

**Using Symbolic Data Analysis to Detect Fraud, Waste, and Abuse  
in Healthcare Insurance Claims Data**

by

Frederick Jonathan Reynolds

A dissertation submitted to the Graduate Faculty of  
Auburn University  
in partial fulfillment of the  
requirements for the Degree of  
Doctor of Philosophy

Auburn, Alabama  
May 2, 2020

Keywords: healthcare, insurance claims, symbolic data analysis,  
insurance fraud, outlier detection

Copyright 2020 by Frederick Jonathan Reynolds

Approved by

Alice E. Smith, Chair, Professor of Industrial and Systems Engineering  
Nedret Billor, Professor of Mathematics and Statistics  
Mark Clark, Professor of Systems and Technology  
Jeffrey S. Smith, Professor of Industrial and Systems Engineering  
David A. Umphress, Professor of Computer Science and Engineering

## ABSTRACT

As the nation's cost of healthcare continues to escalate, so does the exposure to instances of fraud, waste, and abuse. Multiple approaches to detecting these occurrences are in practice today and there are multiple organizations, public and private, which are focused on their identification and reduction. This dissertation investigates symbolic data analysis (SDA) and its applicability to detecting anomalous behavior. SDA is a growing field of study and has implications far beyond what this dissertation will cover. However, driven by the idea that "distributions are the numbers of the future" [1], the core concepts of SDA provide a foundation from which to develop an alternative approach to analyzing healthcare insurance claims data for the presence of anomalous events. The research introduces a symbolic method that investigates data at a higher concept level as opposed to the traditional line level at which most analyses are performed. Simulated datasets and real-world inspired datasets are studied and results between symbolic and centroidal approaches are compared. Results suggest that symbolic anomaly detection techniques perform equally as well as their classical centroidal counterparts when only changes in mean distinguish one set of data from another. When changes are more subtle, particularly when means are equal but the underlying shapes of the distributions are different, the symbolic approach excels. Using the foundational principles of SDA, this dissertation introduces a novel technique to anomaly detection and provides an alternative way of analyzing healthcare insurance claims data for fraud, waste, and abuse.

## ACKNOWLEDGEMENTS

I would like to thank all the people who contributed in some way to the work described in this dissertation.

First and foremost, I would like to express my deep gratitude to my academic advisor, Dr. Alice E. Smith, for providing me with the encouragement necessary to embark on and complete this journey of academic growth. Thank you for your interest in my work and the valuable insights provided along the way. Additionally, I would like to thank my committee members, Dr. Nedret Billor, Dr. Mark Clark, Dr. Jeffrey S. Smith, and Dr. David A. Umphress for agreeing to be on my committee and for providing much needed guidance.

I would like to thank my work colleagues who have encouraged me over the course of my career. The knowledge and experience that I have gained from you helped inform my work and is reflected in the pages of this manuscript.

I'm extremely grateful to my family, including my grandparents, who dedicated their lives to education and instilled the importance of learning in me. Thank you, Aunt Nancy, for helping me with my mathematics homework in middle school and for proofreading this manuscript. Thank you, Mom, for always being there for me. Thanks Dad – I would not have completed this without your support and encouragement. Thank you, Allison, Melissa, and LeighAnn, for your unending support of me. The three of you inspire me every day.

Finally, I'd like to acknowledge my wife, Sissy, for her understanding, love, and patience throughout this project. Thank you. This dissertation is dedicated to you.

# TABLE OF CONTENTS

ABSTRACT.....	ii
ACKNOWLEDGEMENTS.....	iii
LIST OF TABLES.....	vi
LIST OF FIGURES.....	x
LIST OF ABBREVIATIONS.....	xiii
CHAPTER 1: INTRODUCTION.....	1
1.1 Motivation and Contribution.....	4
1.2 Research Objectives.....	5
1.3 Research Methods.....	6
1.4 Limitations of the Research.....	7
1.5 Organization of the Dissertation.....	8
CHAPTER 2: LITERATURE REVIEW.....	9
2.1 Current State of Fraud, Waste, and Abuse in Healthcare.....	9
2.2 Prevalent Methods of Detecting Anomalies in Large Datasets.....	23
2.3 History and Application of Symbolic Data Analysis.....	32
2.4 Cluster Analysis and Symbolic Data Analysis.....	40
2.5 Histogram Binning Methods.....	47
2.6 Practical Application of Symbolic Data Analysis to Large Datasets.....	50
CHAPTER 3: METHODOLOGY AND APPROACH.....	52
3.1 Steps of the Approach.....	52
3.2 Simulated Dataset One.....	53
3.2.1 Selecting a Histogram Binning Method.....	79

3.2.2	Multicollinearity and the Symbolic Approach.....	81
3.2.3	Developing the R Code.....	86
3.3	Simulated Dataset Two .....	89
3.4	Simulated Dataset Three .....	93
3.5	The Iris Flower Dataset .....	98
3.6	Verification and Validation .....	104
3.7	Chapter Summary and Observations .....	107
CHAPTER 4: HEALTHCARE DATASET APPLICATION .....		110
4.1	Ambulance Claims .....	111
4.2	Establishing a Baseline.....	118
4.3	Case 1 – Excessive Mileage .....	131
4.4	Case 2 – Inappropriate Transport Levels .....	140
4.5	Case 3 – Excessive Number of Transports per Beneficiary .....	148
4.6	Case 4 – Incorrect Rate Identification.....	152
4.7	Case 4R - Incorrect Rate Identification Revised .....	155
4.8	Chapter Summary and Observations .....	160
CHAPTER 5: CONCLUSIONS AND PROPOSED FUTURE RESEARCH.....		164
5.1	Conclusions .....	164
5.2	Proposed Future Research.....	166
REFERENCES .....		168

## LIST OF TABLES

Table 1.1:	Classical Data Table [6].....	3
Table 1.2:	Symbolic Data Table.....	3
Table 2.1:	Estimated Sources of Excess Costs in Healthcare [7], [8].....	11
Table 2.2:	Levels of Healthcare Fraud Control [24] .....	22
Table 2.3:	CMS Fraud Detection Model Types [17] .....	24
Table 2.4:	Fraud Detection Types [31] .....	27
Table 2.5:	Individual Matrix [58].....	34
Table 2.6:	Gender Matrix.....	34
Table 2.7:	Instruction Level Matrix .....	35
Table 2.8:	Instruction Level Matrix - Symbolic.....	36
Table 2.9:	Classical Data Table .....	37
Table 2.10:	Symbolic Data Table.....	37
Table 2.11:	Symbolic Data Type Definitions .....	39
Table 2.12:	Binning Methods and Formulas.....	47
Table 2.13:	Results of Binning Calculations.....	48
Table 3.1:	Simulated Dataset One.....	54
Table 3.2:	Simulated Dataset One Descriptors .....	54
Table 3.3:	Simulated Dataset One Bin Results .....	57
Table 3.4:	Bin Representation for V1 .....	58
Table 3.5:	Bin Representation for V2 .....	59
Table 3.6:	Bin Representation for V3 .....	59
Table 3.7:	Pairwise Bin Distances for V1 .....	61

Table 3.8:	Pairwise Bin Distances for V2.....	62
Table 3.9:	Pairwise Bin Distances for V3.....	62
Table 3.10:	Dissimilarity Matrix.....	63
Table 3.11:	Scaled Dissimilarity Matrix .....	64
Table 3.12:	Alpha Value Sensitivity – Simulated Dataset One (Symbolic) .....	69
Table 3.13:	Alpha Value Sensitivity – Simulated Dataset One (Centroidal).....	70
Table 3.14:	Dissimilarity Matrix with AVG DIST Calculation.....	71
Table 3.15:	Threshold Determination Calculation.....	71
Table 3.16:	Simulated Dataset One – Centroidal Approach .....	72
Table 3.17:	Pairwise Distance Table – Centroidal Approach .....	73
Table 3.18:	Scaled Dissimilarity Matrix with AVG DIST Calculation - Centroidal...	73
Table 3.19:	Threshold Determination Calculation - Centroidal.....	74
Table 3.20:	Simulated Dataset One with Categorical Variable Added.....	76
Table 3.21:	Pairwise Distance Table with Categorical Value Added.....	76
Table 3.22:	Dissimilarity Matrix with Categorical Variable Added.....	76
Table 3.23:	Threshold Determination Calculation with Categorical Variable Added.	77
Table 3.24:	Simulated Dataset One with Categorical Variable Added - Centroidal ...	78
Table 3.25:	Pairwise Distance Table with Categorical Value Added - Centroidal.....	78
Table 3.26:	Dissimilarity Matrix with Categorical Variable Added - Centroidal.....	78
Table 3.27:	Threshold Determination Calculation - Centroidal.....	78
Table 3.28:	Binning Methods Compared .....	80
Table 3.29:	Five Group Five Variable Datasets (V and C).....	82
Table 3.30:	Correlation Tables.....	84

Table 3.31:	Symbolic Distance Matrix for Dataset without Correlation .....	85
Table 3.32:	Symbolic Distance Matrix for Dataset with Correlation .....	85
Table 3.33:	Cluster Coordinates - Separate.....	89
Table 3.34:	Three Cluster Dissimilarity Matrix - Centroidal.....	90
Table 3.35:	Three Cluster Dissimilarity Matrix - Symbolic .....	91
Table 3.36:	Cluster Coordinates - Overlaid .....	91
Table 3.37:	Three Cluster Overlaid Dissimilarity Matrix - Centroidal.....	92
Table 3.38:	Three Cluster Overlaid Dissimilarity Matrix - Symbolic .....	93
Table 3.39:	Bloom Set Color Designations.....	103
Table 4.1:	OIG Measures of Questionable Billing [75].....	112
Table 4.2:	Ambulance Claims Data Definitions .....	115
Table 4.3:	Service Code Definitions [76] .....	116
Table 4.4:	Destination and Origin Code Definitions .....	117
Table 4.5:	Baseline Data .....	119
Table 4.6:	Baseline Data Descriptors.....	119
Table 4.7:	Baseline Case Distance Measures - Centroidal .....	127
Table 4.8:	Baseline Case Distance Measures - Symbolic .....	129
Table 4.9:	Performance Metrics for Baseline Dataset .....	129
Table 4.10:	Results Table – Case 1 .....	135
Table 4.11:	Performance Measures – Case 1 .....	136
Table 4.12:	Provider A Statistics with a 400% Increase in Miles.....	139
Table 4.13:	Provider A Statistics with a 50% Increase in Miles.....	139
Table 4.14:	Results Table – Case 2.....	143



Table 4.15:	Performance Measures – Case 2 .....	144
Table 4.16:	Provider A Statistics with a 50% Upcode Action.....	147
Table 4.17:	Provider A statistics with a 10% Upcode Action.....	147
Table 4.18:	Results Table – Case 3.....	151
Table 4.19:	Performance Measures – Case 3.....	152
Table 4.20:	Results Table – Case 4.....	154
Table 4.21:	Performance Measures – Case 4 .....	155
Table 4.22:	Results Table – Case 4R Revised .....	157
Table 4.23:	Performance Measures – Case 4R Revised .....	158
Table 4.24:	Case 4R Binning Results .....	159
Table 4.25:	Case Summary Table .....	163

## LIST OF FIGURES

Figure 2.1:	Sources of Waste in American Healthcare [7], [8] .....	10
Figure 2.2:	Overview of the Process for Fraud Prevention [17] .....	16
Figure 2.3:	Categories of Research Regarding Fraud [22].....	20
Figure 2.4:	Relevant Levels of Categorization Beyond the Individual Claim [23].....	21
Figure 2.5:	Anomalies in a Two-Dimensional Dataset [45].....	29
Figure 2.6:	Symbolic Data Types [58] .....	38
Figure 2.7:	Steps in the Clustering Process [60] .....	41
Figure 2.8:	Agglomerative Versus Divisive Clustering [60].....	43
Figure 2.9:	Dendrogram .....	44
Figure 2.10:	Histogram using Sturges Binning Rule.....	49
Figure 3.1:	Boxplot of Simulated Dataset One .....	55
Figure 3.2:	Alpha using $Q3 + 1.50 * IQR$ .....	66
Figure 3.3:	Alpha using $Q3 + 0.72 * IQR$ .....	66
Figure 3.4:	Distribution of Simulated Dataset One – Categorical Variable.....	75
Figure 3.5:	Five Group Five Variables Without Correlation .....	83
Figure 3.6:	Five Group Five Variables With Correlation.....	83
Figure 3.7:	Paired T-Test of Pairwise Distances .....	86
Figure 3.8:	Screenshot of R Code Report Display .....	88
Figure 3.9:	Three Cluster Scatterplot - Separate .....	90
Figure 3.10:	Three Cluster Scatterplot - Overlaid .....	92
Figure 3.11:	Ten Groups, Random Only .....	94
Figure 3.12:	Ten Groups with Variability Change Across Two Groups.....	95

Figure 3.13:	Ten Groups Introducing Variability Change .....	96
Figure 3.14:	Ten Groups Introducing Mean Change.....	97
Figure 3.15:	Ten Groups with an Asymmetrical Distribution.....	98
Figure 3.16:	3-D Iris Scatterplot.....	99
Figure 3.17:	Iris R Code Centroidal Results .....	100
Figure 3.18:	Iris R Code Symbolic Results .....	100
Figure 3.19:	Distribution of Bloom Color .....	102
Figure 3.20:	Symbolic Table with Bloom Color Added (Modal, Categorical).....	102
Figure 3.21:	Symbolic Table with Bloom Color Added (Set, Categorical) .....	103
Figure 3.22:	R Code Centroidal Results.....	105
Figure 3.23:	R Code Symbolic Results .....	105
Figure 3.24:	R Code Centroidal Results with Categorical Variable .....	106
Figure 3.25:	R Code Symbolic Results with Categorical Variable.....	106
Figure 4.1:	Ambulance CMS 1500 Claim Form .....	114
Figure 4.2:	Graphical Summary of Age .....	121
Figure 4.3:	Pareto Chart of AutoAcc.....	122
Figure 4.4:	Pareto Chart of Gender .....	122
Figure 4.5:	Pareto Chart of Service Code.....	123
Figure 4.6:	Pareto Chart of Origin Code .....	124
Figure 4.7:	Baseline Case Distance Matrix - Centroidal.....	126
Figure 4.8:	Baseline Case Distance Matrix - Symbolic .....	128
Figure 4.9:	Graphical Summary of Miles Traveled from Residence to Hospital.....	132
Figure 4.10:	Graphical Summary of Excessive Mileage.....	133

Figure 4.11: Excessive Mileage with 50% Increase.....	138
Figure 4.12: Pareto Chart of Service Codes .....	140
Figure 4.13: Pareto Chart of Service Codes - Modified.....	141
Figure 4.14: Ten Percent of Volume Upcoded.....	146
Figure 4.15: Screenshot of Simulated Patient with Multiple Trips .....	150

## LIST OF ABBREVIATIONS

ACA	Affordable Care Act
AKS	Anti-Kickback Statute
CMS	Centers for Medicare and Medicaid Services
EDA	Exploratory Data Analysis
FCA	False Claims Act
FPS	Fraud Prevention System
IQR	Interquartile Range
FWA	Fraud, Waste, and Abuse
OIG	Office of Inspector General
SDA	Symbolic Data Analysis
SODAS	Symbolic Official Data Analysis System
SQL	Structured Query Language

# CHAPTER 1

## INTRODUCTION

In 2018, the United States spent \$3.6 trillion on healthcare equaling \$11,172 per person [2]. The National Healthcare Anti-Fraud Association report that some government and law enforcement agencies believe as much as 10% of that spent is due to fraud, waste, and abuse (FWA) [3]–[5]. That amount includes claims for services not needed and not performed. Because most states require claims to be paid within a few weeks after submittal, many payers do not have the resources nor technology to discover and investigate claim invoicing and payment discrepancies in a timely manner. As healthcare spending has increased, so has the volume, variety, and the near real-time availability of the data that describes these payment transactions. Information is becoming available from multiple sources that, if integrated, could provide greater insight into the identification of FWA within the claims payment process. Additionally, the systems that store this data are becoming more scalable, cost effective, and reliable. The confluence of systems accessibility and data availability provides an opportunity to deploy advanced analytical techniques in order to detect discrepancies in bills submitted by and paid to healthcare providers.

While many systems and methodologies have been deployed to detect anomalies in large datasets, this dissertation investigates symbolic data analysis (SDA) and its applicability to such challenges. The premise of SDA analysis is that it summarizes large datasets into higher level classes, or concepts, and then allows for the application of traditional statistical tools on these new classes. This is helpful when the unit of interest is at the concept level as opposed to individual level. For example, when comparing billing

practices of healthcare providers, the unit of interest is not at the individual claim level, but rather the provider level. Performing analyses at the higher order level creates insight that is not apparent at the individual unit level.

This dissertation is significant because it applies a unique approach to improving payment integrity in the healthcare payer environment. The advent of electronic medical records is adding to the vast amount of data available to be analyzed and new approaches must be devised to help with the timely and accurate processing of this data. SDA may provide an alternative way of viewing this data particularly as it applies to accurate payment processing and could enable an entirely new approach to dealing with the issue of detecting instances of FWA.

The purpose of SDA is to extend data mining techniques and traditional statistics to higher level units. When the units become classes, the application of traditional techniques such as clustering, principal component analysis, and regression are still valid but must take on different forms. The following example helps to explain the difference between classical data and symbolic data [6]. Table 1.1 shows a standard data table of soccer players that includes three numerical and two categorical variables.

Table 1.1: Classical Data Table [6]

<b>Player</b>	<b>Team</b>	<b>Age(yr.)</b>	<b>Weight(kg.)</b>	<b>Height(m.)</b>	<b>Nationality</b>
Fernandez	Spain	29	85	1.84	Spanish(Sp.)
Rodriguez	Spain	23	90	1.92	Brazilian(Br.)
Mballo	France	25	82	1.90	Senegalese(Se.)
Zidane	France	27	78	1.85	French(Fr.)

While it may be interesting to examine the makeup of each individual player, it may be of greater interest to understand this data from the perspective of the higher order class of “Team.” For example, a metric of interest may be goals scored in World Cup matches, which is more relevant at the higher order level. Table 1.2 is an example of the symbolic table that can be constructed.

Table 1.2: Symbolic Data Table

<b>Team</b>	<b>Sample Size</b>	<b>Age (yr.)</b>	<b>Weight (kg.)</b>	<b>Height (m.)</b>	<b>Nationality</b>	<b>Goals Scored</b>
Spain	2	[23,29]	[85,90]	[1.84,1.92]	(0.5 Sp.,0.5 Br.)	18
France	2	[25,27]	[78,82]	[1.85,1.90]	(0.5 Fr.,0.5 Se.)	24

The resulting table describes values within the individual cells that are no longer quantitative in nature (age, weight, height) but are instead sets. The variability of nationality is no longer categorical but instead an expression of the frequency of nationalities that occur on the team. SDA is the practice of grouping and summarizing data at the unit level of interest and then extending the principles of traditional statistical



analysis (e.g., clustering, decision trees, factor analysis, regression) to the higher-order data.

The problem of FWA is a costly one in healthcare. Many detection approaches have been implemented with varying degrees of success and while traditional cluster analysis is an obvious choice when it comes to anomaly detection, SDA, and its clustering counterpart may provide new insight into this particular challenge.

## **1.1 Motivation and Contribution**

Most approaches to anomaly detection focus at the raw data level. However, many times the information of interest does not reside at that granular level. If that is the case, then a better approach of analysis would involve a technique that evaluates information at a higher order level, or concept level. For example, a fraud detection system that intends to identify aberrant providers should be looking for outliers at the provider level as opposed to a sub-level of that category. The motivation behind this dissertation is to determine if a symbolic data approach to clustering can provide that type of analysis and in turn be a viable alternative to detecting FWA in our healthcare system.

This research presents a new approach to FWA detection using SDA. Simulated datasets are generated to demonstrate the application of SDA and how it compares to traditional data analysis. The approach is then applied to the readily available and frequently evaluated Iris flower dataset. Additional labels and classes are added and assessed to determine the efficacy of the SDA approach and its ability to distinguish multiple classes within the data. This concept of group identification is then translated to a larger healthcare dataset in order to determine how well it can segregate one class of

providers from another, including the determination of anomalous activity from usual activity. The primary contribution of this research is the development of an alternative approach to unsupervised anomaly detection and its specific application to a large healthcare insurance claims dataset.

## **1.2 Research Objectives**

The limitations that could exist through traditional analytic techniques present an opportunity to study SDA as a viable alternative. This dissertation seeks to answer the following questions:

- How does an analytic approach using symbolic data compare to centroidal methods (arithmetic mean of all points in the sample) in terms of accuracy and ease of implementation?
- What is an appropriate measure that can be used to identify the existence of an outlier group when using symbolic data analysis?
- What are the effects of different binning approaches as they apply to symbolic data analysis?
- How does symbolic data analysis treat multivariate datasets when the input variables contain categorical and continuous data?
- How is symbolic data analysis impacted when label information is applied at the concept level and how does it compare to the application of traditional label information?
- Can symbolic data analysis provide a viable alternative for the detection of fraud, waste, and abuse in a large healthcare insurance claims dataset?

### 1.3 Research Methods

The objective of this dissertation is to evaluate the benefits and practicality of deploying an SDA approach to detect FWA in a healthcare insurance claims dataset. The dissertation begins by defining the steps taken in applying this approach to all datasets. Once described, this same approach is applied to simulated datasets, the publicly available Iris flower dataset, and several real-world inspired healthcare datasets to determine the method's efficacy. Finally, guidelines are developed for the use and interpretation of SDA's application to healthcare datasets.

The research begins by experimenting with simulated datasets where their descriptive nature is known. The foundation of the approach is established here with explanation into how the SDA formulas are derived and how they are applied to the data. An evaluation metric is introduced in this section which allows for the comparison between an SDA approach and a centroidal approach. The method for combining categorical and continuous data is shown along with the computations that can easily be evaluated in spreadsheet form including the construction of distance matrices for both centroidal and SDA methods. A discussion regarding binning techniques is also included.

The research continues with an introduction to the code that was developed in R that will serve as the vehicle for validating the previously mentioned calculations while enabling the study of larger, more complex datasets. The publicly available Iris flower dataset is presented along with discussion of the benefits realized when the user applies additional labeled data at the group level. Contrary to most studies of the Iris flower dataset, this is not an exercise in unsupervised learning but rather a demonstration of how

SDA scores known groups of data and how that effect can be enhanced through augmentation of the data that can only be applied at the group level.

Using the procedures developed for the simulated and Iris flower datasets, the same suite of calculations and analyses are applied to several other datasets in order to determine its effectiveness in identifying anomalous behavior. FWA anomalies are simulated based on the most common events reported in the literature. These datasets also include continuous and categorical data and involve the construction of the appropriate distance matrices. The SDA algorithm is applied to this data to determine its ability to identify anomalous events should they exist. The developed evaluation metric is applied to the results to determine model effectiveness. Traditional centroidal analysis is also performed on the data to serve as a comparison.

The third phase of this research focuses on the impact that SDA can have in the healthcare insurance claims payer environment. Multiple scenarios are tested and the approach's accuracy is assessed and compared to the classical centroidal method.

#### **1.4 Limitations of the Research**

This dissertation studies the viability of an alternate approach to anomaly detection in healthcare datasets using symbolic datasets. In practice these datasets are often large in scale and are inherent with errors and omissions that add to the complexity of the analysis. This research does not include the means and methods required to scrub existing data for any of these problems and leaves the necessary data preparation and feature engineering required in any study of this type solely up to the discretion of the future adopter of this approach.

## **1.5 Organization of the Dissertation**

This dissertation is organized in five chapters. Chapter 1 includes the background and motivation for this dissertation including the objectives to be addressed and the approaches taken. Chapter 2 comprises the literature review. The current state of FWA in the United States healthcare system is reviewed as well as current methods being deployed that are intended to discover anomalies in this type of data. SDA is introduced in this section as well as a discussion of clustering analysis and its relevance to the topic. The chapter concludes with a discussion of current applications to healthcare data using SDA. The methodology and approach are documented in Chapter 3. Items discussed in this chapter include SDA calculations, a non-parametric evaluation score, multicollinearity, scaling and standardization, the introduction of categorical and continuous data, binning approaches and their effect on histogram-valued data. Also described is the R code which was developed to verify and validate the methods being proposed and to test more complex instances that would be difficult to evaluate in a traditional spreadsheet environment. Multiple simulated examples are presented in Chapter 3 that demonstrate the differences between centroidal and symbolic approaches to anomaly detection. Chapter 4 describes the application of the methodology developed to healthcare datasets. Five specific cases are reviewed which are intended to mimic scenarios that could exist in real-world situations. Chapter 5 summarizes the dissertation and recommends areas for future research.

## **CHAPTER 2**

### **LITERATURE REVIEW**

In order to fully and completely explore the application of SDA to a healthcare insurance claims dataset, literature was selected to be reviewed based on its relevance to the following questions:

- What is the current state of fraud, waste, and abuse in today's healthcare environment?
- What are the most prevalent methodologies in use today to detect this behavior and how effective are they?
- What is the history of symbolic data analysis?
- What is cluster analysis and have techniques been developed that combine symbolic data analysis and clustering?
- Has symbolic data analysis been applied to detect fraud in a healthcare insurance claims dataset?

It is the intent of this literature review to highlight previous work in these areas and to provide insight into future areas of study.

#### **2.1 Current State of Fraud, Waste, and Abuse in Healthcare**

The most reliable source of information regarding healthcare expenditures in the United States comes from the Centers for Medicare & Medicaid Services (CMS). In 2018, the nation spent \$3.6 trillion on healthcare expenditures which accounted for 17.7% of the nation's gross domestic product [2]. It is estimated that number will grow at a rate of 5.5% per year and reach \$6.0 trillion spent annually by 2027 [2]. Additionally, studies suggest

that the expenditures in this category do not always go directly to patient care. In 2009, the Institute of Medicine convened a roundtable of experts who concluded that nearly one third of the healthcare costs by the end of that year were considered waste [7]. Figure 2.1 and Table 2.1 summarize the sources of waste in terms of dollars.

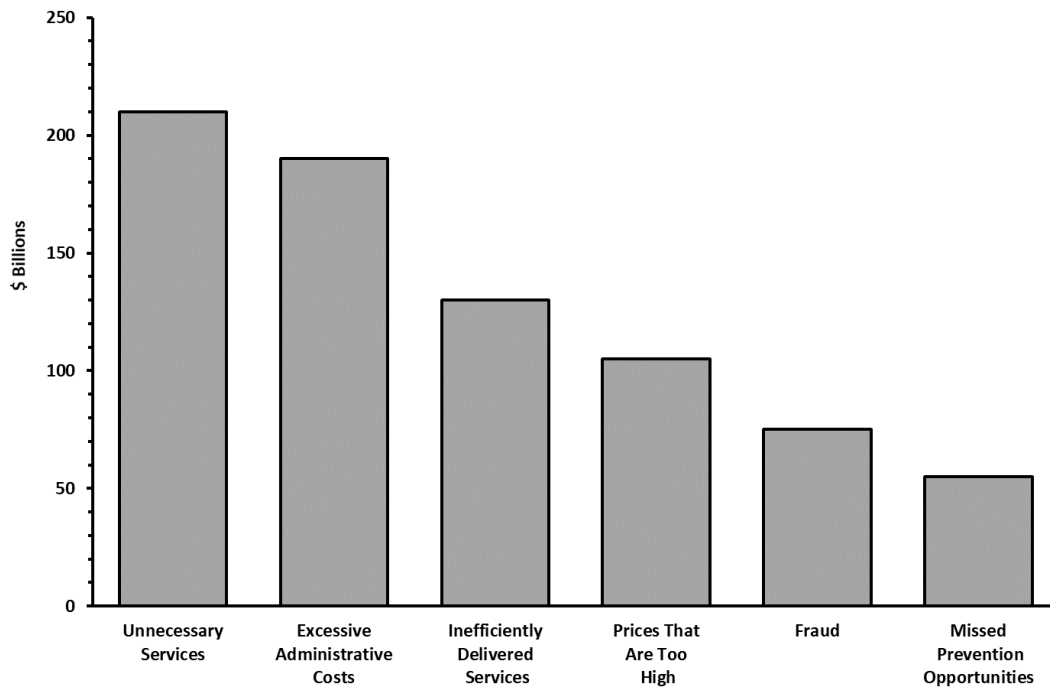


Figure 2.1: Sources of Waste in American Healthcare [7], [8]

Table 2.1: Estimated Sources of Excess Costs in Healthcare [7], [8]

<b>Category</b>	<b>Sources</b>	<b>Estimate of Excess Costs</b>
Unnecessary Services	<ul style="list-style-type: none"> <li>• Overuse – beyond evidence-established levels</li> <li>• Discretionary use beyond benchmarks</li> <li>• Unnecessary choice of higher-cost services</li> </ul>	\$210 billion
Inefficiently Delivered Services	<ul style="list-style-type: none"> <li>• Mistakes – errors, preventable complications</li> <li>• Care fragmentation</li> <li>• Unnecessary use of higher-cost providers</li> <li>• Operational inefficiencies at care delivery sites</li> </ul>	\$130 billion
Excess Administrative Costs	<ul style="list-style-type: none"> <li>• Insurance paperwork costs beyond benchmarks</li> <li>• Insurers’ administrative inefficiencies</li> <li>• Inefficiencies due to care documentation requirements</li> </ul>	\$190 billion
Prices That Are Too High	<ul style="list-style-type: none"> <li>• Service prices beyond competitive benchmarks</li> <li>• Product prices beyond competitive benchmarks</li> </ul>	\$105 billion
Missed Prevention Opportunities	<ul style="list-style-type: none"> <li>• Primary prevention</li> <li>• Secondary prevention</li> <li>• Tertiary prevention</li> </ul>	\$55 billion
Fraud	<ul style="list-style-type: none"> <li>• All sources – payers, clinicians, patients</li> </ul>	\$75 billion

In their 2013 National Training Program, CMS provided guidance regarding Medicare and Medicaid fraud and abuse prevention, detection, recovery, and reporting. CMS defines fraud as the intentional falsification of information [9]. Common types of fraud committed include billing for services not rendered, charging for more expensive



services than services performed (upcoding), performing and charging for services not necessary, misrepresenting treatments as covered by insurance, falsifying diagnoses, and unbundling procedure codes and billing separately [10].

Abuse occurs when sound medical practices are not followed which leads to unnecessary costs, improper payment, or services that are not necessary. Much of the excess costs in the system are due to mistakes and inefficiencies that exist within the current systems. Improper billing, improper payment, and errors in processing contribute heavily to these costs. Additionally, in many cases, waste, and abuse do evolve into fraud. For the purpose of brevity in this dissertation, all instances of fraud, waste, and abuse will subsequently be labeled as fraudulent.

Detection of fraudulent claims can be particularly difficult. Most payers of healthcare services receive claims for services electronically and are required to pay claims in a timely manner per regulatory guidelines. Morris [11] suggested this leads payers to operate in what is considered a “pay and chase” model where payments are made assuming the information represented on the claim is correct. Standard code edits are applied as necessary but very little verification can be done to ensure that the services billed were correct and appropriate [11]. Claims are paid as individual entities and it often requires post-pay analysis for a payer to determine whether a series of claims analyzed together can identify fraudulent billing practices [11].

As a deterrent to fraudulent behavior, the federal government has enacted several laws which are intended to specify criminal and/or civil remedies that can be imposed upon any entity which engages in this behavior [12]. These laws include:

- False Claims Act (FCA);
- Anti-Kickback Statute (AKS);
- Physician Self-Referral Law (Stark Law);
- The Exclusion Statute;
- The Civil Monetary Penalties Law.

The FCA is the government's primary tool to combat fraud against the government. The Act imposes civil liability on any person or entity who knowingly requests reimbursement through a claim for services that are known to be false, where "knowing" is intended to imply deliberate ignorance and/or reckless disregard [12]. An example of this behavior includes submitting a claim for a higher amount than the services that were rendered. Those convicted can be subject to civil penalties that include fines as well as criminal prosecution [12].

The AKS is a law intended to prevent any business or entity from providing any form of reward, payment, or reimbursement in exchange for a recommendation of products or services. It is interesting to note that in some industries, it is acceptable to reward those who provide business referrals. An example of violating the AKS would be a medical services provider accepting below market rates for medical office space in return for referrals to their facility. Penalties may be civil and criminal in nature [12].

The Physician Self-Referral Act, also known as the Stark Law, prevents physicians from referring patients to facilities where that physician may have a financial interest. An example of this may be a physician who has ownership or an investment interest in a radiology center referring a patient to that center when other facilities may be better suited

to deliver care. Penalties are civil in nature and could result in exclusion from participation in all federal healthcare programs [12], [13].

The Exclusion Statute requires those entities convicted of criminal offenses such as Medicare fraud, patient abuse, or illegal distribution of controlled substances be banned from participation in federal healthcare programs. The exclusion list is managed by the Office of the Inspector General (OIG) [12].

The Civil Monetary Penalties Law outlines monetary penalties that are enforced by the OIG. Penalties can be as high as \$50,000 per violation and can include presenting a claim that is known or should be known as false; presenting a claim that is known or should be known that Medicare will not pay; and any violation of the AKS [12].

In addition to enacting legislation to deal with fraudulent activities, the federal government has instituted several key programs. The Affordable Care Act (ACA), signed into law in 2009, includes new provisions for fighting fraud including: tougher sentencing guidelines for those that break the law, enhanced screening for high risk providers and suppliers, advanced fraud detection technology (discussed in the next section) deployed by the CMS, and increased funding over the next ten years to improve anti-fraud efforts.

The Department of Justice and the Department of Health and Human Services jointly created the Healthcare Fraud Prevention and Enforcement Action Team. Strike Force teams focus on “hot spot” locations in the country and identify and apprehend healthcare fraudsters. These teams work with members from the Office of Inspector General, Department of Justice, Federal Bureau of Investigation, and local law enforcement to convict those guilty of healthcare fraud [14].

In 2012, the CMS Program Integrity Command Center was opened which enabled clinicians, data analysts, fraud investigators, and policy experts from multiple government agencies to work collaboratively in one location to quickly develop and deploy advanced fraud detection techniques [15]. The Center's mission is to protect Medicare and Medicaid programs and improve the integrity of the healthcare system through four program areas which include prevention, detection, recovery, and transparency and accountability [16]. The Small Business Jobs Act of 2010 was signed into law to create lending programs that would increase the availability of credit to small businesses. One of the requirements of the law was to further develop and deploy predictive modeling capability to identify FWA in the healthcare system in order to better protect the American taxpayer. Subsequently, in June of 2011, the CMS launched the Fraud Prevention System (FPS). The FPS is a streaming, national service which directs its efforts toward all Medicare Fee-For-Service claims [17]. The system uses predictive models and algorithms to proactively review claims for suspect activity – then prioritizes leads for review and investigation. The result of these efforts prevents public funds from being sent to suspect providers and suppliers [18]. In one example, the FPS identified a provider that was exhibiting high risk billing patterns. An investigative team was sent to the provider location for an unannounced site visit to conduct interviews and review medical records. It was determined that the provider was billing Medicare for services performed by unqualified medical aides. The provider was removed from the Medicare program which prevented \$700 thousand of inappropriate payments [17]. An overview of the Process for Fraud Prevention as defined by CMS is highlighted in Figure 2.2.

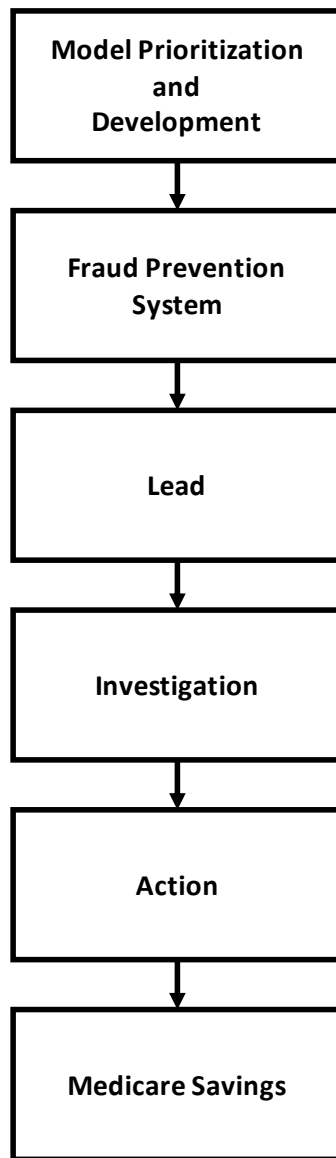


Figure 2.2: Overview of the Process for Fraud Prevention [17]

Model Prioritization and Development refers to the process of determining which type of model (rule based, anomaly detection, prediction, network) best fits the situation and which should be prioritized and deployed. The FPS is the new approach to monitoring activity through the continuous and simultaneous review of relevant data. Results from this approach are used to generate leads and/or highlight potential suspicious activity,

which may lead to investigation, recoupment of funds and, ultimately, savings to the Medicare Program [17].

It is important to note that the tools and technologies used in the detection of fraud are intended to serve as aids in a broader investigation. As indicated in the figure above, the detection of these behaviors, regardless of method, tends to inform an investigation but are rarely enough to be considered in isolation when looking for suspicious activity. When an investigation validates an offense, savings are realized through one or more of the following administrative actions [17]:

- Payment suspension – holds on funds due providers;
- Law enforcement referrals – cases referred to law enforcement for prosecution;
- Overpayment recoveries – seeking refunds from providers where payments exceeded actual amount owed;
- Prepayment edits – contractors review claims before payment is made;
- Auto-denials – computer edits force provider payment denial prior to payment processing;
- Provider revocation – provider’s status precludes them from any form of payment.

In 2013, the FPS was credited with the identification and investigation of 938 suspect providers and suppliers who had patterns of inappropriate billing. The savings associated with these identifications totaled \$210.7 million as certified by the OIG. Planned future enhancements to the FPS tool include [17]:

- Expand and improve models to identify bad actors more quickly and more effectively;
- Expand FPS beyond fraud into waste and abuse;
- Deny claims that are not supported by Medicare policy;
- Identify leads for early intervention by the Medicare Administrative Contractors;
- Evaluate the feasibility of expanding predictive analytics to Medicaid;
- Reduce costs of FPS while applying predictive analytics more effectively and efficiently;
- Share lessons learned and best practices with federal, state, and private partners.

With the enactment of the ACA, seeking efficiencies in the handling of public funds is as important as ever. As noted above, the federal government, through the work of several governmental agencies, plans to continue to develop its competency in the area of FWA detection and prevention.

The private sector has active fraud detection as well and often joins forces with public entities to create solutions. The defense contractor, Northrup Grumman along with Verizon and WellPoint subsidiary National Government Services, were selected by the CMS to develop the FPS previously mentioned. The platform builds on predictive modeling technology used by Verizon to fight fraud. Link, behavioral, and statistical

analysis are used to identify potential fraudulent requests prior to processing [19]. IBISWorld suggests the major commercial providers of fraud detection software and services include ACI Worldwide, FICO and SAS – all of which provide their services globally to public and private organizations. Products include predictive and real-time analytics which support the detection of fraudulent behavior across users, accounts, products, processes, and channels. The fraud detection software industry achieves annual revenues of \$817 million and annual growth of 30.2%. As of 2014, over two hundred businesses were engaged in the work [20]. Many firms work with clients via risk-based pricing and are compensated proportionally with the success of their discoveries. Most focus on approaches best suited to their specific businesses and rarely does one firm dominate the market. The healthcare market is lucrative compared to the markets of telecommunications and finance. While the rate of improper payment in those industries range from 0.1% to 0.2%, it is estimated that the level of improper Medicare payments could be 50-100 times higher [21].

Typically, the parties that commit fraud can be broken into three categories: providers, subscribers, and carriers. Providers include doctors, hospitals, medical equipment providers, ambulance companies, and laboratories. They commit fraud through activities such as billing for services not performed, unbundling services, upcoding, performing medically unnecessary services, and falsification of patient information. Subscribers include patients and patients' employers and they defraud the system by falsifying eligibility records, filing false claims or using other persons' coverage information to illegally claim benefits. Carriers typically refer to insurance companies that pay healthcare insurance claims on behalf of the subscriber or patient and they can falsify



reimbursements or falsify benefit statements [22]. Li et al. suggest that provider fraud is by far the most prevalent and could have the largest impact on the quality and safety of the healthcare system. Li et al. performed a study on the percentage of research papers completed which revealed that “Service Providers” was the number one category of research [22]. Figure 2.3 shows the results of the study.

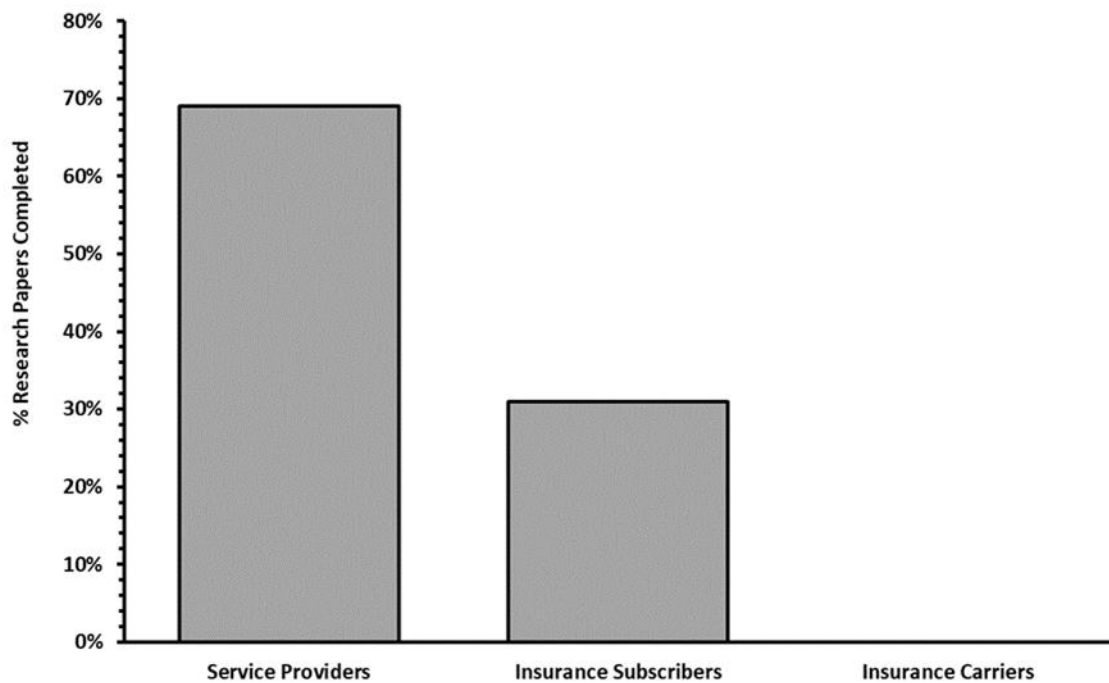


Figure 2.3: Categories of Research Regarding Fraud [22]

Thornton et al. performed a study that further clarified the relationship between patient and providers in the context of understanding fraud. Their work presented several multidimensional models which focused on the importance of evaluating fraudulent activity at levels beyond the individual or claim line level [23].

Figure 2.4 represents a graphic taken from [23] which highlights the critical fields that can apply to most healthcare insurance claims.

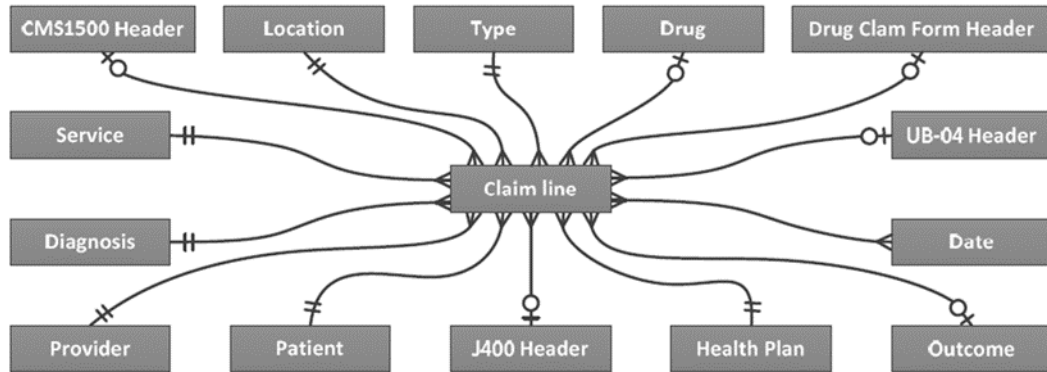


Figure 2.4: Relevant Levels of Categorization Beyond the Individual Claim [23]

Per the notation above, each claim line (of which there could be many) will contain the following singular mandatory fields: patient, provider, diagnosis, service, location, type, and health plan. Multiple dates could be present on each line (e.g., filed, service, paid). Other fields that may exist are singular in nature and may be optional (e.g., headers).

Concepts put forth by Sparrow suggest that the most effective and more challenging fraud detection methods must occur beyond the claim or line level [24]. Sparrow's definitions of these additional levels are cited [23] and displayed in Table 2.2.

Table 2.2: Levels of Healthcare Fraud Control [24]

<b>Level Focus</b>		
Level 1	Single Claim or Transaction	The claim itself and the related provider and the patient.
Level 2	Patient / Provider	One patient, one provider and all their claims.
Level 3	a. Patient	One patient and all its claims and related providers.
	b. Provider	One provider and all its claims and related patients.
Level 4	a. Insurer Policy / Provider	Patients that are covered by the same insurance policy and are targeted by one provider.
	b. Patient / Provider Group	One patient being targeted by multiple providers within a practice.
Level 5	Insurer Policy / Provider Group	Patients with the same policy being targeted by multiple providers within a practice.
Level 6	a. Defined Patient Group	Groups of patients being targeted by providers (e.g., patients living in the same location)
	b. Provider Group	Groups of providers targeting their patients. Groups can be providers within the same practice, clinics, hospitals, or other arrangements.
Level 7	Multiparty, Criminal Conspiracies	Multiparty conspiracies that could involve many relationships.

In summary, the existence of FWA in the healthcare system, regardless of public or private origin, has the effect of increasing the cost and decreasing the quality of care. These inefficiencies in the system artificially drive costs higher which makes affordable care less attainable for many who need it most. New detection techniques, including those introduced by this research, is of interest to both public and private entities. As noted in

the literature, most of the focus regarding fraud occurs at the provider level, yet most detection systems are directed at the transactional claim level. An approach that can provide an alternate way of grouping data and efficiently determine potential aberrant data points is advantageous in combating unnecessary costs. This research will investigate the identification of patterns at a higher level (provider for example) in order to discover incidents of FWA.

## **2.2 Prevalent Methods of Detecting Anomalies in Large Datasets**

Preventing fraud from occurring can be addressed from two different vantage points. The first and most effective is fraud prevention. Preventing a defect from occurring will always be the strongest method of prevention. Fraud detection is the second approach and is an area of study that is continuously evolving. This evolution is often by necessity as the perpetrators of these offenses quickly become familiar with old prevention techniques and begin to seek new tactics with which to attack the system.

Depending on the situation, there may be multiple approaches to detecting fraudulent activity. These approaches can be used independently or in conjunction with each other. The FPS, as introduced in the previous section, focuses on different model types to detect fraud: Rule Based, Anomaly, Predictive, and Network. Table 2.3 is adapted from a CMS graphic depicting these four types along with examples of each [17].

Table 2.3: CMS Fraud Detection Model Types [17]

<b>Model Type</b>	<b>Definition</b>	<b>Example</b>
Rule Based	Filter fraudulent behaviors with rules	Receive a bill containing an identification number that was previously stolen
Anomaly	Detect individual abnormal patterns versus peer group	Receive a bill from a provider with more services billed in a single day than 99% of similar providers of the same service
Predictive	Assess against known fraud cases	Identify a provider that exhibits practices similar to other known fraudulent providers
Network	Discover knowledge through link analysis	Identify a provider that is linked to known fraudulent activity through an address or phone number

Rule based models use previously collected information and known patterns to identify potentially fraudulent activities. One of the problems that exists with rule based approaches is that they tend to quickly become obsolete as perpetrators develop schemes that work around the new “rules.” Anomaly detection models identify occurrences of behavior and compare those incidents to known patterns of activity – then try to determine how different a certain data point or patterns of points need be to be considered “unusual” and worthy of further investigation. Predictive models attempt to look at past cases of known fraud, identify triggering factors that drove the fraudulent behavior, and then attempt to search for similar current conditions which may indicate that fraud would be the expected future outcome. Drawbacks of this method include the limited number of “fraud positive” events from which to draw information from. Social network analysis attempts to link perpetrators through their existing network of relationships. Also referred to as link analysis, the technique evaluates the connections between organizations, transactions, and people in order to discover unusual events.

Palshikar [25] suggests that fraud detection techniques fall into two classes: statistical techniques and artificial intelligence. Statistical techniques include data preprocessing, parameter estimation, model building, time series analysis, and clustering and classification as statistical techniques. Artificial intelligence includes applying data mining to classify and cluster, pattern recognition, machine learning, and neural networks [25].

Bolton [26] focuses on fraud associated with credit card transactions and cites Brause whose database of credit card transactions reveal a fraud rate between 0.1% and 0.2% as compared to the previously stated healthcare fraud rate between 3% and 10% [27]. Bolton and Hand [28] identify two types of statistical fraud detection: supervised and unsupervised. Supervised methods require datasets to be segmented into two classes, fraudulent and non-fraudulent. These require the true knowledge of each of these classes and require that there exist enough samples to populate each of the classes – a situation which is sometimes problematic in fraud detection because known fraudulent events occur infrequently. Typically, in fraud detection problems, the number of legitimate transactions far outweighs the number of fraudulent ones and that can cause misclassification problems. Supervised methods include linear discriminant analysis and logistic discriminant analysis. Rule based methods are also examples of supervised methods that follow the form “if certain conditions exist, then this consequence is enacted.” Unsupervised methods do not require a priori knowledge of the fraudulent labels but instead seek to identify patterns dissimilar from normal activity. Unsupervised methods are employed when no known identifiers exist that could label an event as fraudulent [28].

The most popular unsupervised method is clustering because a priori knowledge of the fraud instance is not required – however this method can perform poorly if improper choices are made when determining the distances between observations. The distance between the points is critical when determining clusters. The most popular distance metric is Euclidean distance – however, its weakness is that it treats each attribute equally in the calculation. This can be problematic when different features are measured on different scales. The Mahalanobis distance is a measure that accounts for this variability [29]. Combining categorical and continuous variables into one good clustering metric can be particularly challenging and problematic, resulting in clusters being formed differently on some variables than on others [26]. Unsupervised methods have attempted to distinguish between local (within group) outliers and overall (global) outliers but it can be difficult to initially define the local domain. Bolton and Hand proposed the concept of Peer Group Analysis to identify local category formation using unsupervised data mining techniques [26]. Gebski et al. proposed a methodology for grouping categorical anomalies [30]. An advantage to using unsupervised methods is that previously undetected types of fraud may be detected whereas supervised methods know only how to identify incidents like those identified. This is a critical concept in insurance fraud as perpetrators are always attempting to stay ahead of the latest detection technique.

Traville [31] performed a review of all fraud detection literature and developed the information in Table 2.4 to outline various fraud detection types and their definitions.

Table 2.4: Fraud Detection Types [31]

Type	Definition	Method	Explanation
Supervised Classification Techniques	Use training sets with prior information on class membership to learn classification patterns	Linear Discrimination	Regression based on a logistic curve
		Support Vector Machines	A kernel method which selects small number of critical boundary instances (support vectors) to construct a separating hyperplane [32]
		Neural Networks	A set of interconnected nodes that imitate the functioning of a brain [33]
		Decision Tree Learning	Methods for building a decision tree for classification
Unsupervised Data Mining Techniques	Do not assume prior class labels of legitimate or fraudulent behavior	Anomaly Detection	Tries to detect outliers that are inconsistent with the remainder of that dataset [34], [35]
		Cluster Analysis	Divide objects into groups (clusters) with objects in a group being similar to one another but dissimilar to the objects in other groups [36]
		Peer Group Analysis	Clusters of similar observations (peer groups) are identified and clustered, subsequently the individual behavior is compared to the cluster's behavior [37]
Statistical Methods	Statistical methods are more model and theory based than data mining methods	Visualization	Allowing users to view the complex patterns or relationships uncovered in the data mining process [38]
		Profiling	Process of modeling the characteristic aspects of the user [39]
		Benford's Law	The distribution of the first-digit number of a lot of natural phenomena like size of companies, telephone lengths, and invoice amounts will have a characteristic non-uniform distribution [40], [41]
Rule Based	Model based on the experience of experts (Bolton et al. 2002b)	Online Analytical Processing	Dynamic ad-hoc multidimensional analysis [42]
		SQL Queries	Queries designed by domain experts

Supervised methods (require prior information to learn classification patterns) include linear discrimination, support vector machines, neural networks, and decision trees.



Unsupervised methods (do not assume prior labels to be classed as fraudulent or non-fraudulent behavior) include anomaly detection, cluster analysis, and peer group analysis. Statistical methods include visualization, profiling, and the application of Benford's Law (also known as the Law of First Digits, which is the finding that the first digit of a series of numbers is more likely to be a lower number and that the distribution of first digits is not uniformly distributed) [43]. Rule based methods include online analytical processing and structured query language (SQL) queries [31].

Real data for testing fraud detection methodologies is often difficult to obtain. Actual data may not contain the specific events that are being tested for or may not have enough positive occurrences. Furthermore, due to the nature of these datasets, companies may be reluctant to share sensitive customer/patient data. Synthetic data offers a range of options that are available to the tester including a higher degree of freedom during testing of the data [44]. The use of synthetic datasets is applicable to both supervised and unsupervised methodologies.

Unsupervised anomaly detection techniques vary but most seek to find instances of outliers. Barnett and Lewis [35] defined an outlier as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data." That is the definition used throughout this dissertation.

Chandola [45] provided the following graphic, represented as Figure 2.5, of outliers in a two-dimensional dataset.

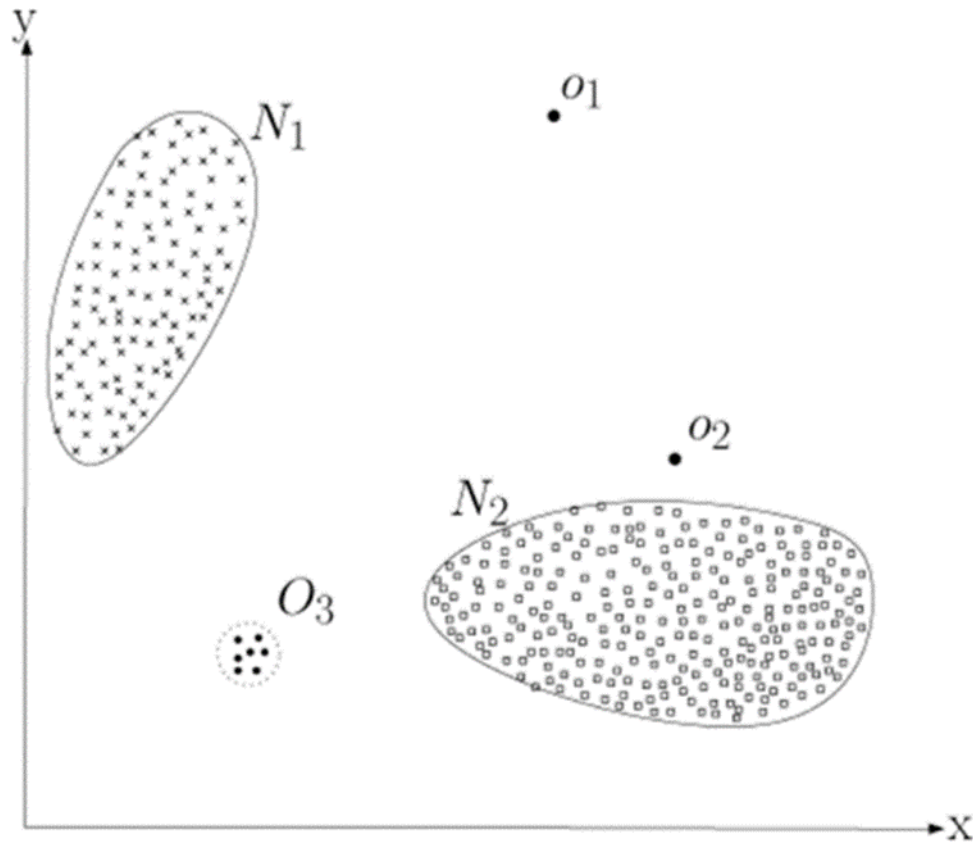


Figure 2.5: Anomalies in a Two-Dimensional Dataset [45]

In Figure 2.5, the areas denoted by  $N_1$  and  $N_2$  would be considered part of a “normal” region while individual points  $O_1$  and  $O_2$  and the set of points labeled  $O_3$  would be considered outliers [45]. The terms anomalies and outliers are often used interchangeably and are used interchangeably in this dissertation.

Chandola [45] discussed the nature of the input data affecting which anomaly detection technique that can be used. The input is usually a set of data instances or records that has corresponding attributes. These attribute types may be binary, categorical, or continuous in nature. Univariate records are those that have only one associated attribute while multivariate records will have more than one attribute – and those attributes can include a mix of attribute types. Most anomaly detection techniques deal with singular record or point data, however, data can be categorized based on existing relationships among the individual instances themselves [46]. Chandola suggested that anomalies may be grouped in the following three categories [45]:

- 1) Point Anomalies – this is the simplest type of anomaly and describes when an individual point can be considered anomalous to the rest of the data.
- 2) Collective Anomalies – describes anomalies that are evident when a collection of instances together is compared to the entire dataset. The individual points themselves may not be anomalous but when grouped together may be [45]. Vatanen explored collective anomalies at greater depth in their application to high energy particle physics [47]. Collective and point anomalies can be transformed to a collective anomaly case through the addition of a context [45].
- 3) Contextual Anomalies – describes data instances that are anomalous within a specific context but may not be otherwise. Contextual anomalies must be records that have an attribute field that describes the context. Many time series datasets conform to contextual anomalies, i.e. outside temperature may be different depending on the time of year. A subfreezing day may be normal in the winter but would be an anomaly if it occurred in the middle of summer. The contextual attribute in this example would be time of year. These events are also referred to as conditional anomalies [48].

He [49] introduced the term semantic outlier which is a point that behaves differently than other data points in the same class, developed an algorithm to cluster

categorical data [50], and introduced a concept called class anomaly detection which is applied to the task of tracking of customer loyalty [51].

Hodge and Austin [52] provided a comprehensive list of applications that rely on outlier detection which includes fraud detection, loan application processing, intrusion detection, activity monitoring, fault diagnosis, and medical condition monitoring. While acknowledging that there is no single solution, it is important that the appropriate technique is suitable for the data presented which will depend on the data type, the existence of labels within the data, the accuracy of the labels, and how outliers are handled once detected [52]. Anomalies are typically reported by using one of two methods. Scores can be applied to anomalous data instances which indicate the likelihood of the record being an outlier. Additionally, labels can be used to assign each instance to a normal or anomalous state [45].

Bolton [28] suggested that “fraud detection is an important area, one in many ways ideal for the application of statistical and data analytic tools, and one where statisticians can make a very substantial and important contribution.”

This dissertation will add to the current literature in the area of contextual anomaly detection. As mentioned in the previous section, the context of provider is often the unit of interest yet techniques using this approach have been limited. Due to the nature of the problem, an unsupervised anomaly detection technique that can incorporate both categorical and numerical data would have many advantages. SDA allows for both data types and is the focus of this dissertation.

### **2.3 History and Application of Symbolic Data Analysis**

Diday and Bock suggested that the origins for SDA was the result of three major influences: exploratory data analysis (EDA), the advent of artificial intelligence, and the concept of numerical taxonomy [53]. EDA was pioneered by the American mathematician John Tukey – most notably by his development of the boxplot in 1969 [54]. EDA suggests a statistical approach that seeks to describe datasets in ways other than the traditional methods of hypothesis testing or formal modeling. His methods ultimately led to advanced computing packages that could describe data, often visually, in ways that could allow for further investigation. Boxplots, histograms, run charts, and scatter plots are common graphical techniques to support EDA and are often used to visually interpret datasets and discover patterns and trends that may not be apparent using classical statistics. It was in this context that Diday suggested the need to extend classical statistical techniques to this new form of descriptive information, specifically as it could be grouped and categorized as symbolic data [53]. Artificial Intelligence was focused on displaying this graphical information in discernable ways but was less focused on being able to explain complex data in simple statistical terms. Diday would go on to develop the Symbolic Official Data Analysis System (SODAS) software platform to address this need. The final influence was Numerical Taxonomy which is a classification system used most frequently in biology where the unit of interest is often at a conceptual level or a “species” level as opposed to the individual unit and uses multiple, equally weighted taxonomic characters (features) to classify groups as similar or dissimilar as opposed to the traditional method of using evolutionary characteristics. The development and advancement of SDA has largely been attributed to two individuals: Edwin Diday and Lynne Billard.

Edwin Diday first introduced the concept of SDA in 1987 while serving as Professor of Computer Science at the University of Paris at Dauphine. He was conducting research into the field of clustering methodologies and realized that by summarizing data within clusters, the inherent characteristics of the data within the cluster are lost. Cluster means were retained but their internal variations were not. This led to the development and study of symbolic data and spawned a new group of students and researchers [55]. Diday has authored or co-authored the three defining texts in this field which include: *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data* [53], *Symbolic Data Analysis and the SODAS Software* [6], and *Symbolic Data Analysis: Conceptual Statistics and Data Mining* [56]. The most recent text, *Symbolic Data Analysis: Conceptual Statistics and Data Mining* [56], was co-authored by Lynne Billard. Billard is a University Professor of Statistics at the University of Georgia and is known for her statistical research, leadership, and advocacy for women in math and science. She has served as president of the American Statistical Association and the International Biometric Society which are the two largest statistical societies in the world. In 2013, she was selected to receive the Florence Nightingale David Award by the Committee of Presidents of Statistical Societies which recognizes female statisticians who exemplify excellence in education, science, and public service [57].

When analyzing a dataset with classical statistics and traditional multivariate analysis, the unit of interest is usually the individual component: an individual entity that can be described by other variables which may be numerical or categorical in nature. Nearly all the descriptions and subsequent data collection activities revolve around this concept. Examples include a person being described by his age, height, weight, hair color,

and eye color. Customers that frequent a business can be described by their sex, age, income amount, educational level, or frequency of trips to the business [58]. In each case, a table can be organized as a data matrix where each cell,  $(i,j)$ , contains a singular value for variable  $j$  which describes individual  $i$ . Table 2.5 is taken from Brito and Noirhomme-Frature [58], and depicts a matrix that describes four individuals, s1-s4. Each individual is assigned two quantitative and two qualitative variables.

Table 2.5: Individual Matrix [58]

<b>Individual</b>	<b># of Children</b>	<b>Weight(kg.)</b>	<b>Gender</b>	<b>Instruction Level</b>
S1	2	52	M	2
S2	1	55	M	3
S3	0	50	M	2
S4	3	60	F	1

While data is commonly displayed and analyzed in this tabular format, the variability and uncertainty that is inherent in the table is often lost when the unit of interest is at a higher group or class level. For example, if the unit of interest above was gender and not the individual, then Table 2.6 can be created to express the characteristics of that classification.

Table 2.6: Gender Matrix

<b>Gender</b>	<b>Sample Size</b>	<b># of Children</b>	<b>Weight(kg.)</b>	<b>Instruction Level</b>
M	3	1	52.3	2.3
F	1	3	60	1

The data could also be presented by the classification of instruction level as shown in Table 2.7.

Table 2.7: Instruction Level Matrix

<b>Instruction Level</b>	<b>Sample Size</b>	<b># of Children</b>	<b>Weight(kg.)</b>	<b>Gender</b>
1	1	3	60	F
2	2	1	51	M
3	1	1	55	M

The example above accurately and simply explains the need for a symbolic data approach to analysis. At the individual level, Table 2.5 is enough when analyzing the dataset at an individual row level. Table 2.6 and Table 2.7 demonstrate the limitation. If we were interested in evaluating the data from a gender perspective, then we could certainly group the data accordingly. However, in doing so, we lose critical information inherent in the original dataset. Simply choosing the average (or mode, or median) for the variables of number of children, weight, and instruction level, we lose information that is relevant to the group variable. Each of those explanatory variables had measures of variability that was lost when summarizing the data at the group level. Table 2.7 illustrates the same limitation if the data were to be analyzed by the variable Instruction Level. A better comparison of individuals by instruction level could be made if we retained the information that was inherent in the original data. SDA provides an approach for analyzing data that considers this inherent variability in the underlying data without an unacceptable loss of information [58].



The literature suggests that this approach is growing in interest in the analysis of datasets when the unit of interest is not the individual record (microdata) but at a higher concept level. For example, a study of credit card purchases would reveal results that may be more meaningful at a higher-level class, like purchaser, store, or demographic area as opposed to the individual purchase itself. Only if the variability of factors within purchaser, store, or demographic area is retained would an analysis comparing each be most useful [58]. In a healthcare insurance claims setting, some information may be better consumed and interpreted at a provider level as opposed to the individual claim level.

Variables which consider this level of information as it applies to a group or class are called symbolic variables. Symbolic variables may be represented in a variety of ways. Table 2.8 represents the data from above, grouped by instruction level but described using symbolic notation.

Table 2.8: Instruction Level Matrix - Symbolic

<b>Instruction Level</b>	<b>Sample Size</b>	<b># of Children</b>	<b>Weight(kg.)</b>	<b>Gender</b>
1	1	3	60	F
2	2	[0,2]	[50,52]	M(100%)
3	1	1	55	M

A general example of the conversion between a classical data table (Table 2.9) and a symbolic table (Table 2.10) below where six individuals (*I1 – I6*) belong to one of two categories (*C1* or *C2*) and can be described by two variables (*Y1* and *Y2*) where *Y1* = *a,b,c* and *Y2* = 1,2,3 [56].

Table 2.9: Classical Data Table

<b>Individual</b>	<b>Concepts</b>	<b>Y1</b>	<b>Y2</b>
I1	C1	a	2
I2	C1	b	1
I3	C1	c	2
I4	C2	b	1
I5	C2	b	3
I6	C2	a	2

Table 2.10: Symbolic Data Table

<b>Concept</b>	<b>Y1</b>	<b>Y2</b>
C1	{a, b, c}	{1, 2}
C2	{a, b}	{1, 2, 3}

Symbolic data may be numerical or categorical and may take on different values. The ontology of symbolic data types was documented by [58] in Figure 2.6 below.

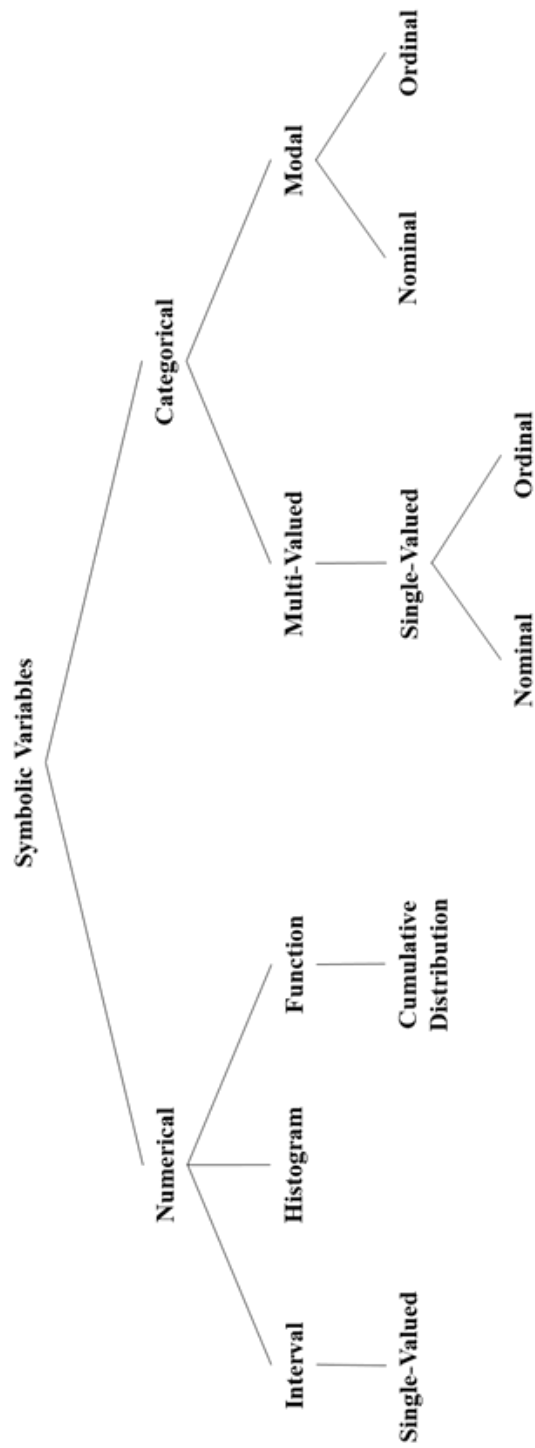


Figure 2.6: Symbolic Data Types [58]

Definitions of each data type and examples are included in Table 2.11.

Table 2.11: Symbolic Data Type Definitions

<b>Type</b>	<b>Description</b>	<b>Example</b>
Numerical	Single-Valued	weight(patient) = 60
	Interval-Valued	height(patient) = [60,75]
	Histogram-Valued	age(patient) = {[0,15], .15; [15-45], .45; [45-90], .40}
	Distribution Function	age(patient) = {[<15], .15; [<45], .60; [<∞], 1.0}
Categorical	Single-Valued	sex(patient) = {male}
	Multi-Valued	insurance coverage(patient) = {bronze, silver, gold}
	Modal	insurance coverage(patient) = {(bronze, .08; silver, .01; gold, .01)}

Another advantage to coding data into its symbolic form is the ability to include additional variables at the concept level which may not make sense or may not be known at an individual level. For example, a classical data table may include individual homeowners within a city. When that data is grouped by city, an additional factor such as percent commuters could be added to the table and provide information that could help in evaluating the concept.

With a symbolic table established, attention turns toward the analysis of the symbolic data table. Unlike classical data, which comprises techniques and tools that have been studied and refined over the past century, SDA statistical analysis is relatively new and the number of available methodologies is still small [59]. Research of the literature

suggests that while this work is just beginning, there has been progress in several traditional areas of statistical analysis including univariate and multivariate descriptive statistics, regression, principal component analysis, and clustering with the latter having received the most work of any of the multivariate methodologies [58].

## **2.4 Cluster Analysis and Symbolic Data Analysis**

Cluster analysis is a method used to group similar items together into entities called clusters where the items in each cluster share characteristics with other items in that cluster – and are very dissimilar from items in other clusters. The first step in cluster analysis is to determine which attributes should be included in the analysis. The second step is to determine which clustering approach to apply to the data. The third step is to determine which measure should be used to gauge similarity (dissimilarity). The fourth step is to determine the number of clusters that you wish to discern – which can be difficult because it involves balancing reducing the ultimate number of clusters with the forfeiture of explanatory information. The final step is to interpret the results of the analysis to determine what each of the resulting groups of clusters mean. Mooi and Sarstedt [60] provide guidance on the previous steps. Figure 2.7 is graphic taken from their work that includes explanations of each step in the clustering process [60].

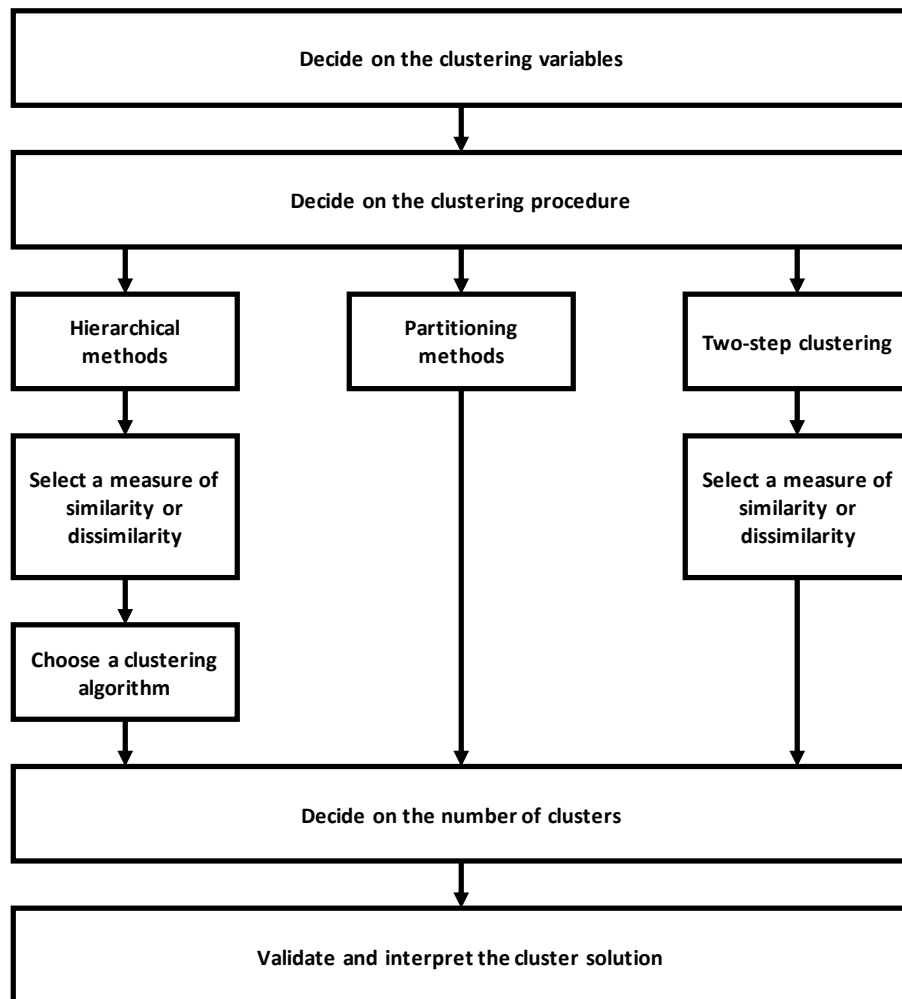


Figure 2.7: Steps in the Clustering Process [60]

The choice of which variables to include in the analysis is an important one. Finding the right quantity and mix of indicator variables can be difficult. The inclusion of too many variables can reduce the designer's ability to see differences. Mooi and Sarstedt cite Formann, 1984 [60] where he recommends a sample size of  $2^m$  where  $m$  equals the number of clustering variables. The designer must also protect against collinearity – two variables representing essentially the same characteristic can lead to that characteristic being artificially overrepresented in the final cluster analysis. Mooi and Sarstedt [60]

suggest that correlations between variables above 0.900 indicate duplicity in the model and should be corrected. It is also recommended that variables of different scales be standardized. Depending on the availability and type of data, the choice may be clear [60].

Tan et al. describes three types of clustering techniques [46].

- 1) Hierarchical vs. Partitional – nested cluster results that start with a single cluster that contains all the instances and flows down to nested partitions.
- 2) Exclusive vs. Overlapping vs. Fuzzy – techniques where the designer chooses whether instances can reside in only one cluster or can overlap into more than one group.
- 3) Complete vs. Partial – techniques where some instances are not assigned a cluster and possibly represent noise or outliers.

Hierarchical techniques are referred to as agglomerative or divisive clustering which has the structure that resembles the following graphic, Figure 2.8, taken from Mooi and Sarstedt [60].

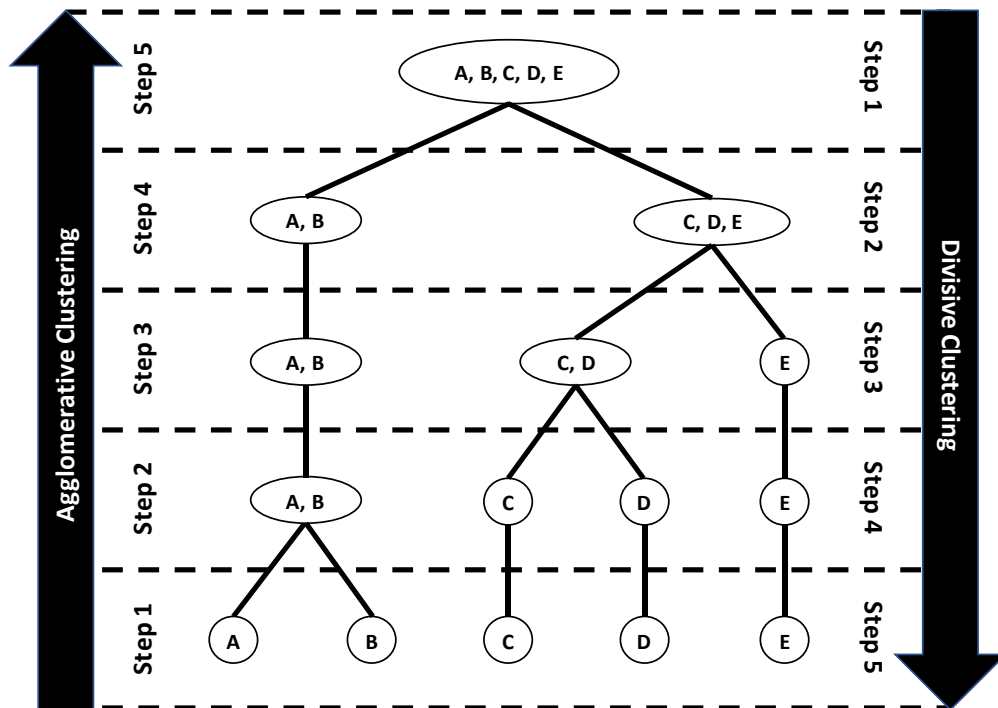


Figure 2.8: Agglomerative Versus Divisive Clustering [60]

The evaluation of the clustering technique depends on the technique employed. Most focus on numerical data and utilize a distance measure to minimize the distance between like objects and maximize the distance between those classes that are different. Distance metrics for categorical data are less common, can be more challenging, and is a topic that is addressed later in the dissertation. The most common method of determining similarity is by assessing the length of the line that connects two different instance points in the dataset using the Euclidean distance. The distances between each pair of points in the dataset are stored in a distance matrix. While the Euclidean distance is the one most often used, other distance measures exist which include the City-block, Chebychev, Angular, Canberra, and Mahalanobis distances.



There are several choices regarding which clustering algorithm to deploy. Some of the more popular algorithms include single linkage, complete linkage, average linkage, and centroidal linkage. Knowledge of the underlying data as well as expected cluster results can help the designer choose the appropriate linkage method.

Results of a cluster analysis are often expressed via a chart called a dendrogram. A visual depiction of a dendrogram is in Figure 2.9.

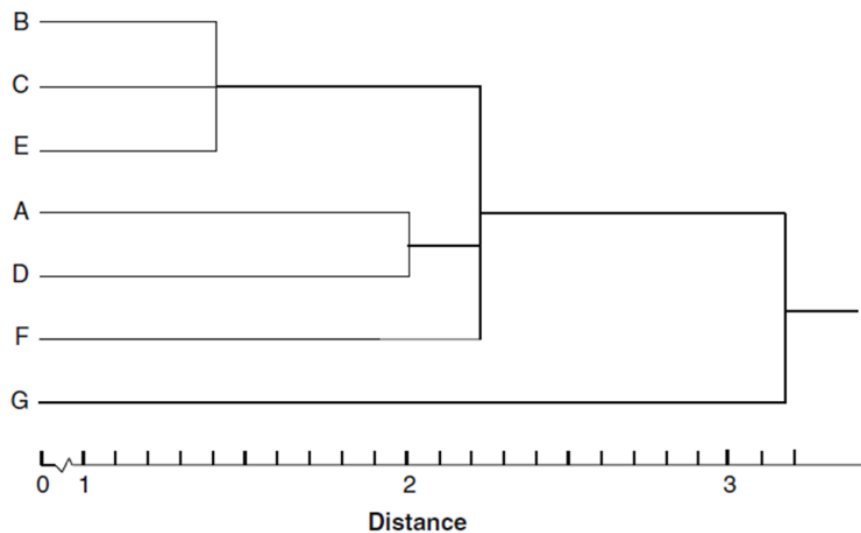


Figure 2.9: Dendrogram

A dendrogram offers a visual tool for being able to determine the number of clusters. Other methods are also available including the variance ratio criterion by Calinski and Harabasz, 1974 as cited by Mooi and Sarstedt [60].

Cluster analysis is frequently used for anomaly detection – or identifying patterns that do not conform to expected behavior. Several factors make this task difficult.

- Defining the region that defines normal is often difficult and the boundary between normal and anomalous is often not precise.
- When anomalies represent fraudulent behavior, the perpetrators are often working hard to disguise their activities as normal so that they fall within accepted boundaries.
- In many areas of study, the definition of “normal” is always changing.
- The characteristics that make an anomaly in one domain are often different when applied to a different domain.
- Data availability for labeled data fields (fraudulent/non-fraudulent) is often an issue.
- The noise in the data frequently masks actual anomalous data.

Because of the many challenges presented by the anomaly detection problem in its most general form, most techniques tend to focus on a specific formulation of the problem given the type of data, the availability of the data, and the type of anomaly being identified [45].

Because of its ability to detect anomalies, cluster analysis is a prime candidate for use in fraud detection. A common metric to measure the output delivered from a fraud detection algorithm is a suspicion score. The higher the score, the more likely that an anomalous data point indicates that fraud has occurred. The ranking of these scores will help prioritize which events should be recommended for further investigation [28].

As previously explained, the aim of clustering is to partition data into homogenous groups in a manner that minimizes within cluster variation and maximizes between cluster variation. That task is the same when applied to symbolic data. Cluster definition and proximity within SDA are defined using dissimilarity measures as in the case of classical

statistical analysis. In most cases, the dissimilarity measures presented for SDA are extensions of their classical counterparts [56].

The literature documents several dissimilarity measures that are frequently applied toward symbolic cluster analysis as well as comparisons of their effectiveness [52], [61]. Billard suggests that two of the most important distance measures are those developed by Gowda and Diday in 1991 and the work by Ichino and Yaguchi in 1994 [56], [62], [63]. A comparison of several dissimilarity measures was conducted by Esposito et al. in support of the ESPRIT Project SODAS [61]. Irpino and Verde recently explored using a Wasserstein distance measure for clustering histogram-valued data and a new approach by Brito and Ichino that creates clusters using a technique called quantile representation [58]. The distance matrix and the method of calculating dissimilarity measures determine the presence or absence of outliers which is the focus of this dissertation.

As in classical clustering, symbolic data clustering can be greatly affected by the variables' scale. De Carvalho et al. suggest several methods to standardize symbolic data in order to compare like scales [64]. Regardless of method, some form of standardization will always be needed in order to obtain an objective, scale-invariant result [58]. De Carvalho et al. suggest several approaches to standardization of interval data [64].

Mali et al. experimented with different validity indices to determine the optimal number of clusters by transforming quantitative validity indices (Normalized Modified Hubert, Davies-Bouldin, Dunn) to a symbolic framework and tests were then performed on several real-life datasets [65].

## 2.5 Histogram Binning Methods

The conversion of a standard dataset to one which is represented symbolically is a critical step in the analysis. When converting continuous variables to histogram-valued variables, the resolution and accuracy of the results would seem to depend on how the data is represented by the histogram. A primary feature of histogram construction is deciding how the data should be divided into bins. There are multiple binning approaches. Several of the more common approaches are summarized in Table 2.12 along with the formulas to determine the number of bins per histogram as well as the bin width to be used [66].

Table 2.12: Binning Methods and Formulas

Method	Notes	# of bins	bin width
<b>Square Root</b>	square root of number of data points	$\sqrt{n}$	$\frac{\max(values) - \min(values)}{\sqrt{n}}$
<b>Sturges</b>	best for normally distributed data, used by MS Excel	$\text{ceil}(\log_2 n) + 1$	$\frac{\max(values) - \min(values)}{\text{ceil}(\log_2 n) + 1}$
<b>Rice</b>	cube root of number of observations	$2 * \sqrt[3]{n}$	$\frac{\max(values) - \min(values)}{2 * \sqrt[3]{n}}$
<b>Scott</b>	uses standard deviation, good for normal	$\frac{\max(values) - \min(values)}{3.5 * \frac{\text{stddev}(values)}{\sqrt[3]{n}}}$	$3.5 * \frac{\text{stddev}(values)}{\sqrt[3]{n}}$
<b>Freedman-Diaconis</b>	uses interquartile range (IQR)	$\frac{\max(values) - \min(values)}{2 * \frac{IQR(values)}{\sqrt[3]{n}}}$	$2 * \frac{IQR(values)}{\sqrt[3]{n}}$

Through the study of different binning techniques, it was observed that many statistical packages and functions (including some packaged R functions) included a feature called “pretty.” For example, a modification to the original Sturges formula which creates a sequence of  $n+1$  equally spaced round values that are chosen to be 1, 2, or 5 times a power of 10 [67]. Since histograms are largely used in practice as a visual tool, this cleans up the bins and produces a more visual appealing representation of the data, as in Figure 2.10. In this dissertation, the accuracy of the histogram is more important than the visual representation of it. Therefore, the “pretty” function of histogram construction was not used and the true bin start and stop points were used per the formulas in Table 2.12.

For illustrative purposes, a sample dataset of 5000 records was generated,  $X \sim N(15.0, 1.0)$ . Table 2.13 depicts the number of bins and bin widths for this random sample.

Table 2.13: Results of Binning Calculations

<b>Method</b>	<b># of bins</b>	<b>bin width</b>
Square Root	71	0.10
Sturges	14	0.50
Rice	34	0.20
Scott	33	0.20
Freedman-Diaconis	44	0.20

The rule to specifically identify the bins and their contents is left-closed, right-open with the final bin being a full closed interval. For example, where  $(a, b, \dots, y, z)$  are real numbers and  $a < b < y < z$ , then the binning rules are:

$$\text{left-closed, right-open:} \quad [a, b) = \{x \mid a \leq x < b\}$$

$$\text{full-closed:} \quad [y, z] = \{x \mid y \leq x \leq z\}$$

In a dataset of  $n$  bins, the first  $n - 1$  bins are left-closed, right open. The final  $n^{\text{th}}$  bin is full-closed. This is the default binning method used for the remainder of this dissertation and is the method coded into the R program introduced in the next chapter.

Figure 2.10 depicts a typical histogram using the Sturges binning rule where a sample of 5000 records resulted in 14 “pretty” bins.

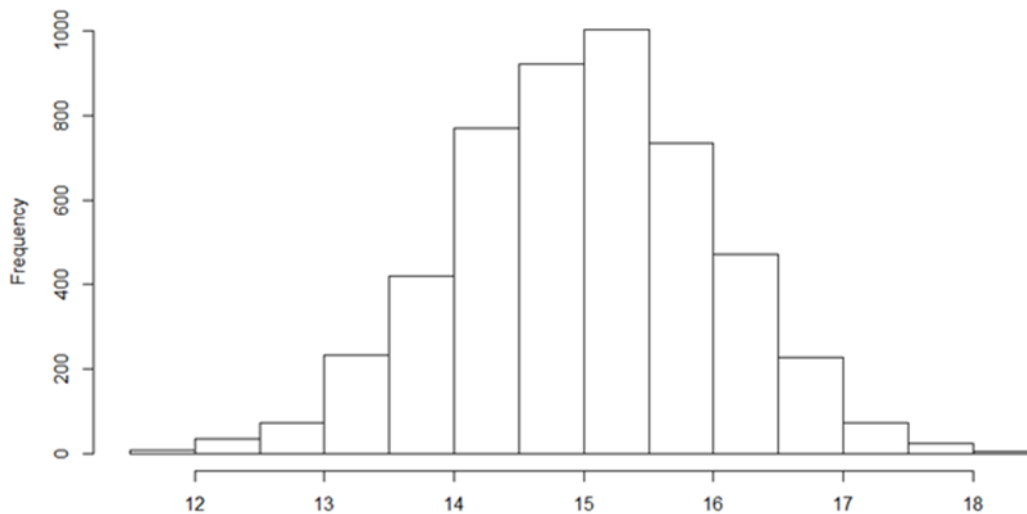


Figure 2.10: Histogram using Sturges Binning Rule

## 2.6 Practical Application of Symbolic Data Analysis to Large Datasets

While a review of the literature did produce several examples and case studies applying SDA to health related data, there was not an instance discovered that applied this methodology to healthcare insurance claims and/or the identification of fraudulent or abusive behavior that could be associated with data of this type. The research discovered an article published in 2009 that applied SDA to structural health monitoring [68]. Although different industry sectors, the approach to discovering departures from normal conditions would seem to be very similar to finding anomalous occurrences in a claims dataset. The article was co-authored by Edwin Diday.

The article explains a symbolic data approach to evaluating structural properties of a railway bridge in France. Monitoring of the physical properties of these structures is critical in order to determine their overall condition. Improper assessment of these conditions could lead to safety and or economic problems. The task is somewhat difficult due to the quantity of data that is collected and the ability to effectively and efficiently evaluate it making the practical application of monitoring methods mostly insufficient. The study proposes a data mining technique that utilizes two streams of data: one is a raw stream of acceleration data directly from sensors on the bridge and the other is a processed set of data that provides a set of modal parameters which includes a frequency reading. SDA is then applied to each of the datasets across three phases of bridge condition which includes: 1) before renovation; 2) during renovation; and 3) after renovation. If differences in the data could be observed, then the results could be used to determine the relative condition of the bridge and whether changes were occurring over time that would signal a weakening of the structure before failure occurred. The experimenters first used classical data analysis

and demonstrated that identification of all three phases of bridge condition could not be achieved. The data was then transformed to symbolic data and was subjected to different clustering techniques. The optimal number of clusters was determined using three evaluation indices as outlined by Milligan and Cooper in their 1985 work [69]. They also proposed a methodology for introducing new data to existing clusters in order to determine the methodology's ability to properly class new information. The results showed that SDA, using the processed modal or categorical data, proved to be an efficient and effective method of classifying and discriminating modifications of the bridge's structure. The hierarchy-divisive and dynamic cloud method outperformed the hierarchy-agglomerative method and overall, the modal data outperformed the raw data [70].



## **CHAPTER 3**

### **METHODOLOGY AND APPROACH**

The purpose of this chapter is to lay out a roadmap that will serve as the foundation for all subsequent experimentation. The basic steps of the analysis are explained, then supported through validation experiments with simulated data. Through these tests, centroidal statistics (assessing samples based on their sample means) and symbolic data statistics are calculated to illustrate the differences between the two and the value that the latter could bring in identifying aberrant data events in healthcare insurance claims data. The simulated examples will provide the detail on how the calculations are made and why certain approaches appear favorable to others. Continuous data, and mixed datasets that include categorical data, are explored. The topic of collinearity is explored, and tests are performed to understand the effects. Continuous data, represented as histogram-valued data, is studied to understand the best histogram binning approach to apply. The calculations for the construction of distance matrices are explained and a non-parametric approach for comparing groups of distance metrics is introduced.

#### **3.1 Steps of the Approach**

Through the research of the literature and subsequent experimentation, a process has been established and can be readily applied to any dataset. The process steps described in this section are uniformly applied to all datasets in this dissertation and are defined as follows:

- 1) Assess the underlying raw data (confirm the source of the data and cleanse the data if required to prepare it for analysis).
- 2) Determine the concept level to be studied (choose the level at which the business problem is best understood).
- 3) Summarize the raw data into concept groups and convert to a symbolic data table.
  - a) Convert continuous data to histogram-valued data if applicable.
    - i) Determine number of bins using the best binning approach.
  - b) Convert categorical data to modal data if applicable.
- 4) Construct the distance matrix.
- 5) Calculate the average distance measures and threshold value using a non-parametric approach.
- 6) Evaluate the result.

A simulated dataset is used to demonstrate the approach above and will provide the foundation for the real-world applications in the following chapter. The calculations for a centroidal and symbolic approach are explained and demonstrated using a spreadsheet process. This dataset will also be used to explain additional findings regarding the introduction of a categorical variable, the development of an outlier detection metric, and the selection of the Sturges binning rule when converting continuous data to histogram-valued data.

### **3.2 Simulated Dataset One**

The first dataset is a four column simulated dataset comprised of three continuous variables and one concept level variable. Column 4 contains the concept level variable and

is made up of four groups. Table 3.1 represents a partial view of the dataset. Table 3.2 provides information about the variables in each column.

Table 3.1: Simulated Dataset One

<b>i</b>	<b>V<sub>1</sub></b>	<b>V<sub>2</sub></b>	<b>V<sub>3</sub></b>	<b>Concept</b>
1	19.87	15.23	13.71	A
2	18.45	13.01	13.87	A
3	17.04	13.56	15.32	A
4	13.21	15.44	15.40	A
5	11.71	15.74	14.43	A
.				
.				
200	13.71	15.37	14.36	D

Table 3.2: Simulated Dataset One Descriptors

<b>Description</b>	<b>Variable Type</b>	<b>Values</b>
V <sub>1</sub> , (A)	Continuous	50 records per concept group, where $X \sim N(15.0, 3.0)$
V <sub>1</sub> , (B-D) V <sub>2</sub> , (A-D) V <sub>3</sub> , (A-D)	Continuous	50 records per concept group, where $X \sim N(15.0, 1.0)$
Concept Level Variable	Categorical	{A, B, C, D}, 50 records each

Each individual cell within each group was randomly generated using the distribution above. A boxplot of the data is shown in Figure 3.1.

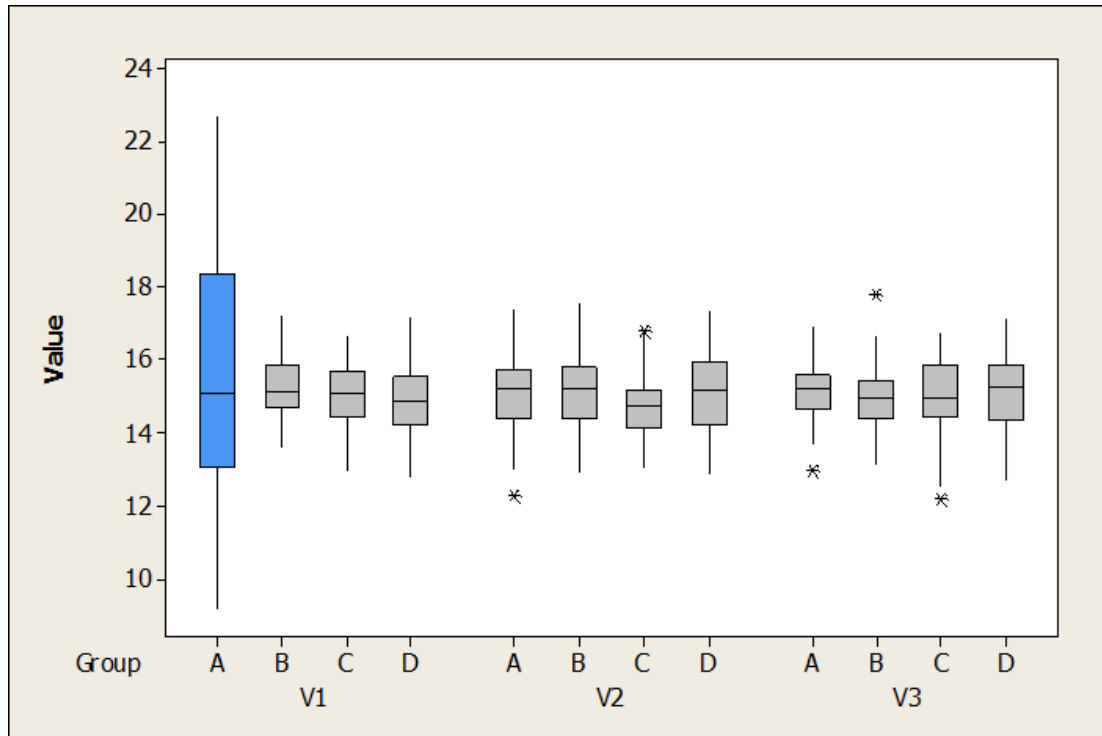


Figure 3.1: Boxplot of Simulated Dataset One

For this dataset and all that follow, a concept level variable needs to be selected. This should be the category or group that is the focus of study and the level at which the researcher intends to observe differences. This research does not attempt to identify which is the appropriate target to study or which records belong to which group. That is left up to the individual researcher and is dependent on the data. It is acknowledged that each dataset can be evaluated using different concept groupings within the data. As described in the example in the introduction, if the experimenter wished to study the difference

between teams as opposed to individual players, then the concept level variable would be the variable that identified the team every player belonged to so that the higher level concept of “teams” could be compared. In Simulated Dataset One, column 4 was selected as the concept level variable, resulting in four groups to be compared.

Once the concept level variable is selected, the standard data table is converted to a symbolic table. This compresses the table of multiple rows of individual records to summarized rows of data without losing the inherent variability of the underlying data. This is an important distinction using this approach. Groups of data may be summarized in multiple ways including singular values, intervals, functions, sets, and modal representation. For the purpose of this dissertation and all subsequent examples, categorical variables are represented as modal variables and continuous variables are represented as histogram-valued data. Presenting continuous variables as histogram-valued data preserves the characteristics of the underlying data which is a feature that is important when attempting to isolate minor inconsistencies within the data.

Simulated Dataset One includes only continuous input variables. Categorical variables are added in subsequent examples. The continuous input variables are converted to histogram-valued variables. In order to construct the histogram, the researcher has multiple options available including which binning methodology to use. For most of the examples in this dissertation, the Sturges binning method was used. A more complete review of binning methods is included later in this chapter.

Simulated Dataset One, with four groups of 50 records each with three continuous input variables resulted in bin quantity and width calculations described in Table 3.3.

Table 3.3: Simulated Dataset One Bin Results

	V1	V2	V3
<i>n</i>	200	200	200
<i>max</i>	22.71	17.54	17.81
<i>min</i>	9.19	12.28	12.17
<i>bin width</i>	1.50	.58	.63
<i># of bins</i>	9	9	9

The number of total rows in the dataset is shown as  $n = 200$  and is the same for all three variables, resulting in an equal number of bins based on the Sturges binning method.

Using the above binning rules, a symbolic data table was created. Below is the resulting four row table based on the number of concepts being evaluated for the variable V1. Each cell in Table 3.4 represents the frequency percentage for the individual bin per the rules above.

Table 3.4: Bin Representation for V1

	1	2	3	4	5	6	7	8	9
A	0.06	0.14	0.06	0.26	0.14	0.04	0.16	0.10	0.04
B	0.00	0.00	0.02	0.50	0.44	0.04	0.00	0.00	0.00
C	0.00	0.00	0.06	0.46	0.48	0.00	0.00	0.00	0.00
D	0.00	0.00	0.14	0.54	0.28	0.04	0.00	0.00	0.00

For example, the 50 records that comprise input variable V1 for Concept A make up a distribution with a 6% representation in the first interval [9.19, 10.69) per the binning rules above. Each group across each input variable will always have a distribution that accounts for 100% of its population and that is consistent across all groups across all variables. Similar representations are made for variables V2 and V3 and are represented in Table 3.5 and Table 3.6.

Table 3.5: Bin Representation for V2

	1	2	3	4	5	6	7	8	9
A	0.02	0.06	0.10	0.14	0.18	0.28	0.14	0.02	0.06
B	0.00	0.06	0.04	0.24	0.14	0.26	0.16	0.06	0.04
C	0.00	0.08	0.16	0.24	0.28	0.10	0.08	0.06	0.00
D	0.02	0.06	0.10	0.14	0.20	0.18	0.16	0.12	0.02

Table 3.6: Bin Representation for V3

	1	2	3	4	5	6	7	8	9
A	0.00	0.02	0.08	0.16	0.32	0.24	0.12	0.06	0.00
B	0.00	0.02	0.10	0.22	0.38	0.12	0.10	0.04	0.02
C	0.04	0.02	0.08	0.22	0.28	0.14	0.14	0.08	0.00
D	0.02	0.04	0.10	0.20	0.14	0.28	0.16	0.06	0.00

The three tables together represent a 4 x 27 symbolic data table that captures all variables by group without losing the embedded information related to the distribution of the data within concept by input variable.



With the symbolic data table complete, a distance matrix can be created that represents the Euclidean distance between every individual pair. Per the literature, there are other methods to calculate distance, but this dissertation will use the Euclidean calculation for every example. Experimenting with different distance calculations may provide an opportunity for future research.

The straight-line distances between each pair of data points within the histogram bin by variable are calculated by

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

where  $\mathbf{p} = (p_1, p_2, \dots, p_n)$  and  $\mathbf{q} = (q_1, q_2, \dots, q_n)$  are two points in Euclidean  $n$ -space.

Individual distances between each possible pair of groups within each designated bin are calculated. For example, the distance between groups A and B (AB) within the second bin for variable V1 is

$$AB(V1, Bin2) = [A(V1, Bin2) - B(V1, Bin2)]^2$$

$$= [0.14 - 0.00]^2$$

$$= 0.020$$

Table 3.7 for variable V1 is shown below, followed by Table 3.8 and Table 3.9 for variables V2 and V3 respectively.

Table 3.7: Pairwise Bin Distances for V1

	1	2	3	4	5	6	7	8	9
AB	0.004	0.020	0.002	0.058	0.090	0.000	0.026	0.010	0.002
AC	0.004	0.020	0.000	0.040	0.116	0.002	0.026	0.010	0.002
AD	0.004	0.020	0.006	0.078	0.020	0.000	0.026	0.010	0.002
BC	0.000	0.000	0.002	0.002	0.002	0.002	0.000	0.000	0.000
BD	0.000	0.000	0.014	0.002	0.026	0.000	0.000	0.000	0.000
CD	0.000	0.000	0.006	0.006	0.040	0.002	0.000	0.000	0.000

Table 3.8: Pairwise Bin Distances for V2

	1	2	3	4	5	6	7	8	9
AB	0.000	0.000	0.004	0.010	0.002	0.000	0.000	0.002	0.000
AC	0.000	0.000	0.004	0.010	0.010	0.032	0.004	0.002	0.004
AD	0.000	0.000	0.000	0.000	0.000	0.010	0.000	0.010	0.002
BC	0.000	0.000	0.014	0.000	0.020	0.026	0.006	0.000	0.002
BD	0.000	0.000	0.004	0.010	0.004	0.006	0.000	0.004	0.000
CD	0.000	0.000	0.004	0.010	0.006	0.006	0.006	0.004	0.000

Table 3.9: Pairwise Bin Distances for V3

	1	2	3	4	5	6	7	8	9
AB	0.000	0.000	0.000	0.004	0.004	0.014	0.000	0.000	0.000
AC	0.002	0.000	0.000	0.004	0.002	0.010	0.000	0.000	0.000
AD	0.000	0.000	0.000	0.002	0.032	0.002	0.002	0.000	0.000
BC	0.002	0.000	0.000	0.000	0.010	0.000	0.002	0.002	0.000
BD	0.000	0.000	0.000	0.000	0.058	0.026	0.004	0.000	0.000
CD	0.000	0.000	0.000	0.000	0.020	0.020	0.000	0.000	0.000

Table 3.10 represents the resulting matrix for all three of the input variables across all four concept variables is below. The matrix, sometimes referred to as a dissimilarity matrix, will always be a square  $n \times n$  symmetric matrix with all elements on the main diagonal equaling zero.

Table 3.10: Dissimilarity Matrix

	A	B	C	D
A	0.00			
B	0.50	0.00		
C	0.55	0.30	0.00	
D	0.47	0.40	0.37	0.00

Where,

$$\begin{aligned}
 \text{Distance (AB)} &= \sqrt{AB(V1, Bin1) + AB(V1, Bin2) + \dots + AB(V3, Bin9)} \\
 &= \sqrt{0.004 + 0.020 + 0.002 + \dots + 0.000} \\
 &= 0.50
 \end{aligned}$$

Scaling is required in order to make the comparison between the centroid and symbolic approaches, ensuring that the units are consistent across methods. The scaling function is derived by dividing each individual cell by the max value of all cells. The formula for the scaling function is:

$$\begin{aligned}
 \text{numerator} &= x \\
 \text{denominator} &= \max(x) \\
 \text{scaled value} &= x / \max(x)
 \end{aligned}$$

Using the scaling function above, the scaled distance matrix is in Table 3.11.

Table 3.11: Scaled Dissimilarity Matrix

	A	B	C	D
A	0.00			
B	0.91	0.00		
C	1.00	0.55	0.00	
D	0.87	0.73	0.67	0.00

Where,

$$\text{Scaled Distance (AB)} = 0.50/0.55 = 0.91$$

Once the distance matrix has been generated, each group within the dataset can be evaluated against all other groups using their distance measures. If all groups look relatively the same with respect to mean and variation, then the distances between each of those groups should be similar. Lower numbers represent similarity. Higher numbers represent dissimilarity. If one or more distance measures are statistically different than the rest, they are flagged as potential anomalies.

In order to assess these differences, an average distance that each group is from the rest of the groups can be calculated. The result is a set of averages that represent the groups by their relative distance to each other. If all groups are similar, then only common cause variation is present and the distribution of these averages should be normally distributed. However, if one or more groups are different in mean and/or distribution, then the resulting distribution of distance measures may not be normal. Because assumption of normality is

not correct in these cases, a non-parametric test was determined to be the best approach to use to identify the presence of anomalies or outliers.

The box and whisker plot is a non-parametric graphical tool that is universally used to depict groups of data. Its most basic construction includes identification of a group's median, first quartile, third quartile and a "whisker" that extends  $1.50 \times$  the interquartile range ( $IQR = Q3 - Q1$ ) below the first quartile and above the third quartile. Points that are outside this region are often considered "outliers" or "anomalies." For this application, only points that extend above  $Q3 + 1.50 \times IQR$  are considered anomalies as their distances appear to be greater than the rest. This is a one-sided test because only greater distances are of importance. Smaller distances or no discernable distance at all suggests similarity within the group. When compared back to a standard normal table, it can be shown that the area of this tail is approximately .0035. Extending the whisker  $.72 \times IQR$  beyond  $Q3$  results in an alpha value of .05. Graphical depictions of both are shown in the Figure 3.2 and Figure 3.3.

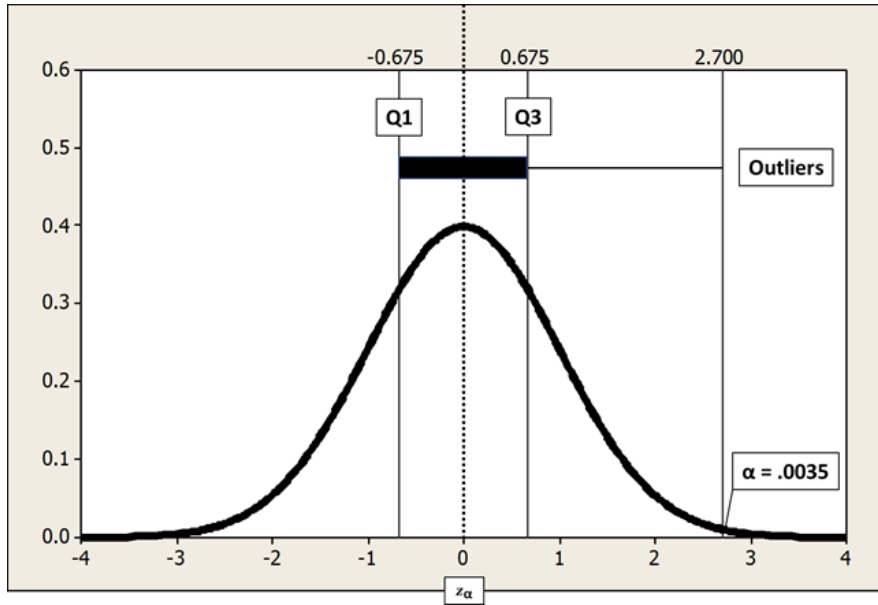


Figure 3.2: Alpha using  $Q3 + 1.50 \cdot IQR$

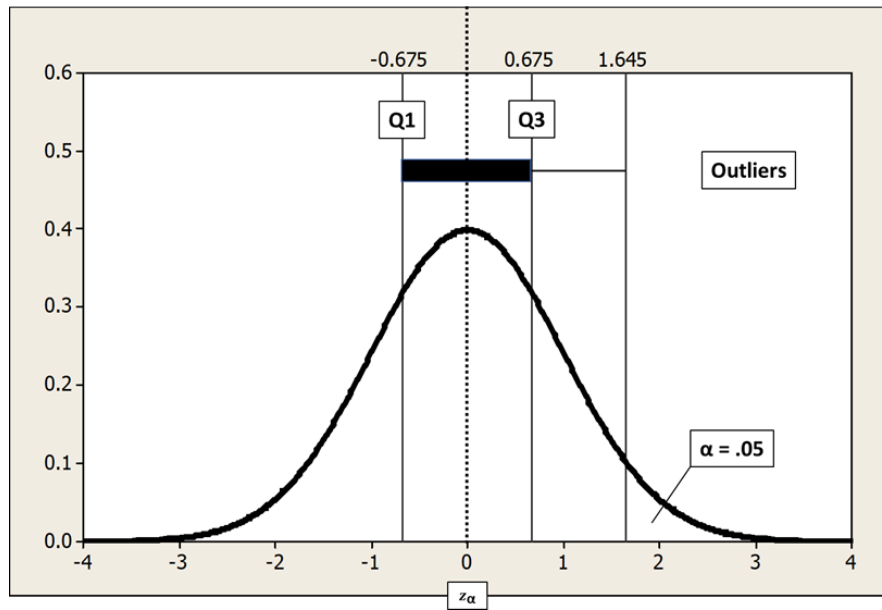


Figure 3.3: Alpha using  $Q3 + 0.72 \cdot IQR$

The decision as to how far to extend the whisker and set the threshold for outlier identification is left up to the researcher. As shown above, an *IQR* multiplier of 0.72 returns an alpha value of 0.05. An *IQR* multiplier of 1.50 returns an alpha value of 0.0035. A reduced range multiplier increases the value of alpha and increases the probability of a Type I error. Type I errors are often called false positives. Type II errors are often called false negatives. In this application, a false positive is identifying an observation as an anomaly, or outlier, when it is not. A false negative is failing to identify an anomaly when it exists. A model that has a high Type I or Type II error rate produces unnecessary costs to the system. Type I error costs may include unnecessary administrative costs associated with researching an event that appears as an outlier when it is not. In the case of fraud detection, an inappropriate investigation may also damage the relationship between the investigating and suspect companies. Excessive Type II error costs results in fraudulent behavior occurring and going undetected. Gadi et al. suggests the importance of understanding the costs of each and that they may not be equally important [71]. The literature offered few examples related to the cost of these types of errors associated with healthcare insurance claims fraud which suggests the topic may present an opportunity for future research. Users of the approach developed herein should choose an alpha value (and resulting threshold value) by balancing the costs of Type I and Type II errors for their particular situation.

In order to study the effect that alpha has on anomaly detection, different levels of alpha were tested using Simulated Dataset One, specifically targeting the anomalous subset of data represented by Group A, V1 as previously defined. As before, the data represented by Group A, V1 follows a distribution where  $X \sim N(15.0, 3.0)$ . As demonstrated later in



the chapter, Group A will be labeled anomalous by the symbolic method because its average distance score will exceed the calculated threshold value designed to detect anomalies. Table 3.12 below shows the alpha level, the multiplier used to obtain the stated alpha, the average distance measure associated with Group A, the calculated threshold value as previously defined, and the difference that Group A's distance measure is from the threshold. For example, when alpha is 0.0035, the multiplier used will be 1.50. For Simulated Dataset One, Group A has an average distance of 0.93 which exceeds the threshold value by 0.04. As shown in Table 3.12, as alpha increases, the threshold value decreases, and vice versa. In this example, alpha must be approximately 0.0002 or less for the model to not signal an anomalous condition. This is a much lower value than the recommended 0.0035 level and much lower than the level of 0.05, which is a commonly used standard.

Table 3.12: Alpha Value Sensitivity – Simulated Dataset One (Symbolic)

<b>Alpha</b>	<b>Multiplier</b>	<b>AVG DIST (A)</b>	<b>Threshold</b>	<b>Difference</b>
0.20000	0.12	0.93	0.80	0.12
0.10000	0.45	0.93	0.82	0.10
0.05000	0.72	0.93	0.84	0.09
0.04000	0.80	0.93	0.84	0.08
0.03000	0.89	0.93	0.85	0.08
0.02000	1.02	0.93	0.86	0.07
0.01000	1.22	0.93	0.87	0.06
0.00500	1.41	0.93	0.88	0.05
0.00350	1.50	0.93	0.89	0.04
0.00250	1.58	0.93	0.89	0.04
0.00100	1.79	0.93	0.90	0.02
0.00050	1.94	0.93	0.91	0.01
0.00020	2.12	0.93	0.92	0.00
0.00010	2.26	0.93	0.93	-0.01
0.00001	2.66	0.93	0.96	-0.03

Simulated Dataset One using a centroidal approach will be demonstrated next in this chapter. The behavior of changing alpha values for that approach are depicted in Table 3.13.

Table 3.13: Alpha Value Sensitivity – Simulated Dataset One (Centroidal)

Alpha	Multiplier	AVG DIST (A)	Threshold	Difference
0.20000	0.12	0.79	0.74	0.05
0.10000	0.45	0.79	0.77	0.02
0.05000	0.72	0.79	0.80	-0.01
0.04000	0.80	0.79	0.81	-0.02
0.03000	0.89	0.79	0.82	-0.03
0.02000	1.02	0.79	0.83	-0.04
0.01000	1.22	0.79	0.85	-0.06
0.00500	1.41	0.79	0.87	-0.08
0.00350	1.50	0.79	0.88	-0.09
0.00250	1.58	0.79	0.88	-0.09
0.00100	1.79	0.79	0.90	-0.11
0.00050	1.94	0.79	0.92	-0.13
0.00020	2.12	0.79	0.94	-0.15
0.00010	2.26	0.79	0.95	-0.16
0.00001	2.66	0.79	0.99	-0.20

Later in this chapter, the results from the symbolic approach and the centroidal approach will be compared. The results will show that the centroidal approach will fail to signal an anomaly at the 0.0035 level. The table above is consistent with that finding and shows that alpha would need to be at a level of 0.10 to signal such an event. The effect of changing alpha levels is the same for both approaches. Higher levels of alpha will reduce the threshold and tend to produce a greater number of false positives.

For the purpose of this dissertation, all tests are set with an *IQR* range multiplier of 1.50 (equating to an alpha value of approximately 0.00350). As previously shown in Figure 3.2 and Figure 3.3, the values for the multiplier and alpha map back to the properties of the

normal curve and can ultimately be adjusted by the researcher to the desired sensitivity. Future research is recommended to more completely explore the impact of alpha on the symbolic approach to anomaly detection.

Revisiting the dissimilarity matrix for the Simulated Dataset One, the full matrix would appear as in Table 3.14 with the average distances from each group excluding the values on the main diagonal.

Table 3.14: Dissimilarity Matrix with AVG DIST Calculation

	A	B	C	D	AVG DIST
A		0.91	1.00	0.87	0.93
B	0.91		0.55	0.73	0.73
C	1.00	0.55		0.67	0.74
D	0.87	0.73	0.67		0.75

Where,

$$\text{AVG DIST (A)} = (0.91 + 1.00 + 0.87) / 3 = 0.93$$

Evaluating the average distances that each group is from every other group in the dataset, the following table of statistics can be calculated, including the threshold value at which to determine the presence of outliers. Table 3.15 displays the threshold statistics.

Table 3.15: Threshold Determination Calculation

<b>Metric</b>	<b>Value</b>
<i>Median</i>	0.75
<i>Q1</i>	0.74
<i>Q3</i>	0.80
<i>Threshold</i>	0.89

In the example above,  $AVG\ DIST(A) = 0.93$  which exceeds the *Threshold* of 0.89. With the non-parametric alpha value set at .0035, it can be assumed that would occur less than 0.35% of the time. Therefore, Group A would be classified as an anomaly with a recommendation to examine it as being different than the other groups. We know this to be true, as V1 of Group A originates from a different distribution.

Several factors could influence a point exceeding the threshold value, including the number of groups being evaluated. Non-parametric tests are generally less powerful than parametric tests and while the test statistic can be calculated with just two groups present, the results for comparison are difficult to assess. Additionally, as with all tests of this nature, a false anomaly could still be identified, although that would be 0.35% of the time or less. Again, the sensitivity of the threshold value could be adjusted based on the rate of false positives and/or false negatives (Type I and/or Type II errors) tolerated in practice.

The primary theme of this research is to compare the use of symbolic histogram-valued data to that of a centroidal approach when evaluating groups of data for outliers. In this example and those that follow, the symbolic approach is compared to the centroidal approach and the results are studied.

Using the Simulated Dataset One, Table 3.16 is constructed and shows the centroids (or means in this case) of each individual group.

Table 3.16: Simulated Dataset One – Centroidal Approach

	V1	V2	V3
A	15.61	15.00	15.15
B	15.24	15.11	15.01
C	15.07	14.73	14.99
D	14.86	15.10	15.13

From this table, Euclidean distances can be calculated as before. The resulting matrix showing the distance between each individual pair is in Table 3.17.

Table 3.17: Pairwise Distance Table – Centroidal Approach

	V1	V2	V3
AB	0.139	0.012	0.019
AC	0.290	0.072	0.026
AD	0.555	0.011	0.000
BC	0.027	0.144	0.001
BD	0.139	0.000	0.015
CD	0.043	0.139	0.021

As previously described, the data can be scaled for comparison purposes. This forces the maximum distance calculated within every dataset to always be 1.00 with the remaining distances expressed in relation to the maximum. Non-parametric statistics were calculated using the scaled data for comparison. The resulting distance matrix and test statistics are in Table 3.18 and Table 3.19.

Table 3.18: Scaled Dissimilarity Matrix with AVG DIST Calculation - Centroidal

	A	B	C	D	AVG DIST
A		0.55	0.83	1.00	0.79
B	0.55		0.55	0.52	0.54
C	0.83	0.55		0.60	0.66
D	1.00	0.52	0.60		0.71

Table 3.19: Threshold Determination Calculation - Centroidal

<b>Metric</b>	<b>Value</b>
<i>Median</i>	0.68
<i>Q1</i>	0.63
<i>Q3</i>	0.73
<i>Threshold</i>	0.88

The difference using SDA can be observed using this example. None of the group's AVG DIST values exceed the threshold, suggesting that there is no difference among them. While we know that the means are the same, we also know that there is a change in variability within V1 of Group A. In this example, only a change in variability is present and the symbolic approach preserves the underlying distributions inherent in the data and relies on that information to calculate distances between the groups.

The treatment of categorical variables in both the centroidal and symbolic approach is similar. In the examples explored in this dissertation, all categorical variables are treated as modal variables, meaning that their values are represented as a percentage of the overall sample. As with the continuous histogram-valued variables, the representation of all categories within a group must equal 1.00. The assignment of dummy variables within the raw dataset are given values of [0,1]. When the centroids or means of these values are computed, the result is a percentage representation of the categorical variable. Displayed below is a depiction of Simulated Dataset One with a categorical CAT1 variable added to the end of the V3 variable as described in the previous example. Therefore, this example includes variation introduced by continuous and categorical variables. Concept A will include the value 'Orange', Concepts B, C, and D will contain the value 'Blue'. For

purposes of the experiment, Concept A was assigned 100% ‘Orange’ and 0% ‘Blue’ but each concept variable could take on partial representations as well. Figure 3.4 is a graphical depiction of the additional categorical variable.

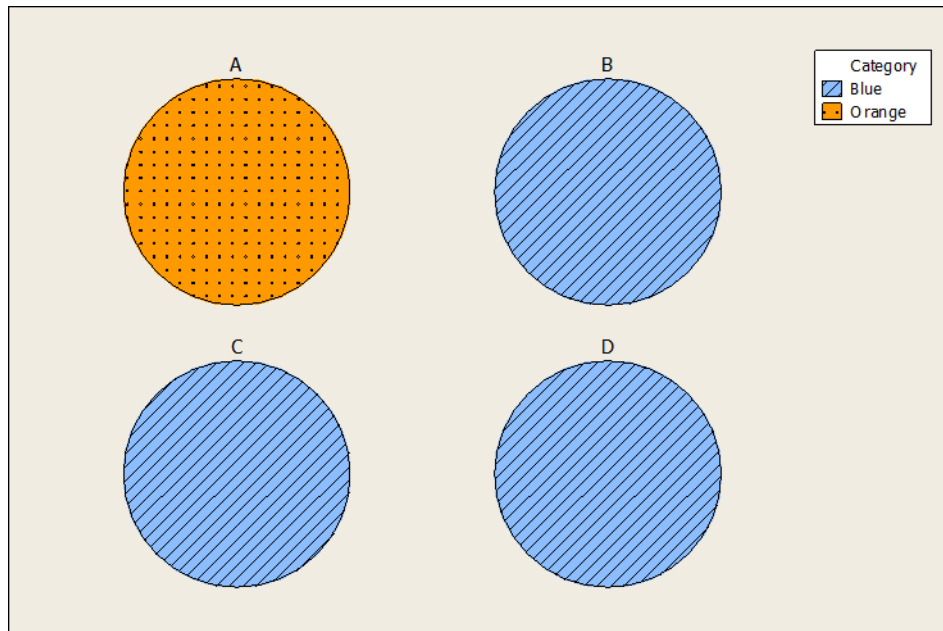


Figure 3.4: Distribution of Simulated Dataset One – Categorical Variable

Below are examples of the symbolic table (Table 3.20), the Euclidean distance table (Table 3.21), the final distance matrix (Table 3.22), and the threshold determination calculation (Table 3.23).

Table 3.20 and Table 3.21 begin with V3, cell five and are only partial representation of the full symbolic table. As expected, when a categorical variable was added to this random dataset in a manner that one concept group is different than the others, the distance measure exceeded the non-parametric threshold and identified that group as an outlier.



Table 3.20: Simulated Dataset One with Categorical Variable Added

	<b>V3</b>					<b>CAT1</b>	
	5	6	7	8	9	Orange	Blue
A	0.32	0.24	0.12	0.06	0.00	1.00	0.00
B	0.38	0.12	0.10	0.04	0.02	0.00	1.00
C	0.28	0.14	0.14	0.08	0.00	0.00	1.00
D	0.14	0.28	0.16	0.06	0.00	0.00	1.00

Table 3.21: Pairwise Distance Table with Categorical Value Added

	<b>V3</b>					<b>CAT1</b>	
	5	6	7	8	9	Orange	Blue
AB	0.004	0.014	0.000	0.000	0.000	1.000	1.000
AC	0.002	0.010	0.000	0.000	0.000	1.000	1.000
AD	0.032	0.002	0.002	0.000	0.000	1.000	1.000
BC	0.010	0.000	0.002	0.002	0.000	0.000	0.000
BD	0.058	0.026	0.004	0.000	0.000	0.000	0.000
CD	0.020	0.020	0.000	0.000	0.000	0.000	0.000

Table 3.22: Dissimilarity Matrix with Categorical Variable Added

	A	B	C	D	<u>AVG DIST</u>
	A		0.99	1.00	0.98
B	0.99		0.20	0.26	0.48
C	1.00	0.20		0.24	0.48
D	0.98	0.26	0.24		0.50

Table 3.23: Threshold Determination Calculation with Categorical Variable Added

<b>Metric</b>	<b>Value</b>
<i>Median</i>	0.49
<i>Q1</i>	0.48
<i>Q3</i>	0.62
<i>Threshold</i>	0.83

In this example, Concept A is clearly different than the rest and represented as such, where  $AVG\ DIST(A) = 0.99$  exceeds  $Threshold = 0.83$ . Its average distance from the rest is great enough to calculate a score greater than the threshold. In practice, Concept A would be singled out for being anomalous and would warrant further investigation. It should also be noted that each variable in the symbolic approach is assigned an equal weight, including the categorical variable. In the Simulated Dataset One example, the three continuous variables and one categorical variable each have a contribution of one and sum to four. Scenarios can be anticipated where this could be modified but that is a topic recommended for future research.

For comparison purposes, the above example was repeated using the centroidal approach. Below are examples of the classic data table (Table 3.24), the Euclidean distance table (Table 3.25), the final distance matrix (Table 3.26), and the threshold determination calculation (Table 3.27).

Table 3.24: Simulated Dataset One with Categorical Variable Added - Centroidal

	V1	V2	V3	Orange	Blue
A	15.61	15.00	15.15	1.00	0.00
B	15.24	15.11	15.01	0.00	1.00
C	15.07	14.73	14.99	0.00	1.00
D	14.86	15.10	15.13	0.00	1.00

Table 3.25: Pairwise Distance Table with Categorical Value Added - Centroidal

	V1	V2	V3	Orange	Blue
AB	0.139	0.012	0.019	1.000	1.000
AC	0.290	0.072	0.026	1.000	1.000
AD	0.555	0.011	0.000	1.000	1.000
BC	0.027	0.144	0.001	0.000	0.000
BD	0.139	0.000	0.015	0.000	0.000
CD	0.043	0.139	0.021	0.000	0.000

Table 3.26: Dissimilarity Matrix with Categorical Variable Added - Centroidal

	A	B	C	D	AVG DIST
A		0.92	0.96	1.00	0.96
B	0.92		0.26	0.24	0.47
C	0.96	0.26		0.28	0.50
D	1.00	0.24	0.28		0.51

Table 3.27: Threshold Determination Calculation - Centroidal

Metric	Value
<i>Median</i>	0.50
<i>Q1</i>	0.49
<i>Q3</i>	0.62
<i>Threshold</i>	0.81

In this case, like the symbolic approach, Concept A is clearly different than the rest and represented as such, where  $AVG\ DIST(A) = 0.96$  exceeds  $Threshold = 0.81$ . Its average distance from the rest is great enough to calculate a score greater than the threshold, producing the same result as the symbolic approach.

### 3.2.1 Selecting a Histogram Binning Method

During this research, multiple histogram binning approaches were assessed. Features of several of the more widely used approaches were introduced and defined in Chapter 2. When comparing one dataset to another by examining their distributions, the method by which those distributions are represented is critical because of the varying bin quantities and bin widths that can be used. As mentioned previously in this chapter, the Sturges binning method was chosen as the default method in this dissertation. In addition to being prominently used in most statistical textbooks and mainstream statistical software, its applicability for use appeared suitable when comparing datasets.

As explained earlier in the chapter, Simulated Dataset One is a four group, three variable dataset where one variable within one group is altered to have greater variability than any of the other group/variable combinations. To better understand the effects of binning and provide a baseline, a separate dataset was generated that contains random noise only, where all groups and variables are distributed as  $X \sim N(15.0, 1.0)$ . There are no anomalies present in the random noise only dataset and there are no distinguishable differences among any of the groups.

Each of the five binning techniques was applied to the random noise only dataset in order to establish baseline distance levels for each group within the dataset. The procedure

was repeated with the anomalous Simulated Dataset One and a threshold value was calculated based on the differences from the random noise only dataset and Simulated Dataset One as shown in Table 3.28. As expected, across each binning approach, Group A was identified as an anomaly. However, the degree to which this identification was made varied depending on the histogram binning approach.

Table 3.28 shows the effect of each binning technique on Group A. Each binning method was applied to the baseline random noise only case and to Simulated Dataset One.

Table 3.28: Binning Methods Compared

Method	Group A Difference from Random Noise Only	Calculated Threshold Value	Difference from Threshold Value
Square Root	0.10	0.02	0.08
Sturges	0.16	0.08	0.08
Rice	0.10	0.02	0.08
Scott	0.09	0.04	0.05
Freedman	0.05	-0.01	0.06

Based on the results above and from similar tests conducted during the course of this research, the Sturges method (without “pretty” breakpoints) was chosen as the default method. The difference from threshold is equal to that of the other binning methods suggesting that it could perform reasonably at discerning anomalistic data from that of random noise.

It should be noted, however, that this is an opportunity for future research. Binning approaches are sensitive to changes in the number of data points being measured and the variability within the data. The test results displayed in Table 3.28 are based on variables

with a sample size of 200. Attempting to discern small changes in larger datasets can present anomalous occurrences that the Sturges method may overlook. This will be further explored in the following chapter where just such a situation arises.

### **3.2.2 Multicollinearity and the Symbolic Approach**

Multicollinearity exists when two or more explanatory variables used for predicting an output are related to each other. That relationship is often expressed in terms of one variable being correlated to another. Perfect correlation exists when one predictor variable can perfectly explain the other. The correlation coefficient is expressed by the symbol  $\rho$  and can range from -1.0 to 1.0. Perfect correlation exists when  $\rho = -1.0$  or  $\rho = 1.0$  – although it rarely exists in naturally occurring data collection. Instead, most explanatory variables have a degree of measurable correlation. It is up to the researcher to determine what value of  $\rho$  constitutes a relationship worth exploring. Hypothesis testing can be used to test the existence of correlation where,

$$H_0: \rho_{xy} = 0 \quad \text{versus} \quad H_a: \rho_{xy} \neq 0$$

To test the effect of collinearity on the symbolic data approach, two random datasets with five groups and five random variables were generated. Table 3.29 represents the 1,000 row structure for both datasets.

Table 3.29: Five Group Five Variable Datasets (V and C)

Group	#	V1, C1	V2, C1	V3, C3	V4, C4	V5, C5
A	200	$X \sim N(10, 1)$	$X \sim N(20, 1)$	$X \sim N(30, 1)$	$X \sim N(40, 1)$	$X \sim N(50, 1)$
B	200	$X \sim N(10, 1)$	$X \sim N(20, 1)$	$X \sim N(30, 1)$	$X \sim N(40, 1)$	$X \sim N(50, 1)$
C	200	$X \sim N(10, 1)$	$X \sim N(20, 1)$	$X \sim N(30, 1)$	$X \sim N(40, 1)$	$X \sim N(50, 1)$
D	200	$X \sim N(10, 1)$	$X \sim N(20, 1)$	$X \sim N(30, 1)$	$X \sim N(40, 1)$	$X \sim N(50, 1)$
E	200	$X \sim N(10, 1)$	$X \sim N(20, 1)$	$X \sim N(30, 1)$	$X \sim N(40, 1)$	$X \sim N(50, 1)$

The two, five group, five variable, 1,000 row datasets were identical with respect to variables V and C with one exception. The first dataset consisted of variables V1 – V5 with no correlation between the variables. The second dataset, which consisted of variables C1 – C5, was generated with a correlation coefficient ( $\rho$ ) of 0.80 between variables C4 and C5.

Figure 3.5 is a graphical depiction of the first dataset (V1-V5) without correlation.

Figure 3.6 is a graphical depiction of the second dataset (C1-C5) with correlation.

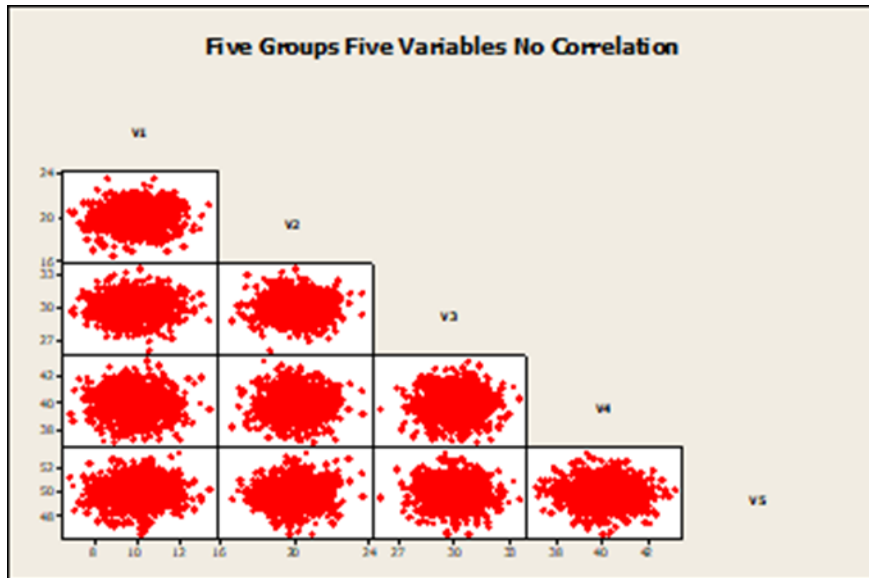


Figure 3.5: Five Group Five Variables Without Correlation

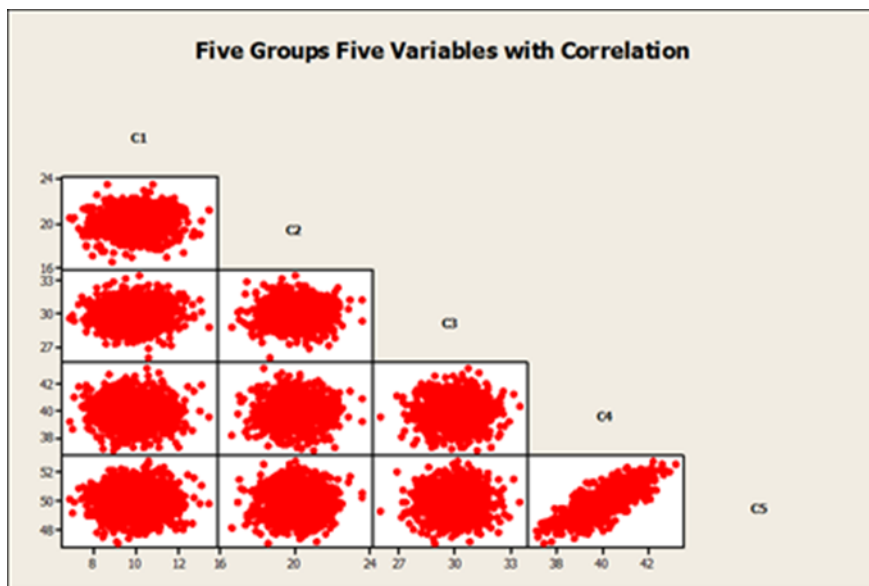


Figure 3.6: Five Group Five Variables With Correlation



Table 3.30 shows the correlation coefficients of all the variables for both datasets where  $\rho_{(C4C5)} = 0.790$ . Note the first number within a cell represents the correlation coefficient and the second number indicates the p-value associated with the test for correlation.

Table 3.30: Correlation Tables

Without Correlation					With Correlation				
	V1	V2	V3	V4		C1	C2	C3	C4
V2	0.005 0.880				C2	0.005 0.880			
V3	0.034 0.281	0.020 0.535			C3	0.034 0.281	0.020 0.535		
V4	-0.019 0.543	-0.005 0.874	0.017 0.595		C4	-0.019 0.543	-0.005 0.874	0.017 0.595	
V5	0.062 0.048	0.046 0.144	0.035 0.274	-0.048 0.126	C5	0.023 0.477	0.024 0.445	0.035 0.268	0.790 0.000
Cell Contents: Pearson correlation P-Value					Cell Contents: Pearson correlation P-Value				

It should be noted that  $\rho_{(V1V5)} = 0.062$ , which signals a weak linear relationship; however, it also appears significant at the .050 level with a measured p-value of .048. This is just an artefact of the random data generation. This artefact is not likely to be observed with larger sample sizes or more stringent p-values than .050.

In order to test the effect of multicollinearity, the symbolic approach was applied to both datasets and the distance matrices were compared. The distance matrices for each group are in Table 3.31 and Table 3.32.

Table 3.31: Symbolic Distance Matrix for Dataset without Correlation

Group	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
<i>A</i>	0.00				
<i>B</i>	0.90	0.00			
<i>C</i>	1.00	0.84	0.00		
<i>D</i>	0.87	0.76	0.93	0.00	
<i>E</i>	0.75	0.82	0.93	0.89	0.00

Table 3.32: Symbolic Distance Matrix for Dataset with Correlation

Group	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
<i>A</i>	0.00				
<i>B</i>	0.88	0.00			
<i>C</i>	1.00	0.83	0.00		
<i>D</i>	0.83	0.90	0.92	0.00	
<i>E</i>	.0.85	0.83	0.99	0.97	0.00

In order to make a comparison, the ten individual pairwise distances were evaluated using a paired t-test. Results of the Minitab test are in Figure 3.7.

	N	Mean	StDev	SE Mean
With Correlation	10	0.9000	0.0675	0.0213
Without Correlation	10	0.8690	0.0784	0.0248
Difference	10	0.0310	0.0599	0.0189

95% CI for mean difference: (-0.0118, 0.0738)  
T-Test of mean difference = 0 (vs not = 0): T-Value = 1.64 P-Value = 0.136

Figure 3.7: Paired T-Test of Pairwise Distances

The results from the test above suggest no significant difference in distances for the presence or the absence of multicollinearity. In other tests performed during this research, similar results were observed which, anecdotally, point to a similar conclusion, that is, the performance of the symbolic method is robust to correlations. As a general rule, multicollinearity is a condition which should be guarded against and certainly future research is needed to more completely explore its impact, if any, on the symbolic approach to anomaly detection.

### 3.2.3 Developing the R Code

Spreadsheet analysis is an excellent way to understand the calculations that drive a new methodology but can be difficult to manage when the cases studied involve higher volume and higher complexity. Building on the foundation established earlier in this chapter, this section describes a tool that was built through the open-source R programming language. R is a readily available statistical software package used by statisticians and data

scientists. Developed in 1993, its ubiquity among professionals in academia and business alike has created a broad user community [72]. Users of R benefit and contribute to this platform through the development of “packages” that extend the features and capability beyond the base application.

While this research does not extend completely into symbolic clustering analysis, the distance matrices and related calculations that are the foundation of that work are also the foundation of this dissertation. Many of the subroutines and function calls that were used in those packages proved helpful with this research, particularly the “RSDA” and “symbolicDA” packages available from the CRAN Repository. Building the R code provided an efficient way to evaluate multiple datasets and make comparisons between the centroidal and symbolic approaches. The test code went through several iterations that included the investigation and research of established public packages as well as custom coding needed to accommodate this dissertation. Custom source functions were developed as separate standalone modules with the ability to be called at multiple times throughout the script. Considered was how the code would handle the previously discussed binning rules, as well as nuances associated with handling continuous and categorical data. The datasets were created in Microsoft Excel or Minitab and then processed using the R code. Raw data was converted to R data frames and a customized scaling function was coded as described in the previous section. Basic descriptive information regarding each dataset was placed in each file and, where applicable, graphical views of the data were generated. Histogram-valued variables were given an ‘H’ designation and categorical variables were coded as ‘M’ or modal. Data tables and distance matrices were created using Euclidean distance as described in the previous chapter. The resulting distance matrix was produced

along with the average distances each group was to the rest. The non-parametric threshold value was calculated as described in the previous chapter and, if one or more of the individual groups exceeded the threshold per the calculations, an anomalous condition was signaled. The display from the R code allowed for the visual assessment of the data and the comparison of the distance matrices between the centroidal and symbolic approaches. Output data was also written to an external file for more in depth analysis. A sample screenshot of one of the R code results reports is in Figure 3.8.

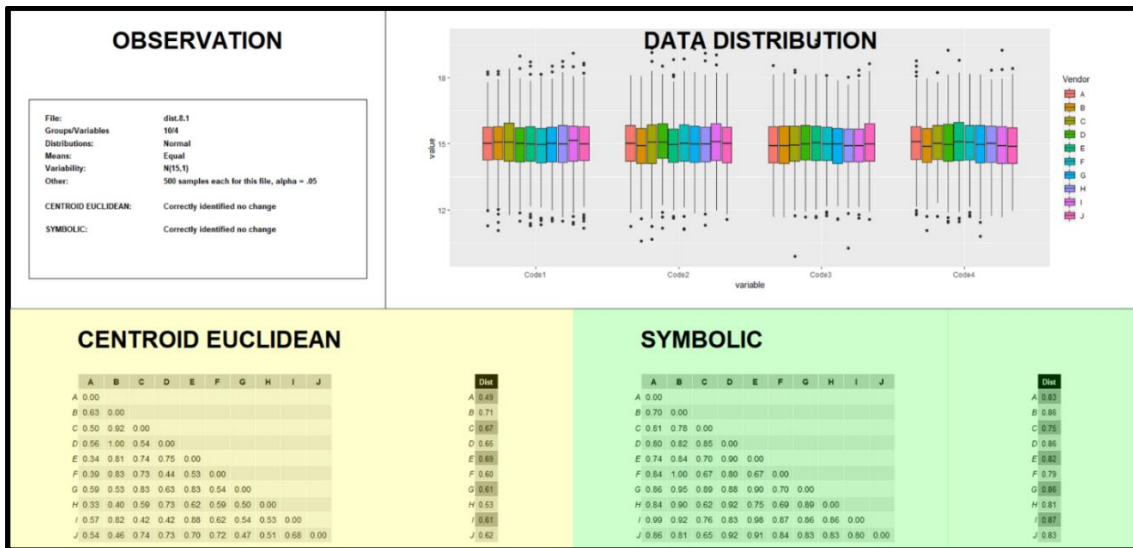


Figure 3.8: Screenshot of R Code Report Display

### 3.3 Simulated Dataset Two

Having established an approach to symbolic analysis and with an efficient method of experimentation, the next step in the dissertation involved testing other datasets to further demonstrate the difference between the two approaches.

We first evaluated a simple, three cluster set. Multiple instances and multiple distributions were evaluated to demonstrate the effect of mean and variability change. Initially, this analysis was performed manually but with the development of the R code, larger sample sizes could be assessed. This example includes simulated data divided into three groups with each group containing 1000 observations. Clusters A and C are uniformly distributed with a range of 10 units in  $xy$  space while Cluster B is normally distributed with a standard deviation of 2.5 units in  $xy$  space. The coordinates of the means in  $xy$  space were designed to create an equilateral triangle and were modeled as depicted in Table 3.33.

Table 3.33: Cluster Coordinates - Separate

Cluster	Cluster Means
A	$\{x,y \mid \bar{x} = 15.00, \bar{y} = 17.01 \}$
B	$\{x,y \mid \bar{x} = 30.00, \bar{y} = 42.99 \}$
C	$\{x,y \mid \bar{x} = 45.00, \bar{y} = 17.01 \}$

Figure 3.9 shows a scatterplot of the clusters.

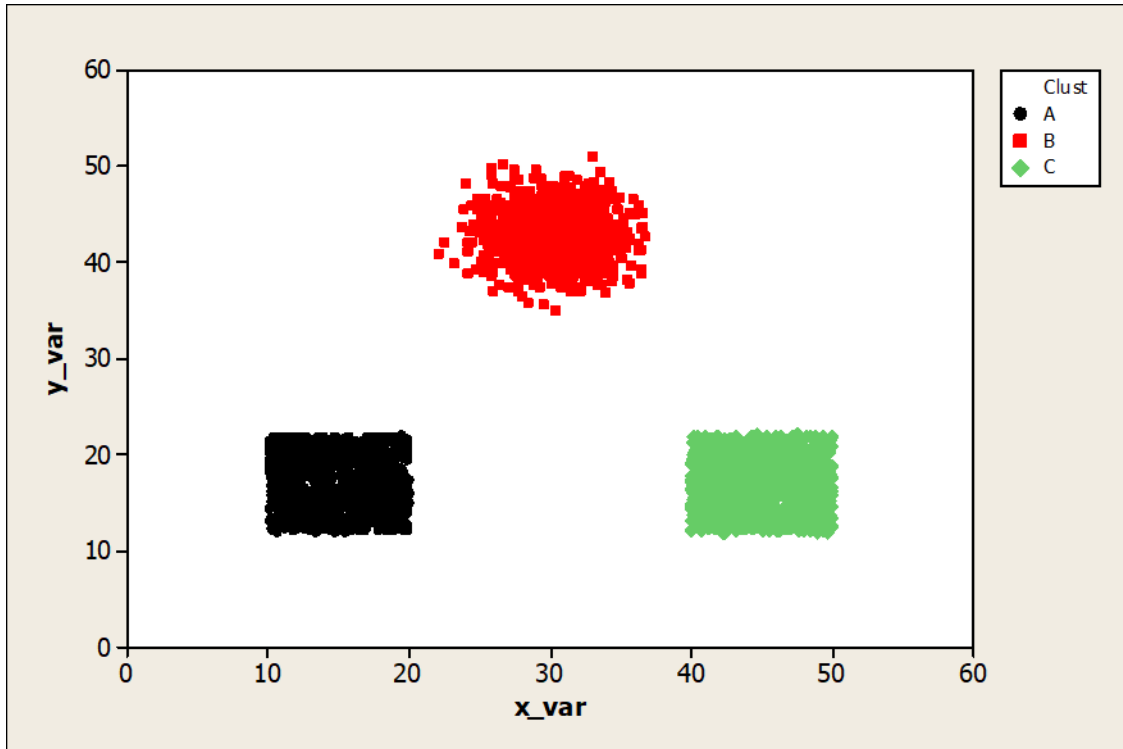


Figure 3.9: Three Cluster Scatterplot - Separate

The resulting distance matrices using the centroidal and symbolic methods are shown in Table 3.34 and Table 3.35.

Table 3.34: Three Cluster Dissimilarity Matrix - Centroidal

	A	B	C	AVG DIST
A	0.00			1.00
B	1.00	0.00		1.00
C	1.00	1.00	0.00	1.00

Table 3.35: Three Cluster Dissimilarity Matrix - Symbolic

	A	B	C	AVG DIST
A	0.00			0.85
B	1.00	0.00		1.00
C	0.70	1.00	0.00	0.85

The centroidal approach shows no difference among the three groups as expected. While the centroids of these clusters are equidistant, the pattern of Cluster B is different from the other two. The symbolic approach correctly shows the greatest difference between Cluster A and Cluster B and between Cluster B and Cluster C. This is not because their centroids are different but because their distributions are. Cluster A and Cluster C appear to be more similar – and they are because they both are generated from the same uniform distribution type. The AVG DIST metric confirms Cluster B is different than the other two.

In contrast, the same three clusters with the same distributions were overlaid equating the  $xy$  means of all three groups. Table 3.36 shows the scatterplot coordinates.

Table 3.36: Cluster Coordinates - Overlaid

Cluster	Cluster Means
A	$\{x,y \mid \bar{x} = 30.00, \bar{y} = 30.00 \}$
B	$\{x,y \mid \bar{x} = 30.00, \bar{y} = 30.00 \}$
C	$\{x,y \mid \bar{x} = 30.00, \bar{y} = 30.00 \}$

Figure 3.10 is a graph of the clusters depicted in  $xy$  space.



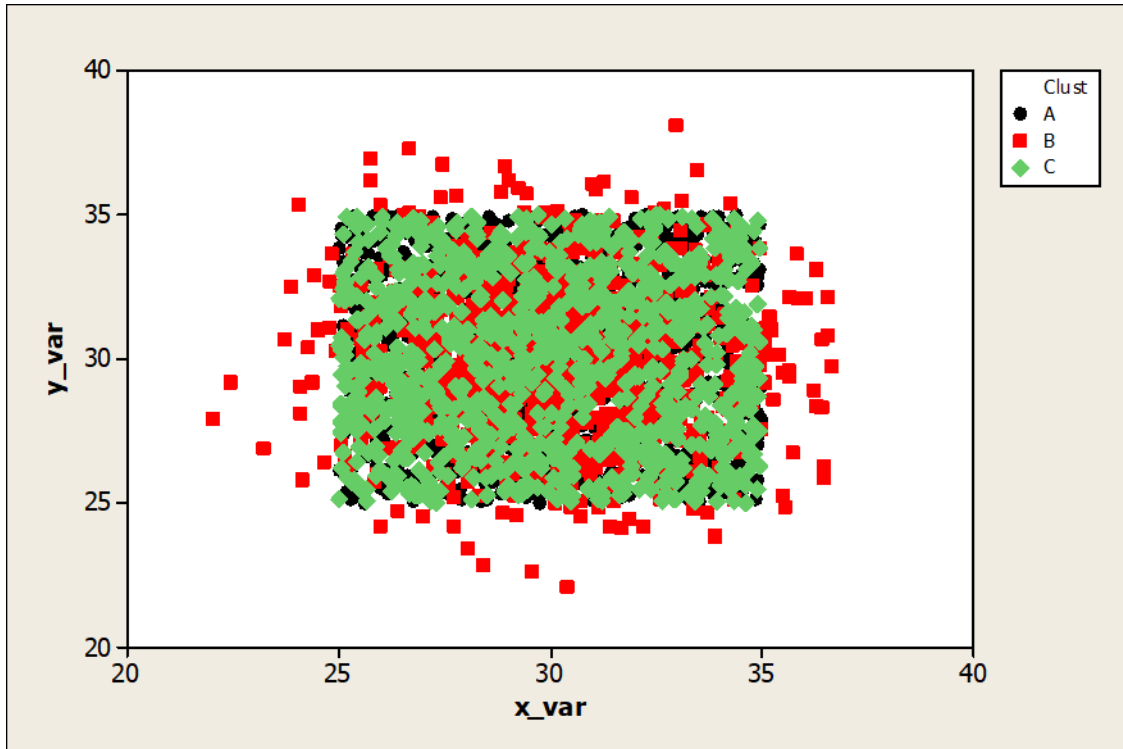


Figure 3.10: Three Cluster Scatterplot - Overlaid

The resulting distance matrices for the centroidal and symbolic methods are in Table 3.37 and Table 3.38.

Table 3.37: Three Cluster Overlaid Dissimilarity Matrix - Centroidal

	A	B	C	AVG DIST
A	0.00			0.82
B	0.79	0.00		0.90
C	0.85	1.00	0.00	0.93

Table 3.38: Three Cluster Overlaid Dissimilarity Matrix - Symbolic

	A	B	C	AVG DIST
A	0.00			0.61
B	1.00	0.00		0.97
C	0.23	0.94	0.00	0.58

When the three clusters are overlaid, the centroidal approach is forced to discern the randomness of the distances between the clusters due to scaling (when really, their means are all statistically the same). The result is that neither of the clusters is of significant distance away from the others, as would be expected when looking at the centroids alone. The symbolic approach does find a difference through a significantly higher AVG DIST for Cluster B than the other two. This represents an important advantage to using the symbolic approach to identify anomalous behavior.

### 3.4 Simulated Dataset Three

The final simulated dataset contains 5000 records. It comprises four continuous input variables that depict the concept level behavior of 10 distinct groups. Multiple instances of this dataset were assessed with the purpose of comparing the symbolic approach to a traditional centroidal approach. Instances of random noise only versus altering one input variable within one group were tested and the results observed. Following are the results of five of the run instances that best demonstrate the findings that were observed across all test situations. In each of the cases that follow, the input is described, and the R code results are displayed followed by a summary of the findings.

The first dataset represents 10 groups of data comprised of four continuous input variables where each variable within a group follows a random normal distribution  $X \sim N(15.0, 1.0)$ . The results of the R code are in Figure 3.11.

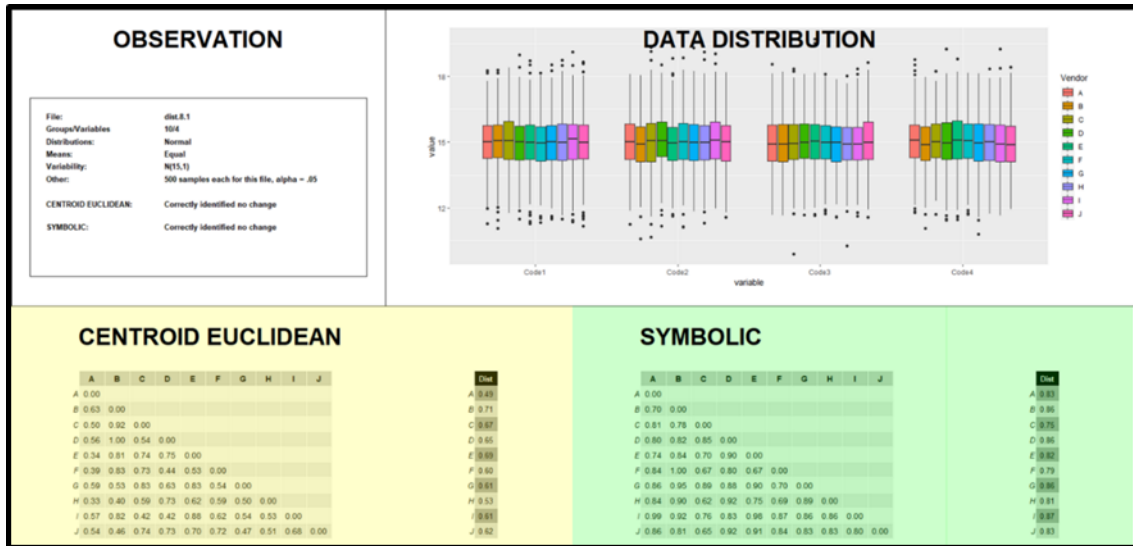


Figure 3.11: Ten Groups, Random Only

In the first test, the centroidal and symbolic approach performed comparably. Medians and thresholds were calculated for both tests and, with only random variation present, no anomalies were identified. Both methods performed as expected.

The second test introduces variation across all four input variables with respect to the last two groups, where their distributions are modified to be  $X \sim N(15.0, 5.0)$ . The results of the R code are in Figure 3.12.

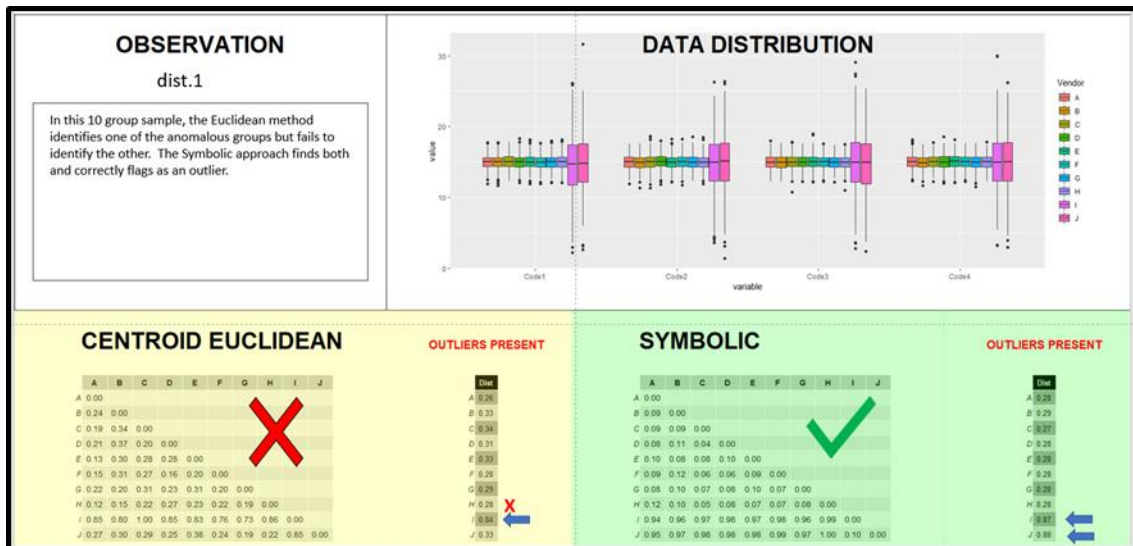


Figure 3.12: Ten Groups with Variability Change Across Two Groups

In this 10-group sample, the centroidal method identifies one of the anomalous groups but fails to identify the other. The symbolic approach finds both and correctly flags them as anomalies.

The third test injects an increase in variance within one input variable within one group. As before, all input variables are  $X \sim N(15.0, 1.0)$  except for Group H, Code 2 which is  $X \sim N(15.0, 5.0)$ . The results of the R code are in Figure 3.13.

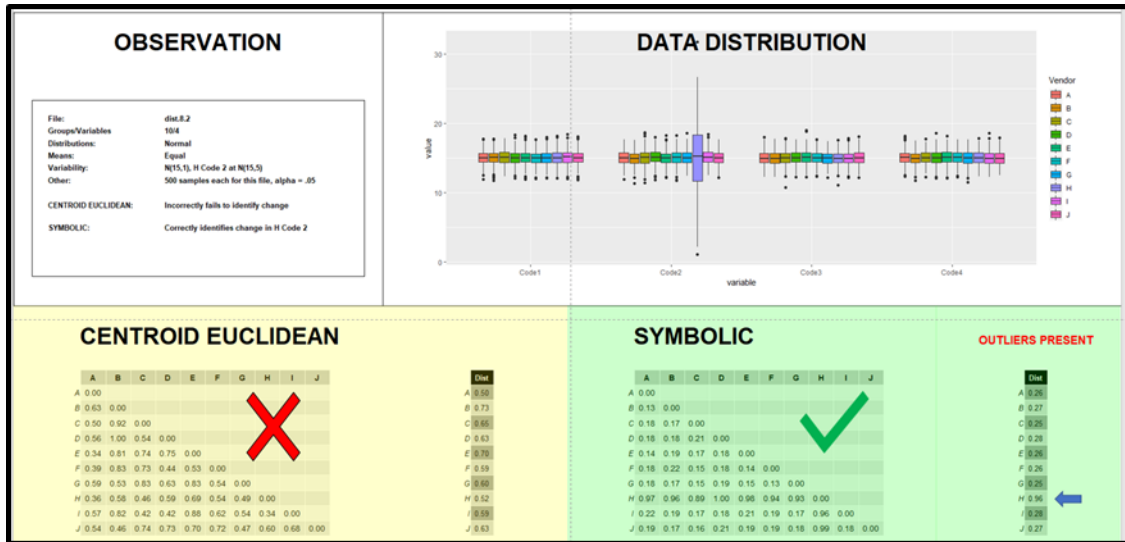


Figure 3.13: Ten Groups Introducing Variability Change

With no change in means across this 10-group sample, the centroidal method fails to identify any change. The symbolic approach does find the anomaly within group H and flags it accordingly.

The fourth test directs a shift within one input variable within one group. As before, all input variables are  $X \sim N(15.0, 1.0)$  with the exception of Group H, Code 2 which is  $X \sim N(20.0, 1.0)$ . The results of the R code are in Figure 3.14.

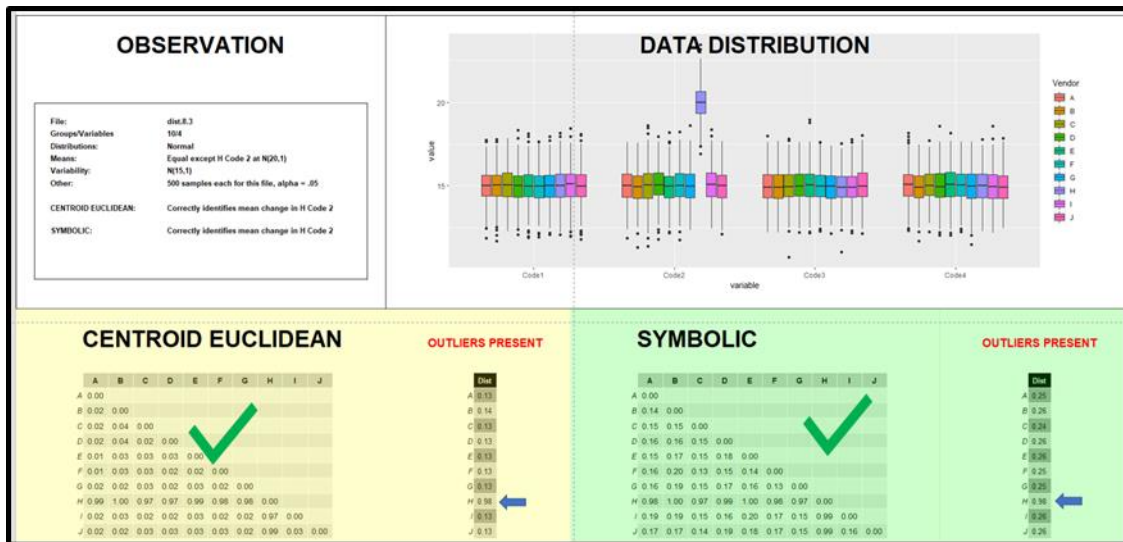


Figure 3.14: Ten Groups Introducing Mean Change

This test seeks to identify only a mean change. The centroidal method relies solely on the mean to determine the distances between the groups which makes it inherently sensitive to outliers when they are present in the data. Per the result above, the centroidal and the symbolic methods properly identified the mean change for Group H, Code 2 and flagged it accordingly.

The fifth test scenario keeps all means across the groups and input variables the same but introduces a change in distribution only. All input variables are  $X \sim N(15.0, 1.0)$  except for Group H, Code 2 which has a mean of 15.0 but is exponentially distributed and is asymmetrical. The results of the R code are in Figure 3.15.

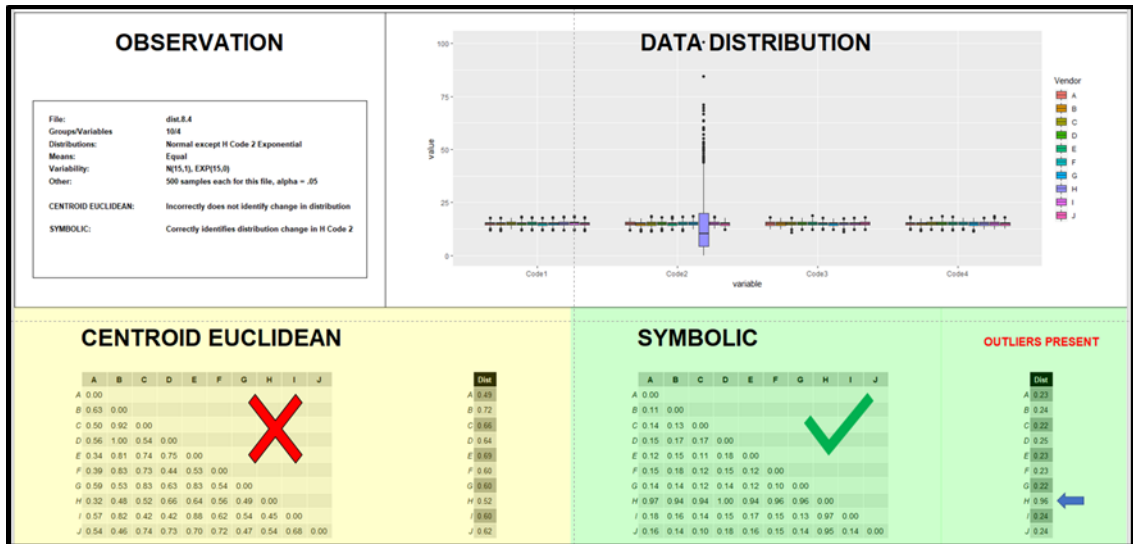


Figure 3.15: Ten Groups with an Asymmetrical Distribution

Per the results above, the centroidal method fails to find the altered Group H/Code2 where the symbolic approach correctly makes the identification.

### 3.5 The Iris Flower Dataset

With the calculations understood and the ability to study other datasets more efficiently, a more complicated but often studied set of data was selected. The Iris flower dataset is a multivariate dataset first studied by Ronald Fisher in 1936 [70]. It has become a classical dataset and test case for those researching statistical classification. It consists of three distinct species of the Iris flower with fifty data points collected of each. Each

sample records four features which are the length and width of the sepals and petals. The set is unique because only two clusters can be identified among the three species if the label of species type is unknown. This research will apply the known labels but compare the distance matrices and resolution observed using the centroidal and symbolic approach. For reference, Figure 3.16 shows a 3D scatterplot that best depicts the challenge presented by the Iris flower dataset.

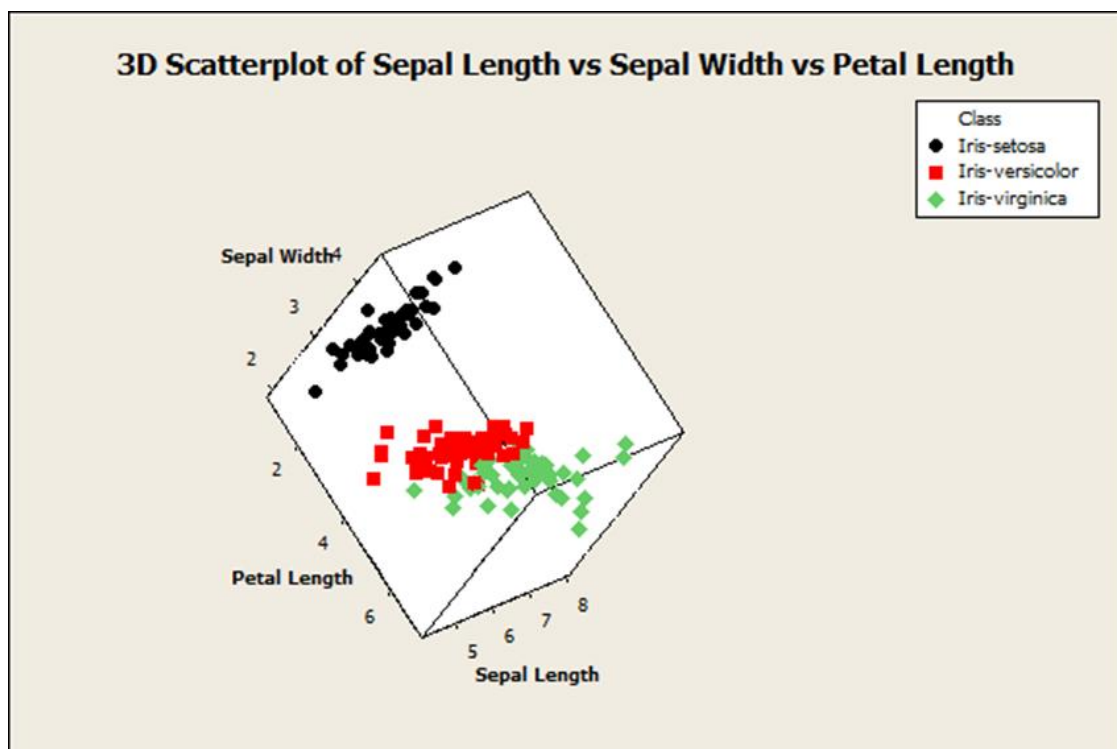


Figure 3.16: 3-D Iris Scatterplot

As with the data from Simulated Dataset One, the Iris flower dataset was assessed manually and assessed through the R code with identical results. Only the results using the R code are shown here for discussion.



The first test was performed using the four continuous variables only: Sepal Length, Sepal Width, Petal Length and Petal Width. The three species of Iris-Setosa, Iris-Versicolor and Iris-Virginica were compared. Figure 3.17 and Figure 3.18 below show the R code result for this test.

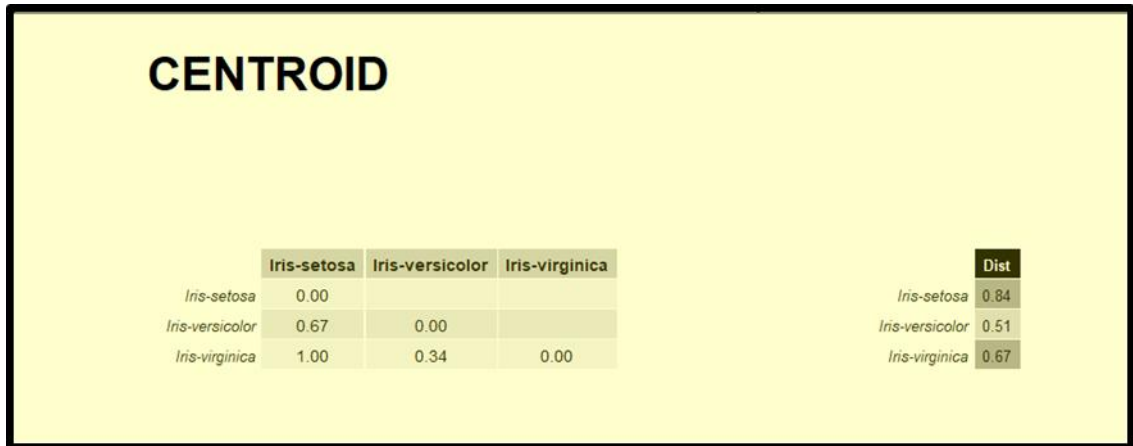


Figure 3.17: Iris R Code Centroidal Results

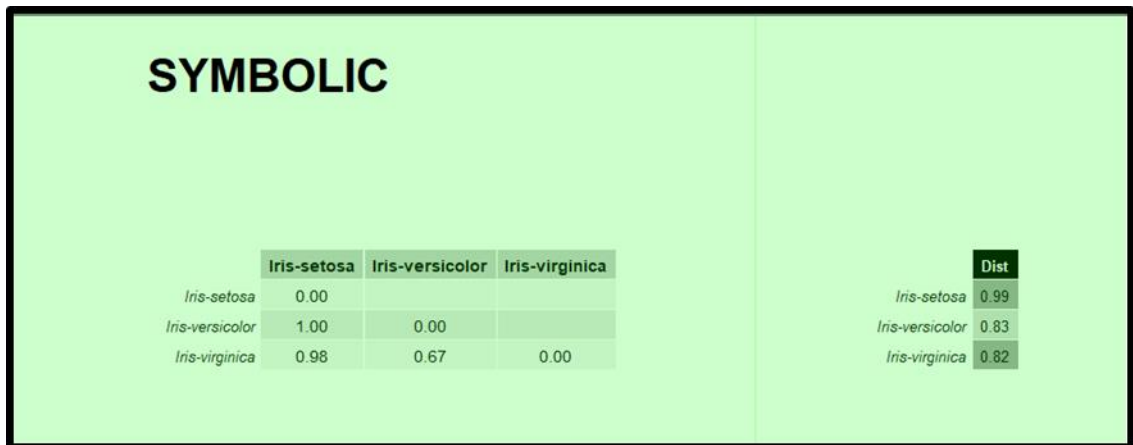


Figure 3.18: Iris R Code Symbolic Results

In order to make the comparisons, the above distance matrices show the scaled distances among the three species. In this case, the expectation for the test is not to identify an outlier but to evaluate the distance scores associated with each approach. Visually, Versicolor and Virginica appear to be close to the same distance away from Setosa. That result is best represented by the symbolic result over the centroidal approach where Versicolor is 1.00 unit away from Setosa while Virginica is 0.98 unit away. Additionally, the challenge of this problem is to find the separation between Versicolor and Virginica. Comparing the Versicolor / Virginica distance in both matrices, it can be shown that the symbolic approach has created a greater “distance” between the two species. Greater distance suggests greater resolution and separation between the two species which has traditionally been difficult to identify.

An added advantage of studying data symbolically at the concept level is that some information only exists at that level. When that information is available, it is critical to include it as part of the source data. When studying the Iris flower dataset and trying to create differences among the three species, it may be helpful to add additional identifying information if available. For example, a researcher may have knowledge of bloom color as an additional categorical variable to include. Figure 3.19 shows a possible distribution of colors by species.

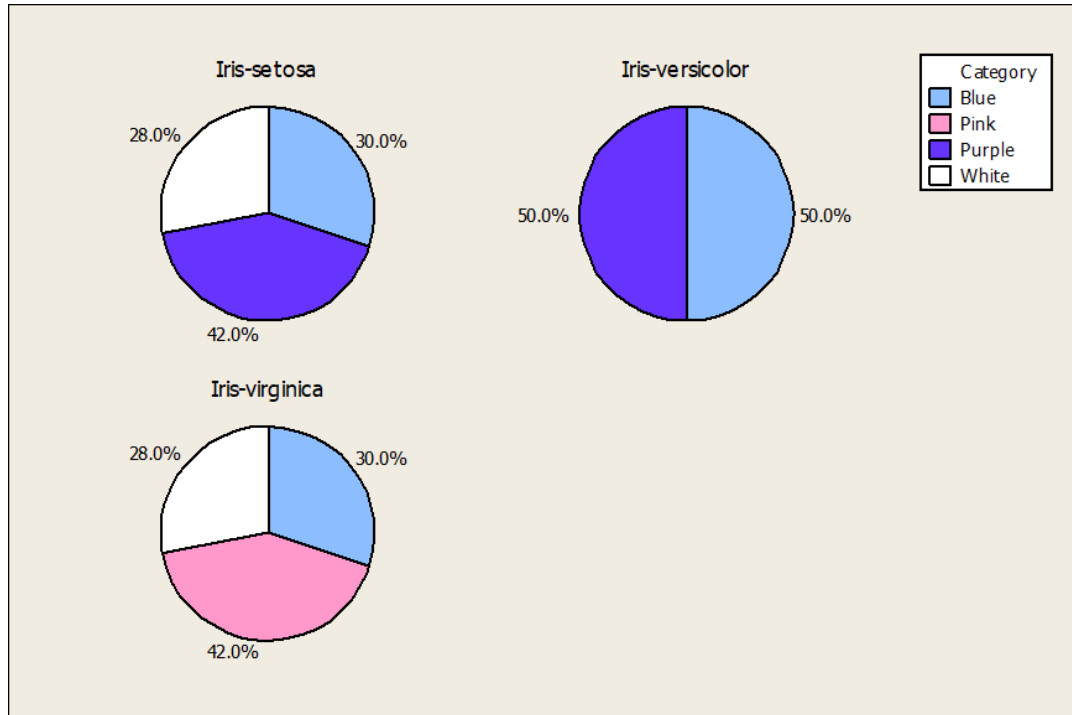


Figure 3.19: Distribution of Bloom Color

The new symbolic results with the categorical variable of bloom color included is shown in Figure 3.20.



Figure 3.20: Symbolic Table with Bloom Color Added (Modal, Categorical)

From Figure 3.20, it can be shown that greater dissimilarity is measured between Iris-versicolor and Iris-virginica, where Distance (Versicolor/Virginica) = 0.77 is greater than the previous Distance (Versicolor/Virginica) = 0.67.

Another benefit of a symbolic table is that information can be added at the concept level when specific row information is not available. Assume that the color of each individual flower bloom was not known, but the set of possible colors was known, those variables could be added as ‘set’ variables as previously explained in the literature review. Table 3.39 shows possible bloom color designations.

Table 3.39: Bloom Set Color Designations

Species	Bloom Color Set
<i>Setosa</i>	{ White, Blue, Purple }
<i>Versicolor</i>	{ Blue, Purple }
<i>Virginica</i>	{ White, Pink, Blue }

Figure 3.21 shows the symbolic results after adding the modal variable sets.



Figure 3.21: Symbolic Table with Bloom Color Added (Set, Categorical)

Categorizing variables as set variables created greater dissimilarity between the Versicolor and Virginica than the previous two examples. Of interest is the observation that the Versicolor / Virginica distance now exceeds Versicolor / Setosa ( $0.95 > 0.90$ ), suggesting that when information is available only at the concept level, its contribution can be significant. In this case, of the five bloom colors that Versicolor and Virginica can take on (Blue, Purple, White, Pink), only one color is shared between them (Blue).

### **3.6 Verification and Validation**

An important part of the research was providing verification and validation: verification that the code was performing correctly and validation that it was providing the relevant information. These two steps were performed by using the code to run the data from the Simulated Dataset One file and verifying and validating that the results obtained were identical to those calculations made by hand and that the conclusions were the same.

Centroidal and symbolic results of running the R code with just continuous variables V1, V2 and V3 are shown in Figure 3.22 and Figure 3.23. Figure 3.22 represents the centroidal approach where the median of the distances equals 0.68 and the threshold is 0.88. Figure 3.23 represents the symbolic approach where the median of the distances equals 0.75 and the threshold is 0.89.

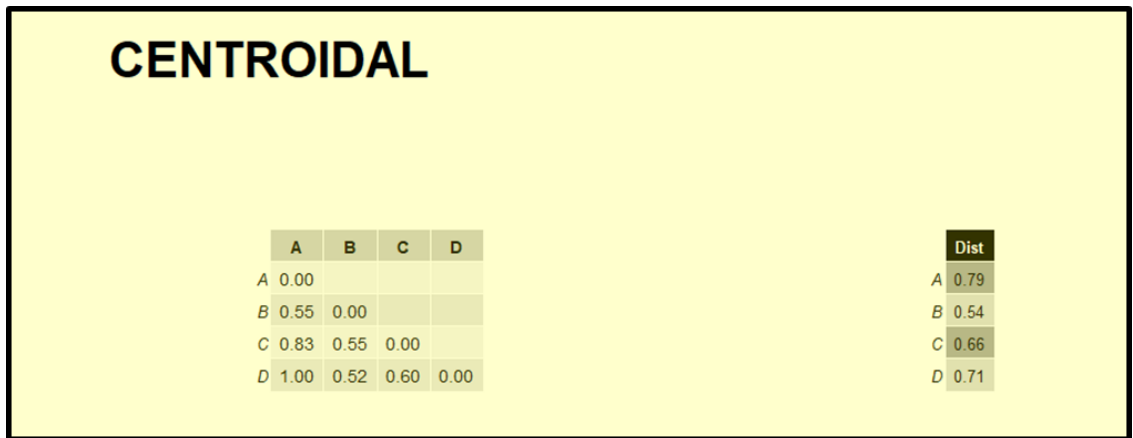


Figure 3.22: R Code Centroidal Results

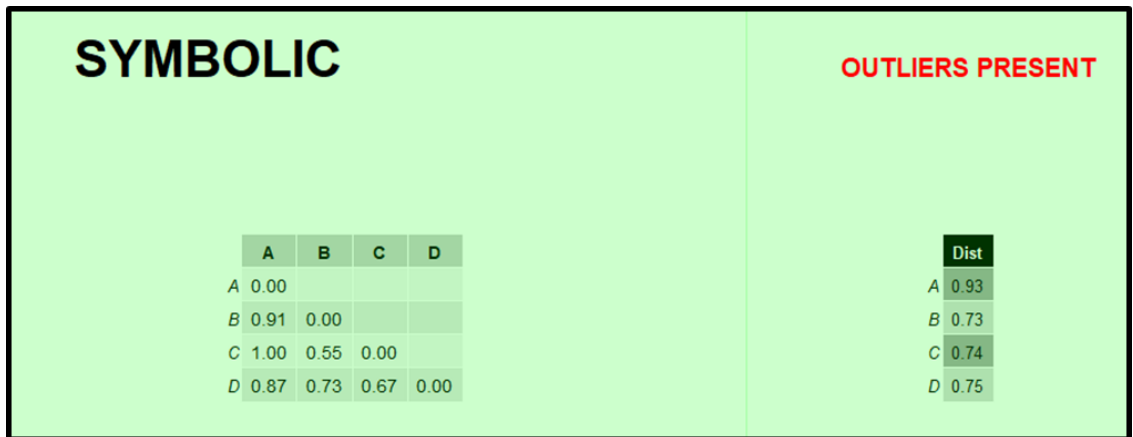


Figure 3.23: R Code Symbolic Results

As observed, the calculations for the dissimilarity matrix, average distance calculations and the threshold values match the manual calculations presented earlier in this chapter. Results confirm that the symbolic approach identifies an anomalous condition with Group A while the centroidal approach fails to do so. The code was programmed to signal an “OUTLIERS PRESENT” condition when any of the groups exceed the threshold level. Both distance matrices match the previous section and the

results obtained through the calculation of the non-parametric threshold metrics are also identical.

Similarly, Figure 3.24 and Figure 3.25 below are the centroidal and symbolic results when adding the additional categorical variable. The median of the centroidal distances equals 0.50 and the threshold is 0.81. The median of the symbolic distances equals 0.49 and the threshold is 0.83.

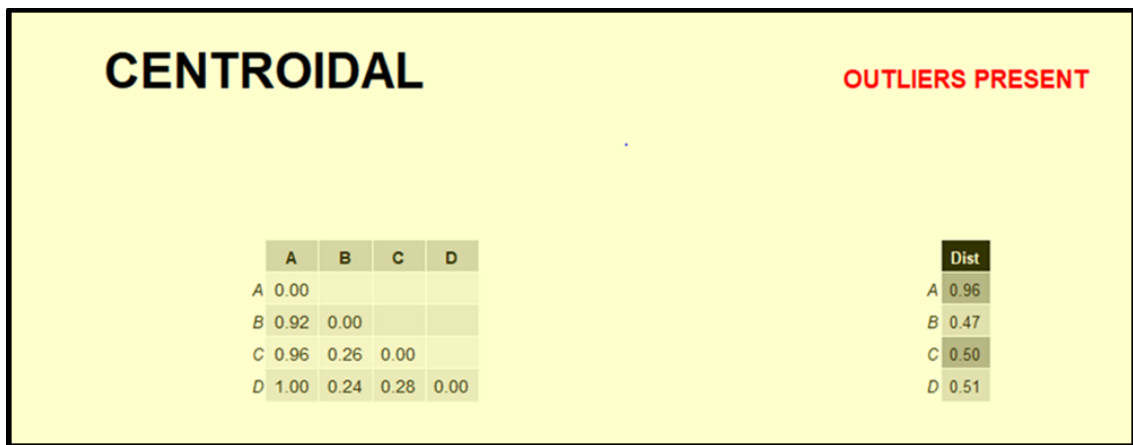


Figure 3.24: R Code Centroidal Results with Categorical Variable

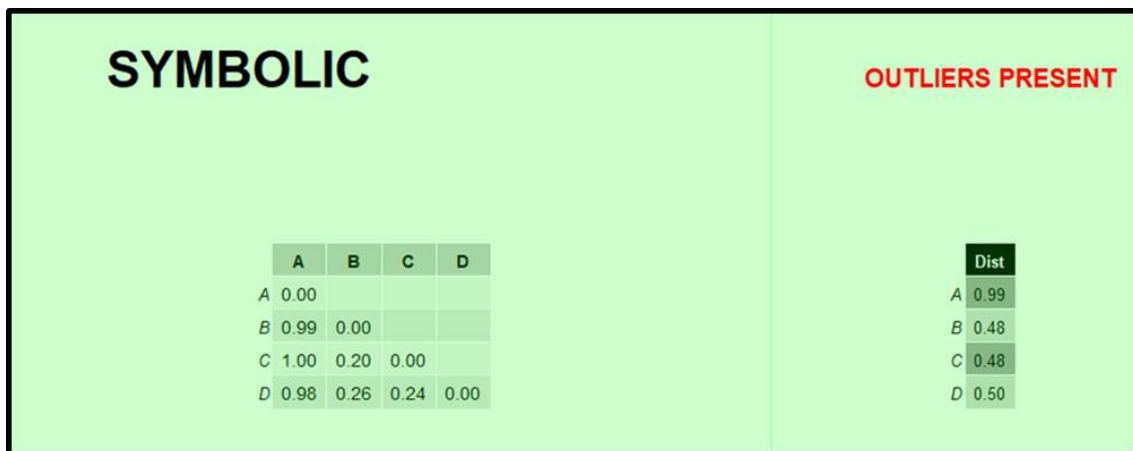


Figure 3.25: R Code Symbolic Results with Categorical Variable

As calculated in the previous chapter and verified through the R code, when a categorical variable is added and when that is done such that one group (Group A) is different than the rest, the calculations for the dissimilarity matrix, the average distance calculation, and the threshold values will change. In the above example, both methods signal an outlier when the categorical information is added.

Matching the R code to the manual spreadsheet calculations performed earlier in the chapter provided the verification and validation needed to move forward with additional experiments, knowing the underlying code and logic were correct.

### **3.7 Chapter Summary and Observations**

The purpose of this chapter was to introduce a new approach to outlier detection, specifically using a method defined as SDA. As opposed to a more conventional approach to outlier analysis using the centroids of datasets to calculate shifts in data patterns, symbolic data preserves the underlying distribution in the data which allows for greater definition and resolution when comparing between groups. In this chapter, a simulated dataset was created to demonstrate the calculations of both methods in order to compare them. The data was randomly generated, and a concept level variable was chosen from which groups could be compared. A traditional flat-file format was converted to a symbolic data table. This process involved converting continuous data into symbolic data represented by histograms. Part of histogram construction involves determining the appropriate binning approach. The Sturges binning method was determined to be the method of choice for this dissertation, although others were assessed. With data in a symbolic format, distance tables were constructed using the Euclidean formula although



other methods for distance calculations are available and may present an opportunity for future research. After converting a symbolic representation of data to a final distance matrix, a non-parametric test was used to determine the presence of outliers. A final iteration of the test was performed to study how categorical variables could be introduced to the data and assessed using this method. The effects of collinearity in variable selection was discussed and evaluated with an example. Also introduced were the tools and techniques required to operationalize symbolic anomaly detection at greater scale. An evaluation tool, the R code, was developed to allow for testing on a broader scale. The code was verified to the manual calculations and the results were validated. The tool was then used to evaluate additional simulated datasets as well as the Iris flower dataset. Additionally, a sample of the code and the noise only dataset was submitted to <https://github.com/rickjr1755-coder/SymbolicScript> for further analysis and development.

Notable observations from this chapter included:

- Construction of scaled distance matrices allow for the comparison of centroidal and symbolic approaches to anomaly detection.
- Greater resolution of differences was observed using the symbolic method.
- Adding categorical data at a concept level can enhance dissimilarity between groups when using the symbolic method.
- Non-parametric anomaly detection metrics, when scaling is applied, allow for comparison between methods.
- Setting  $\alpha = .0035$  for anomaly identification is a good default selection but ultimately at the discretion of the researcher.
- Centroidal and symbolic approaches both identify changes in group means while changes in variability and distribution were best identified by the symbolic approach.

- Collinearity of variables does influence test results but, in the studied examples, did not change the outcome.
- Binning method influences test results, although Sturges method is a good default selection.
- R code reproduces manual calculation and delivers an output that allows for verification, validation, and comparison of methods.

## **CHAPTER 4**

### **HEALTHCARE DATASET APPLICATION**

The previous two chapters of this dissertation have focused on developing a unique approach to anomaly detection and creating a test environment that can be used to identify the presence of unusual behavior. The main purpose of this research is to develop an alternative approach to discovering anomalies and apply this technique to a real-world inspired dataset to determine if anomalies in the form of FWA can be detected. The dataset studied was taken from healthcare insurance claims generated from services provided by ambulance transport suppliers. While the data was patterned after payment profiles of real healthcare providers, the actual cases were artificially generated and do not represent actual events.

The first section of this chapter provides background for choosing this dataset and examines the structure of the data that was studied. A typical ambulance service claim form is presented for reference and relevant data fields that are used throughout this part of the study are defined. A baseline case containing random common cause variability is initially presented and studied in order to create a method of comparison for the subsequent tests. With the baseline established, four common situations where errors arise in processing ambulance claims are presented and evaluated. An anomalistic event is introduced to each new scenario and is studied and compared to the baseline in order to test the suitability of the model and the approach. All instances are evaluated using a centroidal approach and a symbolic approach and are then compared to the baseline. The chapter concludes with a summary of the test results and a discussion regarding the overall performance of the model.

## **4.1 Ambulance Claims**

Due to its continuous growth and vulnerability to FWA, ambulance claims data presents an opportunity to test new approaches to anomaly detection. Ambulance transport is an industry that has changed over the last half century. Forty years ago, ambulances were staffed by volunteers or city fire departments funded by taxpayers. The American Ambulance Association estimates there are approximately 14,000 ambulance services operating in the country. Today, the industry is dominated by private companies and venture capital firms who often have trouble agreeing with insurance companies on how much to charge for services. It is particularly troubling for patients with private insurance, but problems are also present for patients that rely on federally funded healthcare plans like Medicare and Medicaid [73]. In 2015, the OIG from the Department of Health and Human Services conducted a study that concluded Medicare paid over \$50 million in improper payments to ambulance transport companies based on analyzing 7.3 million transport events during the first half of 2012 [74]. The study included analysis of transport destinations, transport levels, and mileage traveled, and concluded with multiple recommendations that included increasing the monitoring levels of ambulance billing [75]. The OIG developed specific measures to identify questionable billing events for ambulance transports. Three of the measures are directly applicable to this dissertation and are defined in Table 4.1.

Table 4.1: OIG Measures of Questionable Billing [75]


Measure	Description
Excessive Mileage for Urban Transports	High average mileage for transports for beneficiaries in urban areas. Such transports may indicate billing for more miles than suppliers actually drove or transports to facilities other than the nearest appropriate facilities.
Inappropriate or Unlikely Transport Level	High percentage of a supplier’s transports with inappropriate or unlikely transport levels given the destinations. Such transports may indicate upcoding or transport levels that were medically unnecessary.
High Number of Transports per Beneficiary	Among suppliers that provided dialysis-related transports, high average per-beneficiary number of transports. Such transports may indicate billing for transports that were medically unnecessary.

The OIG report serves as an excellent starting point from which to begin to test anomaly detection methods in ambulance transport data. While there are multiple approaches to addressing questionable billing practices of ambulance service providers, the following four hypotheses are tested in this dissertation.



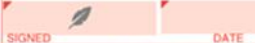
- It was hypothesized that a symbolic data approach could identify an ambulance service provider that charged for more miles as compared to its peer group of providers.
- It was hypothesized that a symbolic data approach could identify an ambulance service provider that more frequently billed for a more expensive level of service as compared to its peer group of providers.
- It was hypothesized that a symbolic data approach could identify an ambulance service provider whose average per-beneficiary number of transports were high as compared to its peer group of providers.
- It was hypothesized that a symbolic data approach could identify an ambulance service provider that incorrectly coded a charge for one of its services as compared to its peer group of providers.

The first three hypotheses stated above directly relate to situations as measured by the OIG. The final test represents an anomalous event that could come from any instance of improper coding on a claim. As demonstrated in the previous section, the symbolic approach is uniquely positioned to identify changes in mean and distribution. It is also able to handle changes in categorical variables, specifically changes in categorical distribution. The instances and hypotheses presented above represent a test of each of these situations.

The purpose of this experiment was to evaluate claim information resulting from episodes of care that required ambulance services. An ambulance claim, regardless of its origin, is required to have certain fields populated. These fields capture the detail required to render payment for the service and to provide a record of patient care. Generally, every claim captures four broad categories of information: patient data, patient health plan data, diagnoses codes, and billing detail. A sample ambulance claim is provided in Figure 4.1.

 **HEALTH INSURANCE CLAIM FORM**  
APPROVED BY NATIONAL UNIFORM CLAIM COMMITTEE (NUCC) 02/12

PICA  PICA

1. MEDICARE <input type="checkbox"/> MEDICAID <input checked="" type="checkbox"/> TRICARE <input type="checkbox"/> CHAMPVA <input type="checkbox"/> GROUP HEALTH PLAN <input type="checkbox"/> FECA BLK LUNG <input type="checkbox"/> OTHER <input type="checkbox"/> <small>(Medicare) (Medicaid) (ID# DoD#) (Member ID#) (ID#) (ID#) (ID#)</small>				1a. INSURED'S I.D. NUMBER (For Program in Item 1) <b>123456789</b>									
2. PATIENT'S NAME (Last Name, First Name, Middle Initial) <b>Doe, John</b>				3. PATIENT'S BIRTH DATE MM DD YY <b>09 06 1965</b>		4. INSURED'S NAME (Last Name, First Name, Middle Initial)							
5. PATIENT'S ADDRESS (No., Street) <b>2242 Ireland Ave</b>				6. PATIENT RELATIONSHIP TO INSURED <input type="checkbox"/> Self <input type="checkbox"/> Spouse <input type="checkbox"/> Child <input type="checkbox"/> Other		7. INSURED'S ADDRESS (No., Street)							
CITY <b>Cincinnati</b>		STATE <b>OH</b>		8. RESERVED FOR NUCC USE		CITY STATE							
ZIP CODE <b>11111</b>		TELEPHONE (Include Area Code) <b>(123) 5551234</b>		8. RESERVED FOR NUCC USE		CITY STATE ZIP CODE TELEPHONE (Include Area Code)							
9. OTHER INSURED'S NAME (Last Name, First Name, Middle Initial)				10. IS PATIENT'S CONDITION RELATED TO:				11. INSURED'S POLICY GROUP OR FECA NUMBER					
a. OTHER INSURED'S POLICY OR GROUP NUMBER				a. EMPLOYMENT? (Current or Previous) <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO				a. INSURED'S DATE OF BIRTH MM DD YY SEX M <input type="checkbox"/> F <input type="checkbox"/>					
b. RESERVED FOR NUCC USE				b. AUTO ACCIDENT? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO PLACE (State)				b. OTHER CLAIM ID (Designated by NUCC)					
c. RESERVED FOR NUCC USE				c. OTHER ACCIDENT? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO				c. INSURANCE PLAN NAME OR PROGRAM NAME					
d. INSURANCE PLAN NAME OR PROGRAM NAME				10d. CLAIM CODES (Designated by NUCC)				d. IS THERE ANOTHER HEALTH BENEFIT PLAN? <input type="checkbox"/> YES <input type="checkbox"/> NO # yes, complete items 9, 9a, and 9d.					
12. PATIENT'S OR AUTHORIZED PERSON'S SIGNATURE I authorize the release of any medical or other information necessary to process this claim. I also request payment of government benefits either to myself or to the party who accepts assignment below.								13. INSURED'S OR AUTHORIZED PERSON'S SIGNATURE I authorize payment of medical benefits to the undersigned physician or supplier for services described below.					
SIGNED  DATE				SIGNED  DATE									
14. DATE OF CURRENT ILLNESS, INJURY, or PREGNANCY (LMP) MM DD YY QUAL				15. OTHER DATE MM DD YY QUAL				16. DATES PATIENT UNABLE TO WORK IN CURRENT OCCUPATION FROM MM DD YY TO MM DD YY					
17. NAME OF REFERRING PROVIDER OR OTHER SOURCE				17a. NPI		18. HOSPITALIZATION DATES RELATED TO CURRENT SERVICES FROM MM DD YY TO MM DD YY							
19. ADDITIONAL CLAIM INFORMATION (Designated by NUCC)								20. OUTSIDE LAB? <input type="checkbox"/> YES <input type="checkbox"/> NO \$ CHARGES					
21. DIAGNOSIS OR NATURE OF ILLNESS OR INJURY Relate A-L to service line below (24E) ICD Ind.								22. RESUBMISSION CODE ORIGINAL REF. NO.					
A. <b>994.7</b>		B.		C.		D.		23. PRIOR AUTHORIZATION NUMBER					
E.		F.		G.		H.		I.					
J.		K.		L.		J.		K.					
24. A. DATE(S) OF SERVICE From MM DD YY To MM DD YY		B. PLACE OF SERVICE	C. EMG	D. PROCEDURES, SERVICES, OR SUPPLIES (Explain Unusual Circumstances) CPT-ICPCS MODIFIER			E. DIAGNOSIS POINTER	F. \$ CHARGES	G. DAYS OR UNITS	H. ICD-9 QUAL	I. RENDERING PROVIDER ID #		
<b>01 01 05 01 01 05 3</b>		<b>3</b>	<b>A0429</b>	<b>RH</b>			<b>1</b>	<b>200 00</b>	<b>1</b>	<b>NPI</b>			
<b>01 01 05 01 01 05 3</b>		<b>3</b>	<b>A0425</b>	<b>ET</b>			<b>1</b>	<b>30 00</b>	<b>6</b>	<b>NPI</b>			
25. FEDERAL TAX I.D. NUMBER		SSN EIN <input type="checkbox"/> <input type="checkbox"/>		26. PATIENT'S ACCOUNT NO. <b>12345</b>		27. ACCEPT ASSIGNMENT? (or print, explain, see back) <input checked="" type="checkbox"/> YES <input type="checkbox"/> NO		28. TOTAL CHARGE \$ <b>230 00</b>		29. AMOUNT PAID \$		30. Rsvd for NUCC Use	
31. SIGNATURE OF PHYSICIAN OR SUPPLIER INCLUDING DEGREES OR CREDENTIALS (I certify that the statements on the reverse apply to this bill and are made a part thereof.)				32. SERVICE FACILITY LOCATION INFORMATION				33. BILLING PROVIDER INFO & PH #					
SIGNED  DATE				a. NPI				b. NPI					

NUCC Instruction Manual available at: www.nucc.org PLEASE PRINT OR TYPE APPROVED OMB-0938-1197 FORM 1500 (02-12)

Figure 4.1: Ambulance CMS 1500 Claim Form

Because of the concise nature of the data on this form, the study of anomalies with ambulance billing proved a reasonable testing ground to assess several situations which the literature suggests are candidates for FWA. The fields, while sometimes complex in description, are limited in number. The data used to test these hypotheses was patterned from a real-world dataset. The fields that ultimately remained in the analysis and their definitions are in Table 4.2.

Table 4.2: Ambulance Claims Data Definitions

<b>Name</b>	<b>Description</b>
Age	Patient age
AutoAcc	Is the claim related to an automobile accident (Y/N)
Destination	The destination code for the trip (H)
Service	Codes representing billed charges (A0425, A0427, A0429, A0433)
ICD	Diagnosis code used to identify the nature of a medical condition
Origin	The origin code for the trip (R, S, E, P, N, J, H)
Sex	Patient gender (M/F)
Charges	Dollars charged for service
Units	# units for billed service (actual mileage if mileage charge, 1 for all others)
Provider	Provider indicator (20 providers denoted A-T)



Service type was limited to four codes. Table 4.3 depicts the codes and their descriptions.

Table 4.3: Service Code Definitions [76]

Service Code	Description
A0425	Ground Mileage
A0427	Advanced Life Support, Level 1 (ALS1) - Includes the provision of medically necessary supplies and services and the provision of an ALS assessment or at least one ALS intervention. An ALS assessment is performed by an ALS crew as part of an emergency response that is necessary because the beneficiary's reported condition at the time of dispatch indicates only an ALS crew is qualified to perform the assessment. An ALS assessment does not necessarily result in a determination that the beneficiary requires an ALS level of transport. In the case of an appropriately dispatched ALS emergency service, if the ALS crew completes an ALS assessment, the services provided by the ambulance transportation service provider or supplier are covered at the ALS emergency level. This is regardless of whether the beneficiary required ALS intervention services during the transport, provided the ambulance transportation itself was medically reasonable and necessary and all other coverage requirements are met. An ALS intervention must be performed by an emergency medical technician-intermediate (EMT-Intermediate) or an EMT-Paramedic in accordance with State and local laws. An ALS1 emergency is an immediate emergency response in which you begin as quickly as possible to take the steps necessary to respond to the call.
A0429	Basic Life Support (BLS) - Includes the provision of medically necessary supplies and services and BLS ambulance transportation as defined by the State where you provide the transport. An emergency response is one that, at the time you are called, you respond immediately. A BLS emergency is an immediate emergency response in which you begin as quickly as possible to take the steps necessary to respond to the call.

A0433	<p>Advanced Life Support, Level 2 (ALS2) - Includes the provision of medically necessary supplies and services, involving:</p> <ul style="list-style-type: none"> <li>•At least three separate administrations of one or more medications by intravenous push/bolus or by continuous infusion (excluding crystalloid fluids) or</li> <li>•At least one of these ALS2 procedures: <ul style="list-style-type: none"> <li>–Manual defibrillation/cardioversion</li> <li>–Endotracheal intubation</li> <li>–Central venous line</li> <li>–Cardiac pacing</li> <li>–Chest decompression</li> <li>–Surgical airway</li> <li>–Intraosseous line</li> </ul> </li> </ul>
-------	--

Table 4.4: Destination and Origin Code Definitions

<b>Destination / Origin Code</b>	<b>Description</b>
R	Residence
S	Scene of Accident
E	Custodial Facility
P	Physician’s Office
N	Skilled Nursing Facility
J	Dialysis Facility
H	Hospital

Table 4.4 shows the origin and destination codes that are pulled from the modifier field on the sample form. For example, RH implies a transport originating at a residence and terminating at a hospital. Age is derived from the birthdate field.

## **4.2 Establishing a Baseline**

Using data from an actual provider database, a sample file was constructed that consisted of 20 randomly designed fictitious providers each exhibiting the same characteristics but with some inherent variability still present. While the providers were fictitious, the distribution of patient ages, service codes, ambulance routing, and dollars paid for services for each of the providers created were based on values that would be expected in a real-world scenario. This patient data was from Medicare recipients only and was used because of the availability of the real-world data to study as well as the publicly available data, previously cited, that could be used for comparison. Table 4.5 is a partial rendering of the dataset used for the baseline case. Table 4.6 provides quantitative and qualitative information for each of the providers as well as information pertaining to the descriptors used for each provider.

Table 4.5: Baseline Data

<b>i</b>	<b>C<sub>1</sub></b>	<b>C<sub>2</sub></b>	<b>C<sub>3</sub></b>	<b>C<sub>4</sub></b>	<b>C<sub>5</sub></b>	<b>C<sub>6</sub></b>	<b>C<sub>7</sub></b>	<b>C<sub>8</sub></b>	<b>C<sub>9</sub></b>	<b>C<sub>10</sub></b>
1	49.12	N	H	A0427	25080	S	F	443.91	1.00	A
2	48.91	N	H	A0427	78650	P	M	464.14	1.00	A
3	76.25	N	H	A0425	71946	E	F	70.9	10.00	A
4	60.36	N	H	A0427	78079	R	M	443.91	1.00	A
5	80.70	N	H	A0425	7804	R	M	63.81	9.00	A
6	88.80	N	H	A0427	71945	R	F	243.91	1.00	A
7	71.50	N	H	A0425	7881	R	M	85.08	12.00	A
8	66.44	N	H	A0427	78650	R	F	243.91	1.00	A
9	60.85	N	H	A0433	7991	R	M	642.49	1.00	A
10	86.54	N	H	A0427	71945	R	F	443.91	1.00	A
.										
.										
.										

Table 4.6: Baseline Data Descriptors

<b>C<sub>i</sub></b>	<b>Description</b>	<b>Variable Type</b>	<b>Values</b>
C <sub>1</sub>	Age	Continuous	[23.2, 106.8]
C <sub>2</sub>	AutoAcc	Categorical	{Y, N}
C <sub>3</sub>	Destination	Categorical	{H}
C <sub>4</sub>	Service	Categorical	{A0425, A0427, A0429, A0433}
C <sub>5</sub>	ICD	Categorical	Multiple
C <sub>6</sub>	Origin	Categorical	{R, S, E, P, N, J, H}
C <sub>7</sub>	Sex	Categorical	{M, F}
C <sub>8</sub>	Charges	Continuous	[7.09, 643.91]
C <sub>9</sub>	Units	Continuous	[1, 53]
C <sub>10</sub>	Provider	Categorical	20 Values {A - T}

In order to create a true baseline to which other instances could be compared, a dataset that contained provider groups with statistically similar characteristics had to be created. Each of the 20 test providers (which comprised of 5,000 cases each) had to exhibit the same general characteristics while maintaining underlying common cause variability. The process for accomplishing this was to generate one “parent” provider dataset that consisted of 100,000 rows. The parent set was modeled using parameters from a typical ambulance service provider. An initial real-world sample comprised of 2,122 cases from a single provider within one specific geographic region was collected. The cases consisted of only Medicare patients with various presenting medical conditions. The data included transports that terminated at a hospital location but could have originated at several locations. The 100,000-row test dataset, 20 providers of 5,000 cases each, was derived from this initial 2,122 case set. The initial set was randomly sampled, with replacement, to generate a 5,000-row provider sample. That process was repeated 20 times using a different random seed each time. The result was a twenty case, 100,000-row “parent” dataset from which to begin the testing. The twenty cases represented fictitious providers and were labeled {A, B, C, ..., S, T}. It should be noted that in each of the examples when an anomaly was introduced by changing a dataset, that change always occurred within ‘Provider A’.

Graphical representations of the descriptor variables of the parent set, and which represent the characteristics of each of the 20 providers, are represented in the charts below.

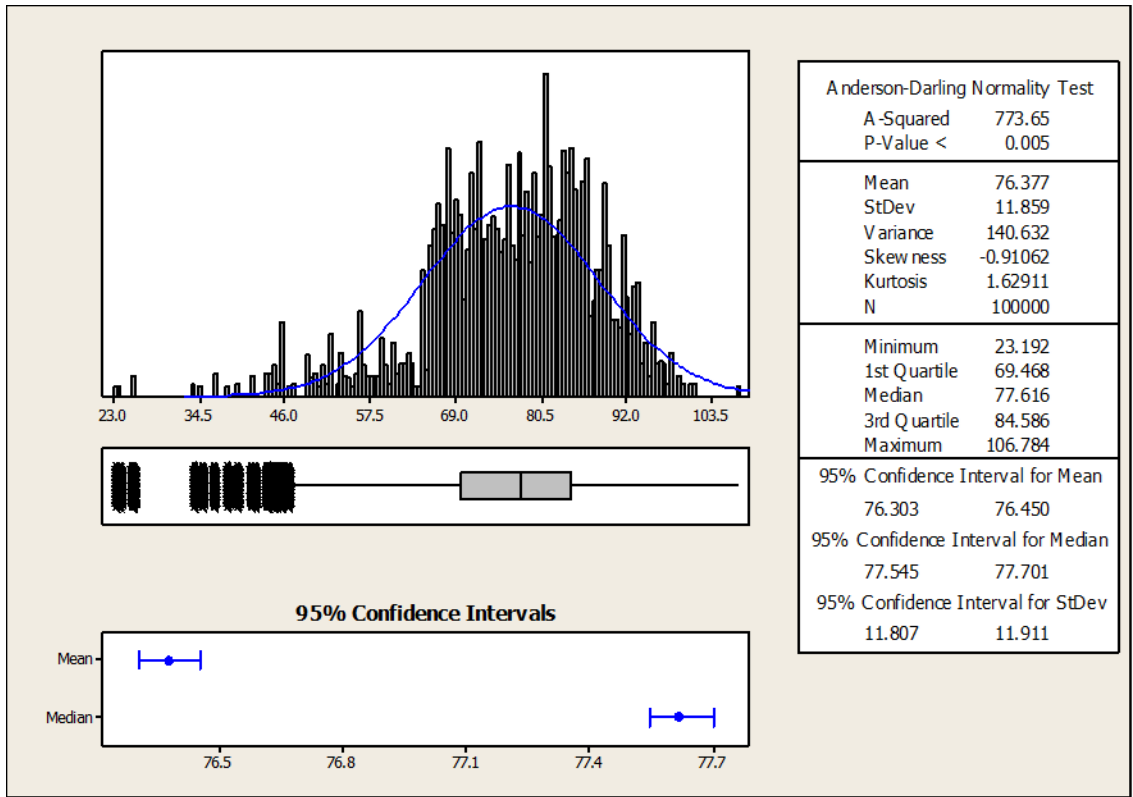


Figure 4.2: Graphical Summary of Age

Figure 4.2 represents a graphical summary of the ages of all Medicare patients. Key statistics from the graphical summary above show 100,000 individual cases with an average age of 76.377. The minimum age is 23.192 and the maximum is 106.784. The typical Medicare beneficiary is over the age of 65, although situations exist for younger people to also qualify for benefits such as people with a disability or those diagnosed with end stage renal disease (kidney failure). Approximately 15% of all Medicare beneficiaries are under the age of 65 [77].

Figure 4.3 and Figure 4.4 reflect characteristics of the 100,000-row sample dataset. Figure 4.3 reflects that only 1.3% of all transport cases were related to an automobile accident. Figure 4.4 displays the percentage of male and female patients.

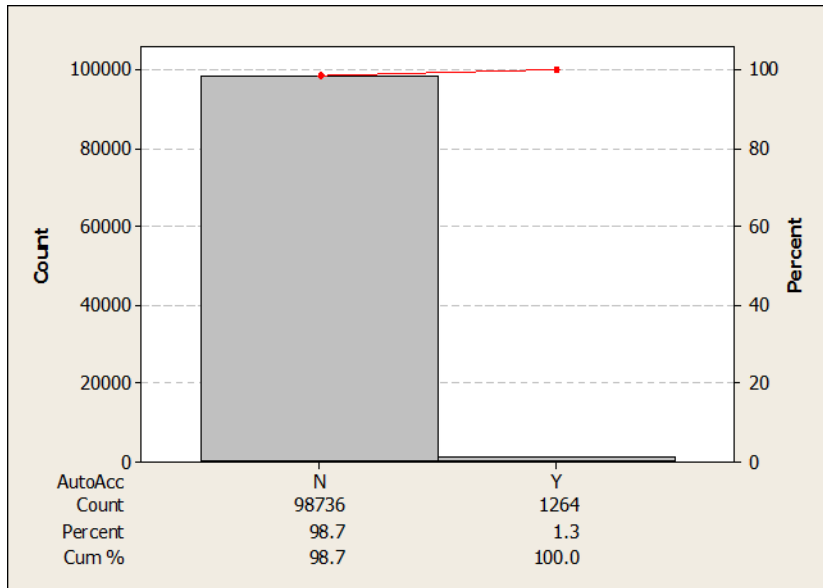


Figure 4.3: Pareto Chart of AutoAcc

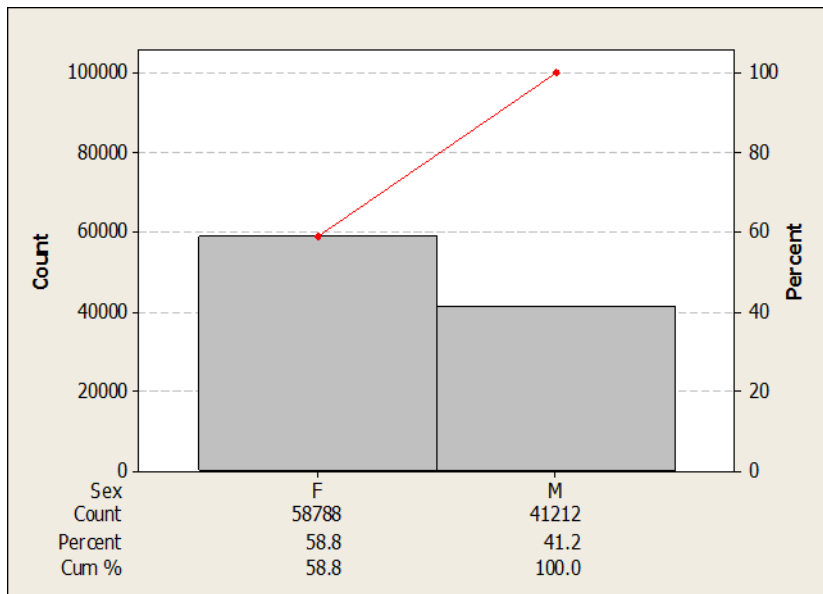


Figure 4.4: Pareto Chart of Gender

Figure 4.5 refers to the service definitions introduced in Table 4.3. Because the mileage code (A0425) is associated with every transport, it represents the greatest representation in the dataset, while code A0427, Advanced Life Support, is the level of service billed most frequently when compared to the other two levels.

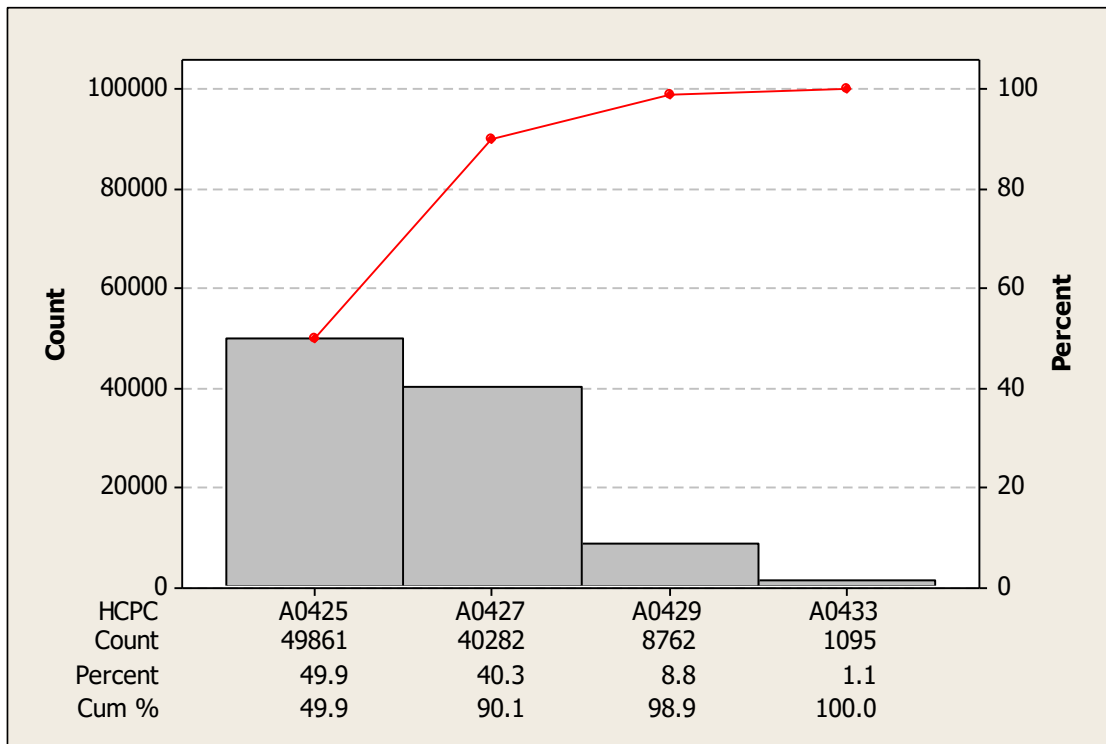


Figure 4.5: Pareto Chart of Service Code



Figure 4.6 reveals that 73.3 percent of all transports originate at the patient's residence per the codes previously defined. The destination for this dataset was limited to 'H', or hospital only.

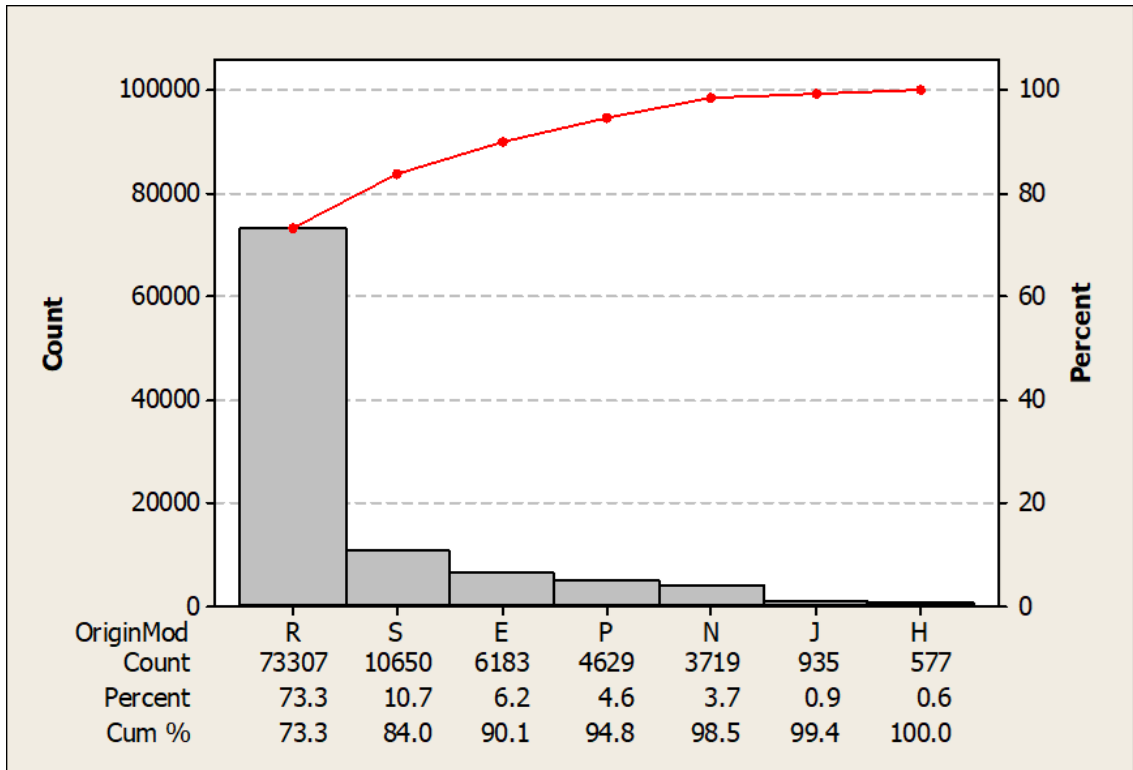


Figure 4.6: Pareto Chart of Origin Code

The first step in testing the hypotheses was to measure the Euclidean distances between pairs of 20 groups in the baseline dataset where only inherent random variability was present. Using the R code, the data was analyzed using provider as the concept variable and Age, AutoAcc, Service, ICD, Origin, Sex, and Units as the input variables. Destination code was kept in the dataset but excluded from all analyses because it has a singular value of H (Hospital) for all cases. Additionally, column C8 (Charges) was excluded from the initial baseline case. Because of its correlation with increasing mileage and upcoding transport levels, it acts as a dependent variable whose influence is not desired when analyzing the underlying causes for change. The Charges variable was added back to the dataset for Case 4 where a price change occurs because it is the only variable to reflect this change.

For all tests in this chapter, the centroidal approach was applied first followed by the symbolic approach. The results were then compared against each other using the metrics previously established. Figure 4.7 is the resulting distance matrix for the baseline dataset created when applying the centroidal approach.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A	0.00																			
B	0.30	0.00																		
C	0.25	0.09	0.00																	
D	0.35	0.20	0.16	0.00																
E	0.41	0.25	0.21	0.08	0.00															
F	0.23	0.12	0.06	0.15	0.21	0.00														
G	0.12	0.23	0.16	0.24	0.30	0.13	0.00													
H	0.39	0.37	0.31	0.20	0.20	0.28	0.29	0.00												
I	0.18	0.42	0.35	0.39	0.43	0.32	0.20	0.32	0.00											
J	0.16	0.46	0.40	0.49	0.54	0.38	0.26	0.48	0.18	0.00										
K	0.35	0.38	0.31	0.23	0.23	0.28	0.26	0.06	0.27	0.43	0.00									
L	0.41	0.26	0.23	0.08	0.06	0.22	0.30	0.18	0.42	0.54	0.22	0.00								
M	0.67	0.41	0.43	0.33	0.30	0.44	0.56	0.46	0.71	0.81	0.51	0.30	0.00							
N	0.30	0.21	0.15	0.10	0.13	0.14	0.18	0.17	0.31	0.42	0.17	0.14	0.41	0.00						
O	0.17	0.16	0.14	0.29	0.34	0.15	0.15	0.40	0.32	0.32	0.38	0.35	0.56	0.26	0.00					
P	0.60	0.38	0.38	0.25	0.20	0.38	0.49	0.35	0.62	0.74	0.39	0.20	0.13	0.32	0.51	0.00				
Q	0.64	0.37	0.39	0.30	0.27	0.41	0.53	0.44	0.68	0.78	0.48	0.28	0.06	0.38	0.52	0.12	0.00			
R	0.27	0.05	0.07	0.20	0.25	0.09	0.19	0.35	0.38	0.42	0.35	0.26	0.43	0.19	0.13	0.39	0.40	0.00		
S	0.87	0.64	0.64	0.52	0.46	0.65	0.75	0.57	0.87	1.00	0.62	0.46	0.26	0.58	0.78	0.27	0.29	0.66	0.00	
T	0.64	0.42	0.42	0.30	0.25	0.43	0.53	0.39	0.67	0.78	0.44	0.25	0.11	0.37	0.55	0.06	0.11	0.43	0.23	0.00

Figure 4.7: Baseline Case Distance Matrix - Centroidal

Once the distances between each pair was calculated, the average distance between each individual provider and the other 19 providers was calculated. As explained in Chapter 3, these are the distances between each individual pair of points in the dataset calculated using the Euclidean method. Table 4.7 represents the final average distance calculations for each of the providers using the centroidal approach. These numbers represent the average distance that each individual provider is from every other provider in the dataset.

Table 4.7: Baseline Case Distance Measures - Centroidal

<b>Provider</b>	<b>Distance</b>	<b>Provider</b>	<b>Distance</b>	<b>Provider</b>	<b>Distance</b>	<b>Provider</b>	<b>Distance</b>
A	0.385	F	0.267	K	0.336	P	0.357
B	0.301	G	0.309	L	0.273	Q	0.393
C	0.271	H	0.327	M	0.416	R	0.290
D	0.256	I	0.424	N	0.260	S	0.586
E	0.269	J	0.505	O	0.341	T	0.388

The same process was then applied using the symbolic approach as previously explained. The next step in establishing the baseline was to apply the symbolic approach to the 100,000-row parent dataset. Figure 4.8 is the resulting distance matrix created when applying the symbolic approach.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A	0.00																			
B	0.70	0.00																		
C	0.72	0.75	0.00																	
D	0.67	0.83	0.76	0.00																
E	0.74	0.79	0.60	0.80	0.00															
F	0.69	0.64	0.76	0.68	0.80	0.00														
G	0.68	0.82	0.56	0.76	0.75	0.69	0.00													
H	0.89	0.98	0.67	0.83	0.68	0.83	0.57	0.00												
I	0.69	0.98	0.82	0.80	0.78	0.77	0.71	0.73	0.00											
J	0.70	0.86	0.59	0.79	0.65	0.76	0.55	0.70	0.69	0.00										
K	0.78	0.87	0.57	0.76	0.61	0.76	0.59	0.54	0.74	0.67	0.00									
L	0.75	1.00	0.80	0.70	0.82	0.82	0.71	0.72	0.71	0.89	0.68	0.00								
M	0.80	0.79	0.72	0.72	0.79	0.83	0.70	0.81	0.91	0.73	0.74	0.87	0.00							
N	0.99	0.85	0.82	0.86	0.82	0.89	0.80	0.70	0.97	0.93	0.71	0.95	0.80	0.00						
O	0.68	0.74	0.60	0.75	0.56	0.67	0.61	0.55	0.70	0.64	0.55	0.73	0.70	0.73	0.00					
P	0.74	0.83	0.60	0.65	0.56	0.70	0.63	0.58	0.69	0.61	0.58	0.68	0.73	0.77	0.54	0.00				
Q	0.86	0.63	0.82	0.85	0.78	0.67	0.83	0.88	0.94	0.84	0.78	0.96	0.78	0.79	0.70	0.72	0.00			
R	0.76	0.59	0.67	0.73	0.71	0.70	0.64	0.76	0.85	0.77	0.73	0.76	0.67	0.68	0.65	0.67	0.65	0.00		
S	0.71	0.86	0.61	0.68	0.63	0.76	0.58	0.54	0.75	0.72	0.53	0.57	0.66	0.75	0.53	0.62	0.82	0.68	0.00	
T	0.88	0.79	0.72	0.73	0.68	0.69	0.75	0.67	0.87	0.80	0.71	0.85	0.77	0.60	0.61	0.63	0.77	0.63	0.67	0.00

Figure 4.8: Baseline Case Distance Matrix - Symbolic

Table 4.8 represents the final distance calculations for each of the providers using the symbolic approach.

Table 4.8: Baseline Case Distance Measures - Symbolic

<b>Provider</b>	<b>Distance</b>	<b>Provider</b>	<b>Distance</b>	<b>Provider</b>	<b>Distance</b>	<b>Provider</b>	<b>Distance</b>
A	0.759	F	0.743	K	0.680	P	0.659
B	0.805	G	0.681	L	0.788	Q	0.793
C	0.691	H	0.718	M	0.765	R	0.700
D	0.756	I	0.796	N	0.812	S	0.667
E	0.713	J	0.730	O	0.644	T	0.727

The performance measures calculated for both the centroidal and symbolic approach for the baseline dataset are in Table 4.9.

Table 4.9: Performance Metrics for Baseline Dataset

<b>Metric</b>	<b>Centroidal</b>	<b>Symbolic</b>
<i>MED</i>	.331	.728
<i>IQR</i>	.117	.082
<i>Threshold</i>	.565	.894

Where, *MED* represents the median value, *IQR* represents the interquartile range, and *Threshold* represents the calculated non-parametric threshold value of the average distances in Table 4.7 and Table 4.8 respectively.

In the centroidal baseline case alone, Provider S has a value of 0.586 which exceeds the calculated threshold value. The occurrence of a value exceeding the threshold is included in this chapter in order to remain consistent with the previous examples. It is important to note, that any identified outlier in the baseline dataset, whether calculated by the centroidal or symbolic approach, is simply a reflection of the common cause variability in the data.

The purpose for establishing a baseline case is to have a randomized dataset from which subsequent tests can be compared to. In a truly randomized set of samples, there will exist common cause variability within each provider. Subsequent hypotheses in this chapter are tested and compared to this set in order to measure the change from what can be assumed to be a normal steady-state scenario. The performance measures for all subsequent tests are calculated for the individual test data as well as for the differences from the baseline. Measurable differences from the established baseline are the ultimate measure of model performance.

### **4.3 Case 1 – Excessive Mileage**

In the 2015 OIG Report, it was noted that one of the areas researched for questionable billing was the recording of excessive mileage: specifically, excessive mileage for urban transports. The report found that 4% of ambulance service providers recorded higher than normal mileage for patients residing in urban areas. These excesses totaled \$7.3 million during the first half of 2012. The average urban transport was calculated as 10 miles while the mileage recorded by the questionable providers was more than three times that level. Reasons included not transporting to the nearest appropriate facility or simply billing for more miles than were driven [75].

The first real-world scenario test was to determine if a symbolic approach could identify this type of anomaly and how would its result compare to a more traditional centroidal approach. Mileage is recorded on the CMS-1500 form using service code A0425. In the original 2,122 case sample, there were 773 instances of trips originating at a residence and terminating at a hospital. The average mileage charged for these trips was 8.3 miles. This is consistent with the data reported in the OIG Report.

In order to simulate an excessive mileage event, the 5,000-row simulated Provider A data was isolated. As expected, its statistical characteristics matched the summary above. In this scenario and the ones that follow in this chapter, anomalistic events were modeled in the Provider A group for convenience although the event could potentially occur in any of the provider groups. In order to simulate additional rows of miles data as accurately as possible, it was necessary to study the underlying distribution of data from the real dataset. The original 773 rows of data containing miles traveled were evaluated. Using Minitab, the distribution was compared to several known distributions, checked for



goodness of fit, and the parameters recorded. The 3-parameter Weibull distribution proved to have the best fit and best matched the data visually, where shape = 1.26857, scale = 8.03881 and threshold = 0.92885. To simulate the anomalous event, a similar distribution was generated which represented an increase in miles billed as described in the OIG Report. The parameters from the fitted distribution were kept the same except for the scale parameter which was increased 400% to match the scenario in the OIG report. The new distribution was a 3-parameter Weibull distribution where shape = 1.26857, scale = 32.155 and threshold = .92855. The values were then rounded to the nearest integer. Figure 4.9 shows the original mileage data with an average of 8.3799 miles traveled. Figure 4.10 is a graphical summary of the altered data.

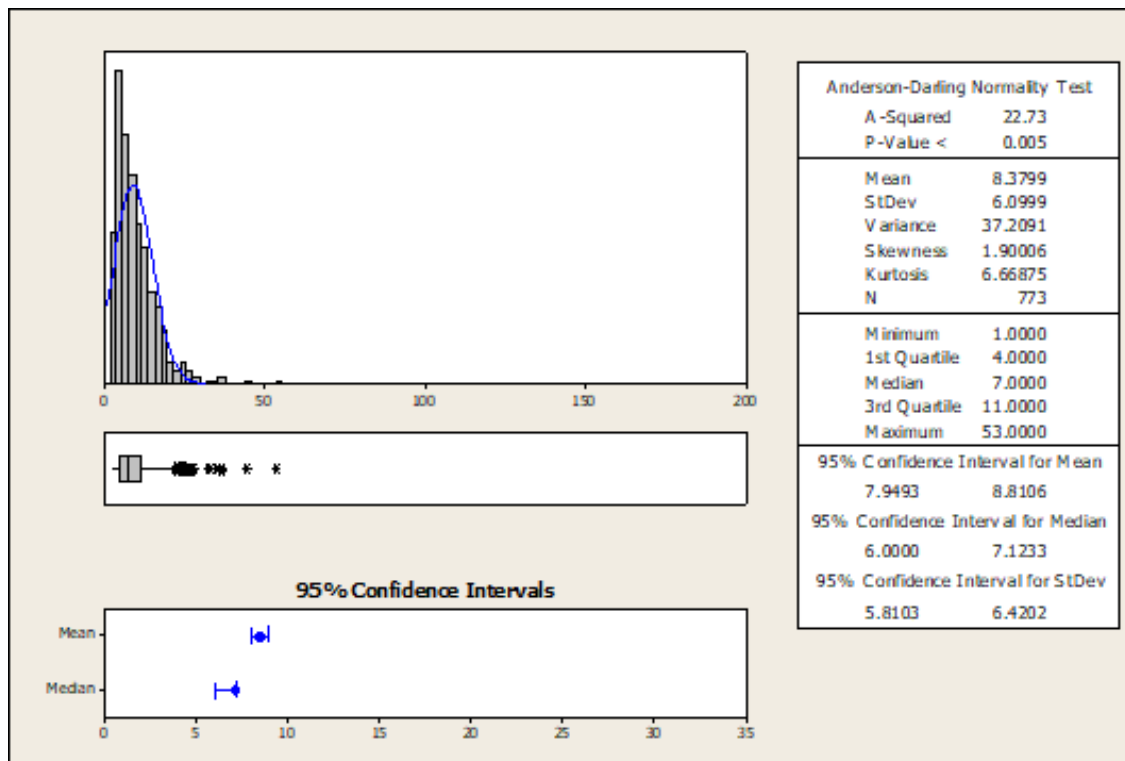


Figure 4.9: Graphical Summary of Miles Traveled from Residence to Hospital

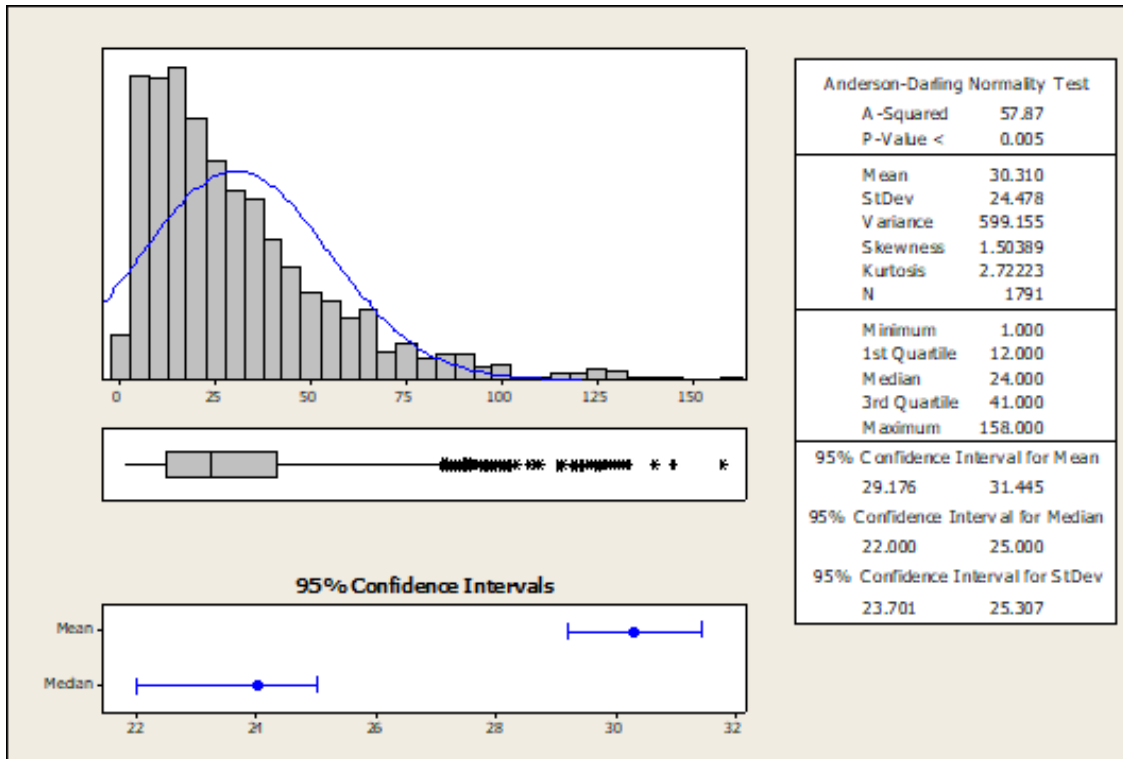


Figure 4.10: Graphical Summary of Excessive Mileage

The initial 1,791 cells of mileage data (units when accompanied by service code A0425) for Provider A were replaced with the altered data represented above. The simulated data retains the same distribution but shifts the mean to 30.310 miles to mimic the scenario explained in the OIG report. The centroidal and symbolic tests were applied to the data to determine if the anomalistic event could be identified.

The process for the analysis is the same throughout this chapter. Distance matrices are created for each approach and the average distance measures are created. For purposes of brevity, the distance matrices will not be displayed for this and the subsequent examples.

The final distance measures are displayed, the threshold metrics are calculated, and the results are compared to each other and the baseline data. Also calculated is the difference (shown as  $\Delta$  in the table) between the baseline value and the scenario data. Positive difference metrics from the baseline suggest a change in the data behavior in the form of greater disparity from the base random case. Table 4.10 is the results table for Case 1: The Excessive Mileage Scenario.

Table 4.10: Results Table – Case 1

Provider	CENTROIDAL			SYMBOLIC		
	Baseline	Case 1	$\Delta$	Baseline	Case 1	$\Delta$
A	0.385	0.988	0.603	0.759	0.965	0.206
B	0.301	0.076	-0.225	0.805	0.236	-0.569
C	0.271	0.073	-0.198	0.691	0.211	-0.480
D	0.256	0.071	-0.185	0.756	0.230	-0.526
E	0.269	0.072	-0.197	0.713	0.218	-0.495
F	0.267	0.073	-0.194	0.743	0.215	-0.528
G	0.309	0.076	-0.232	0.681	0.206	-0.475
H	0.327	0.076	-0.251	0.718	0.204	-0.514
I	0.424	0.085	-0.339	0.796	0.240	-0.555
J	0.505	0.092	-0.413	0.730	0.221	-0.508
K	0.336	0.077	-0.259	0.680	0.211	-0.469
L	0.273	0.072	-0.201	0.788	0.236	-0.552
M	0.416	0.083	-0.333	0.765	0.237	-0.527
N	0.260	0.072	-0.189	0.812	0.241	-0.571
O	0.341	0.079	-0.262	0.644	0.200	-0.444
P	0.357	0.078	-0.279	0.659	0.201	-0.457
Q	0.393	0.082	-0.311	0.793	0.230	-0.562
R	0.290	0.075	-0.215	0.700	0.217	-0.483
S	0.586	0.096	-0.490	0.667	0.203	-0.463
T	0.388	0.081	-0.308	0.727	0.222	-0.505

Table 4.11 summarizes the results table data and computes the threshold value that was used to evaluate the test results. A threshold score was also calculated for the original test data and difference from baseline values ( $\Delta$ ).

Table 4.11: Performance Measures – Case 1

Provider	CENTROIDAL		SYMBOLIC	
	Case 1	$\Delta$	Case 1	$\Delta$
Median	0.077	-0.242	0.220	-0.507
Q1	0.073	-0.308	0.210	-0.534
Q3	0.082	-0.198	0.236	-0.473
IQR	0.009	0.111	0.026	0.061
Threshold	0.096	-0.031	0.275	-0.382

Case 1 involved introducing a provider with an average change in miles billed. The centroidal method identified the change to Provider A. As shown above, Provider A, and S exceeded the calculated threshold for the test case data. Additionally, when compared to its difference from baseline, the centroidal approach appropriately flagged Provider A as exceeding the threshold and identified it as a potential outlier to investigate. When the difference was calculated compared to the baseline, Provider A was the only group to signal as an outlier.

Centroidal Provider A (Case 1) > Centroidal Threshold (Case 1) where,

$$.988 > .096$$

and

Centroidal Provider A ( $\Delta$ ) > Centroidal Threshold( $\Delta$ ) where,

$$.603 > -.031$$

Similarly, the symbolic method appropriately picked up the change in Provider A. Additionally, when compared to its difference from baseline, the symbolic approach appropriately flagged Provider A as exceeding the threshold and identified it as a potential outlier to investigate.

Symbolic Provider A (Case 1) > Symbolic Threshold (Case 1) where,

$$.965 > .275$$

and

Symbolic Provider A ( $\Delta$ ) > Symbolic Threshold( $\Delta$ ) where,

$$.206 > -.382$$

In both cases, when the difference is calculated compared to the baseline, Provider A was the only group to signal as an outlier. The centroidal method and the symbolic method performed equally well in this case.

In the previous example, mileage was increased approximately 400% to match the example stated in the literature. It is conceivable that changes in mileage would occur without such a pronounced change. Figure 4.11 depicts a 50% increase in mileage.

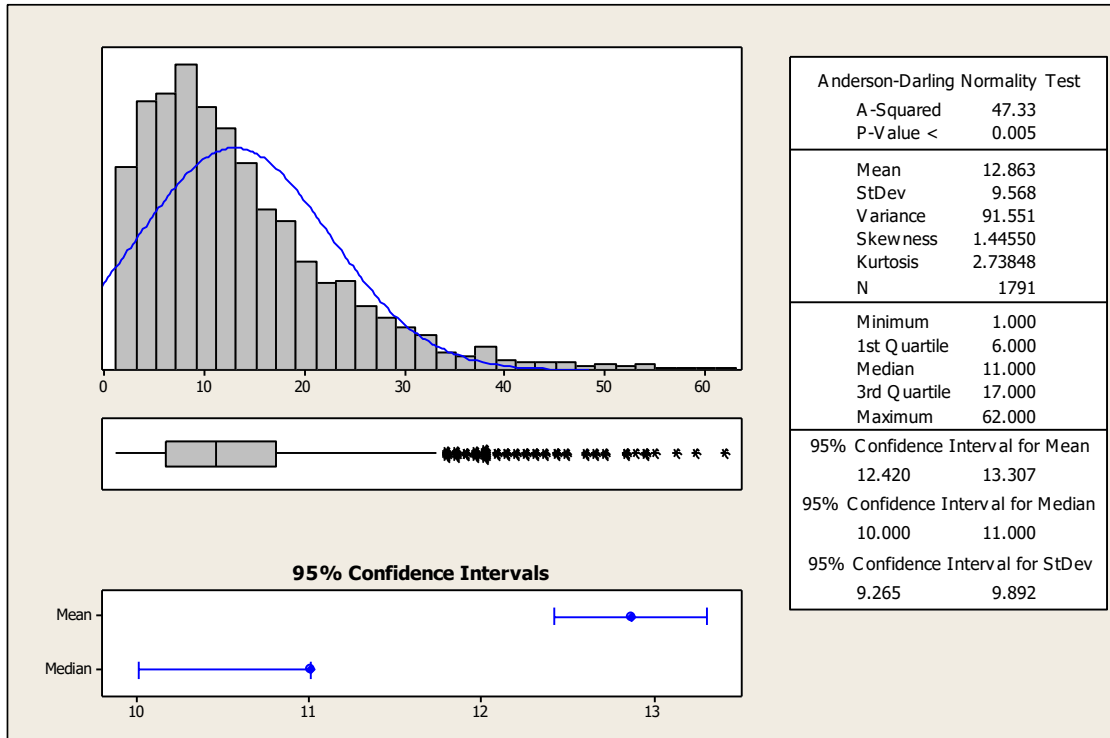


Figure 4.11: Excessive Mileage with 50% Increase

In the data presented above, the minimum mileage traveled is one mile, the maximum traveled is 62 miles, and the average distance traveled is 12.863. Provider A was the altered group with 1,791 mileage data points replaced. The new distribution used was a 3-parameter Weibull where shape = 1.26857, scale = 12.05822 and threshold = .92855. The values were then rounded to the nearest integer. The change in scale from the original distribution reflected a 50% increase in miles traveled. As before,

the centroidal and symbolic tests were applied to the data to determine if the anomalistic event could be identified when the difference in miles traveled was not as pronounced.

The results for the reduced mileage increase were identical to the previous example with both the centroidal and symbolic approaches identifying Provider A as an outlier. While the conclusion was the same, there were differences in the analysis. Specifically, the calculated difference between the anomalous provider, the baseline and the threshold were all reduced suggesting the less exaggerated the difference, the less noticeable the outlier. Table 4.12 and Table 4.13 show the results difference between the 400% and the 50% change in mileage example.

Table 4.12: Provider A Statistics with a 400% Increase in Miles

<b>Metric</b>	<b>Centroidal</b>	<b>Symbolic</b>
<i>Actual Distance Value</i>	.988	.965
<i>Baseline Value</i>	.385	.759
<i>Difference</i>	.603	.206
<i>Difference Threshold</i>	-.031	-.382

Table 4.13: Provider A Statistics with a 50% Increase in Miles

<b>Metric</b>	<b>Centroidal</b>	<b>Symbolic</b>
<i>Actual Distance Value</i>	.937	.901
<i>Baseline Value</i>	.385	.759
<i>Difference</i>	.552	.142
<i>Difference Threshold</i>	-.009	-.171



#### 4.4 Case 2 – Inappropriate Transport Levels

Another instance of questionable billing in the OIG Report focused on suppliers with higher percentages of more expensive transport levels [75]. The second hypothesis tested was to determine if a change in service level distribution could be detected. In the standard provider profile, Basic Life Support Services (A0429) accounted for approximately 8.4% of all transactions. Advanced Life Support – Level 1 (A0427) accounted for approximately 40.7% of all transactions. Figure 4.12 is the distribution of services present in the baseline dataset for Provider A.

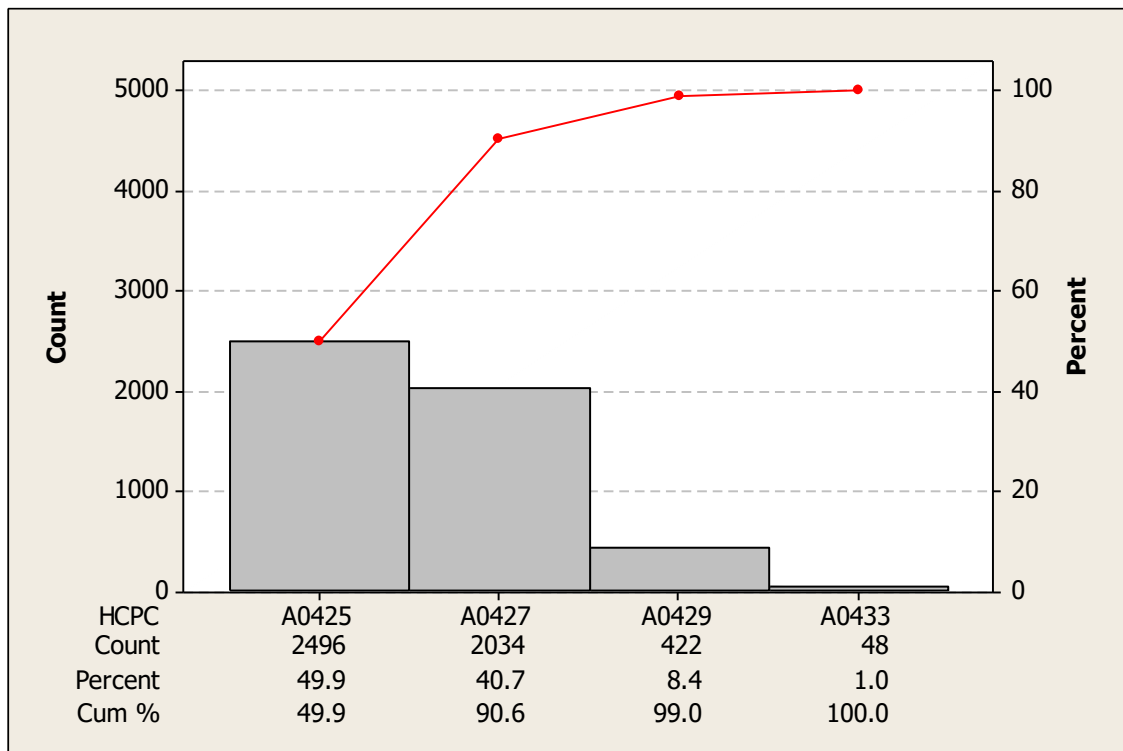


Figure 4.12: Pareto Chart of Service Codes

One of the fraudulent activities highlighted in the literature involved the upcoding from Basic Life Support to the more expensive Advanced Life Support level. Provider A was altered to increase the percentage of Advanced Life Support – Level 1 cases simulating a potential upcode of services. The data was modified by randomly deleting half of the A0429 cases and randomly increasing the A0427 cases by the same amount. Records were modified by row which meant all the accompanying characteristics also changed with a change in service. Figure 4.13 is the distribution of services for the altered dataset.

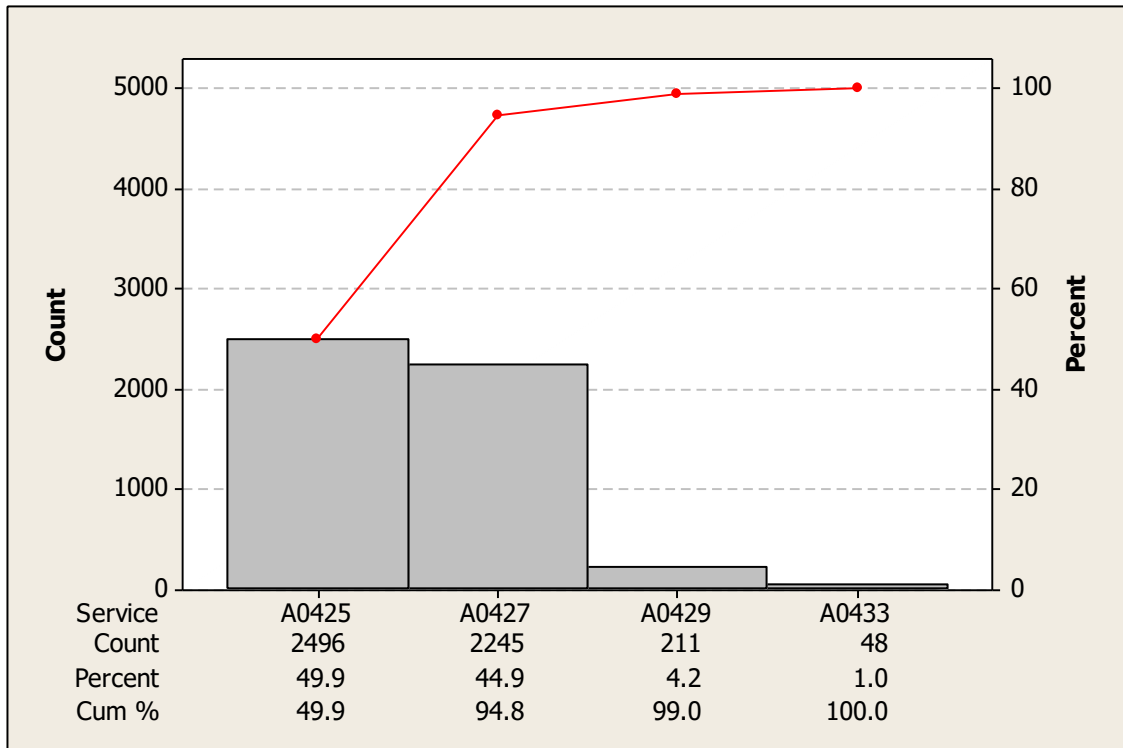


Figure 4.13: Pareto Chart of Service Codes - Modified

The altered dataset reduced the amount of the less expensive A0429 Basic Life Support cases and increased the number of the more expensive A0427 Advanced Life Support cases. Table 4.14 and Table 4.15 are the results table and performance measure table for Case 2: Inappropriate Transport Levels.

Table 4.14: Results Table – Case 2

Provider	CENTROIDAL			SYMBOLIC		
	Baseline	Case 2	$\Delta$	Baseline	Case 2	$\Delta$
A	0.385	0.485	0.100	0.759	0.873	0.114
B	0.301	0.306	0.005	0.805	0.501	-0.304
C	0.271	0.277	0.006	0.691	0.439	-0.252
D	0.256	0.262	0.006	0.756	0.477	-0.279
E	0.269	0.275	0.006	0.713	0.454	-0.259
F	0.267	0.272	0.006	0.743	0.467	-0.276
G	0.309	0.315	0.006	0.681	0.429	-0.252
H	0.327	0.333	0.006	0.718	0.454	-0.263
I	0.424	0.427	0.003	0.796	0.505	-0.290
J	0.505	0.503	-0.001	0.730	0.461	-0.268
K	0.336	0.341	0.005	0.680	0.430	-0.250
L	0.273	0.279	0.006	0.788	0.496	-0.292
M	0.416	0.422	0.006	0.765	0.480	-0.285
N	0.260	0.266	0.006	0.812	0.506	-0.306
O	0.341	0.346	0.005	0.644	0.412	-0.232
P	0.357	0.363	0.006	0.659	0.421	-0.237
Q	0.393	0.398	0.006	0.793	0.493	-0.299
R	0.290	0.296	0.005	0.700	0.440	-0.260
S	0.586	0.592	0.006	0.667	0.424	-0.243
T	0.388	0.394	0.006	0.727	0.462	-0.265

Table 4.15: Performance Measures – Case 2

Provider	CENTROIDAL		SYMBOLIC	
	Case 1	$\Delta$	Case 1	$\Delta$
Median	0.337	0.006	0.462	-0.264
Q1	0.278	0.005	0.436	-0.286
Q3	0.404	0.006	0.494	-0.251
IQR	0.126	0.001	0.057	0.035
Threshold	0.593	0.007	0.580	-0.199

Case 2 involved introducing a provider who had engaged in billing for inappropriate transport levels. When the mix of these service levels is different than the mix of services from similar providers, it could mean a fraudulent event has occurred. A more complex level of service also means a more expensive service. The centroidal method failed to pick up the change in the straight Provider A test data but did correctly make the identification when compared to the baseline.

Centroidal Provider A ( $\Delta$ ) > Centroidal Threshold( $\Delta$ ) where,

$$.100 > .007$$

Similarly, the symbolic method appropriately picked up the change in Provider A. When the difference is calculated compared to the baseline, Provider A is the only group to signal as an outlier. Specifically,

Symbolic Provider A (Case 2) > Symbolic Threshold (Case 2) where,

$$.873 > .580$$

and

Symbolic Provider A ( $\Delta$ ) > Symbolic Threshold( $\Delta$ ) where,

$$.114 > -.199$$

In the previous example, half of the less expensive transport service was moved to the more expensive, Advanced Life Support (A0427), service. Instead of moving 50% of that volume, the next experiment tested the result if approximately 10% (40 cases) were upcoded. Figure 4.14 is a graph of the new distribution for Provider A with after this change was made.

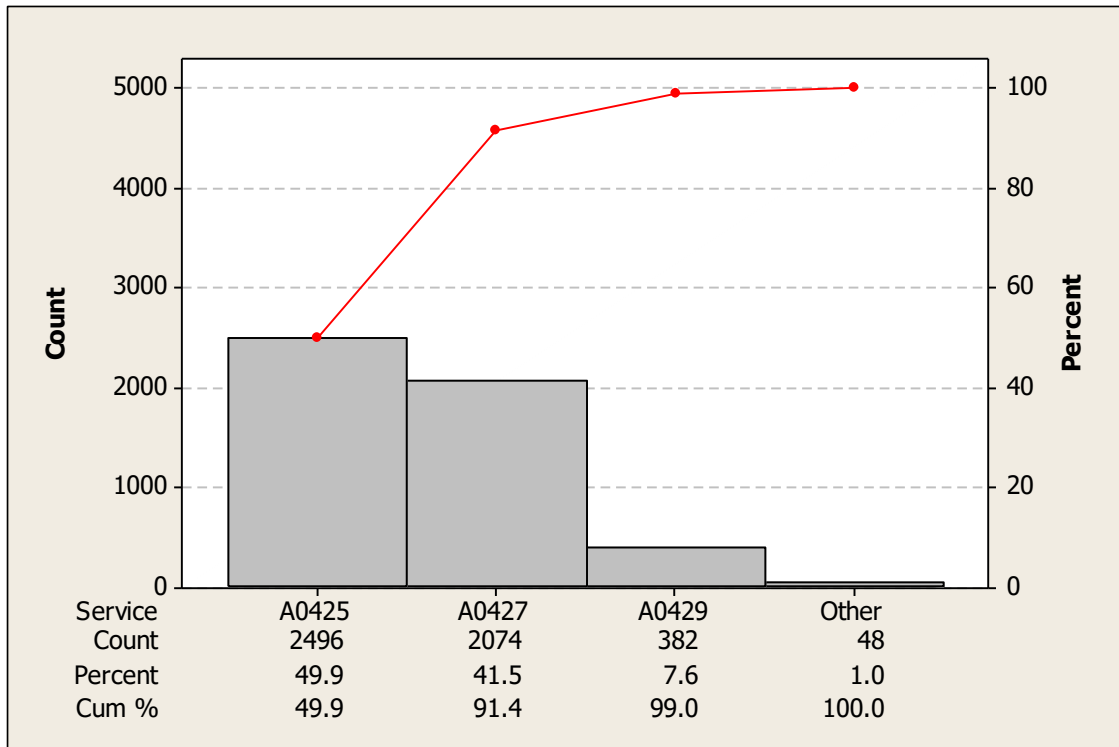


Figure 4.14: Ten Percent of Volume Upcoded

The results for a less significant upcoding action matched the previous example when compared to the difference from the baseline. As in Case 1, while the conclusion was the same, there were differences in the analysis. The calculated difference between the anomalous provider, the baseline, and the threshold were all reduced suggesting the less exaggerated the difference, the less noticeable the outlier. Table 4.16 and Table 4.17 show the results difference between the 50% upcode case and the 10% upcode case.

Table 4.16: Provider A Statistics with a 50% Upcode Action

<b>Metric</b>	<b>Centroidal</b>	<b>Symbolic</b>
<i>Actual Distance Value</i>	.485	.873
<i>Baseline Value</i>	.385	.759
<i>Difference</i>	.100	.114
<i>Threshold</i>	.007	-.199

Table 4.17: Provider A statistics with a 10% Upcode Action

<b>Metric</b>	<b>Centroidal</b>	<b>Symbolic</b>
<i>Actual Distance Value</i>	.398	.791
<i>Baseline Value</i>	.385	.759
<i>Difference</i>	.013	.032
<i>Threshold</i>	.001	-.012

Since this data was scaled, meaningful differences can be compared using the two examples above. When a 50% upcode is present, Provider A exceeds the threshold value by .093 when using the centroidal approach while exceeding the threshold value by .313 using the symbolic method suggesting the difference is more noticeable and measurable using the latter. When the 10% upcode scenario is assessed, the centroidal approach yields a .012 difference while the symbolic method yields a .044 difference from threshold. Both methods showed a tightening toward the difference threshold, but the symbolic method appeared to be more sensitive to the change when comparing the difference to threshold to the baseline value, calculated as



$$\begin{aligned} & \text{Difference from Threshold (Centroidal) / Baseline Value (Centroidal)} \\ & = .012 / .385 = .031 \end{aligned}$$

and

$$\begin{aligned} & \text{Difference from Threshold (Symbolic) / Baseline Value (Symbolic)} \\ & = .044 / .759 = .058 \end{aligned}$$

where,

$$.058 > .031$$

#### **4.5 Case 3 – Excessive Number of Transports per Beneficiary**

A third measure established by [75] addressed the high number of transports per beneficiary. The report looked specifically at transport issues associated with dialysis patients and found suppliers that may have billed for services that were medically unnecessary or for services that did not occur at all. The study found that most dialysis patients received an average of 4 transports during the six-month study period while the suspect suppliers provided 21 transports per beneficiary during that same time period. Three percent of suppliers in the 2012 study had questionable billing patterns relating to this metric totaling \$132.5 million in the first half of that year [75].

The dataset in this dissertation does not look at dialysis patients specifically but can be designed to evaluate the average number of transports per beneficiary and to check for anomalies. While patient data is not directly included in the test dataset, fictitious patient

identifiers were randomly generated and added to the data. The data was filtered to include only the A0425 (mileage) code entries to ensure that each row of data represented one unique trip event. Random patient identifiers were generated and applied to the dataset, with replacement, which allowed for multiple trips to occur within each provider and across providers. The resulting test dataset was a 36,537-row collection of patients seeking transport services across 20 providers. Provider A, for example, had 1,791 trip events utilized by 1,755 beneficiaries resulting in an average of 1.02 trips per beneficiary. With a new variable introduced, a new baseline case was established. Patient ID was included, and Service was excluded (because all were represented by the A0425 mileage code).

To simulate an increased level of transports per beneficiary, Provider A patient data was studied. The data was sorted by patient age, diagnosis code, and miles traveled. A group of 10 similar entries was discovered that each had unique patient identifiers. These 10 cases were assigned the same identifier, simulating one patient within one provider that used the same service 10 times as opposed to the group average of 1.02 transports per beneficiary. A snapshot of the anomalous data within Provider A, identified as patient P8388 appears in Figure 4.15.

1	Age	AutoAcc	Patient	Service	ICD	Origin	Sex	Charges	Units	Provider
1653	91.92	N	P27052	A0425	7197	R	M	21.27	3.00	A
1654	91.98	N	P12824	A0425	78605	R	M	56.72	8.00	A
1655	92.15	N	P8388	A0425	71945	R	F	99.26	14.00	A
1656	92.15	N	P8388	A0425	71945	R	F	99.26	14.00	A
1657	92.15	N	P8388	A0425	71945	R	F	99.26	14.00	A
1658	92.15	N	P8388	A0425	71945	R	F	99.26	14.00	A
1659	92.15	N	P8388	A0425	71945	R	F	99.26	14.00	A
1660	92.15	N	P8388	A0425	71945	R	F	99.26	14.00	A
1661	92.15	N	P8388	A0425	71945	R	F	99.26	14.00	A
1662	92.15	N	P8388	A0425	71945	R	F	99.26	14.00	A
1663	92.15	N	P8388	A0425	71945	R	F	99.26	14.00	A
1664	92.15	N	P8388	A0425	71945	R	F	99.26	14.00	A
1665	92.16	N	P16946	A0425	7850	R	F	21.27	3.00	A
1666	92.16	N	P20819	A0425	7850	R	F	21.27	3.00	A
1667	92.16	N	P27670	A0425	7850	R	F	21.27	3.00	A

Figure 4.15: Screenshot of Simulated Patient with Multiple Trips

Table 4.18 and Table 4.19 are the results table and performance measure table for Case 3: High Number of Transports per Beneficiary. It includes revised baseline numbers, which were also generated.

Table 4.18: Results Table – Case 3

Provider	CENTROIDAL			SYMBOLIC		
	Baseline	Case 3	$\Delta$	Baseline	Case 3	$\Delta$
A	0.337	0.337	3.81E-05	0.843	0.844	4.55E-04
B	0.280	0.280	1.59E-06	0.797	0.796	-9.01E-04
C	0.275	0.275	2.14E-06	0.843	0.842	-9.71E-04
D	0.282	0.282	1.66E-06	0.790	0.789	-9.06E-04
E	0.283	0.283	4.45E-06	0.841	0.840	-9.74E-04
F	0.290	0.290	1.56E-06	0.866	0.865	-9.94E-04
G	0.267	0.267	2.20E-06	0.803	0.802	-9.26E-04
H	0.538	0.538	7.58E-07	0.804	0.803	-9.23E-04
I	0.529	0.529	2.10E-06	0.853	0.852	-9.96E-04
J	0.594	0.594	1.64E-06	0.824	0.823	-9.54E-04
K	0.345	0.345	2.77E-06	0.800	0.799	-9.21E-04
L	0.311	0.311	4.93E-06	0.789	0.788	-9.11E-04
M	0.310	0.310	1.92E-06	0.856	0.855	-1.01E-03
N	0.305	0.305	1.26E-06	0.863	0.862	-1.02E-03
O	0.355	0.355	2.44E-06	0.811	0.810	-9.36E-04
P	0.339	0.339	1.09E-06	0.769	0.768	-8.83E-04
Q	0.362	0.362	1.04E-06	0.776	0.775	-8.91E-04
R	0.256	0.256	1.65E-06	0.784	0.783	-9.31E-04
S	0.254	0.254	1.81E-06	0.767	0.766	-8.76E-04
T	0.389	0.389	1.12E-06	0.808	0.807	-9.23E-04

Table 4.19: Performance Measures – Case 3

Provider	CENTROIDAL		SYMBOLIC	
	Case 3	$\Delta$	Case 3	$\Delta$
Median	0.311	1.73E-06	0.805	-9.24E-04
Q1	0.282	1.49E-06	0.789	-9.72E-04
Q3	0.357	2.26E-06	0.843	-9.05E-04
IQR	0.075	7.73E-07	0.054	6.72E-05
Threshold	0.469	3.42E-06	0.923	-8.04E-04

Case 3 was a test to determine if a beneficiary that exceeded the average trips per provider could be identified using the symbolic approach. A new baseline was established, and the data was assessed using the centroidal approach and the symbolic approach. When the difference to the baseline was observed, the centroidal approach appropriately identified Provider A but also inappropriately flagged two other groups. The symbolic method correctly identified Provider A as being the sole group with an anomaly. The difference was calculated as,

Symbolic Provider A ( $\Delta$ ) > Symbolic Threshold( $\Delta$ ) where,

$$4.55E -04 > -8.04E -04$$

#### 4.6 Case 4 – Incorrect Rate Identification

The final case evaluated was not inspired by a specific example but was a recognition that general administrative errors in billing occur. Intentional or not, these

mistakes contribute to the waste experienced by today's healthcare system. They can be the result of contract misinterpretation, data entry errors, or the result of system errors. This final case used the ambulance transport data to simulate an anomaly of this type.

The hypothesis tested in this experiment was to determine if the symbolic approach could detect an error that was the result of an inappropriate rate. In the baseline case, code A0425 is used for transport mileage. As previously discussed, the Units variable represents the actual mileage billed. Units are multiplied by a flat mileage rate in order to arrive at the final charges. In the base dataset, \$7.09 is the standard mileage rate for all A0425 lines. As in the previous cases, a baseline was generated in order to have a means of comparison. The continuous variable Charges was included as it is the only variable that reflects a pricing error as described.

Pricing error anomalies were added to Provider A to simulate a change. In the base dataset, Provider A had 172 transport trips that were recorded as eight miles in distance. Each of these events were billed at \$7.09, totaling \$56.72 for each line and \$9,755.84 for all lines. The anomalistic pattern was introduced to Provider A by changing the rate billed. Half of the 172 lines were modified to reflect a rate of \$6.09 and a line total of \$48.72. Half were modified to reflect a rate of \$8.09 and a line total of \$64.72. The total for all 172 lines remained \$9,755.84.

The symbolic and centroidal tests were run to determine if the change could be identified. Table 4.20 and Table 4.21 are the results table and performance measure table for Case 4: Pricing Errors.

Table 4.20: Results Table – Case 4

Provider	CENTROIDAL			SYMBOLIC		
	Baseline	Case 4	$\Delta$	Baseline	Case 4	$\Delta$
A	0.221	0.221	1.90E-14	0.727	0.727	0.000
B	0.623	0.623	5.70E-14	0.813	0.813	0.000
C	0.361	0.361	-2.30E-14	0.706	0.706	0.000
D	0.226	0.226	2.10E-14	0.741	0.741	0.000
E	0.219	0.219	3.00E-14	0.716	0.716	0.000
F	0.221	0.221	3.30E-14	0.765	0.765	0.000
G	0.218	0.218	1.80E-14	0.661	0.661	0.000
H	0.350	0.350	2.50E-14	0.742	0.742	0.000
I	0.274	0.274	8.00E-14	0.775	0.775	0.000
J	0.244	0.244	3.50E-14	0.716	0.716	0.000
K	0.217	0.217	2.20E-14	0.673	0.673	0.000
L	0.431	0.431	-5.60E-14	0.767	0.767	0.000
M	0.254	0.254	4.20E-14	0.741	0.741	0.000
N	0.334	0.334	4.60E-14	0.785	0.785	0.000
O	0.221	0.221	1.10E-14	0.657	0.657	0.000
P	0.260	0.260	3.10E-14	0.654	0.654	0.000
Q	0.365	0.365	5.10E-14	0.769	0.769	0.000
R	0.555	0.555	5.91E-14	0.700	0.700	0.000
S	0.234	0.234	7.99E-15	0.658	0.658	0.000
T	0.351	0.351	2.20E-14	0.721	0.721	0.000

Table 4.21: Performance Measures – Case 4

Provider	CENTROIDAL		SYMBOLIC	
	Case 4	$\Delta$	Case 4	$\Delta$
Median	0.257	2.75E-14	0.724	0.000
Q1	0.221	1.87E-14	0.693	0.000
Q3	0.353	4.30E-14	0.765	0.000
IQR	0.132	2.43E-14	0.072	0.000
Threshold	0.552	7.95E-14	0.873	0.000

The initial results for Case 4 were not as expected. Although the differences from the base case were negligible, the centroidal approach incorrectly flagged Provider I as an outlier provider. The symbolic approach failed to find any difference between the base case and anomalistic case and did not flag any provider as anomalous. Differences from the base case distances in the symbolic case were all zero.

#### 4.7 Case 4R - Incorrect Rate Identification Revised

Upon further study, the results from the symbolic run of Case 4 data provided information that prompted another iteration of testing. A zero result across all distances implied that the anomalistic case was being evaluated as equal to the base case even when it is known that the datasets are different. Provider A was modified to contain 172 individual events of charge data that were different.

While the treatment of modal variables is similar in both the centroidal and symbolic approach, continuous variables are evaluated differently. The centroidal method evaluates groups of data using the mean of the group while the symbolic approach



evaluates groups by looking at their distributions. Histograms of each variable within a group are developed and then compared “symbolically” to each other in order to evaluate the degree of difference, if any. In the original analysis of Case 4, the symbolic data results suggested that the comparison of histograms revealed no difference across the 20 groups, yet Provider A should have appeared as different from the rest. Upon further review, the histogram for Provider A in the baseline case was not different than the anomalous dataset. The simulated pricing errors were masked by the calculated width of the bins when preparing the histograms. The changes that occurred between \$6.09 and \$8.09 were all captured within one bin and therefore were not identified. In the previous chapters, the Sturges binning method was chosen as the optimal approach and has been applied consistently in this dissertation. In this final case, when trying to detect a \$2.00 discrepancy (\$8.09 - \$6.09) in 172 rows of a 100,000-row dataset where the total dollars charged ranged from \$6.09 to \$375.77, the Sturges binning method was not sensitive enough. The Scott method was another binning technique evaluated in Chapter 3. It generates nearly four times the number of bins as does the Sturges method. Case 4 was tested again using the Scott method as the binning technique. Table 4.22 and Table 4.23 are the results table and performance measure table for Case 4: Pricing Error - Revised.

Table 4.22: Results Table – Case 4R Revised

Provider	CENTROIDAL			SYMBOLIC		
	Baseline	Case 4R	$\Delta$	Baseline	Case 4R	$\Delta$
A	0.221	0.221	1.90E-14	0.748	0.819	0.071
B	0.623	0.623	5.70E-14	0.822	0.825	0.003
C	0.361	0.361	-2.30E-14	0.730	0.730	0.000
D	0.226	0.226	2.10E-14	0.721	0.726	0.005
E	0.219	0.219	3.00E-14	0.743	0.747	0.004
F	0.221	0.221	3.30E-14	0.756	0.761	0.005
G	0.218	0.218	1.80E-14	0.684	0.688	0.003
H	0.350	0.350	2.50E-14	0.724	0.729	0.005
I	0.274	0.274	8.00E-14	0.781	0.785	0.004
J	0.244	0.244	3.50E-14	0.684	0.687	0.004
K	0.217	0.217	2.20E-14	0.718	0.722	0.005
L	0.431	0.431	-5.60E-14	0.783	0.786	0.003
M	0.254	0.254	4.20E-14	0.731	0.733	0.003
N	0.334	0.334	4.60E-14	0.781	0.784	0.002
O	0.221	0.221	1.10E-14	0.652	0.657	0.004
P	0.260	0.260	3.10E-14	0.697	0.699	0.002
Q	0.365	0.365	5.10E-14	0.765	0.766	0.001
R	0.555	0.555	5.91E-14	0.718	0.723	0.005
S	0.234	0.234	7.99E-15	0.677	0.683	0.006
T	0.351	0.351	2.20E-14	0.737	0.742	0.005

Following is a table that summarizes the performance metrics for the revised test.

Table 4.23: Performance Measures – Case 4R Revised

Provider	CENTROIDAL		SYMBOLIC	
	Case 4R	$\Delta$	Case 4R	$\Delta$
Median	0.257	2.75E-14	0.732	0.004
Q1	0.221	1.87E-14	0.717	0.003
Q3	0.353	4.30E-14	0.771	0.005
IQR	0.132	2.43E-14	0.054	0.002
Threshold	0.552	7.95E-14	0.852	0.008

Employing the Scott binning approach did not change the centroidal calculations or Provider *I* being incorrectly flagged as an outlier. The increased number of bins did provide the resolution needed for the symbolic method to properly identify Provider A as the anomalistic event. Provider A was recognized through the difference calculation where,

Symbolic Provider A ( $\Delta$ ) > Symbolic Threshold( $\Delta$ ) where,

$$.071 > .008$$

This final example highlights the need for additional research in to determining the appropriate binning technique. As mentioned in Chapter 3, the Sturges method is a time-tested binning technique that has been documented in textbooks and coded in to statistical

software applications for many years. While its utility has been proven over time and its effectiveness demonstrated through the examples in this dissertation, it may not always be the best selection. When datasets have a large number of observations or when the difference attempting to be discerned is small, other approaches may work better. Table 4.24 shows the binning parameters for all five studied tests for Case 4R – Incorrect Rate Identification Revised. The variable studied in the case is the dollars charged variable. The variable has 100,000 entries across 20 providers. The minimum charge across all service codes, including the mileage code, was \$6.09. The maximum charge was \$643.91. The difference attempting to be discovered was \$2.00.

Table 4.24: Case 4R Binning Results

	<b>Sqr Root</b>	<b>Sturges</b>	<b>Scott</b>	<b>Rice</b>	<b>Freedman</b>
Number of Bins	316	18	46	93	40
Bin Width	2.02	35.40	13.80	6.87	15.80

A second iteration of Case 4R was run using the Sqr. Root binning approach. Like the Scott method, it identified the anomaly as well. As shown in Table 4.24, when attempting to identify a small change, bin widths greater than the change can mask the difference. Even the Scott method could have missed the change as its bin width of \$13.80 far exceeds the \$2.00 discrepancy.

Choosing the right binning approach is critical to the success of the methodology. The Sturges method can be appropriately applied in most applications and provides the smoothing necessary to properly process inherent noise in data. If the suspected difference

in the data is smaller than the bin width provided by Sturges, then a more discerning binning approach may be needed. The proper binning approach is one that is sensitive enough to identify anomalies in the data but not overly sensitive as to identify common cause variation and call it special cause.

#### **4.8 Chapter Summary and Observations**

The purpose of evaluating this dataset was to test different situations that could exist in a real-world scenario. The available fields and their distributions were modeled after a real provider. A baseline dataset was created to mimic a set of providers that were all similar with only inherent randomness present. Anomalous situations were introduced to represent aberrant behavior that could reasonably be expected to occur with this type of data. These were incidents of excessive mileage being billed, inappropriate service levels being applied, a higher than expected number of transports per beneficiary observed, and a change in the rate or price of a basic service.

The baseline dataset was constructed in this chapter to simulate a steady-state operating scenario. When evaluating real-world situations, common cause variability is almost always present. The baseline case was modeled after a real-world provider but randomized in a way to create 20 similar, but not identical, groups. In some cases, particularly when applying the centroidal approach, this randomization technique produced groups that looked different than the others. Therefore, the more important measure for real-world application was to look for changes to the baseline steady-state case. For each of the cases presented in this chapter, a baseline case was run in addition to the anomalistic case. This provided an opportunity to calculate the outlier score metric for the differences

observed between behavior during steady-state versus behavior when an anomalistic event has occurred.

Case 1 introduced the anomaly of providers billing excessive mileage. It focused on one specific route in the data involving trips from a residence to a hospital. The real data that existed most closely matched the Weibull distribution, therefore, that distribution was used to model the aberrant data. The result was an anomalistic event being placed in Provider A where the average miles for these cases was nearly four times greater and the variability of the data greatly increased. Both tests performed well in this case. The centroidal approach ably identified the change in mean. The symbolic method accurately identified this change as well by clearly capturing and measuring the change in mean and variation through the mapping and comparison of distributions.

In the case of a change in transport levels, both models performed well. In the example, a change in service meant a change in all the other variables being used as input to the model. That meant a modification to the categorical variables of AutoAcc, ICD, Origin and Sex as well as changes to the continuous variables of Age and Units. Both models picked up the differences assigned to Provider A when compared to the baseline and highlighted it as an outlier.

Case 3 presented a challenge for the centroidal model. The differences for each model were minimal but present. The only difference between the anomalistic data and the baseline was the conversion of 10 cells of unique patient identifiers to 10 cells of identical patient identifiers. In a data model that was evaluating five input variables (Age, PatientID, ICD, Sex, Units) there were 500,000 active cells being evaluated (100,000 x 5). Only nine cells were modified to match the tenth which resulted in 10 identical records

representing the same patient. The test was to determine if the models could find such a slight difference. The centroidal approach discovered the difference in Provider A but falsely identified changes in two other providers. The symbolic method correctly identified only the change to Provider A.

The final case, Case 4, proved the most interesting. The variable Charges was introduced back to the model and changes to a rate within the model impacted the final charges billed by Provider A. The total and average charges did not change but the distribution of the continuous variable Charges did change. In the first iteration of the experiment, the centroidal approach erroneously identified the wrong provider as being an anomaly. The symbolic approach failed to find any difference from the baseline. A second run of the test was done, and the Sturges binning method was replaced with the more discriminating binning approaches (Scott and Sqr. Root). In this run, the centroidal result was unchanged, but the symbolic method appropriately identified Provider A as anomalous. While the average of charges across groups had not changed, the distribution had and when there were enough bins in the histograms that depicted the range of charges, the symbolic method identified the anomalous group as different.

Overall, the symbolic model performed well when assessed against these four instances. In a sample of 100,000 rows of data it did well in each case, correctly identifying the presence of a signal indicating the presence of an anomaly. Table 4.25 summarizes the results of the testing done in this chapter.

Table 4.25: Case Summary Table

#	Case	Centroidal Correct	Symbolic Correct
1	Excessive Mileage	YES	YES
2	Inappropriate Transport Levels	YES	YES
3	Excessive Number of Transports per Beneficiary	NO	YES
4	Incorrect Rate Identification	NO	NO
4R	Incorrect Rate Identification Revised	NO	YES



## CHAPTER 5

### CONCLUSIONS AND PROPOSED FUTURE RESEARCH

The purpose of this dissertation was to examine an alternative approach to identifying fraud, waste, and abuse events in healthcare insurance claims datasets. Traditionally, these events are examined at an individual level in search of an outlier or some anomalous occurrence. This research investigated alternatives to looking at the data at a higher “concept” level. The study of SDA provided the foundation for the methods developed in this dissertation. SDA allows for the analysis of concepts while preserving the underlying characteristics of the individual data points. This research examined two approaches, centroidal and symbolic, and compared their effectiveness in identifying anomalous conditions. Continuous and categorical data were tested, and a performance metric was developed in order to compare the two methods. During the research, code was written in R to validate the calculations and to provide a tool that could handle larger datasets. After considering some randomly generated datasets, comparisons of the centroidal and symbolic approaches were performed on larger “real-world” inspired datasets from the healthcare insurance claims industry to determine how well each approach could identify anomalous conditions that could signal the presence of FWA.

#### 5.1 Conclusions

Inspired by the contributions of researchers in the field of SDA, and driven by the conviction that “distributions are the numbers of the future” [1], a new approach to anomaly detection was investigated and several contributions to the body of knowledge in this field were made. It was discovered that symbolic anomaly detection techniques perform equally

as well as their classical centroidal counterparts when only changes in mean or central tendency distinguish anomalistic activity from normal steady-state. When changes are more subtle, particularly when means are equal but the underlying shape of the distribution changes, the symbolic approach excels. The improved resolution alone when compared to centroidal methods makes this approach particularly suited to discovering unusual patterns in the data. Additionally, the benefit of being able to add information at a “concept” level allowed for greater definition and resolution of concepts which helped provide greater distance between an anomaly and the steady-state common cause condition.

A second contribution to this work was the study of multiple histogram binning techniques. While the Sturges binning method remains the most popular technique in histogram construction, and the one adopted for much of this research, there are other techniques available. This seems particularly relevant for this research as the ability to detect changes in distributions can be significantly impacted by histogram construction. In the final example in this research, an alternative histogram approach (Scott Method) demonstrated the value of exploring alternate techniques to driving the resolution required to identify anomalies where they exist.

A third contribution was the customized application of a well-established non-parametric test to several dissimilarity matrices in order to determine a threshold at which to call an event “anomalistic.” This test was applied to classical centroidal tests as well as symbolic tests in order to compare performance.

A final contribution was the development of R code which allowed for testing of complex datasets and application to real-world scenarios, specifically those which involve

potential FWA in healthcare insurance claims datasets. This has been made available online in the public domain to further use and research.

The symbolic approach to anomaly detection appears to have many qualities which make it an excellent candidate for more widespread acceptance across multiple disciplines and industries. As introduced in this dissertation, the robustness of the approach and its ability to handle multiple data types qualify it as a method which should be considered for all types of anomaly detection problems.

## **5.2 Proposed Future Research**

The research introduced in this dissertation can serve as a foundation for additional work using symbolic data to identify anomalistic behavior and more specifically, its application to the identification of FWA. Specific topics that are candidates for additional research include:

- Adjusting the non-parametric score multiplier to optimize Type I and Type II errors in practice.
- Further exploring the impact that multicollinearity could have on the symbolic approach.
- Developing a standard technique to choose the appropriate binning approach for histogram construction depending on the type, volume, and industry specific nature of the data.
- Exploring the effects of alternative distance calculation techniques (including Euclidean which is used herein) and their effects on anomaly detection.
- Investigating the relationship between supervised and unsupervised models by using the symbolic approach in conjunction with labeled datasets and supervised learning techniques.
- Addressing dimensionality reduction as it relates to symbolic data anomaly detection techniques.

- Developing a practical guide to determine “steady-state” as it relates to establishing a baseline to which future activity is compared and to which “anomalous events” are measured against.
- Experimenting with changing variable weights and their contribution to symbolic anomaly detection (all variables considered of equal importance in this dissertation).

## REFERENCES

- [1] B. Schweizer, “Distributions are the numbers of the future.,” in *Proceedings The Mathematics of Fuzzy Systems Meeting University of Naples*, 1984.
- [2] “NHE Fact Sheet | CMS.” <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet> (accessed Mar. 01, 2020).
- [3] A. D. Hackbarth, “Eliminating Waste in US Health Care,” *JAMA*, vol. 307, no. 14, p. 1513, Apr. 2012, doi: 10.1001/jama.2012.362.
- [4] “The \$272 billion swindle | The Economist.” <http://www.economist.com/news/united-states/21603078-why-thieves-love-americas-health-care-system-272-billion-swindle> (accessed Sep. 27, 2016).
- [5] “The Challenge of Health Care Fraud - The NHCAA.” <https://www.nhcaa.org/resources/health-care-anti-fraud-resources/the-challenge-of-health-care-fraud/> (accessed Mar. 01, 2020).
- [6] E. Diday, M. Noirhomme-Fraiture, and Wiley InterScience (Online service), *Symbolic data analysis and the SODAS software*. Chichester, England; Hoboken, NJ: J. Wiley & Sons, 2008.
- [7] Institute of Medicine (US) Roundtable on Evidence-Based Medicine, *The Healthcare Imperative: Lowering Costs and Improving Outcomes: Workshop Series Summary*. Washington (DC): National Academies Press (US), 2010.
- [8] “Wasted Healthcare Spending: A \$750 Billion Opportunity,” *hcldr*, Jul. 15, 2018. <https://hcldr.wordpress.com/2018/07/15/wasted-healthcare-spending-a-750-billion-opportunity/> (accessed Sep. 08, 2019).
- [9] “CMS National Training Program - Centers for Medicare & Medicaid Services.” <http://cms.gov/Outreach-and-Education/Training/CMSNationalTrainingProgram/index.html> (accessed Sep. 02, 2014).
- [10] “Home - The NHCAA.” <http://www.nhcaa.org/> (accessed Apr. 02, 2013).
- [11] L. Morris, “Combating Fraud In Health Care: An Essential Component Of Any Cost Containment Strategy,” *Health Affairs*, vol. 28, no. 5, pp. 1351–6, Oct. 2009.
- [12] C. for Medicare, M. S. 7500 S. B. Baltimore, and M. Usa, “Medicare Fraud and Abuse: Prevention, Detection, and Reporting.” Nov. 18, 2014, Accessed: Feb. 07, 2015. [Online]. Available: <http://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/MLN-Publications-Items/CMS1243333.html>.

- [13] “Comparison of the Anti-Kickback Statute and Stark Law.” United States. Health Care Fraud Prevention and Enforcement Action Team, Office of Inspector General, Accessed: Feb. 07, 2015. [Online]. Available: <https://oig.hhs.gov/compliance/provider-compliance-training/files/StarkandAKSChartHandout508.pdf>.
- [14] ASPA, “HEAT Task Force.” <http://www.stopmedicarefraud.gov/aboutfraud/heattaskforce/index.html> (accessed Feb. 07, 2015).
- [15] P. Budetti, “Command Center Speeds Up Anti-Fraud Efforts,” *The CMS Blog*, Jul. 31, 2012. <http://blog.cms.gov/2012/07/31/command-center-speeds-up-anti-fraud-efforts/> (accessed Feb. 07, 2015).
- [16] “Center for Program Integrity.” [http://www.ahcancal.org/facility\\_operations/integrity/Pages/Center-for-Program-Integrity.aspx](http://www.ahcancal.org/facility_operations/integrity/Pages/Center-for-Program-Integrity.aspx) (accessed Feb. 07, 2015).
- [17] Centers for Medicare and Medicaid Services, “Report to Congress, Fraud Prevention System, Second Implementation Year.” <http://www.stopmedicarefraud.gov/fraud-rtc06242014.pdf> (accessed Feb. 02, 2015).
- [18] S. Agrawal, *Preventing Medicare Fraud: How Can We Best Protect Seniors and Taxpayers?* United States Senate Special Committee on Aging, Washington D.C., 2014.
- [19] “CMS selects Northrop Grumman to implement fraud prevention system | Healthcare Finance News.” <http://www.healthcarefinancenews.com/news/cms-selects-northop-grumman-implement-fraud-prevention-system> (accessed Feb. 07, 2015).
- [20] “Fraud Detection Software Developers in the US Market Research | IBISWorld.” <http://www.ibisworld.com/industry/fraud-detection-software-developers.html> (accessed Feb. 07, 2015).
- [21] “How Ideas From Private Industry Help Combat Medicare Fraud, Waste, And Abuse – Health Affairs Blog.” <http://healthaffairs.org/blog/2013/05/23/how-ideas-from-private-industry-help-combat-medicare-fraud-waste-and-abuse/> (accessed Feb. 07, 2015).
- [22] J. Li, K. Huang, J. Jin, and J. Shi, “A survey on statistical methods for health care fraud detection,” *Health Care Management Science*, vol. 11, no. 3, pp. 275–87, Sep. 2008, doi: <http://dx.doi.org/10.1007/s10729-007-9045-4>.
- [23] D. Thornton, R. M. Mueller, P. Schoutsen, and J. van Hilleberg, “Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analysis Techniques for Fraud Detection,” *Procedia Technology*, vol. 9, pp. 1252–1264, 2013, doi: 10.1016/j.protcy.2013.12.140.

- [24] M. K. Sparrow, *License to steal: why fraud plagues America's health care system*. Boulder, Colo: Westview Press, 1996.
- [25] G. K. Palshikar, "The Hidden Truth - Frauds and their Control: A Critical Application for Business Intelligence," *Intelligence Enterprise*, vol. 5, no. 9, pp. 46-51, May 2002.
- [26] R. J. Bolton, D. J. Hand, and D. J. H, "Unsupervised Profiling Methods for Fraud Detection," in *Proc. Credit Scoring and Credit Control VII*, 2001, pp. 5–7.
- [27] R. Brause, T. Langsdorf, and M. Hepp, "Neural data mining for credit card fraud detection," in *11th IEEE International Conference on Tools with Artificial Intelligence, 1999. Proceedings*, 1999, pp. 103–106, doi: 10.1109/TAI.1999.809773.
- [28] R. J. Bolton and D. J. Hand, "Statistical Fraud Detection: A Review," *Statistical Science*, vol. 17, no. 3, pp. 235–249, Aug. 2002, doi: 10.2307/3182781.
- [29] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, Aug. 2007, doi: 10.1016/j.comnet.2007.02.001.
- [30] M. Gebiski, A. Penev, and R. K. Wong, "Grouping Categorical Anomalies," in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08*, Dec. 2008, vol. 1, pp. 411–414, doi: 10.1109/WIIAT.2008.162.
- [31] R. M. M. Peter Travaille, "Electronic Fraud Detection in the U.S. Medicaid Healthcare Program: Lessons Learned from other Industries.," *Obstetrics and Gynecology International*, 2011.
- [32] A. Sudjianto, S. Nair, M. Yuan, A. Zhang, D. Kern, and F. Cela-Díaz, "Statistical Methods for Fighting Financial Crimes," *Technometrics*, vol. 52, no. 1, pp. 5–19, Feb. 2010, doi: 10.1198/TECH.2010.07032.
- [33] Yufeng Kou, Chang-Tien Lu, S. Sirwongwattana, and Yo-Ping Huang, "Survey of fraud detection techniques," in *IEEE International Conference on Networking, Sensing and Control, 2004*, Mar. 2004, vol. 2, pp. 749-754 Vol.2, doi: 10.1109/ICNSC.2004.1297040.
- [34] F. E. Grubbs, "Procedures for Detecting Outlying Observations in Samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, Feb. 1969, doi: 10.1080/00401706.1969.10490657.
- [35] V. Barnett and T. Lewis, *Outliers in statistical data*. Chichester; New York: Wiley, 1994.

- [36] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature,” *Decision Support Systems*, vol. 50, no. 3, pp. 559-569, Feb. 2011, doi: 10.1016/j.dss.2010.08.006.
- [37] R. J. Bolton and D. J. Hand, “Peer Group Analysis – Local Anomaly Detection in Longitudinal Data,” Department of Mathematics, Imperial College, London, 2001.
- [38] E. Turban, T.-P. Liang, J. E. Aronson, and Sharda, T, *Decision Support Systems and Intelligent Systems*, 9th ed. Upper Saddle River: Pearson Education, 2005.
- [39] T. Fawcett and F. Provost, “Adaptive Fraud Detection,” *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 291–316, Sep. 1997, doi: 10.1023/A:1009700419189.
- [40] T. P. Hill, “A Statistical Derivation of the Significant-Digit Law,” *Statist. Sci.*, vol. 10, no. 4, pp. 354–363, Nov. 1995, doi: 10.1214/ss/1177009869.
- [41] M. Nigrini, *Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations*. John Wiley & Sons, 2011.
- [42] E. F. Codd, S. B. Codd, C. T. Salley, F. Codd, and C. Salley, “Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate,” Jan. 1993, Accessed: Mar. 08, 2020. [Online]. Available: <https://www.scienceopen.com/document?vid=03e6a6e2-7ae8-4c98-86fe-01dffce32cd1>.
- [43] A. Berger and T. P. Hill, “Benford’s Law Strikes Back: No Simple Explanation in Sight for Mathematical Gem,” *Math Intelligencer*, vol. 33, no. 1, pp. 85–91, Mar. 2011, doi: 10.1007/s00283-010-9182-3.
- [44] E. L. Barse, H. Kvarnstrom, and E. Jonsson, “Synthesizing test data for fraud detection systems,” in *Computer Security Applications Conference, 2003. Proceedings. 19th Annual*, Dec. 2003, pp. 384–394, doi: 10.1109/CSAC.2003.1254343.
- [45] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” *ACM Comput. Surv.*, vol. 41, no. 3, 2009, doi: 10.1145/1541880.1541882.
- [46] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Boston: Pearson Addison Wesley, 2005.
- [47] T. Vatanen, M. Kuusela, E. Malmi, T. Raiko, T. Aaltonen, and Y. Nagai, “Semi-supervised detection of collective anomalies with an application in high energy particle physics,” in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, Jun. 2012, pp. 1–8, doi: 10.1109/IJCNN.2012.6252712.



- [48] X. Song, M. Wu, C. Jermaine, and S. Ranka, “Conditional Anomaly Detection,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 5, pp. 631–645, May 2007, doi: 10.1109/TKDE.2007.1009.
- [49] Z. He, S. Deng, and X. Xu, “Outlier Detection Integrating Semantic Knowledge,” in *Proceedings of the Third International Conference on Advances in Web-Age Information Management*, London, UK, UK, 2002, pp. 126–131, Accessed: Feb. 15, 2015. [Online]. Available: <http://dl.acm.org/citation.cfm?id=645941.674192>.
- [50] Z. He, X. Xu, and S. Deng, “Squeezer: An efficient algorithm for clustering categorical data,” *J. Comput. Sci. & Technol.*, vol. 17, no. 5, pp. 611–624, Sep. 2002, doi: 10.1007/BF02948829.
- [51] Z. He, X. Xu, J. Z. Huang, and S. Deng, “Mining class outliers: concepts, algorithms and applications in CRM,” *Expert Systems with Applications*, vol. 27, no. 4, pp. 681–697, Nov. 2004, doi: 10.1016/j.eswa.2004.07.002.
- [52] V. J. Hodge and J. Austin, “A survey of outlier detection methodologies,” *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.
- [53] H. H. Bock and E. Diday, Eds., *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Berlin ; New York: Springer, 2000.
- [54] J. W. Tukey, “Exploratory data analysis,” 1977, Accessed: Sep. 27, 2016. [Online]. Available: <http://xa.yimg.com/kq/groups/16412409/1159714453/name/exploratorydataanalysis.pdf>.
- [55] L. Billard, “Editorial,” *Statistical Analysis and Data Mining*, Mar. 2011.
- [56] L. Billard, *Symbolic data analysis: conceptual statistics and data mining*. Chichester, England ; Hoboken, NJ: John Wiley & Sons Inc, 2006.
- [57] A. Flurry, “UGA’s Lynne Billard selected for Florence Nightingale David Award,” *UGA Today*, Oct. 09, 2013. <https://news.uga.edu/lynne-billard-florence-nightingale-david-award-1013/> (accessed Mar. 02, 2020).
- [58] M. Noirhomme-Fraiture and P. Brito, “Far beyond the classical data models: symbolic data analysis,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 4, no. 2, pp. 157–170, 2011, doi: 10.1002/sam.10112.
- [59] L. Billard and E. Diday, “From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis,” *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 470–487, Jun. 2003.

- [60] E. Mooi and M. Sarstedt, “Cluster Analysis,” in *A Concise Guide to Market Research*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 237–284.
- [61] E. Diday and F. Esposito, “An introduction to symbolic data analysis and the SODAS software,” *Intelligent Data Analysis*, vol. 7, no. 6, pp. 583–601, Dec. 2003.
- [62] K. C. Gowda and E. Diday, “Symbolic clustering using a new similarity measure,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 2, pp. 368–378, Mar. 1992, doi: 10.1109/21.148412.
- [63] M. Ichino and H. Yaguchi, “Generalized Minkowski metrics for mixed feature-type data analysis,” *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 24, no. 4, pp. 698–708, 1994.
- [64] F. de A. de Carvalho, P. Brito, and H.-H. Bock, “Dynamic clustering for interval data based on L 2 distance,” *Computational Statistics*, vol. 21, no. 2, pp. 231–250, 2006.
- [65] K. Mali and S. Mitra, “Clustering and its validation in a symbolic framework,” *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2367–2376, Oct. 2003, doi: 10.1016/S0167-8655(03)00066-7.
- [66] “Histogram – The Ultimate Guide of Binning - AnswerMiner.” <https://www.answerminer.com/blog/binning-guide-ideal-histogram> (accessed Mar. 05, 2020).
- [67] “pretty function | R Documentation.” <https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/pretty> (accessed Mar. 05, 2020).
- [68] A. Cury, C. Crémona, and E. Diday, “Application of symbolic data analysis for structural modification assessment,” *Engineering Structures*, vol. 32, no. 3, pp. 762–775, Mar. 2010, doi: 10.1016/j.engstruct.2009.12.004.
- [69] G. W. Milligan and M. C. Cooper, “An examination of procedures for determining the number of clusters in a data set,” *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985, doi: 10.1007/BF02294245.
- [70] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
- [71] M. F. A. Gadi, X. Wang, and A. P. do Lago, “Credit Card Fraud Detection with Artificial Immune System,” in *Artificial Immune Systems*, vol. 5132, P. J. Bentley, D. Lee, and S. Jung, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 119–131.
- [72] “R : Past and Future History -- A Free Software Project.” [https://cran.r-project.org/doc/html/interface98-paper/paper\\_2.html](https://cran.r-project.org/doc/html/interface98-paper/paper_2.html) (accessed Mar. 06, 2020).

- [73] “Ambulance rides often result in surprise bills,” *Healthcare Finance News*. <https://www.healthcarefinancenews.com/news/ambulance-rides-often-result-surprise-bills> (accessed Jan. 25, 2020).
- [74] A. Goldstein, “Fraudulent ambulance rides: Medicare paid more than \$50 million, IG says,” *The Washington Post*, Sep. 29, 2015.
- [75] “Inappropriate Payments and Questionable Billing for Medicare Part B Ambulance Transports (OEI-09-12-00351; 09/15),” p. 39.
- [76] “Billing for Ambulance Transports | CMS.” <https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Fast-Facts/Ambulance-Transport> (accessed Jan. 24, 2020).
- [77] P. Feb 13 and 2019, “An Overview of Medicare,” *The Henry J. Kaiser Family Foundation*, Feb. 13, 2019. <https://www.kff.org/medicare/issue-brief/an-overview-of-medicare/> (accessed Mar. 05, 2020).