5-2021

KAMILA Clustering for a Mixed-Type Data Analysis of Illinois Medicare Data

Heather Elizabeth Baldacci

Applicant	Heather Baldacci
Thesis Title	KAMILA Clustering for a Mixed-Type Data Analysis of Illinois Medicare Data
Intended Date of Commencement	May 9, 2021

Read, Approved, and Signed by:

Thesis Advisor: ..... Date 5/3/2021

Reader: Christopher Wilson Date 5th May 2021

Certified by: ..... Date .....

For Honors Program use:

Level of Honors conferred: University _____

Departmental _____

KAMILA Clustering for a Mixed-Type Data Analysis of Illinois Medicare Data

Heather Baldacci

May 2021

An Undergraduate Thesis

Presented to The Honors Program

of Butler University

Supervised by Dr. Rasitha Jayasekare

In Partial Fulfillment

of the Requirements for Graduation Honors

Acknowledgement of Sources

For all ideas taken from other sources (books, articles, internet), the source of the ideas is mentioned in the main text and fully referenced at the end of the report.

All material which is quoted essentially word-for-word from other sources is given in quotation marks and referenced.

Pictures and diagrams copied from the internet or other sources are labelled with a reference to the web page or book, article etc.

Special thanks to the Butler University Department of Mathematics, Statistics, and Actuarial Science, the American Mathematical Society, and the Dean of the College of Liberal Arts and Sciences at Butler for supporting this thesis and its presentation at the Joint Mathematics Meeting.

Signed Heather Bultmann Date 5-5-2021

ABSTRACT

KAMILA Clustering for a Mixed-Type Data Analysis of Illinois Medicare Data

Heather Baldacci

May 2021

The Centers for Medicare and Medicare Services (**CMS**) releases annual reports regarding the Market Saturation and Utilization of nationwide Medicare coverage. CMS data provide an opportunity for an in-depth analysis of Medicare usage patterns within the United States that may provide insight into socioeconomic conditions in certain regions. To discover any potential patterns, the **KAMILA** (**KA**y-means for **MI**xed **LA**rge data sets) clustering algorithm has been utilized within the most recent CMS dataset from 2018. Due to the large size of the original data set, the focus of this research has been limited to Illinois Medicare data, grouped by the 102 *counties* in Illinois. The KAMILA algorithm extends the well-known *k-means* clustering algorithm to include mixed-type data by using a weighted semi-parametric procedure. Therefore, it balances the contribution of quantitative and qualitative variables. The optimal number of clusters is decided in-part by the operator of the algorithm with respect to the number of cross-validation runs. After the application of the KAMILA clustering algorithm with both the main CMS dataset and a modified version of it to exclude Cook *County*, two clusters were found with both datasets. This offers insight into the structure of Medicare *Services* in the state of Illinois.

Contents

Chapter 1

Introduction

This section is comprised of the Introduction, the Literature Review, and the Research Objectives.

1.1 Introduction to Medicare

Health insurance as well as health care in the United States are popular topics, especially in today's political environment. Health insurance is an ever-changing field, whether it comes in the form of private or company-supplied health insurance or government-supplied Medicare or Medicaid. Medicare is a federal health insurance program that funds many health care expenses for its beneficiaries ("Facts About Medicare" 2020). It was established in 1965, and is supplied by the Centers for Medicare and Medicaid Services (**CMS**), which is a part of the United States Department of Health and Human Services (HHS). Most beneficiaries are aged 65 and older, but some adults with permanent disabilities or other conditions are granted Medicare benefits. Like Social Security, the majority of United States citizens are able to register for Medicare if they have worked or paid taxes for a minimum amount of time ("Facts About Medicare" 2020).

Medicare itself consists of four parts: A (Hospital Insurance), B (Medical Insurance), C (Medicare Advantage), and D (Prescription drug coverage) ("The Official U.S. Government Handbook" 2020). In 2019, about 60.6 million Americans received coverage through Medicare, and as of 2017, Medicare consisted of 15 percent of federal spending. This number is expected to grow to 18 percent by 2028 (Anderson 2019).

Today, it is particularly useful to look more deeply into Medicare data. The Baby Boomer generation is particularly large compared to the generations that follow it, and this could potentially put a strain on the United

States' government spending. Additionally, in today's day and age, the utilization of health care services is greater than it has ever been before, both with respect to the costs as well as the number of people within the Medicare system. It would be interesting to see how each *Type of Service* that Medicare offers differs from one another.

On a larger scale, Medicare data can delve more deeply into the network of the United States' medical facilities and their potential regional shortcomings. For instance, Long-Term Care Hospitals may be found more prominently in specific counties of a state or regions of the United States. This could point to a lack of financial support for Long-Term Care Hospitals in less densely populated or less affluent areas. Therefore, the analysis of Medicare data could potentially research a positive change in the way that the US healthcare system already operates and bring it to the caliber of other developed countries in regions where data find it is lacking.

In this thesis, the CMS Medicare data will be thoroughly analyzed through the use of the **KAMILA** (**KA**y means for **MI**xed **LA**rge datasets) algorithm, a technique which has not been used as often as other clustering methods. Due to the fact that Butler University is located in the Midwest, the CMS data being utilized for these methods has been narrowed down to Illinois's Medicare data. Illinois is a more highly populated Midwestern state, and it has a multitude of data that can be studied with respect to Medicare data.

1.2 Literature Review

In order to guarantee that the subject of this thesis is unique, other works pertaining to the Centers for Medicare and Medicaid Services (CMS) and the KAMILA clustering algorithm must be thoroughly researched and consulted. Although many people have researched the Medicare field, the majority of these researchers have taken on an approach that pertains more to the field of psychology or into specific aspects of Medicare services. Most of the research does not appear to come from an actuarial science background.

CMS star ratings, an aspect of Medicare, was mentioned by Oxley (2018) in her dissertation about seniors' knowledge of their Medicare Advantage plans. In this case study, the author utilized a sample of twenty senior adults from Florida in order to see whether or not they were aware of the CMS star ratings. The CMS star ratings program is a rating system of the Medicare Advantage (MA) plan, which is a plan offered by a private company from Medicare that contracts with Medicare (Oxley 2018). The dissertation was

more focused on the feelings of the seniors towards the program, which is different from utilizing the CMS dataset that is being used in this thesis.

Researchers at the University of Iowa (Belatti et al. 2014) have also delved into Medicare with a special interest in Total Joint Arthroplasty, which is a surgical process used to restore the function of a joint. This study found that the cost of orthopedic implants is increasing, while there is a decline in physician reimbursements. This contrasting combination could put a strain on the Medicare budget. This research, while it may be fascinating, is more niche than an analysis of the overall usage of Medicare because it pertains more to a specific aspect of Medicare and the budget of Medicare.

Another recent study that includes Medicare data was done in August of 2018 (Chung and Sorensen 2018). This research focuses on hospices operating from 2000-2012 in the United States. They estimated a model of patient demand for hospices and concluded that hospices became more profitable due to competition among hospices, not an increase in the number of hospices available. This dataset utilized hospice data that was supplied by Medicare from 2000 to 2012, hence it is not as recent as the CMS dataset being used in this thesis. It also focused specifically on the pricing of hospice services, not the trends in Medicare data. Thus, this study, while relevant to the hospice side of Medicare, differs drastically from the subject of this thesis.

The KAMILA clustering algorithm has received much attention throughout the past decade, but most of the published work describes how to implement it and the process behind using it for statistical research. Consequently, these works will be thoroughly referenced throughout the data modeling process because they will serve as useful guides to the clustering algorithm processes in this thesis. On the other hand, the CMS dataset that will be used during this thesis has not been fully referenced in any research published prior to this thesis. This will make the application of the clustering algorithms mentioned above into Illinois's CMS data very interesting and new in the field of Medicare research.

1.3 Research Objectives

In June of 2020, the CMS updated its latest "Market Saturation and Utilization Dataset" to include its most recent data from 2018, focusing primarily on Medicare data that pertain to parts A and B of Medicare. It will be referred to as the CMS dataset throughout this thesis. The dataset includes just under 750,000 records that group the data by nation, state, county, and year. For this thesis, this dataset has been narrowed down to

just Illinois's 2018 data by *County*. Illinois data was chosen for this thesis due to personal familiarity with the state.

In order to begin the mining of the CMS dataset for Illinois in 2018, the dataset must be analyzed and properly understood. The variables that will be used in the dataset are the *County*, the *Number of Fee-for-Service Beneficiaries*, the *Number of Providers*, the *Number of Users*, and the *Total Payment* for each respective county, grouped by the *Type of Service*, of which there are 18 different values. The *Types of Service* included in the dataset are those that belong to Medicare parts A and B. The part A *services*, which help to cover Hospital Insurance, consist of Home Health, Hospice, Independent Diagnostic Testing Facility Pt A, Long-Term Care Hospitals, and Skilled Nursing Facility *services*. The part B services, which help cover Medical Insurance, consist of the Ambulance (Emergency & Non-Emergency), Ambulance (Emergency), Ambulance (Non-Emergency), Cardiac Rehabilitation Program, Chiropractic Services, Clinical Laboratory (Billing Independently), Dialysis, Federally Qualified Health Center (FQHC), Independent Diagnostic Testing Facility Pt B, Ophthalmology, Physical & Occupational Therapy, Preventative Health Services, and Psychotherapy *services* ("The Official U.S. Government Handbook" 2020). In summary, it will be interesting to see if some of these *Types of Service* have a higher *Number of Users* or *Providers* depending on the *County* that they are used in, and to see if this offers any insight into the wealth or size of each *County*.

In order to identify patterns of service usages from the dataset, statistical algorithms must be implemented to help group the data in more meaningful ways and enable the reader to see more significant variables or trends in the dataset. To find these underlying natural structures in the data, **cluster analysis** methods must be utilized.

Cluster analysis constitutes a variety of techniques that attempt to identify unknown structures or patterns in a dataset without any initial references (Foss and Markatou 2018). With cluster analysis, the goal is to group the data in a way that observations with certain underlying similarities may be grouped together. As a result, these observations are put together and set apart from the other groups created by clustering. Clustering is an unsupervised learning method, meaning that there is little direction in how to group the variables in a meaningful way. Overall, clustering deals with uncertainties, and it will hopefully be useful when discovering underlying patterns in the CMS dataset on Illinois's 2018 Medicare data.

It is essential to understand the different types of data when selecting a clustering algorithm. Data are typically categorized as qualitative or quantitative. Qualitative data describe characteristics, labels, and levels of ob-

servations, while quantitative data are numerical counts or measurements. Data that consist of both quantitative and qualitative variables are known as **mixed-type** data. The dataset under study for this thesis has both quantitative and qualitative variables. Therefore, the well-known clustering algorithms such as the *k-means* and *hierarchical* clustering algorithms cannot be applied here because they are applied for solely numerical (quantitative) data. In this dataset, the continuous (quantitative) variables are: the *Number of Fee-for-Service Beneficiaries*, the *Number of Providers*, the *Number of Users*, and *Total Payment*. The qualitative variables are the *Counties* and the *Type of Service*, of which there are 18 different possibilities.

The method of clustering that will be enacted with this dataset is the KAMILA (KAy means for MIxed LArge datasets) clustering algorithm. This algorithm uses the mixed-type data, signifying that the KAMILA clustering algorithm can be applied to the CMS dataset. The KAMILA clustering algorithm extends the well-known *k-means* clustering algorithm to include mixed-type data by using a weighted semiparametric procedure (Foss and Markatou 2018). Overall, it will be fascinating to implement a clustering algorithm that is tailored to the nature of the CMS dataset.

In summation, with the CMS dataset, it will be interesting to implement the KAMILA clustering algorithm to uncover any unnoticeable patterns of usage of Medicare's different *Types of Service* (only parts A and B), the *Number of Providers*, and the *Number of Users* in the state of Illinois.

Chapter 2

Data Pre-Processing

This section is comprised of data pre-processing, which includes dimension reduction, outlier analysis, and data visualization.

2.1 Introduction

The first step to the any data mining process is to thoroughly pre-process the data. This is done in order to understand the dataset and prepare it to be used in data mining. Essentially, data pre-processing is the procedure that converts raw data into a format that is ready to be used in core data mining tasks.

Typically data pre-processing consists of activities such as dimension reduction, outlier analysis, data visualization, and data normalization. In order to pre-process data to use in cluster analysis, dimension reduction, outlier analysis, and data visualization will be performed. These steps that lead up to the implementation of the KAMILA algorithm will be explained further throughout this chapter.

2.2 Dimension Reduction

As previously mentioned in the Research Objectives section of Chapter 1, the CMS "Market Saturation and Utilization Dataset" contains 747,944 records. The expansive size of the dataset is due to the CMS's inclusion of grouping the different *Types of Service* by "Nation + Territories", "State," and "County." As a result, much of the data in the dataset are made up of other records that are outside the scope of this thesis, and they must be promptly removed. The decision to look into just Illinois Medicare data was made due to both a personal connection to the state of Illinois, as well as

the fact that the dataset would contain only 1,737 records if only Illinois was analyzed. This value is optimal for testing out the KAMILA clustering algorithm due to its large enough size. An odd value of 1,737 records occurs because not all 102 *Counties* in Illinois offer all 18 different *Types of Service*. For example, Warren County does not have any Long-Term Care Hospitals.

Additionally, there are 32 different variables in the dataset, and many of these variables are combinations or percentages of seven primary variables. For example, there is a variable in the original dataset called the *Percentage of Users out of FFS Beneficiaries*. This variable takes known data from the *Number of Users* and the *Number of Fee-for-Service Beneficiaries* variables. The other 25 variables excluding the seven primary variables are similar in form to this variable, combining the primary variables. Thus, the dataset was narrowed down those seven primary variables, including the *County*, the *Number of Fee-for-Service Beneficiaries*, the *Number of Providers*, the *Number of Users*, the *Total Payment*, and the *Type of Service*. These seven variables are applicable to the KAMILA clustering algorithm because the algorithm is used primarily with mixed-type datasets. In this instance, there are five quantitative variables and two qualitative variables, fulfilling the mixed-type data requirement.

Now that the dataset has gone through the dimension reduction process, it is time to move onto the outlier analysis.

2.3 Outlier Analysis

Outlier analysis is the process of identifying extreme values, or values that are significantly different from the remainder of the dataset. Extreme values typically skew the main results of analyses performed on the dataset. The decision to keep or remove outliers is dependent upon the application of the dataset, as well as how the statistical results are treated. Since the objective of this thesis is to use cluster analysis to identify the natural groupings of the CMS dataset, all outliers will be kept in the dataset. Even though the outliers are being kept in the dataset, it is important to recognize them so that they can be compared with the final results of the clustering process. Therefore, this outlier analysis will be utilized to further understand data and identify any extreme values.

Boxplots will be used to detect outliers in the five quantitative variables of the CMS dataset, as boxplots can only be created using quantitative data. The five quantitative variables are: the *Number of Fee-for-Service Beneficiaries*, the *Number of Providers*, *Number of Users*, the *Number of Dual Eligible*

Users, and the *Total Payment* variables. The following boxplots have been created using the *ggplot2* package in *R* using *RStudio*.

Each variable features two boxplots: one that is zoomed out to display the largest outliers, and one that is zoomed further in to show the aspects of the distributions with respect to each *Type of Service*.

To accommodate for this lack of qualitative data representation, the boxplots have been sorted by each of the 18 different *Types of Service*. Each *Type of Service*'s name is shortened to allow for a clearer output, so the legend for the names is listed in Figure 2.1. In terms of the *County* variable, it will be represented in Section 2.4 through data visualization. It is unrealistic to represent every variable with every *County* in Illinois because there are 102 *Counties*, and the graphs would not be clear enough to supply any insight. That being said, each boxplot will contain a list of the highest outliers. The corresponding *County* and *Type of Service* will be mentioned as well. The *County* size will also be referenced in these lists as well to see which counties to look into during the data visualization section.

Abbreviation	*Type of Service*
AmENE	Ambulance (Emergency & Non-Emergency)
AmE	Ambulance (Emergency)
AmNE	Ambulance (Non-Emergency)
CRP	Cardiac Rehabilitation Program
Chiro	Chiropractic Services
CL	Clinical Laboratory (Billing Independently)
D	Dialysis
FQHC	Federally Qualified Health Center (FQHC)
HH	Home Health
Hosp	Hospice
IDTFA	Independent Diagnostic Testing Facility Pt A
IDTFB	Independent Diagnostic Testing Facility Pt B
LTCH	Long-Term Care Hospitals
Op	Ophthalmology
P&OT	Physical & Occupational Therapy
PHS	Preventative Health Services
Psy	Psychotherapy
SNF	Skilled Nursing Facility

Figure 2.1: Legend for the Abbreviations of the Difference Types of Service used in Outlier Analysis

2.3.1 Variable: *Fee-for-Service-Beneficiaries* by *Type of Service*

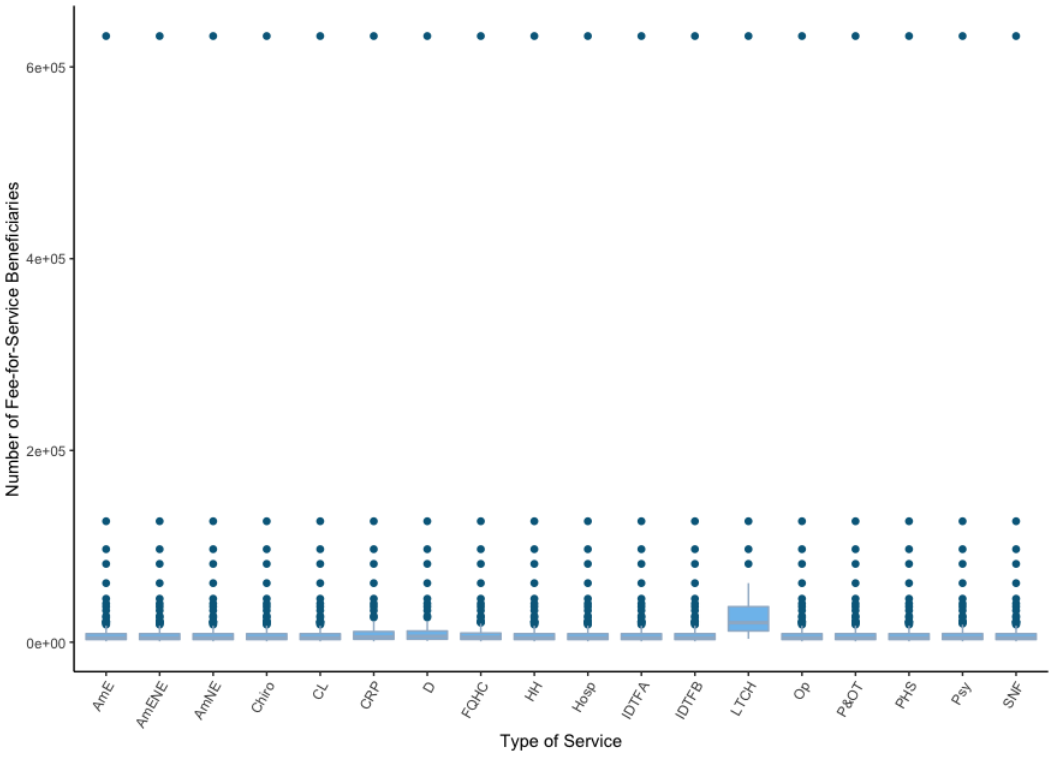

Figure 2.2: Boxplots of Fee-for-Service Beneficiaries by Type of Service

The boxplots above in Figure 2.2 show the *Number of Fee-for-Service Beneficiaries*, grouped by the 18 different *Types of Service*. The *Number of Fee-for-Service Beneficiaries* variable is the number of health care providers that are paid separately for each particular service rendered ("Understanding Fee-for-Service" 1).

Looking at these boxplots, it is clear that the *Fee-for-Service Beneficiaries* variable consists of repeated data values. This is due to the fact that the *Number of Fee-for-Service Beneficiaries* variable has the same value for every *County* in Illinois. The way that the CMS measures it is by *County*, instead of *Type of Service*. For example, the *Number of Fee-for-Service Beneficiaries* in Adams *County* for Dialysis *Services* is 13,234. The *Number of Fee-for-Service Beneficiaries* for every other *Type of Service* in Adams *County* is also 13,234. However, some aspects of the *Type of Service* qualitative variable distributions have slightly different outputs than the others due to their lack

of representation in certain *Counties*.

One of note is the Long-Term Care Hospitals (LTCH) *Type of Service*. This is attributed to the fact that some *Counties* do not offer this Medicare *service*, so they are left out in the *Number of Fee-for-Service Beneficiaries* variable. In these boxplots, it is clear that there is one significant outlier. This outlier is numbered at 632,224 *Fee-for-Service Beneficiaries* in Cook *County* for every *Type of Service* in Cook *County*. Cook *County* is where Chicago is located, so the *Number of Fee-for-Service Beneficiaries* would likely make sense as the highest outlier because Chicago is densely populated. It will to interesting to see if this could be due to either affluence, population size, or both.

To look closer at the patterns of these outliers and the shapes of the distributions, a rescaled version of the boxplots can be found below, in Figure 2.3.

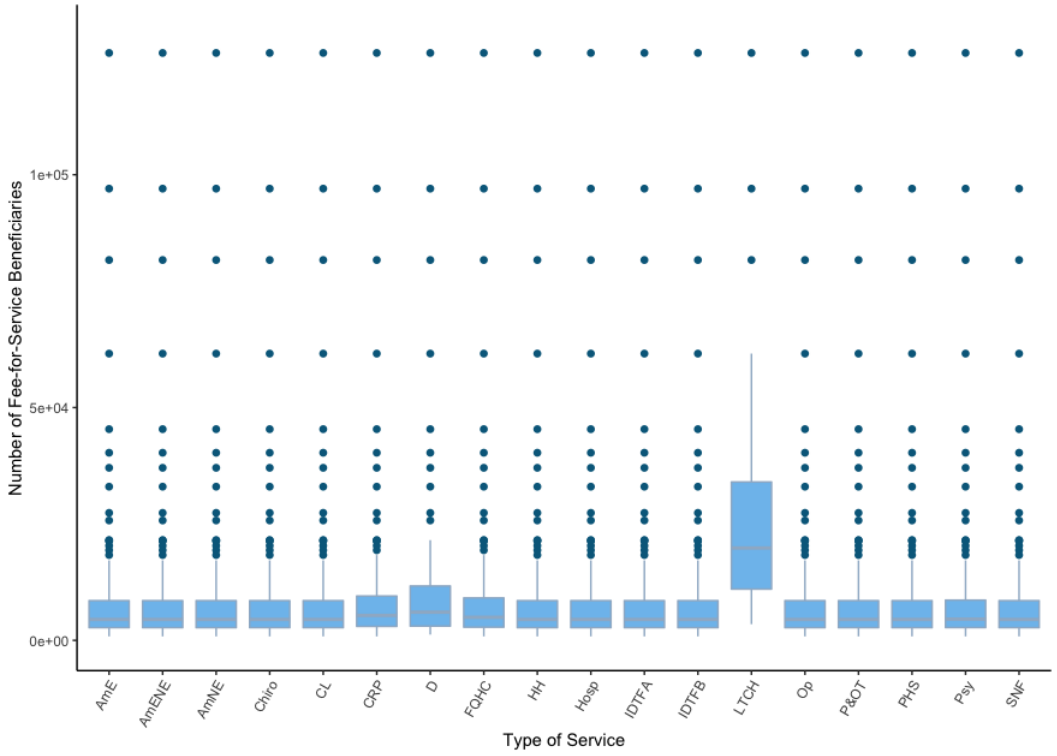

Figure 2.3: Boxplots of Fee-for-Service Beneficiaries by Type of Service Zoomed In

The boxplot above in Figure 2.3 gives a better look at the outliers on a smaller scale, excluding the Cook county outlier at 632,224. Here, it is clear

that there are four other significant outliers of note. The top one is numbered at 126,156 *Fee-for-Service Beneficiaries*, so this indicates that DuPage *County* has the second highest *Number of Fee-for-Service Beneficiaries*. DuPage *County* is also the second most populated *County* in Illinois, after Cook County ("Illinois Counties by Population" 1). The next largest outlier is numbered at 97,014 *Fee-for-Service Beneficiaries*. This is located in Lake *County*, which happens to be the third most populated *County* in Illinois ("Illinois Counties by Population" 1). The next significant outlier is at 81,686 *Fee-for-Service Beneficiaries*. This is for Will County, which happens to be the fourth most populated *County* in Illinois ("Illinois Counties by Population" 1). Lastly, the next significant outlier is at 61,587 for Kane County, which is the fifth most populated *County* in Illinois ("Illinois Counties by Population" 1). A list of the significant outliers from both *Fee-for-Service Beneficiaries* boxplots can be found in Figure 2.4.

Looking more closely at the boxplots, it is evident that the Long-Term Care Hospitals (LTCH) *Type of Service* appears to have a lower number of *Fee-for-Service Beneficiaries*, likely because they require a lot of upkeep and financial aid.

Significant Outlier Value (in *Number of Fee-for-Service Beneficiaries*)	Type of Service	County	Information about County
632,224	Same for All Types of Service	Cook	IL 1st most populated *County*
126,156	Same for All Types of Service	DuPage	IL 2nd most populated *County*
97,014	Same for All Types of Service	Lake	IL 3rd most populated *County*
81,686	Same for All Types of Service	Will	IL 4th most populated *County*
61,587	Same for All Types of Service	Kane	IL 5th most populated County

Figure 2.4: Top Five Outliers in Fee-for-Service Beneficiaries Variable

In summary, looking at both the boxplots and the table above in Figure 2.4, it is evident that the population of each *County* in Illinois carries more weight in the distributions for the *Number of Fee-for-Service Beneficiaries*. It will be interesting to see if this pattern is continued the rest of the boxplots, as well as throughout the clustering process. In the next subsection, the *Number of Providers* outliers will be analyzed.

2.3.2 Variable: *Number of Providers* by *Type of Service*

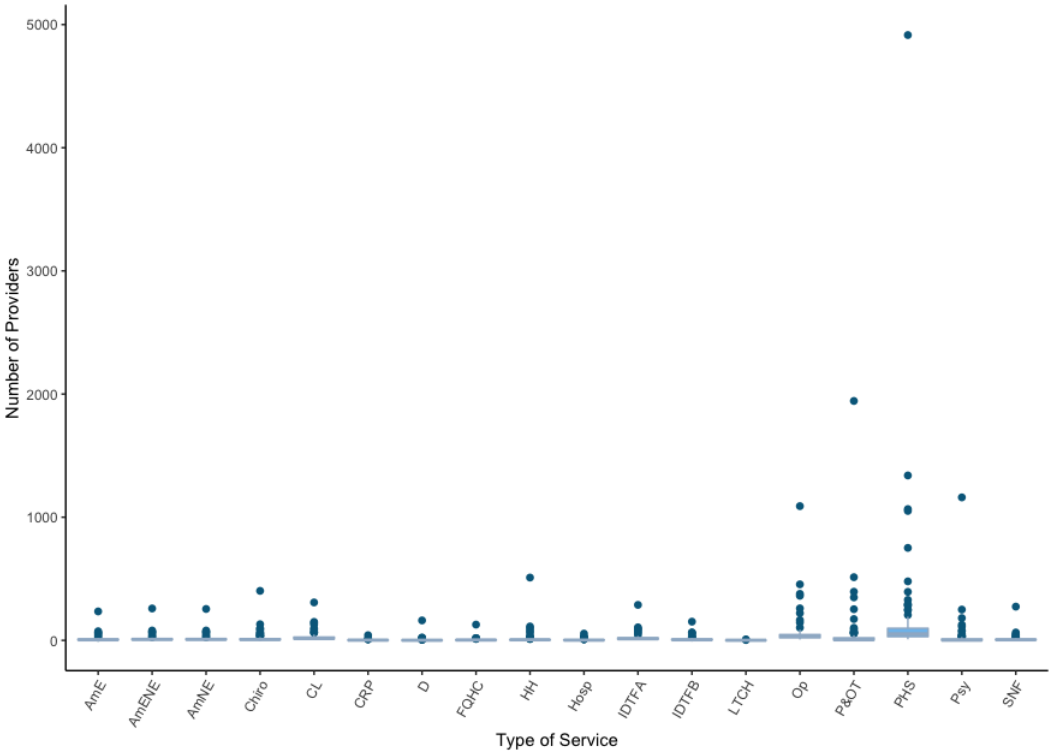

Figure 2.5: Boxplots of Number of Providers by Type of Service

The boxplots shown above in Figure 2.5 display the outliers for the *Number of Providers* of Medicare *services*, grouped by each of the 18 different *Types of Service*. The *Number of Providers* variable is the total number of facilities/practices that provide Medicare *services*. In the graph, there are about eight outliers of note. The highest outlier is located in the Preventative Health Services (PHS) *Type of Service*, with a value of 4,915 Preventative Health Services (PHS) *Providers*. The *County* that this is located in is Cook *County*. The second highest outlier value is at 1,944 *Providers* of Physical & Occupational Therapy (P&OT) in Cook *County*. The third highest outlier value is at 1,339 *Providers* of Preventative Health Services (PHS) in DuPage *County*. The fourth highest outlier value is at 1,161 *Providers* of Psychotherapy (Psy) in Cook *County*. The fifth highest outlier value is at 1,090 *Providers* of Ophthalmology (Op) in Cook *County*. The sixth highest outlier value is at 1,066 *Providers* of Preventative Health Services (PHS) in

Will *County*, followed closely by 1,051 *Providers* of Preventative Health Services (PHS) in Lake *County*. The eighth highest outlier is at 751 *Providers* of Preventative Health Services (PHS) in Kane *County*.

It should be noted the Preventative Health Services (PHS) encompass a number of preventative screening services, such as abdominal aortic aneurysm screening, alcohol misuse screenings and counseling, bone mass measurements, cardiovascular disease screenings, cervical and vaginal cancer screening, depression screening, diabetes screenings, mammograms, glaucoma tests, and many more ("Preventive & Screening Services" 1). It is a very broad *Type of Service*, so it would make sense that there are many *Providers*.

The rescaled version of the boxplots above in Figure 2.5 can be found below, in Figure 2.6, along with the list of outliers for the *Number of Providers* variable.

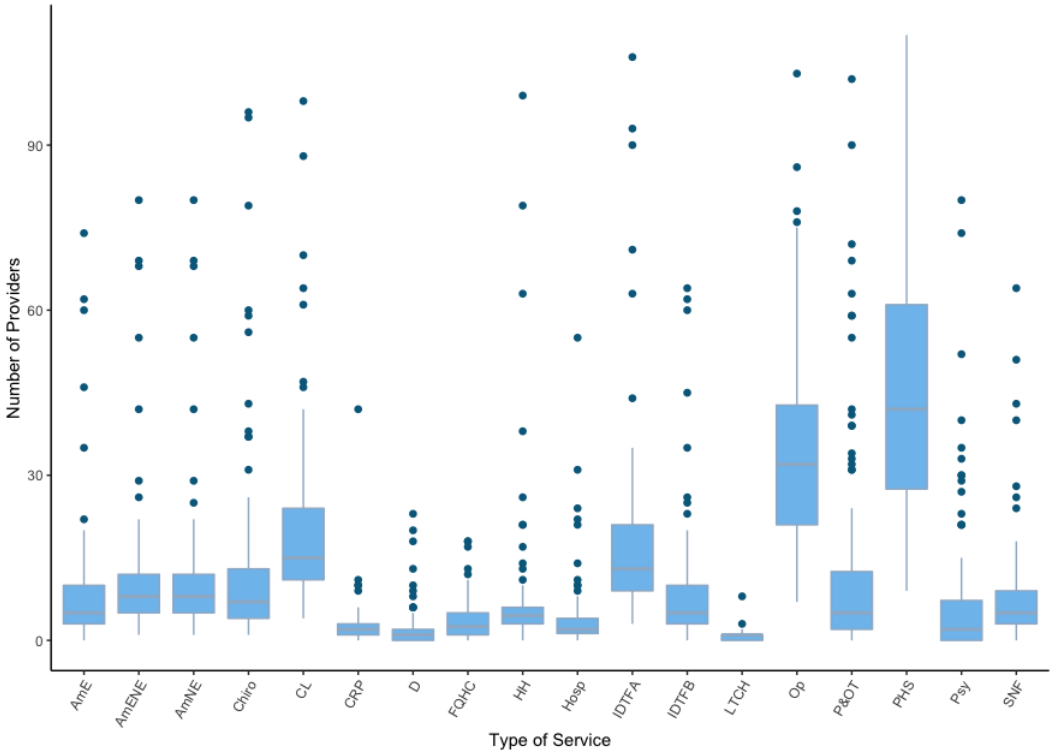

Figure 2.6: Boxplots of Number of Providers by Type of Service Zoomed In

This rescaled boxplot in Figure 2.6 shows the distributions of the *Number of Providers* in terms of the *Type of Service*. It seems like the traditionally more expensive services have fewer *Providers*, and less expensive services have more *Providers*. For instance, Preventative Health Services (PHS), Ophthal-

mology (Op), Clinical Laboratory (CL), and Independent Diagnostic Testing Facility Pt A *Types of Service* have a higher *Number of Providers* compared to the rest of the distributions. Conversely, Long-Term Care Hospitals (LTCH), Dialysis (D), Cardiac Rehabilitation Programs (CRP), and Federally Qualified Health Centers (FQHC) have a smaller *Number of Providers*. Additionally, there is a lot of variability with the *Number of Providers* of Preventative Health Services (PHS) and Ophthalmology (Op).

Significant Outlier Value (in *Number of Providers)*	Type of Service	County	Information about County
4,915	Preventative Health Services	Cook County	IL 1st most populated County
1,944	Physical and Occupational Therapy	Cook County	IL 1st most populated County
1,339	Preventative Health Services	DuPage County	IL 2nd most populated County
1,161	Psychotherapy	Cook County	IL 1st most populated County
1,090	Ophthalmology	Cook County	IL 1st most populated County
1,066	Preventative Health Services	Will County	IL 4th most populated County
1,051	Preventative Health Services	Lake County	IL 3rd most populated County
751	Preventative Health Services	Kane County	IL 5th most populated County

Figure 2.7: Top Eight Outliers in Number of Providers Variable

Overall, judging from the boxplot and the table above (Figure 2.7), there appears to be a pattern of the higher *Number of Provider* values being located in more populated *Counties*. The top five most populated *Counties* in Illinois appear in this list, with Cook *County*, the most populated county, taking the lead for each *Type of Service*. It appears that there are a lot of *Providers* of Preventative Health Services (PHS), Psychotherapy (Psy), and Ophthalmology (Op). This could potentially be attributed to the fact that many family medical practices are represented by these specific fields.

Next, the *Number of Users* variable will be discussed; this variable should be interesting to relate to the *Number of Providers* variable because these two variables are very highly associated with each other.

2.3.3 Variable: *Number of Users* by *Type of Service*

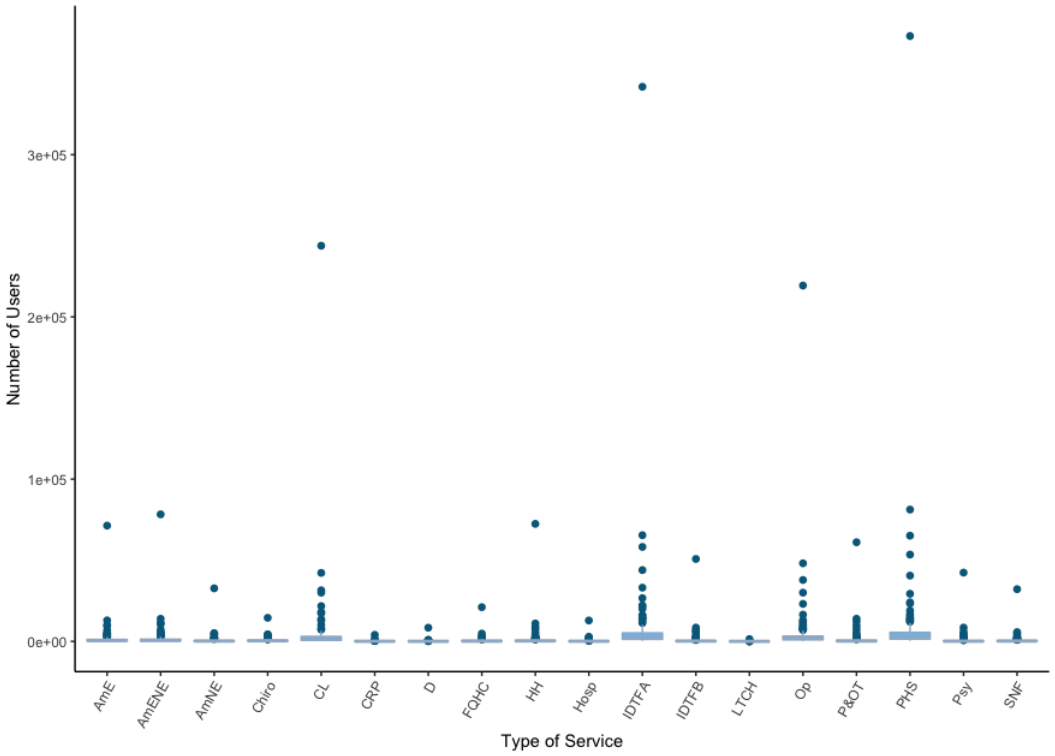

Figure 2.8: Boxplots of Number of Users by Type of Service

The boxplots shown above in Figure 2.8 display the outliers for the *Number of Users*, grouped by each of the 18 different *Types of Service*. The *Number of Users* variable is the number of people with Medicare who use the specific *Types of Service*. At a first glance, it appears that there are many highly valued outliers with the Clinical Laboratory (CL), Independent Diagnostic Testing Facility Pt A (IDTFA), Ophthalmology (Op), and Preventative Health Services (PHS) variables. The only outliers that will be analyzed in this section, however, are the highest seven.

The highest *Number of Users* outlier is 373,118 *Users* of Preventative Health Services (PHS) in Cook *County*. The second highest outlier is at 341,846 *Users* of Independent Diagnostic Testing Facility Pt A in Cook *County*. The third highest outlier is 243,880 *Users* of Clinical Laboratory (CL) in Cook *County*. The fourth highest outlier is 219,308 *Users* of Ophthalmology (Op) in Cook *County*. The fifth highest outlier is at 81,243 *Users* of Preventative Health Services (PHS) in DuPage *County*. The sixth highest

outlier is at 78,304 *Users* of Ambulance (Emergency & Non-Emergency) in Cook *County*. Lastly, the seventh highest outlier is at 72,438 *Users* of Home Health (HH) in Cook *County*.

To look closer at the patterns of these outliers and the shapes of the distributions, a rescaled version of the boxplots can be found below, in Figure 2.9.

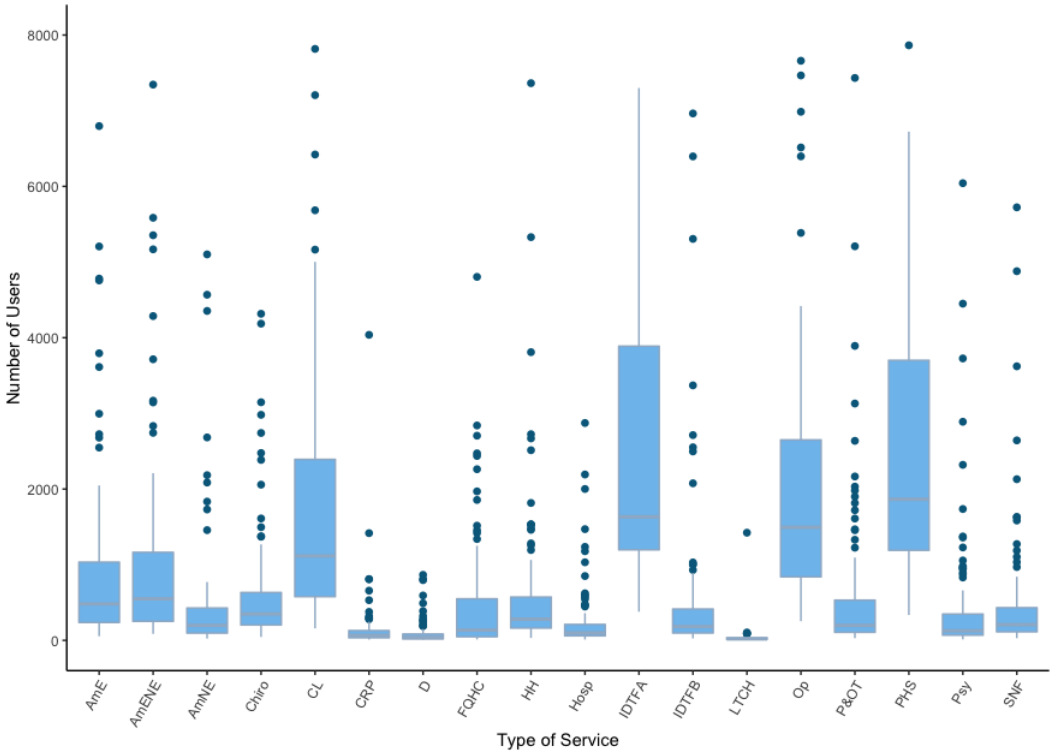

Figure 2.9: Boxplots of Number of Users by Type of Service Zoomed In

The boxplots shown above in Figure 2.9 are the same as before, just with a rescaled y-axis to look more in depth at the boxplots themselves. Long-Term Care Hospitals (LTCH), Cardiac Rehabilitation Programs (CRP), and Dialysis (D) services have a lower *Number of Users*. These *Types of Service* also had the lowest *Number of Providers* in Figure 2.6, which would intuitively makes sense; a *Type of Service* with fewer *Providers* would likely have fewer *Users* as well. Conversely, Independent Diagnostic Testing Facility Pt A (DTFA), Ophthalmology (Op), Preventative Health Services (PHS), and Clinical Laboratory (CL) have a high *Number of Users*. This was also the same in Figure 2.6, where it is clear to see that these four *services* have a high *Number of Providers*.

Significant Outlier Value (in *Number of Users*)	Type of Service	County	Information about County
373,118	Preventative Health Services	Cook	IL 1st most populated County
341,846	Independent Diagnostic Testing Facility Pt A	Cook	IL 1st most populated County
243,880	Clinical Laboratory	Cook	IL 1st most populated County
219,308	Ophthalmology	Cook	IL 1st most populated County
81,243	Preventative Health Services	DuPage	IL 2nd most populated County
78,304	Ambulance (Emergency & Non-Emergency)	Cook	IL 1st most populated County
72,438	Home Health	Cook	IL 1st most populated County

Figure 2.10: Top Seven Outliers in Number of Users Variable

Based on the table above in Figure 2.10, it is clear to see that Cook *County* accounts for six of the seven highest outliers in the boxplot, with Preventative Health Services (PHS) having the highest *Number of Users*. The only other *County* to make the list is DuPage *County*, the second most populated county in Illinois, with Preventative Health Services (PHS). Therefore, it seems evident that Preventative Health Services are by far the most in-demand Medicare Services by *Providers* and *Users* alike. Additionally, these values seem heavily influenced by the population of a *County*.

Thus, it was interesting to see how the *Number of Users* and *Number of Providers* variables appear to be very similar. Now, the next subsection will focus on the *Number of Dual Eligible Users* variable.

2.3.4 Variable: *Number of Dual Eligible Users* by *Type of Service*

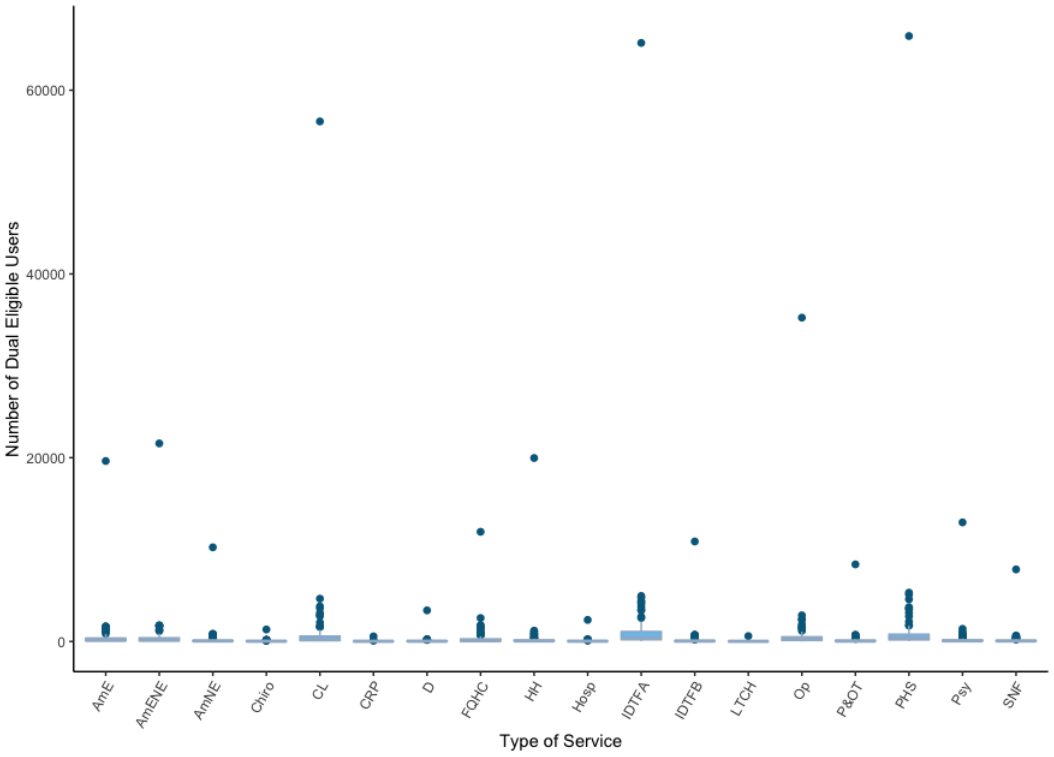

Figure 2.11: Boxplots of Number of Dual Eligible Users by Type of Service

The boxplots shown above in Figure 2.11 display the *Number of Dual Eligible Users* in terms of the 18 different *Types of Service*. A *Dual Eligible User* is someone who qualifies for both Medicare and Medicaid, meaning that they could potentially have a lower level of income and/or a disability to take into consideration. This explains why its boxplots look very similar to the boxplots for the *Number of Users* in Figures 2.8 and 2.9.

Based on the collection of boxplots, there are about seven outliers of major influence will be analyzed more in-depth. The highest outlier is at 65,894 *Dual Eligible Users* of Preventative Health Services (PHS) in Cook *County*. The second highest outlier is at 65,148 *Dual Eligible Users* of Independent Diagnostic Testing Facility Pt A (IDTFA) in Cook *County*. The third highest outlier is at 56,601 *Dual Eligible Users* of Clinical Laboratory (CL) in Cook *County*. The fourth highest outlier is at 35,246 *Dual Eligible Users* of Ophthalmology in Cook *County*. The fifth highest outlier is at 21,557

Dual Eligible Users of Ambulance (Emergency & Non-Emergency) in Cook *County*. The sixth highest outlier is at 19,968 *Dual Eligible Users* of Home Health (HH) in Cook *County*. The seventh highest outlier is at 19,642 *Dual Eligible Users* of Ambulance (Emergency) in Cook *County*.

To look closer at the patterns of these outliers and the shapes of the distributions, a rescaled version of the boxplots can be found below, in Figure 2.12.

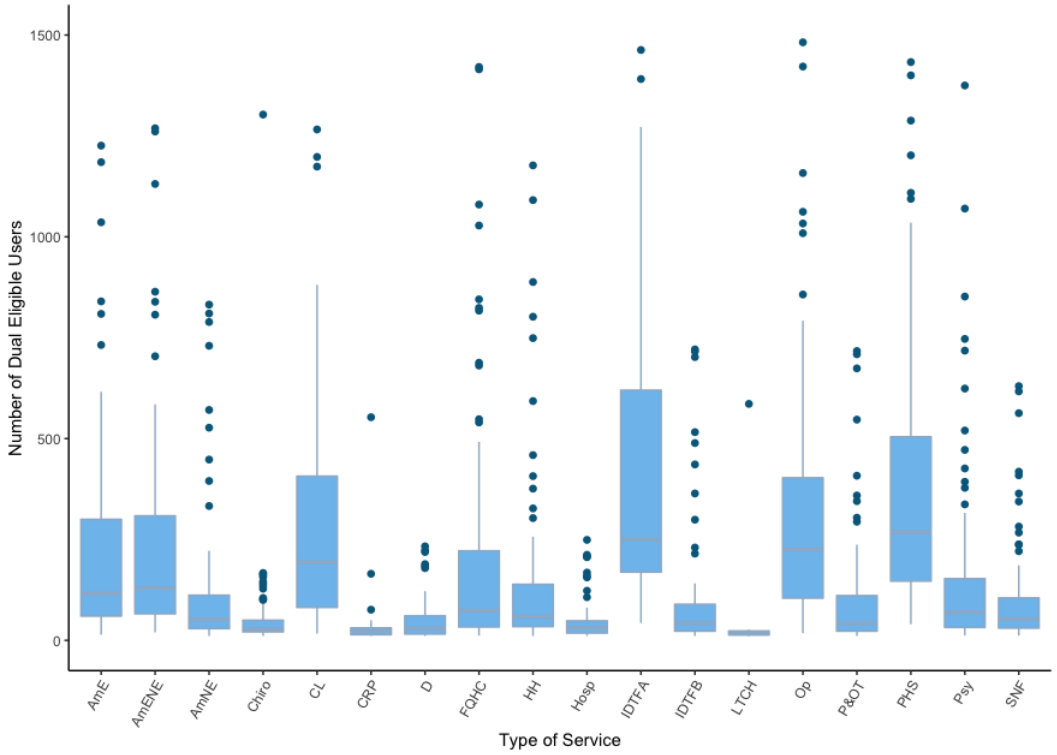

Figure 2.12: Boxplots of Number of Dual Eligible Users by Type of Service Zoomed In

The collection of boxplots in Figure 2.12 is the same as the previous graph in Figure 2.11, expect the y-axis has been rescaled to look more closely at the *Types of Services*. It appears that the Independent Diagnostic Testing Facility Pt A (IDTFA), the Preventative Health Services (PHS), the Clinical Laboratory (CL), and the Ophthalmology (Op) *Types of Service* have higher values and variability with their *Number of Dual Eligible Users*, while the Cardiac Rehabilitation Program (CRP), Chiropractic Services (Chiro), and Long-Term Care Hospitals (LTCH) have very low values and variability with respect to their *Number of Dual Eligible Users*. This pattern appears to be

continued on from the patterns in the boxplots of the *Number of Users* and the *Number of Providers*.

Significant Outlier Value (in *Number of Dual Eligible Users*)	Type of Service	County	Information about County
65,894	Preventative Health Services	Cook	IL 1st most populated County
65,148	Independent Diagnostic Testing Facility Pt A	Cook	IL 1st most populated County
56,601	Clinical Laboratory	Cook	IL 1st most populated County
35,246	Ophthalmology	Cook	IL 1st most populated County
21,557	Ambulance (Emergency & Non-Emergency)	Cook	IL 1st most populated County
19,968	Home Health	Cook	IL 1st most populated County
19,642	Ambulance (Emergency)	Cook	IL 1st most populated County

Figure 2.13: Top Seven Outliers in Number of Dual Eligible Users Variable

Based on the table above in Figure 2.13, it is clear to see that all of the top seven outliers in the *Number of Dual Eligible Users* are found in Cook *County*. This table is very nearly identical to the Figure 2.10 for the *Number of Users*, except it exchanges the *Number of Users* ranking of Preventative Health Services (PHS) in DuPage *County* with Ambulance (Emergency) services in Cook *County*. Again, Preventative Health Services (PHS) are found at the top of the list of outliers, signaling its importance to the *Users* of Americans with Medicare.

In summary, it was interesting to see compare the *Number of Dual Eligible Users* subset with the *Number of Users* and the *Number of Providers* variables. Now, the next subsection will focus on the *Total Payment* variable.

2.3.5 Variable: *Total Payment* by *Type of Service*

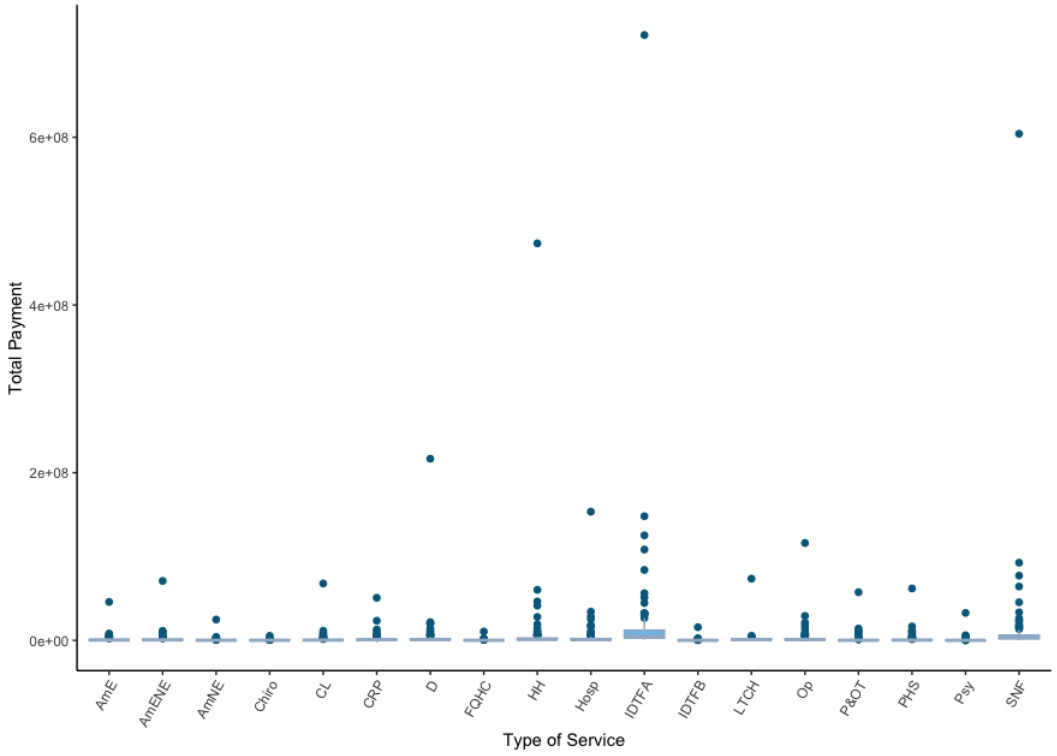

Figure 2.14: Boxplots of Total Payment by Type of Service

The boxplots shown above in Figure 2.14 display the outliers for the *Total Payment*, grouped by each of the 18 different *Types of Service*. The *Total Payment* variable displays the amount of money that the CMS has spent on various Medicare services. In this instance, the highest six outliers will be analyzed more in-depth. First, the highest outlier is at $721,982,816.3 for Independent Diagnostic Testing Facility Pt A (IDTFA) in Cook *County*. The second highest outlier is at $604,240,204 for Skilled Nursing Facilities (SNF) in Cook *County*. The third highest outlier is at $473,502,616.2 for Home Health (HH) in Cook *County*. The fourth highest outlier is at $216,703,570.3 for Dialysis (D) in Cook *County*. The fifth highest outlier is at $153,473,884.1 for Hospice (Hosp) in Cook *County*. Lastly, the sixth highest outlier is at $148,064,564.4 for Independent Diagnostic Testing Facility Pt A (IDTFA) in DuPage *County*.

To look closer at the patterns of these outliers and the shapes of the distributions, a rescaled version of the boxplots can be found below, in Figure 2.15.

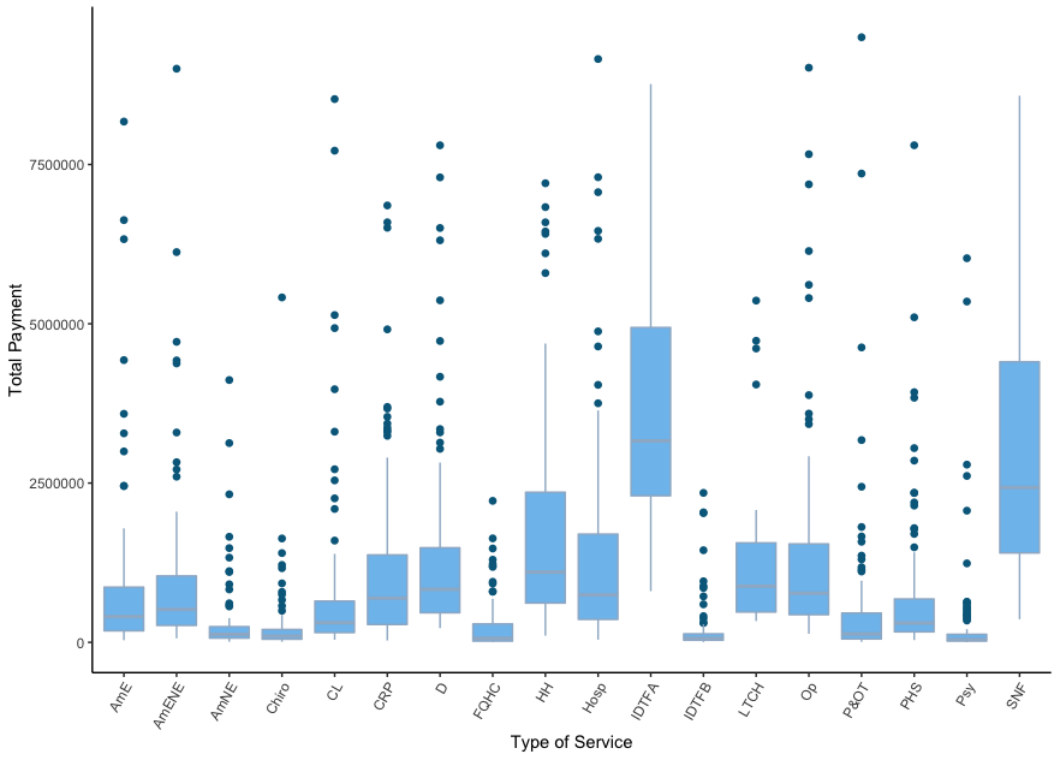

Figure 2.15: Boxplots of Total Payment by Type of Service Zoomed In

The above boxplot collection in Figure 2.15 shows the previous graph in Figure 2.14, but it is just rescaled to display the actual individual distributions of each respective boxplot. Again, it appears that Independent Diagnsotic Testing Facility (IDTFA), Skilled Nursing Facility (SNF), and Home Health (HH) *Types of Service* have significantly higher *Total Payment* values, while Psychotherapy (Psy), Independent Diagnostic Testing Facility Pt B (IDTFB), Federally Qualified Health Centers (FQHC), Ambulance (Non-Emergency), and Chiropractic Services (Chiro) have lower *Total Payment* values.

Significant Outlier Value (in *Total Payment*)	Type of Service	County	Information about County
$721,982,816.3	Independent Diagnostic Testing Facility Pt A	Cook	IL 1st most populated County
$604,240,204	Skilled Nursing Facility	Cook	IL 1st most populated County
$473,502,616.2	Home Health	Cook	IL 1st most populated County
$216,703,570.3	Dialysis	Cook	IL 1st most populated County
$153,473,884.1	Hospice	Cook	IL 1st most populated County
$148,064,564.4	Independent Diagnostic Testing Facility Pt A	DuPage	IL 2nd most populated County

Figure 2.16: Top Six Outliers in Total Payment Variable

Based on the table above in Figure 2.16 of the top six outliers in the *Total Payment* variable, it appears that Independent Diagnostic Testing Facility Pt A services seem to be more costly to Medicare providers, followed by Skilled Nursing Facilities (SNF), Home Health (HH), Dialysis (D), and Hospice (Hosp). While the Independent Diagnostic Testing Facility Pt A (IDTFA) outliers and distributions have been higher for all of the variables, it appears that potentially a combination of its large *Number of Users* and *Number of Providers* may contribute to it being costly to Medicare. The Skilled Nursing Facility (SNF), Home Health, Dialysis, and Hospice *Types of Service* have also not had any outliers of note in the past subsections, so these services must be very costly to the CMS.

In terms of the *Counties*, Cook *County* has followed the trend of having the largest outlier values, and DuPage *County* has also made its way into the list as well. Since all of the collections of boxplots have been analyzed, it is now time to move on to the outlier analysis summary section.

2.3.6 Outlier Analysis Conclusion

In summary, the outlier analysis portion of this thesis helped to uncover the statistics of the most and least utilized *Types of Service* as well as the *Counties* with the highest number of outliers. The *Types of Service* that stood out the most were Preventative Health Services (PHS), Ophthalmology (Op), Independent Diagnostic Testing Facility Pt A (IDTFA), and Clinical Laboratory (CL). This would make sense for the CMS dataset because these

are services that are very popular to the US general public, and they are necessary to many more people than *Services* like Long-Term Care Hospitals, Cardiac Rehabilitation Programs, and Dialysis.

The *Counties* with the highest number of outliers are Cook *County*, where Chicago is located, then DuPage *County*, followed by Lake *County*. These are the three most populous *Counties* in Illinois, so it would make sense that they have a higher *Number of Fee-for-Service Beneficiaries*, *Number of Providers Number of Users*, *Number of Dual Eligible Users*, and *Total Payment*. These *Counties* will be looked at more closely in the data visualization section.

Now that the outliers have been fully analyzed, it is time to move on to the data visualization of the CMS dataset.

2.4 Data Visualization

To further understand the CMS dataset, additional data visualization will be performed. Data visualization is an important technique that is used throughout the data mining process. In addition to the use of identifying outliers, data visualization helps to identify other relationships among data both prior to and after the core data mining task. It can bring the data to life in a way that allows the user to see specific patterns and relationships between variables that could not be seen by looking at raw data and outliers alone. Looking at the relationships between the variables now could also be useful to understand the results of the cluster analysis and compare findings.

With respect to this specific section, the qualitative and quantitative variables will be visualized through bar plots, charts, and scatterplots using the *ggplot2* package in *R*. Out of the many *Counties* listed in Illinois, only the top three affluent, non-affluent, highest-populated, and lowest-populated *Counties* will be looked at. The decision to look into these *Counties* was made because many of the highly populated *Counties* showed up in the outlier analysis section, and it would likely be useful to look at these *Counties* with their polar opposites: the least populated *Counties*. The *Counties* were also sorted by affluence and non-affluence as well because more affluent areas tend to provide more services, while non-affluent areas tend to lack them. Thus, both sides of the Illinois population and affluence spectrums will be represented in this analysis. These *Counties* will be analyzed with bar charts as well as correlation charts for each variable within the *Counties*.

The most affluent *Counties* in Illinois are DuPage *County*, Lake *County*, and McHenry *County* ("Here Are The 10 Richest Counties In Illinois" 1). It is important to note that the *County* where Chicago is located, Cook *County*,

is not included in this category. Conversely, the least affluent *Counties* in Illinois are Alexander *County*, Brown *County*, and Johnson *County* ("List of Illinois Locations by per Capita Income" 1). The most populated *Counties* in Illinois are Cook *County*, DuPage *County*, and Lake *County* ("Illinois Counties by Population" 1). DuPage *County* and Lake *County* are already present in the most affluent *Counties* list, but it is still useful to compare them to Cook *County*. Lastly, the least populated *Counties* in Illinois are Calhoun *County*, Hardin *County*, and Pope *County* ("Illinois Counties by Population" 1).

The quantitative variables will be compared primarily using scatterplots. These scatterplots are designed to compare three different variables, and the scatterplots can be useful to determine if there appear to be any pre-existing correlations between the variables. Correlation does not imply causation; however, understanding a potential relationship between the variables could help to connect the clustering groups after they have been created.

2.4.1 Comparison of Most Affluent, Least Affluent, Most Populated, and Least Populated *Counties' Types of Service*

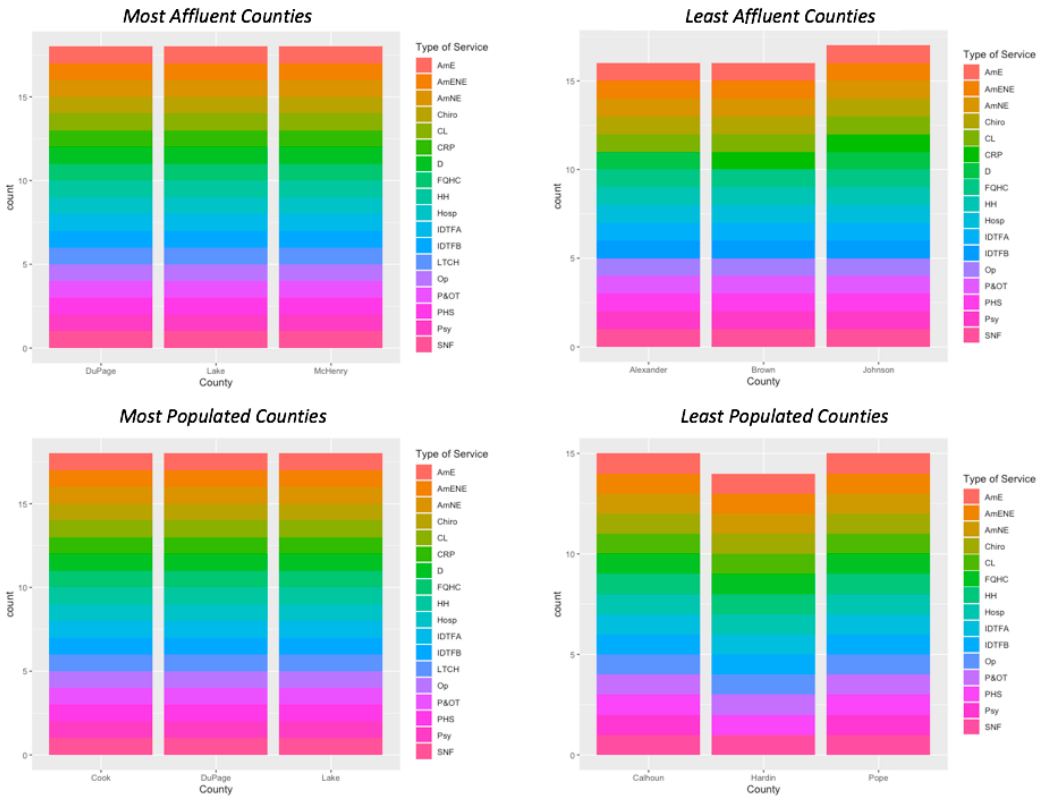

Figure 2.17: Comparing Counties' Types of Services Offered

The bar plots included above in Figure 2.17 serve as a comparative graph of the different *Counties* included in the data visualization section. Each *Type of Service* for each plot is given a specific color code, and the count of those *Types of Service* in each *County* are displayed on the plot. Judging from the graphs above, on the left-hand side, it is clear that the *Counties* that are more populated and more affluent offer all 18 different *Types of Service*. This would intuitively make sense because both population and affluence are associated with the availability of more Medicare resources; the places that have the most people or wealth are the logical places to put *Services* like Long-Term Care Hospitals.

Looking to the right-hand side of Figure 2.17, it is evident that the least populated and least affluent *Counties* in Illinois do not offer all Medicare

Types of Service. In the "Least Affluent Counties" bar plot in the upper right-hand side of Figure 2.17, Johnson *County* offers the most Medicare *Services* at 17 *Services*; it is lacking in Long-Term Care Hospitals. Alexander *County* and Boone *County* only offer 16 Medicare *Services*, both lacking Long-Term Care Hospitals and Ambulance (Emergency) *Services*.

Looking to the lower right-hand side of Figure 2.17 at the "Least Populated Counties" bar plot, these three *Counties* really seem to lack Medicare *Services* compared to the rest of the plots. Calhoun *County* and Pope *County* are lacking in Cardiac Rehabilitation Programs, Long-Term Care Hospitals, and Dialysis *Services*, offering only 15 of the 18 Medicare *Services*. Hardin *County* offers only 14 of the 18 Medicare *Services*, lacking in Cardiac Rehabilitation Programs, Long-Term Care Hospitals, Dialysis, and Psychotherapy *Services*.

Overall, looking at these four bar plots, it is apparent that the population of a *County* might have a bigger impact on whether or not it offers all 18 different types of Medicare *Services*. Next, each of these sets of three *Counties* will be compared with each other regarding the relationships between their quantitative variables.

2.4.2 Comparison of Most Affluent Counties

This includes DuPage *County* (blue), Lake *County* (yellow), and McHenry *County* (Red).

Figure 2.18: Comparing the "Most Affluent Counties"

The chart shown above in Figure 2.18 displays the relationships between the various *Counties* in the left-most column and the uppermost row. Essentially, these two sections both display the same information, just in two different manners. For this section and the rest of the *County* comparison charts in the data visualization section, the top row will be analyzed with respect the *Counties*. The rest of the chart displays the relationships between the different variables in the dataset. It should also be noted that DuPage *County* is most affluent, Lake *County* is in the middle, and McHenry *County* is the least affluent of the "Most Affluent Counties."

Looking into the relationships between the *Counties*, the upper-left-most *County* x *County* panel displays the number of the *Types of Service* offered in each *County*, which was highlighted in the previous chart of Figure 2.17. All 18 *Types of Service* are offered in these three *Counties*.

The second left-most panel in the top row displays the relationship between the *Number of Fee-for-Service Beneficiaries* in each respective *County*. There is only one value for the *Number of Fee-for-Service Beneficiaries* in

each *County*, which explains its strange appearance. In DuPage *County*, the most affluent *County*, there are a little over 120,000 *Fee-for-Service Beneficiaries*. In Lake *County*, there are just under 100,000 *Fee-for-Service Beneficiaries*, and in McHenry *County*, there are a little over 40,000 *Fee-for-Service Beneficiaries*. Thus, it appears that in the instance of these three *Counties* in Illinois, the more affluent the *County*, the higher the *Number of Fee-for-Service Beneficiaries*. This does not imply causation, but is simply stating the fact found with these three specific *Counties*.

The next panel to analyze is the third left-most panel in the top row, which compares the *Counties* with the *Number of Providers*. These data are provided through boxplots, much like the outlier analysis process. DuPage *County* has the highest *Number of Providers* in this subset of *Counties*, as well as the highest values of outliers. Lake *County* is in the middle of the three, and McHenry *County* is significantly lower in value of *Number of Providers* compared to DuPage *County* and Lake *County*.

The next panel to look into is the third right-most column in the top row of the chart, which compares the *Counties* with the *Number of Users*. These distributions are also displayed using boxplots. The distributions are very similar to the *Number of Providers* distributions, but the *Number of Users* for each *County* are higher than those of the *Number of Providers* in each *County*. The ordering of the *Counties* from most to least *Users* is DuPage *County*, then Lake *County*, and then *McHenry* County.

The next panel to analyze is the second right-most column in the top row of the chart, which compares the *Counties* to the *Number of Dual Eligible Users*. Since the *Number of Dual Eligible Users* is a subset of the *Number of Users*, its distributions for each *County* follow the same pattern and ordering from most to least of the *Counties*, just on a smaller scale.

The last panel to analyze is the right-most column in the top row of the chart, which compares the *Counties* to the *Total Payment* of the CMS on Medicare services. These distributions have significantly higher values than all of the previous variables, but the ordering of the distributions still goes from DuPage *County* to Lake *County* to McHenry *County*.

Thus, looking at the comparisons of each *County* in the "Most Affluent Counties" category, it is clear that each value in each variable comparison is significantly high, even when comparing the first-most affluent *County*, DuPage *County*, to the third most-affluent *County*, McHenry *County*. Based on conclusions from these three *Counties*, it is safe to say that in this instance of these three *Counties*, the higher the affluence of the *County*, the higher the values of each distribution per each variable.

The next chart to analyze is the chart displaying the "Least Affluent Counties."

2.4.3 Comparison Least Affluent Counties

This includes Alexander *County* (blue), Brown *County* (yellow), and Johnson *County* (red).

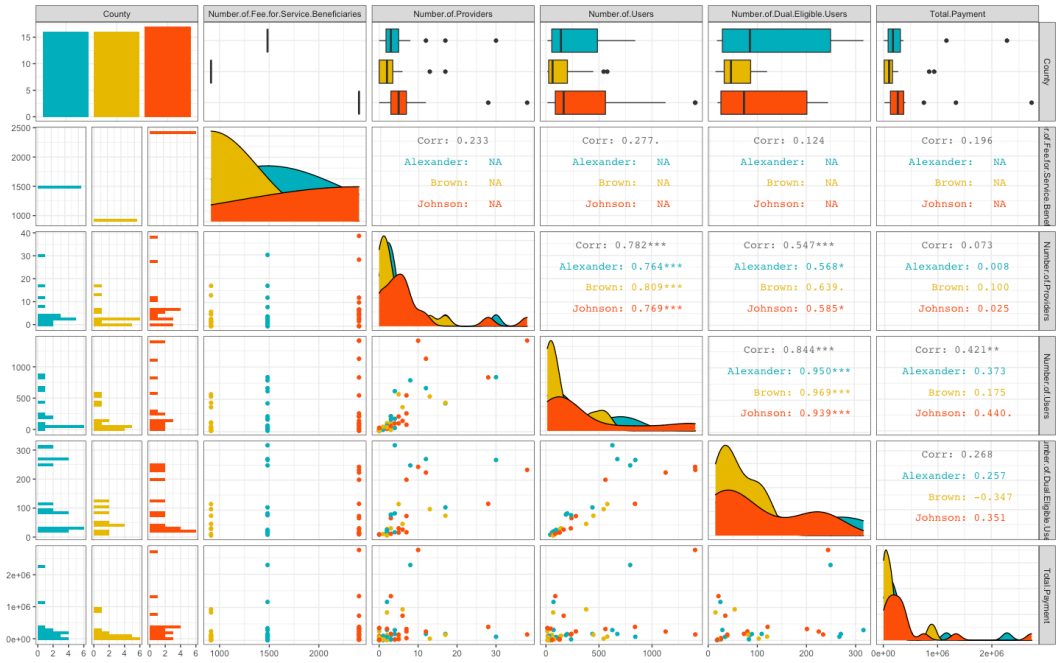

Figure 2.19: Comparing the "Least Affluent Counties"

The chart shown above in Figure 2.19 displays the relationships between the various *Counties* in the left-most column and the uppermost row. The analysis of this section will follow the same guidelines as the analysis of the previous chart in Figure 2.18. It should also be noted that Alexander *County* is least affluent, Johnson *County* is in the middle, and Brown *County* is the most affluent of the three "Least Affluent Counties."

Looking into the relationships between the *Counties*, the upper-left-most *County* x *County* panel displays the number of the *Types of Service* offered in each *County*, which was highlighted in the previous chart of Figure 2.17. Only 15 of the 18 *Types of Service* offered by Medicare are offered in Alexander *County* and Brown *County*, while 16 of the 18 *Types of Service* offered by Medicare are represented in Johnson *County*.

The second left-most panel in the top row displays the relationship between the *Number of Fee-for-Service Beneficiaries* in each respective *County*. There is only one value for the *Number of Fee-for-Service Beneficiaries* in each *County*, which explains its strange appearance. In Alexander *County*, the least affluent *County*, there about 1,500 *Fee-for-Service Beneficiaries*. In Brown *County*, there are less than 1,000 *Fee-for-Service Beneficiaries*, and in Johnson *County*, there are a just under 2,500 *Fee-for-Service Beneficiaries*. Brown *County*, the most affluent of the three "Least Affluent Counties" in Illinois, has the smallest *Number of Fee-for-Service Beneficiaries*, while Alexander *County*, the least affluent of these *Counties* is in the middle of the two other *Counties*. Based off of these three *Counties*, it will be interesting to see the other distributions.

The next panel to analyze is the third left-most panel in the top row, which compares the *Counties* with the *Number of Providers*. These data are provided through boxplots, much like the outlier analysis process. Johnson *County* has the highest *Number of Providers* in this subset of *Counties*, as well as the highest values of outliers. Alexander *County* is in the middle of the three, and Brown *County* is significantly lower in value of *Number of Providers* compared to Johnson *County* and Alexander *County*. This appears to follow the same pattern as the previous comparison of the *Counties* with the *Number of Fee-for-Service Beneficiaries*.

The next panel to look into is the third right-most column in the top row of the chart, which compares the *Counties* with the *Number of Users*. These distributions are displayed using boxplots. The distributions are very similar to the *Number of Providers* distributions, but the distributions for the *Number of Users* for each *County* are on a higher scale . The ordering of the *Counties* from most to least *Users* is Johnson *County*, then Alexander *County*, and then Brown *County*.

The next panel to analyze is the second right-most column in the top row of the chart, which compares the *Counties* to the *Number of Dual Eligible Users*. Since the *Number of Dual Eligible Users* is a subset of the *Number of Users*, its distributions for each *County* follow the same pattern and ordering from most to least of the *Counties*, just on a smaller scale.

The last panel to analyze is the right-most column in the top row of the chart, which compares the *Counties* to the *Total Payment* of the CMS on Medicare services. These distributions have significantly higher values than all of the previous variables, but the ordering of the distributions still goes from Johnson *County* to Alexander *County* to Brown *County*.

Overall, looking at the comparisons of each *County* in the "Least Affluent Counties" category, it is clear that each value in each variable comparison is

significantly low for each distribution. The clear pattern established in this section was that Johnson *County*, the second least-affluent *County* in Illinois, had the highest values in distributions, followed by Alexander *County*, the first least-affluent *County*, and then Brown *County*, the most affluent of the three *Counties*. These results differ greatly from the consensus of the "Most Affluent Counties" analysis.

The next chart to analyze is the chart displaying the "Most Populated Counties." It will be interesting to see if the distributions are similar to the "Most Affluent Counties" results.

2.4.4 Comparison of Most Populated Counties

This includes Cook *County* (blue), DuPage *County* (yellow), and Lake *County* (red).

Figure 2.20: Comparing the "Most Populated Counties"

The chart shown above in Figure 2.20 displays the relationships between the various *Counties* in the left-most column and the uppermost row. The analysis will follow the same structure as the charts in Figures 2.18 and 2.19. It should also be noted that Cook *County* is most populated, DuPage *County* is in the middle, and Lake *County* is the least populated of the "Most Populated Counties."

Looking into the relationships between the *Counties*, the upper-left-most *County* x *County* panel displays the number of the *Types of Service* offered in each *County*, which was highlighted in the previous chart of Figure 2.17. All 18 *Types of Service* are offered in these three *Counties*.

The second left-most panel in the top row displays the relationship between the *Number of Fee-for-Service Beneficiaries* in each respective *County*. There is only one value for the *Number of Fee-for-Service Beneficiaries* in each *County*, which explains its strange appearance. In Cook *County*, the most populous *County*, there are a little over 600,000 *Fee-for-Service Beneficiaries*. In DuPage *County*, there are just under 120,000 *Fee-for-Service Beneficiaries*, and in Lake *County*, there are a little over 100,000 *Fee-for-Service Beneficiaries*. Thus, it appears that in the instance of these three *Counties* in Illinois, the more populous the *County*, the higher the *Number of Fee-for-Service Beneficiaries*. This does not imply causation, but is simply stating the fact found with these three specific *Counties*.

The next panel to analyze is the third left-most panel in the top row, which compares the *Counties* with the *Number of Providers*. These data are provided through boxplots, much like the outlier analysis process. Cook *County* has the highest *Number of Providers* in this subset of *Counties*, as well as the highest values of outliers. DuPage *County* is in the middle of the three, and Lake *County* is the lowest in value of *Number of Providers* of the three *Counties*. While DuPage *County* and Lake *County* are similarly distributed, Cook *County* is significantly higher in the *Number of Providers* than both of them.

The next panel to look into is the third right-most column in the top row of the chart, which compares the *Counties* with the *Number of Users*. These distributions are also displayed using boxplots. The distributions are very similar to the *Number of Providers* distributions, but the *Number of Users* for each *County* are higher than those of the *Number of Providers* in each *County*. The ordering of the *Counties* from most to least *Users* is Cook *County*, then DuPage *County*, and then Lake *County*. The same pattern as with the *Number of Providers* occurs; Cook *County* has significantly higher values than the other two *Counties*.

The next panel to analyze is the second right-most column in the top row of the chart, which compares the *Counties* to the *Number of Dual Eligible Users*. Since the *Number of Dual Eligible Users* is a subset of the *Number of Users*, its distributions for each *County* follow the same pattern and ordering from most to least of the *Counties*, just on a smaller scale.

The last panel to analyze is the right-most column in the top row of the chart, which compares the *Counties* to the *Total Payment* of the CMS on

Medicare services. These distributions have significantly higher values than all of the previous variables, but the ordering of the distributions still goes from Cook *County* to DuPage *County* to Lake *County*, with the latter two *Counties* trailing far behind Cook *County*.

Through looking at the comparisons of each *County* in the "Most Populated Counties" category, it is clear that each value in each variable comparison is significantly high, but it is more evident that each value for each variable compared to Cook *County* is significantly greater than the values of the other two *Counties*, DuPage *County* and Lake *County*. Based on conclusions from these three *Counties*, it is safe to say that in this instance, the higher the population of the *County*, the higher the values of each distribution per each variable.

The next chart to analyze is the chart displaying the "Least Populated Counties." It will be fascinating to see whether or not this section will compare at all to the "Least Affluent Counties" section.

2.4.5 Comparison of Least Populated Counties

This includes Calhoun *County* (blue), Hardin County (*yellow*), and Pope *County* (red).

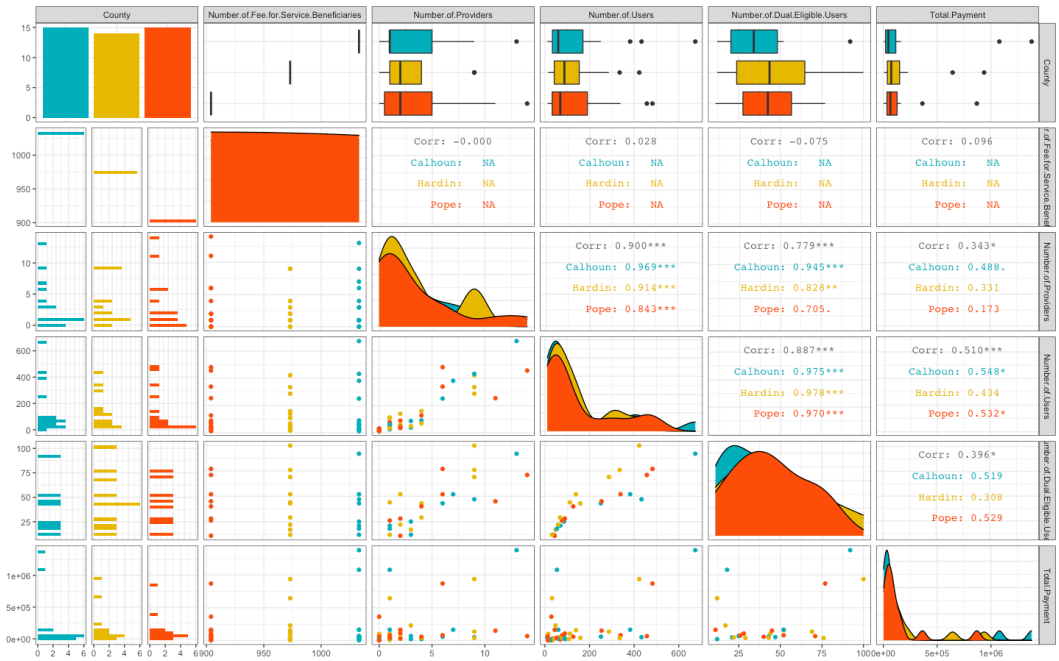

Figure 2.21: Comparing the "Least Populated Counties"

The chart shown above in Figure 2.21 displays the relationships between the various *Counties* in the left-most column and the uppermost row. This chart will be read and analyzed following the process established in Figures 2.18, 2.19, and 2.20. It should also be noted that Calhoun *County* is most populous, Pope *County* is in the middle, and Hardin *County* is the least populous of the three "Least Affluent Counties" in Illinois.

Looking into the relationships between the *Counties*, the upper-left-most *County* x *County* panel displays the number of the *Types of Service* offered in each *County*, which was highlighted in the previous chart of Figure 2.17. Calhoun *County* and Pope *County* offer only 15 of the 18 Medicare *Services*. Hardin *County*, the least populous in Illinois, offers only 14 of the 18 Medicare *Services*.

The second left-most panel in the top row displays the relationship between the *Number of Fee-for-Service Beneficiaries* in each respective *County*. There is only one value for the *Number of Fee-for-Service Beneficiaries* in each *County*, which explains its strange appearance. In Calhoun *County*, the most populous of the three *Counties*, there are a little over 1,000 *Fee-for-Service Beneficiaries*. In Hardin *County*, the least populous of the three, there are just about 975 *Fee-for-Service Beneficiaries*, and in Pope *County*, there are about 900 *Fee-for-Service Beneficiaries*. Surprisingly, the least populous *County*, Hardin *County*, does not have the smallest *Number of Fee-for-Service Beneficiaries*; Pope *County*, the middle *County*, does.

The next panel to analyze is the third left-most panel in the top row, which compares the *Counties* with the *Number of Providers*. These data are provided through boxplots, much like the outlier analysis process. Pope *County* has the highest *Number of Providers* in this subset of *Counties*, as well as the highest values of outliers. Calhoun *County* is in the middle of the three, and Hardin *County* is significantly lower in value of *Number of Providers* compared to Calhoun *County* and Pope *County*. This ordering is a little bit more predictable than the ordering of the *Number of Fee-for-Service Beneficiaries* variable.

The next panel to look into is the third right-most column in the top row of the chart, which compares the *Counties* with the *Number of Users*. These distributions are also displayed using boxplots. The distributions are very similar to the *Number of Providers* distributions, but the *Number of Users* for each *County* are higher than those of the *Number of Providers* in each *County*. The ordering of the *Counties* from most to least *Users* is Pope *County*, then Calhoun *County*, and then Hardin *County*.

The next panel to analyze is the second right-most column in the top row of the chart, which compares the *Counties* to the *Number of Dual Eligible*

Users. Despite the *Number of Dual Eligible Users* being a subset of the *Number of Users*, the distribution of these *Counties* is slightly surprising. Hardin *County* has the highest distribution of *Dual Eligible Users*, with Pope *County* and Calhoun *County* trailing far behind. While it is impossible to draw any conclusions at this stage, this ordering might make sense with respect to the variable. Hardin *County* is the least populous in Illinois, and generally, the lower the population, the lower the affluence. The *Number of Dual Eligible Users* is the number of Medicare users who are qualified for both Medicare and Medicaid, meaning that they are below the poverty line. Thus, this could technically make sense with respect to Hardin *County*.

The last panel to analyze is the right-most column in the top row of the chart, which compares the *Counties* to the *Total Payment* of the CMS on Medicare services. These distributions have significantly higher values than all of the previous variables, and the ordering of the distributions follows that of the comparison of the *Number of Dual Eligible Users* compared to the *Counties*. Thus, Hardin *County* has the highest distribution for *Total Payment*, followed by Pope *County* and then Calhoun *County*.

Overall, looking at the comparisons of each *County* in the "Least Populated Counties" category, it is clear that each value in each variable comparison is significantly low. The analysis of these three *Counties* was definitely surprising with respect to the *Number of Dual Eligible Users* and *Total Payment*, so it will be interesting to see what the cluster analysis results will be.

Now, the analysis of the subsets of *Counties* has concluded, and scatterplots of the quantitative variables will be analyzed.

2.4.6 Scatterplot 1: *Number of Fee-for-Service Beneficiaries* vs. *Number of Users* vs. *Number of Providers*

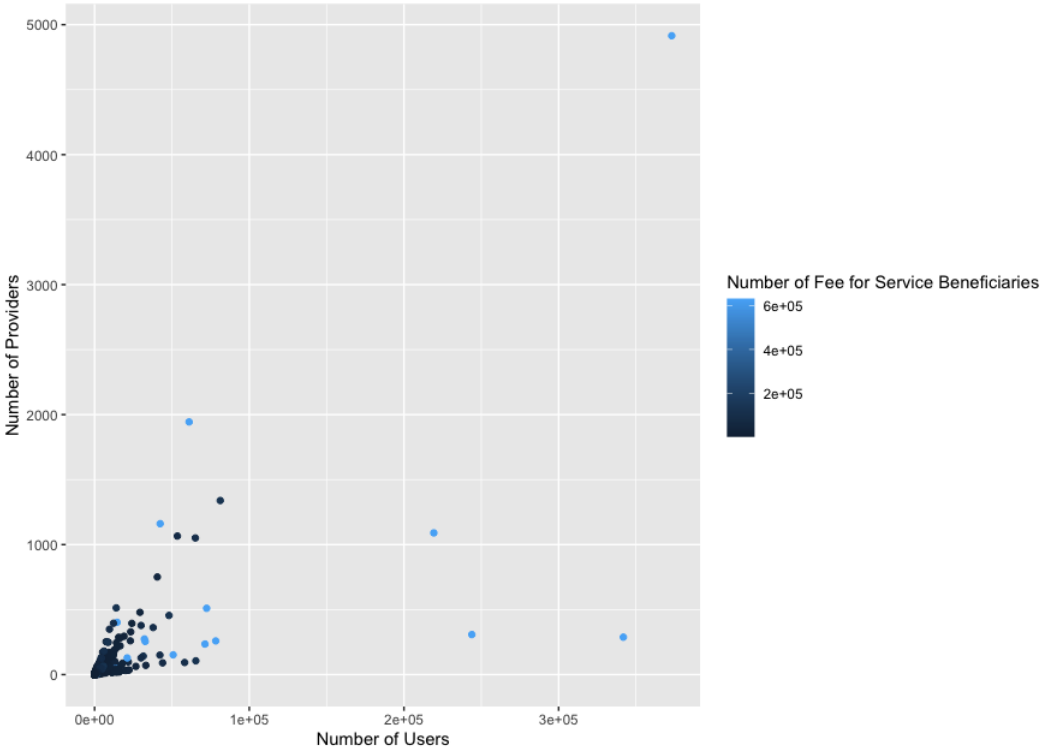

Figure 2.22: Scatterplot comparing the Number of Fee-for-Service Beneficiaries vs. Number of Users vs. Number of Providers

The scatterplot shown in Figure 2.22 compares three quantitative variables: the *Number of Fee-for-Service Beneficiaries*, the *Number of Providers*, and the *Number of Users*. The scatterplot compares the *Number of Users* (on the x-axis) with the *Number of Providers* (on the y-axis), and each corresponding point is assigned a darker blue color for a lower *Number of Fee-for-Service Beneficiaries* and a lighter blue color for a higher *Number of Fee-for-Service Beneficiaries*. There are not that many different colors in this graph because there is only one *Number of Fee-for-Service Beneficiaries* value per *County*.

At a first glance, it appears that there are some positive correlations between the three variables. They do not seem very strong, but they are clear enough to decide on this presence between the variables because higher

Numbers of Users tend to have higher *Numbers of Providers*, as well as greater *Numbers of Fee-for-Service Beneficiaries* because the points become a lighter shade of blue as both the x- and y-axes increase.

A few points, however, detract from these positive correlations. There is a small group of high *Fee-for-Service Beneficiaries* in the bottom left of the graph by the origin; they all also have relatively low *Number of Users* values and low *Number of Providers* values.

One way to look more deeply into these correlations is to analyze the different correlation coefficients, r, between the variables. The correlation coefficient between two variables can detect a linear relationship. The correlation coefficient ranges between -1 and 1, where a value closer to 1 indicates a potential stronger linear relationship, while a value closer to zero indicates a weaker one. When the correlation coefficient is positive, it indicates a positive relationship as well.

The correlation coefficient between the *Number of Providers* and the *Number of Users* is the strongest at 0.7198998. This intuitively makes sense; more *Users* generally indicate more *Providers*. This value, however, could technically be inflated to a higher value by the point to the top right of the graph, indicating a relationship that might not be as strong. The next strongest correlation coefficient is between the *Number of Fee-for-Service Beneficiaries* and the *Number of Users* at 0.6178537. This is further supported by the lighter shades of the points as the *Number of Users* on the x-axis increases. Lastly, the lowest correlation coefficient in the graph is between the *Number of Fee-for-Service Beneficiaries* and the *Number of Providers* at 0.489479. This lower value is evident by the lack of values going up the y-axis.

Thus, the initial assumption was that the more the *Number of Providers* and *Users* the more the *Number of Fee-for-Service Beneficiaries* (and vice versa). With this graph and this specific CMS dataset, this assumption appears to be mostly correct. It will be interesting to see how these values work into the cluster analysis phase of this thesis.

The final plot of the data visualization can be found in the next subsection.

2.4.7 Scatterplot 2: *Number of Providers* vs. *Numbers of Users* vs. *Total Payment*

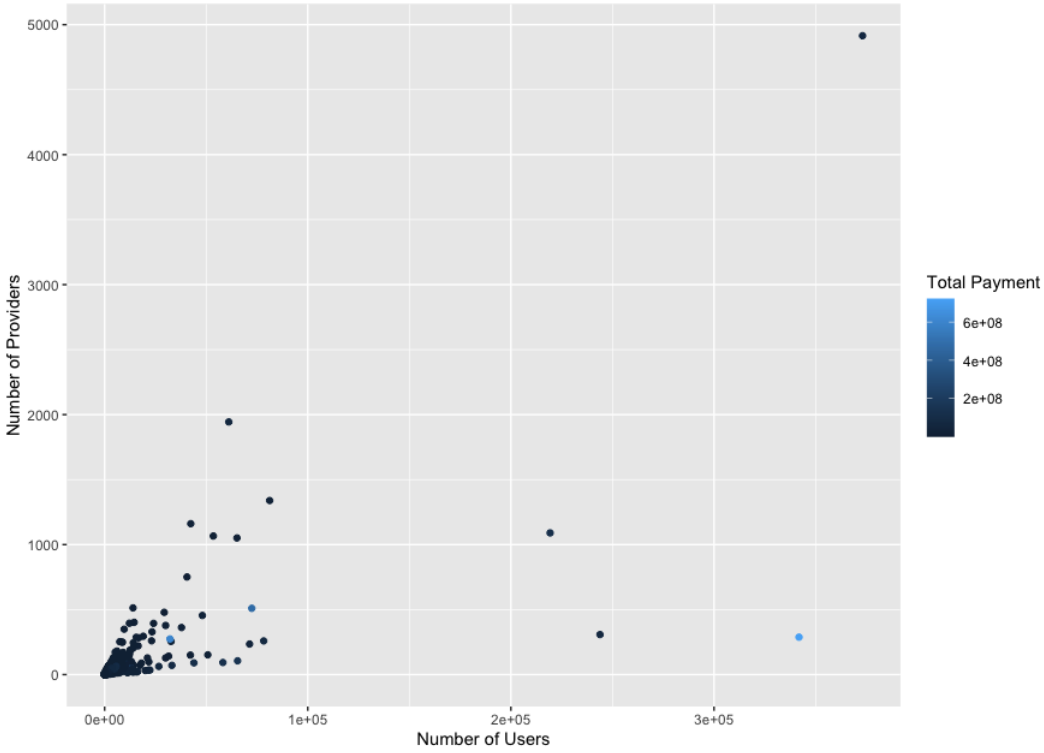

Figure 2.23: Scatterplot comparing the Number of Providers vs. Number of Users vs. Total Payment

The scatterplot shown in Figure 2.23 compares three quantitative variables: the *Total Payment*, the *Number of Providers*, and the *Number of Users*. The scatterplot compares the *Number of Users* (on the x-axis) with the *Number of Providers* (on the y-axis), and each corresponding point is assigned a darker blue color for a lower *Total Payment* and a lighter blue color for a higher *Total Payment*.

Before creating this scatterplot, it was assumed that the three variables would be positively correlated. The relationship between the *Number of Users* and the *Number of Providers* was also already highlighted in the previous scatterplot.

Looking at the graph, it appears that there are some positive correlations between the three variables. They do not appear to be outwardly very strong, but there are definitely enough patterns to say that positive relationships

between the variables are likely to exist.

Similarly to section 2.4.6, it is vital to look into the correlation coefficients between the variables. The correlation coefficient of $r = 0.7198998$ between the *Number of Providers* and the *Number of Users* was already discussed in the previous section. This is still the highest correlation coefficient among these three variables. The second strongest correlation coefficient is at 0.5519889, which portrays the relationship between the *Number of Users* and the *Total Payment* variables. This can be seen along the x-axis. There is a light-blue point, however, at the bottom right of the graph with a large *Total Payment* and *Number of Users* that could be inflating the correlation coefficient value. Thus, the real correlation coefficient could be potentially smaller than 0.5519889. The least correlated variables are the *Number of Providers* and the *Total Payment* variables, with a correlation coefficient of 0.193527. This number might be smaller than the genuine correlation coefficient because of the point to the top-right corner with a small *Total Payment* value, but a large *Number of Providers*. As a result, the genuine correlation coefficient that more accurately represents the data could be slightly larger.

In summary, it will be fascinating to see whether this positive correlation between the *Number of Users*, *Number of Providers*, and *Total Payment* play a more significant role in the clustering process.

2.4.8 Data Visualization Conclusion

In summary, the data visualization section of this thesis helped to bring the qualitative and quantitative variables to life. In this section, the "Most Affluent Counties," the "Least Affluent Counties," the "Most Populated Counties," and the "Least Populated Counties" were all analyzed. Additionally, scatterplots comparing the *Number of Providers* and *Users* with the *Total Payment* and *Number of Fee-for-Service Beneficiaries* variables were created.

With respect to the selected *Counties* in this section, generally, the more affluent *Counties* had higher values of variables than the least affluent *Counties*, and the more populated *Counties* followed the same trend. This intuitively makes sense, because when more people and more wealth are in one area, it is more likely to offer more services on a larger scale.

With respect to the individual scatterplots, the *Number of Providers*, the *Number of Users*, the *Total Payment*, and the *Number of Fee-for-Service Beneficiaries* variables are all positively and decently strongly correlated with each other. In the cluster analysis, it will be fascinating to see how these larger and smaller values are grouped.

Now, it is time to move on to the KAMILA clustering algorithm chapter of this thesis, where the KAMILA algorithm is explained and laid out more in depth.

Chapter 3

The KAMILA Clustering Algorithm

This section is comprised of introductions to the KAMILA clustering algorithm and the KAMILA clustering algorithm process.

3.1 Introduction

Now that the CMS dataset has been thoroughly pre-processed, it is time to look into the inner workings of the KAy–means for MIxed LArge datasets (KAMILA) clustering algorithm. Before actually utilizing the algorithm on the CMS dataset, it is important to delve into the structure of the algorithm itself. This way, it can be clearly understood why the KAMILA clustering algorithm is optimal for the clustering of the CMS dataset. It is imperative to understand the algorithm processes through both an application like *R* and its true mathematical notation.

3.2 The KAMILA Clustering Algorithm

3.2.1 Cluster Analysis Introduction

Cluster analysis is an unsupervised learning technique that attempts to identify unknown structures in a data set without any initial references (Foss and Markatou 2018). An unsupervised learning technique is an approach to analyze data without any clear labels to data or response variables. It differs from supervised learning, which consists of processes like *Multiple Linear Regression* and other *Classification* algorithms. Thus, the algorithm has to utilize the structure of the dataset to draw conclusions. Through this process,

the algorithm strives to sort the data into meaningful groupings based on their "natural" group within the dataset (Foss 2017). These groupings are called clusters. These clusters contain data points that are similar to the rest of the data points in the cluster, but different enough from the other points to not be in another cluster (Markatou 2018).

Cluster analysis is widely popular when used for exploratory data analysis and data summarization, especially with larger datasets (Foss, Markatou, Ray 2019). Cluster analysis "identifies both the number of groups in the data as well as the attributes of such groups" (Foss, Markatou, Ray, Heching 2016). Additionally, it helps to group the dataset in a way that observations with certain underlying similarities that cannot be seen through pre-processing steps are brought to light.

Historically, clustering algorithms have worked with all quantitative variables. Additionally, there are various popular types of clustering algorithms, such as *k-means clustering* and *hierarchical clustering*, which only work with quantitative variables.

Thus, cluster analysis is the optimal approach to uncover the patterns in the CMS dataset. Now, through the KAMILA clustering algorithm description in the next section, it will be clear as to why this specific type of clustering will be particularly of use with potential Medicare trends and groupings of the CMS dataset.

3.2.2 The KAMILA Clustering Algorithm

The KAy-means for MIxed LArge datasets (KAMILA) clustering algorithm is a scalable version of the *k-means* clustering algorithm that was specifically created by Alexander Foss and Marianthi Markatou to be applied to datasets with mixed-type data through the use of a weighted semi-parametric procedure (Foss, Markatou, Ray, Heching 2016). It surpasses the characteristic difficulties of clustering mixed quantitative and qualitative data.

Mixed-type data are data that consist of both **quantitative (continuous)** and **qualitative (categorical)** variables. Thus, the KAMILA algorithm is optimal for the CMS data set, which has two qualitative variables (the *County* and the *Type of Service*) and five quantitative variables (the *Number of Fee-for-Service Beneficiaries*, the *Number of Providers*, the *Number of Users*, the *Number of Dual Eligible Users*, and the *Total Payment*). Mixed-type data are difficult to accommodate into clustering because, "either they require strong parametric assumptions. . . , they are unable to minimize the contribution of individual variables. . . , or they require an arbitrary choice

of weights determining the relative contribution of continuous or categorical variables" (Foss, Markatou, Ray, Heching 2016). Thus, the KAMILA algorithm helps to eliminate these issues and make the clustering process more user-friendly.

In order to work with both mixed-type and large datasets, the KAMILA algorithm extends the well-known *k-means* clustering algorithm and the *Gaussian-multinomial mixture models* (Foss and Markatou 2018).

The *k-means* clustering algorithm is a prototype-based clustering algorithm in which the clusters are formed based on a prototype called a 'centroid'. The centroid is generally the mean or the median of each cluster. The *k-means* clustering algorithm utilizes a distance measure to identify the distance between each observation and centroid within each cluster. The observations are assigned to the cluster of the closest centroid (Tan, Steinbach, Kumar 2016).

Similarly to the *k-means* algorithm, the KAMILA algorithm does not make any strong parametric assumptions regarding quantitative variables. Despite this, the KAMILA algorithm is able to avoid the unbalanced treatment of quantitative and qualitative variables, based on Euclidean distance, found in the *k-means* clustering algorithm (Foss, Markatou, Ray, Heching 2016).

The *Gaussian multinomial mixture model* is a model-based clustering algorithm in which each distribution denotes a cluster. In the *Gaussian multinomial mixture model*, each distribution is a Normal distribution. Each observation is assigned to the distribution which takes highest of the probabilities that belong to one of the distributions (Malkin 2019).

Similarly to the *Gaussian-multinomial mixture models*, the KAMILA algorithm is able to balance qualitative and quantitative variables without the selection of weights, but it is based on a suitable density estimator calculated from the data itself; thus, it reduces the stricter *Gaussian* assumptions (Foss, Markatou, Ray, Heching 2016).

Additionally, there are a number of advantages of the KAMILA algorithm as a byproduct of its combination of these two algorithms. First, the variables are not changed from their original data type; this means that qualitative data are not made quantitative and vice versa. This also ensures an impartial impact on the variable types. The algorithm also avoids restrictive parametric assumptions, and the user does not have to input unique variables weights, but they can if they would like to (Foss and Markatou 2018).

In summary, the KAMILA clustering algorithm is optimal for the CMS dataset not only due to its accommodations for mixed-type data, but also for

its many beneficial improvements of the *k-means* clustering algorithm and the *Gaussian-multinomial mixture models*, as well as its focus towards being more user-friendly. Now that the KAMILA algorithm is more familiar, it is important to explore the algorithm itself in terms of its actual structure.

3.2.3 The KAMILA Clustering Algorithm Process

In this section, the KAMILA algorithm will be worked through, from the introductory model through the algorithm itself.

Creating the Model In order to build the algorithm, it is important to begin with the assumption that the dataset is made up of N independent and identically distributed observations of an $(S+T)$-dimensional vector made up of random variables $(\mathbf{V}^{\perp}, \mathbf{W}^{\perp})^{\perp}$. $(\mathbf{V}^{\perp}, \mathbf{W}^{\perp})^{\perp}$ follow a finite mixture distribution that has H components, where $\mathbf{V}$ is an S-dimensional vector consisting of quantitative random variables and $\mathbf{W}$ is a vector consisting of T qualitative random variables. For these qualitative random variables, the t-th element of $\mathbf{W}$ has L_t qualitative levels, ranging from 1, 2, ..., L_t, with $t = 1, 2, \ldots, T$. The vectors of $\mathbf{V}$ and $\mathbf{W}$ could be dependent, but under the local independence assumption, they are assumed to be independent within any specific cluster (Foss and Markatou 2018).

Since $\mathbf{V}$ is in the h-th cluster, it is modeled as "a vector following a finite mixture of elliptical distributions with individual component density functions" (Foss and Markatou 2018). These independent component density functions are $f_{\mathbf{V},h}(\mathbf{v}; M_h, \sum_h)$, where h guides cluster membership, M_h represents the h-th centroid, and $\sum_h$ is the h-th scaling matrix. Knowing that $\mathbf{W}$ is a part of the h-th cluster, it is modeled as a vector that follows a multinomial finite mixture, each with individual probability mass functions, $f_{\mathbf{W},h}(\mathbf{w}) = \prod_{t=1}^{T} p(w_t; \theta_{ht})$, where $p(\cdot\,;\cdot)$ is the multinomial probability mass function, and θ_{th} is the multinomial parameter vector for the h-th component of the t-th qualitative variable. Knowing that $(\mathbf{V}^{\perp}, \mathbf{W}^{\perp})^{\perp}$ is in the h-th cluster and under the independence assumption, the joint density of $(\mathbf{V}^{\perp}, \mathbf{W}^{\perp})^{\perp}$ is

$$f_{\mathbf{V},\mathbf{W},h}(\mathbf{v}, \mathbf{w}; M_h, \sum_h, \theta_{ht}) = f_{\mathbf{V},h}(\mathbf{v}; M_h, \sum_h) \prod_{t=1}^{T} p(w_t; \theta_{ht}),$$

and the overall density unconditional on cluster membership is given by

$$f_{\mathbf{V},\mathbf{W},h}(\mathbf{v}, \mathbf{w}) = \sum_{h=1}^{H} \pi_h f_{\mathbf{V},\mathbf{W},h}(\mathbf{v}, \mathbf{w}; M_h, \sum_h \theta_{ht}),$$

where π_h represents the previous probability of perceiving the h-th cluster (Foss and Markatou 2018).

Now that the model that the algorithm is based in has been introduced, it is time to look into the *radial kernel density estimation*, which is based on *k-means* clustering.

Estimation of the Radial Kernel Density *Kernel density Estimation* (*KDE*) is non-parametric method of estimating the density function of a random variable. The density function denotes the shape of the distribution of the random variable. *KDE* estimates the density function by applying a weight based on the distance between the data points and smoothing the shape of the curve based on a value called 'bandwidth' (Konlen 2021).

The next step in getting to the KAMILA clustering algorithm is to understand the *radial kernel density*. In this context, the vector $\mathbf{Y}$ refers to the vector $\mathbf{V}$, from the Creating the Model section, within a specific cluster. The KAMILA clustering algorithm utilizes a univariate kernel density estimation scheme, which successfully circumvents the issues of a multivariate kernel density estimator. Generally, multivariate kernel density estimation has the tendency to over-fit data points, and it is more computationally expensive than a univariate kernel density estimation scheme (Foss and Markatou 2018). Thus, it is optimal for the KAMILA algorithm to utilize a univariate kernel density estimation scheme.

Since $\mathbf{Y} \epsilon \mathbb{R}^s$ follows a spherically symmetric distribution with center M, then

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{f_Z(z)\Gamma(\frac{m}{2}+1)}{mz^{m-1}\pi^{m/2}},$$

where

$$z = \sqrt{(\mathbf{y}-M)^{\perp}(\mathbf{y}-M)}, Z = \sqrt{(\mathbf{Y}-M)^{\perp}(\mathbf{Y}-M)}.$$

Relating to the formula above, f_Z is the probability density function of Z. For the KAMILA algorithm, $\hat{f_Z}$ is constructed using a univariate kernel density estimation scheme, which is substituted into the formula above in place of f_Z (Foss and Markatou 2018). Thus, the formula for the KAMILA algorithm is

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{\hat{f_Z}(z)\Gamma(\frac{m}{2}+1)}{mz^{m-1}\pi^{m/2}}.$$

Now that radial kernel density estimators with respect to the KAMILA algorithm have been established, the actual structure of the KAMILA algorithm can be explored in the next section.

The KAMILA Clustering Algorithm The KAMILA algorithm continues on by estimating the unknown parameters of

$$f_{\mathbf{V},\mathbf{W},h}(\mathbf{v},\mathbf{w}) = \sum_{h=1}^{H} \pi_h f_{\mathbf{V},\mathbf{W},h}(\mathbf{v},\mathbf{w}; M_h, \sum_h \theta_{ht}),$$

through an iterative process. At the r-th iteration of the algorithm, let $\mathbf{M}_h^{(r)}$ represent the estimator of the centroid population, h, and let $\boldsymbol{\theta}_{ht}^{(r)}$ represent the estimator of the parameters of the multinomial distribution analogous to the t-th random variable from the population of h.

The first actual step to the KAMILA clustering algorithm is to *initialize* $\mathbf{M}_h^{(r)}$ and $\boldsymbol{\theta}_{ht}^{(r)}$ to find $\mathbf{M}_h^{(0)}$ and $\hat{\boldsymbol{\theta}}_{ht}^{(0)}$. Initializing $\mathbf{M}_h^{(0)}$ for each $h = 1, 2, \ldots, H$ is done with "random draws from the observed continuous data vectors appears to offer a modest advantage over random draws from a uniform distribution with marginal ranges equal to the sample ranges of the data" (Foss and Markatou 2018). The initialization of $\hat{\boldsymbol{\theta}}_{ht}^{(0)}$ for each h and each $t = 1, 2, \ldots, L_t$ is performed through the use of the Dirichlet distribution, with all parameters set equal to one (Foss and Markatou 2018).

Now that the $\mathbf{M}_h^{(0)}$ and $\hat{\boldsymbol{\theta}}_{ht}^{(0)}$ parameters have been *initialized*, the next aspect of the KAMILA algorithm is to repeat the *partition* and *estimation* steps until convergence is obtained. Convergence occurs when the clusters no longer change after each iteration of the algorithm. The *partition* step allocates every data observation to a cluster, and the *estimation* step consists of the re-estimation of the parameters utilizing the new cluster memberships formed during the *partition* step (Foss and Markatou 2018).

The *partition* step starts with the $\mathbf{M}_h^{(r)}$ and $\boldsymbol{\theta}_{ht}^{(r)}$ parameters at the r-th iteration. At the r-th iteration,

$$d_{ih}^{(r)} = \sqrt{\sum_{s=1}^{S} [(v_{is} - \hat{M}_{hs}^{(r)})]^2},$$

denotes the Euclidean distance between the i-th observation to each $\mathbf{M}_h^{(r)}$. The minimum Euclidean distance is then found for the i-th iteration with $z_i^{(r)} = \min_h(d_{ih}^{(r)})$. The formula,

$$\hat{f}_Z^{(r)}(z) = \frac{1}{N g^{(r)}} \sum_{\ell=1}^{N} k\left(\frac{z - z_\ell^{(r)}}{d^{(r)}}\right),$$

estimates the kernel density of the minimum distances. Thus, this is essentially a dimension reduction step; the objective is to look at a singular

univariate density estimator rather than a multivariate density estimator. It also finds the shortest distance between the i-th point and the h-th cluster. In this formula above, k represents the kernel function, and d represents the bandwidth at iteration r (Foss and Markatou 2018). This same function of $\hat{f}_Z^{(r)}$ is utilized in the construction of $\hat{f}_{\mathbf{V}}^{(r)}$, found in the Estimation of Radial Kernel Density section.

The next formula that fits into the algorithm is for $c_{ih}^{(r)}$. With the assumption that the T qualitative variables are independent within the population h, it is possible to compute the probability of observing the i-th vector of qualitative variables (given h population association) as

$$c_{ih}^{(r)} = \prod_{t=1}^{T} p(w_{ih}; \hat{\boldsymbol{\theta}}_{ht}^{(r)}),$$

where $p(\cdot\ ;\ \cdot)$ denotes the multinomial probability mass function (Foss and Markatou 2018).

The final step of the *partition* stage of the KAMILA algorithm is the assign the data objects to clusters. This formula, seen below, for $D_i^{(r)}(h)$ brings all of the other equations together to do so. During this r-th iteration, each i-th observation is assigned to the population h, which maximizes the function:

$$D_i^{(r)}(h) = log[\hat{f}_{\mathbf{V}}^{(r)}(d_{ih}^{(r)})] + log[c_{ih}^{(r)}].$$

The final stage of the KAMILA clustering algorithm is to go through the *estimation* step. The objective of the *estimation* step is to calculate new parameter estimate values for $\mathbf{M}_h^{(r)}$ and $\hat{\boldsymbol{\theta}}_{ht}^{(r)}$. Thus, these new parameters can then be used once again in the next iteration of the algorithm. If the *partition* step yields unchanged clusters from the previous iterations, then the *estimation* step can still be performed, but clusters will be finalized, and the algorithm does not have to repeat itself any further (Foss and Markatou 2018).

Starting out the *estimation* step, during each r-th iteration, the most recent *partition* of the N observations is utilized to compute $M_h^{(r+1)}$ and $\hat{\theta}_{ht}^{(r+1)}$ for all h, s, and t. In this case, $r+1$ represents the next iteration of the algorithm. If $\Omega_h^{(r)}$ represents the set of directories of observations allocated to population h at iteration r, then the parameter estimates can be computed by

$$\mathbf{M}_h^{(r+1)} = \frac{1}{\mid \Omega_h^{(r)} \mid} \sum_{i \epsilon \Omega_h^{(r)}} \mathbf{v}_i,$$

and

$$\hat{\boldsymbol{\theta}}_{ht\ell}^{(r+1)} = \frac{1}{\mid \Omega_h^{(r)} \mid} \sum_{i \epsilon \Omega_h^{(r)}} \mathbf{I}\{w_{ih} = \ell\},$$

where I represents the indicator function and the absolute value represents the cardinality of the set (Foss and Markatou 2018). The calculation of these two parameters is carried out until the clusters remain unchanged.

The kamila *R* Package and the Stopping of Iterations The model of the KAMILA clustering algorithm, the *radial kernel density estimate*, and the algorithm itself, including the *iteration*, *partition*, and *estimation* steps, have been thoroughly outlined. However, there are a few important calculations implemented into the KAMILA algorithm through the **kamila** package in *R*.

The kamila package utilizes a fairly straightforward rule when it comes to stopping a run once the clusters remain unchanged after each iteration. Since the KAMILA algorithm was designed to work with larger sized datasets, the stopping rule requires a higher amount of storage for the comparison of the cluster groupings for two consecutive iterations at a time. This can become very computationally expensive, taking a longer time for the algorithm itself to run as well. As a result of this computational expense, the creators of the algorithm implemented a stopping rule which avoids the cost by using the quantities of

$$\epsilon_{quant} = \sum_{h=1}^{H} \sum_{s=1}^{S} \mid \hat{M}_{h,s}^{(r)} - \hat{M}_{h,s}^{r-1} \mid^k,$$

and

$$\epsilon_{qual} = \sum_{h=1}^{H} \sum_{t=1}^{T} \sum_{\ell=1}^{L_t} \mid \hat{\theta}_{h,t,\ell}^{(r)} - \hat{\theta}_{h,t,\ell}^{r-1} \mid^k .$$

This rule effectively stops when both of these formulas are less than some chosen threshold, which is chosen by the user. This way, ϵ_{quant} and ϵ_{qual} are both equal to zero when the clusters remain unchanged from one iteration to the next (Foss and Markatou 2018).

In summary, now that cluster analysis, mixed-type data, and the KAMILA clustering algorithm have been explored, it is time to move on to the application of the KAMILA clustering algorithm to the CMS dataset to uncover unique clusters for Illinois Medicare data.

Chapter 4

The KAMILA Clustering Algorithm Application to the CMS Dataset

4.1 Introduction

In this chapter of the thesis, the kamila package in *R* will be applied to the CMS dataset. The settings of the kamila package will be discussed, and the CMS data will actually undergo two different cluster analysis processes. First, the CMS dataset will undergo cluster analysis with all of the Illinois data. This dataset will be referred to as the "CMS dataset including Cook *County*".

After this, due to their prominence in the outlier analysis portion of this thesis, all Cook *County* data observations will be removed from the CMS dataset. Since the outliers that pertained to Cook *County* were so much larger than the other *Counties'* outliers, they might interfere with the data output, so it is important to see how cluster analysis goes without the Cook *County* records. Additionally, Cook *County* is home to the Chicago Metropolitan Area, which has a very large population. Chicago's large population could skew the clusters to not represent all of Illinois. This will be referred to as the "CMS dataset excluding Cook *County*." This way, both datasets can be compared after the KAMILA clustering algorithm is applied to them.

4.2 The *RStudio* kamila Package

The KAMILA clustering algorithm will be applied to both CMS datasets through the use of the kamila package in *R*. The kamila package was created by Alexander Foss and Marianthi Markatou, and the package itself was published on March 13th, 2020 (Foss and Markatou 2020). The kamila package's functionality was briefly discussed in the previous chapter, but in this instance, it will be described with respect to its user inputs and arguments.

The first major step to implementing the kamila package on a dataset already read by *R* is to create two different datasets: one containing only qualitative variables and the other containing only the quantitative variables. These two datasets will be referred to as *catDF* (for the qualitative/categorical variables) and *conDF* (for the quantitative/continuous variables) throughout this thesis. Due to the nature of the kamila algorithm working with mixed-type data, it is imperative to separate the two types of variables from the start. After this, the user must utilize the *kamila* function to run the algorithm itself. Throughout this thesis, the function will be assigned the name of *kamRes*. The numerical values assigned to each argument within the function are the numerical values that will be used for both CMS datasets through the kamila package application. The *kamila* function itself looks like this:

kamRes = kamila(conDF, catDF, numClust = 2:20, numInit = 25, calcNumClust="ps", numPredStrCvRun = 10, predStrThresh = 0.5, catBw = 0.05).

It is clear that both the quantitative and qualitative datasets have been included in the *kamila* function through the *conDF* and *catDF* arguments. Additionally, the *numClust*, *numInit*, *calcNumClust*, *numPredStrCvRun*, *predStrThresh*, and *catBW* arguments are included within the kamila function. The *numClust* argument represents the number of clusters that the algorithm outputs (Foss and Markatou 2020). The kamila package is very user-friendly and offers the user myriad options, so for this thesis, the potential number of clusters has been inputted as a range of potential cluster counts. Thus, the *kamila* function can output between two and twenty clusters, as the maximum number of clusters that the package can output is twenty. By entering in a range of numbers, this guarantees that the clustering algorithm will output the true number of clusters with respect to each CMS dataset.

The next argument in the *kamila* function is *numInit*. This argument is the number of initializations used; thus, this represents the number of iterations that the algorithm will go through (Foss and Markatou 2020). The

maximum number of iterations for the algorithm is listed at 25 iterations, so this will be the number that both CMS datasets will go through. Naturally, the algorithm may not need to run all 25 iterations because the clusters may become unchanged on some iterations before then. If that is the case, it will not run all iterations.

The next argument in the *kamila* function is the *calcNumClust* argument. This argument is the method for selecting the number of clusters (Foss and Markatou 2020). In this thesis, this is set to "*ps*", which represents a *prediction strength* method. The "*ps*" method is optimal for the two CMS datasets because it does not output an overwhelming number of clusters, but just enough to recognize natural structures within the dataset itself.

The next argument in the *kamila* function is the *numPredStrCvRun* argument. This argument is the number of runs regarding the prediction strength method. This argument can only be used when the *calcNumClust* is set to equal "*ps*", which is the case in this instance (Foss and Markatou 2020). The maximum value for this input is ten runs, so that was chosen for the application of the CMS datasets.

The next argument in the *kamila* function is the *predStrThresh*, which is the threshold for the prediction strength method. Again, this argument can only be used when the *calcNumClust* is set to equal "*ps*", which is the case in this instance (Foss and Markatou 2020). Generally, with prediction strength, the smaller the threshold, the higher the number of clusters. For this thesis, the prediction strength value should be relatively high at 0.75. This value was found through testing various values with other datasets with the *kamila* function. The value of 0.75 consistently outputted the most ideal results. With this value, the number of clusters outputted will not be too overwhelming, but should be enough to sufficiently analyze to figure out natural structures within the CMS datasets.

The final argument in the *kamila* function is *catBW*, which is set equal to 0.05 for this thesis. This argument shows the bandwidth that is used for the qualitative kernel (Foss and Markatou 2020). The standard assigned bandwidth is 0.025, but in this instance, due to the large size of the CMS datasets, it has been increased to 0.05.

After the *kamila* function is outputted in the *RStudio* console, the most important values to look into are the *nClustbestNClust* values, which output the optimal number of clusters for the specific dataset. The output then lists the prediction strength values for each cluster number, 2 through 20, as was inputted in the *kamila* function earlier. By inputting the *table(kamRes$finaMemb)* function, the user can see the size of each cluster, as well as look into the observations in each cluster.

Now that the kamila package in *R* has been thoroughly analyzed, it is time to see what each CMS dataset outputs when the kamila package is applied to the datasets.

4.3 CMS Dataset Including Cook *County* Algorithm Application

In this section, the previous commands in *R's* kamila package were implemented on the CMS dataset including Cook *County*. For this dataset, the true number of clusters using the KAMILA algorithm with the Illinois Medicare data is only two clusters. Surprisingly, *Cluster 1* has 1,734 records, while *Cluster 2* has only 3 records. Due to the very small number of clusters, as well as the small number of records in *Cluster 2*, a varying number of prediction strength values and binwidth values were applied to the *kamila* function, but they all yielded this exact same result. The prediction strength graph in Figure 4.1 shown below displays how two clusters is the optimal number of clusters because it surpasses the prediction strength threshold of 0.75.

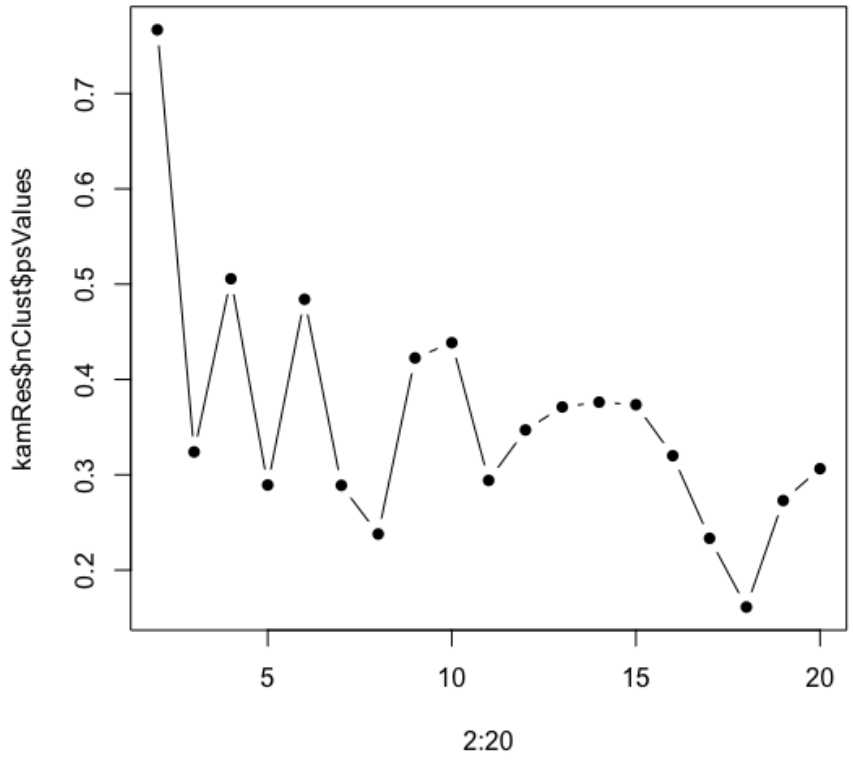

Figure 4.1: Prediction Strength Values for CMS Dataset Including Cook *County*

Now, it is time to go through the same KAMILA clustering application, but with the CMS dataset excluding Cook *County*.

4.4 CMS Dataset Excluding Cook *County* Algorithm Application

Now that the KAMILA clustering algorithm has been implemented with the CMS dataset including Cook *County*, it is time to apply the kamila package to the CMS dataset excluding Cook *County*. With the analysis of this dataset, it will be interesting to see if the number of clusters that the *kamila* function outputs is drastically different.

After implementing the *kamila* function for this dataset, the true number

of clusters is only two clusters, just like the previous CMS dataset. Similarly to the previous CMS dataset, *Cluster 1* contains a larger number of records, at 1,705 records, while *Cluster 2* has only 14 records. These clusters are definitely less drastically separated than the previous CMS dataset, but *Cluster 2* still is significantly smaller than *Cluster 1*. Due to the very small number of clusters, as well as the small number of records in *Cluster 2*, a varying number of prediction strength values and binwidth values were applied to the *kamila* function, but they all yielded this exact same result. The prediction strength graph in Figure 4.4 shown below displays how two clusters is the optimal number of clusters because it surpasses the prediction strength threshold of 0.75. The nature of these values will be more carefully looked into in the next chapter.

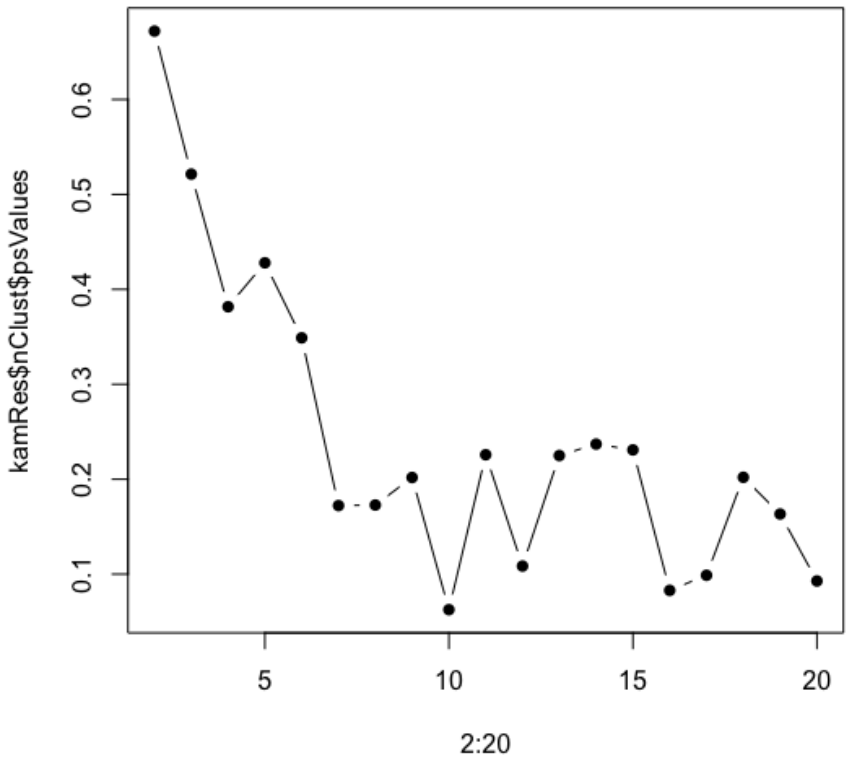

Figure 4.2: Prediction Strength Values for CMS Dataset Excluding Cook *County*

Now that both CMS datasets have gone through the KAMILA clustering algorithm, it is time to delve deeper into their structures, data, and draw potential conclusions regarding Illinois Medicare data.

Chapter 5

Post Analysis and Conclusion

In this chapter, the sets of clusters from both the CMS dataset including Cook *County* and the CMS dataset excluding Cook *County* will be further investigated, and then conclusions regarding Illinois CMS data will be discussed. For the analysis of the two different CMS datasets, the *clusters* will be visualized in the form of bar charts, tables, cluster plots, and scatter plots. Thus, this will allow each *cluster* to be brought to life and will help to uncover potential patterns within each *cluster*.

5.1 CMS Dataset Including Cook *County* Post Analysis

The first dataset that will be analyzed is the CMS dataset including Cook *County*. In the previous chapter, the KAMILA clustering algorithm outputted two different *clusters*. *Cluster 1* contains 1,734 records, while *Cluster 2* is much smaller, with only 3 records.

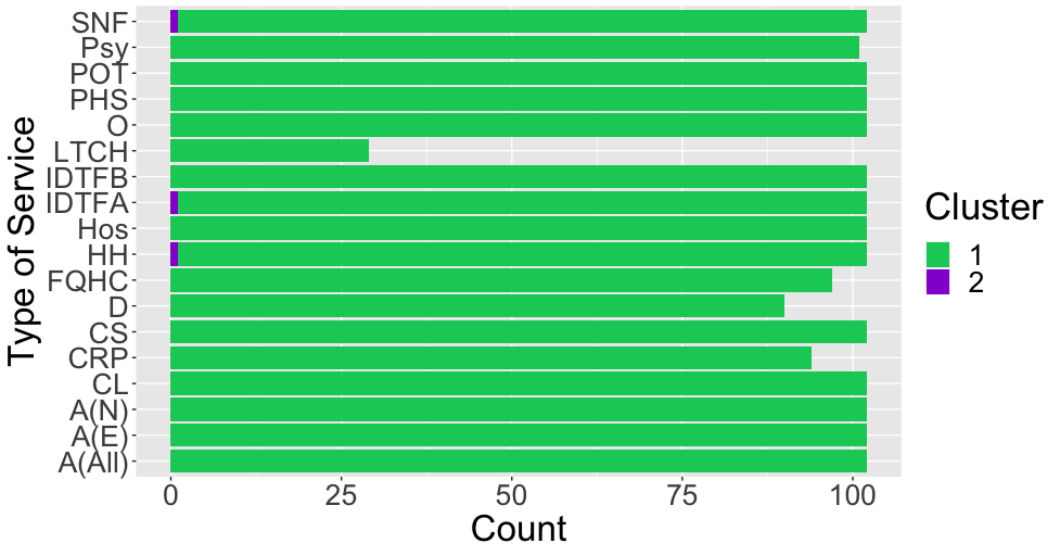

Figure 5.1: Comparative Bar Chart of *Type of Service* in *Clusters 1 and 2* for the CMS Dataset Including Cook *County*

A comparative bar chart is displayed above in Figure 5.1, which truly shows the smaller scale of *Cluster 2* to the *Cluster 1*, as well as the fact that not all *Counties* have every *Type of Service* available.

The second *Cluster* also only contains records with Cook *County*, the *County* that is home to Chicago. In terms of the *Types of Service* represented, the only *Types of Service* are Independent Diagnostic Testing Facility Part A (IDTFA), Skilled Nursing Facility (SNF), and Home Health (HH).

Quantitative Variables per Cluster: Median Values				
	Number of Providers	***Number of Users***	***Number of Dual Eligible Users***	***Total Payment***
Cluster 1	6	330	61	$ 449,456
Cluster 2	288	72,438	19,968	$ 604,240,294

Figure 5.2: Quantitative Comparison of *Clusters 1 and 2* for the CMS Dataset Including Cook *County*

In Figure 5.2, the median values of the clusters are displayed for each quantitative variable type. Median values are included rather than mean values because outliers do not affect the median, but they do affect the mean of the dataset; thus, median values more accurately represent the nature of the data. Additionally, the *Number of Fee-for-Service Beneficiaries* variable

has been left out due to the fact that it repeats for every *Counties' Type of Service*. It is evidently clear that the median values for *Cluster 2*, containing the records with the three *Types of Service* in Cook *County*, are drastically higher than those of *Cluster 1*. This could show that the algorithm found that the higher values in Chicago were the aspect that grouped them together. As a result, this exemplifies the potential importance of removing Cook *County* from the CMS dataset.

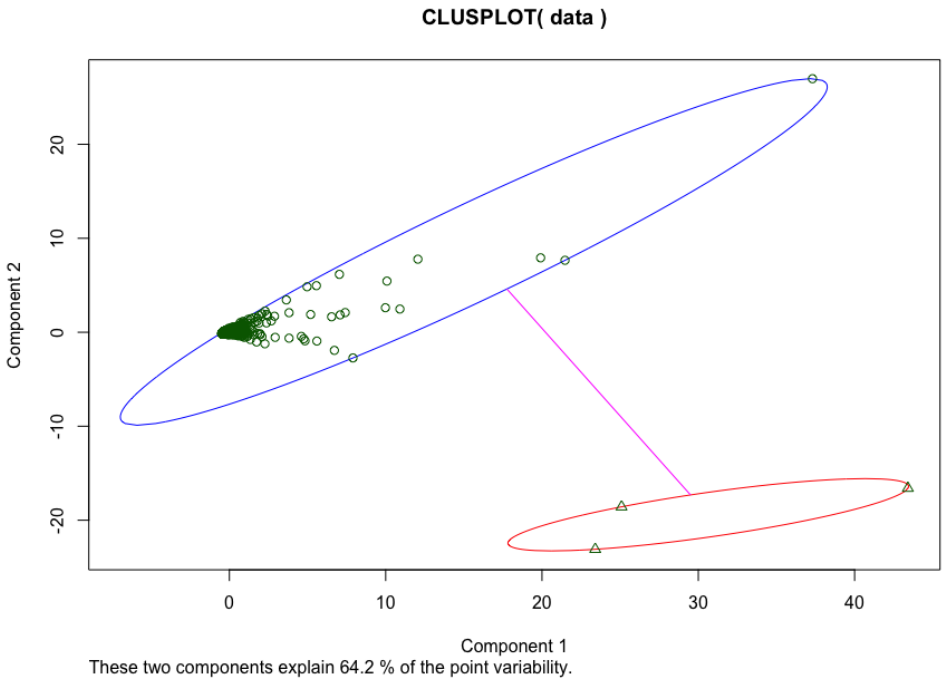

Figure 5.3: Cluster Plot for CMS Dataset Including Cook *County*

The next means of evaluating the current clusters is the utilization of a cluster plot, which can be seen in Figure 5.3. Essentially, a cluster plot is a bivariate plot that visualizes the clustering of the data. Each observation is represented by individual points on the plot. Each *cluster* is encased by an ellipse.

Looking at the cluster plot seen above in Figure 5.3, it is evident that *Cluster 1* is encased in the blue ellipse. Its 1,734 records are represented by the green circular points. The majority of the points are closest to point $(0, 0)$, but there appear to be three points that have strayed farther away from this epicenter. These must be points that did not appear to belong in any particular *cluster*, but the KAMILA algorithm eventually found a better fit with *Cluster 1* after many iterations.

Cluster 2 can be found inside of the red ellipse. The 3 records in *Cluster 2* are represented by the green triangular points. These three points do not appear to be as close to each other as those in *Cluster 1*, but they are definitely far enough away from the *Cluster 1* grouping to be considered a new *Cluster*.

Additionally, it is given in the cluster plot that these two components (*Clusters 1 and 2*) explain 64.2% of the point variability. Variability refers to the spread of the dataset, so these 2 *clusters* are the optimal *clusters* for the KAMILA algorithm because they explain or account for about 64.2% of the variability. Thus, this is the highest amount of variability that can be explained.

Now that the cluster plot of the CMS dataset including Cook *County* has been analyzed, it is time to look at scatterplots representing this data.

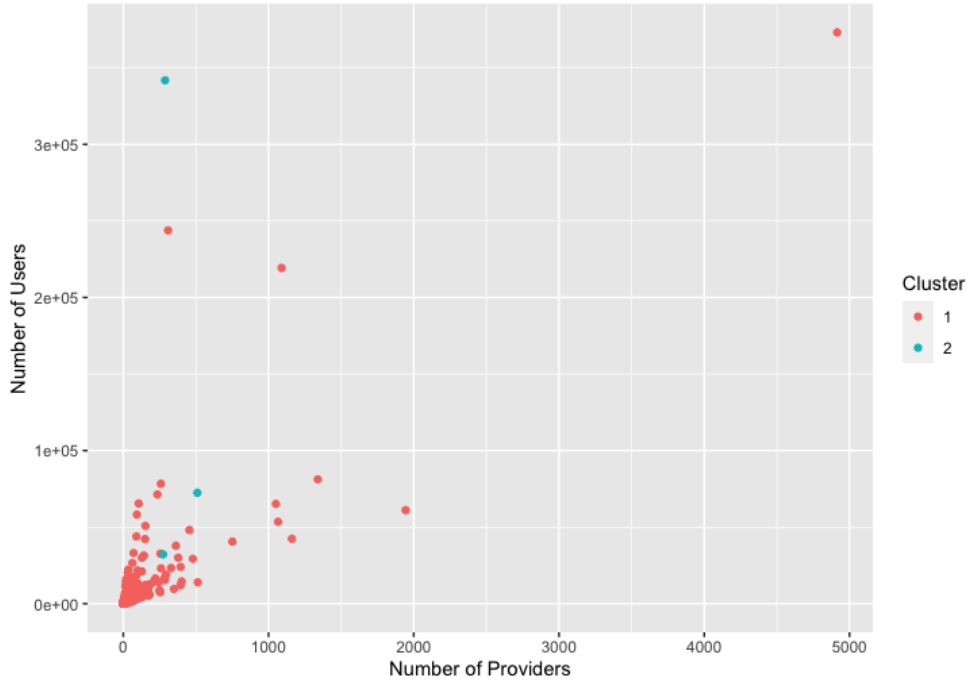

Figure 5.4: Scatterplot: *Number of Providers* vs. *Number of Users* CMS Dataset Including Cook *County*

The scatterplot seen above in Figure 5.4 compares the *Number of Providers* with the *Number of Users* with the CMS dataset including Cook *County*. Records from *Cluster 1* are represented by the red-orange data points, while records from *Cluster 2* are represented by the light blue data points. The

base of this scatterplot was already analyzed in the Data Visualization section of Chapter 2 in Figure 2.22

In the context of the two *clusters*, there does not seem to be a definitive enough pattern between the *Clusters* to draw any concrete conclusions. It does appear that in *Cluster 2*, the *Number of Users* is more variable, but the *Number of Providers* remains in the same range for each data point in *Cluster 2*. The range of the *Number of Providers* in *Cluster 2* is from 274 *providers* to 510 *Providers*, while the *Number of Users* in *Cluster 2* ranges more widely from 32,172 *users* to 341,846 *users*. Thus, it does not appear that the *clusters* were based very heavily on the *Number of Users*.

Now that the *Number of Providers* and the *Number of Users* for each *cluster* have been analyzed, it is time to look at another scatterplot comparing the *Total Payment* and the *Number of Users* by *cluster*.

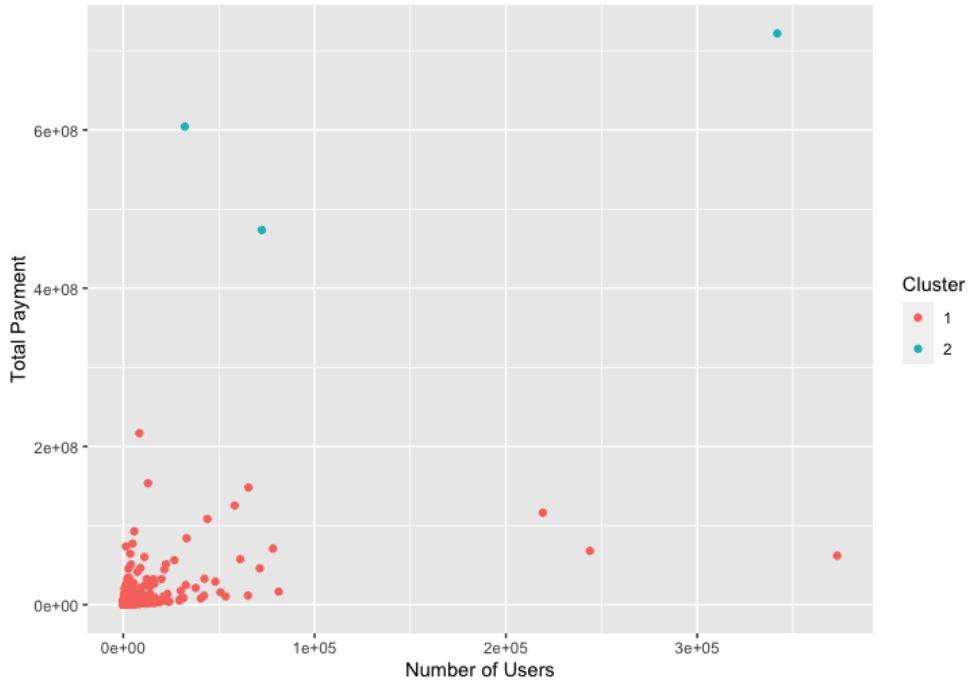

Figure 5.5: Scatterplot: *Total Payment* vs. *Number of Users* CMS Dataset Including Cook *County*

The final plot that will be analyzed with respect to the CMS dataset including Cook *County* is a scatterplot comparing the *Total Payment* and the *Number of Users* by *cluster*, shown above in Figure 5.5. Records from *Cluster 1* are represented by the red-orange data points, while records from *Cluster 2* are represented by the light blue data points.

No variation of this scatterplot has been looked into in any previous chapters, but it is particularly useful to examine with respect to the *Total Payment* variable and the second *cluster*. Looking at the scatterplot, it is evident that *Cluster 2* contains the 3 records with the highest *Total Payment* values. These 3 points' *Total Payment* values range between $473,502,616.2 and $721,982,816.3. Additionally, the *Total Payment* for these 3 records is significantly higher than the *Total Payment* values in *Cluster 1*. This could provide insight that the *Total Payment* variable was useful when determining KAMILA *clusters*.

Now that the CMS dataset including Cook *County clusters* have been analyzed, it is time to look into the CMS dataset excluding Cook *County clusters* to see if the findings follow a similar pattern to the current dataset.

5.2 CMS Dataset Excluding Cook *County* Post Analysis

The second dataset that will be analyzed is the CMS dataset excluding Cook *County*. In the previous chapter, the KAMILA clustering algorithm outputted two different *clusters*. *Cluster 1* contains 1,705 records, while *Cluster 2* is smaller, with only 14 records.

The second *Cluster* contains 3 records from DuPage *County*, 3 records from Lake *County*, 3 records from Will *County*, 2 records from Kane *County*, 1 record from McHenry *County*, 1 record from Madison *County*, and 1 record from Winnebago *County*. These counties all happen to be the second through eighth most populous *Counties* in Illinois, with DuPage *County*, Lake *County*, Will *County*, and Kane *County* being the second through fifth most populous *Counties*, respectively. Thus, despite Cook *County* and Chicago's absence from this CMS dataset, it appears that more populous *Counties* are appearing in the smaller, second *Cluster*.

Figure 5.6: Comparative Bar Chart of *Type of Service* in *Clusters 1 and 2* for the CMS Dataset Excluding Cook *County*

A comparative bar chart is displayed in Figure 5.6, which truly shows the smaller scale of *Cluster 2* to the *Cluster 1*, as well as the fact that not all *Counties* have every *Type of Service* available. With respect to the *Types of Service* represented in *Cluster 2*, the only *Types of Service* were 7 records of Independent Diagnostic Testing Facility Part A (IDTFA), 4 records of Skilled Nursing Facility (SNF), and 3 records of Home Health (HH). These are the exact same 3 *Types of Service* as there were in *Cluster 2* of the CMS dataset including Cook *County*. This could give potential insight into the importance or significance of these three *Services*.

Quantitative Variables per Cluster: Median Values				
	Number of Providers	*Number of Users*	*Number of Dual Eligibile Users*	*Total Payment*
Cluster 1	6	318	59	$ 434,751
Cluster 2	67.5	16,270	1,643	$ 62,242,628

Figure 5.7: Quantitative Comparison of *Clusters 1 and 2* for the CMS Dataset Excluding Cook *County*

In Figure 5.7 the median values of the clusters are displayed for each quantitative variable type. Median values are included rather than mean

values because they have the tendency to be less skewed than means; thus, they more accurately represent the nature of the data. Additionally, the *Number of Fee-for-Service Beneficiaries* variable has been left out due to the fact that it repeats for every *Counties' Type of Service*. It is evidently clear that the median values for *Cluster 2* are significantly higher than those of *Cluster 1*. The median values are not as drastic for *Cluster 2* excluding Cook *County*, but *Cluster 1* for both CMS datasets excluding and including Cook *County* have very similar median values.

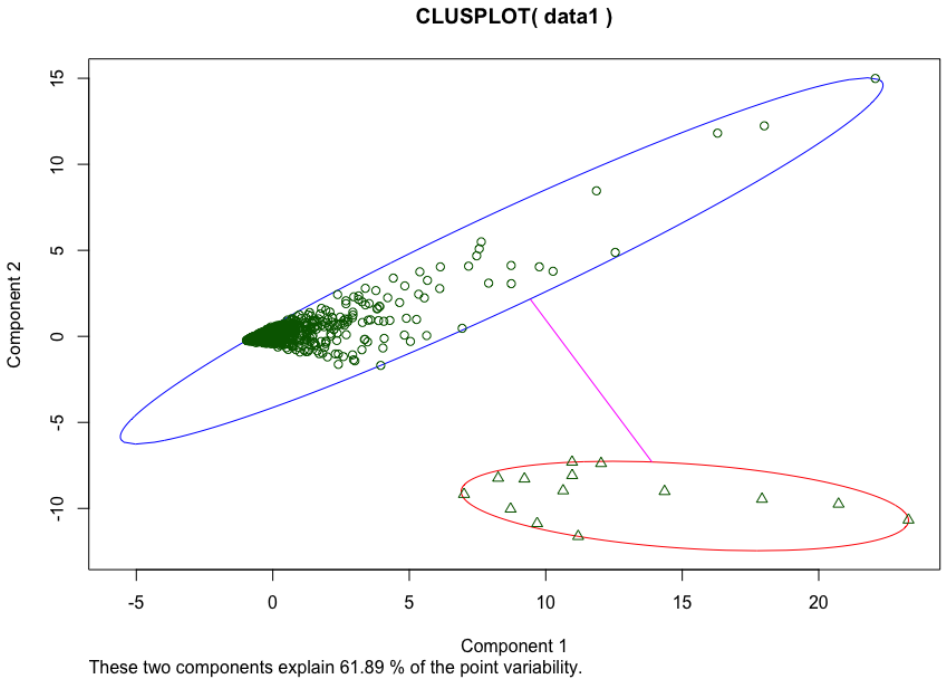

Figure 5.8: Cluster Plot for CMS Dataset Excluding Cook *County*

Now that the medians and *Types of Service* within each *cluster* have been analyzed, it is time to look into the cluster plot of the two *clusters*, shown above in Figure 5.8. Similar to the cluster plot in Figure 5.3, this cluster plot visualizes the two *clusters* and analyzes their variability. Again, each observation is represented by individual points on the plot. Each *cluster* is encased by an ellipse.

In Figure 5.8, *Cluster 1* is encased in the blue ellipse. Its 1,705 records are represented by the green circular points. Like the previous cluster plot with the *CMS* dataset including Cook *County*, the majority of the points are closest to point $(0, 0)$, but there appear to be about 5 points that have

strayed farther away from this epicenter. These must be points that did not appear to belong in any particular *cluster*, but the KAMILA algorithm eventually found a better fit with *Cluster 1* after many iterations.

Cluster 2 is located inside of the red ellipse. The 14 records in *Cluster 2* are represented by the green triangular points. About 10 of the 14 records appear to be in close proximity to each other, with the other 4 records moving further to the right in a nearly straight line. Compared to *Cluster 2* of the CMS dataset including Cook *County*, these records appear to be more closely related to each other, which is more optimal in the context of clustering.

Additionally, it is given in the cluster plot that these two components (*Clusters 1 and 2*) explain 61.89% of the point variability. Variability refers to the spread of the dataset, so these 2 *clusters* are the optimal *clusters* for the KAMILA algorithm because they explain or account for about 61.89% of the variability. Thus, this is the highest amount of variability that can be explained. With this CMS dataset, there is slightly less variability that is explained by the components than in the previous CMS dataset including Cook *County*. Due to this lower value, the result is numerically less optimal than the previous result in the previous section, but the difference is very small, so it does not hold a lot of significance.

Now that the cluster plot of the CMS dataset excluding Cook *County* has been analyzed, it is time to look at scatterplots representing this data.

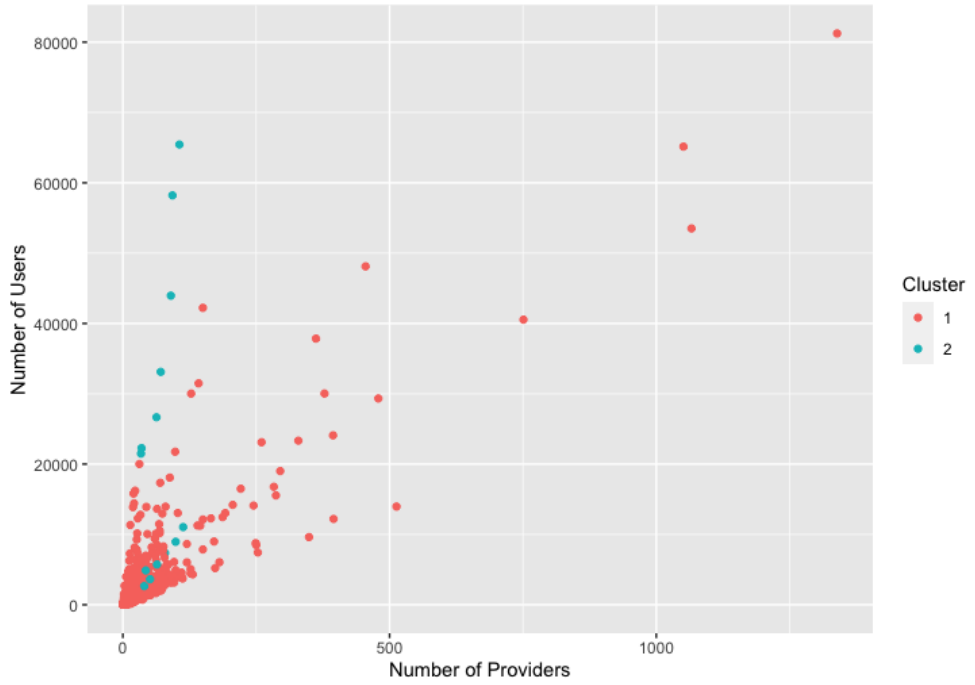

Figure 5.9: Scatterplot: *Number of Providers* vs. *Number of Users* CMS Dataset Excluding Cook *County*

The scatterplot seen above in Figure 5.9 compares the *Number of Providers* with the *Number of Users* with the CMS dataset excluding Cook *County*. Records from *Cluster 1* are represented by the red-orange data points, while records from *Cluster 2* are represented by the light blue data points. The scatterplot with the modified data excluding Cook *County* was not analyzed in a previous chapter; however, the structure of the scatterplot is very similar to the scatterplot in Figure 2.22, just on a smaller scale without Cook *County*.

In the context of the two *clusters*, there does not seem to be a definitive enough pattern between the *Clusters* to draw any concrete conclusions. It does appear that in *Cluster 2*, the *Number of Users* is more variable, but the *Number of Providers* remains in the same range for each data point in *Cluster 2*. This is similar to the findings of Figure 5.4 with the Cook *County* data. The range of the *Number of Providers* in *Cluster 2* is from 34 *providers* to 113 *Providers*. This shows that the *Number of Providers* found in *Cluster 2* were not absurdly high or out of the ordinary when compared to the *Number of Providers* in *Cluster 1*. Thus, these values may have had an impact on the KAMILA clustering algorithm, but there may be another variable that had more influence on the *clusters*.

For the *Number of Users* variable in *Cluster 2*, it ranges more widely from 2,643 *users* to 65,443 *users*. The values for the *Number of Users* had no specific pattern of note within *Cluster 2* and compared to *Cluster 1*. Thus, it does not appear that the *clusters* were based very heavily on the *Number of Users*.

Now that the *Number of Providers* and the *Number of Users* for each *cluster* have been analyzed, it is time to look at another scatterplot comparing the *Total Payment* and the *Number of Users* by *cluster*.

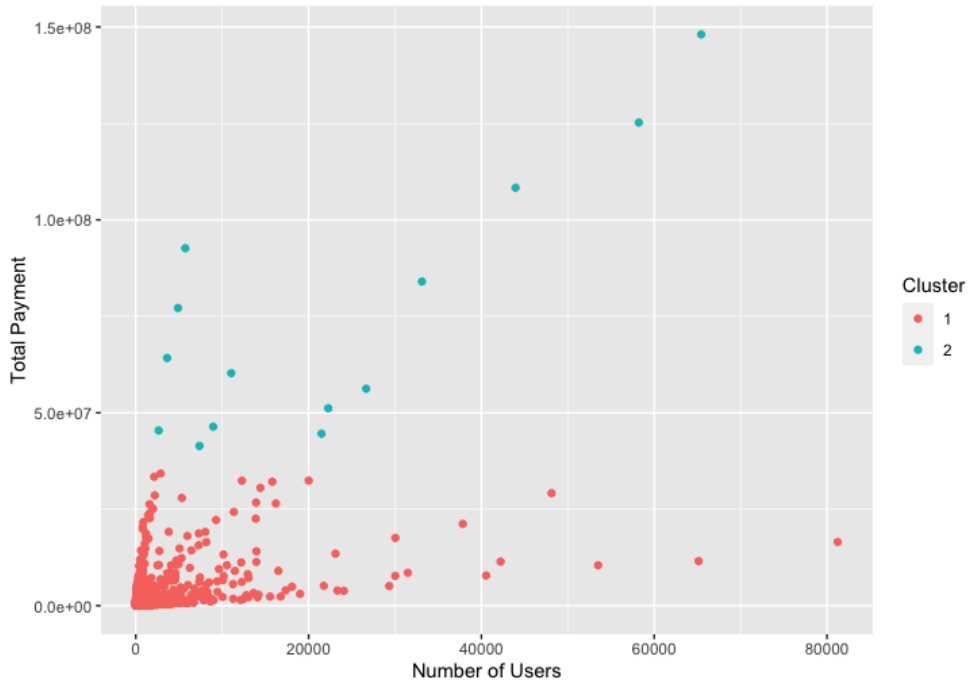

Figure 5.10: Scatterplot: *Total Payment* vs. *Number of Users* CMS Dataset Excluding Cook *County*

The final plot that will be analyzed with respect to the CMS dataset excluding Cook *County* is a scatterplot comparing the *Total Payment* and the *Number of Users* by *cluster*, shown above in Figure 5.10. Records from *Cluster 1* are represented by the red-orange data points, while records from *Cluster 2* are represented by the light blue data points.

No variation of this scatterplot has been looked into in any previous chapters, but it is particularly useful to examine with respect to the *Total Payment* variable and the second *cluster*. Looking at the scatterplot, it is evident that *Cluster 2* contains the 14 records with the highest *Total Payment* values. These 14 points' *Total Payment* values range between $41,410,638.4

and $148,064,564. Additionally, the *Total Payment* for these 14 records is significantly higher than the *Total Payment* values in *Cluster 1*. This could provide insight that the *Total Payment* variable was useful when determining KAMILA *clusters*. Additionally, this pattern is found in Figure 5.5 with respect to the CMS dataset including Cook *County*, so it is likely that the KAMILA clustering algorithm put a lot more emphasis on the *Total Payment* variable compared to the other quantitative variables.

Now that the CMS dataset excluding Cook *County clusters* have been fully analyzed, move on to the Post Analysis Summary.

5.3 Post Analysis Summary

Since the KAMILA clustering algorithm has been applied to the two different variations of the CMS dataset, the two different clusterings have ultimately yielded very similar results, both with and without the presence of Cook *County*.

With respect to the notable records within each datasets' *Cluster 2*, the smaller of the two *clusters* in both instances, there were only three different *Types of Service* present: Independent Diagnostic Testing Facility Part A (IDTFA), Skilled Nursing Facility (SNF), and Home Health (HH). The Independent Diagnostic Testing Facility Part A (IDTFA) *Type of Service* was one of the more frequently *used*, *provided*, and *paid for services* in the Outlier Analysis portion of Chapter 2, but the Skilled Nursing Facility (SNF) and Home Health (HH) variables did not frequently display high distributions for each variable. That being said, these three *Types of Service* had the highest *Total Payment* values in the boxplot collection in Figure 2.15. The records from each respective *Cluster 2* all possessed the highest *Total Payment* values in each respective dataset. This can be seen in Figures 5.5 and 5.10. Thus, it appears that the KAMILA clustering algorithm heavily emphasized the *Total Payment* variable when determining the *clusters*.

Referring to the *Counties* within each *cluster* for both datasets, the more populated *Counties* appeared to be more prevalent in each datasets' second *cluster*. Since both affluence and population were delved into more deeply in the Data Visualization portion of this thesis in Chapter 2, population appears to be more significant in regard to the KAMILA clustering algorithm, as the more populated states with the three *Types of Service* were mentioned. For example, with the CMS dataset including Cook *County*, the only *County* in *Cluster 2* is Cook, the most populated *County* in Illinois ("Illinois Counties by Population" 1). Additionally, with respect to the CMS dataset exclud-

ing Cook *County*, the *Counties* present in *Cluster 2* are DuPage *County*, Lake *County*, Will *County*, and Kane *County*, the second through fifth most populous *Counties*, respectively ("Illinois Counties by Population" 1). Had affluence been a bigger factor, then other *Types of Service* would have likely been represented within the more affluent *Counties*. In the end, however, it would appear that the *Total Payment* and *Type of Service* variable held more weight than the *Counties'* populations.

Now that the two *clusters* from each CMS dataset have been compared and further analyzed, it is time to look into the possible significance of the three *Types of Service* represented in both CMS datasets. According to the CMS, an Independent Diagnostic Testing Facility Part A (IDTFA) is "a facility that is independent both of an attending or consulting physician's office and of a hospital" ("Independent Diagnostic" 1). Thus, each Independent Diagnostic Testing Facility Part A could offer a multitude of different services. Since many people need these different services and due to an Independent Diagnostic Testing Facility Part A being an umbrella term for establishments that offer different services from each other, this could be the reason that there is such a large *Total Payment* variable with respect to this *Type of Service*.

The second most prominent *Type of Service* with respect to each datasets' *Cluster 2* is the Skilled Nursing Facility (SNF) *Type of Service*. According to the CMS, a Skilled Nursing Facility is "care like an intravenous injections that can only be given by a registered nurse or doctor" ("Skilled Nursing Facility" 1). In Illinois, the median monthly cost of a semi-private room in a Skilled Nursing Facility is $5,399 as of 2019 (Witt and Hoyt 1). This is roughly $64,788 per year. The 2019 median monthly cost of a private room in a Skilled Nursing Facility is $6,205, which amounts to roughly $74,460 per year (Wilt and Hoyt 1). It appears that the higher expenses associated with this *Type of Service* might potentially be the reason that it ended up in *Cluster 2* for both CMS datasets.

The final *Type of Service* represented in *Cluster 2* of both CMS datasets is Home Health (HH). According to the CMS, Home Health (HH) services encompasses physical therapy, occupational therapy, part-time or intermittent skilled nursing care, speech-language pathology services, medical social services, part-time or intermittent home health aide services (with personal hands-on care), and injectible osteoporosis drugs for women in the home ("Home Health Services" 1). As of 2019, the average home care hourly rate in Illinois was $22.00 per hour ("Home Care" 1). This could end up being a relatively pricey *Type of Service* over time, especially because Medicare covers up to 90% of the cost.

In summary, it appears that the assignment of each *cluster* for both the CMS dataset including Cook *County* and the CMS dataset excluding Cook *County* sorted the records with a heavy emphasis on the *Total Payment*, *Type of Service*, and *County* variables. This ultimately brings the costs as well as the population of each respective *County* into account with respect to some of the more expensive *Types of Service*.

5.4 Conclusion

Throughout this thesis, the June 2020 version of the CMS's "Market Saturation and Utilization Dataset" went through dimension reduction to include only seven variables in the state of Illinois, outlier analysis, data visualization, and eventually the KAy-means for MIxed LArge datasets (KAMILA) clustering algorithm, where it was separated into two different datasets. One dataset included Cook *County*, while the other excluded Cook *County* due to Cook *County*'s large number of outliers within the original CMS dataset. The overall objective of the thesis was to apply the relatively new KAMILA clustering algorithm, an algorithm that was created by Alexander Foss and Marianthi Markatou to apply clustering to large, mixed-type datasets, to see how the CMS datasets would be grouped together (Foss and Markatou 2018).

For both variations of the CMS dataset, two *clusters* were found within each dataset using the KAMILA algorithm. The results were mostly consistent with each CMS dataset; the second *Cluster* was significantly smaller than the first *Cluster*, with the most populous *Counties* and the most expensive *Types of Service* (Independent Diagnostic Testing Facility Part A (IDTFA), Skilled Nursing Facility (SNF), and Home Health (HH)) being represented. While the results were not exactly optimal, as more *clusters* were initially desired, this does offer insight into the workings of the KAMILA algorithm with respect to the CMS dataset. It also brought to light some of the more expensive *Types of Service* and why they were significant enough to be kept apart from each *Cluster 1*.

This thesis research could potentially be useful as a representative for Medicare services within the Midwest, as Illinois is one of the major centers of the Midwest, showing which *Types of Service* may be too expensive compared to the other *Types of Service*. In conclusion, it would be interesting to delve more deeply into this dataset with respect to other US states and regions.

Chapter 6

References and Bibliography

Anderson, Steve. "A brief history of Medicare in America." medicareresources.org, 1 Sept. 2019, www.medicareresources.org/basic-medicare-information/brief-history-of-medicare/. Accessed 18 February 2020.

Belatti, Daniel A., et al. "Total joint arthroplasty: trends in Medicare reimbursement and implant prices." The Journal of arthroplasty 29.8 (2014): 1539-1544.

Chung, Andrea P., and Alan Sorensen. "For-Profit Entry and Market Expansion in the Hospice Industry." (2018): 1-26.

"Facts About Medicare." ehealth, www.ehealthmedicare.com/about-medicare-articles/facts-about-medicare/. Accessed 18 February 2020.

Foss, Alex, Markatou, Marianthi, Ray, Bonnie, and Heching, Aliza. "A semiparametric method for clustering mixed data." Machine Learning 105.3 (2016): 419-458.

Foss, Alexander Hawthorne. Clustering Methods for Mixed-Type Data. State University of New York at Buffalo, 2017.

Foss, Alexander H., Marianthi Markatou, and Bonnie Ray. "Distance metrics and clustering methods for mixed-type data." International Statistical Review 87.1 (2019): 80-109.

Foss, Alexander H., and Marianthi Markatou. "kamila: clustering mixed-type data in R and Hadoop." Journal of Statistical Software 83.1 (2018): 1-44.

Foss, Alexander, and Marianthi Markatou. "Package 'kamila'." RStudio, 13 March 2020, https://github.com/ahfoss/kamila. Accessed 1 March 2021.

"Here Are The 10 Richest Counties In Illinois." OnlyInYourState, 30 Dec. 2015, www.onlyinyourstate.com/illinois/10-richest-counties-il/. Accessed 10 February 2020.

"Home Care Financial Assistance and Payment Options." Paying For Home Care: Financial Options, Aid and Assistance, www.payingforseniorcare.com/homecare/paying-for-home-care. Accessed 20 March 2021.

"Home Health Services." Medicare.gov the Official US Government Site for Medicare, Center for Medicare & Medicaid Services, www.medicare.gov/coverage/home-health-services. Accessed 20 March 2021.

"Illinois Counties by Population." Illinois Outline, 10 Dec. 2020, www.illinois-demographics.com/counties_by_population. Accessed 10 February 2021.

"Independent Diagnostic Testing Facility (IDTF) Fact Sheet." Cms.gov, Centers for Medicare & Medicaid Services, Aug. 2016, Independent Diagnostic Testing Facility (IDTF) - CMSwww.cms.gov ICN909060-IDTF-Fact-Sheet. Accessed 20 March 2020.

Konlen, Matthew. "Kernel Density Estimation," https://mathisonian.github.io/kde. Accessed 5 April 2021.

"List of Illinois Locations by per Capita Income." Wikipedia, Wikimedia Foundation, 5 June 2020, en.wikipedia.org/wiki/List_of_Illinois_locations_by_per_capita_income. Accessed 10 February 2020.

Malkin, C. "Gaussian Mixture Models Clustering Algorithm Explained," 2019, https://towardsdatascience.com/gaussian-mixture-modelsd13a-5e915c8e. Accessed 5 April 2021.

Markatou, Marianthi. "Clustering Mixed-Type Data." IM&A, 6-9 Nov. 2018, www.ima.umn.edu/2018-2019/SW11.7-9.18/27702. Accessed 19 February 2020.

Oxley, Alicia L. Use of CMS Star Ratings Data by Medicare Beneficiaries:

A Qualitative Exploratory Case Study. Diss. University of Phoenix, 2018: iii, 148-149.

Pita, Andrew. "Market Saturation And Utilization Dataset 2020-01-07." Data.CMS.gov, 7 Jan. 2020, https://data.cms.gov/Special-Programs-Initiatives-Program-Integrity/Market-Saturation-And-Utilization-/Dataset-2020-01-x3vv-caiy. Accessed 23 Jan. 2020.

"Preventive & Screening Services." Medicare.gov the Official US Government Site for Medicare, 2021, www.medicare.gov/coverage/preventive-screening-services. Accessed 12 February 2021.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. "Introduction to data mining." Pearson Education India, 2016. Accessed 5 April 2021.

"The Official U.S. Government Medicare Handbook: Medicare & You." Medicare.gov, Jan. 2020, Medicare and You National Handbook 2020 - Medicare.gov. Accessed 18 February 2020.

"Skilled Nursing Facility (SNF) Care." Medicare.gov the Official US Government Site for Medicare, Center for Medicare & Medicaid Services, www.medicare.gov/coverage/skilled-nursing-facility-snf-care. Accessed 20 March 2020.

"Understanding Fee-for-Service and Value-Based Care." DECO, 2021, www.decorm.com/understanding-fee-for-service-and-value-based-care/. Accessed 11 February 2021.

Witt, Scott, and Jeff Hoyt. "Skilled Nursing Costs: Average Cost of Skilled Nursing Facilities in 2021." SeniorLiving.org, Senior Living, 29 Apr. 2019, www.seniorliving.org/skilled-nursing/cost/. Accessed 20 March 2021.