NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

AN ECONOMIC CLUSTER ANALYSIS OF THE UNITED STATES

by

Tyler R. Goble

June 2022

Thesis Advisor: Ruriko Yoshida
Second Readers: Johannes O. Royset
Adam Perdue,
Texas Real Estate Research Center

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE June 2022	3. REPORT TYPE AND DATES COVERED Master's thesis

4. TITLE AND SUBTITLE
AN ECONOMIC CLUSTER ANALYSIS OF THE UNITED STATES

5. FUNDING NUMBERS

6. AUTHOR(S) Tyler R. Goble

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)

The United States is a large country that has many different areas. The cost of living combined with natural advantages for specific industries pose a difficult problem for individuals looking to find common ground across the United States at scale. Every area requires careful thought and planning by city planners relative to economic development. Past research has determined that economic clusters can be created in order to help decision makers in public office understand various economies; however, no open-source tool has been developed to aid decision makers think through public policy resolutions. Utilizing clustering models, we investigate what economic clusters form, the drivers of these clusters, and lay the ground work for more robust models. The goal of this thesis is to provide public policy decision makers with insights on other metropolitan statistical areas (MSA), encouraging further collaboration and resource sharing to aid in economic growth. Efforts were taken to keep the model simple yet robust, with the understanding that follow-on research can get much more specialized on specific issues. This thesis utilizes clustering techniques in order to determine what MSAs have similar economic outlooks. By identifying these clusters, we provide policymakers with insights on which MSA are comparable to other MSAs, shortening the research process for public policy decisions and promoting collaboration across the country.

14. SUBJECT TERMS
machine learning, generalized network autoregressive time series models, GNAR, networks, economic clusters

15. NUMBER OF PAGES
71

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

AN ECONOMIC CLUSTER ANALYSIS OF THE UNITED STATES

Tyler R. Goble
Captain, United States Marine Corps
BS, United States Naval Academy, 2016

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2022

Approved by: Ruriko Yoshida
Advisor

Johannes O. Royset
Second Reader

Adam Perdue
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The United States is a large country that has many different areas. The cost of living combined with natural advantages for specific industries pose a difficult problem for individuals looking to find common ground across the United States at scale. Every area requires careful thought and planning by city planners relative to economic development. Past research has determined that economic clusters can be created in order to help decision makers in public office understand various economies; however, no open-source tool has been developed to aid decision makers think through public policy resolutions. Utilizing clustering models, we investigate what economic clusters form, the drivers of these clusters, and lay the ground work for more robust models. The goal of this thesis is to provide public policy decision makers with insights on other metropolitan statistical areas (MSA), encouraging further collaboration and resource sharing to aid in economic growth. Efforts were taken to keep the model simple yet robust, with the understanding that follow-on research can get much more specialized on specific issues. This thesis utilizes clustering techniques in order to determine what MSAs have similar economic outlooks. By identifying these clusters, we provide policymakers with insights on which MSA are comparable to other MSAs, shortening the research process for public policy decisions and promoting collaboration across the country.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

BCD	Benchmark Cluster Definitions
BEA	Bureau of Economic Analysis
BLS	Bureau of Labor Statistics
DOD	Department of Defense
FRED	Federal Reserve Economic Data
GDP	Gross Domestic Product
IO	Input Output Clusters
MBP	Material Balance Planning
MSA	Metropolitan Statistical Areas
NPS	Naval Postgraduate School
NAICS	North American Industry Classification System
OEWS	Occupational Employment and Wage Statistics
OECD	Organisation for Economic Co-operation and Development
RSE	Relative Standard Error
TAMU	Texas A&M University Real Estate Research Center
USA	United States of America
USN	U.S. Navy

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

Public policy decision makers have many competing interests. They are typically elected by the people of their communities and have a duty to them to maintain and improve their constituents' quality of life. Looking after those interests is not always a clear cut path, however. When large and small decisions get made it is helpful to have actionable data that can be understood quickly to enable effective due diligence. This thesis sets out to help those decision makers gain insight at the beginning of any project, allowing resources to be allocated in the most efficient manner possible.

The United States is divided up into multiple regions called Metropolitan Statistical Areas (MSA). These MSA are defined as an area that "contain a city of 50,000 or more inhabitants, or contain a Census Bureau-defined urbanized area (UA) and have a total population of at least 100,000 (75,000 in New England)" (United States Census Bureau 1994). Each of these MSA have their own nuances and characteristics. This thesis looks at the similarities and difference in overall number of jobs and pay across the United States.

In this thesis, we have applied multiple clustering methods in order to tease out similarities in the different MSA across the United States. K-Means, Hierarchical, and Spectral clustering methods were used and the results were visualized on "Google My Maps."

Our methods establish a baseline for follow on research to further customize the clusters across the United States. Further research should include additional factors like housing, public works projects, and specialized industries across the United States in order to address specific questions from policy makers. This research when properly applied will save MSA hundreds of thousands of dollars in fees paid out to consulting firms (March 2015).

This thesis utilizes multiple clustering techniques in order to determine what MSA have similar economic outlooks. By identifying these clusters we provide policymakers with insights on which MSA are comparable to their MSA, shortening the research process for public policy decisions and promoting collaboration across the country.

If interested in reproducing the results of this thesis please contact Dr. Yoshida for code utilized in this work at ryoshida@nps.edu.

References

March J (2015) Analysis shows city spent $475,000 on consultants Accessed May 2, 2022, https://www.starnewsonline.com/story/news/2015/04/20/analysis-shows-cityspent-475000-on-consultants/30978948007/.

United States Census Bureau (1994) Geographic reference manual: Chapter 13 metropolitain statistical areas. Accessed April 23, 2017, https://www2.census.gov/geo/pdfs/reference/GARM/Ch13GARM.pdf.

Acknowledgments

I would like to thank my Lord and Savior Jesus Christ for the opportunities that He has given me in this life. It has been a wonderful life up to this point and I am forever grateful.

To my wife, Lindsey, thank you for your constant support. I love you so much and I can't imagine doing anything without you.

To the Marines I have had the pleasure of serving with, thank you. All of the time we got to spend together was some of the most influential times in my life.

And last but not least, to Dr. Yoshida, thank you for advising this thesis. Working with you has truly been a pleasure and the process of writing a thesis was enjoyable because of you.

Semper Fidelis.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

1.1 Problem Statement

This thesis utilizes multiple clustering techniques in order to determine what Metropolitan Statistical Areas (MSA) have similar economic outlooks. By identifying these clusters we provide policymakers with insights on which MSA are comparable to their MSA, shortening the research process for public policy decisions and promoting collaboration across the country.

Currently the United States has a multitude of geographies and cultures that interact with one another in a multitude of ways. There are many ways that these areas can be broken up, and one of the most popular ways is by designating a MSA. These areas have different factors that are tracked by the government and decision makers are constantly trying to educate themselves on the best way to improve the economies in their respective MSA.

Currently decision makers typically have a couple of options when making these decisions. First, they can make an educated guess off of their intuition. Perhaps a civic leader feels like their MSA could use additional public transportation, so they authorize the use of public funds on improving the bus system. Second, they could take a public opinion poll. The public may have desires, and since everyone acts in their own self interest, perhaps they have the best handle on how to best prioritize resources (Federal Reserve Bank of St. Louis 2012). Thirdly, public policy decision makers may choose to hire a consulting firm to analyze their market and provide them with their opinion on the best use of funds.

This thesis is looking to help augment all of these options by utilizing clustering algorithms to give insights into similarities between MSA, enabling public policy decision makers to collaborate with other similar MSA across the country. By doing this public policy will no longer have to reinvent the wheel and we will gain efficiencies for both time and money in the public sector.

The methods used in this thesis are customizable and scalable. Our hope is that this framework will be expanded upon to answer a variety of problems faced by decision makers.

1.2 Structure of This Thesis

This thesis will be broken up into five chapters. Chapter 2 will describe what cluster analysis techniques are relevant to this work, how other researchers worked on this problem set in the past, which economic factors the United States measures that are useful for our research, and how the government currently allocates American tax dollars. Chapter 3 will go into the data set used in this thesis, how we clean the data, discus our clustering methods including K-Means clustering, Spectral clustering, and our methods of visualizing the data. Chapter 4 explains the results of our work, from the different clustering methods to trends across the clusters. Finally Chapter 5 will summarize this research and give suggestions for future contributions.

CHAPTER 2:
Literature and Background

In Section 2.1 we discuss the three types of clustering methods used in this thesis. In Section 2.2 we look at previous work in the field of economic clustering. In Section 2.3 we examine various metrics that are used to describe the economy that are relevant to the factors utilized in this thesis. Finally, in Section 2.4 we explain how the government currently allocates resources.

2.1 What Is Cluster Analysis?

A cluster is defined as a group of observations that are more similar to each other in some way than they are to other observations in the data set. The main goal behind most clustering algorithms is to either minimize the distance between observations within a cluster as well as maximize the distance between clusters (Grootendorst 2021).

In this thesis, we apply three different t ypes o f c lustering a lgorithms: centroid-based, connectivity-based, and spectral-based.

2.1.1 Centroid Based

Centroid-based clustering is the algorithm which picks a centroid of each cluster and then tries to minimize the distance for each individual observation within the cluster to its centroid while at the same time it maximizes the distance between points in different clusters (Yoshida 2021).

The number of clusters is a tuning parameter, that is, the parameter which a user specifies before running the algorithm and which completely depends on an input data set. In order to determine the number of clusters we apply K-means algorithms with different numbers of clusters and select the number of clusters that reduced the sum of squared distances in clusters the most (Gao 2021).

2.1.2 Connectivity Based

Connectivity based clusters use similarities and dissimilarities to connect their observations into clusters. Hierarchical clustering is a form of clustering that is connectivity based. The algorithm links together observations based on their similarity or dissimilarity. At this point partitions are made either top down (divisive) or bottom up (agglomerative) (Viz 2022).

2.1.3 Spectral Based

Spectral-based clusters are rooted in graph theory, where the edges of graphs are utilized in order to determine clusters. This technique is different from simply calculating the distance matrices between observations and is able to handle data in a very flexible way (Fleshman 2019).

2.2 How Others Have Examined Economic Clusters

There have been multiple different ways that economic clusters have been investigated. Our baseline understanding of clustering is provided by a paper titled "Defining Clusters of Related Industries" (Delgado et al. 2016).

In Delgado et al. (2016) the authors define a cluster as a "geographic concentration of industries related by knowledge, skills, inputs, demands and/or other linkages" (pg. 1). The main driving effort behind their research was to develop a clustering method was broader than previously defined methods that could take an industry and see how heavily weighted each industry was in that respective industries cluster. Their methodology enables comparison between geographies in addition to comparison between clustering methods themselves. They also leave room for industry and subject matter expert inputs into their model, building in a factor that enables some subjectivity to be accounted for (Delgado et al. 2016).

The results of their work is the creation of a new set of U.S. Benchmark Cluster Definitions (BCD). These BCDs are a simplification from the North American Industry Classification

System (NAICS). Their BCDs have 51 clusters vice the 778 categories that the NAICS implements. It is worth noting that the Delgado et al. (2016) paper utilized the data within the 2009 NAICS within their model. This means that they are able to still maintain fidelity on the original classifications that NAICS has that are universally accepted. This allows for a more transparent application of their model. Figure 2.1 is the top level codes for the 778 categories from 2022 (United States Census Bureau 2022a).

2022 NAICS

The following table provides detailed information on the structure of NAICS.

Sector	Definition
11	Agriculture, Forestry, Fishing and Hunting
21	Mining, Quarrying, and Oil and Gas Extraction
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale Trade
44-45	Retail Trade
48-49	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate and Rental and Leasing
54	Professional, Scientific, and Technical Services
55	Management of Companies and Enterprises
56	Administrative and Support and Waste Management and Remediation Services
61	Educational Services
62	Health Care and Social Assistance
71	Arts, Entertainment, and Recreation
72	Accommodation and Food Services
81	Other Services (except Public Administration)
92	Public Administration

Figure 2.1. 2022 NAICS Table, Two Digit Codes. The codes are broken up into 20 different broad categories. From those categories more specific industries are then delineated with additional digits. The most specific classification is a six digit code. Using Sector 11: Agriculture, Forestry, Fishing and Hunting as an example, a three digit within that is 112: Animal Production and Aquaculture. Following that Section 1121 goes deeper, classifying activity as Cattle Ranching and Farming. Section 11211 is Beef Cattle Ranching and Farming, including Feedlots. Finally, this category ends with Section 112111: Beef Cattle Ranching and Farming. Notice that Feedlots are no longer included Source: United States Census Bureau (2022b).

Co-location-based clusters are another type of cluster that Delgado discusses. These are defined as "narrowly defined service and manufacturing industries to define clusters, following the principle that co-location reveals the presence of linkages across industries" (Delgado et al. 2016, pg. 6). This type of cluster differentiates between local industry (those that terminate and originate in that market) and traded industries (those that terminate regionally and cross countries). In this type of analysis natural resources like oil and gas that are physically tied to the geographic area of study are excluded due to co-linearity conditions that will arise.

Co-location-based clusters are useful in that they include the geospatial data as a part of their analysis as well as number of jobs in those industries. This has shown to be a helpful tool when quantitative analysis is being performed (Delgado et al. 2016). This thesis will not be using location data, but rather seeing if location patterns can be observed independently of geospatial inputs.

2.3 Different Measures of Economic Activity

Economists have long utilized different methods to gauge the economic health of a country. The topics most relevant to this work are Total Employment, Unemployment, Wage Growth, and Inflation. While this paper utilizes wage and employment numbers to create clusters it is important to have a baseline understanding of what these other metrics are and how they work. A great resource that summarizes all of this information can be found on the Bureau of Labor Statistics (BLS) website in a report titled "The Employment Situation." This report includes all the aforementioned metrics. The Employment Situation typically lags real time by approximately one month.

2.3.1 Total Employment

Total employment is a measure of economic health that is used and is easy to translate to different audiences. Typically this number fluctuates and the BLS has an excellent break down depicted in Figure 2.2.

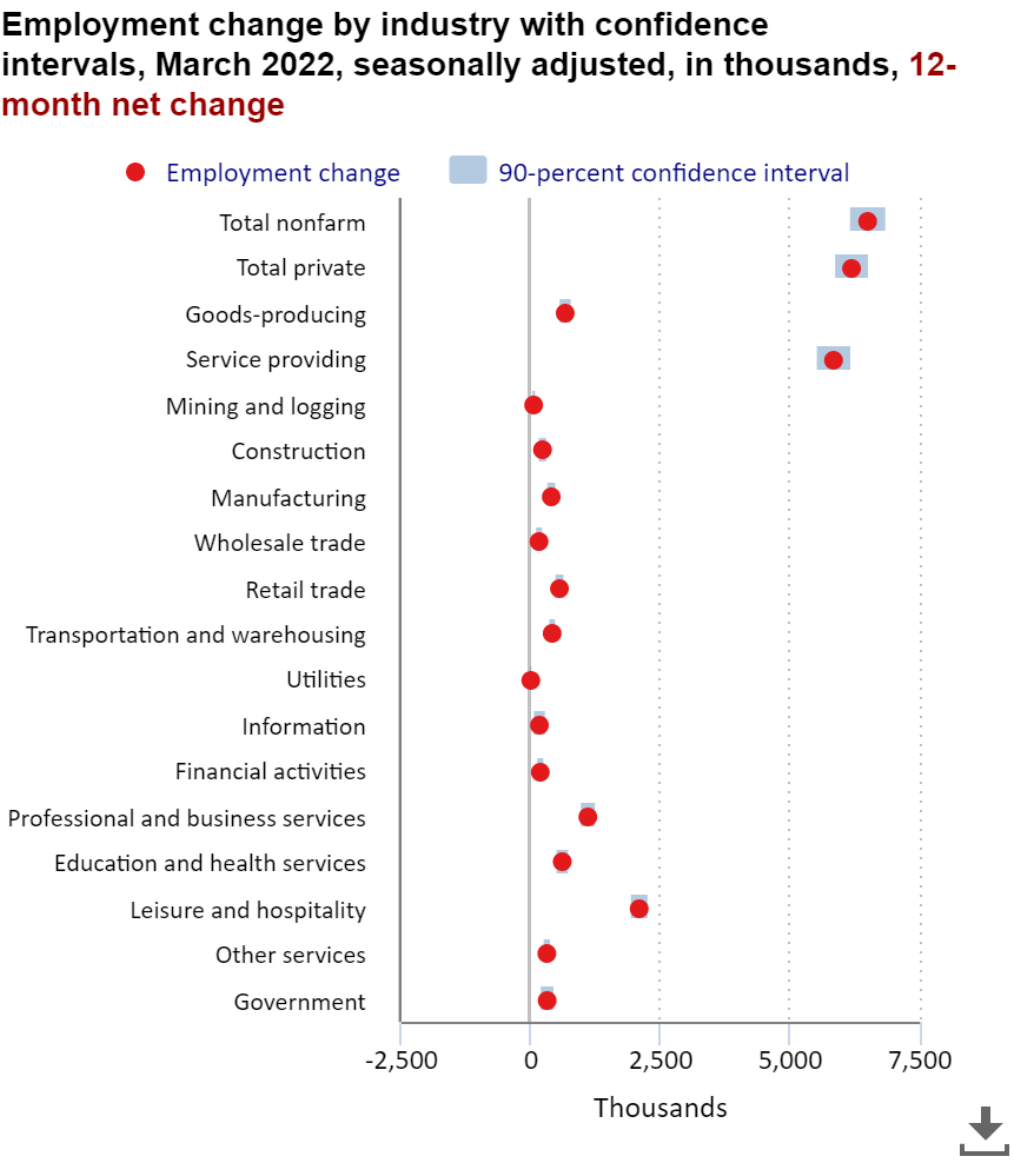

Figure 2.2. Employment Change by Industry. March 2022 12 Month Net Change. This graphic depicts the net change in jobs across the United States for the previous 12 months. There is a 90 percent confidence interval associated with each individual metric. What that confidence interval is telling us is that there is a 90 percent confidence that the true net change in employment is contained by that interval. If the confidence interval does not cross the zero line then we have a statistically significant interval. Source: United States Bureau of Labor Statistics (2022a).

Total nonfarm jobs are defined as all jobs except for government workers, private households, proprietors (individuals working for themselves without an official business entity like a Limited Liability Corporation), and employees for non-profit organizations. This makes up a large portion of the United States employment population.

Total private jobs are all jobs excepting those who are employed by the government at local, state, and national levels. These jobs include the non-profit sector. One of the major factors in this thesis is total number of jobs in an MSA.

2.3.2 Unemployment

Unemployment is very closely related to the total employment numbers, however it is telling a different story. The definition of an unemployed person is that they do not have a job, are actively seeking work within the past four weeks, and are available to take a job (United States Bureau of Labor Statistics 2015).

With a country as large and dis-aggregated as America it is incredibly difficult to get accurate statistics for unemployment. In order to get their numbers the BLS has a sample of approximately 60,000 households or 110,000 individuals. From this subset of Americans additional sampling techniques are applied, and the BLS then is able to produce unemployment data monthly for the United States. Each respondent answers a series of questions that determine whether or not they are actually unemployed. The survey does not ask them if they are unemployed, rather it asks questions like how many hours did you work last week, or if they have been doing anything to find work. By structuring the survey in this way they are able to limit outside bias. An in depth explanation of the way the BLS arrives at unemployment numbers is at the following link: https://www.bls.gov/cps/cps_htgm.htm.

2.3.3 Wage Growth

Another important indicator of economic health is the growth of real wages for workers. Similar to previous discussions on GDP and establishing a base year to control for inflation, the same technique is applied to wages. It does no good for the wage earner to make more money if their wage increase does not keep pace with inflation.

The Federal Reserve Economic Data (FRED) provides excellent data for looking at the real wages in the U.S. economy. Figure 2.3 is a graph of the real median weekly wage earnings.

Figure 2.3. Federal Reserve Economic Data: Real Wages. This chart shows the real weekly earnings for wage and salary earners 16 years old and older. Note the graph condenses the first 90 years post 1900 in order to better visualize more recent data. The grey shaded areas are recessions in the United States. It would appear that real wages have been steadily increasing since 1990, and actually hit a peak in 2020 during the pandemic before dropping back down to pre-pandemic levels. Source: Federal Reserve Economic Data (2022).

The other factors utilized in this thesis are pay levels in each MSA.

2.3.4 Inflation

Inflation is defined as the decline of purchasing power of a given currency over time. The rate at which the purchasing power is being eroded is measured by tracking the cost of a market basket of goods in the economy. Figure 2.4 is a visualization of inflation with data provided by the Organization for Economic Co-operation and Development (OECD). The red star is where the data switches from being historic to predicting the future (Jason Fernando 2022).

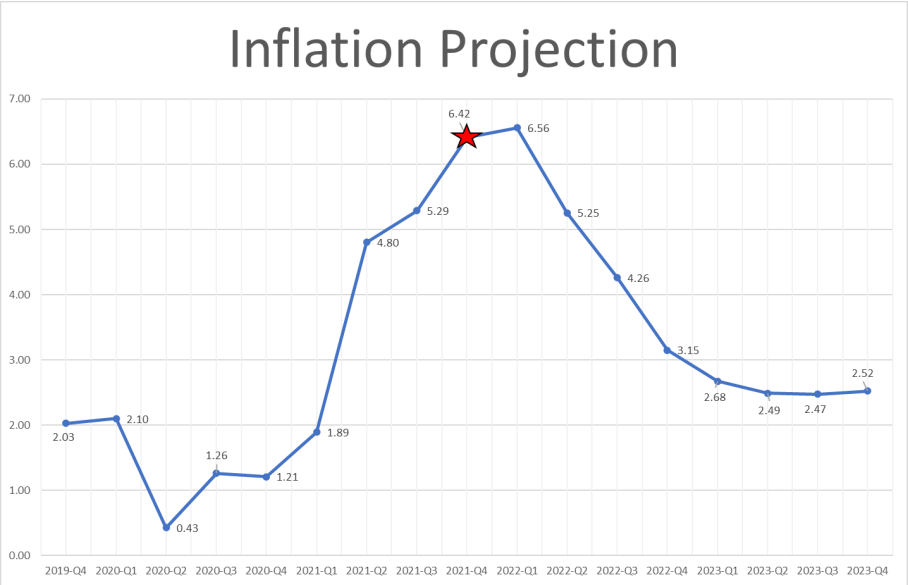

Figure 2.4. Inflation Projections. Adapted from: Trading Economics (2022). Using this data we plotted their projections for inflation in February of 2022. The red stat indicated the point at which their data turned to forecasting vice being recorded historically. At the time of this writing it appears that their forecast is significantly off, with inflation being reported at rates as high as 8.5 percent.

2.4 How Does The Government Allocate Resources?

The government has an enormous task of deciding how to best use scarce resources to best take care of their populations. In the United States, government is thought of in a three tier structure, local, state, and federal. Each level of government has a treasury department that attempts to manage the finances of their respective level in order to enable as much flourishing as possible in the economy (U.S Department of the Treasury 2022).

For organizational purposes, the United States also divides regions up in MSAs. The official Census Bureau definition for an MSA is "..requires the presence of a city of 50,000 or more inhabitants, or as Census Bureau-defined urbanized area (of at least 50,000 inhabitants) and a total population of at least 100,000 (75,000 in New England)." When multiple areas are covered within a single MSA the largest city in that MSA is referenced as the central

city. The overall MSA may be titled with up to three central cities. For example, Albany-Schenectady-Troy, NY is an MSA in New York State (United States Census Bureau 1994).

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3:
Methodology

The purpose behind this thesis is to develop a framework from which to determine what clusters exist among the various MSAs across the United States. First we will go into the original data set, then how we went about cleaning that data. Once the data was processed we will talk through the two clustering algorithms we employed, K-Means and Spectral Clustering. Finally we will discuss the way we utilized "Google My Maps" to create a user friendly visualization that is adaptable to multiple use cases.

3.1 Data Set

The data set was obtained from the BLS as a part of their Occupational Employment and Wage Statistics (OEWS) branch (United States Bureau of Labor Statistics 2022b). The data we used goes from 2015 to 2020. 2020 was the most recently published data set at the time of this thesis.

OEWS is a program that collects information for approximately 800 different occupations across the United States. The data goes from the national level, to state, to MSA, to metropolitan divisions, and nonmetropolitan areas. These segments go from large to small in respect to population. As the data scope becomes more localized, the job descriptions also become more specific. For this thesis we focused on MSA level data.

For each year, the data set consisted of approximately 390,000 rows and 31 columns. Each row was considered an observation for that specific job, within that specific geography.

Figure 3.1 is an overview of the data structure for 2020.

```
<class 'pandas.core.frame.DataFrame'>
Float64Index: 390705 entries, nan to nan
Data columns (total 31 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   AREA           390705 non-null  int64
 1   AREA_TITLE     390705 non-null  object
 2   AREA_TYPE      390705 non-null  int64
 3   PRIM_STATE     390705 non-null  object
 4   NAICS          390705 non-null  object
 5   NAICS_TITLE    390705 non-null  object
 6   I_GROUP        390705 non-null  object
 7   OWN_CODE       390705 non-null  int64
 8   OCC_CODE       390705 non-null  object
 9   OCC_TITLE      390705 non-null  object
 10  O_GROUP        390705 non-null  object
 11  TOT_EMP        390705 non-null  object
 12  EMP_PRSE       390705 non-null  object
 13  JOBS_1000      221444 non-null  object
 14  LOC_QUOTIENT   221444 non-null  object
 15  PCT_TOTAL      162652 non-null  object
 16  H_MEAN         390705 non-null  object
 17  A_MEAN         390705 non-null  object
 18  MEAN_PRSE      390705 non-null  object
 19  H_PCT10        390705 non-null  object
 20  H_PCT25        390705 non-null  object
 21  H_MEDIAN       390705 non-null  object
 22  H_PCT75        390705 non-null  object
 23  H_PCT90        390705 non-null  object
 24  A_PCT10        390705 non-null  object
 25  A_PCT25        390705 non-null  object
 26  A_MEDIAN       390705 non-null  object
 27  A_PCT75        390705 non-null  object
 28  A_PCT90        390705 non-null  object
 29  ANNUAL         15420 non-null   object
 30  HOURLY         727 non-null     object
dtypes: int64(3), object(28)
memory usage: 95.4+ MB
```

Figure 3.1. Overall Data Structure OEWS 2020. The data is relatively user friendly to access and is made publicly available on the BLS website. Note that in its raw form the data is not read in as integers or floating point numbers, therefore some reprocessing must occur prior to analysis. Source: United States Bureau of Labor Statistics (2022b).

The BLS website provides definitions of the columns here:

1. Occupation Title: a descriptive title that corresponds to the SOC code.
2. Group: the level of occupational detail (major group, minor group, broad level, detailed).
3. Employment: the estimated total occupational employment (not including self-

employed).

4. Employment Relative Standard Error (RSE): the RSE of the employment estimate, a measure of the reliability or precision of the employment estimate. The RSE is defined as the ratio of the standard error to the survey estimate. For example, a RSE of 10 percent implies that the standard error is one-tenth as large as the survey estimate.
5. Employment per 1000 jobs: the number of jobs (employment) in the given occupation per 1,000 jobs in the given area.
6. Location Quotient: (State, metropolitan, and nonmetropolitan statistical area estimates only) the ratio of an occupation's share of employment in a given area to that occupation's share of employment in the U.S. as a whole. For example, an occupation that makes up 10 percent of employment in a specific metropolitan area compared with 2 percent of U.S. employment would have a location quotient of 5 for the area in question.
7. Median Hourly Wage: the estimated 50th percentile of the distribution of wages based on data collected from employers in all industries; 50 percent of workers in an occupation earn less than the median wage, and 50 percent earn more than the median wage.
8. Mean Hourly Wage: the estimated total hourly wages of an occupation divided by its estimated employment, i.e., the average hourly wage.
9. Mean Annual Wage: the estimated total annual wages of an occupation divided by its estimated employment, i.e., the average annual wage.
10. Mean RSE: the relative standard error of the mean wage estimates, a measure of the reliability or precision of the mean wage estimates. The relative standard error is defined as the ratio of the standard error to the survey estimate. For example, a relative standard error of 10 percent implies that the standard error is one-tenth as large as the survey estimate.
11. Percentile Wage Estimates: (National estimates only) A percentile wage estimate shows what percentage of workers in an occupation earn less than a given wage and what percentage earn more. For example, a 25th percentile wage of 15.00 indicates that 25 percent of workers (in a given occupation in a given area) earn less than 15.00; therefore 75 percent of workers earn more than 15.00. Note: This percentile wage estimate has since been expanded down to the lowest level in the data set (United States Bureau of Labor Statistics 2022c).

As an example we will look at the Sales Manager occupation in Waco, Texas.

In the case of Sales Managers:

1. Description: Plan, direct, or coordinate the actual distribution or movement of a product or service to the customer. Coordinate sales distribution by establishing sales territories, quotas, and goals and establish training programs for sales representatives. Analyze sales statistics gathered by staff to determine sales potential and inventory requirements and monitor the preferences of customers.
2. Total Employment: 160
3. Jobs per 1000: 1.354
4. Location Quotient: 0.48
5. Median Yearly Wage: 119,010 (United States Bureau of Labor Statistics 2022b).

3.2 Data Cleaning

Once we got familiar with the data set we began the cleaning process. Initially we utilized Python in order to wrangle the data and pair it down to a usable format. After pulling down the data for 2015 to 2020 into our local directory we read each data set into its own data frame utilizing the Pandas package. Upon inspection transformations to the data types were required in order to conduct analysis. Each numeric column needed to be converted into numeric data from the original string format that was given by OEWS.

After converting the data into the proper data type, we then filtered the data down to the MSA level data utilizing various Pandas commands. Once the data was at the MSA level we then utilized the Groupby function on the respective data frames and created a data set for each year that captured the labor market for that year in that geography. Looking at the data and previous research, we decided to use Total Employment, Median Hourly Wage, Annual 25th Percentile Wage, Annual Median Wage, and Annual 75th Percentile Wage in the model. The intent behind these factors was to try and group the various MSA according to the typical wage in that area.

Each years data was then run through the same cleaning process, until the data was ready

to be read into R and run through our clustering code.

3.3 K Means Clusters

Once the data was read into R we utilized an Elbow Plot to determine the number of K-Means Clusters to implement. Figure 3.2 displays the Elbow Plot used in this thesis.

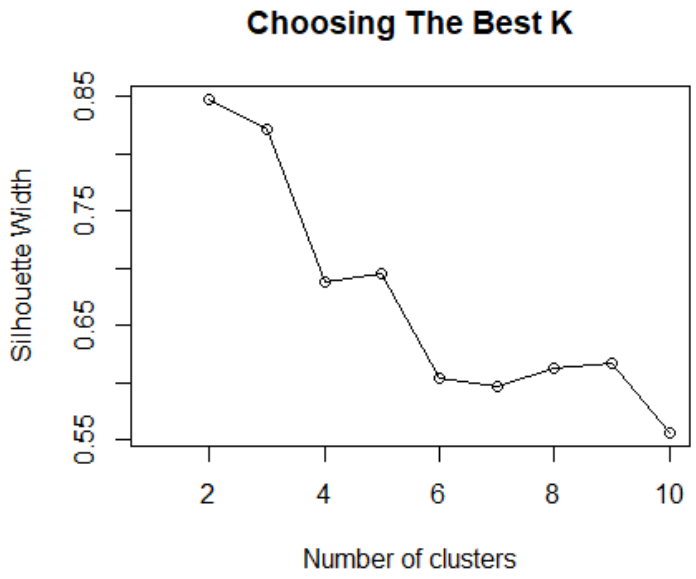

Figure 3.2. 2015 Elbow Graph: There is not a mathematical consensus on how to select the number of clusters. However there are some obvious wrong choices, such as two, eight, and nine. Two is obviously wrong because it has the highest value. Eight and nine are poor choices because the model actually gets worse as the clusters increase from seven.

Upon inspection within each year, six clusters appear to be the most likely number. When looking at the chart that is the point at which the slope begins to lose marginal utility and start to flat line. As discussed in Chapter 2, we could continue to segment the clusters into more and more clusters however we are seeing a decrease in marginal returns for additional clusters at approximately six. To illustrate, we ran that same elbow plot over 100 K's and Figure 3.3 displays the output.

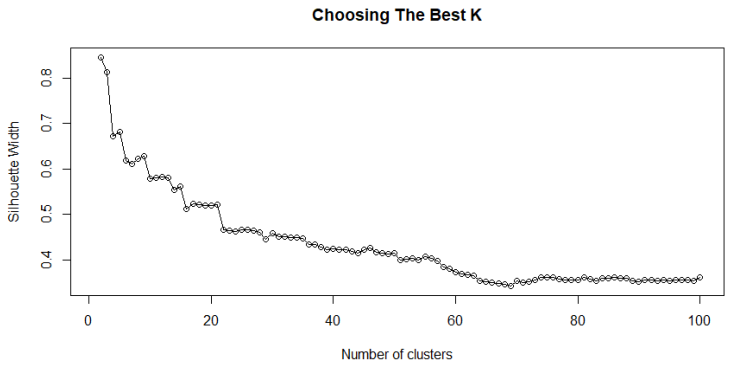

Figure 3.3. 100 K's Elbow Graph: This is where we decided to make a judgement call and stick with six clusters. The true lowest value appears to be around 70 clusters, however at that point concerns arise about the utility of that many clusters.

Six clusters held true throughout the data from 2015 to 2020. Following this test we then ran hierarchical clustering on the data sets to see if our selection of six clusters was reasonable. If it was then we would see a fairly distinct point in the Figure 3.4 that indicates after three splits (or six clusters) we begin to lose fidelity on the clusters.

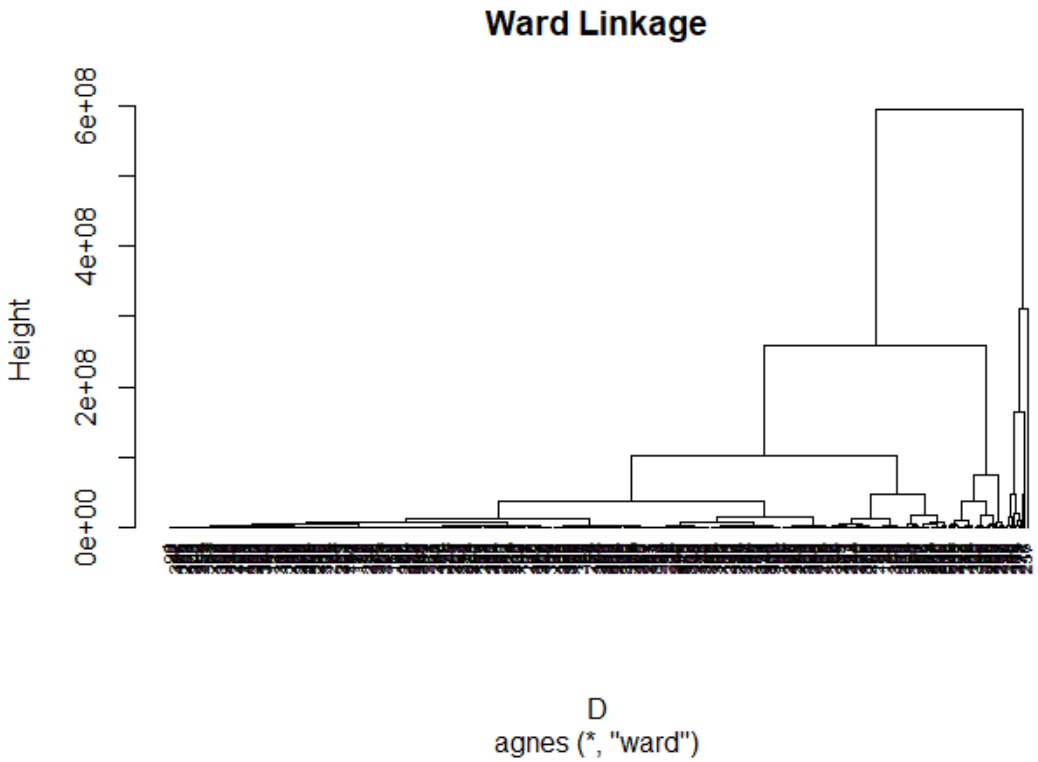

Figure 3.4. Ward Linkage Graph: This output confirms our judgement call from Figure 3.3, past three splits the graph begins to get extremely convoluted and difficult to parse.

Upon inspection of Figure 3.4 we determined that 6 was a reasonable amount of clusters and proceeded to fit the data utilizing K-Means Clustering. Utilizing the pam function (an acronym for Partitioning Around Medoids) we were able to group each year into six clusters. Below is the graph of 2015's output. Note: these charts are attempting to describe in multidimensional space, and look like they are overlapping. In reality the six clusters are separated in the feature space, but limitations on 2-D paper limit Figure 3.5's demonstration ability.

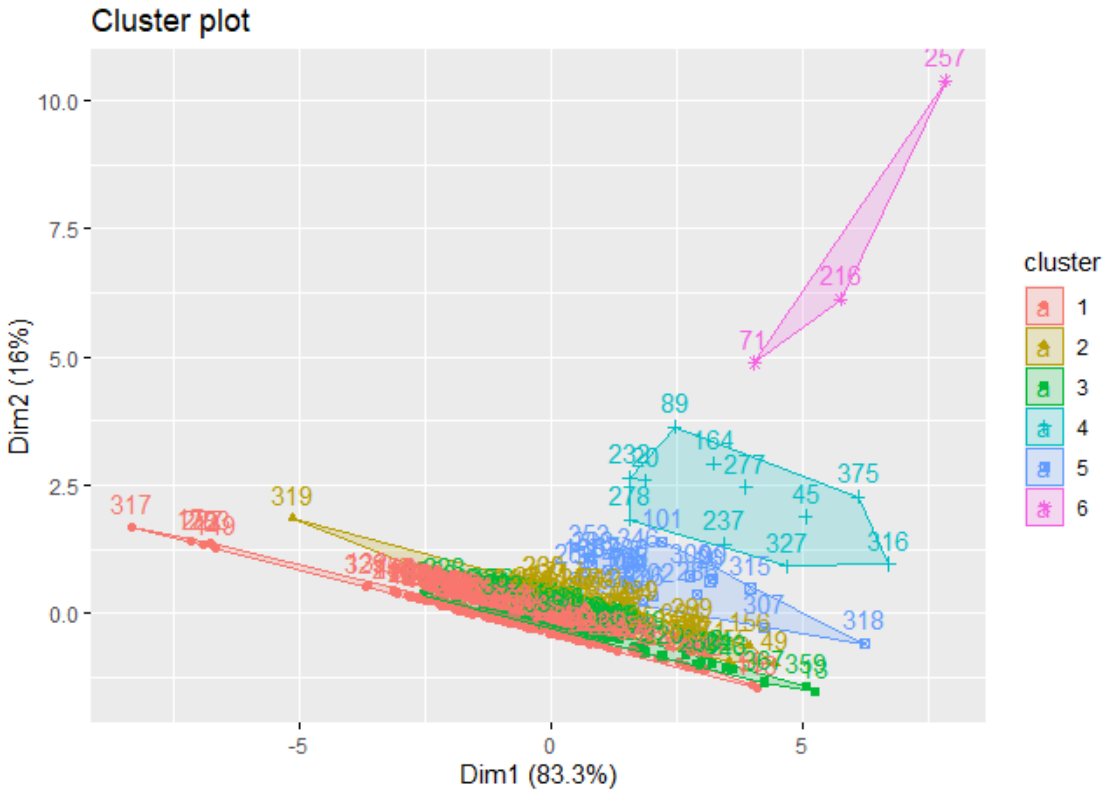

Figure 3.5. K-Means Clustering Graph, 2015 Data. This visual representation is much more limited than what is actually occurring. There are a couple of clusters that are obviously separated from the pack (clusters four and six), however that does not mean that these clusters are smashed on top of one another. With the number of features we utilized, these clusters very well may be extremely distinct.

3.4 Spectral Clusters

Following the K-Means clustering we explored the use of spectral clustering. The results for the 2015 data are in Figure 3.6.

Figure 3.6. Spectral Clustering Graph, 2015 Data. This graph looks very different from the K-Means graph, however more rigorous analysis is required prior to making a total judgement.

Again, this is a fairly poor visualization of the data due to the multidimensional nature of this problem. In order to see better insights of the data we turned to Google Maps to plot the data.

3.5 Data Visualizations

"Google My Maps" provides a very user friendly mapping tool that enables uploading of multiple different types of data from comma separated value sheets to shape files. After we ran the clustering algorithms through all the years, we read the data back into Python and

utilized the Pandas package to create a master data set that included each cluster, by year, by area, and by cluster type. We then uploaded this into Google My Maps.

There are many ways that the user can decide to display the information utilizing "Google My Maps." Figure 3.7 is a screen shot of the results of the 2020 data on the state of Texas.

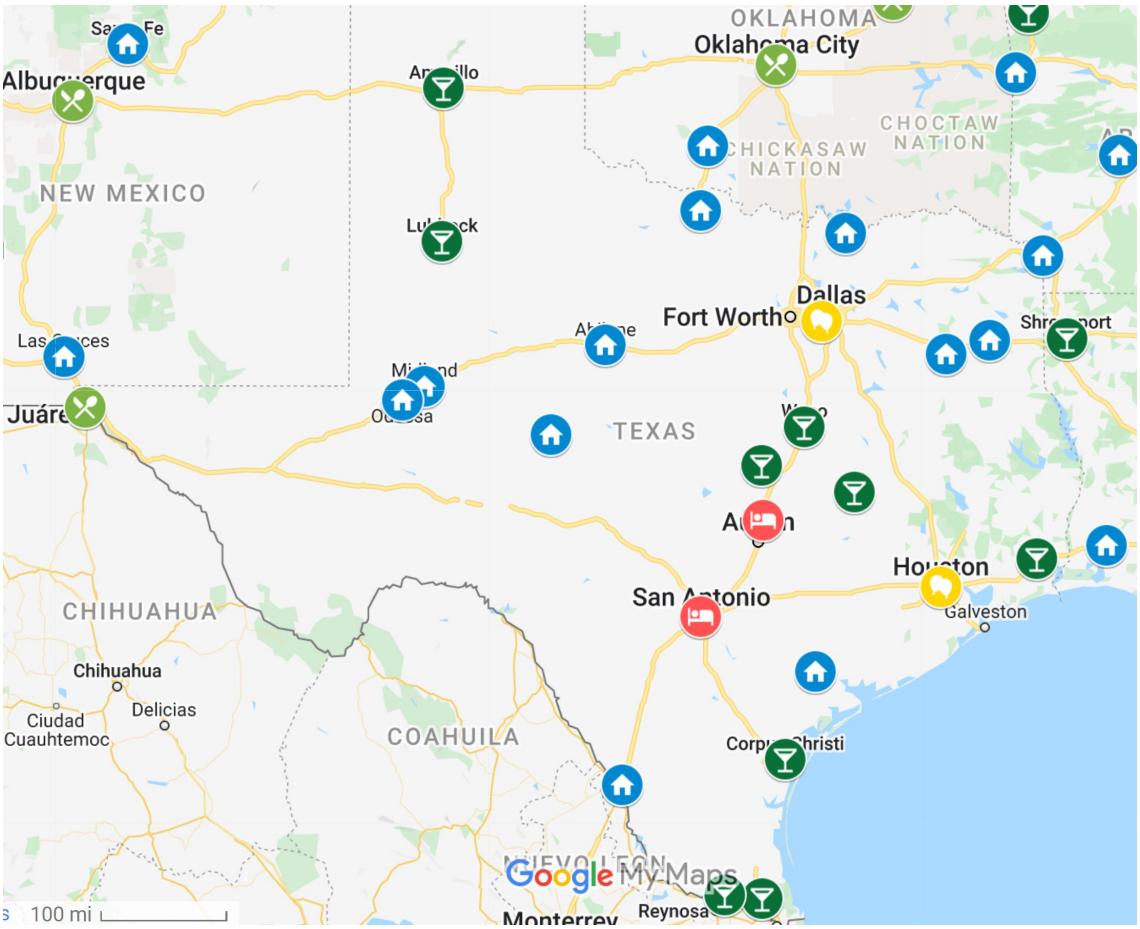

Figure 3.7. Google My Maps Output: Texas 2020 Data. For ease of depiction, the generic symbols for "Google My Maps" was utilized as well as generic colors. The user has the option to make the symbols anything they would like, including custom pictures, logos, and other files so long as they are in JPEG or PNC format. The map is fully interactive as well. The user can zoom out to view the entire world, and zoom all the way in so far as to count the brick pavers at a home.

As the Figure 3.7 depicts, there are some interesting relationships that seem to form between different geographies and wage markets even in Texas. Smaller markets in Texas like Waco, Killeen, and College Station identify as the same cluster, whereas the well long established Houston and Dallas are sharing the same cluster. Newer markets, relatively speaking, like Austin and San Antonio are grouped together.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:
Model Results and Analysis

Sections 4.1 and 4.2 will show very similar results. Both K-Means and Spectral Clustering selected largely the same MSAs for their groupings. This was an interesting finding as the two methods go about clustering in slightly different manners. Section 4.3 will go into detail about various nuances in the data and how clustering looked in the 2020 data set.

4.1 K Clustering Results

K-Means Clustering was used to cluster the data into six different clusters, with varying levels of distribution between clusters. The dominate cluster in 2020 had 213 of 396 MSA in it, or 54 percent of the density of the total data set. The next most dense cluster had 96 of 396 for a 24 percent share of the data. In descending order the rest of the four clusters had 48, 25, 11, and 3 MSA in them for 12, 6, 3, and 1 percent of the data. Figure 4.1 is the overview of the lower 48 states.

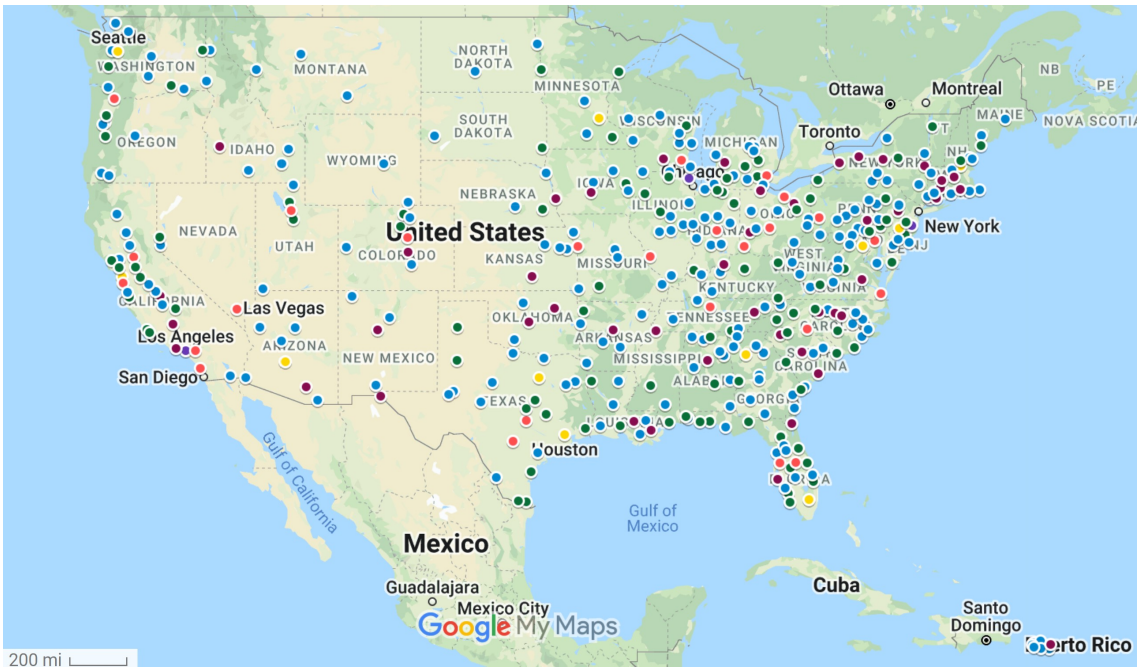

Figure 4.1. Google My Maps Output 2020: United States Lower 48 K Clusters. In order to better view the entire US the icons were changed to different color dots. The predominate color is blue. The majority of the blue dots appear in interior MSAs, which is an interesting component considering no geographic data was used as an input for clustering.

The real power behind this output is interacting with this map and being able to tease out the different factors in each MSA. We will go more into depth on this in Section 4.3.

4.2 Spectral Clustering Results

The Spectral Clustering resulted in identical clusters across the United States. Figure 4.2 displays the lower 48 states and their spectral cluster assignments:

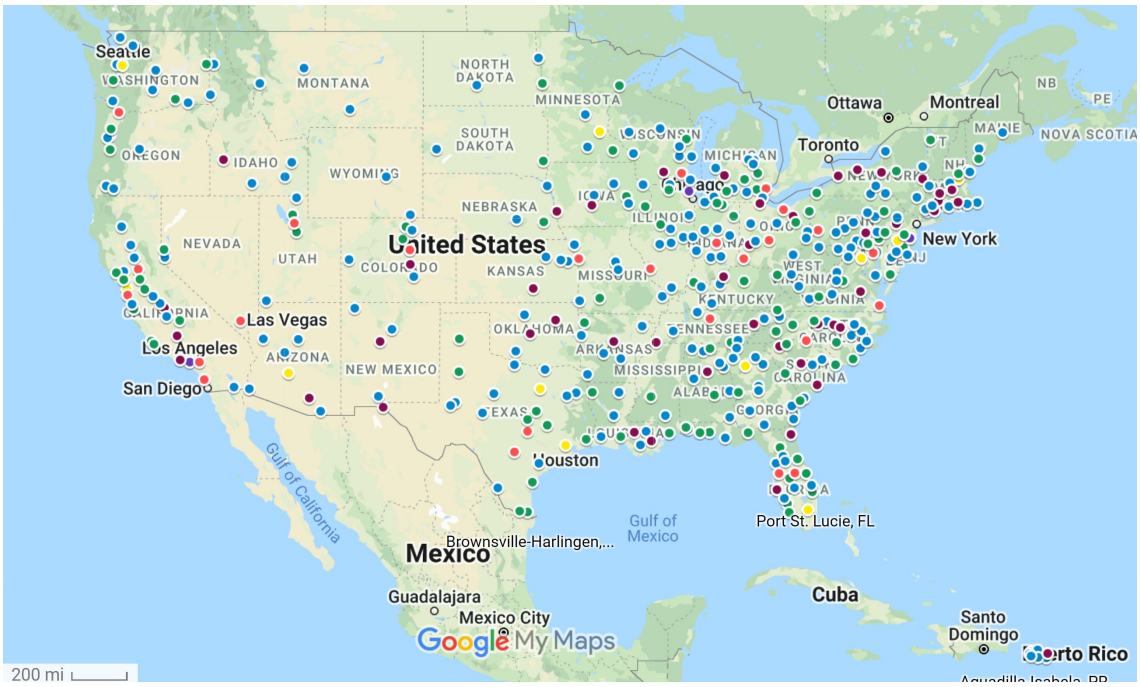

Figure 4.2. Google My Maps Output: United States Lower 48 S Clusters 2020: As alluded to in the previous chapter, the Spectral Clustering ended up producing identical clusters to the K Means clusters.

These overlays give us the baseline for doing more exploratory analysis.

4.3 Google Maps Manipulations

We were interested in seeing if there was a visible pattern between the total number of jobs and the cluster assigned to the respective MSA. To do this we utilized additional filters within "Google My Maps" to investigate. Figure 4.3 is one of the ways we looked closer at this factor.

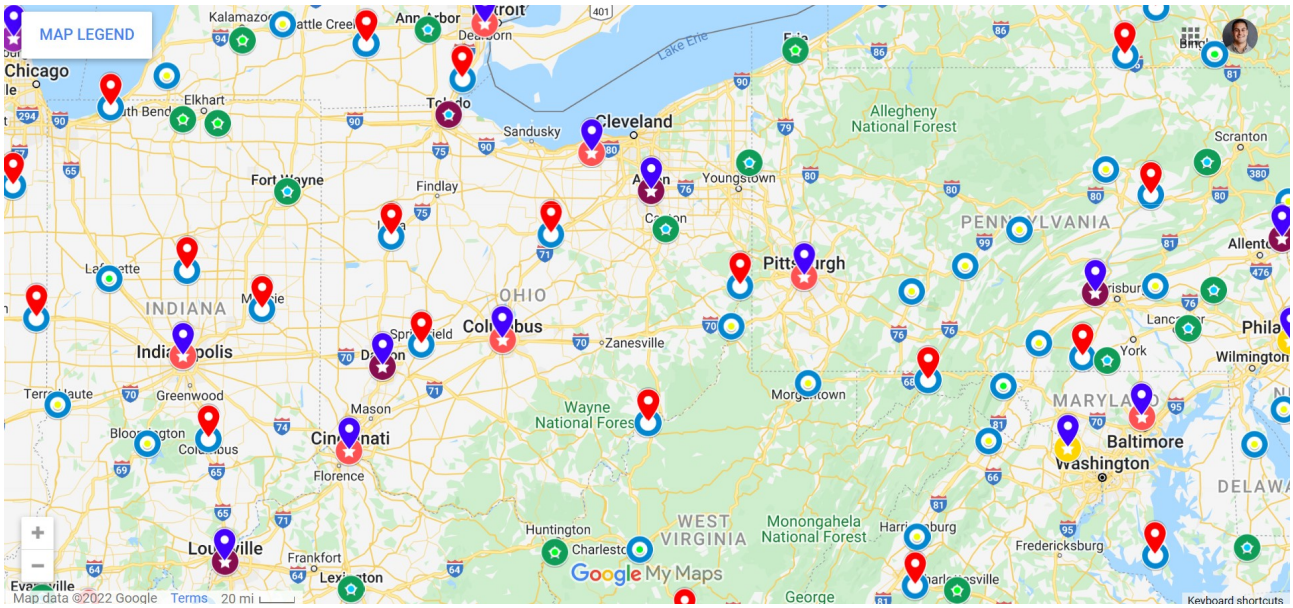

Figure 4.3. Top 20 Percent of Total Job MSA's and K Cluster Overlay: If a point has a red marker on it, then that cluster is in the bottom 20 percent for number of total jobs in that MSA. If the marker is indigo, then it is in the top 20 percent. While this is a close up screen shot of a smaller area for display purposes, when the United States as a whole is examined the majority of smaller job markets are in a single cluster, whereas the larger job markets are in three smaller density clusters.

Figure 4.3 is a sampling of the data. The purple markers are calling out the top 20 percent of total jobs, and the red markers are calling out the bottom 20 percent of total jobs. For clarity's sake we took a screen shot of only the Midwest, but after looking through the map it was evident that the most populated cluster had the preponderance of smaller job markets occupied the same cluster. The larger job markets occupied MSAs in three of the remaining four clusters.

4.4 Movement Within Clusters

Looking at this data across multiple years enabled us to compare movement within clusters to other clusters. We did not investigate why clusters shifted within the data set, but that would make for interesting research. Figure 4.4 and Figure 4.5 are the output for clusters

that experienced a shift in their assigned cluster for both K-Means and Spectral Clustering.

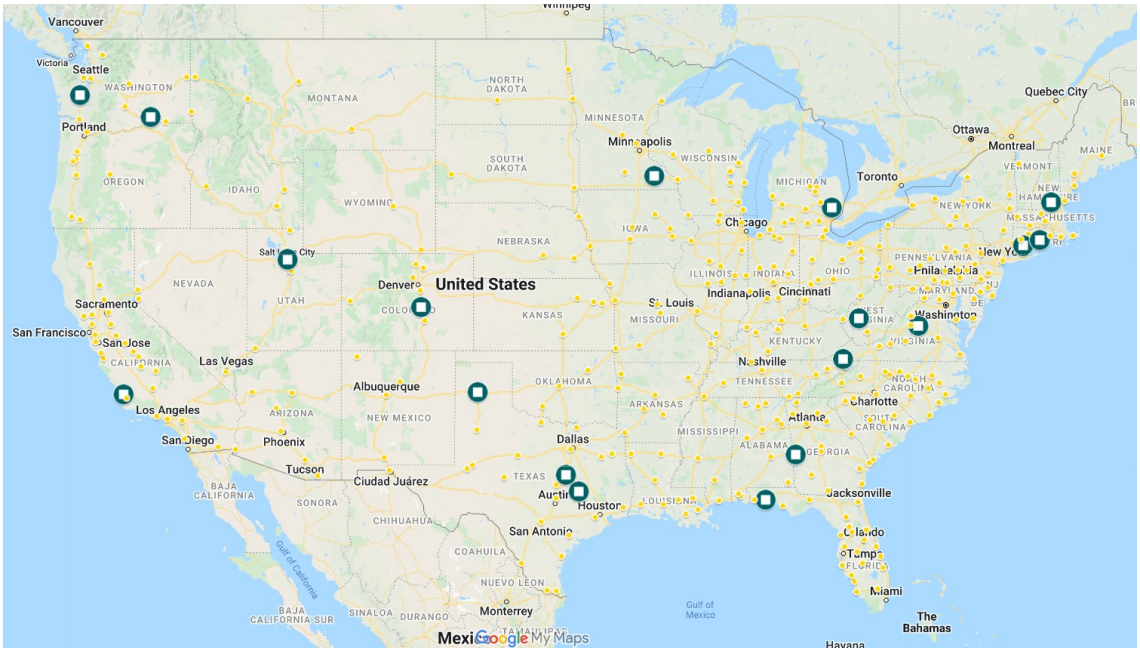

Figure 4.4. K Cluster Changes: 5 Year Outlook. The K Clusters that changed are in the green circles with white squares. All other MSA are yellow dots.

Figure 4.5. Spectral Cluster Changes: 5 Year Outlook: All Spectral Clusters that changed are in orange circles with a white star, the clusters that did not change are green dots. Comparing the two clustering techniques, both saw the same MSA change.

Both cluster methods had identical MSAs changing their respective clusters. Figure 4.6 takes a look at which clusters those belonged to in 2020, the largest cluster only experienced two movements (signified by the blue house icon) whereas the next most dense cluster (signified by the green martini glass) saw 12 of the 18 movers in both K-Means and Spectral Clustering.

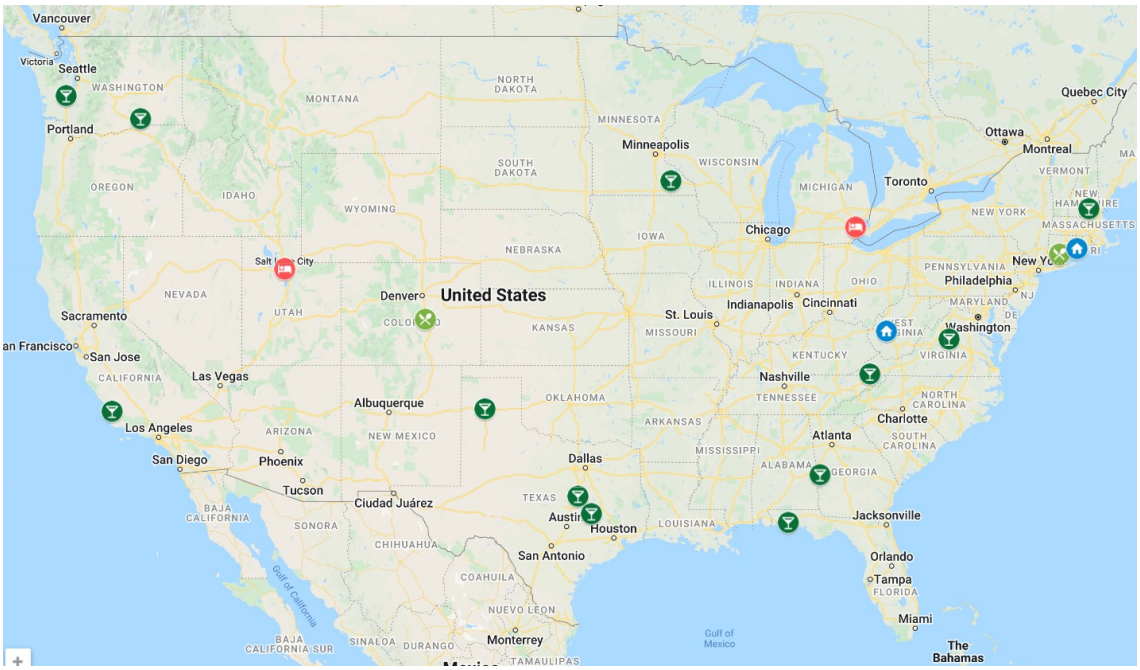

Figure 4.6. Spectral Cluster Changes: 5 Year Outlook with Large Icons. This shows which clusters the respective changing clusters belonged to in 2020. The distribution of changes does not appear to have a readily discernible pattern.

4.5 Trends Across Clusters

Due to the clustering methods selecting the same distributions for 2020, here we will do a deeper analysis on the various factors. First, we will look at the wage factors, then total number of jobs, and finally overall distribution of clusters.

4.5.1 Wage Factors

Looking first at the overall structure of the income data, it is helpful to examine a box and whisker plot of each factor, shown by Figure 4.7.

Figure 4.7. Box and Whisker Plot: Income Factors. Each factor has no more than 12 outliers. With the data set consisting of 396 different MSA that is approximately 3 percent of the data registering as an outlier.

None of the interquantiles overlap with one another. A full summary of the data set is in Tables 4.1, 4.2, 4.3, and 4.4.

Table 4.1. Summary Statistics: 25th Percentile Annual Wages 2020: The data for the 25th percentile is found to be relatively normally distributed with a skewness score of 0. A standard deviation of 5,206 USD and a mean of 41,648 USD and an assumption of normality indicates that the mean is contained in the interval [31,235, 52,060] with 98 percent confidence. The poverty line for a family of four in the U.S. was 26,500 USD in 2021 (Office of the Assistant Secretary for Planning and Evaluation 2022). Source: United States Bureau of Labor Statistics (2022b).

A_PCT25		
Mean	$	41,648
Standard Error	$	262
Median	$	41,284
Mode		#N/A
Standard Deviation	$	5,206
Sample Variance	$	27,104,222
Kurtosis	$	2
Skewness	$	(0)
Range	$	38,250
Minimum	$	20,295
Maximum	$	58,545
Sum	$	16,492,589
Count	$	396

Table 4.2. Summary Statistics: Median Annual Wages 2020: With a skewness score of 0 this data is approximated by the normal distribution. It's 98 percent confidence interval of the mean is contained by the interval [38,421, 64,167]. This shows that the lower end of the median annual earnings has an overlap with the higher end of the 25th percentile, indicating that some median annual wages are similar to the 25th percentile for some MSA.

A_MEDIAN	
Mean	$ 51,294.56
Standard Error	$ 323.44
Median	$ 50,798.33
Mode	#N/A
Standard Deviation	$ 6,436.32
Sample Variance	$ 41,426,270.91
Kurtosis	$ 2.44
Skewness	$ (0.18)
Range	$ 50,044.76
Minimum	$ 23,631.25
Maximum	$ 73,676.01
Sum	$ 20,312,647.70
Count	396

Table 4.3. Summary Statistics: 75th Percentile Annual Wages 2020: This data is also able to be approximated by the normal distribution with a skewness of 0. [48,148, 78,996] is the 98 percent confidence interval for the mean.

A_PCT75		
Mean	$	63,572
Standard Error	$	388
Median	$	63,094
Mode	#N/A	
Standard Deviation	$	7,712
Sample Variance	$	59,476,028
Kurtosis	$	3
Skewness	$	(0)
Range	$	60,393
Minimum	$	28,980
Maximum	$	89,373
Sum	$	25,174,621
Count	396	

Table 4.4. Summary Statistics: Median Hourly Wages 2020: Federal Minimum is 7.25 USD an hour. This data shows that the mean hourly wage in the United States is approximately three times as much as minimum wage (People Ready (2022).

H_MEDIAN	
Mean	24.46
Standard Error	0.15
Median	24.28
Mode	#N/A
Standard Deviation	3.00
Sample Variance	9.02
Kurtosis	2.88
Skewness	-0.28
Range	23.67
Minimum	11.39
Maximum	35.05
Sum	9684.24
Count	396.00

4.5.2 Total Jobs Examined

Total employment factor appears to not come from a normal distribution. Table 4.5 is the summary statistics for this factor.

Table 4.5. Summary Statistics: Total Employment 2020: There is a strong skewness to the right of the data with a score of 7. This means that the majority of the MSA are smaller markets, with some large job numbers skewing the data to the right. That skewness makes the median a more descriptive statistic as it is less sensitive to outliers than the mean.

Total Jobs	
Mean	607238
Standard Error	73413
Median	178795
Mode	65210
Standard Deviation	1460909
Sample Variance	2134256067159
Kurtosis	58
Skewness	7
Range	17627110
Minimum	14370
Maximum	17641480
Sum	240466190
Count	396 height

Looking at Figure 4.8, a histogram of the total employment factor, evidence that this skew exists as well.

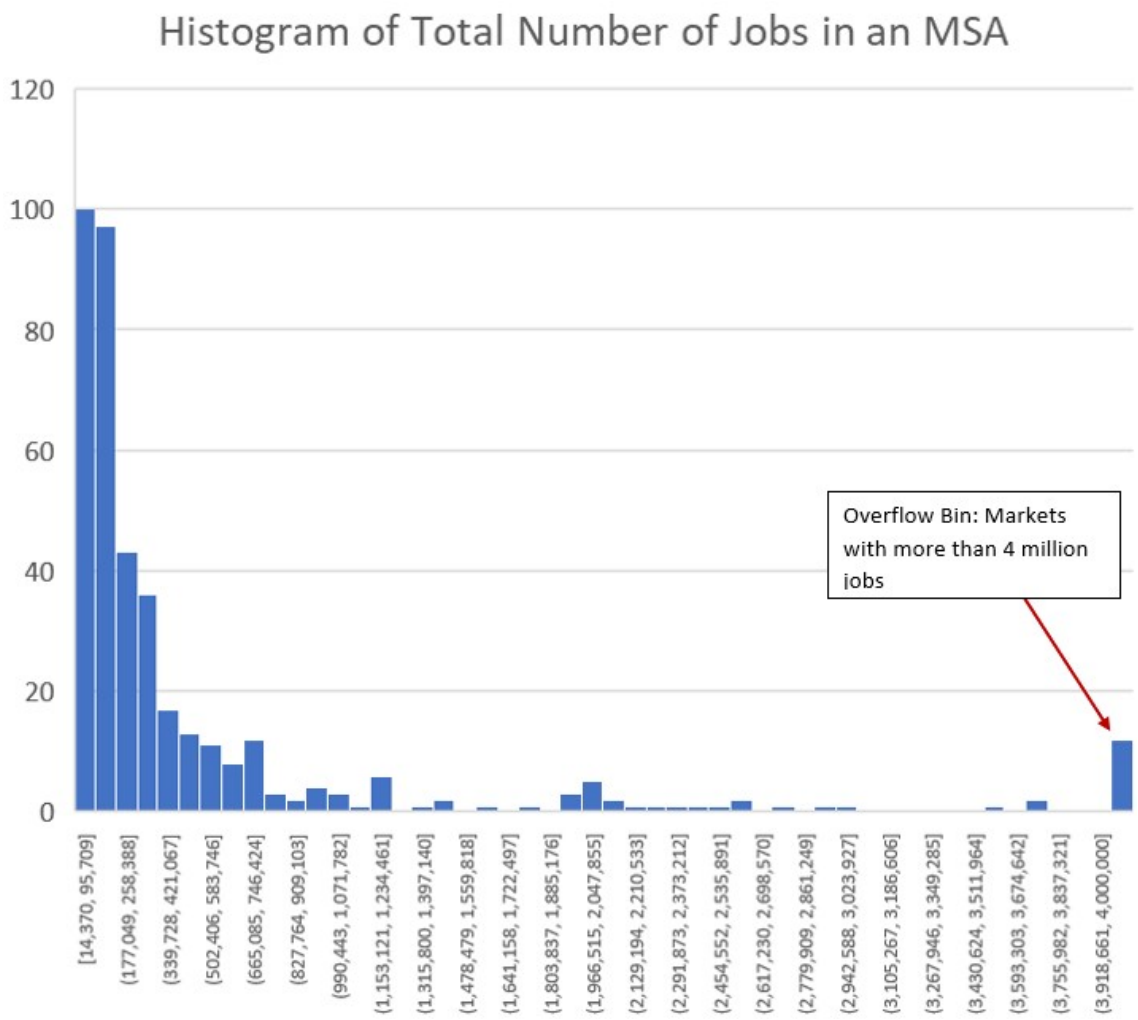

Figure 4.8. Histogram Total Jobs 2020: Note that the overflow bin was utilized in order to better see how many MSA were above 4,000,000 jobs. These markets pull the data to the right when calculating averages.

The majority of MSAs in the United States of America (USA) consist of less than 750,000 jobs. 12 of the 384 MSAs have more than 4,000,000 jobs in them, which drags the data to the right. Further investigation could be done segmenting these bins however that is beyond the scope of this project.

Looking at the individual clusters in Figure 4.9, number of jobs seems to be a strong factor in the cluster assigned.

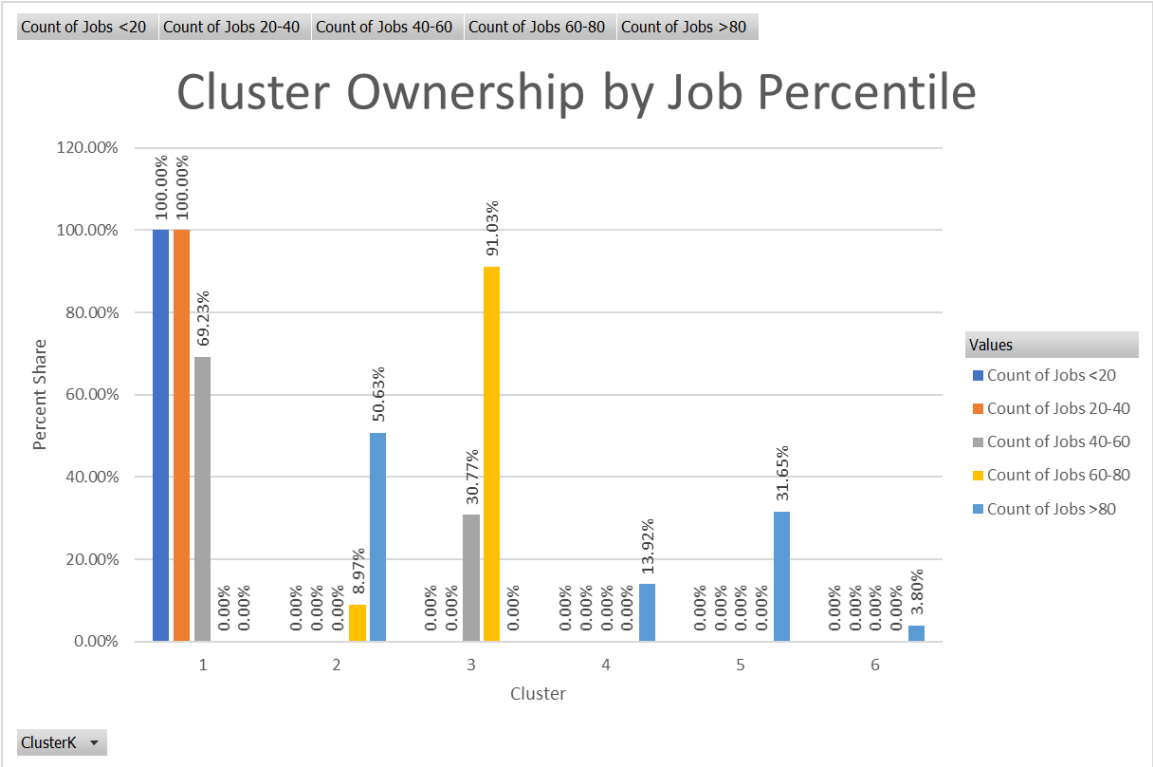

Figure 4.9. Cluster Ownership: This chart breaks down the percentage of MSA that have representation in density of jobs. From left to right, Dark Blue is the lower 0-20 percent of total number of jobs, Orange is 20-40 percent of total number of jobs, Grey is 40-60 percent of total number of jobs, Yellow is 60-80 percent of total number of jobs, and Light Blue is the 80-100 total number of jobs of all the MSA. For example, Cluster 1 has 100 percent of the MSA that are in the lower 40 percent of total jobs. Cluster 3 has a mixture of the 40-60 and 60-80 jobs. The largest job markets are all in four different clusters; 2, 4, 5, and 6 Source: United States Bureau of Labor Statistics (2022b).

Taking a look at income in Figure 4.10, it appears that there is more diversity across clusters.

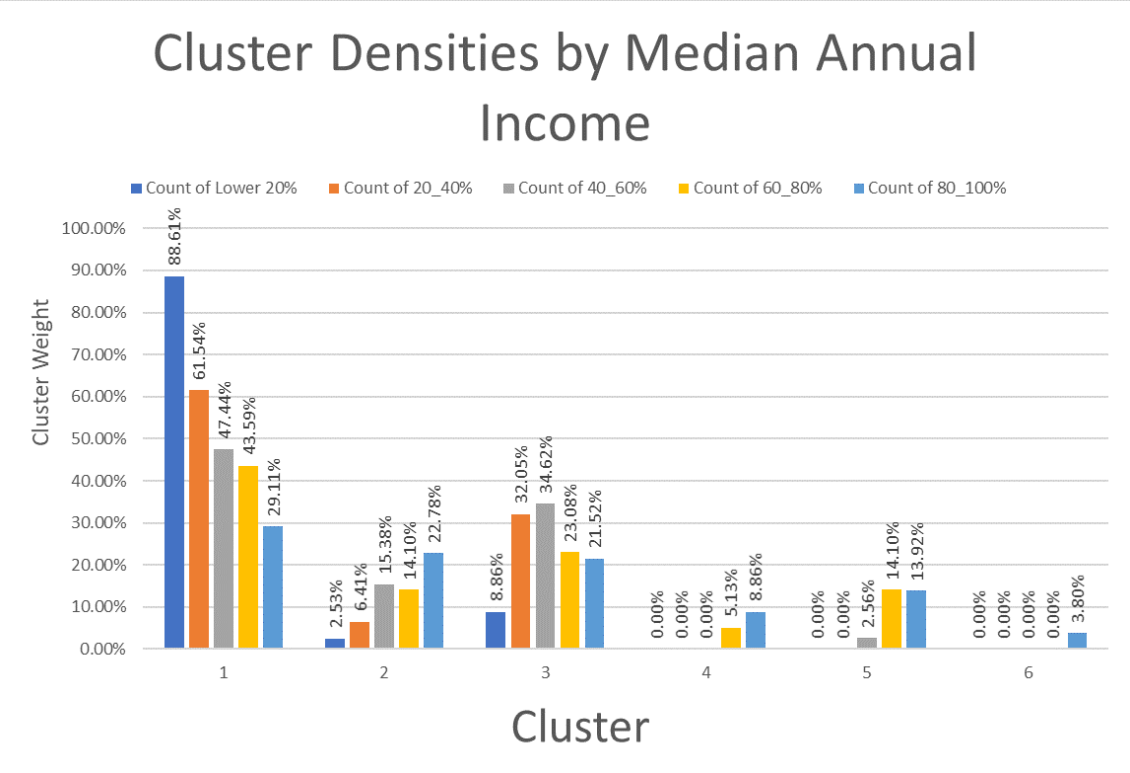

Figure 4.10. Cluster Density Median Income: This chart uses the same terminology as 4.9, displaying the level of income represented in each cluster. From left to right, Dark Blue is the lower 0-20 percent of income level, Orange is 20-40 percent of income level, Grey is 40-60 percent of income level, Yellow is 60-80 percent of income level, and Light Blue is the 80-100 income level of all the MSA. Cluster 1 seems to have a preponderance of the lower income bracket and tapers off as income goes up, however it does have over 25 percent representation in each segment. Cluster 6 has only high income MSA and also contains MSA with a large number of jobs. Cluster 6 contains Los Angeles, Chicago, and New York. Source: United States Bureau of Labor Statistics (2022b).

CHAPTER 5:
Summary, Conclusion, and Future Research

This chapter will conclude this thesis. First we will summarize, then state the conclusion, and have a brief discussion on future research opportunities with this data and methodology.

5.1 Summary

In summary, this thesis explores how different clustering methods could help identify similar MSAs across the United States. The data set that we used came from the BLS and was a part of the OEWS data. The factors used in order to cluster the MSA were percentiles of wages earned as well as total number of jobs in that MSA.

The clustering methods included K-Means Nearest Neighbor algorithms, Hierarchical Clustering, and Spectral Clustering. These different methods all yielded similar results despite the different methods in which they calculate their clusters. Also, the clusters were relatively stable over the past six years worth of data, with very few MSAs changing clusters.

5.2 Conclusion

In conclusion, we have established a method with which to begin more data driven and scalable clustering of economic information. The utilization of open source visualization via "Google My Maps" is extremely helpful and provides a way for non technically inclined individuals to grasp connections in data. This work has only scratched the surface of what is possible with these techniques and follow on research will be important.

5.2.1 K-Means Nearest Neighbor

This algorithm used Euclidean distance to calculate the clusters in the data set. This proved

to be an effective method of clustering the data with our team identifying six clusters throughout the United States.

After identifying the clusters we ran each year from 2015 to 2020 through our code and discovered that 18 clusters had changed in 2016-2020. A change was classified as any cluster that did not stay the exact same from 2016 to 2020. Some clusters changed into a new cluster in the middle of the sample, then reverted to their original cluster. Others ended 2020 in a new cluster from its previous four years. Cluster 3 saw the majority of the movement, with 2020 Cluster 3 consisting of 12 of the 18 MSA movements. These movements are displayed in Figure 5.1.

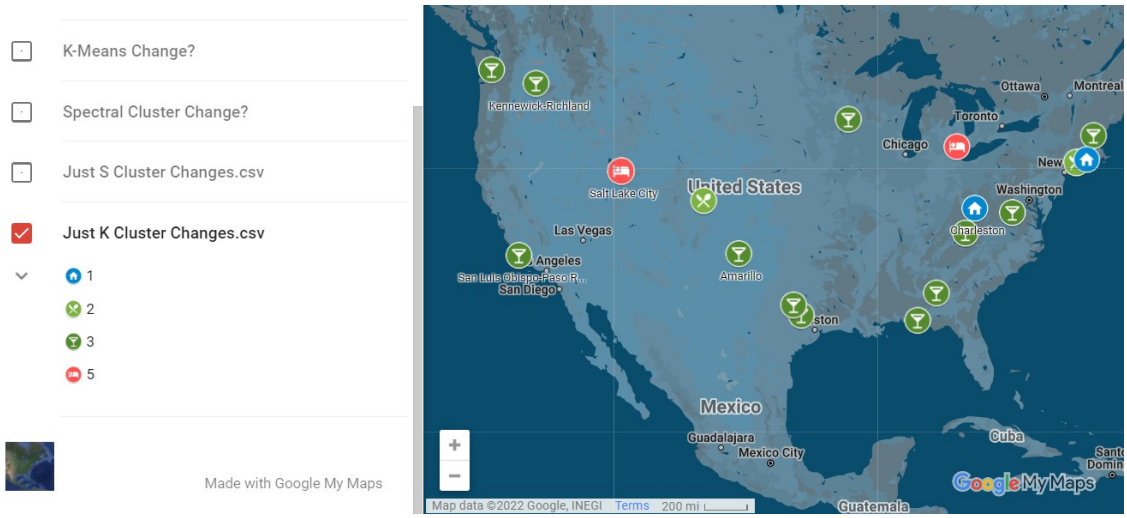

Figure 5.1. K-Means Changes: The Green Martini glass represents Cluster 3. Cluster 3 was the most active amongst changing clusters. The largest cluster, Cluster 1 represented by a Blue House, only saw two MSA change. Cluster 1 consisted of all lower income and majority of the lower total job MSA. Clusters 4 and 6 did not see any movement over the five year period and do not have any plots on the map (United States Bureau of Labor Statistics 2022b).

5.2.2 Spectral Clustering

We found that the results for Spectral Clustering were extremely similar to those of K-Means Nearest Neighbor throughout the data set and identical for 2020 data. This was surprising

to the team, as we expected the Euclidean distance method of K-Means to differ from the Partitioning Around Mediods method used by Spectral Analysis.

When examined closer, the 2015 data seems to have been the largest departure for both data sets, as shown in Figure 5.2.

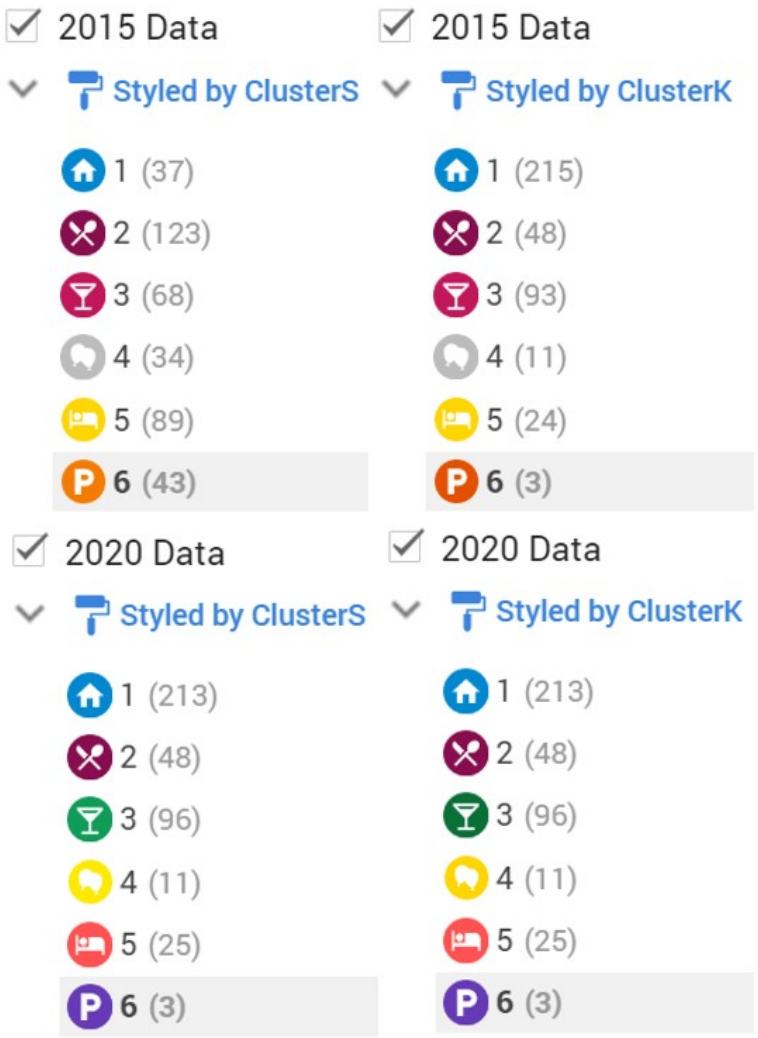

Figure 5.2. 2015 Clusters Compared to 2020 Clusters: It appears there was some sort of anomaly in the data from 2015 when it comes to Spectral analysis. The year produced very different results compared to the rest of the data set. This figure shows the clusters for both Spectral and K-Means clustering in 2015 and 2020. The 2020 results are identical, and the 2015 K-Means results are much closer to 2020 results. It appears that 2015 was an outlier for the data set with respects to Spectral Clustering. Source: United States Bureau of Labor Statistics (2022b).

5.2.3 Hierarchical Clustering

Hierarchical Clustering was predominately used to verify out assumptions. With 394 different MSA it is possible that the number of clusters could be in the double digits. However upon reviewing the dendogram it was clear that our assumption of six clusters was supported by the data.

5.2.4 Income Patterns

Another interesting finding was the layout of high to low income, as well as the amount of pay increases across all MSA. From both metrics Figure 5.3 and Figure 5.4 show earning power is concentrated on the coasts.

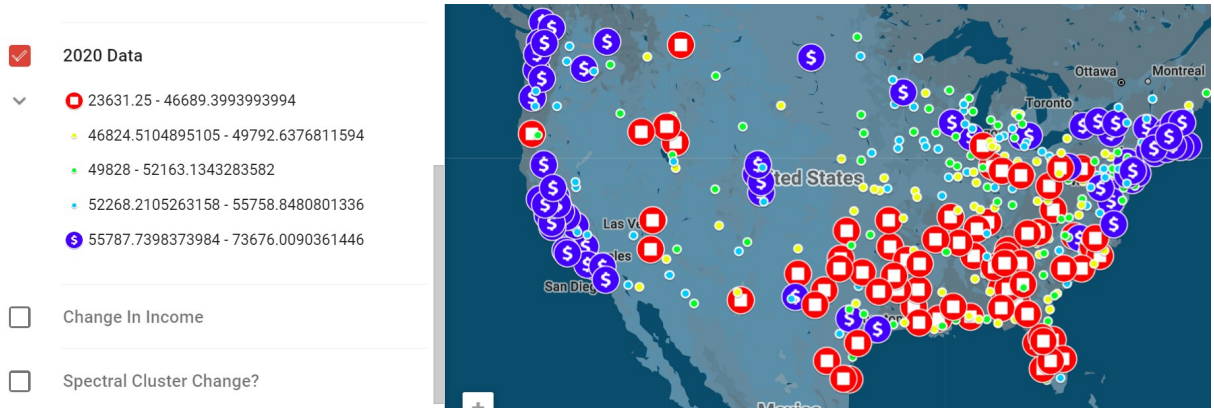

Figure 5.3. Median Income Top and Bottom 20 Percent 2020: The Blue Dollar Signs indicate the MSA that saw the highest median income levels in 2020. The Red Circles with White Squares indicate the markets that were in the lowest 20 percent. Just looking at the map there is a clear pattern of large median incomes in the north east and along the west coast. While individuals do get paid more in those states this is not taking into account cost of living adjustments for the respective MSA. Source: United States Bureau of Labor Statistics (2022b).

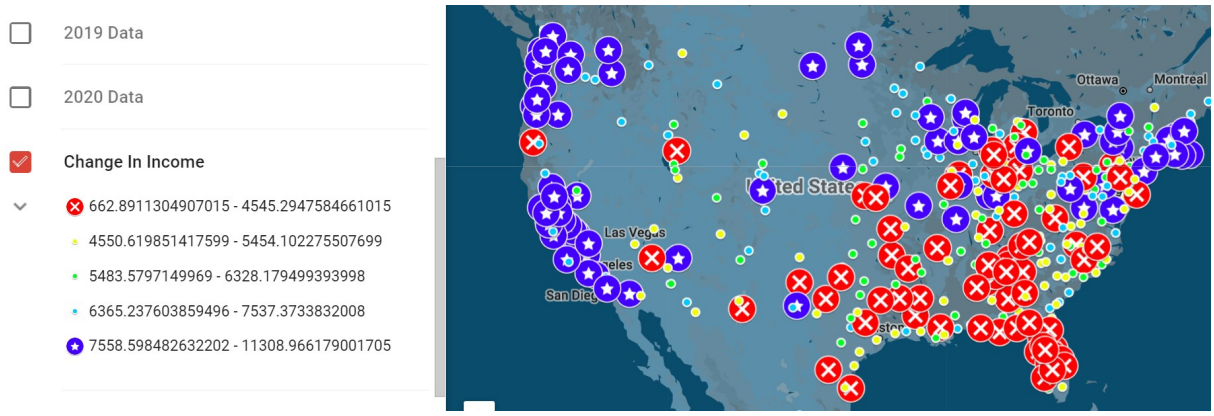

Figure 5.4. Changes in Income Over Six Years: This figure displays Blue Stars for MSA that experiences the top 20 percent of median income growth over the past six years. The Red X indicate the lowest 20 percent in income growth. This map is very similar to Figure 5.3, which is an indication that the higher paying markets also increased their raw amount of pay over the past six years as well. Bismark North Dakota was a top performer in both scenarios, but not geographically close to the rest of the herd relative to income amount and growth. Source: United States Bureau of Labor Statistics (2022b).

5.3 Opportunity for follow on research

There are many opportunities to conduct follow on research on this topic. The United States collects a significant amount of data for MSAs and with the framework we have built expanding the data set and running the code should be explored.

A couple of the more promising avenues that we were unable to investigate involve housing data, unemployment data, more specialized data, and different levels of geography.

5.3.1 Housing Data

Texas A&M University Real Estate Research Center Texas A&M University Real Estate Research Center (TAMU) has very neat and clean data sets for the public to look at detailing various data points like building permits for various types of residential and commercial

real estate. This data would marry up nicely with our current data set and could produce some interesting insights on different MSAs across the United States.

Realtor.com is also a treasure trove of housing data. The data curated is typically very clean and is updated monthly. Supply and demand factors like total number of listings and listing prices are provided as well as other important metrics like days on market and square footage of homes. Zillow.com has similar data as Realtor.com, however they also include estimates of market rent.

This data could be merged and manipulated into a larger data set enabling researchers to tease out how housing clusters across different areas.

5.3.2 Unemployment Data

Unemployment data was also not taken into consideration for this research. The U.S. Census Bureau and Bureau of Labor Statistics have all the requisite data for expansion of this data set. While our research took total number of jobs into account, it did not take the unemployment rate into account.

5.3.3 Specialized Data

If public planners are interested in other factors, our methodology is flexible enough to cluster numerical data so long as the data is collectable for that geographic level. For example, perhaps a city planner is looking to see if it is a good idea to build a new professional sports arena. If they were to add an additional column into the data signifying number of professional sports arenas in each MSA they may be able to focus their due diligence process on MSAs that are similar. This area of future research is infinitely scalable and could be custom tailored to each planners needs.

5.3.4 Different Levels of Geography

This methodology provides a framework to cluster similar data sets at different levels of

granularity. Currently the data was aggregated at the MSA level, however the same process could be done to lower levels like cities or higher levels like states.

Beyond this data set, global clusters could also be developed and be used by the international community in order to help guide international investment and interventions. Nations that are similar in cluster may benefit from cross coordination between one another.

List of References

Delgado M, Porter M, Stern S (2016) Defining clusters of related industries. Journal of Economic Geography.

Federal Reserve Bank of St Louis (2012) The role of self-interest and competition in a market economy - the economic lowdown podcast series. The Economic Lowdown Podcast, accessed May 2, 2022, https://www.stlouisfed.org/education/economic-lowdown-podcast-series/episode-3-the-role-of-self-interest-and-competition-in-a-market-economy.

Federal Reserve Economic Data (2022) Employed full time: Median usual weekly real earnings: Wage and salary workers: 16 years and over. Accessed April 24, 2022, https://fred.stlouisfed.org/series/LES1252881600Q.

Fleshman W (2019) Spectral clustering foundation and application. Towards Data Science, https://towardsdatascience.com/spectral-clustering-aba2640c0d5b#:~:text=Spectral%20clustering%20is%20a %20technique,non%20graph%20data%20as%20well.

Gao J (2021) Clustering lecture 2: Partitional methods. Class Notes, https://cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_partitional.pdf.

Grootendorst M (2021) 9 distance measures in data science. Towards Data Science, https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa.

Jason Fernando (2022) What is inflation? Investopedia, accessed April 24, 2022, https://www.investopedia.com/terms/i/inflation.asp.

March J (2015) Analysis shows city spent $475,000 on consultants Accessed May 2, 2022, https://www.starnewsonline.com/story/news/2015/04/20/analysis-shows-city-spent-475000-on-consultants/30978948007/.

Office of the Assistant Secretary for Planning and Evaluation (2022) Occupational employment and wage statistics: Frequently asked questions. United States Department of Health and Human Services, accessed April 25, 2022, https://aspe.hhs.gov/topics/poverty-economic-mobility/poverty-guidelines/prior-hhs-poverty-guidelines-federal-register-references/2021-poverty-guidelines.

People Ready (2022) Minimum wage set to increase in many U.S. states for 2022. People Ready: A Trueblue Company, accessed April 25, 2022,

https://www.peopleready.com/minimum-wage-set-to-increase-in-many-us-states-for-2022/ #:~:text=Minimum%20wage%20laws%20by%20state%20for%202022 &text=In%20those%20states%2C%20the%20federal,1%2C%202022.

Trading Economics (2022) United States inflation rate. Accessed April 24, 2022, https://tradingeconomics.com/united-states/inflation-cpi#:~:text=US%20Inflation%20Rate%20Tops%20Forecasts,with %20market%20forecasts%20of%208.4%25.

United States Bureau of Labor Statistics (2015) How the government measures unemployment. Accessed April 23, 2022, https://www.bls.gov/cps/cps_htgm.htm.

United States Bureau of Labor Statistics (2022a) Current employment statistics. Accessed April 23, 2022, https://www.bls.gov/ces/.

United States Bureau of Labor Statistics (2022b) Occupational employment and wage statistics. United States Bureau of Labor Statistics, accessed April 24, 2022, Data under "All Data: XLS", https://www.bls.gov/oes/tables.htm.

United States Bureau of Labor Statistics (2022c) Occupational employment and wage statistics: Frequently asked questions. United States Bureau of Labor Statistics, accessed April 24, 2022, https://www.bls.gov/oes/oes_ques.htm.

United States Census Bureau (1994) Geographic reference manual: Chapter 13 metropolitain statistical areas. Accessed April 23, 2017, https://www2.census.gov/geo/pdfs/reference/GARM/Ch13GARM.pdf.

United States Census Bureau (2022a) North american industry classification system. Accessed April 23, 2017, https://www.census.gov/naics/?58967?yearbck=2022.

United States Census Bureau (2022b) North american industry classification system: Search results for: 11. Accessed April 23, 2017, https://www.census.gov/naics/?input=11&chart=2022.

US Department of the Treasury (2022) Role of the treasury. U.S Department of the Treasury, accessed June 1, 2022, https://home.treasury.gov/about/general-information/role-of-the-treasury.

Viz FDT (2022) Dendogram Accessed May 12, 2022, https://www.data-to-viz.com/graph/dendrogram.html.

Yoshida DR (2021) Oa4106 data science short course clustering. Class notes, Advanced Data Analysis, Naval Postgraduate School, September, 2021, Monterey, CA.

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California