COVID-19 Impact: Cluster Analysis

Data Mining Project 2

PREPARED BY

Trevor Dohm, Eileen Garcia, Blake Gebhardt

Executive Summary

The COVID-19 pandemic has presented a significant challenge for policymakers, health professionals, and the public. To tackle this challenge, data mining techniques have been employed to analyze COVID-19 data from various regions of the United States. The primary focus of this report is to utilize clustering techniques to answer key research questions related to the pandemic. Our approach involved collecting data from various sources, including the CDC, state and local health departments, and publicly available data. We employed clustering techniques to identify groups of counties with similar characteristics, such as population density, demographics, and socioeconomic status. By doing so, we aimed to identify counties that were similar in terms of their response to the virus and the makeup of the population.

Our analysis revealed that the infection rate varied significantly across different regions of Texas and that some regions performed better in controlling the spread of the virus than others. Clustering allowed us to group counties with similar characteristics, which helped to identify high-risk areas and hotspots where the virus was spreading rapidly. By understanding the unique characteristics of these regions, policymakers and health professionals can better target their resources and interventions to areas that need them the most. Our report provides valuable insights into the COVID-19 pandemic in the US. By utilizing clustering techniques, we were able to identify groups of counties that share similar characteristics and responses to the virus. This information can be used by the CDC and other health professionals to target their resources and interventions to the areas that need it the most, ultimately helping to control the spread of the virus. First, we analyzed the trend in different areas of the US, including states and counties. We found that the COVID-19 infection rate varied significantly across different regions. We identified regions that did particularly well in controlling the spread of the virus. We used clustering techniques to identify regions with similar characteristics, such as population density, demographics, and socioeconomic status. Finally, using the accumulated information, we came to some conclusions on the development in any given region based on the data of other regions.

Table of Contents

Business Understanding

COVID-19, also known as the novel coronavirus, is a highly infectious respiratory illness that was declared the cause of a global pandemic since its initial outbreak in Wuhan, China in December 2019. It is caused by a virus known as SARS-CoV-2 and is primarily spread through respiratory droplets from an infected individual when they speak, cough, or sneeze. Since it was first discovered, nearly 800 million people across the globe have been infected, and nearly 7 million have died from the disease [1]. Social distancing and the term "flattening the curve" are efforts aimed at slowing the spread of the virus. Social distancing involves staying at least 2 meters away from other people, avoiding large gatherings, and working from home if possible [2]. Social distancing aims to reduce the spread of COVID-19 by limiting contact between others and preventing prolonged contact with respiratory droplets. By staying physically separate, the likelihood of infection by inhaling infected droplets decreases. "Flattening the curve" refers to slowing the rate of new cases so that the healthcare system can withstand the volume of patients. The "curve" refers to the graph of new infections. According to the New York Times, the curve peaked in winter 2020-2021 and winter 2021-2022 [3]. By practicing social distancing and other disease-fighting measures, the rate of new infections will slow, flattening the curve.

Clustering techniques have been employed in the fight against COVID-19 to analyze the spread of the virus, hospitalizations, and available resources such as hospital beds, ventilators, and healthcare workers. The information obtained from clustering analysis is critical in the decision-making process for healthcare officials and policymakers. By analyzing the number of cases, hospitalizations, and deaths, health officials can determine the impact of COVID-19 on the population and identify areas that are most affected. Clustering techniques also allow the identification of hotspots where the virus is spreading rapidly, and resources can be targeted accordingly. By monitoring hospitalization data, healthcare officials can determine the number of hospital beds, ventilators, and healthcare workers needed to effectively treat patients. This information can help decision-makers determine the effectiveness of various measures, such as social distancing and vaccine distribution, and make informed decisions about how to respond to the pandemic. By understanding these factors, we can work towards controlling the pandemic and protecting the health and well-being of individuals around the world, aiming to prevent mass infection in the future.

Data Understanding

U.S. COVID-19 Cases and Census Dataset

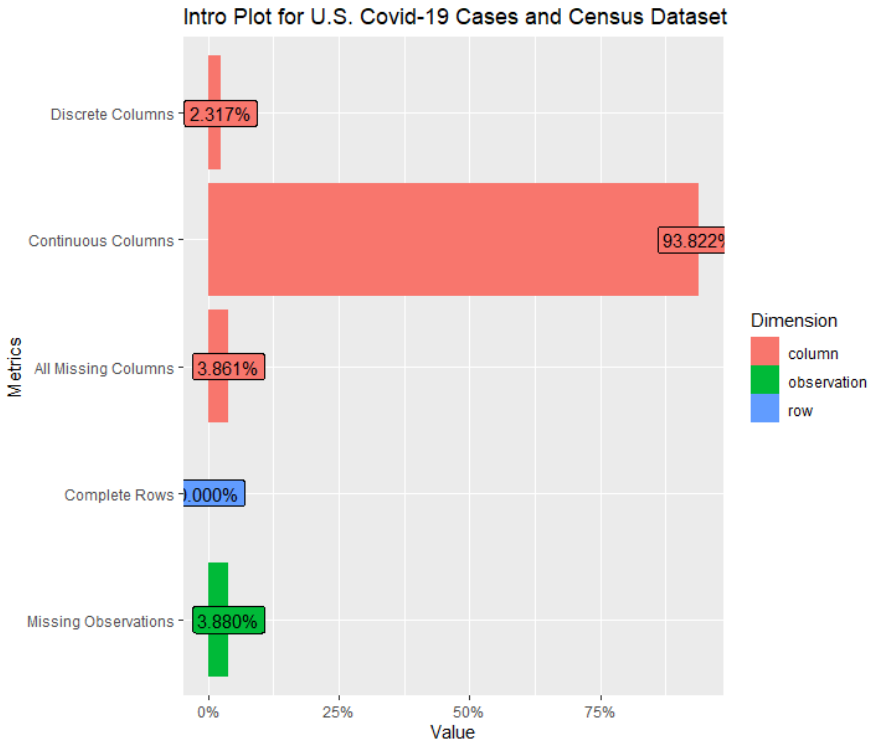

Figure 1. Intro Plot for U.S. Covid-19 Cases and Census Dataset.

The U.S. COVID-19 Cases and Census Dataset is a comprehensive collection of data related to the ongoing COVID-19 pandemic in the United States. This dataset includes critical information on confirmed COVID-19 cases and deaths, as well as a vast range of demographic data such as age, gender, race, and ethnicity obtained from the U.S. Census. This data provides a valuable tool for researchers and policymakers to identify and address disparities and inequalities in COVID-19 outcomes among various populations. To facilitate our analysis, we have chosen to focus exclusively on data points from the state of Texas within the U.S. COVID-19 Cases and Census Dataset. However, as we began to explore the raw dataset visualized in Figure 1, we quickly encountered significant discrepancies that required cleaning and refinement. Our team recognized the importance of thorough cleaning and refinement of the U.S. COVID-19 Cases and Census Dataset to ensure accurate and reliable analysis. As such, we embarked on a meticulous process to identify and eliminate any inconsistencies or errors in the data, which ultimately led to the exclusion of features that were missing entirely from our analysis. However, we encountered some unexpected issues in this process as well, which we will discuss in more detail in the following sections.

Texas County COVID-19 Vaccine Sites Dataset

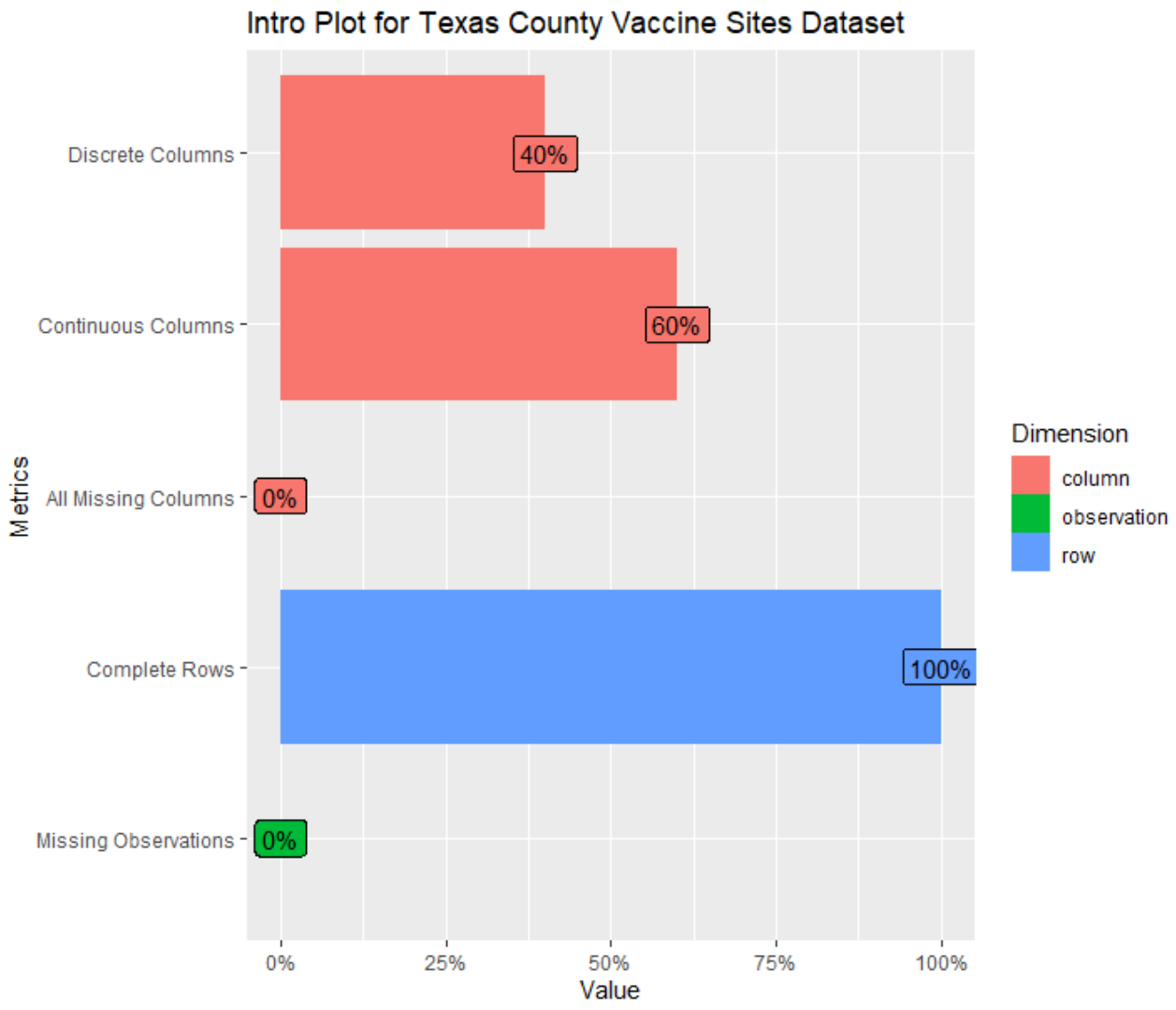

Figure 2. Intro Plot for Texas County Vaccine Sites Dataset.

Using the Texas County Vaccine dataset visualized in Figure 2 combined with the U.S. COVID-19 Cases and Census Dataset, we can apply clustering techniques to identify areas at risk in Texas. By analyzing the vaccine sites per thousand, the total number of sites per county, and the total population of each county, we can group counties with similar vaccination patterns and demographics. We can then compare this information with COVID-19 case data to identify areas at higher risk for infection and death. With this approach, we can gain valuable insights into which regions require more attention and resources to combat the pandemic in Texas. By using this vaccine data, we can identify areas at risk in Texas and develop targeted strategies to combat the spread of COVID-19.

Texas County Coordinates Dataset

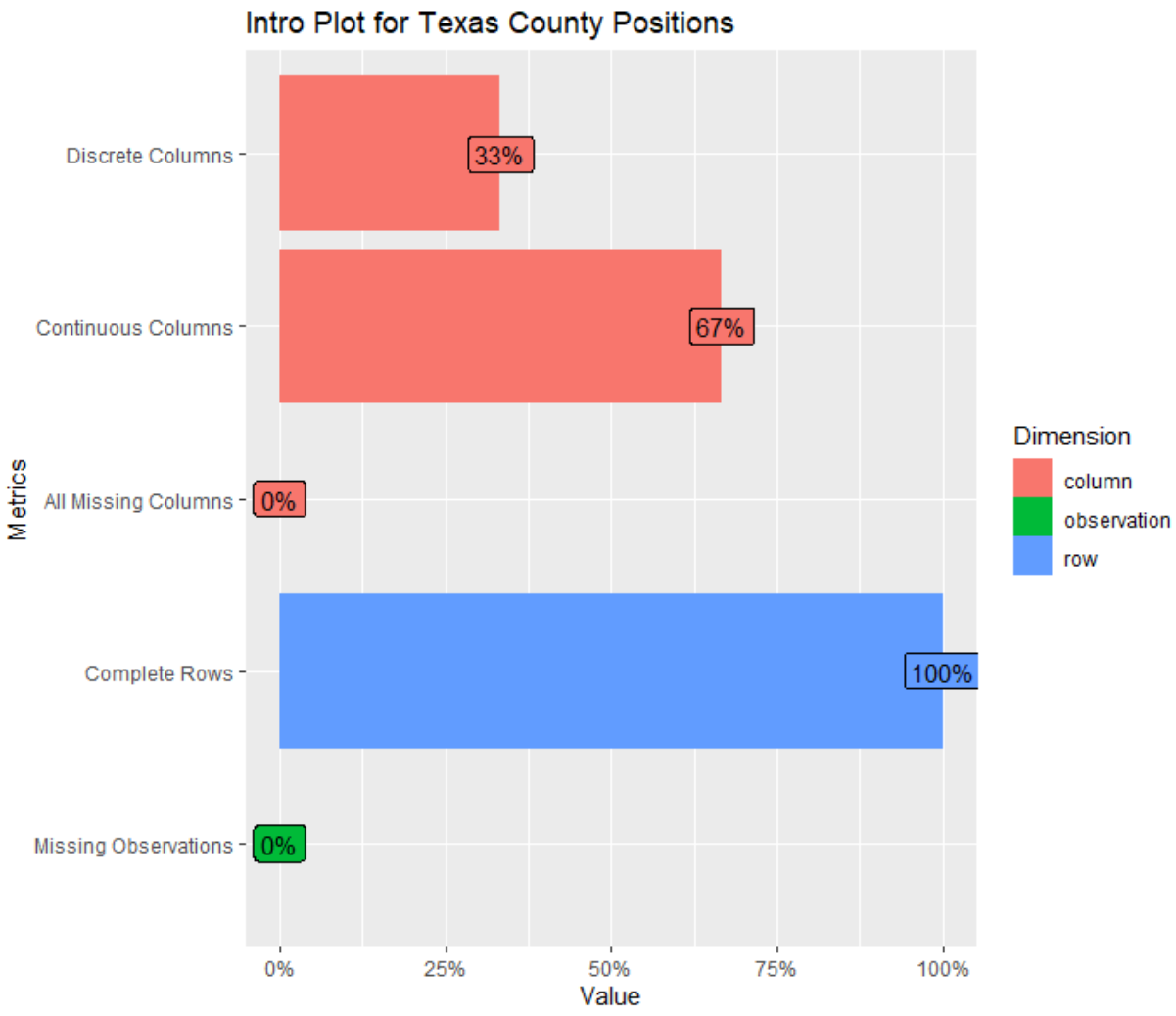

Figure 3. Intro Plot for Texas County Positions.

The Texas County Coordinates Dataset, visualized in Figure 3, is a vital resource containing latitude and longitude coordinates for all counties in the state of Texas. We leveraged this dataset to visualize our clustering data on a Texas map accurately. By combining the clustering results with the county coordinates, we created a visually informative representation of the data that demonstrated how various counties in Texas were grouped based on their characteristics. This approach provided us with valuable insights into the geographical distribution of the clusters and their correlation with different regions of the state. In this context, the Texas County Coordinates Dataset was a critical resource that enabled us to create comprehensive and visually appealing data visualizations. It is worth noting that this dataset contained all complete rows, providing us with a reliable and accurate dataset for our analysis.

Data Cleaning

Initially, we decided to focus on the census and vaccine datasets as they offered valuable information on each Texas county for effective clustering. Our team had the goal of performing clustering on datasets related to Texas counties, and we initially believed that combining the census and vaccine datasets would provide valuable information for this analysis. By merging these datasets, we could gain insight into the demographic and health-related characteristics of each county, resulting in meaningful and interpretable clusters. We transformed the vaccine dataset by converting categorical variables into character factors, scaling numeric values for normalization, and adding a county feature for merging datasets. For the census dataset, we filtered to only include Texas counties, removed and added columns, and calculated new features. However, in the process of removing these data points, we encountered an unexpected issue. The removal of these features led to the loss of some valuable information and insights, which could have potentially contributed to our analysis. As a result, we reevaluated our approach to account for this loss of information and ensure that our analysis remained as comprehensive and accurate as possible. Despite missing data for around 80 counties in the vaccine dataset, we initially decided to continue working with it. We believed that even though we were losing some information, the remaining data still held value and could potentially provide insights that could not be obtained by analyzing the datasets separately. Additionally, we felt we could not discard the vaccine dataset entirely as it provided valuable information about COVID-19. However, we found the results to be suboptimal and switched to a different approach outlined below.

We moved to only use the Census dataset for the dimensionality reduction, but even this ended up removing too many counties from the dataset. Our solution to this was to take the final feature dataset and then subset the original census data upon these features. We had 20 or so important features, with all the rows that we desired. This gave us a complete mapping of Texas, and with the cleaned, important features that we were looking for. After cleaning the Census Dataset, we then removed columns with missing features and computed the correlation matrix between the remaining numeric features. Highly correlated variables are removed from the dataset using a threshold of 0.95, and the correlation matrix was re-computed, yielding a much more workable amount of important features to work with. This dimensionality reduction allowed us to more easily analyze what features we wanted to cluster with, which we will get to later on. In Figure 4, we see a picture of this improved correlation matrix.

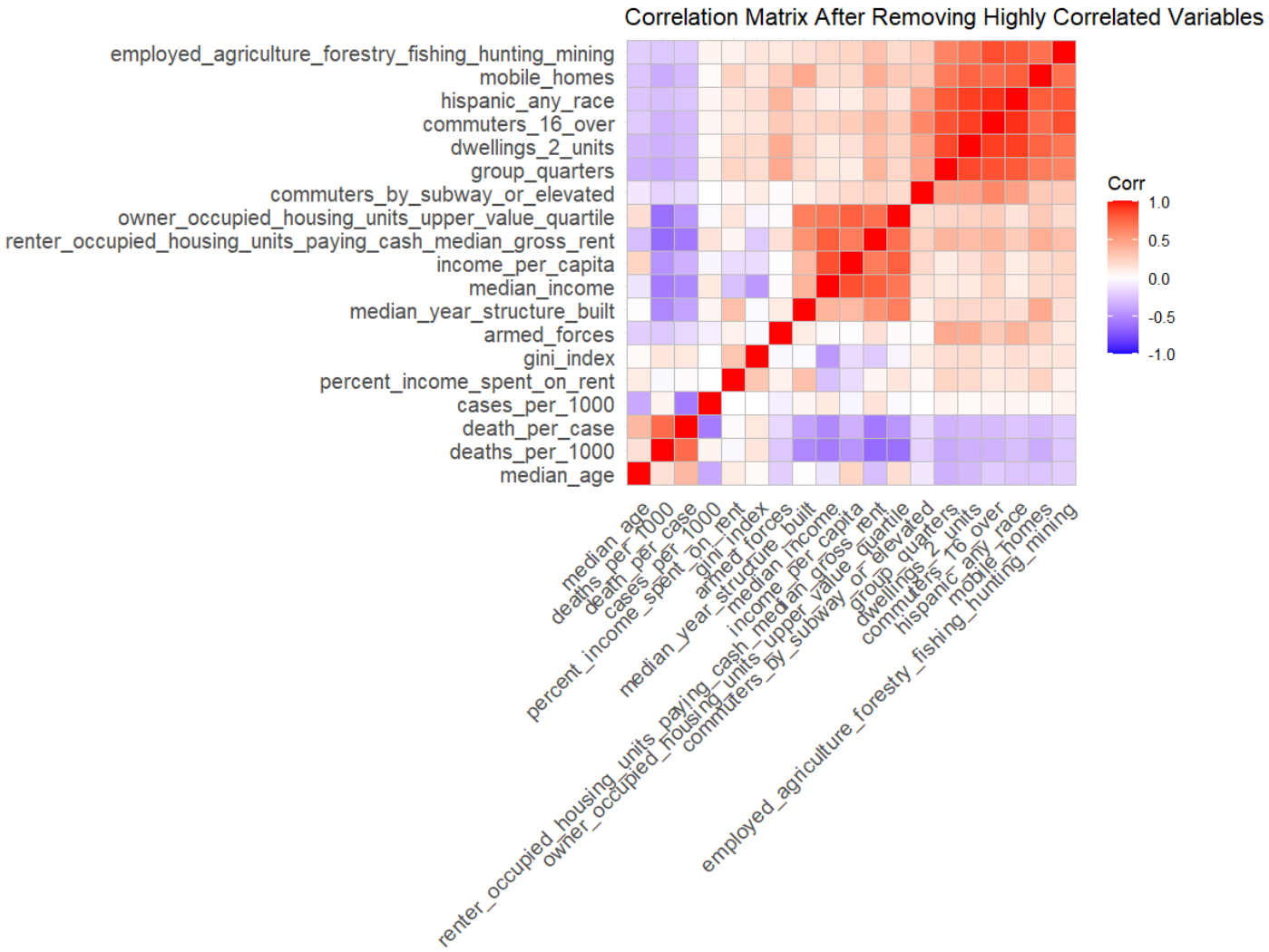

Figure 4. Correlation Matrix After Removing Highly Correlated Variables.

In order to gain a better understanding of the dataset and identify the most important features for clustering analysis, we performed Principal Component Analysis (PCA) on this reduced correlation matrix. The results of the PCA were visualized in a biplot, which allowed us to easily interpret the findings. It is important to note that the colors used in the biplot simply represent the strength of the vector and not any particular category or grouping of data. To determine the most important features in each region, we used cosine similarity analysis. The biplot showed that the variables in the dataset were clearly separated into four distinct quadrants. The fourth quadrant primarily contained discrete values, while the third quadrant contained data related to COVID-19 cases and deaths, which we had previously identified as important variables

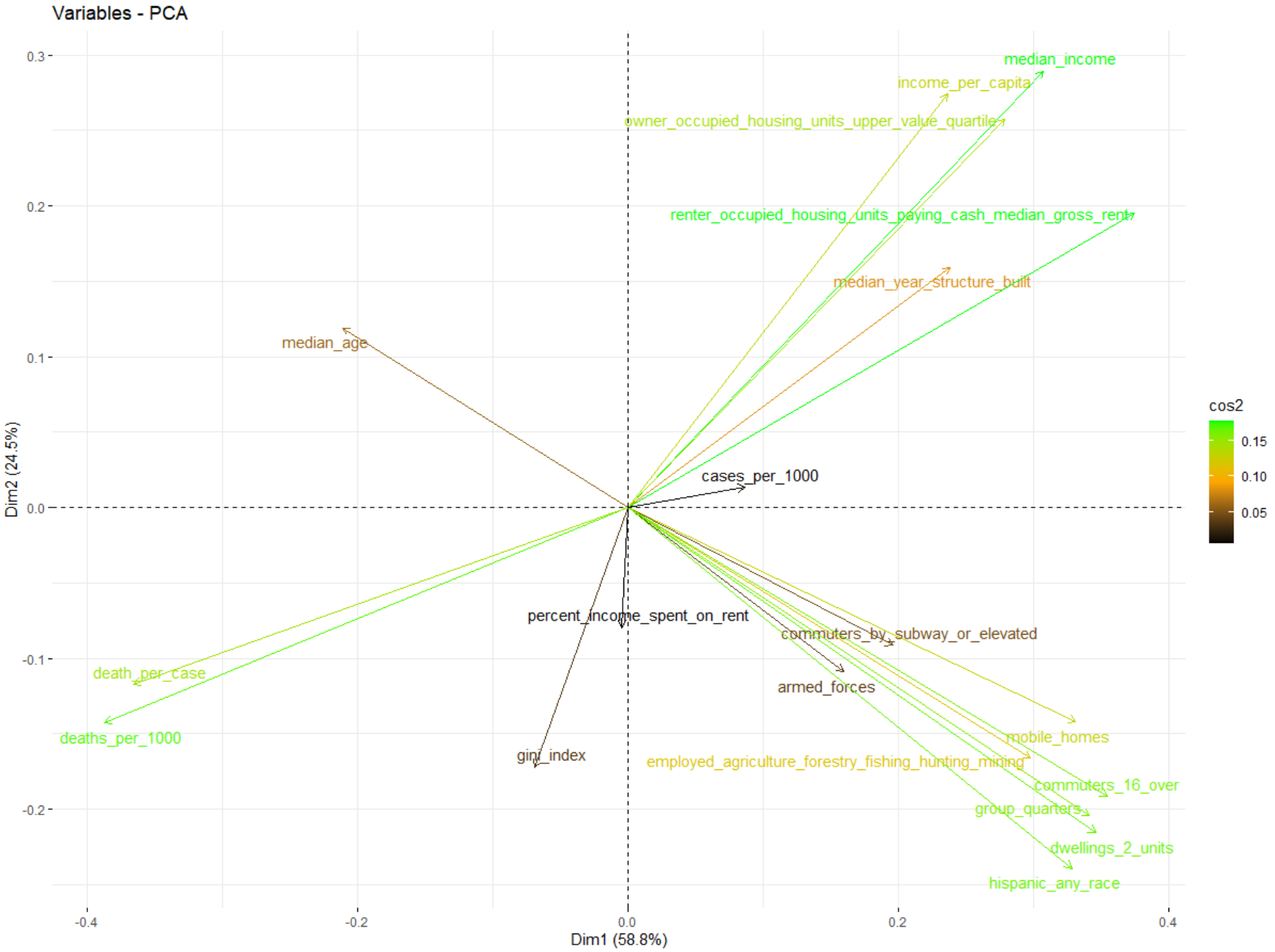

Figure 5. PCA Variables Plot.

According to the results of the PCA biplot analysis of the U.S. COVID-19 Cases and Census Dataset visualized in Figure 5, certain variables appear to be strongly correlated with each other and may be important predictors of COVID-19 outcomes. The variables in the first quadrant, which include income statistics such as median income and income per capita, as well as variables related to housing such as median year structure built and renter-occupied housing units paying cash median gross rent, show a strong positive correlation with other variables in the dataset, particularly those related to COVID-19 outcomes. This suggests that income and housing may play a role in determining COVID-19 outcomes, possibly through factors such as access to healthcare or the ability to practice social distancing.

The second quadrant of the biplot shows only one variable, median age, which is represented by a short brown ray. This suggests that median age is less strongly correlated with other variables in the dataset compared to the variables in the first quadrant. This does not necessarily mean that median age is not a predictor of COVID-19 outcomes, but rather that its relationship with other variables in the dataset may not be as strong. In the third quadrant, we see two very close long green rays, representing deaths per case and deaths per thousand. These variables are strongly correlated with each other and with other variables in the dataset, indicating that they may be important predictors of COVID-19 outcomes. Additionally, there are two somewhat separate short brown rays in this quadrant, representing sites per 1000 people and the Gini index. These variables are less strongly correlated with other variables in the dataset, suggesting that they may be less important predictors of COVID-19 outcomes. Finally, in the fourth quadrant, we see a cluster of variables represented by many close long green rays, including group quarters, total population, Hispanic any race, dwellings 2 units, mobile homes, and employment in agriculture. These variables are strongly correlated with each other and may be important predictors of COVID-19 outcomes in certain regions or populations.

Overall, the results of the PCA biplot analysis suggest that income statistics, housing, and certain demographic factors may be strong predictors of COVID-19 outcomes. Further research is needed to understand the specific mechanisms through which these variables affect COVID-19 outcomes and to develop effective interventions that address these factors. We then took this dataset on which we had performed dimensionality reduction and further removed outliers from it, which, as we discussed before, would end up being an issue. However, since we want to include our full process, we will discuss how we performed the dimensionality reduction as well.

Outlier Removal

Outliers can significantly affect the performance and results of clustering algorithms such as K-means, hierarchical clustering, and DBSCAN. These algorithms rely on the assumption that the data points being clustered are similar to each other and different from outliers. Therefore, outliers can distort the clustering results by pulling the centroids away from the dense areas of the data or causing the algorithm to partition the data incorrectly.

In K-means clustering, for example, outliers can be assigned to a cluster, causing the centroid of that cluster to shift and resulting in suboptimal clusters. Hierarchical clustering is also susceptible to outliers since it is based on the distance between data points, which can be affected by the presence of outliers. DBSCAN is more robust to outliers since it identifies dense areas of the data and ignores sparse regions. However, if an outlier is located in a dense area, it can significantly affect the clustering results.

By producing a box plot of these features, we determined that we had many outliers within our Texas data. Upon examining the box plot produced in Figure 6, we can see that there are a large number of outliers present in the data. Outliers are the data points that fall far outside the range of typical values and can skew our analysis.

As an exercise in outlier identification, we can examine two variables: the Gini index of each county and the number of deaths per thousand due to COVID-19. The Gini index is a measure of income inequality, with higher values indicating greater inequality, while the number of COVID-19 deaths per thousand provides insight into the severity of the pandemic in each county. Counties with high values for both variables may be considered outliers, as they are experiencing both high levels of income inequality and significant COVID-19 mortality rates.

To address this issue, we removed anything more than 3 standard deviations away from the mean. After removing these outliers, we can create a new box plot to visualize the updated data. The new plot visualized in Figure 7 shows the distribution of the 20 features with the outliers removed, allowing for a more accurate analysis of the data. By removing these outliers, we can obtain a better understanding of the true patterns and trends in the data.

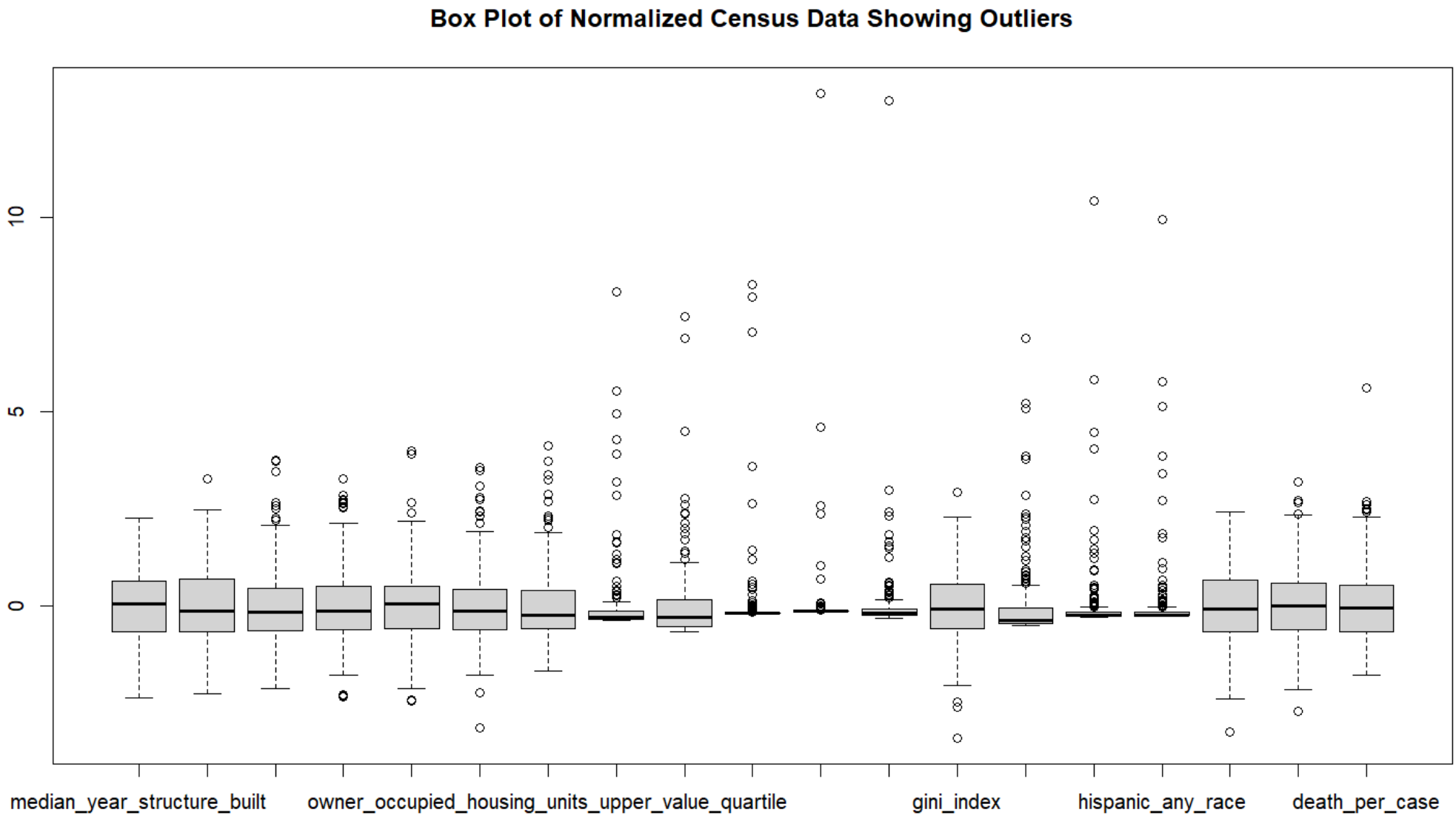

Figure 6. Box Plot of Normalized Census Data Showing Outliers.

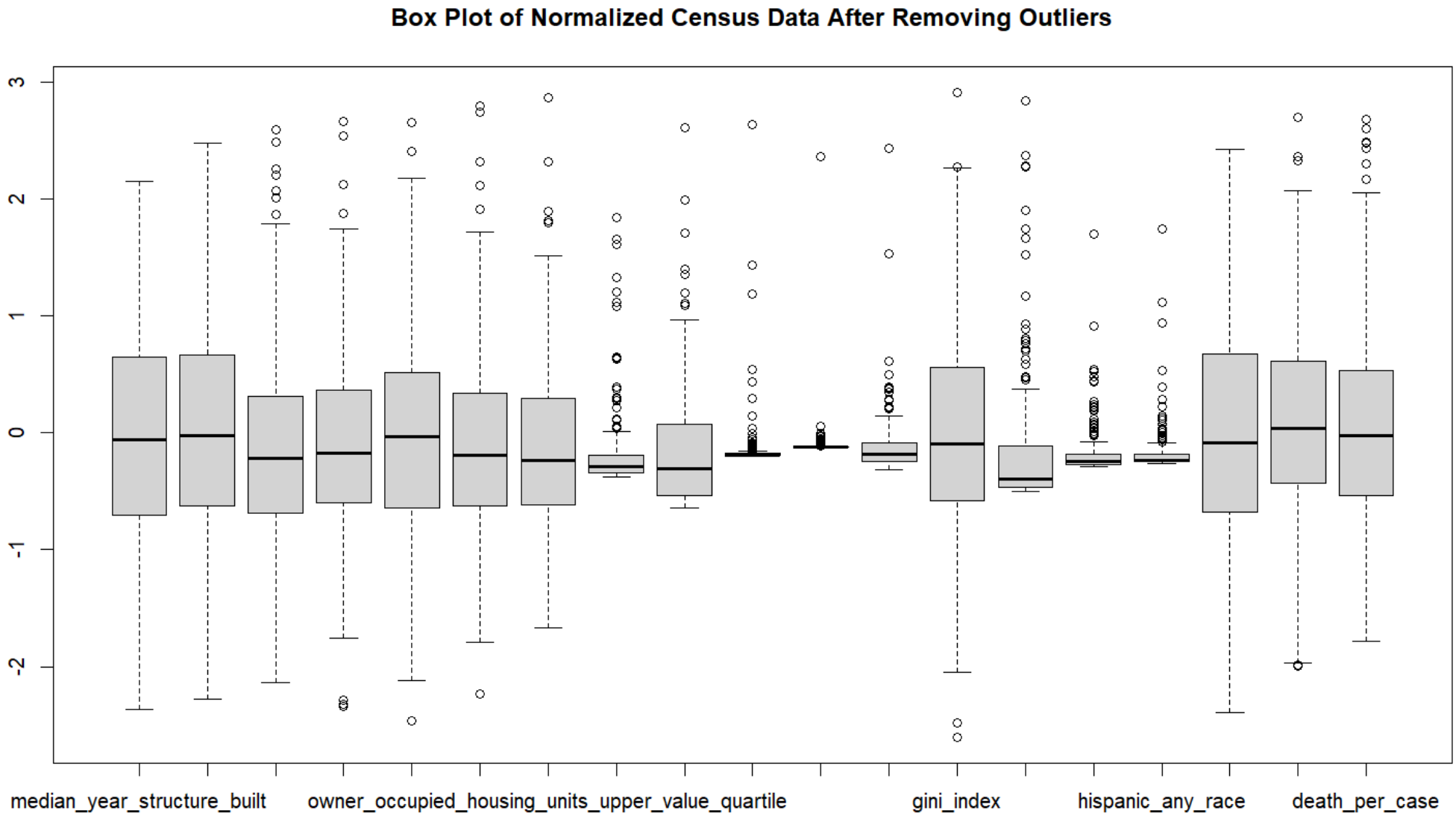

Figure 7. Box Plot of Normalized Census Data After Removing Outliers.

Having been through the spread of COVID-19, this feature selection seems to reflect those groups which we would expect to have been greatly affected by the actions taken to counteract the spread of the virus. For example, it is no surprise that commuters, low-income individuals, and those with a necessity to continue interacting with others during the height of the pandemic would be on this list since they would be most susceptible to COVID-19. Regarding data quality, there were a few problems presented, as visualized in Figure 8. All of the values needed to be numeric, and all null columns needed to be removed. Thus, we cast each of the values to be as such and removed unnecessary columns in the process. As such, we need not analyze each of the features, since they are all relatively similar and self-explanatory by their names alone.

We may also see some important statistics in Table 2 including the mean, median, standard deviation, minimum, and maximum for each of the important features. Since the numbers for each column vary drastically among different important features shown in Table 1 and having two rows with the same number for some columns provides no value for the purpose of this report, we decided against displaying mode. Further, we opted for displaying standard deviation rather than variance as the interpretability of the numbers for variance was difficult since the values were quite large. Finally, we included minimum and maximum rather than range since we can garner more information from the data with these statistics than we could from range. Additionally, the range for most of the features is a trivial calculation since the minimum was commonly 0.00. This feature selection allowed us to analyze the data more accurately.

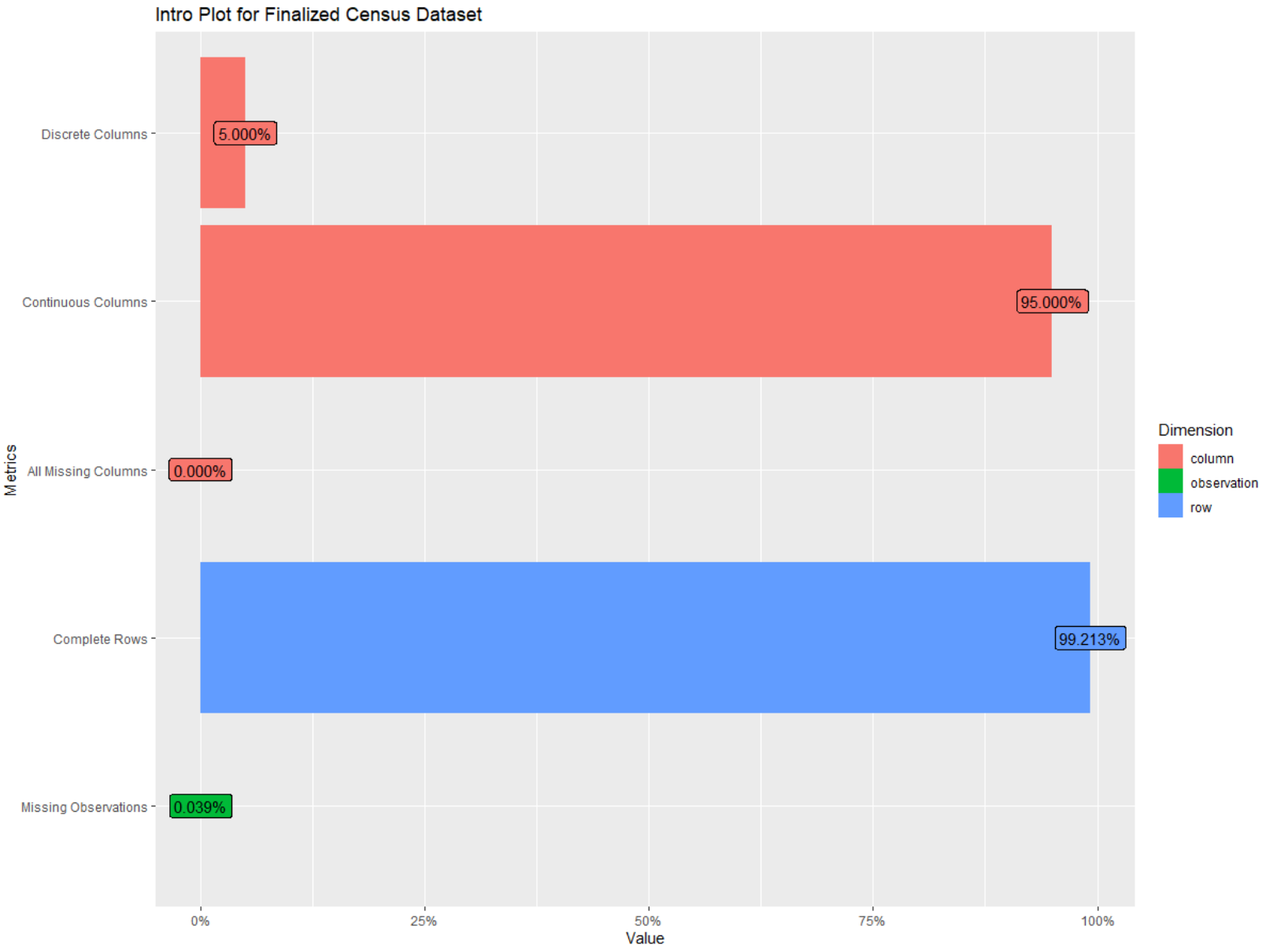

Figure 8. Intro Plot for Finalized Census Dataset.

Important Features

Feature	Scale	Dataset	Description
cases_per_1000	Ratio	Census Data	The number of COVID-19 cases per 1,000 people in a given county.
gini_index	Interval	Census Data	A summary measure of income inequality. 0 represents perfect equality, and 1 represents perfect inequality.
percent_income_spent_on_rent	Ratio	Census Data	The average percentage of income spent by households in a given county.
employed_agriculture_forestry_fishing_hunting_mining	Ratio	Census Data	The number of people employed in the agriculture, forestry, fishing, hunting, or mining industries.
mobile_homes	Ratio	Census Data	The number of mobile homes registered in a county.
hispanic_any_race	Ratio	Census Data	A count of individuals that identify as Hispanic.
dwellings_2_units	Ratio	Census Data	The number of properties with 2 housing units.
group_quarters	Ratio	Census Data	A count of individuals living in group quarters, like college dorms or military barracks.
commuters_by_subway_or_elevated	Ratio	Census Data	A count of individuals who commute by subway or elevated train.
owner_occupied_housing_units_upper_value_quartile	Ratio	Census Data	A count of non-rented housing units that fall within the upper-value quartile within a given county.
renter_occupied_housing_units_paying_cash_median_rent	Ratio	Census Data	The median gross rent paid by renters in a county.
income_per_capita	Ratio	Census Data	The mean income is computed for every individual in a given county.
median_income	Ratio	Census Data	The median income for individuals in a county.
median_year_structure_built	Interval	Census Data	The median construction year in a county.
armed_forces	Ratio	Census Data	A count of individuals in the armed forces.
death_per_case	Ratio	Census Data	A ratio of deaths to COVID-19 cases.
deaths_per_1000	Ratio	Census Data	A count of deaths per 1000 individuals in a county.
median_age	Ratio	Census Data	The median age of individuals in a county.

Table 1. Features Selected For Clustering From Data Sets. All Features are in Euclidean Distance Measures.

Summary Statistics

Feature	Mean	Std. Dev	Min	Pctl. 25	Pctl. 75	Max
median_age	-0.0028	0.95	-2.3	-0.63	0.66	2.5
median_income	-0.086	0.86	-2.1	-0.69	0.31	2.6
income_per_capita	-0.1	0.86	-2.3	-0.6	0.37	2.7
percent_income_spent_on_rent	-0.051	0.94	-2.5	-0.64	0.51	2.7
renter_occupied_housing_units_paying_cash_median_gross_rent	-0.11	0.83	-2.2	-0.62	0.34	2.8
owner_occupied_housing_units_upper_value_quartile	-0.13	0.79	-1.7	-0.62	0.29	2.9
dwellings_2_units	-0.17	0.36	-0.38	-0.34	-0.2	1.8
mobile_homes	-0.15	0.53	-0.65	-0.54	0.074	2.6
armed_forces	-0.14	0.27	-0.19	-0.19	-0.18	2.6
commuters_by_subway_or_elevated	-0.11	0.18	-0.13	-0.13	-0.13	2.4
employed_agriculture_forestry_fishing_hunting_mining	-0.12	0.28	-0.32	-0.25	-0.09	2.4
gini_index	-0.0094	0.98	-2.6	-0.58	0.56	2.9
group_quarters	-0.15	0.58	-0.5	-0.47	-0.11	2.8
commuters_16_over	-0.17	0.22	-0.29	-0.27	-0.19	1.7
hispanic_any_race	-0.17	0.22	-0.26	-0.25	-0.19	1.7
cases_per_1000	0.016	0.99	-2.4	-0.68	0.67	2.4
deaths_per_1000	0.088	0.9	-2	-0.43	0.61	2.7
death_per_case	0.055	0.9	-1.8	-0.54	0.53	2.7

Table 2. Statistical Summary of Important Features.

Data Preparation

Initial Analysis

For the purposes of this report, we focused on cases in the state of Texas. As of the latest update, the state has reported over 5 million cases of COVID-19 and over 75,000 deaths. The dataset provides data at the county level, allowing for a more granular analysis of the pandemic's impact on different regions of the state. At the county level, Harris County, which includes Houston, has reported the highest number of COVID-19 cases and deaths in the state, with over 1 million confirmed cases and over 10,000 deaths. Other counties with high case and death counts include Dallas County, Bexar County, and Tarrant County, which include the cities of Dallas, San Antonio, and Fort Worth, respectively. Prior to performing our clustering analysis, we expect these counties to be in the same cluster. To simplify our analysis, we numbered each of the feature subsets shown in Table 3. Each subset contains different features that may be helpful in clustering the counties of Texas. Table 3 demonstrating the features in each subset can be found below:

Feature Set Number	Features Included
Initial Analysis	• Median Income • Median Age • Percent Income Spent on Rent
Set 1	• Renter occupied housing units paying cash median gross rent • Median age • Percent income spent on rent
Set 2	• Income Per Capita • Owner occupied housing units upper-value quartile • Gini index

Table 3. Features Explored In Report.

Modeling

We explored various clustering techniques on different subsets of the dataset and ultimately utilized three methods: K-means clustering, hierarchical clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

K-means clustering partitions the data into k clusters based on similarity. The algorithm starts by randomly selecting k initial centroids, assigns each data point to the nearest centroid, computes new centroids based on the mean of the points assigned to each cluster, and repeats the process until convergence. The goal is to minimize the distance between data points and their assigned centroid.

Hierarchical clustering creates a hierarchy of clusters based on the similarity between data points. We used agglomerative hierarchical clustering which starts with each data point in its own cluster and repeatedly merges the two closest clusters until only one cluster remains. It is simpler and more intuitive than divisive clustering, which starts with all the data points in one cluster and recursively splits it into smaller clusters until each data point is in its own cluster. Agglomerative clustering is also computationally efficient, making it well-suited for our large datasets.

DBSCAN is a clustering method that groups data points that are close to each other in space and are part of a high-density region. The algorithm defines a radius around each data point and identifies neighboring points within that radius. Points that are close together and have a high density are grouped together, while isolated points are considered noise.

Initial Analysis: Income, Age, Income For Rent

Our initial subset consists of demographic features that can help classify the socioeconomic status of a county. Median income, median age, and the average percentage of income spent on rent are indicators that can provide a general idea of a county's economic status. Counties with high median incomes, high median ages, and low average percentages of income spent on rent can be clustered together as having a high socioeconomic status. This can give us valuable insight into the economic and demographic factors that contribute to a county's socioeconomic status and help us understand the needs and challenges faced by residents in different areas. To cluster the first feature subset, we utilized K-means clustering with a k-value of 4. This allowed us to look for 4 clusters across the state of Texas. The resulting clusters are visualized below:

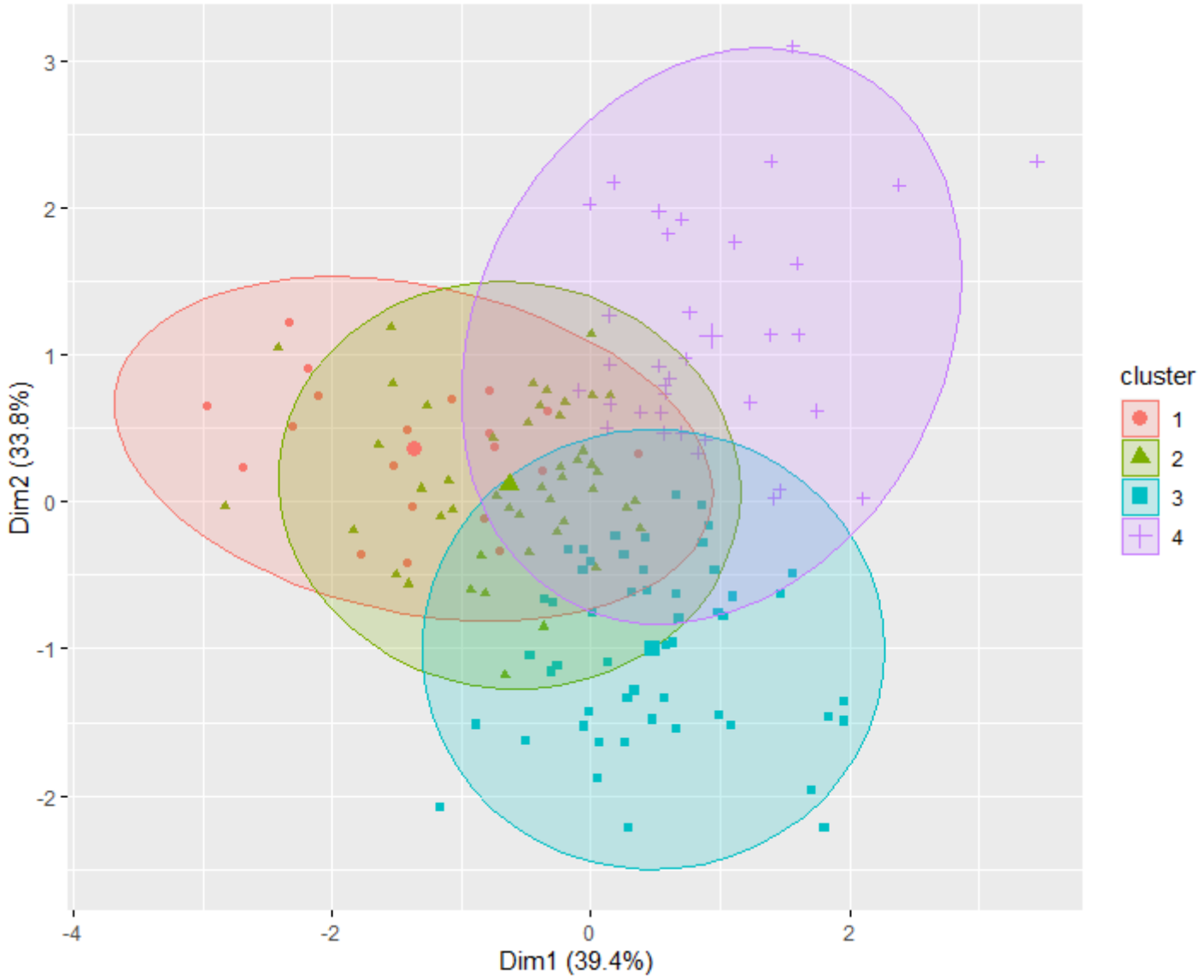

Figure 9. Initial Subset - KMeans Clustering

Figure 9 shows the K-Means Dimension Plot for our Initial Set, which is not clearly clustered. We observed significant overlap between each cluster, and all four clusters covered a relatively large area. In such situations, assigning data points to specific clusters can be difficult, leading to overlapping clusters. K-means clustering has limitations, including sensitivity to the initial cluster centers and the number of clusters. When clusters overlap significantly, trying different clustering algorithms, such as hierarchical clustering or density-based clustering, can be helpful, as we will perform later in this report. It's important to note that the quality of clustering depends on both the algorithm and the dataset. Sometimes, the dataset itself may not have clear and distinct clusters, resulting in overlapping clusters. After joining our cluster data with the Texas Coordinates dataset, we

were able to associate each cluster with its corresponding county and visualize the results on a map. However, it's worth noting that some counties did not report data or were outliers, and these are represented as grey on the map. To ensure that all counties were included in the visualization, we filled in any missing county data with N/A. The resulting map is shown in Figure 10.

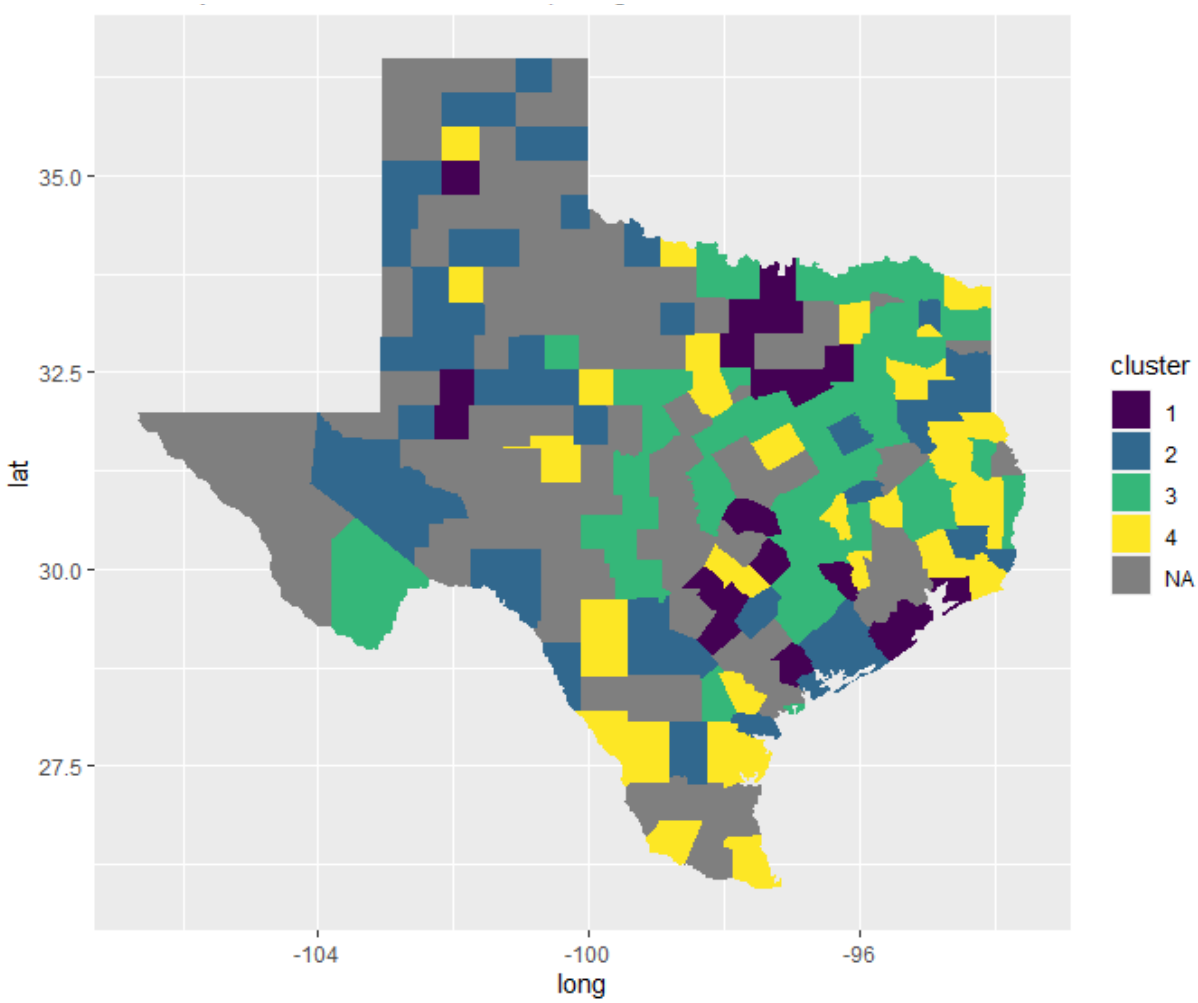

Figure 10. Initial Subset - KMeans Clusters Shown on Texas County Map.

The graph depicting the clustering results is not an ideal representation. A large portion of the Texas counties are greyed out, indicating that they either did not report any data or were classified as outliers. This is not surprising as many counties in West and South Texas have small populations.

However, what is concerning is that most of the major counties are also greyed out, **and for this reason we removed outliers**, as discussed in Outlier Removal. It is possible that these counties have similar feature values, but their size or population is causing them to be treated as outliers. K-means clustering relies heavily on the distance between data points to determine cluster membership. Therefore, outliers can significantly impact the distance metric and lead to separate

cluster assignments or being classified as noise. It may be necessary to explore other clustering algorithms, such as hierarchical or density-based clustering, to address this issue. Additionally, it is important to note that the quality of clustering is dependent on both the algorithm and the dataset, and in some cases, overlapping clusters may be inevitable if the dataset lacks distinct clusters.

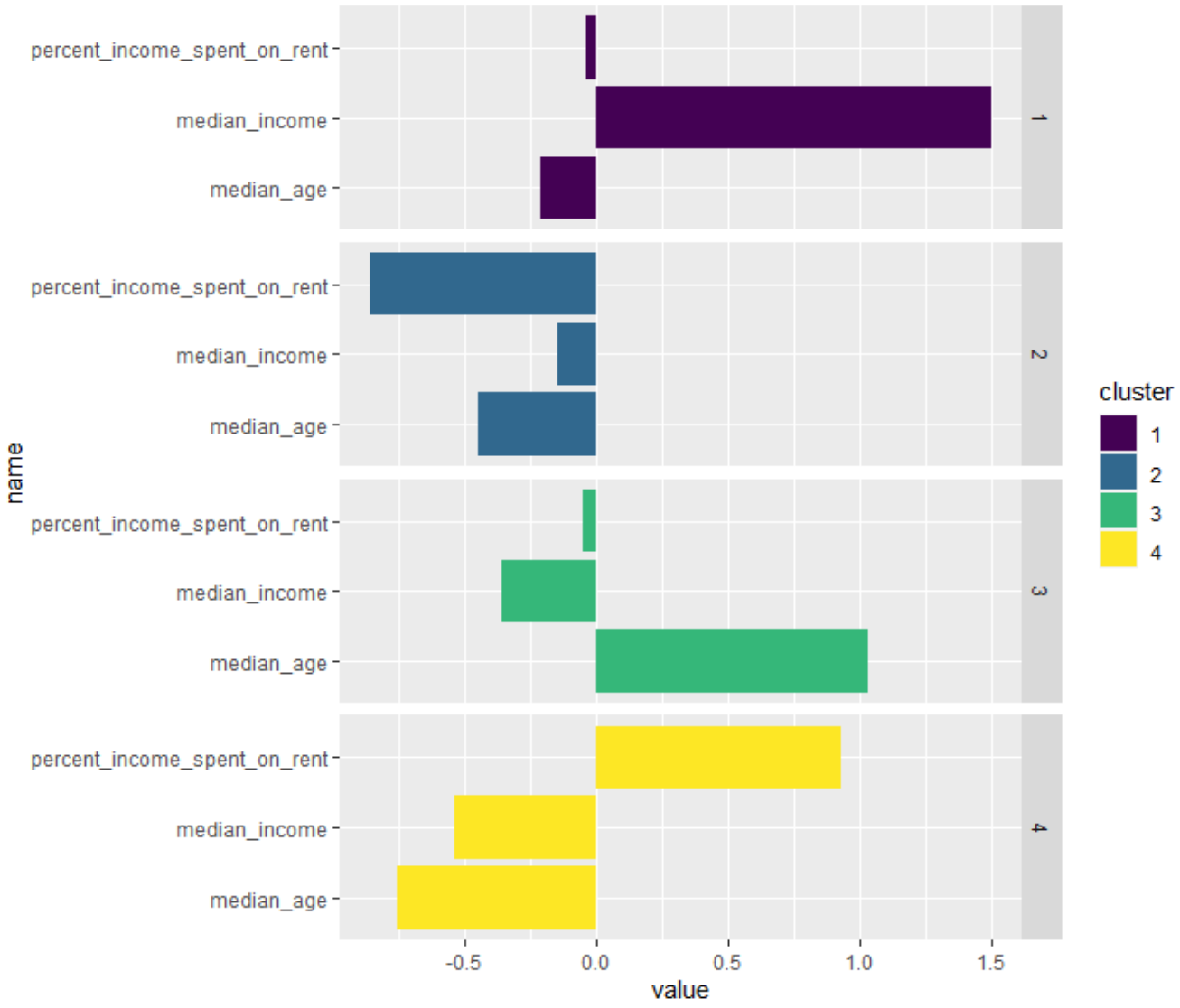

Figure 11. Initial Subset - KMeans Cluster Profiles.

Figure 11 shows the values of the initial subset for each of the clusters. Using this graph, we can get a better understanding of what types of counties are in each cluster. Cluster 1 has a low percentage of income spent on rent, a very high median income, and a low median age. When looking at the map, these counties are primarily suburbs of large cities or medium-sized cities. Just about every Dallas-Fort Worth, Austin, San Antonio, and Houston suburb county belongs to this cluster, as well as medium-sized cities in west Texas like Amarillo and the Midland–Odessa Metroplex. These areas tend

to have higher median incomes due to their proximity to urban job centers and amenities. The low percentage of income spent on rent in this cluster could also be reflective of the fact that housing costs tend to be higher in urban and suburban areas, which means that even though residents in these areas have higher incomes, they may not necessarily be spending a larger proportion of their income on rent. The presence of medium-sized cities in West Texas is interesting, as it suggests that these areas may be experiencing economic growth and attracting a younger, affluent population. The high median income and low median age in this cluster could be reflective of the strong job market in the energy industry in West Texas, as well as the growing healthcare and technology sectors.

Cluster 2 shows a very low percentage of income spent on rent, a low median income, and a low median age. Some possible candidates include rural or economically disadvantaged counties with lower housing costs and a younger population. Most of the counties in Cluster 2 are located in West Texas, with a few in South and East Texas, and in the Houston area, this suggests that the cluster is capturing a distinct geographical region in Texas that has certain socioeconomic characteristics. West Texas is known for its oil and gas industry and has a large rural population. The low median income and low percentage of income spent on rent in this cluster could be reflective of these characteristics. In addition, the presence of counties in South and East Texas could suggest that the cluster is also capturing areas with a similar rural and economically disadvantaged population. The few counties in the Houston area could be due to several reasons. One possibility is that they are located on the outskirts of the city and have a more rural population with lower incomes and housing costs. Another possibility is that they are areas with a high percentage of renters who spend a low proportion of their income on rent.

Cluster 3 shows a low percentage of income spent on rent, a low median income, but a high median age. These counties are primarily located in Northeast and Central Texas, avoiding major metro areas and their suburbs. One possible explanation for this clustering pattern is that the population in these counties may consist of retirees or individuals approaching retirement age who have limited income but own their homes outright or have lower housing costs due to long-term ownership. Furthermore, these counties' locations outside major metro areas and their suburbs may suggest that residents have chosen to live in more rural or less densely populated areas to enjoy a quieter lifestyle or to be closer to family members. The observation that the counties in Cluster 3 are "drive-through counties" suggests that they are located along major highways or thoroughfares, which may have contributed to their low percentage of income spent on rent and potentially their high median age as well. When counties are situated along highways, they may attract more transient or temporary populations, such as truck drivers or other travelers, who may not reside in these areas permanently. As a result, the demand for housing in these counties may be lower than in more densely populated areas, leading to lower housing costs and a lower percentage of income spent on rent.

Finally, Cluster 4 shows a very high percentage of income spent on rent, a fairly low median income, and a fairly low median age. These counties seem to be mostly in East and South Texas. These

counties may be located in areas with a higher cost of living, such as in or near major metropolitan areas or regions with high demand for rental properties. This, combined with a lower median income, may make it difficult for residents to afford housing, leading to a higher percentage of income spent on rent. The lower median age observed in these counties may reflect the presence of younger, working-age individuals who are more likely to rent their homes rather than own them. Moreover, the location of these counties in East and South Texas may reflect the impact of regional economic and social factors. For example, East Texas has a history of agricultural production and manufacturing, while South Texas has a strong energy and petrochemical industry. These industries may draw a younger, working-age population, but these counties may have limited economic opportunities, leading younger individuals to move elsewhere for work, whereas those who remain may struggle to make ends meet.

Due to the dataset missing data from about half of the counties in Texas, the set can be considered bad because it can lead to biased or incomplete results. If the missing data is not missing at random, it can affect the representativeness of the sample and make it difficult to generalize the findings to the larger population. To address this issue, it is often necessary to create better feature subsets by carefully selecting the variables that are most relevant to the research question and ensuring that they are complete for all observations. In this case, we decided to not move forward with this set and instead to create two better feature subsets that contain complete data for all counties in Texas, which will improve the quality and reliability of the analysis.

Subset One: Median Gross Rent, Median Age, Income Spent on Rent

Our initial subset of features, which includes Median Gross Rent, Median Age, and Income Spent on Rent, is particularly relevant for our analysis as it captures important demographic and socioeconomic factors that may influence the transmission and severity of COVID-19. These variables provide valuable information about a county's socioeconomic and demographic profile, including the availability and affordability of housing, the age structure of the population, and the economic resources and financial stability of households. Together, they can give an indication of the overall level of economic development, health risk factors, and access to affordable housing in a county. Clustering counties based on these variables, as visualized in Figure 12 and Figure 13, can help us identify patterns and subgroups that may be more or less vulnerable to the virus. This information can be used to inform targeted public health interventions and resource allocation efforts.

Figure 12. Subset One Pairs Plot.

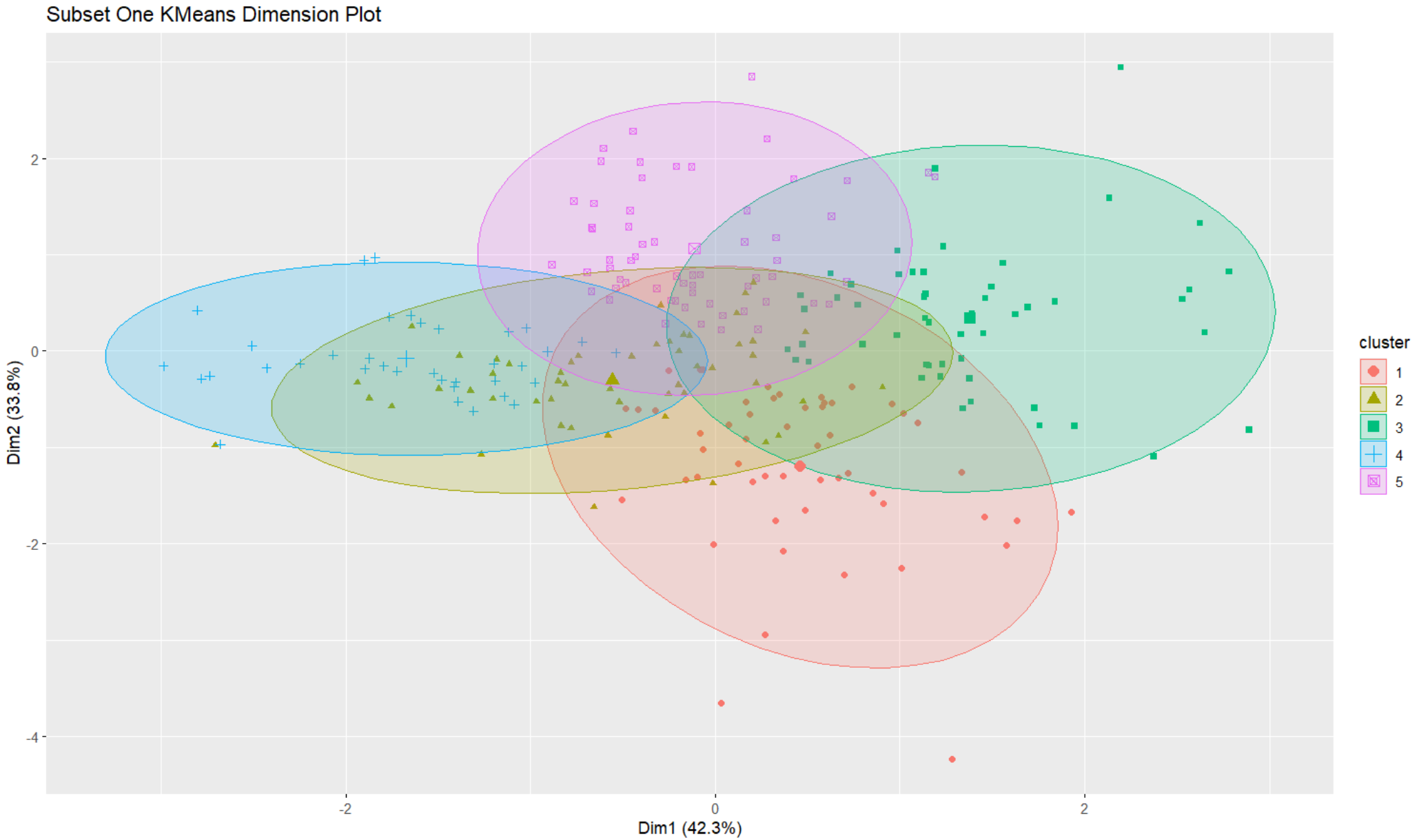

Figure 13. Subset One KMeans Dimension Plot.

Our K-means algorithm determined that the optimal number of clusters was five, which showed a noticeable improvement compared to our initial cluster set, even though there were still some overlaps among them shown in Figure 13.

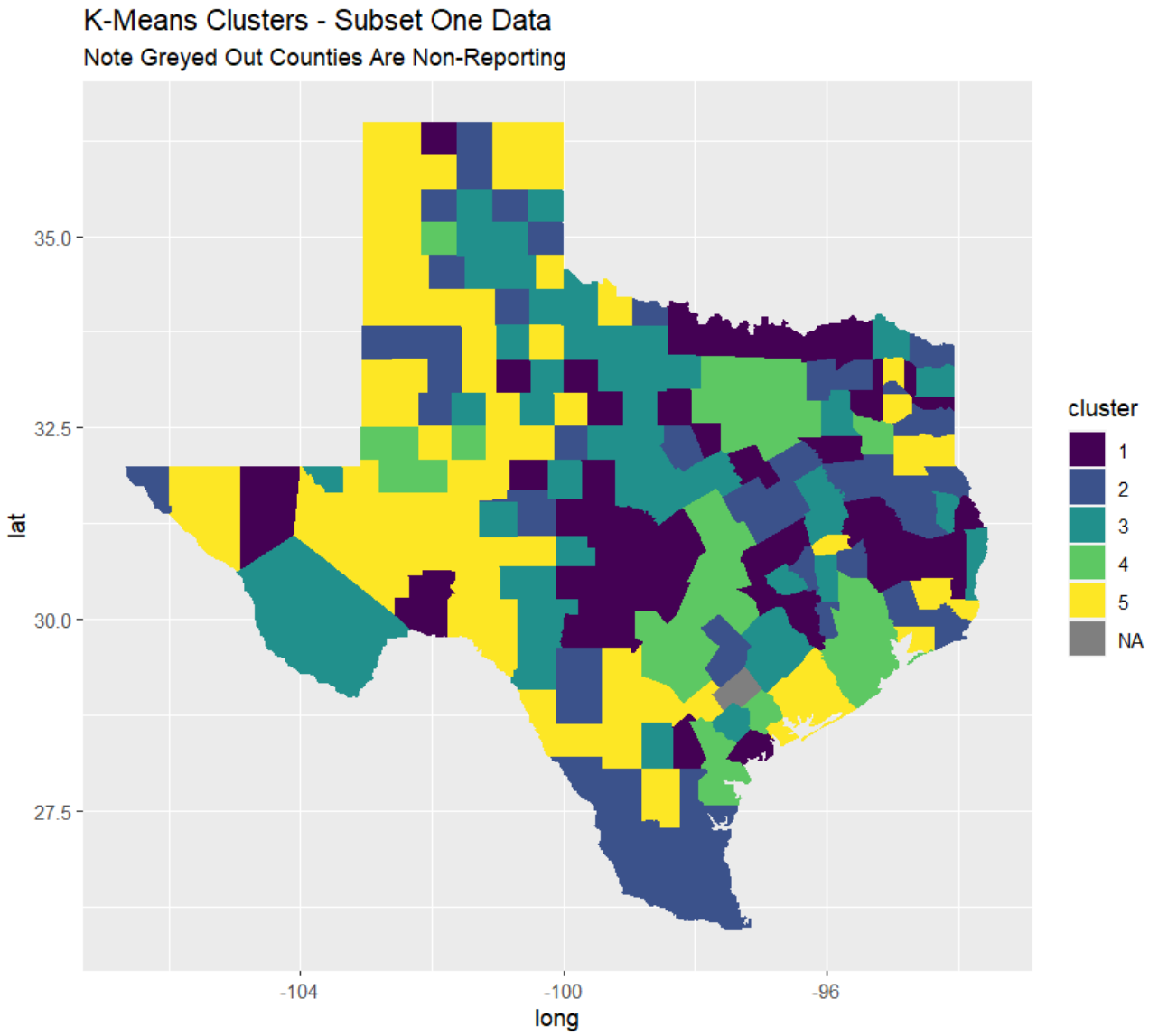

Figure 14. K-Means Clusters - Subset One Data.

This map of Texas in Figure 14 is significantly improved as almost all counties have reported data, and the clusters on the map appear to be logical. Figure 15 shows the K-means cluster profiles of each of our 5 clusters. The profiles align with our previous analysis.

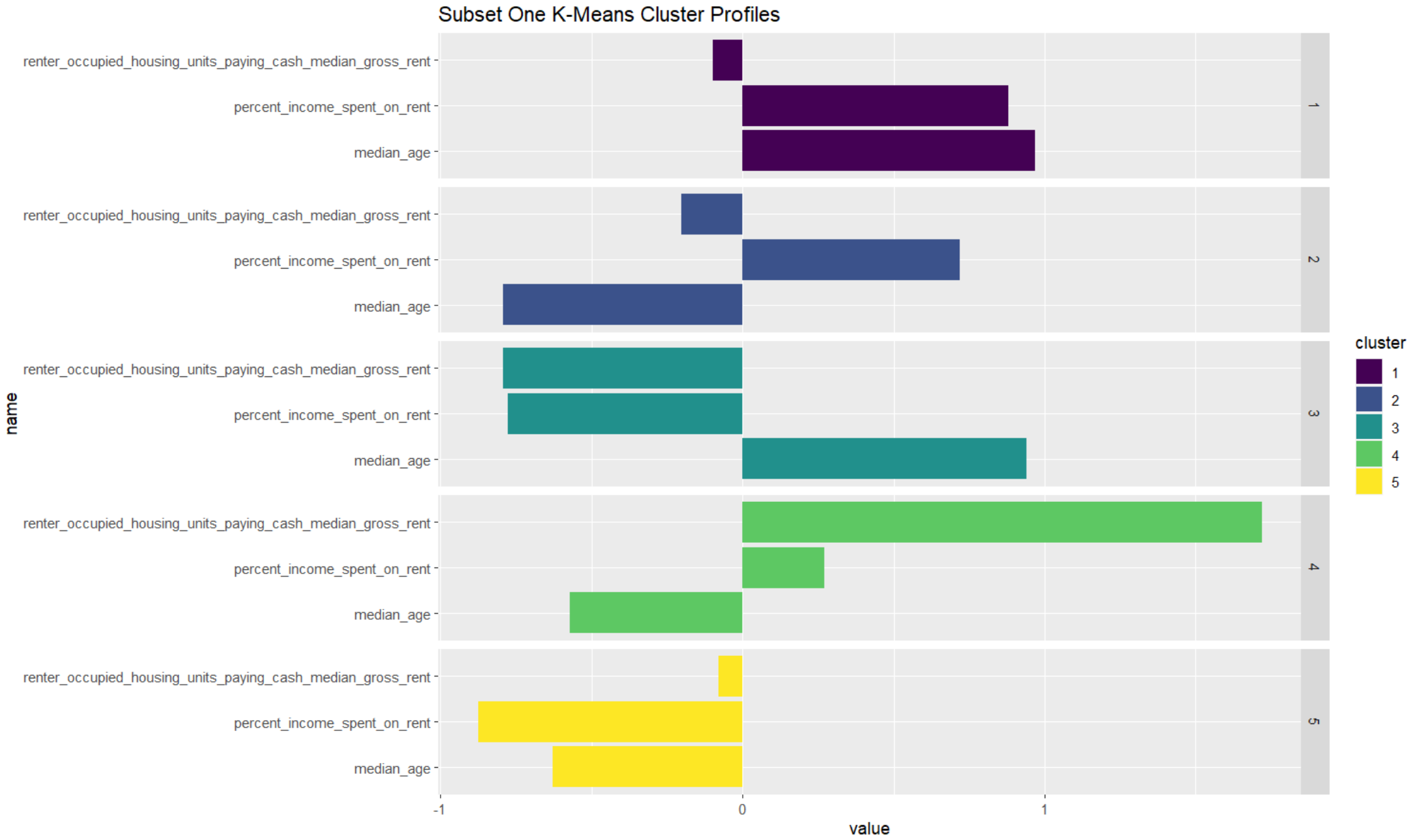

Figure 15. Subset One K-Means Cluster Profile

Cluster One has a slightly lower cash median gross rent, a high percentage of income spent on rent, and a high median age. The high percentage of income spent on rent in Cluster One is likely due to the fact that the population in these rural regions may have lower overall incomes. With lower income, a higher percentage of income may be spent on basic necessities such as housing, resulting in a higher percentage of income being spent on rent. The high median age in this cluster may be due to a number of factors. One potential factor is the lack of job opportunities in rural areas, leading to a younger population leaving for employment elsewhere. Additionally, many of these rural regions may be attractive to retirees seeking a quieter lifestyle away from major urban centers.

Cluster Two has a slightly lower cash median gross rent, a high percentage of income spent on rent, and a low median age. The variability in socioeconomic characteristics across these counties may have contributed to the formation of this cluster. The slightly lower cash median gross rent in this cluster may be due to the fact that many of the counties in this region are less urbanized and have lower overall costs of living. This may lead to lower rent prices in these areas. Additionally, the lower cost of living may attract younger individuals who are just starting out in their careers. The high percentage of income spent on rent in Cluster Two may be due to a variety of factors. One potential factor is that the population in these counties may have lower overall incomes, leading to a higher percentage of income being spent on housing. Additionally, some of the counties in this cluster may have a mix of urban and rural areas, leading to higher housing costs in urbanized areas. The low median age in this cluster may be due to a variety of factors. The lower cost of living and the mix of urban and rural areas may attract younger individuals looking for affordable housing options. Additionally, some of the counties in this cluster may have a higher concentration of colleges and universities, leading to a younger population.

Cluster Three has a very low cash median gross rent, a very low percentage of income spent on rent, and a high median age. This cluster represents the suburbs of suburban counties, which are located around the major urban centers in Texas, such as Houston, Dallas, and San Antonio. The very low cash median gross rent in this cluster may be due to the fact that these suburban areas may have lower overall costs of living compared to the major urban centers they surround. Additionally, the lower rent prices may be attractive to retirees looking to downsize their living arrangements. The very low percentage of income spent on rent in Cluster Three may be due to the fact that the population in these suburban areas may have higher overall incomes. This may result from the proximity to major urban centers, where many of these individuals may work and earn higher salaries. The high median age in this cluster may be due to the fact that these suburban areas may be attractive to retirees looking to downsize their living arrangements. Additionally, these suburban areas may have a lower demand for housing from younger individuals due to the proximity to the major urban centers.

Cluster Four has a very high cash median gross rent, a high percentage of income spent on rent, and a low median age. This cluster represents the major metro areas of Texas and their suburbs, including Dallas-Fort Worth, Houston, Austin, and San Antonio. The very high cash median gross rent in this cluster may be due to the high demand for housing in major urban centers, resulting in higher rent prices. Additionally, the proximity to major employment centers and cultural attractions may make these areas attractive to individuals willing to pay a premium for housing. The high percentage of income spent on rent in Cluster Four may be due to the fact that the population in these areas may have higher overall incomes but also face higher living costs. Additionally, the proximity to major urban centers may result in a higher cost of living and higher housing costs in these areas. The low median age in this cluster may be due to the fact that major urban centers tend to attract younger individuals seeking employment opportunities and cultural experiences. Additionally, the proximity to major colleges and universities may contribute to a younger population.

Finally, Cluster 5 has a low cash median gross rent, a very low percentage of income spent on rent, and a low median age. This cluster represents rural areas in West Texas and South Texas, which are generally less densely populated than other areas of the state. The low cash median gross rent in this cluster may be due to the lower demand for housing in rural areas compared to urban centers. Additionally, the lower overall cost of living in rural areas may contribute to lower rent prices. The very low percentage of income spent on rent in Cluster Five may be due to the fact that the population in these rural areas may have lower overall incomes. However, the lower cost of living in rural areas may offset the lower incomes and result in a lower percentage of income spent on rent. The low median age in this cluster may be due to the fact that younger individuals may be attracted to the lower cost of living and rural lifestyle. Additionally, many rural areas may have a higher concentration of jobs in industries such as agriculture and oil and gas, which may be attractive to younger individuals seeking employment opportunities.

We also performed hierarchical clustering on Feature Set 1. We found that 4 clusters worked best. Figures 16 - 18 below are the resulting clustering plot and map of Texas counties.

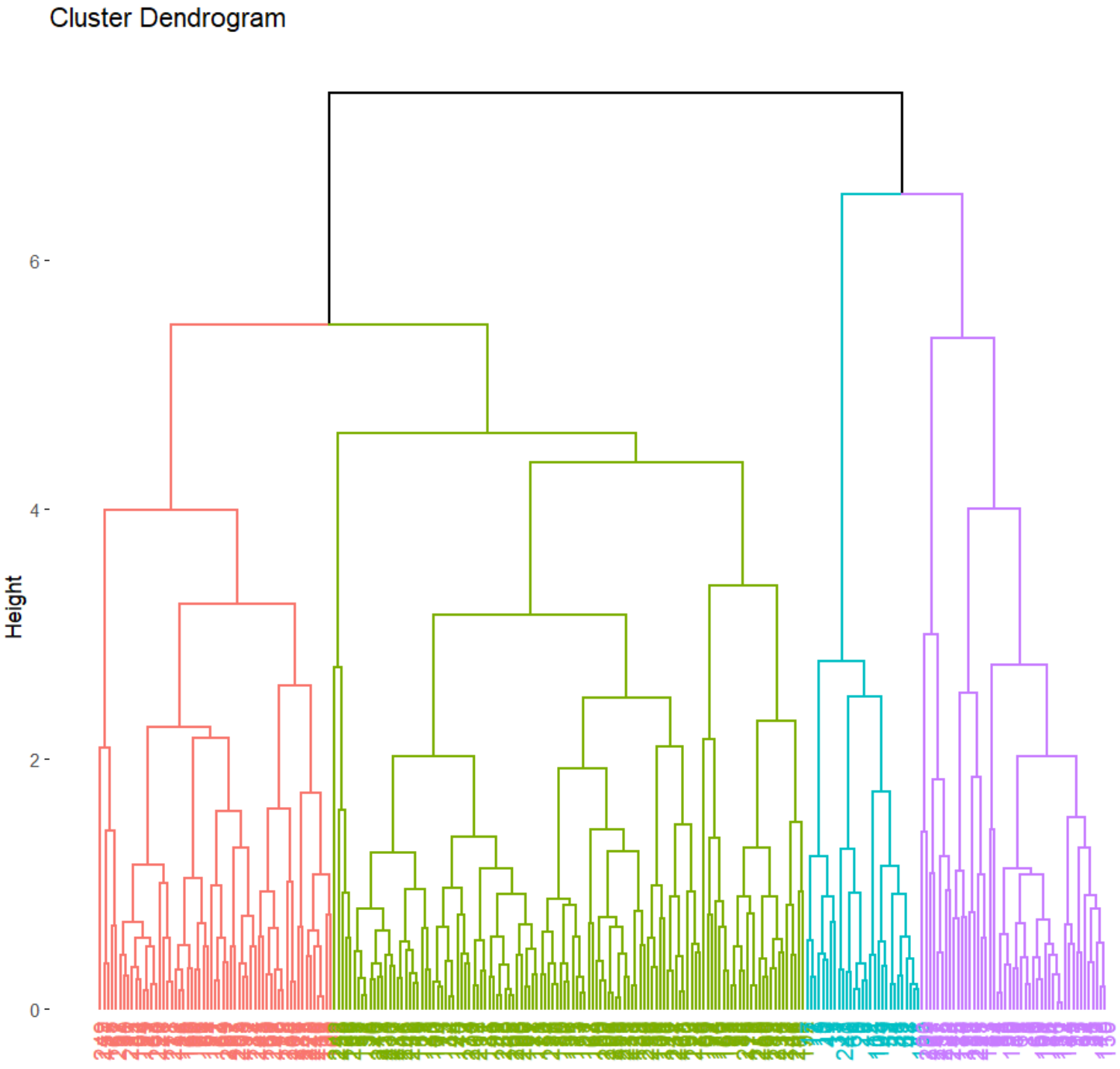

Figure 16. Subset One - Hierarchical Clustering Dendrogram.

Figure 17. Subset One - Hierarchical Clustering Plot.

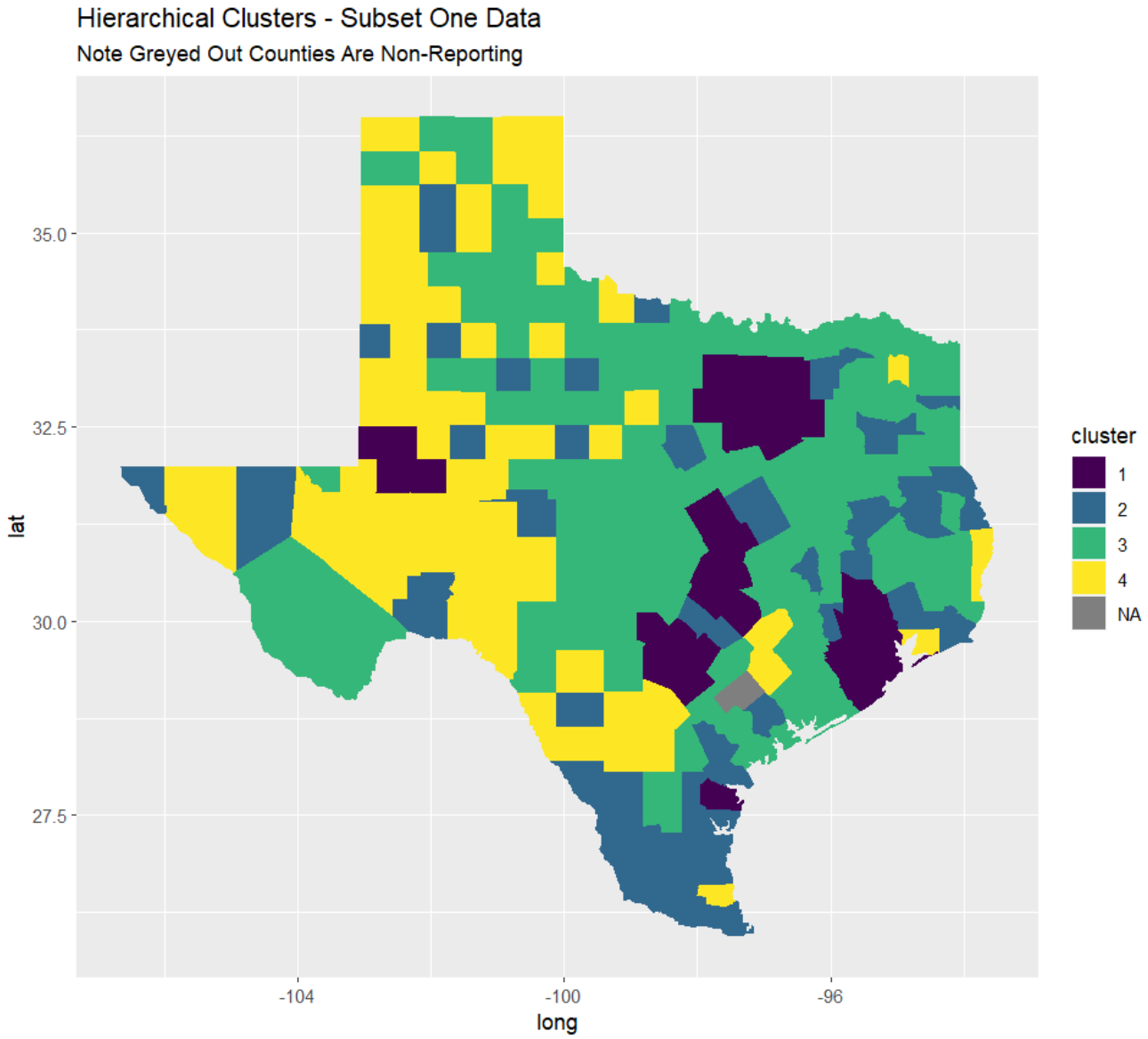

Figure 18. Subset One - Hierarchical Clustering.

Although the clusters may be numbered differently, the hierarchical clustering method found nearly the same clusters as K-means. Cluster 1 of hierarchical corresponds with Cluster 4 of K-means, the major metro areas, and their suburbs. Cluster 2 corresponds with Cluster 2 of K-means, rural counties, and outer suburbs. Cluster 3 corresponds with both Cluster 1 and Cluster 2 of K-means, the rural counties between major metro areas. Finally, Cluster 4 corresponds with Cluster 5 of K-means, the rural counties in West Texas. An analysis of these clusters proves to be similar to that of K-means when taking into account differing numberings. Finally, we performed DBSCAN on this subset. Our data set proved to be too similar for the algorithm to find more than one cluster.

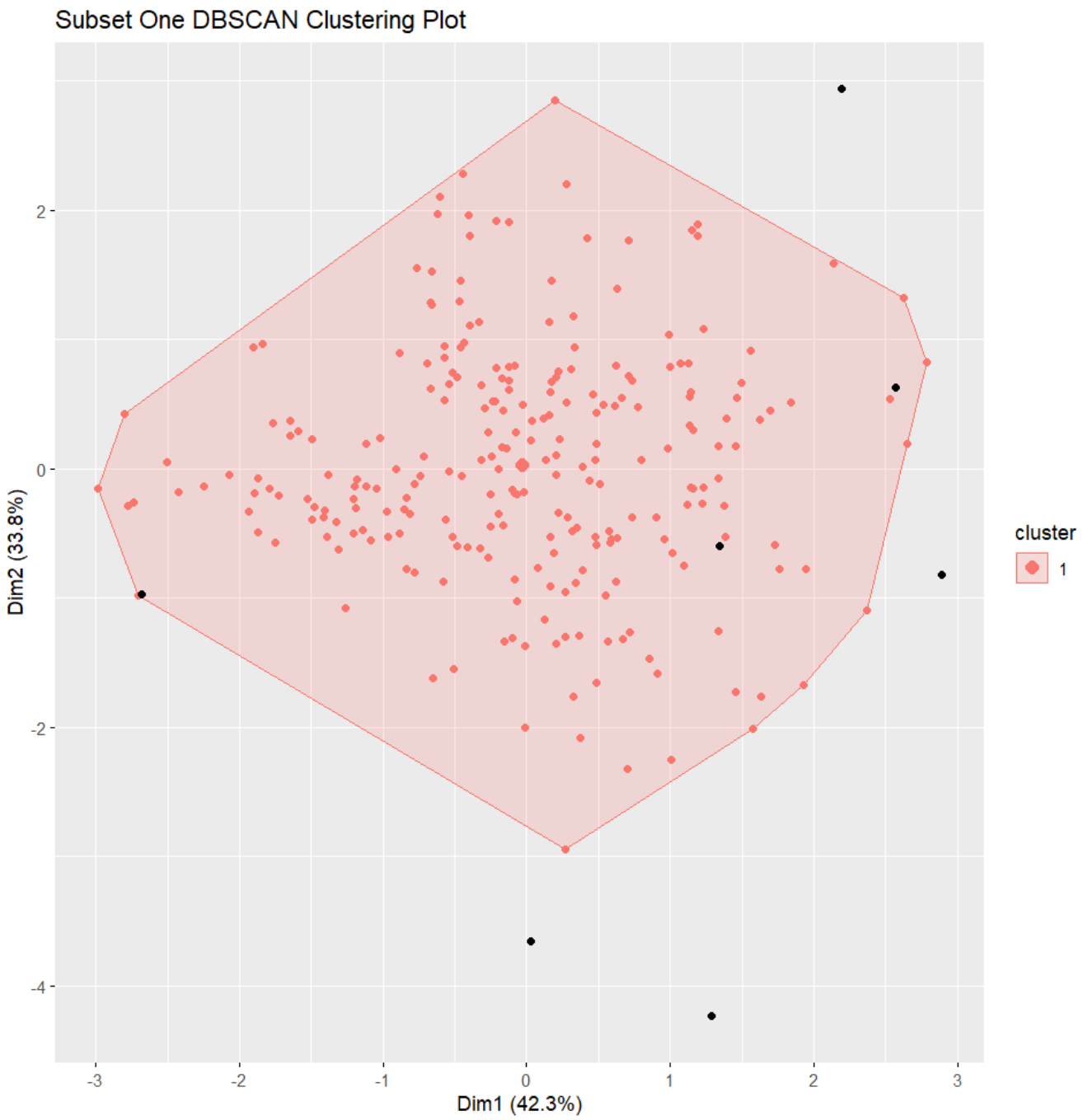

Figure 19. Subset One - DBSCAN Clustering Plot.

One potential explanation for this result is that the features you have selected are not highly correlated with each other, and thus do not provide a strong basis for clustering. For example, while Median Gross Rent and Income spent on rent may be related, Median age may not have a strong relationship with either of these features. Without a clear relationship or pattern among the features, as shown in Figure 19, DBSCAN may not be able to identify meaningful clusters.

Subset One: Internal Validation

To ensure that the number of clusters we selected was appropriate for our analysis, we employed several internal validation techniques. We present the corresponding graphs below in Figures 20-24. Overall, we see our results generally align with the clusterings we picked and when the internal validation methods disagreed, we took the average and used the resulting number of centers.

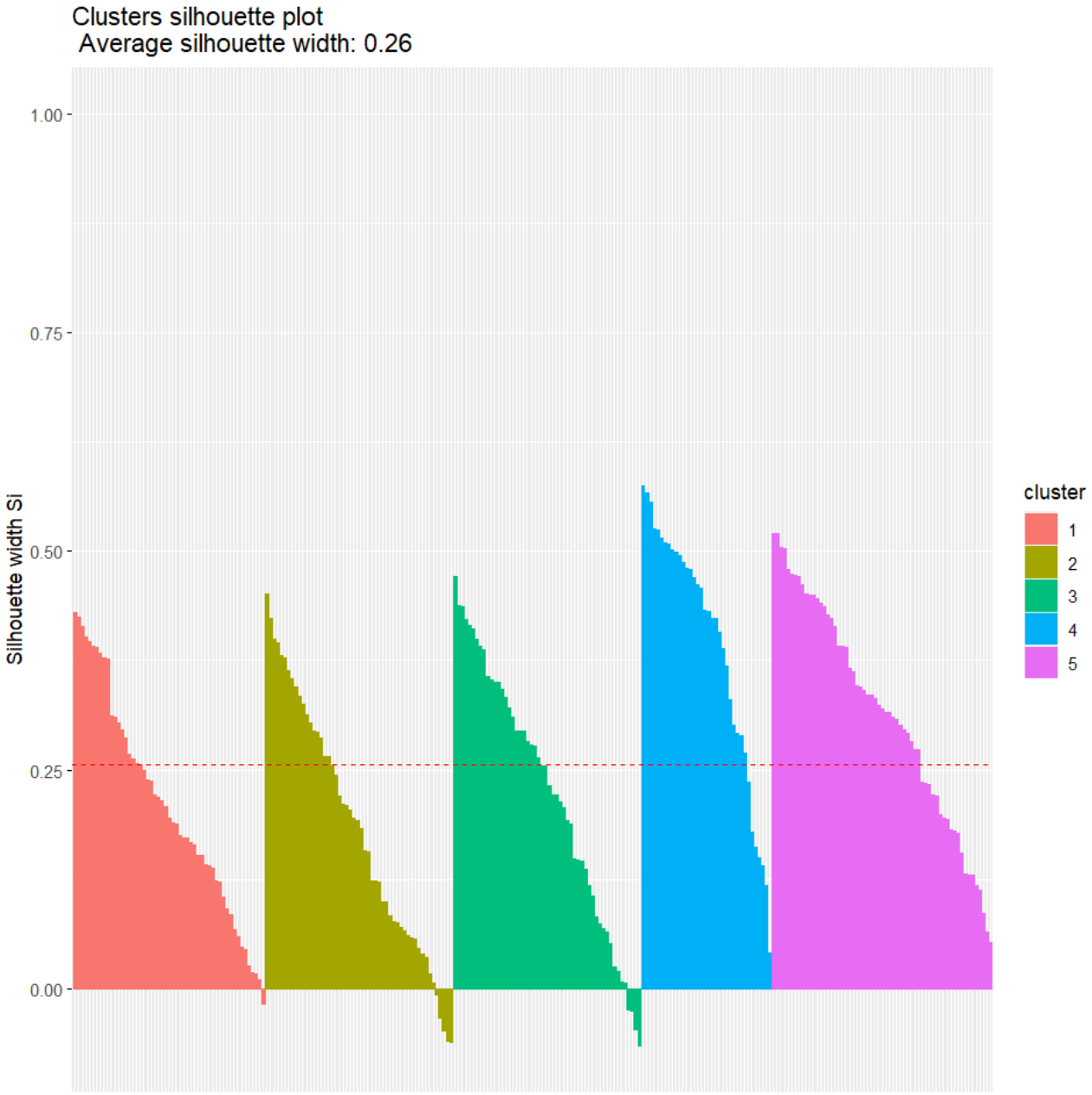

Figure 20. Subset One - Clusters Silhouette Plot.

A silhouette plot, as shown in Figure 20, is a visual representation of the silhouette coefficient for each data point in a cluster analysis. The silhouette coefficient is a measure of how similar an object is to its own cluster compared to other clusters. The coefficient ranges from -1 to 1, where a value of 1 indicates that the object is well-matched to its own cluster, and a value of -1 indicates that it is more similar to other clusters. In a silhouette plot, each data point is represented by a vertical line or bar, with the height of the bar indicating the silhouette coefficient for that point. The bars are arranged by cluster, with each cluster assigned a different color or pattern. The width of the bars is proportional to the number of data points in each cluster. The silhouette plot allows us to visualize the quality of clustering by examining the distribution of silhouette coefficients for each cluster. A good clustering solution will have high silhouette coefficients for all data points, indicating that each point is well-matched to its own cluster and poorly matched to other clusters. Conversely, a poor clustering solution will have low silhouette coefficients, indicating that the data points are not well-clustered and could potentially belong to multiple clusters. As can be seen in our silhouette plot above, the coefficients generally fall within the desired clusters, with some lower coefficients in some places. Overall, this essentially tells us that our clustering was nothing special.

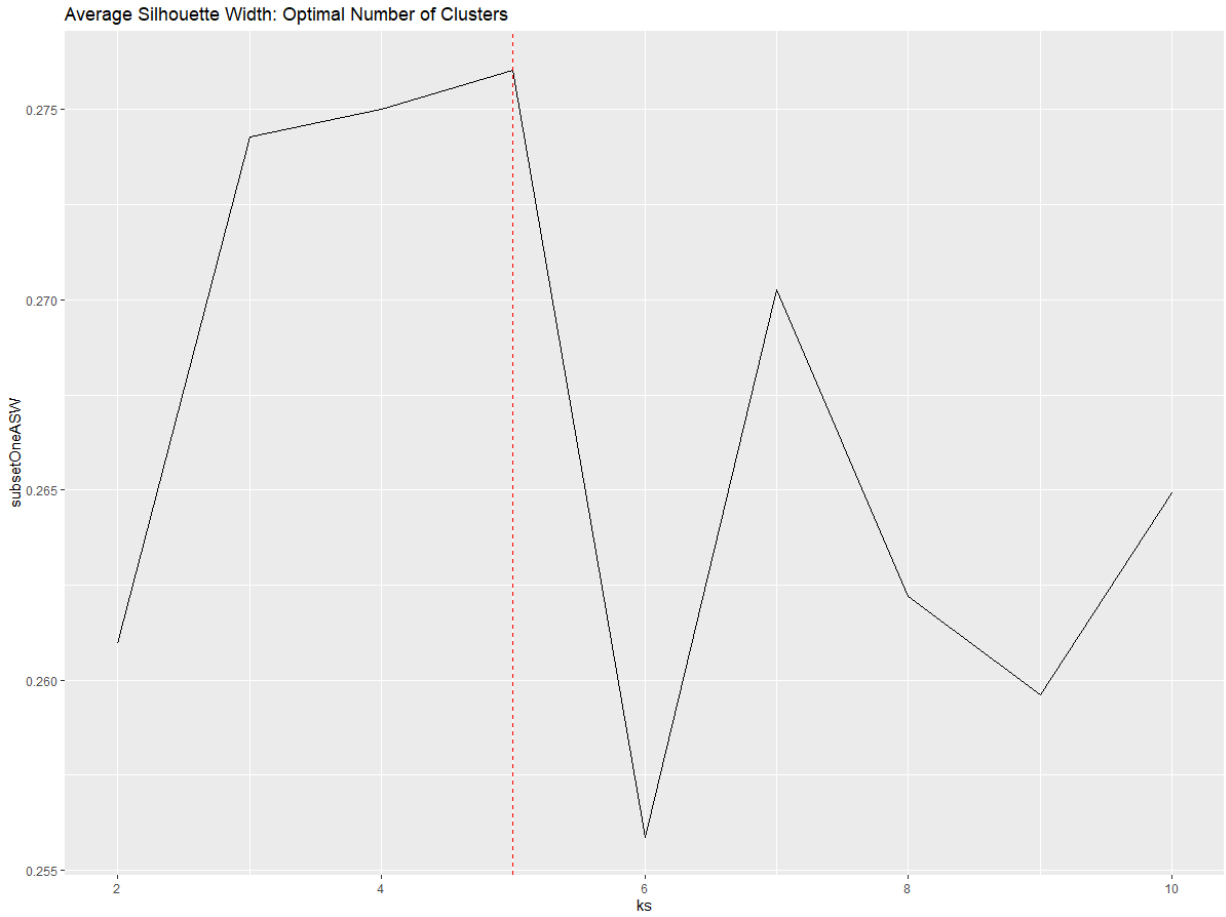

Figure 21. Subset One - Average Silhouette Width - Optimal Number of Clusters.

The average silhouette width, as shown in Figure 21, is a metric used to evaluate the quality of clustering, particularly in K-means clustering. It measures how similar a data point is to its own cluster compared to other clusters. The silhouette width ranges from -1 to 1, with a higher value indicating a better-defined cluster. To determine the optimal number of clusters, we can plot the average silhouette width against the number of clusters. The optimal number of clusters is usually the one that maximizes the average silhouette width, indicating that the clusters are well-separated and distinct. A sharp increase in the average silhouette width followed by a plateau or a decrease suggests that additional clusters are not meaningful and that the clustering algorithm should use the previous number of clusters. Therefore, the optimal number of clusters can be identified by examining the plot of the average silhouette width and selecting the number of clusters with the highest value.

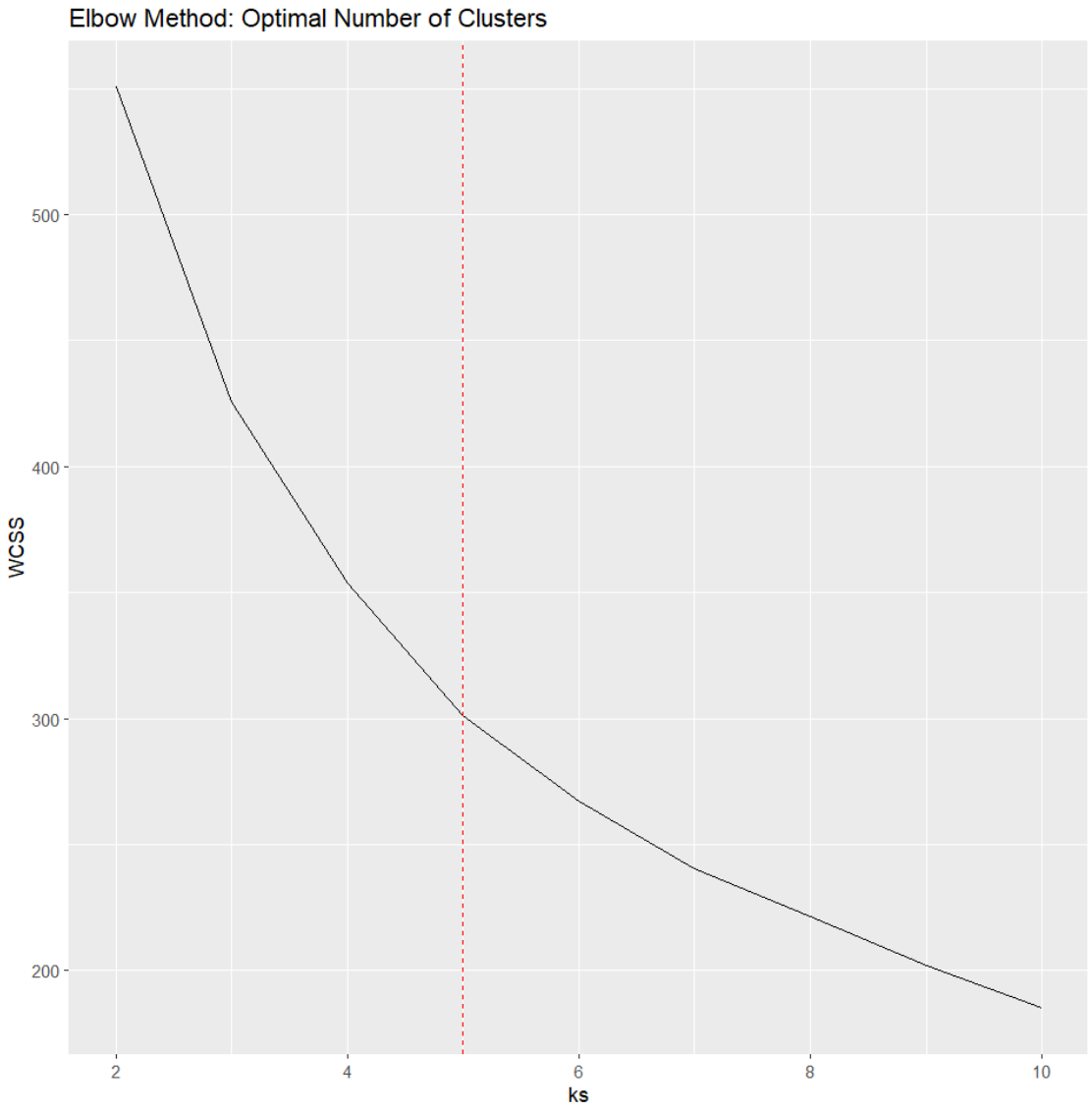

Figure 22. Subset One - Elbow Method: Optimal Number of Clusters.

The elbow method, as shown in Figure 22, is a graphical technique used to determine the optimal number of clusters in a clustering analysis. The method involves plotting the percentage of variance explained by the clusters against the number of clusters, and visually identifying the "elbow" in the plot where the increase in variance explained by adding another cluster begins to level off. To apply the elbow method, the first step is to perform a clustering analysis with a range of different numbers of clusters, typically starting from 2 and increasing incrementally. The percentage of variance explained by each solution is calculated and plotted on a graph. The plot will typically show a downward slope, as the percentage of variance explained increases with each additional cluster. The goal of the elbow method is to identify the point on the plot where the increase in variance explained by adding another cluster starts to level off, forming an elbow shape in the plot. This point is considered the optimal number of clusters for the data set, as it represents the point of diminishing returns where the additional clusters no longer provide a significant increase in the percentage of variance explained. The elbow method is a simple and effective way to determine the optimal number of clusters for a data set. However, it is important to note that the method is not always applicable, and that different data sets may require alternative methods to determine the optimal number of clusters.

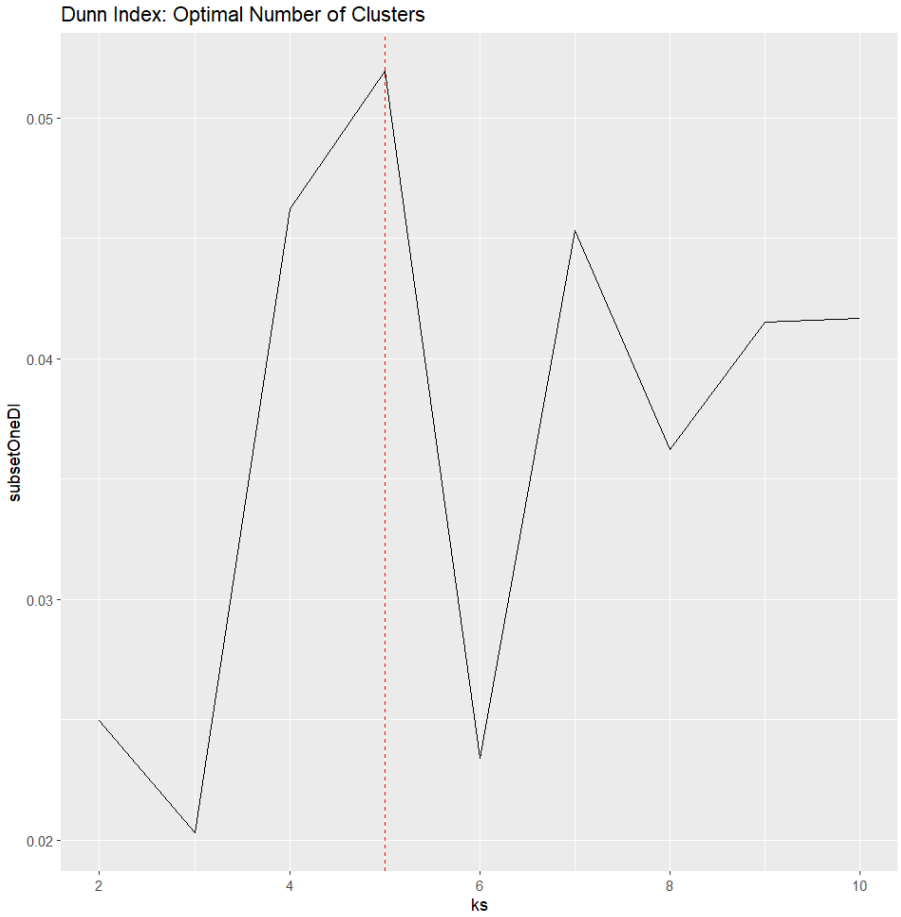

Figure 23. Subset One - Dunn Index: Optimal Number of Clusters.

The Dunn index, as shown in Figure 23, is a clustering validation metric used to evaluate the quality of clustering solutions. It is used to determine the optimal number of clusters by measuring the ratio of the distance between clusters to the diameter of the clusters. The Dunn index measures the minimum distance between clusters divided by the maximum diameter of the clusters. The diameter of a cluster is the distance between the two farthest data points within the cluster. The minimum distance between clusters is the distance between the two closest data points in different clusters. The higher the Dunn index value, the better the clustering solution is considered to be. A high Dunn index value indicates that the distance between clusters is large relative to the diameter of the clusters, suggesting that the clusters are well separated and distinct from one another. To apply the Dunn index, we first perform a clustering analysis with a range of different numbers of clusters. For each clustering solution, we calculate the Dunn index value and choose the solution with the highest value as the optimal number of clusters. The Dunn index is a useful metric for finding the optimal number of clusters, as it takes into account both the distance between clusters and the size of the clusters. However, it is important to note that the method may not always be suitable for certain types of data or clustering algorithms.

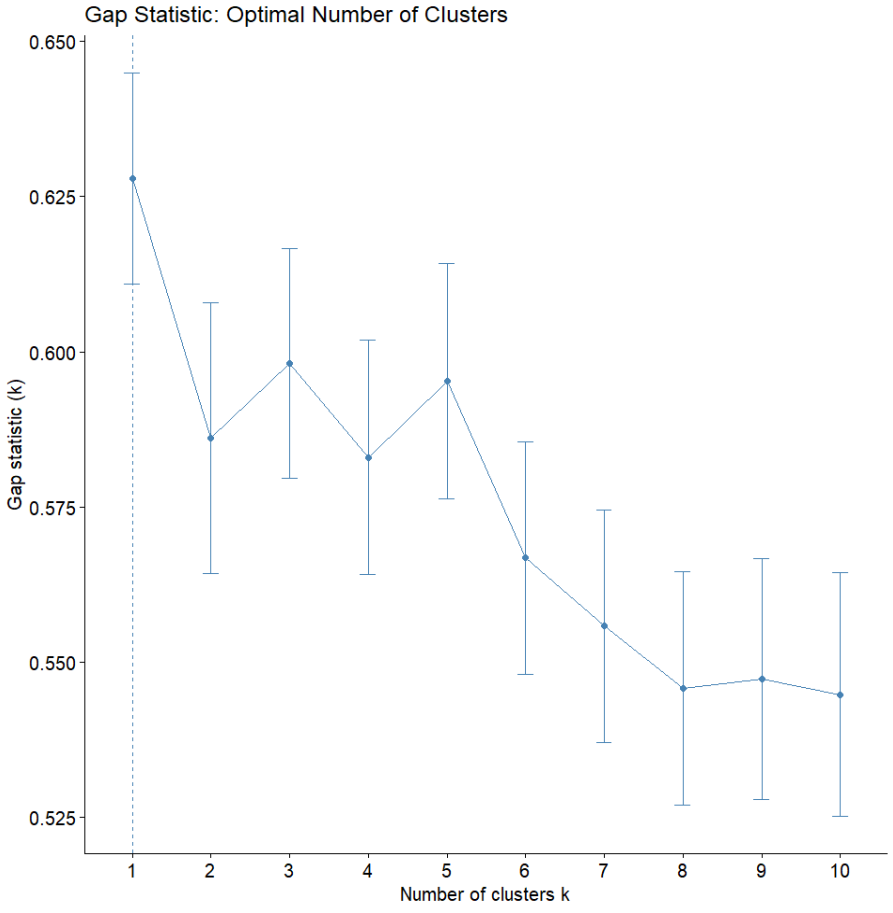

Figure 24. Subset One - Gap Statistic: Optimal Number of Clusters.

The gap statistic method, as shown in Figure 24, involves comparing the within-cluster dispersion of a clustering solution to a set of reference data sets with similar characteristics but without any meaningful clustering structure. It measures the difference between the within-cluster dispersion of the data set being clustered and the expected within-cluster dispersion of the reference data sets. The optimal number of clusters is considered to be the value that maximizes the gap statistic. The first step is to perform a clustering analysis with a range of different numbers of clusters. For each clustering solution, the within-cluster dispersion is calculated and compared to the expected within-cluster dispersion of a set of reference data sets. The expected within-cluster dispersion is calculated as the average within-cluster dispersion of the reference data sets. The gap statistic is then calculated as the difference between the within-cluster dispersion of the data set being clustered and the expected within-cluster dispersion of the reference data sets, multiplied by a factor that takes into account the sample size and the number of clusters being considered. The optimal number of clusters is considered to be the value that maximizes the gap statistic. The gap statistic is a useful method for determining the optimal number of clusters in a clustering analysis, as it takes into account the underlying structure of the data set and the expected within-cluster dispersion of reference data sets. However, the method is not always applicable and may require careful consideration of the reference data sets and the clustering algorithm being used.

Subset Two: Income Per Capita, Upper Quartile Housing, Gini index

Our Subset 2 includes Income Per Capita, Owner Occupied Housing Units Upper Value Quartile, and the Gini Index. Income Per Capita is a measure of the average income earned per person in a given area. This feature can provide insight into the economic well-being of a county and can help identify areas where individuals may be more or less financially stable. Owner Occupied Housing Units Upper Value Quartile is a measure of the highest valued homes in a county that are occupied by their owners. This feature can provide information about the level of wealth and affluence in a given area. Understanding these relationships can help public health officials identify communities where residents may be more likely to have access to resources that can help them avoid infection, such as larger living spaces and access to healthcare. The Gini Index is a measure of income inequality within a population. This feature can provide insight into the distribution of wealth and resources within a given area, which can be an important factor in understanding the potential impact of COVID-19. Understanding the relationship between income inequality and COVID-19 can help public health officials identify communities where certain groups of people may be at higher risk for infection and may require additional resources to prevent and manage outbreaks. On this feature set, we performed both K-means and DBSCAN, as shown in Figures 25-27.

Figure 25. Subset Two Pairs Plot.

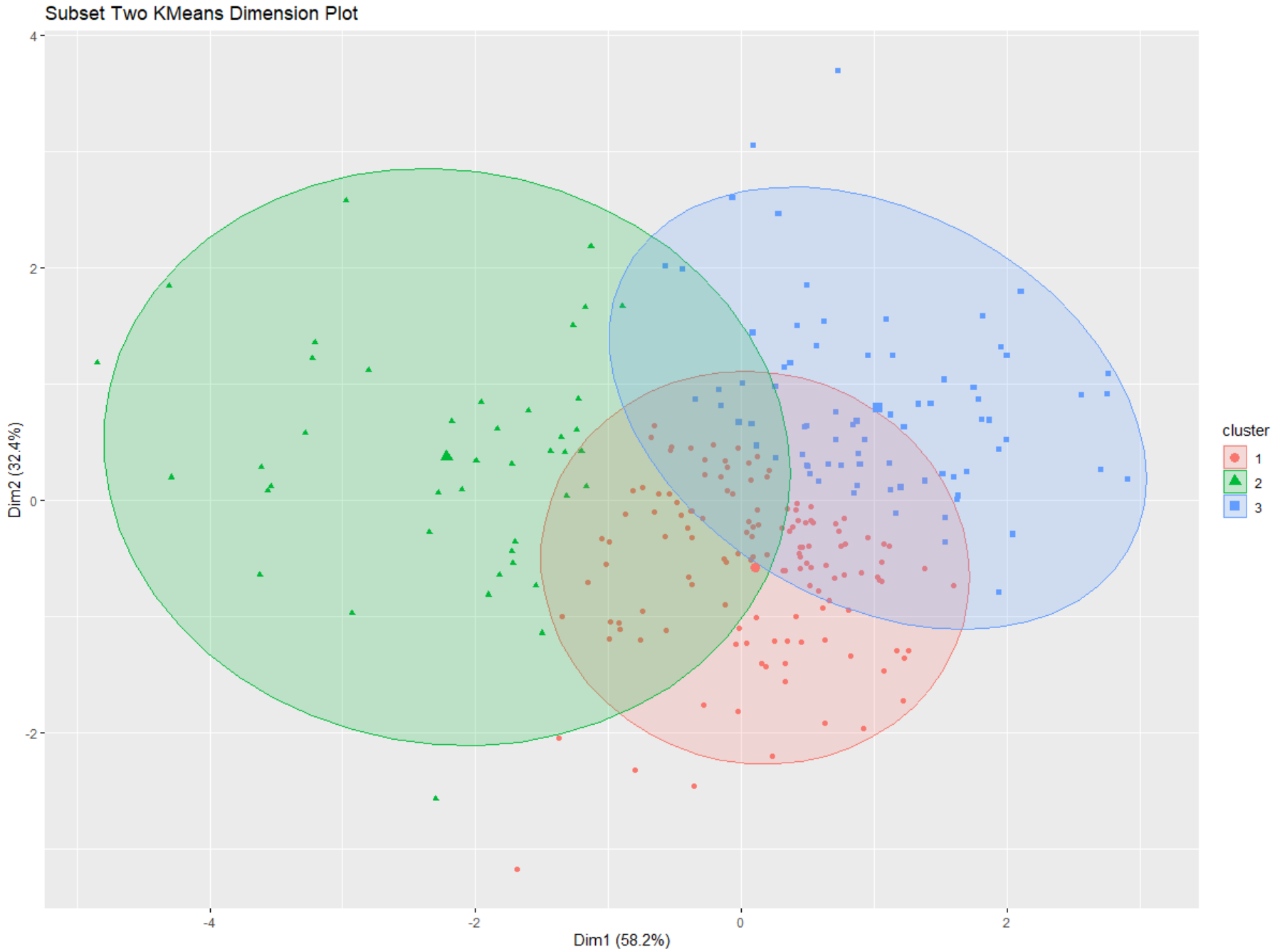

Figure 26. Subset Two - KMeans Dimension Plot.

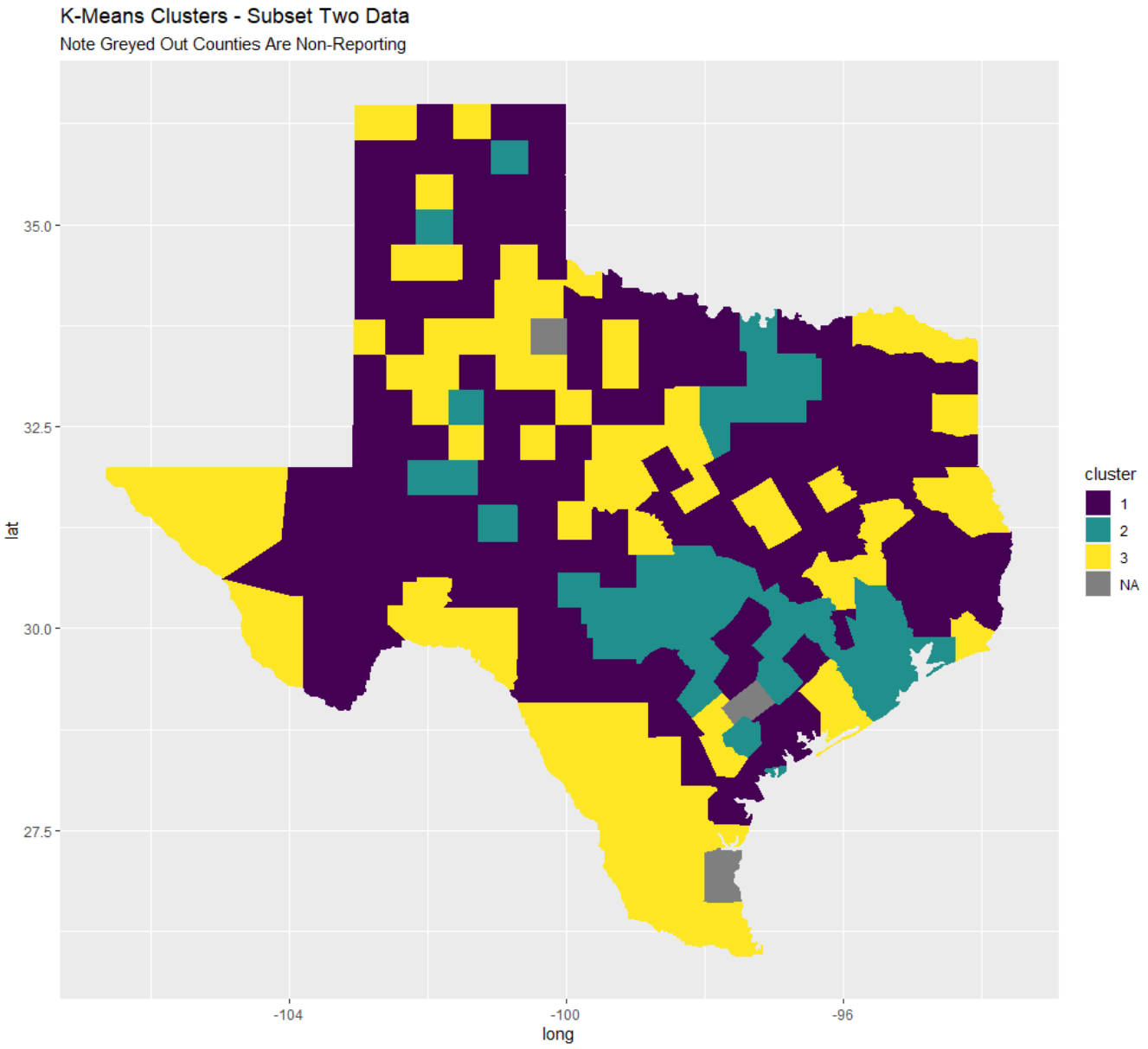

Figure 27. Subset Two - KMeans Clusters.

Similar to Feature Subset 1, the obtained clusters for Feature Subset 2 appear to have a logical interpretation. Figure 28 shows the K-means Cluster Profiles for this subset.

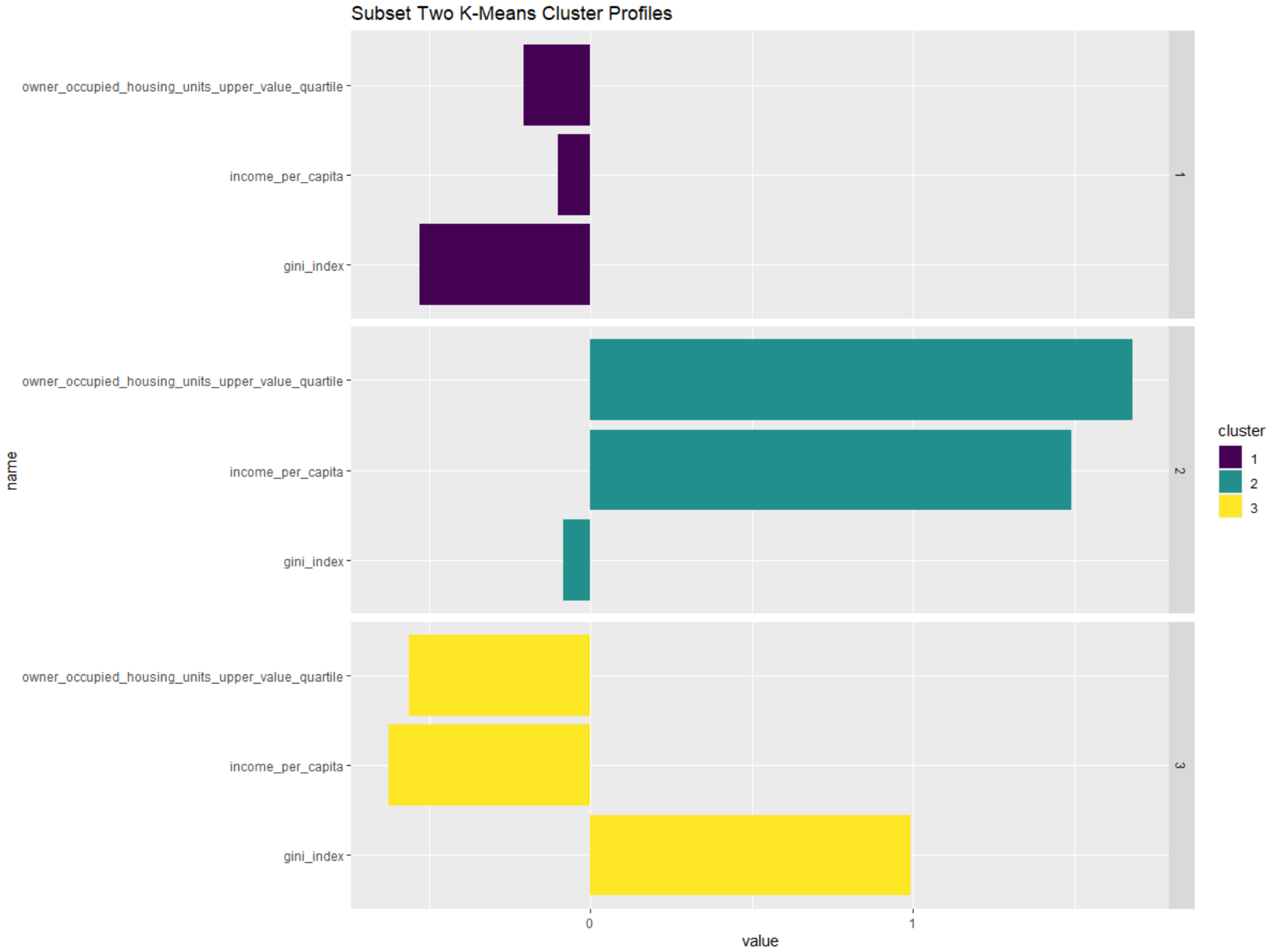

Figure 28. Subset Two - KMeans Cluster Profiles.

Cluster 1 consists of counties with a low number of owner-occupied housing units, a low income per capita, and a very low Gini index. Counties in this cluster are located in South Texas, the far west towards El Paso, and a few counties in the Panhandle. It makes sense that these counties are grouped together because they share similar socioeconomic characteristics. Specifically, these counties tend to have high poverty rates, a lack of affordable housing, and a low concentration of wealth. This cluster likely represents rural or agricultural areas of Texas.

Cluster 2 consists of counties with a very high number of owner-occupied housing units and income per capita, with a slightly low Gini index. Counties in this cluster include major metropolitan areas such as Dallas-Fort Worth, Austin, San Antonio, Houston, Lubbock, and Amarillo. This cluster likely represents urban or suburban areas of Texas with high levels of economic development, where homeownership is relatively common, and the population tends to be more affluent and less unequal.

Cluster 3 consists of counties with very low numbers of owner-occupied housing units and income per capita and a very high Gini index. This cluster likely represents areas of Texas with significant economic and social inequality, where a small number of individuals or corporations hold a disproportionate share of wealth and power. Counties in this cluster could include urban areas with high levels of poverty and economic hardship, as well as rural areas with limited economic opportunities. Our Hierarchical Clustering for Subset 2 produces similar results. We decided to create 4 clusters for reasons that will be discussed later. The resulting plots are shown below in Figures 29-31.

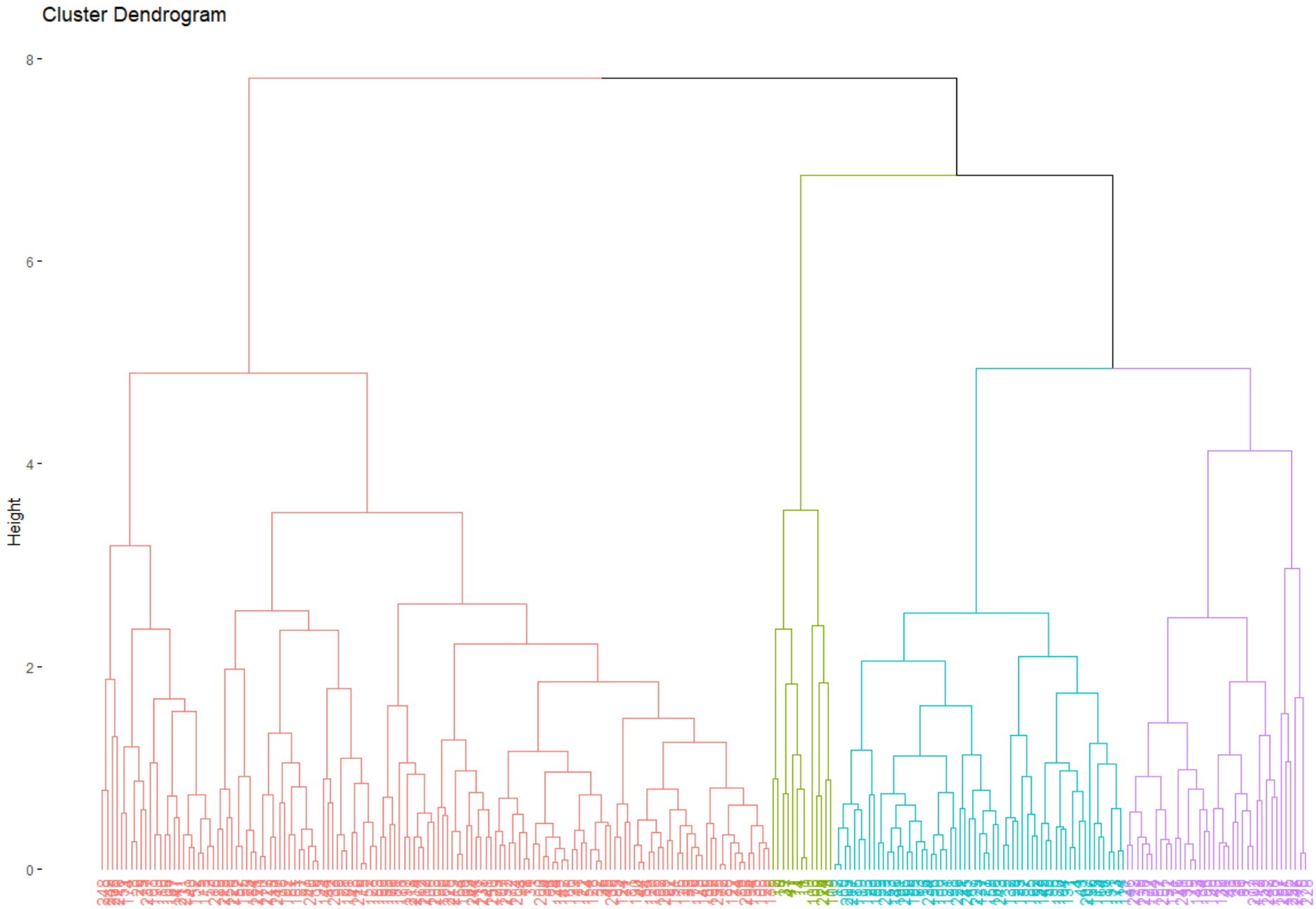

Figure 29. Subset Two Hierarchical Clustering Dendrogram.

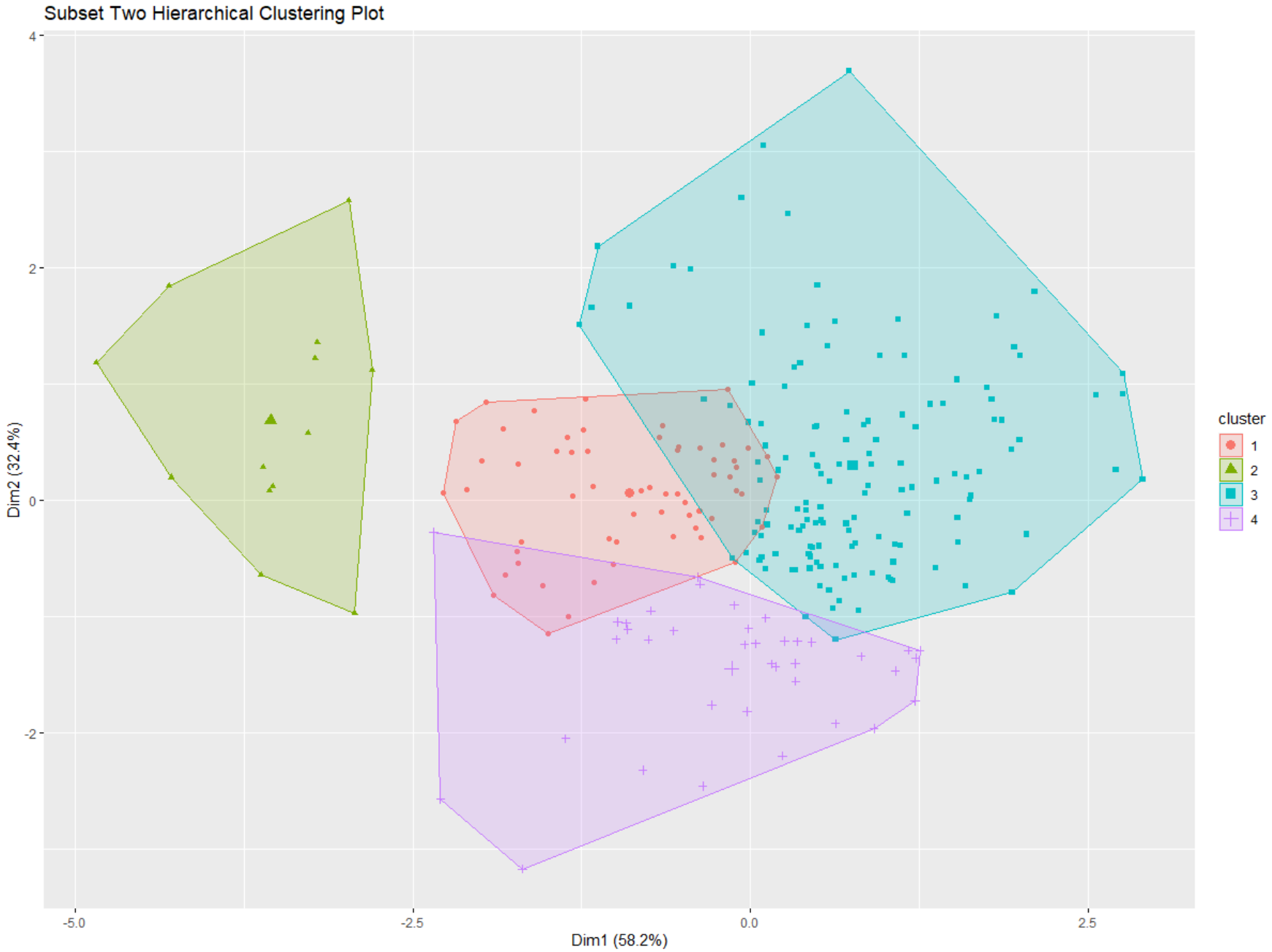

Figure 30. Subset Two Hierarchical Clustering Plot.

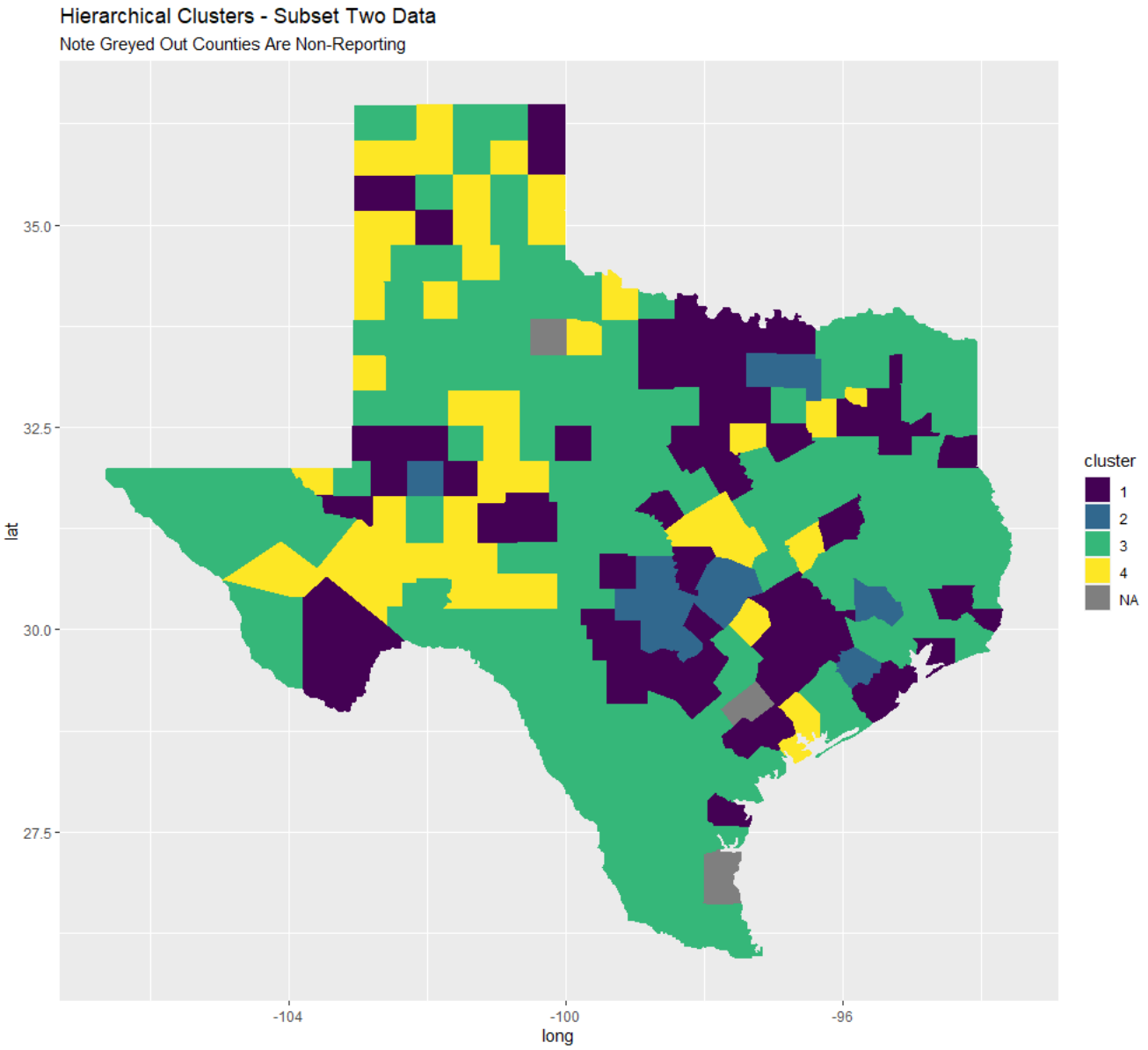

Figure 31. Subset Two - Hierarchical Clusters Texas County Map.

One noticeable difference between the K-means and hierarchical clusterings is the sheer size of Cluster 3 in the hierarchical clustering. It seems to take up the majority of the land area, while in K-means clustering the majority of the land area is split between Clusters 1 and 3. As with Feature Subset 1, the numberings are different between K-means and hierarchical, but the clusters are relatively similar. Cluster 1 in hierarchical corresponds to parts of Cluster 2 in the K-means. Both have counties that tend to be the suburbs of major metropolitan areas. Cluster 2 of hierarchical corresponds with parts of Cluster 2 of K-means. In hierarchical clustering, this cluster represents only the counties of major metro areas. Cluster 3 of hierarchical corresponds to Cluster 1 of K-means. Both these

clusters are of counties that tend to have high poverty rates, a lack of affordable housing, and a low concentration of wealth. Finally, Cluster 4 of hierarchical corresponds to Clusters 1 and 3 of the K-means, mostly the West Texas counties. West Texas counties tend to have lower income per capita and a lower number of upper-quartile housing units compared to counties in other regions of the state. Additionally, the Gini index tends to be higher in West Texas counties, indicating a higher level of income inequality within these communities. Finally, we performed DBSCAN on this feature subset. Unlike our Feature Subset 1, DBSCAN found more than one cluster. However, there is a definite major and minor cluster, as illustrated by the Figure 32 in the graph below.

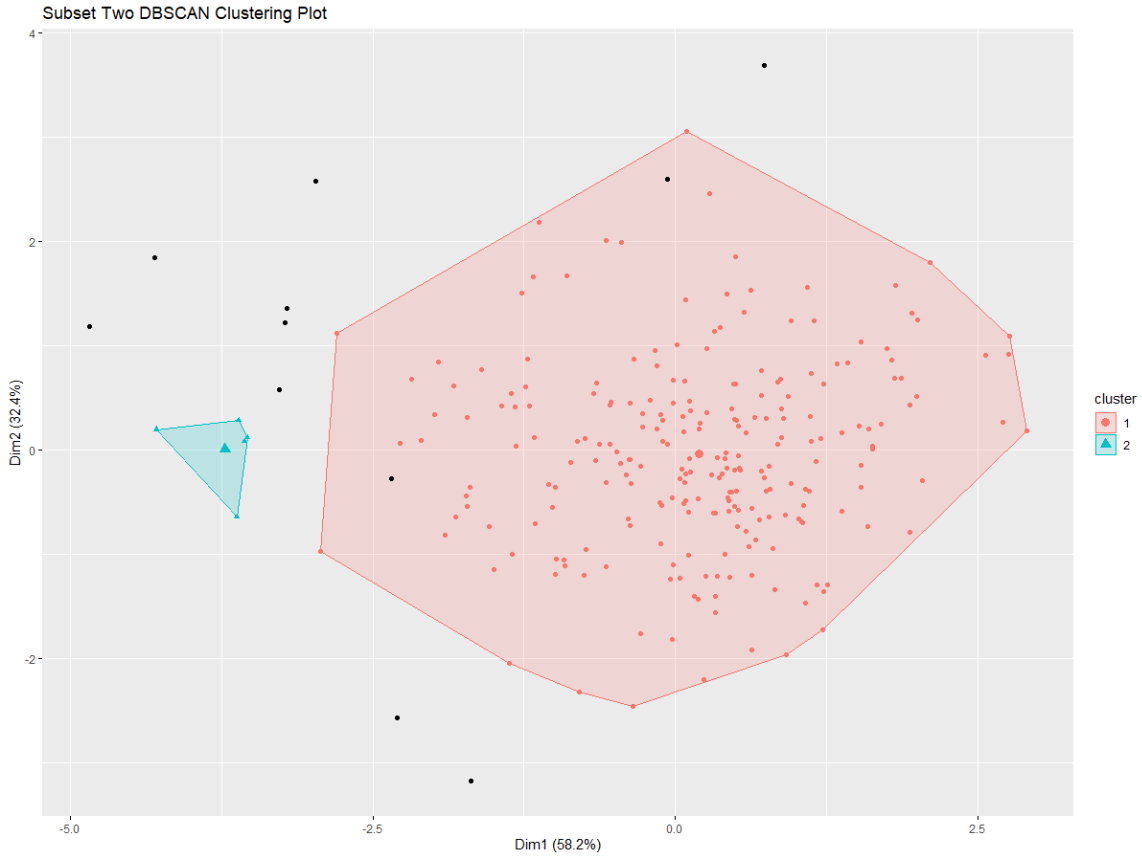

Figure 32. Subset Two - DBSCAN Clustering Plot.

The major cluster likely represents counties with similar socio-economic characteristics, such as moderate to high levels of income per capita, a somewhat high number of upper quartile housing units, and a moderate Gini index. These counties may be located in regions of the state with higher levels of economic development or urbanization, and where homeownership is relatively common. The minor cluster may indeed be outliers that happen to be close to each other in the feature space, and their inclusion in a separate cluster may not necessarily have any meaningful interpretation. Overall, Texas's Gini index hovers around .47, putting it on par with Canada. Drastic changes across the state are somewhat absent, which helps explain the large single cluster.

Subset Two: Internal Validation

Since the plots for each subset exhibit similar characteristics to the ones previously described in Subset One: Internal Validation, we will present them without repeating the descriptions. Figures 33-36 show comparable patterns to those observed before.

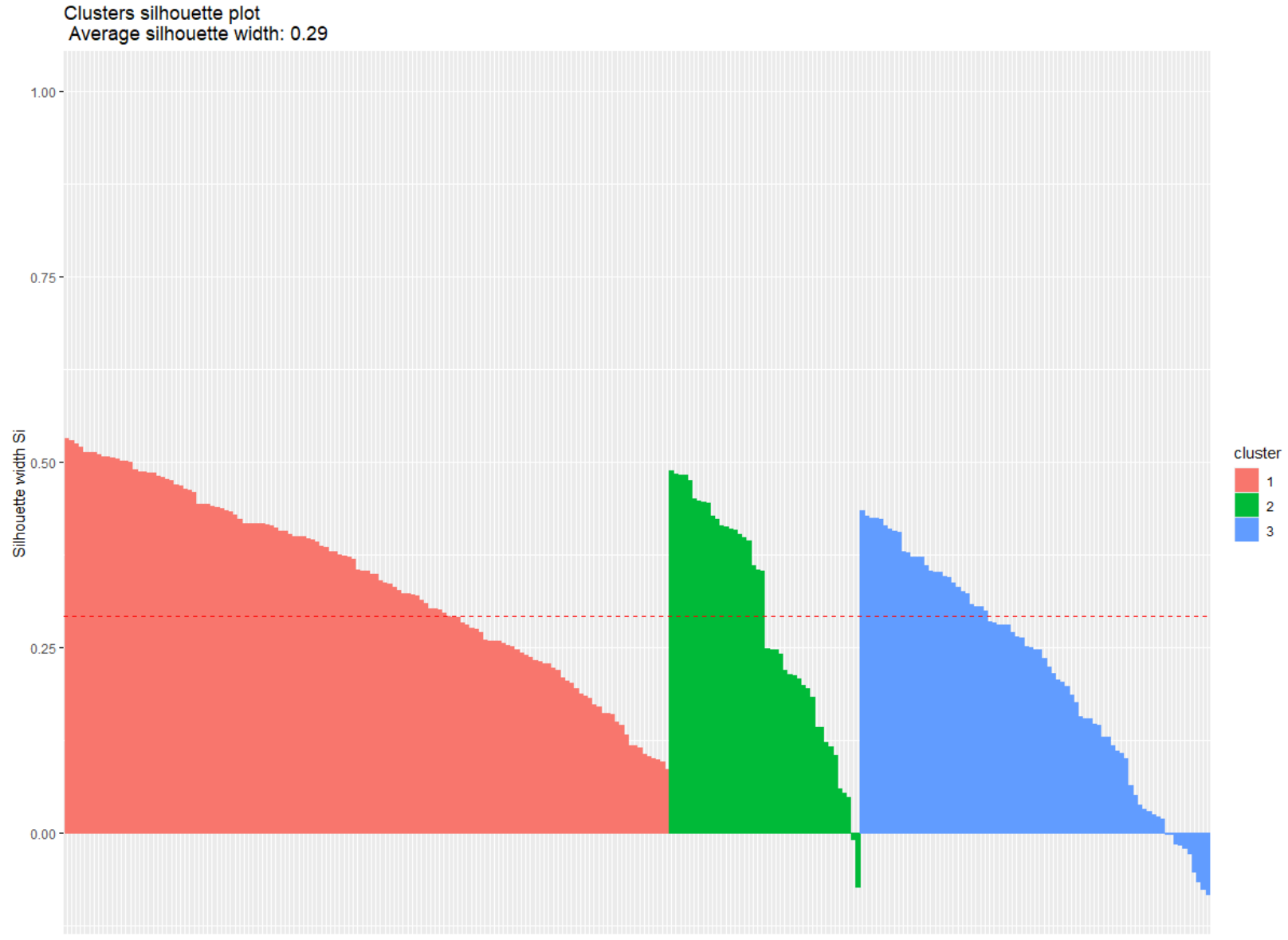

Figure 33. Subset Two - Internal Validation Cluster Silhouette Plot.

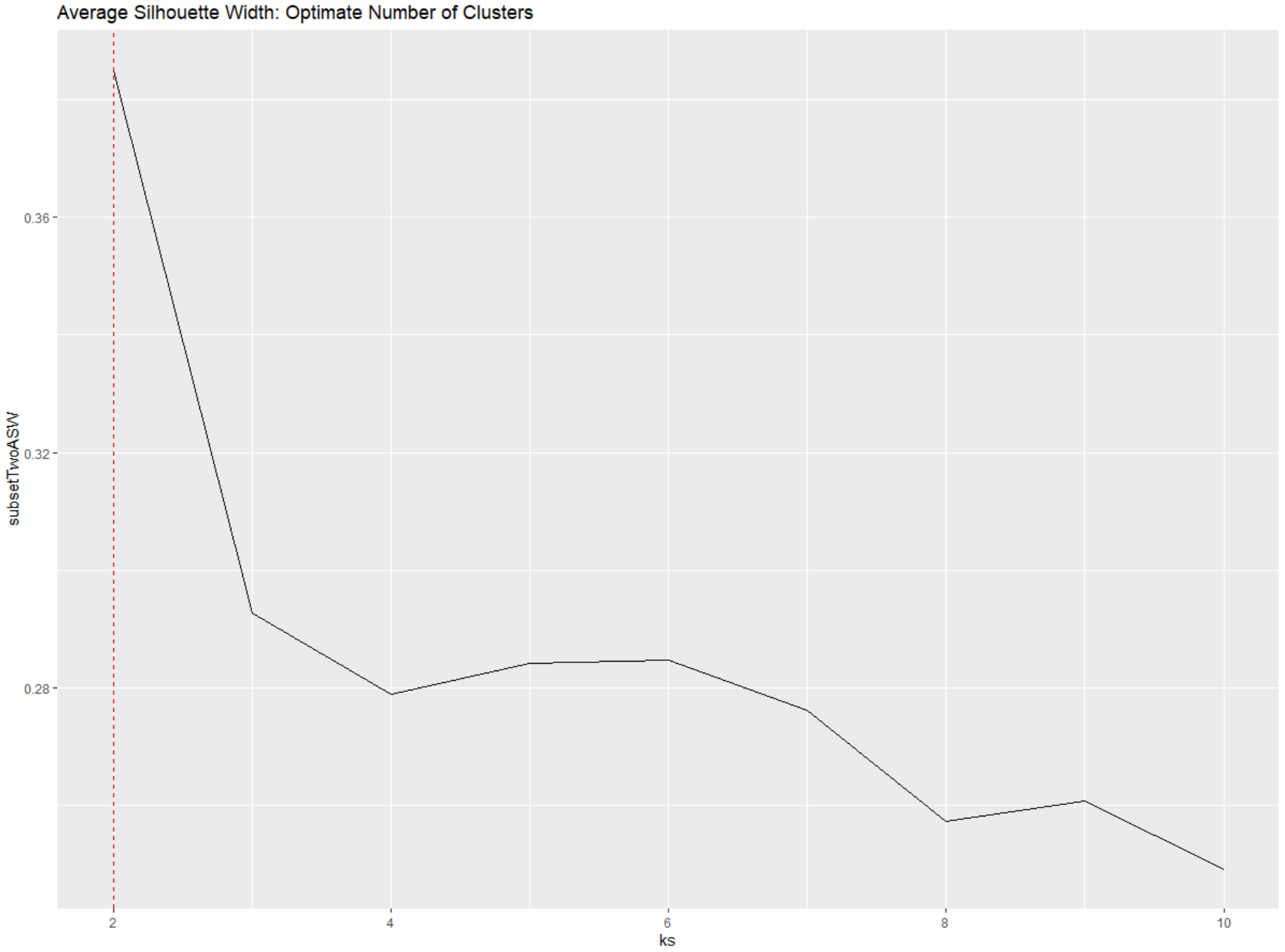

Figure 34. Subset Two - Internal Validation Average Silhouette Width.

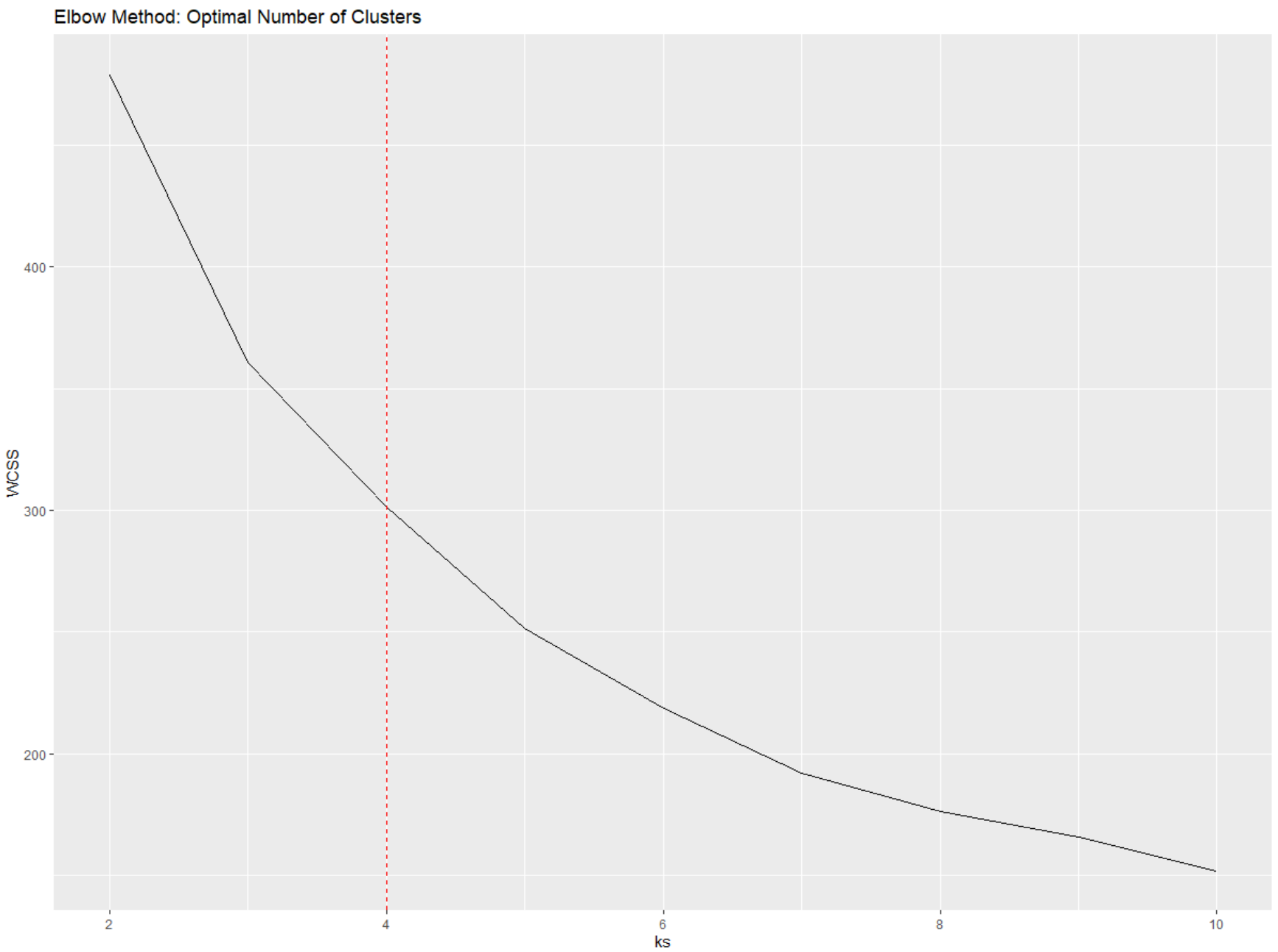

Figure 35. Subset Two - Internal Validation Elbow Method.

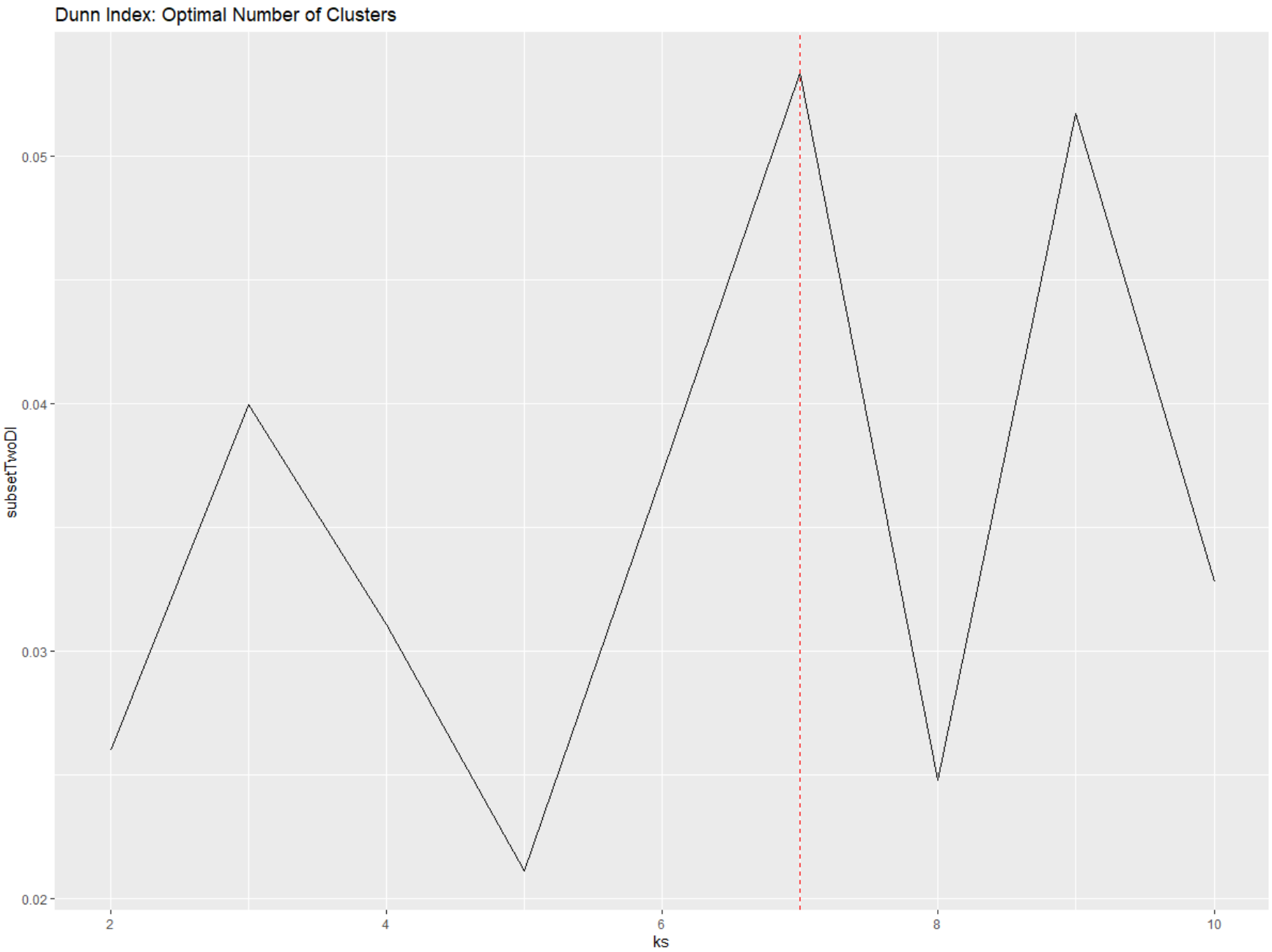

Figure 36. Subset Two - Internal Validation Dunn Index.

Ground Truth And External Validation

The term "ground truth" refers to the reality that we want our data to represent. In our analysis, we selected "deaths per case" as our ground truth feature since it provides a measure of the impact of the virus on each county. To compare the clusterings obtained from subsets one and two, we used the same number of K-means clusters and plotted the resulting clusters on Texas County maps. The maps are shown in Figures 37 and 38.

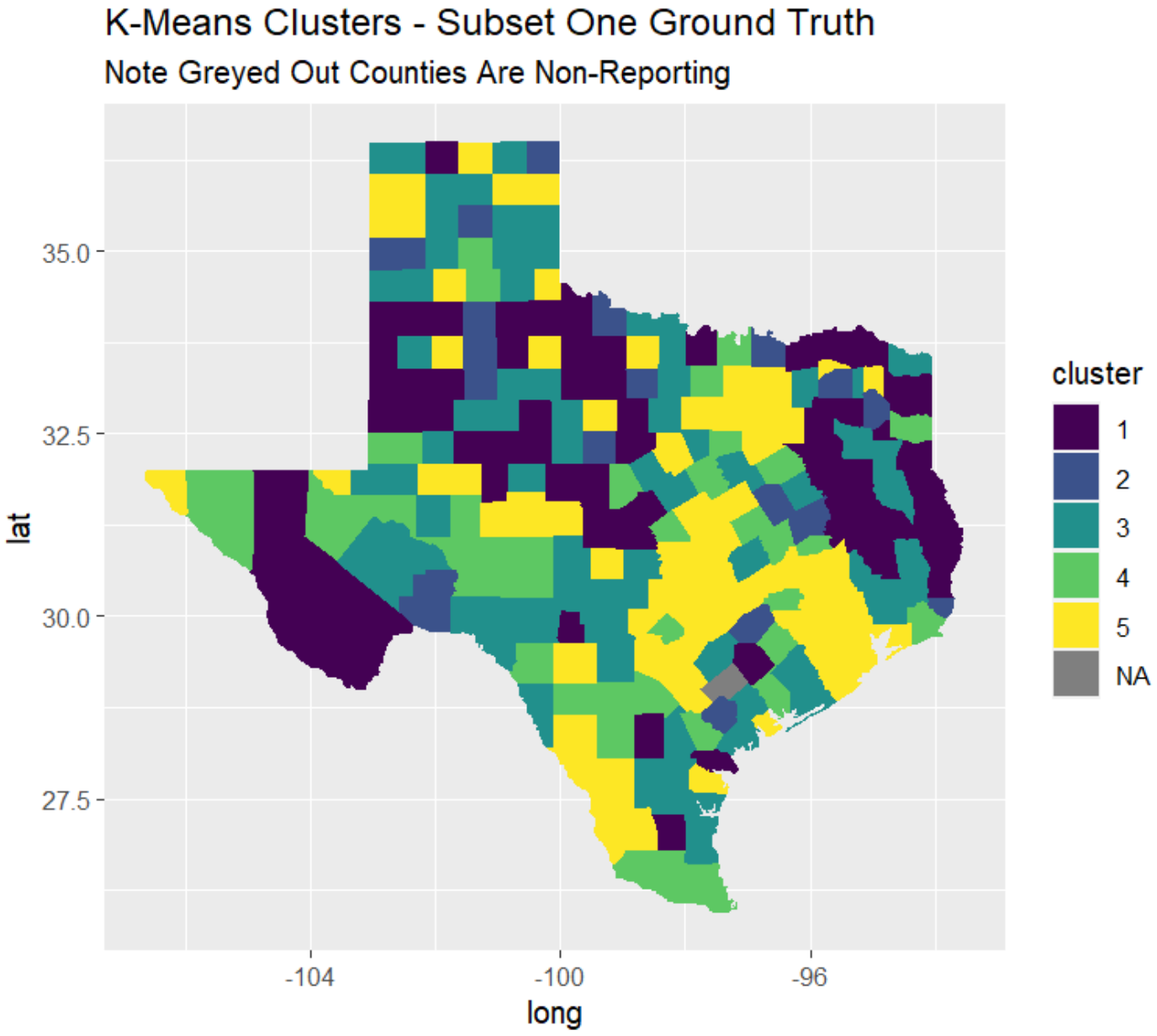

Figure 37. Subset One Ground Truth performed by KMeans Clustering.

We also computed several indices to assess the agreement between our K-means clusterings and the ground truth based on deaths per case. The corrected rand index, the variation of information index, the entropy, and the purity were used for this purpose. The rand index ranges from 0 to 1, where 0 indicates no agreement between the two clusterings and 1 indicates perfect agreement. The variation of information index (VI) measures the distance between two clusters. Entropy refers to the expected information conveyed by identifying the outcome of random trials, where uniform probability has maximum entropy. Purity measures the degree to which clusters contain a single class, so having more clusters makes it easier to achieve higher purity. Table 4 shows the values of these indices for subset one. Our clusterings achieved some similarity to the ground truth, but not significantly so. Essentially, what was grouped similarly can be attributed to population size.

Subset One

	Rand Index	vi	Entropy	Purity
K-Means	0.0829	2.6278	1.8847	0.3773
Hierarchical Clustering	0.07716	2.4012	1.4564	0.5442

Table 4. Subset One External Validation Measures Performed on Ground Truth and Explored Data.

We performed the same calculations for the ground truth using the feature deaths per case for subset two using 3 clusters. These clusters are shown on a Texas county map in Figure 38.

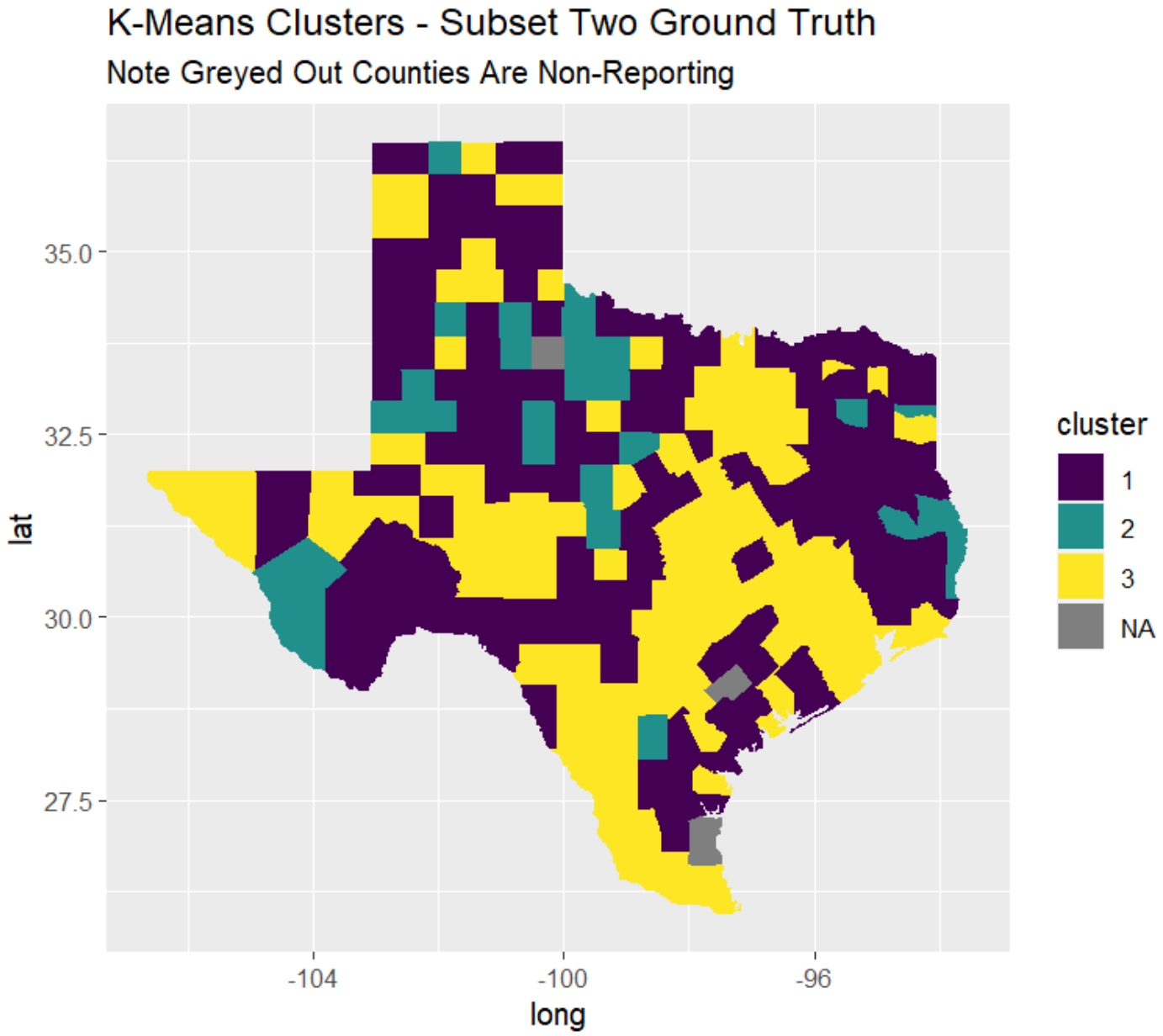

Figure 38. Subset Two Ground Truth performed by KMeans Clustering.

For subset two, the values of these calculations are shown in Table 5. Subset two showed even less similarity to the ground truth than subset one. In fact, the clusterings are significantly different.

Subset Two

	Rand Index	vi	Entropy	Purity
K-Means	0.0329	1.8792	1.4951	0.5257
Hierarchical Clustering	0.0324	1.9577	1.2588	0.4870

Table 5. Subset Two External Validation Measures Performed on Ground Truth and Explored Data.

Through many hours of experimentation with our ground truth clusterings and our explorative clusterings, we were not able to achieve better external validation results. We believe that external validation is difficult to perform on an exploration such as this one because we simply do not know what the 'correct' clusterings might be. However, by choosing a feature that is often characterized by populations and other census features such as deaths per case, we can produce some similarities between the ground truth and our own clusterings.

Evaluation

Regions of Texas and COVID-19

Interestingly, regardless of the clustering method used, the identified clusters were quite similar. We employed three different methods: DBSCAN, k-means, and hierarchical clustering, and all produced comparable results. The findings of the clustering analysis revealed that the major metropolitan counties in Texas, such as Harris, Dallas, Tarrant, Bexar, Travis, and El Paso counties, were grouped together. This outcome was not unexpected, given that these counties have large populations and urban centers. The clustering of these counties highlights the significance of the rural-urban divide in Texas, as counties with larger populations and urban centers tend to exhibit different characteristics than their rural counterparts. The landscape of Texas has undergone significant transformation due to urbanization, with different regions of the state experiencing varying levels of development. It is crucial to consider these geographic distinctions in analyzing the data. The term "Texas Triangle" refers to the area covered by the four major metroplexes of Texas: Dallas-Fort Worth, Austin, San Antonio, and Houston [4]. With over 18 million residents, each of the Texas Triangle metros is among the top six strongest urban areas in the United States in terms of economic power [5]. The Texas Triangle is known for its large urban areas and thriving economies.

Moreover, the clustering of rural counties between metro areas suggests that these counties have similarities in terms of their demographics and socioeconomic status. This is an important finding as it can help policymakers understand the needs of these areas and develop targeted interventions to improve the lives of their residents. The clustering of these rural counties could also be used to promote economic development in these areas, such as the creation of new industries or the expansion of existing ones. Another interesting finding was the clustering of rural counties in West Texas into their own cluster. This suggests that these counties have distinct characteristics that set them apart from other rural counties in Texas. The clustering of these counties could be used to develop specific policies that address the unique needs of these areas. In contrast, West Texas is a sparsely populated region with several small towns and rural communities. The region encompasses a vast area, with few large urban centers, and thus has limited access to medical care. South Texas, on the other hand, is an economically disadvantaged region that is also home to a predominantly Hispanic population. The region has several urban areas, including Laredo, Brownsville, and McAllen, but also has many rural and remote communities. The region has a high prevalence of chronic diseases, such as diabetes and obesity, and limited access to medical care, making its residents more susceptible to severe COVID-19 infections. As of February 2023, the region has reported over 536,000 cases of COVID-19 and more than 9,700 deaths. Many residents of South Texas lack health insurance or have limited access to primary care physicians, and there are few specialists in the region. The region's healthcare system

faces unique challenges due to language and cultural barriers, as well as a shortage of healthcare providers. One limitation of the project is that it only used a limited number of features to cluster the counties. There are many other factors that could be used to cluster the counties, such as education level, unemployment rate, crime rate, and access to healthcare. Future studies could expand on this project by incorporating additional features to create a more comprehensive picture of the clustering of counties in Texas.

Overall, the clustering of counties in Texas based on specific features such as income per capita, owner-occupied housing units upper value quartile, Gini index, median income, median age, and average percent income spent on rent is an important project that can provide valuable insights into the demographics and socioeconomic status of different areas of the state. The findings of the project suggest that there are distinct clusters of counties in Texas, with large metro counties, rural counties between metro areas, and rural counties in West Texas clustering together. The project also found that the clustering of counties is robust and reliable, regardless of the specific method used. The project has some limitations, including the use of a limited number of features, but it provides a solid foundation for future studies on the clustering of counties in Texas.

Texas County Responses - Vaccine Sites

In our final analysis, we delved into how different clusters of Texas counties responded to the COVID-19 pandemic. We specifically focused on the K-means clusters as we had already internally validated that the optimal number of clusters had been used. To evaluate a county's response to the virus, we examined its cases, deaths, and vaccine site data.

To get a better understanding of each cluster's response, we combined the cases and deaths data with vaccine site data, which provided information on the number of vaccine sites per 1000 people in each county. Moreover, to make meaningful comparisons between clusters, we normalized this data. We then calculated the average number of COVID-19 cases, deaths, and vaccine sites per 1000 people for each cluster in subset one and subset two. The results of these calculations are summarized in Tables 6 and 7.

By examining these tables, we can see that certain clusters had lower averages of COVID-19 cases and deaths compared to others, indicating that they were able to better manage the virus. For instance, in subset one, cluster 1 had the lowest average number of cases and deaths per 1000 people, while in subset two, cluster 4 had the lowest average number of cases and deaths per 1000 people. Furthermore, we can observe that some clusters had higher averages of vaccine sites per 1000 people, which suggests that they were able to vaccinate their residents more effectively. In subset one, cluster 3 had the highest average number of vaccine sites per 1000 people, while in subset two, cluster 2 had the highest average number of vaccine sites per 1000 people.

Subset One K-Means Cluster Information

Cluster	Average Cases	Average Deaths	Average Sites per 1000 People
1	291	3.32	-0.2333
2	250	5.02	-0.2169
3	296	4.61	-0.1809
4	334	5.78	0.5191
5	286	5.23	0.1906

Table 6. Subset One K-means Cluster Data for Average Cases, Average Deaths, and Average Sites per 1000 people (after normalization).

We used five K-means clusters for subset one, clustering on the features of median gross rent of individuals renting housing units, median age, and the percentage of income spent on rent. Figure 39 shows the number of counties belonging to each cluster, relative to each other.

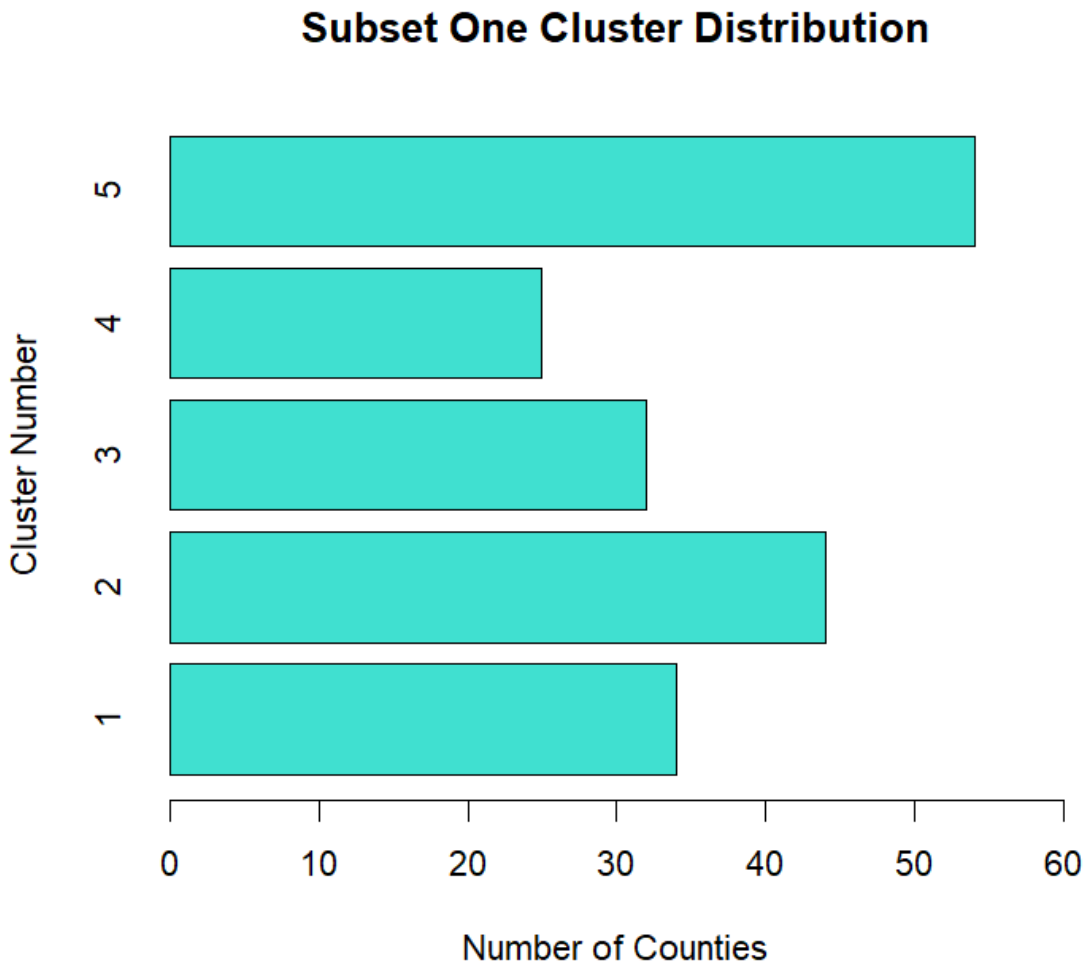

Figure 39. Subset One K-means Number of Counties Within Each Cluster.

Figure 40 visualizes the vaccine site data for each cluster beside each other, where red dots represent the mean in each cluster. It is visually obvious that subset one's cluster 4 had the highest average number of vaccine sites per 1000 people. The table shows the exact value of each of the red dots representing the average number of vaccine sites per 1000 people after normalization. However, upon closer inspection, we noticed that cluster 4 also had the highest values for average cases and average deaths.

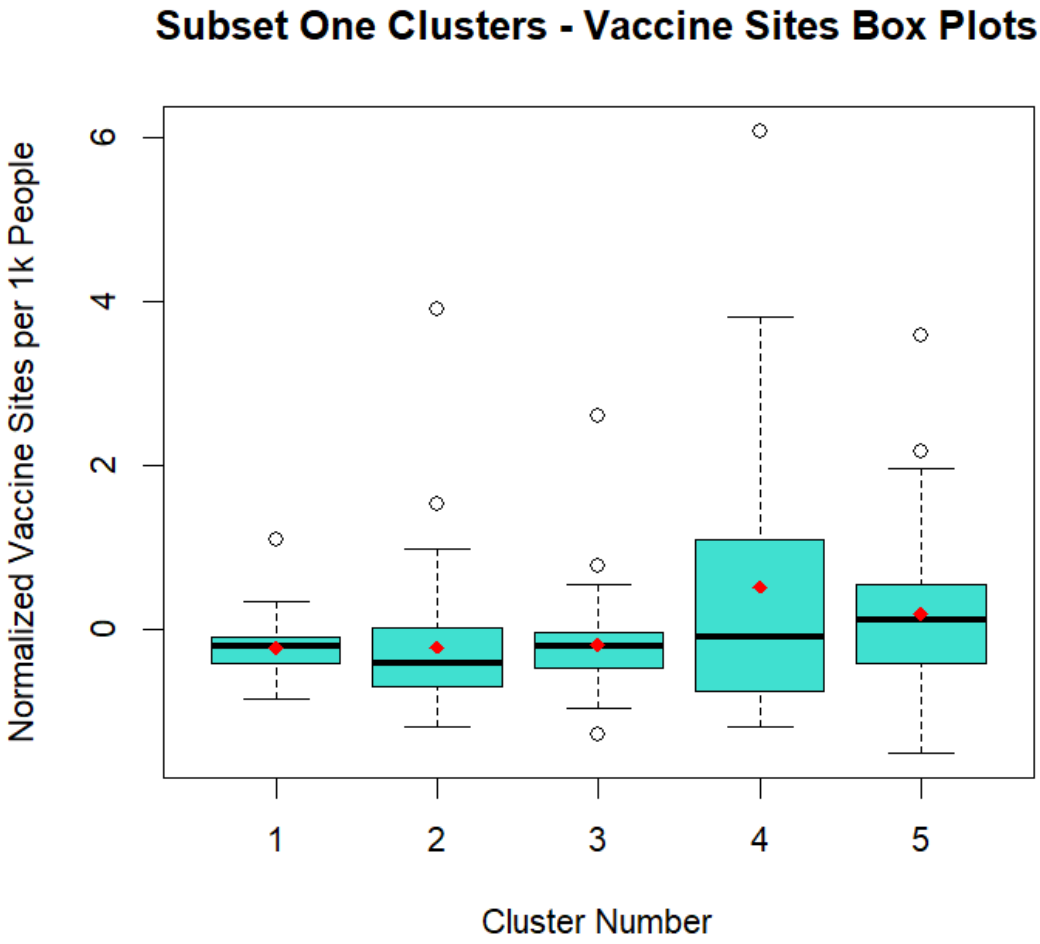

Figure 40. Subset One K-means Vaccine Sites per 1000 People Box Plot. The red dot represents the mean.

On the other end of the scale, we see that subset one's cluster 1 had the least average number of vaccine sites per 1000, the lowest average deaths, and the median average cases. Cluster 1 included counties such as Harris County, Dallas County, and Tarrant County, all counties with high populations as well as higher median gross rent and higher percent income spent on rent. Cluster 4, however, included counties with much lower total populations, such as Stonewall County with about 1000 people, Hardeman County with about 3500 people, and Menard County with about 2000 people. From this information, we can see that less populous counties tended to respond to the virus with more vaccine sites per 1000 people, but because they have a small population this number is inflated and they actually offered fewer vaccine sites in their counties compared to more populous counties like Harris and Dallas counties. By offering fewer actual vaccine sites, these counties did not respond well to the virus, contributing to their high number of average cases and average deaths.

More urban counties like Dallas County and Harris County in cluster 1 provided more actual vaccine sites, but their number of vaccine sites per 1000 people is lower because they simply have large populations in the millions. However, because they offered more vaccine sites, they were able to mitigate more of the harm and produce the lowest average deaths and median average cases. It seems as though living in a more populous and urban county resulted in a better COVID-19 response from the county, likely due to more funding and resources or due to the sheer necessity to mitigate the damage in such populous areas to avoid overwhelming their healthcare system.

For subset two, we used 3 K-means clusters, clustering on the features of income per capita, upper quartile value housing units occupied by the owners, and the Gini index. Figure 41 shows the number of counties belonging to each cluster, relative to each other.

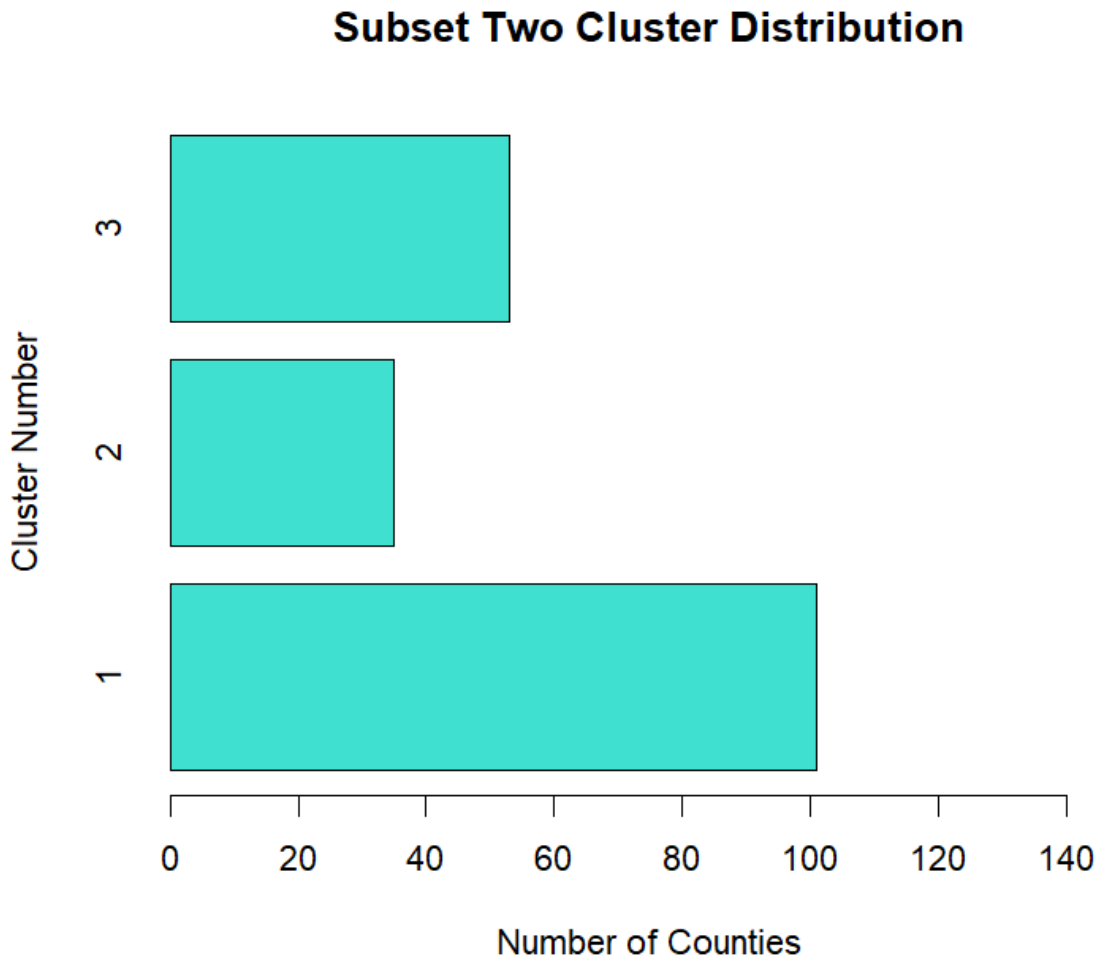

Figure 41. Subset Two K-means Number of Counties Within Each Cluster

Figure 42 visualizes the vaccine site data for each cluster in subset two beside each other in the figure below, where red dots represent the mean in each cluster. As the figure below shows and the table below confirms, the normalized average vaccine sites per 1000 people for each cluster are more similar to one another than in subset one.

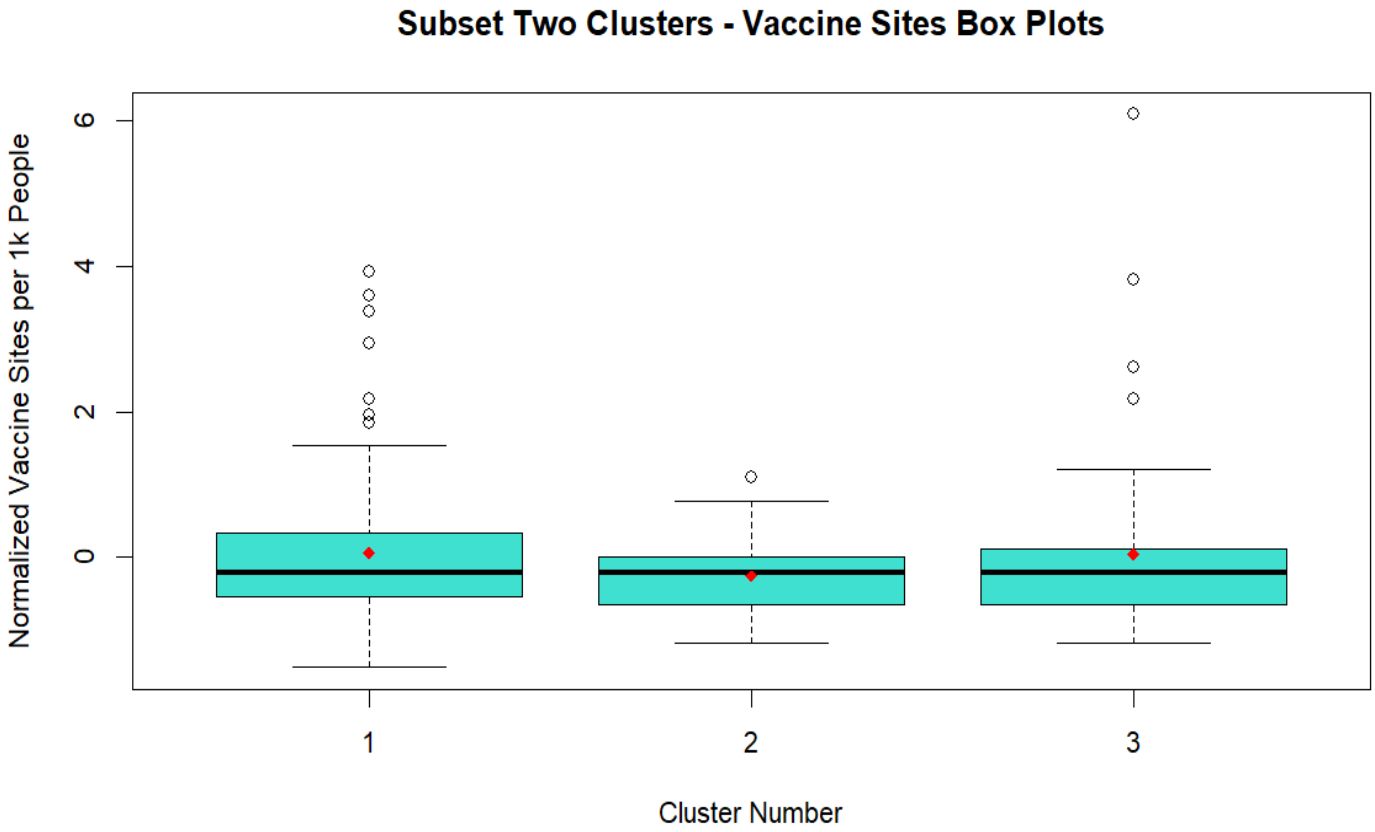

Figure 42. Subset Two K-means Vaccine Sites per 1000 People Box Plot. The red dot represents the mean.

Table 7 shows the exact value of each of the red dots representing the average number of vaccine sites per 1000 people after normalization. Cluster 1 has the highest value for normalized average sites per 1000 people, with cluster 3 coming in a close second. Cluster 2, however, has a value farther away. We looked further into which counties were included in each cluster.

Subset Two K-Means Cluster Information

Cluster	Average Cases	Average Deaths	Average Sites per 1000 People
1	295	5.05	0.0678
2	267	3.25	-0.2541
3	293	5.74	0.0386

Table 7. Subset Two K-means Cluster Data for Average Cases, Average Deaths, and Average Sites per 1000 people (after normalization).

Cluster 2 includes counties such as Dallas, Harris, and Tarrant grouped together again. However, this is expected that they are clustered together since we clustered using the Gini index and the income per capita, which are similar and high in large populous counties. However, as we saw earlier in subset one, the counties in this cluster have the lowest average cases and lowest average deaths. Cluster 1 includes counties such as Upshur, Waller, and Colorado counties. These counties have populations of around 50,000 or fewer and have a lower Gini index, indicating less wealth inequality. A lower population certainly explains the normalized average sites per 1000 people at a higher value than in cluster 2. However, these counties still experienced around the same amount of cases and deaths reported in cluster 3, which had a lower value for the normalized average sites per 1000 people. Once again, cluster 3 grouped together counties with smaller populations and very high Gini Indices, such as Stonewall County. These findings correspond with the conclusion we made from subset one as well. More populous and urban counties seem to respond better to the virus, while less populous communities were among the hardest hit.

Conclusions

In conclusion, this project aimed to analyze data related to the COVID-19 pandemic in different counties across the United States, particularly in Texas. By performing cluster analysis using several methods such as k-means and hierarchical clustering, we were able to group similar counties based on various feature subsets. Our analysis revealed interesting patterns and trends in the data. We found that regardless of the feature subsets and clustering methods used, the clusters were relatively the same. Large metropolitan counties in Texas were clustered together, while rural counties between metro areas were also clustered together. Additionally, rural counties in West Texas were their own distinct cluster.

Our findings have several implications for policymakers and healthcare professionals. By identifying counties with similar characteristics and responses to the virus, we can better target resources and interventions to those areas that need them the most. This can help prevent the spread of the virus and ultimately save lives.

Moving forward, we plan to continue working with COVID-19 data and further explore the questions that we posed at the beginning of this project. We are particularly interested in identifying groups of counties that are more severely hit than others. By performing more cluster analysis using different methods and feature subsets, we hope to gain a better understanding of the factors that contribute to the spread of the virus and the effectiveness of interventions.

Overall, this project highlights the power of data analysis in tackling complex problems such as the COVID-19 pandemic. By using clustering methods to group counties with similar characteristics and responses to the virus, we can gain valuable insights that can inform decision-making and ultimately save lives. We hope that our findings will be useful to policymakers and healthcare professionals as they work to combat the spread of the virus and protect the health and well-being of their communities.

References

[1] “Who coronavirus (COVID-19) dashboard,” *World Health Organization*. [Online]. Available: https://covid19.who.int/. [Accessed: 28-Feb-2023].

[2] “Guidance and tips for tribal community living during COVID-19,” *Centers for Disease Control and Prevention*, 23-Aug-2022. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/community/tribal/social-distancing.html. [Accessed: 28-Feb-2023].

[3] By, “Coronavirus in the U.S.: Latest Map and case count,” *The New York Times*, 03-Mar-2020. [Online]. Available: https://www.nytimes.com/interactive/2021/us/covid-cases.html. [Accessed: 28-Feb-2023].

[4] J. H. Cullum Clark, “The Texas Triangle: A rising megaregion unlike all others,” *George W. Bush Presidential Center*, 19-May-2021. [Online]. Available: https://www.bushcenter.org/publications/the-texas-triangle-a-rising-megaregion-unlike-all-others. [Accessed: 28-Feb-2023].

[5] “Texas triangle,” *Austin Capital Advisors*. [Online]. Available: https://www.austincapitaladvisors.com/texas-triangle. [Accessed: 28-Feb-2023].