Final Report

Olivia Hofmann, Matias Barcelo, and Mike Perkins

2024-11-13

Contents

Problem Description (Business Understanding)

COVID-19 is a highly contagious respiratory illness that first emerged in Wuhan, China in December 2019. COVID-19 entered the United States in January 2020 with the World Health Organization (WHO) declaring COVID-19 a "global health emergency" in March 2020. The virus spreads through respiratory droplets dispersed when someone coughs, sneezes, or even talks. COVID-19 can cause symptoms including those similar to a cold, influenza, or pneumonia with the potential to become very severe and lead to death. The COVID-19 virus overwhelmed healthcare systems and disrupted economies around the world. [1] [2]

The stakeholder for this data analysis is a property developer who is interested in determining the best location in Texas for developing a mixed-use building. The stakeholder's key concern is selecting a county that demonstrates stability and resilience in response to unpredictable events, like the COVID-19 pandemic. The mixed-use building that the stakeholder is looking to develop will have space for a gym, restaurants, pharmacy, and other similar businesses. When deciding where to build this mixed-use building, the stakeholder is looking for insights into which counties in Texas have successfully managed public health crises as situations similar to this would greatly impact the success of the businesses within his building. Every business that would be in the mixed-use building would be heavily reliant on consistent traffic and economic activity. Any change in foot traffic and economic activity would directly impact the success or failure of each business. The analysis will include data on COVID-19 cases, COVID-19 deaths, and the effectiveness of government interventions (such as lock downs and social distancing). This analysis is crucial for the stakeholder to make an informed decision regarding this long-term investment, as counties that respond well to crises are more likely to provide stable environments for growth and development.

Some questions that the stakeholder would like answered are:

- What are the characteristics of counties in Texas that showed resilience during the COVID-19 pandemic, based on COVID-19 case rates?
- What are the economic and social impacts in counties that were more or less affected by the pandemic and how might these influence future development potential?
- How did COVID-19 impact the workplace and employment rates in the various counties?
- Which counties showed consistent consumer foot traffic during the pandemic, indicating stable economic activity?

All of these questions are critical because the answers will help the property developer asses the risk and potential returns on his investment. Data needed to complete this analysis includes COVID-19 data for the state of Texas, COVID-19 date for the entire United States, and COVID-19 mobility data for the world. While these datasets seem broad, each dataset contains necessary features to conduct this analysis, which will be revealed further in the report. By understanding how different counties fared during the pandemic, the developer can make an informed decision regarding where he wants to build, ensuring that the chosen location offers stability and growth potential, even during unforeseen circumstances.

Income Data in Texas Counties

Data Collection, Quality, and Exploration

Objects to Cluster

The objects to be clustered in this analysis are the counties in Texas. To identify which counties demonstrated resilience during the COVID-19 pandemic, income and rent burden metrics will be analyzed alongside general population data. Some key features for clustering include median income, income per capita, a couple rent burden levels, and a few income distribution brackets. These factors provide a comprehensive picture of each county's economic resilience and ability to maintain stability during times of crisis.

By examining income distribution and wealth concentration, we can determine which counties have strong economic foundations. This, in combination with COVID-19 case and death data, will guide the stakeholder in making an informed decision on where to invest in developing a mixed-use building. Counties that managed to sustain consumer traffic and economic activity during the pandemic will likely offer more stability and growth potential for future business ventures.

Features for Clustering

The features analyzed for clustering relate to the category of income and wealth, which are critical for understanding economic resilience. These features include income brackets, median income per capita, rent burden percentages, and population statistics. Each of these features play a significant role in assessing to what capacity the county can withstand a widespread challenge such as the COVID-19 pandemic.

- **Income Levels:** The distribution of households across various income levels can provide insight into a county's overall economic health and resilience.
- **Rent Burden:** High rent burden percentages indicate financial strain on households, which can affect their ability to manage crises effectively.
- **Median Income and Income per Capita:** These metrics serve as broad indicators of wealth within a county. Wealthier counties typically have more resources to navigate economic shocks and support their communities during difficult times.
- **Population:** Including population statistics allows for a more accurate interpretation of COVID-19 impacts by normalizing the number of cases and deaths based on county size.

By clustering counties based on these features, we can identify different income and wealth profiles that may correlate with their resilience during the pandemic. This analysis will enhance our understanding of which counties were better equipped to handle the economic and social disruptions caused by COVID-19, ultimately aiding the stakeholder in making informed investment decisions.

Table of Features and Basic Statistics

Table 1: Basic Statistics of Key Features

Feature	Mean	SD	Min	Max
Median Income	49894.339	12132.676	24794	93645
Income per Capita	24859.020	5240.752	12543	41609
Rent > 50% Income	2976.004	13179.056	0	158668
Rent 30-35% Income	1180.870	5203.838	0	61305
Income < 10,000 USD	2469.768	8601.256	0	98715
Income 50,000-59,999 USD	2945.197	10790.454	3	122390

Income 100,000-124,999 USD	3205.157	11657.055	0	131467
Total Population	107951.228	389476.863	74	4525519

Because there are a lot of features that represent the wealth and income category, features were chosen that represent the most critical dimensions of income distribution and rent burden, while avoiding overly granular breakdowns. This selection captures the distribution of wealth (from low to high incomes), general population data, and rent burden, which are the most relevant features for analyzing the economic stability of a county.

- **Median Income:** This gives a central measure of income distribution in a county.
- **Income per Capita:** Shows wealth distribution on a per-person basis, which complements median income.
- **Rent Over 50 Percent:** This is a key indicator of severe rent burden, which can signify economic strain in a county.
- **Rent 30 to 35 Percent:** This provides a threshold of moderate rent burden.
- **Income Less than $10,000:** Reflects the population in extreme poverty, which is crucial for understanding economic vulnerability.
- **Income $50,000 - $59,999:** Represents household earning within a middle-income bracket, which can provide insight to stability of the county's middle class.
- **Income $100,000 - $124,999:** Indicates a higher income range, reflecting the proportion of relatively affluent residents.

Scale of Measurement

All of the features listed below are ratio scales because they have a true zero point (e.g., zero income, zero population) and allow for meaningful arithmetic operations (e.g., calculating differences, ratios).

Table 2: Measurement Scales for Features

Feature	Scale	Description
Median Income	Ratio	Income in USD
Income per Capita	Ratio	Per capita income in USD
Rent > 50% Income	Ratio	Households paying >50% income in rent
Rent 30-35% Income	Ratio	Households paying 30-35% income in rent
Income <10,000 USD	Ratio	Households earning <10,000 USD
Income 50,000-59,999 USD	Ratio	Households earning 50,000-59,999 USD
Income 100,000-124,999 USD	Ratio	Households earning 100,000-124,999 USD
Total Population	Ratio	Total county population

Measures for Similarity/Distance

For clustering analysis, various measures of similarity or distance can be employed based on the features used. The following measures are particularly relevant:

- **Euclidean Distance:** This is the most widely used distance measure, calculated as the straight-line distance between points in a multi-dimensional space. It is especially effective for continuous numerical data such as income or population figures, where the relationships between data points can be interpreted geometrically. Euclidean distance captures the direct linear relationship between observations, making it intuitive and straightforward for visualizing proximity in clustering contexts. [3]

- **Manhattan Distance:** This measure calculates the distance between two points by summing the absolute differences of their coordinates. Manhattan distance is useful when dealing with outliers or when the scale of measurement varies among features. It reflects a grid-like path, which can be advantageous in scenarios where a more robust metric against extreme values is required. In urban environments, for example, it mirrors the layout of streets. [4]
- **Standardization/Normalization:** When features exhibit wide ranges, normalizing the data before applying distance measures is beneficial. This ensures that each feature contributes equally to the distance calculation, preventing features with larger scales from disproportionately influencing results. [5]

In this analysis, a combination of standardized/normalized distance and Euclidean distance will be utilized. The data will first be standardized to ensure that each feature contributes equally to the distance calculation. The choice of Euclidean distance is justified by its prevalence and effectiveness for income and population data, which typically exhibit continuous numerical characteristics. It provides a clear and meaningful way to measure similarity between counties based on economic and demographic factors.

Normalization/Standardization

Standardization is essential for putting features on a similar scale, enabling meaningful comparisons across variables and preventing features with larger ranges or counts from dominating the analysis—especially in clustering algorithms. Given the wide range of values in the dataset, it was necessary to standardize the numerical features before proceeding with clustering or further analysis. The standardization was done using R and it transforms the data such that each feature has a mean of 0 and a standard deviation of 1. The county name was not standardized since it is a categorical variable. Since standardization is applied to numerical data, this feature was excluded from the process.

Modeling and Evaluation

K-Means Clustering

The K-Means clustering plot shows how Texas counties are grouped into two distinct clusters (1 and 2). Each point on the plot represents a county, and the clusters are visualized using different shapes and colors. The boarder around each cluster provides a visual boundary for each group. This clustering helps uncover patterns among the counties based on their economic resilience during the COVID-19 pandemic.

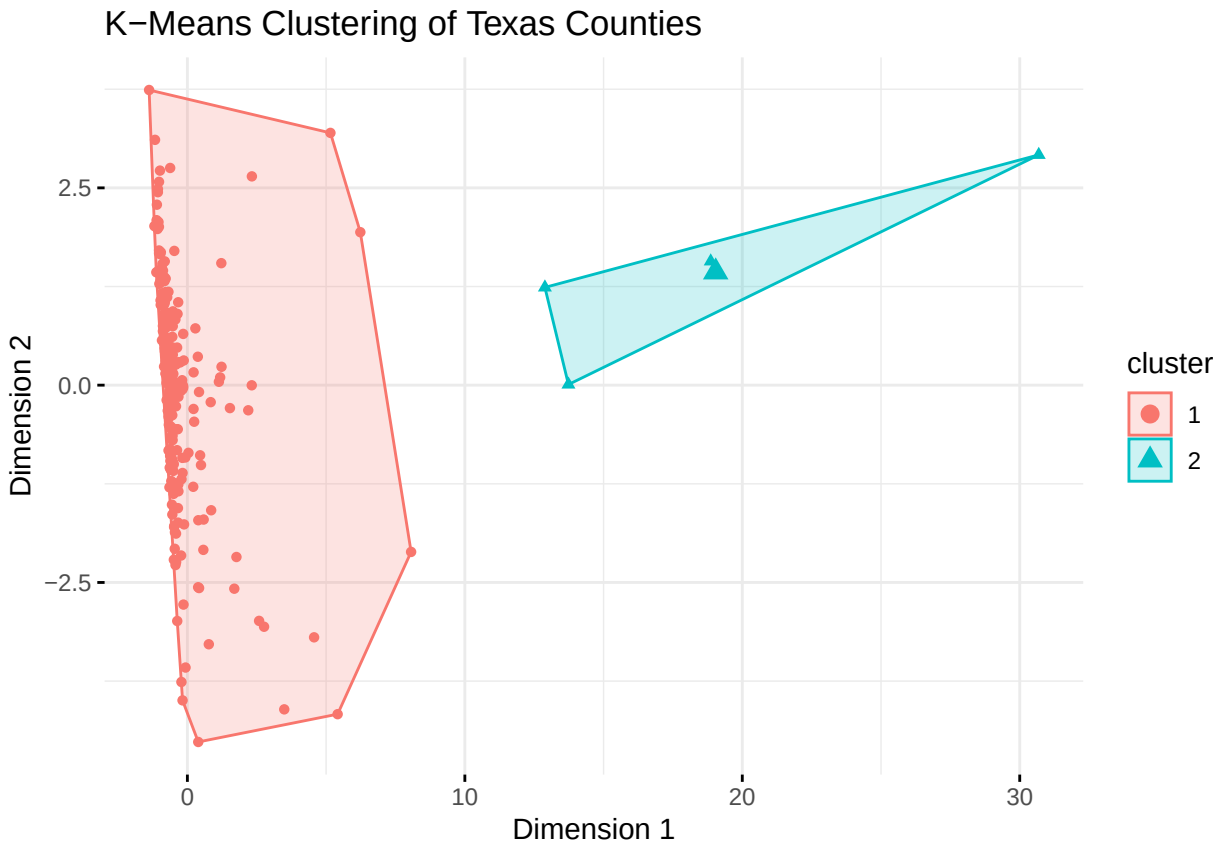


Figure 1: K-Means Clustering of Texas Counties

A summary statistics table is used to provide a detailed breakdown of the average values for key features across the two clusters identified through K-Means clustering. Each cluster represents a distinct group of Texas counties with similar economic, demographic, and pandemic characteristics. The table displays the average median income, income per capita, rent burden levels (both for households spending more than 50% and 30-35% of their income on rent), confirmed COVID-19 cases, deaths, and total population for each cluster.

Table 3: Summary Statistics by Cluster

cluster	Avg Median Income	Avg Income per Capita	Avg Rent > 50%	Avg Rent 30-35%	Avg Confirmed Cases	Avg Deaths	Total Population
1	49780.86	24786.04	1551.7	615.408	5078.896	89.052	65864.8
2	56987.00	29420.25	91995.0	36522.250	217182.500	2529.000	2738352.8

Cluster 1 has a high concentration of points while Cluster 2 captures a much smaller group. This incredibly uneven distribution suggests the clustering is not a great representation of the counties.

- **Average Median Income:** Cluster 1 had an average median income of 47,780.86 USD and Cluster 2 had an average median income of 56,987.00 USD. This shows a very moderate income difference of less than 10,000 USD.
- **Average Deaths:** This is a pretty big discrepancy as Cluster 2 experiences 2,529 average deaths while Cluster 1 only experienced 89. This indicates that Cluster 2 captures a very specific subset of counties with higher COVID-19 mortality.

This clustering does not offer a clear, interpretable division aligned with economic or pandemic impact

metrics, as variation between clusters in largely skewed. This unsupervised K-Means clustering could perform better with supervision.

Suitable Number of Clusters The Elbow Method plots the WSS (Within-Cluster Sum of Squares) for different number of clusters. WSS measures how tightly the data points are grouped around the centroids of the clusters. After a certain point, adding more clusters provides diminishing returns, meaning the reduction in WSS becomes negligible. The optimal number of clusters is found at the "elbow" point, where the rate of decrease in WSS sharply levels off. In the following elbow plot, the elbow occurs around 2 clusters.

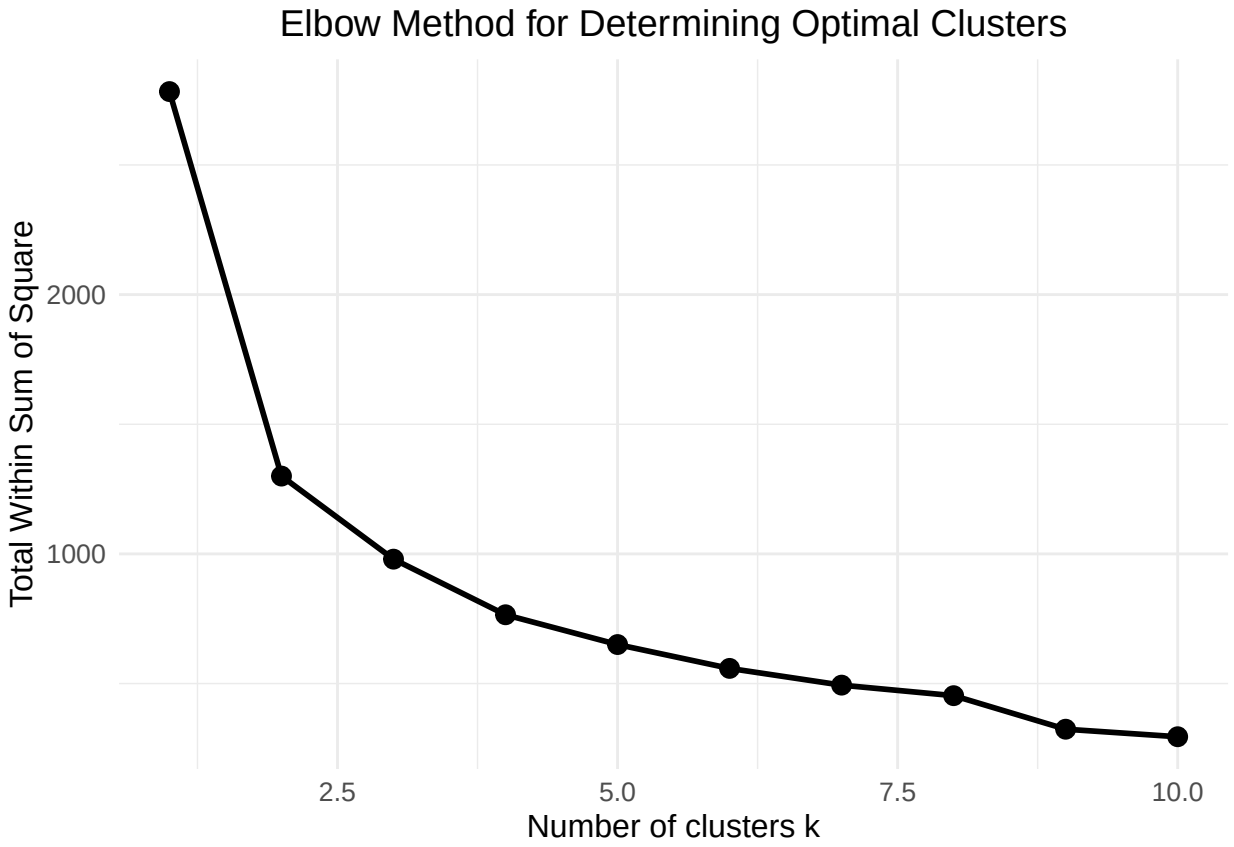


Figure 2: Elbow Method for Determining Optimal Clusters

The Silhouette Method evaluates how well each data point fits within its assigned cluster compared to other clusters. The Silhouette score ranges from -1 to 1, with values close to 1 meaning that the points are well-clustered. In the following Silhouette chart, the peak occurs at 2 clusters.

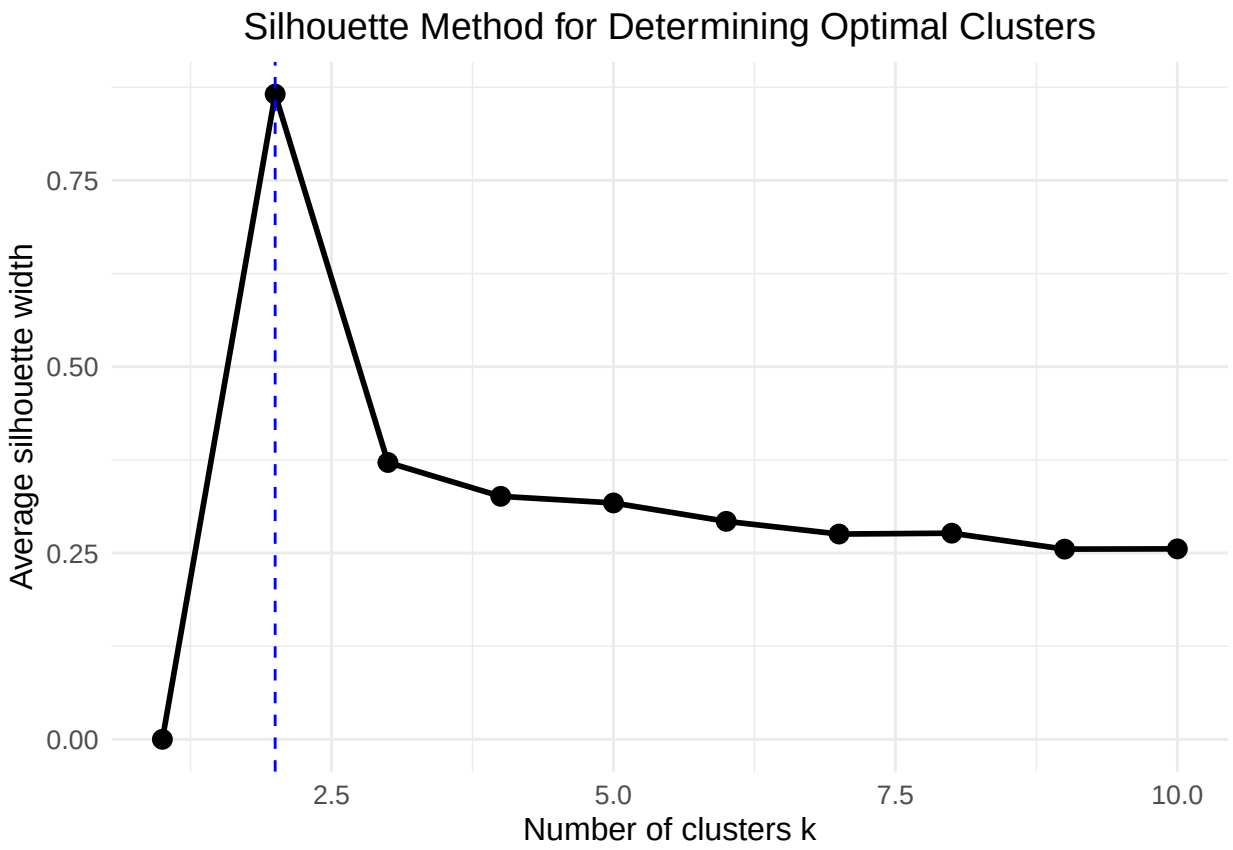


Figure 3: Silhouette Method for Determining Optimal Clusters

After considering both of these models, it was decided to do 2 clusters. Because of the consistency across both methods, 2 clusters was the clear choice.

Unsupervised Evaluation A silhouette plot is used as the unsupervised evaluation to assess the quality and cohesion of clusters generated by the K-Means algorithm. The silhouette width is a metric used to evaluate how well each data point fits within its assigned cluster relative to other clusters. Values near 1 indicate that data points are well-matched to their own cluster and poorly matched to neighboring clusters (high-quality clustering). Values near 0 suggest that the data points lie equally far from two neighboring clusters (uncertainty in clustering assignments).

```
##   cluster size ave.sil.width
## 1       1  250          0.87
## 2       2    4          0.45
```

Figure 4: Silhouette Plot for K-Means Clustering

- **Cluster 1 (Red):** The size of this cluster is 250 points with an average silhouette width of 0.87. This can be interpreted to mean that most of the data points are well-separated from other clusters and they have a high degree of cohesion. Cluster 1 can be defined as compact and well-defined within the data.
- **Cluster 2 (Blue):** The size of this cluster is 4 points with an average silhouette width of 0.45. This can be interpreted to mean that the points within the cluster are less cohesive and lack a clear grouping, leading to a weaker clustering group.

Ground Truth Feature The feature used for the ground truth features is the COVID-19 deaths, comparing the clusters to the death-to-case ratio category (Lower: <0.025, Higher: >0.025). The motivation for choosing this feature as the ground truth feature stems from the goal of examining how wealth and economic conditions impacted the pandemic outcomes.

- The analysis seeks to determine if wealthier counties, indentified through income-realted clustering, exhibit better pandemic performance measured through lower mortality rates relative to confirmed cases.
- By using "Lower" and "Higher" categories, the analysis is simplfied, making it easier to interpret and compare income groups. Additionally, to truly show a comparison between unsupervised and supervised clustering, it was decided to stay consistent with 2 clustering groups.
- Mortality rates serve as a crutial public health indicator, directly reflecting the severity of the pandemic's impact on a county. This feature can provide meaningful insight into how income of a county can indicate resilience and lower mortality rates for a pandemic like COVID-19.

The choice of this feature thus helps explore the correlation between economic factors and the severity of

the pandemic's impact, offering critical and clear insights into the resilience and vulnerabilities of different counties.

```
##
##      Lower Higher
##    1   145    105
##    2     4      0
```

Figure 5: Ground Truth Cluster Comparison

Supervised Evaluation The K-Means clustering plot shows how Texas counties are grouped into two distinct clusters (1 and 2). The features used for clustering are the death_case_ratio (the ratio of COVID-19 deaths to confirmed cases) and income_per_capita. The features were scaled to have a mean of zero and standard deviation of one, making sure that both features contribute equally to the clustering process. The difference between this clustering and the previous K-Means clustering is that this is supervised, meaning that the x and y axis are intentionally chosen to provide a simplified and clear result. The clusters represent groups of counties that share similar characteristics in terms of economic conditions and pandemic impact. Counties within each cluster exhibit more similarity to each other than to those in the other cluster.

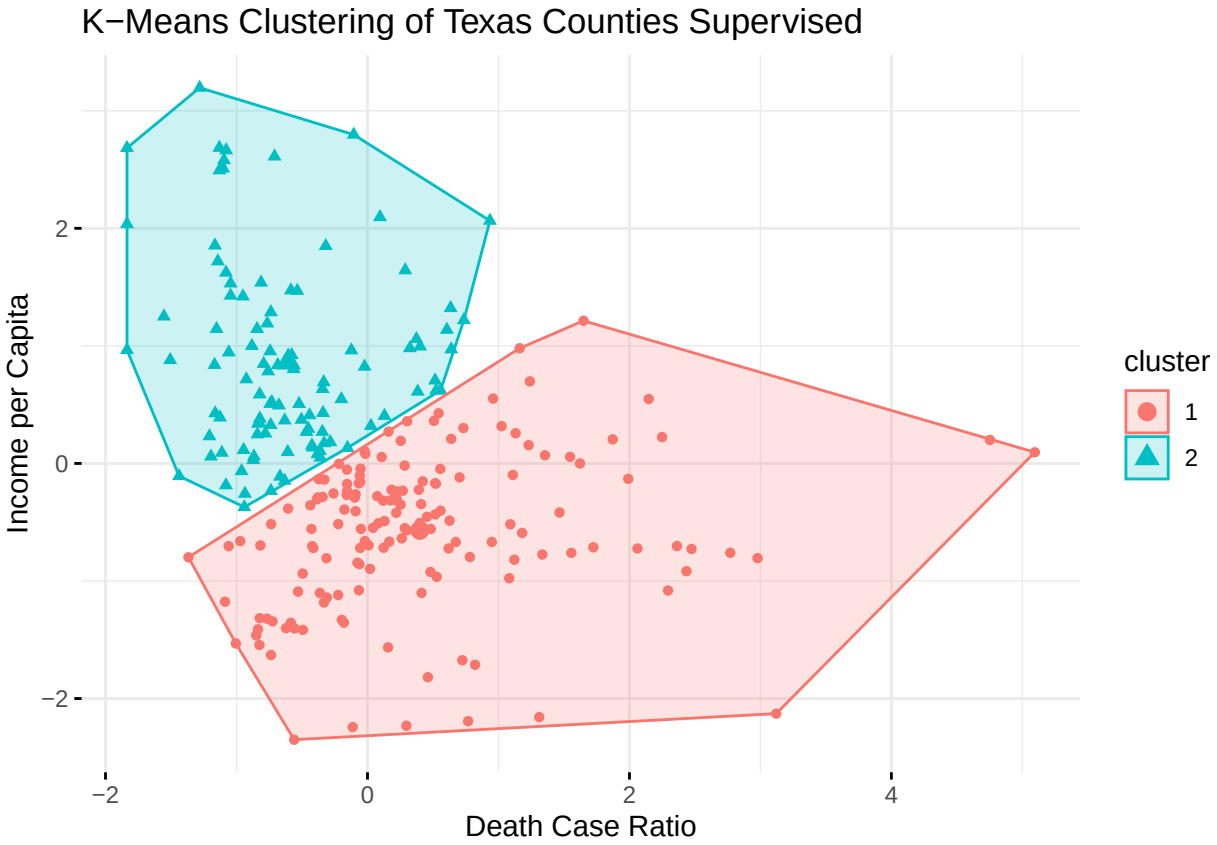


Figure 6: K-Means Clustering of Texas Counties Supervised

A summary statistics table, similar to the previous clustering method, is used to provide a detailed breakdown of the average values for key features across the two supervised clusters identified through K-Means clustering. Each cluster represents a distinct group of Texas counties with more similar economic, demographic, and pandemic characteristics. The table displays the average median income, income per capita, rent burden

levels (both for households spending more than 50% and 30-35% of their income on rent), confirmed COVID-19 cases, deaths, and total population for each cluster.

Table 4: Summary Statistics by Cluster

cluster	Avg Median Income	Avg Income per Capita	Avg Rent > 50%	Avg Rent 30-35%	Avg Confirmed Cases	Avg Deaths	Avg Death Case Ratio	Total Population
1	43937.72	21876.61	804.7255	297.6667	3309.477	80.98693	0.0300654	37969.71
2	58917.73	29376.93	6265.1683	2518.7921	16159.446	197.90099	0.0165567	213962.83

This clustering uses interpretable features: death case ratio and income per capita. This makes the clusters more meaningful, reflecting the economic status directly. The data is more evenly distributed between the two clusters, providing a clearer separation of counties. Clusters 1 and 2 have a relatively even distribution of points, suggesting that the clustering is a good representation of the counties.

- **Average Median Income:** Cluster 1 had an average median income of 58,917.73 USD and Cluster 2 had an average median income of 43,937.72 USD. This shows a very clear differentiation in economic status. The different is around 15,000 USD.
- **Average Death Case Ratio:** Cluster 1 has an average ratio of 0.0166 which is significantly lower than Cluster 2's average ratio of 0.0301. This highlights the relationship between economic conditions and pandemic outcomes more effectively.

The two K-Means clustering analyses (supervised and unsupervised) aim to categorize Texas counties based on economic and pandemic-related features, but they differ significantly in terms of clarity, precision, and interpretability.

Within Clusters 1 and 2, the counties are grouped based on their income and rent burdens. There are three income groups (Low: Income per Capita < 25,000 USD, Middle: 25,000 USD <= Income per Capita < 40,000 USD, High: Income per Capita > 40,000 USD) and two rent burden groups (Low: Rent over 50 Percent <= 5000, High: Rent over 50 Percent > 5000) that are used to provide more detailed comparison of the clusters.

Cluster 1 Analysis

- **Low Income & High Rent Burden:** With an average median income of approximately 39,219.50 USD and an average income per capita of 17,058.50 USD, this subgroup has a death per case ratio of 0.0258 and a relatively high population of about 591,047.25. This suggests that areas with low income and high rent burden may experience significant economic strain and relatively higher mortality rates.
- **Low Income & Low Rent Burden:** This subgroup, with a slightly higher average median income of 43,140.55 USD and income per capita of 21,125.65 USD, has a death case ratio of 0.0279. The total population is notably lower at 23,914.66, which may indicate that smaller populations with low rent burden still faced considerable pandemic challenges.
- **Middle Income & Low Rent Burden:** With the highest average median income and income per capita of Cluster 1, 48,879 USD and 26,590.88 respectively, the death per case ratio is 0.0419 which is the highest in Cluster 1. This could reflect that middle-income regions with low rent burdens still faced significant health challenges, potentially due to other socioeconomic or healthcare access factors.

Cluster 2 Analysis

- **Low Income & High Rent Burden:** With an average median income of 46,262 USD and income per capita of 24,273 USD, this subgroup has a relatively lower death case ratio of 0.0157 compared to its Cluster 1 counterparts. This indicates that economic vulnerability did not translate to equally severe pandemic outcomes across all metrics.

Table 5: Summary Statistics by Subgroups Within Clusters

cluster	income_group	rent_burden_group	Avg Median Income	Avg Income per Capita	Avg Death Case Ratio	Total Population
1	Low Income	High Rent Burden	39219.50	17058.50	0.0257519	591047.25
1	Low Income	Low Rent Burden	43140.55	21125.65	0.0279399	23914.66
1	Middle Income	Low Rent Burden	48876.00	26590.88	0.0418550	18993.50
2	High Income	High Rent Burden	90124.00	41609.00	0.0074628	914075.00
2	Low Income	High Rent Burden	46262.00	24273.00	0.0157121	245720.00
2	Low Income	Low Rent Burden	46604.71	23835.43	0.0117692	26067.43
2	Middle Income	High Rent Burden	62475.78	30879.67	0.0118649	956242.94
2	Middle Income	Low Rent Burden	58966.32	29439.27	0.0182851	41291.97

- **Low Income & Low Rent Burden:** This subgroup has an average median income of 46,604.71 USD and income per capita of 23,835.43 USD, with a low death case ratio of 0.0118. The total population is 26,067.43. The low rent burden appears to mitigate some of the negative effects of low income.
- **Middle Income & High Rent Burden:** With a median income of 62,475.78 USD and income per capita of 30,879.67 USD, this subgroup shows a death case ratio of 0.0119. This indicates a significant economic uplift compared to low-income groups, with moderate resilience in pandemic outcomes despite high rent burdens.
- **Middle Income & Low Rent Burden:** This subgroup has an average median income of 58,966.32 USD and income per capita of 29,439.27 USD, with a death case ratio of 0.0183. The lower rent burden may provide economic stability, but the death case ratio suggests room for improvement in health outcomes.
- **High Income & High Rent Burden:** This subgroup stands out with a high average median income of 90,124 USD and income per capita of 41,609 USD. The death case ratio is the lowest at 0.0075, suggesting that wealthier areas with high rent burdens may have been better equipped to manage the pandemic's impact. The total population in this group is substantial, at 914,075, indicating a dense but resilient economic region.

Higher income levels within Cluster 2 are associated with significantly lower death case ratios, highlighting the advantage of economic stability in managing the pandemic. Conversely, lower income groups in both clusters generally exhibit higher death case ratios. Rent burden appears to be a critical factor in economic vulnerability. However, even within high rent burden subgroups, those with higher income levels (Cluster 2) have better health outcomes. Subgroups with higher populations (e.g., high income, high rent burden areas in Cluster 2) show better resilience, possibly due to better infrastructure, healthcare access, and community resources.

The analysis reveals a clear relationship between income, rent burden, and pandemic outcomes. Wealthier areas, even with high rent burdens, were better at mitigating the negative impacts of COVID-19. These findings emphasize the importance of socioeconomic status and housing stability in public health crises. For stakeholders, this insight can guide investment and development decisions to prioritize areas with economic resilience or consider interventions to support vulnerable regions.

The following visualization illustrates the K-Means clustering results for Texas counties based on two critical

features: the Death Case Ratio and Income per Capita. The clusters are color-coded and outlined with borders, representing the original cluster boundaries from the K-Means algorithm. Each point within the plot is labeled by income group and rent burden status, providing additional context about economic and housing conditions within each cluster.

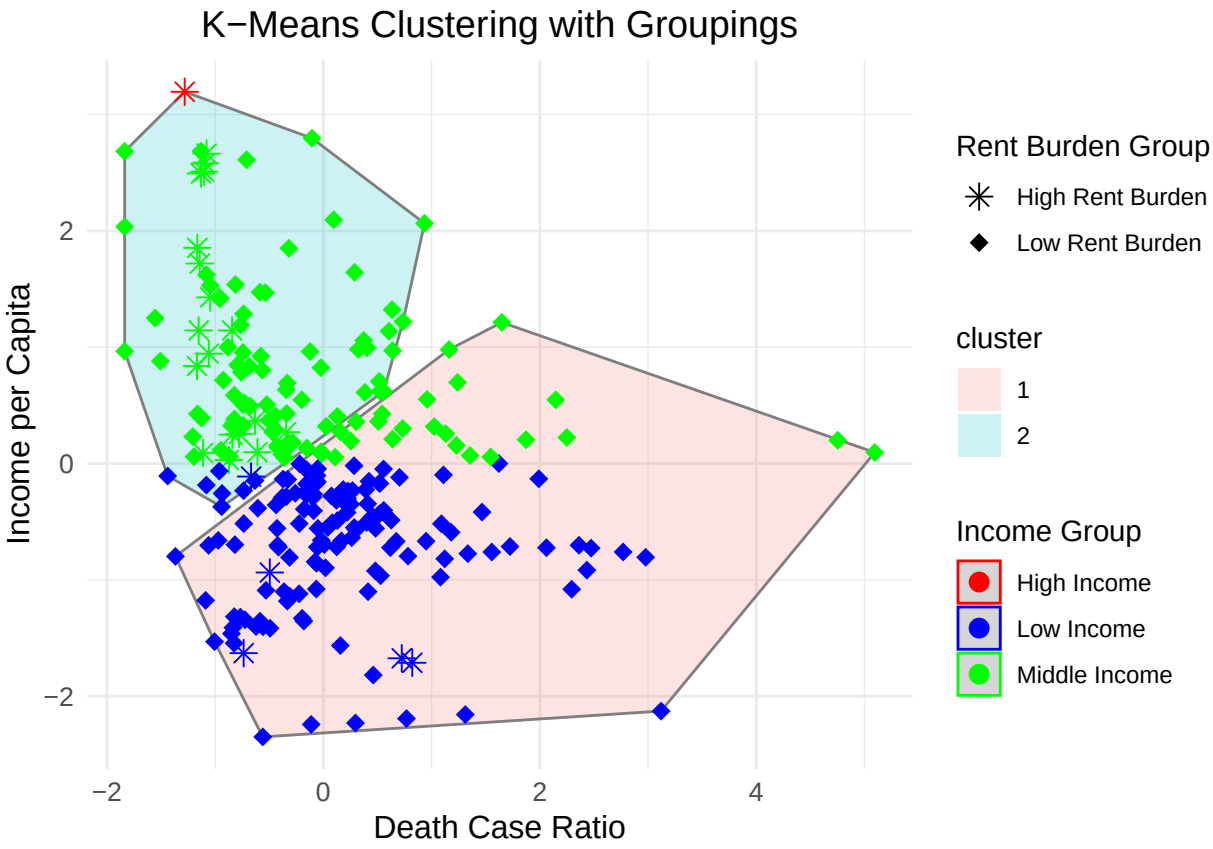


Figure 7: K-Means Clustering with Groupings

Cluster 1 (Red Region) This cluster is characterized by a higher Death Case Ratio and generally lower ot middle Income per Capita. The blue points represent low-income groups, and the density of these points suggests a strong presence of economically vulnerable areas within this cluster. The green points, representing middle-income groups, are present but less dense compared to the low-income group. Notably, this cluster contains both high and low rent burden subgroups, with high rent burden subgroups (indicated by star-shaped markers) mixed throughout. This implies that some areas within this cluster experience compounded economic stress, both in terms of income and rent burden, which could exacerbate health outcomes.

Cluster 2 (Blue Region) This cluster encompasses areas with a lower Death Case Ratio and generally higher Income per Capita. The red points indicate high-income areas, clustered toward the upper end of the income per capita axis, reflecting wealthier regions with better pandemic outcomes. There is a significant presence of green points representing middle-income groups, indicating that this cluster captures a range of moderately affluent areas. These areas seem to have fared better in terms of health outcomes compared to Cluster 1. High-income areas (red points) appear to have a mix of high and low rent burden groups, but even those with high rent burdens display relatively low Death Case Ratios. This suggests that wealthier regions, even with high rent burdens, may have had resources to mitigate the pandemic's effects.

The clustering highlights a strong correlation between income and health resilience. Higher income per capita is associated with lower Death Case Ratios, likely due to better access to healthcare, resources, and infrastructure to manage health crises. Low-income areas, especially those burdened by high housing costs,

Table 6: Purity Scores by Grouping

Grouping	Purity_Score
Income Groups	0.8700787
Rent Burden Groups	0.9055118

appear more vulnerable. The presence of both low and high rent burden groups in Cluster 1 suggests that financial strain could amplify the negative impact of the pandemic. Middle-income areas straddle both clusters, indicating that not all middle-income regions experienced the pandemic uniformly. Factors beyond income, such as healthcare infrastructure, population density, or social support, could influence outcomes.

The following table presents the purity scores for the two different subgroup classifications: Income Groups and Rent Burden Groups.
Purity is a metric used to evaluate the quality of clustering by measuring the extent to which clusters contain data points of a single class. A higher purity score indicates that the clusters are more homogeneous concerning the given grouping.

- **Income:** A purity score of 0.870 indicates that 87.0 percent of the data points within clusters are correctly grouped based on their income level (Low, Middle, or High Income). The relatively high score suggests that the clustering model is effective in distinguishing counties based on income characteristics, but there is still a bit of overlap or misclassification. The presence of overlap could imply that income levels alone do not fully explain the clustering structure.
- **Rent Burden:** The purity score for rent burden classification is 90.6 percent, which is slightly higher than the score for income groups. This suggests that the clusters are even better at grouping counties based on housing affordability stress, indicated by whether they experience high or low rent burdens. This could mean that rent burden is a more distinct factor in the clustering analysis.

Both scores are relatively high, indicating that the K-Means clustering captures meaningful distinctions in the data. Given the high but not perfect purity scores, there may be other unexamined variables influencing the clusters. Additional socioeconomic or demographic factors could be considered in future models to further refine the clustering results. Overall, the analysis suggests that while both income and rent burden are effective for understanding the clustering of Texas counties, housing stress appears to be a particularly significant and differentiating factor. This insight could be valuable for stakeholders aiming to address economic disparities or plan for community resilience.

Heirarchical Clustering

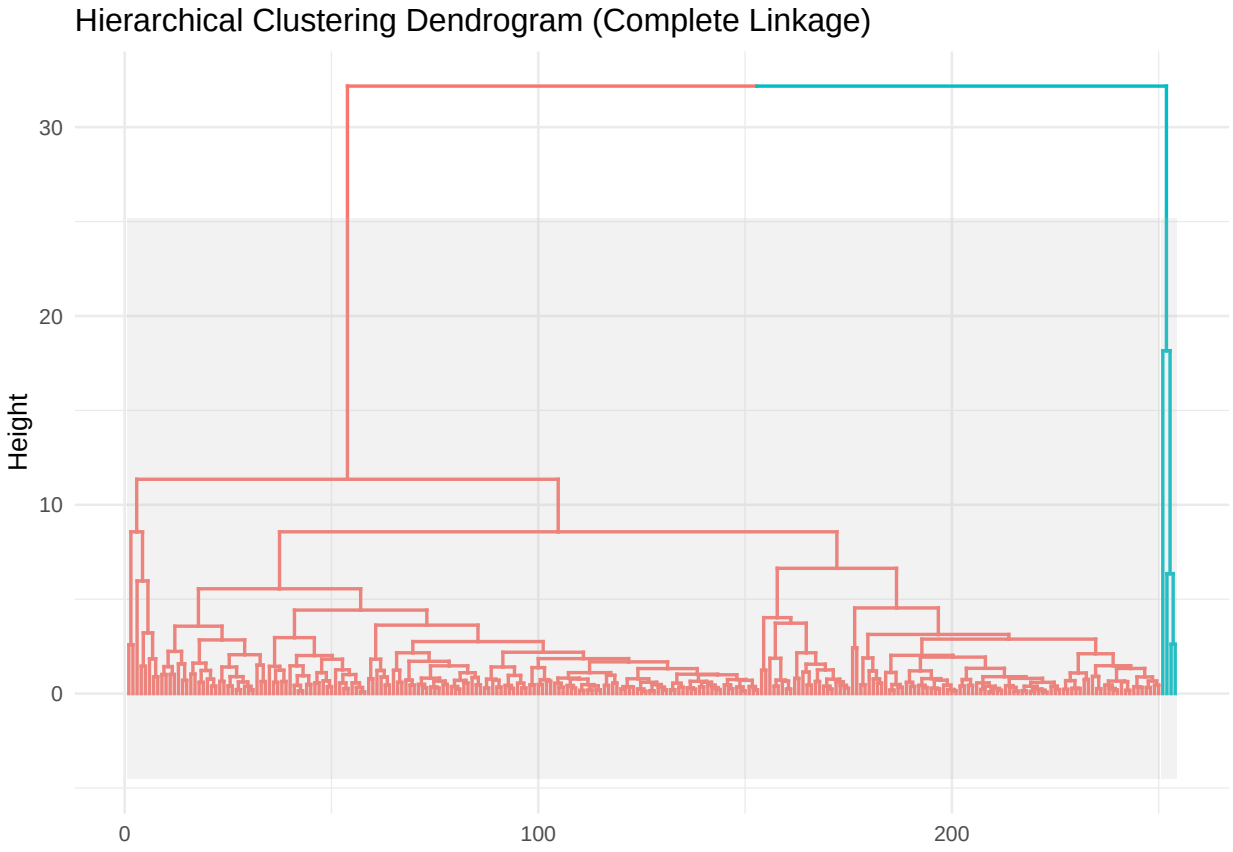
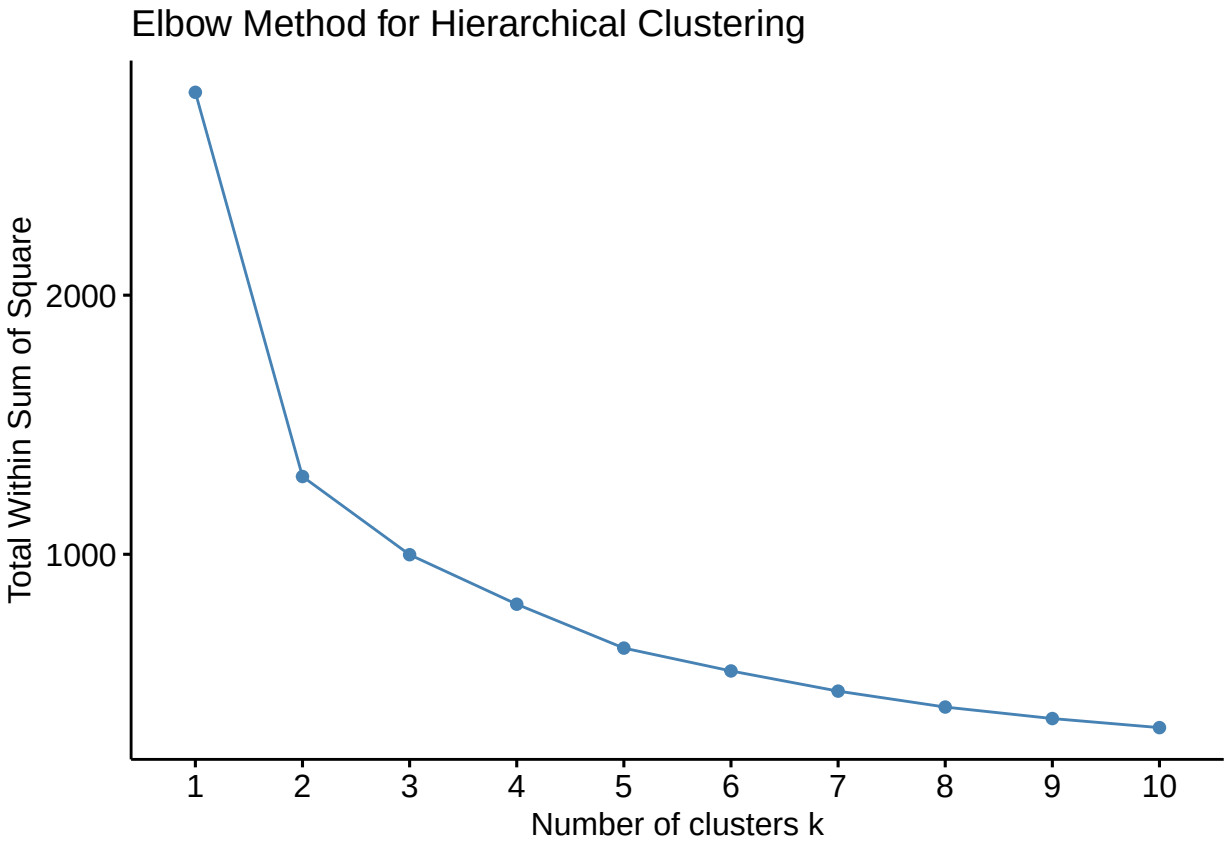


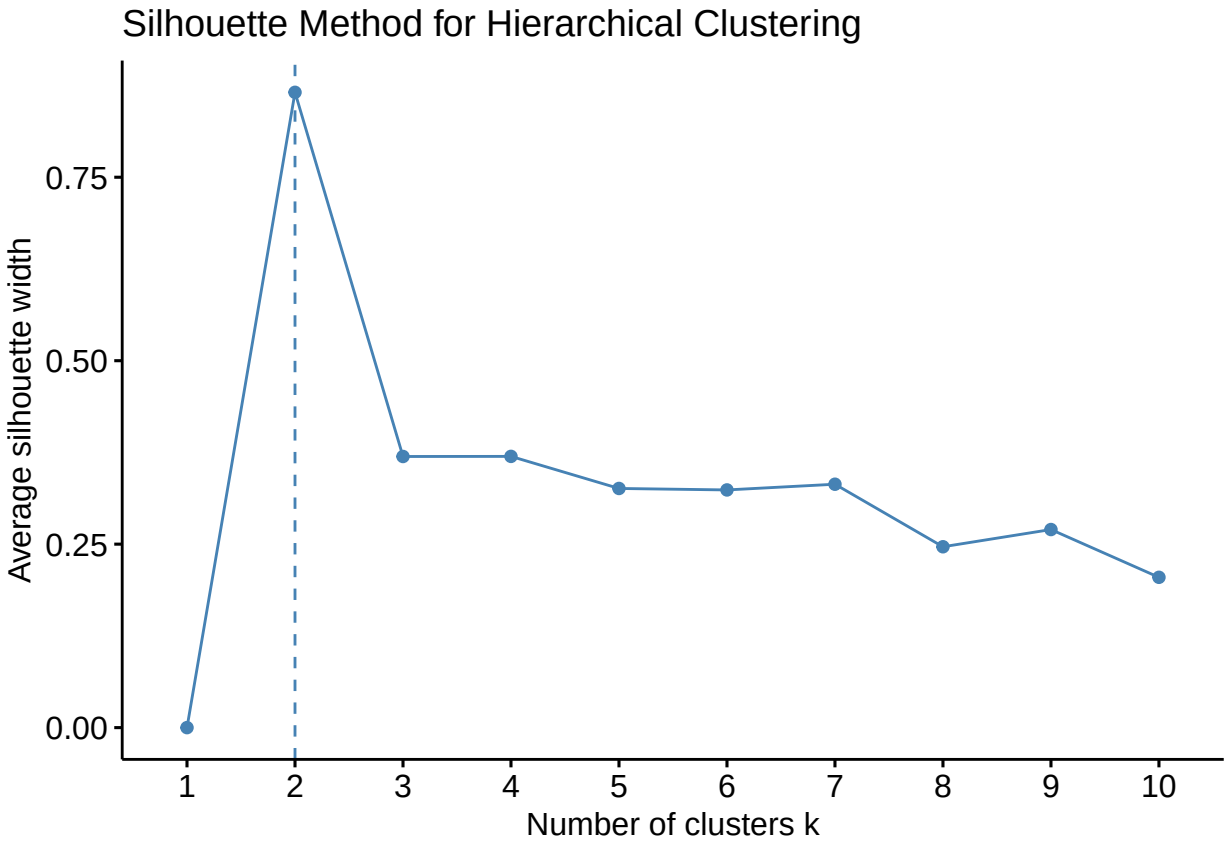
Table 7: Summary Statistics by Hierarchical Cluster

cluster_hc	Avg Median Income	Avg Income per Capita	Avg Rent > 50%	Avg Rent 30-35%	Avg Confirmed Cases	Avg Deaths	Total Population
1	49780.86	24786.04	1551.7	615.408	5078.896	89.052	65864.8
2	56987.00	29420.25	91995.0	36522.250	217182.500	2529.000	2738352.8

Suitable Number of Clusters The Elbow Method plots the WSS (Within-Cluster Sum of Squares) for different number of clusters. WSS measures how tightly the data points are grouped around the centroids of the clusters. After a certain point, adding more clusters provides diminishing returns, meaning the reduction in WSS becomes negligible. The optimal number of clusters is found at the "elbow" point, where the rate of decrease in WSS sharply levels off. In the following elbow plot, the elbow occurs around 2 clusters.



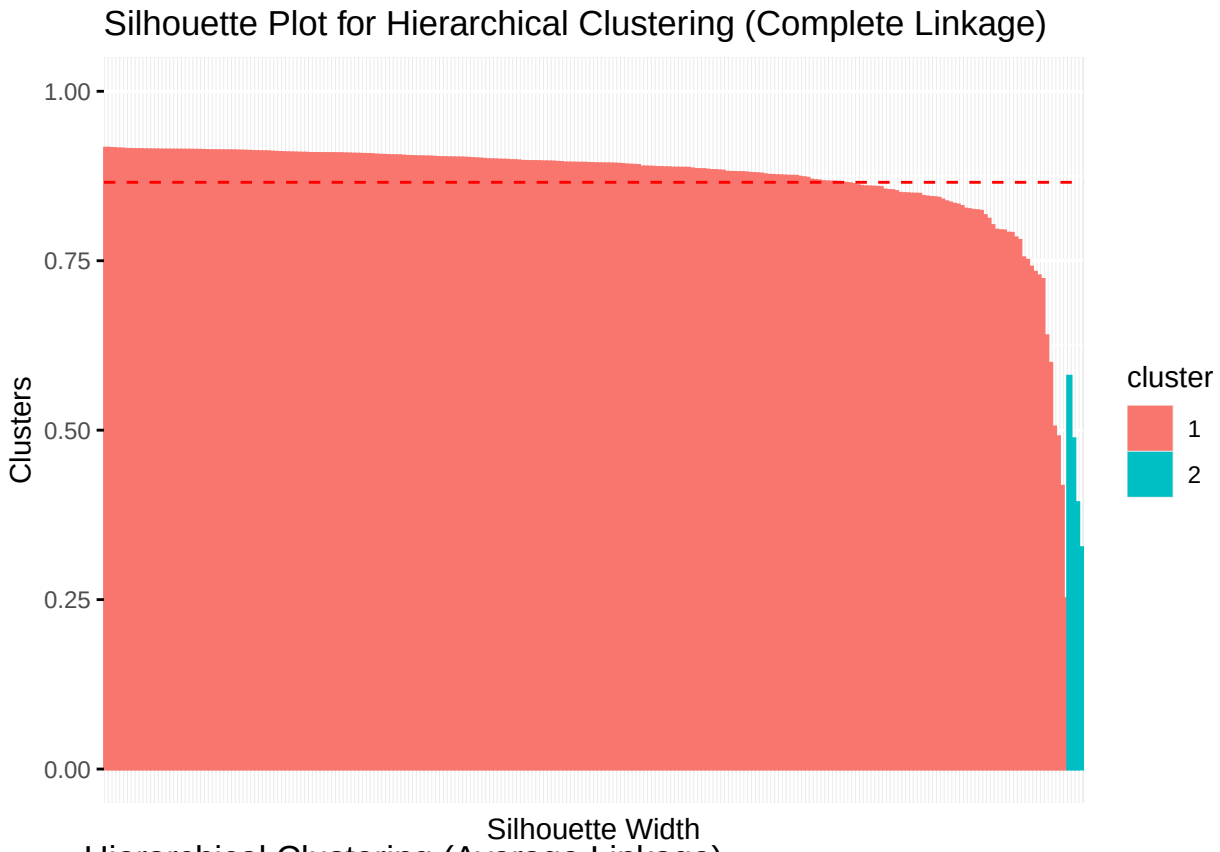
The Silhouette Method evaluates how well each data point fits within its assigned cluster compared to other clusters. The Silhouette score ranges from -1 to 1, with values close to 1 meaning that the points are well-clustered. In the following Silhouette chart, the peak occurs at 2 clusters.

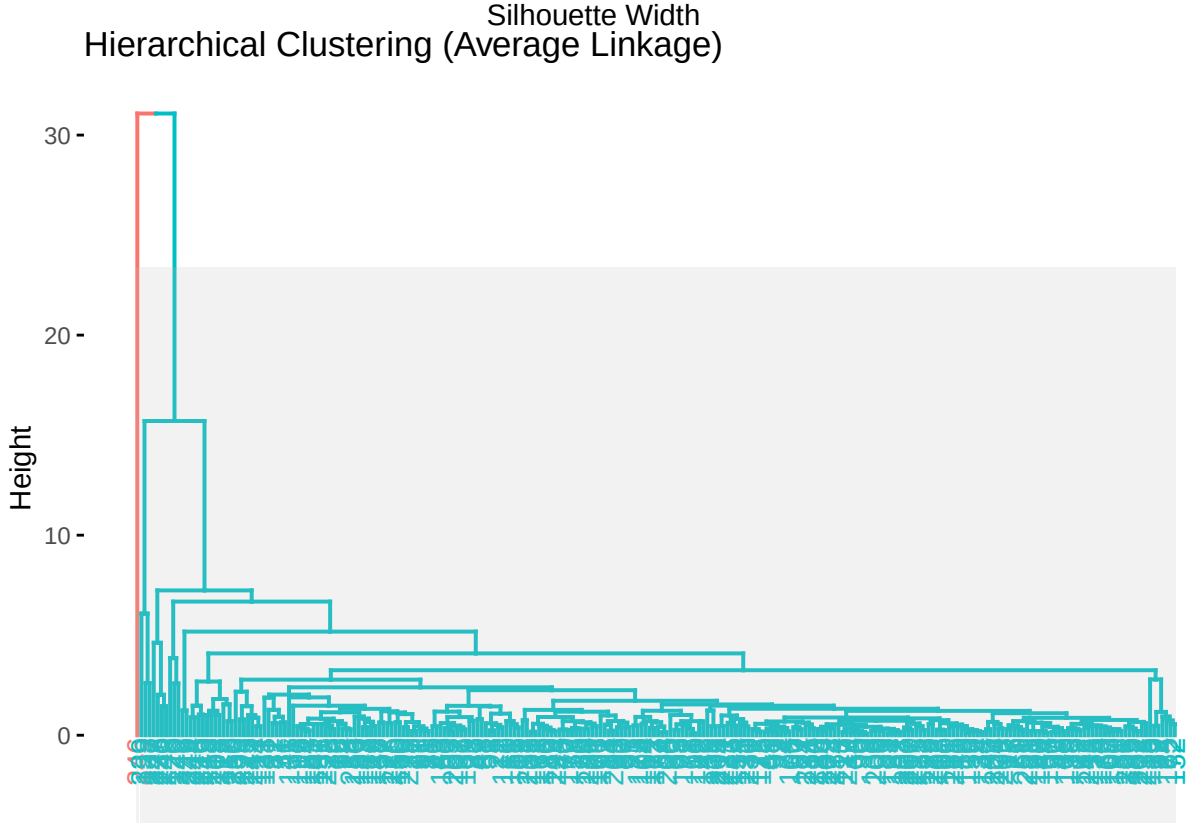


After considering both of these models, it was decided to do 2 clusters. Because of the consistency across both methods, 2 cluters was the clear choice.

Unsupervised Evaluation

```
##   cluster size ave.sil.width
## 1       1  250          0.87
## 2       2    4          0.45
```


Silhouette Plot for Hierarchical Clustering (Complete Linkage)
1.00
0.75
0.50
0.25
0.00
Clusters
Silhouette Width
cluster
1
2


Hierarchical Clustering (Average Linkage)
30
20
10
0
Height

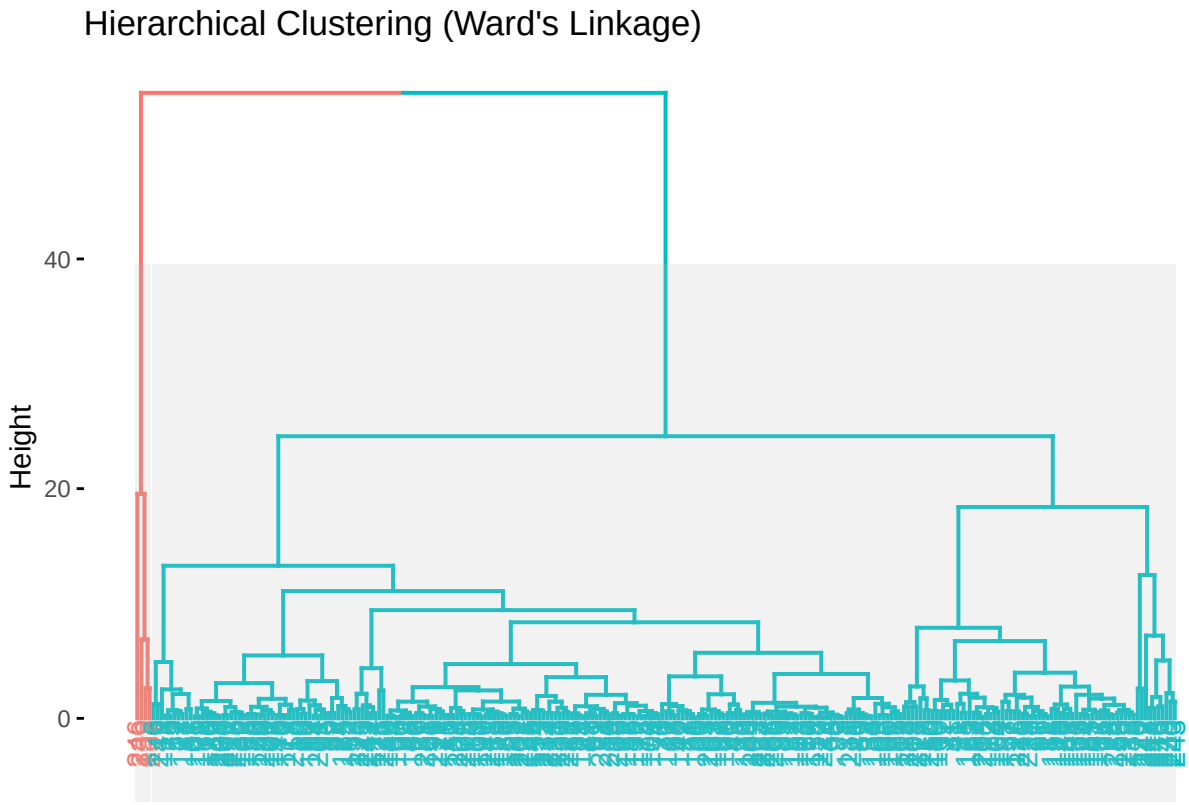
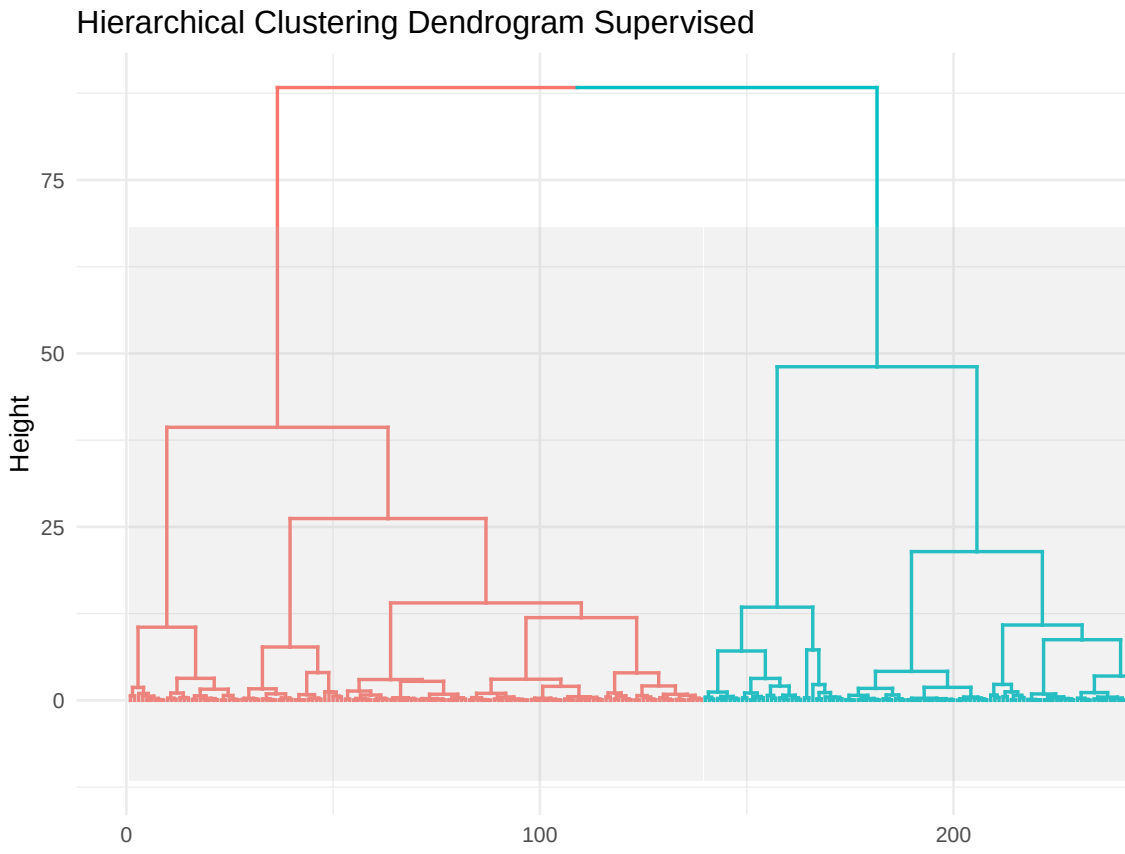


Table 8: Average Silhouette Widths by Linkage Method

Linkage_Method	Avg_Silhouette_Width
Complete	1.984252
Average	1.996063
Ward's	1.984252

Ground Truth Feature

```
##
##      Lower Higher
##   1    145    105
##   2      4      0
```



Supervised Evaluation

Table 9: Summary Statistics by Hierarchical Cluster (Supervised)

cluster_hc	Avg Median Income	Avg Income per Capita	Avg Rent > 50%	Avg Rent 30-35%	Avg Confirmed Cases	Avg Deaths	Avg Death Case Ratio	Total Population
1	42665.18	21229.01	829.9565	304.0609	3558.522	89.34783	0.0327401	39234.2
2	55875.29	27862.27	4751.5108	1906.2878	12440.460	159.02158	0.0180368	164803.4

Population Data in Texas Counties/Layer 2 Clustering

Data Collection, Quality, and Exploration

Objects to Cluster

The first part of the report clustered for the best performing affluent counties using the death to cases ratio and income per capita. Our model takes these optimal clusters of counties (cluster "2" in both methodologies for prior layer of clustering) and applies a second layer of clustering using the same methods.
That is to say, each methodology, K-means and Hierarchical, in this second layer receives cluster "2" from its prior layer and applies itself again (e.g. The K-means method in this layer applies itself to cluster "2" of the K-means result for death to cases ratio and income per capita).

Features for Clustering

The feature set for this second layer on which our K-means and Hierarchical methods apply themselves are population density and COVID-19 cases per thousand. This is done in order to find the highest population density counties with the lowest amount of COVID-19 cases.

- **Population Density** Found by first obtaining the total population for each Texas county using the Tidycensus R package, and then using census.gov's 2023 Geographic info API to retrieve the variable AREALAND_SQMI (land area in square miles) for each county in Texas. The total population for each county was divided by the total land area in square miles to create the population density feature (people per square mile of land).
- **Cases per Thousand** Found by taking total cases from COVID-19 Texas data set and dividing by total population for each county.

Table of Features and Basic Statistics

Table 10: Basic Statistics for Features

Feature	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Total Deaths	0.00	14.00	32.00	135.35	84.75	4024.00
Total Cases	1	505	1393	8854	3652	297629
Total Population	117	6835	18522	112738	51864	4680609
Area in Square Miles	127.2	835.7	908.7	1028.6	1043.4	6183.8
Population Density	0.1749	6.3404	21.8211	119.3629	66.3361	3003.4746
Deaths per Thousand	0.000	1.227	1.781	1.960	2.542	5.838
Cases per Thousand	8.547	60.369	78.495	80.639	97.758	179.111

Scale of Measurement

Table 11: Measurement Scales for Features

Features	Scale	Description
County Name	Nominal	Name of the county
Total Deaths	Ratio	Total Amount of Deaths in the County
Total Cases	Ratio	Total Amount of Cases in the County
Total Population	Ratio	Total Population of the County
Area in Square Miles	Ratio	Area of county in Square Miles
Population Density	Ratio	Population Density in People per Square Mile

Deaths per Thousand	Ratio	Covid Deaths per Thousand Inhabitants
Cases per Thousand	Ratio	Covid Cases per Thousand Inhabitants

Measures for Similarity/Distance

Since the clustering uses K-means and Hierarchical methodologies, Euclidean distance is used. Here are first-five-counties-in-the-data-set's euclidean distance for population density and cases per thousand.

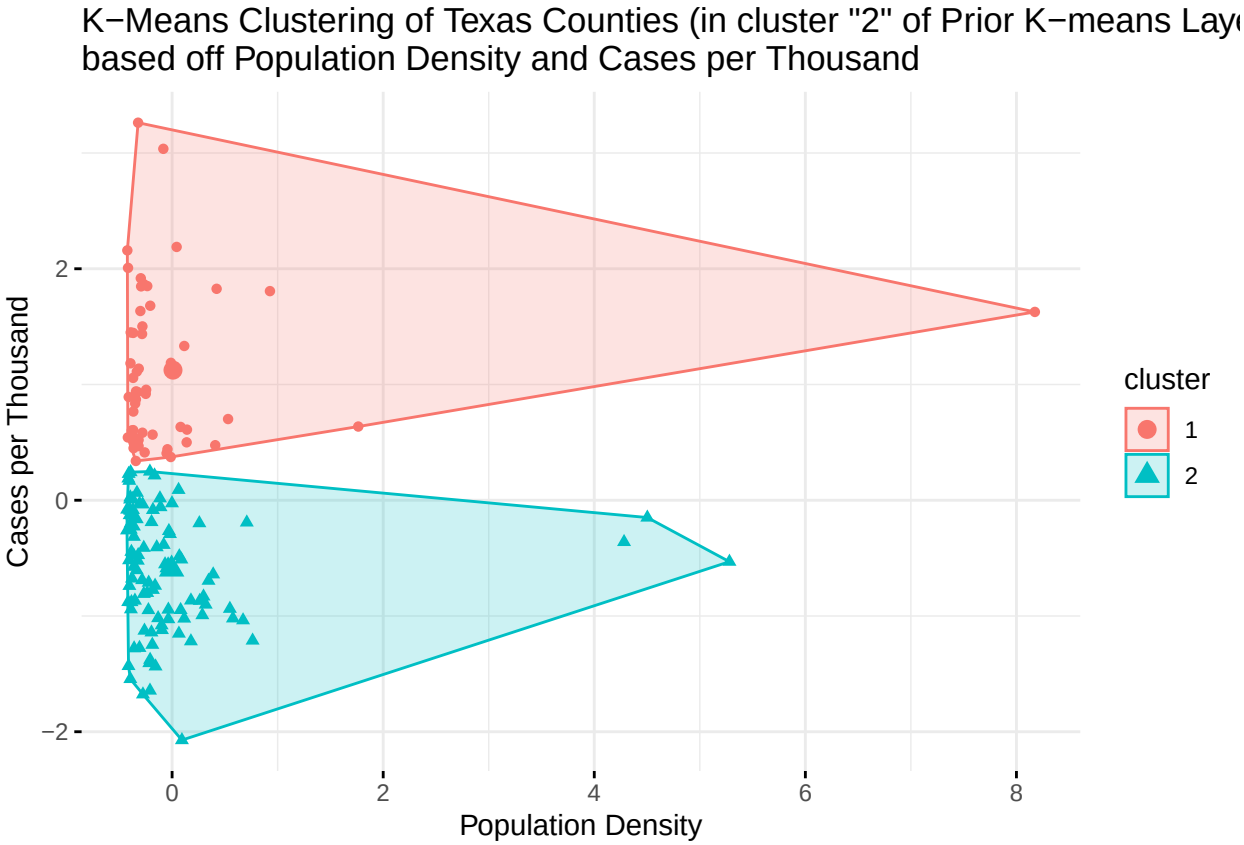
```
##          1         2         3         4
## 2 98.126290
## 3  6.201082 97.470224
## 4 29.604521 70.963217 31.467442
## 5 52.011261 51.166822 49.634707 33.772758
```

Normalization/Standardization

Numeric features in the data set were normalized using R's scale function, which normalizes a distribution using a standard Z-score normalization.

Modeling and Evaluation

K-Means Clustering



The clustering seemingly divides counties into either high cases per thousand or low cases per thousand. Summary Statistics:

- **K-means Cluster 1:** This cluster has a higher average cases and deaths per thousand with an average population density similar to that of the second cluster. This could be attributed as a "COVID-19 low performing" cluster of counties.

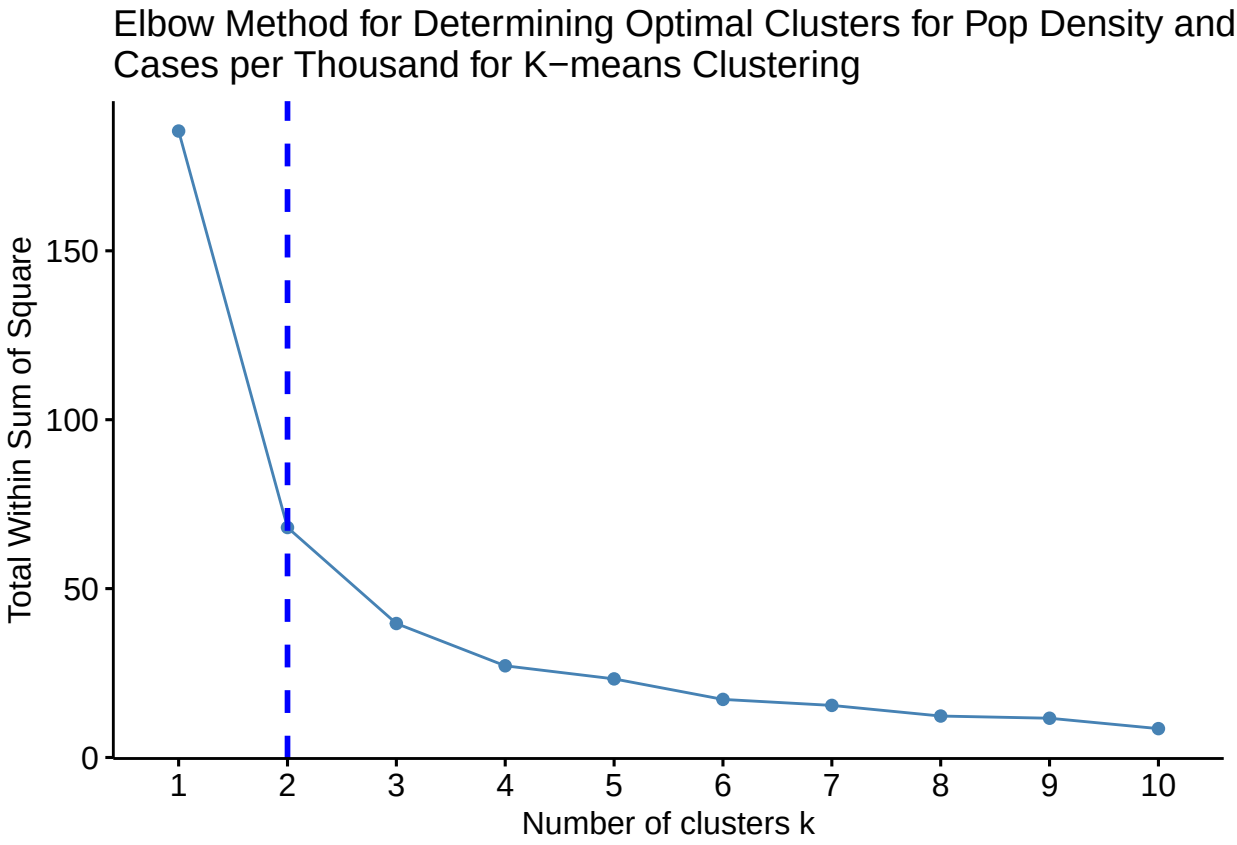
Table 12: Summary Statistics for K-means Cluster based on Population Density and Cases per Thousand

cluster	Avg Cases per thousand	Avg Deaths per thousand	Avg Population Density	Number of Counties
1	116.9643	2.809818	42.08026	53
2	66.9696	2.213788	40.84501	100

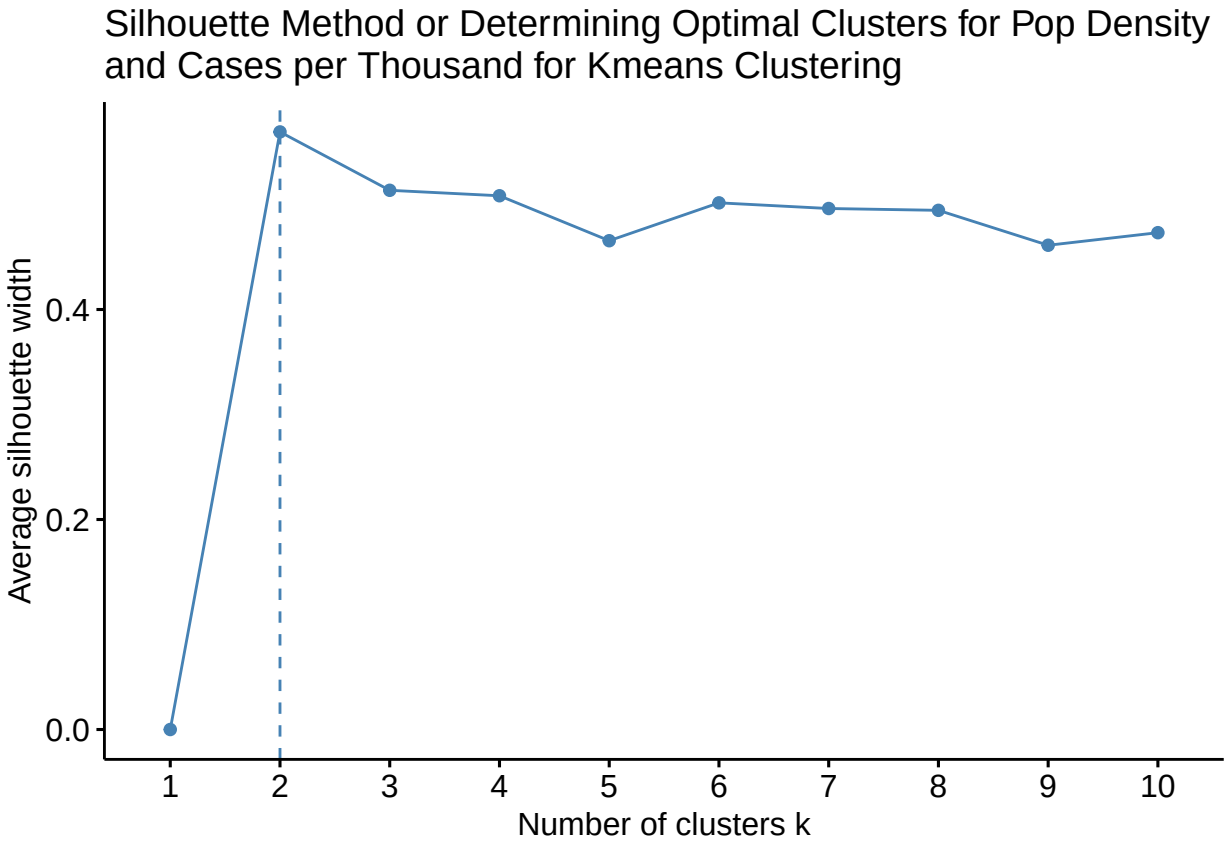
- **K-mean Cluster 2:** This cluster has a lower average cases and deaths per thousand with an average population density similar to that of the first cluster. This could be attributed as a "COVID-19 high performing" cluster of counties.

Suitable Number of Clusters Elbow Method

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

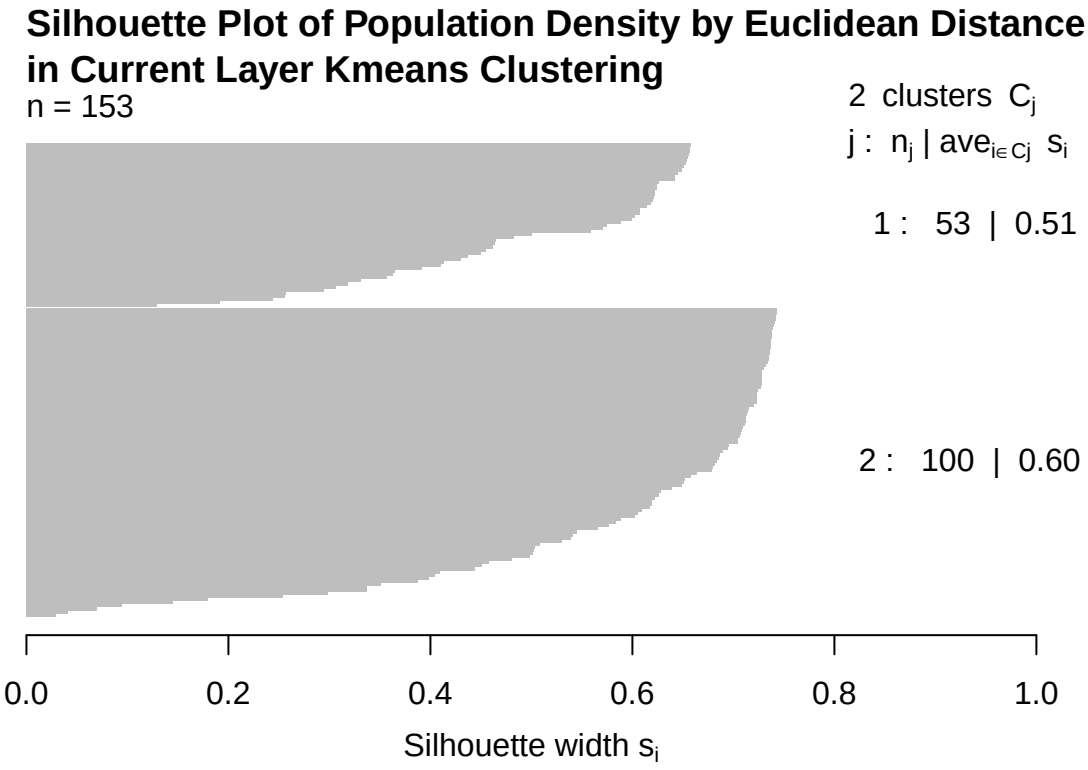


Silhouette Method



Our Elbow and Silhouette methods suggest our optimal amount of clusters for K-means is 2 clusters.

Unsupervised Evaluation Silhouette Width



Average silhouette width : 0.57

Our Average Silhouette width is not close to 1, which means that the centroids may not be as close to the middle of the cluster as they could be; however, the distribution of data points are fair.

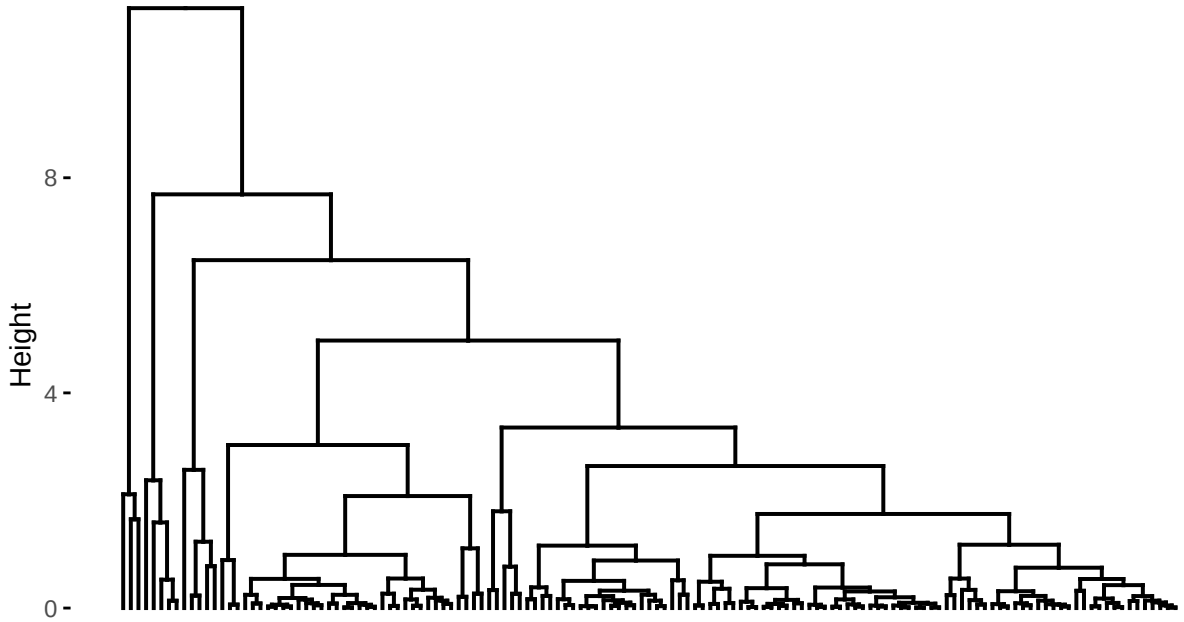
- **K-means Cluster 1:** With 53 points and an average Silhoutte plot of 0.51. The Cohesion is fair.
- **K-means Cluster 2** With 100 points and an average Silhoutte plot of 0.60. The Cohesion of this Silhouette plot could be interpreted as better than the prior cluster but not far better.

Heirarchical Clustering

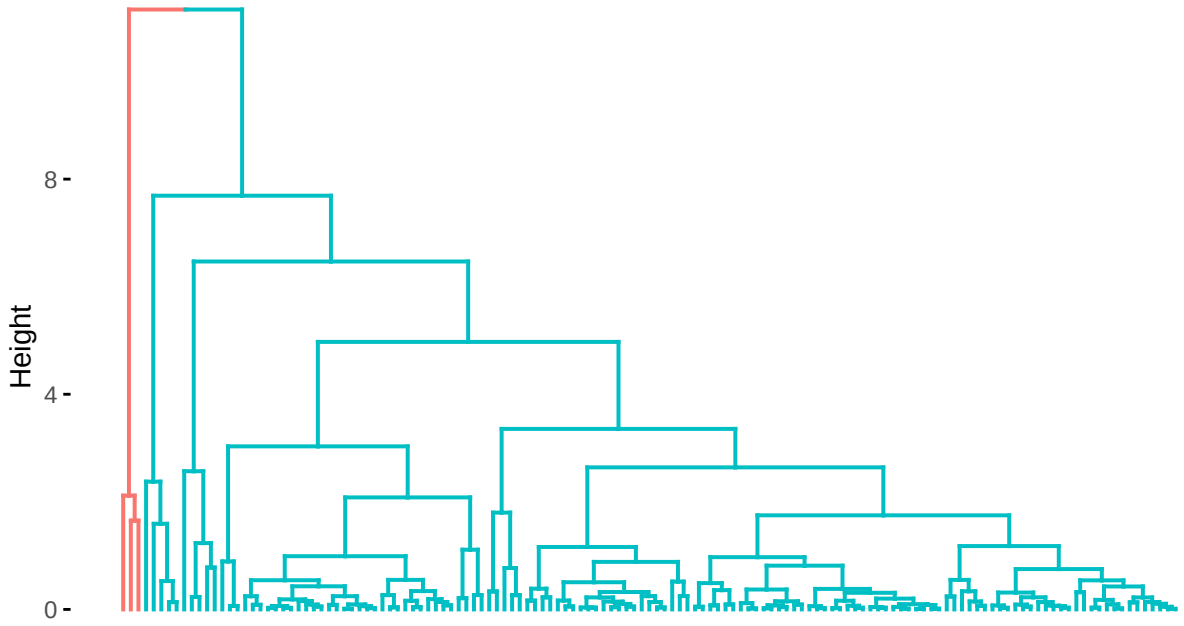
Dendogram

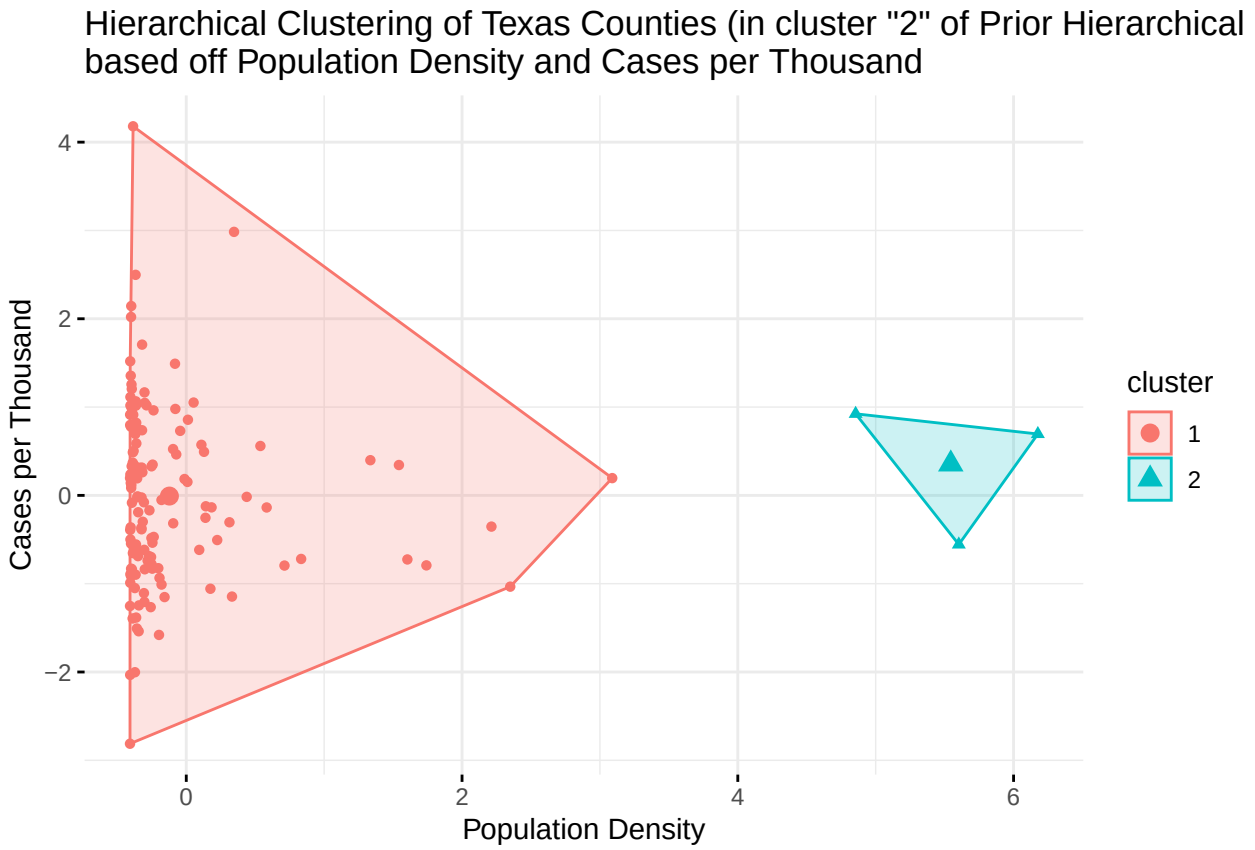
```
## Warning in dist(counties_hierarchical_pop_dens_et_cases_per_k.scaled): NAs
## introduced by coercion
```

Hierarchical Clustering based off Population Density and Cases per Thousand



Hierarchical Clustering based off Population Density and Cases per Thousand For Two Clusters





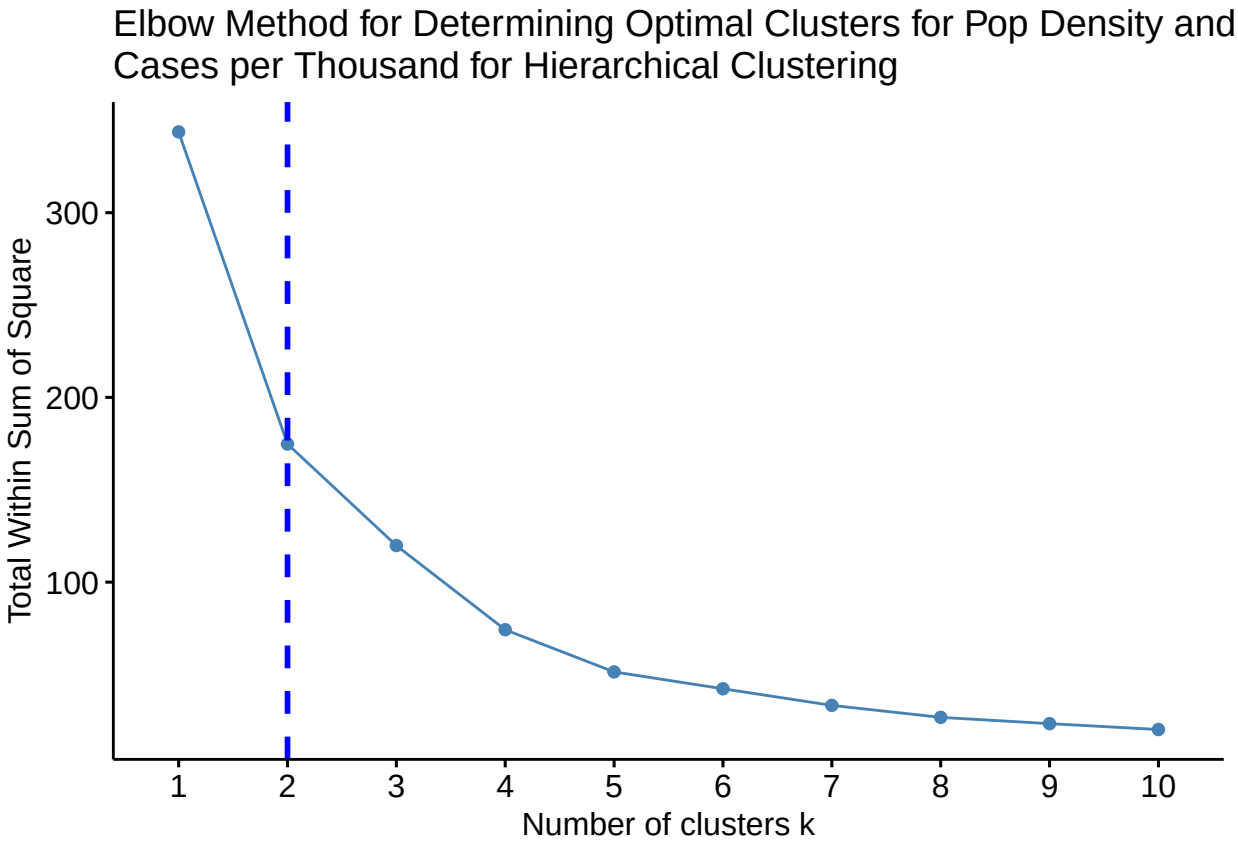
The second layer Hierarchical clustering seemingly divides the data into high and low population density.
Summary Statistics

Table 13: Summary Statistics for Hierarchical Cluster based on Population Density and Cases per Thousand

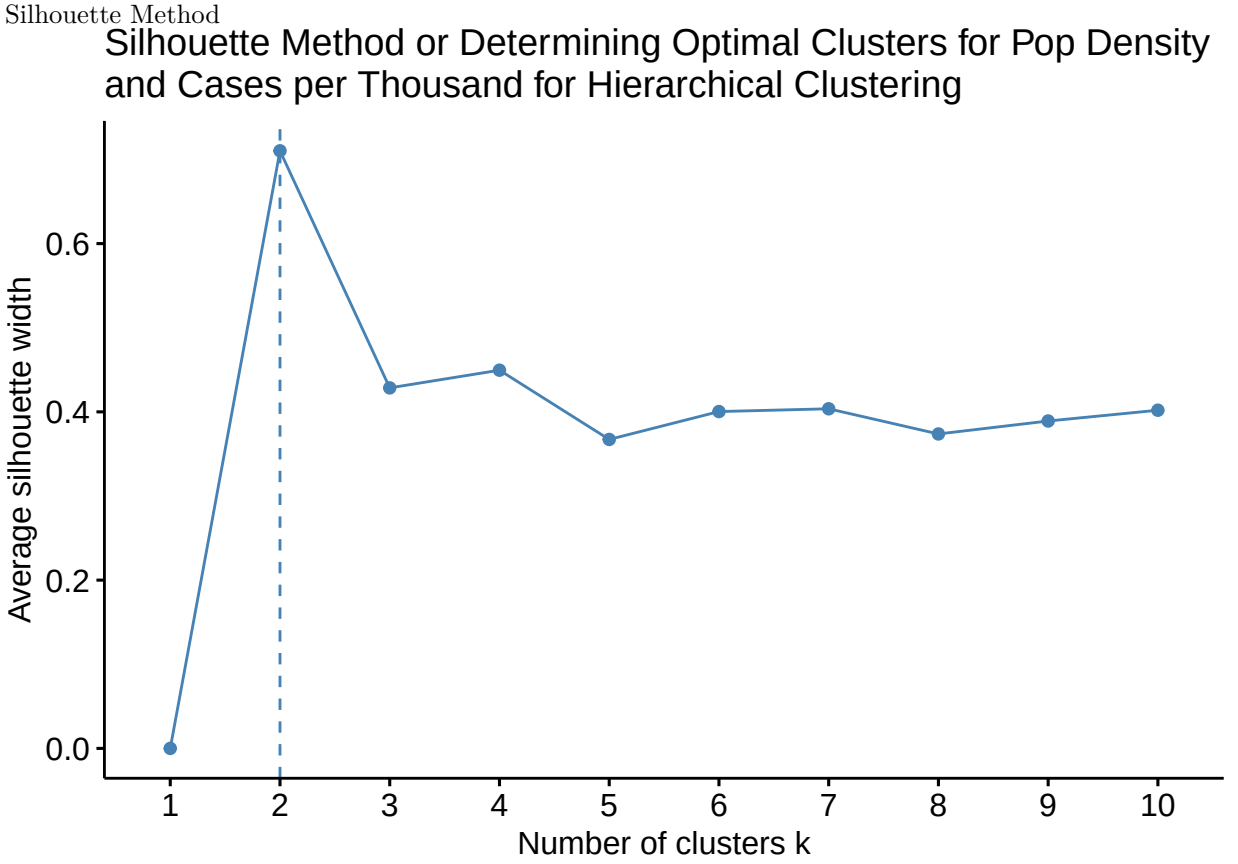
cluster	Avg Cases per thousand	Avg Deaths per thousand	Avg Population Density	Number of Counties
1	76.95841	1.426565	130.5207	136
2	85.79536	0.934339	2715.3660	3

- **Hierarchical Cluster 1:** This cluster consists of most of the counties in the data set. This cluster has lower cases per thousand, but notably a higher deaths per thousand rate. Cluster could be attributed as a "low population density" cluster.
- **Hierarchical Cluster 2:** This is a cluster of the three highest population density counties in the data set and cna be attributed as the "high population density" cluster.

Suitable Number of Clusters Elbow

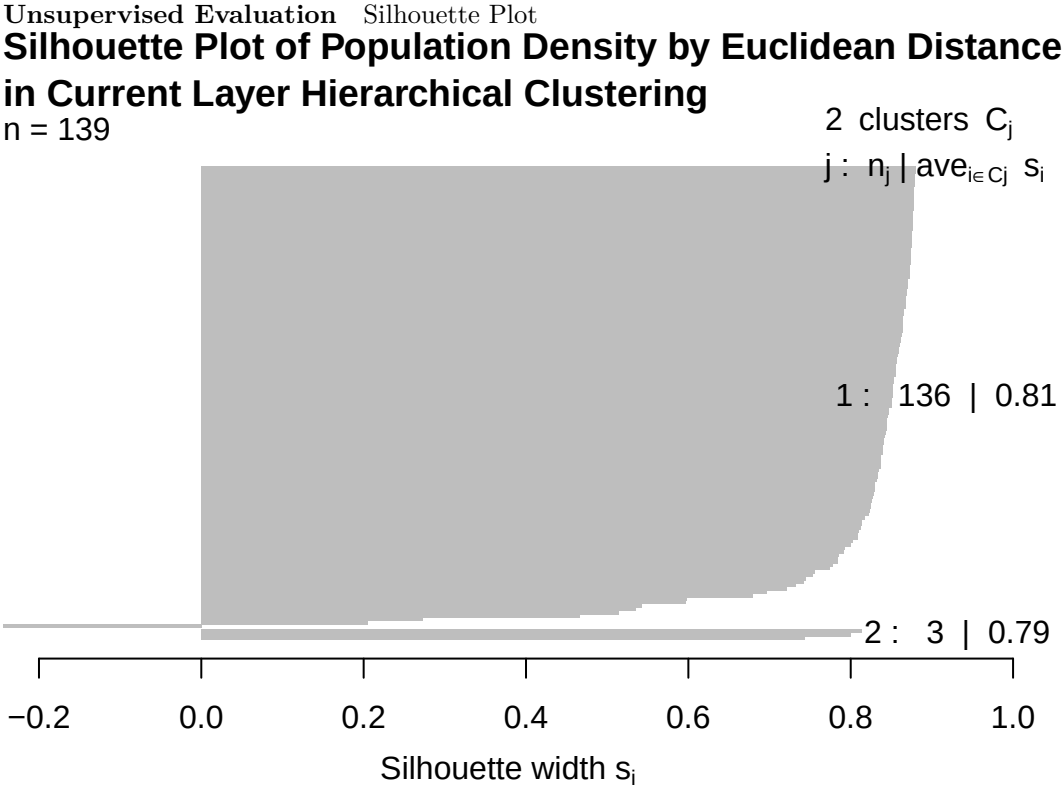


Silhouette Method



Our Elbow and Silhouette methods suggest our Hierarchical Dendogram be cut at 2 clusters.

Unsupervised Evaluation Silhouette Plot



Average silhouette width : 0.81

Our average silhouette widths are close to 1, which means the centroids are close to the center of the clusters; however, the distribution of data points in the cluster are very lop sided in favor of the low population density cluster.

- **Hierarchical Cluster 1:** With 136 points and an average Silhoutte plot of 0.81, this can be interpreted as a high degree of cohesion.
- **Hierarchical Cluster 2** This Silhoutte plot has little data points, but a high average Silhouttee with. This can be interpreted as a fair amount of cohesion.

Ground Truth Feature

Since the model is clustering for various factors such as affluence, COVID performance, and population density, the truth feature of our model will be defined by taking the total sum of a set of weights multiplied by the normalized value of numerical features. Assuming each aspect we are clustering for holds an equal amount of importance to our client, about a third each, the truth feature is defined using the following amount of weight with the following features.

- **Feature 1: Deaths per Thousand** 16.5% Weight
- **Feature 2: Cases per Thousand** 16.5% Weight
- **Feature 3: Median Income** 33.3% Weight
- **Feature 4: Population Density** 33.3% Weight

After normalizing using a z-score distribution, since a lower standard deviation indicates a better performance for both cases and deaths per thousand, the negative of COVID performance features will be subtracted instead of summed.
Mathematically, if our truth feature is defined as β, with each normalized numeric feature defined as f and weight number defined as w, the formula would be:

$\beta = -w_1 f_1 - w_2 f_2 + w_3 f_3 + w_4 f_4$

Supervised Evaluation

Mathematically, our original data set prior to the first layer of clustering contained all counties in Texas, defined as T. The first layer of clustering created four subsets of T, two K-means clusters defined as K_1 and K_2 and two hierarchical clusters defined as H_1 and H_2.

$K_1, K_2, H_1, H_2 \subset T$

The second clusters for both methods were found to be more affluent financially.

For the second layer of clustering the affluent counties, K_2 and H_2, were clustered again using the same methods with respect to another pair of attributes—cases per thousand and population density. With k_1 and k_2 defined as the second layer of K-means clusterings and with h_1 and h_2 defined as the second layer of hierarchical clustering.

$k_1, k_2 \subset K_2$

$h_1, h_2 \subset H_2$

The second layer of K-means clustering divided the counties according to low cases per thousand, k_1, and high cases per thousand, k_2. The second layer of Hierarchical clustering divided the data according to low population density, h_1, and high population density h_2.

Since a client likely wants a subset of counties with low cases per thousand, indicating a better performance during the COVID-19 pandemic, k_1 could be given as a subset of affluent, high COVID-19 performing group of counties.

Since a client likely wants a subset of counties with a higher population density, to make more money serving more people, h_2 could be given as a subset of affluent, high population density group of counties.

To test the "goodness" of our data, for each county in the subset of K_2 and H_2 counties, we will us the β formula on these subsets using aforementioned features, rank the counties from highest to lowest "goodness", divide the set in two at the same magnitude as their methodology subset equivalents, and perform a purity calculation for supervised analysis.

$a_1, a_2 \subset K_2$

$b_1, b_2 \subset H_2$

$|a_1| = |k_1|$

$|a_2| = |k_2|$

$|b_1| = |h_1|$

$|b_2| = |h_2|$

With N being the number of total counties in all clusters, M being the number of clusters being compared, and C and A being the two subsets being compared, and j being the number of subsets A has, the formula for purity is:

$purity = \frac{1}{N} \sum_{i=1}^{M} max_j |C_i \cap A_j|$

For the purity of subset k clusters:

$N = |K_2|, M = 2, C = k, A = a$

$purity \approx 0.653$

For the purity of subset h clusters:

$N = |H_2|, M = 2, C = h, A = b$

$purity \approx 0.978$

With a purity of 1 being a perfect match of clusters, subset k has a slight above average purity, while subset 2 has a near perfect purity. The high purity in subset h can be attributed to the magnitude of h_2 being 3 counties. If 136 of the 139 counties in H_2 are in subset h_1, of course there is going to be a high rate of purity. Therefore, a more interesting comparison for h_2 would be just comparing that cluster of three with the top three "goodness" results of b_2.

$comparison = \frac{1}{3}|h_2 \cap b_2|$

Subset h_2 consists of

-Dallas County

-Harris County

-Tarrant County

Subset b_2 conists of

-Dallas County

-Harris County

-Collin County

Two of the three counties found through our clustering were found also found in our truth feature, which is 66.7%. This is a more accurate metric than purity because of how the data was divided.

Exceptional Work

Data Collection, Quality, and Exploration

Objects to Cluster

Features for Clustering

Table of Features and Basic Statistics

Scale of Measurement

Measures for Similarity/Distance

Normalization/Standardization

Modeling and Evaluation

Clustering _____

Suitable Number of Clusters

Unsupervised Evaluation

Ground Truth Feature

Supervised Evaluation

Clustering _____

Suitable Number of Clusters

Unsupervised Evaluation

Ground Truth Feature

Supervised Evaluation

Recommendations

Discuss how the model can be interpreted and the recommendations based on the findings. Explain the utility for the stakeholders.
After analyzing our model's results, if a client has an interest in opening a business in an affluent, high land population density, and high COVID-19 performing county in Texas, they should consider the following counties.
After taking cluster "1" in the second layer K-means cluster, and sorting from descending order according to population density the three top counties are:

Table 14: Second Layer K-means Clustering Top 3 Counties

county_name
El Paso County
Wichita County
Potter County

The three results in cluster "2" for the second layer hierarchical clustering were:

Table 15: Second Layer K-means Clustering Top 3 Counties

county_name
Dallas County
Harris County
Tarrant County

Describe your results. What recommendations can you formulate based on the clustering results? How do these recommendations relate to the ones already presented in report 1? What findings are the most interesting to your stakeholder?

Conclusion

Summarize the key findings and their relevance to the initial questions.

List of References

[1] “Covid-19,” NFID, https://www.nfid.org/infectious-diseases/covid-19/ (accessed Oct. 8, 2024).
[2] Northwestern Medicine, “Covid-19 pandemic timeline,” Northwestern Medicine, https://www.nm.org/healthbeat/medical-advances/new-therapies-and-drug-trials/covid-19-pandemic-timeline (accessed Oct. 8, 2024).
[3] “10.1 - hierarchical clustering,” 10.1 - Hierarchical Clustering | STAT 555, https://online.stat.psu.edu/stat555/node/85/#:~:text=For%20most%20common%20hierarchical%20clustering,when%20they%20are%20perfectly%20correlated. (accessed Oct. 23, 2024).
[4] “Manhattan distance,” Wikipedia, https://simple.wikipedia.org/wiki/Manhattan_distance (accessed Oct. 23, 2024).
[5] A. Jain, “Normalization and standardization of Data,” Medium, https://medium.com/@abhishekjainindore24/normalization-and-standardization-of-data-408810a88307 (accessed Oct. 23, 2024).

Appendix

Include code snippets, extended tables, or other supplementary information.

Student Contributions

Olivia Hofmann

- Format/Organization of Report (Lead)
- Problem Description (Lead)
- Income Data in Texas Counties (Lead)
- Exceptional Work (Supporter)

Mike Perkins

- Format/Organization of Report (Supporter)
- Exceptional Work (Lead)

Matias Barcelo

- Format/Organization of Report (Supporter)
- Population Data in Texas Counties (Lead)

Extra Graduate Student Work

For each graduate students: Describe your exceptional work in a few sentences.
The graduate students in this group are Olivia Hofmann and Mike Perkins. Both graduate students worked together to ensure the report was held to a high standard and complete the exceptional work clustering.