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, Table 1, 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, Table 2, 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, Figure 1, 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.
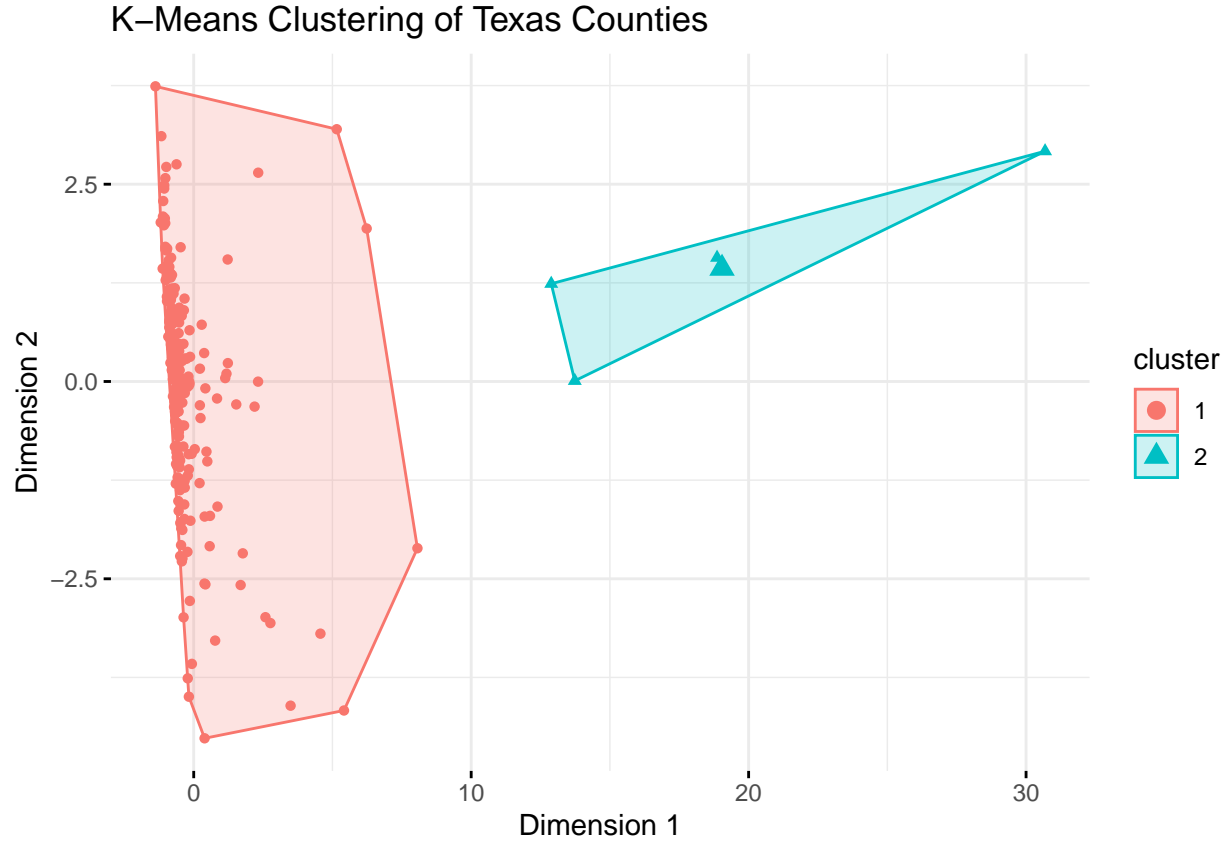
Figure 1: K-Means Clustering of Texas Counties

A summary statistics table, Table 3, 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, Figure 2, 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.
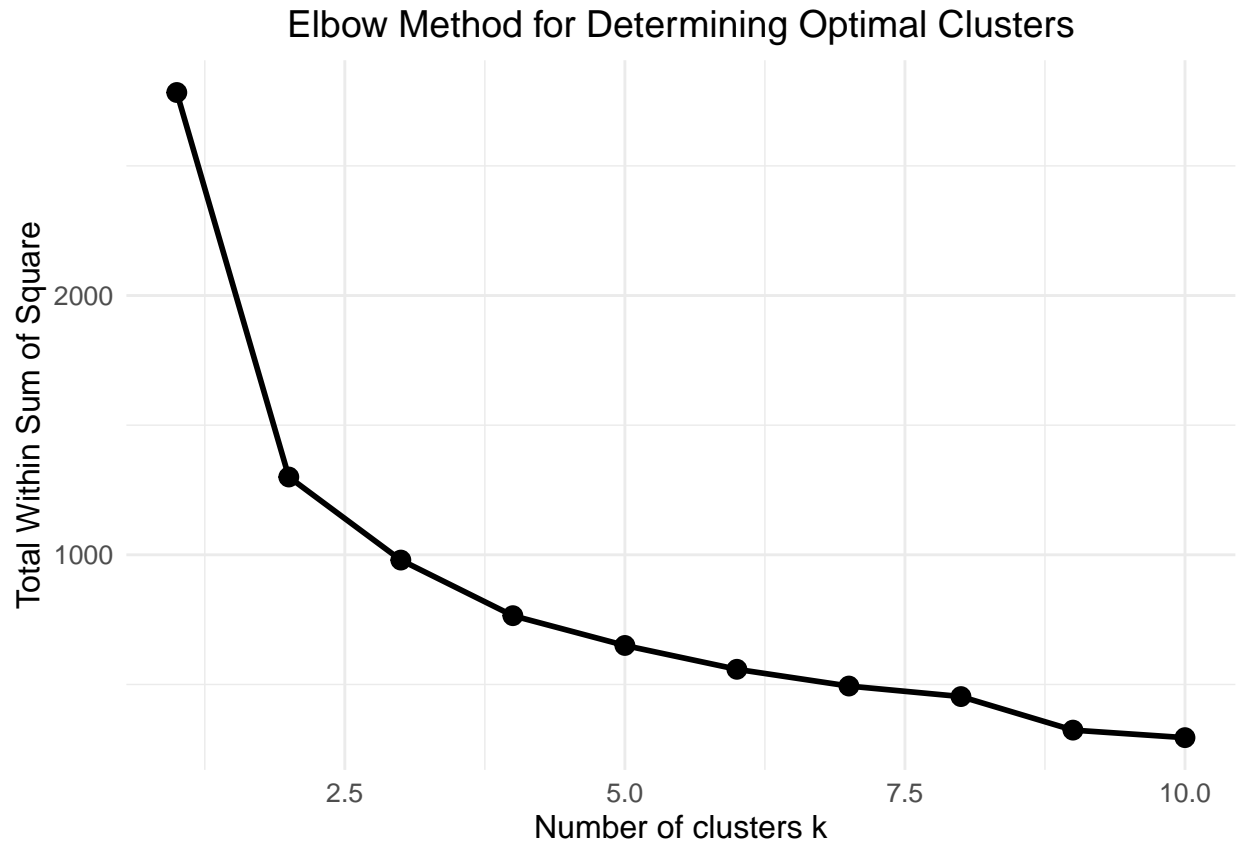

Figure 2: Elbow Method for Determining Optimal Clusters

The Silhouette Method, Figure 3, 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.
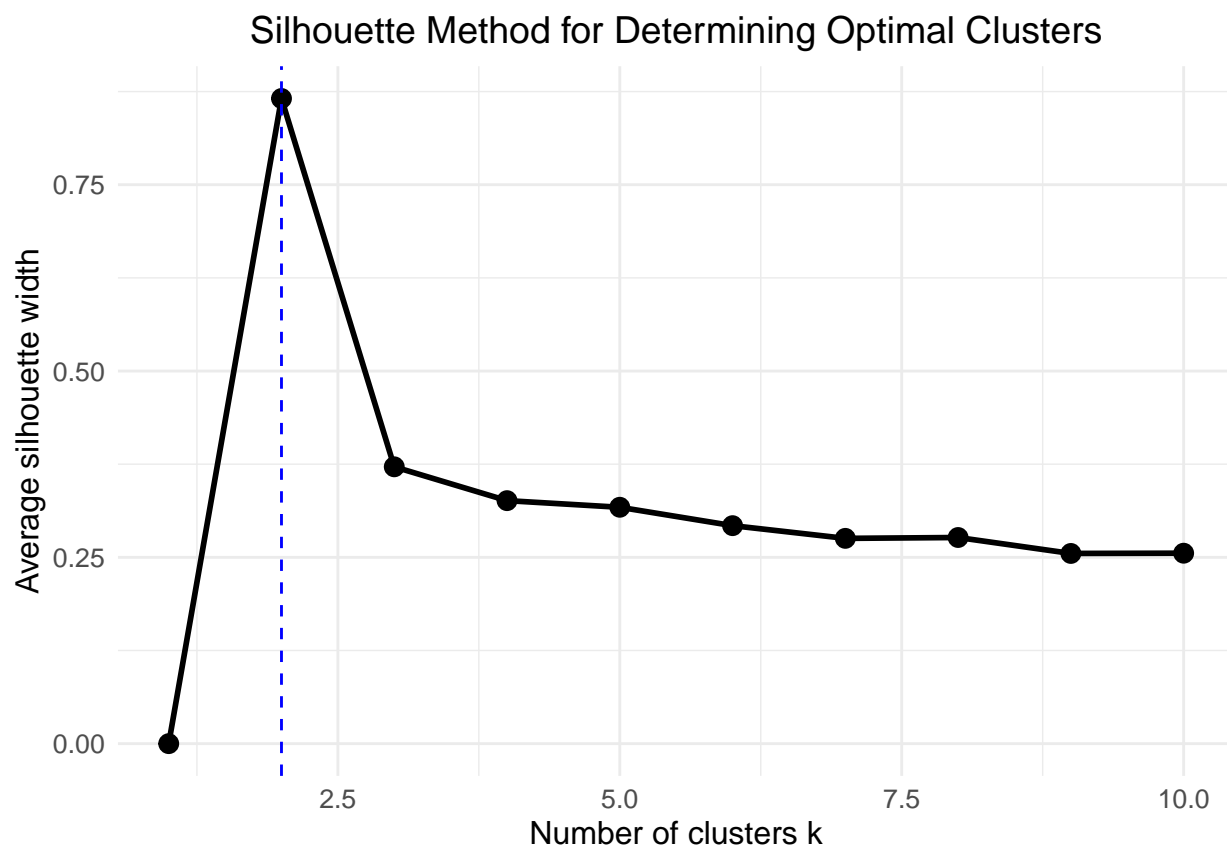
Figure 3: Silhouette Method for Determining Optimal Clusters

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, Figure 4, 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
```

Silhouette Plot for K–Means Clustering

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. This can be seen in Figure 5.

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

Figure 5: Ground Truth Cluster Comparison

Supervised Evaluation The K-Means clustering plot, Figure 6, 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, Table 4, 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

12

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 0.0301 which is significantly higher than Cluster 2's average ratio of 0.0166. 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. This can be seen in Table 5.

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, Figure 7, 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.
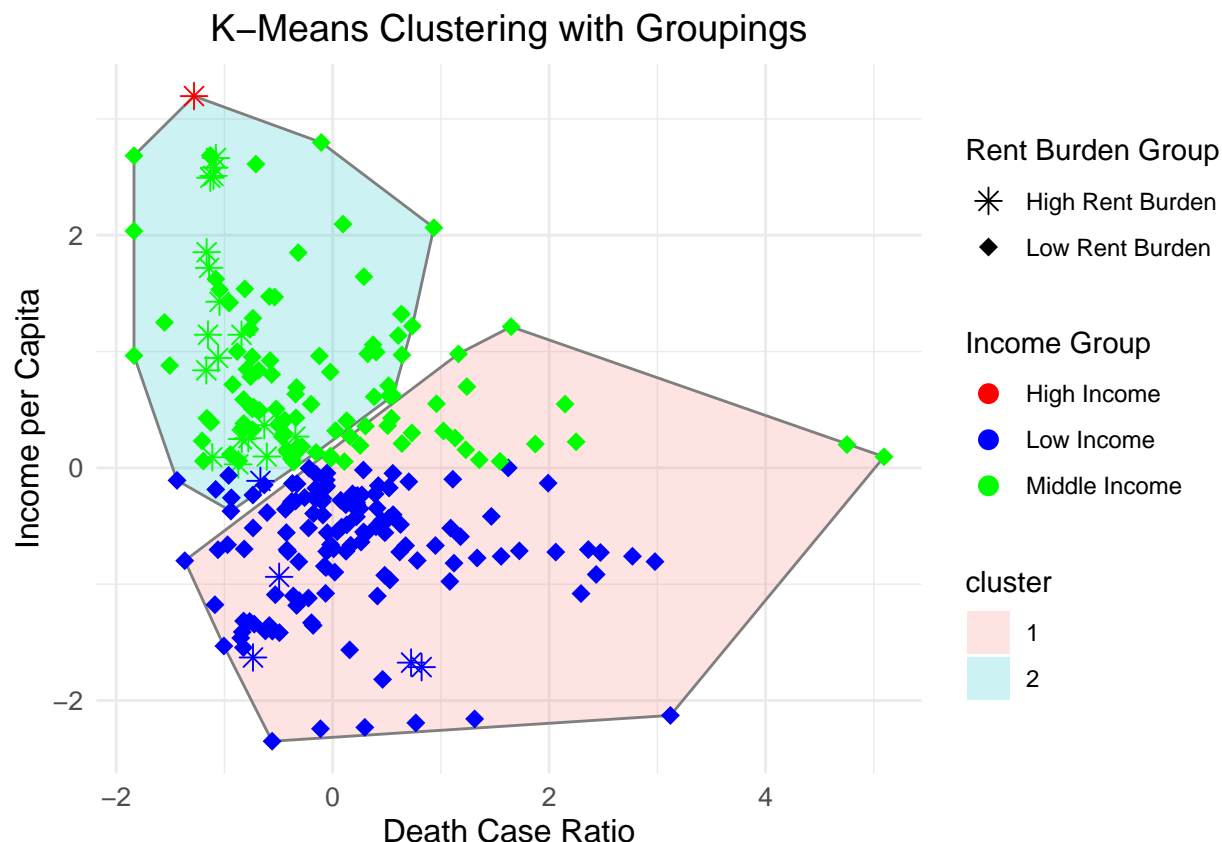

Figure 7: K-Means Clustering with Groupings

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

Table 6: Purity Scores by Grouping

Grouping	Purity_Score
Income Groups	0.8700787
Rent Burden Groups	0.9055118

infrastructure to manage health crises. Low-income areas, especially those burdened by high housing costs, 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, Table 6, 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

The Hierarchical clustering dendrogram, Figure 8, provides a visual representation of how Texas counties are grouped based on their economic and pandemic-related characteristics using complete linkage. In the dendrogram, each leaf represents a county, and the height at which two clusters merge reflects their dissimilarity. The two main clusters are delineated with colored rectangles. The hierarchical clustering analysis offers insights into the hierarchical relationships and distinctions among the counties, emphasizing which regions are more similar or different in terms of economic resilience and COVID-19 impact. This method is useful for exploring nested relationships and understanding the data's structure at various levels of similarity.
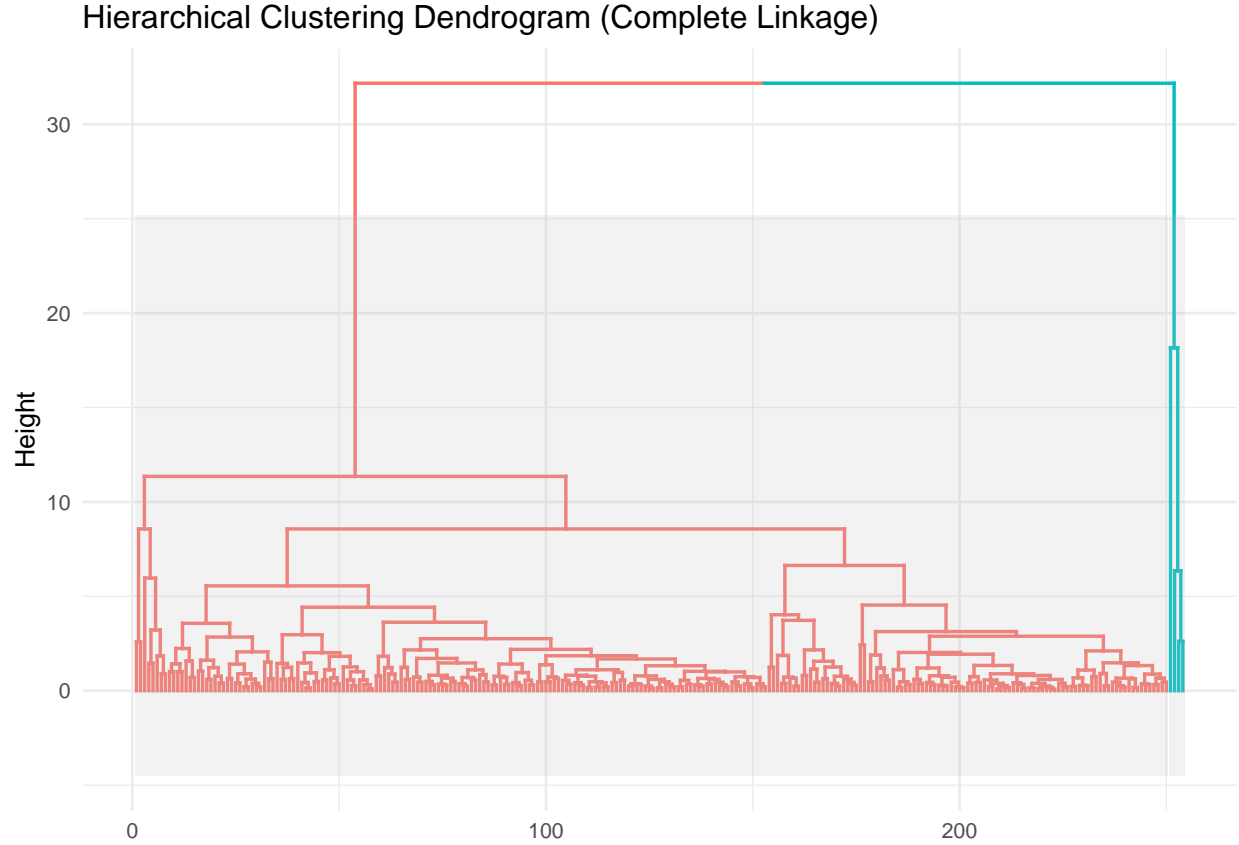
Figure 8: Hierarchical Clustering of Texas Counties

Each cluster represents a distinct group of Texas counties with similar economic, demographic, and pandemic characteristics. The table, Table 7, 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. The summary statistics table provides a comprehensive breakdown of the average values for key economic and pandemic-related features across the two clusters identified through Hierarchical Clustering. This clustering method categorizes Texas counties based on their similarities in median income, income per capita, rent burden, and COVID-19 impacts.

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

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 49,780.86 USD and Cluster 2 had an average median income of 56,987.00 USD. This shows a very moderate income difference of less than 8,000 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 Hierarchical clustering could perform better with supervision.

Suitable Number of Clusters The Elbow Method, Figure 9, 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.
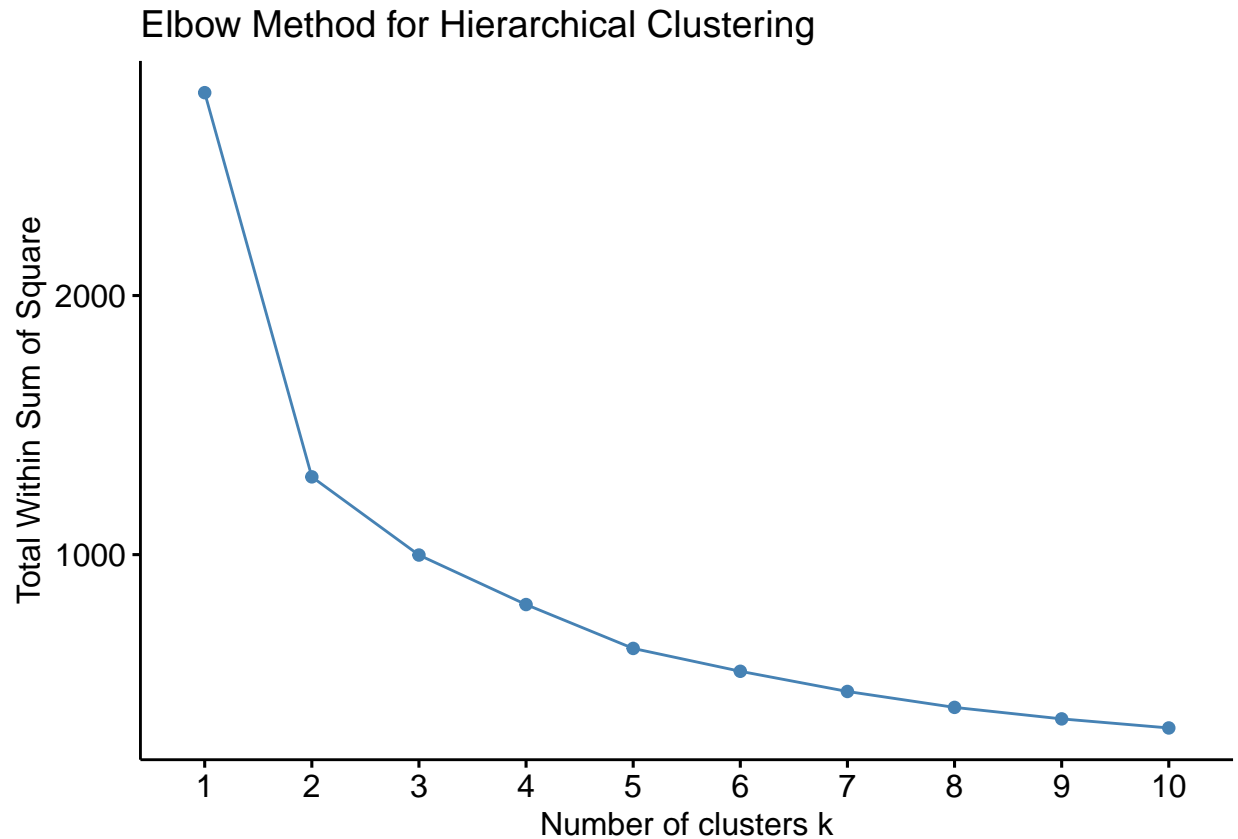

Figure 9: Elbow Method for Hierarchical Clustering

The Silhouette Method, Figure 10, 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.
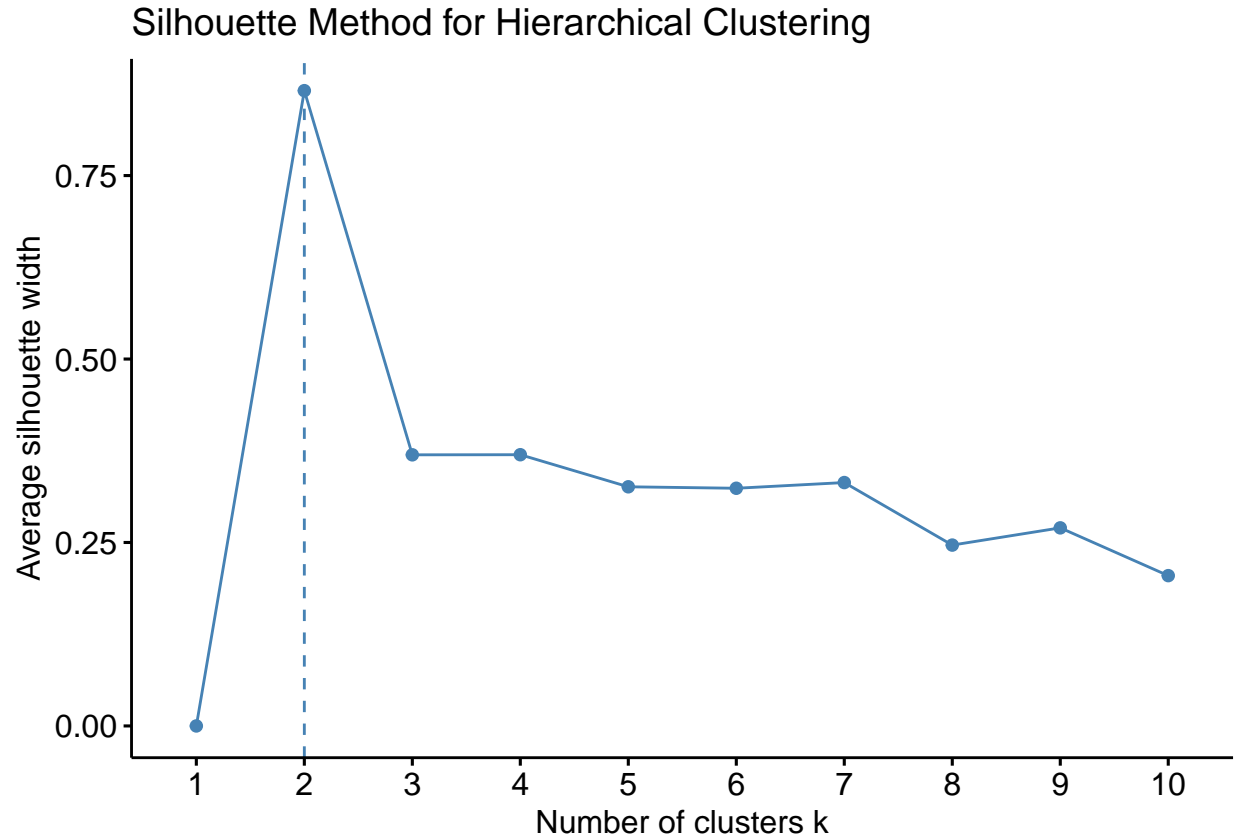
Figure 10: Silhouette Method for Hierarchical Clustering

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 A silhouette plot, Figure 11, is used as the unsupervised evaluation to assess the quality and cohesion of clusters generated by the Hierarchical clustering 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
```

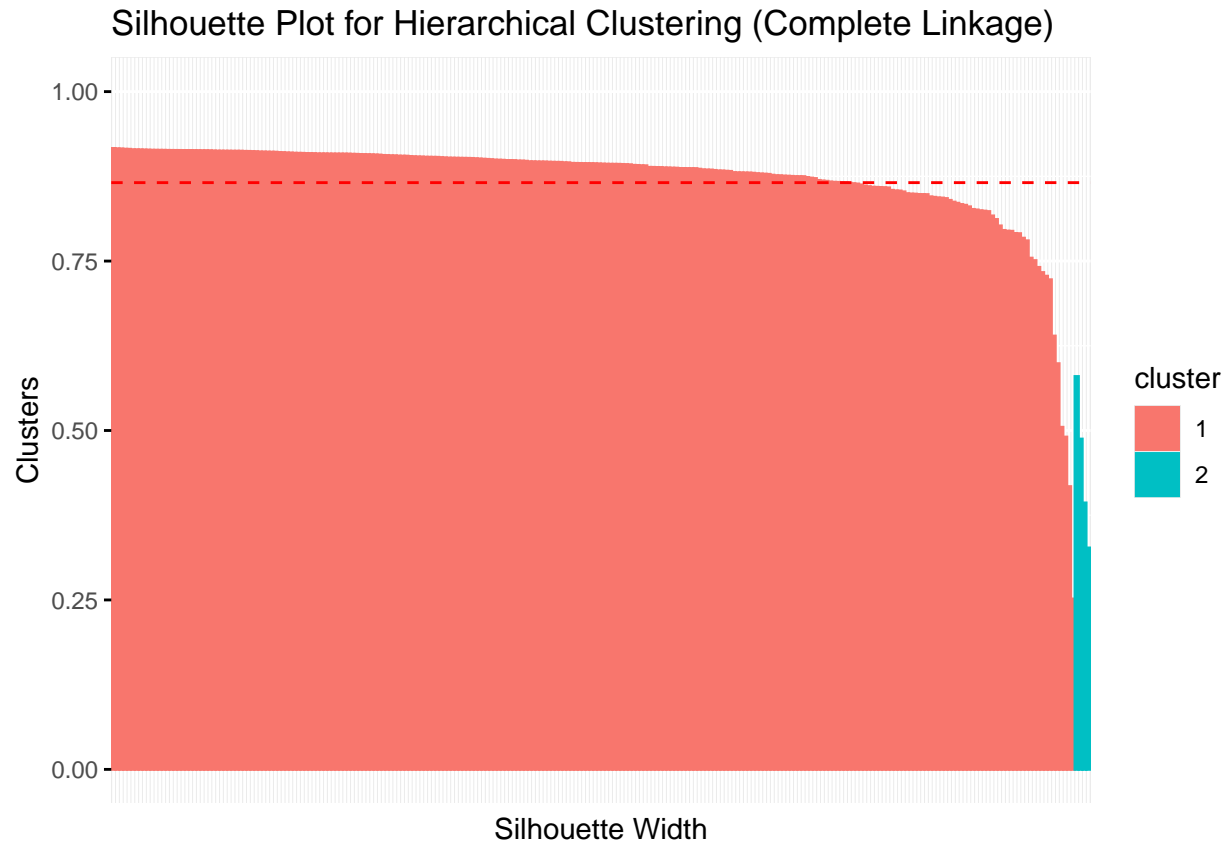# Silhouette Plot for Hierarchical Clustering (Complete Linkage)

Figure 11: Silhouette Plot for Hierarchical Clustering (Complete Linkage)

The following dendrograms, Figure 12, represent hierarchical clustering using two additional linkage methods: Average Linkage and Ward's Linkage. Each method groups Texas counties based on the similarities in economic and pandemic-related features.
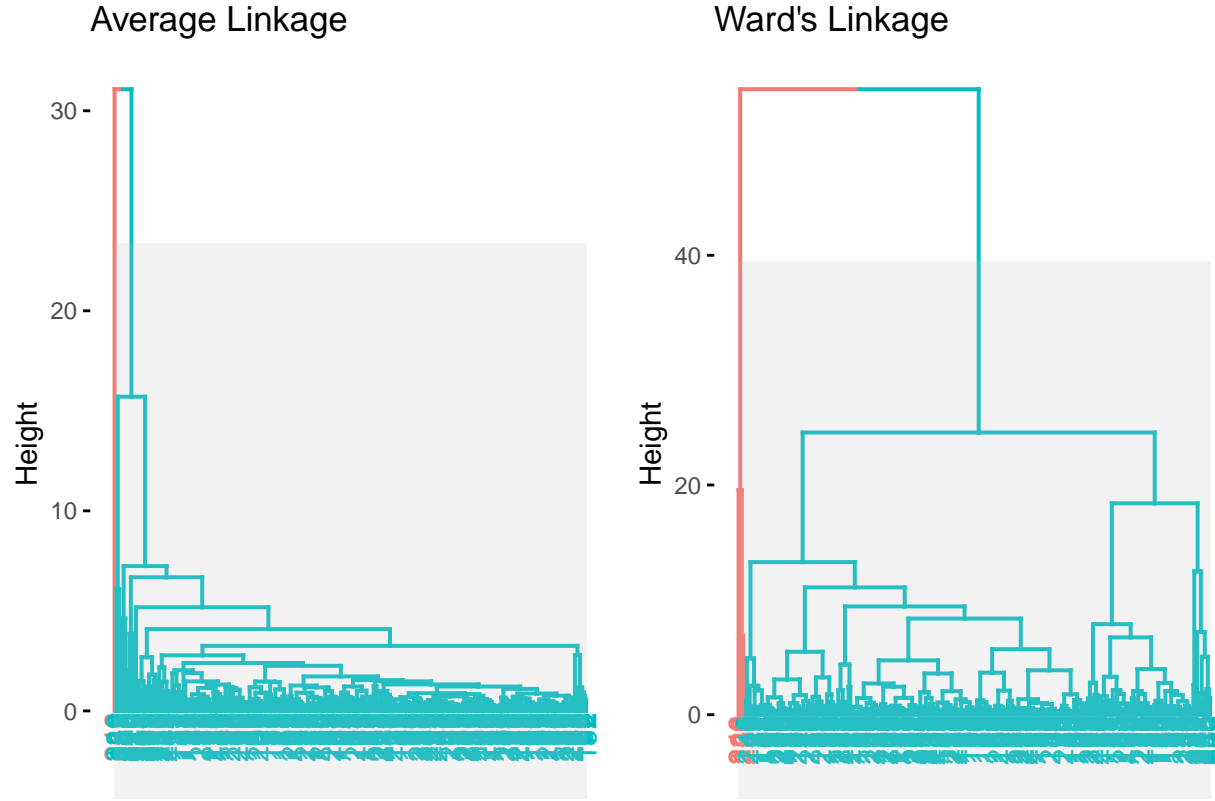
Figure 12: Additional Hierarchical Linkage Methods

Average Linkage The dendrogram shows that the data points merge gradually into larger clusters. The hierarchy has several smaller branches, indicating that clusters are formed by averaging distances between all points across groups. The clusters formed with this method seem less compact, as Average Linkage prioritizes minimizing the average distance between all points within clusters. This can lead to looser groupings and a higher risk of less cohesive clusters.

Wards Linkage Ward's Linkage tends to create more balanced and compact clusters by minimizing the variance within each cluster. The dendrogram shows clearer separation and more even cluster formation compared to Average Linkage. This method produces tighter clusters, ensuring that points within each cluster are more similar. Ward's method is often preferred for its ability to form more meaningful and interpretable clusters, especially when dealing with continuous data. The results suggest that Ward's Linkage might be more suitable for this analysis, providing clearer and more cohesive groups of counties.

The following table, Table 8, provides silhouette scores for three hierarchical clustering linkage methods: Complete, Average, and Ward's. The silhouette score measures how similar an object is to its own cluster compared to other clusters. Scores range from -1 to 1, with higher values indicating better-defined clusters.

Table 8: Average Silhouette Widths by Linkage Method

Linkage_Method	Avg_Silhouette_Width
Complete	0.8657033
Average	0.9017464
Ward's	0.8657033

- **Complete Linkage** (Silhouette Width: 0.8657) This high positive score indicates good clustering quality. The data points are well-matched to their assigned clusters, and there is a clear separation

between clusters. Complete linkage performs well, suggesting distinct groupings in the dataset.

- **Average Linkage** (Silhouette Width: 0.9017) This is the highest score among the methods, indicating excellent clustering quality. The clusters are well-defined, and the separation between them is even clearer than with complete linkage. Average linkage is the best method for this dataset, as it provides the most cohesive and distinct clusters.
- **Ward's Linkage** (Silhouette Width: 0.8657) This score is the same as that for complete linkage and also indicates good clustering quality. The data points are appropriately assigned to their clusters, with clear separation from other clusters.

All three methods perform well, but average linkage stands out as the most effective. The high silhouette scores across methods suggest that the dataset has well-separated, well-defined clusters, making hierarchical clustering a suitable approach.

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. This can be seen in Figure 13.

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

Figure 13: Ground Truth Cluster Comparison

Supervised Evaluation The supervised hierarchical clustering analysis, Figure 14, uses the Ward's linkage method to cluster Texas counties based on two key features: the death case ratio (COVID-19 deaths to confirmed cases) and income per capita. This method minimizes the variance within each cluster, resulting in more compact and well-separated groups. Although previously, the Average linkage was recommended, Ward's method provided the most balanced clustering which is why it was chosen for this analysis. Ward's method had a competative silhouette value, so it can be chosen as the linkage method with confidence.
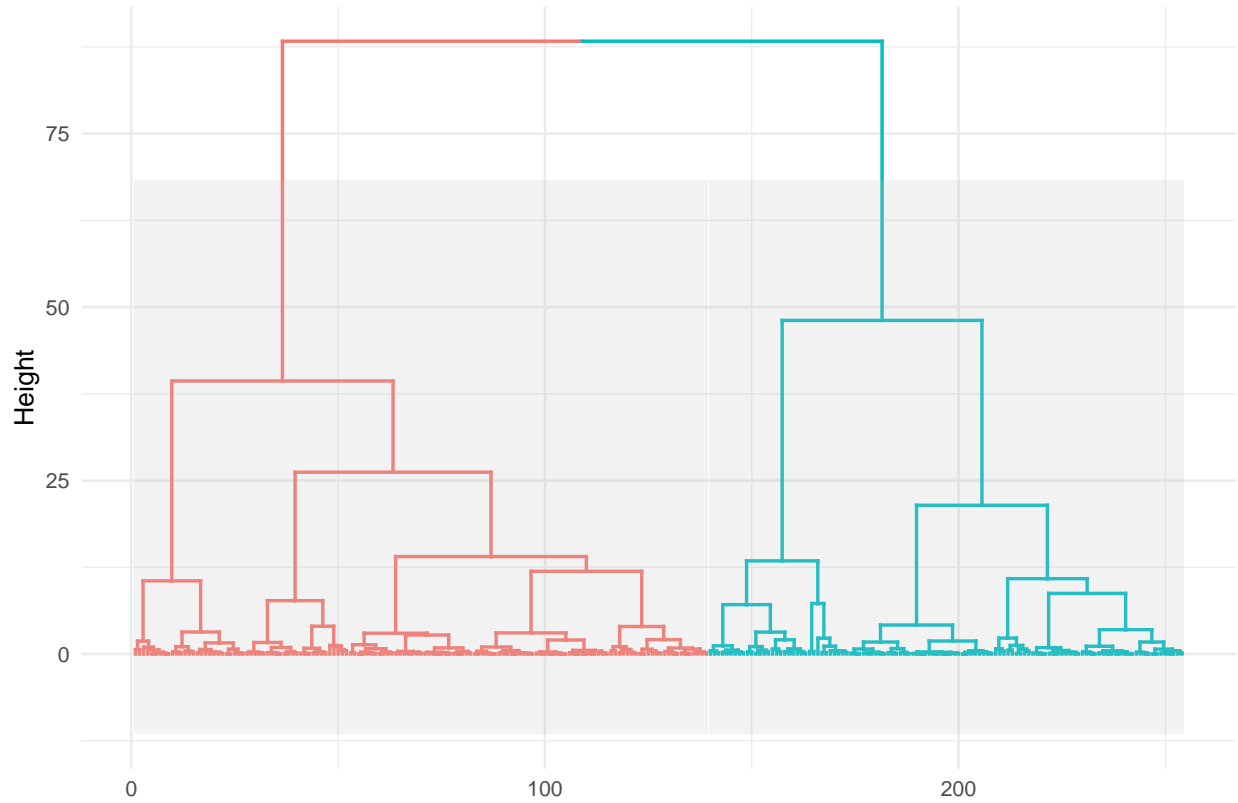
Figure 14: Hierarchical Clustering Dendrogram Supervised

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 dendrogram reveals a clear separation of counties into two clusters. The vertical lines represent the merging of clusters, and the height of these lines indicates the distance between merged clusters. The two clusters are visually distinct, suggesting that the clustering successfully captures variations in economic and pandemic characteristics among Texas counties.

A summary statistics table, Table 9, 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 Hierarchical 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 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

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 has an average median income of 42,665.18 USD and Cluster 2 has an average median income of 55,875.29 USD. This shows a very clear differentiation in economic status. The difference is around 13,000 USD.
- **Average Death Case Ratio:** Cluster 1 has an average ratio of 0.0327 which is significantly higher than Cluster 2's average ratio of 0.0180. This highlights the relationship between economic conditions and pandemic outcomes more effectively.

Cluster 1 Based on the economic indicators, the counties in this cluster are relatively less affluent. Additionally, the average rent burden indicators are significantly lower in this cluster, with few households experiencing high rent burdens. The average number of confirmed cases and deaths are also lower, and the death case ratio is higher at 0.0327. This implies that despite fewer cases, the impact of COVID-19 was relatively severe in terms of mortality.

Cluster 2 The economic indicators for this cluster are showing that the counties in Cluster 2 have greater economic resources and wealth. Counties in this cluster have higher average rent burdens, both in the category of rent greater than 50% of income and rent between 30-35% of income. The lower death case ratio could indicate better healthcare resources or more effective pandemic management.

There is a significant economic disparity between the two clusters, with Cluster 2 comprising wealthier counties. This disparity is also reflected in the rent burden indicators. Wealthier counties (Cluster 2) have a lower death case ratio, possibly indicating that higher income levels and greater resources contribute to better healthcare outcomes and pandemic management. Conversely, less affluent counties (Cluster 1) have a higher death case ratio, even though the absolute number of cases is lower. The hierarchical clustering provides an interpretable division of counties based on economic conditions and pandemic impact. The results highlight the role of economic status in influencing public health outcomes during a pandemic, with wealthier counties generally faring better in terms of mortality rates. This can be seen in Table 10.

Table 10: Summary Statistics by Subgroups Within Hierarchical Clusters

cluster_hc	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	591047.250
1	Low Income	Low Rent Burden	41840.12	20576.07	0.0303888	20872.763
1	Middle Income	Low Rent Burden	49366.14	26944.50	0.0510277	8791.857
2	High Income	High Rent Burden	90124.00	41609.00	0.0074628	914075.000
2	Low Income	High Rent Burden	46262.00	24273.00	0.0157121	245720.000
2	Low Income	Low Rent Burden	47437.43	23190.71	0.0179186	32775.629
2	Middle Income	High Rent Burden	62475.78	30879.67	0.0118649	956242.944
2	Middle Income	Low Rent Burden	57683.40	29041.24	0.0195622	40337.667

Cluster 1 Analysis

- **Low Income & High Rent Burden:** This subgroup comprises counties with lower income levels, with an average median income of 39,219.50 USD and an average income per capita of 17,058.50 USD. These counties experience significant rent burdens, with a total population of 591,047.25. The death case ratio is relatively high at 0.0258, potentially indicating vulnerability to health crises.
- **Low Income & Low Rent Burden:** These counties have slightly higher income levels, with an average median income of 41,840.12 USD and an average income per capita of 20,576.07 USD. Despite having a smaller total population of 20,872.76, economic challenges persist, reflected in a high death case ratio of 0.0304.
- **Middle Income & Low Rent Burden:** This subgroup shows improved economic conditions, with an average median income of 49,366.14 USD and an income per capita of 26,944.50 USD. However, the death case ratio is still high at 0.0510. The subgroup has a relatively small total population of 8,791.86, indicating that these unique characteristics are concentrated among fewer counties.

Cluster 2 Analysis

- **Low Income & High Rent Burden:** These counties have moderate income levels, with an average median income of 46,262.00 USD and an income per capita of 24,273.00 USD. The total population is 245,720.00, indicating that a significant number of residents live under economic stress. The death case ratio is 0.0157, highlighting ongoing health risks.
- **Low Income & Low Rent Burden:** This subgroup, with an average median income of 47,437.43 USD and an income per capita of 23,190.71 USD, benefits from lower rent burdens. Despite this advantage, the death case ratio is slightly higher at 0.0179, and the total population is 32,775.63, reflecting persistent health challenges.
- **Middle Income & High Rent Burden:** Counties in this subgroup have an average median income of 62,475.78 USD and an income per capita of 30,879.67 USD. Although they face significant rent burdens, they exhibit a relatively low death case ratio of 0.0119. The total population of 956,242.94 suggests some economic or health advantages due to greater resources.
- **Middle Income & Low Rent Burden:** With an average median income of 57,683.40 USD and an income per capita of 29,041.24 USD, this subgroup manages rent burdens more effectively. However, the death case ratio remains moderately high at 0.0196, and the total population is 40,337.67, indicating room for improvement in public health.
- **High Income & High Rent Burden:** Counties in this affluent subgroup have an average median income of 90,124.00 USD and an income per capita of 41,609.00 USD. They face substantial rent burdens but have a significantly lower death case ratio of 0.0075, suggesting better health infrastructure or resilience. The total population is substantial at 914,075.00, underscoring the socioeconomic advantages present in these counties.

Higher income groups generally experience lower death case ratios, highlighting the potential impact of economic resources on pandemic resilience. High rent burdens often accompany lower death case ratios in more affluent groups but exacerbate challenges in lower-income counties. The total population figures reflect how widespread each subgroup is, with some subgroups representing larger segments of the Texas population. High population density in wealthier counties could explain the more substantial infrastructure and health benefits.

This analysis underscores the complex interplay between income, rent burden, and public health outcomes. Economic factors appear to significantly influence the death case ratio, with wealthier counties generally faring better during crises like the COVID-19 pandemic.

The hierarchical clustering dendrogram with subgroup colors, Figure 15, provides a comprehensive visualization of how Texas counties are grouped based on two key characteristics: income levels and rent burden.

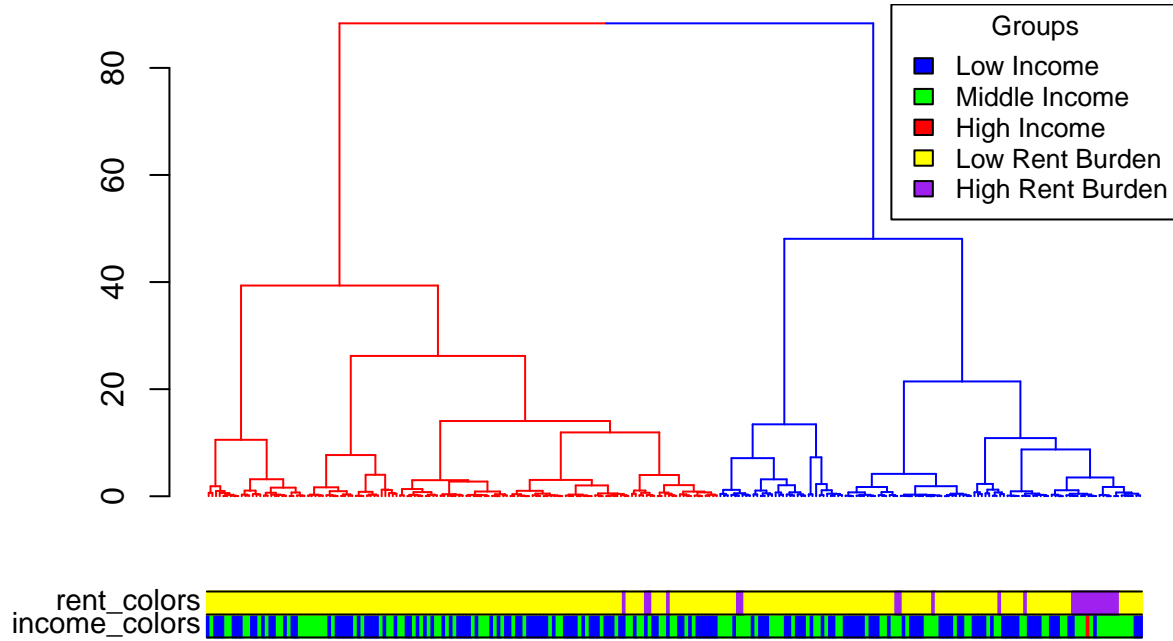# Hierarchical Clustering Dendrogram with Subgroup Colors

Figure 15: Hierarchical Clustering Dendrogram with Subgroup Colors

The dendrogram is constructed using Ward's method, which aims to minimize the variance within clusters. Two primary branches (or clusters) are shown in red and blue, indicating significant differences between these main groups of counties. The height at which branches merge indicates the degree of similarity between clusters. Higher merges suggest less similarity, while lower merges indicate closer relationships between the clusters.

Income Group The blue, green, and red colors at the bottom represent different income levels: (Blue: Low Income, Green: Middle Income, Red: High Income). From the colored bars, it is evident that the red branch cluster primarily contains counties with lower income levels (shown by the dominance of blue and green colors), while the blue branch cluster includes a higher concentration of middle and high-income counties (green and red).

Rent Burden The yellow and purple colors indicate rent burden groups: (Yellow: Low Rent Burden, Purple: High Rent Burden). The presence of purple bars in both clusters suggests that high rent burdens are distributed across counties regardless of income level. However, there seems to be a more considerable amount of yellow (low rent burden) in areas dominated by low and middle-income groups, indicating a possible relationship between lower/middle income and lower rent burden.

The dendrogram effectively highlights income and rent burden variations across Texas counties. The visual distinction between clusters provides a valuable tool for assessing economic vulnerabilities and can guide resource allocation to promote equity and resilience.

The purity scores, Table 11, for the hierarchical clustering analysis reveal insights into how well the clusters correspond to the true labels for 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

Table 11: Purity Scores by Grouping for Hierarchical Clustering

Grouping	Purity_Score
Income Groups	0.7992126
Rent Burden Groups	0.9055118

concerning the given grouping.

- **Income:** A purity score of 0.7992 for Income Groups indicates that approximately 79.92 percent of the counties within the clusters align with their actual income classifications (Low, Middle, or High Income). This suggests a reasonably strong relationship between the hierarchical clustering and the economic classifications. However, there is still some room for improvement in terms of clustering accuracy, which implies that a moderate level of misclassification is present.
- **Rent Burden:** The higher purity score of 0.9055 for Rent Burden Groups suggests that the clustering is more effective at distinguishing counties based on rent burden classifications (Low or High Rent Burden). This means that 90.55% of the counties are correctly grouped according to their rent burden status, indicating a strong correlation between the clusters and the true rent burden categories.

The hierarchical clustering method performs well overall, with a particularly strong association for Rent Burden Groups. The clear delineation between rent burden categories indicates that economic pressure from housing costs is a significant factor that the clustering model can effectively capture. The slightly lower purity score for Income Groups suggests that adding more features or refining the clustering approach could further improve accuracy.

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).

```
## [1] "county_area_combined"

##          county_name area_sqmiles
## 1  Anderson County     1062.617
## 2   Andrews County     1500.686
## 3  Angelina County      797.865
## 4   Aransas County      252.069
## 5     Archer County      903.289
## 6 Armstrong County      909.123

## [1] "county_name"  "area_sqmiles"

## [1] "Duplicate county entries in covid_plus_census:"

## # A tibble: 0 x 2
## # Groups:   county_name [0]
## # i 2 variables: county_name <chr>, total_pop <dbl>

## [1] "Final Texas counties COVID data with population, clusters, and scaled columns:"

## # A tibble: 6 x 14
##   county_name   total_cases total_deaths total_population total_confirmed_case~1
##   <chr>               <int>        <int>            <dbl>                  <dbl>
## 1 Anderson Cou~        5724           84            57747                 -0.100
## 2 Andrews Coun~        1630           37            17577                 -0.232
## 3 Angelina Cou~        6857          206            87700                 -0.0634
## 4 Aransas Coun~         935           29            24832                 -0.255
## 5 Archer County         733           10             8793                 -0.261
## 6 Armstrong Co~         135            6             1929                 -0.280
## # i abbreviated name: 1: total_confirmed_cases_scaled
## # i 9 more variables: total_deaths_scaled <dbl>, kmeans <fct>, hier <int>,
## #   area_sqmiles <dbl>, pop_density <dbl>, deaths_per_k <dbl>,
## #   cases_per_k <dbl>, pop_density_scaled <dbl[,1]>,
## #   cases_per_k_scaled <dbl[,1]>
```

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 ARE-ALAND_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 12: Basic Statistics for Features

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
total_deaths	0	14	32	135	84	4,024
total_cases	0	504	1,391	8,819	3,646	297,629
total_population	74	7,072	18,612	107,951	49,295	4,525,519
area_sqmiles	127.21	835.68	908.68	1,028.62	1,043.45	6,183.76
pop_density	0.11	6.44	21.56	114.01	63.44	2,922.83
deaths_per_k	0.00	1.28	1.80	1.97	2.51	6.28
cases_per_k	13.51	61.61	78.59	81.60	98.90	183.04

Table of Features and Basic Statistics

Scale of Measurement

Table 13: 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

```
## [1] "filtered_hierarchical_county_data_set columns"

##  [1] "county_name"                "total_cases"
##  [3] "total_deaths"               "total_population"
##  [5] "total_confirmed_cases_scaled" "total_deaths_scaled"
##  [7] "hier"                       "area_sqmiles"
##  [9] "pop_density"                "deaths_per_k"
## [11] "cases_per_k"                "pop_density_scaled"
## [13] "cases_per_k_scaled"
```

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] "Euclidean Distance Matrix for Population Density and Cases per Thousand (First Five Counties):"
```

	Andrews County	Aransas County	Archer County	Atascosa County	Austin County
Andrews County	0.000	102.802	9.580	28.326	53.441
Aransas County	102.802	0.000	99.854	77.012	54.895
Archer County	9.580	99.854	0.000	29.973	47.975

|	Andrews County	Aransas County	Archer County	Atascosa County	Austin County
Atascosa County	28.326	77.012	29.973	0.000	36.398
Austin County	53.441	54.895	47.975	36.398	0.000

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

K–Means Clustering of Texas Counties (Cluster '2' of Prior K–means Layer) Based on Population Density and Cases per Thousand
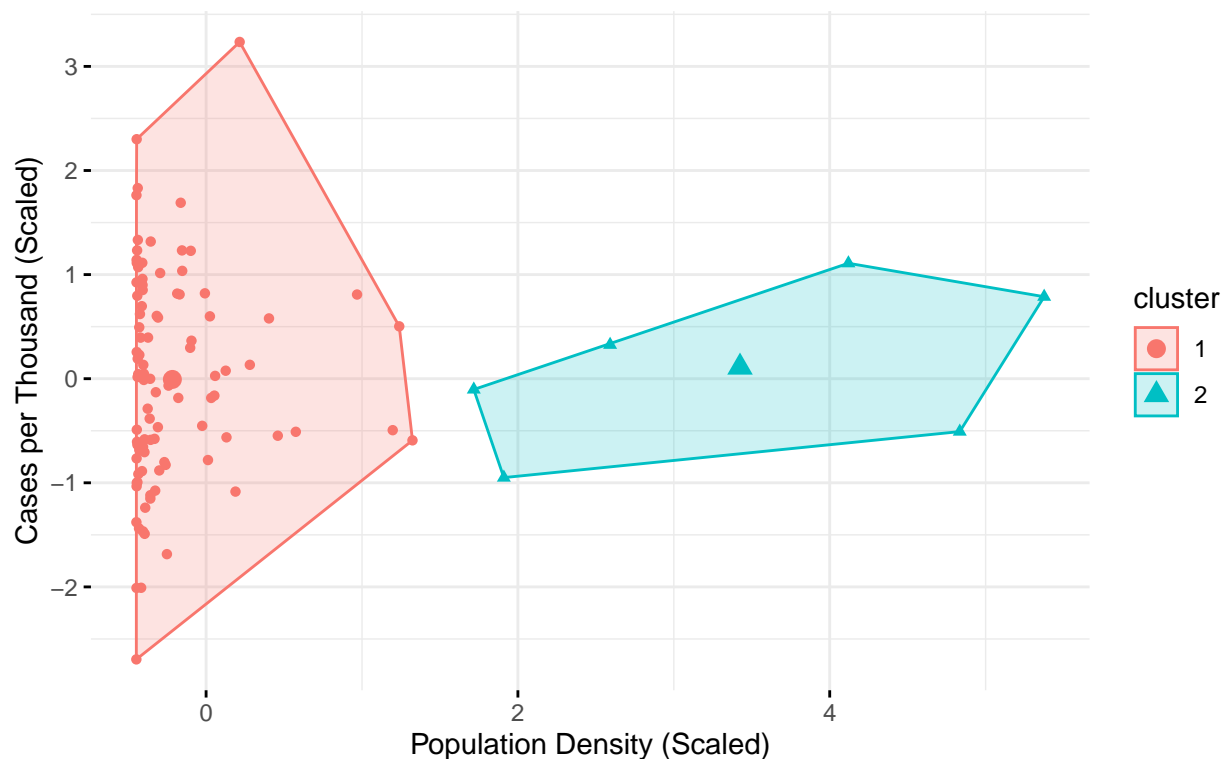
Our K-means clustering seemingly divides the data into low cases-per-thousand and high cases-per-thousand.

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

Unsupervised Evaluation   Silhouette Width
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.
Summary Statistics

As mentioned earlier, the second layer of K-means clustering seemingly prioritized cases-per-thousands over population density.

Ground Truth Feature

Supervised Evaluation

Heirarchical Clustering

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

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
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.
Summary Statistics
As mentioned earlier, the distribution of number of counties in a cluster could be better; however, the three counties found in the second county have a notably low deaths per thousand despite clustering for cases per thousand.

Ground Truth Feature

Supervised Evaluation

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.

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.