Final Report

Olivia Hofmann, Matias Barcelo, and Mike Perkins

2024-11-13

Contents

Problem Description (Business Understanding) **3**

Income Data in Texas Counties **4**
Data Collection, Quality, and Exploration . . . 4
Objects to Cluster . . . 4
Features for Clustering . . . 4
Table of Features and Basic Statistics . . . 4
Scale of Measurement . . . 5
Measures for Similarity/Distance . . . 5
Normalization/Standardization . . . 6
Modeling and Evaluation . . . 6
K-Means Clustering . . . 6
Suitable Number of Clusters . . . 7
Unsupervised Evaluation . . . 9
Ground Truth Feature . . . 10
Supervised Evaluation . . . 11
Heirarchical Clustering . . . 12
Suitable Number of Clusters . . . 12
Unsupervised Evaluation . . . 14
Ground Truth Feature . . . 16
Supervised Evaluation . . . 17

Population Data in Texas Counties **18**
Data Collection, Quality, and Exploration . . . 18
Objects to Cluster . . . 18
Features for Clustering . . . 18
Table of Features and Basic Statistics . . . 18
Scale of Measurement . . . 18
Measures for Similarity/Distance . . . 18
Normalization/Standardization . . . 18
Modeling and Evaluation . . . 18
K-Means Clustering . . . 18
Suitable Number of Clusters . . . 18
Unsupervised Evaluation . . . 18
Ground Truth Feature . . . 18
Supervised Evaluation . . . 18
Heirarchical Clustering . . . 18
Suitable Number of Clusters . . . 18
Unsupervised Evaluation . . . 18
Ground Truth Feature . . . 18

Supervised Evaluation . . . 18

Exceptional Work **19**
Data Collection, Quality, and Exploration . . . 19
Objects to Cluster . . . 19
Features for Clustering . . . 19
Table of Features and Basic Statistics . . . 19
Scale of Measurement . . . 19
Measures for Similarity/Distance . . . 19
Normalization/Standardization . . . 19
Modeling and Evaluation . . . 19
Clustering _____ . . . 19
Suitable Number of Clusters . . . 19
Unsupervised Evaluation . . . 19
Ground Truth Feature . . . 19
Supervised Evaluation . . . 19
Clustering _____ . . . 19
Suitable Number of Clusters . . . 19
Unsupervised Evaluation . . . 19
Ground Truth Feature . . . 19
Supervised Evaluation . . . 19

Recommendations **20**

Conclusion **21**

List of References **22**

Appendix **23**
Student Contributions . . . 23
Extra Graduate Student Work . . . 23

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.

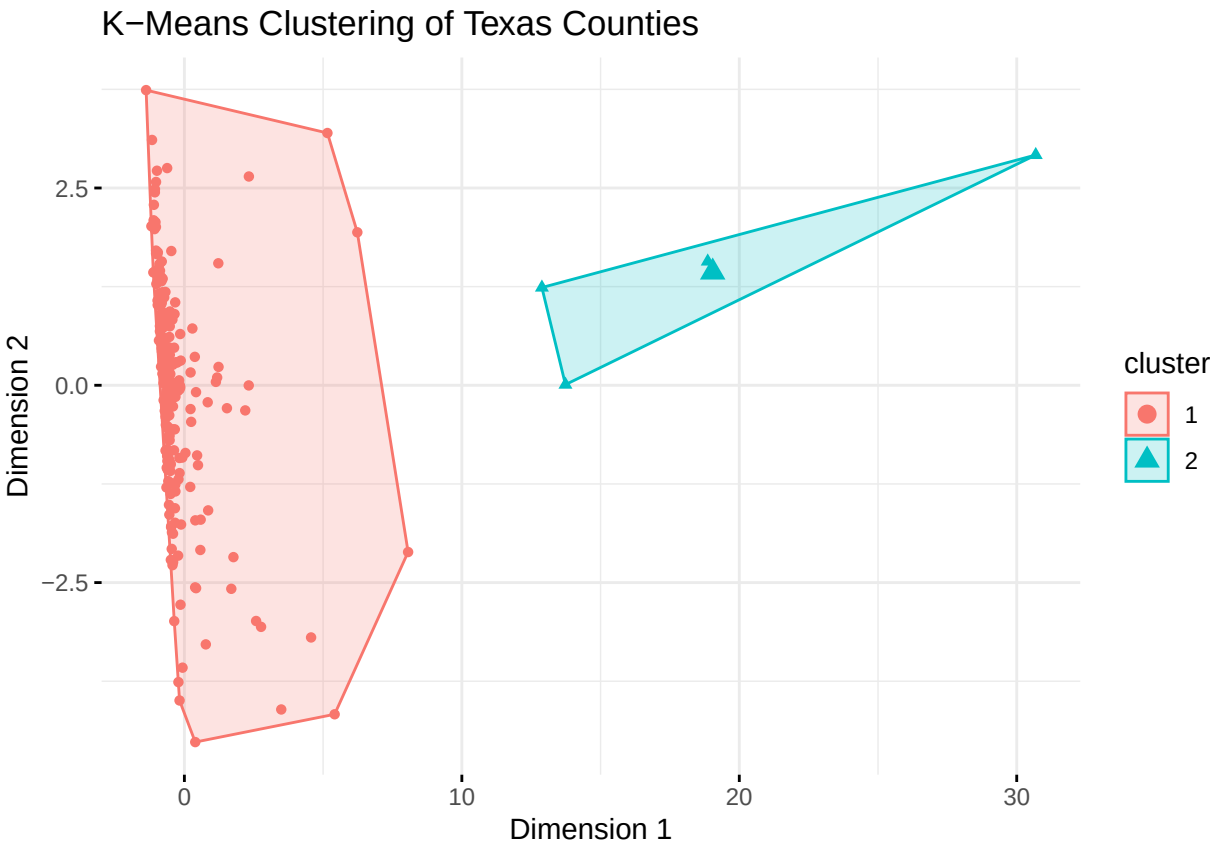


Figure 1: K-Means Clustering of Texas Counties

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.

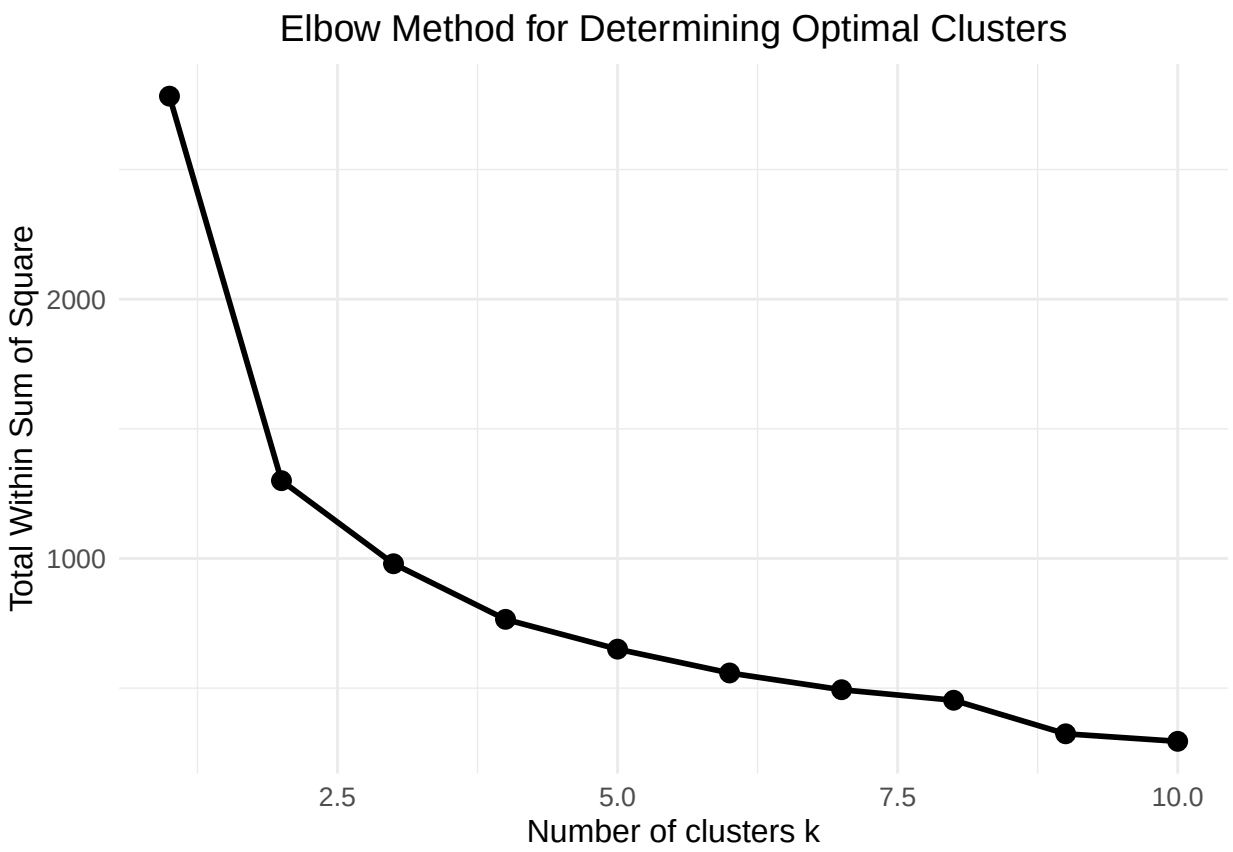


Figure 2: Elbow Method for Determining Optimal Clusters

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.

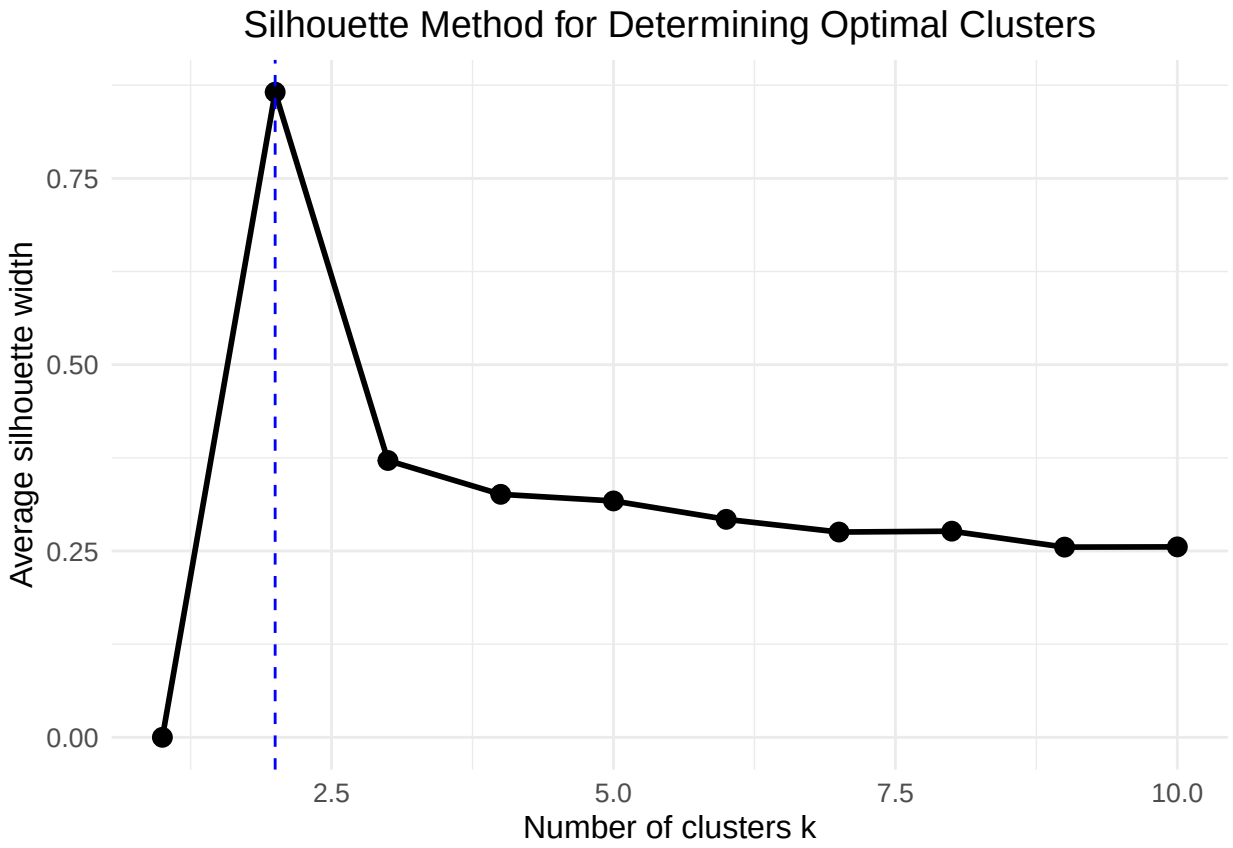


Figure 3: Silhouette Method for Determining Optimal Clusters

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.

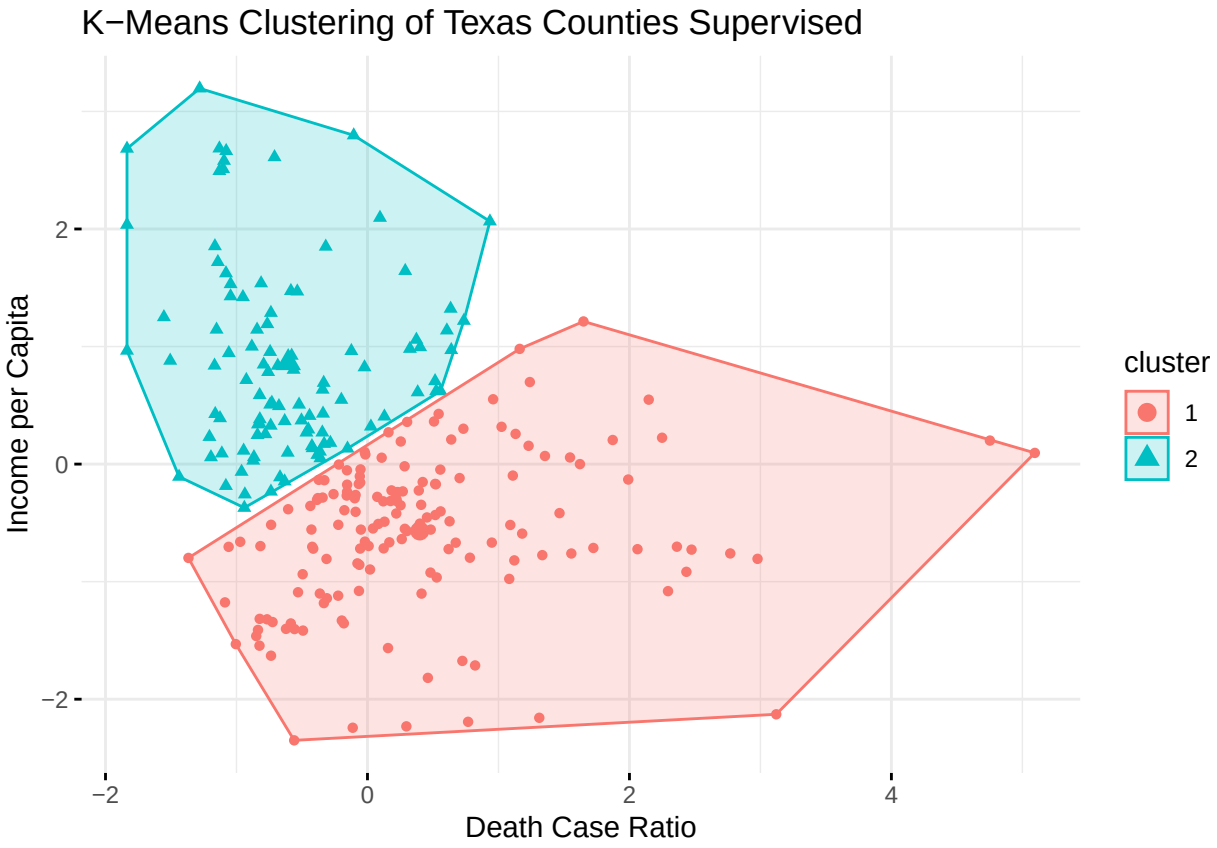


Figure 6: K-Means Clustering of Texas Counties Supervised

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.

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.

Table 4: Summary Statistics by Cluster

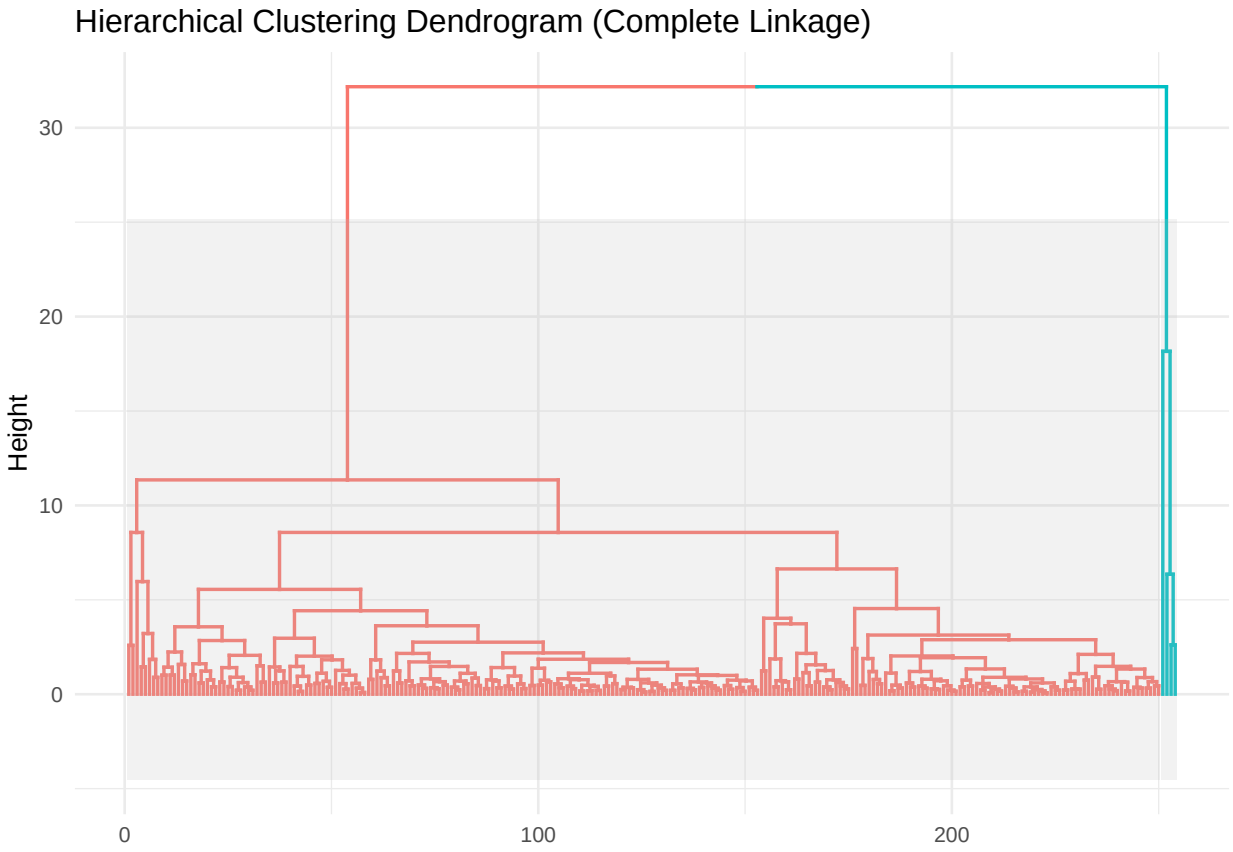
cluster	Avg Median Income	Avg Income per Capita	Avg Rent > 50%	Avg Rent 30-35%	Avg Confirmed Cases	Avg Deaths	Avg Death Case Ratio	Total Population
1	43937.72	21876.61	804.7255	297.6667	3309.477	80.98693	0.0300654	37969.71
2	58917.73	29376.93	6265.1683	2518.7921	16159.446	197.90099	0.0165567	213962.83

- **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 effective;y.

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.

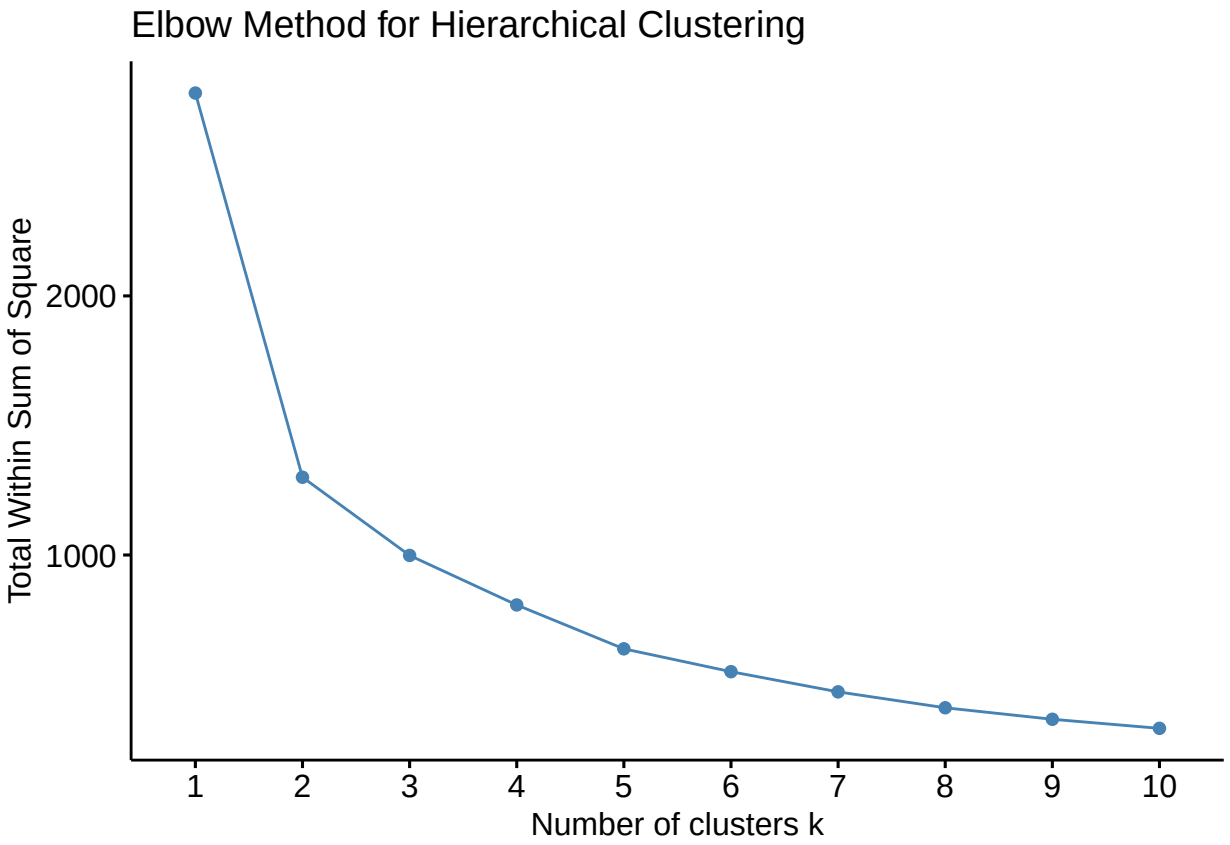
Heirarchical Clustering



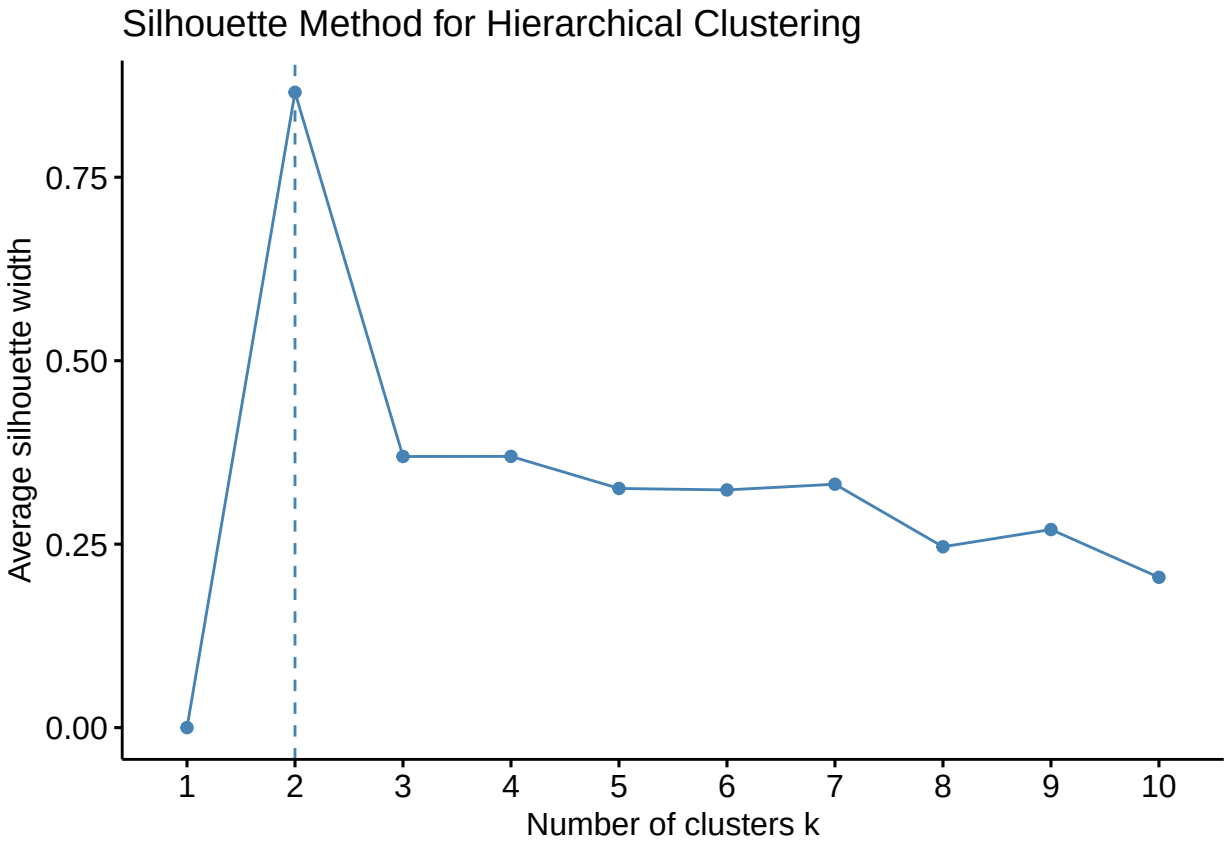
Suitable Number of Clusters The Elbow Method plots the WSS (Within-Cluster Sum of Squares) for different number of clusters. WSS measures how tightly the data points are grouped around the centroids of the clusters. After a certain point, adding more clusters provides diminishing returns, meaning the 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.

Table 5: 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



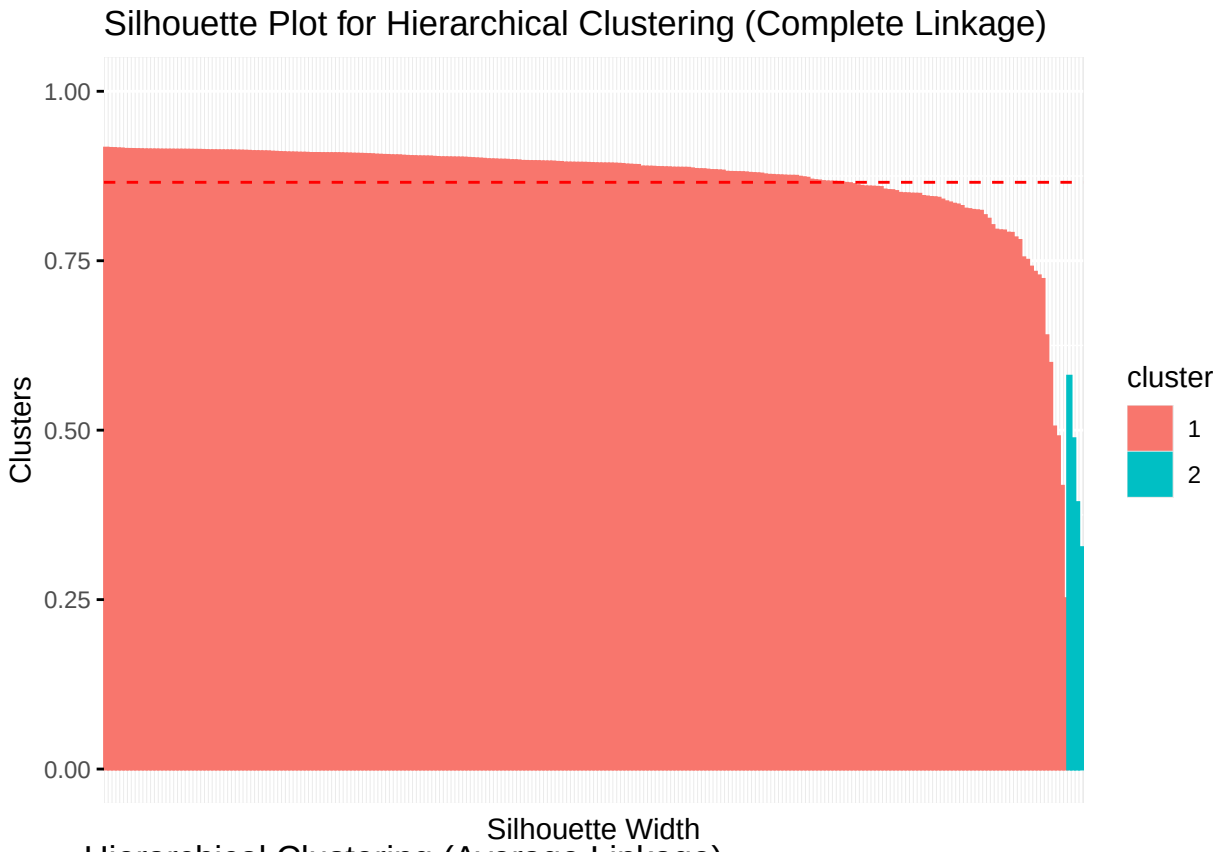
The Silhouette Method evaluates how well each data point fits within its assigned cluster compared to other clusters. The Silhouette score ranges from -1 to 1, with values close to 1 meaning that the points are well-clustered. In the following Silhouette chart, the peak occurs at 2 clusters.

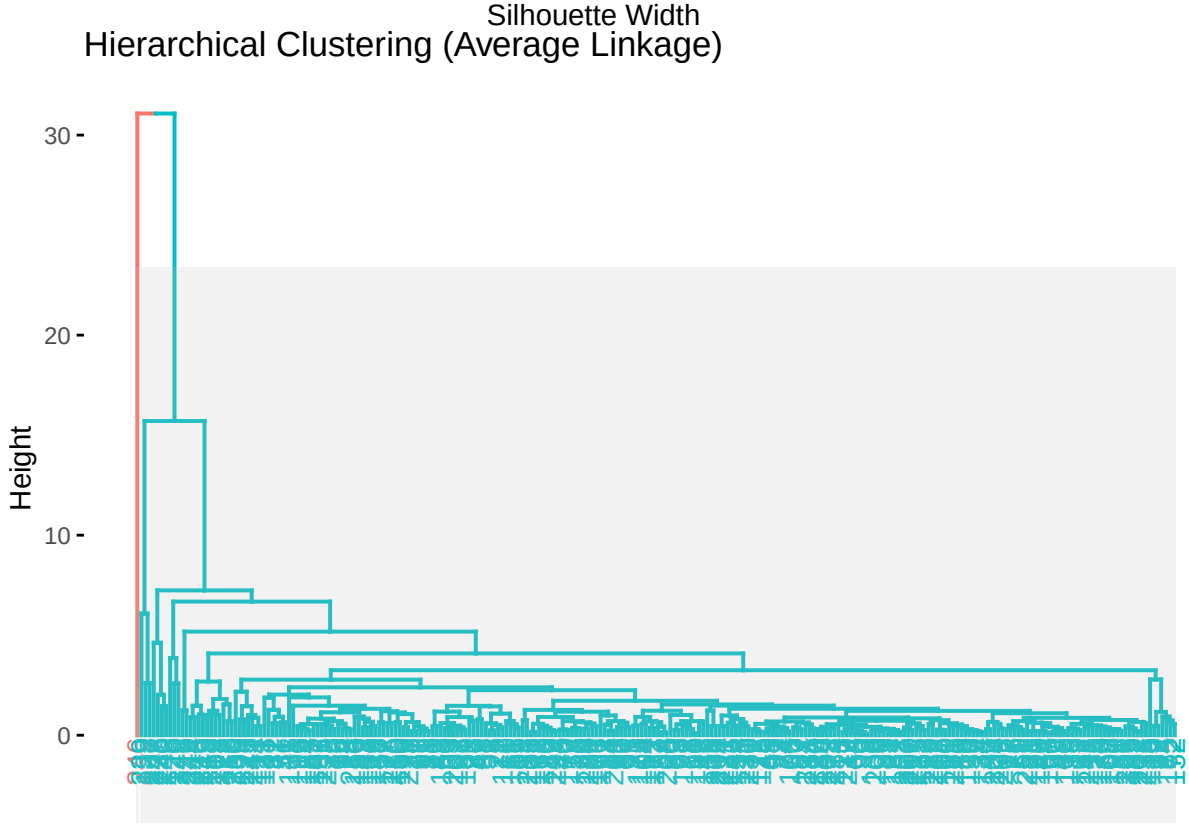


After considering both of these models, it was decided to do 2 clusters. Because of the consistency across both methods, 2 cluters was the clear choice.

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

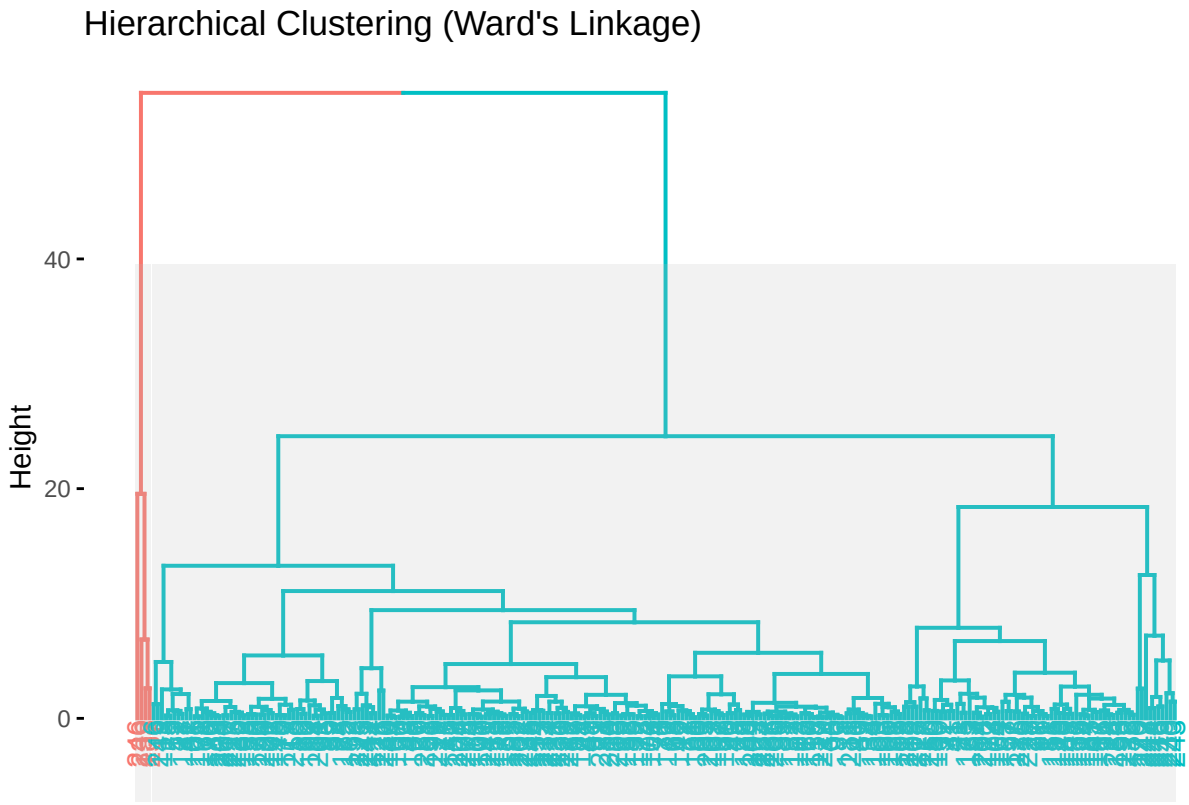
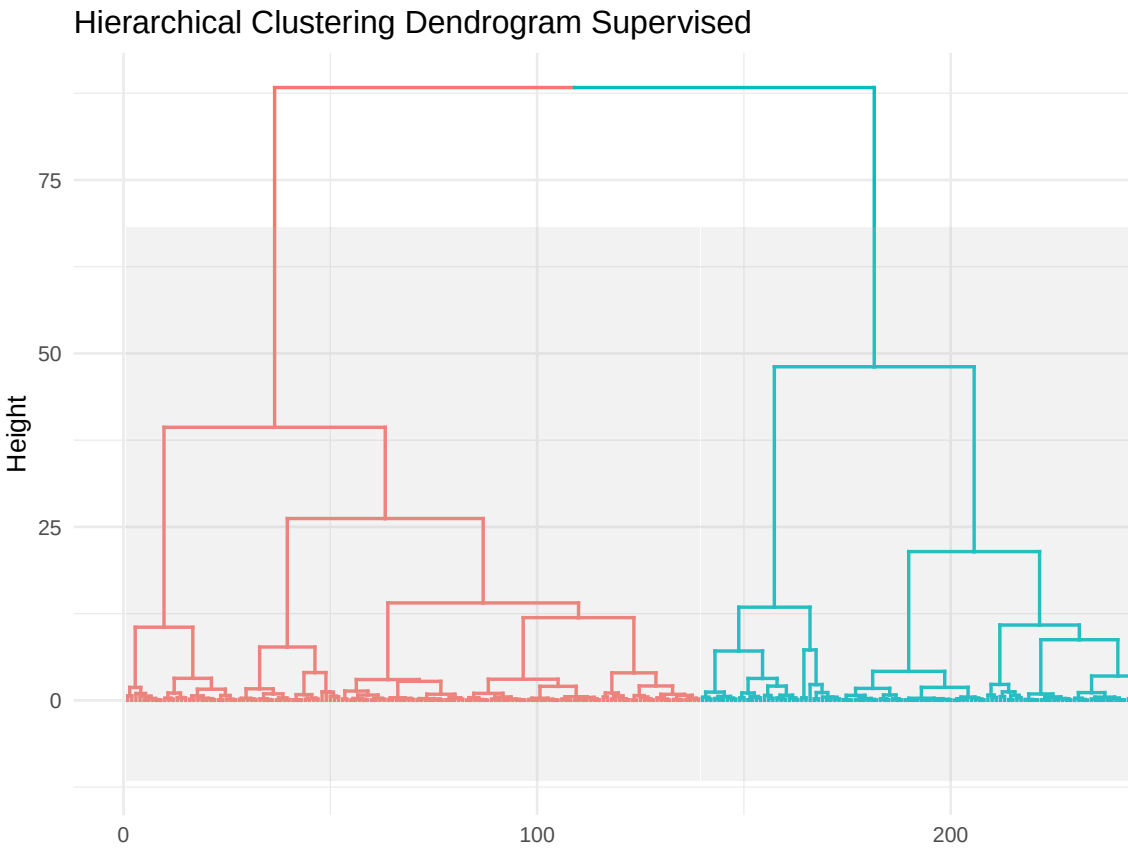


Table 6: Average Silhouette Widths by Linkage Method

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 7: 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

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

K-Means Clustering

Suitable Number of Clusters

Unsupervised Evaluation

Ground Truth Feature

Supervised Evaluation

Heirarchical Clustering

Suitable Number of Clusters

Unsupervised Evaluation

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.