Final Report

Olivia Hofmann and Michael Perkins

2024-11-13

Contents

Problem Statement

The stakeholder, a **property developer**, seeks to identify the best location in Texas for developing a mixed-use building that includes amenities like a gym, restaurants, and a pharmacy. The key concern is selecting a county that demonstrates stability and resilience in response to unpredictable events like the COVID-19 pandemic. This analysis aims to answer the following questions:

Key questions to address include:

- What counties in Texas demonstrated resilience during the COVID-19 pandemic?
- What are the characteristics of these counties in terms of population makeup and economic stability?
- Can we group counties to identify patterns that indicate stability and resilience?

By understanding how different counties fared during the pandemic, the developer can make an informed decision to ensure the chosen location offers long-term stability and resilience against unforeseen circumstances.

Income Data in Texas Counties

Data Collection, Quality, and Exploration

Objects to Cluster

The objects to be clustered are the counties in Texas. We aim to identify counties that exhibited economic resilience during the COVID-19 pandemic by analyzing income and rent burden metrics alongside general population data.

Feature Selection

We selected features that focus on **income, wealth, rent burden, and population**—key factors for assessing economic resilience. These features reflect the economic foundation and stability of each county, critical for understanding responses to crises like COVID-19.

The features used, their scales, and measurement types are as follows:

- **Income Levels**: The distribution of households across various income levels provides insight into a county's overall economic health and resilience.

- **Rent Burden**: High rent burden percentages indicate financial strain on households, affecting their ability to manage crises effectively.

- **Median Income** and **Income per Capita**: These metrics serve as 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 relative to county size.

Each feature is scaled to ensure equal contribution to the clustering process. We use **standardized Euclidean distance** to calculate similarity, ensuring that larger-scale features (e.g., population) do not disproportionately affect cluster formation.

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

Basic Statistics of Features

Table 1: Basic Statistics of Key Features for Clustering Texas Counties

Feature	Mean	SD	Min	Max
Median Income (USD)	49,894.34	12,132.68	24,794	93,645
Income per Capita (USD)	24,859.02	5,240.75	12,543	41,609
Rent > 50% Income	2,976	13,179.06	0	158,668
Rent 30-35% Income	1,180.87	5,203.84	0	61,305
Income < 10,000 USD	2,469.77	8,601.26	0	98,715
Income 50,000-59,999 USD	2,945.2	10,790.45	3	122,390
Income 100,000-124,999 USD	3,205.16	11,657.05	0	131,467
Total Population	107,951.2	389,476.9	74	4,525,519

The table provides an overview of the key features used for clustering Texas counties. Each feature's **mean**, **standard deviation (SD)**, **minimum (Min)**, and **maximum (Max)** values are shown to highlight the range and distribution within the dataset. These statistics offer insight into the variations across counties, helping us understand the economic conditions and resilience factors that may correlate with each county's response to the COVID-19 pandemic.

For instance, features like **median income** and **income per capita** provide a sense of wealth distribution and overall economic stability within each county. Higher values may indicate counties with stronger financial resources, which could contribute to greater resilience. Meanwhile, metrics like **rent burden** (e.g., rent constituting over 50% or 30-35% of household income) reflect financial strain on residents, which might impact a county's ability to handle crises effectively. Lastly, **total population** helps normalize COVID-19 case and death rates, allowing for more accurate comparisons across counties of different sizes.

Analyzing these basic statistics will aid in identifying clusters of counties that share similar economic characteristics, offering the stakeholder valuable insights into which regions may be more suitable for long-term investments.

Scale of Measurement and Similarity Measures

All of the selected features are measured on a **ratio scale**, as they have a true zero point (e.g., zero income, zero population) and allow for meaningful arithmetic operations, such as calculating differences and ratios. The table below defines the scale and provides a brief description of each feature.

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, it is crucial to select appropriate measures of similarity or distance based on the nature of the data. The following measures are particularly relevant:

- **Euclidean Distance**: The most commonly used distance measure, calculated as the straight-line distance between points in a multi-dimensional space. This measure is especially effective for continuous numerical data, such as income or population figures, where relationships between data points can be interpreted geometrically. Euclidean distance provides an intuitive and straightforward approach for visualizing proximity between clusters.
- **Manhattan Distance**: This measure calculates the distance between two points by summing the absolute differences of their coordinates. Manhattan distance is beneficial when dealing with outliers or when the scale of measurement varies among features. It reflects a grid-like path, which can be advantageous when a more robust metric against extreme values is required, as it reduces the impact of outliers.
- **Standardization/Normalization**: When features exhibit wide ranges, standardizing or normalizing the data is essential before applying distance measures. This ensures that each feature contributes equally to the distance calculation, preventing features with larger scales from disproportionately influencing clustering results.

For this analysis, a combination of **standardized Euclidean distance** will be utilized. The data will be standardized to ensure each feature contributes equally to the distance calculation, and then Euclidean distance will be applied. This approach is appropriate given the continuous and numerical nature of income and population data, providing a clear and meaningful way to measure similarity between counties based on economic and demographic factors.

Normalization/Standardization

Standardization was applied to ensure all features were on a comparable scale, which is crucial for clustering algorithms. Without standardization, features with larger ranges or counts could disproportionately influence the analysis. By transforming each numerical feature to have a mean of 0 and a standard deviation of 1, we enable meaningful comparisons across variables. This standardization was conducted in R. The **county name** feature, being categorical, was excluded from this process as it does not require scaling.

Modeling and Evaluation

K-Means Clustering

The K-means clustering analysis divides Texas counties into two distinct clusters, represented by different colors and shapes on the plot. Each point represents a county, and the visual boundary around each cluster highlights the separation between groups. This clustering helps uncover patterns in economic resilience among Texas counties during the COVID-19 pandemic.

The plot below illustrates the clusters formed by K-means, showing how counties are grouped based on economic and demographic features. The analysis enables a clearer understanding of how certain counties exhibit similar characteristics in terms of income, rent burden, and pandemic impacts.

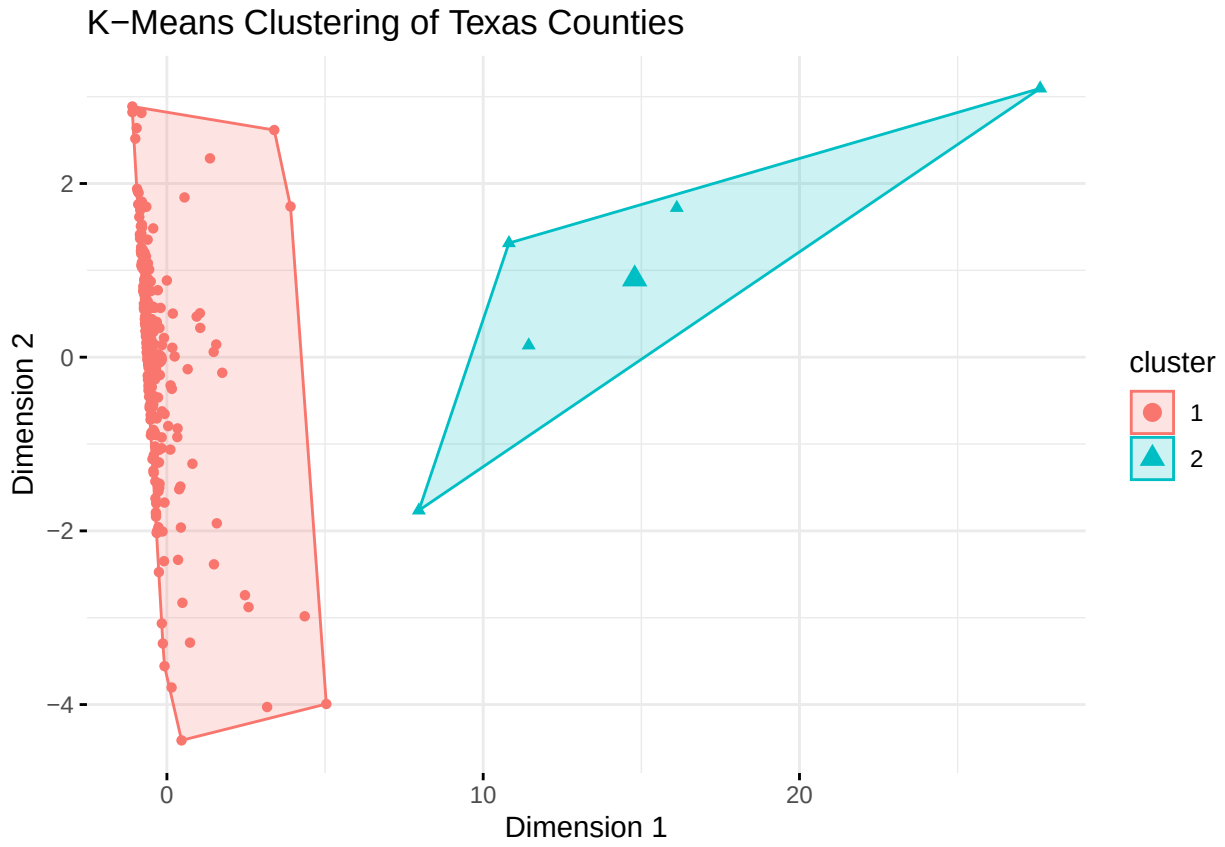


Figure 1: K-Means Clustering of Texas Counties

- **Cluster 1 (Red):** This cluster contains 250 points with an average silhouette width of 0.87, indicating strong cohesion and separation from other clusters. This suggests that Cluster 1 is compact and well-defined within the dataset, providing reliable clustering results for the majority of data points.
- **Cluster 2 (Blue):** This cluster has only 4 points with an average silhouette width of 0.45, indicating lower cohesion and weaker separation from other clusters. This suggests that Cluster 2 may be an outlier group or not well-defined, reflecting weaker clustering performance for these few points.

Overall, while Cluster 1 appears robust, the small size and low silhouette width of Cluster 2 indicate that the clustering may not be entirely effective in capturing distinct patterns across the dataset.

A summary statistics table provides a breakdown of the average values for key features across the two

clusters identified through K-means clustering. Each cluster represents a group of Texas counties with similar economic, demographic, and pandemic characteristics.

The table includes average values for **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 breakdown offers insights into the socioeconomic and demographic differences between the clusters, highlighting patterns that may impact economic resilience and stability within each group of counties.

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	49,706.28	24,729.68	1,366.56	541.67	4,852.43	86.9	61,404.08
2	59,259.60	31,300.20	83,126.40	33,012.80	186,039.60	2,148.4	2,425,999.00

The K-means clustering results suggest two groups of counties, but the distribution is highly uneven. Cluster 1 contains the vast majority of counties (approximately 95%), while Cluster 2 represents only a small subset. This imbalance raises questions about whether the clustering approach effectively captured meaningful differences across Texas counties. The observed patterns show some differences between clusters, but they may be a result of population size rather than distinct economic or pandemic-related characteristics.

- **Average Median Income**: Cluster 1 has an average median income of approximately 49,706 USD, while Cluster 2 has a slightly higher average of about 59,260 USD. Although Cluster 2's median income is somewhat higher, this difference may simply reflect the presence of a few more affluent, populous counties rather than a distinct socioeconomic profile.
- **Average Income per Capita**: The difference in income per capita is also modest, with Cluster 1 averaging around 24,730 USD and Cluster 2 around 31,300 USD. This suggests that while Cluster 2 may include counties with slightly higher economic indicators, these distinctions are not stark and do not strongly differentiate the clusters.
- **Rent Burden**: Cluster 2 has a slightly higher rent burden, with more households spending a significant portion of their income on rent (e.g., 13.67% for households spending over 50% of their income on rent compared to 8.31% in Cluster 1). However, this difference may be incidental rather than indicative of a clear socioeconomic separation.
- **COVID-19 Impact**: The most noticeable difference between clusters is in COVID-19-related metrics. Cluster 2 shows an average of 2,148 deaths, while Cluster 1 has an average of around 87 deaths. This disparity could reflect the concentration of higher population counties in Cluster 2, where higher transmission and mortality rates are expected. However, this metric alone may not provide enough justification for the clustering outcome, as it is heavily influenced by population density rather than underlying resilience or economic factors.
- **Total Population**: Cluster 2 includes counties with significantly larger populations (average 2,426,000) compared to Cluster 1 (61,404). This suggests that the clustering may be driven primarily by population size rather than meaningful economic or pandemic resilience indicators.

Overall, the clustering results highlight some differences in population and COVID-19 metrics but fall short of revealing a clear or actionable division based on economic resilience or pandemic impact. With 95% of counties grouped into a single cluster, this K-means clustering may not provide sufficient insight for distinguishing between counties in a way that aligns with the analysis objectives. Further refinement of the clustering approach—perhaps by adjusting the number of clusters, selecting different features, or exploring alternative clustering algorithms—might be necessary to achieve more meaningful separation among Texas counties.

Suitable Number of Clusters

The **Elbow Method** evaluates the optimal number of clusters by plotting the **Within-Cluster Sum of Squares (WSS)** across various cluster counts. WSS reflects the compactness of clusters, with lower values indicating tighter clusters. The "elbow" point, where additional clusters yield diminishing returns in WSS reduction, is considered optimal. In the elbow plot below, the "elbow" is not distinctly sharp, but the WSS curve flattens around 2 clusters. This suggests that 2 clusters may adequately capture the primary variation in the data, though the lack of a distinct elbow point implies limited clustering structure.

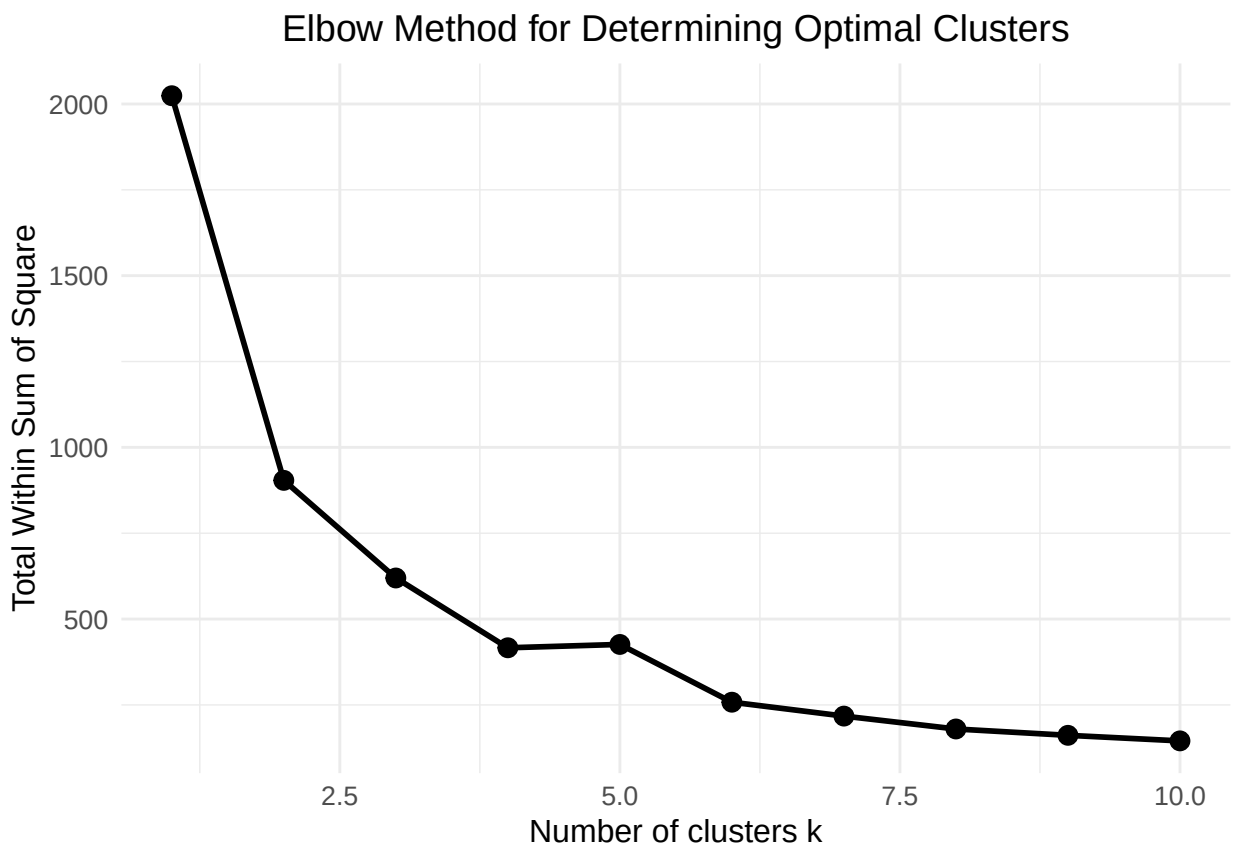


Figure 2: Elbow Method for Determining Optimal Clusters

The **Silhouette Method** measures how similar each data point is to its assigned cluster compared to other clusters. The Silhouette score ranges from -1 to 1, with higher values indicating better clustering cohesion. In the silhouette plot, the peak score occurs at 2 clusters, suggesting this configuration achieves the best separation. However, the modest silhouette values imply only moderate cohesion, indicating that the clustering structure may not be particularly strong in this dataset.

Based on the consistency of both the Elbow and Silhouette methods, 2 clusters were selected as the final clustering solution. Although both methods support this choice, the limited clustering structure observed in the plots suggests that additional refinement may be needed to achieve a more nuanced grouping.

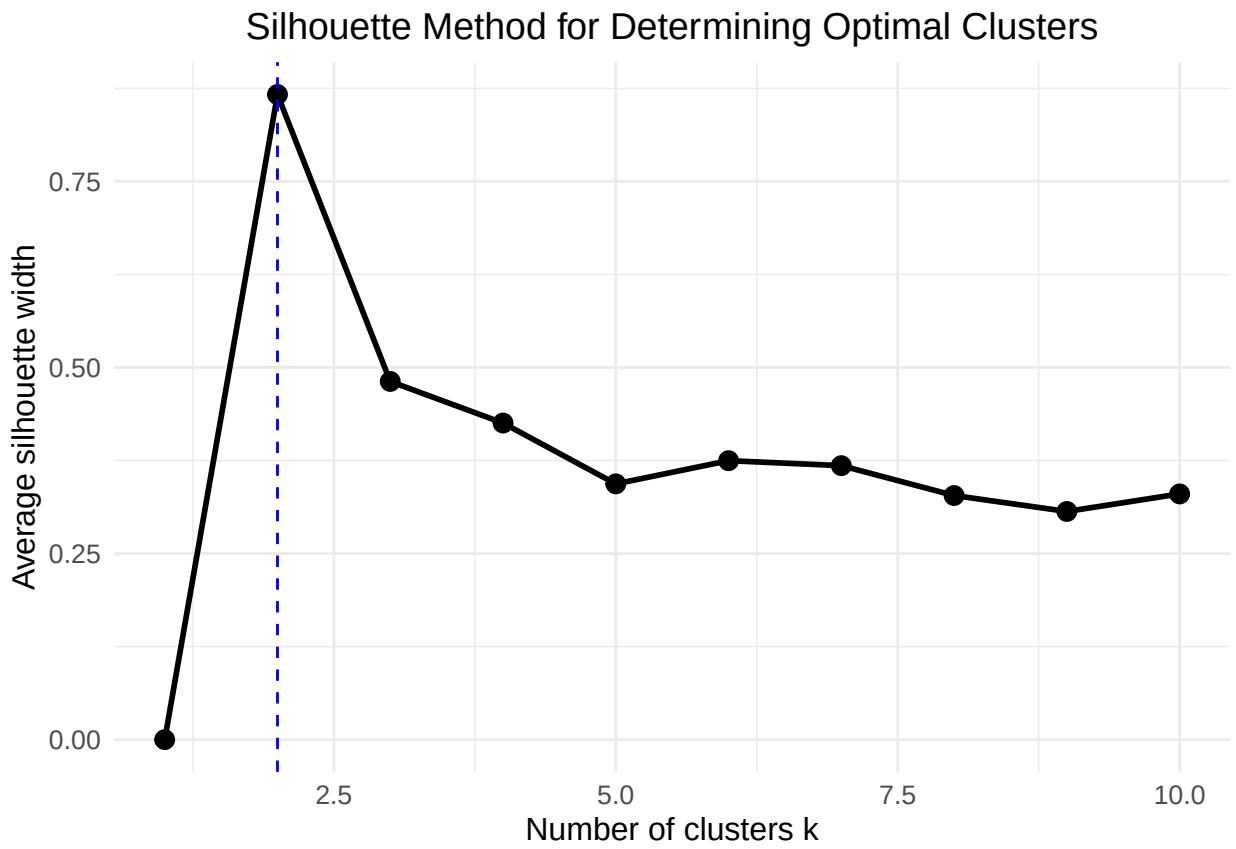


Figure 3: Silhouette Method for Determining Optimal Clusters

Based on the consistency of both the Elbow and Silhouette methods, 2 clusters were selected as the final clustering solution. Although both methods support this choice, the limited clustering structure observed in the plots suggests that additional refinement may be needed to achieve a more nuanced grouping.

Unsupervised Evaluation

A **Silhouette Plot** is used to evaluate the quality and cohesion of the clusters generated by the K-means algorithm. The silhouette width measures how well each data point fits within its assigned cluster compared to neighboring clusters. Values close to 1 indicate that points are well-matched to their own cluster and poorly matched to other clusters, reflecting high-quality clustering. Values near 0 suggest ambiguity, as points lie roughly equidistant from multiple clusters.

Evaluation Using Ground Truth Feature

To assess the alignment between economic clusters and real-world pandemic outcomes, we use the **COVID-19 death-to-case ratio** as a ground truth feature. This ratio, categorized into **"Lower"** (< 0.025) and **"Higher"** (> 0.025), serves as a proxy for pandemic resilience, allowing us to examine whether wealthier counties—identified through income-related clustering—exhibit lower mortality rates relative to confirmed cases.

- This binary categorization simplifies interpretation, making it easier to analyze and compare mortality rates across income groups while maintaining consistency with the 2-cluster structure of the unsupervised analysis.
- Mortality rates, as measured by the death-to-case ratio, provide key public health insights into the severity of the pandemic's impact, which may correlate with economic resilience.

A **contingency table** below compares the clusters generated by K-means with the death-to-case ratio categories, allowing us to evaluate if the clustering effectively captures differences in pandemic outcomes associated with economic conditions.

Table 4: Contingency Table of Clusters and Death Categories

Lower	Higher
144	105
5	0

Supervised Evaluation

This **K-Means clustering** analysis groups Texas counties into two clusters using the **death-to-case ratio** (COVID-19 deaths per confirmed case) and **income per capita**. Both features are scaled to ensure equal contributions to the clustering process, with a mean of zero and a standard deviation of one.

In contrast to the unsupervised approach, this clustering deliberately focuses on the economic conditions and pandemic outcomes to explore their relationship. The resulting clusters highlight counties with similar characteristics in terms of economic status and COVID-19 impact. Counties within the same cluster are more alike in these metrics than those in the other cluster.

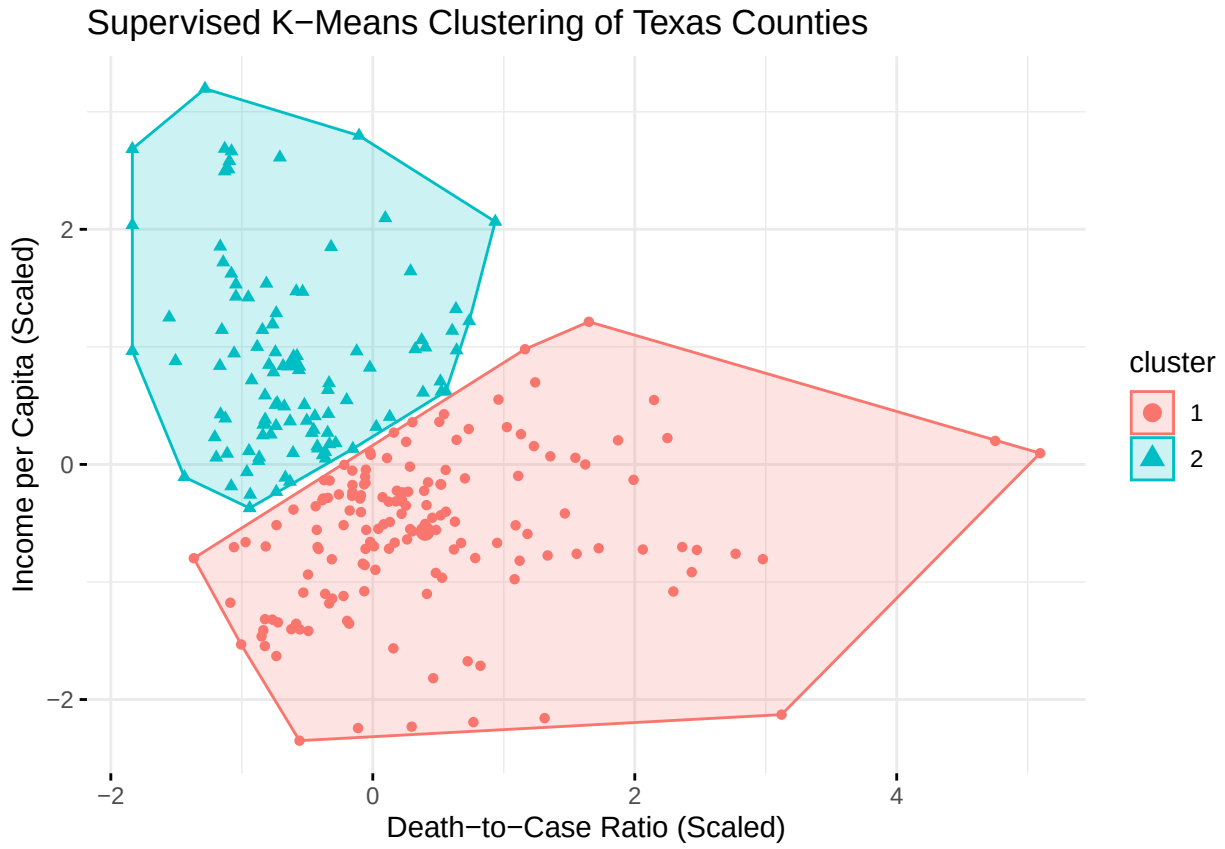


Figure 4: K-Means Clustering of Texas Counties Supervised

A summary statistics table follows, which provides an in-depth breakdown of the average values of key features across the two clusters identified through K-Means clustering. The table includes the **average median income**, **income per capita**, **rent burden levels** (households spending over 50% and between

30-35% of income on rent), **confirmed COVID-19 cases**, **COVID-19 deaths**, and **total population** for each cluster. These statistics help in interpreting the composition and characteristics of the counties in each group, making it easier to assess economic and demographic differences.

Table 5: 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	43,937.72	21,876.61	804.73	297.67	3,309.48	80.99	0.03	37,969.71
2	58,917.73	29,376.93	6,265.17	2,518.79	16,159.45	197.90	0.02	213,962.83

This clustering employs easily interpretable features—**death-to-case ratio** and **income per capita**—which offer a direct reflection of the economic conditions of each county. These features enhance the clusters' interpretability, providing valuable economic insight. The even distribution of data points between the two clusters results in a distinct separation of counties, which indicates that the clustering is a reliable representation of underlying economic and pandemic-related characteristics. The balanced representation of counties in Clusters 1 and 2 suggests that the grouping accurately captures the differences in economic resilience and pandemic outcomes.

- **Average Median Income:** Cluster 1 has an average median income of $58,917.73, while Cluster 2 has an average of $43,937.72. This difference of approximately $15,000 highlights the substantial economic disparity between the two clusters.
- **Average Death-to-Case Ratio:** Cluster 1 has an average ratio of 0.0166, which is considerably lower than Cluster 2's average ratio of 0.0301. This suggests a significant relationship between economic status and the severity of pandemic outcomes, with higher incomes corresponding to better health outcomes.

The two K-Means clustering analyses (supervised and unsupervised) aim to classify Texas counties based on economic and pandemic-related factors, but they vary in terms of their precision, clarity, and ease of interpretation.

Within Clusters 1 and 2, counties are further grouped based on their income and rent burdens. There are three income groups: **Low** (Income per Capita $<$ $25,000), **Middle** ($25,000 $\leq$ Income per Capita $<$ $40,000), and **High** (Income per Capita $>$ $40,000). Additionally, two rent burden groups are defined: **Low Rent Burden** (Rent over 50 Percent $\leq$ 5,000) and **High Rent Burden** (Rent over 50 Percent $>$ 5,000). These categories provide a more nuanced comparison between the clusters and further illustrate the intersection of economic conditions and rent burdens across Texas counties.

Table 6: 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	39,219.50	17,058.50	0.03	591,047.25
1	Low Income	Low Rent Burden	43,140.55	21,125.65	0.03	23,914.66
1	Middle Income	Low Rent Burden	48,876	26,590.88	0.04	18,993.5
2	High Income	High Rent Burden	90,124	41,609	0.01	914,075
2	Low Income	High Rent Burden	46,262.00	24,273.00	0.02	245,720.00
2	Low Income	Low Rent Burden	46,604.71	23,835.43	0.01	26,067.43
2	Middle Income	High Rent Burden	62,475.78	30,879.67	0.01	956,242.94
2	Middle Income	Low Rent Burden	58,966.32	29,439.27	0.02	41,291.97

Cluster Analysis Summary

Cluster 1 Characteristics

- **Low Income & High Rent Burden:** With a median income of around 39,220 USD and income per capita of 17,059 USD, this subgroup has a death-to-case ratio of 0.0258, impacting a larger population (about 591,047). This suggests significant economic strain and relatively higher mortality rates in low-income, high-rent areas.
- **Low Income & Low Rent Burden:** This subgroup, with a slightly higher median income of 43,141 USD and income per capita of 21,126 USD, has a similar death case ratio of 0.0279 but a smaller population of around 23,915. This may indicate that smaller, lower-rent communities still faced pandemic challenges.
- **Middle Income & Low Rent Burden:** With the highest income levels in Cluster 1 (median of 48,879 USD and per capita 26,591 USD), this group exhibits the highest death-to-case ratio of 0.0419, implying that middle-income regions with low rent burden faced considerable health challenges potentially tied to other socioeconomic factors.

Cluster 2 Characteristics

- **Low Income & High Rent Burden:** Median income here is around 46,262 USD with an income per capita of 24,273 USD, and a relatively low death case ratio of 0.0157. Despite economic vulnerability, pandemic outcomes were better than comparable subgroups in Cluster 1.
- **Low Income & Low Rent Burden:** With a median income of 46,605 USD and income per capita of 23,835 USD, this subgroup has a low death case ratio of 0.0118, indicating some mitigation of low income impacts due to a reduced rent burden.
- **Middle Income & High Rent Burden:** This group, with a median income of 62,476 USD and per capita income of 30,880 USD, has a death case ratio of 0.0119. Economic stability here supports moderate resilience in pandemic outcomes.
- **Middle Income & Low Rent Burden:** Featuring a median income of 58,966 USD and per capita income of 29,439 USD, this group has a death-to-case ratio of 0.0183, suggesting that lower rent burden provides economic stability, though health outcomes still lag.
- **High Income & High Rent Burden:** This high-income group, with median income of 90,124 USD and per capita income of 41,609 USD, displays the lowest death case ratio (0.0075). It highlights that wealthier, high-rent areas managed the pandemic effectively, supported by substantial population density (approx. 914,075).

Insights and Implications

Higher income levels in Cluster 2 are linked with significantly lower death case ratios, emphasizing that economic stability aids in pandemic resilience. Conversely, lower-income groups in both clusters tend to show higher mortality rates, with rent burden further intensifying economic strain. Even within high-rent subgroups, those with higher income levels (Cluster 2) fare better in health outcomes. Populous, high-income areas (e.g., high-income, high-rent burden in Cluster 2) show strong resilience, likely due to robust infrastructure and healthcare access.

In summary, wealthier regions, even under high rent burden, mitigated the pandemic's impact more effectively. These findings underscore the importance of socioeconomic stability and housing in managing public health crises, guiding stakeholders toward resilient investment and targeted support for vulnerable areas.

Visualizing K-Means Clustering Results

The following plot shows the K-Means clustering for Texas counties, categorized by Death Case Ratio and Income per Capita. Clusters are color-coded and outlined to reflect K-Means results, with points labeled according to income and rent burden groups for added economic context.

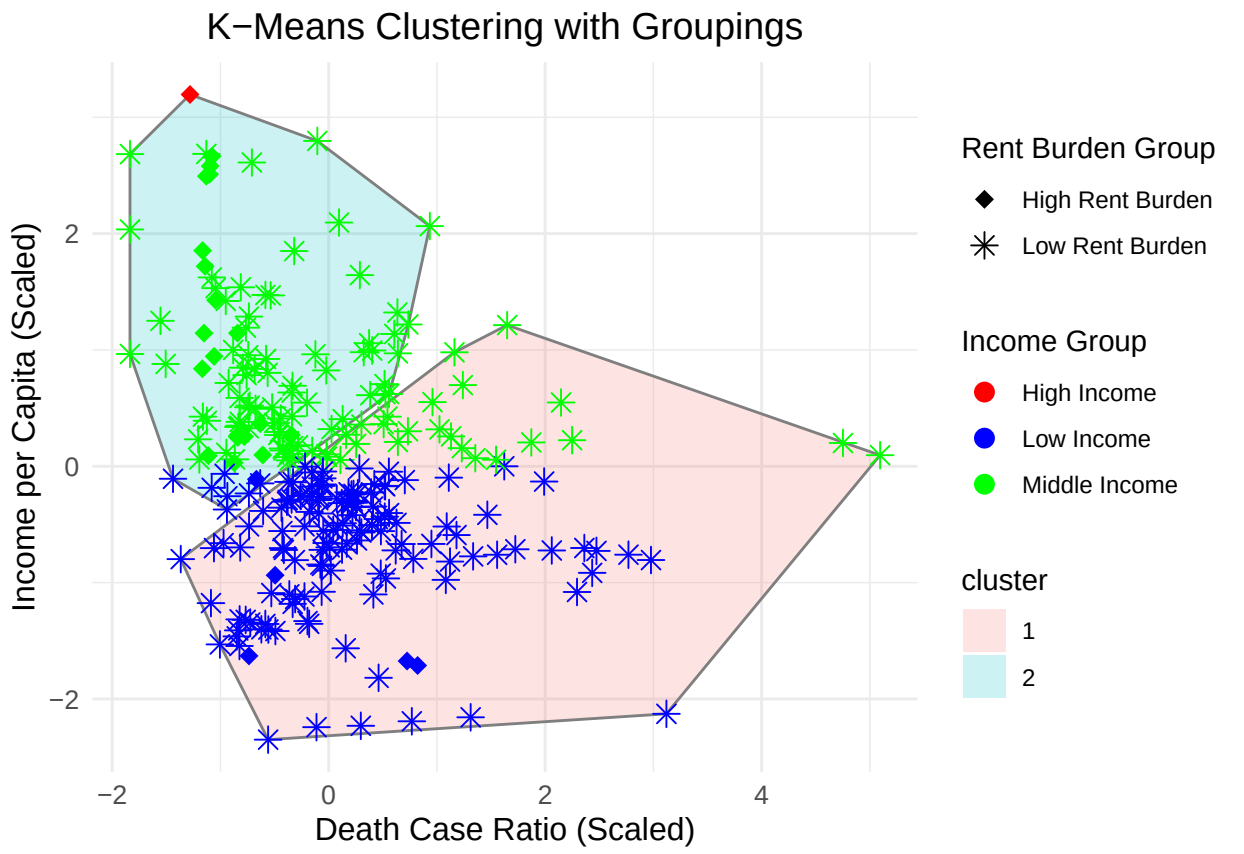


Figure 5: K-Means Clustering with Groupings

Cluster 1 (Red Region)

This cluster is distinguished by a higher Death Case Ratio and generally low to middle Income per Capita. The high density of blue points represents low-income areas, suggesting a concentration of economically vulnerable populations within this cluster. Green points, signifying middle-income groups, are present but less dense. Notably, this cluster includes both high and low rent burden groups, with star-shaped markers indicating high rent burden areas interspersed throughout. The combination of low income and high rent burden suggests compounded financial stress, which may contribute to poorer health outcomes within this cluster.

Cluster 2 (Blue Region)

This cluster represents counties with a lower Death Case Ratio and generally higher Income per Capita. Red points, marking high-income areas, cluster toward the upper end of the income axis, aligning with wealthier regions that experienced better pandemic outcomes. Middle-income areas (green points) are also present, indicating that this cluster captures moderately affluent areas. These areas tended to fare better in terms of health resilience compared to Cluster 1. Even among the high rent burden subgroups in Cluster 2, the Death Case Ratios remain relatively low, suggesting that wealthier regions could leverage resources to mitigate pandemic impacts despite housing cost pressures.

Overall, the clustering results suggest a strong correlation between income level and health resilience. Higher Income per Capita aligns with lower Death Case Ratios, likely due to better access to healthcare, infrastructure, and resources. In contrast, low-income areas, particularly those burdened by high rent, appear more vulnerable. Middle-income areas are present in both clusters, indicating that pandemic outcomes for this group varied widely and may have been influenced by factors beyond income alone, such as healthcare access, population density, and social support networks.

The following table presents the purity scores for the two different subgroup classifications: Income Groups

Table 7: Purity Scores by Grouping

Grouping	Purity.Score
Income Groups	0.870
Rent Burden Groups	0.906

and Rent Burden Groups.

Purity is a metric that evaluates clustering quality by measuring how well clusters align with predefined classes. A higher purity score indicates more homogeneous clusters.

- **Income:** The purity score for income is 0.87, meaning that 87% of data points within clusters are correctly grouped by income level (Low, Middle, or High). This high score suggests that income is a strong factor in clustering, although some overlap exists, indicating that income alone does not fully define the cluster structure.
- **Rent Burden:** With a slightly higher purity score of 0.91, rent burden appears to be an even stronger factor, grouping counties more distinctly by housing affordability stress (High or Low Rent Burden). This suggests that rent burden may be a clearer differentiator in the clustering model.

Overall, the high purity scores indicate that K-Means clustering captures meaningful differences among counties. While both income and rent burden contribute to cluster distinctions, rent burden seems particularly impactful. This insight is valuable for stakeholders focused on economic resilience and addressing housing stress.

Hierarchical Clustering

We will perform hierarchical clustering on the Texas counties dataset to identify patterns and group similar counties based on the selected features. Hierarchical clustering does not require us to specify the number of clusters beforehand and provides a dendrogram that helps visualize the cluster formation at various levels.

Hierarchical Clustering Using Ward's Method

Feature Selection and Data Preparation

We will use the same numerical features as in the K-means clustering for direct comparison:

- Median Income
- Income per Capita
- Rent > 50% Income
- Rent 30-35% Income
- Income <10,000 USD
- Income 50,000-59,999 USD
- Income 100,000-124,999 USD
- Total Population

We will standardize these features to ensure equal contribution to the clustering process.

Computing the Distance Matrix We compute the distance matrix using Euclidean distance, which measures the straight-line distance between points in the feature space.

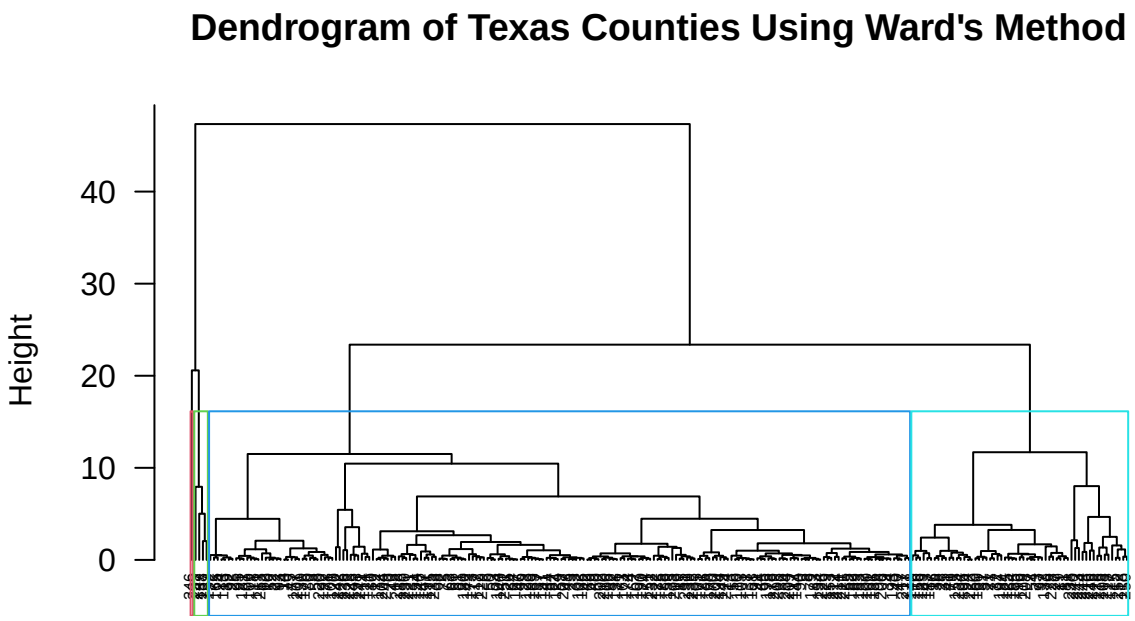


Figure 6: Dendrogram of Texas Counties Using Ward's Method

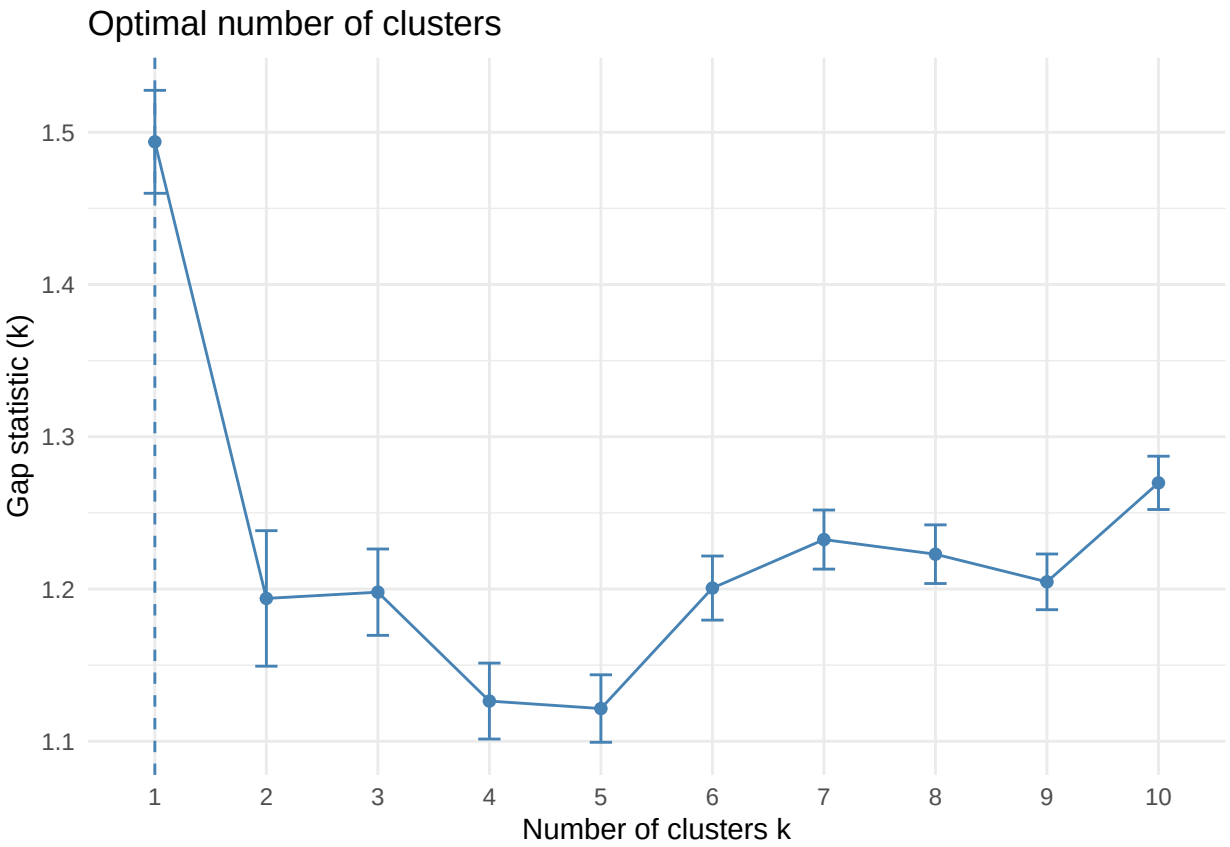
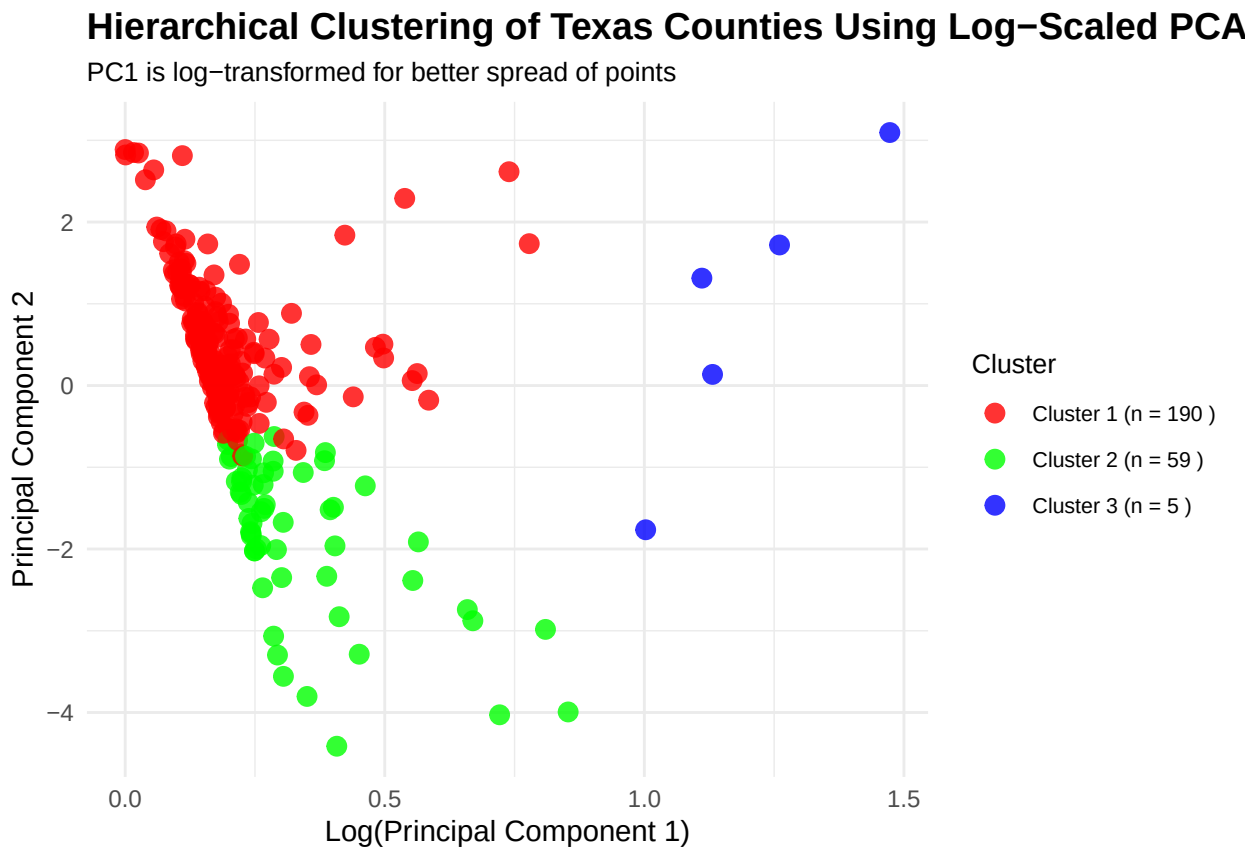
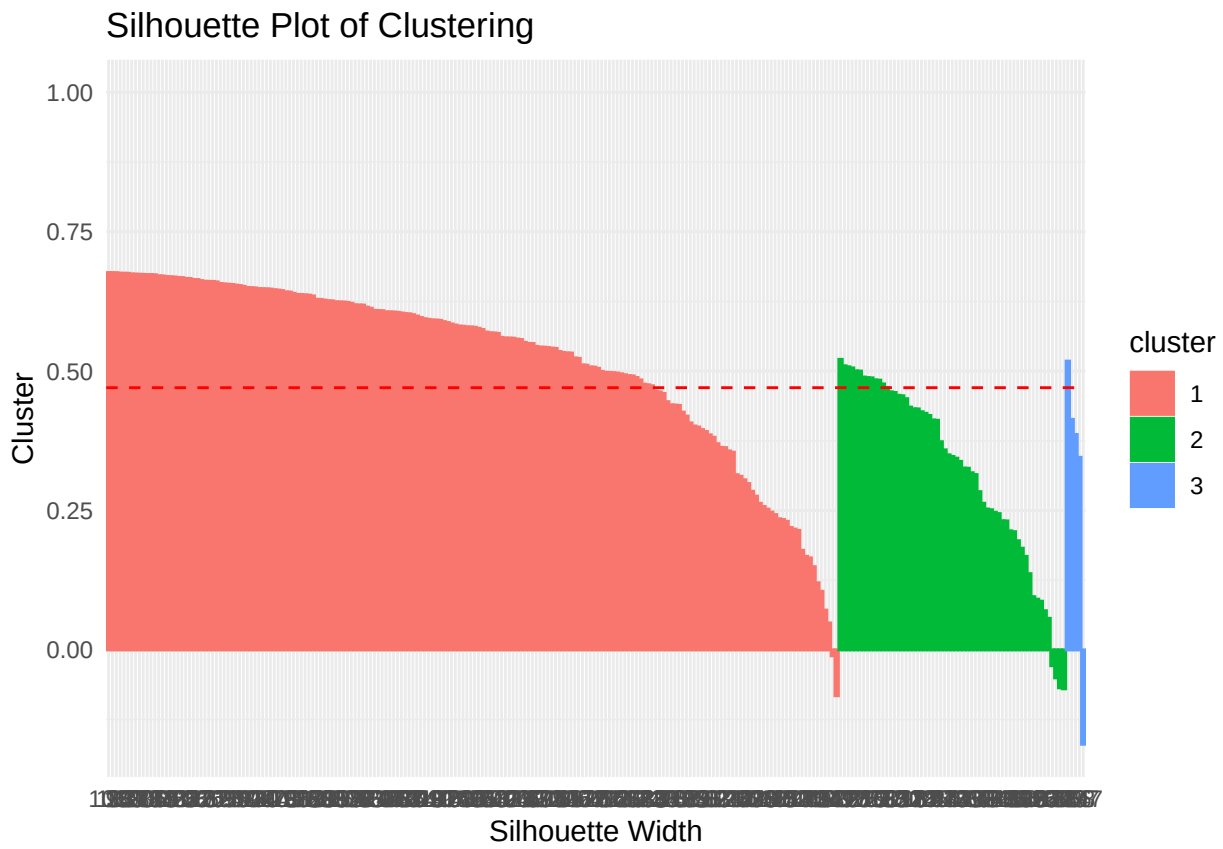


Figure 7: Gap Statistic for Determining Optimal Clusters



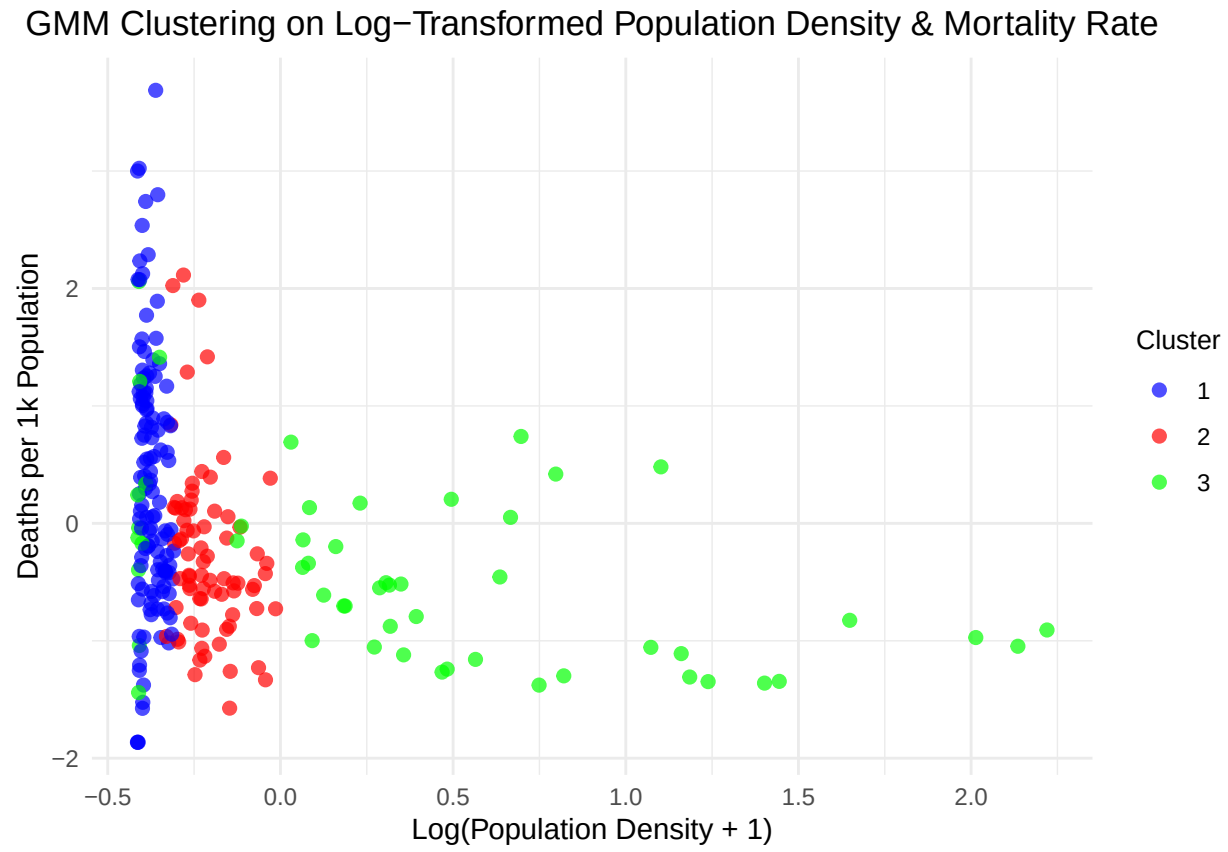

```
##   cluster size ave.sil.width
## 1       1  190          0.52
## 2       2   59          0.32
## 3       3    5          0.30
```

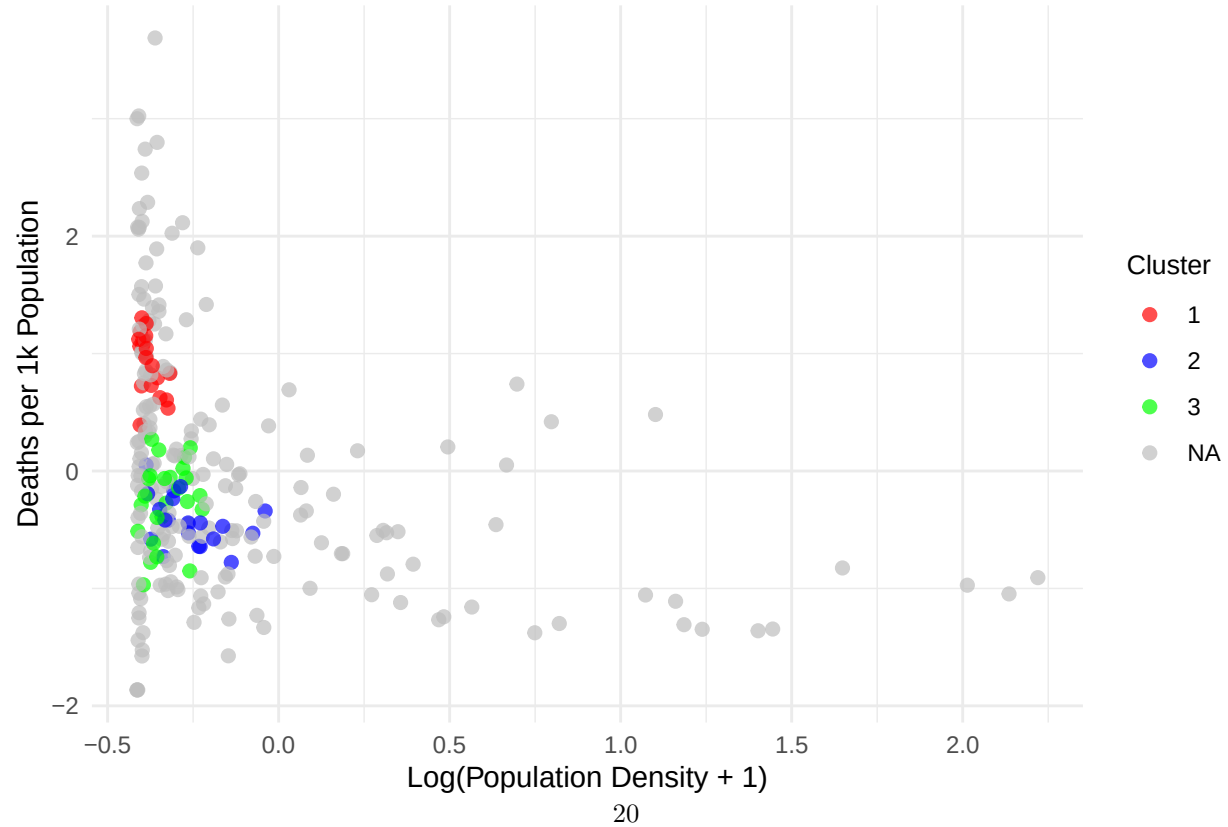



```
## [1] 0.7459644
```

```
## # A tibble: 6 x 8
##   county_name total_deaths total_cases total_population cases_per_k deaths_per_k
##   <chr>              <dbl>       <dbl>            <dbl>       <dbl>        <dbl>
## 1 Anderson C~       -0.127     -0.101           -0.135       0.668       -0.485
## 2 Andrews Co~       -0.243     -0.233           -0.233       0.321        0.0662
## 3 Angelina C~        0.175     -0.0644          -0.0632     -0.0737       0.384
## 4 Aransas Co~       -0.263     -0.255           -0.219       -1.55        -0.726
## 5 Archer Cou~       -0.310     -0.262           -0.257       0.111        -0.778
## 6 Armstrong ~       -0.320     -0.281           -0.273      -0.423         1.06
## # i 2 more variables: area_sqmiles <dbl>, pop_density <dbl>
```

Exceptional Work





Recommendations

Discuss how the model can be interpreted and the recommendations based on the findings. Explain the utility for the stakeholders.
After analyzing our model's results, if a client has an interest in opening a business in an affluent, high land population density, and high COVID-19 performing county in Texas, they should consider the following counties.
After taking cluster "2" in the second layer K-means cluster, and sorting from descending order according to population density the three top counties are:
The three results in cluster "2" for the second layer hierarchical clustering were:
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)

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.