Data Mining Project 2

Olivia Hofmann, Matias Barcelo, Michael Perkins

2024-10-23

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.

Data Preparation

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, rent burden levels, and the distribution of income across different 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	2469.768	8601.256	0	98715
Income $50,000-$59,999	2945.197	10790.454	3	122390
Income $100,000-$124,999	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, the basic statistics were done on a subset of the data. 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 Under 10 Percent:**
- **Rent 25 to 30 Percent:** This provides a threshold of moderate rent burden.
- **Rent Over 50 Percent:** This is a key indicator of severe rent burden, which can signify economic strain in a county.
- **Income Less than $10,000:** Reflects the population in extreme poverty, which is crucial for understanding economic vulnerability.
- **Income $25,000 - $29,999:**
- **Income $45,000 - $44,999:**
- **Income $100,000 - $124,999:**

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	Ratio	Households earning <$10,000
Income $50,000-$59,999	Ratio	Households earning $50,000-$59,999
Income $100,000-$124,999	Ratio	Households earning $100,000-$124,999
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 normalized 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.

Modeling

Normalization

Normalization is essential for standardizing 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 normalize the numerical features before proceeding with clustering or further analysis.

The normalization was based on the total population of each county, and for each numerical column, a corresponding normalized column was created. A new dataset was then constructed, retaining either the normalized or original version of each feature, depending on its relevance. The following features were kept as not normalized:

- **county_name:** A categorical variable representing the county's name. Since normalization is typically applied to numerical data, this feature was excluded from the process.
- **total_pop:** This variable was used as the basis for normalization. Normalizing it would not be meaningful as it serves as the denominator for other variables.
- **median_income:** Already an average measure of income at the county level, this feature did not require normalization because it provides a direct summary of income status rather than a count or proportion.
- **income_per_capita:** Similar to median income, this statistic reflects income averaged per individual and does not need normalization, as it is already scaled relative to the population.

Clustering Analysis

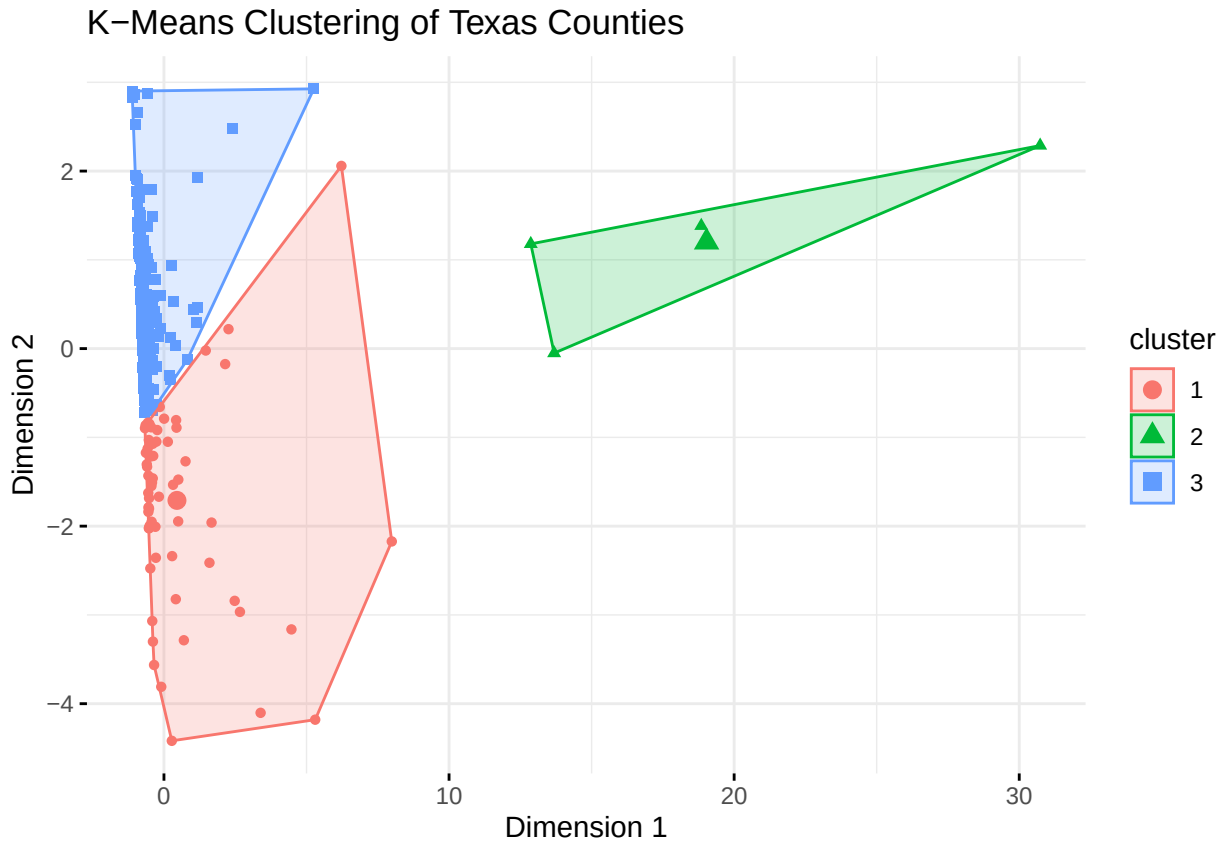
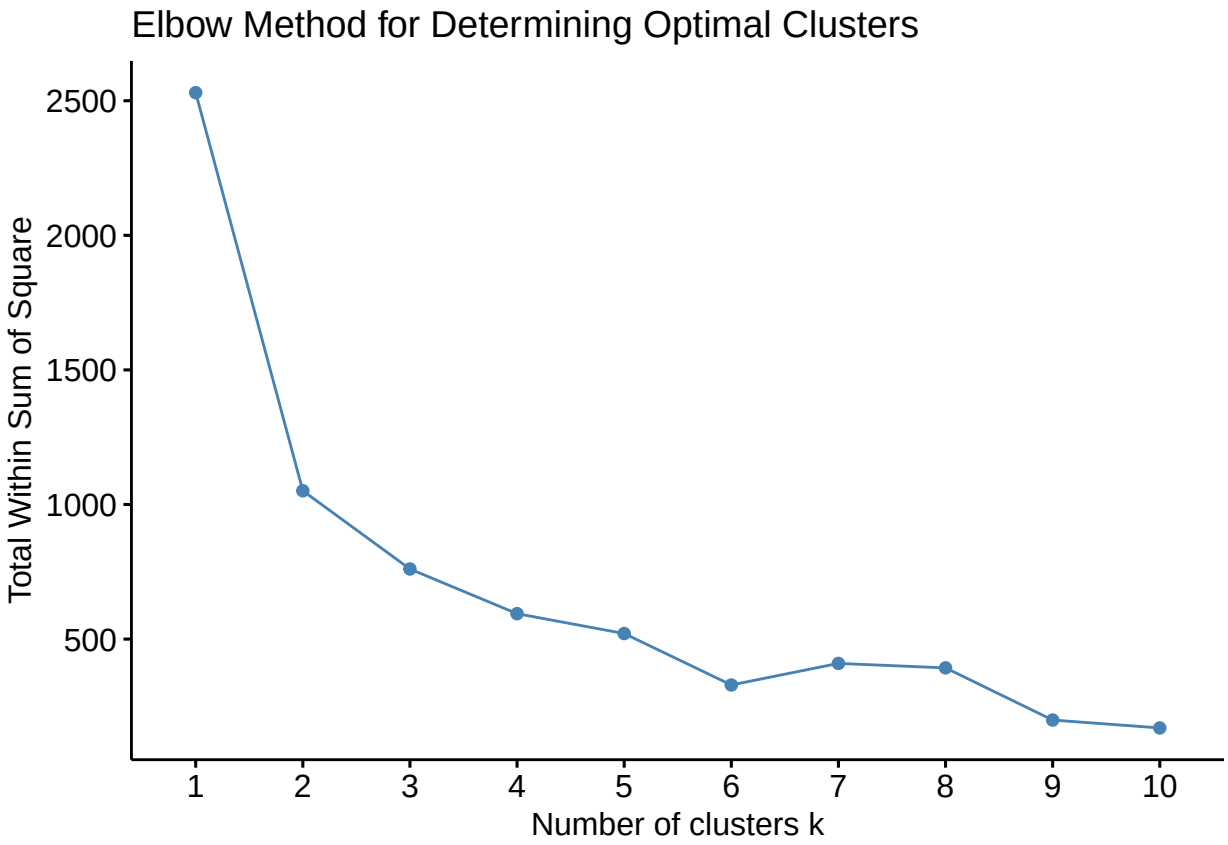
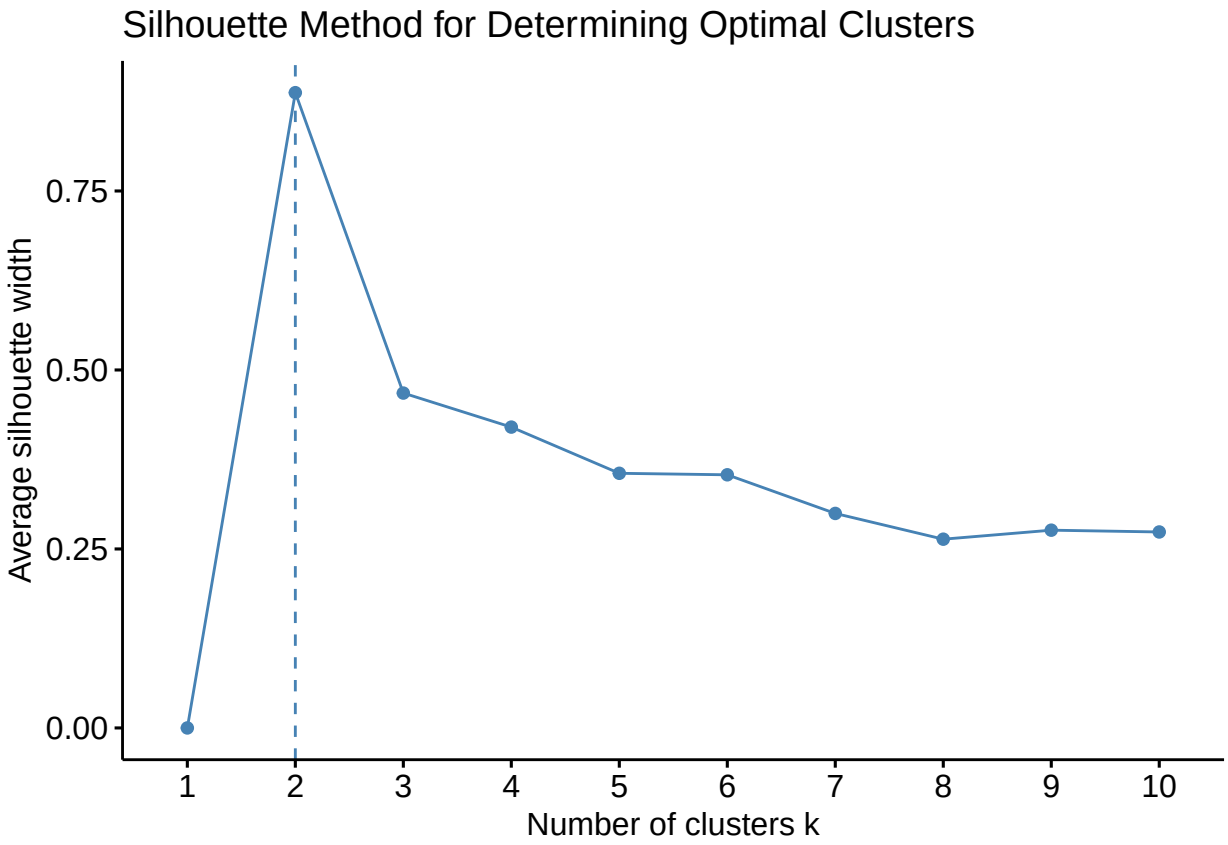


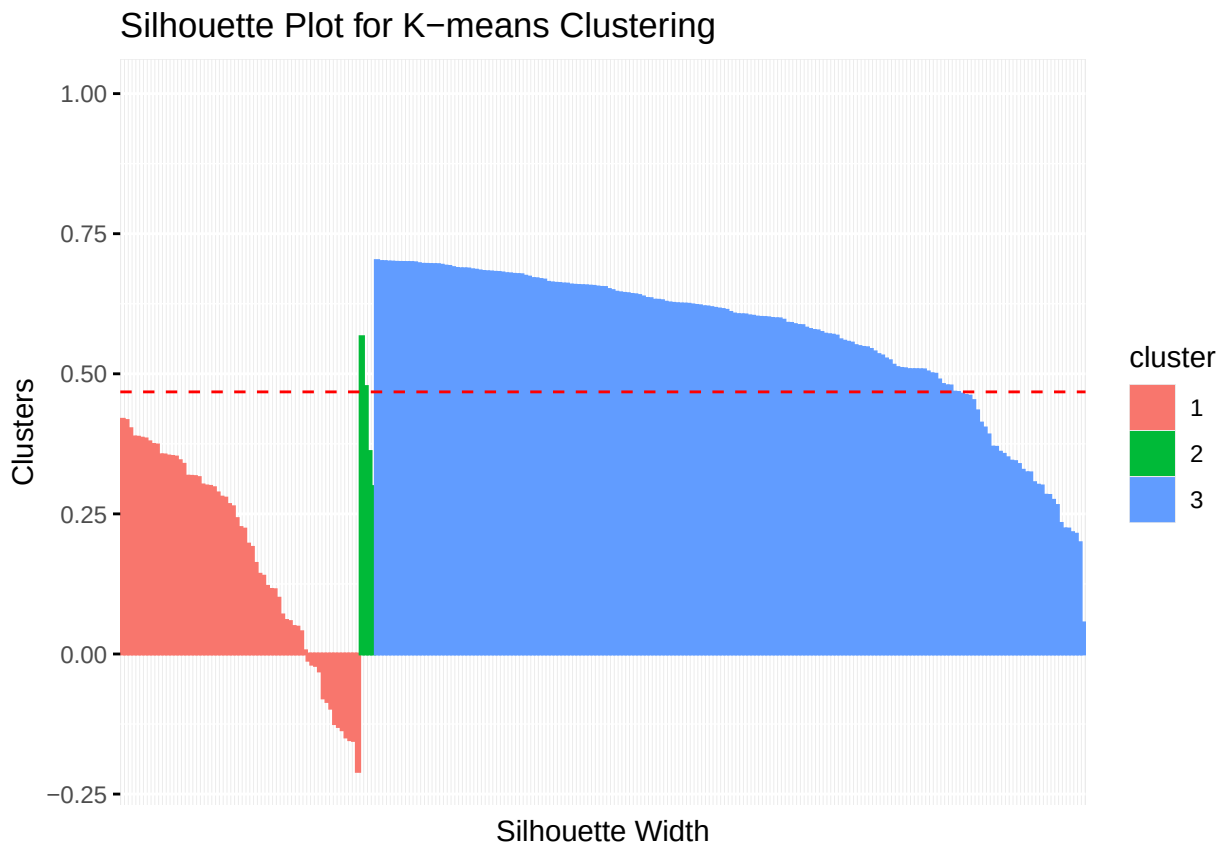
Table 3: Summary Statistics by Cluster

cluster	Avg Median Income	Avg Income per Capita	Avg Rent > 50%	Avg Rent 30-35%	Avg Confirmed Cases	Avg Deaths	Total Population
1	65315.78	31171.44	3603.4603	1566.2063	11541.968	149.47619	151892.60
2	56987.00	29420.25	91995.0000	36522.2500	217182.500	2529.00000	2738352.75
3	44547.17	22634.81	860.4652	295.0856	2901.497	68.69519	36882.18


Elbow Method for Determining Optimal Clusters
Total Within Sum of Square
2500
2000
1500
1000
500
1
2
3
4
5
6
7
8
9
10
Number of clusters k




```
##   cluster size ave.sil.width
## 1       1   63          0.18
## 2       2    4          0.43
## 3       3  187          0.57
```



Hierarchical Clustering Analysis

```
## Warning: The ‘<scale>‘ argument of ‘guides()‘ cannot be ‘FALSE‘. Use "none" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextra/issues>.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

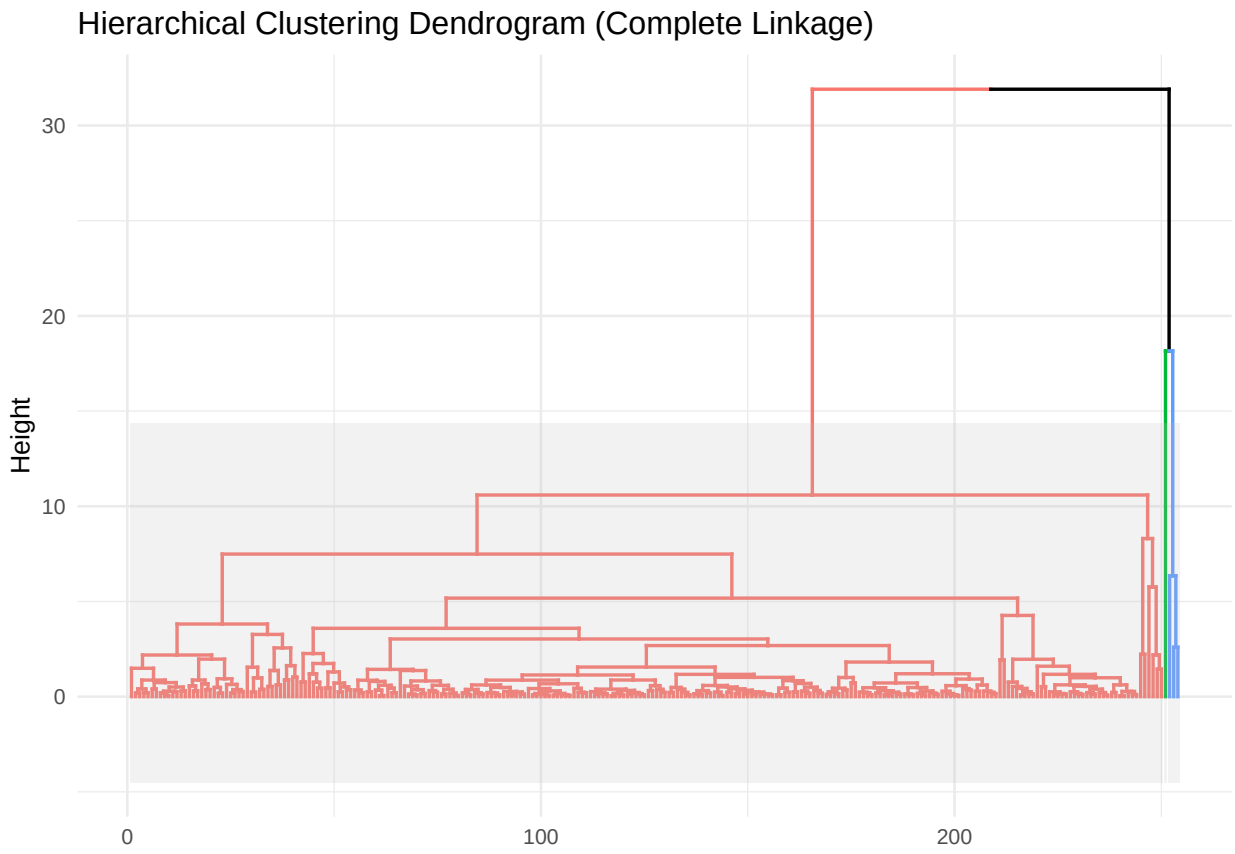
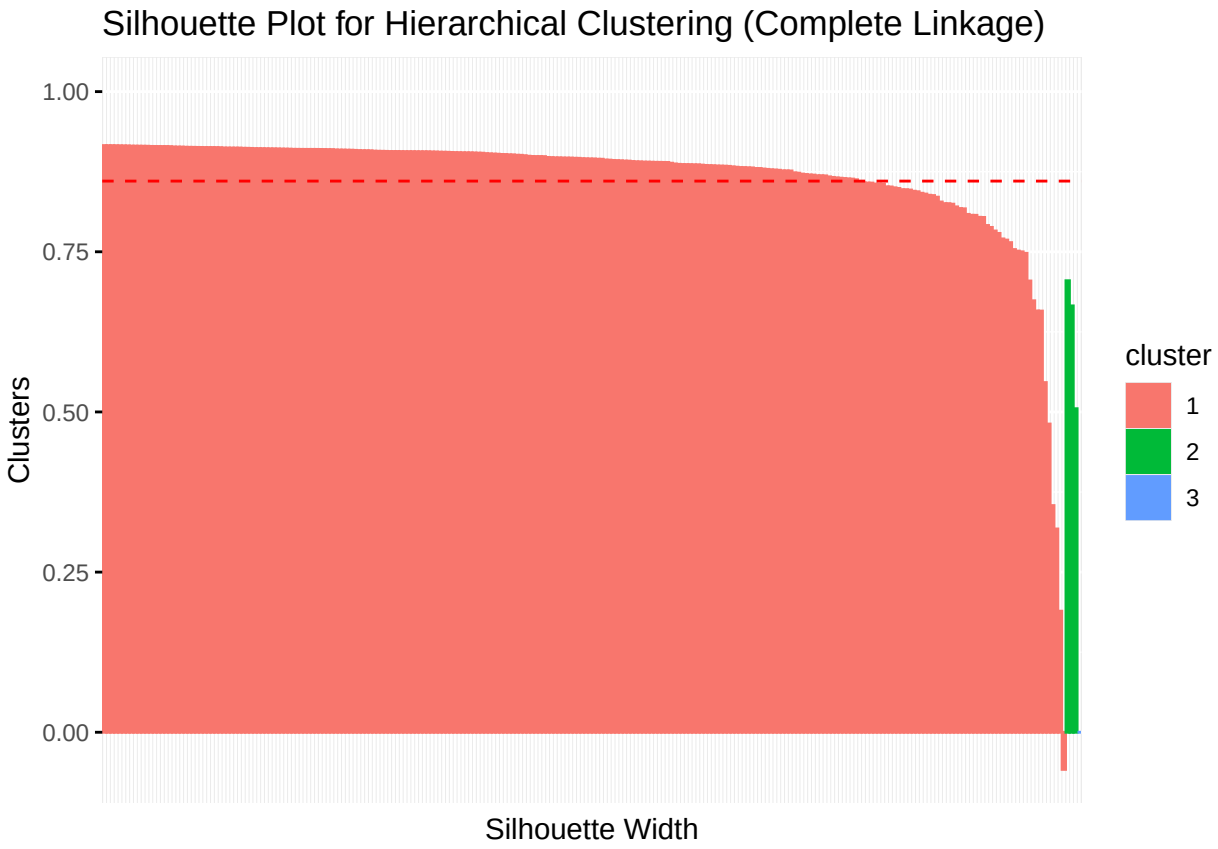
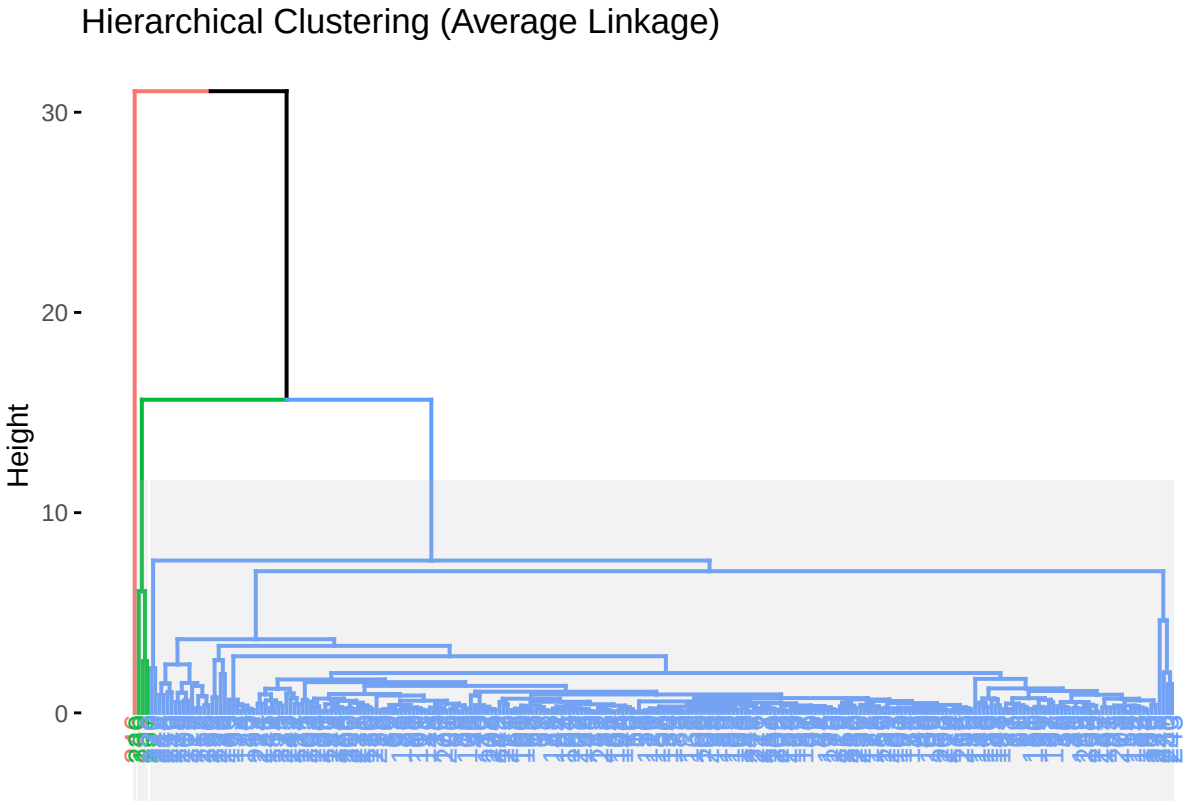


Table 4: Summary Statistics by Hierarchical Cluster

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	1551.70	615.408	5078.896	89.052	65864.8
2	56719.00	28941.67	69770.67	28261.333	194124.667	2097.000	2142630.7
3	57791.00	30856.00	158668.00	61305.000	286356.000	3825.000	4525519.0

```
##   cluster size ave.sil.width
## 1       1  250          0.87
## 2       2    3          0.63
## 3       3    1          0.00
```


Silhouette Plot for Hierarchical Clustering (Complete Linkage)
1.00
0.75
0.50
0.25
0.00
Clusters
Silhouette Width
cluster
1
2
3


Hierarchical Clustering (Average Linkage)
Height
30
20
10
0

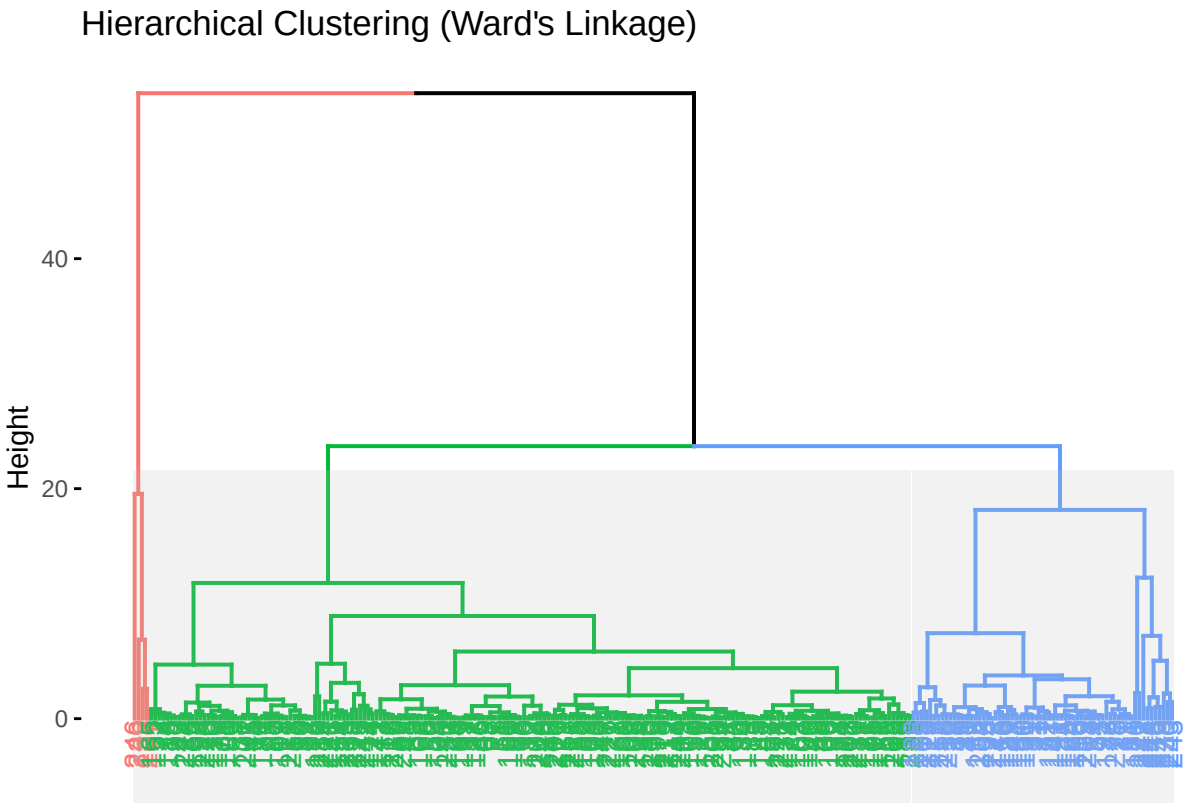


Table 5: Average Silhouette Widths by Linkage Method

Linkage_Method	Avg_Silhouette_Width
Complete	0.8603571
Average	0.8603571
Ward's	0.4651074