Olivia's Portion: Project 2

Olivia Hofmann

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: Table of Features and Basic Statistics

Feature	Description	Mean	Std_Dev	Min	Max
median_income	Median income in the county (USD)	49894.339	12132.676	24794	93645
income_per_capita	Per capita income in the county (USD)	24859.020	5240.752	12543	41609
rent_over_50_percent	Households with rent > 50% of income (%)	2976.004	13179.056	0	158668
rent_30_to_35_percent	Households with rent 30-35% of income (%)	1180.870	5203.838	0	61305
income_less_10000	Households earning <$10,000 (%)	2469.768	8601.256	0	98715
income_50000_59999	Households earning $50,000-$59,999 (%)	2945.197	10790.454	3	122390
income_100000_124999	Households earning $100,000-$124,999 (%)	3205.157	11657.055	0	131467
total_pop	Total population of the county	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 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 $50,000 - $59,999:** This is a middle-income bracket and can act as a proxy for general economic health.
- **Income $100,000 - $124,999:** A higher income bracket that helps assess the presence of wealthier households.
- **Income Less than $10,000:** Reflects the population in extreme poverty, which is crucial for understanding economic vulnerability.

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: Scale of Measurement of Features and Descriptions

Feature	Scale_of_Measurement	Description
median_income	Ratio	Measures income in dollars. Has a true zero (no income).
income_per_capita	Ratio	Measures income per person. Has a true zero.
rent_over_50_percent	Ratio	Number of households paying more than 50% of income in rent.
rent_30_to_35_percent	Ratio	Number of households paying between 30-35% of income in rent.
income_less_10000	Ratio	Number of households earning less than $10,000.
income_50000_59999	Ratio	Number of households earning between $50,000 and $59,999.
income_100000_124999	Ratio	Number of households earning between $100,000 and $124,999.
total_pop	Ratio	Total population count, which has a true zero (no 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

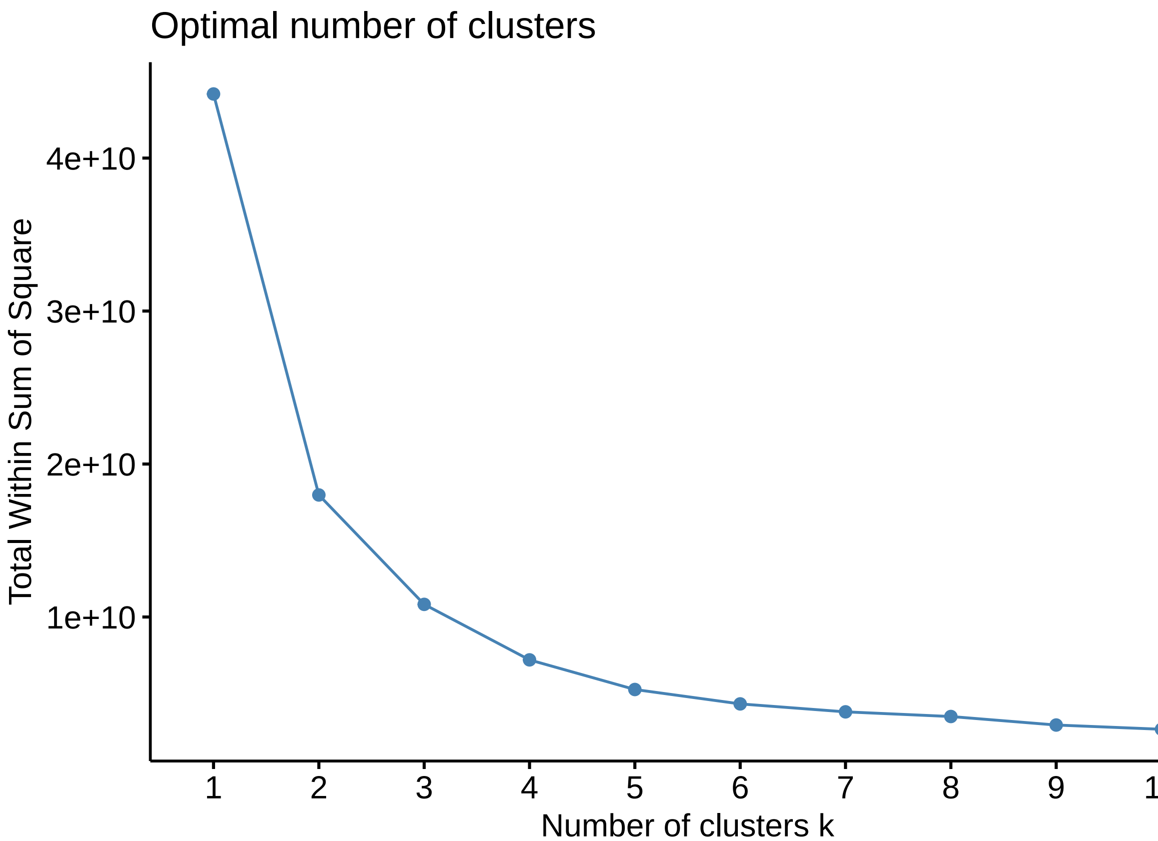
Normalization

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.

Cluster Analysis



K-Means Clustering

Table 3: Summary Statistics for Cluster 1

Feature	Mean	SD	Min	Max
norm_confirmed_cases	0.0743	0.0232	0.0135	0.1167
norm_deaths	0.0011	0.0008	0.0000	0.0033
total_pop	215373.3902	397853.1338	74.0000	1983675.0000
median_income	70856.1951	8611.5309	60275.0000	93645.0000
income_per_capita	32445.8293	4595.9384	20676.0000	41609.0000
norm_rent_burden_not_computed	0.0173	0.0204	0.0026	0.1081
norm_rent_over_50_percent	0.0140	0.0095	0.0000	0.0405
norm_rent_40_to_50_percent	0.0064	0.0051	0.0000	0.0270
norm_rent_35_to_40_percent	0.0046	0.0036	0.0000	0.0152
norm_rent_30_to_35_percent	0.0064	0.0038	0.0000	0.0161
norm_rent_25_to_30_percent	0.0079	0.0052	0.0000	0.0224
norm_rent_20_to_25_percent	0.0121	0.0070	0.0014	0.0405
norm_rent_15_to_20_percent	0.0119	0.0059	0.0000	0.0220
norm_rent_10_to_15_percent	0.0092	0.0042	0.0000	0.0180
norm_rent_under_10_percent	0.0074	0.0091	0.0000	0.0483
norm_income_less_10000	0.0173	0.0083	0.0056	0.0541
norm_income_10000_14999	0.0117	0.0064	0.0000	0.0358
norm_income_15000_19999	0.0124	0.0050	0.0000	0.0270
norm_income_20000_24999	0.0139	0.0061	0.0000	0.0311
norm_income_25000_29999	0.0135	0.0044	0.0060	0.0270
norm_income_30000_34999	0.0158	0.0054	0.0000	0.0331
norm_income_35000_39999	0.0141	0.0073	0.0033	0.0433
norm_income_40000_44999	0.0130	0.0048	0.0000	0.0328
norm_income_45000_49999	0.0121	0.0044	0.0000	0.0247
norm_income_50000_59999	0.0250	0.0086	0.0050	0.0548
norm_income_60000_74999	0.0376	0.0083	0.0225	0.0630
norm_income_75000_99999	0.0535	0.0242	0.0317	0.1892
norm_income_100000_124999	0.0368	0.0119	0.0000	0.0814
norm_income_125000_149999	0.0221	0.0091	0.0000	0.0465
norm_income_150000_199999	0.0255	0.0093	0.0000	0.0484
norm_income_200000_or_more	0.0251	0.0114	0.0079	0.0473
cluster	1.0000	0.0000	1.0000	1.0000

Table 4: Summary Statistics for Cluster 2

Feature	Mean	SD	Min	Max
norm_confirmed_cases	0.0744	0.0227	0.0287	0.1530
norm_deaths	0.0018	0.0008	0.0000	0.0042
total_pop	119723.9083	498915.8423	289.0000	4525519.0000
median_income	51207.8833	4389.6252	44601.0000	62500.0000
income_per_capita	25493.1333	2888.8052	17960.0000	35680.0000
norm_rent_burden_not_computed	0.0184	0.0174	0.0069	0.1834
norm_rent_over_50_percent	0.0161	0.0090	0.0000	0.0443
norm_rent_40_to_50_percent	0.0067	0.0038	0.0000	0.0182
norm_rent_35_to_40_percent	0.0045	0.0031	0.0000	0.0117
norm_rent_30_to_35_percent	0.0064	0.0037	0.0000	0.0152
norm_rent_25_to_30_percent	0.0089	0.0046	0.0000	0.0201

norm_rent_20_to_25_percent	0.0100	0.0050	0.0000	0.0242
norm_rent_15_to_20_percent	0.0118	0.0056	0.0000	0.0309
norm_rent_10_to_15_percent	0.0098	0.0046	0.0000	0.0232
norm_rent_under_10_percent	0.0064	0.0050	0.0000	0.0381
norm_income_less_10000	0.0239	0.0077	0.0000	0.0430
norm_income_10000_14999	0.0200	0.0070	0.0035	0.0526
norm_income_15000_19999	0.0197	0.0066	0.0029	0.0424
norm_income_20000_24999	0.0209	0.0062	0.0069	0.0482
norm_income_25000_29999	0.0200	0.0056	0.0081	0.0453
norm_income_30000_34999	0.0182	0.0052	0.0029	0.0316
norm_income_35000_39999	0.0177	0.0060	0.0052	0.0484
norm_income_40000_44999	0.0178	0.0058	0.0000	0.0343
norm_income_45000_49999	0.0157	0.0049	0.0042	0.0382
norm_income_50000_59999	0.0309	0.0068	0.0109	0.0651
norm_income_60000_74999	0.0368	0.0073	0.0109	0.0737
norm_income_75000_99999	0.0434	0.0076	0.0111	0.0620
norm_income_100000_124999	0.0287	0.0069	0.0053	0.0509
norm_income_125000_149999	0.0153	0.0050	0.0019	0.0303
norm_income_150000_199999	0.0139	0.0045	0.0000	0.0245
norm_income_200000_or_more	0.0124	0.0054	0.0004	0.0346
cluster	2.0000	0.0000	2.0000	2.0000

Table 5: Summary Statistics for Cluster 3

Feature	Mean	SD	Min	Max
norm_confirmed_cases	0.0842	0.0308	0.0231	0.1829
norm_deaths	0.0023	0.0011	0.0008	0.0063
total_pop	45402.5161	130713.9228	564.0000	839539.0000
median_income	38958.1935	5354.3723	24794.0000	46696.0000
income_per_capita	20696.0860	3443.4739	12543.0000	30820.0000
norm_rent_burden_not_computed	0.0225	0.0148	0.0077	0.1223
norm_rent_over_50_percent	0.0175	0.0095	0.0037	0.0696
norm_rent_40_to_50_percent	0.0069	0.0043	0.0000	0.0186
norm_rent_35_to_40_percent	0.0053	0.0034	0.0000	0.0120
norm_rent_30_to_35_percent	0.0068	0.0041	0.0000	0.0208
norm_rent_25_to_30_percent	0.0091	0.0062	0.0000	0.0443
norm_rent_20_to_25_percent	0.0102	0.0069	0.0000	0.0376
norm_rent_15_to_20_percent	0.0104	0.0060	0.0000	0.0341
norm_rent_10_to_15_percent	0.0095	0.0052	0.0000	0.0287
norm_rent_under_10_percent	0.0059	0.0041	0.0000	0.0208
norm_income_less_10000	0.0360	0.0120	0.0132	0.0985
norm_income_10000_14999	0.0271	0.0093	0.0071	0.0599
norm_income_15000_19999	0.0275	0.0095	0.0076	0.0583
norm_income_20000_24999	0.0251	0.0083	0.0111	0.0617
norm_income_25000_29999	0.0225	0.0073	0.0071	0.0475
norm_income_30000_34999	0.0223	0.0079	0.0055	0.0577
norm_income_35000_39999	0.0193	0.0066	0.0015	0.0437
norm_income_40000_44999	0.0187	0.0070	0.0000	0.0426
norm_income_45000_49999	0.0144	0.0055	0.0000	0.0368

norm_income_50000_59999	0.0258	0.0079	0.0087	0.0522
norm_income_60000_74999	0.0322	0.0087	0.0142	0.0574
norm_income_75000_99999	0.0346	0.0098	0.0073	0.0702
norm_income_100000_124999	0.0195	0.0069	0.0035	0.0485
norm_income_125000_149999	0.0111	0.0043	0.0027	0.0288
norm_income_150000_199999	0.0087	0.0040	0.0000	0.0192
norm_income_200000_or_more	0.0074	0.0044	0.0000	0.0269
cluster	3.0000	0.0000	3.0000	3.0000

Perform cluster analysis using several methods (at least k-means and hierarchical clustering) using different feature subsets. Produce at least 4 different clusterings. [30]

How did you determine a suitable number of clusters for each method? [10]

Use unsupervised evaluation to describe and compare the clusterings and the clusters (some visual methods would be good). [10]

Identify a feature you could use as the ground truth to perform supervised evaluation. Compare the clusterings using this method. [10]

Evaluation

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?

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).