**Data and Visualization**

Team 10: Jeevan Rai & Abhilash Narayanan

**Executive Summary:**

This report is on analyzing various aspects of social structures that were collected during the COVID-19 pandemic. These social structures are referred as variables for the dataset COVID-19\_cases\_plus\_census that go through various data mining processes to discover any possible relationships with the number of cases and/or deaths for various geographic regions. There could be many possible reasons that could depict the observed confirmed cases and/or deaths for a particular region during the pandemic. Discovering such insightful relationships and building an effective predictive model are the two primary problems described in the report. From the general public to leaders of various government and non-government institutions, the information described in the report can be beneficial in many ways.

**Table of Contents**

[1 Business Understanding 2](#_Toc181012335)

[2 Data Preparation 3](#_Toc181012336)

[2.1 Statement 3](#_Toc181012337)

[2.2 Statement 3](#_Toc181012338)

[2.3 Statement 3](#_Toc181012339)

[3 Modeling 3](#_Toc181012340)

[3.1 Statement 3](#_Toc181012341)

[3.1.1 Statement 3](#_Toc181012342)

[3.1.2 Statement 3](#_Toc181012343)

[4 Evaluation 4](#_Toc181012344)

[5 Conclusion 4](#_Toc181012345)

[6 List of References 4](#_Toc181012346)

[7 Appendix 4](#_Toc181012347)

# Business Understanding

A widespread disease COVID-19 (also known as coronavirus disease 2019) started at the end of 2019, and it quickly spread throughout the entire world impacting every aspect of human society. It was first identified in December 2019 in Wuhan district in China. Since then, there have been several kinds of studies conducted on the impact of this pandemic. The data on those studies are available for the general public. Among those datasets, we will be looking at four different datasets. The primary focus of the analysis is based on understanding the impact of this pandemic on various aspects of human society all around the world from 2019. These policies were primarily implemented to enforce “social distancing” amongst people. Social distancing involves measures taken to reduce close contact between individuals to slow the spread of infectious diseases such as COVID-19. By implementing measures like social distancing, mask-wearing, and hygiene practices, the goal is to spread out the number of cases over a longer period, resulting in a flatter curve. Additionally, data mining is performed on these datasets to understand any existing relationships between entities. This can be used to predict recurrence of similar pandemic in the near future and the consequences of such situations.

# Data Preparation

## K-means Clustering

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Statistics** |  | **Feature** | **Statistics** |  | **Feature** | **Statistics** |
| county\_name | Length:222  Class :character  Mode :character |  | white\_pop | Min. : 55  1st Qu.: 2580  Median : 7632  Mean : 18563  3rd Qu.: 18465  Max. :162449 |  | female\_0\_20 | Min. : 4.0  1st Qu.: 771.5  Median : 1880.5  Mean : 6437.0  3rd Qu.: 4783.8  Max. :159914.0 |
| male\_pop | Min. : 1  1st Qu.: 436  Median : 1077  Mean : 3619  3rd Qu.: 2544  Max. :107552 |  | black\_pop | Min. : 0.00  1st Qu.: 55.25  Median : 482.50  Mean : 3292.94  3rd Qu.: 2673.00  Max. :85901.00 |  | female\_21\_49 | Min. : 8.0  1st Qu.: 898.5  Median : 2184.0  Mean : 7743.7  3rd Qu.: 5924.0  Max. :164411.0 |
| female\_pop | Min. : 0.00  1st Qu.: 12.00  Median : 25.50  Mean : 77.73  3rd Qu.: 61.75  Max. :2018.00 |  | asian\_pop | Min. : 0.0  1st Qu.: 6.0  Median : 45.0  Mean : 556.8  3rd Qu.: 241.8  Max. :12857.0 |  | female\_50\_above | Min. : 23  1st Qu.: 1184  Median : 3010  Mean : 7664  3rd Qu.: 7389  Max. :136600 |
| hispanic\_pop | Min. : 12  1st Qu.: 1468  Median : 3700  Mean : 19338  3rd Qu.: 9827  Max. :770794 |  | amerindian\_pop | Min. : 0.0  1st Qu.: 5.0  Median : 30.5  Mean : 125.9  3rd Qu.: 113.5  Max. :2344.0 |  | unemployed\_pop | Min. : 0.0  1st Qu.: 121.2  Median : 382.0  Mean : 1197.7  3rd Qu.: 960.5  Max. :27566.0 |
| other\_race\_pop | Min. : 0.00  1st Qu.: 0.00  Median : 0.00  Mean : 31.38  3rd Qu.: 26.75  Max. :477.00 |  | median\_income | Min. :24794  1st Qu.:41537  Median :46412  Mean :47428  3rd Qu.:52295  Max. :80938 |  | employed\_pop | Min. : 39  1st Qu.: 2421  Median : 5354  Mean : 17677  3rd Qu.: 14372  Max. :341350 |
| income\_less\_50K | Min. : 9  1st Qu.: 1111  Median : 2740  Mean : 7589  3rd Qu.: 6524  Max. :148982 |  | income\_50K\_100K | Min. : 19.0  1st Qu.: 534.8  Median : 1427.5  Mean : 4176.5  3rd Qu.: 3627.0  Max. :74071.0 |  | commute | Min. : 66  1st Qu.: 4344  Median : 9925  Mean : 32754  3rd Qu.: 26827  Max. :647479 |
| income\_100K\_150K | Min. : 0.0  1st Qu.: 176.2  Median : 490.5  Mean : 1621.6  3rd Qu.: 1382.8  Max. :25925.0 |  | income\_150K\_more | Min. : 0.0  1st Qu.: 114.8  Median : 274.0  Mean : 984.0  3rd Qu.: 772.2  Max. :14222.0 |  | worked\_at\_home | Min. : 0.0  1st Qu.: 68.5  Median : 190.0  Mean : 598.5  3rd Qu.: 515.0  Max. :15026.0 |
| rent\_under\_50\_percent | Min. : 7.0  1st Qu.: 289.5  Median : 820.0  Mean : 3317.0  3rd Qu.: 2171.0  Max. :73181.0 |  | rent\_over\_50\_percent | Min. : 0.0  1st Qu.: 60.5  Median : 185.0  Mean : 1027.0  3rd Qu.: 599.0  Max. :19775.0 |  | walked\_to\_work | Min. : 0.0  1st Qu.: 47.0  Median : 101.5  Mean : 338.6  3rd Qu.: 242.2  Max. :6964.0 |
| median\_age | Min. :25.80  1st Qu.:34.67  Median :39.20  Mean :39.31  3rd Qu.:43.27  Max. :57.50 |  | male\_0\_20 | Min. : 3  1st Qu.: 837  Median : 2079  Mean : 6785  3rd Qu.: 5274  Max. :163853 |  | male\_21\_49 | Min. : 10  1st Qu.: 971  Median : 2600  Mean : 8189  3rd Qu.: 6300  Max. :167493 |
|  |  |  | male\_50\_above | Min. : 29  1st Qu.: 1112  Median : 2830  Mean : 6771  3rd Qu.: 6769  Max. :109200 |  |  |  |

Statistics of dataset for k-means clustering

Shown above are all the features that are used to perform k-means clustering. The table also shows basic statistics of these features. Only the data points from Texas counties (excluding counties that belong to Dallas Fort Worth, Austin, San Antonio, and Houston) were selected for the dataset. These features were chosen in project 1 to understand possible correlations with COVID-19 confirmed and death cases.

## K-means Clustering (Race & Income Ranges)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Statistics** |  | **Feature** | **Statistics** |  | **Feature** | **Statistics** |
| white\_pop | Min. :-0.642387  1st Qu.:-0.554748  Median :-0.379400  Mean : 0.000000  3rd Qu.:-0.003394  Max. : 4.994073 |  | black\_pop | Min. :-0.36496  1st Qu.:-0.35884  Median :-0.31149  Mean : 0.00000  3rd Qu.:-0.06871  Max. : 9.15558 |  | asian\_pop | Min. :-0.3314  1st Qu.:-0.3278  Median :-0.3046  Mean : 0.0000  3rd Qu.:-0.1875  Max. : 7.3199 |
| hispanic\_pop | Min. :-0.2510  1st Qu.:-0.2321  Median :-0.2031  Mean : 0.0000  3rd Qu.:-0.1235  Max. : 9.7590 |  | amerindian\_pop | Min. :-0.47791  1st Qu.:-0.45894  Median :-0.36216  Mean : 0.00000  3rd Qu.:-0.04717  Max. : 8.41787 |  | other\_race\_pop | Min. :-0.43877  1st Qu.:-0.43877  Median :-0.43877  Mean : 0.00000  3rd Qu.:-0.06472  Max. : 6.23116 |
| income\_less\_50K | Min. :-0.44296  1st Qu.:-0.37855  Median :-0.28334  Mean : 0.00000  3rd Qu.:-0.06225  Max. : 8.26280 |  | income\_50K\_100K | Min. :-0.48025  1st Qu.:-0.42067  Median :-0.31755  Mean : 0.00000  3rd Qu.:-0.06348  Max. : 8.07375 |  | income\_100K\_150K | Min. :-0.49239  1st Qu.:-0.43887  Median :-0.34345  Mean : 0.00000  3rd Qu.:-0.07254  Max. : 7.37936 |
| income\_150K\_more | Min. :-0.4780  1st Qu.:-0.4222  Median :-0.3449  Mean : 0.0000  3rd Qu.:-0.1028  Max. : 6.4301 |  |  |  |  |  |  |

Shown above are the features related to various races and income ranges that were used to perform k-means clustering. The objects that were chosen to cluster were the counties in Texas (excluding counties that belong to DFW metropolitan, Austin, San Antonio, and Houston). Each county is represented by a row, and the clustering is based on features including the number of people in various racial groups and income ranges. These features were chosen because we are interested in understanding the clustering of people living in Texas counties based on their racial demographics and income levels. By clustering based on these features, we aim to identify patterns and groupings that can provide insights into the socio-economic and racial composition of different regions within these counties. This can help in policy making, resource allocation, and understanding demographic trends.

The dataset was scaled to ensure that all of the datapoints lie within similar range. All of the features are on ratio scale because they represent counts of people from various counties that fall under these features. Since we are using k-means clustering, we are using Euclidean distance for similarity/distance. This method was chosen because it tends to work well with a centroid-based approach (i.e. k-means clustering).

## First K-means Clustering

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Statistics** |  | **Feature** | **Statistics** |  | **Feature** | **Statistics** |
| white\_pop | Min. :-0.642387  1st Qu.:-0.554748  Median :-0.379400  Mean : 0.000000  3rd Qu.:-0.003394  Max. : 4.994073 |  | income\_100K\_150K | Min. :-0.49239  1st Qu.:-0.43887  Median :-0.34345  Mean : 0.00000  3rd Qu.:-0.07254  Max. : 7.37936 |  | rent\_under\_50\_percent | Min. :-0.4068  1st Qu.:-0.3720  Median :-0.3069  Mean : 0.0000  3rd Qu.:-0.1408  Max. : 8.5854 |
| male\_21\_49 | Min. :-0.42670  1st Qu.:-0.37657  Median :-0.29161  Mean : 0.00000  3rd Qu.:-0.09854  Max. : 8.31051 |  | commute | Min. :-0.4412  1st Qu.:-0.3834  Median :-0.3081  Mean : 0.0000  3rd Qu.:-0.0800  Max. : 8.2967 |  |  |  |

Shown above are the features related to population of Hispanic men aged 21-49 making 100-150K, commuted to work, and spent more than 50% on rent. These are the features selected to perform first k-means clustering. The objects that were chosen to cluster were the counties in Texas (excluding counties that belong to DFW metropolitan, Austin, San Antonio, and Houston). We want to under the demographic of population that falls under these attributes.

The dataset was scaled to ensure that all of the datapoints lie within similar range. All of the features are on ratio scale because they represent counts of people from various counties that fall under these features. Since we are using k-means clustering, we are using Euclidean distance for similarity/distance. This method was chosen because it tends to work well with a centroid-based approach (i.e. k-means clustering).

## Second K-means Clustering

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Statistics** |  | **Feature** | **Statistics** |  | **Feature** | **Statistics** |
| hispanic\_pop | Min. :-0.2510  1st Qu.:-0.2321  Median :-0.2031  Mean : 0.0000  3rd Qu.:-0.1235  Max. : 9.7590 |  | income\_50K\_100K | Min. :-0.48025  1st Qu.:-0.42067  Median :-0.31755  Mean : 0.00000  3rd Qu.:-0.06348  Max. : 8.07375 |  | female\_21\_49 | Min. :-0.40366  1st Qu.:-0.35719  Median :-0.29011  Mean : 0.00000  3rd Qu.:-0.09495  Max. : 8.17512 |
| worked\_at\_home | Min. :-0.40128  1st Qu.:-0.35535  Median :-0.27388  Mean : 0.00000  3rd Qu.:-0.05596  Max. : 9.67399 |  |  |  |  |  |  |

Shown above are the features related to population of Hispanic women aged 21-49 making 50-100K and worked from home. These are the features selected to perform second k-means clustering. The objects that were chosen to cluster were the counties in Texas (excluding counties that belong to DFW metropolitan, Austin, San Antonio, and Houston). We want to under the demographic of population that falls under these attributes.

The dataset was scaled to ensure that all of the datapoints lie within similar range. All of the features are on ratio scale because they represent counts of people from various counties that fall under these features. Since we are using k-means clustering, we are using Euclidean distance for similarity/distance. This method was chosen because it tends to work well with a centroid-based approach (i.e. k-means clustering).

## Third K-means Clustering

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Statistics** |  | **Feature** | **Statistics** |  | **Feature** | **Statistics** |
| income\_100K\_150K | Min. :-0.49239  1st Qu.:-0.43887  Median :-0.34345  Mean : 0.00000  3rd Qu.:-0.07254  Max. : 7.37936 |  | rent\_under\_50\_percent | Min. :-0.4068  1st Qu.:-0.3720  Median :-0.3069  Mean : 0.0000  3rd Qu.:-0.1408  Max. : 8.5854 |  | male\_50\_above | Min. :-0.524613  1st Qu.:-0.440329  Median :-0.306636  Mean : 0.000000  3rd Qu.:-0.000192  Max. : 7.969696 |
| commute | Min. :-0.4412  1st Qu.:-0.3834  Median :-0.3081  Mean : 0.0000  3rd Qu.:-0.0800  Max. : 8.2967 |  |  |  |  |  |  |

Shown above are the features related to population of men aged over 50 making more than 100K, spent less than 50% on rent, and commuted to work These are the features selected to perform second k-means clustering. The objects that were chosen to cluster were the counties in Texas (excluding counties that belong to DFW metropolitan, Austin, San Antonio, and Houston). We want to under the demographic of population that falls under these attributes.

The dataset was scaled to ensure that all of the datapoints lie within similar range. All of the features are on ratio scale because they represent counts of people from various counties that fall under these features. Since we are using k-means clustering, we are using Euclidean distance for similarity/distance. This method was chosen because it tends to work well with a centroid-based approach (i.e. k-means clustering).

## Fourth K-means Clustering

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Statistics** |  | **Feature** | **Statistics** |  | **Feature** | **Statistics** |
| hispanic\_pop | Min. :-0.2510  1st Qu.:-0.2321  Median :-0.2031  Mean : 0.0000  3rd Qu.:-0.1235  Max. : 9.7590 |  | income\_50K\_100K | Min. :-0.48025  1st Qu.:-0.42067  Median :-0.31755  Mean : 0.00000  3rd Qu.:-0.06348  Max. : 8.07375 |  | male\_21\_49 | Min. :-0.42670  1st Qu.:-0.37657  Median :-0.29161  Mean : 0.00000  3rd Qu.:-0.09854  Max. : 8.31051 |
| commute | Min. :-0.4412  1st Qu.:-0.3834  Median :-0.3081  Mean : 0.0000  3rd Qu.:-0.0800  Max. : 8.2967 |  | rent\_over\_50\_percent | Min. :-0.3775  1st Qu.:-0.3553  Median :-0.3095  Mean : 0.0000  3rd Qu.:-0.1573  Max. : 6.8913 |  |  |  |

Shown above are the features related to population of Hispanic men aged 21-49 making 50-100K, spent more than 50% on rent, and commuted to work These are the features selected to perform second k-means clustering. The objects that were chosen to cluster were the counties in Texas (excluding counties that belong to DFW metropolitan, Austin, San Antonio, and Houston). We want to under the demographic of population that falls under these attributes.

The dataset was scaled to ensure that all of the datapoints lie within similar range. All of the features are on ratio scale because they represent counts of people from various counties that fall under these features. Since we are using k-means clustering, we are using Euclidean distance for similarity/distance. This method was chosen because it tends to work well with a centroid-based approach (i.e. k-means clustering).

# Modeling

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Clustering** | **Features Selected** | **Unsupervised**  **Evaluation** | **Supervised Evaluation** | **Similarity/Distance method** |
| k-means clustering (first clustering) | white\_pop, income\_100K\_150K, rent\_under\_50\_percent, male\_21\_49, commute | Average Silhouette Width  Dunn Index  Pearson gamma  Within cluster sum of squares Elbow Method Gap Statistic | Purity  Entropy | Euclidean distance |
| k-means clustering (second clustering) | hispanic\_pop, income\_50K\_100K, female\_21\_49, worked\_at\_home | Average Silhouette Width  Dunn Index  Pearson gamma  Within cluster sum of squares Elbow Method Gap Statistic | Purity  Entropy | Euclidean distance |
| k-means clustering (third clustering) | income\_100K\_150K, rent\_under\_50\_percent, male\_50\_above, commute | Average Silhouette Width  Dunn Index  Pearson gamma  Within cluster sum of squares Elbow Method Gap Statistic | Purity  Entropy | Euclidean distance |
| k-means clustering (fourth clustering) | hispanic\_pop, income\_50K\_100K, male\_21\_49, commute, rent\_over\_50\_percent | Average Silhouette Width  Dunn Index  Pearson gamma  Within cluster sum of squares Elbow Method Gap Statistic | Purity  Entropy | Euclidean distance |

## K-means Clustering

### First k-means Clustering

### Second k-means Clustering

### Third k-means Clustering

### Fourth k-means Clustering

# Evaluation

[**Modeling and Evaluation**: What type of model do we apply to the data?

Describe why you chose the particular model, model assumption and limitations, what variable you use for the model, and how well the model works. ]

# Conclusion

[Does the project answer the initial questions? Repeat the key findings and why they are important.]

# List of References

USAFacts. (n.d.). COVID-19 case and death counts by state and county. In USAFacts Public Data - COVID-19 US Cases. Retrieved from

<https://console.cloud.google.com/bigquery?p=bigquery-public-data&d=covid19_usafacts&page=dataset&project=crucial-cycling-338005&ws=!1m4!1m3!3m2!1sbigquery-public-data!2scovid19_usafacts>

# Appendix