**Data and Visualization**

Team 10: Jeevan Rai & Abhilash Narayanan

**Table of Contents**

[1 Business Understanding 3](#_Toc182170813)

[2 Data Preparation 4](#_Toc182170814)

[2.1 K-means Clustering 6](#_Toc182170815)

[2.1.1 K-means Clustering (Race & Income Ranges) 6](#_Toc182170816)

[2.1.2 First K-means Clustering 7](#_Toc182170817)

[2.1.2 Second K-means Clustering 7](#_Toc182170818)

[2.1.3 Third K-means Clustering 8](#_Toc182170819)

[2.1.4 Fourth K-means Clustering 9](#_Toc182170820)

[3 Modeling 9](#_Toc182170821)

[3.1 K-means Clustering 10](#_Toc182170822)

[3.1.1 First k-means Clustering 10](#_Toc182170823)

[3.1.2 Second k-means Clustering 14](#_Toc182170824)

[3.1.3 Third k-means Clustering 18](#_Toc182170825)

[3.1.4 Fourth k-means Clustering 23](#_Toc182170826)

[4 Evaluation 29](#_Toc182170827)

[4.1 K-means clustering 29](#_Toc182170828)

[5 List of References 32](#_Toc182170829)

# 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.

This report tries to cluster various Texas counties (excluding counties from Austin, Dallas Fort Worth, San Antonio, and Houston) based on data available for various socio-economic factors such as income, age, working style, etc. Various clustering practices have been experimented to perform this cluster analysis. The report is primarily focused on providing various insights on statistics on COVID-19 confirmed cases and death rates to the Texas Department of State Health Services (DSHS). As the state-level health agency, DSHS is responsible for public health policy, disease surveillance, and health interventions in Texas. The analysis can provide insights into how socio-economic and health factors intersect to influence COVID-19 outcomes across different counties in Texas. This can help DSHS make more informed decisions about policy, resource allocation, and public health strategies. Using various clustering analysis techniques, we want to answer following questions:

* What counties does DSHS need to prioritize for targeted health interventions based on socio-economic and demographic factors?
* What factors (e.g., income, race, working style) contribute to higher vulnerability to COVID-19 cases and deaths in certain counties?
* Where are the counties, that need the highest state level assistance, located? Is there a certain pattern that can be observed on such demography based on various clustering techniques?

# Data Preparation

The dataset “COVID-19\_cases\_plus\_census” is extracted from USAFacts US Coronavirus Database (USAFacts).

The dataset has 259 feature columns and 3142 observations. These observations represent US COVID-19 cases and

death counts for all US states and counties. This is made available for the general public and is hosted in Google BigQuery. To maintain the focus of the analysis limited to specific aspects of the dataset, there were many feature columns removed or updated. For example, the variables providing information on the various income levels could be reduced to income ranges that would still provide enough insights on the financial status of the general public. Thus, the data of those variables were merged into four new income ranges: less than 50K, 50-100K, 100-150K, and 150K-above. This allowed us to drop the previous feature columns and reduce the dimension of the dataset. Following are some more feature processing performed on the dataset:

* 8 variables, containing information on certain percentage of income spent on rent, were merged into 2 new variables (i.e. rent\_under\_50\_percent and rent\_over\_50\_percent)
* variables containing various structures of families were dropped because they were less significant than people’s financial structures
* variables containing counts of males and females for various age groups were merged into new variables that contain counts of males and females for age groups 0-20, 21-49, and 50 above
* dropped variables that contained the counts of people belonging to multiple races and rather maintained emphasis on the counts of people belonging to only one specific race
* dropped variables that contained the counts of people holding various academic qualifications and rather maintained emphasis on income, race, working environment, gender and age
* grouped all the variables that contained the counts of people commuting using various methods and for certain time to work into simply a new group “commute” to have all the counts commuting to work

After all the feature processing completed, there were 3142 observations and 33 variables left in the dataset. The final dataset has no null values. There were also no duplicate observations found in the final dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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 all clusterings

Shown above are all the features that are used to perform all clusterings. 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

## 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 |

Table showing details on clusterings related to modeling

## K-means Clustering

### First k-means Clustering

Before performing k-means clustering with certain number of clusters, we calculated the Hopkins statistic. The value was 0.99 which is close to 1 indicating that the dataset had a strong clustering tendency. Additionally, we also created following plots to visualize distance matrix of the dataset. It’s a heatmap of the distances between observations in the dataset. These plots help us to understand the structure of the data which helps us to identify any natural groupings or patterns in the dataset. This gives us a round idea of how many clusters might be appropriate based on the distributions of the distances. Based on the factoextra plot, we can say that we might need 3 clusters in the k-means clustering for this dataset.

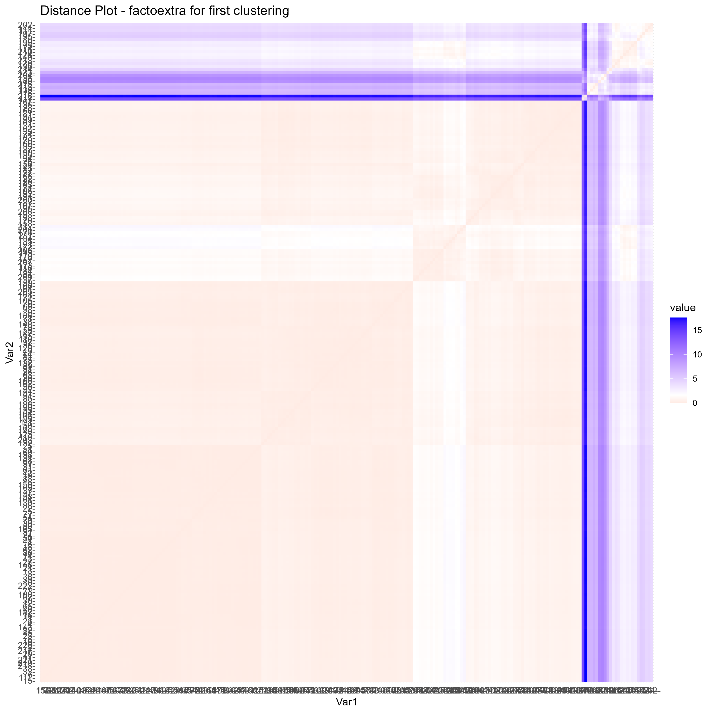


Fig. Distance matrix (left) and factoextra plot (right) for first clustering

**Determine suitable number of clusters**

**Average Silhouette Width**

To determine suitable number of clusters for the dataset, we started with analyzing the average silhouette width for various number of clusters. Based on the plot below, we can see that the highest value is at 2 clusters, and the value starts to decrease rapidly with a greater number of clusters. Thus, based on this assessment, we just need 2 clusters for the dataset.

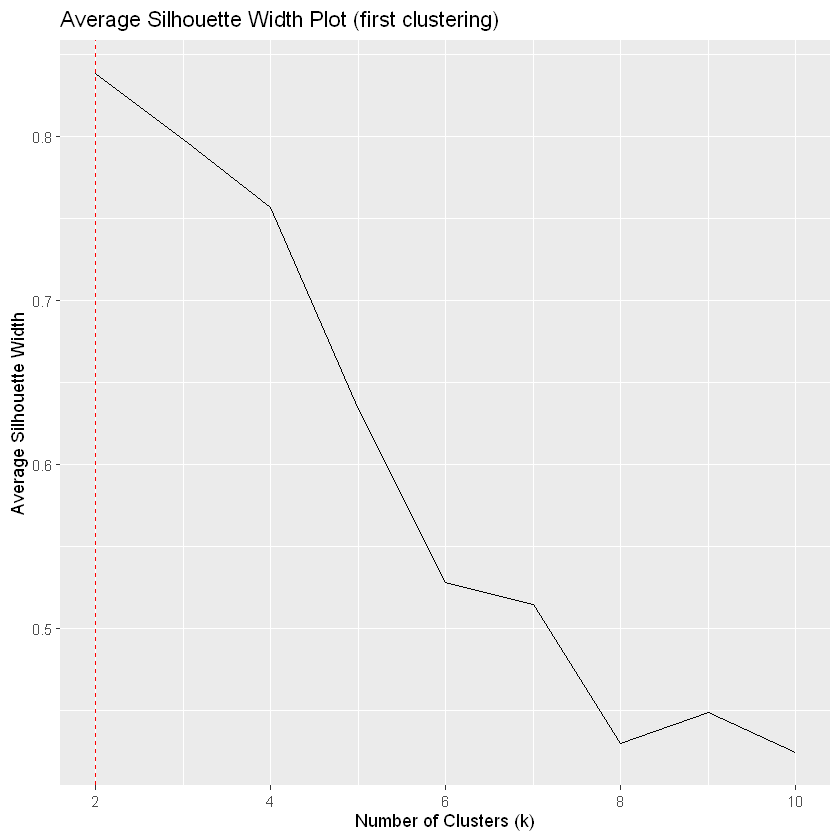


Fig. Average silhouette width plot for first clustering

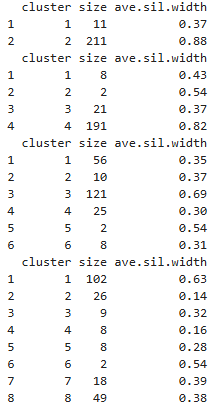
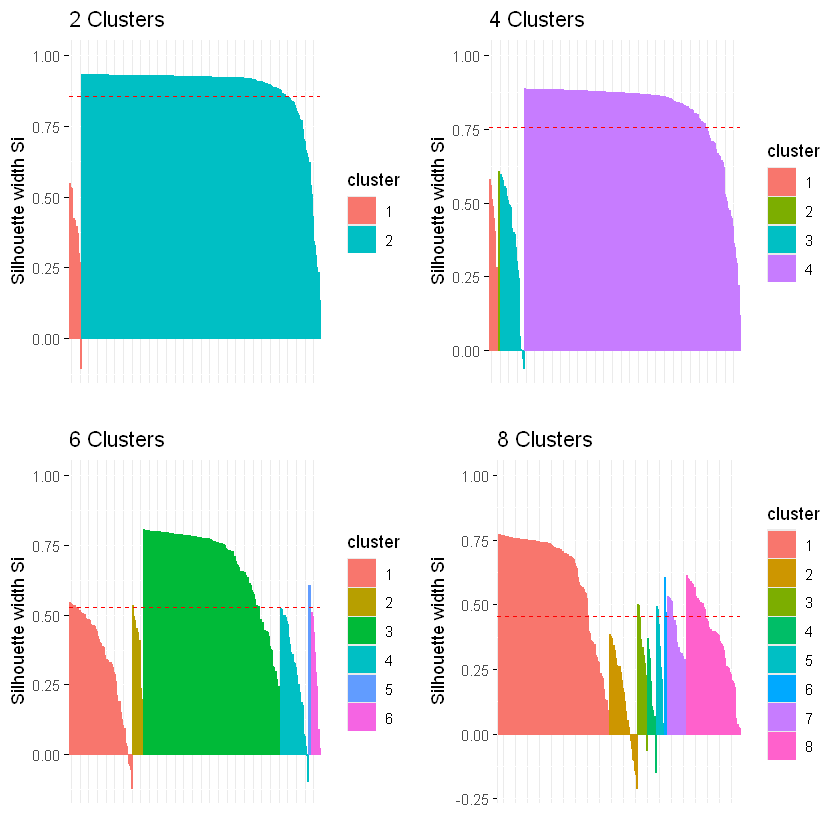


Fig. Silhouette plots for various number of clusters for first clustering

Next, we look at analyzing the silhouette widths and plots for various number of clusters. Based on the above plot, 4 clusters appear to be promising, with two clusters having reasonably high silhouette widths (0.54 and 0.82), but the other two clusters have lower silhouette widths, reducing the overall quality of clustering. Also, the first and second clusters have very few datapoints. 2 Clusters seems to be the optimal choice as it has one cluster with a very high silhouette width (0.88). Although the other cluster has a lower silhouette width (0.37), the simplicity of having fewer

clusters often provide more meaningful insights without overcomplicating the interpretation.

**Elbow Method: Within-Cluster Sum of Square**

Next, we used the Elbow method to visually identify the optimum number of clusters. Based on the elbow plot below, we can see that the within sum of squares value starts to decrease not so significantly after 5 clusters. Since the elbow point (or knee) indicates a point where adding more clusters does not significantly improve the compactness of the clusters, suggesting that the optimal number of clusters has been reached, we can say that the optimum number of clusters is 5.

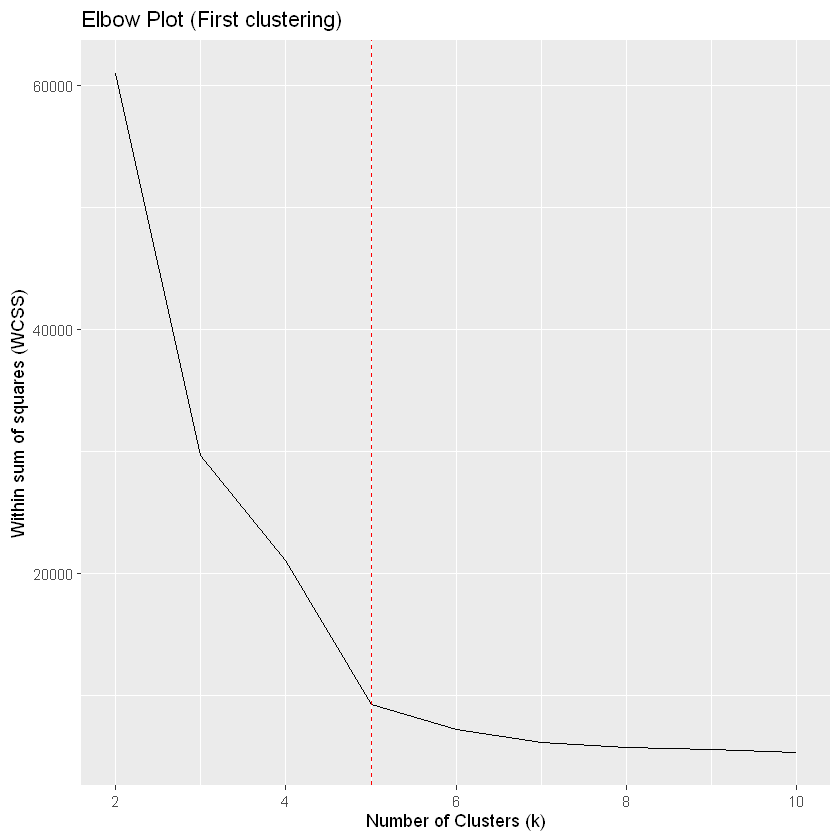


Fig. Elbow plot for first clustering

**Gap Statistic**

Next, we used gap statistic method to find optimum number of clusters. It compares the total within-cluster variation for different numbers of clusters with their expected values under null reference distributions of the data. Based on the plot below, we can see that the max gap value is 1.787 at 11 clusters, and it starts to decrease slightly after that. Thus, we can say that 11 is the optimum number of clusters based on this assessment.

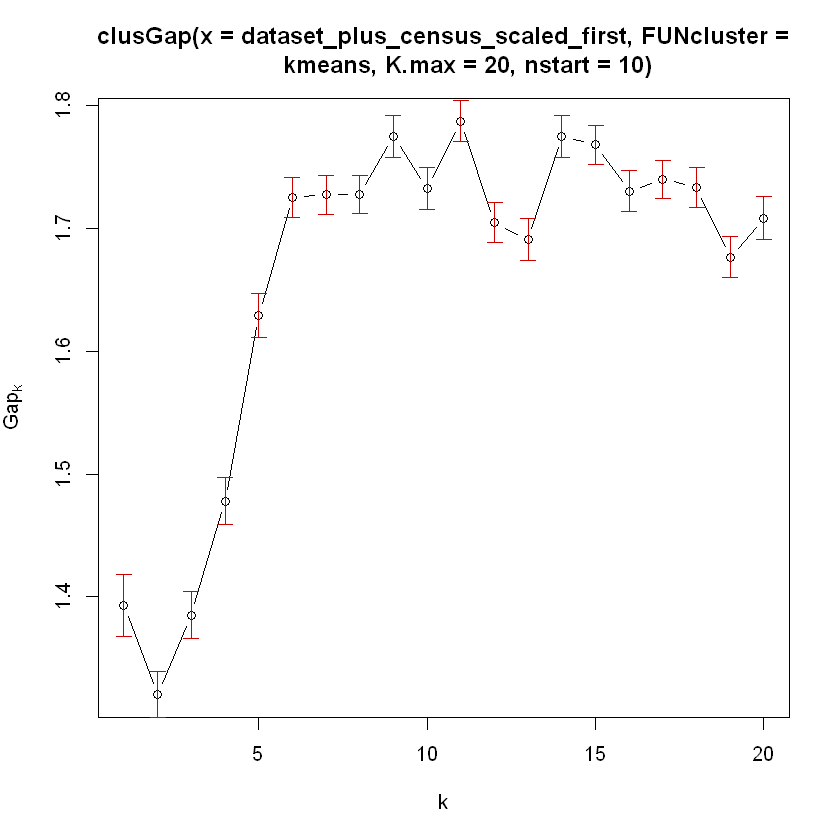
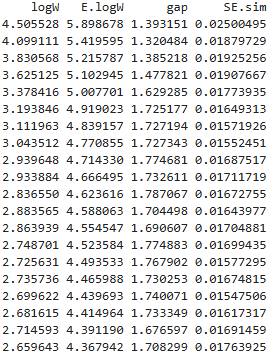
 

Fig. Gap statistic for first clustering

**Unsupervised Evaluation**

Next, we computed various metrics for different number of clusters as shown below. Looking at the sum of squares, we see that the value starts to decrease not so significantly after 5 clusters supporting the observation made on elbow method. Looking at the average silhouette width, we can see that it is max at 2 clusters. The pearson gamma values also indicate that the optimum number of clusters is 2. The dunn index also indicates that 2 clusters are needed. While increasing the number of clusters reduces the within-cluster sum of squares, it also leads to a decrease in the average silhouette width, pearson gamma, and dunn index, indicating lower clustering quality. Thus, we can finally say that the optimum number clusters for the dataset is 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **2 clusters** | **3 clusters** | **5 clusters** | **6 clusters** |
| within.cluster.ss | 397.8061 | 215.1592 | 82.48649 | 71.37958 |
| avg.silwidth | 0.838251 | 0.798405 | 0.633407 | 0.52843 |
| pearsongamma | 0.806838 | 0.771368 | 0.560038 | 0.399343 |
| dunn | 0.042062 | 0.035202 | 0.024113 | 0.014214 |

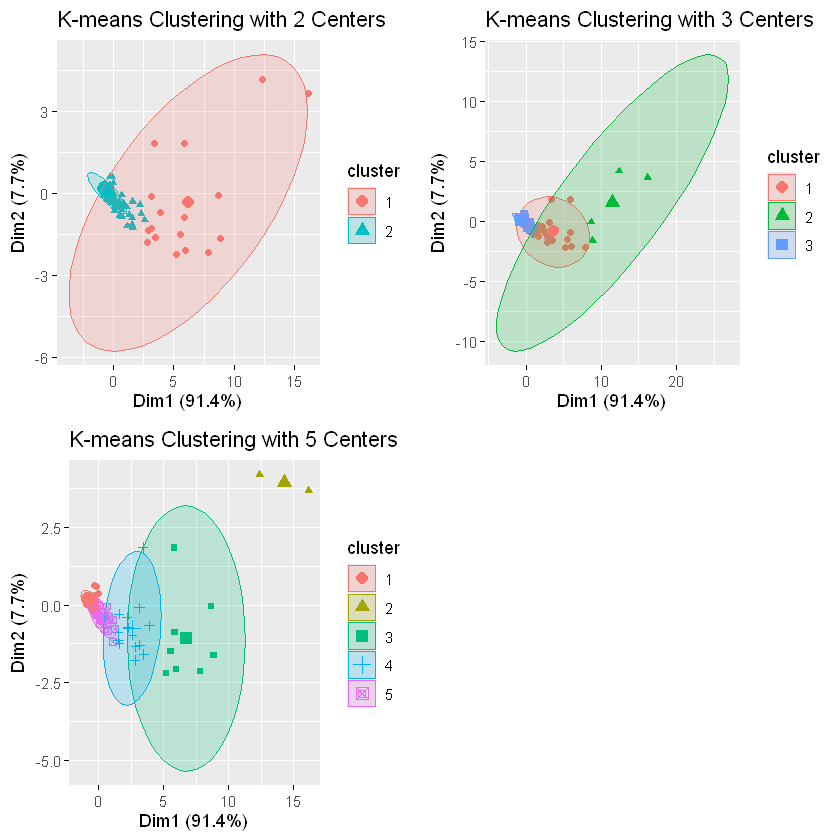
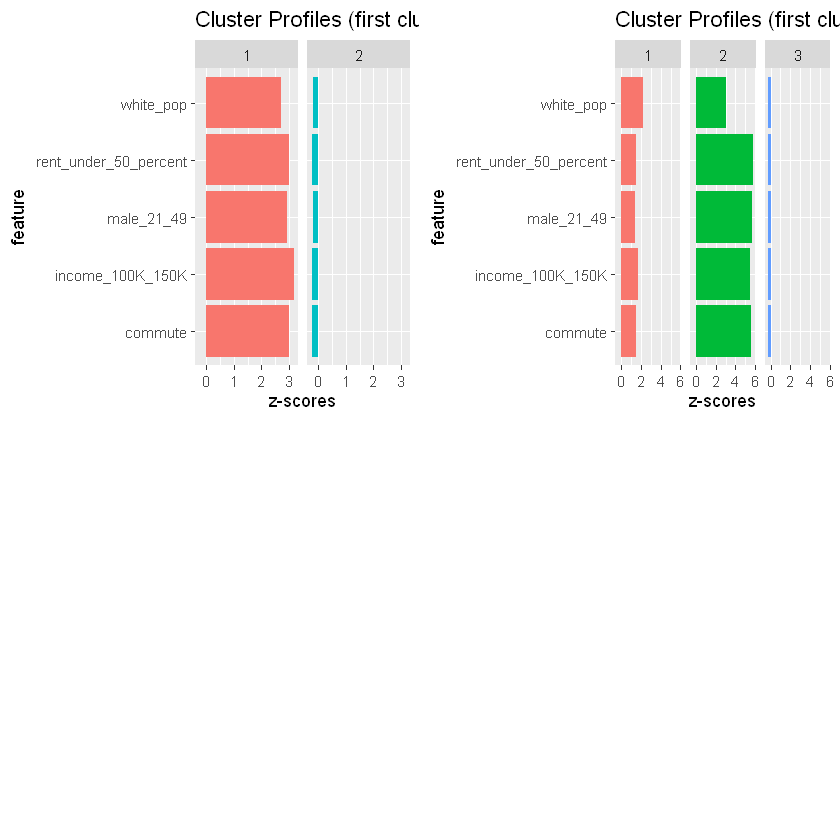
 

Fig. Visualization for first clustering

To further validate the above conclusion, we used visualization of the clustering as shown above. Looking at the clusters on the left, we can see that we only need two clusters. When an additional cluster is added, the clusters don’t properly represent the datapoints. Some of the datapoints that belong to cluster 1 can also be part of cluster 3. Similarly, looking at the plot on the right, we can see that the third cluster has almost same distribution for all the features indicating that it does identify any significant variance of the dataset. It also supports that there are only 2 clusters needed for this dataset. Therefore, based on all of the assessments, we can say that 2 is the optimum number of clusters for the dataset.

### Second k-means Clustering

Before performing k-means clustering with certain number of clusters, we calculated the Hopkins statistic. The value was 0.99 which is close to 1 indicating that the dataset had a strong clustering tendency. Additionally, we also created following plots to visualize distance matrix of the dataset. Based on the factoextra plot, we can say that we might need 2 clusters in the k-means clustering for this dataset.

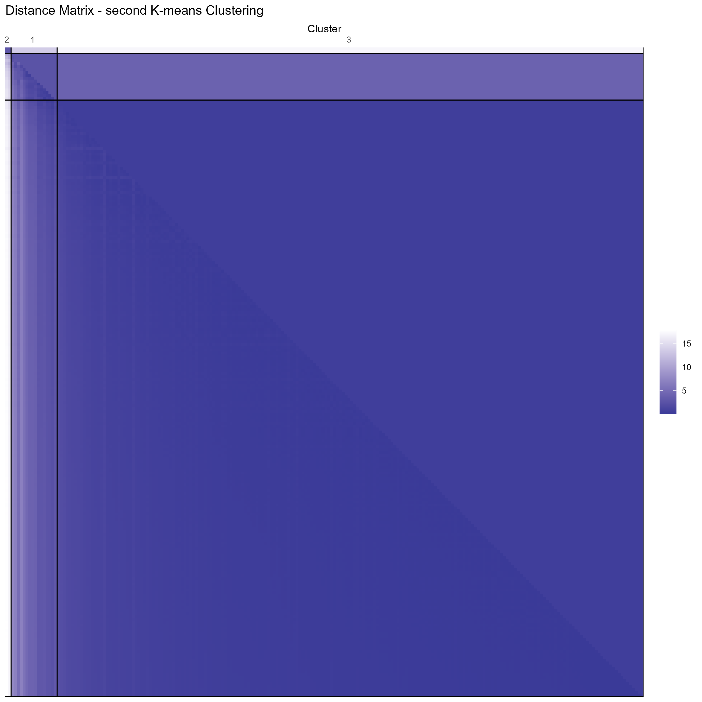
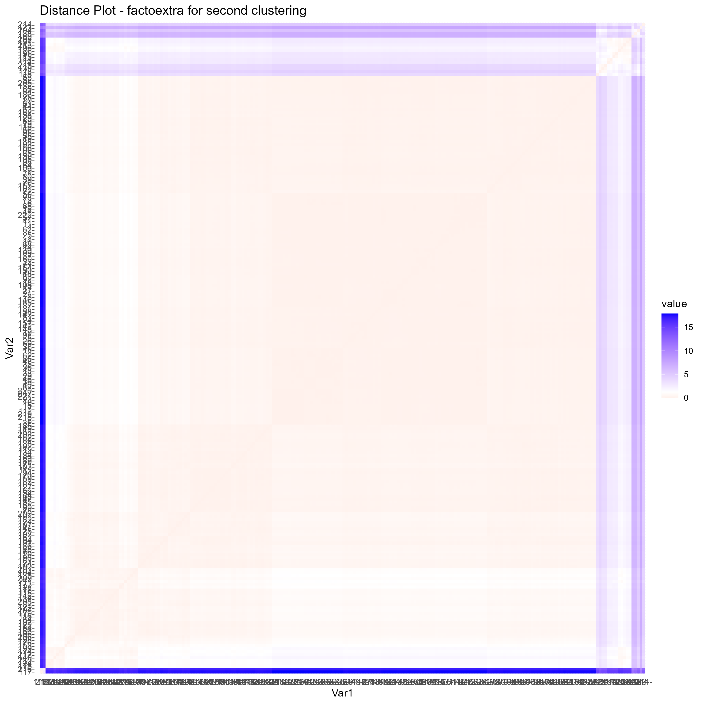
 

Fig. Distance matrix (left) and factoextra plot (right) for second clustering

**Determine suitable number of clusters**

**Average Silhouette Width**

To determine suitable number of clusters for the dataset, we started with analyzing the average silhouette width for various number of clusters. Based on the plot below, we can see that the highest value is at 2 clusters, and the value starts to decrease rapidly with a greater number of clusters. Thus, based on this assessment, we just need 2 clusters for the dataset.

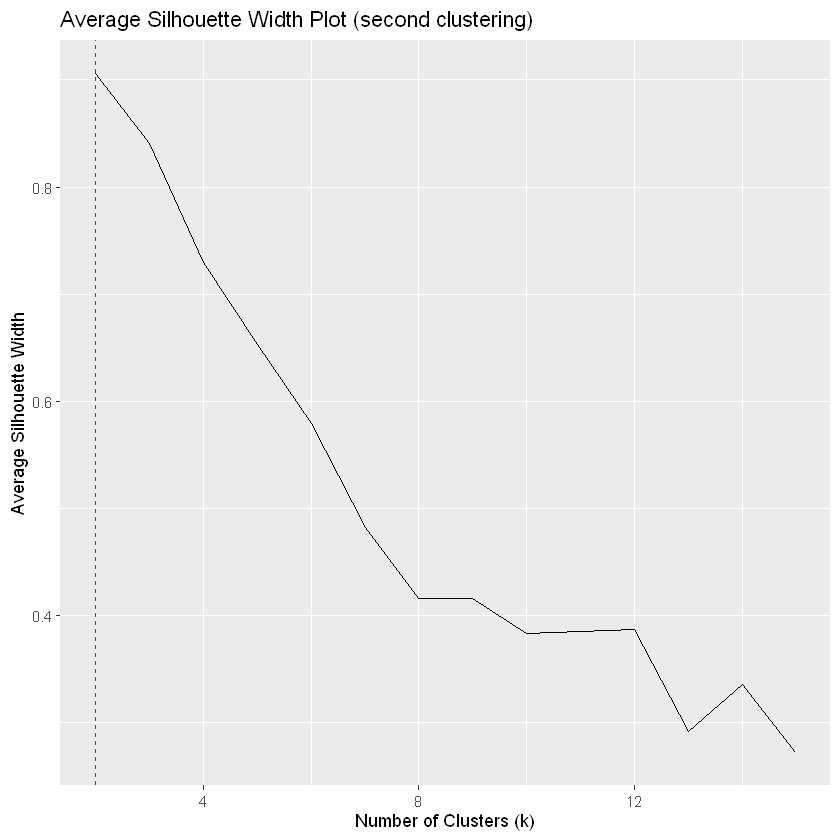


Fig. Average silhouette width plot for first clustering

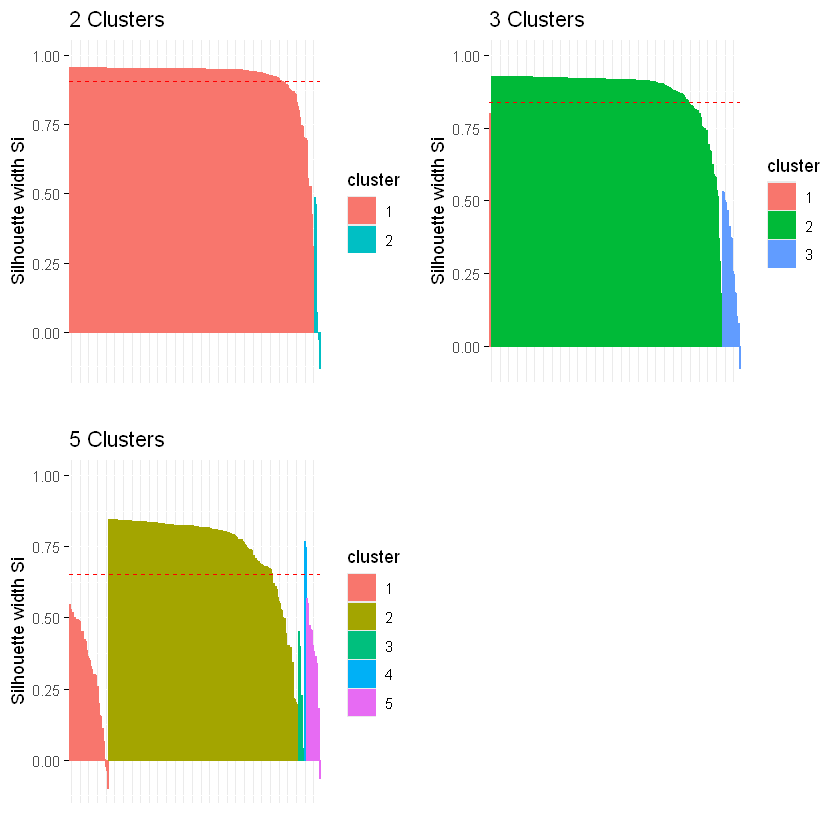
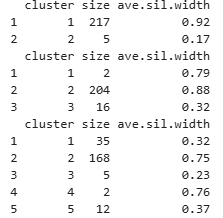
 

Fig. Silhouette plots for various number of clusters for second clustering

Next, we look at analyzing the silhouette widths and plots for various number of clusters. Considering both the

silhouette width and the distribution of cluster sizes, the clustering with 2 clusters seems to be the most appropriate

choice. Cluster 1 has a very high silhouette width (0.92), indicating well-separated clusters. While Cluster 2 is small,

the overall simplicity and high silhouette width for the main cluster make this a reasonable choice.

**Elbow Method: Within-Cluster Sum of Square**

Next, we used the Elbow method to visually identify the optimum number of clusters. Based on the elbow plot below, we can see that the within sum of squares value starts to decrease not so significantly after about 7 clusters. Since the elbow point (or knee) indicates a point where adding more clusters does not significantly improve the compactness of the clusters, suggesting that the optimal number of clusters has been reached, we can say that the optimum number of clusters is 7.

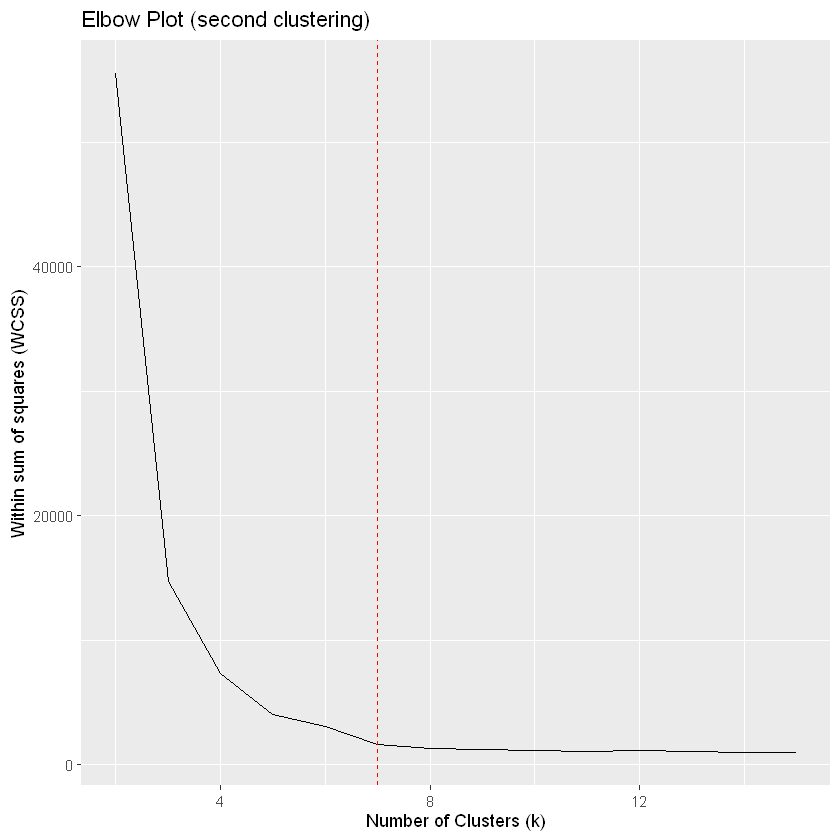


Fig. Elbow plot for first clustering

**Gap Statistic**

Next, we used gap statistic method to find optimum number of clusters. Based on the plot below, we can see that the max gap value is 2.09 at 17 clusters. The gap value appears to be decreasing after that. Thus, we can say that 17 is the optimum number of clusters based on this assessment.

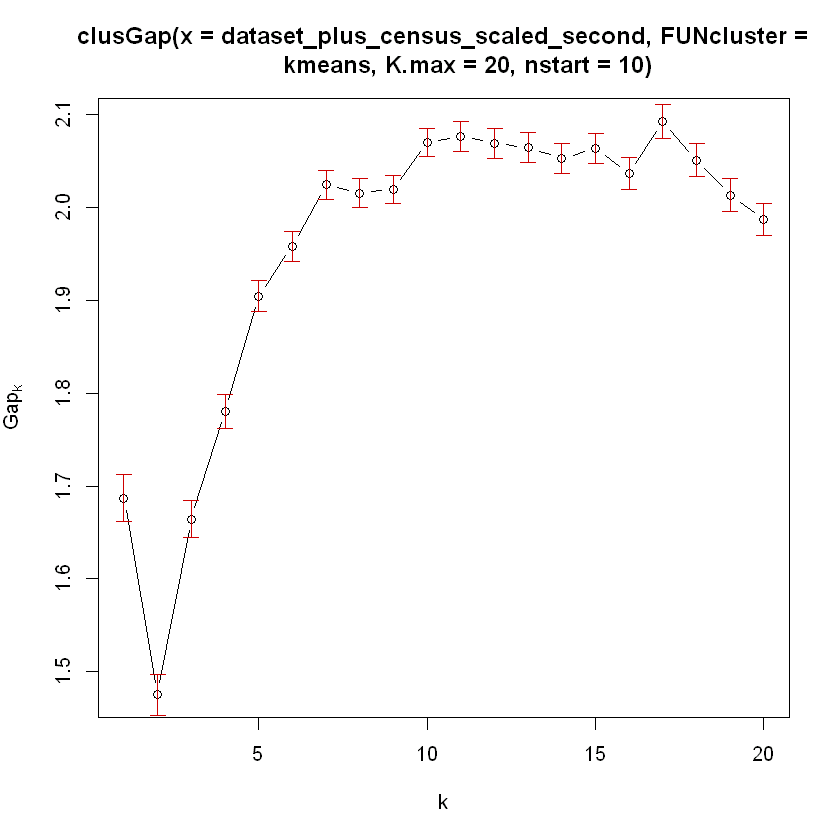
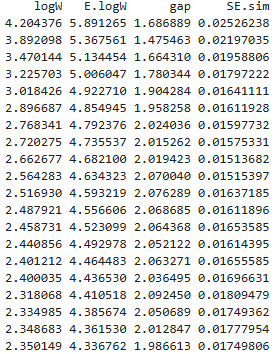
 

Fig. Gap statistic for first clustering

**Unsupervised Evaluation**

Next, we computed various metrics for different number of clusters as shown below. All of the of the metrics are the highest for 2 clusters indicating that the optimum number of clusters is 2 for this dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **2 clusters** | **3 clusters** | **4 clusters** | **6 clusters** |
| within.cluster.ss | 336.9456 | 103.8761 | 64.20614 | 38.06195 |
| avg.silwidth | 0.9055904 | 0.8404076 | 0.729722 | 0.579127 |
| pearsongamma | 0.8086453 | 0.7104225 | 0.5720946 | 0.4022496 |
| dunn | 0.1275271 | 0.07974147 | 0.03998538 | 0.01211304 |

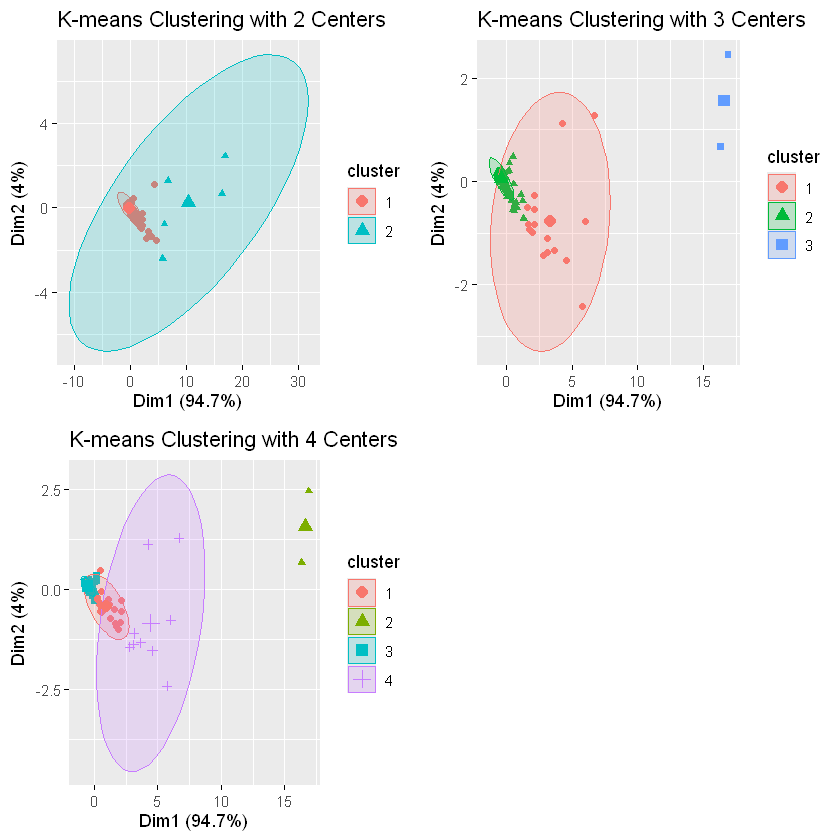
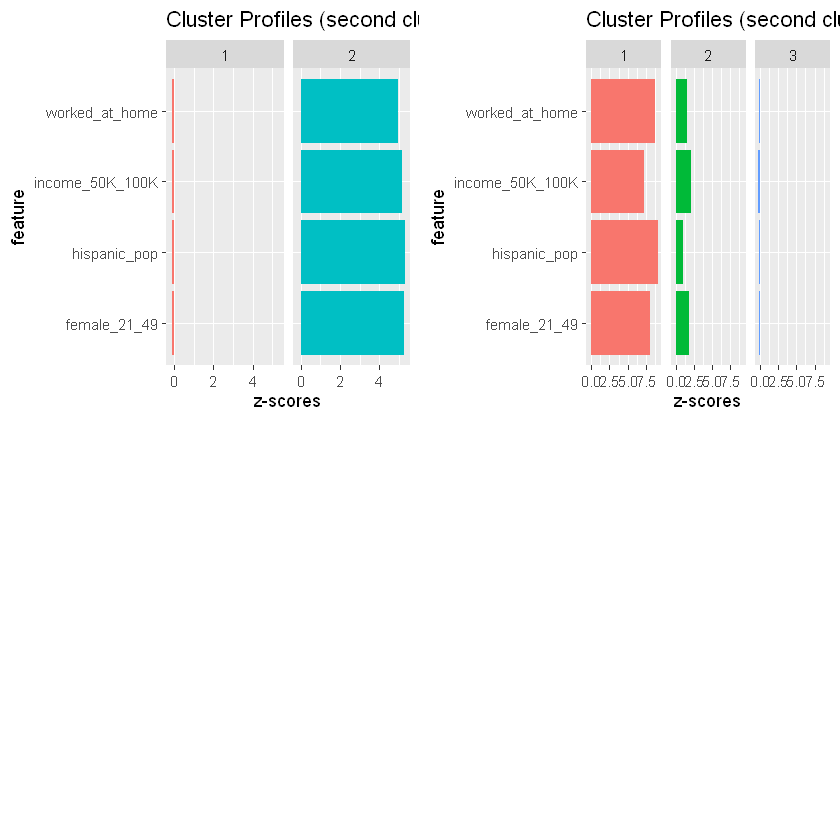
 

Fig. Visualization for first clustering

To further validate the above conclusion, we used visualization of the clustering as shown above. Looking at the clusters on the left, we can see that we only need two clusters. When an additional cluster is added, the clusters don’t properly represent the datapoints. The third cluster seems to cluster outliers. Similarly, looking at the plot on the right, we can see that the third cluster has almost same distribution for all the features indicating that it does identify any significant variance of the dataset. It also supports that there are only 2 clusters needed for this dataset. Therefore, based on all of the assessments, we can say that 2 is the optimum number of clusters for the dataset.

### Third k-means Clustering

Before performing k-means clustering with certain number of clusters, we calculated the Hopkins statistic. The value was 0.99 which is close to 1 indicating that the dataset had a strong clustering tendency. Additionally, we also created following plots to visualize distance matrix of the dataset. Based on the factoextra plot, we can say that we might need 3 clusters in the k-means clustering for this dataset.

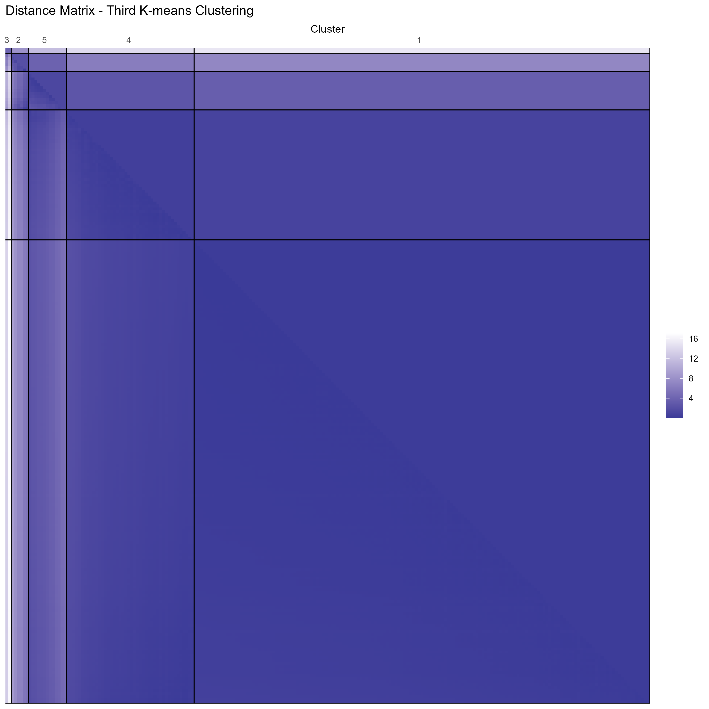
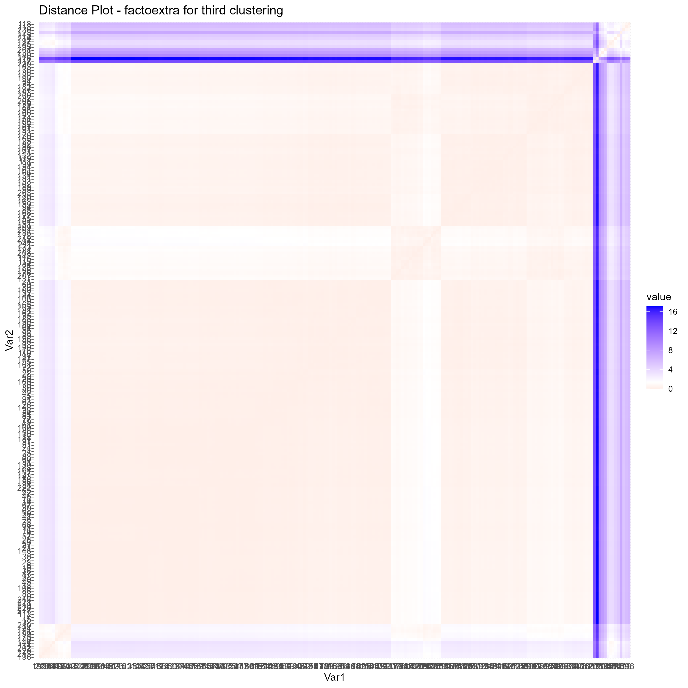
 

Fig. Distance matrix (left) and factoextra plot (right) for second clustering

**Determine suitable number of clusters**

**Average Silhouette Width**

To determine suitable number of clusters for the dataset, we started with analyzing the average silhouette width for various number of clusters. Based on the plot below, we can see that the highest value is at 2 clusters, and the value starts to decrease rapidly with a greater number of clusters. Thus, based on this assessment, we just need 2 clusters for the dataset.

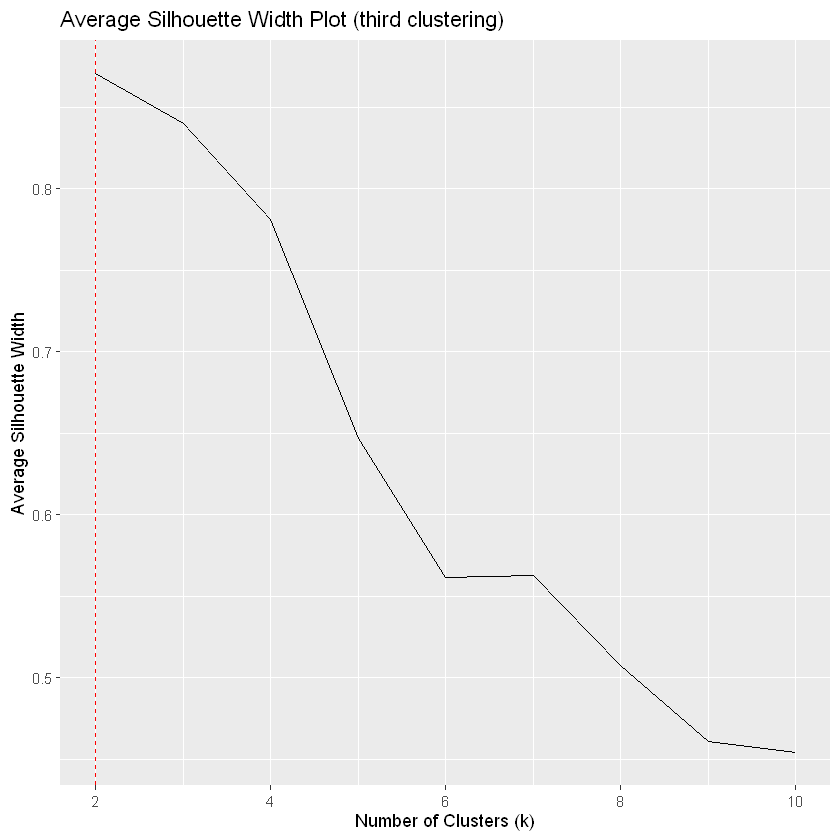


Fig. Average silhouette width plot for first clustering

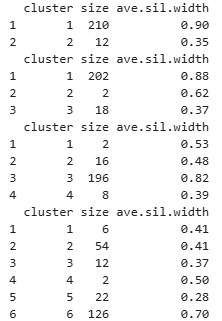
 

Fig. Silhouette plots for various number of clusters for third clustering

Next, we look at analyzing the silhouette widths and plots for various number of clusters. For the clustering with 2

clusters, the average silhouette width for Cluster 1 is very high (0.90), indicating well-separated and compact

clusters. Cluster 2 has a significantly lower silhouette width (0.35), suggesting that it may not be as well-separated or compact. However, the overall silhouette width is strongly influenced by Cluster 1's high value. As we add more

clusters, we can see that the added clusters only cluster very few datapoints. At the same time, the average

silhouette widths also don’t improve much. The second cluster of the clustering with clusters, although smaller and with a lower silhouette width, does not significantly detract from the overall quality of the clustering compared to

higher cluster numbers with more uneven and lower silhouette widths. Therefore, 2 clusters appear to be the

optimum choice.

**Elbow Method: Within-Cluster Sum of Square**

Next, we used the Elbow method to visually identify the optimum number of clusters. Based on the elbow plot below, we can see that the within sum of squares value starts to decrease not so significantly after about 7 clusters. Since the elbow point (or knee) indicates a point where adding more clusters does not significantly improve the compactness of the clusters, suggesting that the optimal number of clusters has been reached, we can say that the optimum number of clusters is 7.



Fig. Elbow plot for third clustering

**Gap Statistic**

Next, we used gap statistic method to find optimum number of clusters. Based on the plot below, we can see that the max gap value is 1.85 at 13 clusters. After that, the gap value either stabilizes or slightly decreases. Thus, we can say that the optimum number of clusters is 13 based on this assessment.

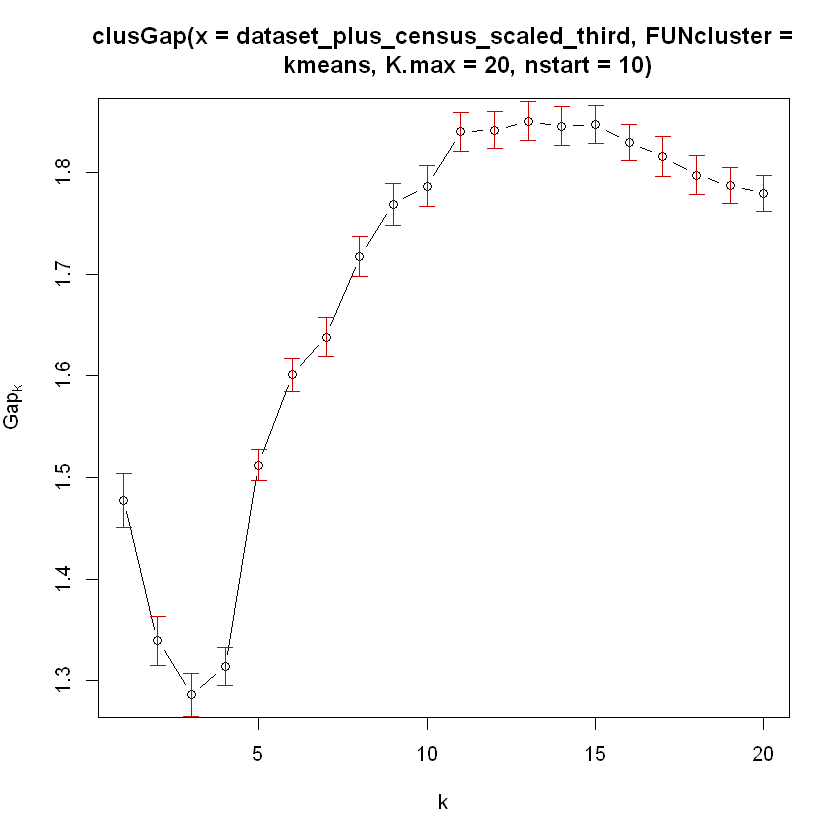
 

Fig. Gap statistic for third clustering

**Unsupervised Evaluation**

Next, we computed various metrics for different number of clusters as shown below. 2 clusters have the highest average silhouette width and Pearson gamma, indicating well-defined and compact clusters. However, 3 clusters have the highest Dunn index, suggesting better separation between clusters. Given the strong performance in both silhouette width and Pearson gamma, 2 clusters appear to be the optimal number of clusters. We further validated this determination by visually analyzing the clustering with various clusters.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **2 clusters** | **3 clusters** | **5 clusters** | **6 clusters** |
| within.cluster.ss | 283.8065 | 129.9854 | 46.36307 | 39.78291 |
| avg.silwidth | 0.8703733 | 0.8400603 | 0.6465164 | 0.5617945 |
| pearsongamma | 0.8150192 | 0.7699686 | 0.4882054 | 0.3815488 |
| dunn | 0.04766757 | 0.08112688 | 0.01734194 | 0.007537748 |

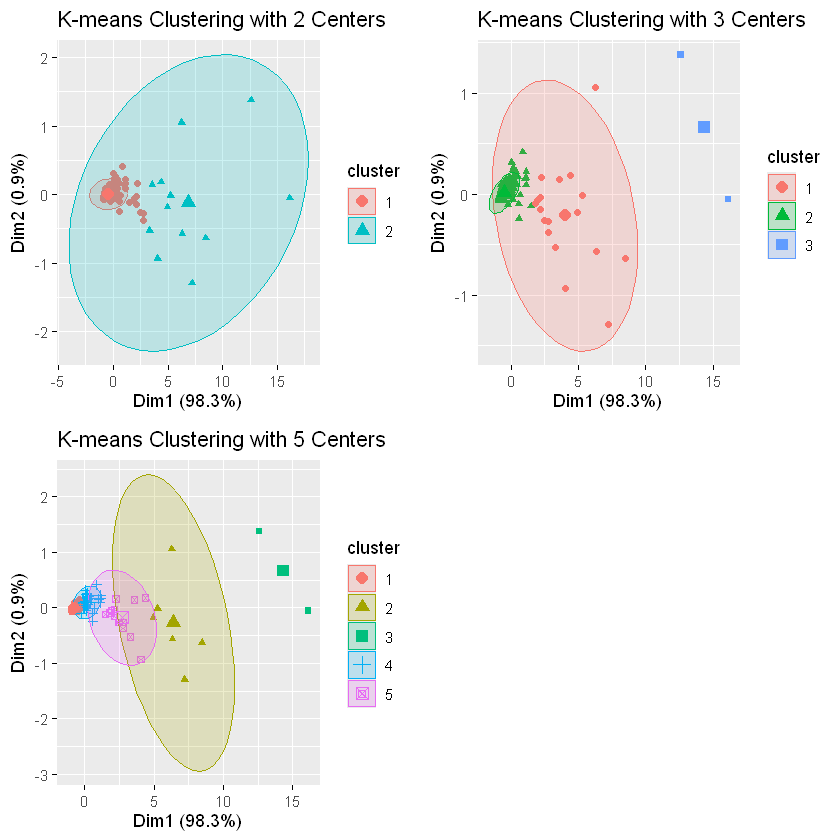
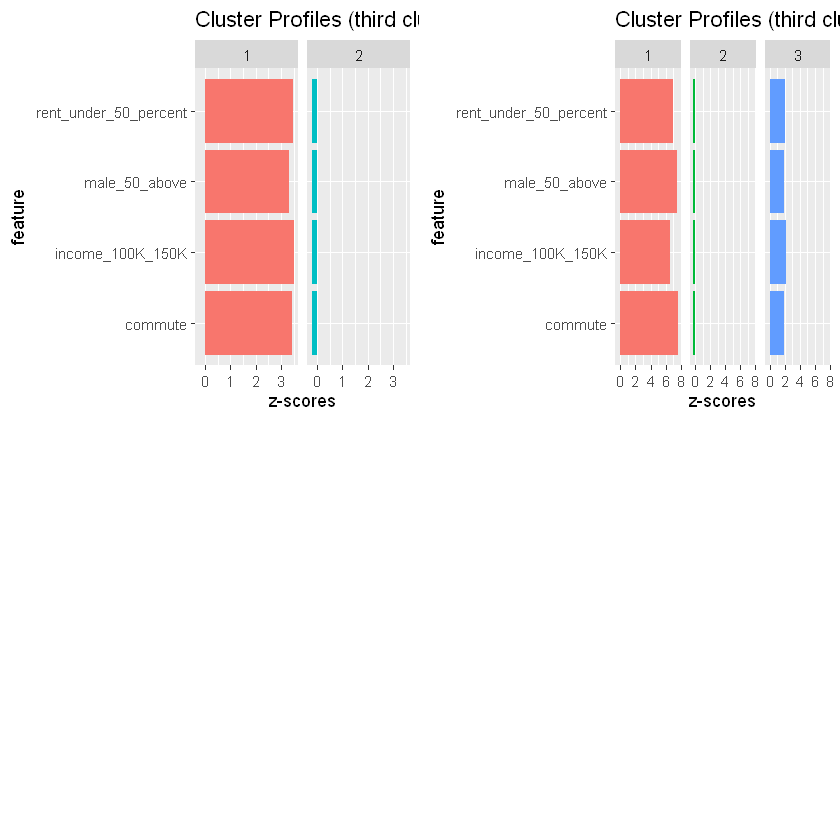
 

Fig. Visualization for third clustering

To further validate the above conclusion, we used visualization of the clustering as shown above. Looking at the clusters on the left, we can see that we only need two clusters. When an additional cluster is added, the clusters don’t properly represent the datapoints. The third cluster seems to cluster outliers. Similarly, looking at the plot on the right, we can see that the second cluster has almost same distribution for all the features indicating that it does identify any significant variance of the dataset. It also supports that there are only 2 clusters needed for this dataset. Therefore, based on all of the assessments, we can say that 2 is the optimum number of clusters for the dataset.

### Fourth k-means Clustering

Before performing k-means clustering with certain number of clusters, we calculated the Hopkins statistic. The value was 0.99 which is close to 1 indicating that the dataset had a strong clustering tendency. Additionally, we also created following plots to visualize distance matrix of the dataset. Based on the factoextra plot, we can say that we might need 3 clusters in the k-means clustering for this dataset.

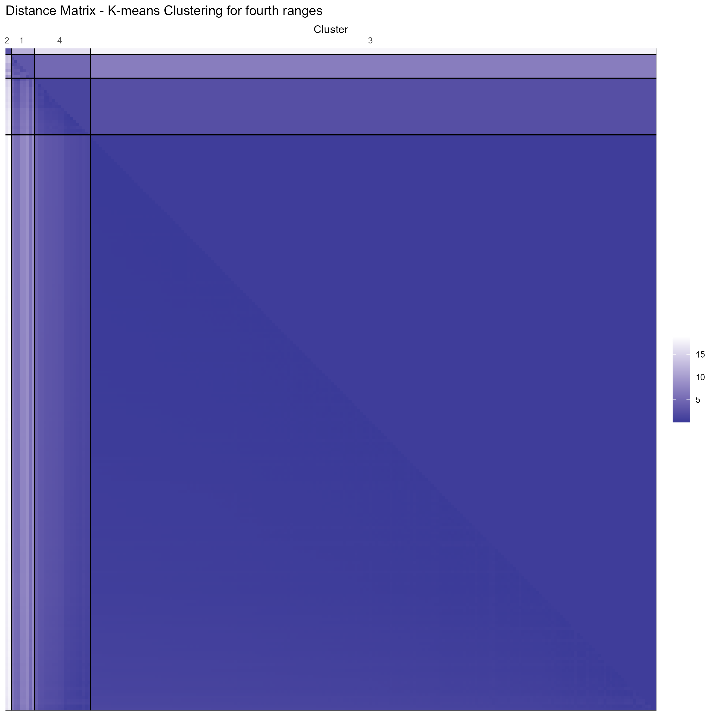
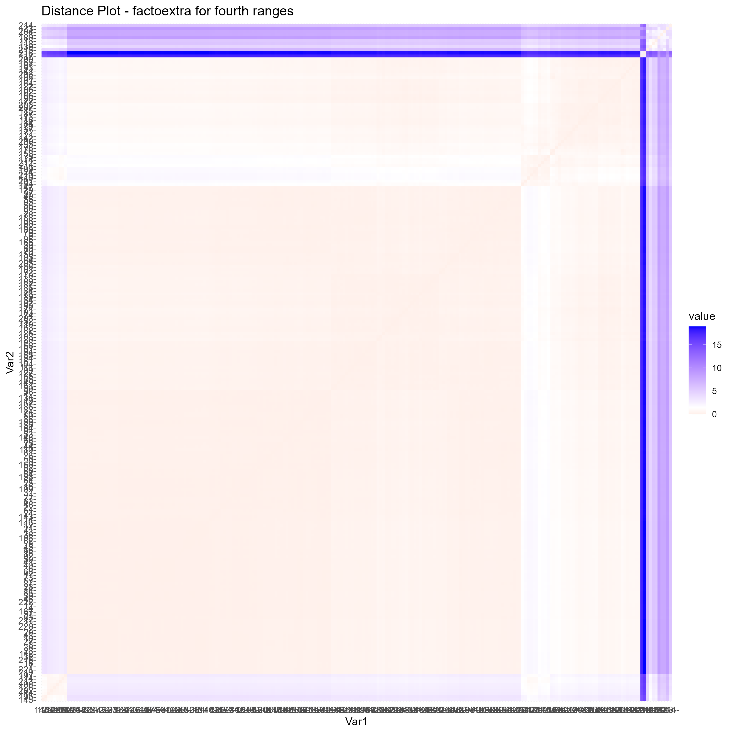
 

Fig. Distance matrix (left) and factoextra plot (right) for fourth clustering

**Determine suitable number of clusters**

**Average Silhouette Width**

To determine suitable number of clusters for the dataset, we started with analyzing the average silhouette width for various number of clusters. Based on the plot below, we can see that the highest value is at 2 clusters, and the value starts to decrease rapidly with a greater number of clusters. Thus, based on this assessment, we just need 2 clusters for the dataset.

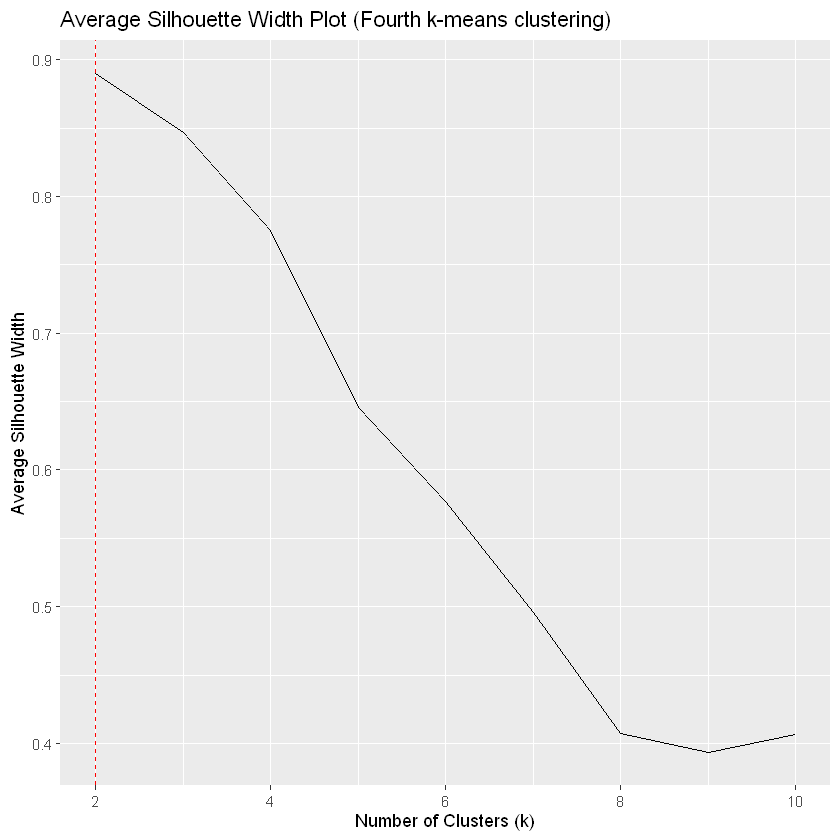


Fig. Average silhouette width plot for first clustering

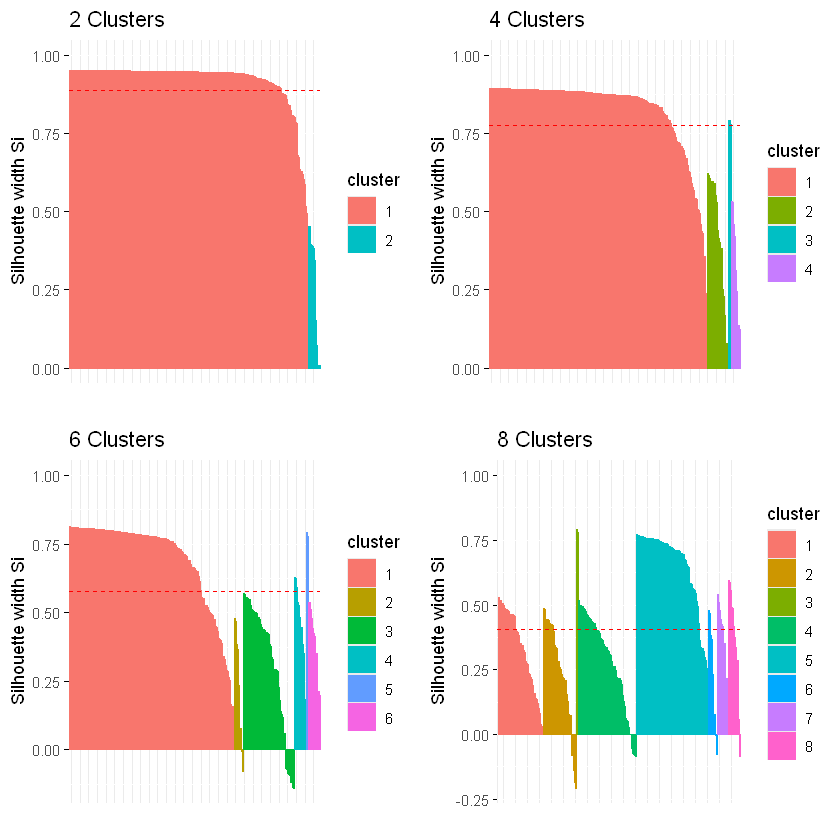
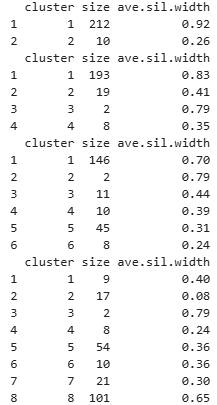
 

Fig. Silhouette plots for various number of clusters for fourth clustering

Next, we look at analyzing the silhouette widths and plots for various number of clusters. For the clustering with 2

clusters, 2 clusters show a very high silhouette width for one cluster but a very low value for the other, suggesting that one cluster is well-defined while the other is not. 4 clusters present a more balanced solution with reasonably high silhouette widths for three clusters (0.83, 0.41, 0.79) and a slightly lower but still reasonable value for one cluster (0.35). However, clusters 3 and 4 of this clustering have very few datapoints. 6 clusters and 8 clusters show more mixed results, with some clusters having low silhouette widths, indicating less well-defined clusters. Therefore, the

optimum number of clusters could not be identified with high confidence using this assessment.

**Elbow Method: Within-Cluster Sum of Square**

Next, we used the Elbow method to visually identify the optimum number of clusters. Based on the elbow plot below, we can see that the within sum of squares value starts to decrease not so significantly after about 4 clusters. Since the elbow point (or knee) indicates a point where adding more clusters does not significantly improve the compactness of the clusters, suggesting that the optimal number of clusters has been reached, we can say that the optimum number of clusters is 4.

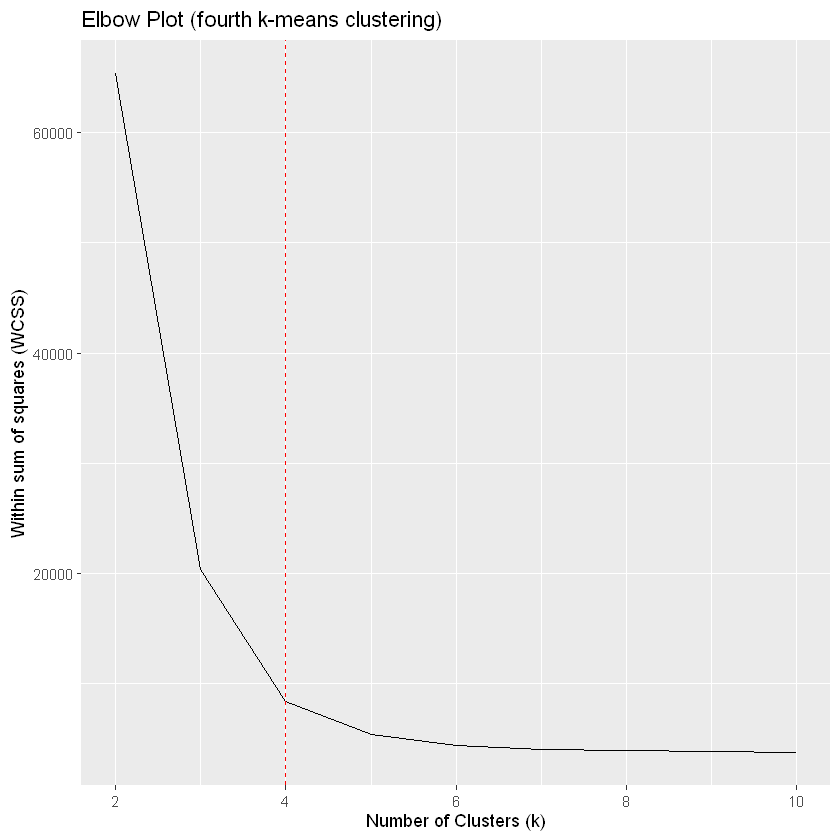


Fig. Elbow plot for third clustering

**Gap Statistic**

Next, we used gap statistic method to find optimum number of clusters. Based on the plot below, we can see that the max gap value is 2.07 at 8 clusters. After that, the gap value either stabilizes or slightly decreases. Thus, we can say that the optimum number of clusters is 8 based on this assessment.

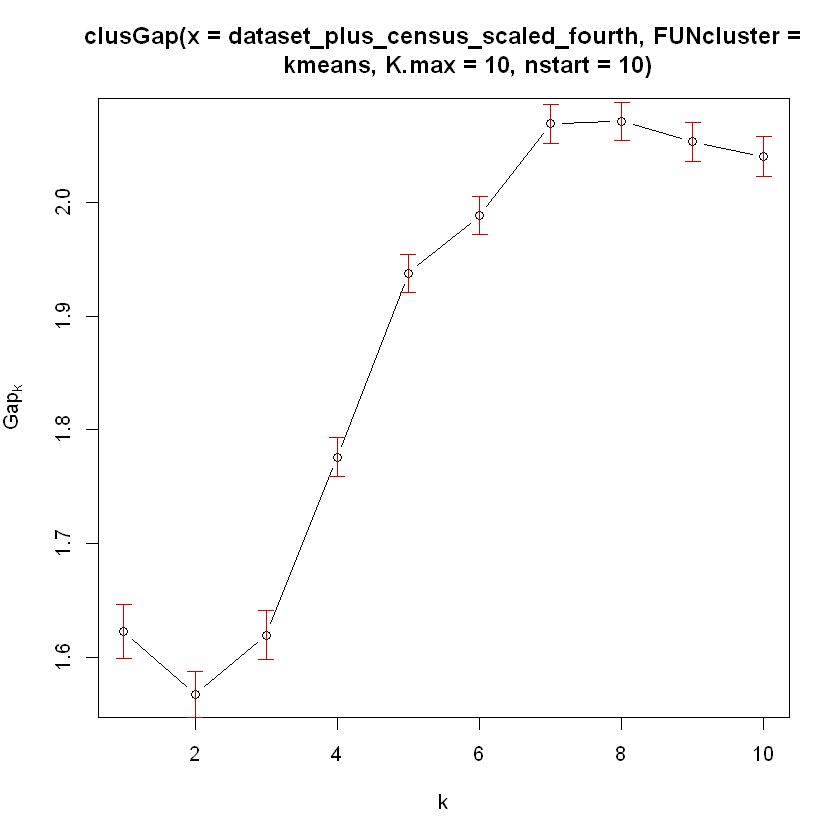
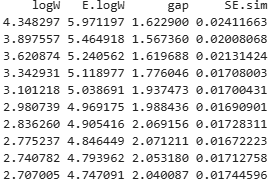
 

Fig. Gap statistic for fourth clustering

**Unsupervised Evaluation**

Next, we computed various metrics for different number of clusters as shown below. Based on the values of various metrics, we can clearly see that all of those values are largest for 2 clusters. Therefore, the optimum number of clusters is 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **2 clusters** | **4 clusters** | **8 clusters** | **10 clusters** |
| within.cluster.ss | 380.2993 | 76.5326 | 50.10439 | 49.19994 |
| avg.silwidth | 0.8896608 | 0.7753996 | 0.4071046 | 0.363984 |
| pearsongamma | 0.8246153 | 0.655894 | 0.2225344 | 0.1924984 |
| dunn | 0.08043983 | 0.05822636 | 0.004683524 | 0.00467794 |

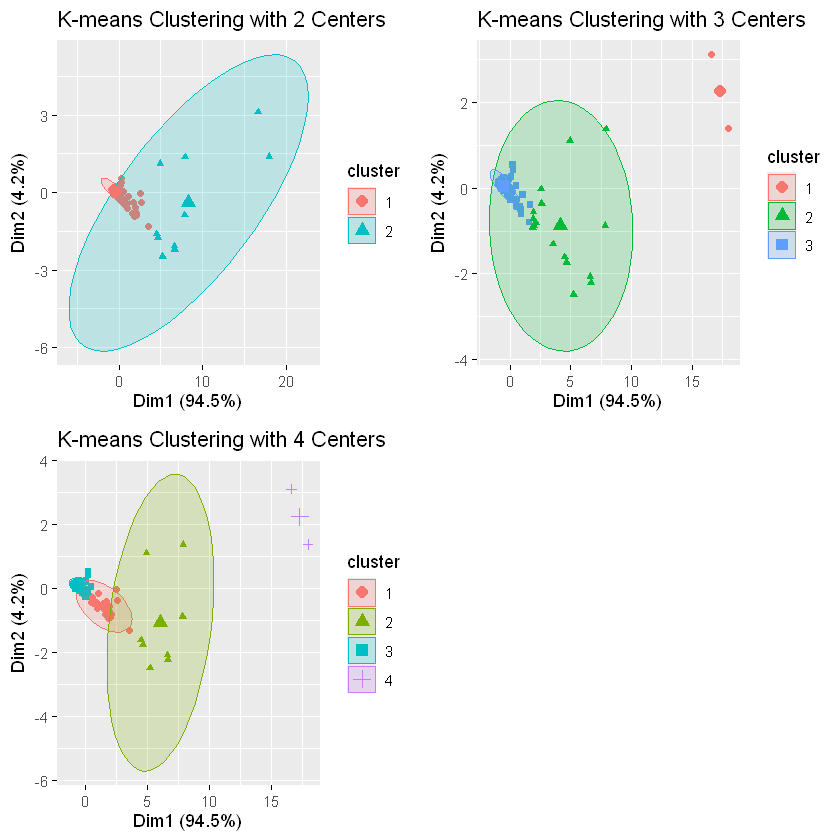
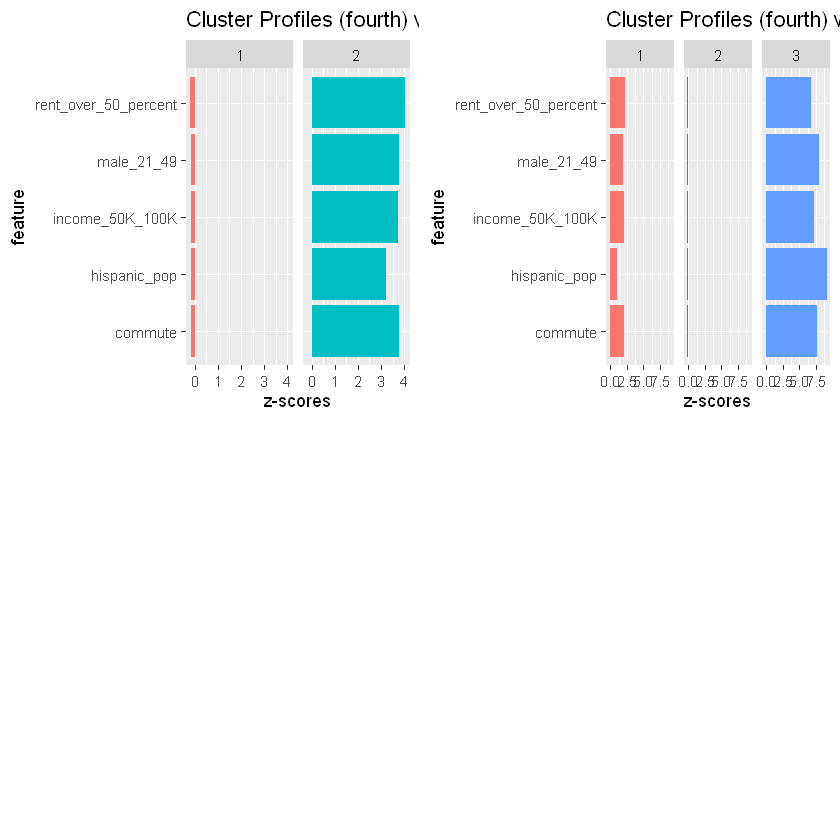
 

Fig. Visualization for fourth clustering

To further validate the above conclusion, we used visualization of the clustering as shown above. Looking at the clusters on the left, we can see that we only need two clusters. When an additional cluster is added, the clusters don’t properly represent the datapoints. The third cluster seems to cluster outliers. Similarly, looking at the plot on the right, we can see that the second cluster has almost same distribution for all the features indicating that it does identify any significant variance of the dataset. It also supports that there are only 2 clusters needed for this dataset. Therefore, based on all of the assessments, we can say that 2 is the optimum number of clusters for the dataset.

**Supervised Evaluation (Compare all four k-means clusterings)**

To perform supervised evaluation, we chose “median\_income” variable as the ground truth. Since the minimum and maximum incomes are 24794 and 80938, we created three labels: low (0-50K), medium (51-75K), and high (76K above). We used purity score and entropy to evaluate how well different clustering align with the predefined income labels (i.e. ground truth). For each cluster of a clustering, the purity score is the proportion of counties in that cluster that are correctly classified into their majority ground truth label. So a higher purity score means that the clustering is more accurate in terms of assigning similar income counties to the same cluster. Similarly, the entropy function is used to measure the disorder or uncertainty in the clustering results with respect to the ground truth labels. It returns a value that quantifies how mixed or pure the clusters are with respect to the known labels. So low entropy score suggests that the clusters are mostly composed of counties with the same income group, which means that the k-means algorithm has effectively separated the counties based on income. Similarly, The Corrected Rand Index (corrected.rand) is a measure of the similarity between two clusterings. So, score of 1 means perfect agreement between the two clusterings (i.e. identical). Likewise, The Variation of Information (vi) is a measure of the difference between two clusterings. So, a smaller vi score indicates more similarity between the two clusterings.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **corrected.rand** | **vi** | **purity** | **entropy** |
| truth | 1 | 0 | 1 | 0 |
| k\_means\_first | 0.010545 | 0.901685 | 0.657658 | 0.460574 |
| k\_means\_second | 0.002179 | 0.819546 | 0.657658 | 0.073128 |
| k\_means\_third | 0.006321 | 0.914865 | 0.657658 | 0.472569 |
| k\_means\_fourth | -0.01869 | 0.891209 | 0.657658 | 0.127927 |

Based on the above table, the first clustering has the highest corrected Rand index, indicating it is the closest to the ground truth. The second clustering has the lowest VI score indicating that it is the closest to the ground truth. All four clusterings have the same purity, so this metric does not help distinguish between them. The second clustering has the lowest entropy indicating it is the closest to the ground truth. Based on the Variation of Information (VI) and Entropy, the second clustering appears to be the closest to the ground truth and thus performed the best clustering.

# Evaluation

### K-means clustering

**First K-means Clustering**

After establishing optimum number of clusters, we generated a heat map of Texas showing all the counties that belonged to one of the two clusters. There are 16 counties that belong to cluster 1, and the name of those counties are shown below. Most of these counties lie far away from the major cities like DFW, Austin, San Antonio, and Houston. We can also see that these counties had higher average confirmed cases per 1000 than the counties that belong to cluster 2. But surprisingly, these counties had lower average deaths per 1000 than those that belong to cluster 2. This could mean that the counties in cluster 1 more confirmed cases but there were effective interventions or better healthcare that kept the mortality rate lower. On the contrary, the counties in cluster 2 could have been less equipped to manage the severity of the virus, or other factors which contributed to higher death rate.

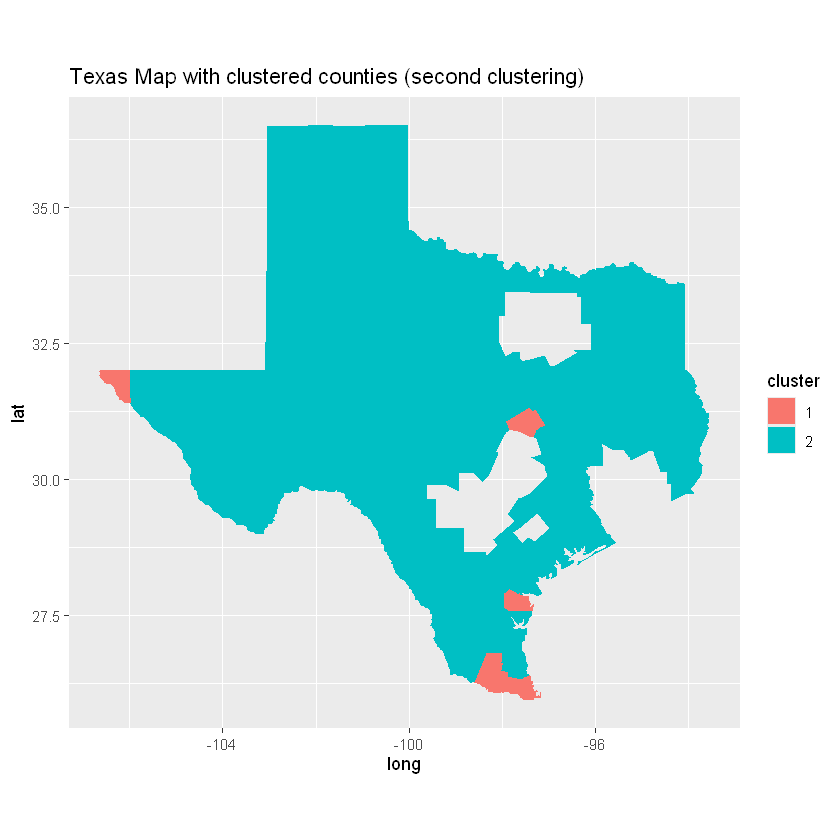
 

|  |  |  |
| --- | --- | --- |
| cluster | avg\_cases\_per\_1000 | avg\_deaths\_per\_1000 |
| 1 | 88.22725 | 1.657642 |
| 2 | 78.07511 | 2.009892 |

Fig. Clusters of counties for first clustering

**Second K-means Clustering**

After establishing optimum number of clusters, we generated a heat map of Texas showing all the counties that belonged to one of the two clusters. There are 5 counties that belong to cluster 1, and the name of those counties are shown below. Most of these counties lie far away at the edges of the state. We can also see that these counties had higher average confirmed cases per 1000 than the counties that belong to cluster 2. Additionally, these counties also had higher average deaths per 1000 than those that belong to cluster 2. This could mean that the counties in cluster 2 had less equipped to manage the severity of the virus, or other factors contributing to higher confirmed and death rates.



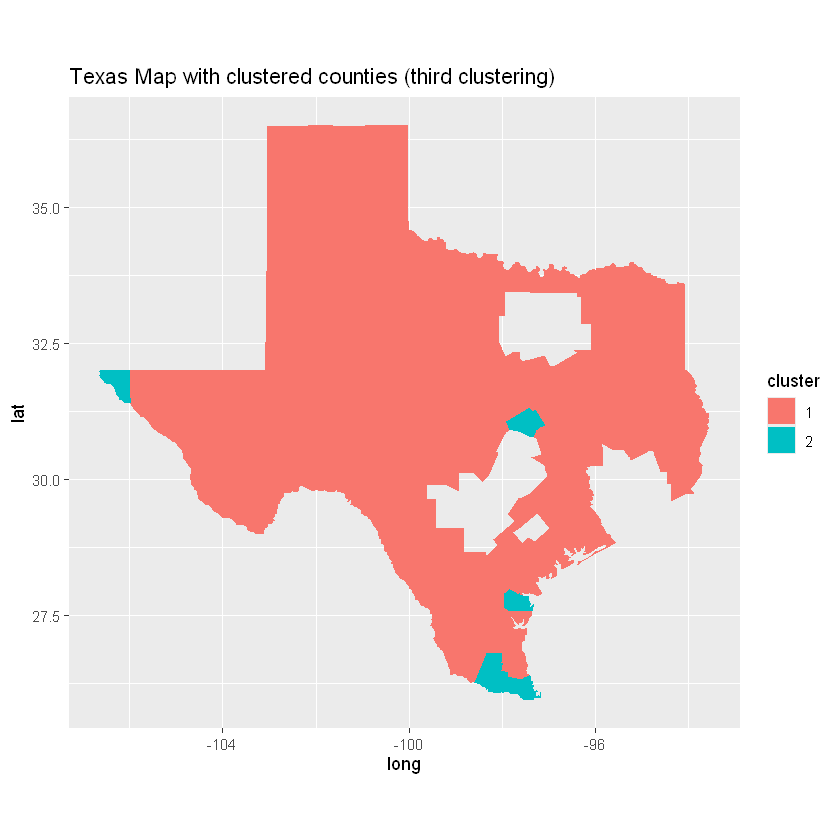


|  |  |  |
| --- | --- | --- |
| cluster | avg\_cases\_per\_1000 | avg\_deaths\_per\_1000 |
| 1 | 98.03851 | 2.363770 |
| 2 | 78.63196 | 1.981056 |

Fig. Clusters of counties for second clustering

**Third K-means Clustering**

After establishing optimum number of clusters, we generated a heat map of Texas showing all the counties that belonged to one of the two clusters. There are 5 counties that belong to cluster 1, and the name of those counties are shown below. Most of these counties lie far away at the edges of the state. Interestingly these are the same 5 counties that were in same cluster (i.e. cluster 1) in second k-means clustering. We can also see that these counties had higher average confirmed cases per 1000 than the counties that belong to cluster 1. However, these counties had slightly lower average deaths per 1000 than those that belong to cluster 1. This could mean that the counties in cluster 2 were able to control the deaths better than those counties that belong to cluster 1 despite having higher average confirmed cases.



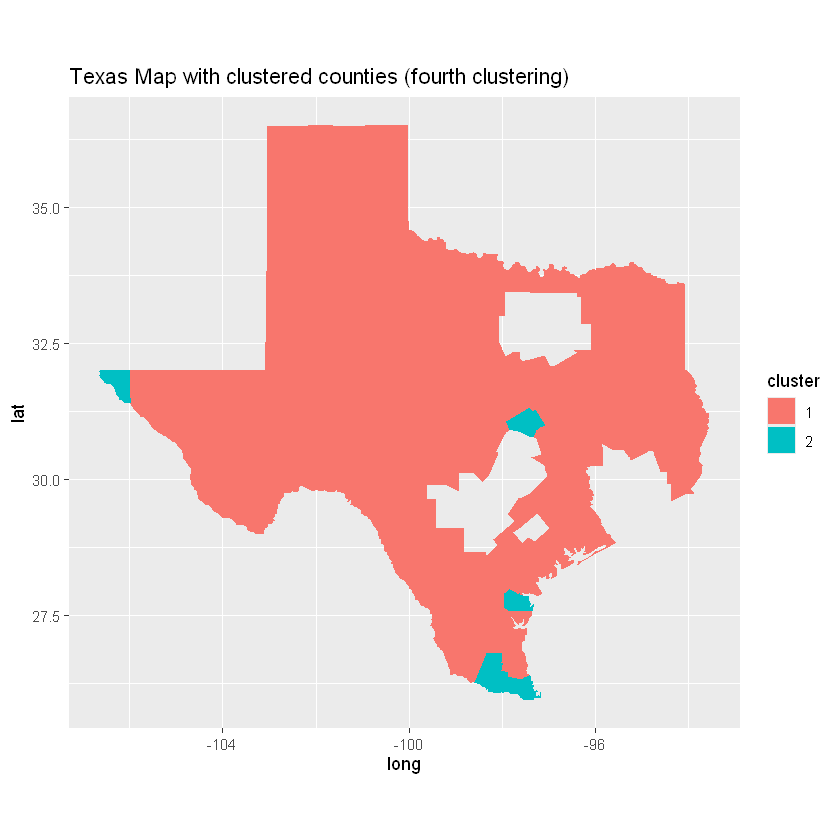


|  |  |  |
| --- | --- | --- |
| cluster | avg\_cases\_per\_1000 | avg\_deaths\_per\_1000 |
| 1 | 78.72391 | 1.986519 |
| 2 | 82.40390 | 1.897068 |

Fig. Clusters of counties for third clustering

**Fourth K-means Clustering**

After establishing optimum number of clusters, we generated a heat map of Texas showing all the counties that belonged to one of the two clusters. There are 5 counties that belong to cluster 2, and the name of those counties are shown below. Most of these counties lie far away at the edges of the state. Interestingly these are the same 5 counties that were in same cluster (i.e. cluster 2) in third k-means clustering. However, it is quite interesting to see that the average cases are higher in counties that belong cluster 1 while this was opposite in third clustering (i.e. higher in counties that belong to cluster 2). Same is true for the average deaths. This indicates that if the features/attributes that were considered for fourth clustering are chosen, the metrics on average confirmed cases and deaths are exactly opposite if the features from third clustering are chosen.





|  |  |  |
| --- | --- | --- |
| cluster | avg\_cases\_per\_1000 | avg\_deaths\_per\_1000 |
| 1 | 82.40390 | 1.897068 |
| 2 | 78.72391 | 1.986519 |

Fig. Clusters of counties for fourth clustering

# 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>