Data Mining Project Report

Project 2: Clustering

Student Names: Xiaoyingzi Liu

Mahsa Sheikhihafshejani

Abdulraheem Alobaidi

Project Group 4

University: Southern Methodist University

Date: 04-5-2021

ABSTRACT

In this report, the impact of the Covid-19 pandemic in Texas and the data similarities among the counties are analyzed. First, the data are collected and divided into three groups based on the different features. The first group of the data is related to the demographic composition such as family households, median age, median income and so on. The second group of data is based on the available resources such as the number of physicians, nurses, and hospital beds. The third group of the data is for special features such as social distancing response, which is the overall changes in the mobility from the baseline, death rate, average confirmed cases, and average death. The data is pre-processed to remove the outliers and replace the missing values in some features with the mean values. Subsequently, the data is clustered to analyze the similarity among the counties in Texas using different algorithm such as K-means clustering, hierarchical clustering, and PAM clustering. The internal validation and external validation are performed to analyze the performance of different clustering algorithms. Lastly, the results of the clustering are evaluated to find out some recommendations that might be helpful to slow down and reduce the impact of Covid-19 pandemic.

Content

1 INTRODUCTION 1

2 DATA PREPARATION 2

2.1 DATA DESCRIPTION 2
2.2 VERIFYING DATA QUALITY 4
2.3 STATISTICAL SUMMARY 4
2.4 OUTLIER REMOVAL AND MEAN IMPUTATION 5
2.5 DISTANCE MEASURES 7

3 MODELING 7

3.1 CLUSTERING ANALYSIS ON DATA RELATED TO DEMOGRAPHIC COMPOSITION 7
3.1.1 K-Means Clustering 7
3.1.2 Hierarchical Clustering 9
3.1.3 Partitioning Around Medoids Clustering (exceptional work) 11
3.1.4 Internal Validation 11
3.1.5 External Validation 13
3.2 CLUSTERING ANALYSIS ON DATA RELATED TO HOSPITAL RESOURCES (EXCEPTIONAL WORK) 14
3.2.1 K-Means Clustering 15
3.2.2 Hierarchical Clustering 16
3.2.3 Partitioning Around Medoids Clustering 17
3.2.4 Internal Validation 18
3.2.5 External Validation 20
3.3 CLUSTERING BASED ON SPECIFIC FEATURES 20
3.3.1 Spread rate and commuters by public transportation 20
3.3.2 Death per case, social distancing response, and hospital beds 23

4 EVALUATION 24

4.1 EVALUATION ON DATA RELATED TO DEMOGRAPHIC COMPOSITION 24
4.2 EVALUATION ON DATA RELATED TO HOSPITAL RESOURCES 25
4.3 EVALUATION ON DATA BASED ON SPECIFIC FEATURES 26

5 CONCLUSION 28

1 Introduction

Categorization methods called clustering can help us to find a dataset structure and explore our data. By analyzing this structure, we can better understand the behavior of the data. Clustering is grouping specific objects based on their characteristics and their similarities. As for data mining, this method divides the data into different groups that are best suited for the particular analysis using certain algorithm. Clustering algorithm is used extensively to organize and categorize data [1] and reduce the dimensionality of the data set with the object of finding distinct groups in the data set [2]. Clustering can be done as unsupervised, semi-supervised, or supervised [3]. Unsupervised clustering is a well-established method that finds the structure in unlabeled data. In this project, our focus is on the unsupervised learning on the Covid-19 data set to group data based on their different characteristics and see how this categorizing can differentiate the impact of Covid-19 on different counties in Texas. Different clustering methods can differ based on the type of data and the volume of data [4].

The impact of Covid-19 pandemic on the counties in Texas is analyzed in this report. The main goal of this report is to understand the similarities among the different counties in Texas in term of demographic features, hospital resources, and the other Covid-19 features such as the confirmed cases, death rate, spread rate, and social distancing response. Clustering analysis on the data set is important to identify counties that are similar to each other based certain features and group them together. Then, recommendations can be provided to effectively mitigate the impact of Covid-19 on each group of counties. In this report, multiple clustering methods are used and compared to find the best clustering method. The performance of each clustering algorithm is evaluated by internal and external validations.

In the following sections, the data set for this report is presented and pre-processed in the data preparation section. Modeling section presents different clustering algorithms that are used to analyze the similarities among counties in Texas based on different combination of features used in the dataset. Moreover, the quality of the clustering algorithms is analyzed by performing internal and external validation in the modeling section. The results from clustering are analyzed in the evaluation section. Finally, the main conclusions are presented in Section 5.

2 Data Preparation

2.1 Data Description

The data file contains three different data sets including US Covid-19 cases plus census data, Texas Covid-19 cases and Google global mobility report. US Covid-19 cases plus census data provides total US COVID-19 case and death counts by state and county until 2021-03-14, and census data provides useful information about the population. Texas Covid-19 cases data set provides COVID-19 case and death counts by county in Texas from 2020-01-22 to 2021-03-19. It also provides changes and trends by county. Google global mobility report aims to provide insights into what has changed in response to policies aimed at combating COVID-19. The reports chart movement trends over time by geography, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. For each category in a region, reports show the changes which compares mobility for the report date to the baseline day.

For this project, we focus on Covid-19 cases in Texas counties and try to group similar counties in order to find patterns. First of all, we are interested in counties that are similar in terms of makeup of the population. If we find counties with similar demographic composition, then they might exhibit a similar spread of the infection. A description of selected features regarding to demographic composition is given in Table 1. Second, we are also interested in whether counties might exhibit a similar spread of the infection if they have similar resources, for example hospital resources. A description of selected features regarding to hospital resources is given in Table 2. Note that the variable population density was created since we used density instead of total population. Moreover, the average number of confirmed cases or deaths increased during 03-13-2021 to 03-19-2021 in each county were created using Texas Covid-19 cases data. The distancing response was created using Global Mobility data. The features for comparison are presented in Table 3.

Table 1: The description of selected features regarding to demographic composition

Feature	Scale of Measurements	Description
county_name	Nominal	Name of the county
pop_density	Ratio	The population density of the county in people per square mile
family_households	Ratio	Total number of households contain families
median_age	Ratio	Median age
median_income	Ratio	Median income
unemployed_pop	Ratio	Population that are classified as unemployed
commuters_by_public_transportation	Ratio	Number of commuters by public transportation

Table 2: The description of selected features regarding to resources

Feature	Scale of Measurements	Description
county_name	Nominal	Name of the county
pop_density	Ratio	The population density of the county in people per square mile
physicians_and_dentists	Ratio	Total physicians and dentists in each county
registered_nurses	Ratio	Total registered nurses in each county
hospital_beds	Ratio	Total physicians and dentists in each county

Table 3: The description of created features for comparison

Feature	Scale of Measurements	Description
cases_per_1000	Ratio	Total confirmed cases per 1000 people in the total population
deaths_per_1000	Ratio	Total deaths per 1000 people in the total population
death_per_case	Ratio	Percentage of number of deaths per total confirmed cases
distancing_response	Ratio	The overall average changes in mobility from the baseline in each county
avg_confirmed	Ratio	The average number of confirmed cases increased during 03-13-2021 to 03-19-2021
avg_deaths	Ratio	The average number of deaths increased during 03-13-2021 to 03-19-2021

2.2 Verifying Data Quality

The quality of the data can be explored in following aspects: Are there missing values? Are there duplicate data? Are there outliers and are those mistakes? The data has 254 observations and 16 variables. Duplication of data will not be a problem in this data set. There were some missing values in variables *physicians_and_dentists*, *registered_nurses*, *hospital_beds* and *distancing_response*. The imputation method used was mean imputation. The mean of variables that contains missing values was calculated and used to replace all missing values in those variables.

2.3 Statistical Summary

The statistical summary of all selected features listed in Table 1 and Table 2 are given below. Note that *county_name* is excluded since the data type is nominal and it has no mode.

Table 3: The statistical summary of selected features

Feature	Mode	Frequency	Mean	Median	St.Dev	Min	Max
median_age	------	------	39.0	38.6	6.0	25.8	57.5
family_households	------	------	25,828	4,474	91,572	19	1,066,649
median_income	------	------	49,894	48,311	12,132	24,794	93,645
unemployed_pop	------	------	3090	481	12,157	0	149,192
commuters_by_ public_ transportation	0	97	737	10	4,675	0	57,933
physicians_and_ dentists	0	96	35	3	157	0	2,159
registered_ nurses	676	75	676	138	2,246	5	28,094
hospital_beds	343	77	343	104	828	5	7923
distancing_ response	-11.44	35	-11.4	-11.4	8.2	-54.9	18.5
cases_per_1000	------	------	98.5	96.7	28.8	13.5	194.9
deaths_per_1000	2.14	4	2.5	2.4	1.2	0	8.3
death_per_case	0.02	96	0.03	0.02	0.014	0	0.09
pop_density	------	------	114.1	21.5	337.0	0.1	2,929.2
avg_confirmed	0	64	15.6	1.0	65.0	0	746.6
avg_deaths	0	135	0.5	0	1.8	0	19.5

2.4 Outlier Removal and Mean Imputation

The first step is to remove the outliers form the data. By plotting the lof figure (Figure 1), we can set the lof to 2.2 and remove the outliers.

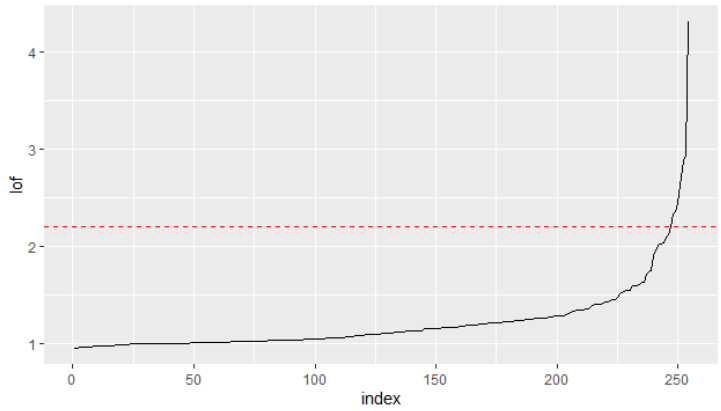

Fig. 1: lof for outlier removal

Form 254 counties, 7 counties are removed. Bell County, Bexar County, Dallas County, Harris County, Hidalgo County, Loving County, McMullen County are the counties that are recognized as outliers and should be removed from the dataset. These seven counties are shown in Figures 2. From these figures, we see that when we plot the outliers for in three dimensions, for this specific combination, some points are not outliers and they are outlier with respect to other features.

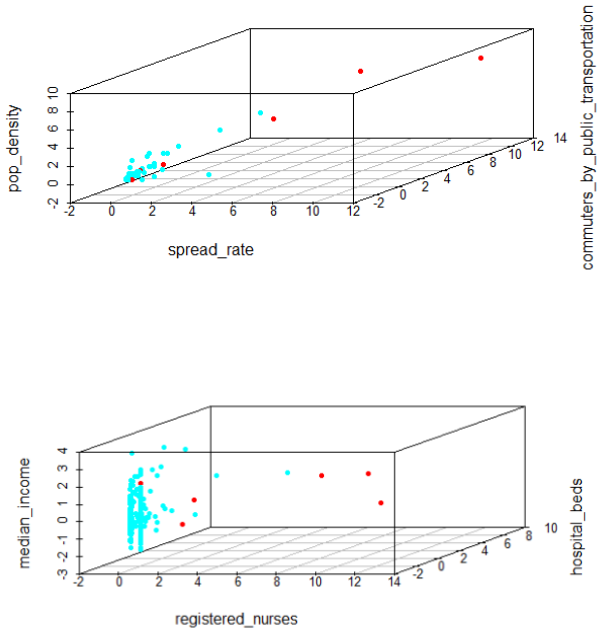

Fig. 2: Outliers with respect to three features

2.5 Distance Measures

All features in the objects are in ratio scale. Therefore, the distance measures can be obtained using Minkowski distance such as Manhattan distance, Euclidean distance, and maximum distance. The largest three counties in Texas (Harris, Dallas, and Tarrant) are selected to compare the three distance measures. Table 2 shows the distance between the largest three counties using different distance measures. As it can be seen, the Harris county is closer to Dallas county while it is very dissimilar to Tarrant county in all the three distance measures.

Table 4: The distance between the Harris, Dallas, and Tarrant counites

Distance Measure	County	County	
		Harris	Dallas
Manhattan distance	Dallas	11.449915	-
	Tarrant	17.425246	8.290812
Euclidean distance	Dallas	3.780920	-
	Tarrant	5.774283	3.239836
Maximum distance	Dallas	1.825742	-
	Tarrant	2.674061	2.092457

3 Modeling

3.1 Clustering Analysis on data related to demographic composition

In this section, the features that are related to demographic composition are selected to cluster the counties in Texas to analyze the similarity of the data. The features for this clustering include population density, family households, median age, median income, unemployed population, and commuters by public transportation.

3.1.1 K-Means Clustering

Determine Number of Clusters

The optimal number of clustering is determined by looking at the knee point in which the within cluster sum of squares (WSS) is dramatically reduced. Figure 3 show the values of WSS in each number of clusters. As it can be seen, the WSS decreases as the number of clusters increases.

The knee point occurs when the number of clusters is five. Therefore, the K-means clustering is performed to obtain five number of clusters.

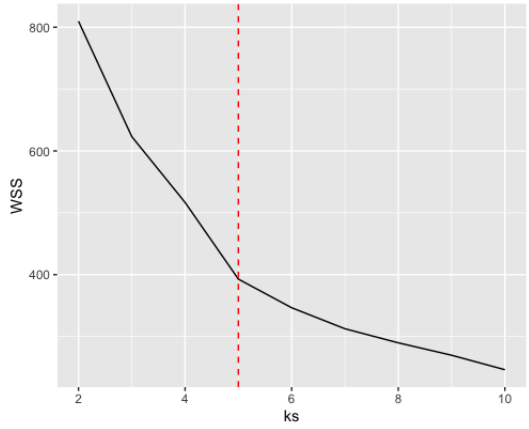

Fig. 3: The within cluster sum of squares (WSS) in each number of clusters

Results and discussion

Figure 4 shows the cluster profiles obtained by K-means clustering. The counties are clustered in five groups. Total 247 observations were included and for cluster 1 to 5, the size was 8, 97, 45, 2, 95 respectively. Cluster 1 and cluster 4 are similar in term of the features that are above or below the average; however, the range of values are different. Only median age is above the average in cluster 2 and only median income is above the average in cluster 3. The median age and median income are below the average in cluster 5 and the remaining variables are on the average. Table 5 shows the summary of the five clusters obtained by K-means clustering.

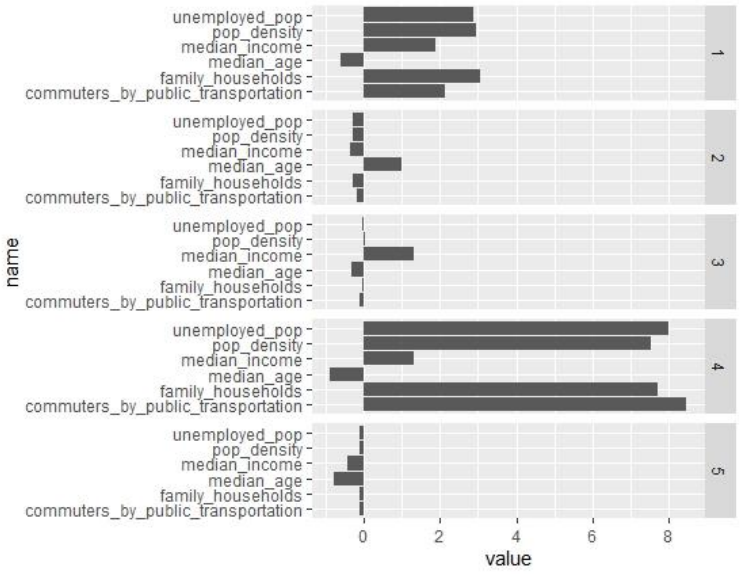

Fig. 4: The cluster profiles by K-means

Table 5: The summary of the clusters in term of the variables being above, below, or on the average

Clusters	Features on the average	Features above the average	Features below the average
Cluster 1	-	- population density - family households - median income - unemployed population - commuters by public transportation	- median age
Cluster 2	-	- median age	- population density - family households - median income - unemployed population - commuters by public transportation
Cluster 3	- unemployed population - population density - family households - commuters by public transportation	- median income	- median age
Cluster 4	-	(above average but more than Cluster 1) - population density - family households - median income - unemployed population - commuters by public transportation	- median age
Cluster 5	- population density - family households - unemployed population - commuters by public transportation	-	- median income - median age

3.1.2 Hierarchical Clustering

Determine Number of Clusters

We will look at different figures for methods Within Sum of Squares, Average Silhouette Width, and Dunn Index to determine the number of clusters. As shown in the Figure 5, Within Sum of Squares, Average Silhouette Width and Dunn Index provide different number of clusters due to the nature of the data which is not well behaved for clustering. Therefore, we choose 5

number of clusters in the hierarchical clustering algorithm here to compare it with K-means clustering.

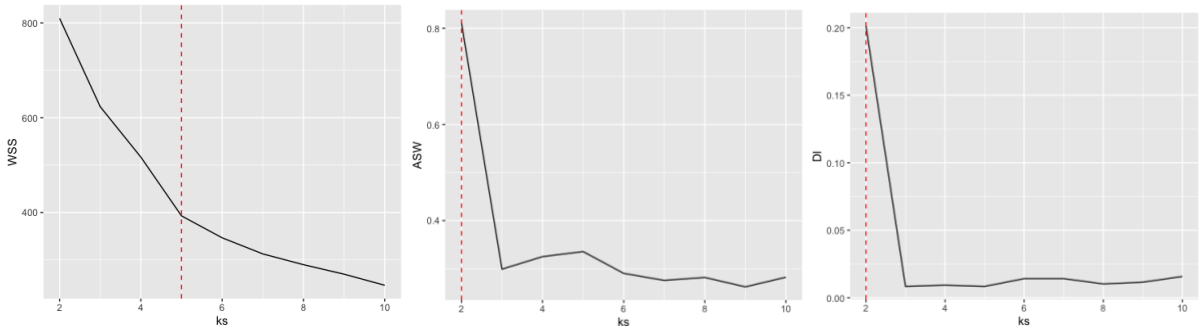

Fig. 5: The WSS, ASW and Dunn Index in each number of clusters

Results and Discussion

Figure 6 shows the cluster profiles obtained by hierarchical clustering. The counties are clustered in five groups. All the features in Cluster 1 seem to on the average except median income below the average. In Cluster 2, median income is above the average and median age below the average, other features are on the average. Cluster 3, Cluster 4 and Cluster 5 are similar in term of the features that are above or below the average; however, the range of values are different. For example, the feature commuters by public transportation is well above the average in Cluster 5.

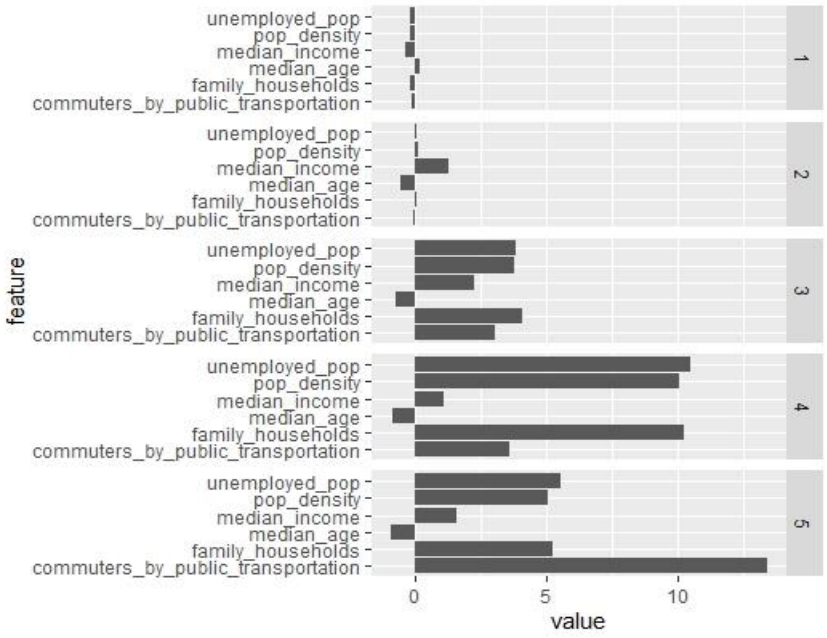

Fig. 6: The cluster profiles by hierarchical clustering

3.1.3 Partitioning Around Medoids Clustering (exceptional work)

In this section, partitioning around medoids (PAM) algorithm is used to cluster the data based on the features related to demographic composition. The number of clusters is chosen to be 5 clusters to be comparable with K-means and hierarchical clustering. Figure 7 shows that the cluster profile using PAM clustering. In Cluster 1, the median age is above the average while the other features are below the average. Cluster 1 obtained by PAM is similar to Cluster 2 obtained by K-means clustering. In Cluster 2, the median income is the only features that is above the average. In Cluster 3, all the features are below the average. Cluster 4 is similar to Cluster 1; however, the range of values are larger in Cluster 4 than in Cluster 1. Interestingly, Cluster 5 is similar to Cluster 1 and Cluster 4 in term of the features that above or below that average, but the values in Cluster 5 is larger than in Cluster 1 and Cluster 2.

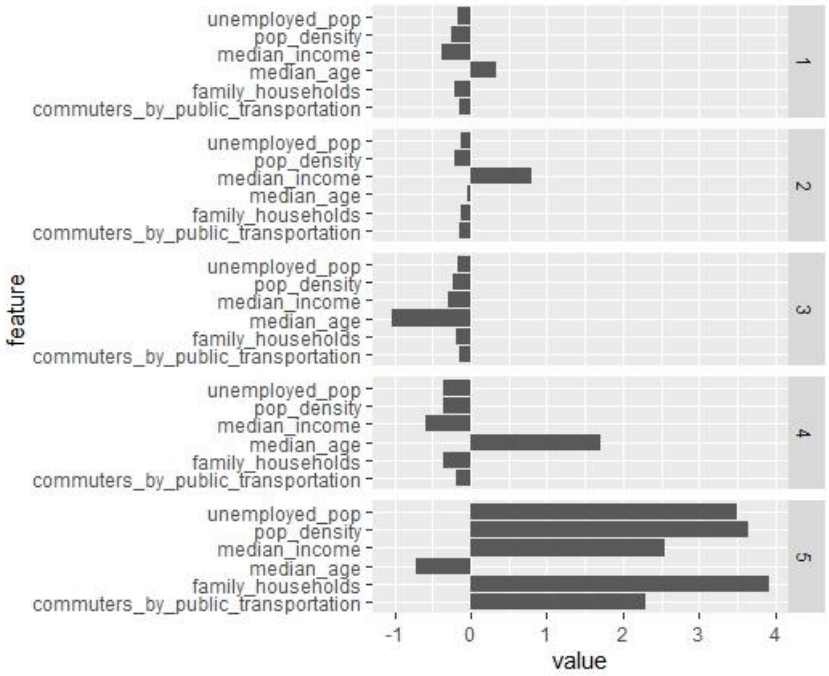

Fig. 7: The cluster profiles by PAM clustering

3.1.4 Internal Validation

We compared clustering quality by looking at the within cluster sum of squares and the average silhouette width. The results of the comparison are presented in Table 6, and the silhouette plot of K-Means Clustering is presented in Figure 8. We can tell that the average

silhouette width of K-Means Clustering and PAM clustering are 0.34 and 0.26, respectively. While the average silhouette width of hierarchical clustering is 0.75 which is much closer to 1. Therefore, the results of hierarchical clustering fit the data better. In addition, the visualization of distance matrix is presented in Figure 9. It seems that Clusters 2, 3, 5 are similar to each other while Cluster 4 is very dissimilar that the other clusters.

Table 6: The results of comparison

	Km	Hc_complete	Hc_single	PAM
within.cluster.ss	392.747	623.716	813.742	522.393
avg.silwidth	0.336	0.747	0.718	0.256

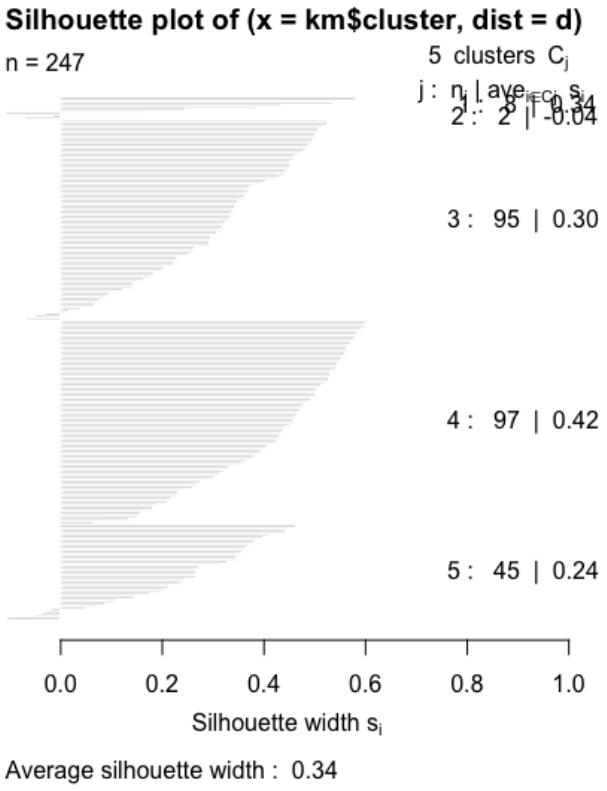

Fig. 8: Silhouette plot of K-means

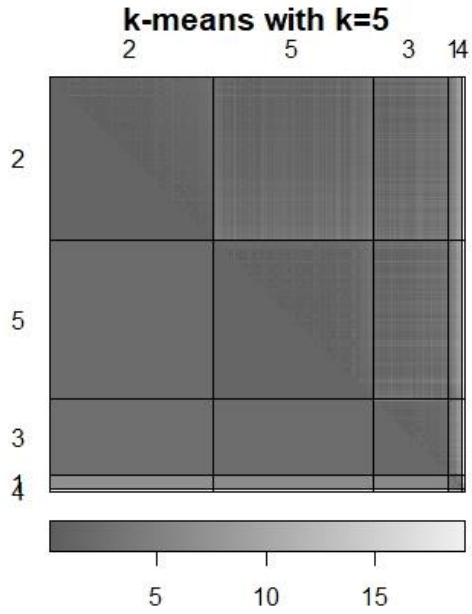

Fig. 9: Distance Matrix of K-means

3.1.5 External Validation

To evaluate the clustering algorithm used in section 3.1, the external validation is performed using the death per case in the counties as the ground truth information. The goal is to measure which cluster labels match the external supplied class labels. The measure indices for comparison are corrected rand index (ARI), variation of information (VI), entropy, and purity.

Figure 10 shows the ground truth which is the counties in Texas grouped into high death per case and low death per case. Table 7 shows the results for the external validation. The corrected rand index is closed to zero which shows that the clustering obtained by K-means clustering, hierarchical clustering, and PAM clustering have no relation with the ground truth information. Clustering obtained by hierarchical clustering has the lowest variation of information (1.066). On the other hand, the PAM clustering has highest variation of information (1.88). The value of entropy shows that the clustering obtained by hierarchical clustering has the lowest impurity (0.397) while the value of purity shows the hierarchical clustering has the highest value of 0.89. Therefore, the external validation shows that the hierarchical clustering outperforms the K-means clustering and PAM clustering in term of purity of the clustering and the variation of information. On the other hand, the PAM clustering has the lowest performance among the other clustering algorithm.

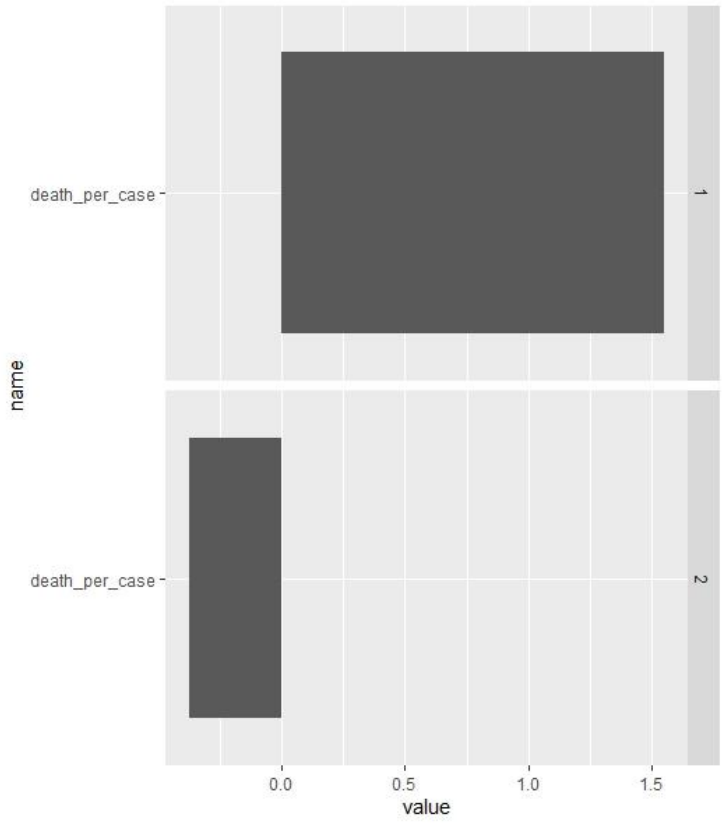

Fig. 10: The ground truth which is the counties in Texas grouped into high death per case and low death per case

Table 7: The results of external validation

Clustering Algorithm	Corrected Rand Index (ARI)	Variation of Information (VI)	Entropy	Purity
kmeans_5	-0.0237	1.644	0.743	0.171
hc_5	-0.0989	1.066	0.397	0.889
PAM_5	0.00856	1.884	1.047	0.187

3.2 Clustering Analysis on data related to hospital resources (exceptional work)

In this section, the features that are related to hospital resources are selected to cluster the counties in Texas to analyze the similarity of the data. The features for this clustering include population density, number of physicians/dentists, number of registered nurses and number of hospital beds.

3.2.1 K-Means Clustering

Determine Number of Clusters

The optimal number of clustering is determined by looking at the knee point in which the within cluster sum of squares (WSS) is dramatically reduced. Figure 11 show the values of WSS in each number of clusters. As it can be seen, the WSS decreases as the number of clusters increases. The knee point occurs when the number of clusters is four. Therefore, the K-means clustering is performed to obtain four number of clusters.

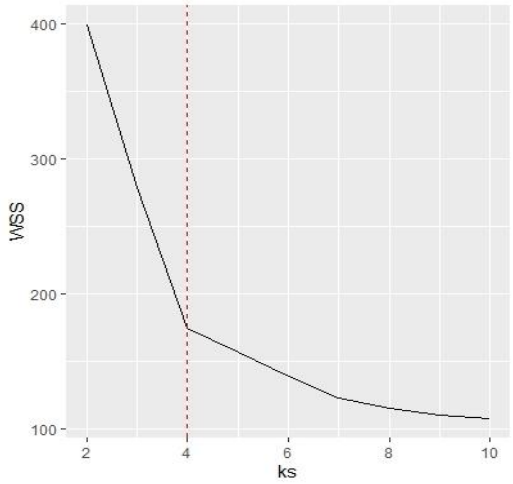

Fig 11: The within cluster sum of squares (WSS) in each number of clusters

Results and Discussion

Figure 12 shows the cluster profiles obtained by K-Means Clustering. The counties are clustered in four groups. Total 247 observations were included and for cluster 1 to 4, the size was 138, 12, 94, 3 respectively. All the features in Cluster 1 are below the average. In Cluster 2, only physicians/dentists is slightly below the average while the other features are above the average. In Cluster 3, population density is slightly below the average while the other features are slightly above the average. Lastly, all the features are well above the average in Cluster 4.

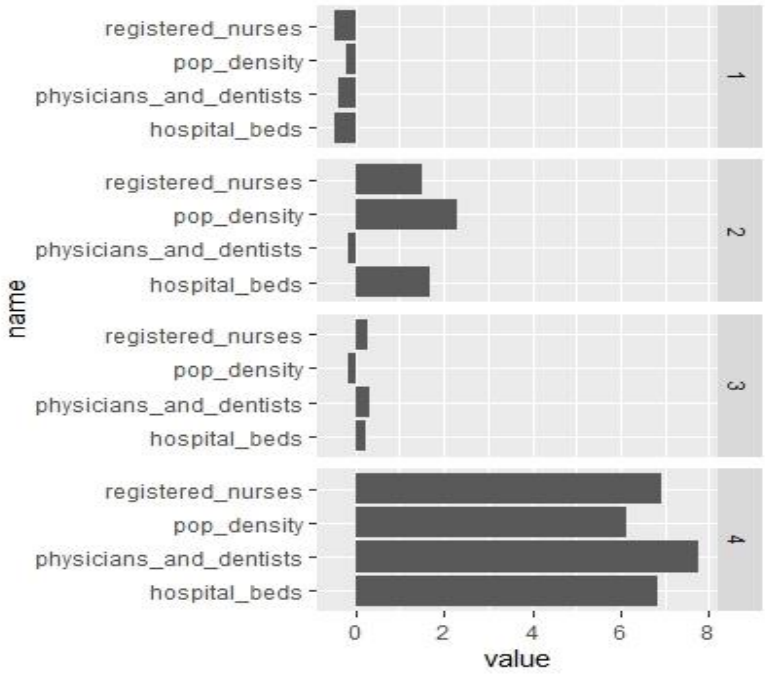

Fig. 12: The cluster profiles obtained by K-means clustering

3.2.2 Hierarchical Clustering

Determine Number of Clusters

We will look at different figures for methods Within Sum of Squares, Average Silhouette Width, and Dunn Index to determine the number of clusters. As shown in the Figure 13, Within Sum of Squares, Average Silhouette Width and Dunn Index provide different number of clusters due to the nature of the data which is not well behaved for clustering. Therefore, we choose 4 number of clusters in the hierarchical clustering algorithm here to compare it with K-means clustering.

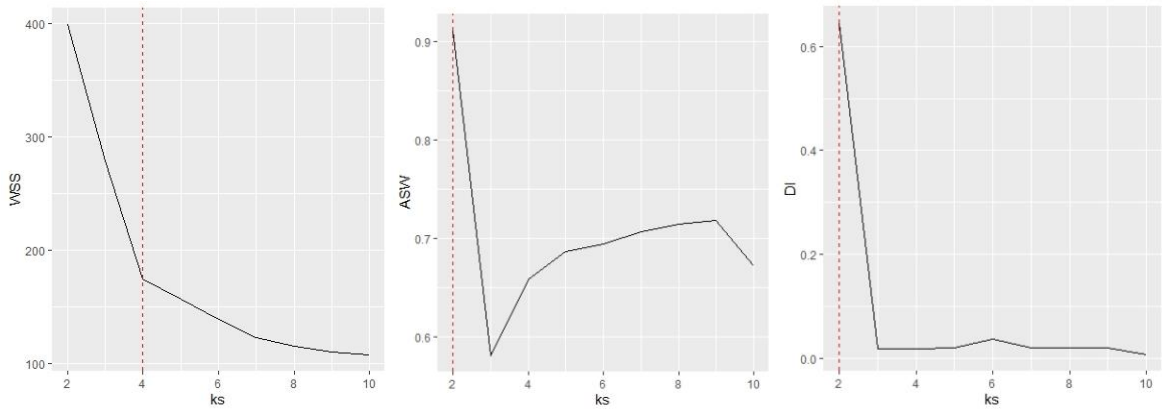

Fig. 13: The WSS, ASW and Dunn Index in each number of clusters

Results and Discussion

The clustering profiles obtained by hierarchical clustering are shown in Figure 14. In Cluster 1, the hospital resource features and population are below the average. In Cluster 2, the physicians and dentists are below the average while the other features are above the average. In Cluster 3 and Cluster 4, the hospital resources are above the average. However, the value ranges are different in Cluster 3 and Cluster 4. It seems that Cluster 4 has higher population and hospital resources than Cluster 3. It can be concluded the hospital resources are higher for the high populated counties.

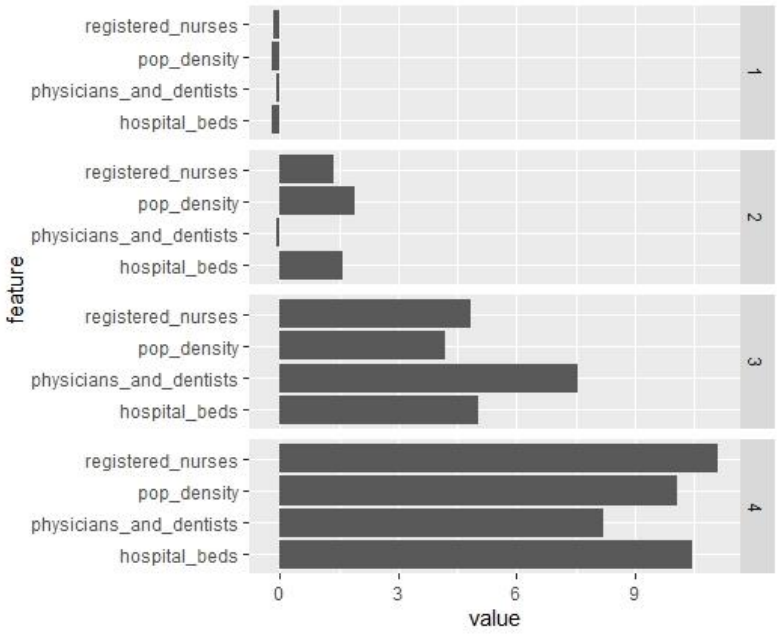

Fig. 14: The cluster profiles obtained by hierarchical clustering

3.2.3 Partitioning Around Medoids Clustering

In this section, the PAM algorithm is performed to obtained four clusters for the hospital resources data. Figure 15 shows the cluster profiles obtained by PAM clustering. In Cluster 1, all features are below the average while the features are above the average in Cluster 4. Meanwhile, the population is below the average in Cluster 2, while physicians and dentists are below the average is in Cluster 3. It seems that the clusters obtained by PAM is similar to the

clusters obtained by the K-means clusters while it is quite different than the clusters obtained by hierarchical clustering.

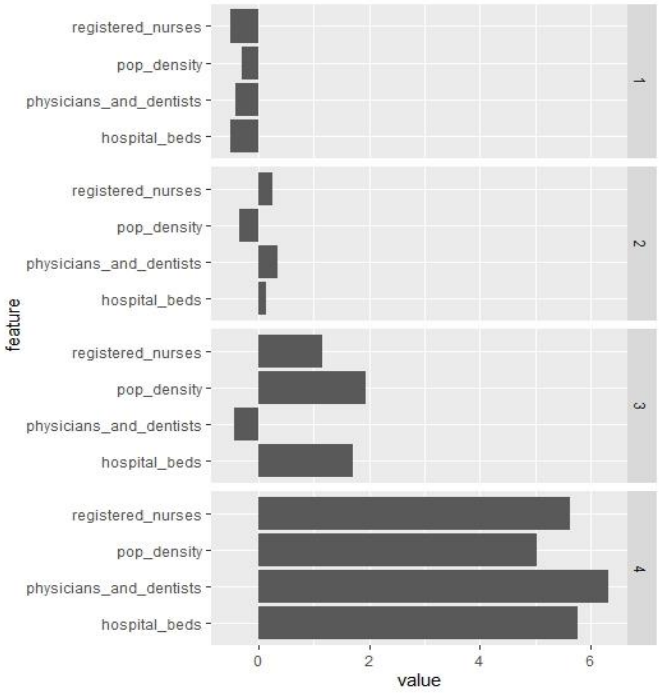

Fig. 15: The cluster profiles obtained by PAM clustering

3.2.4 Internal Validation

We compared clustering quality by looking at the within cluster sum of squares and the average silhouette width. The results of the comparison are presented in Table 8, and the silhouette plot of K-Means Clustering is presented in Figure 16. We can tell that the average silhouette width of K-Means Clustering and PAM clustering are 0.659 and 0.665, respectively. On the other hand, the average silhouette width of hierarchical clustering is 0.72 which is much closer to 1. Therefore, the results of hierarchical clustering fit the data better. In addition, the visualization of distance matrix is presented in Figure 17. It seems that Cluster 1 and Cluster 3, are similar to each other while Cluster 4 is very dissimilar that the other clusters.

Table 8: The results of comparison

	Km	Hc_complete	Hc_single	PAM
within.cluster.ss	174.558	188.607	322.937	174.904
avg.silwidth	0.659	0.719	0.882	0.665

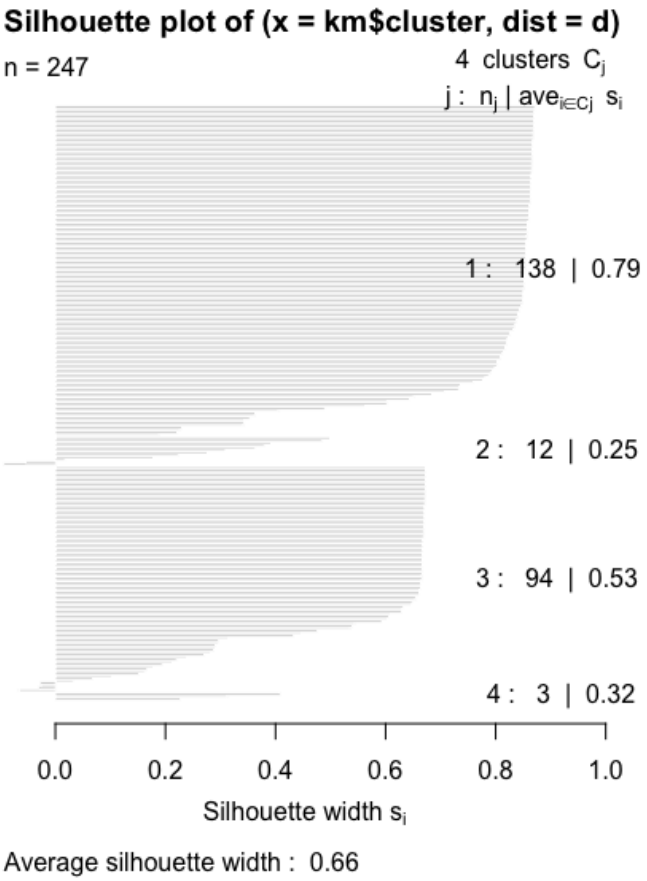

Fig. 16: Silhouette plot of K-means

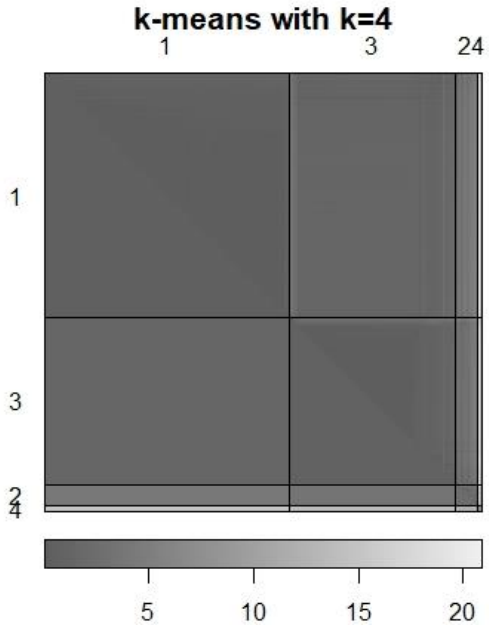

Fig. 17: Distance Matrix of K-means

3.2.5 External Validation

In this section, the external validation is performed to measure which cluster labels match the external supplied class labels. Similar to Section 3.1.5, the death per case is used as the ground truth information which is shown previously in Figure 10.

Table 9 shows the results for the external validation. The corrected rand index is closed to zero which shows that the clustering obtained by K-means clustering, hierarchical clustering, and PAM clustering have no relation with the ground truth information. Clustering obtained by hierarchical clustering has the lowest variation of information (0.785). On the other hand, the PAM clustering has highest variation of information (1.3789). The value of entropy shows that the clustering obtained by hierarchical clustering has the lowest impurity (0.163) while the value of purity shows the hierarchical clustering has the highest value of 0.9637. Therefore, the external validation shows that the hierarchical clustering outperforms the K-means clustering and PAM clustering in term of purity of the clustering and the variation of information. On the other hand, the K-means clustering has the lowest performance among the other clustering algorithm in term of purity index.

Table 9: The results of external validation

Clustering Algorithm	Corrected Rand Index (ARI)	Variation of Information (VI)	Entropy	Purity
kmeans_4	0.00123	1.372	0.691	0.302
hc_4	-0.0591	0.785	0.163	0.964
PAM_4	0.00449	1.379	1.121	0.491

3.3 Clustering based on specific features

3.3.1 Spread rate and commuters by public transportation

Optimal number of clusters are determined by looking at the different methods Within Sum of Squares, Average Silhouette Width, Dunn Index and Gap Statistic in Figure 18. Since we are getting different number of clusters suggested by these methods, we chose four cluster

and then increase the number of clusters to see how the clustering changes with different number of clusters.

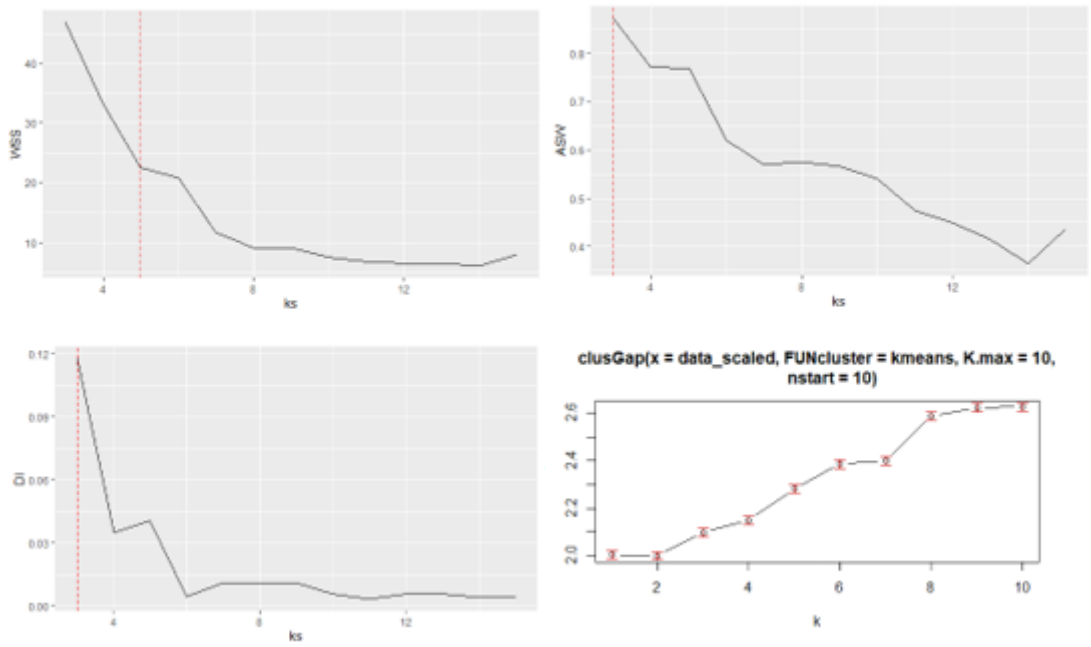

Fig. 18: Optimal number of clusters using different methods

Figure 19 shows the relation between different variables in the first combination. Variables have positive correlations and by increasing one variable, other ones also increase. Also, we see that almost all of the points for each variable are around the minimum value and less points have large value of each variable. Because of this fact, we will have almost all of data in one cluster and other clusters contains less points.

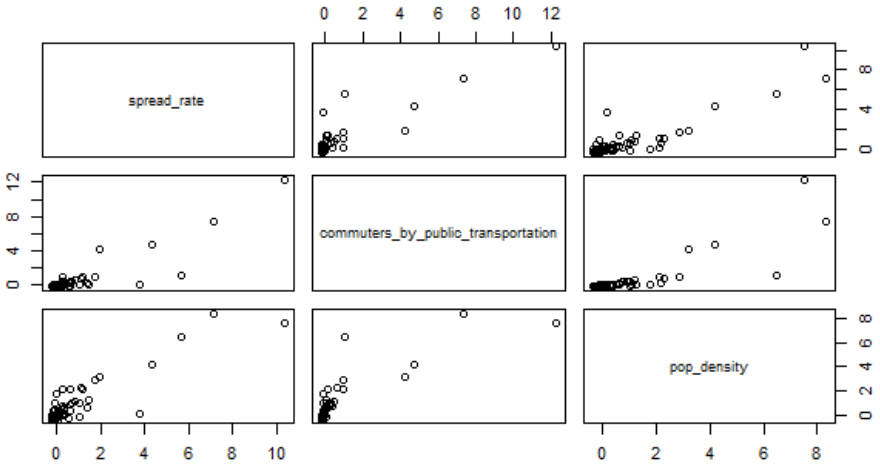

Fig. 19: pair plot for the first combination

Considering four clusters for the K-Means method, we get the following clusters: Four clusters are shown in Figure 20 and there is only one point in one cluster which may have to be distinguished as outlier. As we mentioned before in pair plot, since almost all of points are around the minimum points, we have a dense cluster and then a smaller cluster for larger values and then less dense clusters.

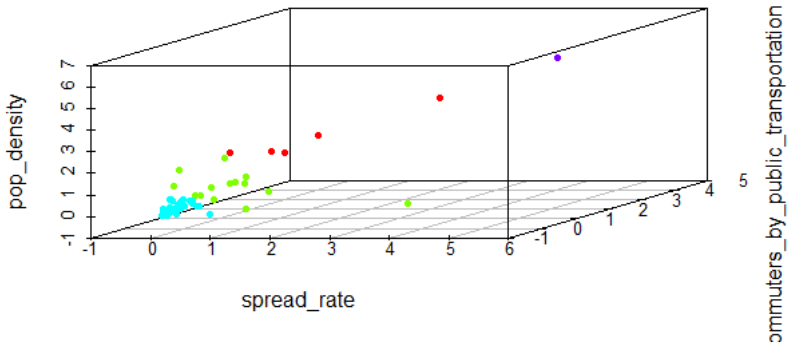

Fig. 20: four clusters for K-Means

The scaled number of spread rates, population density and commuter by public transportation for each cluster is shown in Figure 21. For the first cluster which contains one point that corresponds to Tarrant county, we have the largest number of each features for all three variables. The next cluster contains points with less values for these three variables and then we have two other clusters one with less values for the variables and the other with negative values.

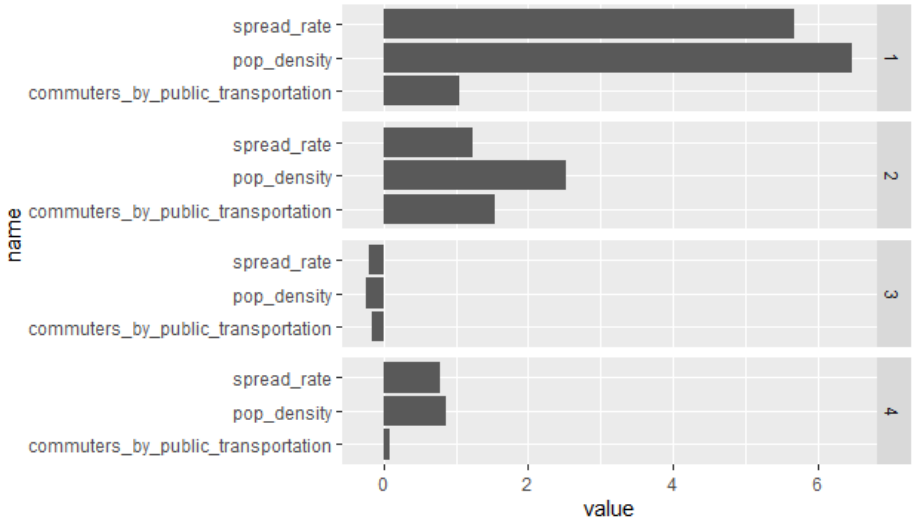

Fig. 21: Cluster profiles obtained by K-Means clustering

3.3.2 Death per case, social distancing response, and hospital beds

In this section, the data for hospital beds, social distancing response, and death per cases are clustered to analyze the similarity of the counties in Texas based on these special features. Figure 22 shows the clustering profiles obtained by K-means algorithm. Cluster 1 has 4 counties in which the hospital beds are above the average while the social distancing response and death per case are below the average. Cluster 2 contains 52 counties that the hospital beds are slightly below the average, the social distancing is below the average, and the death per case is on the average. The majority of the counties in Texas is in Cluster 3 in which the hospital beds on the average, social distancing response is above the average and the death per case is below the average. Lastly, Cluster 4 has 41 counties where death per case is very high above the average, the social distancing is slightly above the average while the hospital beds is below the average. It seems that the majority of counties in Texas have death per case below the average. These counties have hospital beds on the average. However, 41 counties have the death per case is above the average. For these counties, the hospital beds are below the average and the social distancing is slightly above the average. Therefore, the proper response to reduce the death per case might be by increasing the hospital beds in these counties and enforcing additional restrictions.

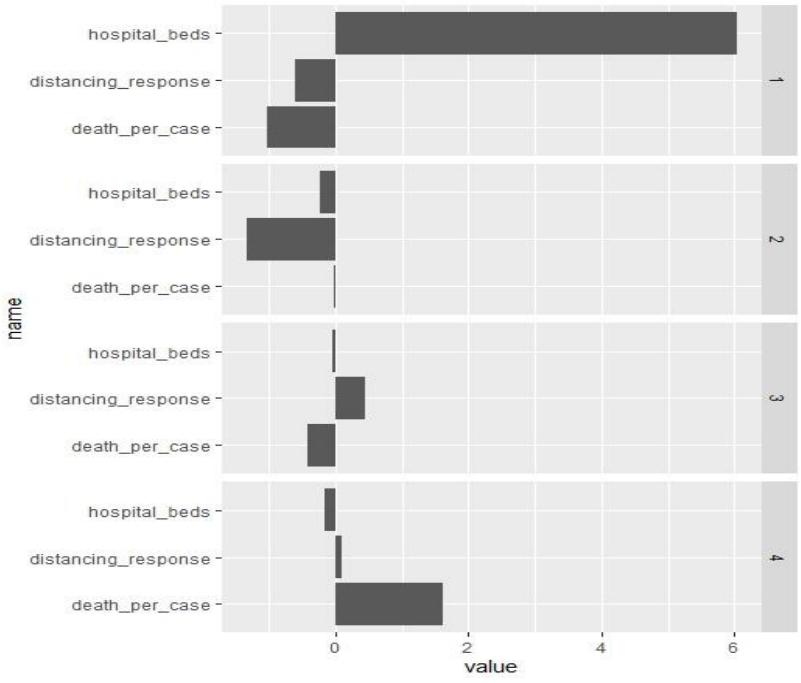

Fig. 22: Cluster profiles obtained by K-Means clustering

4 Evaluation

4.1 Evaluation on data related to demographic composition

Figure 23 visualizes the clusters based on Texas county map. We can see that majority of the counties are in cluster 2 and 5. Next, we want to check if some statistics, such as confirmed cases /deaths rates (per 1000 people), social distancing response rate and trend for last week are different in different clusters. The comparison is shown in the Table 10.

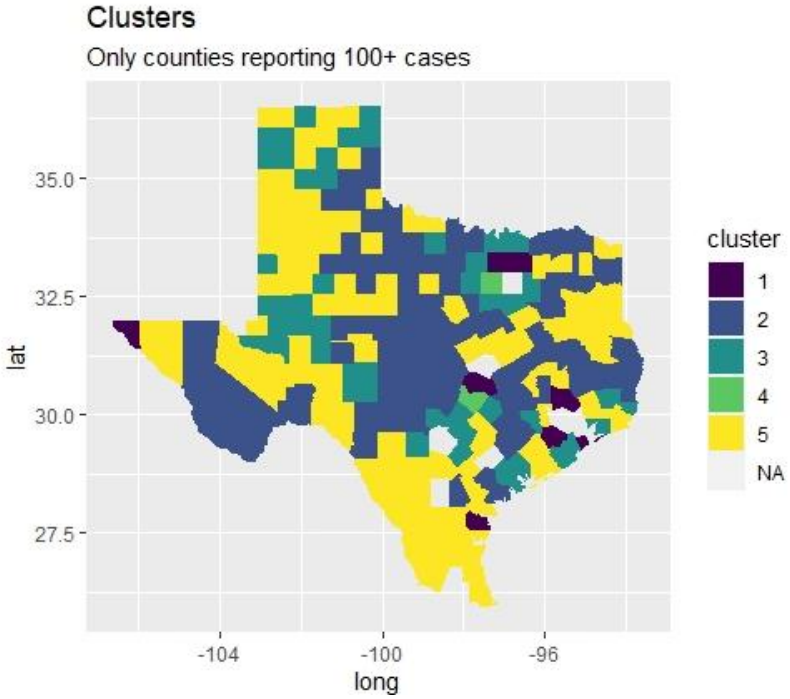

Fig. 23: Clusters based on Texas county map

Table 10: Comparison for the specific features in different clusters

Cluster	Avg_cases	Avg_deaths	Avg_distancing_reponse	Avg_cases_lastweek
1	92.3	2.22	-9.07	14.6
2	102	2.69	-11.5	4.85
3	88.9	2.43	-9.70	5.21
4	126	2.90	-4.58	5.42
5	98.9	2.52	-12.2	12.8

From the table above, we can tell that Cluster 4 has the highest average confirmed cases per 1000 people while Cluster 3 has the lowest average confirmed cases per 1000 people. All Clusters have similar average deaths. We also noticed that from March 13th to 19th, Cluster 1 has the highest number of average increased confirmed cases. Therefore, we recommend that people in those counties should always wear masks in public to prevent the transmission of the virus. Sending more vaccines to Cluster 1 would also be helpful.

4.2 Evaluation on data related to hospital resources

Figure 24 visualizes the clusters based on Texas county map. We can see that majority of the counties are in cluster 1 and 3. Next, we want to check if some statistics, such as confirmed cases /deaths rates (per 1000 people), social distancing response rate and trend for last week are different in different clusters. The comparison is shown in the Table 11.

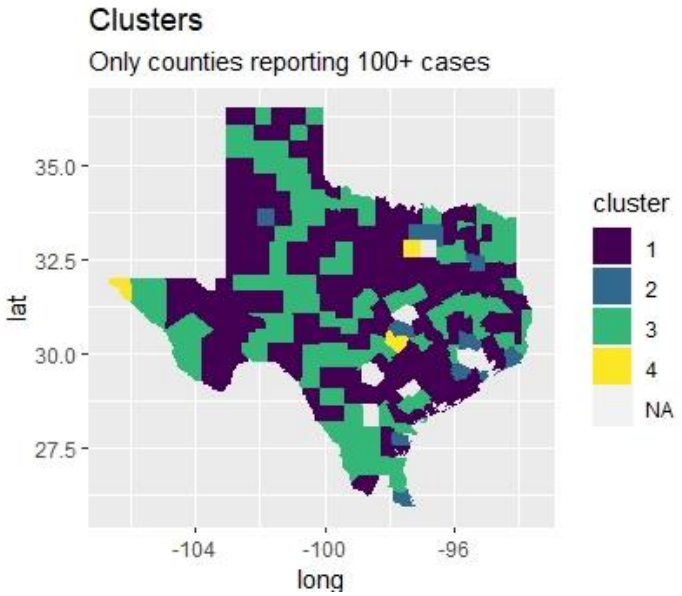

Fig. 24: Clusters based on Texas county map

Table 11: Comparison for the specific features in different clusters

Cluster	Avg_cases	Avg_deaths	Avg_distancing_reponse	Avg_cases_lastweek
1	95.6	2.56	-11.5	8.03
2	94.5	2.20	-11.0	12.8
3	102	2.62	-11.5	8.21
4	115	2.94	3.1	3.61

From the table above, we can tell that Cluster 4 has the highest average confirmed cases per 1000 people and the highest average deaths per 1000 people. However, interestingly, we noticed that from March 13th to 19th, Cluster 4 has the lowest number of average increased confirmed cases. We may guess that counties in Cluster 4 controlled the transmission of the virus recently. We also noticed that Cluster 2 has the highest number of average increased confirmed cases. Therefore, we recommend that people in those counties should always wear masks in public to prevent the transmission of the virus. Sending more vaccines to Cluster 2 would also be helpful.

4.3 Evaluation on data based on specific features

The K-means clustering is shown in the Texas map in Figure 25. The smallest cluster which contains Tarrant county is shown by the green color. The cluster that contains most of the counties is shown by yellow color. The PCA plot is shown in Figure 26. One cluster is the Tarrant county which is shown by blue color. Most of the counties are in the small cluster which is shown by purple color.

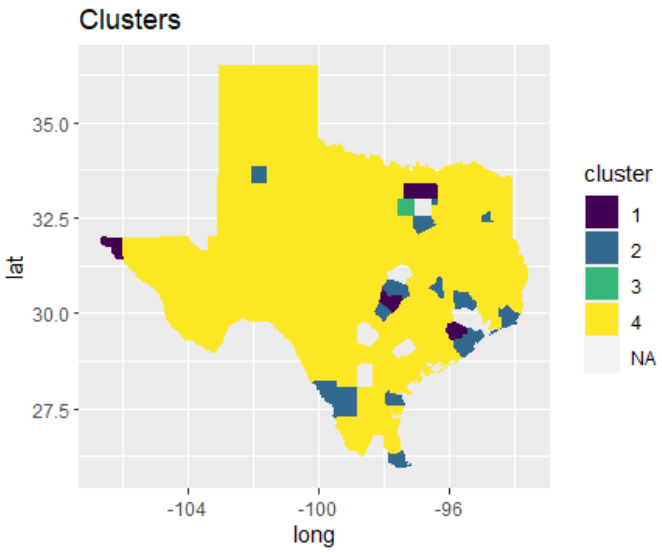

Fig. 25: Texas map showing clustering

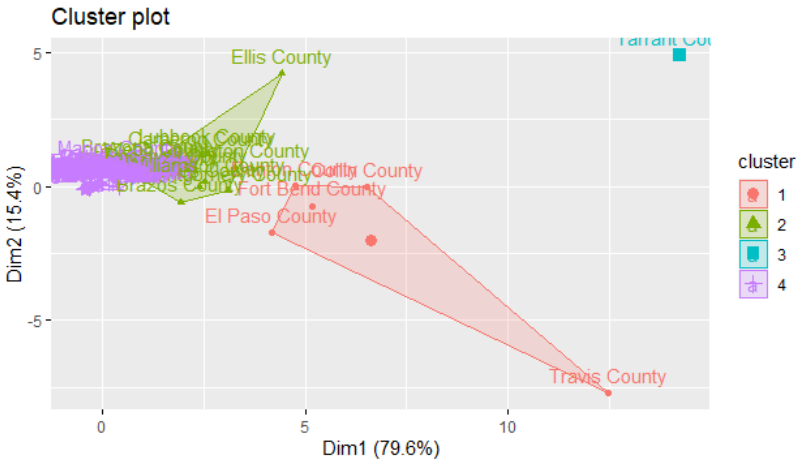

Fig. 26: PCA showing clusters

Clustering with different number of clusters is shown by PCA plot in Figure 27. Considering three clusters, we get a large cluster and two small cluster and by increasing the number of clusters to four clusters, the largest cluster breaks into two clusters. By increasing the number of clusters, it seems that five is a good choice for K-means clustering because we see that by increasing the number of clusters to six, the newly added cluster contains only one point.

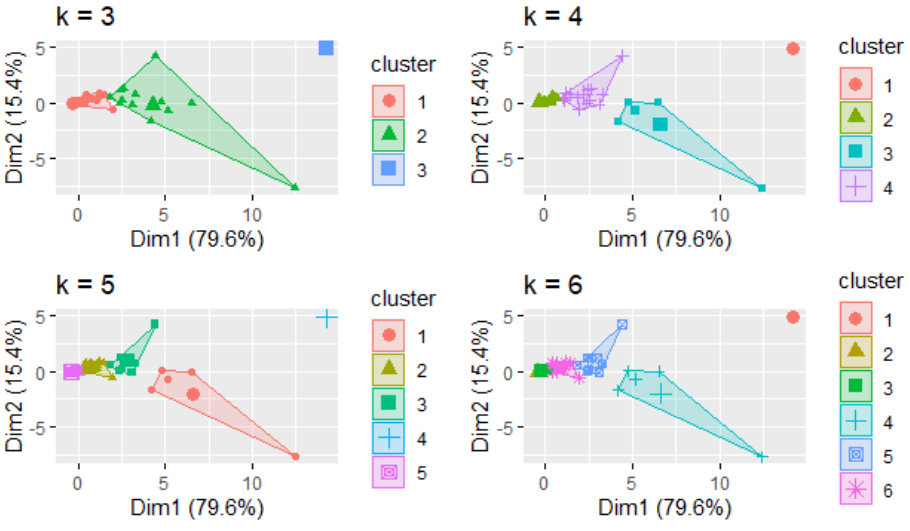

Fig. 27: PCA showing different number of clusters

5 Conclusion

In this report, clustering analysis is performed on the Covid-19 data for Texas. The objective is to understand the similarities among counties in Texas based on certain features. K-means clustering, hierarchical clustering, PAM clustering algorithms are applied to group the counties based on the demographic features, hospital resources, and specific features such as spread rate, social distancing response and death per rate.

The internal validation results show that the hierarchical clustering has the highest average silhouette width. Therefore, hierarchical clustering is better fit the data than K-means clustering or PAM clustering. The external validation results show that the hierarchical clustering outperforms the K-means clustering and PAM clustering in term of purity of the clustering and the variation of information for the same number of clustering.

The results show that the spread rate is high for counties that have high populations and large commuters by public transportation. Enforcing some restrictions for those counties and encourage people to work from home might be the proper response to slow down the spread of Covid-19. The results also show that only four counties have higher number of hospital beds than the average. The death per case for these counties are below the average. The majority of counties in Texas have death per case below the average. These counties have hospital beds on average. However, 41 counties have the death per case above the average. The hospital beds are below the average for these counties while the social distancing response is slightly above the average. For these counties, increasing the hospital beds and enforcing additional restrictions might be the proper response to reduce the death rate.

References

[1] M. Verma, M. Srivastava, N. Chack, A. K. Diswar, and N. Gupta, “A Comparative Study of Various Clustering Algorithms in Data Mining,” Int. J. Eng. Res. Appl. www.ijera.com, vol. 2, no. 3, pp. 1379–1384, 2012.

[2] P. K. Jain and R. Pamula, “Two-Step Anomaly Detection Approach.”

[3] M. Z. Hossain, M. N. Akhtar, R. B. Ahmad, and M. Rahman, “A dynamic K-means clustering for data mining,”

[4] S. Rashid, A. Ahmed, I. Al Barazanchi, and Z. A. Jaaz, “Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set,” Period. Eng. Nat. Sci., vol. 7, no. 2, pp. 448– 457, 2019.