DATA MINING – PROJECT 2

United States COVID-19 Cluster Analysis

Varun Singh

Table of Contents

Figures

Business Understanding

The Covid-19 pandemic brought the entire globe to a standstill. Covid-19 is a contagious disease which spreads quickly. The R_0 for COVID-19 is a median of 5.7, which means it can spread to 5-6 people from a person who has contracted the virus. The pandemic not only resulted in loss of life, and it also caused immeasurable economic losses, people lost their businesses, jobs, some industries were labeled as non-essential, so they were not allowed to operate at all. It put immense pressure on the global medical infrastructure, health care workers were overworked and exhausted throughout the globe and resulted in shortage of life saving critical medical equipment. All of this was done with the aim of lowering the mortality rate.

This was a once in a lifetime event that previously happened almost a century ago. There was a lack of understanding of diseases like we have today, and we didn't have data mining tools then. The entire global leadership was struggling to find ways to mitigate this disaster during the peak phase of the pandemic, but slowly we used these tools we were missing last time to flatten the curve. Data Mining techniques to visualize data and the effects of the pandemic helped governments, NGO's or think tanks to come up with ways to handle and minimize the impact of the pandemic. While the medical experts were trying to fight disease, Computer scientists and

Data scientists were trying to study the data and help make general and government policy decisions to help minimize the effects until a solution could be found, which would help end the pandemic and make it manageable by developing a vaccine or a drug.

This project report is meant to explore the census covid dataset and find clusters of different counties in the US. And the domain we are interested in is Economic Science; wealth distribution among population, earning and consumption, employment industry, cost of living, etc. are some of the topics we were interested in exploring.

We will explore multiple features from the provided datasets and use clustering. We will use K-means clustering, Hierarchical clustering and Partitioning around medoid algorithms on economic and employment related features and wish to find patterns in the data and uncover hidden insights in the data. The aim is to only use features and learn about the data without any labels, unsupervised learning.

The principal objective of this report is to partition counties based on different features, who are more at risk of contracting or dying from Covid-19. We look at data about people living in poverty, counties whose per capita income is low, and then later look at the data about covid cases and mortality caused by Covid-19 to these counties.

Data Description and Feature Selection

The US_covid-19_census data set is a comprehensive data set with various features and applications well beyond the impact of COVID-19. We were interested and focused on several features from the whole dataset, so we picked features for clustering and used them in two different ways i.e. Economic dataset and the Employment dataset. In the Economic dataset, we used features such as median_income, income_per_capita, median_rent, gini_index, Households_public_asst_or_food_stamps, poverty, etc. for performing the clustering techniques. Similarly, In the employment dataset, we used employed_agriculture forestry_fishing_hunting_mining, employed_construction, employed_instruction, etc. We used these features because these features helped us clearly identify how counties responded to the COVID-19 virus and also helped to find the counties that were immensely affected by the virus. Moreover, Data quality was good, and no missing values were found for the features used.

We selected two different kinds of features from the dataset and did clustering with different algorithms on these features separately. The feature types are -

- **Economic features** - We wanted to cluster the counties on the basis of economic features, such as terms of the economic indicators and then we wanted to divide(cluster) counties based on data about people employed in different kinds of settings. The features in this subset were - median_income, income_per_capita, gini_index, median_rent, households_public_asst_or_food_stamps and poverty.
- **Employment-related features** - We wanted to cluster the counties on the basis of the employment industry of people in various counties. The reason behind this is that there was a huge debate at the time about different kinds of job industries. Certain people were able to work from home, while some were being forced to return to office, some were doing essential jobs and also a lot of people lost their businesses. The features in this subset were - employed_agriculture_forestry_fishing_hunting_mining, employed_arts_entertainment_recreation_accommodation_food, employed_construction, employed_education_health_social, employed_finance_insurance_real_estate, employed_information, employed_other_services_not_public_admin, employed_manufacturing, employed_public_administration, employed_retail_trade, employed_science_management_admin_waste, employed_transportation_warehousing_utilities, employed_wholesale_trade.

Column Names and their characteristics:

employed_agriculture_forestry_fishing_hunting_mining(Scale: Ratio): Numerical data representing the population employed in the area of agriculture, forestry, fishing, hunting, and Mining. It's on a ratio scale, allowing for meaningful comparisons.

employed_arts_entertainment_recreation_accommodation_food(Scale: Ratio): A numerical column indicating the employed population in arts, entertainment, recreation, accommodation, and food. It's on a ratio scale, allowing for meaningful comparisons and calculations.

employed_construction(Scale: Ratio): A numerical data column in the area of construction and it allows comparisons because of ratio scale.

employed_education_health_social(Scale: Ratio): A numerical column indicating the employed population in education, health, and social. It's on a ratio scale, allowing for meaningful comparisons.

employed_finance_insurance_real_estate(Scale: Ratio): Numerical columns indicating the employed populations in finance, insurance, and real estate. It’s on a ratio scale, allowing for meaningful comparisons and calculations.

employed_information(Scale: Ratio): Numerical column information of the population who are employed. It is on a ratio scale.

employed_manufacturing(Scale: Ratio): This column likely contains numerical values representing the employed population in manufacturing. It's on a ratio scale as it has a true zero point, and ratios between values are meaningful.

employed_other_services_not_public_admin(Scale: Ratio): A numerical data column of the other services employed not public admin and it allows comparisons because of ratio scale.

employed_public_administration(Scale: Ratio): Numerical column of the population who are employed in public administration. It is on a ratio scale.

employed_retail_trade(Scale: Ratio): Numerical column employed in retail trade and It's on a ratio scale, allowing for meaningful comparisons and calculations.

employed_science_management_admin_waste(Scale: Ratio): Numerical data representing the ratio scale employed in the area of science, management, admin, and waste. It's on a ratio scale, allowing for meaningful comparisons.

employed_transportation_warehousing_utilities(Scale: Ratio): Numerical column indicating the employed in transportation, warehousing utilities and It's on a ratio scale, allowing for meaningful comparisons and calculations.

employed_wholesale_trade(Scale: Ratio): This column represents the ratio scale data of employed in the area of wholesale trade. It's on a ratio scale, allowing for meaningful calculations.

confirmed cases(Scale: Ratio): This numerical column represents counts of confirmed COVID-19 cases. The data is on a ratio scale, meaning it has a true zero point, and ratios of values are meaningful (e.g., one county having twice as many cases as another).

deaths (Scale: Ratio): This column represents counts of COVID-19-related deaths. It's also on a ratio scale.

median_income(Scale: Ratio): A numerical column representing the median income of residents in each county. It's on a ratio scale, allowing for meaningful comparisons and calculations.

income_per_capita(Scale: Ratio): This numerical column represents the income per capita in each county. It's on a ratio scale.

median_rent(Scale: Ratio): A numerical column representing the median rent in each county. It's on a ratio scale, providing a true zero point and meaningful ratios between values.

gini_index: A numerical column indicates the gini index in each county. The Gini index of 0 represents perfect equality, while an index of 1 implies perfect inequality.

households_public_asst_or_food_stamps(Scale: Ratio): This numerical column represents the households of public asst or food stamps in each county. It's on a ratio scale.

poverty(Scale: Ratio): Numerical data column of poverty in each county. It's on a ratio scale, allowing for meaningful comparisons and calculations.

nonfamily_households(Scale: Ratio): It represents the numerical column of non-family households in each county and it's on a ratio scale.

family_households(Scale: Ratio): It represents the numerical column of family households in each county and it's on a ratio scale.

Statistical Summary

A statistics summary gives information about the data in a sample. It can help understand the values better. It may include the total number of values, minimum value, and maximum value, along with the mean value and the standard deviation corresponding to a data collection. With this, you can understand the trends, outliers, and distribution of values in a data set. This is especially useful when dealing with large amounts of data as it can help in analyzing the data better. This information can be utilized to steer the rest of the analysis and derive more information about a data set. These are values that are calculated based on the sample data and do not go beyond the data on hand.

A statistical summary of COVID-19 encapsulates critical data points essential for understanding the impact of the virus. This includes the total number of confirmed cases and deaths, providing a basic overview of the severity of the outbreak. Economic feature information such as median_income, median_rent, income_per_capita, gini_index, and poverty data sheds light on how the virus affects diverse populations in various counties. Moreover, Employment related features like employed_agriculture_forestry_fishing_hunting_mining, employed_arts_entertainment_recreation_accommodation_food, employed_construction, employed_education_health_social, employed_finance_insurance_real_estate, employed_information, employed_other_services_not_public_admin,

employed_manufacturing, etc provide context to understand the employment situation in the Covid times in different kinds of job industries.

Feature Columns:

Column names	Scale	Mean	Median	Min	Max
employed_agriculture_forestry_fishing_hunting_mining	Ratio	896.9	477.5	0.0	74082.0
employed_arts_entertainment_recreation_accommodation_food	Ratio	4642.5	863.0	0.0	535155.0
employed_construction	Ratio	3044	811	0.0	276879
employed_education_health_social	Ratio	11070	2448	2	989093
employed_finance_insurance_real_estate	Ratio	3153.5	443.5	0.0	294319.0
employed_information	Ratio	1010	133	0.0	213966
employed_manufacturing	Ratio	4926.0	1420.5	0.0	475291.0
employed_other_services_not_public_admin	Ratio	2346	516	0.0	289152
employed_public_administration	Ratio	2236	534	0.0	152274

employed_retail_trade	Ratio	5463.7	1283.5	2.0	506091.0
employed_science_management_admin_waste	Ratio	5410.9	668.0	0.0	614276.0
employed_transportation_warehousing_utilities	Ratio	2444.8	579.5	0.0	270211.0
employed_wholesale_trade	Ratio	1287	241	0.0	165717
confirmed_cases	Ratio	7558.9	1916.5	0.0	1002614.0
deaths	Ratio	124.8	31.0	0.0	13936.0
median_income	Ratio	49754	48066	19264	129588
income_per_capita	Ratio	26040	25273	9334	69529
median_rent	Ratio	563.4	510.5	140.0	1879.0
gini_index	Ratio	0.4448	0.4423	0.3271	0.5976

households_public_asst_or_food_stamps	Ratio	5032.4	1480.5	0.0	333729.0
poverty	Ratio	14529	4120	10	1688505
nonfamily_households	Ratio	12898	3168	12	1091276
family_households	Ratio	24920	6604	15	2203922

Distance Measures

Many Clustering approaches use distance measures to assess the similarity or difference between pairs of data points. The distance measures depend on the type and nature of the dataset. In the case of COVID-19 data, we can measure the distance based on features such as deaths and total cases.

There are several ways to calculate the distance measures such as Euclidean distance and Manhattan distance.

Euclidean distance(Partition Around Medoids) is used for continuous variables i.e. total deaths and total COVID cases. For example, to calculate the Euclidean distance between different counties based on total cases and deaths.

It is an ordinary distance calculated between two points. It is one of the most used methods in cluster analysis. It is calculated by the root of squared differences between two points in a coordinate.

In Partition Around Medoids, random K medoids are chosen and the distance between points is calculated using Euclidean distance to form clusters. Then we change the medoids based on SSE (sum of squared error). These steps are repeated in the algorithm until the SSE is minimal and optimal clusters are achieved.

where coordinates x and y of a point are the pc1 and pc2 generated using the economic data set features.

The formula of the Euclidean distance is -

$\sqrt{(x2 - x1)^2 + (y2-y1)^2}$

where x2 = total cases of county 2 and x1 = total cases of county 1

y2 = total deaths of county 2 and y1 = total deaths of county 1

Manhattan distance(Partition Around Medoids) is also similar but more appropriate than euclidean distance as it involves absolute differences.

We have replaced the Euclidean distance with Manhattan distance here to calculate the distance and assign each data point to its nearest medoid. Manhattan distance is always suitable for data in grid structure which calculates the distance along axes.

Manhattan distance given two data points (x1,y1) and (x2,y2) in a grid two dimensional space is given as :

$|x2 - x1| + |y2 - y1|$

where coordinates x and y of a point are the pc1 and pc2 generated using the economic data set features.

Ward Linkage function(Hierarchical):

This function is mainly used in Hierarchical clustering algorithms to reduce the variance within the cluster. The ward linkage function goal is to minimize the increase in total variance within the merged cluster. The merge occurs to minimize the squared error or variance between the data points within the combined cluster. Here the data points references to the coordinate points of PC1 and PC 2 of the economic features. The variance of these data points are calculated and checked for minimum variance to merge the two clusters.

ward linkage is calculated as follows :

Increase in variance $= \sum(i \in C1 \cup C2) \; || xi - m ||^2$

The new mean m is calculated as the weighted average of the individual means m 1 and m 2 based on the sizes of the clusters

new mean $m = (n1m1 + n2m2) \; / (n1 + n2)$

Ward linkage $= ((n1n2) / (n1 + n2)) * || m1 - m2 || ^2$

Average linkage(Hierarchical) :

Average linkage is another method used in Hierarchical clustering. This method is different from the Ward linkage method. In this method dissimilarity between two clusters is calculated by finding the average of pairwise distance between the data points in the cluster. The distance is calculated between pairs of points one from each cluster using standard techniques like Euclidean or Manhattan distance. Average is calculated for all these distances. Using these averages, merging of clusters is done accordingly.

Average Linkage method Formula :

$(1 / n1 . n2) . \sum i(1,n1) \sum j(1, n2)$ distance(xi1,xj2)

where n1 and n2 are the number of points in the two clusters, and xi1 and xj2 are the i-th point from the first cluster and j-th point from the second cluster, respectively. Average linkage tends to produce more balanced and spherical clusters compared to other linkage methods

Verifying Data Quality

Ensuring data quality is a careful process that checks for accuracy, completeness, and reliability in a dataset. It requires thorough validation and cross-referencing of each data point using trustworthy sources to confirm its accuracy.

For columns like confirmed_cases, deaths, median_income, income_per_capita, median_rent, households_public_asst_or_food_stamps, poverty, nonfamily_households, family_households we have dropped NA values and used averages for columns containing absolute values by 1000* total population.

employed_agriculture_forestry_fishing_hunting_mining: This column reveals the distribution of employment ratios in agriculture, forestry, fishing, hunting, and mining across various counties. Each county is uniquely identified, with distinct values representing the workforce engagement in these specific sectors. The numeric ratios are complete, providing insights into the employment landscape related to primary industries.

employed_arts_entertainment_recreation_accommodation_food: In this column, unique values represent the employment ratios in arts, entertainment, recreation, accommodation, and food services for each county. There are no duplicates, ensuring accurate identification. The numeric ratios are complete, offering a comprehensive view of the employment patterns in these sectors across diverse counties.

employed_construction : column illustrates the employment ratios in the construction sector for different counties. Each entry is distinct, signifying a unique county, and there are no missing values. The numeric ratios provide a clear picture of the workforce engagement in the construction industry, contributing to a comprehensive understanding of employment dynamics.

employed_education_health_social : Distinct values in the employed_education_health_social column represent employment ratios in education, health, and social services across various counties. Each county is uniquely identified, and the absence of missing entries ensures data completeness. The numeric ratios offer insights into the workforce distribution in critical service sectors, contributing to a nuanced understanding of employment trends.

employed_finance_insurance_real_estate : The column showcases employment ratios in finance, insurance, and real estate for each county. Uniqueness in entries and absence of missing values ensure data accuracy and completeness. The numeric ratios provide valuable insights into the distribution of employment within the financial and real estate sectors across diverse counties**.**

employed_information : Distinct values in the column represent employment ratios in the information sector for each county. Each entry uniquely identifies a county, and there are no missing values, ensuring data completeness. The numeric ratios offer insights into the employment landscape related to information services, contributing to a comprehensive understanding of workforce engagement.

employed_manufacturing : The column illustrates the distribution of employment ratios in the manufacturing sector across various counties. Each entry uniquely identifies a county, and there are no missing values. The numeric ratios provide insights into the manufacturing workforce, contributing to a comprehensive analysis of employment patterns.

employed_other_services_not_public_admin : Distinct values in the column represent employment ratios in other services, excluding public administration, for each county. Each entry is unique, and there are no missing values, ensuring data completeness. The numeric ratios offer insights into the distribution of employment within specified service sectors

employed_public_administration : The employed_public_administration column showcases employment ratios in public administration for each county. Uniqueness in entries and absence of missing values ensure data accuracy and completeness. The numeric ratios provide insights into the workforce engaged in public administration across diverse counties.

employed_retail_trade : Distinct values in the column represent employment ratios in the retail trade sector for each county. Each entry is unique, and the absence of missing values ensures data completeness. The numeric ratios offer insights into the distribution of employment within the retail trade industry, contributing to a comprehensive understanding of workforce engagement.

employed_science_management_admin_waste : The column illustrates the distribution of employment ratios in science, management, administration, and waste services across various counties. Each entry is unique, and there are no missing values. The numeric ratios provide insights into the workforce engaged in these specified sectors, contributing to a comprehensive analysis of employment patterns.

employed_transportation_warehousing_utilities : Distinct values in this column represent employment ratios in transportation, warehousing, and utilities for each county. Each entry uniquely identifies a county, and there are no missing values. The numeric ratios offer insights into the distribution of employment within these specified sectors, contributing to a comprehensive understanding of workforce engagement.

employed_wholesale_trade : This column showcases employment ratios in wholesale trade for each county. Uniqueness in entries and absence of missing values ensure data accuracy and completeness. The numeric ratios provide insights into the distribution of employment within the wholesale trade industry across diverse counties.

confirmed_cases : Distinct values in this column represent counts of confirmed COVID-19 cases for each county. Each entry is unique, and there are no missing values, ensuring data

completeness. The numeric ratios offer insights into the magnitude of confirmed cases across diverse counties.

deaths : The column illustrates counts of deaths due to COVID-19 for each county. Uniqueness in entries and absence of missing values ensure data accuracy and completeness. The numeric ratios provide insights into the magnitude of deaths across diverse counties.

median_income : Distinct values in this column represent median income figures for each county. Each entry is unique, and there are no missing values, ensuring data completeness. The numeric ratios offer insights into the income distribution across diverse counties.

income_per_capita : The income_per_capita column showcases income per capita ratios for each county. Uniqueness in entries and absence of missing values ensure data accuracy and completeness. The numeric ratios provide insights into the per capita income distribution across diverse counties.

median_rent : Distinct values in this column represent median rent figures for each county. Each entry is unique, and there are no missing values, ensuring data completeness. The numeric ratios offer insights into the rent distribution across diverse counties.

households_public_asst_or_food_stamps : Distinct values in the households_public_asst_or_food_stamps column represent ratios of households receiving public assistance or food stamps for each county. Each entry is unique, and there are no missing values, ensuring data completeness. The numeric ratios offer insights into the distribution of households receiving public assistance or food stamps across diverse counties.

poverty : The poverty column illustrates poverty ratios for each county. Uniqueness in entries and absence of missing values ensure data accuracy and completeness. The numeric ratios provide insights into the prevalence of poverty across diverse counties, contributing to a comprehensive analysis of socio-economic conditions.

nonfamily_households : Distinct values in this column represent ratios of non-family households for each county. Each entry is unique, and there are no missing values, ensuring data completeness. The numeric ratios offer insights into the distribution of non-family households across diverse counties, contributing to a comprehensive understanding of household structures.

Data Preprocessing -

Before we begin clustering, I want to inform the reader of this report that the report is organized on the basis of the two types of features that we extracted from the data. One is concerned about the economic indicators of a county, and the other one is about the people employed in different industries. So first we present Economic Features, use clustering algorithms to do unsupervised learning and then we move on to Employment Industry Features and use the same clustering algorithms we used previously for Economic features.

Scaling - We used the scaling function in R to scale the values of the selected features.

Outlier Removal - We used an isolation forest for outlier removal on both economic and employment feature types. Isolation forest is an anomaly detection algorithm. It isolates outliers by randomly selecting a feature and then it randomly selects a split value,
a split value which lies somewhere in between the maximum and minimum values of the selected feature. A score is assigned to each sample and we used that score to declare a threshold. The below table shows the summary of the scores column of the Economic features.

summary(scores) of isolation forest

MIN	1st Quartile	Median	Mean	3rd Quartile	MAX
6.785	20.256	21.894	21.147	22.922	24.164

Dimension Reduction - We used principal component analysis(PCA) on the economic features. We found that PC1 and PC2 retained quite a lot of information from the original 6 features that we selected so we decided to go ahead with it for this feature type.

PCA for Economic feature type

PC1	PC2	PC3	PC4	PC5	PC6
60.529709	21.116355	10.459585	4.205852	2.338064	1.350435

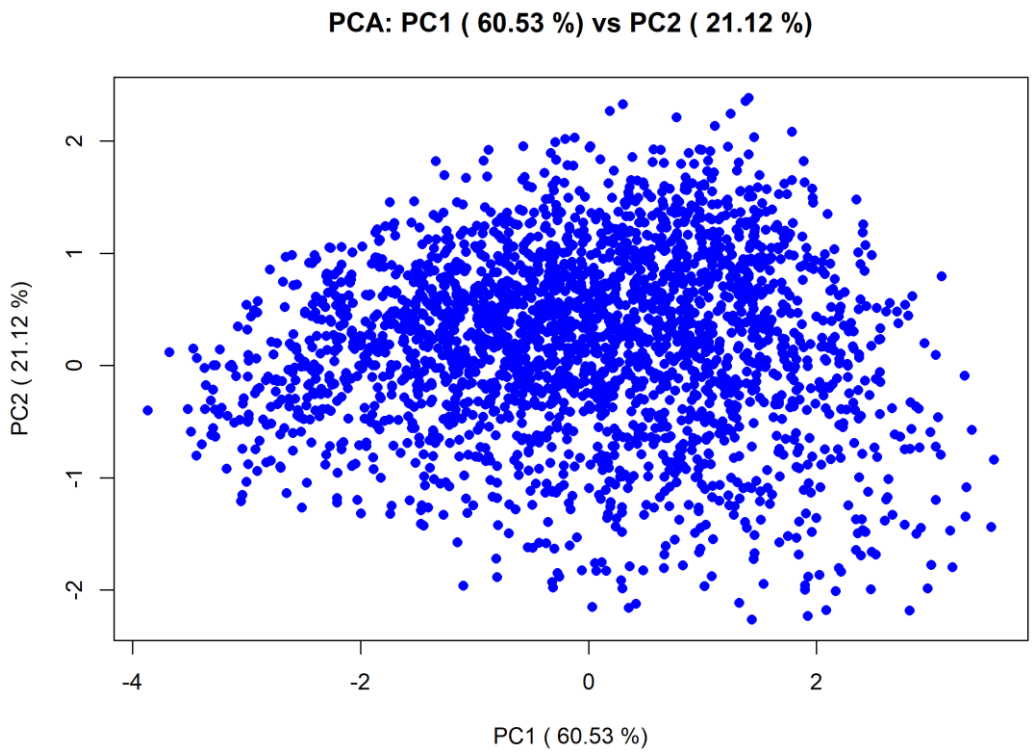


Fig-1: PCA on Economic Features

For the employment features we did do PCA dimension reduction but found the results to be unsatisfactory to move forward with this data preprocessing step for this dataset. PC1 and PC2 were able to retain approximately 81% of the information in the case of economic features, but it took Principal components 1-8 to achieve the same level of retention of information after dimension reduction. Below figure shows a scree plot, to retain 80% of the information after dimension reduction of Employment features we had to use 8 dimensions, for visualization and interpretation we decided to go with a different dimension reduction algorithm.

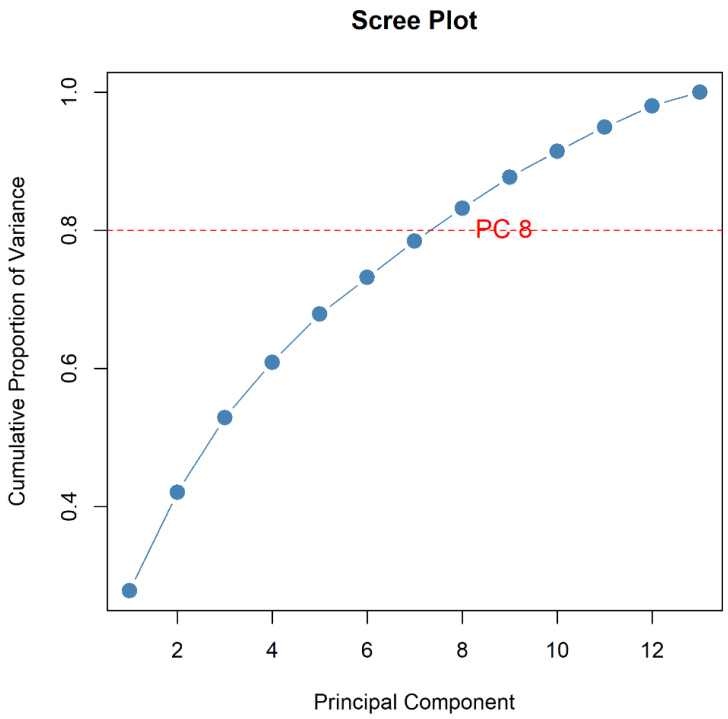


Fig-2: Scree Plot of Cumulative Variance vs Principal Component

PCA for Employment feature type

PC1	PC2	PC3	PC4	PC5	PC6	PC7
27.794511	14.280584	10.815815	7.999027	6.979652	5.288609	5.250046

PC8	PC9	PC10	PC11	PC12	PC13
4.796715	4.483991	3.753464	3.483187	3.077333	1.997064

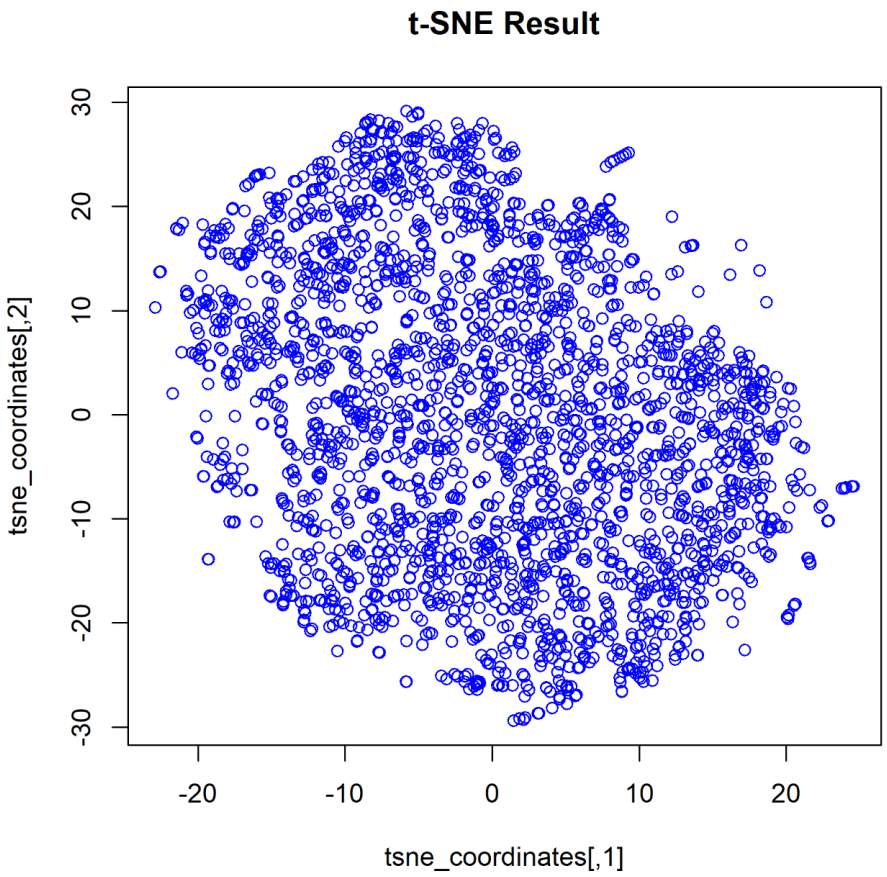


Fig-3: t-SNE on Employment industry features

We decided to use t-SNE (t-Distributed Stochastic Neighbor Embedding) for the Employment industry features. This algorithm works well for high dimensional data, and while PCA focuses on the data on a global scale, t-SNE focuses on preserving local similarities in the data. This was a useful advantage to have because we didn't know how well the clustering algorithms might work on the employment industry features.

Determining the number of clusters

Determining the number of clusters is one of the most important tasks in clustering. While we were sure we wanted two clusters from the Economic features, we were not so sure about the Employment industry features.

Economic Features -

The reason for being sure about two clusters for the economic features is that we wanted to divide counties into two groups on the basis of features that provide some information about the general economic condition and prosperity of the county. But still we decided to use statistical methods to validate our initial decision about choosing k=2.

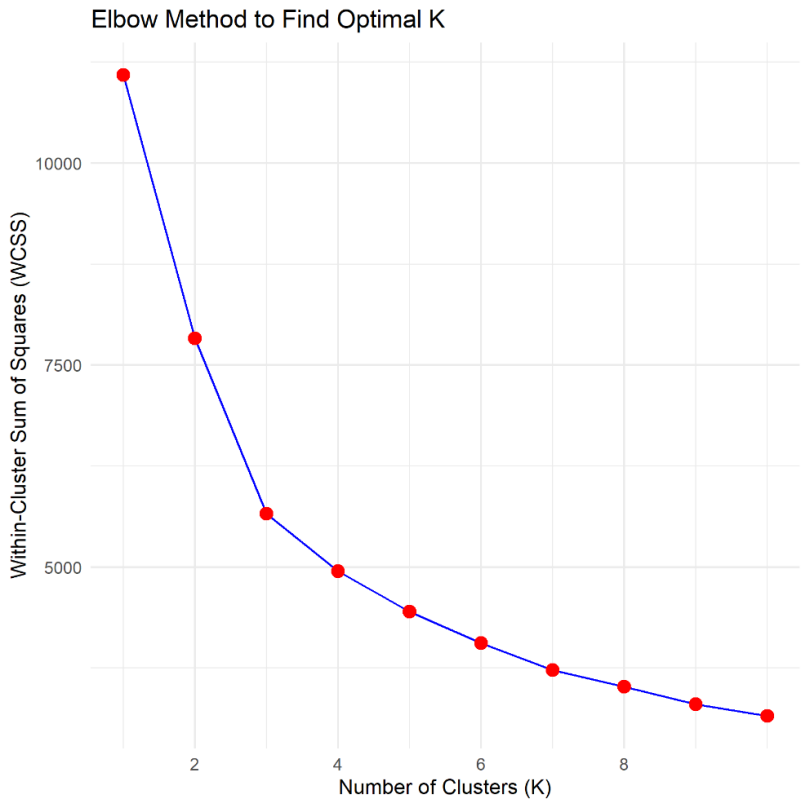


Fig-4: Elbow method to find optimal K

We used the elbow method to find the number of clusters. This method aims to provide a plot of WCSS, within the cluster sum of squares. This is the square of the distance between the centroid of the cluster and each point in the cluster. When there is a sudden drop in the WCSS we select that point on the x-axis as the number of clusters for our clustering task since after

increasing the number of clusters after that point, there is not a significant decrease in the rate of decrease of WCSS.

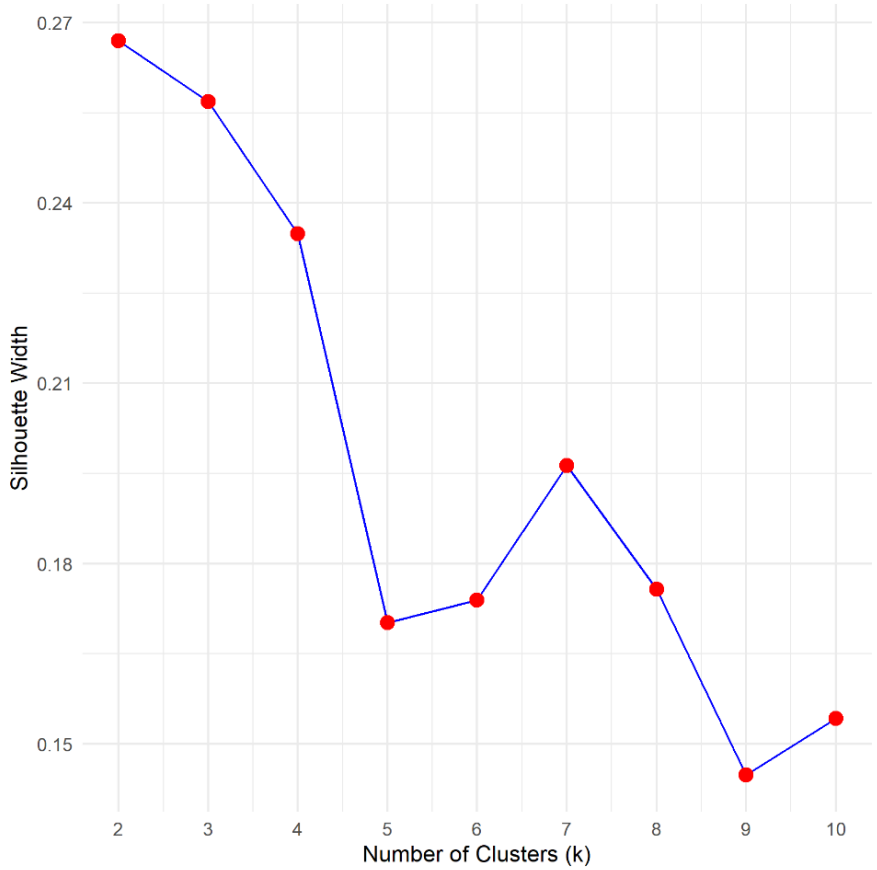


Fig-5: Silhouette Analysis to determine clusters

Silhouette analysis also helps in determining the number of clusters. It runs usually for a range of different values of k(number of clusters) and determines the similarity of points in each cluster for every value of k. Then using a matrix of silhouette scores for each value of k this above plot helps in determining the number of clusters. The highest value of the silhouette coefficient is found when k=2. This is also the number of clusters we had in mind when we set out to do clustering on these features.

Employment Industry Features -

We were not quite as sure about the employment industry features as we were about the economic indicators when we selected these features. This is in essence a pure unsupervised learning task that we undertook since for this data we didn't have a classification of 'rich' and 'poor' counties at the back of our minds.

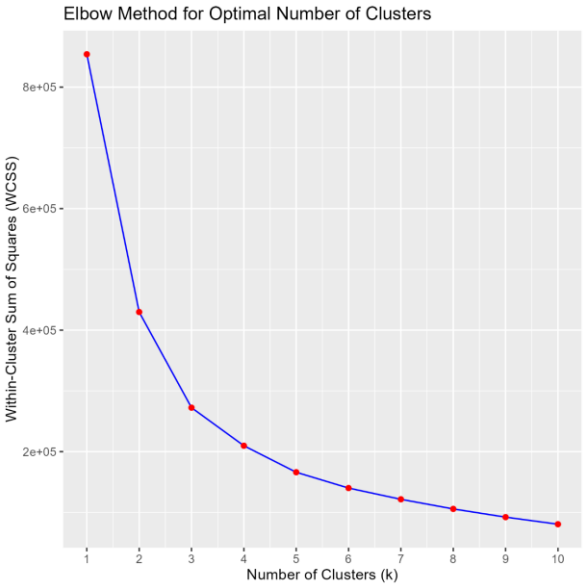


Fig-6: Elbow Method for Optimal Clusters

When we used the elbow method on the subset containing the employment industry features, we found the elbow point at k=3. The above figure shows that at k=3, there is a sudden drop in WCSS at that point.

We also used the silhouette method to determine the optimum number of clusters for this category of features. The below figure shows the plot of silhouette scores for different numbers of clusters. The maximum value for the silhouette score was at k=10, but we didn't want 10 clusters for 13 features. We feel that 10 is just too many clusters for this clustering and we wanted to avoid overfitting and wanted a more general model from our data and the selected clustering algorithm. So we decided to go with k=3, that is create 3 clusters for the employment industry features.

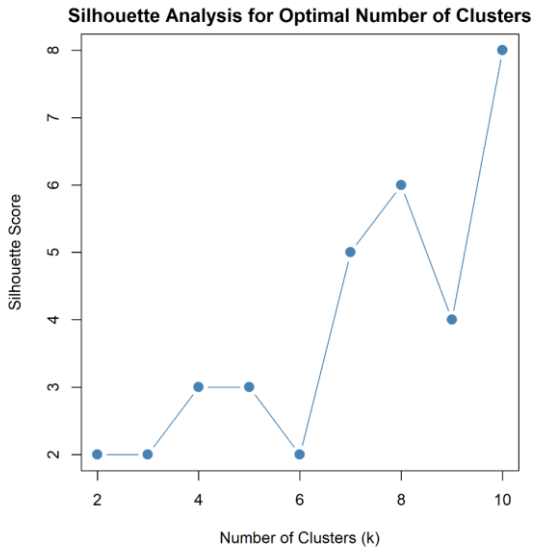


Fig-7: Silhouette Analysis for optimal number clusters

K-means clustering - Economic Features

We used the k-means () function in R for clustering our subset of data containing economic features. The below plot shows clustering done using the K-means clustering algorithm for the economic features. The means () method in R takes in two inputs, one is a matrix on which we did all the data preparation. We scaled the data, used an isolation tree to remove outliers from the data, and then used PCA for dimension reduction because visualizing so many dimensions would not have been feasible. After all these steps to prepare our data, we fed a matrix containing PC1 and PC2 values as columns to the means () function.

Another input for the function is the number of centers, this parameter takes in the number of clusters that we need in the data. From silhouette analysis, we know that we will get the most clear clustering as compared to any other value, and our requirement while undertaking this task of clustering was also the same. We wanted to make two clusters, which contained well-prepared data, to build a generalized model, which would classify a county as rich or poor based on the features we selected.

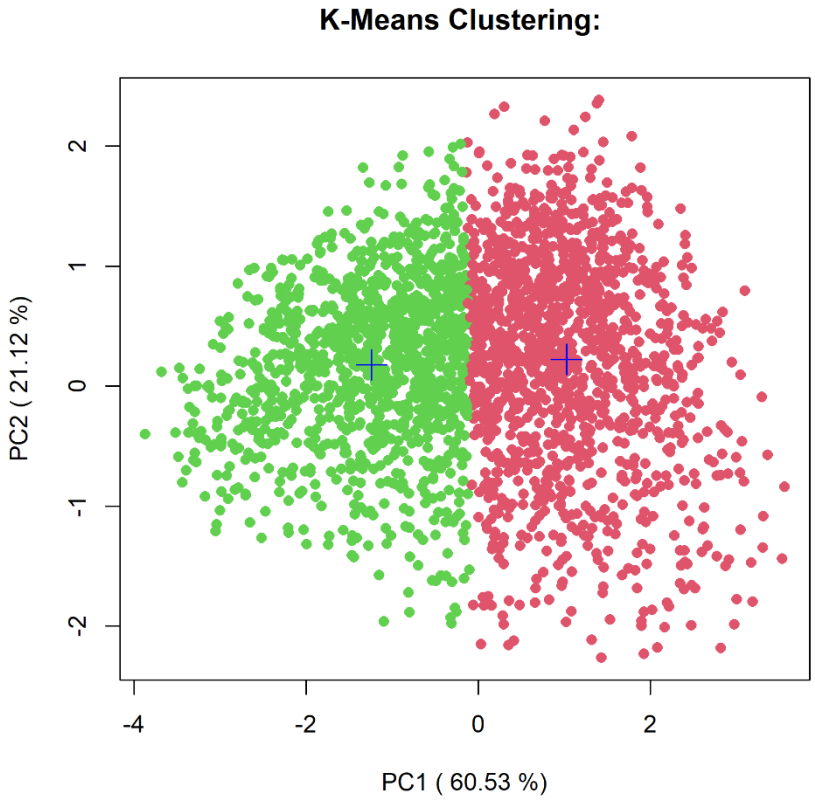


Fig-8: K-means clustering on economic features

The clustering plot above shows two clusters formed, indicated by green and red colors; blue points indicate the centroid of those clusters.

kmeans_result - Below are some of the results we got by doing K-means clustering on the Economic features.

```
K-means clustering with 2 clusters of sizes 1390, 1040


Cluster means:
        pc1        pc2
1  1.079497 -0.2406876
2 -1.313239 -0.1163619


Within cluster sum of squares by cluster:
[1] 2535.635 1758.999
 (between_SS / total_SS =  44.3 %)
```

Evaluating Clusters -

After we were finished with clustering the counties into two, we attached the cluster information to the original subset. Then on the basis of the formed cluster, we computed the average of the initially selected features such as median rent, median income, income per capita, and poverty. The below graph shows the average of these values for both clusters. We could not have been happier with the clusters formed because this clustering task gives out conspicuous patterns in the data.

If we observe the below graph, we find the features income per capita and median income features have a very high value, almost twice, in Cluster 2 compared to Cluster 1. Median rent is also high for these counties, these are counties where people earn way more than in the other cluster. Consequently, the cost of living indicator median rent is also approximately 1.5x in cluster 2 as compared to cluster 1.

When we look at the features such as poverty and households on food stamps, then we see the opposite, this is where cluster 1 has almost twice the values as compared to cluster 2.

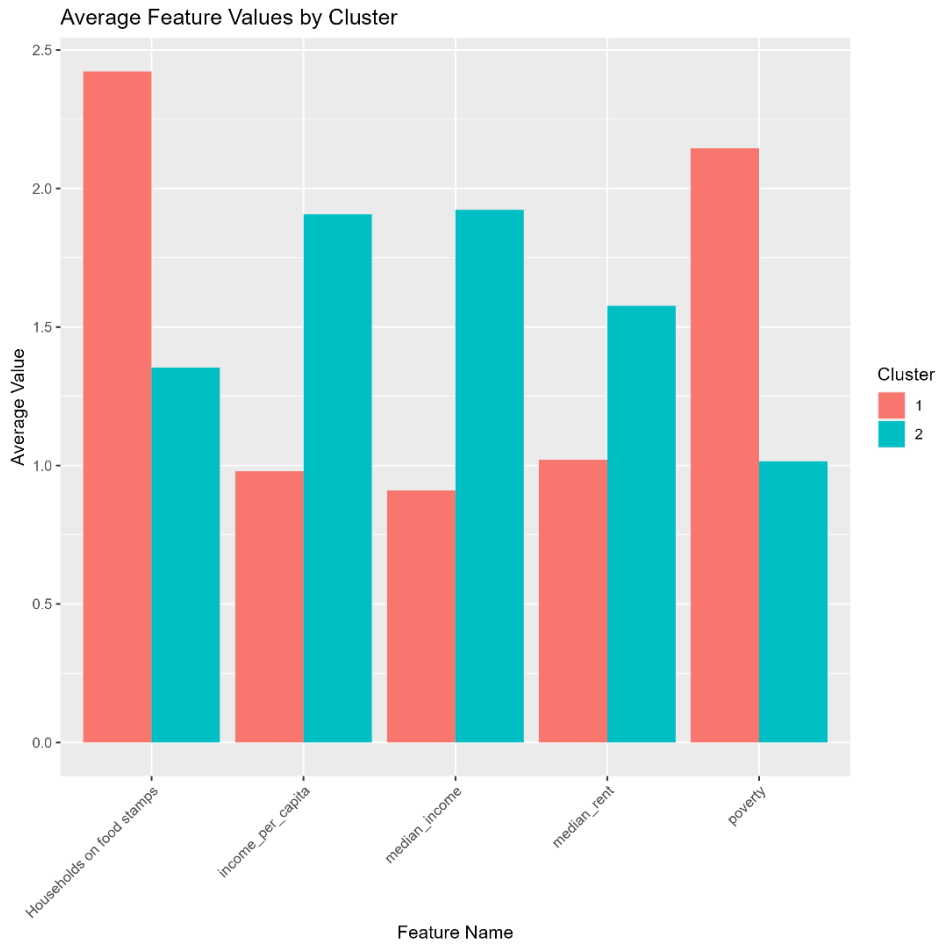


Fig-9: Average feature values by cluster

By this visual analysis, we find that our clustering was done quite efficiently since it divided counties based on economic features and successfully found a pattern in the data. The above figure clearly shows the difference between the two clusters. Cluster 2 contains the 'rich' counties, while Cluster 1 contains the 'poor' counties.

Since the dataset contains COVID information apart from census data, we were also interested in looking at the confirmed cases and deaths caused by COVID-19 for the two clusters that we created. If we look at the below figures. The bar graph on the left shows the number of covid cases per 1000 people and the other one shows the number of deaths due to covid per 1000 people. The findings are a little surprising, While the number of confirmed cases per 1000 people was higher in cluster 1 as compared to cluster 2. But for the number of deaths per 1000 people, it was the opposite. In this figure, we can see that the counties in Cluster 1 fared better than Cluster 2(rich) counties.

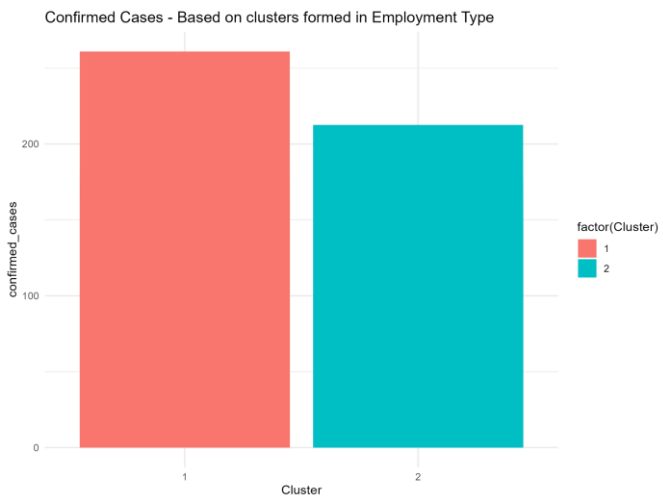


Fig-10: Confirmed cases in Employment type

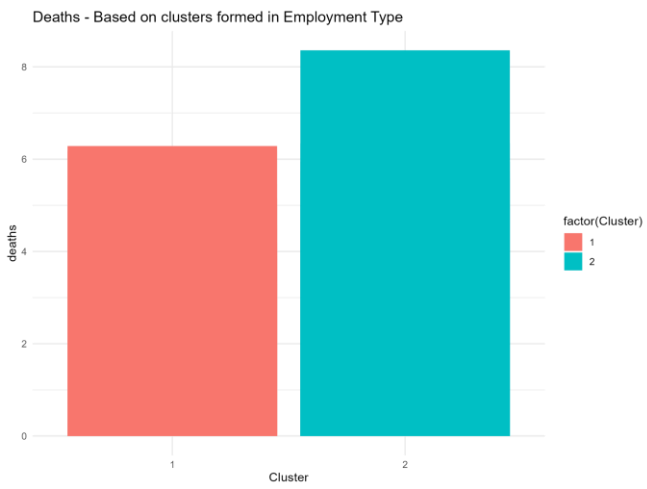


Fig-11: Total deaths in Employment type

Hierarchical Clustering – Economic Features

The hierarchical clustering works by treating every sample as a separate cluster. And then it computes the similarity and dissimilarity using a metric such as Euclidean, Manhattan, etc. This metric is also known as linkage. We experimented with various linkages such as euclidean, manhattan, averages, Ward function, correlation etc.

We found that Ward and Average linkages performed the best to form clusters for the economic features. So we proceeded with these two methods.

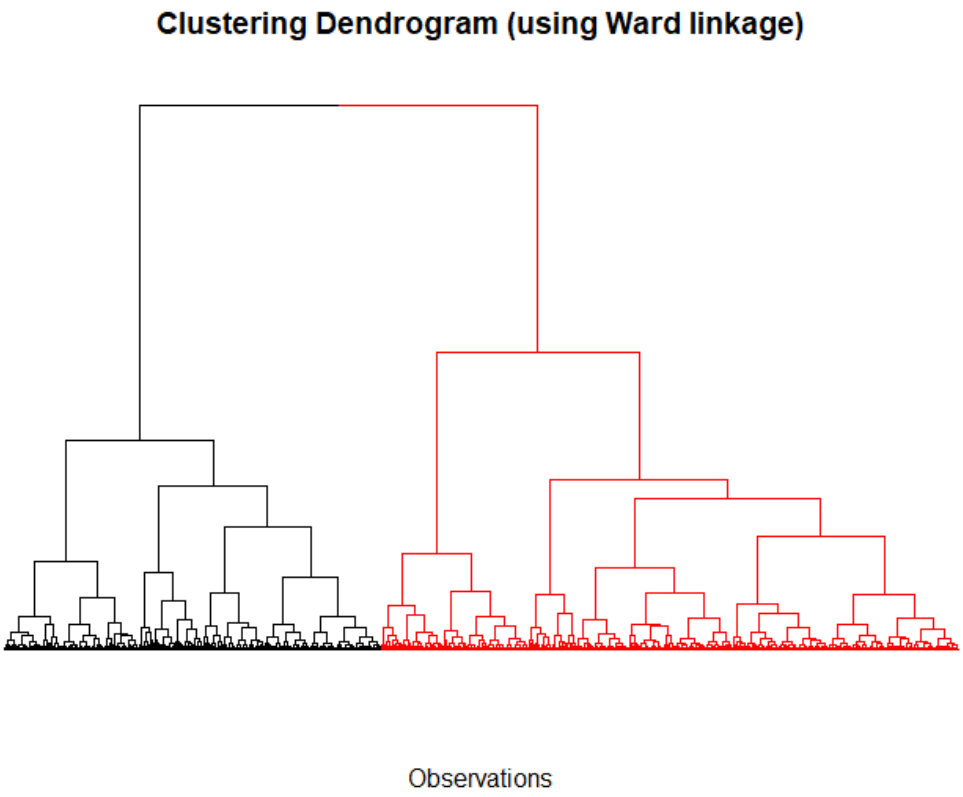


Fig-12: Clustering Dendrogram using Ward linkage

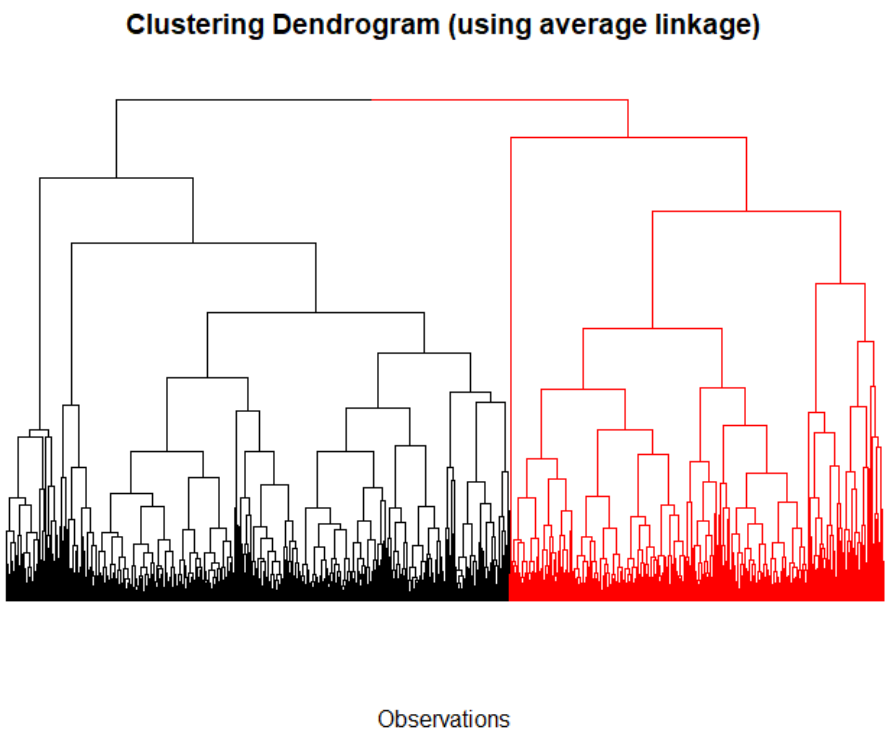


Fig-13: Clustering Dendrogram using average linkage

When we look at the below visuals to analyze the clusters using both Average and Ward Linkages we find there is only a slight difference in the clusters for counties near the center of the plot.

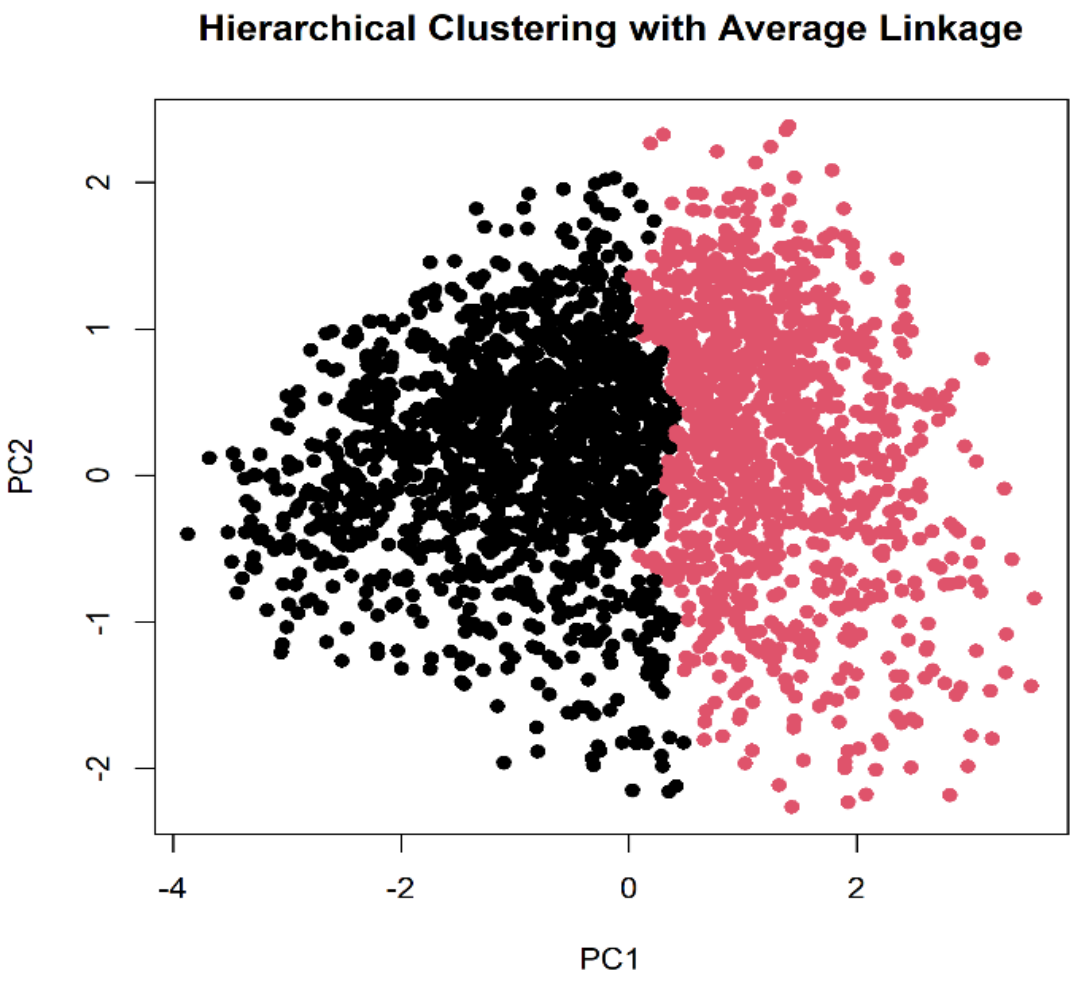


Fig-14: Hierarchical Clustering with Average linkage

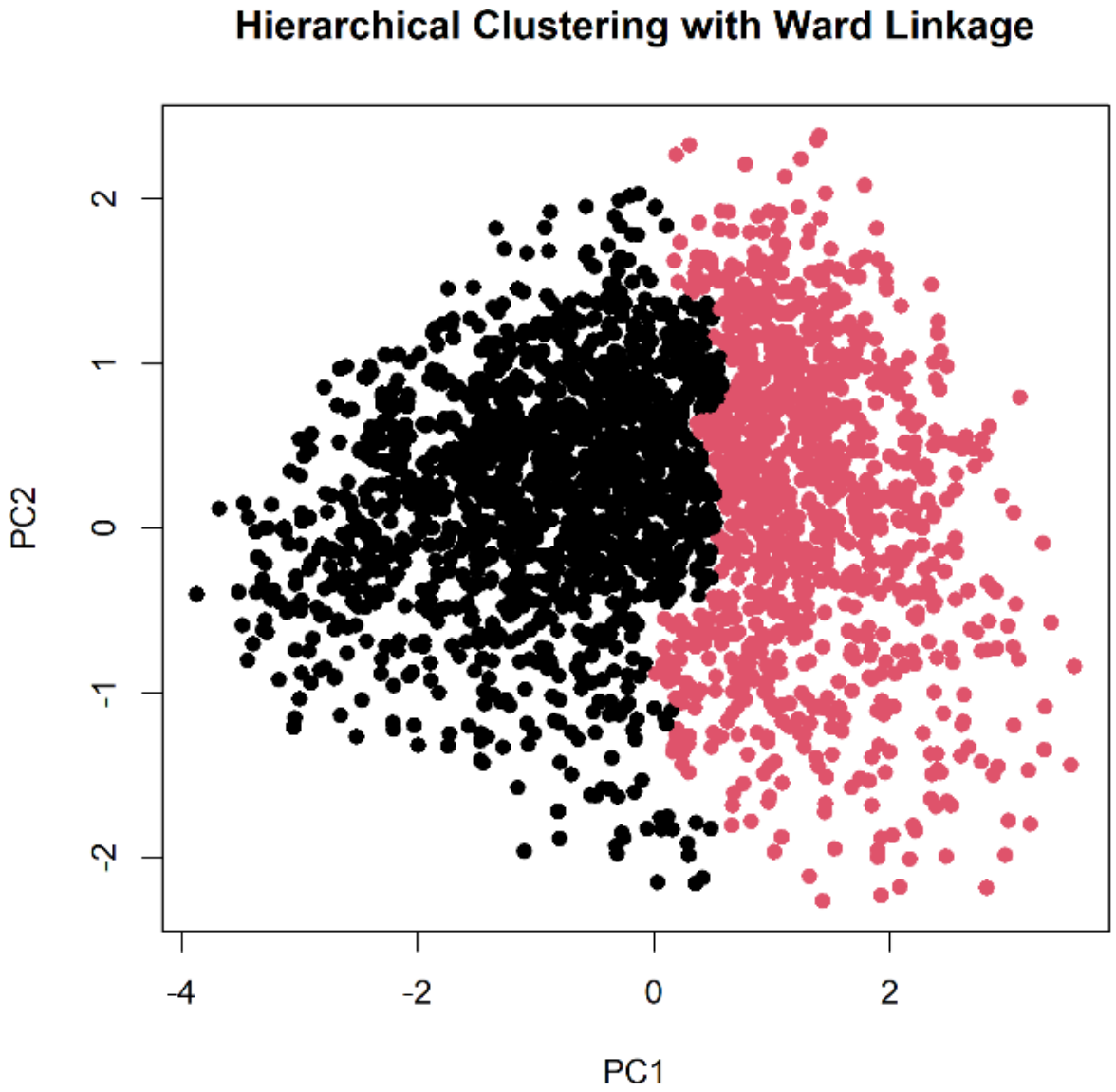


Fig-15: Hierarchical Clustering with Ward linkage

After looking at the clusters created by both the Average and Ward linkage methods, we look at the features and comparison of clusters, for both the Hierarchical clusters we created using the Average and Ward linkage methods.

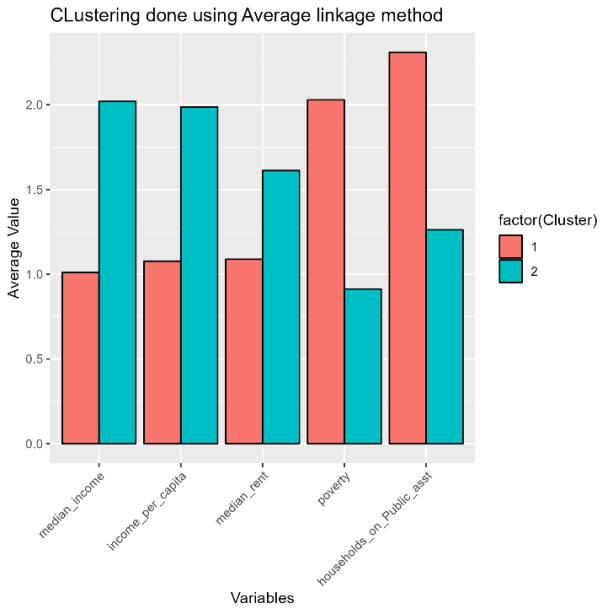


Fig-16: Clustering done using Average linkage

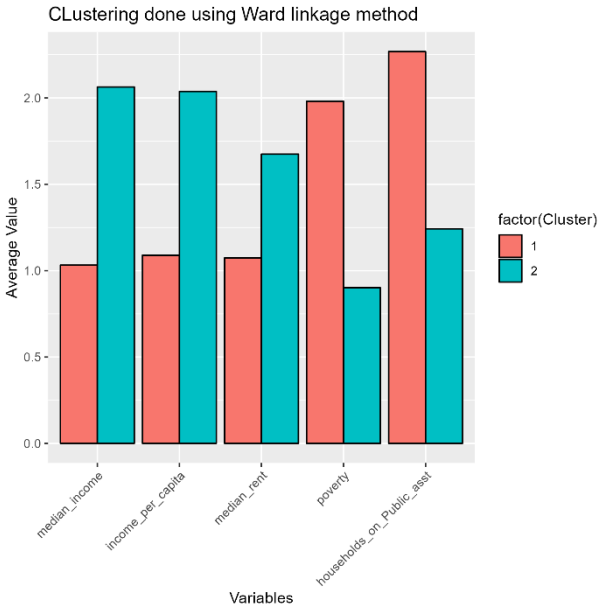


Fig-17: Clustering done using Ward linkage

When we analyze the above bar graphs individually, we find that Cluster 1 has more poverty and households depending upon government assistance and food stamps. While Cluster 2 does well in terms of median and per capita income. Also, the median rent(cost of living) is higher in Cluster 2.

The above observation is true for both the linkage methods, Average and Ward. If we compare the two methods on how clearly the two divided counties in the US, then there is little to separate the two linkage methods.

Now we want to look at the graphs below, to look at how the 'rich' and 'poor' counties did in terms of covid cases and deaths due to Covid.

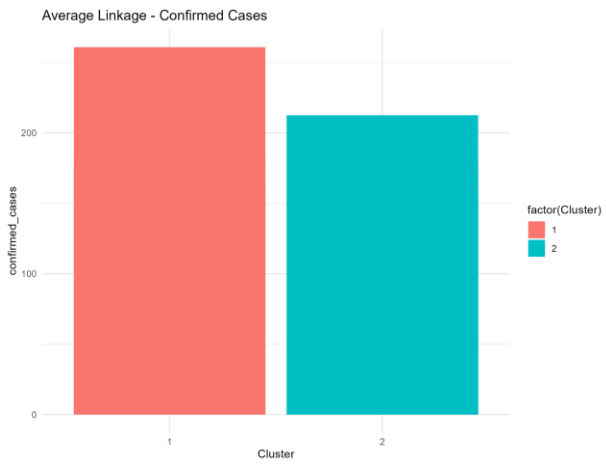


Fig-18: Confirmed cases vs cluster using average linkage

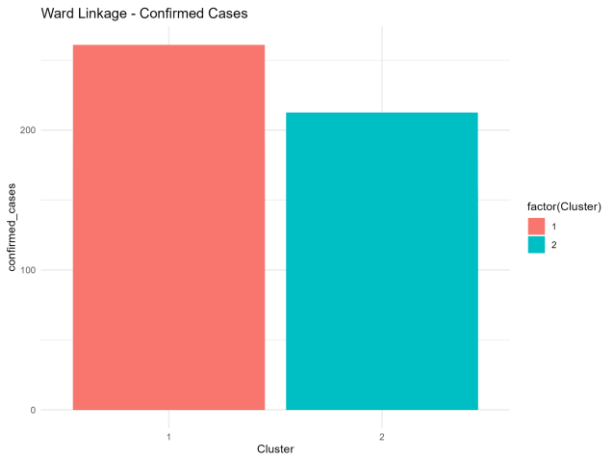


Fig-19: Confirmed cases vs cluster using ward method

Again, we find there is little difference between the results of the two linkage methods in terms of total covid cases and deaths due to covid. Figures above show the number of covid cases were higher in Cluster 1 in both the cases, while using Average and Ward linkages. But we find the same pattern, what we saw in k-means clustering. The confirmed cases per 1000 people in 'poor'(Cluster 1) counties are higher while the death rate per 1000 people is higher in 'richer'(Cluster 2) counties. Figures below shows that Cluster 2 has a higher death rate due to covid.

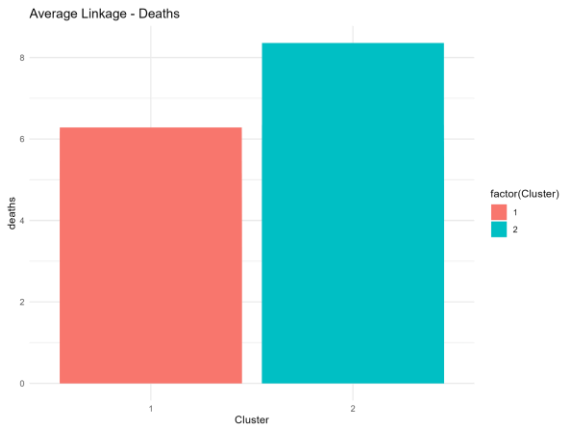


Fig-20: Deaths vs cluster using average linkage

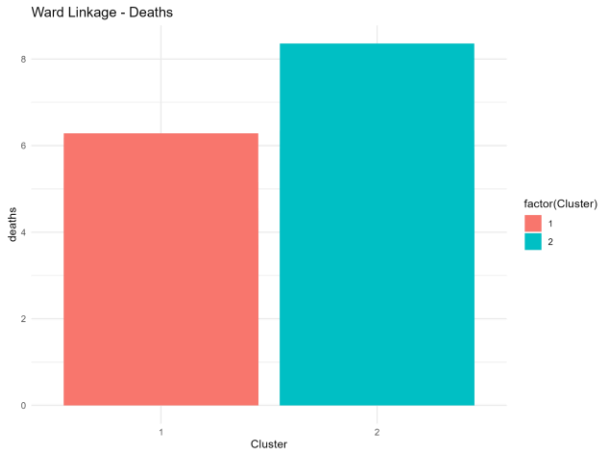


Fig-21: Deaths vs cluster using ward method

PAM(Partition around Medoids) - Exceptional Work

While trying to find the third clustering algorithm for exceptional work we tried a lot of different clustering algorithms. We used DBSCAN as well and hoped to find good results from it but that didn't work out as expected.

So we tried Partitioning around medoids clustering method. This algorithm is similar to K-means in terms of the procedure of the algorithm. But a major difference is that PAM uses medoids instead of centroids like K-mean clustering does. Unlike the centroid, the medoid is an actual sample in the data.

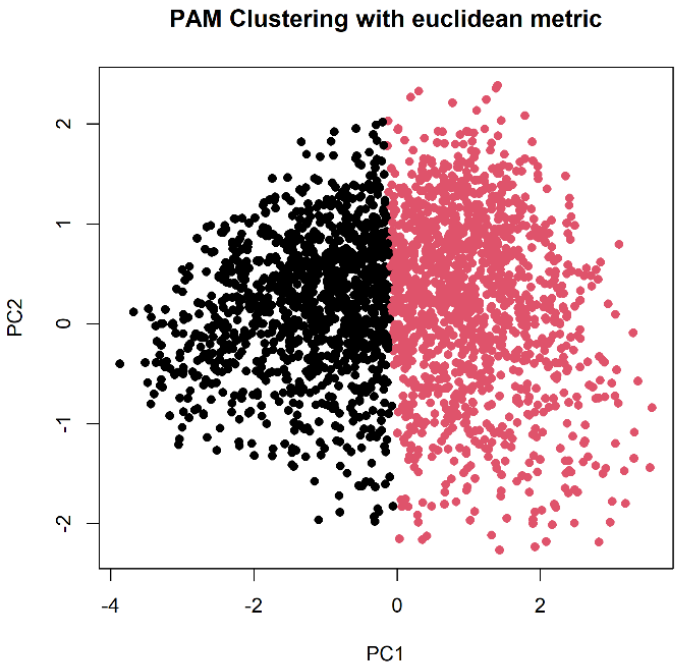


Fig-22: PAM Clustering with Euclidean metric

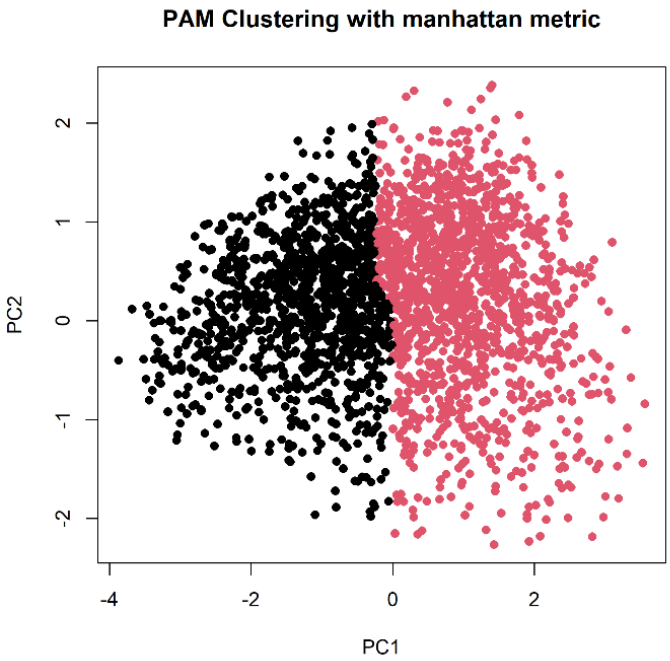


Fig-23: PAM Clustering with Manhattan metric

We used Euclidean and Manhattan metric to form clusters in the Economic features data. The above two figures show how the clusters are formed using the two methods. There is only a slight difference in the middle of the plot where some counties might differ in terms of which cluster they belong to using the two distance measures. But overall, the clustering and the model looks generalized, with no overlaps.

Now we will take a look at the figures below to take a deep dive into the features in both the clusters.

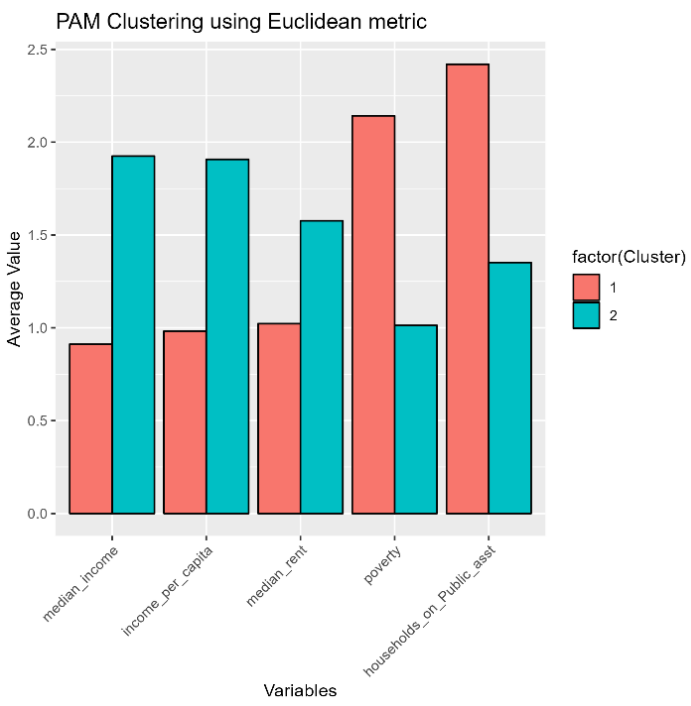


Fig-24: PAM Euclidean metric(Avg value vs Variables)

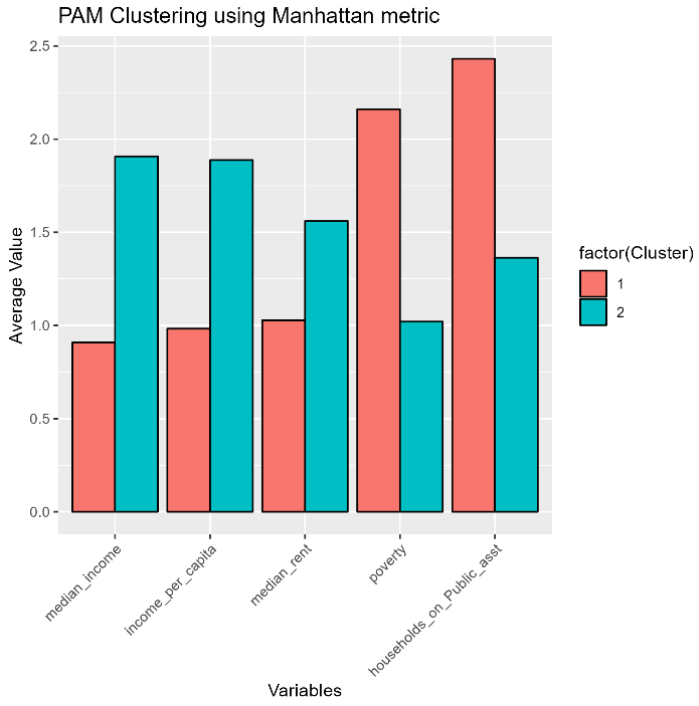


Fig-25: PAM Manhattan metric(Avg value vs Variables)

The above figures show the clusters and the average values of the features. We see the same pattern again, features related to poverty, median income and per capita income are higher in Cluster 2. And Cluster 1 is higher in poverty and households depending upon assistance. We again find the clusters divide counties into rich and poor categories.

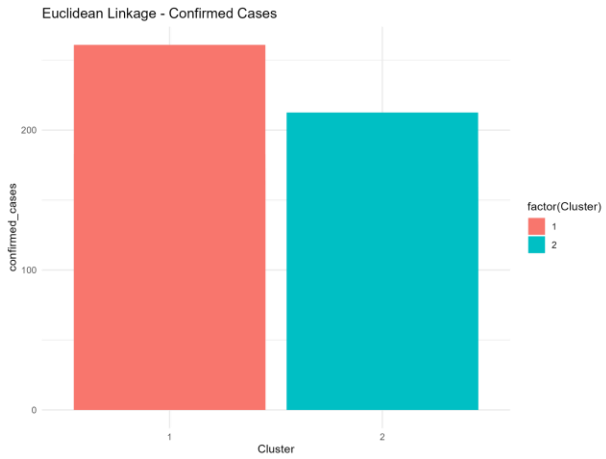


Fig-26: Euclidean linkage Confirmed cases

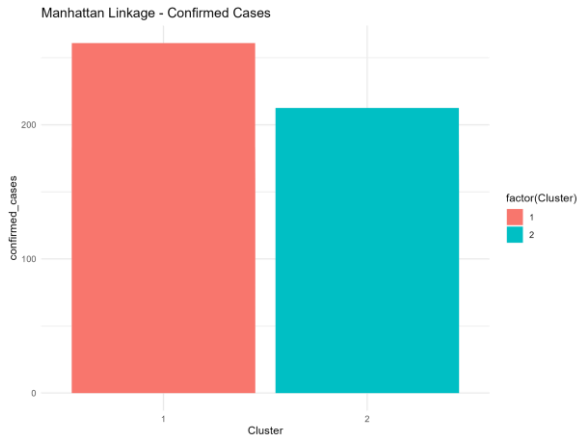


Fig-27: Manhattan linkage Confirmed cases

Now we look at the confirmed cases and death statistics of the counties which are divided into two clusters based on euclidean and Manhattan distance. Visually there is little we can say about the comparison of the two different clusters that we made using manhattan and

Euclidean distances. There is little difference between the clusters formed by both distance metrics.

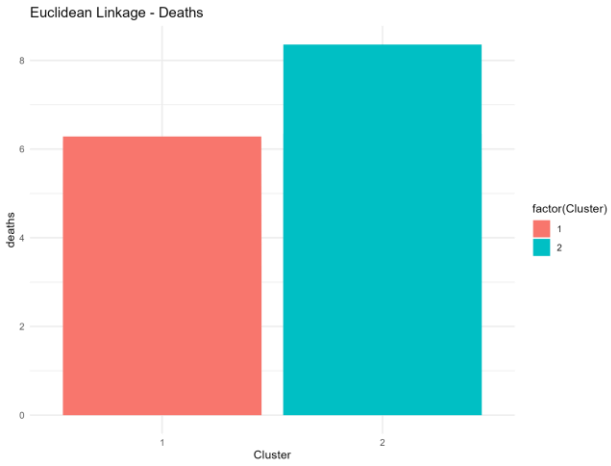


Fig-28: Euclidean linkage Deaths vs clusters

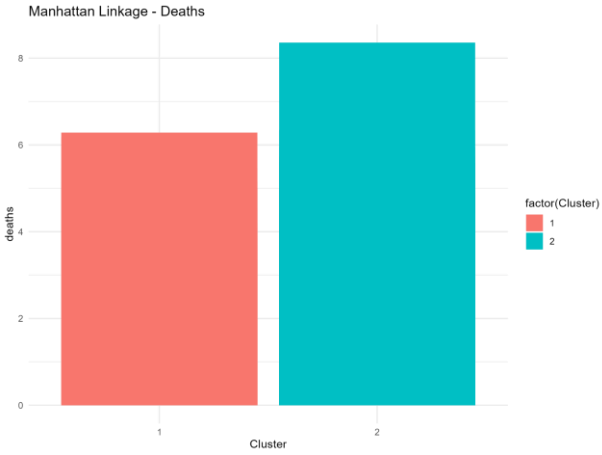


Fig-29: Euclidean linkage Deaths vs clusters

We find the same pattern with this clustering method, the number of cases per 1000 people is higher in Cluster 1(poor counties), but the number of deaths per 1000 people is higher in Cluster 2(richer counties).

K-means Clustering - Employment Data

We wanted to use clustering on features that were concerned about the employment industry of the residents of a county. So, we took these features to create an Employment Industry subset. We had planned on using the same data preprocessing pipeline created for the Economic features, but the PCA analysis failed to do the job for these features. We went ahead with t-SNE dimension reduction.

To determine the number of clusters for this dataset, we used silhouette analysis and found varying answers every time. On each run the silhouette coefficients had a different value for optimum k. Below is a clustering that we did on Employment Industry features, based on the optimum value of k=8, which is one of the values we got by silhouette coefficient analysis.

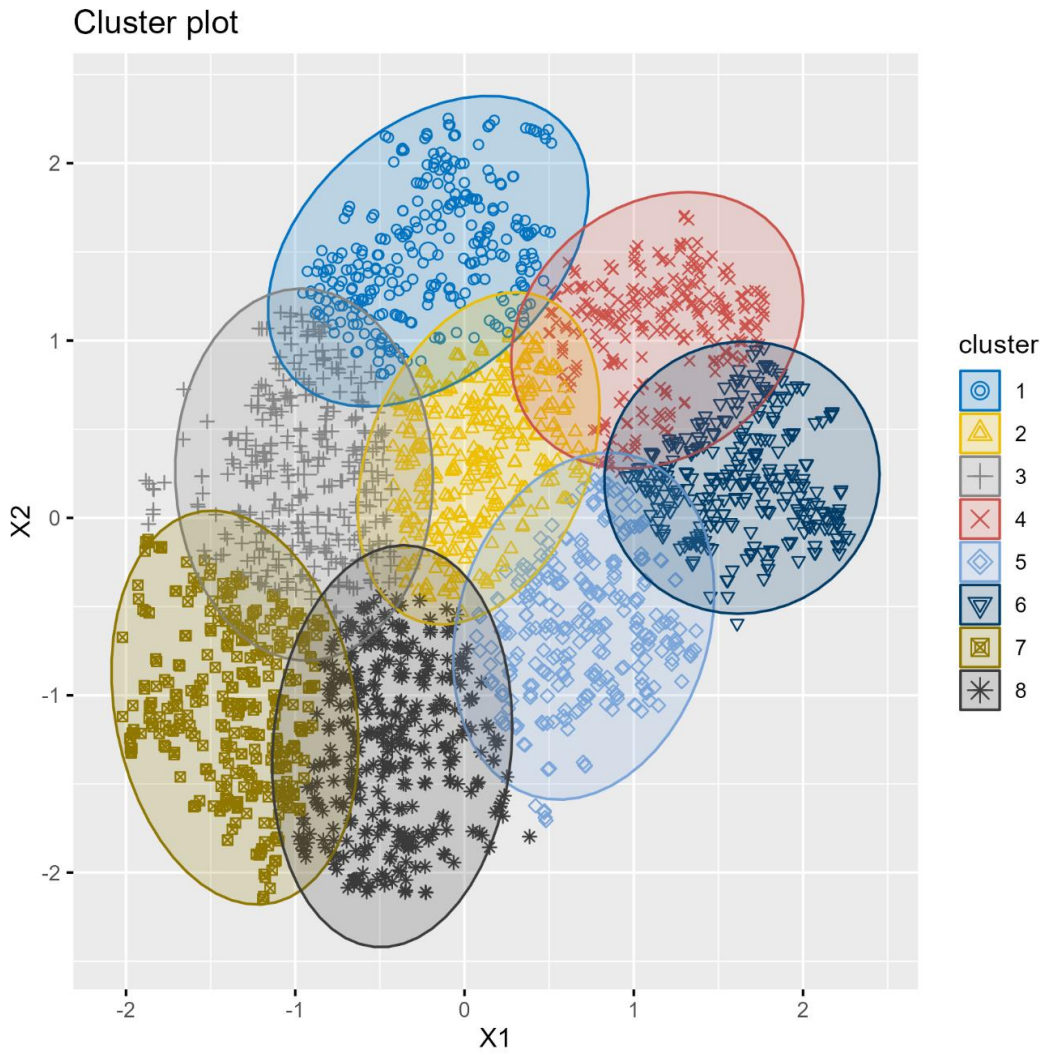


Fig-30: Cluster Plot with 8 clusters

We decided not to move forward with 8 clusters. This is because we had 13 dimensions in the original data which we can't use to visualize or find patterns efficiently. So we decided to move forward with k=3, which was the value that elbow point provided.

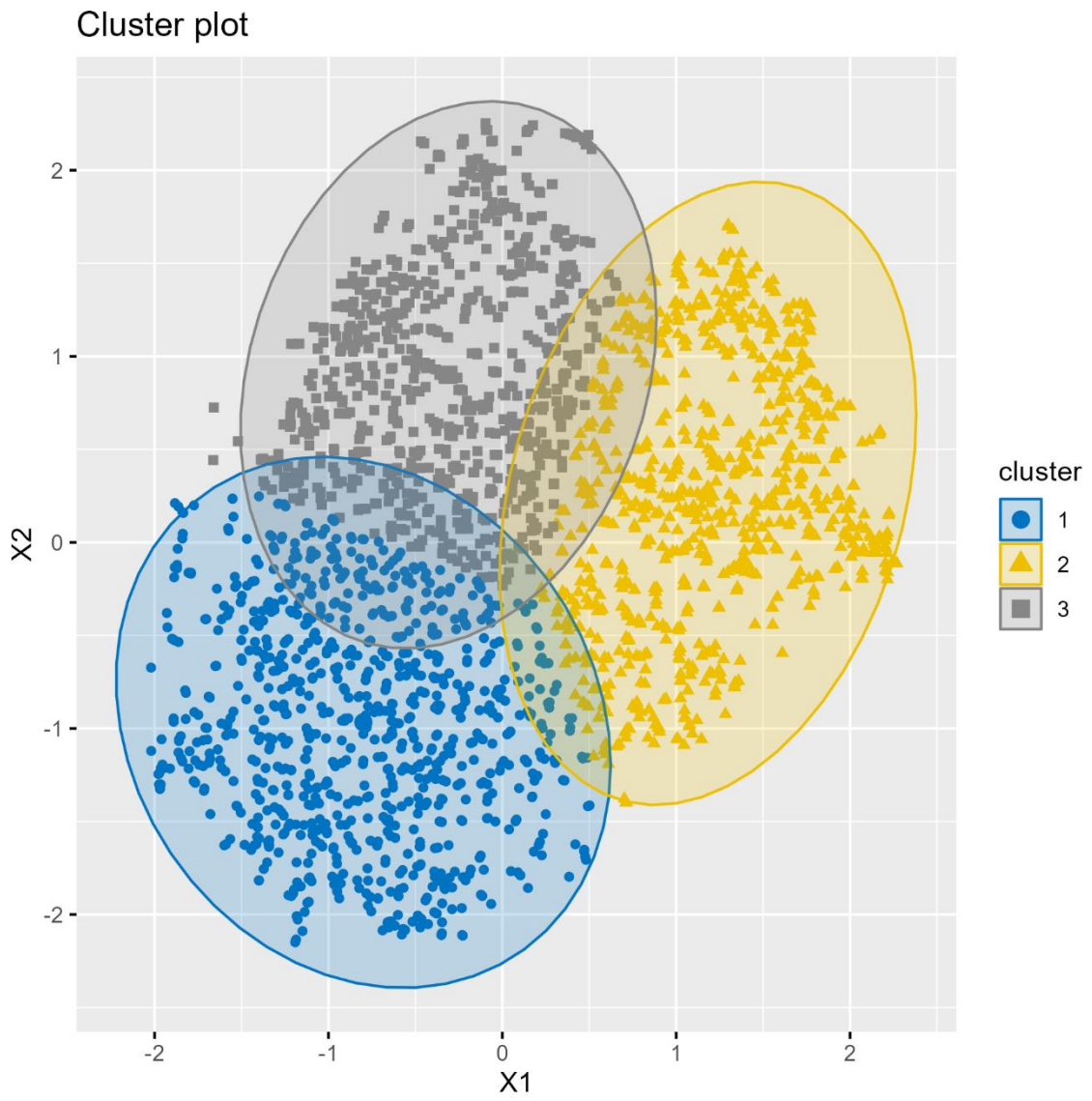


Fig-31: Cluster Plot with 3 clusters

Above figure shows K-means clustering for the Employment industry features, here k=3(from elbow point). As we can see the clusters are slightly overlapping near the center, but there is also significant generalization in terms of points being far apart from other clusters.

Below bar graph shows a comparison of the average values of the features for each of the three clusters formed using K-means. Each bar represents the number of people employed in a industry, per 1000 people in the county.

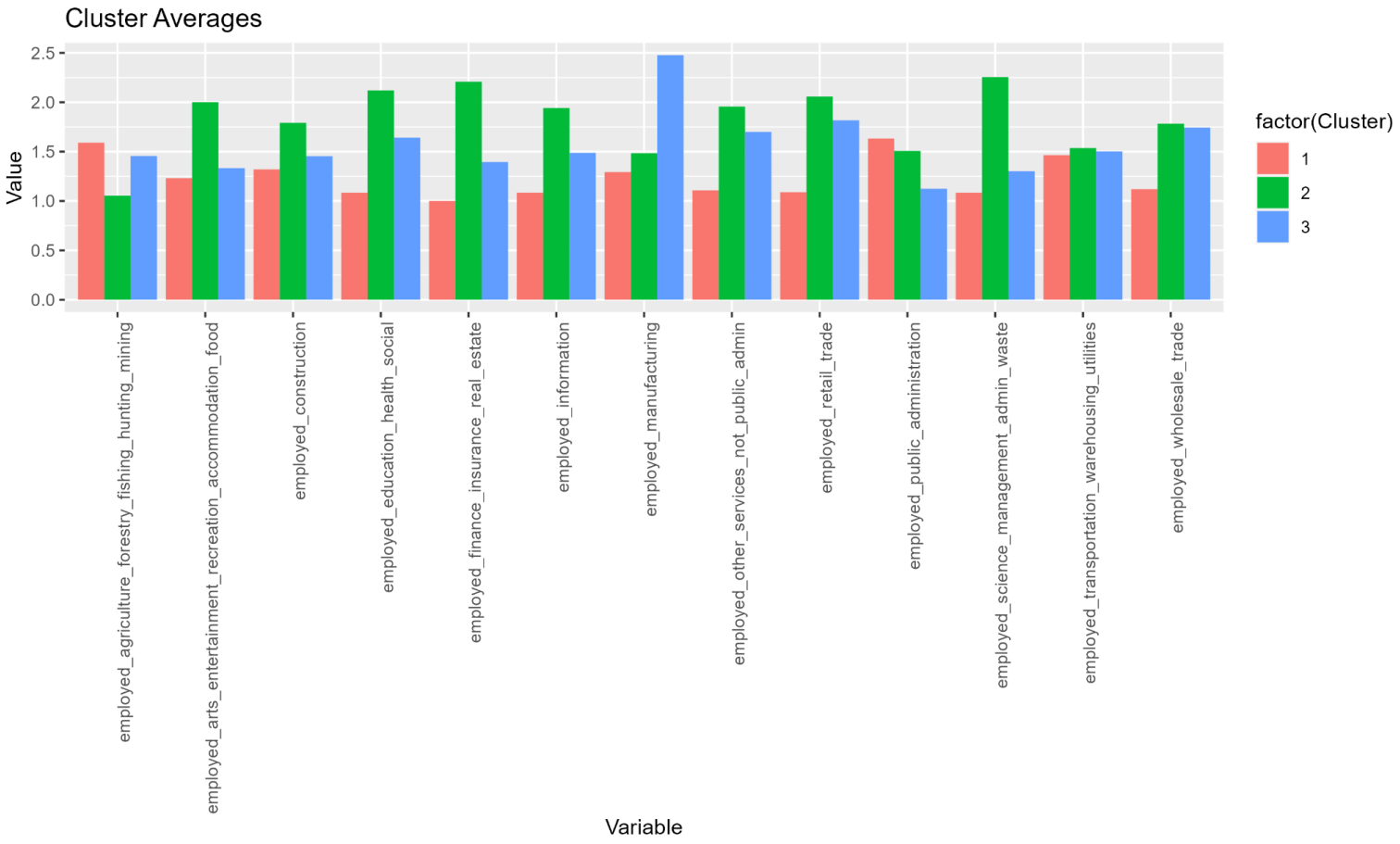


Fig-32: Bar graph comparison of different clusters with features average value.

In the below figure, we see how the clusters did in terms of actually fighting covid19. We see that counties in Cluster 3 had the most cases per thousand people. When we look at the

counties in Cluster 3, we notice that a high number of people work in construction, wholesale trade and retail trade.

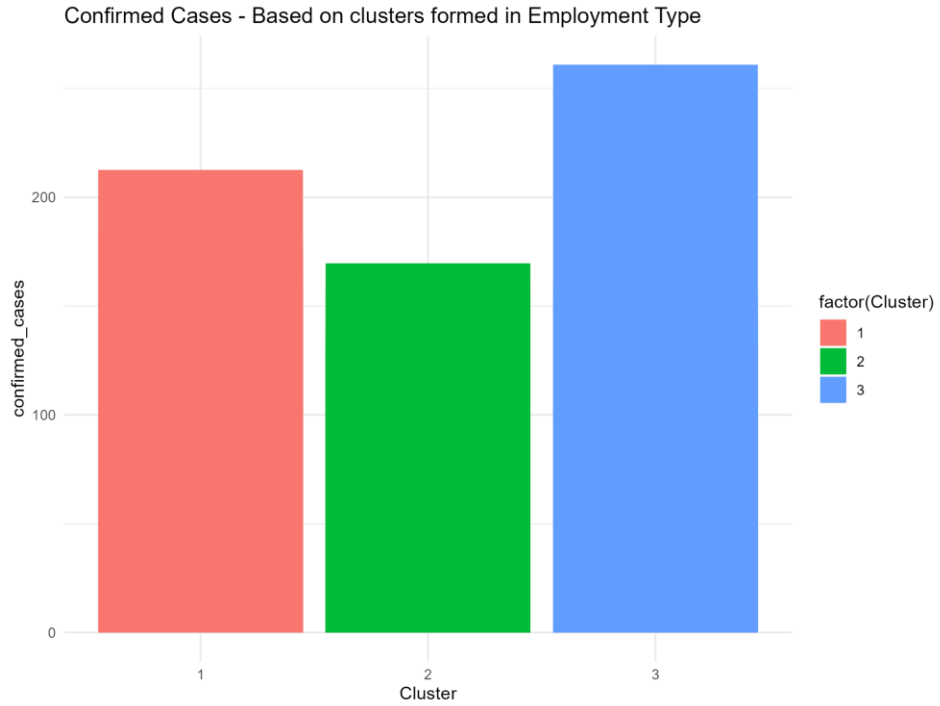


Fig-33: Confirmed cases vs clusters(Employment type)

Now when we look at the number of deaths per 1000 people for the clusters. We see that Cluster 1 has the most deaths per thousand people. If we look at Figure 32, we see that counties in Cluster 1, have the most number of people employed in agriculture, fishing, mining, forestry, hunting, and public administration.

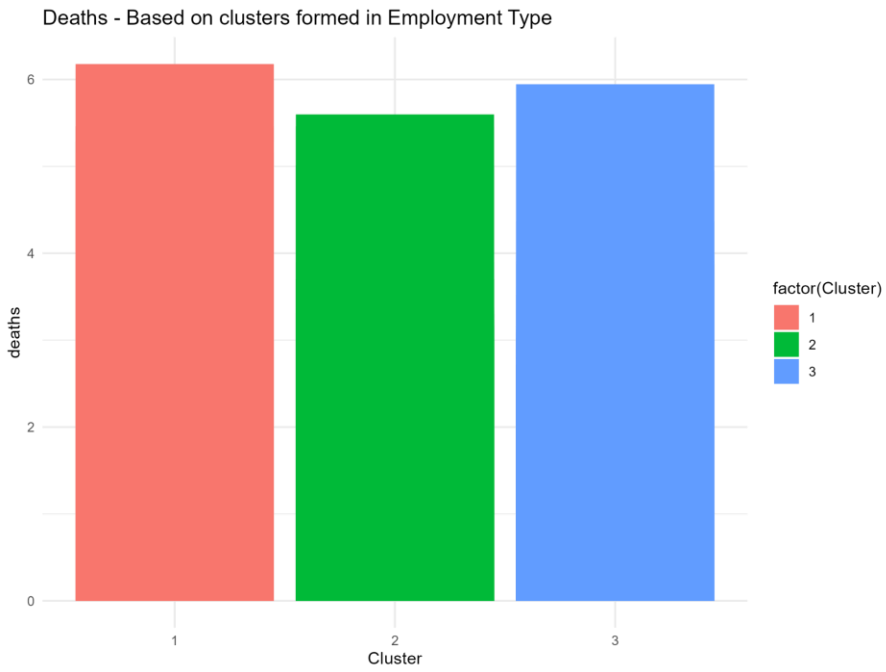


Fig-34: deaths vs clusters(Employment type)

Hierarchical Clustering – Employment Data

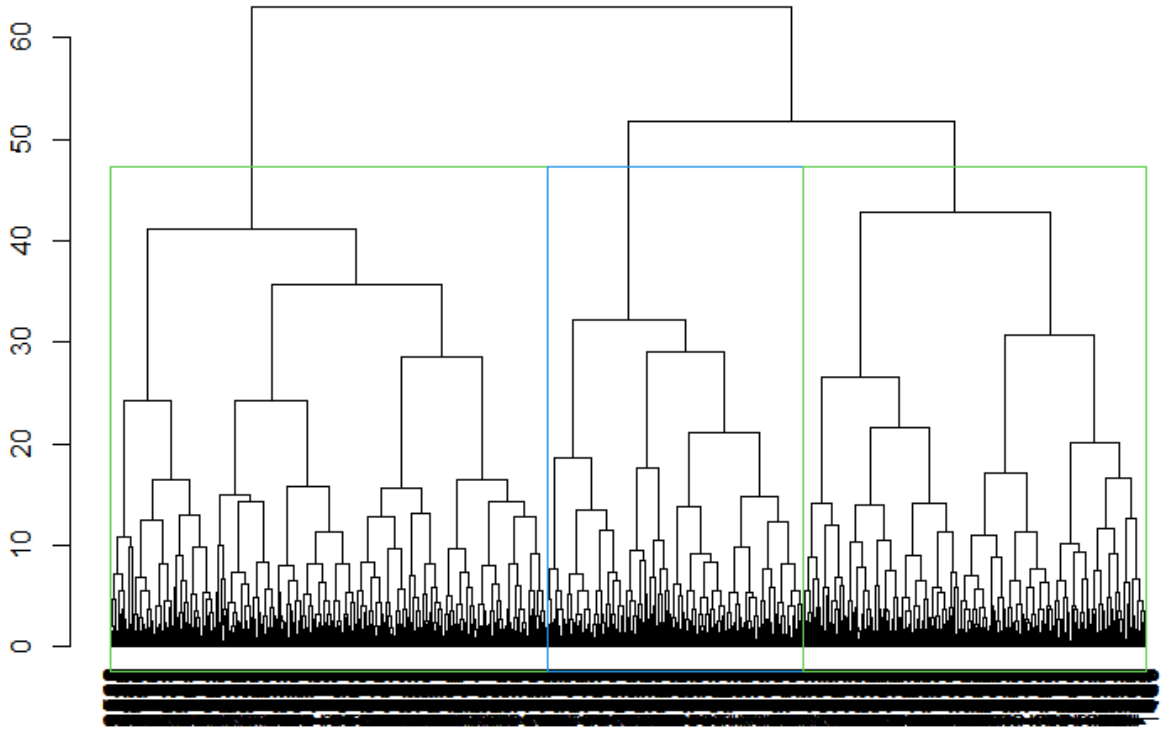


Fig-35: Hierarchical Clustering in Employment Data

Again we used Average Linkages and Ward Method to create two sets of clusters. In the below figures we see that there is some overlap of points in the Average linkages clustering while Ward method seems to do a good job of separating the points in the clusters.

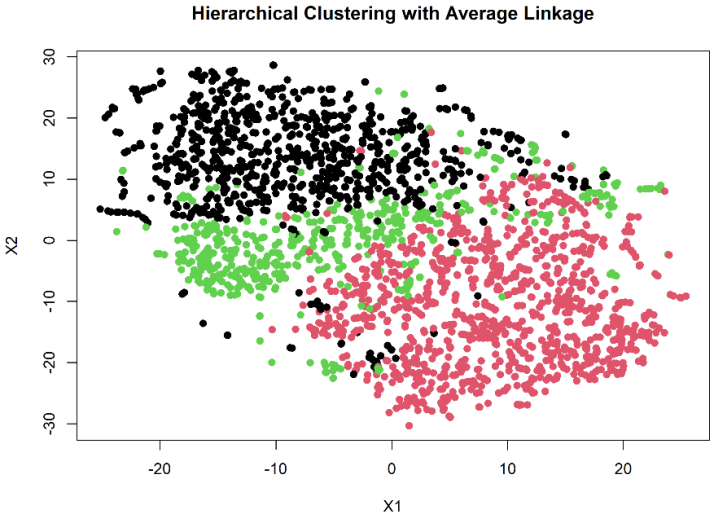


Fig-36: Hierarchical Clustering with Average linkage

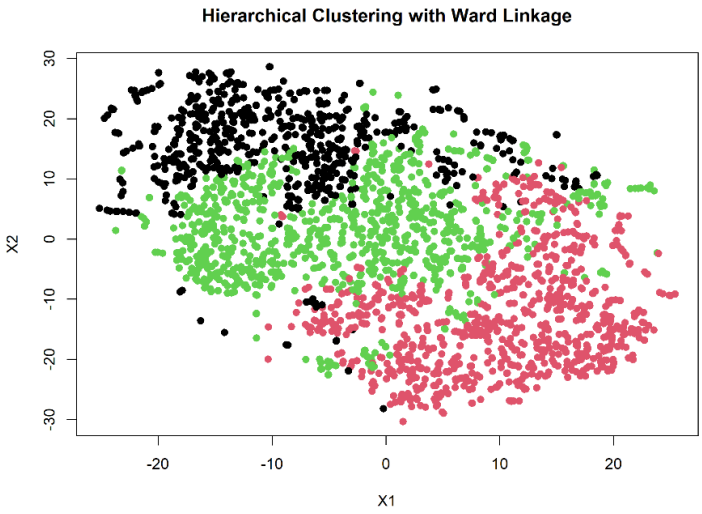


Fig-37: Hierarchical Clustering with Ward linkage

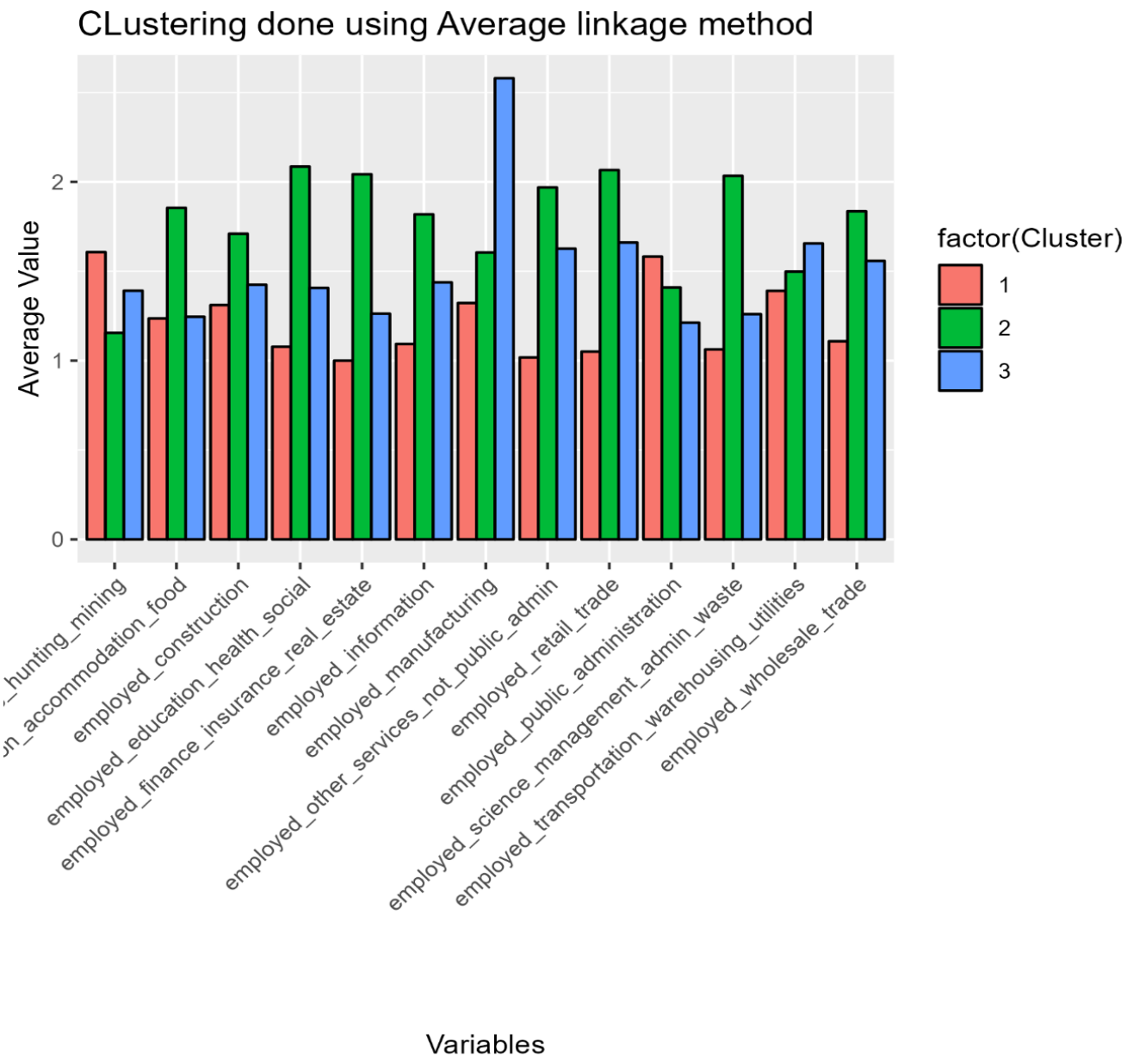


Fig-38: Clustering using average linkage method(Avg value vs employment variables)

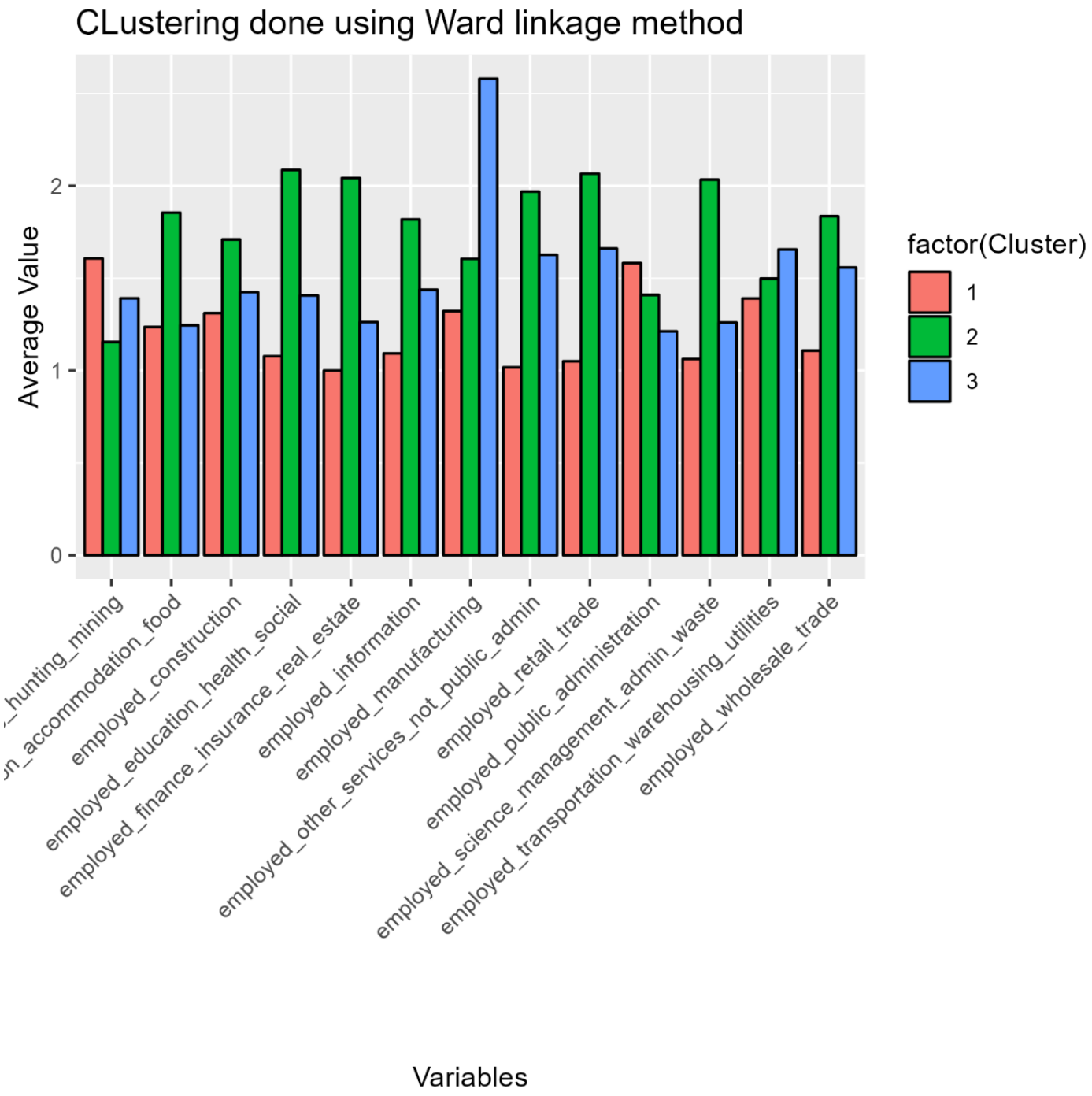


Fig-39: Clustering using Ward linkage method on(Avg value vs employment variables)

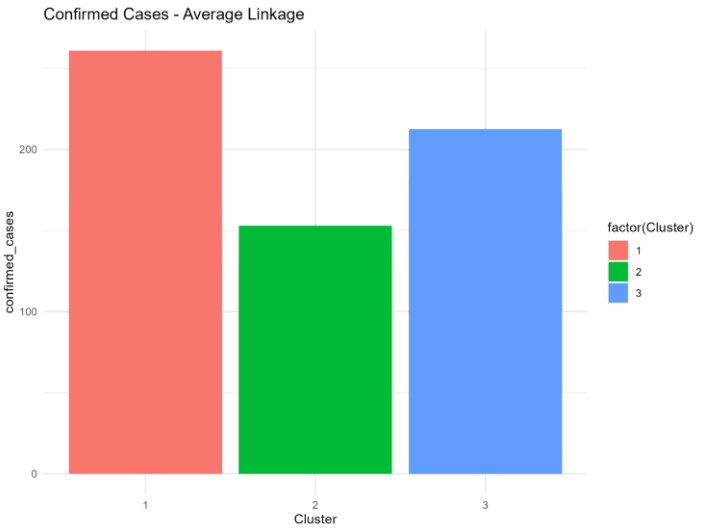


Fig-40: Average linkage(confirmed cases per 1k vs cluster)

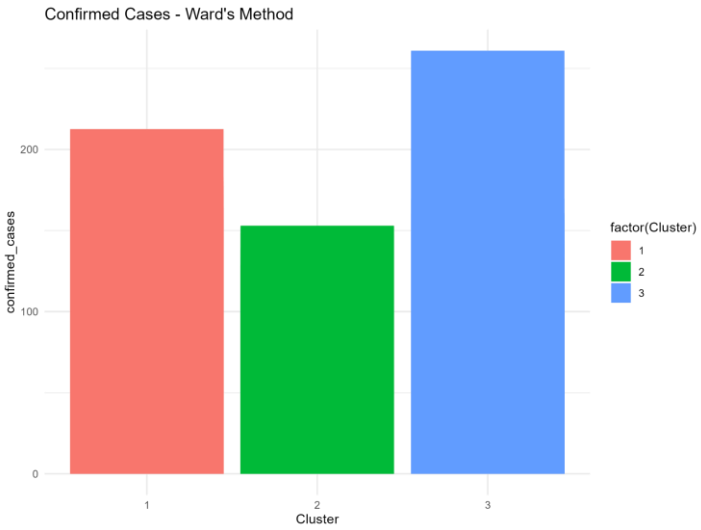


Fig-41: Ward's method(confirmed cases 1k vs cluster)

The above graphs show the confirmed cases per 1000 people, we see how there is a difference in clusters using Average and Ward linkage. Average Linkage clustering says that Cluster 1 had the most cases, while clusters created using Ward linkage indicate that the most number of cases were found in counties of Cluster 3.

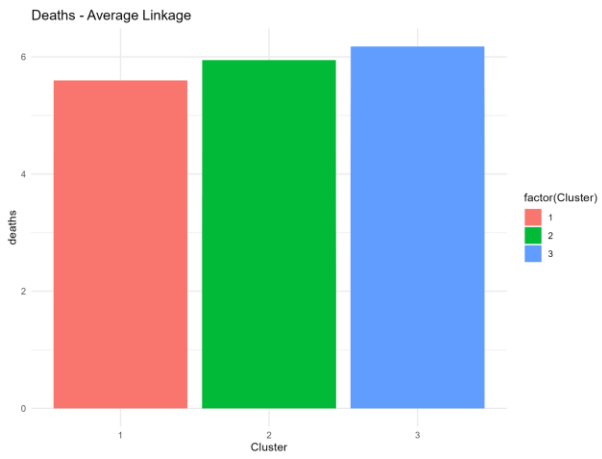


Fig-42: Average linkage(deaths per 1000 vs cluster)

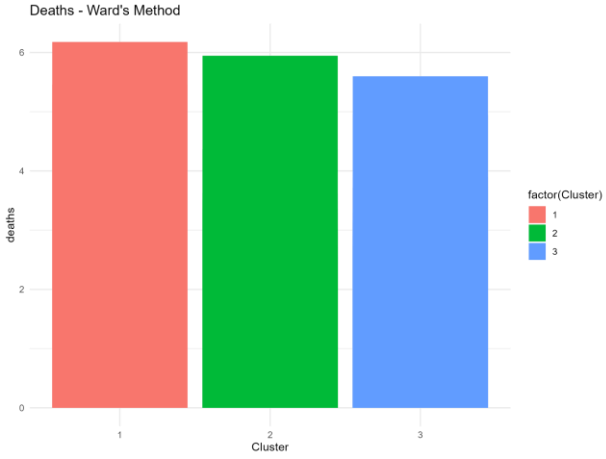
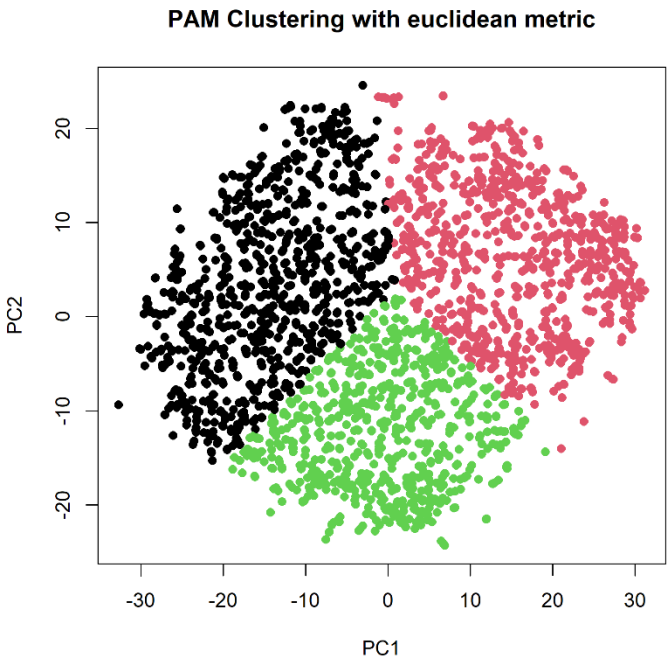


Fig-43: Ward's method(deaths per 1000 vs cluster)

The above graphs show deaths per thousand people in clusters formed using Average and Ward linkage respectively. In Clustering done using Average linkage, Cluster 3 had the highest death rate, while using Ward's method we see that Cluster 1 did the worst.

PAM(Partition around Medoids) - Exceptional Work



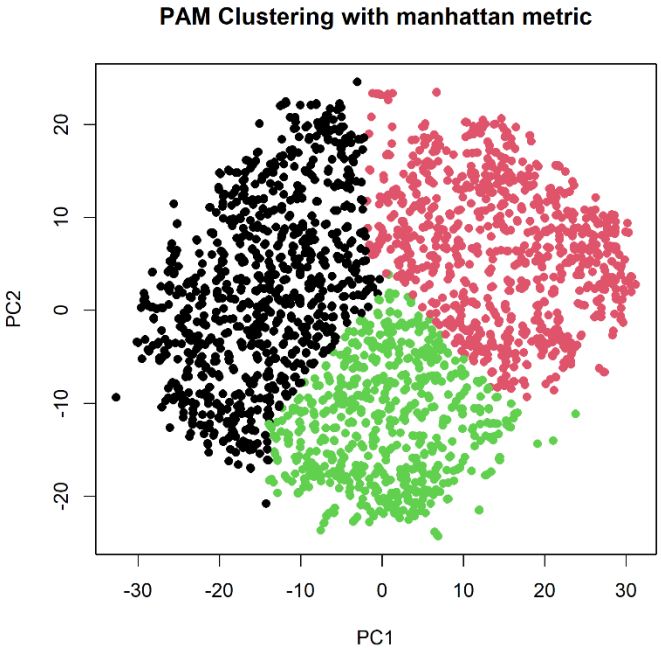
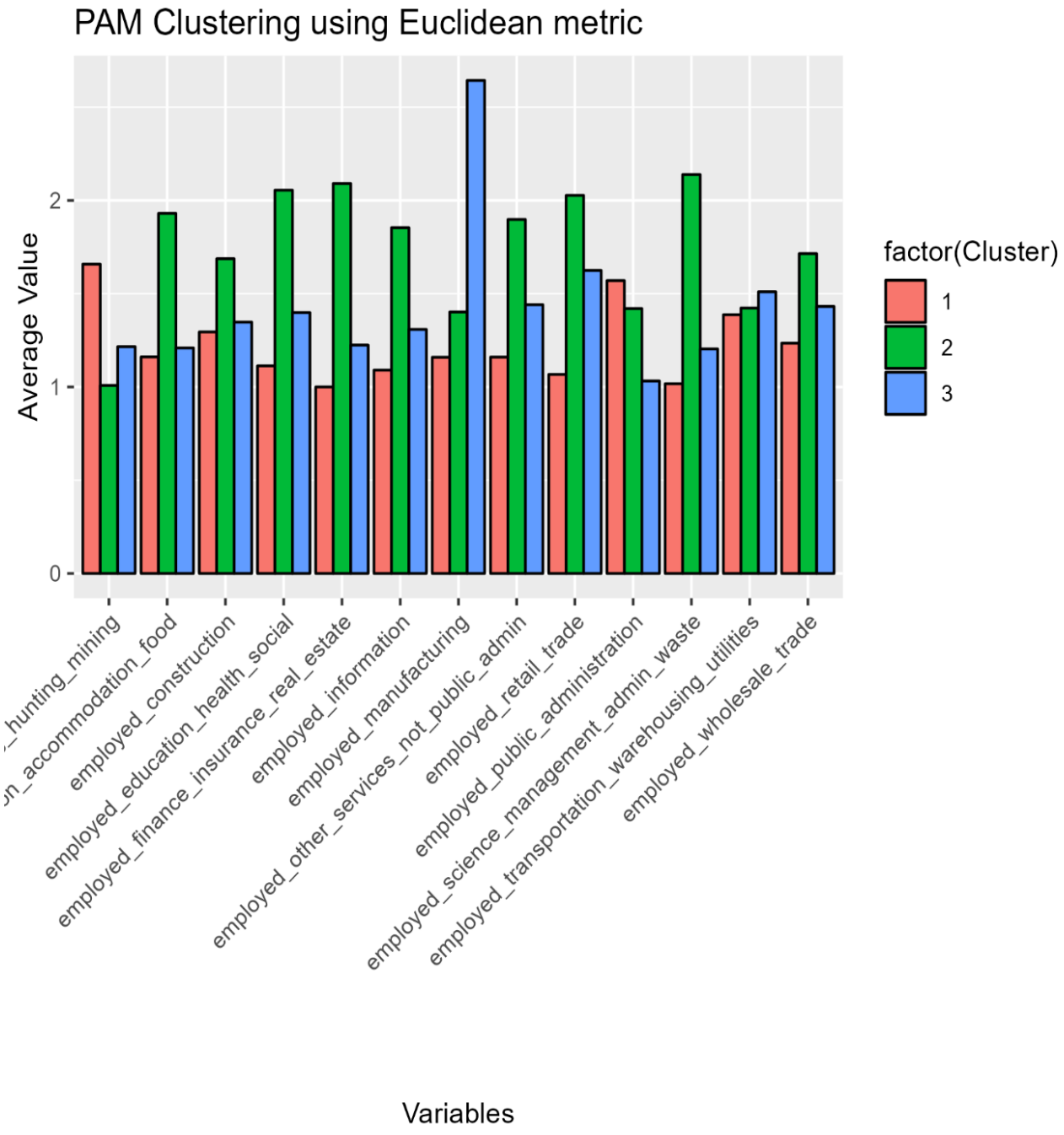


Fig-44: PAM Clustering with Euclidean metric(employment)

Fig-45: PAM Clustering with manhattan metric(employment)

We used Euclidean and Manhattan metric to create two sets of clusters for the Employment Industry type features. There is again, only a slight difference between the two around the edge of the clusters. Apart from that there is little noticeable difference between the two.

Fig-46: PAM Clustering using Euclidean metric, Avg value vs employment variables(below)



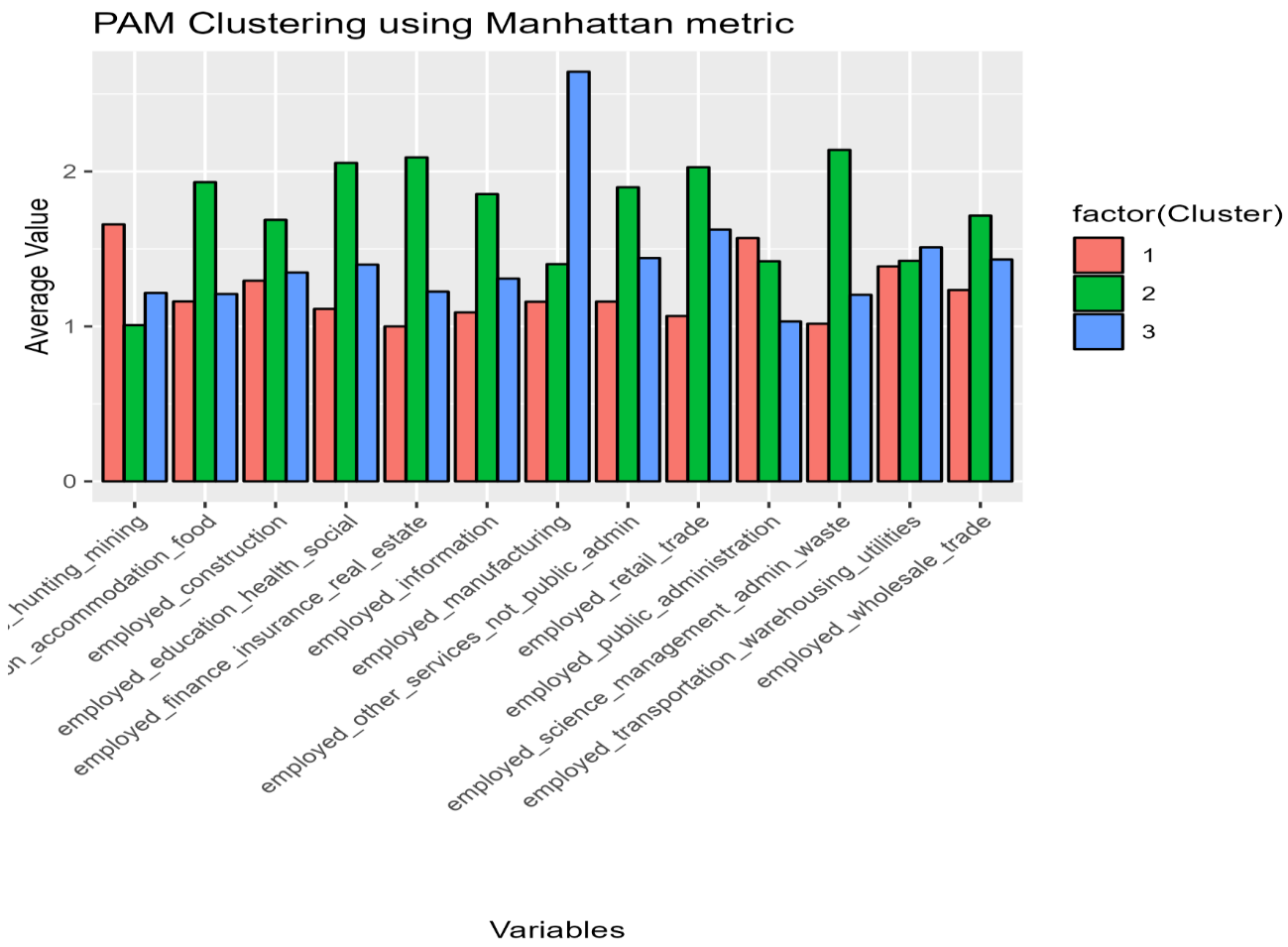


Fig-47: PAM Clustering using Manhattan metric (Avg value vs employment variables)

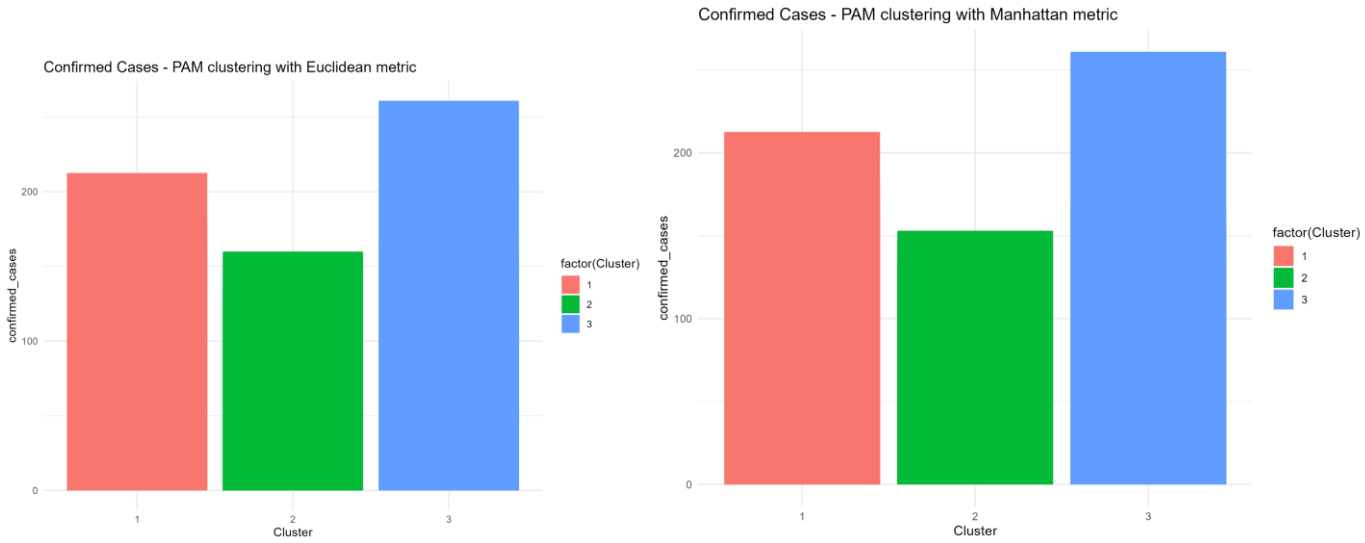
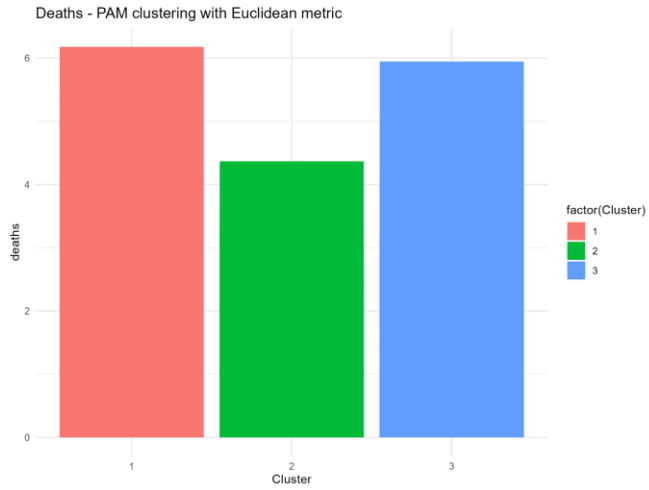


Fig-48: PAM Clustering Euclidean metric (confirmed case)

Fig-49: PAM Clustering Manhattan metric (confirmed case)

The number of confirmed cases per 1000 people is similar in both the cases, as shown in figures above. So is the number of deaths per 1000 people, shown in figures below.



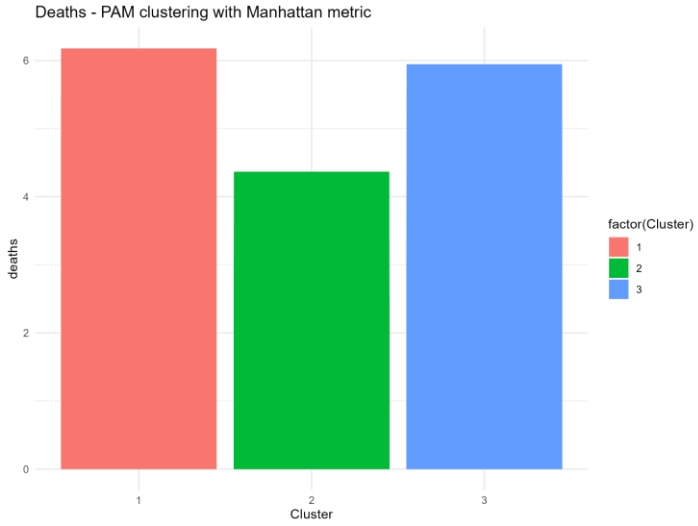


Fig-50: PAM Clustering Euclidean metric (deaths)

Fig-51: PAM Clustering manhattan metric (deaths)

We see that cluster 3 had the most cases, while cluster 1 had the highest death rate.

Validation using Silhouette Analysis -

While we created a lot of clusters using various algorithms and different parameter values, we had to validate our results using some metric. While we did a lot of analysis to validate clusters while we were working on them to make sure whatever inferences we are making are accurate. But we also looked at the statistics associated with clusters.

For the Economic Features, we were able to successfully validate our model using domain knowledge. We verified that the clusters we created were not arbitrary, and there was good generalization in terms of categorizing counties into 'rich' and 'poor'.

Apart from this we also did silhouette analysis to verify if the clusters formed are well separated or not.

The below plot shows silhouette width for the K means clustering of Employment features that we did.

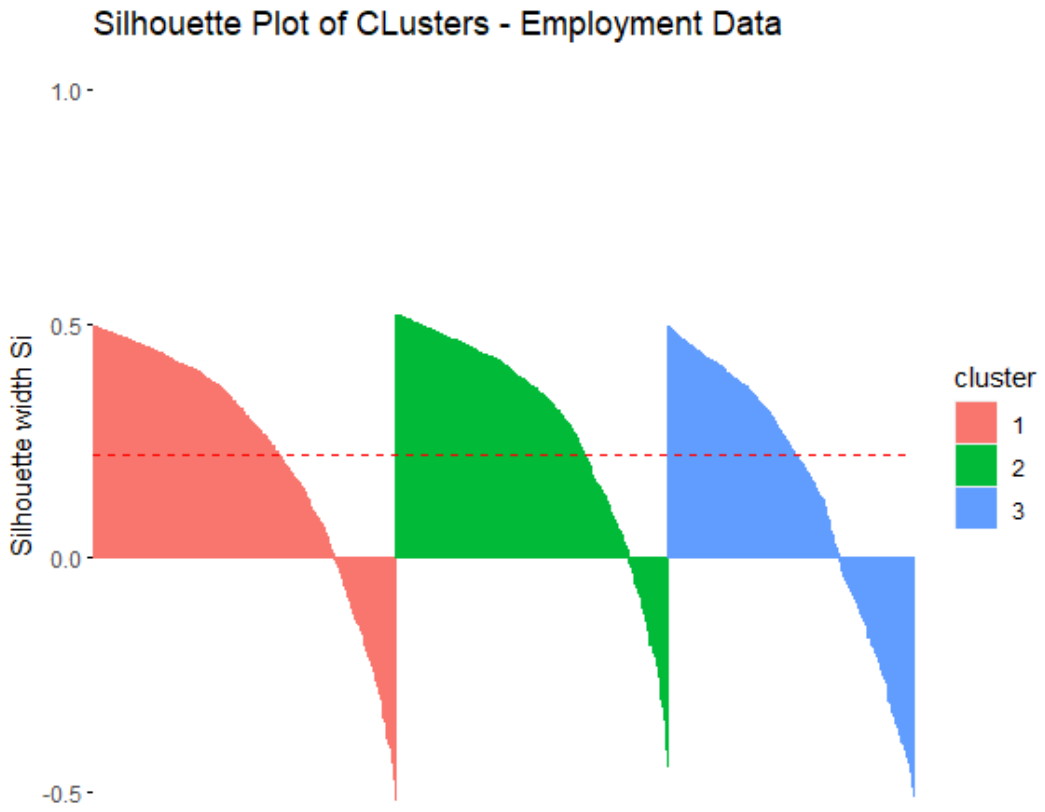


Fig-52: Silhouette Plot for K-means clustering on Employment Features

We see from the below graphs Hierarchical clustering did better to separate samples into clusters for the economic features compared to the Employment features.

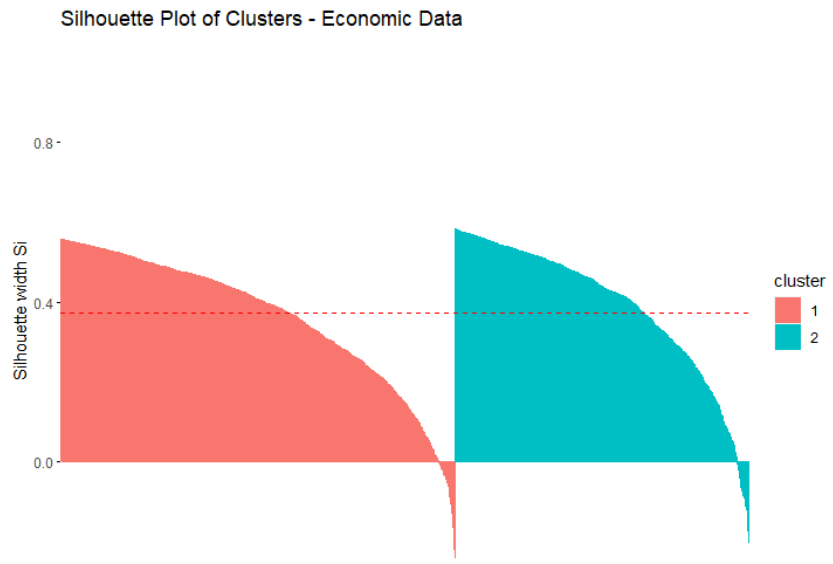


Fig-53: Hierarchical Clustering of Economic features

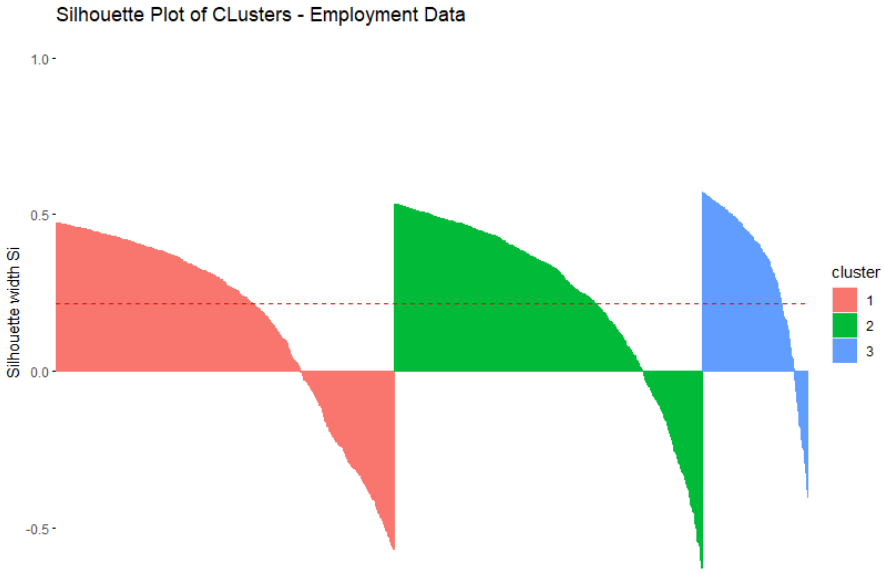
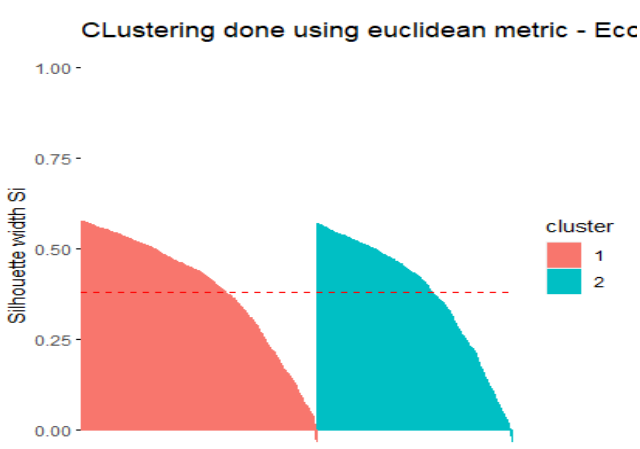


Fig-54: Hierarchical Clustering of Employment Features

Below Figures show clustering done using Partition around Medoids on Economic features for both Euclidean and Manhattan metrics. The Clusters formed by both metrics are very similar according to the below figures, indicating the silhouette width of each cluster.



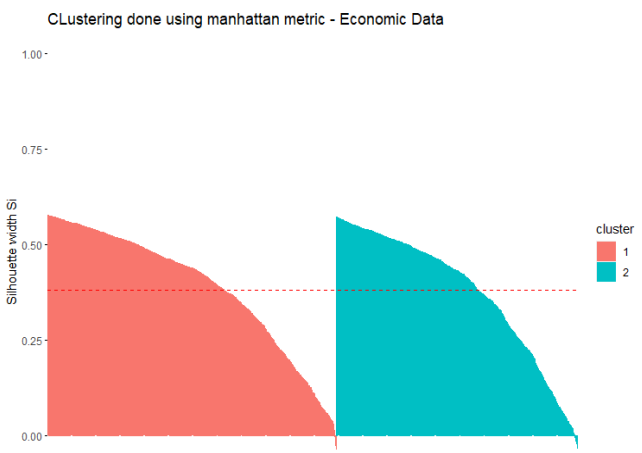


Fig-55: Clustering using Euclidean metric (economic data)

Fig-56: Clustering using Manhattan metric(economic data)

Below figures show PAM clustering done on Employment Industry features, using Euclidean and Manhattan metric.

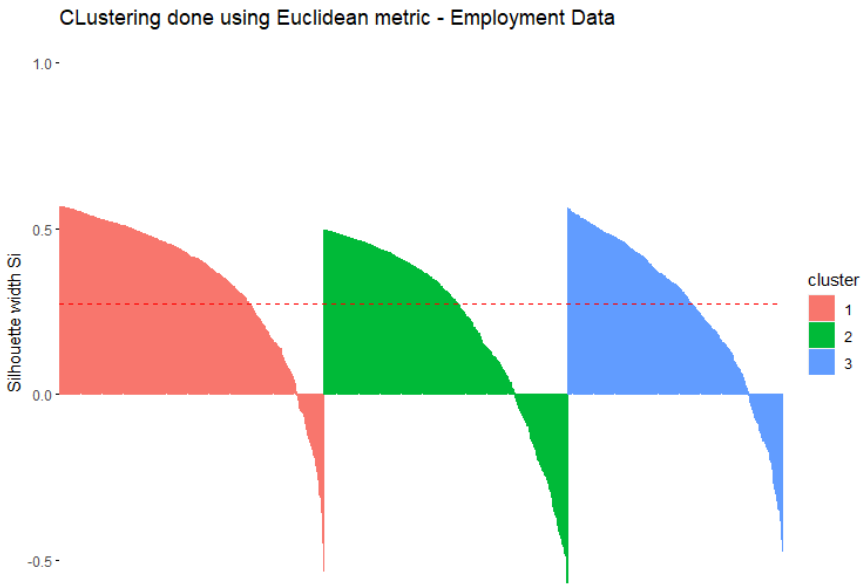


Fig-57: PAM Cluster ing using Euclidean metric (employment data)

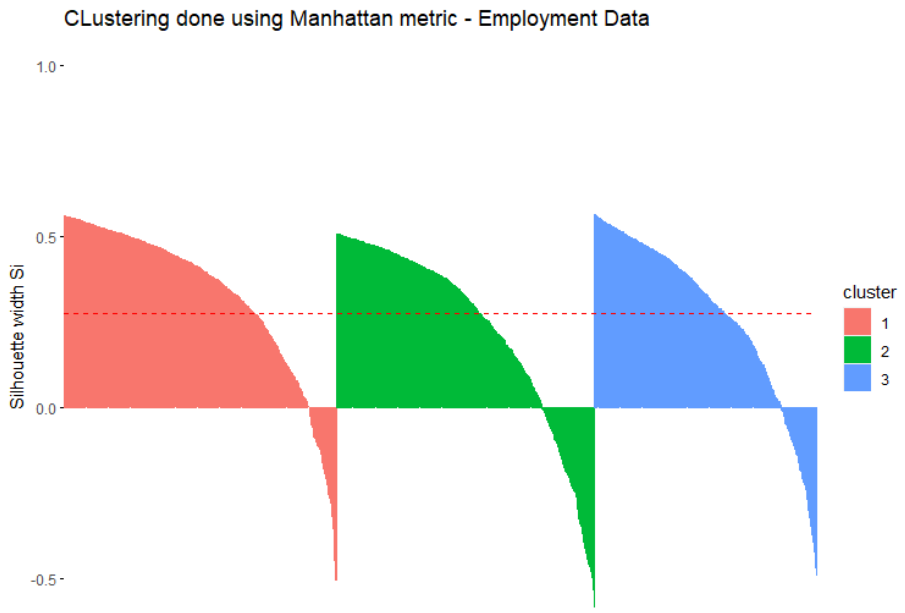


Fig-58: PAM clustering done using Manhattan metric(employment data)

There is not a noticeable difference between the two distance metrics that we used. But the plot clearly shows that Cluster 2 had points which overlapped with cluster 1 and 3, for both the metrics.

Conclusion -

We successfully used different clustering algorithms to implement unsupervised learning in the Covid-19 and census datasets. We successfully identified the features related to the domain we were interested in, which is the Economy formerly known as Economic Sciences.

We performed cluster analysis using K means, Hierarchical, and PAM clustering. All three algorithms worked perfectly for the Economic Feature types, we were able to validate using bar graphs displaying the average values of the features in the subset of the data we used. The counties characterized as 'poor' had significantly low per capita income and median income and there were significantly more households dependent on assistance and the number of people living in poverty in these counties was also significantly higher.

We also found out that the number of cases was higher in 'poor' counties while the death rate was higher in 'rich' counties.

For the other subset of features, the Employment Industry type features, that we selected, the motivation behind the selection of this feature was to identify at-risk job industries, and which job industry was more likely to suffer from COVID-19. We successfully identified counties that had more people working in at-risk industries, although the model created for this partition task was not as good as the one created for the Economic Feature subset.

References -

https://www.r-bloggers.com/2020/05/how-to-determine-the-number-of-clusters-for-k-means-in-r/

https://databank.worldbank.org/metadataglossary

https://www.r-bloggers.com/2021/12/how-to-use-the-scale-function-in-r/

https://r-charts.com/part-whole/hclust/

https://zimanaanalytics.medium.com/how-to-create-a-dendrogram-in-r-programming-c0eac78ac77

https://www.datacamp.com/tutorial/hierarchical-clustering-R

https://stackoverflow.com/questions/2310913/how-do-i-manually-create-a-dendrogram-or-hclust-object-in-r

https://stackoverflow.com/questions/38891392/how-to-change-the-color-of-dendrogram-for-each-group-in-a-cluster

https://r-graph-gallery.com/29-basic-dendrogram.html

https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/clustering-algorithms-evaluation-r/tutorial/

https://www.r-bloggers.com/2021/04/cluster-analysis-in-r/

https://www.statmethods.net/advstats/cluster.html

https://www.datanovia.com/en/courses/advanced-clustering/