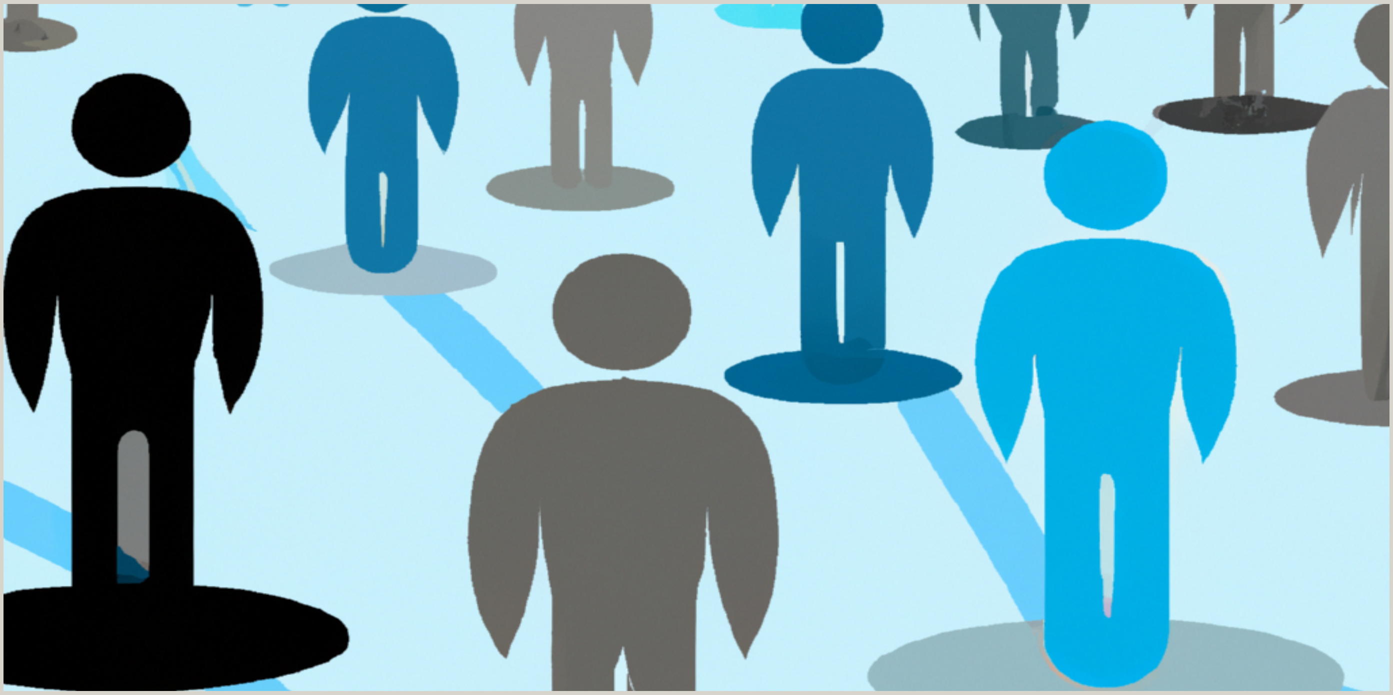


Part 2: COVID-19 Clustering Analysis for Policy Recommendations

By Lawrence Lim

Table of Contents

Introduction

❖ Abstract

With COVID vaccines on the verge of being released from Pfizer, Moderna, and J&J, there are questions regarding where exactly vaccines should be distributed and the particular demographics that should receive the vaccines. Recently, polling has suggested that over half of people in the US are interested in receiving the vaccine, which would amount to over 180 million people. With only 3 million vaccines available in the first batch and up to 100 million shots produced by April, the total number of Americans interested in receiving the vaccine cannot do so until the end of 2021 or later. With vastly greater demand than supply, the state of Texas has tasked me to determine which subset of individuals should receive the vaccine first by grouping based on demographic attributes. For this study, I decided to use K-means clustering to partition the cities/states into a pre-specified number of clusters based on deaths and other socioeconomic factors. For example, the clusters could be based on factors such as population density, age distribution, and income level. This would allow for

the identification of areas that are most affected by the pandemic and help in developing targeted interventions and policies. Furthermore, Hierarchical Clustering allowed me to create a hierarchy of clusters based on the similarity between counties. The hierarchy could then be used to identify clusters of counties that are most similar in terms of their infection numbers and demographics. This could help in identifying regional patterns in the spread of the virus and understanding the underlying factors that contribute to these patterns.

❖ Business Understanding

The group interested in the study will be the Texas Department of Health which is responsible for the transportation and rollout of the vaccine to different communities. Furthermore, Health departments in other states and the national CDC might also be interested in the conclusions from this research in influencing policies on a state and national level regarding vaccine distribution planning for 2021. It is important to note that the CDC is responsible for distributing the limited national supply of vaccines between the various states. Thus, this research is essential in allowing the CDC to fully understand Texas' vaccine needs as one of the largest and most populous states in the nation.

The Texas Department of Health will want to understand which individuals and which districts should be allowed to get the vaccine first. One particular factor to consider that the CDC has already explored is age. In a study where 18-29-year-olds are the reference group, individuals that are 65 and older are only about 60-80% as likely to contract COVID. However, individuals 65-74 years old are 60 times more likely to die and five times more likely to be hospitalized if infected. And those over 85 years old, individuals 350 times more likely to die from COVID and 15 times more likely to require hospitalization because of COVID complications. This existing research suggests that if we were to cluster different counties in Texas based on the median population age, as compiled online and appendable to our existing Texas dataset, we would find that counties with an older population will have a greater ratio of total deaths compared to one with a younger population.

Data Understanding and Preparation

❖ Introduction to the Given Data

As it pertains to the national data, the dataset that I will utilize stems from my project 1 dataset, as this project is building off project 1. Furthermore, I already selected the attributes from project 1 that are not duplicates or merely smaller sub-categories of another overarching attribute. For example, the % with a graduate degree is an often used metric to define the educational background of a particular district in comparison with others. Thus, the other attributes, like the % that graduate high school and the number of people in particular grade levels, are unnecessary attributes that should be eliminated. Furthermore, the median income and Gini index both capture the true middle income of the population, along with the relative distribution of wealth and wealth inequality. Median income and Gini index are also statistics that can be used to compare completely different cities, unlike the # individuals in poverty, which was removed as it depends on population size and thus was not a valid metric. It is important to note that we will remove additional features in the feature selection phase. The following national data attributes that I will start with are listed below:

Column	**Data Type**	**Description**	**Example**	**Potential Use Cases of Features in Clustering**
County Name	nominal	This is the name of the particular county.	Tarrant County	This is a nominal variable and cannot be directly used in clustering algorithms. However, it can be used as a label to identify clusters after the clustering has been performed.
State	nominal	This is the state abbreviation the county resides in.	TX	This is a nominal variable and cannot be directly used in clustering algorithms. However, it can be used as a label to identify clusters after the clustering has been performed.
Population_1_year_and_over	ratio	This is the population of a county that is over one year old	1000000	This is a ratio variable that could be used to group counties based on their population size.
Confirmed	ratio	This is the number of confirmed cases (as reported on January 19th, 2021).	1201	This is a ratio variable that could be used to group counties based on their COVID-19 prevalence.
Deaths	ratio	This is the number of deaths (as reported on January 19th, 2021).	321	This is a ratio variable that could be used to group counties based on their COVID-19 severity.

median_year_structure_built	ordinal	This is the median year that a home was built in for a particular county.	1972	This is an ordinal variable that could be used to group counties based on the age of their housing stock.
percent_income_spent_on_rent	ratio	This is the average % amount of income spent on rent for residents in the county.	52%	This is a ratio variable that could be used to group counties based on their housing affordability.
Percent_male	ratio	This is the percentage of males in the county	50%	This is a ratio variable that could be used to group counties based on their gender demographics.
median_income	ratio	This is the median income of an individual in the county.	$60,000	This is a ratio variable that could be used to group counties based on their income level.
% graduate degree or more *	ratio	This is the % of individuals that have a graduate degree or higher.	41%	This is a ratio variable that could be used to group counties based on their educational attainment.
aggregate_time_travel _work	ratio	This is the total amount of time (in minutes) spent by individuals traveling to work on January 19th.	1668430	This is a ratio variable that could be used to group counties based on their commuting patterns.
Genie index	Ratio	This is an index that measures inequality in a particular county.	0.417	This is a ratio variable that could be used to group counties based on their level of income inequality.

For the state data, we will utilize the attributes listed below:

Column	**Data Type**	**Description**	**Example**	**Potential Use**

				Cases of Features in Clustering
County Name	nominal	This is the name of the specified county	Dallas	Specify the county that we are looking at.
Total Deaths	ratio	This is the number of deaths (from 1/22/2020 to 1/25/2021)	7000	could help identify if certain age groups are more vulnerable to the severe effects of the virus.
Total Confirmed Cases	ratio	This is the number of deaths (from 1/22/2020 to 1/25/2021)	130000	Could be a metric in measuring how different age ranges compare with regards to their susceptibility or likelihood to contract the virus. For example, data on the number of COVID-19 cases by age range could be analyzed to determine if certain age groups are more vulnerable to the virus than others.
Median Population Age	ratio	This is the median population age as reported by the Texas Association of Counties	43	Could use this as grouping mechanism to determine identify associations between different age ranges as it pertains to the # of COVID cases or deaths

*****Note that the FIPS Code attributes for State and County have been eliminated as we are not aiming to display a geographic map of our data.**

- Feature Selection
    - Correlation Coefficient

The first means of additional feature reduction can be achieved through the use of correlation analysis. In this case, we want to see which particular features have a high and similar correlation/impact on the dependent variable. This is important as if two features have a high correlation, we might consider dropping one of the features to prevent multicollinearity, depending on what that feature is. To determine whether a highly correlated feature is indeed redundant, I will calculate the value inflation factor between the features. It is important to note that I have chosen deaths as my target attribute in the national dataset, as it is the best factor to quantify vaccine lethality.

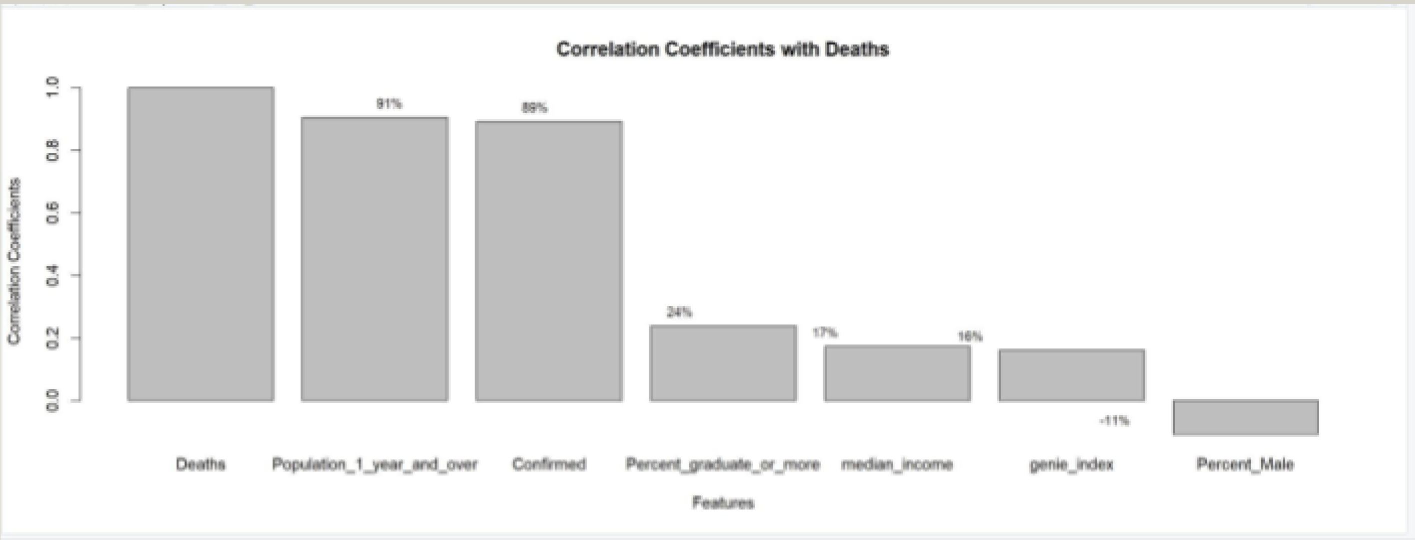


Figure 1: Correlation Coefficient between Applicable Variables

Figure 1 shows the relationship between the target variable (Death) and other applicable attributes that we are considering eliminating. The variables like State and County are not considered as they are necessary variables, considering the context of the problem is to determine the particular districts that need the vaccine. If we were to delete them, there would be no way to determine changes in factors that relate to geographic location. Excluding death, we see that the only attribute of concern is the Population_1_year_and_over and confirmed cases. This is not surprising considering that the number of COVID deaths will often be higher in areas with greater populations and areas that have more people infected. The confirmed cases can be considered a multicollinear variable in this case; however, I feel that population does have importance in the dataset when considering the question of the people and places that should get the vaccine first. Therefore, I will keep the population attribute for now, and see if other feature selection methods determine if it has value.

Figure 2 shows us the VIF results for each of the attributes. VIF is a means of determining the multicollinearity within a dataset between the different attributes. Typically, a VIF above 5 may be a cause for concern, but

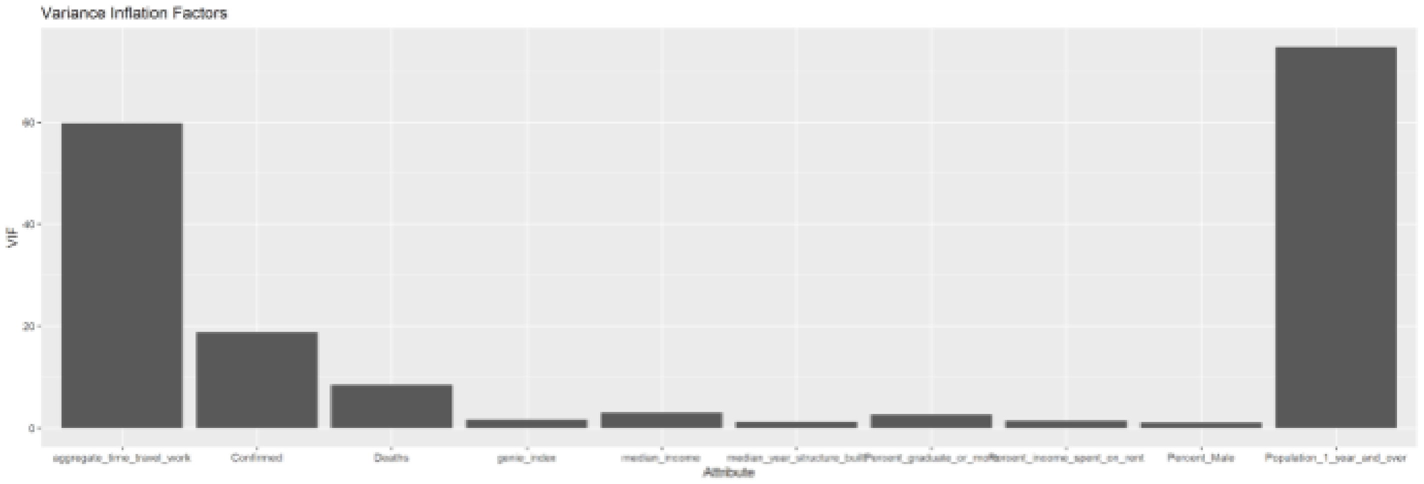


Figure 2: VIF Results

again, the context of the overall goal must be considered. The aggregate time to travel to work and population are shown overall as variables with extremely high VIF, suggesting that they are highly correlated with each other. These two variables thus may be multicollinear with one another. Therefore, we can remove aggregate travel time to work and keep just population size.

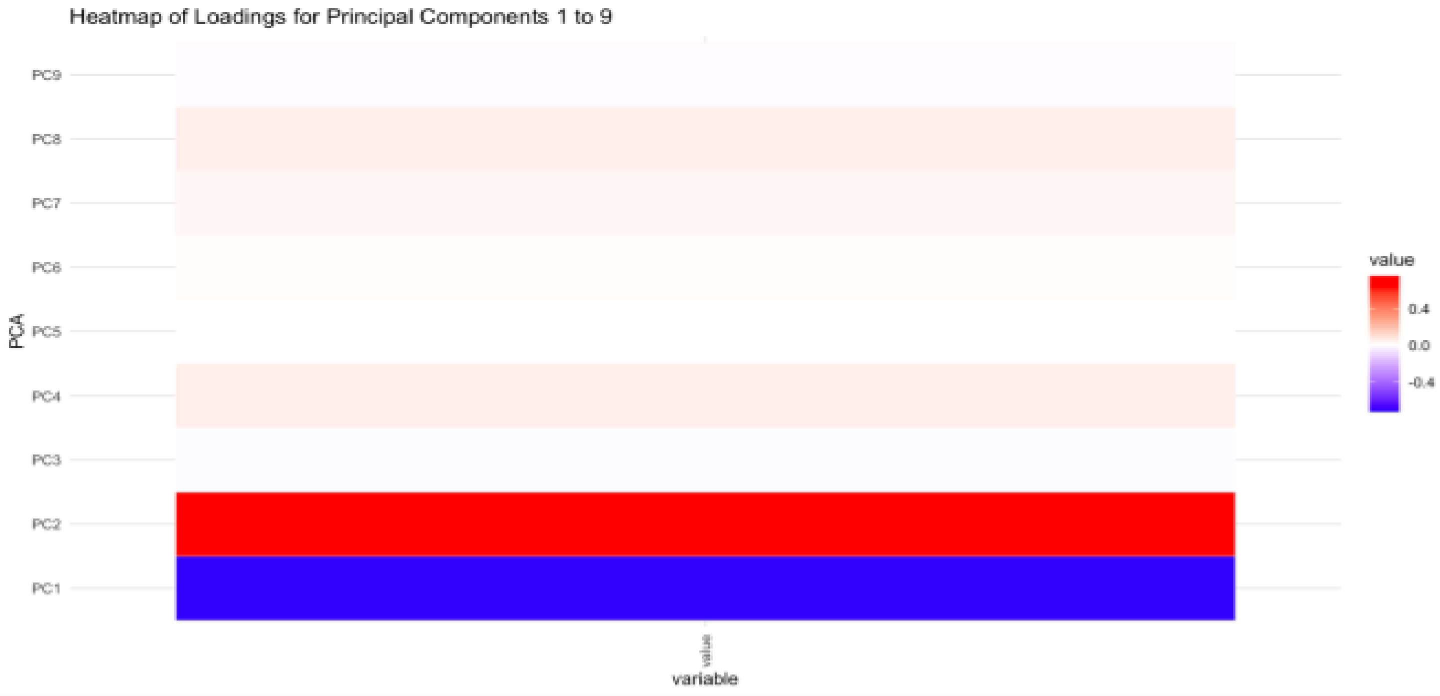
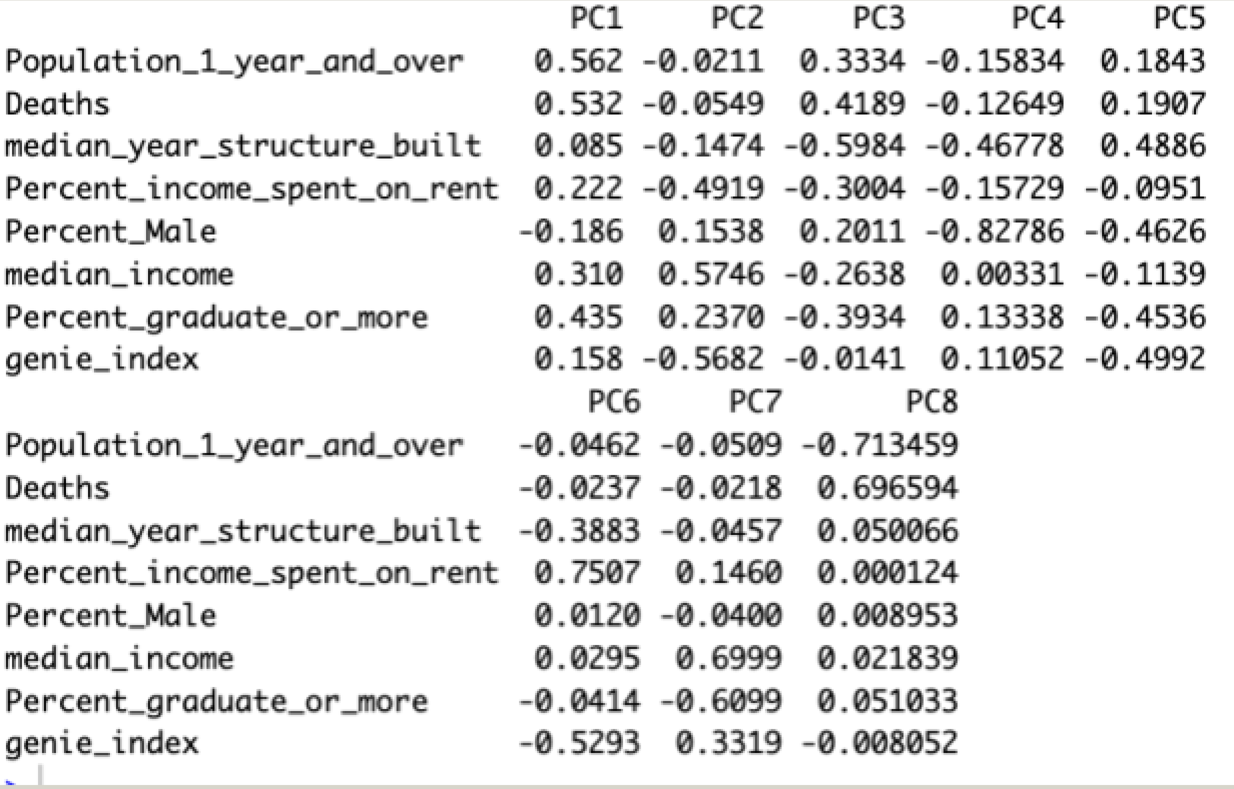


Figure 3: PCA Results: How is variation in the data explained by the variables?

- ➢ Principal Feature Analysis (i.e., Dimensionality Reduction)

PCA is an important factor as it allows us to see the relationship between the original variables by looking at the loadings. PCA allows us to answer the question regarding how linear combinations of the variables, which form each of the components, explain the overall variation in the data. Figure 3 below depicts the principal components for the attributes, excluding confirmed cases and aggregate time to travel to work. The two brightly colored bars for PCA 1 and 2 with other less colored components suggest that those two components explain the majority of the variation in the dataset. This can be considered a positive finding as if the brightly colored principal components have a strong relationship with some or all of the attributes, it suggests that specific attributes may perform favorably when clustering. Figure 4 shows the contribution of attributes to each component. This is significant as it allows us to perform dimensionality reduction on the dataset. Thus, we can determine the features that contribute the most to the set and select them for clustering as they capture the underlying patterns in the data. The table suggests that the features that we have selected are good choices for clustering. In PC1, we see that population, deaths, income spent on rent, median income, and the % with a graduate degree all contribute to the variance in the data. For PC2, the other large component of the data, we see that all the attributes excluding population and deaths contribute moderately or greatly to the data.



```
                                PC1     PC2     PC3      PC4     PC5
Population_1_year_and_over    0.562 -0.0211  0.3334 -0.15834  0.1843
Deaths                        0.532 -0.0549  0.4189 -0.12649  0.1907
median_year_structure_built   0.085 -0.1474 -0.5984 -0.46778  0.4886
Percent_income_spent_on_rent  0.222 -0.4919 -0.3004 -0.15729 -0.0951
Percent_Male                 -0.186  0.1538  0.2011 -0.82786 -0.4626
median_income                 0.310  0.5746 -0.2638  0.00331 -0.1139
Percent_graduate_or_more      0.435  0.2370 -0.3934  0.13338 -0.4536
genie_index                   0.158 -0.5682 -0.0141  0.11052 -0.4992
                                 PC6     PC7       PC8
Population_1_year_and_over   -0.0462 -0.0509 -0.713459
Deaths                       -0.0237 -0.0218  0.696594
median_year_structure_built  -0.3883 -0.0457  0.050066
Percent_income_spent_on_rent  0.7507  0.1460  0.000124
Percent_Male                  0.0120 -0.0400  0.008953
median_income                 0.0295  0.6999  0.021839
Percent_graduate_or_more     -0.0414 -0.6099  0.051033
genie_index                  -0.5293  0.3319 -0.008052
```

Figure 4: Loading Matrix

➢ Final Feature Selection Choices:

Given the reduction accomplished before, the final list of features after the process of feature elimination and modification will be listed below. Names have been simplified/renamed for simplicity:

National Dataset:

Column	Data Type	Description	Example
Population_1_year_and_over	ratio	Population of county	120000
Deaths	ratio	# of deaths in a county	56
median_year_structure_built	ordinal	The median year a structure was built in a county	1976
Percent_income_spent_on_rent	ratio	% of income spent on rent, on avg, in county	52%
Percent_Male	ratio	% of males in county	50%
median_income	ratio	Median income in county	60000

Percent_graduate_or_more	ratio	% of county population with masters degree or more	31%
genie_index	ratio	a statistical measure that quantifies income inequality within a population.	41.4
County_ID	nominal	Unique county identifier #	39

❖ Determination of the Scale of Measurement

The proper measurement option depends on the problem context and the dataset. The national dataset consists of demographic and socioeconomic factors that are compared by the magnitude of their values. Thus, there is a benefit to selecting cosine similarity, which is more generally favored when dealing with quantifying text-based info. However, though our dataset is dominated by numerical attributes, it is also important to consider the categorical fields of county_name and state. This along with the fact that our dataset has higher dimensionality with more than 3 attributes would suggest that Manhattan distance is a better choice. One way to see the difference between these two options is by determining the pairwise distance between all data points using both distance measures and comparing the resulting distance matrices. To achieve this, I must convert the categorical variables to numeric identifiers and then perform a pairwise calculation. Though the obvious solution would be to simply have each unique state and unique county be its own unique number, it is important to consider the fact that two counties can have the same name in separate states. To ensure that counties with the same name are not given the same number, it would be best to concatenate the name and state into one column and use the concatenation to create the unique identifying number. Another scale of measurement that should be considered given the context of the problem is the Minkowski distance as is mixes both Euclidean and Manhattan distances in a manner that allows it to emphasize larger distances between features while also being resilient to outliers and larger variation in the dataset.

The bar plot in Figure 5 shows a comparison of the correlation values between the original data and the Euclidean and Manhattan distance matrices. The plot indicates which method is more correlated with the original data by showing the higher correlation value for that method. As expected, we see that the Manhattan distance is more effective in capturing the underlying structure of the data than the Euclidean distance metric. Furthermore, the Minkowski distance performs the worst of the three. This could be because, with p=3, the Minkowski distance may be too sensitive to the differences between the various attributes in the dataset, causing Minkowski not to consider the overall patterns as well. Given the comparison, we will go with Manhattan as our scale of measurement.

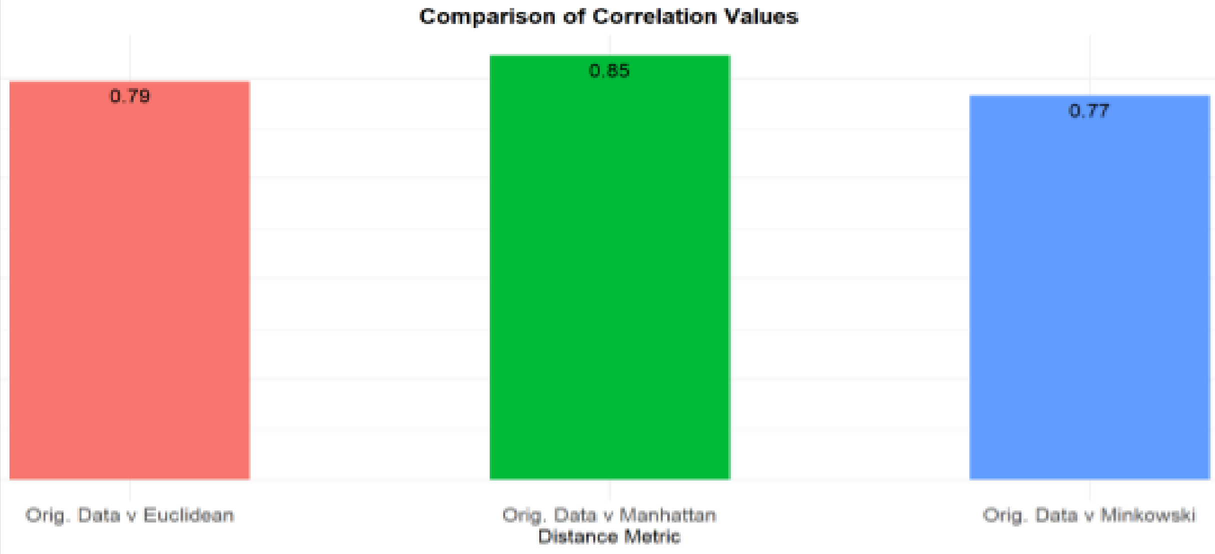


Figure 5: Comparing Representation Performance of Euclidean vs Manhattan Distance vs Minkowski

Modeling

- Determining the Suitable Number of Clusters:

In order to determine the suitable number of clusters, I have relied on three methods. Firstly, finding the average silhouette score at each K value, and secondly, utilizing the elbow method. The average silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to 1, where a higher score indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters. Thus, a higher Average Silhouette at a certain K value could be indicative of the ideal number of clusters to use. Figure 6 shows the average score for each number of clusters that we add. As we add clusters, we see that the data is becoming over-segmented and that the clusters are becoming too small. Thus, these smaller sub-clusters end up having points that have greater and greater relations with neighbors versus a particular point's own cluster. The maximum silhouette score occurs at K=2, which means that 2 clusters may be enough to well represent the data in clustering. In other words, it suggests that the data can be divided into two distinct clusters that are well-separated from each other.

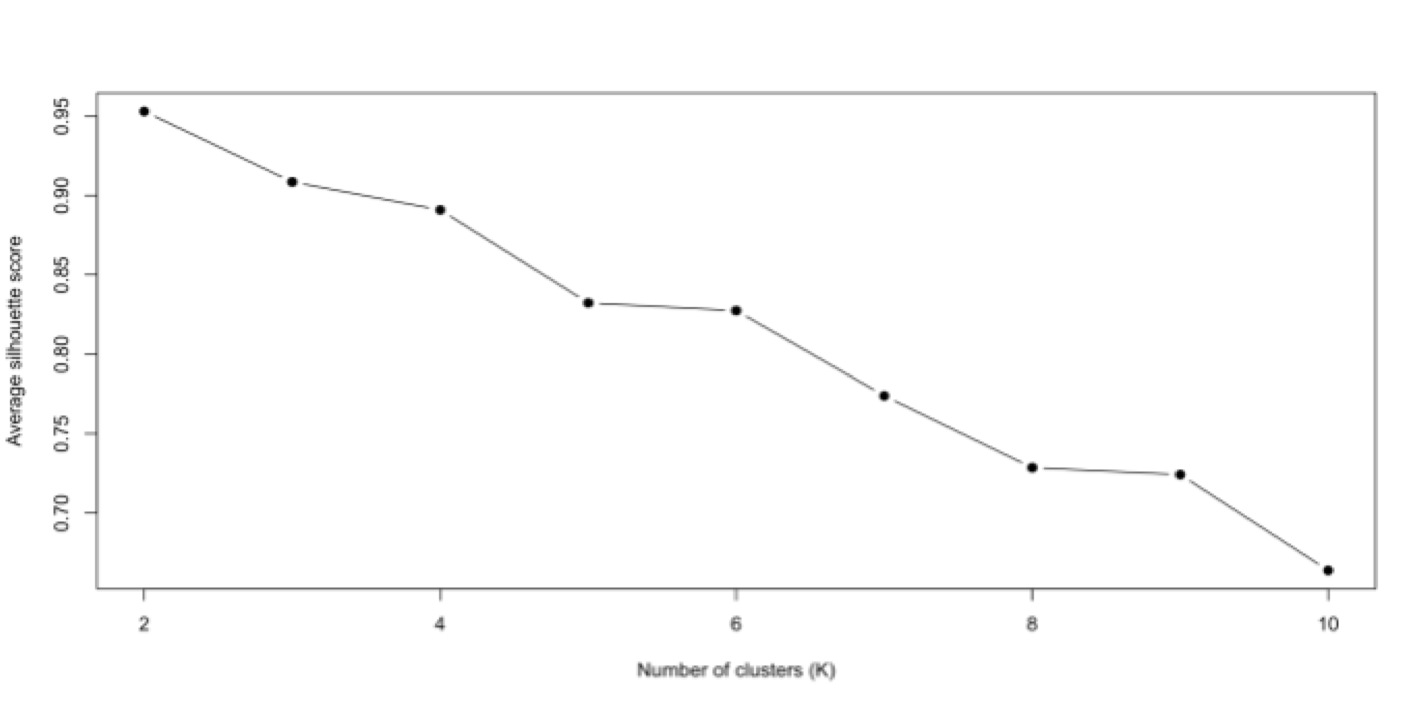


Figure 6: Average Silhouette Score

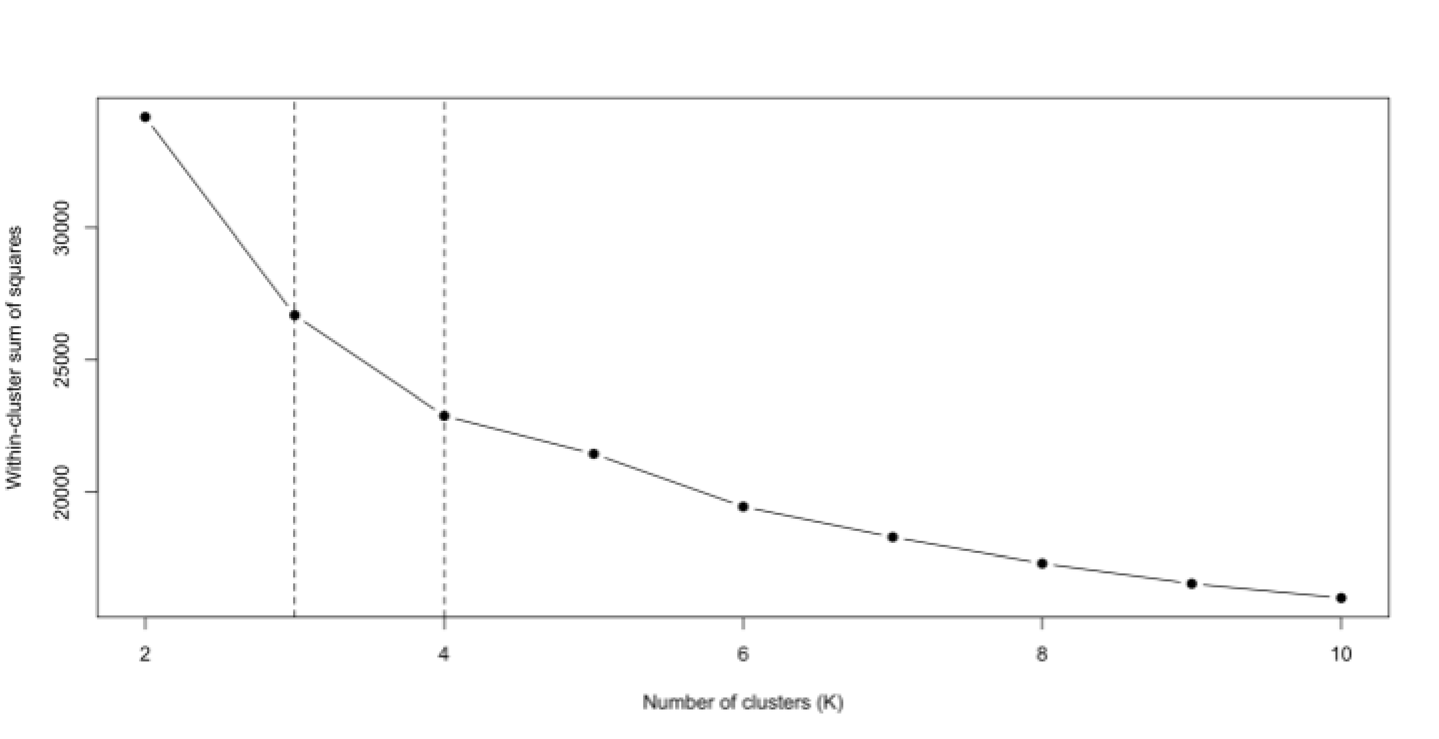


Figure 7: Elbow Method

The elbow method is another technique that should be considered. The main differentiator between the Silhouette method and the elbow method is that the elbow method measures the sum of the squared Euclidean distances between each point in a cluster and its centroid, and it measures how compact the clusters are. The elbow graph ultimately allows us to isolate the point where there is enough data to account for the variation in the data, but not enough where overfitting occurs that ruins adaptability.

Figure 7 shows the cluster sum of squares depending on the # of clusters. It represents the sum of the squared distances between each point in a cluster and the centroid of that cluster, where the centroid is considered roughly the middle average of a cluster. The vertical dotted lines show the two “elbows: in the graph at k=3 and k=4, suggesting that 3 or 4 clusters should be used to minimize the variance for each cluster, ensuring that the clusters are compact and well separated. Because the silhouette score and elbow method give different results, it is important to consider what each method does in choosing a k value. The elbow method is based on the within-cluster sum of squares (WSS) and looks for a point where the WSS stops decreasing rapidly as the number of clusters increases. This method may be appropriate for data where the clusters are compact and well-separated, and where the goal is to minimize the variance within each cluster. The silhouette method instead measures the degree of separation between clusters and looks for a number of clusters that maximizes the mean silhouette width. This method may be more appropriate for data where the clusters are less well-defined or where the goal is to maximize the separation between clusters. Given we haven’t actually begun clustering yet, we may choose to do one more test to determine the optimal amount of clusters.

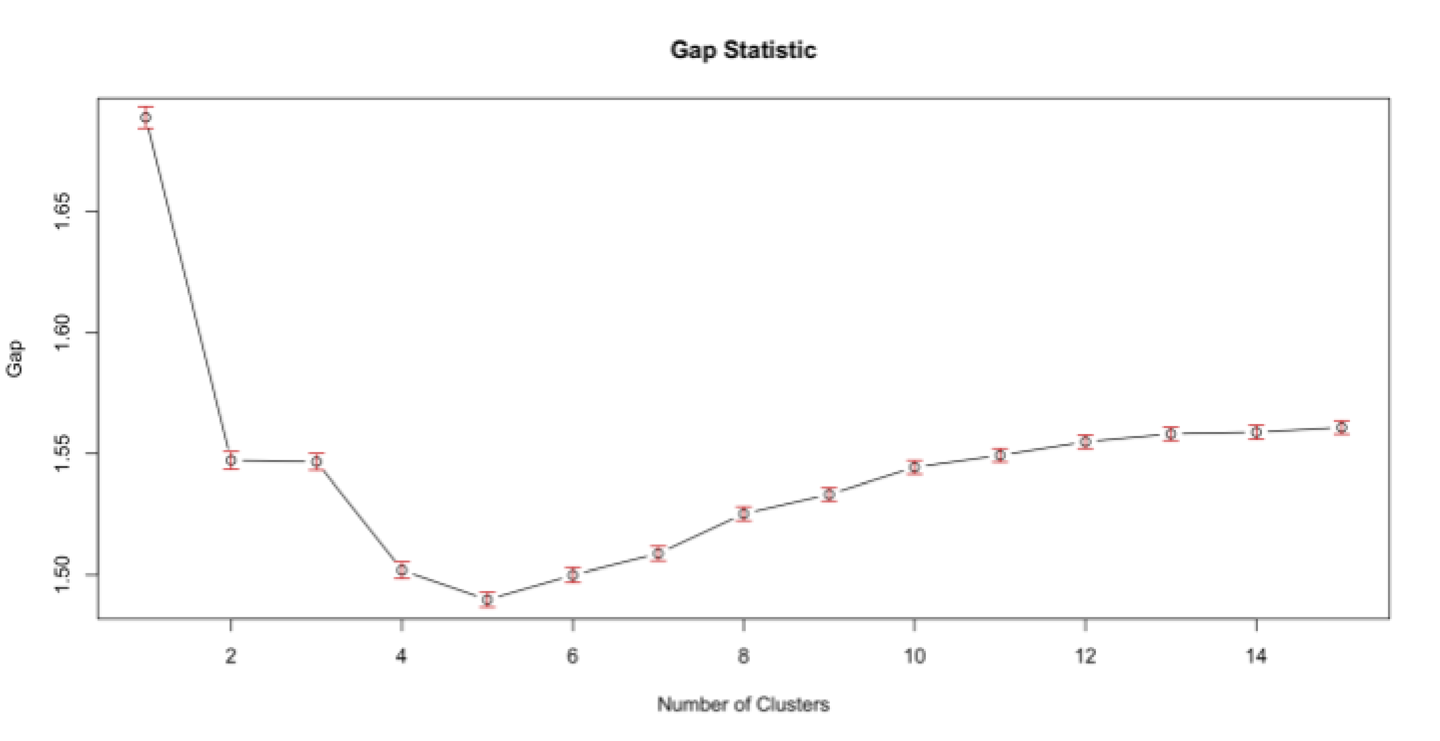


Figure 8: Gap Statistics Result

The final test I performed for determining the suitable number of clusters was plotting the gap statistics. One benefit provided by this method is that it is better at dealing with noise and increased variability in the COVID-19 statistics to determine the clusters. Looking at Figure 8, we are given the standard error for each # of clusters. The optimal number of clusters is the smallest value of k such that the Gap statistic for that k is greater than or equal to the Gap statistic for k+1 minus the SE for k+1. This means we need to find the part of the graph where the SE first begins to level. The graph below shows us that a k value of 2 or 3 is preferred, but not 4. With Figures 6 through 8, we can see that k=3 remains a solid option that is recommended by the gap method, and elbow method, while also maintaining a high silhouette score.

❖ K-Means Clustering

➢ The Results of Clustering

The clustering graph shows the results of the k-means clustering algorithm with 3 clusters on two dimensions that represent linear combinations of the 10 attributes. There are 1369 counties in the first cluster, 1292 in the second, and 338 in the third. The graph displays the scatterplot of the standardized data, where each point represents a county, and the color of the point indicates the cluster to which it belongs. The ellipses surrounding the points represent the 95% confidence intervals for each cluster. From the graph, we can observe that the counties are clustered into 3 groups, with major overlap between the groups (particularly clusters 1 and 3). This suggests that the feature values of cluster 3 are similar to those found in cluster 1. The % next to Dim1 and Dim2 represents the % of the variation in the data explained by the two principal components. We can see that dimension 1 captures a large amount of variation within the data.

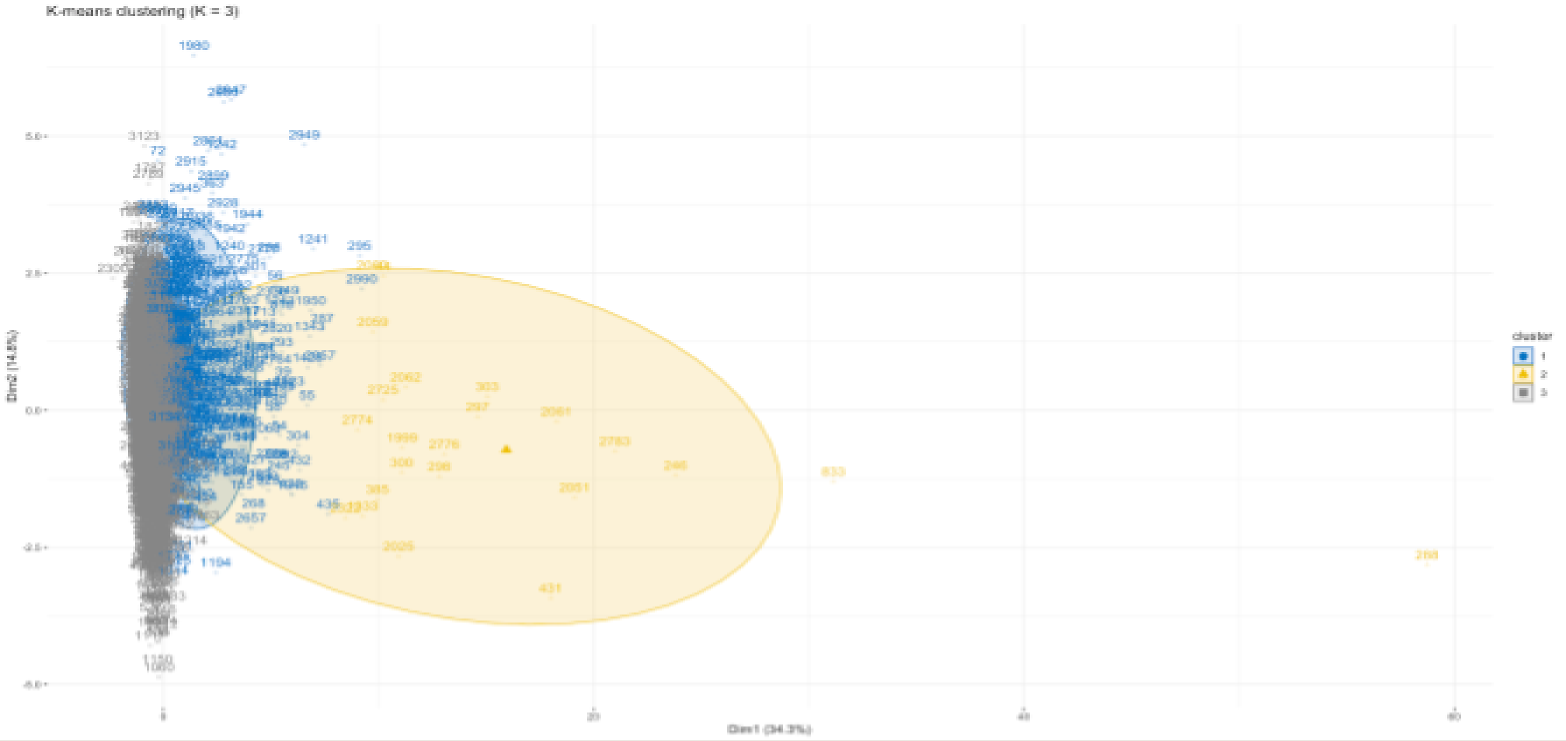


Figure 9: K Means Clustering

The clustering summary below shows the mean and standard deviation for each of the clusters for each attribute. This summary allows us to compare the different clusters and identify any differences or similarities between them. Interesting statistics to note include the fact that cluster 3 has a lower median_year_structure_built mean and standard deviation compared to clusters 1 and especially 2. This could suggest that cluster 1 contains counties with older structures on average. Additionally, we can see that cluster 3 has a negative mean for Percent_income_spent_on_ rent and Deaths, indicating that this cluster has a lower number of deaths and a lower percentage of income spent on rent. Cluster 2 had higher means for Deaths, median_year_structure_built, and Percent_income_ spent_ o n_rent. This suggests that these variables may be important in distinguishing the counties in cluster 2 from the other counties. Cluster 3, on the other hand, never had a negative mean for all attributes except Population and Percent_Male. We can further see that cluster 3 consists of counties with lower populations, lower incomes, less education, and less money spent on housing compared to the other clusters. The fact that cluster 3 has fewer deaths than the dataset average does suggest that income and education as attributes do not explain the data and contribute to deaths as much as population size. This is further proven demonstrated by Figures 10 and 11 below.

Cluster Summary: K-Means

Attribute	Mean	SD
Population_1_year_and_over	0.5826188 7.9697314 0.2069930	0.9501528 5.6596138 0.1414483
Deaths	0.3875369 8.2681642 -0.1657900	0.8725480 6.0719881 0.1421648
median_year_structure_built	0.5709521 -0.5687047 -0.1234767	0.9820240 1.4544142 0.9522714
Percent_income_spent_on_rent	0.4360995 1.0853960 -0.1087095	0.7052185 0.7121902 1.0259891
Percent_Male	-0.31433281 -0.51088878 0.07578452	0.52573370 0.39553676 1.06818941
median_income	1.2360357 0.9802191 -0.2882589	1.1581257 1.3321226 0.6925997
Percent_graduate_or_more	1.5245501 1.2945879 -0.3563546	1.1742852 1.4386374 0.4878489
genie_index	0.11913432 1.16519683 -0.03792633	1.05272981 1.04284160 0.97883880
County_ID	0.05926835 0.16113268 -0.01490326	0.99897962 0.96762156 1.00029989

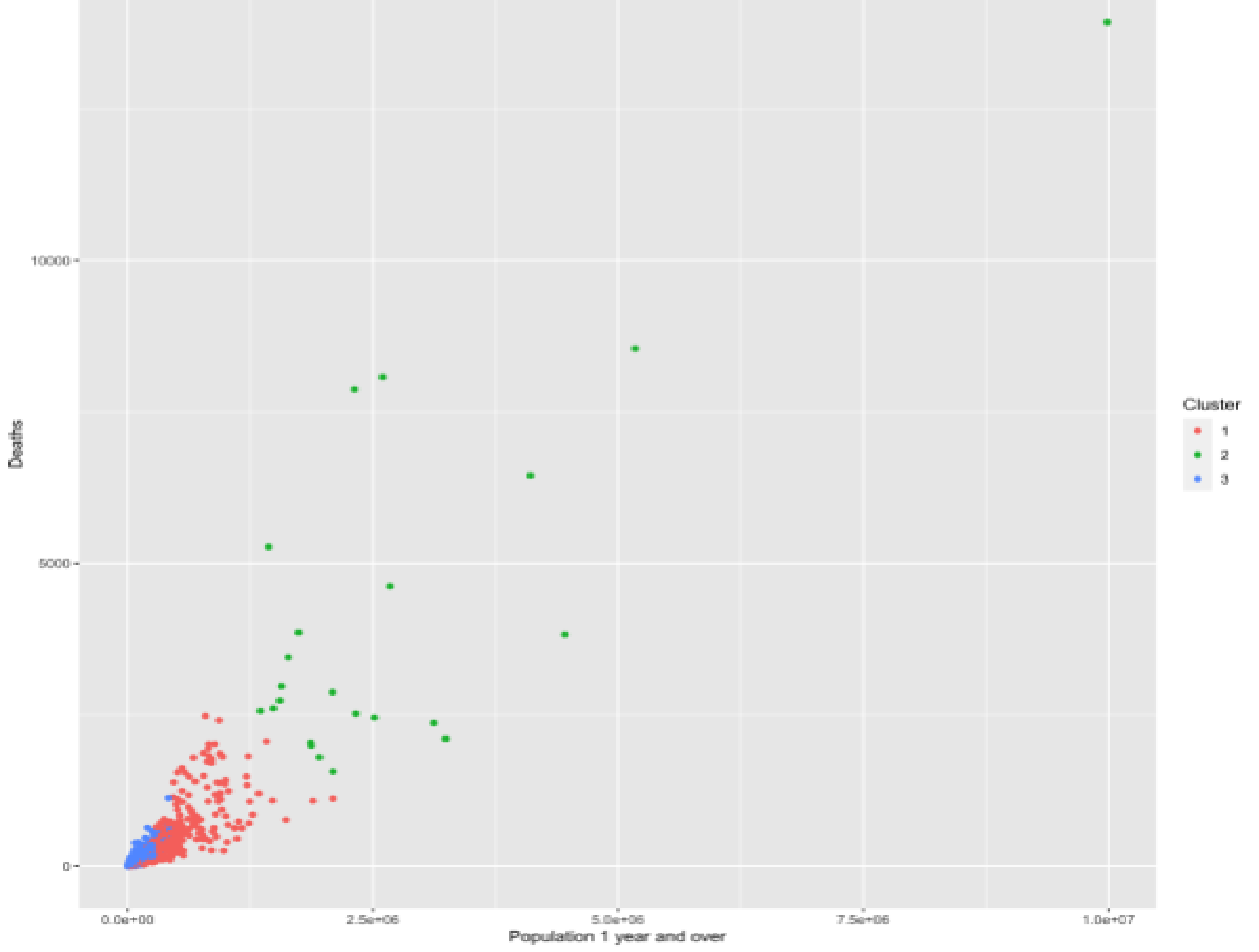


Figure 10: Population v. Deaths by Cluster

Figure 10 shows us how cluster 2 (which was the yellow cluster in Figure 9) is separated from the other clusters by population size and death. This is not surprising as we learned in the data preparation phase that these two variables are significant contributors in determining the values for each data point and have a positive correlation with each other. It is important to note that the third cluster overlaps with the first, but we can now see that it differentiates itself by being representative of smaller counties with fewer deaths. Looking at Figure 11, we see that the median year of the age of buildings in a county (which is representative of the overall age of the county’s infrastructure) shows that counties with older buildings have a far more variable amount of COVID deaths while modern counties with builds from the 1990s and onwards have low deaths with virtually no high outliers. Though the cause cannot be concluded from associations, it is possible that areas with older buildings are more likely to be areas with less advanced health infrastructure, fewer health department resources, and a higher Gini index.

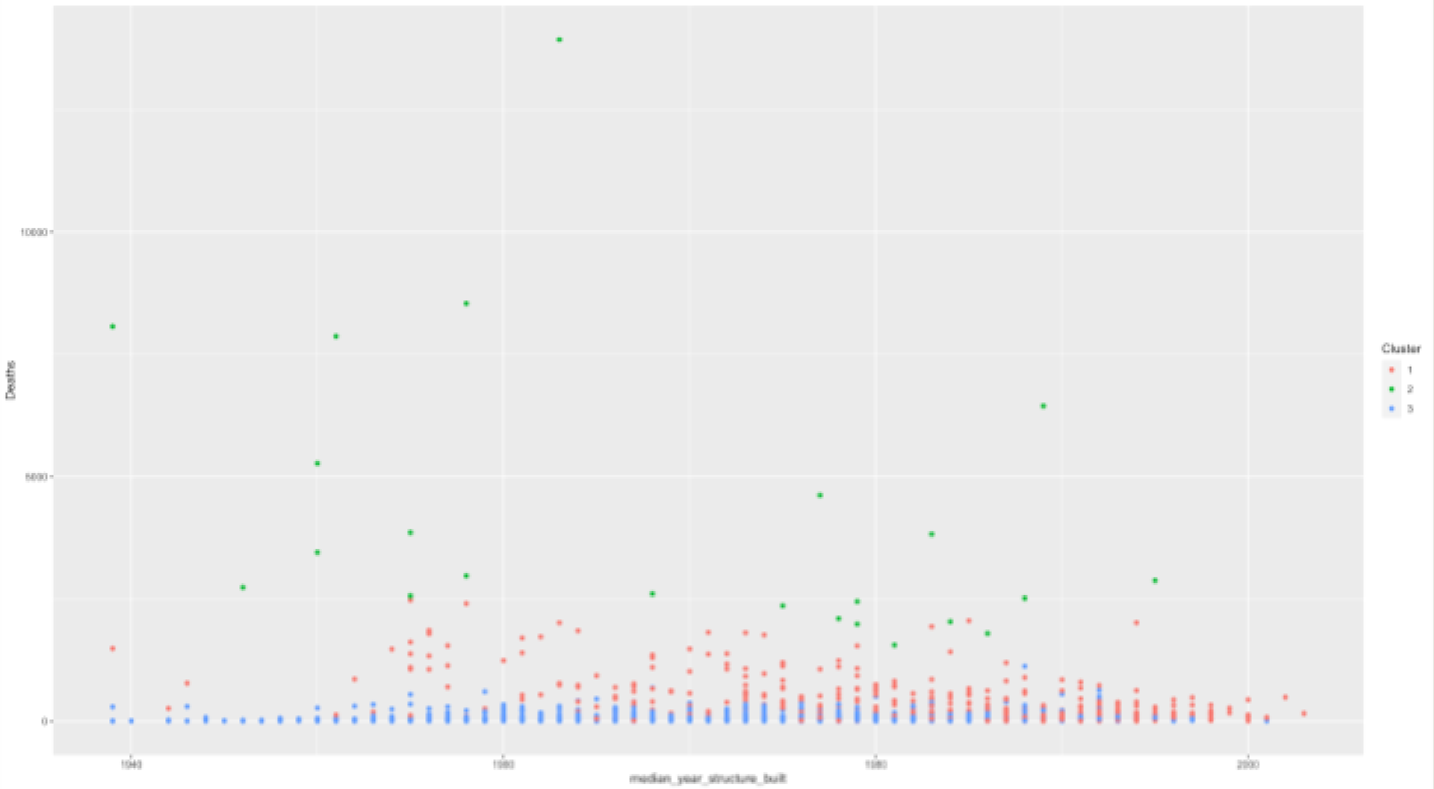


Figure 11: Median_year_structure_built v Deaths

➢ Internal Validation Method and Results

The internal validation method that I chose to utilize is the Dunn statistic. This measure evaluates the compactness of each cluster relative to the separation between clusters. A higher Dunn index may be an indicator of better clustering. Looking at the clustering statistics, I found that the Dunn Index is -0.14 which suggests that the clustering solution may not be ideal. Furthermore, the average Silhouette width for clusters 1 and 3 was -0.90 and -0.64, which is not surprising considering that both clusters overlap. This also suggests that there are data points in cluster 1 that are more similar to points in cluster 3 and vice-versa. Another statistic of concern is the Pearson gamma, which is -0.001. This statistic measures the correlation coefficient between the pairwise distances in the original feature space and the pairwise distances in the reduced feature space (i.e., after clustering). Thus, A value close to zero suggests that there is no significant correlation between the original distances and the reduced distances. This is backed up by the s-statistic, which demonstrates the level of stability for the clustering algorithm. We find a stability of -4.18, which suggests that the clustering solution is less stable than if we were to create random partitions in our clustering. Given the performance of our clustering model, one resolution that I tried was changing the scale method from Z-score standardization, which scales the features of the data to have a mean of 0 and a standard deviation of 1, to min-max scaling, which normalizes the data between the range of 0 and 1. However, this did not improve clustering results by much, with there still being significant overlap between the clusters. The Dunn statistic rose from -0.14 to 0 and the silhouette widths to -0.02 and -0.04. However, this still is not high enough to be able to conclude that the clusters are distinguished from each other. One other method to solve this issue of performance would be to try a different clustering algorithm, which I will do next.

❖ Hierarchical Clustering

➢ Implementation

The means of determining the number of clusters to create is similar for both hierarchical clustering and k-means clustering. However, one other means of determining the number of clusters that I can consider is a dendrogram, which displays the hierarchical relationships between the data points and can be visually inspected to identify natural clusters in the data. When the dendrogram begins to split off and as the height increases, we can identify the ideal number of clusters. In essence, The number of clusters can be determined based on the number of vertical lines that can be drawn through the dendrogram without crossing any horizontal lines. Each vertical line corresponds to a cluster solution. Figure 12 below suggests that 2 clusters remain a good option while 3 clusters may not be as good an option, as shown by the red and blue lines cutting through two vertical lines representing two distinct clusters. Therefore, for hierarchical clustering, I will use 2 clusters. Regarding the normalization technique, I have learned that min-max is a better option than z-standardization as COVID data includes many outliers regarding population size and COVID deaths. And regarding the distance method, I will keep Manhattan distance as my method.

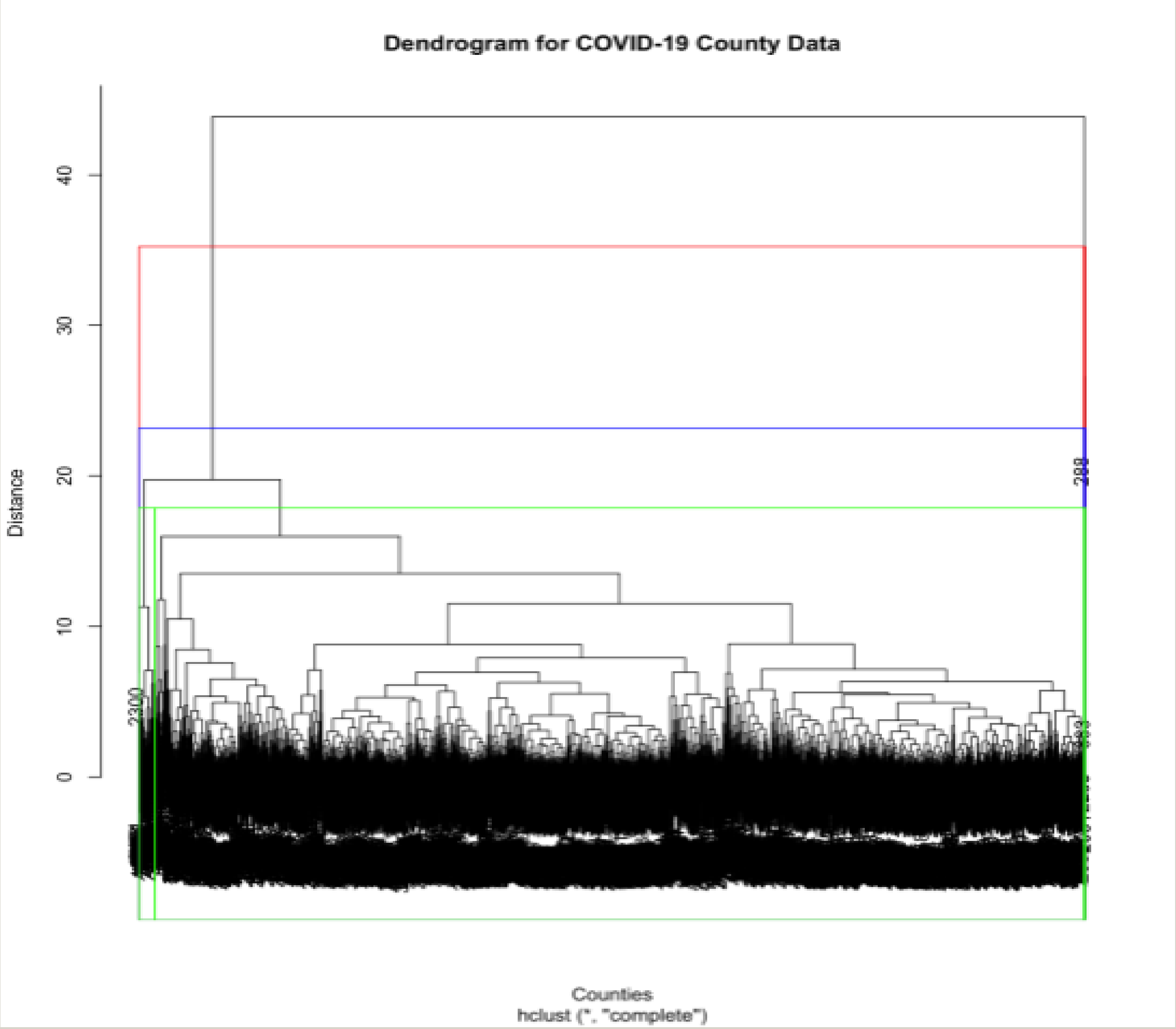


Figure 12: Dedrogram

➢ Results and Internal Validation Method

Having decided on the implementation, we can now create an agglomerative model with 2 clusters shown in our dendrogram and can visualize different combinations of 2 attributes.

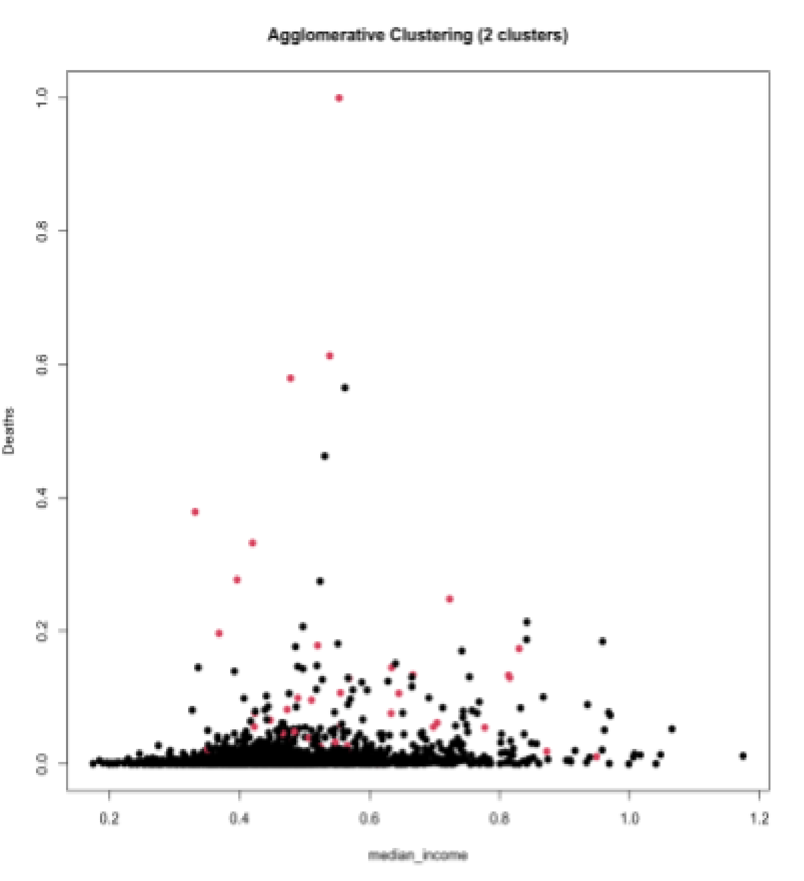


Figure 13: Agglomerative Clusters: Median income v. Deaths

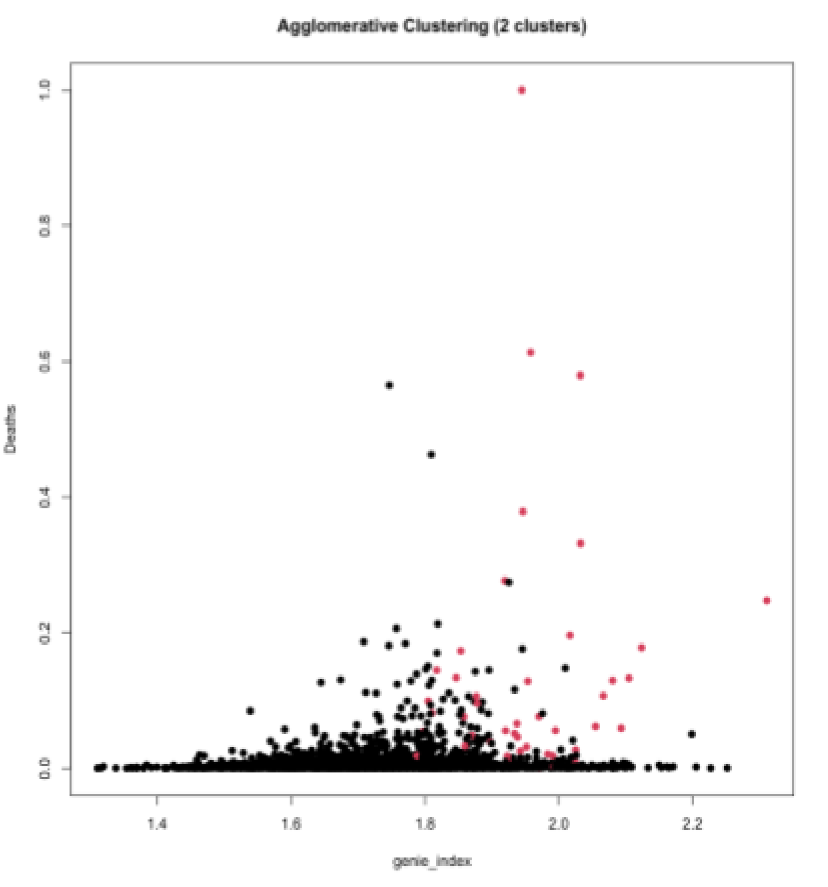


Figure 14: Gini_Index v. Deaths

For the above figures, I decided to utilize median income and gini index as a method of determining whether two distinct clusters relating deaths and median income can be formed. We find that while genie_index and median income both result in the clusters overlapping, genie index better represents the distinctions between the two clusters as we see that a county with a higher gini index will have a higher chance of also having a higher amount of deaths than average and are also more likely to be a part of the 2nd cluster. As it pertains to the context of the goal of clustering, this suggests that whether a place has higher income inequality should be a factor that contributes to whether a county gets the vaccine sooner or later.

Looking at some of the other agglomerative models for different attributes, I see population appears to be the only attribute that shows a clear distinction between the attributes. To get a better overview of the distinctions between the two clusters, I will get the mean and standard deviation for each attribute for both clusters.

<u>Cluster Summary: Hierarchal</u>

Attribute	**Mean**	**SD**
Population_1_year_and_over	0.008994818	0.023813384

	0.114269126	0.160353785
Deaths	0.007462492 0.134401709	0.022082359 0.189156671
median_year_structure_built	30.8680973 30.6604167	0.1722693 0.2109612
Percent_income_spent_on_rent	0.72400819 0.80518135	0.10906220 0.05673892
Percent_Male	1.33568038 1.29100670	0.06170896 0.02148067
median_income	0.4498931 0.5677050	0.1170693 0.1440410
Percent_graduate_or_more	0.16337505 0.38202928	0.09 0.13
genie_index	1.7178108 1.9531408	0.1311267 0.1018338
County_ID	0.5023055 0.3708843	0.2887962 0.2623353

The cluster summary shows the mean and standard deviation of various attributes of the two clusters created through hierarchical clustering. In general, the clusters seem to differ in most of the attributes. Cluster 2 has a higher mean value of population, deaths, median year structure built, percent income spent on rent, median income, percent graduate or more, genie index, and a lower mean value of percent male. On the other hand, Cluster 1 has a lower mean value of the above attributes and a higher mean value of percent male. As it pertains to mean, the largest % change in differences from one cluster to another cluster occurs with the percentage of individuals with a graduate degree or more, suggesting that the two clusters differentiate significantly based on that statistic. In fact, a graduate degree seems to be a greater determiner of whether you are in cluster 1 or 2 than even population or death rate.

It is also important to note that, unlike our previous clustering with k-means, the median year of a structure is not a differentiator as the overall age and distribution of the average building age for clusters 1 and 2 is virtually identical.

Figure 15 gives us the Dunn index for both clusters. Though there is no definite definition of what is a good and bad index value, the fact that both of the Dunn index values are >0 suggests that our hierarchical clustering may be a better option than K-means. Another statistic to determine internal validation would be the average silhouette score:

the average silhouette score for Cluster 1 is 0.3472666, which is a fairly good score indicating that the data points are well-clustered. The score for Cluster 2 is 0.2494469, which is not as good as Cluster 1, but still reasonable. But most importantly, these values are markedly better than the negative avg. silhouette scores received for k-means analysis. Thus, where the K-means suffered from severe overlapping and misclassification of points in the wrong cluster, hierarchical clustering has been able to successfully differentiate the points in each cluster based on the attributes.

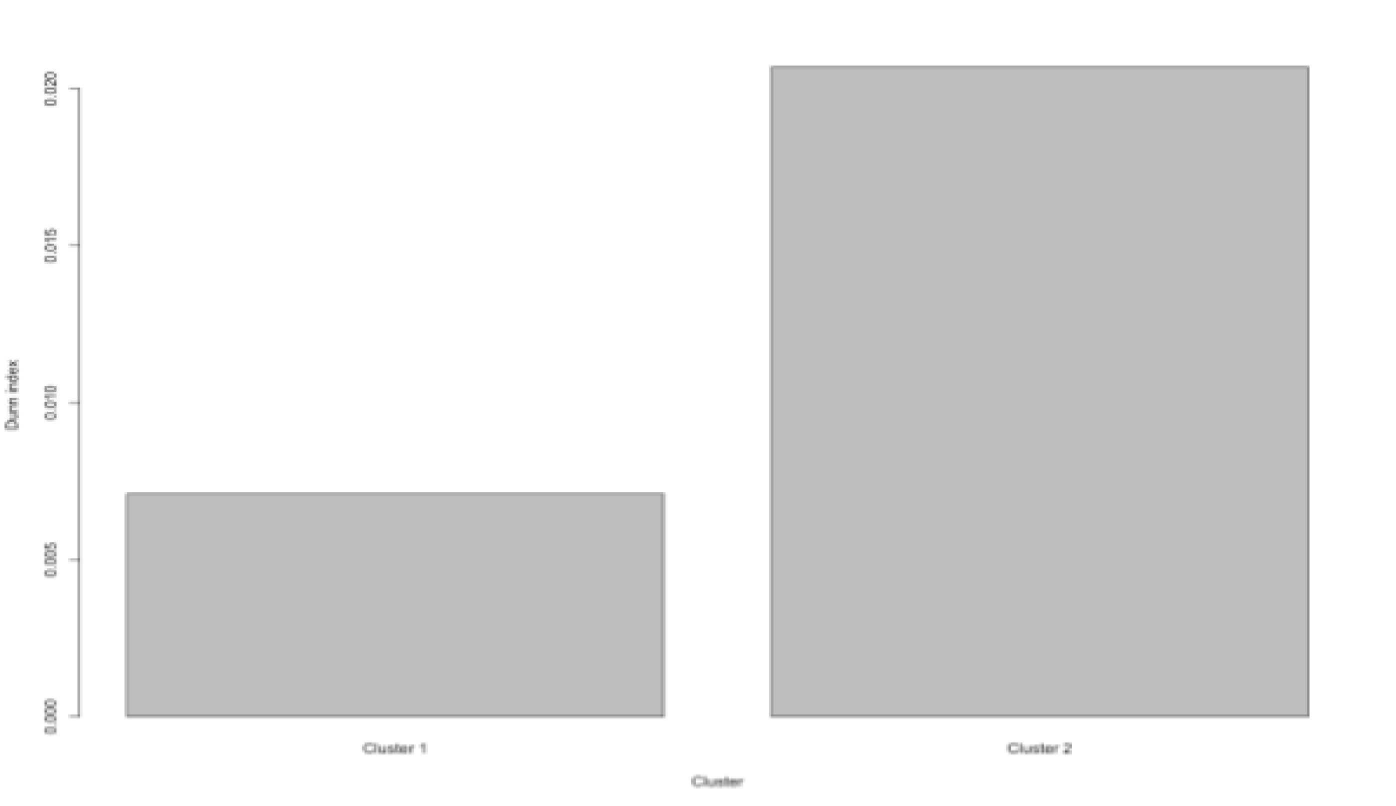


Figure 15: Dunn Indices for Hierarchical Cluster

- ❖ Fuzzy Clustering
  - ➢ Implementation

Determining the number of clusters for fuzzy clustering is not as straightforward as in other clustering methods, such as K-means or hierarchical. In fuzzy clustering, each data point can belong to multiple clusters to varying degrees. However, one method of determining the number of clusters is the elbow method which relies on the SSE from the centroid to the data to determine the optimal number of clusters. In Figure 16, the FPC is a ratio of the sum of the squares of membership values to the sum of the squares of distances from the centroid. In other words, it measures the degree to which a data point is close to its own cluster center relative to the distance from all other cluster centers. We see that the points that minimize overlap but maximize cluster separation are where the # of clusters is 3 or 4, which is the same result we got for the k-means elbow method.

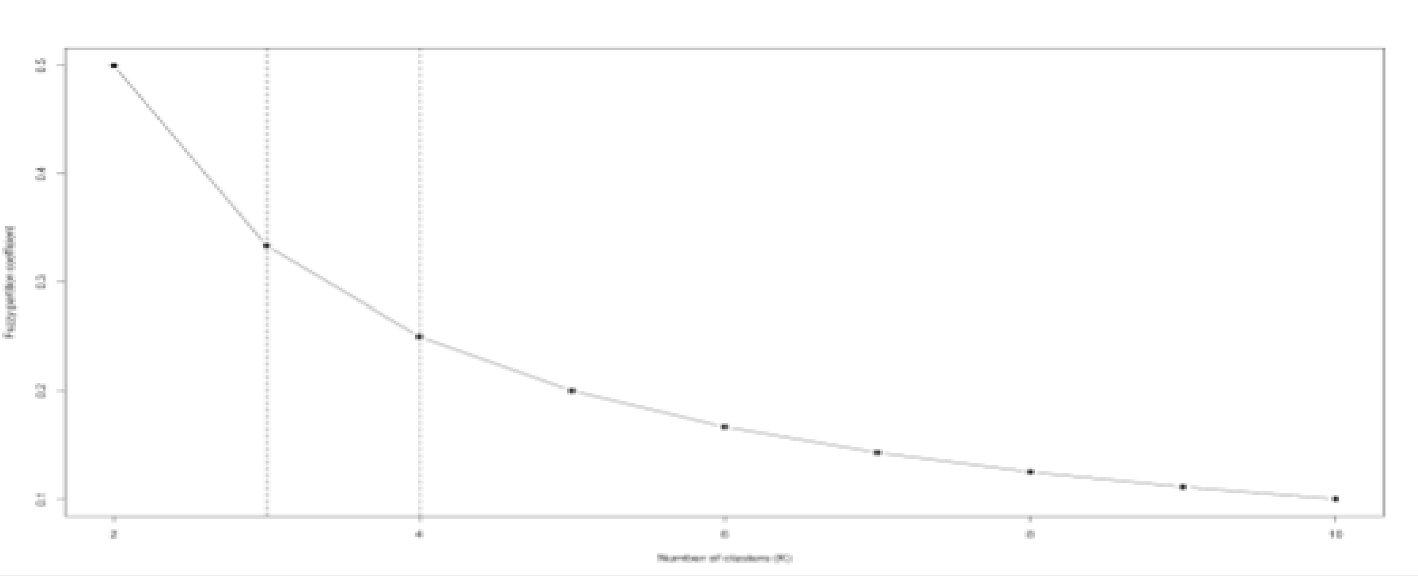


Figure 16: Elbow Method on Fuzzy Cluster

Another method of determining the optimal number of clusters is the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Considering the dataset, I note that my data has a lot of attributes and many rows, which means it is a high-complexity model. This is significant as BIC is known to perform worse on complex models while AIC is better. And since there is not too much redundancy in my model and the risk of overfitting, the benefits that would make me choose BIC are not valid in this instance. Therefore, I will display the AIC values below. The fact that the AIC values increase linearly per the number of clusters shows that the smallest possible amount is the leftmost number of clusters, two. However, it is important to know that this contradicts the elbow method above. Thus, it is important to note that AIC takes into account both the goodness of fit and the complexity of the model, while the elbow method only considers clustering quality regarding SSE difference. Thus, AIC may be better relied on for data of high complexity at the expense of the fit of the model while elbow, though certainly a valid method, would have greater credence in simple models. Therefore, for this clustering, I will still go with 2 clusters.

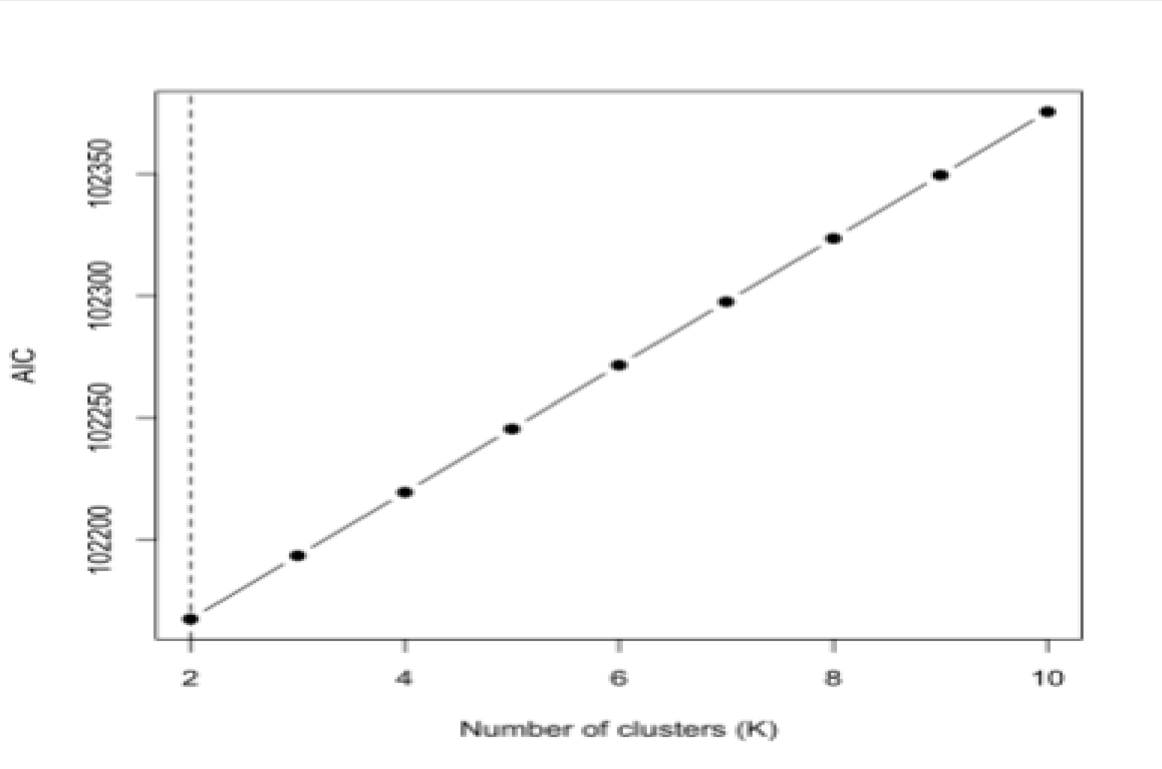


Figure 17: AIC v. Number of Clusters

➢ The Results and Internal Validation

Clustering Statistics:

	Cluster 1	**Cluster 2**
Membership Mean	0.4999703	0.5000297
Membership SD	0.001323484	0.001323484

Cluster Size	1505	1494
Average Silhouette widths	0.1743263	0.1253001
Individual silhouette widths (minimum)	-0.04276	
Individual silhouette widths (1st quartile)	0.07175	
Individual silhouette widths (median)	0.14682	
Individual silhouette widths (mean)	0.14990	
Individual silhouette widths (3rd quartile)	0.22824	
Individual silhouette widths (maximum)	0.35535	

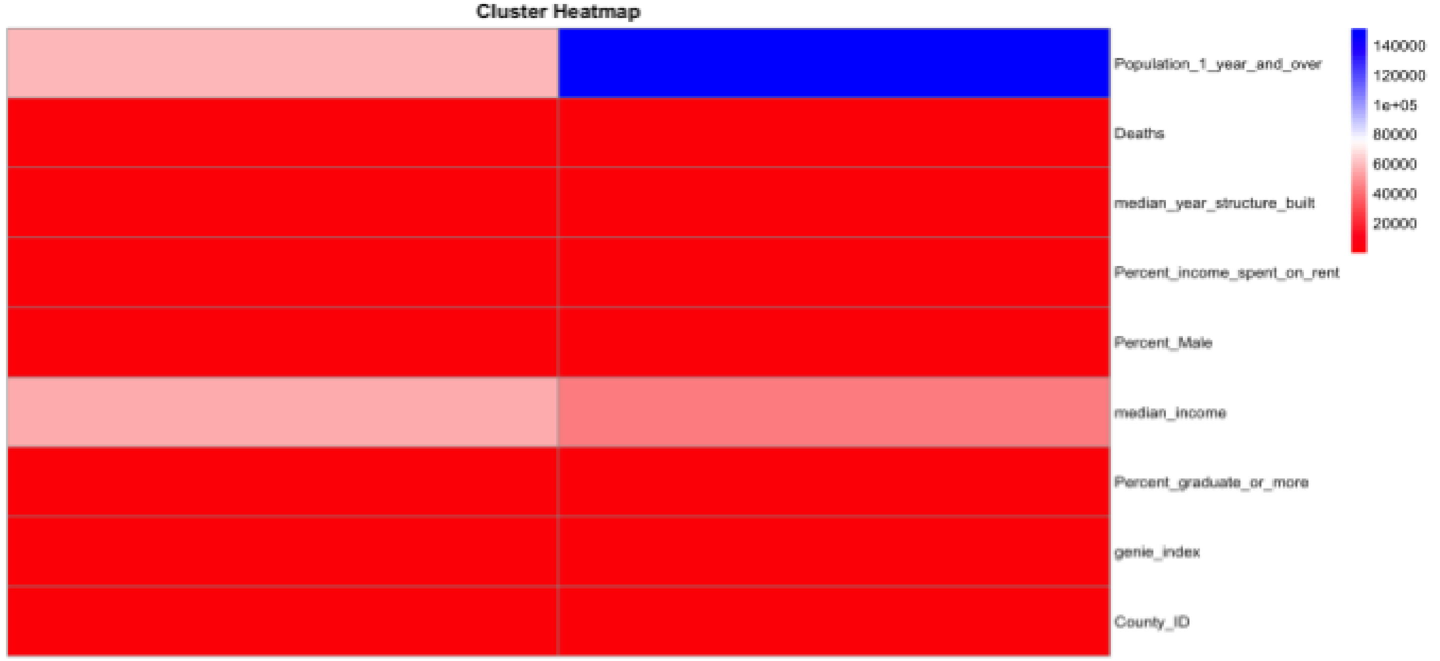


Figure 18: Fuzzy Clustering Heatmap

The heat map above shows the results of the clustering by depicting the distribution of each feature within the cluster. Red indicates a high value, blue indicates a lower value, and white indicates a moderate value for that particular feature. We can see that the colors across all clusters are mainly red, which suggests that the values between colors may actually be quite similar. Thus, we can conclude that all the variables in bright red are probably not good attributes to distinguish the clusters. However, we do see that median income and population are two attributes that have differences between the two clusters, with the 2nd cluster having a low population amount and moderately-high median income and the first cluster having a moderately high population and less median income.

In order to compare the performance of fuzzy clustering in relation to prior clustering methods, we can look at the average silhouette width of 0.17 and 0.12 respectively. These average silhouette widths suggest that the clusters may not be well separated, but are not as intermixed as K-means. However, hierarchical clustering has been shown to have much higher silhouette widths. Another area of concern for this clustering method is shown through the membership mean, which is almost identical between the two clusters. This suggests that the two clusters are similarly well-defined and that there is not a clear separation between them based on the attributes.

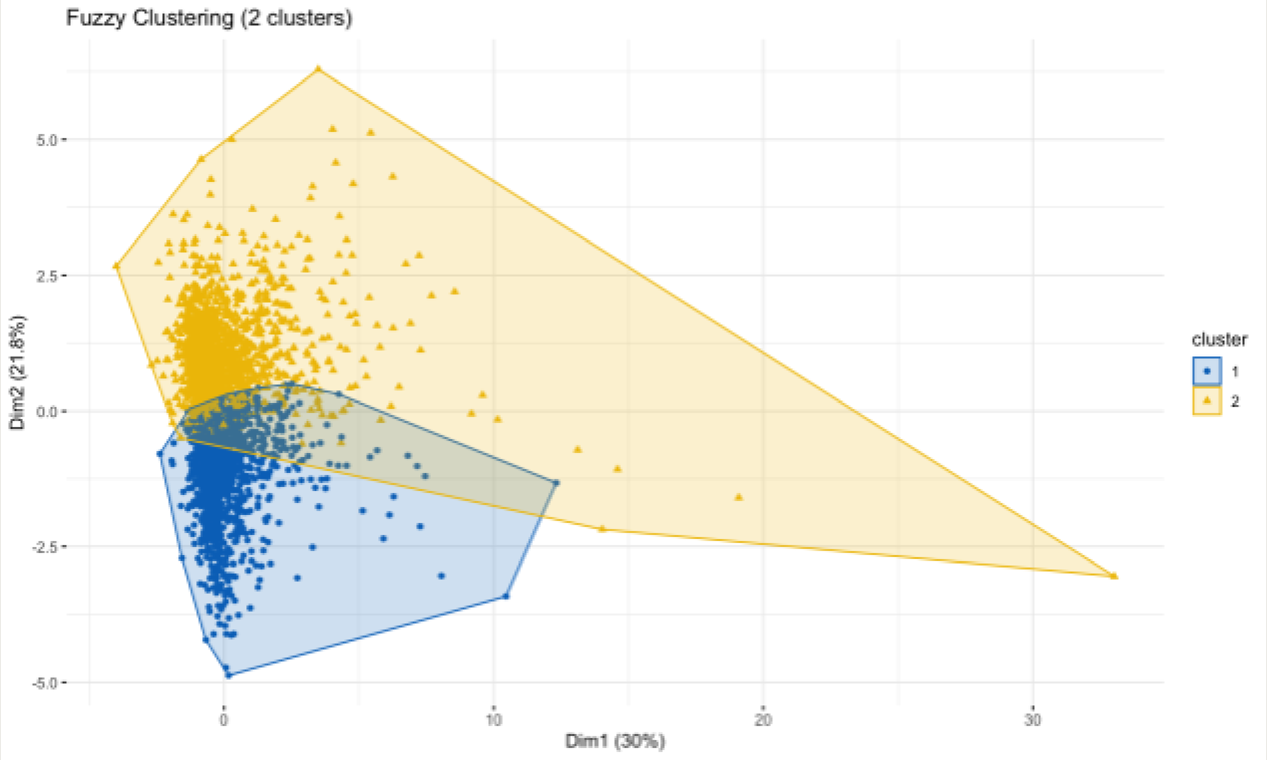


Figure 19: Fuzzy Clustering:

Figure 19 shows the results of 2 fuzzy clusters. We confirm that many of the counties do fall within the overlap between the two clusters, which suggests that a large portion of the points cannot be properly distinguished. That being said, 30% and 21.8% state that the principal components are able to explain a small, yet not insignificant amount of the variance within the data.

❖ External Validation

➢ K-Means

For external validation, I decided to use deaths as the grounding truth for my dataset due to the goal of finding those who are in need of the vaccine the most. The first statistic that I utilized for external validation was the adjusted rand score, which indicates the amount of similarity between two labels with 1 being heavily related, 0 being random, and -1 being disagreement. In the case of my k-means clustering score, the ARI score of 1.243292e-05 indicates that the clustering algorithm is not able to accurately identify the true groupings of the data, and the clustering results are practically random. This may be due to various factors, such as the choice of the clustering algorithm, and the selection of features. It is important to note that K-Means was our least-performing model of the three.

➢ Fuzzy

For fuzzy clustering, I decided to utilize a PCA plot to visualize the clustering results and the ground truth labels in a low-dimensional space. The data points are projected onto the first two or three principal components, which capture the majority of the variation in the data. The color or shape of the data points is used to indicate their cluster membership or ground truth label. Figure 20 shows each county with the features represented in a two-dimensional PCA. The counties with higher deaths are shown by the larger circles. One characteristic that I notice is that the more we move along Dim1, the greater the number of deaths is. And when we look at this in conjunction with Figure 19, we see that cluster 2 appears to be “capturing” the area around those counties that have a larger number of deaths overall.

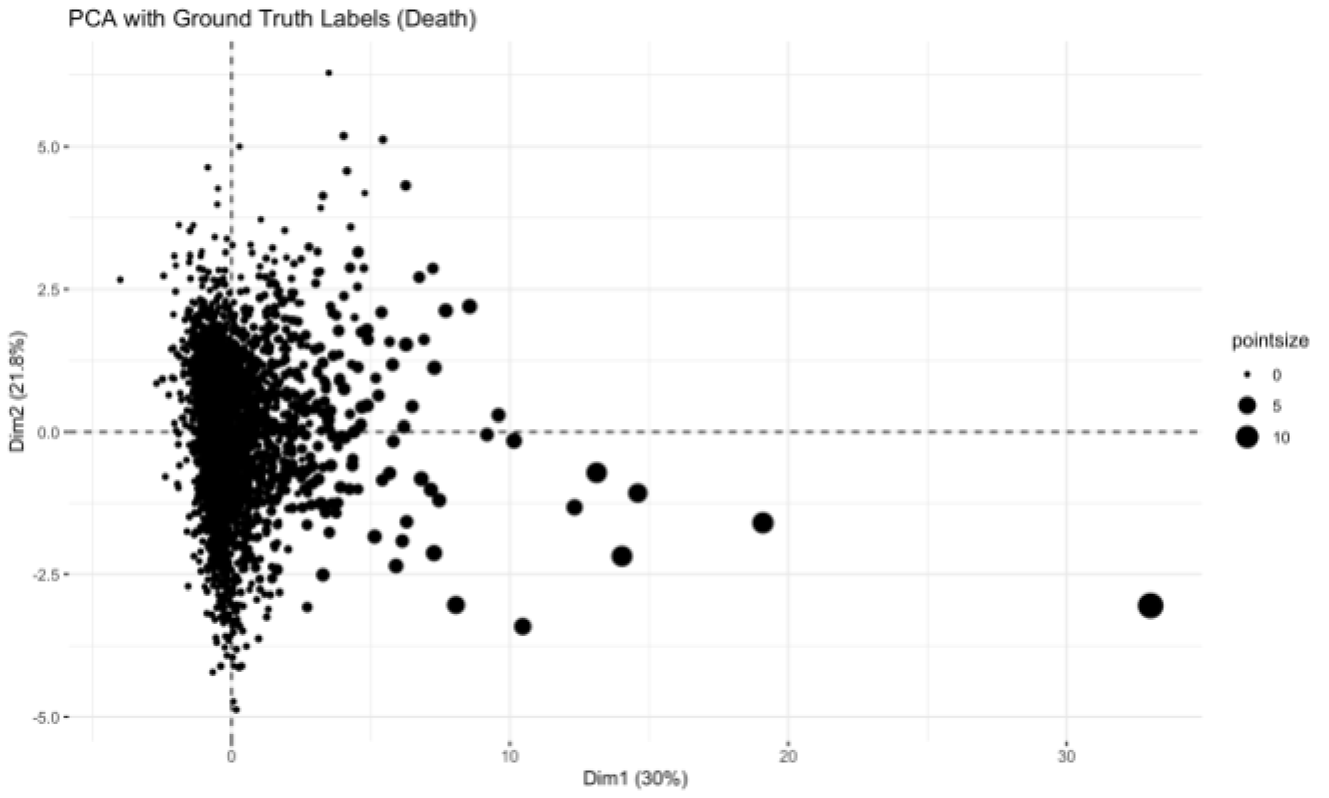


Figure 20: PCA for Fuzzy Clustering

➢ Hierarchical Clustering

For hierarchical clustering, I have decided to start off by finding the adjusted rand index to measure the similarity between the clustering results and the ground truth labels. An ARI score of 0.38 shows that my hierarchical clustering algorithm clustering has some degree of similarity with the ground truth labels, but there is also some degree of mismatch. In other words, the clustering is not perfect, but it is better than random.

Confusion Matrix

Truth Labels	High	Low
High prediction	619	2333
Low prediction	8	39

The Confusion Matrix above shows the represents the performance of a Hierarchal clustering model that predicts whether the number of deaths is in the "High" or "Low" category based on the features. From this confusion matrix, we see that only a total of 658 of the truth labels have been classified correctly while 2341 have been classified incorrectly. Though this may seem terrible, it is important to note that there were only 8 instances where the model predicted low but the actual county deaths were high. On the other hand, there were 2333 false positives- where the model though the amount of COVID was high but it was indeed low. This is important as it is far better to have resources available to districts that end up struggling less than anticipated (false positive) than not having the resources available in a district that ends up with a COVID outbreak with many deaths (false negative).

Overall Evaluation

❖ Interesting Findings

The first essential finding that I can conclude from the clusterings is that population is a key trait that can distinguish one cluster from another in k-means and fuzzy clustering. The K-means method in Figure 10 shows that perhaps we can split the data into 3 separate clusters based on the size of the population, which corresponds to the number of deaths (low, medium, or high). And when looking at the fuzzy clustering heatmap, we can see that the only differentiator that the clustering algorithm was able to identify for the given attributes is whether the population is classified as low or high. Factors like the genie index and median income are worse factors than population but are the two other actors that show a small level of success. This is because the clustering algorithm is able to identify some of the counties with higher genie index or lower median income as being in another cluster with generally higher COVID deaths. In comparing the different clustering performances, the clustering results and statistics appear to point to Hierarchical clustering having the best performance due to higher silhouette values and noticeable differences between the two clusters for 6 of the attributes with the greatest differences on percent_graduate degree or more, genie_index, and population size- with the cluster with higher gini index and lower % graduate degree having higher population and having more deaths.

❖ Business-Related Significance

With the population being the clear distinguishing factor, the Texas health department should focus on maximizing vaccine distribution to local drug stores and healthcare facilities in those areas. However, given the fact that factors like the genie index were identified as an even greater differentiator between clusters, we may also want to consider income inequality not just between different counties, but also within counties. For instance, within the city of Dallas, the Highland Park area is considered quite well-off and even ranks as one of the top 10 wealthiest communities in the entire United States in a Bloomberg survey analysis. On the other hand, South Dallas has districts with extremely high wealth inequality caused by gentrification. Thus, Texas' health department should also target stores located in parts of a county that have less wealth to ensure that communities with less resources to control virus spread and individuals with less income to be able to afford treatment, are able to be less susceptible to getting the virus.

Works Cited

Covid-19: First vaccine given in US as roll-out begins - BBC News

Risk for COVID-19 Infection, Hospitalization, and Death By Age Group | CDC

Texas Counties: Median Age (txcip.org)

https://dallas.culturemap.com/news/city-life/03-15-18-highland-park-top-10-richest-places-in-us/