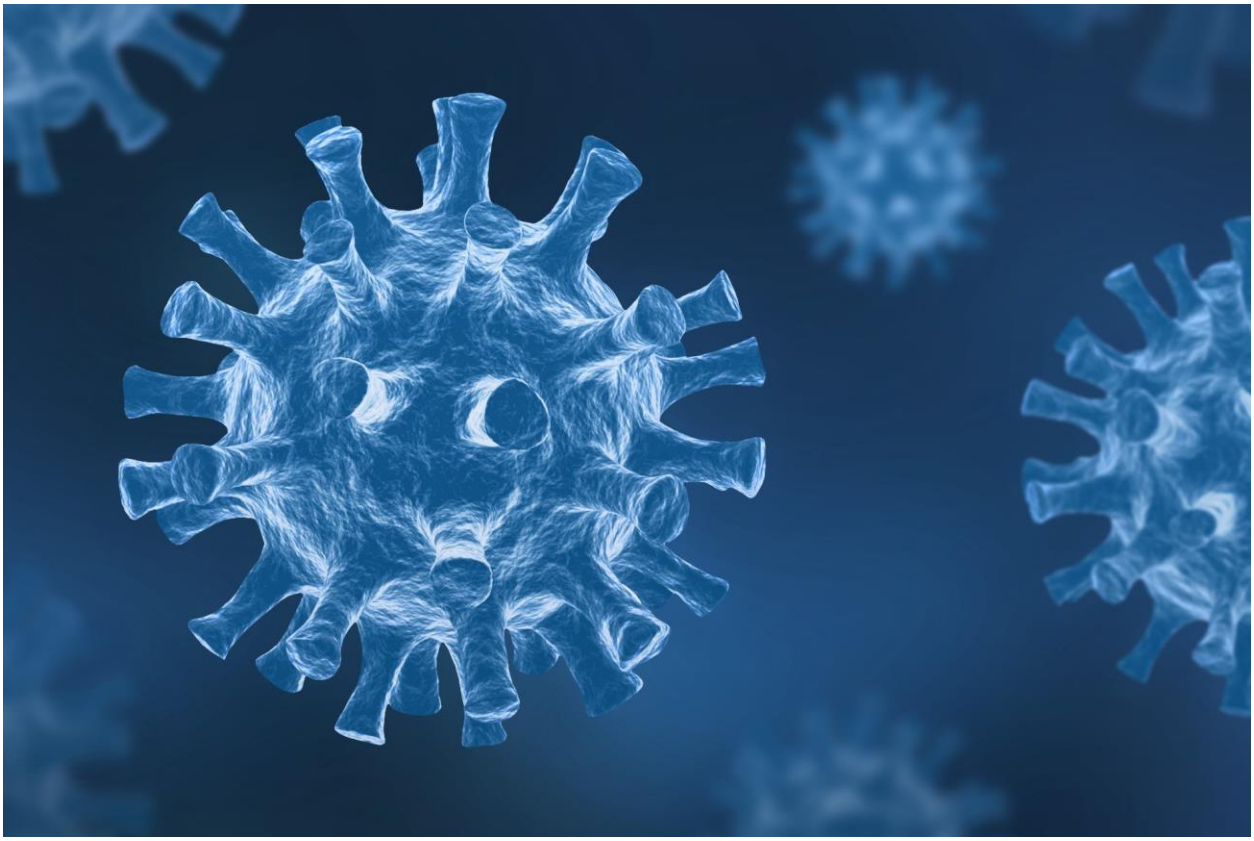
# Clustering COVID-19 Data

*A further look at the data relating to COVID-19 in Texas*



**James Zhai, Ryan Capó, Will Clark**

5 April 2021

EMIS 5331

# EXECUTIVE SUMMARY

COVID-19 has proven extremely difficult to control, despite governments' best efforts to eradicate the virus. In dealing with the virus, it seems that a "one-size-fits-all" method does not work well. Our goal is to find similarities of different counties through clustering to find different groups of counties that have similar demographics and a similar impact by COVID. In doing so, no longer will the Texas government have to implement the same hopeful precautions for everyone. We will specifically target different types of counties with a treatment that should be most effective for them. This allows the government to be more flexible and effective in creating policies that will be most productive in dealing with COVID-19.

In our study, we found that counties closer to big cities geographically show lower deaths and cases on average when compared to counties farther from these big cities. In response to this, healthcare should be made more accessible in areas that are farther from these cities, as well as potentially making vaccines more readily available in more remote counties.

# DATA PREPARATION

## Features Description

**PopDensity:** refers to the number of people per unit of area—calculated by dividing the total population by the area covered by the population. In this case, the data set contains data for population density of each county in Texas. *PopDensity* is on a ratio scale.

**MedianAge:** refers to the *median age* in a county in Texas. In other words, this variable represents the middle value in a range of ages in each county in Texas. *MedianAge* is on a ratio scale in this case because age 0 does not exist.

**LowEdu%:** refers to the percentage of people in a county in Texas that fall under the *lower education* criteria. The criteria for *lower education* is people who have completed education up to elementary school. *LowEdu%* is on a nominal scale. Note that this variable only accounts for adults between ages 45 - 64 in Texas.

**MidEdu%:** refers to the percentage of people in a county in Texas that fall under the *middle education* criteria. The criteria for *middle education* is people who have completed education from 9th - 12th grade. *MidEdu%* is on a nominal scale. Note that this variable only accounts for adults between ages 45 - 64 in Texas.

**HighEdu%:** refers to the percentage of people in a county in Texas that fall under the *higher education* criteria. The criteria for *higher education* is people who have completed education up to at least college. *HighEdu%* is on a nominal scale. Note that this variable only accounts for adults between ages 45 - 64 in Texas.

**MedianIncome:** refers to the *median income* in a county in Texas, or the middle value in a range of incomes of how much income each county has. The median is a more accurate representation for income as opposed to the mean because the median helps to omit outliers. *MedianIncome* is on a ratio scale.

**DeathRate:** refers to the total deaths in one county divided by the total deaths for the total population. *DeathRate* is on a ratio scale.

**CaseRate:** refers to the total cases in one county divided by the cases for the total population. *CaseRate* is on a ratio scale.

**AvgTimeToWork:** refers to the average time it takes for an individual in a county to travel to work. *AvgTimeToWork* is on an interval scale because there is order involved and the difference between two values is meaningful.

**PublicCommute%:** refers to the percentage of people in a county that utilize public transportation to

commute to work. *PublicCommute%* is on a nominal scale.

**LongCommute%:** refers to the percentage of people in a county that have a long commute time to work. The criteria for a *long commute* is people who take at least 45 minutes to get to work. *LongCommute%* is on a nominal scale.

## Features Statistical Summary

| Variable | Scale | Range | Median | Mean | Variance | Std. Dev. | Max | Min |
|---|---|---|---|---|---|---|---|---|
| **PopDensity** | Ratio | 1,119.8 | 8.31 | 43.63 | 16,612 | 128.89 | 1198 | 0.042 |
| **MedianAge** | Ordinal | 31.70 | 38.55 | 39.02 | 35.60 | 5.97 | 57.50 | 25.80 |
| **MedianIncome** | Ratio | 68,851 | 48,311 | 49,894 | 1.47e8 | 12,132 | 93,645 | 24,794 |
| **LowEdu%** | Nominal | .156 | .061 | .060 | .00072 | .026 | .156 | 0 |
| **MidEdu%** | Nominal | .585 | .256 | .271 | .008 | .093 | .585 | 0 |
| **HighEdu%** | Nominal | .661 | .680 | .668 | .008 | .094 | 1.00 | .338 |
| **CaseRate** | Ratio | .169 | .073 | .079 | 0 | .026 | .182 | 0 |
| **DeathRate** | Ratio | .006 | .001 | .018 | 0 | .00098 | .006 | 0 |
| **AvgTimeToWork** | Interval | 10.54 | 8.91 | 9.33 | 5.68 | 2.38 | 15.23 | 4.68 |
| **PublicCommute%** | Nominal | .269 | .121 | .126 | .002 | .044 | .269 | 0 |
| **LongCommute%** | Nominal | .563 | .330 | .335 | .016 | .127 | .625 | .061 |

*Table 1: Statistical Summary for Clustering Features*

The features we chose to use for clustering are listed above in Table 1. In project 1, we analyzed correlations between various levels of education and confirmed COVID-19 cases and deaths in each county in Texas. For this project, we aimed to expand on that. Therefore, we chose additional variables like *PopDensity*, *AvgTimeToWork*, *PublicCommute%*, and *LongCommute%* to identify groups within these variables that we could cluster to give us meaningful data. In doing so, the addition of these variables allowed for us to find other trends that are associated with confirmed cases and deaths. The trends that

our team discovered in these variables are further detailed in the *modeling* and *evaluation* sections.

Table 1 also provides a statistical summary for our chosen features, breaking down the most important stats such as range, median, mean, variance, standard deviation, maximum, and minimum. The table highlights a number of important stats our team deemed significant to discuss. For example, the median case rate is .073 and the median death rate is .001. These two values illustrate how the death rate is very low when compared to the case rate. This is important because our clustering analysis consists of comparing the two rates with other variables like *MedianAge* and *MediumIncome*. Now, the values in the table are not as important as the values clustered on the graph. Rather, the statistical summary table not only allows for us to gain a stronger insight into our features, but also allows for us to visualize which variables we might be able to correlate for clustering.

Another example is looking at the mean values of *LowEdu%*, *MidEdu%*, and *HighEdu%* in Table 1. Looking at these values, we can see that *HighEdu%* has the largest percentage. From this we are able to infer that the mean of higher education for all counties in Texas is 66.8%. With this information, we are able to explore potential correlations. For example, it is intriguing to look at how these counties of higher education are further affected (referring to case and death rate) by variables like public and long commute percentages. This methodology can also be applied to counties with lower and middle education backgrounds, which can also be cross correlated with population density. The statistical summary table essentially aids in creating ideas for potential correlations. This is the process our team followed to further investigate our clustering analysis on COVID-19 data in Texas.

# MODELING

## Section 1: Civilians' Age, Income, and Educational Background in each county

To find the relationship between people's ages, income and educational backgrounds in each county, the report uses the median age, median income, and the distribution of educational diplomas for people between 45-64 in Texas. The distribution of educational diplomas has splitted into three parts: Low Educational Background (<9 grade), Middle Educational Background (9 grade to high school diploma), and High Educational Background (college degree). To find the Population density, the program will use the countys' total population divide by their area in square kilometers.

After building those datasets, the program will put those data into one data frame, with the county's population density and all the data related to COVID-19. The plot of correlation has shown some sort of relation between data in the frame, and it may indicate some effect of variables on the similarity of counties in clustering.

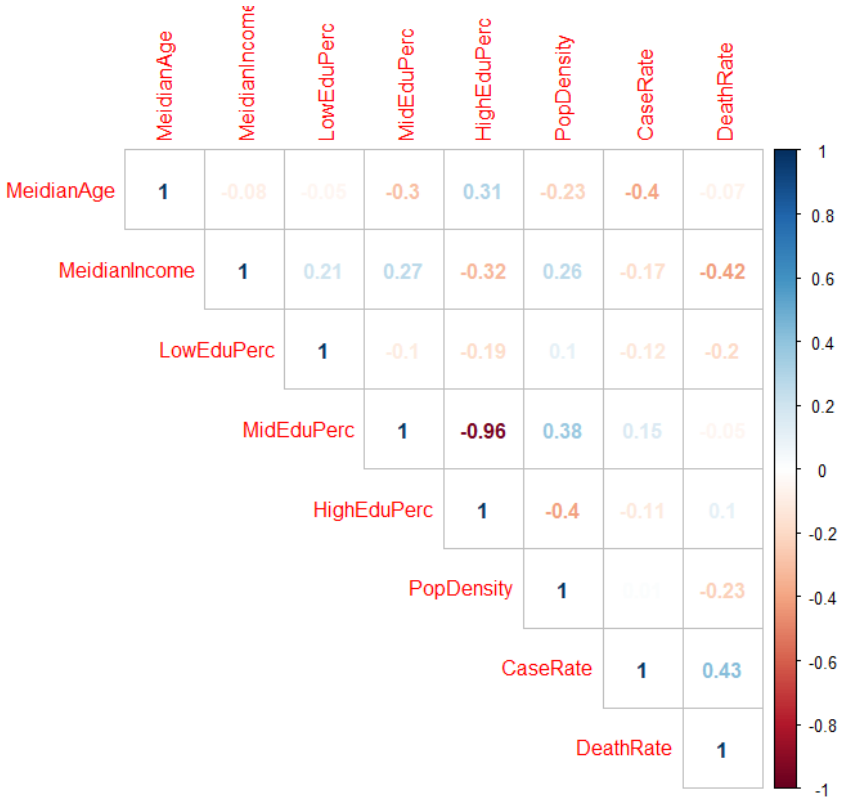



*Figure 3.1: Correlation for Income-Education data frame*

Then, the K-means method clusters the data. Since there are multiple elements in the data frame that will be evaluated together, the report uses the clustering profile to represent the outcome.

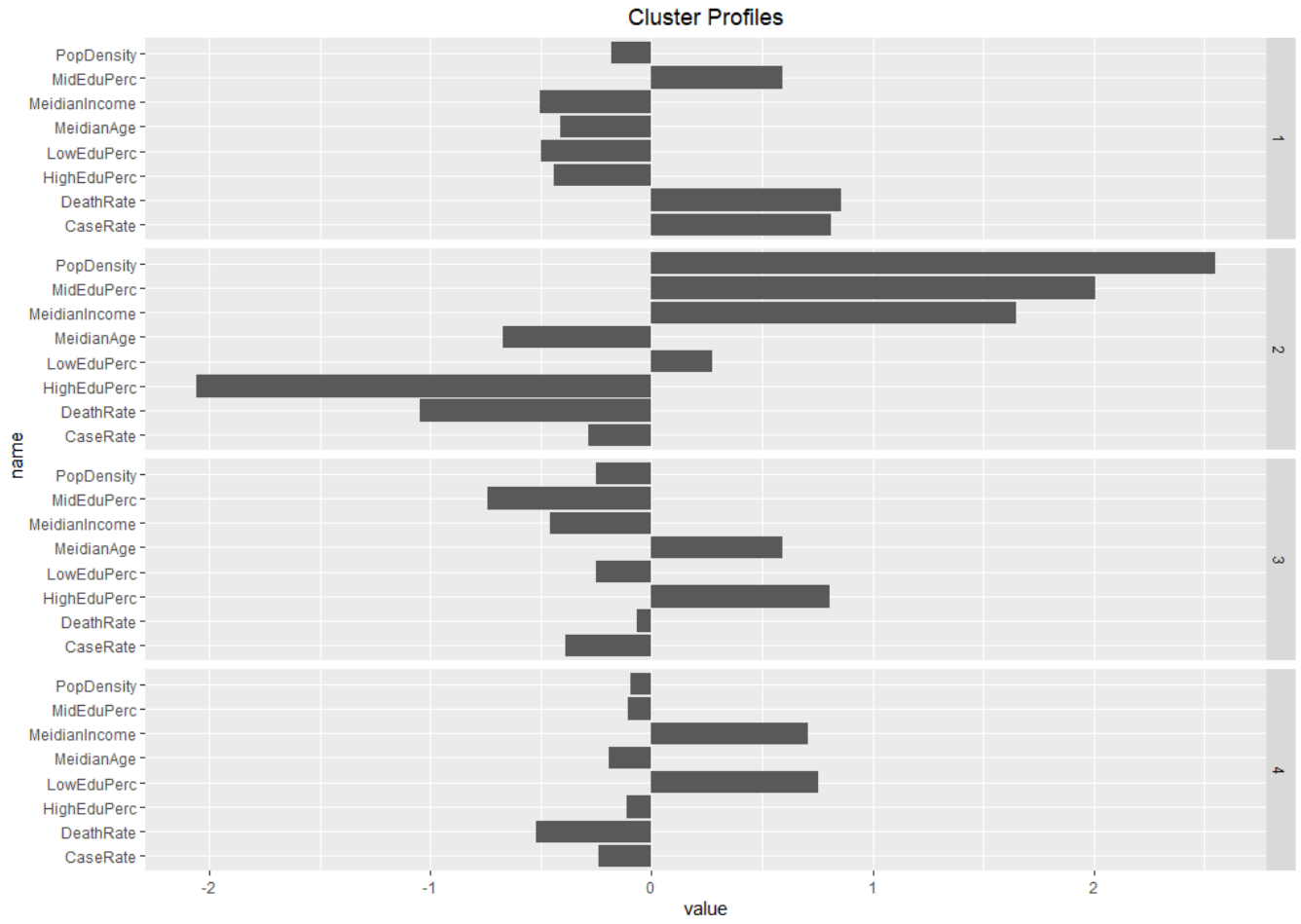



*Figure 3.2: Profiling different clusters*

The number of clusters could be 3 or 4; a smaller number will not visualize each cluster's difference. On the other hand, a higher number of clusters will make the characters of each cluster random and arbitrary, which will also be hard to determine the difference.

For further analysis and visualization, the program will find the mean of the percentage of confirmed cases and the mean of the percentage of death cases for each cluster; it will group each cluster and find the mean of each COVID-19 result.

| Cluster | Percentage of confirmed cases | Percentage of death cases |
|---|---|---|

| | | |
|---|---|---|
| **1** | 0.0998 | 0.00275 |
| **2** | 0.0694 | 0.000866 |
| **3** | 0.0675 | 0.00173 |
| **4** | 0.0713 | 0.00135 |

Furthermore, to let the cluster's characteristics correspond with each county, the program puts all the clusters into the map of Texas. It uses the Texas map from the "ggmap" and "maps" libraries. After inserting each county's name according to the index, the data frame links the map by using the county's name. In the map, the county which does not have a cluster may be caused by a failure in map link.

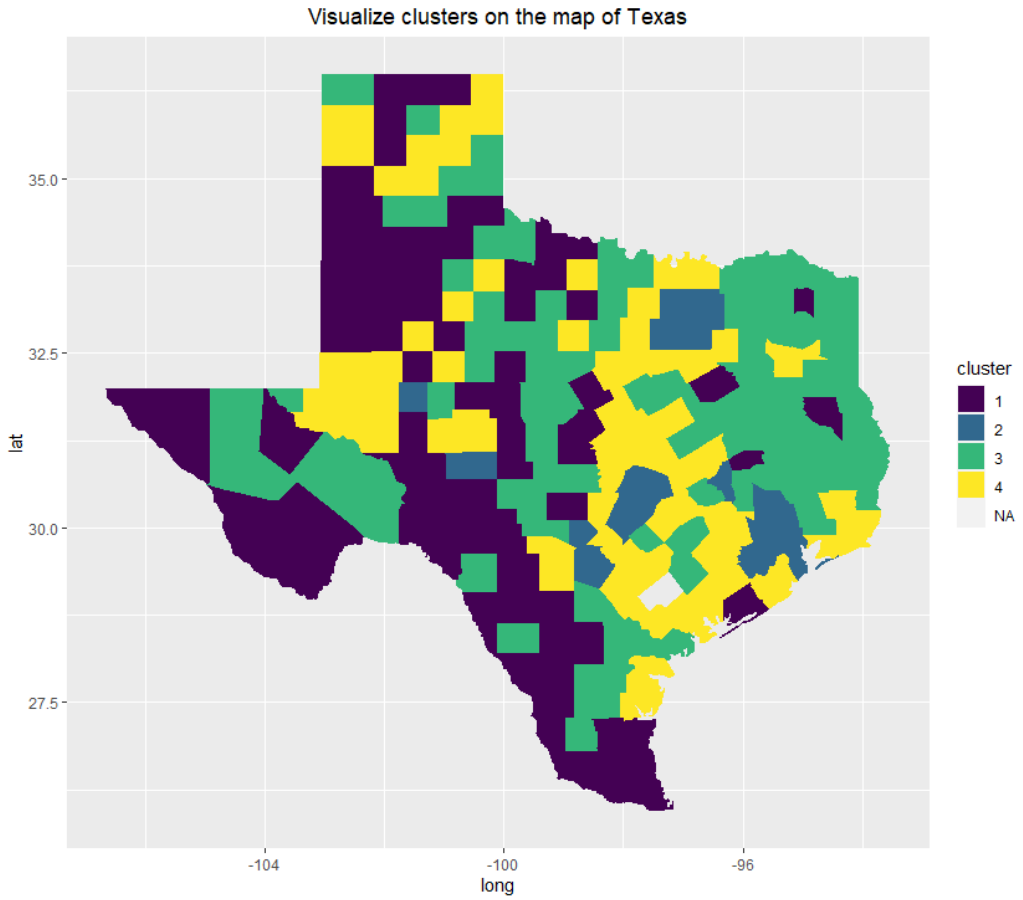



*Figure 3.3: Geographical map of clustered counties*

## Section 2: Commute and Public Transportation

The second section focuses on people's commute time and the percentage of people who choose public transportations. The data frame is called "CommInc", and it has the countys' basic information such as name, area, and population density. Besides, it finds the percentage of long commutes. It uses the number of people whose commute time from 5 minutes to more than 90 minutes. In this report, the term *long commute* is defined as a commute that takes more than 45 minutes. Therefore, the percentage of long commutes is the percentage of people who commute greater than 45 minutes every day. Besides, the average time to work uses the aggregate time to work divided by the total population since the aggregate time to work is the total time for people in Texas to commute every day. Finally, another vector called "Public_Commute_Percent" is the percentage of people who take the subway, bus, and poolcar.

In the first part of this section, the program evaluates the relationship between people's median income and their average time travel to work. After building the plot graph, the program uses the Hierarchical Clustering (average method) to find the clusters. In this graph, the clustering method uses three clusters because it represents three different areas: one with low median income and a short time to work, one with medium median income and various travel time to work, and the last one with high medium income along with high commute time. These three parts may refer to different counties' characteristics from metropolitan to countryside.

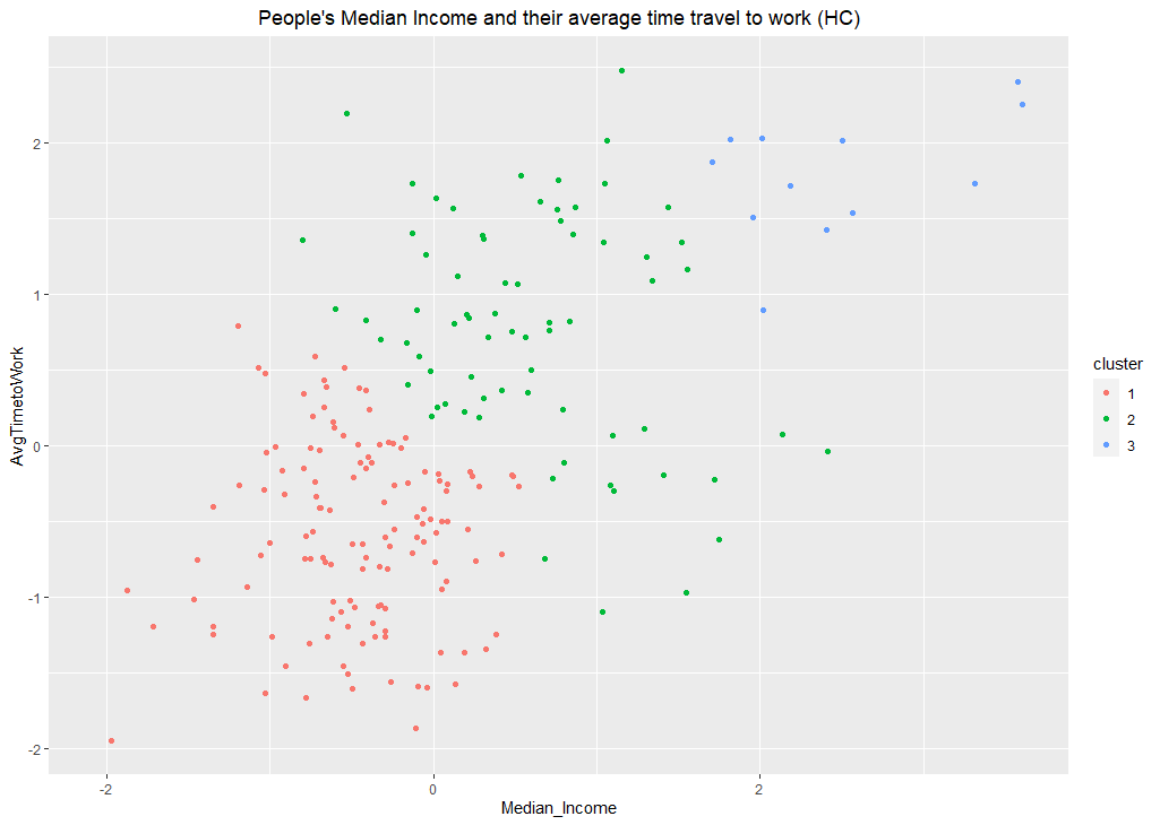


*Figure 3.4: Clusters of median income and average commute to work (Hierarchical clustering)*

Another way to evaluate the connection between median income and the average time travel time to work is using Partitioning Around Medoids (PAM). For the same scaled data set, the program uses *PAM* with three cluster points. Compared to the Hierarchy Clustering, *PAM* vertically splits the plot into three parts; the red part represents the group with few median incomes; the blue part refers to the groups with median income around the mean, and the green part shows the county whose people have high median income. The plot of *PAM* indicates three different situations compared to those in the Hierarchy Clustering.

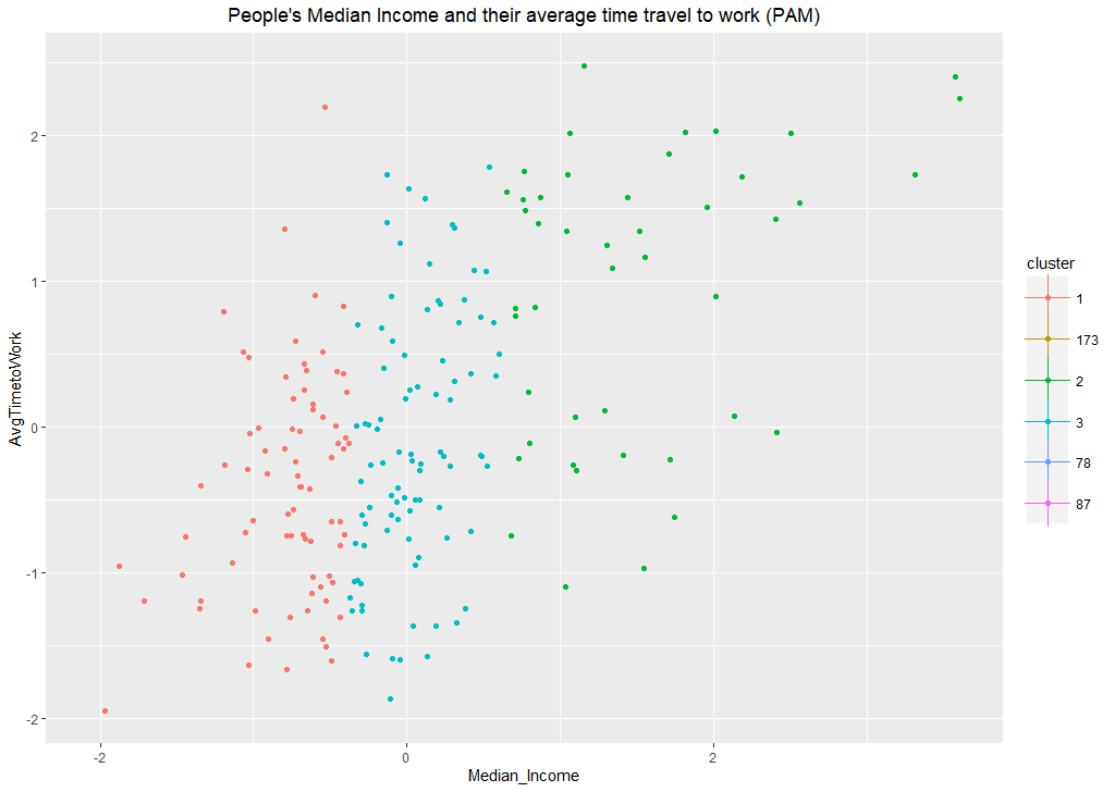



*Figure 3.5: Clusters of median income and average commute to work (Partitioning around medoids)*

In the second part of this section, the program seeks how the county's population density corresponds to the percentage of public transportation people. Nevertheless, in the plot of population density, some counties may have very high population density like Dallas and Harris, which are shown at the right part of the graph. Due to those extreme situations, it would be hard to evaluate the graph because most of the counties will be pushed to the left of the graph, which formed a vertical straight line. Thus, to represent the connection in a clear format, the program only evaluates the counties with

population densities smaller than 30.


*Figure 3.6: Un-cleaned population density vs. percentage of public transportation usage*

After cleaning the extreme conditions, the program finds clusters by using both Hierarchy Clustering and Partitioning Around Medoids. In Hierarchy Clustering, all the counties separate into three parts because its shape looks like a *T* laying on the floor. When the population density is low, counties reflect two different situations in the percentage of public transportation. On the other hand, when the population density becomes high, the percentage gets close to the mean. Under the Partitioning Around Medoids clustering method, the result has reflected the same outcome as the one from Hierarchy Clustering. Therefore, it would be reasonable to argue that a county's population density may have three different relations with the percentage of people who take public transportation, depending on its circumstance.

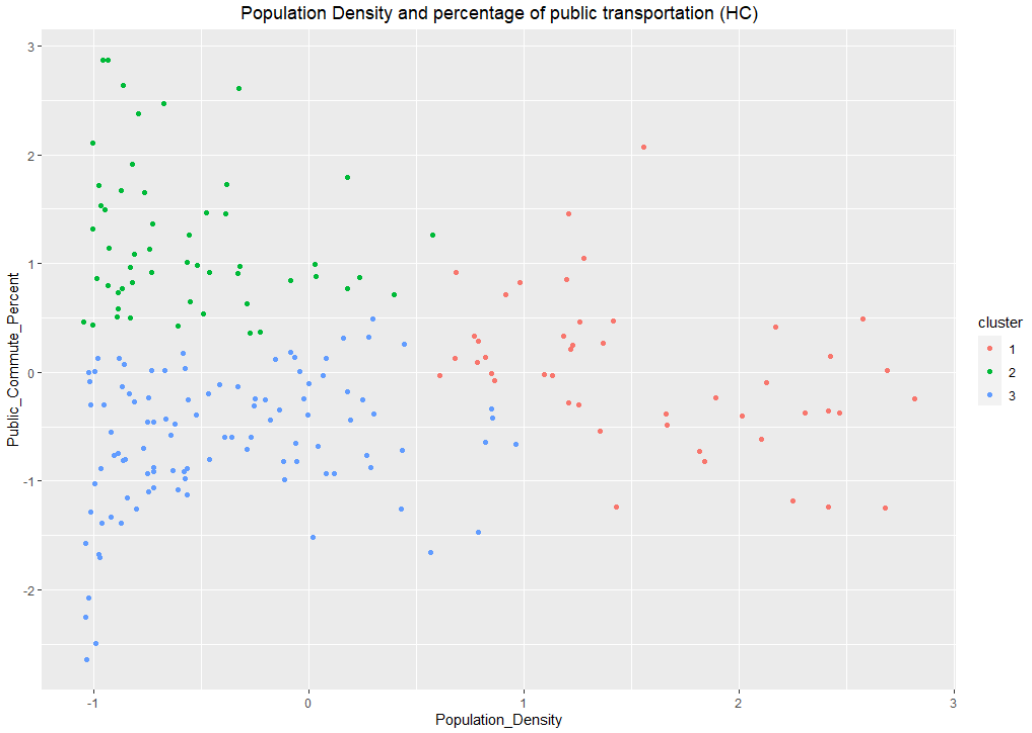



*Figure 3.7: Population density and public transportation usage (HC)*

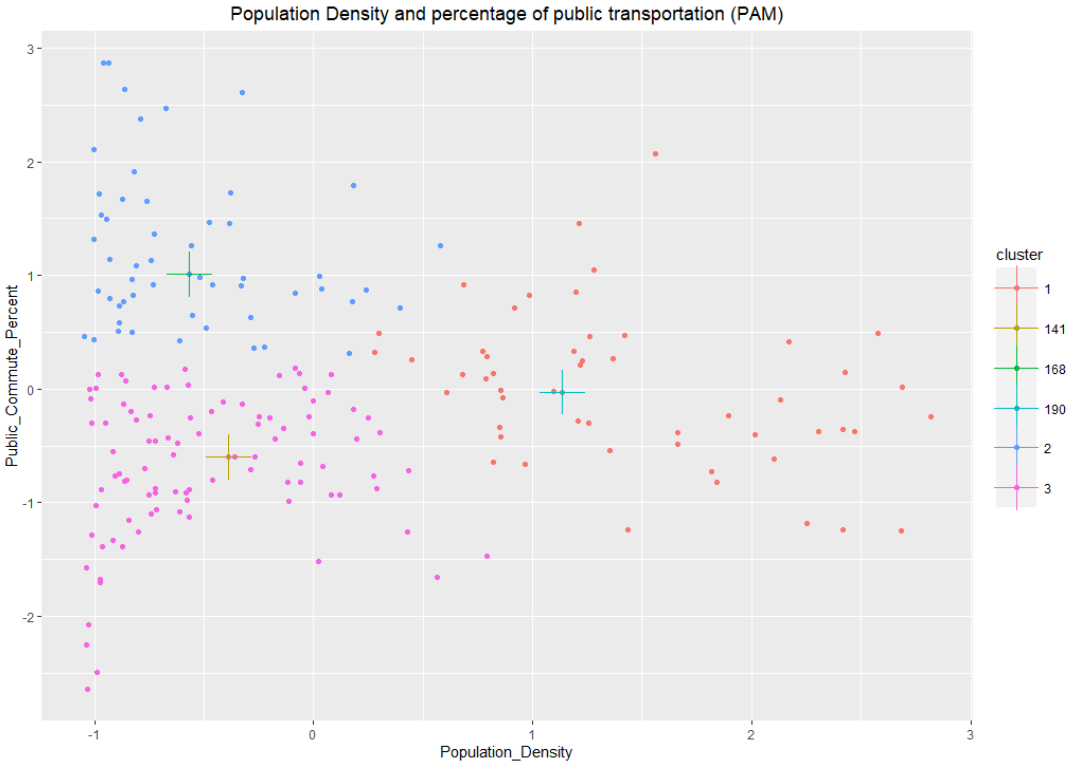



*Figure 3.8: Population density and public transportation usage (PAM)*

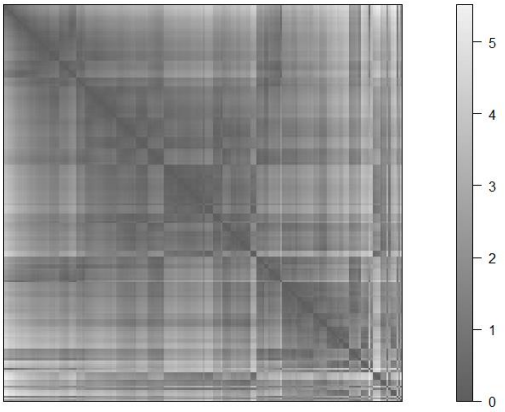



*Figure 3.9: Similarity matrix for the clustering of population density/public transportation*

The clustering analysis Visual Analysis for Cluster Tendency Assessment (VAT) reflects a certain amount of tendency as a block of structure. However, the block is not shown clearly in the graph, indicating that the similarity in the clustering groups is not evident.

The third part will find the population density and how it corresponds to the percentage of people who take long commute time (> 45 minutes).

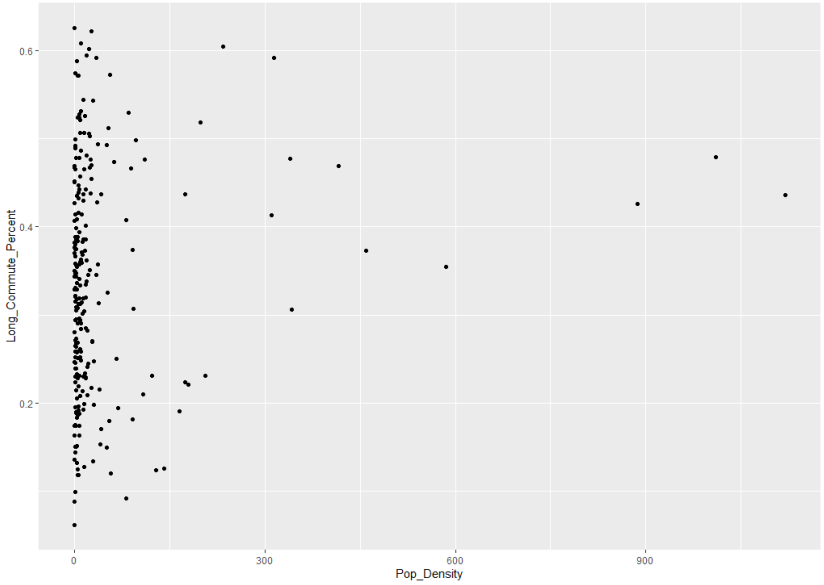



*Figure 3.10: Percentage of people with >45 minute commute vs population density (Uncleaned)*

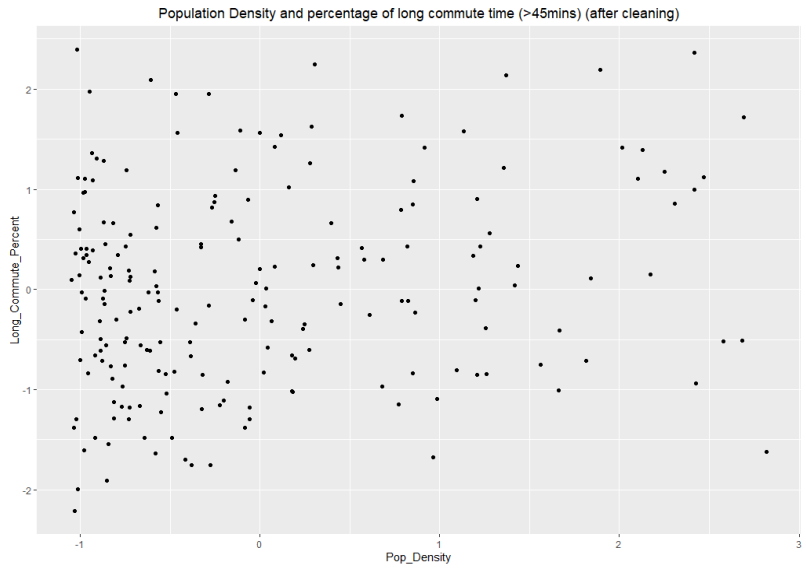



*Figure 3.11: Percentage of people with >45 minute commute vs population density (Cleaned)*

After removing the counties with extreme high population density, it finds clusters using Hierarchy Clustering and Partitioning Around Medoids. Although in this part, it is difficult to find centers and determine how many clusters it will have since it is not apparent. However, it is clear that there is a cluster at the bottom left of the plot and a cluster at the mid-left. Therefore, those points may reflect some similarity, and the number of clusters will be set to three.

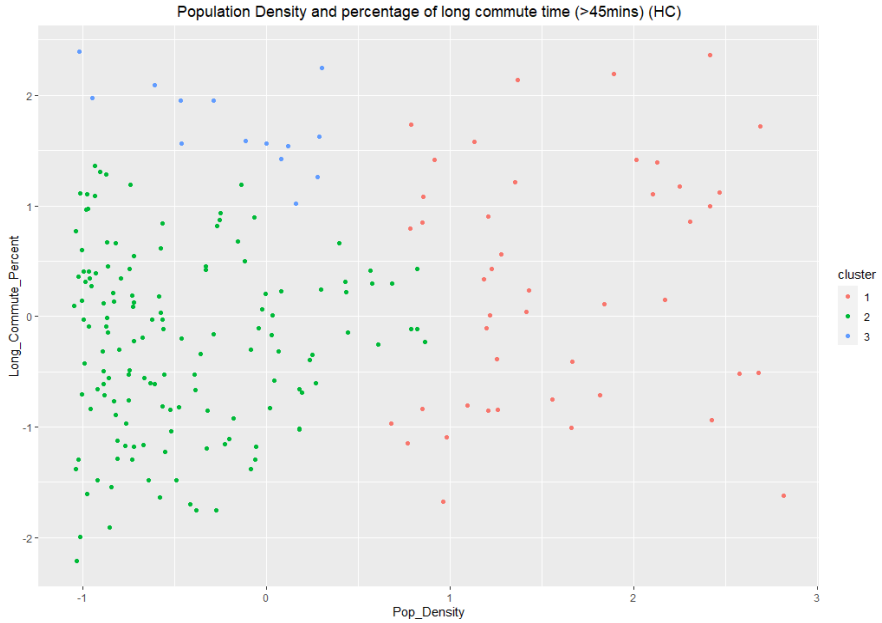



*Figure 3.12: Percentage of people with >45 minute commute vs population density (HC)*

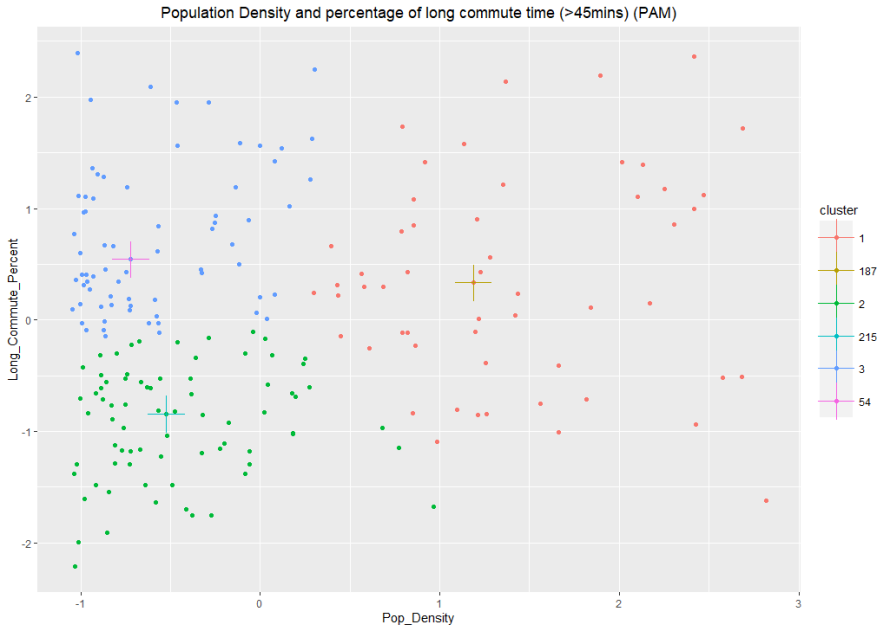



*Figure 3.13: Percentage of people with >45 minute commute vs population density (PAM)*

When the report validates the clusters, the *VAT* result indicates that the similarity between counties on the population density and commute time may be weak; there may be few connections between these elements.

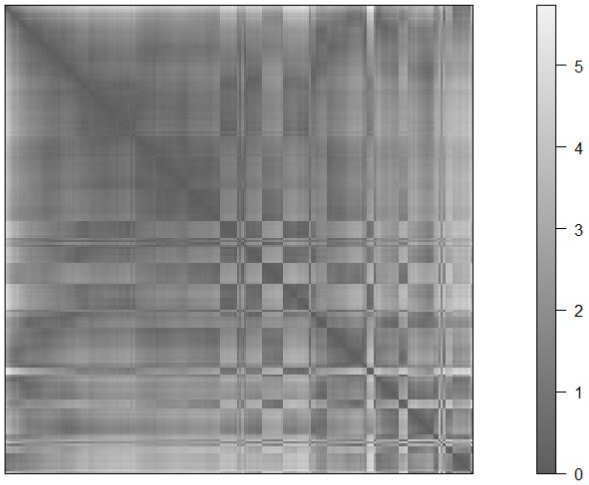


*Figure 3.14: Similarity matrix for Percentage of people with >45 minute commute vs population density*

## Section 3: Median age and median income

In this section, the report seeks the similarity between counties on civilians' median age and

median income. The clusters are explored by using Hierarchy Clustering, Partitioning Around Medoids, and Gaussian Mixture Models.

For the Hierarchy Clustering, it uses three clusters with complete methods because it would be easy to identify at least two clusters around the middle of the plot, and with some outlines around the major cluster, the number of clusters could be 1-3.


*Figure 3.15: Clusters of median age and median income (HC)*

The Partitioning Around Medoid method has a similar outcome compared to the one from Hierarchy Clustering.

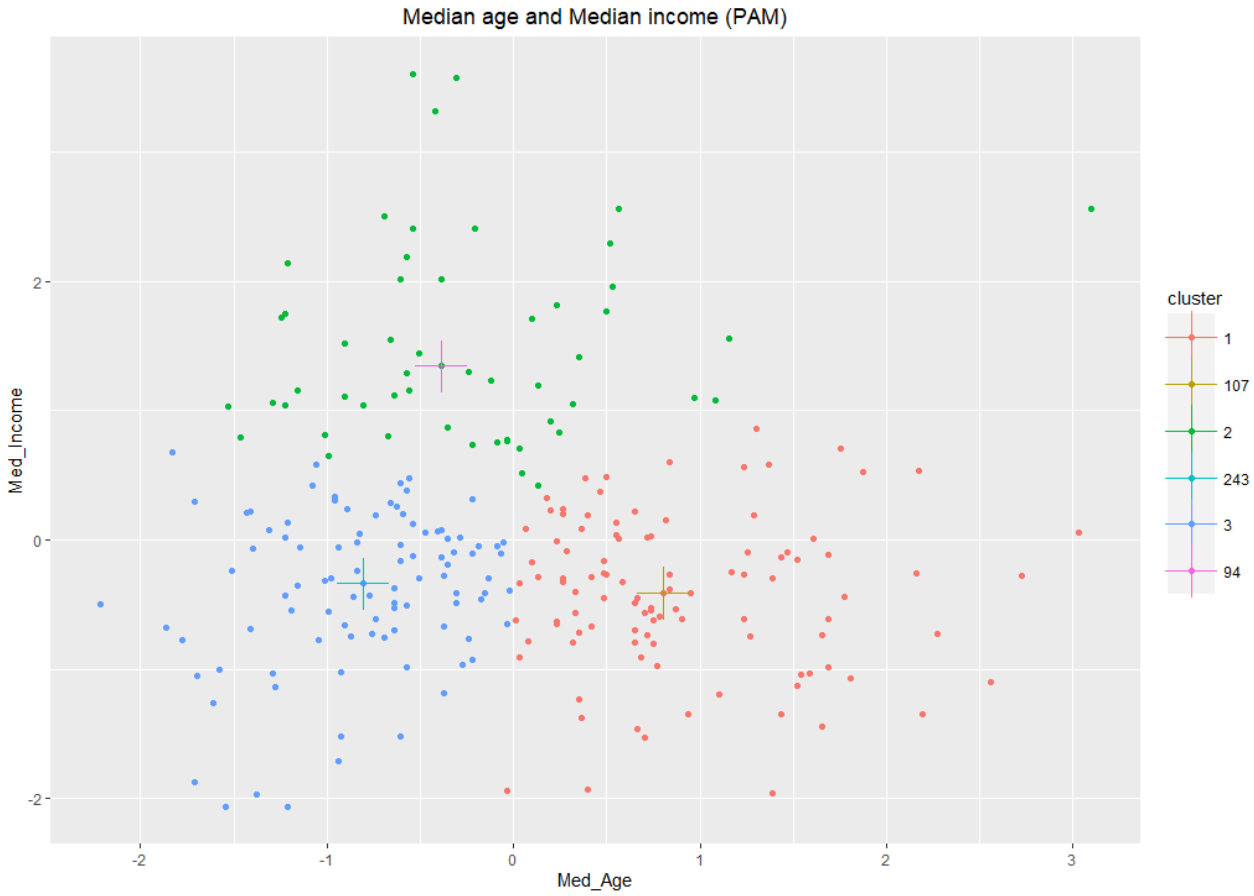



*Figure 3.16: Clusters of median age and median income (PAM)*

Finally, the program finds the clusters by using the Gaussian Mixture Models (Figure 3.17). The Gaussian Mixture Model utilizes the Bayesian Information Criterion to find the number of clusters so that *k* will be settled automatically. In this plot, the number of clusters becomes 2. Moreover, instead of setting three clusters in the middle of crowds, the Gaussian Mixture Model sets the clusters at the center with two flat ellipses; it makes the green group the outlier of the blue group.

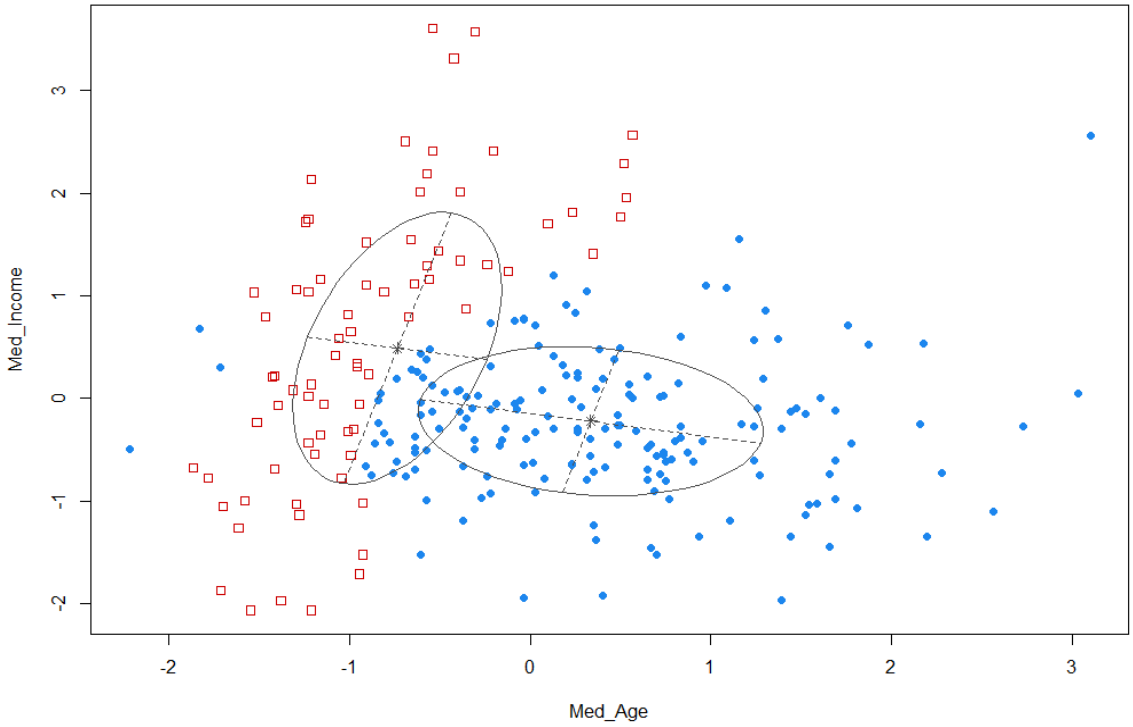



*Figure 3.17: Gaussian mixture model of median income and median age*

The validation of clustering in this section has shown an apparent distance matrix in the plot below. That is because most of the counties have similar median age and median income data, so a smaller distance between the counties and the cluster center may lead to a good clustering result.

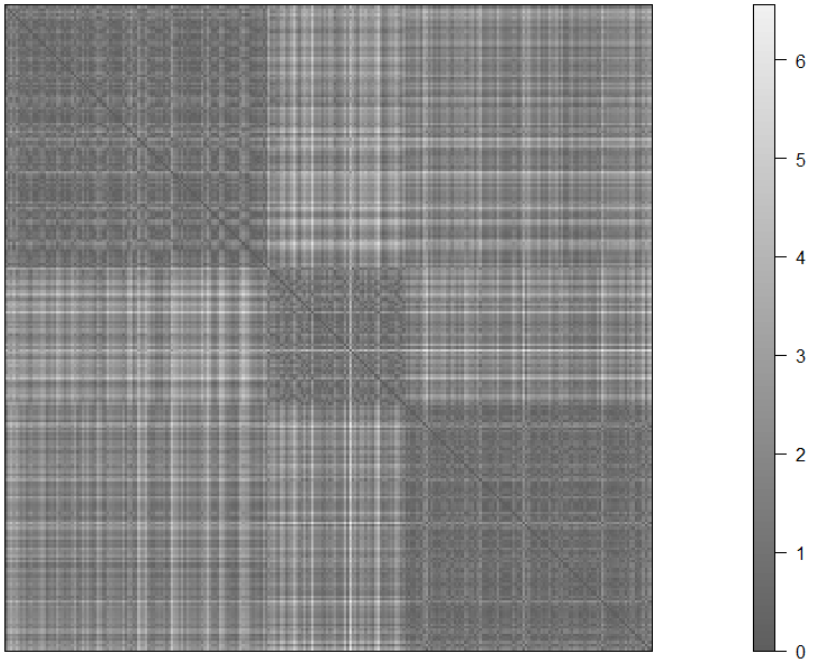



*Figure 3.18: Similarity matrix for clusters of median age and median income*

## Section 4: County's population density and its COVID-19 situation

Could each county's population density reflect some similarity toward the spread of COVID-19? This section identifies the connection between population density and the average confirmed cases and death cases of each county. The similarity in the result of the spread of COVID-19 could give people some idea about how population density becomes a key.

In the first part, the program builds the plot with the county's population density and its percentage of infected people. The percentage refers to the number, which is the total number of confirmed cases divided by the total population.

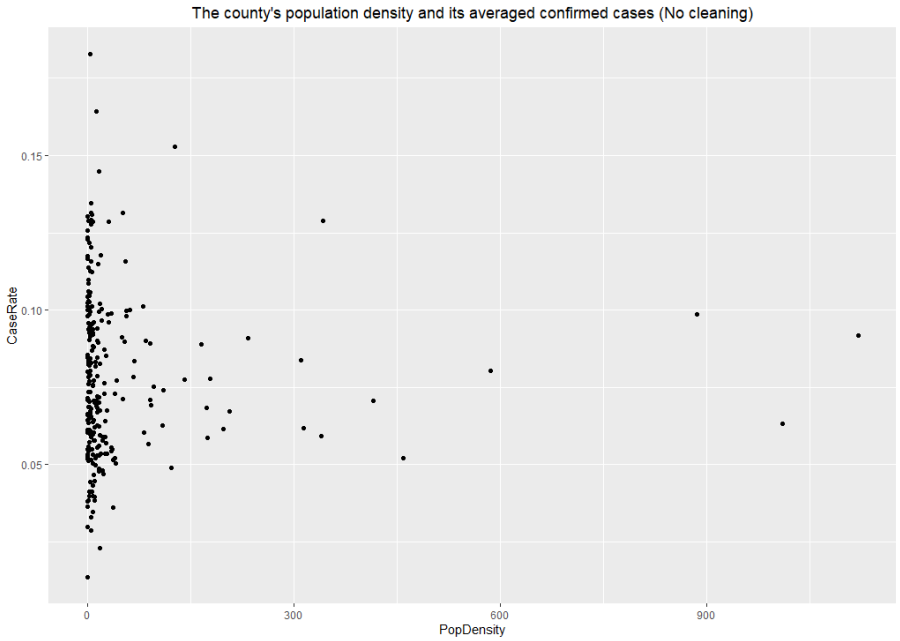



*Figure 3.19: Confirmed cases vs. population density (Un-cleaned)*

Since some counties hold extremely high population density like Dallas and Harris, it would be difficult to do further analysis. Therefore, the report discards those counties and focuses on the rest. Therefore, the program discards counties with a population density higher than 50 from the plot.

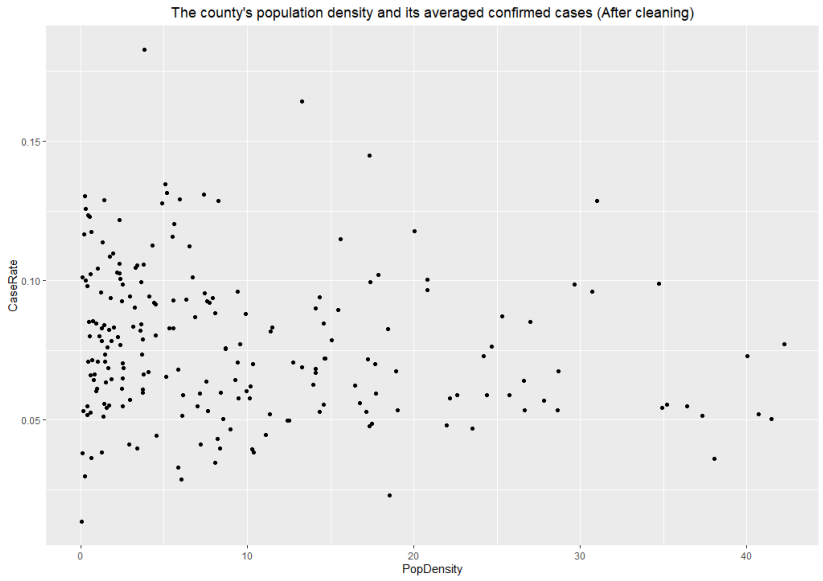



*Figure 3.20: Confirmed cases vs. population density (Cleaned)*

The analysis starts with a K-means clustering. It chooses two clusters because there is only one

obvious cluster on the left side of the plot. However, many points in the plot, which represent Texas counties, are randomly distributed on the plot's right-hand side. It would also be challenging to determine the correct number of clusters, in which that number could be 1 or 2.

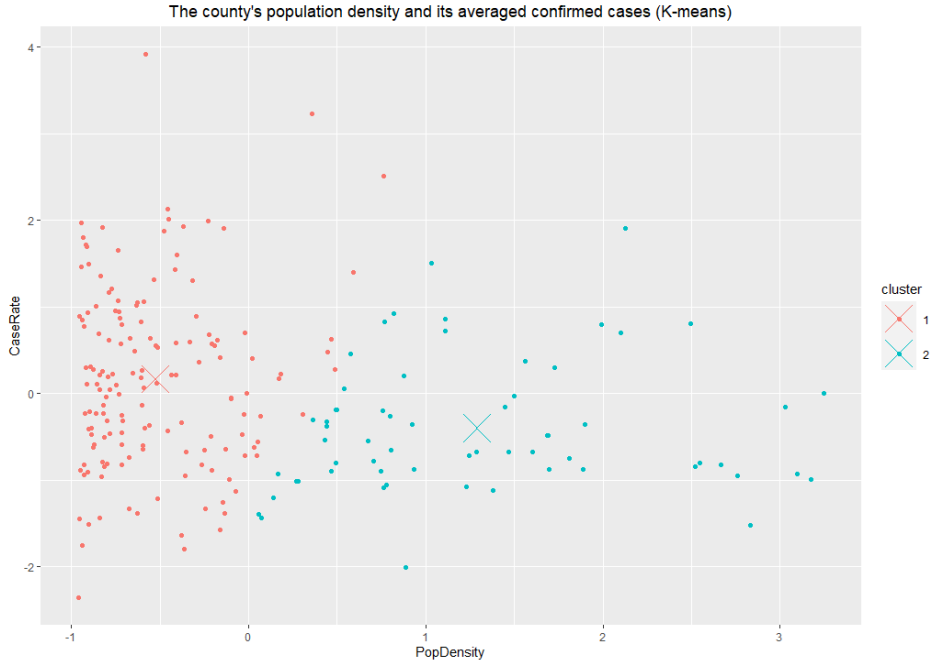



*Figure 3.21: Average confirmed cases vs. Population density clusters (K-means)*

The Partitioning Around Medoid clustering method also shows a similar result. In the plot, all the plots have formed a triangle-shaped area: when the population density is low, the normalized case rate could vary from -2 to 2. On the other hand, when the population density increases, the confirmed case gets closer to the mean.

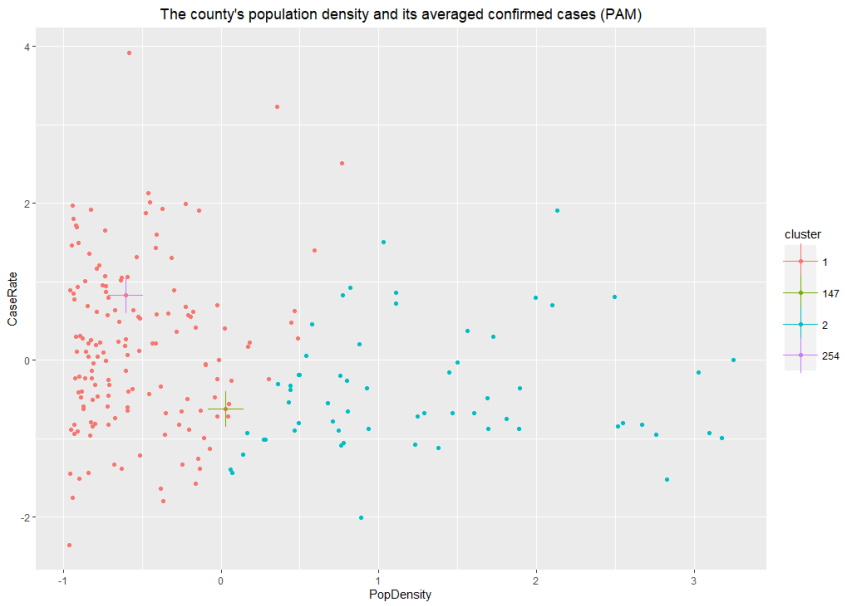



*Figure 3.22: Population density vs. average confirmed cases (PAM)*

When the program uses the similarity matrix to validate the clustering, the outcome shows that the randomly distributed points in the plot did not form a clear cluster, especially when the case rate is high; it would be hard to infer its county's corresponding population density. Therefore, Texas counties may not have a similarity of characteristics on population density and confirmed case rate.

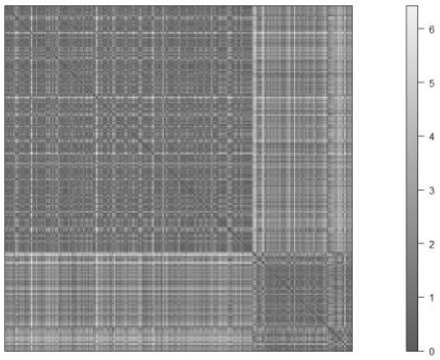


*Figure 3.23: Similarity matrix for clusters population density vs. average confirmed cases*

The second part uses the same method to find counties' similarities in population density and its death rate under the pandemic. Although neither the confirmed case rate nor the death rate is data related to each county's demographic characteristics, they could help the analysis on comparing and contrasting the counties' medical system and facilities in the hospitals. Moreover, it may also reflect how people in different counties react toward the pandemic.

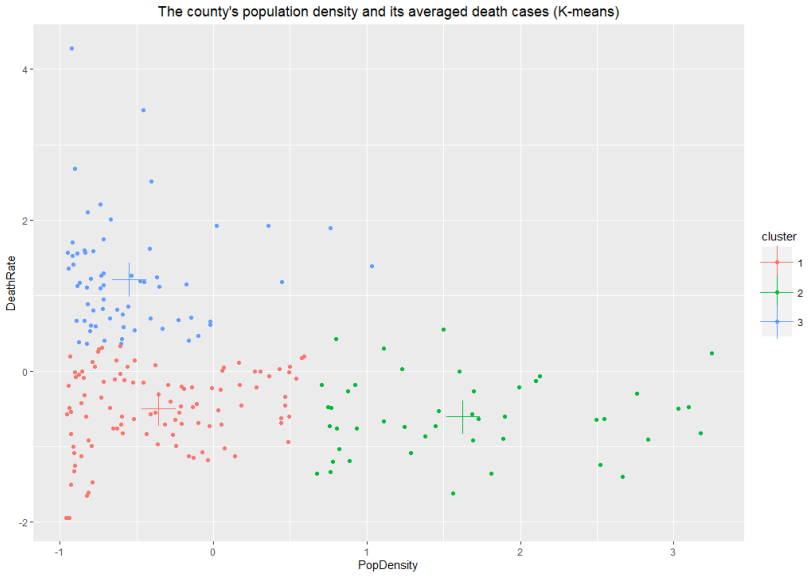



*Figure 3.24: Population Density and average number of deaths (K-means)*

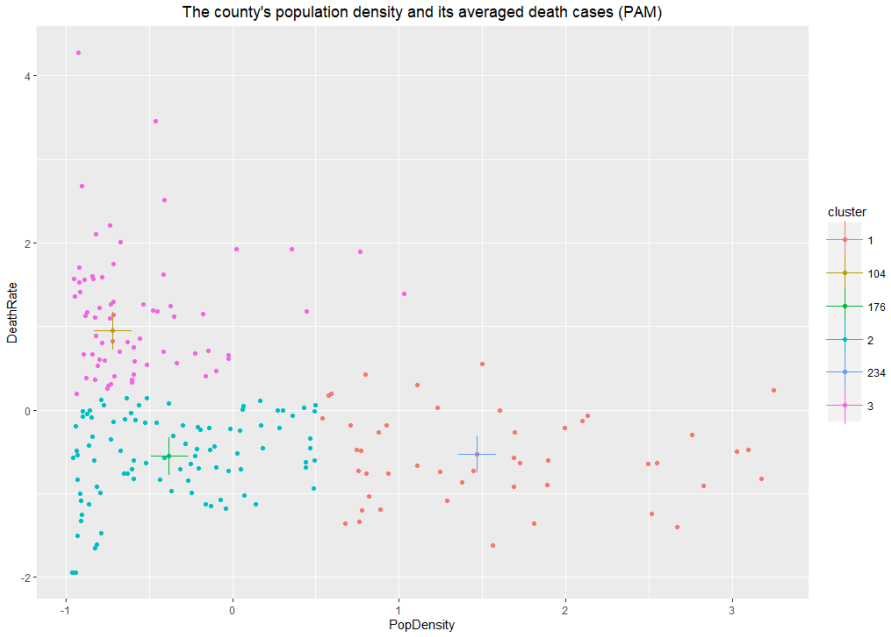



*Figure 3.25: Population Density and average number of deaths (PAM)*

Unlike the plots in the first part, the connection between the county's population density and the death rate has formed a relatively better clustering. In both K-means and PAM clustering methods, they use three cluster points because all the plot points are in a triangle shape. There is a group at the bottom left of the plot and two around it. The program could also use two clusters regarding the points around the left-hand side as a whole. Thus, the number of clusters could be 2 to 3.

The validation result from the similarity matrix also indicates a better clustering compared to

the one from part, which refers to  counties having similar population density. Their reaction toward people with severe COVID-19 symptoms may also be similar. This result can point out how different types of counties work on hospitalization, medical intensive care, or other acts toward treating patients.

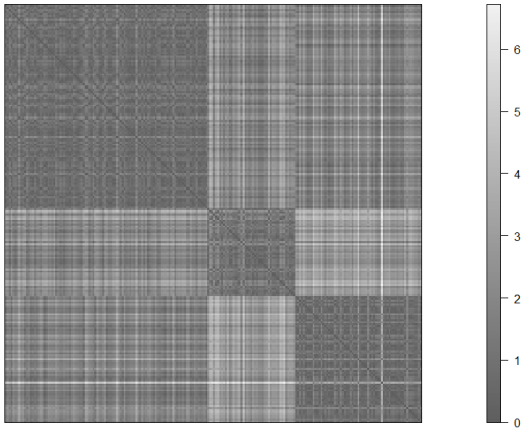



*Figure 3.26: Similarity matrix for Population Density and average number of deaths*

## EVALUATION

Looking at all of the clusters provided some interesting new relationships between unexpected places. For example, Figure 3.12 portrays the relationship between population density and long commute times. Large cities have high population densities and thus more traffic. This is expected. However, the data shows that cities with smaller population densities (typically smaller towns farther from cities) also have some high commute times. We infer that this is likely from having a longer distance to go to work. This could show that although many people choose not to live in a large city, their job requires them to go into an office in a city. This could increase the transmission of COVID from larger cities to smaller towns. This is important to note as it means even as COVID numbers might slow for Texas as a whole, those numbers might slow down because of isolation of small towns who are able to limit interaction due to smaller population density. It is also important to note the COVID numbers in larger cities where many people will commute to work. This could mean businesses need to stay in a state of work-from-home longer than initially expected. It could also mean that restaurants and bars should keep mask mandates despite numbers of confirmed cases and deaths not rising. By opening everything up with no masks, smaller towns who might experience lower numbers due to social distancing and limited social interactions might see increases from members of their population visiting larger cities for recreational activities or dining.

One of the most helpful graphs is Figure 3.2 as seen earlier in the report. This shows the four clusters and their relative demographics. Figure 3.3 helps visualize where these clusters actually are and helps us infer and draw conclusions about possible methods of handling the virus.

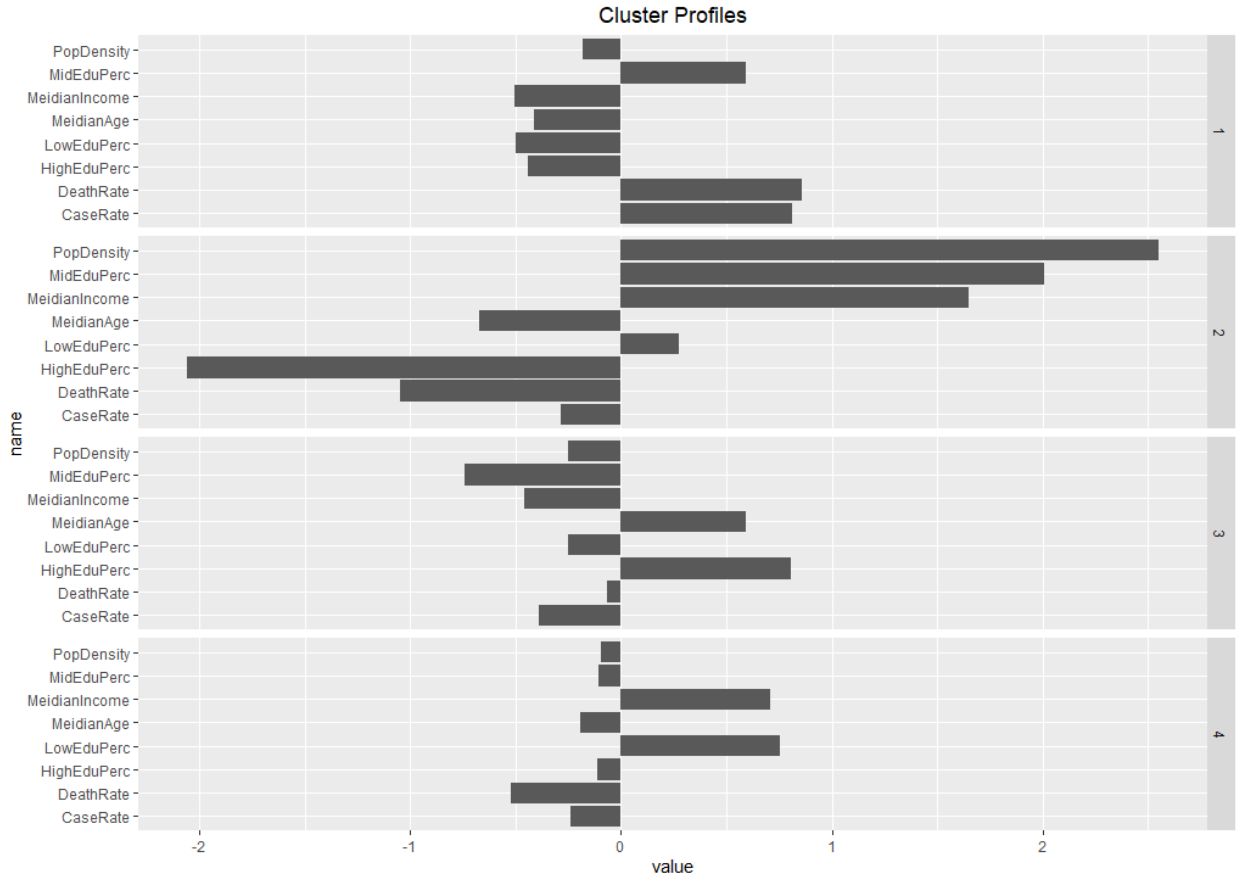



*Figure 4.1: Profiling different clusters*

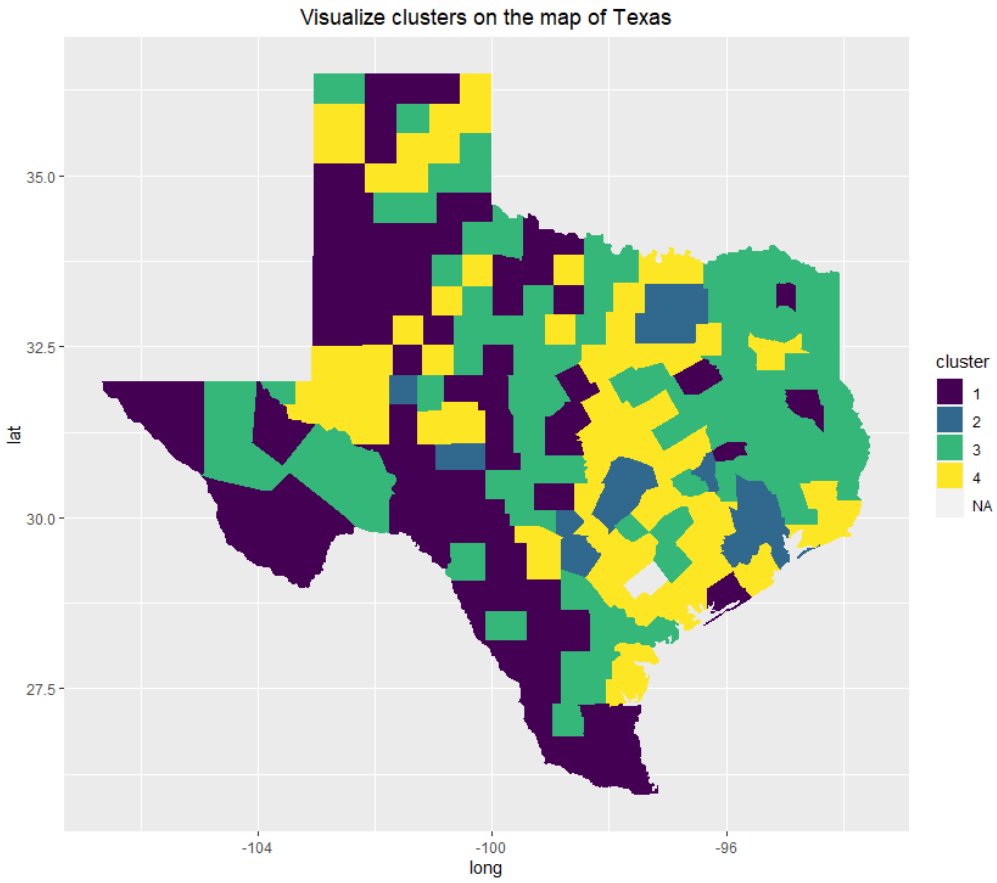



*Figure 4.2: Geographical map of clustered counties*

In Figure 4.1, we see that cluster 1 has a lower population density and low median income. Some of these counties are along the border of Texas and Mexico and in the panhandle of Texas. Some of these counties, like Lamb County, have very small populations and very little healthcare ability. For example, the hospital in Lamb County has a maximum patient capacity of 42 patients [1]. While dealing with a global pandemic where many hospitals are overrun with patients, a county hospital with a maximum capacity of 42 patients is inadequate to deal with the overload of patients requiring attention. This is reflected in the percentage of cases and deaths.

However, in looking at counties in cluster 2, these counties have a much higher population density. Some of these counties include Dallas, Tarrant, Harris, and Travis. Although these counties have large population densities, their case and death rates are low comparatively. However, across Dallas County and Tarrant County, there are over 80 hospitals [2]. This abundance of healthcare facilities means those who need treatment are able to find it easily.

The third group of clusters from Figure 4.1 shows counties with similar population densities to counties from cluster 1, but they have a much larger population with higher education and a higher median age. Most of the values are very similar, except for age, education, cases and deaths. This seems

representative of counties with older populations who are possibly further ahead in their careers and with a higher income. How do we resolve the average number of cases and deaths? It seems that because the population is significantly older than the average population, these people probably know they are more at risk for COVID. This means those populations likely took more precautions individually to limit their exposure to the virus.

The last cluster has more of a median population density, and they are typically found closer to big cities. They have a high median income, likely from commuting to bigger cities for work. They tend to have less education than average, but still have a lower number of cases and deaths on average. This is likely due to their proximity to larger cities with abundant healthcare options.

One challenge that has become evident is the difficulty to use employment demographics to analyze COVID. Due to the number of people who commute from one county to another for work, these numbers become less relevant in how to deal with COVID. What seems to become relevant is the relationship between counties and their healthcare opportunities. Counties with abundant healthcare available seem to experience less deaths and cases on average despite their large populations.

## CONCLUSION

The number of cases and deaths experienced on average in a county seem to be most affected by the number of healthcare facilities available to the county. Counties with higher populations tend to have more healthcare facilities available, and counties with lower populations tend to have less health care options. However, the counties with smaller population densities that fall closer to larger cities geographically also experience lower numbers, likely due to their proximity to abundant healthcare facilities.

Our team would suggest using more resources to make healthcare available to counties in cluster 1 who have smaller populations but much less accessibility to healthcare. This could mean prioritizing sending vaccines to smaller counties who are experiencing higher COVID numbers since they are experiencing more deaths on average.

We would also suggest waiting to see decreases in cases in counties with large cities before opening businesses at full capacity due to the large number of people who commute to work across counties.

## REFERENCES

[1] https://littlefieldtexas.net/203/Lamb-Healthcare-Center

[2] https://www.ariacremation.com/dallas-hospitals/