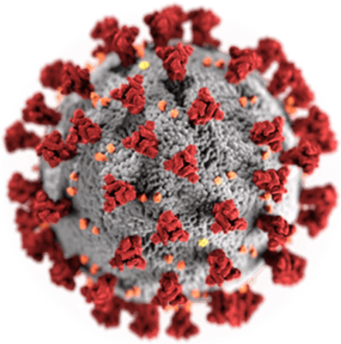


Project 2
Clustering
COVID-19 Dataset

CS 5331 - Data Mining

Authors:

Liam Lowsley-Williams

Emily Fashenpour

Harrison Noble

I. Executive Summary

In early 2020, the United States was introduced to the novel COVID-19 virus which rapidly swept across the nation, infecting and killing thousands of US citizens. States and counties quickly began documenting case numbers and the death toll on a day to day basis to better track the effects of this virus. When infection and death numbers are cross referenced with census data about the certain characteristics of populations, interesting conclusions can be made regarding how the virus spreads in different parts of the country. The focus of this report is to examine COVID-19 and census data from the state of Texas in order to group, or cluster, similar counties and look for patterns. Some general questions we are hoping to answer regarding the found patterns are: what counties have similar demographic and socioeconomic characteristics and how did they respond to the virus? What are the differences between the groupings of counties? Can we identify groupings of counties that are more severely affected by COVID-19 than others? And can we make any recommendations to speed up vaccination in certain counties based on the findings presented in our analysis? This report is written for government healthcare entities in the state of Texas in hopes that they will consider our recommendations and distribute COVID-19 vaccines to the counties affected the worst. The analysis reveals counties with high poverty rates, low education levels, high median age, and low income levels are more severely affected by COVID-19 in terms of fatality rates and deaths per 1000 people. These counties are often found along the Texas/Mexico border and closer towards central Texas. Therefore, we recommend that COVID-19 vaccines should be distributed to counties along the border and to low income counties with high poverty rates and low education levels.

Table of Contents

II. Data Preparation

Features we are Using for Clustering

For this project we modified our dataset from Project 1 to include more census data to provide some more interesting clusters. We used data from the combined datasets “COVID-19 cases TX” and “COVID-19 cases plus census” to generate one single dataset for our clustering needs. From the “COVID-19 cases TX” we extracted the case information for the last recorded day in the dataset, in this case it was 01/25/2021. From the “COVID-19 cases plus census” we extracted population information (total and per race), median income, income per capita, number of commuters by public transportation, education levels, and number of people in poverty. We also only used data for counties in Texas from the “COVID-19 cases plus census” dataset since we are limiting our focus to Texas on this report. **Because Texas has a vast population range at the county level, we decided to convert many features to “per 1000 people” to better normalize the data and give a better understanding of the relative density of an attribute in a county. Without this, the data could skew to either the few very large counties with high populations, or the many very small counties with small populations**. Below is the description of the features for the combined dataset we will be using for analysis.

Table 1: Description of features used for cluster analysis

Feature	DataType	Description
county_fips_code	Nominal	Identification number that uniquely identifies geographical area for county
county_name	Nominal	Name of county in US
cases_per_1000	Ratio	Cumulative number of confirmed cases per 1000 people at 01/25/2021
deaths_per_1000	Ratio	Cumulative number of deaths per 1000 people at 01/25/2021
fatality_rate	Ratio	Deaths per 1000 divided by cases per 1000
commuters_public_1000	Ratio	Number of people that take public transportation per 1000 people
poverty_1000	Ratio	Number of people that live below the poverty line per 1000 people
white_per_1000	Ratio	Number of people identifying as White per 1000 people
black_per_1000	Ratio	Number of people identifying as African American per 1000 people
asian_per_1000	Ratio	Number of people identifying as Asian per 1000 people
hispanic_per_1000	Ratio	Number of people identifying as Hispanic per 1000 people
amerindian_per_1000	Ratio	Number of people identifying as American Indian per 1000 people

other_per_1000	Ratio	Number of people identifying as race other than White, African American, Asian, Hispanic, or American Indian per 1000 people
median_income	Ratio	Median income
income_per_capita	Ratio	Income per-capita
median_age	Ratio	Median age
associates_per_1000	Ratio	Number of people with an associate’s degree per 1000 people
bachelors_per_1000	Ratio	Number of people with a bachelor's degree per 1000 people
high_school_per_1000	Ratio	Number of people with a highschool diploma per 1000 people
ged_per_1000	Ratio	Number of people with a GED per 1000 people

Below, in Table 2, is the statistical summary of the selected features in our chosen dataset. As mentioned above, this dataset contains information from both the census dataset and the cumulative cases as of 01/25/2021 for each country in the state of Texas.

Table 2: Statistical summary of select features in our dataset

Feature	DataType	Mean	Median	St. Dev	Variance	Min	Max	Range
county_fips_code	Nominal	---	---	---	---	---	---	---
county_name	Nominal	---	---	---	---	---	---	---
cases_per_1000	Ratio	81.6	78.59	27.08	733.57	13.51	183.04	169.53
deaths_per_1000	Ratio	1.97	1.79	1.04	1.08	0	6.28	6.28
fatality_rate	Ratio	0.02508	0.02395	0.01335	0.00017	0	0.09167	0.09167
commuters_public_1000	Ratio	1.361	0.63	2.32	5.4	0	17.31	17.31
poverty_1000	Ratio	155.91	155.22	57.33	3,287.3	28.24	383.53	355.29
white_per_1000	Ratio	565.59	595.12	211.5	44,735.18	6.35	923.58	917.23
black_per_1000	Ratio	61.47	36.35	65.99	4,355.55	0	337.43	337.43
asian_per_1000	Ratio	9.79	4.74	19.36	374.89	0	192.21	192.21
hispanic_per_1000	Ratio	345.18	263.78	232.33	53,979.04	34.54	991.85	957.31
amerindian_per_1000	Ratio	3.32	2.25	5.06	25.68	0	54.05	54.05
other_race_per_1000	Ratio	0.83	0.11	2.31	5.33	0	24.8	24.8
median_income	Ratio	49,894.34	48,311	12,132.68	147,201,815	24,794	93,645	68,851

income_per_capita	Ratio	24,859.02	24,284.5	5,240.75	27,465,478	23,543	41,609	29,066
median_age	Ratio	39.02	38.55	5.97	35.59	25.8	57.5	31.7
associates_per_1000	Ratio	42.27	43.29	13.17	173.48	5.00	89.04	84.04
bachelors_per_1000	Ratio	85.63	80.08	33.18	1,101,57	19.5	212.83	193.33
high_school_per_1000	Ratio	171.68	171.21	39.11	1,529.82	80.93	292.36	211.43
ged_per_1000	Ratio	212.72	214.89	46.89	2,199.41	98.36	337.77	239.41

The reason that we chose to use these features for clustering is because we believe that they will give us interesting insights as to how COVID-19 affected each county in regards to their socio-economic characteristics. Using data such as median income, poverty, commuters by public transportation, population by racial group, and other features listed above should help us identify possible cluster relationships and understand how COVID-19 affected the counties within Texas given their population makeup pulled from the census.

Scale of Measurement for the Features

As seen above, all our data except county name and county FIPS code are ratio data types. Additionally, all the features that are related to population sizes (e.g. the racial populations) are measured on a "per 1000 people" scale to normalize the data and properly compare larger counties to smaller counties. Before running cluster analysis on a subset of selected features, we also do Z-score normalization which further "squishes" the scale of the data. Because Z-score can have both negative (below the mean) and positive (above the mean) values, Z-score normalization transforms our ratio data into interval data. **Since our data is on the Z-score interval scale before running cluster analysis, Euclidean distance is an appropriate distance measure as each feature is now using the same scale and different feature values can be compared.**

III. Modeling

Cluster Analysis

Below is our cluster analysis on five different clusters that we decided to look at. The majority of these clusters relate to clustering on demographic information which we then correlate to average fatality rate, cases, and deaths within those clusters.

A. Cluster 1: Poverty per 1000, Public Commuters per 1000, and Median Age

The first cluster that we chose to look at included the features poverty per 1000 people, public transportation commuters per 1000 people, and median age. We chose to look at these features and attempt to cluster them as we figured there would be some relation between them. We figured that counties with high poverty may use more public transportation and wanted to see how this may relate to case counts, death counts, and fatality rates in the clusters. Furthermore, we also added median age to try and see if there may have been any correlation between poverty and public transportation. We performed K-Means clustering and Hierarchical clustering in order to obtain a more holistic picture of the clustering. Our explanation to the number of clusters we chose will be explained later in this report. Below you can see our cluster summaries as depicted in Figures 1.1 and 1.2 followed by the actual counties mapped to the clusters in Figures 1.3 and 1.4 and finally the average case, death, and fatality rates for the clusters in Tables 1.5 and 1.6.

Fig 1.1: K-Means Summary

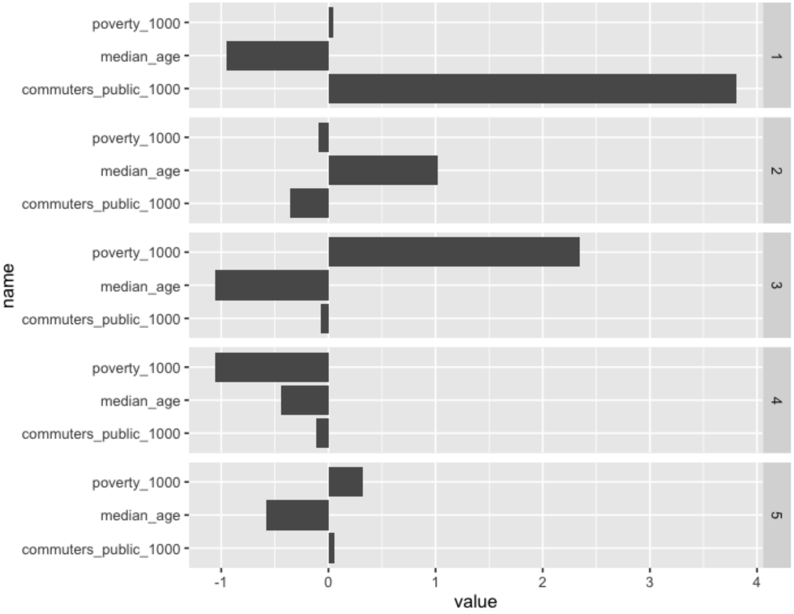


Fig 1.2: Hierarchical Summary

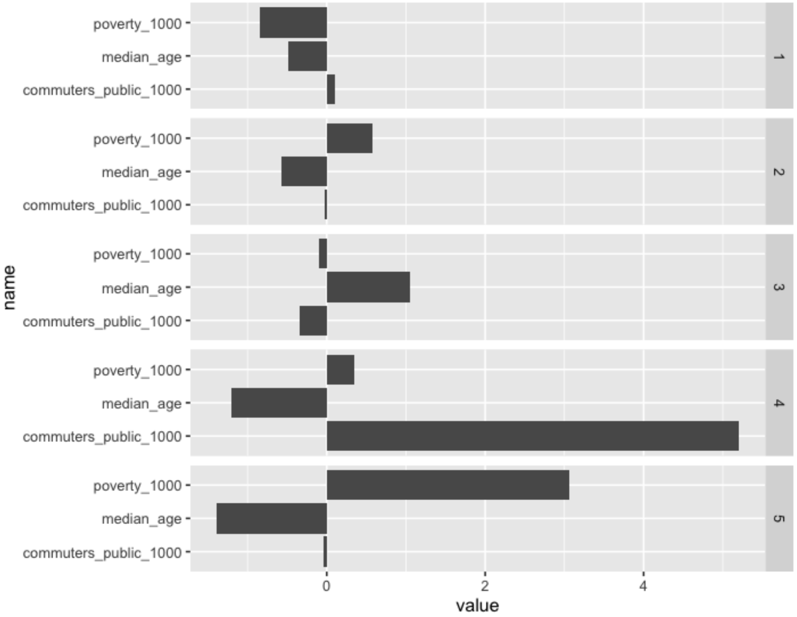


Fig 1.3: K-Means Clusters Map

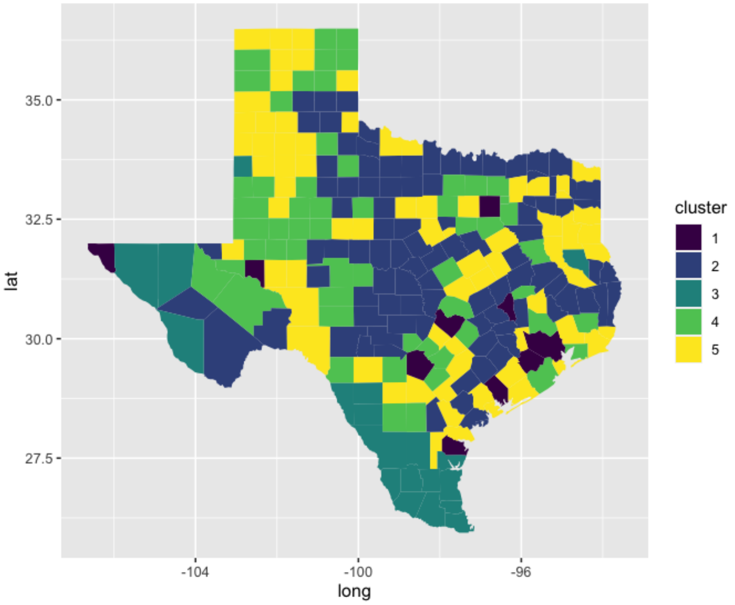


Fig 1.4: Hierarchical Clusters Map

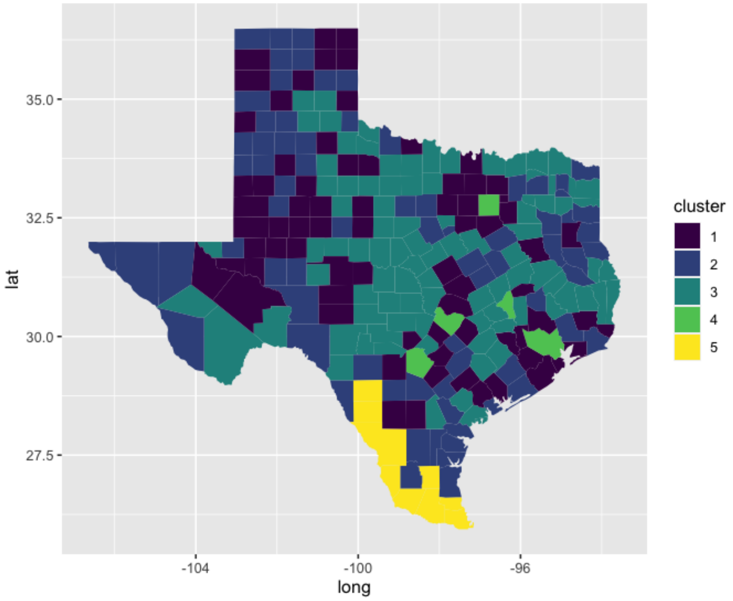


Table 1.5: K-Means Clusters

Cluster	Avg. Cases Per 1000	Avg. Deaths Per 1000	Avg. Fatality Rate
1	87.3	1.29	0.0141
2	69.2	1.96	0.0287
3	98.6	2.72	0.0302
4	83.5	1.63	0.0197
5	91.0	2.16	0.0247

Table 1.6: Hierarchical Clusters

Cluster	Avg. Cases Per 1000	Avg. Deaths Per 1000	Avg. Fatality Rate
1	82.3	1.38	0.0167
2	87.6	2.21	0.0265
3	66.9	1.85	0.0288
4	72.4	0.857	0.0119
5	98.2	2.65	0.0288

Taking a look at K-Means clustering first in Figures 1.1, 1.3, and Table 1.5, we can see that there were some correlations but they were not exactly what we expected. We see that in cluster 1 there was some poverty and high use of public transportation and those clusters resided in large cities. Cluster 1 also had the lowest fatality rate which is interesting to see as it had the highest public transportation use. **We expected the areas with high public transportation use to be more dangerous with higher fatality rates but that was not the case. This could be due to public transportation being limited or shut down during the pandemic.** Regardless, we did not expect to see such an occurrence. Moving on to cluster 3 we can see that this cluster actually had the highest fatality rate of all the other clusters, it was followed closely by cluster 2. What is interesting about cluster 3 is that it had a high concentration of impoverished people. **This seems to suggest that areas with high poverty, which also seem to be located along the Mexico/U.S. border, have higher fatality rates and imply that they are dangerous locations to reside in**

with regards to COVID-19. Cluster 2 on the other hand was our second deadliest cluster with the second highest fatality rate of the clusters. This cluster had the highest median age but lower commuters and poverty levels. The cluster also seemed to reside in more rural areas towards the center of Texas and far from large cities. **This was another interesting correlation because it seems to tell us that the fatality rate was worse in areas where older people resided, this makes sense though as older people are more at risk from contracting the virus and suffering from it fatally.** When we cross referenced these results with those from our hierarchical clustering, we saw a similar correlation which further backed up our claims made above.

B. Cluster 2: Cases/Deaths per 1000 and Fatality Rate

For our second cluster, we chose to continue to look at poverty and cases/deaths/fatality rates but we changed it up a bit this time. This time we decided to cluster on cases per 1000 people, deaths per 1000 people, and fatality rates. We chose to cluster these features since we later looked at the statistical averages of the clusters in regards to average poverty, median income, and income per capita. This allowed us to look at a somewhat flipped version of what we saw in our first cluster seen above. What we wanted to see here was that with the clusters that had higher cases/deaths per 1000 people and higher fatality rates we would see a higher average poverty and lower average median income/income per capita. This would reinforce our understanding that areas of high poverty and low income would be more dangerous with regards to COVID-19 fatality rates and deaths/cases per 1000. Below you can see that we performed both K-Means and Hierarchical clustering in order to cross reference our results to provide a more holistic view on the clustering. Like with cluster 1 above, our explanation to the number of clusters we chose will be explained later in this report. Below you can see our cluster summaries as depicted in Figures 2.1 and 2.2 followed by the actual counties mapped to the clusters in Figures 2.3 and 2.4 and finally the average poverty, median income, and income per capita for the clusters in Tables 2.5 and 2.6.

Fig 2.1: K-Means Summary

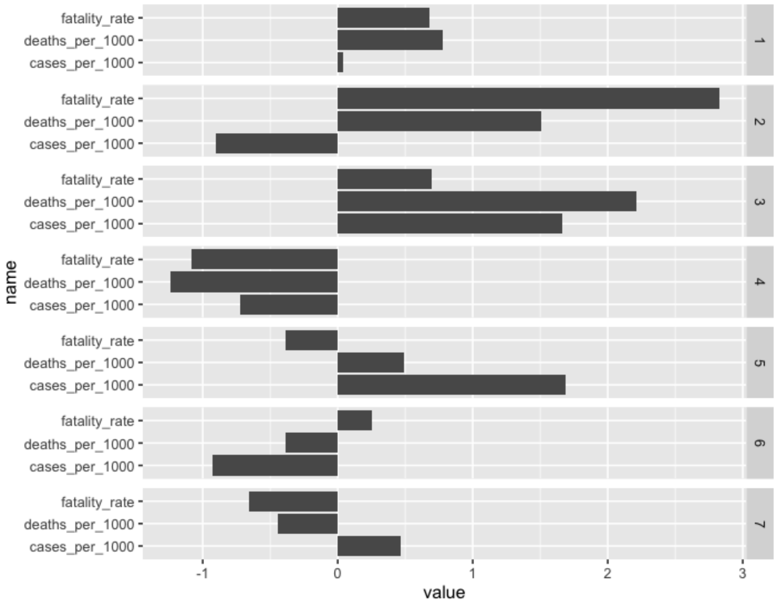


Fig 2.2: Hierarchical Summary

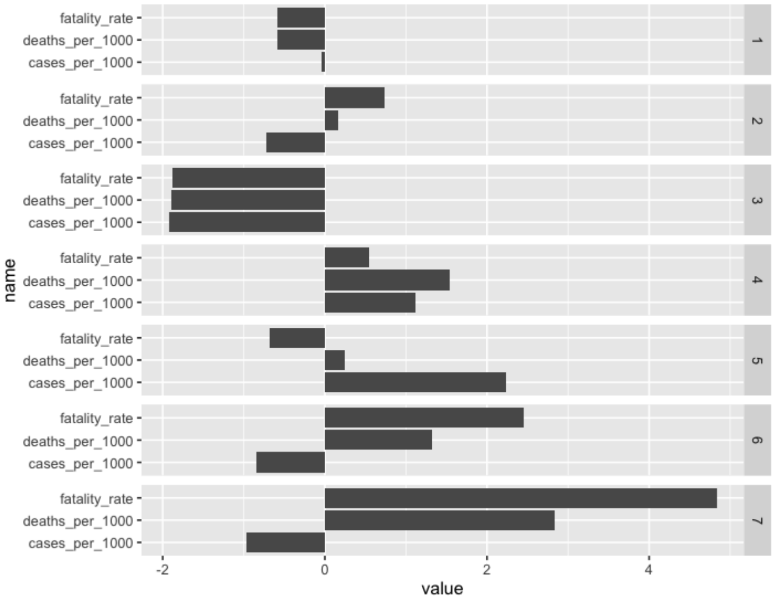


Fig 2.3: K-Means Clusters Map

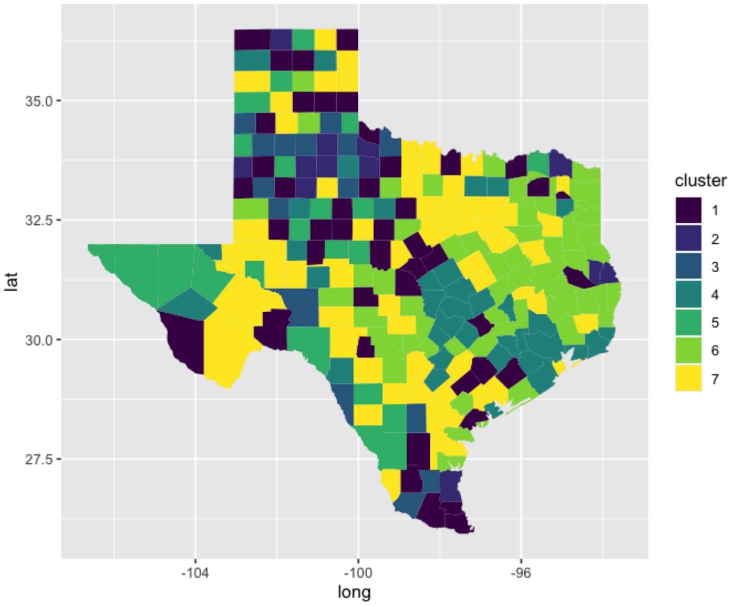


Fig 2.4: Hierarchical Clusters Map

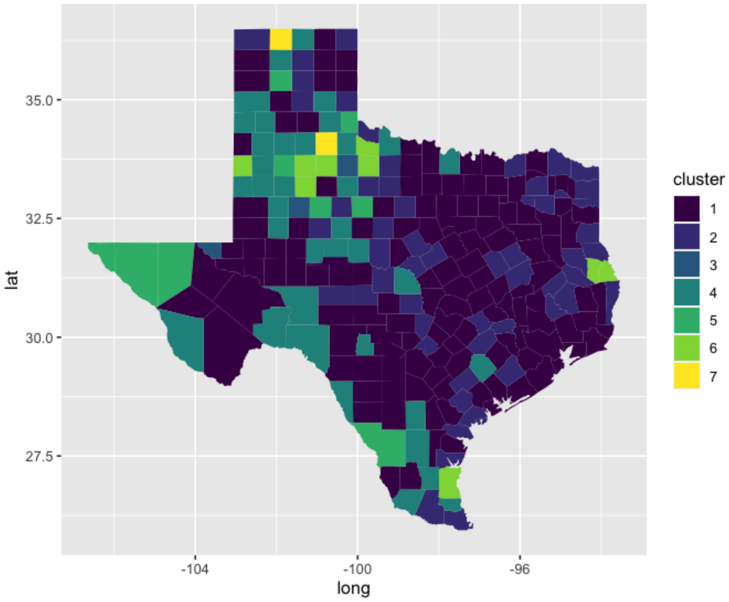


Table 2.5: K-Means Clusters

Cluster	Avg. Poverty Per 1000	Avg. Median Income	Avg. Income Per Capita
1	167	$45,235	$23,438
2	167	$40,924	$21,602
3	201	$42,897	$21,414
4	117	$62,348	$29,674
5	186	$43,749	$20,614
6	156	$48,233	$25,066
7	148	$53,013	$25,753

Table 2.6: Hierarchical Clusters

Cluster	Avg. Poverty Per 1000	Avg. Median Income	Avg. Income Per Capita
1	147	$53,285	$26,040
2	176	$45,524	$23,872
3	110	$75,497	$35,086
4	206	$40,264	$20,656
5	255	$38,610	$17,270
6	184	$36,259	$19,781
7	134	$47,780	$25,633

Taking a look at K-Means clustering first in Figures 2.1, 2.3, and Table 2.5, we can see that there were some correlations and they resulted as we expected. Looking at the summary, we can see that cluster 4 has a low fatality rate, a lower number of deaths per 1000, and a lower number of cases per 1000. We can also see in Table 2.5 that cluster 4 had the highest average median income and lowest average poverty per 1000 people out of all the clusters. Looking at the county map of Texas, we can see several counties in cluster 4 are near large cities like Dallas, Austin, and Houston. **Counties with a higher median income averages and lower poverty per 1000 people tend to have lower fatality rates. This is likely due to better access to resources and since the counties tend to be near large cities.** We can also see the opposite is true as well. In cluster 2, which has a high fatality rate, high number of deaths per 1000 people and low numbers of cases per 1000 people, has the lowest average median income and one of the higher levels of poverty per 1000 people. These counties tend to be located away from large cities. **Counties with lower median income averages tend to have higher fatality rates. This is likely due to limited access to resources. This further reinforces the idea that counties with lower income and high poverty are more at risk from the virus. One recommendation to combat this is to make sure counties with higher poverty rates and limited access to resources, like healthcare, are receiving vaccine doses.** Similar results are seen from the hierarchical clustering, with slightly varying averages across the clusters and map.

C. Cluster 3: Racial Populations per 1000

For our third cluster, we chose to look at racial populations per 1000 people and see if there was any correlation between a particular area that had a high concentration of a particular race and the average cases, deaths, and fatalities rates of those clusters. For our races that were distinguished in this section, we had White, Black, Hispanic, Asian, American Indian, and other races. What we were interested in seeing here was if one particular race had a high association with fatality rates and cases/deaths and where large concentrations of those races were located. The reason we are interested in looking at this is because during the epidemic, it was an area of discussion for COVID-19 response as headlines frequently stated that minority groups suffered from the virus more harshly than white people did. Furthermore, in our previous cluster analysis we saw that there were high fatality rates along the Mexico/US border, thus seeing if these areas were highly concentrated with hispanic people may assist us in understanding whether or not fatality rates were higher with them than with other races. Below you can see that we performed both K-Means and Hierarchical clustering in order to cross reference our results to provide a more holistic view on the clustering. Our explanation to the number of clusters we chose here will be explained later in this report. Below you can see our cluster summaries as depicted in Figures 3.1 and 3.2 followed by the actual counties mapped to the clusters in Figures 3.3 and 3.4 and finally the average case, death, and fatality rates for the clusters in Tables 3.5 and 3.6.

Fig 3.1: K-Means Summary

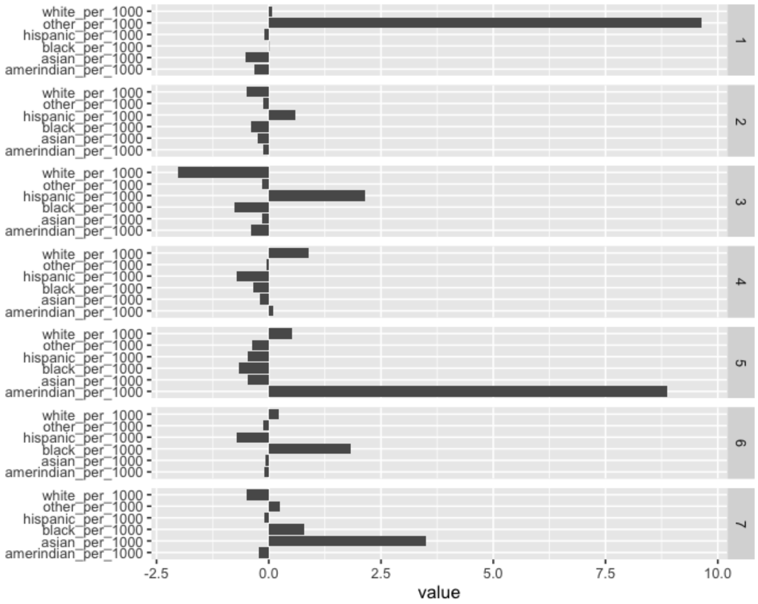


Fig 3.2: Hierarchical Summary

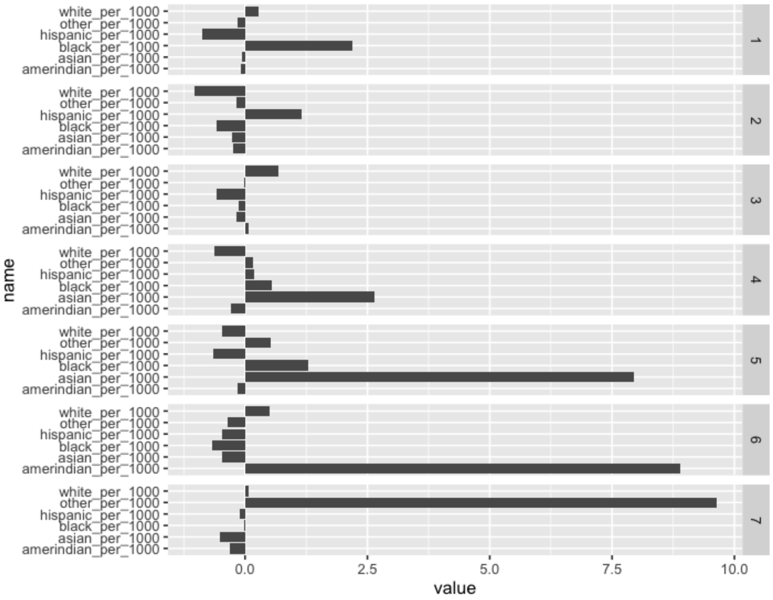


Fig 3.3: K-Means Clusters Map

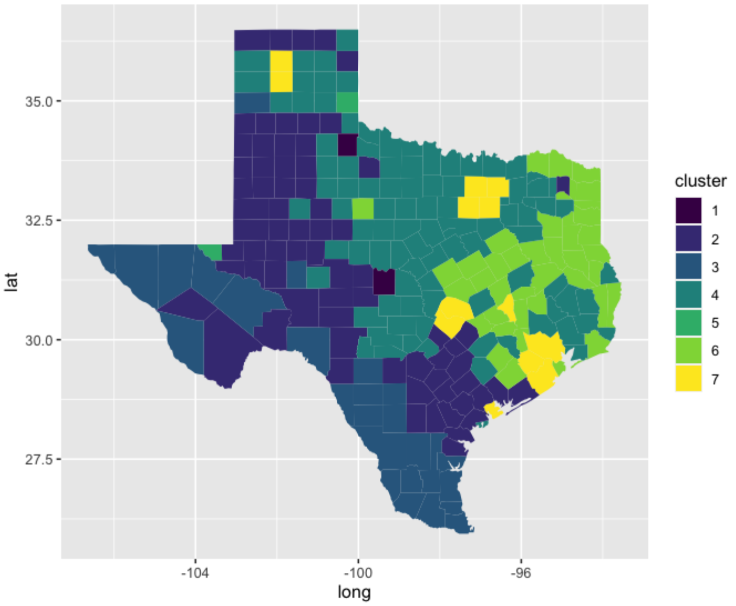


Fig 3.4: Hierarchical Clusters Map

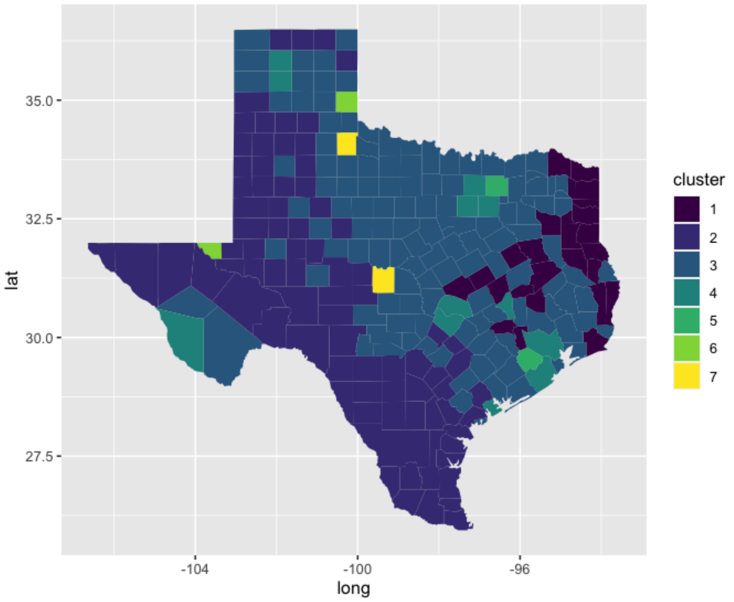


Table 3.5: K-Means Clusters

Cluster	Avg. Cases Per 1000	Avg. Deaths Per 1000	Avg. Fatality Rate
1	88.7	3.20	0.0344
2	91.0	2.30	0.0262
3	106.0	2.54	0.0256
4	74.1	1.81	0.0254
5	46.4	1.32	0.0167
6	68.7	1.66	0.0260
7	80.8	1.03	0.0120

Table 3.6: Hierarchical Clusters

Cluster	Avg. Cases Per 1000	Avg. Deaths Per 1000	Avg. Fatality Rate
1	60.3	1.61	0.0288
2	95.9	2.38	0.0262
3	74.4	1.72	0.0242
4	74.6	1.08	0.0139
5	68.0	0.636	0.00941
6	31.9	0.740	0.00933
7	79.0	2.73	0.0329

Taking a look at K-Means clustering first in Figures 3.1, 3.3, and Table 3.5, we can see that there were some correlations and they resulted somewhat as we expected. First looking at the cluster with the highest fatality rate, we saw that this cluster only included 2 counties and the **predominant race in those counties was listed as other. This was not something we anticipated and was a little strange to observe. We assumed this to be a special case (possibly an outlier) and did not pay as much attention to it as it was not super relevant to what we were looking for.** Therefore, we will put it aside for now. Moving on to the next clusters with the highest fatality rates, we saw that these clusters included 2, 6, and 3. All three of these clusters had fatality rates that were roughly the same all within 0.06% of each other.

Looking at these clusters, we began to see more of what we initially thought we would see. These three clusters contained high concentrations of hispanic and black people. **This would reinforce our understanding that high concentrations of Hispanic and Black populations were located in areas of high fatality rates, implying that those races are in a more dangerous environment with regards to COVID-19.** Furthermore, we saw a correlation to some of the analysis that we saw above with regards to location. Clusters 2 and 3 included high concentrations of hispanic people and were typically located on the map closer to the border between Mexico and The U.S. **This further reinforced our understanding that areas near the border were more negatively affected by COVID-19 with respect to fatality rates and that providing relief to those areas first should be a priority.** This could be due to the impoverished characteristics of the area as noted above. Lastly, the cluster with the lowest fatality rate was cluster 7 which contained high concentrations of Asian people and were typically located in the large cities across Texas. This was an interesting finding as Asians are considered a minority group in the U.S. however they seemed to respond better with regards to fatality rates. **However, the low fatality rate observed in these areas may not be a result of the high concentrations of Asian people living there but instead is a result of the fact that these locations likely had better medical resources and infrastructure to respond to the epidemic.** Regardless, it was still an interesting find nonetheless. When we cross referenced the results with those that we found using Hierarchical Clustering, we essentially saw the same results only the fatality rates were adjusted slightly. The only major difference we saw here was that cluster 6 which was predominantly american indian had a much lower fatality rate along with cluster 5 which was predominantly asian. The reason for this could be due to those clusters including a small number of counties in the first place, a change in including or not including as little as one more county could vastly change those numbers. Otherwise, the results agreed across both clustering methods.

D. Cluster 4: Poverty per 1000, Hispanics per 1000, and Median Age

For our fourth cluster, we chose to look at poverty per 1000 people, hispanics per 1000 people, and median age to see if there was any correlation between areas with high concentrations of these features and the average cases, deaths, and fatalities rates. The reason we chose to look at these particular features was because in our previous clusterings, we saw a trend among hispanic people, older people, and impoverished people with higher fatality rates. Due to this, we wanted to see if any particular cluster of these features showed the highest fatality rate so that we could better understand where priority should lay in terms of assisting the areas that need the help the most. Just like with our other clusters, below you can see both K-Means and Hierarchical clustering which we used in order to cross reference our results to provide a more holistic view on the analysis. Our explanation to the number of clusters we chose here will be explained later in this report. Below you can see our cluster summaries as depicted in Figures 4.1 and 4.2

followed by the actual counties mapped to the clusters in Figures 4.3 and 4.4 and finally the average case, death, and fatality rates for the clusters in Tables 4.5 and 4.6.

Fig 4.1: K-Means Summary

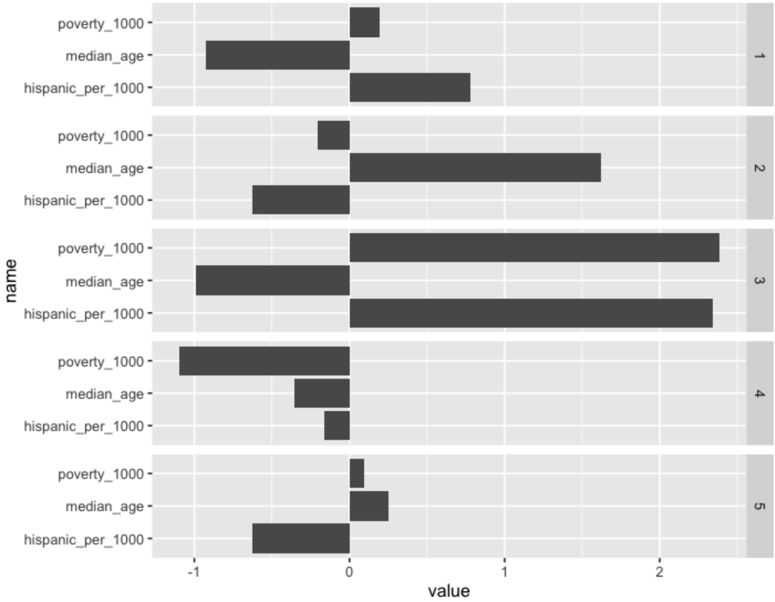


Fig 4.2: Hierarchical Summary

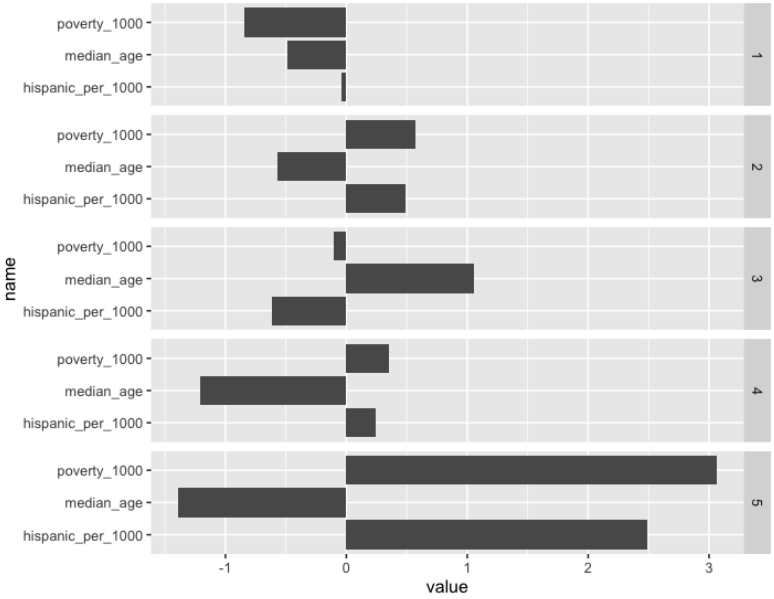


Fig 4.3: K-Means Clusters Map

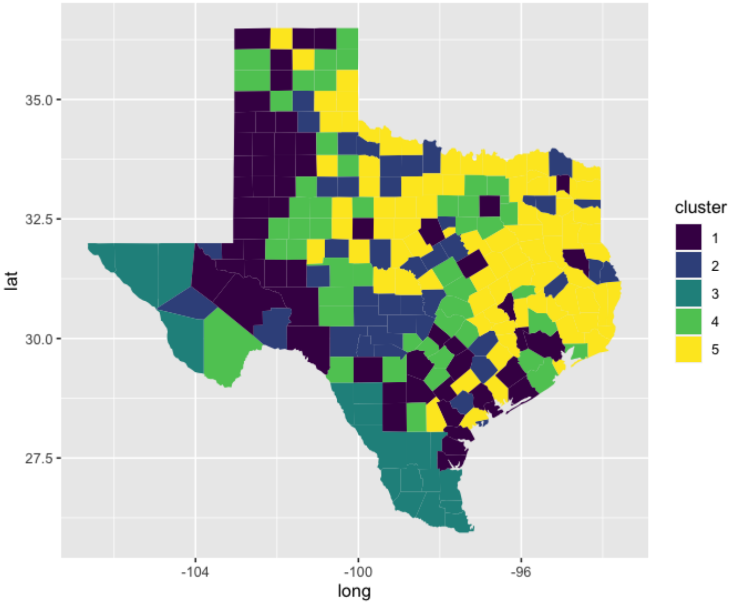


Fig 4.4: Hierarchical Clusters Map

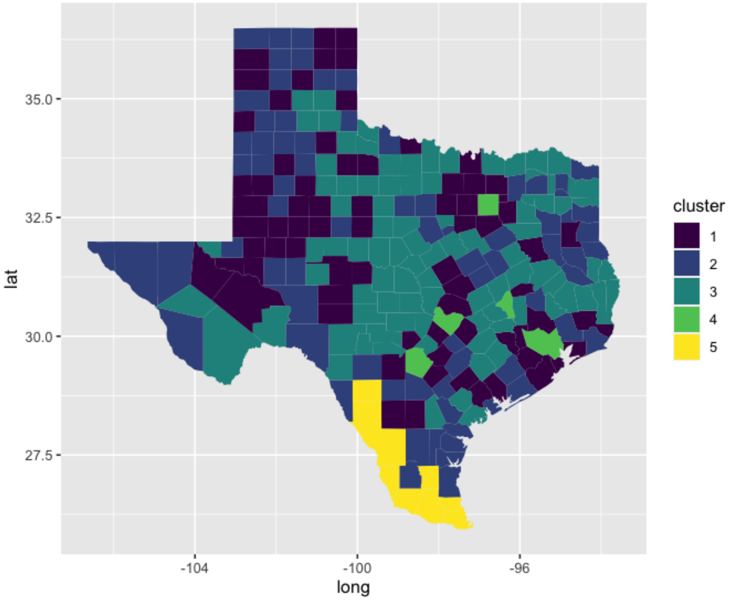


Table 4.5: K-Means Clusters

Cluster	Avg. Cases Per 1000	Avg. Deaths Per 1000	Avg. Fatality Rate
1	96.3	2.30	0.0239
2	69.0	2.03	0.0301
3	105.0	2.66	0.0275
4	80.1	1.43	0.0182
5	73.6	1.88	0.0268

Table 4.6: Hierarchical Clusters

Cluster	Avg. Cases Per 1000	Avg. Deaths Per 1000	Avg. Fatality Rate
1	82.3	1.38	0.0167
2	87.6	2.21	0.0265
3	66.9	1.85	0.0288
4	72.4	0.857	0.0119
5	98.2	2.65	0.0288

Taking a look at K-Means clustering in Figures 4.1, 4.3, and Table 4.5, we can see that there were some interesting correlations. Looking first at the cluster with the highest fatality rate, cluster 2, we could see that this cluster contained a large concentration of older people. This cluster was also located in more rural areas closer to the center of texas and scattered near the panhandle and border to Mexico. This correlation was exactly what we were looking for when initially analyzing these results. **It reinforces the understanding that the areas with older populations had a higher fatality rate and that providing assistance to those locations first would be a good idea.** Moving on to the next most fatal cluster, cluster 3, we can see that this cluster had high concentrations of hispanic and impoverished people. **Again, this reinforces some of the results we saw earlier implying that areas along the border that included large concentrations of impoverished and hispanic people had high fatality rates and providing assistance to these areas should be a priority as well.** Another interesting cluster was cluster 5 which had a somewhat high fatality rate and it turns out that this cluster had an above average median age and poverty. **It also just so happens that where this cluster falls on the map is also where we saw high concentrations of black people which could reinforce what we looked at earlier with regards to minority groups.** Lastly, the least fatal cluster was cluster 4 which contained lower concentrations of impoverished, older, and hispanic people. **This cluster was located primarily near large cities and implies that these locations were safer to be in with regards to COVID-19 fatalities. Again, as mentioned above, this could be likely due to the areas being more wealthy and having better medical care to support victims of the COVID-19 epidemic.** When cross referencing the K-Means results with the Hierarchical results, we saw similarities that backed our initial assumptions. The only change we saw was with cluster 4 which had the lowest fatality rate and it did not quite match the matching cluster we saw in K-Means. This was likely due to the fact that one or two counties were included/not included which changed the rate for that cluster and thus brought the fatality rate down. That being said, the results we were most interested in, age and impoverish/hispanic people, were unchanged.

E. Cluster 5: Education Level and Poverty per 1000

The fifth cluster includes the following features: poverty per 100 people, high school including GED per 1000 people, high school diploma per 1000 people, bachelor's degree per 1000 people, and associates degree per 1000 people. These features were chosen to find relationships between level of education and poverty and see how those factors affected average cases per 1000 people, average deaths per 1000 people, and average fatality rates across the counties in each cluster. It was expected that the clusters with higher levels of education would have somewhat low levels of poverty and therefore lower deaths per 1000 peoples and lower fatality rates and this trend would be seen in counties with large cities. For example, Houston in Harris county. It was also expected clusters with lower levels of education would have higher levels of poverty and higher fatality rates. The reasoning behind the expected relation between high education, low poverty, and fatality rates and lower education, high poverty, and high fatality rates is explained by access to healthcare and other resources. High education, for the most part, means access to better jobs, better pay, and better resources, like healthcare. Both K-Means and Hierarchical clustering were used to gather a detailed understanding of the relationships between the attributes. Later in the report, it will be explained why 7 clusters were chosen for the optimal number of clusters. The cluster summaries can be seen in Figures 5.1 and 5.2. The counties mapped to their clusters can be seen in Figures 5.3 and 5.4. The average cases per 1000, deaths per 1000, and fatality rates for each cluster can be seen in Table 5.5 and 5.6.

Fig 5.1: K-Means Summary for Cluster 5

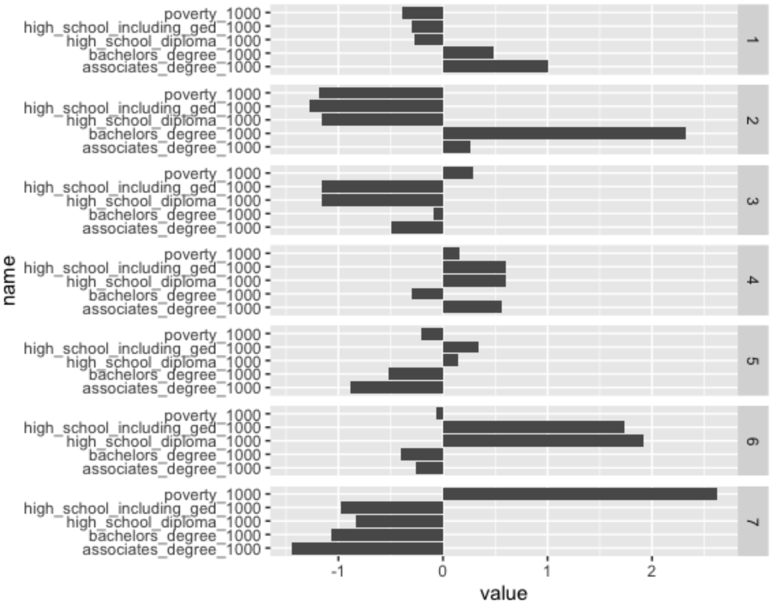


Fig 5.2: Hierarchical Summary for Cluster 5

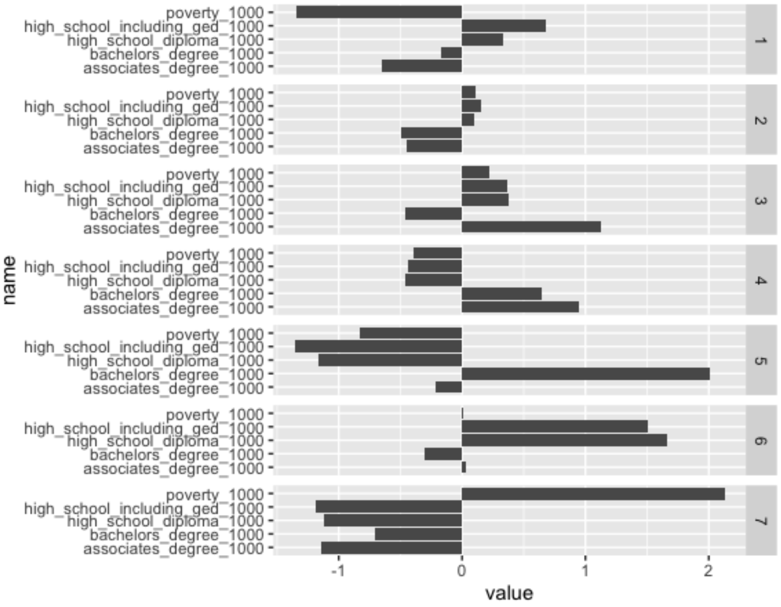


Fig 5.3: K-Means Map for Cluster 5

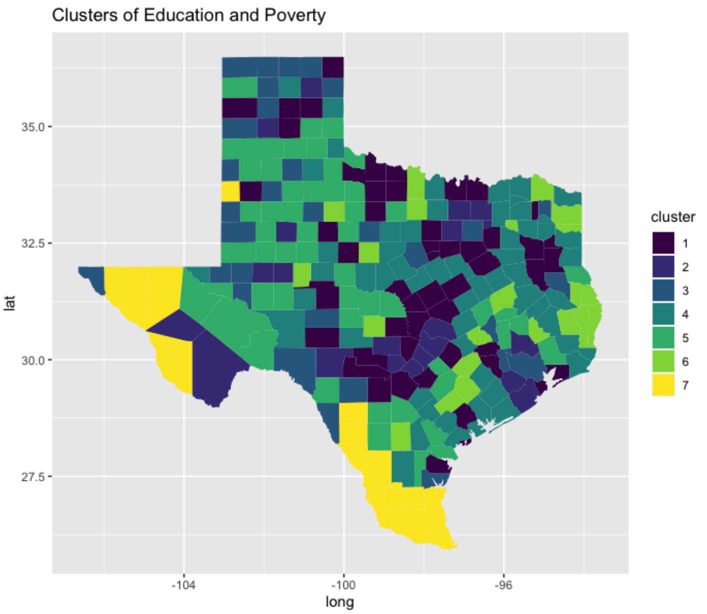


Fig 5.4: Hierarchical Map for Cluster 5

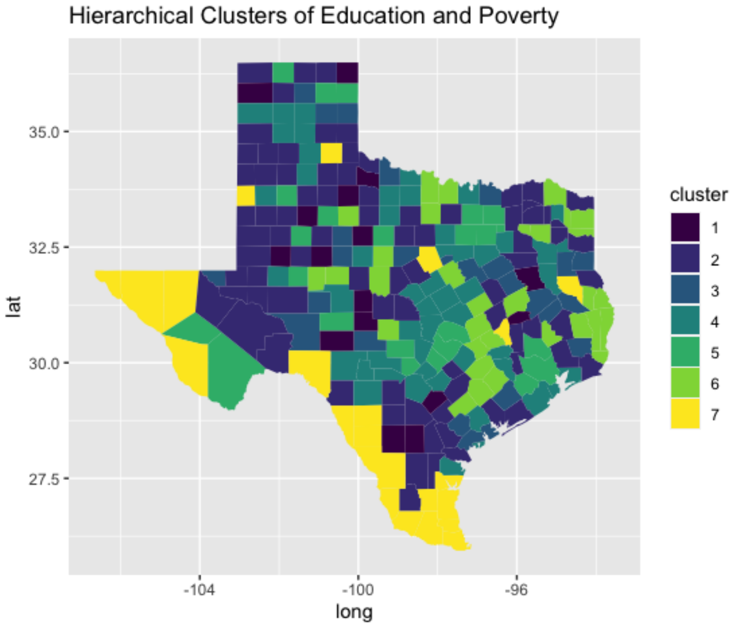


Table 5.5: K-Means Clusters for Cluster 5

Cluster	Avg. Cases Per 1000	Avg. Deaths Per 1000	Avg. Fatality Rate
1	76.1	1.72	0.0238
2	74.5	0.902	0.0120
3	96.3	2.13	0.0236
4	73.9	1.83	0.0251
5	90.7	2.39	0.0270
6	66.4	2.18	0.0339
7	100	2.74	0.0303

Table 5.6: Hierarchical Clusters for Cluster 5

Cluster	Avg. Cases Per 1000	Avg. Deaths Per 1000	Avg. Fatality Rate
1	83.4	2.04	0.0258
2	80.7	2.00	0.0252
3	73.0	1.79	0.0240
4	76.2	1.49	0.0208
5	75.5	0.853	0.0121
6	62.2	1.84	0.0315
7	98.6	2.66	0.0294

First, we will look at the K-Means clustering in Figures 5.1, 5.3, and Table 5.5. We can see the resulting clusterings show what was somewhat expected. In the summary (Figure 5.1), cluster 1 and 2 show that there are higher numbers of people with a bachelors and associates degree and lower levels of poverty. When we look at the county map of Texas (Figure 5.3), we can see these counties tend to be counties with large cities or counties near large cities. It was a surprise to see such a low level of poverty in cluster 2 because counties with large cities tend to have a great diversity of people and skills, however it was not too surprising because jobs requiring high levels of education tend to be in large cities. Table 5.5 also tells us the fatality rate of cluster 2 is the lowest out of all the others, with a rate of 0.0120. Cluster 1 also has one of the lower fatality rates, with a rate of 0.0238. **Higher educated people tend to be located in or near countries with large cities. These counties tend to have lower levels of poverty and fatality rates, likely because higher educated people are qualified for better paying jobs and have access to better resources and healthcare.** Cluster 7 has very low levels of education, with few people having received a high school GED. We can also see cluster 7 counties tend to be located away from large cities along the Texas/Mexico border. These counties also have high levels of poverty and cluster 7 has one of the highest average fatality rates out of the other clusters, with a rate of 0.0303. Cluster 6 has the highest fatality rate (0.0339). The education level in these counties show many people have a high school diploma or GED and are located farther away from large cities. **Lower educated people tend to be located in counties farther away from large cities and these counties tend to have higher levels of poverty and fatality rate. This is likely due to limited access and ability to pay for healthcare.** Similar results are seen from the hierarchical clustering, with slightly varying averages across the clusters and map.

Number of Clusters for Each Method

In this section we will analyze how we determined a suitable number of clusters for each of the above clustering methods. This section is broken down by each method.

A. Cluster 1: Poverty per 1000, Commuters per 1000, and Median Age

For cluster 1 we looked at the within sum of squares, average silhouette width, and Dunn Index charts of the scaled subset of data to identify what the best number of clusters would be. For this cluster, we ultimately chose 5 clusters. We chose 5 clusters because on the WSS graph (Fig 1.7) we saw a potential knee that would indicate an ideal cluster number. Furthermore, on the DI graph (Fig 1.9), we saw a peak at 5 which further indicates that is the best cluster number. We did see an issue when trying to use the ASW graph (Fig 1.8) since it didn't seem to provide the best results, so we ignored it when choosing. After clustering with K-Means, we did see a within cluster sum of squares by cluster percentage of 66% which was not the best but still not bad. Furthermore, after analyzing the results, we believe that we saw the best clustering of the data with this number of clusters.

Fig 1.7: Within Sum of Squares

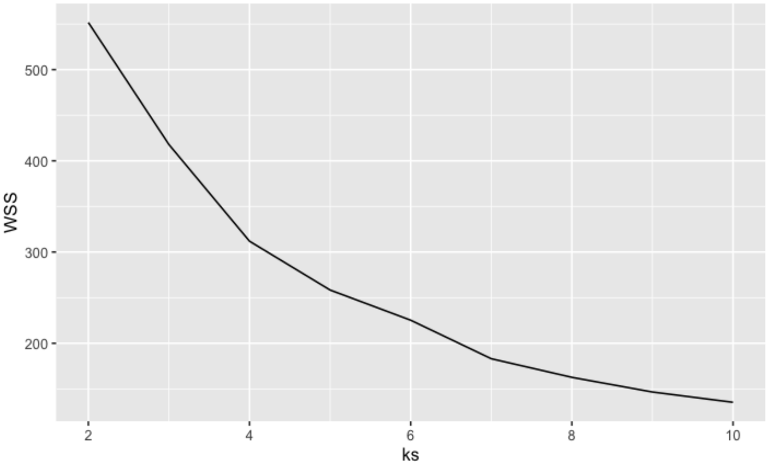


Fig 1.8: Average Silhouette Width

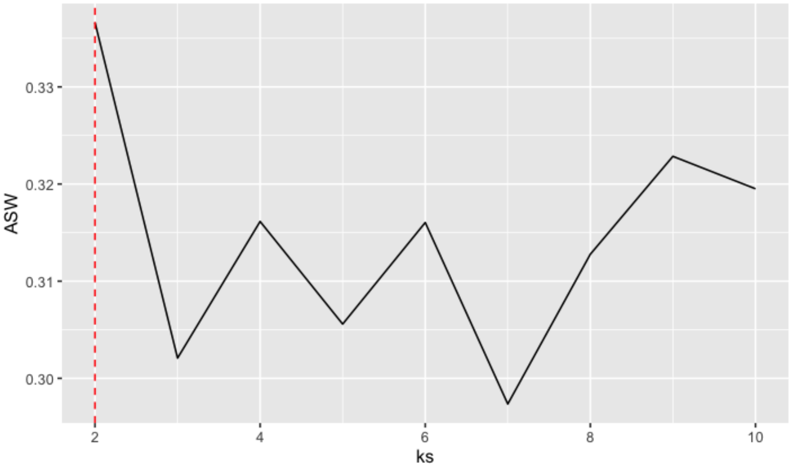
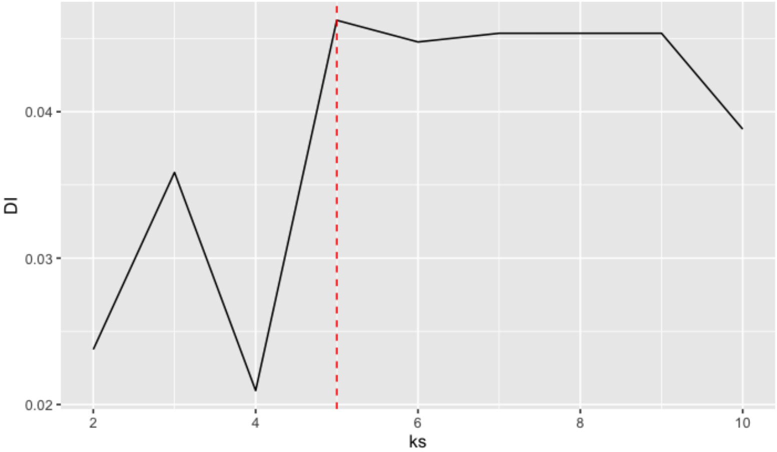


Fig 1.9: Dunn Index



B. Cluster 2: Cases/Deaths per 1000 and Fatality Rate

For cluster 2 we looked at the within sum of squares, average silhouette width, and Dunn Index charts of the scaled subset of data to identify what the best number of clusters would be. For this cluster, we ultimately chose 7 clusters. We chose 7 clusters because on the WSS graph (Fig 2.7) we saw a potential knee that would indicate an ideal cluster number. Furthermore, on the ASW graph (Fig 2.8) we saw a peak at 7 which further indicates that 7 is an ideal number of clusters. Lastly, on the DI graph (Fig 2.9), we saw a local peak at 7. It wasn't the global peak but it was a peak nonetheless and agreed with what we saw in our other graphs. After clustering with K-Means, we did see a within cluster sum of squares by cluster percentage of 79.9% which was pretty good. Furthermore, after analyzing the results, we believe that we saw the best clustering of the data with this number of clusters.

Fig 2.7: Within Sum of Squares

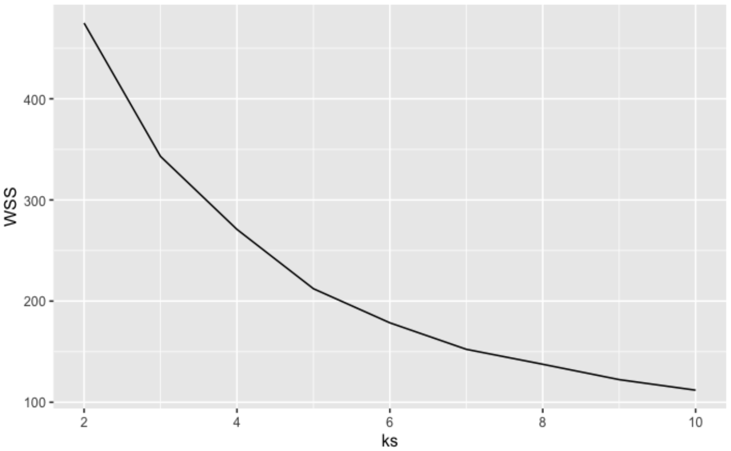


Fig 2.8: Average Silhouette Width

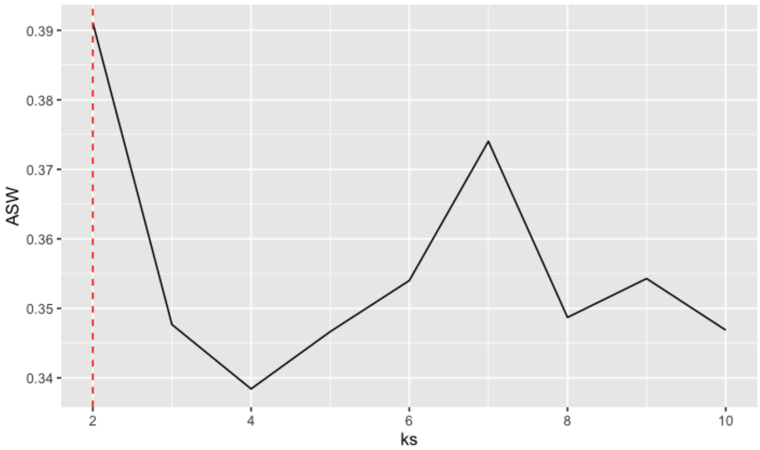
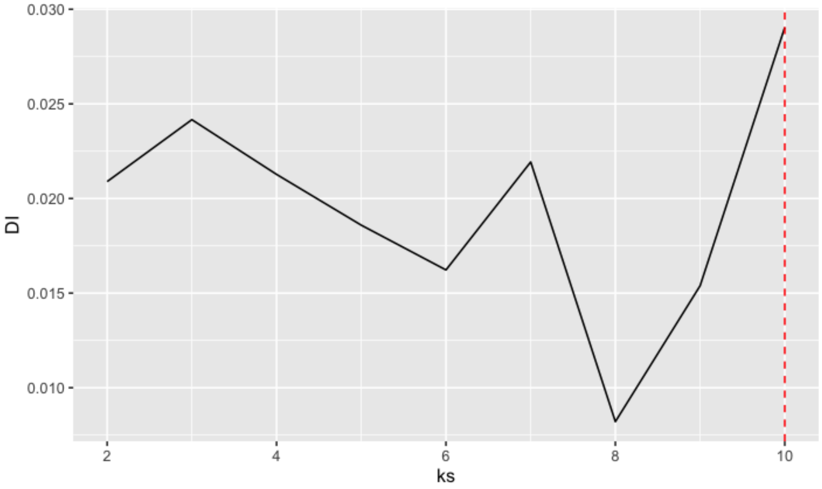


Fig 2.9: Dunn Index



C. Cluster 3: Racial Populations per 1000

For cluster 3 we looked at the within sum of squares, average silhouette width, and Dunn Index charts of the scaled subset of data to identify what the best number of clusters would be. For this cluster, we ultimately chose 7 clusters. We chose 7 clusters because on the WSS graph (Fig 3.7) we saw a potential knee that would indicate an ideal cluster number. Furthermore, on the ASW graph (Fig 3.8) we saw a local peak at 7 which was not far off from the global peak at 5. Lastly, on the DI graph (Fig 3.9), we saw another local peak at 7. It also was better than the global peak at 9 when compared with the other graphs. Overall, we saw the highest numbers and best knee across the various graphs with a cluster size of 7. Furthermore, after clustering with K-Means, we did see a within cluster sum of squares by cluster percentage of 75.1% which was pretty good. After analyzing the results, we believe that we saw the best clustering of the data with this number of clusters.

Fig 3.7: Within Sum of Squares

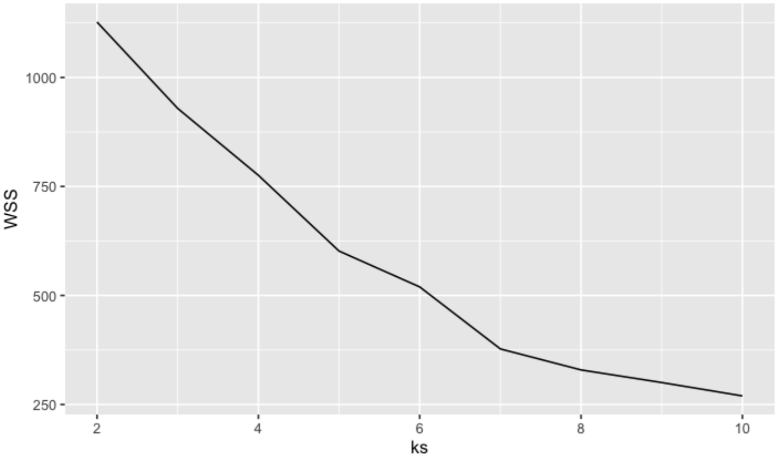


Fig 3.8: Average Silhouette Width

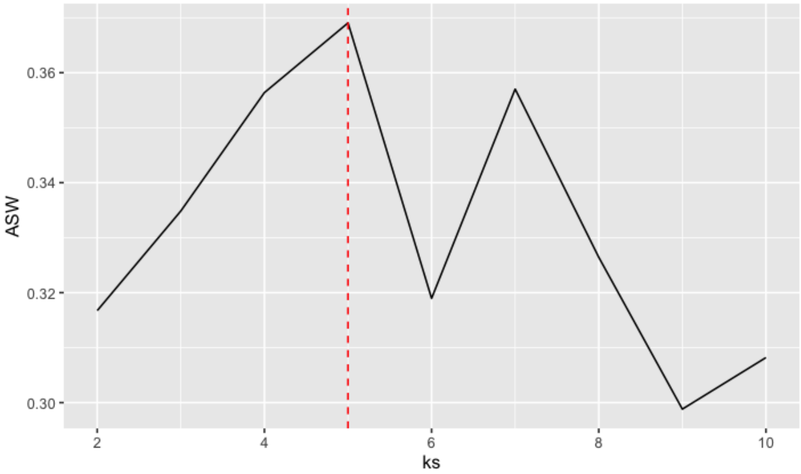
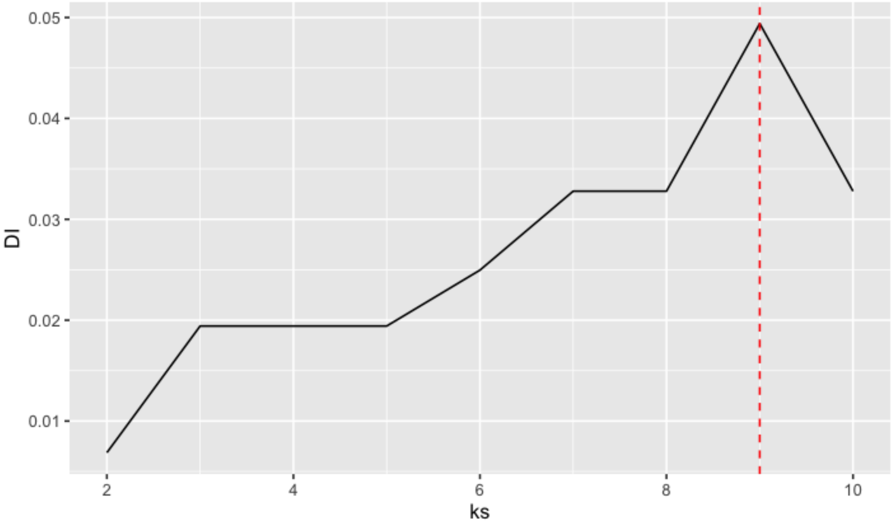


Fig 3.9: Dunn Index



D. Cluster 4: Poverty per 1000, Hispanics per 1000, and Median Age

For cluster 4 we looked at the within sum of squares, average silhouette width, and Dunn Index charts of the scaled subset of data to identify what the best number of clusters would be. For this cluster, we ultimately chose 5 clusters. We chose 5 clusters because on the WSS graph (Fig 4.7) we saw a potential knee that would indicate an ideal cluster number. Furthermore, on the ASW graph (Fig 4.8) we saw a local peak at 5 which was better than the global peak at 2. Lastly, on the DI graph (Fig 4.9), we didn't see too much useful information here so we somewhat ignored it. Overall, these graphs were not very helpful in assisting us in finding the best cluster number. We chose 5 because we figured it was the best cluster number to choose based on all the graphs, considering that they didn't really agree with each other. After clustering with K-Means, we did see a within cluster sum of squares by cluster percentage of 71.5% which was ok, and after analyzing the results, we believe that we saw the best clustering of the data with this number of clusters.

Fig 4.7: Within Sum of Squares

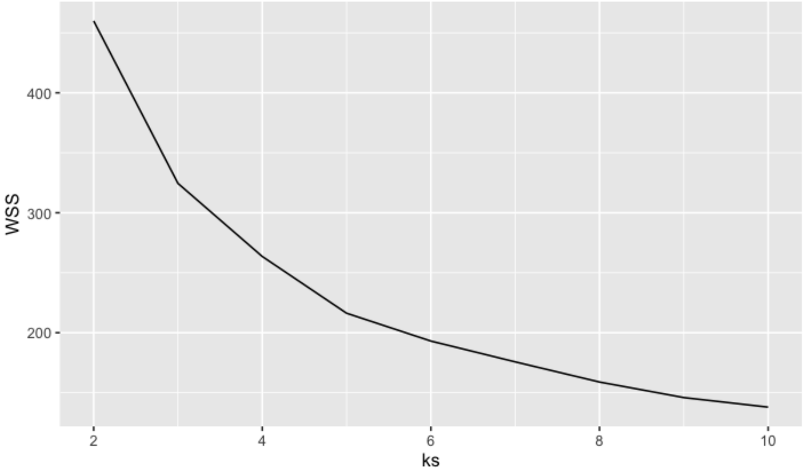


Fig 4.8: Average Silhouette Width

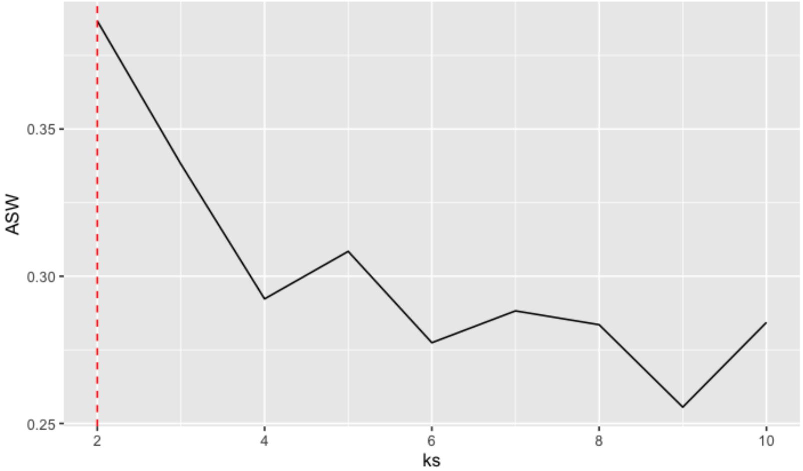
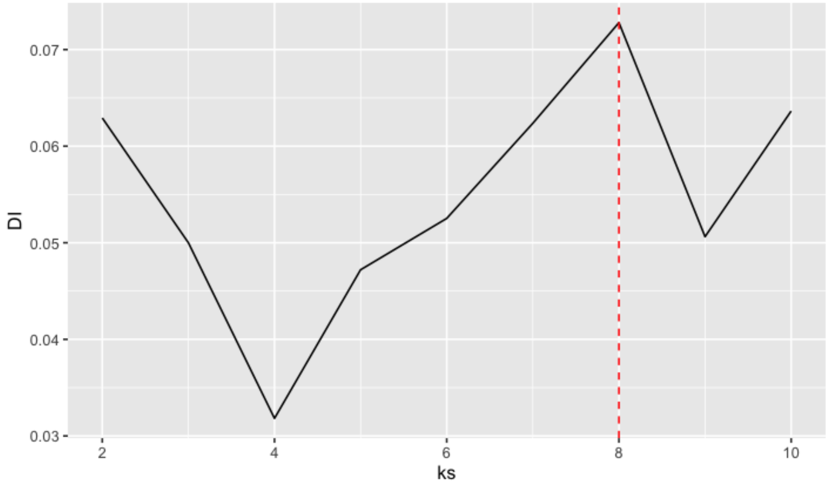


Fig 4.9: Dunn Index



E. Cluster 5: Education Level and Poverty per 1000

Lastly, for cluster 5 we looked at the within sum of squares, average silhouette width, and Dunn Index charts of the scaled subset of data to identify what the best number of clusters would be. For this cluster, we ultimately chose 7 clusters. We chose 7 clusters because on the WSS graph (Fig 5.7) we saw a potential knee that would indicate an ideal cluster number. On the ASW graph (Fig 5.8), a global peak was given at 2 and a local peak given at 5 and 6. On the DI graph (Fig 5.9), we saw what was close to a peak at 7 clusters. When comparing the ASW peaks to the WSS knees, we decided 2, 5, and 6 were not the ideal cluster numbers. The DI graph gave a local peak at 7 clusters, which is where a knee appeared on the WSS graph. Overall, the DI graph and ASW graphs had conflicting peaks. We chose 7 because we figured it was the best cluster number to choose based on all the graphs, considering that they didn't really agree with each other. After clustering with K-Means, we did see a within cluster sum of squares by cluster percentage of 67.2% which was ok, and after analyzing the results, we believe that we saw the best clustering of the data with this number of clusters.

Fig 5.7: Within Sum of Squares

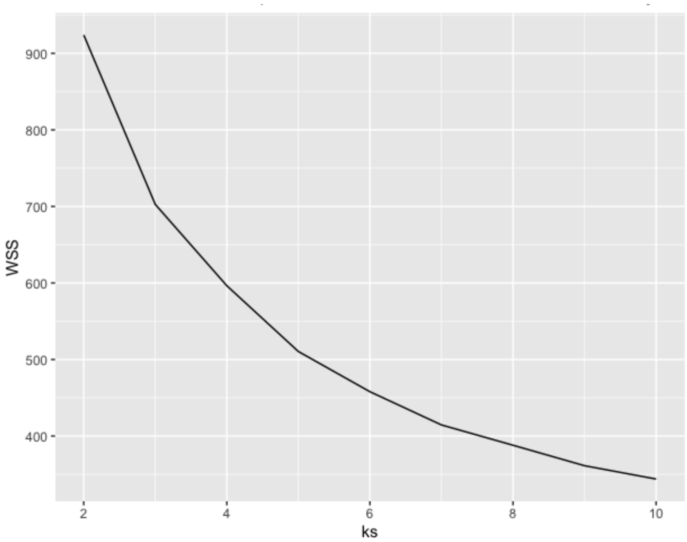


Fig 5.8: Average Silhouette Width

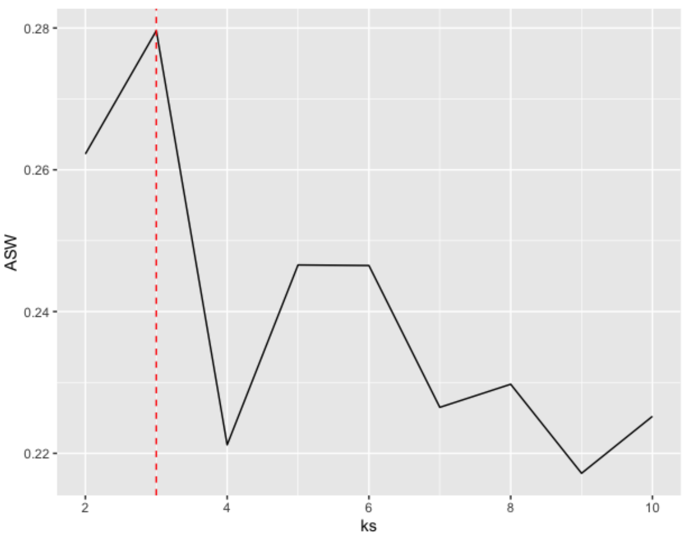
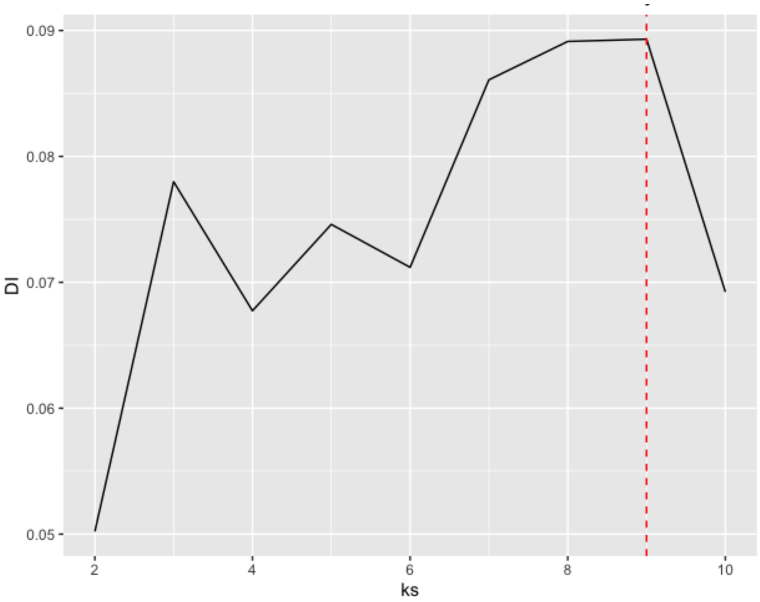


Fig 5.9: Dunn Index



Internal Validation Measures to Describe and Compare Clusterings

Below is some analysis on the internal validation measures we used to describe and compare the clusterings that were performed. For each of the clusters we provided a dissimilarity plot that indicates a closer similarity within clusters by darker colors and an average silhouette plot which indicates a closer similarity within clusters by a higher average silhouette width.

A. Cluster 1: Poverty per 1000, Commuters per 1000, and Median Age

For cluster 1 we saw that for 5 clusters, our average silhouette width was 0.3 and our similarity within clusters in Fig 1.10 was quite similar. However, for clusters 1, 3, and 4, we do have some instances with negative silhouette widths. This means these counties are possibly incorrectly clustered. Overall, a large majority of the silhouette widths are positive so we deem this clustering method adequate. Therefore, we believe that our number of clusters here for kmeans was good and we could rely on the assumptions drawn from these clusters.

Fig 1.10: Dissimilarity Plot

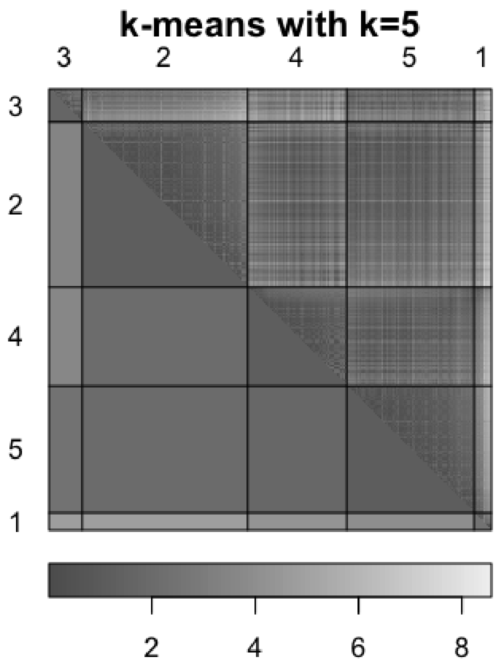
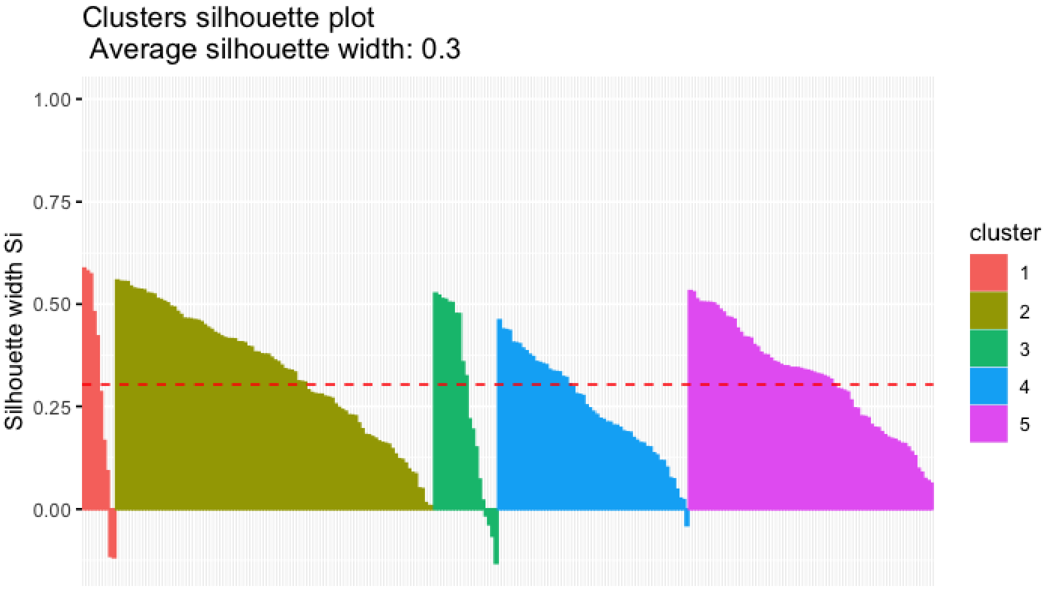


Fig 1.11: Average Silhouette Plot



B. Cluster 2: Cases/Deaths per 1000 and Fatality Rate

For cluster 2 we saw that for 7 clusters, our average silhouette width was 0.36 and our similarity within clusters in Fig 1.10 was quite similar and had dark colors (depicting similarity). We can see that there is one instance in both cluster 3 and 5 where a county has a negative silhouette width. However, only having 2 incorrectly clustered counties will not severely impact our conclusions drawn from this clustering. Therefore, we believe that our number of clusters here for K-Means was good (better than cluster 1) and we could rely on the assumptions drawn from these clusters.

Fig 2.10: Dissimilarity Plot

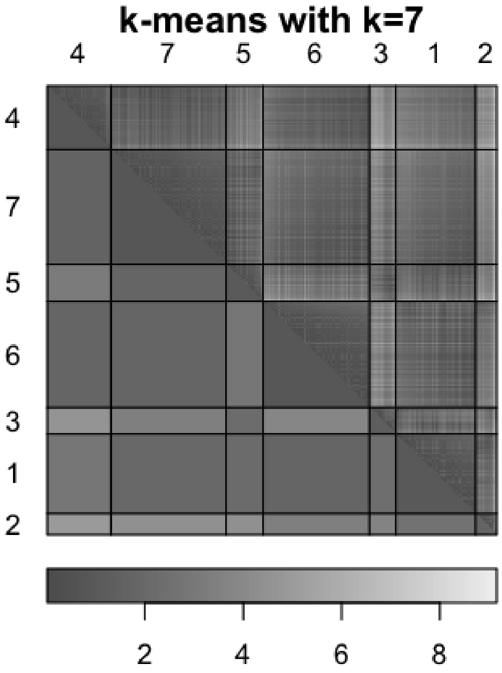
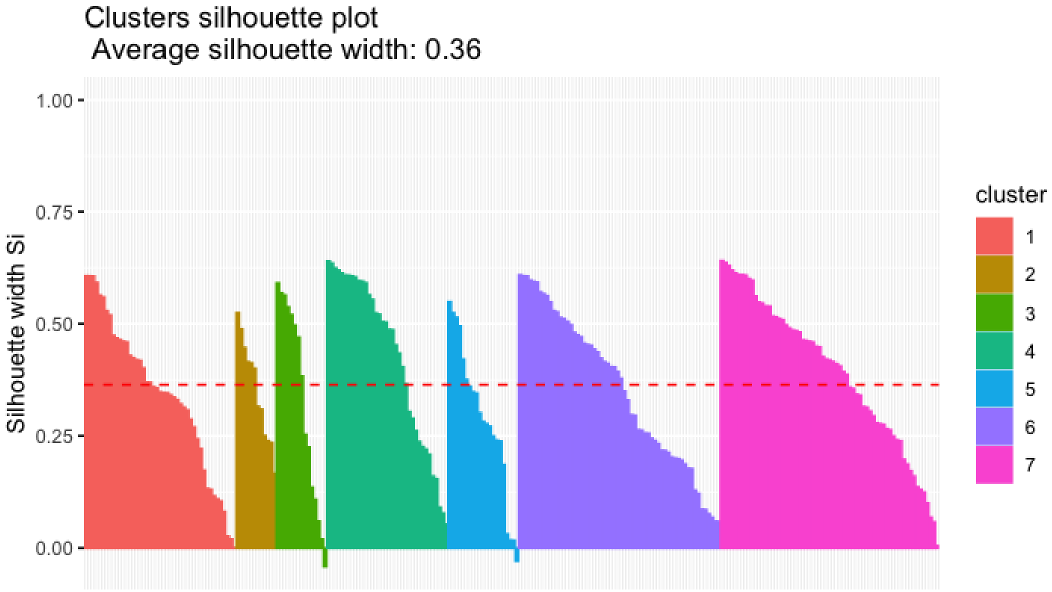


Fig 2.11: Average Silhouette Plot



C. Cluster 3: Racial Populations per 1000

For cluster 3 we saw that for 7 clusters, our average silhouette width was 0.36, which was the same as cluster 2, and our similarity within clusters in Fig 1.10 had darker colors depicting similarity within clusters. Overall, clusters 1, 3, and 5 have very strong silhouette widths and we only see 2 instances with negative silhouette widths (one in cluster 4 and one in cluster 7). We believe the silhouette widths in clusters 1 and 5 are so high because there are only 2 counties in each of these clusters, possibly meaning that these counties are outliers. We believe the clusterings in this cluster method are very good overall. Therefore, we believe that the number of clusters here for K-Means was good and we could rely on the assumptions drawn from these clusters.

Fig 3.10: Dissimilarity Plot

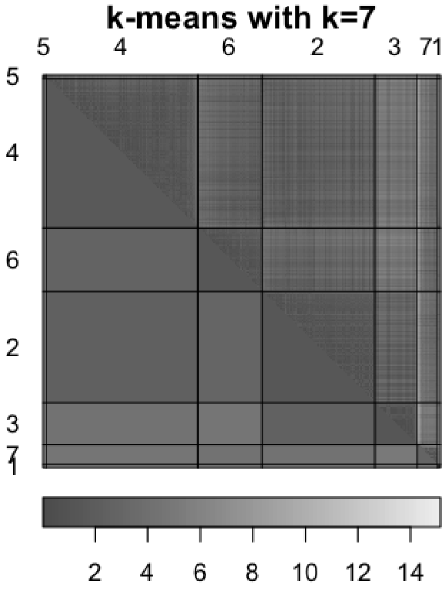
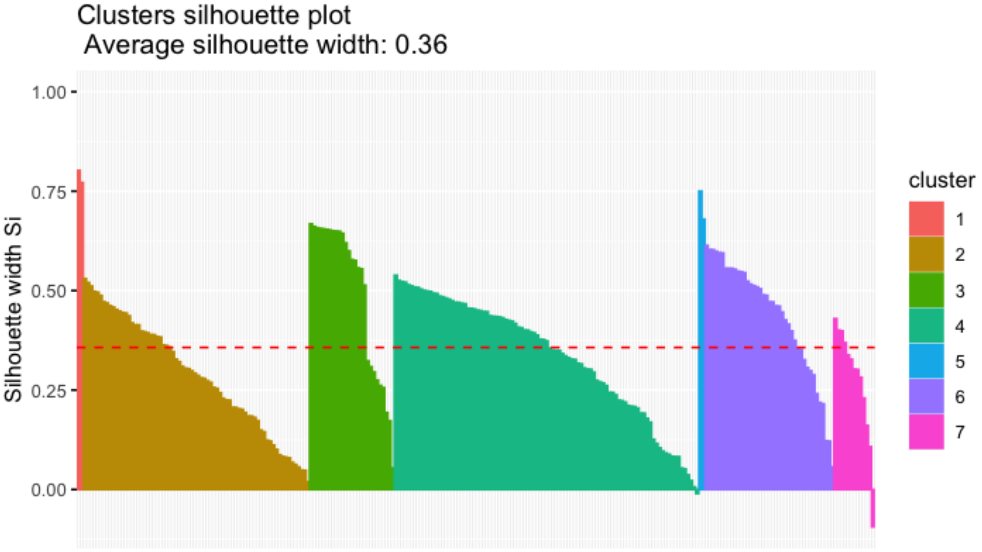


Fig 3.11: Average Silhouette Plot



D. Cluster 4: Poverty per 1000, Hispanics per 1000, and Median Age

For cluster 4 we saw that for 5 clusters, our average silhouette width was 0.31 and our similarity within clusters in Fig 1.10 had decent similarity within clusters. With this clustering method, we can see that both clusters 1 and 4 have two instances of negative silhouette width and cluster 2 has four instances of negative silhouette width. However, the overall silhouette width in these clusters seem to be good. Therefore, we believe that the number of clusters here for K-Means was good and we could rely on the assumptions drawn from these clusters.

Fig 4.10: Dissimilarity Plot

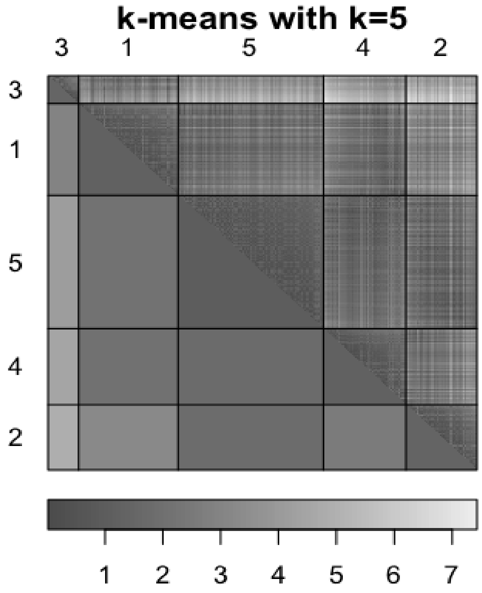
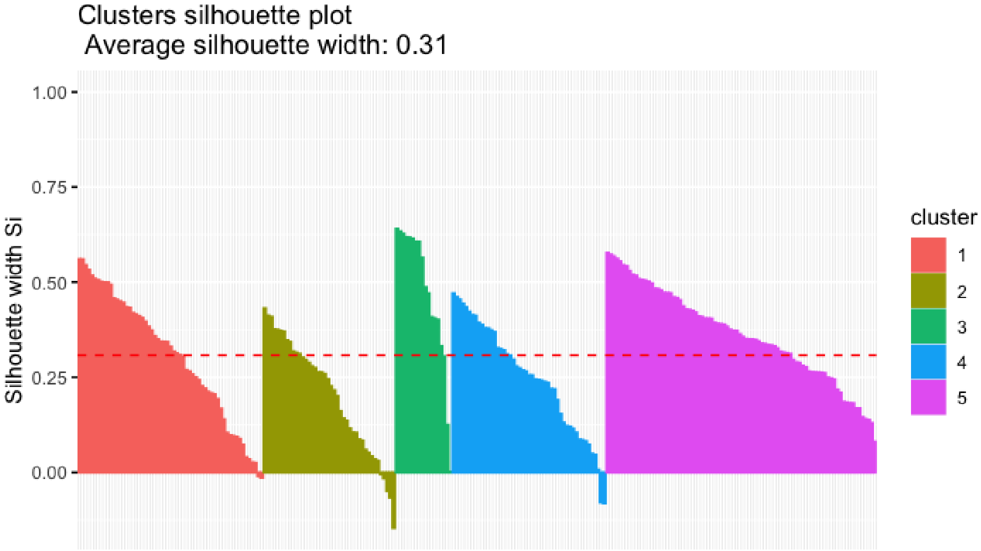


Fig 4.11: Average Silhouette Plot



E. Cluster 5: Education Level and Poverty per 1000

For cluster 5, we had an average silhouette width of 0.22, which is not a high number but could be brought down by the negative attributes that seem to not fit best in their cluster. Overall, we can see that these attributes share similarities by the darker spots on the dissimilarity plot. We believe the number of instances with negative silhouette widths may be due to the number of features we looked at in this clustering method. However, the vast majority of the instances have decent positive silhouette widths and the values of the negative instances aren't large enough to make us reconsider this clustering method. We believe the number of clusters that were chosen was adequate and we can reasonably rely on the assumptions drawn from the clusters.

Fig 5.10: Dissimilarity Plot

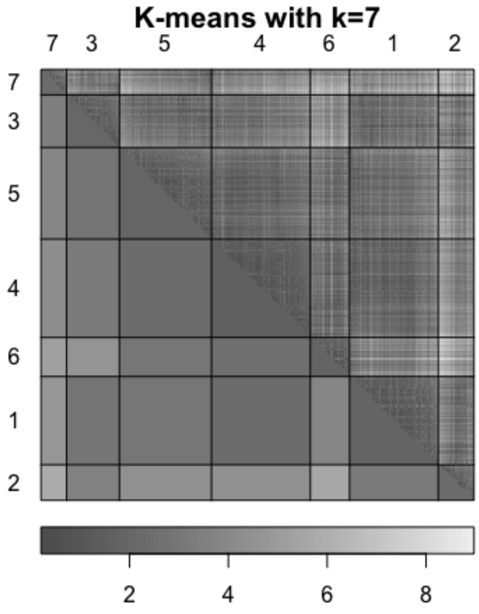
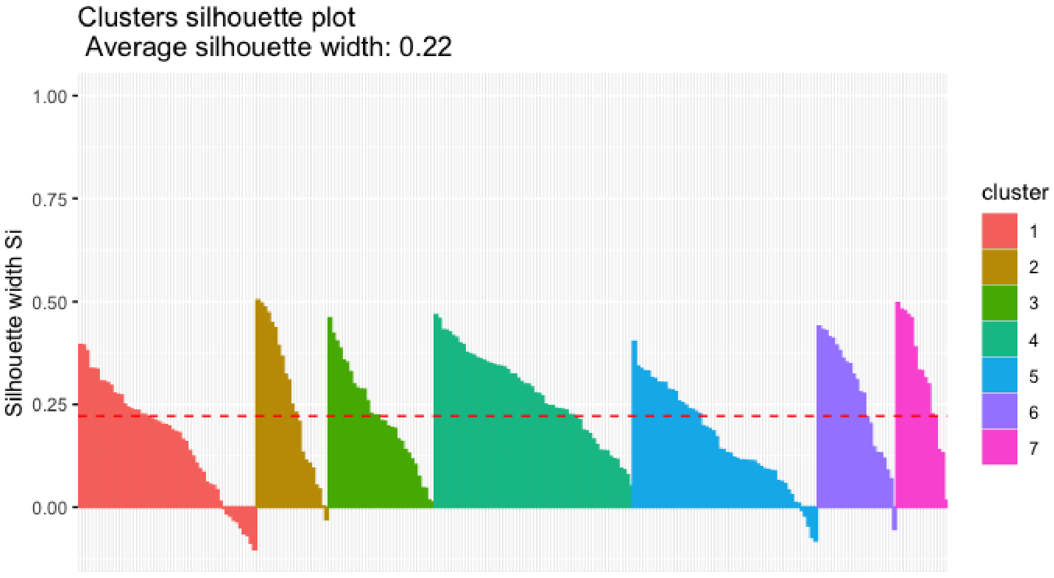


Fig 5.11: Silhouette Plot



IV. Evaluation

Below are the key findings that we observed when conducting this report:

- We expected the areas with high public transportation use to be more dangerous with higher fatality rates but that was not the case. This could be due to public transportation being limited or shut down during the pandemic.
- We found that areas with high poverty, which also seem to be located along the Mexico/U.S. border, have higher fatality rates and imply that they are dangerous locations to reside in with regards to COVID-19.

- We found that the fatality rate was worse in areas where older people resided, this makes sense as older people are more at risk from contracting the virus and suffering from it fatally.
- We found that counties with higher average median incomes and lower poverty per 1000 people tended to have lower fatality rates. This is likely due to better access to medical care since the counties tend to be near/in large cities.
- We found counties with high concentrations of Hispanic and Black populations were located in areas with high fatality rates, implying that those races are in a perilous environment with regards to COVID-19.
- We found a low fatality rate in areas with high concentrations of Asian people. However, this was likely due to the fact that these areas were also located in/near large cities where better medical care could be obtained and not due to their race.
- We found that higher educated people tend to be located in or near countries with large cities. These counties tend to have lower levels of poverty and fatality rates, likely because higher educated people are qualified for better paying jobs and have access to better resources and healthcare.
- We found that lower educated people tend to be located in counties farther away from large cities and these counties tend to have higher levels of poverty and fatality rate. This is likely due to limited access and ability to pay for healthcare.

From these findings, we found several correlations particularly interesting. Some of these were listed in the executive summary at the beginning of the report. Overall the most interesting finding we saw was that areas along the U.S. and Mexico border were the ones to respond the worst to the epidemic with regards to fatality rates. Furthermore, we saw a correlation with this and hispanic populations per 1000 in those areas, signifying that hispanic populations suffered worse from the virus than other demographic groups. Additionally, we saw another correlation with poverty and hispanic populations that went in hand with the fatality rates we saw. **From this finding we were able to make the understanding that hispanic populations tend to be impoverished and especially so along the Mexico border making their response to the virus worse than other areas in Texas**. Another interesting, but somewhat expected, finding we saw was that in areas with a higher median age, we saw a higher fatality rate. **This was interesting because, as one might suspect, it implies that older people are more at risk of fatally contracting the virus. This finding helps emphasize that older populations should be the first to receive the vaccine along with impoverished and/or hispanic people.** Moving on, another finding we saw was that areas with the most commuters by public transportation were not actually the most dangerous with regards to fatality rates. **This was interesting as one could assume that public transportation would make the spread far easier and result in worse fatality rates, however, that was not the case**. Lastly, one more interesting finding we saw was that in areas with higher educated people, we saw a lower poverty and fatality rate, and in areas with lower educated people, we saw a higher poverty and fatality rate. **This was interesting to**

us as it allowed us to make the understanding that educated people likely have better jobs and have access to better health and medical care if need be. It also may imply that educated people with higher performing jobs had more flexibility in how they worked and would not have to necessarily work face-to-face like with lower income jobs (store clerk, grocery bagger, etc.). Overall, we observed these four key findings to be the most interesting and help us understand how a response should be curated towards the virus with regards to assistance and availability to the vaccine.

V. Conclusion

County based census data is a good potential indicator for the spread and containment of COVID-19 within the respective county. In this report, we provide an in-depth analysis on various demographic and socioeconomic relations between counties in Texas. We utilized K-Means and Hierarchical clustering methods in order to produce such analysis and cross referenced our results with county based COVID-19 data to determine how different demographic and socioeconomic groupings handled the virus. The following conclusions and recommendations can be made after analyzing the data:

- To combat further COVID-19 deaths in Texas, we must make sure counties with higher poverty rates and limited access to resources, like healthcare, are receiving vaccine doses.
- Areas with low education levels are more likely to suffer fatally from the virus due to lower income levels and limited access to healthcare. Therefore, these areas must have high priority in receiving doses of the COVID-19 vaccine.
- Areas near the border were more negatively affected by COVID-19 with respect to fatality rates compared to other areas in Texas. Providing relief to those areas first should be a priority.
- Areas with a higher median age are also more likely to suffer fatally from the virus and thus they should also be a priority when distributing the vaccines and providing relief.

VI. References

[1] COVID-19 cases plus census dataset. https://smu.instructure.com/files/4270322/download?download_frd=1

[2] COVID-19 cases TX dataset. https://smu.instructure.com/files/4270321/download?download_frd=1