Project 2

Henry Lambson, Jonas Moros, Blake Miller

CS 7331

3/31/2023

Introduction and Data Preparation

Abstract

The primary objective of this research paper is to investigate the influence of race and socio-economic status on the transmission and severity of COVID-19 through the application of clustering analysis. Our study will be centered on the state of California, which offers a unique case due to its varied population density across different regions. To accomplish this goal, we will conduct a clustering analysis using U.S. Census data and COVID-19 statistics. Specifically, we will employ both k-means clustering and hierarchical clustering to identify patterns and relationships between the variables of interest.

Business Understanding

The business understanding for this research project is to gain insight into the impact of race and socio-economic status on the transmission and severity of COVID-19 in California. By analyzing U.S. Census data and COVID-19 statistics through clustering analysis, we aim to identify patterns and relationships between these variables. This information can be useful for businesses operating in California, as it can provide them with a better understanding of which demographic groups may be most affected by the pandemic. This information can be used to tailor business strategies and operations to better serve the needs of these groups, while also potentially mitigating the spread of COVID-19. Additionally, this research can help inform public health policies and interventions aimed at reducing disparities in COVID-19 outcomes.

Describe which features you want to use for clustering and why?

Our team is currently focused on analyzing the specific features and demographics of California to gain a deeper understanding of the impact of COVID-19 on various groups within the state. In particular, we have identified two main groups of features for analysis. The first group centers around race, and includes variables such as ethnicity, gender, and median age. By examining these variables, we aim to explore the varying impacts of COVID-19 on different demographic groups, including identifying which genders and age groups were most affected.

The second group of features that we are analyzing revolves around income demographics, which includes variables such as median income, income per capita, median rent, poverty, and commuters by public transportation. Alongside these variables, we are also incorporating gender demographics to further understand the impact on different groups within the state. By utilizing these features, we hope to gain insight into the ways in which income and transportation factors have influenced the impact of COVID-19 on various communities in California. Table 1 will show the features that will be used.

Table 1: Features

Feature	Data Type	Description
county_name	Nominal	Name of the county
confirmed_cases	Ratio	Number of confirmed cases in each county
deaths	Ratio	Number of deaths in each county
median_age	Ratio	Median age of the county
cases_per_1000	Ratio	Number of cases per 1000 people in each county
deaths_per_1000	Ratio	Number of deaths per 1000 people in each county
death_per_case	Ratio	Deaths per case in each county
black_per_1000	Ratio	Number of black people per 1000 people in each county
white_per_1000	Ratio	Number of white people per 1000 people in each county
asian_per_1000	Ratio	Number of asian people per 1000 people in each county
hispanic_per_1000	Ratio	Number of hispanic people

		per 1000 people in each county
amerindian_per_1000	Ratio	Number of American Indian people per 1000 people in each county
median_income	Ratio	Median income for each county
income_per_capita	Ratio	Income per capita for each county
median_rent	Ratio	Median rent for each county
poverty	Ratio	Number of people in poverty in each county
commuters_by_public_transportation	Ratio	Number of people who commute via public transportation in each county

Outliers

There is only one outlier county in the California dataset, Los Angeles County. We initially thought that we should leave this county in the dataset, since it accounts for approximately 25% of the population of California, but after clustering with it in the dataset, we found that it either heavily skewed a cluster containing it and other counties, or took up an entire cluster on its own. Because of this, we ended up deciding to remove Los Angeles County from the dataset, and saw significant improvements in our clustering.

What is the scale of measurement of the features and what are appropriate distance measures?

We have chosen to use the following clustering methods: K-Means and Hierarchical. For each clustering method, we ran both of our feature data subsets through each method. All the features that we are clustering have a ratio scale of measurement. Since our feature scale of measurements are all ratios, we are using Euclidean distance measurement.

Modeling and Cluster Analysis

Model Types:

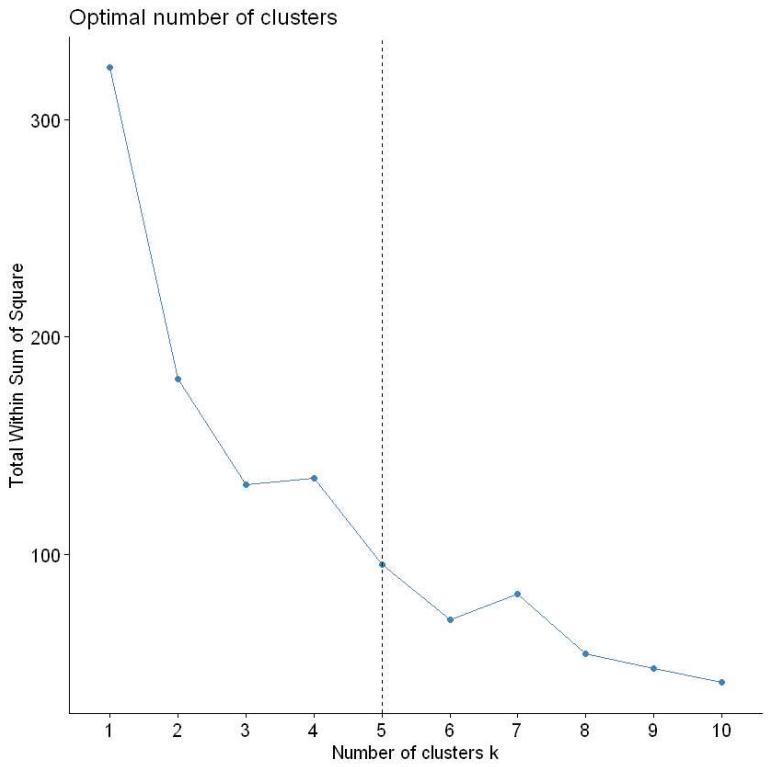
Two K-Means Clustering on Two Different Feature Data Sets

Two Hierarchical Clusters on Two Different Feature Data Sets

Determining Number of Clusters

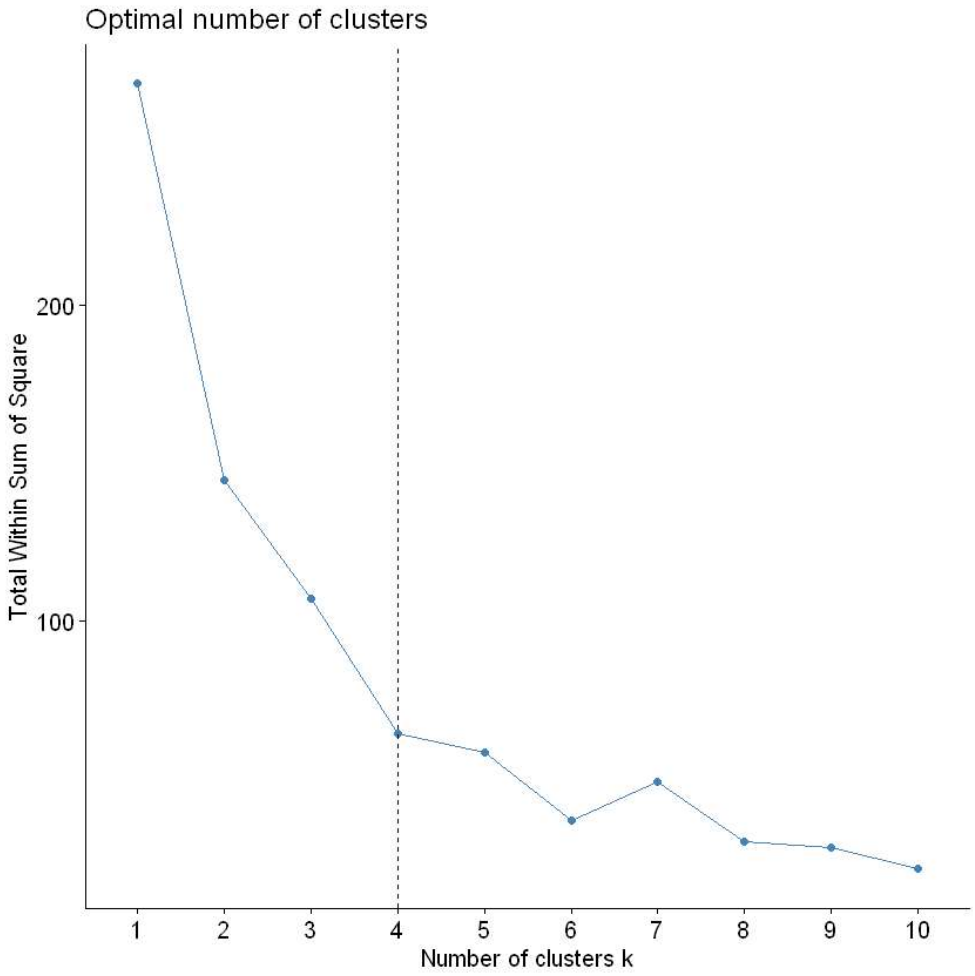
In order to conduct the cluster analysis of the feature subsets, we first need to determine how many clusters to use for each subset. To this end, we use the fviz_nbclust() function in order to take into account the within cluster sum of squares, average silhouettes, and gap statistics. Figure 1 will show the plot for the racial demographics subset, and Figure 2 will show the plot for the socio-economic subset.

Figure 1: Optimal Number of Clusters Racial



In this graph, we are looking for the inflection point of the curve in order to determine the number of clusters that is optimal for this dataset. From this graph, we determined that the inflection point, and optimal number of clusters is 5. For both k-means and hierarchical clustering on the racial demographics subset, we will be using 5 clusters.

Figure 2: Optimal Number of Clusters Socio-Economic

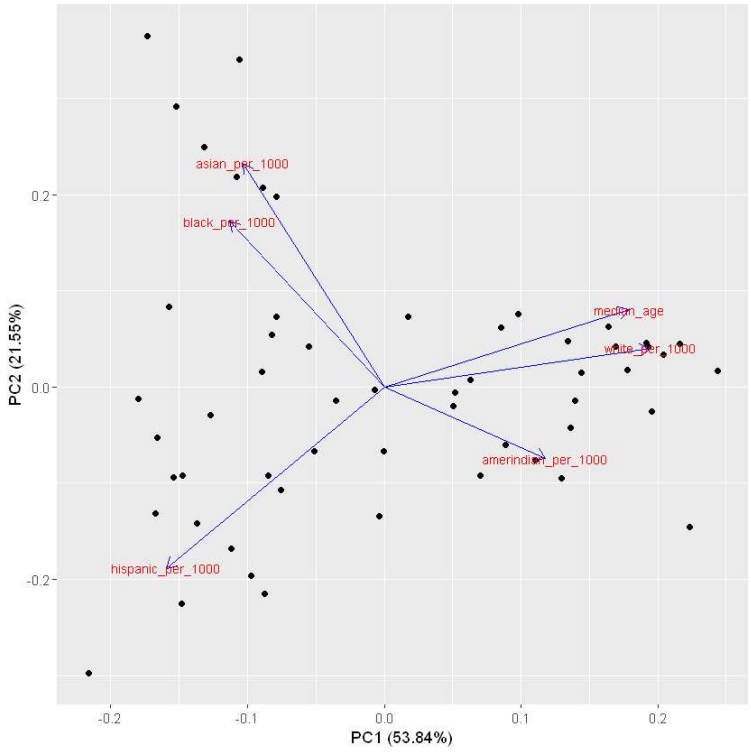


Looking at this plot, we can see that the inflection point of the curve is at 4 clusters. For k-means and hierarchical clustering on the socio-economic subset, we will be using 4 clusters.

PCA Analysis

To get a deeper look into how each of the variables in our dataset might affect our clusterings, we performed principal component analysis (PCA) on each subset of data. PCA identifies the components in a dataset which explain the most variation in the data, called principal components. Through looking at these principal components and which variables contribute to the variation in the dataset, we can see how variables might be correlated and how they might affect the clusterings created.

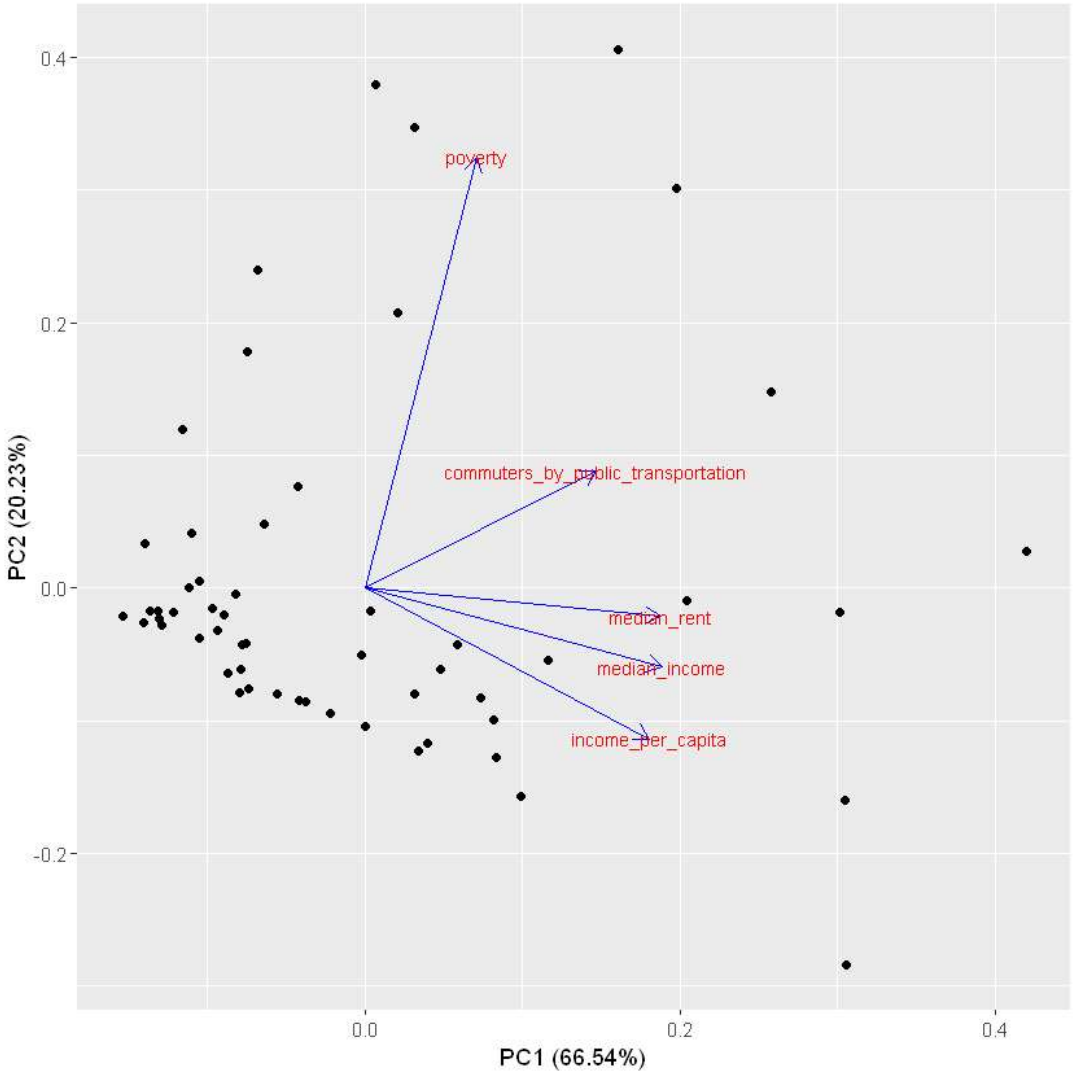
Figure 3: PCA Analysis on Race/Age Variables



As we can see from the graph above, the first principal component (PC1) explains 53.84% of the variance in the dataset. All variables influence PC1 some, but the median age, white population per 1000, and hispanic population per 1000 influence PC1 the most. PC2 explains 21.55% of the variance in the dataset, and is most influenced by black population per 1000, asian population per 1000, and hispanic population per 1000. We can also see from the plot that there are some positive and negative correlations. The vectors representing median age and white population per 1000 are very close, meaning they are positively correlated with one another, and the vector representing hispanic population per 1000 is at a very large angle

with both of them, meaning they are negatively correlated [2]. We also can see a positive correlation with the Asian and black populations per 1000, and they have a negative correlation with the American Indian population.

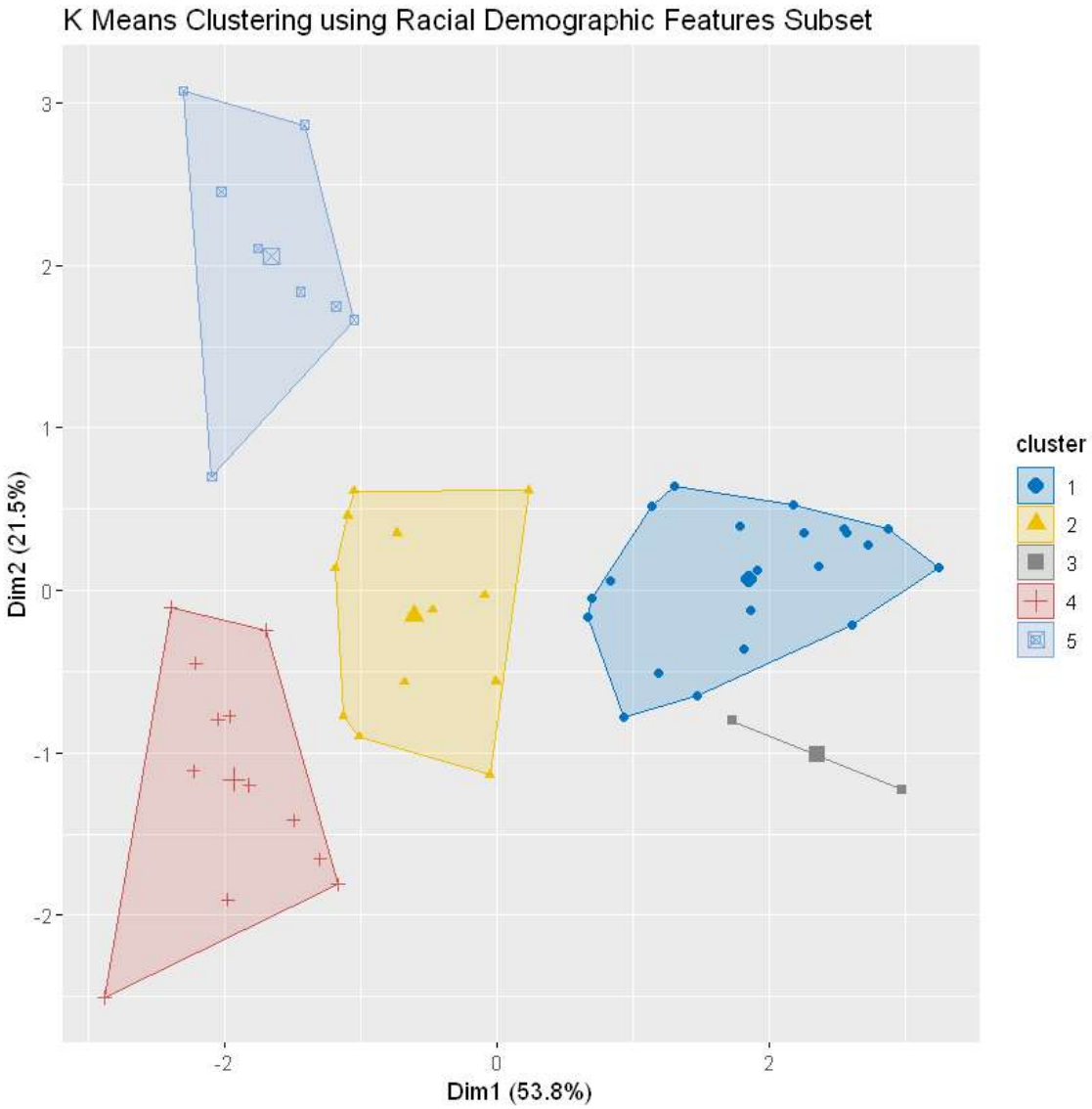
Figure 4: PCA Analysis on Economic Variables



As we can see from the graph above, the first principal component (PC1) explains 66.54% of the variance in the dataset. The median rent, median income, and income per capita are the variables which influence PC1 the most. PC2 explains 20.23% of the variance in the dataset and is almost solely influenced by poverty. We can also see from the plot that there are some positive correlations between median rent, median income, and income per capita, which makes sense as they are all related variables.

K Means Clustering Analysis (Euclidean distance)

Figure 5: K-Means Racial Clustering



Can you use a feature as the ground truth and perform external validation?

For our external validation, we looked at the number of cases per 1000 and deaths per 1000 for each cluster. This will give us insight into whether our clusterings have any relevance to the number of cases and deaths, which could help lawmakers and local governments figure out what groups are most affected by COVID-19.

Figure 6: K-Means Racial Silhouette Plot

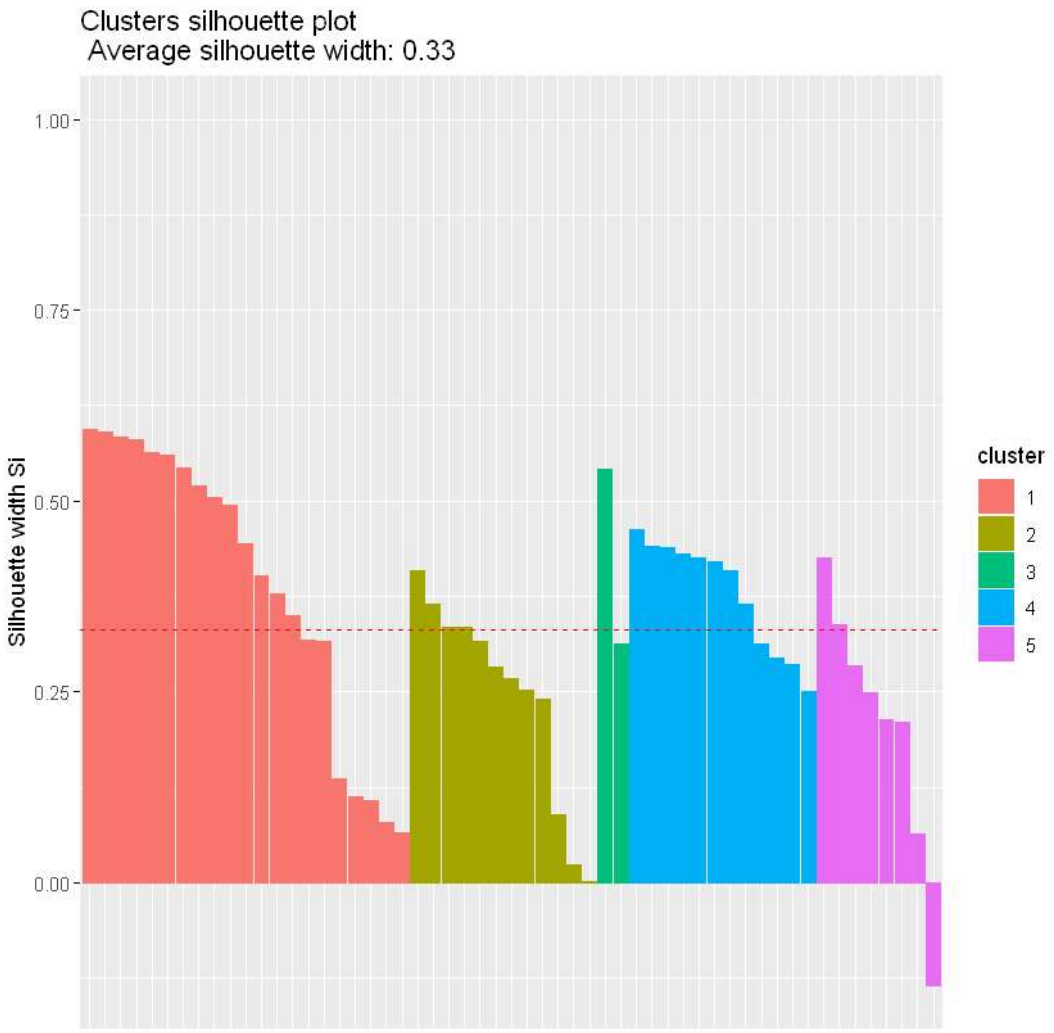


Table 2: K-means Racial Ground Truth

Ground Truth:

Clusters	Deaths Per 1000	Cases Per 1000
1	0.4557677	42.65174
2	0.6135756	71.06307
3	0.8333627	39.26763
4	0.9750980	97.95994
5	0.5146325	50.64852

We can see in Figure 6 that using K means clustering on our racial demographic subset with the optimal number of clusters found from Figure 1 gives us good results. The clusters visually look separated and all seem to have a similar number of points, other than cluster 3 which only contains 2 points. As an internal cluster validation method, we used silhouette width. We can see above that our average silhouette width is 0.33, which indicates a weak structure, but a structure nonetheless [3].

Figure 7: K-Means Economic Clustering

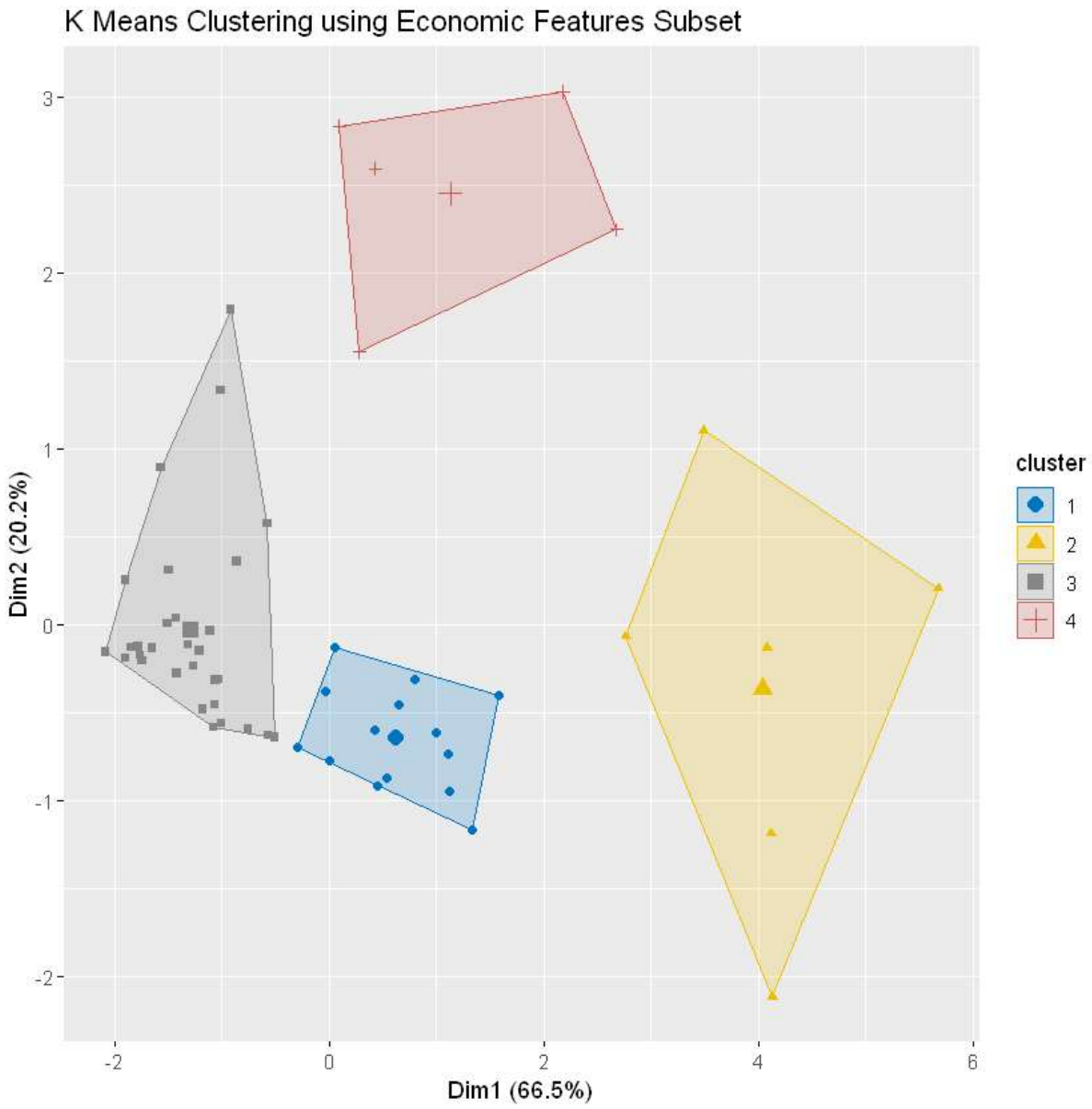


Figure 8: K-means Economic Silhouette Plot

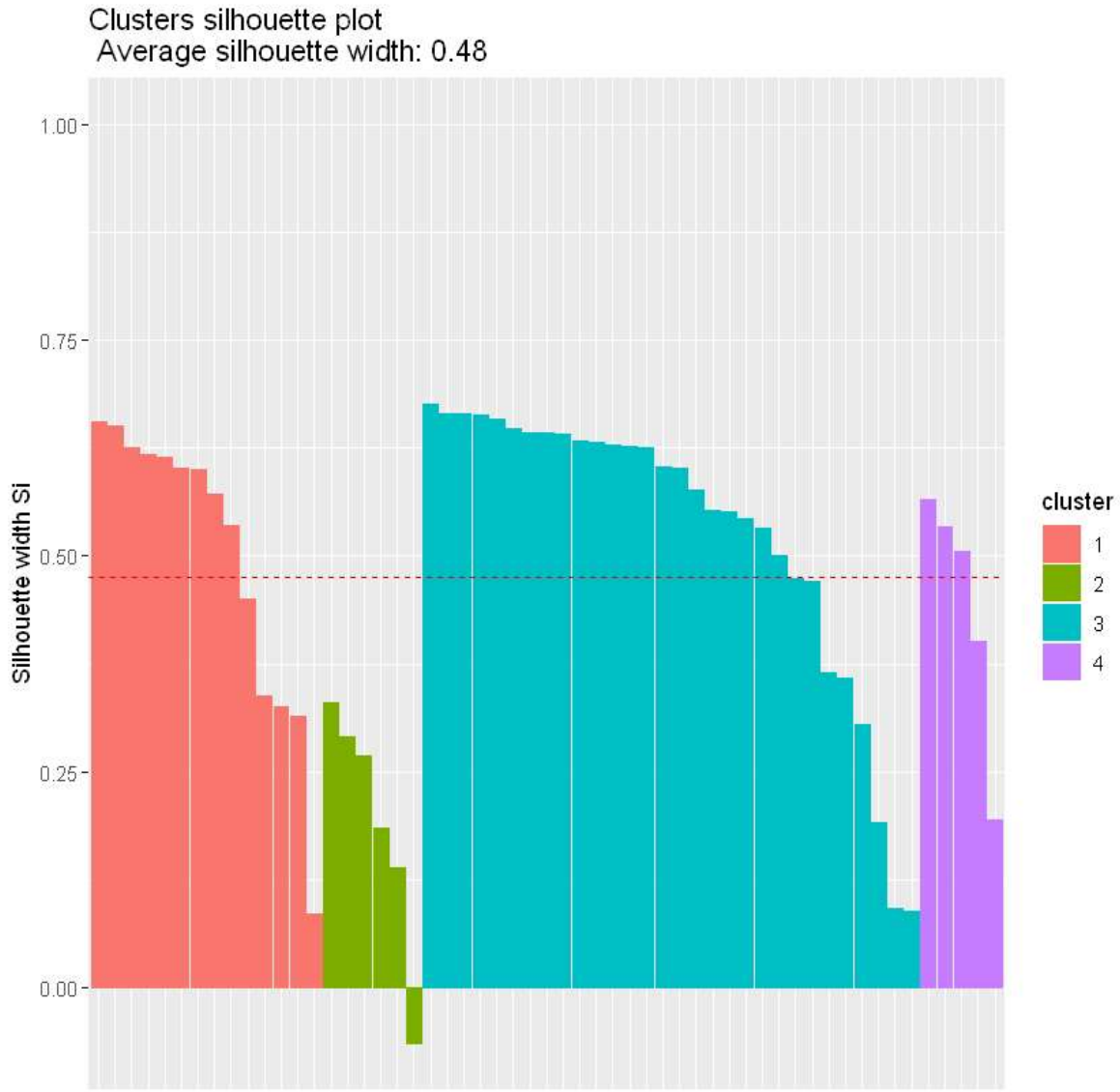


Table 3: K-Means Economic Ground Truth

Ground Truth:

Cluster	**Deaths Per 1000**	**Cases Per 1000**
1	0.4668890	57.49874
2	0.4518307	43.04868
3	0.7084882	64.48844
4	0.7833832	81.95166

Using K means on our economic features subset with the optimal number of clusters from Figure 2, we get the scatter plot shown in Figure 7. Each cluster seems to be separated well, but there is an uneven number of points in each clustering. Cluster 1 and 3 have many dense points, while clusters 2 and 4 have less points that are more spread out. The average silhouette width of 0.48 confirms that this clustering has a reasonable structuring.

Figure 9: Hierarchical Clustering Racial

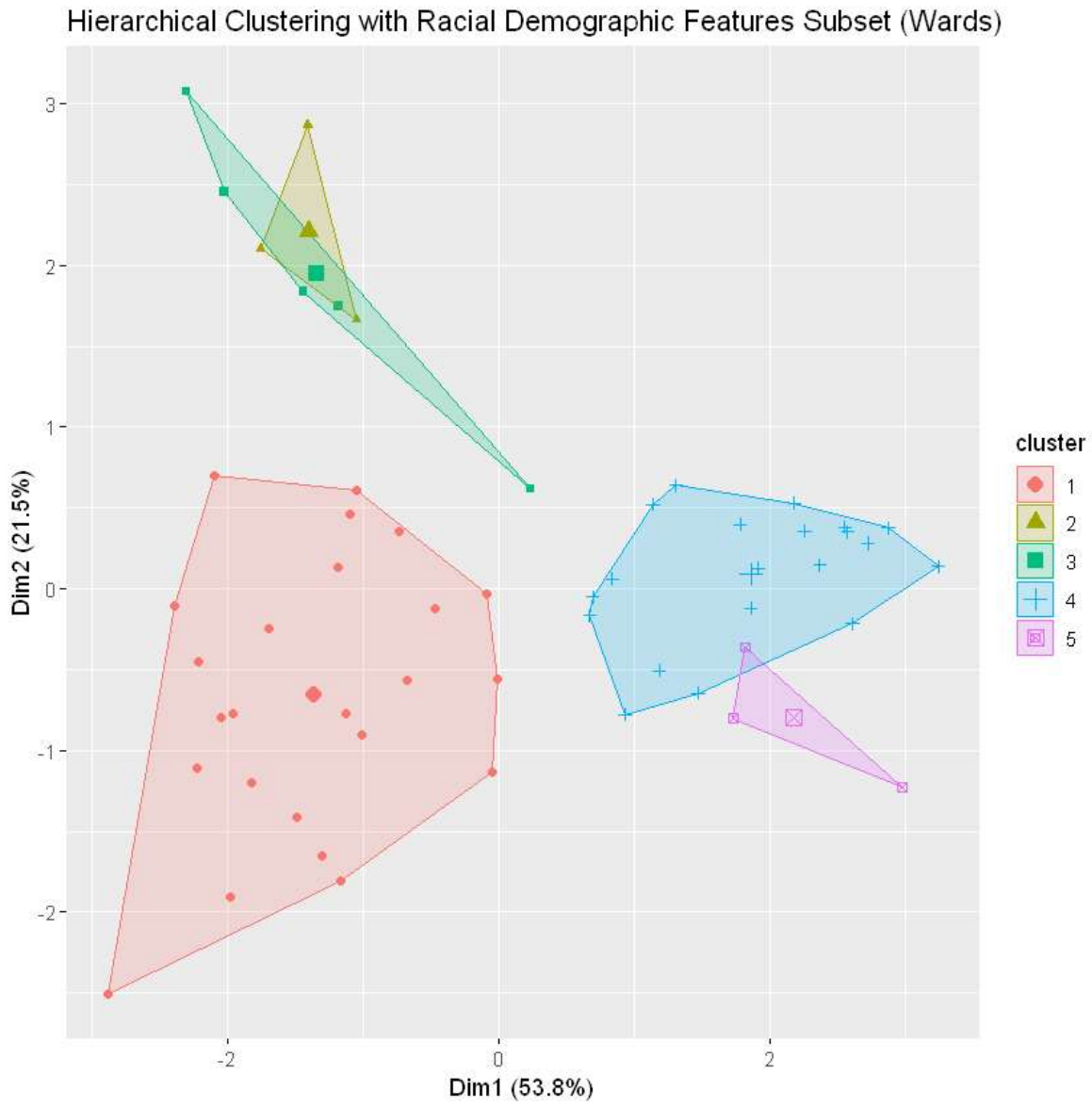


Figure 10: Hierarchical Racial Silhouette Plot

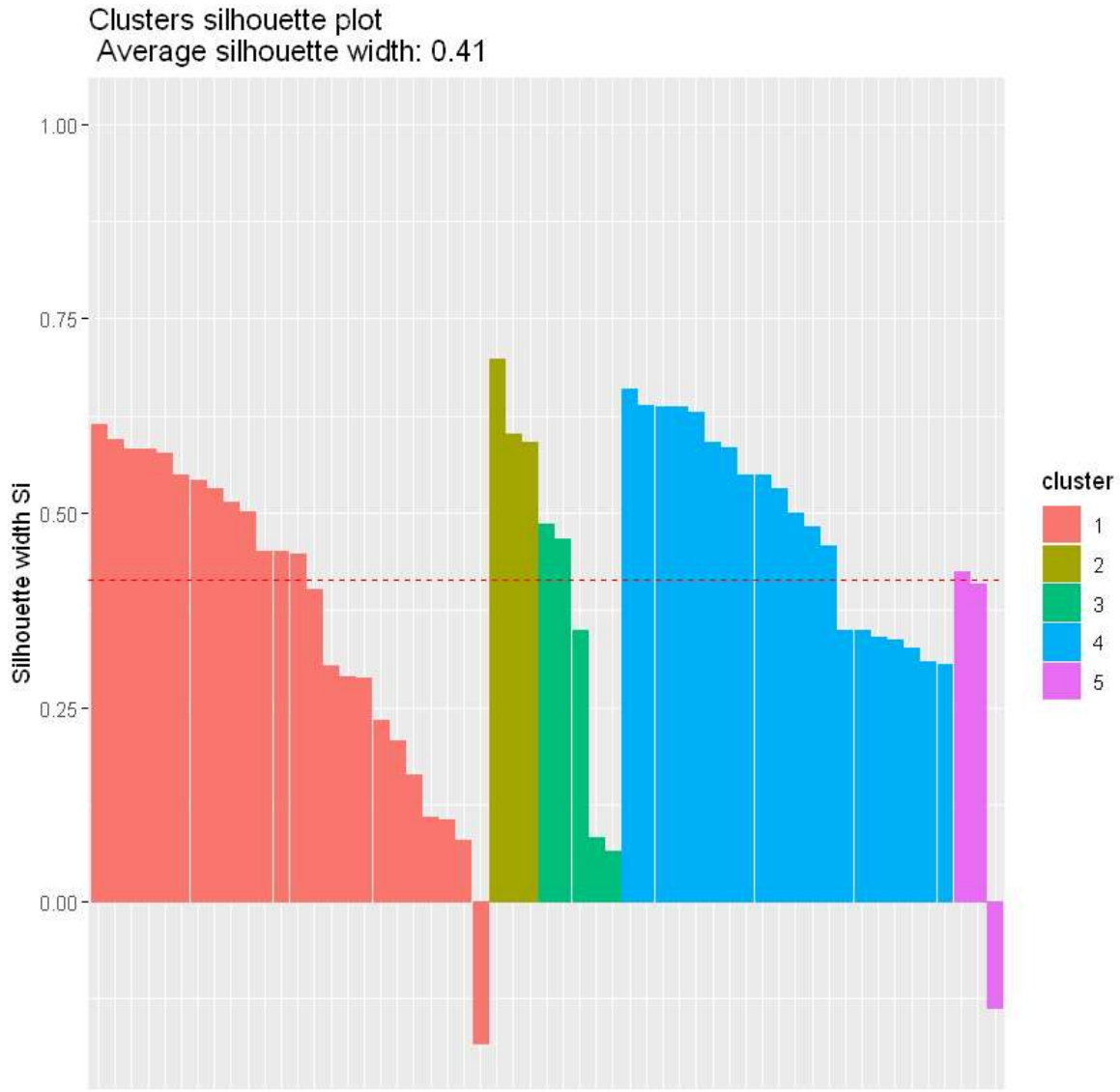


Table 4: Hierarchical Racial Ground Truth

Ground Truth:

Clusters	Deaths Per 1000	Cases Per 1000
1	0.8201635	81.52677
2	0.4170773	41.84514

3	0.4491979	70.25728
4	0.4696993	43.91968
5	0.6146201	31.94269

Using hierarchical clustering on the racial demographic features with the optimal number of clusters, we get the scatter plot shown in Figure 9. The clusters are decently spread out, but there is a bit of overlap in the clusters. Clusters 2 and 3 overlap quite a bit and are both centered very close together. Also, Cluster 5 has one point which is inside the bounds of cluster 4. The average silhouette width is 0.41, indicating a reasonable structure.

Figure 11: Hierarchical Clustering Economic

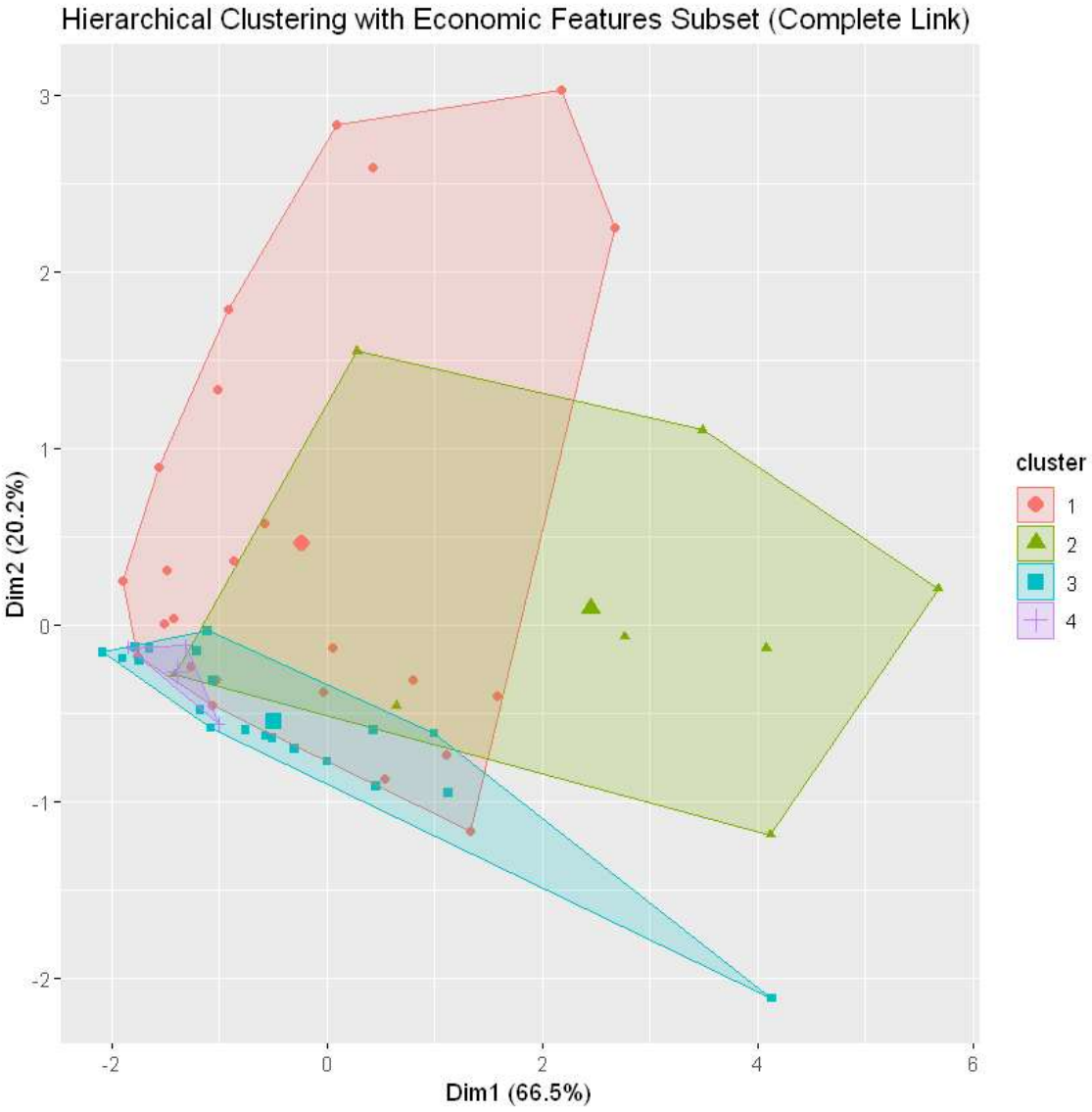


Figure 12: Hierarchical Economic Silhouette Plot

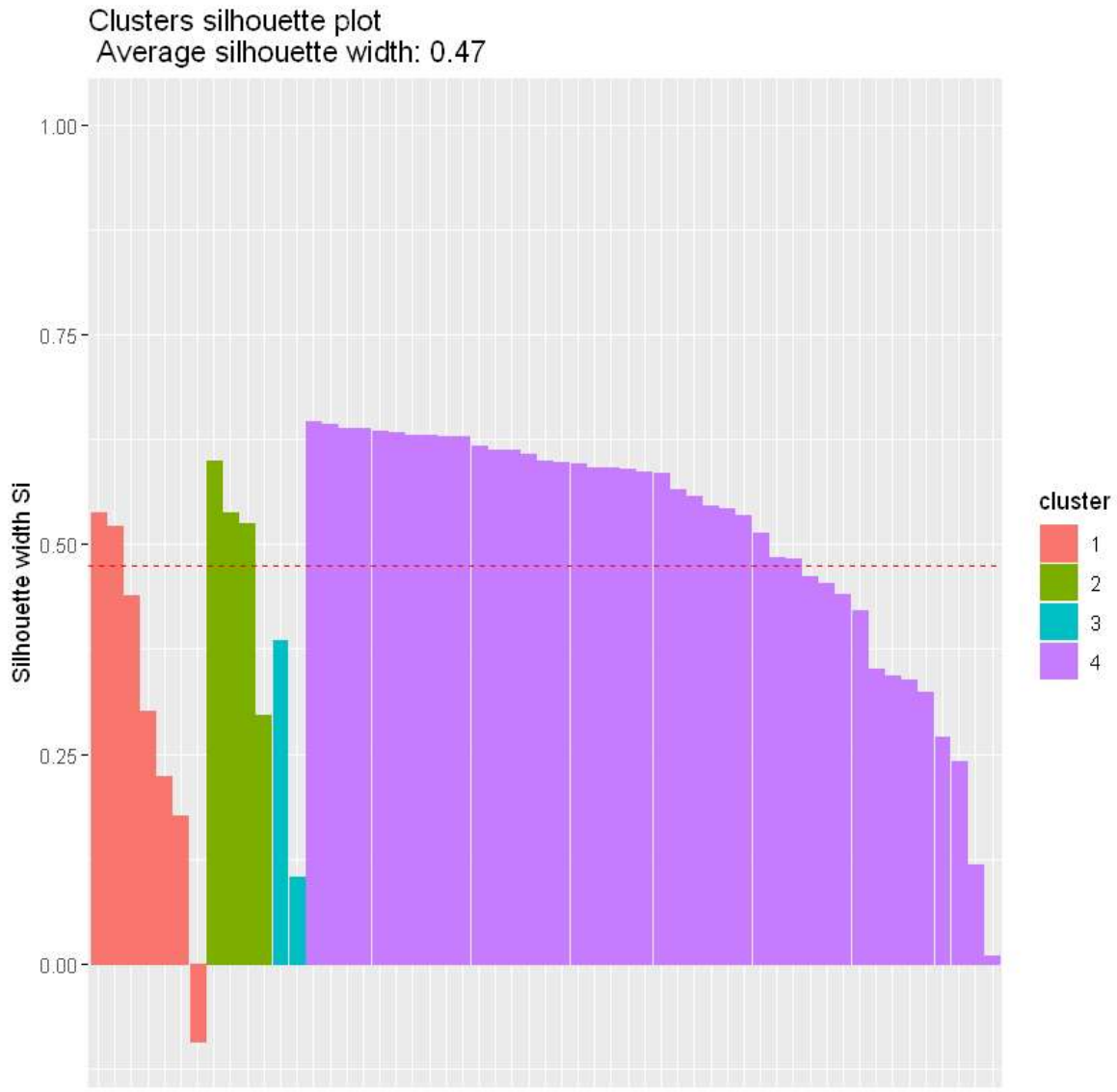


Table 5: Hierarchical Economic Ground Truth

Ground Truth:

Clusters	Deaths Per 1000	Cases Per 1000

1	0.7884437	83.65890
2	0.4847533	45.66911
3	0.3859854	37.80781
4	0.6235453	61.04242

Using hierarchical clustering on the economic subset of features with the optimal number of clusters, we get the scatter plot shown in Figure 11. We can see that the clustering looks a bit chaotic, every cluster has an overlap and some have multiple overlaps. While the average silhouette width of 0.47 indicates that a reasonable structure has been found, the visual clustering does not indicate a very strong structure at all.

Cluster Visualization With Heat Maps

Figures 13,14,15 and 16 will visualize the clusters on heatmaps of California. Figures 13 and 14 will be k-means and hierarchical clustering of the racial subset, while Figures 15 and 16 will be k-means and hierarchical clustering of the economic subset.

To begin labeling the clusters for each feature subset, we will analyze the means of each feature in each cluster. Table 6 will show the means for k-means clustering on the racial demographics subset, Table 7 will show the means for hierarchical clustering on the racial demographics subset, Table 8 will show the means for k-means clustering on the socio-economic subset, and Table 9 will show the means for hierarchical clustering on the socio-economic subset.

Figure 13 Heatmap: K-Means Using Demographic/Racial Feature Set

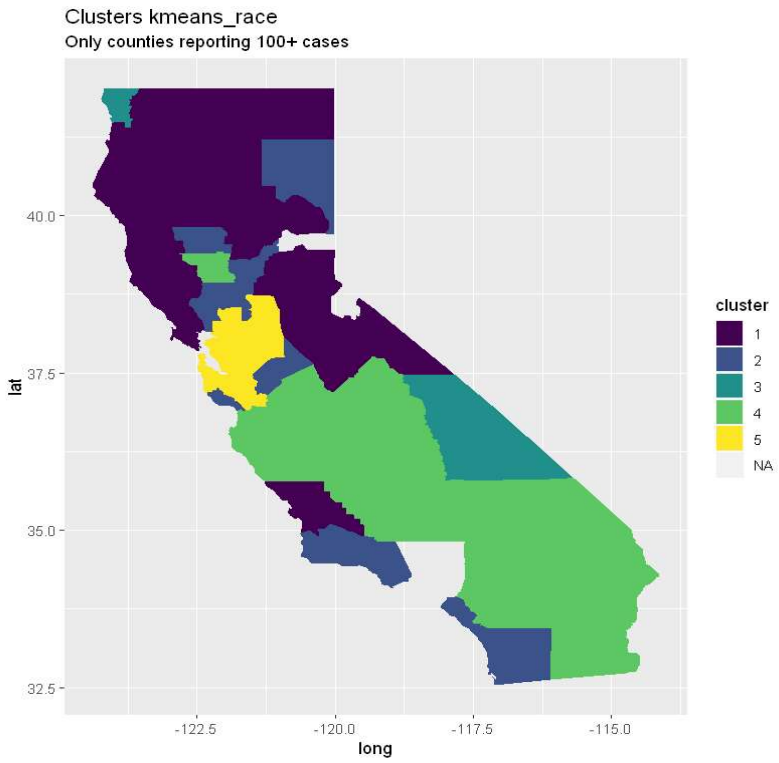


Table 6: Cluster Means K-means Racial

Cluster	Median Age	White per 1000	Black per 1000	Asian per 1000	Hispanic per 1000	American Indian per 1000
1	0.9637065	1.0841726	-0.597727	-0.573596	-0.850618	0.2628242
2	-0.5582103	-0.218155	-0.102739	0.1377239	0.2305123	-0.3306668
3	-0.4800069	0.4439970	-0.539782	-0.615301	-0.568490	3.9873135
4	-1.0165223	-1.206281	0.2334877	-0.319741	1.5413401	-0.4083589
5	-0.2876324	-0.820297	1.5078573	1.9325428	-0.282783	-0.5782034

From these means, we can begin to identify what each cluster is looking for and begin labeling them.

Cluster 1: Characterized by a high white population, a high median age and a low minority population.

Cluster 2: This cluster is racially diverse, with no significant demographic standing out. It also has a somewhat low median age.

Cluster 3: Characterized by an extremely high American Indian population in comparison to the other clusters.

Cluster 4: Characterized by a low median and a high hispanic population.

Cluster 5: Characterized by a high black and asian population, and a low white and American Indian population.

From these cluster classifications, and the heatmap of the clustering, we can see where each racial demographic is mostly populated. Beginning with cluster 1 which focuses on the white population and high median age, we can see that the majority of the very north of California belongs to this cluster. Since this area is mostly rural, it makes sense that the majority of the population is white. This cluster also contains the area closer to central California, which would include San Luis Obispo county, which contains approximately 88% white population. Cluster 2 does not have any truly defining features, but it does show to be racially diverse. This cluster includes areas such as San Diego and Santa Barbara, along with some areas in northern California. Cluster 3's most prominent feature is American Indian per 1000, which is shown on the heatmap to be the very northwest county, Del Norte County, and the county that borders southern Nevada, Inyo County. Del Norte has approximately 10% American Indian population, and Inyo has approximately 14%, numbers that are significantly higher than the other counties in California. Cluster 3 contains only these two counties because they are outliers when it comes to American Indian population. Cluster 4 shows that inland in southern California there is a high amount of hispanic people, and a low amount of white people. This group contains Fresno County and San Bernardino County, both of which have approximately 55% hispanic population. Cluster 5 shows that the San Francisco Bay Area has large Asian and black populations. San Francisco has approximately 37% Asian population, and Contra Costa County has approximately 10% black population, which is much more

than many of the other counties in California. All county demographic statistics were taken from reference 1.

Figure 14: Heatmap Hierarchical Cluster Using Demographic/Racial Feature Set

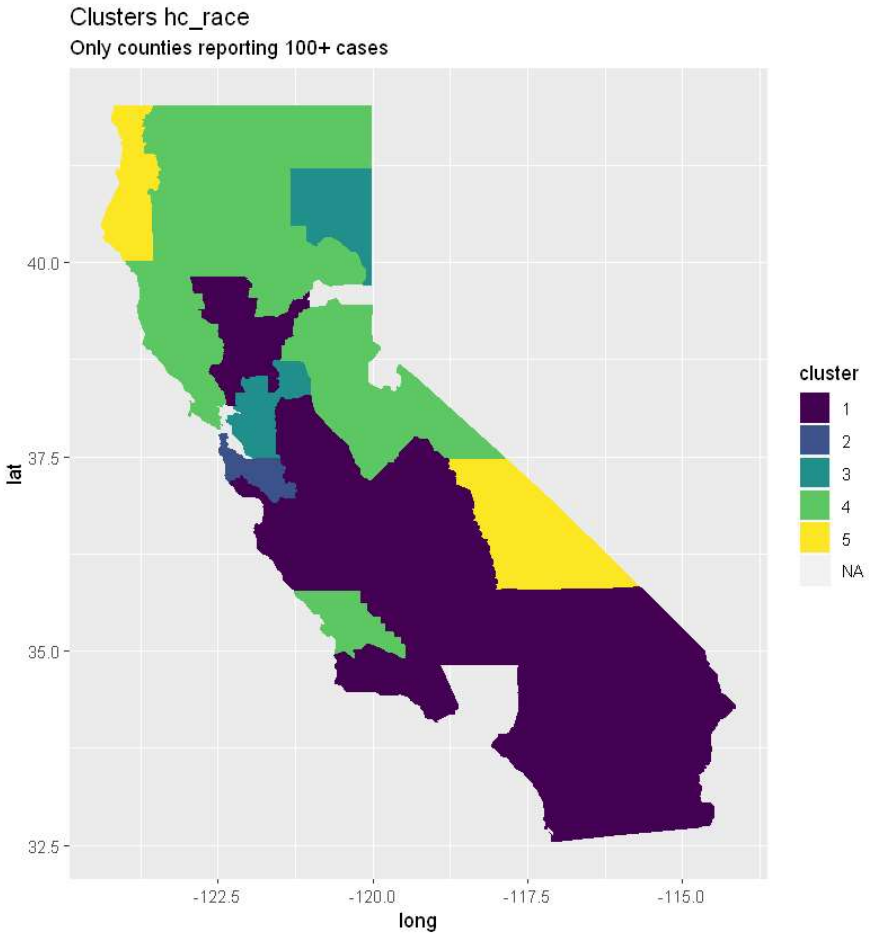


Table 7: Cluster Means for Hierarchical Racial

Cluster	Median Age	White per 1000	Black per 1000	Asian per 1000	Hispanic per 1000	American Indian per 1000
1	-0.8027325	-0.7827651	0.03754548	-0.02405339	0.9379542	-0.4277643
2	-0.1373114	-0.8596847	0.10894023	3.00230246	-0.4702810	-0.6205102

3	-0.3040675	-0.4580406	2.48078608	0.96930131	-0.4200199	-0.2732124
4	1.0235677	1.0860701	-0.59499155	-0.57629554	-0.8387910	0.1912956
5	0.2421656	0.6447388	-0.57733718	-0.58400720	-0.7413793	3.2226748

Cluster 1: Characterized by a large hispanic population and a low median age.

Cluster 2: Characterized by a low white population and a high asian population

Cluster 3: Characterized by a high black population, and a slightly high asian population

Cluster 4: Characterized by a high median age, a high white population and a low hispanic population

Cluster 5: Characterized by a high American Indian population and a slightly high white population, with low hispanic, asian and black populations.

Now to look at the hierarchical clustering for the racial demographics subset, starting with cluster 1, we can see similarities to cluster 4 of the k-means clustering. This cluster consists of counties with a high density of hispanics, and low median age. The areas of the map that this cluster covers is very similar to cluster 4 of the k-means clustering, with the exception that the hierarchical cluster contains the entirety of Southern California, including San Diego County and Orange County. This cluster also extends slightly further north than the k-means cluster. Cluster 2 is characterized by a high Asian population, which shows similarities to cluster 5 of the k-means clustering. For this cluster, the East Bay was not included as it was in the k-means clustering, which explains why this cluster is not looking for a high black population. Cluster 3 in the hierarchical clustering does contain the East Bay counties, which explains why it is characterized by a high black population. These counties contain cities such as Oakland and San Leandro, both of which are high in black population, Oakland especially so. Cluster 4 shows to be high in white population and median age, and covers very similar counties to cluster 1 in the k-means clustering. It mainly spans northern California, and includes San Luis Obispo county, just as cluster 1 of the k-means clustering did. Cluster 5 is looking for a high density of American Indians, similar to cluster 3 of k-means. The main difference between the two is that this cluster contains Humboldt County along with

Del Norte and Inyo. Humboldt county has 6.7% American Indian population, which is lower than Del Norte and Inyo, but is still higher than the other counties in California.

Figure 15: Heatmap K-Means Cluster Using Economic Feature Set

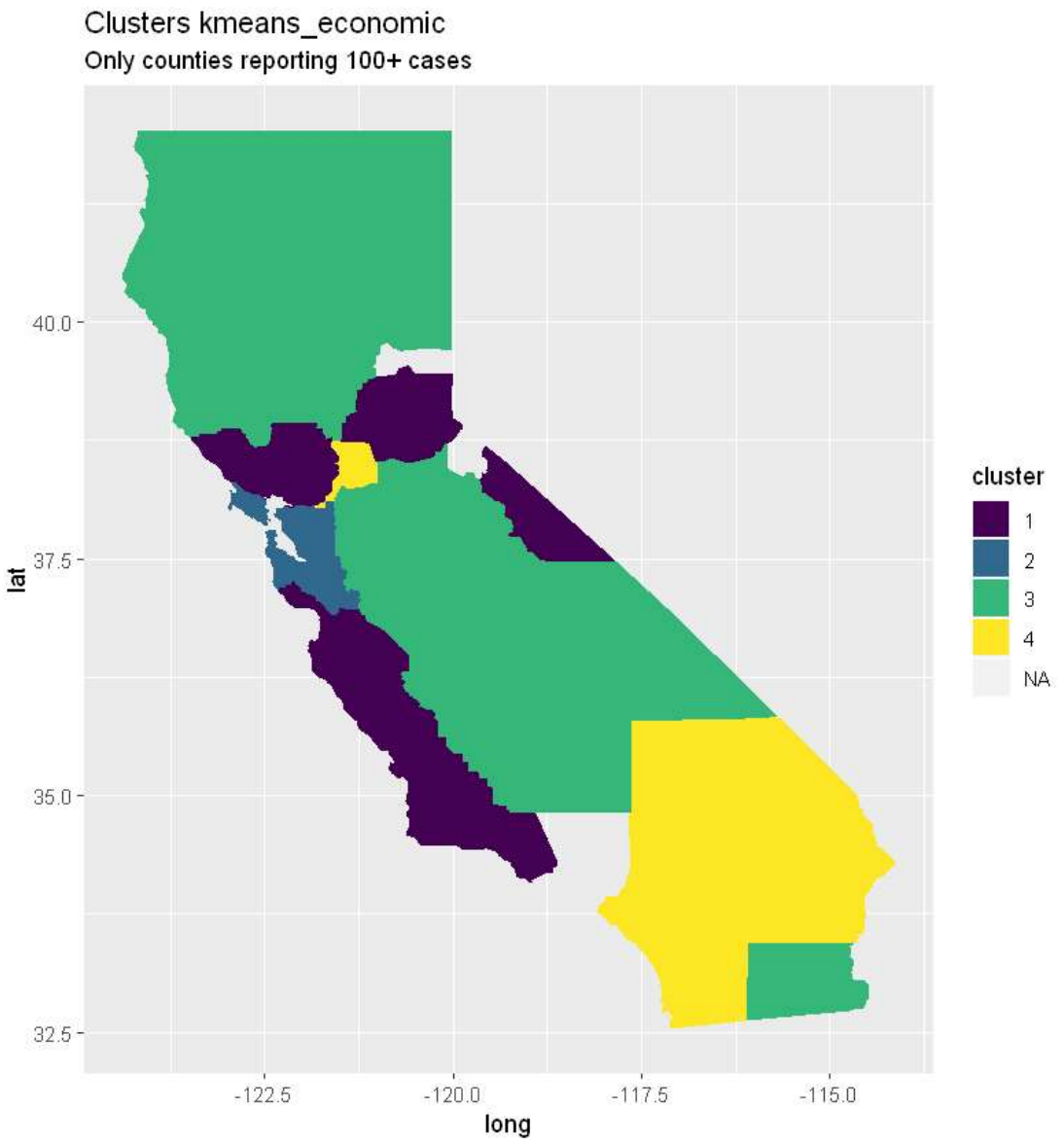


Table 8: Cluster Means for K-means Economic

Cluster	Median Income	Income per capita	Median Rent	Poverty	Commuters by Public

					Transportation
1	0.5585589	0.39330153	0.6137536	-0.3576054	-0.2873322
2	2.0595064	2.15721589	1.8794068	0.2841103	2.0959933
3	-0.7177187	-0.60840589	-0.7554339	-0.3339846	-0.3647994
4	0.2709399	-0.03946803	0.5588049	2.6642703	0.4781349

Cluster 1: Characterized by a low amount of poverty and high values for median income, income per capita and median rent.

Cluster 2: Characterized by very high median income, income per capita, median rent and commuters by public transportation

Cluster 3: Characterized by low median income, income per capita and median rent

Cluster 4: Characterized by a very high amount of poverty.

Cluster 1 shows low values for the amount of poverty, and high values for median income, income per capita, and median rent. This cluster mainly covers the coastline of California, along with some counties that border Nevada. The counties on the coastline are broken up by the Bay Area, which is one of the most expensive places to live in the world. Some of the notable counties included in this cluster are Santa Cruz, Monterrey, San Luis Obispo, and Santa Barbara. These four counties all have a median household income of over $82,000, with Santa Cruz having $96,000. Cluster 2 shows very high values for median income, income per capita, and median rent. This cluster is exclusively the Bay Area, and includes San Francisco County, San Mateo county, and Marin County. The median rent in San Mateo County is approximately $2,600, and all three of these counties have median household incomes of over $125,000. These three counties are by far the most expensive place to live in California, with San Mateo county having the highest rent and highest median income at $136,000. Cluster 3 shows low median income, income per capita and median rent. This cluster spans the rural areas of California, including the very north of the state, and the valley area in the middle of the state. All of the counties in these areas have median household incomes ranging from $50,000 to $65,000, and median rent at around $1000. These counties do not

contain any major cities, and therefore, their rent and income are low. Cluster 4 shows very high amounts of poverty in comparison to the other clusters. This cluster is mainly southern California, with the exclusion of Imperial county (bottom right), and also contains Sacramento county in the north. When looking at the statistics for poverty in these areas, Kern county has approximately 19% of its population in poverty. This number is not an extreme amount greater than counties not in this cluster as the cluster means would suggest, but it is still higher.

Figure 16: Heatmap Hierarchical Cluster Using Economic Feature Set

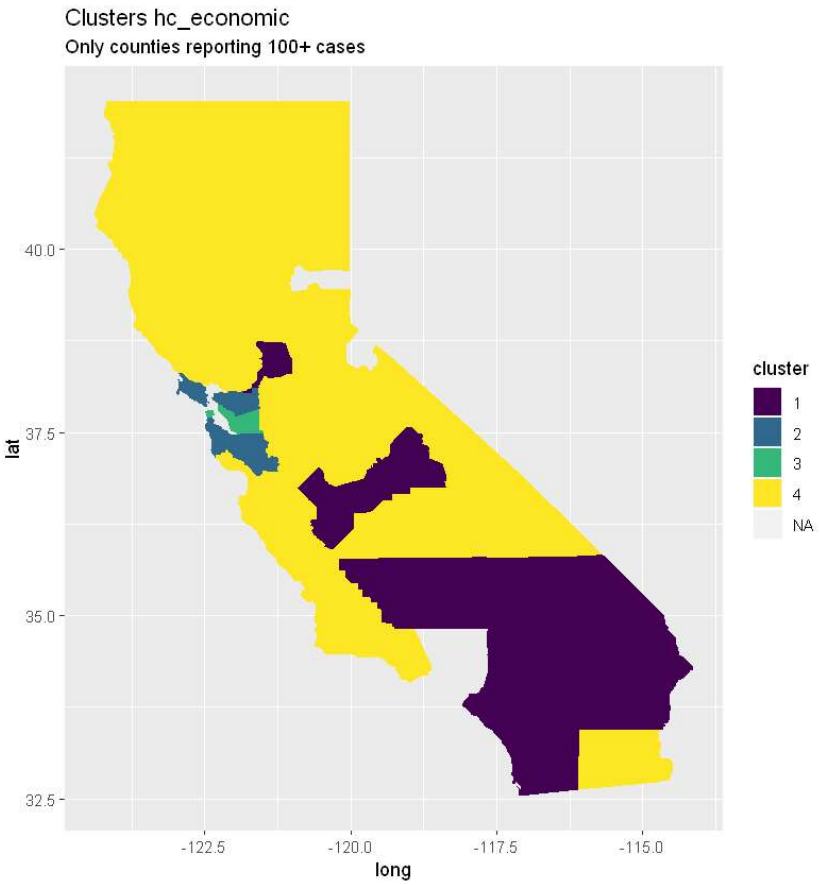


Table 9: Cluster Means for Hierarchical Economic

Cluster	Median Income	Income per Capita	Median Rent	Poverty	Commuters by Public Transportation

1	0.008231289	-0.2688131	0.1998063	2.2865591	0.2675234
2	2.254324085	2.2410344	2.0738644	0.1165623	0.9104101
3	1.669870905	1.9895788	1.4904917	0.6192062	4.4671598
4	-0.295587076	-0.2633715	-0.3017877	-0.4216804	-0.3440148

Cluster 1: Characterized by high poverty

Cluster 2: Characterized by very high median income, income per capita, and median rent

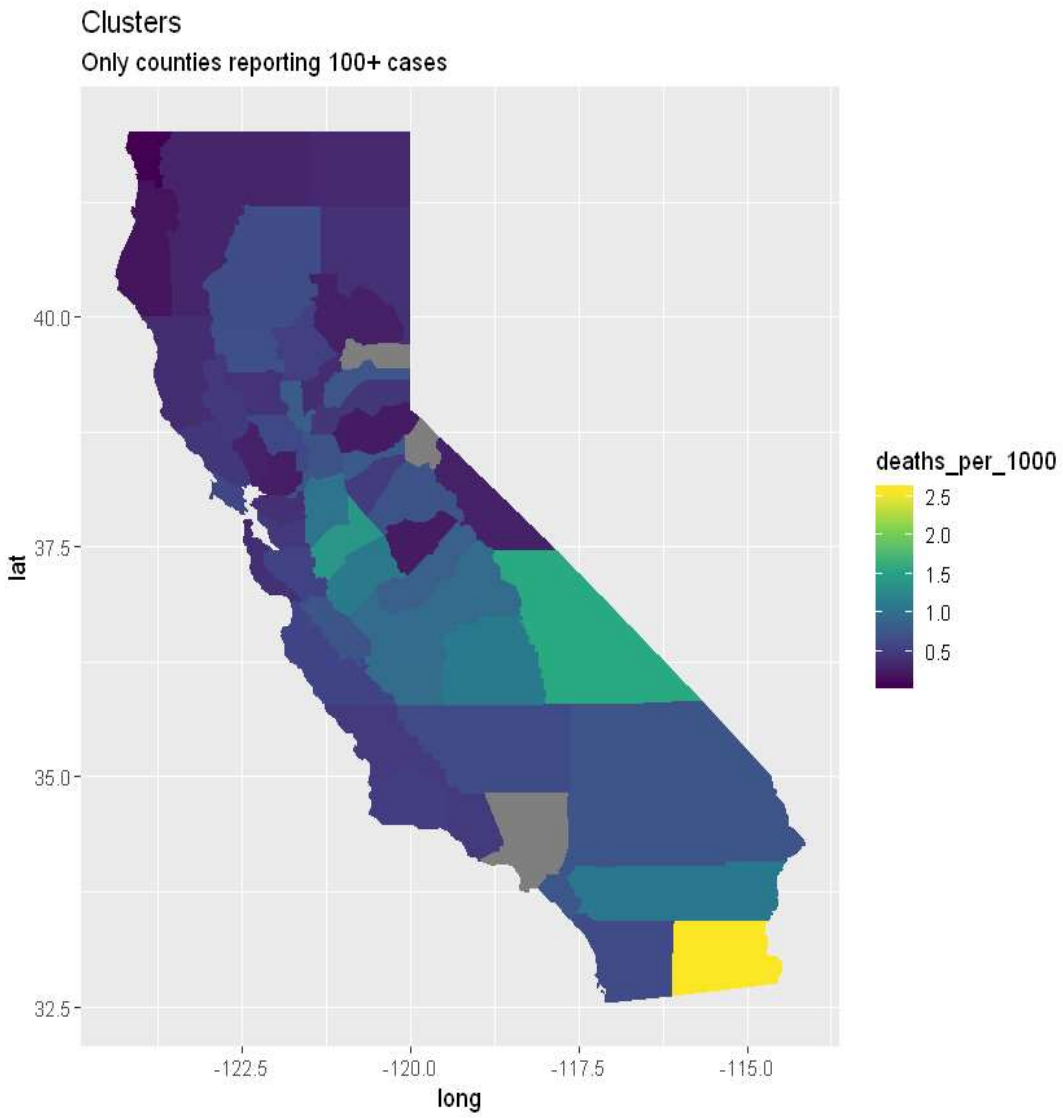
Cluster 3: Characterized by high median income, income per capita, median rent, and very high commuters by public transportation

Cluster 4: Characterized by low poverty, and relatively low of the other features.

Comparing the hierarchical clustering to the k-means clustering, we can see that cluster 1 contains the very high amount of poverty that cluster 4 of the k-means clustering did. The main difference between the two clusterings is that the hierarchical cluster also includes Fresno county in central California. Fresno has a poverty rate of 19.4%, which is above that of Kern county. This shows that in the k-means clustering, the low median income, income per capita, and rent of Fresno county took precedence over its high poverty rate, which caused it to not end up in the cluster looking for high poverty. Cluster 2 is very similar to cluster 2 of k-means clustering, with the exception that the hierarchical clustering does not contain the East Bay, and does not contain San Francisco. The exclusion of the East Bay makes sense, as it has a lower median income, income per capita and rent than the counties that are on the Peninsula. The exclusion of San Francisco from this cluster is interesting, as it does have lower median income, income per capita, and rent than San Mateo County and Marin County, but not by too big of a margin. Cluster 3 is different from all of the clusters in the k-means clustering, as its primary feature is commuters by public transportation. This explains why San Francisco was excluded from cluster 2, as it has a very high amount of commuters using public transportation. The other counties in this cluster are in the East Bay, which also have cities that are high in public transportation, such as Oakland. Cluster 4 is shown to have low values of every feature in this subset, and covers all of the other counties in California. This is by far the largest cluster, which makes it difficult to identify individual counties that show off what this cluster is looking for. There is not

much in common between the coastal counties and the northern counties, as this cluster is trying to suggest.

Figure 17: County Heat Map Showing Deaths per 1000 People



To compare our clustering findings to the amount of deaths in California, we will be analyzing this heatmap. The clear outlier in this map is Imperial County in the south east, with just over 2.5 deaths per 1000 people. As shown in the racial demographics clustering, Imperial County has a high population density of hispanic people, and a low median age. From the census data in Reference 1, we can see that Imperial County has a population of approximately 86% hispanic, by far the

most in the entire state. This is unsurprising, considering it is on the border of Mexico. However, its neighbor county, San Dieno, only has a hispanic population of approximately 36%, a stark difference. To add onto this, Imperial county's population is approximately 30% foreign born, which means that a large population of the people who live there are immigrants from Mexico, illegal and legal. These two statistics are the main outlier of Imperial county, as it shares economic similarities to other counties, as shown by the socio-economic clusterings, with the exception of a median income of about $49,000, which is lower than most other counties in its cluster. This leads us to believe that the death rate in Imperial county was as high as it was due to a large, poor, hispanic population containing immigrants from Mexico.

Looking at the county with the next highest amount of deaths, we can see that Inyo county on the border of Nevada had around 2.0 deaths per 1000 people. Recalling from the racial demographics clustering, Inyo county had the highest density of American Indians in the state, at around 14% of the population. However the other counties that were clustered with high American Indian populations, Humboldt and Del Norte counties, do not share this same high death rate. Both Humboldt and Del Norte share the lowest rate of deaths per 1000 people in the state, as shown by the dark blue color in the north west of the state. This leads us to believe that a high density of American Indians is not the cause for Inyo County's high death rate. Both k-means and hierarchical clusterings of the economic subset group Inyo county in the cluster with the lowest median income, median rent, and income per capita. However, when inspecting the data on Inyo county, it does not seem to have any significant outliers in these areas when compared to other counties in this cluster. Because of this, we believe that the high death rate is not completely related to the areas we investigated, and therefore do not feel confident in classifying a particular feature as the reason for the death rate. That being said, as shown in the heatmap, the entirety of the valley area of California has ranging from around 1.0 to around 2.0 deaths per 1000 people. As shown in economic clusterings, these counties are among the poorest in the state, which would give its residents less access to any potential medical attention they would need.

Now looking at the counties with the least amount of deaths, we can see very dark areas in the Northwest, on the Peninsula in the Bay Area, and in the mountainous region to the East. Beginning with the Northwest, this region has a high population density of American Indians as discussed earlier, but they are also predominantly white. These counties do not have a high percentage of hispanic, black or asian people in them. This could very easily not be the reason for why their death rates are so low, as will be shown by the Peninsula, but it is worthwhile to mention, since everything in the state should be compared back to Imperial County. The Peninsula,

which contains San Mateo County, Santa Cruz, and San Francisco are also on the very low end of deaths per 1000 people. These three counties were shown to be part of the clusters with the highest amount of median income, median rent, and income per capita, with San Mateo County being the highest in the entire state. We believe that this is the main driving force behind why there were so few deaths in these counties. The people here were able to afford the medical services they needed, and thus had a lesser death rate. These counties are also very high in population density of asians, which could suggest that asian people are less susceptible to COVID-19. The counties in the Sierra Nevada mountain range also had very low deaths per 1000 people. These counties, like the Northwest counties, contain a very high population of white people, and low populations of asian and black people. Again, race has not proven to be a leading factor when it comes to deaths to COVID-19 but it is necessary to mention.

Overall, with the exception of Imperial County and a few others, California counties had a low death rate across the entire state. It is worth bringing up that Los Angeles county, which contains approximately 25% of California's population, has been excluded from this analysis because it is such an extreme outlier.

Exceptional Work: Partitioning Around Medoids

Partitioning around medoids is a method that is quite similar to K-means, but instead of creating clusters around centroids, which are essentially newly created points, it creates clusters around medoids. Medoids are the points that are most central to a cluster, and represent the average point in a cluster. PAM clustering is useful for datasets with a small number of clusters, or noisy data. PAM is also less sensitive to outliers compared to K-means.

We used PAM clustering on both of our data subsets with the optimal number of clusters found above.

Figure 18: PAM Clustering Racial Subset

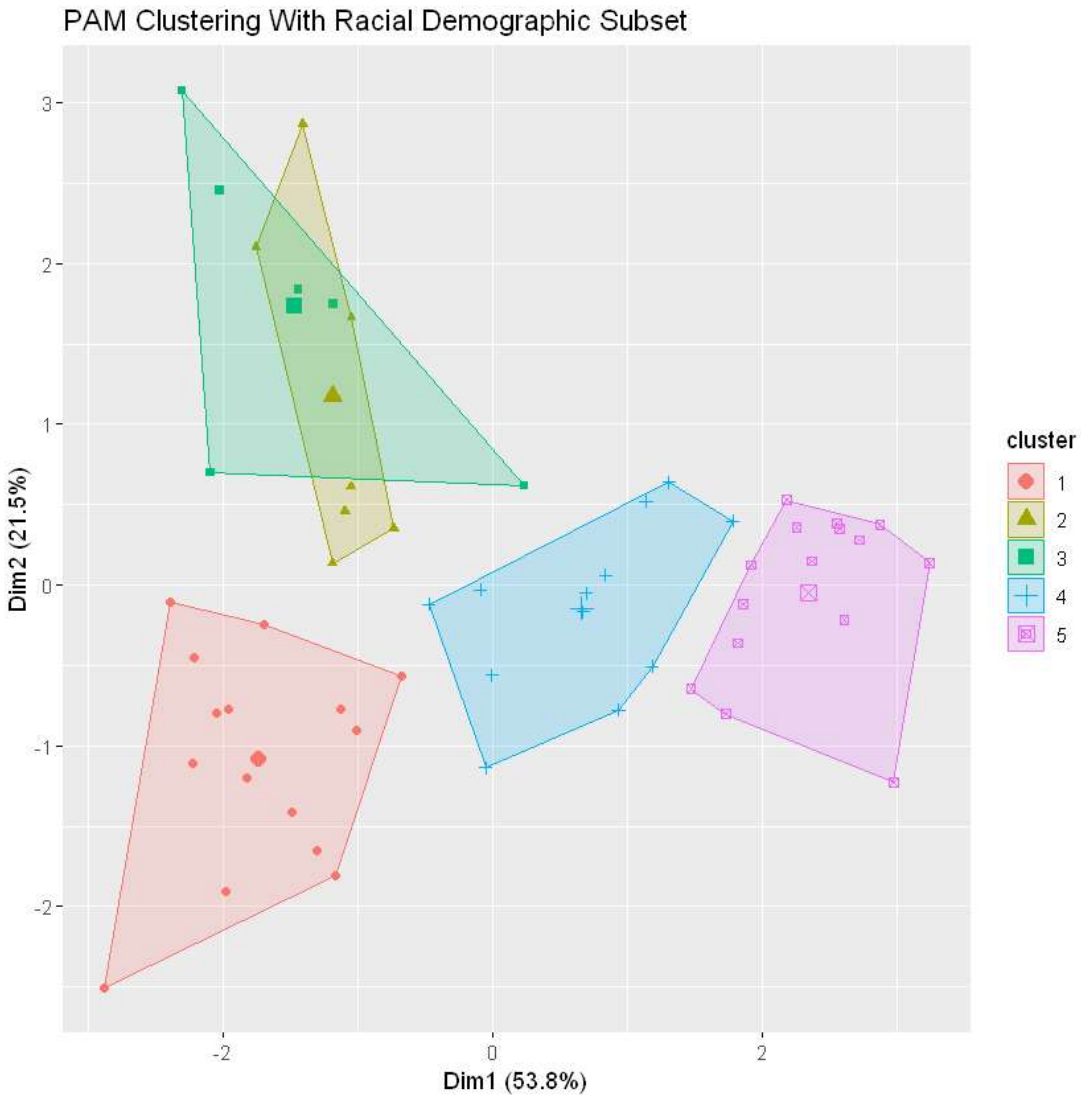


Figure 19: PAM Racial Silhouette

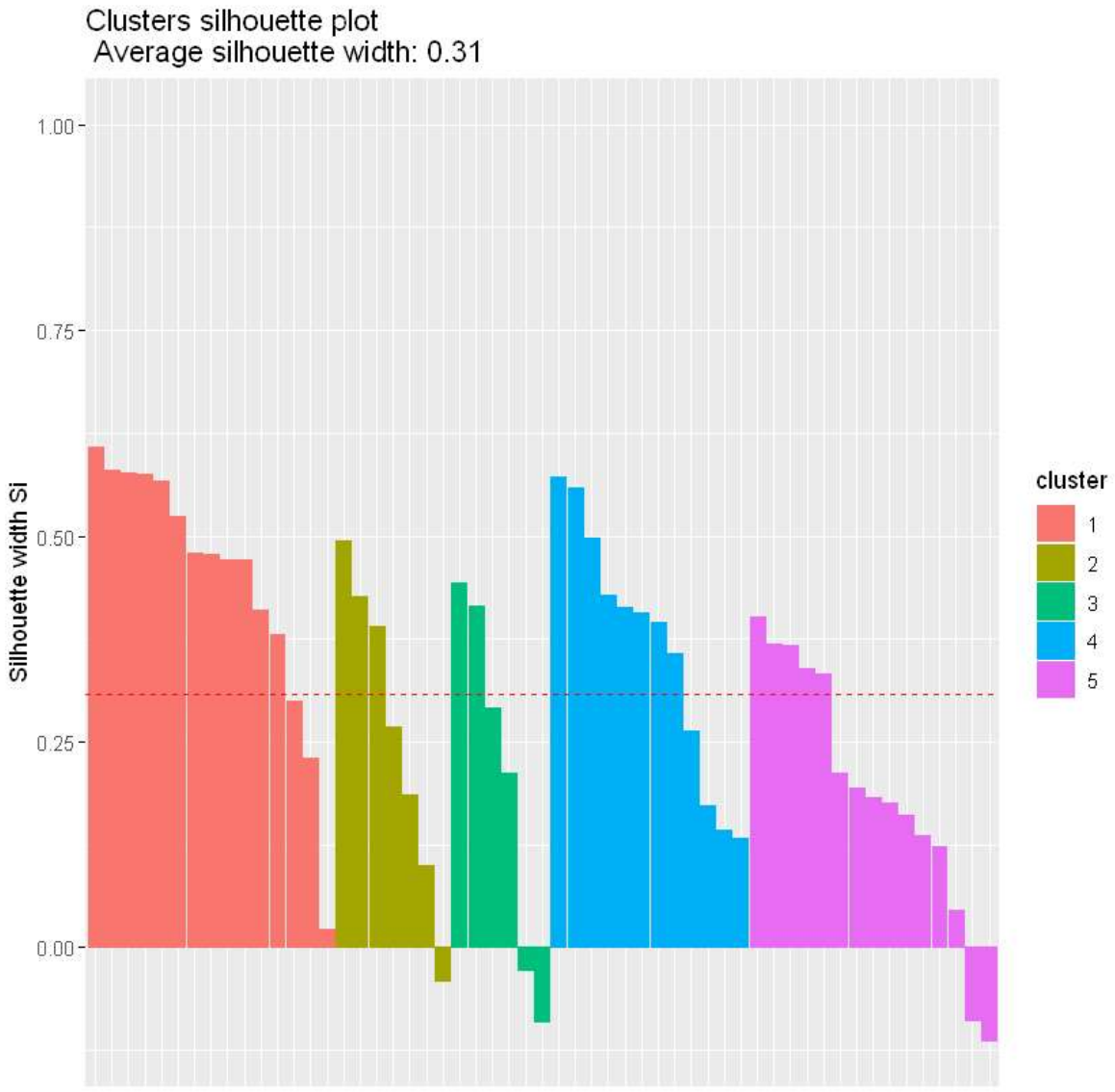


Table 10: PAM Racial Ground Truth

Ground Truth:

Cluster	Deaths Per 1000	Cases Per 1000
1	0.9404522	92.21818
2	0.5804953	55.55402
3	0.5464867	71.75972
4	0.4547119	54.04297
5	0.5008852	37.09741

We can see in Figure 19 that using PAM clustering on our racial demographic subset with the optimal number of clusters found from above gives some good clusterings. The clusters visually look separated other than cluster 2 and 3 which overlap each other and have a small number of datapoints. For internal validation, the average silhouette width of 0.31 shows that a pretty weak clustering structure was found. But, looking at the ground truth table for the clusters, we can see that cluster 1 was able to identify a group with much higher death per 1000 and cases per 1000 than the other groups. This could be used to inform the local government about a group with a higher risk of harm from COVID-19.

Figure 20: PAM Clustering Economic Subset

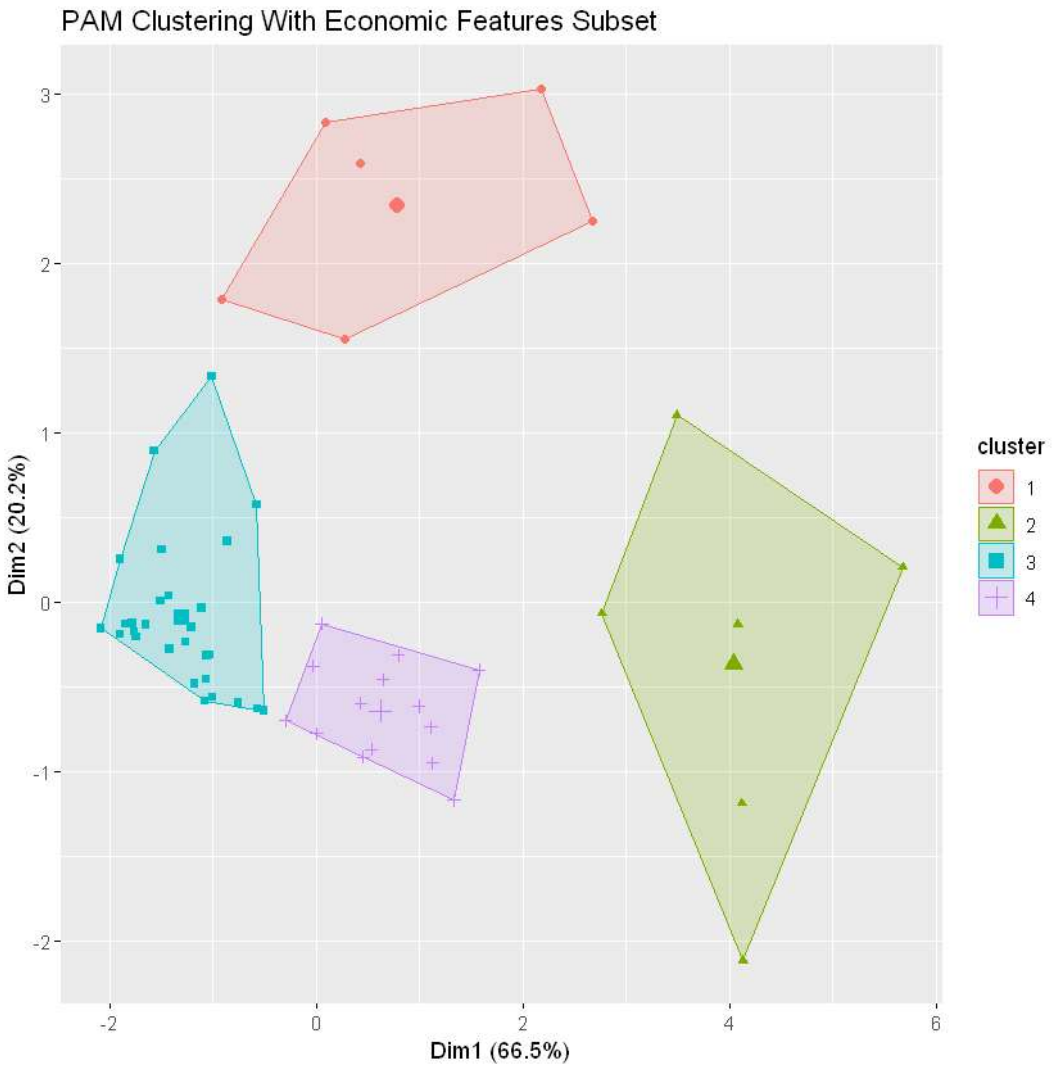


Figure 20: PAM Economic Silhouette

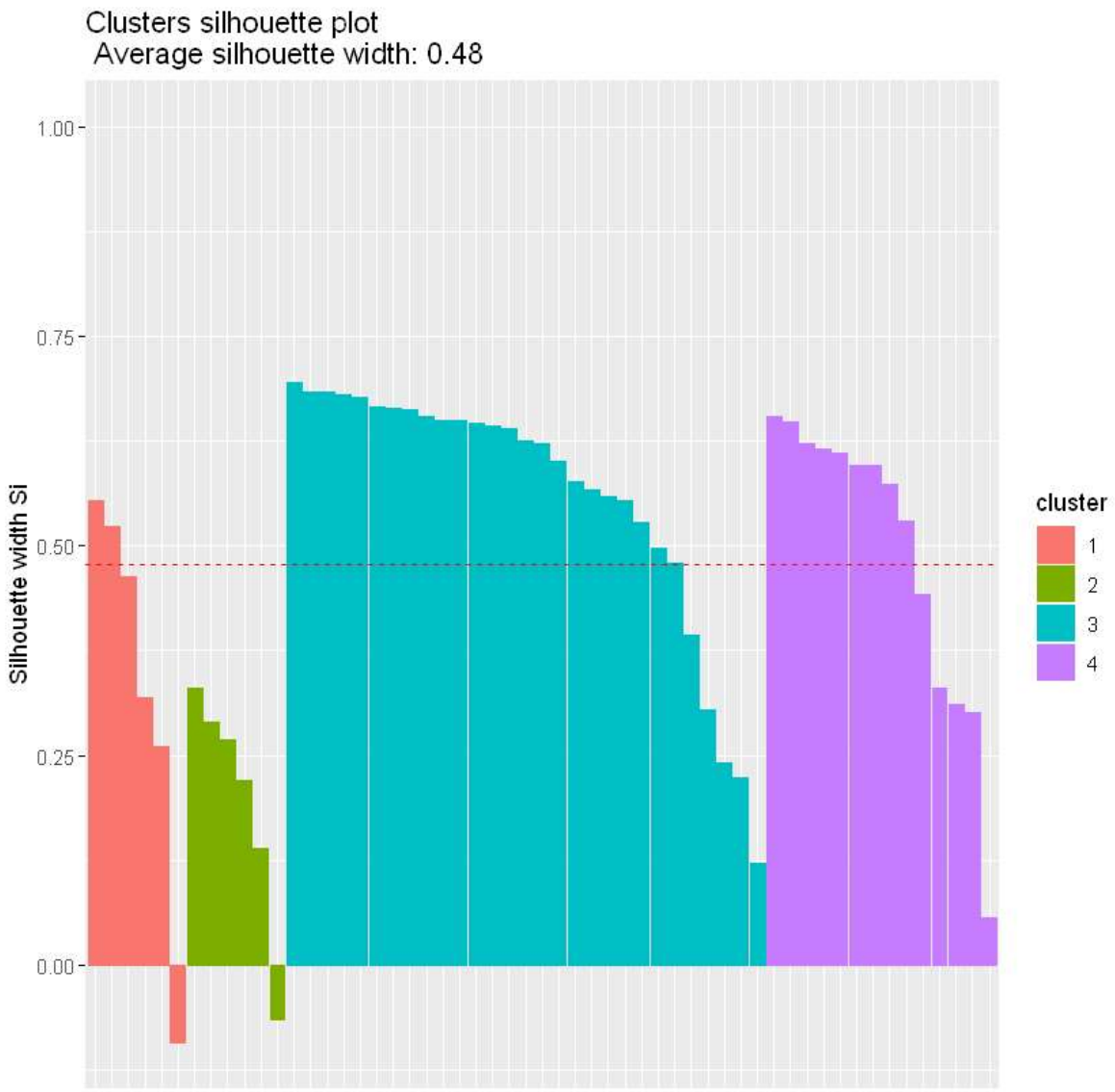


Table 11: PAM Economic Ground Truth

Ground Truth:

Cluster	Deaths Per 1000	Cases Per 1000

1	0.8126905	82.44216
2	0.4518307	43.04868
3	0.6998420	63.78478
4	0.4668890	57.49874

We can see in Figure 20 that using PAM clustering on our economic data subset with the optimal number of clusters found from above gives clusterings which are very similar to the K-means clusterings. The clusters all look separated, but have an uneven number of points in each cluster. The average silhouette width of 0.48 indicates the strongest clustering of all of the clusterings so far. Looking at the ground truth table for the clusters, we can see again that cluster 1 was able to identify a group with much higher death per 1000 and cases per 1000 than the other groups.

Evaluation & Conclusion

From all of our clustering algorithms used, we can see that some performed well and others performed poorly. In terms of our external validation variables of cases and deaths per 1000, K-means clustering of the racial demographic subset performed the best of all of the clustering algorithms. The K-means clustering of the racial demographic subset was able to identify the clustering with the highest deaths per 1000 and cases per 1000, a clustering with low cases and high deaths per 1000, and a clustering with low cases and low deaths per 1000. These clusterings could give lawmakers more information about what demographics are at low or high risk of harm from COVID-19. When looking at the death rates of the counties in California, the main takeaway is that wealthy counties will have the lowest death rates, while poorer, predominantly non-white counties will have a higher death rate.

References

[1] https://www.census.gov/quickfacts/delnortecountycalifornia

[2] https://blog.bioturing.com/2018/06/18/how-to-read-pca-biplots-and-scree-plots/

[3] https://web.archive.org/web/20111002220803/http://www.unesco.org:80/webworld/idams/advguide/Chapt7_1_1.htm