Multivariate Analysis of Social Determinants and COVID-19 Outcomes

Alex Ho and Eduardo Landin

STAT 541 - May 2021

1 Background

The COVID-19 pandemic has demonstrated the world's overall lack of preparedness to fight a global health threat. In order to better prepare for future pandemics and public health challenges, it is important to identify the factors that can help mitigate or exacerbate the impacts of a global health crisis.

We decided to analyze the relationship between a large number of social factors (pre-pandemic demographics, health outcomes, and health resource availability) and COVID-19 case counts and deaths. Dr. Bin Yu created a repository with demographic and health data for the majority of counties in the U.S. (Bin Yu et al). This data set, when combined with cumulative case counts and death counts from the New York Times served as our primary source of data for the analysis that follows.

2 Dimensionality Reduction and Data Augmentation

Dr. Yu's data set tracked 103 variables for each county in the U.S. these variables included:

- 6 identifying variables: e.g. state name and county name.
- 7 geographical identifiers: e.g. longitude and latitude.
- 38 population variables: e.g. total population and number of females aged 40-45.
- 9 social outcome related variables: e.g. unemployment rate and high school graduation rate
- 8 health resource availability variables: e.g. number of hospitals, social vulnerability index percentile and number of ICU beds.
- 20 health outcome variables: e.g. stroke mortality and percentage of adults with obesity.
- 15 social distancing and mobility variables: e.g. estimated percentage of people who never wear masks and date when public schools were closed.

In order to reduce the number of variables and simplify the data analysis process, we decided to remove variables with a high number of NA values (over 80% missing values). We also identified and removed the variables that exhibited very little variability. For example, we found that most counties closed down public schools and enacted travel bans at around the same date. With so little variability, there was little reason to try to study the impact that these variables would have on the vastly different health outcomes of each county.

Even after removing these variables, we still had over 65 columns. We managed to decrease this number down to 37 by removing some of the population variables. We did this because the different population variables were highly correlated with one another and because some of the population estimates from 2010 could be replaced with the population estimates from 2018. In the end we decided to keep the total number of males in the county, the total number of females in the county and the estimated population in the 65+ age group. These three values are estimates from 2018 and so they would be more recent than the other 2010 census estimates included in the data set. Furthermore, these variables serve as indirect measures of the male-to-female ratio in the county, the total county population, and the proportion of older citizens in the county. As a result, there was no need to keep other population variables that also served as indirect measures of these same demographic variables.

After removing these columns, we then removed all of the counties that had any missing values or where not in both the Yu data set and the New York Times data set. This reduced the number of counties to about 2500 (there are about 3000 counties in the contiguous U.S).

After reducing the number of variables, we divided the total case counts and death counts by the total population. We then created categorical variables for the case counts. A value of 1 (red in the LDA plots below) was assigned to counties with case counts below the 1st quartile (low case count), a value of 2 (blue in the LDA plots below) was assigned to counties with case counts

between the 1st and 2nd quartile (medium case counts) and a value of 3 (green in the LDA plots below) was assigned to counties with case counts above the 2nd quartile (high case counts). This process was then repeated for the death counts, as well.

Finally, the entire data set was centered and scaled. This was accomplished by computing the mean and standard deviation of each column, subtracting each entry by the corresponding column mean, and dividing that quantity by the corresponding standard deviation.

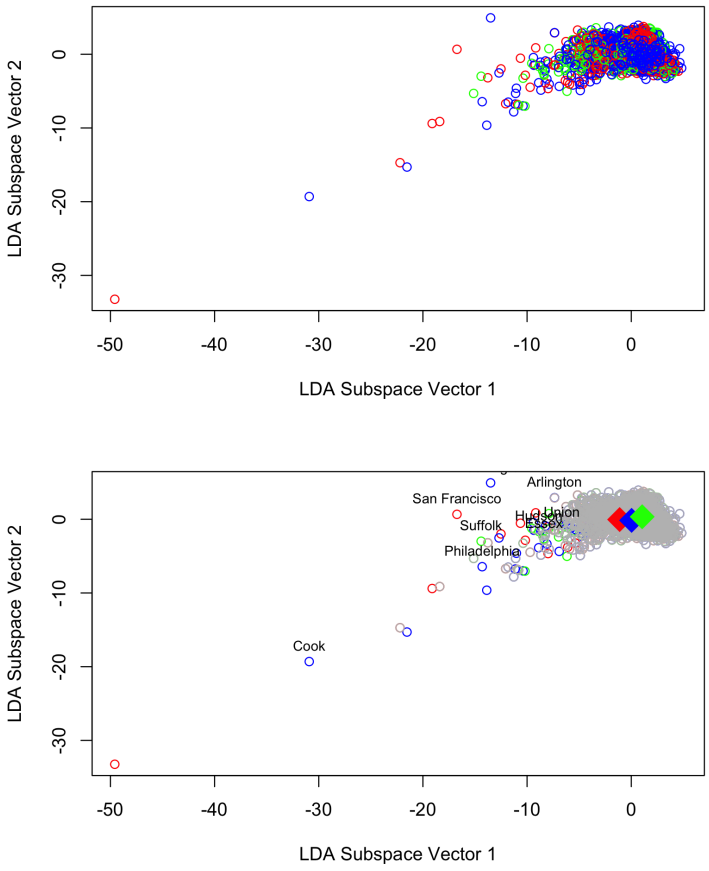
3 U.S. Counties Linear Discrimination Analysis

3.1 U.S. County Cases

We start with the case data for all counties in the United States, excluding New York state. From 1 we can immediately see that there is very little separation between the groups and their means with a sparse trail of points separate from the main cluster of points. These points do not all belong to any one case level group, and when viewing only the high-density counties it is also apparent that density also does not explain these points.

3.2 U.S. County Deaths

The death data in 2 for all counties in the United States shows a similar story to the case data. There is a predominant cluster of points consisting of all three death level groups with little separation having means that are very close. There is again the trail of sparse points separate from the main cluster, however this time it is more evenly symmetric about zero along the second LDA subspace vector compared to the cases plot 1 which has almost all of the trailing points in the negative values for the second LDA subspace vector. We also can notice that population density does not alone explain these trailing points, unlike in the Texas level data in 3 and 4. Overall, though, the data for both the cases and deaths have the same shape using counties across the United States.



(a) Counties with low population density are colored grey.

Figure 1: U.S. case by county data. Red is the lowest range consisting of the bottom quartile, blue is the middle two quartiles, and green is the fourth quartile. Projected into LDA subspace using means of low (red), medium (blue), and high (green) groups.

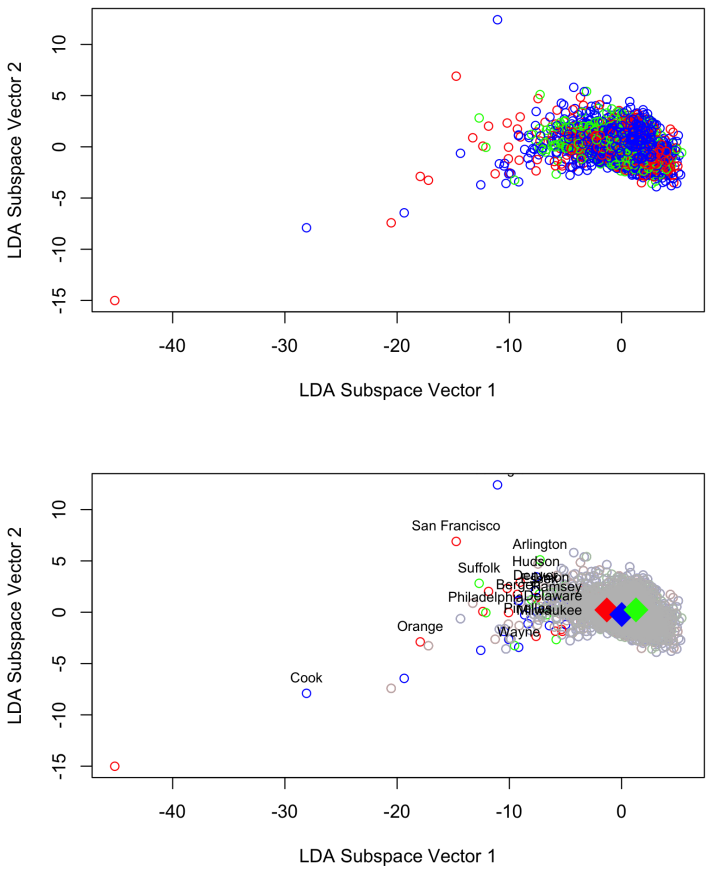


Figure 2: U.S. death by county data. Red is the lowest range consisting of the bottom quartile, blue is the middle two quartiles, and green is the fourth quartile. Projected into LDA subspace using means of red, blue, and green groups.

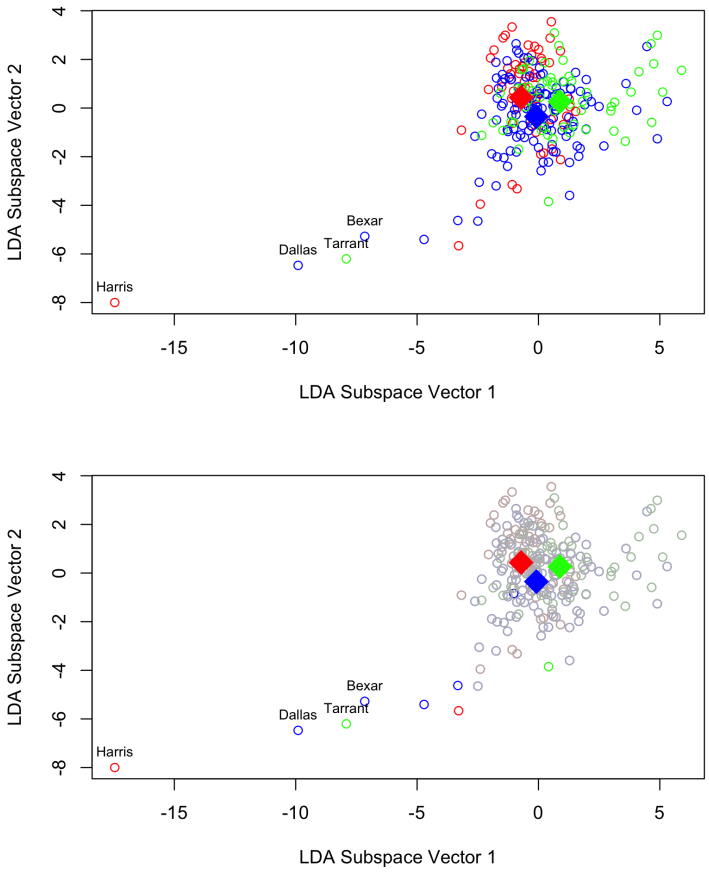


Figure 3: Texas case by county data. Red is the lowest range consisting of the bottom quartile, blue is the middle two quartiles, and green is the fourth quartile. Projected into LDA subspace using means of red, blue, and green groups.

4 Texas Counties Linear Discrimination Analysis

4.1 Texas County Cases

Next we move on to using only Texas counties to explore COVID case data in 3. With an order of magnitude fewer data points, the plot is much clearer however a similar trend to the United States level data emerges. The means for each of the case level categories are very close together, and the groups also do not show clear separation. With the colored and labeled data points, one could

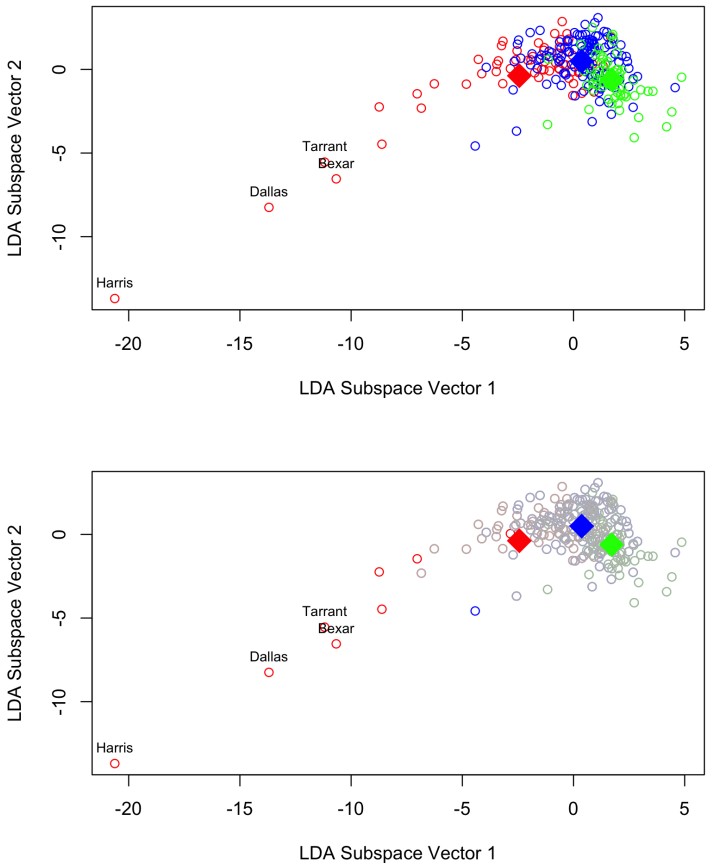


Figure 4: Texas death by county data. Red is the lowest range consisting of the bottom quartile, blue is the middle two quartiles, and green is the fourth quartile. Projected into LDA subspace using means of red, blue, and green groups.

attempt to find groups with high level (green) points being in the upper right, while middle level (blue) is central and low level (red) to the left. However, a priori without labeling and coloring, there is no clear grouping. When highlighting the high population density counties, there is a prominent trend that that they make up the sparse trailing points.

4.2 Texas County Deaths

The deaths by Texas counties plot shown in 4 illustrates the clearest trends of any of the plots. While the means for each death level group are close, they are the most divergent compared to the Texas case plots and the national level plots. With coloring, we can more easily see that the red low death level points all are to the left side of the plot and that they make up almost all of the trailing points separate from the main cluster. We can also note that the blue middle level points all cluster towards the center, and the green high level points collect slightly below the blue middle level points and are also more off to the right of the plot. We can also see that in the bottom plot that almost all of the high population density counties are made up of the low death level points, with just a single middle level point being considered high population density by our cutoff. This trend makes sense given background information that more population dense counties are more likely to have several hospitals leading to easier access to better healthcare compared to rural, low population density areas. With the labels we can also see that the four largest metropolitan counties are part of the trailing, high population density and low death level points.

5 Texas Counties Hierarchical Clustering

Since the LDA subspace was not successful at separating the three groups, we decided to perform hierarchical clustering and K-means clustering. Our goal was to identify new clusters and thus identify what made counties similar and different from one another.

We performed hierarchical clustering using complete linkage. The resulting Dendogram (see appendix) showed that the counties could be easily split into two clusters: one very large one with most of the counties and another smaller cluster containing 4 counties. Figure 5 shows the result of using two clusters (left) and three clusters (right). In order to plot the data in 2 dimensions,

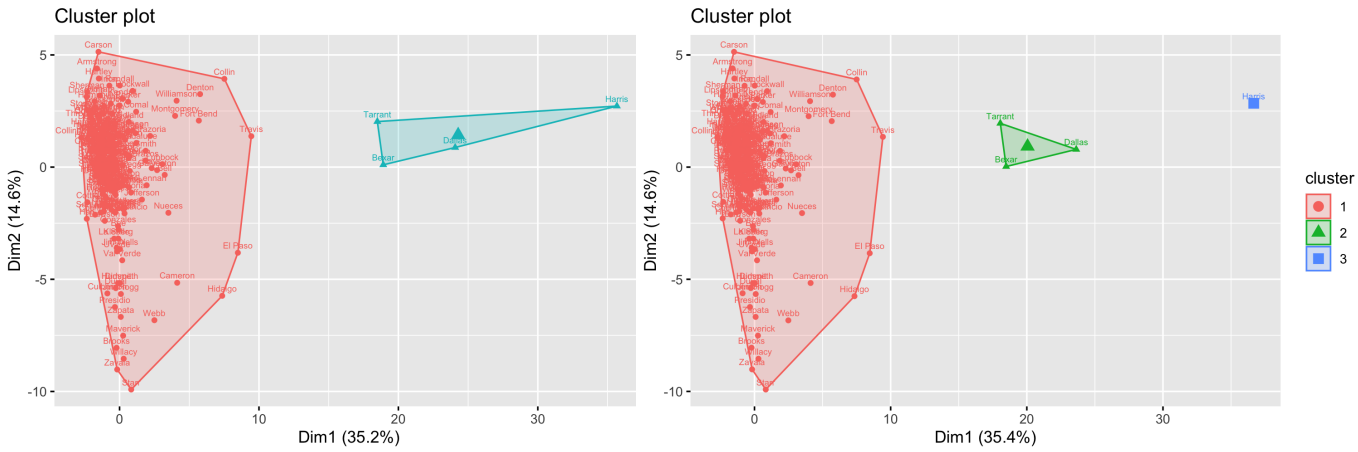


Figure 5: Hierarchical Clustering of Texas Counties Using 2 and 3 Clusters

the counties have been projected onto the subspace spanned by the first two principal components (the value in the brackets specifies the percentage variance provided by the principal component).

In the two cluster graph we can see that the four most densely populated counties (Bexar, Dallas, Harris, and Terrant) are once again seperated from the other counties. This echoes the results from the LDA subspace plot.

In the three cluster graph we see that Harris county has been identified as an outlier within the densely populated county cluster and so it has been labeled as its own separate cluster. Although we cannot visualize the data in 39 dimensions, this behavior seems quite reasonable when looking at the PCA subspace plot: hierarchical clustering with complete linkage tends to form spherical clusters and so perhaps in an attempt to make the second cluster more spherical, Harris county was isolated into its own cluster.

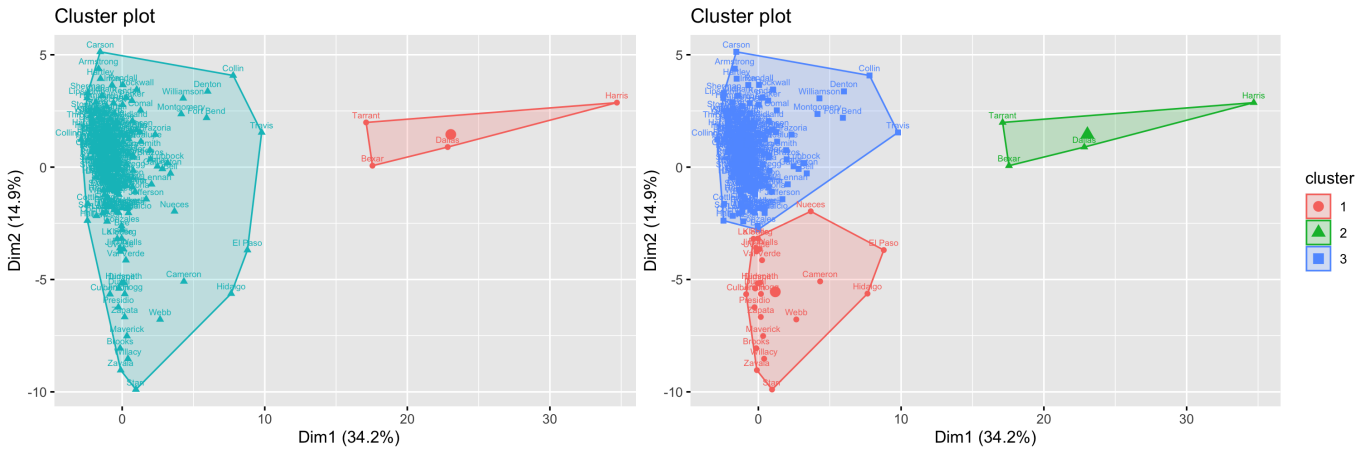


Figure 6: K-Means Clustering with Texas Counties

6 Texas Counties K-Means Clustering

Figure 6 shows the results of performing K-Means clustering with $k = 2$ (left) and $k = 3$ (right). As seen in the plot, using $k = 2$ yields near identical results to performing hierarchical clustering with two clusters.

However, when we try clustering with $k = 3$, k-means does not move Harris county into its own cluster. Instead, it breaks off the large cluster containing a majority of the counties into two smaller and seemingly more spherical clusters.

The table 7 shows the value of the means for each of the clusters for the $k = 3$ case. Recall that all variables have been scaled and centered. Also note that the total cases and total death values were divided by the total population of the county before being scaled and centered.

We can see that the green group has a much higher population density than the other two groups. When looking at the group means we also found that this group tended to have a greater number of medical resources (e.g. hospitals and ICU beds). This result makes sense as urbanized areas likely have greater medical resources than smaller rural areas.

Group	total cases	total deaths	Population Density	Dem-Rep Ratio	Unemployment Rate
Blue	1.058	0.647	0.008	2.483	1.327
Green	0.149	-0.955	6.596	1.196	-0.132
Red	-0.113	-0.050	-0.120	-0.281	-0.136

Figure 7: K-means clustering group means by variable for K = 3. See 6 for plot.

The table 7 shows some of the variables where there is the greatest difference between the blue cluster and the red cluster. We can see that although the two clusters have similarly low population densities, the Democrat to Republican ratio and the unemployment rate is very different in these two groups. The blue cluster also had a significantly higher average percentage of single parent households (1.252 vs. - 0.140), a significantly higher average percentage of people in poverty (1.608 vs. -0.162). This subset of variables and the differences in the means would seem to indicate that the blue cluster generally has worse social outcomes than the red cluster, however, more rigorous analysis and testing is required before making any conclusions.

7 Conclusion

The LDA subspace did not effectively separate the data according to the labels of low, medium and high case counts. This result could be the effect of poor categorization on our behalf, or evidence to show that there are few underlying factors that contribute to COVID case and death rates across the United States. Using the principal components along with clustering algorithms yielded the most interesting results which showed signs of social outcome and medical access differences across Texas.

Although the results above show promising directions for further exploration and analysis, it is important to note that it is too soon to make any decisive conclusions about the data. More formal statistical analysis–in the form of confidence intervals, hypothesis testing and uncertainty quantification–is required in order to truly understand the differences between the clusters. This work serves better as a staring exploration point to discover more research questions to answer for a more detailed statistical analysis.

Finally, in order to reason about the relationship between the various variables and the case counts, regression modeling and variable selection is necessary.

8 Appendix

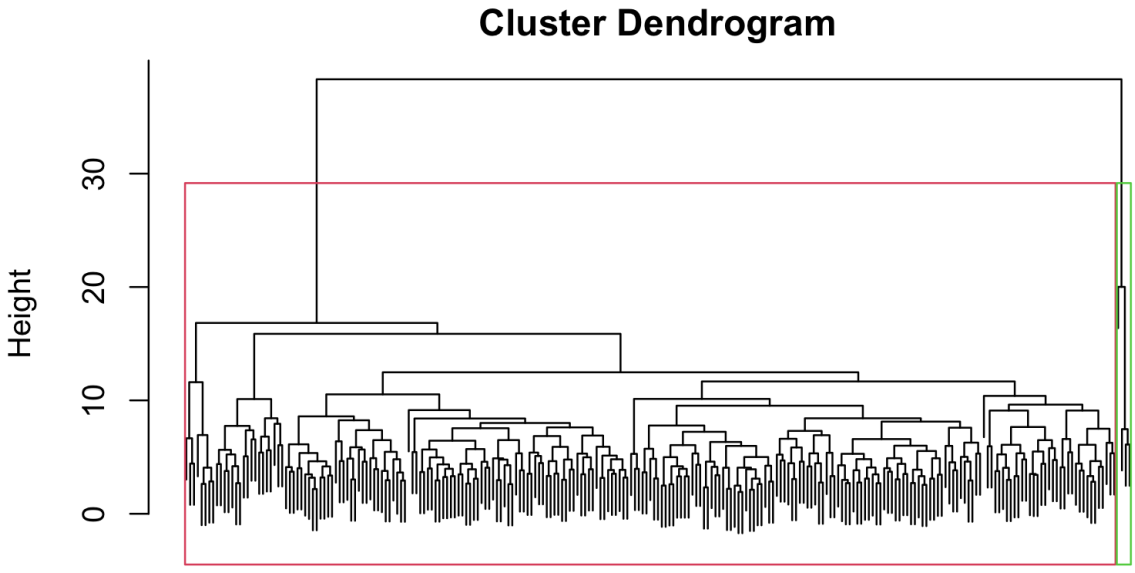


Figure 8: Dendogram resulting from complete linkage hierarchical clustering with K = 2

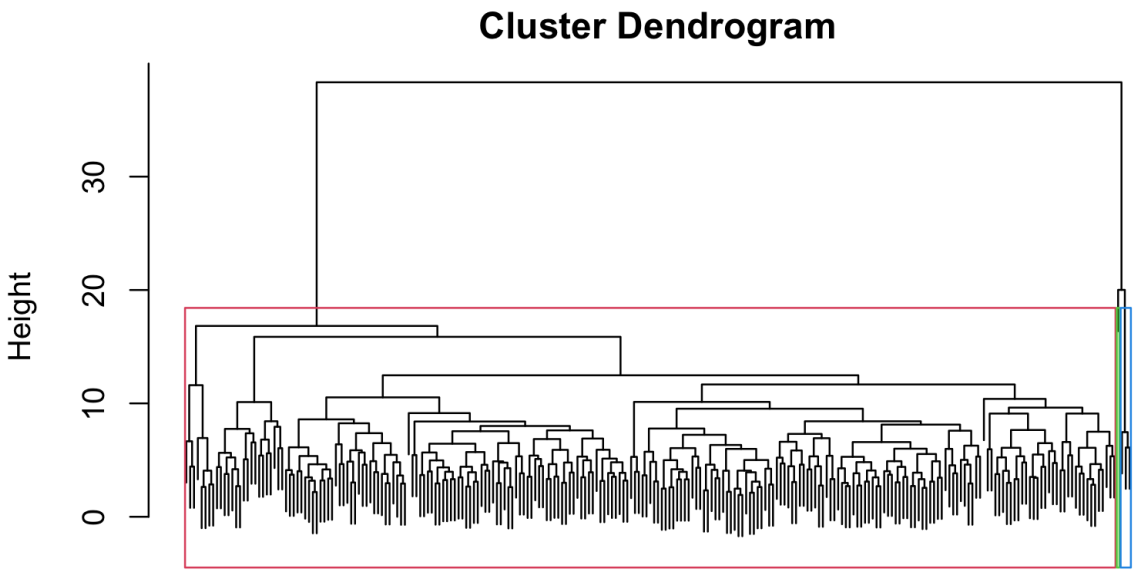


Figure 9: Dendogram resulting from complete linkage hierarchical clustering with K = 3

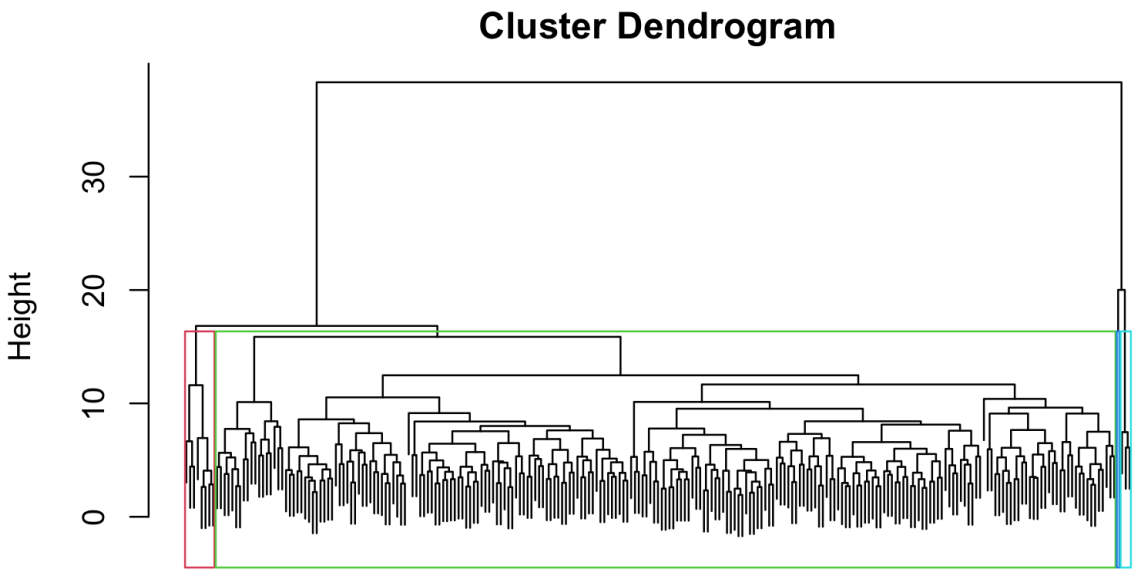


Figure 10: Dendogram resulting from complete linkage hierarchical clustering with K = 4