STA 4990 Project 2

Kyle Folbrecht, Anthony Mastrangelo, Michael Pena

May 11, 2023

Overview

In this project we are not attempting to find a response or fit a model to a data-set, we are attempting to cluster the variables together in order to set ourselves up for success when we need to predict a response variable, using other methods such as Random Forests or Support Vector Machines. We will go over the clusters that we created, why we decided on the amount of clusters that we decided upon, and how to define each of the clusters. In this dataset the variables are as follows:

- **Sex** = Binary Variable, 0 = Male and 1 = Female,
- **Martial.Status** = Binary Variable, 0 if not, 1 if they are married
- **Age** = Age in years
- **Education** = 0 if unknown/other, 1 if highschool, 2 if university, 3 if grad school
- **Income** = Annual Income in US dollars
- **Occupation** = 0 if unemployed/unskilled, 1 if skilled employee/official, 2 if management/self-employed/officer
- **Settlement.Size** = 0 if small city, 1 if mid-sized city, 2 if big city

We have also created variables in order to normalize the data, by converting our two numerical variables to:

- **div_age** = Normalized Age data by taking the function 1/Age.
- **quantile_income** = Converting income into a categorical variable, 2 if low income, 3 if middle-class income, 4 if upper-class income.

Initial Analysis

Given that we only have two numerical variables, namely income and age, the plot that we need to analyze the clusters on should be that plot, since all of the other categorical plots will just be used in order to classify each of the clusters. In the raw data graph, we have also inserted the average income of the dataset, in order to show where the most concentrated part of the data would be. This concentrated portion of data ranges from the ages of 20 to 40 centered around the average income.

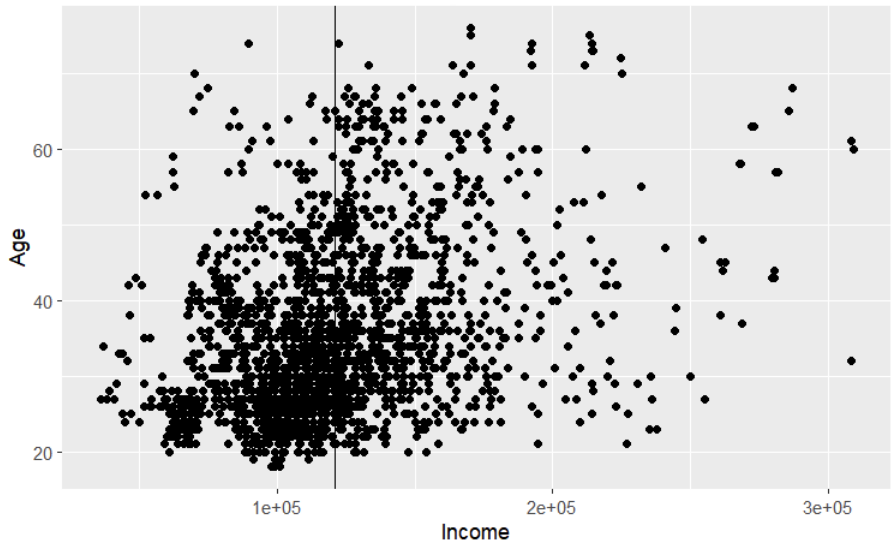

Figure 1: Raw Data, Income vs Age

Reinforcing the idea above, when plotting a histogram of age and income, we can see the data for each is right-skewed. Once again we have added the average value for age and income. We can see that the average age of our data-set is 36-38, and the average income of our data-set is approximately 120k.

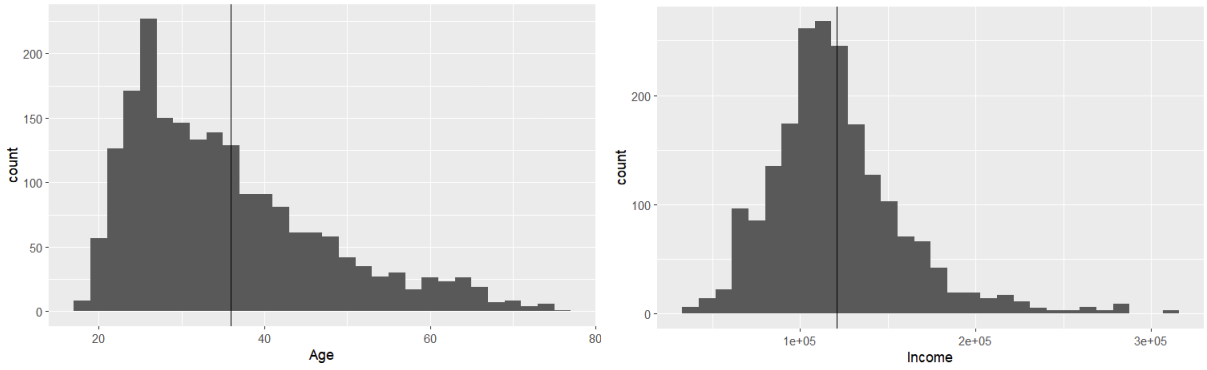

Figure 2: Distributions of Numerical Variables

There will be instances in our code where we used normalized versions of the variables as described above in order to obtain more accurate clusters; however, none of the data that will be shown in this paper will use those aforementioned variables

Clustering

When looking at a scree plot derived from our data, there were multiple points that seemed like they could be the "elbow". The most obvious one is at 6, but there are slight elbows at 4 and 5 as well.

Figure 3: Scree Plot

We decided to split our data in 4, 5 and 6 clusters. What was found was that when we had 6 clusters in our data, they were not visibly separate from the rest of the pack. The counts of the clusters were even as shown although the disparity is not apparent while plotting the data. Though the aforementioned concentrated area in our plot was evenly split, the spread out values in our data were mixed between 3 different clusters

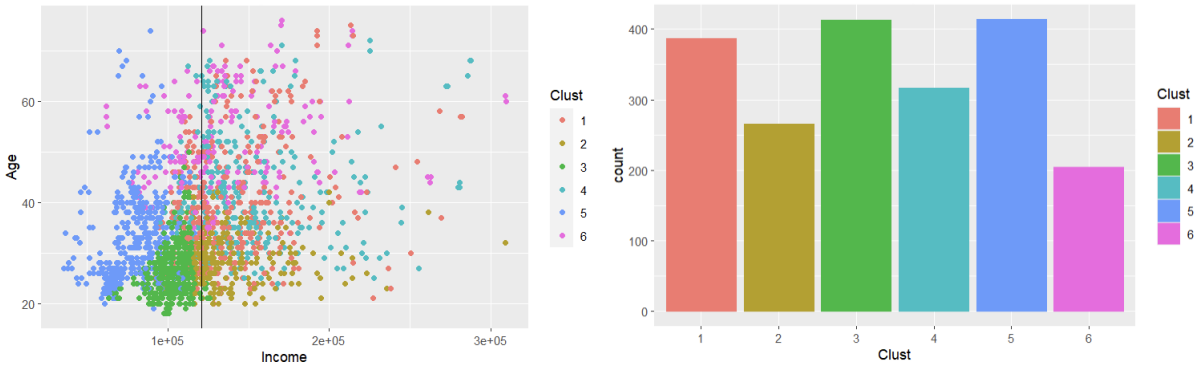

Figure 4: 6 Cluster Analysis

Having 4 clusters produced the exact opposite effect, where the data that was spread around the plot was evenly split in different clusters, but the concentrated area in the middle was very mixed. Unlike the 6 clusters, one cluster in our data had more values in it compared to the others, as it included most of the points that were spread away from the "center" of our data. The amount of datapoints in each cluster decreases as shown.

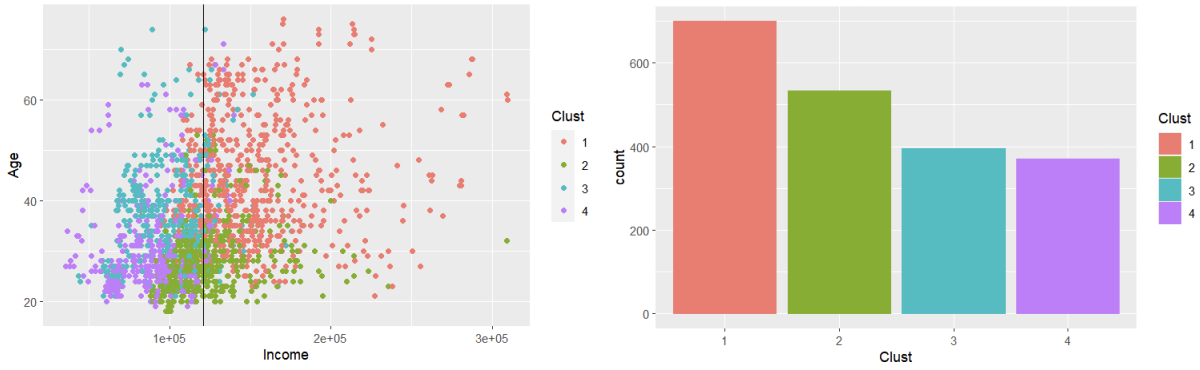

Figure 5: 4 Cluster Analysis

In the end, we decided to go with 5 clusters as it gave the most interpretable data whilst splitting the "center" of our graph in a very apparent way. Though the data points that were spread away from the most concentrated part of our plot seemed to be mixed within different clusters, the majority of the data points between each cluster had a characteristic to the majority of them; although, it wasn't exactly a "perfect" split.

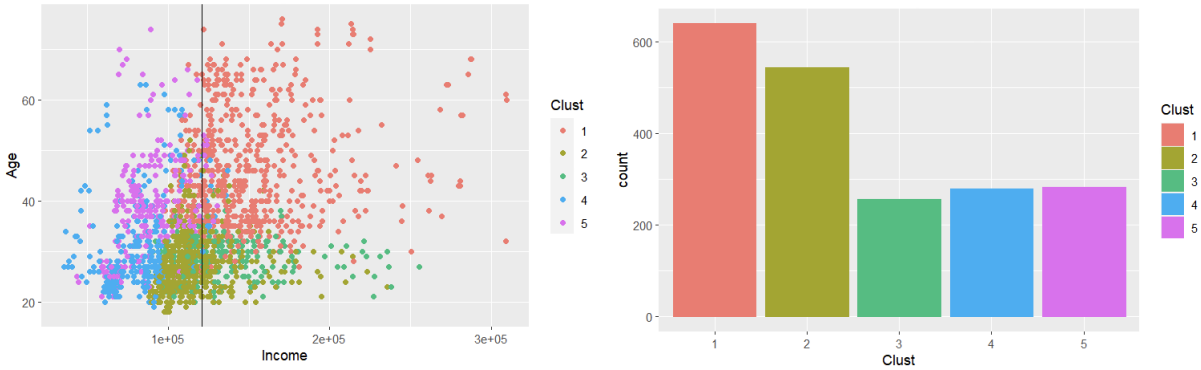

Figure 6: 5 Cluster Analysis

When clustering like this we can see that clusters 1 and 2 hold a majority of the data, and clusters 3, 4 and 5 have approximately the same amount of data points in each. We can see that the majority of the concentrated portion of our data has clustered together instead of being shared between multiple clusters as shown previously. As mentioned, this clustering isn't a "perfect" split as you can see that clusters 4 and 5 seem to follow the same trends; some of the data in cluster 3 is part of our "center".

Identifying Clusters

Clusters 4 and 5 are the most closely related of all the others. Both clusters are typically unemployed, live in small cities, and are part of the lower class, which are all factors that relate to each other.

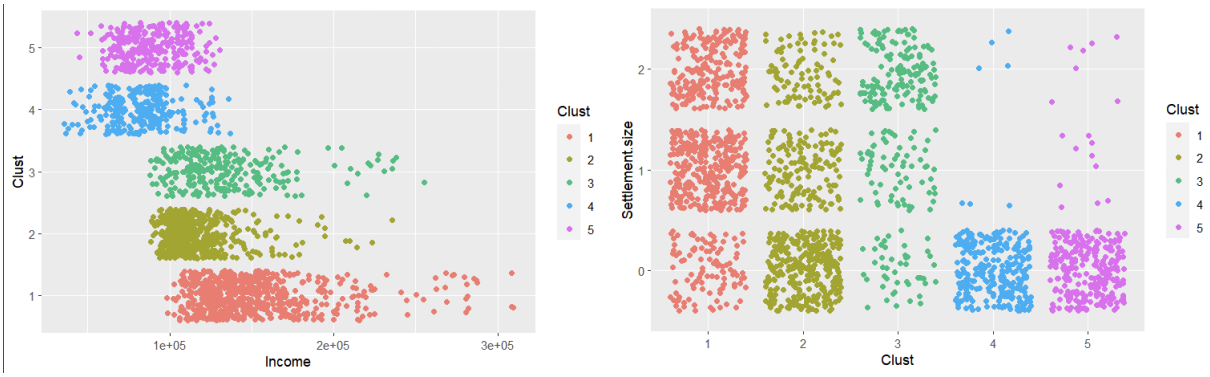

Figure 7: Income by Clusters

The main difference in these two clusters lies in their martial status and gender distributions. People in cluster 4 are Married Females, so they can be classified as "Stay at Home Mothers", whereas the people in cluster 5 are unmarried men, so probably men working minimum wage jobs such as McDonald's

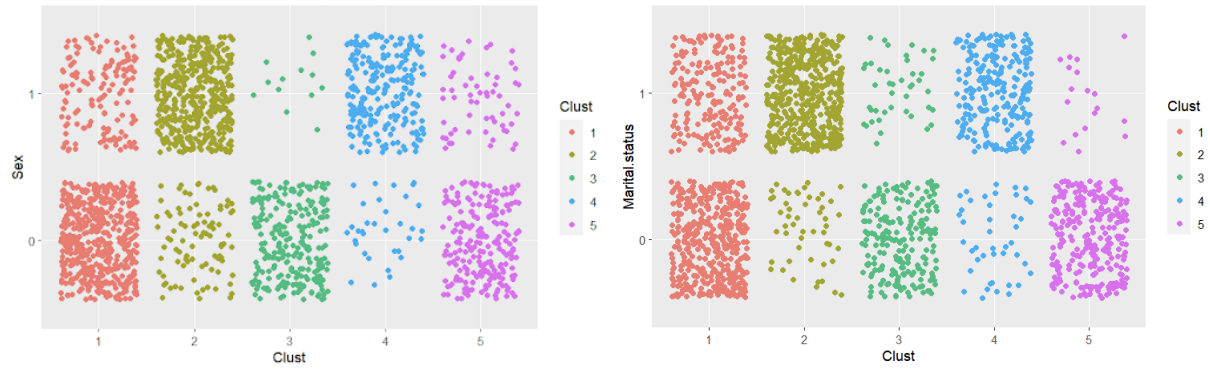

Figure 8: Demographics of Clusters

As shown above, cluster 3 falls under the same demographic as cluster 5, namely they are unmarried men. The main difference in this cluster compared to the others is the average age of this cluster. Where all the other clusters seem to center themselves around the aforementioned average age of 36-38, this cluster is typically the younger generation, with the maximum age being around 40 and the youngest being around 20. This cluster also has the lowest education on average, being the cluster with the largest amount of uneducated individuals.

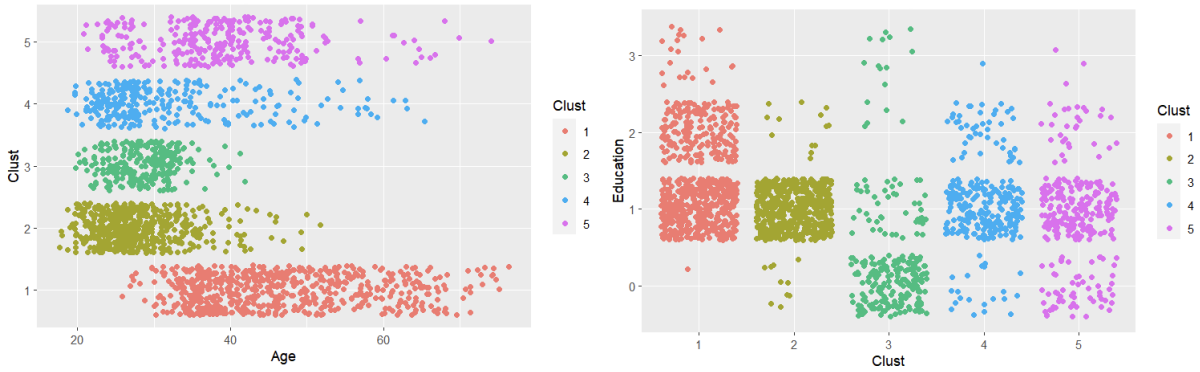

Figure 9: Age of Clusters

Cluster 2 also has a lower age distribution compared to its other clusters. This cluster is also comprised of mainly married women, much like cluster 4. The main difference with this cluster is that it's average education is less than the other clusters, with a majority of the data being in high school and none of the data in this cluster going to grad school.

Finally, our largest cluster, cluster 1 has a majority of the data in it. However, it seems to diverge from the rest of our data. As shown in previous graphs, this cluster has the greatest amount of people in college, the highest average age, and the highest average income, making it fit to the majority of the outliers. In the end, we are left with the Income vs Age plot looking like:

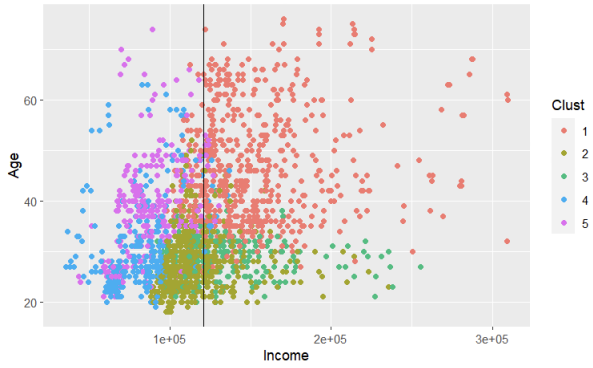

Figure 10: Final Model

Conclusion

In the end, we decided that the best way to visualize and separate the data was with 5 clusters. The 1st cluster seem to be the people with "Old Money" boasting the highest average age and highest average income. Cluster 2 are the working class married women, none of which have gone to grad school and live typically in smaller cities. Cluster 3 are the uneducated younger generation, boasting the lowest average age and education, yet the majority of the data here live in big cities. Cluster 4 are the "stay at home mothers" who have the lowest income and live in small cities. Cluster 5 is unmarried men who live in small cities. We chose to have 5 clusters since it would give us the best division of data compared to its neighboring amount of clusters 4 and 6 and centralized most of the data around the average age and income.