Demographics Report

Qian Fang

Among the chi-square dataset, I chose Demographics as my dataset for exploration and further analysis by using the statistical tools. I have three main parts in my report and they are data cleaning, EDA analysis, K-means clustering.

**DATA CLEANING**

First, I use *is.na* function to check whether there are missing values in my dataset and then I found that there is no missing value from the result. However, from the Define data values dataset, it indicates that data with “-2222” means no data available. In the summary of poverty, I found that there is a “-2222” in Hawaii, so I set it to 0, indicating that it is not an extreme value, but a missing value.

Second, most of the data is percentage, so the percentage of age group 0-85 should covers 90% or above 90% of whole population. The age ranges 19 to 64 should cover large parts of whole population. After checking for all of these, I can be sure that age data is ok and there is no need to clean.

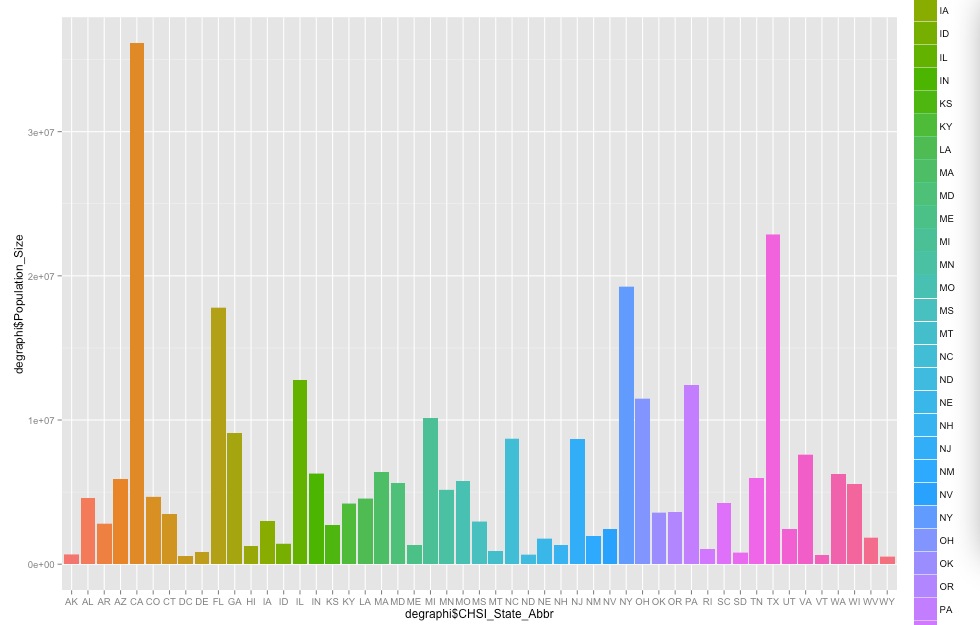
Third, I checked for race data. The percentage of race, white, Black, Asian, Hispanic should exceed 50% of whole population, and then I found that there are several counties that are not satisfied with this conditions. I looked at them and found that these counties are in the border of the United States, so it was reasonable that their population consists lots of people with other races. Thus, the race data is ok.

**EDA analysis**

In my EDA analysis, I used *ggplot* to see the relationship between States and population size, poverty, White, Black, Asian, and Hispanic. Here are some graphs from R output.

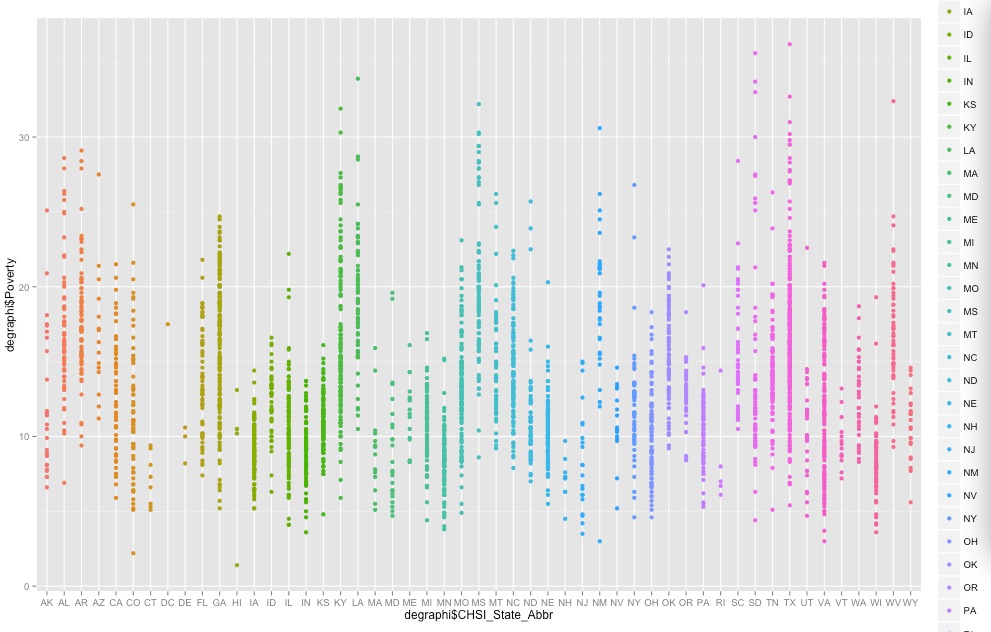
Here are some examples from my R output.

The first graph is the histogram of States and Population size.



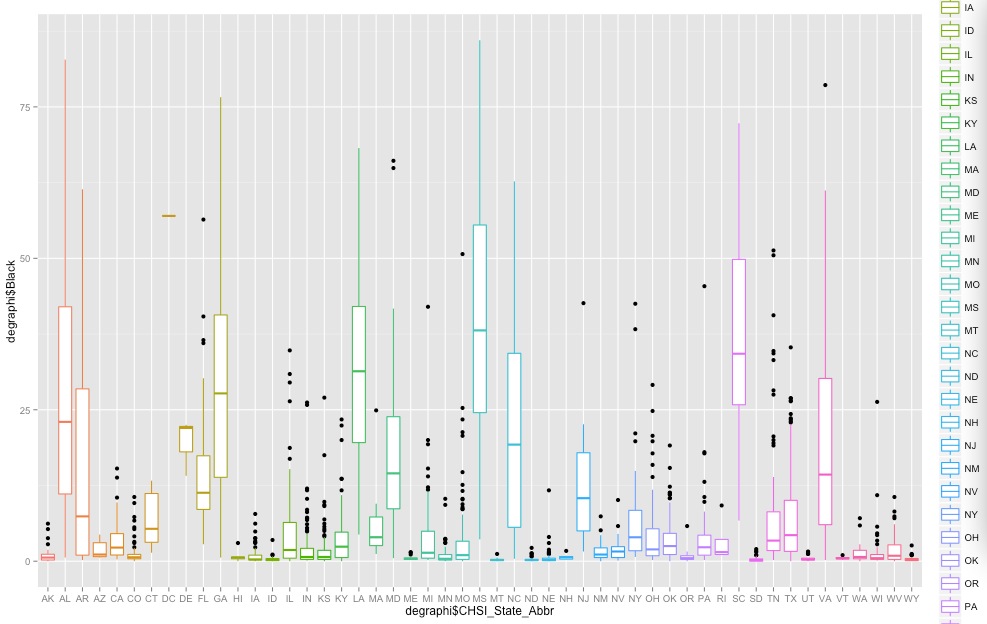
From the graph, we can see that CA has the largest population size, followed with TX and NY, while WY has the smallest population.

The second graph is plot of States and Poverty:



From the graph, we can see each dot represent each county within States. TX has comparatively high poverty rates than that of any other states.

The third plot is the boxplot of States and Black people



From the graph, we can see how black people are distributed in each state.

**Cluster K means**

Since my demographics dataset contains categorical variables, I cannot use *kmeans* directly, so I made a data frame of States and Population size. I used *kmeans(data2, centers=3)* to divide data to 3 clusters with similarity. Here is the graph.

