Capstone Projects 1: Exploratory Data Analysis

Daniel Lee

July 16, 2018

For the exploratory data analysis portion, I started by visualizing the distribution of the crude mortality rate caused by opioid overdose 2016 (county level). I created a boxplot and a plot of histogram and CDF overlaid on top of each other. From this, I can see that the distribution is skewed right.

Then, I created a correlation heat map of all the variables. Then, I created scatterplots for top 12 pairs of variables with highest correlation values. From this, I can see that population estimate and GQ estimates are the most correlated.

Then, I visualized all the correlation values between crude mortality rate and the other continuous variables by creating a bar plot and scatter plots. From this, I see that overall death rate is the most correlated with correlation value of 0.52.

Then, I performed a hypothesis test of correlation to see if the observed correlation values are statistically significant. I can see that all of the observed correlation values are statistically significant.

Then, I created violin plots and overlaid bee swarm plots of all the continuous variables except crude mortality rate. From this, I can see that many of the distributions are skewed right.

Conclusion:
It seems as if the the counties with higher mortality rates caused by opioid addiction have the following:

- Higher overall mortality rate
- Have more adults with high school diplomas only
- Higher opioid prescription rates
- Decrease in population size in 2016
- Lower median household income

The reason why counties with higher mortality rate caused by opioid overdose tend to come from counties with higher overall mortality rate should be investigated further. One reason that these counties have such high mortality rates could be that the population in these counties are older. Another reason could be that these counties may not have good access to healthcare for the residents. This can be something on which we can do more research.

Further research could also be done to investigate as to why counties that have high percentage of the population with only high school diplomas have high opioid overdose mortality

rate. This can be because the population in these counties are older. Or, people in these counties may not have access to higher education for some reason.