Capstone Project 1: Data Storytelling
Daniel Lee

For the storytelling portion, I started by asking the question and explaining briefly the number of deaths caused by opioid overdose in America from 1999 to 2016. Then, I created a bar chart of all the deaths by year to show the increase. On top of the bar chart, I overlayed a line plot.

Afterwards, I explained why we should care about the increasing opioid deaths. I included a 3-minute Youtube video of a testimony of someone who overcame opioid addiction.

Next, I generate three bar plots that show the total number of deaths by opioid overdose from 1999 - 2016 by demographic factors. One is the number of opioid overdose deaths grouped by race. Second is the number of deaths grouped by gender. Third is the number of deaths grouped by age groups.

Next, I create a US map visualization that contains gradient of colors that represent the corresponding crude rate levels on a county level.

Next, I create a bar plot of the top ten counties with the highest mortality rate caused by opioid overdose.

Then, I explained that I want to solve a prediction problem. The prediction problem is that I want to predict a county's rate of death caused by opioid overdose in 2016 based on a county's median household income, population estimate, unemployment rate, poverty rate estimate, educational attainment, and opioid prescription rate by health care providers.

After, I included a description of the datasets used for this project.

Then, I explained the variables that are in the dataset.

Then, I begin the exploratory data analysis. I begin by displaying the distribution of the crude death rate by county. I display a boxplot as well as a histogram with a density line overlaid on top of the histogram. I also have the ecdf of the crude mortality rate overlayed on top of the histogram.

Then, I create a heat map of the correlations among all the continuous variables. The heat map reveals that there's quite a number of variables that are highly correlated.

Next, I create a bar chart of absolute value of correlation values between crude death rate and all other continuous variables. In the bar chart are the nine variables with the highest absolute value of correlation values. Along with the bar chart, I display a dataframe table of the nine variables with highest absolute value of correlation when paired with crude death rate. I choose nine variables because nine is a nice number of variables to create scatter plots paired with crude death rate, which I do next.

After creating nine scatter plots, I then create nine violin plots using the same variables I used to create scatter plots. I do this to examine the distributions of the plots to see if there's any patterns that I'm missing.

After this, I conclude that there seems to be a strong relationship between crude death rate and the overall death rate of a county. The reason why this is the case should be investigated further. Perhaps these counties are older in age. Or, perhaps these counties do not have good access to healthcare.