Capstone Project 1: Data Visualization

https://github.com/mdubow/Springboard-Data-Science-Career-Track/blob/master/Capstone-Project-1/Data-Visualization/Capstone%201%20Data%20Visualization.ipynb

During the previous step of the project, data wrangling, I started with two datasets (one for prescriptions in the United States for the years 2006 to 2012, the other for deaths resulting from opioid overdose), cleaned them, then merged them together and calculated prescription and overdose rates per state per year. It should be noted that the prescription dataset used here encompasses only the first 10 million rows of the overall dataset. Therefore, any calculations or conclusions based on the prescriptions statistics may not be accurate. However, the process expounded below should work regardless of the final data. Now that our data is tidy, it's time to create visualizations in order to explore the data better.

After importing all necessary processing modules and files, my first action was to look at the "top offenders" for prescription rate and overdose rate with a bar graph. I decided to specifically use a horizontal bar graph, in part because our x-ticks, being the names of the states, were too long to be effectively displayed while avoiding text overlap and remaining aesthetically pleasing (in my opinion). Although I originally displayed all 51 regions (50 states plus D.C.) in my graphs, I decided to simplify our view to just the top 10 using the .head function. Examining these graphs reveals an interesting insight: the states with the most pills per person (New England) are not at all the states with the highest overdose deaths (Appalachia and the Southwest). Clearly, there is some other factor at play here.

I also attempted to create a stacked bar graph for prescription and overdose rates in the years 2006, 2009, and 2012. If successful, this would have demonstrated the growth of both rates over time. The top states did not have consistent growth over this time period - some only entered the top states in one year, others actually decreased over time. As a result, the stacked graphs were difficult to read, and perhaps not the best way to visually inspect the data.

Since I was attempting to see the change in data over time - a trend - it seemed that a line graph would be a more effective way to accomplish this. These graphs for both prescription and overdose rates show an overall upward trend, although both also decreased slightly between 2011 and 2012. Next, I wanted to see if I could find a positive correlation between average overdose rate and average prescription rate from 2006 to 2012 . At first I tried a line graph of overdose rate versus prescription rate,

which produced a graph with a strange downward hitch at the end, representing the decline in rates from 2011 to 2012. Thus I realized my mistake - when attempting to visualize correlations, the better method is not a line graph but a scatter plot. The succeeding scatter plot for these rates appeared to show this correlation more clearly, but I wanted a fit line to confirm this. I imported the stats module from the Scipy library, then used it to calculate the equation of the fit line for this data. This resulting graph showed a regression with a slope of 32.95.

I was wondering if I could reduce the prescription rates and overdose rates into a single quantity which could be used to quantify how deadly opioid prescriptions are in a given state. I dubbed this metric the “pain ratio” and calculated it by dividing overdose rate by prescription rate. Graphing this pain ratio over time gave another curious insight: even though both overdose rate and prescription rate increased from 2006 to 2012, the overall trend of the pain ratio is negative. This indicates that opioid prescriptions have actually become less harmful over time.

Putting the top 5 states for pain ratio into one graph mostly reveals a similar trend. The top state for pain ratio, Alaska, nearly breaks the scale of the graph. An explicit examination of the data for Alaska shows that in 2008 the state had an absurd pain ratio of about 300 in comparison to the average of about 5.4. Although it seems like an error, this number is well corroborated by the data - 94 deaths and 31 prescriptions, leading to respective overdose and prescription rates of 13.7 and 4.5. In fact, Alaska is the only state to have a higher overdose rate than prescription rate at any point in this time period, also accomplishing this in 2006. This could be a result of faulty data entry or due to the incomplete sample, as mentioned in the first paragraph above.

To display the range of values in the dataset, my first instinct was to create box and whisker plots of prescription and overdose rates. However, these proved to be ineffective at showing where the bulk of the values lie, so I also created violin plots. These new plots also showed outliers for all years well above average values, which we saw earlier in the bar graphs.

My final visualization was to create heat maps overlaid on a geographical map of the United States in order to demonstrate the disparity of values in different parts of the country. Thankfully, the Plotly module graph_objects can create such maps quite easily. In order for this function to understand the categorical regions of the data, it was necessary to create a dictionary of state names and abbreviations (used during the data wrangling step) and utilize a dictionary comprehension along with the .map function so that the state names would be converted back into abbreviations. The heat maps

confirm what the bar graphs told us: prescriptions are centered in New England and overdoses in West Virginia, Kentucky, and the Southwest.

Because prescription rates are not enough to predict overdose rates in a given state, perhaps the next step in this project is to integrate more demographic data and use it to build a machine learning model that can better predict rates based on a wider variety of background factors (i.e. income, sex, race, etc.).