Kusmana – DS 2002 Reflection Paper

A primary challenge I faced was the data cleaning process, followed by variable selection and additional research. The original format of the sentencing data set, which is directly obtained from ICPSR, is a .tsv file, and my following step is to convert it into .csv. Since the data contained 574 variables and over 76,000 observations, I had to optimize the data by filtering it to only contain relevant columns/variables. I had to consult the codebook and call the U.S. Sentencing Commission for help with choosing the correct variables, since some overwrite each other. This, and research to understand the Guidlines and how different factors interact with each other when it comes to creating sentencing outcomes, took up most of my time with the project – perhaps even more than the python coding (with learning new plotly packages), visualization, and presentation parts.

For the purpose of narrowing down the population of interest, I then the data to include only individuals convicted of trafficking cocaine, both crack and powder, and no other drugs. To avoid unbalanced data in the population, I further omitted defendants with a Criminal History Category of less than six, which is calculated based on their criminal history points. Finally, I removed an outlier because the observation has a cocaine quantity of 200,000 grams – and everything from that point are observations below 50KG.

Yet the unbalanced data still persists. Out of 1909 powder cocaine observations, only 53 defendants went to trial – which is 2.7%. Out of 1138 crack cocaine observations, 56 defendants went to trial – which is 4.8%. These figures are consistent with the national aveage, where only 2-4% of defendants convicted of federal offenses annually rejected plea deals and took the case to trial. This means when making the visualizations and graphs, I was not able to isolate observations based on Plea/Trial status and had to combine both groups together, which sometimes confounded the interactions between plots. To address this issue, I grouped the Plea/Trial group using different colors. For instance, in the weight vs sentence length scatterplot for each race, I coded red for Trial and blue for Plea. That way, I could still see how the two groups were sentenced differently.

Another challenge is displaying my visualizations from python. I used plotly to make interactive charts with hover information to show specific information about each observation – such as the type of drug, sentence length, and defendants' demographic information – which is especially relevant for the scatterplots. I was not able to display this in the google slides and jupyter notebook. For this reason, I included a code to convert the interactive charts into HTML (included in the GitHub), so user would be able to use the slider for the charts for each racial group and use the hover information to better understand each observation, namely the outliers.

A lesson I learned when it comes to technical challenge is using new python packages for plotly and data visualization (matplotlib). I had to consult stackoverflow and, for disclosure, ChatGPT to troubleshoot bugs in my codes. I learned that we could convert a python file into HTML, which allows the plotly interactive graphs to work. I would likely use this feature for dashboarding or for simple data analysis at work and future projects. In addition, if I could do the project differently, I would combine my charts (which are parts of EDA) with a simple regression analysis to see if my assumptions about Plea/Trial status holds true – that it is one of the most significant variables affecting sentence length, perhaps even more influential on the y-variable than drug weight.

I gained research and python skills while working on this project, primarily with integrating my knowledge of the Sentencing Guidelines, the law, and the data with data visualization, coding, and presentation. I also was able to practice my skill in explaining statistical and sentencing concepts to a general audience – during my internship, this was a valued ability because it allowed us to connect with shareholders and share our findings in an understandable way. I appreciate that in this class, we were able to work on that skill while also combining it with actual python and data analysis.