Evan Edmunds

DSC 530

Final Paper

My dataset contains the information of the total crime count in Russia, the counts of various subsets of crime categories, and has them all labelled by month from 2003-2020. The questions that I wanted to answer with this dataset is whether there exists a strong correlation between different crime categories and whether a regression model could be created with a crime category predicting total crime counts.

My analysis was proven for both questions. Through the creation of a correlation matrix and scatterplots between Total\_crimes and all other variables. These illustrated that the it is possible that a strong correlation exists between Serious crimes and Theft crimes. A hypothesis test of permuting the data confirmed that the correlation between the variables existed and was statistically significant, with a p-value of 0. The Regression model made for Theft predicting Total\_crimes was shown to be strong as well, with an R-squared value of .9023, explaining over 90% of the variance between the two variables.

I missed the opportunity of doing a time series analysis due to the time constraints of the course. I also believe that there were other analyses that could be done with a deeper dive into the data. I would have liked the chance to predict future values with my regression model, but I was, again, constrained by time. I don’t think that there are many other variables that would be useful to this dataset, other than additional subsets of crime to be added.

I had originally assumed that most of the data would be normally distributed, but within different ranges (rarer crimes would have smaller counts and vice versa). I was not expecting to find variables with Pareto-like distributions or bimodal distributions. The challenges I faced were primarily with the histograms. I could not figure out how to define the number of bins with the thinkplot file. This meant I had to learn how to do it using pyplot from matplotlib. I still don’t fully understand how to implement a time series, so I will be researching that on my own after the semester concludes.

I enjoyed working on this project and all the challenges with it.