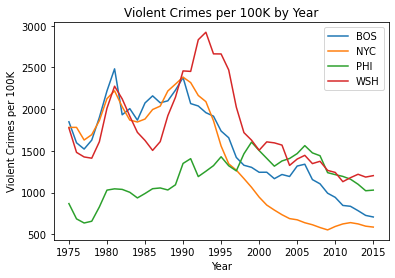
Frankie Way and Brian Waters

Final Project Report

Analyzing crime data can be a great way into looking into which cities in the United States are safe, which are cleaning up the streets, and which are falling by the wayside. To analyze crime data, we gathered a data source from data.world that compiled violent crime sums and standardized rates from 1975-2015. This data set provided us with information on numerous metropolitan areas, with standardized rates of homicides, rapes, robberies, and aggravated assaults and overall violent crime (per 100K of population), the sum totals of each measurement.

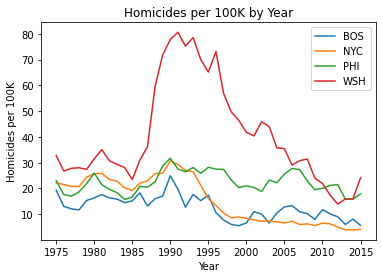
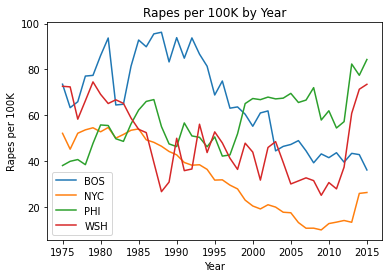
The data for the most part was already cleaned and ready for use, but in order to analyze our four selected northeastern cities (New York, Boston, Philadelphia, and Washington, D.C.) we had to do a little bit of manipulation to get the data ready for use. This data set was perfect because it provided clean data on a variety of cities that we could use to analyze crime trends over a 40-year time frame. We did have to sort the data by value and then we created 4 new data frames, one for each city. To analyze trends in violent crime we looked at the distribution for all cities. The distributions are clearly non-normal, allowing us to use non-parametric tests, to test for statistical differences.

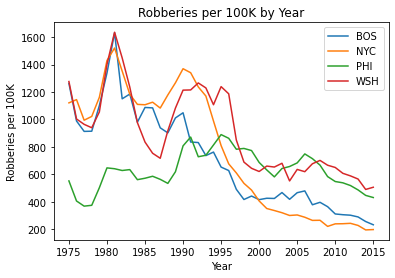
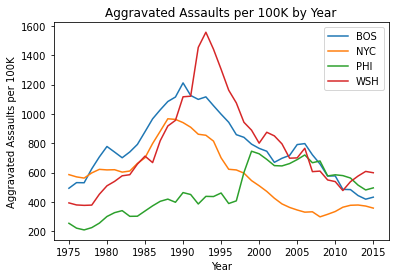
We used a test called the Wilcoxon Rank Sum test to see if there were differences in each cities crime trends (distributions). In simpler terms, the test results will show whether or not two cities crime trends are systematically different from one another. We ran 6 tests, one for each combination of cities and found 3 differences in distributions. Those being Boston/Philadelphia, New York City/Washington, D.C., and Philadelphia/Washington, D.C..

After identifying the differing distributions, we set out to see what type of crime, if any one in particular, caused the variability over the 40-year period. Using a spearman rank correlation test, we were able to identify the correlation between any type of crime in two cities over the designated time frame. The output from this test is a number between -1 and 1 that shows co-movement between two variables. Overall, we found that aggravated assaults and robberies were most influential (based on the correlation coefficient) in determining whether a distribution was statistically different from one another. This could simply be because they are weighted more heavily (the occur more often than rapes or murders).

We achieved our goal of analyzing crime in these four northeastern cities by performing two different types of hypothesis testing for non-parametric data. We were able to identify which cities had distributions that were unlike that of another city as well as determine which types of violent crime were correlated with one another. More often than not, murder trends were more likely to move in the same direction, while rape, aggravated assault, and robbery trends were either uncorrelated (meaning there was little to no co-movement between city data) or negatively correlated, meaning they would move in opposite directions (which would explain why their distributions of violent crime were systemically different from one another).

Plots of types of violent crime by cities:





Data source can be obtained at the following link:

<https://data.world/carlvlewis/u-s-metro-areas-violent-crime-rates-by-type-1970-2015>