Inferential Statistics for Capstone 1

The amount of inferential stats to be done with this data was limited at best, since showing how closely related two independent variables were was not an end goal in this project, and there were far too many independent to dependent variable combinations to test for significance of each one with t tests, so inferential statistics for the project is as follows: checking the crime rates, both violent and nonviolent are statistically significantly different from known national averages, seeing if there is a statistically significant difference between the rates of violent and nonviolent crime rates themselves, and in addition, looking at single correlation levels of variables that would be used later on in the machine learning portions.

Results analysis

A main focal point of the project is the specific dependent variables of violent crime per 100,000 people and nonviolent crimes since these are rates that can be easily compared regardless of actual population size.

I decided to run 3 different types of statistical tests detailed as follows, to try to understand the dependent variables better. The first is a T-test on two related scores. In this case, I am assuming that despite them not being literally the same variable, that there would be some relationship hence *related* between the rate of violent and nonviolent crimes. Here I assume, in that in general, cities with higher crime rate sin general would have equally high numbers of violent and nonviolent crime. The null hypothesis is that the rate of violent crimes and nonviolent crimes have identical values. After running the test, we see a relatively large t-statistic and a p-value of zero, which leads to the rejection of the null hypothesis. What this means is there is a statistically significant difference between the two, therefore a high violent crime rate doesn't necessarily mean that the town/city has a higher non-violent crime rate as well.

I reran the test with the assumption that the two crime rates are completely independent. Once again, we are testing for an equal average value. AS expected, we once again ran into a low p value of 0.0 leading to a rejection of the null hypothesis of averages being equal. However, it is interesting to note that I did end up with a slightly smaller t-statistic. This leads me to believe that it is better to interpret the two crime rates to be more independent from each other than related.

In addition to the two sample T-tests, I also decided to run some one-sample tests with each of the rates to determine how significant the rates were to Statista's national crime rate average for the year of crime data that we have, 1995, as well as their property crime rate which fits all the categories of nonviolent crime as a rate per 100,000. These natural averages are good to use to see if the dataset is a true representation of the nation. As the results indicates, it is different enough to reject eh null hypothesis that the sample is equal to the mean. However, it is interesting to note that despite the p-values being low enough to reject that the t-statistics are low, meaning that it is not obscenely far from the expected values compared to other possible data.

Last but not least, I decided to correlate the independent variables to the two dependent variables to find the pearson r's to find which variables seem to be more correlated to the end result of crime rates.

Strong correlations could mean that they would have a higher weight towards getting a closer result when doing predictions. This was important to sort through as there were hundreds of independent variables, and while the machine learning algorithms would do a lot of the work for me in determining the most important in the group, it would also be good to know for future data creation which variables are most important.