Capstone 1 - Statistical Analysis

To answer my primary research questions, I employed inferential statistical analysis to evaluate the relationship between my target variable, stop outcome, and six key features: driver gender, driver race, driver age, police department, weekday, and violation type.

Driver gender

I wanted to investigate if there was a significant difference in stop outcome ratios (written warnings: citations) among male and female groups that wasn't due solely to a class imbalance in my dataset. To do this, I first outlined my null and alternative hypothesis:

Ho: The ratio of female and male stop outcomes are the same.

Ha: The ratio of female and male stop outcomes are different.

Next, I generated 10,000 bootstrap replicates of stop outcomes for both males and females. Each replicate is the difference between the female ratio and the male ratio. I then calculated the lower and upper limit for a 95% confidence interval: [0.2936638, 0.349116]. Since this interval does not contain zero, I can reject the null hypothesis. To visualize these results, I plotted a histogram of these values with the percentiles marked.

To further test this relationship with more traditionally metrics, I repeated the same analysis using the difference in means rather than the difference in ratios.

Ho: The mean of female and male stop outcomes are the same.

Ha: The mean of female and male stop outcomes are different.

Since stop outcome is binary, I took the mean of the sum of stop outcomes for each gender, which allowed for a comparison in the quantity of written warnings for each group (written warnings = 1 in the hot encoded dataset). For this analysis, the 95% confidence interval was [0.04180389, 0.04934824], which again does not contain zero, and thus I was able to reject this null hypothesis as well. Similarly, I plotted a histogram of these results with percentiles marked. Both histograms appear to mimic a normal distribution.

Finally, to test the dependence of the two relationships, I ran a Chi-squared test. The results show that driver gender and stop outcome are dependent, with $p = 0.000$.

Driver Age

Because I split driver age into three categories: early, middle, and late, the best statistical test to compare age group and stop outcome is a chi-squared test. The results show that the two variables are dependent, with $p = 0.000$.

Driver race, Police department, weekday and violation type.

I ran four more chi-squared tests with these variables against stop outcome to ensure a relationship between the features and my target variable. All four were found to be dependent with $p = 0.000$.

Because most of my variables are categorical, the statistical tests are less intensive in nature. However, these results have given me valuable reassurance that my chosen features have a dependent relationship with my target variable.