Only a handful of variables which provided more insightful data during the EDA section were selected for statistical analysis. These variables showed a more promising relationship with the *Severity* variable than some of the others. This is a list of the variables:

* Day of the week
* Hour
* Month
* Date
* Address type
* Junction type
* Speed
* Road length
* Road congestion

Since the target variable, *Severity Description*, is a categorical variable, two different statistical inference techniques were used - the **Chi-squared test for independence for comparing two categorical variables** and **difference of mean two-sided test for comparing a categorical variable to a numerical variable**. The Jupyter Notebook for the analysis can be found [here](https://github.com/saychelsea11/Springboard-Projects/blob/master/Capstone1_Project/Statistical_analysis.ipynb).

1. **Inference for comparing categorical variables using Chi-squared distribution**

The chi-squared test for association tests the null hypothesis that the two categorical variables are independent. This test compares the observed data to a model which assumes that the variables are independent. As a result, if the hypothesis fails, the variables are proven to be not independent of each other. This could also indicate a possible relation between the two categorical variables and could help in selecting statistically significant variables for the final model. The procedure for this test is as follows:

* State null and alternate hypothesis
* Find the degrees of freedom
* Calculate expected frequency count of sample variables
* Calculate the Chi-Squared statistic
* State the significance level
* Determine the p-value for statistical significance test

As an example of the procedure, we used the *Day of the Week* variable against *Severity Description*. In this case, the null hypothesis stated that the severity and rate of collisions during certain days of the week are independent. The alternate stated that severity and rate of collisions during certain days of the week are not independent.

The degrees of freedom (6 in this case) is determined from the number of rows and columns in the contingency table. The expected frequencies of the variables are calculated for each category and compared with the observed frequencies which finally yields the Chi-Squared statistic, which came out to be 6.0895. Using this value and the degrees of freedom, the p-value was determined to be 0.456377. Since this value was greater than the significance level of 0.05, the null could not be rejected indicating that the two variables were independent.

Using the same technique, the *Month* variable was found to be dependent with *Severity* with a p-value of 0.004. The Hour variable had a p-value of less than 0.00001 which also was a statistically significant result. Both Address Type and Junction Type were also statistically significant variables with p-values of less than 0.00001 as well. The *Date* variable that stores the days of the month, yielded a p-value of 0.99, which was not statistically significant.

1. **Inference for comparing severity with numerical variables**

This technique is useful for comparing the means of the severity categories across different numerical variables to determine if any of the variables can explain the variability in severity. For this process, three different techniques were explored:

* *Bootstrap sampling simulations* - sampling with replacement
* *Normal distribution sampling simulations* - assumes data is normally distributed
* *Permutation sampling simulations* - no prior assumption regarding distribution of data

The statistical relation between the *Speed* variable and *Severity* was explored using all three techniques and it was determined that all of them yielded similar results whereby validating each approach. For analyzing the other variables, permutation sampling was used since it does not make any assumptions regarding the distribution of the data.

This technique combines two samples and randomly sorts the data, following which the data is split into two permutation samples with the same size as the original two sets. This ensures that the two sets have the same mean, used as the condition for the null hypothesis. The sampling is done repeatedly and a sampling distribution of the difference of means is created.

Outline of the process:

* Null hypothesis: The difference in mean of the two samples is 0
* Alternate hypothesis: The difference in mean of the two samples is not 0
* Data is split into severe and non-severe sets
* Permutation sampling is conducted
* P-value is determined for statistical significance

Inference for the *Speed* variable yielded a p-value of 2.055e-13 which was statistically significant and indicated a possible distinction between the severity categories. *Road Length* and *Road Congestion* had p-values of 0.1344 and 8.7887e-06 respectively with only the latter being statistically significant.