**Statistical Question/Hypothesis:**

For my hypothesis, I wanted to find out if percentages of driver involved in fatal collisions who were alcohol-impaired may have influence on the percentages of driver involved in fatal collisions who were speeding.

**Outcome of your EDA:**

The outcome of my EDA was that there was no significant relationship between these two variables since the p-value was 0.491 which is > 0.05, it failed to reject the null hypothesis. The percentages of driver involve in fatal collisions who were speeding will can increase or decrease in the event of with or without alcohol-impairment. However, as for correlation from scatter plot result, there is a what seems like a very weak positive correlation between these two variables, but it is to weak and there are many other confounding factors to determine any significance or infer causation between these two variables.

**What do you feel was missed during the analysis?**

There are other variables in this dataset that I thought it was irrelevant to my hypothesis, so I did not include it in the data analysis such as insurance premiums, state and losses incurred by insurance companies for collisions per insured driver variable. With the addition of those variables, it may show some causation from any unforeseen correlation. Aside from that, I feel like I did not fully analyze my dataset by individual State since the percentage between these variables may vary between different States. For some States, the percentage between these variables may correlate while others may not.

**Were there any variables you felt could have helped in the analysis?**

The variable State could have helped in my analysis if I included, because there may be variations in percentages among States. Some may have high percentages, and some may have lower percentages which can skew the datasets. As a result, it may create a false negative error in my analysis. By including State variable, I feel like the analysis would be a lot better and I could have gotten a full picture on how percentages of my variables differ among different States.

**Were there any assumptions made you felt were incorrect?**

I assumed that other factors like NotDistracted, and NoPrevAccidents could have play a role in influence the number of fatal collisions per billion miles. In addition, I assumed people will have a higher chance of speeding with alcohol impairment, but it turns out there were no correlations, or I did not see the correlation as much as I expected. Also, there may be some outliers left behind even after I cleaned the data.

**What challenges did you face, what did you not fully understand?**

I understood and was able to apply Python in the different statistical calculations such as PMF, CDF, Analytical Distribution, Hypothesis Testing, Regression Analysis, etc. The thing I did not fully understand is sometimes I get this settingwithcopywarning for Pandas that I am trying to slice and dice the dataset, but I did not do anything to the original dataset. The warning sometimes appears and disappears whenever I re-run the code which is something I don’t understanding. This dataset from Kaggle has helped me by applying theoretical concepts into real world applications and overall, this class has given me a chance to use a lot of the Python packages I never used before to be able to do EDA from importing the data to run regression analyses.