Final Project Summary

For my Statistical question/hypothesis I decided to try and find the relationship between the percentage of drivers Involved In fatal collisions who were alcohol-impaired and the percentage of drivers Involved In fatal collisions who were speeding. While the EDA did not find much correlation, it was interesting looking at the scatter plots and seeing a concentration of data plots towards the 25-35 percent range for fatal accidents in impaired drivers. The question and variables I picked had the highest correlation, but the overall correlation was low as to point out low significance or causation.

My chosen tests and regression analysis could be better if I was more experienced in choosing the correct test to find better significance in the variables. What I did not understand though was the significance of my p-value in my Permutation test. I got a p-value of 0.485 and am unsure if this is showing me a lesser or greater significance in these variables. There may be other variables I did not use that could have helped such as the car Insurance premiums variable or the losses incurred by insurance companies for collisions per insured driver variable. These could show some causation from unforeseen correlation.

I don’t think there were many assumptions other then the chosen tests and regression analysis that could have been incorrect. There were some large challenges to this assignment. The first being that it was the first assignment that intended for us to use our own data and combine methods we used in the book and what we learned to do EDA. Besides not understanding the significance of the Permutation test I believe the resources we needed to guide us through the project was there for us. Only a little googling needed. Overall this class has been very useful and challenging for me as a student and someone who wants to learn more about being a data scientist.