I completed the following works in this project.

1. Data manipulation

2. Pre-analysis

3. Logistic regression

4. KNN algorithm

5. Wilcoxon sign rank test

Data manipulation

This project used the accident, vehicle, and person datasets in Fatality Analysis Reporting System (FARS). These three datasets need to be merged to create the analytical dataset.

YEAR and ST_CASE are the key variables to merge the accident and vehicle datasets. YEAR, ST_CASE and VEH_NO are the key variables to merge the vehicle and person datasets.

Reference: myself. I worked for the National Highway Traffic Safety Administration for eight years.

Pre-analysis

This is a part of exploratory data analysis. The bar chart, histogram and box plot were to present the frequency and the rate of rollovers in different features.

The chi-square test and one-way ANOVA were used to examine the significant association between the target and features.

Reference: myself. I have a master degree in Statistics (University of Maryland, 2012) and a master degree in Mathematical Finance (University of Pittsburgh, 2008), and I know what test I should use.

Logistic regression

The pd.get_dummies was applied on the categorical variables when building the logistic regression model. The sm.Logit can display the full results (estimated parameter, z-statistic, and p-value), but LogisticRegression can only provide the accuracy score.

Reference: https://towardsdatascience.com/logistic-regression-model-fitting-and-finding-the-correlation-p-value-z-score-confidence-8330fb86db19

About 30 percent of the Python code was learned from the above website.

KNN algorithm

The preprocessing.LabelEncoder can create the artificial distance for categorical variables when using the KNN algorithm.

Reference: https://faculty.nps.edu/sebuttre/home/Research/KnnCat/ordsdoc.html

I learned the KNN algorithm from the above website. There is no Python code in the above link.

Wilcoxon sign rank test

This test was used, since the probability distribution of the accuracy scores was unknown. Wilcoxon sign rank test is a non-parametric test, and its testing power is usually low.

Reference: Myself.