
# US Crash Accidents

## Capstone Sprint 2

By : John Narvaez

Date: Aug 18, 2021

# Prelude

← NEWS

## NHTSA: Traffic Crashes Cost America $340 Billion in 2019

Agency releases new study examining the cost of motor vehicle crashes, injuries and fatalities

January 10, 2023 | Washington, DC

Motor vehicle crashes cost American society $340 billion in 2019, the National Highway Traffic Safety Administration announced today. The agency's new report, "The Economic and Societal Impact of Motor Vehicle Crashes, 2019," examines the costs of one year of crashes that killed an estimated 36,500 people, injured 4.5 million, and damaged 23 million vehicles.

# Prelude

Traffic crashes cost taxpayers $30 billion in 2019, roughly 9% of all motor vehicle crash costs. This is the equivalent of $230 in added taxes for every household in the United States.

These losses include medical costs, lost productivity, legal and court costs, emergency service costs, insurance administration costs, congestion costs, property damage, and workplace losses. These figures include both police-reported and unreported crashes.

When quality-of-life valuations are considered, the total value of societal harm from motor vehicle crashes in 2019 was nearly $1.4 trillion.

The report includes new data on the total value of seat belt use. From 1975 to 2019, seat belt use saved 404,000 lives and prevented $17.8 trillion in societal harm.

# Overview:

- What is my dataset ?

Row - 7M Columns - 46

Data collected from February 2016 - March 2023

Objective :

To determine if some road features can predict the Severity of accidents. Ranked 1 to be the least up to 4 being the severe one.

# Road Features:

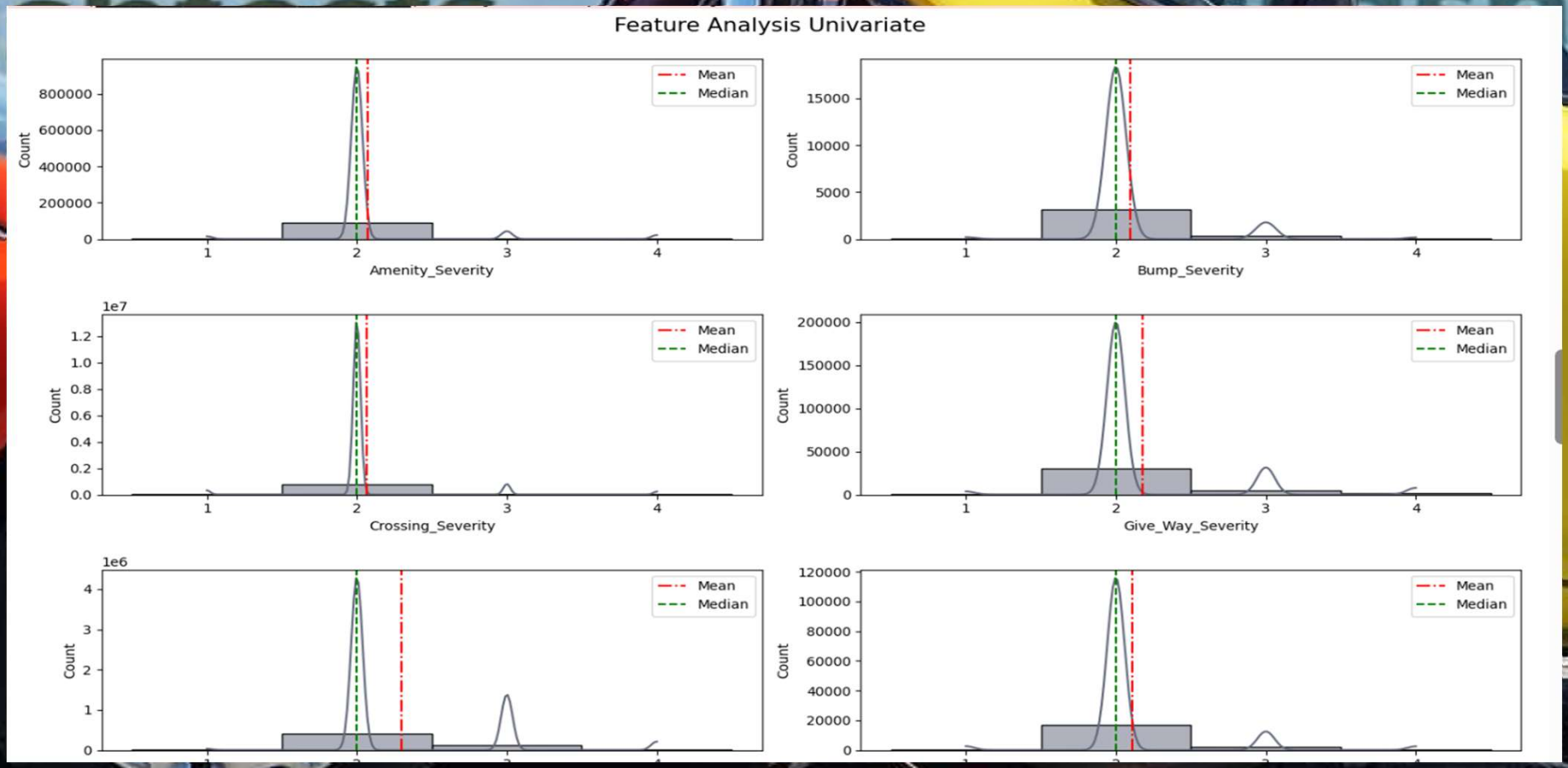


# Some Data Preparation

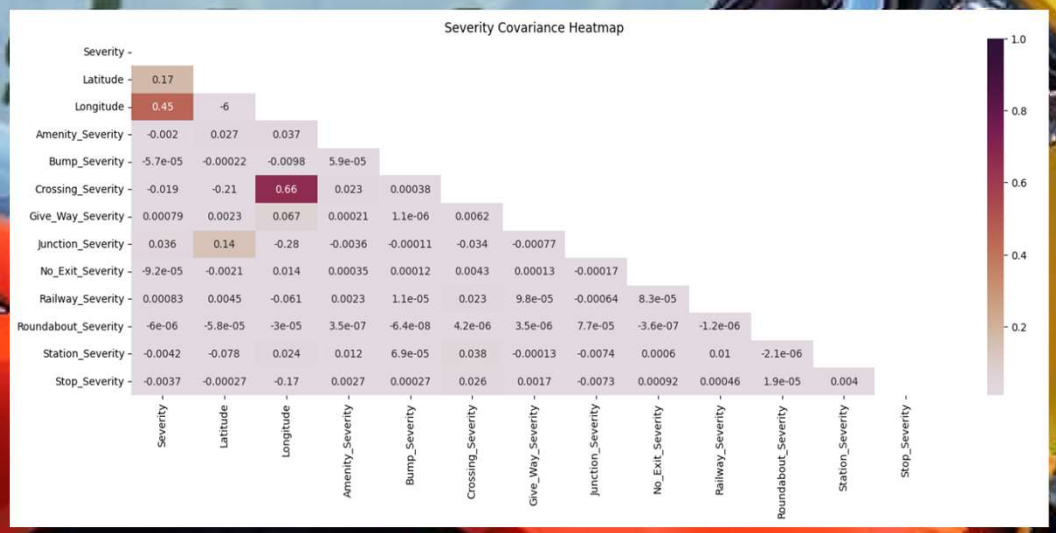
Added new combined features

| Severity | Bump | Crossing | Create | Bump Severity | Crossing Severity |
|---|---|---|---|---|---|
| 1 | True | False | | 1 | 0 |
| 4 | False | True | | 0 | 4 |
| 3 | False | True | | 0 | 3 |
| 1 | True | True | | 1 | 1 |
| 2 | True | False | | 2 | 0 |

# Feature Analysis



# Covariance and Correlation





There seem **no strong multicolinearity** in independent features.

But noticed that some have same rating.

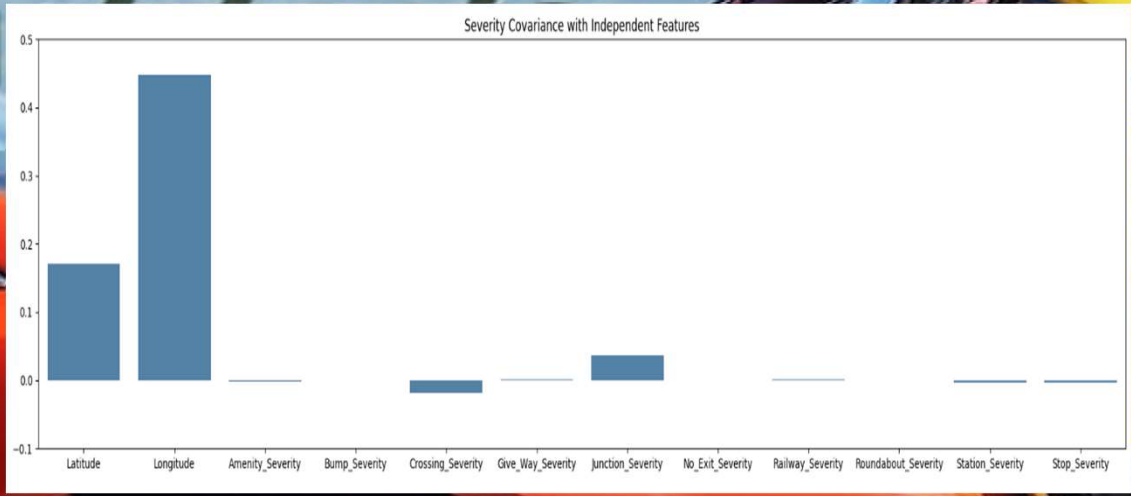
**1.Railway_Severity & Station_Severity** - 0.15

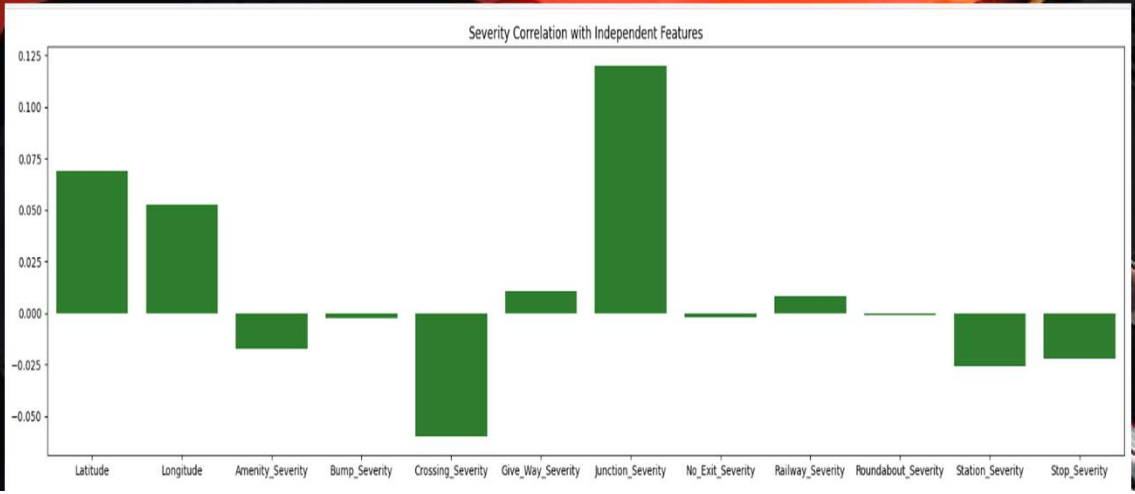
**2.Crossing_Severity & Station_Severity & Railway_Severity** - 0.17

**3.Amenity_Severity & Crossing_Severity** - 0.15

First two relates possibly about a train station nearby.

# Covariance and Correlation




- **Positive correlation are:**
- Give_Way_Severity,
- Junction_Severity,
- Railway_Severity,
- Latitude, Longtitude

- **Negative correlation are:**
- Amenity_Severity, Bump_Severity, Crossing_Severity, No_Exit_Severity, Roundabout_Severity, Station_Severity, Stop_Severity ty'

# Detecting MultiCollinearity

| | Features | VIF Scores |
|---|---|---|
| 0 | Severity | 1.243907 |
| 1 | Amenity_Severity | 1.052348 |
| 2 | Bump_Severity | 1.001426 |
| 3 | Crossing_Severity | 1.216149 |
| 4 | Give_Way_Severity | 1.009367 |
| 5 | Junction_Severity | 1.096052 |
| 6 | No_Exit_Severity | 1.006963 |
| 7 | Railway_Severity | 1.055067 |
| 8 | Roundabout_Severity | 1.000170 |
| 9 | Station_Severity | 1.088753 |
| 10 | Stop_Severity | 1.041732 |

Ordinary Least Square Result Severity

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Severity | **R-squared:** | 0.018 |
| **Model:** | OLS | **Adj. R-squared:** | 0.018 |
| **Method:** | Least Squares | **F-statistic:** | 1.409e+04 |
| **Date:** | Wed, 16 Aug 2023 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 11:21:27 | **Log-Likelihood:** | -5.3361e+06 |
| **No. Observations:** | 7717272 | **AIC:** | 1.067e+07 |
| **Df Residuals:** | 7717261 | **BIC:** | 1.067e+07 |
| **Df Model:** | 10 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 2.2067 | 0.000 | 1.13e+04 | 0.000 | 2.206 | 2.207 |
| **Amenity_Severity** | -0.0119 | 0.001 | -15.665 | 0.000 | -0.013 | -0.010 |
| **Bump_Severity** | -0.0135 | 0.004 | -3.507 | 0.000 | -0.021 | -0.006 |
| **Crossing_Severity** | -0.0368 | 0.000 | -134.428 | 0.000 | -0.037 | -0.036 |
| **Give_Way_Severity** | 0.0474 | 0.001 | 41.732 | 0.000 | 0.045 | 0.050 |
| **Junction_Severity** | 0.0909 | 0.000 | 321.731 | 0.000 | 0.090 | 0.091 |
| **No_Exit_Severity** | 0.0089 | 0.002 | 5.537 | 0.000 | 0.006 | 0.012 |
| **Railway_Severity** | 0.0476 | 0.001 | 54.737 | 0.000 | 0.046 | 0.049 |
| **Roundabout_Severity** | -0.0895 | 0.015 | -6.125 | 0.000 | -0.118 | -0.061 |
| **Station_Severity** | -0.0213 | 0.001 | -39.978 | 0.000 | -0.022 | -0.020 |
| **Stop_Severity** | -0.0175 | 0.001 | -34.612 | 0.000 | -0.018 | -0.016 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 2623656.319 | **Durbin-Watson:** | 1.431 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 7876549.707 |
| **Skew:** | 1.802 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 6.391 | **Cond. No.** | 89.7 |

**Based on the p-values** associated with every variable, we can see that all features seem **significant in predicting the Severity rating** (p-values are > 0.05 are considered insignificant).

**$R^2$ is 0.018** has a very small value explaining the variance of Severity.
The **coeff** are almost close to the correlation shown before except that the **No_Exit_Severity** is on the **positive side.**

# Issues

In summary, the covariance direction is hardly to determine as the independent variable are closely situation to zero. Likewise also difficult to interpret the correlation looking at the heatmap; which features are correlated.

So I decided to use the Variance Inflation Factors (VIF) formula. In a perfect multicollinearity the value should be 1. Result values are below 2 however closely nearing to 1.

That is moderately or slightly correlated (Multicollinear) between features.

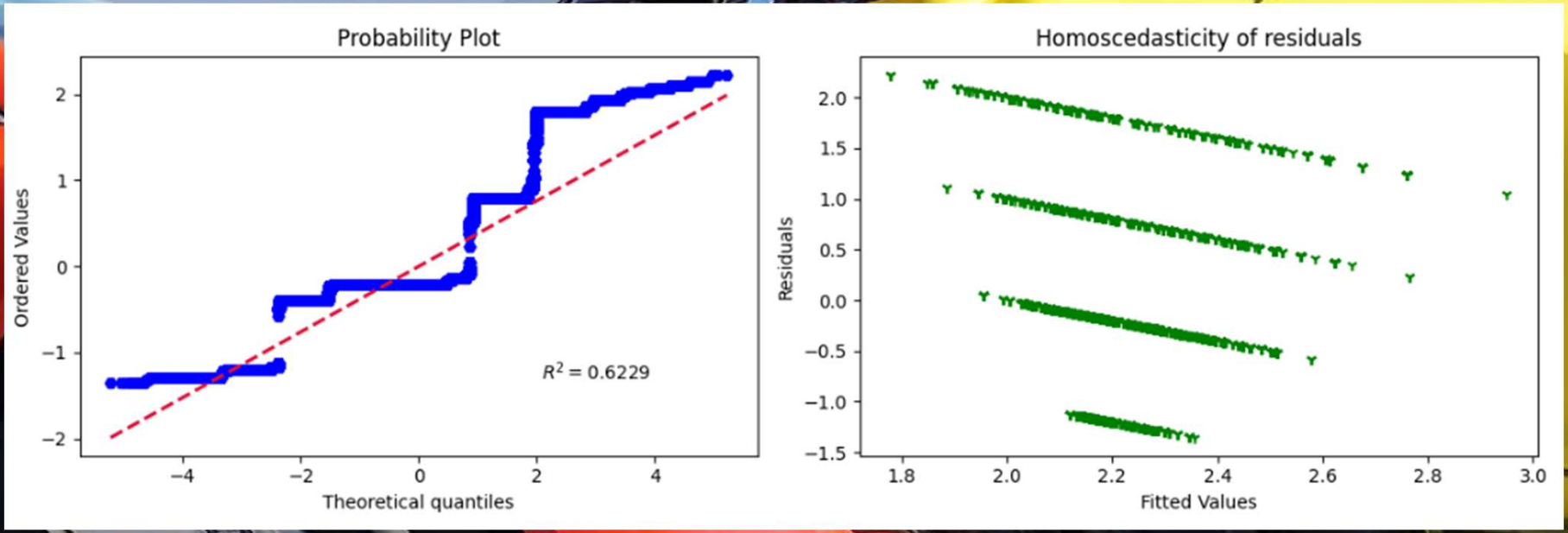
# Issues

Build a regression model using the statsmodel.api Linear Regression $R^2$ value is 0.018 that could explain the variation in the target variable (Severity). The coefficient are near zero value. The p-value is less than 0.05 that suggest independent variables are significant in determine the Severity.

The probability plot (q-q plot) are not uniform. Ordered values matches the Theoretical quantile at -3, -1, 1 and close to 2.

With the Homoscedacity plot is also difficult to interpret and my research with such is subjective. Comparing with other that such is not a violation of homoscedacity.

# Probability and Residuals



# Summary

In this first try of understanding the data and perhaps there are things that I did not consider or overlooked. The $R^2$ square is abnormally low.

Steps:

1. Re-check again the data preparation.

2. Try other than the statsmodel module and try other modelling.