

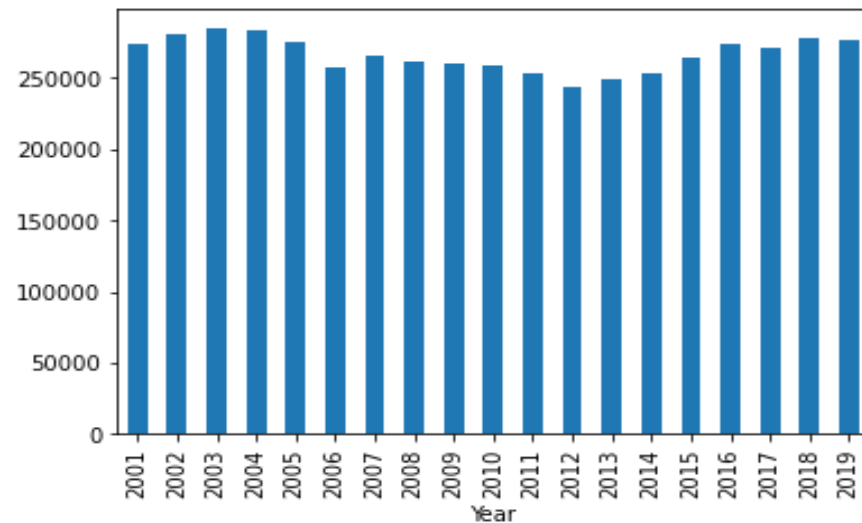
The Data Analysis of Traffic Accidents in Camden County, New Jersey (2017-2019)

Ning Shangguan

Data Source of Traffic accidents in New Jersey: Department of Transportation

https://www.state.nj.us/transportation/refdata/accident/crash_statistics.shtm

Total Crashes 2001-2019

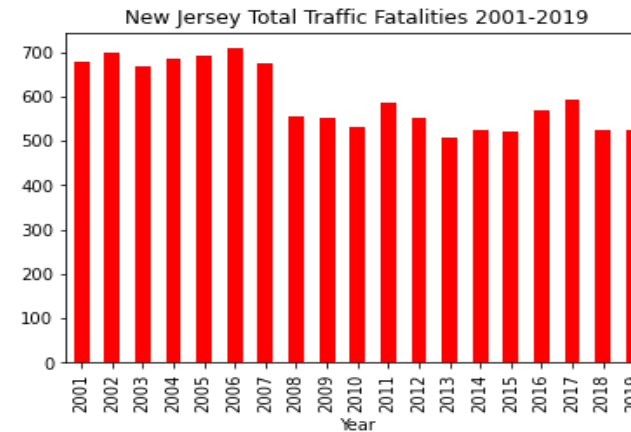
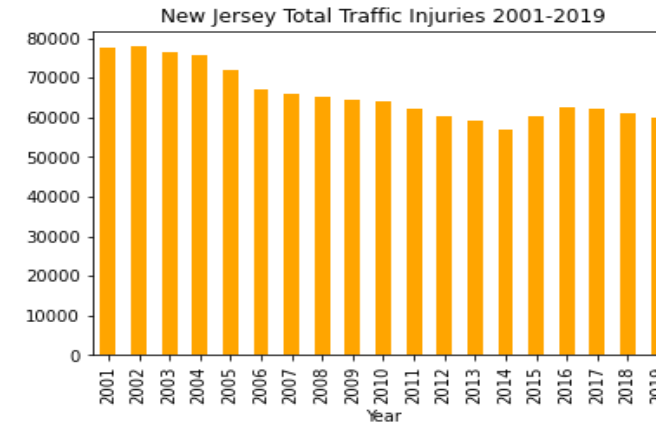


Year 2001: 274,110

Year 2019: 276,861

Population 2001: 8.49 million

Population 2019: 8.88 million



Year 2001: 77,397 injuries, 667 fatalities

Year 2019: 59,850 injuries, 524 fatalities

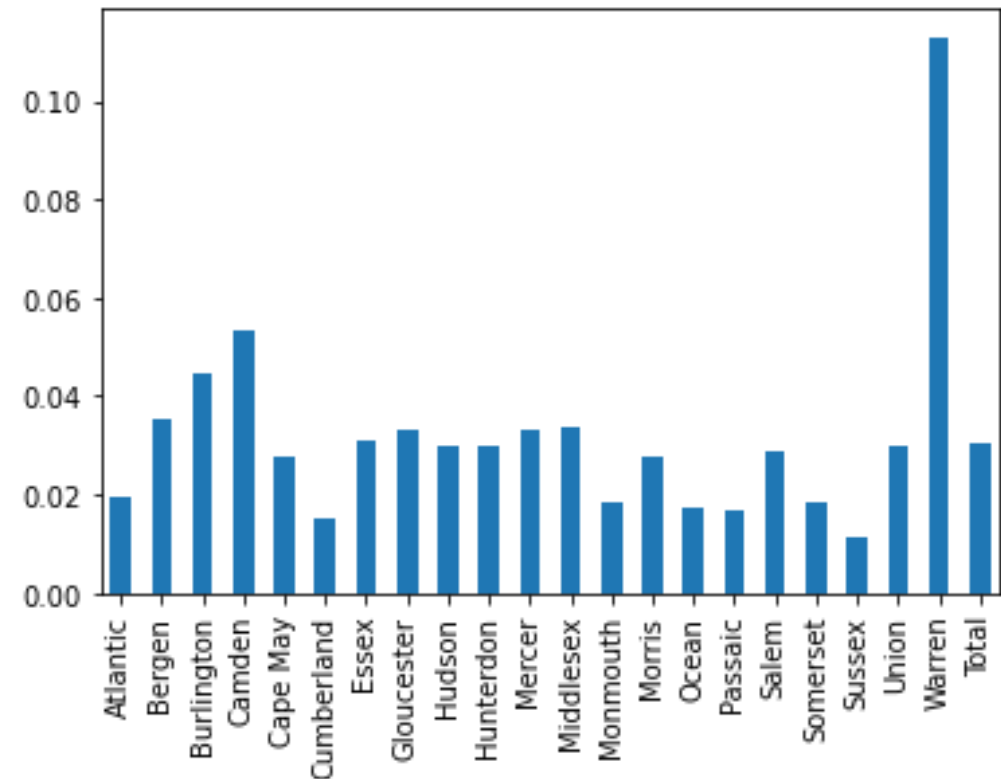
The Problems with the Kaggle Dataset:

The Kaggle website provides a 3 million traffic accidents of US from February 2016 to December 2020 at <https://www.kaggle.com/sobhanmoosavi/us-accidents>

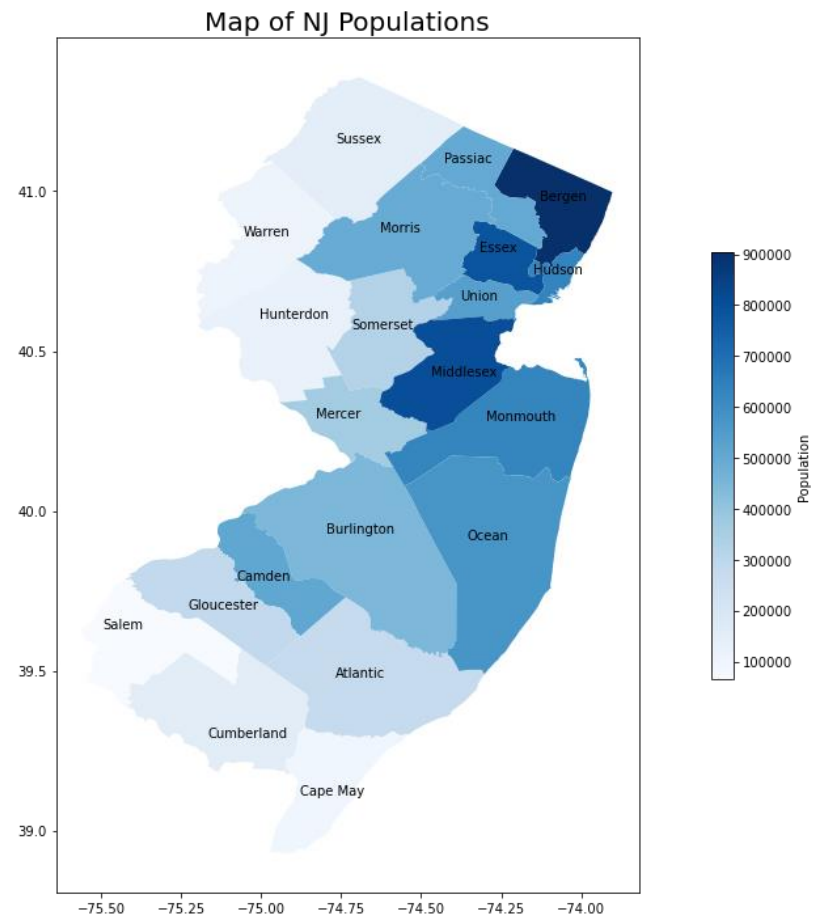
Problems:

1. It is estimated by the NHTSA that 6 million car accidents happen in the U.S. each year.
(<https://cdan.nhtsa.gov/tsftables/National%20Statistics.pdf>)
2. In Year of 2019 alone, NJ DOT reported 276,861 traffic accidents. The Kaggle dataset only has 8,435 cases. The sampling ratio is only 3.05% .
3. The sampling ratio of each county varies significantly. There are 21 counties in NJ. The highest (Warren) is 11.3% while the lowest (Sussex) is only 1.15%.

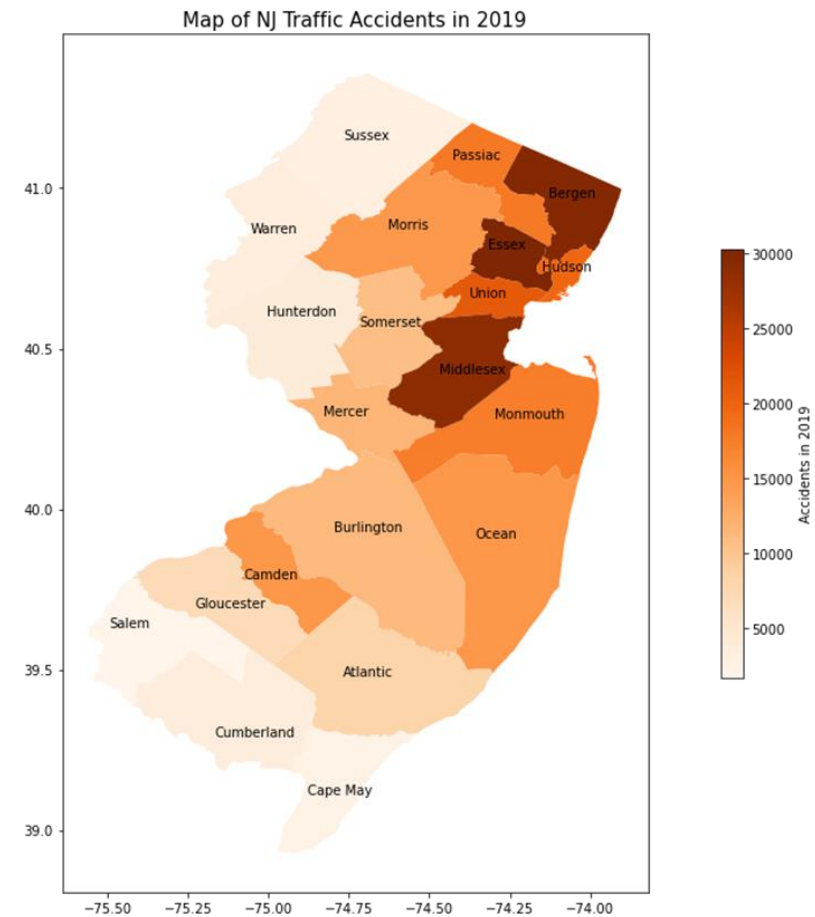
Kaggle Sampling Ratios of NJ Counties



NJ Population and Traffic Accidents



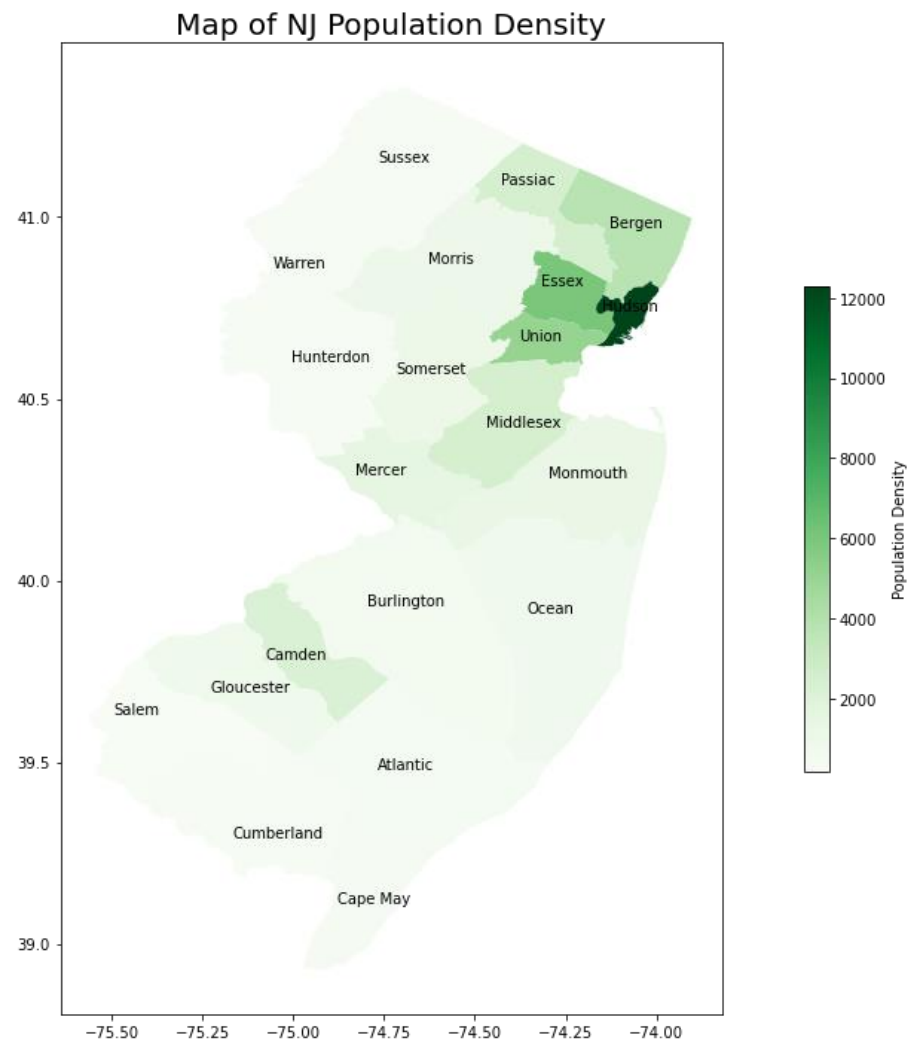
Bergen: 930K, Salem: 63K



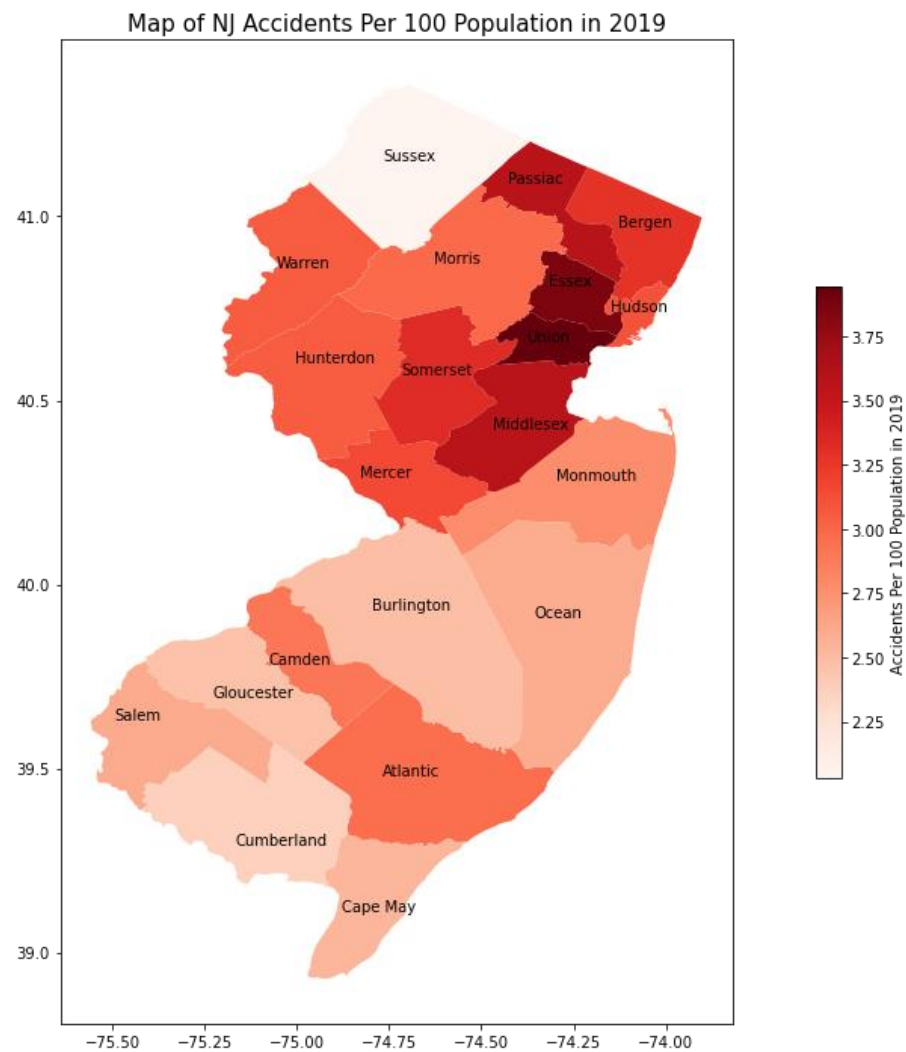
Bergen: 29.7K, Salem: 1.7K

NJ has a population of almost 9 millions and an area of 8722 square miles (22,588 square kilometers), making it the most densely populated state in US.

Population Density and Accidents per 100 Populations

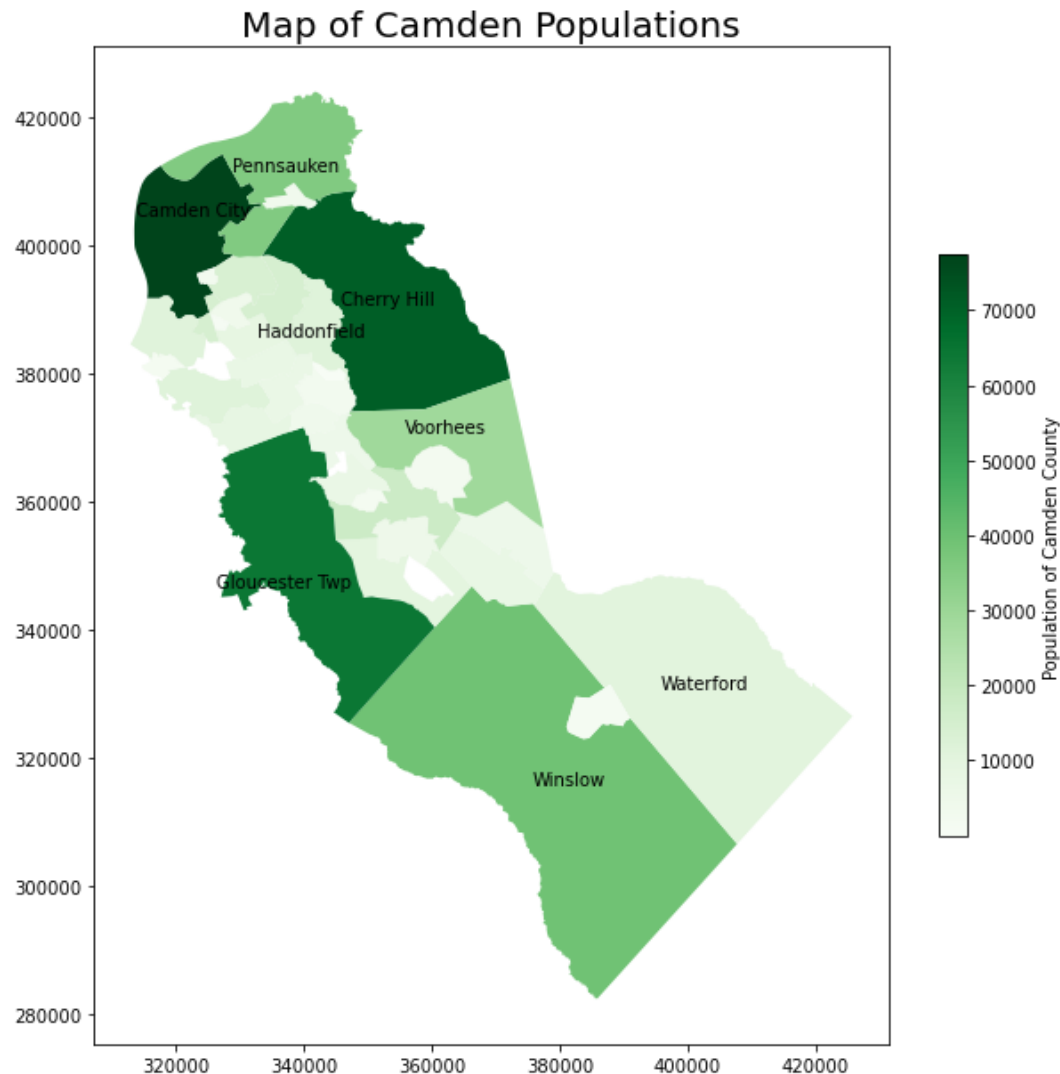


Highest: Hudson 14000 per sq. miles
Lowest: Salem 188 per sq. miles
Camden: 2290 per sq miles



Highest: Union 3.95
Lowest: Sussex, 2.03
Camden: 2.91, Hudson: 3.11

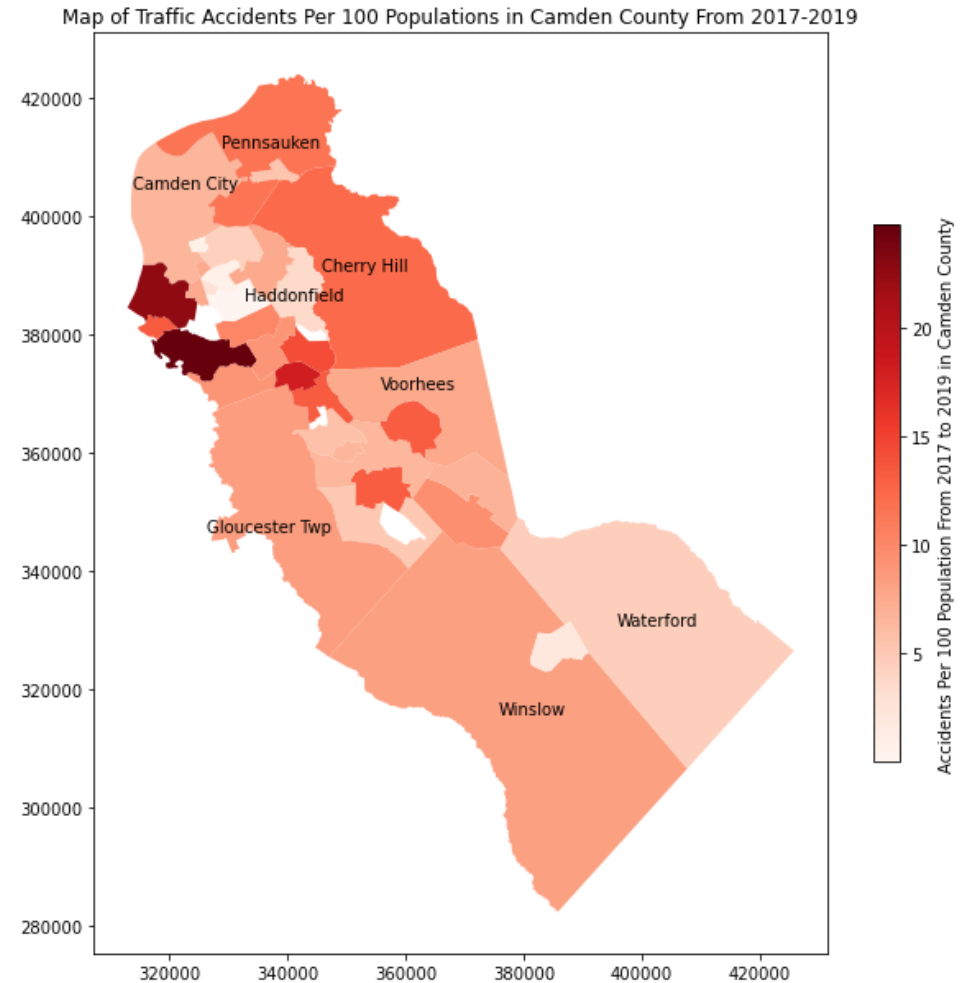
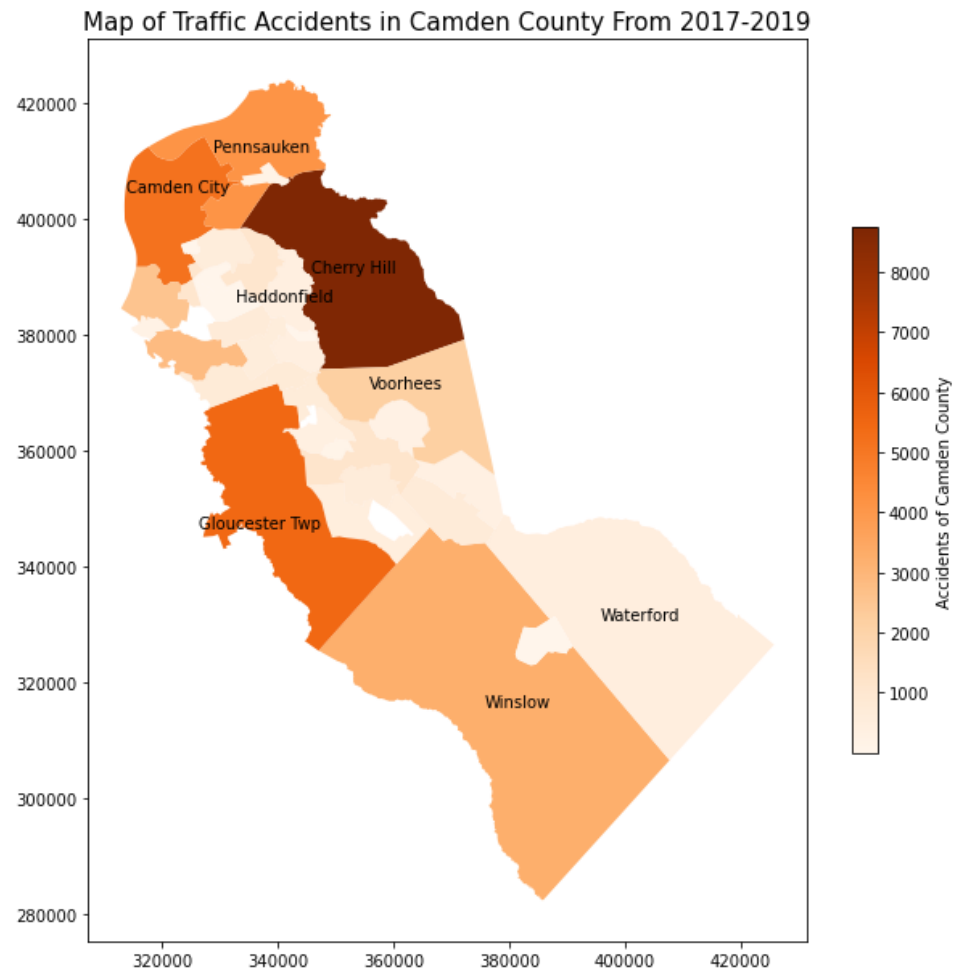
Camden County of NJ:



Camden County has a population of 506,471, an area of 227.293 sq mi (588.69 km²).

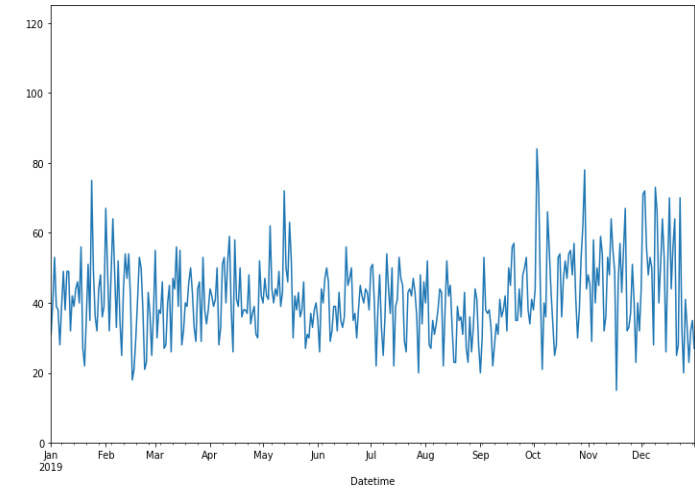
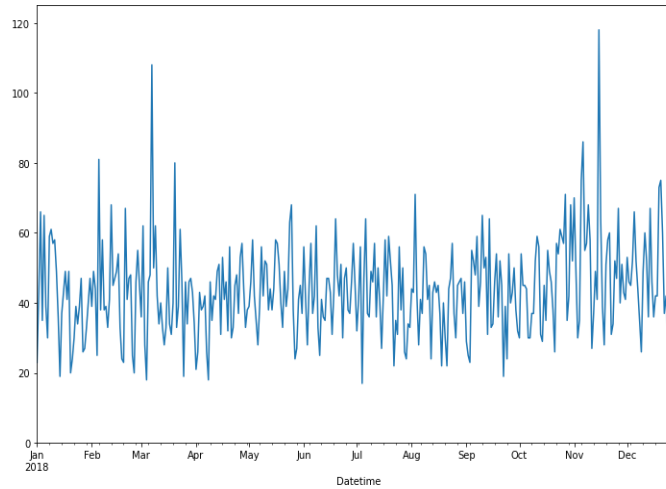
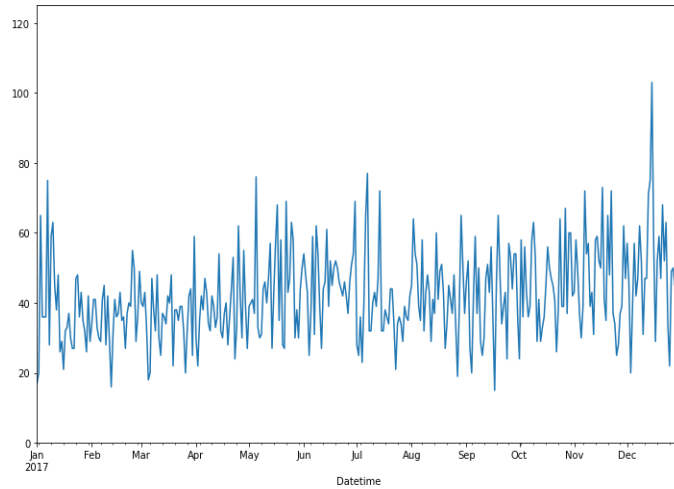
Camden City (pop.: 77,000), Cherry Hill (pop.: 71,000) and Gloucester Twp (pop.: 64,000) are 3 largest municipalities.

Traffic Accidents in Camden County:



The total accidents of Camden County during 2017-2019 are 15,176, 15,755 and 14,950. The traffic injury cases are 4,003, 4,088 and 3,784 while fatal cases are 40, 42 and 45.

Daily Accidents in Camden County (2017-2019):



Counted from the dataset from NJ DOT: 46519 cases from 2017-2019

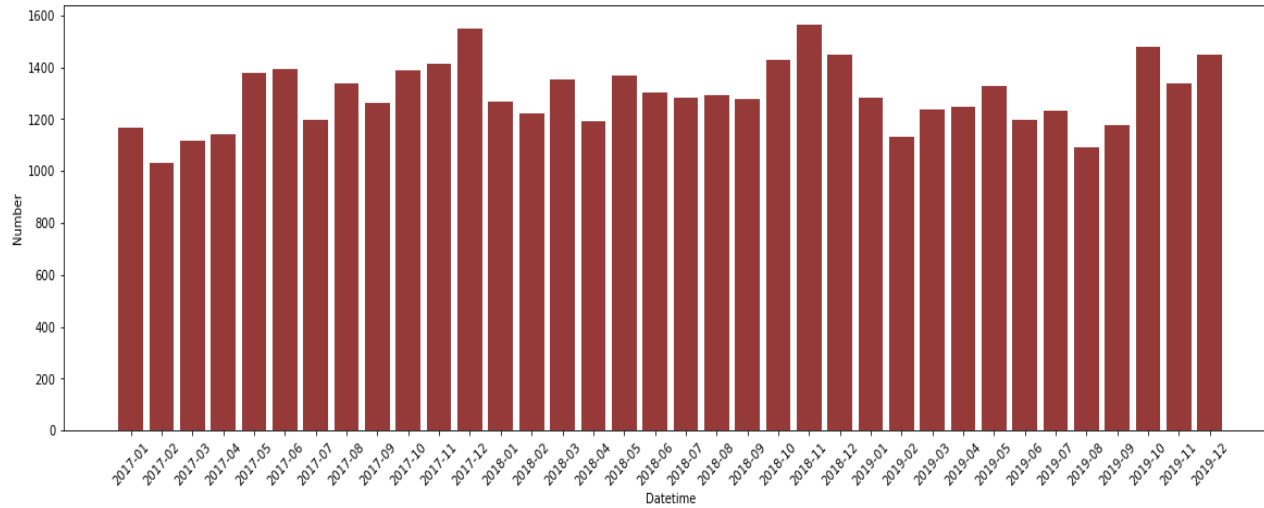
Daily Average: 42.5 per day

Highest 3 days during 2017-2019: 2017-12-15: 103, 2018-03-07: 108, 2018-11-15: 118
Most of the cases reported snow conditions.

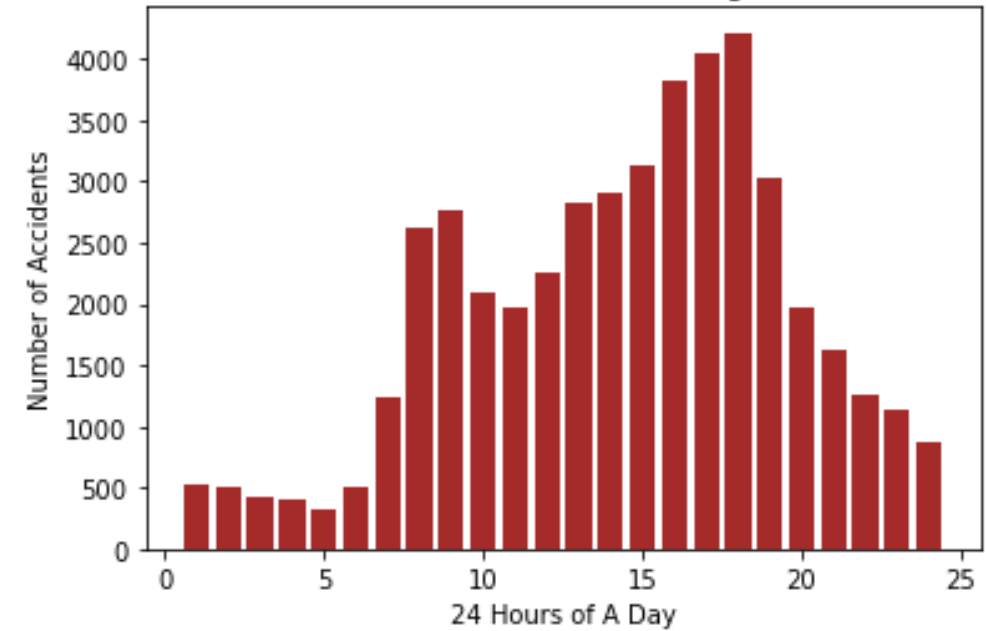
Lowest 4 days during 2017-2019: 2017-09-17 (Monday, 15 cases), 2019-11-17 (Sunday, 15)
2017-01-01 (17), 2018-07-04 (17)

Accidents vs. Time

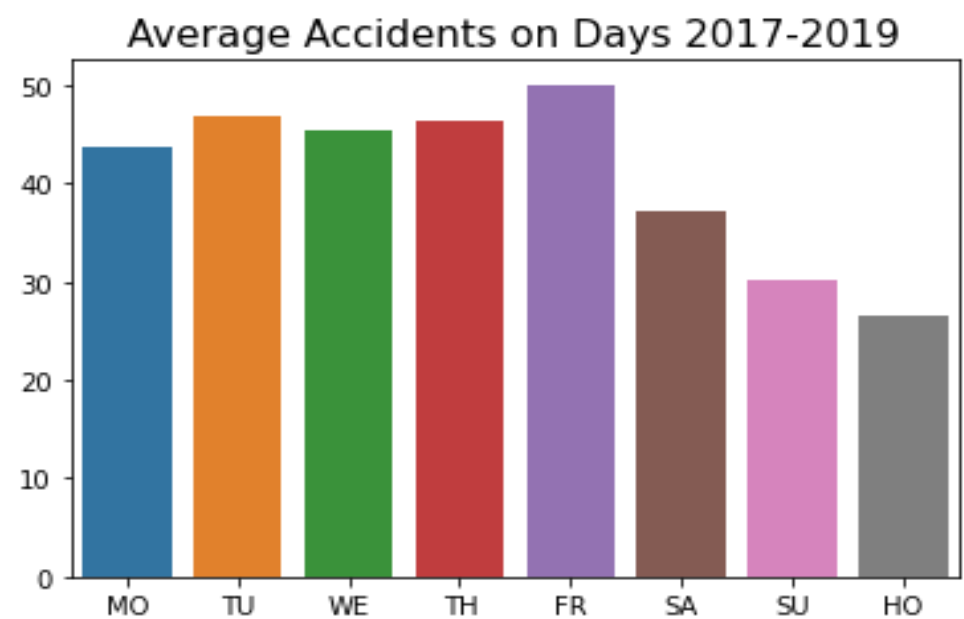
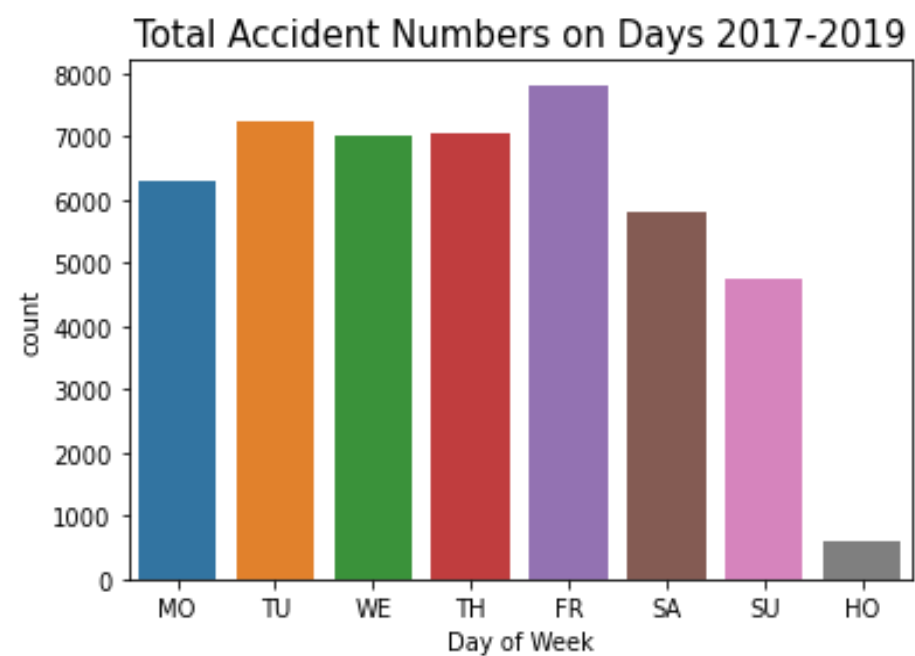
Accident Numbers in Each Month during 2017-2019



Total Accidents in Each Hour During 2017-2019

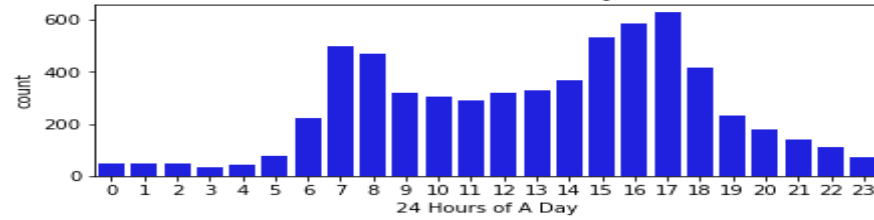


Accident Occurrences In Different Days of Week

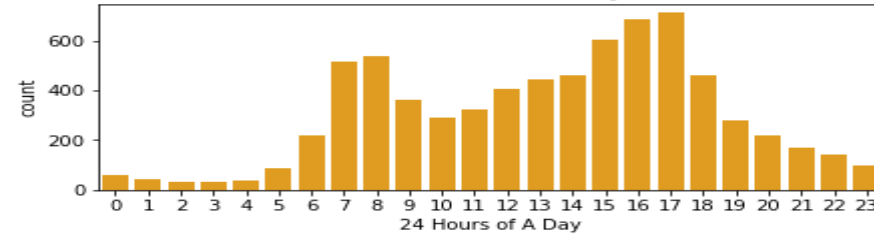


Accidents on Different Days

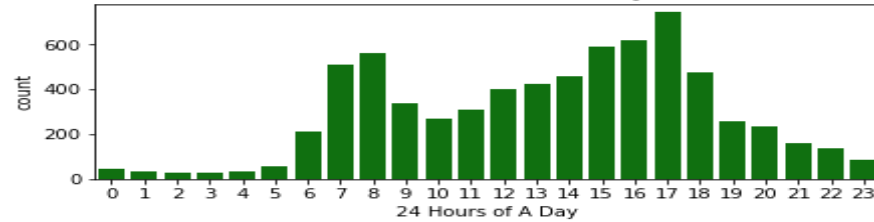
Accident Numbers on Mondays 2017-2019



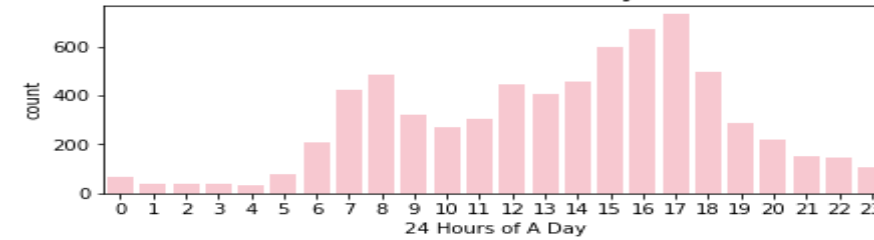
Accident Numbers on Tuesdays 2017-2019



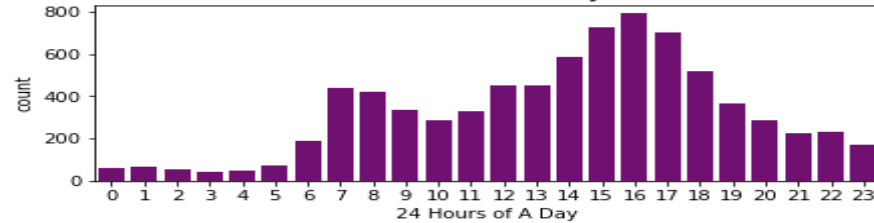
Accident Numbers on Wednesdays 2017-2019



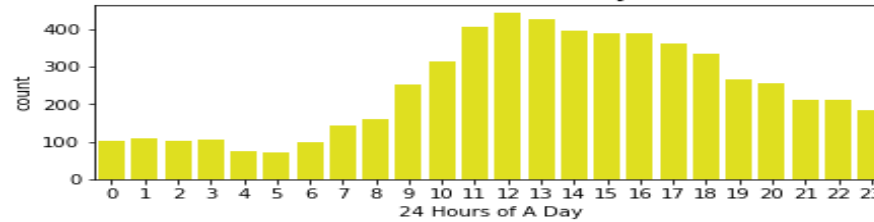
Accident Numbers on Thursdays 2017-2019



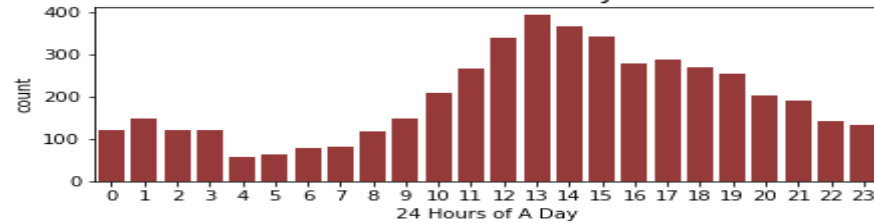
Accident Numbers on Fridays 2017-2019



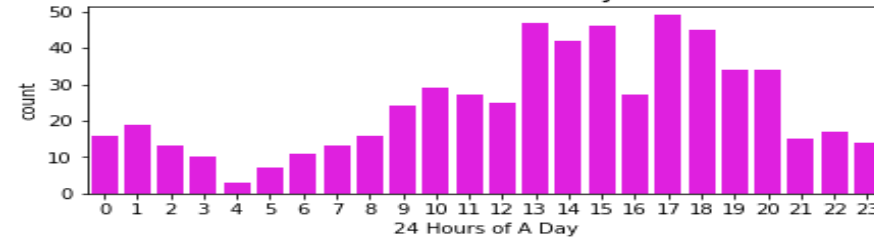
Accident Numbers on Saturdays 2017-2019



Accident Numbers on Sundays 2017-2019



Accident Numbers on Holidays 2017-2019

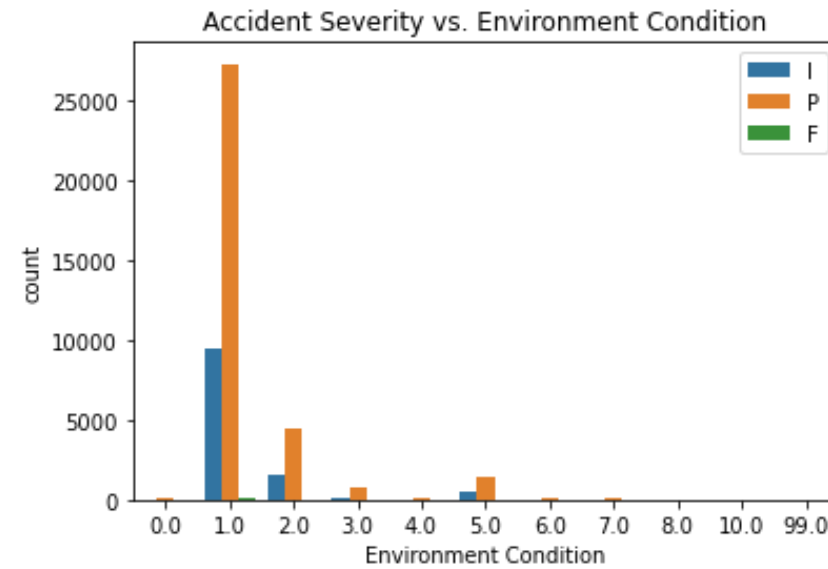
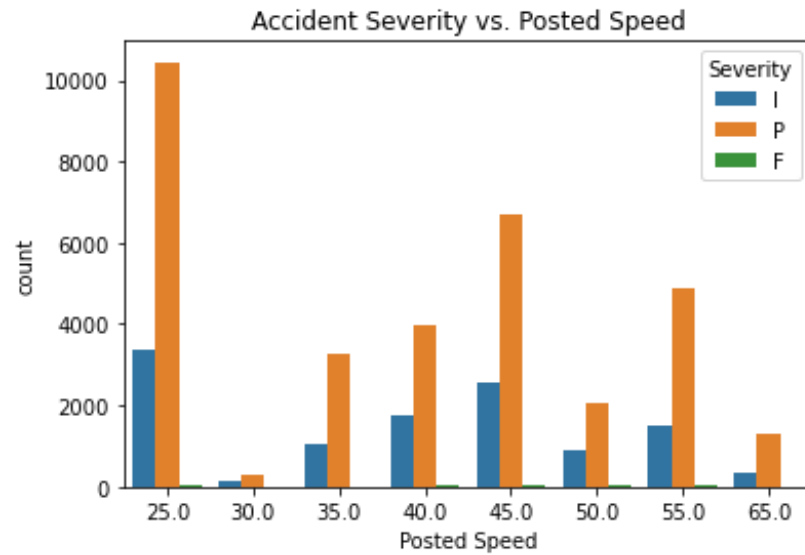
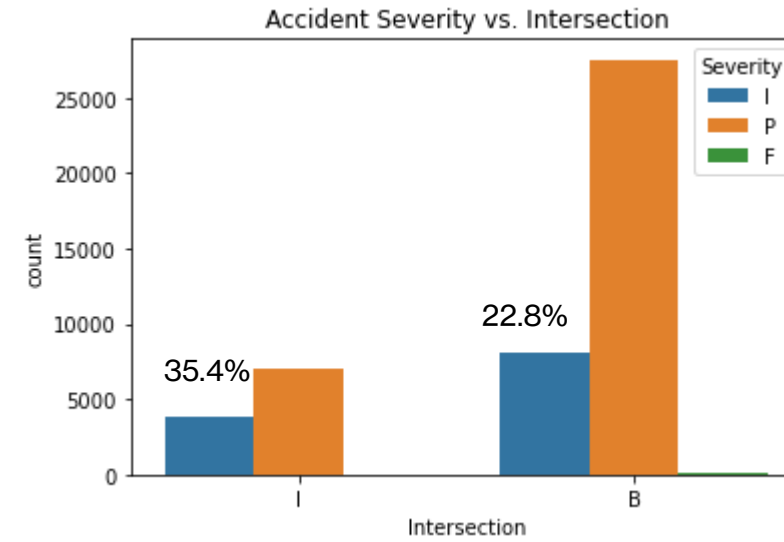
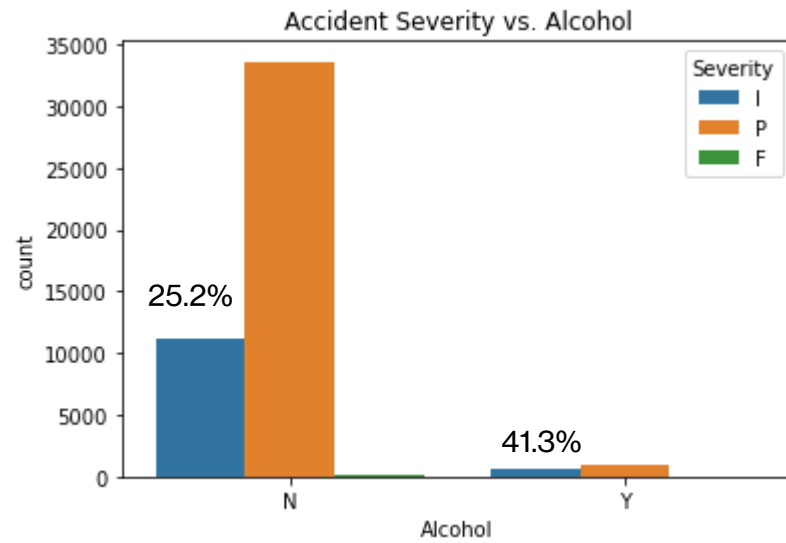


Crash Severity Research

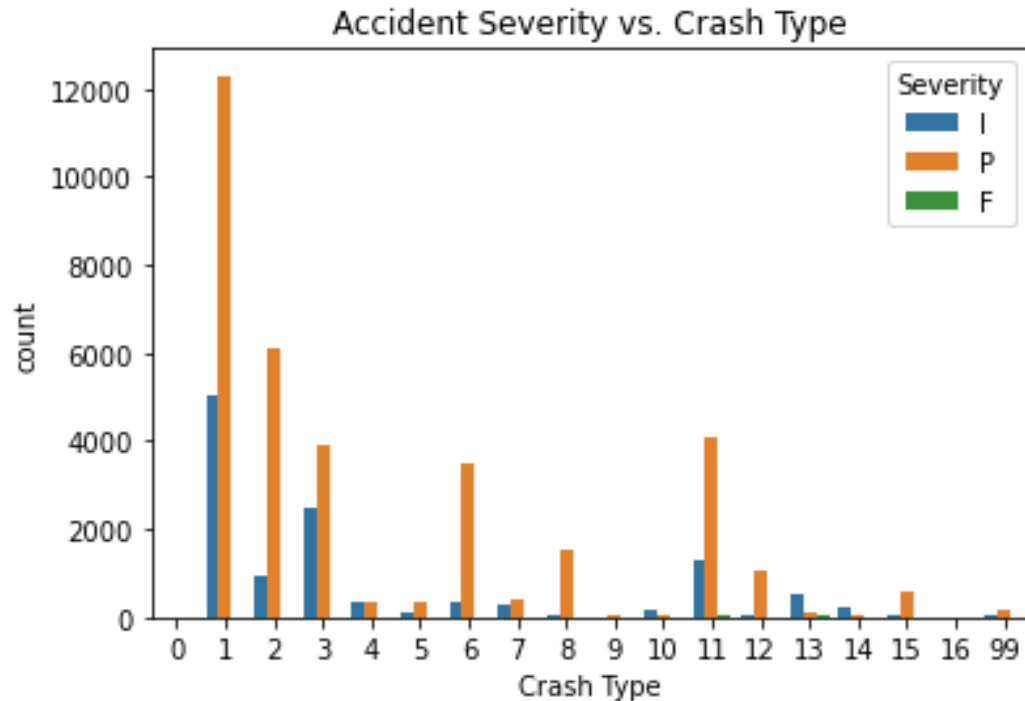
- During 2017-2019, among 46536 cases:
- Property Damage (P) : 34543
- Injury (I): 11853
- Fatality (F): 140

Need to understand which factors can more likely cause injuries or fatalities

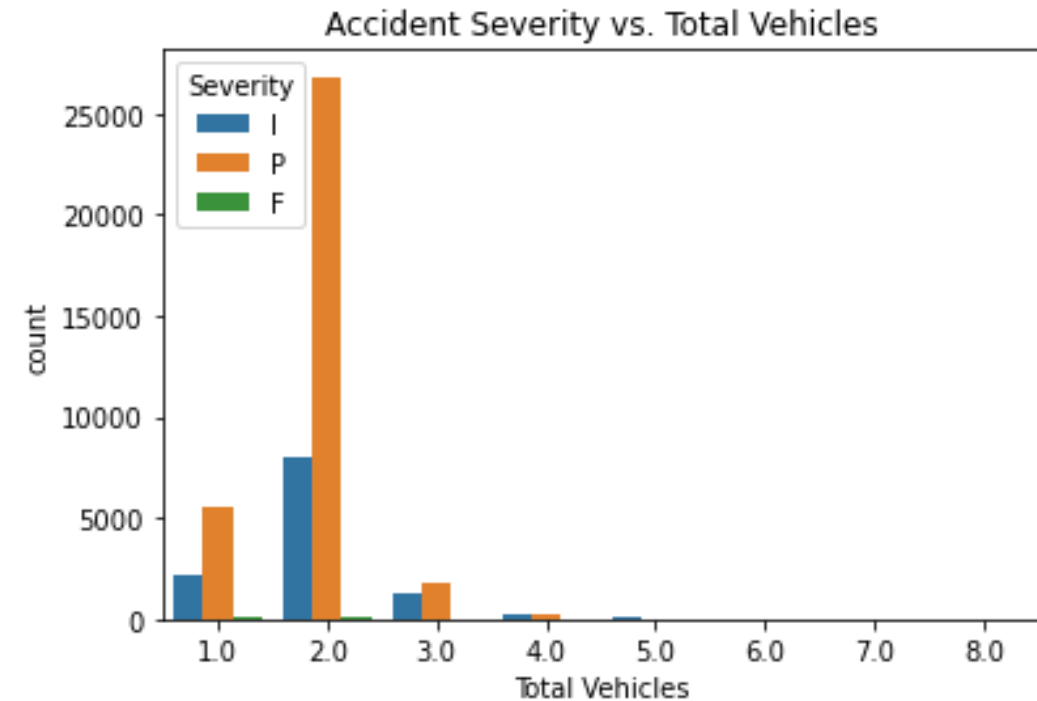
Impact Factors



Factors Significantly Matter



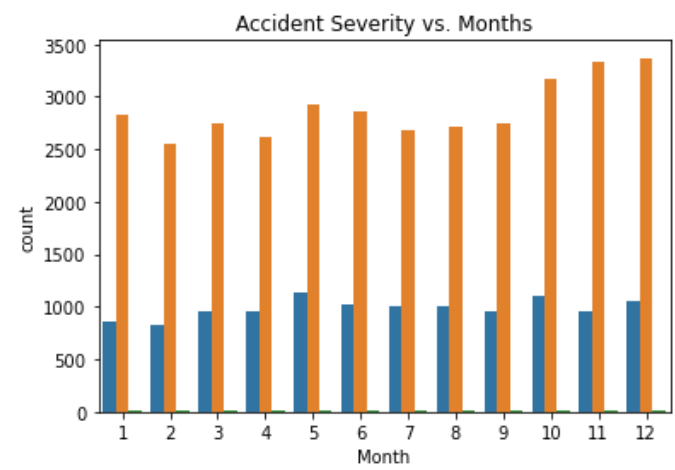
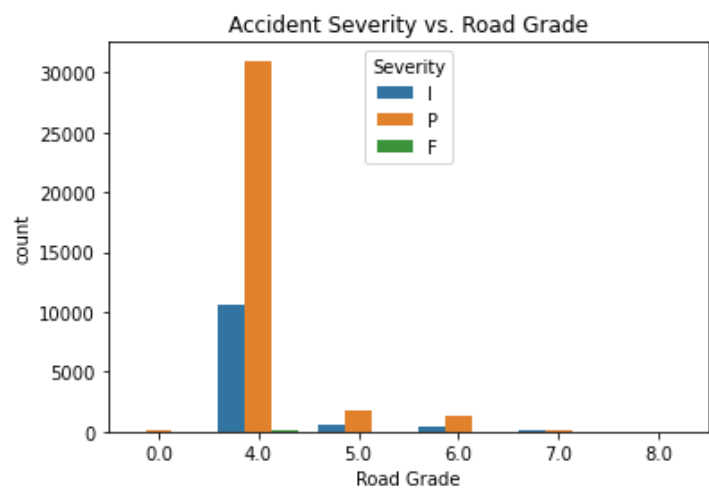
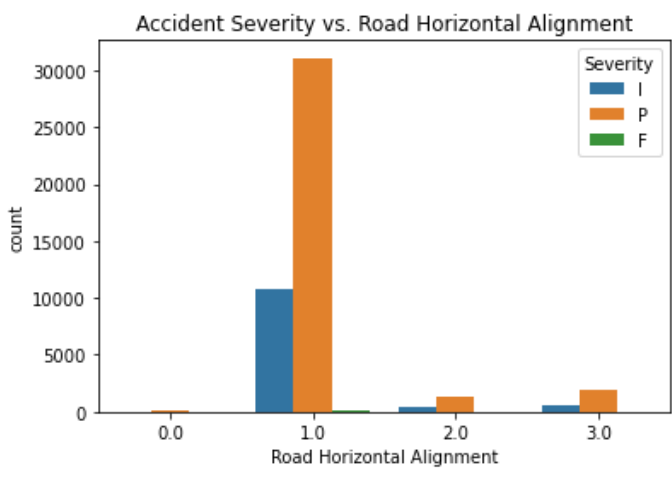
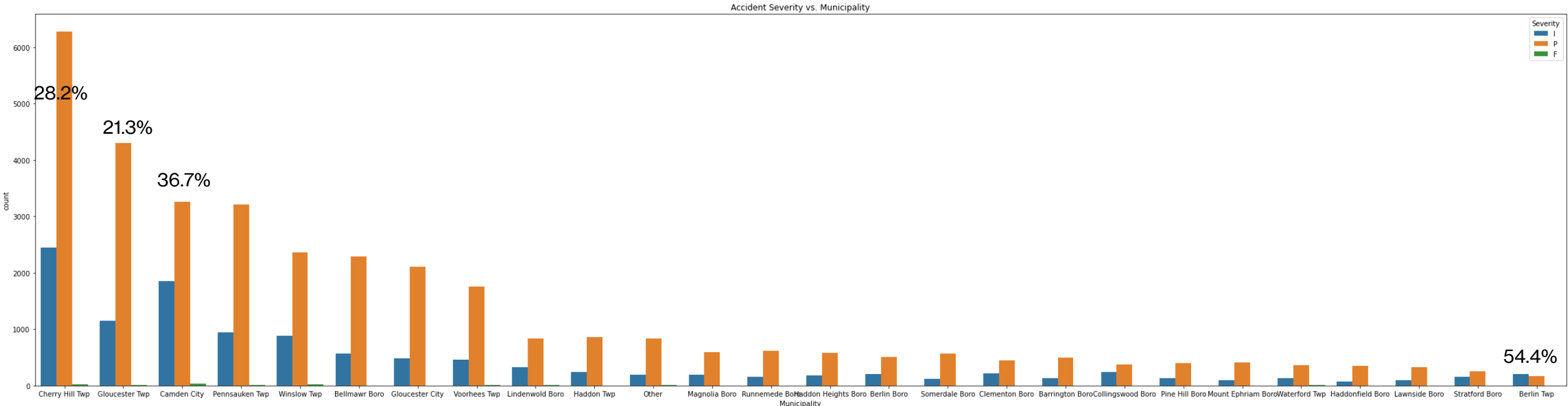
1. Same direction (Rear end) 29.0%
2. Same direction (Side swipe) 13.2%
3. Right Angle
4. Opposite Direction(Head on, Angular) 49.2%
7. Left Turn, U Turn 41.9%
8. Backing 8.6%
12. Animal: 3.8%
13. Pedestrain 85.9%
14. Pedalcyclist 79.3%



Crashes involving 3 or more vehicles likely cause injury or death

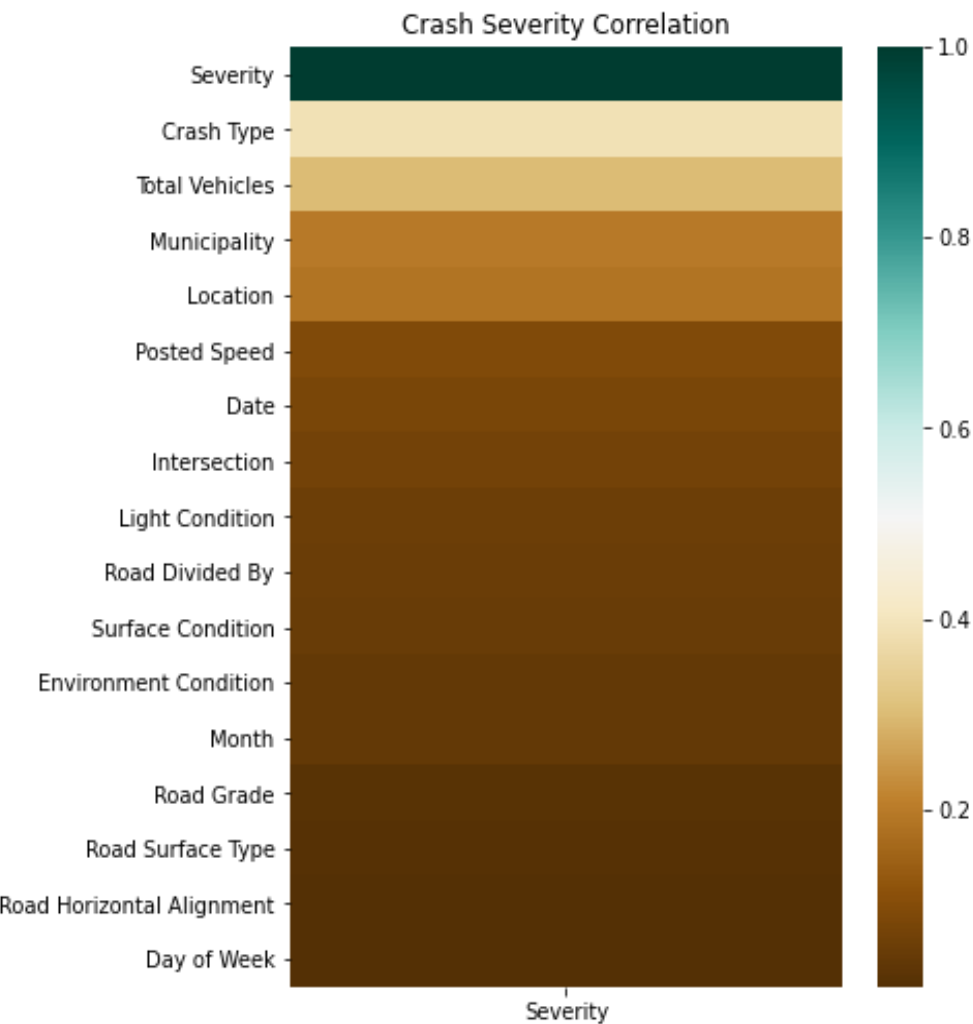
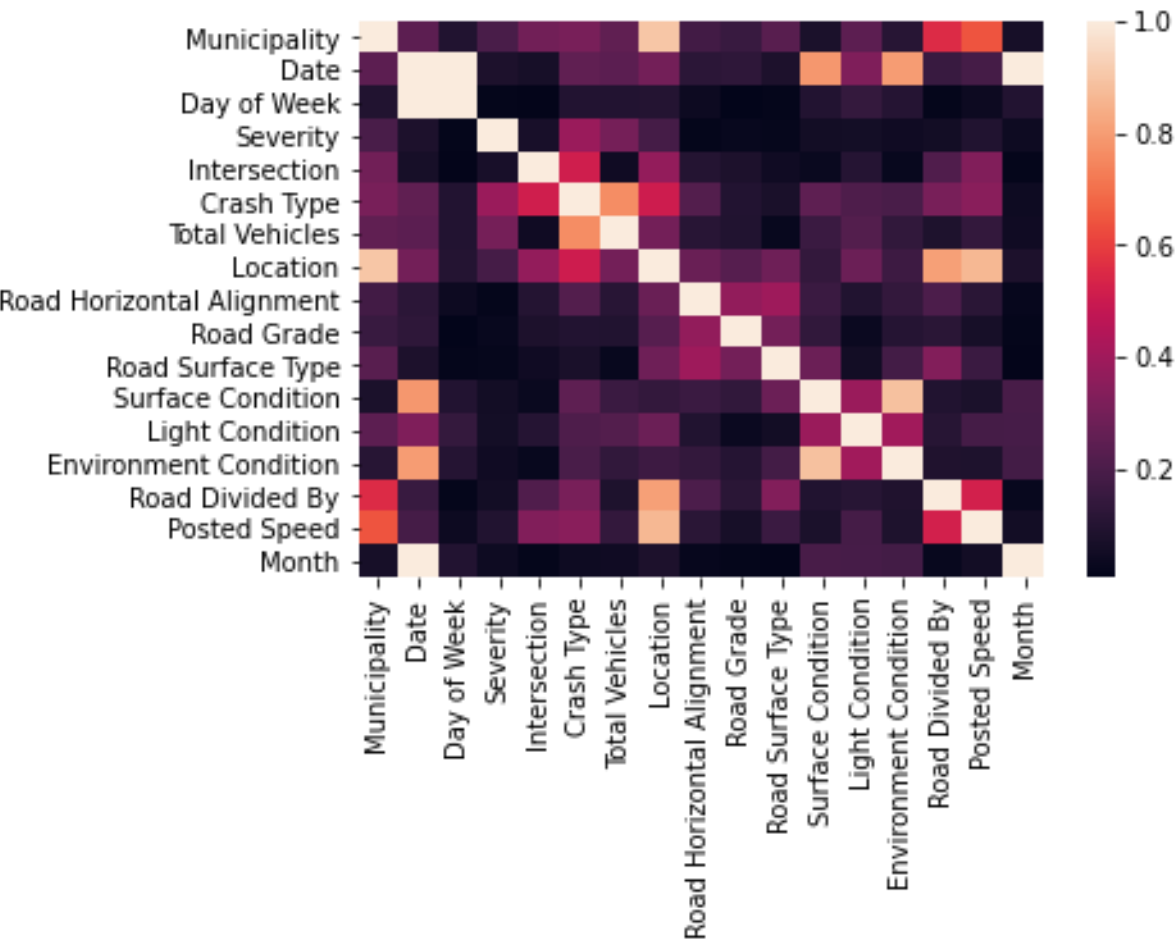
- 1: 29.1%
- 2: 23.1%
- 3 or more: 43.8%

Location Matters!

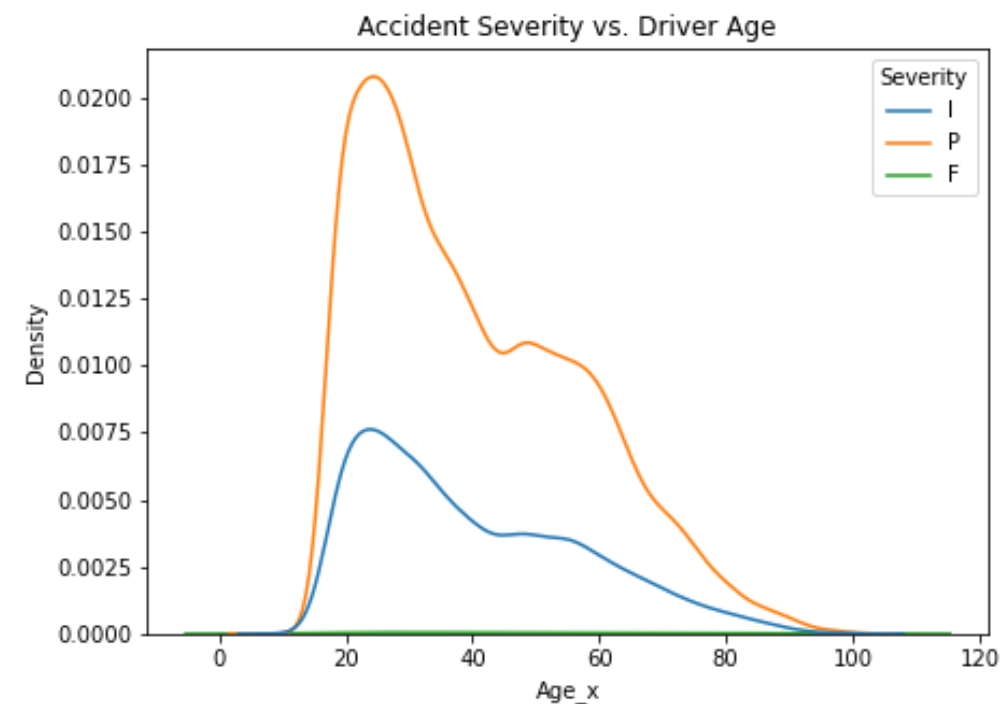
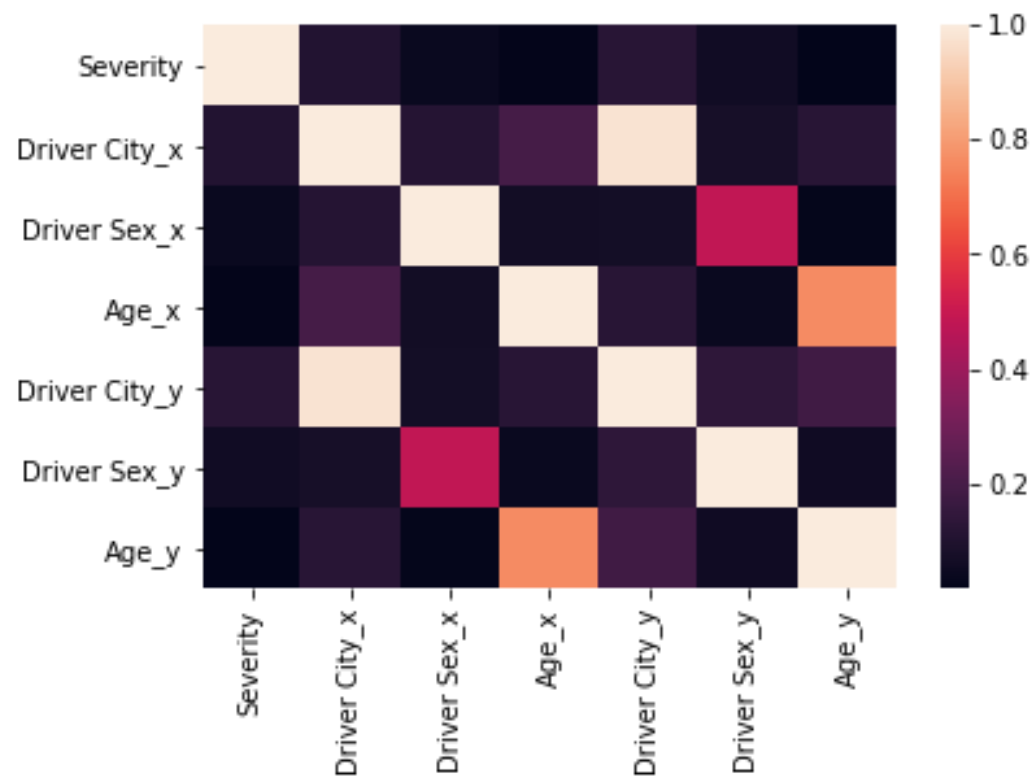


Phi_K Correlations

Assign accident severity of 'P' as 0, 'I' and 'F' as 1.

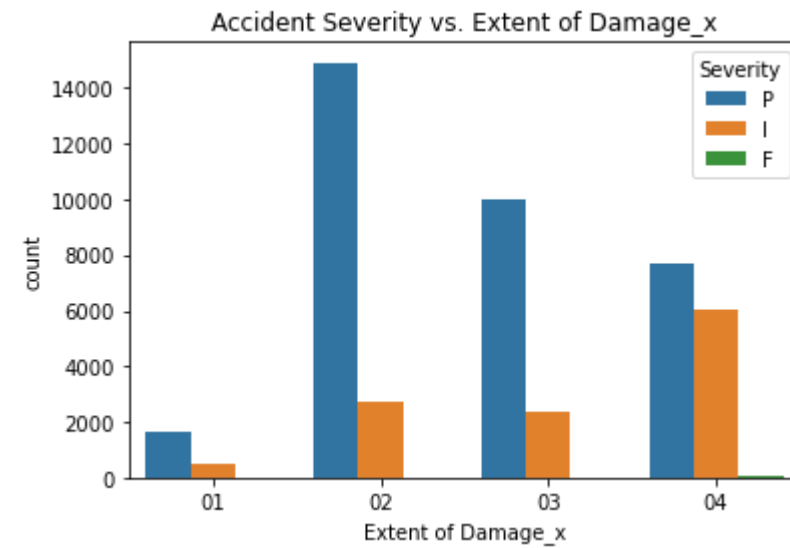
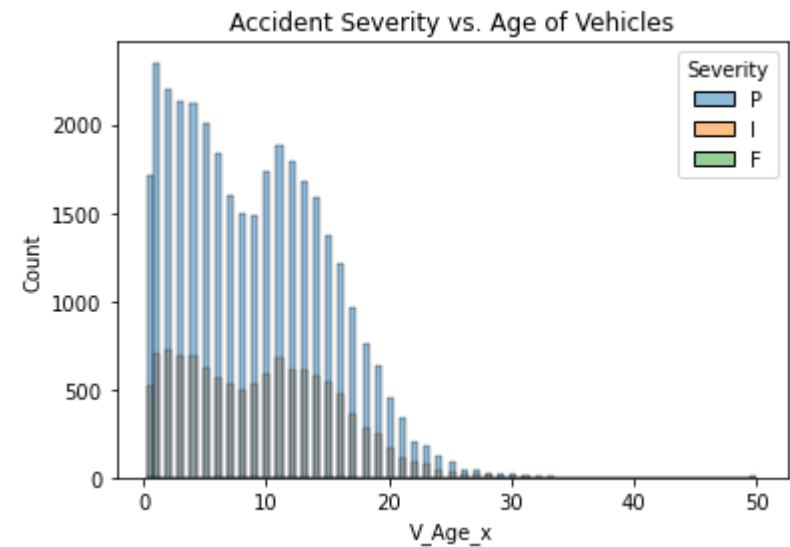
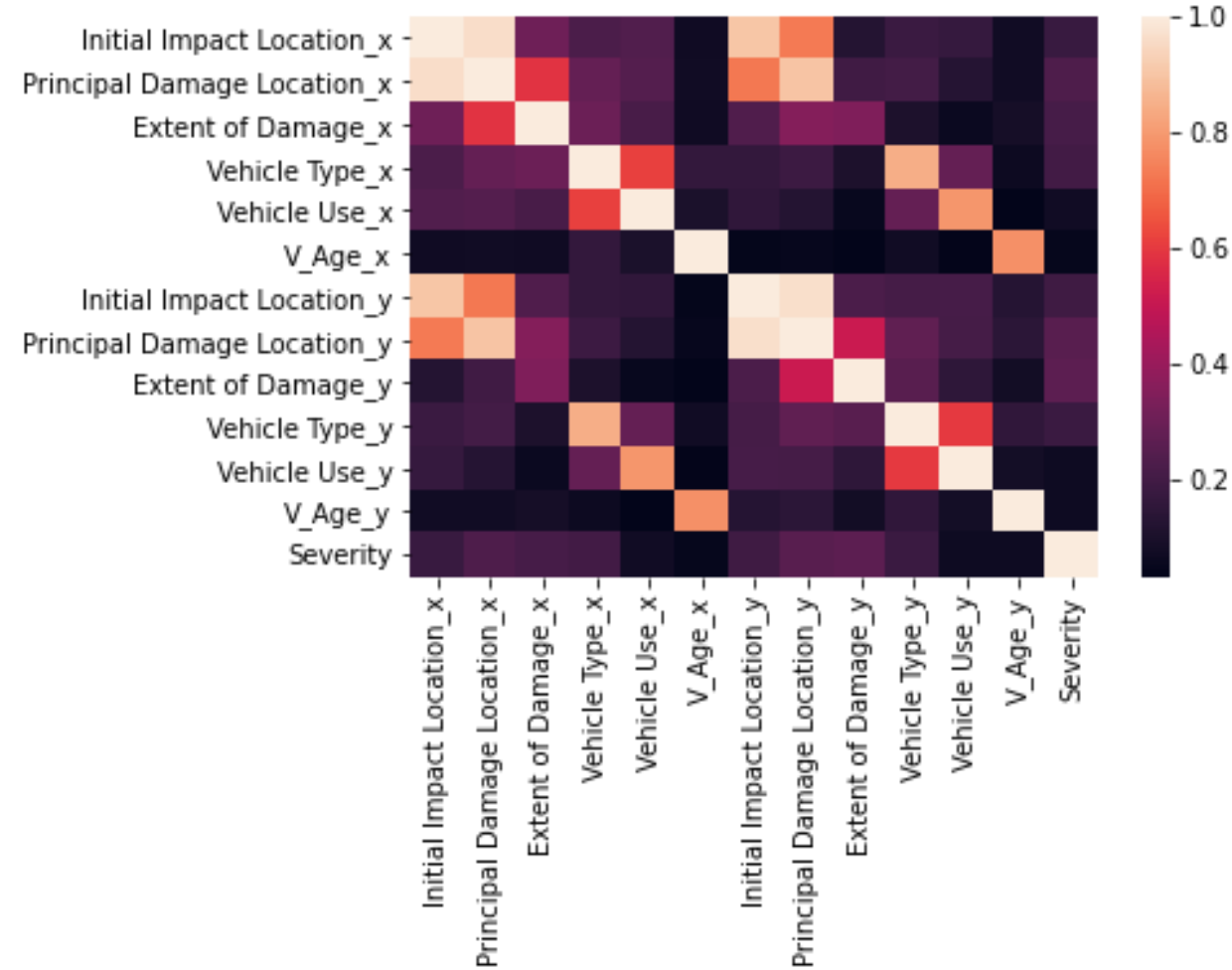


Drivers' Features Correlations



Drivers' age is not significantly related to the accident severity

Vehicles' Feature Correlations



Machine Learning: Predict the Severity of the Traffic Accident

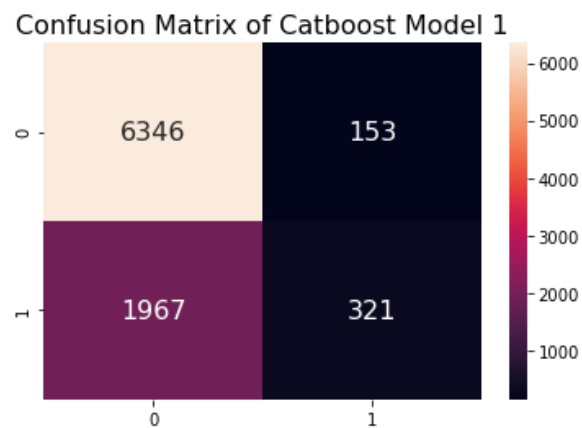
1. Catboost-1

Selected Features:

Categorical: 'Municipality', 'Intersection', 'Crash Type', 'Total Vehicles'

Numeric: Posted Speed, Month

	Precision	Recall	F1-score	Support
0	0.76	0.98	0.86	6499
1	0.69	0.14	0.23	2288
Accuracy			0.76	8787
Macro average	0.73	0.56	0.54	8787
Weighted Average	0.74	0.76	0.69	8787



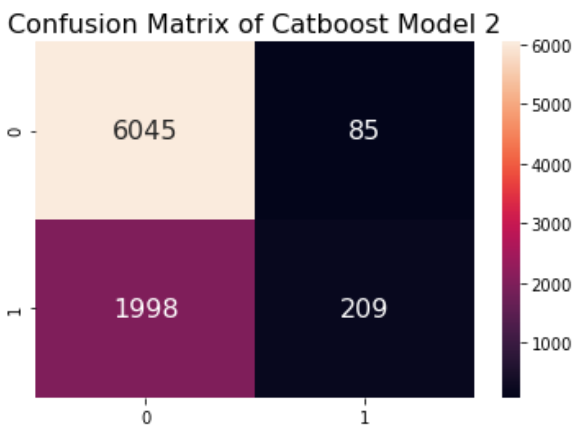
2. Catboost-2

Selected Features:

Categorical: 'Intersection', 'Crash Type', 'Total Vehicles', 'Location', 'Light Condition', 'Environment Condition', 'Road Divided By', 'Driver Sex_x', 'Driver Sex_y'

Numeric: 'Posted Speed', 'Month', 'Driver Sex_x', 'Age_x', 'Driver Sex_y', 'Age_y'

	Precision	Recall	F1-score	Support
0	0.75	0.99	0.85	2207
1	0.71	0.09	0.17	6130
Accuracy			0.75	8337
Macro average	0.73	0.54	0.51	8337
Weighted Average	0.74	0.75	0.69	8337

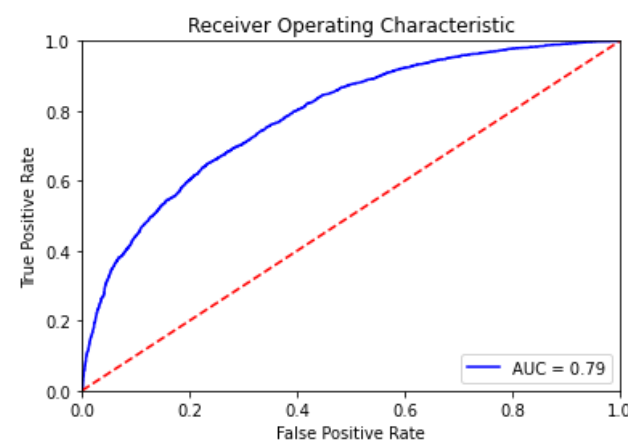


3. Catboost-3

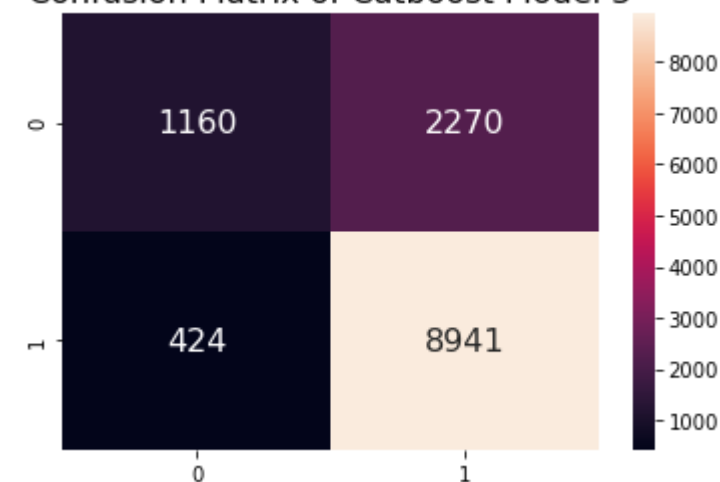
Selected Features:

Categorical: 'Municipality', 'Intersection', 'Crash Type', 'Total Vehicles', 'Environment Condition', 'Road Divided By', 'Initial Impact Location_x', 'Principal Damage Location_x', 'Vehicle Type_x', 'Vehicle Use_x', 'Initial Impact Location_y', 'Principal Damage Location_y', 'Vehicle Type_y', 'Vehicle Use_y'.
Numeric: None

	Precision	Recall	F1-score	Support
0	0.80	0.95	0.87	3430
1	0.73	0.34	0.46	9365
Accuracy			0.79	12795
Macro average	0.76	0.65	0.67	12795
Weighted Average	0.78	0.79	0.76	12795



Confusion Matrix of Catboost Model 3



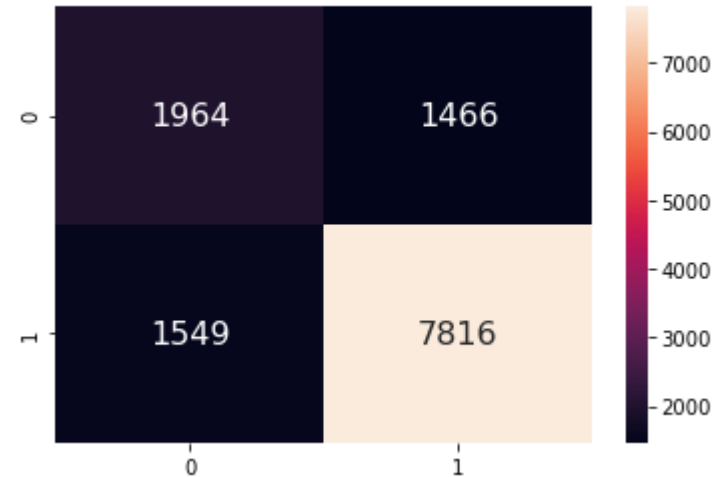
Solving the Data Imbalance Problem

Synthetic Minority Oversampling Technique (SMOTE)

0: 22187, 1: 7667  0: 22187, 1: 22187

	precision	recall	f1-score	support
I	0.56	0.57	0.57	3430
P	0.84	0.83	0.84	9365
accuracy			0.76	12795
macro avg	0.70	0.70	0.70	12795
weighted avg	0.77	0.76	0.77	12795

Confusion Matrix of Catboost Model 4

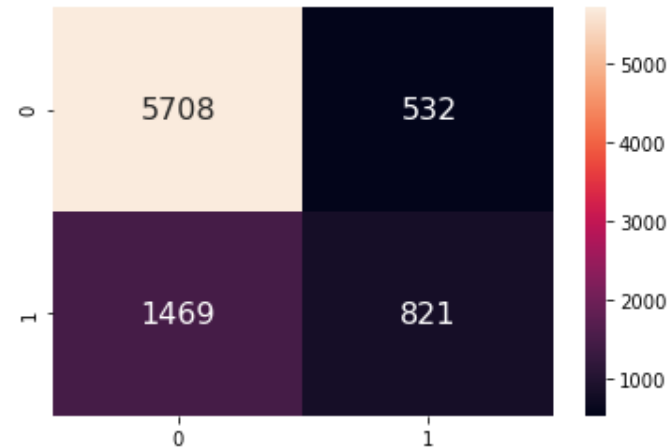


Random Forest Model

Train data : 0: 25312
 1: 8807

	precision	recall	f1-score	support
0	0.80	0.91	0.85	6240
1	0.61	0.36	0.45	2290
accuracy				0.77 8530
macro avg				0.70 0.64 0.65 8530
weighted avg				0.74 0.77 0.74 8530

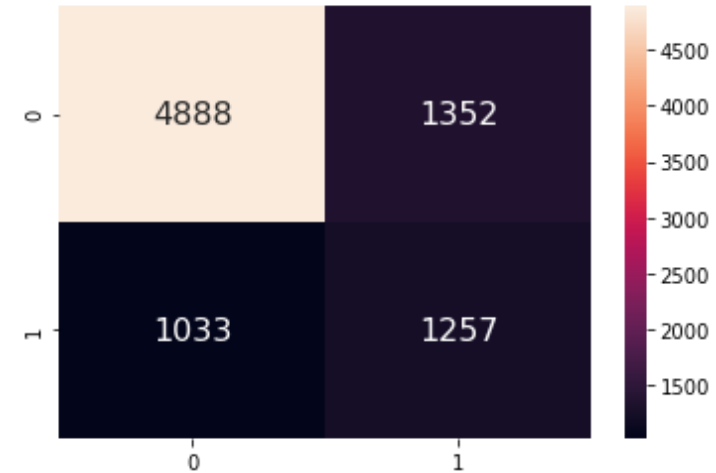
Confusion Matrix of Random Forest Model 1



Train data after SMOTE: 0: 25312
 1: 25312

	precision	recall	f1-score	support
0	0.83	0.78	0.80	6240
1	0.48	0.55	0.51	2290
accuracy				0.72 8530
macro avg				0.65 0.67 0.66 8530
weighted avg				0.73 0.72 0.73 8530

Confusion Matrix of Random Forest Model 2

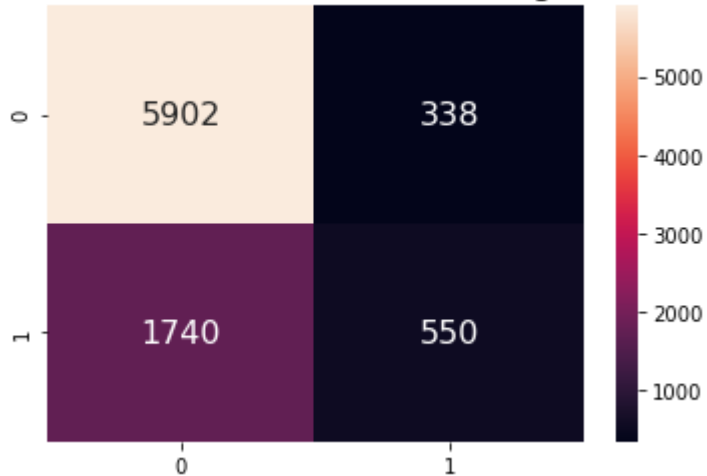


Gradient Boosting Model

Train data : 0: 25312
 1: 8807

	precision	recall	f1-score	support
0	0.77	0.95	0.85	6240
1	0.62	0.24	0.35	2290
accuracy				0.76 8530
macro avg				0.70 0.59 0.60 8530
weighted avg				0.73 0.76 0.71 8530

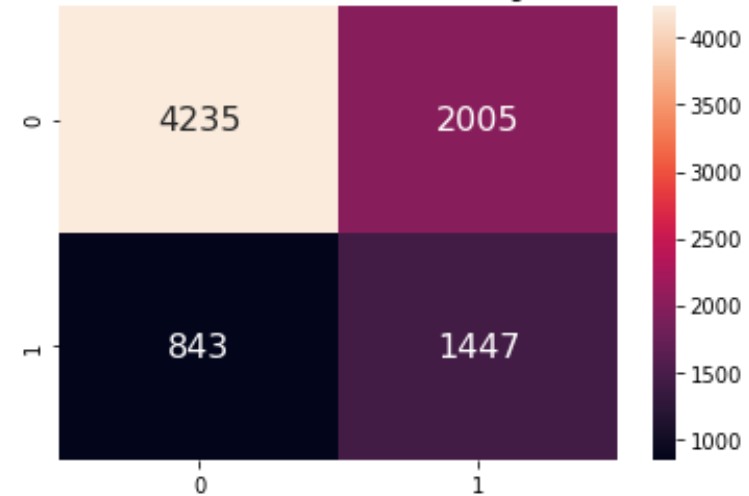
Confusion Matrix of Gradient Boosting Model 1



Train data after SMOTE: 0: 25312
 1: 25312

	precision	recall	f1-score	support
0	0.83	0.68	0.75	6240
1	0.42	0.63	0.50	2290
accuracy				0.67 8530
macro avg				0.63 0.66 0.63 8530
weighted avg				0.72 0.67 0.68 8530

Confusion Matrix of Gradient Boosting Model 2



Summary

1. In the state of New Jersey, the total number of traffic crashes remain unchanged during the past 20 years, the injuries and deaths from traffic crashes declined.
2. The traffic accident occurrences vs. hours of a day, weekdays, holidays and months were studied.
3. A classification model to predict the severity (causing injury/death or just property damage) of a traffic accident were established. It can achieve close to 80% accuracy.
4. The features mostly likely influence injury or death are crash type, vehicle damage extent, vehicle impact location and municipality.
5. Due to too much missing data of the exact location of accidents and lack of the detailed weather conditions, I only use the municipality as the location feature and no weather feature, future study can focus more on the exact locations and the detailed weather conditions to find ways of better predictions.