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

Ning Shangguan

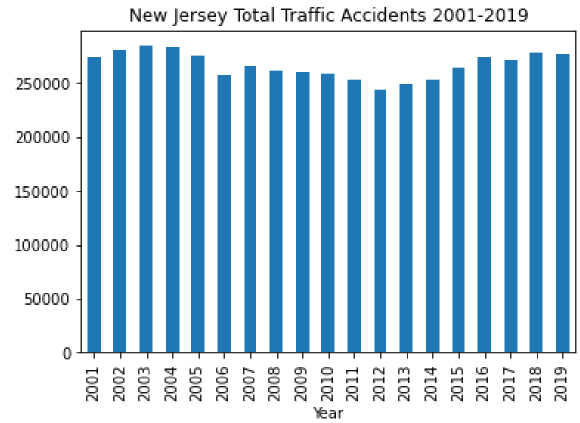
Although the rate of traffic accident fatality rate declined significantly during the past 50 years, traffic accidents remain to be one of the major causes of death in US. As a resident of Camden County of New Jersey, I am interested in the local traffic accident statistics. The questions I would like to ask are: What is the traffic crash rate in my local area, how it compare to those of the other parts of the state and the whole country? What is the traffic accident peak time in a day, a week, a month, and a year? What factors affect the traffic injury and death? What kind of accidents are more likely to cause personal injury or fatalities?

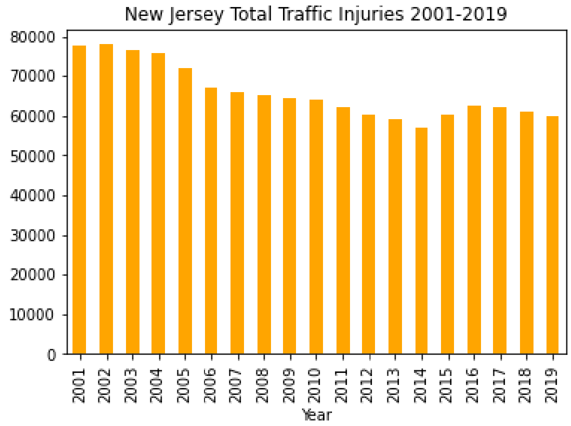
I decided to do a data research to study the traffic accidents in Camden County, New Jersey to find the answers to the questions.

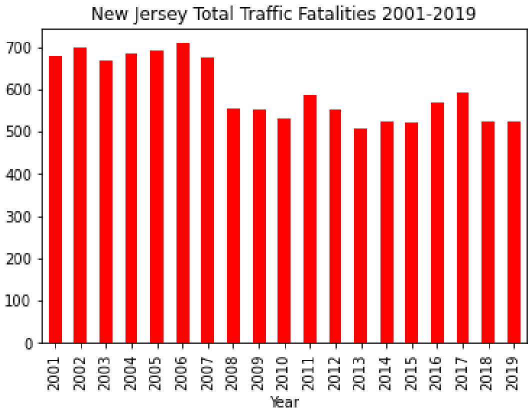
1. Data

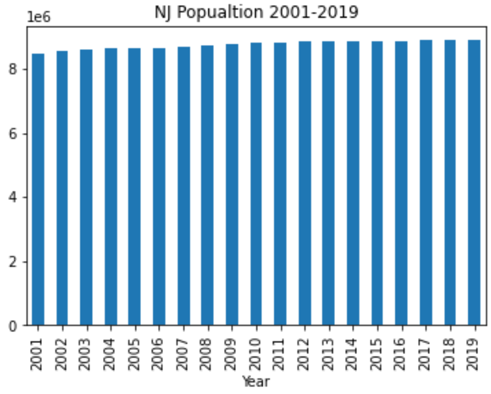
The department of New Jersey provide the summary of the traffic crashes data from 2001-2019 at https://www.state.nj.us/transportation/refdata/accident/crash_statistics.shtm for each counties (21 total) of the state.

Let I looked at the overall summary (crashes, injuries and fatalities) of the state of New Jersey:


New Jersey Total Traffic Accidents 2001-2019


New Jersey Total Traffic Injuries 2001-2019

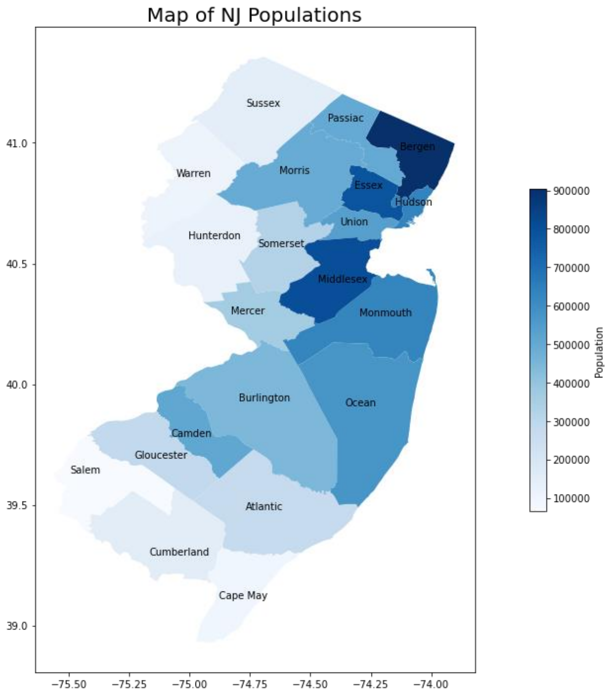

New Jersey Total Traffic Fatalities 2001-2019

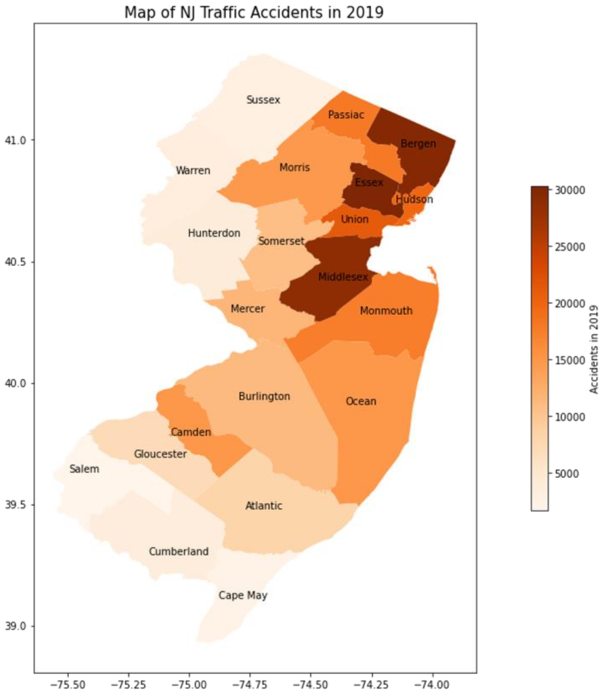

NJ Popualtion 2001-2019

While the total traffic crashes remaining unchanged (Year 2001: 274,110, Year 2019: 276,861), the injuries (Year 2001: 77,397, Year 2019: 59,850) and fatalities (Year 2001: 667, Year 2019: 524) showed a significant drop during the 19 years' span. The population of New jersey during this period did not change much: it was 8.49 million in 2001 and 8.88 million in 2019.

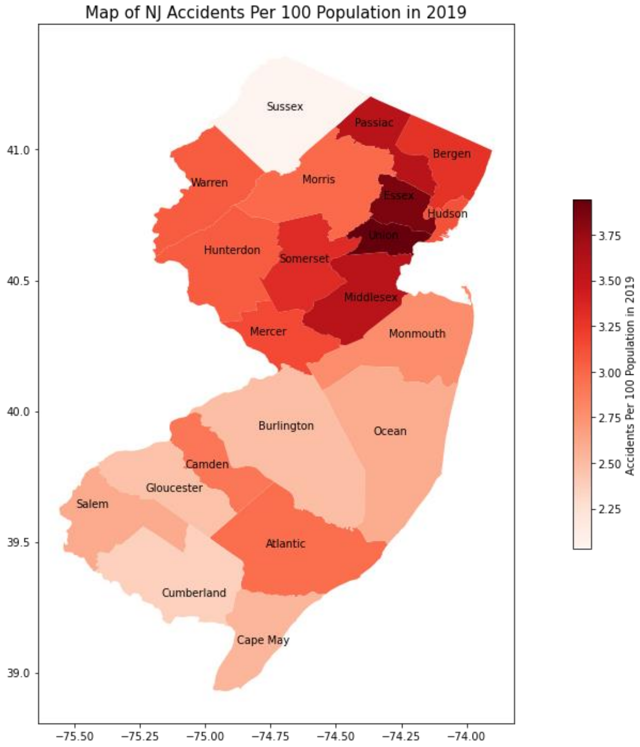
My understanding of the data of the traffic crashes of NJ is: People's driving habits did not change much, the overall traffic accidents rate remained almost same during 2001-2019. The cars had become safer, so the overall injury and fatality rate decreased significantly.

The maps below show populations of each county in NJ and the accidents in 2019. The traffic accident number in each county is directly related to the populations of the county: Bergen County has the highest population (930K) and the highest number of traffic accidents in 2019 (29.7K). Warren County has the lowest population (63K) and lowest accident number in that year (1.7 K). Camden County is in the middle range among the 21 counties of NJ: 50.6K population and 15 K accidents in 2019.



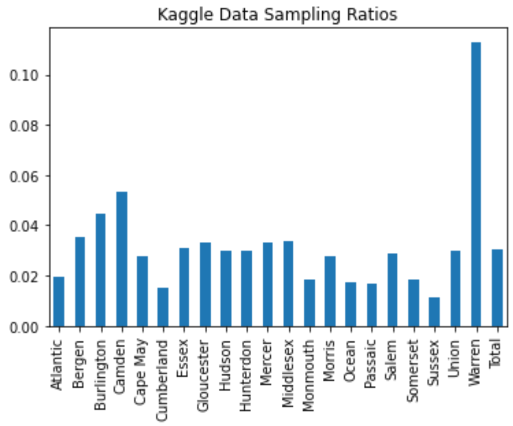


The map below shows the accidents per 100 populations do not vary much from each county of the state. The Union County has the highest of 3.95 and the Sussex County has the lowest of 2.03. My guess is there are many corporate offices in the Union County and people need to drive there to work. The ratio of actual vehicle number to the resident population is higher in Union County. Sussex County is mostly mountainous area and not many corporate offices sit there. Although the population and population density vary a lot among the counties, the difference of accidents per 100 populations is relatively small.



The Kaggle Dataset:

On the other hand, it is worth noting that 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. It is a clean dataset without missing values. But the estimation by the NHTSA says that 6 million car accidents happen in the U.S. every year, 3 million cases in almost 5 years is only a small portion of the total accidents. If we look closer at the state of NJ, the Kaggle dataset provided 8,435 cases in year of 2019 while the DOT of NJ reported 276,861 accidents. The sampling ratio is only 3.05%. If we look at each county, it become even more problematic. As the graph below shows, the sampling ratio from each county varies significantly: the lowest is Sussex County, only 1.15% of the cases were sampled into the Kaggle dataset, the highest is Warren County with a 11.3% ratio. My conclusion is the Kaggle dataset is good for practice of the machine learning skills but not good for the real data project research.



2. Data Cleaning

NJ Department of Transportaion provides Crashes, Drivers, Vehicles, Occupants and Pedestrians datasets in text format for each year at https://www.state.nj.us/transportation/refdata/accident/rawdata01-current.shtm. First, I converted the Crashes tables (2017-2019) to the Pandas DataFrames.

I. Traffic Crashes

There were 15233 records in the total crash summary of Camden County of NJ in 2019. Each row had 38 Columns. They are: 'Number', 'County', 'Municipality', 'Date', 'Day of Week', 'Time', 'Police Station', 'Killed', 'Injured', 'P Killed', 'P injured', 'Severity', 'Intersection', 'Alcohol', 'Hazmat', 'Crash Type', 'Total Vehicles', 'Location', 'Route', 'Std Rte Identifier', 'MilePost', 'Read System', 'Road Character', 'Road Horizontal Alignment', 'Road Grade', 'Road Surface Type', 'Surface Condition', 'Light Condition', 'Environment Condition', 'Road Divided By', 'T Traffic Control Zone', 'Unit', 'Cross Street Name', 'Posted Speed', 'First Harmful Event', 'Latitude', 'Lo', 'Cell Phone' and 'Other Property Damage'.
Some of the features had many missing values, they were:

```
Police Station                  6657
Road Character                 15233
Unit                            3895
Cross Street Name               3680
Route                           3018
Latitude                       12019
Lo                             12019
Other Property Damage          13761
```

The total is 15233, the ‘Road Character’ is full of missing values, it must be dropped.
The latitude and longitude are valuable features. But with about 80% of the values were missing, it had to be dropped. The Route’s values were overlapped with Location. It was dropped. The other features listed above were also dropped.

The ‘Cell Phone’ and ‘Hazmat’ features were extremely unbalanced and did not seem to impact the injury ratio, they were dropped.
‘Cell Phone’ values count: ‘Hazmat’ values count:

```
N    15136                    N     15418
Y       96                    Y         4
```

The other columns were dropped: 'T Traffic Control Zone', 'Unit', 'MilePost', 'First Harmful Event', ,'Std Rte Identifier', 'Read System’.

The data of 2017 and 2018 were converted to the Pandas DataFrame and concatenated with 2019 DataFrame. Totally I had 46,977 entries and 25 columns.

Date and Time missing values count:

```
Date    279
Time    441
```

All the rows have Date or Time missing values were dropped. Date and Time were merged to one column.

Finally, I had a traffic crash table of 46,536 entries and 24 columns.

II. Drivers

First, I looked at the data of 2017 Camden County Drivers (involved in traffic accidents). The text file was converted to the Pandas DataFrame. It was 29805 entries and 22 columns.

I decided to drop the 'Charge 1', 'Summons 1', 'Charge 2',   'Summons 2', 'Charge 3', 'Summons 3', 'Charge 4', and 'Summons 4' columns because they could leak the accident severity information.

The 'Driver State', 'Driver Zip Code' and 'Driver License State' columns were dopped due to the duplicate information. The ‘Driver DOB’ were replaced with the Driver’s ‘Age’ feature.

The drivers’ data entries were almost twice as many as the crash table entries, because most of the crash cases involved 2 vehicles, some even 3 or more. The Vehicle Number values count:

```
1:15405, 2:12803,  3: 1325,  4: 226,  5: 35,  6:9,  7: 1,  NJ: 1
```

So there were more than 1500 accidents involving 3 or more vehicles. But there were still some accidents only involving one vehicle. To deal with the problem, I only took vehicle number 1 and 2, split the table and merged them with the same case number. Then every accident case had 2 drivers involved in the accidents. For the cases with only 1 driver (missing value for the other driver), I duplicated the driver 1 information to the driver 2 or vice versa.

After the drivers’ data processing, I still had 43988 entries (2017-2019) and 7 columns. The drivers’ data table could be merged with Crash table by sharing the same case number.

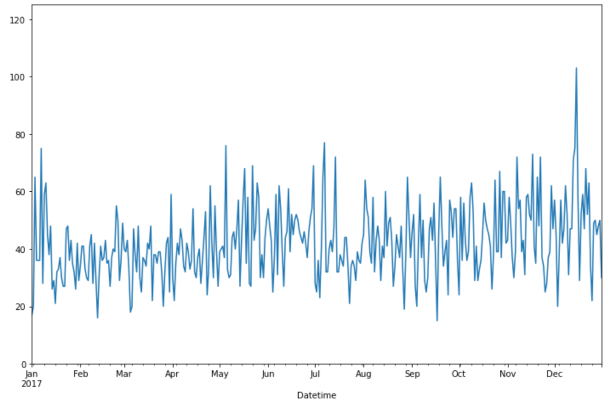
III. Vehicles

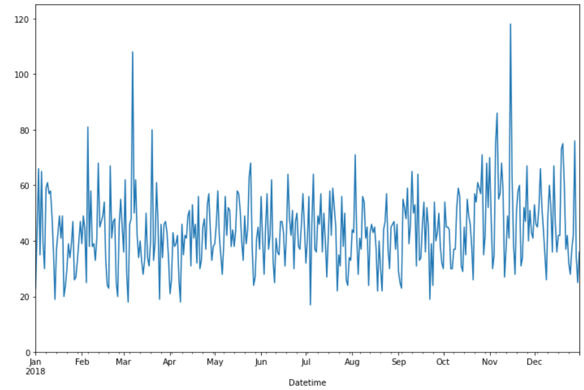
Same as the driver’s data, most of the accidents involved 2 vehicles, it is not clear which vehicle was the guilty side. I did the same processing to the vehicles’ table as the drivers’ table: split and merge, duplicate if there was missing value. I dropped the columns with more than one third of missing values or obviously was not related to the accident severity.

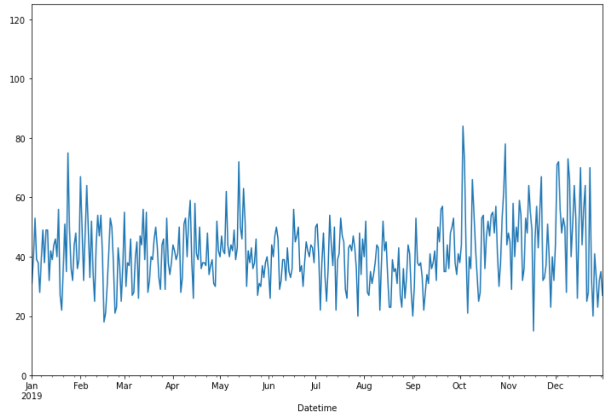
Totally the vehicles’ table had 46034 entries and 15 columns.

3. Data Wrangling

For the total 46536 traffic accident cases in 2017-2019, the daily average number was 42.5 per day. The graphs below show the number of daily cases in 2017, 2018 and 2019.







Among 3 years, there were 3 days with exceptionaly high number of accidents:

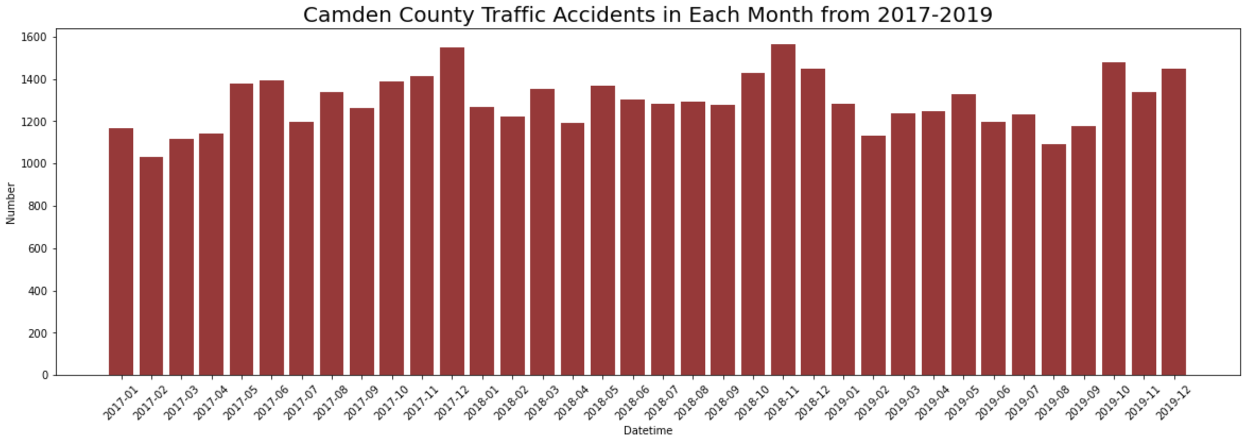
2017-12-15: 103,   2018-03-07: 108,   2018-11-15: 118

Most of the cases in 3 days had an environmental condition of number 3, which meaned snow. It becasme clear that snow weather condtions caused high accident numbers on those 3 days.

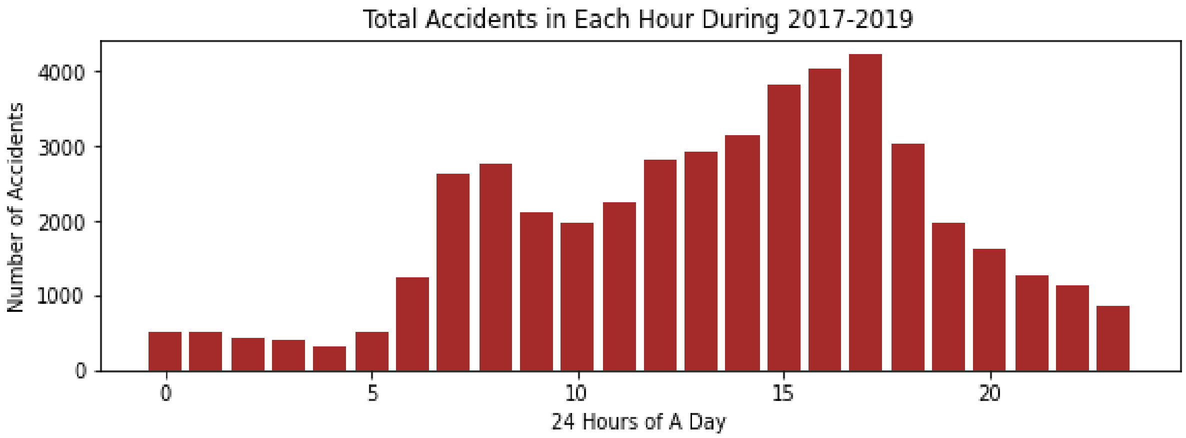
The least accident days were:

2017-09-17 (Monday, 15 cases), 2019-11-17 (Sunday, 15), 2017-01-01 (17) and 2018-07-04 (17)

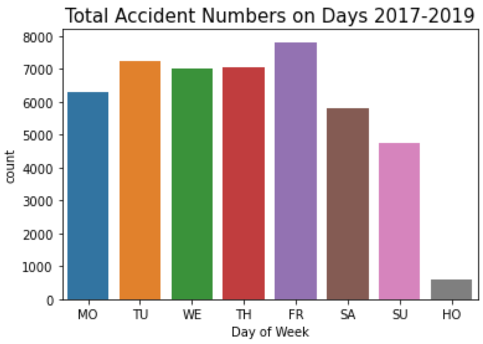
The graph below showed the monthly cases were beween 1000-1500. November and December were a relatively peak time during the whole year.

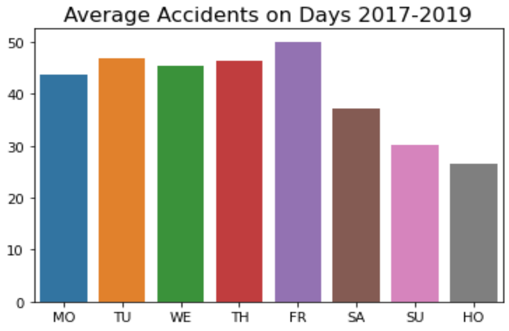


If we look at each hour of a day, we can see the late afternoon (16 -18) is the peak time of the accident occurrence during the whole day, there is a minor morning peak between 7-9. The early morning time (4-5) is the lowest accident time.

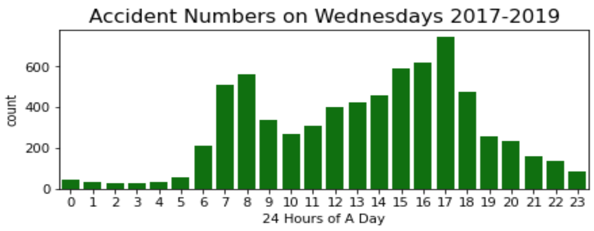


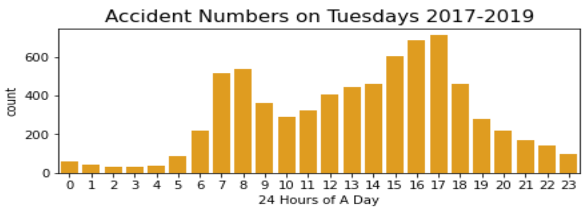
The above summary is based on every day of the 3 years. If we look at the accident occurrence on different weekdays, we can find Fridays have the highest number of accidents, both on total and average.

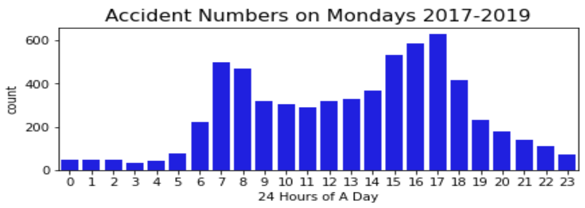


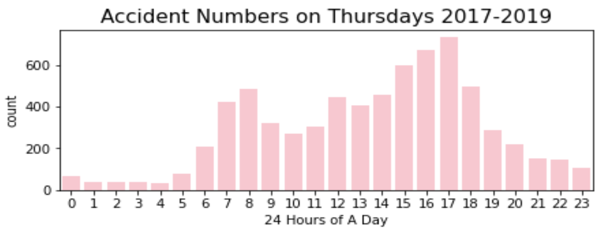


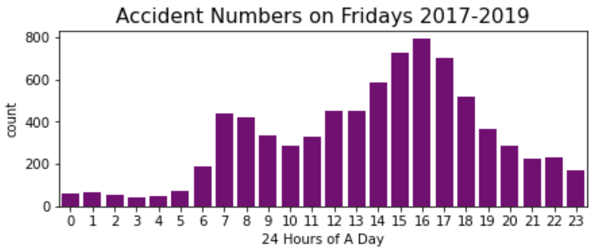
The accidents on every hour of the day (from Monday to Sunday and Holidays) are shown in the graphs below. We can see different patterns: Monday to Thursday share the similar accident occurrences with the double peaks: the major peak in late afternoon and the minor peak in early morning. Fridays see the peak time during 16:00-17:00 instead of 17:00 to 18:00 on Mondays to Thursdays. Saturdays, Sundays, and Holidays is similar to weekends: not having the morning peak and the peak time of the whole day is around noon. Those 3 types of days have more share of accidents at nights and in early mornings.

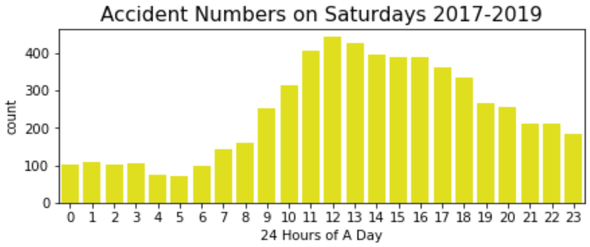


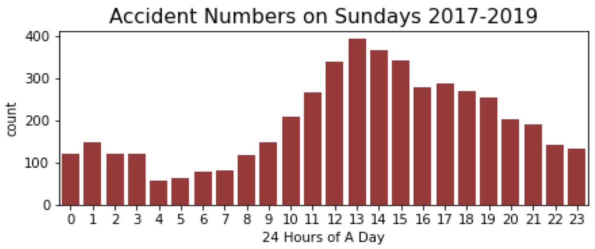


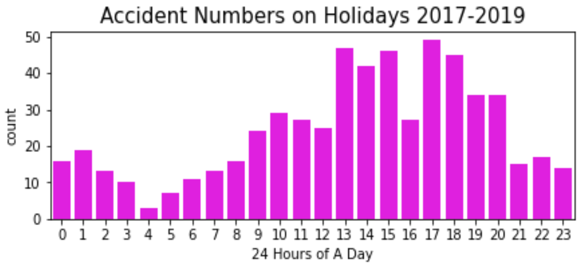












4. Exploratory Data Analysis and Initial Findings

1) We are particularly interested in the severity of an accident, does the accident only cause property damage or cause personal body injury? What kind of accidents are more likely to cause injury and even fatality?

   Among the 46,536 traffic accidents in the Camden County during 2017-2019, I did a value count.

   Accident severity values count:

```
P     34543
I     11853
F       140
```

'p' means property damage, 'I' means injury and 'F' means fatality. From the data above, we can see 25.8% of the accidents caused injury or death. (From now on, we call the ratio of the injuries and fatalities as ' the Ratio'.

The total fatality was 149 since some accidents incurred more than 1 fatality.

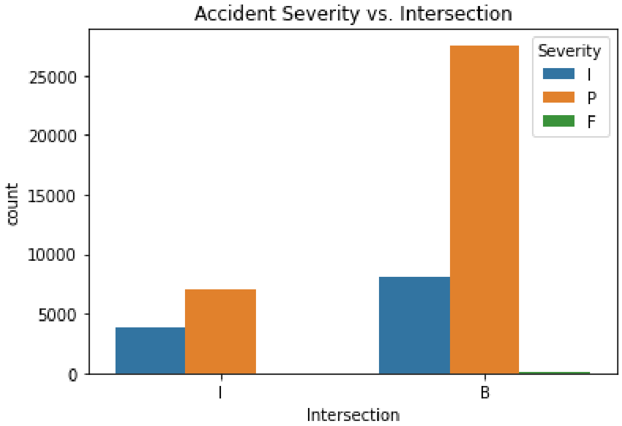
2) Next, we will look at the different features' influence on the severity of the accidents.

   A. The accident location: intersection or non-intersection?

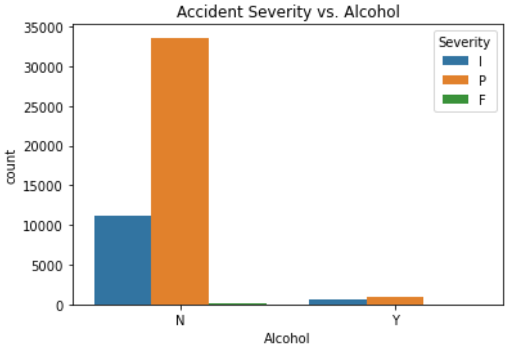
If an accident happened at an intersection ('I') , the ratio of the 'I' or 'F' is 35.4%.

If an accident happened at a non-intersection ('B'), the ratio of the 'I ' or 'P' is 22.8%

The location of the accident matters.



B. Is alcohol involved in the accident?

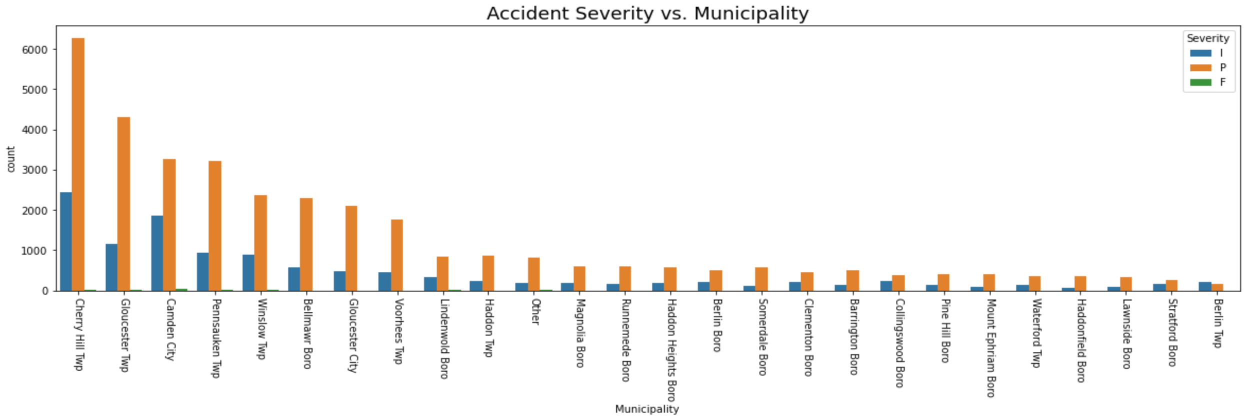


Only 3.4% of the cases involved alcohol. If alcohol is involved, the Ratio is 41.3%, if not, the Ratio is 25.2%. Although alcohol affects the Ratio, but the percentage of alcohol positive is too low to have a significant influence.

C. Municipality

The municipality of the accident location really matters!

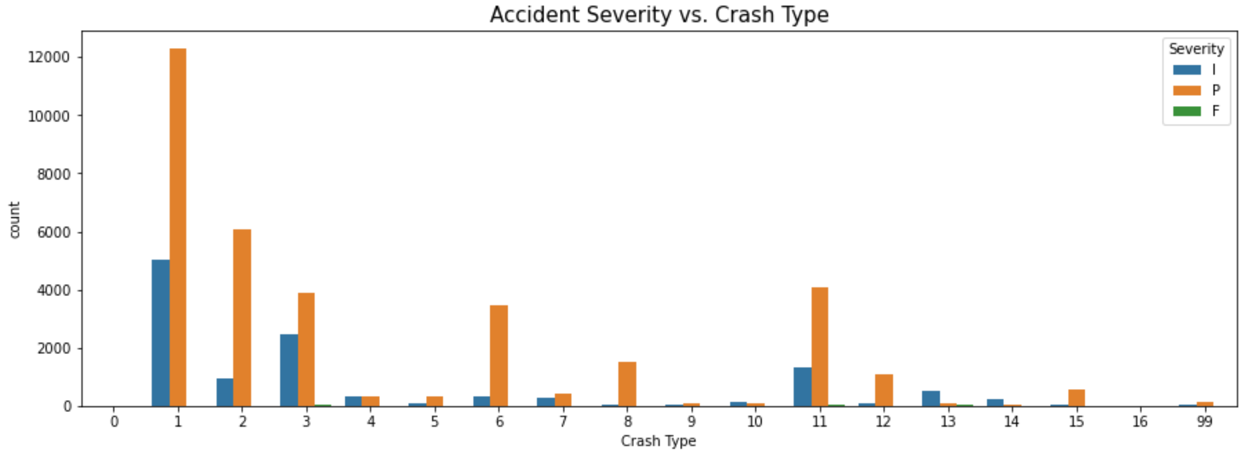
The Ratio at Cherry Hill Twp is 28.2%, Gloucester Twp: 21.3% , Camden City: 36.7%, and Berlin Twp: 54.4%. Why Berlin Twp had such a high Ratio is worth further studying.



D. Crash Type

The Crash Type has a big impact on the Ratio. Some of the number of the crash type are:

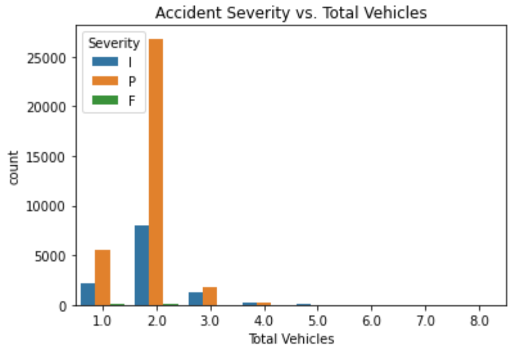
1. Same direction (Rear end) 29.0%
2. Same direction (Side swipe) 13.2%
3. Right Angle 39.0%
4. Opposite Direction (Head on, Angular) 49.2%
6. Backing 8.6%
7. Left Turn, U Turn 41.9%
8. Animal 3.8%
13. Pedestrain 85.9%
14. Pedalcyclist 79.3%



It is easy to understand the accidents involving pedestrian had a high Ratio of 85.9% and the accidents involving animals only had a low Ratio of 3.8% of personal injuries or death.

It is also worth noting the total fatalities from the accidents involving pedestrians (Type 14) were 49 in Camden County from 2017 to 2019. Although the accident type 13 only accounted for **1.37%** of the all traffic accidents (538 in 46,536), the fatalities accounted for an unproportionally high of **33.1%** (49 in 148). Type 13 accident is the most dangerous accident type!

E. Total Vehicles



The total number of vehicles involved in the traffic accident is another indicator of the severity of the accident. The Ratios of accidents of different numbers of vehicles are:

1 vehicle accidents: 29.1%

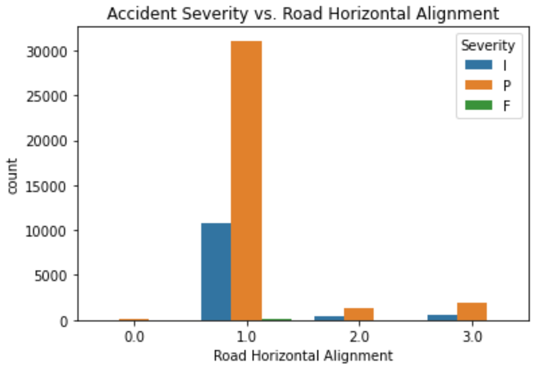
2 vehicles accidents: 23.1%

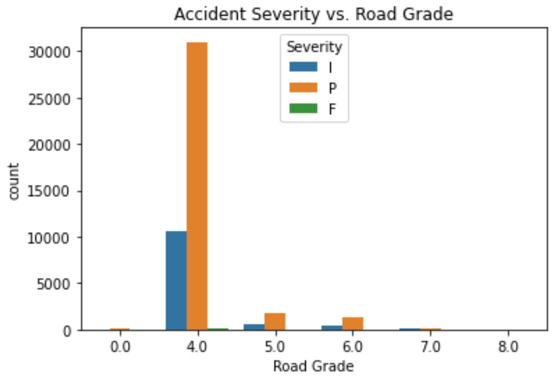
3 or more vehicles accidents: 43.8%

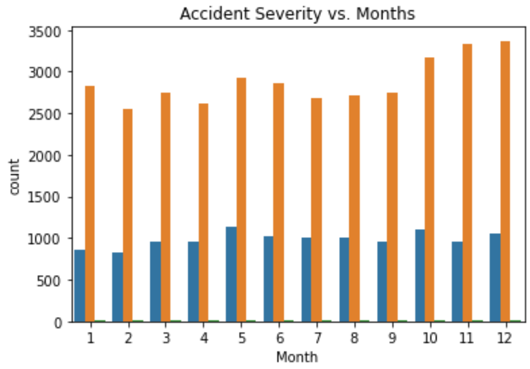
Pedestrian related accidents mostly involved only 1 vehicle, this type of accidents made the Ratio of the 1 vehicle accidents higher.

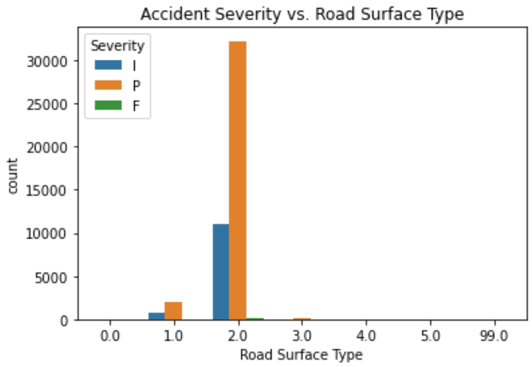
F. Other Crashed related features

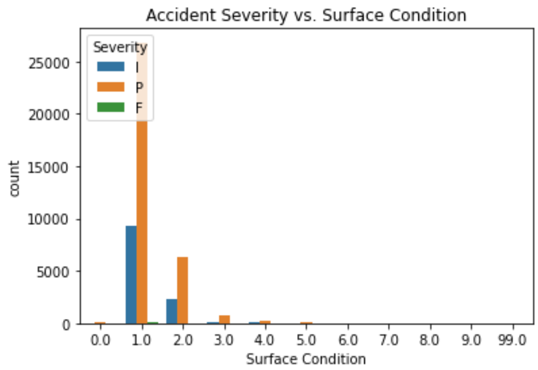
All the other crash related features were reviewed and they seemed to be not significantly related to the severity of the accidents.

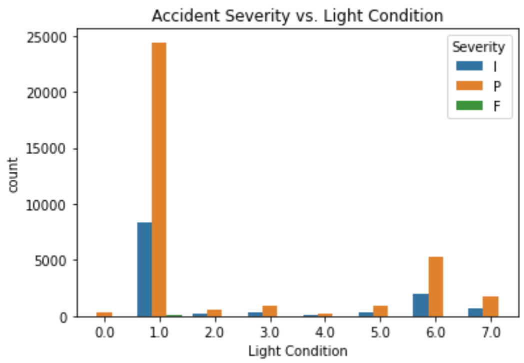


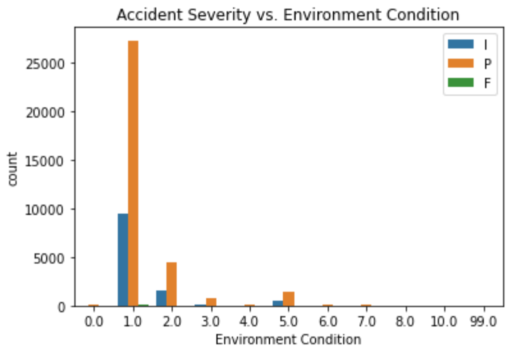


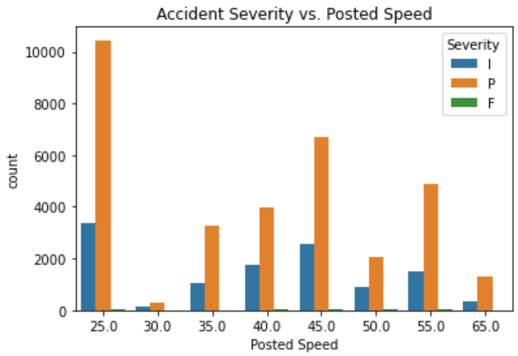








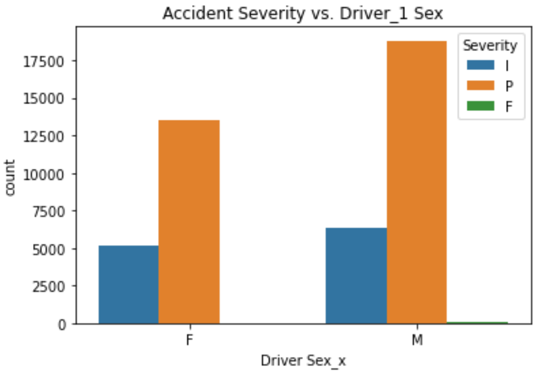


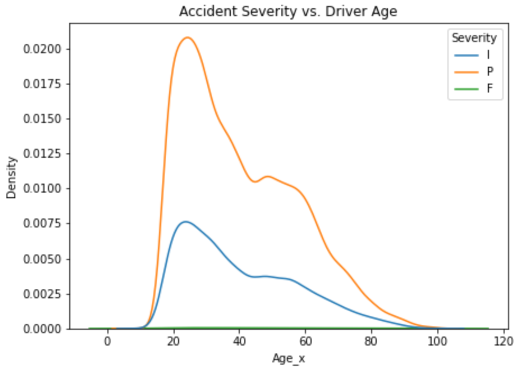


It is a little surprising that the vehicle speed (posted speed) is not significantly related to the severity of the accidents. I guess it is because at relatively low speed it is more often to have opposite direction crashes and pedestrian related crashes.

G. Drivers' features

Drivers' age and sex do not have much influence on the severity of the accidents. As shown on the graphs below, female drivers-1 have 27.8% of the injury or fatality ratio while male drivers-1 have 25.6% ratio. The right graph shows relationship between the driver-1 age and the accident severity, the age is not a significant feature for the severity of the accidents (drivers of all the ages seem to have almost same Ratios).

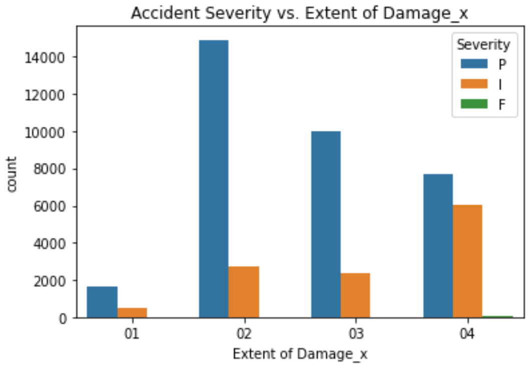


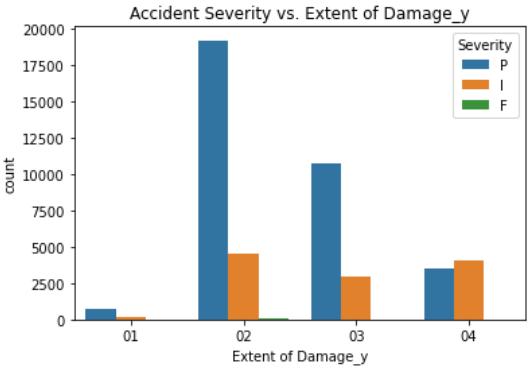


H. Vehicles Features

a. Extent of Damage

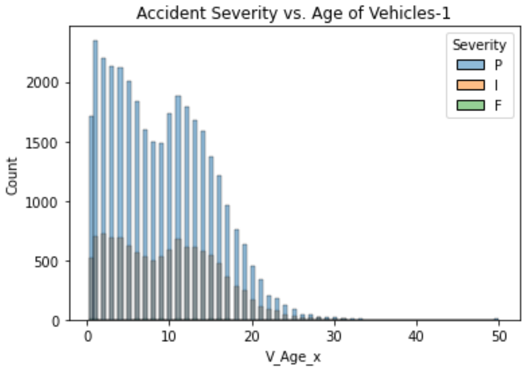
The damage to the vehicles is an obvious indicator to the severity od the accidents. The more damage to the vehicles can lead to higher possibility of the injury and fatality.

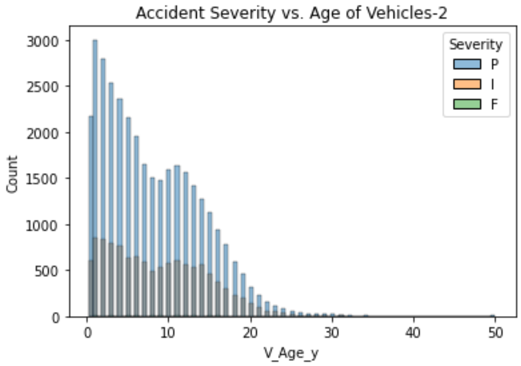




b. Ages of Vehicles

Like the ages of drivers, the ages of vehicles seem not to be a strong indicator of the severity of the accidents. The graphs show both the ages of vehicle-1(x) and vehicle-2(y) have no significant impact on the severity of the accidents.

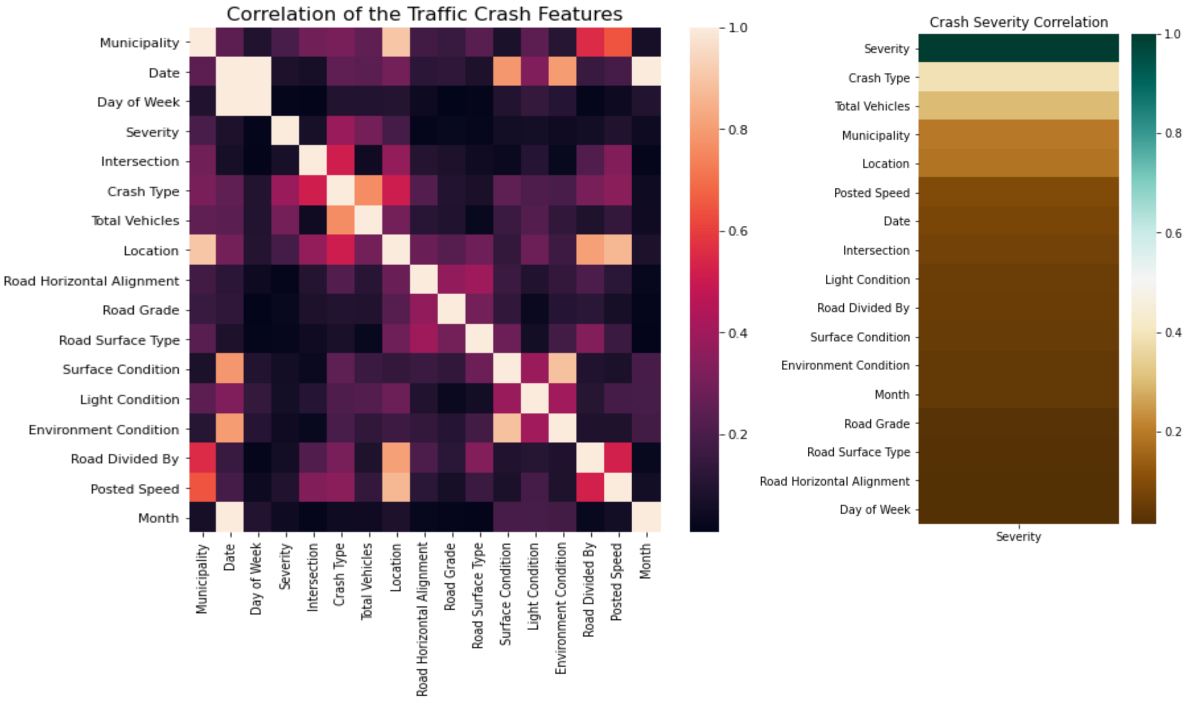




c. Other Vehicle features

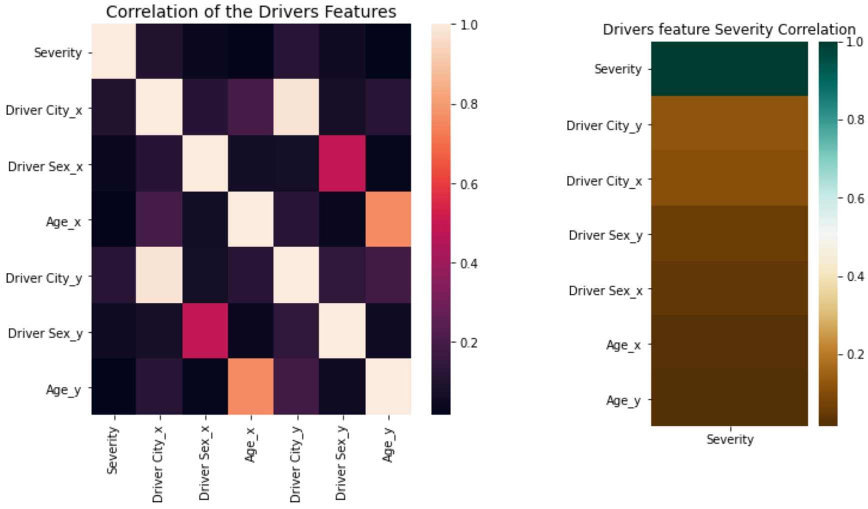
Other features such as 'Initial Impact Location', 'Principal Damage Location', ‘Vehicle Type' also have some influence on the severity of the accidents.

3) Since most of the features were categorical, I used a Phi-K correlation to see the correlation between these features. First I looked at the crash features.

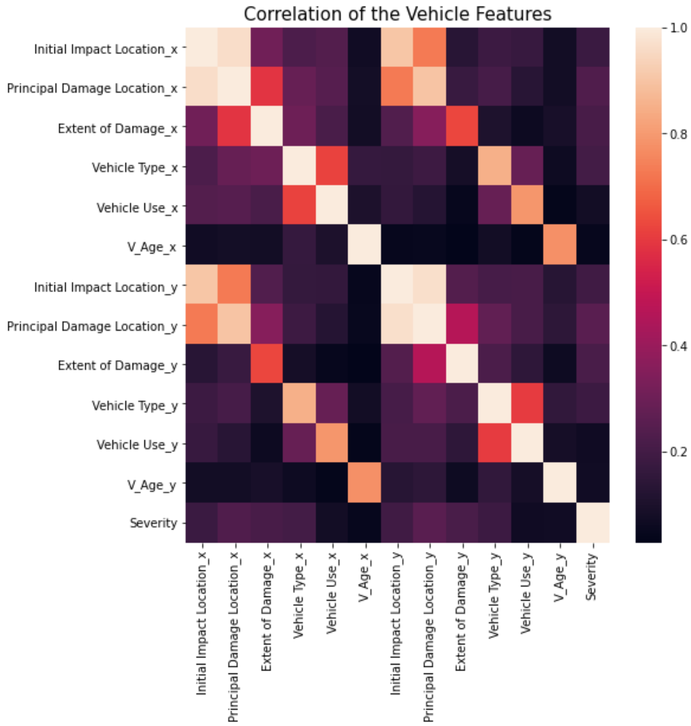


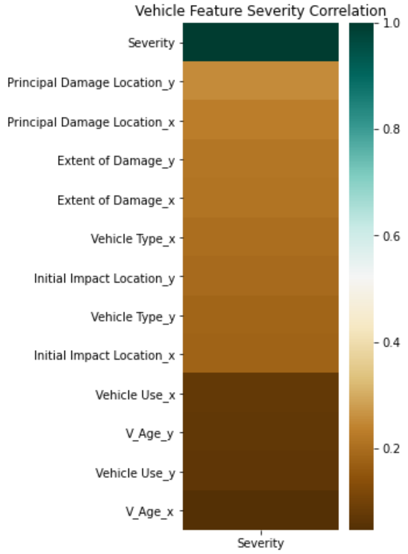
The left side of the above graph shows the correlations between all the features while the right side only shows the relationship between the severity of the crash and other features. It is consistent with my previous analysis, the crash type, total vehicles and municipality are top 3 features which has most influence on the severity. The location is an overlapping feature with municipality.

The PHI K correlations show drivers' feature have not much influence.



Finally, I looked at the vehicles' features:





The vehicle features do have some influences on the severity of the crashes.

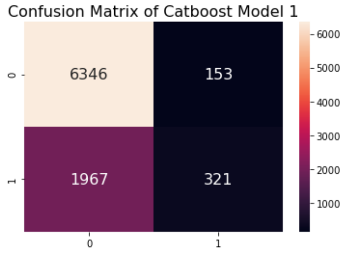
5. Machine Learning and Predictions

1) The preprocessing started after the explorative data analysis. I assigned the value 1 for the ‘I’ and ‘p’ of the accident severity, if it is only property damage ‘p’, then it is 0.
   Since most of the features are categorical, I chose CatBoost as the first choice of the algorithm for the prediction model. It is binary classification, Severity = 0 or 1 as the target value.

a. First I only chose the crash features: 'Municipality', 'Severity', 'Intersection', 'Crash Type', 'Total Vehicles', 'Road Surface Type', 'Road Divided By', 'Posted Speed', 'Month'.
   The categorical features are 'Municipality', 'Intersection', 'Crash Type' and 'Total Vehicles'.
   The numeric features are 'Posted Speed' and 'Month'.
   I still had 42,558 entries after dropping the null values. The data were randomly split to training set (80%) and test set (20%).
   The CatBoost worked and the results were:

```
              precision    recall  f1-score   support

           0       0.76      0.98      0.86      6499
           1       0.68      0.14      0.23      2288

    accuracy                           0.76      8787
   macro avg       0.72      0.56      0.54      8787
weighted avg       0.74      0.76      0.69      8787
```

The confusion matrix is:

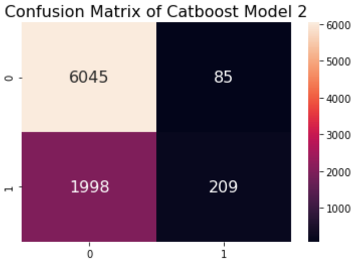


b. Second time I added drivers' features: 'Driver Sex_x', 'Age_x', 'Driver Sex_y', 'Age_y' and used 'location' instead of 'municipality'.

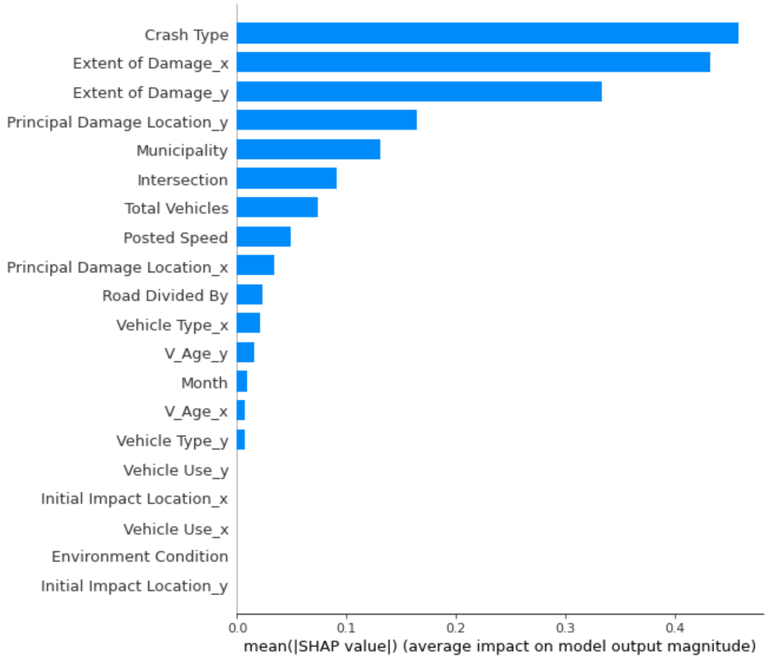
The Catboost's results are:

```
              precision    recall  f1-score   support

           0       0.75      0.99      0.85      6130
           1       0.71      0.09      0.17      2207

    accuracy                           0.75      8337
   macro avg       0.73      0.54      0.51      8337
weighted avg       0.74      0.75      0.67      8337
```

The confusion matrix is:



SHAP Analysis:



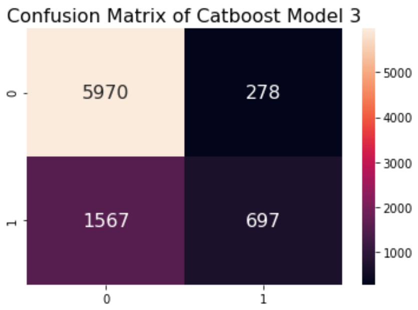
The second Catboost model test results were even worse than the first one because I did not choose the right features, and the SHAP analysis proved it.

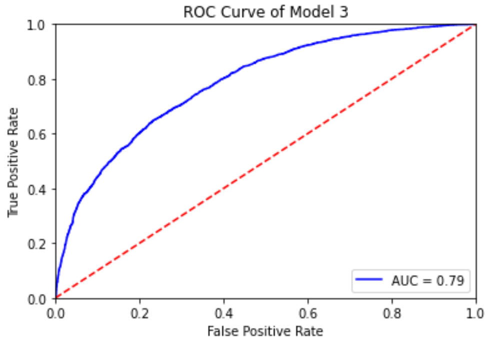
c. Third time I added vehicles features, the features I used were 'Severity', 'Intersection', 'Crash Type', 'Total Vehicles', ‘Environment Condition', 'Road Divided By', 'Posted Speed', 'Month', 'Initial Impact Location_x', 'Principal Damage Location_x', 'Extent of Damage_x', 'Vehicle Type_x', 'Vehicle Use_x', 'V_Age_x', 'Initial Impact Location_y', 'Principal Damage Location_y', 'Extent of Damage_y', 'Vehicle Type_y', 'Vehicle Use_y', 'V_Age_y'.

   With these features, the results were:

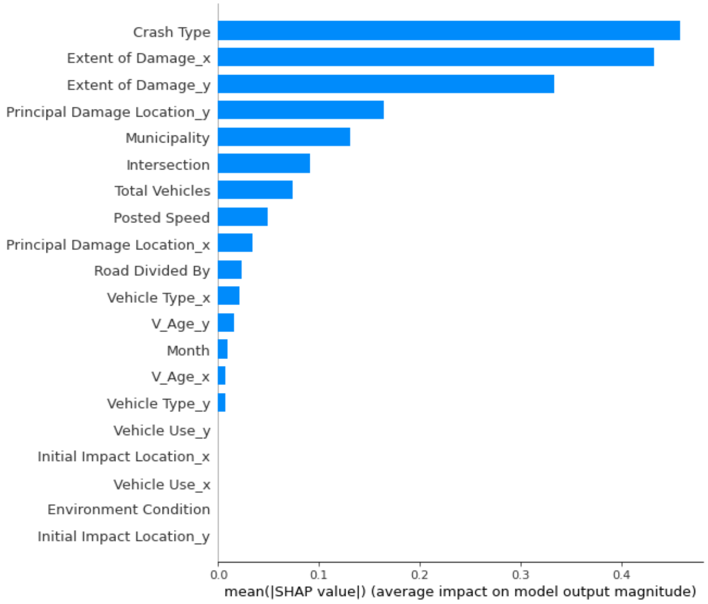
```
              precision    recall  f1-score   support

           0       0.79      0.96      0.87      6248
           1       0.71      0.31      0.43      2264

    accuracy                           0.78      8512
   macro avg       0.75      0.63      0.65      8512
weighted avg       0.77      0.78      0.75      8512
```

The confusion matrix and ROC curve:





SHAP analysis:



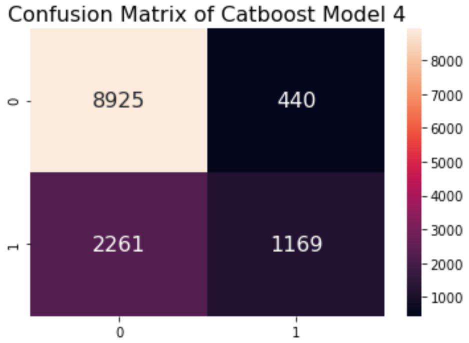
d. The fourth time I changed the features to:

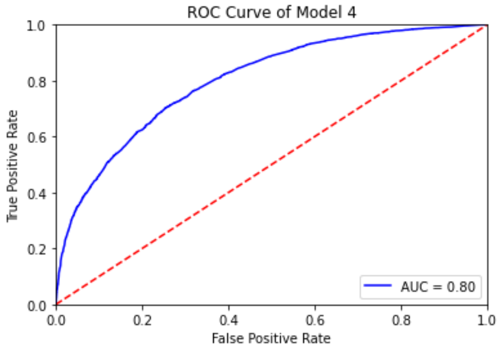
'Municipality', 'Intersection', 'Crash Type','Total Vehicles', 'Initial Impact Location_x', 'Principal Damage Location_x', 'Extent of Damage_x', 'Vehicle Type_x', 'Initial Impact Location_y', 'Principal Damage Location_y', 'Extent of Damage_y' and 'Vehicle Type_y'.

Compared with the previous time, I added ‘Municipality’ but dropped ‘Environment Condition', 'Road Divided By', 'Posted Speed', 'Month', 'V_Age_x' and ‘V_Age_y’.

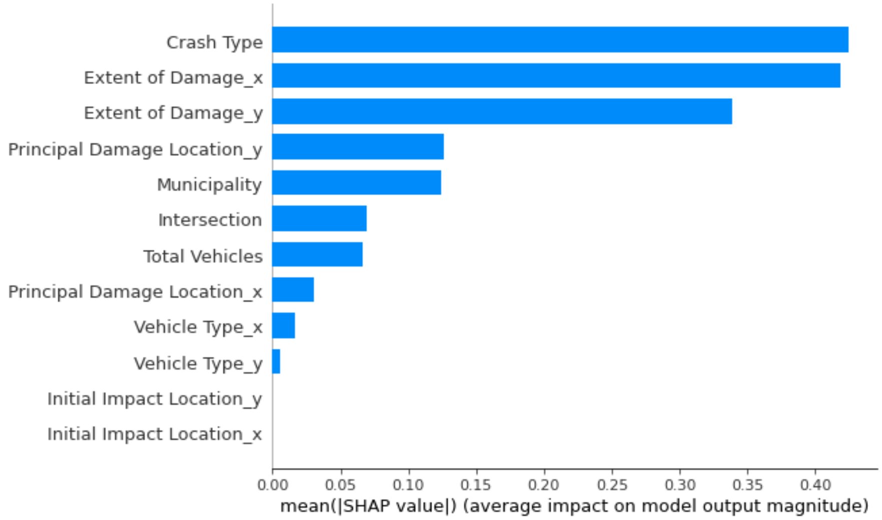
The results improved again:

```
              precision    recall  f1-score   support

           0       0.80      0.95      0.87      9365
           1       0.73      0.34      0.46      3430

    accuracy                           0.79     12795
   macro avg       0.76      0.65      0.67     12795
weighted avg       0.78      0.79      0.76     12795
```





The SHAP:

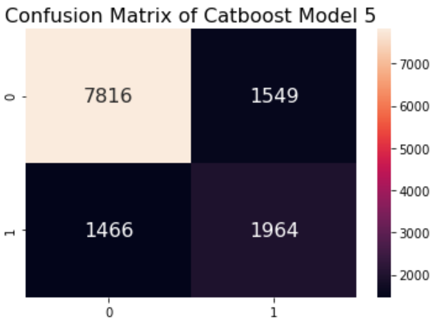


e. SMOTE method for imbalanced classification

Although the results improved, the problems remain. The model was biased against minor class. The recall of 1 class was 0.34. To improve the minor class recall, I used the SMOTE method to make a synthetic balanced data and converted the training data from 22187 (0)/ 7767 (1) to 22187 (0)/22187(1). I used the same features as the previous model.

With a balanced training data, the results were also more balanced:

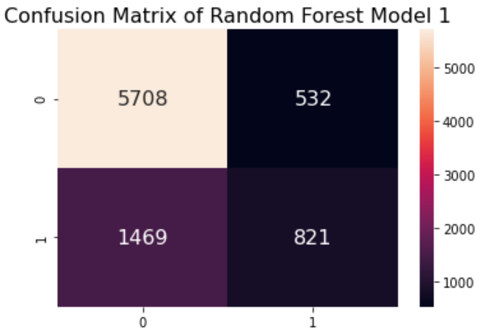
```
              precision    recall  f1-score   support

           0       0.84      0.83      0.84      9365
           1       0.56      0.57      0.57      3430

    accuracy                           0.76     12795
   macro avg       0.70      0.70      0.70     12795
weighted avg       0.77      0.76      0.77     12795
```



f. Random Forest Model

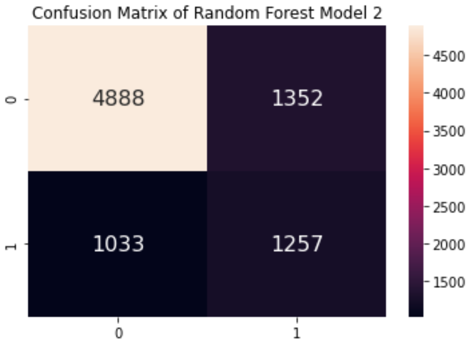
I used the binary encoding to convert categorical variables to binary digits. The training data was 25312 ‘0’ and 8807 ‘1’. The model training rate was significantly faster when I used the binary digits. The results were as follows:

```
              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
```



It is still unbalanced, so the smote method was used again. The training data was converted from 25312/8807 to 25312/25312. After the model training, the test results were:

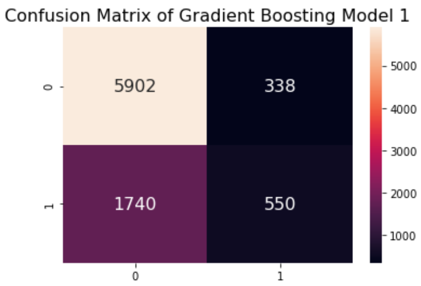
```
              precision    recall  f1-score   support

           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
```



g. Gradient Boosting Model

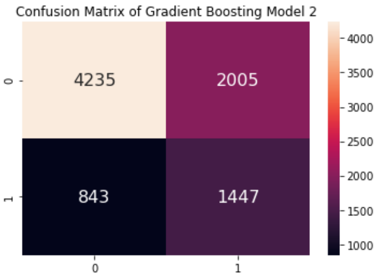
The Gradient Boosting Model used the same training data as the random forest model. Before the SMOTE, the test results were as follows:

```
              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
```

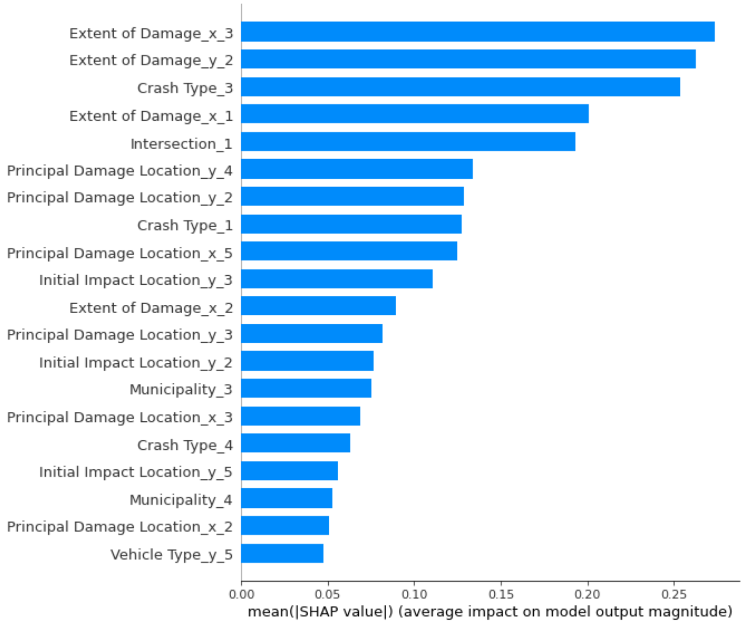

Confusion Matrix of Gradient Boosting Model 1

After SMOTE:

```
              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

SHAP of the Gradient Boosting Model:



6. Compare the Models

It is obvious that the SMOTE method is necessary to improve the minor class prediction. After SMOTE, the comparisons of the test results of CatBoost, Random Forest and Gradient Boosting are as follows:

	CatBoost	Random Forest	Gradient Boosting
Accuracy	0.76	0.72	0.67
Classification Table	Precision Recall 0 0.84 0.83 1 0.56 0.57	Precision Recall 0 0.83 0.78 1 0.48 0.55	Precision Recall 0 0.83 0.68 1 0.42 0.63

It is obvious that CatBoost worked best in this mostly categorical feature prediction.

7. Summary

In this data research project, I have found and accomplished:

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 significantly. The number of accidents in each county is closely related to the population of each county.
2. The traffic accident occurrences vs. hours of a day, days of a week, holidays and months were studied.
3. A classification model to predict the severity (causing injury/death or just property damage) of a traffic accident has been established, it can achieve close to 80% accuracy. The Catboost model with SMOTE method worked best to have a balanced prediction and highest 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.