IBM Capstone Project – Predicting car accident severity in Seattle city.

By

SHASHANKA RANGI

- In this project, we developed models that can predict the occurrence of an accident and it's severity based on number of factors such as environmental factors, driver behavior, traffic conditions and location. The results from the model predictions prove that the information from the data can be used for the prediction of future accidents with a given set of conditions.

- Data needed for this project was provided by the IBM data science coursera course.

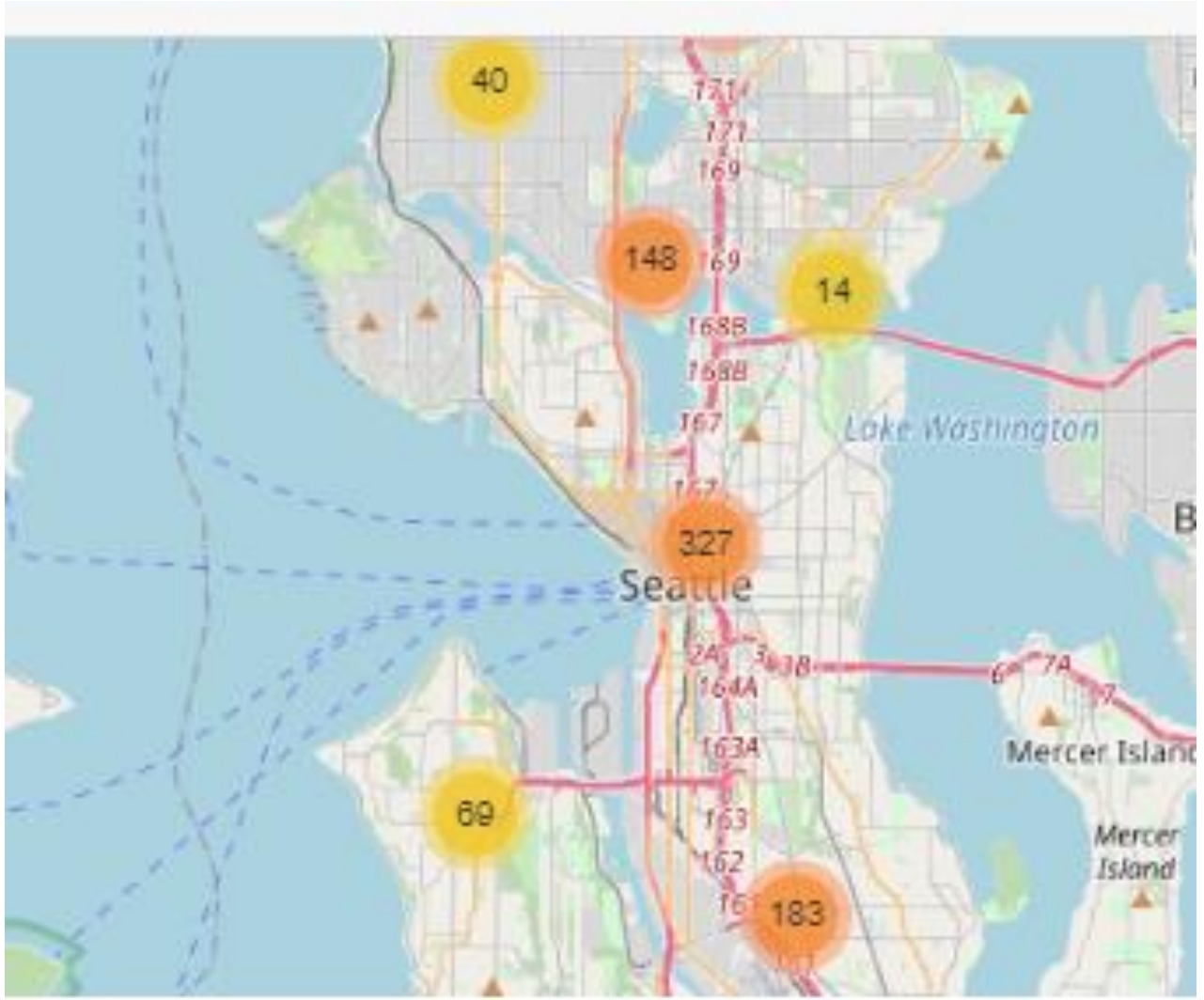
- The data consists of 37 independent variables and 194,673 rows. The dependent variable, "SEVERITYCODE", contains numbers that correspond to different levels of severity caused by an accident represented by 1 and 2.

  Severity codes are as follows:
  1: Very Low Probability — Chance or Property Damage
  2: High Probability — Chance of injury and fatality.

- Date cleaning was performed, and important features needed for the prediction were selected. The dataframe now has, one target variable severitycode and 13 independent variables i.e. address type, collision type, pedestrian count, pedestrian and cycle count, vehicle count, junction type, inattentionind(whether the driver is paying the attention or not), whether the driver is under influence or not, weather conditions, road conditions, light conditions, speeding, whether the car has hit parked car or not and hour of the day when the accident severity is high.

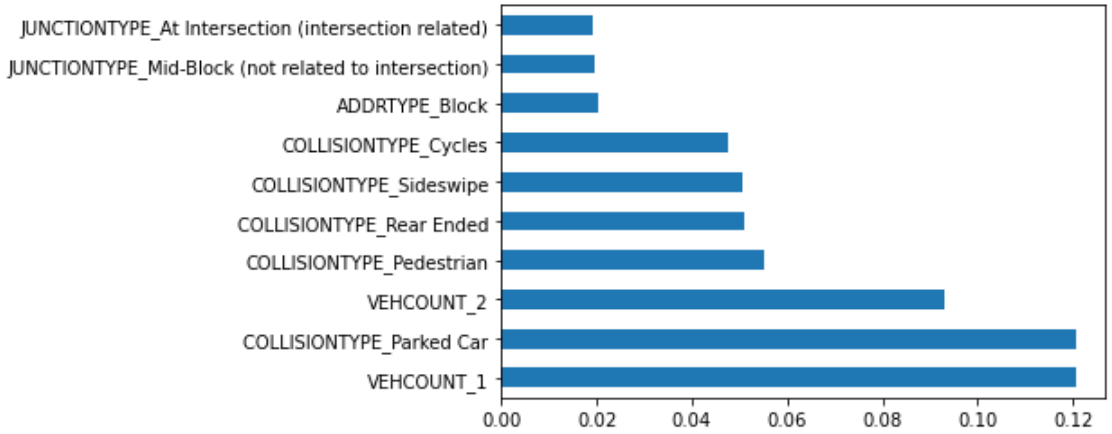
Visualizing collisions on the map.

A quick look at the map tells us that although there are accidents happening everywhere around the city, there seems to be more accidents concentrated in the center of the city. Let's explore further and understand the reason for accident severity.



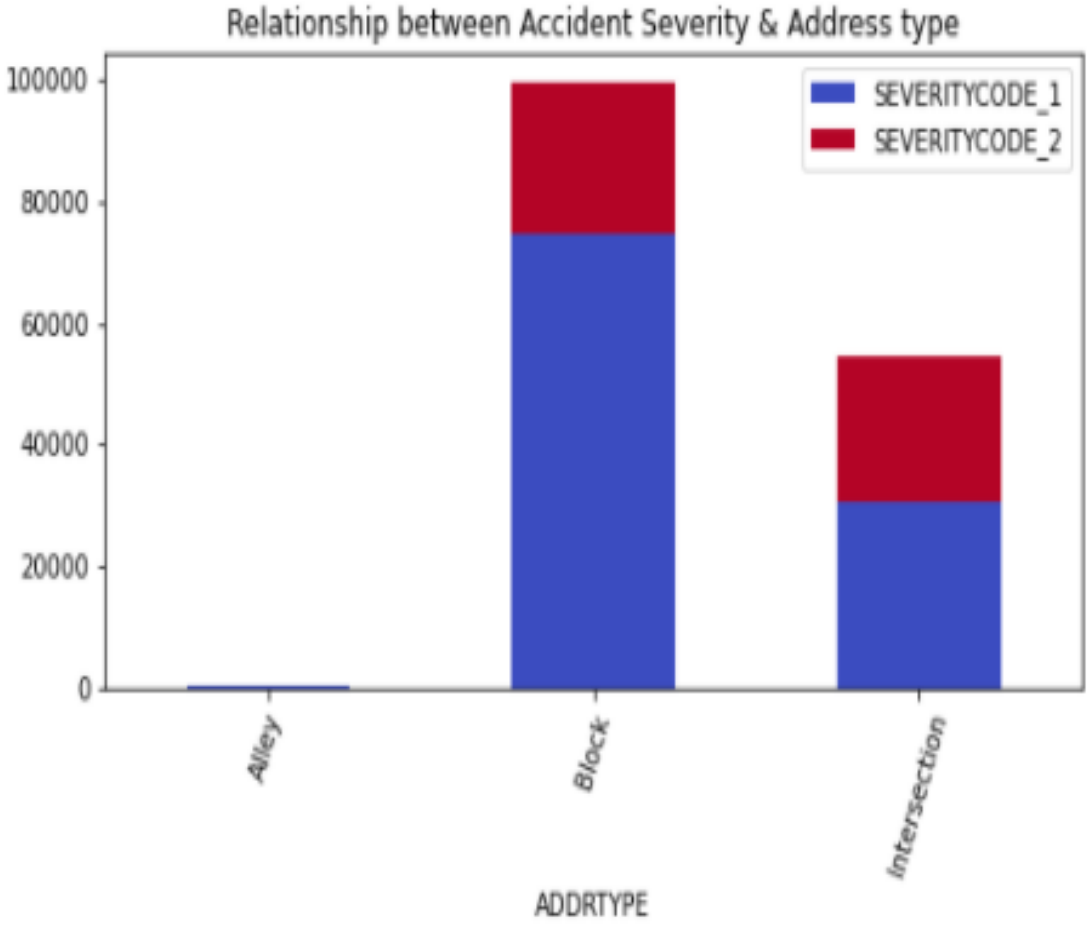
Important features

The graph on the right-hand side shows the features with high accident severity code.



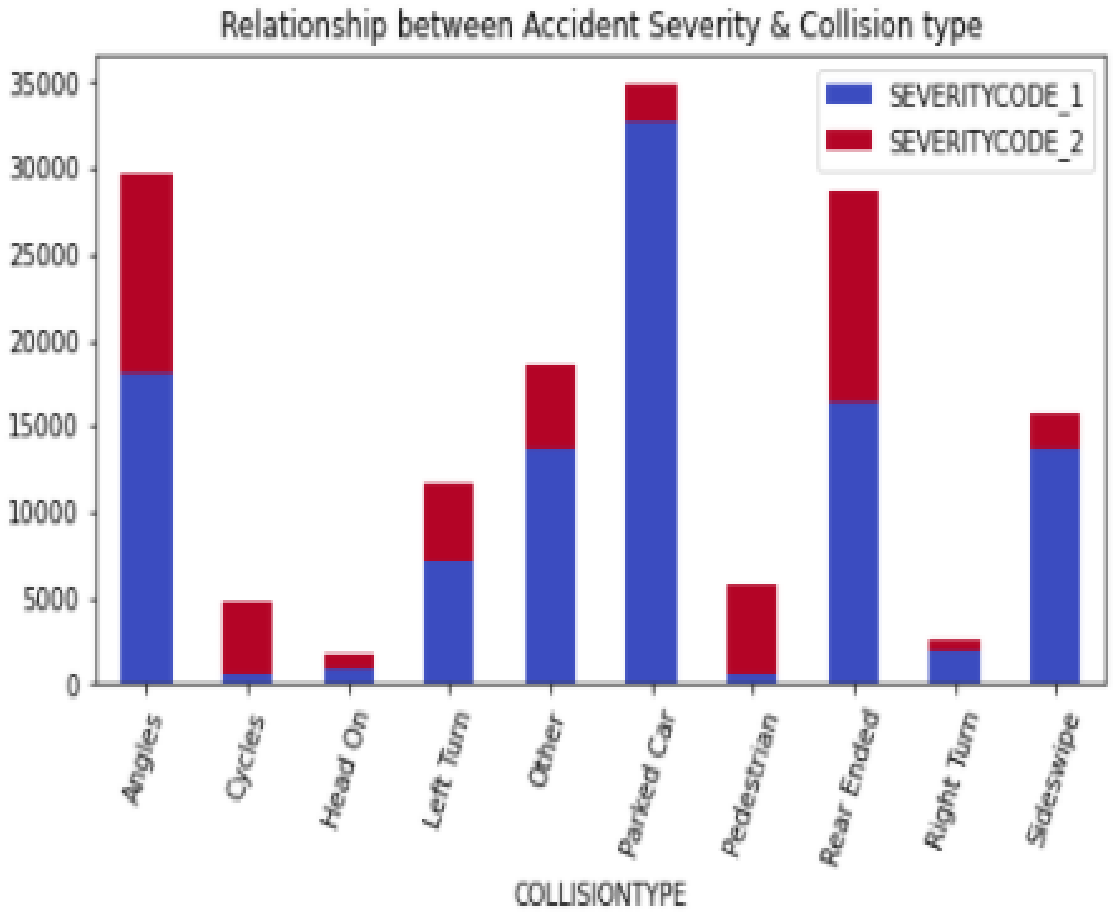
Exploring the relationship between address type and severitycode

From the graph, although it appears that there are more accidents happening near blocks, the accident severity seems to be high near the intersections. Traffic monitoring authorities can use this data to enforce more traffic regulations near blocks and intersections.



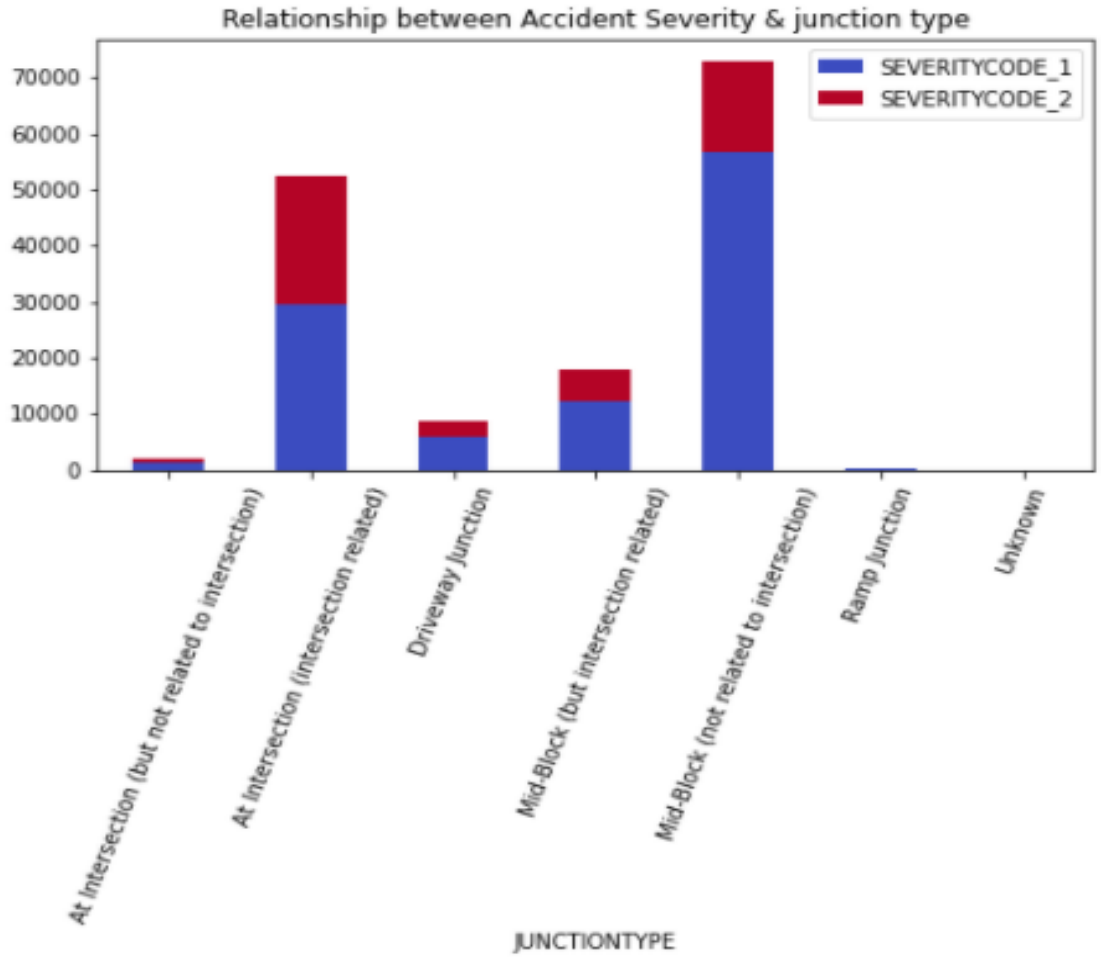
Exploring the relationship between Collision type and severitycode

From the collision graph, we can see that majority of the collisions happen with a parked car, but the severity of the accidents is high when collision happens from rear-end or angles. Car manufacturing companies can use this information to manufacture car models with more safety features like auto-park, parking sensors, lane departure warnings, auto-braking etc.



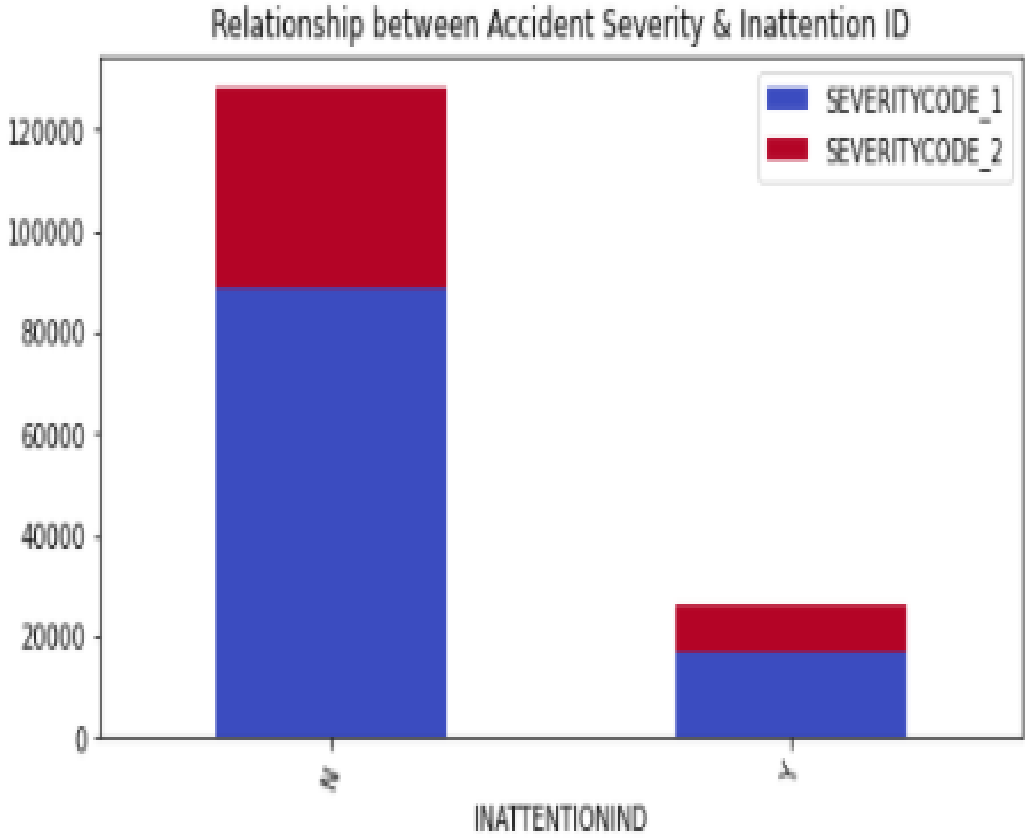
Exploring the relationship between junction type and severitycode

From the graph, we can note that majority of the accidents at intersection and mid-block, with high severity being at the intersections. Like discussed above in the Address type graphs, more traffic enforcements should be in place at intersections to avoid accidents and its severity.



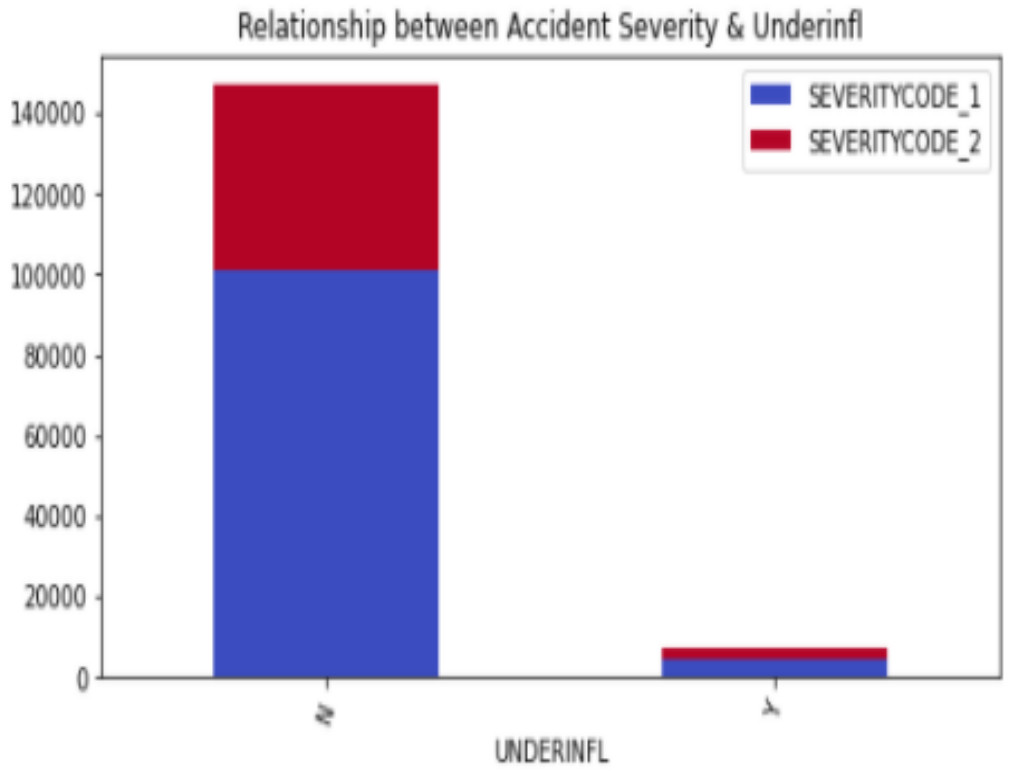
Exploring the relationship between driver's attention to road and severitycode of the accident

It is interesting to note that not many accidents happen due to driver's inattention. And, also with accidents happening due to drivers' inattention the severity looks low.



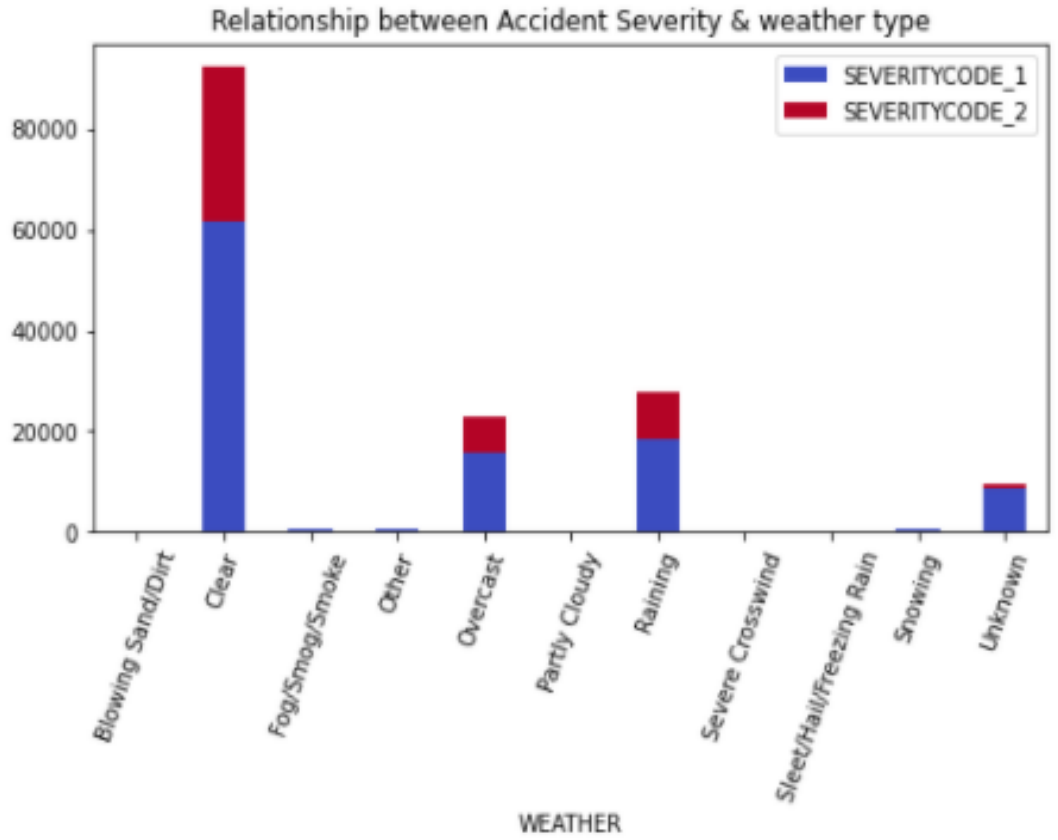
Exploring the relationship between influence of alcohol and severitycode

Like driver's inattention, the accidents happening under influence also appears to be low with low severity.



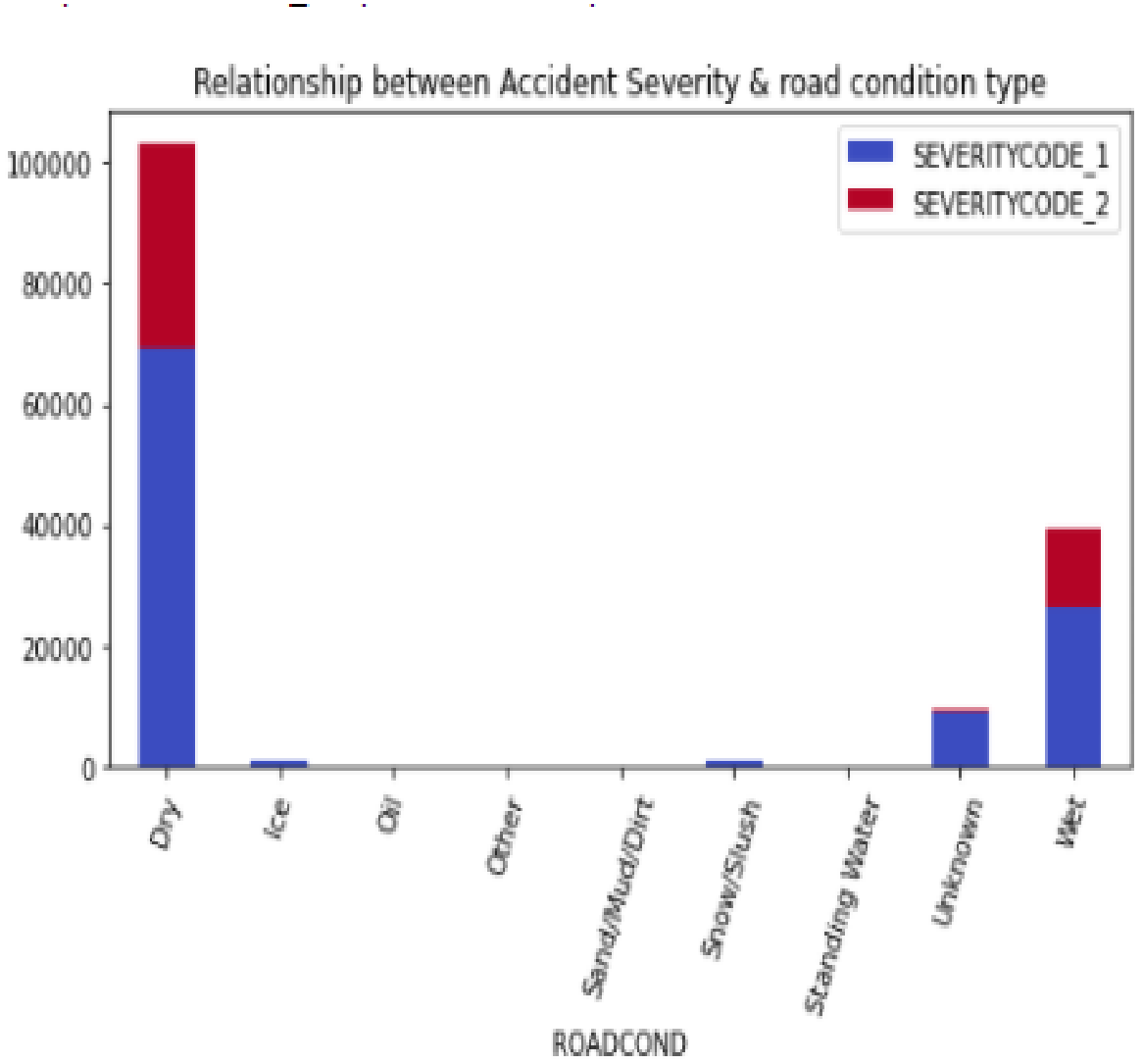
Exploring the relationship between weather condition and severitycode

From the bar graph, surprisingly majority of the accidents happen on a clear day and with high severity. Rainy days and overcast days also have some accidents.



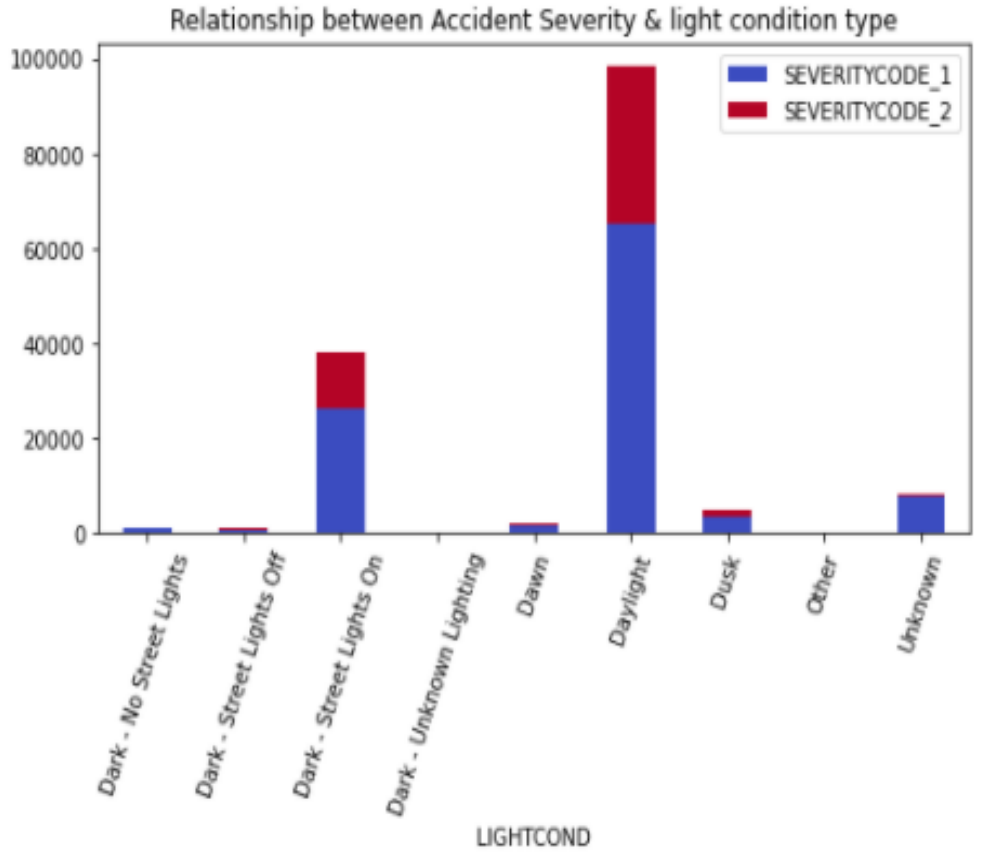
Exploring the relationship between road condition and severitycode

Just like the weather, more accidents and with high severity happen on a dry day. From the weather and road conditions graphs, we can see that light is one of the contributing factors for the accidents. Drivers should consider wearing sunglasses or other appropriate safety equipment when the weather is clear, dry and sunny to avoid being in an accident.



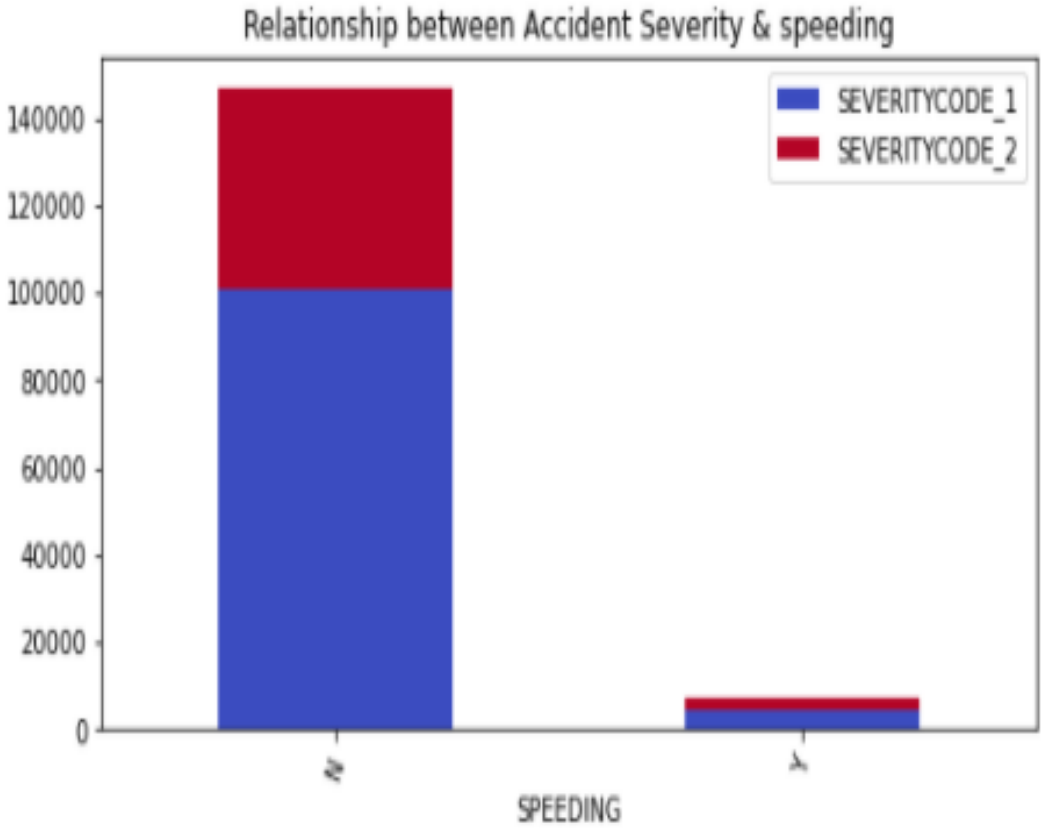
Exploring the relationship between light condition and severitycode.

This graph gives us another reason to believe that light is contributing to the accidents, as we can see more accidents happen in the daylight and the severity is also high. Second high variable, dark-street lights on also adds to the fact that light is the very important cause for the accidents.



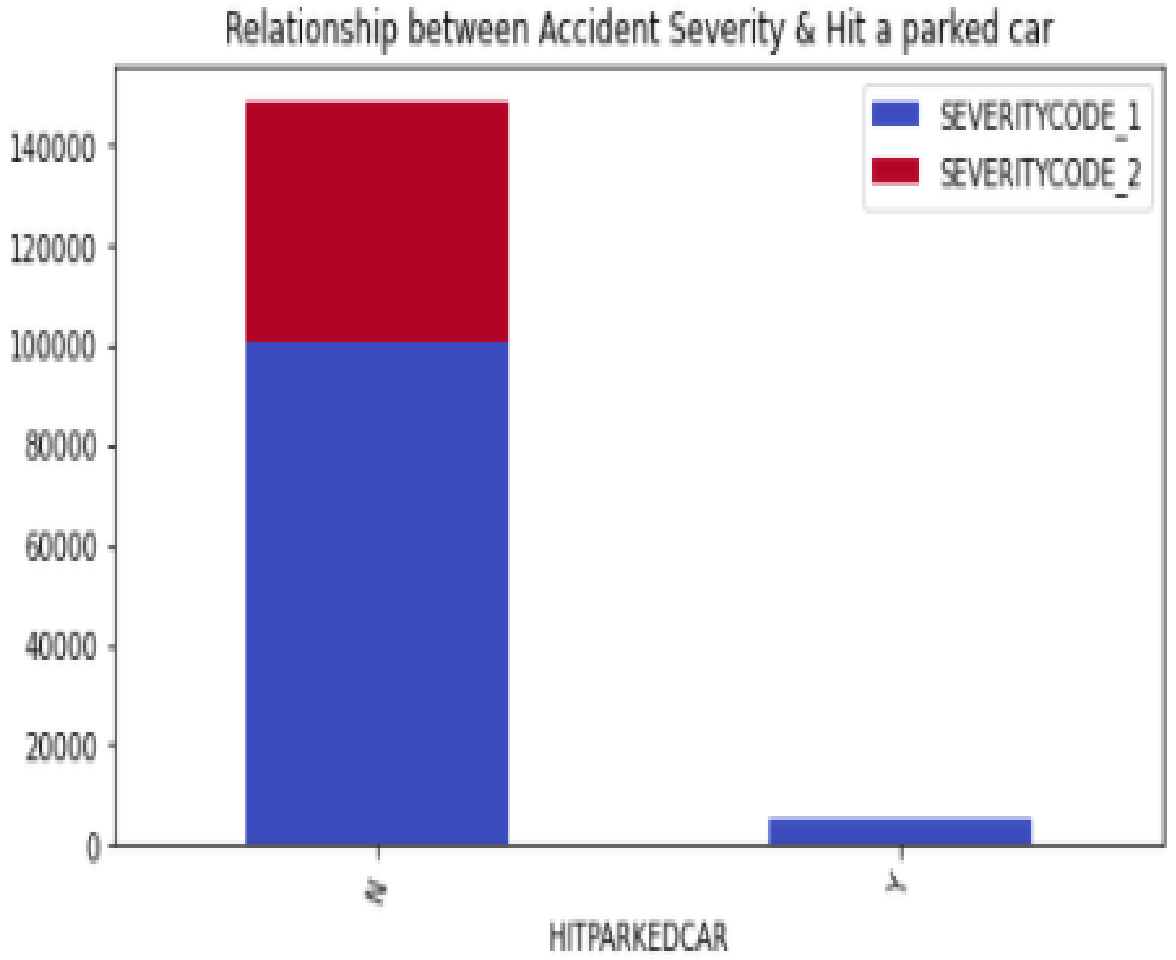
Exploring the relationship between speeding and severitycode

Not many accidents have occurred due to speeding, and the severity is also low.



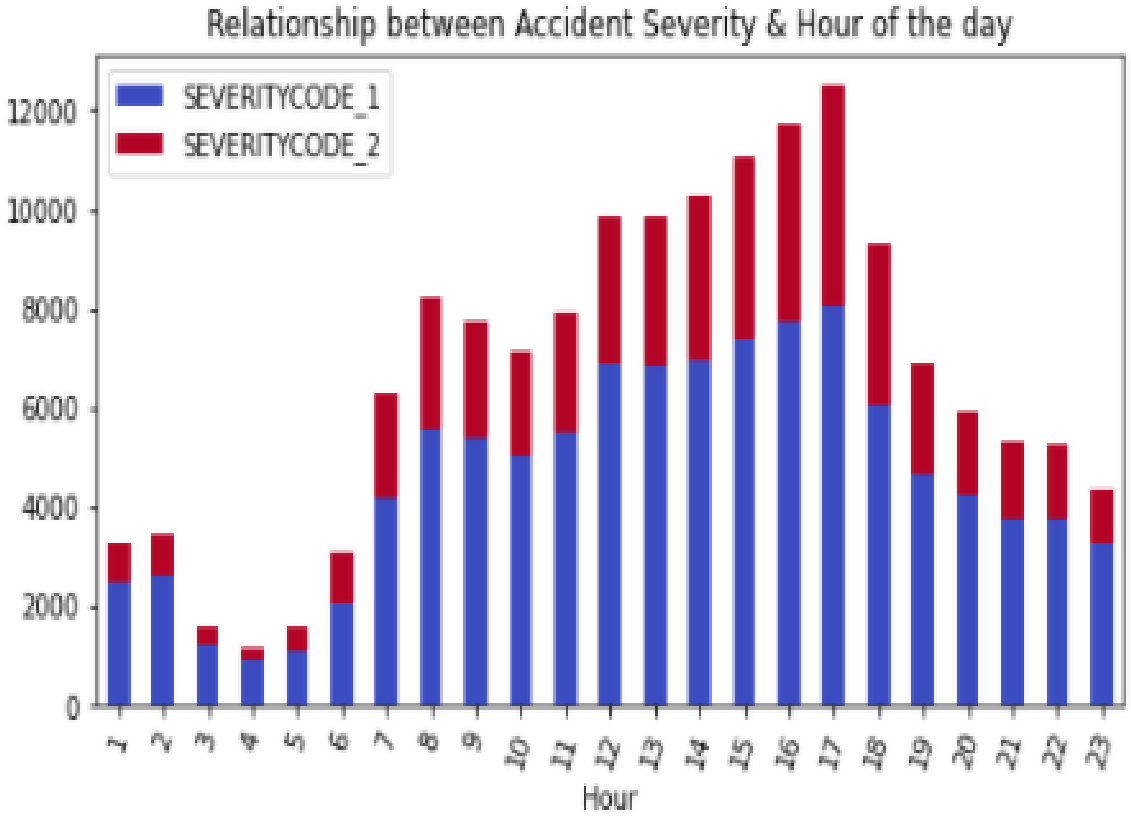
Exploring the relationship between whether the car is parked or not and severitycode

Not many accidents have occurred when the car is parked, and the severity is also low. Majority accidents happened when the cars are in motion, so another reason to suggest cars should be designed with safety as number one priority and strict enforcement will traffic rules will help in reducing the accidents.



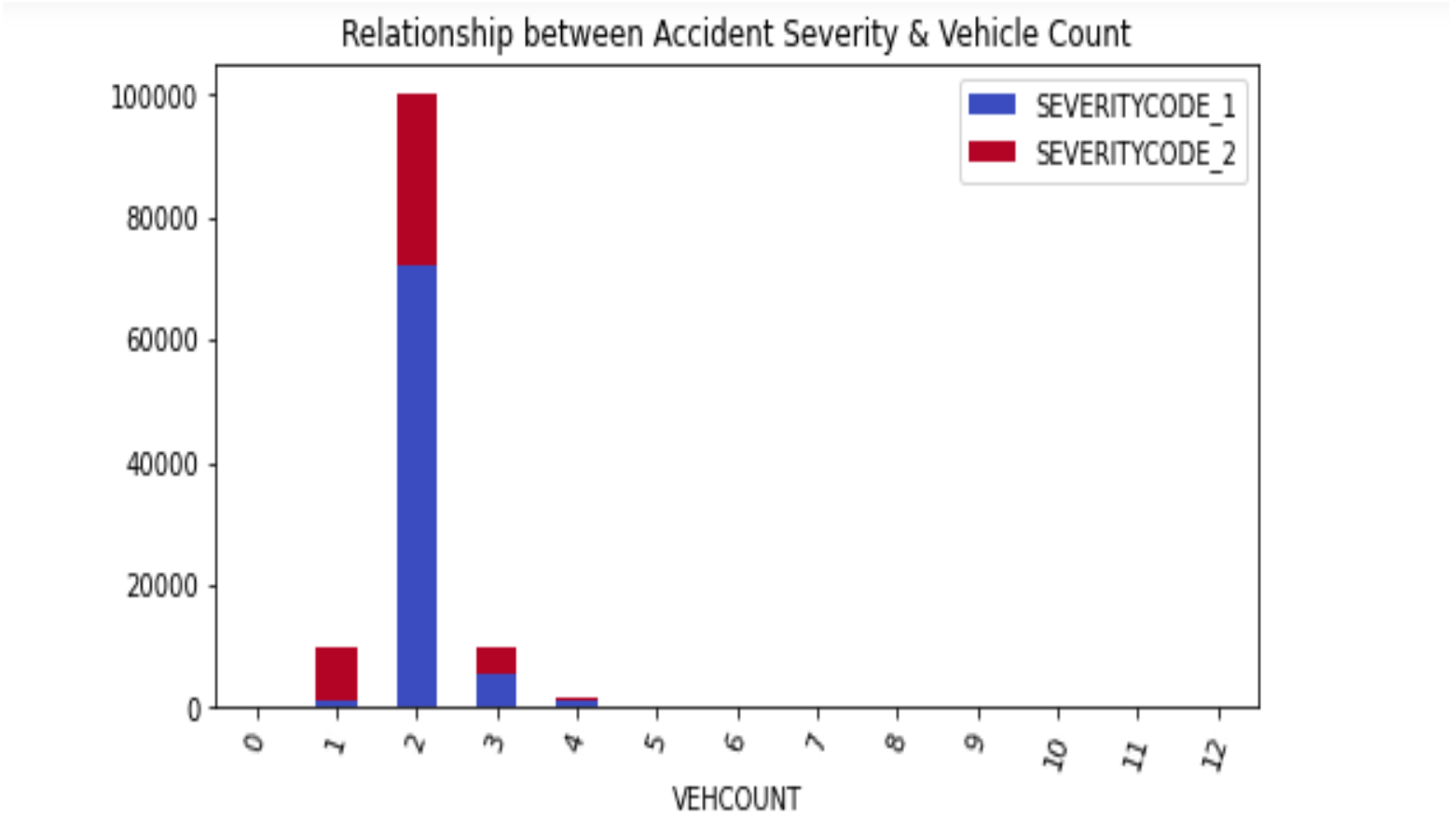
Exploring the relationship between hour of the day and severitycode.

The graph tells us that, accidents happen between 7AM to 7PM and the peaks can be observed during the rush hour 2PM to 5PM.



Exploring the relationship between number of vehicles involved and severitycode

Accident severity seems to be high when two to three vehicles are involved.



Modelling:

The data has been trained and models have been developed using three different algorithms:

- Logistic regression
- K-NN classifier
- Decision Tree Classifier

Model evaluation

Confusion matrix and accuracy values can be found on the right hand side.

```
Confusion matrix of Logistic Regression:
LogisticRegression()
Accuracy score of logistic regression is  0.7246654699738904
Confusion matrix of K-NN classifier:
[[13035  2992]
 [ 4600  3885]]
Accuracy score of K-NN classifier is  0.6902741514360313
Confusion matrix of Decision Tree classifier:
[[13035  2992]
 [ 4600  3885]]
Accuracy score of Decision Tree classifier is:  0.6902741514360313
```

Results and Future work

- From the model evaluation, it appears that logistic regression model has a better accuracy in prediction the test results when compared to other models.
- As a future work , trained models from this work can be deployed onto various traffic monitoring applications used by police, health and emergency response teams and also can be used to develop apps to alter drivers on road and weather conditions to predict accident severity with a given set of conditions.
- Also, car manufacturing companies can use these data to develop car models with more safety options.