# 1. Introduction

(A description of the problem and a discussion of the background.)

## 1.1. Background

The commuting habits determine the appearance, social relationships, and, to some extent, the individual's health. The reason being State and local governments well maintain the overwhelming majority of roads, cars are the number one mode of transportation in the United States.

The Seattle Department of Transportation (SDOT) develops, maintains, and operates a transportation system that promotes the mobility of people and goods, and enhances the quality of life, environment, and economy of Seattle. However, there are thousands of car accidents that are occurring every day in Seattle. Alcohol and drug impairment, driver's distractions, weather, speeding, road conditions, vehicle malfunction, and many can cause road accidents.

## 1.2. Problem

## Most people experience the aftereffect of a car collision at some point in their lives, but everyone is not clear on how to avoid it. Several uncontrollable factors cause car collisions. There are three car collisions stages: car accident, vehicle damage, and human injury. If we do not solve or reduce the number of car collisions, we might face more fatalities or different kind of severities.

## 1.3 Interest

Data that might help convey car collision reasons based on the controllable and uncontrollable factors that might help the Seattle department of transportation, Police, Insurance companies, and local people.

By making insightful decisions based on data, the model can benefit the target audience to control and reduce car collisions.

# 2.Data Traversing and Cleaning

(A description of the data and how it will be used to solve the problem.)

## 2.1 Data source

Seattle Department of Transportation collects all the collision data and records by traffic records. Data includes all the types of collisions like angles, sideswipe, parked cars. These collisions will be displayed at the intersection or mid-block of a segment. The timeframe of data is from 2004 to the present.

I have downloaded collision data from <http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.csv>

and saved into Collisions.csv.

## 2.2 Data cleaning (Data Preprocessing)

The data was import for analysis from the Seattle Department of transportation. The downloaded data is stored temporarily into the collision.csv file before starting data analysis; the data must be clean.

There were several problems with the dataset which needed to be fixed before starting the data analysis. Firstly, the dates field was not in ML algorithm accepted format, as INCDATE was not useful, so dropped the INCDATE column, and the INCDTTM column is split into three columns: Month, Weekday, and hour.

Below are a few dropped few columns which are not considered while creating a model.

INCDATE', 'INCDTTM','OBJECTID', 'INCKEY', 'COLDETKEY', 'INTKEY', 'SEGLANEKEY', 'CROSSWALKKEY','EXCEPTRSNDESC', 'SEVERITYDESC', 'SDOT\_COLDESC', 'ST\_COLDESC', 'LOCATION' and SDOT\_COLCODE.

PEDROWNOTGRNT, SPEEDING and INATTENTIONIND variables have missing values which are substituted by 'Unknown'.

As Machine learning algorithms takes only categorical values (numerical values) so it is imported to convert all below variables into categorical value and later checked if all the columns have int data type or not.

'SEVERITYCODE','ADDRTYPE','COLLISIONTYPE', 'JUNCTIONTYPE', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SPEEDING', 'ST\_COLCODE', 'HITPARKEDCAR'.

# 3. Exploratory Data Analysis

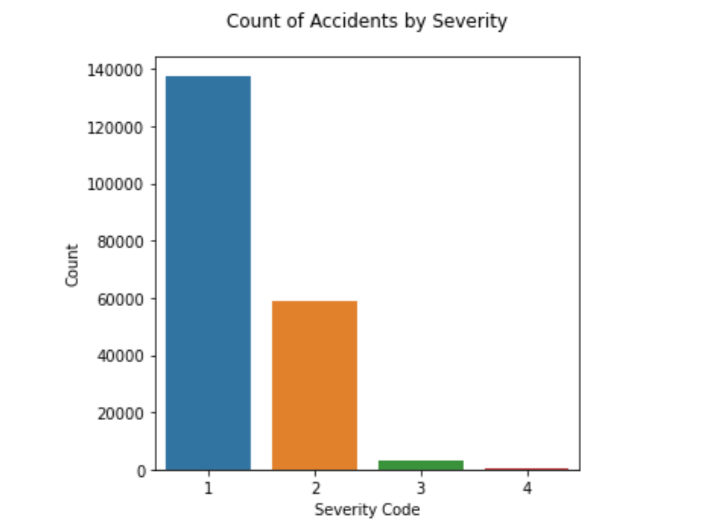
## 3.1 Calculation of target variable

I chose the SEVERITY\_CODE as the target variable and found out the counts of each code.

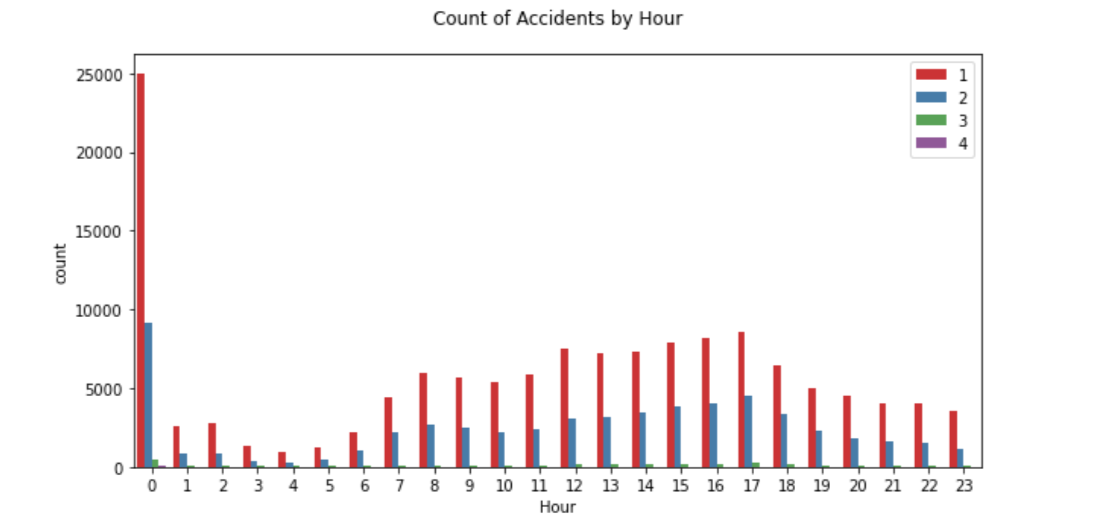
A code that corresponds to the severity of the collision:

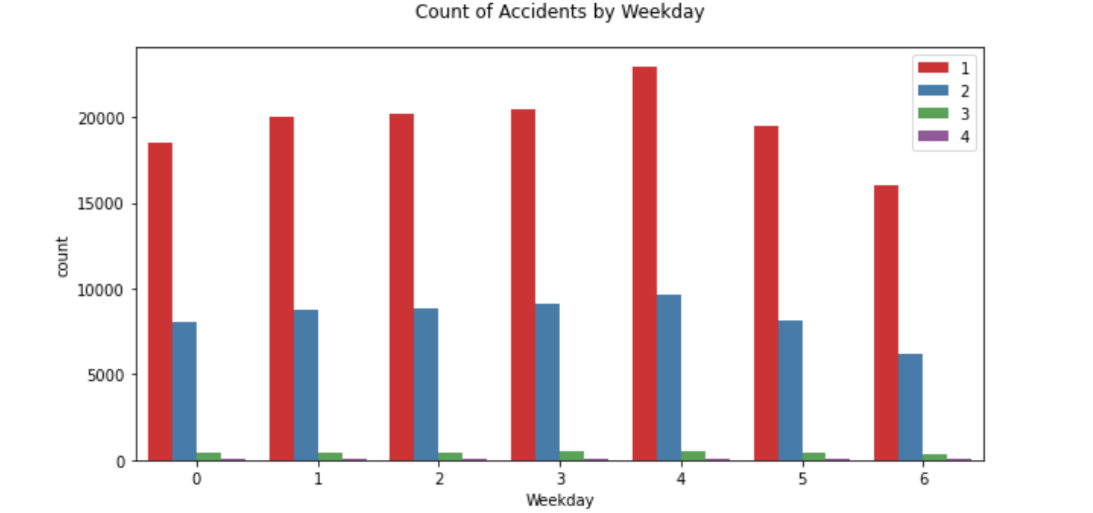
* 3—fatality
* 2b—serious injury
* 2—injury
* 1—prop damage
* 0—unknown

## 3.2 Relationship between count of accident as per each severity code.



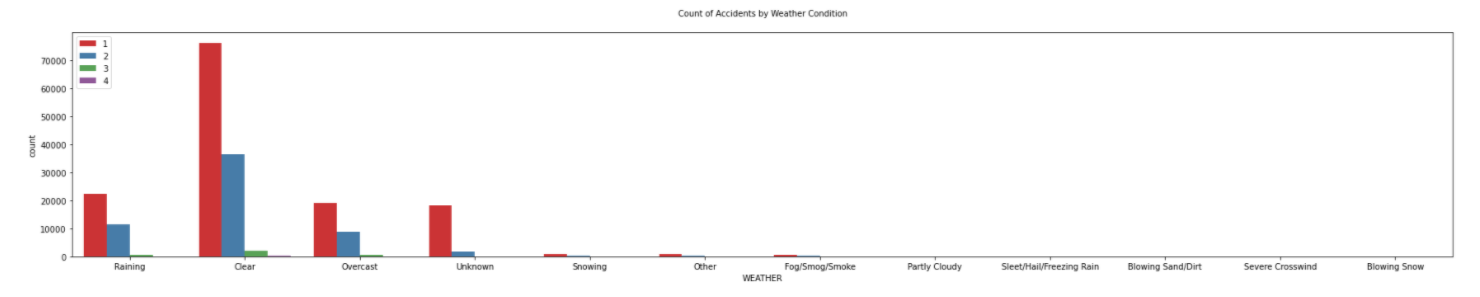
## 3.3 Relationship between date (Hour, weekday and month) and accident count.



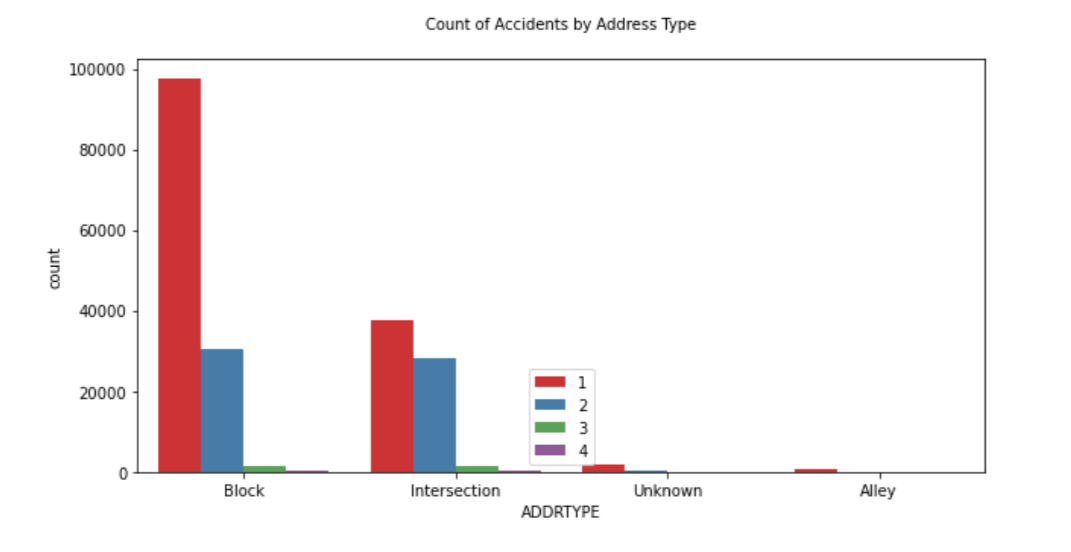




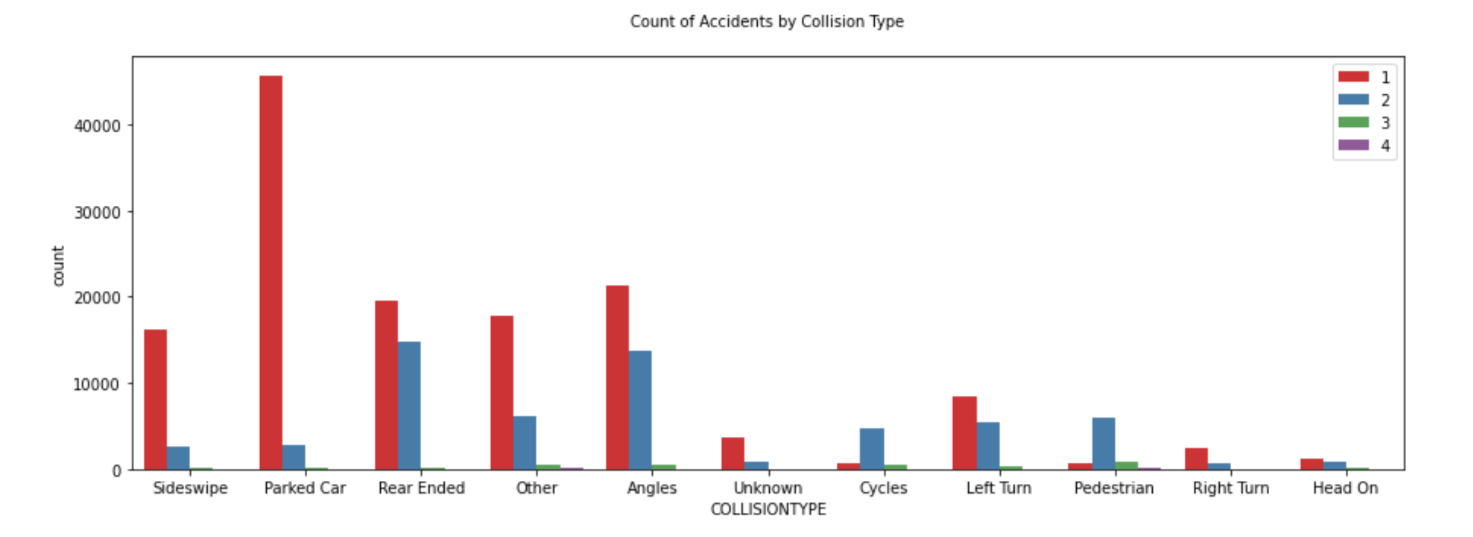
## 3.4 Relationship between weather and accident count.



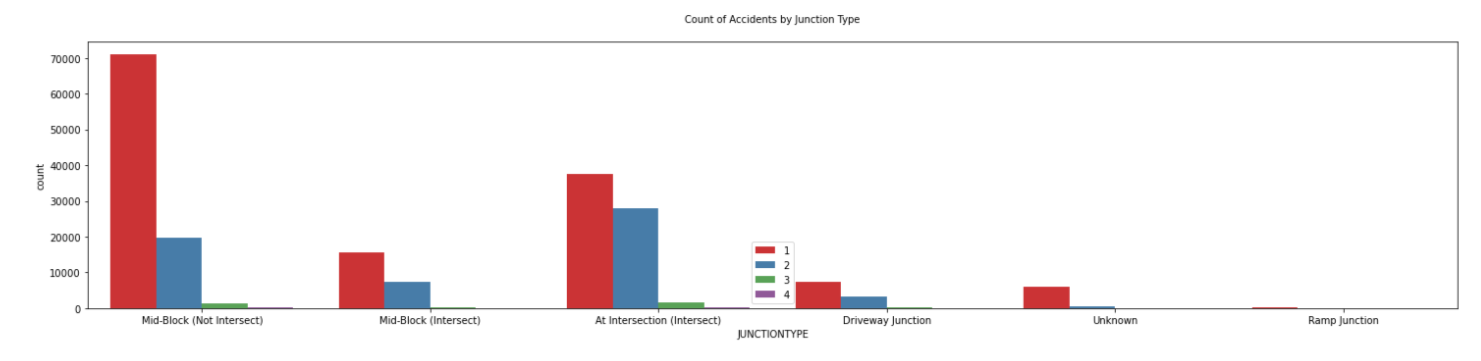
## 3.5 Relationship between Address type and accident count.



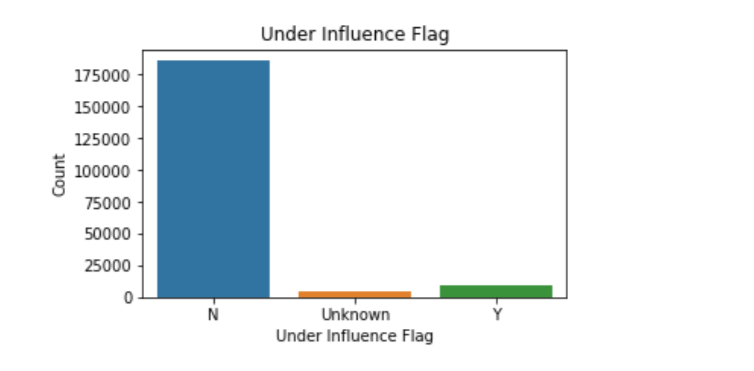
## 3.6 Relationship between collision type and accident count.



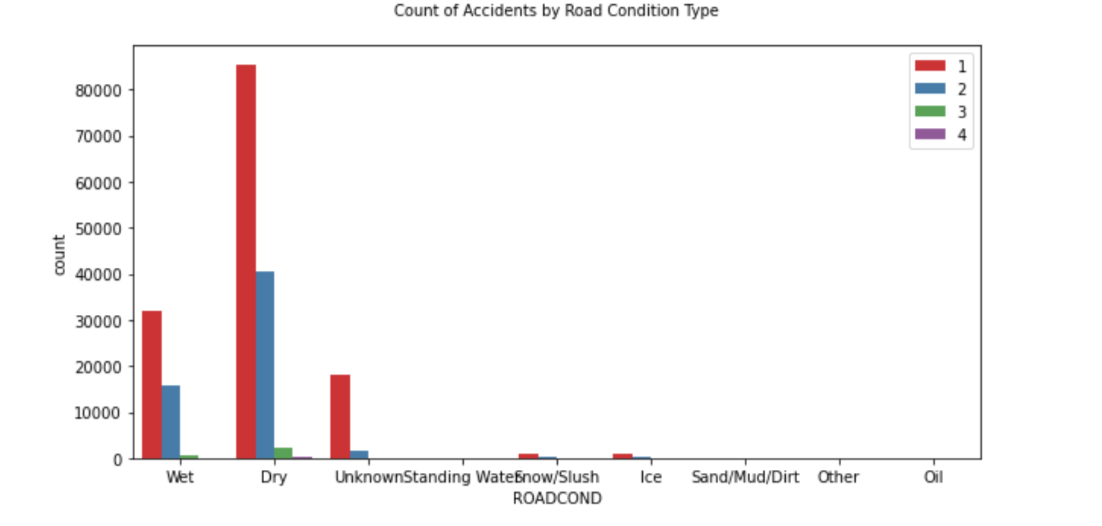
## 3.7 Relationship between junction type and accident count.



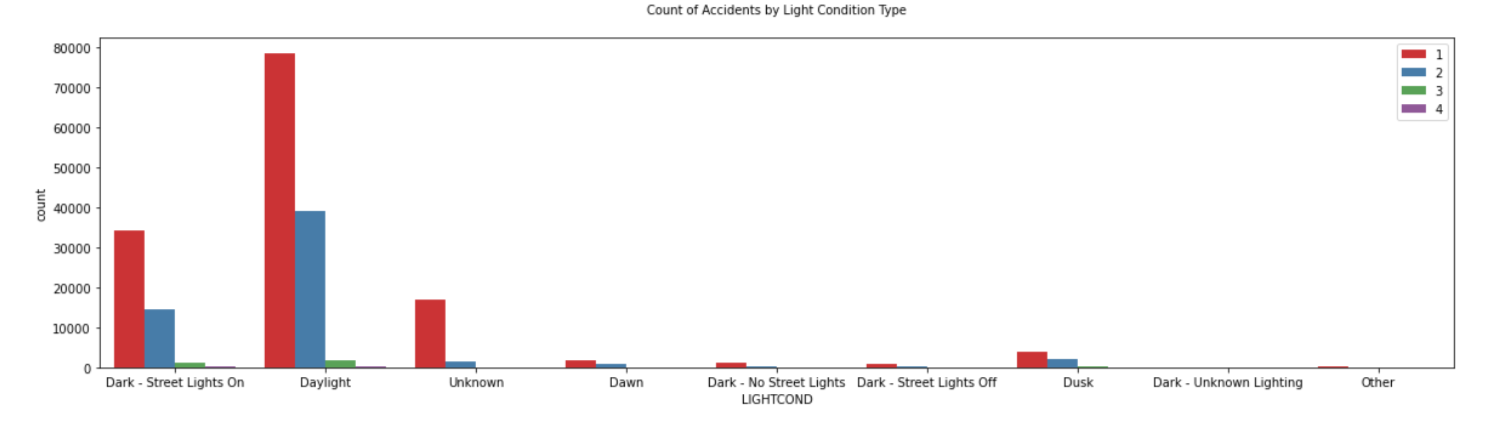
## 3.8 Relationship between accident under influence and accident count.



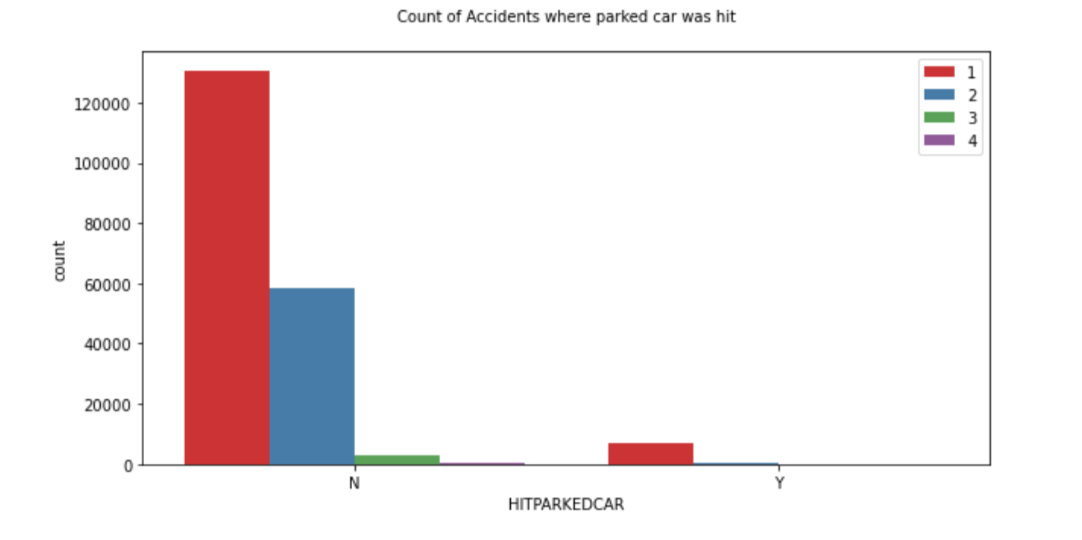
## 3.8 Relationship between road conditions and accident count.



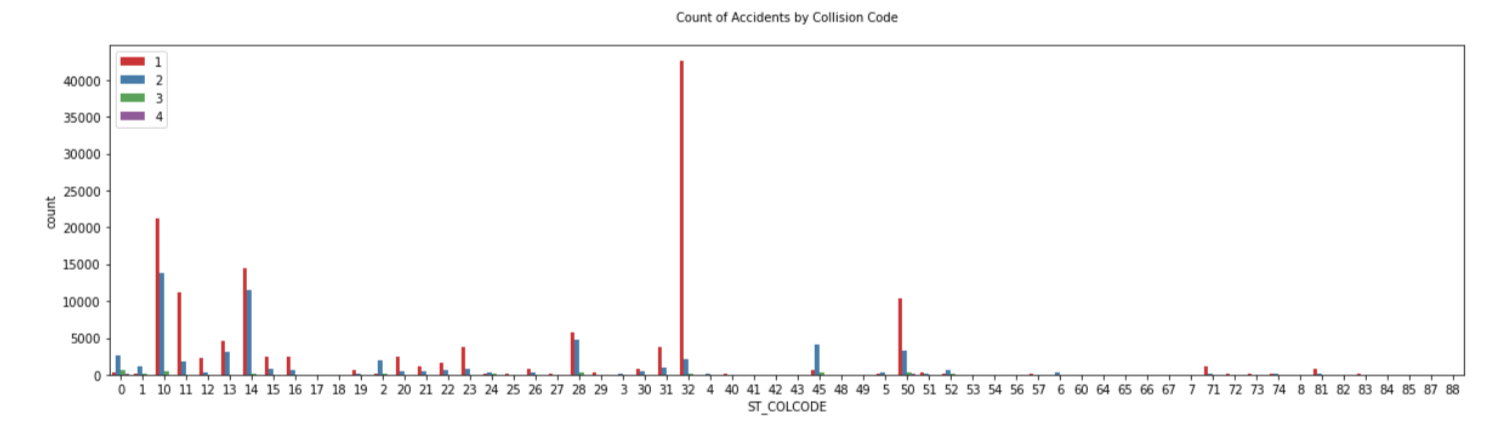
## 3.9 Relationship between light conditions and accident count.



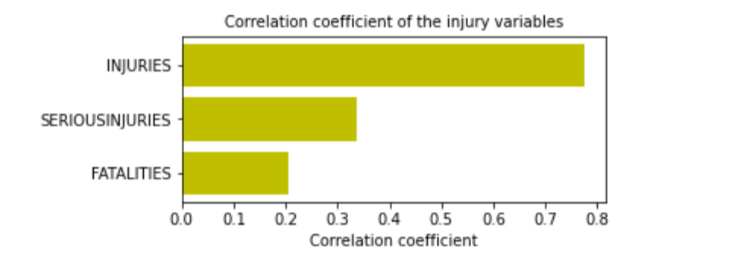
## 3.10 Relationship between hit parked car and accident count.



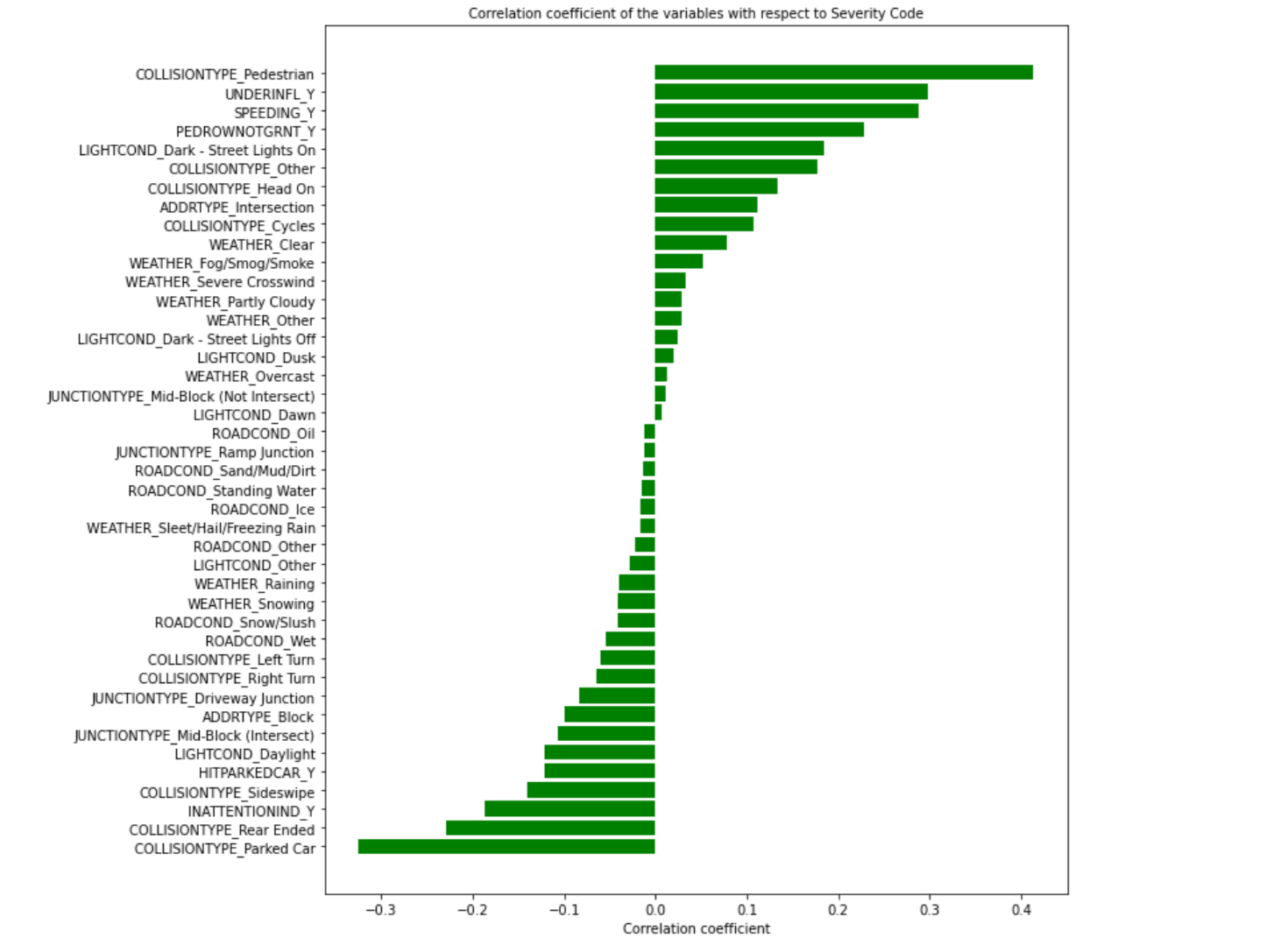
## 3.11 Relationship between hit parked car and accident count.



## 3.12 Correlation between Injuries, serious injuries and fatalities.

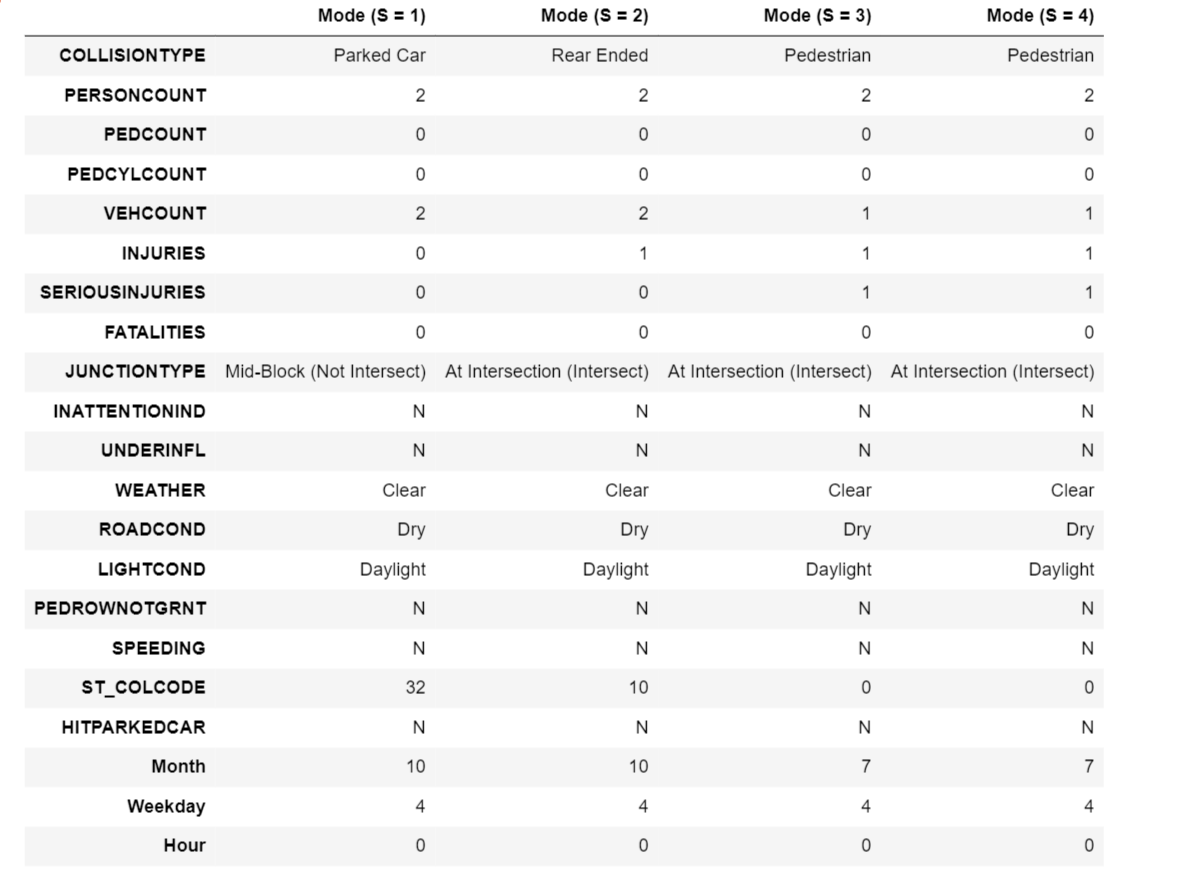


## 3.13 Correlation coefficient between all the variables and fatalities.



## 3.14 Finding out the mode with respect to each feature

It will be interesting to see the mode (highest frequency) values of each feature with respect to the severity codes.



# 4. Predictive Modeling

## 4.1 Train Test Split

* Training set—a subset to train a model.
* Test set—a subset to test the trained model.

Split the data into X\_train, X\_test, y\_train, and y\_test, after standardizing the input features.

## 4.2 Comparison of Different Classification Models

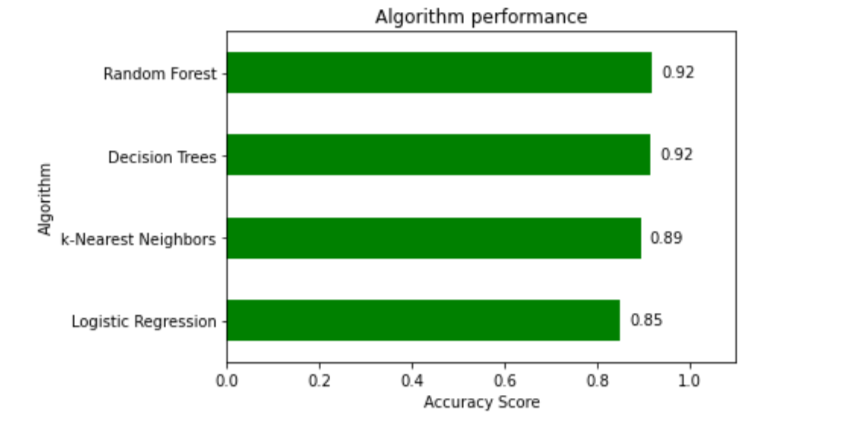
Accuracy score below 4 models are calculated in the Jupyter notebook. As per the calculations Random Forest and Decision Tree have similar accuracy.

#### **1. Logistic Regression**

#### **2.KNN**

#### **3.Decision Tree**

#### **4.Random Forest**



# 5. Conclusions

# In this study, I analyzed the relationship between Severity code and different attributes in the collision data. I identified light conditions, road conditions, and junction types as the most important causes of a car collision. I built both regression models and classification models to predict how car collisions can be avoided. These models can be beneficial to SDOT, police, and even regular commuters in several ways. For example, it could help SDOT estimate the improvement area, like installing more traffic signals on particular junction types or imposing speed limits in more collision causing areas, giving more attentions to pedestrians and parked cars. The model and graph can also help commuters be more alert while choosing roads, driving in different light conditions, and avoiding more collision causing junction types.

# 6. Future directions

We can conclude that the Random Forest is the best model in this scenario (Decision Tree is the second best, but very close to Random Forest). An interesting point to note here is that the top important features are somewhat different between Random Forest and the Decision Tree models.

Following the Random Forest model, we see that special attention needs to be given to pedestrians, speeding, collision with a parked car, rear-ended collision, under influence. The collision codes 50 (Struck Fixed Object), 32 (One Parked - One Moving), 10 (Entering at Angle) are the most frequent ones.

**Future Scope**

Models in this study mainly focused on individual features. The relations between the key features and accident severity can be further studied in details. Different data balancing techniques can be applied and evaluated. Development of a much more complex real-time accident risk prediction model. These interactions data are obviously more difficult to extract and quantify, but if optimized, could bring significant improvements to the models.

**GitHub link -**https://github.com/SHIVANIGODE/Coursera\_Capstone/blob/master/Assignment\_Accident\_Severity.ipynb