**Predicting the severity of a car accident**

## 1.Introduction:

### 1.1.Background:

Car accidents are quite possible on the road.The consequences of road accidents might be severe like injuries for involved persons. It also causes terrible traffic jams which delays others' routines.The roads are shut down in case of severe crashes. It is important to predict the road accidents to avoid such situations.

### 1.2.Problem:

Given the road and weather conditions, we intend to warn the chances of car accidents and how severe it could be, so that the driver can drive carefully. It will also help the driver to change the travel plans if able to. Based on the car accident severity data provided, I intend to identify the chances of getting into a road accident.

### 1.3.Interest:

Obviously a car driver wants to avoid the road accidents, which has severe impacts like injuries or issues in his/her insurance. Insurance companies might also be interested in predicting the chances to work on their reserves for a specified period or for any other reasons.

## 2. Data acquisition and Cleaning

### 2.1. Data Sources:

The collision history of seattle was downloaded from Seattle.gov site and has been used for further studies. This dataset has the details of car accident history of Seattle

### 2.2. Data Cleaning:

The dataset almost had 0.22M records. I decided to consider only the records that have reported at least 5 accidents at the same location, which reduced the size of my data to 0.177M.

The dataset had some redundant columns which I decided to remove.

| Kept columns | Dropped columns | Reason for dropping |
|---|---|---|
| INCDTTM | INCDATE | Incident date is available as part of Incident time column |
| EXCEPTRNCODE | EXCEPTRSNDESC | This feature is intended to show whether the record has enough information to explain the accident, the value -NEI says it does not have enough information. Though these NEI records |

| | | |
|---|---|---|
| | | state that these do not have enough information in one or the other fields, some of the records still have the information like how the road condition/weather was, at the time of accident. So I decided to keep these records too, planning to replace the insufficient information with mean or frequency. |
| INJURIES, SERIOUSINJURIES, FATALITIES | | Not required for this study, hence dropped |

There were 19 features which had missing values. I decided to replace these missing values by the mean or frequency.

**2.3 Feature Selection:**

I was interested in checking the features like Weather, Location, Junction Type, Road condition,Light Condition,person count,vehicle count and Speeding have any correlation with the target severity level to identify the severity of the car accident.

- I dropped the records for which Road condition, Light condition and Weather information are not available, as these features will not be helpful to predict the accident severity if these information are not available. Replacing by its frequency will not give the accurate information.
- The features X,Y are the coordinates of the location, so the missing values were droppedas they may not be required for analysis.
- The feature AddressType was a block or intersection which was replaced by its frequency.
- The feature intkey was replaced by its mean.
- EXCEPTRSNCODE shows whether the incident has enough information or not. I changed the existing records to have either 0(enough information) or 1(not enough information) values.Missing values are assumed to have enough information and changed to be 0.
- The feature COLLISIONTYPE took valid values like how the collision happened, rear ended, right turn, other, etc. For the missing values, I used ‘Other’ as the option as we do not know how the collision happened and other seemed most appropriate.
- JUNCTIONTYPE explained the type of junction like midblock, intersection , ramp etc. The missing values of Junctiontype were changed to Unknown type.
- The Nan values of ‘in attention indicator’ (INATTENTIONIND) were replaced by 0 and the existing yes values were replaced by 1.

- UNDERINFL indicator takes the values 0,1,Y and N along with missing values. The missing values were replaced by 0.
- Missing values of the feature PEDROWNOTGRNT were replaced by ‘N’.
- Missing values of SDOTCOLNUM were replaced by its average value.
- The Speeding indicator was set to Y if speeding is true. The missing values were set to N.
- Collision code (ST_COLCODE) missing values were replaced by 0.
- Collision description (ST_COLDESC) missing values were replaced by a general value “other”.
- The target variable SeverityCode was ordered to have values from 0 to 3 only.
- I changed the datatype of the features EXCEPTRSNCODE, INATTENTIONIND and ST_COLCODE to integer. My dataset looked much better now.

**3. Methodology:**

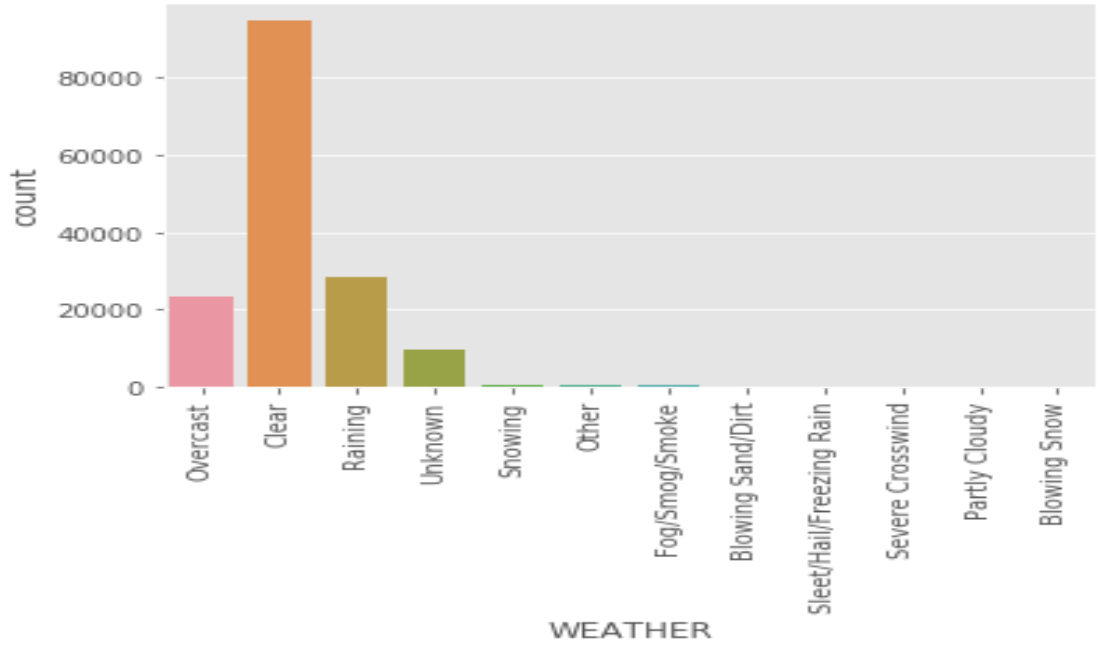
**3.1 Exploratory Data Analysis**

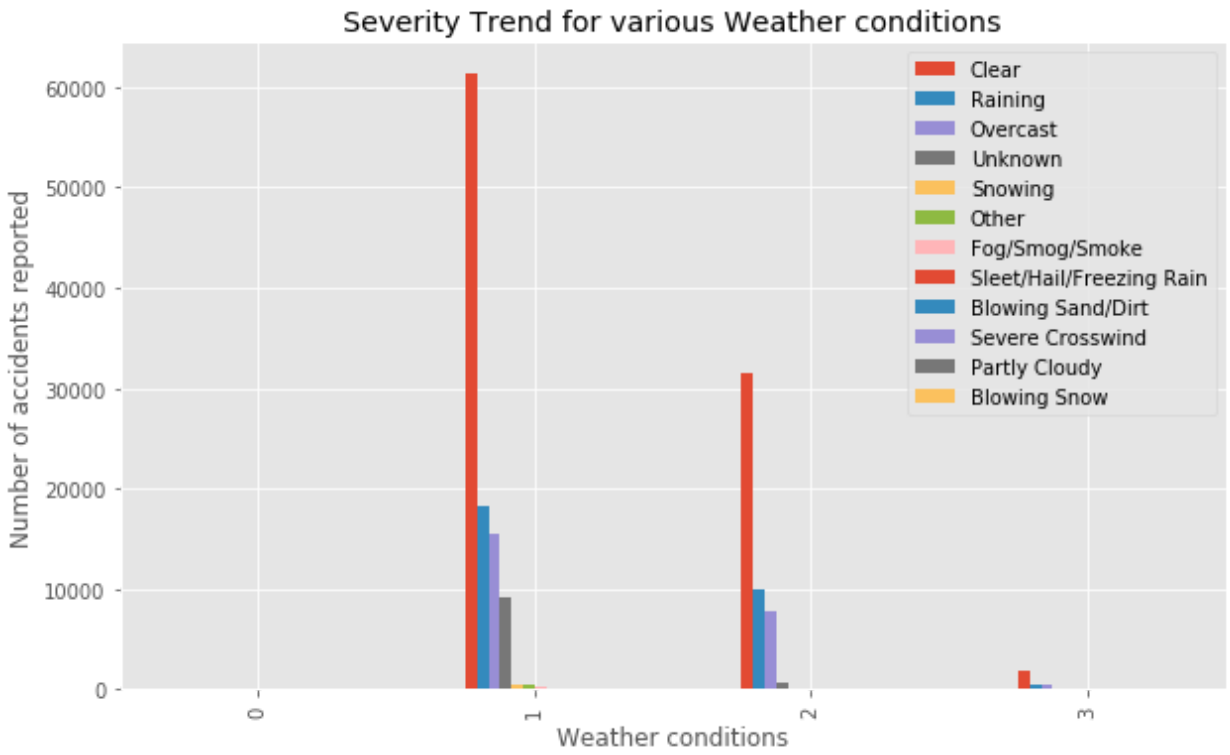
**3.1.1 Calculation of target variable**

Severity Code was in the dataset and assumed to be the target variable. Severity code values ranged from 0 to 3 based on the level of severity(unknown,prop damage, injury, serious injury and fatality) It had a non numeric value for serious injury which was adjusted to have numeric value.

**3.1.2 Relationship between Weather and Severity Code**

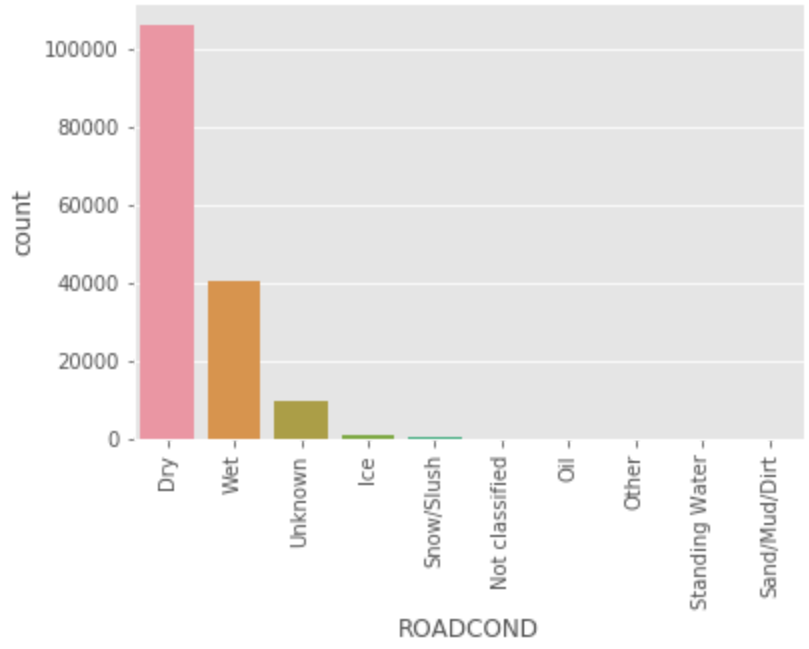
When the weather is clear,most of the accidents have been reported. Raining and overcast have also been contributing main factors of accidents.

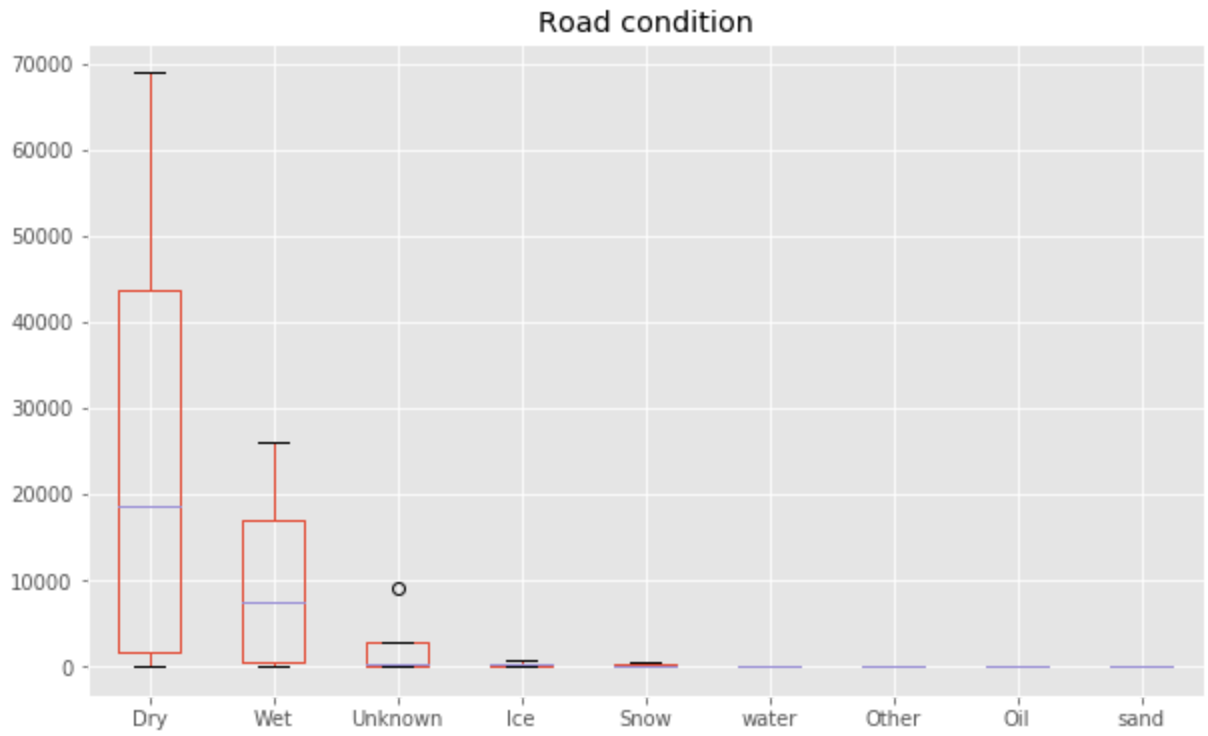


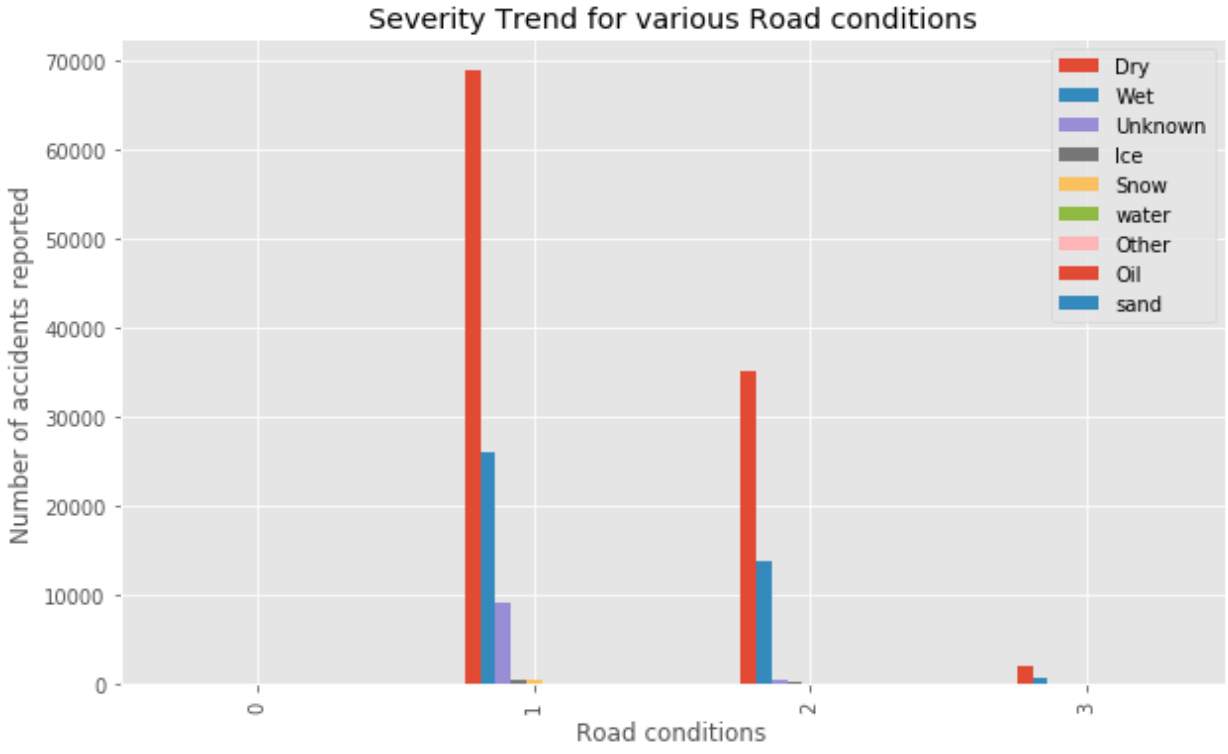



### 3.1.3 Relationship between Road condition and Severity Code

Most of the accidents were observed in Dry road conditions with considerable severity. Wet conditions also observed at the time of accidents.

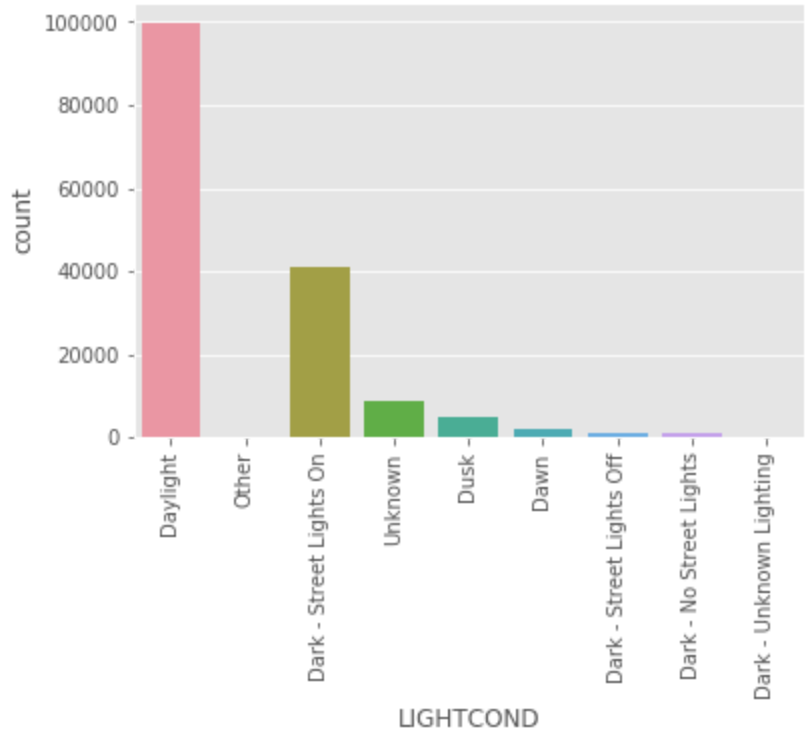



Road condition
70000
60000
50000
40000
30000
20000
10000
0
Dry
Wet
Unknown
Ice
Snow
water
Other
Oil
sand


Severity Trend for various Road conditions
Number of accidents reported
70000
60000
50000
40000
30000
20000
10000
0
Dry
Wet
Unknown
Ice
Snow
water
Other
Oil
sand
0
1
2
3
Road conditions

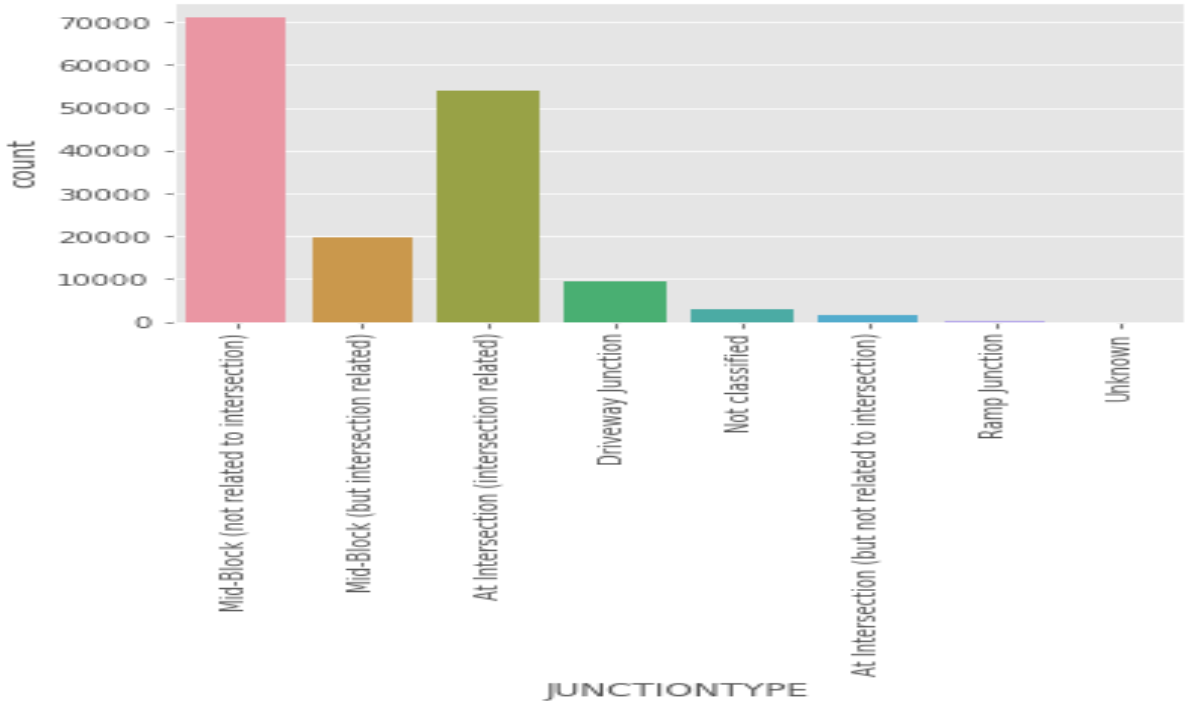
### 3.1.4 Relationship between Light Condition and Severity Code

Majority of the accidents have happened during daylight time.



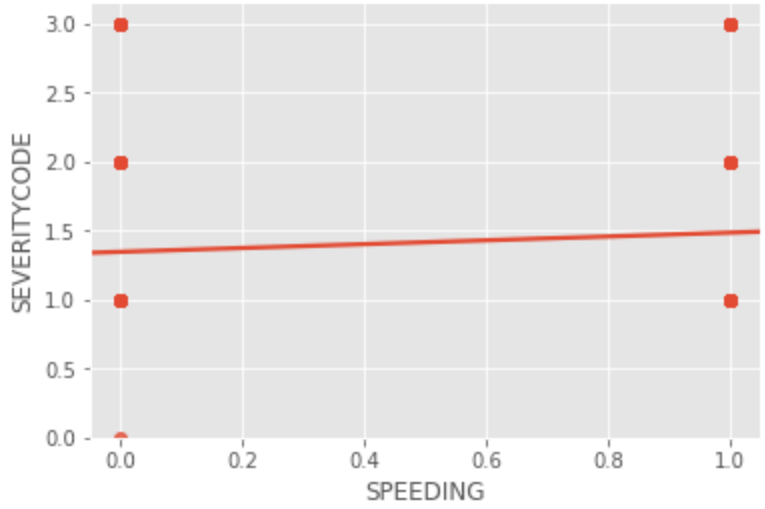

### 3.1.5 Relationship between Junction type and Severity Code

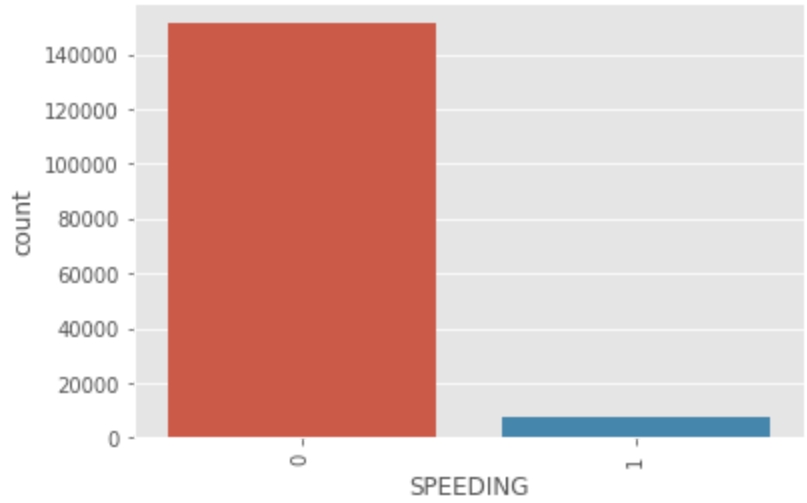
The highest number of accidents are either related to mid-block or intersection related.

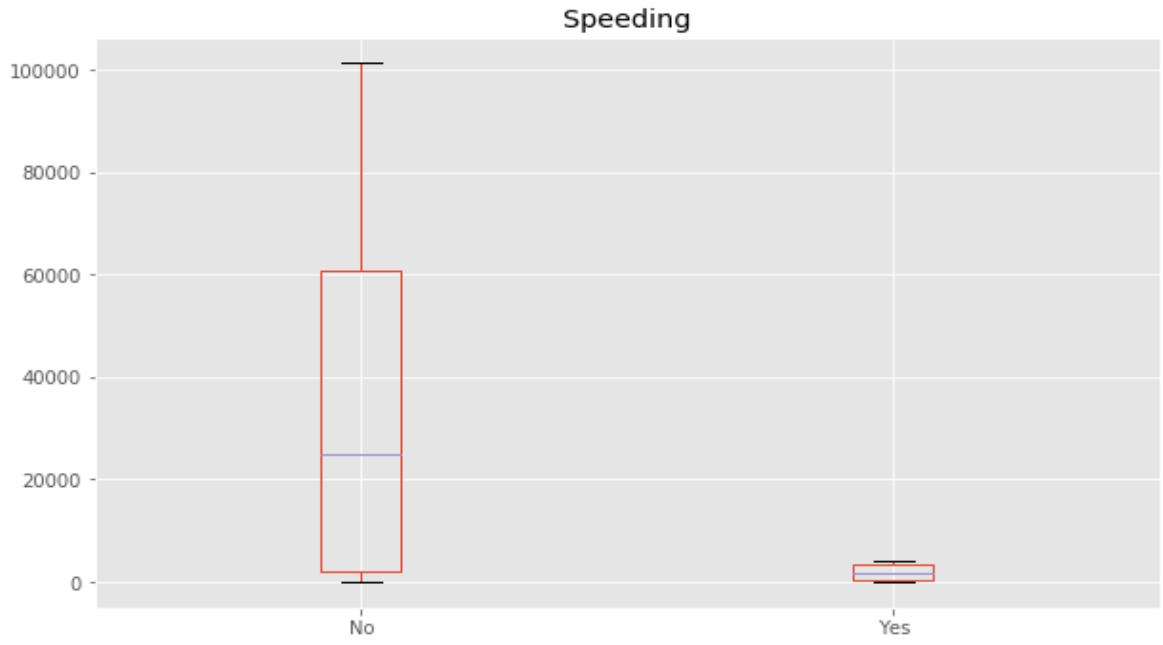


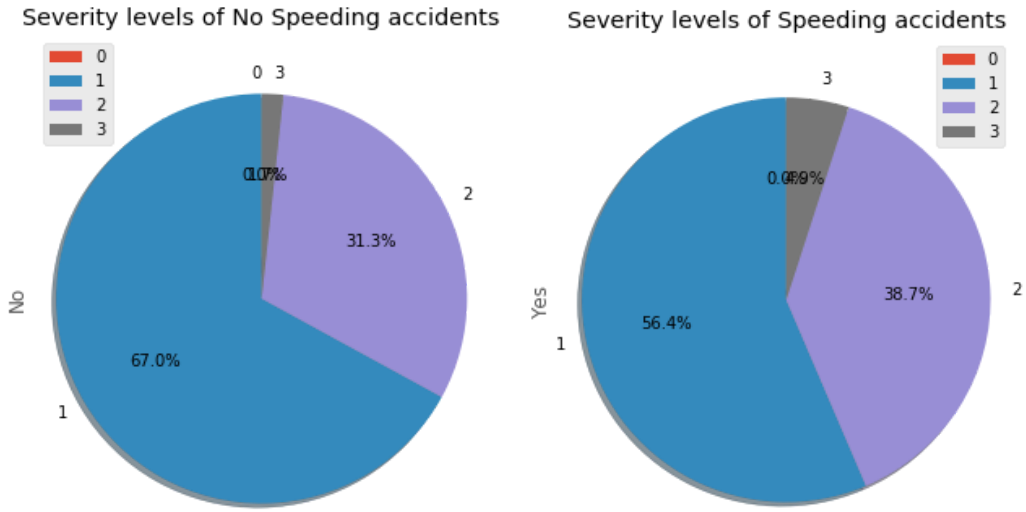
### 3.1.6 Relationship between Speeding and Severity Code

Speeding and severity code has positive correlation as anticipated. The data shows if the speeding increases, a slight increase in the severity of the accident as well. That is , if the reported accident has an observation of speeding, 4.9% of fatality accidents (sev 3) have been observed, whereas fatality accidents are just 0.7% if speeding is not reported.









## 4.Results

### 4.1.Machine Learning model:

There are two types of models available, regression and classification. Regression provides additional information for the observation if required, whereas classification provides the chances/probabilities of the severity of the car accident in this case. So I decided to use a classification model.

#### 4.1.1 Classification models

I used logistic regression to predict the severity of the car accidents, which varied from 0 to 3(highest level).The features like weather, road condition, light condition, person count, x and y coordinates, junction type and speeding were considered to predict the severity code. 20% of data was used for the test and data was classified.

The evaluation metrics score for Logistic regression is as follows:

| Measure | Value |
|---|---|
| Jaccard Similarity score | 0.688347 |
| F1 score | 0.63 |
| Log loss | 0.61 |

## 5.Discussion

In this study, I analysed the relationship between the car accident severity and their accidents observed data like weather condition, road condition, junction type etc. I identified these features are very important in deciding the car accident severity. I also built a classification model, Logistic Regression, to predict the car accident severity. This will be helpful in predicting the severity of car accidents.

## 6. Conclusions and Future directions:

I was able to achieve ~63% accuracy in predicting the car accident severity. I can still try to see whether any other classification models can be applied and how it performs the severity prediction.