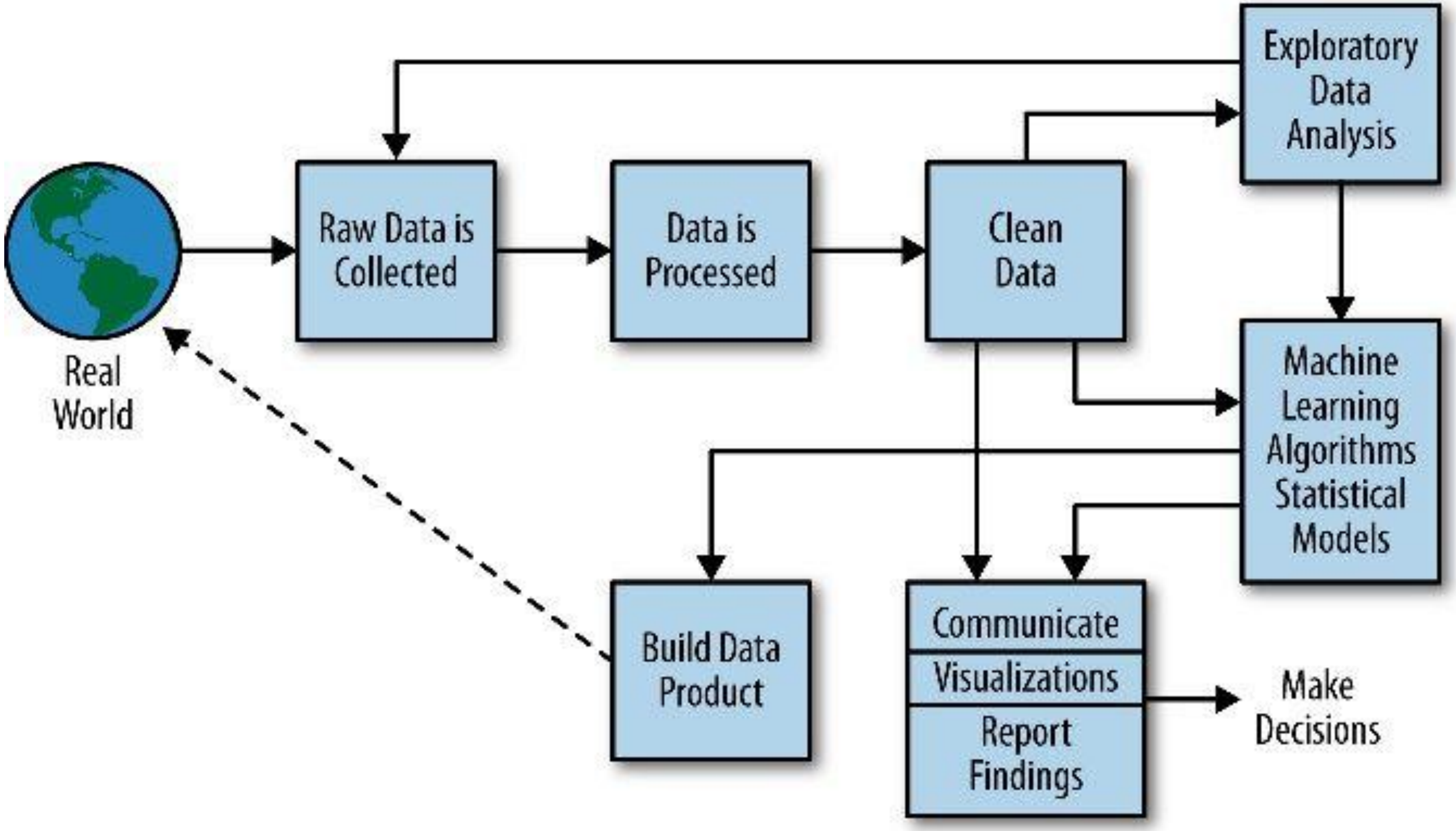
IBM DATA SCIENCE CAPSTONE PROJECT

DATA ANALYSIS OF CAR CRASHES OCCURRED IN SEATTLE BETWEEN THE YEARS 2004 AND 2020

By,
Dhaval Chaurasia

Data Science Methodology



Business Understanding

- The objective of this project is to develop a machine learning model to predict the severity of an accident depending on various circumstances.
- Circumstances such as weather condition, road conditions, light conditions etc.
- The target audience will be the government, the emergency response services, the police and the driver.

Data Collection

- The data was collected from Government of Seattle website.
- The data came with a PDF file which contains the description about the attributes.
- There are 37 attributes and 194673 rows of data.
- The data set had information about the car crashes that occurred between 2004 and 2020.

Data Processing

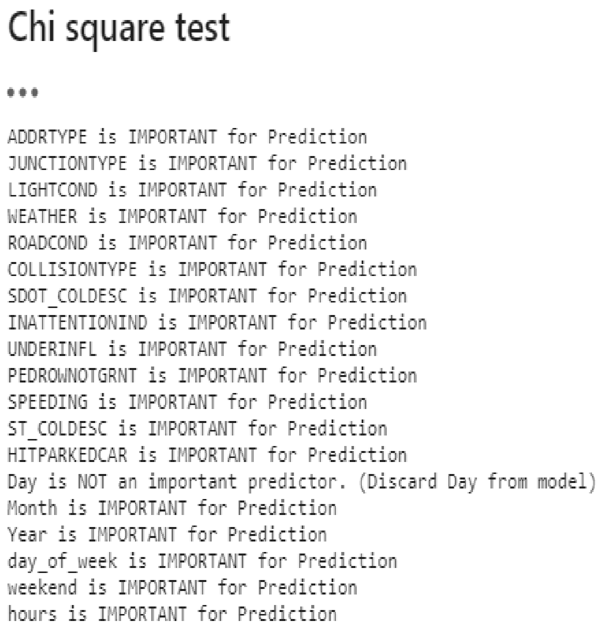
- Dropping duplicate columns such as "SEVERITYCODE.1“.
- Removing redundant data such as “INCKEY”, “OBJECTID” etc.
- The columns “EXCEPTRSNCODE” and “EXCEPTRSNDESC” were entirely empty therefore it was deleted.
- All the null values were entirely dropped.
- Target variable “SEVERITYCODE” had data in 1 and 2 formats. It was converted to 0 and 1 format.
- Data imbalance was corrected.
- Data normalization was done.

Attribute Selection

- Target variable is SEVERITYCODE

0	Property Damage
1	Injury

- Attributes for modelling

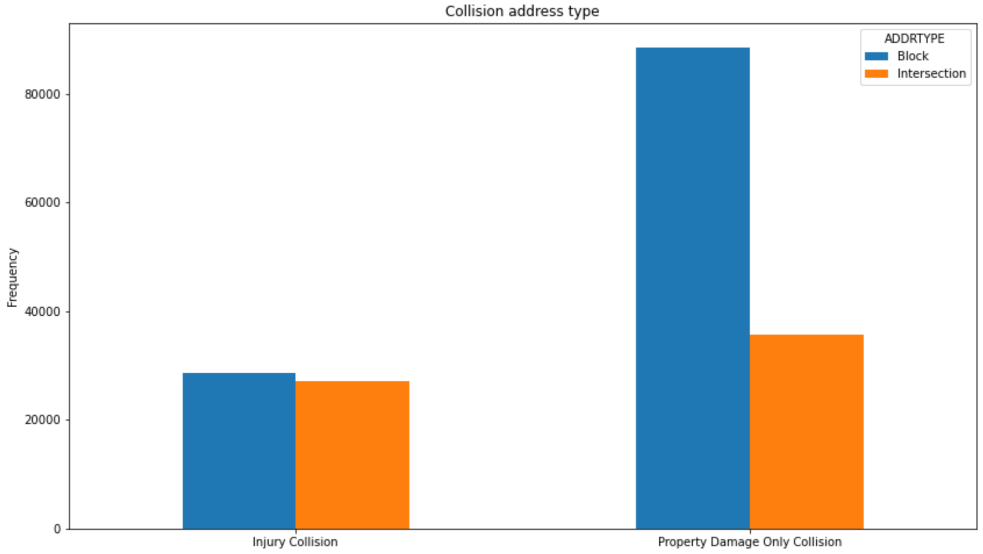


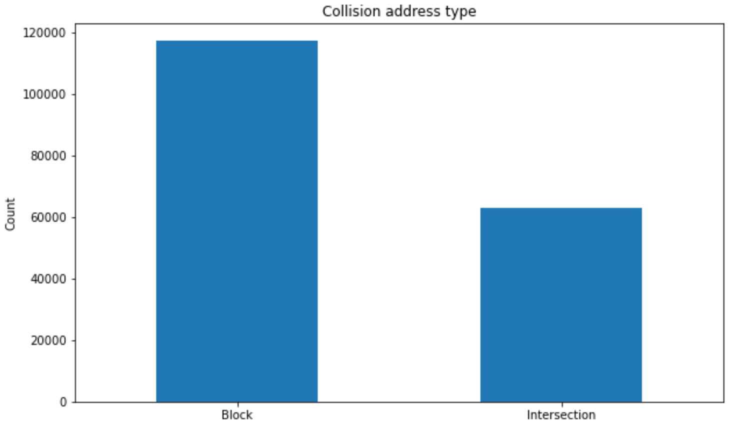
Chi square test

```
...
ADDRTYPE is IMPORTANT for Prediction
JUNCTIONTYPE is IMPORTANT for Prediction
LIGHTCOND is IMPORTANT for Prediction
WEATHER is IMPORTANT for Prediction
ROADCOND is IMPORTANT for Prediction
COLLISIONTYPE is IMPORTANT for Prediction
SDOT_COLDESC is IMPORTANT for Prediction
INATTENTIONIND is IMPORTANT for Prediction
UNDERINFL is IMPORTANT for Prediction
PEDROWNOTGRNT is IMPORTANT for Prediction
SPEEDING is IMPORTANT for Prediction
ST_COLDESC is IMPORTANT for Prediction
HITPARKEDCAR is IMPORTANT for Prediction
Day is NOT an important predictor. (Discard Day from model)
Month is IMPORTANT for Prediction
Year is IMPORTANT for Prediction
day_of_week is IMPORTANT for Prediction
weekend is IMPORTANT for Prediction
hours is IMPORTANT for Prediction
```

Exploratory Data Analysis

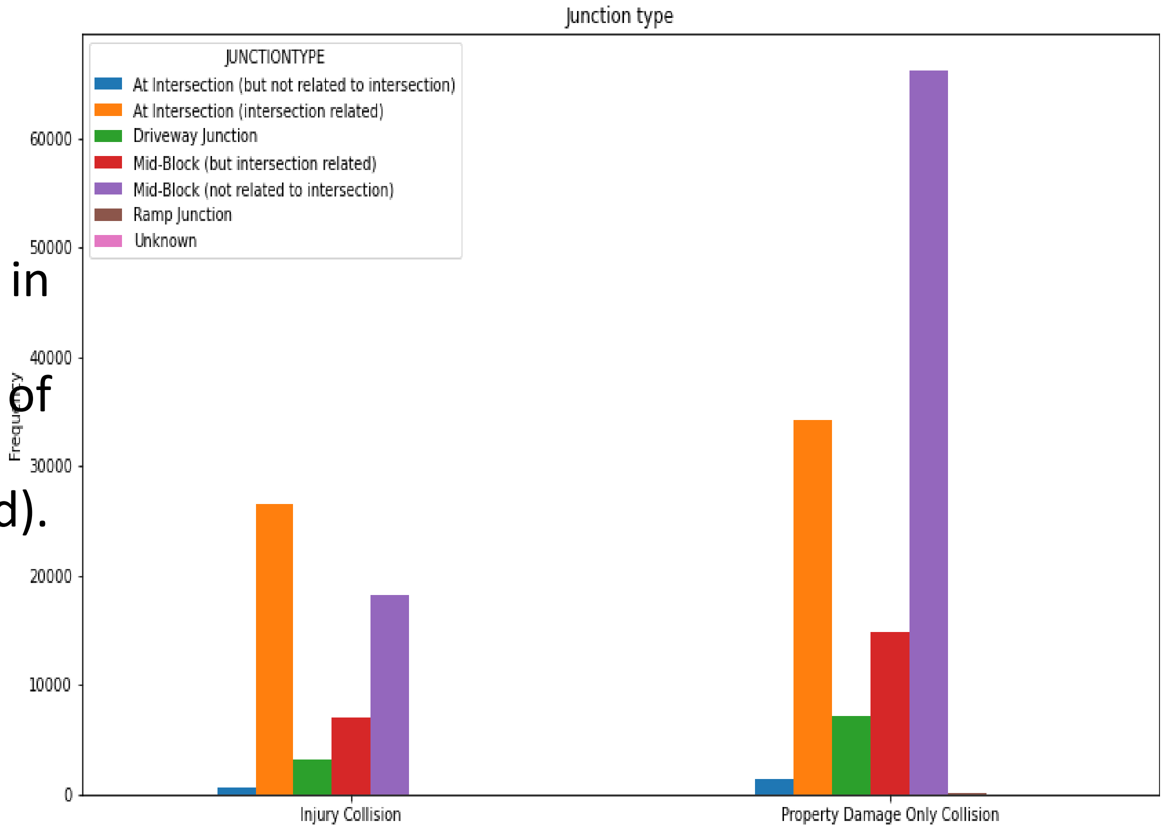
- Relationship between collision address type and severity of an accident.
- majority of accidents takes place around the blocks rather than at intersection. Close to 120K accidents happened at or around the block compared to about 60k at intersection.





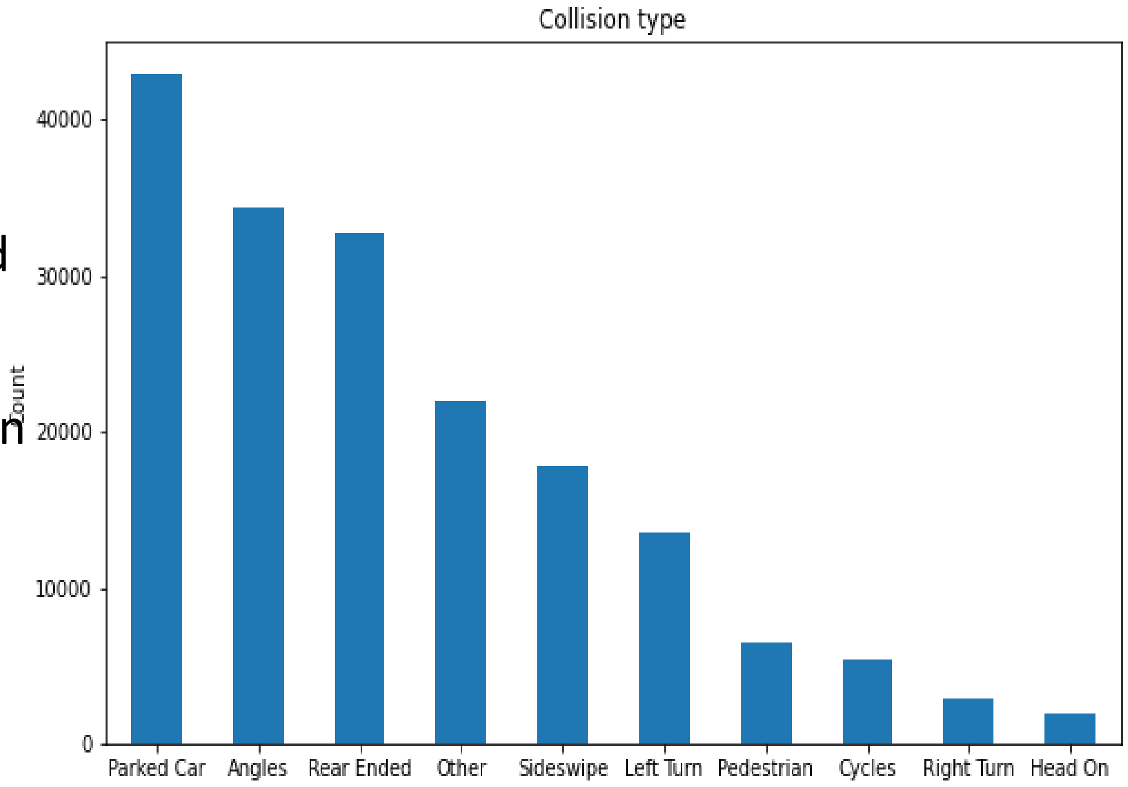
Exploratory Data Analysis

- Relationship between junction type and severity of an accident
- Majority of accident (>60K) which resulted in property damage, happened at Mid-Block (not related to intersection) whereas most of the accidents (>25K) that caused injuries happen at intersection (intersection related).



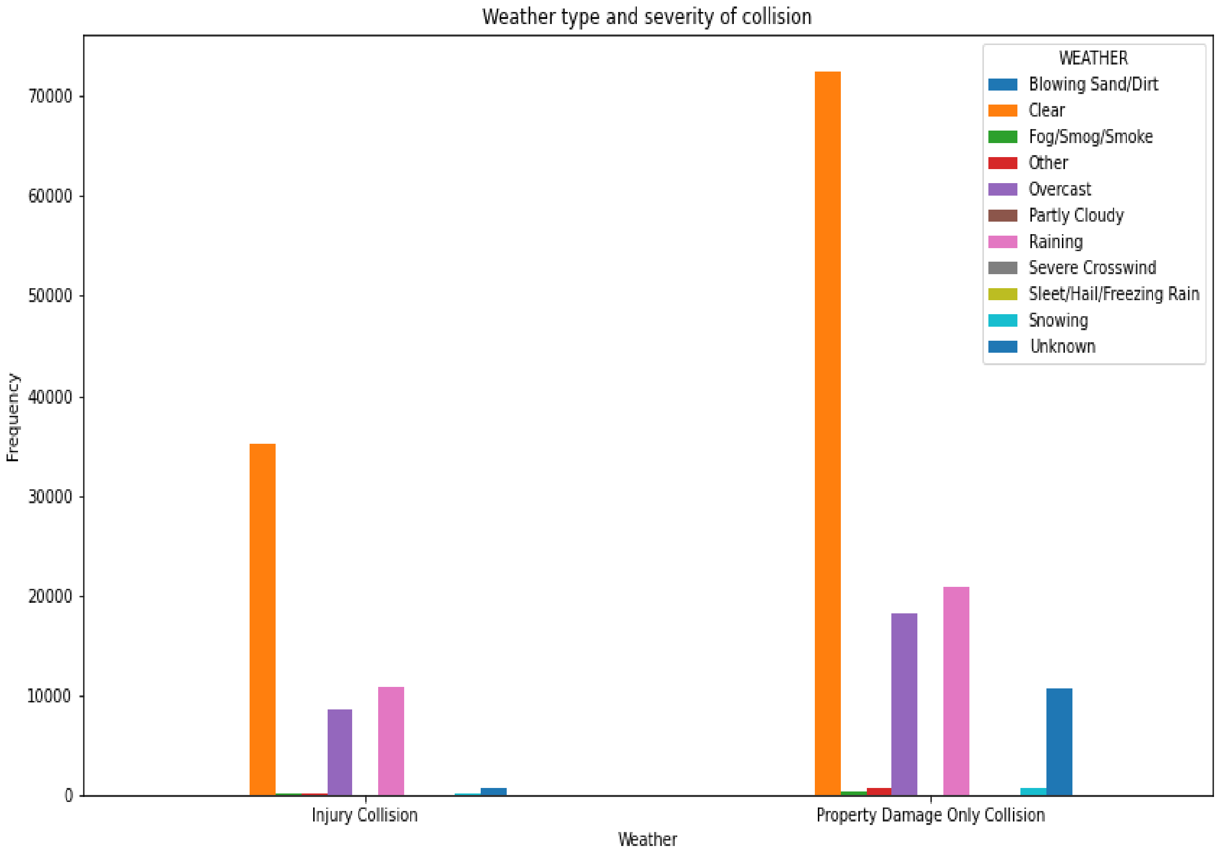
Exploratory Data Analysis

- Relationship between collision type and severity of an accident
- In Figure 4 shows 40K accidents that involved a parked car, resulted in property damage while less than 5K resulted in injury. Rear ended and sideswipe caused more injury than other type of collision.



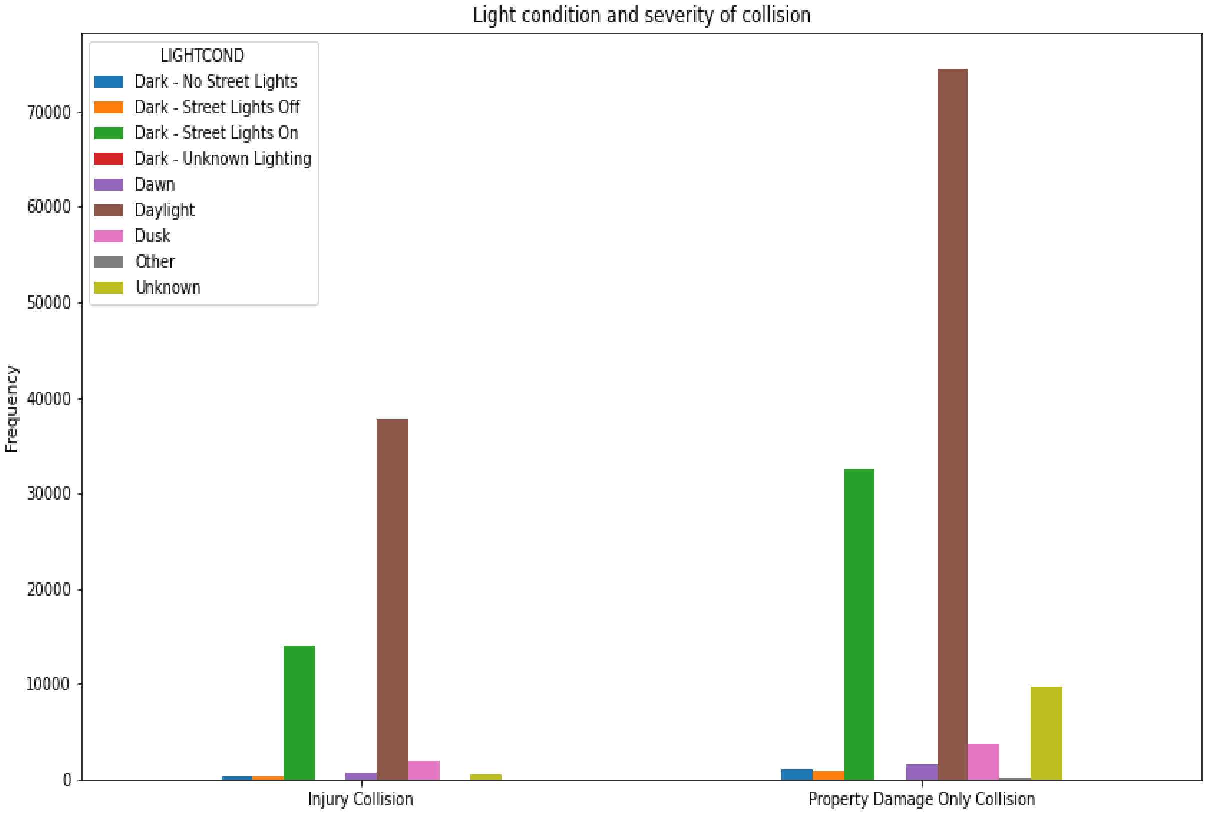
Exploratory Data Analysis

- Relationship between weather type and severity of an accident



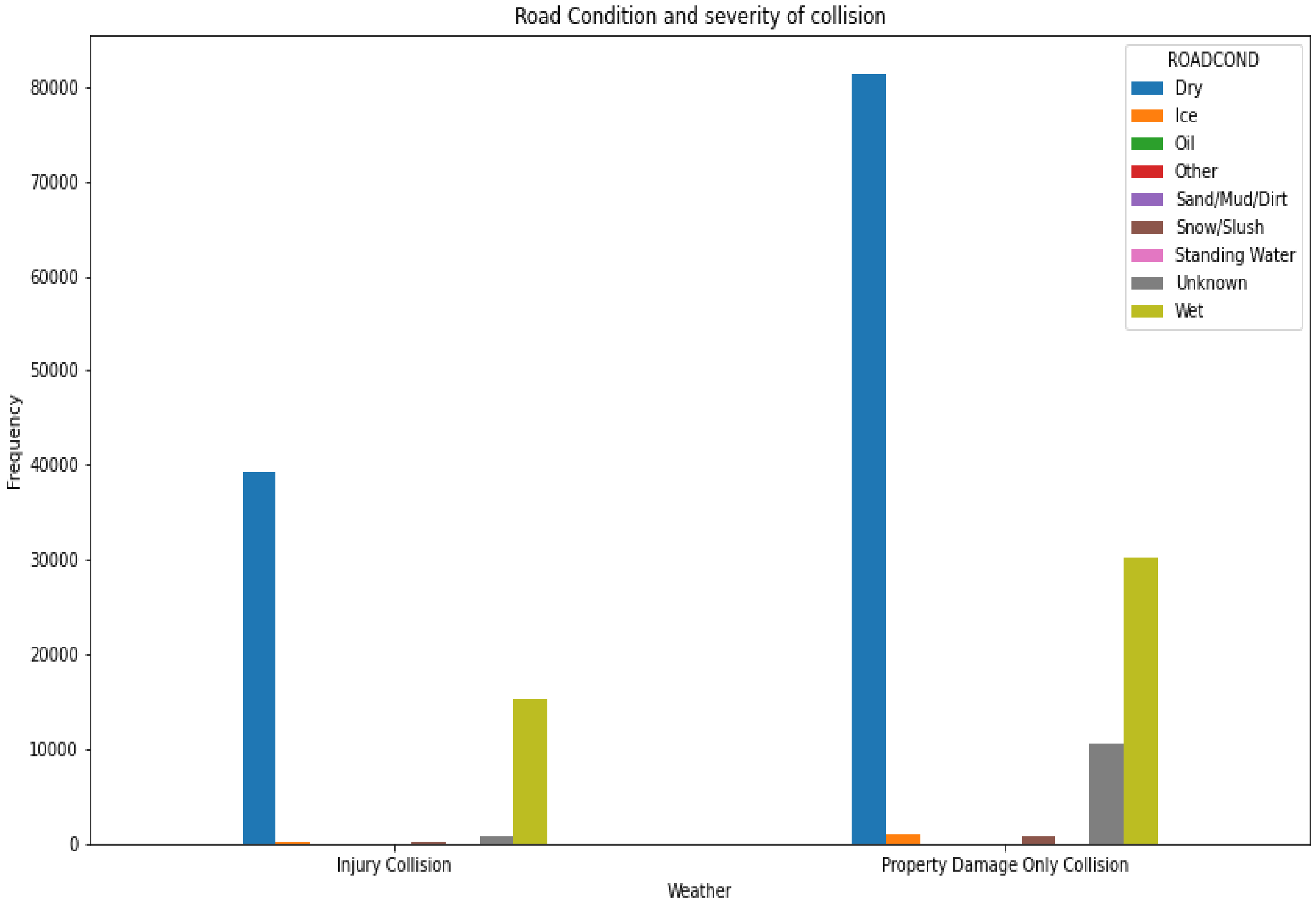
Exploratory Data Analysis

- Relationship between light condition and severity of an accident
- Majority of (~110k) accident happened in a day light conditions. Close to 50k accident happened in Dark (streetlight on) conditions of which around 15k resulted in injury and 35k resulted in property damage.



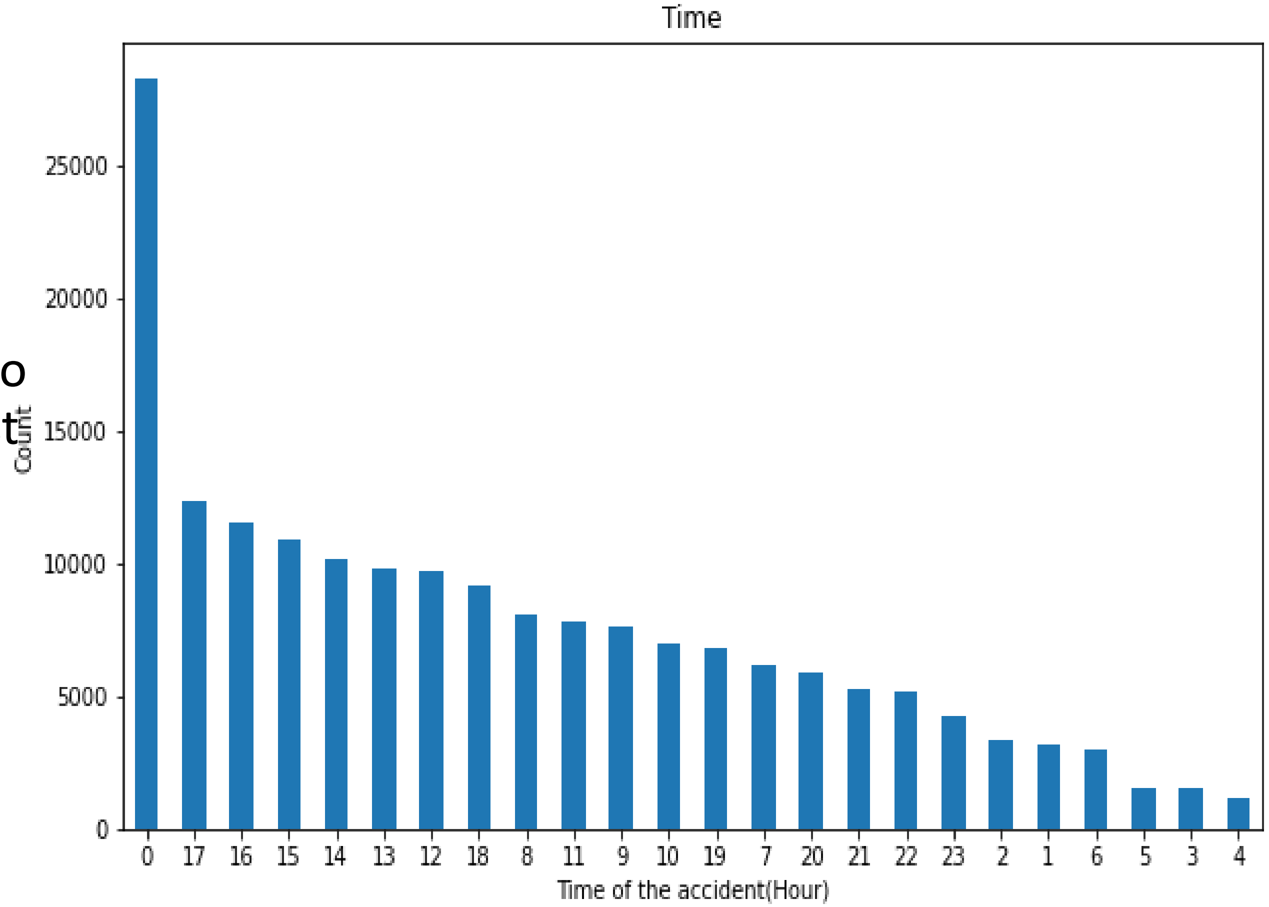
Exploratory Data Analysis

- Relationship between road condition and severity of an accident



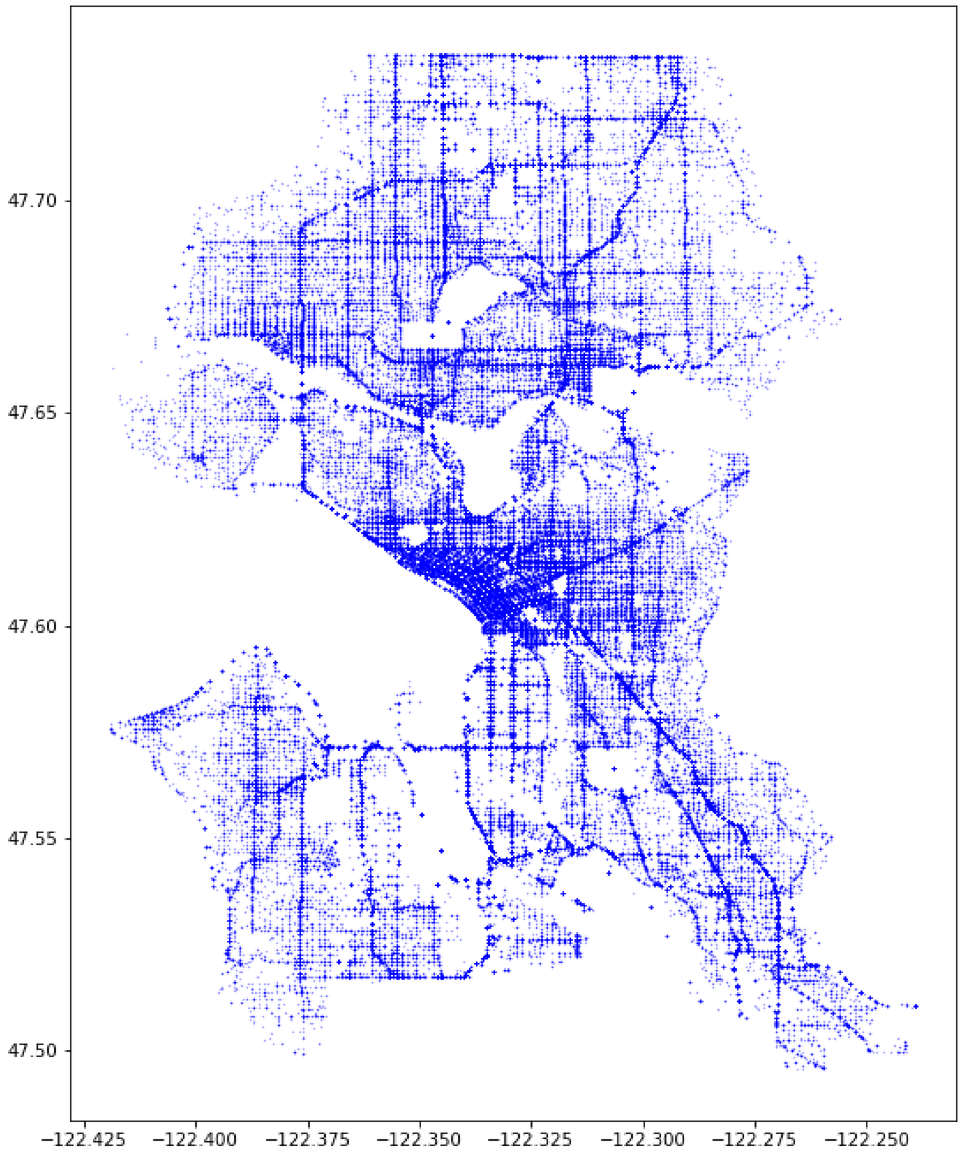
Exploratory Data Analysis

- Relationship between time and accident count.
- Figure 8 shows us that most of the accident happened at mid night (>25k). This was followed by accident occurred during 14hrs to 17hrs. This is where office hours end for most of the people.



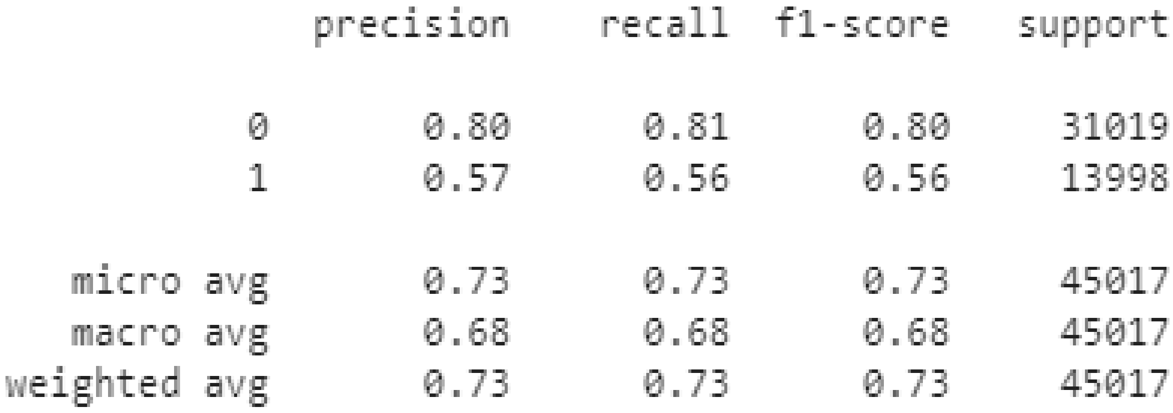
Exploratory Data Analysis

- Scatter plot of X and Y coordinates of various accidents
- The dense blue area in the centre is downtown Seattle.



Machine learning Model

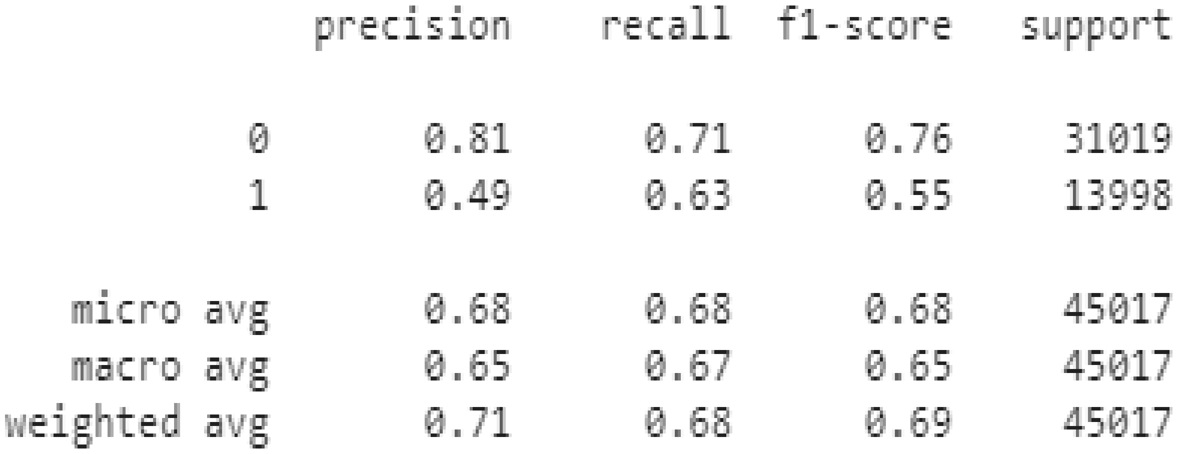
- Decision tree model achieved an accuracy score and Jaccard similarity score of 0.73. Recall, Precision and f1 score is shown below.



```
              precision    recall  f1-score   support

           0       0.80      0.81      0.80     31019
           1       0.57      0.56      0.56     13998

   micro avg       0.73      0.73      0.73     45017
   macro avg       0.68      0.68      0.68     45017
weighted avg       0.73      0.73      0.73     45017
```

Machine learning Model

- Logistic regression model achieved a Jaccard score of 0.68 and log loss of 0.58. Recall, Precision and f1 score is shown below.

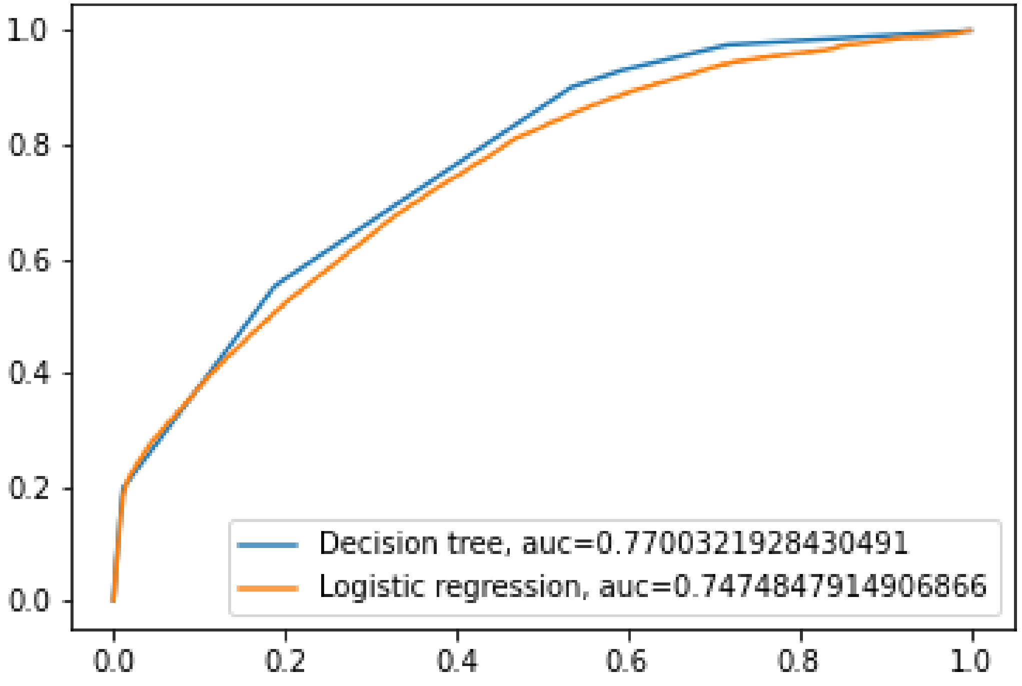


```
              precision    recall  f1-score   support

           0       0.81      0.71      0.76     31019
           1       0.49      0.63      0.55     13998

   micro avg       0.68      0.68      0.68     45017
   macro avg       0.65      0.67      0.65     45017
weighted avg       0.71      0.68      0.69     45017
```

ROC (Receiver Operating Characteristic Curve).

- It is the plot between the True Positive Rate (y-axis) and False Positive Rate (x-axis). It is a performance measurement for classification problem at various threshold value.

- AUC (Area under curve) represents the degree of separation. An excellent model has an AUC close to 1 which means it has good measure of separability.



Conclusion and future directions

- Model was built and desired accuracy was achieved
- The accuracy can be significantly improved by adding few more attributes such as gender of driver, age of driver, engine capacity, type of vehicle, number of casualties etc.
- With the help of this model, improving road infrastructure and awareness programs about road safety and rules, Government can reduce the number of cases that result in loss of life and property.

Presentation End

Thank You!