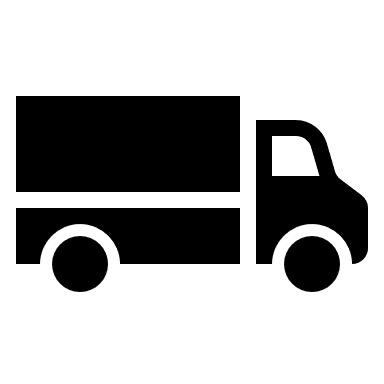
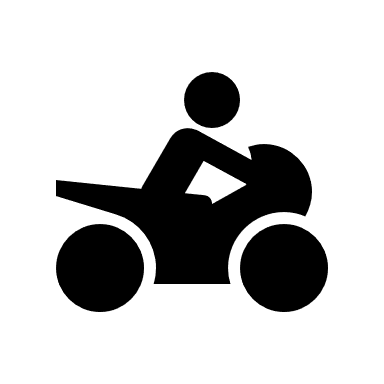
## 

**** ****

**SEATTLE TRAFFICK DATA ANALYSIS AND RISK PREDICTION**

**IBM Final Capstone Project**

**Puja Misra**

**Introduction**

This project is our Final submission to IBM Data Science Professional Certificate course on Coursera. The goal of the project is to detail and use Data Science tool set for Predictive analysis.

We will be working on a real-life problem and demonstrate how Machine Learning can help us predict and process the value by applying the learned skills.

**Business Understanding**

* **Background:**

According to 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours. Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people. This number has stayed relatively steady for the past decade. According to 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours. Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people. This number has stayed relatively steady for the past decade.

* **Problem Statement:**

As we see in the background statement above, the numbers of fatal crashes are having an upward trend or has been steady for past decade for Seattle as per WSDOT .

A car parked on a beach

Description automatically generated

* **Objective of the Project:**
* The purpose of the project is to gather the data and determine what causes the accident and the attributes that leads to the severity.
* Through data visualization and machine learning algorithm we will be analyzing a significant range of attributes, including weather conditions, road condition, speeding, special events, roadworks, traffic jams among others and we will try to predict what are the conditions that can contribute to high severity accidents which may cause loss of life or loss of property .WSOT can use the model to take precaution to minimize the loss of property and life.
* Reducing the insurance cost and Preventing fatalities
* **Stakeholders:**
* Government Officials
* Emergency Responders (911 dispatchers)
* Common People
* Insurance Companies

**Data understanding**

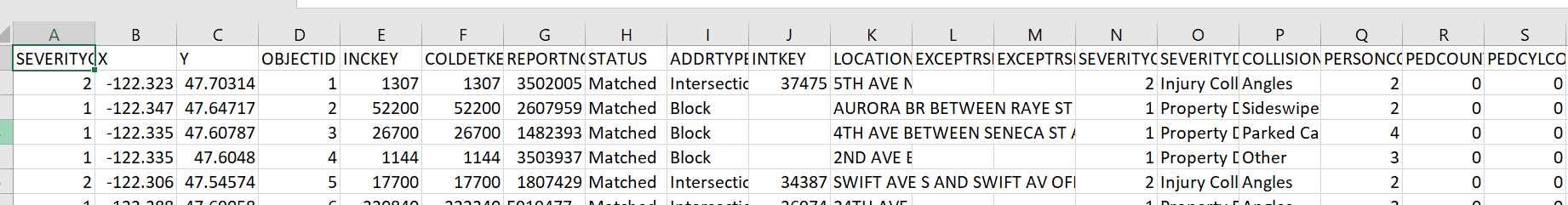
We chose the public data from open source available with labeled columns and attributes and observations data to help us do our analysis better.

Sample Data Below

Link to the Data

<https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv>

https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions\_OD.pdf



The data consists of 40 independent variables and 221738 rows. The dependent variable, “SEVERITYCODE”, contains numbers that correspond to different levels of severity caused by an accident from 0 to 4.

Severity codes are as follows:

0: Unknown

1: Property Damage

2: Injury

2b: Serious Injury

3: Fatality

**Data Preparation**

# Public Traffic data for Seattle city USA is available from the Open source (Link Mentioned above)

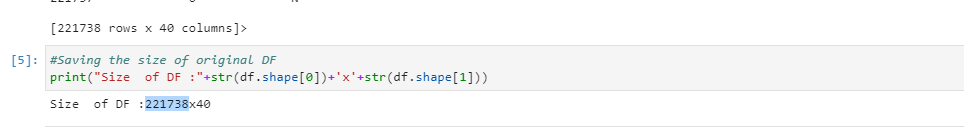
# After the data has been extracted, keeping the columns required in the data frame.

# Excluding the rows with null values.

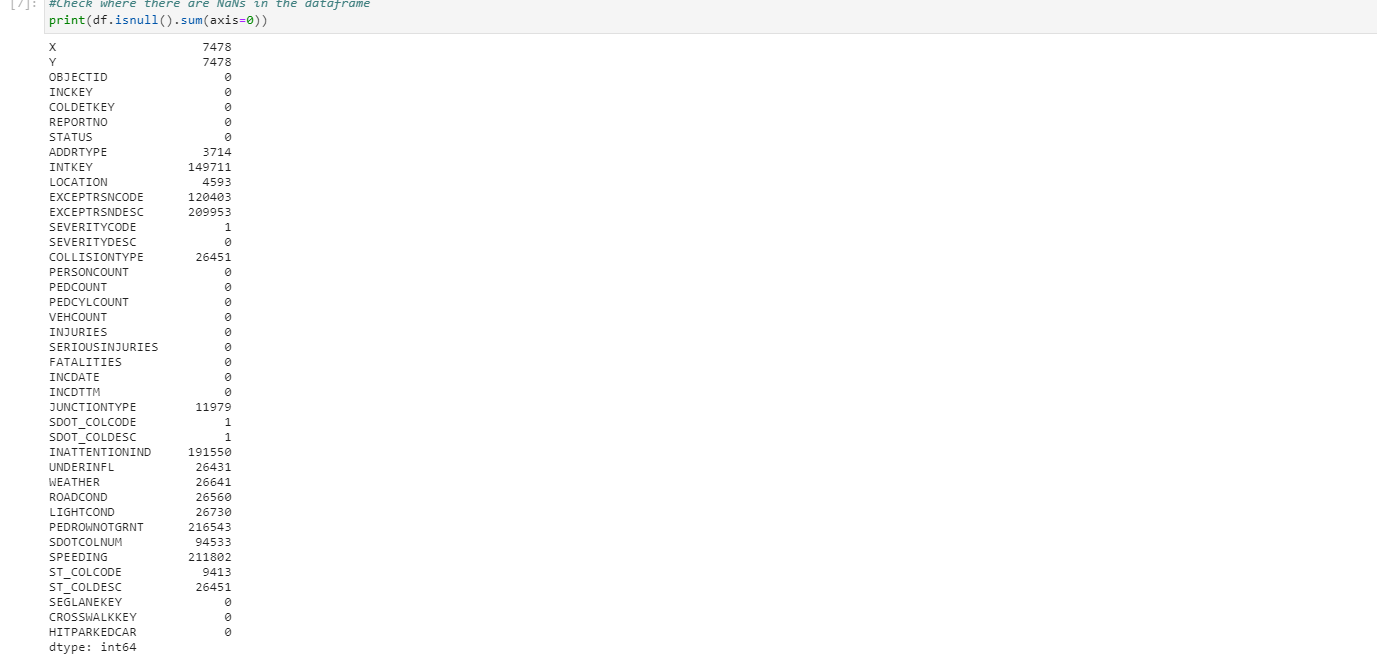
# Transform the data type for analysis.

* Load the Data to data frame.

Original Size of the Data frame

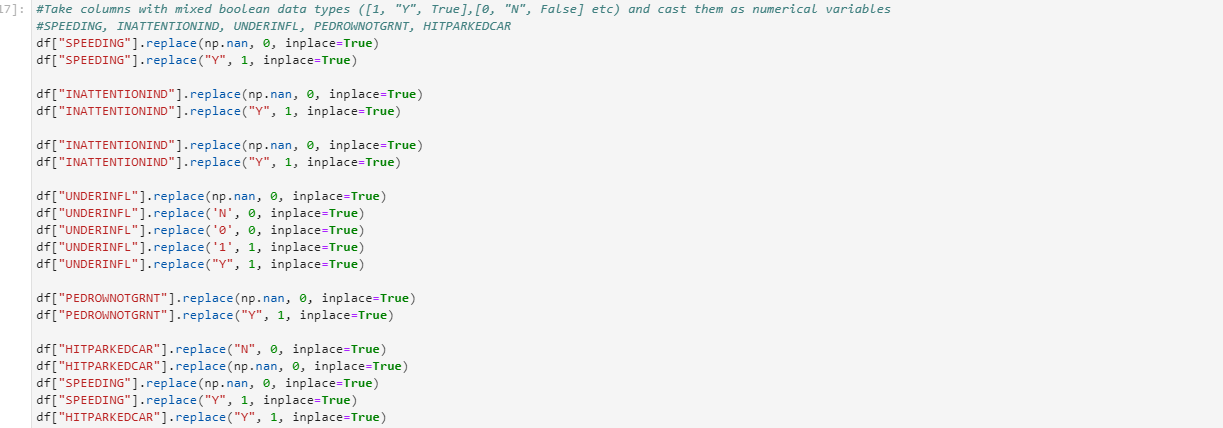


Missing Values from the Data frame



Preparing the Data after doing the Data Cleansing

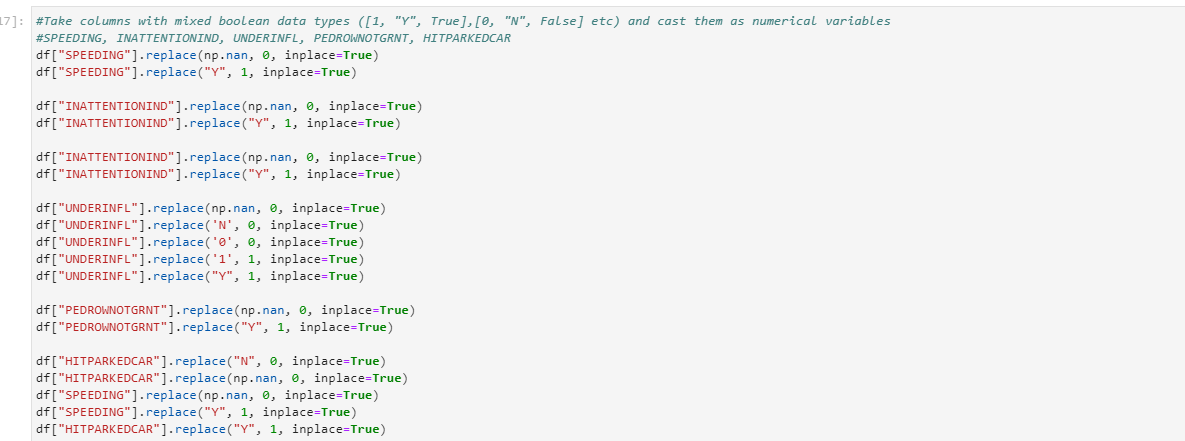


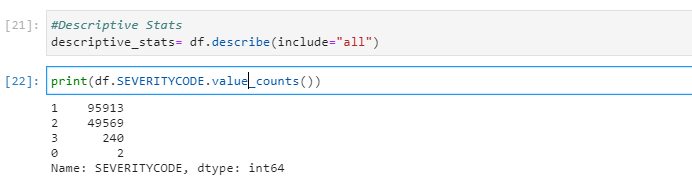


Data Cleansing-Dropping the unwanted columns.

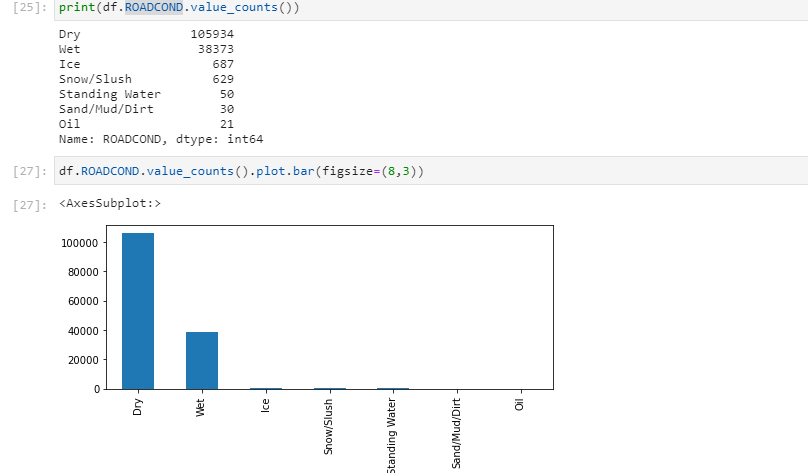


Casting the columns to right datatype (numerical variables) for calculations.



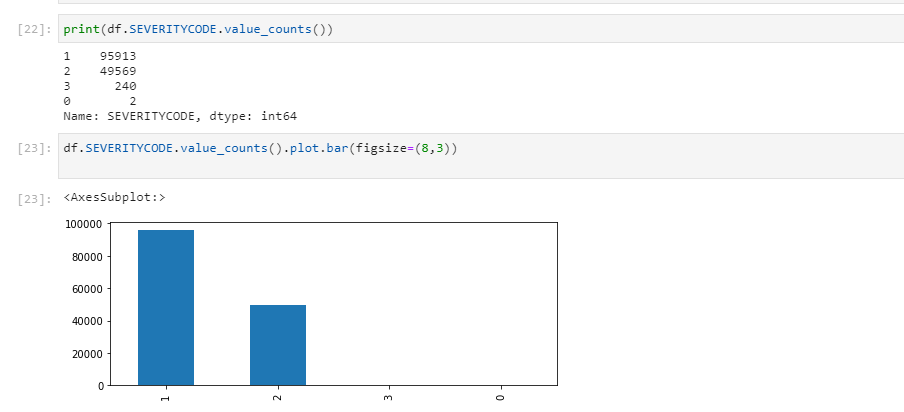


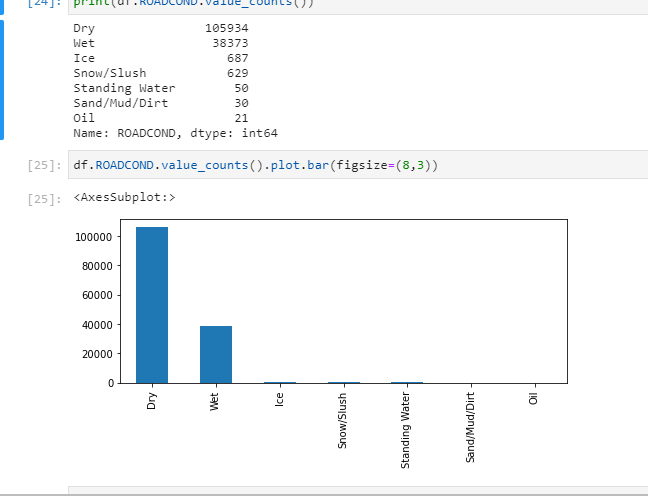
Count Based on the Road Condition

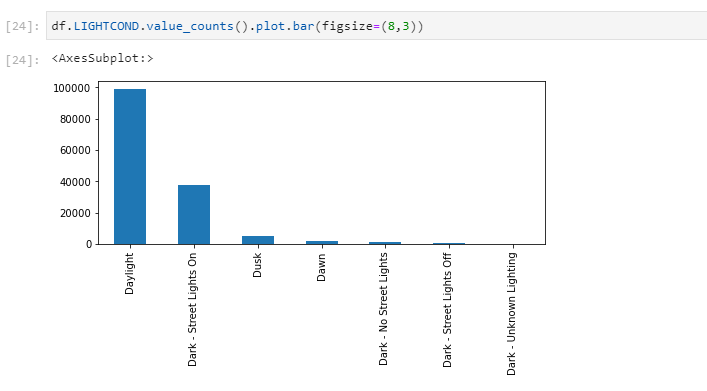
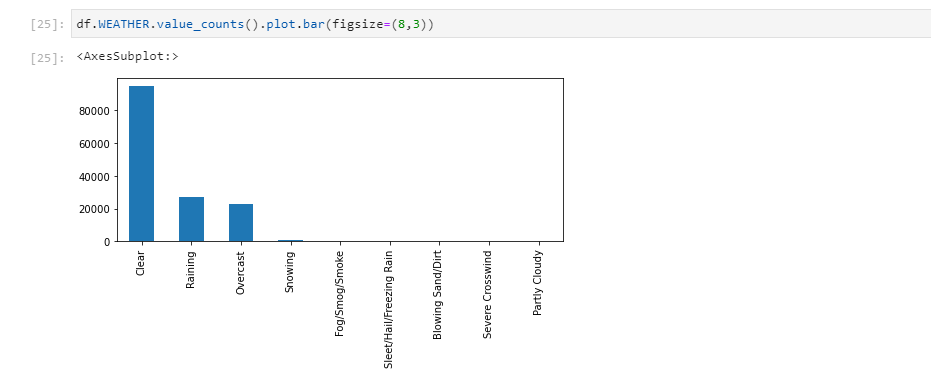


**Exploratory Data Analysis**

We will run a value count on road (‘ROADCOND’) and weather condition (‘WEATHER’) to get ideas of the different road and weather conditions. We will also check the value count on light condition (’LIGHTCOND’), to see the breakdowns of accidents occurring during the different light conditions. The results will then be used for data modeling.



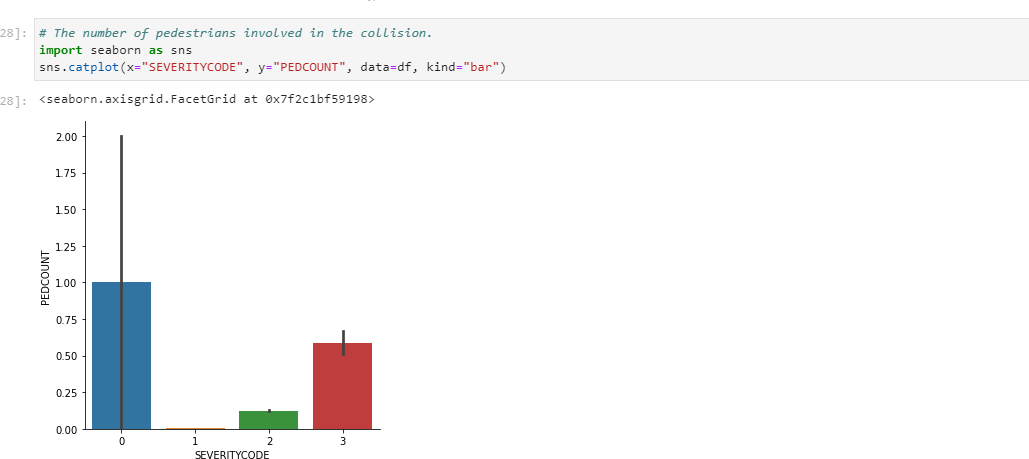




We built the catplot with seaborn library below to check how the severity Is impacted by various attributes.

# Number of pedestrians involved in the collision (PEDCOUNT) and Severity

# Severity 2 (Injury) accidents has happened to Pedestrians, compared to Sev 1 accidents



# Number of people involved in the collision (PERSONCOUNT) and Severity[¶](https://render.githubusercontent.com/view/ipynb?commit=9d23f29c27aede2302c67e94475b547df4a918ac&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f61767262616472692f436f7572736572615f43617073746f6e655f41737369676e6d656e742f396432336632396332376165646532333032633637653934343735623534376466346139313861632f73656174746c656163636964656e747373657665726974792e6970796e62&nwo=avrbadri%2FCoursera_Capstone_Assignment&path=seattleaccidentsseverity.ipynb&repository_id=293546239&repository_type=Repository#Number-of-people-involved-in-the-collision-(PERSONCOUNT)-and-Severity)

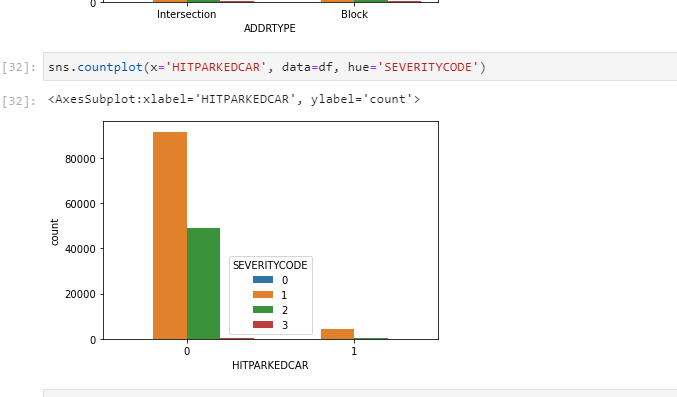
# The data shows that the severity of the accident is high with person count.

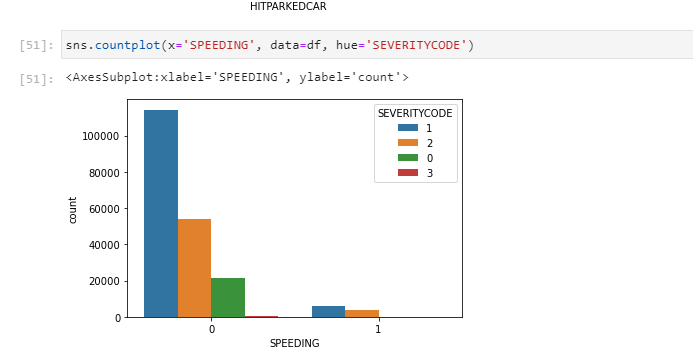
# 

# The below data shows accidents are happening more near the blocks and less at intersection. Severity 2 is almost same on Block and on Intersection

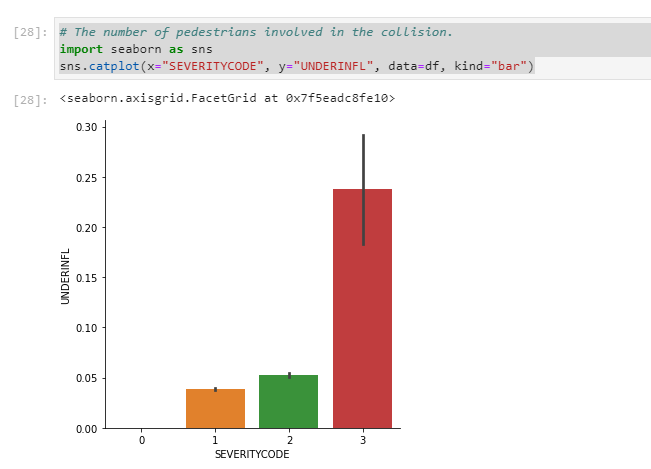
# 

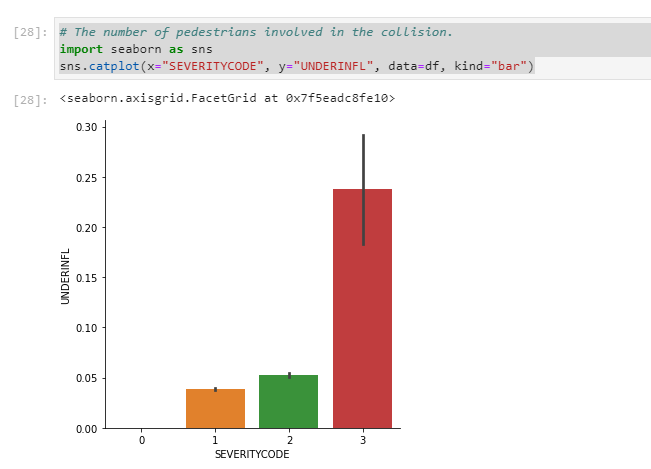
* Number of cases of accidents when car was parked



Looking at the Speeding cases, there were 9381 speeding cases.

* Figure shows the high number of accidents when people were under influence

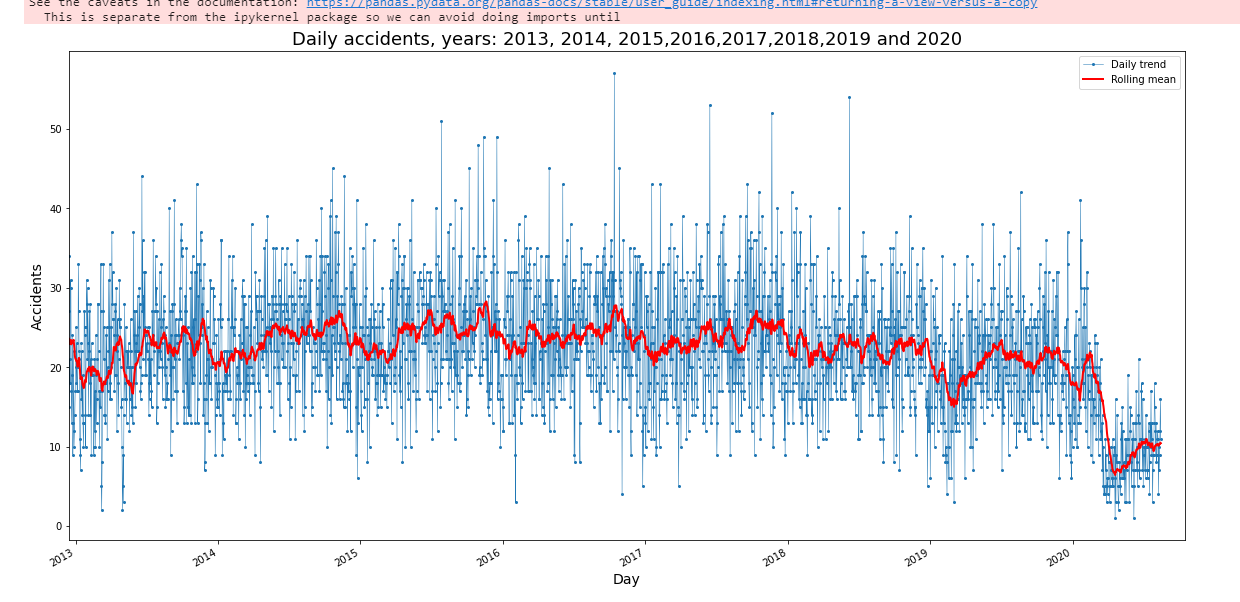




We further created an Incident dataset. Following features were used: "Incident Date", "Incident Time","CollisionDescription”,”Weather”,”Road condition”,”Light Condition” indicating the Incident Detail.

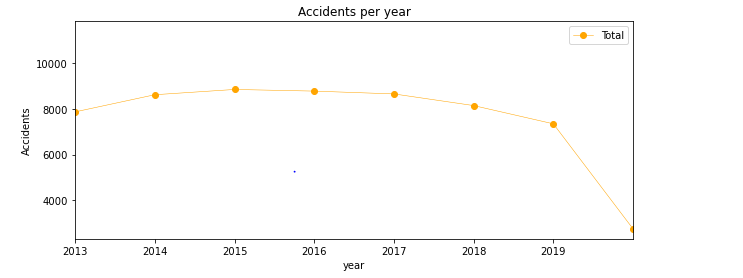


The Below graph shows the daily accidents from 2013 till 2020 with high around end of 2016 and beginning of 2017 and the number of accidents low in 2020.



Plotting the accidents Per year, confirms the same trend as above



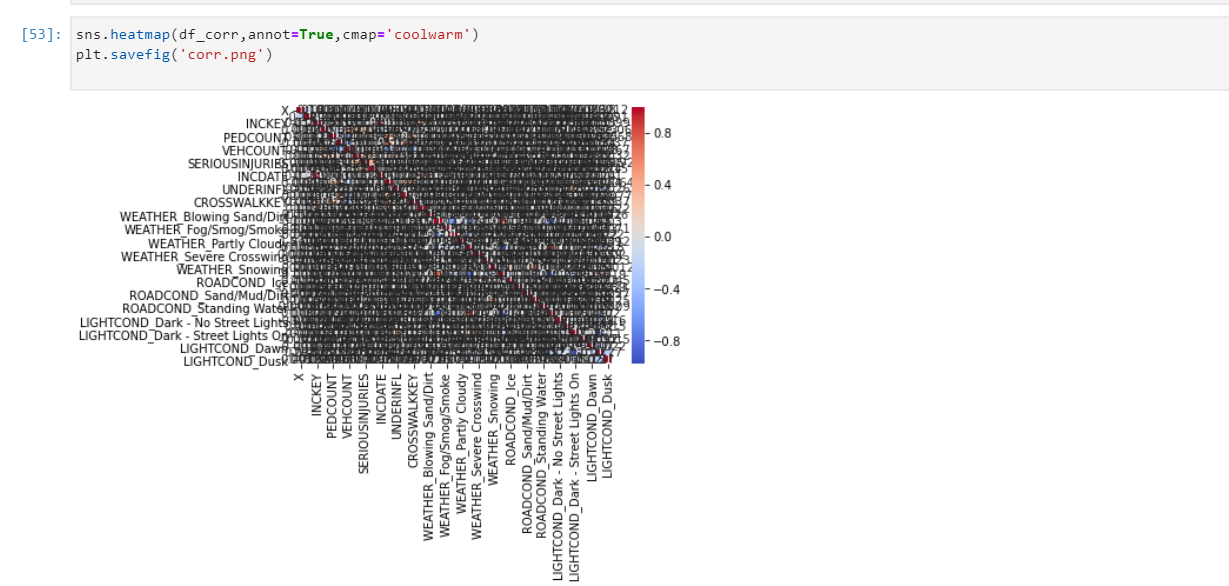


**Modeling**

* Using Numpy, Scarlar, Linear regression on the clean transformed data (Shown Above)
* After importing necessary packages and splitting preprocessed data into test and train sets, for each machine learning model, we will build and evaluate the model with the techniques as follow:
* Dataframe with features below created [“ADDRTYPE”,”COLLISIONTYPE,”JUNCTIONTYPE”,”WEATHER”,”ROADCOND”,”LIGHTCOND”,”UNDERINFL”,”HITPARKEDCAR”]

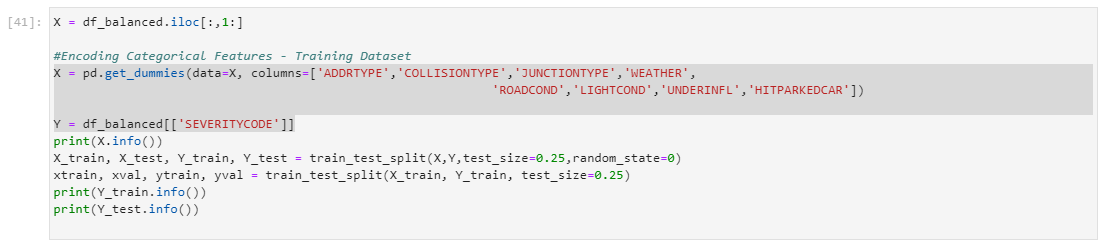


Heatmap

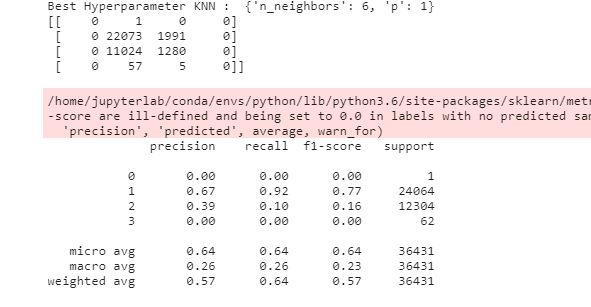


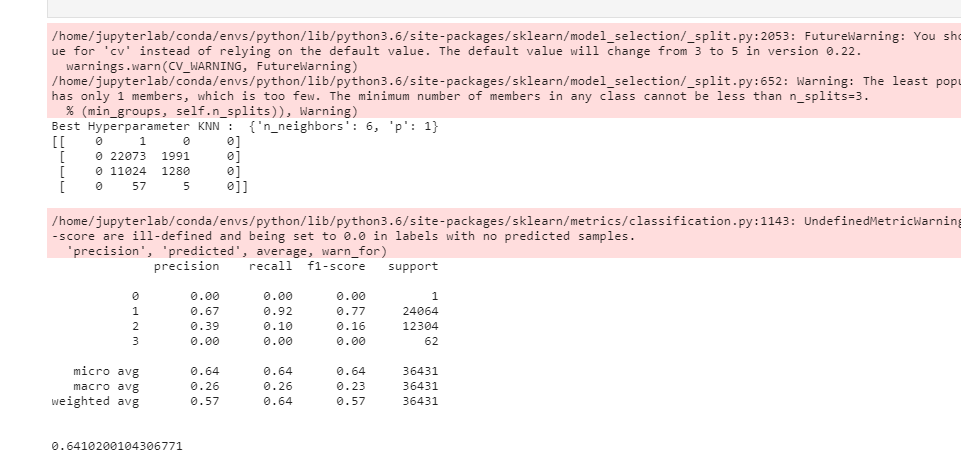
Machine Learning Models

* Github as a repository and running Jupyter Notebook are used to process data and build Machine Learning models
* Python and its popular packages such as Pandas, NumPy and Sklearn is used for determining the accuracy.
* The dataset x and y are constructed.After normalization they are split into x\_train,y\_train,x\_test and y\_test using train\_test split.75% of the data is used for training and 25% is used for testing as below

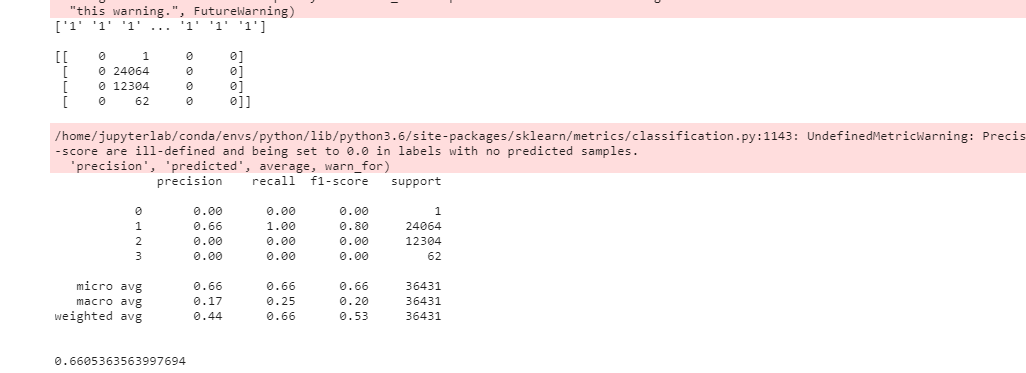


* K Nearest Neighbour (KNN)

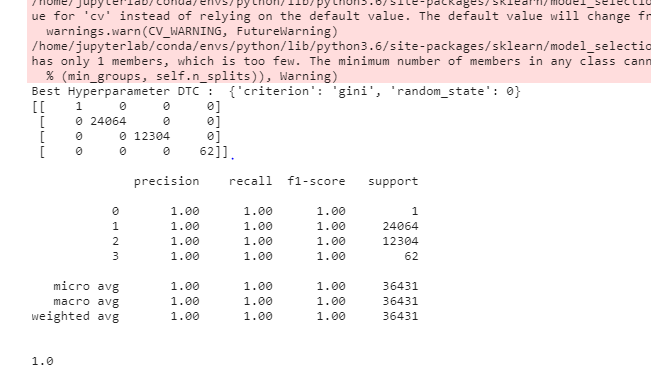




* LINEAR REGRESSION



* DECISION TREE



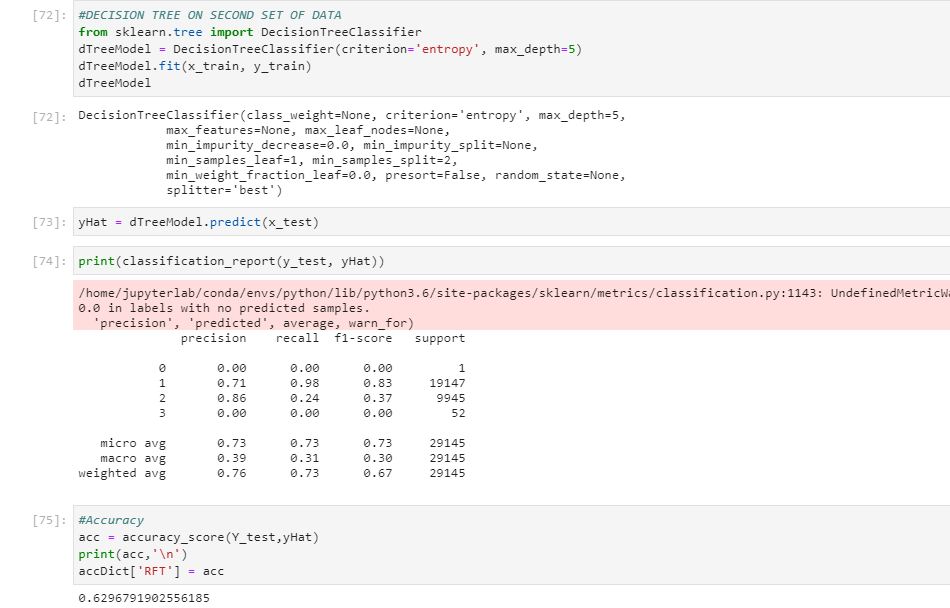
Based on the data sample of train and test data we see there is an Overfitting, meaning our model is learning the noise from the data and its ability to generalize the results is very low. In this case you have a small training error but very large validation error. So, we will try to prune the data set and run the model again.

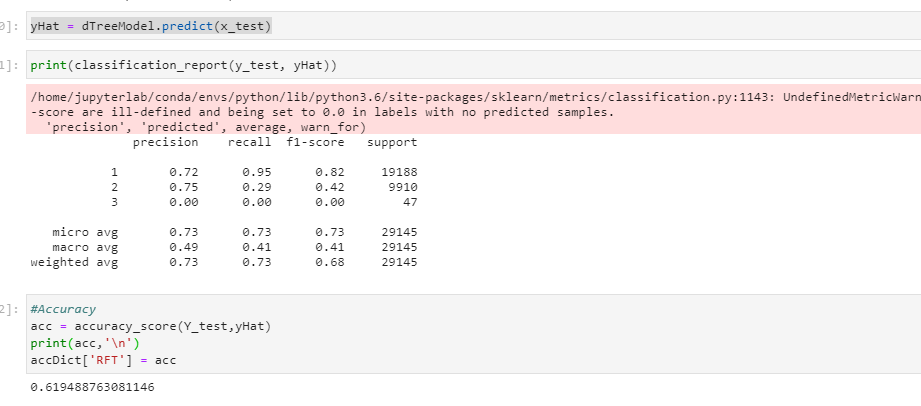
We changed our dataset and started again to find the accuracy and which model fits the best.

Dataframe:data\_clean

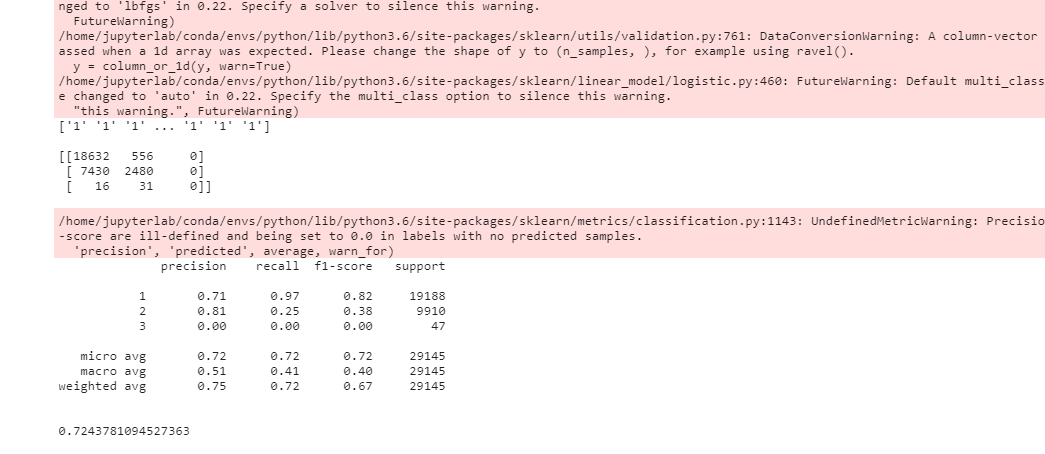


* DECISION TREE

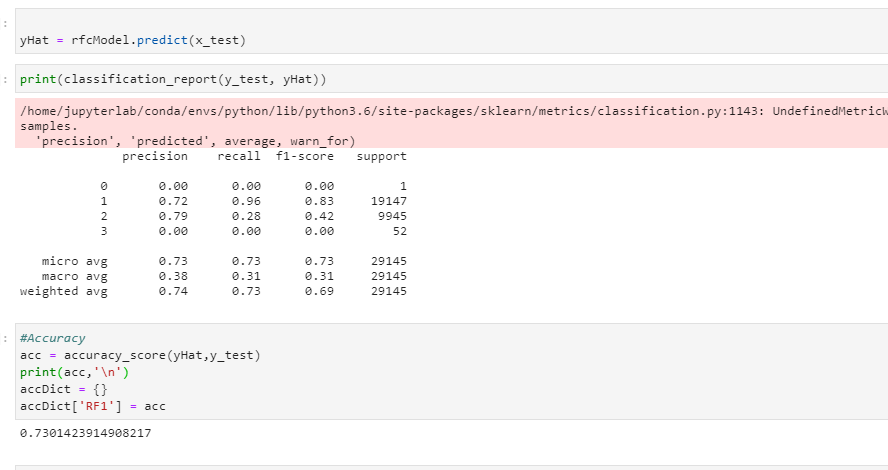




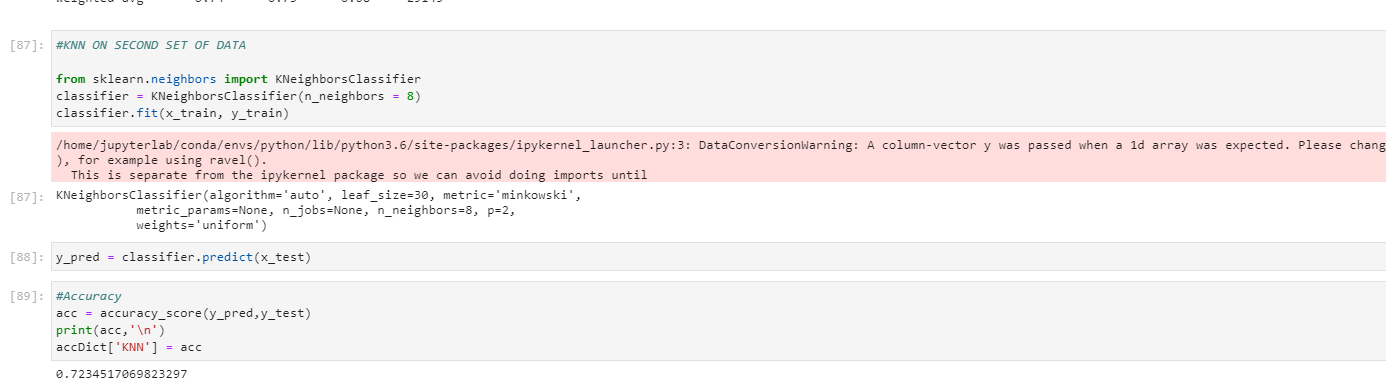
* LINEAR REGRESSION



* RANDOM FOREST



* KNN



**CONCLUSION**

We showed the impact of weather, road condition,road intersection,underinfluence on the two classes property damage (class 1) or injury (class 2) with graphs.

Except DecisionTree the other three models have a close accuracy of 72-73% which means that the model has trained well but we can have more accuracy with additional features and by having more data for severity code and missing description.

KNN,Logistic Regression, Random forest have an accuracy of 72-74% ,this is possibly the similarity of the features for both types of accidents(1 and 2)

Though with 73% of accuracy from Random Forest we can say that the model has trained well and performs well on the testing as well as the trained data.