Introduction / Problem Description :

In the United States and throughout much of the world, car accidents are a leading cause of serious injury and death. In fact, in the U.S. alone, at least 38,800 people were killed in motor vehicle collisions in 2019. One of the worst affected city from these motor vehicle accidents in USA is Seattle.

Seattle is s a seaport city on the West Coast of the United States. It is the largest city in the state of Washington, as well as the largest in the Pacific Northwest. As of the latest census, there were 713,700 people living in Seattle. Seattle residents get around by car, trolley, streetcar, public bus, bicycle, on foot, and by rail. With such bustling streets, it's no surprise that Seattle sees car accidents every day.

The results of these accidents are mostly severe. People loose their live or suffer from serious injuries which remain there for very long, sometimes. These people who survive traffic accidents can face hefty medical bills, thousands of dollars in lost wages and property damage, pain and suffering, and lost quality of life. A collision could cause permanent disability, stripping the victim of the life he or she might have had.

Car accidents happen for a host of reasons, including behavioral, environmental, and situational. The behavioral reasons inlcude the carelessness from the driver's side like - Drug and Alcohol impairment, Over Speeding, Driver's Distraction, etc. There can be some environmental reasons as well like- bad lightning conditions, rainy and windy weather. Vehicle sudden malfunctioning can be another reason for accidents.

A small number of car accidents are inevitable and can't be prevented. Most of them, however, could at least be prevented, and many result from poor decisions by someone who should have done better.

It would be great if real-time road conditions and alerts can be provided to estimate the trip safeness. In this way, it can be decided beforehand if the driver will take the risk, based on reliable information. For example, if the driver is going towards a route where there is windy and rainy weather, and low lightning condition, and he gets a prior notification about it, he can try to be more alert whle driving on that route, or even try to change the route, if feasible.

Through this project, I will be trying to assess the severity and different reasons for these accidents and will try to figure out the ways in which these figures can be brought down significantly.

The target audience of the project, who will be most benefitted by the outcome of this project are - Seattle people, local Seattle government, police, rescue groups, and the insurance companies as well.

Data Description :

Data, if enquired propely, can give us a lot of insights. The dataset that we will be working to analyze the accident data is provided by Seattle Police Department(SPD) and recorded by Traffic Records. The dataset contains all types of collisions between 2004 to Present . This dataset is updated weekly and contains 194673 rows and 38 columns.

Out of these 38 columns, first column is 'SEVERITYCODE', which is our dependent variable. Rest of the 37 columns are the attributes. All the attibutes are not important for our model, and so from these 37 attributes, we will choose some of the important attributes for our model. They consist of :

- **ADDRTYPE** : Collision address type: Alley, Block, or Intersection
- **LOCATION** : Description of the general location of the collision.
- **PERSONCOUNT** : The total number of people involved in the collision.
- **PEDCOUNT** : The number of pedestrians involved in the collision.
- **PEDCYLCOUNT** : The number of bicycles involved in the collision.
- **VEHCOUNT** : The number of vehicles involved in the collision.
- **INJURIES :** The number of total injuries in the collision.
- **INCDTTM** : The date and time of the incident.
- **FATALITIES :** The number of fatalities in the collision.
- **JUNCTIONTYPE** : Category of junction at which collision took place.
- **UNDERINFL** : Whether or not a driver involved was under the influence of drugs or alcohol.
- **WEATHER** : A description of the weather conditions during the time of the collision.
- **ROADCOND** : The condition of the road during the collision.
- **LIGHTCOND** : The light conditions during the collision.
- **SPEEDING** : Whether or not speeding was a factor in the collision (Y/N).
- **SEGLANEKEY** : A key for the lane segment in which the collision occurred.
- **CROSSWALKKEY** : A key for the crosswalk at which the collision occurred.
- **HITPARKEDCAR** : Whether or not the collision involved hitting a parked car.

We will further process the dataset in order to make it apt for our use.

The dataset we will be using can be downloaded from the link provided by Coursera.

Methodology :

In this section, I will discuss and describe the exploratory data analysis that I did on the dataset, the data prepararion methods I used, and what machine learnings were used and why.

Data Preprocessing :

As described above in the Data Section, I started with 11 variables inluding the dependent variable – 'SEVERITYCODE'.

I explored the unique value(s) of each column to know what types of values are there in each column. During exploring those values, I came across certain columns like 'INCDTTM', 'LIGHTCOND', 'WEATHER', etc. which had some values as mentioned – 'Unknown'. These values were uncertain and so they were of no use to us. So, I dropped all those rows having any value as – 'Unknown'.

```
Shape before dropping rows :  (194673, 11)
Shape after dropping rows :  (175763, 11)
```

Then, there was a column 'UNDERINFL' with inconsitency in data entry. There was 'Y' and 'N', and 1 and 0 as well. So, I converted 'Y' to 1, and 'N' to 0.

In similar way, there wee some inconsistency in two other columns – 'LIGHTCOND' and 'INATTENTIOIND' which were corrected.

After all these, it was time to manage the missing values. Two columns – 'UNDERINFL', and 'INATTENTIONIND' had some missing values, which were replaced with 0.

All the other columns having missing vlues were left as it is, and then all those rows having any missing value were dropped.

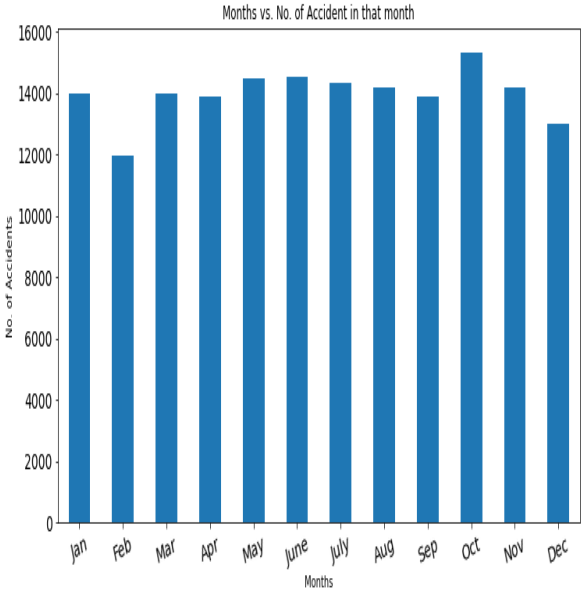
```
data.isna().sum()
```

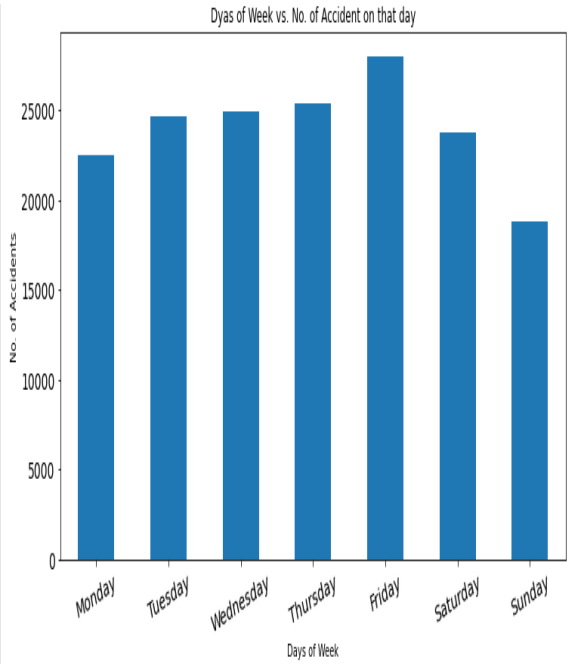
```
SEVERITYCODE       0
ADDRTYPE           0
COLLISIONTYPE      0
VEHCOUNT           0
INCDTTM            0
JUNCTIONTYPE       0
UNDERINFL          0
INATTENTIONIND     0
WEATHER            0
ROADCOND           0
LIGHTCOND          0
dtype: int64
```

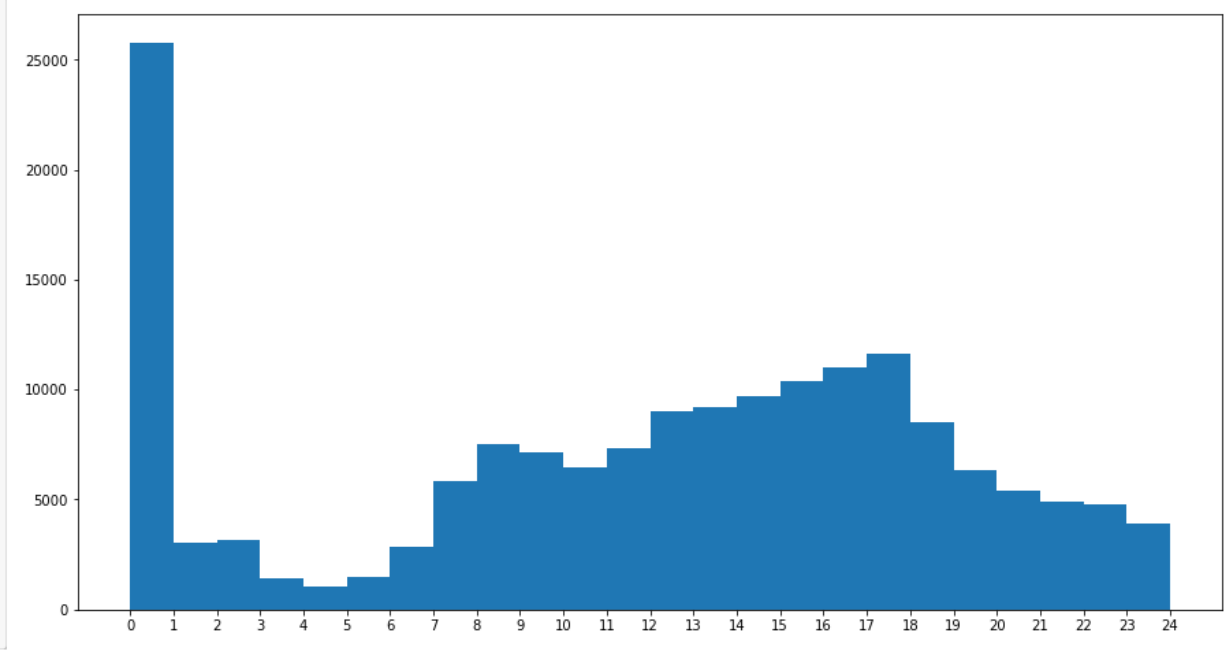
Data Visualization :

After the data cleaning was done, it was the time to visualize to data in form of graphs and plots.

I started with visualizing that whether the no. of accidents has anything to do with the time of the day, or any particulay day, or month, and also the trend of accidents in each year.







From the above plots, I was able to make some observations. Firstly, a positive thing was that every year(except 1-2 in between), the no. of accidents were reducing. Then in terms of week days, more no. of accidents ere observed on Friday, and similarly in terms of months, maximum accidents were there in month of October. At last, as expected most of the accidents happened in the rush hours of day and evening. But the maximum accidents happened in between 12-1am in midnight. The no. was very high.

Similary, plots were drawn to visualize the impact of each feature on the no. of accidents, and I made some interesting observations which I will discuss in details in later 'Observation' section.

Data Preparation for training models :

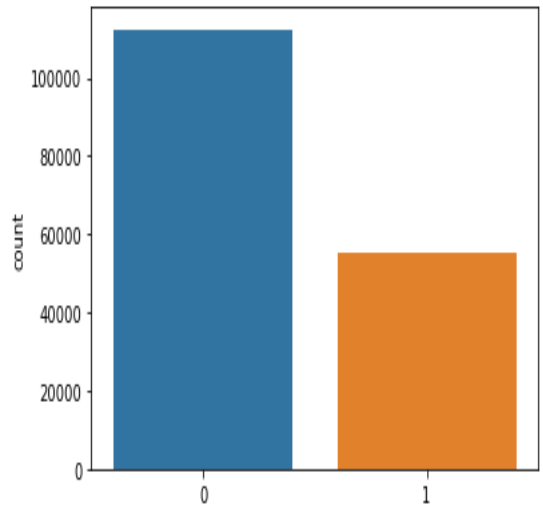
After the visualization phase was over, it was time to prepare our dataset for modelling. I came to this step after data visualization step so that any manipulation don't disturb the actual observation from the real dataset.

I started with converting the categorical values into numerical values. This was necessary because most of the algorithms don't work with categorical data. I used 'one-hot encoding' for conversion.

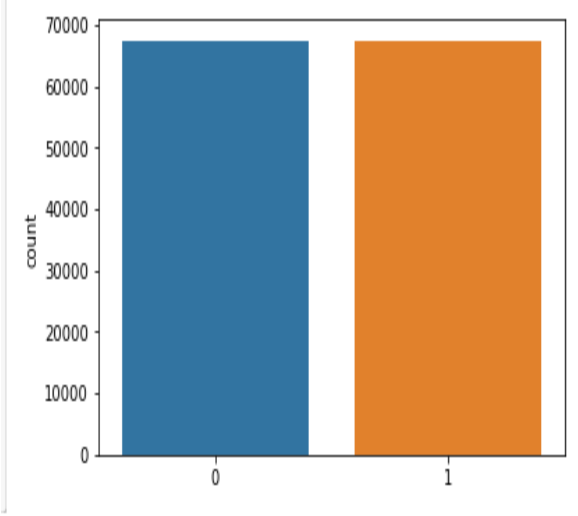
Finally, it was time to attend the imbalance of our dataset. Our dataset was highly imbalanced and going forward without attending it would have produced a biased model in favour of majority class. I used oversampling of minority class followed by undersampling of majority

class. This way I didn't loose much information and also the dataset didn't become bias because of oversampling. I used **imblearn** library for this.

```
sns.countplot(y);
```




```
sns.countplot(y);
```



Training Models :

Now, we are ready for training our model. But before that, we will split our training set in training and test data set in ratio of 70-30. 70% of data was used for training set and 30% of data was used as test set. After this data was normalized.

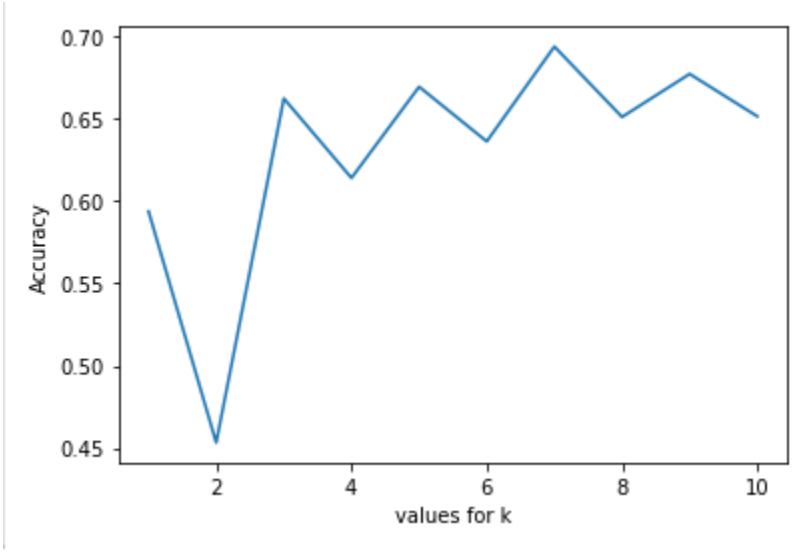
```
Train set: (94388, 44) (94388,)
Test set: (40452, 44) (40452,)
```

The algorithms that I used for tarining my models are : K-Nearest Neighbors, Decision Tree Classifier, SVM classifier, and Logistic Regression. To measure the accuracy of my model, I used three performance metrics – jaccard similarity score, F1 score, log loss.

KNN Classifier

KNN classifier is one of the most simple algorithm for classification task. The most important task while using KNN to carefully choose the value of k. For this, I choose 10 values from 1 to 10 and tested my model on each value of k to choose the optimal value. After checking for each

value of k, the most optimal value was k = 7 with 69.37% accuracy. I used this for my final model using KNN.



Decision Tree Classifier

Decision Tree Classifier is a tree-structure based algorithm. Depth of the tree is the hyperparameter for this algorithm. After trying different values for d(depth), I choose d=4.

```
Depth =  4 : 0.7213826303121176
```

Then I trained my model on **SVM Classifer** and **Logistic Regression Classifier,** which are both simple, and easy to implement as well as yields very good result most of the time.

Result :

After training my models using four different algorithm, it was time for prediction of test_data to check the accuracy of our model and to decide which model is giving us the best result.

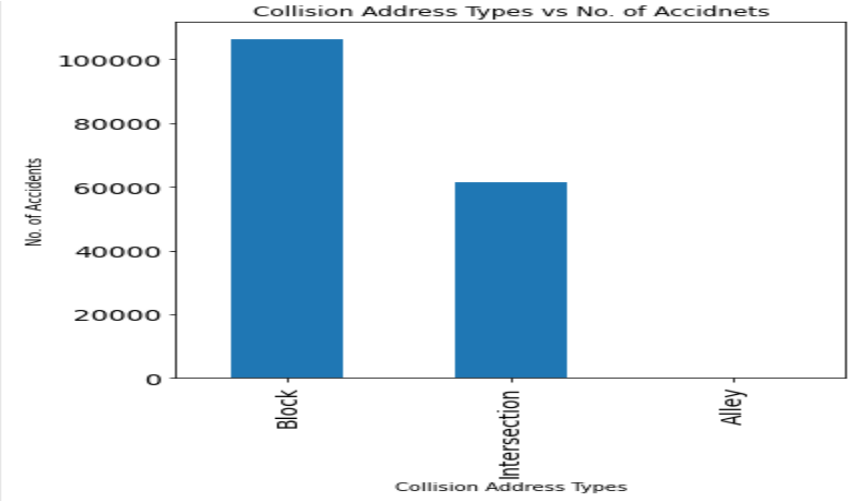
	Algorithm	Jaccard	F1-score	LogLoss
0	KNN	0.53	0.65	NaN
1	Decision Tree	0.56	0.63	NaN
2	SVM	0.55	0.67	NaN
3	LogisticRegression	0.55	0.67	0.58

From the above table, we can see that all algorithms are giving very similar accuracy results for the performace metrics we used. But, if we have to chhose one, then we can go for Logistic Regression. It is simple to implement ans also faster to implemennt.

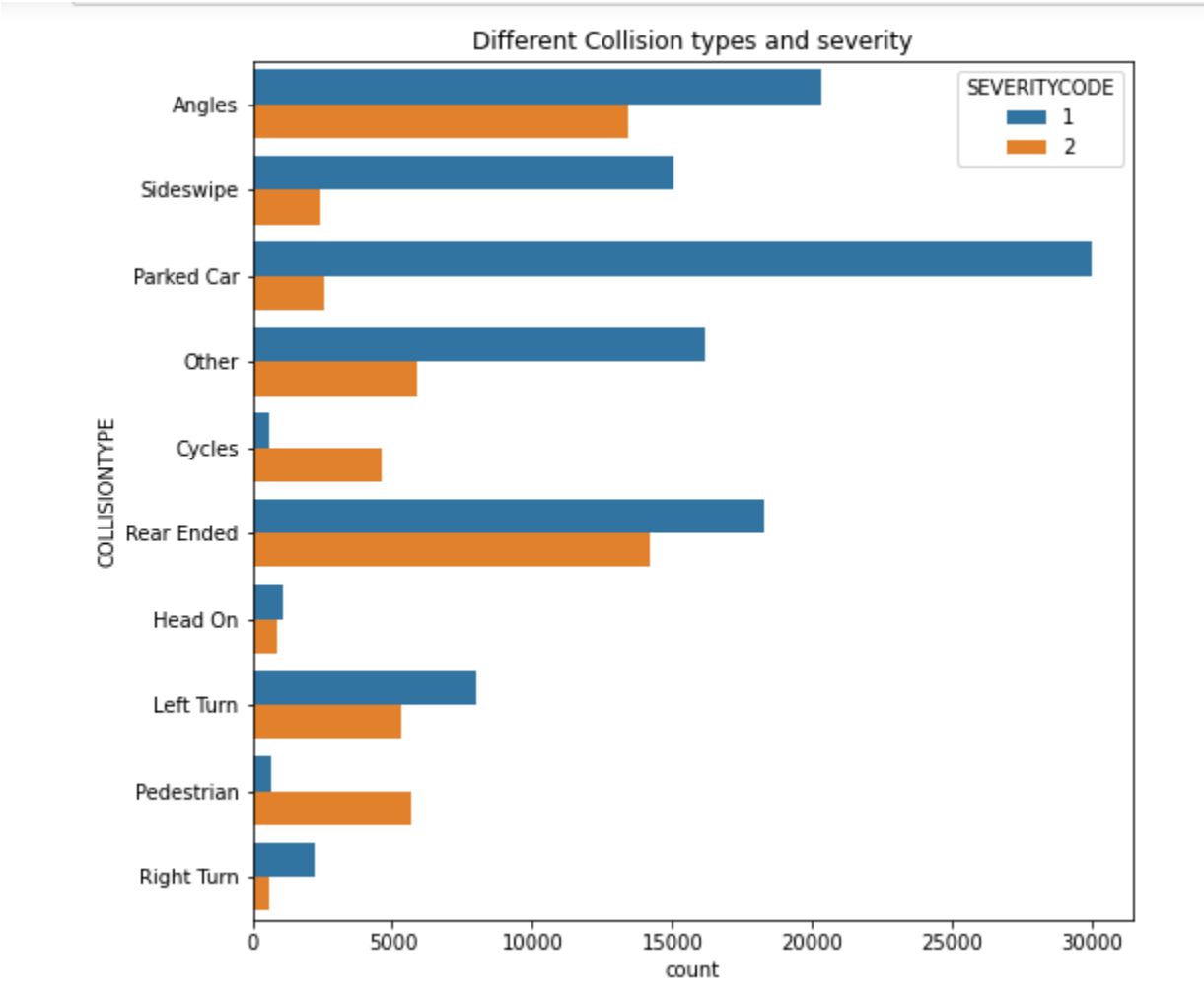
Discussion / Observations:

After all the tasks completed, it's time to discuss the observation that I made. Well, there were quite few interesting observations. Let's start with the relation between no. of accidents and timestamp. As discussed earlier, one of the postive observation was that the no. of accidents has been reducing almost every year. In 2020, it is very low, most probably because of lockdown imposed. Also, as expected most of the accidents happened in the rush hours of day and evening. These can be reduced by a better traffic arangement in these rush hours. October was one month having maximum no. of accidents. It can be figured out that whether there is any specific reason for that and if it can be solved.

Also, I found that most of the accidens where happening at Intersection and Block types. This is again something that can be improved by a better traffic management and better implementation of road signals.

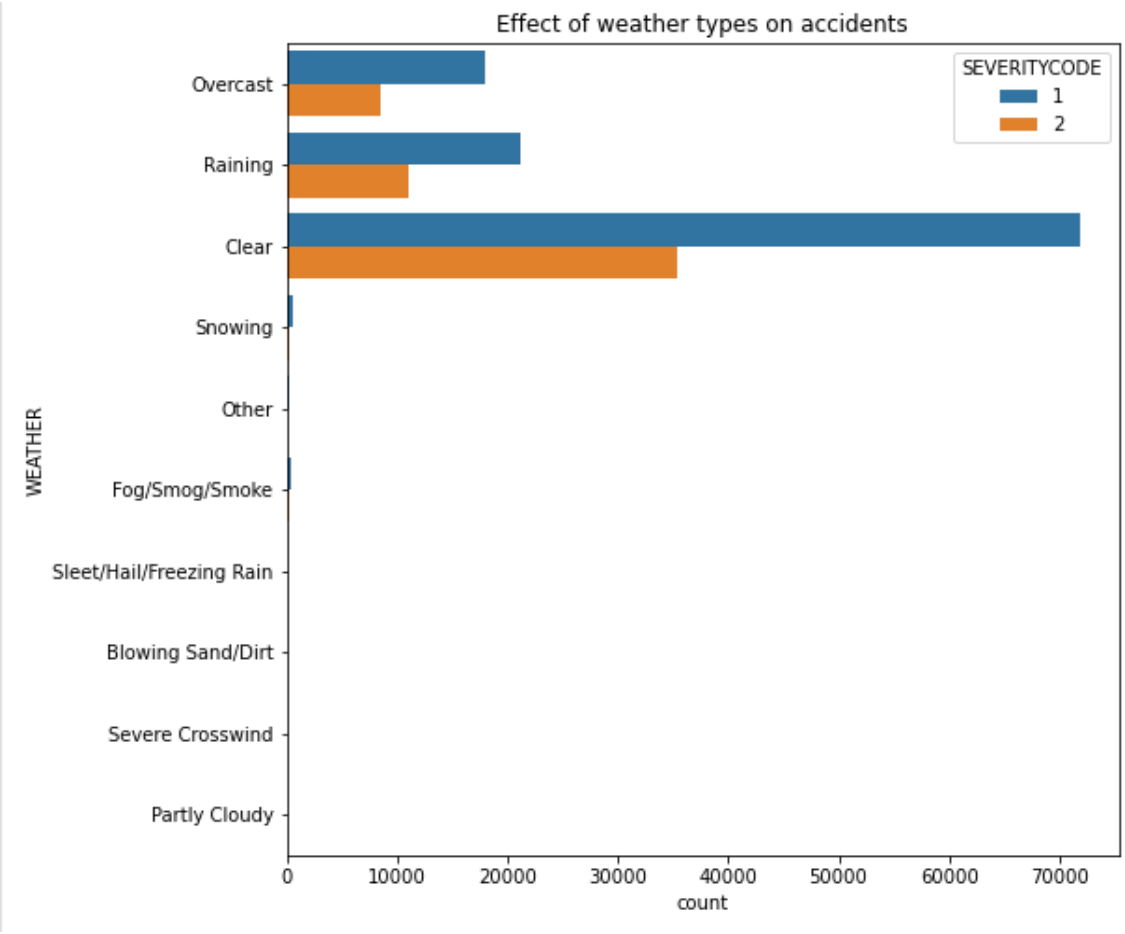


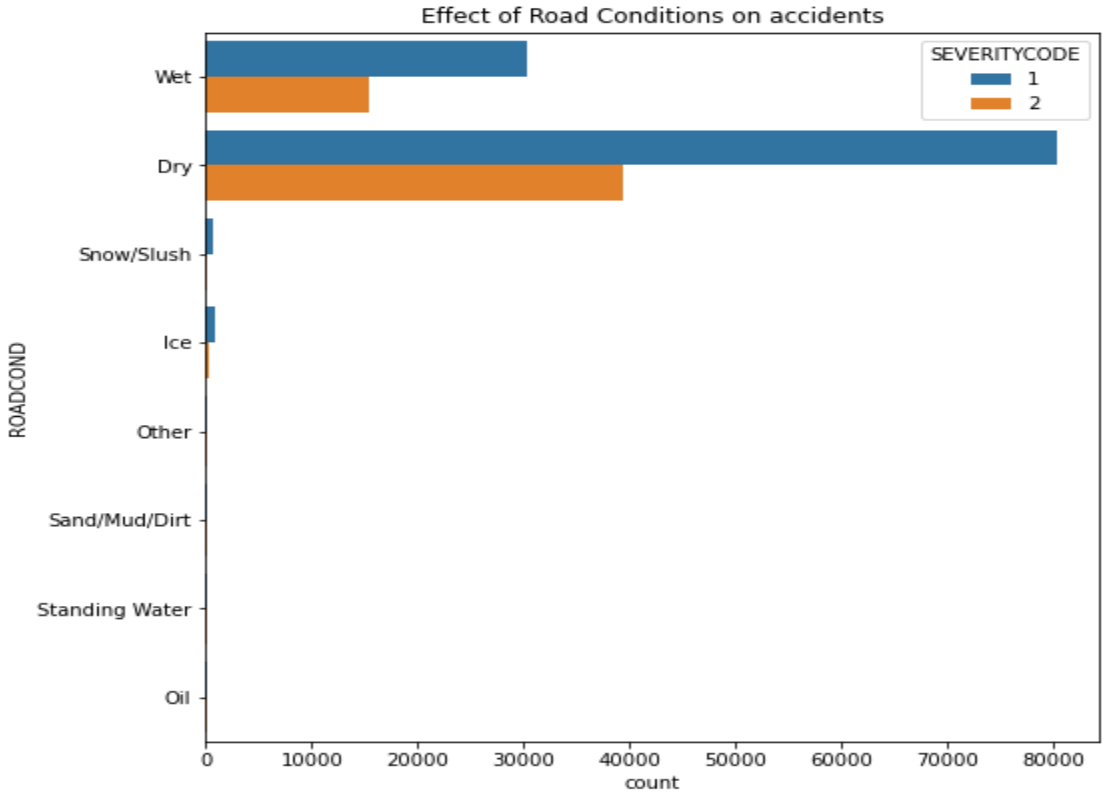
Also, I found that many cases were there in which parked cars were damaged, or it was a rear end collision. This simply means that the accidents happened due to the carelessness of a single person in most of the cases .

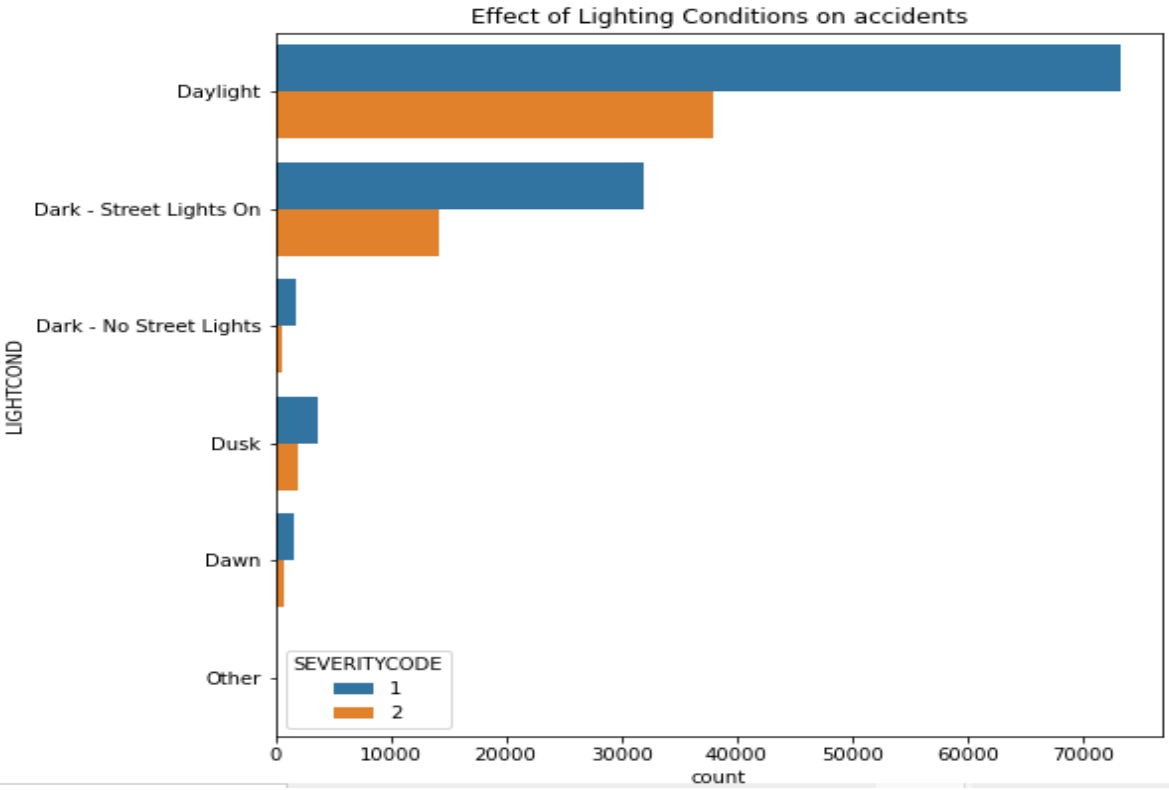


The most interesting finding for me was the impact of weather and lighting condition on no. of accidents. Before beginning the visualization step, I was thinking that bad weather like rainy and windy weather, or snow, or absence of street lights must be playing a huge role in the no. of accidents. But for my surprise, most of the accidents happened on clear day and dry road, and that too in daylight. Although, there were significant number of accidents on wet road, or in overcast conditions, or in absence of steet lights in dark, we can't simply ignore that maximum accidents happened when the conditions were apt for driving. This again implies towards the carelessness of the drivers. The drivers need to be trained properly and made responsible.Also, the street light should be properly maintained which can reduce the no. of accidents significantly. One another way of protecting the drivers from accidents on days when there is bad weather,or heavy traffic on road, is by sending them alerts in advance about the

adverse conditions. So that either the driver can take proper care while driving in that way, or if possible canchange their way to some alternate route.




Effect of Road Conditions on accidents
SEVERITYCODE
1
2
Wet
Dry
Snow/Slush
Ice
Other
Sand/Mud/Dirt
Standing Water
Oil
ROADCOND
0
10000
20000
30000
40000
50000
60000
70000
80000
count


Effect of Lighting Conditions on accidents
Daylight
Dark - Street Lights On
Dark - No Street Lights
Dusk
Dawn
Other
LIGHTCOND
SEVERITYCODE
1
2
0
10000
20000
30000
40000
50000
60000
70000
count

Concluion:

I was able to make a lot of intersting observations about the no. of accidents and their severity after analysis of the dataset. Overall, it was evident that most of the accidents were happening because of human carelessness, which canbe significantly improved by proper training of drivers and awareness among them. But, there were also a lagre no. of cases of missing street lights which needs to be improved. The traffic management also needs to be improved in rush hours and road signals should be propely insatlled to prevent the accidnets at intersection and blocks. Also, as discussed earlier, an alert system can be developed for the drivers to alert them of heavy tarffic or adverse weather condtions on their route, in advance. All these measures can help in reducing the no. of accidents an can prevent a lot of lives and property loss.

With this, I would like to conclude my project.

Thankyou for studying this report and for analyzing my work. I hope you would have liked it.