INTRODUCTION

BACKGROUND: In the modern world road traffic accidents are very common in each and every part of the world. Seattle is one the busiest cities and as of 2017, there have been more than 11,000 motor vehicle-involved **collisions** per year. These collisions not only lead to high property damage but may result in injuries and in worst case scenarios even death. In 2017, a total **of** 187 fatal and serious **injury collisions** were reported on **Seattle** streets. The causes behind the collisions may range from physical factors like road conditions, lighting, weather, time of day etc to human factors like inattention, speeding, under influence. Various efforts and steps can be taken in order to minimize these collisions. Every city has devised certain traffic rules and regulations to help in this regard. Even, WHO has given certain recommendation to be followed by the government agencies to have an impact on the rate collisions and more importantly to reduce the cost of property damage as well as of life. Even small regulations sometimes have much greater impact.

PROBLEM: To predict the severity of collision based on the various physical and human factors. And make recommendations and take necessary actions based on those predictions like improving the road conditions, lighting, regulations on speed in various area etc.

STAKEHOLDERS:

General Public- vehicle drivers and pedestrians to take necessary precautions in certain circumstances.

Seattle Traffic department- to prepare plans regarding improving road conditions, lightning, speeding limits, etc

Car manufacturing industry- To increase car re-enforcements at the different sites of cars which are frequently involved and especially responsible for the most serious injuries and fatalities.

Weather department- to make necessary updates and caution messages for different locations depending upon weather.

Health and paramedical departments- to take necessary steps to provide faster emergencies services in the areas more prone to accidents to reduce the loss of life.

DATA.

Data utilized for this analysis was downloaded from Kaggle Seattle collision dataset as csv file.

http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab\_0.csv

DATA UNDERSTANDING AND CLEANING:

The dataset includes data regarding the severity of collisions and various parameters associated with it in the city of Seattle from year 2004 till now.

After exploring the data, we found to have 40 parameters, among those

|  |
| --- |
| 'OBJECTID', |
| 'INCKEY', |
| 'COLDETKEY', |
| 'REPORTNO', |
| 'STATUS', |
| 'EXCEPTRSNCODE' |
| 'EXCEPTRSNDESC', |
| 'SDOT\_COLCODE', |
| 'PEDROWNOTGRNT |
| 'ST\_COLCODE', |
| 'SDOTCOLNUM', |
| 'SEGLANEKEY', |
| 'CROSSWALKKEY', |
| 'HITPARKEDCAR’ |

were found to be least relevant.

Parameters like 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'EXCEPTRSNCODE','SDOT\_COLCODE', 'ST\_COLCODE','SDOTCOLNUM','SEGLANEKEY', 'CROSSWALKKEY' mostly consisted of some unique number or code assigned to each case which did not add any more useful information to the analysis.

Parameters “INCDTTM” and ‘INCDATE” were utilised to form columns like “Year”, “Month”, “Weekday” “Hour” and “TIMEOFDAY”. TIMEOFDAY column was created to see relation ship between the severity of accidents and the time of the day like early morning, morning, noon, evening, night and late night. Columns “INCDTTM” AND “INCDATE” were also eventually dropped.

There were lot of parameters with very high number of missing values. Dropping all those values would have resulted in loss of lot of information. So we tried to include as much as data as possible for the analysis.

'UNDERINFL’ column had values of “N”,” Y”,”1”,”0”. So “1” and ”0” were converted to “Y” and “N” respectively.

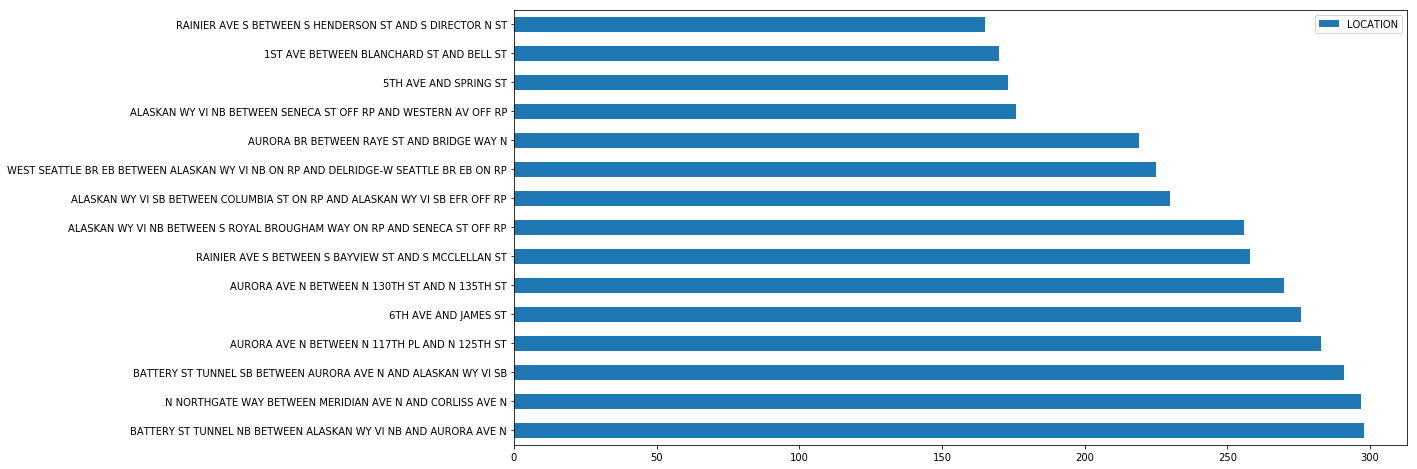
‘INATTENTION’ and ‘SPEEDING’ columns were found to have only “Y” values. So, it was safe to consider that all NAN values would be “N”.

All the NAN values for parameters '' ADDRTYPE, 'SEVERITYCODE', 'SEVERITYDESC', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'SDOT\_COLDESC', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'SPEEDING','ST\_COLDESC were converted to either “Unknown” or “not mentioned” or “not stated” value.

After all this there were missing values for only the parameters “X”,” Y” and

“LOCATION”.

Before dropping the rows with NAN values for these parameters we analysed the top15 locations with maximum number of accidents.

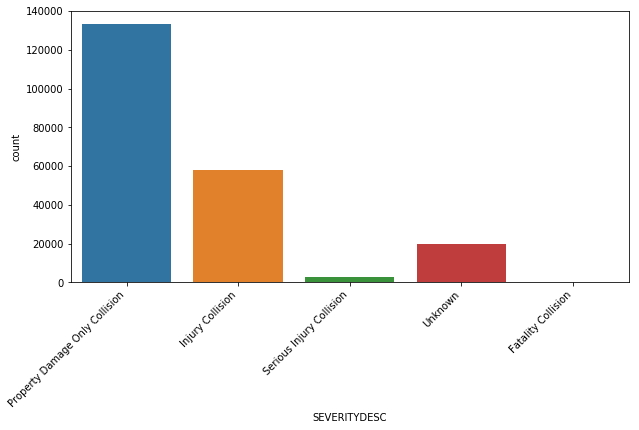


At the end we had the following parameters:

|  |  |
| --- | --- |
| PARAMETERS | Description. |
| ‘X’ | Latitude |
| Y | longitude |
| LOCATION | location |
| COLLISIONTYPE | Type of collision rear, side swipe etc |
| ADDRTYPE | Whether collision occurred at block or intersection |
| SEVERITYCODE | Code to describe severity of collision |
| SEVERITYDESC | Description of severity, property damage, injuries, serious injuries, fatalities etc. |
| JUNCTIONTYPE | Type of junction, intersection, ramp, etc |
| SDOT\_COLDESC | Type of collision with a car, pedestrian, sidehit etc. |
| INATTENTIONIND | Whether the collision occurred due to inattention |
| UNDERINFL | Whether the collision occurred to the person being under influence |
| WEATHER | Condition of weather at the time of accident |
| ROADCOND | Condition of road at the time of accident |
| LIGHTCOND | Status of lighting at the time of accident |
| SPEEDING | Whether the collision occurred when the vehicle was speeding. |
| ST\_COLDESC |  |
| Weekday | Weekday the collision occurred |
| Hour | The hour the collision occurred |
| Year | The year the collision occurred |
| Month | The month in which the collision occurred |
| TIMEOFDAY | Time of the day at the time of accident , morning, noon, evening, night, late night etc |
| PERSONCOUNT | Number of persons involved in accident |
| PEDCOUNT | Number of pedestrians involved in accident |
| PEDCYCLIST COUNT | Number of cyclists involved in accident |
| VEHICLE COUNT | Number of vehicles involved in accident |

SEVERITY DESCRIPTION: Property damage was seen more frequent than the injuries especially fatalities with creates the imbalanced dataset. Thus, for machine learning we will have to create a balanced dataset which can be done either by

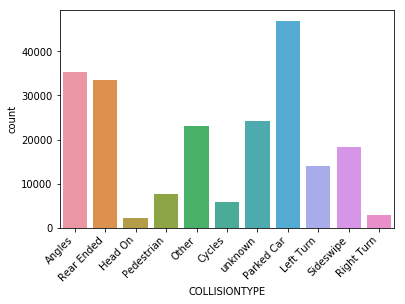
oversampling the minority dataset or under sampling the majority dataset or both.



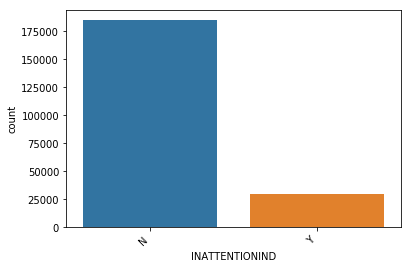
ADDRTYPE



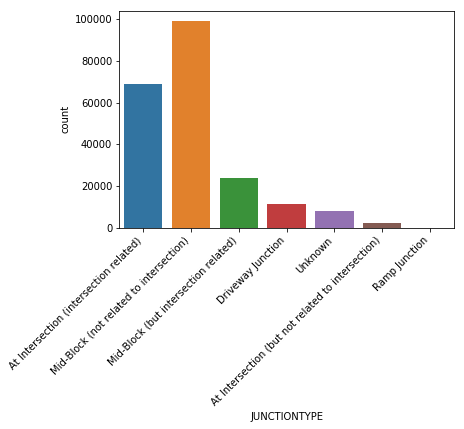
COLLISIONTYPE



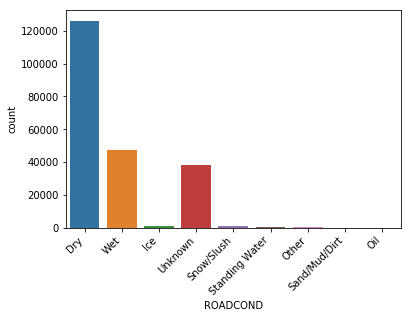
INATTENTION



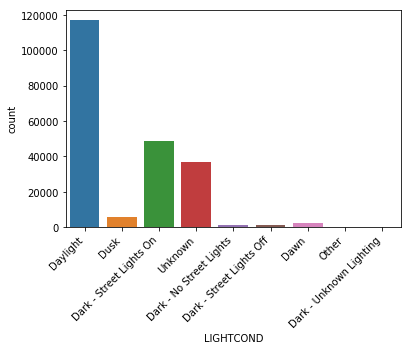
JUNCTIONTYPE



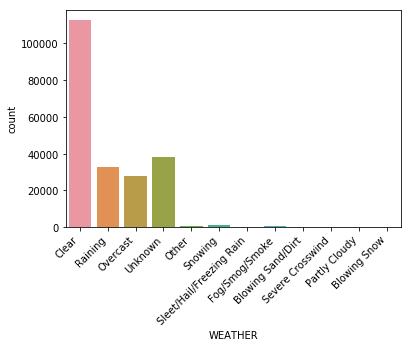
ROADCONDITION



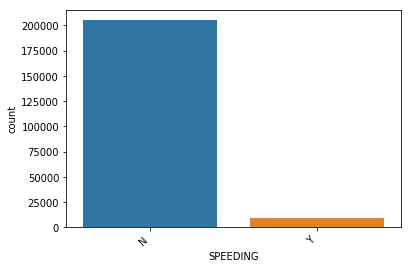
LIGHT CONDITION



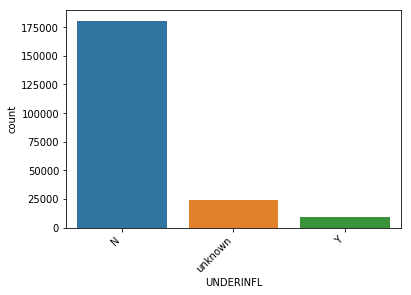
WEATHER:



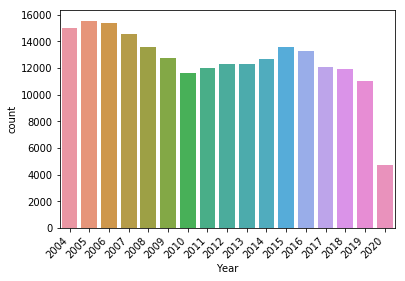
SPEEDING



UNDER-INFLUENCE



YEARWISE COLLISION



TIME OF THE DAY:

