Accident Severity Analysis (IBM Capstone)

1. Introduction

1.1 Problem & Background Description

Motor vehicle accidents can happen quite suddenly, and all too often lead to fatal results. They can involve cars, trucks, motorcycles, and single or multiple vehicles. Some of these deaths occur at the scene of an accident, while others cause injuries that lead to fatalities after the victims receive medical care.

In one report, the World Health Organization (WHO) articulates that the proportion of road accidents fatalities to total deaths in the world has grown by 2.2% from 1.255 million deaths in 2012. The accidents rate was increased by 0.3% from 2000 to 2012 despite the efforts served in terms of better road and laws enforcement. This is how WHO ranks road accidents as the 9th leading cause of deaths with 17.7 per 100,000 population in the world, which is sadly close to that of dangerous diseases like diabetes, diarrhea, HIV/AIDS, etc. In addition, approximately 30 to 50 million population are either injured or permanently disabled every year. Moreover, road accidents every year cause great financial havoc of $518 billion and so costing countries 1% to 2% of their individual GDP alone.

Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

1.2 Target Audience

These analyses can be quite beneficial to the following-

1. Daily commuter's in saving time.
2. Road Traffic control department for making policies & understanding traffic in different areas.
3. Government for making useful policies to reduce chances of accidents.

2. Data acquisition and cleaning

2.1 Data sources

For the purpose analysis, Data used in this capstone project was provided by IBM.

2.2 Data cleaning

The collisions dataset includes statistics of accidents from 2004 to starting quarter of 2020. The dataset contains 194673 records in 38 columns. The first column in the dataset in “**SEVERITYCODE**” which is a labeled data and describes the severity of an accident, remaining columns have different attributes such as “**ROADCOND**” which describes about the condition of road at the time of accident etc.

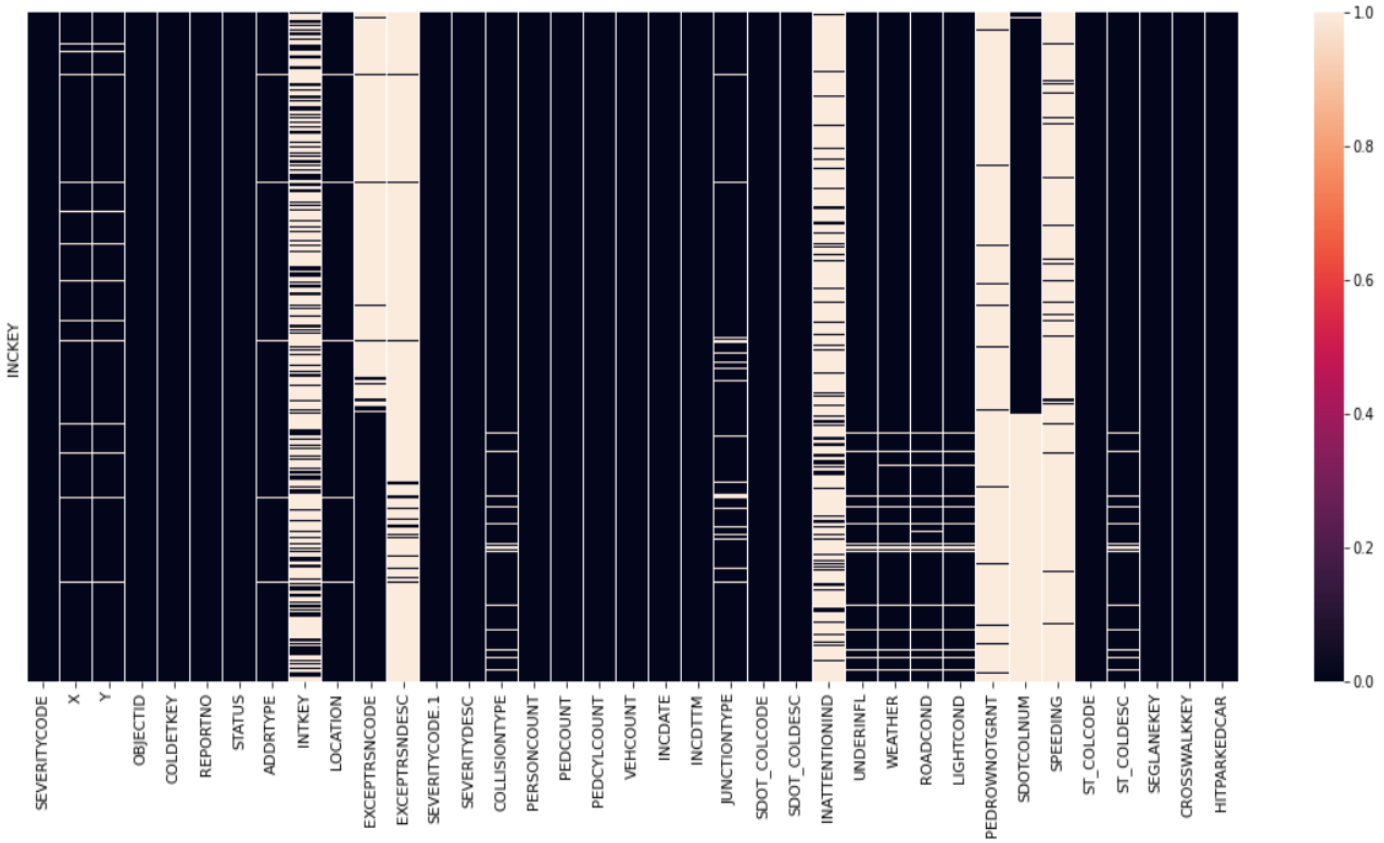
There are several problems with this data that needed to be fixed.

- The dataset has so many missing values in some of the attributes.

- There is a duplicate column of “**SEVERITYCODE**” as “**SEVERITYCODE1**”.
- Datatype of several attributes is not correct.
- Several attributes are not useful in building a machine learning model.
- Dependent variable “**SEVERITYCODE**” has unbalanced data, this can make our model biased.
- There are some inconsistent values in **‘UNDERINFL’.**

Most of the features in the dataset were useful in making a Machine learning model, but they needed to be cleaned in the preprocessing phase.

For handling missing values First, I used Seaborns Heatmap to plot all the missing values.



From the above figure we can conclude that the location coordinates **'X','Y'** values miss at the same time, **'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND'** and **'COLLISIONTYPE'** values miss at the same time, the other values miss at random. To remove all the missing values, I used pandas dropna method to drop all the missing values from the dataset.

There were several features such as **'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SDOTCOLNUM', 'SPEEDING'** etc. Which are not useful to create a ML mode, hence these features were dropped. **‘SPEEDING’** was dropped because it had 185340 missing values. After the selection of useful features there were 19 features left in total.

To convert the data type of **‘SDOT_COLCODE’** and **‘ST_COLCODE’** I used astype() to convert them to integer values. For **‘INCDATE’,** I used pandas to_datetime function. And three new features **(“YEAR”, “MONTH”, “WEEKDAY”)** to the dataset by extracting from **“INCDATE”.**

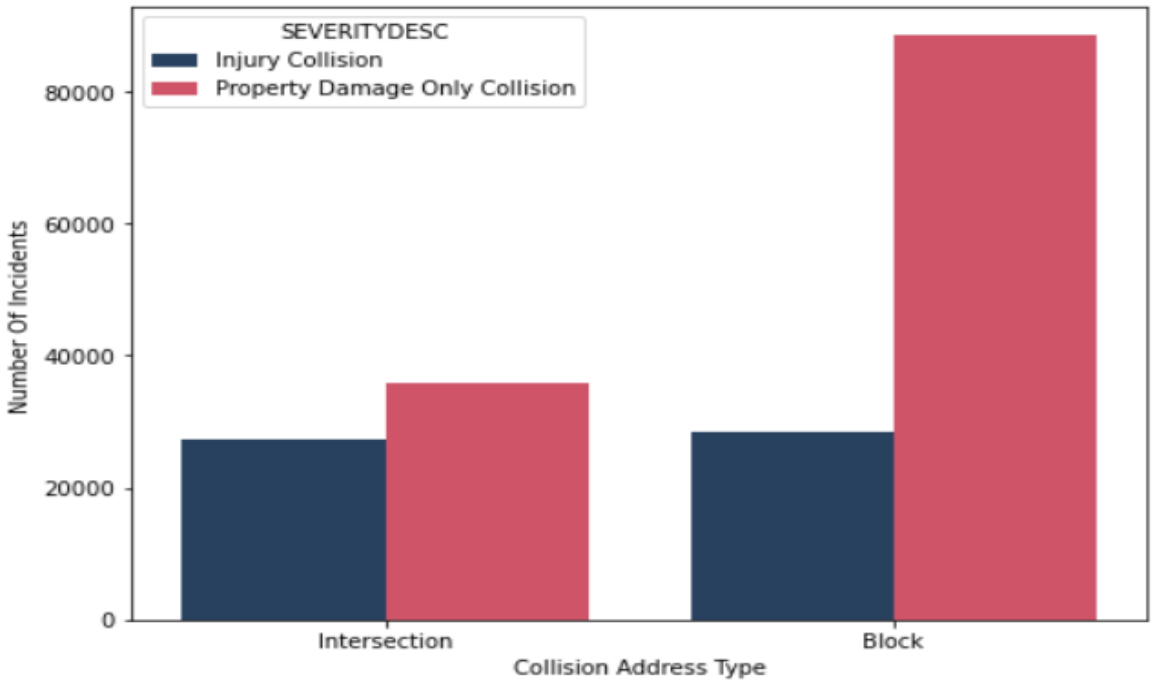
To deal with inconsistency in **‘UNDERINFL’,** I replaced the values “**Y” and “N” to “1” and “2”.**

After the preprocessing phase there were total 180067 rows and 22 columns.

3. Exploratory Data Analysis

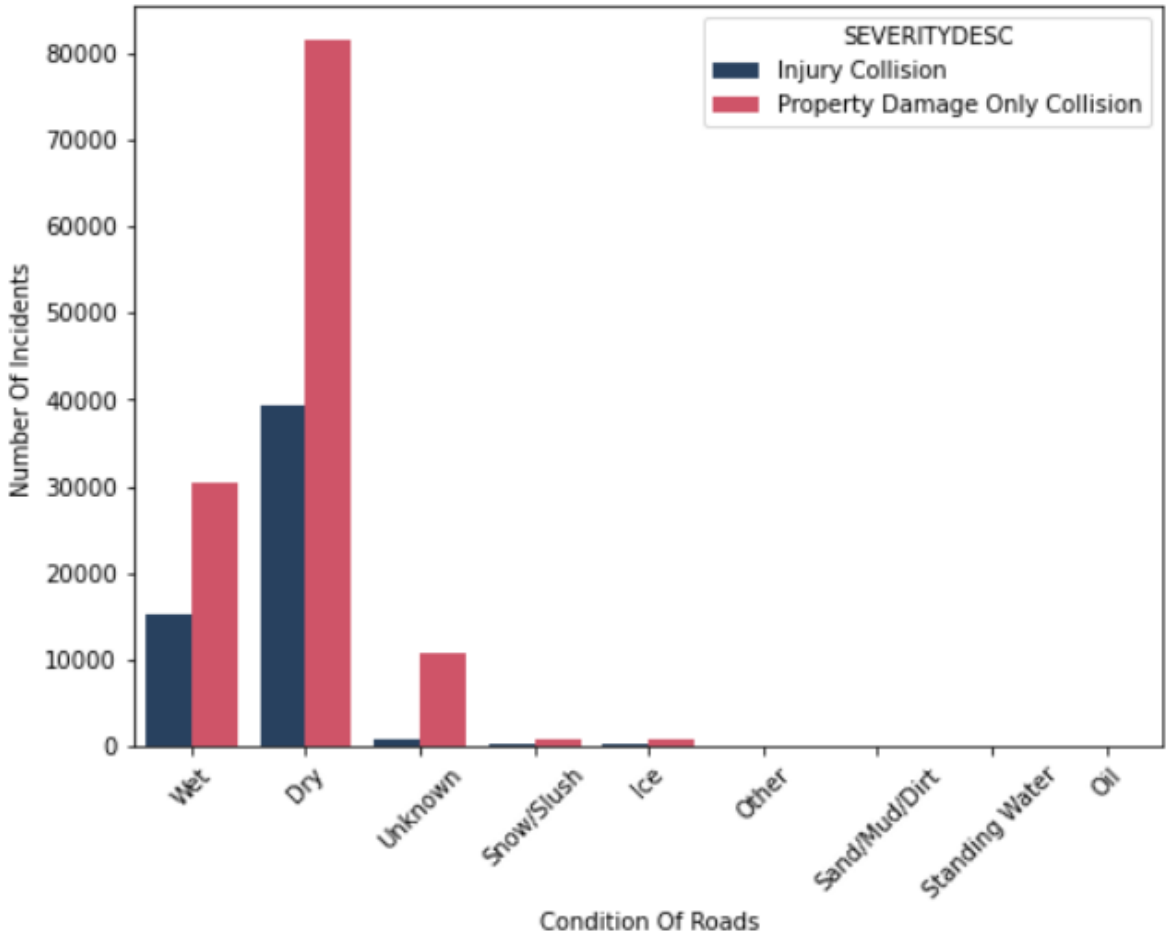
Most of the features in our cleaned dataset are categorical, hence we will use seaborn library to visualize our data.

Relationship between Collision address type and Severity



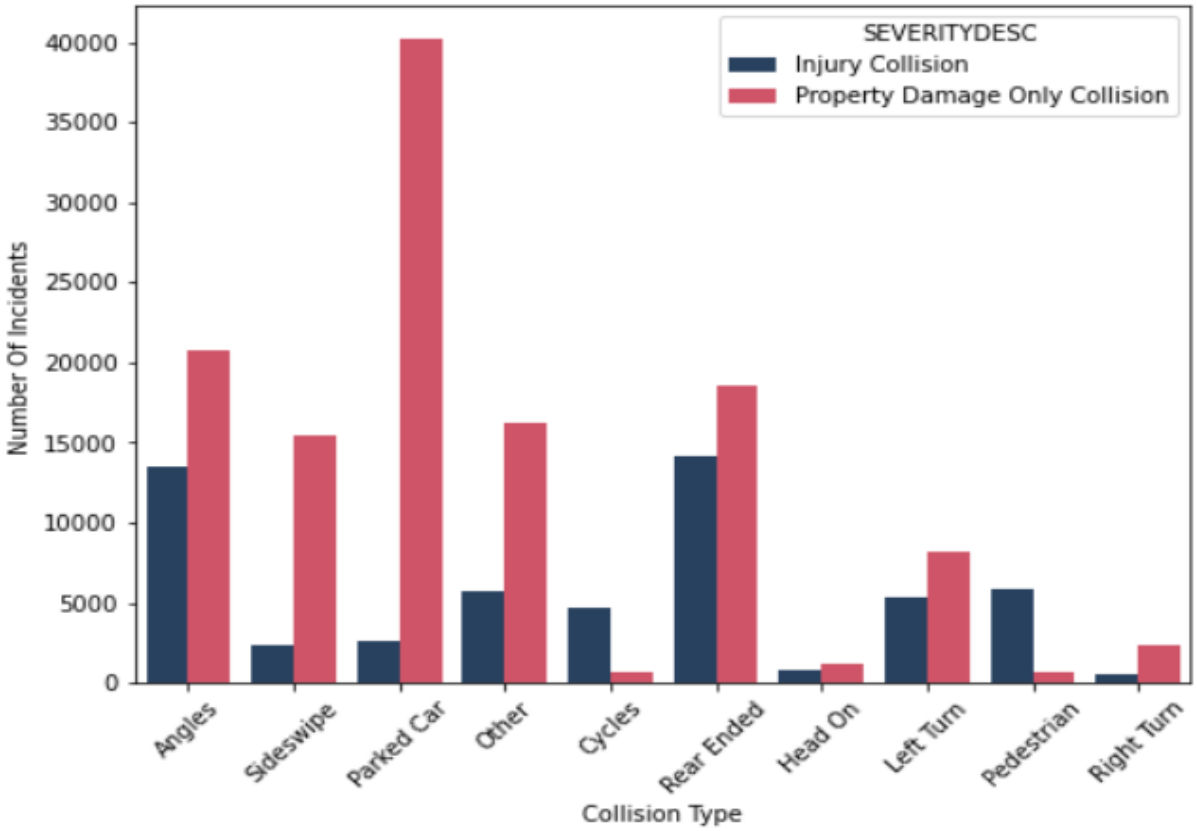
From the above figure we can say that Block leads to more accidents than intersections, Injury collision is almost the same in both the Collision address types i.e. around 25,000 cases, whereas property damage is more recorded in Block address with more than 80,000 cases recorded as compared to around 38,000 for intersection.

Relationship between Condition of Roads and Severity



From the figure above we can observe that more accidents happen on dry roads which are more than 120,000 followed by wet roads with closed 55,000.

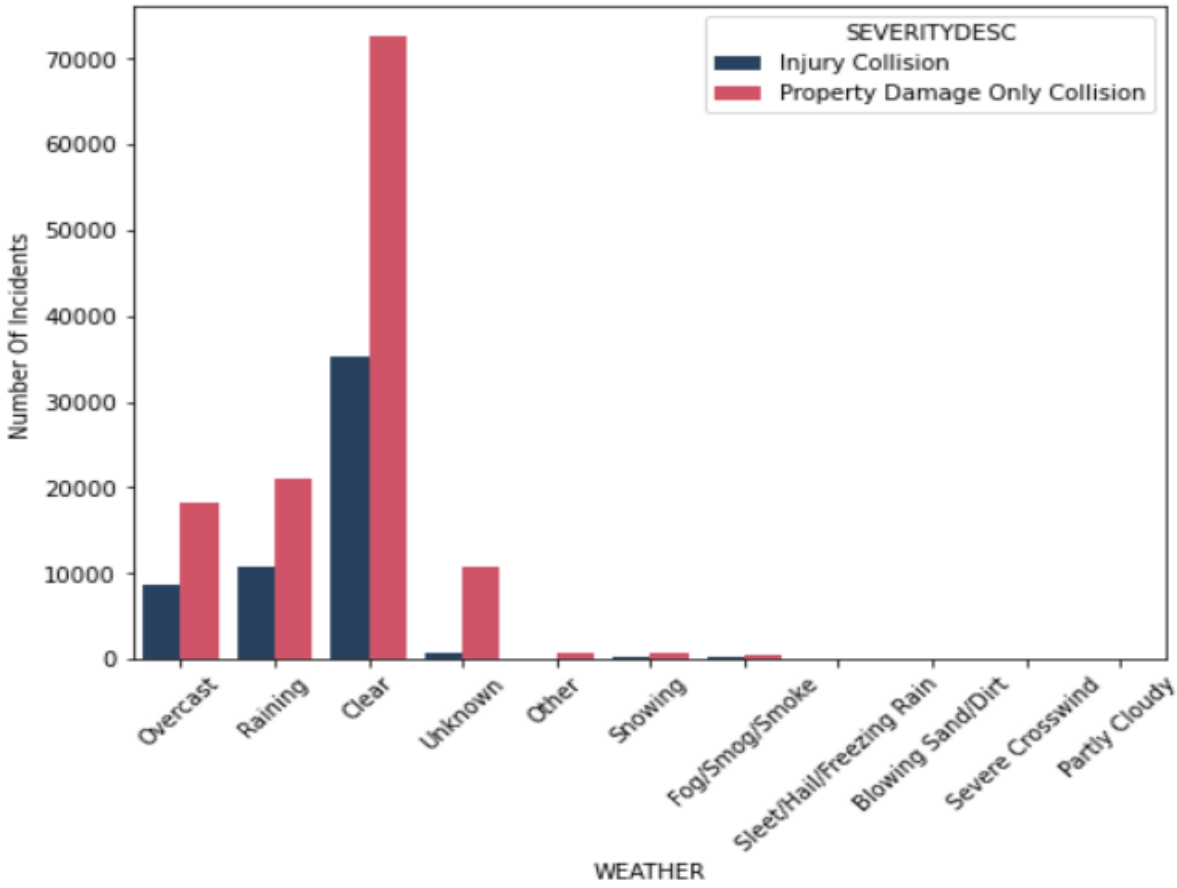
Relationship between Collision Type and Severity



From the figure above we can observe that most collisions occur with a parked car (45,000 cases) followed by Angles (around 35,000 cases), Rear Ended (around 34,000 cases), Other (around 20,000 cases).

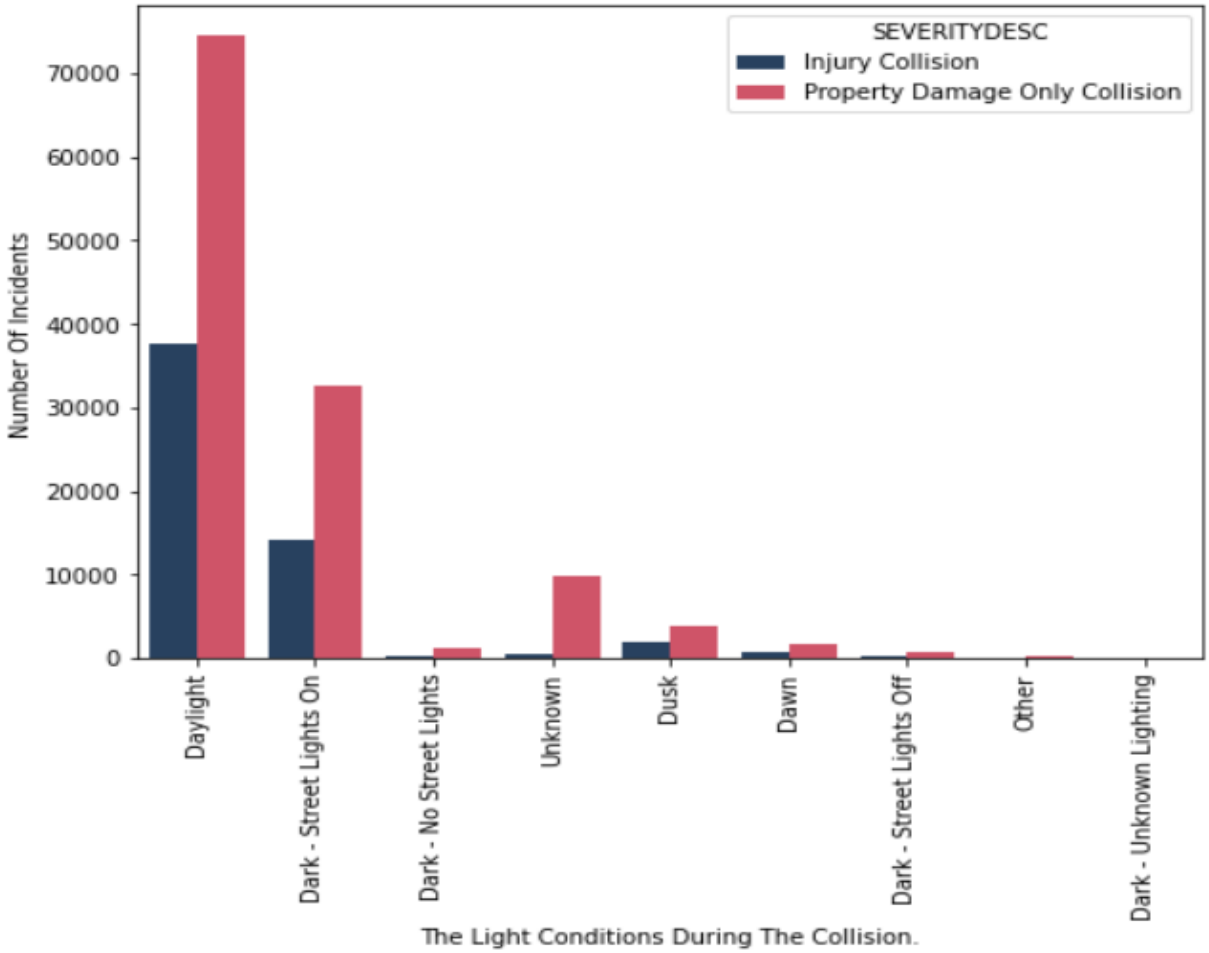
Mostly Property Damage is the Severity type in all these collisions except for the Pedestrian where Injury collision are recorded more (around 7,000 cases).

Relationship between Weather and Severity



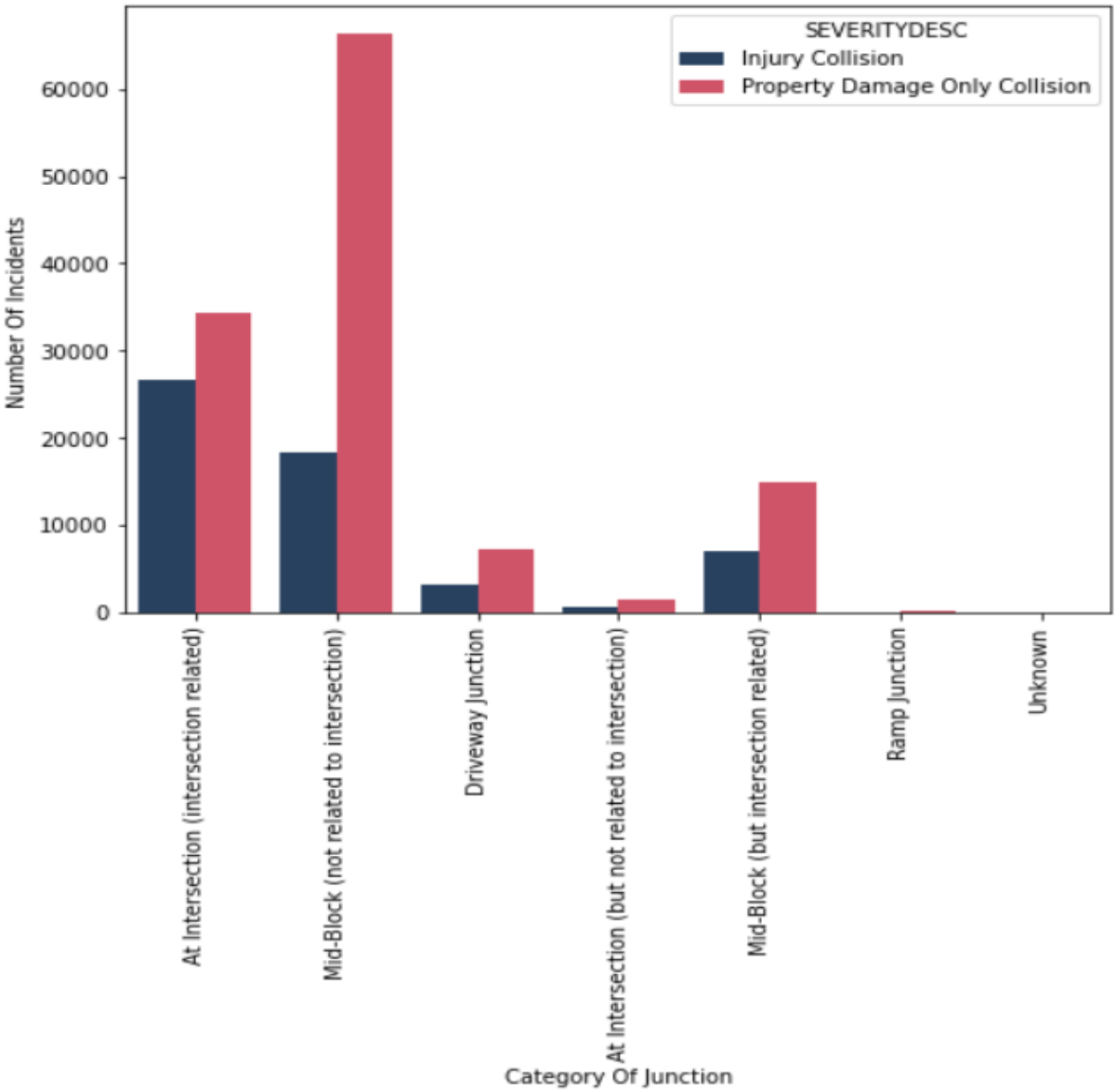
From the figure above we can observe that most incidents are recorded in the clear weather with around 100,000 cases followed by Raining and Overcast. Weather conditions like Snowing and Fog/Smoke, Severe Crosswind does not lead to high number of accidents.

Relationship between Condition of Light during accident and Severity



From the figure above we can observe that most accidents happen during Daylight or in Dark- street lights on, this indicates that bad light conditions do not lead to more accidents.

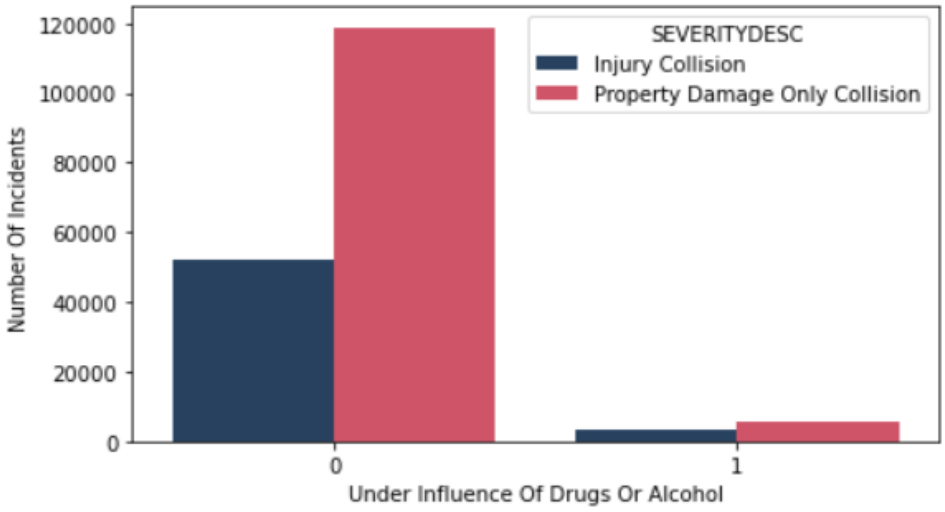
Relationship between Junction Type and Severity



From the figure above we can observe that Mid-block leads to more accidents with more than 80,000 cases recorded, Followed by At Intersection (Intersection Related) with around 56,000 cases recorded.

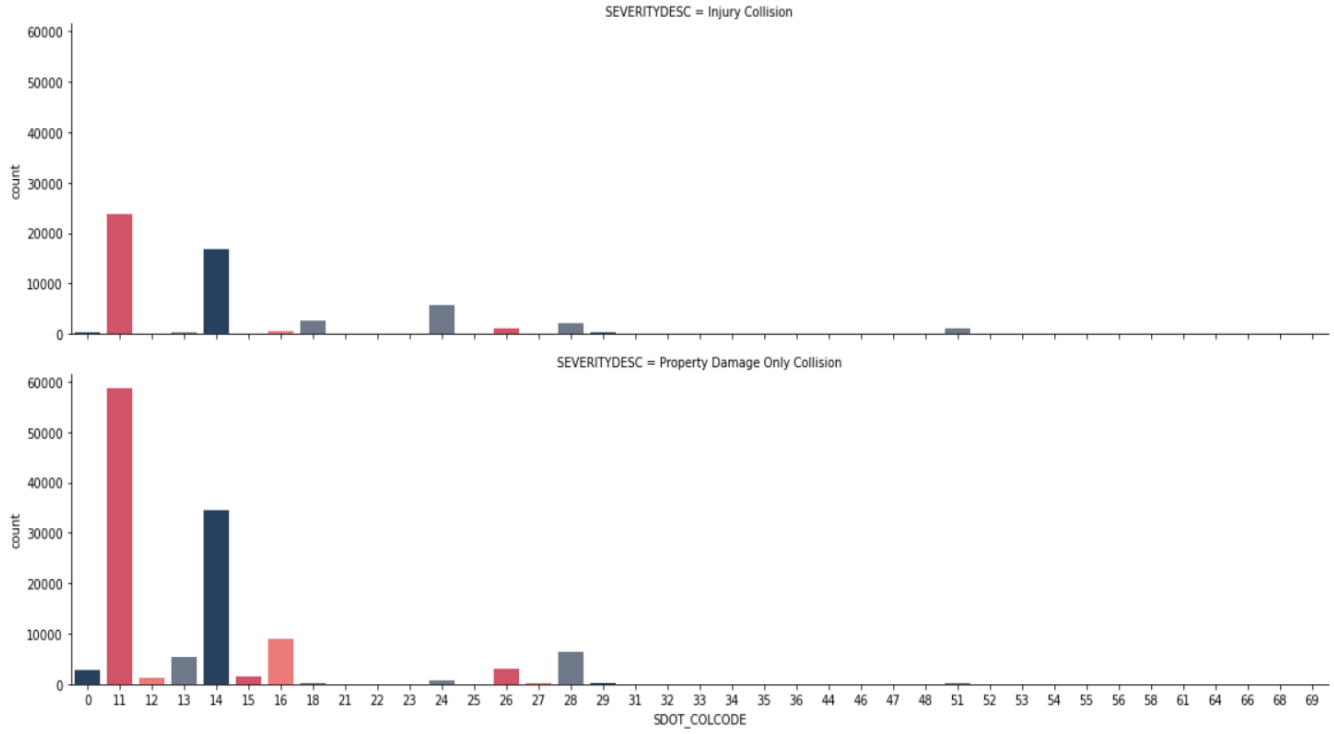
Property damage is most reported at all junction Types.

Relationship between Influence of Drugs and Severity



From the figure above we can observe that Drugs or Alcohol does not lead to more accidents in general.

Relationship between SDOT_COLCODE and Severity

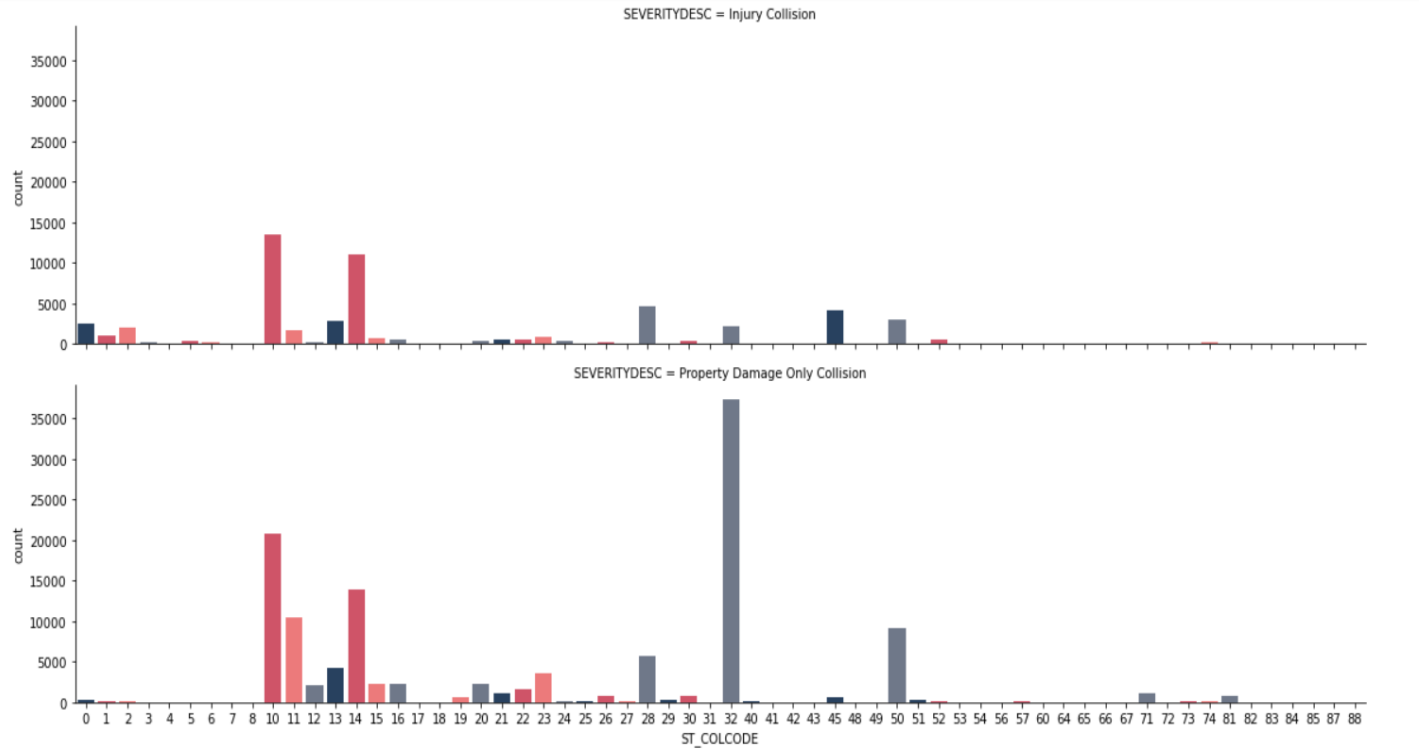


From the figure above we observe the following are the most common collision types-

a) 11 - Motor Vehicle struck Motor Vehicle, Front End at Angle,
b) 14 - Motor Vehicle struck Motor Vehicle, Rear End,
c) 16 - Motor Vehicle struck Motor Vehicle, Left Side Swipe,
d) 24 - Motor Vehicle struck Pedestrian.

11 & 14 are recorded most, whereas 16 causes more property damage & 24 causes more Injuries.

Relationship between ST_COLCODE and Severity

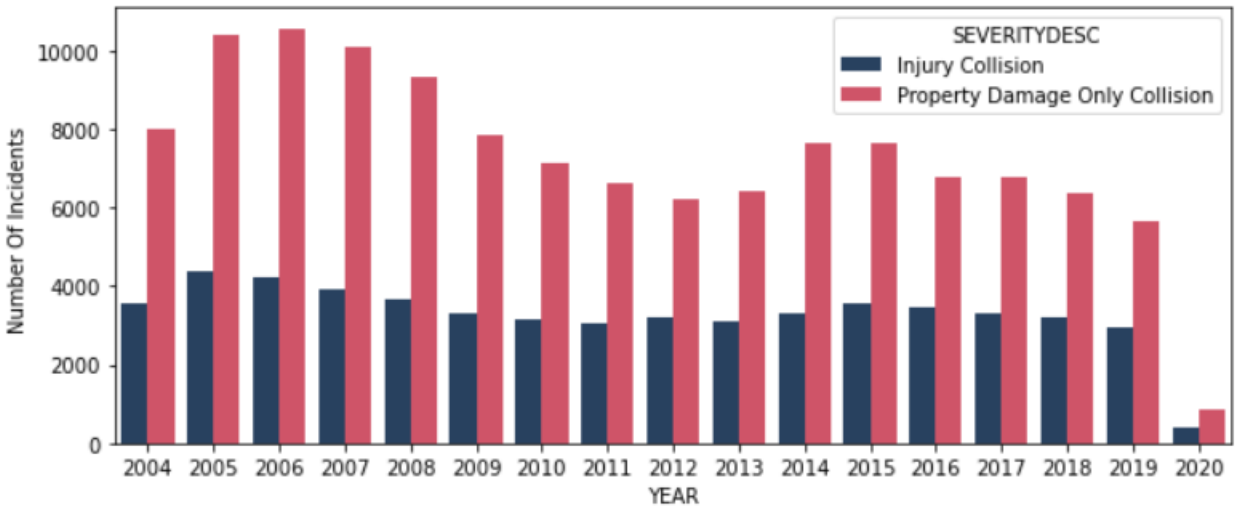


From the figure above we observe the following are the most common collision types-

a) 10 - Entering at Angle,
b) 14 - From same direction – Both going straight – One stopped – Rear End
c) 32 - One Parked – One Moving.

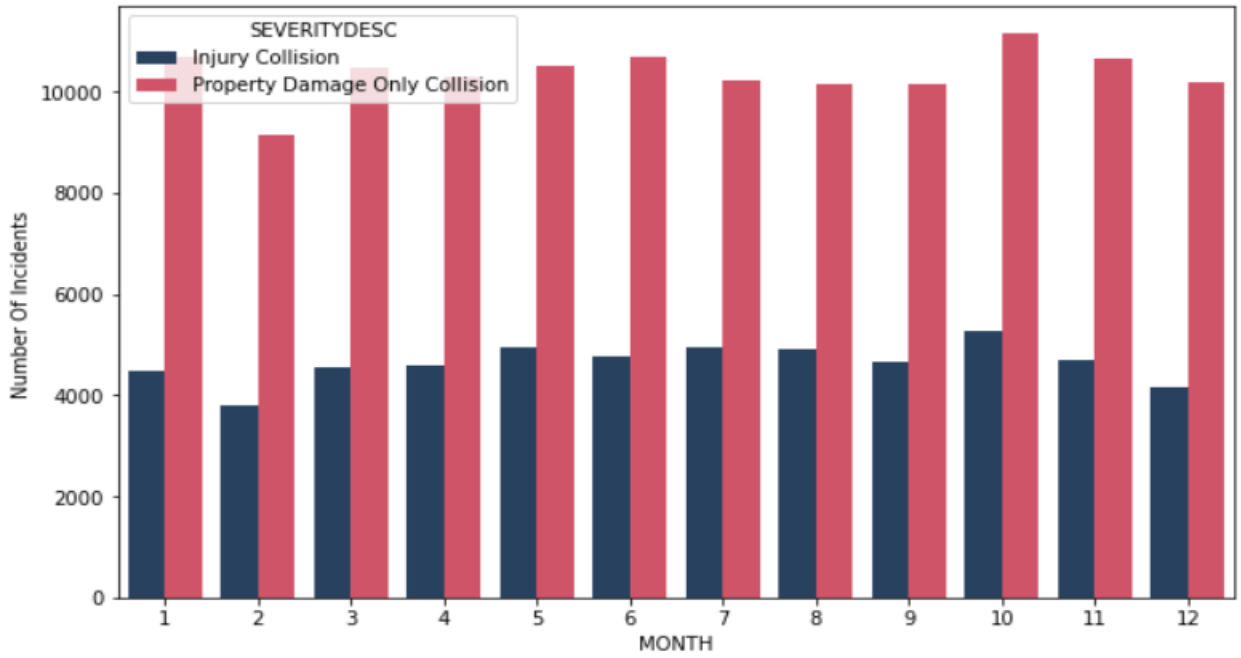
10 & 14 causes more Injuries and are extremely dangerous for human life, whereas 32 causes more Property Damage.

Relationship between Year and Severity



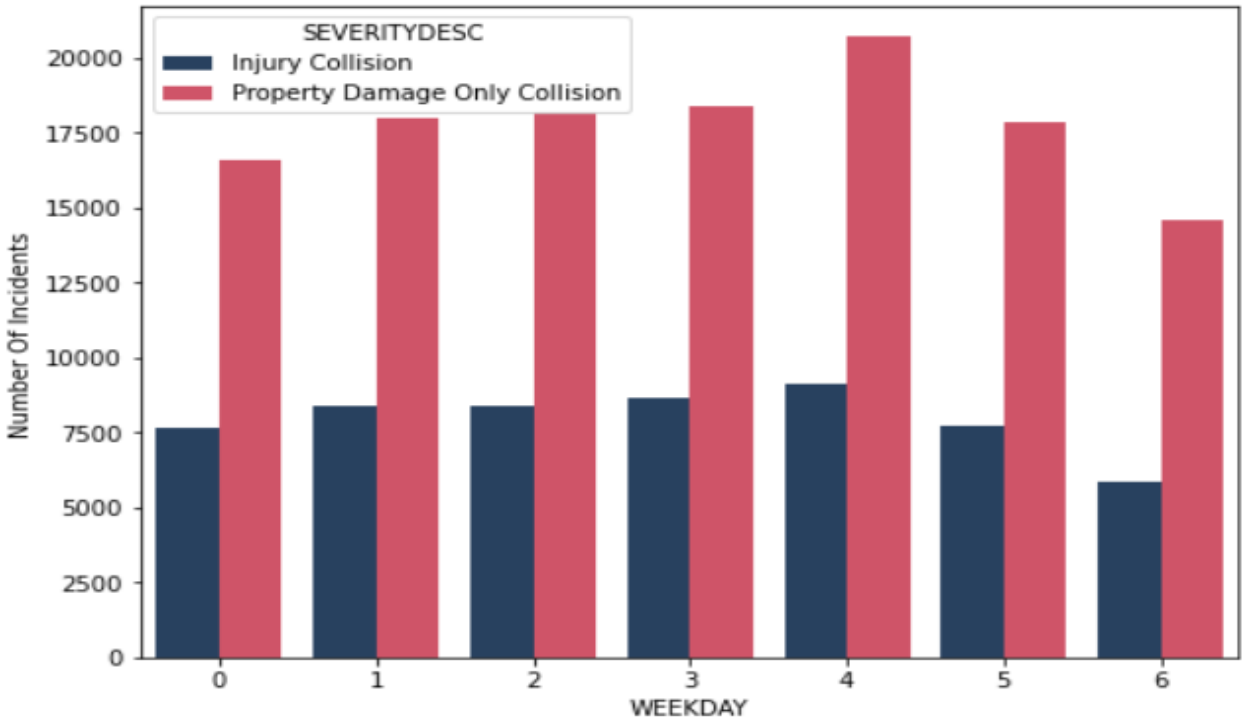
From the figure above can observe that number of accidents are gradually decreasing per year. It may be because of the improved safety features in cars and awareness among public about safety measures.

Relationship between Month and Severity



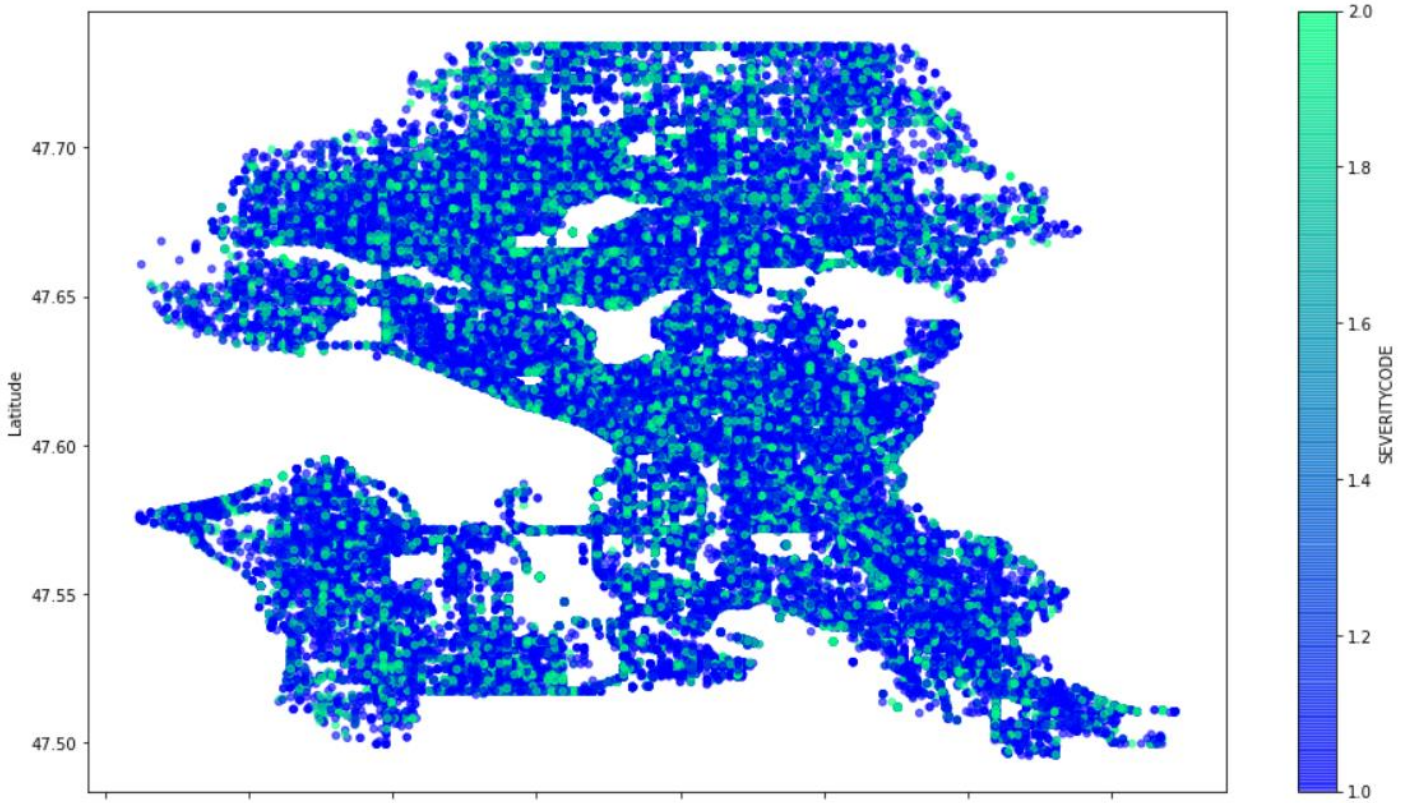
From the figure above we can observe that although the number of accidents is almost the same in every month, but October and January recorded slightly more incidents.

Relationship between Weekday and Severity



From the figure above we can observe that Thursday recorded more incidents than any other day.

Relationship between Latitude and Severity



4. Model Building

4.1 Data Preparation

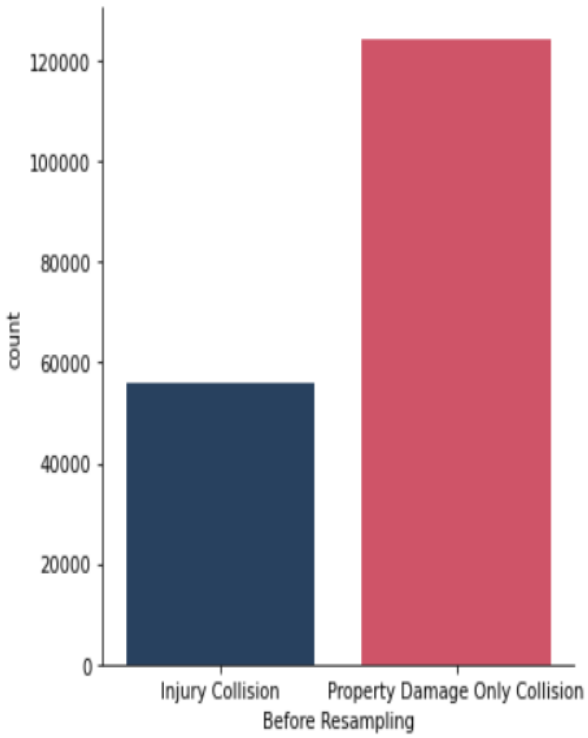
Before building a model, we needed to choose relevant columns that will help us in building a good Machine learning model.

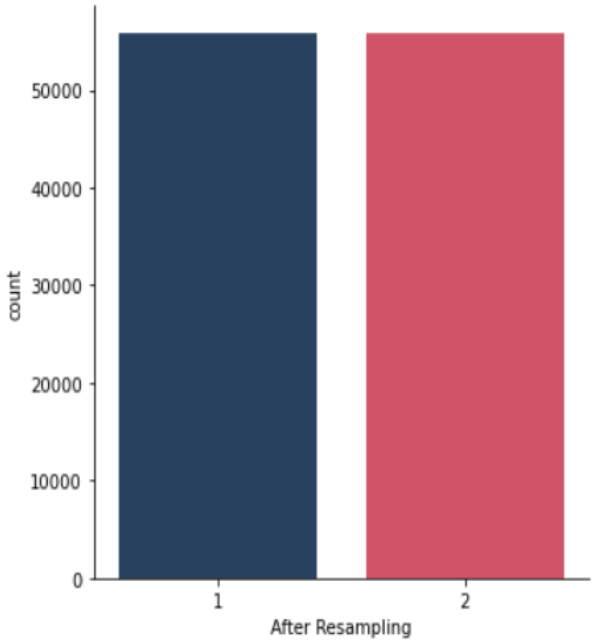
So, we dropped **“SEVERITYDESC”, “INCDATE”, “YEAR”, “MONTH”, “WEEKDAY”, “PEDCOUNT”, “PEDCYLCOUNT”** from the database as these are not useful in building our model.

For building a machine learning model we need to convert categorical variables to continuous variables, for that we used Pandas get_dummies() function.

4.2 Balancing Dataset

What is Data Imbalance? Data Imbalance usually reflects an unequal distribution of classes within a dataset. For example, in a credit card fraud detection dataset, most of the credit card transactions are not fraud and a very few classes are fraud transactions. In our dataset **“SEVERITYCODE”** feature has imbalanced labels, where **“1 for Injury collision”** has 55,809 values and **“2 for Property damage”** has 124258 values. This can create a biased ML model.





To tackle this situation, we used **imblearn library**. In that we used NearMiss and SMOTE one by one to balance the dataset and chose NearMiss as it gave more accuracy to the model. After resampling we got 55,809 values in both the labels.

4.3 Modeling

The problem at hand was of classification, hence We chose two famous classification algorithm's, Logistic Regression and Decision tree to make a ML model.

4.3.1 Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.).

The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.

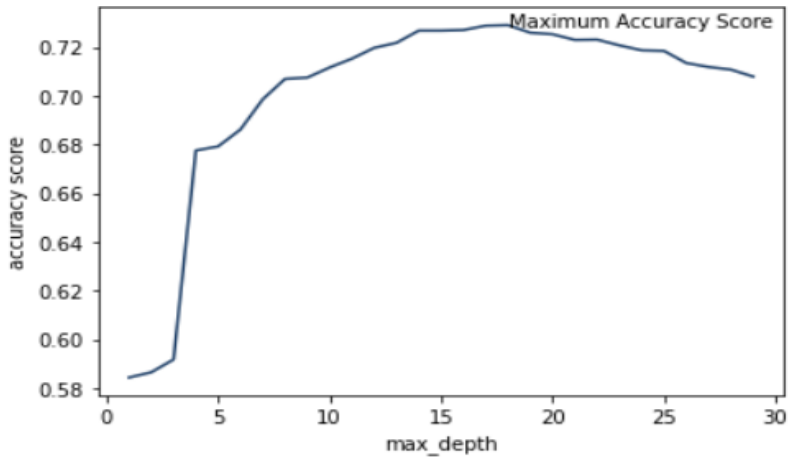
I used the Logistic Regression on training set and performed predictions on the test set. And then compared the true value with the predicted value.

4.3.2 Decision Tree

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute as shown in the above figure. This process is then repeated for the subtree rooted at the new node.

I applied Decision tree algorithm to the training set and used criterion as “entropy” and max depth as 18. To find the max depth I applied for loop to get results of the depth that gets us the maximum accuracy.



Max Depth of 18 shows better accuracy

4.4 Evaluation

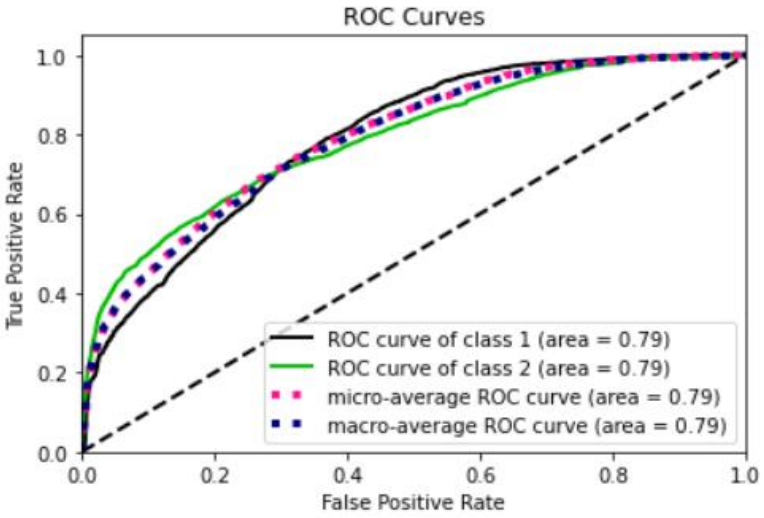
To evaluate the accuracy of ML model, I used different accuracy scores of **Sklearn library**.

4.4.1 Logistic Regression Performance

Below are the different accuracy scores of Logistic Regression Model.

Logistic Regression's Accuracy: 0.7089529952816103
Logistic Regression's LogLoss: 0.542224535278289
Logistic Regression's F1-Score: 0.7082943087673719
Average_neg_mean_absolute_error: -0.29749654430962985

ROC CURVE

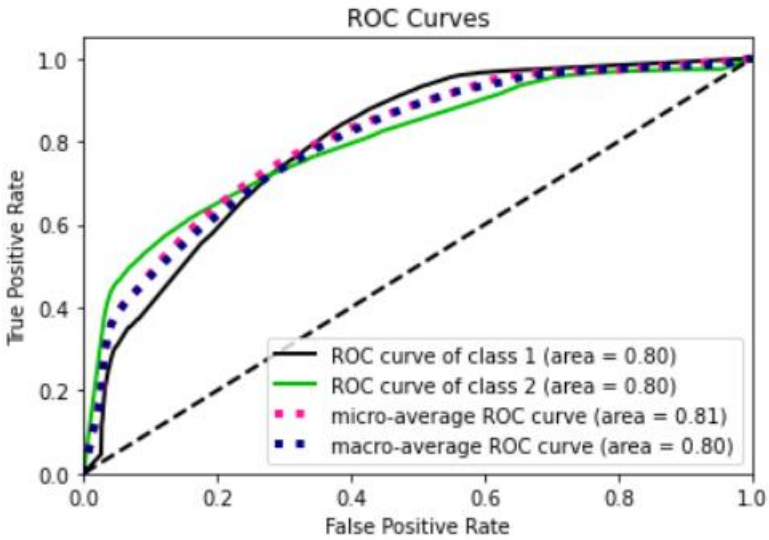


4.4.2 Decision Tree Performance

Below are the different accuracy scores of Decision Tree Classifier.

DecisionTree's Accuracy: 0.7281251866451651
DecisionTree's F1-Score: 0.7249841884639812
Average_neg_mean_absolute_error: -0.2799109199815697

ROC CURVE



5. Discussion

Following are the interesting observations me made from the result of Exploratory Data Analysis and Modeling-

- It was observed that more accidents occurred at blocks rather than at intersections. Property damage was more recorded at Blocks, but both had similar number of Injury collisions
- It was observed that most of the accidents took place in daylight and dry weather, bad lighting and snowy weather did not result in more accidents as we expected.
- It was observed that most accidents were recorded on Thursday and least were recorded on Saturday, maybe it was because many people prefer to stay at home on weekends.
- It was observed that most injuries were caused when Motor vehicle struck pedestrian.

- It was also observed that every year the number of accidents is gradually slowing down, it may be because of the improved safety features in cars, new government policies for safety and increased awareness among public, which is a good sign.

6. Conclusion

In this study, I analyzed the relationship between Severity of an accident based on different factors. I identified that most of the accidents recorded had resulted in more property damage than Injuries.
I built Logistic regression and decision tree model to predict the severity of an accident. Both models performed well with over 70% accuracy, but Decision Tree was better among both with 72.8% accuracy. These models can be extremely useful for drivers and different organizations in making useful decisions to reduce the chances of an accident happening. For example, it could help drivers to drive more carefully if there are more chances of incurring an Injury, or traffic managing department can use this model to locate & manage the high accident areas. Overall, this model can be particularly useful for various organizations and individuals in saving human lives and reducing property damage.