Applied Data Science Capstone

Data Analysis of Car Collision Dataset

Introduction

The total number of personal vehicles in Seattle city is estimated to be over 400,000 vehicles. The car population is estimated to have been doubled from the year 2010. With the increase in car ownership there has been an increase in car accidents as well. According to the National Highway Traffic Safety Administration of the USA suggests that the economical and societal harm from car accidents cost up to $871 billion in a single year.

Business Problem

This capstone project aims to predict how severity of accidents can be reduced based on some factors. Correlation between various factors and number of accidents is analyzed using data visualization.

And finally, KNN and Logistic Regression algorithms are used to predict the probability of occurrence of future car collisions.

Data Feature Selection

A new data frame (df) is created with the selected attributes which would be used for this Analysis. The selected attributes are:

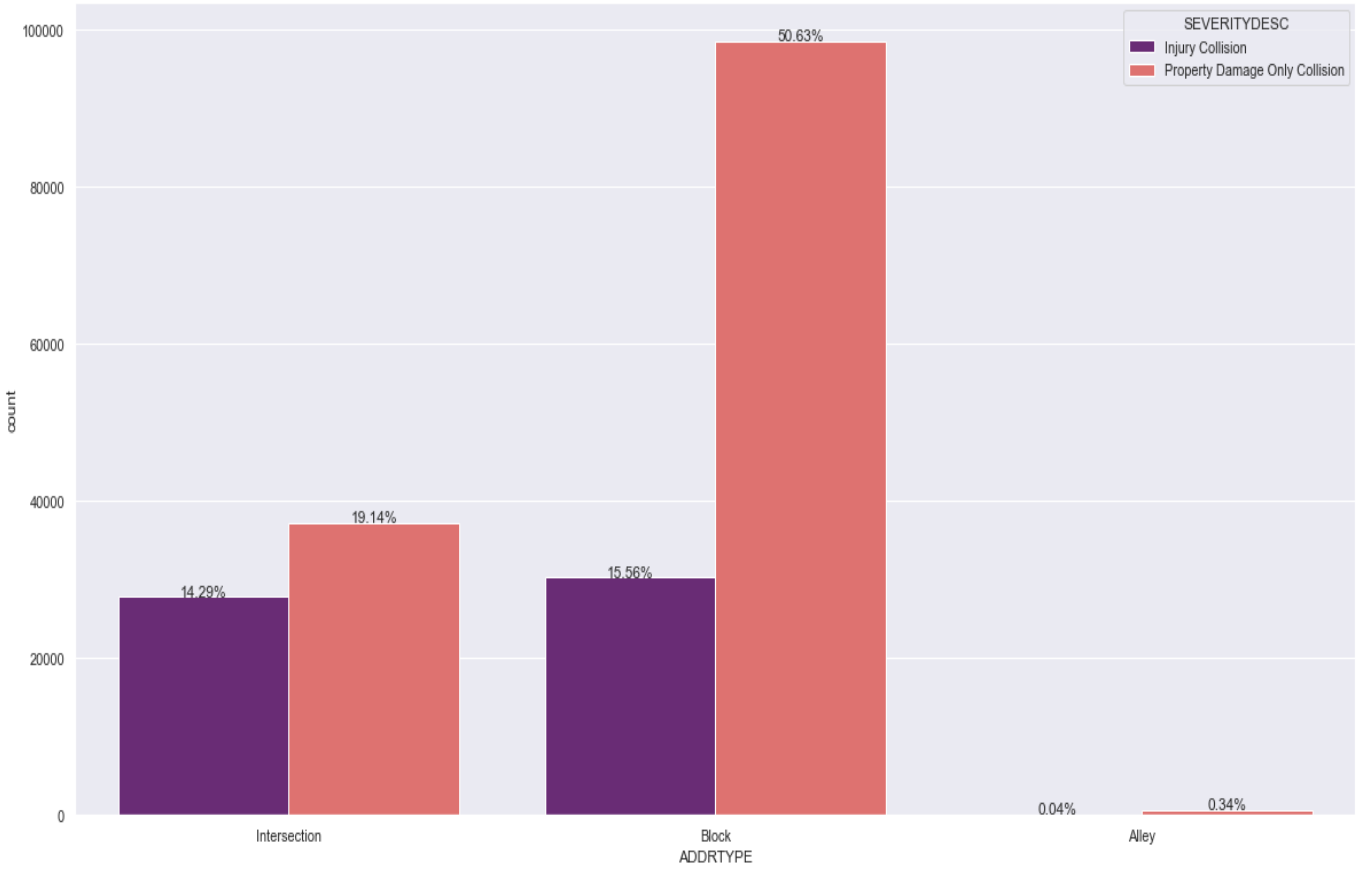
SEVERITYCODE, SEVERITYDESC, ADDRTYPE, INCDATE, INCDTTM, JUNCTIONTYPE, INATTENTIONIND, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, SPEEDING, HITPARKEDCAR.

Missing values are identified in the data frame and are replaced by the value with most frequency.

Binning the data and transforming continuous numerical variables into discrete categorical "bins" for grouped analysis.

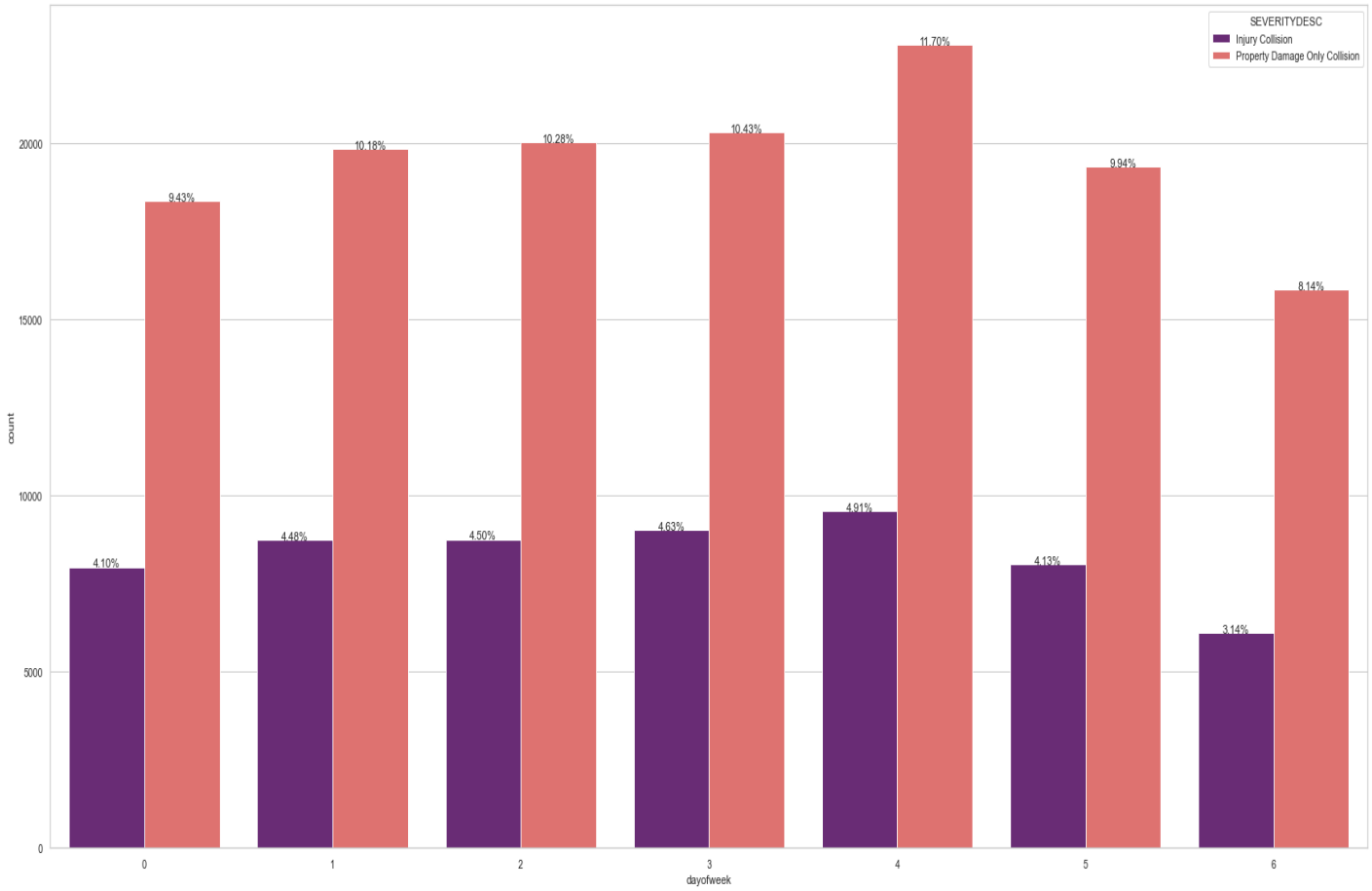
Data Visualization:

1. Car Collisions at Each Type of Address Mentioned



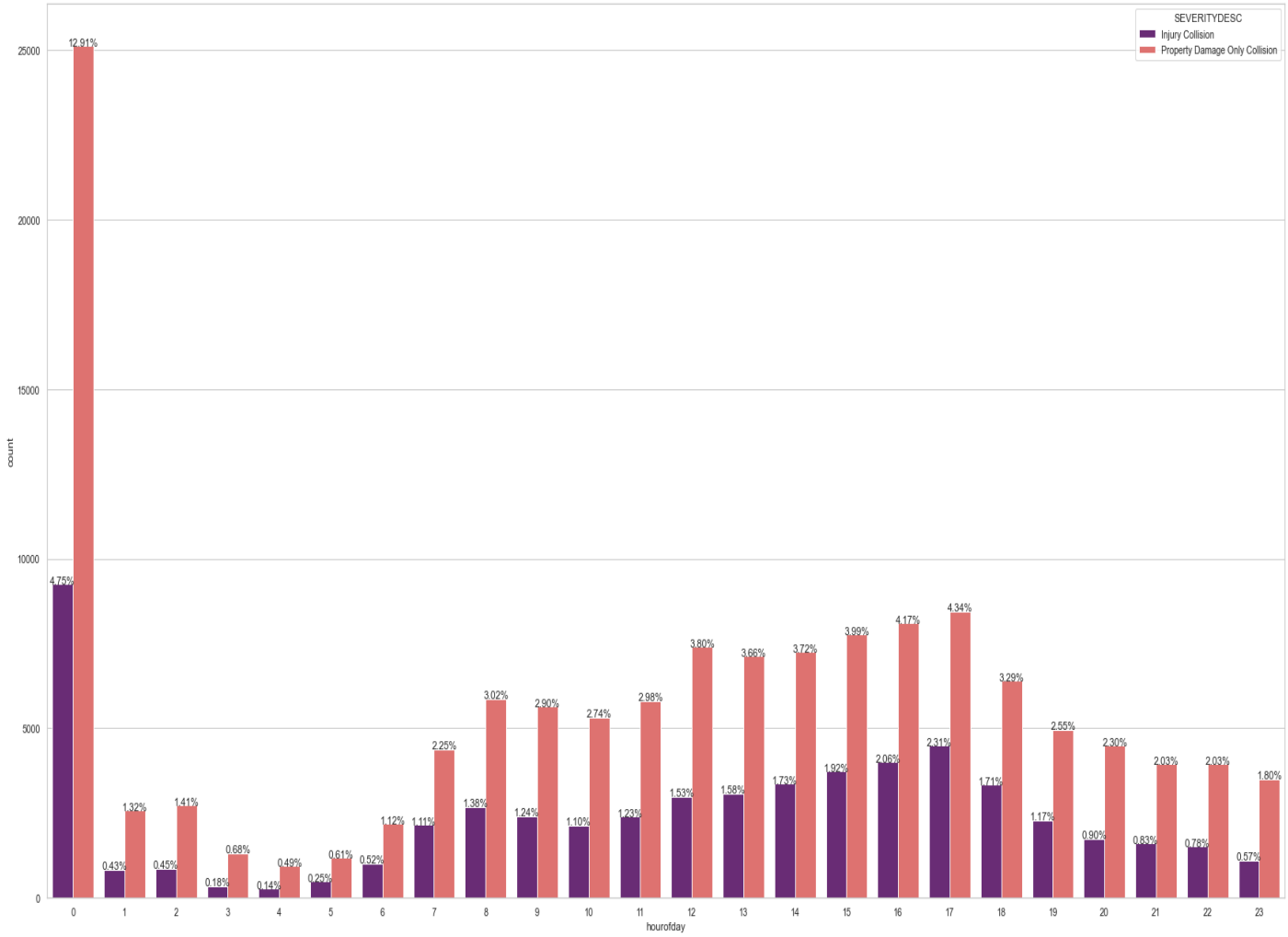
Car collisions at Blocks is the highest followed by collisions at Intersections and lastly at Alleys. Interesting point to be noted here is that of the total collisions at Intersections, a large number of accidents result in Injury when it is compared proportionally with collisions at Blocks.

2. Car Collisions by Day of Week (Monday-Sunday is 0-6)



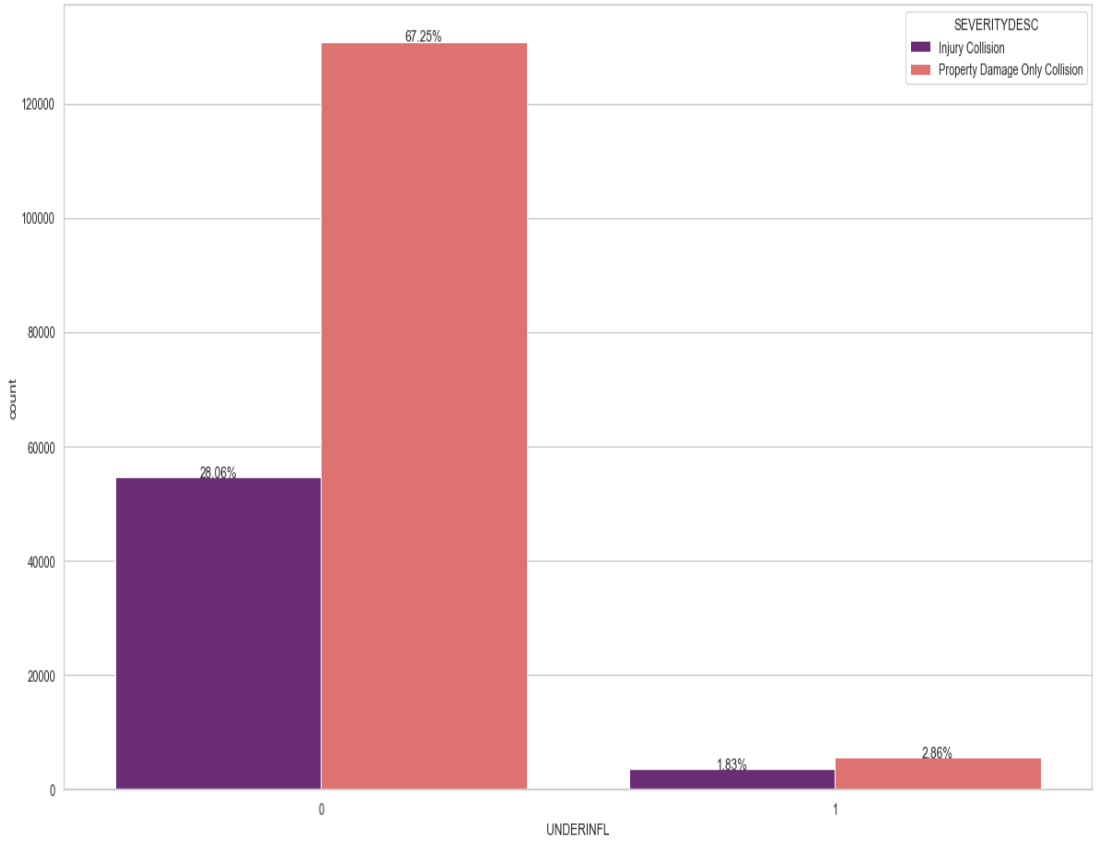
The highest number of car collisions occur on Fridays.

3.Car Collisions by Hour of the Day



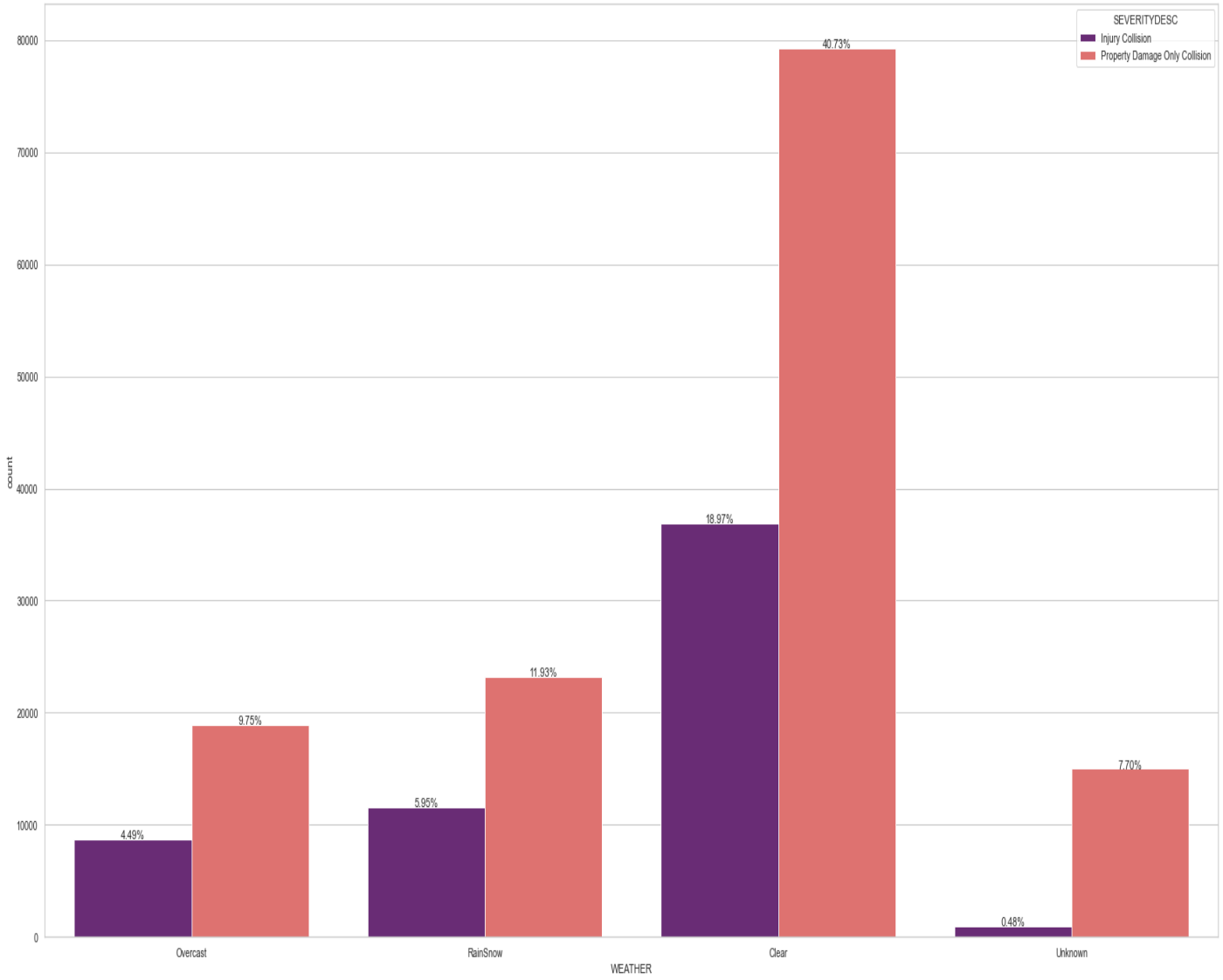
The highest number of collisions by far occur at 12am.

4. When Driver was under the Influence of Drugs or Alcohol (0 is No, 1 is Yes)



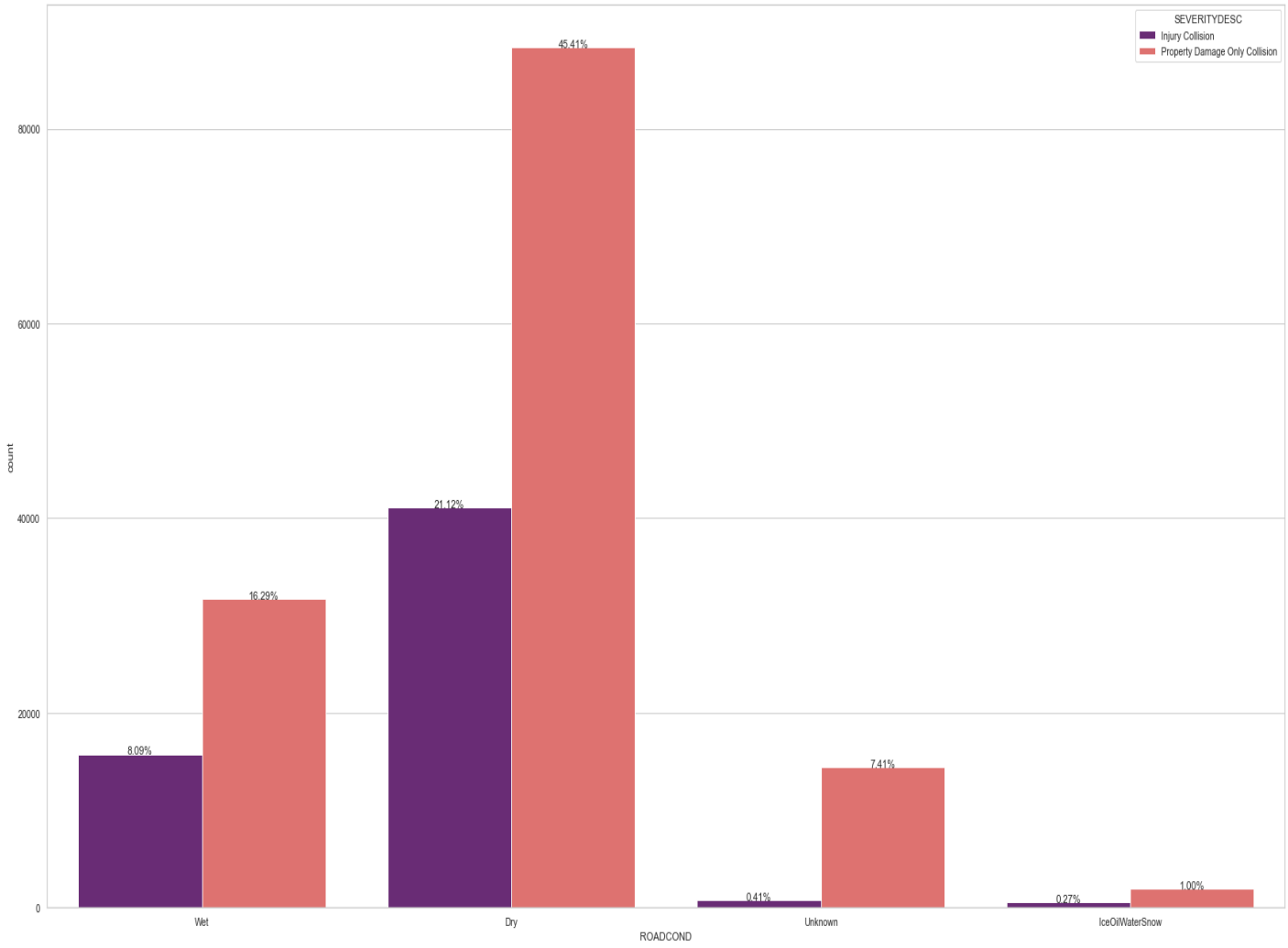
Most of the car collisions occur when the driver is not under the influence of any drug or alcohol but it should be observer that when the driver is under the influence of any drug or alcohol the percentage of that collision to cause injury is high proportionally.

5.Car Collisions by Weather Conditions



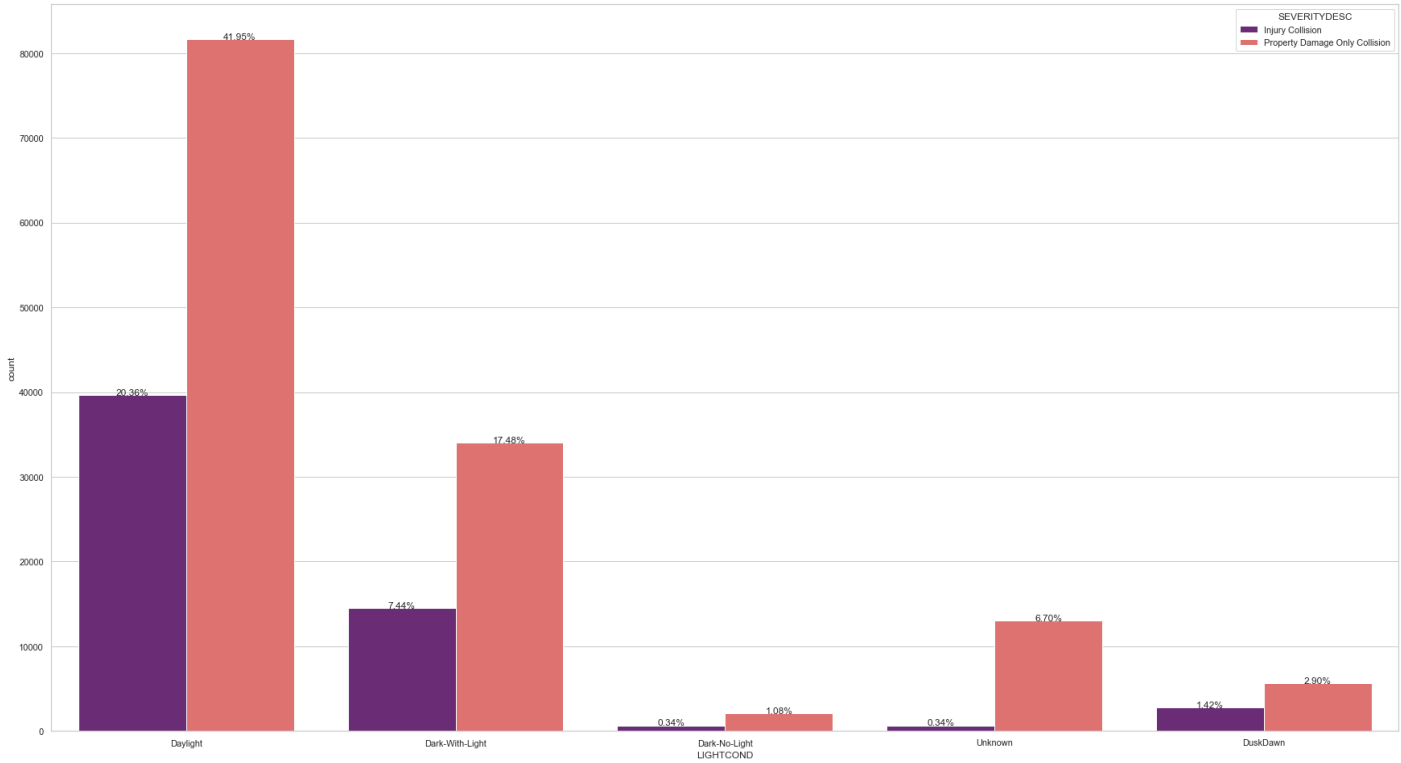
Most of the car collisions occur on clear days but it should be noted that more collisions cause injury on when there is an overcast or rain/snow.

6.Car Collisions by Road Conditions



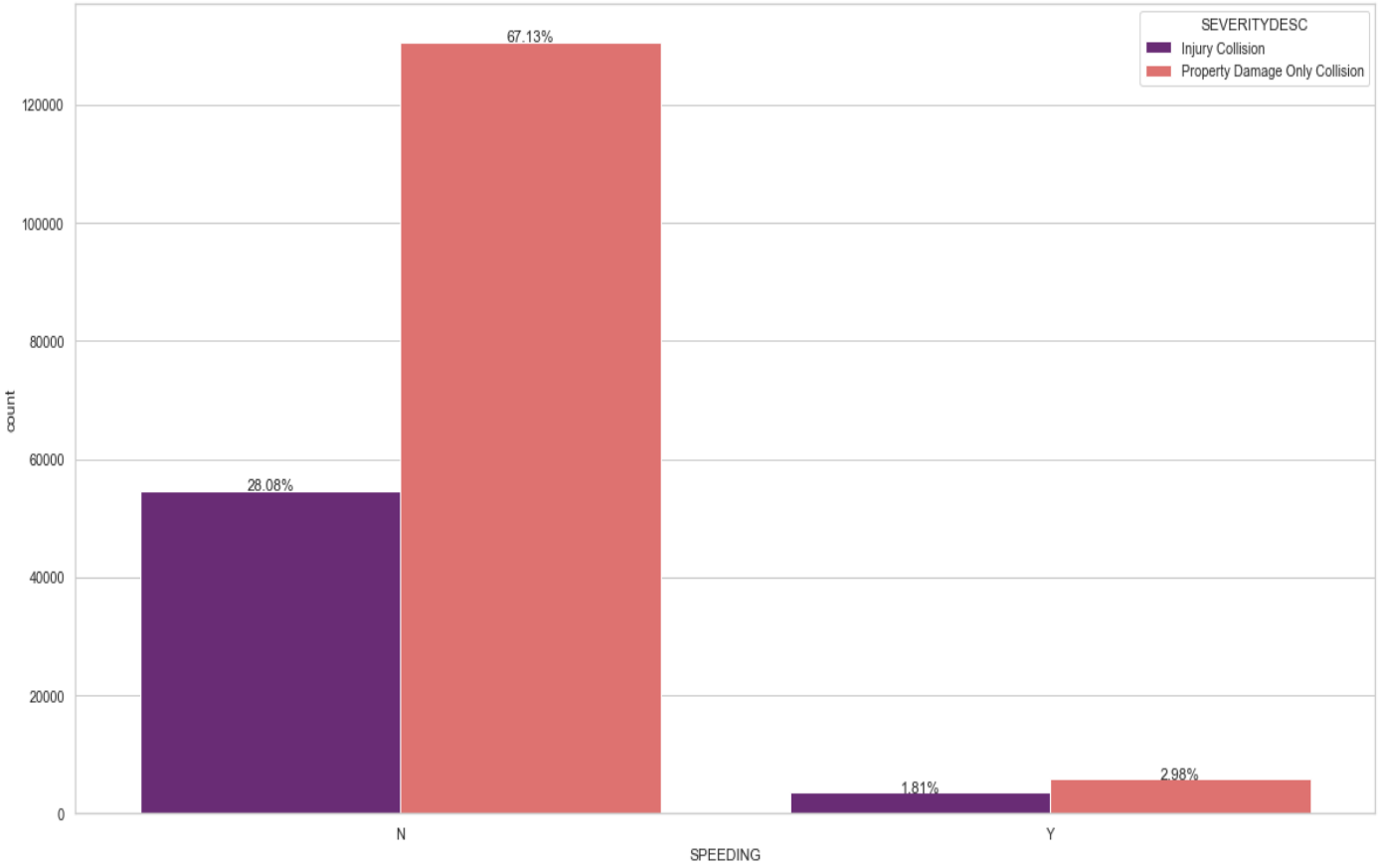
Most of the car collisions occur on dry roads followed by wet and lastly Ice/Snow roads.

7.Car Collisions by Light Conditions



Most of the car collisions occur when there is day light.

8.Car Collisions when Speeding or Not



Most of the car collisions occur when the driver is Not speeding but it should be noted that in case of collision due to speeding the more collisions results in injury when compared to the driver who was not speeding.

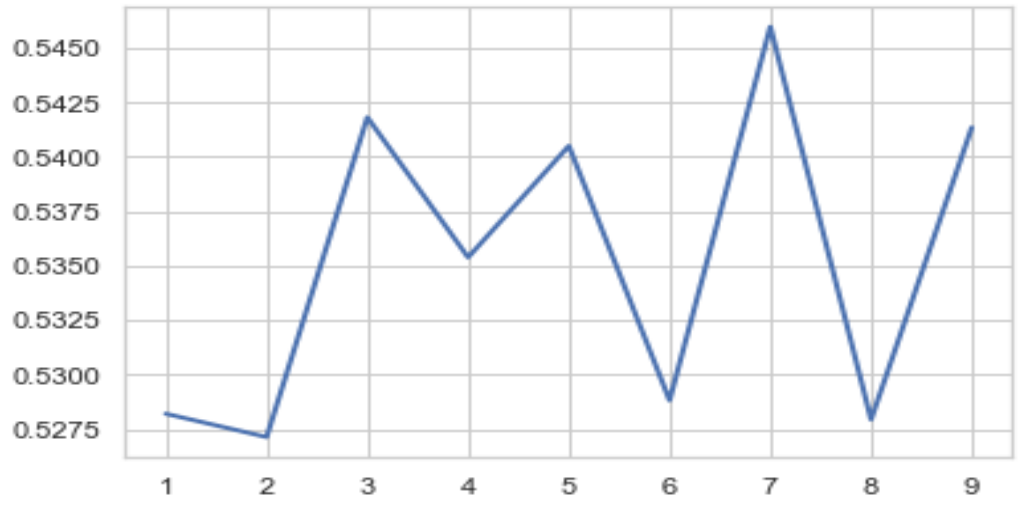
Modeling

Training dataset is 60% of the total data.

Testing dataset is 40% of the total data.

K Nearest Neighbor (KNN)

The best K as shown in the figure, for the model where the highest elbow bend exists is at K=7.

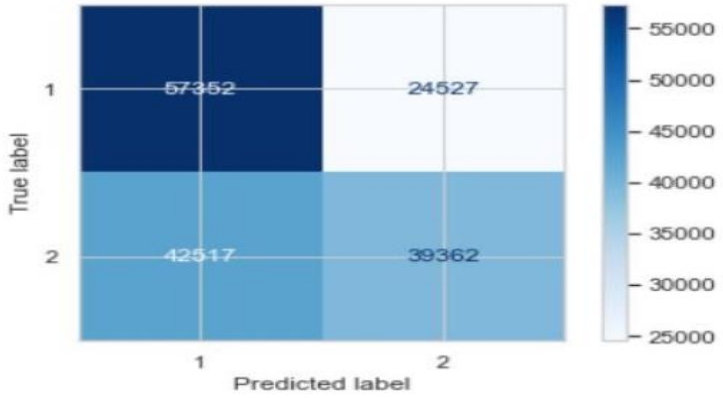


Train Data Confusion Matrix:

```
[[57352 24527]
 [42517 39362]]
              precision    recall  f1-score   support

           1       0.57      0.70      0.63     81879
           2       0.62      0.48      0.54     81879

    accuracy                           0.59    163758
   macro avg       0.60      0.59      0.59    163758
weighted avg       0.60      0.59      0.59    163758
```

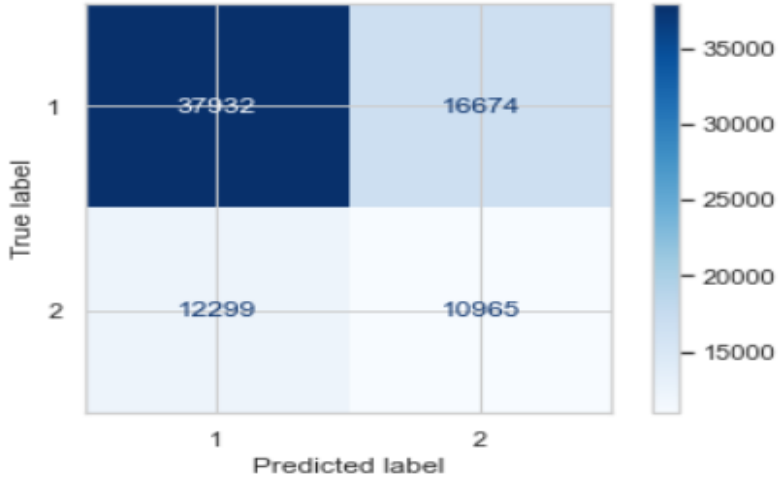


Test dataset Confusion Matrix:

```
[[37932 16674]
 [12299 10965]]
              precision    recall  f1-score   support

           1       0.76      0.69      0.72     54606
           2       0.40      0.47      0.43     23264

    accuracy                           0.63     77870
   macro avg       0.58      0.58      0.58     77870
weighted avg       0.65      0.63      0.64     77870
```



Logistic Regression

The C = 0.01 used for regularization strength.

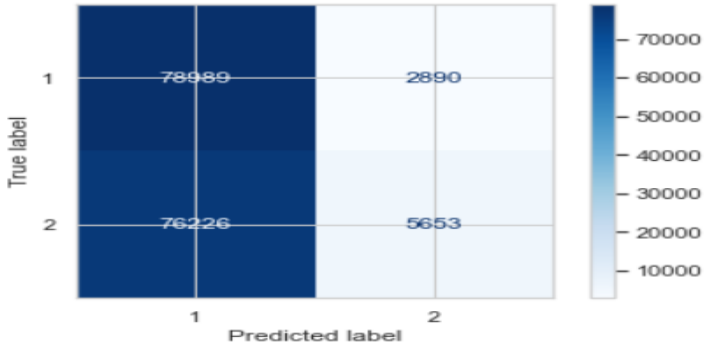
The solver used is liblinear.

Train Data Confusion Matrix:

```
[[78989  2890]
 [76226  5653]]
              precision    recall  f1-score   support

           1       0.51      0.96      0.67     81879
           2       0.66      0.07      0.13     81879

    accuracy                           0.52    163758
   macro avg       0.59      0.52      0.40    163758
weighted avg       0.59      0.52      0.40    163758
```

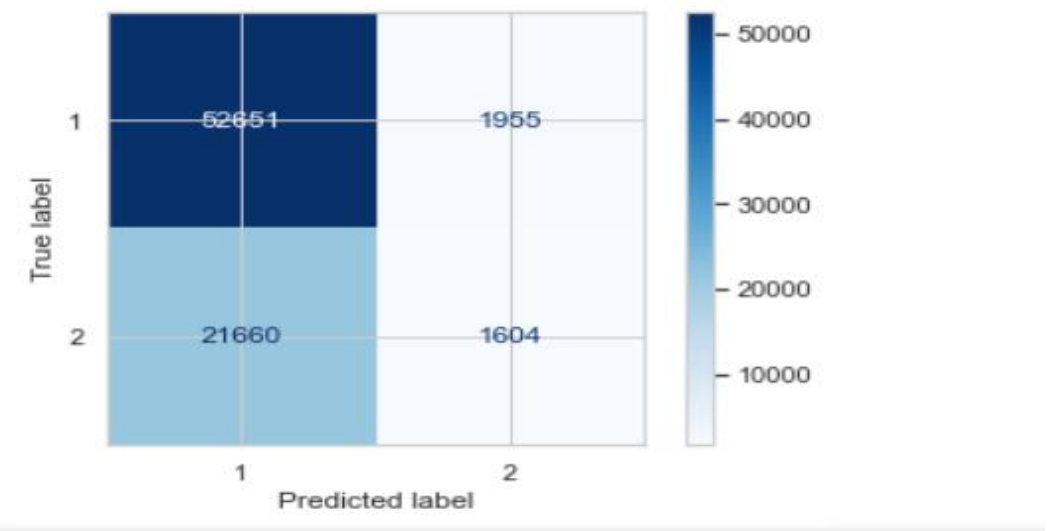


Test Data Confusion Matrix:

```
[[52651  1955]
 [21660  1604]]
              precision    recall  f1-score   support

           1       0.71      0.96      0.82     54606
           2       0.45      0.07      0.12     23264

    accuracy                           0.70     77870
   macro avg       0.58      0.52      0.47     77870
weighted avg       0.63      0.70      0.61     77870
```



Conclusion:

By comparing the two models by their f1 Score, Precision and Recall, we can determine the accuracy of the two models perform individually for each output of the target variable.

It would be better to use Logistic regression model for the prediction of occurrences of car collisions based on the features that have been selected.

Algorithm	1 / 2	f1 Score	Precision	Recall
KNN	1	0.72	0.76	0.69
	2	0.43	0.40	0.47
Logistic Regression	1	0.82	0.71	0.96
	2	0.12	0.45	0.07