# PREDICTING THE SEVERITY OF COLLISIONS

Jessica Teng

---

# INTRODUCTION

## Background:

This is a Capstone Project for Coursera's IBM Data Science Professional Certificate. The project utilizes the provided shared dataset to analyze and predict the severity of collisions/car accidents occured in Seattle, USA.

## Goal/Problem:

Car accidents occur on a daily basis, and the severity of the accident can range from no injury to fatality. Serious injuries and the loss of precious family members from car accidents can negatively impact an individual's and the family members' lives. Some injuries can also lead individuals blind, paralyzed, or amputated. With the number of accidents and deaths rising, it is imperative to improve road safety and reduce the number of occurrences by developing a prediction model to predict car accidents severity.

## Audience:

This is mainly targeted towards residents in Seattle. The presented information is especially important to commuters, drivers, health professionals, police officers, EMTs, and pedestrians who are directly participating in their own safety and the safety of others. By knowing the relationships between special road/weather conditions along with other factors and their likelihood of causing car accidents/collisions can prevent accidents from happening. Decreasing the number of accidents can also decrease the chances of unnecessary traffic jams and the cost of innocent lives.

---

# DATA DESCRIPTION

The dataset contains 194674 rows and 38 columns with 37 attributes. The collisions and car accidents in the dataset were all provided by the Seattle Police Department and recorded by Traffic Records. **The timeframe of the recorded collisions were from 2004 to the present day with weekly updates.**

The **target variable** is called "SEVERITYCODE" with 0 being unknown, 1 being property damage, 2 being injury, 2b being serious injury, and 3 being fatality.

After cleaning up the data by dropping the irrelevant columns, the following **16 attributes/columns remain:**

- SEVERITYCODE

- ADDRTYPE
- SEVERITYDESC
- COLLISIONTYPE
- PERSONCOUNT
- PEDCOUNT
- PEDCYCLOUNT
- VEHCOUNT
- INCDATE
- INCDTTM
- INATTENTIONIND
- UNDERINFL
- WEATHER
- ROADCOND
- LIGHTCOND
- SPEEDING

The above attributes are directly related to the causes and consequences of the accidents/collisions. This includes when and where the accidents took place. Accidents could be because the driver was speeding, under the influence of alcohol or drugs, and texting or not paying attention. In addition, drivers could also be affected by the different lighting and weather conditions. Furthermore, the attributes note how many vehicles, pedestrians, bicycles, and individuals were involved with each of the accidents.

---

# METHODOLOGY & ANALYSIS

This section is carried out by using several functions to check data and count values, followed by grouping different columns into tables and plotting them for a better visualization for the analysis of trends.

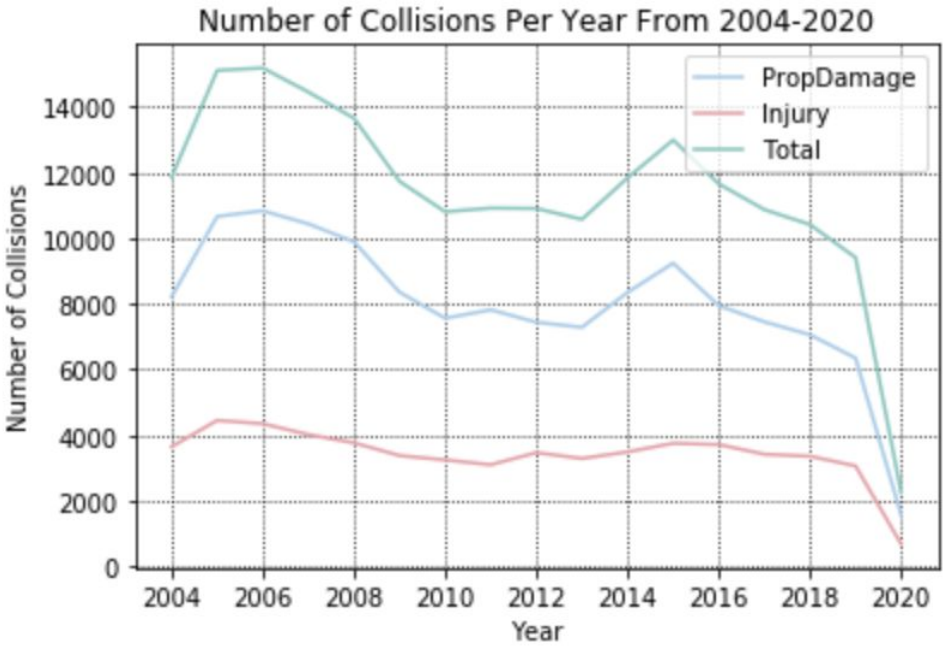
Irrelevant columns were dropped and **.shape()** and **.head()** functions were used to read the data and see the array dimensions. Then, numerous lines of codes were ran to get information on the total number of each type of collisions and the different causes /conditions involved in the collisions.

```
Number of collisions with property damage:  136485
Number of collisions with injury:  58188
Number of collisions with fatality:  0
```

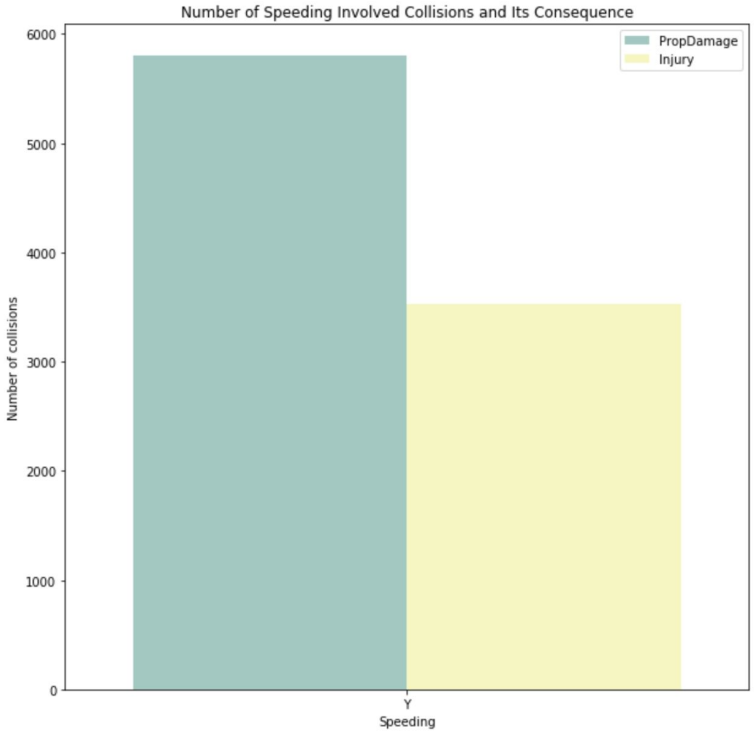
```
Number of collisions where the driver was speeding:  9333
Number of collisions during the day:  116137
Number of collisions during the night with street lights:  48507
Number of collisions during the night with no street lights:  1537
Number of collisions during the night:  50044
Number of collisions during dusk:  5902
Number of collisions during dawn:  2502
Number of collisions where the road was dry:  124510
Number of collisions where the road was wet:  47474
Number of collisions where the driver was under the influence:  9121
Number of collisions where the driver was not paying attention:  29805
```

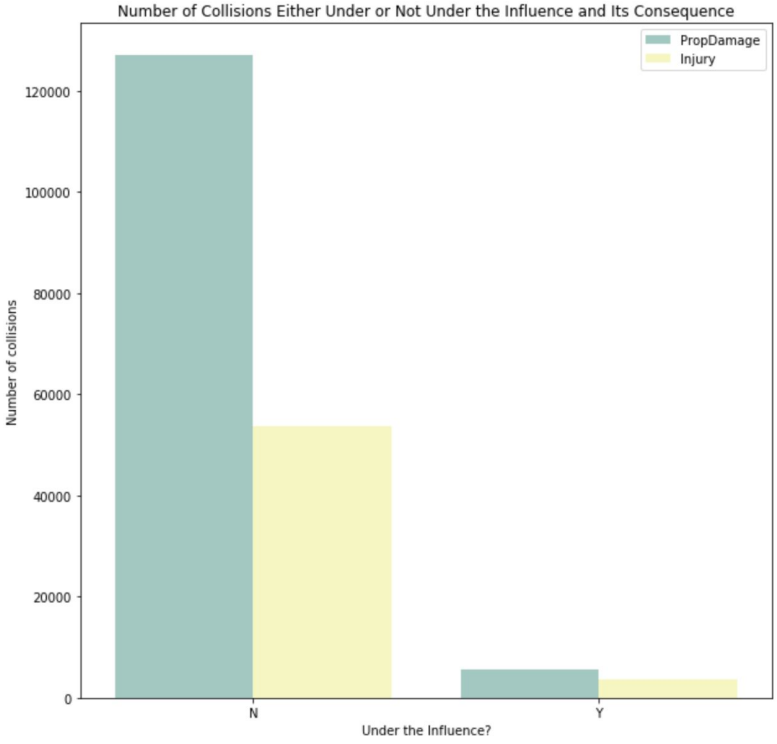
The dates and the times were also put together and converted for easier access. Collisions with respect to the SEVERITYCODE throughout the years from 2004 - 2020 were the first to be grouped and graphed to examine the relationships. A “Total” column was also added to combine for a total number of collisions per year.

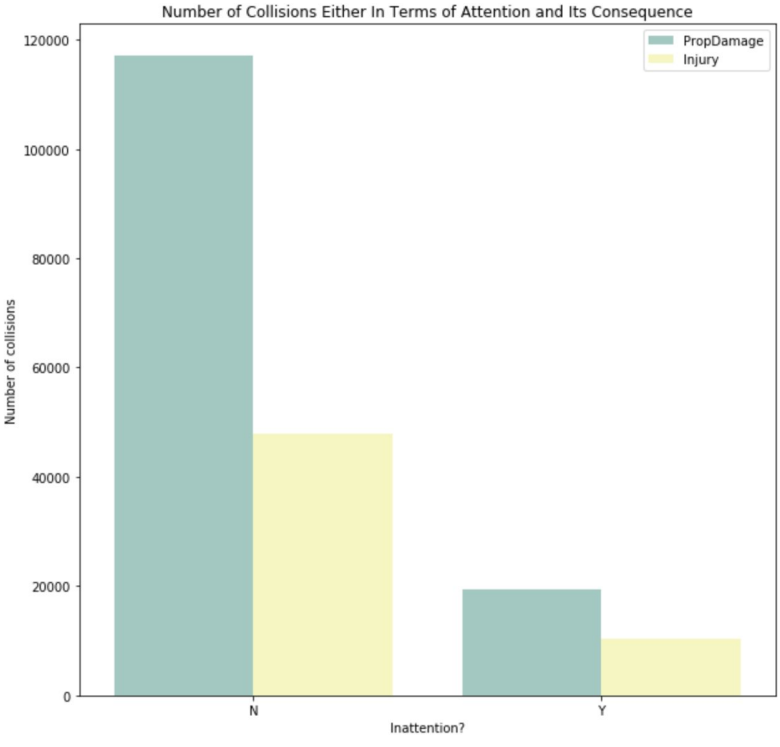
| SEVERITYCODE | PropDamage | Injury | Total |
|---|---|---|---|
| Year | | | |
| 2004 | 8218 | 3647 | 11865 |
| 2005 | 10665 | 4450 | 15115 |
| 2006 | 10838 | 4350 | 15188 |
| 2007 | 10439 | 4017 | 14456 |
| 2008 | 9893 | 3767 | 13660 |
| 2009 | 8356 | 3378 | 11734 |
| 2010 | 7563 | 3245 | 10808 |
| 2011 | 7820 | 3099 | 10919 |
| 2012 | 7440 | 3467 | 10907 |
| 2013 | 7287 | 3290 | 10577 |
| 2014 | 8351 | 3490 | 11841 |
| 2015 | 9243 | 3752 | 12995 |
| 2016 | 7945 | 3714 | 11659 |
| 2017 | 7454 | 3419 | 10873 |
| 2018 | 7061 | 3358 | 10419 |
| 2019 | 6350 | 3062 | 9412 |
| 2020 | 1562 | 683 | 2245 |

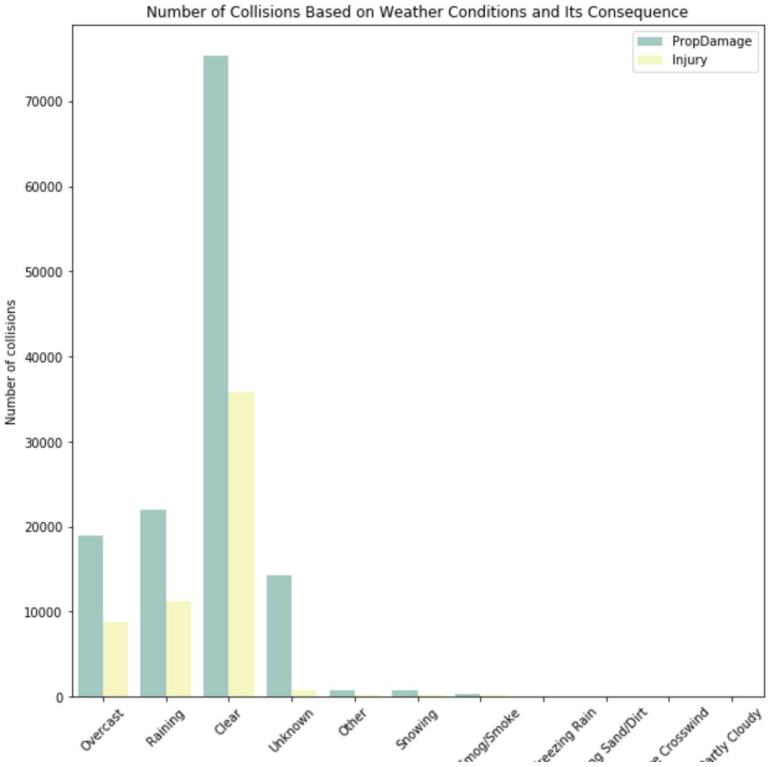


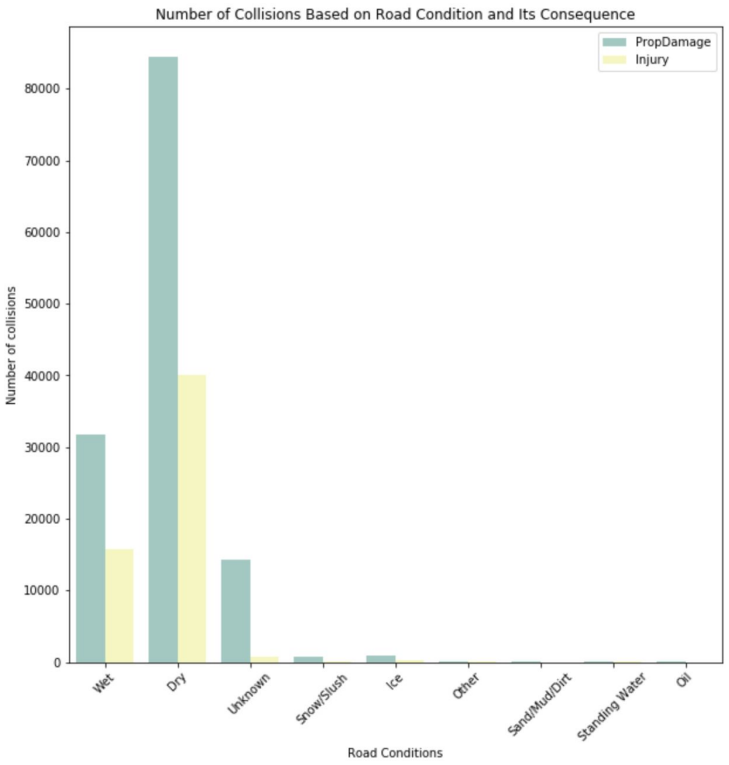

The same concepts and relationships were examined for different potential causes/conditions. For example, collisions where the drivers were under the influence, not paying attention, and speeding. Collisions where the road was wet due to different weather conditions. Collisions where the vision of the drivers was unclear due to different light conditions or time of the day.

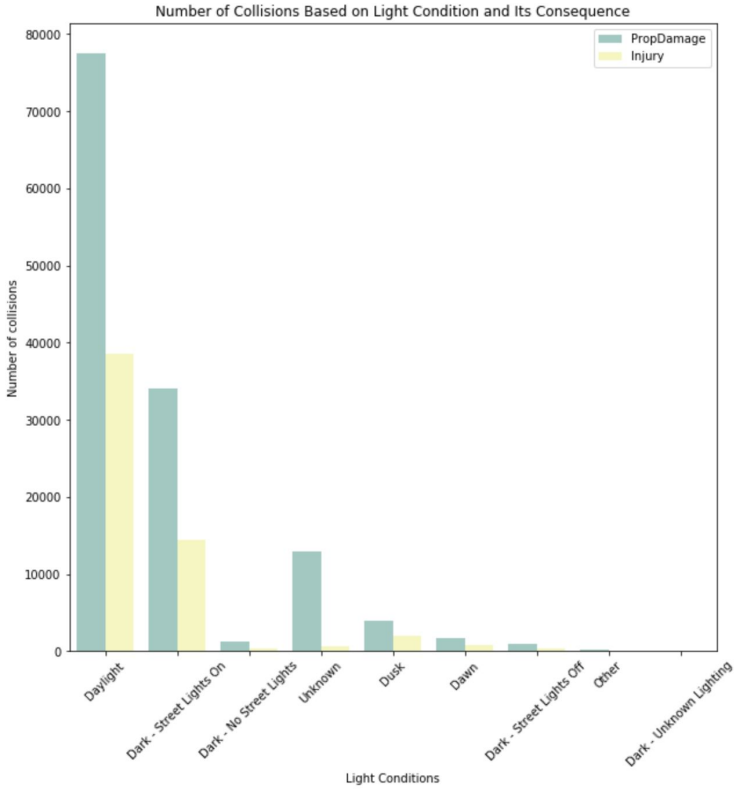

Number of Speeding Involved Collisions and Its Consequence
PropDamage
Injury
Number of collisions
Speeding
Y
0
1000
2000
3000
4000
5000
6000


Number of Collisions Either Under or Not Under the Influence and Its Consequence
PropDamage
Injury
Number of collisions
Under the Influence?
N
Y
0
20000
40000
60000
80000
100000
120000

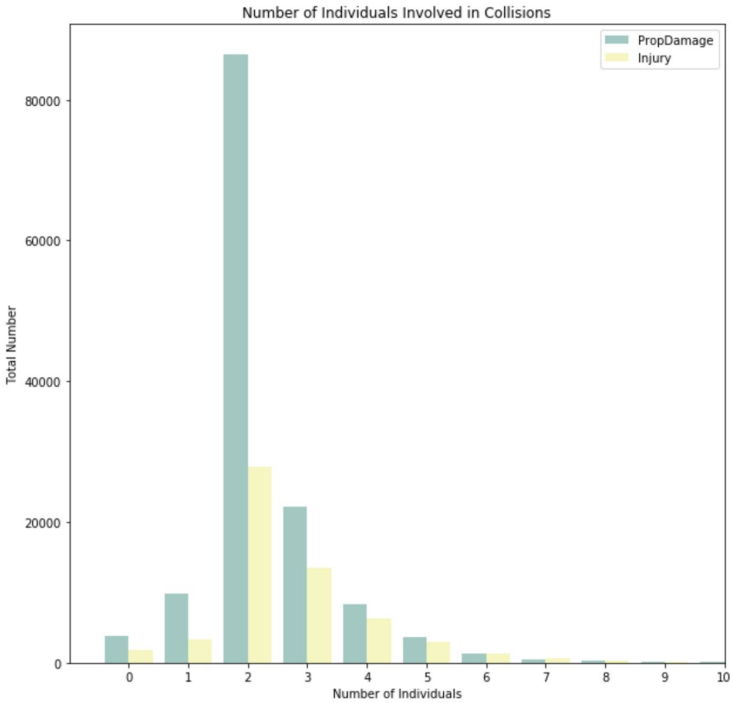

Number of Collisions Either In Terms of Attention and Its Consequence
PropDamage
Injury
Number of collisions
Inattention?
N
Y
0
20000
40000
60000
80000
100000
120000

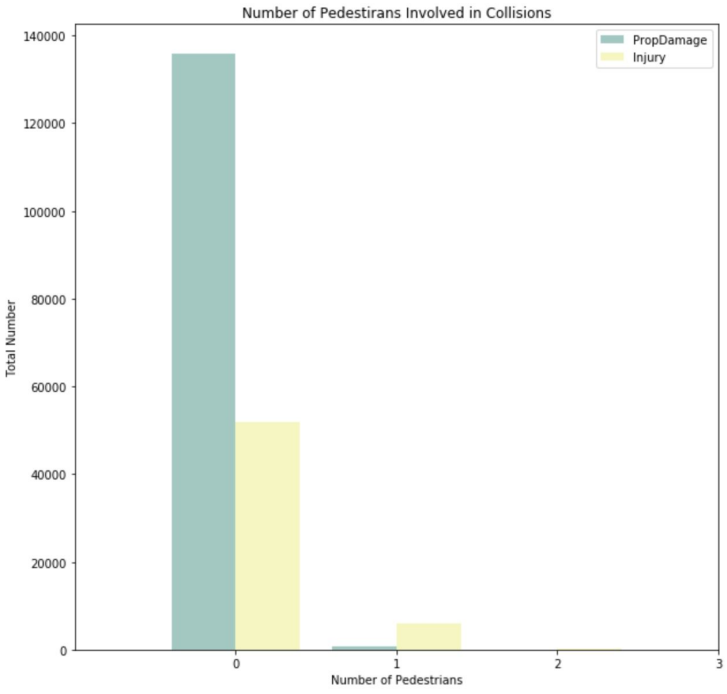

Number of Collisions Based on Weather Conditions and Its Consequence
PropDamage
Injury
Number of collisions
0
10000
20000
30000
40000
50000
60000
70000
Overcast
Raining
Clear
Unknown
Other
Snowing

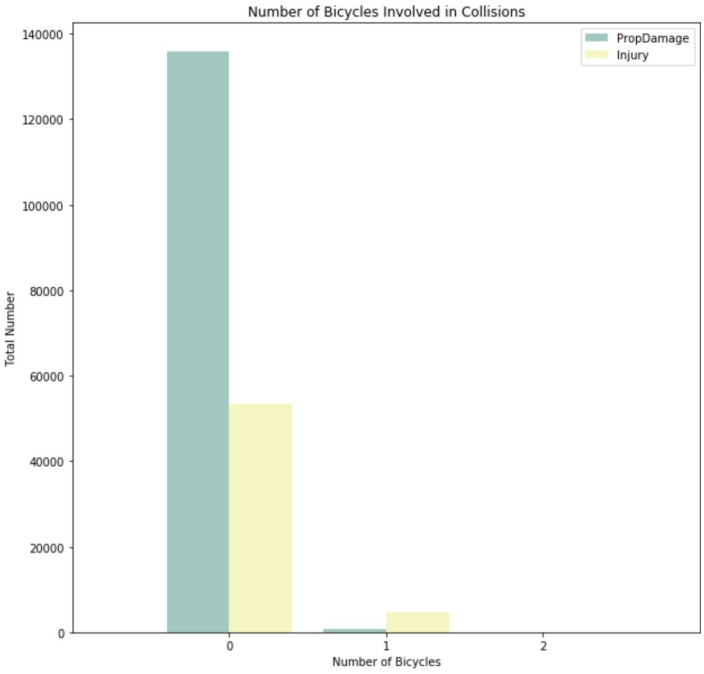


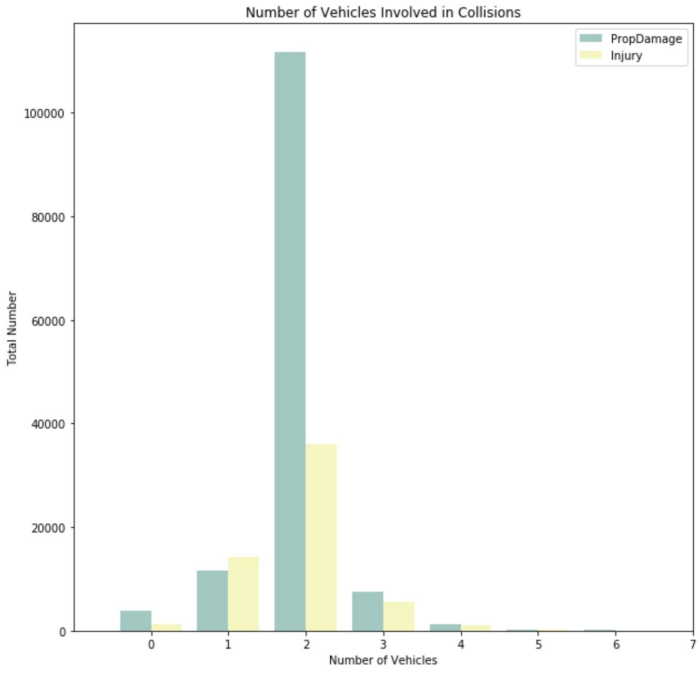




The total number of vehicles, pedestrians, bicycles, and individuals involved in the collisions with respect to SEVERITYCODES were also examined.

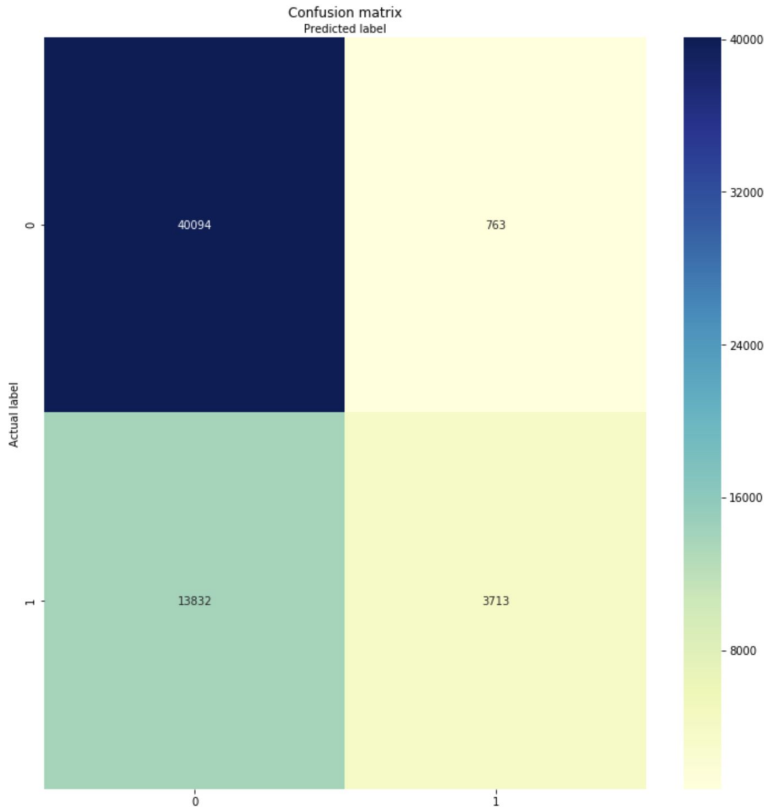





Logistic Regression was the classification algorithm used to predict the severity of car collisions since the data was binary. The processes involved were pre-processing and selection, training testing and splitting the dataset, model development, fitting the model with data, prediction, and the evaluation of the model using a confusion matrix. The confusion matrix is then created using heatmap for better visualization.



---

# RESULTS/DISCUSSION

The model built was used to predict the severity of car collisions. The SEVERITYCODE with 1 being “Property Damage” and 2 being “Injury”. Logistic Regression resulted with a score of about 0.75. From the model and the numerous graphs for visualizations, relationships and trends were examined between the years, different causes, and different conditions during the collision. Most of the collisions happened with property damage, drivers were not under the influence, drivers were not speeding, and drivers were paying attention. In addition, the three most common weather conditions during collisions were when the weather was clear, raining, and overcasting. The road condition was mainly dry when the collisions happened, followed by wet road condition. Most collisions happened during the daylight, followed by at night but with street lights on. Furthermore, collisions most often involved 2 individuals and 2 vehicles with no pedestrians and no bicycles involved.

```
Accuracy:  0.7500941748570255
```

---

# CONCLUSION

The model built was useful in predicting if the collision would result in property damage or injury based on certain conditions (road conditions, weather conditions, driver’s attention...etc.), however, the model can definitely be improved upon to build a model with better accuracy. The model can also improve if there is more relevant and precise information, and if there are less “NaN” or missing values.