Predicting and Mitigating Car Accident Severity

Alyssa Heare

Coursera – IBM DATA SCIENCE- Applied Data Science Capstone

**Abstract**

The following study was conducted to analyze car accident severity. The United States has a high accident rate and the severity of such accidents can be life threating. The data utilized was historic data from the city of Seattle. Seattle is a growing city, which still has a large vehicle travel sector and yet similar infrastructure to other cities across the United States, making it a feasible chose for this study. The study’s data was manipulated and analyzed for indications of what caused such severity and if severity level could be predicted using machine learning tools such as classification methodology to diminish such accidents. This study provided insight as to the causes of higher severity and accident levels and recommended actions that could be taken based on the projections and examined data.

**Predicting and Mitigating Car Accident Severity**

In 2018, the NHTSA has reported there were less than 40,000 traffic related deaths. The NHTSA also reported over 2.7 million accidents where individuals sustained road related injuries! Can there be a way to reduce or avoid such accidents, or reduce the severity of the accidents that are occurring?

In this study I will be analyzing data from the Seattle Department of Transportation. Seattle is one of the fastest growing cities in the United States with its inner metro area population of four million, and the numbers continue to grow causing a fair amount of traffic congestion and accidents. Due to its weather conditions Seattle has been nicknamed “The Rainy City”. The proximity to its western mountain ranges causes Seattle to experience an average of 152 rainy days ranging from typical rainfall to light showers. Later in the year, during the winter months and beginning of the year is when Seattle experiences the majority of its rainfall. Although, heavy snow fall is less common during these months.

This study will hopefully reveal any insight into what can be done to reduce or even potentially eliminate accidents from occurring. It should be noted that during the past few years state legislators and municipalities have implemented goals to eliminate traffic related deaths and injuries. Their efforts have lowered deaths and major injuries by 28 percent. This study will hopefully reveal insight into what strategies legislation and its municipalities can take to hopefully prevent accidents from occurring.

# Methodology and Materials

The original dataset from the Seattle department of transportation was studied and analyzed for various factors which could provide insight to traffic accident severity and predicting what factors could mitigate such injuries.

The dataset was first normalized, and all null values were eliminated. All the factor sets were considered and projected and examined in bar graphs to determine which factors were the top contributors to be used as predictor variables.

The variables were then weighted based upon whether or not the factor could be controlled or possessed a specific impact on the outcome (severity of accident). These factors were then chosen based on their weighted and logical impacts on accident severity. The data was also set to correlations t determine if they had any effect on the severity outcome. These factors were used to build various machine learning models with the end goal to be to produce the specific severity codes: 0-unknown, 1- property damage, 2-injury, 2b- serious injury, and 3- fatality.

Variables such as city code, days of the week, and months and other non-essential data were eliminated. When analyzing data variables, it was evident that they were to similar to be considered great factors.

The use of various machine learning tools was used to determine if there was an ability to predict the severity of an accident and when or where such accidents would most likely occur. The most reasonable machine learning technique could provide such insight and provide a more promising effect was the Decision Tree model. The model was split, trained, and tested which proved to be a viable predictor of accident severity.

Another machine learning model used was logistic regression. Logistic Regression can be used as a noise tolerant classification method. The data set although has some noise it was seen as a logical method to use as well. The data was again split, trained, and tested at various C levels to conclude which would best produce promising results.

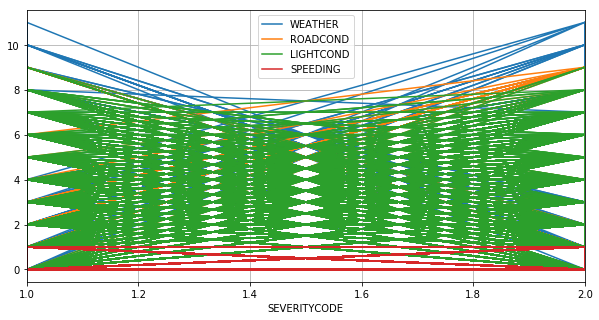
KNN, (K-nearest neighbor), was also used but was omitted due to seeming to be unfitting for the data set as a model for predictions.

**Results**

After careful analyzation, considerations, and multiple testing of different models the results showed that **Data observations:**

1. More injuries involving prop. Damage vs. Injury
2. Collisions happened in or near downtown or highway
3. Collisions without injury are well distributed.
4. Most common persons involved is 2
5. Rare for pedestrians to be involved
6. More collisions happen on Friday, least on Sunday- with the average accidents per day being very similar to the same.
7. More collisions happen in October and less in Dec, but average accidents per month were very similar to the same.
8. Clear weather caused more occurrences of collisions with rainy days being the second highest.
9. Dry road conditions showed to be more prevalent to road accidents then wet.
10. Speeding accounted for main reason for collisions, along with carelessness and most likely poor infrastructure.

Speeding, weather, road and light conditions are good features to predict collision severity.

***CHART SHOWS FEATURES AND THEIR OVERLAP OF IMPACT ON SEVERITY CODE***.

The correlation data was to prove that these factors directly impacted the severity code for the data set, and they showed positive correlation scores all above .98. showing they are good variable to build models form.

**EXAMPLE CORR. CODE:**

df['SEVERITYCODE'].corr(df['WEATHER'])

Out[18]:

0.9999996387652679

In [19]:

df.corr()

Out[19]:

|  | **SEVERITYCODE** | **WEATHER** |
| --- | --- | --- |
| **SEVERITYCODE** | 1.0 | 1.0 |
| **WEATHER** | 1.0 | 1.0 |

In [26]:

df['SEVERITYCODE'].corr(df['SPEEDING'])

Out[26]:

0.9999996387652679

In [27]:

df.corr()

Out[27]:

|  | **SEVERITYCODE** | **SPEEDING** |
| --- | --- | --- |
| **SEVERITYCODE** | 1.0 | 1.0 |
| **SPEEDING** | 1.0 | 1.0 |

After implementing machine learning techniques, the Jaccard scores for both the decision tree and logistic regression models proved to be near 0.70.

**JACCARD SCORES FOR BOTH THE DECISION TREE AND LOGISTIC REGRESSION MODELS:**

##DT

jaccard\_similarity\_score(Y\_test, dtyhat)

Out[33]:

0.7008664086846341

##DT F1

f1\_score(Y\_test, dtyhat, average='macro')

Out[34]:

(0.532314927278956, 0.5000580523239426, 0.41250769110713076, None)

##LR

jaccard\_similarity\_score(Y\_test, lryhat)

Out[12]:

0.7009691448922982

*#logloss*

yhat\_prob = lr.predict\_proba(X\_test)

log\_loss(Y\_test, yhat\_prob)

Out[14]:

0.6049390411925653

With an F1-Score of .54 , which we can expect that these where the maximum values achieved when weighing out precision against recall. Therefore, the decision tree model is a good predictor model for accident severity. The logistic regression log loss score was at 0.6 with a C depth of 5. Although it was not as good of a predictor as the decision tree model it would still be viable for information on the matter. The models were ran with suggested information which the outcome was already known and produced the correct severity when tested with poor weather conditions , while speeding, and on a wet road; the severity code produced at code of 2- injury without being disabling or fatal. While on clear weather days without speeding, on a dry road the severity level was determined to be 1- property damage no injury reported.

KNN, (K-nearest neighbor), was also used but seemed to be unfitting for the data set as a model for predictions. Although it the scores for such model seem relevant the information was too tightly similar to guarantee the results were completely or fairly accurate.

### Discussion

The overall outcome of the results would suggest that on days when visibility is reduced more sever accidents were likely to take place. Although more accidents occurred on clear days the severity was less debilitating. Also, when persons are speeding the more likely the severity code was in the upper quadrant of severity.

When considering the results, it seems evident that to fix or diminish accident severity that during busy times of day mornings and late afternoons, during raining or reduced visibility all highway lanes should be open including HOV. Allowing for the less cramped lanes which can lead to collisions. Also, during those times, message boards displaying delays and speed changes would be utilized to make travelers aware of halted or slow traffic.

Although the weather cannot be controlled how local and state municipalities respond to the studies results could heavily impact severity in the future. Distractions and poor infrastructure could play a very large roll in where these accidents take place. Highway expansions or investing in visual enhancements to make drivers more aware of their surroundings and traffic news could lead to less severity.

As speeding remains a major contributor a larger portion of the fatality scores, state legislation could follow recommendations to enhance fines for speeding, and/ or longer probationary or suspension time for reckless driving speeds. Considering the majority of accidents were due to rear-ending other vehicles, law enforcement could be more vigilant and visible to the public during peek times of accidents. This tactic could deter such poor behaviors such as following to close to forward vehicles, speeding, and use of distracting materials.

**Conclusion**

Vehicular crashes can lead to many unsettling outcomes. The major contributors to crash severity range from weather conditions to driver actions: speeding or distractions. These are then most common causes of accidents examined that directly correlate with the severity outcome. There are many ways to predict how sever a crash could be that have been discussed to provide conformation as to what conditions can cause certain accidents. There are also, many recommendations that can be made to lawmakers, law enforcement, the Transportation Departments, and to individuals. To simply educate drivers and individuals is not enough as it is relatively common knowledge that those variables can clear the way for crashes to happen. There is really only one way a person can truly prevent such crashes and injuries and it is by being more aware and giving themselves a constant reminder to be cautious and obey all traffic guidelines.

References

REDMOND. (2020, August 10). 25 Things To Know Before Moving to Seattle In 2020 | Should I Relocate To Seattle? On The Go Moving and Storage. <https://onthegomoving.com/moving-to-seattle/>

Statistics Solutions. (n.d.). *What is Logistic Regression?* Retrieved September 10, 2020, from https://www.statisticssolutions.com/what-is-logistic-regression/#:%7E:text=Logistic%20regression%20is%20the%20appropriate,variable%20is%20dichotomous%20(binary).&text=Logistic%20regression%20is%20used%20to,or%20ratio%2Dlevel%20independent%20variables.

Wikipedia contributors. (2020, September 17). *Seattle*. Wikipedia. https://en.wikipedia.org/wiki/Seattle

Footnotes

NONE

Tables

Table 1

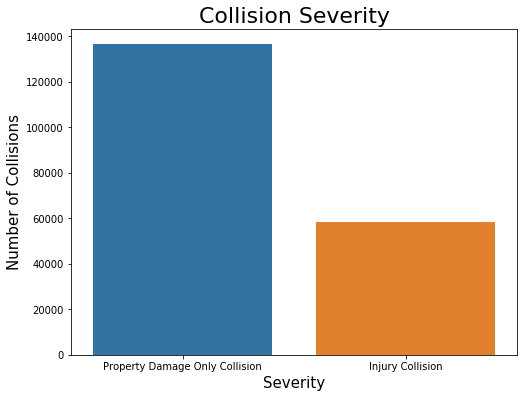
ORIGINAL DATA AFTER DROP VALUES

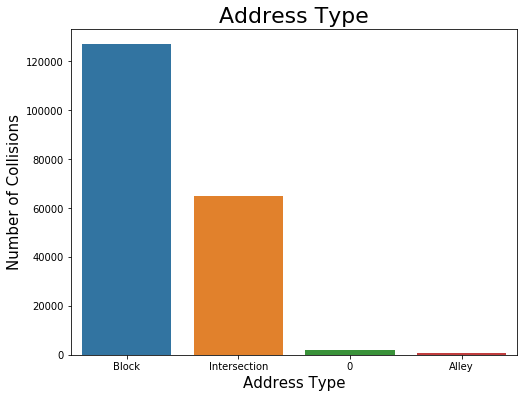
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| df1.head() |  |  |  |  |
| Out[3]: |  |  |  |  |
| SEVERITYCODE | WEATHER | ROADCOND | LIGHTCOND | SPEEDING |
| 0 | 2 | Overcast | Wet | Daylight |
| 1 | 1 | Raining | Wet | Dark - Street Lights On |
| 2 | 1 | Overcast | Dry | Daylight |
| 3 | 1 | Clear | Dry | Daylight |

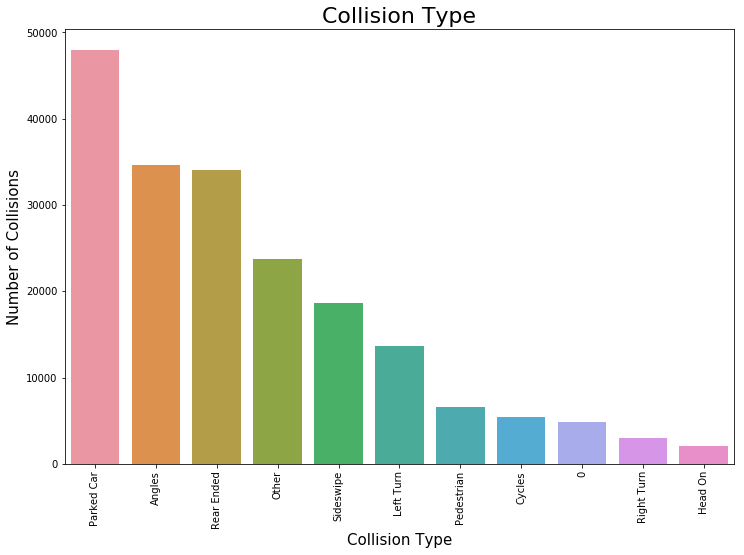
NORMALIZATION DATA AFTER DROP VALUES

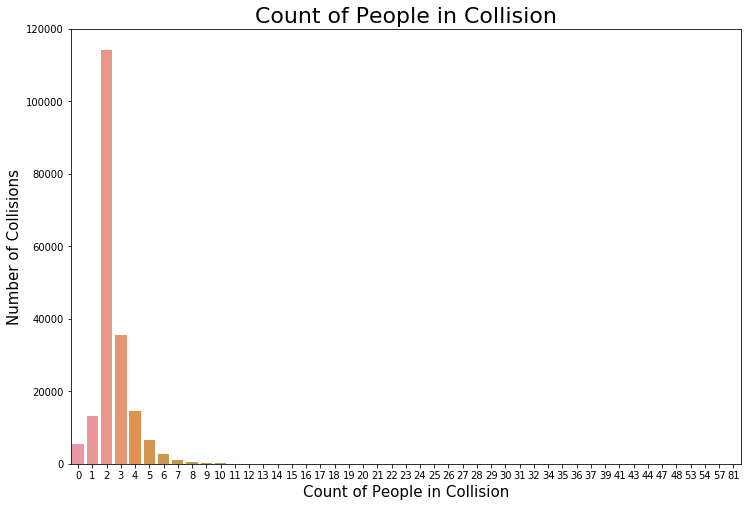
|  | ***SEVERITYCODE*** | ***WEATHER*** | ***ROADCOND*** | ***LIGHTCOND*** | ***SPEEDING*** |
| --- | --- | --- | --- | --- | --- |
| ***0*** | *2* | *3* | *2* | *1* | *0* |
| ***1*** | *1* | *2* | *2* | *2* | *0* |
| ***2*** | *1* | *3* | *1* | *1* | *0* |
| ***3*** | *1* | *1* | *1* | *1* | *0* |
| ***4*** | *2* | *2* | *2* | *1* | *0* |

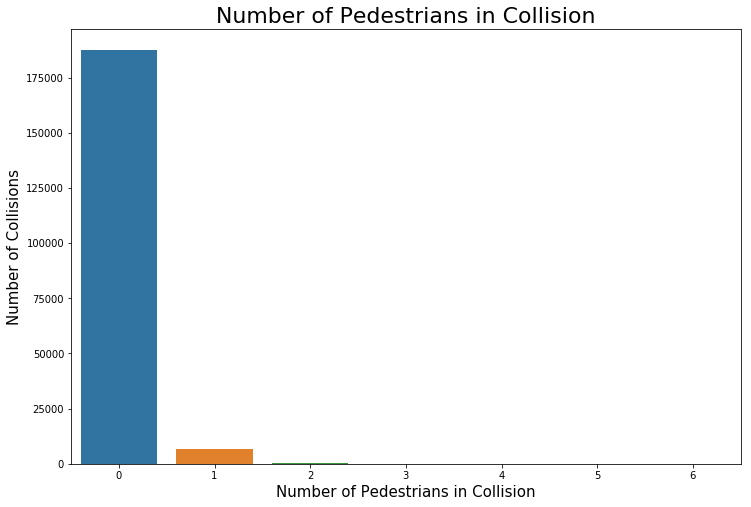
bar\_plot(df['SEVERITYDESC'], 'Severity', 'Number of Collisions', 'Collision Severity', size=(8,6))

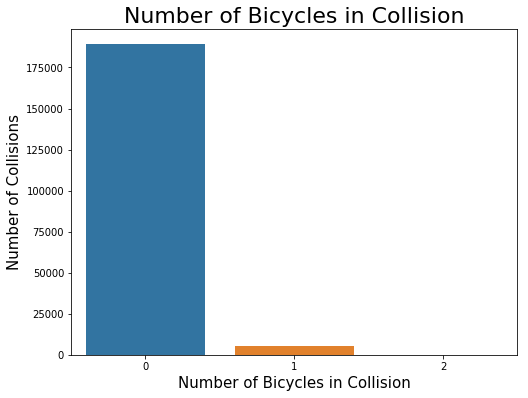
**

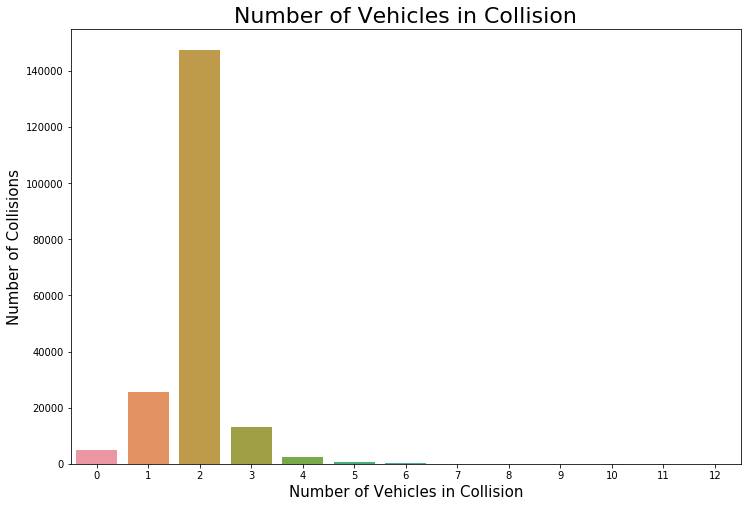
**

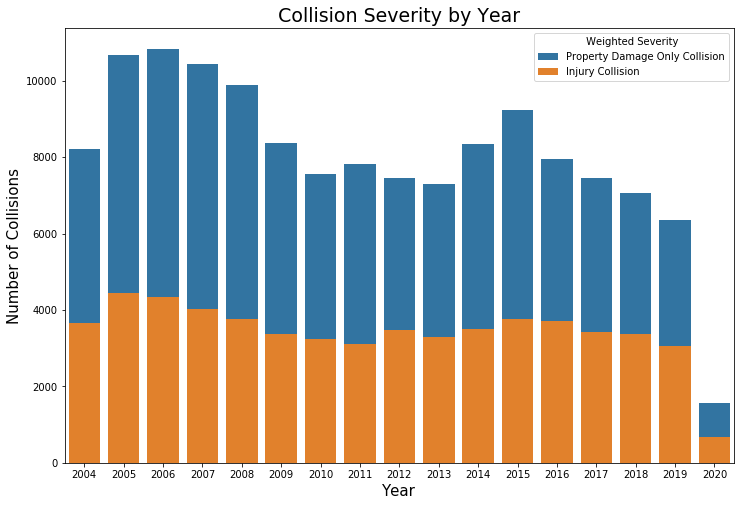
**

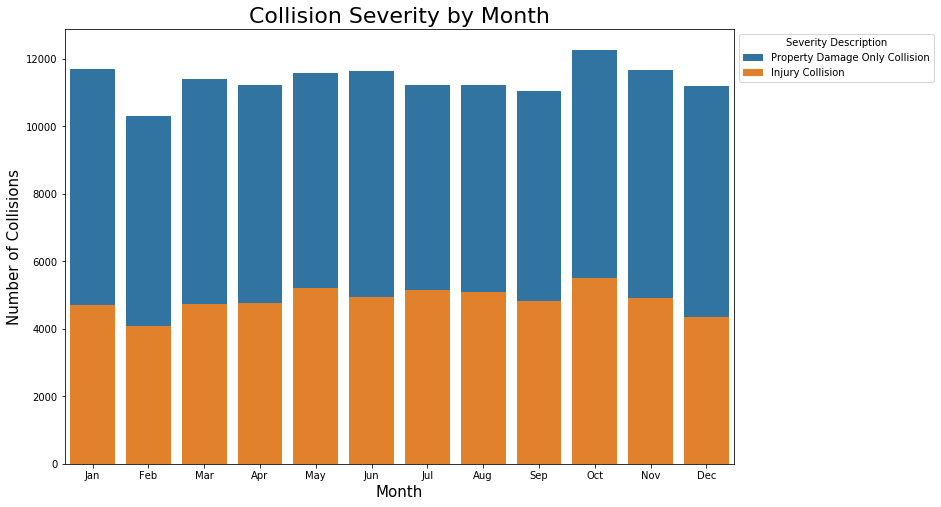
**

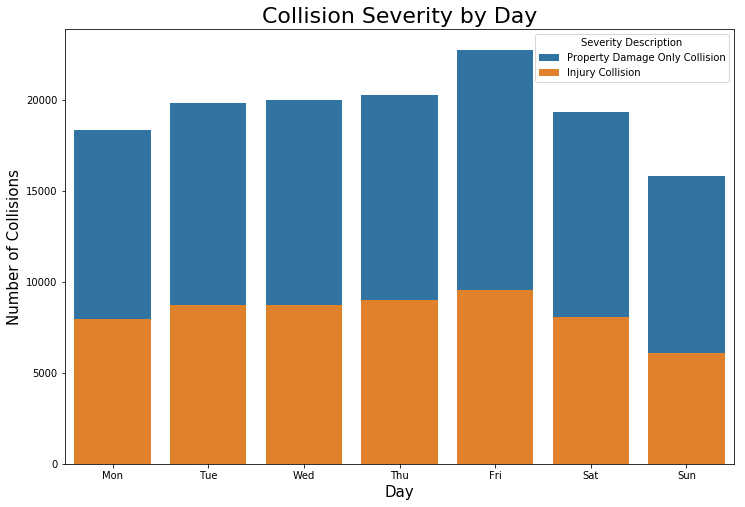
**

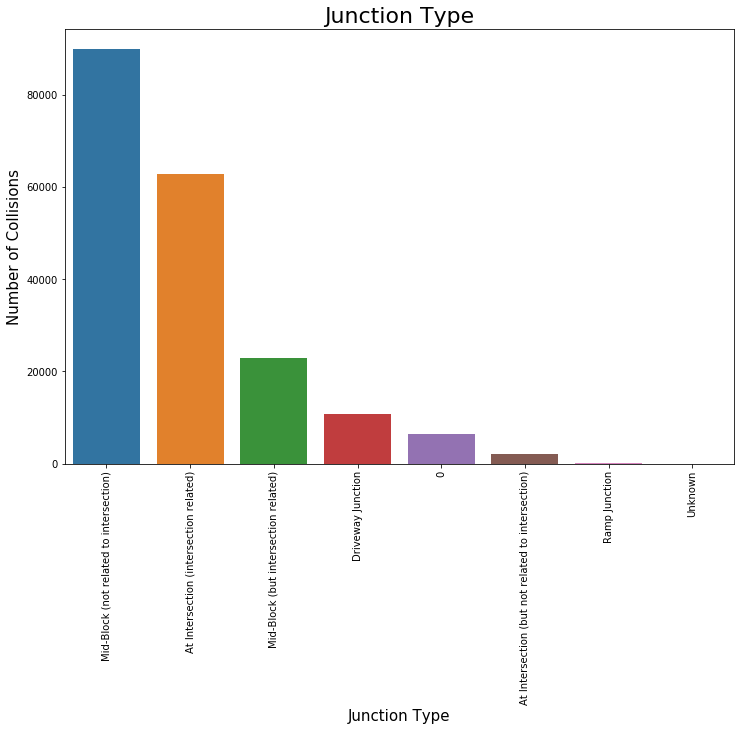
**

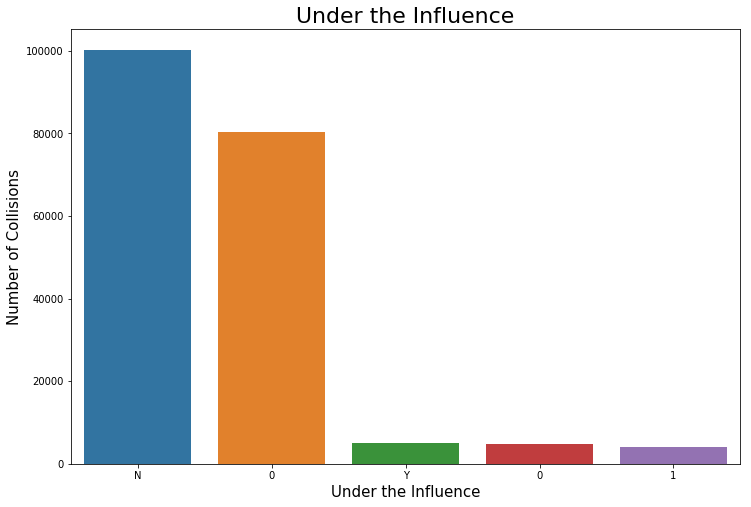
**

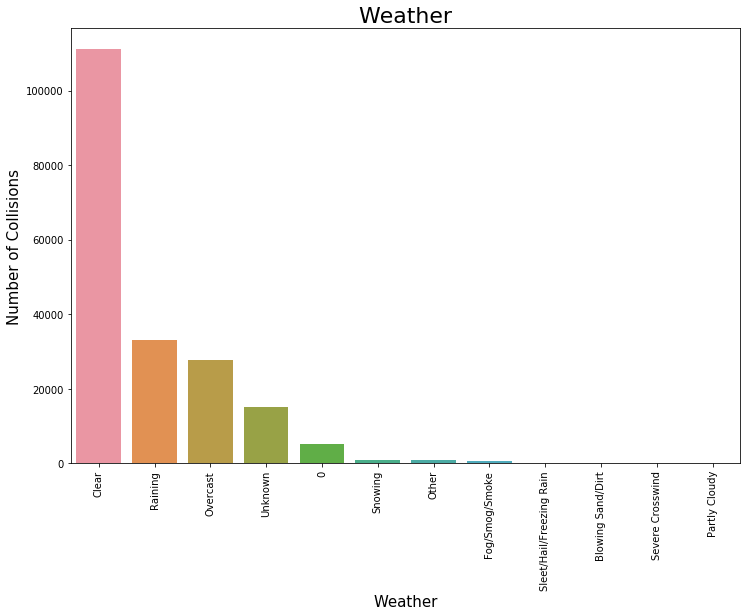
**

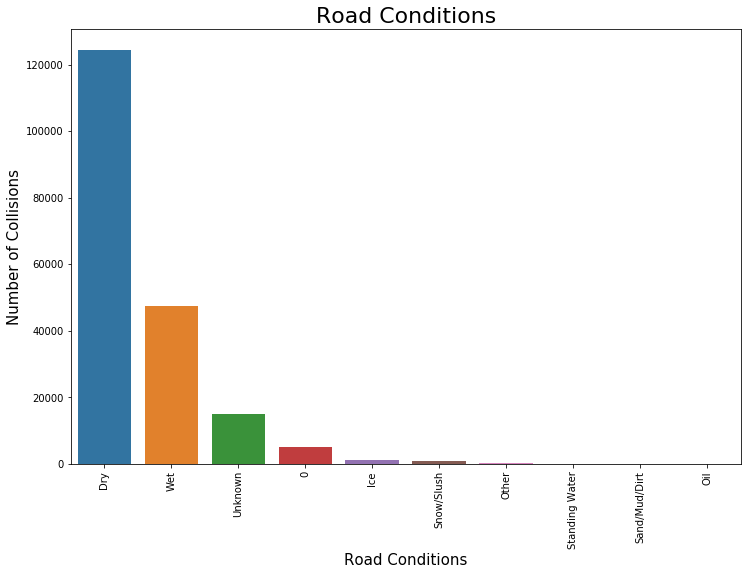
**

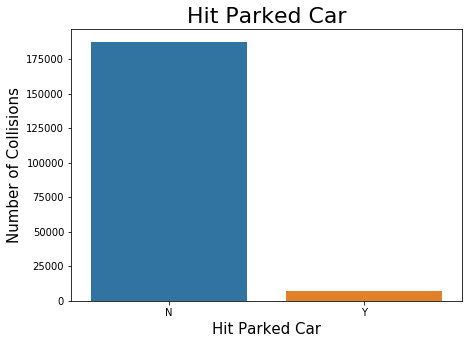
**

**

**

**

**

**