Study of Severity of US Car Accidents Using Machine Learning

Shambhu KC

September 21, 2020

1 Introduction

1.1 Understanding the problem

Road accident is one of the leading causes of death in many developed as well as in developing countries. According to Centers for Disease Control and Prevention (CDC), approximately 1.35 million people are killed each year worldwide on roadways. And, another 50 million people are injured which involves serious damage to their body parts. In addition to causing serious physical damage, the road accidents also cause impact on the traffic; in some cases the accident causes several hours of delay. Long hours of traffic is something that everybody wants to avoid. Hence, reducing the roadways accidents have become an important public safety challenge. While it is not possible to shutdown the whole transportation system to prevent those deaths, since it has direct impact on the economic and social development of any countries. However, the risk can be minimized by taking several preventive actions such as providing more road safety guidelines, improving the road conditions, and imposing new regulations for drivers. In order to take those preventive actions, it is first necessary to understand what is actually causing the problem. Hence the question that we are asking here is "can we build a model that learns from the historical data and predict the severity of future accidents given the involving the factors?" From this model, we can learn the impact of different factors on the level of severity. Those factors can be road conditions, weather conditions, daytime, safety status, driving conditions. After we have clear idea of which factor is leading more to

severe accidents, we can then lay out several recommendations to minimize those risks. In addition, depending on the severity level, we can also predict how long the traffic is likely to be delayed by that accident. This is very important for those people who are looking to reach to their destination on time.

1.2 Objective

This model would be beneficial to anyone driving in the road at any time. Since, the main goal of this model is to predict the severity of accident and the delays it can cause to the traffic, it can notify others driving through the same road, in the real-time, about the accident and the possibility of delays they might experience ahead. Based on that they can decide whether to proceed in the same road or take an alternate route. As the model will also cover impact of the different weather conditions and driving conditions on the severity of the accident, it would also be useful to giving recommendations to general public to decide whether it is safe to drive given the current situations.

2 Understanding the Data

The data that we are going to utilize for this project is countrywide car accident dataset, which includes the data collected from February 2016 to June 2020 covering 49 states of the USA. This dataset was extracted from kaggle uploaded by Sobhan Moosavi. There are about 3.5 million accident records. This dataset includes many intrinsic and contextual attributes such as location, time, level of severity, time taken to clear the traffic, weather conditions including many other information. The important part of this dataset is that it contains the most recent reports of accidents as well. Since, our main focus is on the severity of the accident, the following features from the dataset are very useful:

- **Severity**: It shows the severity of the accidents in terms of number between 1 and 4, where 1 indicates the least impact on traffic, whereas 4 indicates a significant impact on the traffic.
- **Start Time**: It shows the start time of the accident.
- **Temperature(F)**: It shows the temperature at the place of accident (in Fahrenheit).

- **Wind Chill(F)**: It shows the wind chill (in Farenheit).
- **Visibility(mi)**: Shows visibility (miles).
- **Wind Speed (mph)**: Shows wind speed (in miles per hour).
- **Precipitation(in)**: Shows precipitation amount in inches, if there is any.
- **Amenity**: It indicates presence of amenity in a nearby location.
- **Crossing**: It indicates presence of crossing in a nearby location.
- **Give Way**: It indicates presence of give way in a nearby location.
- **Junction**: It indicates presence of junction in a nearby location.
- **Traffic Signal**: It indicates presence of traffic signal in a nearby location.
- **Crossing**: It indicates presence of railway in a nearby location.
- **Stop**: It indicates presence of stop in a nearby location.
- **Station**: It indicates presence of station in a nearby location.
- **Sunrise Sunset**: It shows the period of the day (i.e. day or night) based on sunrise/sunset.

3 Data Preparation

Initially, the dataset had 3,513,617 rows and 46 columns. As I was not going to explore the severity in terms of the location (i.e., which states or counties has more severe accidents), so all the columns related to location are dropped. In addition, some of the main features for this study such as Precipitation, Temperature, Wind Chill were found to be missing values. Hence, all the rows containing a NAN values are also dropped. I attempted to replace the missing value with column average. However, after assigning the mean value for NAN values, I realized that it also changes the mean value for each severity level. Hence, instead of using a hypothetical value, which may or may not be correct, I decided to drop the entire row containing a NAN value.

After dropping the unnecessary columns and the rows with NAN, the dataset had a shape of (1282849, 32).

I was also interested in exploring if there is any trend in the number of accidents related to the time of the day or day of the week. But, there were no separate columns displaying those information. Hence, by using the python datetime module and the start time column, both the time and day of week are extracted.

4 Exploratory Data Analysis and Visualization

First of all, by using the bar plot, I explored total number of accidents by severity listed in the dataframe. It was found that the majority of the accidents are of severity level 2, followed by level 3 and 4.

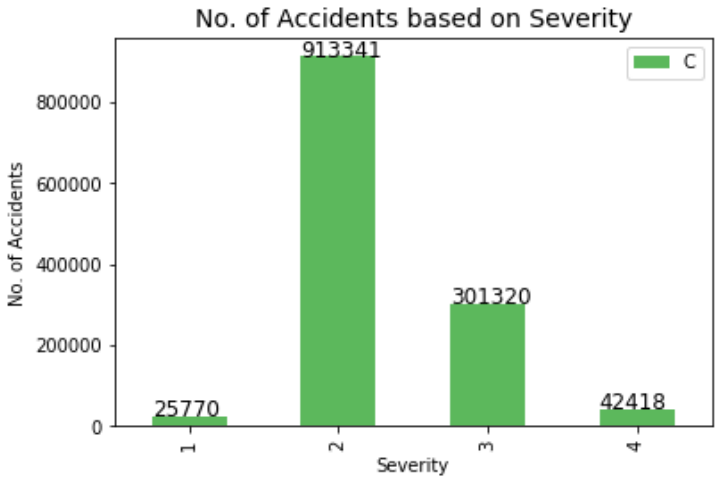

Figure 1: abced

Before building any model, it was necessary to understand the relation between each of the features and the severity of the accidents. Various visualization as well as statistical measures were utilized to study the extent of correlation. The bar graph (as shown in figure 2) shows that on a typical workday there were more accidents than on the weekends. The more accidents on a weekday can be related to the more number of vehicles on the

road.

It can also been seen that during daytime there are more accidents than

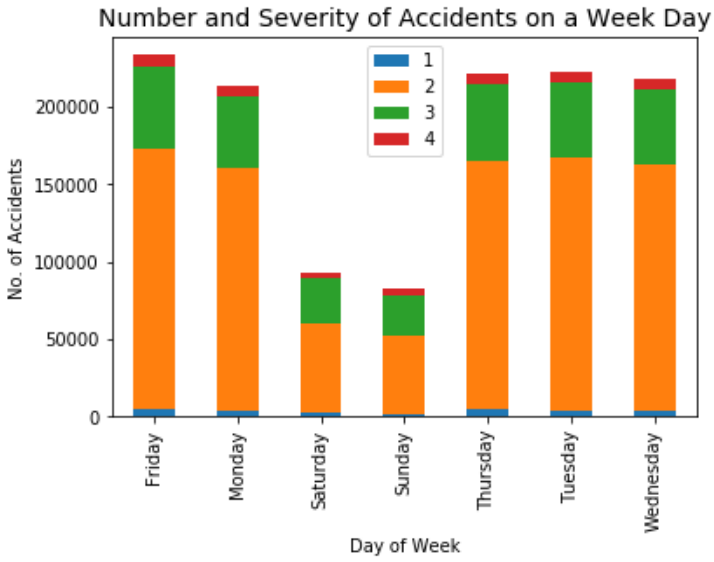

Figure 2:

during nighttime, which is again due to the fact that there were more vehicles on the road during daytime, see fig.3. On a typical workday, the time between 8 - 9 am and 5 - 6 pm is usually a heavy traffic hours due to office starting time and ending time. That means at that time, there will be more vehicles on the road increasing the probability of an accident. The exact trend can be observed from the line plot shown in fig. 4. The number of accidents peaked around 8 am and again at 5 pm. Also, there were more accidents in the day as compared to night.

I was also curious to know if the proximity of traffic objects fuels the accidents. I used a pie chart, as shown in figure 5 to study how the number of accidents varies with respect to the traffic signals or other objects. It can be seen that the majority of accidents occurred near the traffic signal, crossing, and junction. This actually makes sense because often people don't want to wait for any crossing crowds or the incoming red lights, which stimulates them to speed up and cross the intersection. Unfortunately, such an action results in accidents in many cases.

I also used Pearson correlation method to calculate the coefficient of corre-

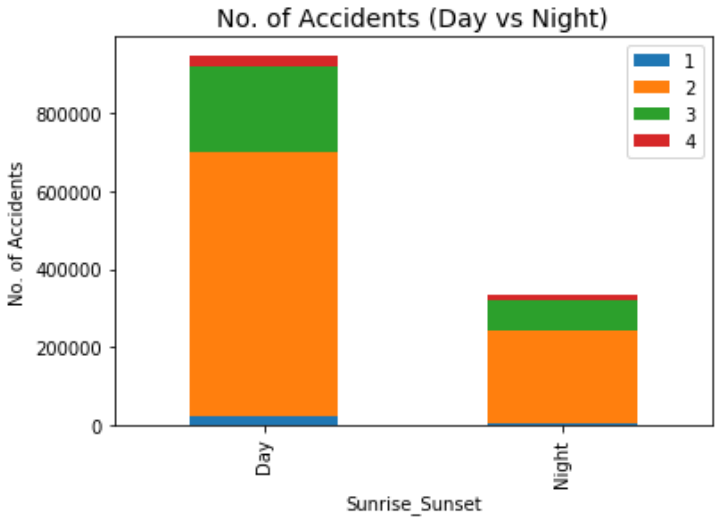

Figure 3:

lation between the features of interest and the accident severity. A heat map used to display the possible correlation, as shown in figure 6. Although the correlation is not very strong, it is sufficient to argue that there some level of dependency between the accident severity and the features of our interest.

Then to explore the impact of different weather condition, box plot along with estimation of mean were utilized. From these box plots, it can be argued that the weather condition has influence on the seriousness of an accident. Typically, when the outside temperature is low, when there is more rain, and when the visibility is low, more serious accidents had occurred (see figure 7).

5 Build a Model

From the data visualization and from exploratory data analysis, it is clear that the selected features have influence on the severity of the accident. Hence, these features are used to build a model by using machine learning algorithm. In this project, I have utilized two algorithms; Decision Tree and K Nearest Neighbors. Both algorithms are supervised learning approach and works best in the case of a categorical response variable which takes discrete values. The sckit-learn library for python programming has many in

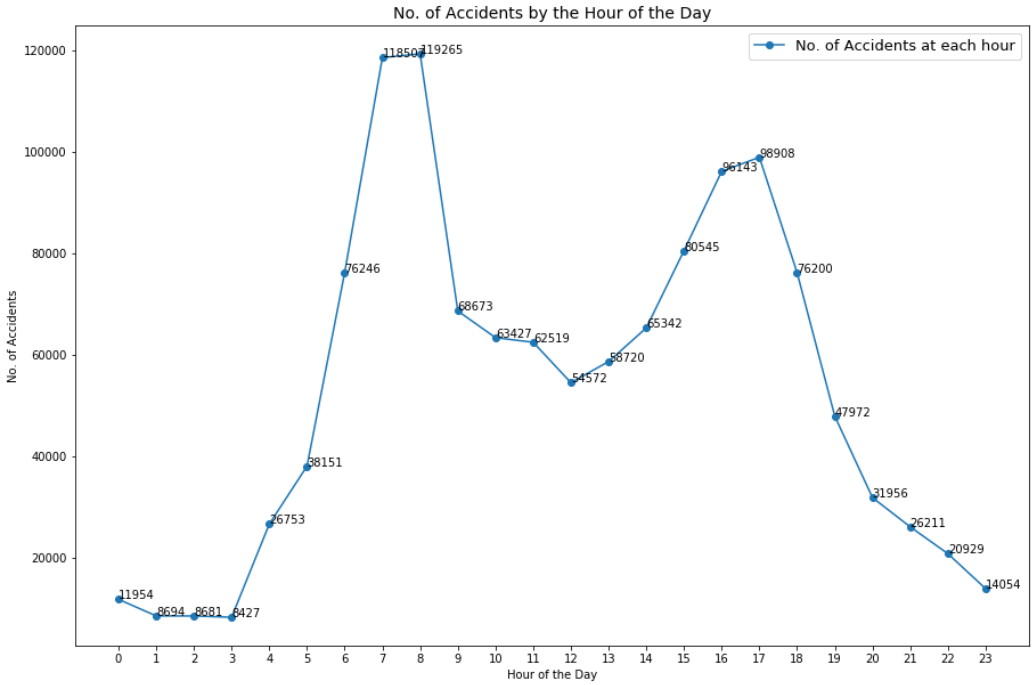

Figure 4:

built functions, which really makes the model building much simpler. First of all, the entire dataset was split into train and test set. The model learns from the train set and it's accuracy is then performed on the test dataset. In this project, to find the best set of parameters for each classification method, I utilized GridSearchCV method and then compared the different evaluation metrices such as accuracy-score, f1-score, precision-score, recall-score.
The results are shown below.

6 Result

In summary, it is found that the severity of an accident depends on many factors. When there are more vehicles on the road, such as during rush hours, the probability of an accident is higher. The probability of accident is also found be higher in proximity of the traffic objects. The low temperature and low visibility, more precipitation can result more severe accidents.

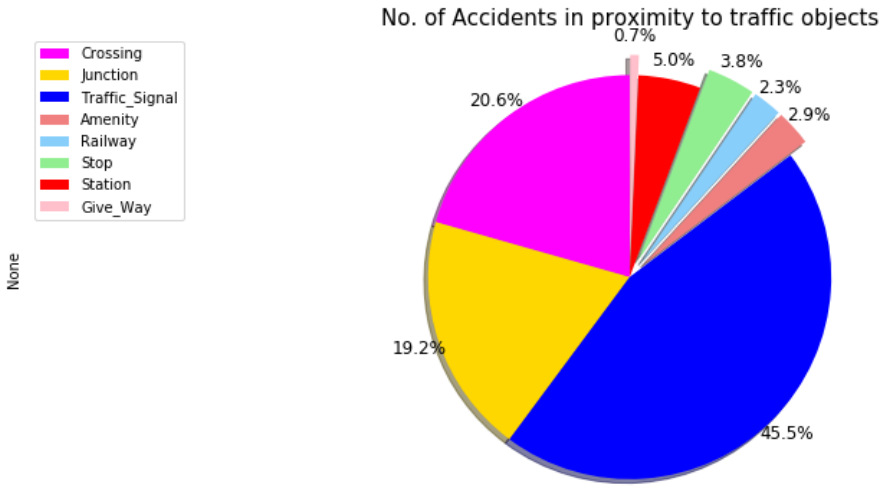

Figure 5:

7 Recommendation

Based on the analysis of historical data, I can provide following recommendations. Since more accidents have occurred in proximity of intersection or traffic lights, more stringent rules should be enforced at those places. The drivers should be more cautious when it is rush hour and when the weather condition is not ideal.

8 Future Work

In this project, I analyzed the severity in terms of the impact on the traffic. However, it is also necessary to study the severity in terms of fatalities and how other features such as sex of the driver, speeding condition, vehicle type impact the seriousness of the accident. Such a study would be helpful to provide more detailed guidance on the road safety. Hence, my future goal is to acquire data covering those information and extend the current model.

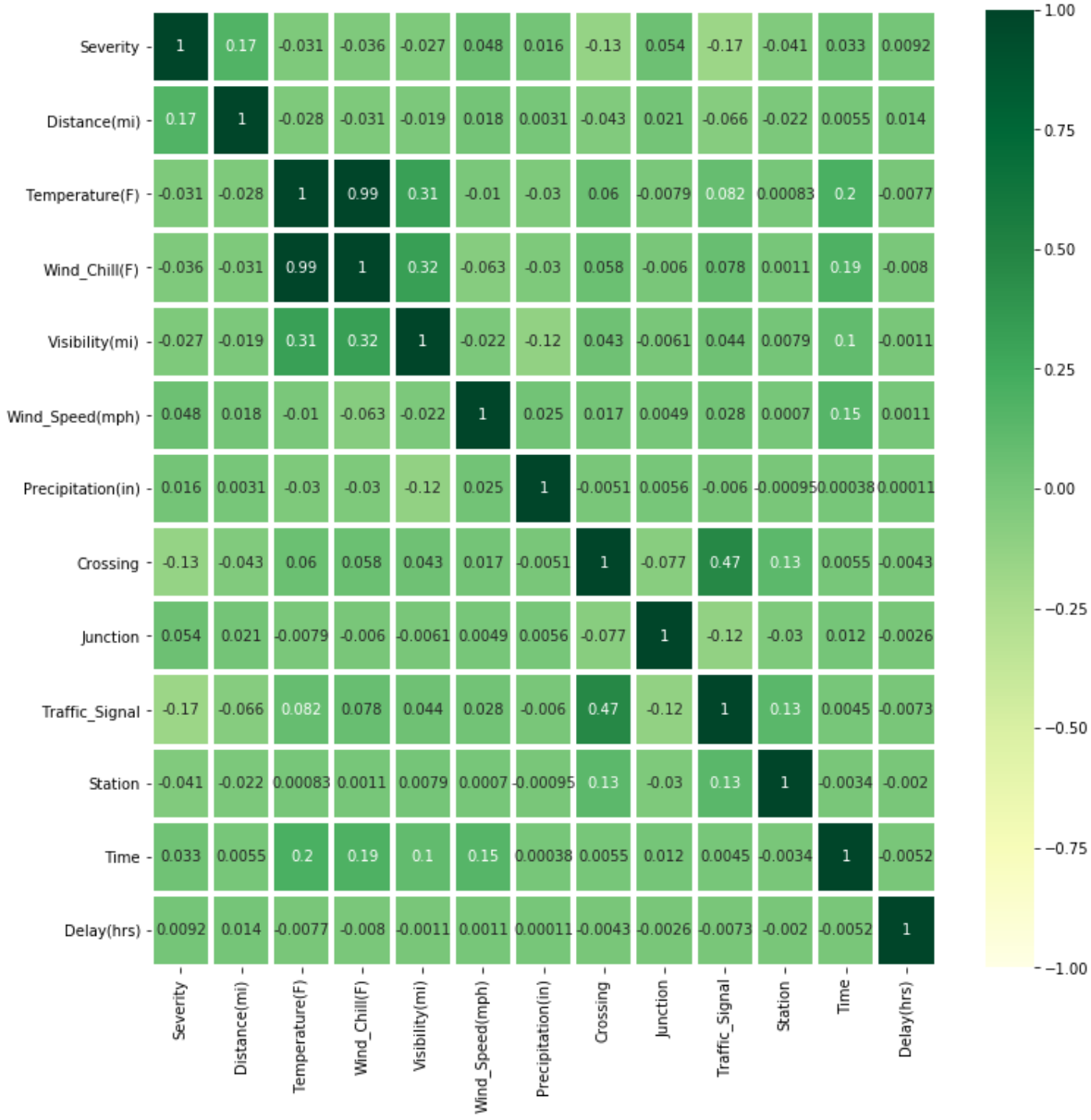
Severity 1 0.17 -0.031 -0.036 -0.027 0.048 0.016 -0.13 0.054 -0.17 -0.041 0.033 0.0092
Distance(mi) 0.17 1 -0.028 -0.031 -0.019 0.018 0.0031 -0.043 0.021 -0.066 -0.022 0.0055 0.014
Temperature(F) -0.031 -0.028 1 0.99 0.31 -0.01 -0.03 0.06 -0.0079 0.082 0.00083 0.2 -0.0077
Wind_Chill(F) -0.036 -0.031 0.99 1 0.32 -0.063 -0.03 0.058 -0.006 0.078 0.0011 0.19 -0.008
Visibility(mi) -0.027 -0.019 0.31 0.32 1 -0.022 -0.12 0.043 -0.0061 0.044 0.0079 0.1 -0.0011
Wind_Speed(mph) 0.048 0.018 -0.01 -0.063 -0.022 1 0.025 0.017 0.0049 0.028 0.0007 0.15 0.0011
Precipitation(in) 0.016 0.0031 -0.03 -0.03 -0.12 0.025 1 -0.0051 0.0056 -0.006 -0.00095 0.00038 0.00011
Crossing -0.13 -0.043 0.06 0.058 0.043 0.017 -0.0051 1 -0.077 0.47 0.13 0.0055 -0.0043
Junction 0.054 0.021 -0.0079 -0.006 -0.0061 0.0049 0.0056 -0.077 1 -0.12 -0.03 0.012 -0.0026
Traffic_Signal -0.17 -0.066 0.082 0.078 0.044 0.028 -0.006 0.47 -0.12 1 0.13 0.0045 -0.0073
Station -0.041 -0.022 0.00083 0.0011 0.0079 0.0007 -0.00095 0.13 -0.03 0.13 1 -0.0034 -0.002
Time 0.033 0.0055 0.2 0.19 0.1 0.15 0.00038 0.0055 0.012 0.0045 -0.0034 1 -0.0052
Delay(hrs) 0.0092 0.014 -0.0077 -0.008 -0.0011 0.0011 0.00011 -0.0043 -0.0026 -0.0073 -0.002 -0.0052 1
Severity
Distance(mi)
Temperature(F)
Wind_Chill(F)
Visibility(mi)
Wind_Speed(mph)
Precipitation(in)
Crossing
Junction
Traffic_Signal
Station
Time
Delay(hrs)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Figure 6:

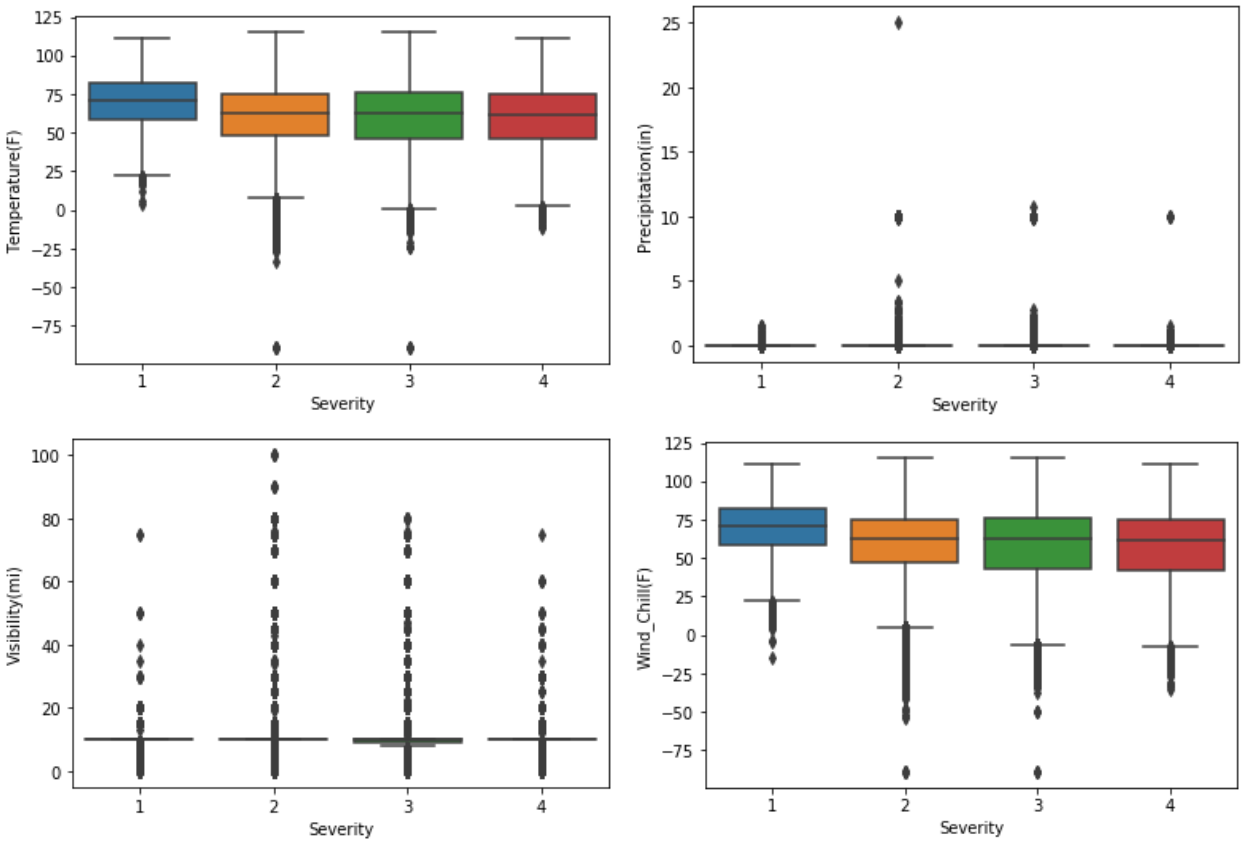
Temperature(F)
Severity
Precipitation(in)
Severity
Visibility(mi)
Severity
Wind_Chill(F)
Severity

Figure 7: