A Model for Warning Drivers of Upcoming Serious Car Accidents Using Real-Time Generated Data

Nick Chatziefthymiou

September 17, 2020

1. Introduction

Car accidents are one of the leading causes of death worldwide, especially among children and young adults. Approximately 54 million people sustained injuries causes by road traffic crashes in 2013 with many of them leading to disabilities while 1.4 million of them were fatal. The study also suggests that deaths from car accidents in 1990 were estimated to be 1.1 million showing an increase in fatalities. More than half of all deaths in road traffic accidents occurred among vulnerable road users such as cyclists, motorcyclists and pedestrians according to the World Health Organization. Car accidents is a major worldwide problem with social, economic and health consequences for the people. A significant number of scientific studies attempt to address this issue in order to find solutions spanning from policies to reduce the frequency and severity of the accidents to algorithms predicting the probability of occurrence and the risk of an incident. In this study we will focus on making real-time predictions on the severity of an accident given a set of environmental variables by implementing a machine learning model in order to warn drivers of potentially dangerous roads. The algorithm can be developed into an application for smartphones or be embedded into GPS devices as well as in electronic road signs which will be prompting the drivers to be more attentive or even avoid certain roads.

2. Accident Dataset

The goal of this section is to indicate the sources where the data has been collected from as well as describe the meaning of each feature in the acquired dataset. It is of great importance that any data collected for the development of models intended for public use comes from reliable sources, especially when its use concerns public safety. The dataset used in this study comes from seattle.gov, the official website of the city of Seattle and provides collision data recorded from 2004 – 2020.

2.1 Data Description

There are 38 features present in the dataset and a total of 194673 records of car accidents. The first few columns provide several identification keys unique for each incident, this information is not usable for the purpose of this paper therefore it will be ignored in the analysis. The location of the accident is described by three columns, two of which refer to the latitude and longitude values while the third one contains the address. The date and time when the incident occurred is given by two columns, one containing only the date while the other containing the date and time of the collision. Next there is information on whether the accident occurred at an intersection, a block or an alley, the type of the junction and the unique keys of each intersection, crosswalk road segment where the accident occurred. There are several features showing variables such as the weather, the road condition, the lighting condition, whether the driver was inattentive, speeding was under the influence of alcohol or the pedestrian right of way was not granted. The dataset also provides information regarding the number of pedestrians, cyclists and vehicles involved in the accident and whether there was a collision with parked cars. There is information regarding the State Collision Code which uniquely describes each type of collision using a numeric value. Finally, the severity of each accident is given as a binary variable taking the value of 1 if the accident was less serious and the value of 2 if the accident was severe, this column will be used as the target data for the model to predict.

2.2 Feature Selection

The goal of this study is the development of a model that predicts the severity of an accident using real-time data, this puts some limits on the type of features that can be used since not all of the dataset's columns contain information that can be obtained in real-time. Locational variables such as intersections codes, junction types etc. as well as information regarding weather, road condition and speed limits can be acquired through the use of GPS, news and speedometers. Features containing information regarding the type of the collision cannot be used since there is no way to know this before the accident occurs. Furthermore, data on whether the driver is inattentive or under the influence of alcohol would be unknown to a machine such as electronic road signs or GPS devices therefore they should be excluded from the set of usable features.

There are more variables for which the methodology of obtaining real-time data is ambiguous. These include the number of vehicles, pedestrians, cyclists and whether parked cars were involved in an accident. These features can be converted into binary, this would mean that they will provide information on whether or not they were involved in an accident. By analyzing the frequency of vehicles, pedestrians and cyclists passing from certain locations as well as the number of parked cars in that location, signals of these frequencies passing certain limits can be generated.

2.3 Data Preparation

The process of data preparation includes removing or filling missing values, converting features into certain formats and making sure the data is balanced. This step is required in most algorithms in order for them to be implemented correctly and can significantly boost their accuracy. The dataset contains features which vary on their data types as well as a lot of missing values. Most of the features contained either discrete numeric or categorical values, the later were converted into discrete numeric values so that they can be used in the algorithm. Certain features contained only one unique value and a large number of missing values, for cases such as these missing values were converted into their own category. Such an example would be the column showing whether a car involved in an accident was speeding or not, the only unique value of this column was 'Y' which most probably ment 'YES', when a row wasn't equal to 'Y' the value was missing, this indicated that the empty rows should take the value 'N' which would stand for 'NO' so that the feature would become binary. Some features contained a large number of unique values many of which had only a small number of appearances in the dataset, these values were merged with larger categories of similar characteristics. There were three variables containing information on the exact location of the accident, from these features only the ones containing the latitude and longitude were used while the third one was completely dropped. In order to convert the (lat, lon) coordinates into discrete numeric values the k-means clustering algorithm was used on the two features (Y, X) generating six distinct location clusters. The remaining rows containing missing values were removed.

There were two columns with information regarding the date and time of the accidents. One of them only contained data on the date of the accident while the other also contained data on the exact time it occurred. In order to analyze date and time in more detail, four new rows containing the month, the day of the month, the day of the week and the hour of the accident were generated while the initial columns were dropped from the dataset.

When looking at the frequency of the severity of the accidents it is obvious that the dataset is unbalanced in favor of the less severe ones and therefore biased. After the preprocessing step was done, about 67% of the remaining data had a severity of 1 while only 33% had a severity of 2. In order to tackle this problem a technique called down-sampling was used, a random sample without replacement of size equal to the number of rows containing a severity value of 2 was taken from the rows containing the value 1, the end result was a dataset with an equal amount of the two distinct values of the target data. The final dataset contains 20 columns one of which is the target data and 109380 accident records which is large enough for training and testing the model.

3. Methodology

In order to understand the candidate features and choose those most relevant to the prediction of accident severity, a number of graphs and tables were used. It would be interesting to find out whether one or more unique values of a feature are correlated with higher or lower frequency of severe accidents compared to the general frequency. To achieve that, stacked bar charts showing the percentage of severe versus minor accidents for each unique value of the candidate features were used. Since the data is unbalanced a mark has been placed on each bar indicating the frequency of severe accidents throughout the dataset, this can be used as a way to

compare the frequency of specific categories belonging to candidate features against the dataset frequency. It is clear that when the meeting point of the two stacked bars of a specific category of a feature shows significant distance from the marker, this category is more related to either severe or minor accidents, on the other hand if the stacked bars meet near the marker then the significance of that category for the prediction of severity is very low. Graphs 1.1 – 1.10 show this frequency comparison between severe and minor cases, for each feature there are some categories which show at least a small difference from the marker while other categories remain close to the dataset frequency. More specifically, **graphs 1.1** and **1.2** illustrate how accidents occurring at intersections are likely to be more severe than accidents occurring at blocks or alleys. Regarding the locations on the map a conclusion can be made that there is slightly higher chance of severe accidents at location cluster 4 as shown in **graph 1.4** and **map 1**. The road conditions can also affect the severity of an accidents according to **graph 1.3**, the data shows that when there is oil on the road the chance of having a serious accident increases while standing water, wet roads and snow decrease the severity of the accident. One would expect that dark roads relate to serious accidents but according to **graph 1.5** it's usually the opposite. **Graphs 1.6, 1.7, 1.8, 1.9** illustrate the relationship of date and time to the severity of an accident, there is no significant correlation except perhaps with time since severe accidents tend to be less likely during 1:00am to 5:00am. **Graphs 1.10, 1.11, 1.13, 1.14, 1.16** show the dramatic increase in the probability of having a serious accident when bicycles or pedestrians are involved as well as when the accidents occur at road segments or when the right of way is not granted. According to **graph 1.12** an accident is likely to be less severe if parked cars were involved while **graph 1.15** shows that speeding slightly increases the probability of serious accidents.

The correlation between the candidate features and the target data can indicate whether or not a feature can be accurately used in a model. A heatmap showing the correlation between features and target data as well as among the features themselves is shown in **graph 2.1**. The metric used for measuring correlation was Cramer's V since the final form of the features contain numeric data representing categories. As shown in the graph, the intersection where the accident occurred is the most significant feature when it comes to predicting the severity of an accident. Other features related to severity include the junction's type, whether the accident occurred at an intersection, a block or an alley, whether parked cars, bicycles or pedestrians were involved, the crosswalk where the incident occurred, whether it was a road segment or not and the number of vehicles involved. The rest of the features don't show significant correlation with the target data.

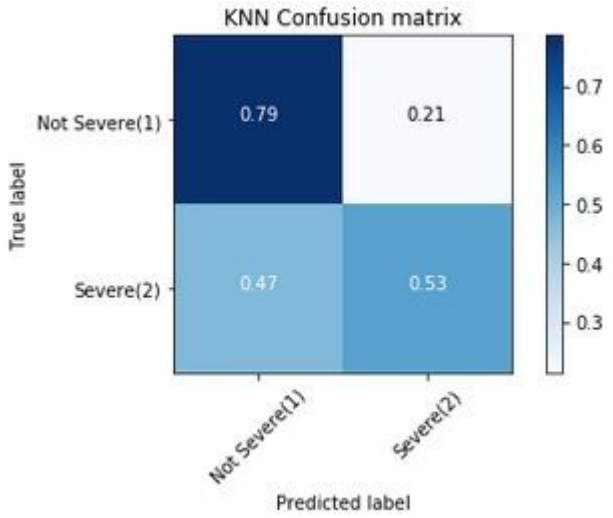
The problem requires a binary classification algorithm that can predict whether an accident will be severe or minor. There are a number of different algorithms to chose from. The algorithms chosen for this project are the k-nearest neighbors, a simple but effective supervised learning algorithm, logistic regression, good for predicting binary classes and a decision tree. Finally, random forest which implements a group of decision trees to improve the prediction accuracy was also implemented. Due to the large size of the dataset, the support vector machine was left out since it would perform poorly in terms of speed.

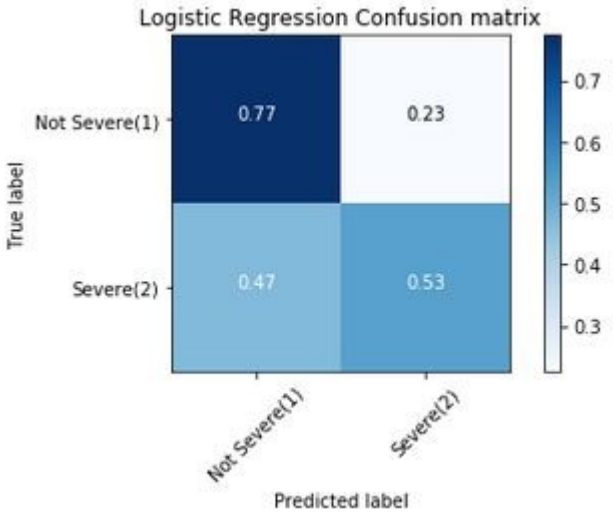
4. Results

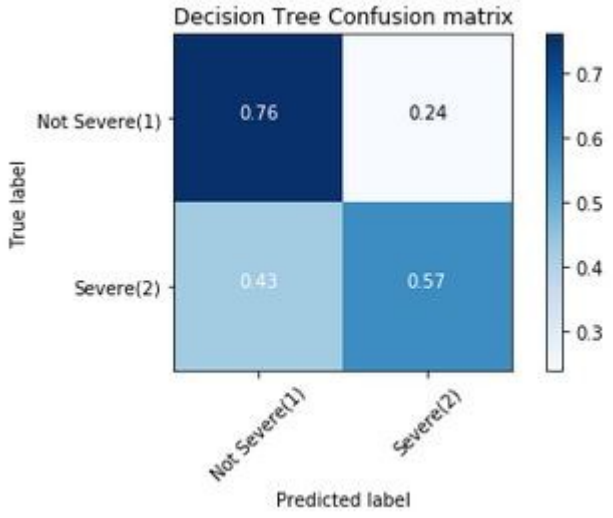
Overall the results of both the decision tree and the random forest as well as the two other algorithms used for comparison show some ability in predicting the severity of a car accident, however the evaluation metrics indicate that further analysis of the subject is required if a reliable model is to be produced. The algorithm that performed the most accurate predictions was random forest with a Jaccard score of 0,5409 and an f1-score of 0.7020 as shown in table 1, the confusion matrix indicates that although it's doing well in predicting minor accidents with 78% success rate, there is only an accuracy of 56% when predicting severe cases which is the primary goal of this model. The results coming from the rest of the algorithms used were slightly worse with K-Nearest Neighbors predicting 79% of minor cases and 53% of severe cases correctly, giving it a Jaccard score of 0.5379 and an f1-score of 0.6995 followed by the decision tree with 76% and 57% prediction accuracy on minor and severe cases respectively while its Jaccard score was 0.5355 and its f1-score was 0.6975. Logistic regression performed worse than the other algorithms with 77% of minor accidents classified correctly and only 53% of severe accidents classified as severe, this resulted in a Jaccard score of 0.5295 and an f1-score of 0.6924 while the log loss was quite high at 0.6046, another sign of its inability to deliver accurate predictions.

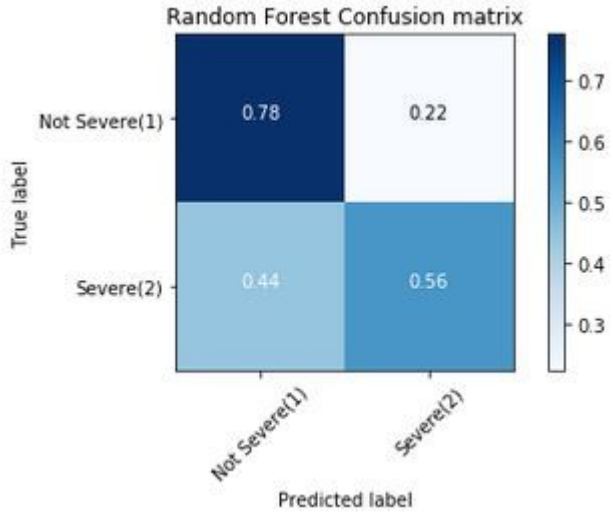
Table 1

Model	Jaccard Index	F1-Score	Log Loss
K Nearest Neighbors	0.5379	0.6995	-
Logistic Regression	0.5295	0.6924	0.6046
Decision Tree	0.5355	0.6975	-
Random Forest	0.5409	0.7020	-









5. Discussion

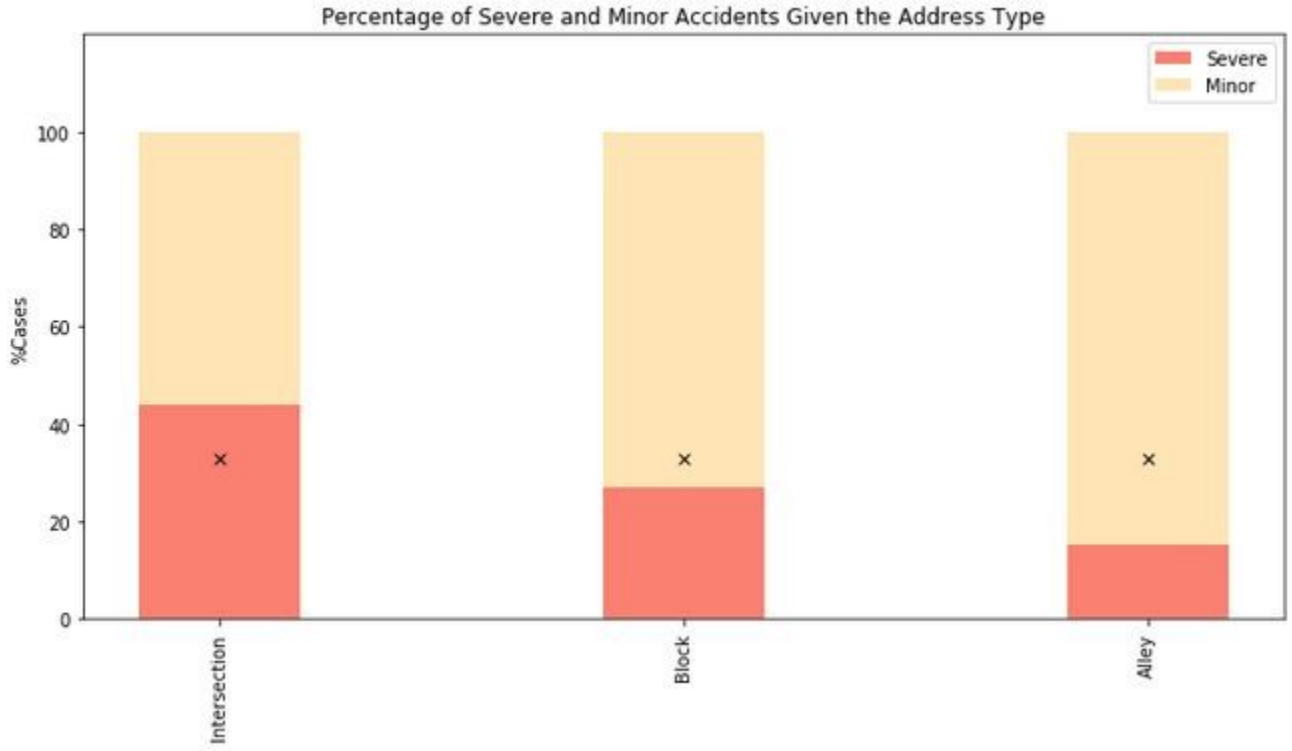
The goal of this research is to develop a model that can be used to warn drivers of the possibility of being involved in a severe accident in real-time. The ability of the final model to predict the severity of a possible car accident based on real-time generated data doesn't seem to be as accurate as one would expect. This could be due to the lack of more relevant data such as the type of the collision which is impossible to predict before an accident occurs. While the model can accurately predict some of the cases, overall it is not applicable yet and further research must be done using larger datasets and different algorithms such as deep neural networks which might be able to pick up patterns that simple machine learning algorithms such as the ones used here cannot.

6. Conclusion

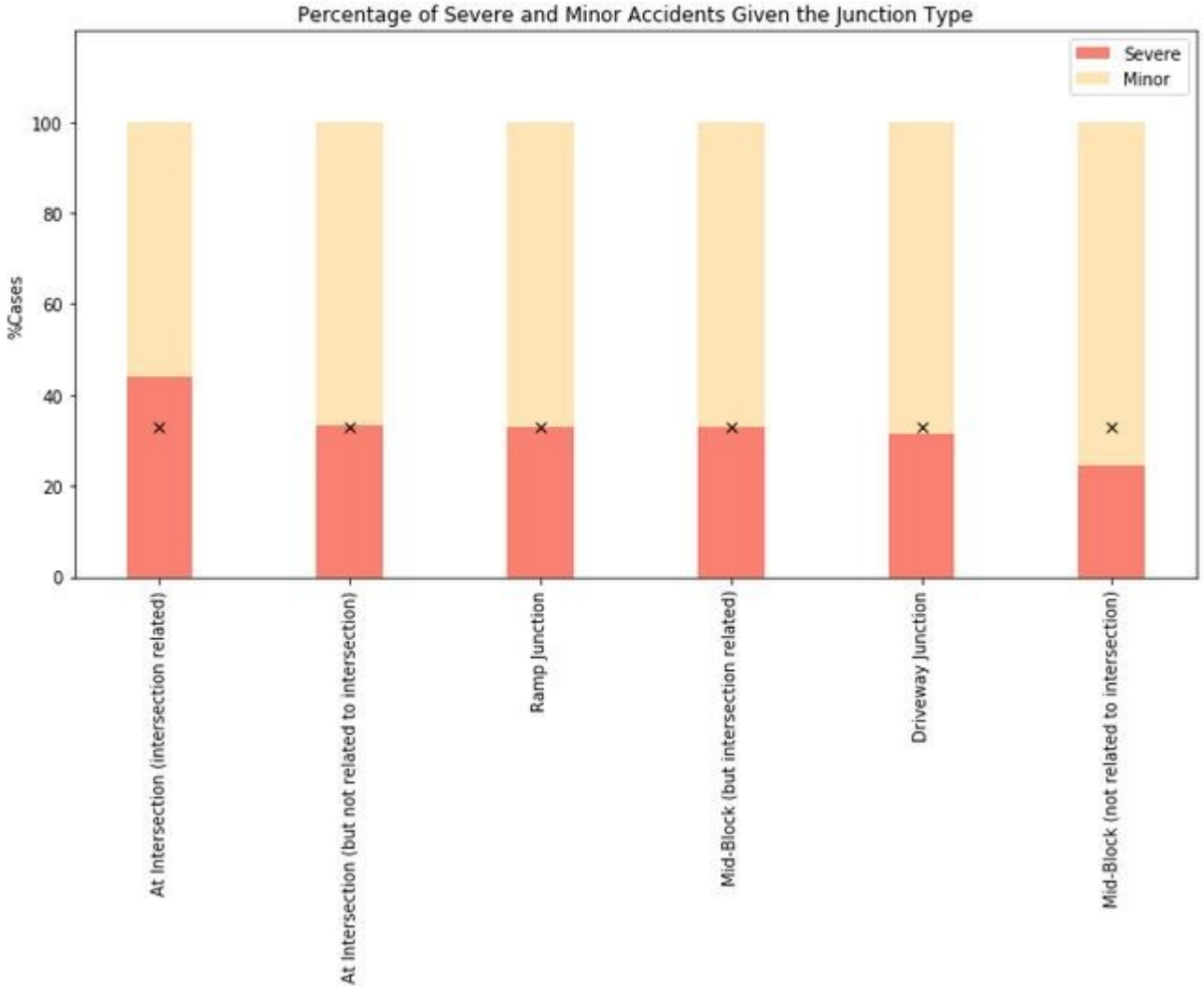
Road traffic accidents is one of the most common causes of death worldwide and therefore a serious problem governments must deal with. The goal of this report is to produce a real-time model for predicting the severity of probable accidents and warning the drivers in order to avoid certain locations or actions. Using data recorded in Seattle, an attempt was made to understand the different factors leading to serious accidents in order to develop a machine learning model that classifies possible accidents as serious or minor. The final model implements the random forest algorithm using the nine most significant features in accident severity prediction. The results showed some level of accuracy in predicting safe zones and therefore warning drivers for possible dangerous roads to avoid but further research must be done in order to develop an applicable model.

Appendix

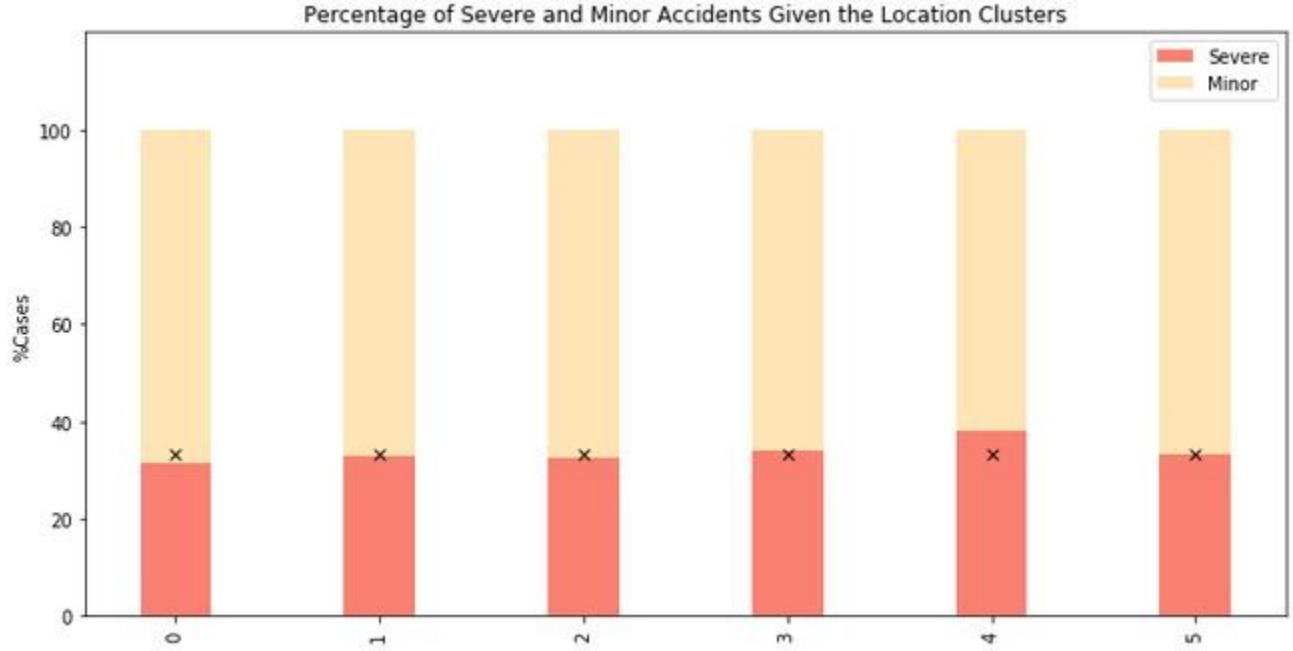
Graph 1.1



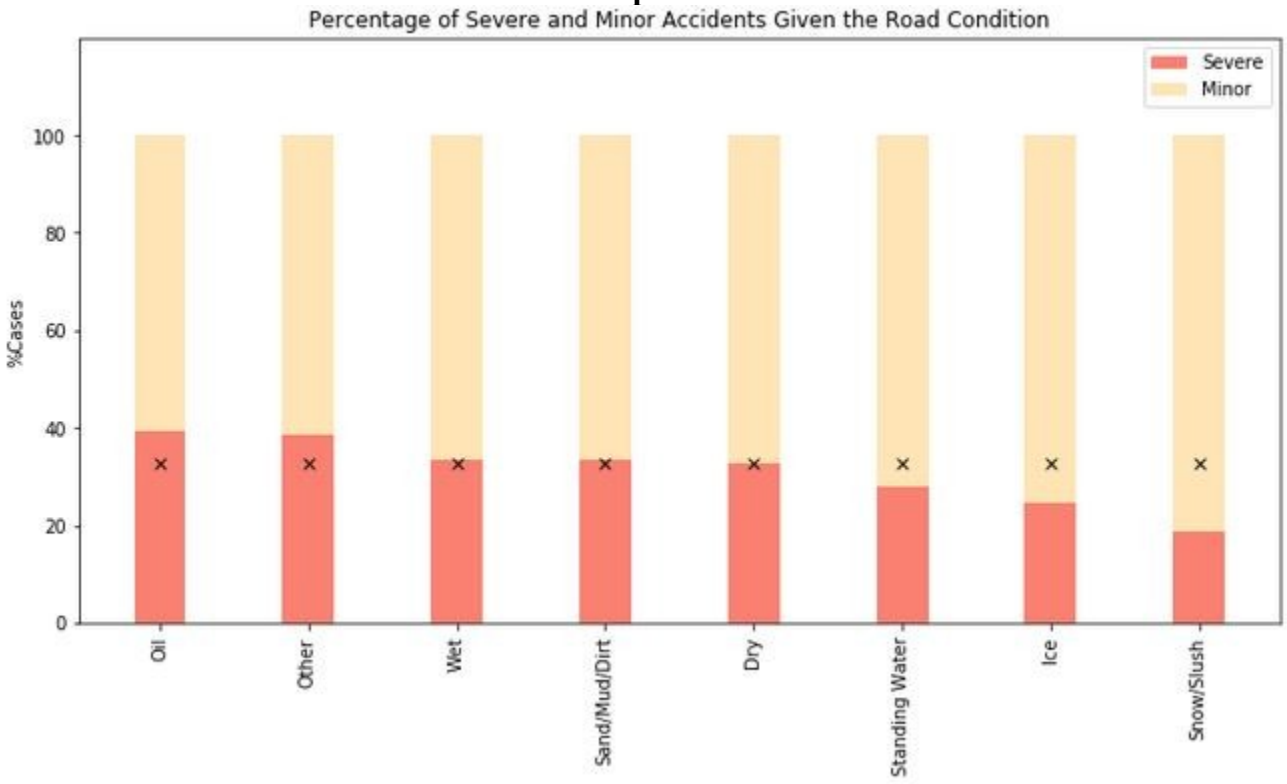
Graph 1.2



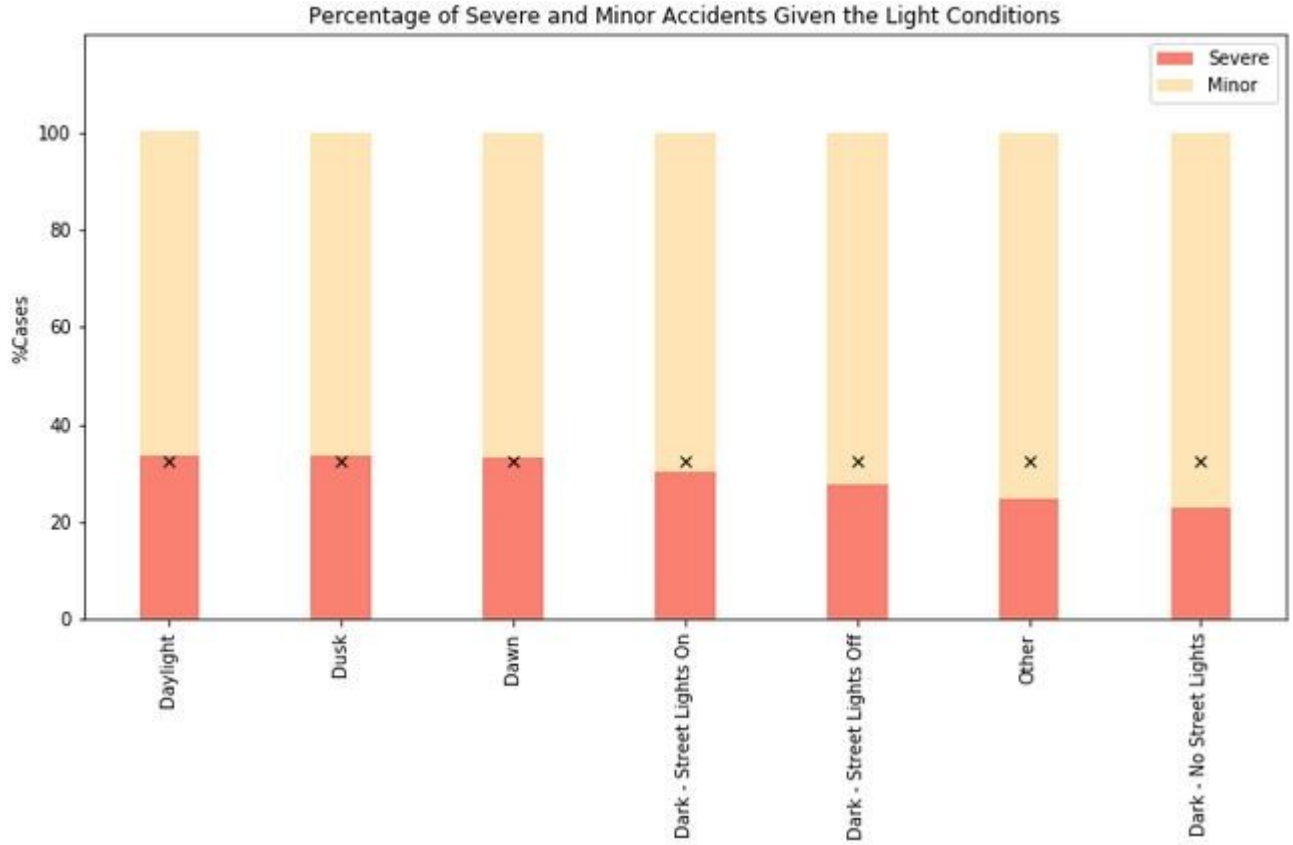
Graph 1.3



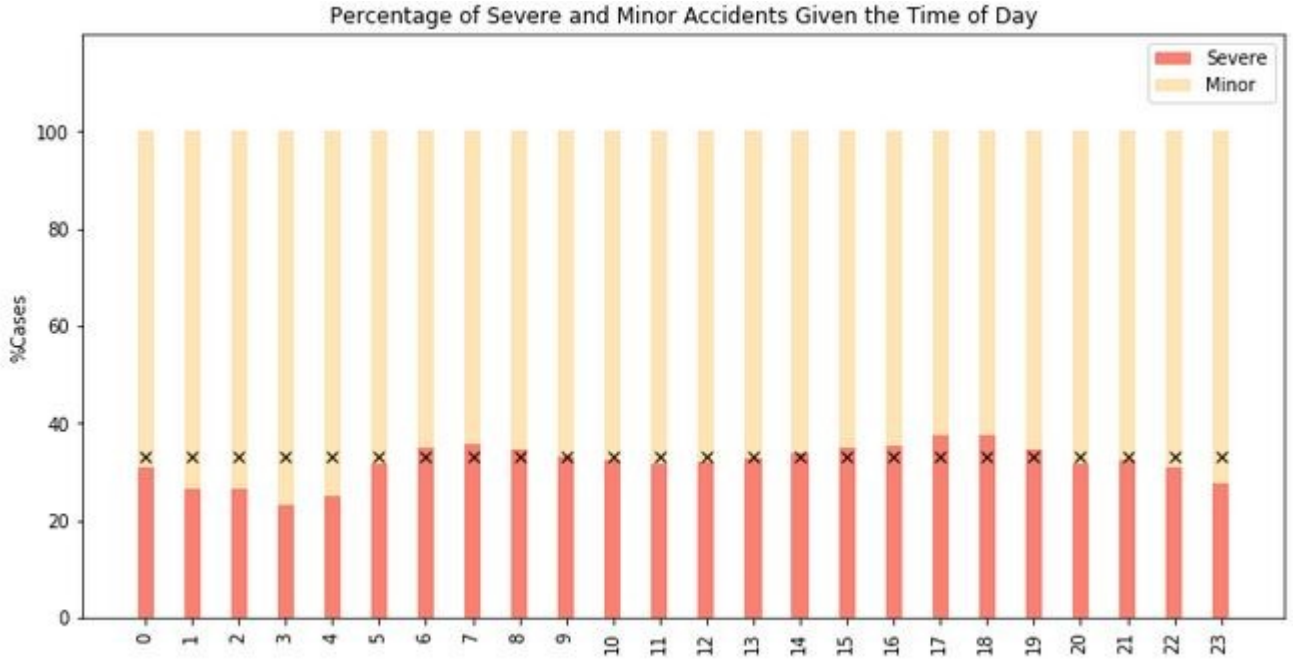
Graph 1.4


Percentage of Severe and Minor Accidents Given the Road Condition
Severe
Minor
%Cases
100
80
60
40
20
0
Oil
Other
Wet
Sand/Mud/Dirt
Dry
Standing Water
Ice
Snow/Slush

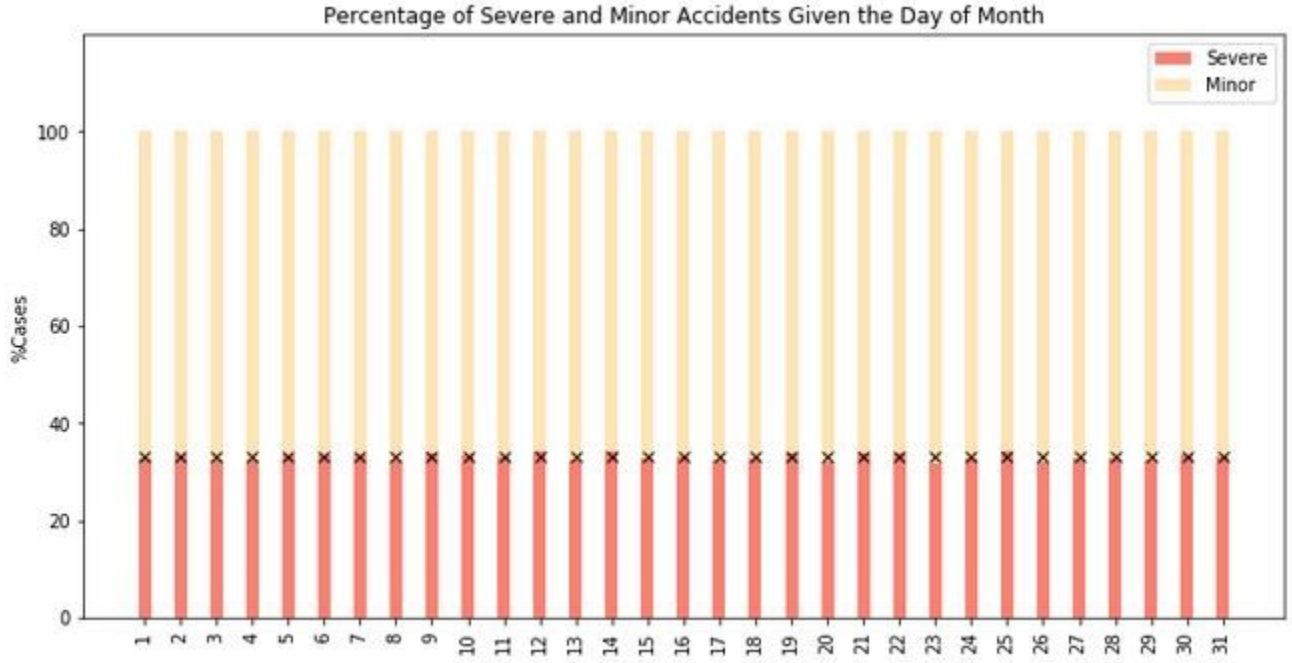
Graph 1.5



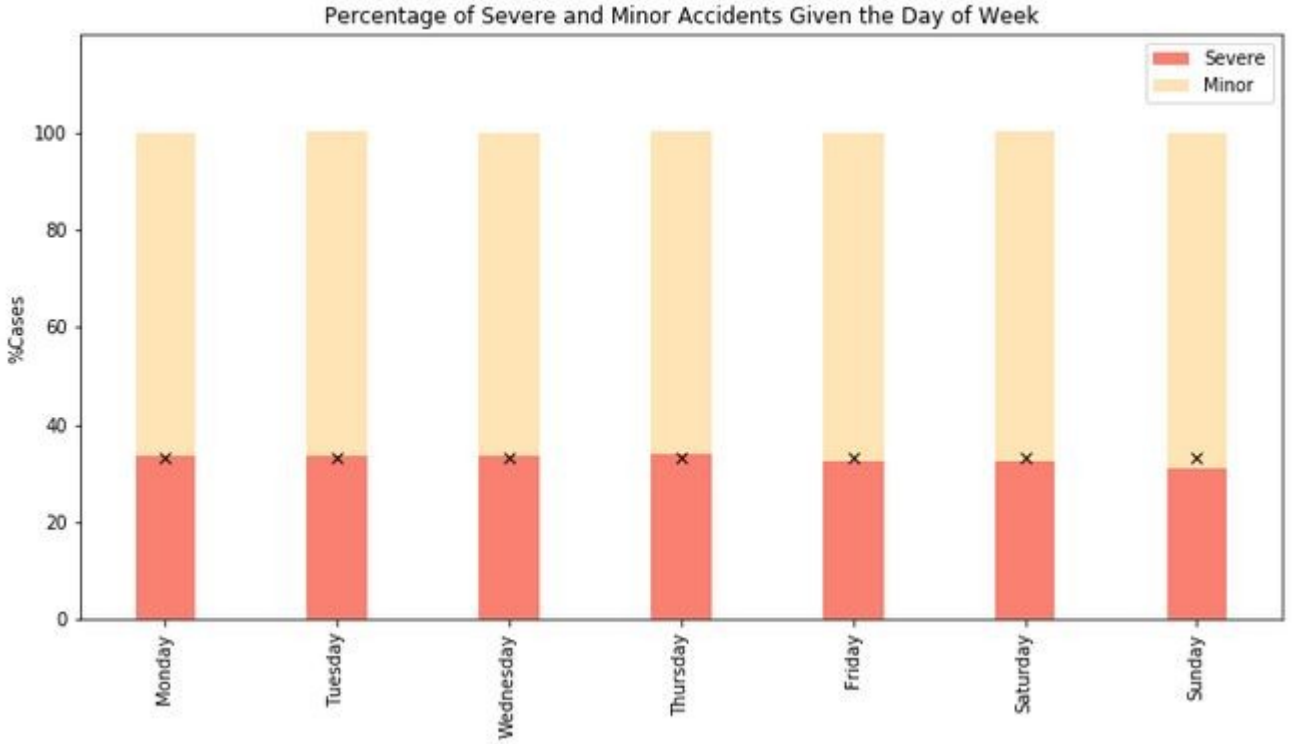
Graph 1.6



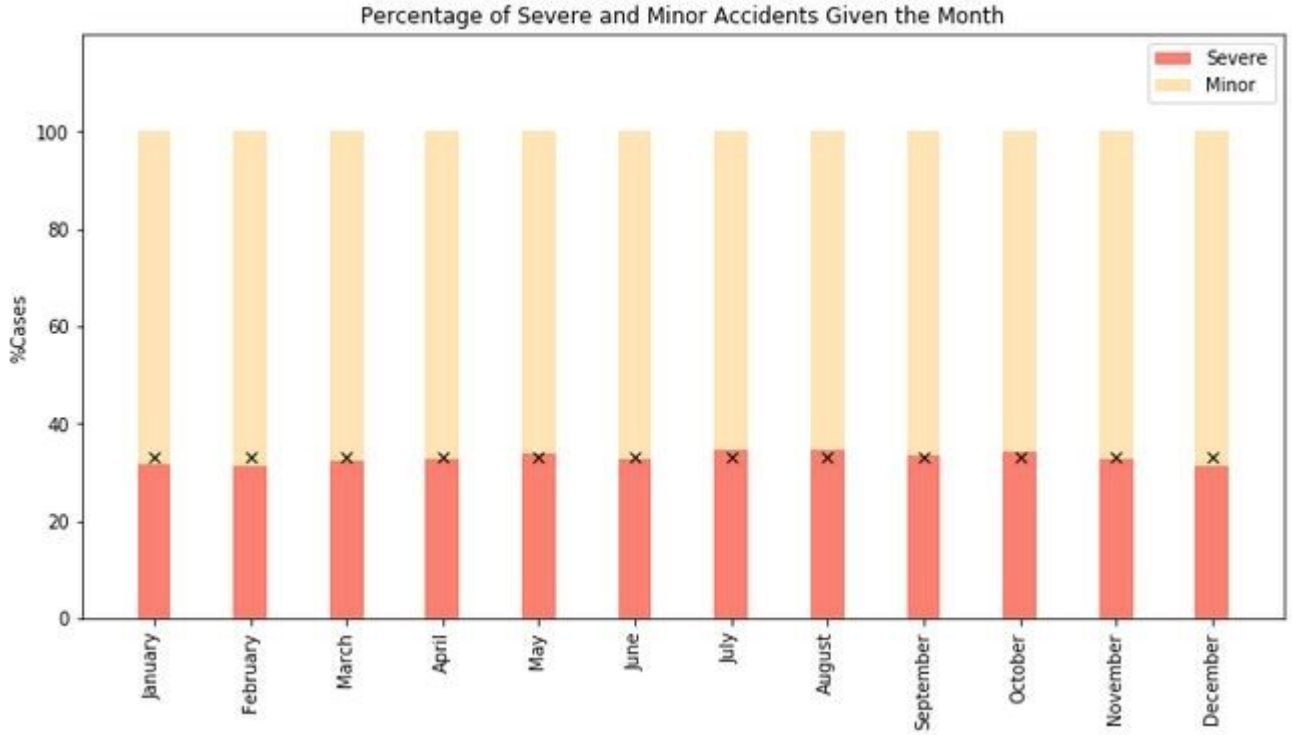
Graph 1.7



Graph 1.8

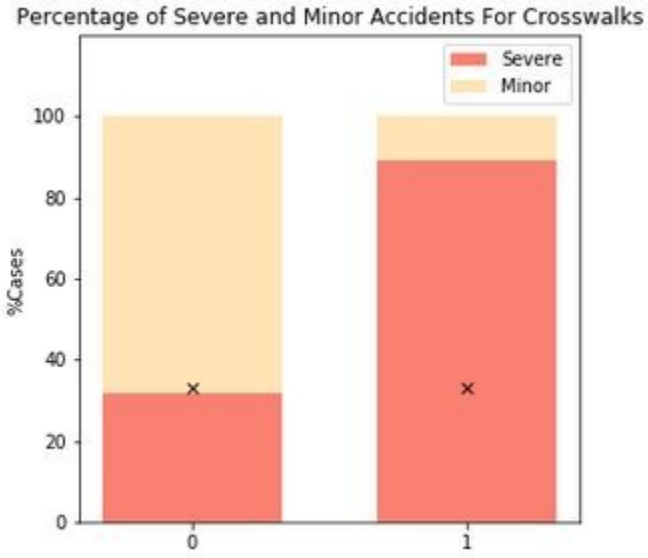


Graph 1.9

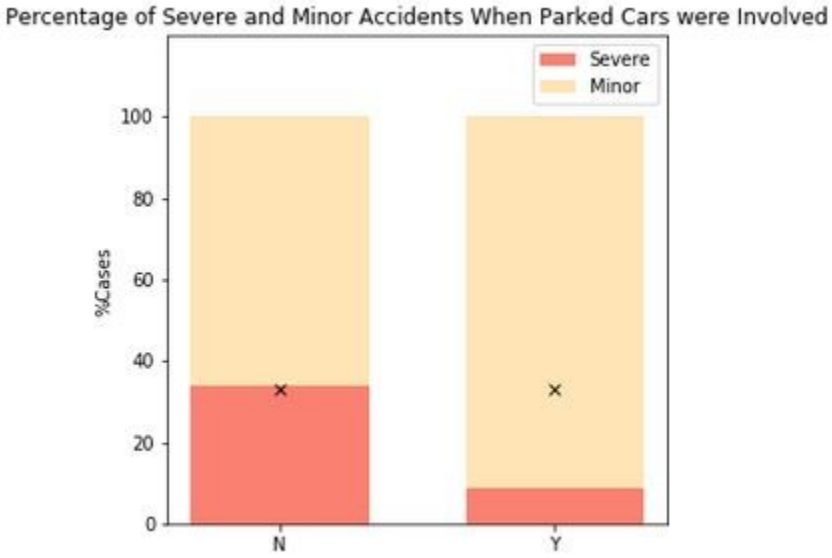


Graph 1.10

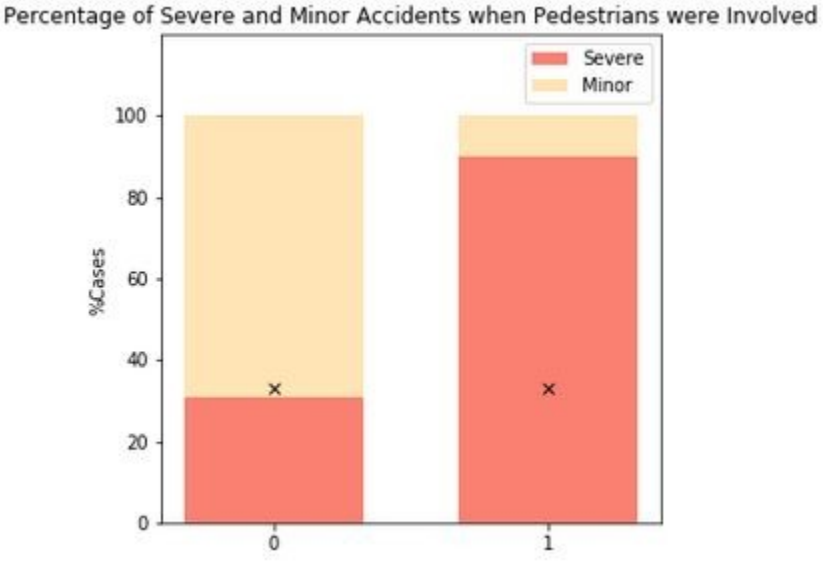
Graph 1.11



Graph 1.12



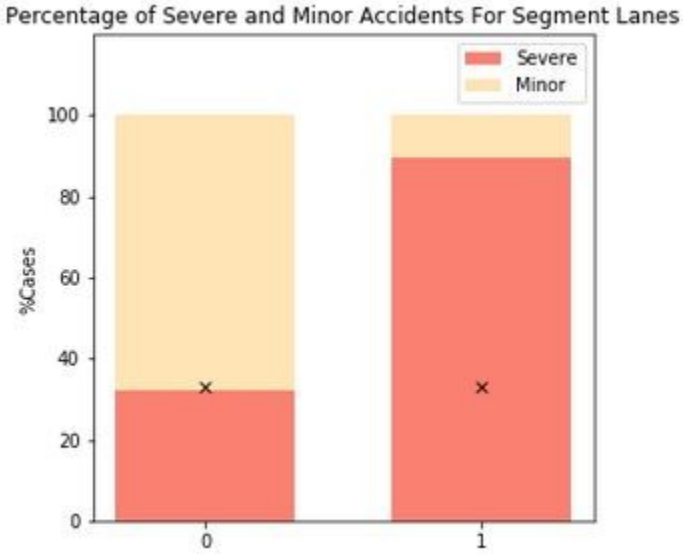
Graph 1.13



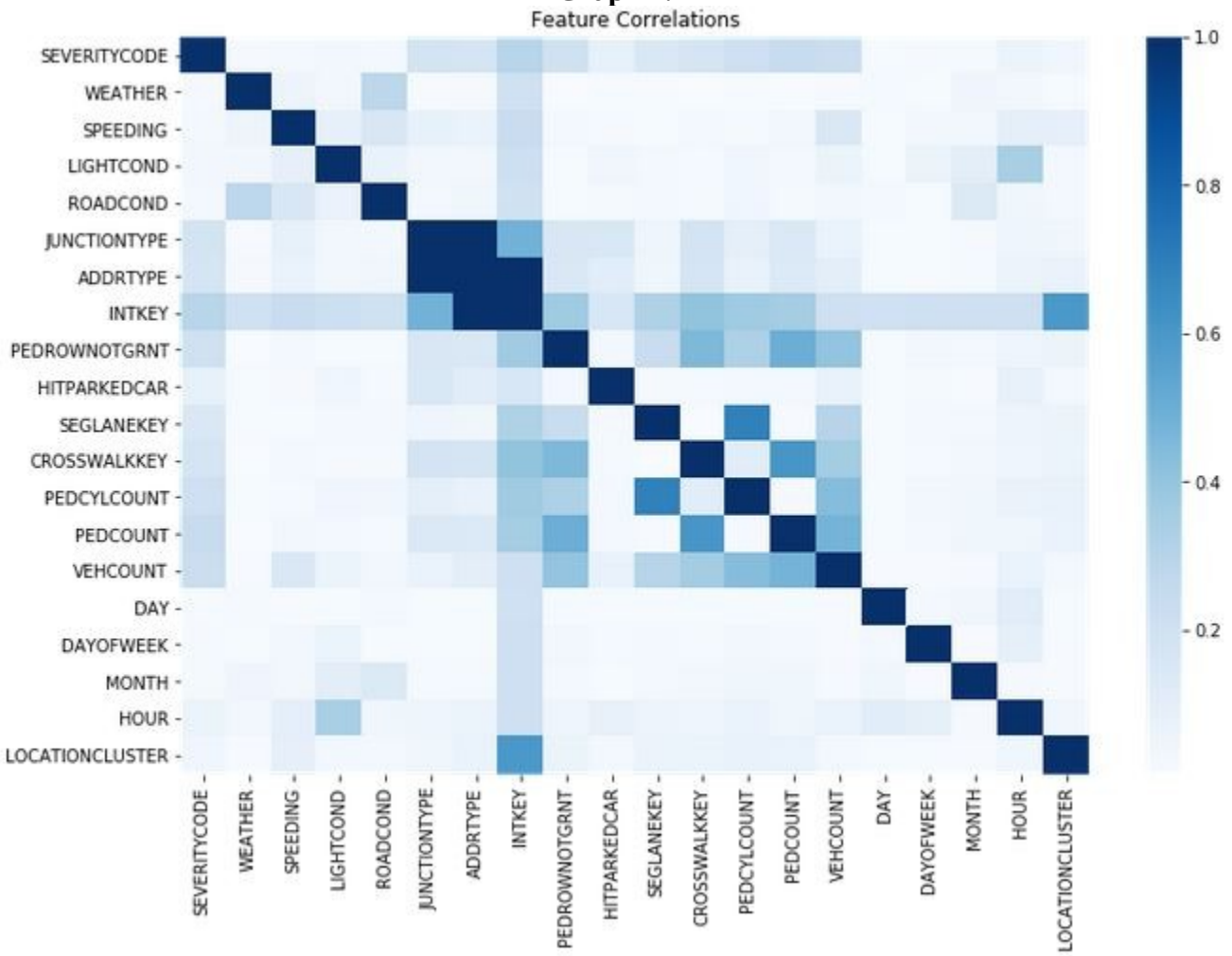
Graph 1.14

Graph 1.15

Graph 1.16



Graph 2.1


Feature Correlations
SEVERITYCODE
WEATHER
SPEEDING
LIGHTCOND
ROADCOND
JUNCTIONTYPE
ADDRTYPE
INTKEY
PEDROWNOTGRNT
HITPARKEDCAR
SEGLANEKEY
CROSSWALKKEY
PEDCYLCOUNT
PEDCOUNT
VEHCOUNT
DAY
DAYOFWEEK
MONTH
HOUR
LOCATIONCLUSTER
1.0
0.8
0.6
0.4
0.2

Map 1: Centroids of the Location Clusters