NYC Vehicle Crashes

Anuj Chheda, Ruzan Sasuri

ABSTRACT

In the modern era, vehicles have become an integral part of our society. It allows people to move relatively large distances in relatively short amounts of time. However as is with most advantages there are also many disadvantages. The most significant disadvantage of having so many vehicles is that they are driven by humans who are prone to error. These errors can lead to significant catastrophes at the speeds at which vehicles travel, including property damage, injury, and death. It is due to this that it is becoming vital to study all kinds of vehicular accidents and use the information gained by these studies to make vehicular travel safer.

Using data analytics, vehicle manufacturers can build safer vehicles; lawmakers can create better infrastructure and vehicle rules to decrease the number of vehicular accidents. The information from this analysis can also be used by the general public themselves as rules or guidelines to help save their own lives and the lives of other people around them while they drive a vehicle. This project uses the NYPD Motor Vehicle Collisions database from the NYC Open data website.

1. OVERVIEW

For our project, we are using the NYPD Motor Vehicle Collisions dataset from the NYC Open data website. We would like to find some information from this dataset that can help us understand the trends in vehicular collisions in New York City and find some information from this dataset that can help save people's money, property and more importantly their lives. We plan on using the dataset to find streets and corners where many accidents take place and use this information can help navigation applications to provide better routing information that is both fast and safe for a vehicle to travel through.

We plan to divide this dataset into training, testing, and validation which has 70/15/15% of the data respectively, and we will use the output from the validation dataset to measure the performance of our model. To evaluate the model we used, we will check the procedure and results of our model for any logical errors we made. We also need to clean our data and filter a few rows from our data to ensure a right balance between performance and accuracy.

In section 2 we will speak about related works from different sources of previous research done on the same topic. Section 3 will discuss the ethical concerns of our project. Section 4 will talk about the different possible business cases of our project. Section 5 speaks about the preprocessing work we will perform on the data such as cleaning, feature selection and using data mining methods. Section 6 will speak about the issues we had with the remaining data. Section 7 will speak about the various algorithms we plan to use on our data. The final sections will contain any extra information or interesting facts we can find through our data.

2. RELATED WORKS

2.1 Effects of winter precipitation on automobile collisions, injuries, and fatalities in the United States

Black, Alan et al[5] studied vehicular crash data from multiple cities across the United States of America to find the increase in the probability of vehicular crashes in winter due to precipitation(snow, snow grains, ice crystals, ice pellets, or freezing rain).

The author chose the cities to study based on the average amount of winter precipitation that occurs every year.

Rather than set a fixed time criteria like other papers had done, the author recorded the number of collisions that occurred during the period of the entire winter precipitation event. In this way, he managed to find differences in collisions between larger and smaller winter precipitation events.

The authors then prioritized snow over all other winter precipitation events. This would mean for example if snow and freezing rain occurred at the same time the precipitation event was categorized as snow. Any event that did not have snow was categorized as an ice event.

Using this data, the authors plotted the spatial distribution of the risk of collision due to winter precipitation on a map of the United States of America to visualize the risks. The results can be found in Figure 1.

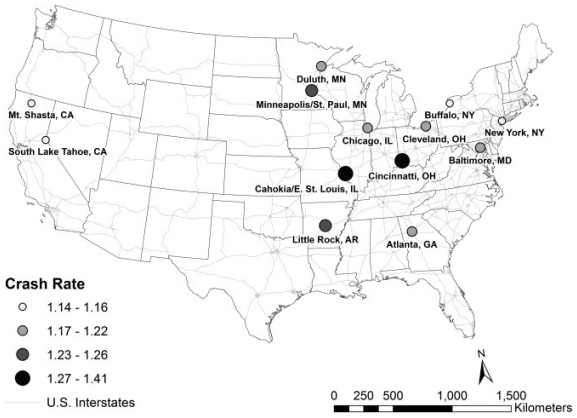

Figure 1: Location of the 13 study cities and the spatial distribution of the relative risk of winter precipitation related crash.[5]

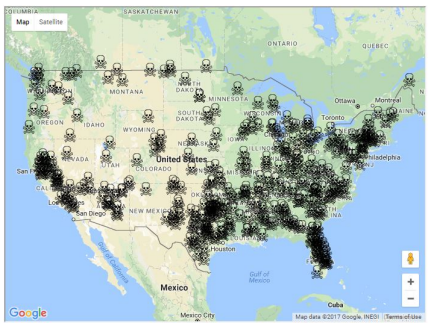

Figure 2: Traffic Incidents, having 3 or more fatalities.[6]

2.2 Traffic Fatalities – Getting Started

Kaggle user Mike Chirico analyzed the 2015 Traffic Fatalities database found on Kaggle, in his notebook called Traffic Fatalities – Getting Started[6]. He created a basic introduction to analyzing a dataset related to vehicular crashes. Considering the number of datasets contained in the database, the user was able to perform some in-depth study into the causes and risk relations for vehicular crashes across the United States of America. Using the availability of information, he mapped the cases of 3 or more fatalities occurring in a traffic collision as shown in Figure 2.

Using the data available he could also draw pie charts to show the relationship between drinking and driving and the weather to the number of fatalities as shown in Figure 3.

2.3 Traffic Fatalities in 2015

Kaggle user Abigail Larion analyzed the 2015 Traffic Fatalities database found on Kaggle, in her notebook called Traffic Fatalities in 2015[8]. She used the number of traffic fatalities in each state as the setting for the shade of red as

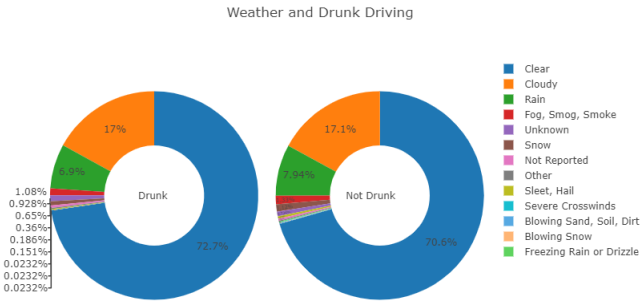

Figure 3: Pie graph for type of weather for drunk vs not drunk fatalities[6]

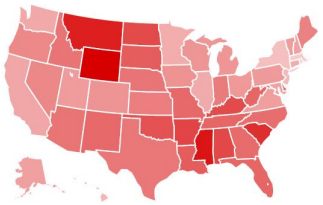

Figure 4: Traffic Fatalities per 100,000 People in United States (2015) [8]

shown in figure 4. In this way, she showed us that the states in the northwest and southeast parts of the United States of America had the darkest colorings of all the states. This shows that these states reported the highest traffic fatalities in 2015. Wyoming was the state with the darkest shade of red. This means that Wyoming was the leading state when it came to traffic fatalities in 2015.

2.4 Traffic accident segmentation by means of latent class clustering

Depaire, Benoit et al[7] found that traffic accident data often takes all types of traffic accidents into consideration. This leads to inevitable relationships between attributes to be latent. He mentions that Vaent, Francesca et al[11] found that Sundays and holidays lead to a higher injury rate for truck drivers than other days. However, when analyzing the entire dataset of accidents, such a relation was not easily deduced. Furthermore, Ulfarsson and Mannering[10] in their research concluded that a seemingly insignificant attribute in a dataset could become significant by splitting the data based on the type of accident.

To avoid this, the authors proposed a method of clustering the data to split the data into various types of accidents that can make the possible relationships more apparent, rather than always using domain knowledge to split the data based

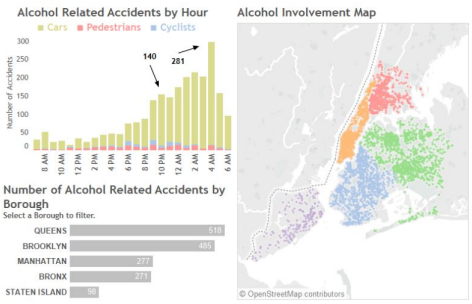

Figure 5: Graph for time of alcohol related collisions[9]

on the type of accident. The authors cluster the data such that the homogeneity of each cluster is high, but the heterogeneity of all the cluster prototypes is high.

2.5 Exploring NYC Vehicle Crash Data in Tableau

Maureen O'Donnell used the dataset found on OpenNYC to find relationships between various accident influencing factors such as time of the day, level of intoxication, weather conditions, and the number of reported accidents due to these contributing factors in her article Exploring NYC Vehicle Crash Data in Tableau[9]. By analyzing the data in a bar graph with the relation of time and vehicle to the number of collisions as shown in figure 5, she found that most accidents occur between 10 pm and 4 am. This might be due to lower visibility and sleepiness. The reason can also be used to prove that the more people consume alcohol before and during these hours.
In the article, author groups the 48 different contributing factors into six different groups and provides analysis about which types of reasons are responsible for more number of accidents. The most notable among those reasons are distractions and physical ailments contributing to approximately 2/3rd of all the accidents. The author also discusses how the day of the week might affect the number of accidents occurring which we will talk about later in the report. They also discuss the number of accidents happening at each hour of the day. One might assume that most of the accidents might happen after it goes dark, later in the evening, but the contrary is true. The number of accidents grows steadily after 10 am, and peaks at 4 pm and then the frequency starts dropping till 4 am.

2.6 Traffic Safety Facts

NHTSA's National Center for Statistics and Analysis produced a report[2] with facts, figures and trends related to motor vehicle crashed in 2015. The most important feature of this article is a small graph included in the article which plots the correlation between various factors such as Vehicular Miles Traveled(VMT), average monthly temperature, new passenger vehicle registration, average gas prices and U.S. unemployment, and the number of traffic fatalities. From the graph in fig 6, we can observe some contributing factors to the rise of Vehicular collisions:

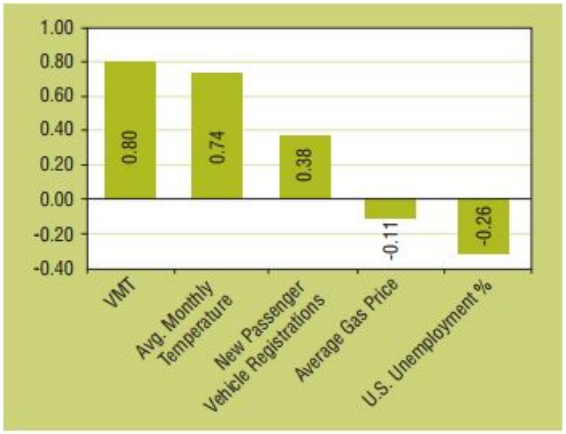

Figure 6: 5-Year Correlation Between Fatalities and Economic and Other Indicators, 2011âĂŞ2015[2]

1. As the total number of miles traveled by all registered cars increases so does the fatality. This is because the more a car travels, the less safe it gets due to general wear and tears.

2. As the average monthly temperature increase, the number of fatalities increases. This is because the temperature affects the mood of the driver and increases the chances of tire bursts.

3. As the number of new cars registered increases, the number of fatalities increase. This is because of the number of cars on the streets or rather the number of new drivers on the streets increases, which leads to higher chance of collision.

There are also some factors that lead to a decrease in the number of vehicular collision fatalities:

1. As the gas price increases, the number of fatalities decreases. This is because gas becomes more expensive, hence fewer people tend to buy gas(or drive).

2. As the unemployment percent increases, the number of fatalities decreases. This is because the number of people who can afford cars and gas also decreases.

2.7 Independent Statistical Analysis

In his independent analysis on the NYC motor collisions data from Jan to April 2015, Mr. Adhokshaja calculates the frequency of each of the causes leading to injuring or killing a person as we can see in figure 7. We can see that the leading causes of accidents are failure to yield the right way, driver

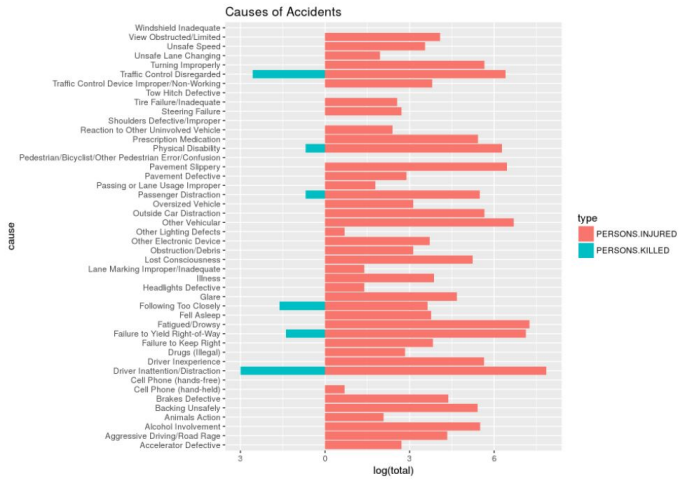

Figure 7: Causes for Accidents[4]

drowsiness, turning improperly, traffic control disregard and driver inattention. This means that just by following the rules, we can avoid the majority of these accidents. Failure to yield right way, driver inattention and traffic disregard are also the leading causes of accidents resulting in death. This metric shows that by providing stricter laws and spreading more awareness might lead to considerable decreases in the number of accidents occurring in New York City.

3. ETHICAL CONCERNS

As we have shown earlier, the only ethical concern is the misinterpretation or misrepresentation of the analysis that our project has conducted on the data. This can lead to problems if performed by men with legislative authority. For example, someone can easily misrepresent the data found in Figure 3, ad state that the clearer the sky is the more likely you're to crash, but conveniently leave out the fact that there will be more vehicles on the road when the sky is clear.

4. BUSINESS CASE

By analyzing vehicular collision data, we can find out the leading causes of accidents, accident hot spots. Emergency services need to respond to an accident with promptness and accuracy. We can say then say that emergency services would be one of the primary industries which need this kind of information. Using this information, they can decrease response time by knowing the location and time hot spots of collision and can provide faster treatment to those injured in a collision by predicting what caused the collision or the outcome of the collision.

Traffic and road regulators can also benefit from this kind of analysis. Legislators can construct better traffic laws to decrease the number of collisions or at the very least decrease the number of fatalities that occur in a collision. Navigation

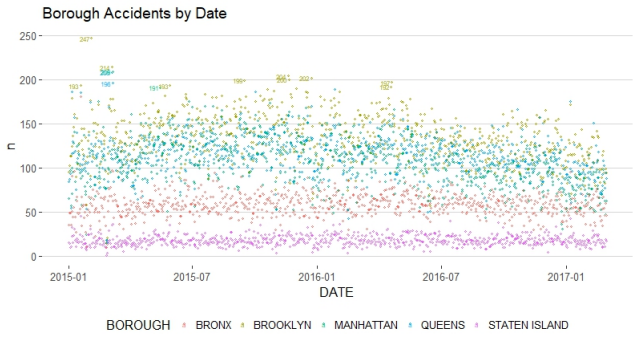

Figure 8: Time plot for the number of collisions every day in each borough

applications can also be improved using this analysis by recommending safer routes to follow at certain times of the day or for certain weather conditions. Unfortunately, this business case is a double-edged sword as it will lower the number of collisions and fatalities taking place which decreases the amount of data available for research and analysis.

From our database, we have been able to find that there exists a relation between the street name and the type of vehicles involved in the collisions around that street. Also, there is a correlation between the type of the vehicle and the factor that the vehicle plays in the collision. This information might be used by the insurance companies to charge higher premiums to people working or living in on related streets or driving these vehicles. There are ethical ramifications for usage for

5. PRE-PROCESSING

This data is obtained from the NYC OpenData. This is part of a VisionZero initiative started by the NYPD to reduce the number of deaths and ultimately injuries caused by motor vehicle collisions. The data obtained had recorded collisions starting from mid-2014 till three days ago. All the collisions occurring in the NYC area is being uploaded every weekday and is uploaded the website three days after the collision. Initially, we decided to perform our analysis over a 2 year and 2 month period starting from 1St, Jan 2015 until 28th Feb 2017. After finishing up an initial analysis and visualizing the data, we decided to perform our initial clustering tests in the two months with the highest amount of collisions as can be seen from figure 8.

Initial analysis of the number of collision per population of each of the burrows would yield figure 9 on the chosen data. NYC has only five boroughs and figures 10 reveals that it has six boroughs. On analysis, we find that lot of the data is missing from the database.

5.1 Data Cleaning

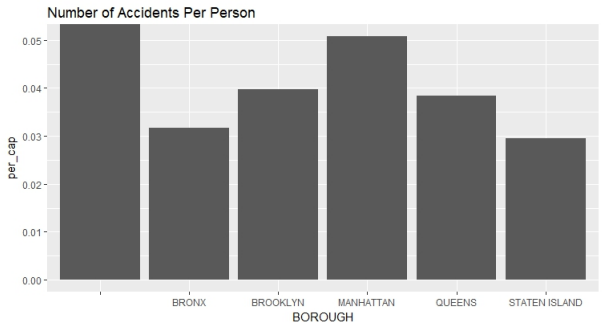

Figure 9: Plot for the number of collisions per population in each borough

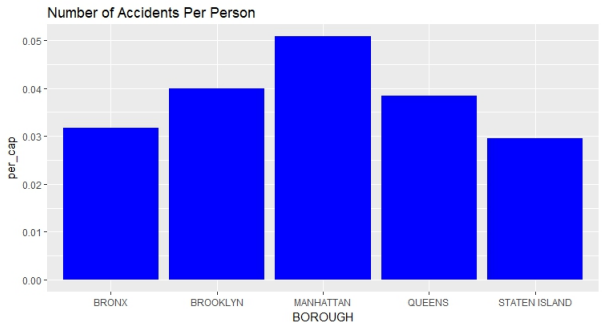

Figure 10: Plot for the number of collisions per population in each borough

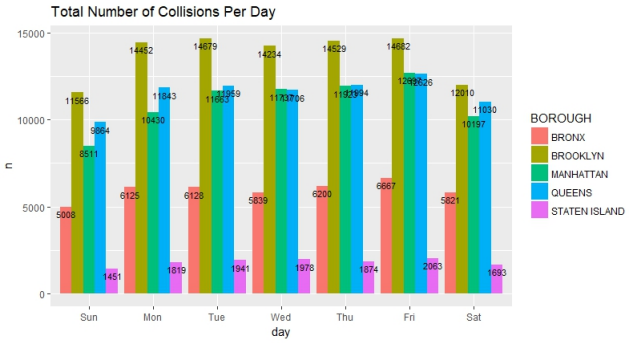

Figure 11: Number of collisions on every day in each borough

As the data was obtained directly from the government website, it was not clean and had a lot of missing values, bad values, and irrelevant attributes. There are different ways to work around this problem. One of the approaches we followed was to use the zip codes to predict which borough. Afterwards, we ended up with many records which would not have zip-code along with the borough name from the record entry. We decided to use latitude and longitude with K-nearest neighbors to predict which borough this record would belong to but some of the records would not even have latitude and longitude recorded in the data. This was a tiresome process with not much change in the number of records that were not labeled for the borough.
Hence, we started by removing only those items which affected the analysis and then we would restore the values. After finishing the initial analysis, we removed all the rows which had any of the attributes missing in the dataset. This reduced the size of our observations from 477K to just over 300K. After cleaning the missing values, and performing the visualizations to find patterns visible. After visualizing the data for two months, we found bad values among some of the attributes. E.g.: Some of the data records has missing Latitude and Longitude values of where the collision occurs. After removing those records, some other records had 0 as latitude and 0 as longitude in the data. This would indicate that some of the accidents in NYC were happening in the Gulf of South Africa. After removing those records, we did not encounter any bad data from other records till now.
After removing all the bad and missing data, we would get the following graph for the number of collisions per population in each borough.

5.2 Attribute Selection

Our data has 29 attributes including the unique key for each of the data record. Initial analysis of the data is performed on a reduced dataset of two months (March and April 2015). There is no inherent structure to this data that would allow us to cluster or classify the data. Hence, we introduced our classification to this dataset. Depending on whether a person was injured or killed or no harm was done, we divided the dataset into three groups (KILLED, INJURED and SAFE) and classified the whole data on this condition. We calculated OneR and K-means clustering rule for among all the attributes. All the initial rules were highly obscure and would select the attribute which would have all the unique values and would yield a data classification accuracy of over 95%. Along with removing Unique Key, we had to remove the date of the accident because that value spans over a more extensive range. Hence, we reconstructed the date and introduced two new features, a month and day of the week. This would provide us insight into how the day of the week would affect the number of collisions going up or down. To show this, we created this graph as shown in figure 11.
After initial analysis, we started removing all the attributes which would either not contribute towards the prediction of whether given the conditions would result in a loss of life or injury to a person. Among the attributes we removed were Off street, on street and cross street names. Due to lack of documentation, it was difficult for us to interpret what these three street names signified in the junction at which the collision took place. However, from the covariance matrix in figure 12, we can see that the type of the vehicle is dependent on the factor that car played in the collision.

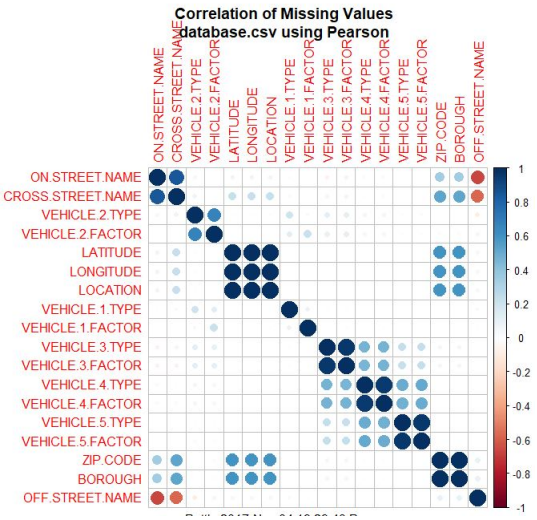

Figure 12: Covariance Matrix

From the lack of values or missing data, there is no reliable correlation between them. Due to this reason, we decided to remove these attributes from consideration for the classification. Similarly, we can also observe that there is a relation between the street names and the type of car involved in the collision. Hence, we can look at the street name and predict the kind of vehicles driving in the vicinity and predict which type of accident is probably in that region and how many cars might be involved and what are the chances of a person getting hurt.

To perform K-Nearest Neighbors Classification and RandomForest Classification we used Scikit-Learn algorithms. A drawback of Scikit-Learn algorithms is that most built-in algorithms cannot process categorical values. The ideal way to deal with this problem is to employ an encoding scheme called the OneHotEncoding Scheme. Given below is a simple example to illustrate the working of the algorithm:

Name
A
B
C

After Encoding:

A	B	C
1	0	0
0	1	0
0	0	1

The final features we selected are as follows:

1. **ID:** This is the ID of the incident as recorded in the original dataset. The ID is ignored while running any classification or clustering algorithm as it does not provide any value to classification algorithms.

2. **Day:** This is the day of the week on which the collision took place. Here 0 is Monday, and 6 is Sunday. Python's datetime package contains functions to find the day of the week for a particular date in any standard date format. This feature is One-Hot encoded since the days of the week are cyclic(i.e., Monday comes after Sunday), and one can not say that "Tuesday" > "Monday."

3. **Time:** This is the hour number in the 24-hour format during which the collision took place. It is a set of numbers from 0 to 23. This feature is One-Hot encoded as the hour number is cyclic(i.e., 0 comes after 23) and saying that 3 pm < 4 pm gives us no real information.

4. **Borough:** This stores the index number of the borough in which the collision took place. We arrange the Boroughs randomly as follows:

   (a) 0 = Manhattan

   (b) 1 = The Bronx

   (c) 2 = Queens

   (d) 3 = Brooklyn

   (e) 4 = Staten Island

   This feature is one-Hot encoded as its value is decided randomly as mentioned above and we can not say that 3 > 0(i.e., "Brooklyn" > "Manhattan).

5. **Latitude:** This stores the latitude(rounded to the 7th decimal place) at which the collision took place. This feature is directly taken from the original dataset.

6. **Longitude:** This stores the longitude(rounded to the 7th decimal place) at which the collision took place. This feature is directly taken from the original dataset.

7. **Class:** As mentioned above, the class is decided based on the harm to any persons involved in the collision as given in the original dataset. The classes are assigned as follows:

   (a) KILLED(2) = If at least one person involved in the collision was killed.

   (b) INJURED(1) = If none of the people involved in the collision was killed, but at least one person involved was injured.

   (c) SAFE(0) = If none of the people involved in the collision was killed or injured.

These features after One-hot encoding is used as the complete dataset in the K-Nearest Neighbors Classifier and Random Forest Classifier algorithms.

6. ALGORITHMS

For the classification algorithms, we decided whether the collision led to "KILLED"(2), "INJURED"(1) or "SAFE"(0).

Table 1: Confusion Matrix for the 0R rule

	Safe	Injured	Killed
Safe	76209	0	0
Injured	16406	0	0
Killed	67	0	0

Table 2: Confusion Matrix for the 1R rule

	Safe	Injured	Killed
Safe	76139	70	0
Injured	16166	240	0
Killed	65	2	0

6.1 0R Rule

The 0R rule performs classification based on the mode of the classifier. In the case of our data, the OR rule or the mode of the classes is 0, or the 0R rule show us that in most accidents, the people involved will neither be injured nor will there be fatalities. This was found using WEKA.
The accuracy of the 0R rule came to 82.23%. the confusion matrix can be seen in Table 1.

6.2 1R Rule

The 1R rule performs classification by only considering one attribute in the dataset other than the class. The attribute to be considered is chosen as the model with the least error. The attribute chosen for the 1R rule in our dataset is "VEHICLE.2.TYPE" as chosen by WEKA. The 1R rule is as follows:
PASSENGER VEHICLE ⇒ SAFE
UNKNOWN ⇒ SAFE
SPORT UTILITY/STATION WAGON ⇒ SAFE
BUS ⇒ SAFE
VAN ⇒ SAFE
TAXI ⇒ SAFE
BICYCLE ⇒ INJURED
OTHER ⇒ SAFE
PICK-UP TRUCK ⇒ SAFE
SMALL COM VEH(4 TIRES) ⇒ SAFE
LIVERY VEHICLE ⇒ SAFE
LARGE COM VEH(6 OR MORE TIRES) ⇒ SAFE
MOTORCYCLE ⇒ SAFE
AMBULANCE ⇒ SAFE
FIRE TRUCK ⇒ SAFE
SCOOTER ⇒ INJURED
PEDICAB ⇒ SAFE
These rules show us that even using the 1R rule we still can not classify the instance as causing a loss of life if we only consider the type of the 2nd vehicle involved in the crash.
The accuracy of the 1R rule came to 82.41%. The confusion matrix can be seen in Table 2.

6.3 Decision Stumps

A decision stump is a single decision node or an individual decision that the classifier needs to make to classify the result, which gives the minimum error. Since the best attribute using 1R rule mentioned above is "VEHICLE.2.TYPE" as chosen by WEKA, WEKA will choose the same attribute for its decision stump. We used 10-fold cross-validation and 80/20 split to validate the decision stump.
WEKA's decision stump showed us that if the "VEHICLE.2.TYPE" is a bicycle, the probability of injury is great. If not a bicycle, the probability that a person is safe is high. If there was only one vehicle involved in the crash, the probability of being safe or injured is almost the same. In all 3 cases, the probability that the person will die is the lowest.
We ran WEKA's decision stump once more, and it showed us that if the "VEHICLE.2.FACTOR" is "PASSENGER DISTRACTION," the probability of injury is great. If not a "PASSENGER DISTRACTION," the probability that a person is safe is high. If there was only one vehicle involved in the crash, the likelihood of being safe or injured is almost the same. In all 3 cases, the probability that the person will die is the lowest.
With the second run of the decision stump(which makes more sense), the accuracy came to 82.26%. The confusion matrix can be seen in Table 3.

Table 3: Confusion Matrix for a decision stump

	Safe	Injured	Killed
Safe	76150	59	0
Injured	16314	92	0
Killed	264	0	0

6.4 J-Rip

J-Rip is a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP[3]. This algorithm is used by WEKA to create streamlined decision trees. WEKA produces a JRIP model which is almost the same as the decision stump it produced. That is to say that the JRip model decides the class based on whether or not the "VEHICLE.2.TYPE" is a bicycle. If it is a bicycle, it classifies it as "INJURED." If it is not a bicycle, it classifies it as "SAFE."
The accuracy of the JRip rule came to 82.17%. The confusion matrix can be seen in Table 4.

Table 4: Confusion Matrix for JRip

	Safe	Injured	Killed
Safe	76047	162	0
Injured	16293	113	0
Killed	67	0	0

6.5 K-means

K-Means is a method of vector quantization for finding the prototype of clusters where the number of clusters is k or the number inputted by the user. K-means starts with k random cluster centroids. It then allocates all the points to the nearest centroid. It then calculates the new cluster centroid. Once that is done it continues to do the last two steps till either the change in cluster centroids is too little to matter, or a certain number of iteration limit is reached. We chose the Manhattan distance for this algorithm.
We fed our dataset into WEKA and using K-means where k=5(the number of boroughs). Using this, our cluster distribution comes to:

1. Cluster 0: 23968 instances(26% of the total instances).
2. Cluster 1: 24196 instances(26% of the total instances).
3. Cluster 2: 11592 instances(13% of the total instances).
4. Cluster 3: 27474 instances(30% of the total instances).
5. Cluster 4: 5452 instances(6% of the total instances).

6.6 Hierarchical Clustering

We performed hierarchical clustering on the data of a single month after removing all the rows which had missing values using R. The number of records for the grouping was 27000 and as we can see from the dendrogram generated that there are four visible clusters inherent to the data. The same can be concluded from the earlier clustering example of K means which created almost equal sized bins when the value of K was set to 5. Since the dendrogram is so dense, we can not see the point when k=5. Still, we can verify that 5 is the best k value as it is the number of boroughs and when we use k=5 in the K-Means algorithm, we find cluster prototypes in which each prototype represents a district.

6.7 K-Nearest Neighbors Classifier

The k-Nearest Neighbors algorithm performs classification based on the mode of the classes of the k nearest points to the point to classify in the feature space.
We used Scikit-Learn's KNeighborsClassifier algorithm to fit our features into a model and use that model to predict whether a new collision would likely result in at least one death, no deaths but at least one injury, or no injury or death.

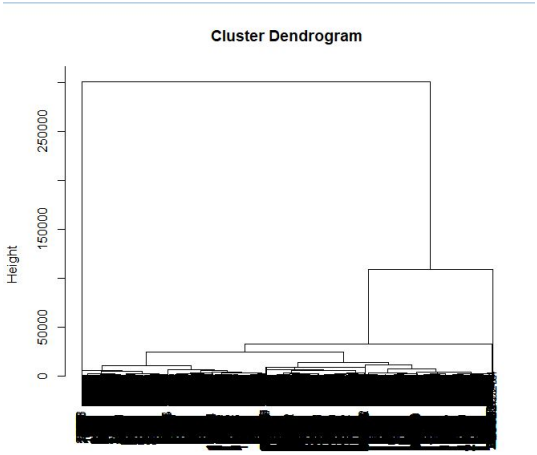

Figure 13: Dendrogram for month March

The KNeighborsClassifier algorithm requires a few parameters to fit the model. These are as follows:

1. **n_neighbors:** The number of neighbors to consider. This is the k-value of the algorithm. To ensure a good fit, we need to find the best k-value to use so that the model is not over-fit or under-fit. To figure this out, we tried for k values from 1 to 100 that are not multiples of 3. We did this to avoid breaking ties. The accuracy for each k-value was then mapped in a graph as shown in Figure 14. This shows us that the best k-value to use is ten as it is the point of diminishing returns(knee-point). It is set to this as any k value less than 13 gives a poor accuracy. Any k value greater than ten, on the other hand, does not significantly increase the accuracy of the classifier, but it requires many more computations than a k value of 10.
2. **weights:** We leave this to the default value of "uniform," i.e., no feature is given a greater weight-age than another.
3. **algorithm:** We set this to the default of "auto," i.e., the library will decide which algorithm to use based on the actual data added while fitting the model.
4. **p:** We set this to 1 as along with the default metric of "Minkowski" The model will use the L1 Minkowski Norm or the Manhattan Distance to measure any distances as required.
5. **metric:** We set this to the default value "Minkowski."

For the remaining parameters, we let the package decide the best values for the model.

We start our algorithm by splitting the data and classes into two parts. We use 70% of our dataset as the training

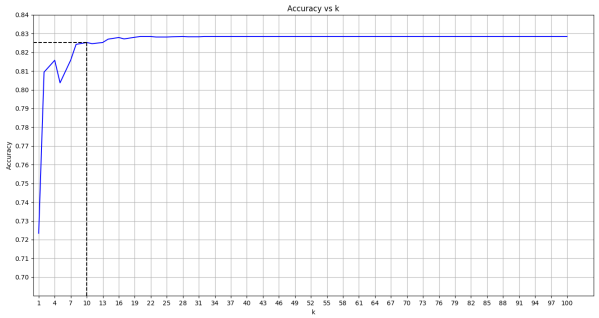

Figure 14: Accuracy vs. k for K-Nearest Neighbors

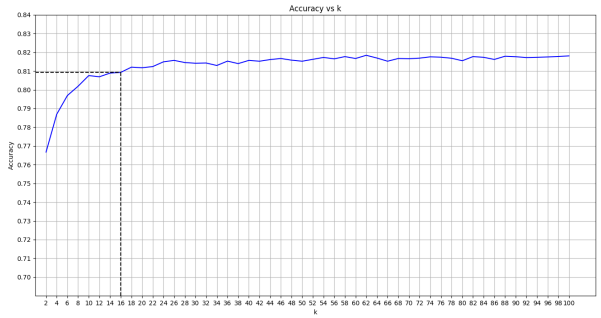

Figure 15: Accuracy vs. k for Random Forest

data and 30% of or data as test data. Then as mentioned above, we try the K-Nearest Neighbors algorithm for k values ranging from 1 to 100, ignoring all k that are multiples of 3 to break ties. For each k value, we then initialize the KNeighborsClassifier from Scikit-Learn. We fit the model using our training data and classes. We then use the model to predict the classes for our test data. If the expected value of the class is the same as the original class value, we add to our real positive count. The accuracy of a special k is then calculated as the number of true positives divided by the number of records. Finally, we plot a graph as shown in Figure 14.

Figure 14 contains a graph of the accuracy vs. the k value for all k values that are used. The blue line displays a knee curve, and the black dashed line shows us the point of diminishing returns.

To summarize the best model we could fit in K-Nearest Neighbors used a k value of 10 and gave us an accuracy of 0.825 or 82.5%.

6.8 Random Forest Classifier

The Random Forest Classifier algorithm is an ensemble learning method to build a classification model. This means that the Random Forest Classifier creates n(a preset number) decision tree classifiers, which classify the data based on its nodes. The algorithm then finds the mode of the n predicted classes and uses that value as the predicted class of the record.

We used Scikit-Learn's RandomForestClassifier algorithm to fit our features into a model and use that model to predict whether a new collision would likely result in at least one death, no deaths but at least one injury, or no injury or death.

The RandomForestClassifier algorithm requires a few parameters to fit the model. These are as follows:

1. **n_estimators:** The number of trees in the forest. This means that the algorithm will create k(the value of the parameter) number of trees. To ensure a good fit, we need to find the best k-value to use so that the model is not over-fit or under-fit. To figure this out, we tried for k values from 1 to 100. The accuracy for each k-value was then mapped in a graph as shown in Figure 15, but here we only display each even value of k to avoid overlapping labels for the x-axis. This shows us that the best k-value to use is 16 as it is the point of diminishing returns(knee-point). It is set to this as any k value less than 16 gives a poor accuracy. Any k value greater than 16 on the other hand, does not significantly increase the accuracy of the classifier, but it requires many more computations than a k value of 16.

2. **criterion:** This is the function that measures the quality of a split. We leave this as its default value "gini" which uses the Gini impurity to measure the quality of a split.

3. **max_features:** We leave this as the default value of "auto" which chooses the max features as the square root of the number of features.

4. **max_depth:** We leave this as the default value "None" as we do not want to limit any tree to a certain depth.

5. **min_samples_leaf:** We leave this as the default value of 1. This ensures that the tree does not create a leaf node when it can split the node.

6. **max_leaf_nodes:** We leave this as the default value of "None." This prevents the tree from not splitting nodes when the number of leaf nodes exceeds a certain number.

For the remaining parameters, we let the package decide the best values for the model.

We start our algorithm by splitting the data and classes into two parts. We use 70% of our dataset as the training data and 30% of our data as test data. Then as mentioned above, we try the Random Forest classifier algorithm for k

values ranging from 1 to 100. For each k value, we then initialize the RandomForestClassifier from Scikit-Learn. We fit the model using our training data and classes. We then use the model to predict the classes for our test data. If the predicted value of the class is the same as the original class value, we add to our true positive count. The accuracy of a particular k is then calculated as the number of true positives divided by the number of records. Finally, we plot a graph as shown in Figure 15.

Figure 15 contains a graph of the accuracy vs. the k value. We display the average of the accuracy of an odd k value and an even k value as the accuracy of the even k value to avoid overlapping labels on the x-axis. The blue line displays a knee curve, and the black dashed line shows us the point of diminishing returns.

To summarize the best model we could fit in the Random Forest Generator used a k value of 16 and gave us an accuracy of 0.81 or 81.0%.

7. ANALYSIS

We started with the original unclean data for a second analysis so that we could find some new patterns which we did not find during the previous analysis. We are aware of the confirmation bias, and we did not want to fall victims to it. This time we did not remove any of the street names. We only kept vehicle one factor and vehicle one type for our analysis. We cleaned the data again without removing the missing fields. We introduced new features such as the month and year of the accident, the day of the week, time of the day and season. For the season field, December-February as Winter, March-May as Spring, June-August as Summer and September-November as Fall. We also removed the data of the year 2017 from our data. Thus, we only kept data from Jan 2015 to Dec 2016 for the current analysis as keeping the original data skewed the results in favor of more number of accidents in the winter season because we had two extra months of data for this particular season. We can observe from the figure that highest number of collisions occur during the summer season followed by fall, spring and winter.

After plotting for each season, we plotted the total number of collisions in each borough during the 2-year span and can see that Brooklyn has the highest number of accidents followed by Queens, Manhattan, The Bronx and Staten Island. We know from the previous analysis that Manhattan has the highest number of per person accidents. This is due to the fact that a lot of people from rest of the boroughs travel to Manhattan for work. Another reason might be the traffic in Manhattan contributing to the total number of accidents in that borough.

We can observe that Friday has the highest number of collisions followed by Wednesday and Thursday. Something peculiar is that the least number of accidents occur over

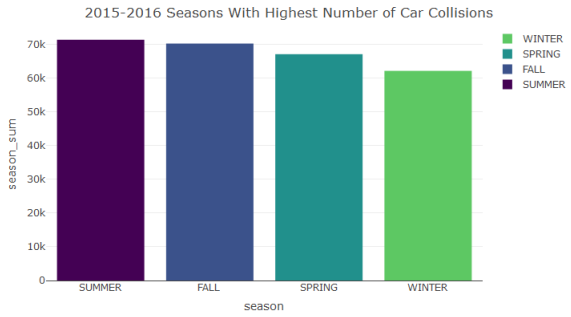

Figure 16: Car collisions each season for 2015-2016

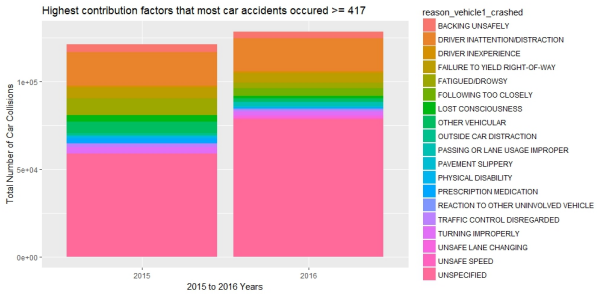

Figure 17: Reasons for car accidents above the threshold

the weekend followed by Monday. One of the reasons behind this may be that people do not go out much during the weekend and tend to stay in and rest. Another reason might be that people are not in the rush to get to work or somewhere else hence they tend to pay more attention while driving. Although we have not explored it this time, it would be interesting to find the leading causes of accidents during the weekend and what type of vehicles are involved in those collisions. We have found that the average number of vehicular collisions happening each day are 417. Hence, we tried to find out the reasons that would lead to high number of collisions on certain days and low number of collisions on other days and plotted two graphs to differentiate between the reasons.

We can observe from that graph that on high collision reasons, major factors are distractions, driver inattention, failure to lead right-of-way, drowsy, driving too closely. Hence, we can conclude that just by raising awareness we have the ability to avoid these accidents. These accidents can be avoided if the driver is paying more attention while driving. While reasons have less number of car accidents, the major factors are alcohol involvement, animal action, brakes defective, road rage, drugs. The other factors are random factors which cannot be controlled easily. We can notice two different patterns on this different types of days. If the driver is not paying attention or is being lazy, then that will lead to higher probability of accidents. For future work, we can find

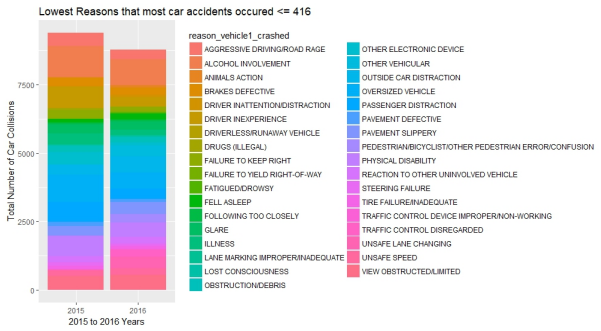

Figure 18: Reasons for car accidents below threshold

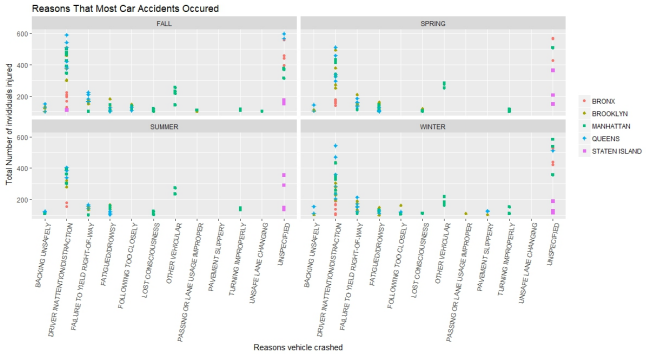

Figure 19: Car collisions for each borough

the average accident per day and then divide the data into days having a higher number of accidents than usual and days having a lower number of accidents than usual. Then perform an analysis to find the major factors on these two different kinds of days. This might offer some insight into why certain days tend to have a higher number of accidents than others.

We plotted the highest contributing reasons for each season in the five boroughs and can see patterns. Some reasons like improper lane usage, following too closely and slippery road are not observed during the summer season. Also, driver inattention during summer is the lowest as compared to the rest of the seasons. Failure to yield right-of-way is also the lowest during the summer season. Driver inattention is at its peak during the fall season. We can also see that driver inattention is highest in Queens as compared to the rest of the boroughs. Queens also leads all the districts in backing reason slowly. Reasons for this might be lack of proper policing in that borough or paucity of punishments. We can observe that most of the driver fatigue leading accidents are seen in Brooklyn. Turning improperly can only be found in Manhattan. No other borough finds this reason leading to crash. Majority of the data collected for Staten Island has unspecified as the reason. This is a data gap as we do not know what the leading causes of accidents in Staten Island are. During the fall season, the number of crashes in Queens jumps by a considerable amount as compared to the rest of

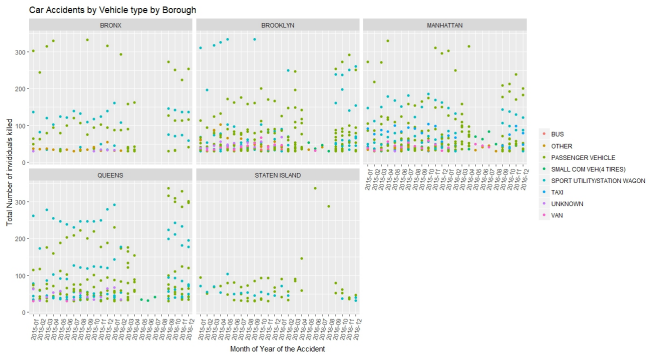

Figure 20: Vehicle type in each borough

the seasons. We can conclude that it is relatively safer to not drive during fall in Queens.

After analyzing the leading causes of accidents in each borough, we decided to look for patterns in the vehicle types involved in the accidents and separated them by month in each of the five districts. The first thing that we can observe from this graph is that there is enormous data gap of four months from may to August of 2016. This data gap can be blamed on the data cleaning. During the data cleaning phase for the initial analysis and time graph construction, a majority of the cases recorded during these months had some missing values which lead to a large number of records being removed from the dataset. Although, we do not know the reason behind the sudden increase in the missing values among the recorded accidents. It suddenly takes a dip in the number of crashes for some of the boroughs, and no accidents are observed in The Bronx. For the remaining months of the year, we can see that passenger vehicle is the leading vehicle involved in accidents in The Bronx and Manhattan. Similarly, Sports utility/station wagons are the prominent types of vehicles involved in collisions in Brooklyn and Queens. Although, in the latter half of the year 2016, passenger vehicles were the leading cause instead of sport utility vehicles in Queens. A similar flip can also be observed in Brooklyn as well. Taxis contribute to a lot of accidents in Manhattan. This is more than the rest of the boroughs.

We now analyze pedestrians in which borough are more likely to be involved in an accident. From Figure 20, we can see that Brooklyn has the most number of pedestrians injured. This means that streets of Brooklyn are more dangerous to walk on than the rest of the boroughs. Queens has the second highest number of people among the five boroughs and is then understandable to have a higher number of pedestrians being injured as compared to the rest of the districts. This graph shows that there is nothing out of the ordinary in the data found. Even without the plot, most people would've been able to guess which borough would have the highest number of pedestrians getting injured based on the population of the district. Although, it is always better

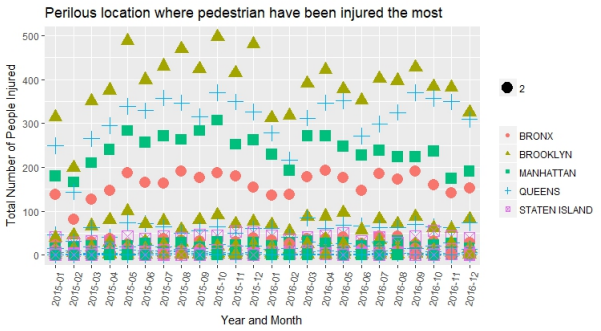

Figure 21: Number of people injured in each borough

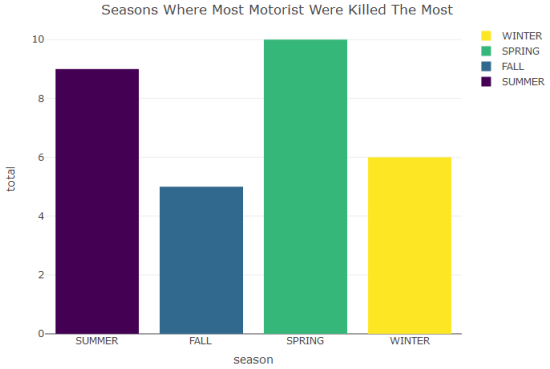

Figure 22: Motorist Killed by season

to be sure and confirm that the data backs up such claims.

We analyzed the reasons leading to a motorists death. From the figure, we can say that Spring is the most dangerous time to drive a motorcycle. Also, it is twice as likely to get killed by riding a bike in Spring than in Fall. But this is the bane of using relative statistics. Although the statement is correct, it does not draw the entire picture regarding the safety of driving motorcycles. We include this graph just to highlight the point where certain conclusions can be drawn from data without stating about that actual probabilities. Manhattan has observed just a single motorist death during 2015-2016. It is safer than Staten Island for motorists. The number of deaths has reduced considerably in 2016 as compared to 2015. Four motorcycle deaths were observed during September 2015 which was the peak for motorcyclists death. After that, we can observe a graduate decline in the deaths across all five boroughs.

After plotting for each season, we made a plot for the number of collisions in each borough during each month which we can observe in the figure. We can see that May and July have the highest number of accidents in each month. We will complete an in-depth analysis of both the peak months to find any patterns which might be useful. First, we compared the total number of accidents in both the months of 2015 and

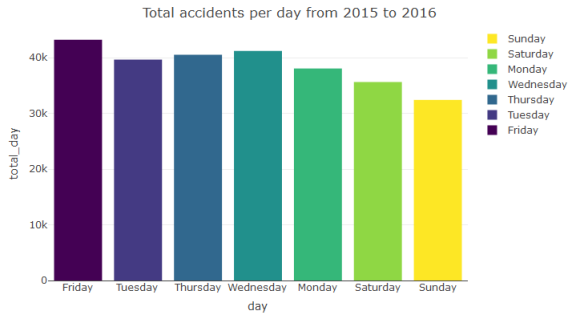

Figure 23: Total accidents on each day of the week

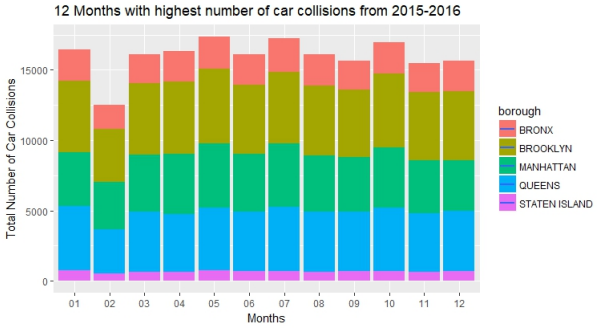

Figure 24: Car collisions each month

2016 and compared them. We can observe a general trend of reduction in the number of accidents every year. One of the reasons behind this trend is the Zero initiative started by the mayor of NYC in 2014. This effort was modeled after the government of Sweden began a similar action. The goal of the initiative is to reduce the number of traffic fatalities to half its quantity by 2025.[1]

From the comparison between the graphs of two months, we can observe that Brooklyn has the higher number of accidents than the rest of the boroughs. Also, both the months peaked around the same date, i.e., towards the end of the month in 2015. It would be worth looking into why did both the month peak towards the end of the month. Also, during May, the number of accidents in Brooklyn dipped more

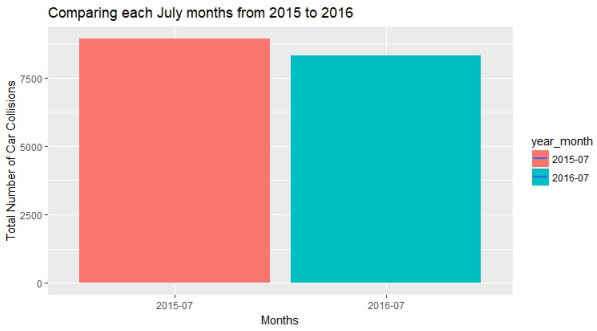

Figure 25: Collisions in July

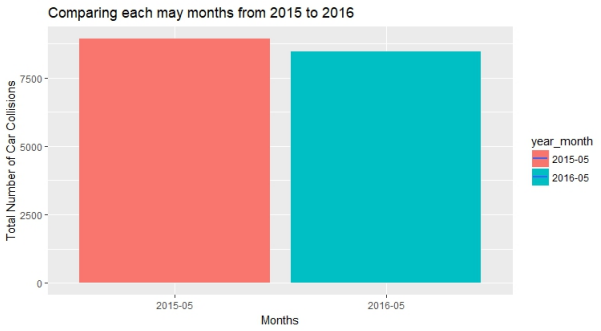

Figure 26: Collisions in May

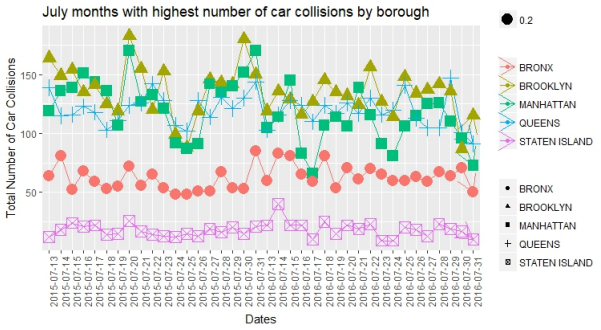

Figure 27: Peak collision dates in July month

than usual, and Manhattan peaked on 15th of may. It is very unusual to have a more significant amount of crashes in Manhattan as compared to Brooklyn. Although it does happen around the same data again next year, the difference between them is huge in 2015 and is minimal in 2016. That seems to be an anomaly. We can easily see the weekend dip trend in both the graphs. In the July peak collision graph, we can see a dip in the number of accidents for two days which are weekend and then a peak for the next five days followed again by a dip.

While exploring the data, this is something unusual we noticed. Even though the number of accidents for the month of may decreased going from 2015 to 2016, the number of collisions increased in Queens, Staten Island, and The Bronx. All these increases were offset by the decrease in the number of accidents in Manhattan.

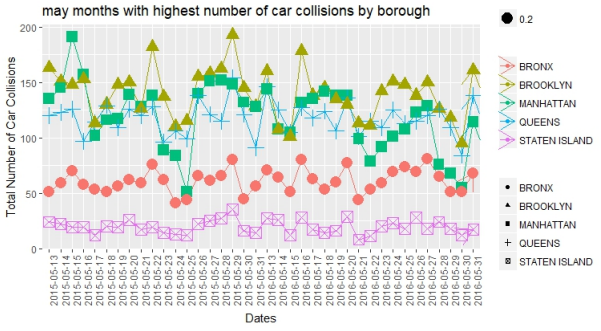

Figure 28: Peak collision dates in May month

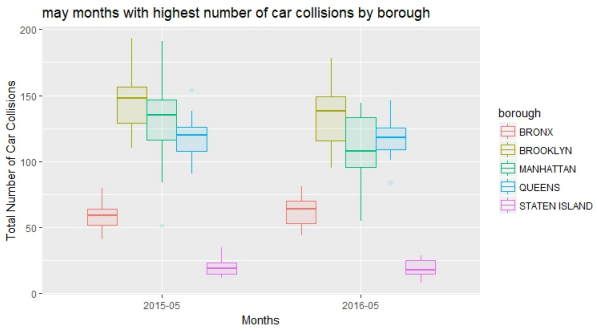

Figure 29: Collisions in May by borough

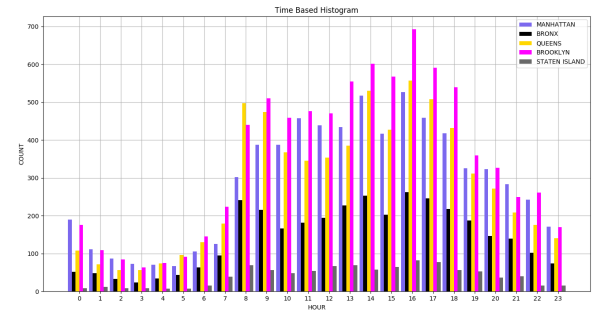

Figure 30: A histogram of the number of collisions in each borough at each hour

From these two time histograms, we can observe that from 4 am, the number of accidents decreases until hitting a local minimum around 9 am dipping a bit at 10 am. After 10 am, it again increases till it peaks at around 4 pm after which it decreases for the next 12 hours till 4 am. It is interesting to note that the probability of accidents peaks at 4 pm. One of the reasons may be that people are trying to rush and get to the place they are headed towards and reach their destination before the 5 pm rush hour hits. Other people might try to leave early for the same reason and might be in a hurry. Another reason might be that after finishing the work day, people are tired and might not be paying much attention while driving or walking. All of these factors contribute towards the peak at 4 pm. It is better to skip the rush hour and wait a couple of hours before driving. Just waiting for two more hours would decrease the probability of accidents to half the likelihood at 4 pm.

We plot all the accidents in during the two years 2015-2016 against the NYC map. This plot just shows that over the course of time, the accidents cover most of the city and we cannot see any visible clusters.

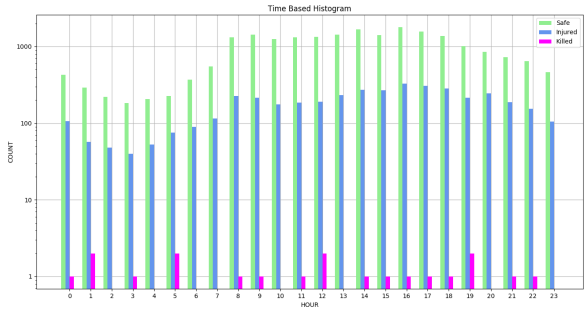

Figure 31: A time histogram for the number of collisions for each class

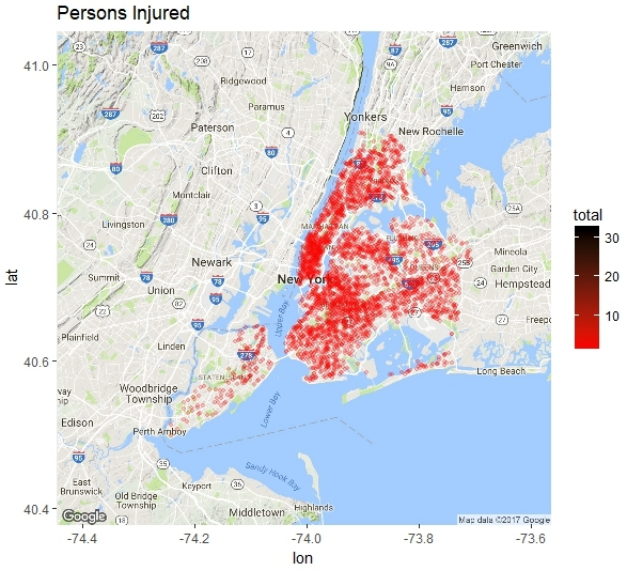

Figure 32: Graph of latitude vs. longitude vs. class for a collision

8. CONCLUSION

We found a great business need for our project in the emergency response industry. Unfortunately, we could not identify the causes of collision fatalities. Instead, we came to many exciting conclusions given the dataset we had. We saw that using the 0R rule, and most vehicular collisions lead to no loss of life or injury. In fact, people in most accidents have a low probability of being injured and an extremely low probability of dying. Of course, any fatality in a vehicular crash should be avoided, and we believe that a more in-depth analysis of the cases of a vehicular accident leading to a fatality should be undertaken. This is unfortunately difficult to ascertain from the dataset we use because the instances of death are too low to provide any significant value to our data. Our research instead can be used by legislators an emergency response personnel to decrease vehicular collision response time and vehicular collision injury/fatality count. Due to the VisionZero initiative, the number of collisions is reducing year by year. May and July are the least-safest months for motor collisions in New York City. The number of traffic collisions peak at 4 pm. It is relatively safe to drive around the city from 6 pm to 6 am. Brooklyn is the most likely borough to have an accident.

8.1 Learning

During the span of this project, we gained a plethora of knowledge and came to many conclusions related to our dataset and machine learning in general. We learned how much time data cleaning takes. Data cleaning requires the cleaner to make a lot of decisions regarding what to change, what to keep and what to delete from the database. After that comes the feature selection. Which features to choose is very important to the data and the model we are trying to fit it. If we have oblique data, we need to carry out Principle Component Analysis to create usable features. Visualizing the data always helps in deciding the features we need to use. Choosing on which colors to use and the contrast between them as well as determining on a background color or image is an essential part that many data people overlook. This is the difference between a graph being able to convey a lot of information with a glance, rather than requiring the user to pay more attention to find the patterns in the figure. There are always a large variety of models that can be discovered from a chart, if enough effort is put into finding them, be it by creating features or understanding more complex figures. When we first started classification, we had to invent our classes to develop a business case for the data. Many times it is necessary to think about the business cases first and then choose the data, as well as the features, as that might ease the effort of completing the task. Data cleaning takes up a huge amount of time and sometimes is 50% or more of the work. Given enough information related to the task at hand, it is always possible to derive meaningful inferences

about the data. If there is more data available regarding the kind of accidents, demographics of the driver, total damages due to the collision, we can find even more use cases for the data. Even though we are working on old data and a lot of work has been carried out on data, there is always the possibility of finding new meaningful information.

9. REFERENCES

[1] Jrip-pentaho data mining. https://en.wikipedia.org/wiki/Vision_Zero_(New_York_City).

[2] Research note: 2015 motor vehicle crashes. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812318.

[3] Vision zero(new york city). http://wiki.pentaho.com/display/DATAMINING/JRip.

[4] Adhokshaja, P. Exploratory data analysis. https://www.kaggle.com/adhok93/exploratory-data-analysis/notebook.

[5] Black, A. W., and Mote, T. L. Effects of winter precipitation on automobile collisions, injuries, and fatalities in the united states. *Journal of transport geography 48* (2015), 165–175.

[6] Chirico, M. Traffic fatalities – getting started. https://www.kaggle.com/mchirico/traffic-fatalities-getting-started/notebook.

[7] Depaire, B., Wets, G., and Vanhoof, K. Traffic accident segmentation by means of latent class clustering. *Accident Analysis & Prevention 40*, 4 (2008), 1257–1266.

[8] Larion, A. Traffic fatalities in 2015. https://www.kaggle.com/abigaillarion/traffic-fatalities-in-2015/notebook.

[9] O'Donnell, M. Exploring nyc vehicle crash data in tableau. https://www.interworks.com/blog/modonnell/2015/08/26/exploring-nyc-vehicle-crash-data-tableau.

[10] Ulfarsson, G. F., and Mannering, F. L. Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. *Accident Analysis & Prevention 36*, 2 (2004), 135–147.

[11] Valent, F., Schiava, F., Savonitto, C., Gallo, T., Brusaferro, S., and Barbone, F. Risk factors for fatal road traffic accidents in udine, italy. *Accident Analysis & Prevention 34*, 1 (2002), 71–84.