Accident Prevention and Preparation (Group 6: AP&P)

Uzmabanu Kapadia EMPLID: 23864352 ukapadi000@citymail.cuny.edu
Tanmim Ahmmed EMPLID: 24060614 tahmmed000@citymail.cuny.edu
Sohail Ahmad EMPLID:23921990 sahmad005@citymail.cuny.edu

Abstract:

This project aims to aid in the prevention and preparation of accidents. The goal of this project is to have improved emergency response times and strategies to reduce accidents. We plan to achieve a trained model using the motor collision dataset which is crucial for understanding and analyzing traffic accident patterns. Through this model, we can work to predict the most accidents likely to happen based on a given region, time, or season. The anticipated model here would aid in public safety measures and city planning to help reach our end goal of preventing and preparing against accidents.

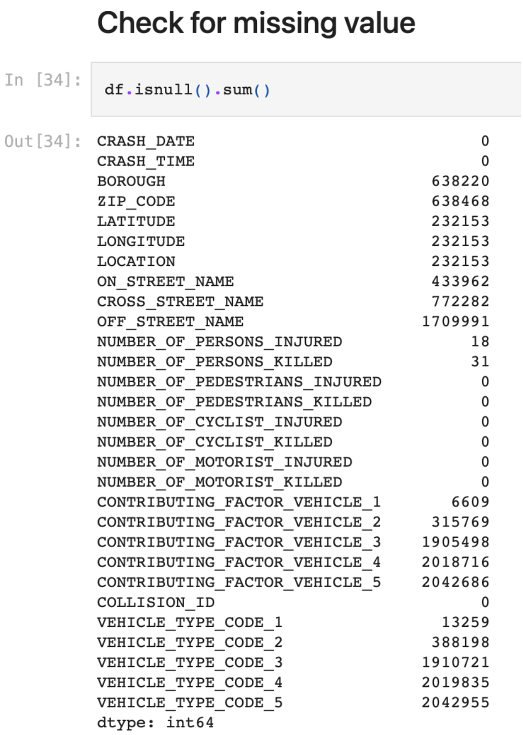
Introduction:

When driving one can be affected by various situations. These driving conditions can include traffic, location, or the time/season which can impact the accident statistics and often increase it. However, the city is usually not prepared to quickly respond to the increased accidents that can happen due to fluctuations in the number of drivers on the road. In some areas which are prone to accidents, it might make situations even worse causing road blockages and creating additional issues for drivers. Currently, it takes about nine minutes for aid to arrive however, such fluctuations would increase the response time which affects death rates in accidents. Here, we want to examine ways to help prevent against or prepare for such conditions overall to reduce accidents and have quicker response time through preparation.

For this project, we decided to use the Motor Vehicle Collisions dataset from NYC Open Data. This dataset includes date/time, location, causes, vehicle types, injuries, and fatalities in the crash. The dataset provides granular details suitable for in–depth analysis and real-time updates. This dataset was used to create a logistic regression model to help with accident predictions. The model was first trained where the features and targets were selected and then the data was split into train and test sets and defined. Then, the model selection and prediction was done on the test set after which the model was evaluated.

Pre-processing:

The dataset being used here, the motor vehicle collision dataset, was a fairly clean dataset with minimal cleaning required. However, some cleaning done for this dataset was checking for any missing values. This was done to ensure that our dataset did not consist of a great deal of null data that would need to be removed. In the image below it can be seen that for the columns from the dataset being used for this project, the majority of them contained little to no null values.

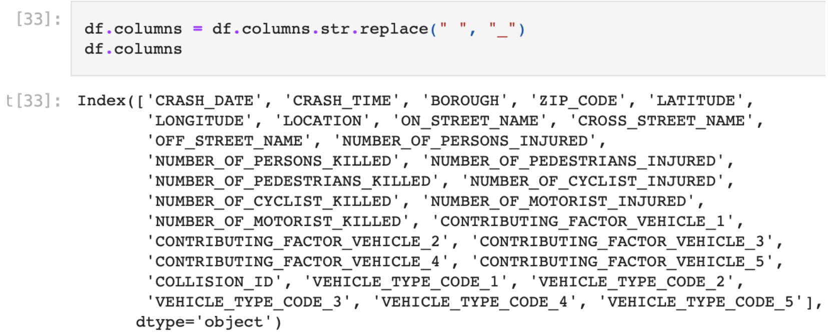


Check for missing value

```
In [34]:  df.isnull().sum()

Out[34]:  CRASH_DATE                             0
          CRASH_TIME                             0
          BOROUGH                           638220
          ZIP_CODE                          638468
          LATITUDE                          232153
          LONGITUDE                         232153
          LOCATION                          232153
          ON_STREET_NAME                    433962
          CROSS_STREET_NAME                 772282
          OFF_STREET_NAME                  1709991
          NUMBER_OF_PERSONS_INJURED             18
          NUMBER_OF_PERSONS_KILLED              31
          NUMBER_OF_PEDESTRIANS_INJURED          0
          NUMBER_OF_PEDESTRIANS_KILLED           0
          NUMBER_OF_CYCLIST_INJURED              0
          NUMBER_OF_CYCLIST_KILLED               0
          NUMBER_OF_MOTORIST_INJURED             0
          NUMBER_OF_MOTORIST_KILLED              0
          CONTRIBUTING_FACTOR_VEHICLE_1       6609
          CONTRIBUTING_FACTOR_VEHICLE_2     315769
          CONTRIBUTING_FACTOR_VEHICLE_3    1905498
          CONTRIBUTING_FACTOR_VEHICLE_4    2018716
          CONTRIBUTING_FACTOR_VEHICLE_5    2042686
          COLLISION_ID                           0
          VEHICLE_TYPE_CODE_1                13259
          VEHICLE_TYPE_CODE_2               388198
          VEHICLE_TYPE_CODE_3              1910721
          VEHICLE_TYPE_CODE_4              2019835
          VEHICLE_TYPE_CODE_5              2042955
          dtype: int64
```

Additionally, as a part of preprocessing all the spaces in the column names were replaced with an underscore as a way to make it easier to use the data columns. Below it can be seen how the replace statement was used to change “ ” to “_” and the new names with the underscore displayed below the code.



```
 [33]:  df.columns = df.columns.str.replace(" ", "_")
        df.columns

t[33]: Index(['CRASH_DATE', 'CRASH_TIME', 'BOROUGH', 'ZIP_CODE', 'LATITUDE',
              'LONGITUDE', 'LOCATION', 'ON_STREET_NAME', 'CROSS_STREET_NAME',
              'OFF_STREET_NAME', 'NUMBER_OF_PERSONS_INJURED',
              'NUMBER_OF_PERSONS_KILLED', 'NUMBER_OF_PEDESTRIANS_INJURED',
              'NUMBER_OF_PEDESTRIANS_KILLED', 'NUMBER_OF_CYCLIST_INJURED',
              'NUMBER_OF_CYCLIST_KILLED', 'NUMBER_OF_MOTORIST_INJURED',
              'NUMBER_OF_MOTORIST_KILLED', 'CONTRIBUTING_FACTOR_VEHICLE_1',
              'CONTRIBUTING_FACTOR_VEHICLE_2', 'CONTRIBUTING_FACTOR_VEHICLE_3',
              'CONTRIBUTING_FACTOR_VEHICLE_4', 'CONTRIBUTING_FACTOR_VEHICLE_5',
              'COLLISION_ID', 'VEHICLE_TYPE_CODE_1', 'VEHICLE_TYPE_CODE_2',
              'VEHICLE_TYPE_CODE_3', 'VEHICLE_TYPE_CODE_4', 'VEHICLE_TYPE_CODE_5'],
             dtype='object')
```

Modeling:

Our logistic Regression model achieved an overall accuracy of 0.78 on the evaluation dataset. This accuracy represents the proportion of correctly predicted instances out of the total instances. While accuracy is an important metric, it's essential to delve deeper into other evaluation metrics to gain a better understanding of our model's performance.

Precision: Precision measures the accuracy of positive predictions. For class 0, the precision is 0.78, indicating that 78% of predicted severe accidents were correct. However, for class 1, the precision is 0.00, suggesting that the model struggles to correctly identify severe accidents.

Recall: Recall quantifies the ability of the model to capture all instances of a particular class. For class 0, the recall is 1.00, meaning the model effectively identifies non-severe accidents. However, for class 1, the recall is 0.00, indicating a complete failure to identify severe accidents.

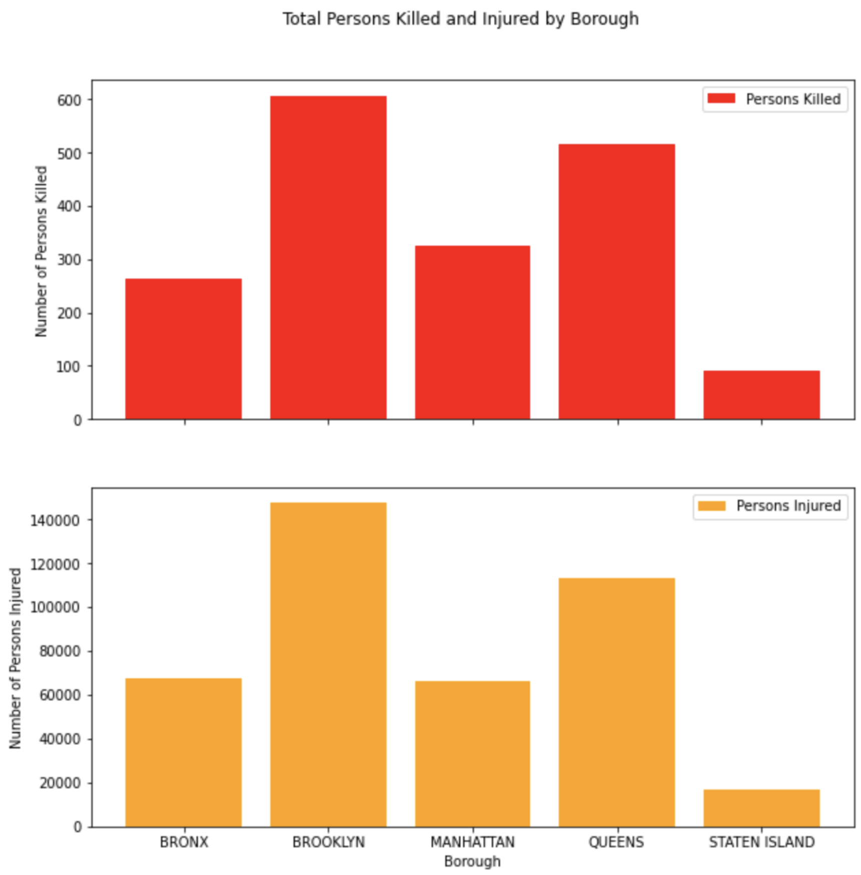
F1-Score: The F1-score is the harmonic mean of precision and recall. For class 0, the F1-score is 0.88, reflecting a balance between precision and recall. For class 1, the F1-score is 0.00, suggesting poor performance.

Support: Support represents the number of actual occurrences of each class in the specified dataset. In this case, there are 271,009 instances of class 0 and 77,270 instances of class 1.

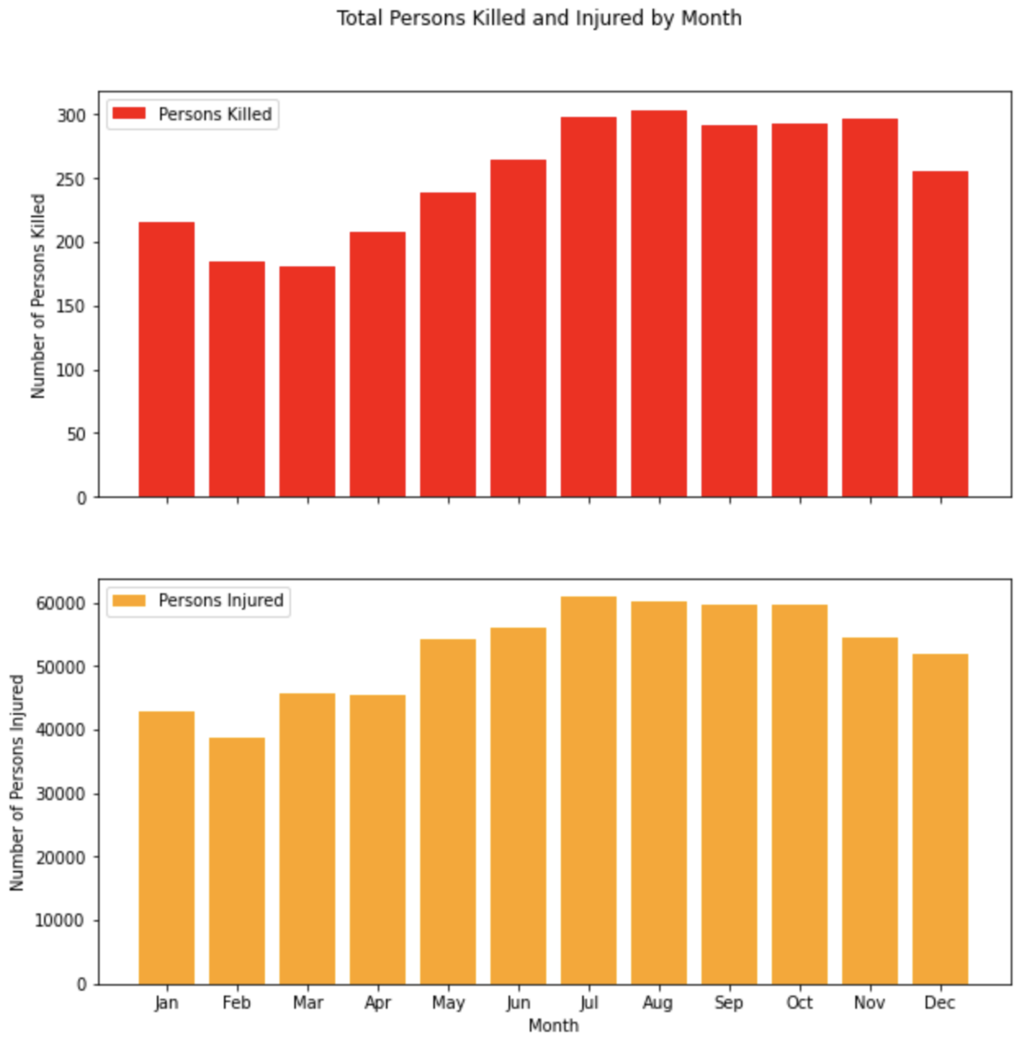
The classification report reveals imbalanced performance between the two classes, with the model excelling at predicting non-severe accidents (class 0) but performing poorly on severe accidents (class 1). The low precision, recall, and F1-score for class 1 indicate a significant challenge in identifying severe accidents. The low scores for predicting severe accidents (class 1) in the classification report indicate that the model faces challenges in correctly identifying and predicting these instances. Several factors may contribute to low scores such as class imbalance, data quality, and model evaluation metrics.

Analysis:

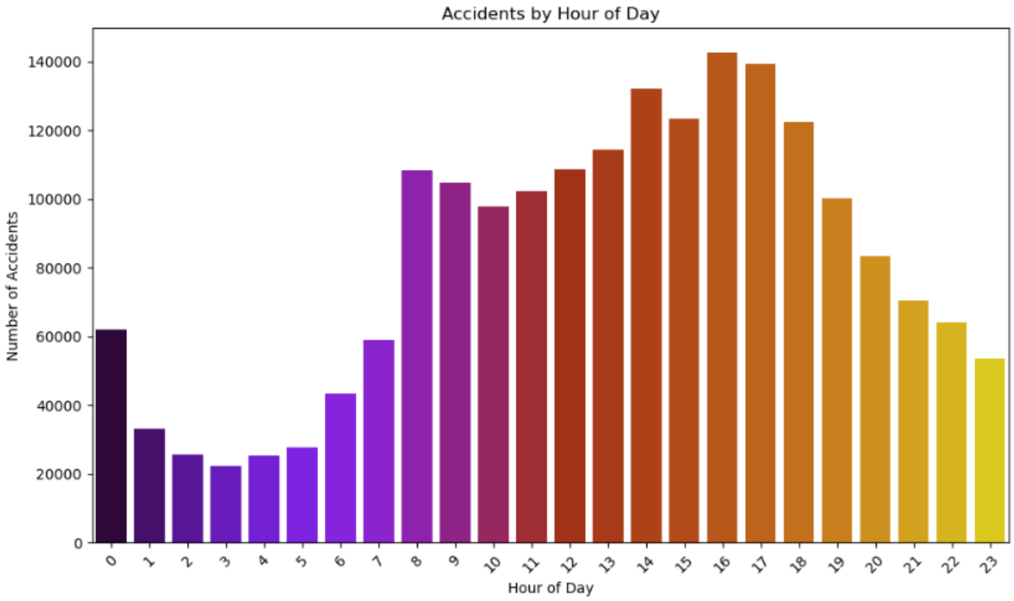
For this project, various different techniques were used to help with accident prevention. A Lot of the data was visualized to see common trends in the accidents. Here, we will be going through the visualized data to analyze these trends.

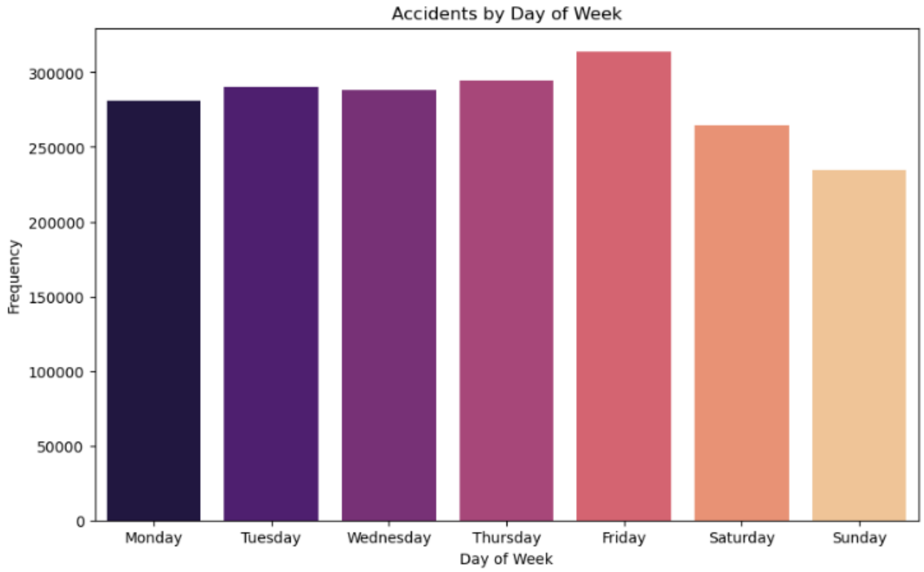


In the image above we can see the graphs of how many people were killed and injured grouped by the different New York City Boroughs. From this, the borough with the most accidents leading to death and injuries is Brooklyn (approx. 600 deaths and 140,000 injuries) and the second most being Queens (approx. 500 deaths and 120,000 injuries). This is likely due to the traffic in these boroughs and them having certain accident hotspots. Looking at this it is safe to say the boroughs requiring the greatest amount of emergency response and preparedness on a regular time are Brooklyn and Queens.



These tables show the total number of people killed and injured grouped using the different months. From these graphs we can see that during the summer months of July and August the collision rates are the highest with deaths ranging from 275-300 and injured persons approximating around 60,000. This portrays that the greatest amount of care required from first responders and the drivers on the road is during the summer holidays times as they have the most collisions.

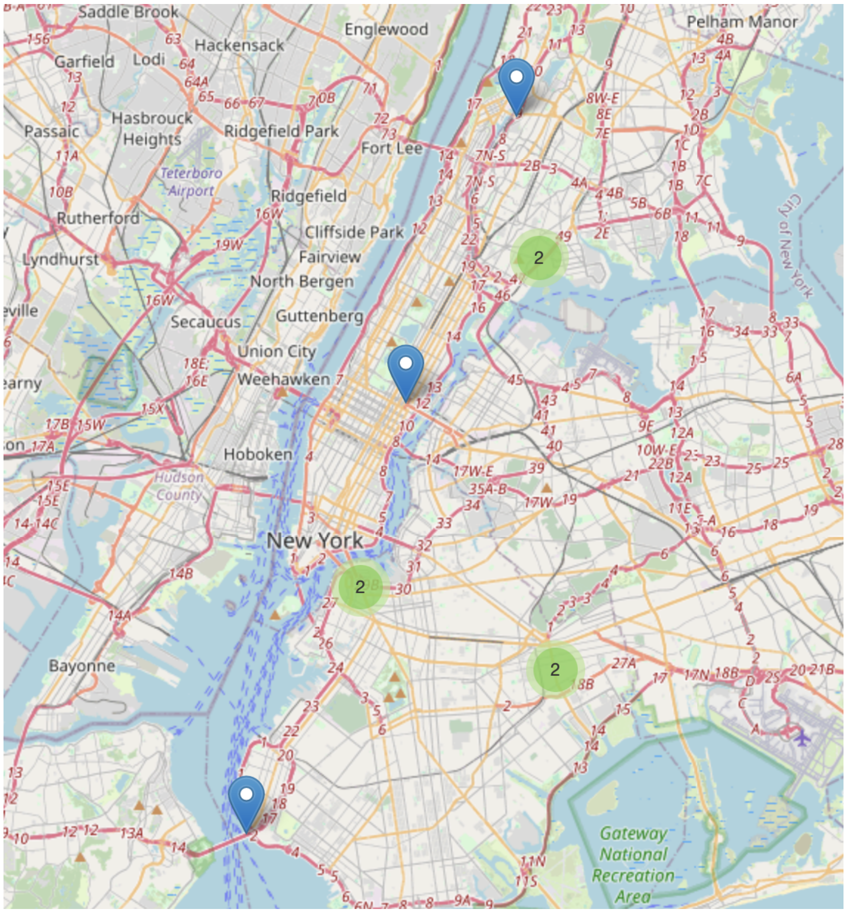




The generated visualizations detail the incidence of traffic accidents segmented by hour of the day and by day of the week. From the first graph, “Accidents by Hour of Day” a discernible pattern emerges; the frequency of accidents escalates progressively in the morning hours, reaching a peak during the afternoon around 15:00 to 18:00, before tapering off towards the late evening. This suggests a correlation with rush hour traffic, where increased vehicular congestion likely contributes to a high number of traffic accidents.

The second graph, “Accidents by Day of Week” presents a comparative view of accident occurrences across the different days of the week. A consistent level of traffic related incidents is observed from Monday through Thursday, with a noticeable uptrend on Friday and the highest concentration on Saturday, and the figures decline on Sunday. The elevated accident rate on Saturdays may imply convergence of different factors, including increased leisure travel and varied driving patterns during the weekend.

The insights underscore the heightened need for vigilance and alertness during late afternoons and Saturdays, when accident rates surge. This data can be crucial for first responders and policy makers to strategize on resource allocation and for drivers to exercise increased caution during these identified peak times.



In terms of location, we can see that some of the accident hotspots include Verrazano Narrows Bridge with an accident count of 670. Queensboro bridge with an accident count of 474, Bruckner Expressway near cypress avenue with an accident count of 597 and Bruckner Blvd near Hunts Point Avenue has an accident count of 467. Some other hotspots include Major Deegan Expressway which has an accident count of 685, Tillary street in Brooklyn has an accident count of 646, Linden Blvd in Brooklyn has an accident count of 466, and lastly Atlantic Avenue in Brooklyn has an accident count of 513. Through this we are able to understand some areas and intersections which may be prone to accidents and can take preventative measures to reduce accidents as well as increase emergency response in these areas.

Summary:

Overall, using our logistic regression model and the data visualization we were able to analyze and understand the accident patterns through New York City to help prevent and prepare for accidents. Through our Logistic regression model which has a 79% accuracy we are able to predict instances of accidents to help with prevention and preparation. Additionally through the various data visualizations we were able to help create an overall analysis on what areas were the most likely to have accidents, the months in which most accidents occurred, what days had the most accidents, as well as during what times of the day did most of these accidents occur. Through this the accidents patterns such as most accidents happen in Brooklyn and Queens, with most accidents occuring in the summer months of July and August, as well as the most popular day being Fridays during 4-5 pm time slot. This creates an in-depth understanding on when most care is needed from first responders and when drivers should be most cautious. In conclusion, gaining knowledge on this through the visualizations and the trained model gives insights to help with being prepared and preventing accidents.