SI 618 Final Report

Unraveling New York City's Traffic Accidents in 2019: Insights and Impactful Factors for Safety Enhancement

Group jiaou-yxmin: Jiaou Chen, Yuexin Min

Motivation

As urbanization accelerates, the number of vehicles continues to increase, leading to frequent road traffic accidents that pose a threat to public safety and result in property losses. To reduce the occurrence of road traffic accidents and minimize the impact on people's lives and property, it is imperative to gain a deep understanding of the characteristics and influencing factors of these accidents. Only by doing so can we better identify and comprehend the key factors causing accidents, optimize traffic management measures, enhance road safety, and reduce both the frequency and severity of traffic accidents.

To help address the problem of increasing traffic accidents, we intend to investigate the possible factors related to traffic accidents, which may include the type of road, speed limit, weather, time, and type of vehicles. We would also like to investigate the relationship between the possible factors related to the direct reason for the accident that is included in our data set.

Based on our motivations, we had generated several questions to research:

- When and where do accidents happen most frequently in New York?
- How can the severity of traffic crashes be categorized based on injuries and fatalities in the dataset, and what are the most common causes of crashes?
- What could be the possible factors of the accidents impacting the severity of an accident?

After doing some background research, we decided to use the Gradient Boosting method for modeling in our data analysis part regarding the third problem above, which is a common method in the field of traffic accident research (Cuenca et al., 2018; Ahmed et al., 2021).

Dataset Description

Primary Data:

Our primary dataset is the information on motor vehicle collisions in New York from 2012 to 2023, which is collected by the New York City Police Department (NYPD) from MV-104AN forms, created by the police to record the accident data.

The analysis will primarily focus on the dataset of the year 2019 due to the discernible impact of the COVID-19 pandemic on traffic volume. The decrease in tourists and a tendency for individuals to remain at home during this period notably contributed to a reduction in the number of reported accidents. Subsequent years may not present as representative a picture due to the evident decrease observed post-2019. The key features we'll look at include the date and time that the crash happened, the street that the crash happened on, the borough where the accidents happened, the number of deaths & injuries of the participants, the vehicle type of each participant, and the contributing factor of the accident from each participant.

The size of the dataset is 2026647 records/411.1MB (in total), and there were 211486 records in 2019.

The URL of this dataset is: https://www.kaggle.com/datasets/tush32/motor-vehicle-collisions-crashes

Secondary Data:

The secondary dataset is collected so that we can investigate possible factors influencing the accident severity.

1. New York City Speed Limit Data

   Our first secondary dataset records information about all streets in New York. It comes from the New York City Department of Transportation (NYCDOT). It includes some key features, for example, the street name and speed limit of the street. We expect the speed limit to be in integer type and the street name and type to be in string. The estimated size of this dataset is 22000 kb. The URL of the dataset is:
   https://www.kaggle.com/datasets/splacorn/speed-limits-in-nyc-taxi-playground-challenge
   The format of this dataset is csv, which requires pandas to read csv to transfer this data into the data frame.
2. New York City Weather Data

   Our second secondary dataset is collected from the National Weather Service. It contains daily data on the minimum temperature, maximum temperature, average temperature, precipitation, new snowfall, and current snow depth in 2019. We anticipate figuring out the relationship between precipitation, current snow depth, and the frequency of happening times of certain types of accidents. The data type of these two fields may be integer and float. The estimated size of this dataset is 4 kb. The URL of the dataset is https://www.kaggle.com/datasets/alejopaullier/new-york-city-weather-data-2019
   The format of this dataset is csv. The format of this dataset is csv, which requires pandas to read csv to transfer this data into the data frame.

Data Manipulation

Data Cleaning:

Primary Data:

First, we read the primary data set (the crash data) from the csv file. Then we filtered the primary data by year. We only looked at data in 2019, as after 2019, the traffic volume was impacted by COVID-19.

We drop missing values for several columns. These include the crash date, crash time, on-street name and contributing factor 1, and vehicle type 1. We fill in missing values with 0 for the number of deaths and injuries. Also, we identify and drop records where a contributing factor exists but the corresponding vehicle type is missing, and vice versa.

Weather Data:

We select only the 'date', 'precipitation', 'new_snow', and 'snow_depth' columns from the entire weather dataset. Then we fill the missing values with 0, namely the nan value. We also found that the dataset uses "T" for missing values, so we also replaced 0 with T.

Speed Limit Data:

We first drop the missing values in all columns. Then we find that for each street, there may exist several speed limit road segments, so we compute the mean speed for a given street.

Data Conversion:

Date conversion for Merging:

To merge the three data frames, we made several conversions and added several new features to our data. First, we convert the date column of both the primary data frame and weather data frame from string to timestamp. Then we applied the left join to the primary data frame and weather data frame. Then we merge the primary data with the street data based on the "ON STREET NAME" and "street".

Data conversion for visualization:

For further visualization, we extract the Month and Day from the 'CRASH DATE' in the crash_total. We then get the standard form of the time in a day by applying to_datetime for 'CRASH TIME'. We also map the date to the day of the week.

Due to the complexity of modeling actual time in the data analysis phase, we introduced a new column named "time interval." This column maps hours into distinct time segments: morning spans from 5 am to 12 pm, afternoon encompasses the period from 12 pm to 5 pm, evening spans from 5 pm to 9 pm, and night extends from 9 pm to 5 am the following day.

Challenges:

One big challenge we encountered was merging the speed limit data frame and the primary data frame. We find that there are only several streets that can be matched successfully. So we did a data conversion to the street columns of the primary data frame and the street data frame.

We extract the space from the street name column of both datasets and make all letters in lowercase. Successfully, 130 thousand rows of data have been successfully retained.

Data Visualization and Analysis

Visualization and Analysis of Place Factors

1. **Initially, we utilize Folium to visualize the crash density across the New York City map using a heat map.**

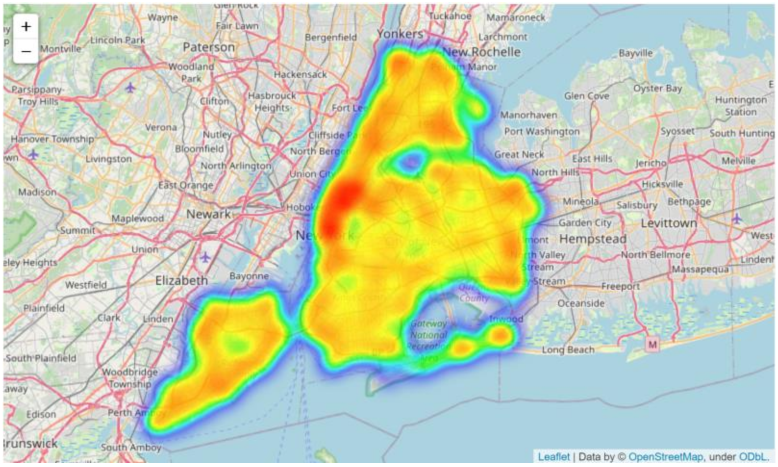

Fig. 1. Heat map of the crash density in NYC

As shown in Fig. 1, crash density rises in tandem with color temperature, it becomes evident that the areas marked in red, northwest New York primarily near Manhattan, depict the locations where traffic accidents occur most frequently. Areas distant from the center experience fewer traffic accidents, aligning with common knowledge that as Manhattan is a concentration of economic activity, there is a high volume of pedestrian and vehicular traffic, resulting in a higher density of traffic accidents.

2. **Then we used a bar plot to show the counts of crashes in each state.**

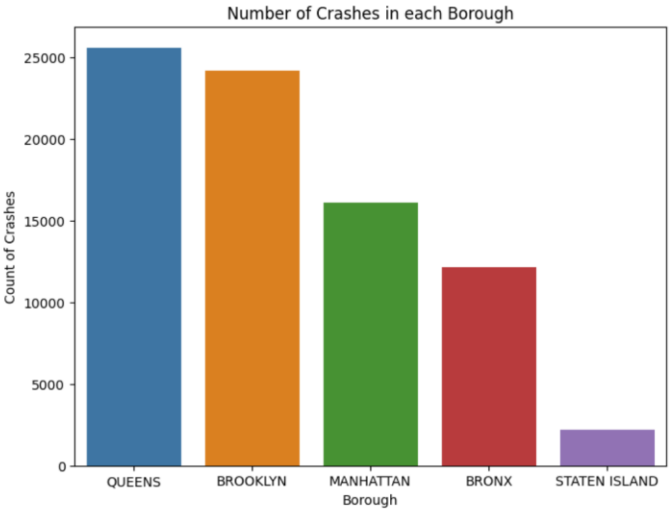

Fig. 2. Number of crashes in each borough

We find from Fig. 2 that the districts with the most traffic crashes are Queens, Brooklyn, and Manhattan. Even though Manhattan has the highest accident density, the Queens and Brooklyn districts have a larger area than Manhattan, which results in a higher traffic accident count.

Visualization and Analysis of Time Factors

1. **We first create visualizations of crash counts concerning hourly, daily, and monthly time intervals.**

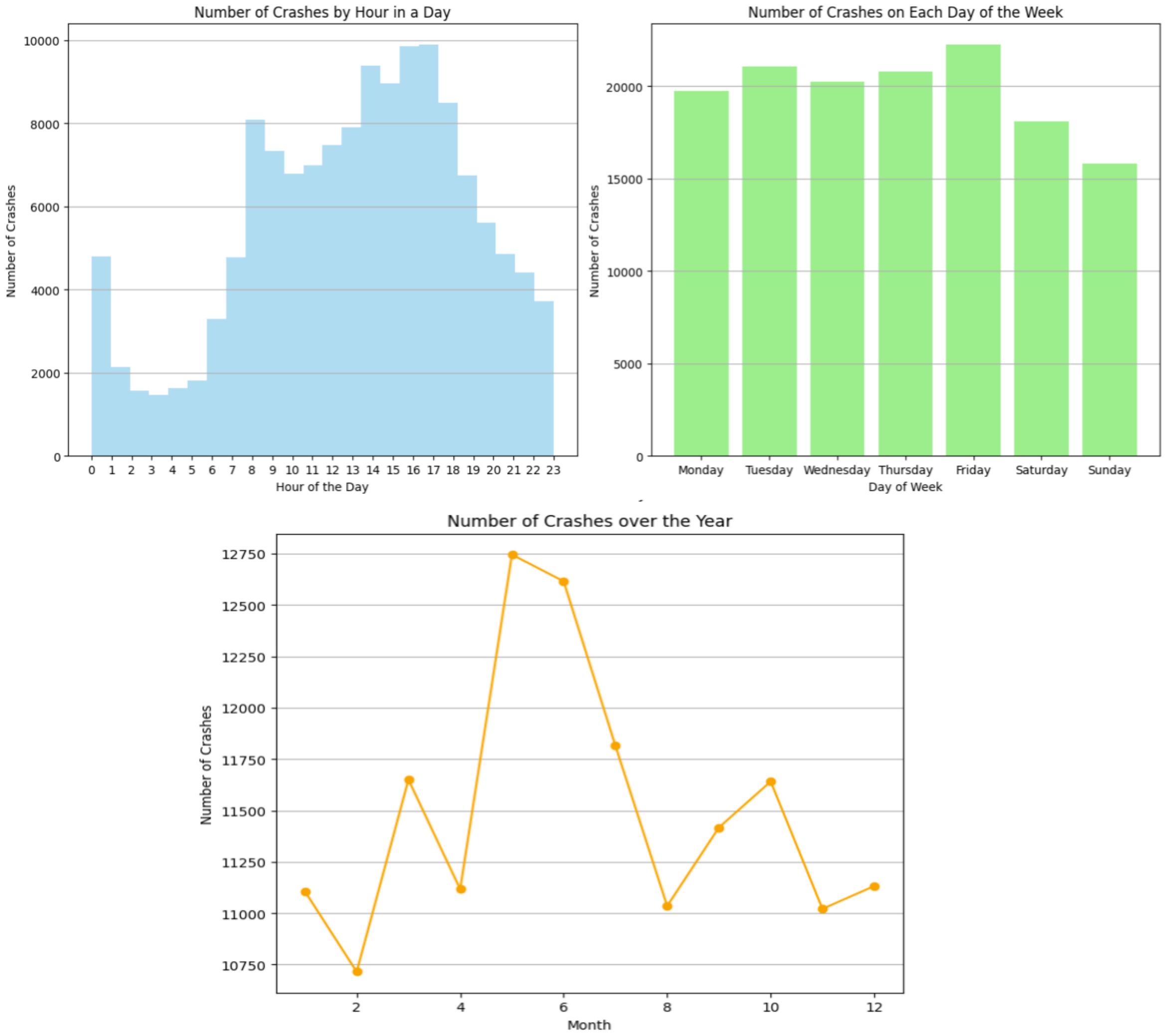

Fig. 3. Plots of crash counts concerning hourly, daily, and monthly time intervals.

From Fig. 3, we can find that:

1. In a single day, traffic crashes mainly concentrate from 8 am to 9 am, and 2 pm to 5 pm, which are the rush hours. This is reasonable due to the high traffic volume during morning and evening peak hours, prone to traffic accidents.
2. In a week, the number of crashes is high on weekdays (especially Friday), and low on weekends (especially Sunday). This may be because the traffic flow is

large and concentrated in the morning and evening peaks on weekdays, leading to traffic accidents.

3. In a year, the number of crashes is high in May and June, and low in February. This may be due to tourist peaks or some specific events.

2. **We use a heatmap to visualize the counts of crashes given a specific combination of time intervals and day of the week.**

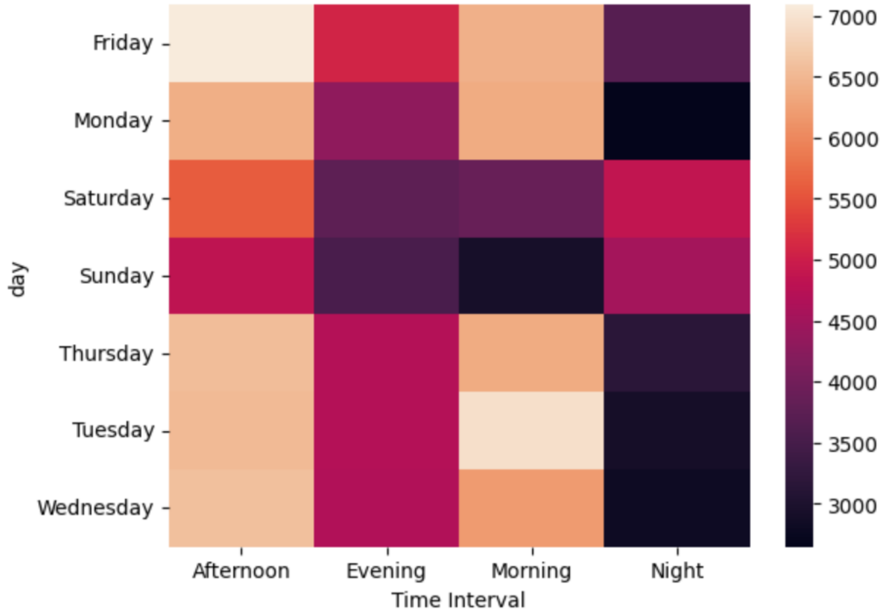

Fig. 4. Heatmap of the counts of crashes by day and time Interval

Through Fig. 4 we find that traffic accidents most happened on Friday afternoon and Tuesday morning. Even though Morning tends to be a time when traffic accidents are likely to happen, the frequency of traffic accidents happening on Sunday Morning is relatively low, which aligns with the common knowledge.

Crash Factor Analysis

1. Common Cause Analysis

We use bar plots to visualize the top 5 most common causes of crashes. As there are 5 contributing factors at most for each crash in our original dataset, we decided to count the contributing factor of vehicle 1 (which is the main participant of the traffic crash) as the cause of the crash.

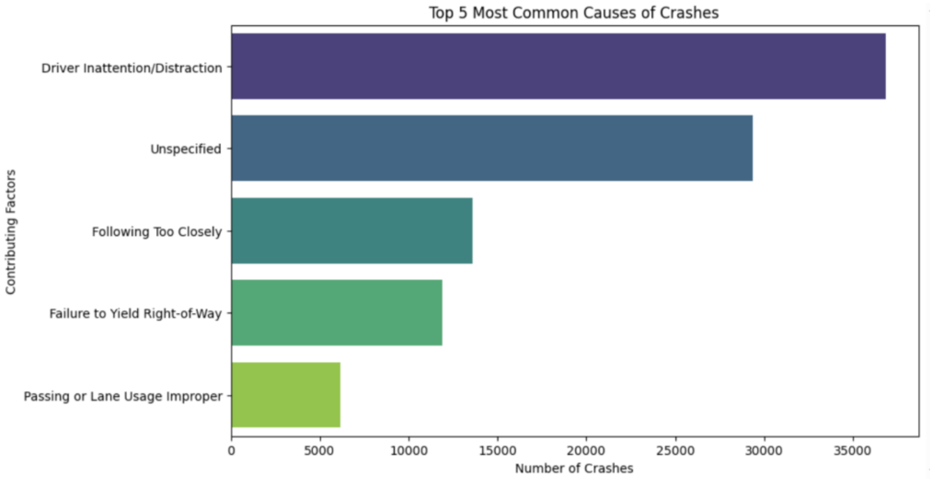

Fig. 5. Top 5 most common causes of crashes

We can find from Fig. 5 that driver inattention or distraction is the most common cause of traffic crashes.

2. Participants Analysis

We then calculate and visualize the sum of injuries and fatalities categorized by participant types (Pedestrians, Cyclists, and Motorists) using bar plots.

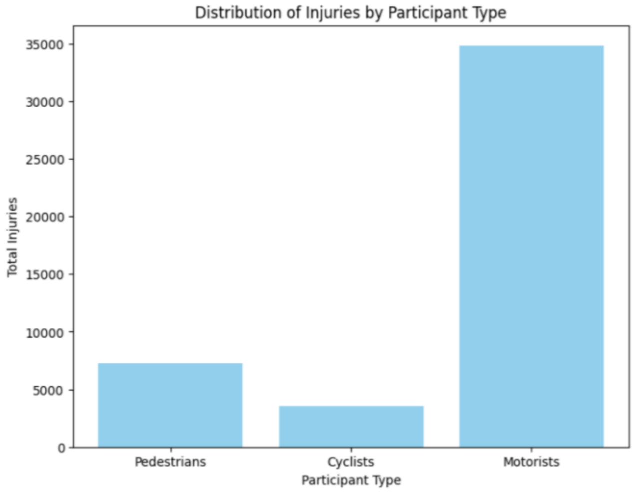

Fig. 6. Distribution of injuries by participant type

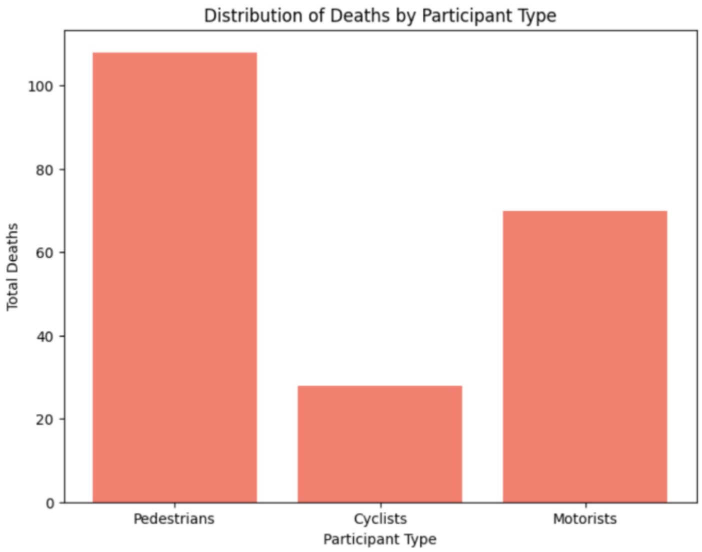

Fig. 7. Distribution of deaths by participant type

From Fig. 6, we can find that motorists have the highest count of injuries, this might be because most traffic crashes happen between motor vehicles, according to common sense. However, as we can find from Fig. 7, pedestrians have the highest number of deaths. This indicates that pedestrians are vulnerable groups and are more likely to die in traffic crashes. So, there needs to be more protection for vulnerable people like pedestrians on the road.

Data Analysis

In the data analysis part, we use the Gradient Boosting method for modeling and find several important features regarding the data.

Feature Engineering:

First, we classify the severity of the crashes based on the number of persons injured and killed. To be specific, we identify the crash to be 'severe' if there are more than 0 people dead or more than 2 people injured. We identify the severity to be 'moderate' if there are more than 0 people injured. Else, for those crashes with no one dead or injured, we identify the crash's severity to be 'minor'.

Initially, we translate severity labels into numerical values. We choose specific attributes from the primary dataset, including crash times, and incorporate additional data from secondary sources, such as speed limit and weather conditions, as our selected features. Additionally, we utilize one-hot encoding to process categorical data, such as 'day' and 'Time Interval'.

Training and Testing:

In this part, we first split the train and test data from the dataset, then apply gradient boosting to predict the severity of the crashes. The test size occupies 33% of the entire dataset.

We also visualize the importance of the features to see the leading features affecting the severity of crashes.

The feature importance in gradient boosting is shown below:

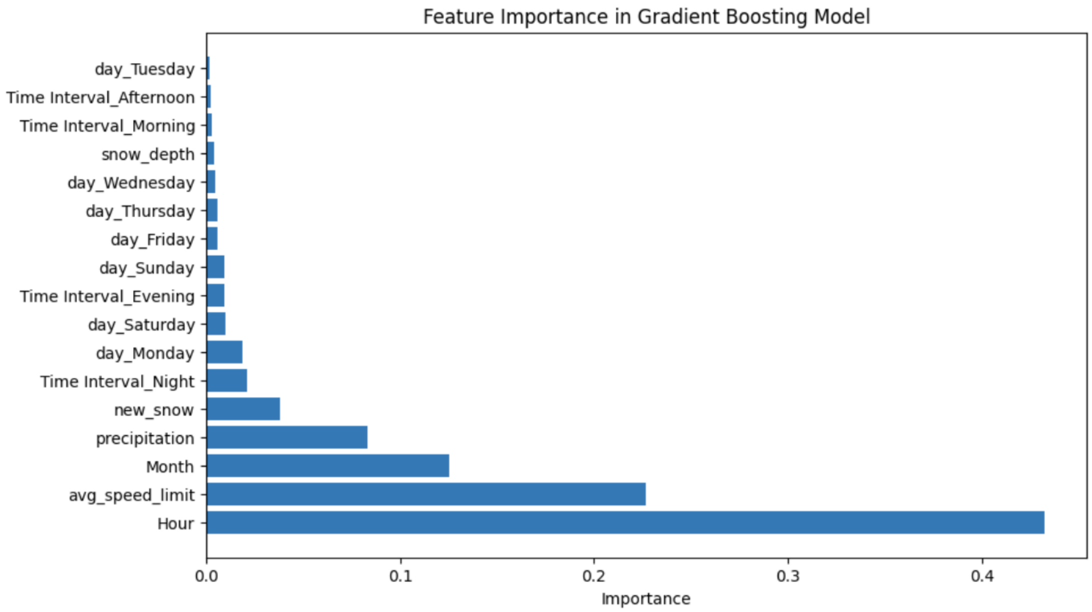

Fig. 8. Feature importance in Gradient Boosting Model

As we can see from Fig. 8, Hour, the average speed limit of the street, Month, precipitation, and new_snow are the top 5 leading features affecting the severity of crashes.

Finally, we use accuracy_score from sklearn to obtain the accuracy of our model based on the test set and get the accuracy of 0.758.

Conclusion

Based on our data analysis, we are able to answer the questions we raised at the beginning.

- First, according to Fig. 1, accidents occur most frequently near Manhattan. These areas are shown in red and correspond to Manhattan's high economic activity and traffic volume. Although Manhattan has the highest accident density, districts such as Queens and Brooklyn, which have larger areas, also experience a significant number of accidents.

- We also found that the frequency of accidents peaks during morning rush hours from 8 am to 9 am and between 2 pm to 5 pm, particularly on weekdays, especially Fridays. In contrast, the number of accidents decreases notably on weekends, particularly on

Sundays. Moreover, higher accident occurrences are observed in May and June, whereas February witnesses fewer accidents.

- Second, injuries and fatalities were considered when evaluating the severity of traffic crashes. Driver inattention or distraction emerged as the most common cause of crashes. Moreover, while motorists suffer the highest count of injuries due to crashes between vehicles, pedestrians face a higher risk of fatalities, highlighting their vulnerability in traffic accidents.

- Last, In our data analysis, we found that several factors significantly impact the severity of crashes. Key contributors include the time of day, average street speed limit, month, precipitation, and new snowfall. These factors play a vital role in predicting the severity of accidents. These insights were derived from our analysis of the dataset containing New York City's motor vehicle collision information from 2019, considering factors such as crash times, street details, weather conditions, and severity labels. The analysis, conducted to understand accident patterns and contributing factors, aimed to aid in enhancing road safety and reducing the frequency and severity of traffic accidents in the city.

Statement of Works

We first worked together to search for the data set and decide on the topic, then we decided who was responsible for one specific part of this project.

Yuexin Min is mainly responsible for data visualization, data analysis, and additional interpretation for the notebook. Jiaou Chen is mainly responsible for data manipulation and writing the report. Although the coding is done by each person, throughout the whole process of coding, we would make outlines for each step together. In the data manipulation part, we first decided what feature could be used as the key to join two tables. When we encountered challenges with matching the street name, we worked together to find the possible reasons and look for solutions to unify the form of keys to join two tables. In the data visualization part and data analysis, we decided together what could be the possible factors affecting the severity. We discussed some conversions on data. We decided how to categorize the hour into several data categories.

After finishing the report, we checked the content of the report and notebook together, making some modifications and fixing grammatical errors.

References:

L. G. Cuenca, E. Puertas, N. Aliane and J. F. Andres, "Traffic Accidents Classification and Injury Severity Prediction," 2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE), Singapore, 2018, pp. 52-57, doi: 10.1109/ICITE.2018.8492545

S. Ahmed, M. A. Hossain, M. M. I. Bhuiyan and S. K. Ray, "A Comparative Study of Machine Learning Algorithms to Predict Road Accident Severity," 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS), London, United Kingdom, 2021, pp. 390-397, doi: 10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00069.