Syracuse University
School of Information Studies

Final Project Report
on
Analyzing Motor Vehicle Collisions: Insights from Crash Data

IST462-SCRIPTING FOR DATA ANALYSIS

Chenze Chen, Ding(Wesley) Chen, Mason Freer
12/12/2023

Introduction

Transportation is a fundamental part of modern society, but it also carries significant risks, particularly the occurrence of motor vehicle collisions. Such incidents lead to human suffering, loss, or even a negative impact on the economy. Our project aims to have a scrutiny into motor vehicle collision analysis with the goal of identifying key trends and insights to enhance road safety.

This project seeks to analyze a dataset of motor vehicle collisions reported in New York City and identify their primary contributing factors.The research examines several aspects such as time-related trends, correlation between different variables, and geographical distribution. By focusing on such aspects, we believe the study provides a special perspective on the factors that influence motor vehicle accidents. This analysis is essential as it can guide policymakers, urban planners, and public safety officials in imposing effective strategies to decrease the frequency and severity of such incidents.

Our project stands on data analysis, utilizing data processing and analytical techniques mainly in Python. With this approach, the project aims to provide practical insights and contribute to the broader discussion on promoting road safety and preventing accidents. As we explore the data and its analysis, our report illuminates the multifaceted aspects of motor vehicle accidents, offering a data-driven pathway towards safer roads and communities.

Methodology

This section outlines the methodological approach applied for the data cleaning and transformation processes, crucial for ensuring the accuracy and reliability of our analysis.

1. Data Cleaning

The initial step in our methodology focused on the removal of incomplete records, particularly in the context of the 'NUMBER OF PERSONS INJURED' and 'NUMBER OF PERSONS KILLED' columns.The rationale behind this decision was clear. We think the absence of data in these key fields reduces the significance of the corresponding records for our analysis. Here are what we did:

- Identifying and removing rows with missing values in both 'NUMBER OF PERSONS INJURED' and 'NUMBER OF PERSONS KILLED' columns
- Verification of the removal process was conducted by examining the sum of null values after process was completed
- An additional review of the NA rows was undertaken to ensure that their removal was justified

2. Data Transformation

Following the cleaning phase, we proceeded with data transformation. This stage was important in refining the dataset for more focused analysis. Steps included the followings:

- Elimination of columns considered irrelevant for our analysis. These included 'ON STREET NAME', 'CROSS STREET NAME', 'OFF STREET NAME', 'COLLISION_ID', and 'LOCATION'.
- Checking for null values was conducted to ensure the integrity of the dataset.
- Identification and removal of duplicate records were carried out, further purifying the data for analysis.
- The dataset's categorical data was refined by correcting typographical errors in 'CONTRIBUTING FACTOR VEHICLE 1', such as replacing 'Illnes' with 'Illness' and standardizing entries like 'Drugs (Illegal)' to 'Drugs (illegal)'.

3. Outlier Detection

In order to ensure the statistical robustness of our dataset, we utilized boxplot visualizations to detect and assess outliers in the columns for 'NUMBER OF PERSONS

INJURED' and 'NUMBER OF PERSONS KILLED'. This was a important step for identifying extreme values that could potentially skew our analysis. Our examination found no significant outliers in these categories.

4. Correlation Analysis

As part of our exploratory data analysis, we conducted a correlation analysis to identify potential relationships between different variables. This was done through computing a correlation matrix and visually depicting the resulting matrix through a heatmap to make identification of any notable correlations.

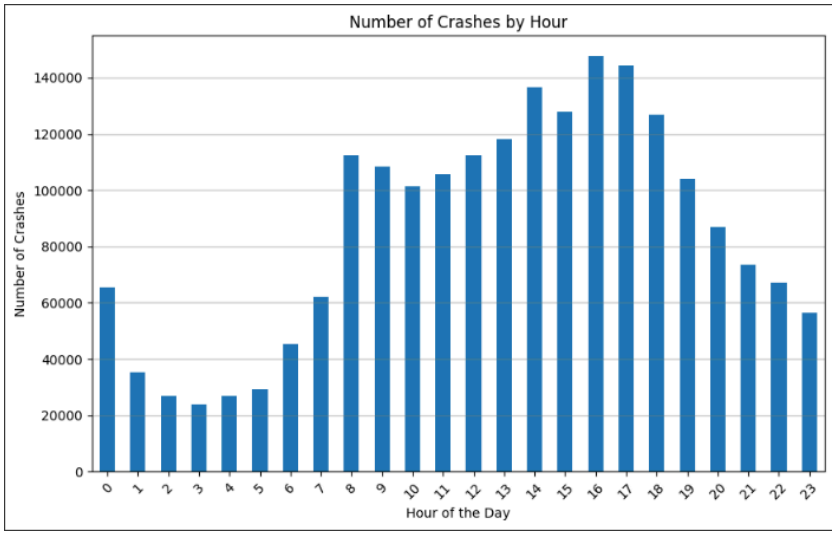
Data Analysis and Insights

1. Time Analysis

In this section, we examine the time-related aspects of car crashes in our data to find patterns and trends. We analyze different time periods: Hour, Day, Month, and Year, and their respective number of crashes and impact on individuals.

1. Hour and Number of Crashes

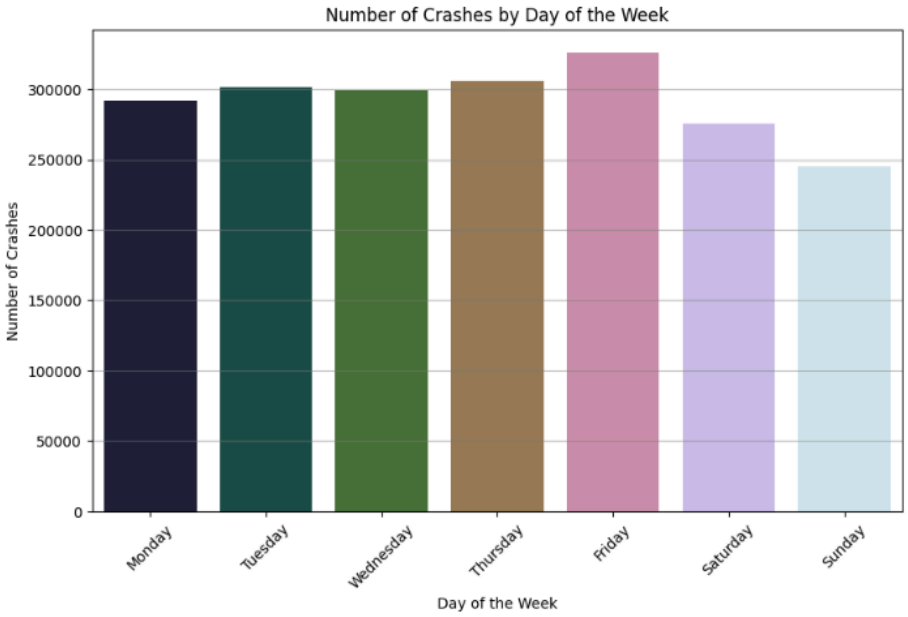
- A gradual increase in crashes is observed from the early hours, peaking around 8 AM, likely corresponding to morning rush hour.
- Between 10 AM and 3 PM, a significant rise in crashes is noted, with the peak occurring between 2 PM and 3 PM. This suggests a high incidence of crashes during early to mid-afternoon.
- Post 5 PM, a decline in crashes is evident, becoming more pronounced after 7 PM, with the lowest frequency occurring post 10 PM.



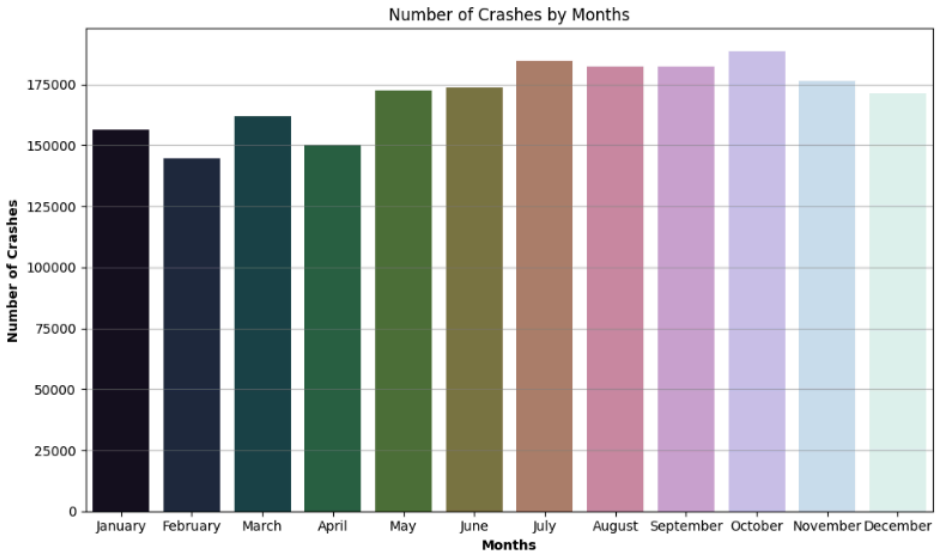
2. Day and Number of Crashes

- Fridays experience the highest number of crashes, possibly due to increased traffic from weekend plans or end-of-week commuting.

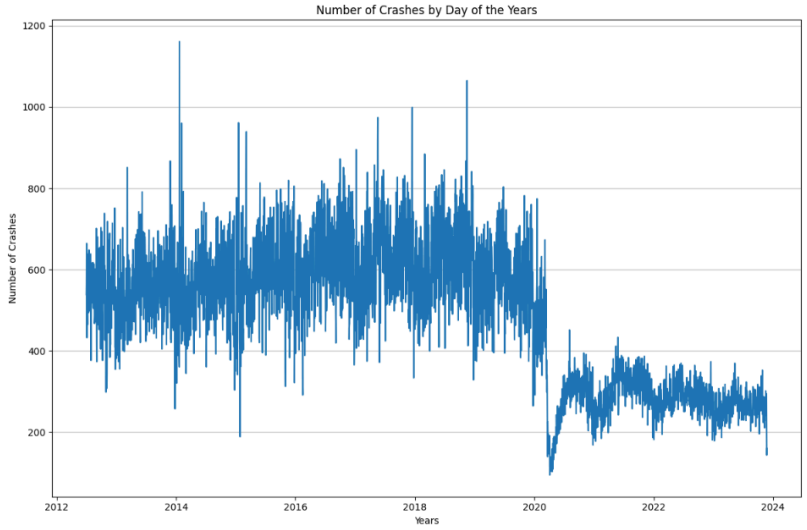
- Sundays show the lowest crash numbers, with a noticeable decrease from Saturday, likely due to reduced commuter traffic and fewer vehicles on the road overall.
- A progressive increase in crashes is seen as the week advances, peaking on Friday.



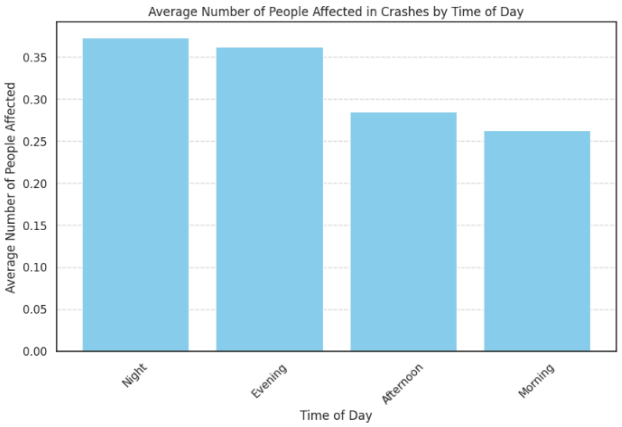
3. Month and Number of Crashes
   - A seasonal variation in crashes, with summer months (June, July, August) and December reporting higher crash numbers. This could be due to increased travel during summer holidays and challenging winter driving conditions in December.
   - Lower crash numbers in February and November might reflect reduced travel activity or other seasonal influences.

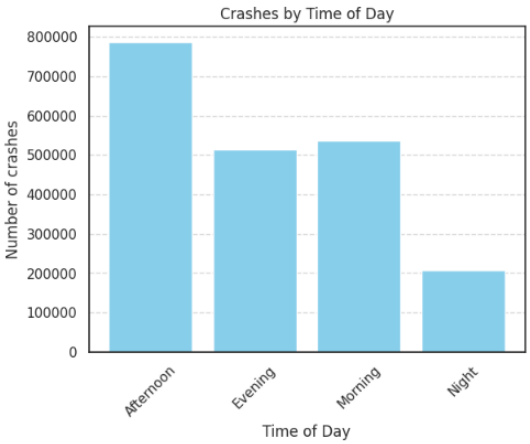


4. Year and Number of Crashes
    - A noticeable overall decline in crashes post-2020, possibly due to improved road safety measures or external factors like the COVID-19 pandemic.
    - Recurring periods of higher and lower crash numbers suggest potential seasonal trends or cyclic events affecting crash rates.

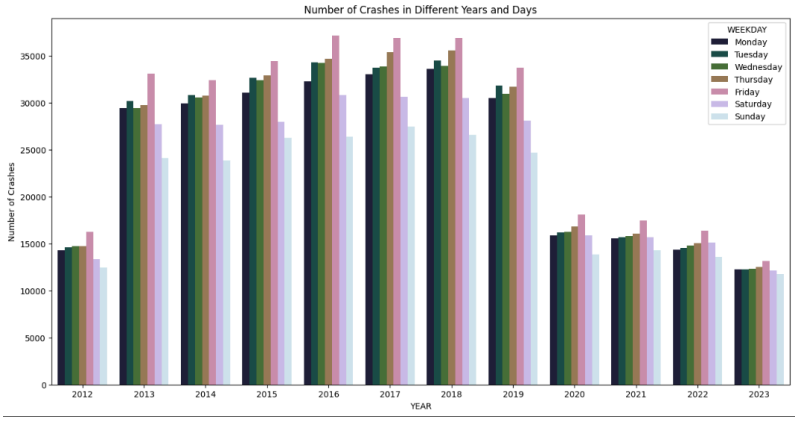


5. Time and People Affected
    - Nighttime records the highest average number of people affected per crash, potentially due to factors like reduced visibility and driver fatigue.
    - Mornings have the lowest average impact, indicating that crashes during this time may be less severe or involve fewer people.

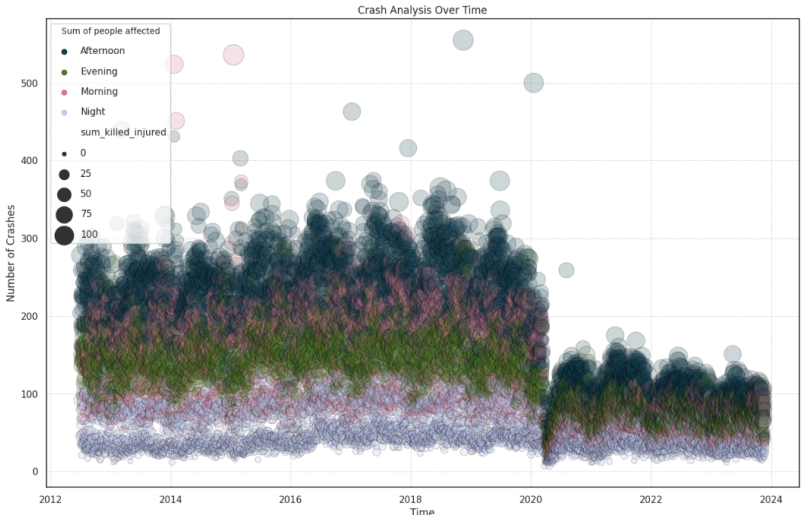




6. Year, Day, and Number of Crashes
    - Relative stability in crash numbers from 2012 to 2019, with no significant year-over-year fluctuations.
    - Fridays consistently record high crash numbers, while mid-week days like Wednesday and Thursday also report higher crashes compared to the start of the week.
    - The decline starting in 2020 aligns with the onset of the COVID-19 pandemic, reflecting reduced traffic and consequently fewer crashes.



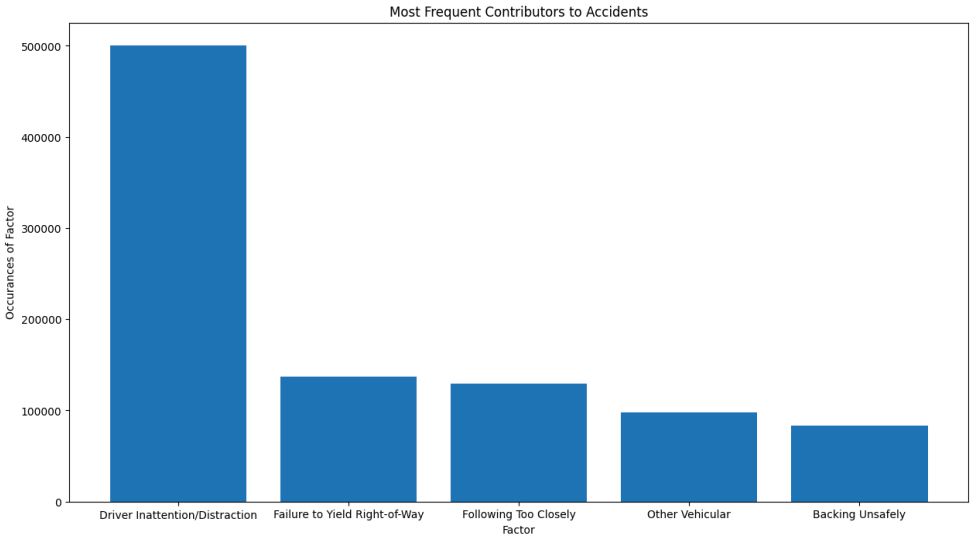
7. Year, Time, People Affected, and Number of Crashes
    - No clear pattern in the severity of crashes concerning the time of day, as indicated by larger bubbles (representing higher numbers of people affected) scattered throughout the plot.
    - Afternoon and Evening periods exhibit more frequent crashes compared to Morning and Night.



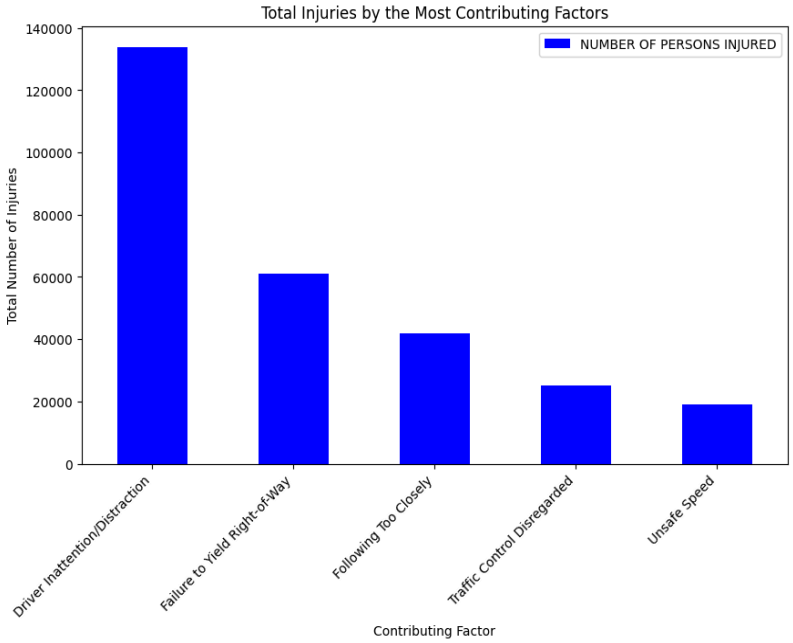
2. Contributing Factors Analysis

In this section, we analyzed the most contributing factors related to all of the motor vehicle collisions in NYC.

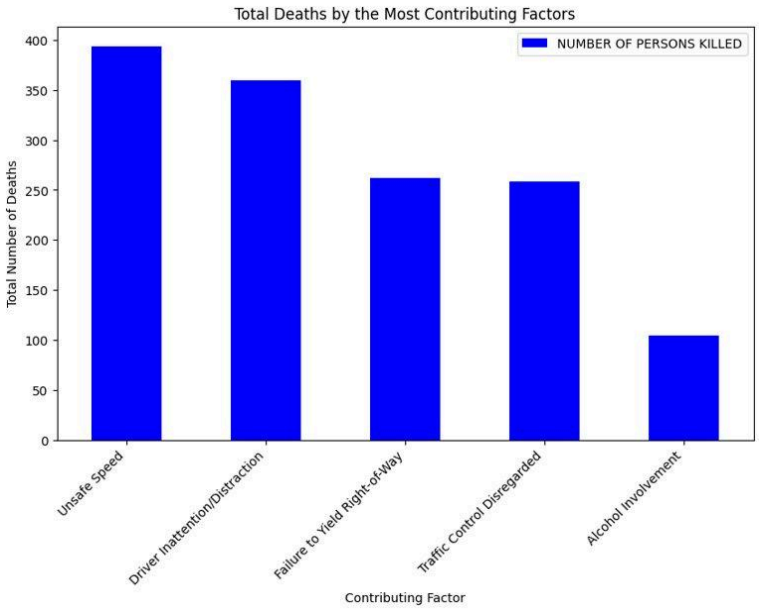
1. Distribution of Contributing Factors
    - Unspecified was the largest contributing factor, i.e the reason behind that vehicle being involved in the collision is unknown.
    - Filtering out unspecified contributing factors, driven inattention was the most common contributor to accidents, followed by a failure to yield to a right of way, following too closely, other vehicles, and backing out unsafely.



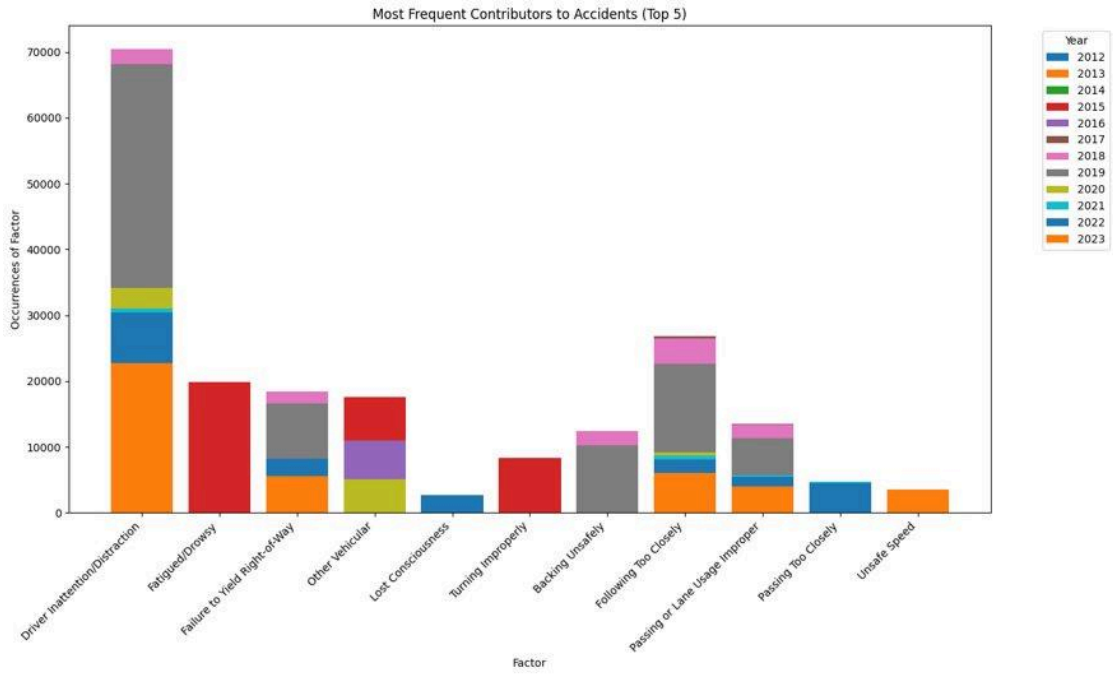
2. Distribution of Total Injuries by Contributing Factors
    - Driver Inattention resulted in the most injuries at approximately 130k injuries
    - ‘Unsafe Speed’ and ‘Traffic Control Disregarded’ were less common contributing factors in the initial distribution but were in the top 5 contributing factors in injuries. These are high risk behaviors that are likely to cause injury to themselves or others.



3. Distribution of Total Deaths by Contributing Factors
    - Unsafe Speed resulted in the most deaths at approximately 395 deaths total
    - 'Unsafe Speed,' 'Traffic Control Disregarded,' and 'Alcohol Involvement' were less common contributing factors in the initial distribution but were in the top 5 contributing factors in deaths. These activities are high risk behaviors, especially alcohol involvement that are likely to cause accidents that get someone killed.



4. Most Contributing Factors over Time Distribution
    - Distribution that showcases the top 5 contributing factors each year since 2012 in a stacked bar chart
    - Shows that in years like 2015, drowsiness was a significant factor for accidents, but not so prevalent in other years.
    - Helpful for stakeholders to understand historically what factors contributed most to accidents and understand what to change.



5. 2023 Contributing Factor Analysis
    - Isolated accidents that occurred in 2023 to discover the most prevalent contributing factors in recent times.
    - ‘Driver Inattention,’ ‘Following Too Closely,’ ‘Failure to Yield Right of Way,’ ‘Passing or Lane Usage Improper,’ and ‘Unsafe Speed’ were the top 5 contributing factors.

3. Geospatial Analysis

The data frame utilized for geospatial analysis of collision hotspots is "crash_data," resulting from previous data cleaning and transformation. To enhance the dataset's quality, we drop rows with missing values in columns such as 'BOROUGH,' 'ZIP CODE,' 'LATITUDE,' and 'LONGITUDE.' The 'CRASH DATE' column is converted to datetime, and a new dataframe

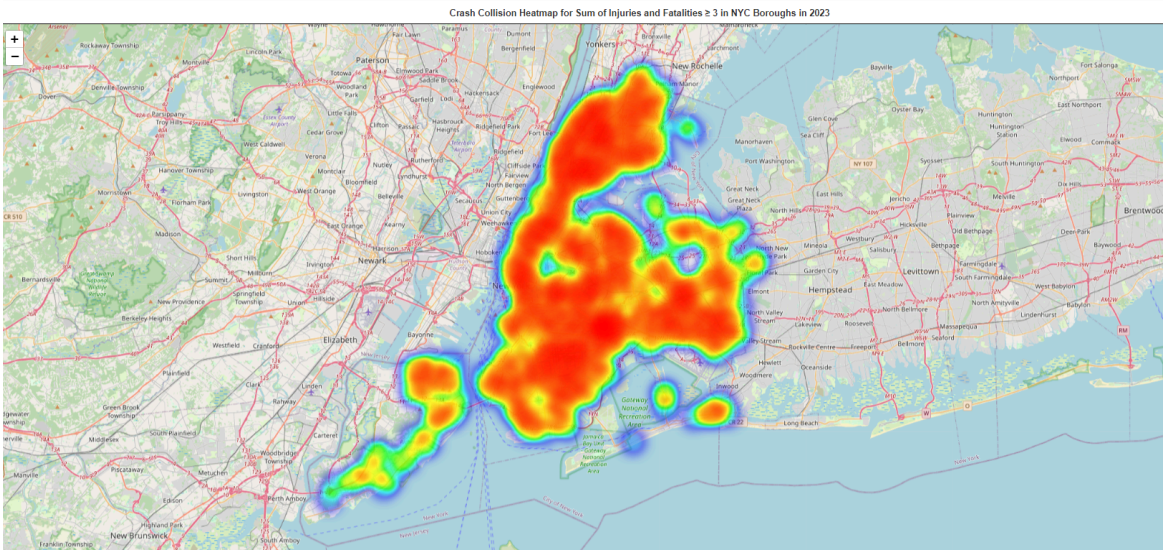
named "crash_data_2023" is created by copying from crash_data, filtering for the year 2023. We further refine crash_data_2023 by summing columns 'NUMBER OF PERSON INJURED' and 'NUMBER OF PERSON KILLED' to generate a new column, "sum_killed_injured." The dataframe is then filtered for entries where "sum_killed_injured >= 3.0," resulting in "crash_data_2023_filtered," focusing on accidents with this severe impact.

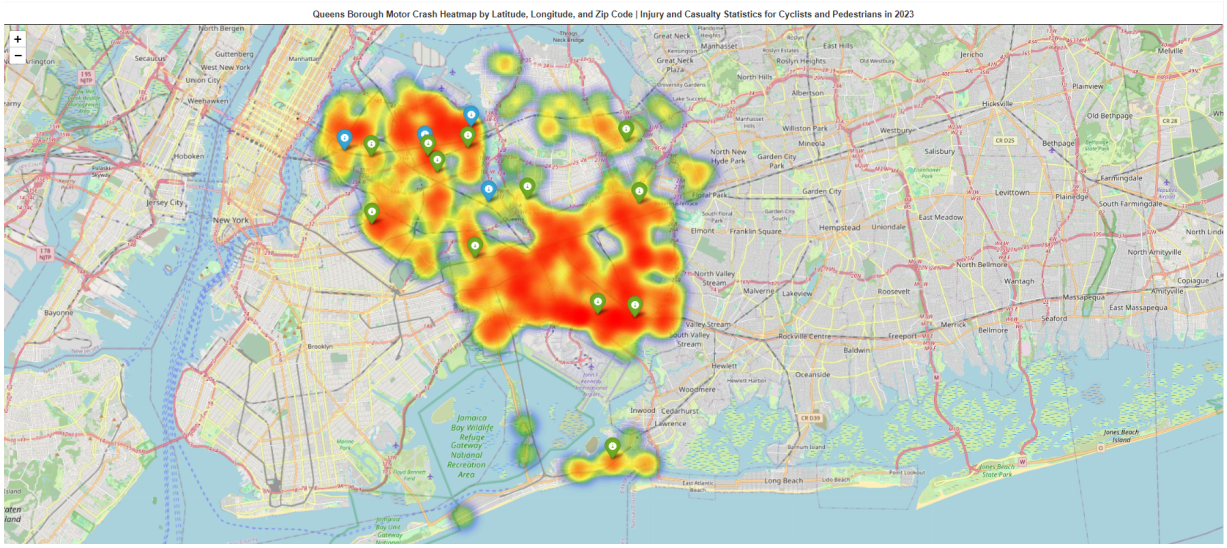
For data visualizations, we utilize the Folium library to create an interactive map illustrating the locations of motor vehicle collisions with injuries and fatalities in New York City during 2023. The map is centered at latitude 40 and longitude -74 with an initial zoom level of 12. A HeatMap layer is generated through lists of (latitude, longitude). Observing the collision heatmap, indicating a sum of injuries and fatalities greater than or equal to 3 in the NYC boroughs in 2023, reveals Queens borough's elevated collision rate compared to other boroughs. Consequently, we decided to conduct a detailed analysis of the Queens crash data.

Initially, we filter the preprocessed crash data for Queens and calculate the sums of injuries and fatalities for cyclists, pedestrians, and motorists separately. A base map centered around Queens is created using the Folium library. Rows with missing values in relevant columns are dropped, and ZIP CODE data is processed for better representation. Similar to our approach for the five NYC boroughs earlier, the script generates a HeatMap layer color-coded by ZIP CODE, illustrating the concentration of incidents. Notably, the code iterates through Queens crash data, adding markers for locations with cyclist and pedestrian incidents, colored in blue and green, respectively.

The resulting map offers an interactive visualization of motor crash statistics in Queens, providing insights into areas with notable cyclist and pedestrian incidents in 2023. Notably, most incidents occur on the left side of the Queens borough. For future analysis, we plan to investigate the streets marked by these incidents to gain a deeper understanding of the road

situation and explore other potential factors contributing to collisions.





Syracuse University
School of Information Studies

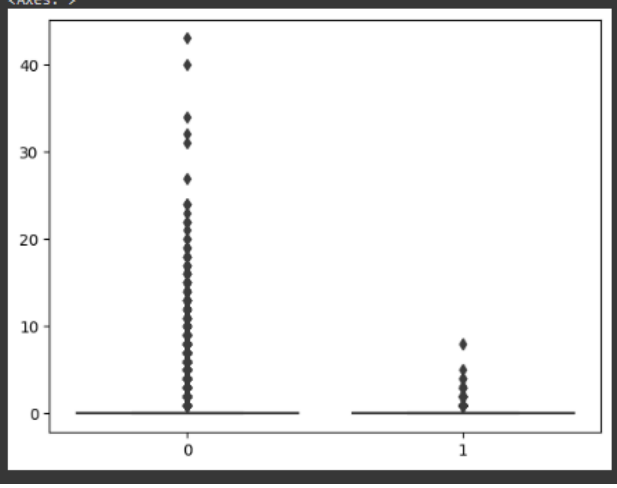
Appendices

```
# checking for outliers through boxplot


sns.boxplot(data=[data_cleaned_transformation['NUMBER OF PERSONS INJURED'], data_cleaned_transformation['NUMBER OF PERSONS KILLED']])
#data.boxplot(column=['NUMBER OF PERSONS INJURED'])
#data.boxplot(column=['NUMBER OF PERSONS KILLED'])

# No outliers present
```

```
<Axes: >
```




```
[ ] # check for correlation

correlation = data.corr()
correlation.style.background_gradient(cmap='coolwarm')
```

```
<ipython-input-14-19e44d87eeb1>:3: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  correlation = data.corr()
```

	LATITUDE	LONGITUDE	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	COLLISION_ID
LATITUDE	1.000000	-0.963426	-0.007853	-0.001107	-0.004022	-0.002014	-0.004503	0.000489	-0.004705	0.000181	-0.036946
LONGITUDE	-0.963426	1.000000	0.008372	0.001147	0.004269	0.001975	0.003801	-0.000509	0.005370	-0.000066	0.036562
NUMBER OF PERSONS INJURED	-0.007853	0.008372	1.000000	0.014814	0.277108	0.002159	0.173395	-0.001235	0.906171	0.020714	0.062793
NUMBER OF PERSONS KILLED	-0.001107	0.001147	0.014814	1.000000	0.005597	0.691431	-0.004031	0.278299	0.014597	0.661858	0.004652
NUMBER OF PEDESTRIANS INJURED	-0.004022	0.004269	0.277108	0.005597	1.000000	0.011018	-0.032406	0.000585	-0.066278	-0.002941	-0.002079
NUMBER OF PEDESTRIANS KILLED	-0.002014	0.001975	0.002159	0.691431	0.011018	1.000000	-0.002830	0.021304	-0.000935	0.005353	0.000679
NUMBER OF CYCLIST INJURED	-0.004503	0.003801	0.173395	-0.004031	-0.032406	-0.002830	1.000000	0.001642	-0.051469	-0.003689	0.028931
NUMBER OF CYCLIST KILLED	0.000489	-0.000509	-0.001235	0.278299	0.000585	0.021304	0.001642	1.000000	-0.001875	-0.000239	0.002793
NUMBER OF MOTORIST INJURED	-0.004705	0.005370	0.906171	0.014597	-0.066278	-0.000935	-0.051469	-0.001875	1.000000	0.024057	0.055352
NUMBER OF MOTORIST KILLED	0.000181	-0.000066	0.020714	0.661858	-0.002941	0.005353	-0.003689	-0.000239	0.024057	1.000000	0.004478
COLLISION_ID	-0.036946	0.036562	0.062793	0.004652	-0.002079	0.000679	0.028931	0.002793	0.055352	0.004478	1.000000