Data Exploration and Risk Assessment of Vehicle Collisions in NYC

Student: Dawood Aijaz
Matriculation Number: 23282229

1.Introduction:

Traffic accidents in major cities pose significant risks to public safety, and New York City is no exception. This project focuses on analyzing motor vehicle collisions in NYC to understand patterns and factors that contribute to these incidents.
The primary goal is to answer the question: "**What are the primary factors contributing to motor vehicle collisions in New York City?**" Through data analysis and visualization, we will explore temporal patterns (such as time of day and seasonal variations), spatial distributions (identifying accident hotspots), and various contributing factors (like driver behavior and environmental conditions).

2.Used Data:

2.1.Dataset Description:

The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police-reported motor vehicle collisions in NYC. The police report (MV104-AN) must be filled out for collisions where someone is injured or killed, or where there is at least $1000 worth of damage

2.2.Dataset Details:

- Source: NYC Open Data platform, Motor Vehicle Collisions - Crashes | NYC Open Data
- Dataset Name: "Motor Vehicle Collisions - Crashes"
- Agency: Police Department (NYPD)
- Rows: 2.14M
- Columns: 29

Licenses:
- License: NYC OpenData. Motor Vehicle Collisions - Crashes | NYC Open Data
- Why Allowed: This license permits unrestricted use, provided attribution is given.

The data pipeline processes crashes where either an injury/fatality occurred or property damage exceeded $1,000, as per NYPD reporting requirements (MV104-AN).

1. Temporal Information:
   - CRASH DATE: Standardized datetime format
   - CRASH TIME: Formatted time values
2. Location Data:
   - LATITUDE/LONGITUDE: Cleaned geographic coordinates (null values removed)
   - BOROUGH: Categorized with 'Unknown' for missing values
   - ZIP CODE: Standardized to 5-digit format (defaulting to '00000' if missing)
3. Severity Metrics (all converted to integers):

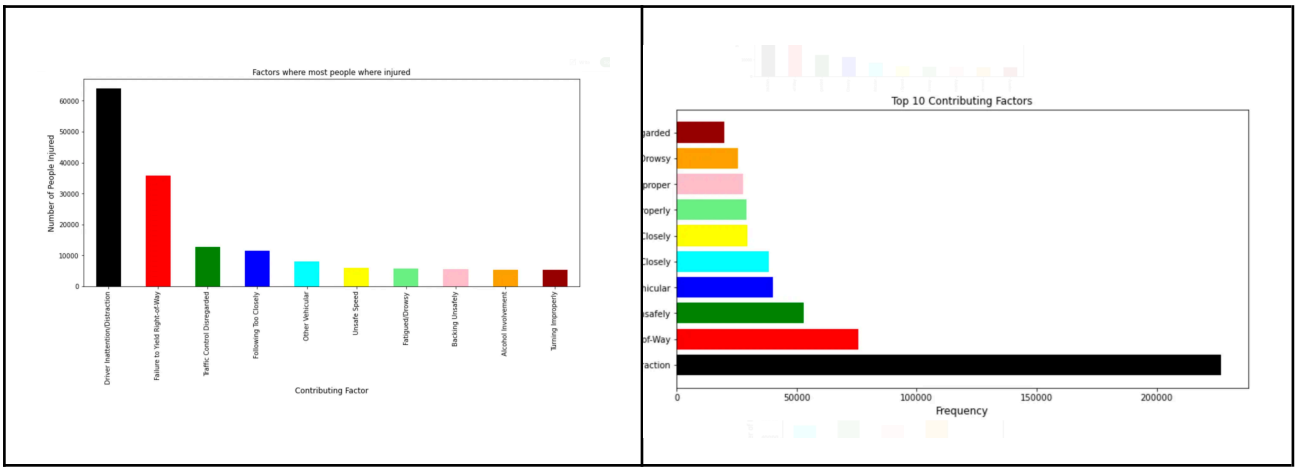
- NUMBER OF PERSONS INJURED/KILLED
- NUMBER OF PEDESTRIANS INJURED/KILLED
- NUMBER OF CYCLIST INJURED/KILLED
- NUMBER OF MOTORIST INJURED/KILLED

The pipeline implements specific data quality measures, including handling missing values, standardizing formats, and removing records with invalid geographic coordinates.

3.Analysis:

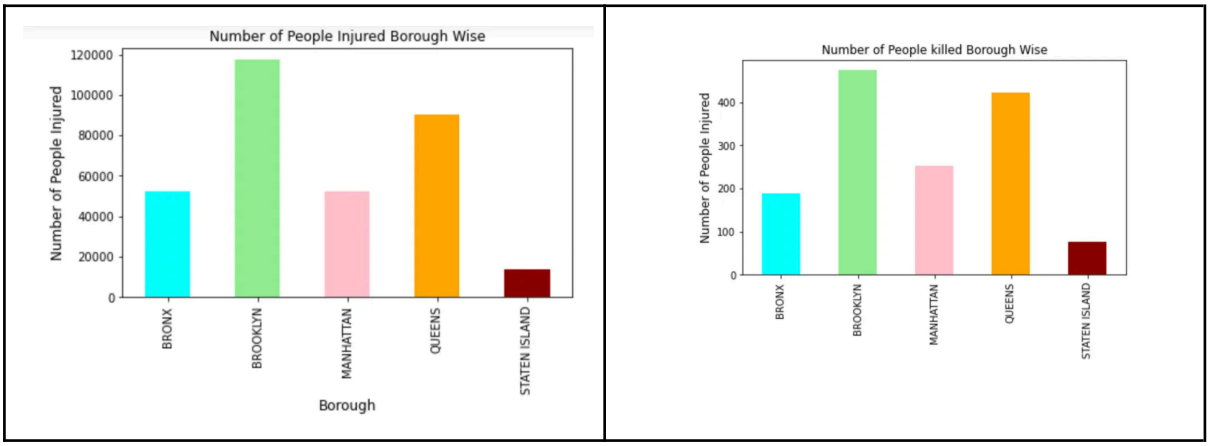
The primary objective of conducting Exploratory Data Analysis (EDA) and visualization is to enhance the understanding of a dataset asking the underlying questions.

1.Which factors contributed in highest injuries/deaths and what is the frequency of contributing factor?



The plot indicates that the average number of people injured in collisions is around 6,500. It highlights the top 10 contributing factors so that effective measures can be taken to address the most significant causes. From these factors, it appears that driver inattention, failure to yield the right-of-way, and disregard for traffic control contribute to the highest number of injuries.

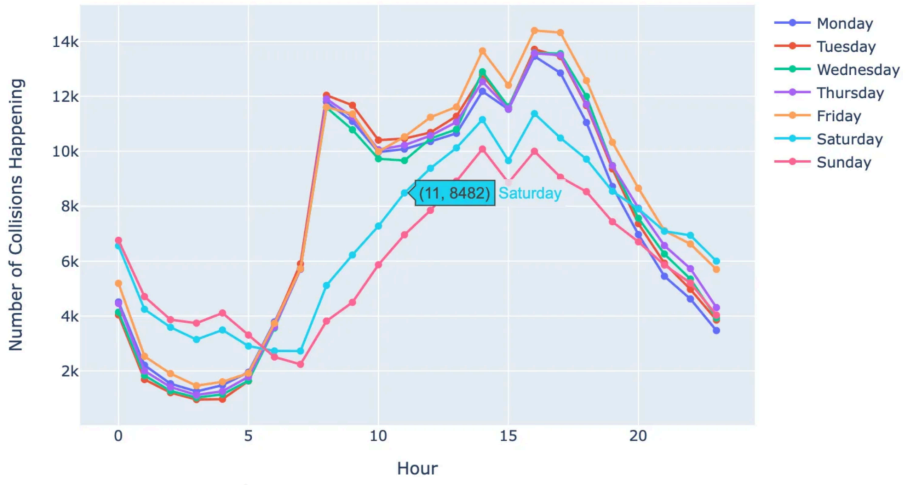
2.Which Borough has maximum collisions happening and how many people were injured and killed?



The two bar plots indicate that Brooklyn experiences the highest number of fatalities and injuries, followed by Queens. Upon reviewing various published articles, I found that Brooklyn's numerous

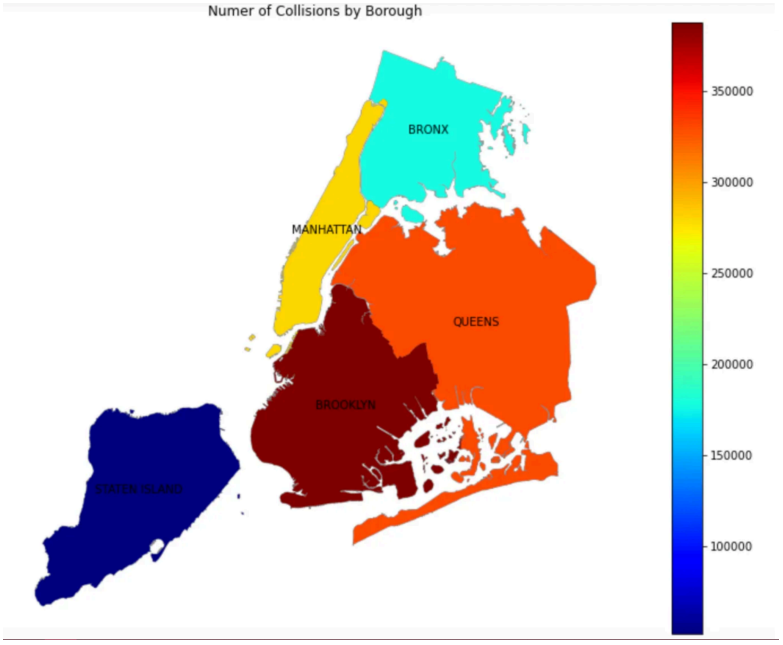
dangerous intersections and frequent driver inattention contribute to its high collision rate. Meanwhile, Staten Island has the fewest collisions, likely due to its population density being roughly one-fifth that of Brooklyn.

3.What is weekly analysis of collisions taking place in boroughs by hours?



Collisions predominantly occur on weekdays (Monday–Friday), likely due to heavier traffic from work and routine errands. They peak between 9–11 AM—the morning rush hour—and then decrease before rising again in the evening, when people are returning from work, ultimately tapering off later in the day.

4.What is the number of collisions for each borough and how to view it on the NYC map?



When I used the jet colormap, dark red highlights clearly showed that Brooklyn and Queens Borough had the highest number of collisions.

5.Which streets are frequently facing collisions?

Aim was to identify the most hazardous streets in NYC by using a Counter to generate key–value pairs of street names and their collision frequencies. After sorting the results in descending order to focus on the most dangerous streets, I created a word cloud based on the Counter output. The size of each word in the cloud indicates how frequently collisions occur there: the larger the word, the higher the number of incidents.

4.Conclusion:

Through comprehensive data exploration and predictive modeling on NYC collision data, several critical insights emerge. Weekday peak hours—morning (9–11 AM) and evening rush hour—consistently show elevated collision rates, driven by heavy traffic and driver inattention. Borough-level data indicates Brooklyn and Queens experience the highest number of accidents, likely due to dense population and a larger volume of vehicles on the road. Weather conditions, especially cloudy skies, also correlate with higher collision frequencies, underscoring how reduced visibility affects driving safety.

By pinpointing these patterns and factors—such as time of day, borough density, and weather—targeted interventions can be implemented to prevent collisions, including improved traffic management, driver awareness campaigns, and road infrastructure enhancements. Ultimately, this data-driven approach offers a tangible pathway toward reducing accidents and promoting safer roadways across New York City.