**6356 Case Project 1**

**(Fall 2024)**

**Analysing Urban Traffic Collisions (New York):**

**Insights into Time, Location, and Contributing Factors for Safer Roads**

**Introduction:**

Motor vehicle collisions are a persistent issue in urban environments, particularly in a bustling metropolis like New York City. With millions of vehicles, pedestrians, and cyclists navigating its streets daily, the risk of accidents is ever-present. This project seeks to analyze traffic collision data from New York City to uncover patterns and insights related to when and where collisions are most likely to occur, the severity of these incidents, and the contributing factors such as speeding or failure to yield. By examining key variables such as the time of day, location, and types of vehicles involved, this analysis aims to provide actionable insights that can help city planners, traffic management authorities, and policymakers improve road safety and reduce accidents across the five boroughs.

There are many questions which can be answered with so many variables in the picture, we have selected the following questions to research as they can answer and solve some very important issues. They are:

**Q1) Which boroughs in New York City experience the highest frequency of motor vehicle collisions, and what are the potential contributing factors in these areas? Based on the boroughs what were the top worst incidents and what are their main contributing factors?**

**Q2) How does the time of day (Day vs. Night) impact the frequency and severity of collisions, and What was trend over the years?**

**Q3) What are the most common contributing factors (e.g., speeding, failure to yield, impaired driving) in collisions, and how do these factors differ between boroughs or vary with the time of day?**

**Q4)** **How does the type of vehicle involved in a collision (e.g., car, truck, motorcycle) correlate with the severity of the accident, and are certain vehicle types more likely to result in fatalities or serious injuries?**

**DATABASE AND ITS STRUCTURE:**

**What does the data describe?**

The dataset provides detailed records of motor vehicle collisions that occurred in New York City. It captures critical information about each collision, including the date and time of the incident, the location (borough, ZIP code, street, and intersection), the number of people involved (pedestrians, cyclists, motorists), and whether any injuries or fatalities occurred. Additionally, it includes contributing factors for the collisions, such as speeding, Alcohol consumption, failure to yield, and impaired driving, along with the types of vehicles involved. The dataset also classifies collisions by whether they occurred during the day or night and on which day of the week.

**Where did the data come from?**

The data was sourced from the Kaggle Open Database, which hosts a variety of datasets provided by organizations, governments, and researchers. This specific dataset on New York City motor vehicle collisions (2015-2017) is publicly available for analysis and has been widely used for traffic safety research.

Website link: <https://www.kaggle.com/datasets/nypd/vehicle-collisions>

**What is the size and structure of the data?**

The dataset consists of approximately [number of rows] rows and [number of columns] columns. Each row represents an individual collision, and the columns include both numerical and categorical variables. Key columns include `CRASH DATE`, `CRASH TIME`, `BOROUGH`, `ZIP CODE`, `NUMBER OF PERSONS INJURED`, `NUMBER OF PERSONS KILLED`, and various contributing factors for the collision. Additional columns such as `DAY\_NIGHT` and `CRASH\_DAY` were derived to enhance the analysis. Despite being comprehensive, the dataset does not include weather data, which could have been a useful factor in understanding environmental impacts on collision rates.

Total number of rows: 477,732 and Columns: 31

**Methodology**

**What steps did you take to download, clean, import, and otherwise transform the data?**

The dataset was first downloaded from the Kaggle Open Database as a CSV file. After obtaining the file, the following steps were taken:

**Data Cleaning:** The dataset was inspected for missing or invalid values. For instance, entries with missing location data (such as latitude, longitude, or ZIP code) or extreme outliers were reviewed. Unnecessary columns (such as irrelevant attributes or ones with too many missing values) were removed to streamline the analysis.

**Note:** We missed some good attributes like Weather and Day of the week for the dataset, Hence we decided to add rows i.e. “day\_night” and “crash\_day” to have some data and get valuable insights for the research we planned to do and answer the questions we came up with.

**Queries for the new columns:**

Adding the attributes:

**ALTER** **TABLE** mg\_collisions2

**ADD** **COLUMN** ***day\_night*** **VARCHAR**(10),

**ADD** **COLUMN** ***crash\_day*** **VARCHAR**(10);

Generating values for the Day and Night:

**UPDATE** mg\_collisions2

**SET** ***day\_night*** = **CASE**

**WHEN** **EXTRACT**(**HOUR** **FROM** **"time"**) **BETWEEN** 6 **AND** 18 **THEN** **'Day'**

**ELSE** **'Night'**

**END**;

Generating values for the Day of the week (Crash day):

**UPDATE** mg\_collisions2

**SET** ***crash\_day*** = **TO\_CHAR**("date", **'Day'**);

**Note:** These values were derived based on the existing data from Date and Time.

**Data Transformation:** To facilitate more detailed analysis, additional columns were created:

A DAY\_NIGHT column was generated based on CRASH TIME to classify whether the collision occurred during the day or night.

A CRASH\_DAY column was derived from CRASH DATE to categorize collisions by the day of the week.

**Data Formatting:** The date and time columns were converted to a uniform format (YYYY-MM-DD for CRASH DATE and HH:MM for CRASH TIME). Any text columns were normalized to ensure consistency (e.g., borough names were standardized to highways/freeways where it was empty).

**What DBMS(s) did you use?**

The dataset was imported into DBeaver, a multi-platform database management tool, where SQL queries were used for data analysis and exploration. DBeaver was connected to a PostgreSQL database for efficient querying and large-scale data manipulation.

Initially, the data was explored in MongoDB, but later transitioned to SQL-based analysis in DBeaver due to the preference for structured queries for this project and better visualizations.

**What other tools did you use?**

* **Python (pandas, matplotlib):** Python was used for preliminary data exploration, cleaning, and visualization. The pandas library helped manipulate the dataset, and matplotlib was used to generate charts and graphs for visual analysis.
* **Excel:** Basic data inspections and quick checks were done in Excel before the data was imported into DBeaver.
* **Jupyter Notebooks:** Python code was developed in Jupyter Notebooks for data cleaning and visualization purposes, allowing for an interactive development process.
* **Visualization Tools:** Charts and graphs generated from Power BI were used to illustrate findings in the final report.

**Queries:**

**Q1) Which boroughs in New York City experience the highest frequency of motor vehicle collisions, and what are the potential contributing factors in these areas? Based on the boroughs what were the top worst incidents and what are their main contributing factors?**

Before getting started with the case analysis and diving into the data, we found a lot of NULL values in the “borough” attribute. When we started to find the reason and decided to clean the data, we realized that it was for all the incidents that took place on a Highway/freeway which are usually between these places and hence were not categorized.

Once we understood the value of the Null data, we updated the values with the value:

“Highways and Freeways”. Below is the update command used to do the same.

**UPDATE** mg\_collisions2

**SET** borough = **'Highways and Freeways'**

**WHERE** borough **IS** **NULL**;

**Query 1: (Basic analysis)**

**Finding the numbers of incidents in each borough in New York**

**SELECT borough,**

**COUNT(\*) AS *total\_accidents*,**

**SUM(personsinjured) AS *total\_injured*,**

**SUM(personskilled) AS *total\_deaths***

**FROM mg\_collisions2**

**WHERE borough IS NOT NULL**

**GROUP BY borough**

**ORDER BY *total\_accidents* DESC;**

**A screenshot of a graph

Description automatically generated**

*This query groups the data by borough and calculates the total number of accidents, total number of people injured, and total number of deaths for each borough in New York. By grouping the results by borough and summing up the incidents, we get a clear view of how different areas are impacted by motor vehicle collisions. The query also orders the results by the total number of accidents in descending order, so the borough with the highest accident count is listed first.*

**A screenshot of a computer

Description automatically generatedQuery 2: Finding the number of incidents on Each day of the week in specific boroughs**

**SELECT *crash\_day*, COUNT(\*) AS *total\_accidents***

**FROM mg\_collisions2**

**GROUP BY crash\_day**

**ORDER BY *total\_accidents* DESC;**

**A screenshot of a computer

Description automatically generated**

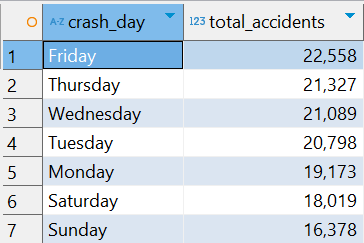
**SELECT *crash\_day*, COUNT(\*) AS *total\_accidents***

**FROM mg\_collisions2**

**WHERE borough = 'STATEN ISLAND'**

**GROUP BY crash\_day**

**ORDER BY *total\_accidents* DESC;**

****

**SELECT *crash\_day*, COUNT(\*) AS *total\_accidents***

**FROM mg\_collisions2**

**WHERE borough = 'Highways and Freeways'**

**GROUP BY crash\_day**

**ORDER BY *total\_accidents* DESC;**

*These queries provide the total number of accidents for each day of the week by grouping the data based on the day of the crash (crash\_day). The COUNT(\*) function is used to count the number of incidents for each day, and the results are ordered in descending order based on the number of accidents.*

**Query 3: Top 2 Leading Causes of Accidents per Borough, Including Deaths and Injuries**

**WITH *RankedCauses* AS (**

**SELECT borough,**

**"vehicle1factor" AS *contributing\_factor*,**

**COUNT(\*) AS *total\_incidents*,**

**SUM(personsinjured) AS *total\_injured*,**

**SUM(personskilled) AS *total\_deaths*,**

**ROW\_NUMBER() OVER (PARTITION BY borough ORDER BY COUNT(\*) DESC) AS *rank***

**FROM mg\_collisions2**

**WHERE borough IS NOT NULL**

**AND "vehicle1factor" IS NOT NULL**

**AND "vehicle1factor" NOT IN ('Unspecified', '')**

**GROUP BY borough, "vehicle1factor"**

**)**

**SELECT borough,**

***contributing\_factor*,**

***total\_incidents*,**

***total\_injured*,**

***total\_deaths***

**FROM *RankedCauses***

**WHERE *rank* <= 2**

**ORDER BY borough, *total\_incidents* DESC;**

**A screenshot of a data

Description automatically generated**

*This query identifies the top two leading contributing factors for motor vehicle accidents in each borough, based on the number of incidents. It also provides the associated total injuries and deaths for each factor. The ROW\_NUMBER() function is used to rank the contributing factors per borough, and the WHERE rank <= 2 filter ensures that only the top two factors are displayed for each borough.*

**Query 4: (Complex analysis)**

**Based on the boroughs what were the top 10 worst incidents and what are their main contributing factors?**

**SELECT borough,**

**date,**

***crash\_day*,**

***day\_night*,**

**"vehicle1factor" AS *main\_contributing\_factor*,**

**personsinjured + personskilled AS *total\_severity*,**

**personsinjured,**

**personskilled**

**FROM mg\_collisions2**

**WHERE borough IS NOT NULL**

**AND "vehicle1factor" NOT IN ('Unspecified', '')**

**ORDER BY *total\_severity* DESC**

**LIMIT 10;**

**A screenshot of a computer

Description automatically generated**

*This query identifies the top 10 worst incidents across different boroughs in New York based on total severity, which is calculated as the sum of injuries (personsinjured) and deaths (personskilled). It also shows the main contributing factor for each incident, along with details like the date, day of the week (crash\_day), and whether the incident occurred during the day or night.*

**Q2) How does the time of day (Day vs. Night) impact the frequency and severity of collisions, and What was trend over the years?**

**Query 5: Analysing the trend over the years from 2015 to 2017.**

**SELECT EXTRACT(YEAR FROM date) AS year,**

**COUNT(\*) AS total\_accidents,**

**SUM(personsinjured) AS total\_injured,**

**SUM(personskilled) AS total\_deaths**

**A screenshot of a computer

Description automatically generatedFROM mg\_collisions2**

**WHERE date IS NOT NULL**

**GROUP BY year**

**ORDER BY year ASC;**

*This query analyses the yearly trend in traffic accidents from 2015 to 2017, focusing on the total number of accidents, total injuries, and total deaths. By grouping the data by year and summing the number of incidents, injuries, and fatalities, this query helps in understanding how accident frequency and severity have changed over time.*

**Query 6: Find Total Incidents for Day/Night, Most Common Days, Most Common Borough, and Total Deaths and Injuries:**

**WITH *DayNightData* AS (**

**SELECT *day\_night*,**

**COUNT(\*) AS *total\_incidents*,**

**SUM(personsinjured) AS *total\_injured*,**

**SUM(personskilled) AS *total\_deaths***

**FROM mg\_collisions2**

**WHERE *day\_night* IS NOT NULL**

**GROUP BY day\_night**

**),**

***CommonDays* AS (**

**SELECT *crash\_day*,**

**COUNT(\*) AS *day\_count*,**

**ROW\_NUMBER() OVER (ORDER BY COUNT(\*) DESC) AS *rank***

**FROM mg\_collisions2**

**WHERE *crash\_day* IS NOT NULL**

**GROUP BY crash\_day**

**),**

***CommonBorough* AS (**

**SELECT *day\_night*,**

**borough,**

**COUNT(\*) AS *borough\_count*,**

**ROW\_NUMBER() OVER (PARTITION BY *day\_night* ORDER BY COUNT(\*) DESC) AS *rank***

**FROM mg\_collisions2**

**WHERE borough IS NOT NULL**

**GROUP BY day\_night, borough**

**)**

**SELECT *DND*.day\_night,**

***DND*.*total\_incidents*,**

***DND*.*total\_injured*,**

***DND*.*total\_deaths*,**

***CD*.crash\_day AS *most\_common\_day*,**

***CB*.borough AS *most\_common\_borough***

**FROM *DayNightData* *DND***

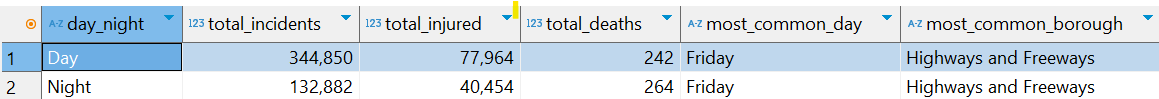
**LEFT JOIN *CommonDays* *CD***

**ON *CD*.*rank* = 1 -- Get the most common day**

**LEFT JOIN *CommonBorough* *CB***

**ON *DND*.day\_night = *CB*.day\_night AND *CB*.*rank* = 1 -- Get the most common borough for each time period**

**ORDER BY *DND*.day\_night;**



*This query breaks down traffic incidents based on whether they occurred during the day or night. It also identifies the most common day of the week for accidents and the borough with the highest number of incidents for both day and night periods.*

**Query 7: Complex analysis of incidents categorized by day and night for each borough and seeing the total injured and total fatalities**

**SELECT borough, *day\_night*,**

**COUNT(\*) AS *total\_incidents*,**

**SUM(personsinjured) AS *total\_persons\_injured*,**

**SUM(personskilled) AS *total\_persons\_killed*,**

**SUM(pedestriansinjured) AS *total\_pedestrians\_injured*,**

**SUM(pedestrianskilled) AS *total\_pedestrians\_killed*,**

**SUM(cyclistsinjured) AS *total\_cyclists\_injured*,**

**SUM(cyclistskilled) AS *total\_cyclists\_killed*,**

**SUM(motoristsinjured) AS *total\_motorists\_injured*,**

**SUM(motoristskilled) AS *total\_motorists\_killed***

**FROM mg\_collisions2**

**WHERE borough IS NOT NULL**

**GROUP BY borough, day\_night**

**ORDER BY borough, *total\_incidents* DESC;**

**A screenshot of a computer

Description automatically generated**

*This query provides a breakdown of traffic incidents in each borough, distinguishing between incidents that occurred during the day and night. The query also includes detailed information about the different types of people involved (pedestrians, cyclists, motorists) and the number of injuries and fatalities for each group.*

**Q3) What are the most common contributing factors (e.g., speeding, failure to yield, impaired driving) in collisions, and how do these factors differ between boroughs or vary with the time of day?**

**Query 8: Find Stats for the Top 10 Most Common Contributing Factors**

**WITH *FactorStats* AS (**

**SELECT "vehicle1factor" AS *contributing\_factor*,**

**COUNT(\*) AS *total\_accidents*,**

**SUM(personsinjured) AS *total\_injured*,**

**SUM(personskilled) AS *total\_deaths***

**FROM mg\_collisions2**

**WHERE "vehicle1factor" IS NOT NULL**

**AND "vehicle1factor" NOT IN ('Unspecified', '')**

**GROUP BY *contributing\_factor***

**)**

**SELECT *contributing\_factor*,**

***total\_accidents*,**

***total\_injured*,**

***total\_deaths***

**FROM *FactorStats***

**ORDER BY *total\_accidents* DESC**

**LIMIT 20;**

**A screenshot of a computer

Description automatically generated**

*This query identifies the top contributing factors to motor vehicle collisions, showing the total number of accidents, total injuries, and total deaths associated with each factor. By excluding unspecified values, the query focuses on the most impactful factors and provides insight into which behaviours or conditions are most commonly linked to traffic accidents.*

**Query 8: Analysing top incidents which had alcohol involvement (Interesting question)**

**SELECT borough,**

**date,**

***crash\_day*,**

***day\_night*,**

**"vehicle1factor" AS *main\_contributing\_factor*,**

**personsinjured + personskilled AS *total\_severity*,**

**personsinjured,**

**personskilled**

**FROM mg\_collisions2**

**WHERE borough IS NOT NULL**

**AND "vehicle1factor" ILIKE '%Alcohol%'**

**ORDER BY *total\_severity* DESC**

**LIMIT 10;**

**A screenshot of a computer

Description automatically generated**

*This query identifies the top 10 most severe incidents where alcohol involvement was a contributing factor. It includes the borough, date, day of the week, and whether the incident occurred during the day or night. The severity of each incident is calculated as the total number of injuries and deaths.*

**Q4)** **How does the type of vehicle involved in a collision (e.g., car, truck, motorcycle) correlate with the severity of the accident, and are certain vehicle types more likely to result in fatalities or serious injuries?**

**Query 9:**

**SELECT "vehicle1type" AS *vehicle\_type*,**

**COUNT(\*) AS *total\_accidents*,**

**SUM(personsinjured) AS *total\_injured*,**

**SUM(personskilled) AS *total\_deaths***

**FROM mg\_collisions2**

**WHERE "vehicle1type" IS NOT NULL**

**GROUP BY *vehicle\_type***

**ORDER BY *total\_accidents* DESC;**

**A screenshot of a report

Description automatically generated**

**Query 10:**

**SELECT "vehicle1type" AS *vehicle\_type*,**

**COUNT(\*) AS *total\_accidents*,**

**SUM(personsinjured) AS *total\_injured*,**

**SUM(personskilled) AS *total\_deaths*,**

**(SUM(personsinjured) + SUM(personskilled))::decimal / COUNT(\*) AS *severity\_rate*,**

**(SUM(personskilled)::decimal / (SUM(personsinjured) + SUM(personskilled))) \* 100 AS *fatality\_percentage***

**FROM mg\_collisions2**

**WHERE "vehicle1type" IS NOT NULL**

**GROUP BY *vehicle\_type***

**ORDER BY *severity\_rate* DESC;**

**A screenshot of a data

Description automatically generated**

*This query analyzes vehicle types based on their severity rate (injuries + deaths per accident) and fatality percentage (deaths as a percentage of total injuries and deaths). The result is sorted by severity rate to highlight which vehicle types are involved in the most severe accidents.*

**Interesting Findings in Our Traffic Collision Analysis**

A colorful pie chart with white text

Description automatically generatedA graph with a blue line

Description automatically generatedOur analysis of traffic collisions in New York City revealed that while Highways and Freeways saw over 139,000 incidents, resulting in 37,758 injuries and 188 deaths, Brooklyn had the highest overall accident rate with 104,961 incidents, 28,799 injuries, and 108 deaths. This is driven by Brooklyn's large population, dense traffic, and busy commuter routes like the BQE and Belt Parkway, which experience heavy congestion. The borough’s mix of vehicles, pedestrians, and cyclists further contributes to its high accident count, highlighting the need for targeted safety measures and improved road infrastructure.

A graph of a crash

Description automatically generatedA graph of blue and yellow bars

Description automatically generatedThe sharp decline in traffic incidents from 2016 to 2017 can be attributed to several key initiatives in New York City's traffic safety strategy. In 2014, the Vision Zero plan was introduced, aimed at eliminating traffic deaths and reducing accidents. By 2016, key measures like reducing the citywide speed limit to 25 mph, installing speed cameras near schools, and redesigning dangerous intersections were fully in place, contributing to the drop in incidents by 2017. The NYPD also increased enforcement of traffic violations, particularly targeting distracted driving, speeding, and failure to yield to pedestrians—major contributors to accidents. Additionally, the rise of ridesharing services like Uber and Lyft likely reduced instances of drunk driving, particularly on weekends. Together, these factors likely explain the significant reduction in traffic incidents in 2017.

Nighttime accidents, though fewer (132,882 incidents), are deadlier, with 264 deaths compared to 242 daytime deaths. Reduced visibility, driver fatigue, and impaired driving contribute to this. Interestingly, while nighttime crashes are more severe, daytime incidents are more frequent, with 344,850 accidents, likely due to heavier traffic. Fridays, especially Friday nights in New York, stand out as the most dangerous, driven by the city's vibrant nightlife. Happy hours, crowded events, and increased alcohol consumption lead to more impaired drivers on the road. The combination of tired commuters, recreational drivers, and heavy congestion makes Friday nights particularly risky, highlighting the need for stronger DUI enforcement and safety A graph of a graph

Description automatically generated with medium confidencemeasures.

A screenshot of a computer

Description automatically generatedOur analysis revealed that Driver Inattention/Distraction is the leading cause of collisions, accounting for 69,474 accidents, resulting in 17,545 injuries and 66 deaths. The widespread use of smartphones and in-car technologies likely exacerbates this issue, making distracted driving a persistent risk. Failure to Yield Right-of-Way also caused 21,852 accidents, leading to 8,953 injuries and 34 deaths, highlighting the need for better traffic education and law enforcement to reduce preventable collisions.

Interestingly, while alcohol-related accidents are less frequent, they tend to be among the most severe. For instance, an alcohol-related crash on December 6, 2016, resulted in 11 injuries. Despite New York’s vibrant nightlife and common assumptions about alcohol-related dangers, these incidents don’t lead to as many deaths as one might expect. Although they are serious, the fatality rate in alcohol-related crashes remains relatively low compared to other causes, making them dangerous but less lethal overall.

A screenshot of a computer

Description automatically generatedA graph with purple squares

Description automatically generatedWhen analyzing the types of vehicles involved in accidents, passenger vehicles accounted for the largest share of incidents, with 295,532 accidents. However, what stands out is the significantly higher fatality percentage for motorcycles. Despite having only 2,539 total accidents, motorcycles accounted for 37 deaths, resulting in a 2.48% fatality rate. This makes motorcycles far more dangerous than vehicle types like bicycles or passenger vehicles, which had much lower fatality percentages. The higher risk associated with motorcycles could be attributed to their lack of protective enclosures, the higher speeds at which they are often operated, and the challenges of visibility on the road. This highlights the urgent need for improved rider education, safety gear enforcement, and greater awareness from other road users. The rising use of motorcycles in NYC is likely driven by their ability to navigate congested traffic and ease parking challenges.

In addition, certain vehicle types, like large commercial vehicles (e.g., trucks with six or more tires), although involved in fewer overall accidents (4,860 incidents), had a much higher fatality percentage. With 20 deaths, large trucks show a 3.33% fatality rate, making them particularly dangerous in collisions, especially when interacting with smaller vehicles. This reflects the inherent risks associated with large trucks, such as their size, weight, and longer stopping distances. These findings point to the need for stricter safety regulations and specialized training for drivers of large commercial vehicles to minimize fatal outcomes in collisions.

**Conclusion:**

Our analysis of New York City traffic collisions highlights key factors contributing to accidents and their severity. Driver inattention/distraction leads the causes, exacerbated by smartphones and in-car technologies, while alcohol-related accidents, though less frequent, are notably severe, underscoring the need for stronger DUI enforcement, particularly on weekend nights. Motorcycles have a disproportionately high fatality rate despite fewer accidents, driven by the inherent risks of riding, while large commercial vehicles pose significant dangers in collisions with smaller vehicles, making stricter regulations and driver training essential. These findings point to the need for targeted safety interventions, better traffic management, and public awareness to reduce both the frequency and severity of accidents, ultimately making NYC’s roads safer for all.

**By Team B**

**Chinni Raj Shubhang Reddy - 2307764**

**Taha Mandviwala - 2306538**

**Lohitha Kolli - 2319201**

**Nitish Kumar Boovaraghan - 2310427**

**Tawfeeq Hasan - 2262071**