**6356 Case Project 2**

**(Fall 2024)**

**Analysing Urban Traffic Collisions (New York): Insights into Safety Trends, Vulnerable Road Users, and Contributing Factors**

A bicycle and food on the ground

Description automatically generated**Introduction:**

Traffic collisions are a major concern in New York City, especially given the city’s dense population, diverse transportation methods, and the daily interactions of vehicles, cyclists, and pedestrians. This project focuses on analyzing key aspects of these incidents, such as their impact on cyclists, the monthly trends in collision frequencies and severities, and the risks associated with taxi-involved collisions. Understanding these patterns is crucial to addressing specific safety challenges faced by vulnerable groups like cyclists and identifying seasonal and situational factors that influence high-severity accidents. For instance, in the first half of 2024, New York City witnessed the highest number of traffic fatalities since the Vision Zero program’s inception nearly a decade ago, with 61 pedestrian deaths, marking a 27% increase from the same period last year.

Additionally, by exploring the prevalence of unspecified contributing factors, this analysis highlights gaps in reporting that could hinder effective intervention. The ultimate goal is to provide actionable insights to improve road safety, protect lives, and create a more efficient and safer urban environment in New York City.

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) How do collision frequencies and severities involving cyclists vary across different boroughs, and what are the most common contributing factors in these incidents?**

**Q2) How do traffic collisions in New York City vary by month in terms of frequency, severity, injuries, and fatalities, and which months show the highest proportion of high-severity accidents?**

**Q3) How do taxi-involved collisions impact traffic safety in New York City, and what are the trends in their frequency, severity, and contributing factors across different boroughs, times of day, and months?**

**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`, `CRASH\_DAY`, ‘SEVERITY\_CATEGORY’ and ‘INCIDENT\_MONTH’ 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'**);

Adding the Month attribute:

ALTER TABLE mg\_collisions2

ADD COLUMN incident\_month VARCHAR(15);

Generating values for the Month of the incident:

UPDATE mg\_collisions2

SET incident\_month = TO\_CHAR(date, 'Month');

Add the severity\_category Column:

ALTER TABLE mg\_collisions2

ADD COLUMN severity\_category VARCHAR(10);

Generating values for the severity\_category:

UPDATE mg\_collisions2

SET severity\_category = CASE

WHEN (personsinjured + personskilled) <= 2 THEN 'Low'

WHEN (personsinjured + personskilled) BETWEEN 3 AND 5 THEN 'Moderate'

WHEN (personsinjured + personskilled) > 5 THEN 'High'

ELSE 'Unknown' -- For NULL or missing data

END

WHERE personsinjured IS NOT NULL AND personskilled IS NOT NULL;

**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.

A INCIDENT\_MONTH column was derived from CRASH DATE to categorize collisions by month of the year.

A SEVERITY\_CATEGORY column was derived from the total\_injured and total\_deaths column to help us analyze and divide these incidents into 3 categories i.e. LOW, MODERATE AND HIGH.

**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. Also used Neo4j to create a graph analysis to explore new findings.

**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) How do collision frequencies and severities involving cyclists vary across different boroughs, and what are the most common contributing factors in these incidents?**

**Query 1: Analysis of Cyclist Collisions by Borough**

**SELECT borough,**

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

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

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

**FROM mg\_collisions2**

**WHERE (cyclistsinjured > 0 OR cyclistskilled > 0)**

**AND borough IS NOT NULL**

**GROUP BY borough**

**ORDER BY total\_accidents DESC;**

**A screenshot of a computer

Description automatically generated**

*This identifies the number of cyclist-related accidents, injuries, and fatalities in each borough, highlighting the areas most affected by cyclist collisions. Results are sorted by total accidents in descending order.*

**Query 2: Vehicle Types Most Involved in Cyclist Collisions**

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

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

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

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

**FROM mg\_collisions2**

**WHERE (cyclistsinjured > 0 OR cyclistskilled > 0)**

**AND "vehicle1type" IS NOT NULL**

**GROUP BY vehicle\_type**

**ORDER BY total\_accidents DESC;**

**A screenshot of a computer

Description automatically generated**

*This identifies the types of vehicles most frequently involved in cyclist accidents. It calculates the total accidents, injuries, and fatalities associated with each vehicle type and orders the results by the highest number of accidents.*

**Query 3: Cyclist Incidents by Day of the Week with Severity and Fatality Analysis**

**SELECT day\_night,**

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

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

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

**(SUM(cyclistsinjured) + SUM(cyclistskilled))::decimal / COUNT(\*) AS severity\_rate,**

**(SUM(cyclistskilled)::decimal / (SUM(cyclistsinjured) + SUM(cyclistskilled))) \* 100 AS fatality\_percentage**

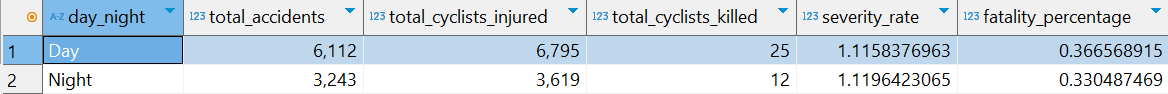
**FROM mg\_collisions2**

**WHERE (cyclistsinjured > 0 OR cyclistskilled > 0)**

**AND day\_night IS NOT NULL**

**GROUP BY day\_night**

**ORDER BY total\_accidents DESC;**

****

**Query 4: Analysis by the day of the week**

**SELECT crash\_day AS day\_of\_week,**

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

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

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

**(SUM(cyclistsinjured) + SUM(cyclistskilled))::decimal / COUNT(\*) AS severity\_rate,**

**(SUM(cyclistskilled)::decimal / (SUM(cyclistsinjured) + SUM(cyclistskilled))) \* 100 AS fatality\_percentage**

**FROM mg\_collisions2**

**WHERE (cyclistsinjured > 0 OR cyclistskilled > 0)**

**AND crash\_day IS NOT NULL**

**GROUP BY day\_of\_week**

**ORDER BY total\_accidents DESC;**

**A screenshot of a graph

Description automatically generated**

*This query analyzes cyclist-related accidents by day of the week and their severity. It calculates the total accidents, injuries, fatalities, severity rate, and fatality percentage for each day. The results highlight which days see the most cyclist incidents and the severity of those incidents.*

**Q2) How do traffic collisions in New York City vary by month in terms of frequency, severity, injuries, and fatalities, and which months show the highest proportion of high-severity accidents?**

**Query 5:**

**SELECT TO\_CHAR(date, 'Month') AS month,**

**EXTRACT(MONTH FROM date) AS month\_number,**

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

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

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

**SUM(CASE WHEN severity\_category = 'Low' THEN 1 ELSE 0 END) AS low\_severity,**

**SUM(CASE WHEN severity\_category = 'Moderate' THEN 1 ELSE 0 END) AS moderate\_severity,**

**SUM(CASE WHEN severity\_category = 'High' THEN 1 ELSE 0 END) AS high\_severity**

**FROM mg\_collisions2**

**GROUP BY month, month\_number**

**ORDER BY month\_number;**

**A screenshot of a computer

Description automatically generated***This query examines traffic collisions in New York City by month, analyzing their frequency, injuries, fatalities, and severity levels (low, moderate, high). It identifies patterns in collisions across the year, highlighting months with the highest number and severity of accidents. Results are sorted by month.*

**Query 6: Contributing Factors for High-Severity Accidents in August**

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

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

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

**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 EXTRACT(MONTH FROM date) = 8**

**AND severity\_category = 'High'**

**AND "vehicle1factor" IS NOT NULL**

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

**GROUP BY contributing\_factor**

**ORDER BY severity\_rate DESC;**

**A screenshot of a computer

Description automatically generated**

*This query identifies the contributing factors for high-severity accidents in August, calculating the total accidents, injuries, fatalities, severity rate, and fatality percentage for each factor. It helps pinpoint key causes of severe incidents during this month.*

**Query: Day/Night Analysis for January and August**

**SELECT EXTRACT(MONTH FROM date) AS month,**

**day\_night,**

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

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

**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**

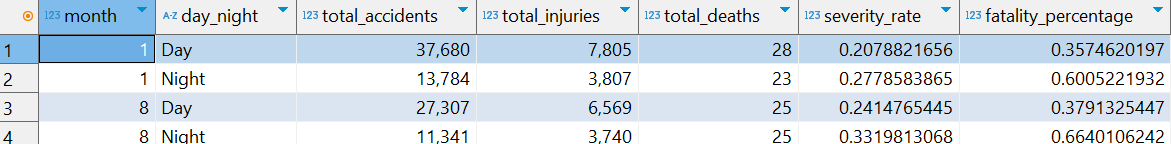
**FROM mg\_collisions2**

**WHERE EXTRACT(MONTH FROM date) IN (1, 8)**

**AND day\_night IS NOT NULL**

**GROUP BY month, day\_night**

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

****

This query compares traffic accidents during the day and night in January and August. It calculates total accidents, injuries, fatalities, severity rate, and fatality percentage for each time period, highlighting the impact of time on collision patterns in these months.

**Q3) How do taxi-involved collisions impact traffic safety in New York City, and what are the trends in their frequency, severity, and contributing factors across different boroughs, times of day, and months?**

**SELECT TO\_CHAR(date, 'Month') AS month,**

**EXTRACT(MONTH FROM date) AS month\_number,**

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

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

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

**FROM mg\_collisions2**

**WHERE "vehicle1type" ILIKE '%Taxi%'**

**GROUP BY month, month\_number**

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

**A screenshot of a data

Description automatically generated**

This query analyzes taxi-involved collisions by month, calculating the total number of accidents, injuries, and fatalities. It identifies months with the highest frequency of taxi-related incidents, helping to understand seasonal trends in taxi safety.

**Query: To find reasons behind the unspecified factor which is majority of the dataset.**

**SELECT**

**borough,**

**day\_night,**

**crash\_day AS day\_of\_week,**

**severity\_category,**

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

**COUNT(CASE WHEN "vehicle1factor" = 'UNSPECIFIED' THEN 1 END) AS unspecified\_accidents,**

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

**ROUND((COUNT(CASE WHEN "vehicle1factor" = 'UNSPECIFIED' THEN 1 END)::decimal / COUNT(\*)) \* 100, 2) AS unspecified\_percentage**

**FROM mg\_collisions2**

**WHERE borough IS NOT NULL**

**AND day\_night IS NOT NULL**

**AND crash\_day IS NOT NULL**

**AND severity\_category IS NOT NULL**

**AND "vehicle1type" IS NOT NULL**

**GROUP BY borough, day\_night, crash\_day, severity\_category, "vehicle1type"**

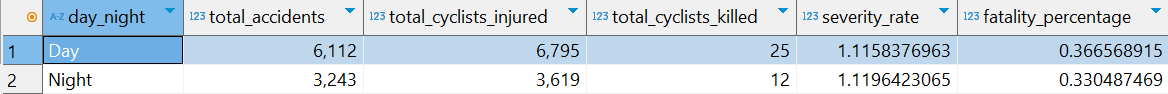
**ORDER BY unspecified\_percentage DESC, unspecified\_accidents DESC;**

**A screenshot of a computer

Description automatically generated**

This query examines traffic collisions with unspecified contributing factors, grouped by borough, time of day, day of the week, severity category, and vehicle type. It calculates the number and percentage of accidents with unspecified factors, highlighting gaps in data reporting and areas requiring better data accuracy.

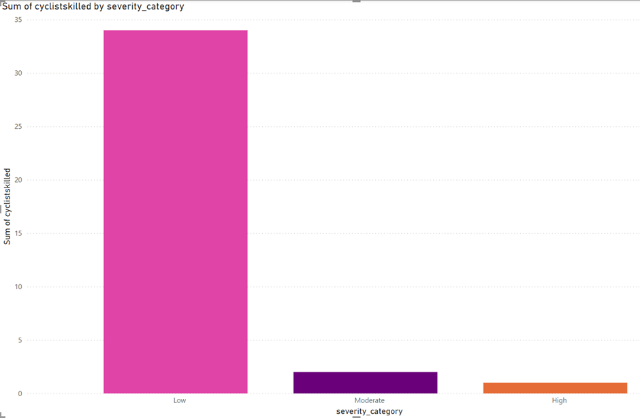
**Interesting Findings in Our Traffic Collision Analysis**

****A graph of a number of blue rectangles

Description automatically generatedA screenshot of a computer

Description automatically generatedNew York City’s traffic collision data reveals significant insights into safety challenges and areas for improvement. Cyclist-related incidents are particularly alarming, with Brooklyn reporting the highest number of such collisions—3,025 incidents leading to 3,313 injuries and 11 fatalities. The borough's dense population and widespread cycling infrastructure contribute to these high numbers.

A graph with a line going up

Description automatically generatedPassenger vehicles were most commonly involved in cyclist-related collisions, with 4,979 cases citywide. Nighttime collisions, though less frequent than daytime incidents, had a higher severity rate of 1.12, likely due to reduced visibility and riskier behaviors. Wednesdays stood out as the day with the most cyclist-related collisions, indicating potential patterns tied to midweek commuting. However, a substantial portion of these incidents had unspecified contributing factors, underscoring the need for more detailed data collection to better understand and mitigate risks.

Each line here represents the month of the year i.e. Jan, Feb…, Dec.

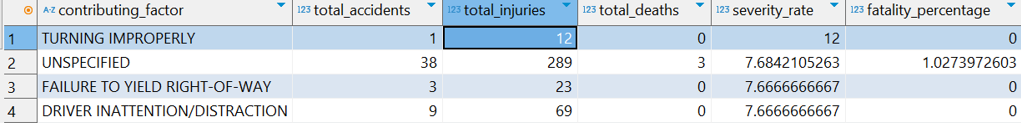
Monthly variations in collisions provide further insights. January recorded the highest number of incidents at 51,464, largely attributed to icy roads and adverse winter weather conditions. These factors reduce vehicle control and increase risks for all road users. Meanwhile, August showed the highest number of high-severity accidents, with 62 such cases. This may be due to heightened summer road activity, including tourism, outdoor events, and increased pedestrian and cyclist presence. The intense summer heat might also contribute to driver fatigue or aggression, adding to the risks. The one incident which caused more than 12 people getting injured spiked the severity of indidents overall in the month of August. These findings emphasize A graph showing a red line

Description automatically generated**A screenshot of a computer

Description automatically generated**the importance of tailored seasonal safety measures, such as enhanced road maintenance in winter and traffic management strategies during the summer months.

The reason behind the spike in the month of August:

A pie chart with different colors

Description automatically generated

Taxi-involved collisions revealed critical patterns, particularly in Manhattan, where taxis play a central role in urban transportation. January saw the highest number of taxi-related incidents with 2,188, corresponding to the post-holiday rush and heavy winter traffic. While December had fewer incidents, it recorded the highest fatalities, with four deaths. Most taxi-related collisions occurred during the day, likely due to higher daytime traffic and taxi activity. Despite this, the severity rates were consistent across day and night, suggesting that other factors, such as traffic density or road conditions, contribute equally. The frequent absence of detailed contributing factors in taxi-related collisions highlights a gap in data reporting that could limit the effectiveness of safety interventions.

A screenshot of a computer

Description automatically generatedA graph of a number of people

Description automatically generatedA graph of blue and orange bars

Description automatically generatedThe issue of unspecified contributing factors is a recurring theme in the dataset. These factors accounted for 235,726 incidents, resulting in 54,717 injuries and 224 fatalities. The lack of detailed information makes it challenging to identify and address specific causes. Unspecified factors were particularly prevalent in boroughs like Brooklyn and Manhattan, where dense traffic and complex intersections may complicate precise reporting. This points to the need for enhanced data collection practices, such as mandatory reporting of contributing factors and the use of technology like traffic cameras to improve accuracy. By reducing the reliance on unspecified categories, safety measures can become more targeted and effective, addressing the root causes of collisions more directly.

**A screenshot of a computer

Description automatically generated**

**Conclusion:**

Our analysis of New York City traffic collisions underscores critical areas for intervention to improve safety and reduce fatalities. Cyclist-related incidents remain a significant concern, particularly in Brooklyn, where dense traffic and extensive cycling infrastructure contribute to high collision rates. The prevalence of unspecified contributing factors limits our understanding of the causes, highlighting the need for more accurate data collection. Seasonal trends reveal heightened collision risks in winter months like January, likely due to icy conditions, while summer months like August show an increase in high-severity accidents, driven by tourism, outdoor activities, and heat-related fatigue. Taxi-involved collisions, particularly concentrated in Manhattan, demonstrate unique risks tied to the city's heavy reliance on taxis for daily commutes, with fatalities peaking in unpredictable ways, such as in December. Addressing these issues requires a multi-faceted approach, including enhanced reporting practices, seasonally tailored safety measures, stricter enforcement of traffic laws, and public awareness campaigns. These findings underscore the importance of data-driven strategies to ensure safer streets for all New Yorkers.

**By Team B**

**Chinni Raj Shubhang Reddy - 2307764**

**Taha Mandviwala - 2306538**

**Lohitha Kolli - 2319201**

**Nitish Kumar Boovaraghan - 2310427**

**Tawfeeq Hasan - 2262071**