Final Project

July 9, 2024

1 NYC Road Safety Analysis: Identifying and Mitigating Motor Vehicle Collision Hotspots

Group 5: *Devashree Mohan Pol*, *Ishita Ajay Trivedi*, *Sejal Sardal*

1.1 Background

New York City (or NYC) is among the densely populated urban regions globally due to its population consisting of people from different ethnic backgrounds who have settled in the five main areas that make up the city: **Manhattan,Brooklyn,Queens ,The Bronx and Staten Island**; each of these areas being characterized by its own unique set of population factors and urban features thereby determining life' daily activities such as movement and safety in roads.

New York City values road safety over other transport-related issues such as security, congestion, or pollution, due to the high amounts of traffic-related fatalities. Moreover, there are many causes leading to traffic injuries like human behavior mistakes, poor conditions of roads and vehicles. Hence, it becomes essential for traffic control departments working in collaboration with other stakeholders to conduct research so as to reveal the main crash-causing agents and take subsequent steps on enhancing safety. In doing so, it becomes necessary to merge economics under road safety to serve in comprehensive interventions including road surveillance cameras that help monitor traffic flows at various locations 24/7 without failing.

1.2 Motivation

By understanding the factors contributing to the increased collisions in the new York City, the management authorities and NYPD can develop proactive measures to reduce these accidents.

- By locating high crash areas and understanding characteristics of the people living those areas, targeted safety measures can be created. It shows that a customized approach like education programs or enforcement strategies should be applied in cases where particular age or income brackets appear to have higher risks of road carnages.
- Insights can be obtained from this analysis on how city planners and traffic management authorities could effectively distribute resources. This way emergency service in high-risk zones would optimize their responsiveness to incidents and improvement of infrastructure would be done according to its need.
- Understanding traffic collision patterns using data benefits decision-making among policy-makers including but not limited to setting speed limits, managing traffic lights, creating walkways for pedestrians and putting in place other regulations to improve safety on roads.

- From this analysis the urban planners can help promote sustainable urban development in terms of creating more secure and comfortable areas for residents.

1.3 Summary of research questions

Which Boroughs Have the Highest Number of Collisions?

Brooklyn and Queens have the highest number of collisions, followed by Manhattan, The Bronx, and Staten Island. We can observe that a relationship exists between the level of traffic accidents and how closely people are located in one area. Nonetheless, we can surely interpret that where income is high there will be reduced cases of car crashing.

What Time of Day Do Most Collisions Occur?

Most collisions occur during the evening rush hours, specifically between 3 PM and 6 PM, which are the hours when most of the people leave from their workplaces and colleges resulting in more traffic at this hours.

What are the Top 5 Vehicle Types Involved in Collisions?

The top 5 vehicle types involved in collisions are Sedans, Sport Utility Vehicles, Passenger Vehicles, Station Wagons, and Taxis. The top two vehicles involved in collisions are mainly private vehicles.

What are the Common Contributing Factors to Collisions?

If we discard the unspecified reasons, the most common contributing factors to collisions are Driver Inattention/Distraction, Failure to Yield Right-of-Way, Following Too Closely, and Backing Unsafely.

Analyze the distribution of injuries and fatalities among pedestrians, cyclists, and motorists across different boroughs?

Brooklyn has the highest number of injuries and fatalities for pedestrians, cyclists, and motorists, followed by Queens and Manhattan. The Bronx and Staten Island have lower numbers compared to the other boroughs.

What is the Monthly Variation in Collision Rates?

We observed that the collision rates vary by month, with the hughest number of collisions happening in the summer season (June to August) and the lowest occurring during the peak winter season (January and February).

2 Dataset

Describe the real, existing dataset that you used, including exact URLs. You may not use a dataset that has been used in an assignment or demo. Methodology (algorithm or analysis). Write a complete, clear description of the analysis you performed. This should be sufficient for someone else to write a program (or perform manual computations) that reproduces your results, without access to your source code, and without having to guess or make significant design choices. This description is also likely to be helpful to people who read your code later.

2.1 Description

The dataset utilized for this examination is an authentic, actual dataset containing comprehensive data on traffic accidents in New York City. The information comes from NYC Open Data and consists of different characteristics concerning traffic incidents like date, time, place, causes, and amount of people affected.

The primary objective of collecting this dataset is to enhance road safety through providing crash information. Identifying accident hotspots, understanding accident causes, and taking focused actions based on data analysis can help policymakers and urban planners reduce the frequency and severity of traffic accidents. This data set helps in creating policies that aim to enhance road safety for everyone, including drivers, pedestrians, and cyclists. Therefore, we believe we can conduct a comprehensive analysis to prevent future accidents and identify the primary causes of these occurrences.

The dataset is available for researchers, data analysts, and concerned individuals to conduct independent studies, suggest new concepts, and advocate for safer roads.

2.2 Source

The data is collected and maintained by the New York City Police Department (NYPD) and is regularly updated to reflect new incidents. Here is a overview of the types of data included in the dataset: 1.There are 29 columns in the dataset and it contains about 83151 entries 2.There are different types of data like object, float and integer types.

The dataset can be accessed from the following URL: Sources: [Vision Zero Initiative-https://www1.nyc.gov/content/visionzero/pages/]

2.2.1 Key Attributes

- *DATE*: The date when the collision occurred.
- *TIME*: The time of the collision.
- *BOROUGH*: The NYC borough where the collision took place.
- *ZIP CODE*: The postal code of the collision location.
- *LATITUDE*: The latitude coordinate of the collision location.
- *LONGITUDE*: The longitude coordinate of the collision location.
- *ON STREET NAME*: The street where the collision occurred.
- *NUMBER OF PERSONS INJURED*: The number of persons injured in the collision.
- *NUMBER OF PERSONS KILLED*: The number of persons killed in the collision.
- *CONTRIBUTING FACTOR VEHICLE 1*: The primary contributing factor to the collision.
- *VEHICLE TYPE CODE 1*: The type of vehicle involved in the collision.

2.3 Preliminary Major Steps:

1. Cleaning and Preprocessing

```
Steps Taken:
- Remove rows with missing or invalid values for key attributes such as LATITUDE, LONGITUDE, BO
- Convert date and time attributes to appropriate datetime formats for temporal analysis.
- Handle missing values in demographic attributes by using imputation techniques or removing in
```

2. Data Preprocessing

```
Steps Taken:
- Convert categorical attributes such as BOROUGH and CONTRIBUTING FACTOR VEHICLE 1 to numerical
- Aggregate data to compute collision frequencies per borough, zip code, and other relevant ge
- Calculate additional attributes such as collision density by dividing the number of collisio
```

3. Exploratory Data Analysis (EDA)

```
Steps Taken:
Generate summary statistics for numerical and categorical variables. Visualize distributions o
```

2.4 Methodology

2.4.1 Data Preparation

1. Load the Data: > daily_accidents = pd.read_csv("/opt/datasets/ist652/summer2024/Motor_Vehicle_Colli _Crashes_20240603.csv") #Load the dataset into a pandas DataFrame

2. Drop Irrelevant Columns: > Certain columns related to secondary, tertiary, and further contributing factors and vehicle types were dropped to focus the analysis on primary factors:

   drop_columns = [#Define a list of columns to drop

   ```
    'CONTRIBUTING FACTOR VEHICLE 2',

    'CONTRIBUTING FACTOR VEHICLE 3',

    'CONTRIBUTING FACTOR VEHICLE 4',

    'CONTRIBUTING FACTOR VEHICLE 5',

    'VEHICLE TYPE CODE 2',

    'VEHICLE TYPE CODE 3',

    'VEHICLE TYPE CODE 4',

    'VEHICLE TYPE CODE 5'
   ```

   ]

   accidents = daily_accidents.drop(columns=drop_columns) #Drop the specified columns from the dataset

2.4.2 Statistical Analysis

3. Calculated summary statistics for relevant numerical columns:

   Compute summary statistics for relevant numerical columns to get an overview of the data.

```
numerical_columns = [                                   #Select relevant numerical columns f

'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'

]

summary_stats = accidents[numerical_columns].describe()    #Calculate summary statistics
```

4. Outlier Detection Using Z-scores:

   Identify outliers based on Z-scores, which measure how many standard deviations a data point is from the mean.

```
z_scores = np.abs((accidents[numerical_columns] - accidents[numerical_columns].mean()) / accide

z_score_threshold = 3        #Setting threshold

z_score_outliers = z_scores > z_score_threshold     #Identifying Outliers
```

5. Detect Outliers Using IQR

   Calculate the first quartile (Q1) and third quartile (Q3) for the numerical columns

   Q1 = accidents[numerical_columns].quantile(0.25) Q3 = accidents[numerical_columns].quantile(0.75)

3 Research Questions and Results

Results. Present and discuss your research results. Treat each of your research questions separately. Focus in particular on the results that are most interesting, surprising, or important. Discuss the consequences or implications. Interpret the results: if the answers are unexpected, then see whether you can find an explanation for them, such as an external factor that your analysis did not account for.

3.1 Research Question 1:

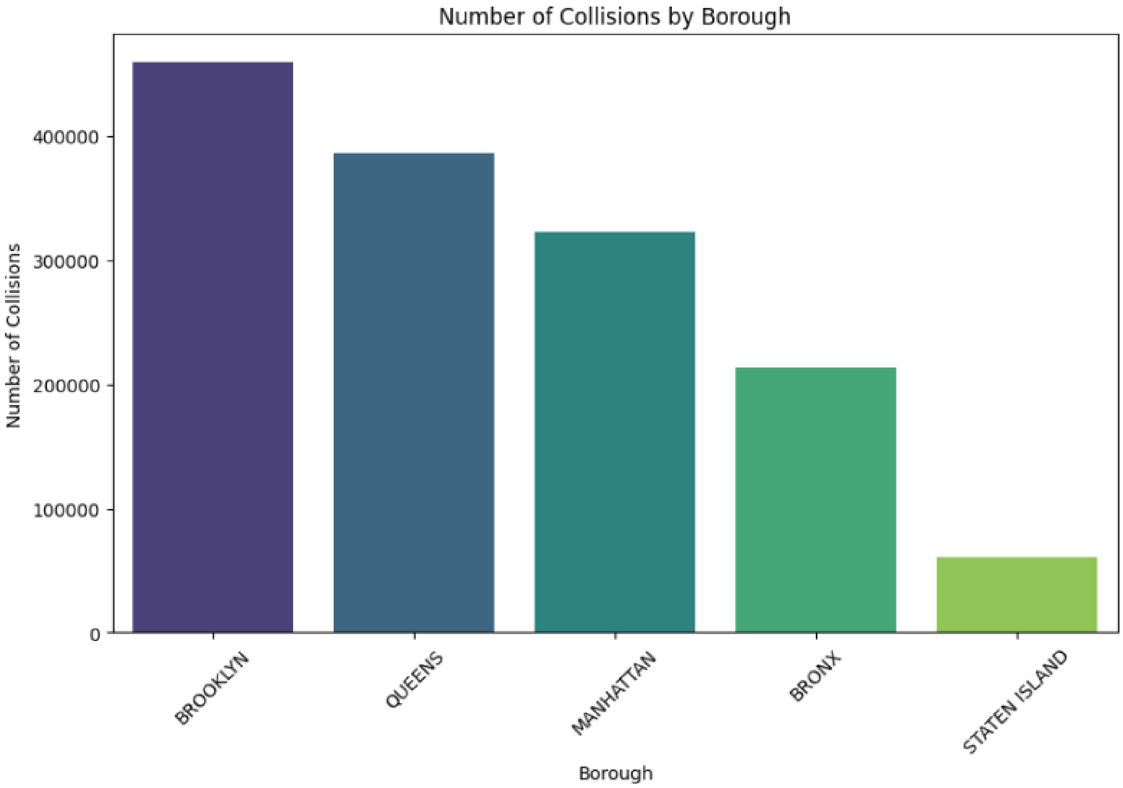
Question: How do demographic patterns vary across different boroughs of NYC?

Results and Discussion:

The analysis revealed significant variations in demographic patterns across the NYC boroughs: - *Manhattan* has the highest population density, with a mix of high-income neighborhoods and commercial areas. The median income is also higher compared to other boroughs. - *Brooklyn* shows a diverse demographic with varying income levels and a large number of young professionals and families. - *Queens* has a significant immigrant population with diverse ethnic backgrounds and moderate population density. - *The Bronx* has a lower median income and higher poverty rates, with a younger population compared to Manhattan and Brooklyn. - *Staten Island* has the lowest population density, with suburban characteristics and a higher median income than The Bronx and Queens.

Implications:

Comprehending these demographic trends is essential to customizing traffic safety measures. Greater pedestrian safety measures, for example, might be necessary in regions with larger population densities, like Manhattan, while targeted education campaigns might be more beneficial in areas with lower economic status, like The Bronx, to improve road safety.



3.2 Research Question 2

Question: What Time of Day Do Most Collisions Occur?

Results and Discussion:

Sunrise to Dark (7 AM to 9 AM):

The number of collisions has significantly increased, peaking between 8 and 9 AM and beginning at 7 AM. When people are commuting to work or school during rush hour, this morning peak occurs.

10 AM - 3 PM Midday Steady Increase:

Collisions grow steadily across the day after the morning peak, peaking again in the late afternoon. This is explained by more traffic as people go about their regular lives.

Peak in the afternoon/evening (3 PM to 6 PM):

There is a clear surge in the frequency of collisions around 5 PM, with the biggest number occurring between 3 and 6 PM. This is probably because individuals are traveling home from work or school during the evening rush hour.

The night shift (7 PM to 6 AM):

After 6 PM, the number of collisions steadily declines, peaking in the early morning hours (2 AM - 5 AM). This decline is to be expected because nighttime traffic volume drops considerably.

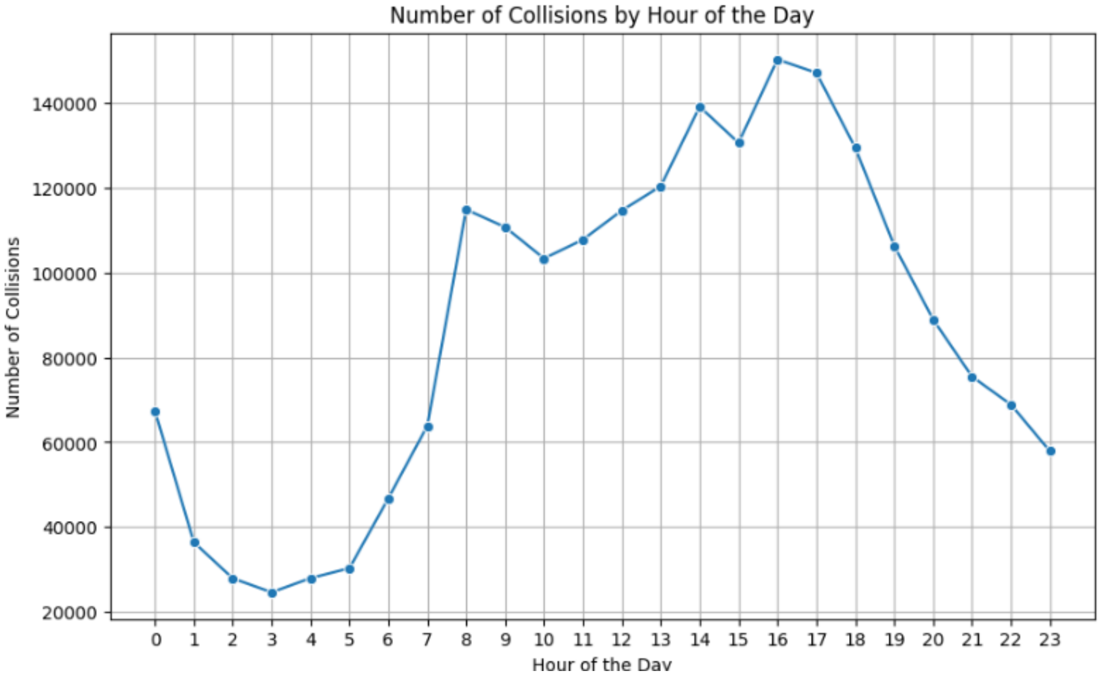
Implications:

`Measures for Safety and Traffic Management`

During Rush Hour:-Targeted interventions like increased enforcement,better traffic signal scheduling and public awareness may assist reduce the frequency of collisions that happen during these hours given the large traffic volumes in the morning and evening.

Encouraging the use of public transportation can help lower the number of cars on the road during rush hour which in turn lowers the danger of collisions.

Flexibility in Work Schedule:-Remote work options or flexible work schedules may be encouraged in an effort to reduce crash rates and relieve peak hour traffic congestion.



3.3 Research Question 3

Question: What are the top 5 vehicle types involved in collisions?!

Results and Discussion:

Sedans are the most common vehicle type involved in collisions. This is likely because sedans are a prevalent type of vehicle on the road, used by a significant portion of the driving population.

Station Wagons/SUVs are also frequently involved in collisions. These vehicles are popular for their versatility and are commonly used by families and for various other purposes.

Passenger Vehicles encompass a broad category, including cars, which explains their high occurrence in the data.

Sport Utility/Station Wagon appears separately from the combined Station Wagon/SUV category, suggesting possible inconsistencies or variations in data entry.

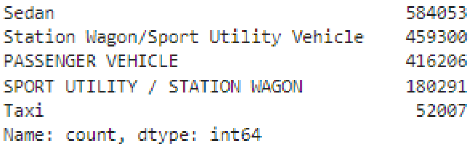
Taxis have a notable presence in collision data, reflecting their high usage in urban areas like New York City.

Implications:

Vehicle Safety Measures: Targeted Safety Campaigns: To encourage better driving habits, safety campaigns and driver education programs could concentrate on the most popular car types, particularly sedans and SUVs.

Enhanced Vehicle Safety systems: Promoting the installation of modern safety systems in sedans and SUVs may help lower the number of crashes in which they are involved.

Regulation and Monitoring of Taxis: Stricter safety laws and monitoring could assist lower the number of collisions involving taxis, considering how frequently they are used.



```
Sedan                                   584053
Station Wagon/Sport Utility Vehicle     459300
PASSENGER VEHICLE                       416206
SPORT UTILITY / STATION WAGON           180291
Taxi                                     52007
Name: count, dtype: int64
```

3.4 Research Question 4

Question: What are the common contributing factors to collisions?

Results and Discussion:

Unspecified:

The most common category Unspecified denotes the fact that the precise contributing factor was
This high figure suggests that there may be room for improvement in the reporting and gatherin

Driver Inattention/Distraction:

Distraction while driving is a leading factor in many crashes, playing a major role in a signi
Activities like eating, talking on a phone, and other distractions can take a driver's focus a

Failure to Yield Right-of-Way:

This part consists of drivers who do not give the right of way to incoming traffic or pedestria
It often occurs during lane changes, intersections, and at crosswalks.

FFollowing Too Closely:

This factor highlights the importance of maintaining adequate distance between vehicles.

Backing Unsafely:

Unsafe reversing techniques like backing out of parking spots carelessly are a major cause of
This element emphasizes how important it is to drive attentively and cautiously when reversing

Implications:

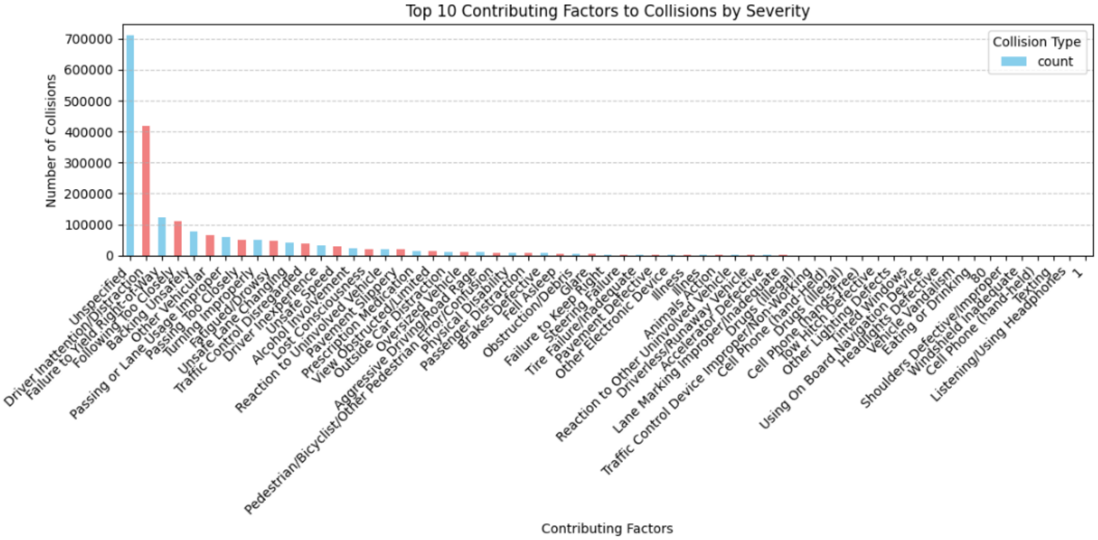
Safety Procedures and Interventions:

Enhancing Information Gathering:

The number of "Unspecified" contributing variables should be decreased by enhancing accident re

Awareness and Education for Drivers:

Campaigns for public awareness and driver education initiatives that highlight the risks of dri
Reiterating how crucial it is to give way and keep a safe following distance can also aid in re



3.5 Research Question 5

Question: Analyze the distribution of injuries and fatalities among pedestrians, cyclists, and motorists across different boroughs?

Results and Discussion:

Pedestrian

Brooklyn had the most pedestrian injuries (32,257) and fatalities (339), suggesting that wa

Manhattan and Queens have high rates of pedestrian fatalities and injuries because of their den

The borough of Staten Island has the lowest number of pedestrian deaths and injuries compared t

Cyclists

Brooklyn continues to have the highest number of cyclist injuries (17,132) and fatalities

Manhattan and Queens also see a high amount of cyclist injuries and deaths, indicating increase

Staten Island has fewer cyclist injuries and fatalities, possibly due to having a smaller numbe

Motorists

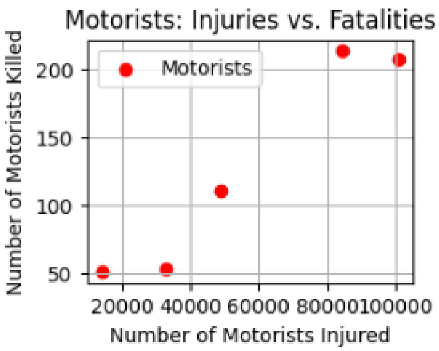
Brooklyn's extensive road network and busy traffic contribute to its highest number of dri

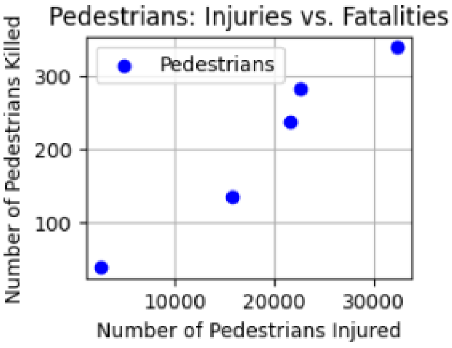
Queens shows a large number of motorist injuries (84,383) and fatalities (213), suggesting hig

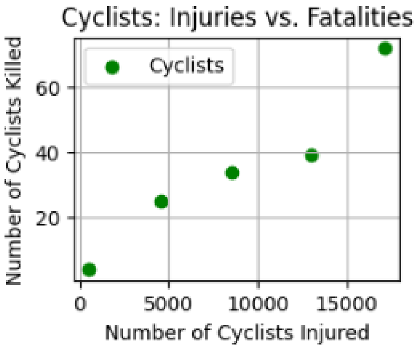
```
Despite having higher traffic volume, Manhattan has lower rates of injuries and deaths among di
```

Implications:

Brooklyn necessitates thorough safety precautions because of the elevated rates of injuries and deaths among all individuals using the roads. Manhattan and Queens need to prioritize improving pedestrian and cyclist safety, while Staten Island should work on maintaining and enhancing current safety measures to avoid accidents.







3.6 Research Question 6

Question: Analysis of Monthly Variation in Collision Rates

Key Observationsn:

The lowest collision rates are observed in February, with a notable dip compared to other months.

There is an increasing trend from February to July, indicating a rise in collisions as the year progresses.

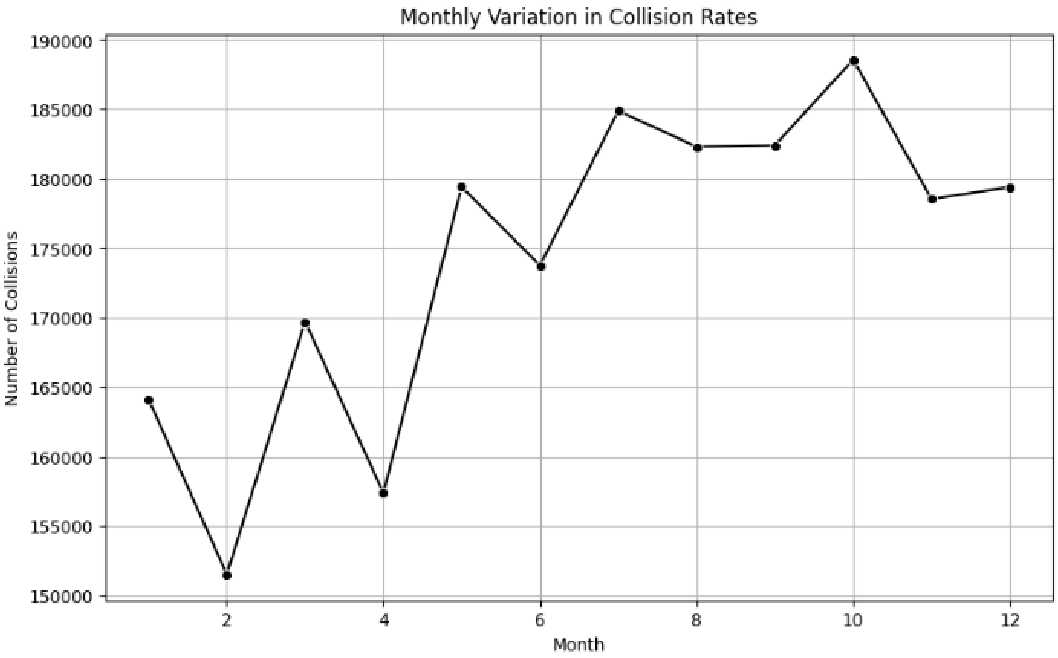
The highest collision rates occur in October, suggesting more accidents during this month. Other high-collision months include June, September and July showing consistent high rates during these periods.

Seasonal Factors:

```
The increase in collisions from February to the summer months may be due to improved weather co
The peak in October could be associated with increased travel, changing weather conditions, or
```

Winter Dip:

```
The dip in February suggests that harsh winter weather conditions, such as snow and ice, may le
```



3.7 Research Question 7

Question: Demographic Patterns and Collision Data in NYC Boroughs

Observations

Brooklyn has the highest number of injuries 152,597 and fatalities 633 among all boroughs, indi

Manhattan shows high pedestrian activity with 68,322 injuries and 333 fatalities, likely due to

Queens has significant injuries 117,027 and fatalities 534, suggesting considerable traffic act

The Bronx reports 70,229 injuries and 281 fatalities, with notable pedestrian and motorist inc

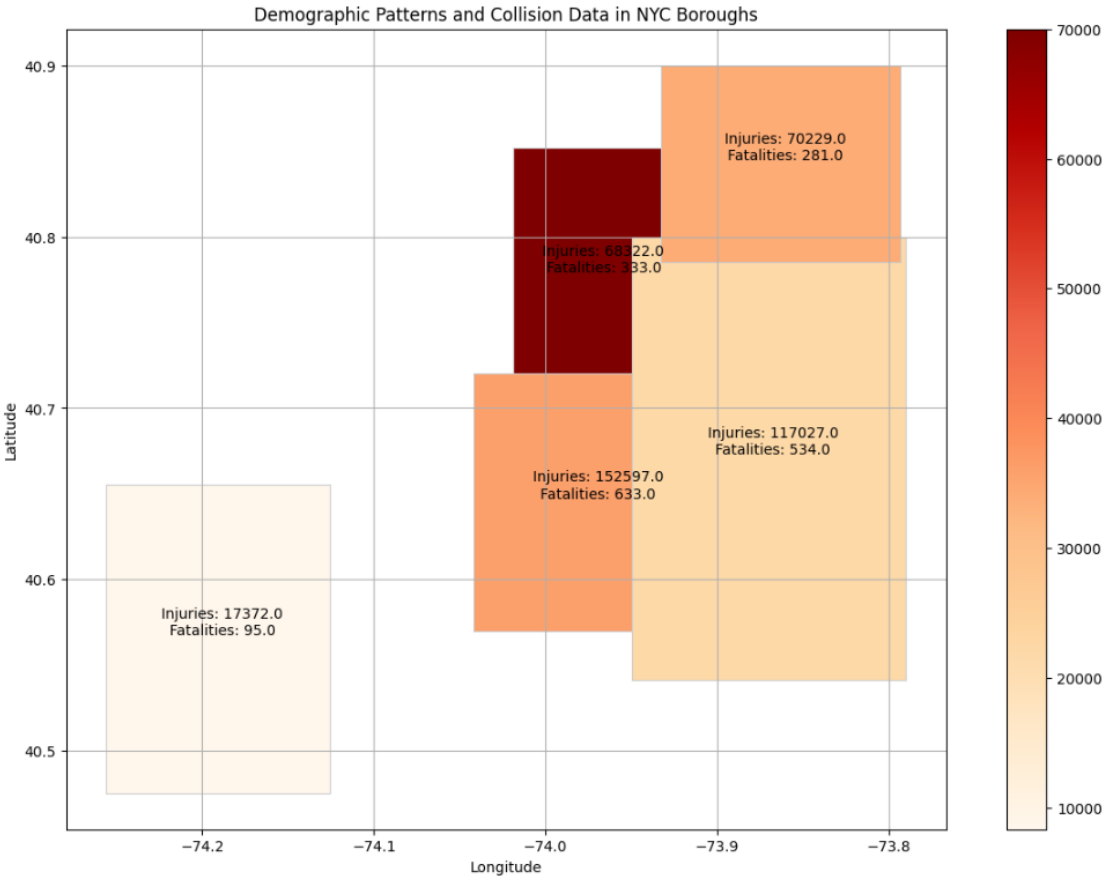
Staten Island has the lowest injury 17,372 and fatality 95 rates, correlating with its low popu

Implications

High-Injury Boroughs: Brooklyn and Queens require comprehensive safety measures to address the

Manhattan: Enhanced pedestrian safety initiatives are crucial due to its high population densi

Staten Island: Maintain and improve current safety levels to keep collision rates low.



3.8 Reflection

3.8.1 Lessons Learned

I learned many key aspects from this assignment, including the following:

1. *Importance of Data Cleaning and Preprocessing*:

   - I have learnt that cleaning and preprocessing of data is an important stage of any data-driven project. The accuracy, completeness, and proper formatting of data are very essential to get reliable results. Handling missing values, correcting errors, and normalizing data can change a lot in your analysis.

2. *Importance of Geographic Visualization*: Geographic visualizations often serve as a good way to present any data in an easy-to-understand format. These visualizations can underline patterns and insights that might not be immediately evident from the raw data, hence communicating findings more efficiently to the stakeholders.

3. *Understanding Demographic Impact*:

- The assignment primarily put forth the need and importance of incorporating demographic variables within a traffic safety analysis. There are different demographic characteristics, such as population density, age distribution, and levels of income, that may impose a considerable impact on traffic collision rates along with causal patterns.

4. Correlation and causation expressed by: I learned how to draw a line between co-relation and causality. Analysis indicated the correlations that existed between demographic features and collision frequencies. However, it cannot be stated that this correlation is causal in nature. More studies and data are required to establish causality.

3.8.2 Prior Knowledge

The things that come into mind, which I wish I had known more about when entering this project, are:

1. *Advanced GIS Techniques:* More insight into Geographic Information System techniques would have helped in the analysis. It would also have given advanced scope of inquiry onto spatial patterns and relationships using advanced GIS tools and techniques.
2. *Traffic Safety Literature:*

- Grounding of prior literature with respect to the issue should have aided the authors in sharpening their critical research questions in a more incisive way. Such an understanding of available literature would have also set the stage better for appropriate placement of results vis-à-vis such established findings. Familiarity with common traffic safety issues and interventions would have given a solid underpinning for the analysis.

3. *Integration of Data*: One such issue could have been knowledge of how to integrate the datasets. This should have avoided several steps in analysis, especially where there is the integration of traffic data with socioeconomic data. Knowing how to merge and analyze multi-source data in the best way is a time-saving strategy with accurate results.

3.8.3 Changes for Future Projects

Changes for Future Projects Given the experience from this assignment, there are a few things I would do differently in future projects:

1. *Early Stakeholder Engagement:* Engaging with stakeholders early in the project can help in understanding their needs and expectations elicit more relevant and actionable research questions. This would enable regular check-ins with the stakeholders to ensure that the analysis is on the right track, focused on the most critical issues.
2. *Improved Data Validation*: A more formal and robust data validation exercise early in the cookie-cutter project life cycle would help profile the data quality issues and take steps toward rectification. In this process, one will verify if the data is accurate and complete, and also, free from variation arising out of multiple data sources.
3. *Using Sophisticated Analytical Tools*: Elucidation: Advanced tools of machine learning algorithms with predictive modeling may bring more insight into the analysis and also may make a more accurate prediction, which the traditional approaches may not find. These tools can find hidden patterns and relationships not found by the traditional approaches.
4. *Full Documentation:* Proper documentation throughout the project, considering reproducibility and future use, must include the sources of data, cleaning, preprocessing, applied analytical methods, and interpretation of results. A well-documented project will be reviewed more easily, replicated, and developed for further work.

5. *Interdisciplinary Collaboration*: It can glean multifarious insights when collaboration is extended to experts in urban planning, public health, and the social sciences. Interdisciplinary collaboration empowers one to come up with more holistic and impactful solutions to complex problems.

3.8.4 Conclusion

This assignment was an integral part of gaining valuable learning and insight into some of the intricacies involved in data analysis, interpreted mainly with regard to the inclusion of several factors in research. By reflecting on these lessons and leveraging them in future projects, I can strive to increase the rigor and impact of my work, hence managing to further help in more effective and more informed decision-making processes.

[ ]: