Mid Term Project - Analyzing New York City Data with SQL, Python, and Version Control

Project Report: Motor Vehicle Collision

1. Introduction

The objective of this project is to empower a comprehensive exploration, analysis, and insight generation from real-world data concerning motor vehicle collisions in New York City. By working with open datasets from the NYC Open Data platform, we gain a deeper understanding of factors affecting road safety, vehicle dynamics, and demographic attributes involved in urban traffic incidents. This project focuses on applying common data analysis and collaborative software development techniques, including SQL for data merging, Python for analytical and visualization tasks, and version control to enable teamwork and versioning throughout the project lifecycle.

Motor vehicle collisions in a densely populated urban environment like New York City have significant implications on public safety, policy making, and urban planning. By examining and integrating data across three datasets—crashes_df, person_df, and vehicles_df—we aim to draw a holistic picture of collision incidents, capturing details on the circumstances, individuals, and vehicles involved. This project will involve data preprocessing, cleansing, and merging these datasets on the common identifier COLLISION_ID. Following this, exploratory data analysis (EDA) techniques will be applied to gain insights, detect patterns, and identify any outliers. We will leverage SQL for efficient data merging and data retrieval, while Python will facilitate data processing, statistical analysis, and visualization, allowing for a detailed exploration of factors associated with collisions.

2. Project Data Flow

2.1 Dataset Overview

2.1.1 Motor Vehicle Collision - Crashes

The Motor Vehicle Collisions crash table provides detailed records of NYC crashes reported by police. Each row represents a crash event based on form MV-104AN, required for incidents involving injury, fatality, or over $1,000 in damage. The data, which is preliminary and updated based on revised details, forms part of the NYPD's broader traffic safety efforts under the TrafficStat program. Established in 1998, TrafficStat applies CompStat principles to improve road safety, with data initially entered manually in the Traffic Accident Management System (TAMS) until the implementation of the Finest Online Records Management System (FORMS) in 2016. FORMS enables officers to electronically log comprehensive crash details, supporting detailed traffic safety analysis and the citywide Vision Zero initiative to eliminate traffic fatalities.

2.1.2 Motor Vehicle Collision - Vehicles

The Motor Vehicle Collisions vehicle table provides details on each vehicle involved in NYC crashes, with data recorded since April 2016 when crash reporting became electronic. Each row represents one vehicle in a crash, using information from form MV-104AN, required for incidents involving injury, fatality, or significant damage. The NYPD's TrafficStat program, started in 1998, adapted CompStat principles for traffic safety. Data collection was initially managed through the Traffic Accident Management System (TAMS) until the 2016 launch of the Finest Online Records Management System (FORMS), which electronically records detailed crash data for advanced safety analysis, supporting the Vision Zero goal to eliminate traffic fatalities.

2.1.3 Motor Vehicle Collision - Person

The Motor Vehicle Collisions person table records details for each individual involved in a crash (e.g., driver, occupant, pedestrian) since April 2016, when NYC crash reporting transitioned to an electronic system. Data comes from form MV-104AN, mandatory for incidents involving injuries, fatalities, or significant damage. The NYPD implemented TrafficStat in 1998 to apply CompStat principles to traffic incidents, initially collecting limited data via the Traffic Accident Management System (TAMS). In 2016, FORMS replaced TAMS, enabling detailed, electronic crash data collection to support Vision Zero's goal of eliminating traffic fatalities through comprehensive safety analysis.

2.2 Dataset Structure

2.2.1 Motor Vehicle Collisions - Crashes Table:

The dataset contains 2.13 million records with 29 columns, where each row represents a motor vehicle collision. Key columns include:

- CRASH DATE: Date of the collision.
- CRASH TIME: Time of the collision.
- BOROUGH: Borough where the collision occurred.
- ZIP CODE: Postal code of the incident.
- LATITUDE & LONGITUDE: Geographic coordinates of the collision (EPSG 4326).
- LOCATION: Latitude and Longitude pair.
- ON STREET NAME: Street on which the collision occurred.
- CROSS STREET NAME: Nearest cross street to the collision.
- OFF STREET NAME: Address of the incident if known.
- NUMBER OF PERSONS INJURED: Total persons injured in the collision.
- NUMBER OF PERSONS KILLED: Total persons killed in the collision.

- NUMBER OF PEDESTRIANS INJURED/KILLED: Counts of pedestrians injured or killed.
- NUMBER OF CYCLIST INJURED/KILLED: Counts of cyclists injured or killed.
- NUMBER OF MOTORIST INJURED/KILLED: Counts of motor vehicle occupants injured or killed.
- CONTRIBUTING FACTOR VEHICLE 1–5: Contributing factors for up to five vehicles.
- COLLISION_ID: System-generated unique record code (Primary Key).
- VEHICLE TYPE CODE 1–5: Type of vehicle involved (e.g., Car/SUV, Motorcycle, Truck/Bus).

2.2.2 Motor Vehicle Collisions - Vehicle Table:

The dataset contains 4.28 million records with 25 columns, where each row represents a motor vehicle involved in a crash. Key columns include:

- UNIQUE_ID: System-generated unique record code (Primary Key).
- COLLISION_ID: Crash identification code linking to the Crash table (Foreign Key).
- CRASH_DATE: Date of the collision.
- CRASH_TIME: Time of the collision.
- VEHICLE_ID: Unique vehicle identification code.
- STATE_REGISTRATION: State where the vehicle is registered.
- VEHICLE_TYPE: Type of vehicle (e.g., Car/SUV, Motorcycle, Truck/Bus).
- VEHICLE_MAKE: Vehicle make (brand).
- VEHICLE_MODEL: Vehicle model.
- VEHICLE_YEAR: Manufacturing year of the vehicle.
- TRAVEL_DIRECTION: Direction in which the vehicle was traveling.
- VEHICLE_OCCUPANTS: Number of occupants in the vehicle.
- DRIVER_SEX: Driver's gender.
- DRIVER_LICENSE_STATUS: License status (e.g., Licensed, Permit).
- DRIVER_LICENSE_JURISDICTION: State issuing the driver's license.
- PRE_CRASH: Pre-crash action (e.g., Going Straight, Making Right Turn).
- POINT_OF_IMPACT: Initial point of impact on the vehicle.
- VEHICLE_DAMAGE: Primary location of damage on the vehicle.
- VEHICLE_DAMAGE_1, VEHICLE_DAMAGE_2, VEHICLE_DAMAGE_3: Additional damage locations.
- PUBLIC_PROPERTY_DAMAGE: Whether public property was damaged (Yes or No).
- PUBLIC_PROPERTY_DAMAGE_TYPE: Type of public property damaged (e.g., Sign, Light Post).
- CONTRIBUTING_FACTOR_1 & CONTRIBUTING_FACTOR_2: Factors contributing to the collision.

2.2.3 Motor Vehicle Collisions – Person Table:

The dataset contains 5.52 million records with 21 columns, where each row represents a person involved in a crash (driver, occupant, pedestrian, bicyclist, etc.). Key columns include:

- UNIQUE_ID: System-generated unique record code (Primary Key for Person table).
- COLLISION_ID: Crash identification code matching the unique ID in the Crash table (Foreign Key).
- CRASH_DATE: Date of the collision.
- CRASH_TIME: Time of the collision.
- PERSON_ID: Identification code for each person involved.
- PERSON_TYPE: Role of the person (e.g., Bicyclist, Motor Vehicle Occupant, Pedestrian).
- PERSON_INJURY: Injury status (e.g., Injured, Killed, Unspecified).
- VEHICLE_ID: Vehicle identification code associated with the person (Foreign Key to the Vehicle table).
- PERSON_AGE: Age of the person.
- EJECTION: Whether the person was ejected, partially ejected, or not ejected from the vehicle.
- EMOTIONAL_STATUS: Observed emotional state (e.g., Apparent Death, Unconscious).
- BODILY_INJURY: Specific injured body areas (e.g., Head, Neck).
- POSITION_IN_VEHICLE: Seating position within the vehicle (e.g., Driver, Front Passenger).
- SAFETY_EQUIPMENT: Safety equipment used (e.g., Lap Belt, Air Bag).
- PED_LOCATION: Pedestrian's location relative to the intersection.
- PED_ACTION: Pedestrian's activity during the crash (e.g., Walking with the Signal).
- COMPLAINT: Type of physical complaint (e.g., Concussion).
- PED_ROLE: Role of the pedestrian (e.g., Pedestrian, Witness).
- CONTRIBUTING_FACTOR_1 & CONTRIBUTING_FACTOR_2: Factors contributing to the collision.
- PERSON_SEX: Gender of the person.

2.3 Data Preprocessing and Cleaning

2.3.1 Loading the Dataset

In this step, the Motor Vehicle Collision [Vehicles, Person, Crashes] datasets are imported into a pandas DataFrame. The dataset is converted from its original format into CSV for convenient access and sharing. Key functions are used to gain an understanding of the data's structure and content:

- **Loading the Data**: The dataset is read into a pandas DataFrame.
- **Understanding the Data**:
    - **Size**: Provides the total number of data elements (rows × columns).
    - **Shape**: Displays the dataset's dimensions, indicating the number of rows and columns.
    - **Info**: Shows data types for each column and the count of non-null values per column.
    - **Describe**: Provides summary statistics (e.g., mean, minimum, maximum) for numerical columns.

This initial examination gives a foundational overview of the dataset, helping to identify missing values, data types (numerical, categorical), and areas that may require data cleaning before further analysis.

2.3.2 For Motor Vehicle Collision - Crashes

Handling Missing Values

Missing values were identified and addressed. Columns like NUMBER OF PERSONS KILLED and NUMBER OF PERSONS INJURED had NaN values replaced with 0 and converted to int64. The BOROUGH column, with 31% missing values, was filled with 'Unspecified'.

Removing Duplicates and Irrelevant Columns

Duplicate records based on COLLISION_ID were identified and removed. Irrelevant columns such as ZIP CODE, LATITUDE, and street names were dropped to simplify the dataset for analysis, keeping only the essential BOROUGH and LOCATION columns.

Standardizing Contributing Factors

Misspellings in the CONTRIBUTING FACTOR VEHICLE 1 and CONTRIBUTING FACTOR VEHICLE 2 columns were corrected, and missing values were filled with 'Unspecified' for consistency.

2.3.3 Motor Vehicle Collision - Person

Handling Missing Values

- **Null Values**: Missing values in critical columns were handled. For example, the PERSON_SEX column was filled with 'U', while other columns such as EJECTION, EMOTIONAL_STATUS, and BODILY_INJURY were filled with 'Unknown'.
- **Dropping Columns**: The columns CONTRIBUTING_FACTOR_1, CONTRIBUTING_FACTOR_2, PED_ACTION, and PED_LOCATION were dropped due to the high percentage of missing values (over 98%).
- **Handling Null in Specific Columns**: The PERSON_ID column, which had missing values, was dropped entirely. Columns like VEHICLE_ID and PERSON_AGE were filled with 0 and converted to int64 type.

Removing Duplicates

Duplicate records were identified based on the UNIQUE_ID and removed. This ensures that each record is unique within the dataset.

Merging Crash Date and Time

The CRASH_DATE and CRASH_TIME columns were merged into a single DateTime column for consistency. The CRASH_TIME column was then dropped.

2.3.4 Motor Vehicle Collision - Vehicle

Handling Missing Values

- **Null Values**: Missing values in columns such as CONTRIBUTING_FACTOR_1, CONTRIBUTING_FACTOR_2, and PUBLIC_PROPERTY_DAMAGE were filled with 'Unspecified'. The DRIVER_SEX column was filled with 'U'.
- **Dropping Columns**: The columns PUBLIC_PROPERTY_DAMAGE_TYPE and VEHICLE_MODEL were dropped due to the high percentage of missing values (over 99%).
- **Converting Data Types**: The VEHICLE_YEAR and VEHICLE_OCCUPANTS columns were filled with 0 and converted to int type.
- **Filling Object Columns**: Remaining object columns (excluding the ones filled previously) were filled with 'Unknown'.

Handling Duplicates

- **Duplicate Rows**: No duplicate rows were found in the dataset.
- **Duplicate UNIQUE_ID**: No duplicate UNIQUE_ID entries were found, ensuring each record is unique.

Handling Date and Time

- **Date Conversion**: The CRASH_DATE was converted to a datetime object, and only the date part was retained.
- **Combining Date and Time**: The CRASH_DATE and CRASH_TIME columns were combined into a single DateTime column for consistency. The CRASH_TIME column was then dropped.

3. SQL Queries

3.1 Collision Analysis: To analyze collisions by type and count.

Objective

The objective of this query is to identify the frequency of each vehicle type involved in collisions. By analyzing the `VEHICLE TYPE CODE 1` column in the `crashes` table, the query counts occurrences of each unique vehicle type in accidents. This can reveal which types of vehicles are most frequently involved in crashes, which may help in identifying patterns in vehicle type and accident risk.

Explanation of the Query

- **`SELECT "VEHICLE TYPE CODE 1"`**: Selects the primary vehicle type involved in each collision.
- **`COUNT(*) AS type_count`**: Counts the total number of records for each vehicle type, effectively measuring the number of accidents each vehicle type was involved in.
- **`FROM crashes`**: Specifies the source table, `crashes`.
- **`GROUP BY "VEHICLE TYPE CODE 1"`**: Groups the data by each unique vehicle type, so the count will be calculated for each individual type.
- **`ORDER BY type_count DESC`**: Orders the results in descending order based on `type_count`, so the most frequently involved vehicle type appears at the top.

Inference

By running this query, we gain insight into the distribution of vehicle types involved in crashes. For example, if we see that certain types of vehicles, like **SUVs** or **motorcycles**, have a higher `type_count`, we might infer that these vehicle types are more frequently involved in accidents compared to others. This information could be indicative of underlying factors such as vehicle size, driver demographics, or other behavioral patterns specific to certain vehicle types.

Conclusions

1. **High-Risk Vehicle Types**: If specific vehicle types, such as motorcycles or trucks, appear at the top of the list with high counts, they might be more prone to accidents. This could be due to factors like increased exposure to hazards, more challenging handling, or increased usage.

2. **Inform Policy and Safety Initiatives**: The results can inform regulatory bodies or traffic safety agencies. For instance, if motorcycles frequently appear in crash data, the findings could support implementing targeted safety programs, stricter regulations, or specialized infrastructure (like dedicated motorcycle lanes).

3. **Insight for Insurance Companies**: Insurance companies could use the insights to adjust premiums or develop targeted insurance plans for higher-risk vehicle types, potentially offering incentives for safe driving in these categories.

4. **Further Analysis**: The query's results could motivate additional analysis on specific vehicle types with high accident counts to understand underlying causes, such as driver demographics, road conditions, or common accident types.

3.2 Location analysis: Accident frequency by location¶

Objective

The objective of this query is to analyze the distribution of traffic accidents across different boroughs. By counting the number of crashes occurring in each borough, we aim to identify which areas experience the highest frequency of accidents. This information can help highlight high-risk areas, inform targeted safety measures, and support resource allocation for traffic management.

Explanation of the Query

- **`SELECT BOROUGH`**: Selects the borough where each crash occurred.
- **`COUNT(*) AS accident_count`**: Counts the total number of crashes within each borough, effectively measuring accident frequency by area.
- **`FROM crashes`**: Specifies the source table, `crashes`, which contains records of individual accidents.
- **`GROUP BY BOROUGH`**: Groups the data by each unique borough, so that the count is calculated for each borough separately.
- **`ORDER BY accident_count DESC`**: Orders the results in descending order based on `accident_count`, so the borough with the highest number of accidents appears at the top.

Inference

The query reveals which boroughs have the highest and lowest numbers of traffic accidents. Boroughs with a higher `accident_count` may face unique factors that contribute to a higher accident frequency, such as:

- Higher traffic volume or population density
- Presence of busy intersections or high-risk roads
- Inadequate traffic infrastructure or signage
- Increased pedestrian or cyclist activity

This information helps in understanding the spatial distribution of traffic accidents and could be used to identify boroughs that require more focused attention for traffic safety initiatives.

Conclusions

1. **Identification of High-Risk Areas**: Boroughs with a high `accident_count` can be considered high-risk areas for traffic collisions. Traffic safety authorities may use this information to allocate resources, such as increased patrolling, installation of traffic signals, and improved road design.

2. **Improvement of Traffic Safety Measures**: High-accident boroughs may benefit from targeted safety measures such as additional signage, improved pedestrian crossings, or traffic-calming strategies like speed bumps. This could be a focal point for urban planners and policymakers to enhance traffic safety.

3. **Support for Public Awareness Campaigns**: Knowing which boroughs have higher accident rates can help in developing localized public awareness campaigns focused on safe driving and pedestrian behavior in specific high-risk areas.

4. **Further Research Opportunities**: This query serves as a starting point for deeper analysis. For example, further investigation could explore accident types, times of day, or vehicle types that contribute to higher accident rates in specific boroughs.

3.3 Contributing factors and their related severities

Objective

The objective of this query is to examine the relationship between the primary contributing factors of vehicle crashes (`contributing factor vehicle 1`) and the resulting severity of injuries (`Person_injury`). By counting the occurrences of each injury severity level for each contributing factor, we aim to identify which factors are most often associated with severe or fatal injuries. This analysis can help prioritize safety measures and interventions for the most dangerous driving behaviors or conditions.

Explanation of the Query

- **`SELECT "contributing factor vehicle 1"`**: Retrieves the main contributing factor associated with each crash from the `crashes` table (e.g., speeding, driver inattention).
- **`Person_injury`**: Selects the severity of injury (e.g., Injured, Killed) from the `person` table.
- **`COUNT(*) AS severity_count`**: Counts the number of occurrences for each unique combination of contributing factor and injury severity, giving the frequency of each injury type for each factor.

- **`FROM crashes JOIN person ON crashes.COLLISION_ID = person.COLLISION_ID`**: Joins the `crashes` and `person` tables on the `COLLISION_ID`, allowing us to analyze injury details alongside crash contributing factors.
- **`GROUP BY "contributing factor vehicle 1", Person_injury`**: Groups the data by each unique pair of contributing factor and injury severity, so we get counts for each combination.
- **`ORDER BY "contributing factor vehicle 1", severity_count DESC`**: Orders results first by the contributing factor and then by severity count in descending order, so the most frequent injury levels per factor are shown at the top.

Inference

This query reveals which contributing factors are associated with different levels of injury severity. By analyzing the count of each injury type per contributing factor, we can make several inferences:

- **High-Severity Factors**: Contributing factors with a high `severity_count` for fatal or serious injuries might indicate particularly hazardous conditions or driver behaviors.
- **Common vs. Severe Factors**: Some factors may appear frequently (e.g., minor distractions) but result in mostly non-serious injuries. Conversely, less common factors (e.g., high-speed impact) might correlate with severe injuries.
- **Insights on Prevention**: Identifying the most frequent contributing factors for specific injury types can guide traffic authorities to target the most impactful causes of accidents through prevention programs.

Conclusions

1. **Targeted Interventions for High-Severity Factors**: If certain contributing factors (e.g., speeding, impaired driving) consistently result in severe injuries or fatalities, traffic safety authorities might prioritize these for intervention. For instance, more stringent speed control and impaired driving campaigns can be implemented.

2. **Enhanced Awareness Campaigns**: For frequently occurring but less severe contributing factors (e.g., distracted driving), public awareness campaigns could be created to educate drivers about the risks, potentially reducing the overall accident rate.

3. **Prioritizing Road Safety Measures**: The analysis can also support infrastructure improvements. For example, intersections or high-traffic areas prone to accidents caused by visibility issues or other specific contributing factors can be re-engineered to mitigate risk.

4. **Further Analysis and Safety Research**: This query provides foundational insights that could lead to more detailed research. For example, additional analysis could explore whether certain contributing factors correlate with specific demographics or times of day, allowing for even more targeted safety measures.

3.4 Highest contributing factor of highest number of crashes

Objective

The objective of this query is to examine the relationship between the primary contributing factors of vehicle crashes (`contributing factor vehicle 1`) and the resulting severity of injuries (`Person_injury`). By counting the occurrences of each injury severity level for each contributing factor, we aim to identify which factors are most often associated with severe or fatal injuries. This analysis can help prioritize safety measures and interventions for the most dangerous driving behaviors or conditions.

Explanation of the Query

- **`SELECT "contributing factor vehicle 1"`**: Retrieves the main contributing factor associated with each crash from the `crashes` table (e.g., speeding, driver inattention).
- **`Person_injury`**: Selects the severity of injury (e.g., Injured, Killed) from the `person` table.
- **`COUNT(*) AS severity_count`**: Counts the number of occurrences for each unique combination of contributing factor and injury severity, giving the frequency of each injury type for each factor.
- **`FROM crashes JOIN person ON crashes.COLLISION_ID = person.COLLISION_ID`**: Joins the `crashes` and `person` tables on the `COLLISION_ID`, allowing us to analyze injury details alongside crash contributing factors.
- **`GROUP BY "contributing factor vehicle 1", Person_injury`**: Groups the data by each unique pair of contributing factor and injury severity, so we get counts for each combination.
- **`ORDER BY "contributing factor vehicle 1", severity_count DESC`**: Orders results first by the contributing factor and then by severity count in descending order, so the most frequent injury levels per factor are shown at the top.

Inference

This query reveals which contributing factors are associated with different levels of injury severity. By analyzing the count of each injury type per contributing factor, we can make several inferences:

- **High-Severity Factors**: Contributing factors with a high `severity_count` for fatal or serious injuries might indicate particularly hazardous conditions or driver behaviors.
- **Common vs. Severe Factors**: Some factors may appear frequently (e.g., minor distractions) but result in mostly non-serious injuries. Conversely, less common factors (e.g., high-speed impact) might correlate with severe injuries.

- **Insights on Prevention**: Identifying the most frequent contributing factors for specific injury types can guide traffic authorities to target the most impactful causes of accidents through prevention programs.

Conclusions

1. **Targeted Interventions for High-Severity Factors**: If certain contributing factors (e.g., speeding, impaired driving) consistently result in severe injuries or fatalities, traffic safety authorities might prioritize these for intervention. For instance, more stringent speed control and impaired driving campaigns can be implemented.

2. **Enhanced Awareness Campaigns**: For frequently occurring but less severe contributing factors (e.g., distracted driving), public awareness campaigns could be created to educate drivers about the risks, potentially reducing the overall accident rate.

3. **Prioritizing Road Safety Measures**: The analysis can also support infrastructure improvements. For example, intersections or high-traffic areas prone to accidents caused by visibility issues or other specific contributing factors can be re-engineered to mitigate risk.

4. **Further Analysis and Safety Research**: This query provides foundational insights that could lead to more detailed research. For example, additional analysis could explore whether certain contributing factors correlate with specific demographics or times of day, allowing for even more targeted safety measures.

3.5 Total Injuries by Contributing Factor and Vehicle Type: Finding the total injuries for each combination of contributing factor and vehicle type

Objective

The objective of this query is to analyze the relationship between the primary contributing factors of vehicle collisions (`CONTRIBUTING_FACTOR_1`) and the types of vehicles involved (`VEHICLE_TYPE`), focusing on incidents where injuries occurred. By counting the number of injuries for each combination of contributing factor and vehicle type, this query aims to identify which factors and vehicle types are most frequently associated with injuries in crashes. This information can support targeted traffic safety interventions and contribute to understanding the risk profiles of different vehicle types and behaviors.

Explanation of the Query

- **`SELECT v."CONTRIBUTING_FACTOR_1"`**: Retrieves the primary contributing factor for each vehicle involved in the collision (e.g., speeding, distracted driving).
- **`v."VEHICLE_TYPE"`**: Retrieves the type of vehicle involved in the collision (e.g., car, truck, motorcycle).

- **`COUNT(p.PERSON_INJURY) AS total_injuries`**: Counts the number of injuries (`PERSON_INJURY = 'Injured'`) for each combination of contributing factor and vehicle type.
- **`FROM vehicles AS v JOIN person AS p ON v.COLLISION_ID = p.COLLISION_ID`**: Joins the `vehicles` and `person` tables on `COLLISION_ID`, allowing us to analyze contributing factors and vehicle types alongside injury data.
- **`WHERE p.PERSON_INJURY = 'Injured'`**: Filters to include only records where `PERSON_INJURY` is classified as 'Injured'.
- **`GROUP BY v."CONTRIBUTING_FACTOR_1", v."VEHICLE_TYPE"`**: Groups results by each unique pair of contributing factor and vehicle type to obtain counts for each combination.
- **`ORDER BY total_injuries DESC LIMIT 10`**: Orders results by the total number of injuries in descending order and limits the output to the top 10 results, showing the highest-risk combinations of vehicle type and contributing factor.

Inference

The results of this query provide insights into which vehicle types are most frequently involved in collisions with injuries for specific contributing factors. Potential inferences include:

- **High-Injury Contributing Factors**: Some contributing factors may be associated with a high number of injuries across multiple vehicle types, indicating particularly dangerous behaviors or conditions.
- **Risky Vehicle Types**: Certain vehicle types, such as motorcycles or trucks, may appear more frequently in injury-causing incidents for specific contributing factors. This could reflect increased vulnerability or risk in certain types of collisions.
- **Pattern Identification for Safety Measures**: By understanding the combinations of vehicle types and contributing factors that frequently result in injuries, traffic safety authorities can target these areas with tailored safety measures.

Conclusions

1. **Identify High-Risk Combinations**: If certain vehicle types (e.g., motorcycles) and contributing factors (e.g., speeding, distracted driving) consistently result in injuries, these combinations can be flagged as high-risk. Safety campaigns and enforcement efforts can prioritize these specific combinations.

2. **Targeted Safety Interventions**: Based on the findings, authorities might focus on stricter regulation or awareness campaigns around high-risk behaviors. For instance, if large trucks frequently have injury-causing collisions due to impaired visibility, adding specialized signage or restricted lanes might mitigate these incidents.

3. **Vehicle-Specific Awareness Programs**: Insights on risky vehicle types may justify targeted safety programs. For example, if motorcycles show high injury counts with certain factors, there might be an opportunity to emphasize awareness around motorcycle visibility and safety for other drivers.

4. **Opportunities for Future Research**: This query provides a baseline for further analysis. For instance, authorities might want to investigate additional factors such as time of day, location, or road conditions to build a more comprehensive risk profile.

3.6 Top Crash Dates with the Most Injuries by Contributing Factor: Finding the dates with the highest number of injuries for each contributing factor in the crashes dataset.

Objective

The objective of this query is to analyze the relationship between the date of each crash, the primary contributing factor in each collision (`CONTRIBUTING_FACTOR_1`), and the number of injuries (`PERSON_INJURY`). By identifying the dates with the highest number of injuries for each contributing factor, this query provides insights into patterns and trends in accident injuries over time. This information can help inform safety measures and targeted interventions to address specific contributing factors associated with injuries on particular dates.

Explanation of the Query

- **`SELECT c."CRASH DATE"`**: Selects the date of each crash from the `crashes` table.
- **`v."CONTRIBUTING_FACTOR_1"`**: Retrieves the primary contributing factor for each vehicle in the collision from the `vehicles` table (e.g., speeding, driver inattention).
- **`COUNT(p."PERSON_INJURY") AS total_injuries`**: Counts the number of injuries (where `PERSON_INJURY = 'Injured'`) associated with each combination of crash date and contributing factor.
- **`FROM crashes AS c JOIN person AS p ON c.COLLISION_ID = p.COLLISION_ID`**: Joins the `crashes` and `person` tables on `COLLISION_ID` to analyze the injury details for each crash.
- **`JOIN vehicles AS v ON c.COLLISION_ID = v.COLLISION_ID`**: Joins the `vehicles` table on `COLLISION_ID` to bring in contributing factors associated with each collision.
- **`WHERE p."PERSON_INJURY" = 'Injured'`**: Filters the data to include only records where `PERSON_INJURY` is classified as 'Injured'.
- **`GROUP BY c."CRASH DATE", v."CONTRIBUTING_FACTOR_1"`**: Groups the results by each unique combination of crash date and contributing factor.
- **`ORDER BY total_injuries DESC LIMIT 20`**: Orders the results by the number of injuries in descending order, limiting the output to the top 20 entries with the highest injury counts.

Inference

The query results provide insights into the dates and contributing factors that are most frequently associated with injuries in crashes. Potential inferences include:

- **High-Risk Dates for Specific Contributing Factors**: Certain dates may show high injury counts associated with specific contributing factors, suggesting an increased accident risk during those times. These could correlate with specific events, holidays, or seasonal factors.
- **Recurrent Dangerous Factors**: Contributing factors such as speeding or driver inattention may appear frequently on high-injury dates, highlighting consistent risk factors that lead to injuries.
- **Temporal Trends**: If specific dates or times of the year show repeated high injury counts, it may suggest patterns, such as increased risk during weekends, holidays, or adverse weather conditions.

Conclusions

1. **Identify High-Injury Dates for Targeted Interventions**: If certain dates exhibit a high number of injuries, traffic authorities can consider specific interventions, such as increased patrolling or public safety announcements, around those dates to prevent accidents.

2. **Focus on High-Risk Contributing Factors**: Consistent contributing factors with high injury counts (like speeding or distracted driving) could be targeted for stricter enforcement measures, awareness campaigns, or driver education programs.

3. **Support for Seasonal or Event-Based Safety Measures**: If injury-prone dates align with specific times of the year (e.g., holidays or events), implementing seasonal traffic management and public awareness programs may help reduce risks. Authorities can increase visibility and cautionary signage or enforce speed limits around these dates.

4. **Further Research on Accident Patterns**: This query lays the groundwork for additional research, such as analyzing accident times, weather conditions, or road types associated with these high-risk dates and contributing factors, to gain a fuller understanding of accident dynamics.

3.7 Number of persons injured based on their position in the vehicle.

Objective

The objective of this query is to analyze the relationship between the position of individuals within a vehicle (`POSITION_IN_VEHICLE`) and the number of injuries resulting from collisions. By summing the total number of injuries per position, this query aims to identify which seating positions in the vehicle are most frequently associated with injuries in crashes. This information can be useful for understanding vehicle safety dynamics and prioritizing safety improvements for specific seating positions.

Explanation of the Query

- **`SELECT p.POSITION_IN_VEHICLE`**: Retrieves the position of each individual within the vehicle (e.g., driver, front passenger, rear passenger) from the `person` table.
- **`SUM(c."NUMBER OF PERSONS INJURED") AS total_injuries`**: Sums the number of injuries from the `crashes` table for each seating position. This provides the total injuries associated with each seating position across all crashes.
- **`FROM person AS p JOIN crashes AS c ON p.COLLISION_ID = c.COLLISION_ID`**: Joins the `person` and `crashes` tables on `COLLISION_ID`, allowing analysis of injuries by individual seating positions.
- **`WHERE c."NUMBER OF PERSONS INJURED" > 0`**: Filters the data to include only records where injuries occurred in the crash.
- **`GROUP BY p.POSITION_IN_VEHICLE`**: Groups the results by each seating position within the vehicle, so the total injuries are calculated for each unique position.
- **`ORDER BY total_injuries DESC`**: Orders the results by the total number of injuries in descending order, showing the seating positions most frequently associated with injuries at the top.

Inference

The results of this query can provide insights into the relative safety of different seating positions within a vehicle. Key inferences include:

- **High-Risk Seating Positions**: Certain seating positions may show a higher number of injuries, suggesting they are more vulnerable during collisions. For example, the front passenger or driver positions may have more exposure to impact zones, making them more injury-prone.
- **Potential Benefits of Safety Equipment**: If certain seating positions (such as rear seats) have fewer injuries, this may indicate the effectiveness of safety equipment like seatbelts and airbags in these areas.
- **Occupant Vulnerability**: The query may reveal that occupants in specific positions, such as the driver, are at greater risk, likely due to exposure in frontal impacts.

Conclusions

1. **Focus on Enhanced Safety for High-Risk Positions**: If certain positions, such as the driver or front passenger seat, are consistently associated with higher injury counts, vehicle manufacturers and safety regulators might prioritize additional safety features or protections for these positions. These could include reinforced airbags, improved seatbelt mechanisms, or enhanced crash impact absorption for those areas.

2. **Encouragement of Rear-Seat Usage**: If rear seating positions show lower injury rates, public awareness campaigns could encourage passengers to sit in these safer areas. Parents, for example, may be encouraged to place children in rear seats whenever possible.

3. **Data-Driven Recommendations for Safety Improvements**: This analysis supports data-driven decisions to improve vehicle safety features tailored to specific seating positions. For instance, automakers might introduce new safety technologies in higher-risk seating

areas, or regulatory agencies might implement new standards focused on injury-prone positions.

4. **Further Research on Injury Causes by Position**: This query serves as a baseline for further analysis, such as studying the types of injuries most commonly associated with each position or correlating injuries by seating position with specific crash types (e.g., frontal vs. rear-end collisions).

3.8 Number of persons injured based on their position in the vehicle.

Objective

The objective of this query is to analyze the number of unique collisions involving vehicles registered in different states. By counting distinct `COLLISION_ID`s for each `STATE_REGISTRATION`, this query identifies which states have the highest number of vehicles involved in collisions. This information can provide insights into collision patterns based on vehicle registration locations and may reflect traffic density, interstate travel patterns, or regional driving behaviors.

Explanation of the Query

- **`SELECT STATE_REGISTRATION`**: Retrieves the state in which each vehicle involved in a collision is registered.
- **`COUNT(DISTINCT COLLISION_ID) AS collision_count`**: Counts the unique `COLLISION_ID`s for each state, measuring the total number of distinct collisions in which vehicles from each state were involved.
- **`FROM vehicles`**: Specifies the `vehicles` table as the data source.
- **`GROUP BY STATE_REGISTRATION`**: Groups the results by each unique state registration, so the collision count is calculated separately for each state.
- **`ORDER BY collision_count DESC`**: Orders the results by the total number of collisions in descending order, showing states with the highest collision involvement at the top.

Inference

The query results highlight the states with the highest number of vehicles involved in collisions, offering several inferences:

- **High-Collision States**: States appearing at the top of the list have vehicles involved in the most collisions. This may reflect a combination of high population density, greater vehicle ownership, or higher levels of interstate travel.
- **Influence of Traffic Density**: States with large metropolitan areas or popular travel routes might show higher collision counts, suggesting that these areas may have higher traffic density and, consequently, more collisions.

- **Interstate Travel Patterns**: If states located along major highways or border regions appear frequently, this might indicate significant interstate travel, where vehicles from various states interact on common routes.

Conclusions

1. **Resource Allocation for High-Collision States**: States with higher collision counts may benefit from targeted safety interventions, such as increased road patrolling, enhanced signage, or educational programs focused on collision prevention. This data could guide the allocation of resources to high-risk areas.

2. **Further Analysis of High-Involvement States**: States consistently showing high involvement in collisions might warrant further investigation into underlying factors, such as traffic density, road infrastructure, or regional driving habits. Insights from such analyses could support state-specific traffic management and safety strategies.

3. **Support for Regional Safety Initiatives**: Regions with frequent vehicle collisions could implement region-specific safety initiatives. For example, states with high interstate travel may benefit from targeted campaigns on safe long-distance driving practices, or states with major cities may prioritize urban traffic safety.

4. **Identification of Cross-State Patterns**: The query results can reveal patterns in cross-state vehicle involvement. For instance, states with high numbers of out-of-state collisions might collaborate on interstate safety campaigns or advocate for consistent traffic regulations across borders.

3.9 Number of injuries based on the status of usage of safety equipments.

Objective

The objective of this query is to analyze the relationship between the type of safety equipment used (`SAFETY_EQUIPMENT`) and the resulting injury outcomes (`PERSON_INJURY`) in vehicle collisions. By counting occurrences of each injury type for each type of safety equipment, this query aims to reveal the effectiveness of various safety equipment in reducing injury severity. This analysis can inform safety standards and encourage the use of effective safety equipment to minimize injuries.

Explanation of the Query

- **`SELECT SAFETY_EQUIPMENT`**: Retrieves the type of safety equipment used by each individual involved in the collision (e.g., seatbelt, helmet, child safety seat).
- **`PERSON_INJURY`**: Retrieves the injury outcome for each person (e.g., Injured, Killed, No Injury).
- **`COUNT(*) AS injury_count`**: Counts the number of occurrences for each combination of safety equipment and injury outcome, indicating how frequently each injury type occurs for each type of safety equipment.

- **`FROM person`**: Specifies the `person` table as the source of data.
- **`GROUP BY SAFETY_EQUIPMENT, PERSON_INJURY`**: Groups the results by each unique pair of safety equipment and injury outcome, allowing for counts to be calculated for each combination.
- **`ORDER BY SAFETY_EQUIPMENT, injury_count DESC`**: Orders the results first by `SAFETY_EQUIPMENT` and then by `injury_count` in descending order, so the most frequent injury types for each safety equipment type appear at the top.

Inference

This query provides insights into the effectiveness of different types of safety equipment in preventing or mitigating injuries. Potential inferences include:

- **Effectiveness of Specific Safety Equipment**: Equipment types with lower counts of severe injuries (such as deaths) may be more effective in protecting individuals during collisions.
- **Risk Associated with Lack of Safety Equipment**: If certain injury types (like severe injuries or fatalities) are more frequent among people not using safety equipment, it underscores the importance of using safety measures.
- **Vulnerabilities of Certain Safety Equipment**: Some safety equipment may be associated with higher injury counts, suggesting that it might be less effective in certain crash scenarios or that proper usage could be a factor.

Conclusions

1. **Promote the Use of Effective Safety Equipment**: If certain types of safety equipment (such as seatbelts) are associated with fewer or less severe injuries, this supports promoting these types of equipment through public safety campaigns and education initiatives.

2. **Strengthen Safety Regulations**: Based on the analysis, regulatory authorities could implement or enhance laws requiring effective safety equipment. For example, if helmets are associated with fewer head injuries in motorcyclists, authorities might mandate helmet use in all regions.

3. **Identify Need for Improved Safety Equipment**: If specific types of safety equipment (like airbags) are still associated with high injury counts, further research may be needed to improve their design or effectiveness. Automotive manufacturers could be encouraged to enhance these features to improve safety outcomes.

4. **Encourage Proper Use of Safety Equipment**: High injury counts associated with certain safety equipment may indicate improper usage rather than a lack of effectiveness. This insight could lead to educational campaigns that teach people how to use safety equipment correctly (e.g., ensuring that seatbelts and child safety seats are properly secured).

3.10 Analyzing relationship between pre-crash situation and contributing factor based on the number of injuries in each collision¶

Objective

The objective of this query is to analyze the relationship between the actions or circumstances just before a crash (`PRE_CRASH`) and the primary contributing factor to the crash (`CONTRIBUTING_FACTOR_1`), with a focus on the resulting number of injuries. By aggregating the total number of injuries for each unique combination of pre-crash actions and contributing factors, this query aims to identify the most dangerous combinations of driver actions and external factors that lead to injuries. This information can help in developing targeted interventions for accident prevention and injury reduction.

Explanation of the Query

- **`SELECT v.PRE_CRASH`**: Retrieves the action or situation of the vehicle just before the crash occurred (e.g., going straight, turning, stopping).
- **`v.CONTRIBUTING_FACTOR_1`**: Retrieves the primary contributing factor to the crash from the `vehicles` table (e.g., speeding, distracted driving).
- **`SUM(c."NUMBER OF PERSONS INJURED") AS total_injuries`**: Calculates the total number of persons injured for each combination of pre-crash action and contributing factor.
- **`FROM vehicles AS v JOIN crashes AS c ON v.COLLISION_ID = c.COLLISION_ID`**: Joins the `vehicles` and `crashes` tables on `COLLISION_ID`, allowing analysis of injury data alongside pre-crash actions and contributing factors.
- **`WHERE c."NUMBER OF PERSONS INJURED" > 0`**: Filters the data to include only cases where at least one injury occurred.
- **`GROUP BY v.PRE_CRASH, v.CONTRIBUTING_FACTOR_1`**: Groups the results by each unique pair of pre-crash action and contributing factor, allowing the calculation of injury totals for each combination.
- **`ORDER BY total_injuries DESC`**: Orders the results by the total number of injuries in descending order, showing the combinations most frequently associated with injuries at the top.

Inference

This query provides insights into how specific pre-crash actions and contributing factors contribute to injury outcomes. Potential inferences include:

- **High-Risk Pre-Crash and Contributing Factor Combinations**: Certain combinations of pre-crash actions (e.g., making a left turn) and contributing factors (e.g., driver inattention) may result in higher numbers of injuries.

- **Risk of Certain Driver Behaviors**: Driver behaviors such as speeding or failing to yield may appear frequently with certain pre-crash actions, suggesting that these behaviors are particularly dangerous in specific contexts.
- **Patterns in Crash Severity**: Some combinations may reveal patterns in crash severity, suggesting that certain pre-crash actions combined with specific contributing factors consistently lead to more severe injury outcomes.

Conclusions

1. **Target High-Risk Behavior Combinations**: If specific combinations of pre-crash actions and contributing factors (such as making a turn while distracted) are consistently associated with high injury counts, these combinations could be prioritized for safety interventions. For example, public awareness campaigns could emphasize the dangers of distraction during complex maneuvers like turning.

2. **Safety Enhancements at Intersections and High-Risk Areas**: If actions like turning or lane changing are often associated with injuries due to specific contributing factors (such as failing to yield), authorities could focus on improving intersection design, signage, or signals to reduce the likelihood of these incidents.

3. **Inform Policy on Speeding and Driver Attention**: The query results may reveal that high-speed actions combined with driver inattention or aggressive driving result in more injuries. Policymakers could consider enforcing stricter penalties for speeding or aggressive driving in certain areas or under specific conditions.

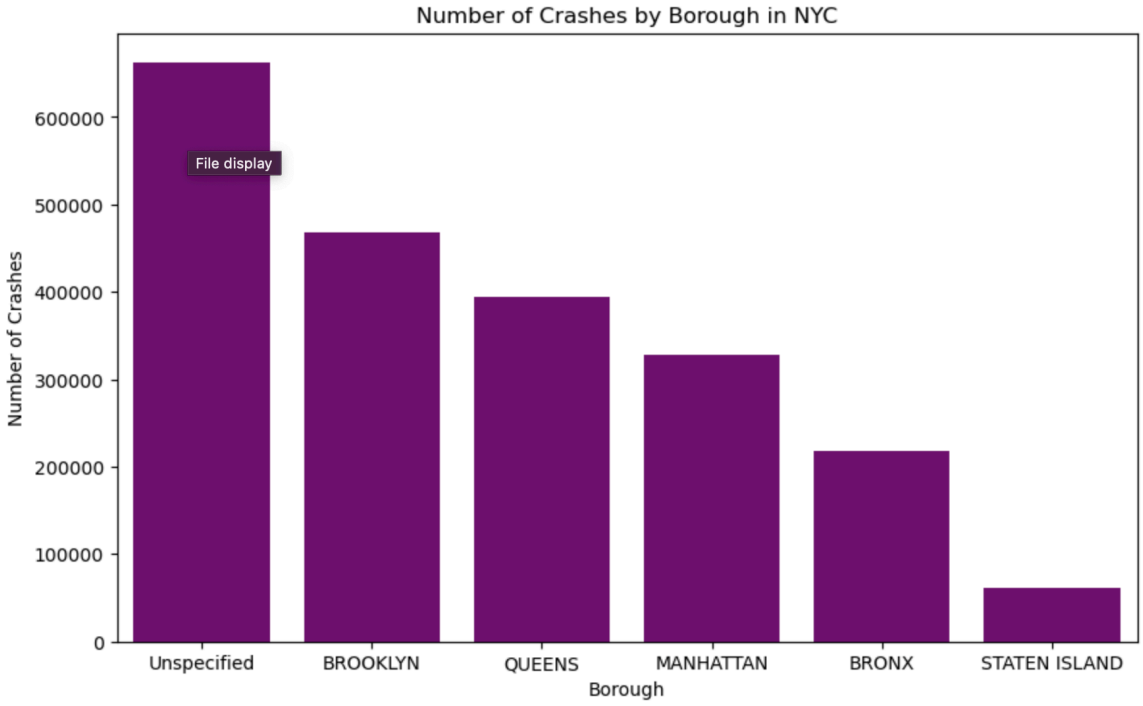
4. **Further Research into High-Injury Combinations**: This query provides a foundation for more detailed analysis, such as exploring crash severity by time of day, weather conditions, or road characteristics. These insights could support targeted interventions in specific environments, such as school zones or high-speed highways.

Python Data Analysis

Using pandas, matplotlib, and seaborn, three visualizations were created to analyze crash patterns and demographics:

1. **Crash Distribution by Borough**

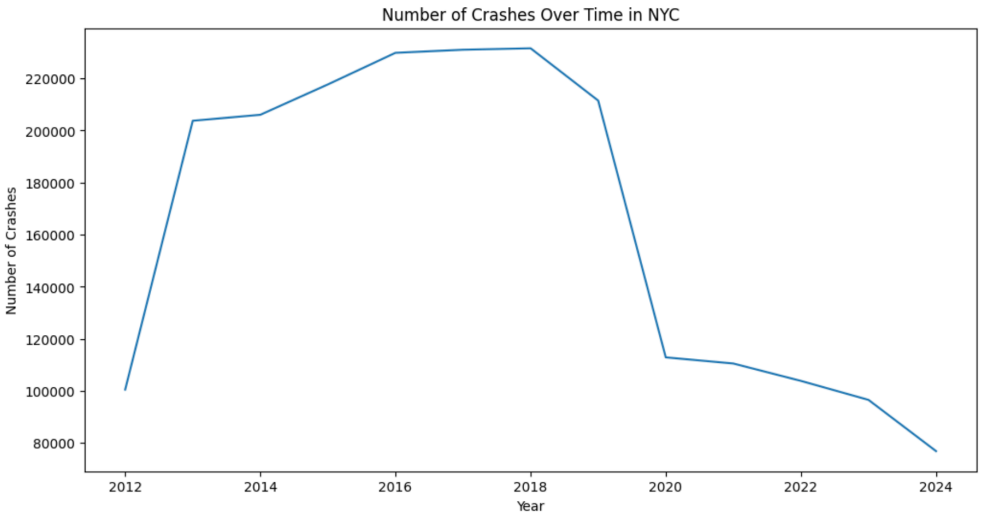
- **Objective**: To identify which boroughs in New York City experience the highest number of motor vehicle crashes.
- **Approach**: Aggregate crash data by borough and calculate the total number of incidents in each. This analysis helps prioritize safety initiatives and resource allocation by pinpointing areas with higher crash rates.
- **Expected Insights**: Boroughs with more traffic congestion or higher population density, such as Brooklyn or Manhattan, might have more crashes. This analysis provides a geographic perspective on crash distribution, guiding local government efforts to implement safety measures in high-risk areas.



This chart indicates which boroughs require more traffic safety measures.

2. Crash Trends Overtime:

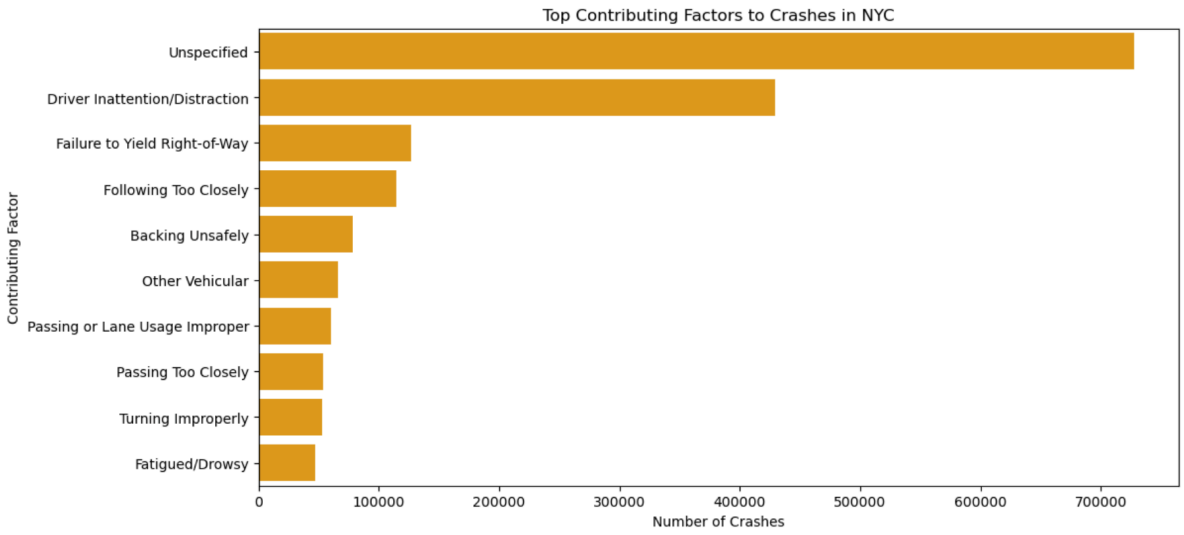
- **Objective**: The number of crashes rose steadily from 2012 to 2016, stabilized around 2016-2019, and then declined sharply starting in 2020, likely due to COVID-19 restrictions reducing traffic.
- **Approach**: Aggregate the number of crashes by year and plot the trend to analyze changes over time, identifying periods of increase, stability, and decline.
- **Expected Insights**: The impact of external events (e.g., COVID-19) on crash frequency. Opportunities to study traffic patterns or safety initiatives that could explain fluctuations in crash numbers.



This graph filters crash records within a specific date range (in this case, the year 2012-2024) and groups them by each year to obtain the crash counts. This allows us to track trends over time, helping identify years with unusually high crash rates.

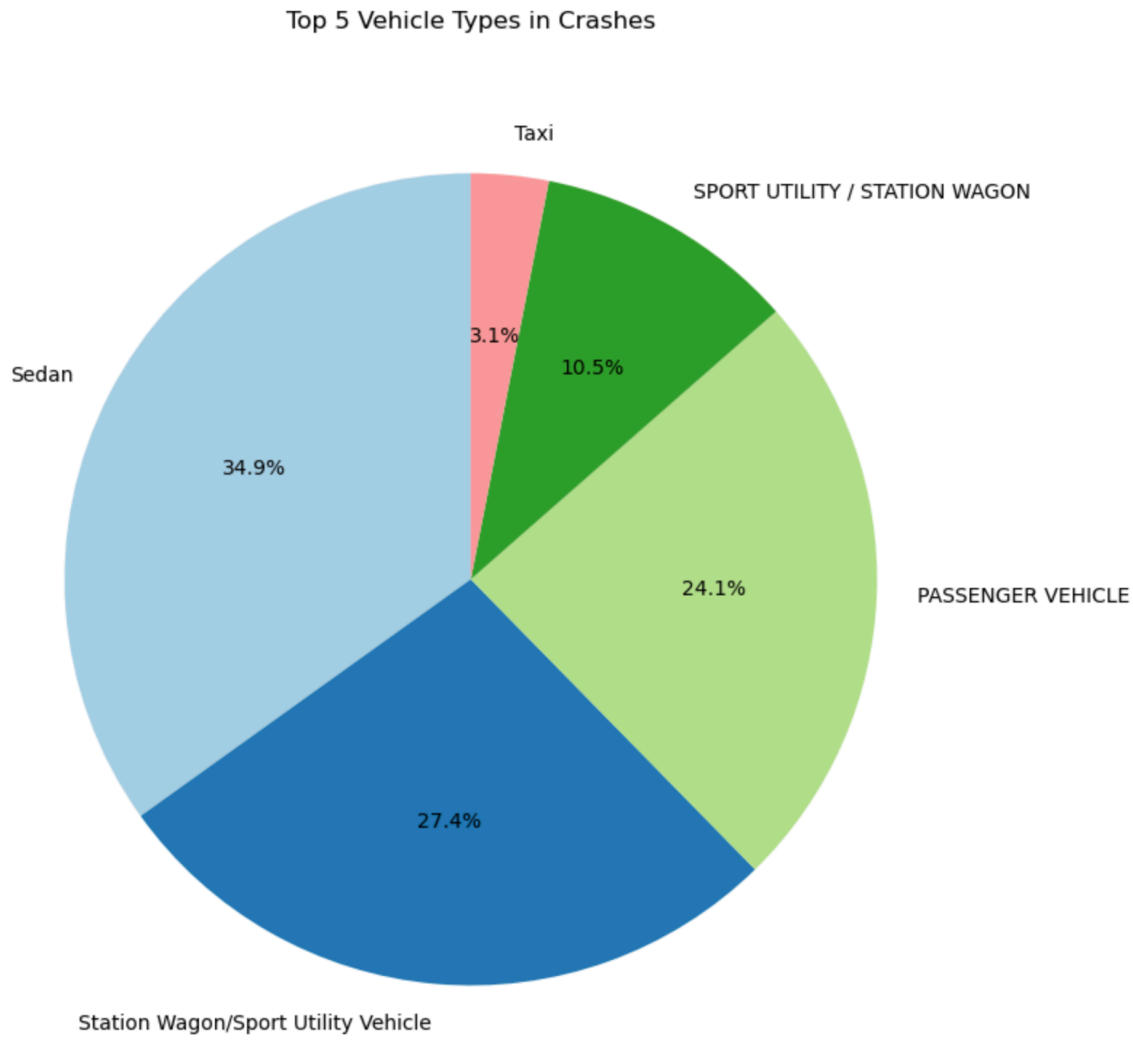
3. Top Contributing Factors to Crashes

- **Objective**: To uncover the primary factors leading to crashes, helping identify behavioral or environmental causes that increase crash likelihood.
- **Approach**: Analyze the contributing factors recorded in the data (e.g., distracted driving, alcohol impairment) and rank them based on frequency. This involves examining the data fields dedicated to crash causes and counting how often each factor appears.
- **Expected Insights**: Common contributing factors could include driver inattention, failure to yield, or adverse weather conditions. Identifying these helps shape policies targeting high-frequency causes, such as campaigns against distracted driving.



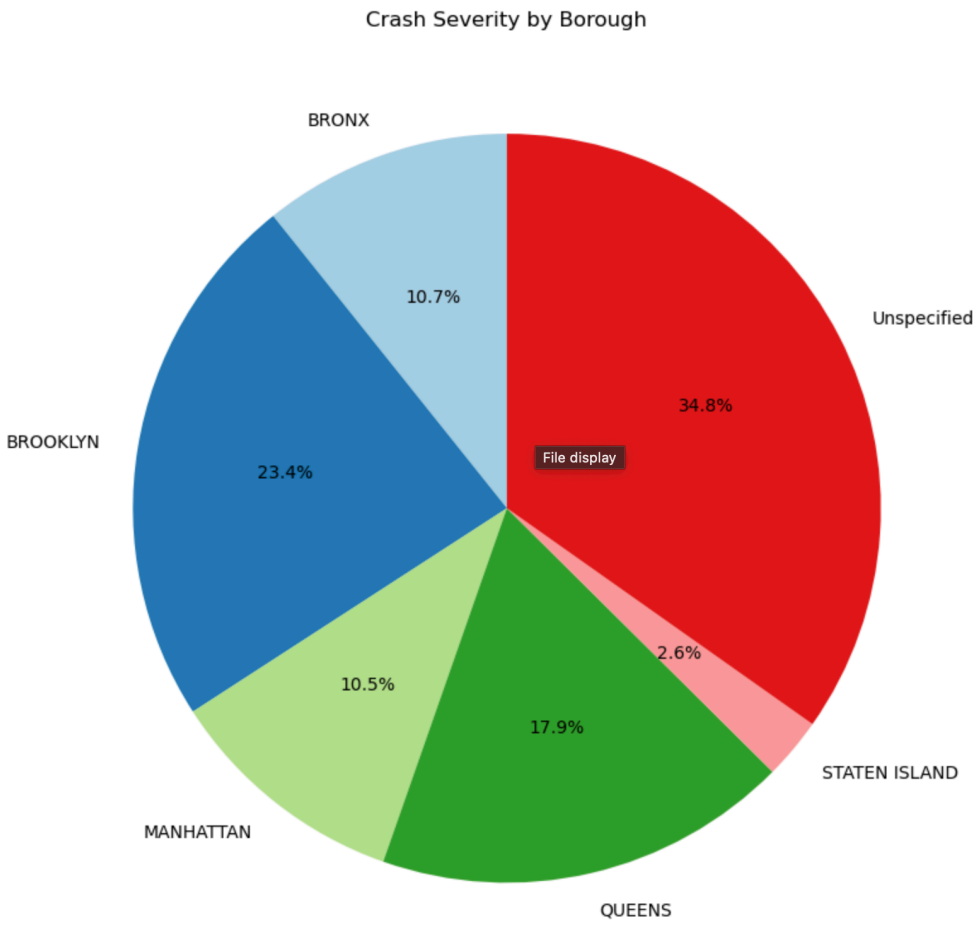
4. Vehicle Type Involvement

- **Objective**: To determine which types of vehicles (e.g., sedans, motorcycles) are most frequently involved in crashes.
- **Approach**: Group crashes by vehicle type and calculate the frequency of each. This analysis may reveal trends in the types of vehicles that tend to be involved in accidents, whether due to their size, maneuverability, or other factors.
- **Expected Insights**: Sedan and Station Wagon/Sport Utility vehicle might be more common due to their popularity, but motorcycles may show a higher risk of severe crashes. Insights could guide recommendations on vehicle-specific safety features or awareness campaigns for drivers of high-risk vehicles.



5. Severity of Crashes

- **Objective:** To identify boroughs with the highest crash frequencies and the impact of missing "Unspecified" data to guide targeted safety measures.
- **Approach:** Calculate and visualize each borough's crash percentage in a pie chart to highlight areas needing traffic management.
- **Expected Insights:** High-crash boroughs (e.g., Brooklyn) may need more safety resources, and the large "Unspecified" segment suggests data gaps.



Version Control

Using Git for collaborative development ensures structured project management:

- We used **Branches** for features and data analysis stages.
- We used **Merges** for changes from collaborators.
- We uploaded **Documentation** of contributions and changes to maintain project clarity and track collaborative efforts.

Observations

Initially, we attempted merging the data with 1,000 entries but found no common elements. We then scaled up to 10,000 and 100,000 entries, discovering only 5 common elements. Consequently, we decided to run the entire dataset, which took more than 3 hours; however, the system crashed. Subsequently, we tried processing 20 million entries due to the substantial size of our dataset, but the system crashed again. Finally, at 10 million entries, we successfully processed 6,000 common entries and proceeded with our analysis.

Limitations of the Project

1.Data Quality and Completeness: The accuracy and completeness of the datasets, including crashes, person, and vehicles, are critical. Missing or inaccurately recorded data on factors such as injury severity, vehicle types, and contributing factors may affect the reliability of insights. Incomplete fields in PRE_CRASH, SAFETY_EQUIPMENT, and CONTRIBUTING_FACTOR_1 could lead to skewed or incomplete conclusions.
2.Scope of Contributing Factors: The available data may not account for all potential contributing factors, such as road conditions, weather, or infrastructure quality, which are crucial for a comprehensive analysis of traffic accidents. Without this context, conclusions might overlook significant external factors.
3.Contextual Factors in Behavior Analysis: Behavioral factors such as DISTRACTED DRIVING or AGGRESSIVE DRIVING are often self-reported or observed after the incident, which can introduce reporting bias. This can affect the analysis of behavior-related factors contributing to crashes.
4.Lack of Individualized Data: Aggregating data at a group level, such as by borough or vehicle type, may overlook unique individual characteristics that could influence crash outcomes, such as driving experience, age, or personal driving habits.
5.Memory consumption: Due to the size of the merged data, the memory consumption is increased which affects the system's efficiency.

Challenges and Solutions

1. **Data Volume**: Managing large datasets led to memory issues. Solution: Employing data chunking and efficient database queries to optimize performance.
2. **Data Quality**: Missing and inconsistent data posed challenges. Solution: Developed data-cleaning scripts to automate preprocessing and improve dataset reliability.

Future Steps

1. **Enhanced Analysis**: Integrating more datasets (e.g., traffic flow or weather data) could provide deeper insights into crash factors.
2. **Predictive Modeling**: Implementing machine learning models could help predict high-risk areas and times, improving traffic safety initiatives.

Conclusions

1.High-Risk Contributing Factors: Certain behaviors, such as distracted driving and speeding, show a clear correlation with severe injuries. This reinforces the importance of targeted enforcement and awareness campaigns to mitigate these behaviors.
2.Significance of Safety Equipment: The analysis supports the role of safety equipment in reducing injury severity. Findings may encourage policymakers to strengthen safety equipment regulations and educate the public on proper equipment use, particularly for high-risk seating positions.
3.Impact of Pre-Crash Actions: Certain pre-crash actions, such as making a left turn or lane changes, emerge as particularly risky, especially when combined with high-risk behaviors. This suggests that infrastructure improvements, such as clearer lane markings and signage, could enhance safety in these scenarios.
4.Identification of High-Injury Locations: Geographic patterns in accident data, such as high injury counts in specific boroughs, can inform regional traffic safety initiatives, optimized patrolling, and infrastructure improvements.