NYC Crash Compass: Navigating Traffic Safety with Data

MGMTMSA 405: DATA MANAGEMENT

GROUP 23
PROJECT 6

BAO, DENNY
KASHI, SHASHANK MANJUNATHA
MA, GUOYAO
WANG, SUMMER
YEUNG, CHI LING (AUDREY)

Table of Contents

Executive Summary

In this project we aim to build a data warehouse and analytical dashboard based on the Motor Vehicle Collisions Crashes dataset from New York City. The dataset, sourced from NYC OpenData, includes NYPD-reported information on motor vehicle collisions.

The project requires loading and analyzing the data in Snowflake, designing a dimensional model, and developing SQL queries to calculate key performance indicators (KPIs) on crashes by area, contributing factors, vehicle types, and injury/fatality rates.

Interactive Tableau dashboards were created to visualize the KPIs across time period and geography. The dashboards provide an analytical tool to derive insights for improving traffic safety, aligning with NYC's Vision Zero initiative.

Challenges included the dataset size and the preliminary nature of the NYPD data. This final report and the Tableau dashboards demonstrate the application of data warehousing and visualization techniques to develop an analytics solution from a real-world dataset.

Some important insights from our analysis are:

- **Crashes by Zip Code and Borough**
  - **High Crash Zone:** The Bronx, particularly zip code 10463, has a significantly higher number of crashes compared to other areas. This highlights a need for targeted safety measures in this specific location.
  - **Borough Level Analysis:** Brooklyn leads in total crashes, but a deeper dive by zip code within boroughs (like focusing on 10463 in the Bronx) can provide more granular insights for resource allocation.

- **Split of Injuries by Contributing Factor**
  - **Data Quality Gap:** "Unspecified" is a leading contributing factor, indicating a need for improved data collection on crash causes. This will enable better targeted interventions.
  - **Other Vehicular Issues:** "Other Vehicular" is a major cause of injuries across all vehicles. Further breakdown of this category (e.g., unsafe lane changes) is crucial to develop specific safety campaigns.

- **Number of People Killed by Contributing Factors**
    - **Driver Inattention:** Distracted driving is the second leading cause of fatalities, highlighting the importance of public awareness campaigns and stricter regulations on phone use while driving.
    - **Speeding and Right-of-Way:** "Unsafe speed" and "failure to yield" are significant factors. This calls for stricter enforcement measures like increased speed checks and targeted campaigns on right-of-way rules.

- **Number of People Injured by Vehicle Combination**
    - **Unknown Vehicle Type 2:** A large portion of crashes involve a passenger vehicle and an unknown second vehicle. Improved data collection on the type of vehicle involved in these crashes is essential for better understanding crash trends.
    - **Two-Wheeled Vehicle Safety:** Two-wheeled vehicles are involved in a significant number of injuries. This highlights the need for safety campaigns targeting scooter, bike, and motorcycle riders, encouraging safe driving practices and protective gear use.

Project Statement

The main goal of our project is to implement a dimensional model in Snowflake, a cloud-based data warehousing platform, to enable efficient analysis of the crash data. The dimensional model consists of fact and dimension tables, allowing for the calculation of KPIs relevant to understanding and improving traffic safety in New York City.

The KPIs are the following:

1. Crashes by area, analyzing the frequency of crashes by borough or zip code over time.
2. Contributing factors, examining the number of injuries and fatalities associated with each contributing factor, such as unsafe speed or slippery pavement.
3. Contributing factors, examining the number of people killed associated with each contributing factor, such as unsafe speed or slippery pavement.
4. Vehicle types, investigating the number of injuries associated with different vehicle type combinations involved in crashes.
5. Injury/fatality rates, calculating the percentage of crashes resulting in injuries or fatalities by borough, zip code, and time period.

To visualize the KPIs, interactive dashboards are developed using Tableau. These dashboards will allow users to look into the crash data across different dimensions, such as geography and time, to uncover patterns and insights. The dashboards will be a useful tool for data-driven decision-making, aligning with New York City's Vision Zero initiative to eliminate traffic fatalities and improve overall road safety.

Data Literacy

The dataset we used in this project pertains to vehicle collisions and crashes in New York City. There are two primary methods of loading data into Snowflake: the first is uploading the data to an S3 bucket and then accessing this bucket within Snowflake to import the data and create a new table. The second method is to load the data directly from the desktop. At first glance, we understand that this dataset includes over 2 million rows and is 417 MB in size. Given the considerable size of the file, we decided to opt for the first method—copying the file to an S3 bucket—before loading it into Snowflake, as this approach is more reliable for handling a large dataset. Moreover, it allows us to take advantage of Snowflake's bulk loading capabilities, which are designed to work efficiently with cloud storage services like S3.

We conducted a data exploration to better understand the dataset. It consists of 29 columns, many of which contain a significant number of null values. We started the data cleaning process by removing some null values. Given that we are provided with both zip codes and locations (i.e., coordinates), we created a mapping from coordinates to zip codes and used it to fill in the missing zip code values based on rows that already had zip code data. We then applied the same process to the borough data. By creating a mapping from coordinates to boroughs, we were able to fill in missing borough values as well. After these steps, approximately 250,000 rows still had empty borough and zip code values. Considering the total number of records is over 2 million, we decided to remove the rows that remained null for the borough and zip code, as they constituted less than 15% of the total dataset. Next, we filled null values in the "NUMBER OF PERSONS INJURED" with the sum of the individual injury columns (i.e., the number of pedestrians injured, plus the number of cyclists injured, plus the number of motorists injured), and we did the same for "NUMBER OF PERSONS KILLED" using the sum of the individual fatality columns (i.e., the number of pedestrians killed, plus the number of cyclists killed, plus the number of motorists killed). This allowed us to measure the impact of the collisions and crashes more precisely afterwards.

The table below shows the column names along with the corresponding number of null values before and after the cleaning process.

Column Name	Number of Null Values (Before Cleaning)	Number of Null Values (After Cleaning)	Number of Null Values (After Handling Null Values)
CRASH DATE	0	0	0
CRASH TIME	0	0	0
BOROUGH	641181	0	0
ZIP CODE	641429	0	0
LATITUDE	232765	231236	0
LONGITUDE	232765	231236	0
LOCATION	232765	231236	0
ON STREET NAME	436626	367300	0
CROSS STREET NAME	777123	572213	0
OFF STREET NAME	1716767	1462597	0
NUMBER OF PERSONS INJURED	18	0	0
NUMBER OF PERSONS KILLED	31	0	0
NUMBER OF PEDESTRIANS INJURED	0	0	0
NUMBER OF PEDESTRIANS KILLED	0	0	0
NUMBER OF CYCLIST INJURED	0	0	0
NUMBER OF CYCLIST KILLED	0	0	0
NUMBER OF MOTORIST INJURED	0	0	0
NUMBER OF MOTORIST KILLED	0	0	0
CONTRIBUTING FACTOR VEHICLE 1	6688	0	0
CONTRIBUTING FACTOR VEHICLE 2	318213	273747	0
CONTRIBUTING FACTOR VEHICLE 3	1914038	1677802	0
CONTRIBUTING FACTOR VEHICLE 4	2027881	1767492	0
CONTRIBUTING FACTOR VEHICLE 5	2052056	1787009	0
COLLISION_ID	0	0	0
VEHICLE TYPE CODE 1	13430	0	0
VEHICLE TYPE CODE 2	391684	337396	0
VEHICLE TYPE CODE 3	1919315	1681423	0
VEHICLE TYPE CODE 4	2029019	1768164	0
VEHICLE TYPE CODE 5	2052328	1787148	0

To better figure out the causes of collisions, it is important to know the types of vehicles involved in the incidents. We identified 1,115 unique entries in the 'Vehicle Type Code 1' column; however, due to the probable manual entry by law enforcement, this column contains numerous entries with ambiguous descriptions. Therefore, we decided to classify these entries into ten distinct categories:

1. **Passenger Vehicle:** This includes passenger vehicles that are for personal use, such as sedans and sport utility vehicles/station wagons etc.
2. **Commercial Vehicle:** This encompasses commercial vehicles such as delivery vehicles operated by companies such as FedEx and USPS etc.
3. **Truck:** This category covers all types of trucks.
4. **Emergency Vehicle:** This includes vehicles like ambulances and fire trucks.
5. **Bus:** This category encompasses all types of buses, including school buses and public transit buses.
6. **Two-Wheeled Vehicle:** This includes two-wheeled vehicles such as bicycles, scooters, motorcycles, etc.
7. **Construction Vehicle:** This includes heavy machinery such as drilling rigs and bulldozers.
8. **Recreational Vehicle:** This covers vehicles used for leisure activities, such as motorhomes, travel trailers, and camping trailers.
9. **Specialized or Miscellaneous Vehicle:** This includes vehicles like street sweepers and cement mixers that do not fit into other categories.
10. **Unknown/Other:** This is for entries that do not fit into any of the categories above.

This categorization was also applied to the 'Vehicle Type Code 2' data. It is important to note that our study focuses exclusively on collisions involving two vehicles, as these represent the majority of cases. Below is a summary of the vehicles categorized under codes 1 and 2.

Vehicle Bucket	Number of Vehicles	
	Vehicle Type Code 1	Vehicle Type Code 2
Passenger Vehicle	1566137	1142125
Commercial Vehicle	48898	51086
Truck	57627	55676
Emergency Vehicle	8966	5324
Bus	33114	28414
Two-Wheeled Vehicle	33302	71236
Construction Vehicle	895	1148
Recreational Vehicle	97	136
Specialized or Miscellaneous Vehicle	1219	1300
Unknown/Other	44175	100589

Finally, we replaced any missing values in “LATITUDE” and “LONGTITUDE” columns with -999.0000 and all null values in other columns with “NA” or “Unknown/Other.” This last step was critical because it allowed us to completely fill out every row in our fact table with the correct corresponding dimension table key values. In SQL, a null does not equal a null, so the join conditions we used to join our dimension tables together to insert the keys into the fact table would have resulted in many missing key values in the fact table had we not replaced the nulls.

Data Dictionary

We have developed a data dictionary to document the structure and content of the dataset. This data dictionary offers a detailed description of the data attributes, as well as data type, maximum character length (where applicable), numeric scale (where applicable), nullability, and a relevant description. Additionally, we specify the tables in which these columns are located. These tables have been identified and generated based on dimensional modeling and the Entity-Relationship (ER) diagram discussed below.

Table	Attribute	Data Type	Character Maximum Length	Numeric Scale	Is Nullable?	Description
COLLISION_FACT	COLLISION_ID	NUMBER	7		NO	Unique identifier for the collision table
COLLISION_FACT	DATE_KEY	NUMBER	9		NO	Unique identifier for the date table
COLLISION_FACT	TIMEOFDAY_KEY	NUMBER	38		NO	Unique identifier for the time of day table
COLLISION_FACT	LOCATION_KEY	NUMBER	38		NO	Unique identifier for the location table
COLLISION_FACT	VEHICLE_KEY	NUMBER	38		NO	Unique identifier for the vehicle table

COLLISION_FACT	CONTRIBUTING_FACTOR_KEY	NUMBER	38		NO	Unique identifier for the contributing factor table
COLLISION_FACT	NUMBER_OF_PERSONS_INJURED	NUMBER			NO	Injured person count
COLLISION_FACT	NUMBER_OF_PERSONS_KILLED	NUMBER			NO	Fatality count
COLLISION_FACT	NUMBER_OF_PEDESTRIANS_INJURED	NUMBER			NO	Injured pedestrian count
COLLISION_FACT	NUMBER_OF_PEDESTRIANS_KILLED	NUMBER			NO	Pedestrian fatality count
COLLISION_FACT	NUMBER_OF_CYCLIST_INJURED	NUMBER			NO	Injured cyclist count
COLLISION_FACT	NUMBER_OF_CYCLIST_KILLED	NUMBER			NO	Cyclist fatality count
COLLISION_FACT	NUMBER_OF_MOTORIST_INJURED	NUMBER			NO	Injured motorist count
COLLISION_FACT	NUMBER_OF_MOTORIST_KILLED	NUMBER			NO	Motorist fatality count
DATE_DIM	DATE_KEY	NUMBER	9		NO	Unique identifier for the date table
DATE_DIM	CRASH_DATE	DATE			NO	Date of the collision

TIMEOFDAY_DIM	TIMEOFDAY_KEY	NUMBER	38		NO	Unique identifier for the time of day table
TIMEOFDAY_DIM	CRASH_TIME	TIME			NO	Time of the collision
TIMEOFDAY_DIM	HOUR	NUMBER	38		NO	Hour number
TIMEOFDAY_DIM	AM_PM	TEXT			NO	AM or PM
TIMEOFDAY_DIM	HOUR_24	NUMBER	2		NO	Hour number from 1 to 24
LOCATION_DIM	LOCATION_KEY	NUMBER	38		NO	Unique identifier for the location table
LOCATION_DIM	BOROUGH	TEXT	255		NO	Borough name
LOCATION_DIM	ZIP_CODE	TEXT	10		NO	Zip code of the collision
LOCATION_DIM	LATITUDE	NUMBER		8	YES	Latitude where the collision occurred
LOCATION_DIM	LONGITUDE	NUMBER		8	YES	Longitude where the collision occurred
LOCATION_DIM	LOCATION	TEXT	255		YES	Geographic coordinate of the collision
LOCATION_DIM	ON_STREET_NAME	TEXT	255		YES	Street where the collision occurred

Dimensional Modeling & ER Diagram

To define the fact and dimension tables, we followed the four steps as learned in class. The first step is selecting the business process. For this project, our aim is to analyze vehicle collisions and crashes in New York City, which includes details about the crash date and time, types of vehicles involved, and factors that contributed to the crashes, among others. Next, we declared the grain. The grain of the fact table is a single record of a collision event. Each row in the fact table represents one unique collision incident, including all measurable quantities and associated details. The third step is identifying the dimensions. We developed the dimension tables based on the Key Performance Indicators (KPIs), as illustrated in the table below.

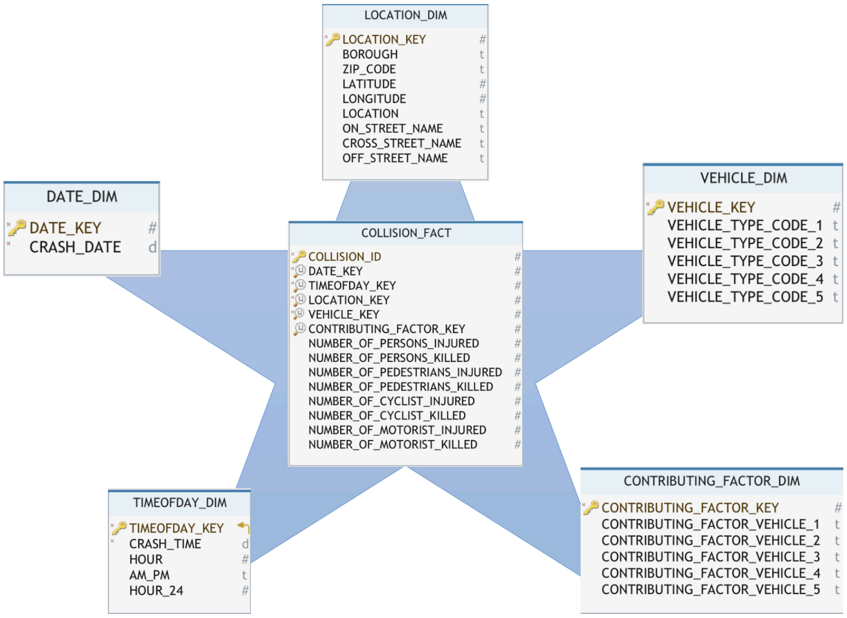
KPIs	Dimension Tables Required
Crashes by area on a given period of time	Location dimension
	Date dimension
	Time of day dimension

Number of people injured by contributing factor	Contributing factor dimension
Number of people killed by contributing factor	Contributing factor dimension
Injuries/deaths per type of vehicle in the crash	Vehicle dimension
Injuries/deaths vs crashes rate	Location dimension

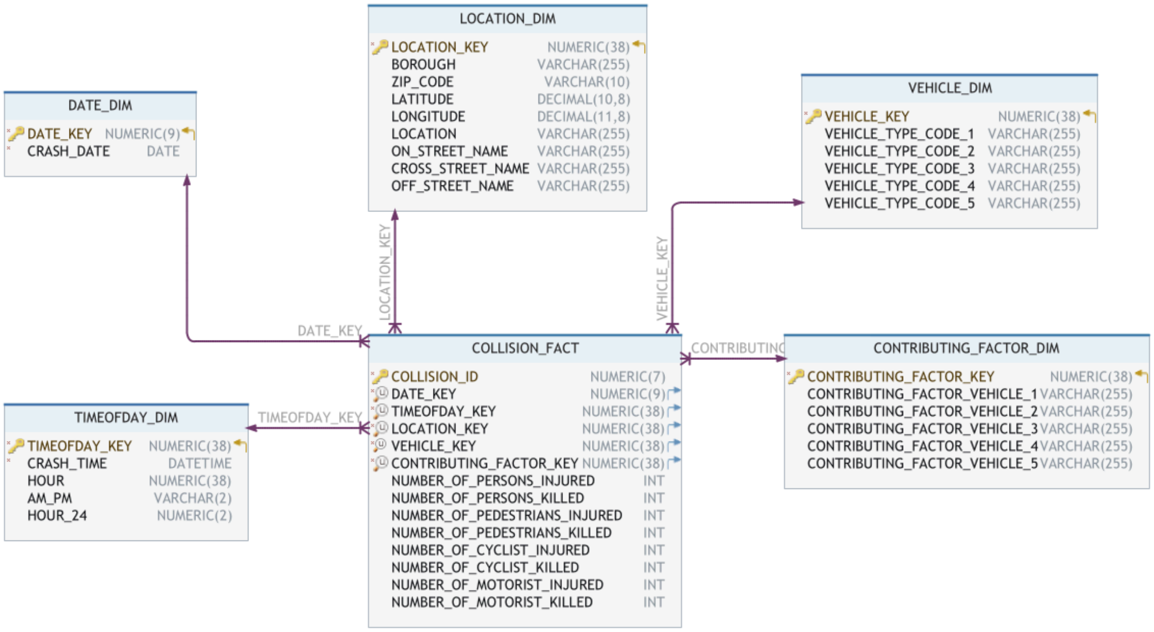
The last step is to identify the fact table. It should contain measurable, numeric information related to each collision, thus, we created a collision fact table including the following columns:

- Collision ID
- 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

The fact and dimension tables are structured in a star schema configuration:



The Entity-Relationship (ER) diagram is as follows:



LOCATION_DIM
LOCATION_KEY NUMERIC(38)
BOROUGH VARCHAR(255)
ZIP_CODE VARCHAR(10)
LATITUDE DECIMAL(10,8)
LONGITUDE DECIMAL(11,8)
LOCATION VARCHAR(255)
ON_STREET_NAME VARCHAR(255)
CROSS_STREET_NAME VARCHAR(255)
OFF_STREET_NAME VARCHAR(255)
DATE_DIM
DATE_KEY NUMERIC(9)
CRASH_DATE DATE
VEHICLE_DIM
VEHICLE_KEY NUMERIC(38)
VEHICLE_TYPE_CODE_1 VARCHAR(255)
VEHICLE_TYPE_CODE_2 VARCHAR(255)
VEHICLE_TYPE_CODE_3 VARCHAR(255)
VEHICLE_TYPE_CODE_4 VARCHAR(255)
VEHICLE_TYPE_CODE_5 VARCHAR(255)
LOCATION_KEY
VEHICLE_KEY
DATE_KEY
CONTRIBUTING
COLLISION_FACT
COLLISION_ID NUMERIC(7)
DATE_KEY NUMERIC(9)
TIMEOFDAY_KEY NUMERIC(38)
LOCATION_KEY NUMERIC(38)
VEHICLE_KEY NUMERIC(38)
CONTRIBUTING_FACTOR_KEY NUMERIC(38)
NUMBER_OF_PERSONS_INJURED INT
NUMBER_OF_PERSONS_KILLED INT
NUMBER_OF_PEDESTRIANS_INJURED INT
NUMBER_OF_PEDESTRIANS_KILLED INT
NUMBER_OF_CYCLIST_INJURED INT
NUMBER_OF_CYCLIST_KILLED INT
NUMBER_OF_MOTORIST_INJURED INT
NUMBER_OF_MOTORIST_KILLED INT
CONTRIBUTING_FACTOR_DIM
CONTRIBUTING_FACTOR_KEY NUMERIC(38)
CONTRIBUTING_FACTOR_VEHICLE_1 VARCHAR(255)
CONTRIBUTING_FACTOR_VEHICLE_2 VARCHAR(255)
CONTRIBUTING_FACTOR_VEHICLE_3 VARCHAR(255)
CONTRIBUTING_FACTOR_VEHICLE_4 VARCHAR(255)
CONTRIBUTING_FACTOR_VEHICLE_5 VARCHAR(255)
TIMEOFDAY_DIM
TIMEOFDAY_KEY
TIMEOFDAY_KEY NUMERIC(38)
CRASH_TIME DATETIME
HOUR NUMERIC(38)
AM_PM VARCHAR(2)
HOUR_24 NUMERIC(2)

Data Transformation

We created the following dimension tables based on the full dataset:

- **DATE_DIM:** We added an auto-generated number as the primary key for this table, named "date_key," and obtained the "crash_date" from the dataset.
- **TIMEOFDAY_DIM:** We added an auto-generated number as the primary key for this table, named "timeofday_key," obtained the "crash_time" from the dataset, and added columns such as “hour”, “AM_PM”, “HOUR_24” etc.
- **LOCATION_DIM:** We added an auto-generated number as the primary key for this table, named "location_key," and included "borough," "zip_code," "latitude," "longitude," "location," "on_street_name," "cross_street_name," and "off_street_name" columns from the dataset in this table.
- **VEHICLE_DIM:** We added an auto-generated number as the primary key for this table, named "vehicle_key," and included "vehicle_type_code_1" to "vehicle_type_code_5" columns from the dataset in this table.
- **CONTRIBUTING_FACTOR_DIM:** We added an auto-generated number as the primary key for this table, named "contributing_factor_key," and included "contributing_factor_vehicle_1" to "contributing_factor_vehicle_5" columns from the dataset in this table.

Next, we created the fact table. As mentioned above, we included "collision_id" and traffic safety metrics, such as counts of injuries and fatalities, in the fact table. Then, we joined it with the five dimension tables using the conditions below:

- Joined with **DATE_DIM** on "crash_date" and “date”
- Joined with **TIMEOFDAY_DIM** on “crash_time” and "HOUR_24"
- Joined with **LOCATION_DIM** on "borough," "zip_code," "latitude," "longitude," "location," "on_street_name," "cross_street_name," and "off_street_name"
- Joined with **VEHICLE_DIM** on "vehicle_type_code_1" to "vehicle_type_code_5"
- Joined with **CONTRIBUTING_FACTOR_DIM** on "contributing_factor_vehicle_1" to "contributing_factor_vehicle_5"

By doing this, the fact table is joined with all five dimension tables, providing sufficient data to construct KPIs.

KPIs and Visualizations

The first KPI measures the number of crashes by zip code and borough. The top 3 zip codes with the highest number of crashes are 10463, 11207,11236, with 903,893; 124,611; 83,108 crashes, respectively. Particularly, the area with zip code 10463, which is primarily located in Bronx County, NY, has a markedly high number of crashes compared to other areas. The second bar chart reveals that Brooklyn leads in the number of crashes with over 2 million incidents, followed by the Bronx with over 1.8 million. For a more in-depth analysis, we can examine other metrics such as contributing factors to explore the primary causes of these crashes in the Bronx and Brooklyn and develop targeted strategies to mitigate them.

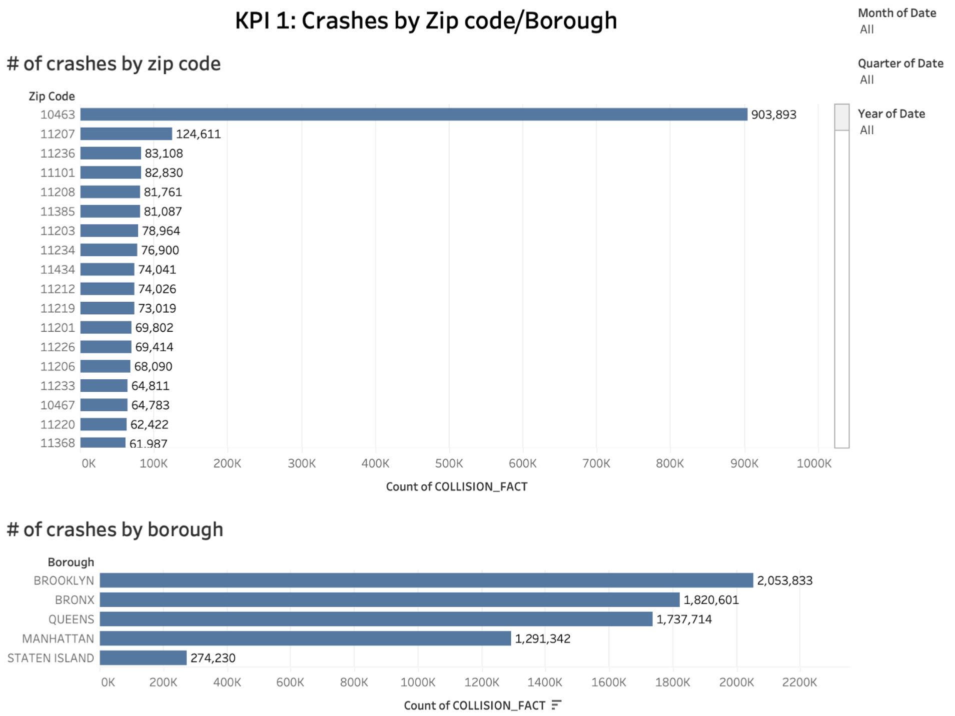


Figure 1

For the split of number of people injured split by the contributing factor (KPI 2), we see that the most common contributing factor across all vehicles are unspecified, other vehicular, following too closely, and driver inattention. As we do not have an accurate idea of the contributing factor data point for most collisions, we might have to pay more attention to gathering this datapoint in greater detail to come up with remedial action for these causes.

- For Vehicle 1 contributing factor, the leading factor to causing injuries is unspecified (28,208), followed by failure to yield (2,176) and traffic control (644,068).
- For Vehicle 2, ‘Other vehicular’ is the leading contributing factor causing injuries (512,360) followed by failure to yield (108,068) and traffic control (3.312).
- For Vehicle 3, ‘Other vehicular’ is again the leading contributing factor causing injuries (1,831,284), followed by following too closely (277,400) and pavement slippery (83,900).
- For Vehicle 4 too, ‘Other vehicular' is the leading contributing factor causing injuries (2,039,916), followed by driver inattention (83,900) and fatigued/drowsy (83,900).

Overall: Other vehicular seems to be the leading contributing factor causing injuries across all the vehicles. This could be due to various reasons such as unsafe lane changes or not following the rules of the road.

However, it is important to note that this data only shows the number of people injured and doesn't provide any details on the severity of the injuries. Additionally, the contributing factors may not be mutually exclusive, so one crash may have multiple contributing factors.

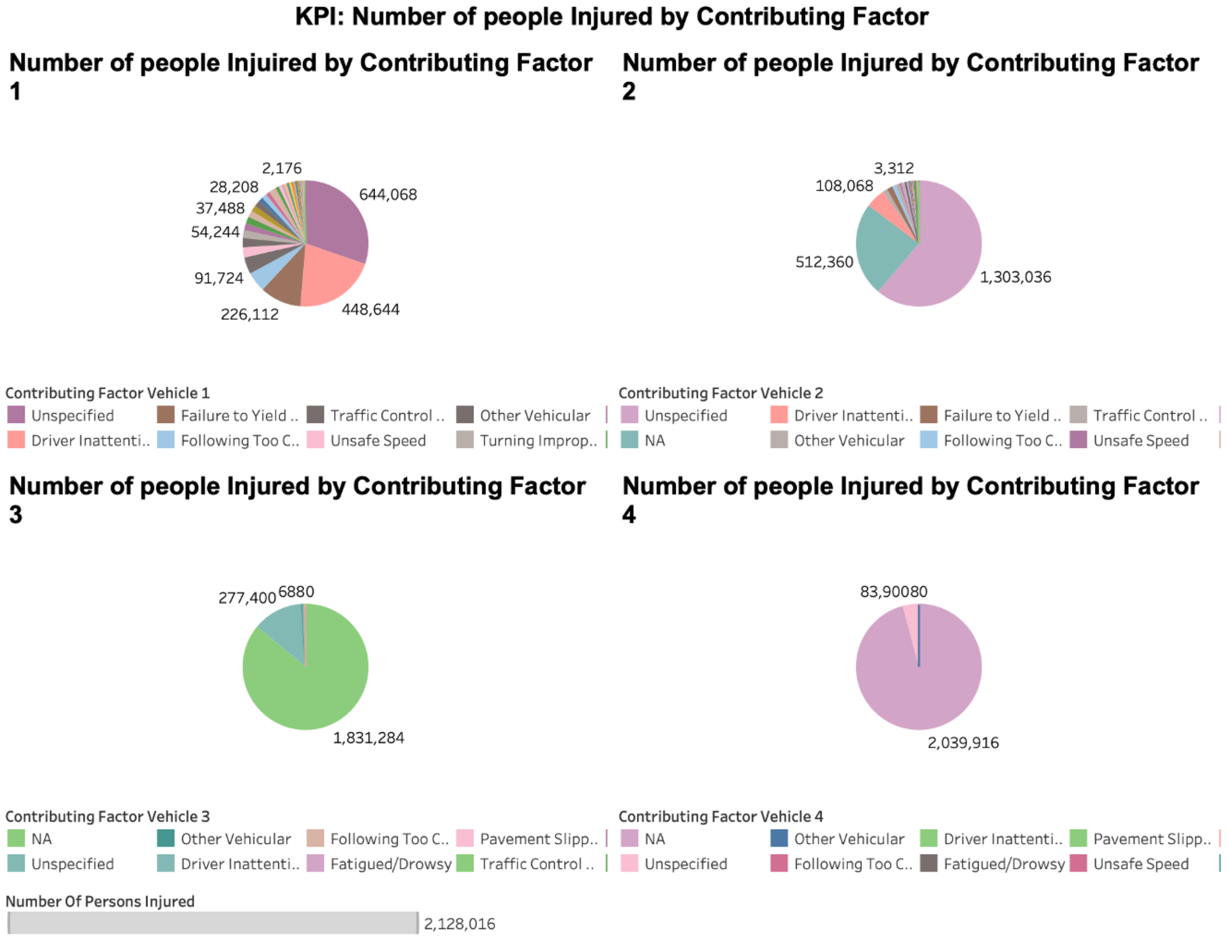


Figure 2

KPI 3 details the number of fatalities attributable to contributing factors, with a focus on factors 1 and 2 due to the relative completeness of these data sets. Out of the 10,084 total fatalities recorded, a substantial proportion remains categorized under 'unspecified' reasons. However, it's evident that “driver inattention/distraction” is the second most common cause of fatalities, underscoring the need for initiatives that enhance driver concentration on the road. “Unsafe speed” comes in as the next considerable factor, with “failure to yield” and “disregard of traffic control” also contributing to the death toll. To tackle these issues, a multifaceted approach is necessary. Public campaigns and stricter regulations should target the reduction of driver distraction, with a particular emphasis on minimizing mobile device usage through public service announcements and the enforcement of hands-free technology laws. Additionally, to deal with the dangers of driving too fast, we may need to enforce speed limits better, by doing more speed checks and putting speed cameras in places where accidents happen often.

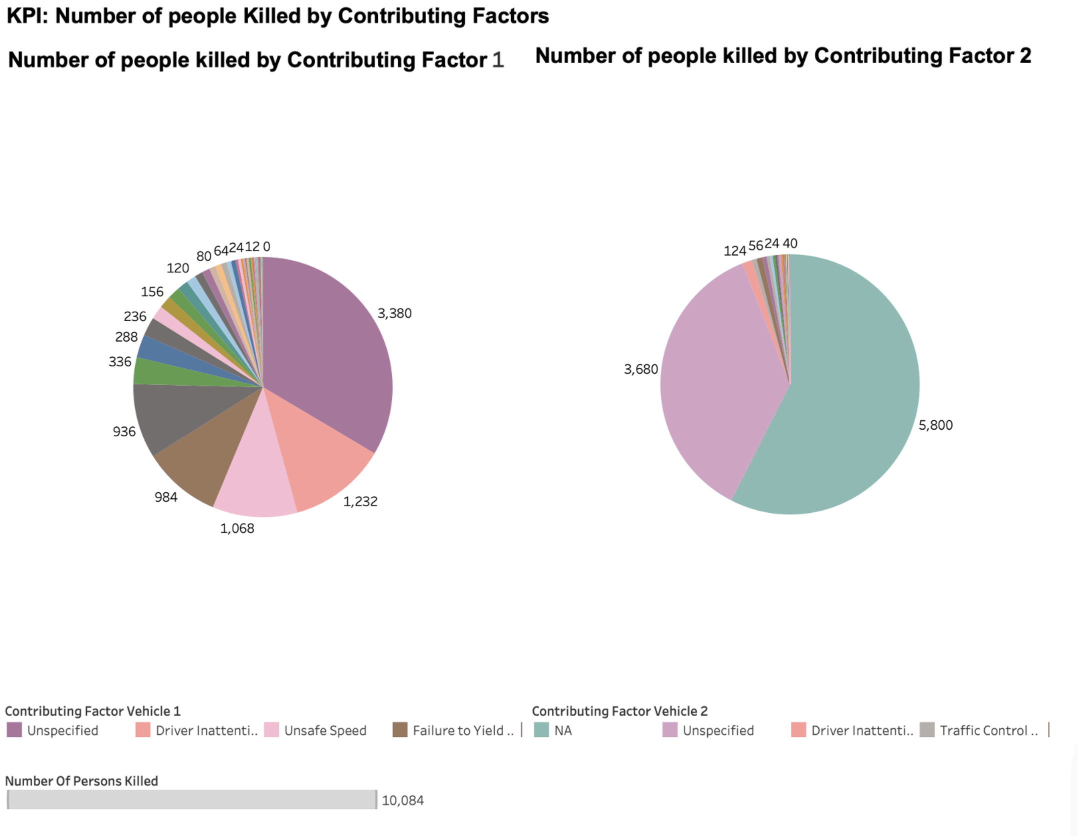


Figure 3

For KPI 4, the vast majority of injuries stemmed from crashes that involved a passenger vehicle. This is expected, since passenger vehicles make up the majority of vehicles on the road. Interestingly, the second most common combination of vehicle combinations in a crash are passenger vehicles and unknown/other. This means that many crashes involve a passenger vehicle only or the type of the other involved vehicle is often not reported. Since this makes up such a large percentage of the injuries, identifying the actual vehicle type 2 involved in these crashes would allow us to gain much more accurate insights into crash trends. Also, the visualization shows that two-wheeled vehicles are involved in the second highest number of injuries. This category can include vehicle types such as scooters, bikes, and motorcycles. This insight can be used to encourage operators of two-wheeled vehicles to practice safer driving and to wear protective equipment.

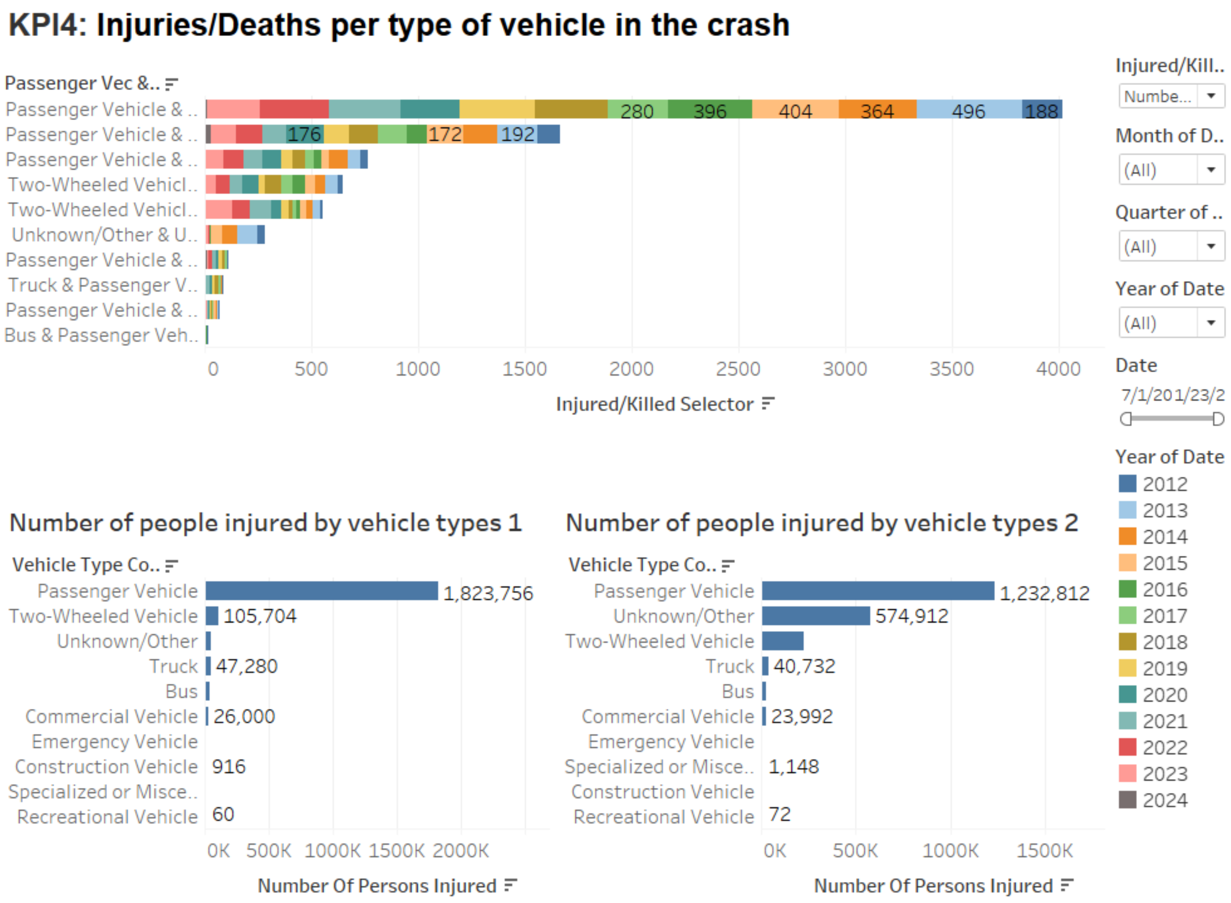
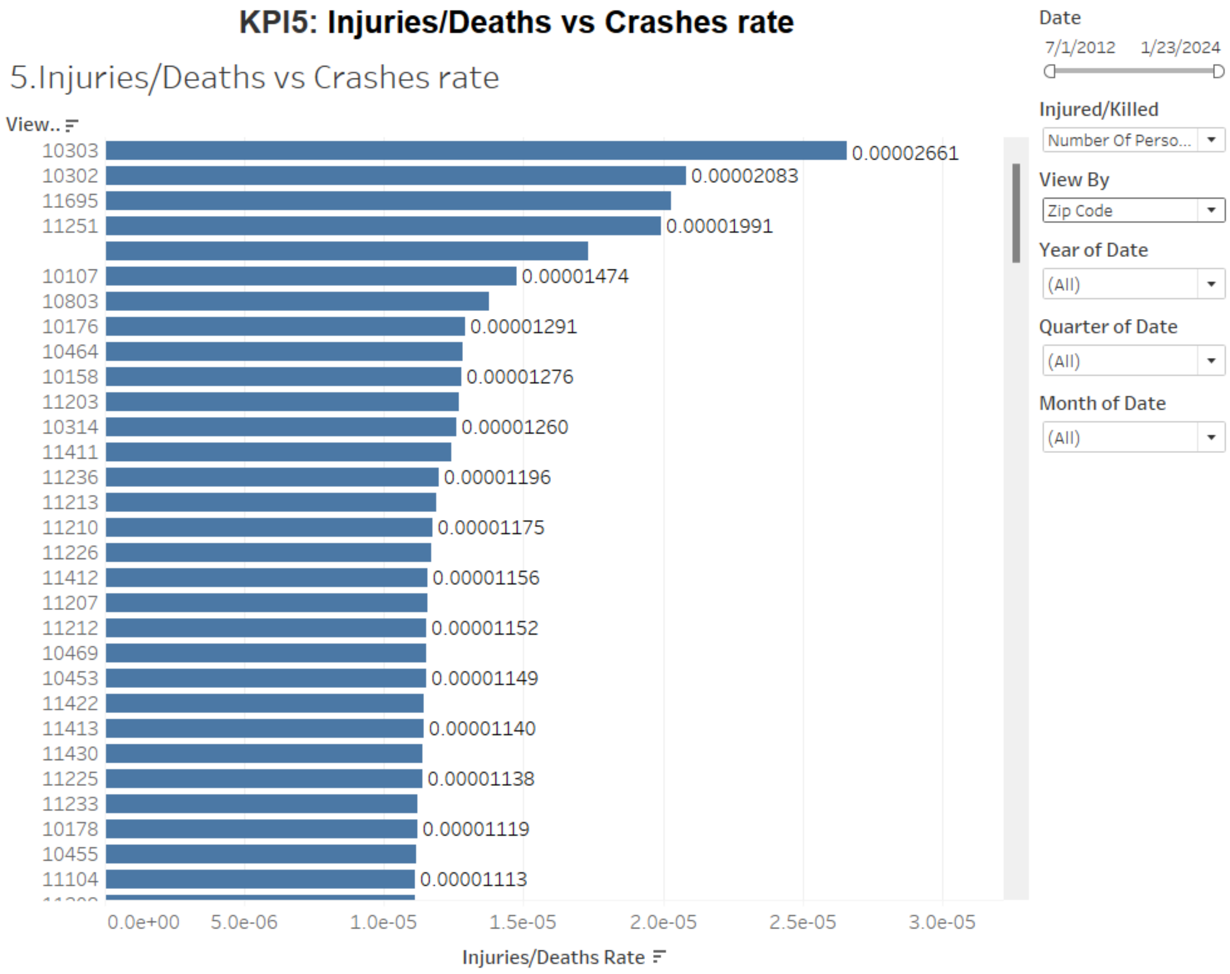


Figure 4

The distinct patterns revealed by the "Injuries/Deaths vs. Crashes Rate" KPI highlight Staten Island and Queens as areas of particular concern for both injuries and fatalities, with certain zip codes like 10303 and 10302 standing out. This data may suggest localized issues, potentially involving road design, traffic flow, or specific behaviors prevalent in these neighborhoods. The high injury rates in Brooklyn and Bronx, along with Staten Island, emphasize the need for comprehensive reviews of street safety, especially considering the Vision Zero goals to reduce traffic-related deaths. The uniformity in death rates across Queens, Bronx, and Staten Island implies that fatal accidents, while less frequent, occur at a consistent rate, pointing to systematic risks that affect these boroughs equally.

To address these findings, a two-fold strategy is essential. First, interventions must be tailored to the unique challenges of the highlighted areas: in neighborhoods with high injury rates, this might include increasing pedestrian safety measures and traffic calming efforts, whereas areas with high fatality rates might benefit from aggressive speed enforcement and improved protection for vulnerable road users. Second, enhancing data collection efforts is imperative for pinpointing contributing factors more precisely, which will allow for more targeted and effective solutions. With a clearer understanding of the data, NYPD and traffic safety agencies can focus on implementing and monitoring interventions that will not only serve immediate safety needs but also contribute to the long-term success of initiatives aimed at eliminating traffic fatalities and severe injuries citywide.



Date
7/1/2012 1/23/2024

Injured/Killed
Number Of Perso...

View By
Zip Code

Year of Date
(All)

Quarter of Date
(All)

Month of Date
(All)

Project Challenges

We found that most of the challenges we faced had more to do with the data cleaning, preprocessing, and transformation rather than the actual data visualization.

Data Entry Standardization
One of the main challenges we faced during the project was generating the “Injuries/deaths per type of vehicle in the crash" KPI. The main obstacle was that the values found in the “Vehicle Type” columns were not standardized. The values seemed to be from open-ended manual data entries by law enforcement. Some values were incoherent or incomplete. To illustrate, the “vehicle_type_code_1” column has 1,115 unique values alone. In order to better summarize and visualize the data, we grouped the vehicles into 10 broad categories. This allowed us to see the injuries/deaths per type of vehicle in a more concise manner.

Dimensionality
Another big challenge was deciding how to handle the dimensionality and high number of null values in the vehicle and contributing factor columns. Since the raw data had five vehicle type columns, for example, if a crash only involved 2 vehicles, only the first 2 columns would be populated. As a result, vehicle_type_code_3, vehicle_type_code_4, and vehicle_type_code_5 columns contained mostly null values. We decided on only visualizing data from vehicle_type_code_1 and vehicle_type_code_2, as well as contributing_factor_vehicle_1 and contributing_factor_vehicle_2 since these two columns were mostly complete. Also, including all 5 columns in our visualizations would result in too many dimensions and would make the visualizations too difficult to understand. We felt that just including the first 2 vehicle columns in our visualizations was sufficient in showing any trends.

Missing Zip Code and Borough Data
An important challenge that we encountered but resolved relatively easily was the missing zip code and borough data. The raw dataset had over 640,000 missing values in each of these columns. To resolve this issue and ensure data integrity, mapped coordinates to zip code and borough using the existing data, and used this mapping to fill out over 400,000 of the missing values in each column.

Handling Remaining Null Values
Finally, when creating the fact table, we realized that the presence of null values in our dimension tables prevented us from filling out the fact table with the dimension table keys. This is because the null values were not captured in our join statement conditions. To resolve this, we replaced all remaining null values with either "NA" or "Unknown/Other" depending on the context of the corresponding column.

Note
To show how we resolved these issues more clearly, we included the Python notebook we used in our project submission.

Project Limitations

As mentioned above, one of the project limitations we encountered was the high amount of missing data. This prevented us from including vehicle type code 3-5 and contributing factor vehicle 3-5 in our visualizations. We also had to remove about 12% of the rows from the raw data set due to missing zip code and borough data. As a result, our analysis may be incomplete or skewed.

Another project limitation was that the dataset did not include personal, demographic, or socioeconomic data. It also did not include contextual data of the crashes, such as the amount of traffic or the weather conditions. These factors can all contribute to motor vehicle crashes, and including them in the dataset could give us a more holistic view and help us better understand the underlying factors that influence collision rates.

Finally, it is possible that not all motor vehicle collisions that occurred in New York City were recorded in this dataset, especially those that did not include any injuries or deaths. This underreporting can lead to an inaccurate data analysis and an underestimation of the number of collisions.

Appendix

1. **Finals_2024_Group23_Project6_DataCleaning.ipynb**: Python notebook used to clean the NYC Motor Vehicle Collision dataset.
2. **Finals_2024_Group23_Project6_report.sql**: SQL code used to upload the dataset from S3 bucket into Snowflake and create the dimension and fact tables.
3. **Finals_2024_Group23_Project6_Queries.sql**: Sample SQL queries for the 5 required KPIs.