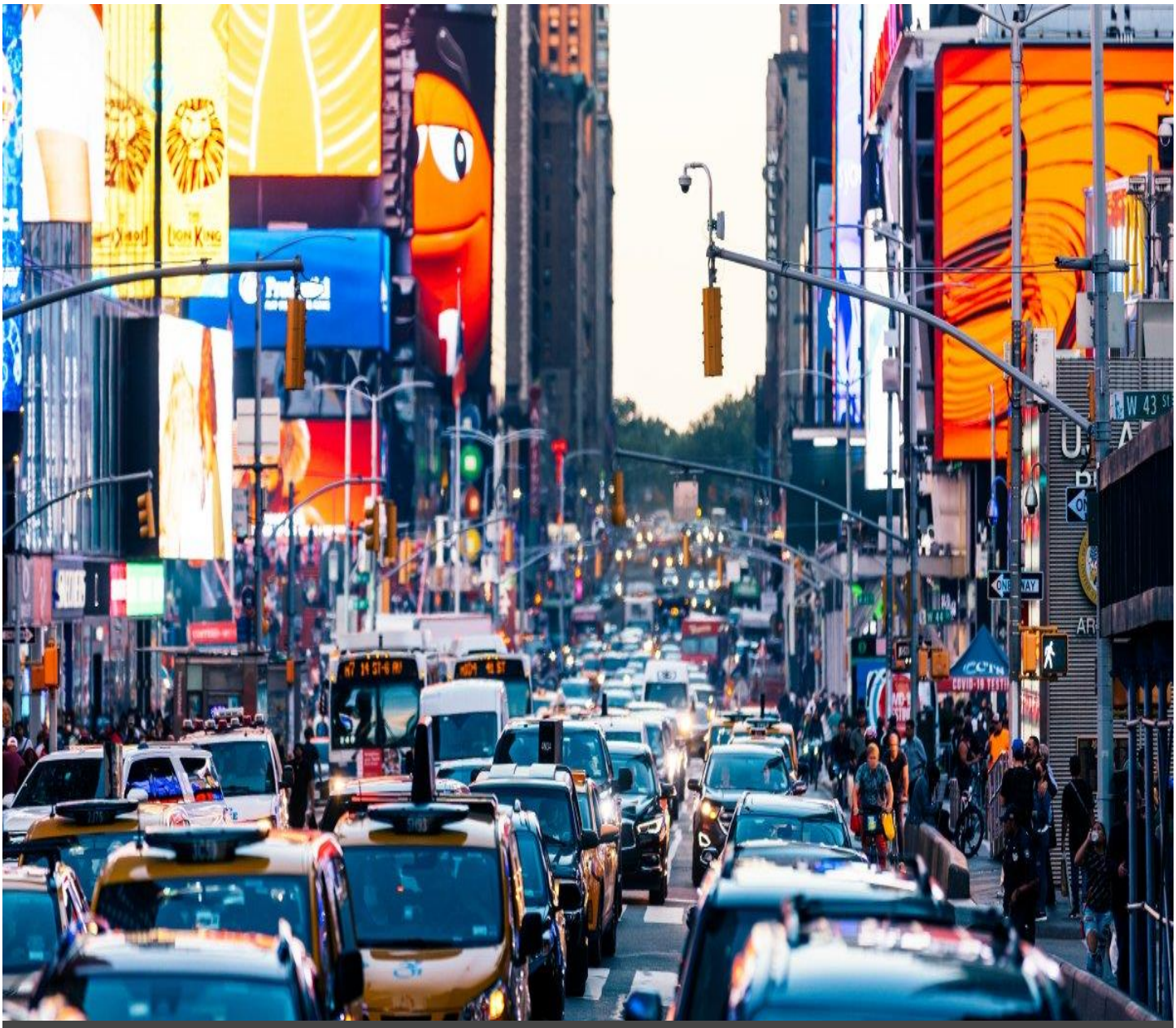


NY Motor Vehicle Collision Analytics Final Project

UCLA Anderson School of Management – Group 12

Nicole Lai, Stephanie Chen, Armaan Dhanda, Hamza Shahab Shafqat

MGMTMSA405 Data Management

Table of Contents

Table of Contents 1

Executive Summary 2

Project Statement 3

Data Dictionary 4

Dimensional Modelling and ERD 6

- Step 1: Define the business or event process 6
- Step 2: Define the grain 6
- Step 3: Identify the Dimension tables 6
- Step 4: Define the Fact table 7

Data Transformation 8

Produce KPIs Using FACT and DIM Tables 9

Data Visualization 10

Project Challenges 17

Appendix 19

Executive Summary

This comprehensive project aimed at analyzing and understanding the patterns, causes, and effects of motor vehicle collisions in New York from 2012 to the present. At the core of this initiative is the development of a sophisticated data warehouse designed to integrate and meticulously analyze collision data, with a particular focus on key performance indicators (KPIs) such as the geographical distribution of crashes, the number of people injured or killed by contributing factors, and the correlation between vehicle types and collision outcomes.

To accomplish this, the project delineates a series of methodical steps including the cleaning and incorporation of the collision dataset into the Snowflake platform, the creation of a data dictionary to standardize terms and measures, the construction of a dimensional model to facilitate efficient data querying and analysis, and the employment of Tableau to generate interactive dashboards. These dashboards provide stakeholders with actionable insights, underpinning data-driven decision-making processes within New York's transportation and public safety sectors. The overarching goal is to leverage these insights to forge safer roadways and devise more effective traffic management strategies, thereby curtailing the prevalence and severity of motor vehicle collisions.

Furthermore, the analysis acknowledges the police reporting requirement for collisions resulting in injury, death, or substantial property damage (exceeding $1000), while also suggesting that minor incidents, often excluded from reports, may contribute to a broader understanding of distracted driving behaviors. The study points to the use of smartphones while driving—especially in high-traffic areas and by drivers of high-impact vehicles like SUVs and station wagons—as a significant factor in the increasing trend of vehicular collisions.

In conclusion, the project not only sheds light on the current state of motor vehicle collisions in New York but also offers strategic recommendations for the government and relevant stakeholders. These include the enactment of stricter law enforcement measures, such as elevated fines and more severe penalties, coupled with educational campaigns emphasizing the importance of undivided attention while driving. Together, these recommendations aim to mitigate the rising trend of vehicle collisions, enhancing road safety for all New Yorkers.

Project Statement

In the bustling streets of New York, motor vehicle collisions pose significant challenges to public safety, traffic management, and emergency response services. With an ever-increasing number of vehicles on the road, understanding the dynamics of these collisions becomes crucial for city planners, policymakers, and the general public. **The goal of this project is to harness the power of data analytics to provide insights into the patterns, causes, and effects of motor vehicle collisions in New York from 2012 to the present**.

To achieve this, we propose the development of a comprehensive data warehouse that integrates detailed collision data, **focusing on key performance indicators (KPIs)** including:

- Crashes by Area
- Number of People Injured by Contributing Factor
- Number of People Killed by Contributing Factor
- Injuries/Deaths per Type of Vehicle in the Crash
- Injuries/Death vs. Crashes Rate

This data warehouse will be the foundation for advanced analytics and visualization, enabling stakeholders to make informed decisions based on empirical evidence.

The project involves several key steps:

1. Cleaning and loading the collision data set into Snowflake
2. Creating a data dictionary
3. Building a dimensional model to support efficient querying and analysis
4. Using Tableau to build interactive dashboards for KPIs

This project not only aims to enhance our understanding of motor vehicle collisions in New York but also to help support data-driven decision-making within the city's transportation and safety departments. By providing access to actionable insights, we hope to contribute to developing safer roads and more effective traffic management strategies, ultimately reducing the frequency and severity of motor vehicle collisions in New York.

Data Dictionary

The NYC Motor Vehicle Collisions data set (publicly available on the city of New York's website under public safety data) contains information from all police-reported motor vehicle collisions in NYC. **The police report is required to be filled out for collisions where someone is injured or killed, or where there is at least $1000 worth of damage.** The dataset has **only one table (represented as NY_CRASHES**) that contains details on the crash events, and each row represents a crash event. The table has crash event data from 1st July 2012 to 23rd January 2024. The following tables contain metadata and column information respectively for the NY_CRASHES data:

TABLE NAME	INITIAL ROW COUNT	FINAL ROW COUNT	SIZE
NY_CRASHES	2,061,020	1,994,818	493,931 KB

COLUMN NAME	ORDINAL POSITION	DATA TYPE	NUMERIC SCALE	NULLABLE	COLUMN DESCRIPTION	PRIMARY KEY
CRASH_DATE	1	DATE		NO	Occurrence date of collision	
CRASH_TIME	2	TIME		NO	Occurrence time of collision	
BOROUGH	3	TEXT		NO	Borough where collision occurred	
ZIP_CODE	4	NUMBER	0	NO	Postal code of incident occurrence	
LATITUDE	5	NUMBER	6	NO	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)	
LONGITUDE	6	NUMBER	6	NO	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)	
ON_STREET_NAME	7	TEXT		YES	Street on which the collision occurred	
CROSS_STREET_NAME	8	TEXT		YES	Nearest cross street to the collision	
OFF_STREET_NAME	9	TEXT		YES	Street address if known	

NUM_PERSONS_INJURED	10	NUMBER	0	YES	Number of persons injured	
NUM_PERSONS_KILLED	11	NUMBER	0	YES	Number of persons killed	
NUM_PEDEST_INJURED	12	NUMBER	0	YES	Number of pedestrians injured	
NUM_PEDEST_KILLED	13	NUMBER	0	YES	Number of pedestrians killed	
NUM_CYCL_INJURED	14	NUMBER	0	YES	Number of cyclists injured	
NUM_CYCL_KILLED	15	NUMBER	0	YES	Number of cyclists killed	
NUM_MOTOR_INJURED	16	NUMBER	0	YES	Number of vehicle occupants injured	
NUM_MOTOR_KILLED	17	NUMBER	0	YES	Number of vehicle occupants killed	
CONT_FACTOR_VEH_1	18	TEXT		YES	Factors contributing to the collision for designated vehicle	
CONT_FACTOR_VEH_2	19	TEXT		YES	Factors contributing to the collision for designated vehicle	
CONT_FACTOR_VEH_3	20	TEXT		YES	Factors contributing to the collision for designated vehicle	
CONT_FACTOR_VEH_4	21	TEXT		YES	Factors contributing to the collision for designated vehicle	
CONT_FACTOR_VEH_5	22	TEXT		YES	Factors contributing to the collision for designated vehicle	
COLLISION_ID	23	NUMBER	0	NO	Unique record code generated by system	✓
VEH_TYPE_CODE_1	24	TEXT		YES	Type of vehicle	
VEH_TYPE_CODE_2	25	TEXT		YES	Type of vehicle	
VEH_TYPE_CODE_3	26	TEXT		YES	Type of vehicle	
VEH_TYPE_CODE_4	27	TEXT		YES	Type of vehicle	
VEH_TYPE_CODE_5	28	TEXT		YES	Type of vehicle	

Dimensional Modelling and ERD

We implemented the 4-step process to model our dimension and fact tables:

Step 1: Define the business or event process

The process flow initiates with the occurrence of a motor vehicle collision. Each collision event is meticulously documented and recorded in the "Motor Vehicle Collisions" (MVC) database table, where each row encapsulates the details of a specific collision incident. This data repository is enriched with information sourced from police-reported motor vehicle collisions across New York City.

This goal is to conduct in-depth traffic safety analyses to support the goal of eliminating traffic fatalities. At the core of this business context lies the MVC table, which is the primary entity for deriving insights into traffic safety and accident management. Each record within the MVC table signifies a discrete collision event, encompassing crucial details such as collision date, time, location, contributing factors, vehicle types, casualties, and injury/death counts.

Using this business context, we can think of the most significant entity which we would want to measure and gain insights from. This will be our **grain** that would be used for further analytics.

Step 2: Define the grain

The grain of a table refers to the level of detail or granularity at which each record in the table represents a distinct and unique observation or event. In the case of the NY_CRASHES table in the provided SQL schema, the grain can be determined as follows:

- Each record in the NY_CRASHES table represents a single crash event.
- The primary key of the table is COLLISION_ID, which is unique for each crash event.
- Attributes such as date, time, location, contributing factors, vehicle types, casualties, and injury/death counts provide detailed information about each crash event.
- The table captures all relevant details at the level of individual crash incidents.

Therefore, we will retain the grain of our FACT table at the level of individual crash events. Each record in the table corresponds to a single crash, and all the attributes associated with that record provide specific details about that particular crash event.

Step 3: Identify the Dimension tables

For the use case of the NY crashes application, given the business process flow mentioned above, we can conceptualize the dimension tables as follows:

- **Location Dimension** (LOCATION_DIM): Holds details about crash locations such as borough, latitude, longitude, and street names.
- **Contributing Factor Dimension** (CONT_FACTOR_DIM): Lists contributing factors to crashes.
- **Vehicle Type Dimension** (VEH_TYPE_DIM): Contains types of vehicles involved in crashes.
- **Casualty Dimension** (CASUALTY_DIM): Contains information about casualties including the number of pedestrians, cyclists, and motorists injured or killed.
- **Date Dimension** (DATE_DIM): Provides date-related attributes for time-based analysis.

Step 4: Identify the Fact table

- **Crash Fact Table** (CRASH_FACT): Stores the core crash data including collision ID, date/time, location, contributing factors, vehicle types, casualties, and injury/death counts. Foreign keys from the dimension tables (LOCATION_DIM, CONT_FACTOR_DIM, VEH_TYPE_DIM, CASUALTY_DIM, DATE_DIM) are used to establish relationships.

These tables are represented in a Star Schema as an Entity-Relationship Diagram (*Figure 1*):

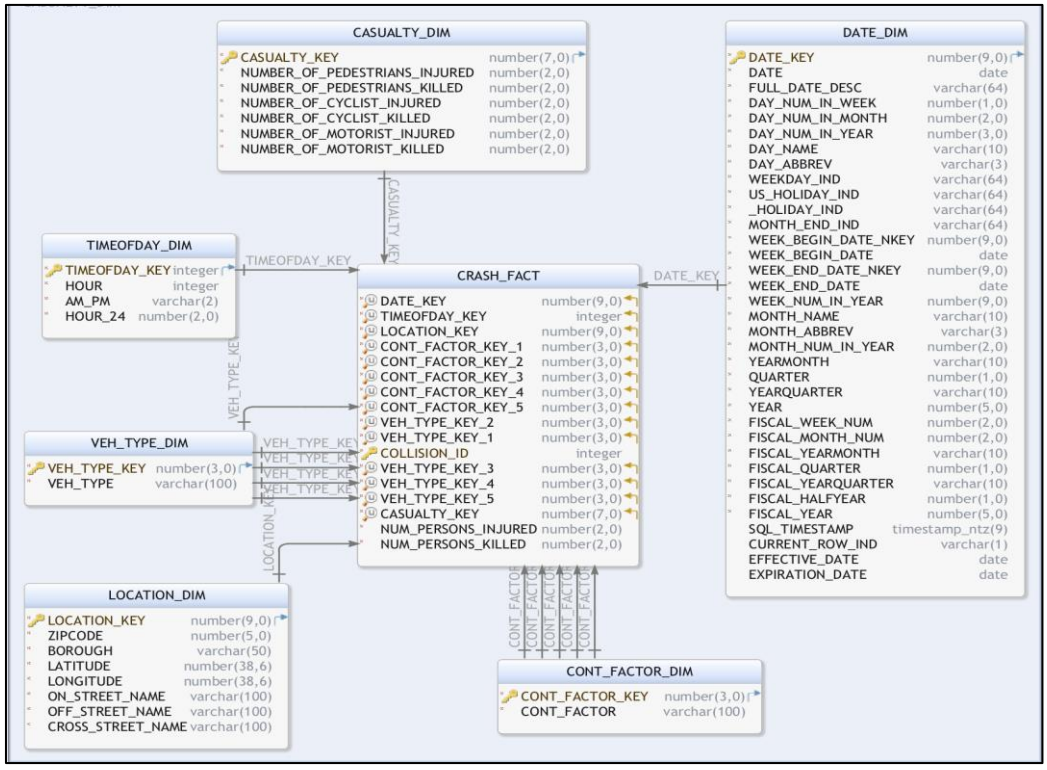


Figure 1: Star Schema

Data Transformation

The raw dataset contained 2,061,020 observations and 29 columns. We first explored the data to identify missing values in the columns pertinent to our proposed dimensions and fact table, and ultimately our KPI calculation. **There were null values for *latitude, longitude, zip code,* and *borough* along the tune of ~20% of the data (across all these columns).** This caused a problem because a lot of our analysis was geographic in nature and based on this location information. **Therefore, we performed data cleaning to insert missing column values based on other available data for the crash event**. As a result, we were able to retain 1,994,818 rows in our final NY_CRASHES table and **only had to drop 3.2% of the observations[1].**

After importing the final clean data onto Snowflake, we created the dimension and fact tables and wrote SQL queries[2] to populate them using the NY_CRASHES table. The methodology adopted for populating some of the dimension tables is as follows:

- **CASUALTY_DIM:** We assigned the casualty key in this table to unique combinations of the columns *NUM_PEDEST_INJURED,* NUM*_PEDEST_KILLED, NUM_CYCL_INJURED, NUM_CYCL_KILLED, NUM_MOTOR_INJURED and NUM_MOTOR_KILLED* in the NY_CRASHES table.
- **CONT_FACTOR_DIM:** We assigned the contributing factor key in this table to unique contributing factors across all five *CONT_FACTOR_VEH* columns in the NY_CRASHES table.
- **LOCATION_DIM:** We assigned the location key in this table to unique combinations of the columns *ZIP_CODE, BOROUGH, LATITUDE, LONGITUDE, ON_STREET_NAME, OFF_STREET_NAME and CROSS_STREET_NAME* in the NY_CRASHES table.
- **VEH_TYPE_DIM:** We assigned the vehicle type key in this table to unique vehicle types across all five *VEH_TYPE_CODE* columns in the NY_CRASHES table.

The final configuration of the Snowflake data warehouse is as follows (see Figure 2):

	name	database_name	schema_name	kind	comment	cluster_by	rows	bytes
1	CASUALTY_DIM	MYDB	NY_CRASHES	TABLE			187	3584
2	CONT_FACTOR_DIM	MYDB	NY_CRASHES	TABLE			56	2048
3	CRASH_FACT	MYDB	NY_CRASHES	TABLE			1994818	32560640
4	DATE_DIM	MYDB	NY_CRASHES	TABLE	Type 0 Dimension Table Housing Calendar and Fiscal Year		15001	693760
5	LOCATION_DIM	MYDB	NY_CRASHES	TABLE			438217	6029312
6	NY_CRASHES	MYDB	NY_CRASHES	TABLE			1994818	54471168
7	TIMEOFDAY_DIM	MYDB	NY_CRASHES	TABLE			25	1536
8	VEH_TYPE_DIM	MYDB	NY_CRASHES	TABLE			48	1536

Figure 2: The final configuration of the Snowflake Data Warehouse

[1] We dropped the column (latitude, longitude) pair since we already had separate columns for latitude and longitude, rendering this redundant. This is why the NY_CRASHES table has 28 columns.

[2] Please see the SQL codes for the fact and dimension tables in the attached file:
Finals_2024_Group12_Project6_Fact&DimTables.sql

Produce KPIs Using FACT and DIM Tables

Five main KPIs calculated using the fact and dimension tables are as the following[3]:

1. **Crashes by Area:** This KPI measures the number of crashes by zip code/borough, in a given period (month, quarter, year, etc.). The calculation for this KPI involved the *CRASH_FACT*, *LOCATION_DIM*, and *DATE_DIM* tables.

2. **Number of People Injured by Contributing Factor:** This KPI measures the number of people injured in crash events for each type of contributing factor (e.g. unsafe speed, pavement slippery, etc.), in a given time period. Calculation for this KPI involved the *CRASH_FACT*, *CONT_FACTOR_DIM*, and *DATE_DIM* tables.

3. **Number of People Killed by Contributing Factor:** This KPI measures the number of people killed in crash events for each type of contributing factor (e.g. unsafe speed, pavement slippery, etc.), in a given time period. Calculation for this KPI involved the *CRASH_FACT*, *CONT_FACTOR_DIM*, and *DATE_DIM* tables.

4. **Injuries/Deaths per Combination of Vehicle Types:** This KPI measures the number of people injured/killed per the combinations of the type of vehicles involved (for example van & bike, sedan & bus, etc.), in a given period. Calculation for this KPI involved the *CRASH_FACT*, *VEH_TYPE_DIM*, and *DATE_DIM* tables.

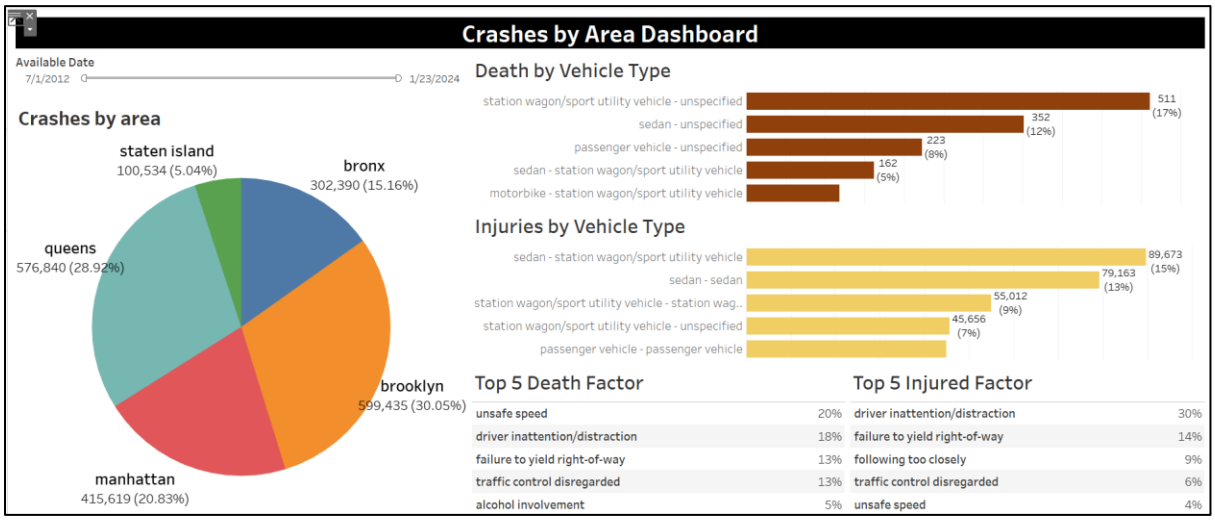
5. **Injury/Death Rate (% of Crashes):** This KPI measures the rate of people injured/killed vis-à-vis the number of crashes, calculated as (number of people injured or killed / number of crashes) x 100, for a given borough/zip code in a specific period. Calculation for this KPI involved the *CRASH_FACT*, *LOCATION_DIM*, and *DATE_DIM* tables.

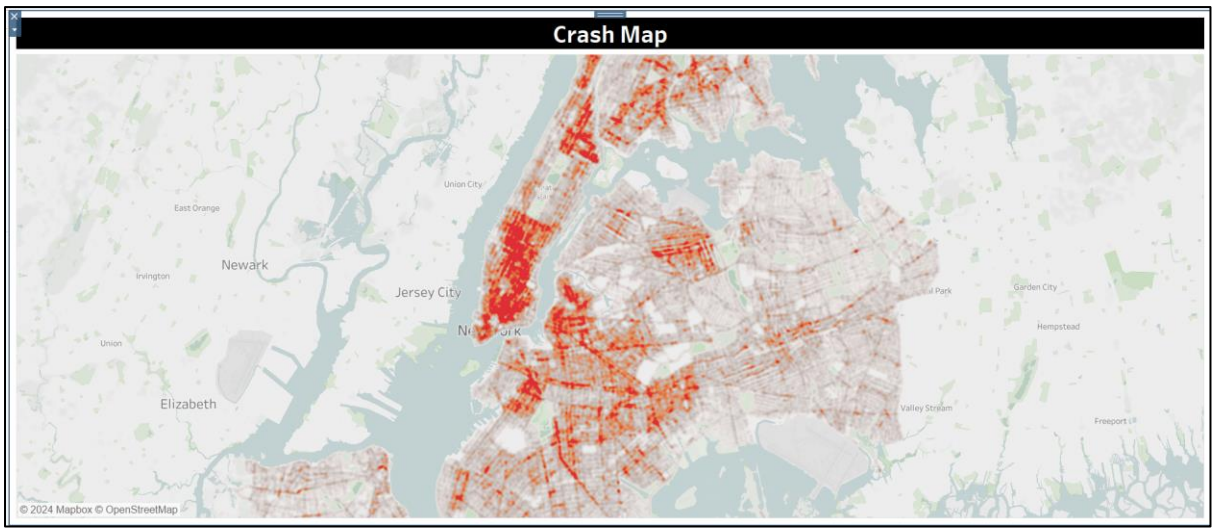
[3] Please see the SQL codes for each KPI in the attached file: *Finals_2024_Group12_Project6_KPIs.sql*

Data Visualization

To enhance the visualization and presentation of our analysis for the key performance indicators (KPIs) associated with the New York collision dataset, **we have developed five dashboards, each dedicated to a specific KPI**. Furthermore, for each KPI, we have introduced a "Dashboard Feature" section, which elucidates the dashboard's utility, and a "Dashboard Insights" section, where we derive and present insights from the collision dataset. **In conclusion, we have synthesized all relevant findings to formulate a comprehensive recommendation for the Mayor of New York on strategies to ameliorate the city's collision circumstances**, informed by historical data and thorough analysis.

1. Crashes by Area Dashboard + Crash Map





Dashboard Features:

In the *Crash by Area* dashboard, the breakdown of crashes by area is depicted through a pie chart. Users can adjust the date range with the "Available Date" slider, facilitating a deeper

understanding of area-specific crashes within selected timeframes. This dashboard features several dynamic components: "Crashes by Area" (pie chart), "Deaths by Vehicle Types" (bar chart in the upper right), "Injuries by Vehicle Types" (bar chart in the bottom right), and the "Top 5 Death/Injury Factors," all of which update based on the timeframe chosen. Additionally, for granular statistical analysis of a particular borough, users can interact with the dashboard by clicking on the respective pie chart segment. This action triggers an update of all related visualizations on the right-hand side to reflect data specific to that borough.

In the *Crash Map* dashboard, we have mapped all recorded crashes onto the New York map to provide a comprehensive visual representation of crash occurrences. This is achieved through a density (heat) map, offering users an immediate sense of where crashes are most concentrated within the city.

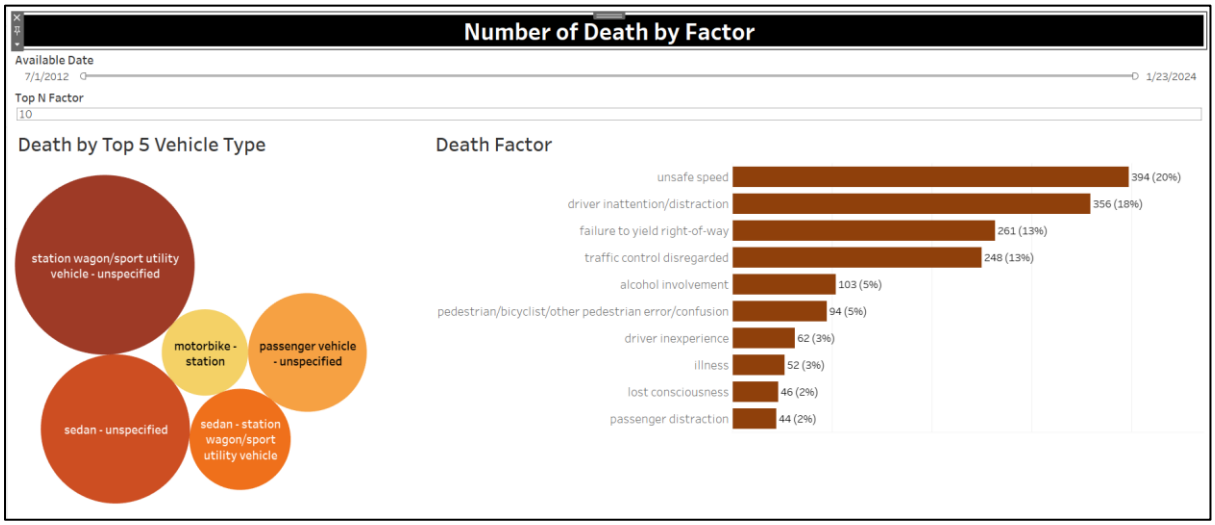
Dashboard Insights:

As illustrated in the *Crash by Area* dashboard, a significant proportion of crashes occurred in Brooklyn (30.05%), Queens (28.92%), and Manhattan (20.83%). Furthermore, the *Crash Map* reveals that within these boroughs, high-traffic districts such as Lower Manhattan, Brooklyn Heights, and Williamsburg are particularly prone to collisions. **This correlation between traffic density and collision frequency underscores the heightened risk of accidents in busier areas**.

An analysis of the *Top 5 Death and Injuries Factors* provides deeper insights into the reasons behind these incidents. Additionally, the *Top Death/Injuries by Vehicle Type* section indicates that station wagons/sport utility vehicles are most frequently involved in collisions. **The larger size of these vehicles likely contributes to the severity of impacts, resulting in higher death and injury rates compared to other vehicle types involved in collisions**.

The primary factors contributing to collisions, as identified, include drivers' disregard for laws and carelessness, manifesting in behaviors such as unsafe speed, driver inattention/distraction, and following too closely. In response to these findings, **it is recommended that the Mayor of New York consider implementing stricter enforcement of traffic laws**. Such measures could include more severe consequences for violations, aiming to compel drivers to adhere to traffic regulations and maintain full attention while driving, thereby reducing the frequency of collisions over time.

2. Number of People Killed by Contributing Factor Dashboard



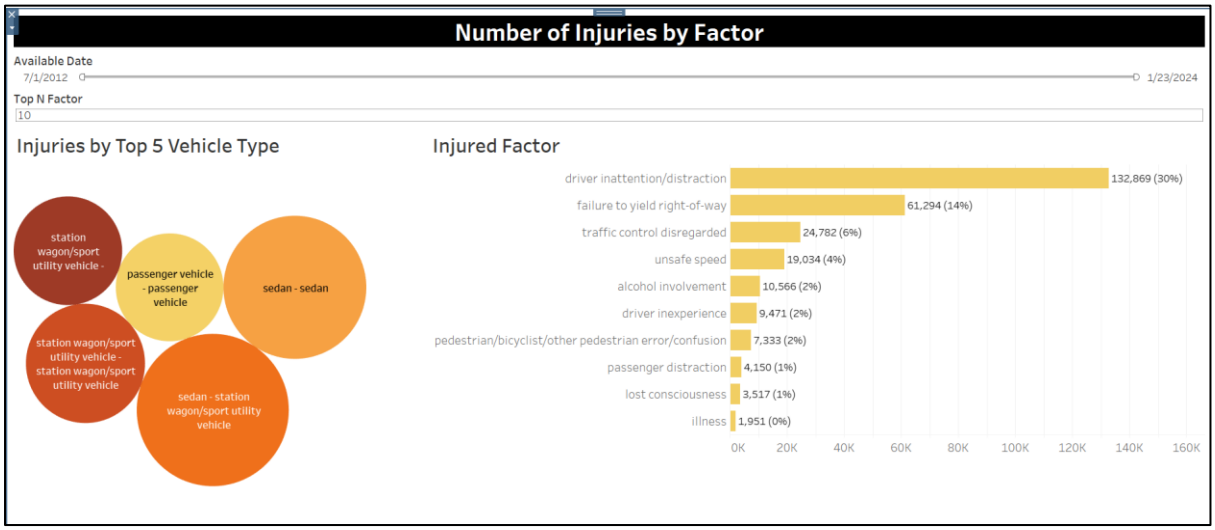
Dashboard Features:

In the dashboard, users are afforded the flexibility to adjust the date range using the "Available Date" slider located at the top of the dashboard. Within the "Top N Factor" section, users can specify their desired number of top factors for analysis. The visualization on the left-hand side displays the Top 5 vehicle types associated with lethal collisions. By interacting with a segment of the pie chart, the corresponding death factors on the right-hand side will update to reflect the top death factors attributable to the selected vehicle types. In the absence of any selection from the Top 5 Vehicle Types, the dashboard will default to showing the top N death factors, providing users with a comprehensive overview of the key elements contributing to fatalities.

Dashboard Insights:

Based on the current dashboard, it is evident that station wagons and sport utility vehicles (SUVs) are the vehicle types most frequently associated with fatalities. Furthermore, the leading causes of these fatal accidents are identified as unsafe speed (20%) and driver inattention/distraction (18%). Similar to the finding in the previous dashboard, the larger size of station wagons and SUVs likely exacerbates the impact of collisions, contributing to a higher fatality rate, particularly when combined with high speeds and driver inattention. **This data underscores the urgent need for stricter traffic laws in New York City**. By enforcing regulations more rigorously, the city aims to enhance driver focus and compliance with the law, thereby safeguarding not only the drivers themselves but also the wider community.

3. Number of People Injured by Contributing Factor Dashboard



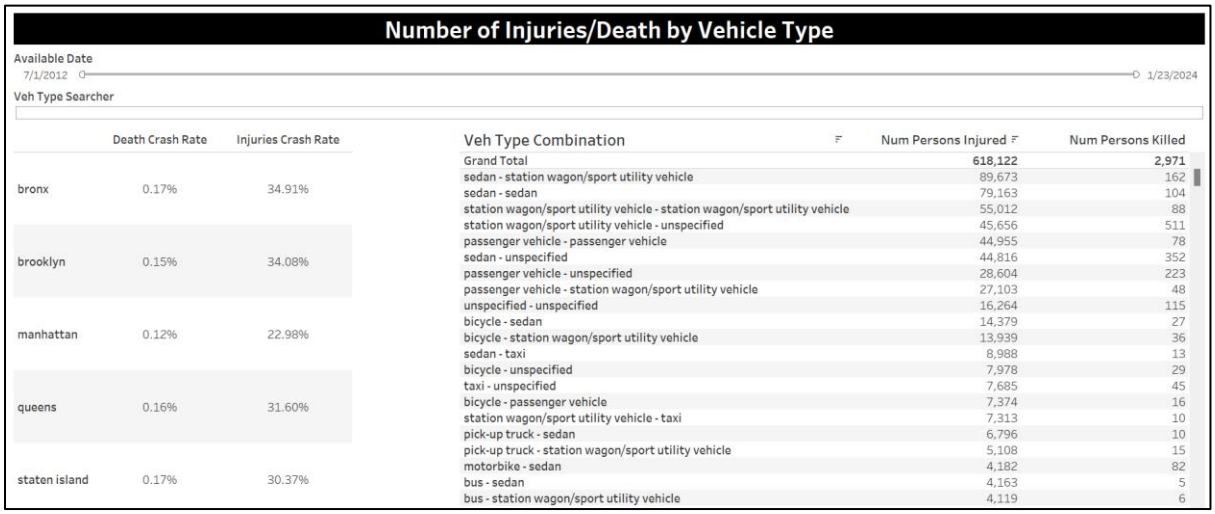
Dashboard Features:

In the dashboard, similar to the previous dashboard, users are afforded the flexibility to adjust the date range using the "Available Date" slider located at the top of the dashboard. Within the "Top N Factor" section, users can specify their desired number of top factors for analysis. The visualization on the left-hand side displays the Top 5 vehicle types associated with collisions that cause injuries. By interacting with a segment of the pie chart, the corresponding injury factors on the right-hand side will update to reflect the top injury factors attributable to the selected vehicle types. In the absence of any selection from the Top 5 Vehicle Types, the dashboard will default to showing the top N injury factors, providing users with a comprehensive overview of the key elements contributing to injuries.

Dashboard Insights:

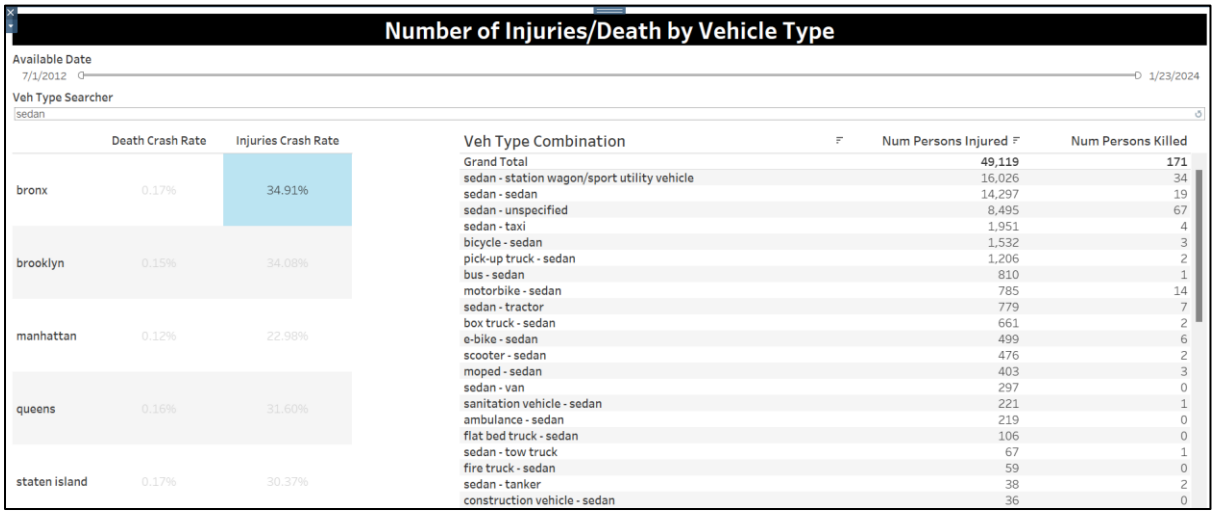
In the dashboard analysis, station wagons, and SUVs emerge as the predominant vehicle types involved in injury collisions. This trend likely stems from the greater impact these larger vehicles can inflict during accidents. A notable distinction, however, lies in the leading cause of these injuries: driver inattention/distraction, accounting for 30% of incidents, stands out as the most significant factor. **This finding could be related to the extensive use of smartphones over time while on the road and again underscores the critical need for the New York City government to implement effective strategies to ensure that drivers remain fully attentive while on the road**.

4. Injuries/Deaths by Vehicle Types Dashboard



	Death Crash Rate	Injuries Crash Rate
bronx	0.17%	34.91%
brooklyn	0.15%	34.08%
manhattan	0.12%	22.98%
queens	0.16%	31.60%
staten island	0.17%	30.37%

Veh Type Combination	Num Persons Injured	Num Persons Killed
Grand Total	618,122	2,971
sedan - station wagon/sport utility vehicle	89,673	162
sedan - sedan	79,163	104
station wagon/sport utility vehicle - station wagon/sport utility vehicle	55,012	88
station wagon/sport utility vehicle - unspecified	45,656	511
passenger vehicle - passenger vehicle	44,955	78
sedan - unspecified	44,816	352
passenger vehicle - unspecified	28,604	223
passenger vehicle - station wagon/sport utility vehicle	27,103	48
unspecified - unspecified	16,264	115
bicycle - sedan	14,379	27
bicycle - station wagon/sport utility vehicle	13,939	36
sedan - taxi	8,988	13
bicycle - unspecified	7,978	29
taxi - unspecified	7,685	45
bicycle - passenger vehicle	7,374	16
station wagon/sport utility vehicle - taxi	7,313	10
pick-up truck - sedan	6,796	10
pick-up truck - station wagon/sport utility vehicle	5,108	15
motorbike - sedan	4,182	82
bus - sedan	4,163	5
bus - station wagon/sport utility vehicle	4,119	6



	Death Crash Rate	Injuries Crash Rate
bronx	0.17%	34.91%
brooklyn	0.15%	34.08%
manhattan	0.12%	22.98%
queens	0.16%	31.60%
staten island	0.17%	30.37%

Veh Type Combination	Num Persons Injured	Num Persons Killed
Grand Total	49,119	171
sedan - station wagon/sport utility vehicle	16,026	34
sedan - sedan	14,297	19
sedan - unspecified	8,495	67
sedan - taxi	1,951	4
bicycle - sedan	1,532	3
pick-up truck - sedan	1,206	2
bus - sedan	810	1
motorbike - sedan	785	14
sedan - tractor	779	7
box truck - sedan	661	2
e-bike - sedan	499	6
scooter - sedan	476	2
moped - sedan	403	3
sedan - van	297	0
sanitation vehicle - sedan	221	1
ambulance - sedan	219	0
flat bed truck - sedan	106	0
sedan - tow truck	67	1
fire truck - sedan	59	0
sedan - tanker	38	2
construction vehicle - sedan	36	0

Dashboard Features:

As with all dashboards in this comprehensive analysis, a date slider is positioned at the top of the dashboard, allowing users to customize the time frame for their analysis. For examining the number of injuries and deaths associated with a specific vehicle type, users can enter the desired type. The outcomes are presented in two main sections: (1) on the left-hand side, the relative rates of death and injuries across the five boroughs are displayed, and (2) on the right-hand side, the absolute numbers of individuals injured or killed are shown. Should the user opt not to specify a vehicle type, the results will aggregate data across all vehicle types.

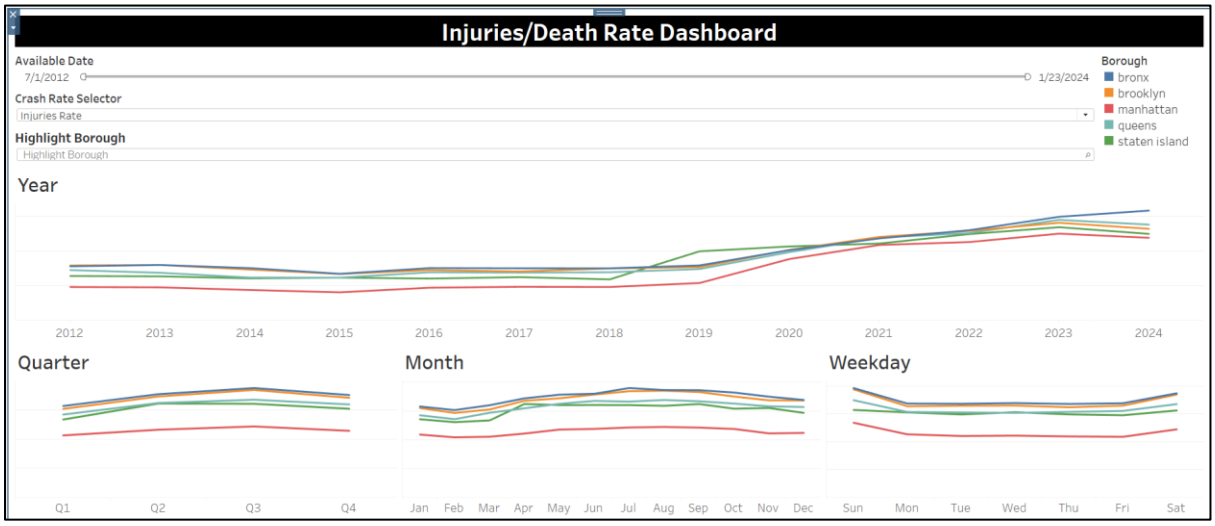
To illustrate the dashboard's functionality, consider the second screenshot as an example. Selecting a specific segment (e.g., the injury crash rate in the Bronx) will prompt an update on the right-hand side, displaying the number of collision-induced injuries specific to the Bronx. This dynamic feature enhances the user's ability to perform detailed analyses of collision data based on vehicle type and location.

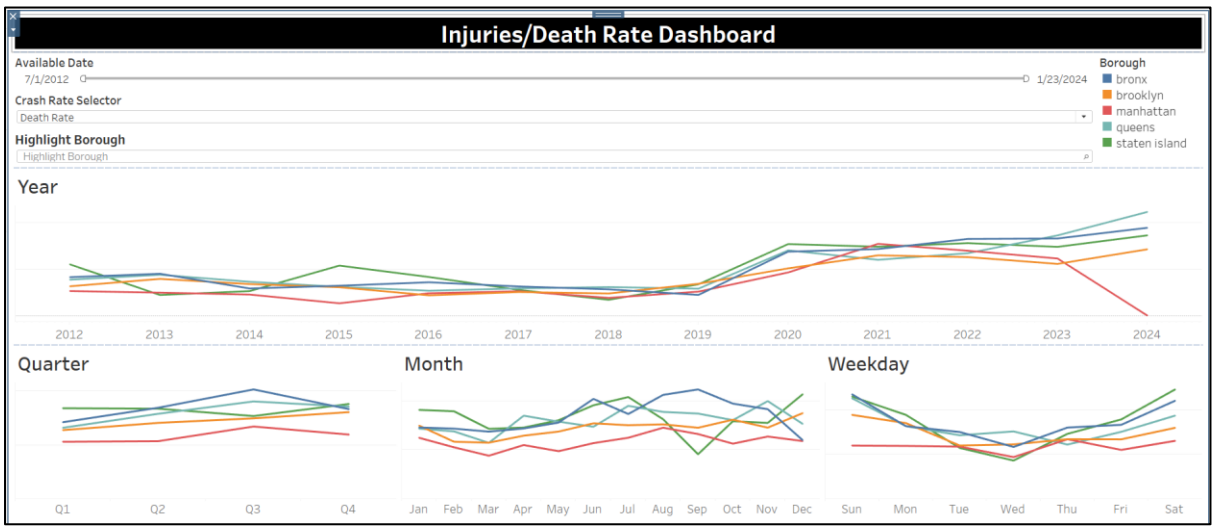
Dashboard Insights:

Upon analyzing collision data without filtering for specific vehicle types, it becomes apparent that the Bronx and Staten Island exhibit the highest rates of fatal crashes, while the Bronx and Brooklyn have the highest rates of injury crashes. These findings suggest a critical need for targeted traffic regulation enhancements, particularly in the Bronx, due to its significant rates of both fatal and non-fatal collisions. **To effectively reduce these rates, officials should prioritize interventions in the Bronx borough.**

Furthermore, the analysis indicates that station wagons/SUVs and sedans are frequently involved in these incidents, pointing to these groups of drivers as primary targets for regulatory focus. Implementing specific safety measures and regulations for drivers of these vehicle types could significantly impact the overall safety of road users in the Bronx, potentially reducing the incidence of both injuries and fatalities.

5. Injuries/Death vs. Crash Rate Dashboard





Dashboard Features:

This deep dive analysis includes dashboards equipped with a date slider at the top, allowing users to adjust the time frame to their preferred range of dates. In the "Crash Rate Selector," users can choose between viewing the injury rate or the death rate. Furthermore, the "Highlight Borough" feature enables users to select one or multiple boroughs to observe how injury or death rates have fluctuated over time.

Dashboard Insights:

The data visualizations reveal a concerning trend of increasing injury rates across all boroughs, with the Bronx registering the highest rate of injuries resulting from crashes. Seasonal analysis indicates a higher likelihood of injury collisions during summer and fall, periods characterized by increased leisure activities (e.g., summer vacations, travel, and potentially increased instances of drunk driving following National Holidays), leading to less responsible driving practices. This theory is corroborated by the observation that weekends are more prone to accidents resulting in injuries.

Regarding the death crash rate, while the overall insights mirror those of the injury crash rate, Staten Island presents a notable exception with a significant decrease in the death crash rate in September. This trend may be attributed to a reduction in tourists and visitors post-summer, leading to fewer collisions with fatal outcomes.

Conclusion

In summary, upon integrating our comprehensive analyses, it becomes imperative for the Mayor of New York to implement and enforce pertinent legislation across all boroughs to regulate driver behaviors and mitigate the risk of injuries or fatalities in collisions**. Given the critical situation, particular emphasis should be placed on the busiest areas in Manhattan, Brooklyn, and the Bronx as the primary areas of focus**. Efforts should specifically target drivers of sport wagons/SUVs and sedans, introducing measures such as driver safety campaigns, alongside the enforcement of stricter penalties for the use of smartphones while driving and for speeding violations.

Project Challenges

Utilizing a large government dataset to extract insights concerning motor vehicle collisions in NYC presented a few challenges. These can be categorized broadly into the following areas, with descriptions of how we solved them:

Data Cleaning

The raw dataset contained messy data including NULL values for critical columns such as *LATITUDE, LONGITUDE, ZIP_CODE* and *BOROUGH*. In total, there were over 400,000 rows of data that had missing values for one or more of the above-mentioned columns. Our first approach was to use geopy, the Python client for geocoding web services, to fill in missing latitude and longitude values using the addresses from column *ON_STREET_NAME* (if available). Subsequently, we would then use the latitude and longitude values to fill in the missing zip codes from geopy using the Google Maps API. However, since we needed to pull data from a web server, and our dataset had over 2 million rows, this was a painstakingly slow process and after running code for a great number of hours with little progress, we decided to take another approach.

We realized that we were taking an inefficient approach by iterating over all the rows to fetch missing values from geopy and filling them in the data frame. Instead, we created a dictionary mapping the distinct *ON_STREET_NAME* and *(LATITUDE, LONGITUDE)* pairs in the existing data, and used this dictionary to fill in a majority of the missing values that had similar *ON_STREET_NAME* values. We were now left with only a few distinct *ON_STREET_NAME* values that did not have a corresponding *(LATITUDE, LONGITUDE)* pair in the data frame. For these values, we then used geopy to extract the coordinates.

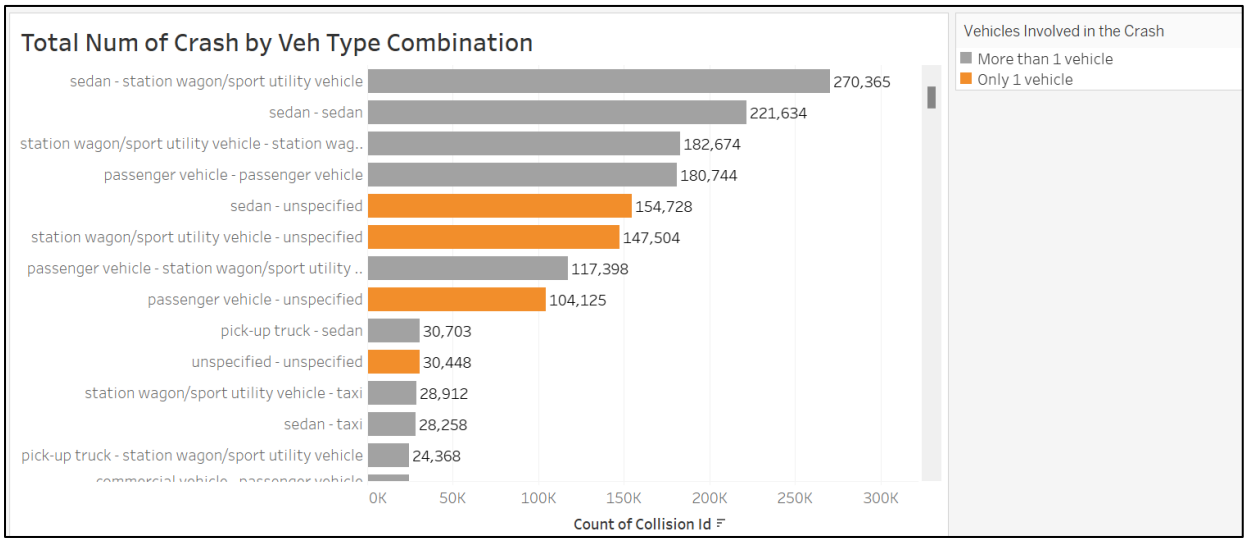
A similar method was then adopted for filling in the missing values for the ZIP_CODE column. The *BOROUGH* column was also filled by creating a dictionary of the *ZIP_CODE* and *BOROUGH* pairs in the existing data. Ultimately, we dropped all of the remaining rows that had missing values for any of the location columns, or those that had 0 values for *LATITUDE* and/or *LONGITUDE*, since that represented incorrect coordinates. We only had to drop 66,202 rows from the data, which represents 3.2% of the observations.

For the contributing factor and vehicle type columns, we had many distinct values (over 1000 for vehicle type) due to spelling mistakes, different names being used for the same category and garbage values. For these columns, we manually cleaned the distinct values (combining similar categories and categorizing the garbage values as ‘unspecified’) and used a dictionary to replace the data with the final set of distinct categories. In total, we had 56 different contributing factors and 48 different vehicle types in the end.

Uploading Data to Snowflake

Since our dataset was only ~500 MB and was required to be uploaded once, we decided to directly ingest it into Snowflake instead of using Amazon S3, which is more efficient for automating data ingestion at faster speeds. Data ingestion into Snowflake from the local computer was another challenge. Files over 250 MB cannot be ingested using Snowflake's web interface. Instead, we had to download and install Snow SQL (command line interface) for uploading the .csv file to the data warehouse stage, from where we copied it into the NY_CRASHES table. This entailed research on the ingestion process and learning CLI commands for Snow SQL.

Appendix



Total Number of Crashes by Combinations of Vehicle Type Dashboard