Business Data Processing and Business Intelligence

Executive Summary: Business Data Processing and Intelligence Work by Jørgen Leiros

Jørgen Maurstad Leiros, as part of his MSc in Business Administration and Data Science, conducted a data-driven analysis of **New York City's traffic incidents** to uncover actionable insights aimed at reducing injuries and fatalities. His work applied **data science methodologies** to process, model, and visualize large-scale traffic data, contributing to evidence-based policymaking.

In this project, I worked with a dataset containing 2.2 million records across 32 columns. Due to its size, the data had to be fetched in chunks into the Python environment. After thorough cleaning, I organized the data into separate CSV files that matched a predefined SQL schema, allowing for efficient bulk import instead of inserting one record at a time. This experience gave me practical insight into the challenges of managing and transforming large-scale data.

Key areas of focus in this project include:

1. **Traffic Data Analysis for Safety Measures** – Collected and analyzed NYC traffic incident data from the NYPD and external sources, including weather and geospatial data. He structured the data using **PostgreSQL** and applied **dimensional modeling (star schema)** to optimize database queries for Power BI visualization.
2. **Modeling and Statistical Insights** – Implemented **linear regression** to assess the impact of weather and population density on fatality rates. Findings suggested that **motorcycle-related crashes** and **traffic control device failures** were significant risk factors, while adverse weather conditions also increased fatality risks.
3. **Dashboard Development for Decision-Making** – Designed an interactive **Power BI dashboard** for the NYC Department of Transportation, allowing policymakers to analyze trends, contributing factors, and crash severity by borough and time of day.
4. **Key Findings and Policy Implications** – Identified critical areas for intervention, including **speed regulations, motorcycle safety measures, and improvements in traffic control devices**. The study emphasizes the need for **targeted safety campaigns, improved road infrastructure, and stricter enforcement of traffic laws**.

Through this work, Jørgen has demonstrated expertise in **data processing, predictive analytics, and business intelligence**, showcasing his ability to transform complex datasets into actionable insights for urban planning and safety improvements.

Introduction

In 2021, road traffic injuries accounted for an estimated 1.2 million deaths globally, ranking as the 11th leading cause of mortality worldwide (IHME, 2024). While diseases and infections dominate the top ten causes of death, the prevalence of road-related fatalities underscores a critical public health and safety challenge. Rolison et al. highlights human factors as the main cause of traffic accidents with environmental and infrastructural problems also leading to a significant number of accidents (2018). With regards to the severity of a road accident, it was noted that excessive speed was the root cause of 30 % of fatalities. These contributing factors are important to consider when developing safety initiatives, legislation, law enforcement and setting up and maintaining traffic control devices.

New York City, with its dense population, high pedestrian volume, and complex transportation network, presents unique challenges for ensuring road safety. Despite ongoing initiatives, traffic injuries and fatalities remain a pressing concern with 1 175 traffic fatalities in 2022, reflecting the need for effective, evidence-based interventions (DiNapoli, 2024). While the total number of accidents has decreased, fatalities have risen, highlighting that while some progress has been made, there remains a critical need for data-driven insights to better understand and address the causes of more severe accidents resulting in injury and death. This study aims to bridge this gap by analyzing NYC traffic incident data to uncover actionable insights that can inform targeted safety measures.

In this project we have collected traffic incident data as reported by the NYPD, performed analysis to discover insights into important factors contributing to severe crashes and created a dashboard where we visualize our findings. The dashboard is designed to support executive staff at the New York City Department of Transportation in making informed decisions regarding traffic initiatives, as well as the maintenance of roads and traffic control devices. By presenting data on the primary contributing factors of traffic accidents and their severity, the dashboard aims to enhance decision-making processes and identify critical areas for intervention, leading to the following problem statement:

How can analysis of New York City's traffic crash data inform government policies to reduce the number of crashes and fatalities across the city?

Conceptual Framework

We used the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework to guide our project, selecting it for its iterative and flexible approach (IBM, 2021). The framework structured our work, beginning with problem definition and data understanding, followed by preparation, modeling, and evaluation to uncover insights into accident severity and contributing factors. While CRISP-DM provided an overarching structure for the entire project, we applied the ETL (Extract, Transform, Load) process specifically within the data preparation stage to ensure data quality and readiness for analysis. The business and problem definition has been laid out in the introduction and the following sections will address activities and considerations for each of the subsequent stages of the CRISP-DM framework.

Data Collection and Understanding

The data used for this project consists primarily of data from incident reports about traffic accidents as reported by the New York City Police Department from 2012 to the present day. This data is collected from a RESTful API, which is part of the Socrata Open Data API (SODA), commonly used by governments and organizations to share open datasets. As part of the RESTful architecture, we interact with the API by making HTTP requests that conform to the constraints of the framework. In this case, one of the key constraints is a limit on the number of rows returned per call, which is restricted to 1,000 rows of data per request. As the dataset contains 2.4 million rows, we bypass the default limit of 1,000 rows by increasing the limit to 10,000 rows (the maximum allowed) and repeatedly calling the API until all records are retrieved.

Additional data sources were used to aid our analysis and help in the data preperation stage. Specifically, geospatial data files were downloaded from The City of New York's official website for data resources to aid in the correct labeling of New York Boroughs and weather data was collected as we suspected that weather could have an impact on road safety. The decision to download files for borough labeling instead of collecting the data from a web API was based on the need for multiple shapefiles with interdependencies. Additionally, since this is static information that does not require frequent updates, using an API call to fetch updated data was deemed unnecessary. The weather data, however, was collected through an open-source weather API which is part of the Open-Meteo Archive API. This API also operates on a RESTful architecture, allowing interaction with specified parameters such as location (latitude and longitude), date range, and weather metrics. The weather data was then merged with the main dataset based on the date value.

The relevant data for our analysis was imported into a PostgreSQL database, structured into separate tables following an analytical database design for optimized querying and analysis. The Python library Psycopg2 was used for data transfer, and the database was connected to

Power BI to create our interactive dashboard. Due to size of the dataset, csv files were created in Python and copied into the database. The primary and foreign key constraints were added in PGAdmin4.

We included weather data in our analysis based on a suspicion that it might influence both the number and severity of collisions. This suspicion was supported early in our analysis when we observed that warmer months tended to have more collisions. The rationale is further supported in a study by Bergel-Hayat et al. (2013) exploring the impact of weather on road accident risk, which states that number of collisions and casualties can be significantly affected by weather conditions.

Datasets	Attributes
Crash data	Crash Time, Crash Date, Number of Persons Killed, Number of Persons Injured, Vehicle Type, Contributing Factor, Number of Vehicles Involved, Borough, Latitude, Longitude
Weather data	Average Temperature, Minimum Temperature, Maximum Temperature, Total Precipitation, Total Snowfall, Total Rainfall, Temperature Category, Precipitation Category

Table 1: Datasets used in the project and relevant features for each.

Data Preparation

The first step of our data preprocessing focused on identifying and handling missing values. Missing values were largely due to variations in standards when filling out incident reports or the absence of data for specific columns. For example, columns like *contributing_factor_vehicle_5* and *vehicle_type_5* were left blank if there were only four vehicles involved in an incident. To address this, missing values in these columns were replaced with zero to indicate the absence of a fifth vehicle. Remaining missing values were found in columns that were not relevant to our analysis, such as *off_street_name,* which were dealt with by removing the columns entirely. Similarly, the *longitude* and *latitude* columns had 8% missing values that could not be imputed, necessitating the removal of the observations containing missing values for this column.

Another column which was important for our analysis, which had a substantial number of missing values, was the *borough* column. Furthermore, we found that the *borough* value for some observations did not match the longitude and latitudes. To improve data quality, shapefiles were used to correct mislabeled boroughs in the borough column and to fill in missing values. The shapefile contains borough boundaries and is used in conjunction with the longitude and latitude values from the crash dataset. A spatial join assigns borough names to incidents by determining whether their locations fall within the boundaries defined in the shapefile.

After addressing missing values, we focused on creating more descriptive feature names and generating new features by binning data into categories and aggregating values. Feature names were converted to lowercase for consistency, and ambiguous names were replaced with more descriptive alternatives. In addition, time-of-day data was simplified by rounding recorded times to the nearest hour. Weather features, which initially consisted of continuous values such as average daily temperature and precipitation, were transformed into categorical bins. While continuous data was useful for identifying correlations, categorizing temperature and precipitation into bins made these features easier to visualize. Finally, a new feature was created to reflect the magnitude of an incident by counting the number of vehicles involved, checking how many of the *vehicle_type_code* columns contained non-zero values.

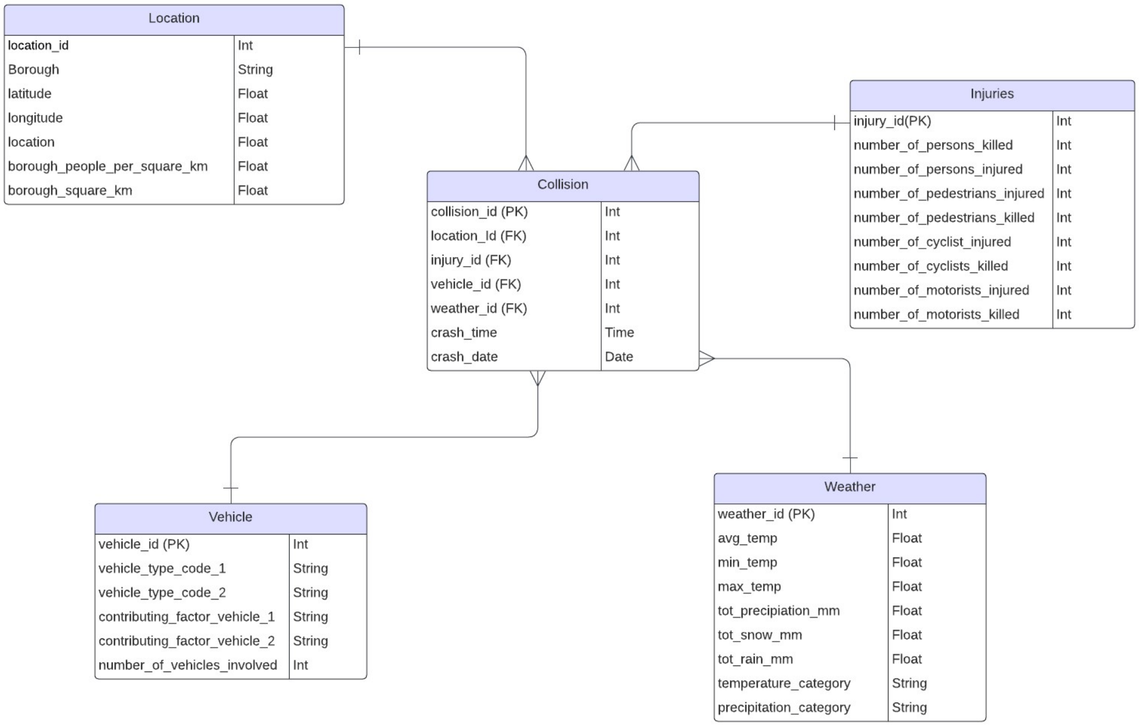
Free-text entry fields posed a unique challenge due to their inconsistent and diverse values. Notably, fields such as *vehicle_type_code* (1–5) and *contributing_factor_vehicle* (1–5) were highly relevant to our analysis but contained a wide variety of entries. To manage this, rows with values appearing fewer than 100 times were excluded, and the remaining entries were grouped into broader, more meaningful categories to facilitate analysis. The remaining number of observations was 1,453,168 and the final features which were loaded into PostgreSQL can be seen in our data dictionary [Appendix A].

Modeling

In addition to the cleaning activities presented above, we used linear regression to assess the impact of weather and population density on the number of fatalities. The population density feature was created based on the recorded population density for each borough. Based on the coefficients of the linear regression model, we found that both the *borough_people_per_square_km* and *min_temp* columns had statistically significant impacts on the number of fatalities. However, the small magnitude of these coefficients suggests that their influence, while statistically detectable, is unlikely to have meaningful practical implications.

Database Structure

To streamline visualization in Power BI we separated our data into appropriate tables making up a star schema according to dimensional modeling standards with a fact table and surrounding dimension tables before loading the data into PostgreSQL. Our subject of analysis is naturally *collision*, making this our fact table with surrounding dimension tables providing descriptions for each crash (Kimball & Ross, 2002, p. 19). We identified *weather*, *vehicle*, *injury* and *location* as the most natural dimension tables and the resulting schema can be seen in figure 1. Another possible dimension table could have been *time* but as this would have a minimal impact on storage and computation, and consisted of only two variables, we decided to leave this in the fact table.



Location	
location_id	Int
Borough	String
latitude	Float
longitude	Float
location	Float
borough_people_per_square_km	Float
borough_square_km	Float

Injuries	
injury_id(PK)	Int
number_of_persons_killed	Int
number_of_persons_injured	Int
number_of_pedestrians_injured	Int
number_of_pedestrians_killed	Int
number_of_cyclist_injured	Int
number_of_cyclists_killed	Int
number_of_motorists_injured	Int
number_of_motorists_killed	Int

Collision	
collision_id (PK)	Int
location_Id (FK)	Int
injury_id (FK)	Int
vehicle_id (FK)	Int
weather_id (FK)	Int
crash_time	Time
crash_date	Date

Vehicle	
vehicle_id (PK)	Int
vehicle_type_code_1	String
vehicle_type_code_2	String
contributing_factor_vehicle_1	String
contributing_factor_vehicle_2	String
number_of_vehicles_involved	Int

Weather	
weather_id (PK)	Int
avg_temp	Float
min_temp	Float
max_temp	Float
tot_precipiation_mm	Float
tot_snow_mm	Float
tot_rain_mm	Float
temperature_category	String
precipitation_category	String

Figure 1: Schema showing relations and key structure as well as variable and datatype for each table.

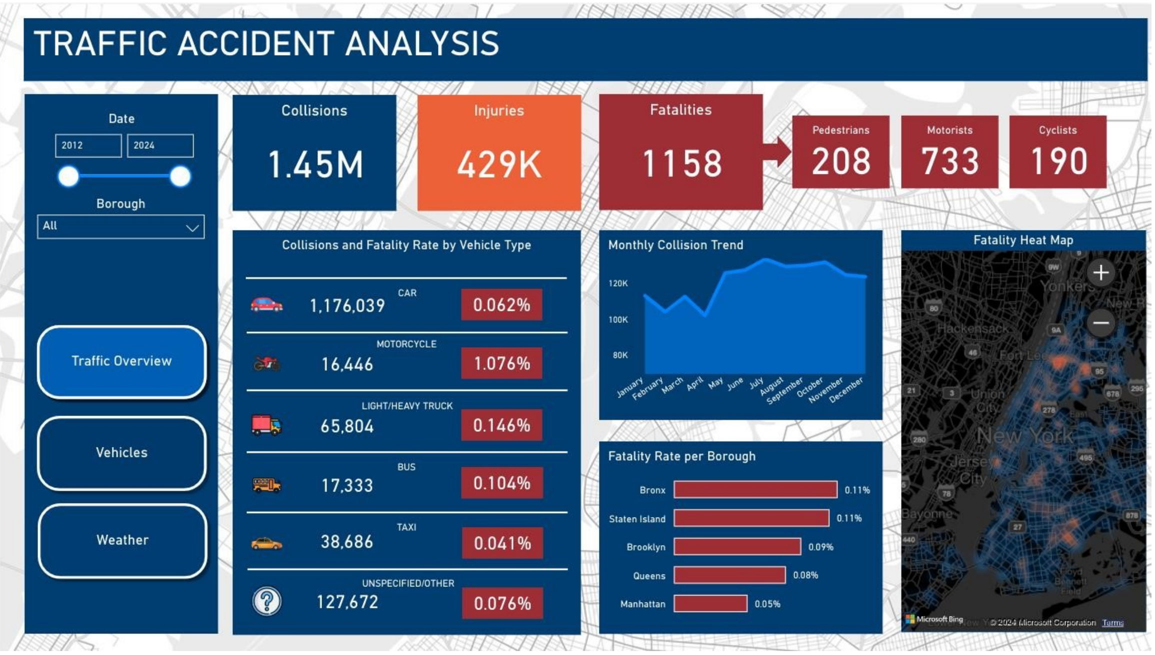
The reasoning for creating a star schema as opposed to a flat file or operational database is that the purpose of this database is visualizing findings of our analysis in a dashboard. A star schema aids the task of visualization by making querying easier and providing intuitive relations making it easier to navigate the vast amount of data (Kimberly & Ross, 2002, p. 22). An operational database aims to support transactional tasks where data is added to the database iteratively, one record at a time. For this use case, the database needs to be highly normalized, following rules for normalization which are intended to remove redundancy and improve integrity (Microsoft, 2023). While an analytical database structure is the most suitable for our current use case, adhering to normalization standards would be crucial if our dashboard were designed to visualize real-time data. Given that over 300 collisions are reported on an average day in New York City, reducing redundancy and optimizing query performance becomes essential for handling such high-frequency data efficiently.

Results

After conducting an initial analysis and modeling, we proceeded to visualize the data and insights. While the modeling showed little evidence of a correlation between the various features in our data and the number of fatalities, our goal is to provide a comprehensive overview of traffic incidents in New York City. In addition to an initial overview of the magnitude and location of crashes and resulting fatalities in the first page of our dashboard, we provide two pages to examine possible contributing factors. In addition to presenting contributing factors as reported by the NYPD, we investigate vehicle type and weather – factors that are supported by existing literature as influential in determining fatality rates (Ma et al., 2019). On the left-hand side of each view, we have included a navigation menu and filters for date and borough, allowing executives at New York City Department of Transportation to drill down into specific details as needed.

Traffic Accident Overview

The first page of the dashboard provides an overview of traffic accident data in New York City. Key metrics including collisions, injuries and fatalities are prominently displayed alongside a breakdown of the fatalities by road user type. Additionally, a container to the left of the dashboard highlights collision and fatality rates by vehicle type reflecting a disproportional risk related to motorcycles. A monthly trend graph tracks average collisions over the year visualizing a trend of fewer collisions in the colder months. A bar chart shows the fatality rate per borough combined with A fatality heat map which visualizes accident hotspots across the city. By looking into the locations with the highest number of fatalities, the target audience might get valuable insight into important contributing factors such as speed limit informing possible preventative measures.



TRAFFIC ACCIDENT ANALYSIS

Date: 2012 – 2024

Borough: All

Traffic Overview

Vehicles

Weather

Collisions 1.45M

Injuries 429K

Fatalities 1158 → Pedestrians 208, Motorists 733, Cyclists 190

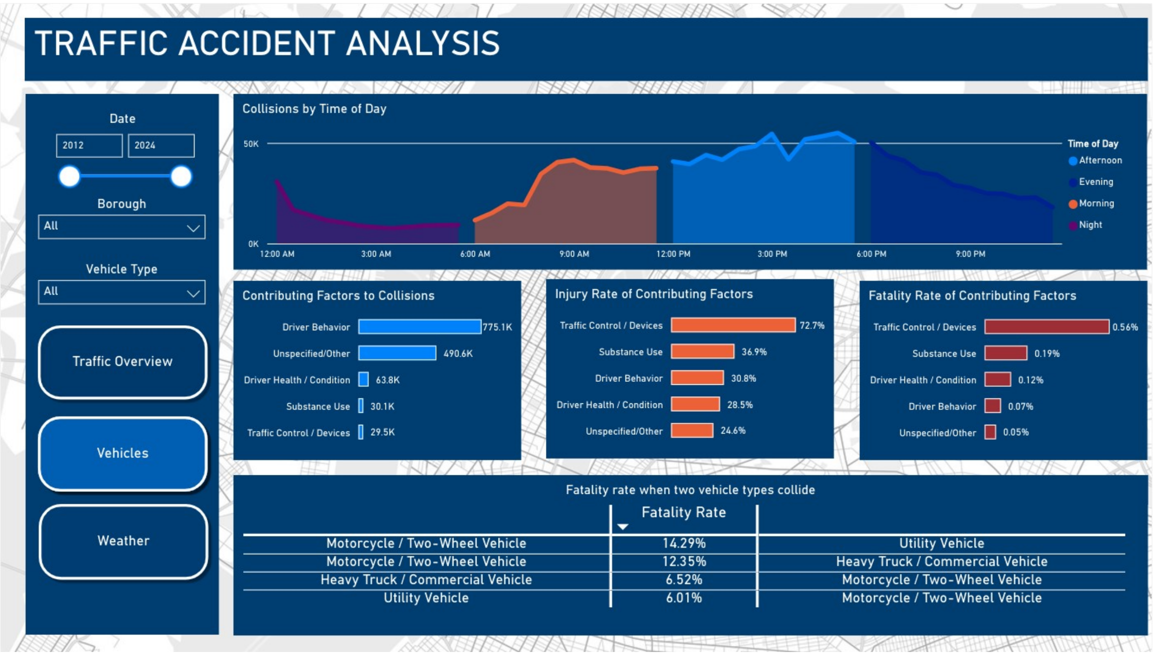
Collisions and Fatality Rate by Vehicle Type

CAR	1,176,039	0.062%
MOTORCYCLE	16,446	1.076%
LIGHT/HEAVY TRUCK	65,804	0.146%
BUS	17,333	0.104%
TAXI	38,686	0.041%
UNSPECIFIED/OTHER	127,672	0.076%

Figure 2: First view of our dashboard providing an overview of historical traffic incident data in New York City.

Contributing factors and Vehicle Impact

The second page of the dashboard focuses on analyzing collisions in New York City by time of day, contributing factors, and vehicle interactions. A time-of-day trend chart reveals higher collision rates during morning and afternoon hours, with a noticeable decline at night. Bar charts provide insights into contributing factors, highlighting driver behavior and traffic control/device issues as leading causes, with their respective injury and fatality rates. Additionally, a table examines fatality rates when specific vehicle types collide, emphasizing the high risk associated with motorcycle interactions. This page helps the target audience identify high-risk periods, common contributing factors, and vehicle type dynamics, offering actionable insights for improving traffic safety and reducing fatalities.



Fatality rate when two vehicle types collide

	Fatality Rate	
Motorcycle / Two-Wheel Vehicle	14.29%	Utility Vehicle
Motorcycle / Two-Wheel Vehicle	12.35%	Heavy Truck / Commercial Vehicle
Heavy Truck / Commercial Vehicle	6.52%	Motorcycle / Two-Wheel Vehicle
Utility Vehicle	6.01%	Motorcycle / Two-Wheel Vehicle

Figure 3: Second view provides more granular view into contributing factors and fatality per vehicle type.

Weather Impact

The final page of the dashboard explores the relationship between weather conditions, temperature, and traffic accidents in New York City. Bar charts display the distribution of collisions across different temperatures and weather conditions, with warmer and milder temperatures showing higher collision counts. The fatality rate is also highest for the warmer days as well as for very rainy conditions, highlighting increased risk during adverse weather. A stacked bar chart of the contributing factors further breaks down collision causes by weather conditions, emphasizing driver behavior as a dominant factor across all conditions. We also see a large increase in the environmental / road conditions factor for both snowy and very snowy days. Finally, a trend chart visualizes collisions, injuries, and fatalities throughout the year, revealing some seasonal variations and enabling stakeholders to link weather patterns with traffic safety outcomes. This page provides actionable insights into how environmental factors influence traffic risks, supporting better planning and prevention strategies.

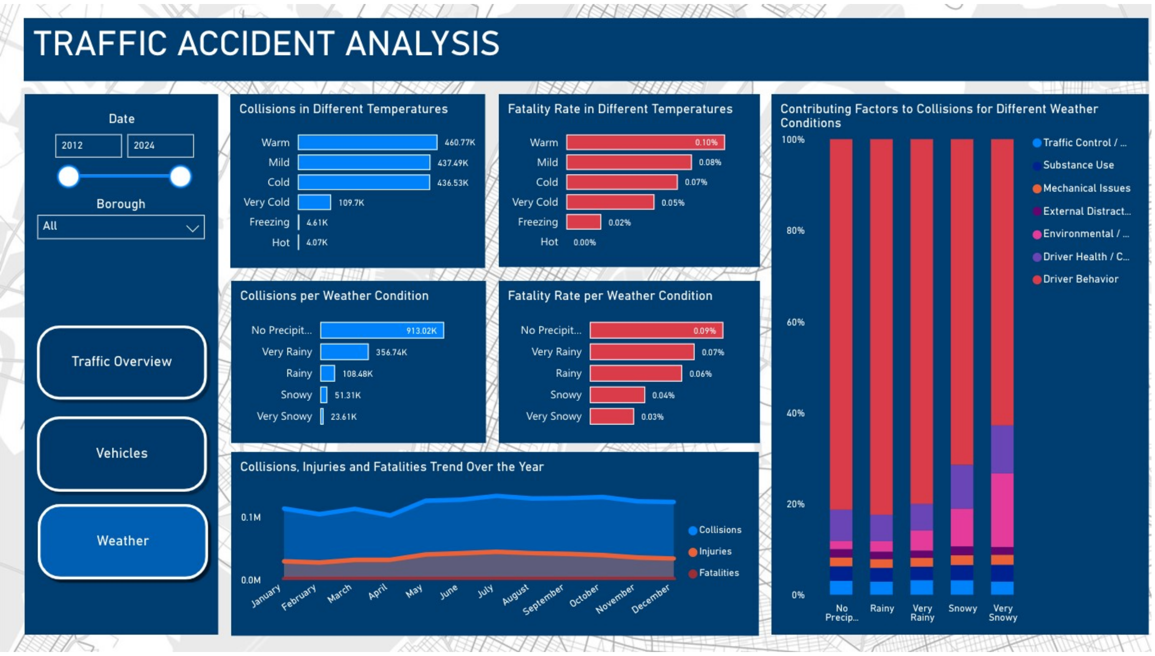


Figure 4: Third view provides insight into the correlation between weather and fatalities.

Discussion

From investigating the contributing factors to severe traffic incidents leading to injury and fatality, we have made some interesting discoveries which could help inform new policies and better allocation of resources. Initial analysis and modeling efforts revealed little significant patterns but when visualizing the data and looking at fatality rate rather than number of fatalities, we were able to uncover valuable insight. From the first page of our dashboard, one of the most interesting insights we can see is the fatality rate of crashes involving motorcycles compared to any other type of vehicle. Exploring this insight further in the next page reveals that collisions between motorcycles and heavy vehicles have the highest fatality rate, suggesting that motorcyclists are at risk. Another interesting insight from the first page is a slight trend of higher fatality rate in the less densely populated areas and more frequently on main roads confirmed when you zoom in on the heat map. Although the reason for this must be investigated further, one possible explanation is that vehicles travel at higher speeds in these areas.

When looking at the distribution of both vehicles, contributing factors and weather conditions it is clear why it can be difficult to find significant correlation between our features and the target variable of fatalities. Even though motorcycles have the highest fatality rate, they are involved in less than a 60th of the crashes cars are involved in, and while snow might be the leading environmental cause of fatalities, there are a lot more days with no precipitation, making the impact of this finding less significant when looking at over 2 million records. *Driver behavior* is also listed as the contributing factor for the vast majority of incidents while *traffic control / devices* have a much higher fatality rate. If we consider the severity of the events investigated in this report, it should be clear that factors related to a high fatality rate should be dealt with despite them making up a relatively small proportion of the total traffic incidents in the last 12 years.

The analysis highlights driver behavior as the most significant contributing factor to collisions and fatalities, a broad category that includes unsafe speed, failure to yield, and other risky actions. This suggests a critical need for further investigation into the demographics and characteristics of drivers involved in crashes, including factors such as age, driving history, and potential impairment. Understanding who is crashing and why would enable more targeted interventions, such as tailored safety campaigns, stricter enforcement of speeding laws, and educational initiatives to address specific high-risk behaviors.

Among our concrete findings, traffic control devices—under the direct purview of the Department of Transportation—stand out as the contributing factor with the highest injury and fatality rates. This underscores the need for DOT to assess the placement, functionality, and visibility of these devices to ensure they are improving safety rather than inadvertently contributing to accidents. Additionally, the disproportionate risk associated with motorcycles and increased fatalities during adverse weather conditions highlight the need for targeted safety campaigns, improved road infrastructure for all weather conditions, and enhancements like better drainage and snow/ice treatments. These measures could significantly improve traffic safety and reduce fatalities in New York City.

Conclusion

The insights highlighted in this report can guide the Department of Transportation in shaping policies to reduce crashes and fatalities across New York City. By identifying highrisk behaviors, vehicle types, and environmental conditions, the analysis provides a foundation for targeted preventative measures and infrastructure improvements. However, limitations such as potential gaps in data quality, underreporting of certain factors, and the broad categorization of driver behavior must be addressed to refine policy recommendations. Further investigation into specific driver demographics, the effectiveness of traffic control devices, and the role of weather conditions can help develop more precise and impactful government policies to enhance traffic safety and save lives.

References

IHME, Global Burden of Disease (2024) – with minor processing by Our World in Data. “Acute hepatitis” [dataset]. IHME, Global Burden of Disease, “Global Burden of Disease - Deaths and DALYs” [original data].

U.S. Census Bureau. (2024) New York City, New York. Retrieved 18. 11. 2024 from https://data.census.gov/table/DECENNIALPL2020.P1?g=160XX00US3651000

DiNapoli, T. P. (27. 06. 2024) Motor Vehicle Fatalities Rise Sharply in NY. Office of the New York State Comptroller. https://www.osc.ny.gov/press/releases/2024/06/dinapoli-motorvehicle-fatalities-rise-sharply-ny

Kimball R. & Ross M. (2002). The data warehouse toolkit: the complete guide to dimensional modeling (2nd ed.). Wiley.

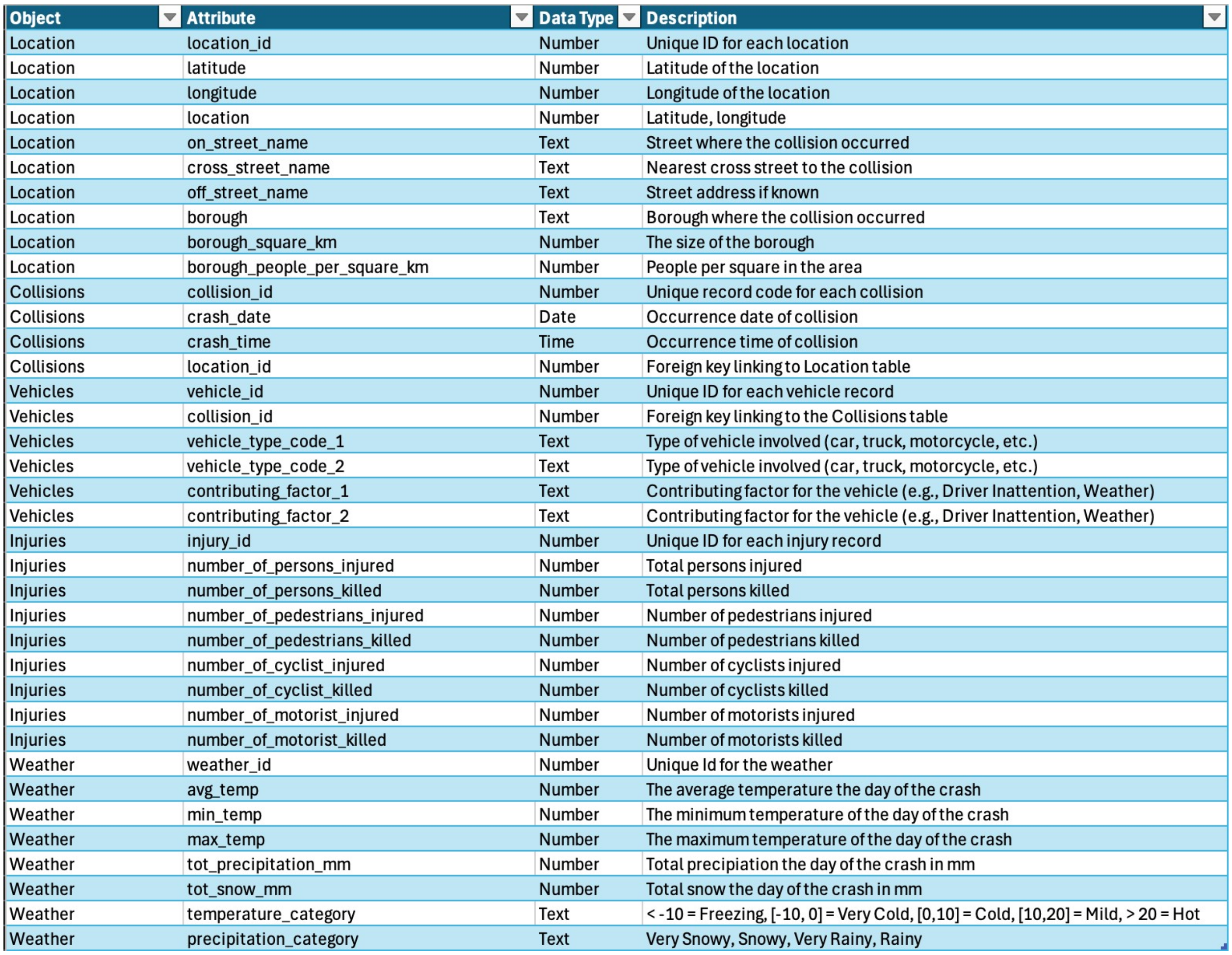
Microsoft (2023), Description of the database normalization basics. Retrieved from https://learn.microsoft.com/en-us/office/troubleshoot/access/database-normalizationdescription

J. Ma, Y. Ding, J. C. P. Cheng, Y. Tan, V. J. L. Gan and J. Zhang. (2019), "Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective," in IEEE Access, vol. 7, pp. 148059-148072, 2019, doi: 10.1109/ACCESS.2019.2946401.

IBM (2021), CRISP-DM Help Overview. Retrieved 02. 12. 2024 from https://www.ibm.com/docs/pt-br/spss-modeler/saas?topic=dm-crisp-help-overview

Appendix

A) Data Dictionary



Object	Attribute	Data Type	Description
Location	location_id	Number	Unique ID for each location
Location	latitude	Number	Latitude of the location
Location	longitude	Number	Longitude of the location
Location	location	Number	Latitude, longitude
Location	on_street_name	Text	Street where the collision occurred
Location	cross_street_name	Text	Nearest cross street to the collision
Location	off_street_name	Text	Street address if known
Location	borough	Text	Borough where the collision occurred
Location	borough_square_km	Number	The size of the borough
Location	borough_people_per_square_km	Number	People per square in the area
Collisions	collision_id	Number	Unique record code for each collision
Collisions	crash_date	Date	Occurrence date of collision
Collisions	crash_time	Time	Occurrence time of collision
Collisions	location_id	Number	Foreign key linking to Location table
Vehicles	vehicle_id	Number	Unique ID for each vehicle record
Vehicles	collision_id	Number	Foreign key linking to the Collisions table
Vehicles	vehicle_type_code_1	Text	Type of vehicle involved (car, truck, motorcycle, etc.)
Vehicles	vehicle_type_code_2	Text	Type of vehicle involved (car, truck, motorcycle, etc.)
Vehicles	contributing_factor_1	Text	Contributing factor for the vehicle (e.g., Driver Inattention, Weather)
Vehicles	contributing_factor_2	Text	Contributing factor for the vehicle (e.g., Driver Inattention, Weather)
Injuries	injury_id	Number	Unique ID for each injury record
Injuries	number_of_persons_injured	Number	Total persons injured
Injuries	number_of_persons_killed	Number	Total persons killed
Injuries	number_of_pedestrians_injured	Number	Number of pedestrians injured
Injuries	number_of_pedestrians_killed	Number	Number of pedestrians killed
Injuries	number_of_cyclist_injured	Number	Number of cyclists injured
Injuries	number_of_cyclist_killed	Number	Number of cyclists killed
Injuries	number_of_motorist_injured	Number	Number of motorists injured
Injuries	number_of_motorist_killed	Number	Number of motorists killed
Weather	weather_id	Number	Unique Id for the weather
Weather	avg_temp	Number	The average temperature the day of the crash
Weather	min_temp	Number	The minimum temperature of the day of the crash
Weather	max_temp	Number	The maximum temperature of the day of the crash
Weather	tot_precipitation_mm	Number	Total precipiation the day of the crash in mm
Weather	tot_snow_mm	Number	Total snow the day of the crash in mm
Weather	temperature_category	Text	< -10 = Freezing, [-10, 0] = Very Cold, [0,10] = Cold, [10,20] = Mild, > 20 = Hot
Weather	precipitation_category	Text	Very Snowy, Snowy, Very Rainy, Rainy