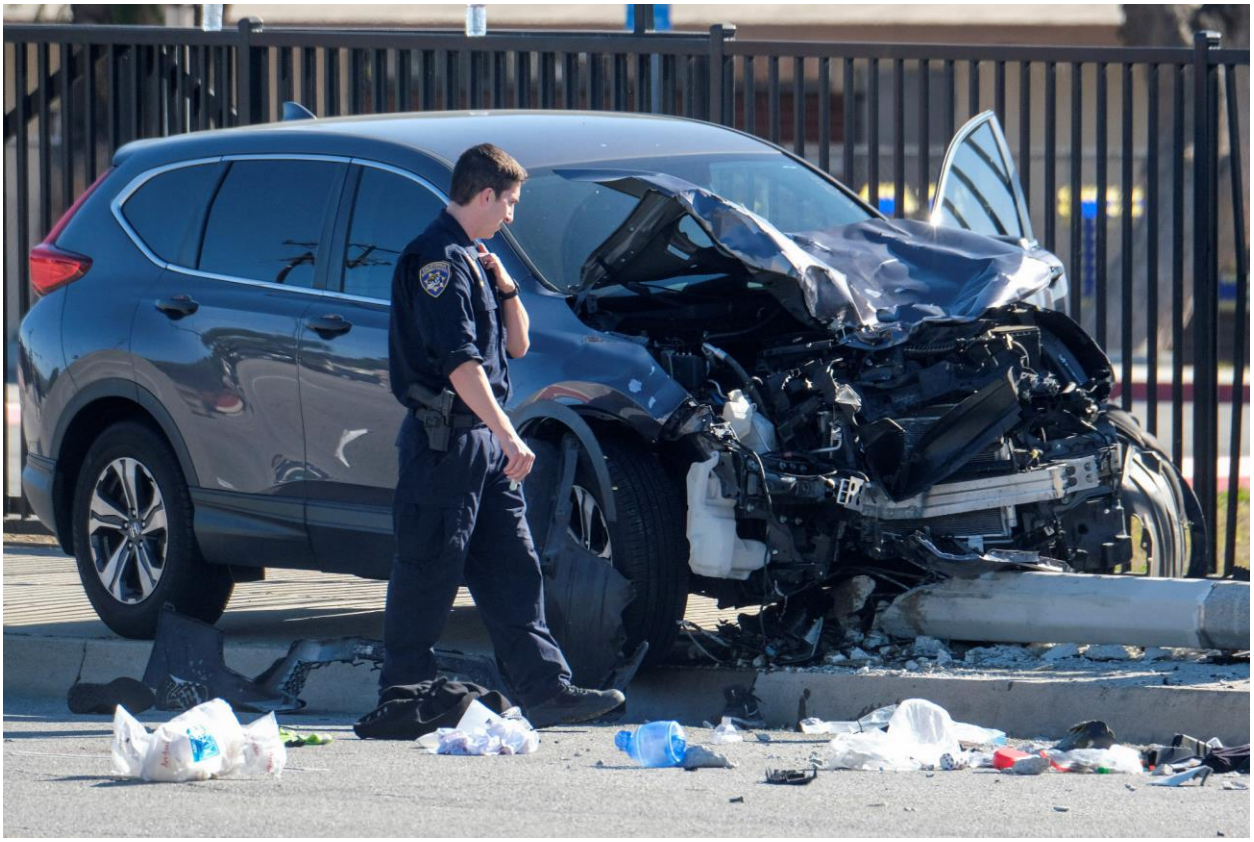
Introduction to Data Science 2023
Group B2

Preliminary report on data-analysis project

Finding patterns in NYC car crashes



Contents

Business Understanding

Background

This project aims to analyze the Motor Vehicle Collisions data, which includes details from police-reported motor vehicle collisions in New York City. Each record in the dataset represents a unique crash event, offering a comprehensive view of the factors and circumstances surrounding these incidents.

Business Goals

The primary goal is to use data on motor vehicle collisions to accurately predict the severity (mostly in terms of lethality) of motor vehicle collisions. By understanding the patterns and factors which contribute to severe crashes, this analysis hopes to provide a proof-of-concept for an automated system which could assist emergency services in the allocation of resources. If successful, such a system might also be used to inform policy decisions and improve road safety measures.

Business success criteria

Success in this project will be measured by the ability to:

- Accurately identify patterns linking vehicle data to crash lethality.
- Establish patterns between the timing, location and severity of crashes.
- Develop a predictive model which is able to accurately assess crash severity based on the previously mentioned factors.

Inventory of resources

Data: Access to the Motor Vehicle Collisons dataset provided by the NYC Police Department.

Software: Analytical tools and libraries. For the purposes of this project, Python and Python libraries will be used.

Requirements, Assumptions, Constraints

Requirements: Data handling procedures which account for the deadline of the data analysis project.

Assumptions: The dataset is comprehensive and accurately reflects the crash events in NYC.

Constraints: Limited information provided on the participating vehicles in collisions. Limited information in regards to the reason for the crash. Given that the dataset contains preliminary collision information (from police reports), the data might not be complete for some collision events (e.g. a participant dying later, due to injuries sustained in the collision).

Terminology

Collision: The dataset used in this analysis is based on police reported collisions provided by the NYC Police Department. The police report (MV104-AN) is required to be filled out for collisions where someone is injured or killed, or where there is at least $1000 worth of damage.

Costs and Benefits

Costs: Costs for this analysis, in terms of data processing costs, can be considered negligible.

Benefits: Improved public safety and potential reductions in road accidents and associated costs.

Data Mining Goals

The goals of data-mining are as follows:

- To find patterns and relationships between vehicle types and crash lethality.

- To find patterns between the time, location, and severity of crashes.
- To develop a predictive model for estimating crash severity based on the previously mentioned factors.

Data Mining Success Criteria

The success of data-mining for this analysis is based on:

- Identification of significant patterns in factors (vehicle data, time, location) provided by the dataset and severity of collisions.
- Development of a predictive model which can, with reasonable accuracy, predict crash severity.
- Insights gained into factors which lead to more severe collisions.

Data Understanding

Data Requirements

This data analysis aims to find patterns in crash lethality in regards to vehicle types, time of crash, and location. Therefore, the required data must include:

- Details about vehicles involved in the collision.
- Information about the crash (fatalities, injuries).
- Time and location of each collision event.

Data Availability

The public dataset provided by the City of New York, composed of police reports from the NYPD, contains the necessary details for this analysis. The records for collision events include information about the vehicles involved, crash specifics including fatalities and injuries, and the time and location of the collision.

Selection Criteria

All of the data required for our analysis is taken from the collisions dataset provided by the city of New York. This dataset contains 29 columns.

For our purposes, we will use all columns except for Latitude and Longitude. Those two columns are not strictly necessary as the Location column already contains a Latitude, Longitude pair.

Describing Data

The source of the dataset is the City of New York. It contains 29 columns, which are as follows:

COLUMN NAME	DESCRIPTION	TYPE	FORMAT
CRASH DATE	Occurrence date of collision	Date & Time	MM/DD/YY
CRASH TIME	Occurrence time of collision	Plain Text	24H format E.g 14:42
BOROUGH	Borough where collision occurred	Plain Text	E.g BROOKLYN
ZIP CODE	Postal code of incident occurrence	Plain Text	E.g 11226
LATITUDE	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)	Number	E.g 40.65584
LONGITUDE	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)	Number	E.g -73.950134
LOCATION	Latitude , Longitude pair[1]	Location	E.g (40.65584°, -73.950134°)

[1] A collision at an intersection located on a precinct border may appear in the list of intersections of any precinct that shares that intersection,
but will appear in only one precinct list. Collision location is captured as the nearest intersection.
A collision on a highway, bridge or tunnel which occurs at or near a precinct border may appear in the list of highways, bridges or tunnels of any precinct
that shares that location, but will appear in only one precinct list. Collision location is captured as the nearest highway/bridge/tunnel reference/mile marker.

ON STREET NAME	Street on which the collision occurred	Plain Text	E.g STATEN ISLAND EXPRESSWAY
CROSS STREET NAME	Nearest cross street to the collision	Plain Text	
OFF STREET NAME	Street address if known	Plain Text	
NUMBER OF PERSONS INJURED	Number of persons injured	Number	
NUMBER OF PERSONS KILLED	Number of persons killed	Number	
NUMBER OF PEDESTRIANS INJURED	Number of pedestrians injured	Number	
NUMBER OF PEDESTRIANS KILLED	Number of pedestrians killed	Number	
NUMBER OF CYCLIST INJURED	Number of cyclists injured	Number	
NUMBER OF CYCLIST KILLED	Number of cyclists killed	Number	
NUMBER OF MOTORIST INJURED	Number of vehicle occupants injured	Number	
NUMBER OF MOTORIST KILLED	Number of vehicle occupants killed	Number	
CONTRIBUTING	Factors contributing	Plain Text	Full list[2]

[2] AGGRESSIVE DRIVING/ROAD RAGE, ALCOHOL INVOLVEMENT, BACKING UNSAFELY, CELL PHONE (HAND-HELD), DRIVER INATTENTION/DISTRACTION, DRIVER INEXPERIENCE, DRUGS (ILLEGAL), EATING OR DRINKING, ERR/CONFUSN PED/BIKE/OTHER PED, FAILURE TO KEEP RIGHT, FAILURE TO YIELD RIGHT-OF-WAY, FATIGUED/DROWSY, FELL ASLEEP, FOLLOWING TOO CLOSELY, ILLNESS, LOST CONSCIOUSNESS, OTHER ELECTRONIC DEVICE, OTHER UNINVOLVED VEHICLE, OUTSIDE CAR DISTRACTION, PASSENGER DISTRACTION, PASSING OR LANE USAGE IMPROPER, PASSING TOO

FACTOR VEHICLE 1-5	to the collision for designated vehicle		Contributing factors are listed when known.
COLLISION_ID	Unique record code generated by system. Primary Key for Crash table.	Number	E.g 4023867
VEHICLE TYPE CODE 1-5	Type of vehicle based on the selected vehicle category	Plain Text	(ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)

At 2.05M rows, the dataset is sufficiently large for our analysis. It also includes sufficient information about each collision for us to draw conclusions from.Exploring Data

The dataset we are using for this analysis can be considered sufficiently comprehensive for our purposes. While there is a noticeable amount of missing data, especially for columns relating to the streets on which or off which the collision took place, this is not a significant problem for our analysis.

From a quick examination it is apparent that most collisions only include two vehicles. This is proven by the fact that the Vehicle Type columns from 3-5 are mostly empty. This is expected.

Data Quality

While the dataset we are using is not without flaws, the quality issues present in it are not major enough to disqualify it as a data source for our analysis. The size of the dataset is large enough that any rows with issues disqualifying it for a given analytical task can be disregarded for that specific task (e.g. while a row with insufficient location data can't be

CLOSELY, PHYSICAL DISABILITY, PRESCRIPTION MEDICATION, TRAFFIC CONTROL DISREGARDED, TURNING IMPROPERLY, UNSAFE LANE CHANGING, UNSAFE SPEED

used for drawing conclusions regarding locations, it can still be used for analysis of patterns regarding time or some other factor).

Project Plan

Tasks

For our data analysis project, there are x tasks which need to be accomplished:

1. Find patterns between types of vehicles involved in collisions and collision severity.
2. Find patterns between collision severity and the time of collisions.
3. Find patterns between collision severity and the location of collisions.
4. Create a heatmap of collisions overlaid onto a map of NYC.
5. Train a model to predict the severity of crashes based on location, time and contributing vehicle data.

In our analysis, each member of the team will contribute to each of these tasks. The estimated contribution time for each member will be about 7 hours for tasks 1-3 and 5. Task 4 relies on data which we have already extracted, so the estimated contribution time for each member is 1 hour.

Methods and Tools

For our data analysis, we plan to use the Python programming language and available libraries related to machine learning and data analysis.