Motor Vehicle Collision Analysis

Abijith Pradeep (ap8246)
Damaraju Venkatesh
Jaya Sabarish Reddy Remala

Introduction

Traffic accidents are serious issues, that can possibly cause disabilities, injuries, and even fatalities.

In order to decrease the number of accidents, we need to understand and analyze the traffic accidents dataset.

Big Data ecosystem has the ability to store, manipulate, analyze, and mine large traffic accident datasets and can drive knowledge creation that can help decision-makers reduce the number of accidents.

Problem Formulation

Given the dataset of Motor Vehicle Collisions, provide statistical answers by searching within the data to provide useful insights for motor vehicle accident prevention.

1) Are number of accidents increasing or decreasing over years?
2) Can we identify the hotspots across NYC for accidents?
3) What are major causes for accidents?
4) Which category of people on the road are most affected because of accident?
5) Are new laws and regulations effective in limiting the accidents?

Methodology - Data Pre-Processing

1. **Null values:** A record from here will be ignored only if collision_id is null else the column will be considered and can provide useful information.
2. **Borough filling:** Using Geospatial libraries and analysis we could determine the Borough of a particular collision_id given its latitude and longitude values.
3. **Vehicle Group:** Grouping vehicle using vehicle_make and vehicle_type even when either one of them is null.

Methodology - Data Analysis

- Categorical Analysis
- Vehicle Data Analysis
- Contributing Factor Analysis
- Hotspot Analysis

Categorical Analysis

1. For Accidents involving pedestrians, based on the pedestrian action reported, count of accidents are calculated
2. For all the categories chosen the contributing factor causing major number of accidents is calculated
3. Distribution of accidents across all boroughs is analysed
4. Accident count is normalised by dividing it with the population of the borough.
5. Injuries and Deaths comparison over years and time of day..
6. Impact of new legislations as part of Vision Zero are also analysed.

Vehicle Analysis

1. Created Vehicle groups as:
   a. Small, Light, Medium, Heavy, Other (Unknown), Unspecified
2. Analysis was done filtering out the Unspecified group
3. Created a UDF on the above groups to return strings based on the presence of a group.
4. Filtered out with license_status column and analysed for unlicensed drivers, which indirectly checks out as crime.

Contributing Factor Analysis

- There are several factors that were recorded in the dataset that causes an accident.
- Almost >65 factors contributes the happening of all accidents in Bronx, Brooklyn, Manhattan, Queens and Staten Island.
- In order to visualize, we further classified to 20 types such as
- **Biological Effects** - Fatigued/Drowsy, Lost Consciousness, Fell Asleep, Illness, Shoulders Defective/Improper, Physical Disability
- **Driver Incompetence** - Backing Unsafely, Passing or Lane Usage Improper, Turning Improperly, Unsafe Lane Changing, Driver Inexperience, Aggressive Driving/Road Rage, Eating or Drinking
- The above classification has been done via **spark sql** and analysed via **Plotly** to find the CF Factor types those played a vital role wrt each borough.

Results and Inferences

1. Pedestrian Data Inferences
2. Cyclist Data Inferences
3. Motorist Data Inferences
4. Vehicle Data Inferences
5. Contributing Factor Inferences

Pedestrian Inferences

1) Crossing with Signal - highest mentioned reason for pedestrians which means negligence of drivers is the root cause
2) Lack of insight on Right of Way is also a major contributing factor
3) Brooklyn had the highest number of accidents while Staten Island has the least
4) On normalising the accident count with population, Queens has the highest ratio
5) Highest number of accidents are happening in the timeframe of 4 PM - 7 PM

Cyclist Inferences

1) Brooklyn had the highest number of accidents happening over all other borough.
2) On normalising the count with population Manhattan is found the most dangerous borough for Cyclists.
3) Most number of delivery vehicles being used in Manhattan is one of the reason, while the complex street structure and busy traffic being other reasons
4) Similar to Pedestrian data highest number of accidents are happening in the timeframe of 4 PM - 7 PM
5) While the trend of death causing accidents has been oscillating over years, the number of injuries causing accidents has been increasing

Motorist Inferences

1) Surprisingly Staten Island had the highest ratio of accidents per population
2) Manhattan and Queens followed Staten Island respectively
3) Brooklyn and Bronx had the least ratio of accidents per population
4) Movement of large vehicles is highest in the borough of Queens, could be one of the reason of death and injury causing accidents
5) Cabs moving in and out of borough is highest in Manhattan borough, leading to highest number of accidents in the borough

Vehicle Inferences

1. Light Mode vehicles had the highest overall number of accidents, about 1.8M, which was 50% of the total vehicles present.
2. After the second analysis the above inference made sense, as we found Light mode vehicles colliding with each other or other pedestrians/obstacles as the highest number, around 600k collisions out of the 1.8M unique collisions, which contributed around 33%.
3. Contributing factors of unlicensed drivers showed that driver inattention or carelessness caused a major number of accidents, around 5800 which was 28% of the total unlicensed driver accidents.

Contributing Factor Inferences

1. Based on trend, **Manhattan**, the cause of accidents tops with the following Contributing Factors Driver Inattention, Driver Incompetence, Following closely, Other Vehicle Intervention and Biological Effects. If we deep dive into the above causes, as our analysis states the facts are that due to an increase in the population density, being the heart of the state the people who live here prefer to have personal vehicles to commute. Having more corporate offices and the world's biggest shopping malls may lead to attracting tourists and huge no of residents during the festive season.

2. The drivers who were responsible for accepting the Lyft services will always try to meet the expected drop-off time to satisfy their customers. To achieve that most of them will overlook the traffic signals, pedestrian rules and bumper-bumper traffic.

3. **Queens** - tops the list for the factors, **Right Way Rule** and the **Traffic Rule Avoidance.** We found that interesting because the population is lesser than Manhattan and Brooklyn but queens has most no of single way roads, motorist accidents exceeds Brooklyn and ignorance.

CF Classification Stats:

Below were the top 7 classification factors that were responsible for **69%** of over all accidents in NYC.

1. **Driver Inattention** (**M**>**B**>**Q**>**BX**>**S**)
2. **Driver Incompetence** (**M**>**B**>**Q**>**BX**>**S**)
3. **Following Too Closely** (**M**>**B**>**Q**>**S**>**BX**)
4. **Right Way Rule** (**Q**>**B**>**M**>**BX**>**S**)
5. **Other Vehicle Intervention** (**M**>**B**>**Q**>**BX**>**S**)
6. **Biological Effects** (**M**>**B**>**Q**>**BX**>**S**)
7. **Traffic Rule Avoidance** (**Q**>**B**>**M**>**BX**>**S**)

M- Manhattan, **B** - Brooklyn, **Q** - Queens, **BX** - Bronx, **S** - Staten Island

Challenges & Limitations

1) We tried to correlate the accident data with weather data in NYC. But the data we found was insufficient to provide any conclusions
2) There are few columns for which off street or cross street names mentioned which could be leveraged to fetch the borough, but it was too ambiguous to derive.
3) Grouping vehicles was tough with only a single given column (Vehicle_Type) so we had incorporate another column (Vehicle_Make) in order to group the vehicles better.
4) Lot of repeated reasons under the contributing factor column which made it difficult to classify without grouping them

Web Interface

https://projec.streamlit.app/

Above URL is our web interface containing all the interactive plots and analysis related to our project.

Github Repository: https://github.com/sabarishreddy99/BigData_CS-GY-6513_Fall2023_Proj

Thank You!