Data 603: Big Data Processing
Data integration and Predictive analysis on Motor Vehicle Collision in NYC
Professor : Najam Hassan

By :

Mohammed Abdul Rasheed Khan
Suhana sheik
Dixith Kumar Bandari
Lokesh Chava

Introduction

- We see many deaths related to accidents in day-to-day life.
- To understand the factors of these accidents we will delve deeper into the data collected in the New York city.
- There are three different datasets: crashes, vehicles and persons in Data.gov
- We will merge these datasets and look at meaningful insights of these datasets.
- We will visualize many scenarios such as when most of the accidents occur, what are the factors causing these accidents, at what time a greater number of accidents occur, and reasons for the accidents and fatalities causes in accidents.
- This data will be helpful for the city’s traffic polices to avoid accidents and take necessary measures to avoid accidents and reduce deaths.

Objectives

- The purpose of this project is to apply big data analytics on the dataset of New York state motor vehicle collisions.
- The aim is to provide a more comprehensive and dynamic framework for evaluating the collisions and their time frames with all the details about the crash, location vulnerabilities and estimating damages.
- Integrate the data from different datasets like collisions, vehicles and persons.
- Perform Data Integration, Predictive Analysis, Visualization, Reports stating the recommendation to the Government to avoid further incidents and the locations to focus on.

Dataset Overview

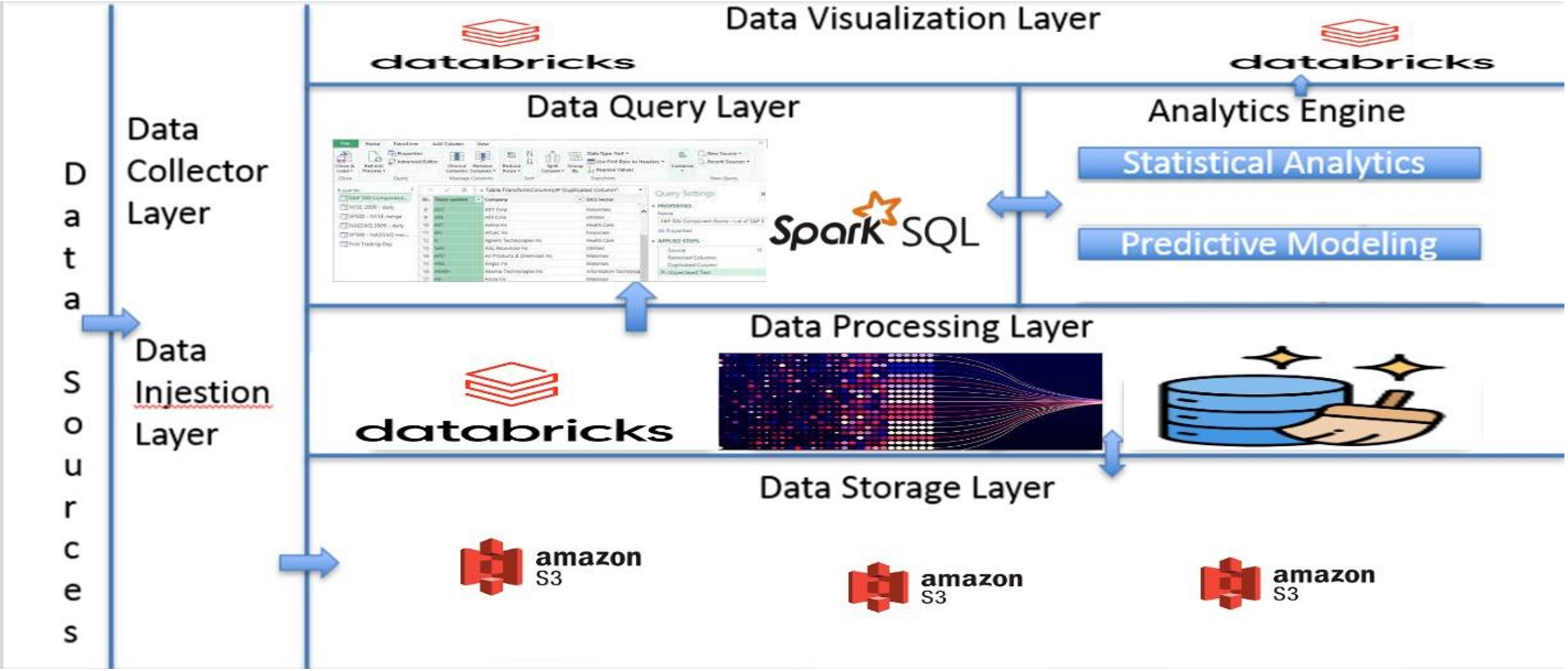
Data Source: https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes/resource/b5a431d2-4832-43a6-9334-86b62bdb033f.

The data set is fetched from “data.gov” which is a legitimate United states government website.

The data access is set to public and has different kinds of Datasets collected directly from the government organizations and private organizations functioning under the US government.

The data set has been updated lastly in September, 2023 and has over 2 million rows of data.

Tools and Technologies



Data Sources

Data Collector Layer

Data Injestion Layer

Data Visualization Layer

Data Query Layer

Analytics Engine

Statistical Analytics

Predictive Modeling

Data Processing Layer

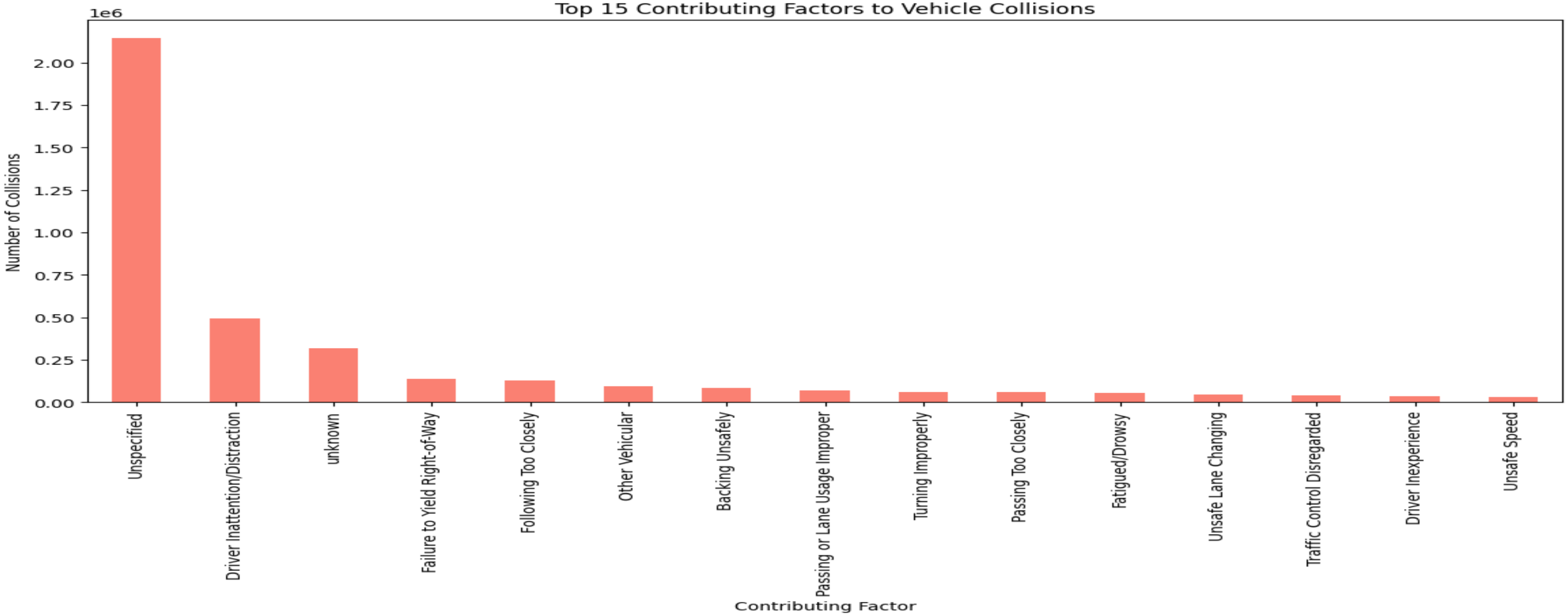
Data Storage Layer

- Implementation of cloud storage using amazon S3 database.
- Imputation of data sets into data bricks platform for data Pre - processing and integration.
- Apache spark session is used for data stimulation integration of datasets into a single data frame.
- Pandas, Seaborn and Sklearn Python packages are used for analysis and visualization of data.
- Databricks dashboards created to generate reports with meaningful insights from the analyzed data.

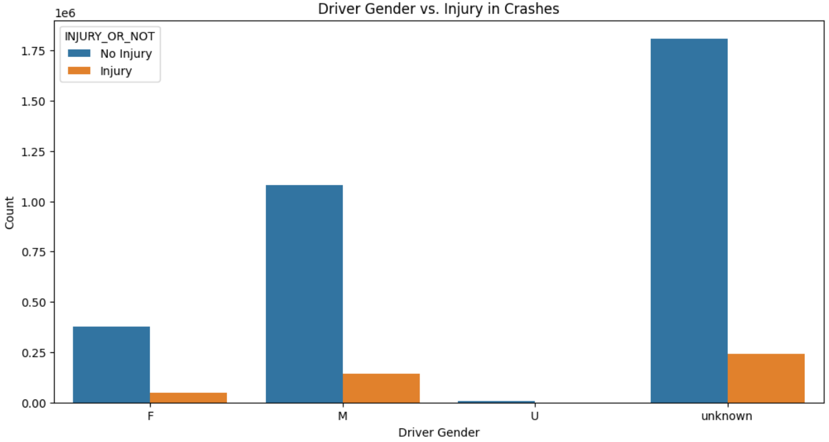
Insights and Recommendations

- In most of the crash's road occupants are main reason for the incidents followed by Bicyclists.
- Friday is the most accident occurring day among all days.
- 10PM to 12PM is the most accident happening time zone.
- Male stands as the primary culprit for crashes influenced by distractions, drugs, alcohol and aggressive driving
- Crashes with vehicle defects includes Accelerator defective as the primary reason followed Breaks defective
- Human factors include emotional influence, illness and hormonal imbalance
- Second part collision factors are right of the way, other vehicles using headphones mostly committed by the female group.

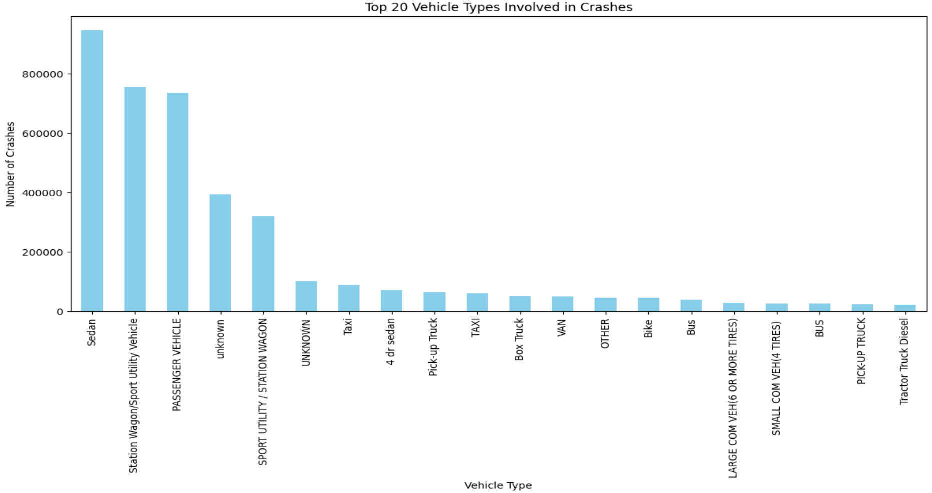
- Top vehicle models includes sedan, SUV and passenger vehicles
- Motorist persists more injuries at 72.8%, followed by pedestrians at 18.4% and cyclists at 8.8%
- Top zip codes are 11207, 11101, 11236
- 2020 has all time low accidents with a stable count of crashes from 2021 till date and 2016 to 2019 being the most accident happen period.



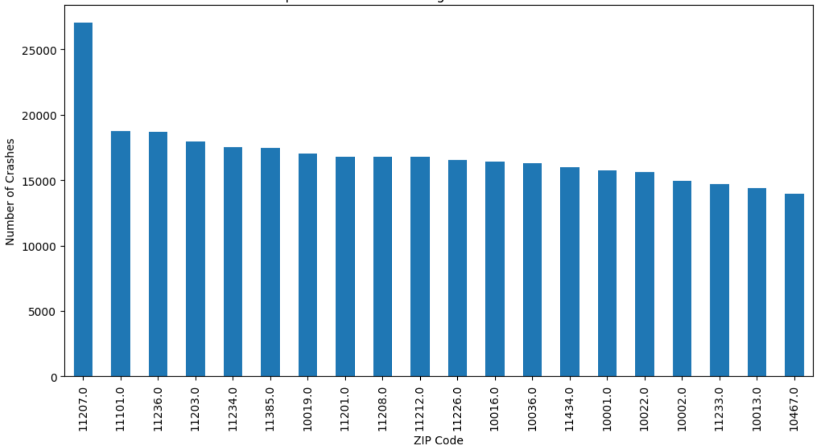
Driver Gender vs. Injury in Crashes

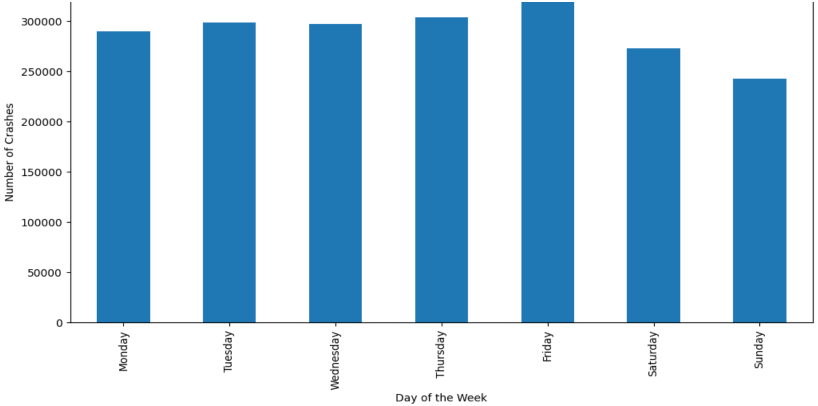


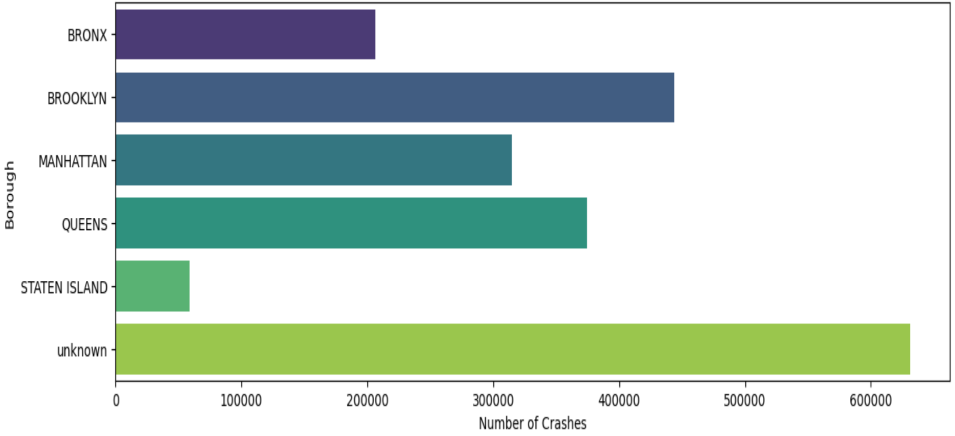
Top 20 Vehicle Types Involved in Crashes



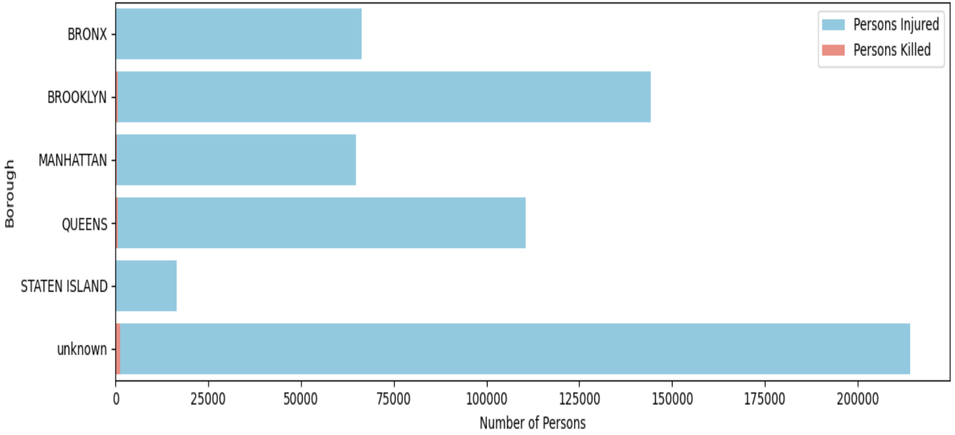
Top 20 ZIP Codes with Highest Number of Crashes



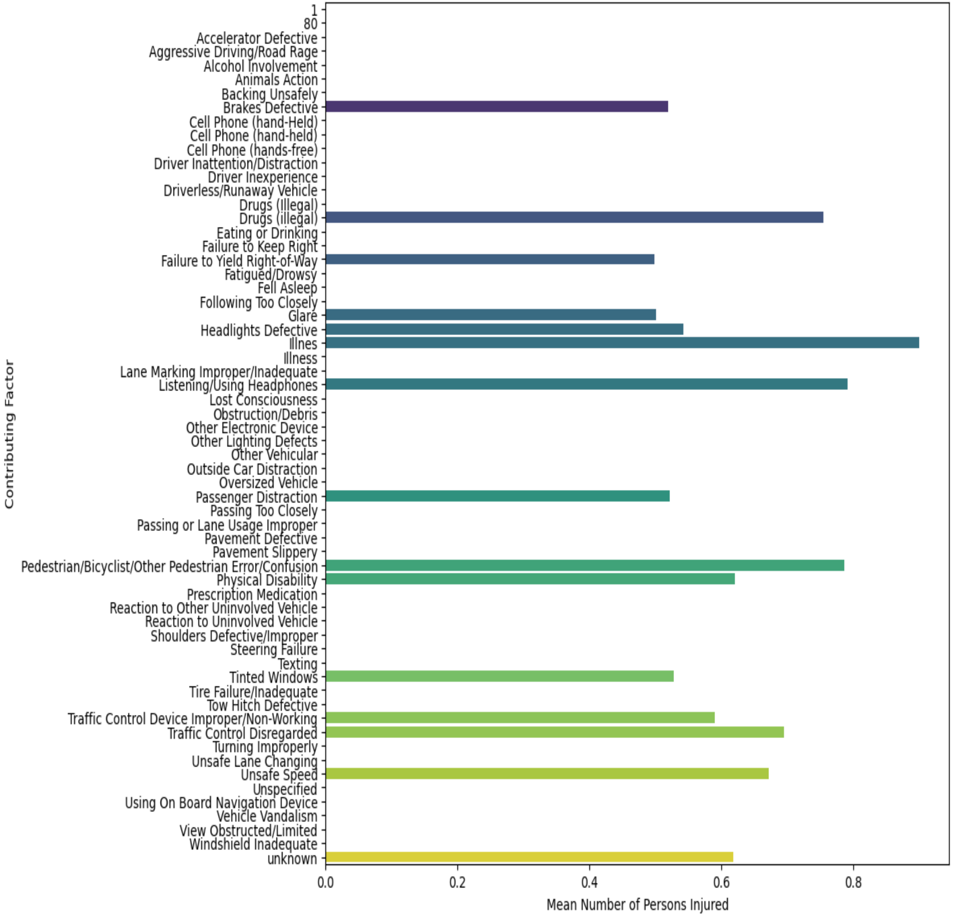




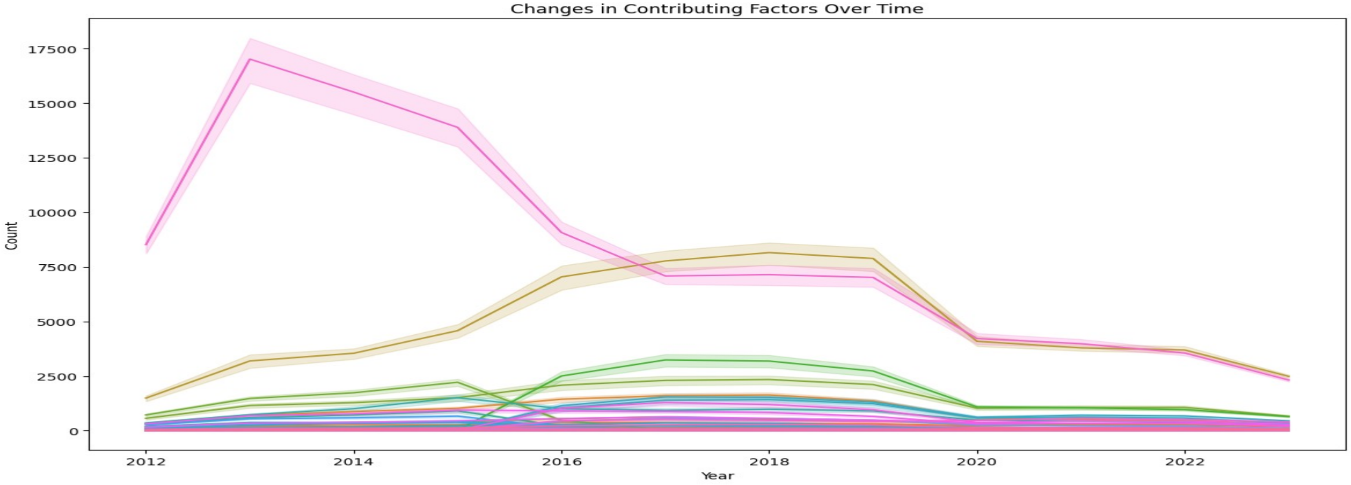
Total Number of Crashes Across Different Boroughs

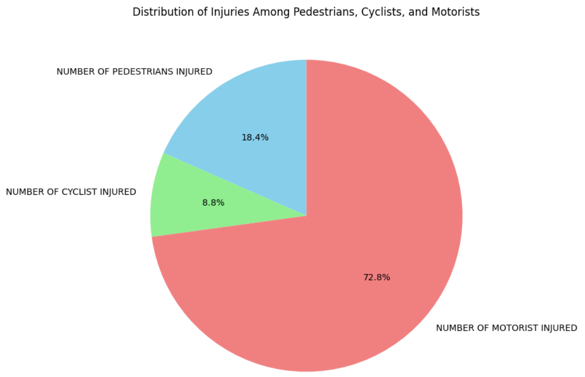


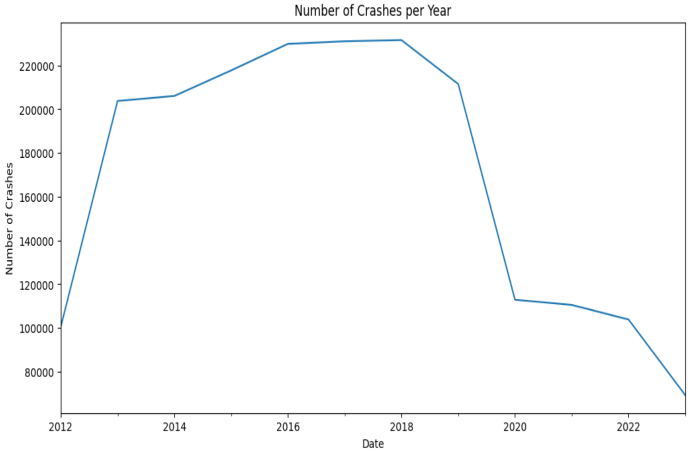
Average Number of Persons Injured or Killed Across Different Boroughs



Top 15 Contributing Factors and Mean Number of Persons Injured


Changes in Contributing Factors Over Time
Count
17500
15000
12500
10000
7500
5000
2500
0
2012
2014
2016
2018
2020
2022
Year
Contributing Factor
1
80
Accelerator Defective
Aggressive Driving/Road Rage
Alcohol Involvement
Animals Action
Backing Unsafely
Brakes Defective
Cell Phone (hand-Held)
Cell Phone (hand-held)
Cell Phone (hands-free)
Driver Inattention/Distraction
Driver Inexperience
Driverless/Runaway Vehicle
Drugs (Illegal)
Drugs (illegal)
Eating or Drinking
Failure to Keep Right
Failure to Yield Right-of-Way
Fatigued/Drowsy
Fell Asleep
Following Too Closely
Glare
Headlights Defective
Illnes
Illness
Lane Marking Improper/Inadequate
Listening/Using Headphones
Lost Consciousness
Obstruction/Debris
Other Electronic Device
Other Lighting Defects
Other Vehicular
Outside Car Distraction
Oversized Vehicle
Passenger Distraction
Passing Too Closely
Passing or Lane Usage Improper
Pavement Defective
Pavement Slippery
Pedestrian/Bicyclist/Other Pedestrian Error/Confusion
Physical Disability
Prescription Medication
Reaction to Other Uninvolved Vehicle
Reaction to Uninvolved Vehicle
Shoulders Defective/Improper
Steering Failure
Texting
Tinted Windows
Tire Failure/Inadequate
Tow Hitch Defective
Traffic Control Device Improper/Non-Working
Traffic Control Disregarded
Turning Improperly
Unsafe Lane Changing
Unsafe Speed
Unspecified
Using On Board Navigation Device
Vehicle Vandalism
View Obstructed/Limited
Windshield Inadequate
unknown


Distribution of Injuries Among Pedestrians, Cyclists, and Motorists
NUMBER OF PEDESTRIANS INJURED
18.4%
NUMBER OF CYCLIST INJURED
8.8%
72.8%
NUMBER OF MOTORIST INJURED


Number of Crashes per Year
Number of Crashes
220000
200000
180000
160000
140000
120000
100000
80000
2012
2014
2016
2018
2020
2022
Date

Future Opportunities

- Advanced Analytics for Prediction
- Integration with Autonomous Vehicles
- Real-Time Data Analysis
- Infrastructure Planning
- Improved Emergency Response
- Behavioral Analysis

Challenges

- Data Quality and Standardization
- Handling Big Data
- Privacy Concerns
- Interoperability
- Ethical Considerations
- Cybersecurity Risks

References

- *Motor vehicle collisions - Crashes - Comma separated Values file - catalog*. (n.d.). https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes/resource/b5a431d2-4832-43a6-9334-86b62bdb033f
- Data 603 class work and reference materials.

Thank You
Any question?