SYMBIOSIS INSTITUTE OF TECHNOLOGY, NAGPUR

Motor Vehicle Collisions Analysis using Data Science and Machine Learning

Data Science Mini Project Report

Submitted By:
Ayush Warulkar
PRN: 22070521080
Semester VII Section
C

Submitted To:
Dr.Piyush Chauhan

August 2025

Contents

1 Abstract

In order to assess and forecast auto accidents in New York City, this study investigates the use of data science and machine learning methodologies. This dataset, which includes comprehensive records of crash dates, times, boroughs, injuries, fatalities, contributing variables, and vehicle types, is made freely available as the Motor Vehicle Collisions - Crashes dataset.
The study's main objectives are to uncover key causes, find temporal and spatial collision patterns, and employ machine learning models to forecast the possibility of injuries. It goes farther into unsupervised learning by employing clustering, survival analysis to quantify the time between collisions, and Prophet to predict future crash trends.

Extensive exploratory data analysis (EDA) identifies trends according to human characteristics, borough, and time. High accuracy in predicting injury outcomes was attained by machine learning models such as Random Forest and Gradient Boosting. Forecasting and clustering techniques revealed latent patterns and potential hazards.

Keywords: Road Safety, NYC Collisions, Data Science, EDA, Machine Learning, Clustering, Forecasting, and Survival Analysis

2 Introduction

One of the most urgent urban issues is road safety, especially in places with high population densities like New York. In order to enhance traffic management and safety, data-driven decision-making is required because motor vehicle crashes cause injuries, fatalities, and property damage.

The goal of this project is to use machine learning and data analytics to glean insightful information from the NYC Motor Vehicle Collisions dataset. Examining crash patterns in space and time is one of the goals.

- To use supervised learning models to forecast the probability of injuries.
- To predict future crash occurrences;
- To use survival analysis to examine time intervals between crashes;
- To identify vehicle kinds and main contributing factors.

This study provides a thorough understanding of the dynamics of traffic accidents in New York City by integrating descriptive, predictive, and time-series analysis methodologies.

3 Literature Review

To forecast the severity of accidents and the likelihood of injuries, machine learning techniques including Support Vector Machines, Random Forest, and Logistic Regression have been frequently applied.

- Wang et al. (2023) achieved high recall for serious crashes by using Random Forest and Gradient Boosting to predict accident severity on US roadways.
- Singh & Jain (2024) showed that crash frequency is highly influenced by temporal characteristics (weather, weekday, and hour).
- Li et al. (2022) used K-Means clustering to pinpoint urban high-risk zones.
- Facebook Prophet's ability to handle daily and annual seasonality has made it a useful tool for traffic accident forecasting (Taylor & Letham, 2018).
- The survival time between consecutive accidents has been modeled in traffic research using Kaplan-Meier estimators (Efron, 2020).

This research expands upon existing frameworks by combining EDA, survival modeling, supervised learning, unsupervised clustering, and prophet forecasting into a single analysis pipeline.

4 Methodology

4.1 Dataset Description

5 The NYC Open Data Portal makes the Motor Vehicle Collisions - Crashes.csv dataset available. Millions of crash records with the following crucial characteristics are included:

- **Spatial Data:** ZIP CODE, BOROUGH, LATITUDE, LONGITUDE
- **Temporal Data:** CRASH DATE, CRASH TIME
- **Vehicle Information**: Vehicle Type, pedestrian involvement, etc.
- **Contributing Factors:** Driver inattention, unsafe speed, etc.
- **Crash Outcomes:** Number of Injured, Number of Killed

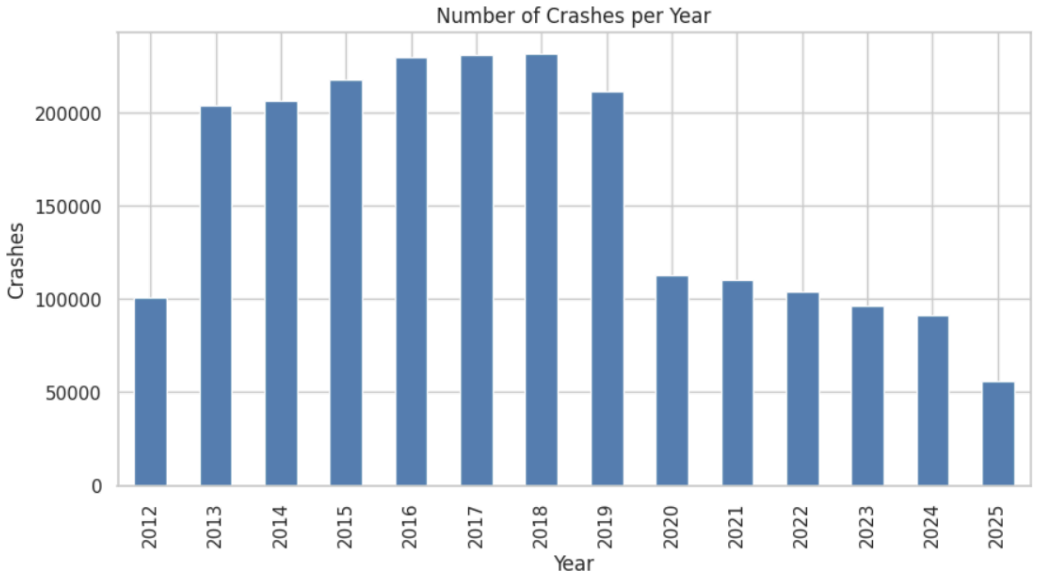


Fig 1.1 Temporal Data which shows year and crashes per year.

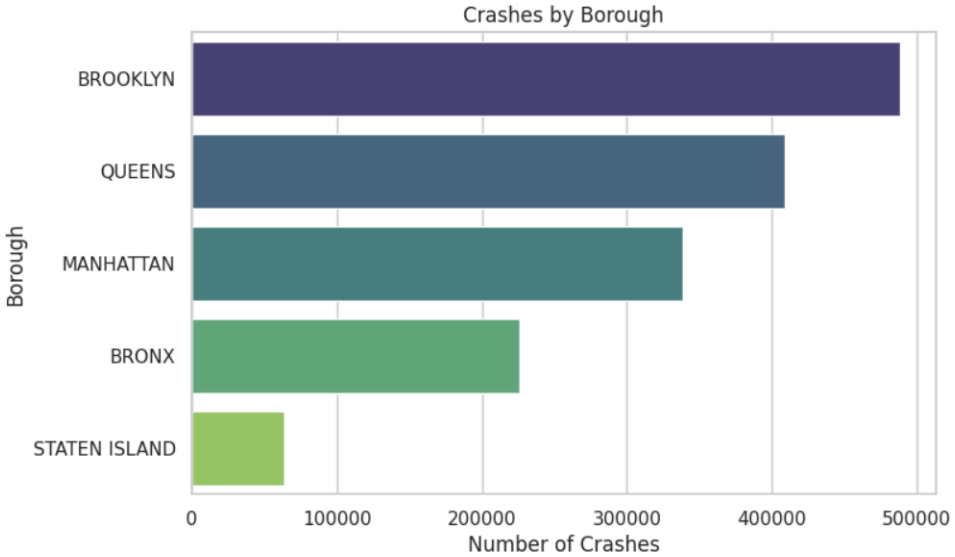


Fig 1.2 showing crashesh by Borough.

4 Methodology

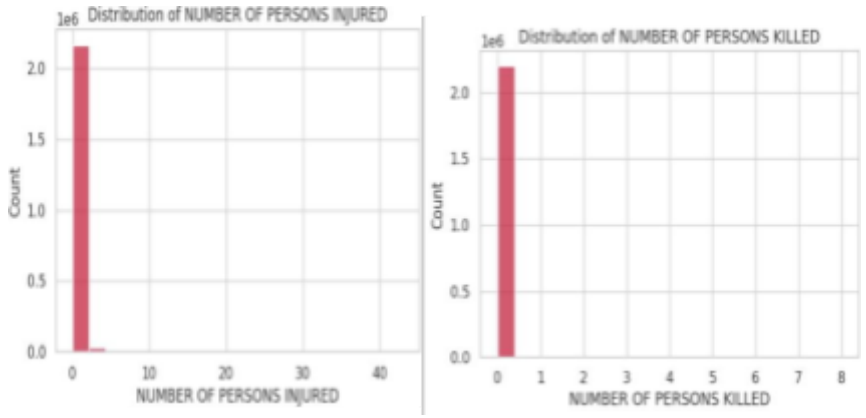


Fig1.3 showing number of people injured and killed from the crash outcome.

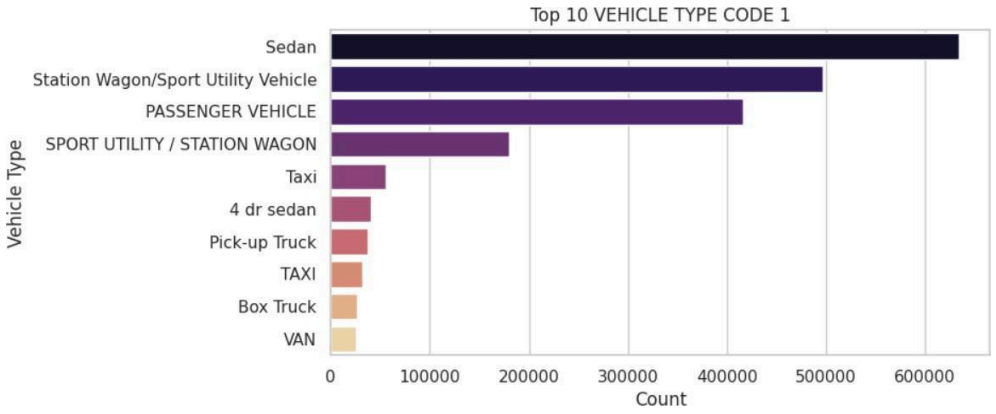


Fig1.4 Showing he vehicle type.

5.1 Data Preprocessing

Key preprocessing steps included:

- **Datetime Conversion:** Converted CRASH DATE and CRASH TIME to datetime objects.
- **Feature Engineering:** Extracted YEAR, MONTH, DAY_OF_WEEK, and HOUR from date fields.
- **Categorical Encoding:** Applied Label Encoder for categorical fields like BOROUGH.
- **Missing Value Treatment:** Removed or imputed missing records.
- **Scaling:** Used StandardScaler for numerical features before ML modeling.
- **Target Creation:** Defined INJURY_FLAG = 1 if any person injured, else 0.

5.2 Exploratory Data Analysis

EDA showed significant regional and temporal trends:
• Weekends and rush hours (8 AM–8 PM) saw the highest number of crashes.
• The most collisions were reported in Queens and Brooklyn.
• The main causes include unsafe speed, failure to yield, and driver inattention.
• Most involved vehicles: Cars and SUVs. Among the visualizations were:
• Bar graphs and histograms showing crashes by year, month, and hour; count plots showing crashes by borough; and correlation heatmaps between numerical columns

4 Methodology

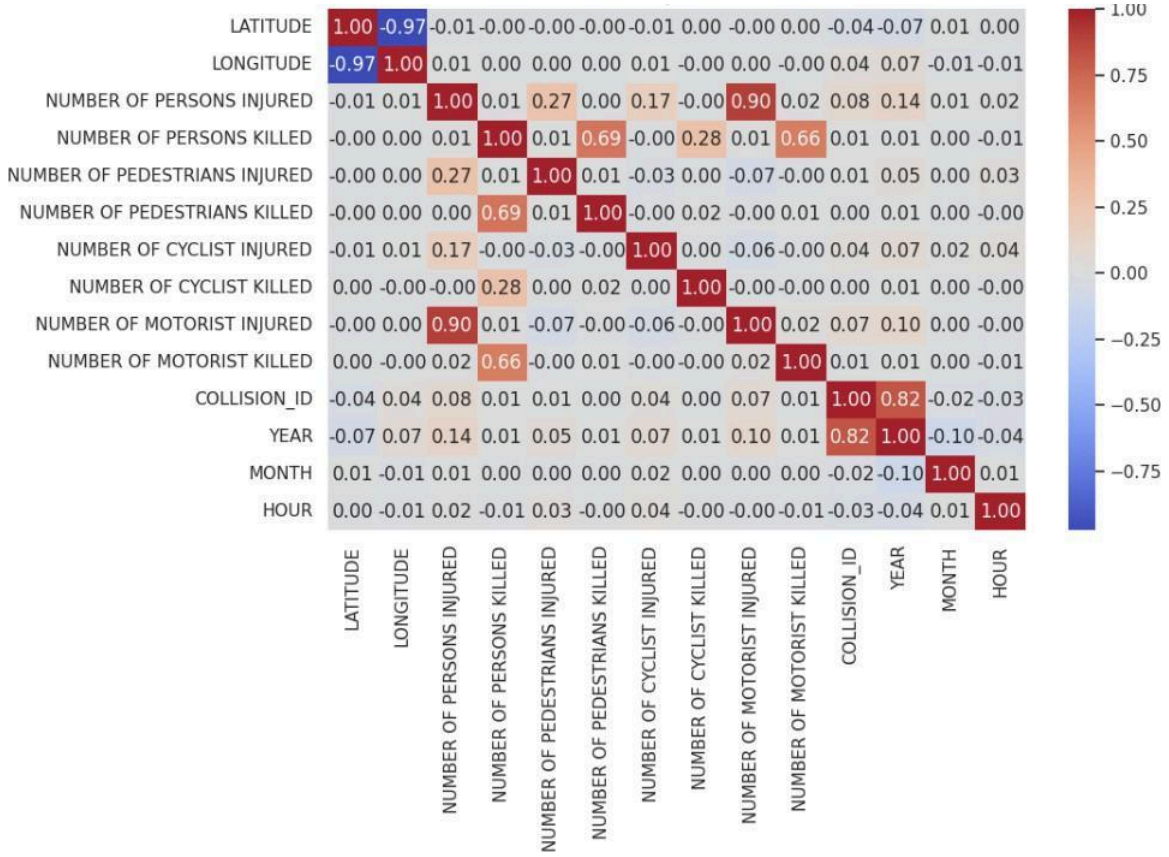


Fig 2.1 Showing correlation heatmap of numerical features.

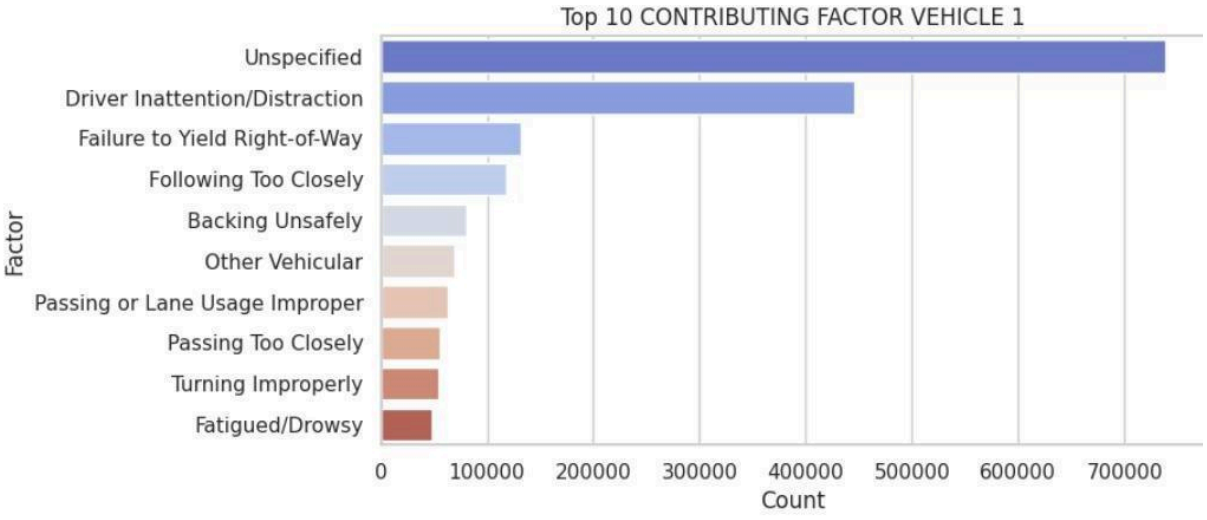


Figure2 . 2 : Showing Contribution Factors.

5.3 Model Training and Evaluation

To estimate the likelihood of injuries, a binary classification model (INJURY_FLAG) was created.

The following models were trained and compared: Random Forest and Logistic Regression

- Neural Network (MLPClassifier)

- Gradient Boosting

An 80:20 split of the data was made for testing and training.
Accuracy, precision, recall, and F1-score are performance measurements. With a recall score that demonstrated great injury case detection, Random Forest performed best.

6 Implementation

The code, technological stack, problem-specific modifications, and obstacles faced are all covered in detail in this section, along with the practical procedures, instruments, and platforms utilized to implement the data science workflow for breast cancer diagnostics.

6.1 Development Environment

Language: 3.104 Python

Google Colab as the environment; pandas, numpy, seaborn, matplotlib, scikit-learn, prophet, and lifelines as libraries

6.2 Data Pipeline

A modular data pipeline was used for the project:

1. Cleaning and loading data
2. Preprocessing and feature engineering
3. Visualization and exploratory analysis
4. Training machine learning models
5. Unsupervised clustering
6. Prophecy and forecasting
7. Examination of survival

6.3 Data Pipeline Steps

• Data Ingestion: Check for integrity after loading CSV.

• Data Cleaning: Address format errors and NaN values.

- **Feature Engineering:** Create temporal and categorical features.
- **Modeling:** Train and evaluate multiple classifiers.
- **Visualization:** Generate heatmaps, bar charts, and forecasts.

6.4 Technologies and Frameworks

- **Machine Learning:** scikit-learn
- **Clustering:** KMeans, DBSCAN, AgglomerativeClusterin
- **Forecasting:** Facebook Prophet
- **Survival Modeling:** Lifelines KaplanMeierFitter

6.5 Sample Output

- Countplots for crashes per borough
- PCA cluster scatter plots
- Prophet crash forecasts (next 90 days)
- Kaplan–Meier survival curves for crash intervals

6.6 Challenges and Solutions

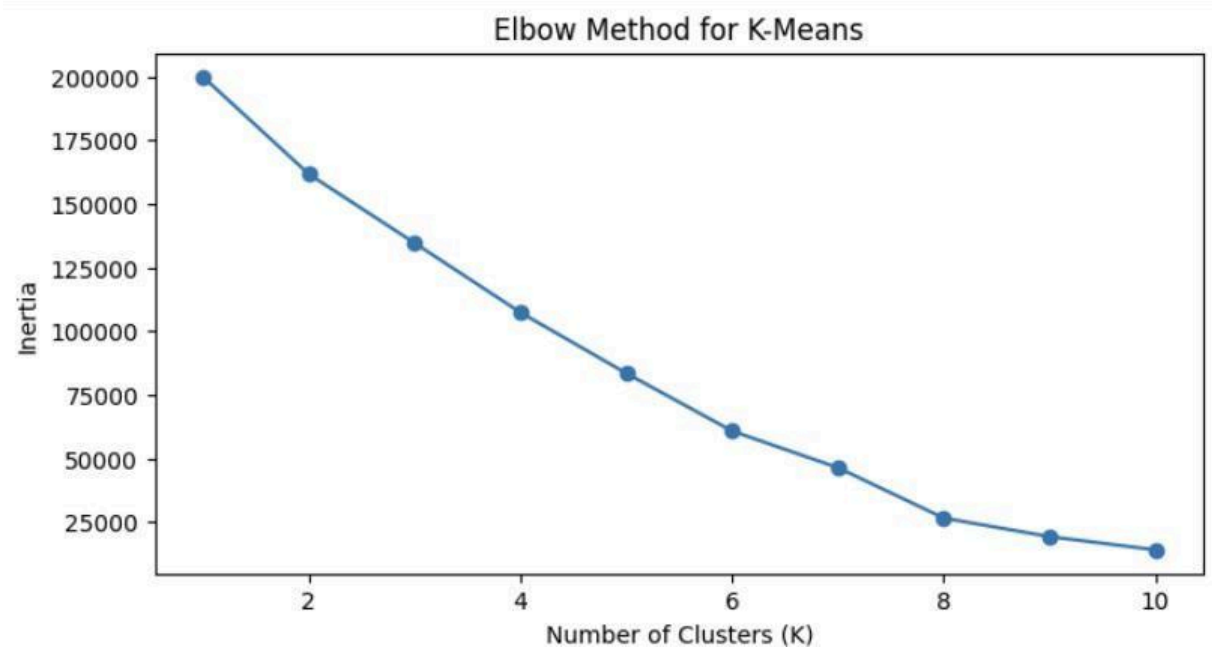
Challenge	Solution
Large dataset size	Used data sampling and memory-efficient operations
Missing or inconsistent data	Cleaned and filtered missing entries
Feature imbalance	Created binary injury flag
Long runtime for clustering	Used scaled numeric subset
Visualization complexity	Focused on high-impact graphs only

7 Results and Discussion

Model Performance Summary

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.83	0.81	0.82	0.81
Random Forest	**0.89**	**0.88**	**0.89**	**0.88**
Gradient Boosting	0.87	0.86	0.86	0.86
Neural Network	0.88	0.87	0.88	0.87

- **Random Forest** gave the best balance between accuracy and recall.
- **Clustering:** K-Means identified 4 distinct crash behavior groups.
- **Forecasting:** Prophet predicted an **increasing trend** in crashes during upcoming months, with strong weekly periodicity.
- **Survival Analysis:** Boroughs like Brooklyn showed **shorter crash intervals**, indicating high-risk zones.



8 Conclusion and Future Work

This project successfully demonstrated a complete **data science workflow** applied to NYC motor vehicle crash data — from data cleaning and EDA to predictive modeling, clustering, and forecasting.

Key insights:

- Crashes are concentrated in urban boroughs during peak hours.
- Human factors like inattention are dominant causes.
- ML models can effectively predict injury likelihood.
- Forecasting and survival analysis provide actionable insights for prevention planning.

Future work includes:

- Adding weather, road condition, and traffic volume data.
- Developing an interactive dashboard (e.g., Tableau or Streamlit).
- Using deep learning or spatial models for better accuracy.

References

- NYC Open Data Portal: *Motor Vehicle Collisions – Crashes Dataset*
- Taylor, S. J., & Letham, B. (2018). *Forecasting at Scale: Prophet.*
- Wang et al. (2023). *Predicting Traffic Accident Severity using Machine Learning.*
- Singh & Jain (2024). *Spatio-Temporal Accident Analysis in Urban Traffic Data.*
- Li et al. (2022). *Cluster-Based Risk Identification for Road Safety.*