SAFE ROADS INITIATIVE

Motor Vehicle Collision Analysis

Presented by Ayush Warulkar

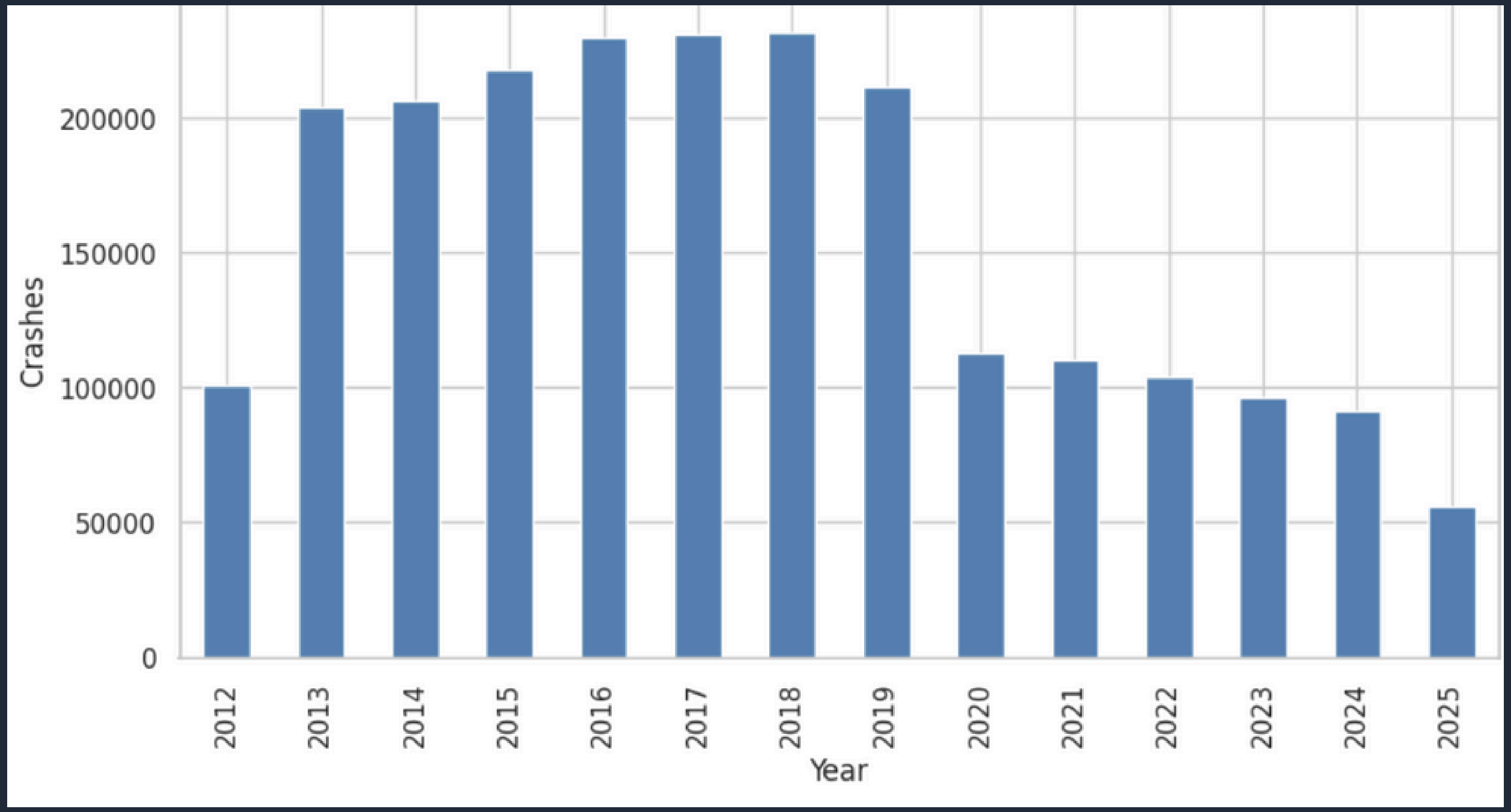
Introduction to Collision Analysis

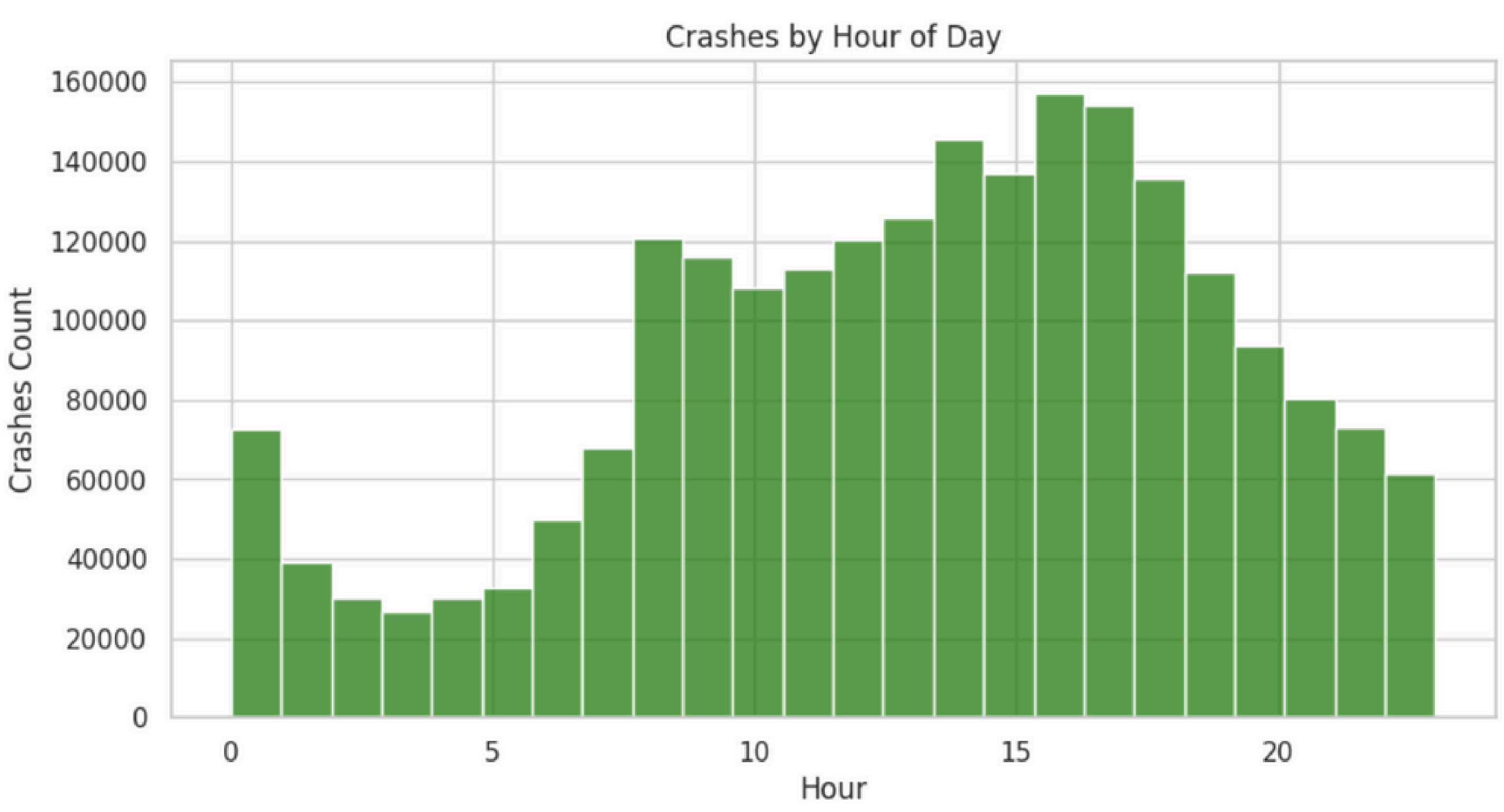
NYC has thousands of motor vehicle crashes annually.

Understanding crash patterns can reduce injuries and fatalities.

Data science provides tools to analyze and forecast collisions.

The goal is to extract insights and predict accident risks.





Objectives of the Project

To explore NYC crash data using Exploratory Data Analysis (EDA).

To identify key factors contributing to accidents.

To analyze temporal and spatial trends in crashes.

To predict injury likelihood using machine learning models.

To forecast future crashes using time series forecasting.

Dataset Description

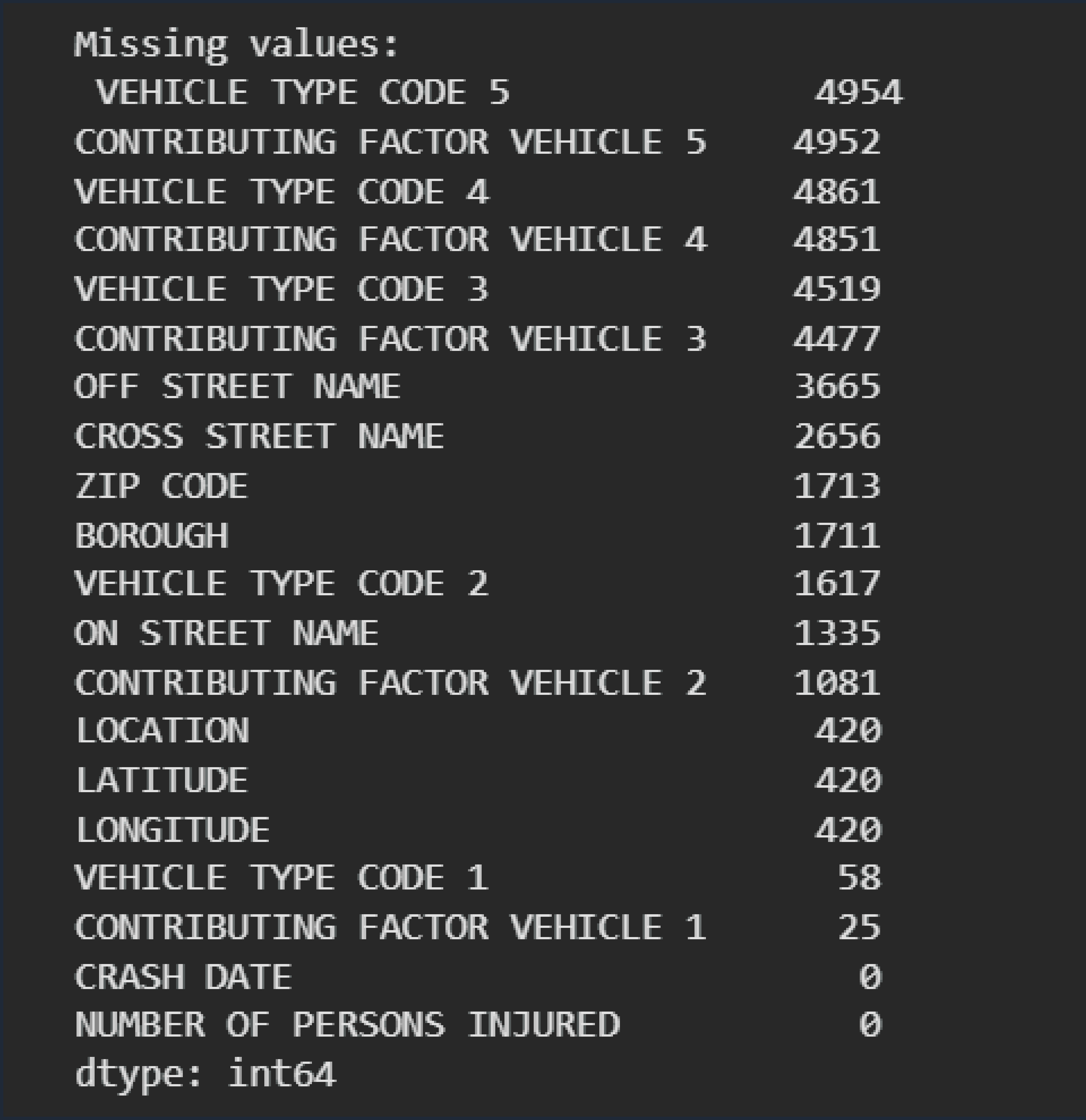
Dataset: NYC Motor Vehicle Collisions – Crashes (from NYC Open Data)

Records: ~1.8 million+ entries

Columns: Date, Time, Borough, Vehicle Type, Contributing Factors, Injuries, Fatalities, etc.

Time Period: 2012 to present

Data Source: data.gov



```
Missing values:
 VEHICLE TYPE CODE 5              4954
CONTRIBUTING FACTOR VEHICLE 5    4952
VEHICLE TYPE CODE 4              4861
CONTRIBUTING FACTOR VEHICLE 4    4851
VEHICLE TYPE CODE 3              4519
CONTRIBUTING FACTOR VEHICLE 3    4477
OFF STREET NAME                  3665
CROSS STREET NAME                2656
ZIP CODE                         1713
BOROUGH                          1711
VEHICLE TYPE CODE 2              1617
ON STREET NAME                   1335
CONTRIBUTING FACTOR VEHICLE 2    1081
LOCATION                          420
LATITUDE                          420
LONGITUDE                         420
VEHICLE TYPE CODE 1                58
CONTRIBUTING FACTOR VEHICLE 1      25
CRASH DATE                          0
NUMBER OF PERSONS INJURED           0
dtype: int64
```

Data Preprocessing & Feature Engineering

Handled missing values and removed duplicates.

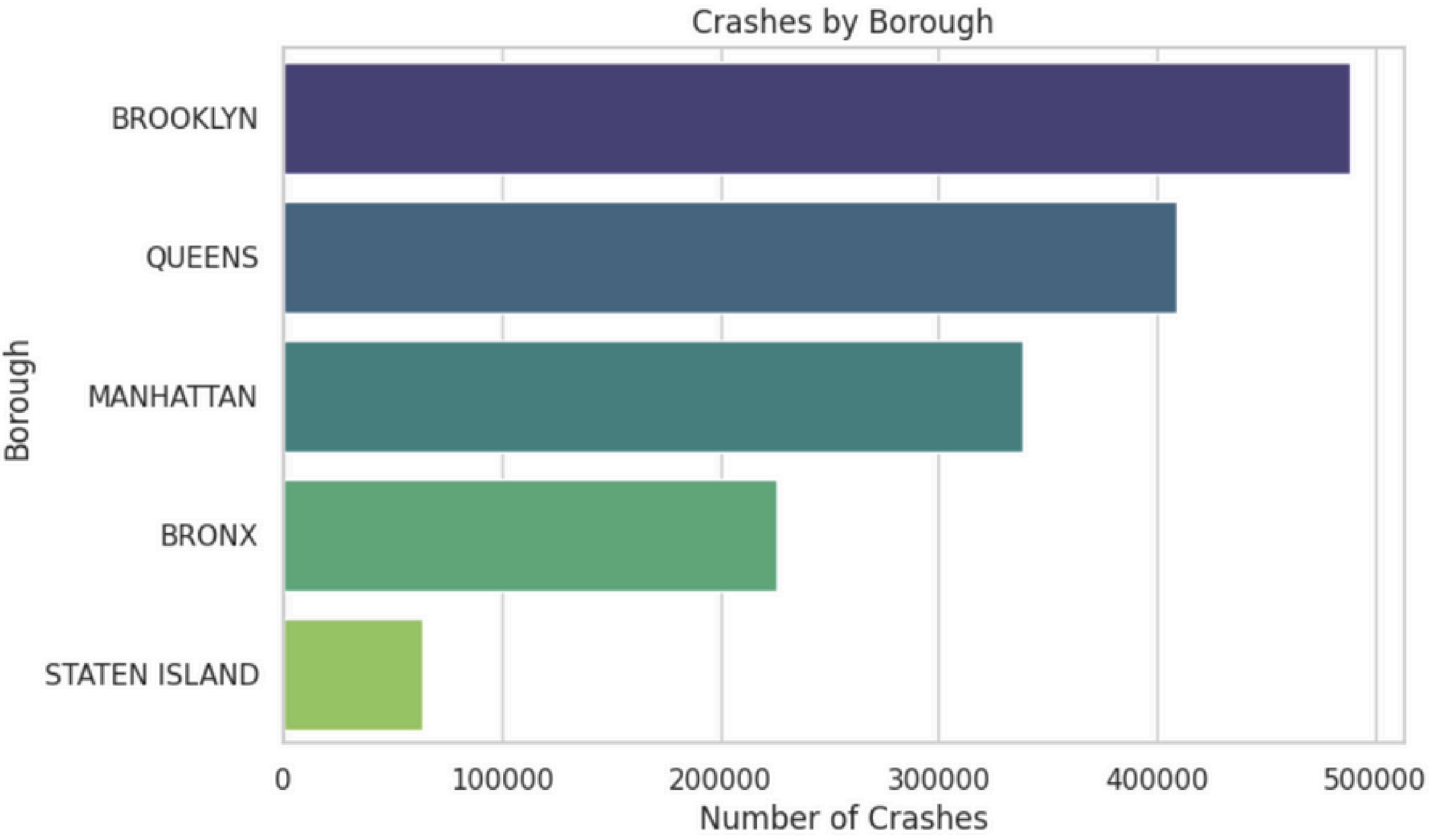
Converted CRASH_DATE and CRASH_TIME into datetime format.

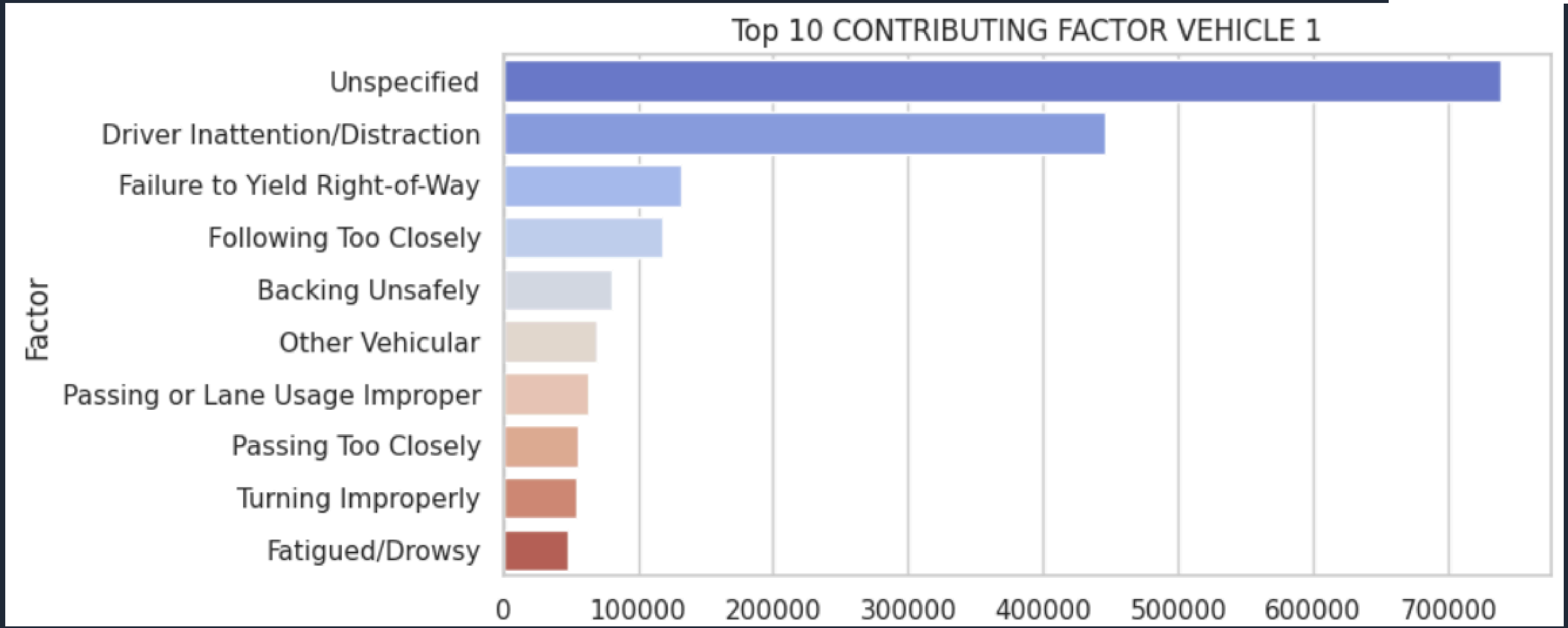
Extracted Year, Month, Hour, and Day of Week.

Standardized categorical fields like Borough and Vehicle Type.

Exploratory Data Analysis (Part 1)

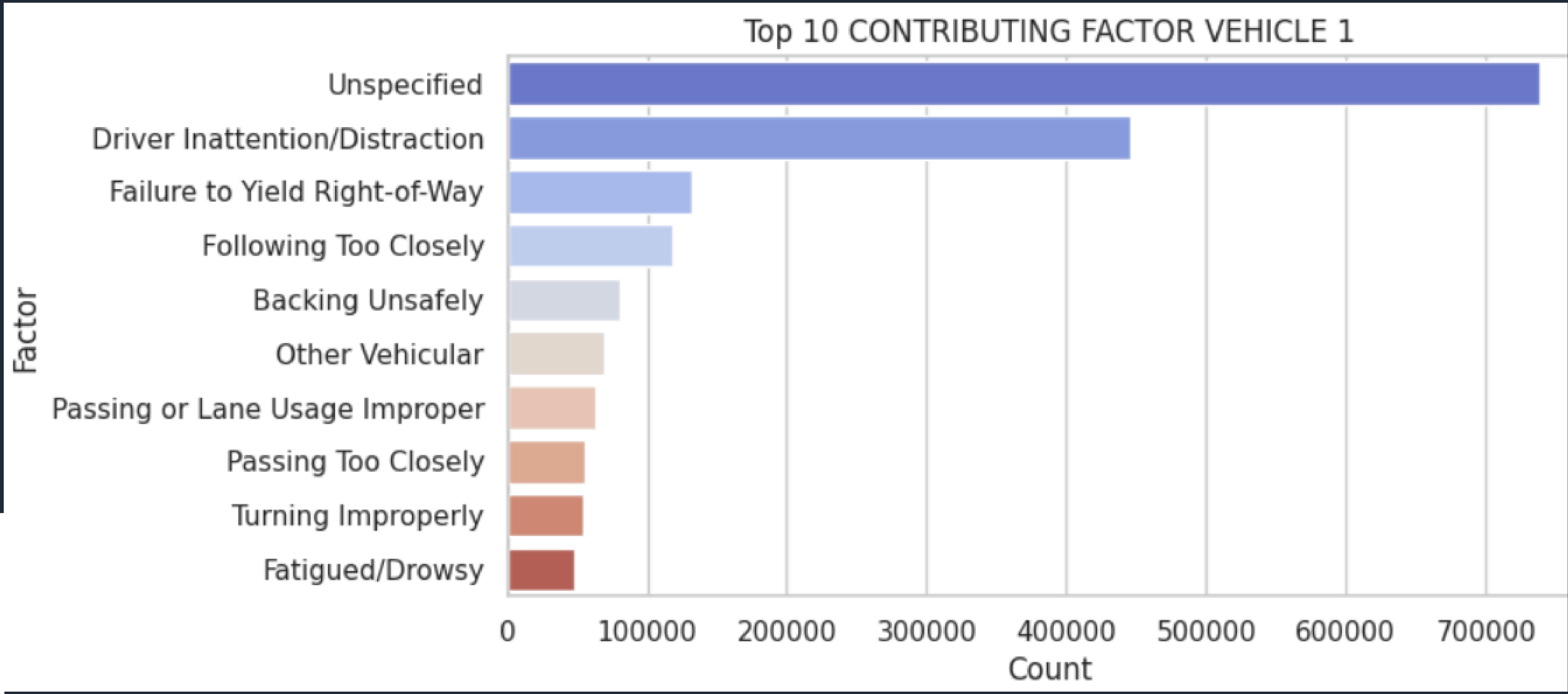
- Crash count distribution by Borough.
- Hourly and Weekly patterns.
- Most common contributing factors.

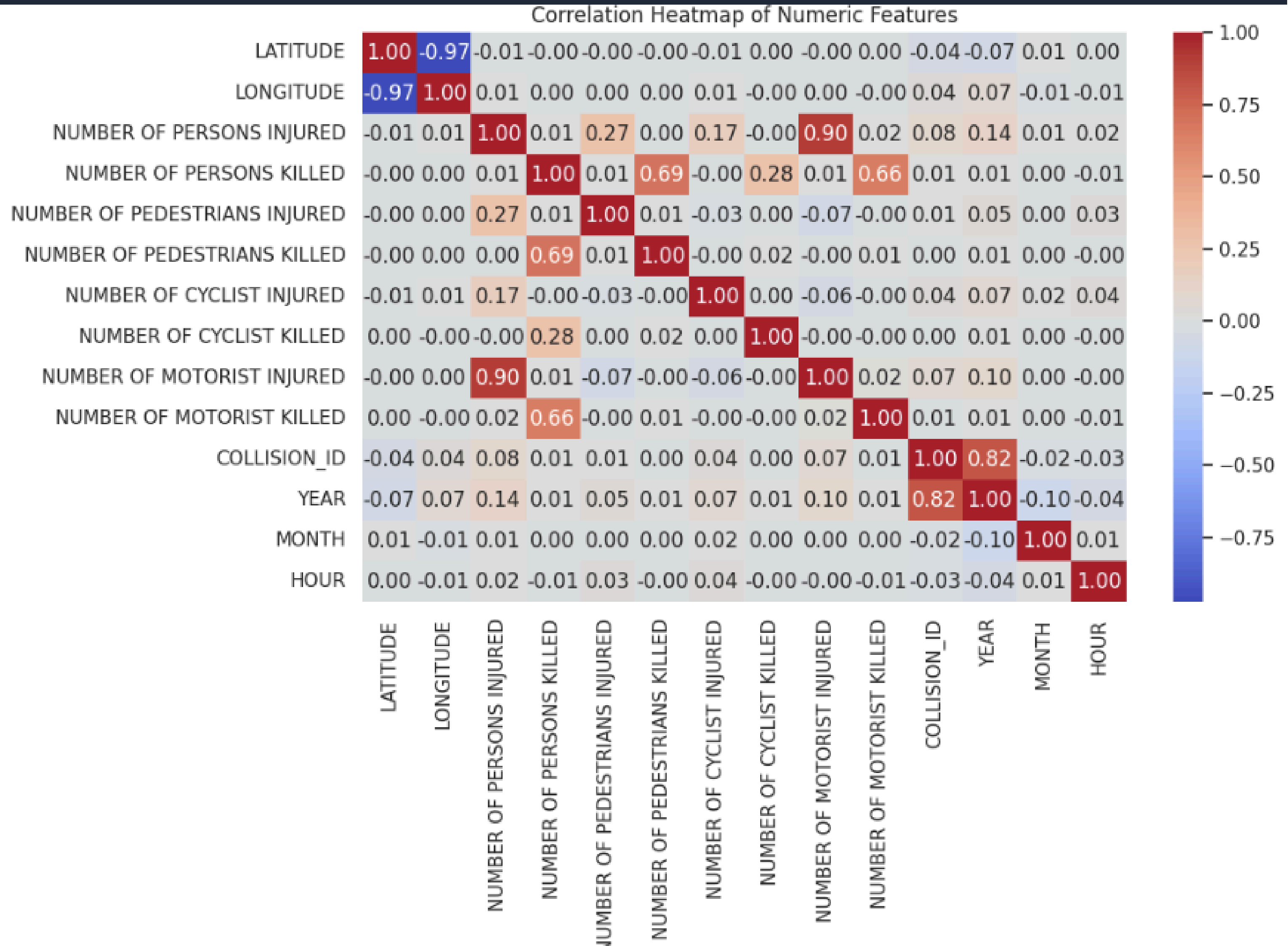




Exploratory Data Analysis (Part 2)

- Vehicle types vs injury severity.
- Correlation among numeric fields.
- Identifying relationships using heatmap.




Correlation Heatmap of Numeric Features

	LATITUDE	LONGITUDE	NUMBER OF PERSONS INJURED	NUMBER OF PERSONS KILLED	NUMBER OF PEDESTRIANS INJURED	NUMBER OF PEDESTRIANS KILLED	NUMBER OF CYCLIST INJURED	NUMBER OF CYCLIST KILLED	NUMBER OF MOTORIST INJURED	NUMBER OF MOTORIST KILLED	COLLISION_ID	YEAR	MONTH	HOUR
LATITUDE	1.00	-0.97	-0.01	-0.00	-0.00	-0.00	-0.01	0.00	-0.00	0.00	-0.04	-0.07	0.01	0.00
LONGITUDE	-0.97	1.00	0.01	0.00	0.00	0.00	0.01	-0.00	0.00	-0.00	0.04	0.07	-0.01	-0.01
NUMBER OF PERSONS INJURED	-0.01	0.01	1.00	0.01	0.27	0.00	0.17	-0.00	0.90	0.02	0.08	0.14	0.01	0.02
NUMBER OF PERSONS KILLED	-0.00	0.00	0.01	1.00	0.01	0.69	-0.00	0.28	0.01	0.66	0.01	0.01	0.00	-0.01
NUMBER OF PEDESTRIANS INJURED	-0.00	0.00	0.27	0.01	1.00	0.01	-0.03	0.00	-0.07	-0.00	0.01	0.05	0.00	0.03
NUMBER OF PEDESTRIANS KILLED	-0.00	0.00	0.00	0.69	0.01	1.00	-0.00	0.02	-0.00	0.01	0.00	0.01	0.00	-0.00
NUMBER OF CYCLIST INJURED	-0.01	0.01	0.17	-0.00	-0.03	-0.00	1.00	0.00	-0.06	-0.00	0.04	0.07	0.02	0.04
NUMBER OF CYCLIST KILLED	0.00	-0.00	-0.00	0.28	0.00	0.02	0.00	1.00	-0.00	-0.00	0.00	0.01	0.00	-0.00
NUMBER OF MOTORIST INJURED	-0.00	0.00	0.90	0.01	-0.07	-0.00	-0.06	-0.00	1.00	0.02	0.07	0.10	0.00	-0.00
NUMBER OF MOTORIST KILLED	0.00	-0.00	0.02	0.66	-0.00	0.01	-0.00	-0.00	0.02	1.00	0.01	0.01	0.00	-0.01
COLLISION_ID	-0.04	0.04	0.08	0.01	0.01	0.00	0.04	0.00	0.07	0.01	1.00	0.82	-0.02	-0.03
YEAR	-0.07	0.07	0.14	0.01	0.05	0.01	0.07	0.01	0.10	0.01	0.82	1.00	-0.10	-0.04
MONTH	0.01	-0.01	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	-0.02	-0.10	1.00	0.01
HOUR	0.00	-0.01	0.02	-0.01	0.03	-0.00	0.04	-0.00	-0.00	-0.01	-0.03	-0.04	0.01	1.00

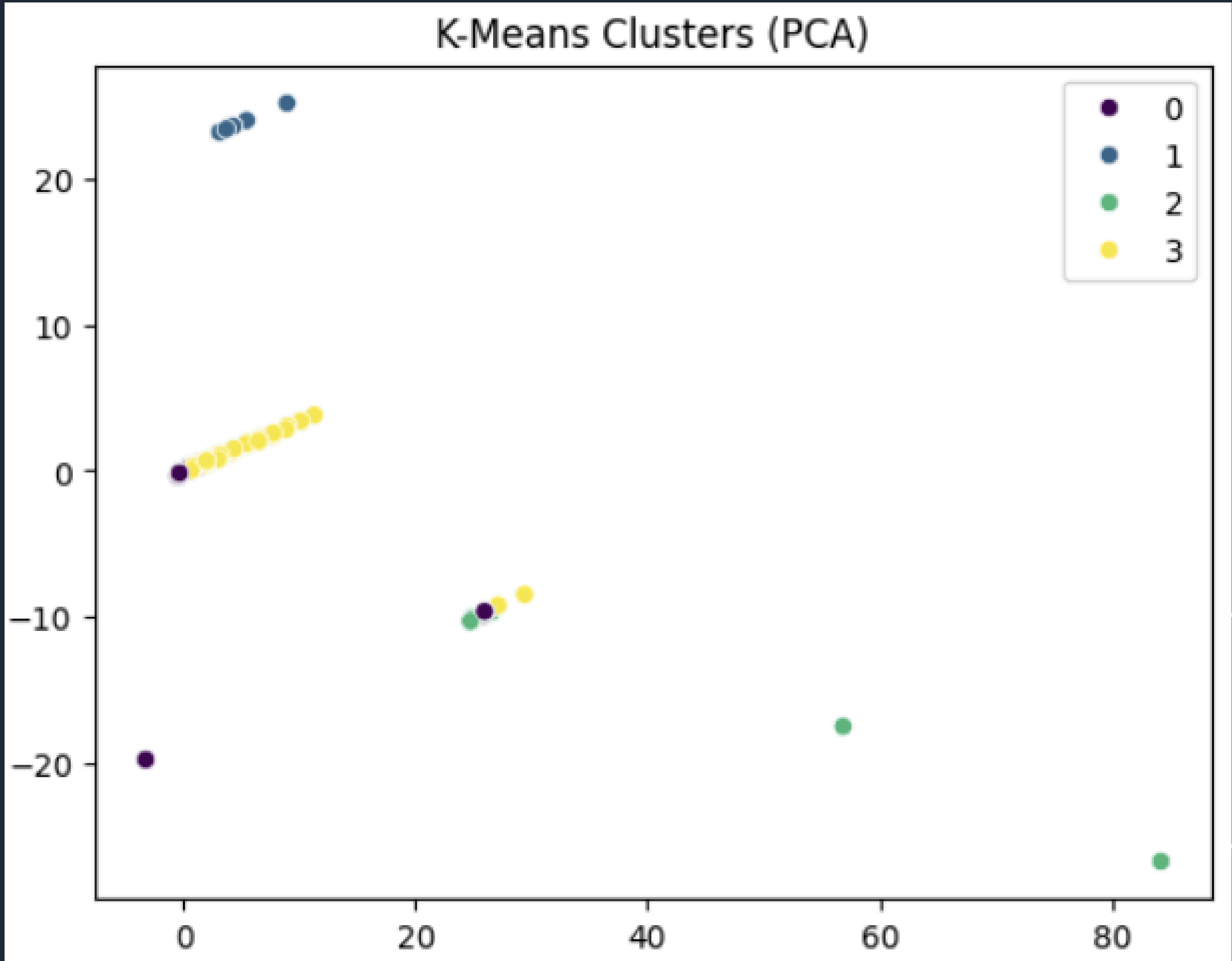
Machine Learning & Clustering

Created binary variable: Injury Occurred (Yes/No).

Models used: Logistic Regression, Random Forest, Gradient Boosting.

Best model: Random Forest (highest accuracy).

K-Means Clustering identified 4 crash behavior clusters.

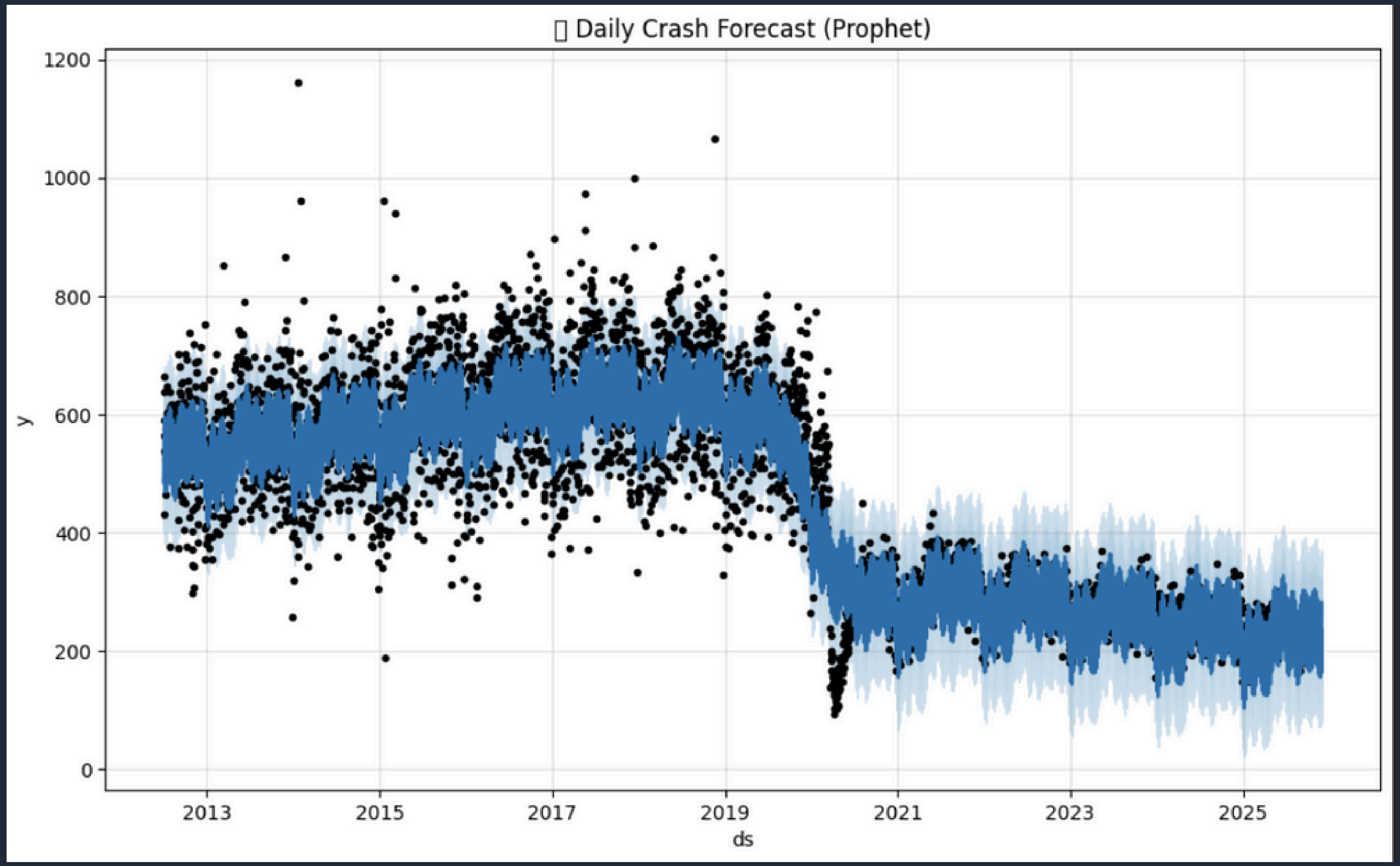


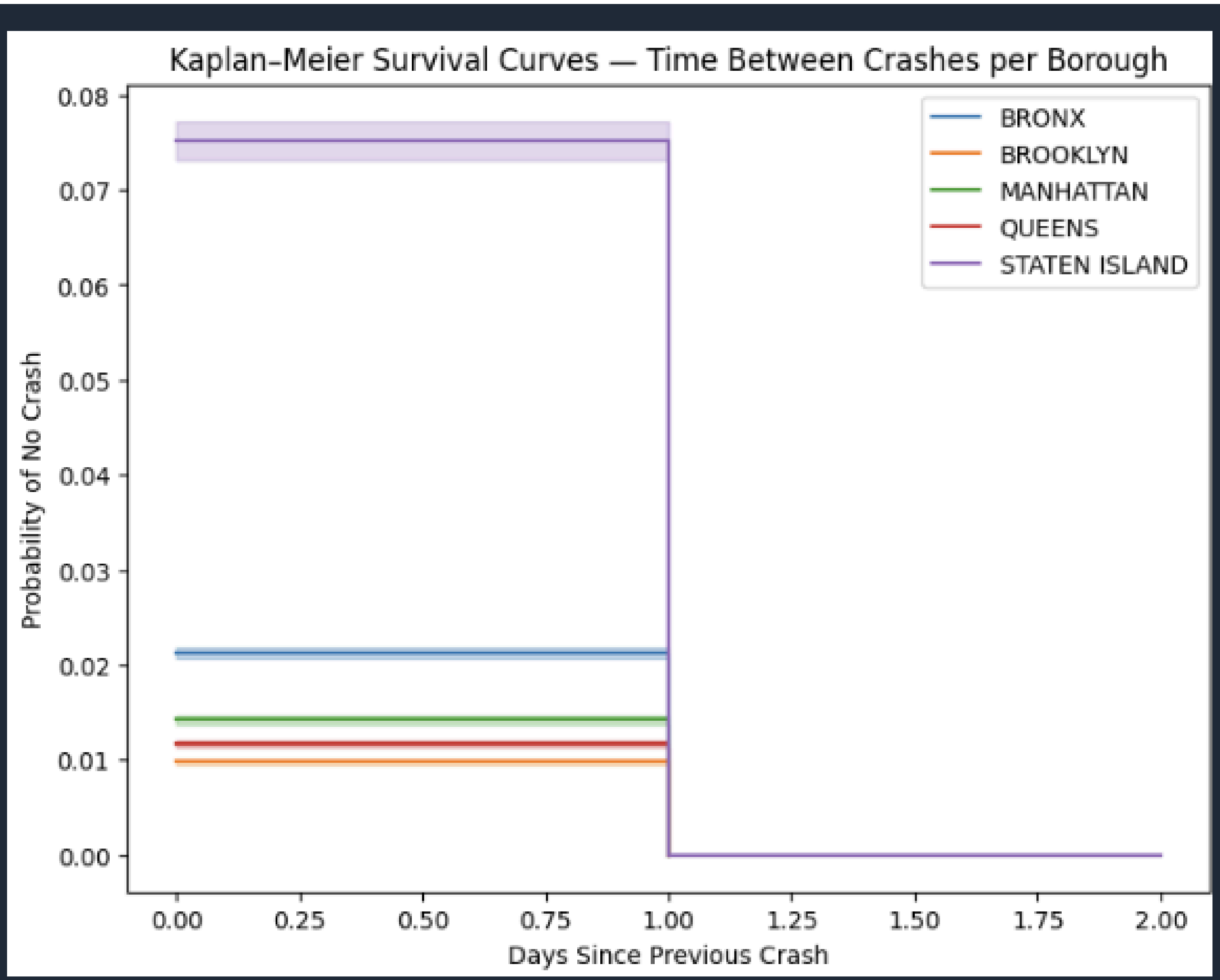
Forecasting & Survival Analysis

Forecasted crash trends using Facebook Prophet.

Captured weekly and yearly seasonality patterns.

Survival analysis (Kaplan–Meier) used to estimate time between crashes per borough.





Conclusion

Brooklyn & Queens are most accident-prone.

Rush hours and weekends show more crashes.

Driver inattention and speeding are major causes.

ML & Forecasting provide data-driven insights.

THANK YOU