Exploratory Data Analysis and Time Series Analysis based Prediction of Automobile Crashes in NYC

Gokul Gandhikumar
A69036649

Ninad Ekbote
A69026968

***Abstract*—A comprehensive dataset of all the automobile crashes in New York City (NYC) from 2012 to 2024 have been collected and provided by the New York City Police Department (NYPD). Exploratory Data Analysis (EDA) focusing on identifying patterns and trends in automobile crash distributions across various boroughs of NYC, including daily and monthly trends, as well as correlations between crash frequencies in different boroughs was performed. The daily automobile crash numbers for the next two years were predicted using a time series analysis based Autoregressive Integrated Moving Average (ARIMA) model. The findings and methodologies presented in this work can enhance decision-making processes and aid in the development of targeted interventions to reduce traffic-related incidents in NYC and similar urban areas.**

I. Introduction

Road traffic crashes are a major global concern, leading to significant fatalities, injuries and economic losses. Automobile accidents pose significant challenges to urban safety and mobility, particularly in densely populated areas like New York City (NYC). The comprehensive analysis of traffic accident data is crucial for developing effective strategies to enhance road safety and urban planning. This work presents an in-depth analysis of the Motor Vehicle Collisions dataset provided by the New York City Police Department (NYPD), covering crashes from 2012 to 2024. This dataset with records of 2,132,793 crashes across NYC's five boroughs, serves as a foundation for identifying patterns, trends and potential risk factors associated with traffic accidents.

This work involves two phases: Exploratory Data Analysis (EDA) and Time Series Analysis based Predictions. The EDA phase focuses on obtaining key insights into crash distributions across various dimensions. The distribution of crashes throughout the day, monthly trends by borough, and correlations between crash frequencies across different areas of the city are analyzed deeply.

Finally, time series analysis techniques are applied to develop predictive models for future crash occurrences. This involves the application of ARIMA (Autoregressive Integrated Moving Average) model for forecasting. The proposed model was able to predict daily crash frequencies for the years 2025 and 2026, providing valuable foresight for NYPD, urban planners and policymakers.

By combining EDA insights with predictive modeling, this work aims to contribute to the ongoing efforts in improving road safety in large cities such as NYC. The findings and methodologies presented in this work can augment decision-making processes, aid in the development of targeted interventions and contribute to the reduction of traffic-related incidents in large urban metropolises.

II. Dataset

The Motor Vehicle Collisions dataset [1] from the City of New York provides detailed information about traffic accidents, offering valuable insights into road safety and urban planning. Collected by the NYPD, the dataset includes crash details such as date, time, location, contributing factors, vehicle types, and injury severity. The dataset spans several years and covers accidents across all five boroughs of New York City. The dataset also includes metadata such as weather conditions, lighting, and road surface status, which are crucial for understanding the factors influencing accidents. These details are available for 2,132,793 crashes that have taken place in NYC from 2012 to 2024.

This dataset is open-sourced and accessible to the public, enabling policymakers, urban planners, and safety organizations to better understand and address traffic-related issues in the city. It is a critical resource for initiatives aimed at reducing traffic fatalities and improving road safety through data-driven strategies. Data scientists can leverage this data for various applications, including traffic pattern analysis, accident prediction models, and evaluating the impact of urban infrastructure changes on road safety.

III. Exploratory Data Analysis

A. Distribution of Crashes across a Day

The plot in Fig. 1 focused on identifying trends in crash times across a 24-hour period, visualized through a histogram that offers insights into hourly crash frequencies. The results showed a clear bimodal distribution, with two prominent peaks occurring during the day. The first increase was observed during the morning rush hours (7 AM to 9 AM), aligning with the start of work and school commutes when vehicular traffic density tends to spike. The second, more pronounced peak occurred during the evening commute period (4 PM to 7 PM), a time window characterized by heightened road usage as commuters return home, often compounded by driver fatigue, lower attentiveness, and increased congestion.

Crash frequencies were notably reduced during late-night hours (midnight to 5 AM), reflecting minimal road activity during this period. However, despite the overall lower frequency, crashes occurring during these hours could have

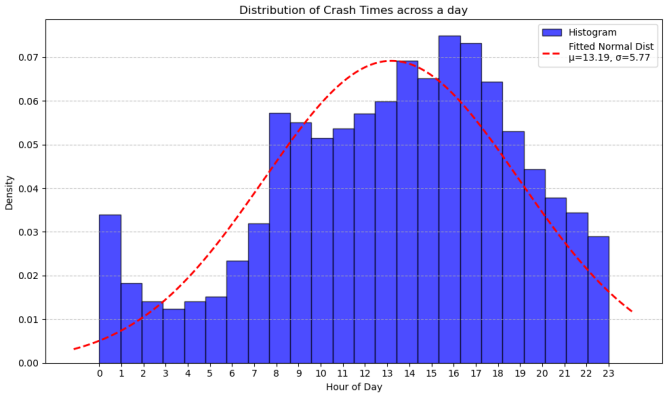

Fig. 1. Distribution of Crash Time in a Day

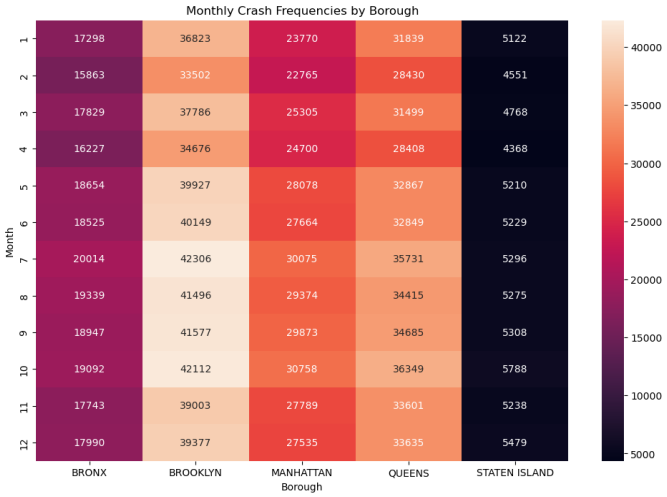

Fig. 2. Monthly Crash Frequencies by Borough

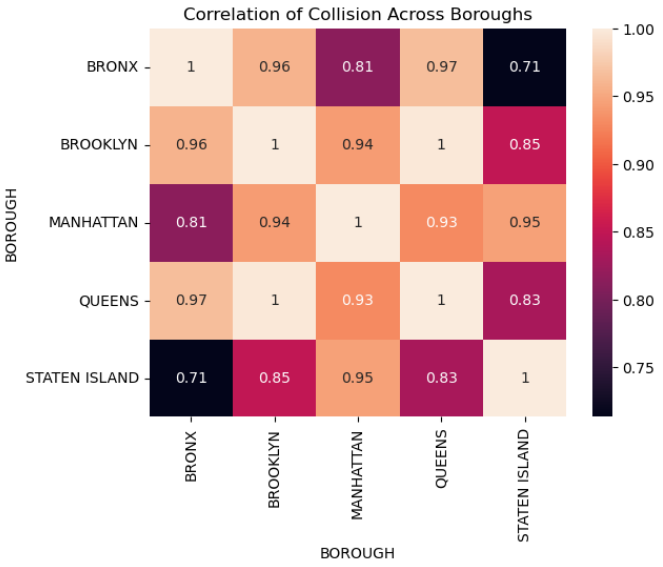

Fig. 3. Correlation Matrix of Collision Frequencies Across Boroughs

higher severity due to factors such as speeding, rash driving, and limited visibility. A gradual increase in crash incidents was also observed starting from early morning hours (5 AM to 7 AM), likely coinciding with the beginning of daily vehicular activity, including commercial transport operations. These findings highlight the strong temporal correlation between crash occurrences and daily human activity patterns, emphasizing the influence of traffic volume, driver behaviour, and time-dependent factors such as lighting conditions and roadway congestion. The bimodal peaks suggest that targeted interventions during these high-risk windows are critical. It is also observed that the distribution of crashes on day follows a normal distribution. When, the data is fitted on a normal curve, it has the a mean at 1:19 pm and a standard deviation of 5.7 hours.

B. Monthly Crash Trends by Borough

The heatmap in Fig. 2 visualizes monthly automobile crash frequencies by borough in NYC. Distinct patterns of crashes across boroughs and months can be observed from the plot. Brooklyn consistently records the highest number of crashes throughout the year, peaking in July (40,508). Queens follows Brooklyn's crash counts, showing a similar rise during the summer months. Manhattan and the Bronx experience relatively lower crash frequencies, with counts ranging between 15,000–29,000 monthly. Staten Island, being the least populated, exhibits the lowest crash counts, ranging from 4,100 to 5,500. Overall, crashes peak during summer months (May–October), due to increased road activity and travel during this period.

C. Correlation of Crash Frequency across Boroughs

The heatmap in Fig. 3 shows the correlation between crashes taking place in different boroughs of NYC. The values of correlation ranges from -1 to 1. Values close to 1 indicate a strong positive correlation, meaning that as crashes increase in one borough, they tend to also increase in the other. Values close to -1 indicate a strong negative correlation, and values around 0 suggest no significant correlation. For example, the correlation between crashes in Manhattan and Brooklyn is 0.94 and it is close to 1. This implies that if the crashes increases in Manhattan, then the number of crashes in Brooklyn will also increase.

IV. Time Series Analysis

Time series analysis is a statistical technique that is used to examine and interpret data points collected sequentially over time. It focuses on identifying trends and seasonality patterns in sequential data to extract insights and make predictions. In particular, to perform time series analysis, the data points in the dataset should be collected at consistent intervals. These kinds of datasets are essential for extracting seasonality trends and understanding the evolution of variables over time.

Time series analysis involves two main steps [2]. Firstly, good predictions can only be made when a variable exhibits strong stationarity. To test whether the prediction variable is

stationary, hypothesis testing methods like Augmented Dickey-Fuller (ADF) test are utilized. Further, the variable needs to be made stationary if it fails the stationarity test. Finally, future predictions of time series data are made using various models such as ARIMA, exponential smoothing, and spectral analysis.

In this work the number of crashes that happens in NYC on each particular date for the next two years is predicted. The dataset consists of the automobile crashes that have taken place in NYC from 2012 to 2024 (till date). The predictions of number of crashes that can happen over the next two years (2025 and 2026) are made using time series analysis.

A. Stationarity Testing

In order to predict the average number of crashes on each day using time series prediction we need to ensure that the input variable data we are predicting is stationary. Stationary data refers to a time series data whose statistical properties, such as mean, variance, and autocorrelation, remain constant over time [3]. It does not have trends, seasonality, or time-dependent changes, making it predictable and suitable for many time series models. Stationarity simplifies analysis by ensuring consistent patterns unaffected by time shifts or transformations. In cases when the prediction variable is non-stationary, some statistical data augmentation techniques needs to be applied to convert the given input data to stationary data.

The stationary of our variable, average number of crashes, can be tested by applying hypothesis testing. The null hypothesis is that the variable is non-stationary and the alternate hypothesis is that the variable is stationary.

The Augmented Dickey-Fuller (ADF) test [4] is the most commonly used statistical method to determine whether a time series is stationary or not. The test works based on the ADF equation given below:

$$\Delta_{y_t} = \gamma y_{t-1} + \alpha \sum_{i=1}^{n} \Delta_{y_{t-i}} + \epsilon \tag{1}$$

Where, γ is the unit root, Δ_{y_t} is the difference between subsequent values in series, α is a constant hyperparameter and ϵ accounts for noise. A time series becomes stationary only when $\gamma < 0$ in the ADF equation. Thus, the null and alternate hypothesis transforms into:

$$H_o \rightarrow \gamma \geq 0 \tag{2}$$

$$H_A \rightarrow \gamma < 0 \tag{3}$$

On performing the ADF test for the average number of crashes in the dataset, the p-values shown in Fig. 4 are obtained. Since, the computed p-value of 0.61 is very much larger than our significance level of 0.05, the null hypothesis H_o is accepted. It implies that the given prediction variable is non-stationary.

The first order time series differencing techniques can be used to remove non-stationarity in a time series data. It involves replacing the each data point y_t by the difference between the previous data point as shown below.

Test Statistic	-1.332266
p-value	0.614284
#Lags Used	29.000000
Number of Observations Used	4481.000000
Critical Value (1%)	-3.431810
Critical Value (5%)	-2.862185
Critical Value (10%)	-2.567113

Fig. 4. Inital ADF test results

Test Statistic	-1.701824e+01
p-value	8.493913e-30
#Lags Used	3.200000e+01
Number of Observations Used	4.477000e+03
Critical Value (1%)	-3.431811e+00
Critical Value (5%)	-2.862186e+00
Critical Value (10%)	-2.567114e+00

Fig. 5. ADF test results after performing first order differencing

$$y_t \coloneqq y_t - y_{t-1} \tag{4}$$

The ADF test is again conducted after performing first order differencing and the results obtained are shown in Fig. 5. The p-value obtained now is very much less than the significance value of 0.05. So, the null hypothesis H_o is rejected and the alternate hypothesis H_A is accepted. Therefore, the input variable now is stationary.

B. Prediction using ARIMA Model

The ARIMA model [5] is a commonly used statistical method for analyzing and forecasting time series data. It has three components: Autoregressive (AR), Integrated (I) and Moving Average (MA). The AR part models the dependency between an observation and a specified number of lagged observations. The MA component captures the relationship between an observation and residual errors from a moving average model. The I component ensures stationarity by differencing the data a specified number of times. The AR equation is given below:

$$\hat{y_t} = \phi_1 y_{t-1} + \cdots + \phi_p y_{p-1} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \tag{5}$$

Where, y_t is the observation at time t, p is the number of lag terms considered for AR step, ϕ are the AR coefficient, q is the number of lag terms considered for MA step, θ are the AR coefficient and ϵ_t is the error at t. In this work, the number of both the lagged terms p and q were considered to be 45 while training the ARIMA model.

The above ARIMA model was trained and the future values of average crashes per day in NYC was predicted. The predictions and the actual value are plotted in Fig. 6. The black dots represents the actual number of crashed on that day. the dark blue line represents the predicted number of

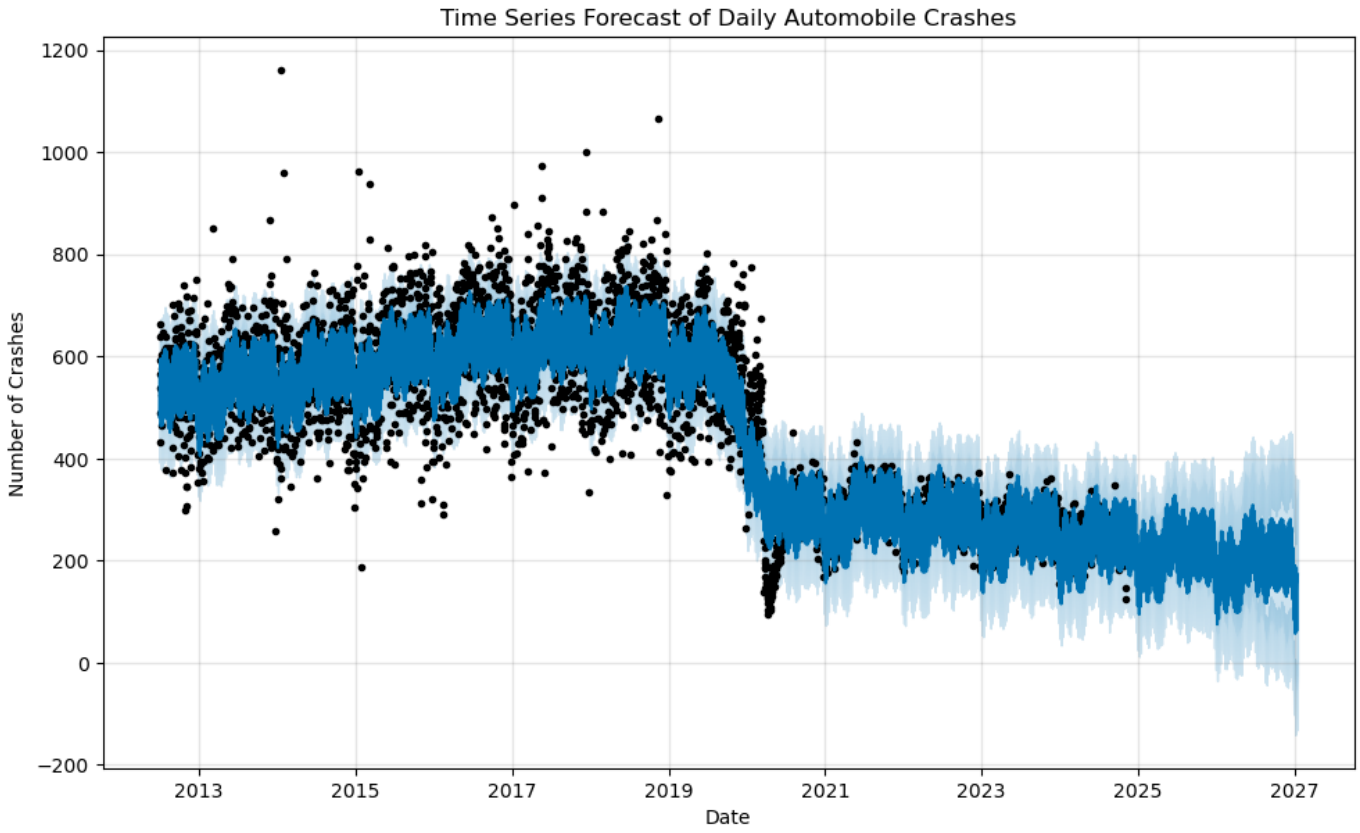

Fig. 6. Time Series Prediction

crashes by the ARIMA model. The light blue area around the dark blue line represents the confidence level of the model's forecast. This area is narrower in the initial years and its gets wider as we near the future predictions. This indicates that the uncertainty increases for future predictions. The predictions have been made in the plot for the years 2025 and 2026.

To evaluate the model, the dataset was divided into a training and testing dataset. The number of crashes for the first 1825 days or 5 years was used for training and then the rest of the data for the next 900 days was used for testing. Then, the Mean Absolute Error (MAE) for the predicted values was calculated. MAE is defined as the average of the absolute differences between the predicted values and the actual values. Mathematically, it can be expressed as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

The MAE is plotted in Fig. 7 and the average value of MAE for the predictions was 69.74. Since, the average number of crashes per day is 579, the average margin of error in prediction is 11.7%.

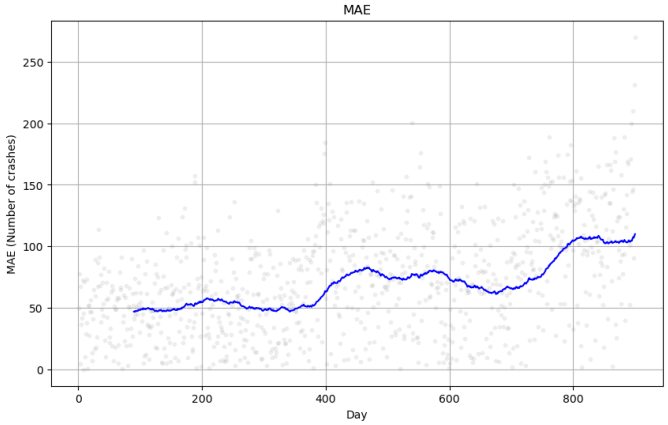

Fig. 7. MAE for prediction

V. Conclusion

The yearly, monthly and daily crashes across NYC boroughs from 2012 to 2014 were analyzed thoroughly. It was observed that crash frequencies follow a bimodal distribution throughout the day with peaks during morning and evening rush hours, Brooklyn and Queens consistently records the highest number of crashes with both boroughs experiencing peak incidents during summer months and there is a strong positive correlation between crash frequencies across different boroughs. A time series analysis based ARIMA model was proposed and trained successfully to predict the daily crashes in 2025 and 2026. The ARIMA model performed reasonably well with an average margin of error of 11.7% MAE of 69.74 crashes.

These findings provide valuable insights for urban planners, policymakers and the NYPD to develop targeted interventions and improve road safety in huge cities like NYC.

Source Code Availability

The code used for obtaining the results in this work is open-sourced and it can be accessed at https://github.com/gokulg02/Time-Series-Analysis-based-Prediction-of-Automobile-Crashes-in-NYC.

References

[1] New York City Police Department. (n.d.). Motor vehicle collisions - crashes [Dataset]. https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data.

[2] Shumway, Robert H., David S. Stoffer, and David S. Stoffer. Time series analysis and its applications. Vol. 3. New York: springer, 2000.

[3] Kwiatkowski, Denis, et al. "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?." Journal of econometrics 54.1-3 (1992): 159-178.

[4] Elliott, Graham, Thomas J. Rothenberg, and James H. Stock. "Efficient tests for an autoregressive unit root." (1992).

[5] Taylor, Sean J., and Benjamin Letham. "Forecasting at scale." The American Statistician 72.1(2018):37-45.