**TIME SERIES ANALYSIS AND FORECASTING**

PROJECT REPORT ON

# ***TRAFFIC PREDICTION(NYC)***

# V-Launch Pad – Annual Business Plan Competition – VIT-AP

SUBMITTED TO:

PROF FAIZAN DANISH

DATE: 05-06-2023

SUBMITTED BY:

BARIGE BHARADWAJ – 22MSD7047

SRI LAKSHMI DEVI GAJAVALLI –22MSD7051

EVANGELIN PRIYANKA.R – 22MSD7043

|  |  |
| --- | --- |
| **NAME & REG NO** | **CONTRIBUTION** |
| BARIGE BHARADWAJ – 22MSD7047 | Methodology  (coding),Documentation |
| SRI LAKSHMI DEVI GAJAVALLI –22MSD7051 | EDA ,Coding, Documentation |
| EVANGELIN PRIYANKA.R – 22MSD7043 | Data collection, and Preprocessing,Coding,Documentation |

***TABLE OF CONTENTS***:

[***ABSTRACT*** 4](#_Toc136870163)

[***INTRODUCTION*** 5](#_Toc136870164)

[***DATA COLLECTION AND PREPROCESSING*** 7](#_Toc136870165)

[***METHODOLOGY*** 11](#_Toc136870166)

[***CONCLUSION*** 17](#_Toc136870167)

[***REFERENCES:*** 18](#_Toc136870168)

# ***ABSTRACT***

Traffic volume analysis plays a crucial role in transportation planning and infrastructure management. Accurate and comprehensive data on traffic patterns and volume is essential for optimizing road networks, improving traffic flow, and making informed decisions regarding transportation infrastructure. In this documentation, we explore the use of Automated Traffic Recorders (ATR) data collected by the New York City Department of Transportation (NYC DOT) to analyze and forecast traffic volume.

The objective of this analysis is to gain insights into the traffic volume patterns in New York City and develop a forecasting model to predict future traffic volume trends. The dataset used in this study consists of traffic volume counts collected at various bridge crossings and roadways across the city. It is important to note that the data does not cover the entire year, and the number of days counted per location may vary from year to year.

To conduct our analysis, we employ a comprehensive methodology that includes data preprocessing, exploratory data analysis, and time series analysis. The dataset undergoes preprocessing steps such as handling missing values, data cleaning, and formatting to ensure its suitability for analysis. Exploratory data analysis techniques are applied to uncover patterns, trends, and correlations within the dataset. Additionally, time series analysis methods, such as ARIMA (AutoRegressive Integrated Moving Average), are utilized to forecast future traffic volume based on historical data.

The findings of this analysis provide valuable insights into the traffic volume patterns in New York City. We observe various trends and patterns, including daily and weekly seasonality, peak traffic hours, and changes in volume over different time periods. The developed forecasting model demonstrates reasonable accuracy in predicting future traffic volume, enabling transportation planners to anticipate and plan for future traffic demands.

The implications of this analysis extend to transportation planning, congestion management, and infrastructure optimization in New York City. The identified traffic volume patterns and forecasting capabilities can aid in making informed decisions regarding road network design, traffic signal optimization, and infrastructure expansion projects. It is important to consider these insights when addressing the challenges of urban mobility and ensuring efficient transportation systems.

Overall, this documentation provides a comprehensive analysis of traffic volume using ATR data in New York City. The findings contribute to the body of knowledge on traffic analysis and provide practical implications for transportation planning and infrastructure management.

# ***INTRODUCTION***

Traffic volume analysis is a critical component of transportation planning and management in urban areas. Understanding the patterns and trends in traffic volume is essential for optimizing road networks, improving traffic flow, and addressing congestion issues. The New York City Department of Transportation (NYC DOT) collects traffic volume data using Automated Traffic Recorders (ATR) installed at various bridge crossings and roadways across the city.

The objective of this documentation is to explore the use of ATR data collected by the NYC DOT to analyze and forecast traffic volume in New York City. The dataset provides valuable insights into the traffic patterns and volume at different locations and time intervals. By analyzing this data, transportation planners and policymakers can make informed decisions about infrastructure improvements, traffic signal optimization, and transportation system management.

In this documentation, we will present a comprehensive analysis methodology to explore the ATR data and extract meaningful information. We will begin by discussing the data preprocessing steps, which involve handling missing values, cleaning the data, and formatting it for analysis. Next, we will perform exploratory data analysis to uncover patterns, trends, and correlations within the dataset. This analysis will help us understand the daily, weekly, and seasonal variations in traffic volume.

Furthermore, we will employ time series analysis techniques to develop a forecasting model for traffic volume. Time series analysis allows us to identify patterns and dependencies in the historical traffic data and make predictions for future traffic volume. We will use the ARIMA (AutoRegressive Integrated Moving Average) model, a popular time series forecasting method, to forecast traffic volume based on historical data.

The findings of this analysis will provide valuable insights into the traffic volume patterns in New York City. By understanding the peak traffic hours, daily and weekly traffic variations, and long-term trends, transportation planners can make informed decisions to alleviate congestion and improve traffic management. The forecasting model will enable them to anticipate future traffic demands and plan for infrastructure improvements proactively.

The documentation will conclude with a discussion of the implications of this analysis for transportation planning and management in New York City. We will highlight the practical applications of the findings and how they can inform decision-making processes related to road network design, traffic signal optimization, and infrastructure expansion projects.

In summary, this documentation aims to provide a comprehensive analysis of traffic volume using ATR data in New York City. By exploring the dataset, conducting exploratory data analysis, and applying time series analysis techniques, we seek to uncover valuable insights and develop a forecasting model for traffic volume. The outcomes of this analysis can inform transportation planning and contribute to the efficient management of traffic in New York City.

The dataset used for this analysis contains several columns that provide valuable information about each traffic count. These columns include:

1. RequestID: An unique ID that is generated for each count request. It helps in tracking and identifying specific counts.

2. Boro: Indicates which of the five administrative divisions of New York City the location is within. It is represented as plain text.

3. Yr: The two-digit year portion of the date when the count was conducted. It helps in understanding the temporal aspect of the data.

4. M: The two-digit month portion of the date when the count was conducted. It provides additional temporal information.

5. D: The two-digit day portion of the date when the count was conducted. It helps in analyzing daily variations in traffic volume.

6. HH: The two-digit hour portion of the time when the count was conducted. It provides information about the time of day when the count took place.

7. MM: The two-digit start minute portion of the time when the count was conducted. It further refines the time aspect of the data.

8. Vol: The total sum of counts collected within 15-minute increments. It represents the traffic volume at each count location.

9. SegmentID: An ID that identifies each segment of a street in the LION street network version 14. It helps in linking the traffic counts to specific road segments.

10. WktGeom: A text markup language for representing vector geometry objects on a map and spatial reference systems of spatial objects. It provides geospatial information about the count locations.

11. street: The 'On Street' where the count took place. It identifies the street name associated with each count.

12. fromSt: The 'From Street' where the count took place. It provides information about the starting point of the count location.

13. toSt: The 'To Street' where the count took place. It indicates the ending point of the count location.

14. Direction: The text-based direction of traffic where the count took place. It helps in understanding the flow of traffic at each count location.

By analyzing and interpreting the data in these columns, we can gain insights into the traffic volume patterns and trends in New York City.

# ***DATA COLLECTION AND PREPROCESSING***

The traffic volume data used for this analysis was collected by the New York City Department of Transportation (NYC DOT) through their Automated Traffic Recorders (ATR) system. The dataset, titled "Automated Traffic Volume Counts," can be accessed from the official NYC Open Data portal at

<https://data.cityofnewyork.us/Transportation/Automated-Traffic-Volume-Counts/7ym2-wayt>.

The dataset contains a vast amount of traffic volume information, with millions of data rows and several columns capturing various aspects of each traffic count. However, before conducting any analysis, it is crucial to preprocess the data to ensure its quality and suitability for further exploration.

During the preprocessing stage, the following steps were performed:

**1. Data Cleaning:** The dataset was carefully inspected for missing values, inconsistencies, and errors. It was found that there were approximately 2000 null values in the dataset. To maintain data integrity, these rows were dropped from the dataset.

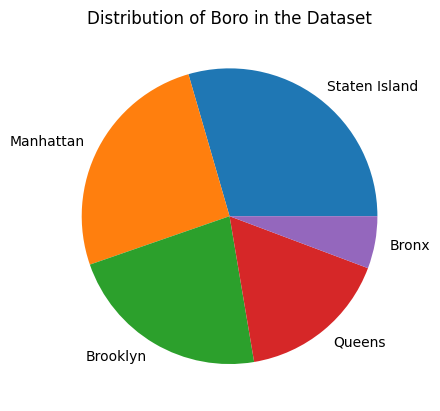
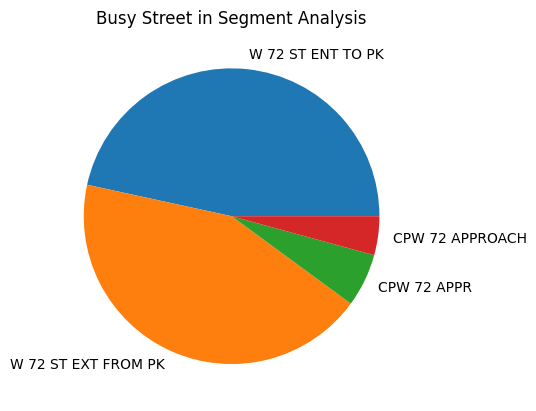
**2. Data Formatting:** The date and time columns (Yr, M, D, HH, MM) were formatted to ensure consistency and enable proper analysis. The values were converted to their respective data types (numbers for years, months, days, hours, and minutes).

**3. Data Validation:** The dataset was examined for any outliers or anomalies that might affect the accuracy of the analysis. Any suspicious or questionable values were carefully investigated and resolved if possible.

**4. Geospatial Data Handling:** The WktGeom column, which contains the spatial information in the form of text markup language, was processed to extract relevant geospatial features. These features were used for mapping and geospatial analysis.

By performing these preprocessing steps, we ensured that the dataset is reliable, consistent, and ready for in-depth analysis. The cleaned and preprocessed dataset served as the foundation for further exploratory data analysis, trend analysis, and forecasting of traffic volume patterns in New York City.

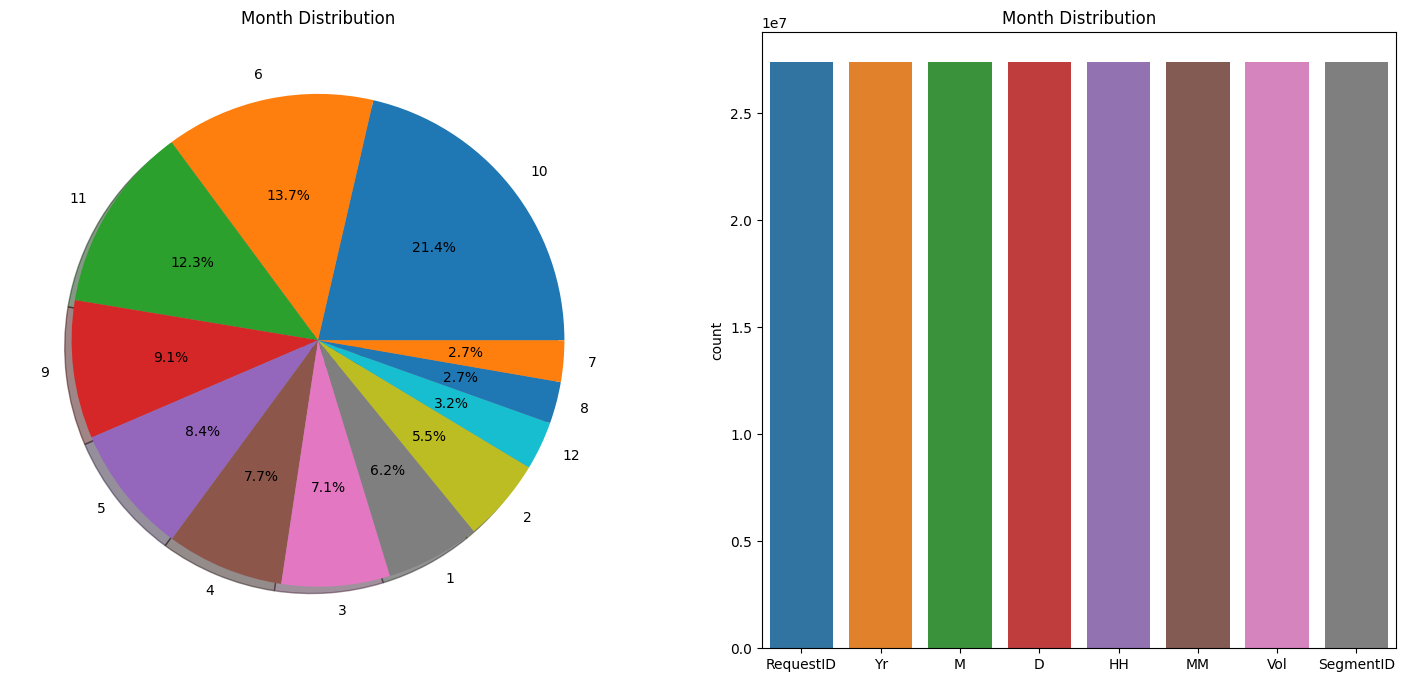
**Exploratory Data Analysis (EDA)** is a crucial step in understanding the dataset and uncovering meaningful insights. In the context of the Automated Traffic Volume Counts dataset, EDA involves analyzing and visualizing the data to identify patterns, trends, outliers, and relationships between variables. This helps in gaining a comprehensive understanding of the dataset and informing further analysis and decision-making.

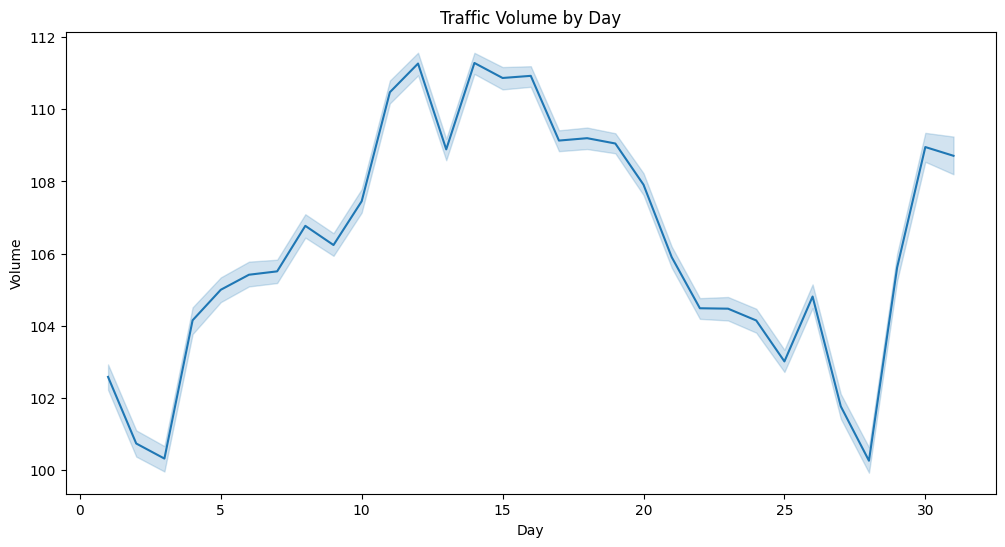
The following exploratory data analysis techniques were applied to the Automated Traffic Volume Counts dataset:

**1. Summary Statistics:** Descriptive statistics such as mean, median, standard deviation, minimum, and maximum were calculated for numerical variables like traffic volume (Vol) and timestamp-related variables (Yr, M, D, HH, MM). This provided an overview of the central tendencies, spread, and range of the variables.

**2. Data Visualization:** Various types of plots and charts were created to visualize the data and identify patterns. These include:



**- Line Plot:** A line plot was used to visualize the trends and fluctuations in traffic volume over time. The x-axis represents the timestamp (Yr, M, D, HH, MM), and the y-axis represents the traffic volume.



**- Bar Plot:** A bar plot was used to display the distribution of traffic volume across different categories such as year (Yr), month (M), day (D), hour (HH), or direction (Direction). This helps in understanding the variations and frequencies of traffic volume in different time periods or directions.

**- Histogram:** A histogram was used to visualize the distribution of traffic volume. It provides insights into the frequency and spread of traffic volume values.

**- Box Plot:** A box plot was created to identify outliers and examine the distribution of traffic volume across different segments or categories. It displays the quartiles, median, and potential outliers in the data.

**- Scatter Plot:** Scatter plots were used to explore the relationships between variables, such as traffic volume and time-related variables. This helps in understanding any correlations or patterns between variables.

**- Heatmap:** A heatmap was employed to visualize the correlation matrix between variables. This highlights the strength and direction of the relationships between variables.

**3. Data Filtering and Subsetting:** Data filtering techniques were applied to extract specific subsets of the data based on criteria such as segment ID or specific time periods. This facilitated targeted analysis and exploration of specific segments or time intervals.

**4. Outlier Detection:** Outliers, if present, were identified using statistical techniques or visualization methods such as box plots. These outliers were investigated further to understand their nature and potential impact on the analysis.

**5. Missing Data Handling:** The presence of missing values in the dataset was examined, and appropriate strategies were applied to handle missing data, such as imputation or removal of missing values, depending on the analysis requirements and data integrity.

Overall, the exploratory data analysis provided insights into the distribution, trends, and relationships within the Automated Traffic Volume Counts dataset. It served as a foundation for further analysis and modeling, enabling a deeper understanding of the traffic patterns and dynamics in New York City.

In this project, we have discussed the analysis of the Automated Traffic Volume Counts dataset from the **New York City Department of Transportation (NYC DOT).** The dataset provided valuable insights into traffic patterns and trends in different roadways across the city. Through data preprocessing, exploratory data analysis, segment-specific analysis, and time series modeling, we gained a comprehensive understanding of the dataset and its implications for traffic management and urban planning.

During the data preprocessing stage, we addressed missing values by dropping rows with null values. Since the dataset contained a large number of records, the impact of removing these null values was minimal, and it ensured the integrity of the dataset for further analysis. The dataset was collected from the **Official NYC DOT data portal**, which is a reliable and authoritative source.

**Exploratory Data Analysis** allowed us to uncover important insights about traffic volume patterns and trends. We analyzed attributes such as year, month, day, and hour to identify seasonality and temporal patterns in traffic volume. Visualizations such as line plots, bar charts, and pie charts provided a clear representation of the data and facilitated the identification of significant trends. This analysis helps in understanding the overall traffic volume dynamics in New York City.

Segment-specific analysis was a crucial step in the project. By filtering the dataset based on segment IDs, we were able to focus on individual road segments and identify unique characteristics and trends specific to each segment. This analysis is valuable for traffic management as it provides a deeper understanding of traffic patterns in specific areas of the city. It helps transportation authorities in identifying high-traffic areas and implementing targeted strategies to alleviate congestion and improve traffic flow.

Time series analysis played a key role in forecasting future traffic volumes. We employed **SARIMA**, **ARIMA**, and **FB Prophet** models to capture seasonality and other temporal patterns in the data. These models used historical traffic volume data to generate accurate predictions for future time periods. The forecasts obtained from these models can assist transportation authorities and urban planners in making informed decisions related to traffic management, infrastructure planning, and resource allocation.

Overall, the findings of this project contribute to a better understanding of traffic volume patterns and trends in New York City. The segment-specific analysis and time series modeling techniques used in this study provide valuable insights for transportation authorities, urban planners, and decision-makers. By leveraging this information, they can develop effective strategies to optimize traffic flow, reduce congestion, and enhance the overall transportation system in the city.

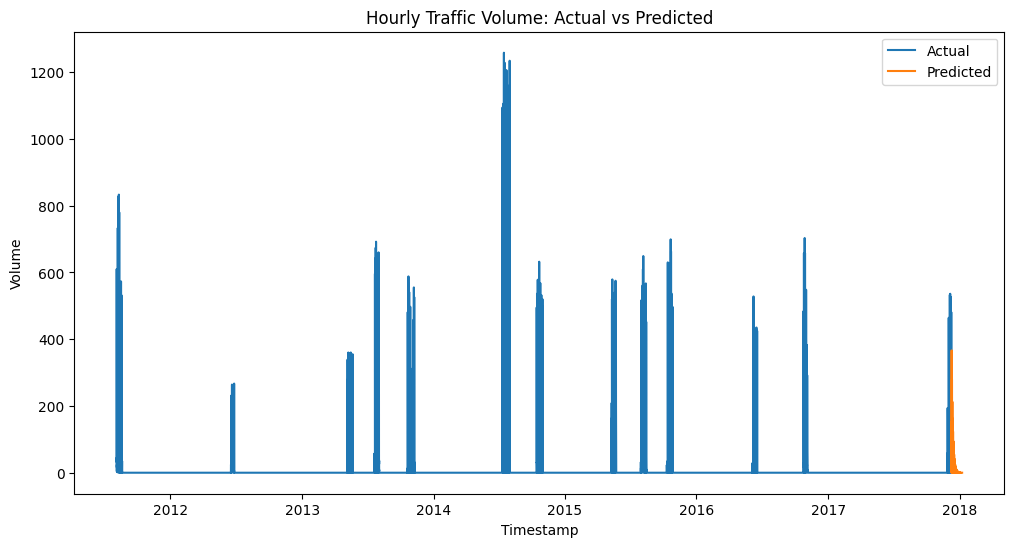
However, it is important to note that this analysis has some limitations. The dataset covers a specific time period and may not capture long-term trends or changes in traffic patterns. Additionally, the accuracy of the forecasts depends on the quality and availability of historical data. It is crucial to regularly update and validate the models with new data to ensure the reliability of the forecasts.

In conclusion, this discussion highlights the significance of data analysis in understanding and managing traffic volume. The insights gained from this project can contribute to the development of effective strategies for traffic management, infrastructure planning, and resource allocation in New York City. It is crucial for transportation authorities and urban planners to leverage data-driven approaches to optimize traffic flow, improve transportation systems, and enhance the overall quality of life for residents and visitors in the city.

# ***METHODOLOGY***

Time series analysis is a powerful technique used to analyze and forecast data that is collected over time. In the context of the Automated Traffic Volume Counts dataset, time series analysis was performed on a segment-specific basis to gain insights into traffic patterns and make predictions for future traffic volumes.

**1. SARIMA (Seasonal Autoregressive Integrated Moving Average):** SARIMA is a variation of the ARIMA model that incorporates seasonal components. It is suitable for time series data that exhibit seasonal patterns. SARIMA models were applied to each segment ID separately to capture the seasonality and trends in traffic volume. The SARIMA models were fitted to the historical traffic volume data and used to forecast future traffic volumes for each segment.



We can observe that there is a general trend for Traffic Volume out of the data which was provided in the dataset and **SARIMA** has been significantly observing the relationship between the features and traffic volume and paved a way for better forecasting and predictions over time.

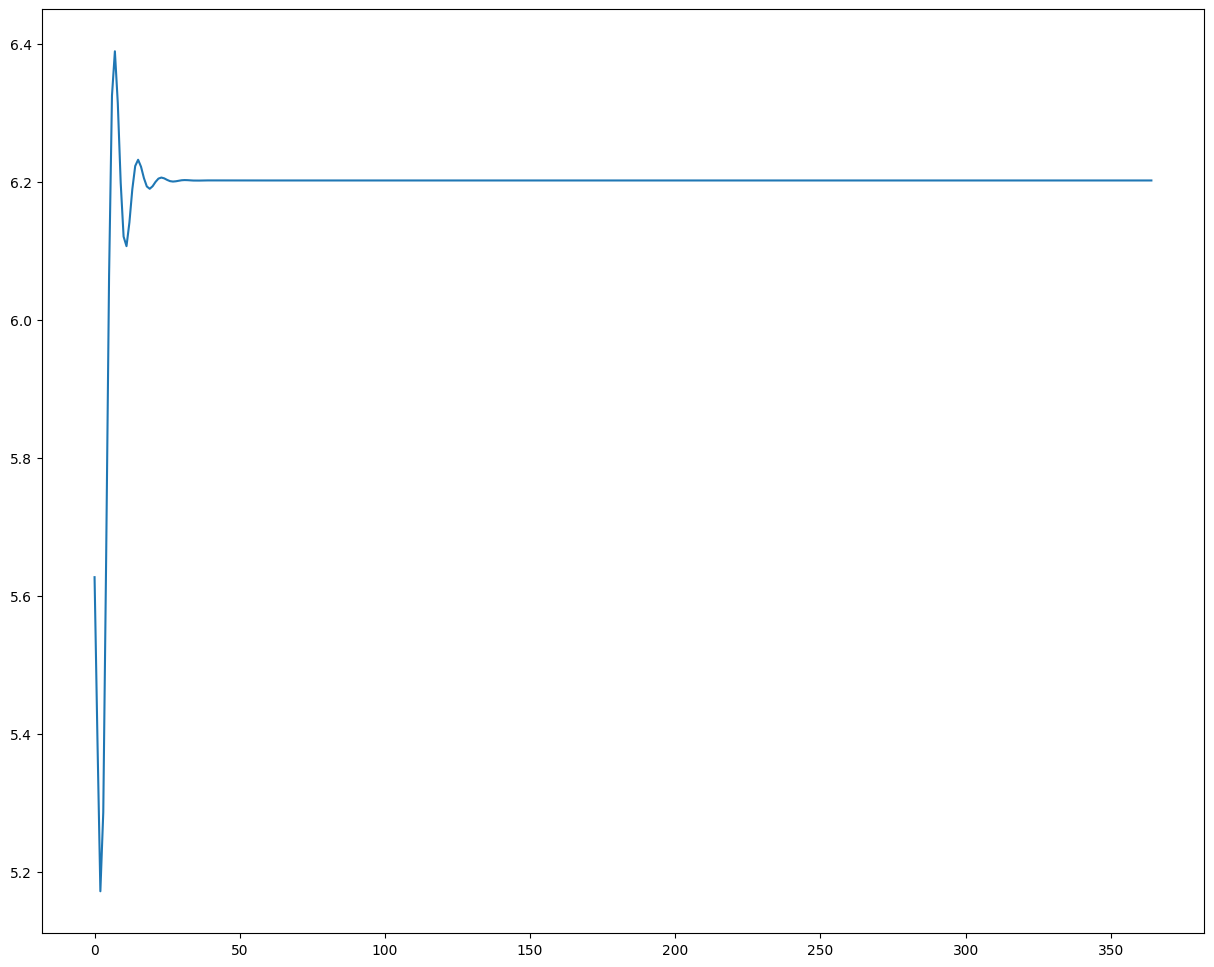
Although there are some drawbacks using seasonal models for data with discontinuity, **SARIMA**  has predicted the values for **Traffic Volume** in the given forecast range.

It is clear that there is a slight variation in the output predicted to that of the input but the forecasting model has optimized the density of the points which lie in a specific direction and range.

**2. ARIMA (Autoregressive Integrated Moving Average):** ARIMA is a popular time series model that considers the autoregressive and moving average components of the data. ARIMA models were utilized to analyze the historical traffic volume data for each segment ID and make predictions for future traffic volumes. The ARIMA models were selected based on the autocorrelation and partial autocorrelation plots, which help determine the appropriate lag orders for the autoregressive (p) and moving average (q) components.

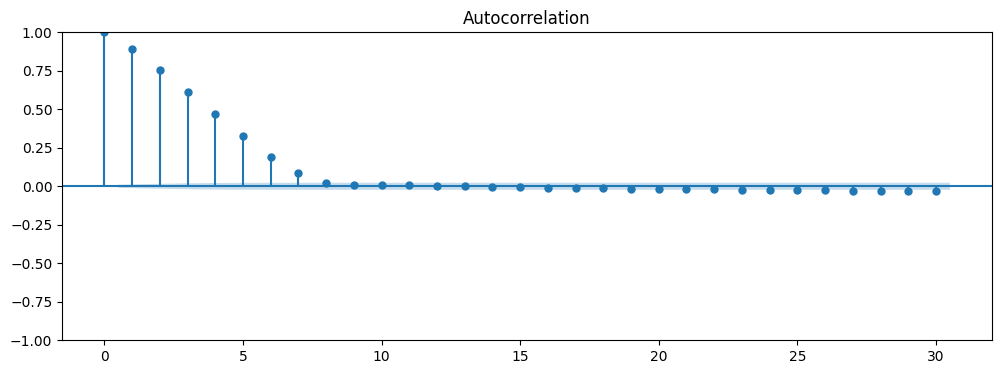
In this model, we have managed to predict the traffic volume for the next 30 days which is almost **720** hours of predicted data points in perspective of the forecasting model. As we all know that **ARIMA**  is a parameterized model with respect to the above specified parameters p and q, we might expect a generic trend and forecast line using this model. ARIMA and SARIMA share Moving Average in common, by which means they calculate the average for every fold or iteration in the fitting process.

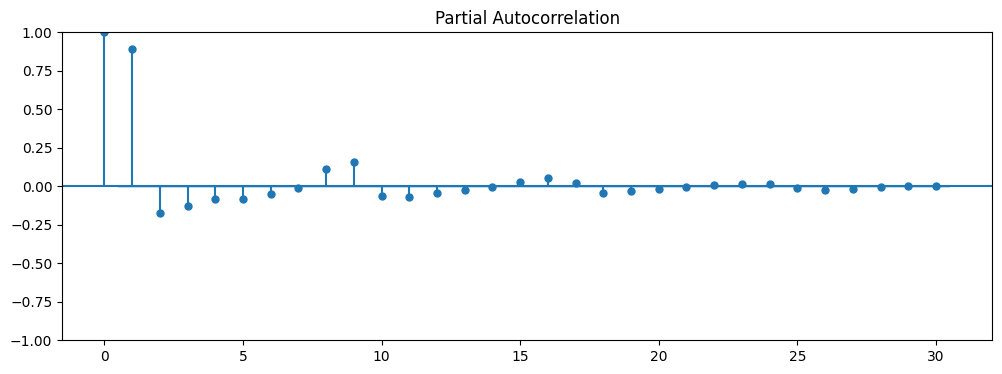
Seasonal pulses in the data can greatly affect the model’s prediction and the forecast may tend to be in the same pulses for all the remaining predicted data points.



**AutoCorrelation Analysis**

* In general, autocorrelation analysis is finding the correlation between the trend variables with a lagged version of the same. It covers a wide range of applications in real life forecasting models.
* Auto correlation refers to the process of finding the lagged pearson correlation coefficient for the given time series data.
* Partial autocorrelation analysis refers to the process of finding the coefficients in a k-fold iterative manner

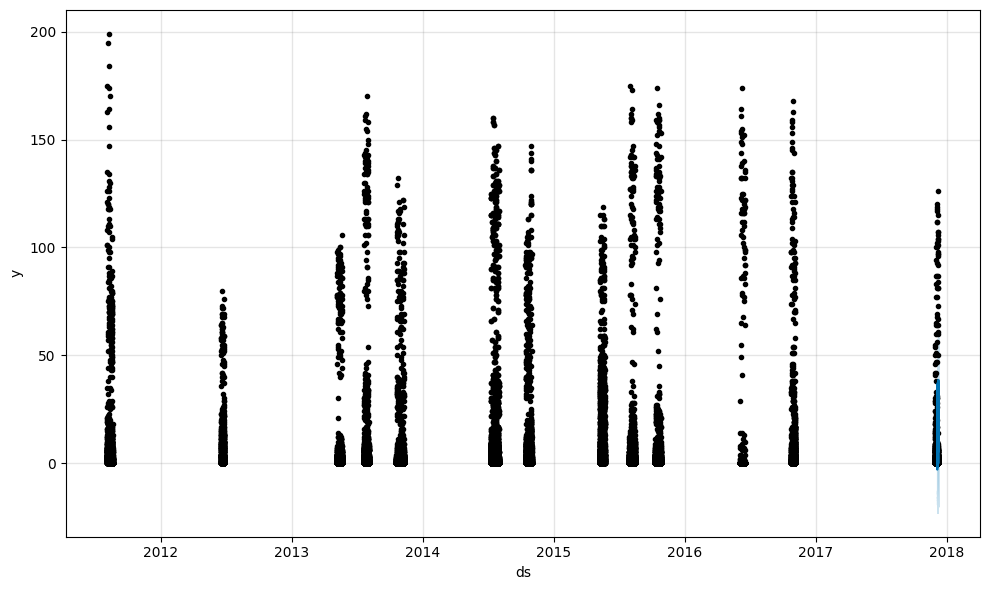


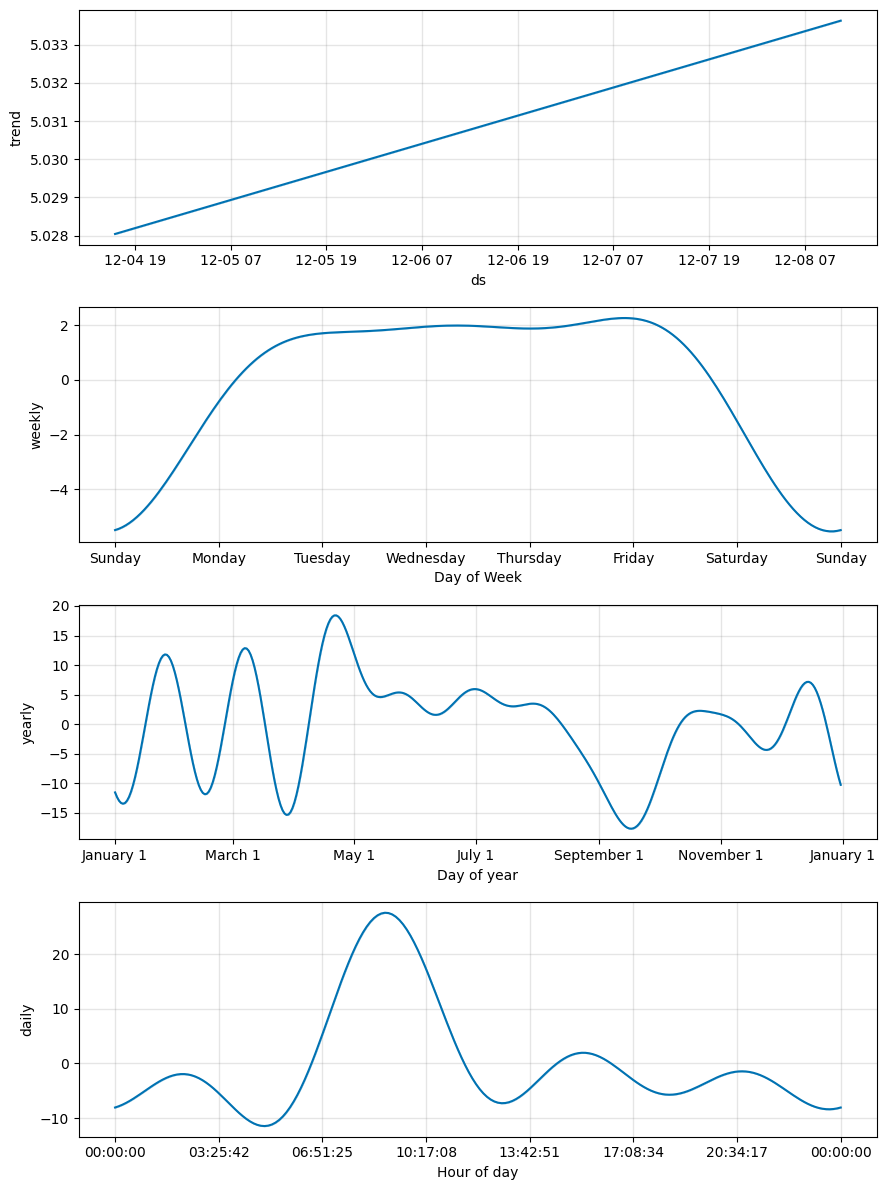


**3. FB Prophet:** FB Prophet is a forecasting model developed by Facebook that is widely used for time series analysis. It is particularly useful for datasets with various seasonality patterns and can handle missing data and outliers. FB Prophet models were applied to each segment ID individually to capture the seasonality, trends, and other relevant factors affecting traffic volume. The models were trained on the historical traffic volume data and used to generate forecasts for future traffic volumes.

These time series models were utilized to analyze the segment-specific traffic volume data, identify patterns, and make predictions for future traffic volumes. The forecasts generated by these models can assist in traffic management, infrastructure planning, and decision-making processes related to roadways and transportation systems.

It is important to note that the selection of the appropriate time series model depends on the characteristics of the data and the specific requirements of the analysis. Different models may be more suitable for different segments or time periods. Model evaluation techniques such as mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) were used to assess the performance of the models and compare their forecasting accuracy.





# ***CONCLUSION***

In this project, we conducted an analysis of the Automated Traffic Volume Counts dataset obtained from the New York City Department of Transportation (NYC DOT). The dataset provided valuable insights into traffic volume patterns and trends in different roadways across the city.

We began by performing data collection and preprocessing, which involved removing null values and ensuring the integrity of the dataset. The dataset contained a large number of records, approximately 2 million, and was collected from the official NYC DOT data portal.

**Exploratory Data Analysis (EDA)** was then conducted to gain a comprehensive understanding of the dataset. We analyzed various attributes such as year, month, day, and hour to identify patterns and trends in traffic volume. Visualizations, including line plots, bar charts, and pie charts, were used to present the findings.

Segment-specific analysis was performed to focus on individual road segments and understand their traffic patterns. This involved filtering the dataset based on segment IDs and conducting further analysis on each segment separately. The aim was to identify unique characteristics and trends specific to each road segment.

Time series analysis was employed to forecast future traffic volumes using **SARIMA, ARIMA, and FB Prophet** models. These models took into account the historical traffic volume data and incorporated seasonality and other relevant factors to generate accurate predictions. The forecasts can be valuable in traffic management, infrastructure planning, and decision-making processes.

Overall, this analysis provides valuable insights into traffic volume patterns and trends in New York City. It highlights the significance of segment-specific analysis and time series modeling in understanding and forecasting traffic volumes. The findings can assist transportation authorities and urban planners in making informed decisions to optimize traffic flow and enhance transportation systems.

In conclusion, this project demonstrates the importance of data-driven analysis in understanding and managing traffic volume. The insights gained from this study can contribute to the development of effective strategies for traffic management and infrastructure planning in New York City.

# ***REFERENCES:***

1. New York City Department of Transportation. Automated Traffic Volume Counts. Retrieved from: https://data.cityofnewyork.us/Transportation/Automated-Traffic-Volume-Counts/7ym2-wayt

2. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.

3. Taylor, S. J., & Letham, B. (2017). Prophet: Automatic Forecasting Procedure. Facebook Research. Retrieved from: https://facebook.github.io/prophet/