Application of Time Series Techniques in Uber NYC Data

Introduction

What is Time Series?

A time series is a sequence of data points recorded at successive time intervals, often used to analyze trends, patterns, and dependencies over time. Mathematical and statistical analysis performed on this kind of data to find hidden patterns and meaningful insight is called time series analysis. Time-series modeling techniques are used to understand past patterns from the data and try to forecast future horizons.

Components of Time Series Analysis

1. **Trend:** The long-term movement in data, indicating an overall increase, decrease, or stability.
2. **Seasonality:** Regular, predictable fluctuations occurring at fixed intervals (e.g., daily, weekly, or yearly patterns).
3. **Cyclic Patterns:** Long-term fluctuations that do not have a fixed period, often influenced by external factors like economic conditions.
4. **Irregular Variations:** Random or unpredictable variations that cannot be explained by trends or seasonality.

Applications of Time Series Analysis

Time series analysis has widespread applications across various industries, including:

- **Finance:** Stock market prediction, interest rate forecasting, and risk management.
- **Healthcare:** Disease prediction, patient monitoring, and outbreak forecasting.
- **Climate & Environment:** Weather forecasting, climate change modeling, and disaster prediction.
- **Retail & Inventory Management:** Demand forecasting, sales analysis, and stock optimization.
- **Transportation & Logistics:** Traffic prediction, ride-hailing demand forecasting, and fleet management.
- **Energy Sector:** Load forecasting, power demand analysis, and resource allocation.

In this study, focus lies on forecasting the demand for Uber with Time Series approach and for that purpose, **Uber Pickups in New York City** has been considered.

About the Dataset

Name: Uber Pickups in New York City

This directory contains data on over 4.5 million Uber pickups in New York City from April to September 2014, and 14.3 million more Uber pickups from January to June 2015. Trip-level data on 10 other for-hire vehicle (FHV) companies, as well as aggregated data for 329 FHV companies, is also included.

Focus of our study - Uber trip data from 2014 (April - September), separated by month, with detailed location information.

There are six files of raw data on Uber pickups in New York City from April to September 2014.This Data set is used for Time series Forecasting. The files are separated by month and each has the following columns:

- Date/Time : The date and time of the Uber pickup
- Lat : The latitude of the Uber pickup
- Lon : The longitude of the Uber pickup
- Base : The TLC base company code affiliated with the Uber pickup

Source: Kaggle

Learning about Uber

Uber is a leading ride-hailing company that has transformed urban transportation by providing an app-based platform connecting riders with drivers. Founded in **2009**, Uber quickly expanded worldwide, offering convenient, on-demand rides at competitive prices.

Uber in NYC

New York City is one of Uber's busiest markets, with **millions of daily rides** serving commuters, tourists, and residents. Uber has significantly impacted NYC's transportation landscape, competing with taxis and public transit while adapting to local regulations. The Uber NYC dataset used in this study provides valuable insights into **ride demand patterns, peak hours, and seasonal variations** across different times and locations.

Methodology

Problem Statement

The objective of this study is to visualise, explore, and experiment with various time series forecasting models to predict Uber NYC pickups demand efficiently. We aim to analyze patterns, trends, and seasonality in the time series and evaluate the performance of different forecasting models - ARIMA, Single ES, Double ES, Triple ES, and SARIMA.

Steps Involved:

1. **Importing Data:** Loading the dataset for analysis.
2. **Data Preprocessing:** Cleaning and structuring the dataset to remove inconsistencies and missing values.
3. **Data Visualization:** Exploring trends, patterns, and anomalies using graphical representations such as line charts and histograms.
4. **Splitting the Dataset:** Dividing the dataset into training (90%) and testing (10%) subsets to ensure effective model evaluation.
5. **Model Training and Testing:** Applying different forecasting models to train on historical data and test their predictive accuracy.
6. **Comparing Models:** Evaluating models based on performance metrics such as RMSE (Root Mean Squared Error).
7. **Output Analysis:** Interpreting forecasting results to understand model effectiveness and practical applicability.
8. **Conclusion:** Interpretating the final outcomes and concluding the paper.

Which Forecasting Models are used?

Exponential Smoothing

Exponential Smoothing is a widely used forecasting method that applies a weighted average of past values, with more recent observations having a greater influence on future predictions than older ones. It is a special case of the weighted moving average method, where the weights decline geometrically over time. This approach ensures that the most recent data points contribute more significantly to the forecast, making it effective for short-term predictions.

Types of Exponential Smoothing

1. **Simple Exponential Smoothing (SES):** Suitable for data without trends or seasonality, where future values are predicted based on weighted averages of past observations.
2. **Double Exponential Smoothing (Holt's Model):** Accounts for trends but not seasonality by incorporating a trend-adjustment factor to improve prediction accuracy.

3. **Triple Exponential Smoothing (Holt-Winter's Method):** Handles both trend and seasonal variations by incorporating a seasonal component in addition to trend adjustments.

ARIMA

ARIMA (AutoRegressive Integrated Moving Average): Used for non-seasonal time series data, incorporating:

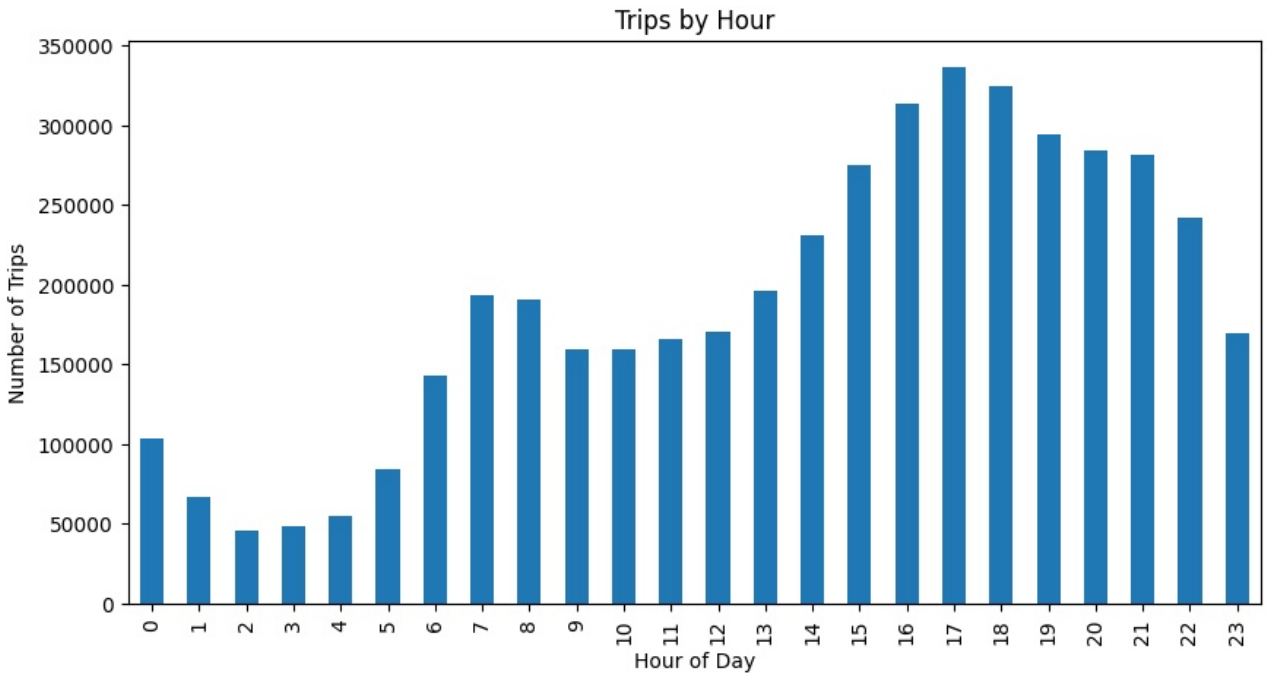
- AutoRegression (AR): Using past values for predictions.
- Moving Average (MA): Using past errors to smooth predictions.
- Differencing (I): Making the series stationary by removing trends and fluctuations.

SARIMA

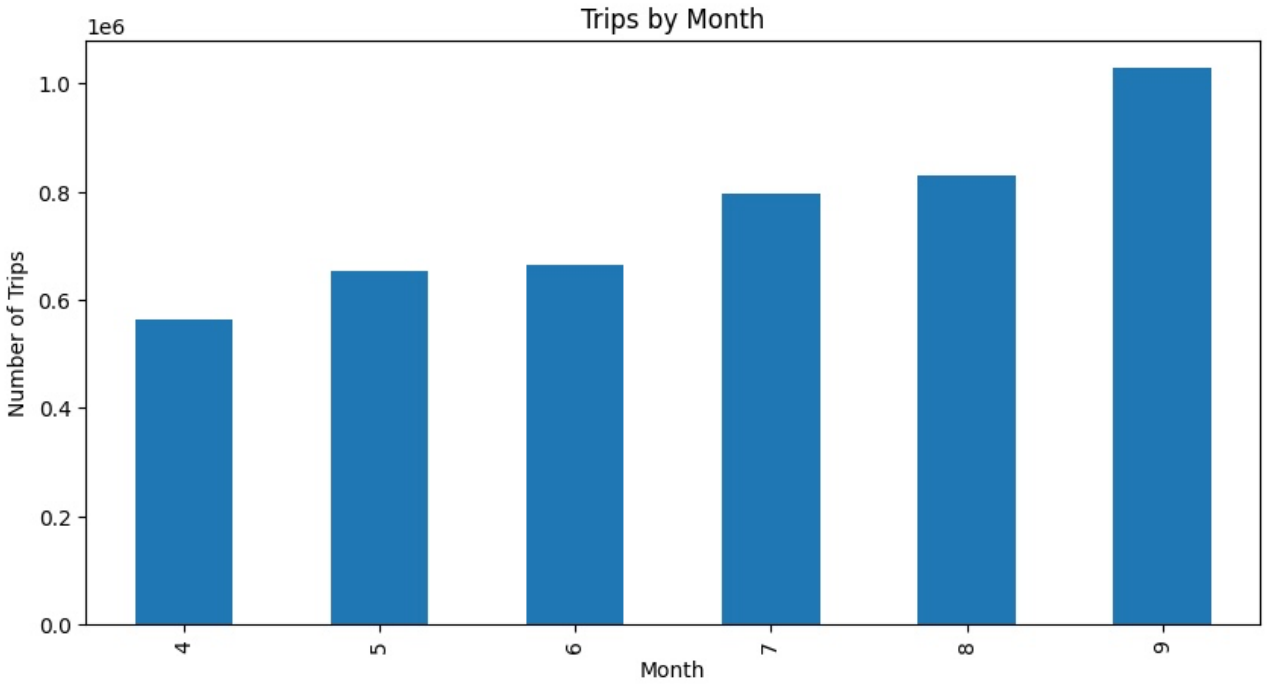
SARIMA (Seasonal AutoRegressive Integrated Moving Average): Extends ARIMA by incorporating seasonal components, making it more effective for datasets with periodic fluctuations.

Data Visualization Insights

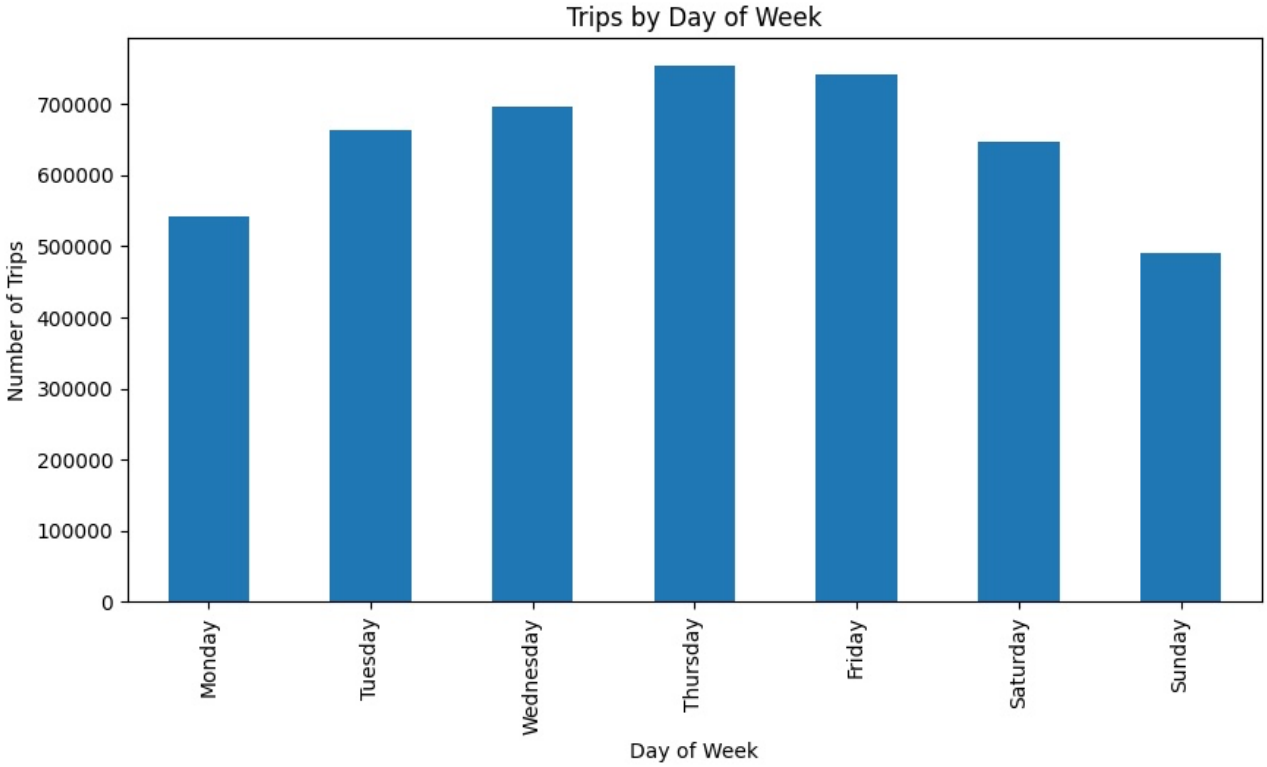
Analyzing the Uber NYC dataset revealed several important insights regarding ride demand, peak hours, and variations across different days and months.



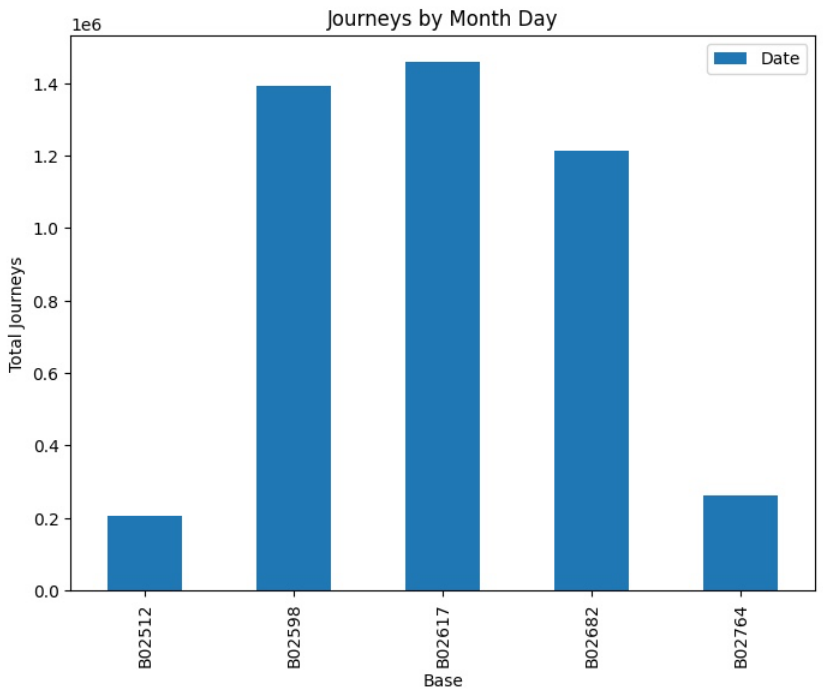
A clear daily pattern shows peak hours around late evenings and early nights across all days.



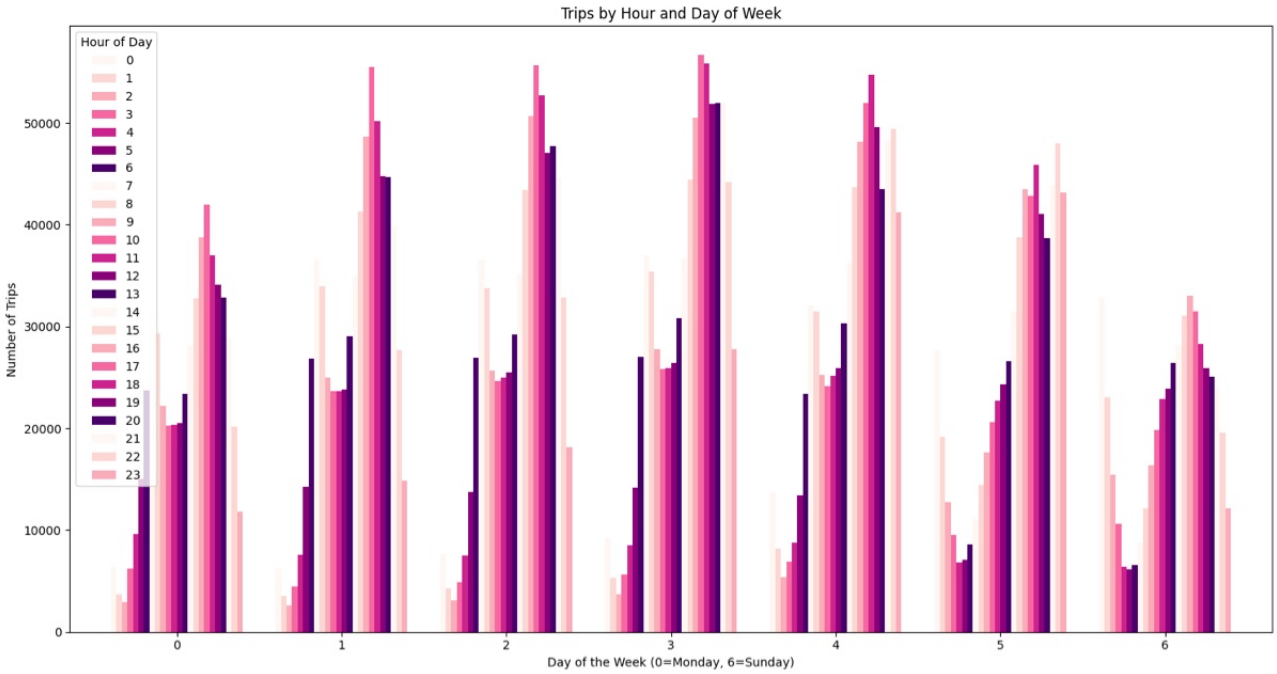
The number of Uber pickups increases steadily from April to September with the highest demand in September. This upward trend suggests **seasonal variations**, potentially due to increased tourism, changes in commuter behavior, or local events.



The highest number of Uber pickups occurred on **Thursdays and Fridays**, showing a surge in demand leading up to the weekend. **Sundays and Mondays had the lowest ride activity**, indicating reduced commuting and nightlife activity at the beginning of the week.

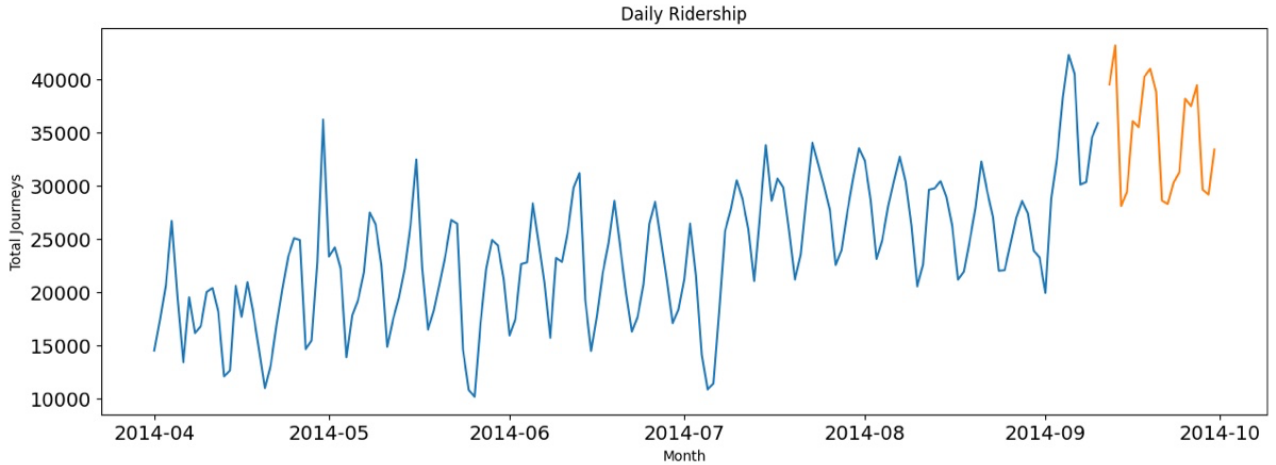


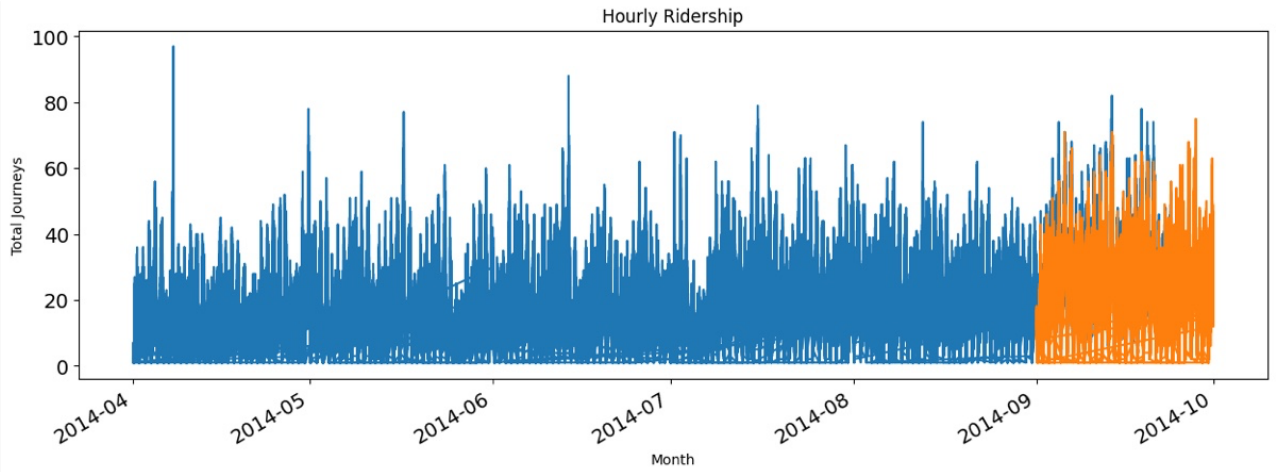
B02617 (Weiter) and B02598 (Hinter) have the highest number of trips, meaning they are the most active bases. These bases likely serve **high-demand areas, such as downtown NYC and commercial hubs**.



The visualization shows Uber trip volumes by hour and day of the week, highlighting distinct patterns in ride demand. Weekdays (Tuesday–Friday) exhibit two peak periods: **morning rush hours (7 AM – 9 AM)** and **evening rush hours (5 PM – 8 PM)**, indicating heavy commuter usage. In contrast, weekends (Saturday and Sunday) see higher demand in **late evening and night (8 PM – 1 AM)**, likely due to entertainment and social outings, with Sunday experiencing a gradual decline.

Training and Testing





Significance of 90-10 Split

A 90-10 split means that 90% of the dataset is used for training the model, while 10% is reserved for testing. This ratio is commonly used in time series forecasting for several reasons:

- Time series models rely on historical patterns. A larger training set (90%) ensures the model learns trends, seasonality, and variations effectively.
- The remaining 10% test set helps assess how well the model generalizes to unseen data. Prevents overfitting by ensuring the model is not just memorizing past values.

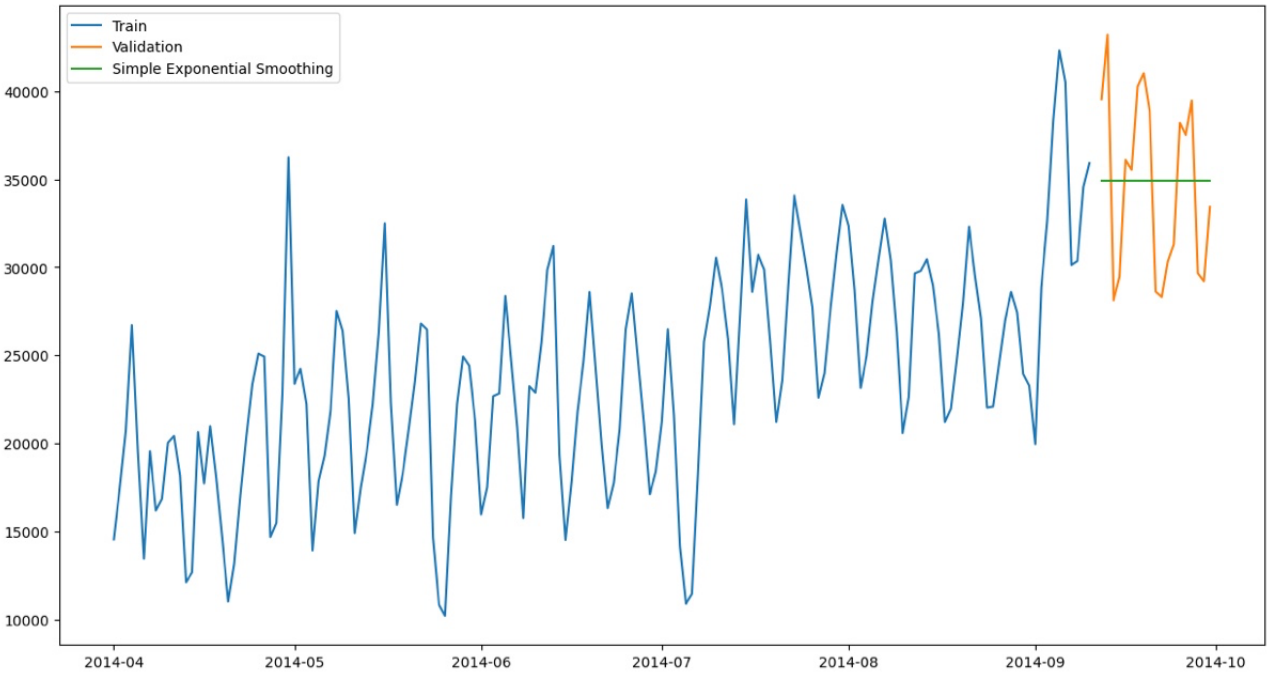
Forecasting Result

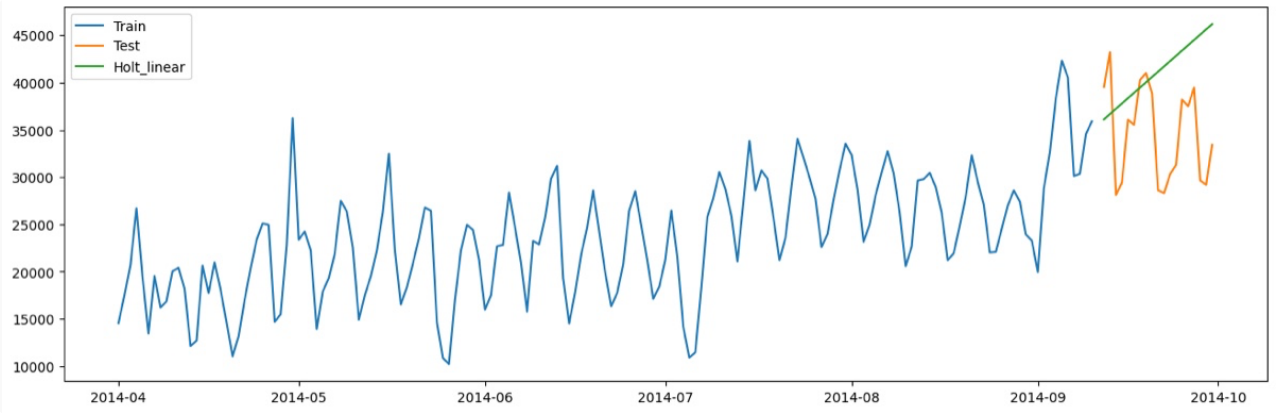
Code File

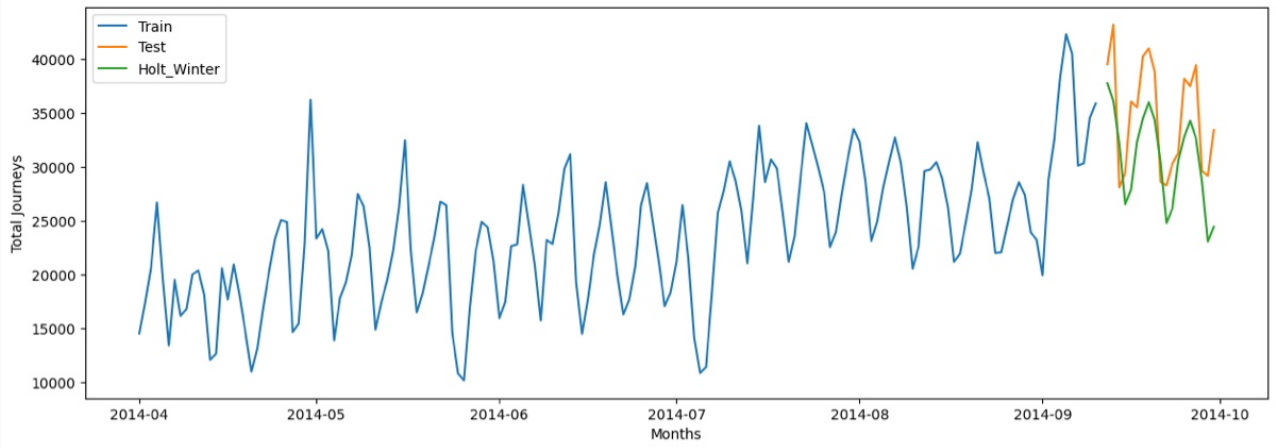
The forecasting models applied to the Uber NYC dataset aimed to predict future ride demand using various time series techniques. The models were evaluated based on their ability to capture trends, seasonality, and irregular variations, with performance measured using Root Mean Squared Error (RMSE).

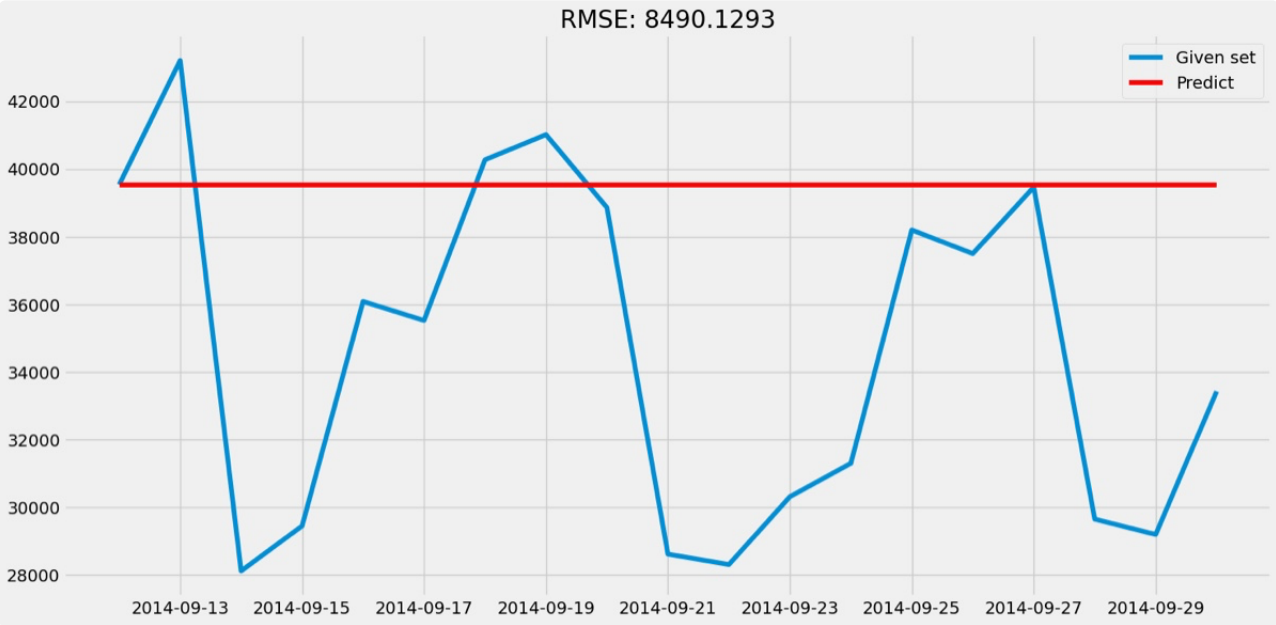
Model Performance Comparison

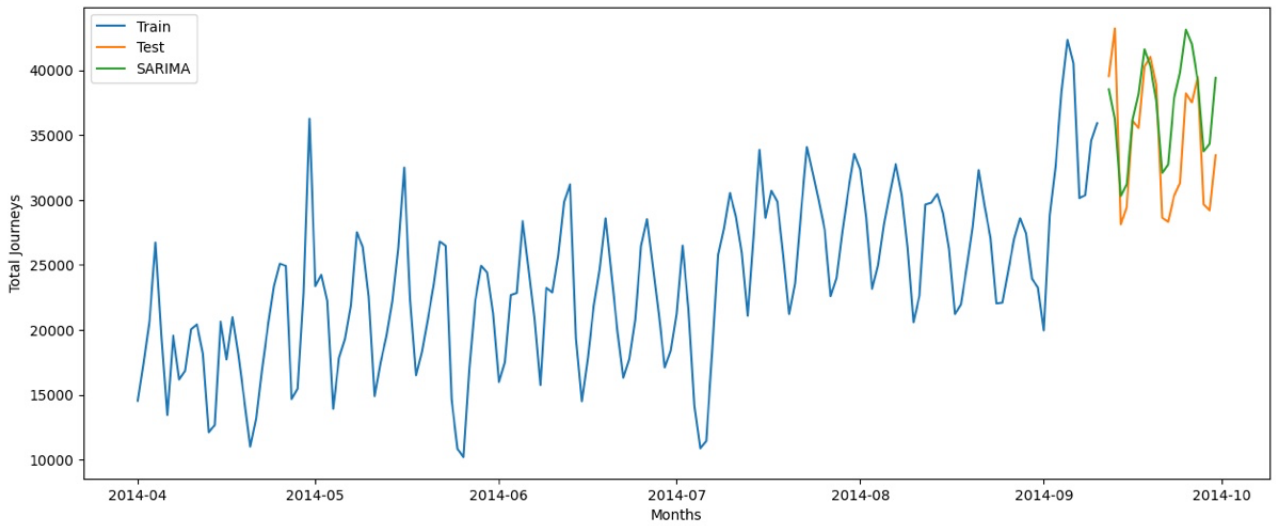
Model	RMSE (Root Mean Squared Error)
Simple Exponential Smoothing	7885.06
Double Exponential Smoothing	9185.89
Triple Exponential Smoothing	**4978.63**
ARIMA	8490.13
SARIMA	**4302.07**

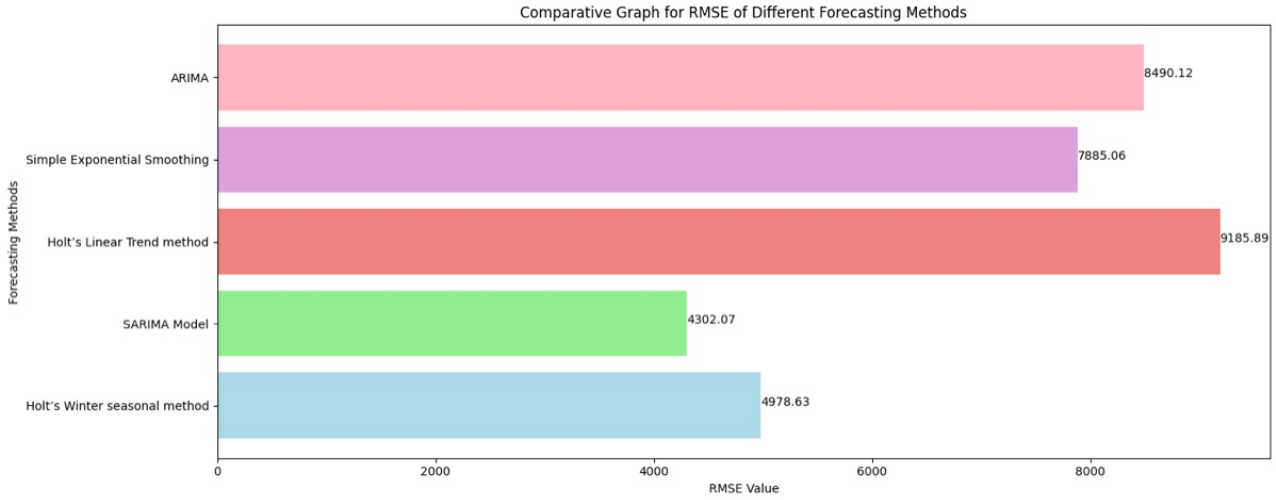





Train
Test
Holt_Winter
Total Journeys
Months
2014-04
2014-05
2014-06
2014-07
2014-08
2014-09
2014-10


RMSE: 8490.1293
Given set
Predict
2014-09-13
2014-09-15
2014-09-17
2014-09-19
2014-09-21
2014-09-23
2014-09-25
2014-09-27
2014-09-29


Train
Test
SARIMA
Total Journeys
Months
2014-04
2014-05
2014-06
2014-07
2014-08
2014-09
2014-10



Conclusion

Holt Winter's and SARIMA with small RMSE values 4978.63 and 4302.7 respectively gave the best result for our dataset compared to other models used. These models effectively captured both trend and seasonality in Uber pickup data, making them ideal choices for demand prediction. By using these models, Uber can optimize resource allocation, improve customer satisfaction, and enhance operational efficiency.

Future Work

- Extending analysis to more recent Uber data to observe shifts in demand patterns over time.
- Exploring advanced machine learning techniques, such as deep learning models (LSTMs), to improve prediction accuracy.
- Integrating external factors (weather, events, traffic data) into forecasting models to enhance predictions.
- Conducting a comparative study between time series forecasting and machine learning-based predictive analytics.

Time Series Semester VI Report

Prepared by: Asma Gite (D021), Sauleha Khan (D031)
Mentored by: Dr. Suresh Pathare