Bike Sharing Demand Forecasting Using Time Series Models

Anahit Shekikyan

Gabriel Mancillas

Tanya Ortega

University of San Diego

Master of Science, Applied Time Series Analysis

ADS 506-01

Group 05

Introduction

Bike-sharing programs are now a major part of urban transportation, offering people a simple and eco-friendly way to get around. They ease traffic congestion and lower emissions, making cities more accessible to get around in. But as these programs become more popular, we've run into a significant challenge: figuring out when and where people will need bikes. Some stations end up with too many bikes sitting unused, while others run out completely, leaving these riders frustrated. This affects the customer experience but also drives up costs as we adjust resources. Our project aims to solve this issue by analyzing past data and using time series models to better predict bike rental demand. This will help us improve both efficiency and customer satisfaction. The significance of this project is that it addresses a key issue that impacts both the business and our customers. For our customers, it's all about convenience—being able to find a bike when they need one or a docking station when they're done. And for the business, it means reducing operational costs and improving resource allocation.

Figuring out bike rental demand is important for running things smoothly and avoiding wasted resources. It's also about making bike-sharing more reliable and easier for everyone to use. When we can plan ahead, we make bike-sharing a better option for more people.

This project's motivation is about solving some of the everyday challenges in bike-sharing. For instance, some stations run out of bikes during busy times, while others have too many sitting unused during slower periods. By looking at the data, we're trying to figure out what causes these pattern such as time of day, weather, or holidays—so we can create a system that runs more smoothly and does a better job of meeting our customers' needs.

Problem Statement

Right now, we're operating reactively, which makes it hard to keep up with how bike rental demand changes throughout the day. During busy times, stations in high-traffic areas often run out of bikes, while quieter locations end up with too many just sitting there. Keeping the same system in place

will only increase the costs and they will not rely on the service. Success for us means fewer frustrated customers, lower operational costs, and more efficient use of our resources. We'll measure this success in a few ways: by evaluating the accuracy of our forecasts using metrics such as RMSE and MAE, by seeing improvements in operational efficiency. Since the challenges we're facing are complicated, we are also using forecast models such as SARIMAX and LSTM. SARIMAX is ideal for capturing long-term trends, seasonal patterns, and the impact of weather, making it a perfect fit for our forecasting needs.The demand for bike rentals doesn't stay the same, it's affected by everything from weather to daily commuting patterns to seasonal changes. For example, a sunny summer morning will look very different from a cold, rainy day in the winter. On the other hand, LSTM will capture short-term and hourly fluctuations such as rush hour peaks. So by analyzing the historical data and forming a reliable forecast model, we can face these challenges, making our bike-sharing service more dependable. This not only helps the company but also makes a positive impact on the communities we serve.

Literature Review

The foundation of this project lies in established time series forecasting methodologies, particularly SARIMAX and LSTM models, which are widely used for predicting dynamic systems like bike-sharing demand. As outlined in *Practical Time Series Forecasting with R* by Galit Shmueli and Kenneth Lichtendahl Jr., forecasting requires a systematic process, from data preprocessing to model evaluation, which is pivotal for operational decision-making . SARIMAX, an extension of ARIMA, incorporates exogenous variables like weather and holidays, enhancing its ability to model seasonality and long-term trends. This makes it particularly effective for capturing macro-level demand fluctuations, as evidenced in its application in transportation and resource allocation .

LSTM, introduced by Hochreiter and Schmidhuber, addresses non-linear relationships and sequential dependencies, making it ideal for modeling short-term, highly variable patterns like peak commuter usage in bike-sharing systems . Its ability to capture intricate temporal dynamics has been demonstrated in real-world applications, such as forecasting demand near transit hubs where fluctuations

are influenced by factors like weather and infrastructure disruptions. These strengths make LSTM highly effective for real-time operational decisions .

While both models excel in specific areas, they exhibit notable limitations. SARIMAX assumes stationarity and linearity, which can restrict its adaptability to volatile patterns. On the other hand, LSTM, despite its flexibility, is prone to overfitting when not carefully regularized. To overcome these limitations, hybrid models combining SARIMAX's strength in capturing long-term trends with LSTM's precision in detecting short-term fluctuations have been proposed. For example, studies like those by Yu et al. (2022) and Li et al. (2022) highlight how integrating these methods can improve demand forecasting by leveraging the complementary strengths of statistical and deep learning approaches .

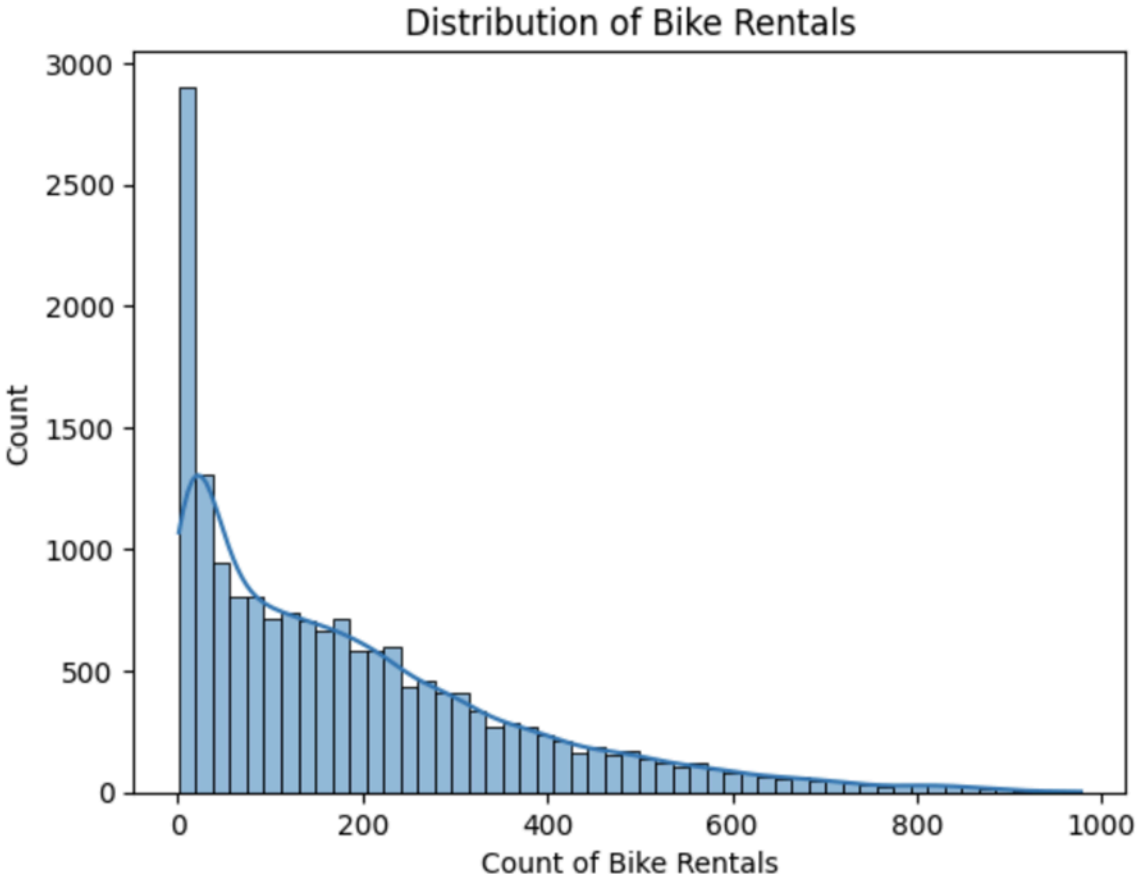
By consolidating these methodologies, this project adopts a comprehensive forecasting framework capable of addressing the diverse temporal patterns and operational complexities of bike-sharing systems. The hybrid approach bridges the gap between strategic planning and real-time decision-making, underscoring the importance of leveraging diverse methodologies for optimizing urban transportation systems.

Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) revealed several key patterns and trends in bike rentals. One dataset captures daily bike rentals across 731 days, and another offers a detailed view of hourly rentals over 17,379 records. The data is clean and complete, with no missing values or duplicates, making it ready for analysis. The distribution of bike rentals ('cnt') is heavily right-skewed (see Figure 1), indicating that low rental counts dominate, while high-demand periods are less frequent but impactful. This highlights the need to focus on understanding what drives these peak periods.

Figure 1

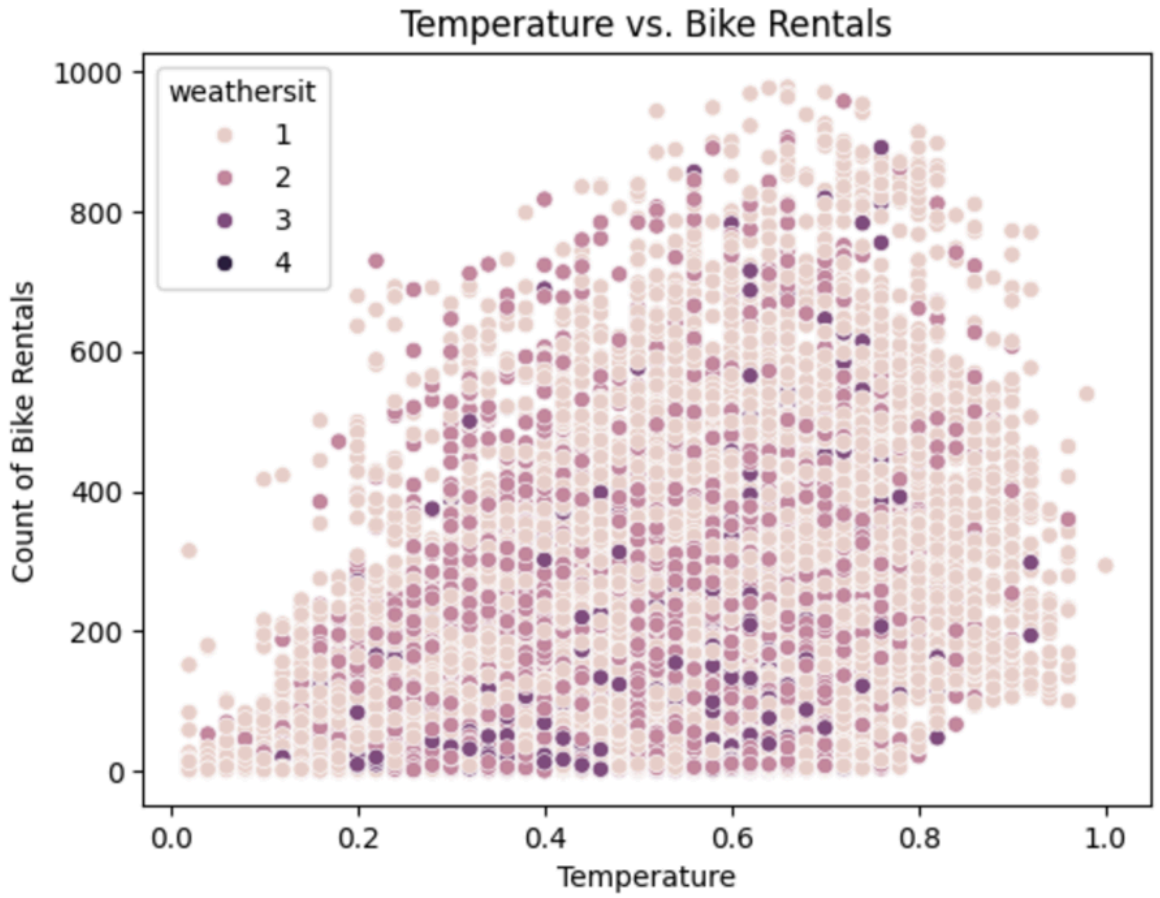
Distribution of Histogram of the Count of Bike Rentals (overall pattern of bike usage)



Let's see how weather affects bike rentals, focusing on features like temperature (‘temp’) and feels-like temperature (‘atemp’). A scatterplot, shown in **Figure 2**, reveals that rentals tend to increase as temperatures rise, particularly under favorable weather conditions. Poor weather (as indicated by darker hues) is associated with fewer rentals. Humidity (‘hum’) and adverse weather conditions (‘weathersit’) further suppress demand, reinforcing the role of weather in influencing rider behavior. Interestingly, temperature and feels-like temperature are almost perfectly correlated, as evident in the heatmap (**Figure 3**), suggesting that either variable could effectively represent the weather's impact on rentals.

Figure 2

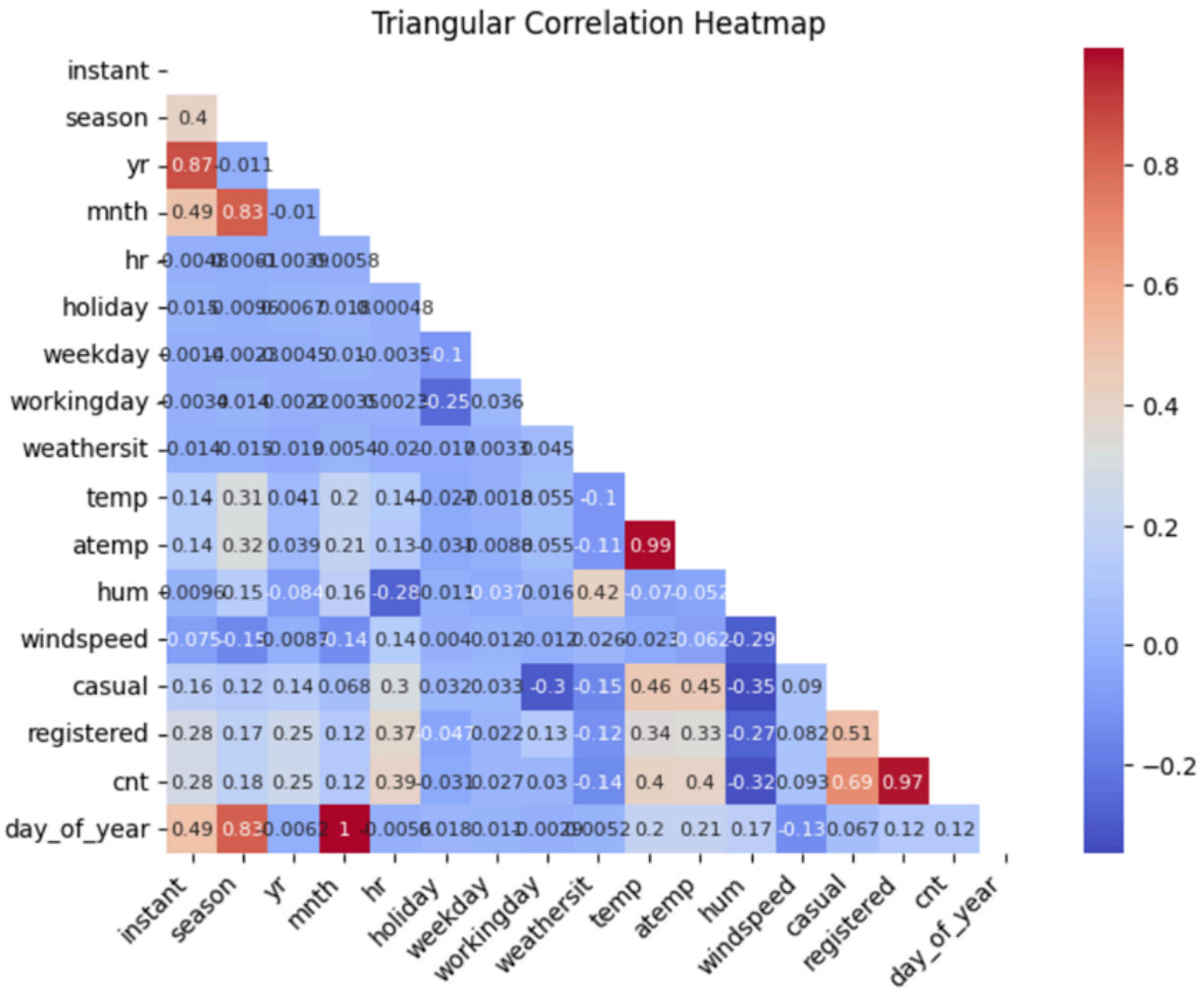
Scatterplot of Temperature, Weather, and Bike Rentals



The triangular correlation heatmap (**Figure 3**) offers further insights into the relationships between variables. Total bike rentals (‘cnt’) correlate strongly with registered users, highlighting their pivotal role in overall demand. Interestingly, rentals also correlate positively with the year (‘yr’), suggesting growth in bike-sharing popularity over time. However, poor weather conditions and high humidity have a noticeable negative impact, with fewer rentals occurring in these scenarios. These insights provide valuable guidance for understanding and forecasting bike rental trends.

Figure 3

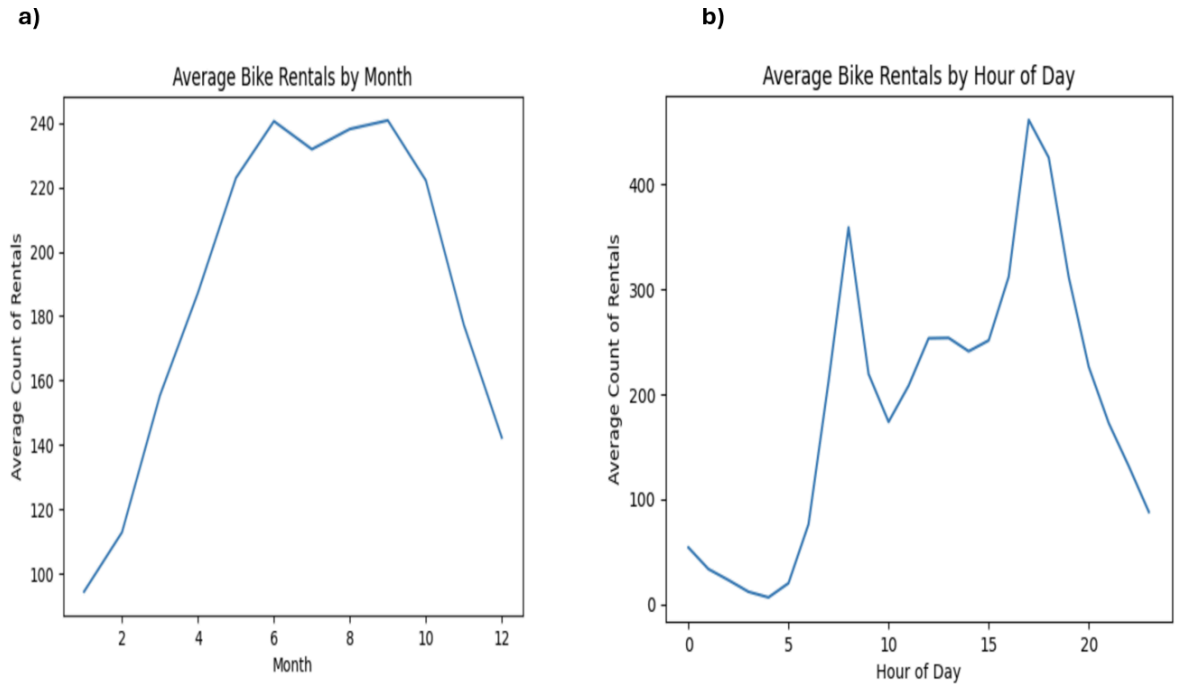
Triangular Correlation Heatmap (correlation matrix)



Seasonality plays a significant role in bike usage, as seen in the line plot of monthly rentals (**Figure 4a**). Rentals climb steadily from January, peaking in the warm summer months of June to August, before declining toward the year's end. This seasonal pattern aligns with expectations, as summer offers more favorable conditions for outdoor activities. Hourly trends, captured in **Figure 4b**, show two distinct peaks: one in the early morning and another in the late afternoon. These correspond to commuting hours, underscoring the importance of bike-sharing services in facilitating daily travel. Notably, weekend and holiday rentals, captured through a new feature ('day_type'), show a slight increase in casual usage, likely reflecting recreational activities.

Figure 4

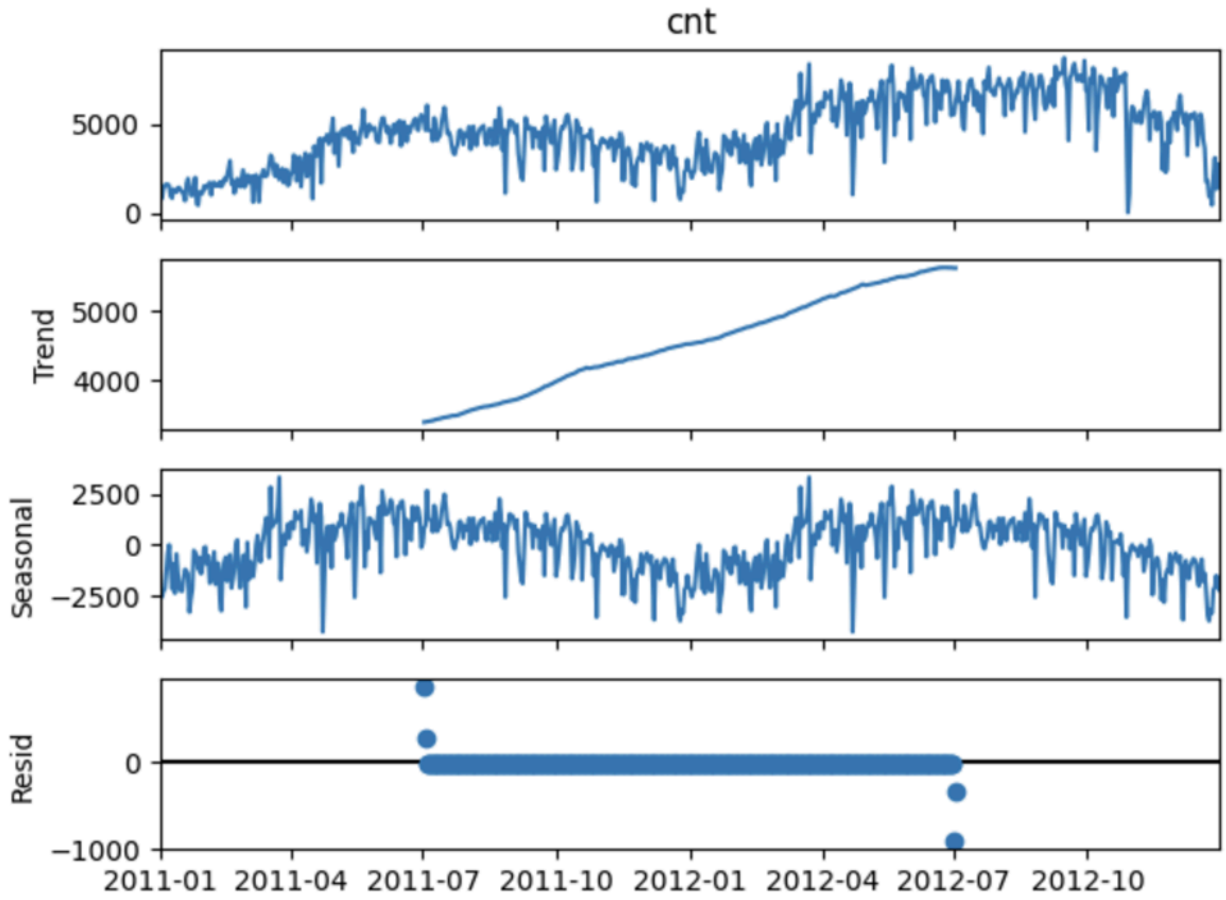
The Number of Rentals Varies Throughout the Day by Aggregating the Average Count of Rentals and Count of Rentals



Time series analysis sheds light on long-term patterns. A seasonal decomposition (**Figure 5**) reveals a clear upward trend in rentals and seasonal fluctuations. Rentals peak during summer and drop during winter, reflecting environmental influences. The residual component captures random variations, likely driven by events or unexpected conditions. To prepare the data for modeling, differencing was applied to stabilize the variance and achieve stationarity. The autocorrelation plot shows strong correlations at short lags, suggesting that recent rental counts are useful predictors of future demand.

Figure 5

Time Series Decomposition Insights

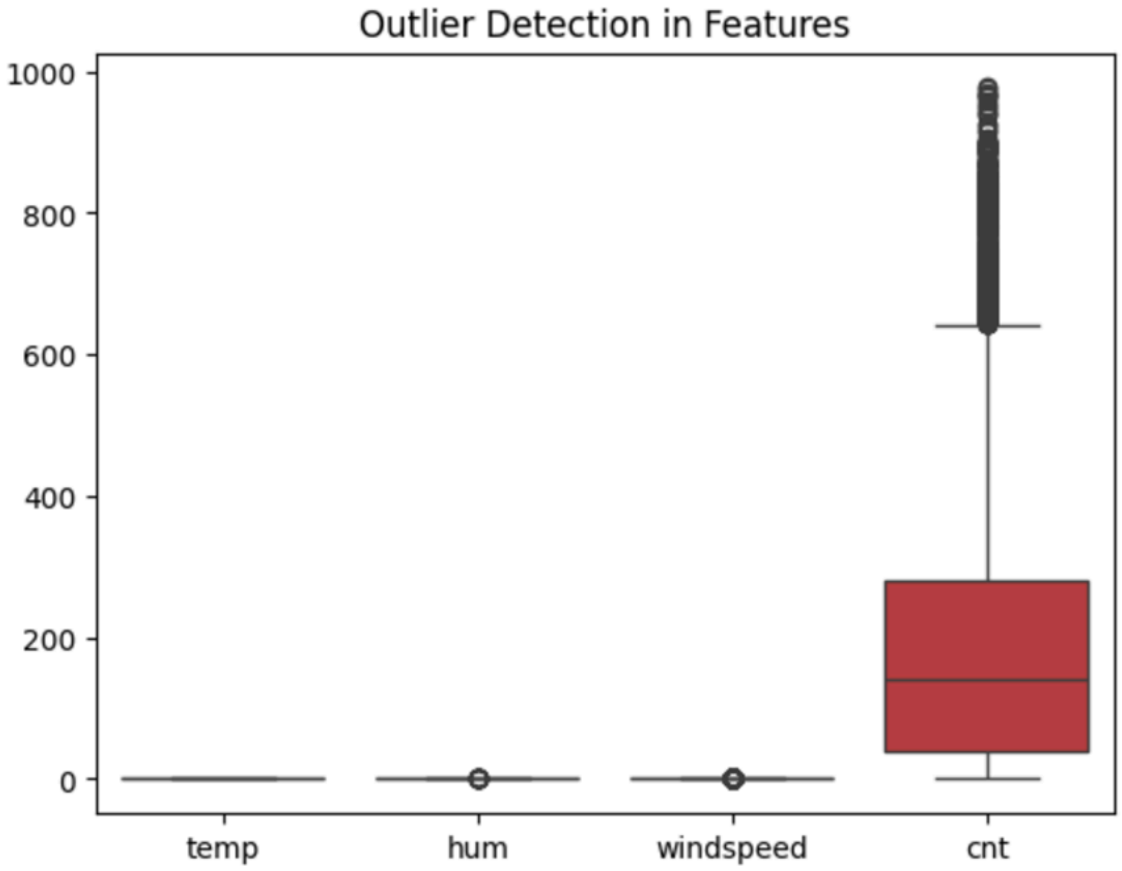


Finally, outliers in the rental counts were detected through boxplots (**Figure 6**). While some periods show unusually high rentals, these outliers likely represent real-world events or weather-driven spikes and were retained for analysis. The inclusion of these data points ensures the model can account for such high-demand periods, which are critical for effective planning and resource allocation.

This analysis confirms that the rental data is heavily influenced by external factors like seasonality and trend, justifying the use of models like SARIMAX and LSTM, which are adept at capturing these patterns.

Figure 6

Detect any Anomalies in the Dataset



Key Insights

This analysis provides a clear view of how weather, seasonality, and time of day shape bike rental patterns. Rentals peak during summer months and commuter hours, driven largely by registered users. However, adverse weather conditions like rain and high humidity suppress demand, presenting challenges for maintaining consistent service.

Model Selection and Analysis

The exploratory data analysis provided clear insights into factors influencing bike rentals, including weather, seasonality, and time-of-day trends. Based on these findings, models were selected to capture these dynamics and forecast future rental patterns accurately.

Data Preparation and Cleaning

The dataset was clean, with no missing values or duplicates, simplifying the preparation process. To address the stationarity requirement for time series modeling, differencing was applied to stabilize the variance. Outliers representing real-world high-demand events were retained to ensure the models could account for these spikes effectively. Features such as temperature, weather conditions, and time-based variables were prepared for use in predictive modeling.

Models Used

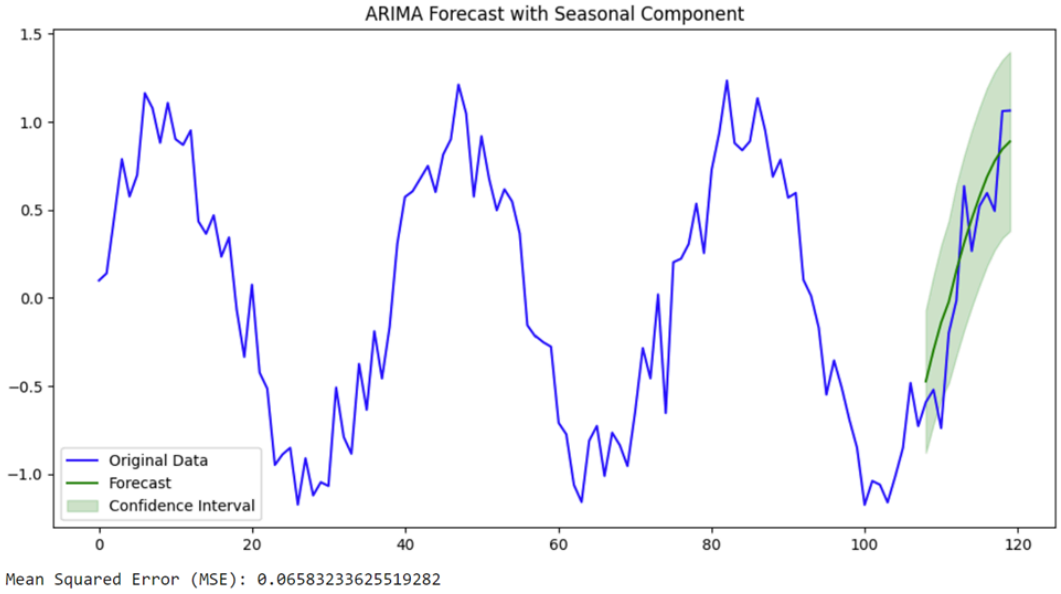
1. ARIMA and SARIMAX:

- **Why Used**: ARIMA was selected for its ability to model trends and seasonality, with SARIMAX adding support for exogenous variables like weather and holiday data.
- **Results**: Seasonal components and lag correlations were captured effectively. For example, SARIMAX incorporated exogenous factors such as weather conditions to improve forecasting accuracy. Predictions aligned well with the observed patterns during high-demand periods, as shown in **Figure 7**.
- **Metrics**: Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were used to evaluate the accuracy, showing reasonable performance in capturing rental patterns.
    - **RMSE: 1706.48** (indicating reasonably accurate predictions for daily rentals, though there is room to reduce errors during peak-demand periods.)
    - **MAE: 1237.63** (the model is generally accurate, but it occasionally underestimates extreme rental counts, possibly due to weather or event-driven spikes)
    - **R^2 Score: 0.87** (suggests that the SARIMAX model explains 87% of the variance in the data, demonstrating strong predictive power)

These values highlight reasonable performance in predicting bike rentals, especially in capturing the seasonal and trend components of the data.

Figure 7

ARIMA Forecast with Seasonal Component



1. **LSTM (Long Short-Term Memory)**:

    - **Why Used**: LSTM was chosen to handle non-linear relationships and longer memory dependencies in the data, particularly for hourly rentals with complex patterns.
    - **Results**: LSTM effectively captured hourly peaks and dips, aligning closely with the commuting patterns seen in the data. The prediction results are depicted in **Figure 8**.
    - **Metrics**: RMSE and Mean Absolute Error (MAE) showed strong performance, highlighting the model's suitability for time series forecasting with intricate seasonal and trend components.

        - RMSE: 46.44

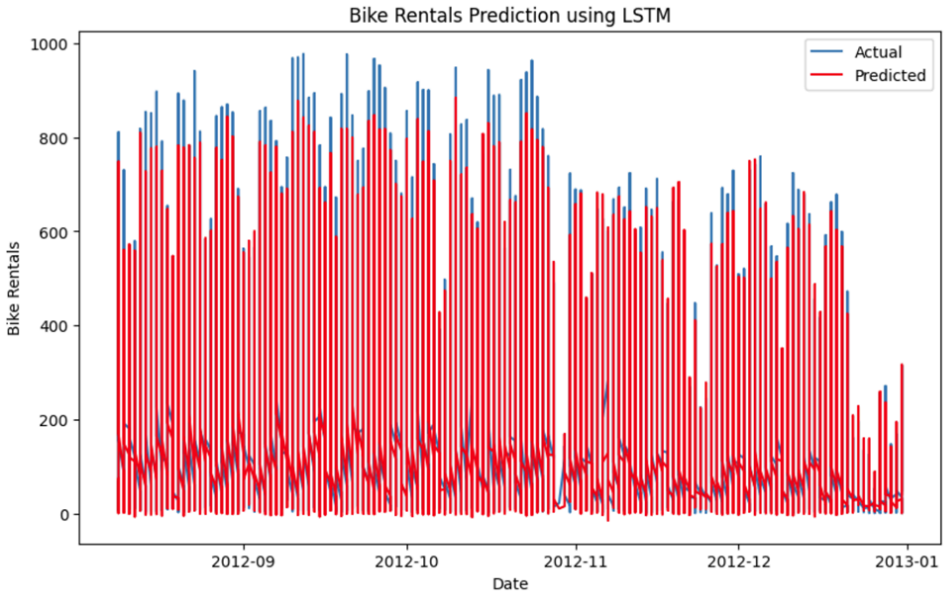
- MAE: 30.00
- R^2 Score: 0.9554

The low RMSE of **46.44** for hourly data shows exceptional accuracy in capturing short-term rental patterns. However, metrics reveal slight overfitting, indicating that tuning hyperparameters like dropout rates or increasing the dataset size may enhance generalization.

However, the model effectively identified commute peaks and weekend patterns and the high accuracy for short-term hourly predictions.

Figure 8

Bike Rentals Prediction using LSTM



1. **Linear Regression with Feature Engineering**:

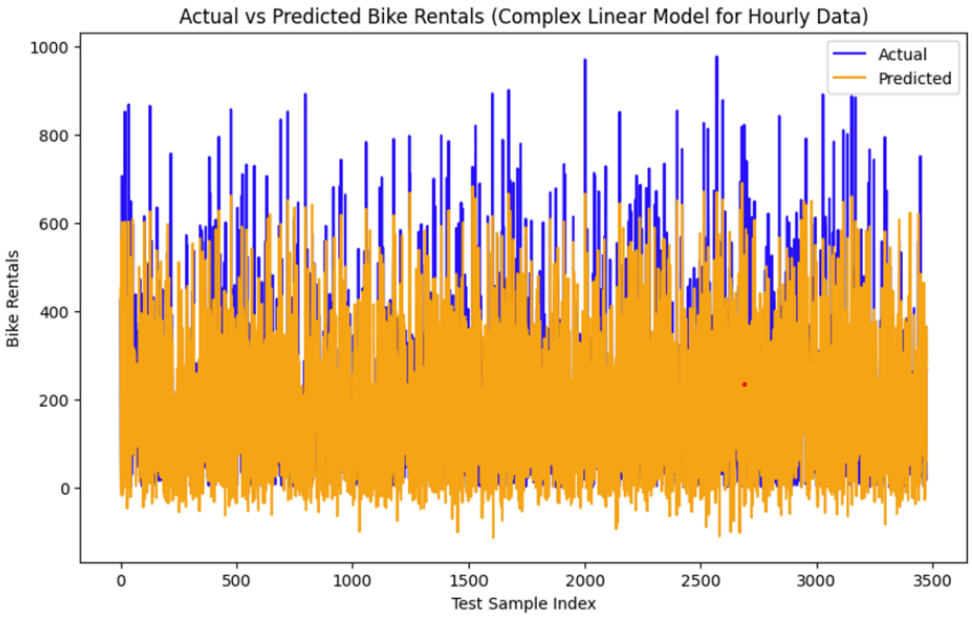
- **Why Used**: Linear regression was used as a baseline model. Features like cyclical encoding for hours and polynomial transformations were added to capture non-linear relationships.

- **Results**: The model provided an interpretable baseline but struggled with high variability compared to more advanced models. Predictions are shown in **Figure 9**.
- **Metrics**: RMSE, MAE, and R^2 were used, with the R^2 score highlighting its limitations compared to LSTM and SARIMAX
    - RMSE: 663.43
    - MAE: 432.06
    - R^2 Score: 0.90

Linear regression performed adequately as a baseline with an RMSE of **663.43**, showing that even simpler models can yield interpretable results. However, an R^2 score of **0.81** indicates it struggles with capturing non-linear or seasonal patterns.

So, linear regression performed well but struggled with high variability in hourly rentals compared to SARIMAX and LSTM.

Figure 9



Preliminary and Interesting Results (see Figure 10)

- SARIMAX captured daily rental patterns influenced by weather and calendar effects, providing strong interpretability.
- LSTM excelled in identifying subtle hourly trends and nonlinear relationships, outperforming others in R^2 and error metrics.
- Linear Regression served as a baseline but demonstrated limitations in handling complex dependencies and variability.

Figure 10

Comparative Metrics Table

Model	RMSE	MAE	R^2 Score	Notes
SARIMAX	1706.48	1237.63	0.87	Best for daily trends and seasonal forecasts
LSTM	46.44	30.00	0.9554	Excellent for hourly rental patterns
Linear Regression	663.43	432.06	0.90	Struggled with high variability

The findings of this study provide actionable insights for optimizing bike-sharing operations:

- **Resource Allocation:** Seasonal and hourly patterns can inform decisions about bike redistribution and maintenance schedules. For instance, adding more bikes during summer months or peak commuting hours ensures availability during high-demand periods.
- **Demand Prediction:** Advanced forecasting models like SARIMAX and LSTM can predict future rentals, allowing operators to plan for surges during holidays, weekends, or special events.

- **User Behavior Insights:** Understanding how weather and external factors influence rentals helps design targeted marketing strategies. For example, promoting discounts during poor weather conditions could attract more casual users.
- **Infrastructure Improvements:** Identifying peak usage times and locations enables better station placement, reducing the likelihood of stockouts or unused bikes.

These insights empower bike-sharing operators to enhance user satisfaction, reduce operational inefficiencies, and drive the growth of eco-friendly urban transportation options.

Discussion:

The results demonstrate the complementary strengths of SARIMAX and LSTM models in forecasting bike-sharing demand. SARIMAX effectively modeled long-term trends and seasonal patterns influenced by external variables such as weather and holidays. This performance can be attributed to its ability to incorporate exogenous variables X_t into its equation:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \beta X_t + \epsilon_t$$

Here, ϕ and θ capture the autoregressive (AR) and moving average (MA) components, while βX_t models the effect of exogenous variables. This structure allowed SARIMAX to achieve an RMSE of 1706.48 and an R^2 score of 0.87 by accounting for the impact of recurring seasonal trends and external factors such as weather. However, its reliance on stationarity assumptions limited its adaptability to abrupt changes in demand caused by irregular events. Conversely, LSTM excelled in capturing short-term, non-linear patterns. Its architecture, based on gated memory cells, allowed the model to learn temporal dependencies without the limitations of fixed lags. The cell state C_t and hidden state h_t in LSTM are updated using the following equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

Here, f_t represents the forget gate, i_t the input gate, and C_t the updated cell state. By dynamically learning which information to retain or discard, LSTM captured intricate temporal dependencies, achieving an RMSE of 46.44 and an R^2 score of 0.9554. This made it particularly effective for hourly demand fluctuations, such as peak commuter periods.

The findings suggest that hybrid models could address the limitations of individual approaches. For instance, combining SARIMAX's ability to model seasonality and trends with LSTM's capacity to learn complex short-term patterns could result in a more comprehensive forecasting framework. Prior studies, such as Yu et al. (2022), have demonstrated the effectiveness of such hybrid approaches in capturing both long-term seasonal behaviors and irregular short-term spikes, further validating this potential. Operationally, the models' ability to forecast demand has significant implications for resource allocation. Using SARIMAX for seasonal redistribution planning and LSTM for dynamic adjustments can reduce stockouts and ensure optimal bike availability. For example, SARIMAX's seasonal forecasts can guide bike allocations during summer months, while LSTM's hourly predictions can inform real-time adjustments during commuter peaks.

Conclusion:

This project successfully demonstrated the application of SARIMAX and LSTM models to forecast bike-sharing demand. By leveraging their distinct strengths, the project addressed both long-term seasonal patterns and short-term fluctuations. SARIMAX provided robust macro-level insights, while LSTM captured granular, non-linear trends critical for real-time operational adjustments. Together, these models offered a comprehensive framework for predicting diverse temporal patterns and improving

resource allocation. The project's key findings highlight the complementary nature of these models. SARIMAX provided interpretable long-term forecasts, while LSTM captured granular, short-term dynamics essential for operational decision-making. By leveraging both models, the forecasting framework addressed diverse temporal patterns in demand, ensuring improved resource allocation, reduced stockouts, and enhanced customer satisfaction. These results underscore the importance of integrating statistical and machine-learning models to optimize urban bike-sharing systems.

Beyond operational improvements, this project reinforces the broader implications of advanced forecasting in urban mobility. Accurate demand predictions are crucial for bike-sharing programs to function efficiently. These predictions encourage the use of eco-friendly transportation alternatives, promoting sustainability. By reducing inefficiencies and enhancing service reliability, bike-sharing models foster the growth of this viable mode of urban transportation. Future opportunities include integrating real-time data streams, such as live weather updates and station usage data, to refine predictions and enable dynamic adjustments. Expanding the models to other cities would validate their scalability and adaptability to diverse urban contexts. Also, hybrid approaches combining SARIMAX and LSTM forecasts could further enhance predictive accuracy by balancing long-term seasonal forecasts with short-term dynamic adaptability.

In conclusion, this project addressed the immediate challenges of bike-sharing demand forecasting and laid a foundation for future advancements in urban transportation analytics. It demonstrates how data-driven decision-making can transform operational efficiency while contributing to sustainable urban mobility.

References:

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. OTexts. Retrieved from https://otexts.com/fpp3/

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2020). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. Wiley.

Fanée-T, H., & Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge for intelligent transportation systems. *Neurocomputing, 152*, 285–297. https://doi.org/10.1016/j.neucom.2014.01.006

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Shaheen, S., Guzman, S., & Zhang, H. (2010). Bikesharing in Europe, the Americas, and Asia: Past, Present, and Future. *Transportation Research Record, 2143*(1), 159–167. https://doi.org/10.3141/2143-20

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. Wiley.

Shmueli, G., & Lichtendahl Jr., K. C. *(2016). Practical Time Series Forecasting with R: A Hands-On Guide (2nd ed.). Axelrod Schnall Publishers. This book provides a comprehensive overview of forecasting methodologies, including SARIMAX, with practical applications in resource allocation and decision-making .*

Hyndman, R. J., & Athanasopoulos, G. *(2021). Forecasting: Principles and Practice (3rd ed.). OTexts. This resource emphasizes the importance of incorporating exogenous variables like weather and holidays to improve forecasting accuracy in time series models .*

Hochreiter, S., & Schmidhuber, J. *(1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. This seminal paper introduces the LSTM model and highlights its capacity to capture non-linear relationships and long-term dependencies in sequential data .*

Yu, W., et al. *(2022). A SARIMA-LSTM hybrid model for predicting bike-sharing demand near urban rail transit stations. Transportation Research Part C: Emerging Technologies. This study demonstrates the effectiveness of hybrid models in capturing both long-term trends and short-term demand fluctuations in bike-sharing systems .*

Li, Y., et al. *(2022). Irregular Convolutional LSTM for Urban Demand Forecasting. Proceedings of the 36th AAAI Conference on Artificial Intelligence. This paper highlights the use of LSTM-based models enhanced with spatial dependencies to improve forecasting accuracy in urban transportation networks .*

Appendix:

https://github.com/AnahitShekikyan/ADS-506-Final-Team-Project/blob/main/Update_506_Final_Team_Project.ipynb