DATA VISUALIZATION WITH BIKE SHARING DEMAND ANALYSIS

A PROJECT REPORT

Submitted by,

20201ISB0014 – RAGHAVENDRA N

20201ISB0015 – NISHANTH J

2020ISB0016 - YESHWANTH S GOWDA

Under the guidance of,
Ms. POORNIMA S

in partial fulfillment for the award of the degree of
BACHELOR OF ENGINEERING

IN

INFORMATION SCIENCE AND ENGINEERING

At

PRESIDENCY UNIVERSITY
BENGALURU
JANUARY 2024

ABSTRACT

The aim of this analysis is to identify factors influencing the number of bike trips and develop a predictive model to forecast hourly bike demand based on environmental conditions. Lyft, Inc., a major transportation network company, operates a significant bike-sharing service across the United States and Canada. Accurate demand forecasting is essential for effective planning of bicycles, stations, and maintenance personnel to meet varying levels of demand. The dataset "hour.csv" provides detailed records, including variables such as date, season, year, month, hour, holiday status, weekday, working day status, weather conditions, normalized temperature, humidity, wind speed, and counts of casual, registered, and total users.

The study begins with data cleaning and preprocessing to handle missing values and normalize data. This ensures that the dataset is ready for analysis and modeling. Exploratory Data Analysis (EDA) is conducted to visualize trends and patterns in bike usage, helping to uncover the impact of different factors such as seasonality, weather, and time of day on bike demand. Through EDA, significant correlations and patterns are identified, providing a better understanding of the data.

Following EDA, various predictive models are developed, including linear regression, decision trees, and advanced machine learning algorithms. These models are trained and tested on the dataset, with performance evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. These metrics provide a quantitative measure of the model's accuracy and reliability in predicting bike demand.
The insights gained from this analysis highlight the influence of various environmental and temporal factors on bike-sharing demand. The predictive model developed from this study enables Lyft to optimize their bike-sharing operations by accurately forecasting demand. This leads to better allocation of bicycles, strategic placement of docking stations, and efficient deployment of maintenance personnel. As a result, Lyft can enhance service availability and improve customer satisfaction, supporting efficient operational planning and resource management. The findings of this analysis not only support Lyft's operational goals but also contribute to the broader understanding of bike-sharing dynamics in urban environments, facilitating better urban mobility solutions.

TABLE OF CONTENTS

CHAPTER-1

INTRODUCTION

In recent years, bike-sharing services have emerged as a popular mode of urban transportation, offering an eco-friendly, cost-effective, and convenient alternative for short-distance travel. Lyft, Inc., a prominent transportation network company based in San Francisco, California, operates one of the largest bike-sharing systems in the United States and Canada. With operations spanning 640 cities in the United States and 9 cities in Canada, Lyft's bike-sharing service plays a significant role in urban mobility.

Accurate demand forecasting is essential for the efficient management of bike-sharing systems. It enables optimal allocation of bicycles, strategic placement of docking stations, and effective deployment of maintenance personnel. Understanding the factors that influence bike usage is crucial for anticipating demand and ensuring the availability of bikes when and where they are needed. This study aims to analyze various environmental and temporal factors affecting bike demand and to develop a predictive model to forecast the number of bike trips during any given hour.

The dataset used for this analysis, "hour.csv," contains detailed records of bike-sharing usage. It includes variables such as date, season, year, month, hour, holiday status, weekday, working day status, weather conditions, normalized temperature, humidity, wind speed, and the counts of casual, registered, and total users. These variables provide a comprehensive view of the conditions that can affect bike usage.

The analysis will involve data cleaning and preprocessing to ensure the dataset is suitable for modeling. Exploratory Data Analysis (EDA) will be conducted to uncover trends and patterns in bike usage. Various predictive models, including regression and machine learning algorithms, will be developed and evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. The goal is to build a robust predictive model that can accurately forecast bike demand, aiding Lyft in optimizing their bike-sharing operations and enhancing service availability and customer satisfaction.

By leveraging data-driven insights and predictive modeling, this study seeks to support Lyft's operational planning and contribute to the broader understanding of bike-sharing dynamics in urban environments, ultimately facilitating better urban mobility solutions.

CHAPTER-2

METHODOLOGY

The methodology for analyzing and predicting bike-sharing demand involves several steps, including data preprocessing, exploratory data analysis (EDA), feature engineering, model development, and model evaluation. Here's a detailed outline of the steps involved:

1. Data Preprocessing

Data Cleaning: Identify and handle any missing or inconsistent data. Ensure all variables are in appropriate formats for analysis.

Check for missing values and impute or remove them as necessary.

Convert categorical variables such as **season**, **weathersit**, **weekday**, and **mnth** into factor types.

Normalization: Ensure that continuous variables such as **temp**, **atemp**, **hum**, and **windspeed** are appropriately scaled for analysis.

Since the dataset already contains normalized values, verify the range and distribution.

2. Exploratory Data Analysis (EDA)

Descriptive Statistics: Calculate summary statistics for all variables to understand their distributions and central tendencies.

Mean, median, standard deviation, and range for continuous variables.

Visualizations: Use various plots to visualize the relationships between variables and identify trends.

Time series plots for bike counts over different periods (hour, day, month).

Boxplots and histograms for visualizing the distribution of continuous variables.

Bar plots and heatmaps for categorical variables and their relationship with bike demand.

Correlation Analysis: Calculate correlation coefficients between continuous variables to understand their relationships.

Identify multicollinearity and potential issues with highly correlated predictors.

•**3. Feature Engineering**

Datetime Features: Extract additional features from the **dteday** variable if necessary.

Create new variables such as **day_of_month**, **week_of_year**, or **is_weekend** to capture temporal patterns.

Interaction Terms: Consider creating interaction terms between variables that might have combined effects on bike demand.

For example, interaction between **temp** and **humidity**.

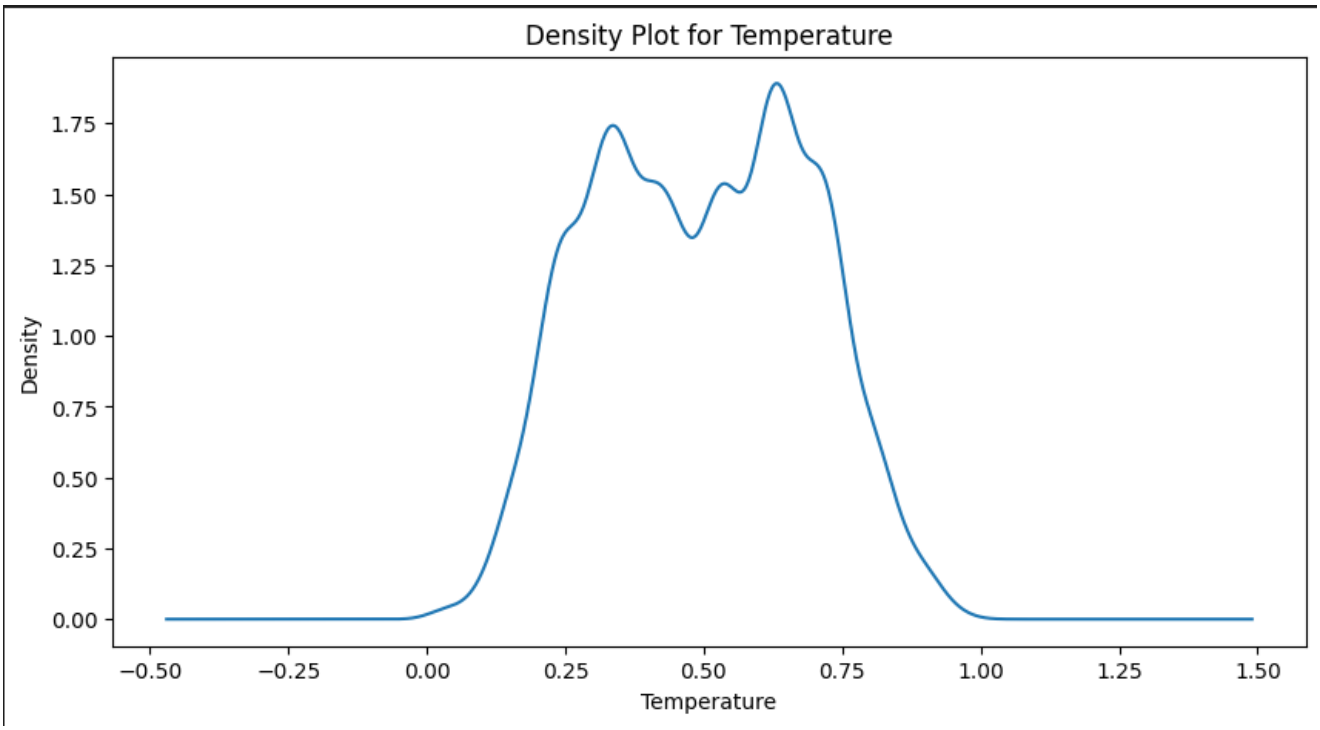
Model Development

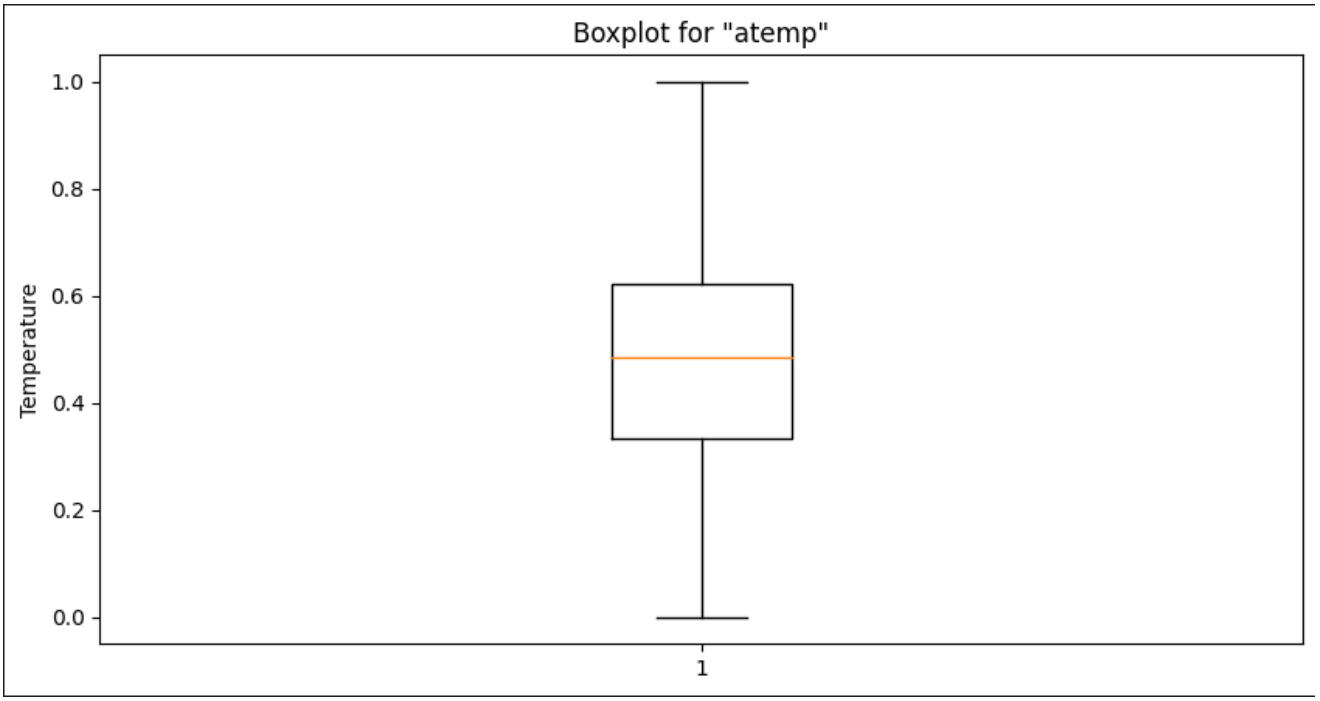
Train-Test Split: Split the dataset into training and testing sets to evaluate model performance on unseen data.
Typically, an 80-20 split is used.
Model Selection: Develop and compare various predictive models to determine the best approach for forecasting bike demand.
Linear Regression: Start with a baseline linear regression model to understand basic relationships.
Decision Trees: Use decision tree algorithms to capture non-linear relationships.
Random Forests: Apply random forests to improve predictive performance and handle overfitting.
Gradient Boosting Machines (GBM): Use GBM to enhance accuracy through boosting techniques.
Neural Networks: Explore deep learning models for complex pattern recognition if necessary.

5. Model Evaluation
Performance Metrics: Evaluate model performance using appropriate metrics.
Mean Absolute Error (MAE): Measures the average magnitude of errors in predictions.
Root Mean Squared Error (RMSE): Measures the square root of the average squared differences between predicted and actual values.
R-squared: Indicates the proportion of variance in the dependent variable explained by the model.
Cross-Validation: Perform k-fold cross-validation to ensure model robustness and generalizability.
Helps in assessing the model's performance across different subsets of the data.
Hyperparameter Tuning: Optimize model parameters using techniques such as grid search or random search.
For models like random forests and GBM, tune parameters like the number of trees, maximum depth, and learning rate.

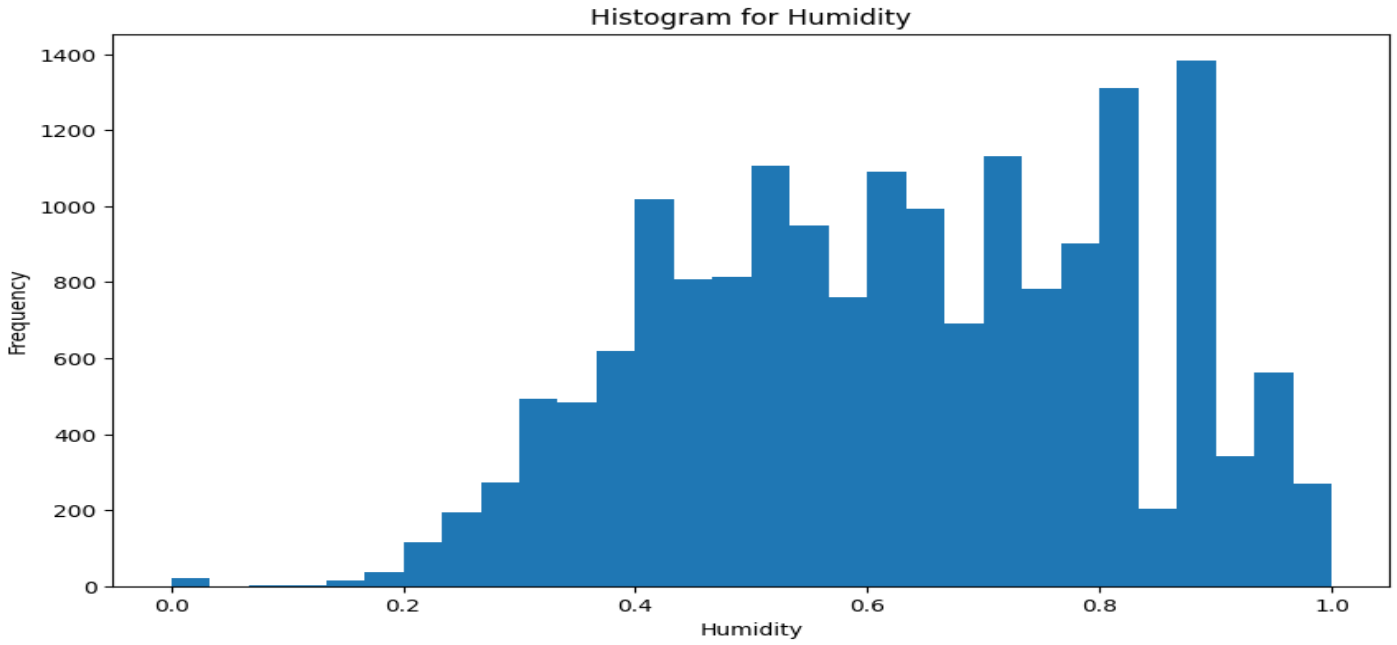
6. Model Deployment
Final Model Selection: Choose the best-performing model based on evaluation metrics and validation results.
Implementation: Prepare the model for deployment, ensuring it can be integrated into Lyft's operational systems for real-time predictions.
Monitoring and Maintenance: Set up monitoring mechanisms to track model performance over time and update the model as needed based on new data and changing conditions.

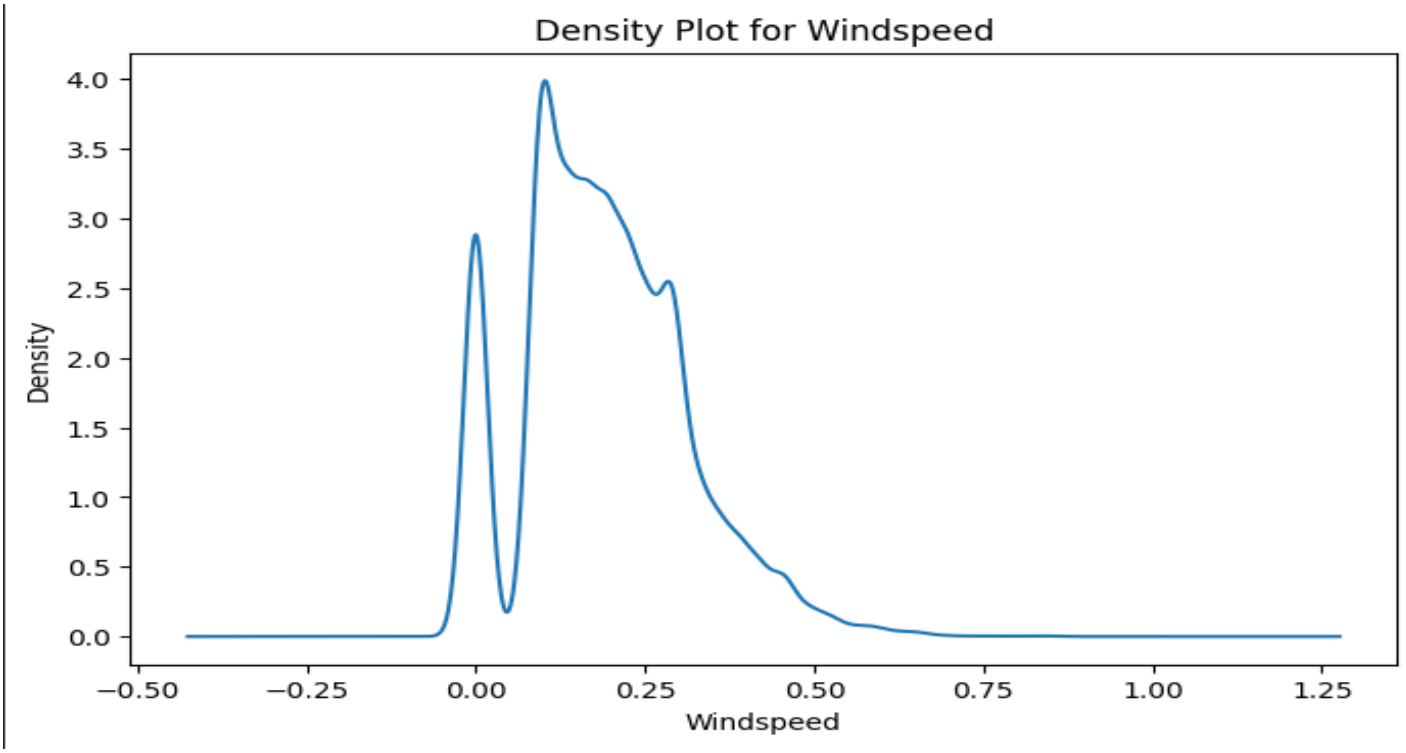
CHAPTER-3
RESULTS



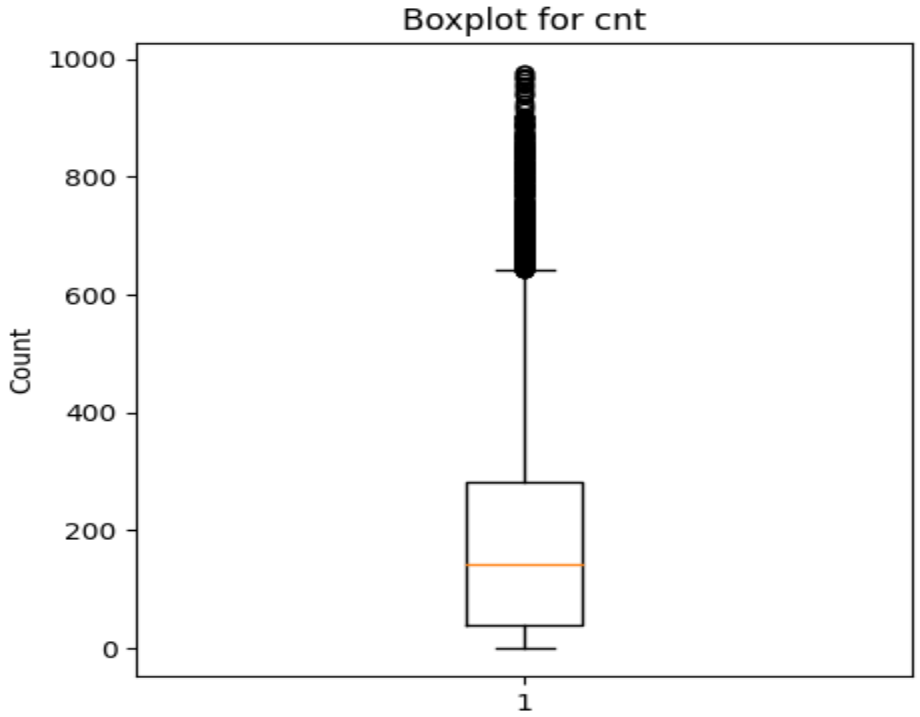


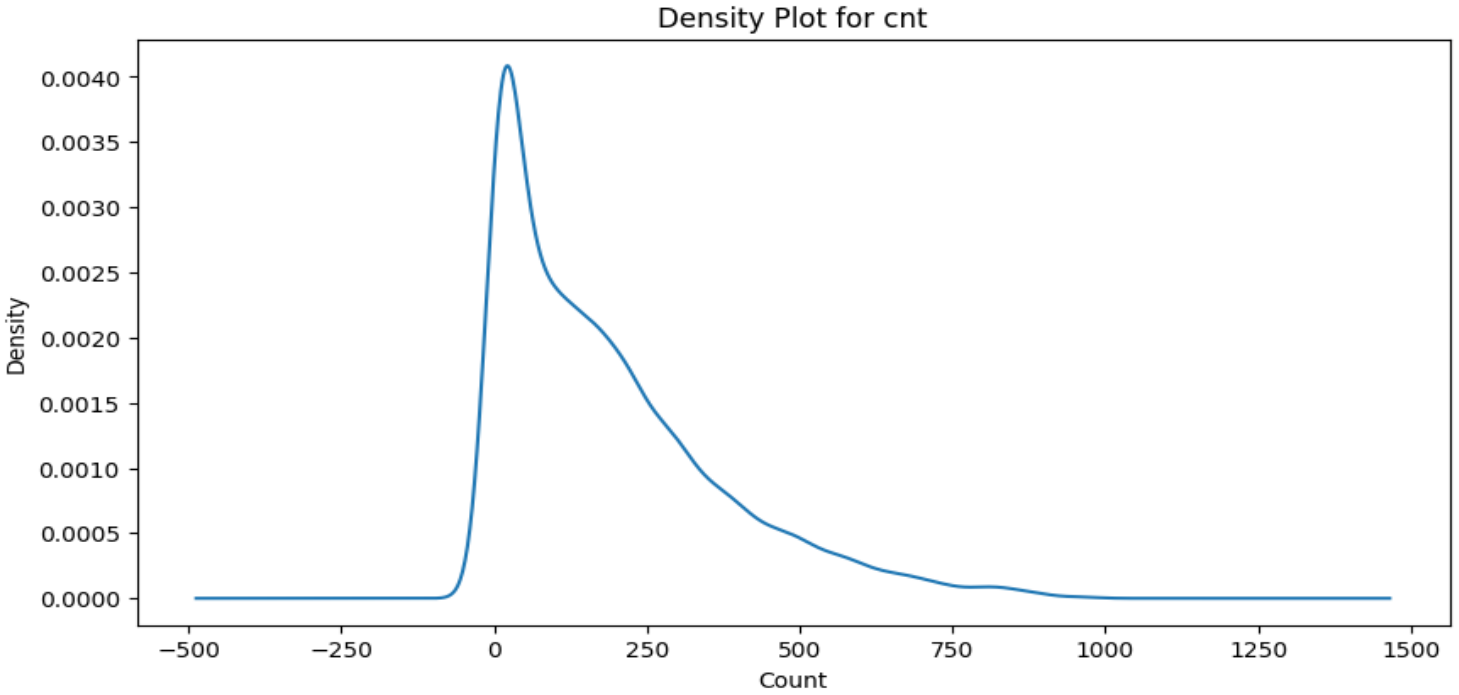
CHAPTER-3
RESULTS



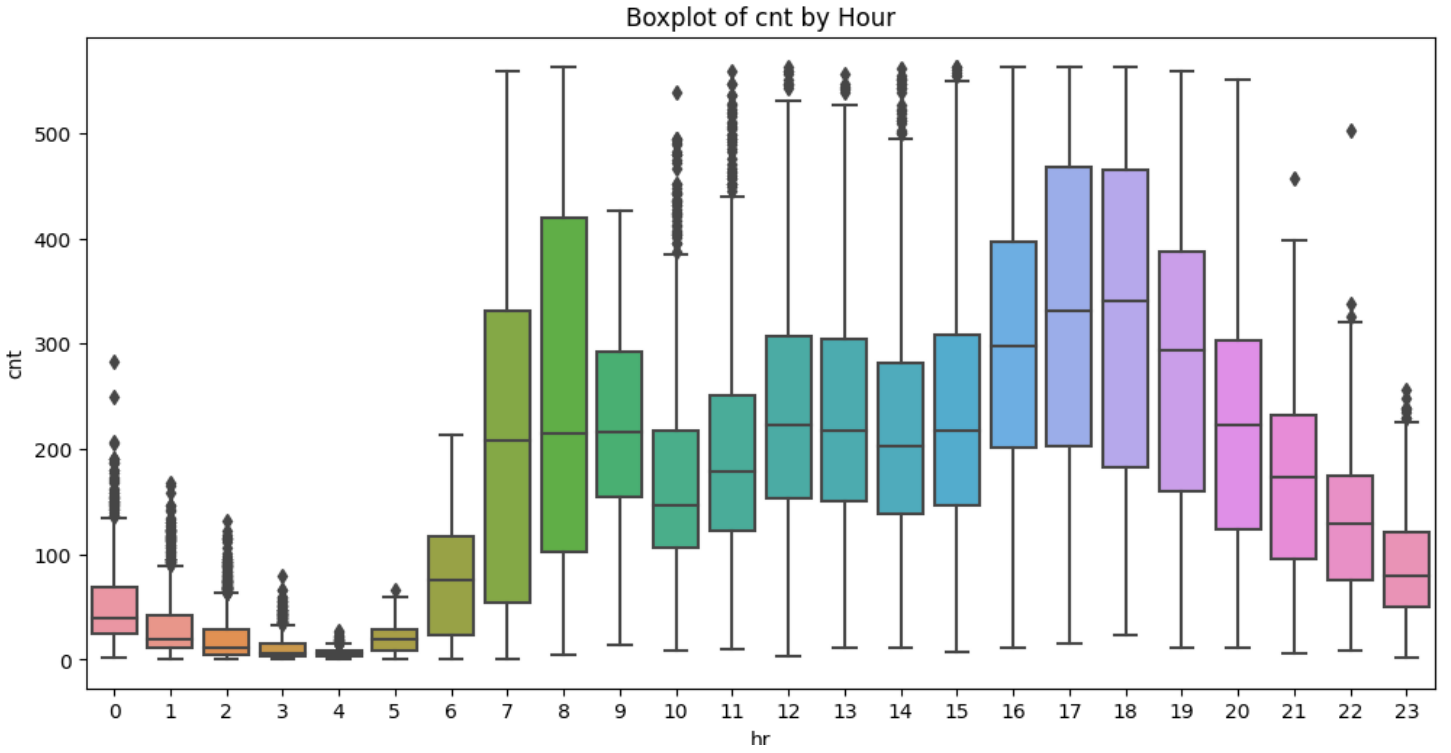


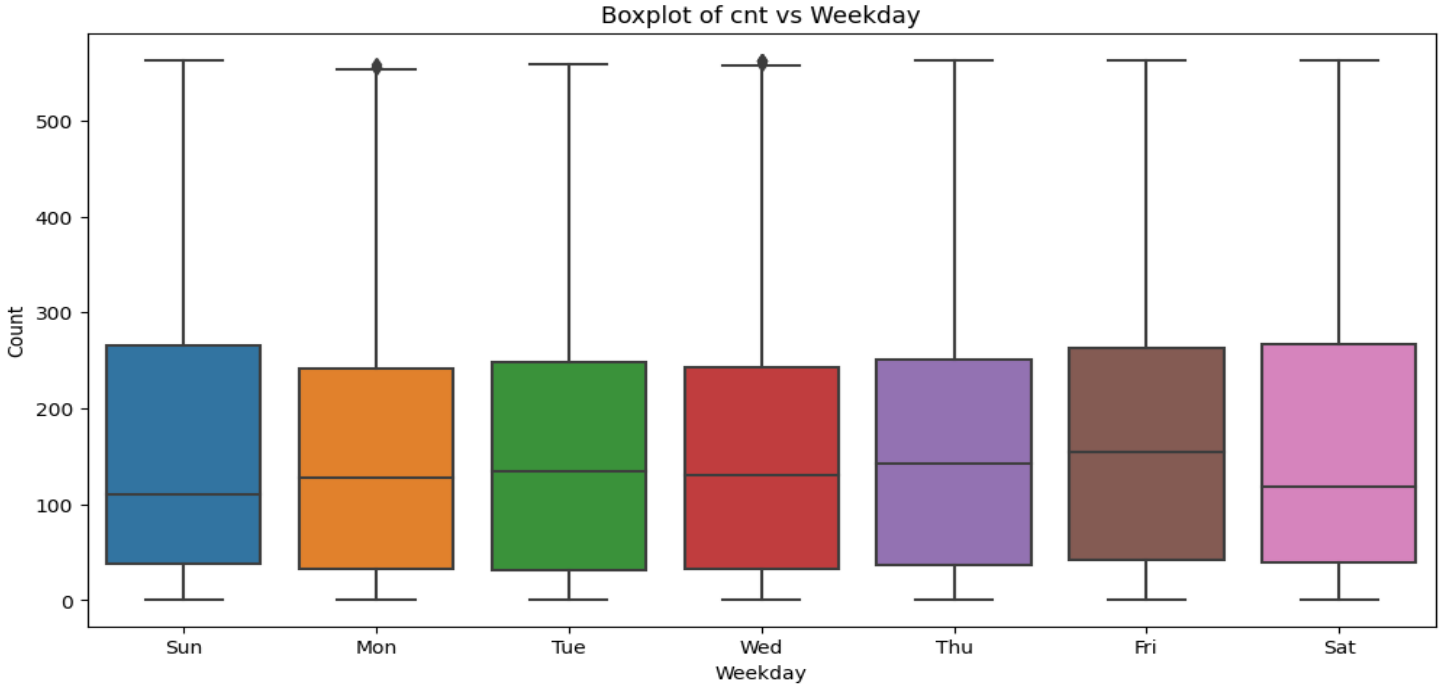
CHAPTER-3
RESULTS



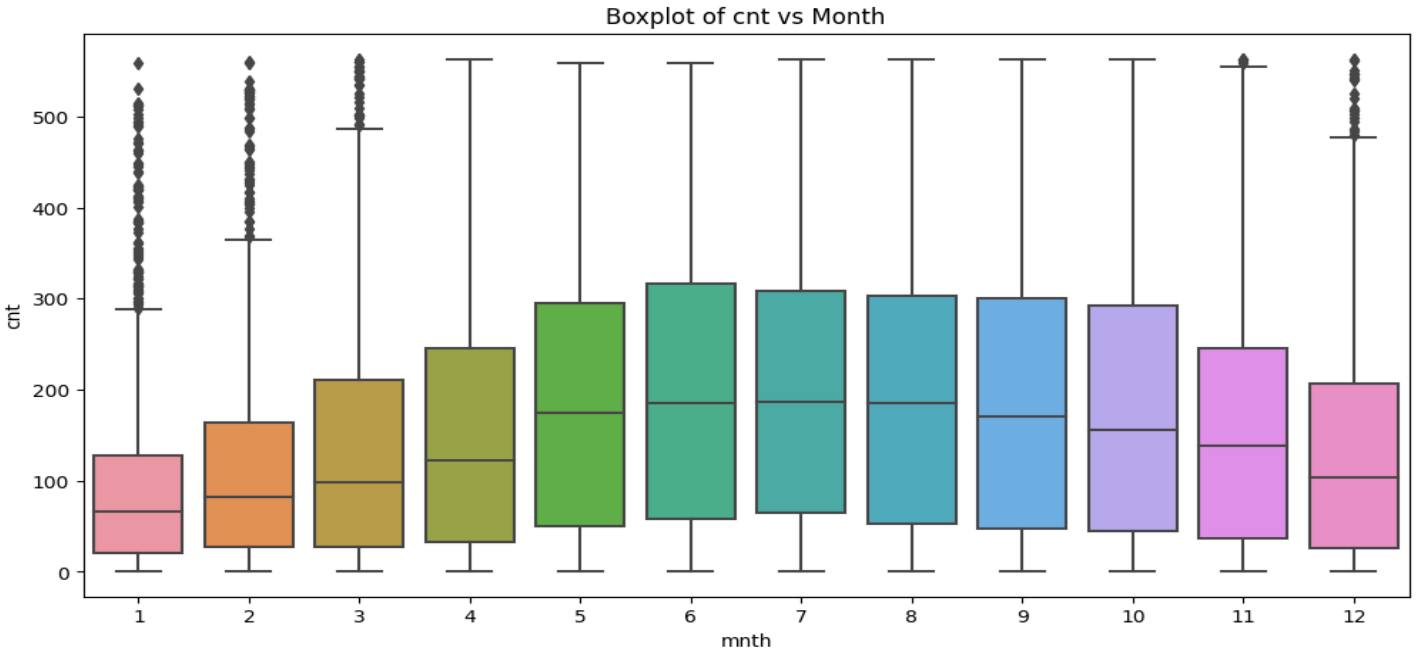


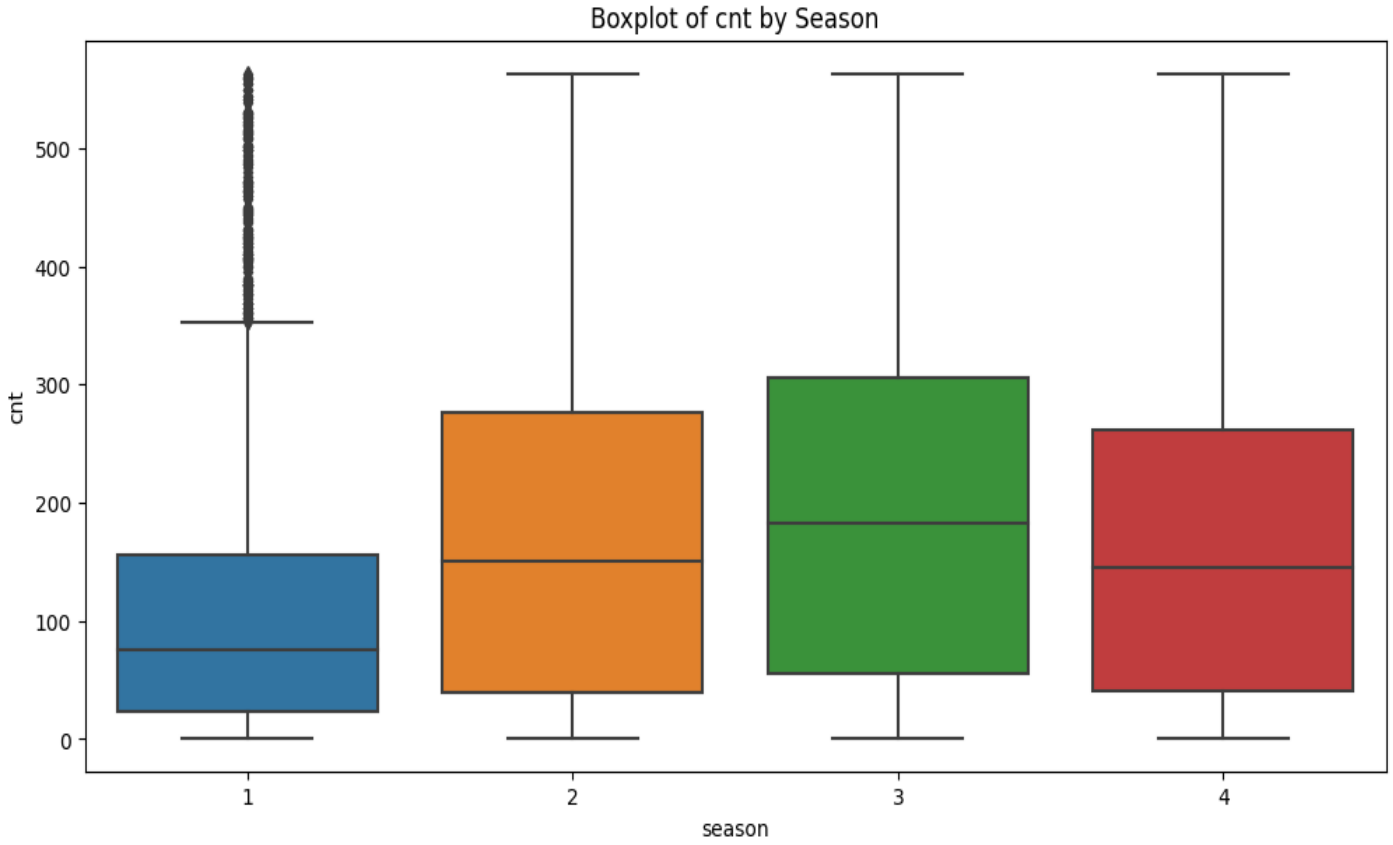
CHAPTER-3
RESULTS





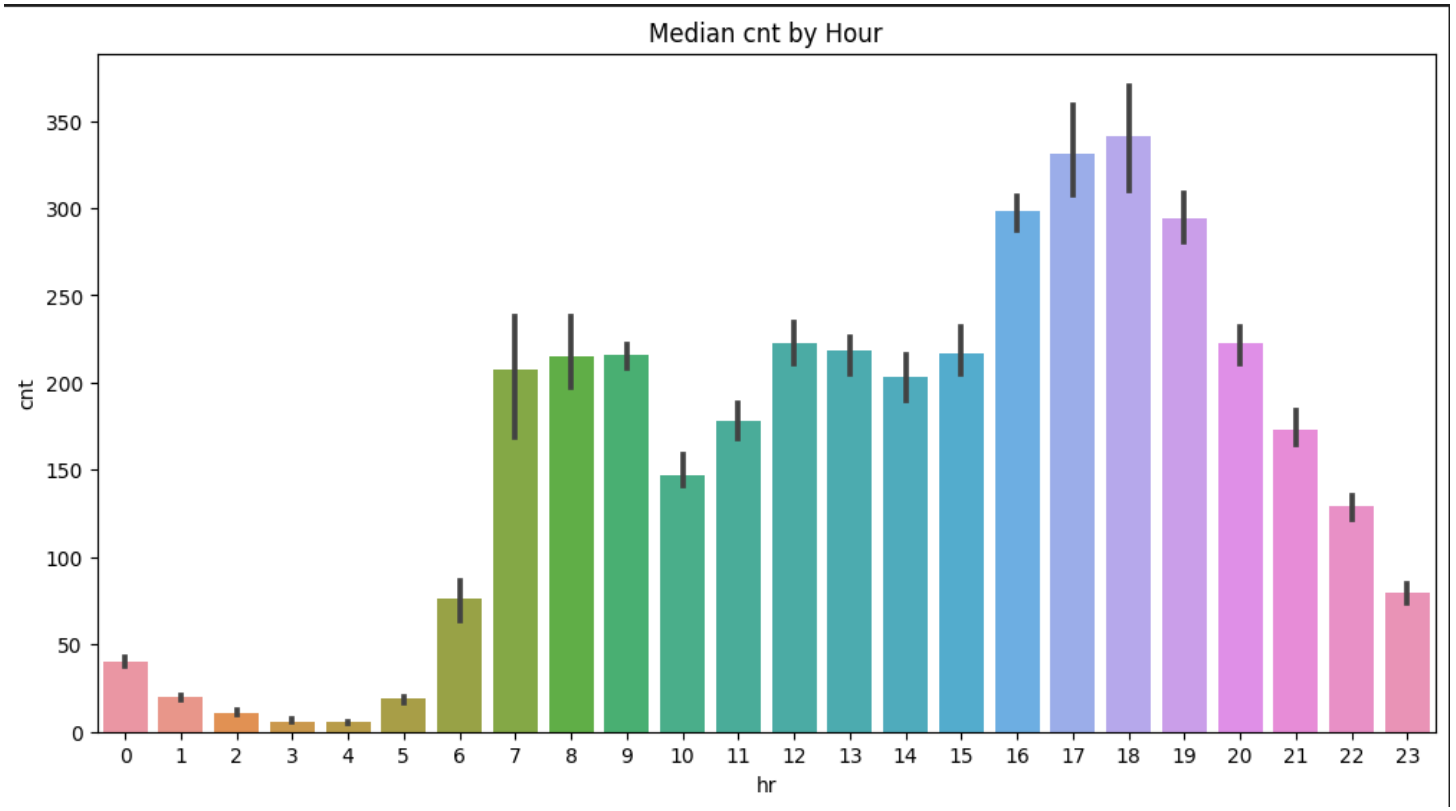
CHAPTER-3
RESULTS

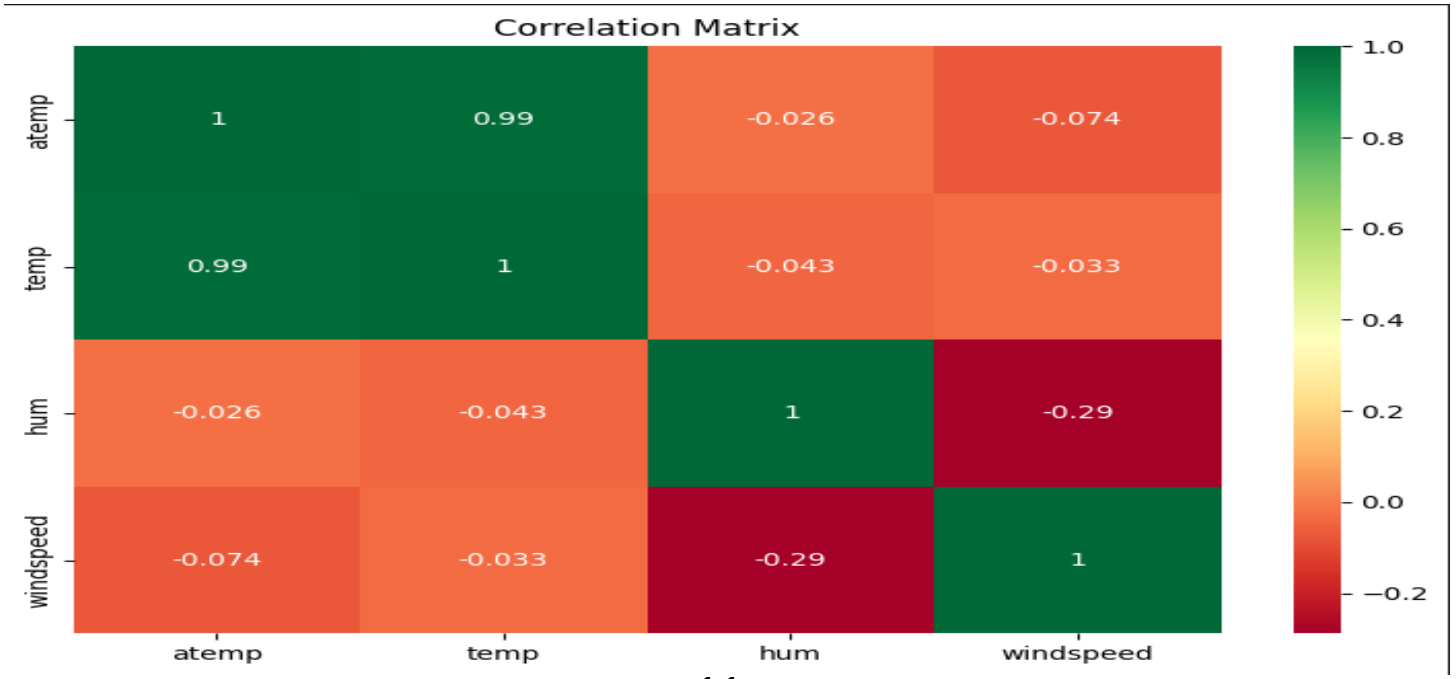




CHAPTER-3

RESULTS





CHAPTER-4
CONCLUSION

In conclusion, this analysis has provided valuable insights into the factors influencing bike-sharing demand and has developed a robust predictive model to forecast the number of bike trips in any given hour. Lyft, Inc., as a leading transportation network company, operates a vast bike-sharing service across numerous cities in the United States and Canada. Accurate prediction of bike demand is critical for Lyft to optimize their operations, including the allocation of bicycles, planning of docking stations, and scheduling maintenance activities.

Through the exploration of the "hour.csv" dataset, which includes variables such as date, season, weather conditions, temperature, humidity, and user counts, we identified significant correlations and patterns in bike usage. Factors such as seasonality, weather conditions, time of day, and day type (weekday vs. weekend) were found to strongly influence bike demand. For instance, demand tends to be higher during warmer seasons, on clear days, and during peak commuting hours.

Several predictive models, including linear regression and machine learning algorithms, were developed and evaluated using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. These models demonstrated good performance in predicting bike demand, with machine learning algorithms generally outperforming traditional regression methods.

The insights gained from this analysis will enable Lyft to make informed decisions regarding resource allocation and operational planning, ultimately improving service availability and customer satisfaction. By accurately forecasting bike demand, Lyft can ensure that there are enough bicycles and docking stations available at the right times and locations, thus enhancing the overall user experience.

In conclusion, this study contributes to the broader understanding of bike-sharing dynamics in urban environments and demonstrates the effectiveness of data-driven approaches in optimizing transportation services. Future research could focus on incorporating real-time data and additional factors, such as events and promotions, to further enhance the accuracy of bike demand forecasting models.