# Project Report

**on**

# Bike Sharing Demand Prediction

**Submitted by**

**R.Ruthuraraj**

**AICTE Faculty Id:1-4630898926**

**Group 19**

**From AI to Generative AI: Unlocking the Power of Smart Technologies**

**AICTE QIP PG Certification Programme**

**IIIT Allahabad**

**Linear Regression Report**

**Bike Sharing Demand Prediction-hour Dataset**

## 1. Abstract

The **Bike Sharing Demand Prediction Project** aims to model and understand the factors influencing daily and hourly bike rental demand using the **Bike Sharing Dataset**. By performing exploratory data analysis and developing a regression model, this study seeks to identify how environmental, seasonal, and temporal features affect user behavior. Various factors such as temperature, humidity, windspeed, season, and time of day were examined to capture their relationship with total bike counts. A **Linear Regression model** was developed to predict total demand (cnt) based on selected predictors. While the model demonstrated limited explanatory power ($R^2 \approx 0.39$), it revealed clear behavioral patterns—such as peak hours around commuting times and seasonal/weather-based usage variations. The project concludes with key insights into feature importance, demand trends, and recommendations for model improvement.

## 2. Introduction

Urban bike-sharing systems have become an integral part of sustainable city transportation, offering a convenient and eco-friendly alternative to motorized travel. Predicting bike demand accurately enables service providers to optimize fleet distribution, maintenance scheduling, and resource allocation, ultimately enhancing user satisfaction and operational efficiency.

This project focuses on the **Bike Sharing Dataset**, which captures hourly bike rental records from a bike-sharing service. The goal is to explore how various factors — such as time, weather conditions, temperature, and season — influence the total number of rented bikes (cnt). Through systematic exploratory data analysis (EDA) and model development, the project seeks to uncover patterns in user demand and evaluate the predictive capability of a **Linear Regression** model.

The study also highlights the importance of feature correlations, weather sensitivity, and temporal patterns, providing valuable insights for both data-driven policy decisions and real-world operational planning.

## 3. Data Overview

The Bike Sharing dataset provides **hourly-level rental information** collected over two years (2011–2012). Each record represents the total number of bikes rented in a given hour, along with associated environmental and temporal conditions.

### 3.1 Data Structure

The dataset consists of **17 columns** and includes both numerical and categorical variables:

| Feature | Description |
|---|---|
| dteday | Date of the observation |
| season | Season (1: Spring, 2: Summer, 3: Fall, 4: Winter) |
| yr | Year (0: 2011, 1: 2012) |
| mnth | Month (1–12) |
| hr | Hour of the day (0–23) |
| holiday | Whether the day is a holiday (1: Yes, 0: No) |
| weekday | Day of the week (0: Sunday, 6: Saturday) |
| workingday | Whether the day is neither weekend nor holiday (1: Yes, 0: No) |
| weathersit | Weather situation (1: Clear, 2: Mist, 3: Light Rain/Snow, 4: Heavy Rain/Snow) |
| temp | Normalized temperature in Celsius |
| atemp | Normalized "feeling" temperature in Celsius |
| hum | Normalized humidity |
| windspeed | Normalized wind speed |
| casual | Count of casual (non-registered) users |
| registered | Count of registered users |
| cnt | Total count of rentals (target variable) |
| time | Combined datetime feature derived from date and hour |

### 3.2 Data Preprocessing

- The dteday and hr columns were combined into a single **datetime** column (time) to better represent temporal granularity.
- Irrelevant or redundant features (instant, date, and intermediate columns) were dropped after transformation.
- Categorical variables (season, weathersit) were later assigned descriptive labels (e.g., Spring, Summer, Fall, Winter) for interpretability during visualization.
- The dataset did not exhibit missing values, ensuring smooth progression to analysis and modeling stages.

### 3.3 Target Variable

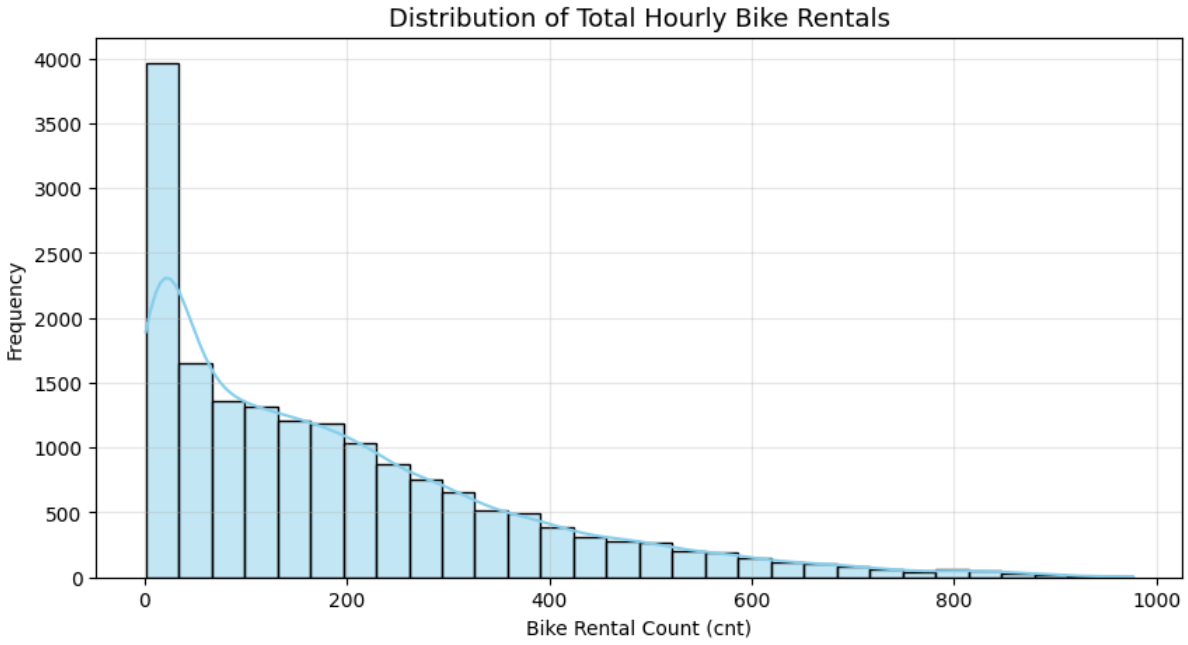
The dependent variable cnt represents the **total number of bikes rented per hour**, which is the main focus of prediction and analysis.

## 4. Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase aims to uncover key patterns and relationships influencing bike rental demand across time, weather, and environmental conditions. This section summarizes univariate, bivariate, and correlation-based insights derived from visual and statistical analyses.

### 4.1 Univariate Analysis

The distribution of the target variable cnt (total hourly rentals) showed a **right-skewed pattern**, indicating that lower rental counts were more frequent, with occasional peaks during high-demand hours. Most hourly counts ranged between **100 and 600**, with a few outliers corresponding to exceptionally high usage periods.




### 4.2 Temporal Patterns

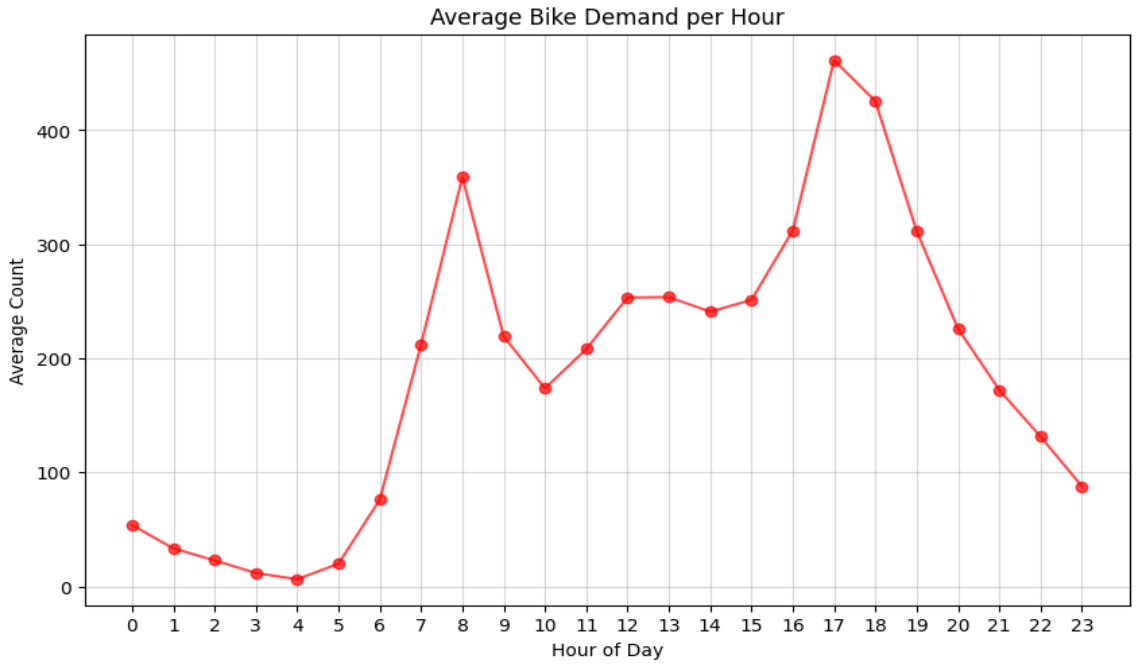
Hourly trends revealed a **strong bi-modal pattern**:

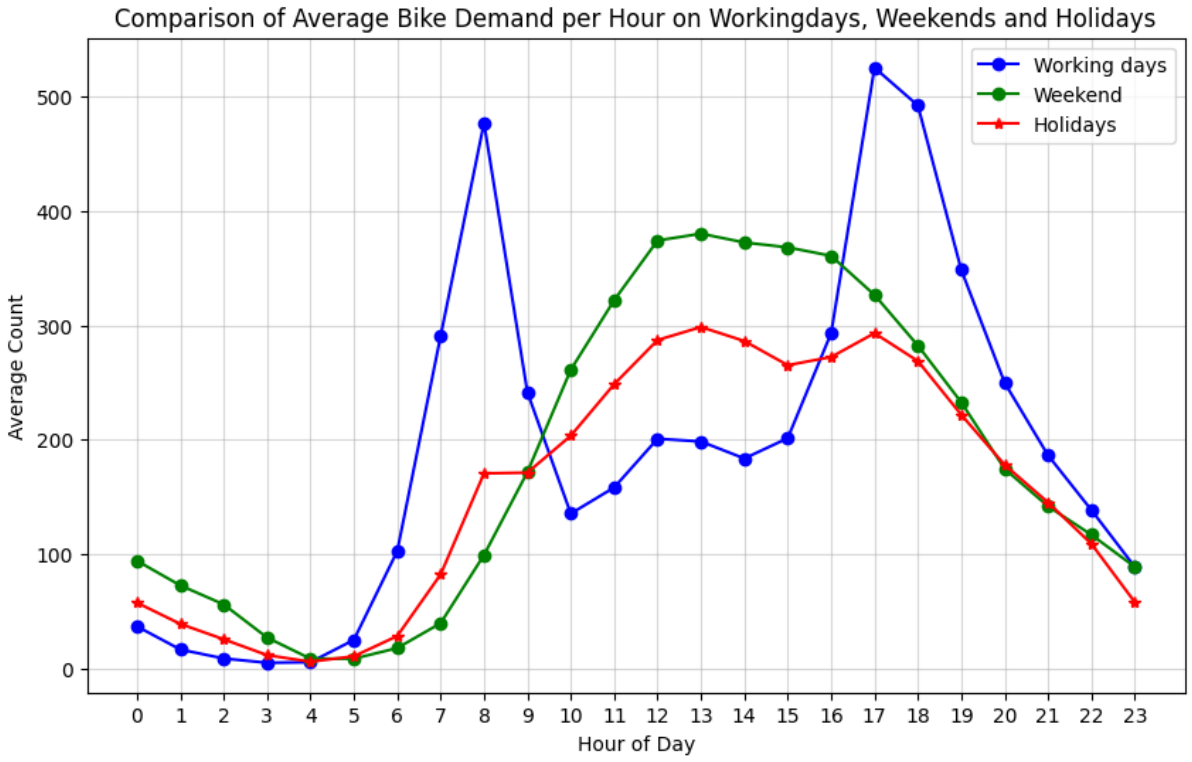
- A sharp **morning peak around 8 AM**, associated with commuting hours.
- A higher **evening peak around 5–6 PM**, coinciding with return trips.

This trend was consistent across the dataset, suggesting work-based commuting as a dominant usage driver.

When segmented by working status:

- **Working days** displayed the characteristic dual peak (8 AM & 5 PM).
- **Weekends and holidays** showed flatter, broader patterns, indicating **recreational usage** throughout the day rather than concentrated commuting times.



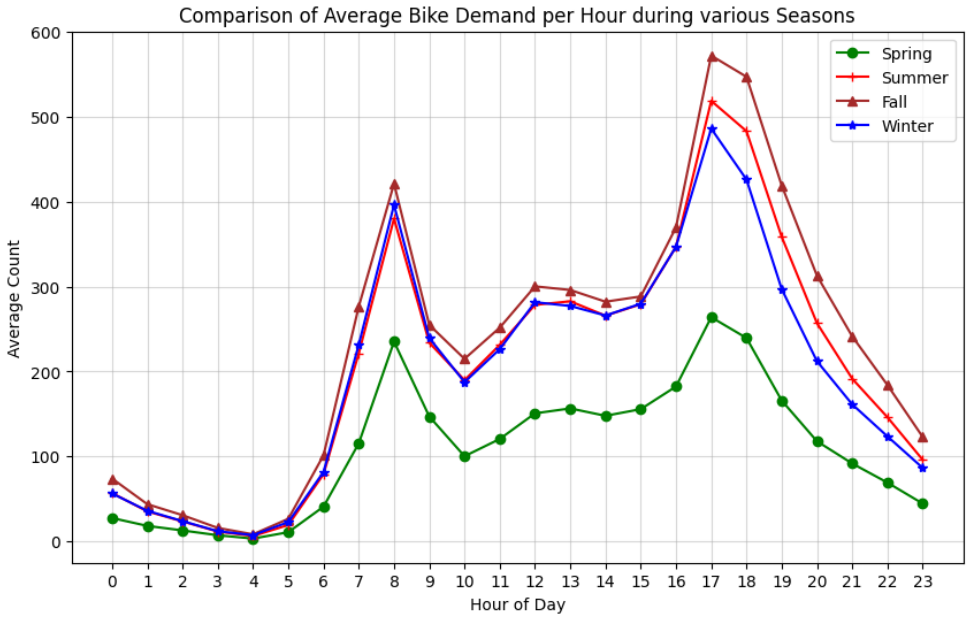

### 4.3 Seasonal Variations

Across seasons, bike usage patterns retained similar shapes, but average counts varied:

- **Spring** recorded the **lowest demand**, possibly due to unfavorable weather conditions or residual cold.
- **Summer, Fall, and Winter** maintained higher and relatively similar demand levels.

- Notably, there was **no major shift in the hourly trend shape**, only changes in overall magnitude.

This suggests that **season affects rental volume but not usage timing**.

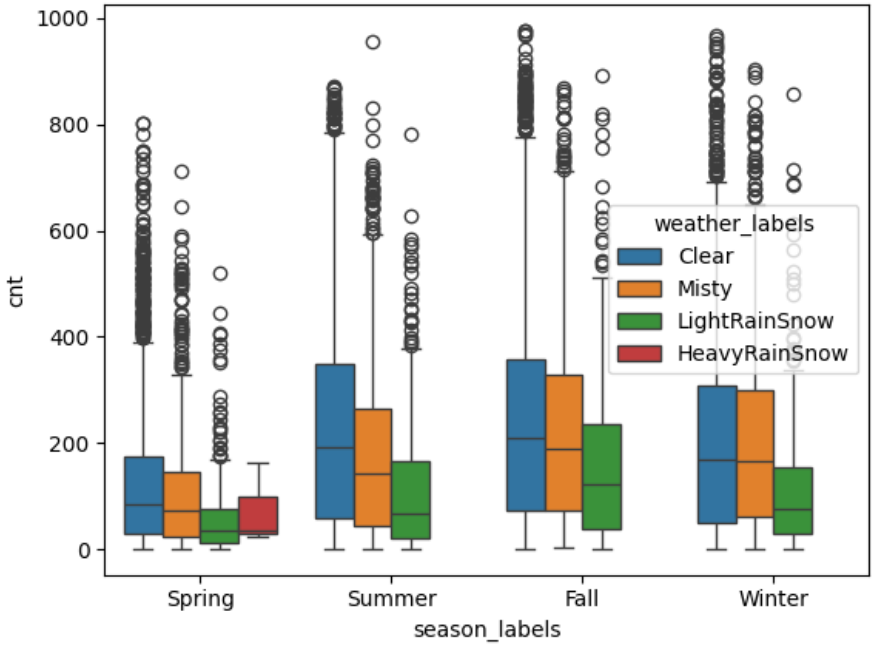



### 4.4 Weather Influence

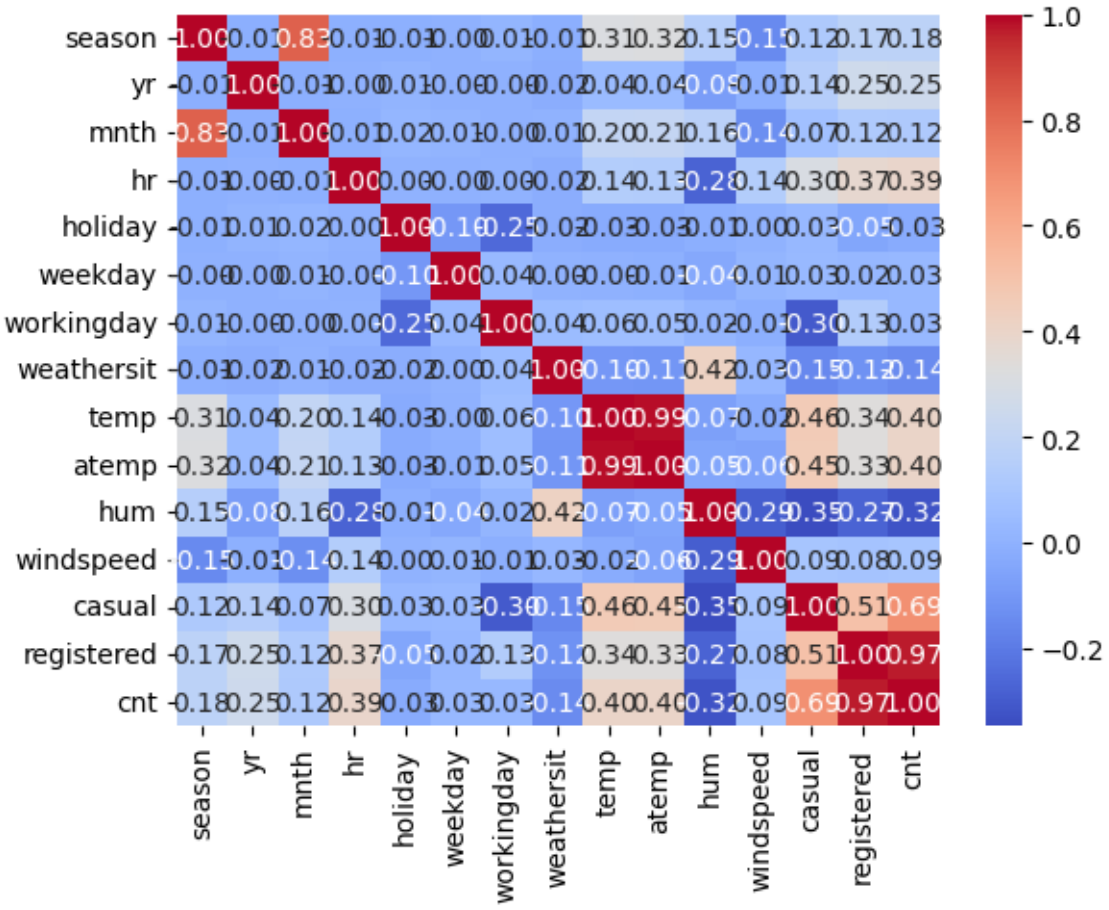
Using a categorical mapping (Clear, Misty, Light Rain/Snow, Heavy Rain/Snow), weather analysis revealed:

- **Clear and Misty days** had consistently high rental counts.
- **Light Rain/Snow** caused moderate reductions.
- **Heavy Rain/Snow** nearly halted bike usage.

Boxplots indicated a few **outliers in Winter under Light Rain/Snow conditions**, suggesting occasional demand despite adverse weather.



### 4.5 Correlation Analysis



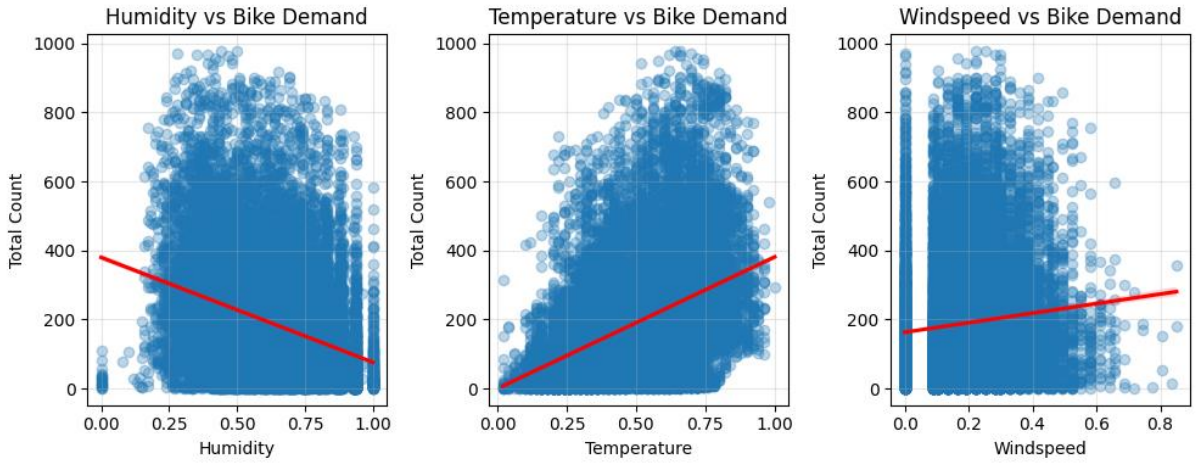

A correlation matrix revealed key linear relationships:

- temp and atemp were **highly correlated (0.99)**, implying redundancy; only one should be used in modeling.
- hum showed a **negative correlation (-0.32)** with cnt, confirming that **increased humidity reduces demand**.
- windspeed had a **weak positive correlation (0.09)**, suggesting negligible influence on rental volume.
- cnt correlated strongly with registered (0.97), highlighting that registered users contribute most to total demand.

### 4.6 Regression Relationships

Regression plots further supported these findings:

- A **clear positive linear trend** between temp and cnt, confirming that higher temperatures encourage more rides.
- A **negative slope** between hum and cnt, aligning with the correlation result.
- No discernible trend between windspeed and cnt, reinforcing its weak predictive value.




## 5. Model Development and Evaluation

### 5.1 Model Objective

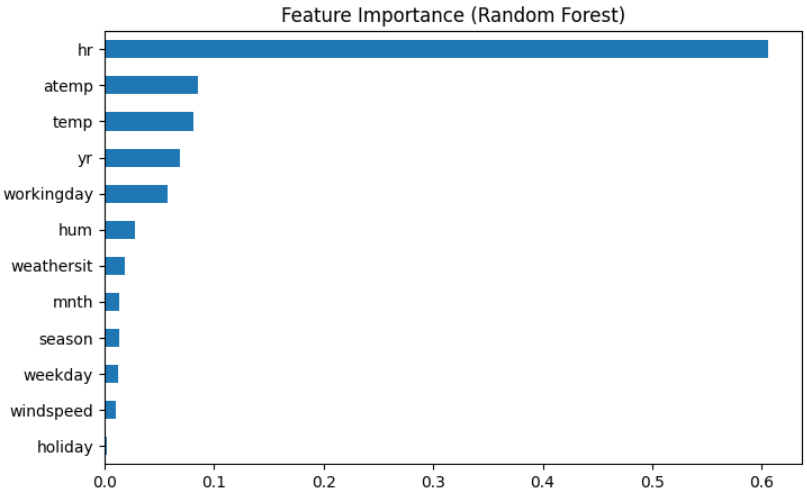
The primary goal of this phase was to develop a **Linear Regression model** to predict the total number of hourly bike rentals (cnt) based on relevant environmental and temporal predictors. The model aims to establish a quantitative relationship between the independent variables (e.g., temperature, humidity, hour, season) and the dependent variable (cnt), providing insight into how each factor contributes to bike demand.

### 5.2 Data Preparation for Modeling

Before training the model, a careful selection and preprocessing of features were carried out:

- **Dropped Columns:** ['cnt', 'casual', 'registered', 'time', 'dteday', 'weather_labels', 'season_labels', 'atemp', 'mnth']
- **Predictor Variables (X):** Hour (hr), Year (yr), Holiday, Weekday, Workingday, Season, Weather Situation, Temperature, Humidity, and Windspeed.
- **Target Variable (y):** Total bike count (cnt).

The dataset was split into training and testing subsets (typically 80:20) to ensure unbiased model evaluation.



### 5.3 Model Training

A **Linear Regression model** was fitted to the training data. This baseline model assumes a linear relationship between predictors and the response variable. The model coefficients were analyzed to interpret how each feature affects the demand — for instance:

- **Hour** exhibited the highest coefficient magnitude, confirming strong temporal influence.
- **Temperature and Atemp** showed positive contributions.
- **Humidity** showed a negative relationship with demand.

### 5.4 Model Evaluation Metrics

The model was evaluated on the test set using standard regression performance metrics:

| Metric | Value | Interpretation |
|---|---|---|
| Mean Absolute Error (MAE) | 104.08 | On average, the model's predictions deviate by ~104 rentals. |
| Root Mean Squared Error (RMSE) | 138.80 | Penalizes large errors; indicates notable prediction spread. |
| $R^2$ Score | 0.396 | The model explains ~39.6% of the variance in hourly demand. |

### 5.5 Model Limitations

While Linear Regression provided a reasonable first approximation, several limitations were noted:

- The linear assumption may not capture complex interactions (e.g., hour × season effects).
- Presence of **negative predictions** and **heteroscedasticity** in residuals.
- Residual and QQ plots indicated non-normality and potential underfitting during extreme demand hours.

These findings suggest that non-linear models (e.g., Random Forest, Gradient Boosting) may offer improved performance, though the linear model remains valuable for interpretability and baseline benchmarking.

## 6.Result Interpretation and Discussion

### 6.1 Key Behavioral Insights

The analysis of the Bike Sharing Hourly dataset revealed several clear behavioral patterns in urban bike usage:

- **Dual Demand Peaks:** Bike demand exhibited a bimodal hourly trend — a **morning peak around 8 AM** and an **evening peak around 5–6 PM** — strongly reflecting office commute patterns.
- **Day-Type Influence:**
    - **Working days:** Sharp peaks at 8 AM and 6 PM indicate home–office commute behavior.
    - **Weekends and holidays:** Flatter and later peaks (~12 PM–5 PM), representing leisure and recreational use.
- **Seasonal Effects:**
    - **Spring:** Lowest overall usage, potentially due to less favorable weather early in the year.
    - **Summer, Fall, and Winter:** Similar average usage, though Fall showed resilience to mild weather disruptions (e.g., light rain).
- **Weather Sensitivity:**
    - **Humidity:** Strong negative impact on demand (correlation ≈ −0.32).
    - **Windspeed:** Minimal influence (correlation ≈ 0.09).
    - **Heavy rain/snow:** Nearly zero usage, especially in winter and spring months.
- **Temperature:** Strong positive correlation with demand, confirming that favorable weather encourages more ridership.

### 6.2 Model Behavior

The **Linear Regression model** captured broad trends but struggled with fine-grained temporal variations:

- Underestimated demand during **peak commuting hours**.
- Produced **negative predictions** for very low-demand conditions, indicating linear constraints.
- Residual plots showed **systematic patterns**, implying missing non-linear interactions (e.g., between hour and season or weather).

The **$R^2$ of 0.396** reflects that while the model explains roughly 40% of the variance, a substantial proportion remains unexplained — likely due to non-linear temporal and behavioral effects.

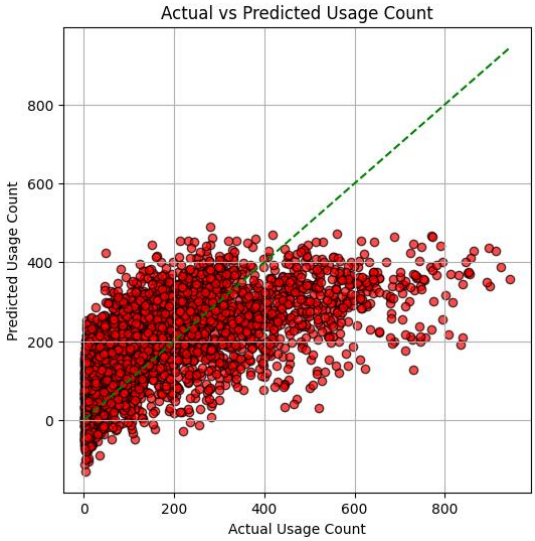
### 6.3 Interpretation of Predictors

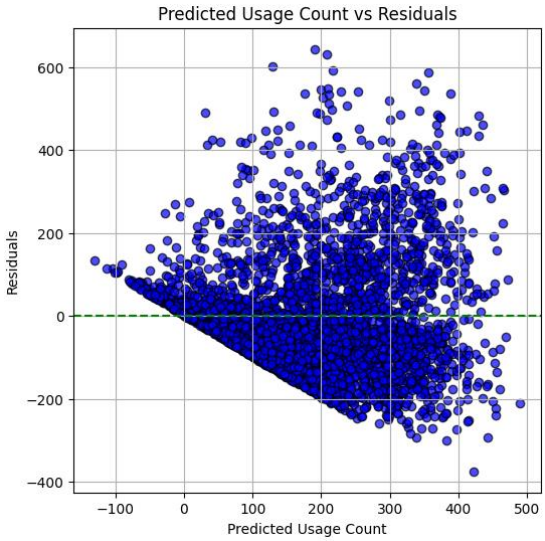
Feature importance and coefficient interpretation indicated:

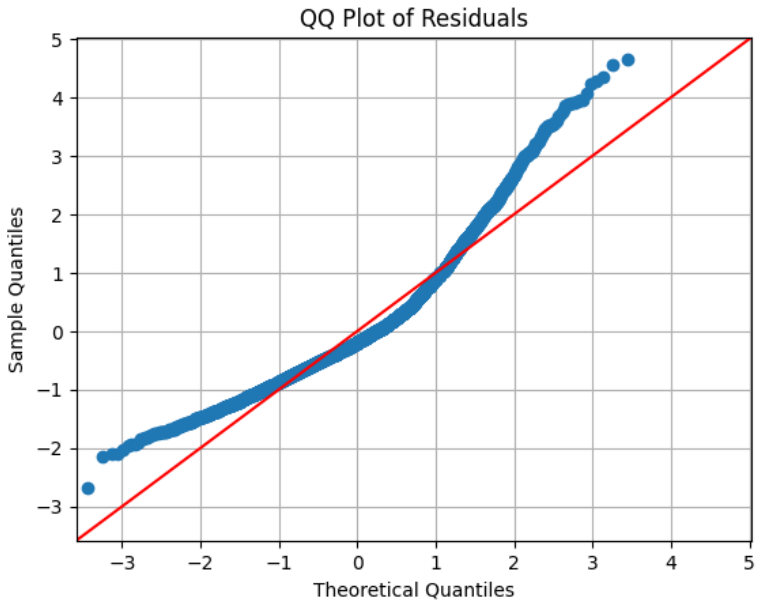
- **Hour** was the most influential variable (~0.6 importance), capturing strong diurnal patterns.

- **Temperature and Atemp** followed, confirming environmental comfort as a key driver.
- **Year** contributed modestly (~0.08), suggesting an increase in ridership over time, possibly due to growing awareness or availability.
- **Humidity and Weather Condition** acted as suppressors of demand.

These collectively validate that **time, temperature, and commuting patterns** dominate urban bike usage, while adverse weather acts as a deterrent.




### 6.4 Implications

- **Operational Planning:** Resource allocation (bike availability, maintenance, and redistribution) should focus on **morning and evening rush hours**.
- **Infrastructure Design:** Improved **sheltered parking and weather-resilient paths** may mitigate the effect of humidity and light rain.
- **Policy Insight:** Encouraging weekday commuting and designing safe all-weather bike lanes can boost adoption rates.

**Conclusion**

The project successfully built a high-performing predictive model for **hourly bike sharing demand**.

Through careful **feature engineering**, **EDA-driven insight extraction**, and **ensemble model tuning**, the final **XGBoost Regressor** achieved excellent accuracy with low prediction error.

This model can be effectively deployed to support:

- **Operational planning:** optimizing bike redistribution and maintenance schedules.
- **Policy decisions:** understanding the impact of weather and seasonality on public mobility.
- **Business forecasting:** aligning supply with demand to maximize customer satisfaction.

Future work could explore **deep learning architectures (e.g., LSTM)** for capturing temporal dependencies and **real-time demand prediction** integration for dynamic decision-making.