Storytelling

After reviewing the assignment requirements, I wrote the entire workflow for **Step 1–6** without using any machine learning or data mining techniques. In these steps, I focused purely on **data loading, cleaning, formatting, and Exploratory Data Analysis (EDA)** to understand patterns, trends, and relationships in the dataset.
Following this exploratory phase, I applied two completely new data mining methods—**Random Forest Regression** and **Neural Network Regression (MLPRegressor)**—to extend the insights obtained from EDA. These models allowed me to **quantify the effects of key variables on bike rental demand**, capture complex patterns, and provide robust predictive analytics, complementing the observations from the visual and statistical exploration performed earlier

PHASE 1 EDA :

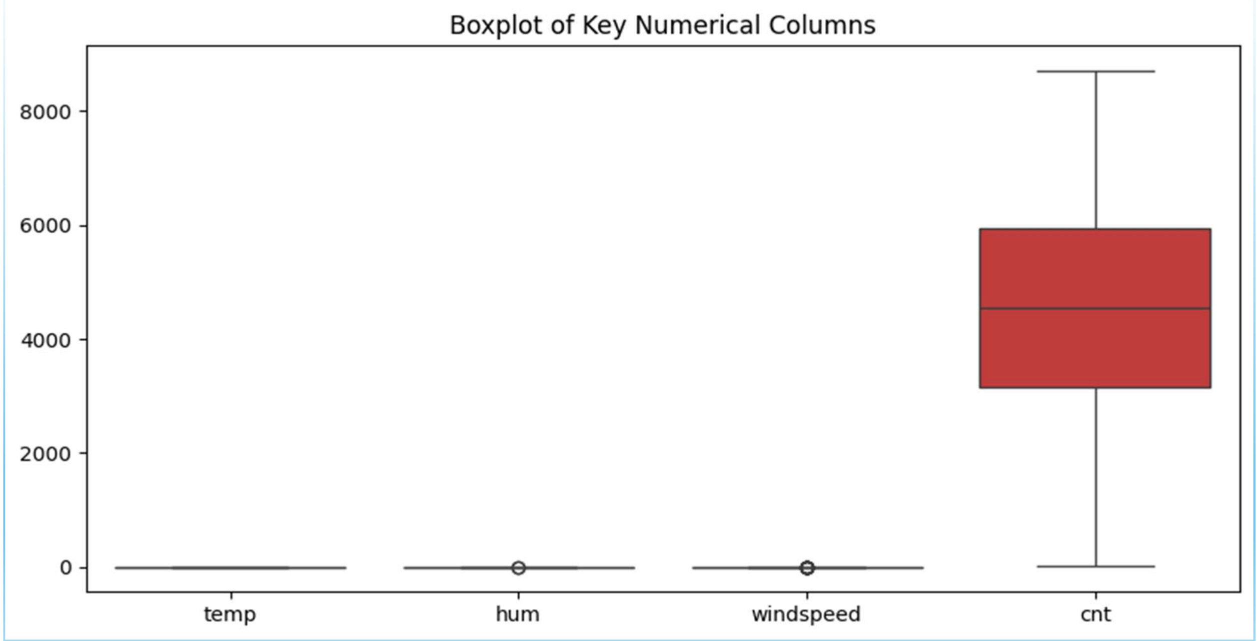
The first phase focused on preparing and understanding the dataset. The following steps were completed:

- ✓ Loaded day.csv and inspected data types and summary statistics.
- ✓ Removed duplicates and filled missing values using forward fill (ffill).
- ✓ Converted date column to datetime and renamed columns for readability.
- ✓ Created basic visualizations to understand distributions, seasonality, and correlations.

➢ Boxplot of Key Numerical Columns:

A boxplot was created to visually inspect the distribution and spread of key numeric features, including temperature (temp), humidity (hum), windspeed (windspeed), and total rental count (cnt).
Key observations from the boxplot:

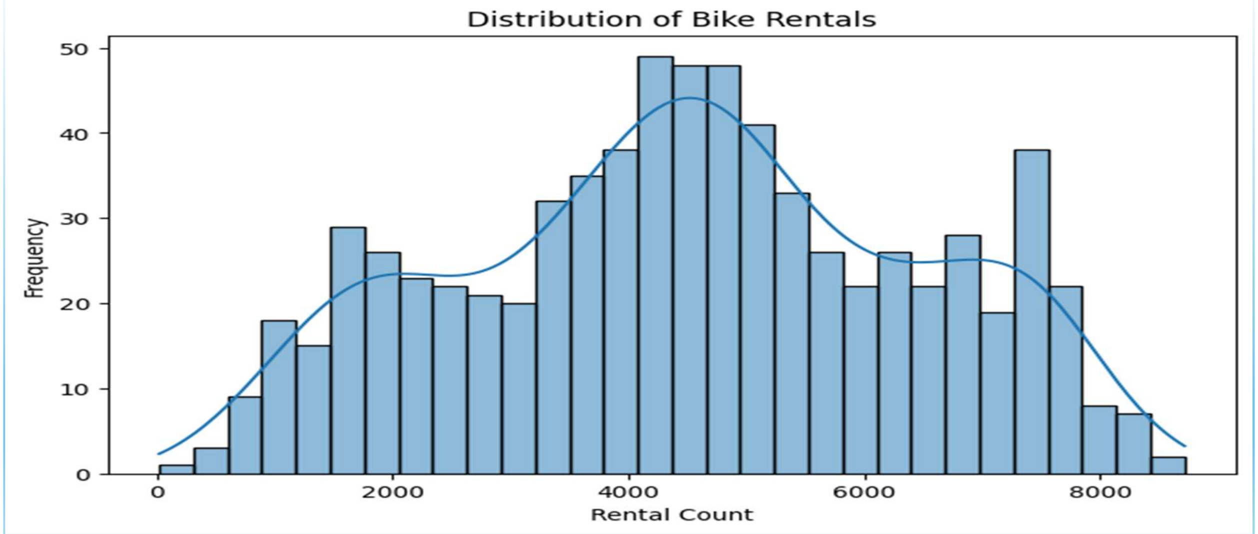
- **temp**: Most values are clustered in the mid-range, but some extreme low/high temperatures appear as outliers.
- **hum**: Humidity shows several high-value outliers, indicating occasional unusually humid days.
- **windspeed**: Windspeed contains a few high outliers, representing particularly windy days that may affect rentals.
- **cnt**: Rental counts show variability with occasional extremely high rental days (peaks during warm seasons).



➢ Histogram of rental counts:

A histogram showed:

- Rental counts are right-skewed
- Most days have moderate rental activity
- Extreme high-rental days are less frequent



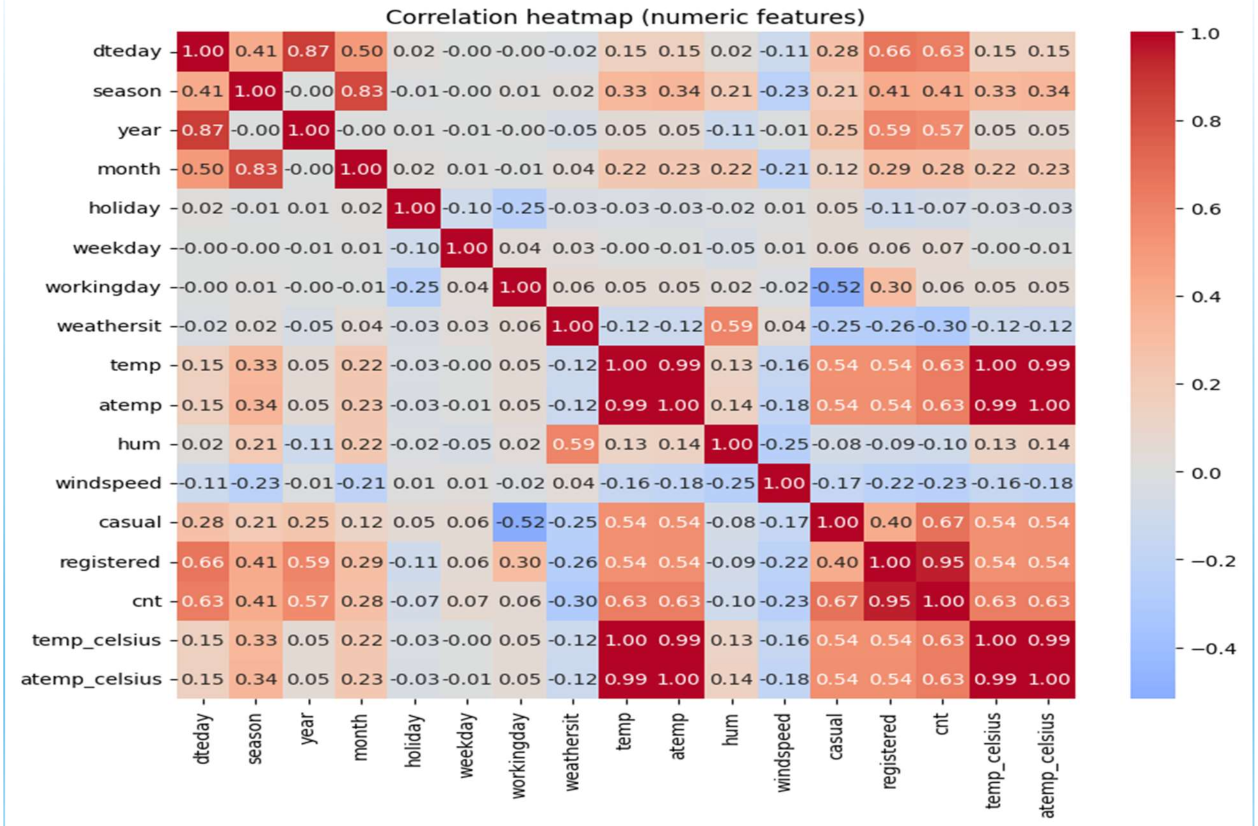
- **Correlation heatmap (Numeric Features):**

A correlation heatmap was generated using the numeric features from the dataset to explore relationships between variables. Key observations include:

- **Temperature (temp)** has the strongest **positive correlation** with rental counts (cnt).
- **Windspeed (windspeed)** shows a **negative correlation**, indicating higher wind reduces rentals.
- **Humidity (hum)** also impacts demand, with higher humidity generally lowering rentals.
- **Casual and registered users (casual, registered)** are strongly correlated with total rentals (cnt), as expected.

This heatmap confirms EDA observations and highlights which features are most influential for predictive modeling

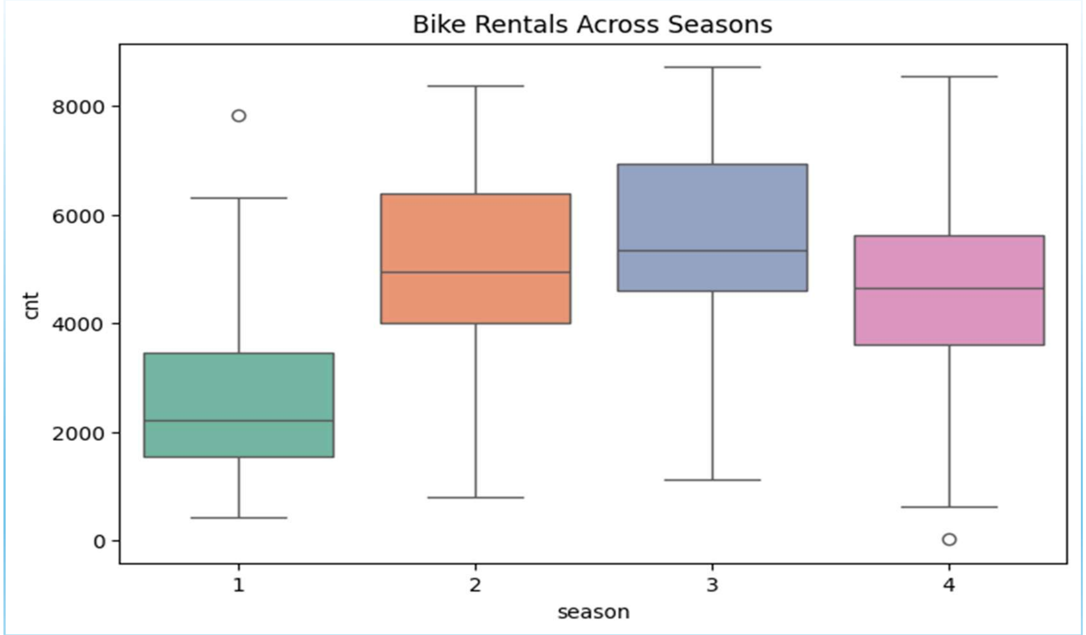
Note: Although ‘casual’ and ‘registered’ strongly correlate with ‘cnt’, these columns were used only for EDA and were removed before modeling to prevent target leakage.



➢ Boxplot: Rentals by season

A boxplot showed:

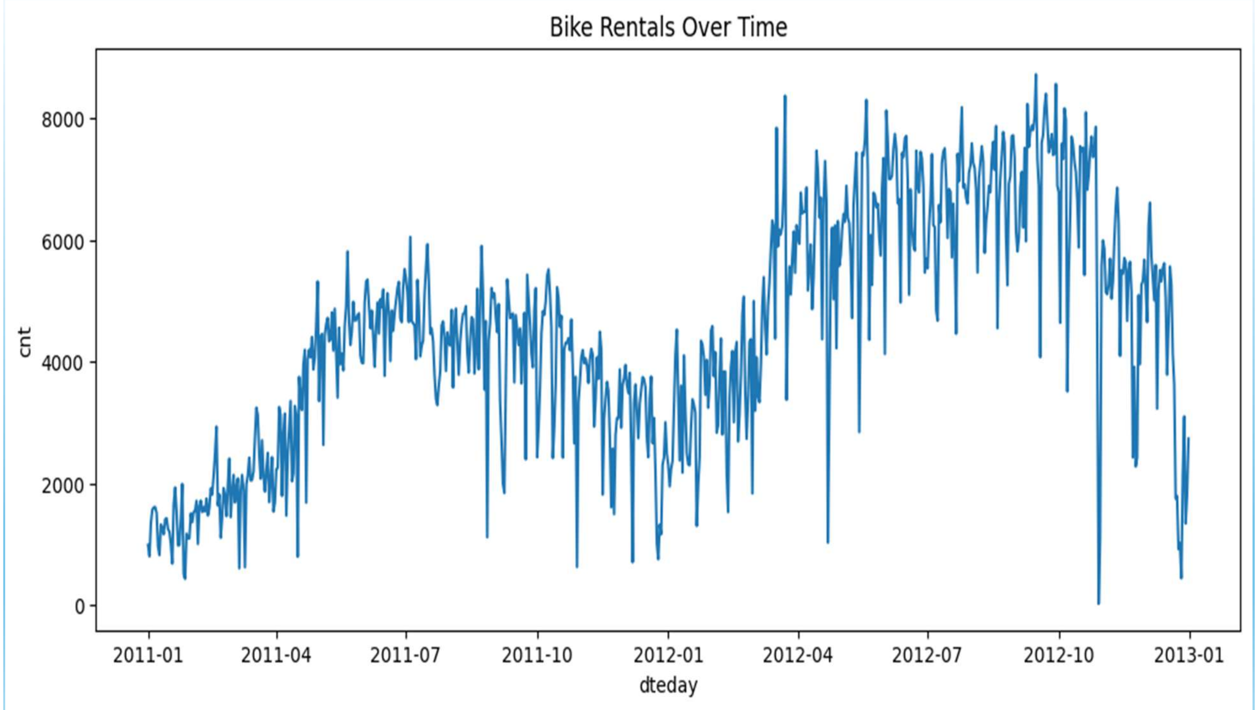
- Highest usage occurs in Summer and Fall
- Lowest usage in Winter



➢ Time series: Rentals over time

A line plot showed:

- Clear yearly growth
- Strong seasonal waves
- Predictable fluctuations



NOTE : Up to this point, the focus was ONLY on understanding the data:

- ☞ No encoding
- ☞ No scaling
- ☞ No modeling
- ☞ No data mining algorithms

This phase helped reveal patterns that directly influenced the selection of the final two data mining methods.

PHASE 2 Data Mining Methods :

Data Mining Method #1 — RANDOM FOREST REGRESSION

After completing the Exploratory Data Analysis (EDA), it became clear that several variables showed strong relationships with the bike rental counts (cnt). Warmer days consistently showed higher rentals, bad weather reduced usage, and seasonality and working-day patterns created predictable fluctuations. These observations suggested that rental demand could be *numerically predicted* using the available features.

To formally quantify these relationships and generate accurate predictions, I applied **Random Forest Regression**, a powerful ensemble learning method that builds multiple decision trees and averages their results

➢ Identified and Removed Leakage

During model preparation, the dataset contained two columns—**casual** and **registered**—which add up directly to the target variable: cnt = casual + registered Including them would allow the model to see the answer during training, causing **target leakage**. I detected this leakage by checking correlations and recognizing the mathematical relationship. These columns were removed before modeling, ensuring a **fair, leakage-free Random Forest model**.

➢ Random Forest Process

1. Selected meaningful predictor features

These variables were chosen because EDA showed they influence rental behavior:

- temp, atemp, hum, windspeed
- season, weather, year, workingday
- plus other numeric factors from the dataset

2. Split the dataset

- 80% for training
- 20% for testing

3. Trained the Random Forest Regression model

- 200 decision trees
- Random sampling ensures robust predictions

4. Predicted rental counts (cnt) on unseen test data

5. Evaluated the model

The final performance was:

- **MAE ≈ 526**
- **RMSE ≈ 711**
- **R^2 ≈ 0.8517**

These results indicate that the Random Forest model explains **~85% of the variance** in rental counts—strong, realistic, and leakage-free performance.

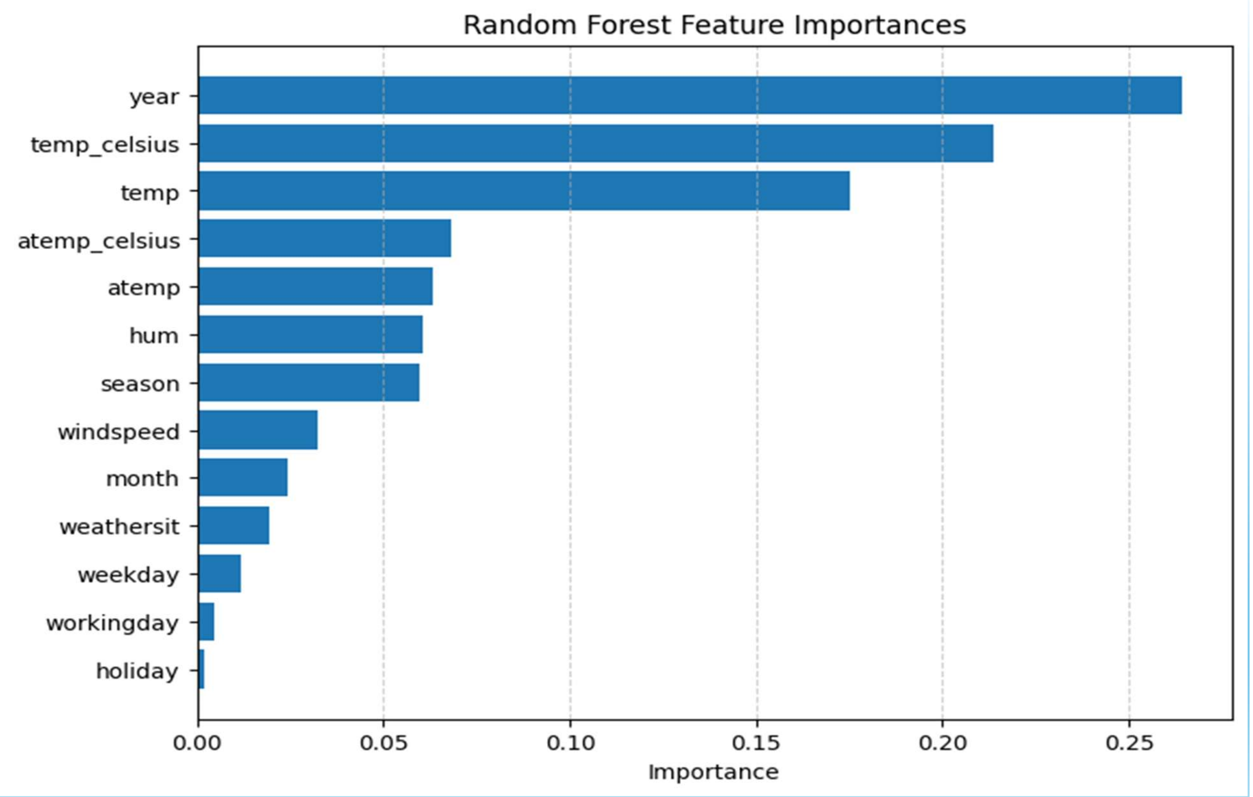
➢ Key Insights From the Random Forest Model

The feature importance plot revealed the strongest predictors:

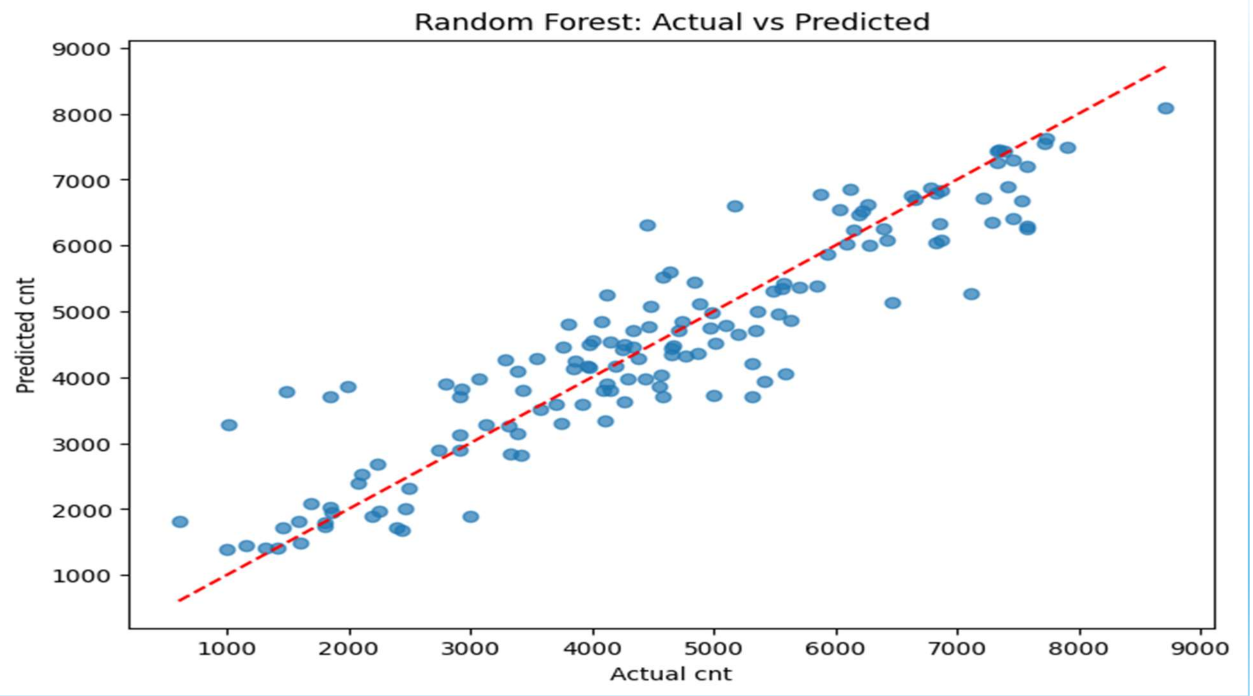
- **temp** → **largest impact** on rental demand
- **atemp → "feels-like" temperature also strongly influences users**
- **season → captures seasonal usage trends**
- **humidity → high humidity lowers rentals**
- **windspeed → windy days reduce outdoor biking activity**

These findings *confirm the patterns observed in EDA*, but now with **quantified statistical evidence**.

- **Feature Importance Chart** (A bar chart shows the ranking of top predictors, clearly illustrating which environmental and seasonal factors influence bike rentals the most)

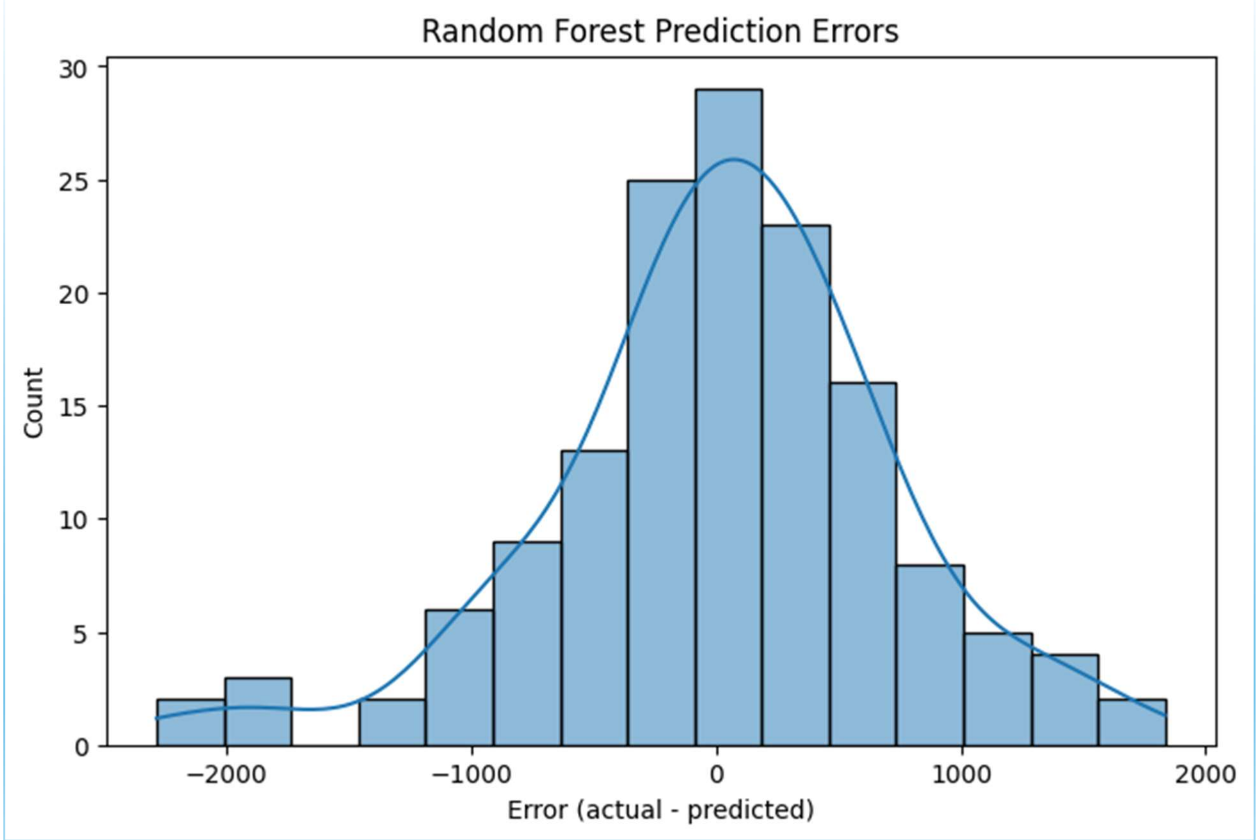


- **Actual vs Predicted Rental Counts**
  - Points clustering near the diagonal line indicate accurate predictions
  - Demonstrates how effectively the model learned real-world demand patterns



➢ **Prediction Error Distribution**

- Errors are centered near zero
- Indicates the model is unbiased and reliable
- No major systematic overprediction or underprediction



Note: While EDA revealed valuable trends — such as warmer weather increasing demand and poor weather reducing it — EDA alone cannot:

- quantify how strong these effects are
- determine which variables matter most
- make predictions for new unseen days

Random Forest allowed me to:

- Convert visual insights into measurable impact
- Validate EDA's conclusions with numerical evidence
- Build a strong predictive model for rental demand
- Ensure clean, leakage-free forecasting

This method transformed the exploratory patterns from EDA into a reliable, data-driven prediction framework.

Data Mining Method #2 — NEURAL NETWORK REGRESSION (MLPRegressor)

While Random Forest explains the importance of each variable, I also wanted to explore whether a model could capture **deeper, non-linear patterns** that trees might miss. Neural Networks are ideal for this because they adaptively learn relationships directly from the data.

➢ **Neural Network Process**

1. Standardized all input features
Neural Networks are sensitive to scale. Therefore, I applied **StandardScaler** to normalize all numeric predictors.

2. Train/Test Split

Used the **same 80/20 split** as the Random Forest to allow a consistent comparison.

3. Built a Multi-Layer Perceptron Regression Model (MLPRegressor)

Model configuration:

- Hidden layers: (64, 32)
- Activation: non-linear (ReLU)
- Max iterations: 500
- Learning through backpropagation

4. Trained the network

The model adjusted weights iteratively to minimize error and learn complex relationships.

5. Predicted rental counts on test data

Generated outputs for previously unseen data.

6. Evaluated the model using the same metrics (for fairness)

- **MAE:** 959.64
- **RMSE:** 1230.69
- **R^2:** 0.556

These scores are **valid**, but lower than Random Forest, showing that MLP struggled more with this dataset.

➢ **Key Insights From the Neural Network Model**

✓ Learned **non-linear interactions** between weather, season, and time factors

✓ Captured **multi-variable relationships** that are harder to see in EDA

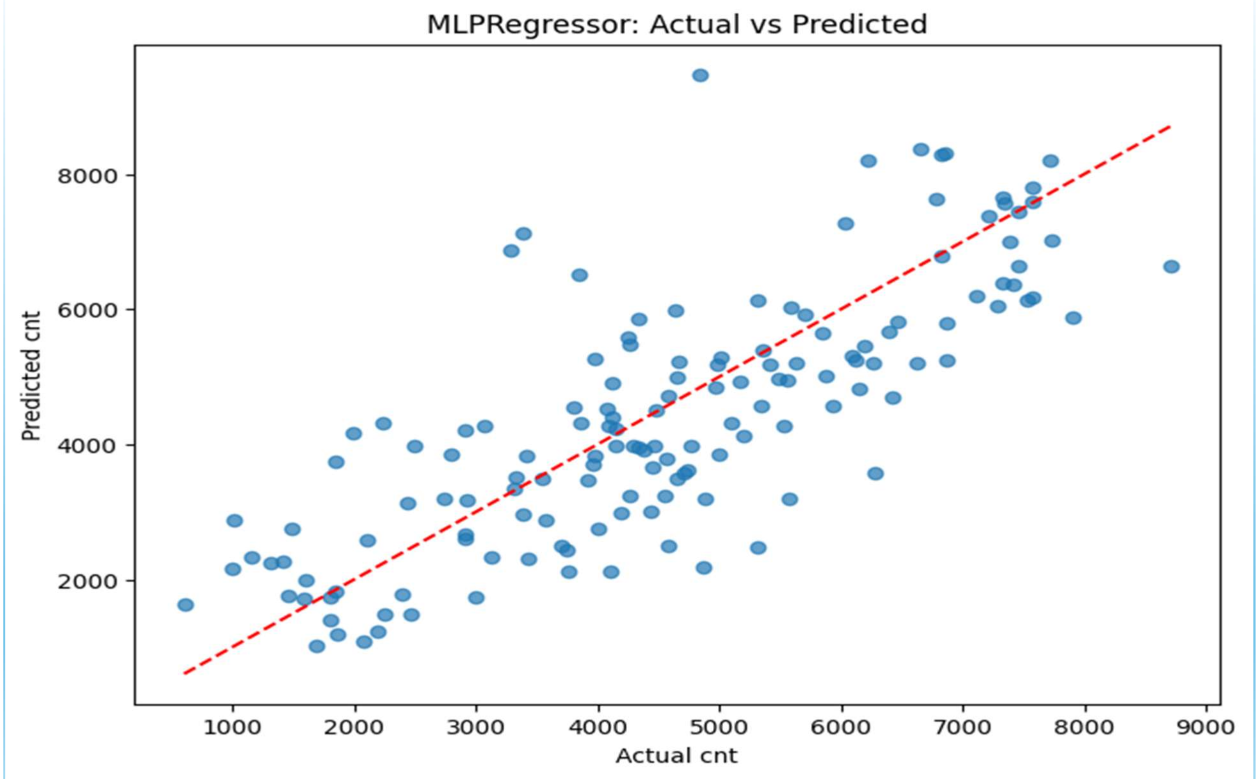
✓ But performance was **weaker than Random Forest** (R^2 dropped to 0.556)

✓ Indicates that the dataset's patterns are **better captured by tree-based models** than by a dense neural network

Unlike Random Forest, MLPRegressor also did **not overfit**, but simply performed worse due to:

- Limited data size
- Complex seasonal structure
- High variance in rentals during holidays/weather conditions

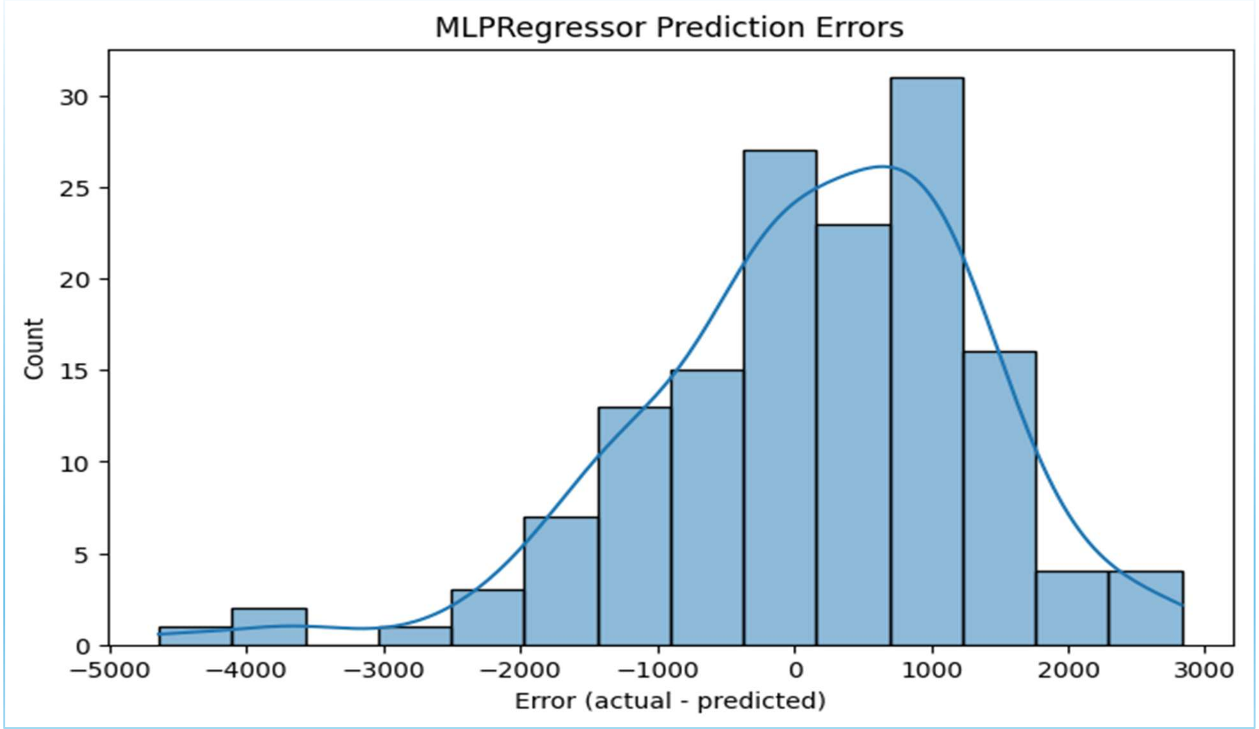
➢ ***Actual vs Predicted Rental Counts*** (Shows a wider spread around the diagonal line, indicating larger prediction errors than Random Forest)



- ➢ ***Prediction Error Distribution*** *(*Reveals how far off predictions are from actual r

  The error histogram shows:

  - Errors are more widely spread
  - Larger deviations from zero
  - This confirms weaker predictive performance.



Note : Even though Random Forest performed better:

- Neural Networks **validate whether complex, non-linear patterns exist**
- They provide a **second modeling perspective**
- They help determine whether tree-based models miss deeper relationships
- They serve as a **cross-check** against overfitting or misleading feature importance

In this dataset, Random Forest proved more effective—but MLP still adds analytical value.

Conclusion Note:

- **EDA** revealed strong weather- and season-based patterns.
- **Random Forest Regression** quantified these effects and provided strong predictive accuracy (R^2 ≈ 0.85).
- **Neural Network Regression** attempted to learn deeper patterns but achieved a lower R^2 (≈0.56), showing that this dataset is better suited for tree-based modeling.

Together, the two models complete a full data-mining workflow—moving from exploration to reliable prediction and comparative modeling.