Bike Rental Demand Analysis: Impact of Weather and Weekends – Cem Özbaş 34453

Project Overview

This project investigates how **weather conditions** and **day type** (weekend vs weekday) influence the **bike rental demand** in New York City.
By integrating trip records from Citi Bike and historical weather data from Open-Meteo, we perform an exploratory and statistical analysis to identify how external factors affect the number of daily rides and train predictive models to forecast ride volume trends.

This study aims to: - Measure the impact of rainy weather on bike usage - Analyze behavioral patterns between weekdays and weekends - Quantify statistical significance of observed trends - Predict March 2025 trip demand using ML models trained on January–February data and make comparison with real world Match 2025 data

Datasets

1. Citi Bike System Data

The primary dataset comes from **Citi Bike**, New York City's bike-sharing service.
It includes detailed ride-level information:

- **Start Date & Time**: When the trip began
- **End Date & Time**: When the trip ended
- **Start & End Stations**: Location IDs
- **User Type**: Subscriber vs casual user
- **Trip Duration**

📌 Source: Citi Bike System Data

---

2. Open-Meteo Historical Weather API

Historical weather conditions were retrieved from **Open-Meteo**'s archive API.
The daily weather information contains:

- **Date**
- **Max and Min Temperatures** (°C)
- **Precipitation Sum** (mm)
- **Rainy Day Label** (>1mm rainfall considered "rainy")

📌 Source: Open-Meteo Historical Weather API

Data Collection Period

- **Training Period:** January 1 – February 28, 2025 (60 days)
- **Test Period:** March 1 – 31, 2025 (31 days)
- **Location:** New York City, USA

Research Questions

- Do rainy days significantly reduce daily bike ride numbers?
- Is there a meaningful difference between weekday and weekend bike usage?
- How do daily trip counts fluctuate over time during the winter months?
- Can we predict trip volume using weather and calender features?

Hypotheses

1. Rain Impact Hypothesis

- **Null Hypothesis (H_0):** Rainy days have no significant effect on bike rentals.
- **Alternative Hypothesis (H_1):** Rainy days significantly decrease bike rentals.

2. Weekend Impact Hypothesis

- **Null Hypothesis (H_0):** Weekends have no significant effect on bike rental numbers.
- **Alternative Hypothesis (H_1):** Weekends significantly influence bike rental numbers compared to weekdays.

Methods

1. **Data Cleaning and Merging**
   - Parsed timestamps and extracted dates
   - Merged bike rental data with daily weather conditions
   - Labeled days as rainy or dry, and weekend or weekday
2. **Exploratory Data Analysis (EDA)**
   - Plotted total daily bike trips
   - Highlighted rainy days in time series
   - Compared trip distributions across rain and weekend categories using boxplots
3. **Statistical Hypothesis Testing**
   - **Independent two-sample t-tests** to compare:
     - Rainy days vs dry days
     - Weekends vs weekdays
   - Calculated p-values and interpreted significance at $\alpha = 0.05$
4. **Machine Learning Prediction**
   - Trained two models using Jan-Feb data:

- **Random Forest Regressor**
- **Linear Regression**

– Input features: precipitation, temperature, weekday, weekend flag, temperature range
– Target: daily trip count

📈 Visualizations

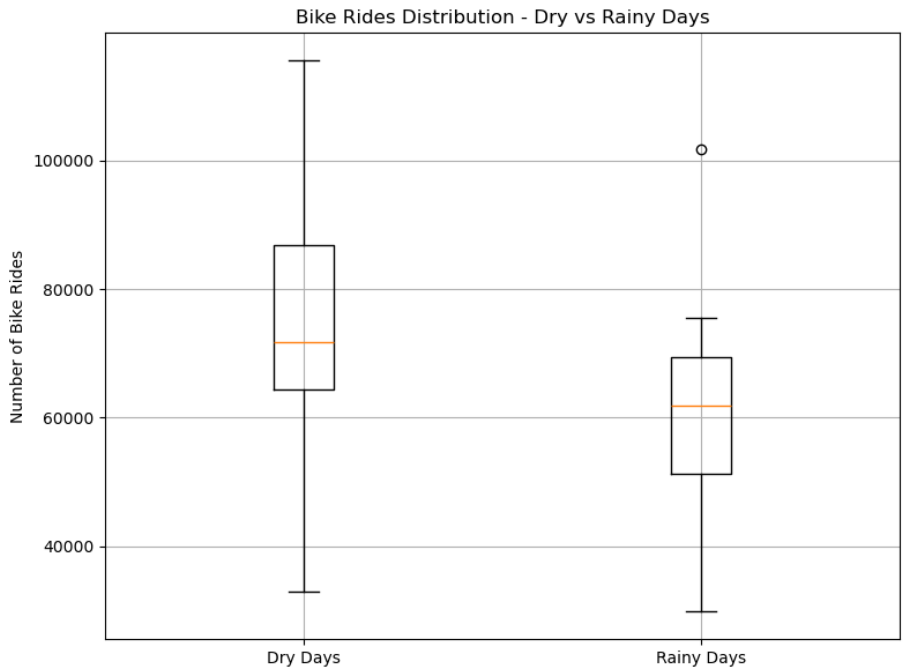
- **Daily Trip Counts (Jan-Feb)** (rainy days highlighted in red)
- **Boxplot:** Dry Days vs Rainy Days
- **Boxplot:** Weekdays vs Weekends
- **ML Prediction Chart:** Actual vs predicted for March

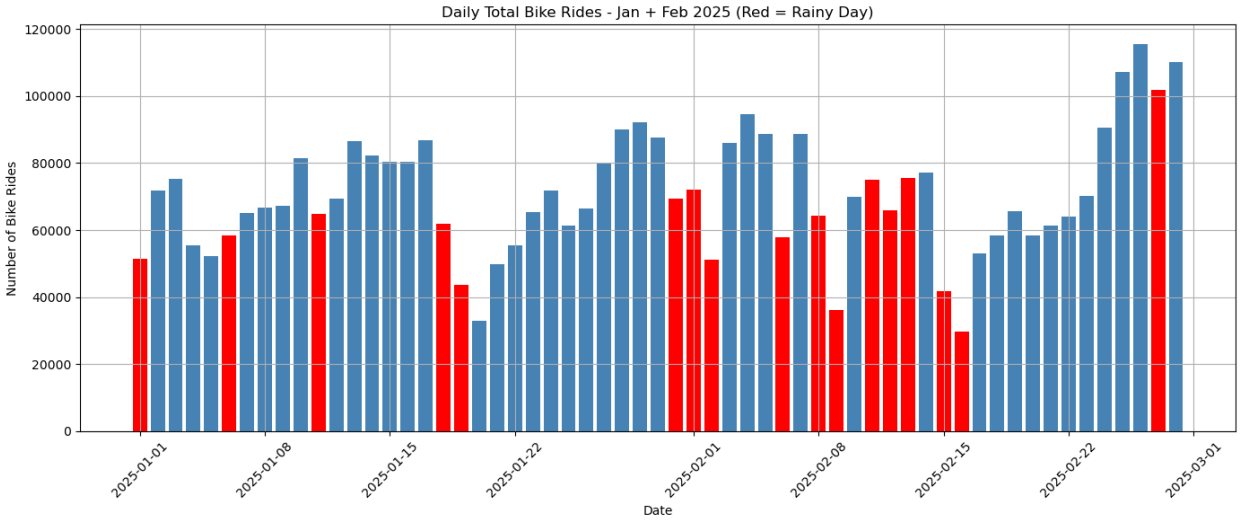
Results Summary

1. Rain Effect

- **T-statistic**: -2.95
- **P-value**: 0.0062
- **Avg rider on dry days:** ~74,634
- **Avg rider on rainy days:** ~60,052
- **Estimated drop due to rain:** ~19.5%

Interpretation:
Since p < 0.05, the difference is **statistically significant** Rainy days **reduce bike rentals by approximately 19.5%**

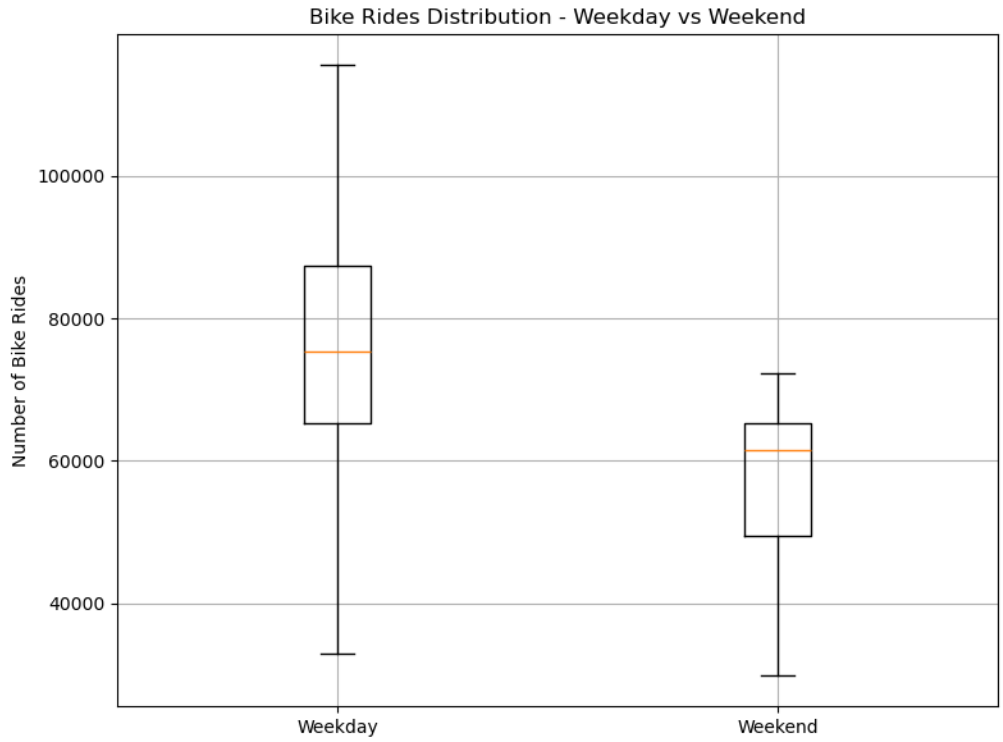




2. Weekend Effect

- **T-statistic**: -4.57
- **P-value**: 0.0001

Interpretation:
Since p < 0.05, the difference is **statistically significant** Bike rental patterns differ between weekdays and weekends — with **weekdays showing higher usage**, which is likely related that in winter months there are less leisure rides on weekends compared to the number of commuters in weekdays.



March Demand Prediction Using Machine Learning

To test the predictive power of machine learning, models were trained using **January + February 2025** data and then tested on **March 2025** data.

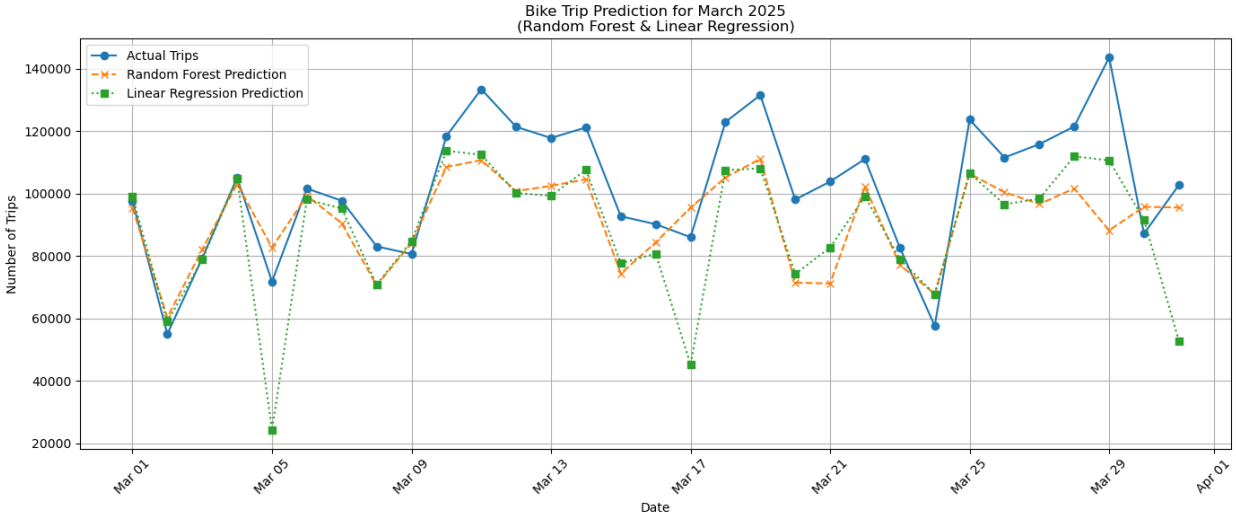
Two models were applied: - **Random Forest Regressor** - **Linear Regression**

Model Edvaluation Metrics

- **Random Forect Regression**
  - R^2 Score: 0.3243
  - MAE (Mean Absolute Error): 13,820.8 trips
- **Linear Regression**
  - R^2 Score: 0.1133
  - MAE (Mean Absolute Error): 15,368.8 trips

Interpretation - While the **R^2 values** may appear low, the models — especially **Random Forest** — still produced **reasonably accurate** predictions in absolute terms. - An average daily prediction error of **~13,800** trips suggests that the model captured meaningful patterns in the data, particularly on **non-extreme weather days**. - The **visual alignment** of predicted vs actual values indicates **trend-following behavior**, even if the models could not explain all the variance.

Visualization - A comparison chart of actual vs predicted trips for March is provided below:



Limitations and Future Work

Limitations: - Only 3 winter months analyzed (not generalizable to warmer seasons) - No distinction between user types (commuters vs tourists) - ML models trained on small dataset, prone to overfitting

Future Work: - Extend study to a full calendar year for seasonal trends. - Analyze user types separately (commuters vs tourists). - Explore additional ML models