Analysis of Factors Influencing Bike Rentals in a Shared Bicycle System
INFO 6105 Final Project Report

Introduction

Bike-sharing systems are transforming urban mobility, offering sustainable and convenient alternatives for commuting and leisure activities. Capital Bikeshare, a prominent bike-sharing service, provides an extensive dataset capturing daily and hourly bike rentals alongside weather, temporal, and seasonal factors. Analyzing this dataset can help uncover critical insights into rental patterns, optimize system operations, and improve user experience.

This analysis focuses exclusively on the daily dataset (day.csv), providing aggregated daily rental counts. By studying the impact of weather conditions, seasonality, and temporal trends, this analysis aims to answer key questions such as:

- How do weather conditions (e.g., temperature, humidity, wind speed) influence daily bike rental patterns during different seasons?
- Are there significant differences in mean bike rental numbers between weekdays and weekends across seasons?

To achieve this, the study employs descriptive statistics, visualization techniques, and machine learning models to extract actionable insights and evaluate predictive performance.

Dataset Overview

The day.csv dataset will be the sole focus of this analysis. This file contains daily aggregated data spanning from 2011 to 2012 and includes 731 records with 16 attributes. These attributes capture a mix of temporal, weather-related, and behavioral variables, as well as the overall rental counts.

Temporal Attributes:

- dteday: Date of the record (YYYY-MM-DD).
- season: Categorical attribute (1 = Winter, 2 = Spring, 3 = Summer, 4 = Fall).
- yr: Binary indicator for the year (0 = 2011, 1 = 2012).
- mnth: Month of the year (1 = January to 12 = December).
- weekday: Day of the week (0 = Sunday to 6 = Saturday).
- holiday: Binary indicator (1 = Holiday, 0 = Non-Holiday).
- workingday: Binary indicator (1 = Working Day, 0 = Weekend or Holiday).

Weather Attributes:

- weathersit: Categorical variable representing weather conditions:
    - 1 = Clear, Few Clouds.
    - 2 = Misty, Cloudy.
    - 3 = Light Snow, Light Rain.
    - 4 = Heavy Rain, Thunderstorms.

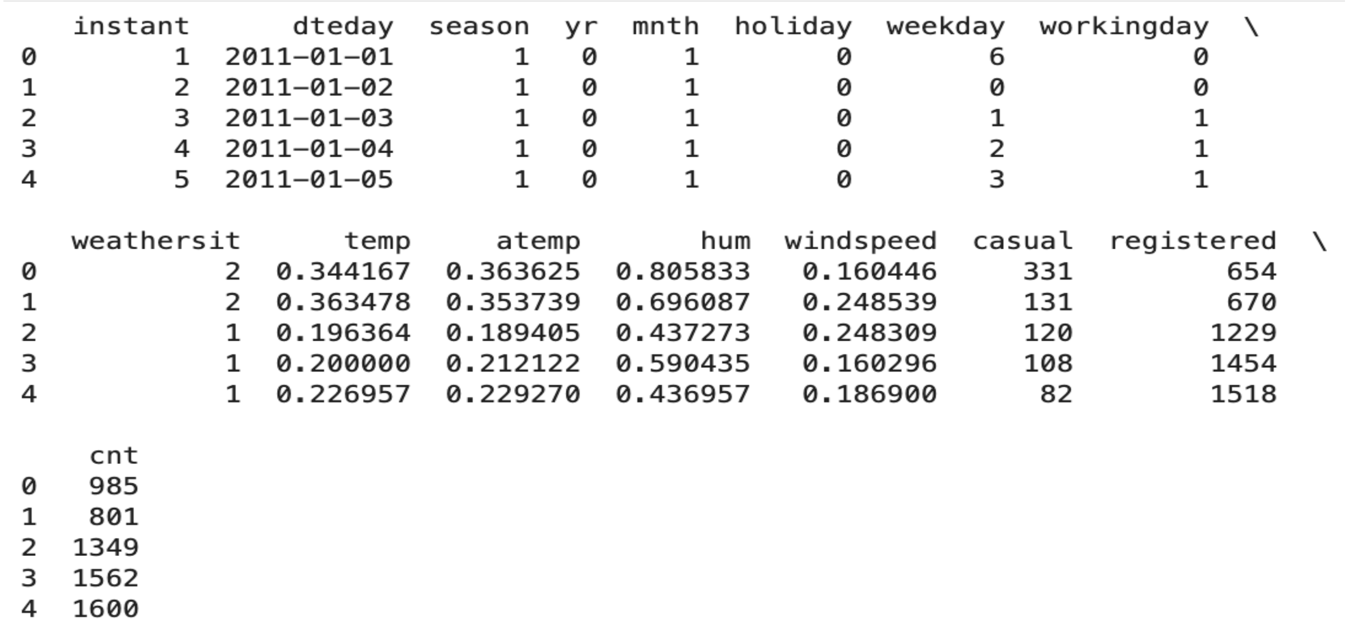
- temp: Normalized temperature in Celsius (scaled to [0, 1]).
- atemp: Normalized "feels like" temperature (scaled to [0, 1]).
- hum: Normalized humidity (scaled to [0, 1]).
- windspeed: Normalized wind speed (scaled to [0, 1]).

Target Variables:

- cnt: Total number of daily bike rentals (sum of casual and registered).
- casual: Rentals by casual users (non-registered).
- registered: Rentals by registered users.

Rationale for Using the Daily Dataset

The day.csv file provides a higher level of aggregation, making it ideal for understanding broad patterns, such as the influence of seasonal changes and weather conditions. While the hourly dataset (hour.csv) offers finer granularity, the daily dataset aligns better with the goals of this study, focusing on long-term trends and seasonal effects.



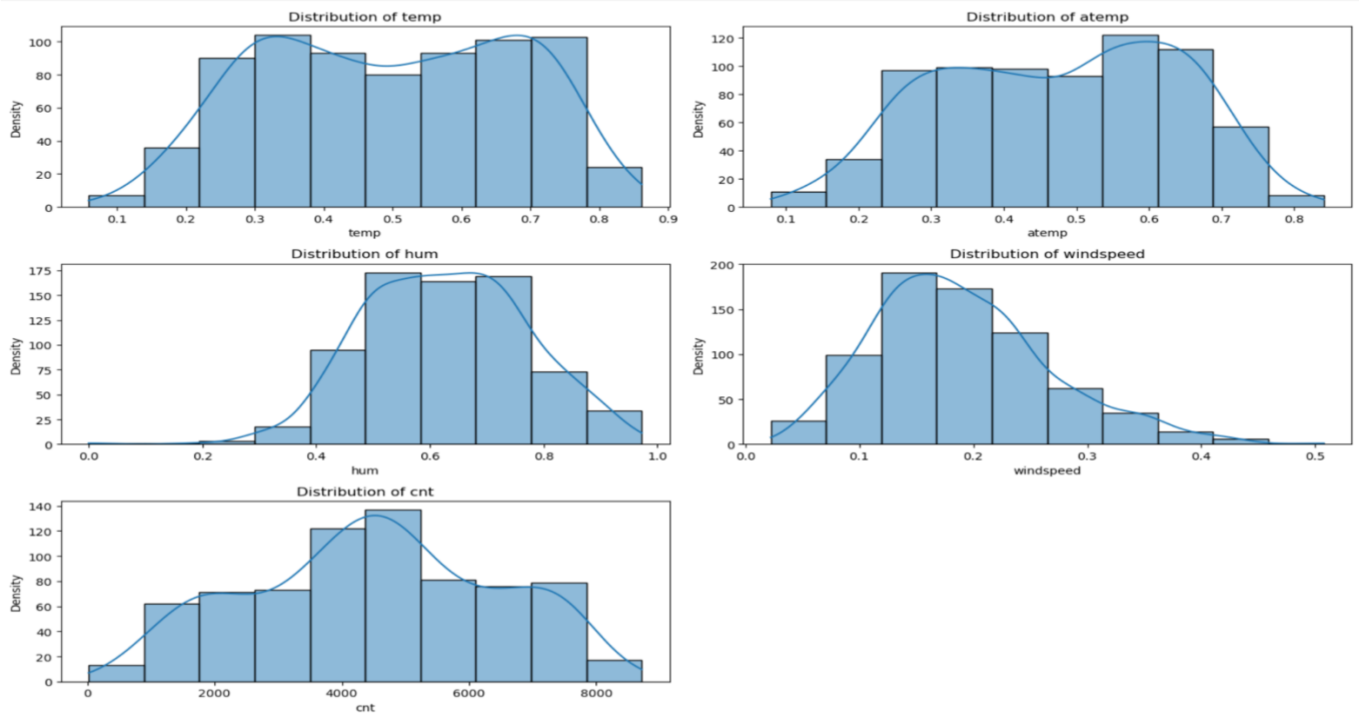
```
   instant      dteday  season  yr  mnth  holiday  weekday  workingday  \
0        1  2011-01-01       1   0     1        0        6           0
1        2  2011-01-02       1   0     1        0        0           0
2        3  2011-01-03       1   0     1        0        1           1
3        4  2011-01-04       1   0     1        0        2           1
4        5  2011-01-05       1   0     1        0        3           1

   weathersit      temp     atemp       hum  windspeed  casual  registered  \
0           2  0.344167  0.363625  0.805833   0.160446     331         654
1           2  0.363478  0.353739  0.696087   0.248539     131         670
2           1  0.196364  0.189405  0.437273   0.248309     120        1229
3           1  0.200000  0.212122  0.590435   0.160296     108        1454
4           1  0.226957  0.229270  0.436957   0.186900      82        1518

    cnt
0   985
1   801
2  1349
3  1562
4  1600
```

Exploratory Data Analysis (EDA)

The dataset underwent several checks to ensure its readiness for analysis. Missing values were examined using isnull().sum(), and the absence of missing data was visually confirmed through a missingness matrix (msno.matrix). The structure and uniqueness of columns were verified using nunique() and shape, confirming that the dataset was clean and well-organized. To streamline the dataset, the instant column, which served as an arbitrary index with no analytical value, was removed.

The distributions of key variables were analyzed through visualizations and descriptive statistics. Histograms were used to understand the overall spread of each variable, providing insights into their shapes and potential multimodality. For instance, temperature (temp) and feels-like temperature (atemp) histograms revealed a bimodal distribution, reflecting seasonal peaks likely tied to summer and winter. Humidity (hum) histograms showed a slight right skew, with most values in the moderate-to-high range, while wind speed (windspeed) histograms exhibited a strong right skew, indicating that most days experienced low wind speeds with a few higher outliers. Box plots were also employed to assess variability and detect outliers within key attributes.

Finally, the bike rental count (cnt) showed a multimodal distribution in its histogram, suggesting peaks during high-demand seasons. These analyses provided a foundation for further exploration, highlighting temperature and seasonal patterns as key factors influencing bike rentals, with humidity and windspeed having less prominent roles.

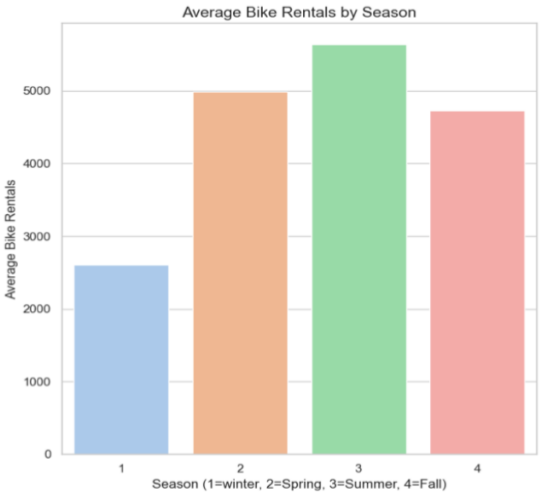


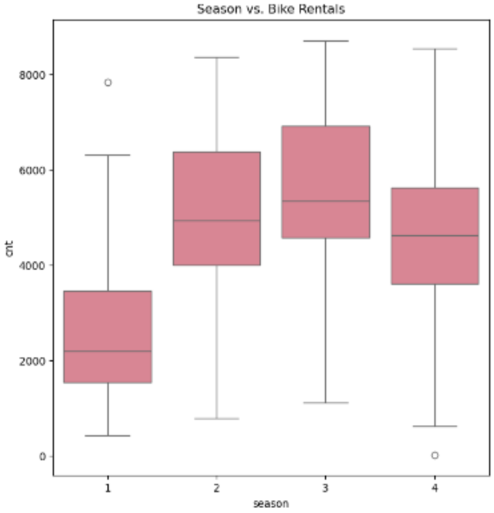
Seasonal and Temporal Analysis

Understanding seasonal and temporal factors is crucial for revealing the underlying patterns in bike rental behavior. Several visualizations and analyses were conducted to explore these patterns and extract meaningful insights.

Seasonal Trends

Bike rentals show distinct seasonal patterns, with the lowest rentals observed in Winter (Season 1) and the highest in Summer (Season 3). Spring (Season 2) follows Summer in popularity, as moderate and pleasant weather encourages outdoor activities. Conversely, Winter rentals remain low, likely due to harsh weather deterring bike usage. Monthly trends reveal a steady increase in rentals from January, peaking during the summer months (June, July, and August), and tapering off toward December.

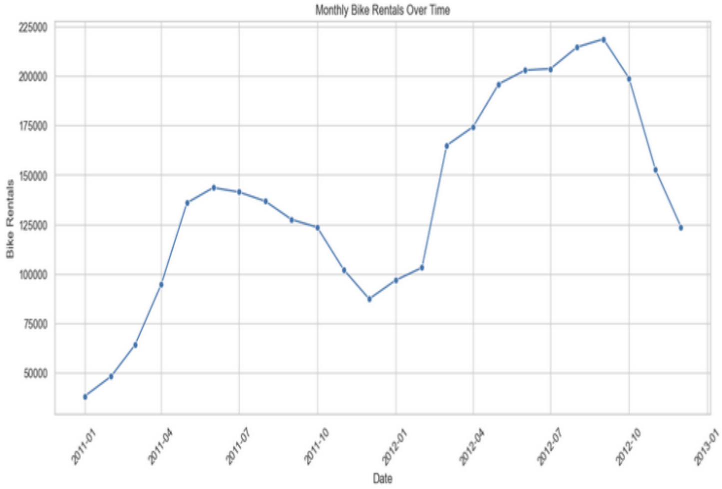


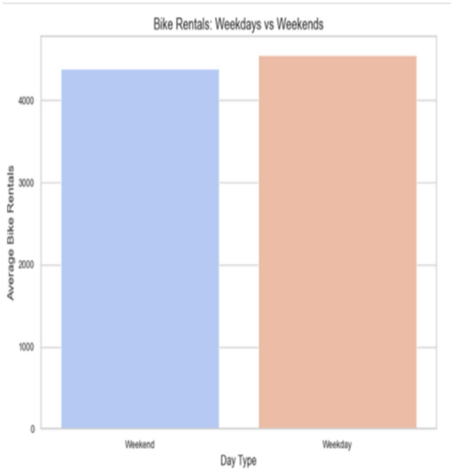


Temporal Trends

Yearly and monthly rental patterns, examined through line plots, revealed a significant increase in rentals in 2012 compared to 2011, indicating the growing popularity of the bike-sharing program. Monthly trends showed that rentals were highest in July and September, reflecting increased demand during summer months, likely influenced by events or vacations.

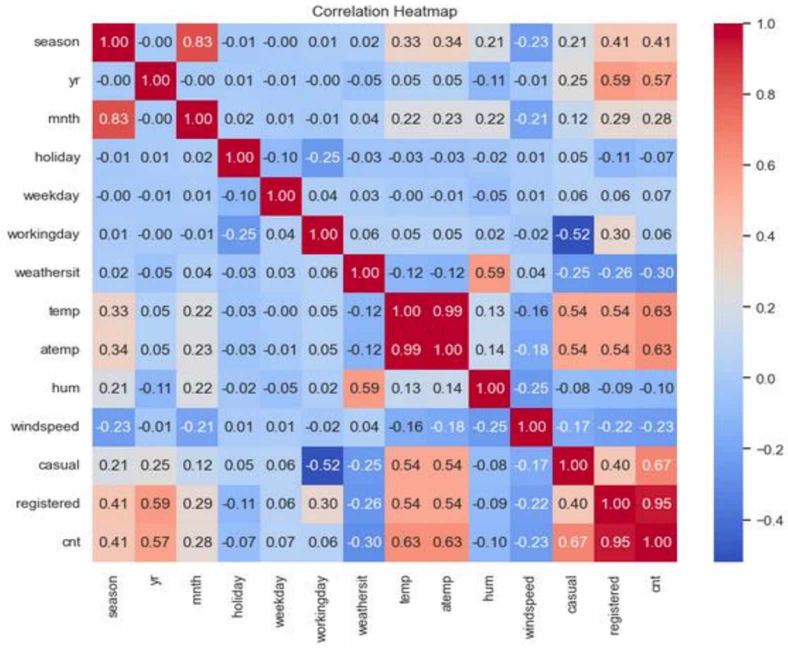
Analyzing weekday versus weekend rentals revealed slightly higher rentals on weekdays, as shown by bar plots and violin plots, suggesting that bikes were primarily used for commuting. Weekend rentals showed greater variability, likely reflecting recreational or leisure activities. These patterns provide valuable insights into user behavior, enabling targeted strategies for weekday commuters and weekend leisure users.

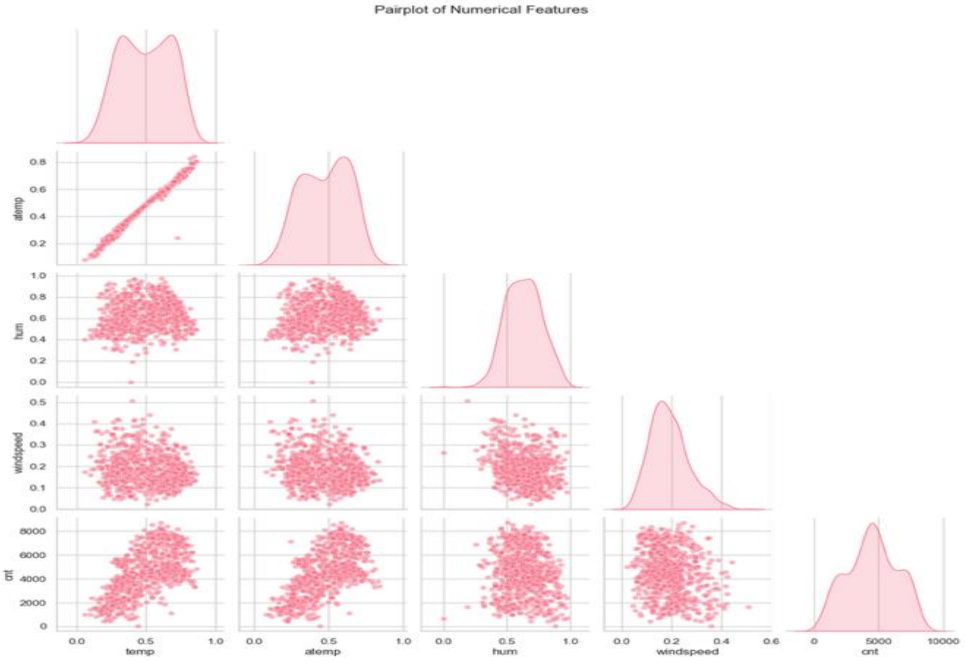




Exploring Variable Relationships

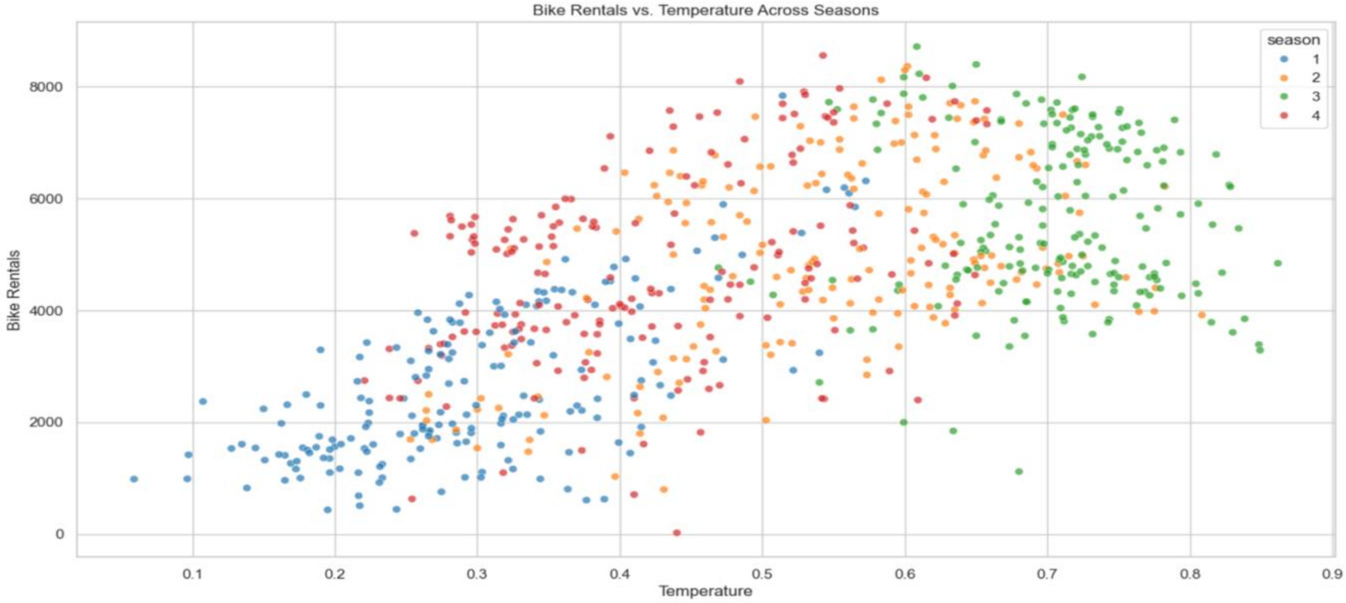
Correlation analysis revealed strong positive relationships between bike rentals (cnt) and temperature-related features (temp and atemp), indicating that higher temperatures encourage bike usage. Wind speed and poor weather conditions showed negative impacts, deterring rentals. **Pairplots** confirmed a strong positive relationship between temperature and rentals, while humidity and wind speed had weaker effects. The high correlation between temp and atemp suggested potential redundancy, aiding in feature selection. These insights highlight the influence of environmental factors on user behavior and support system optimization.





Weather's Influence on Rentals

Visual analysis, including scatter plots of temperature against rentals and violin plots showing rental distribution across seasons and weather conditions, highlighted that temperature is the most significant weather factor. Rentals increased with rising temperatures, particularly in Summer and Fall. In contrast, poor weather conditions such as light or heavy rain significantly reduced rentals, showcasing the sensitivity of bike usage to adverse weather. Wind speed and humidity were found to have minimal effects, reinforcing temperature as the dominant driver of bike rental patterns.



Weekday vs. Weekend Rental Patterns

To explore differences in rental behavior between weekdays and weekends, bar plots, box plots, and facet grids were used. The analysis revealed that rentals were consistently higher on weekdays, particularly in Winter and Fall, suggesting a focus on weekday commuting during these seasons. Statistical tests provided further validation:

- **ANOVA:** No significant differences were found between weekday and weekend rentals in most seasons (P-value > 0.05), although Winter approached significance, hinting at a possible dominance of weekday rentals.
- **Kruskal-Wallis:** This non-parametric test identified a significant difference in Winter rentals between weekdays and weekends (P-value ≈ 0.042), indicating a distinct weekday preference in colder months.

Outlier Detection and Handling

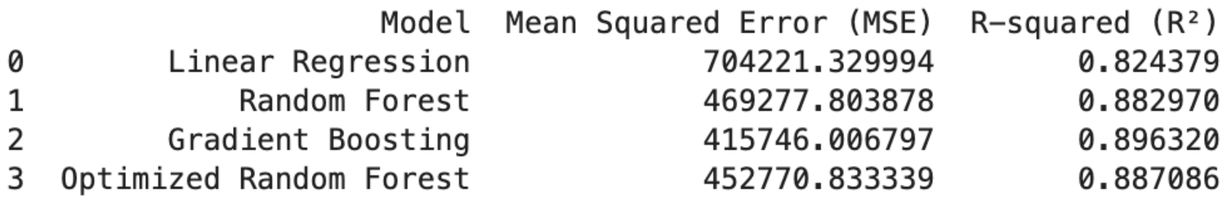
Outliers were detected using the Interquartile Range (IQR), with 14 data points identified and removed to reduce noise and variability in key metrics. This step ensured cleaner analysis and improved model performance by eliminating extreme values that could distort patterns and predictions.

Feature Scaling and Encoding

Continuous variables were standardized to ensure equal weight during modeling, while categorical features such as season and weather conditions were one-hot encoded for better representation. These preprocessing steps enhanced the model's ability to capture relationships between features and rentals, leading to more reliable predictions.

Modeling and Evaluation

Three machine learning models—Linear Regression, Random Forest, and Gradient Boosting—were used to predict bike rentals. Gradient Boosting outperformed the others, achieving a **Mean Squared Error (MSE) of 415,**746 and an **R-squared (R^2) value of 0.896**, indicating excellent predictive performance. SHAP analysis revealed that temperature was the most influential predictor, followed by seasonality and weather conditions, underscoring the importance of these features in rental patterns.



```
                    Model  Mean Squared Error (MSE)  R-squared (R²)
0       Linear Regression             704221.329994        0.824379
1           Random Forest             469277.803878        0.882970
2       Gradient Boosting             415746.006797        0.896320
3  Optimized Random Forest            452770.833339        0.887086
```

Conclusion

Weather's Influence on Bike Rentals

Temperature is the most significant factor affecting bike rental patterns. A strong positive correlation exists between temperature and the total number of rentals, with warmer temperatures driving higher bike usage. This trend is most pronounced in Spring and Summer, which see the highest rental volumes due to favorable weather conditions. In contrast, rentals are significantly lower in Winter, even on warmer days, likely due to reduced demand during the colder season. Humidity and wind speed show weak negative correlations with rentals, suggesting that while these factors have some impact, their effects are much less significant than temperature. Seasonal variations further highlight that Spring and Summer are peak seasons for rentals, followed by Fall, while Winter consistently records the lowest rental volumes.

Weekday vs. Weekend Rentals Across Seasons

Analysis of weekday versus weekend rentals reveals no significant differences in most seasons, including Spring, Summer, and Fall, where P-values from statistical tests exceed the threshold for significance. Rentals are consistently high in these seasons, irrespective of the day type, likely driven by a combination of commuting and leisure activities. However, Winter shows a potential difference, with higher rentals on weekdays compared to weekends, as indicated by a P-value of approximately 0.042 in the Kruskal-Wallis test. This trend suggests that weekday rentals in Winter are influenced by commuter behavior, while leisure-related weekend rentals are less prevalent in the colder months. Overall, seasonal and weather factors have a much more substantial impact on rental patterns than day type.

References

- UCI Bike Sharing Dataset. (https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset).
- McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference.
- Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment*. *Computing in Science & Engineering*.
- Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*.
- Waskom, M., et al. (2014). *Seaborn: Statistical Data Visualization*.
- Missingno Library. (https://github.com/ResidentMario/missingno).
- Conover, W. J. (1999). *Practical Nonparametric Statistics*. Wiley.
- Anderson, T. W., & Darling, D. A. (1954). *A Test of Goodness of Fit*. *Journal of the American Statistical Association*.

Video: https://drive.google.com/drive/folders/1c0Jb-uY5dCkl-BgB6E4AQpbWrNbXO3hc?usp=drive_link