Ddareungi Bike Demand Analysis

AGENDA

- Overview
- Data Understanding
- Exploratory Data Analysis
- Hypothesis Testing and Regression
- Recommendations
- Future Vision

Addressing Demand Forecasting and Supply Optimization Challenges for Seoul's Ddareungi Bike-Sharing System

About Ddareungi

Ddareungi, also known as "Seoul bike" in English, is a bike-sharing system in Seoul, South Korea. The system was launched in October 2015 and has experienced rapid expansion since then. Key features of the system include:

- 1500 bike docking stations operating 24/7 throughout Seoul
- Stations located in high-traffic areas like subway entrances, bus stops, and public offices
- Users can check bike availability, rent, and return bikes using a mobile app

Business Complication

As the bike-sharing system grows in popularity, Ddareungi faces a crucial challenge:

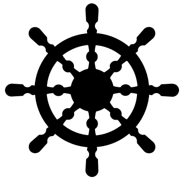
- Ensuring a stable supply of rental bikes to meet demand
- Minimizing waiting times for users
- Accurately estimating the required number of bikes at each hour to maintain efficient service

Overarching Question

How can we accurately estimate the hourly bike rental demand in Seoul to ensure a stable supply of rental bikes?

Key Objectives for the Analysis

Understand Demand Drivers



- Conduct a comprehensive analysis to identify and quantify the key factors affecting bike rental demand, such as weather conditions, time of day, seasons, and holidays
- This will help uncover patterns and correlations that influence usage, providing insights into customer behavior and operational challenges

Develop Predictive Insights

- Leverage advanced machine learning techniques to build robust models capable of accurately forecasting hourly bike rentals
- These predictive insights will enable Ddareungi to anticipate fluctuations in demand, proactively address supply gaps, and improve overall operational efficiency

Enhance Operational Strategies

- Translate analytical findings into actionable recommendations for optimizing bike availability across docking stations
- This includes strategies for dynamic bike redistribution, weather-based planning, and targeted promotions, aimed at ensuring resource efficiency and enhancing customer satisfaction

Comprehensive Overview of Ddareungi Bike-Sharing System Data

The dataset provided by Ddareungi gives comprehensive hourly records of bike rentals from December 2017 through November 2018. The dataset contains 8,760 records with 14 variables, including both temporal and weather-related features

- The 14 variables capture temporal, weather, and operational aspects
- Key variables influencing bike rental demand are Rented Bike Count (the target variable), Hour, Seasons, and weather-related features such as Temperature, Rainfall, and Snowfall
- Additional categorical variables, such as Holiday and Functioning Day, provide context about demand variability

	Rented Bike Count	Hour	Temperature (°C)
count	8760	8760	8760
mean	705	11.5	13
std	645	6.922582	13
min	0	0	-17.8
25%	191	5.8	3.5
50%	504	11.5	13.7
75%	1065	17.3	22.5
max	3556	23	39.4

Transforming Data to Ensure Integrity and Consistency

For the analysis of Seoul's bike-sharing data, the dataset was inherently clean, with no missing or duplicate values, providing a solid foundation for analytical modeling. However, transformations and feature engineering were performed to ensure the data was uniform, comprehensive, and optimized for machine learning models

Feature Engineering

Temporal Variables:

- Created derived variables to enhance temporal analysis, such as:
  - is_weekdays: A binary feature indicating weekends versus weekdays.
- Deleted Variables such as Date (after extracting days of the week and month) and Dew point temperature to remove redundancy

Seasonal Features:

- The Seasons variable was encoded into four binary variables to allow the models to independently assess the impact of each season on bike rentals

Normalization

Continuous variables such as Temperature, Humidity, Visibility, and Wind Speed were scaled to standardize ranges and reduce variability, improving model interpretability and stability during training

Encoding Categorical Variables

Variables such as Holiday and Functioning Day were converted into binary numeric representations, ensuring consistency in handling categorical data

Logarithmic Transformation

Used log transformation to balance and normalize skewed data for improving the accuracy of the statistical model

Analyzing Key Correlations Between Weather, Time, and Bike Rental Demand in Seoul's Ddareungi System

Observations from Correlation Matrix

Strong Positive Correlations:

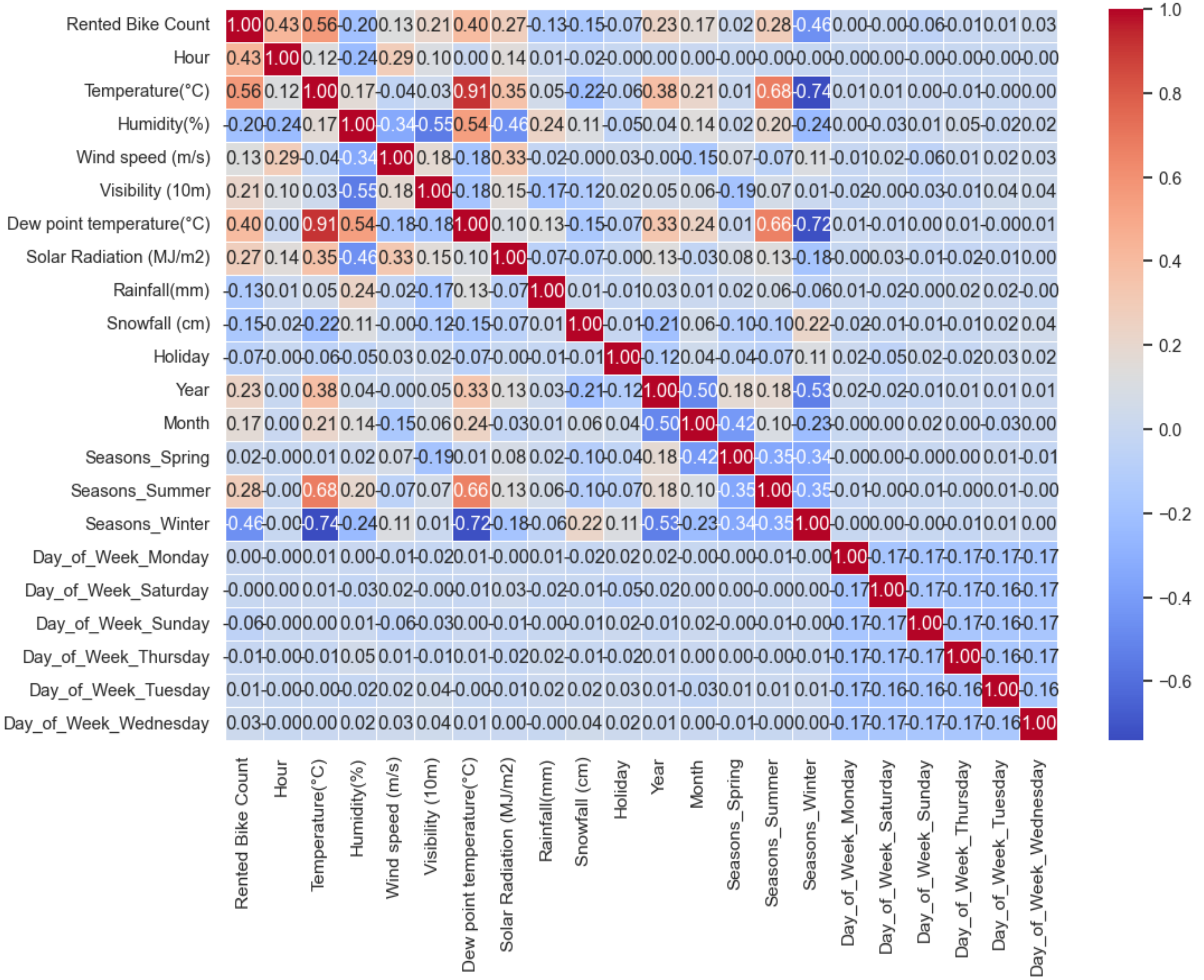
1. **Temperature:** Higher temperatures are associated with increased bike rentals, likely due to favorable weather conditions encouraging outdoor activities
2. **Hour:** Certain hours, particularly during rush hours, strongly correlate with higher rentals, highlighting time-of-day-driven demand patterns

Negative Correlations:

1. **Rainfall and Snowfall:** A negative impact on demand, as adverse weather discourages outdoor commuting and leisure activities

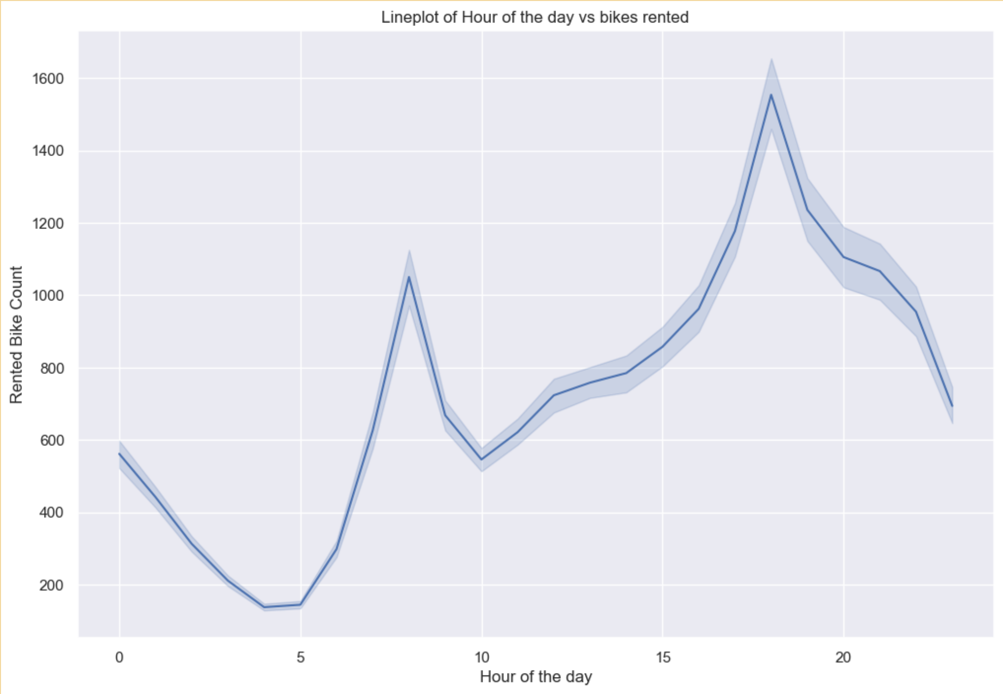
Inter-variable Relationships:

1. **Temperature and Visibility:** Better visibility often coincides with higher temperatures, as clear weather conditions are conducive to increased outdoor activity

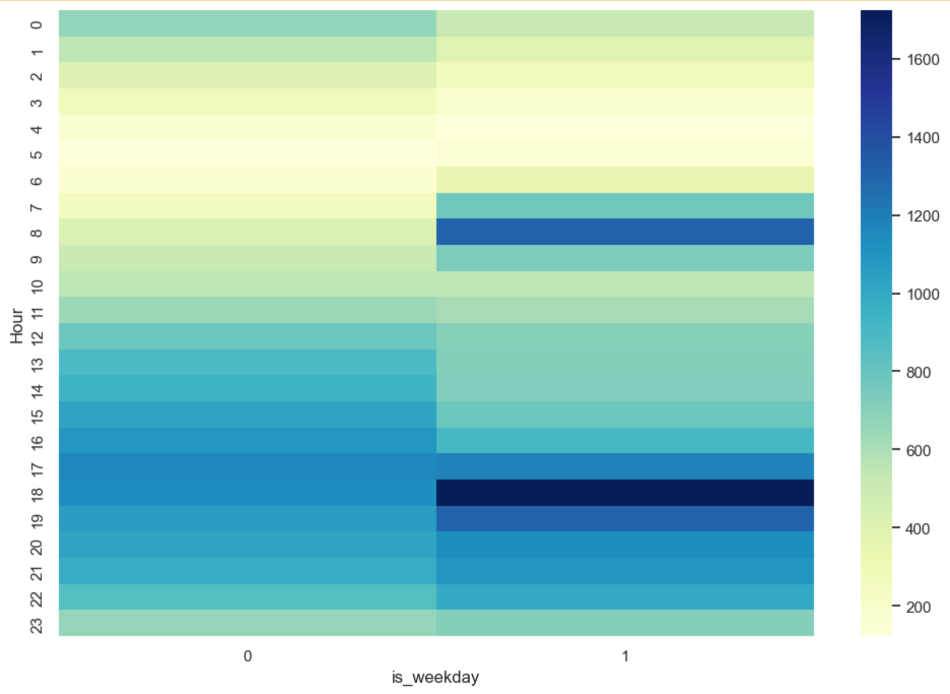


Visualizing Hourly Bike Rental Trends by Comparing Weekday and Weekend Patterns Using Line Graphs and Heatmaps

Line plot Hour of the Day vs. Bikes Rented



Heatmap of Weekday vs Hour of the day



- The line plot reveals distinct peaks in bike rentals during the morning rush hour (7–9 AM) and evening rush hour (5–7 PM)
- Lower demand is observed during nighttime (11 PM–5 AM), indicating minimal usage for non-commuting purposes

- Rentals are significantly higher on weekdays compared to weekends
- The heatmap highlights that weekday usage aligns closely with commuting hours, whereas weekend usage is more evenly distributed, reflecting leisure activities

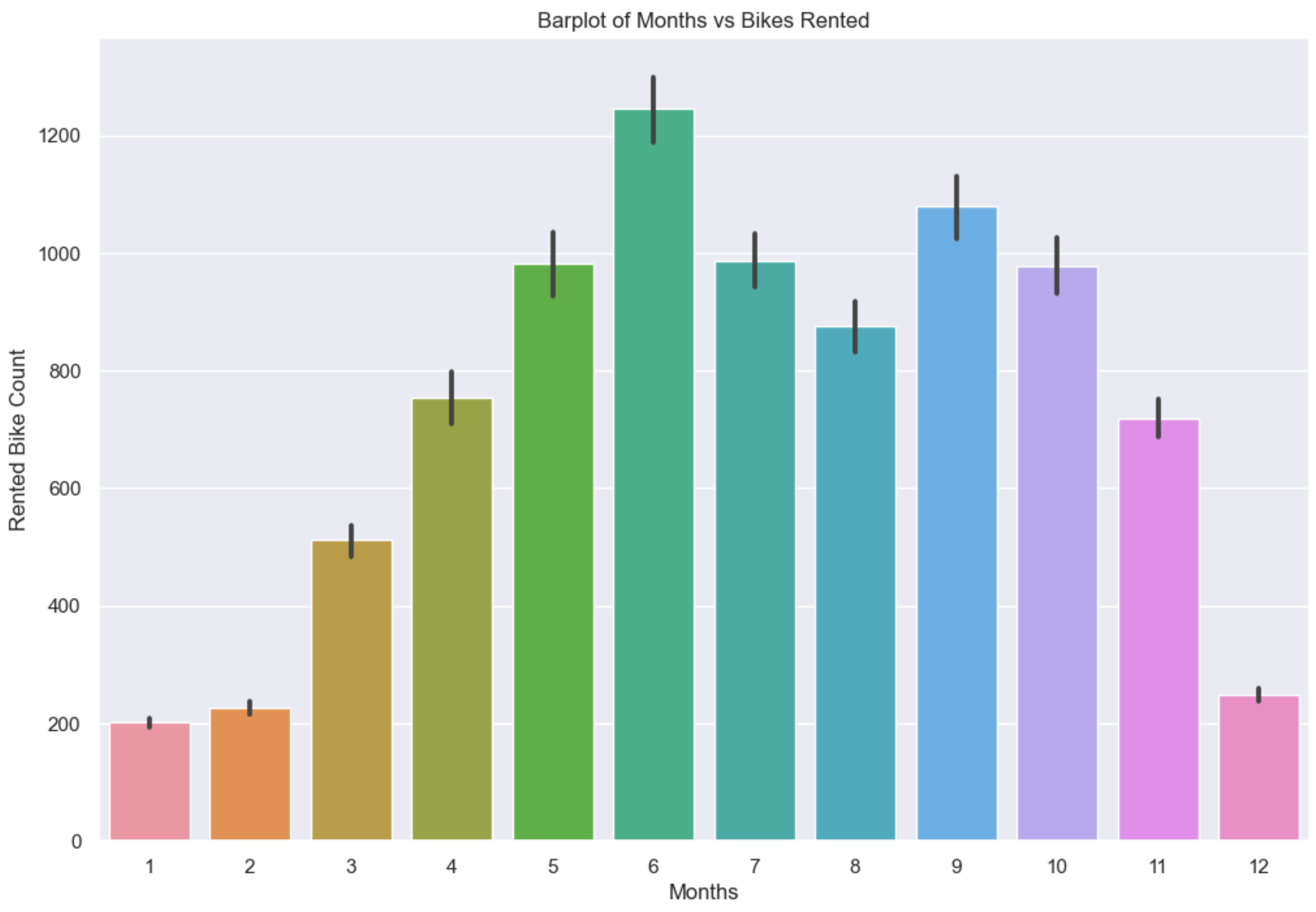
Visualizing Monthly Bike Rental Trends with Peak Summer Usage and Winter Decline

Seasonal trends:

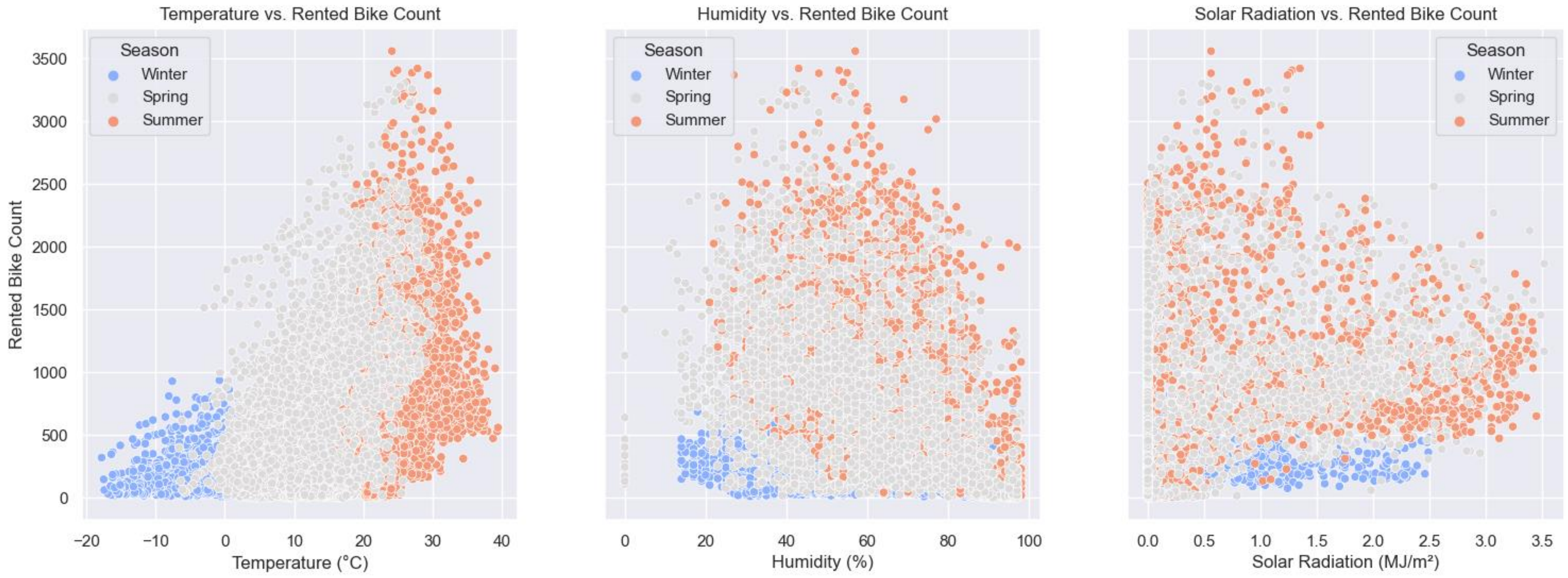
- There is a clear seasonal pattern in bike rentals, with higher usage during warmer months and lower usage during colder months
- Peak rental periods occur during summer months (June, July, August), with the highest rentals typically in July
- The lowest rental periods are during winter months (December, January, February), with December or January usually having the fewest rentals

User type differences:

- Registered users tend to rent bikes more consistently throughout the year, though their usage also increases in warmer months
- Casual users show much more pronounced seasonal variation, with very low usage in winter months and significantly higher usage in summer months



Analyzing the Impact of Weather-Related Variables on Bike Rental Demand Across Seasons



Observations:

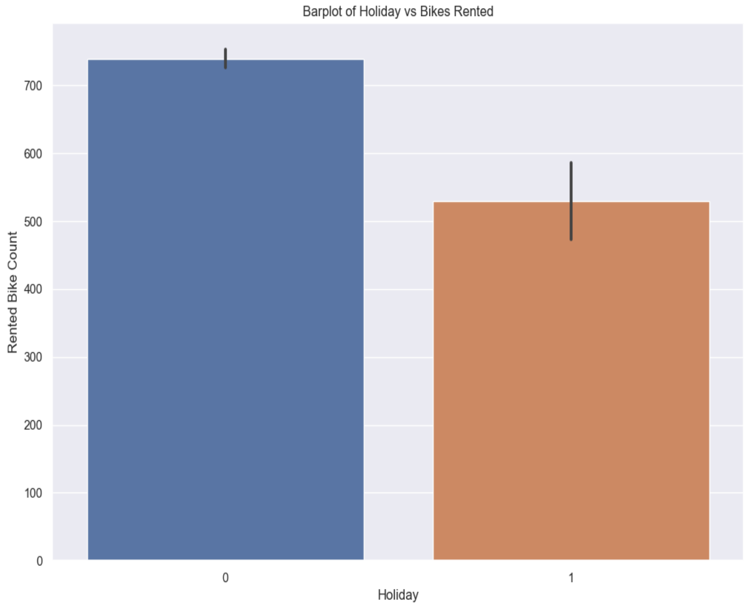
- - Bike rentals increase with rising temperatures
- - Higher humidity correlates with fewer rentals, but the effect is weaker than temperature
- - More solar radiation (sunnier weather) is linked to increased bike rentals

Relationship with the Months vs Bikes Rented Barplot

- The bar plot shows higher rental numbers during summer months (June, July, August) when temperatures are typically higher, days are longer (more solar radiation), and humidity is often lower.
- Conversely, winter months (December, January, February) show the lowest rental numbers, corresponding to colder temperatures, less sunlight, and potentially higher humidity

Identifying Key Factors Affecting Bike Rental Demand Through Hypothesis Testing

Hypothesis Test	Results
Mean of the Rental Bikes test (One Sample T-Test) H_0: The true population mean of hourly bike rentals is equal to 500 bikes H_1: The true population mean of hourly bike rentals is not equal to 500 bikes	T-Statistic: 32.82 P-Value: 1.61×10^{-222} H_0 is rejected
Holiday vs. Non-Holiday Bike Rentals (Independent Samples T-Test) H_0: No significant difference in bike rentals between holidays and non-holidays H_1: Significant difference	F-Statistic: 8.81 P-Value: 0.0 H_0 is rejected
High vs. Low Humidity Days (One-Way ANOVA) H_0: No significant difference in rentals between high and low humidity days H_1: Significant difference	F-Statistic: 10.15 P-Value: 6.03×10^{-127} H_0 is rejected



Conclusion #1:

- The actual average of 729.16 bikes rented per hour is significantly higher than the hypothesized value of 500 bikes. The extremely low p-value provides strong statistical evidence that this difference is not due to chance

Conclusion #2:

- The null hypothesis is rejected for the holiday vs. non-holiday test, indicating a statistically significant difference in bike rentals between holidays and non-holidays. On average, there are fewer bike rentals on holidays compared to non-holidays

Conclusion #3:

- Humidity levels significantly affect bike rental patterns, with different humidity levels associated with varying rental numbers

Assessing the impact of environmental factors on bike rentals using Multiple Linear Regression

Critical Business Insights

Weather Impact

- Temperature is the strongest positive driver (+145 additional rentals per °C)
- Rainfall (-23.24%) and humidity (-26.18%) significantly decrease rentals

Time Patterns

- Hourly trends show consistent rental patterns (+4.01% per hour)
- Holidays see 45% fewer rentals than regular days
- Summer shows the highest rental activity (+ 39.36% compared to autumn)

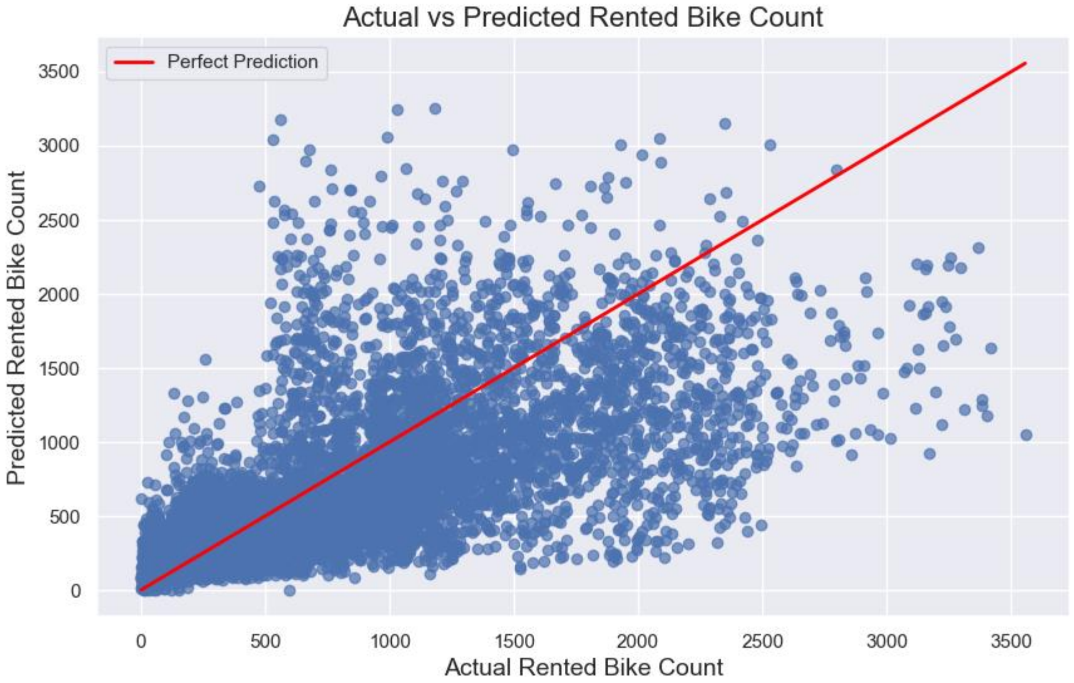
Model Reliability

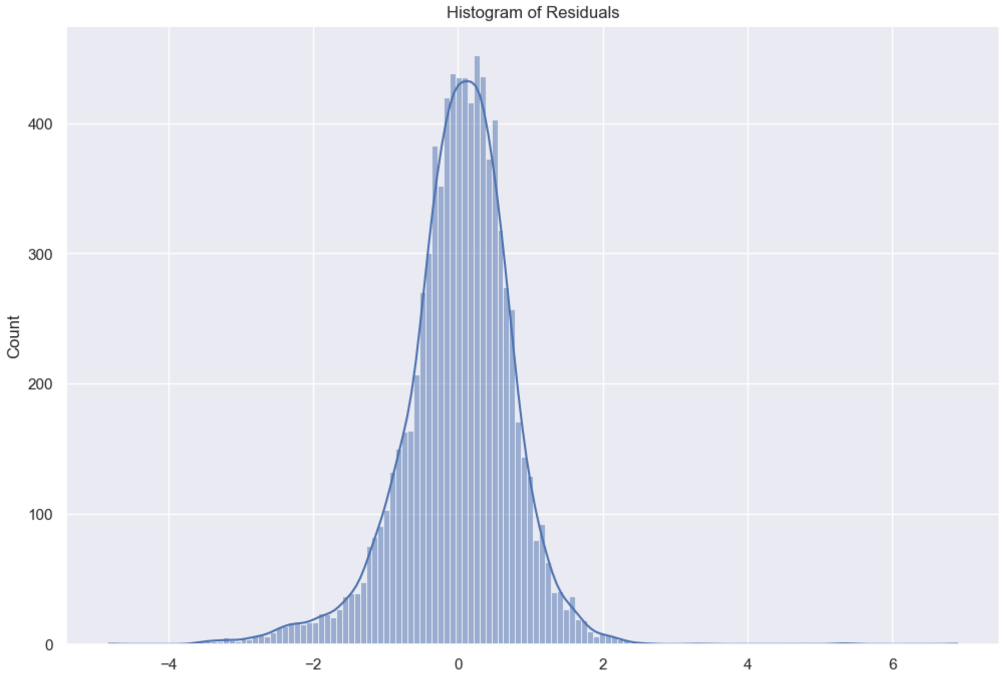
- The model explains 57.7% of the variance in bike rentals (R-squared = 0.577), indicating that temperature, hour, holidays, and weather conditions are strong predictors of rental demand

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Log_Rented_Bike_Count   R-squared:                       0.577
Model:                               OLS   Adj. R-squared:                  0.577
Method:                    Least Squares   F-statistic:                     1154.
Date:                   Sun, 17 Nov 2024   Prob (F-statistic):               0.00
Time:                           18:06:17   Log-Likelihood:                -9579.0
No. Observations:                   8465   AIC:                         1.918e+04
Df Residuals:                       8454   BIC:                         1.926e+04
Df Model:                             10
Covariance Type:               nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                       5.6262      0.022    251.906      0.000       5.582       5.670
Hour                        0.0393      0.001     31.560      0.000       0.037       0.042
Temperature(°C)             0.7878      0.013     60.075      0.000       0.762       0.813
Humidity(%)                -0.3036      0.013    -23.344      0.000      -0.329      -0.278
Visibility (10m)            0.0441      0.010      4.284      0.000       0.024       0.064
Solar Radiation (MJ/m2)    -0.0569      0.011     -5.181      0.000      -0.078      -0.035
Rainfall(mm)               -0.2640      0.008    -31.274      0.000      -0.281      -0.247
Snowfall (cm)              -0.0196      0.009     -2.292      0.022      -0.036      -0.003
Holiday                    -0.4495      0.038    -11.752      0.000      -0.524      -0.375
Seasons_Summer             -0.3319      0.026    -12.748      0.000      -0.383      -0.281
is_weekday                  0.1720      0.018      9.533      0.000       0.137       0.207
==============================================================================
Omnibus:                     1017.712   Durbin-Watson:                   0.528
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4245.220
Skew:                          -0.538   Prob(JB):                         0.00
Kurtosis:                       6.298   Cond. No.                         63.4
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Evaluating Model Performance by Comparing Actual vs Predicted Bike Rentals and Analyzing Residual Distribution





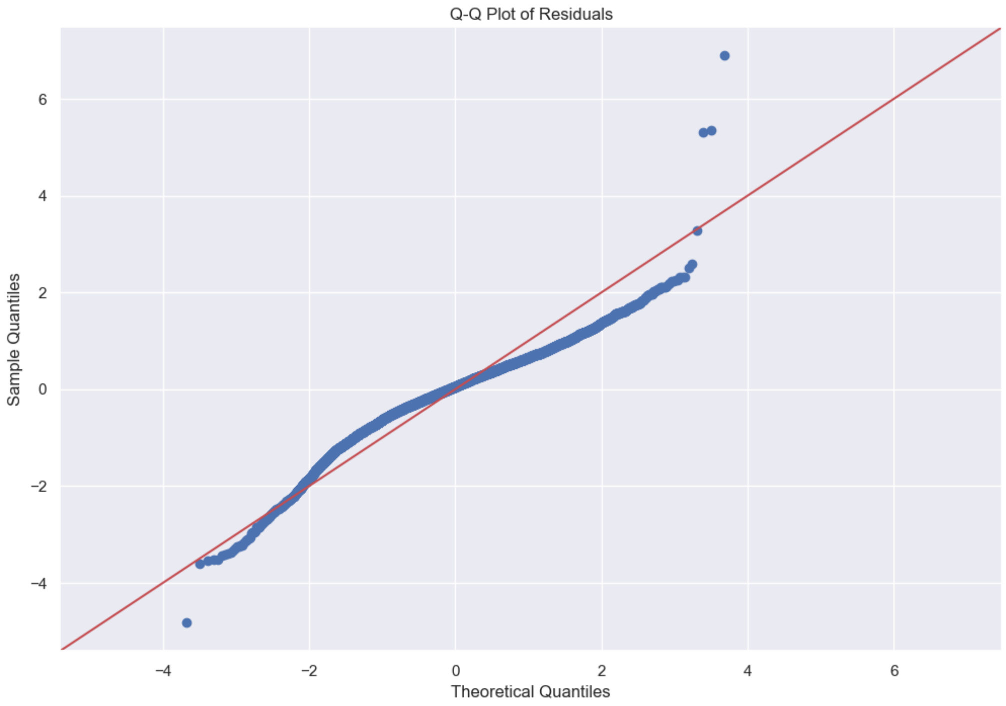
Purpose: Compare the actual bike rentals to the predicted rentals from the model

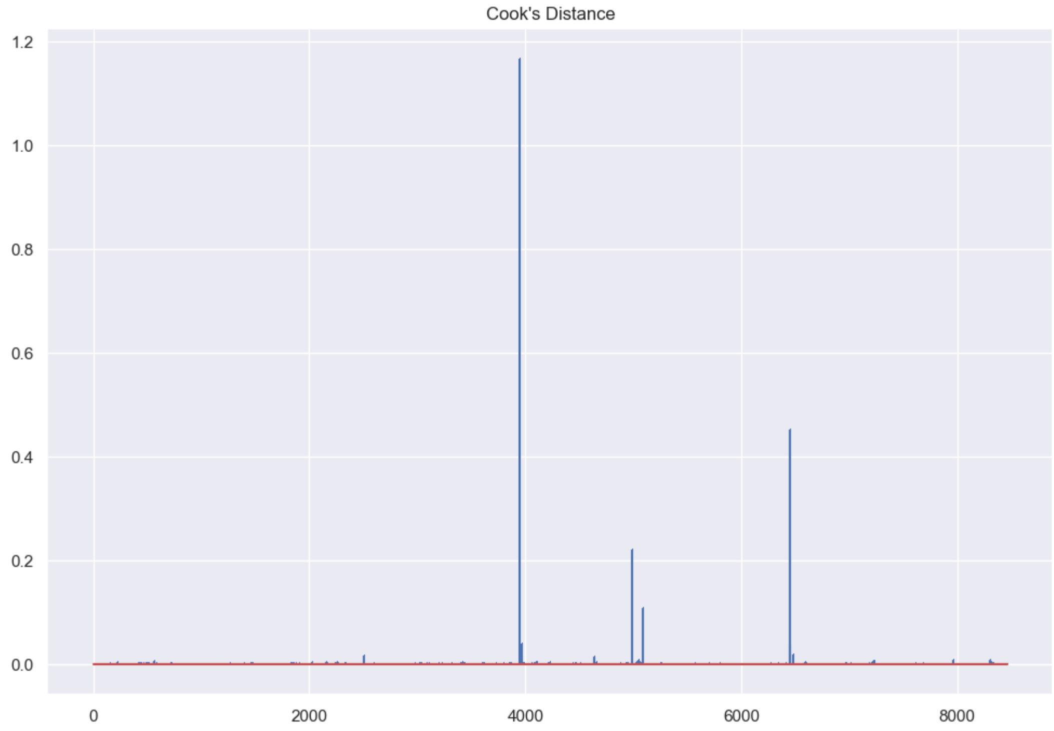
Key Insight: The red line indicates perfect predictions. The model performs well for lower rental counts but struggles with higher counts, underestimating demand during peak periods

Purpose: Displays the differences (residuals) between predicted and actual rentals to show the model's prediction accuracy

Key Insight: The residuals are mostly centered around zero, indicating that most predictions are close to the actual values. However, there are some outliers, meaning that for some instances, the model is far off in its predictions

Assessing Model Assumptions by Evaluating Residual Normality with Q-Q Plot and Identifying Influential Data Points Using Cook's Distance





Purpose: This plot checks if the residuals follow a normal distribution, which is important for model accuracy.

Key Insight: Most points align with the red line, indicating normal distribution of residuals, but outliers at both extremes show the model struggles with extreme rental values.

Purpose: Identifies data points that have a large influence on the model's predictions

Key Insight: Most data points have low Cook's distance and little influence on the model, The single spike that extends above 1.0 in the plot indicates a highly influential observation in the dataset

Key Insights from VIF values

Hypothesis Test	VIF Factor

Initial VIF Calculation:

- The first VIF calculation included several variables, such as Seasons (Spring, Summer, Winter), Month, and Year, which showed high VIF values
- High VIF values suggest multicollinearity, meaning these variables are highly correlated with each other or other predictors in the model
- Multicollinearity can lead to unstable coefficient estimates and make it difficult to assess the true effect of each variable

	VIF Factor	features
0	0.00	const
1	1.21	Hour
2	5.18	Temperature(°C)
3	2.65	Humidity(%)
4	1.30	Wind speed (m/s)
5	1.70	Visibility (10m)
6	1.94	Solar Radiation (MJ/m2)
7	1.07	Rainfall(mm)
8	1.13	Snowfall (cm)
9	1.03	Holiday
10	18.20	Year
11	22.56	Month
12	14.57	Seasons_Spring
13	6.27	Seasons_Summer
14	28.69	Seasons_Winter
15	inf	is_weekday
16	inf	is_weekend

Refined VIF Calculation:

- After detecting multicollinearity in the first calculation, some variables were removed or combined to reduce redundancy.
- The second VIF calculation shows much lower values across all variables. The highest VIF is now 7.49 for the constant term, which is acceptable.
- This indicates that multicollinearity has been reduced, and the model is more stable.

	VIF Factor	features
0	7.76	const
1	1.19	Hour
2	2.63	Temperature(°C)
3	2.55	Humidity(%)
4	1.27	Wind speed (m/s)
5	1.59	Visibility (10m)
6	1.93	Solar Radiation (MJ/m2)
7	1.07	Rainfall(mm)
8	1.10	Snowfall (cm)
9	1.01	Holiday
10	1.96	Seasons_Summer
11	1.00	is_weekday

Interpretation from the Latest VIF Calculation:

- Temperature (VIF = 2.63): This variable has a low VIF, meaning it does not have strong multicollinearity with other predictors.
- Humidity (VIF = 2.55) and Solar Radiation (VIF = 1.93) also show low VIF values, indicating they are independent enough to remain in the model without causing issues.
- Constant Term (VIF = 7.76): While this is higher than other variables, it is still within an acceptable range and does not pose a significant threat to model stability.

Optimizing Bike Redistribution, Operational Efficiency, and Pricing Strategies to Enhance Customer Satisfaction and Reduce Costs

Optimize Bike Redistribution

Time-Based Distribution:

- Focus on peak hours (5-7 PM) with average 1,200 bikes/hour demand
- Reduce fleet in low-demand hours (11 PM-4 AM) for maintenance

Weather-Optimized Operations:

- Increase fleet by 30% during optimal conditions (Temperature >20°C, Humidity <60%)
- Reduce operations during rainfall (data shows 23.24% decrease in demand)

Operational Improvements

Maintenance Schedule:

- Schedule major maintenance during predicted low-demand periods
- Implement preventive maintenance before peak summer season

Technology Integration:

- Real-time weather monitoring system
- Predictive demand modeling based on multiple variables

Dynamic Pricing Strategy

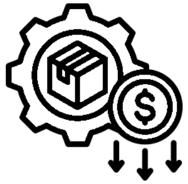
Weather-Based Pricing:

- Premium rates during optimal weather conditions
- Special winter rates to encourage ridership during cold months

Time-Sensitive Pricing:

- Peak hour surcharge (5-7 PM)
- Early bird discounts (6-8 AM)
- Holiday-specific pricing strategies (data shows 36.16% lower demand)

High Customer Satisfaction



Reduction in Operation Cost

Increase in Operational Efficiency

Exploring Future Opportunities by Enhancing Pricing Strategies, Operational Efficiency, and Predictive Maintenance for Sustainable Growth

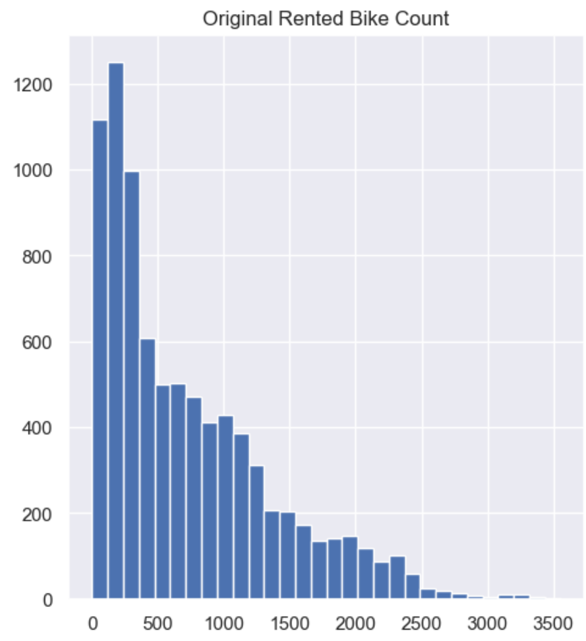
Additional Business Question	Required Data	Analytics
How do different pricing strategies affect usage patterns	Transaction history, pricing experiments data, subscription data	Price sensitivity analysis, subscription behavior modeling
Which stations consistently face bike shortages or surpluses?	Real-time station inventory, POI (Points of Interest) data, station capacity	Geospatial analysis, capacity utilization modeling
Can we predict maintenance needs before breakdowns occur?	Maintenance records, bike usage history, bike model specifications	Predictive maintenance modeling, equipment lifecycle analysis

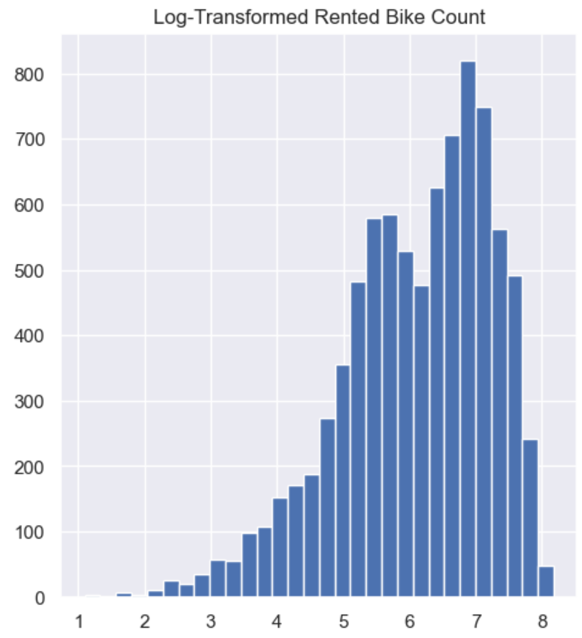
Appendix

Appendix 1: Performed Log-transformation to Improve the Accuracy Of the Statistical Models

Appendix 2: Further insight into the relationship between temperature, rented bike count, and hour of the day

Appendix 1: Performed Log-transformation to Improve the Accuracy Of the Statistical Models

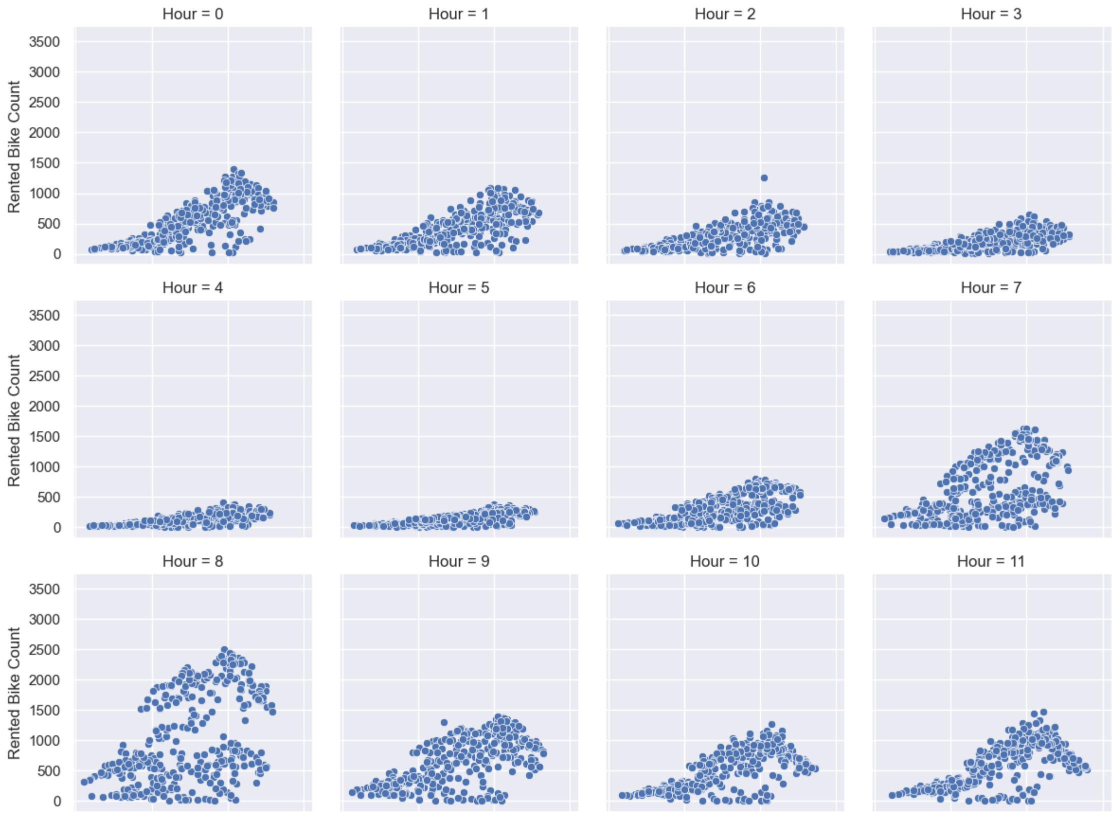




Original Distribution
- Most rental counts are concentrated in the lower range (0-1000 bikes)
- Long right tail extending to 3500 bikes (Positive Skewed)
- Shows extreme values that could skew statistical analyses
- Indicates potential outliers in high rental periods

Log-Transformed Distribution
- More balanced and normally distributed
- Easier to identify patterns and relationships
- Better suited for statistical modeling
- Helps in meeting assumptions for linear regression analysis

Appendix 2: Further insight into the relationship between temperature, rented bike count, and hour of the day



Time-Dependent Temperature Sensitivity:
The slope of the relationship between temperature and rentals varies by hour, indicating that the temperature sensitivity of bike rentals changes throughout the day.

Rental Patterns:
These peak hours also show a wider spread of rental counts, indicating other factors beyond temperature influencing rentals during these times