Regression Analysis on Bike Sharing Demand

An independent project by: Antonius Jose

Objective

The aim of this study is to create a regression model to predict the hourly demand for bike rentals and determine which variables provided have the greatest influence on bike rental demand.

Dataset

The dataset holds bike-sharing data from startups located in Korea that includes various factors like weather conditions, time of day, and public holidays to predict the demand for bike rentals in Seoul.

Attribute	Meaning
Index	Used as a row identifier
Date	Date of the observed information
Rented Bike Count	Number of bikes rented
Hour	Time of day
Temperature(C)	Air temperature measured in Celsius
Humidity(%)	Humidity of the air expressed as a percentage
Wind speed (m/s)	Air movement speed in meters per second
Dew point temperature(C)	The temperature at which condensation begins in the air
Solar Radiation (MJ/m2)	Solar energy received per square meter
Rainfall(mm)	Precipitation level measured in milimeters
Snowfall (cm)	Snow accumulated measured in centimeters
Seasons	The season the date belongs to
Holiday	Indicator showing whether the day is a public holiday
Functioning Day	Indicator showing whether bike operations was active

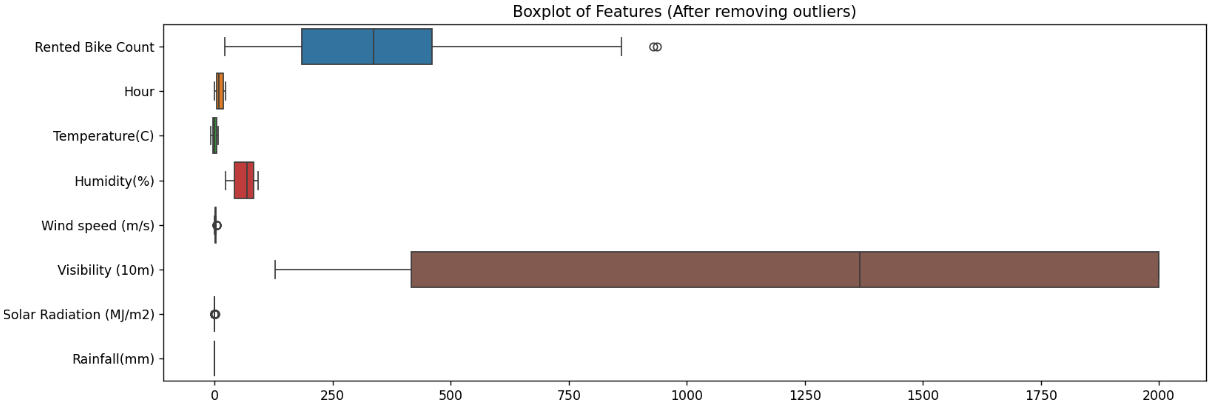
Data Cleaning

Removed Attributes

In the data cleaning process, the columns: date, index, Seasons, Holiday, Functioning Day, Snowfall (cm), and Dew point temperature(C) were removed because they were redundant or did not have a significant impact for the regression analysis.

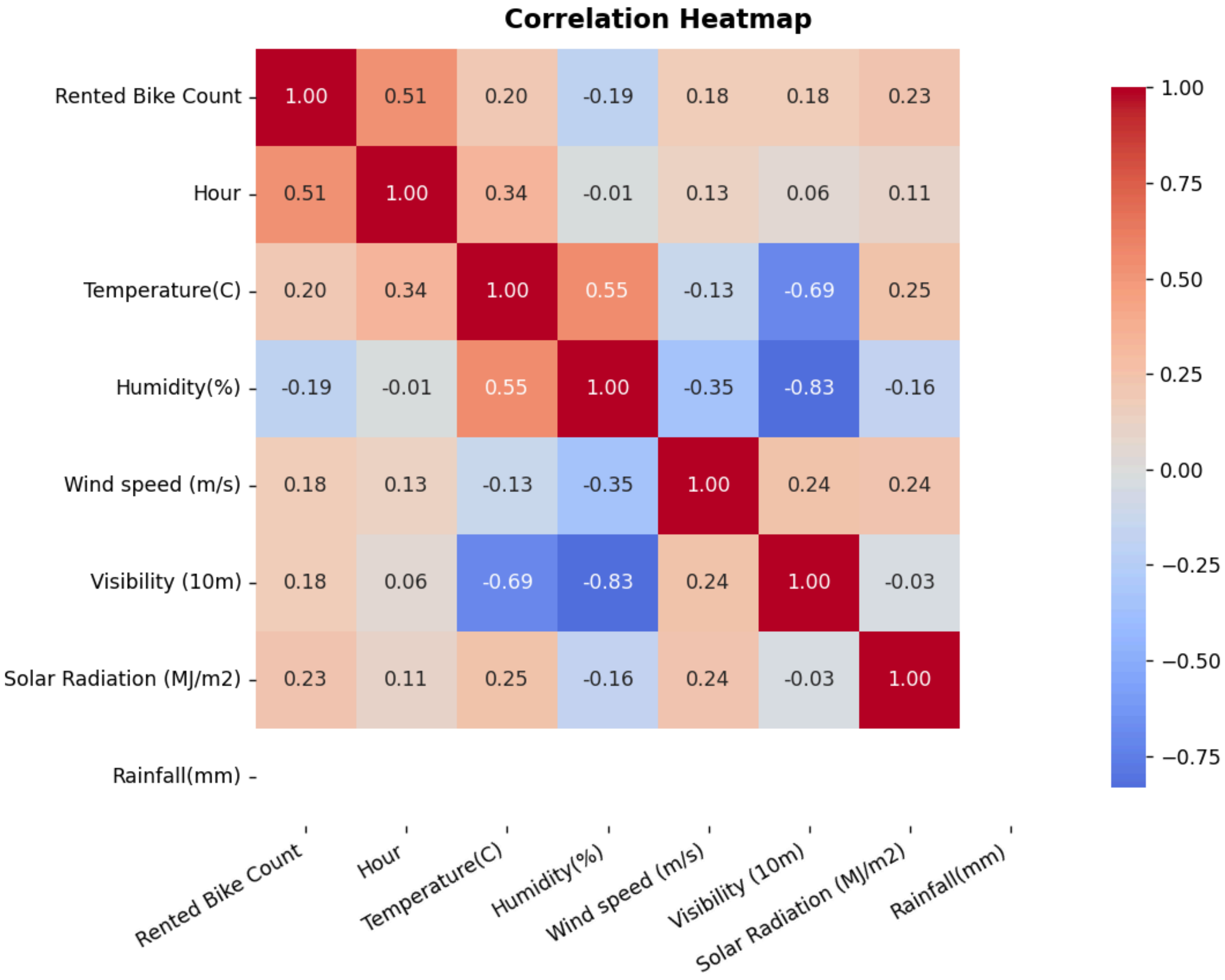
Outliers

The outliers were detected by determining the interquartile range and any value that was above or below 2xIQR was removed to ensure accuracy and prevent inflating or deflating the regression output. The result was the dataset has been reduced to 80 rows from 100 rows. The boxplot below illustrates the summary of each attribute after removing outliers.



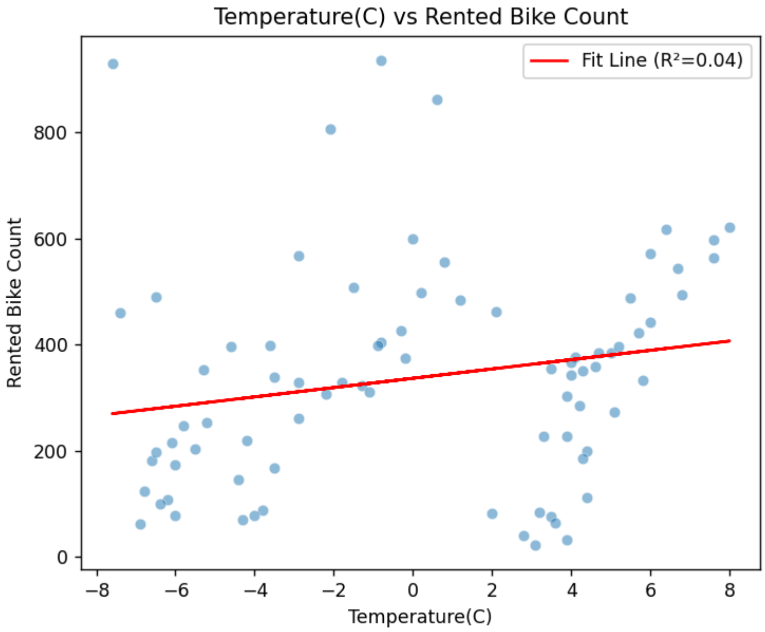
Exploratory Data Analysis

Before the model is trained, multiple visualizations were created to identify the correlations between variables and which ones were the best predictors for the dependent variable (Rented Bike Count).

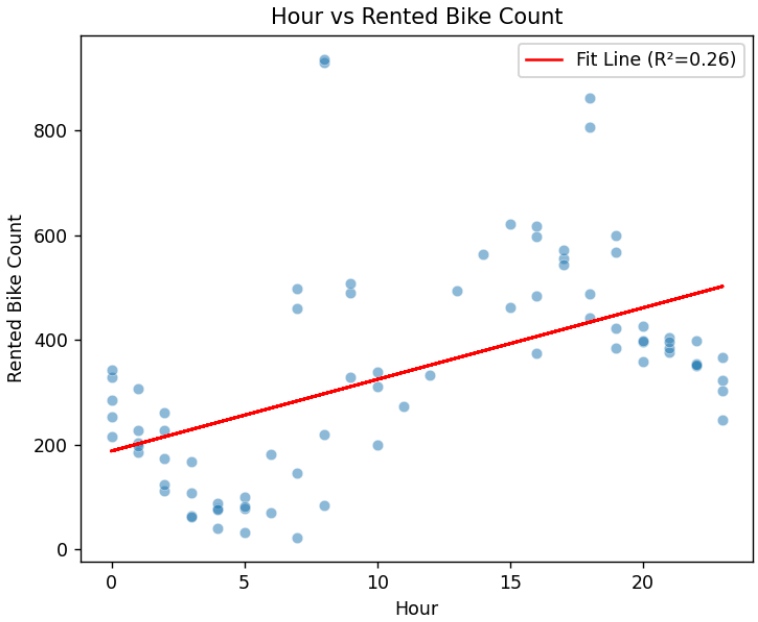


The correlation heatmap shows us the linear association between variables. The key takeaways are that Hour, Temperature and Solar radiation have a positive relationship with rental demand while the attributes wind speed have a negative relationship. In this graph, rainfall is left blank because its rare occurrence in the dataset makes it difficult to determine its correlation.

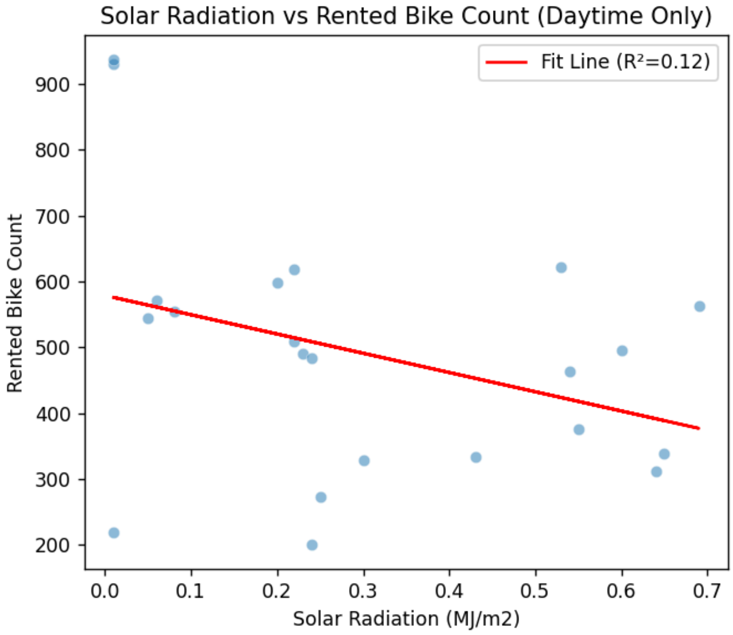
Attribute Analysis



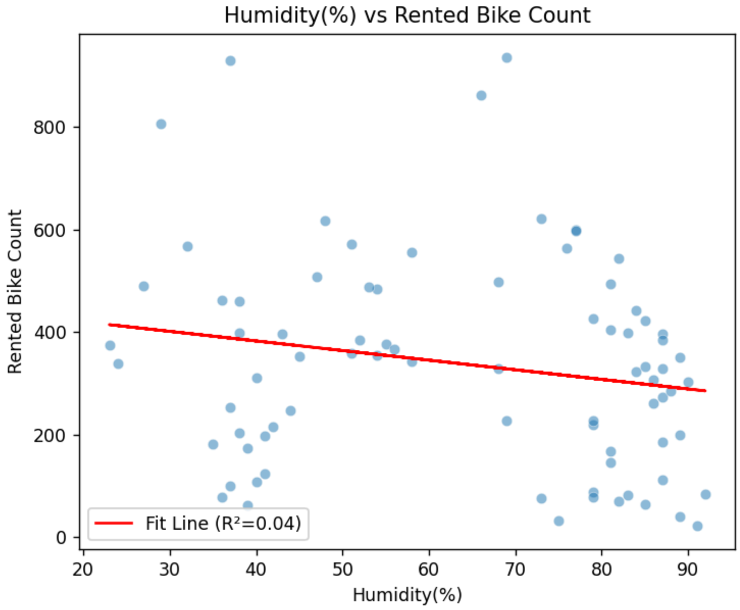
The graph suggests that temperature does not have a strong relationship with rental demand and while warmer temperatures experience higher demand, the relationship is not linear.



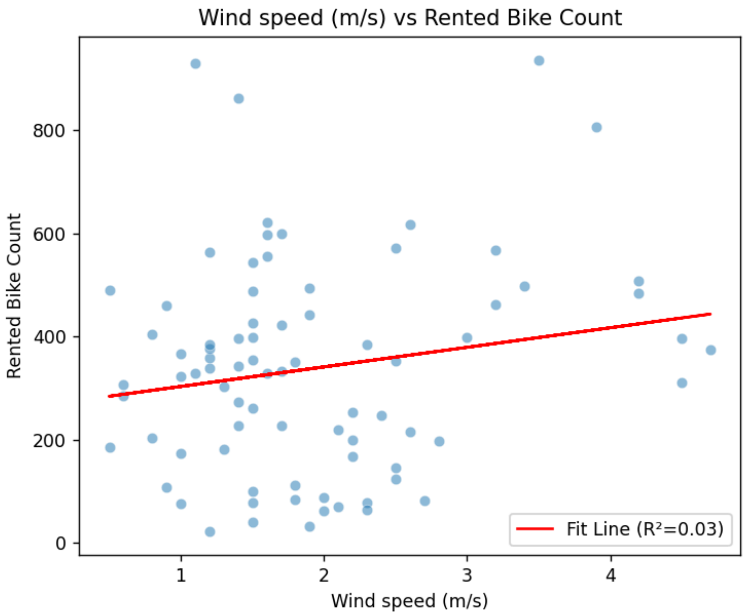
Among all the other variables, Hour has displayed the highest r2 score and can be said to have the strongest correlation with rental demand. The points in the graph indicate that demand rises as morning approaches and peaks during evening hours before declining towards nighttime.



The solar radiation has been adjusted to only show daytime results. It can be observed that demand rises as daylight begins and peaks when the temperature is mild but begins to fall as conditions become hot. It can be concluded that the relationship is not linear and solar radiation does not play a big role in determining demand.



The fit line indicates that humidity has a negative relationship with rental demand which could be explained by increasing humidity providing unfavorable conditions for biking hence, a decline in demand. Overall, the relationship is not linear and humidity is a weak indicator of rental demand.



The data points in this graph seem to be scattered randomly, not showing any type of trend. This suggests that wind speed has a minimal to zero impact on determining rental demand, and the relationship is not linear.

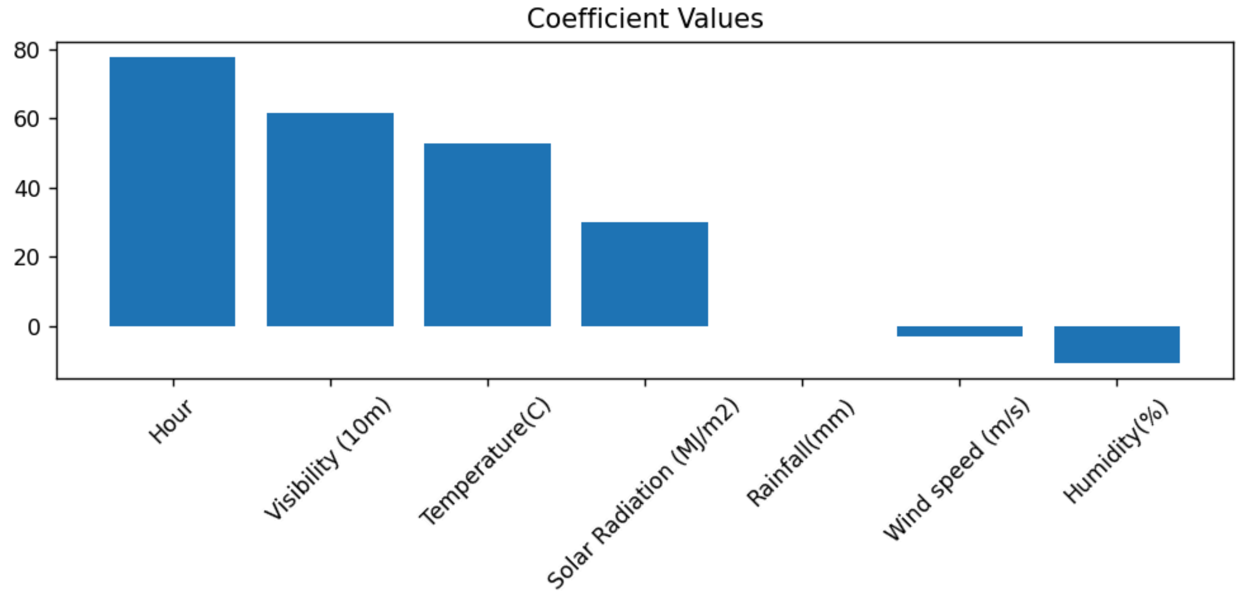
All in all, the graphs have shown us that hour has the greatest relationship with rental demand but also that there is no single indicator for rental demand and that most of the variables do not have linear characteristics. However, this could suggest that rental demand is dependent on a variety of factors occurring in that specific date.

Multiple Regression Analysis

The multiple linear regression model was developed to predict hourly bike rental demand with the previously mentioned attributes as the independent variables which included Hour, Temperature, Humidity, Wind Speed, Solar Radiation, Visibility, and Rainfall, while the dependent variable acted as the Rented Bike Count. The dataset was split into training and testing portions in a 3:1 ratio, and the input features were standardized using the StandardScaler so that the data will have a mean of 0 and a standard deviation of 1 to ensure a consistent scale across variables.

The performance evaluation showed a Mean Absolute Error (MAE) of 95.10, indicating that on average the model's predictions deviated from the actual rental count by around 95 bikes per hour. The R^2 value of 0.50 showed that approximately half of the variation in rental demand could be explained by the selected features. Additionally, the average percentage error of 35.32%, a moderate level of prediction accuracy. These results indicate that the model may be useful for understanding general demand patterns influenced by time and weather conditions but still requires improvements to be able to have better prediction ability of rental demands.

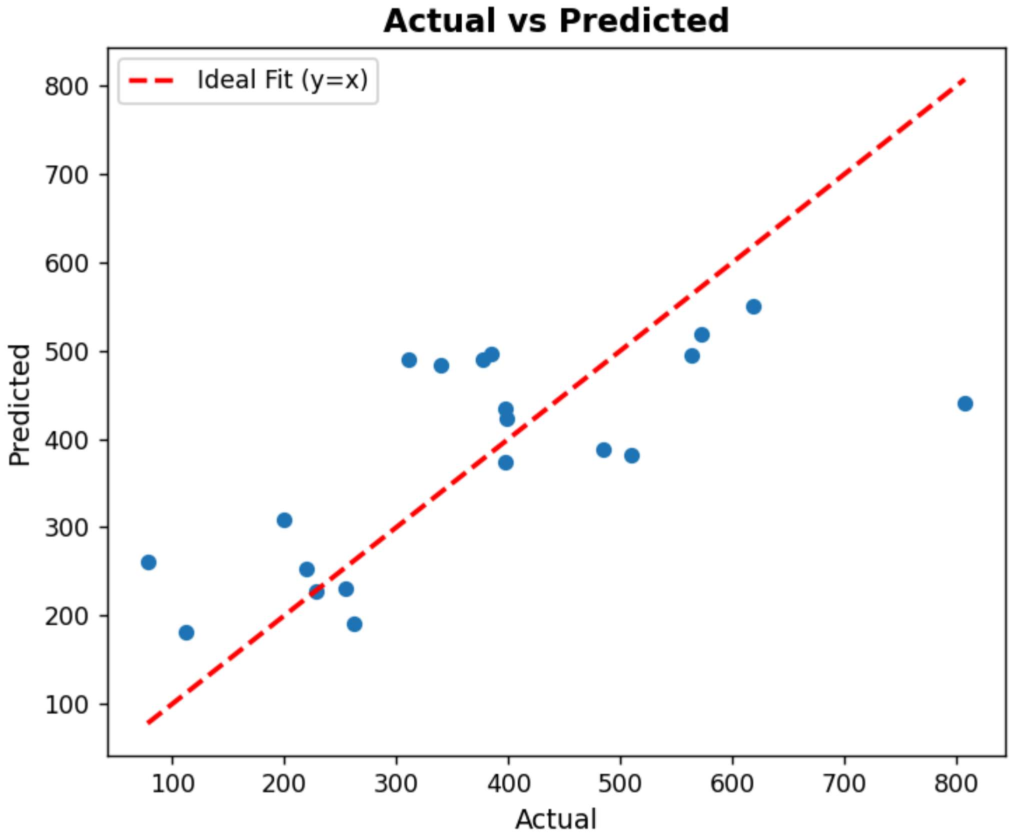
Coefficient Analysis

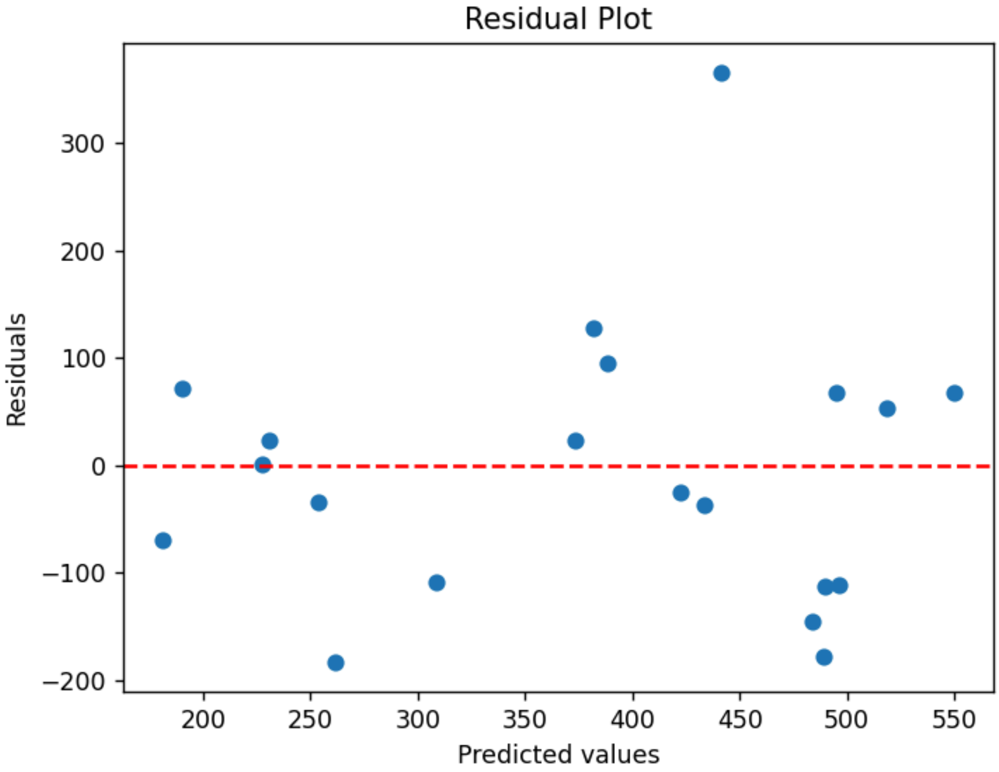


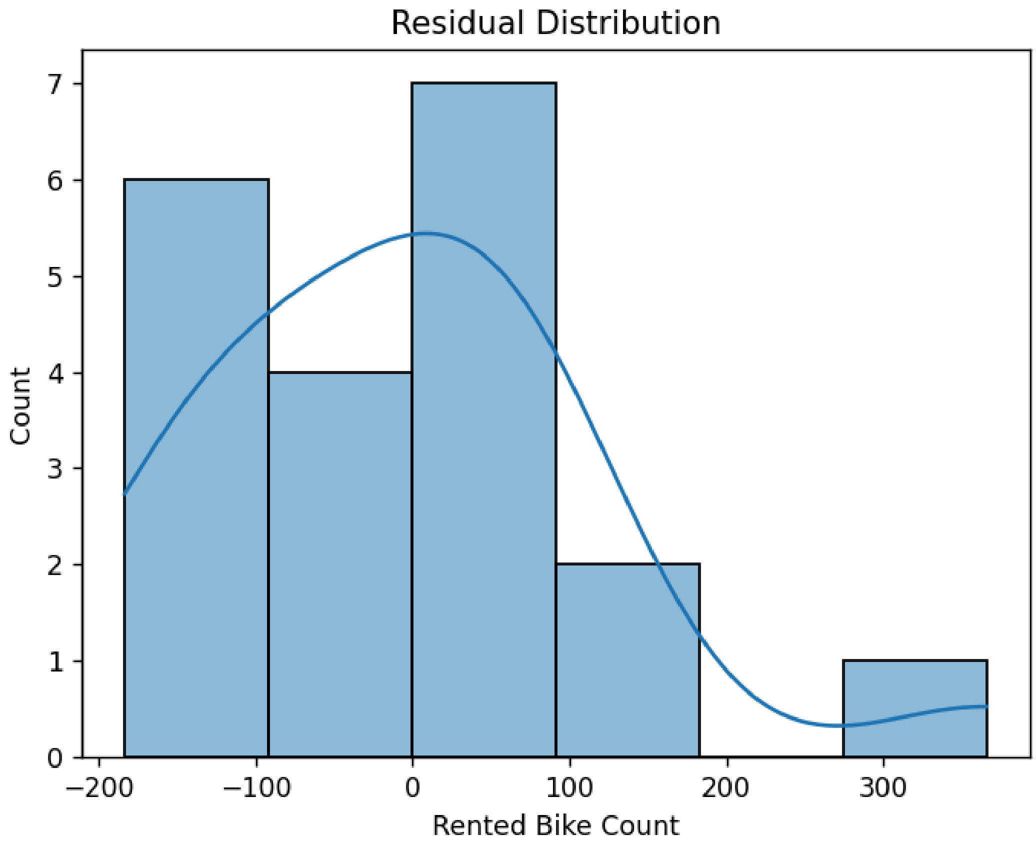
The coefficient values indicate the relative influence and direction of each variable on rental demand. Hour has the strongest positive effect, showing that bike rentals rise substantially during later hours of the day. Visibility and temperature also show notable positive contributions, meaning clearer conditions and warmer weather are associated with higher rental counts. Solar radiation contributes positively as well, indicating that daylight intensity aligns with increased usage. In contrast, humidity and wind speed show small negative effects, suggesting that uncomfortable or windy conditions slightly reduce demand. Rainfall shows effectively no influence in this dataset, implying that either rainfall occurrences were too limited or rental behavior did not vary meaningfully during recorded rainy conditions

Validation

The residual diagnostic graphs produced indicate that the model generally adheres to linear regression assumptions. The Actual vs Predicted plot shows points aligning near the identity line, indicating that the model captures broad variation in rental demand. However, the residual plot displays no visible pattern which means that there is no clear heteroscedasticity or omitted nonlinear structure. Moreover, the residual distribution is centered near zero but not perfectly symmetric. The tallest concentration is near zero, and the second tallest on the far-left side with a short right tail indicates mild positive skew. This suggests that there is some underestimation of higher rental counts, but not to the extent that it invalidates the model.


Actual vs Predicted
Ideal Fit (y=x)
Predicted
800
700
600
500
400
300
200
100
Actual
100
200
300
400
500
600
700
800


Residual Plot
Residuals
300
200
100
0
−100
−200
Predicted values
200
250
300
350
400
450
500
550



Conclusion & Limitations

The analysis shows that bike rental demand is influenced by multiple factors rather than a single dominant variable with the hour of the day having the strongest effect, reflecting typical commuting or leisure behavior patterns. Furthermore, it can be suggested that the presence of warmer temperatures, higher visibility, and greater solar radiation can be associated with increased rental counts due to favorable weather conditions encouraging bike usage. In contrast, humidity and wind speed show slight negative effects, suggesting that environmental discomfort reduces demand, though their influence is comparatively small.

The multiple linear regression model achieved an R^2 of approximately 0.50, meaning it explains about half of the variation in rental demand. This value may be suitable for general understanding of demand trends but is still lacking for actual forecasting. Lastly, the model diagnostics show largely random residuals centered near zero, supporting the appropriateness of the linear model structure.

It is important to note that the dataset has important limitations that affect the model's accuracy and ability. For example, the data is heavily concentrated in the winter season and contains no holiday variation, limiting the model's ability to capture seasonal or situational demand patterns which explains rainfall's rare occurrence and minimal influence to rental demand. After cleaning, the dataset size is also small, only being around 80 which reduces the reliability of coefficient estimates.

Future improvements that should be considered should begin by expanding the dataset to include multiple seasons like Spring, Summer and Autumn and to cover a greater range of weather conditions like Rain or Sunny. This would greatly help the model to have access to possibly more variables with strong correlations which have been excluded in this case like Season and Holidays. Finally, there may be other modelling approaches that should be explored as they could be more effective in explaining the key factors behind rental demand and forecasting it such as polynomial regression or tree-based algorithms.