TU Dortmund

Introductory Case Studies

Project 3: Regression Analysis of Bike Rentals

Lecturers:
Prof. Dr. Crystal Wiedner
Prof. Dr. Philipp Doebler
Prof. Dr. Rouven Michel
Prof. Dr. Marlies Hafer

Author: Siddhartha Karki (Matr. No : 238092)

Group number: 4

Group members:
Sagar Basnet

December 15, 2024

Contents

1 Introduction

Bike-sharing systems gain widespread popularity as an environmentally friendly and efficient mode of urban transportation. Managing these systems requires accurate predictions of bike demand to ensure optimal resource allocation and user satisfaction. This project explores the factors influencing daily bike rentals in a bike-sharing system by applying regression analysis techniques. The primary objective is to construct a predictive model for bike demand based on weather conditions, temperature, humidity, and temporal factors.

In this project, the primary objective is to identify key predictors of daily bike rentals and develop a robust predictive model. Using a dataset of 731 observations, descriptive analyses are performed to identify patterns and relationships between variables. A correlation matrix and heatmap examine the relationships among continuous variables, including temp, atemp, hum, windspeed, and cnt. Scatter plots are subsequently employed to visualize specific pairwise relationships. For categorical variables such as month, weekday, workingday, and weathersit, box plots are utilized to assess their associations with the dependent variable (cnt). These visualizations provide valuable insights into how different variables influence bike rentals.

A linear regression model is initially developed using all predictors, and polynomial terms are included to capture non-linear effects, particularly for the temperature variable. The backward selection method, guided by Akaike Information Criterion (AIC) values, refines the predictor set and improves model accuracy. To ensure the reliability of the final model, diagnostic evaluations are conducted, including residual analysis to check for linearity and heteroscedasticity, and multicollinearity assessments using variance inflation factors (VIF).

The report is organized as follows: Section 2 provides a comprehensive description of the dataset and its structure. Section 3 details the statistical methods and correction techniques applied in the analysis. Section 4 presents the results and offers an in-depth discussion of their interpretation. Finally, Section 5 concludes the report by summarizing the key findings, highlighting their implications for regression analysis, and suggesting potential avenues for future research.

2 Problem Statement

2.1 Data Set and Data Quality

This project examines a dataset from a bike-sharing system, obtained from the UCI Machine Learning Repository. The dataset comprises 731 daily records and includes nine variables: one dependent variable, the total daily bike rentals, and eight independent variables. The details about the variables are tabulated below:

Table 1: Characteristics of Variables in the Dataset

Variables	**Data Type**	**Description**
cnt	Numeric	Number of bike rentals on the corresponding day. This is the primary dependent variable representing the total count of bike shares.
mnth	Categorical	Month of the year (1 = January, ..., 12 = December). It represents the time period within the year of the rental data.
weekday	Categorical	Day of the week (0 = Sunday, ..., 6 = Saturday). This indicates the specific day within the week.
workingday	Categorical	Dummy variable: 1 if the day is neither a weekend nor a holiday, otherwise 0. It highlights whether the day is a typical workday.
weathersit	Categorical	Categorical variable representing the weather situation (1 = Clear, ..., 4 = Heavy Rain/Snow). It reflects the impact of weather on bike rentals.
temp	Numeric	Temperature in Celsius, normalized between 0 and 1. It represents the measured atmospheric temperature.
atemp	Numeric	Perceived or "feeling" temperature in Celsius, normalized between 0 and 1. This adjusts for human perception of temperature.
hum	Numeric	Humidity level, normalized between 0 and 1. This indicates the moisture content in the air.
windspeed	Numeric	Windspeed, standardized between 0 and 1. It measures the velocity of wind on the given day.

The dataset exhibits high data quality, with no missing or invalid entries identified. Continuous variables such as temp, atemp, hum, and windspeed are normalized or standardized, ensuring comparability across different scales. Temporal variables (month,

weekday) are appropriately encoded as integers, while categorical data (weathersit, workingday) are stored in easily interpretable formats.

The project's primary goal is to understand the relationship between the number of bike rentals and influencing factors using regression techniques. Specific objectives include exploring relationships between variables through descriptive analysis, developing a regression model that accounts for potential nonlinearities in temperature, selecting the best subset of predictors using backward selection and AIC, and evaluating the model using diagnostic tools such as residual plots and variance inflation factors. The findings aim to provide actionable insights for improving the management and planning of bike-sharing services.

3 Statistical Methods

3.1 Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable y and one or more independent variables $x_1, x_2, \ldots, x_k$. The objective of linear regression is to find the linear equation that best fits the data, enabling prediction of the dependent variable based on the values of the independent variables (Seber and Lee, 2012).

3.1.1 Model and Assumptions

Let y represent the dependent variable, and $x_1, x_2, \ldots, x_k$ represent the independent variables (predictors). The linear regression model assumes that the relationship between the dependent and independent variables is linear and can be described by the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

Where:

- β_0 is the intercept,
- $\beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients associated with the independent variables $x_1, x_2, \ldots, x_k$,

- ϵ is the error term, representing random noise or unobserved factors.

The error term ϵ is assumed to follow a normal distribution with a mean of zero and constant variance σ^2, i.e.,

$$\epsilon \sim N(0, \sigma^2)$$

The coefficients $\beta_0, \beta_1, \ldots, \beta_k$ are estimated using Ordinary Least Squares (OLS). The method aims to minimize the sum of squared residuals:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Where $\hat{y}_i$ is the predicted value for the i-th observation and is given by:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

The estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are computed by solving the normal equation:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Where:

- X is the matrix of independent variables (including a column of ones for the intercept),
- y is the vector of observed dependent values,
- $\hat{\beta}$ is the vector of estimated coefficients.

The assumptions of linear regression models are:

1. Linearity: The relationship between the dependent variable y and the independent variables $x_1, x_2, \ldots, x_k$ is linear. The model assumes the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

2. Independence of Errors: The error terms $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are assumed to be independent of each other, meaning no correlation exists between the errors of different observations.

3. Homoscedasticity: The variance of the error terms is constant across all values of the independent variables. This means the spread of the residuals (errors) is uniform across the range of predicted values:

$$\text{Var}(\epsilon_i) = \sigma^2 \quad \text{for all} \quad i$$

4. Normality of Errors: The error terms are assumed to be normally distributed:

$$\epsilon \sim N(0, \sigma^2)$$

3.2 Dummy Encoding

In regression analysis, categorical variables cannot be directly included in their original form. To address this, dummy encoding is commonly used, transforming a categorical variable with k categories into $k-1$ binary (dummy) variables. Each dummy variable takes the value 1 if an observation belongs to a specific category and 0 otherwise (Garavaglia and Sharma, 1998).

For a categorical variable z with k categories ($z \in \{A_1, A_2, \ldots, A_k\}$), the dummy variables $z_1, z_2, \ldots, z_{k-1}$ are defined as:

$$z_m = \begin{cases} 1 & \text{if } z = A_m, \\ 0 & \text{otherwise,} \end{cases}$$

where $m = 1, 2, \ldots, k-1$. One category, A_k, is omitted and serves as the *reference category* (typically the most frequent or contextually significant). In the regression model, these dummy variables enable comparisons of each category with the reference group. The model for an outcome y can be expressed as:

$$y_i = \alpha_0 + \alpha_1 z_{i1} + \alpha_2 z_{i2} + \cdots + \alpha_{k-1} z_{i,k-1} + \varepsilon_i,$$

where α_0 is the intercept, α_m are the coefficients of the dummy variables, and ε_i is the error term. The equation illustrates how each category (through its dummy variable) affects the outcome relative to the reference category. By using dummy variables, the regression model can quantify the impact of categorical factors such as region, season, or product type on the dependent variable.

3.3 Goodness of Fit

The goodness of fit measures how well the regression model explains the variability in the dependent variable. It evaluates the model's ability to predict outcomes and provides insights into its effectiveness.The most commonly used measure of goodness of fit is the coefficient of determination (R^2), adjusted R^2 (Schermelleh-Engel et al., 2003).

3.3.1 R-Square and Adjusted R-Square

The most commonly used measure of goodness of fit is the coefficient of determination (R^2), which quantifies the proportion of the variance in the dependent variable that is explained by the independent variables. An R^2 value closer to 1 indicates a good fit, while a value closer to 0 suggests that the model does not explain much of the variability (Schermelleh-Engel et al., 2003). The mathematical expression for R^2 is:

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}$$

where $\hat{y}_i$ denotes the predicted values, y_i are the actual observed values, and $\bar{y}$ is the mean of the observed data.

However, R^2 can be misleading, particularly when comparing models with different numbers of predictors, as it tends to increase with the addition of variables, even if they do not meaningfully improve the model. To address this, the adjusted R^2 modifies R^2 by introducing a penalty for the number of predictors, making it more robust for model comparison. Its formula is:

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p}$$

Here, n is the total number of observations, and p is the number of predictors in the model, including the intercept. Adjusted R^2 accounts for the complexity of the model and only increases if the added predictors enhance the model's explanatory power. By balancing the model's fit and complexity, adjusted R^2 provides a more reliable criterion for selecting between competing models.

3.4 Hypothesis Testing

Hypothesis testing is used to determine the statistical significance of predictors in a linear regression model, ensuring that only variables with meaningful contributions to the response variable are included. The null hypothesis (H_0) states that a predictor's regression coefficient (β_j) is zero, implying no impact on the response, while the alternative hypothesis (H_1) proposes that $\beta_j \neq 0$. A t-test is commonly employed for this purpose, with the test statistic calculated as

$$t_j = \frac{\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}}.$$

This statistic follows a t-distribution with $n - p$ degrees of freedom, where n represents the number of observations, and p denotes the total number of predictors.

The null hypothesis is rejected if the absolute value of the test statistic exceeds the critical value derived from the t-distribution at a specified significance level (α), typically 0.05. Rejecting H_0 indicates that the predictor has a significant effect on the response variable. This method helps in constructing reliable regression models by retaining only the predictors that have a substantial influence, thereby minimizing the risk of overfitting and excluding irrelevant variables (Newey and McFadden, 1994).

3.5 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a model selection criterion that balances goodness of fit and model complexity by penalizing models with more parameters. It is based on the concept of likelihood, which reflects how well a model explains the observed data (Chakrabarti and Ghosh, 2011). The general formula for AIC is:

$$AIC = -2 \cdot \log(L) + 2k$$

where L is the likelihood of the model (representing how well the model fits the data) and k is the number of parameters estimated in the model, including intercepts and coefficients.

The term $-2\log(L)$ rewards the model for its goodness of fit, while the term $2k$ penalizes models that have more parameters, thus reducing the likelihood of overfitting. The model

with the lowest AIC value is generally preferred, as it suggests a good trade-off between model complexity and goodness of fit.

To identify a suitable subset of explanatory variables for predicting bike rental counts (`cnt`), backward selection was employed. This method starts with a full model and iteratively removes the least contributing variables based on the Akaike Information Criterion (AIC), balancing model complexity and goodness of fit.

The initial model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon,$$

where Y is the dependent variable (`cnt`), $X_1, X_2, \ldots, X_k$ are explanatory variables, $\beta_0, \beta_1, \ldots, \beta_k$ are coefficients estimated by least squares, and $\epsilon \sim N(0, \sigma^2)$ is the error term.

The AIC, defined as

$$\text{AIC} = -2 \cdot \log(L) + 2k,$$

evaluates models, where L is the likelihood and k is the number of parameters. Lower AIC values indicate better models. The algorithm iteratively removes variables with the least contribution (highest p-value) until no further AIC improvement occurs.

3.6 Backward Selection

To identify a suitable subset of explanatory variables for predicting bike rental counts (`cnt`), backward selection was employed. This method starts with a full model and iteratively removes the least contributing variables based on the Akaike Information Criterion (AIC), balancing model complexity and goodness of fit (Akaike, 1998).

The initial model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon,$$

where Y is the dependent variable (`cnt`), $X_1, X_2, \ldots, X_k$ are explanatory variables, $\beta_0, \beta_1, \ldots, \beta_k$ are coefficients estimated by least squares, and $\epsilon \sim N(0, \sigma^2)$ is the error term.

The AIC, defined as

$$\text{AIC} = -2 \cdot \log(L) + 2k,$$

evaluates models, where L is the likelihood and k is the number of parameters. Lower AIC values indicate better models. The algorithm iteratively removes variables with the least contribution (highest p-value) until no further AIC improvement occurs.

4 Statistical Analysis

In this section, the results derived from the application of the methods outlined in the previous sections are thoroughly analyzed and discussed. The aim is to evaluate the findings in the context of the stated objectives and ensure they address the requirements of the provided tasks effectively.

4.1 Variables Relationship Analysis

To understand the relationships between the variables, the dependent variable Count (Rentals) is examined first. From Figure 1, it is evident that the average number of

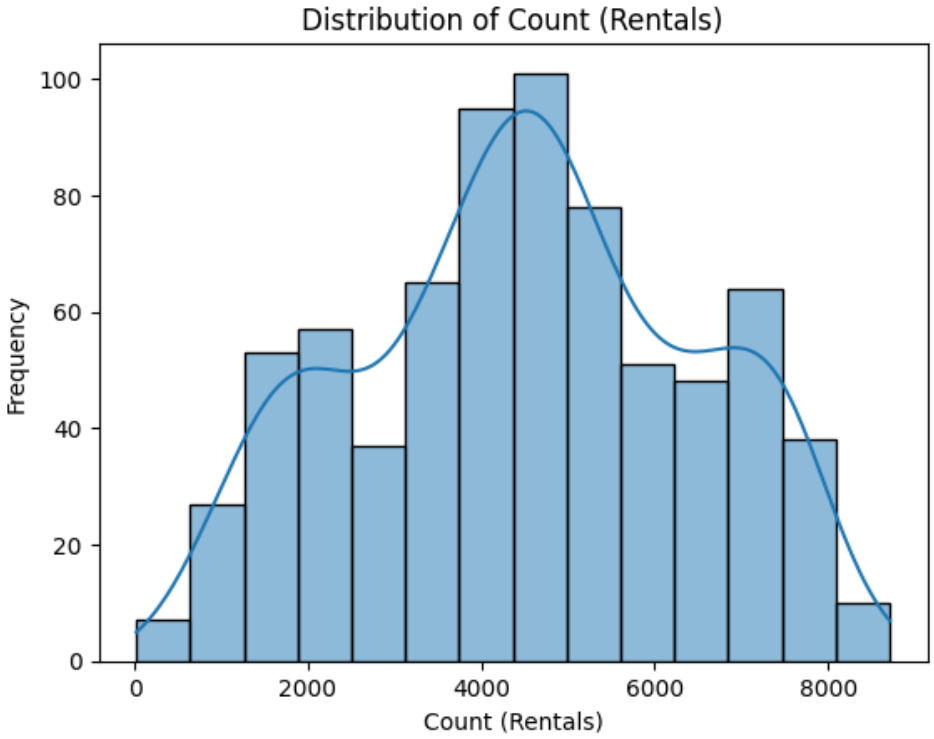


Figure 1: Distribution of Rentals

bike rentals is 4,504.34 across 731 days. The maximum and minimum rentals were

recorded at 8,714 and 22, respectively. The distribution of Count (Rentals) appears approximately symmetric and bell-shaped, indicating a near-normal distribution. This observation is further supported by the skewness value of -0.047, suggesting that the variable is approximately symmetric with only a slight left skew.

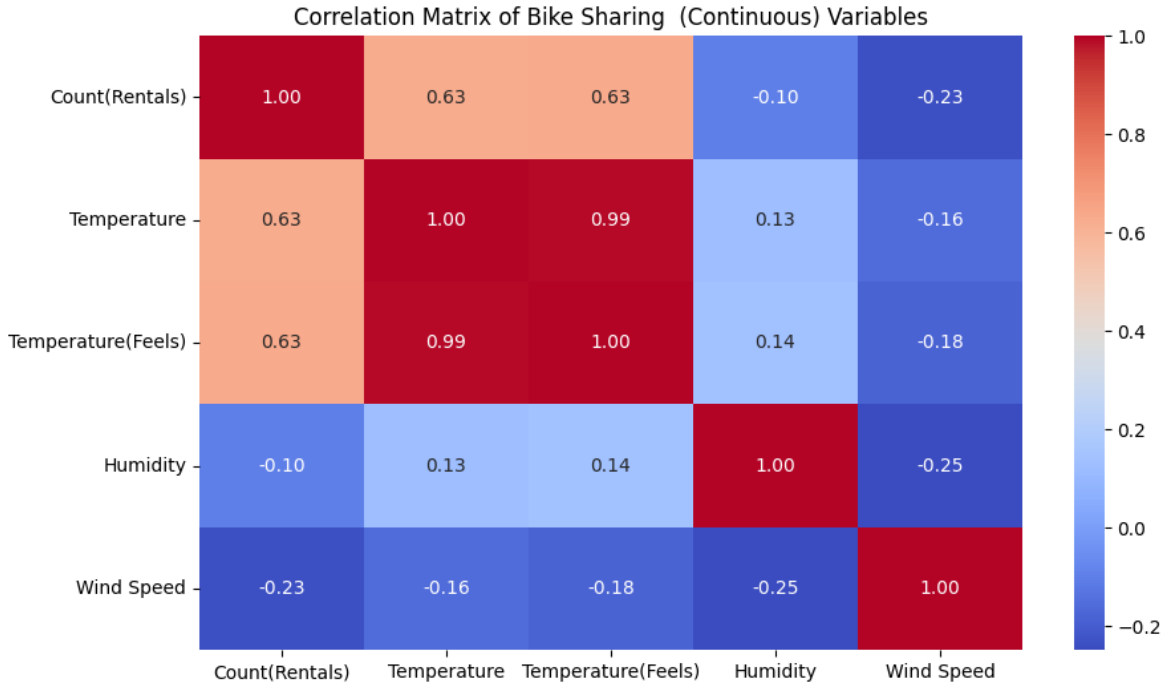


Figure 2: Heatmap showing correlation between Count(Rentals), Temperature, Temperature(Feels), Humidity, Wind Speed

The Figure 2 depicts the correlation between the variables Count (Rentals), Temperature, Feels-like Temperature, Humidity, and Wind Speed. The correlation matrix provides insights into the general relationships among these variables. We are particularly interested in examining the relationship between the dependent variable Count (Rentals) and other independent variables. The matrix reveals that Temperature and Feels-like Temperature have a strong positive correlation (0.63) with Count (Rentals). This suggests that warmer temperatures are associated with higher bike rental counts. In contrast, Humidity and Wind Speed show weak negative correlations (-0.10 and -0.23, respectively), indicating that higher humidity and wind speeds may slightly reduce bike rentals.

Based on the boxplots in Figures 5, 6, and 7 in the Appendix, the relationships between categorical variables (months, weekdays, and working days) and bike rentals show notable trends. Bike rentals are highest in the warmer months, particularly July and

August, with counts exceeding 7,000, while colder months like January and February see the lowest counts, around 2,000 to 3,000, reflecting the influence of seasonal temperature variations. Rentals across weekdays remain fairly consistent, with slight increases on Wednesday and Friday (over 5,500 rentals), while Sunday shows reduced activity, averaging about 4,500, possibly due to fewer commutes. Rentals on working days are slightly higher than on non-working days, averaging around 5,000 and 4,800 rentals, respectively. Despite this, variability is similar across both categories, with counts generally ranging between 3,000 to 7,000 and occasional spikes beyond 8,000, suggesting a balance of commuting and leisure usage.

4.2 Modeling Nonlinear Relationship and Predictor Variable Selection

To model the relationship between the dependent variable Count (Rentals) and the other independent variables, a linear regression model was initially considered. However, upon examining the relationship between the response variable and temperature, it became evident that a non-linear relationship exists, as shown in Figure 8 (scatter plot in Appendix). To capture this non-linearity, a squared term for temperature (temp_sq) was introduced, representing the quadratic relationship between temperature and bike rentals. The decision to use a quadratic term, rather than a cubic or higher power, was based on the fact that quadratic relationships often provide sufficient flexibility to capture non-linearity without overfitting the model, especially when the dataset is small.

Categorical variables such as mnth (month), weathersit (weather situation), and weekday were converted into dummy variables. This transformation allows for the inclusion of these categorical variables in the regression model, enabling the analysis of their impact on bike rentals. By using dummy variables, we can examine how bike rentals vary across different months, weather conditions, and weekdays.The final model includes the quadratic term for temperature (temp_sq) and the dummy variables for the categorical factors, capturing the relationship between bike rentals and various predictors.

4.3 Best Subset Selection

To identify the most suitable subset of explanatory variables for predicting bike rentals (cnt), the backward selection method is employed. This approach is particularly benefi-

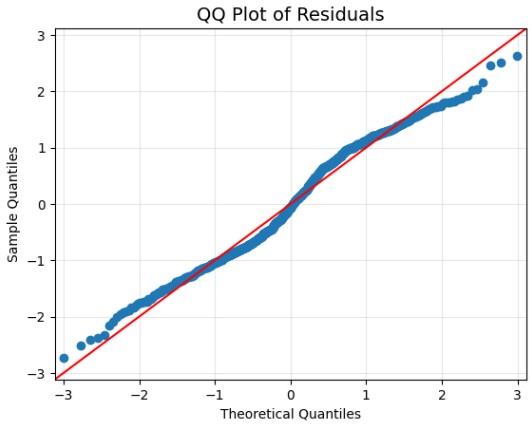
cial for datasets with a moderate number of predictors, as it starts with the full model and systematically eliminates the least significant variables, reducing the risk of omitting relevant predictors prematurely. The Akaike Information Criterion (AIC) was used as the guiding criterion for model selection because it effectively balances model fit and complexity by penalizing unnecessary predictors. AIC was favored over alternative criteria as it prioritizes predictive accuracy over model parsimony, aligning well with the objectives of this analysis.

The backward elimination method led to the selection of the following significant predictors: workingday, temp, temp_sq, hum, windspeed, weekday_6, weathersit_2, weathersit_3, and several months (mnth_2, mnth_3, mnth_9, mnth_10). The resulting regression model has an AIC of 12486.372, indicating a good balance between model fit and complexity. From the Table 3 in Appendix, it is clear that The R-squared value of 0.606 shows that approximately 60.6% of the variation in bike rentals is explained by the selected predictors. Key predictors include workingday (coefficient = 363.66, p-value = 0.003), which positively affects bike rentals, and temp (coefficient = 23580.45, p-value $<$ 0.001), which has a strong positive impact. Temp_sq (coefficient = -17830.07, p-value $<$ 0.001) indicates a non-linear relationship, with higher temperatures initially increasing rentals, but diminishing at extreme values. Hum (coefficient = -3394.14, p-value $<$ 0.001) and windspeed (coefficient = -4423.28, p-value $<$ 0.001) negatively influence rentals. Additionally, weekday_6 (coefficient = 478.70, p-value = 0.003) and weather conditions (weathersit_2: coefficient = -254.63, p-value = 0.036; weathersit_3: coefficient = -1937.72, p-value $<$ 0.001) significantly impact rentals. The confidence intervals indicate precise estimates, and the low p-values confirm the statistical significance of these predictors.

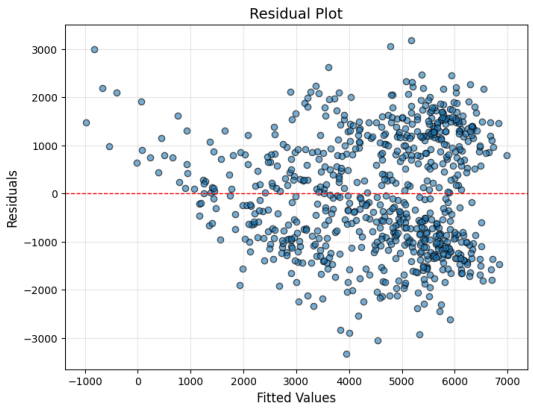
4.4 Model Evaluation and Multicollinearity Check

Once the model is selected using backward selection and the AIC value, in this section, we checked for patterns of linearity, heteroskedasticity, and normality using the residual and QQ plots, and also examined multicollinearity using the variance inflation factor (VIF).

From the QQ plot in Figure 3(a), it appears that the points align closely with the diagonal line, indicating that the residuals approximately follow a normal distribution.



(a) QQ-plot for selected model



(b) Residual Plot for selected model

Figure 3: QQ Plot and Residual Plot for Model Evaluation

However, slight deviations at the tails may suggest minor departures from normality, though these are not severe enough to undermine the model.

Similarly, from the residual plot in Figure 3(b), the residuals are scattered randomly around the horizontal line at zero, suggesting that the assumption of linearity is fulfilled. However, the spread of residuals appears slightly uneven across the range of fitted values, with more variability observed at the extremes, which shows potential heteroscedasticity. The Table 2 provides further insights into multicollinearity among

Feature	**VIF**
const	89.966433
workingday	1.595405
temp	41.922899
temp_sq	42.445530
hum	1.944287
windspeed	1.184429
weekday_6	1.581027
weathersit_2	1.595972
weathersit_3	1.299028
mnth_2	1.211902
mnth_3	1.173800
mnth_9	1.132480
mnth_10	1.194533
mnth_11	1.202954

Table 2: Variance Inflation Factor (VIF) for predictors

predictors. Most predictors exhibit acceptable VIF values, with the majority below 5, in-

dicating low multicollinearity. Notably, the quadratic term for temperature (temp_sq) and the temperature variable itself have higher VIF values of 42.4 and 41.9, respectively, because temp_sq is derived directly from temp. This type of multicollinearity is expected and acceptable in models using polynomial terms, as its purpose is to improve the model's fit by capturing non-linear effects.

4.5 Final Model Formulation and Evaluation

The final regression model is selected via AIC-guided backward elimination method. The AIC score is calculated for all subsets of co-variates, and the results of the top 3 models are listed in Table 4 in the Appendix section. The final regression model looks like:

$$\text{cnt} = \beta_0 + \beta_1 \cdot \text{workingday} + \beta_2 \cdot \text{temp} + \beta_3 \cdot \text{temp\_sq} + \beta_4 \cdot \text{hum} + \beta_5 \cdot \text{windspeed} + \beta_6 \cdot \text{weekday\_6} +$$

$$\beta_7 \cdot \text{weathersit\_2} + \beta_8 \cdot \text{weathersit\_3} + \beta_9 \cdot \text{mnth\_2} + \beta_{10} \cdot \text{mnth\_3} + \beta_{11} \cdot \text{mnth\_9} + \beta_{12} \cdot \text{mnth\_10}$$

While most predictors show linear relationships, there are signs that the model's assumption of linearity may not fully hold for variables like temperature, particularly with the quadratic term temp_sq. The residual plot suggests that as the fitted values increase, the spread of residuals becomes uneven, indicating potential heteroscedasticity. This suggests that the model may be less reliable at higher levels of bike rentals, where predictions could become less accurate due to increased variability.

In terms of multicollinearity, the Variance Inflation Factor (VIF) analysis reveals that most features have acceptable VIF values, well below the threshold of 5. However, both temp and temp_sq exhibit much higher VIF values (42.4 and 41.9, respectively), which is expected due to the polynomial relationship between temperature and bike rentals. This indicates a multicollinearity issue between these variables, but it is considered acceptable in the context of the model's goal to capture non-linear effects. Despite these challenges, the model is robust and performs well, but addressing heteroscedasticity and potential outliers could further improve its predictive power.

5 Summary

This project aimed to develop a regression model to predict bike rentals (`cnt`) based on key variables, including weather conditions, temperature, and day of the week. Descriptive analysis showed that the average number of bike rentals was 4,504.34, with temperature exhibiting a strong positive correlation (0.63) with rentals. As temperature increases, bike rentals tend to rise. Weather conditions also had significant effects on bike usage, while humidity and windspeed showed weaker negative correlations.

To capture the non-linear relationship between temperature and bike rentals, a quadratic term for temperature (`temp_sq`) was introduced. This allowed the model to account for the fact that bike rentals increase with temperature up to a certain threshold, beyond which they may plateau or decrease. The final regression model, selected via backward elimination guided by AIC, included significant variables such as working day (`workingday`), temperature (`temp`), squared temperature (`temp_sq`), humidity (`hum`), windspeed (`windspeed`), weekday 6 (`weekday_6`), weather conditions 2 and 3 (`weathersit_2`, `weathersit_3`), and months 2, 3, 9, and 10 (`mnth_2`, `mnth_3`, `mnth_9`, `mnth_10`). The model's R-squared value of 0.606 indicated that 60.6% of the variance in bike rentals was explained by these variables. Temperature, both linearly and quadratically, played a central role in influencing demand, with rentals increasing as temperature rose, up to a point.

Despite explaining a significant portion of the variance, the model evaluation revealed some issues. Residual plots suggested potential heteroscedasticity, meaning the model's accuracy may be less reliable for extreme rental values. This suggests that the model could be improved, particularly for days with unusually high or low rental counts. Multicollinearity between temperature and its squared term was observed, which was expected due to the non-linear relationship. However, it did not significantly affect the model's overall performance.

For future improvements, exploring alternative techniques like weighted least squares regression could address heteroscedasticity. Additionally, regularization methods such as Lasso or Ridge regression could help manage multicollinearity and improve model generalizability. Further analysis of factors like local events, public holidays, or demographic trends could also provide deeper insights and refine the model for better predictive capabilities.

Bibliography

Akaike, H. (1998). *A New Look at the Statistical Model Identification*, pages 215–222. Springer New York, New York, NY.

Chakrabarti, A. and Ghosh, J. K. (2011). Aic, bic and recent advances in model selection. *Philosophy of statistics*, pages 583–605.

Garavaglia, S. and Sharma, A. (1998). A smart guide to dummy variables: Four applications and a macro. In *Proceedings of the northeast SAS users group conference*, volume 43.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Schermelleh-Engel, K., Moosbrugger, H., Müller, H., et al. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online*, 8(2):23–74.

Seber, G. A. and Lee, A. J. (2012). *Linear regression analysis*. John Wiley & Sons.

Appendix

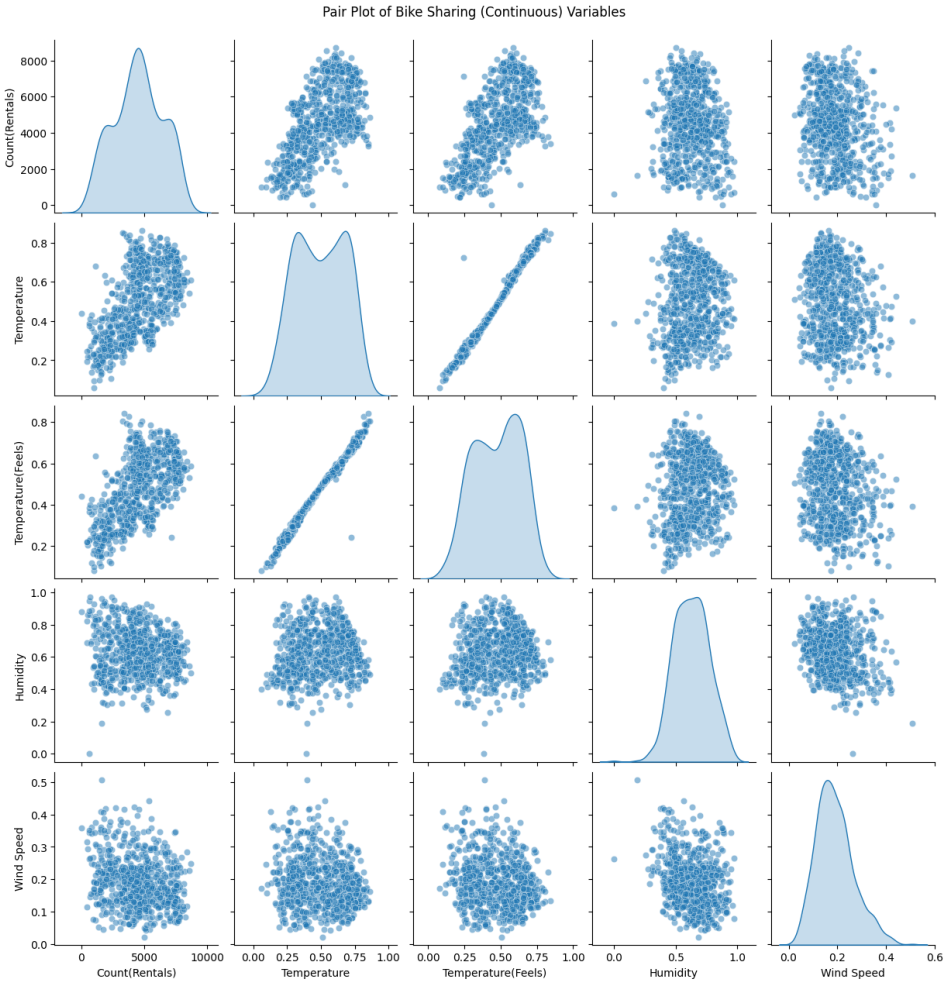


Figure 4: Pairplot showing correlation between Count(Rentals), Temperature, Temperature(Feels), Humidity, Wind Speed

Table 3: Regression Results Table

Variable	**Coef**	**Std Err**	**t**	**P> |t|**	**[0.025]**	**[0.975]**
const	539.8762	430.394	1.254	0.210	-305.106	1384.858
workingday	363.6569	123.279	2.950	0.003	121.627	605.687
temp	23580.000	1606.114	14.682	0.000	20400.000	26700.000
temp_sq	-17830.000	1623.115	-10.985	0.000	-21000.000	-14600.000
hum	-3394.135	444.533	-7.635	0.000	-4266.878	-2521.393
windspeed	-4423.279	637.657	-6.937	0.000	-5675.178	-3171.380
weekday_6	478.6963	162.679	2.943	0.003	159.313	798.080
weathersit_2	-254.6270	121.195	-2.101	0.036	-492.566	-16.688
weathersit_3	-1937.7227	309.609	-6.259	0.000	-2545.571	-1329.874
mnth_2	-536.5799	186.298	-2.880	0.004	-902.336	-170.824
mnth_3	-335.0068	176.454	-1.899	0.058	-681.435	11.421
mnth_9	577.5532	175.922	3.283	0.001	232.168	922.938
mnth_10	613.4052	178.005	3.446	0.001	263.931	962.879
mnth_11	292.7606	181.314	1.615	0.107	-63.209	648.730

Table 4: Top 3 Models Based on AIC

Model	**AIC**	**Selected Features**
1	12486.372	const, workingday, temp, temp_sq, hum, windspeed, weekday_6, weathersit_2, weathersit_3, mnth_2, mnth_3, mnth_9, mnth_10, mnth_11
2	12486.622	const, workingday, temp, temp_sq, hum, windspeed, weekday_1, weekday_6, weathersit_2, weathersit_3, mnth_2, mnth_3, mnth_9, mnth_10, mnth_11
3	12487.011	const, workingday, temp, atemp, temp_sq, hum, windspeed, weekday_1, weekday_6, weathersit_2, weathersit_3, mnth_2, mnth_3, mnth_9, mnth_10, mnth_11

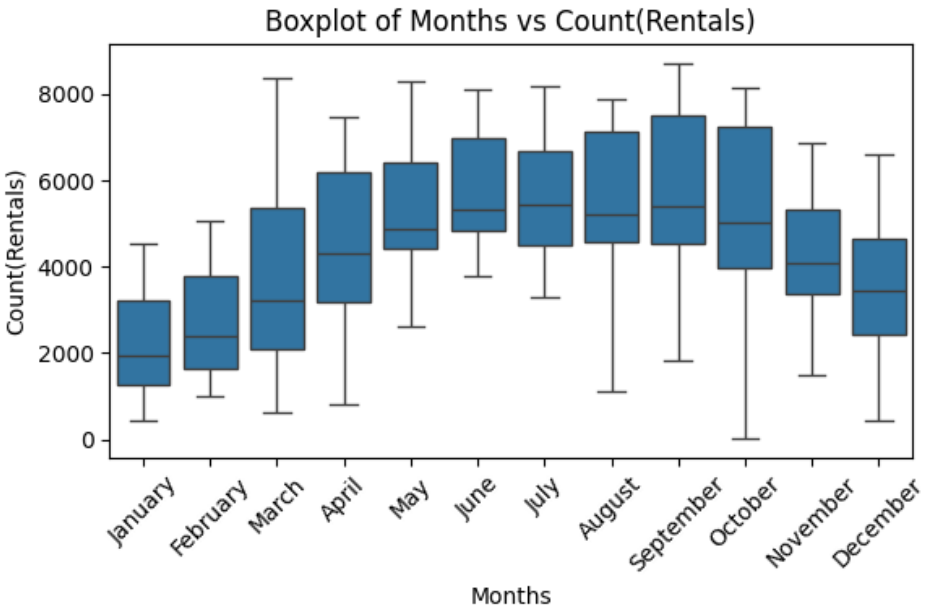


Figure 5: Box plot for Months vs Count(rentals)

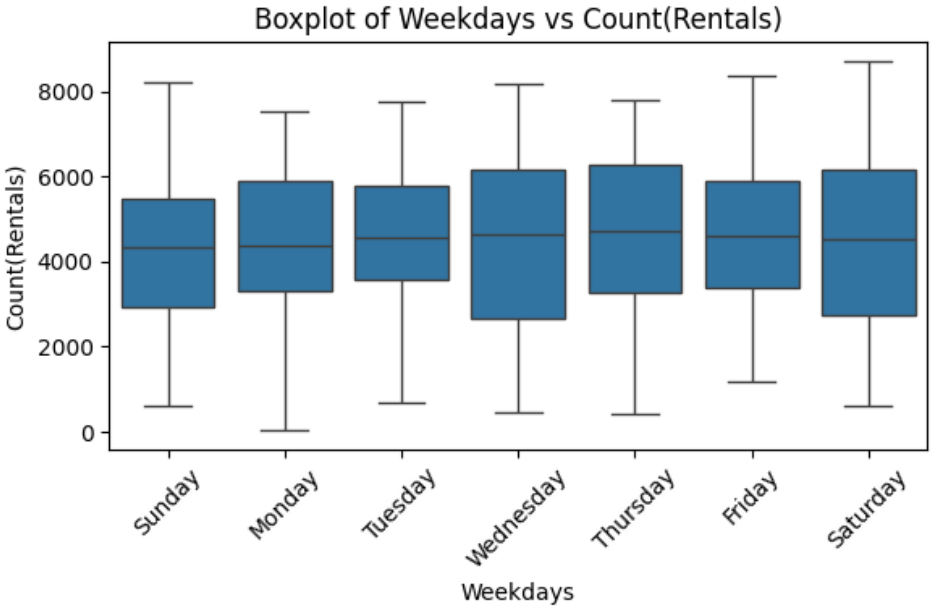


Figure 6: Box plot for Days vs Count(rentals)

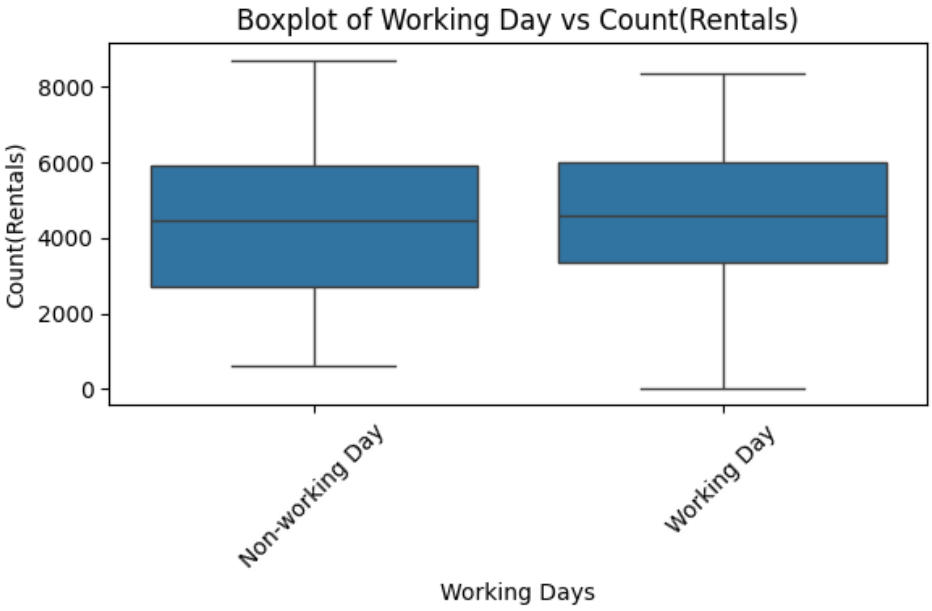


Figure 7: Box plot for Working Days vs Count(rentals)

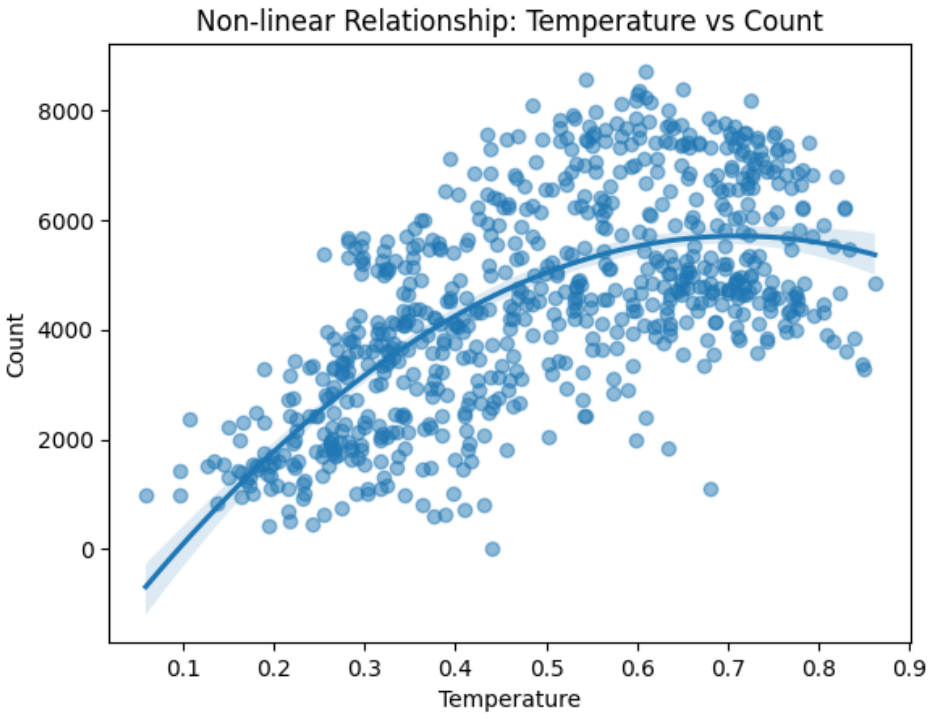


Figure 8: Box plot for Working Days vs Count(rentals)