Project

SS9859A

Regression

Predicting Bike Sharing

Author:
Aryan Rezanezhad
Parinaz Zarei

Student number:
251386495
251390152

Friday 8th December, 2023

Contents

1 Abstract

The bike-sharing system project explores the dynamics of individuals borrowing and returning bikes between company docks, with corresponding payments. The analysis focuses on a dataset reflecting the demand for the bike-sharing system across multiple variables, such as season, year, month, holiday, weekday, working day, and temperature. Notably, the dataset, previously utilized by Bike India, served as the foundation for adapting business strategies and plans to mitigate the impact of the Covid-induced pandemic on revenues. Our investigation utilized regression modeling techniques, revealing that our final model effectively explains 93% of the variation in the response variable.

2 Introduction

Bike sharing systems offer an affordable or even free means for users to navigate urban areas. Users can conveniently pick up a bicycle from one of the city's distributed stations and return it elsewhere. While these systems have advanced, incorporating sensors to track user interactions, their management and data utilization often fall short, resulting in insufficient bicycle availability at stations. The growing demand for car rentals has prompted the modernization of bike sharing systems through automated processes. Worldwide, there are over 500 thousand bicycles in various bike-sharing programs, attracting significant interest for their potential to alleviate traffic, environmental, and health concerns. Currently, Spain, Italy, and China lead as bike-friendly countries with 132, 104, and 79 programs, respectively. The number of metropolitan areas adopting bike-friendly systems is on the rise [1].

This report focuses on bike sharing data, specifically collected for the years 2011 and 2012 in both hourly and daily scales. Our analysis centers on the daily time series, spanning 730 days, where we examine 10 out of 14 variables due to the correlation of bike-sharing rental processes with seasonal conditions like weather, precipitation, temperature, holidays, and days of the week. The dataset originates from the two-year historical records of the Capital Bike Share system in Washington D.C., USA.

The objective of this project is to utilize a linear regression model to analyze the dataset and predict the number of daily rentals within each time frame. Our investigation aims to unveil the relationship between the dependent variable (response variable) and environmental and seasonal variables acting as predictors. All relevant code is included in the attachment.

3 Literature review

Bike-sharing data has been the subject of various studies. Yan Pan et al. have proposed a real-time prediction method for bike renting and returning in different city areas for future periods. Their approach relies on historical, weather, and time data, constructing a network of bike trips and applying a community detection method to identify two communities with the highest demand for shared bikes. The model's evaluation, based on the Root Mean Squared Error, demonstrates superior performance compared to other deep learning models, as evidenced by the comparison of their RMSEs [2]. Other studies have explored different aspects of bike-sharing data. Xudong Wang et al. developed a regression model with spatially varying coefficients to investigate the impact of land use, social-demographic factors, and transportation infrastructure on bike-sharing demand at various stations. Unlike existing geographically weighted models, they defined station-specific regression and employed a graph structure to ensure nearby stations had similar coefficients. Using data from the BIXI service in Montreal, they revealed spatially varying patterns in regression coefficients and highlighted areas more sensitive to marginal changes in specific factors [3].

In a different approach, researchers have explored the increasing trend of metropolitan cities adopting bike-friendly systems. The attractiveness of this data for research lies in its association with major cities offering this system alongside public transport services such as the subway. This report specifically focuses on bike-sharing data collected for the years 2011 and 2012, encompassing both hourly and daily scales, totaling 3,807,587 records. The analysis centers on the daily time series spanning 731 days, considering 10 out of 14 variables due to correlations with seasonal conditions, including weather, precipitation, temperature, holiday, and day of the week [1].

The study represents a fundamental research effort aimed at enhancing the efficiency of bike-sharing services and determining optimal locations for bike stations. Researchers conducted a thorough examination of bike usage characteristics and analyzed factors influencing demand using a multiple regression model. The findings from the analysis of bike usage characteristics indicate a higher rate of bike usage compared to the installation rate of public bike stations near parks in the city. Additionally, the demand for bike usage during weekends surpasses that of weekdays, suggesting that the primary purpose of bike usage may be recreational activities. Notably, the return rate at the same location as the rental station is relatively high. Furthermore, the study highlights a biased bike usage pattern in specific areas, specifically Dunsan and Yuseong, as bike-sharing stations are not evenly distributed [4].

Focused on applied regression analysis, this reference delves into both linear and non-linear regression techniques. It provides insights into the practical aspects of regression modeling, making it a valuable resource for those looking to understand and apply regression methods, including polynomial regression.[5].

This book is a comprehensive introduction to statistical learning, covering a range of machine learning techniques, including decision trees, random forests, and support vector machines. It is beneficial for those seeking a broader perspective on regression using modern machine learning approaches [6].

Focusing on nonparametric regression, this reference explores methods such as kernel regression. It is particularly relevant for understanding techniques that don't assume a specific functional form, providing flexibility in capturing patterns in the data, which can be valuable for bike-sharing datasets [7].

This classic work introduces generalized linear models (GLMs), extending the scope beyond linear regression to handle a broader class of problems. With a focus on probability distributions and link functions, it is an essential reference for researchers and practitioners interested in addressing diverse regression scenarios [8].

4 Preprocessing

4.1 Structure of the data

Below, you'll find the structure of our data, revealing a total of 16 variables. Our objective is to predict the variable "Casual" by leveraging the other variables within the model. This dataset spans 731 days, encompassing the collected information.

```
'data.frame':   731 obs. of  16 variables:
 $ instant   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ dteday    : chr  "2011-01-01" "2011-01-02" "2011-01-03" "2011-01-04" ...
 $ season    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ yr        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ mnth      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ weekday   : int  6 0 1 2 3 4 5 6 0 1 ...
 $ workingday: int  0 0 1 1 1 1 1 0 0 1 ...
 $ weathersit: int  2 2 1 1 1 1 2 2 1 1 ...
 $ temp      : num  0.344 0.363 0.196 0.2 0.227 ...
 $ atemp     : num  0.364 0.354 0.189 0.212 0.229 ...
 $ hum       : num  0.806 0.696 0.437 0.59 0.437 ...
 $ windspeed : num  0.16 0.249 0.248 0.16 0.187 ...
 $ casual    : int  331 131 120 108 82 88 148 68 54 41 ...
 $ registered: int  654 670 1229 1454 1518 1518 1362 891 768 1280 ...
 $ cnt       : int  985 801 1349 1562 1600 1606 1510 959 822 1321 ...
```

Figure 1: An Overview of Key Variables and Observations/ Output of R

- **instant**: Index or a unique identifier for each day
- **dteday**: The date of the record
- **season**: Categorical data representing the season
- **yr**: Year of the record. It's encoded as 0 for the first year in the dataset and 1 for the second year.
- **mnth**: The month of the year (1 to 12)
- **holiday**: Binary data indicating whether the day is a holiday (1) or not (0)
- **weekday**: Categorical data representing the day of the week
- **workingday**: Binary data indicating whether the day is a working day (1) or not (0)
- **weathersit**: Categorical data representing the weather situation
- **temp**: The normalized temperature in Celsius
- **atemp**: The normalized feeling temperature in Celsius
- **hum**: The normalized humidity level
- **windspeed**: The normalized wind speed
- **casual**: The count of casual users renting bikes
- **registered**: The count of registered users renting bikes
- **cnt**: The total count of bike rentals (casual PLUS registered).

4.2 Data Summary

Figure 1 illustrates the distribution of the "Casual" variable. The mean value for this variable is 848.2, with the 1st and 3rd quantiles at 315.5 and 848.2, respectively. The minimum value recorded was 2. The frequency of rental occurrences was higher at the 3rd quantile, gradually decreasing thereafter. Another significant predictor in our study is the "temp" variable, ranging from a minimum of 0.059 to a maximum of 0.861. Subsequently, a summary of each variable in the study is provided.

```
    instant         dteday            season           yr             mnth          holiday          weekday
 Min.   :  1.0   Length:731         Min.   :1.000   Min.   :0.0000   Min.   : 1.00   Min.   :0.00000   Min.   :0.000
 1st Qu.:183.5   Class :character   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.: 4.00   1st Qu.:0.00000   1st Qu.:1.000
 Median :366.0   Mode  :character   Median :3.000   Median :1.0000   Median : 7.00   Median :0.00000   Median :3.000
 Mean   :366.0                      Mean   :2.497   Mean   :0.5007   Mean   : 6.52   Mean   :0.02873   Mean   :2.997
 3rd Qu.:548.5                      3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:10.00   3rd Qu.:0.00000   3rd Qu.:5.000
 Max.   :731.0                      Max.   :4.000   Max.   :1.0000   Max.   :12.00   Max.   :1.00000   Max.   :6.000
   workingday      weathersit         temp             atemp              hum          windspeed          casual
 Min.   :0.000   Min.   :1.000   Min.   :0.05913   Min.   :0.07907   Min.   :0.0000   Min.   :0.02239   Min.   :   2.0
 1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.33708   1st Qu.:0.33784   1st Qu.:0.5200   1st Qu.:0.13495   1st Qu.: 315.5
 Median :1.000   Median :1.000   Median :0.49833   Median :0.48673   Median :0.6267   Median :0.18097   Median : 713.0
 Mean   :0.684   Mean   :1.395   Mean   :0.49538   Mean   :0.47435   Mean   :0.6279   Mean   :0.19049   Mean   : 848.2
 3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:0.65542   3rd Qu.:0.60860   3rd Qu.:0.7302   3rd Qu.:0.23321   3rd Qu.:1096.0
 Max.   :1.000   Max.   :3.000   Max.   :0.86167   Max.   :0.84090   Max.   :0.9725   Max.   :0.50746   Max.   :3410.0
   registered       cnt
 Min.   :  20   Min.   :  22
 1st Qu.:2497   1st Qu.:3152
 Median :3662   Median :4548
 Mean   :3656   Mean   :4504
 3rd Qu.:4776   3rd Qu.:5956
 Max.   :6946   Max.   :8714
```

Figure 2: Statistical Summary for Bike Rental Variables/ Output of R

4.3 Finding

The dataset comprises 730 rows and 16 columns. With the exception of one column, all others are either of float or integer type. The specific column mentioned is of date type. Upon inspecting the data, it appears that certain fields possess categorical characteristics, despite being in integer/float format. We will conduct a thorough analysis to determine whether to convert these fields to categorical or retain them as integers.

4.4 Eliminating unnecessary and redundant columns from the dataset.

The initial dataset comprises a total of 16 variables.

instant: Record index

dteday: Date

season: Season (1: spring, 2: summer, 3: fall, 4: winter)

yr: Year (0: 2011, 1:2012)

mnth: Month (1 to 12)

hr: Hour (0 to 23)

holiday: Weather day is holiday or not (extracted from Holiday Schedule)

weekday: Day of the week

workingday: If day is neither weekend nor holiday is 1, otherwise is 0.

weathersit: (extracted from Freemeteo)

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

temp: Normalized temperature in Celsius. The values are derived via $(t - t_{\min})/(t_{\max} - t_{\min})$, $t_{\min} = -8$, $t_{\max} = +39$ (only in hourly scale)

atemp: Normalized feeling temperature in Celsius. The values are derived via $(t - t_{\min})/(t_{\max} - t_{\min})$, $t_{\min} = -16$, $t_{\max} = +50$ (only in hourly scale)

hum: Normalized humidity. The values are divided by 100 (max)

windspeed: Normalized wind speed. The values are divided by 67 (max)

casual: Count of casual users

registered: Count of registered users

cnt: Count of total rental bikes including both casual and registered

Following a preliminary assessment of the data and the data dictionary, the following variables are identified for removal from subsequent analysis:

The variables identified for removal based on a high-level examination and the data dictionary are as follows:

- **instant:** It serves as an index value.
- **dteday:** This column includes the date, but since separate columns for 'year' and 'month' are already present, it can be omitted.
- **cnt and registered:** Both these columns contain the count of bike bookings by different customer categories. Since the objective is to determine the total count of bikes, regardless of specific categories, these columns can be excluded. Additionally, a new variable has been created to capture the ratio of these customer types.

We will save the modified dataframe as 'Bike_Sharing_New' to preserve the original dataset for any future analysis or validation. Dummy variables will be created for five categorical variables: 'yr', 'mnth', 'weekday', 'season', and 'weathersit'. Figure 3 illustrates the distribution of bike rentals.

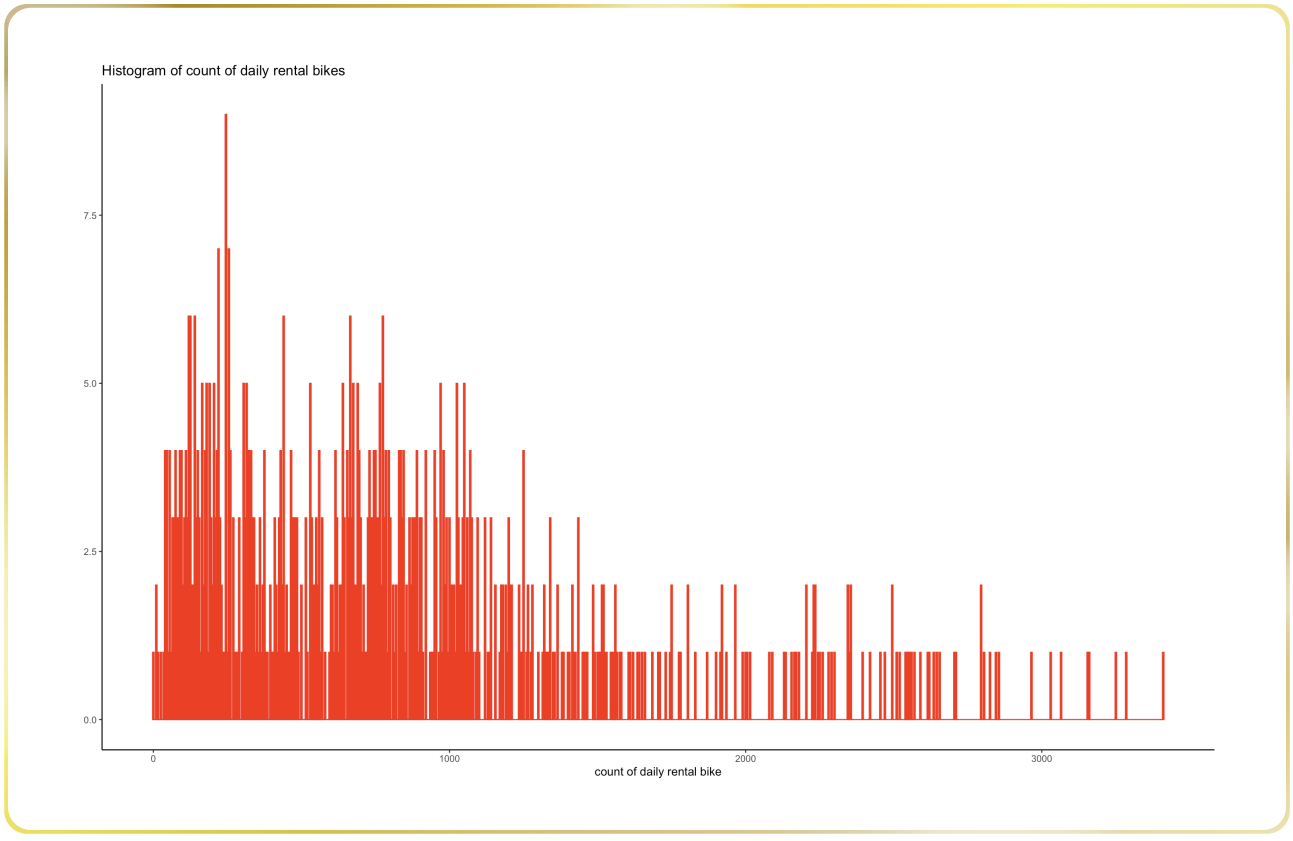

Figure 3: Daily Bike Rental Counts: A Histogram Analysis/ Output of R

5 Collinearity

Collinearity refers to a linear relationship between predictor or independent variables in regression analysis. It becomes a concern when there is a significant correlation between two potential predictor variables, leading to issues such as an increase in the p-value (reduction in significance level) of one predictor variable when another is included in the regression model, or a high variance inflation factor (VIF). The VIF quantifies the degree of collinearity, with values around 1 or 2 indicating essentially no collinearity and values of 20 or higher suggesting extreme collinearity. Multicollinearity poses a problem in multiple regression models as the input variables are interdependent, making it challenging to assess the individual impact of each independent variable on the dependent variable. Although multicollinearity doesn't diminish a model's overall predictive power, it can result in regression coefficients that are not statistically significant, resembling a form of double-counting in the model.
While high multicollinearity complicates the estimation of relationships between independent and dependent variables, it doesn't directly diminish the model's predictive power. However, it makes it harder to determine the unique influence of each variable on the dependent variable when multiple independent variables are closely related or measure similar aspects. In such cases, the underlying effect they measure is accounted for across variables, making it challenging to attribute specific influences.
Scatterplots are useful for visualizing variables with low or high collinearity. In Figures 4,5, there is evident high collinearity between the variables 'atemp' and 'temp,' emphasizing the need to address or consider this issue in the multiple regression model.

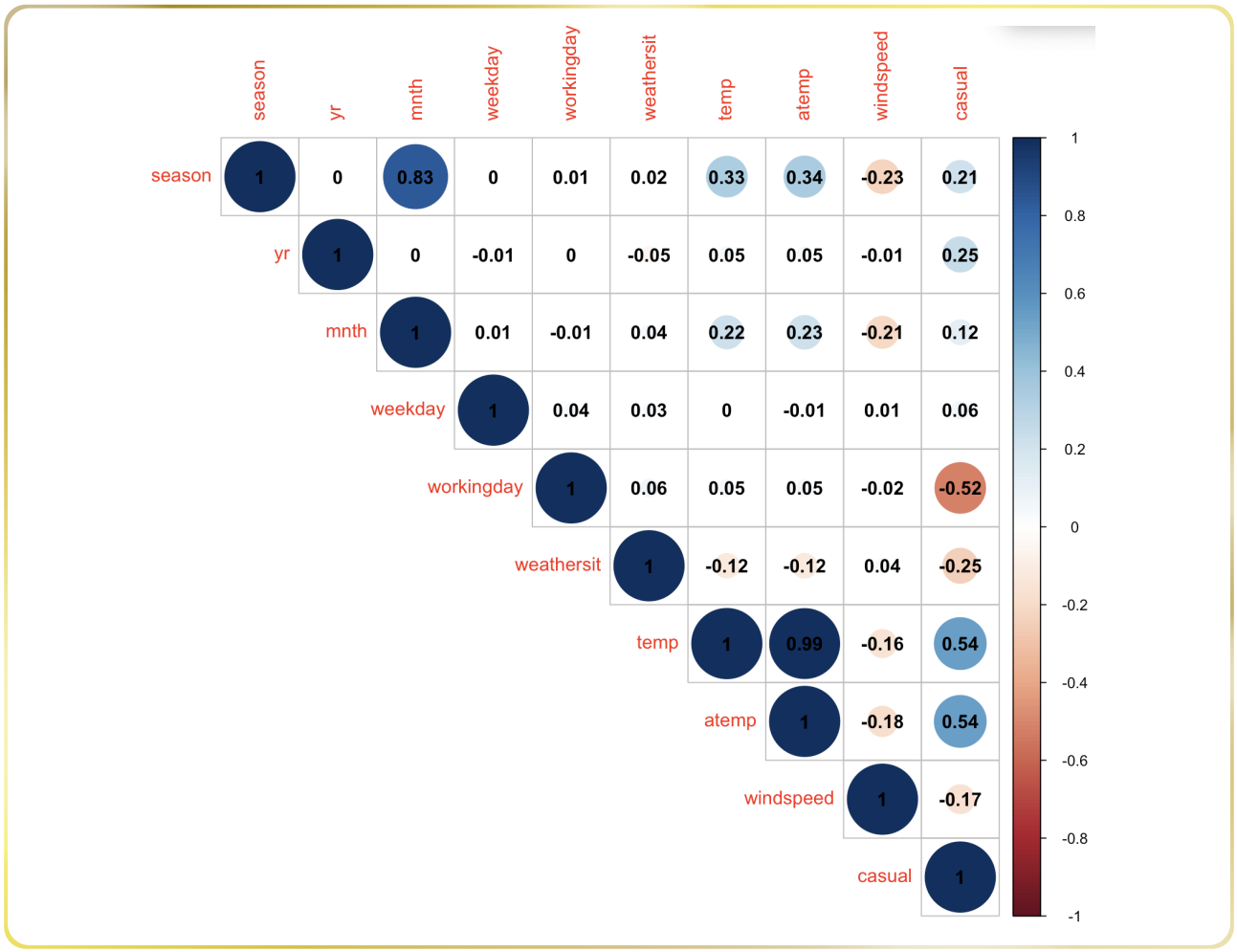

Figure 4: Correlation Matrix, visualizing Relationships and Trends/ Output of R

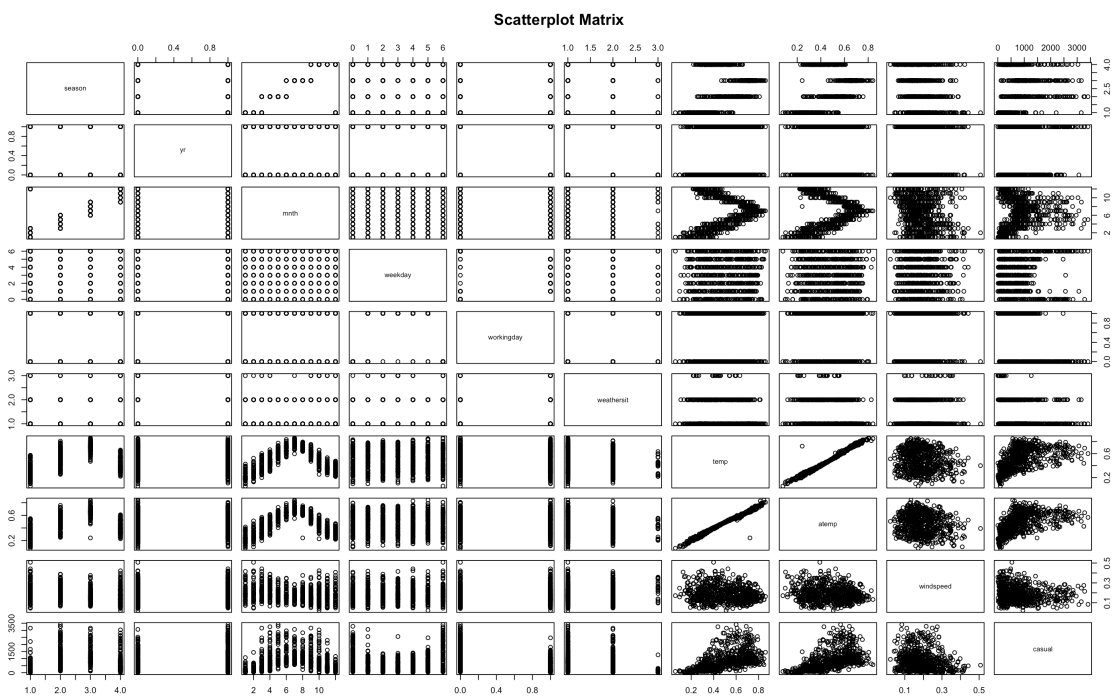

Figure 5: Scatterplot Matrix for Bike Rental Dataset Variables

5.1 Investigate collinearity

In R, the VIF() function is employed to calculate the Variance Inflation Factor (VIF) for variables in the model. In Figure 5, VIF values exceeding 10 are considered noteworthy. Upon analysis, it has been observed that three variables, namely 'temp', 'atemp', and 'season_3', have VIF greater than 10. Given the high correlation between 'temp' and 'atemp' revealed in the earlier scatter plot, a decision has been made to eliminate two predictors, 'atemp' and 'season_3', from the analysis. This step aims to address multicollinearity concerns and enhance the stability of the regression model.

```
   temp    atemp season_3
      2        3        6
```

Figure 6: Output of R

In the subsequent analysis, we have reported key metrics of the model, including R-squared, Adjusted R-squared, F-statistics, and p-value. Notably, in Figure 6 the regression model explains about 74% of the variability observed in the target variable. The small p-value indicates that at least one of the predictors is significant in the model, underscoring the model's overall statistical significance.

```
R squared is 0.7392109
 Adjusted R squared is 0.7291949
 F-statistics is 73.80241
 P_value is smaller than 2.2e-16
```

Figure 7: Output of R

6 Checking model assumptions

6.1 Breusch-Pagan test

The Breusch-Pagan Test (BP test), introduced by Trevor Breusch and Adrian Pagan in 1979, is employed to assess the presence of heteroscedasticity in a regression model. The test formulates the following null and alternative hypotheses:

Null Hypothesis (H_0) : Homoscedasticity is present (residuals are distributed with equal variance).

Alternative Hypothesis (H_1) : Heteroscedasticity is present (residuals are not distributed with equal variance).

The test involves the following steps:

1. Fit the regression model.
2. Calculate the squared residuals of the model.
3. Fit a new regression model using the squared residuals as response values.
4. Calculate the Chi-Square test statistic X^2 as $n \times R^2_{\text{new}}$, where:
   - n : Total number of observations.
   - R^2_{new} : R^2 of the new regression model that uses squared residuals as response values.
5. If the p-value corresponding to this Chi-Square test statistic, with p (number of predictors) degrees of freedom, is less than a chosen significance level (typically $\alpha = 0.05$), then the null hypothesis is rejected, indicating the presence of heteroscedasticity in the regression model. Otherwise, if the p-value is greater than α, the null hypothesis is not rejected, and homoscedasticity is assumed.

In the specific result of the `bptest` (Figure 8), it is noted that the p-value is very small. Therefore, the null hypothesis is rejected, providing evidence that heteroscedasticity is present in the regression model.

```
        studentized Breusch-Pagan test

data:  model_1
BP = 146.93, df = 26, p-value < 2.2e-16
```

Figure 8: Results of the Studentized Breusch-Pagan Test for Heteroscedasticity on Model 1 /Output of R

6.2 Shapiro-Wilk Test

The Shapiro-Wilk test assesses whether a sample comes from a normally distributed population. The null hypothesis for this test is that the population is normally distributed. If the p-value is less than the chosen alpha level, typically set at 0.05, the null hypothesis is rejected, suggesting evidence that the tested data do not follow a normal distribution. Conversely, if the p-value is greater than the chosen alpha level, the null hypothesis cannot be rejected, indicating that the data may be reasonably assumed to come from a normally distributed population.
In the provided output of the Shapiro-Wilk test (Figure 9), the small p-value indicates evidence against the null hypothesis. Therefore, it suggests that there is reason to reject the assumption that the data is drawn from a normally distributed population.

```
        Shapiro-Wilk normality test

data:  resid(model_1)
W = 0.96135, p-value = 5.833e-13
```

Figure 9: Shapiro-Wilk Test for Normality of Residuals in Model 1 /Output of R

After addressing collinearity concerns, a regression model was constructed for the response variable using other predictors. However, the results prompted a more thorough examination of the data, as indications of violations in the equality of variance and normality were observed through the Breusch-Pagan (BP) and Shapiro-Wilk (S-W) tests. Consequently, further scrutiny was conducted to investigate the presence of outliers in the data.

6.3 Cook's Distance

Cook's distance, denoted as D, is a metric used in regression analysis to identify influential outliers among predictor variables. It provides a measure of the impact of each observation on the regression model, combining leverage and residual values. Higher values of Cook's distance indicate greater influence of a particular observation on the model.
Interpretations of Cook's distance vary, and there is no universally accepted cutoff point. One approach suggests investigating points with a Cook's distance exceeding $\frac{4}{n}$, where n is the number of observations. Alternatively, some authors propose examining any "large" D, with a consensus that values above 1 indicate influential observations, while values above 0.5 may also warrant attention. Any value that significantly deviates from others should be thoroughly investigated.
In the provided code result (Figure 10), observations considered influential points are identified based on the threshold of $\frac{4}{n}$. The output indicates the presence of 55 influential points in the dataset.

```
> which(CD > 4/length(CD))
 26  65 105 106 184 185 204 239 247 266 282 302 340 341 358 442 443 448 449 456 457 462 463 470 471 472 478 505 513 514 519 533 534
 26  65 105 106 184 185 204 239 247 266 282 302 340 341 358 442 443 448 449 456 457 462 463 470 471 472 478 505 513 514 519 533 534
540 541 546 547 551 554 555 569 596 611 624 632 638 645 659 666 681 693 695 722 726 729
540 541 546 547 551 554 555 569 596 611 624 632 638 645 659 666 681 693 695 722 726 729
> length(which(CD > 4/length(CD)))
[1] 55
```

Figure 10: Identifies 55 influential data points via Cook's distance

An outlier is an observation point that significantly deviates from the fitted regression line, typically characterized by a large residual. The term "large" is context-dependent and is relative to the variable units. To standardize residuals and define outliers, the residuals are divided by an estimate of their standard deviation. In this context, an observation with a standardized residual larger than 2 is considered an outlier.
Following the treatment of outliers, a regression model was redeveloped. The results indicate an improvement in normality, but the assumptions of both normality and variance equality still did not hold. Given the violation of the variance equality assumption, a decision was made to transform the response variable. Two transformation approaches, namely log-transformation and Box–Cox transformation, were employed to identify the most suitable transformation.
In the subsequent analysis, the results of Breusch-Pagan (BP) and Shapiro-Wilk (SW) tests are presented. Both tests yield very small p-values, indicating a violation of the assumptions of equality of variance and normality. This suggests the need for further exploration and potential adjustments in the model.

7 Transformations

7.1 Logarithmic Transformation

In cases where there is a non-linear relationship between independent and dependent variables, and the original data exhibits skewness, log-transformation is employed to achieve monotonicity. This transformation replaces each variable x with log(x) . Following the log-transformation of the response variable (Figure 11), the results indicate very small p-values for both the Breusch-Pagan (BP) and Shapiro-Wilk (SW) tests. This suggests that even after the log-transformation, the assumptions of equality of variance and normality are still violated. Further considerations and adjustments may be necessary to address these issues in the model.

```
> bptest(model_3)

        studentized Breusch-Pagan test

data:  model_3
BP = 82.058, df = 26, p-value = 1.006e-07

> shapiro.test(resid(model_3))

        Shapiro-Wilk normality test

data:  resid(model_3)
W = 0.87144, p-value < 2.2e-16
```

Figure 11: R output

7.2 Box Cox transformation

The Box-Cox transformation is a method used to transform non-normally distributed dependent variables into a more normal shape. Normality is a crucial assumption for various statistical techniques, and applying the Box-Cox transformation allows for a broader range of tests to be applied to the data.
Named after statisticians George Box and Sir David Roxbee Cox, who collaborated on a 1964 paper introducing the technique, the Box-Cox transformation is particularly useful in situations where the normality assumption is not met.
Following the application of the Box-Cox transformation to the model (Figures 12, 13), it is observed that the p-values for both the Breusch-Pagan (BP) and Shapiro-Wilk (SW) tests have increased. This suggests that the transformation has helped address the violations of the assumptions of equality of variance and normality, indicating an improvement in the model's fit to the data.

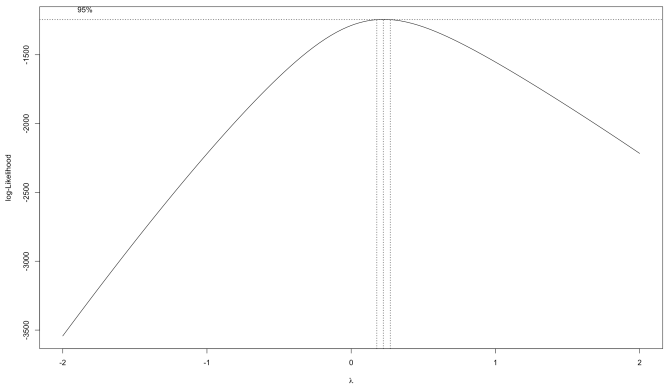

Figure 12: Data normalization via Box-Cox transformation / R output

```
> bptest(model_4)

        studentized Breusch-Pagan test

data:  model_4
BP = 34.835, df = 26, p-value = 0.1153

> shapiro.test(resid(model_4))

        Shapiro-Wilk normality test

data:  resid(model_4)
W = 0.99378, p-value = 0.01306
```

Figure 13: Test diagnostics for regression assumptions / R output

8 Variable Selection

The objective of this step is to identify a set of variables that optimally fit the model, enabling accurate predictions. Using the Akaike Information Criterion (AIC) as the measure, a set of variables was chosen through the stepwise method. The resulting model, based on the selected variables, is presented below (Figures 14, 15), with the lowest AIC observed at -35.78. This model is deemed to provide the best balance of fit and complexity according to the AIC measure.

```
Call:
lm(formula = ((casual^(lambda) - 1)/(lambda)) ~ temp + workingday +
    yr_1 + mnth_2 + mnth_4 + mnth_10 + weathersit_3 + weathersit_2 +
    mnth_5 + mnth_3 + mnth_9 + mnth_6 + mnth_8 + mnth_7 + mnth_11 +
    mnth_12 + weekday_3 + weekday_6 + weekday_2 + weekday_4 +
    weekday_1 + windspeed, data = Bike_Sharing_Dummy)

Residuals:
     Min       1Q   Median       3Q      Max
-2.47262 -0.62699  0.03344  0.68556  2.18304

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.53961    0.23593  40.435  < 2e-16 ***
temp          9.60524    0.57088  16.825  < 2e-16 ***
workingday   -2.64808    0.12961 -20.431  < 2e-16 ***
yr_1          1.59063    0.07896  20.146  < 2e-16 ***
mnth_2        0.60047    0.19634   3.058  0.00233 **
mnth_4        4.35094    0.23630  18.413  < 2e-16 ***
mnth_10       4.11410    0.23906  17.209  < 2e-16 ***
weathersit_3 -5.31605    0.56371  -9.430  < 2e-16 ***
weathersit_2 -1.08145    0.08372 -12.918  < 2e-16 ***
mnth_5        4.07282    0.27278  14.931  < 2e-16 ***
mnth_3        3.35836    0.20886  16.080  < 2e-16 ***
mnth_9        3.55397    0.28997  12.256  < 2e-16 ***
mnth_6        3.40127    0.31831  10.685  < 2e-16 ***
mnth_8        3.10383    0.32664   9.502  < 2e-16 ***
mnth_7        2.57599    0.35080   7.343 7.00e-13 ***
mnth_11       2.86274    0.20597  13.899  < 2e-16 ***
mnth_12       1.55321    0.20332   7.639 8.94e-14 ***
weekday_3    -1.48692    0.14016 -10.609  < 2e-16 ***
weekday_6     0.79848    0.15187   5.258 2.05e-07 ***
weekday_2    -1.19736    0.13965  -8.574  < 2e-16 ***
weekday_4    -1.04900    0.13764  -7.621 1.02e-13 ***
weekday_1    -0.80265    0.13174  -6.092 2.01e-09 ***
windspeed    -2.93049    0.53905  -5.436 7.98e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9561 on 587 degrees of freedom
Multiple R-squared:  0.9255,    Adjusted R-squared:  0.9227
F-statistic: 331.4 on 22 and 587 DF,  p-value: < 2.2e-16
```

Figure 14: Summary of linear regression model with coefficients and diagnostics/ R output

```
> bptest(model_5)

        studentized Breusch-Pagan test

data:  model_5
BP = 29.166, df = 22, p-value = 0.1402

> shapiro.test(resid(model_5))

        Shapiro-Wilk normality test

data:  resid(model_5)
W = 0.99312, p-value = 0.006746
```

Figure 15: Test diagnostics for regression assumption / R output

After the stepwise variable selection, our model is as follows:

```
lm(((casual^(lambda) - 1)/(lambda)) ~ temp + workingday + yr_1 +
mnth_2 + mnth_4 + mnth_10 + weathersit_3 + weathersit_2 + mnth_5 +
mnth_3 + mnth_9 + mnth_6 + mnth_8 + mnth_7 + mnth_11 + mnth_12 +
weekday_3 + weekday_6 + weekday_2 + weekday_4 + weekday_1 + windspeed)
```

During the examination of interaction effects between predictors, the interaction between 'temp' and 'workingday' was found to be significant. However, the Variance Inflation Factor (VIF) of this term is very high. Subsequently, a quadratic term of 'temp' was introduced into the model, and as indicated in the results, this quadratic term is deemed an important predictor in our model.

9 Semi-Final Model

The statement explains that the p-values from the Breusch-Pagan test for heteroscedasticity and the Shapiro-Wilk test for normality in the final statistical model exceed the 0.05 threshold, suggesting that the model's residuals are homoscedastic (have constant variance) and normally distributed, thereby satisfying key regression assumptions. Additionally, it mentions that an ANOVA test was conducted to evaluate the significance of including a quadratic term in the model, with the results indicating that the quadratic term substantially contributes to the model's ability to predict daily rental counts. Essentially, the inclusion of the quadratic term helps capture the non-linear relationships in the data, which improves the model's predictive accuracy.

10 Quadratic effects

To enhance the understanding of the quadratic term's influence on the response variable, we executed an ANOVA test contrasting the null and alternative hypotheses:

- H_0: The model sans the quadratic term ($\text{temp}^2 = 0$) is sufficient.
- H_1: The model benefits from the inclusion of the quadratic term.

The ANOVA test resulted in a p-value much lower than the conventional alpha level of 0.05, which leads us to reject H_0 in favor of H_1. This outcome endorses the quadratic term's significance, thereby confirming its integral role in enhancing the model's predictive accuracy and elucidating the underlying non-linear relationship between temperature and daily rentals.

```
Call:
lm(formula = ((casual^(lambda) - 1)/(lambda)) ~ temp + workingday +
    yr_1 + mnth_2 + mnth_4 + mnth_10 + weathersit_3 + weathersit_2 +
    mnth_5 + mnth_3 + mnth_9 + mnth_6 + mnth_8 + mnth_7 + mnth_11 +
    mnth_12 + weekday_3 + weekday_6 + weekday_2 + weekday_4 +
    weekday_1 + windspeed, data = Bike_Sharing_Dummy)

Coefficients:
 (Intercept)          temp    workingday          yr_1        mnth_2        mnth_4
      9.5396        9.6052       -2.6481        1.5906        0.6005        4.3509
     mnth_10  weathersit_3  weathersit_2        mnth_5        mnth_3        mnth_9
      4.1141       -5.3160       -1.0814        4.0728        3.3584        3.5540
      mnth_6        mnth_8        mnth_7       mnth_11       mnth_12     weekday_3
      3.4013        3.1038        2.5760        2.8627        1.5532       -1.4869
   weekday_6     weekday_2     weekday_4     weekday_1     windspeed
      0.7985       -1.1974       -1.0490       -0.8026       -2.9305
```

Figure 16: Fit the first model without the quadratic term / R output

```
Call:
lm(formula = ((casual^(lambda) - 1)/(lambda)) ~ temp + workingday +
    yr_1 + weathersit_3 + weathersit_2 + weekday_6 + mnth_3 +
    windspeed + weekday_3 + weekday_2 + weekday_4 + weekday_1 +
    mnth_10 + mnth_9 + mnth_8 + mnth_11 + mnth_4 + mnth_5 + mnth_6 +
    mnth_7 + mnth_12 + I(temp^2), data = Bike_Sharing_Dummy)

Coefficients:
 (Intercept)          temp    workingday          yr_1  weathersit_3  weathersit_2
      7.0001       25.7570       -2.6474        1.5009       -5.3742       -1.1440
   weekday_6        mnth_3     windspeed     weekday_3     weekday_2     weekday_4
      0.7943        2.5976       -3.2355       -1.4359       -1.1501       -1.0003
   weekday_1       mnth_10        mnth_9        mnth_8       mnth_11        mnth_4
     -0.7915        3.1894        3.0257        3.2137        2.0136        3.4323
      mnth_5        mnth_6        mnth_7       mnth_12     I(temp^2)
      3.4119        3.2811        3.0853        0.8401       -17.6794
```

Figure 17: Fit the second model with the quadratic term / R output

```
> bptest(model_6)

        studentized Breusch-Pagan test

data:  model_6
BP = 29.166, df = 22, p-value = 0.1402

> shapiro.test(resid(model_6))

        Shapiro-Wilk normality test

data:  resid(model_6)
W = 0.99312, p-value = 0.006746
```

Figure 18: BP and Shapiro Test for model without the quadratic term / R output

```
> bptest(model_7)

        studentized Breusch-Pagan test

data:  model_7
BP = 57.771, df = 23, p-value = 7.971e-05

> shapiro.test(resid(model_7))

        Shapiro-Wilk normality test

data:  resid(model_7)
W = 0.99742, p-value = 0.4652
```

Figure 19: BP and Shapiro Test for model with the quadratic term / R output

```
> anova(model_6, model_7)
Analysis of Variance Table

Model 1: ((casual^(lambda) - 1)/(lambda)) ~ temp + workingday + yr_1 +
    mnth_2 + mnth_4 + mnth_10 + weathersit_3 + weathersit_2 +
    mnth_5 + mnth_3 + mnth_9 + mnth_6 + mnth_8 + mnth_7 + mnth_11 +
    mnth_12 + weekday_3 + weekday_6 + weekday_2 + weekday_4 +
    weekday_1 + windspeed
Model 2: ((casual^(lambda) - 1)/(lambda)) ~ temp + workingday + yr_1 +
    mnth_2 + mnth_4 + mnth_10 + weathersit_3 + weathersit_2 +
    mnth_5 + mnth_3 + mnth_9 + mnth_6 + mnth_8 + mnth_7 + mnth_11 +
    mnth_12 + weekday_3 + weekday_6 + weekday_2 + weekday_4 +
    weekday_1 + windspeed + I(temp^2)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    587 536.60
2    586 466.78  1    69.824 87.659 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 20: ANOVA to compare model with-without quadratic term / R output

The ANOVA has yielded a p-value that is significantly lower than the threshold of 0.05. Consequently, we can dismiss the null hypothesis, H0. This leads us to incorporate the quadratic term into our statistical model.

```
> summary(model_6)$r.squared
[1] 0.9254771
> summary(model_6)$adj.r.squared
[1] 0.9226841
> summary(model_7)$r.squared
[1] 0.9351743
> summary(model_7)$adj.r.squared
[1] 0.9326299
```

Figure 21: ANOVA to compare model with-without quadratic term / R output

$Model_6$ explains 92.5 % of the variance in the dependent variable, while $model_7$ explains 93.5 %, slightly improving upon the explanatory power of the previous model. R-squared is a statistical measure that indicates the percentage of the variance in the dependent variable that can be explained by the independent variables in the model. The output suggests that our model's predictions improve when we include a term that involves a number quadratic term. Therefore, so we choose model with quadratic term.

11 Graphs confirm the validity of our model

- Residual vs Fitted Plot

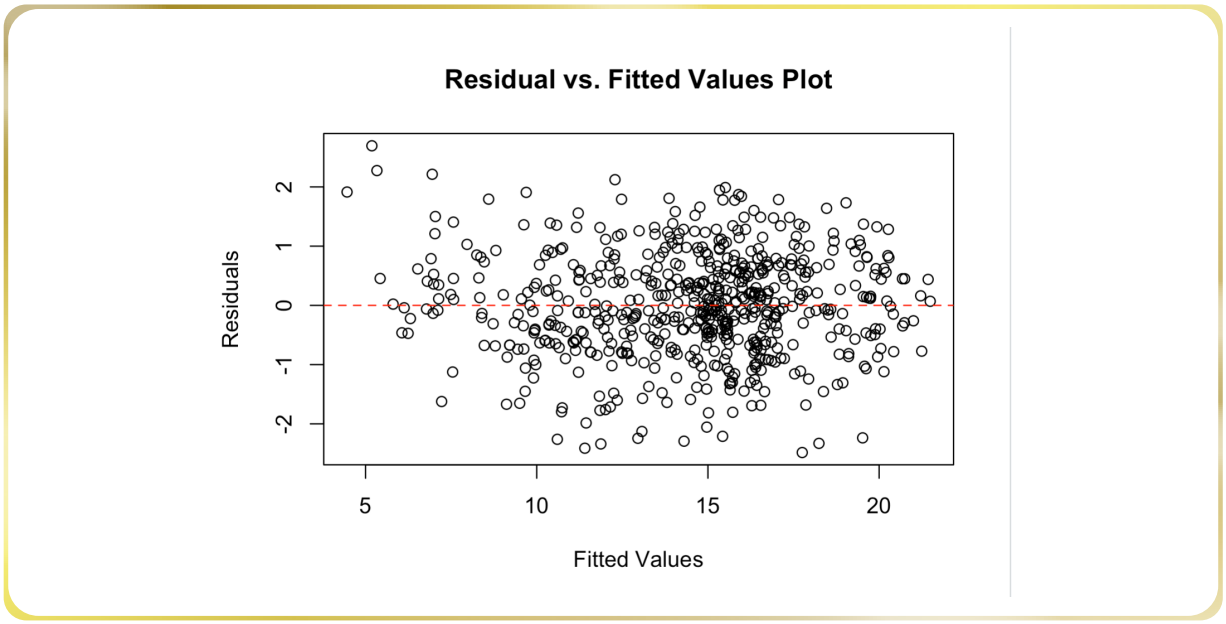

Figure 22: Residuals vs. Fitted Values plot / R output

The plot suggests that the residuals are scattered randomly and do not form any particular pattern as they vary with the fitted values. This random scatter is a good sign that the model's assumption of a linear relationship between the independent variables and the dependent variable is reasonable. It means the model is likely fitting the data without missing any obvious trends or patterns.

- Standardized Residuals vs Theoretical Quantiles

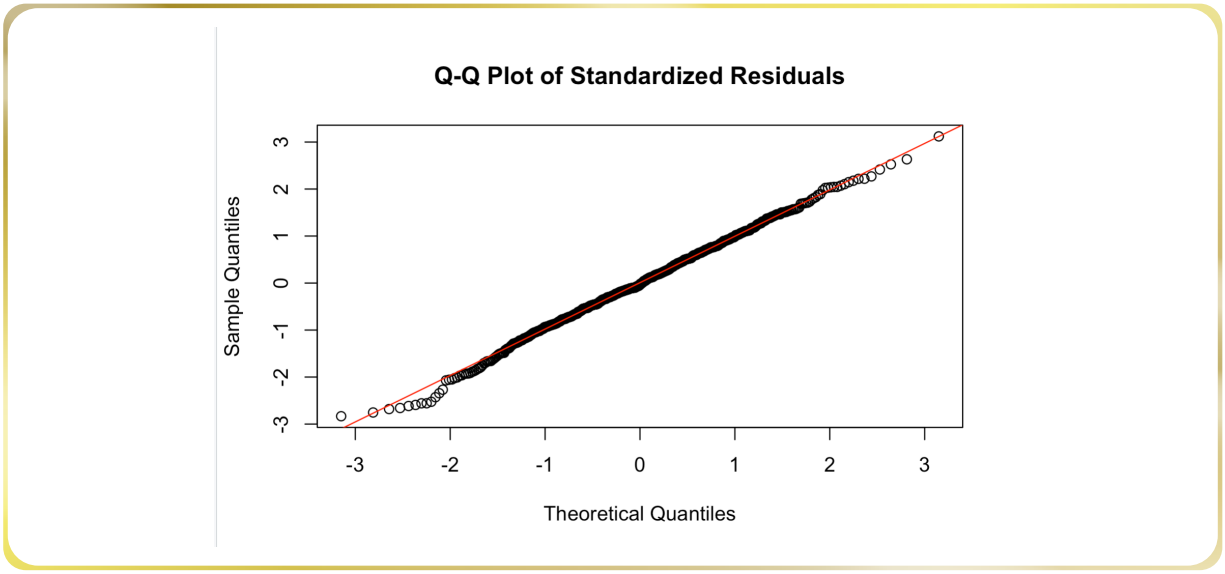

Figure 23: Standardized Residuals vs Theoretical Quantiles / R output

In the provided plot, most of the data points fall along the straight line, which implies that the residuals of the model follow a normal distribution quite closely. This suggests that the normality assumption for the linear regression model's residuals is met.

- Residual-Fitted Value Plot

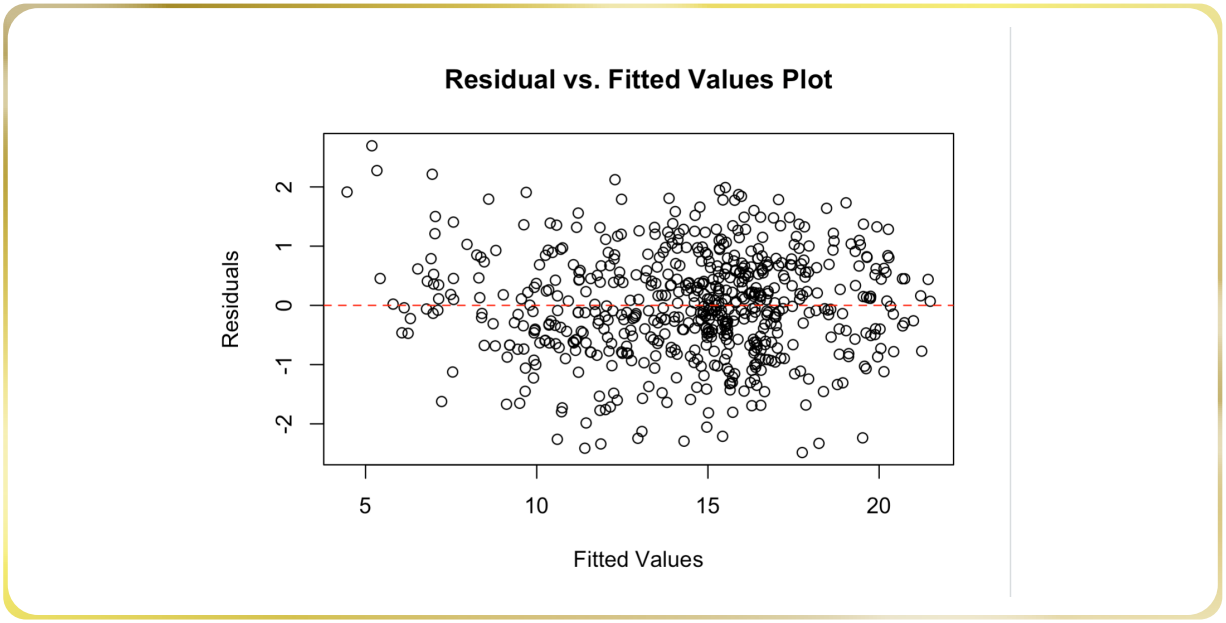

Figure 24: Residuals vs. Fitted Values plot / R output

From the plot, it seems that the residuals are scattered randomly around the horizontal line at zero, without forming any distinct patterns. This kind of spread is generally a good indication that the residuals have constant variance and the model is well-fitted.

- Scale-Location Plot

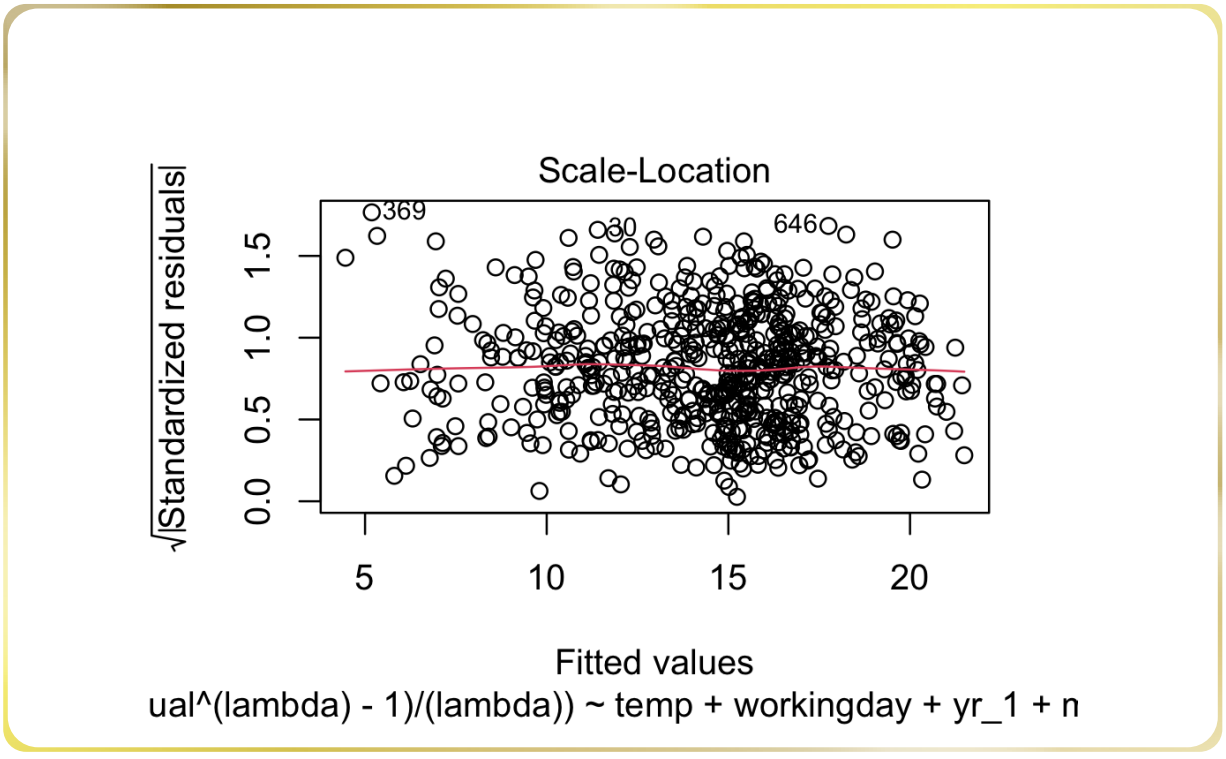

Figure 25: Scale-Location plot displaying homoscedasticity of residuals in a linear regression model / R output

This uniform spread suggests that the model's errors are consistent across all values, which is a good sign that the model is fitting the data well.

- Residuals vs. Leverage plot

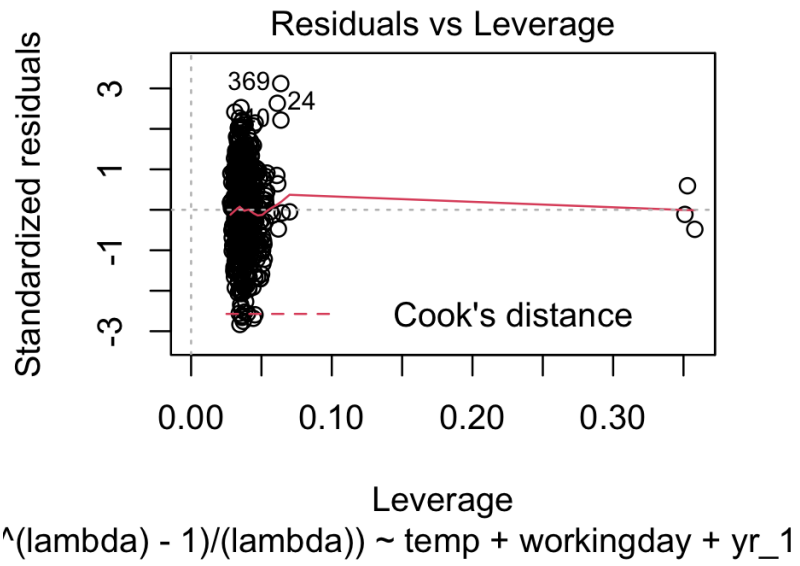

Figure 26: The Residuals vs. Leverage plot / R output

Most points cluster in the center, indicating low leverage. However, points far from the center line, especially those beyond the Cook's distance lines (dashed lines), may have a disproportionate impact on the model. The few points with high leverage or large residuals warrant closer examination as potential outliers or influential data points.

12 Conclusion

In our analysis, we have crafted a predictive model that effectively identifies the main factors influencing bike rental numbers. This allows bike rental businesses to leverage these insights, align their operational strategies accordingly, and potentially enhance their revenue. While our model is robust, we acknowledge the utility of alternative modeling approaches such as logistic regression, random forest, and XG Boosting, which could also serve to predict bike sharing demand. Moreover, a Poisson regression model is another viable method that could be applied to our dataset, particularly for its ability to predict counts. Our findings are significant, with the model accounting for 93% of the variability in casual bike rental usage, indicating a strong predictive performance.

References

[1] F. H. Fanaee-T and S. J. Gama, "Event labeling combining ensemble detectors and background knowledge," *Progress in Artificial Intelligence*, vol. 2, pp. 113–137, 2014.

[2] Y. Pana, R. C. Zheng, J. Zhangaa, and X. Yao, "Predicting bike sharing demand using recurrent neural," *Procedia Computer Science*, vol. 147, pp. 562–566, 2019.

[3] X. Wang, Z. Cheng, M. Trépanier, and L. Sun, "Modeling bike-sharing demand using a regression model with spatially varying coefficients," *Journal of Transport Geography*, vol. 93, 2021.

[4] M. sik and Yunseung, "Analysis of the affecting factors on the bike-sharing demand focused on daejeon city," *KSCE Journal of Civil _ Environmental Engineering Research*, vol. 5, pp. 1517–1524,

[5] L. S. Fahrmeir L Kneib T, "Regression - modelle, methoden and anwendungen," *Springer*, vol. 2, 2009.

[6] H. T. James G Witten D, "An introduction to statistical learning," *Springer*, vol. 3, 2013.

[7] CloudyML, "Bike sharing demand project," *YouTube Channel*, vol. 4, 2022.

[8] N. Bansal, "Regression techniques on a bike-sharing dataset," *YouTube Channel*, vol. 3, 2019.

13 Appendix

```
#Libraries
library(ggplot2)
library(fastDummies)
library(lmtest)
library(MASS)
library(faraway)
library(car)

#Data
Bike_Sharing <- read.csv("/Users/rezanejad/Desktop/Reg Project/
                                    day.csv",header=T)
#Structure of Data
str(Bike_Sharing)
summary(Bike_Sharing)

#Cleaning Data
Bike_Sharing_New = Bike_Sharing[,c("season", "yr", "mnth", "weekday",
            "workingday", "weathersit", "temp",
            "atemp", "windspeed","casual")]
dim(Bike_Sharing_New)
Bike_Sharing_Dummy <- dummy_cols(Bike_Sharing_New, select_columns =
    c("season","yr","weekday","mnth","weathersit"), TRUE)
Bike_Sharing_Dummy <- subset(Bike_Sharing_Dummy, select =
    -c(season_1,weekday_0,yr_0,mnth_1,weathersit_dim(Bike_Sharing_Dummy)
#Histogram
qplot(Bike_Sharing_New$casual, geom="histogram", binwidth = 5,
        main ="Histogram of count of daily rental bikes",
        xlab ="count of daily rental bike", fill=I("red"),
        col=I("red"))+ theme_classic()

#Pair Plot
pairs(~.,data=Bike_Sharing_New,main="Scatterplot Matrix")
library(corrplot)
par(mfrow=c(1,1))
cor_matrix <- cor(Bike_Sharing_New,method = "pearson",use =
            "complete.obs")
corrplot(cor_matrix,method = "circle", type = "upper" ,
            addCoef.col = "black")

#Calculating VIF
model = lm(casual ~ ., data = Bike_Sharing_Dummy)

#cor(Bike_Sharing_Dummy)which(vif(model)>10)
Bike_Sharing_Dummy = subset(Bike_Sharing_Dummy, select =
            -c(atemp,season_3))
r_squared <- summary(model)$r.squared
adjusted_r_squared <- summary(model)$adj.r.squared
f_statistic <- summary(model)$fstatistic[1]
```

```
p_value <- summary(model)$fstatistic[4]
cat("R squared is", r_squared,"\n","Adjusted R squared is",
        adjusted_r_squared,"\n","F-statistics is", f_statistic,"\n",
        "P_value is smaller than 2.2e-16")

#Model evaluation
model_1 = lm(casual ~ ., data = Bike_Sharing_Dummy)

#Checking equal variance assumption
bptest(model_1)

#Checking normality assumption
shapiro.test(resid(model_1))

#Cook's distance
CD = cooks.distance(model_1)
which(CD > 4/length(CD))

#Cook's distances with threshold = 4/n
out_i = which( CD > 4/length(CD))
Bike_Sharing_Dummy = Bike_Sharing_Dummy[-out_i,]

#Logarithmic transformation
model_3 = lm(log(casual) ~ ., data = Bike_Sharing_Dummy)
bptest(model_3)
shapiro.test(resid(model_3))
out_i = which(cooks.distance(model_3)>4 / length(cooks.distance(model_3))
Bike_Sharing_Dummy = Bike_Sharing_Dummy[-out_i,]

#Box Cox Transformation
bc = boxcox(model_1)
lambda <- bc$x[which.max(bc$y)]

#Fit new linear regression model using the Box-Cox transformation
model_4 <-lm(((casual^lambda-1)/lambda) ~ ., data = Bike_Sharing_Dummy)
out_i = which(cooks.distance(model_4)>4 / length(cooks.distance(model_4))
Bike_Sharing_Dummy = Bike_Sharing_Dummy[-out_i,]
model_4 <- lm(((casual^lambda-1)/lambda) ~ ., data = Bike_Sharing_Dummy)
bptest(model_4)
shapiro.test(resid(model_4))

#Variable Selection
model_5= step(lm(((casual^(lambda) - 1)/(lambda))~1,
        data = Bike_Sharing_Dummy),scope =((casual^(lambda) -
        1)/(lambda)) ~ temp + workingday + yr_1 + season_4 +
        weathersit_2 + mnth_3 + weathersit_3 + mnth_4 +
        weekday_6 + mnth_10 + mnth_9 + mnth_5 + windspeed +
        weekday_3 + mnth_11 + weekday_2 + weekday_4 + weekday_1 +
        mnth_8 + mnth_6 + mnth_7 + mnth_12 + mnth_2 ,
        direction = "both",trace = 0)
model_5 = lm(((casual^(lambda) - 1)/(lambda)) ~ temp + workingday +
```

```
        yr_1 +mnth_2 + mnth_4 + mnth_10 + weathersit_3 +
        weathersit_2 + mnth_5 + mnth_3 + mnth_9 + mnth_6 +
        mnth_8 + mnth_7 + mnth_11 + mnth_12 + weekday_3 +
        weekday_6 + weekday_2 + weekday_4 + weekday_1 +
        windspeed ,data = Bike_Sharing_Dummy)
summary(model_5)
bptest(model_5)
shapiro.test(resid(model_5))

#Semi_Final_Model and after that.
#Final model
model <- lm(((casual^(lambda) - 1) / lambda) ~ temp +
        workingday + yr_1 + mnth_2 + mnth_4 + mnth_10 +
        weathersit_3 + weathersit_2 + mnth_5 + mnth_3 +
        mnth_9 + mnth_6 + mnth_8 + mnth_7 + mnth_11 +
        mnth_12 + weekday_3 + weekday_6 + weekday_2 +
        weekday_4 + weekday_1 + windspeed, data =
        Bike_Sharing_Dummy)

#Quadratic effect
#model without quadratic form
model_6 <- lm(formula = ((casual^(lambda) - 1)/(lambda)) ~ temp +
        workingday + yr_1 + mnth_2 + mnth_4 + mnth_10 +
        weathersit_3 + weathersit_2 + mnth_5 + mnth_3 + mnth_9 +
        mnth_6 + mnth_8 + mnth_7 + mnth_11 + mnth_12 + weekday_3 +
        weekday_6 + weekday_2 + weekday_4 + weekday_1 + windspeed,
        data = Bike_Sharing_Dummy)
summary(model_6)
bptest(model_6)
shapiro.test(resid(model_6))

#Fit the first model without the quadratic term
model_6 <- lm(formula = ((casual^(lambda) - 1)/(lambda)) ~ temp +
        workingday + yr_1 + mnth_2 + mnth_4 + mnth_10 +
        weathersit_3 + weathersit_2 + mnth_5 + mnth_3 + mnth_9 +
        mnth_6 + mnth_8 + mnth_7 + mnth_11 + mnth_12 + weekday_3 +
        weekday_6 + weekday_2 + weekday_4 + weekday_1 + windspeed,
        data = Bike_Sharing_Dummy)

#Fit the second model with the quadratic term
model_7 <- lm(formula = ((casual^(lambda) - 1)/(lambda)) ~ temp +
        workingday + yr_1 + mnth_2 + mnth_4 + mnth_10 +
        weathersit_3 + weathersit_2 + mnth_5 + mnth_3 + mnth_9 +
        mnth_6 + mnth_8 + mnth_7 + mnth_11 + mnth_12 + weekday_3 +
        weekday_6 + weekday_2 + weekday_4 + weekday_1 + windspeed +
        I(temp^2), data = Bike_Sharing_Dummy)
bptest(model_7)
shapiro.test(resid(model_7))
bptest(model_6)
shapiro.test(resid(model_6))
anova(model_6, model_7)
```

```
summary(model_6)$r.squared
summary(model_6)$adj.r.squared
summary(model_7)$r.squared
summary(model_7)$adj.r.squared
summary(model_7)$fstatistic
pf(summary(model_7)$fstatistic[1], summary(model_7)$fstatistic[2],
        summary(model_7)$fstatistic[3], lower.tail = FALSE)
plot(x = fitted(model_7), y = residuals(model_7), xlab = "Fitted
        Values", ylab = "Residuals", main = "Residual vs. Fitted
        Values Plot")
abline(h = 0, col = "red", lty = 2)
std_residuals <- rstandard(model_7)
qqnorm(std_residuals, main = "Q-Q Plot of Standardized Residuals")
qqline(std_residuals, col = "red")
scale_location_plot <- plot(model_7, which = 3)
plot(model_7, which = 5)
abline(h = c(0.5, 1), col = "red", lty = 2)
influencePlot(model_7, id.method = "identify", main = "Influence
        Plot with Cook's distance")
```