STA 319 2.0

Advanced Regression Analysis

Project

Name : R.H.T.O.Wickramarathne

Index No: AS2018581

Table of Contents

List of Figures

List of Tables

Introduction

Background of the problem

Bike sharing systems are the new generation of traditional bike rentals where the whole process from membership, rental and return has become automatic. The user can easily rent a bike from a particular position and return to another place through these systems. Currently, there are about over 500 bike-sharing programs around the world which are composed of over 500 thousand bicycles. Today, there exists great interest in these systems due to their essential role in traffic, environmental and health issues. These systems could provide an alternative mode of transportation for many people, whether for work or leisure. Despite giving many benefits to their users, they have some disadvantages. Apart from interesting real-world applications of bike-sharing systems, the characteristics of data generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns the bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of the crucial events in the town could be detected via monitoring these data.

Table 1:Data description

Variable	**Description**
instant	record index
dteday	date
season	Winter, spring, summer, autumn (numeric 1, 2,3,4)
yr	Year (2011, 2012) numeric values (0, 1)
mnth holiday	Months (numeric values 1 to 12) whether a day is holiday or not (1 for holiday, 0 for not holiday)
weekday	day of the week (numeric values 1 to 7)
workingday	if day is neither weekend nor holiday is 1, otherwise is 0
temp	The normalized temperature in Celsius. The values are divided by 41 (max)
weathersit	Clear, Few clouds, partly cloudy, partly cloudy (numeric value 1) Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist (numeric value 2) Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (numeric value 3) Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog (numeric value 4)
atemp	Normalized air temperature in Celsius. The values are divided by 50 (max)
hum	Normalized humidity. The values are divided by 100 (max)
windspeed	Normalized wind speed. The values are divided by 67 (max)
cnt	count of total rental bikes including both casual and registered

Objectives

- Make a prediction of bike rental count daily based on the environmental and seasonal settings.
- Make a validation of anomaly or event detection.

Methodology

This project aims to investigate the relationship between the factors that may influence the number of rented bikes and their mobility in a city.

This project's implementation phase has been investigated in three major sections. The first section conducts exploratory data analysis. The second section performs data analysis and predictive model and fits the model, and the third section is detected anomalies and events then validate the model. The data was loaded and processed using R software.

Variables in the data set were investigated as environmental and seasonal variables, and their essential parameters were calculated and summarized. Environmental variables are temperature, humidity, wind speed and weather sit. The season variable with its four levels (winter, spring, summer, autumn) is one of the factors which could have some potential effect on the number of rented biked throughout the year. And these variables were explored using histograms, boxplots, and scatterplots. Outliers were detected using boxplots. But outliers were not removed since there can be extreme cases in environmental variables.

The data analysis was done by using few steps and used 2011 data set. First, identify the effects of multicollinearity. It was illustrated using matrix correlations for qualitative and quantitative variables. An absolute correlation coefficient of >0.7 among two or more predictors indicates the presence of multicollinearity.

To determine the possible relationship among variables then discard few variables due to multicollinearity effect, the multilinear regression model is built. The MLR model represented as following if there are p number of predictors,

$$Y=\beta_1X_1+\beta_2X_2+\beta_3X_3+\ldots+\beta_pX_p+\varepsilon_i$$

Where Y describes the dependent variable and $X_1,X_2,\ldots,X_p$ are independent variables, $\beta_1,\beta_2,\ldots,\beta_p$ are called regression coefficients and ε is known as the error term.

Model was built using lm () function in R . Since qualitative variables were in the model, the factor() function was used to covert the categorical variables into factors. Then the summary function is used to obtain the outputs.

Then the backward selection method was used to find the best subset of predictor variables and set 0.15 as α_R(alpha to remove) and removed variables with the highest p-values and greater than the α_R. After choosing the best subset of predictor variables, conduct residual analysis using histogram,

normal QQ-plot and residual vs fitted values plot to check whether the target variable is satisfied assumptions of residuals. Shapiro-Wilcoxon test was conducted to examine the normality of error terms.

Hypothesis to be tested for Shapiro-Wilcoxon test:

H_0 : Residuals follow normal distribution

H_1 : Residuals do not follow normal distribution.

Decision rule : Reject H_0 if p-value < 0.05 (α)

Next, Cook's distance bar chart used to detect influential observations and influential cases were removed from the data set. Then re-fit the model according to the new data set. Again, conduct the residual analysis for the new model.

Afterwards, using correlogram of ACF and PACF found the order of serial correlation. A test called the Breusch-Godfrey test is performed using bgtst() in order to detect autocorrelation.

Assume ε_t follows an AR(p) process and let ρ be the i^{th} autocorrelation in the residuals

Hypothesis to be tested:

$H_0 : \rho_1 = \rho_2 = \cdots = \rho_p = 0$ and H1: At least one $\rho_i \neq \rho_j$

Decision rule :Reject Ho if p-value < 0.05

Then Cochrane-Orcutt method was used to adjust autocorrelation. And it was the best-fitted model for our data set. After that, conduct residual analysis to check assumptions on the residuals.

Finally, the third section has validated the model using the 2012 data set as test data set. It was done by obtaining MSE for the 2011 data set, and the fitted model is used to predict each case in the new data set (2012 data set) and calculate the mean of squared prediction error(MSPR). Then compare these two values. Suppose MSPR is fairly close to the MSE of the fitted model. In that case, the MSE of the fitted model is not seriously biased and gives an appropriate indication of the model's predictive ability.

Exploratory data analysis

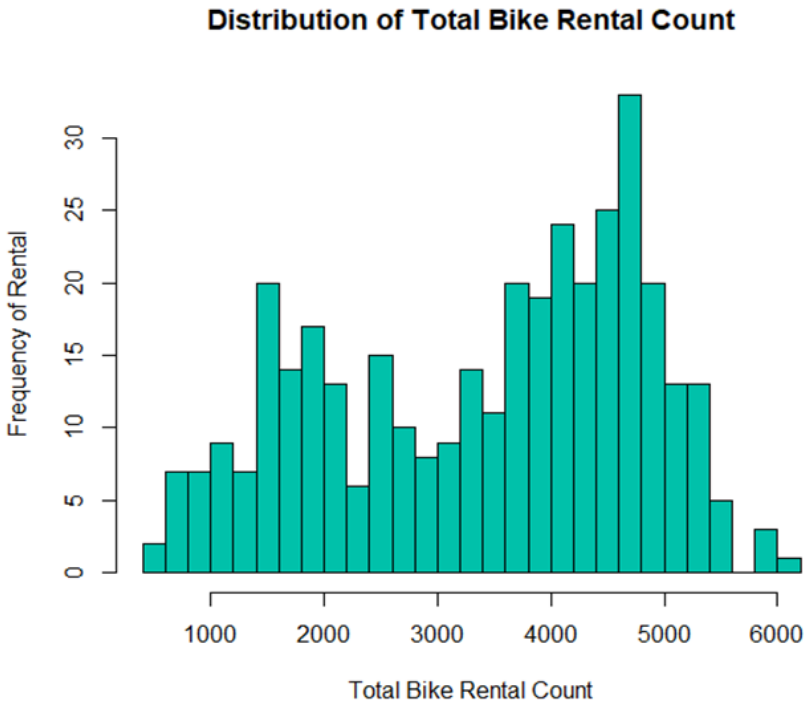

Figure 1: Distribution of total bike rental count

Figure 1 shows the number of total rented bikes following a left-skewed distribution.

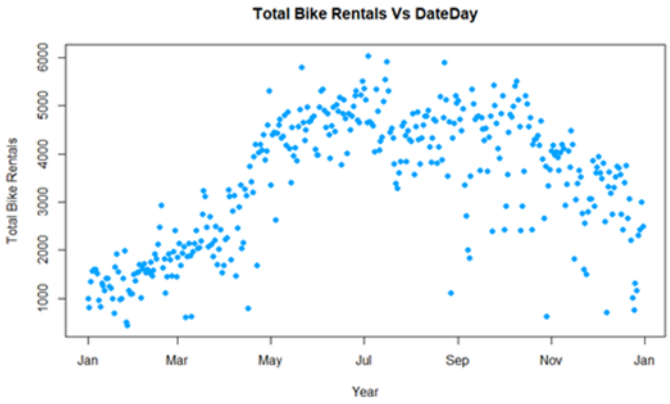

Figure 2:Total bike rental count vs DateDay

The graph depicts the relationship between the Total Bike Rentals(cnt) and the Year variable. During the summer and fall seasons, total number of rented bikes are high.

Exploring quantitative variables.

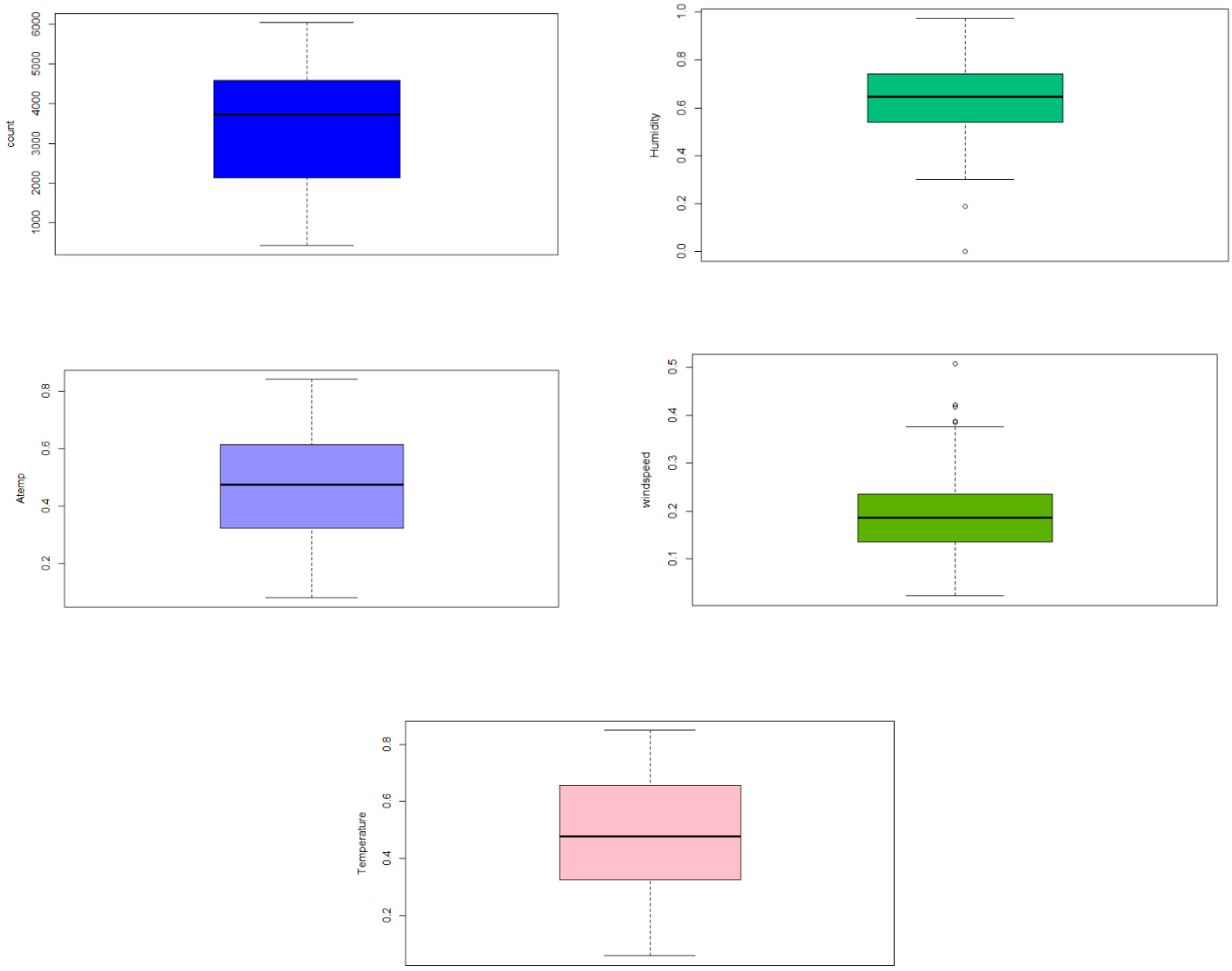

Figure 3:Distribution of quantitative variables

According to the box plots, there are no outliers present in the number of rental bikes, temperature and feeling temperature. But outliers are present in humidity and wind speed. Moreover, the distribution of the e temperature and feeling temperature are approximately same.

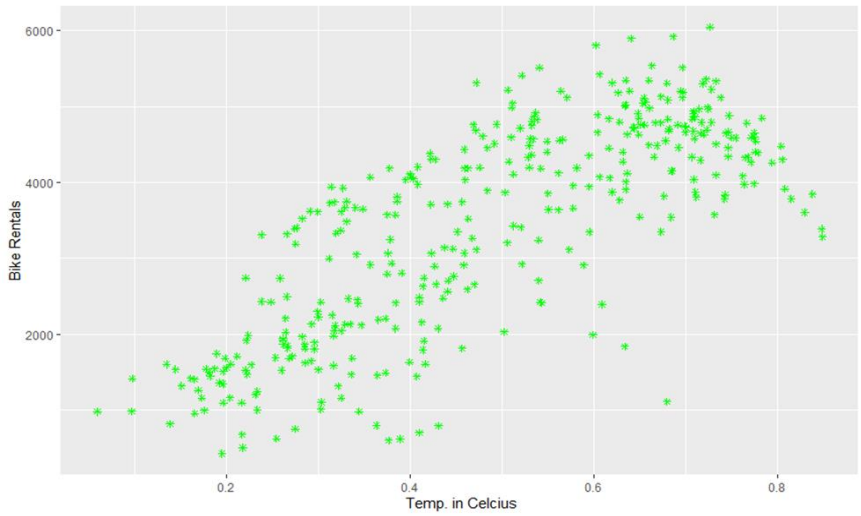

Figure 4:Relationship between bike rental and temperature

Figure 4 illustrates an uphill pattern as the move from left to right, this indicates a positive relationship between the number of total rented bikes and temperature. As the temperature increase, the bike rental tends to increase.

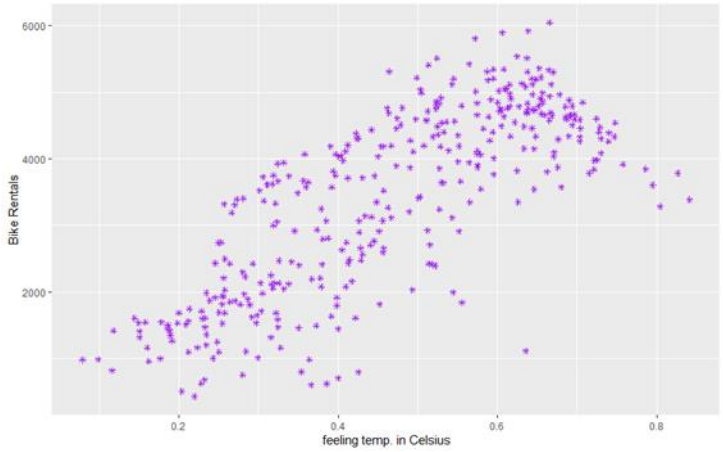

Figure 5:Relationship between bike rental and feeling temperature

The graph shows an uphill pattern as the move from left to right, this indicates a positive relationship between the number of total rented bikes and feeling temperature. As the feeling temperature increase, the bike rental tends to increase.

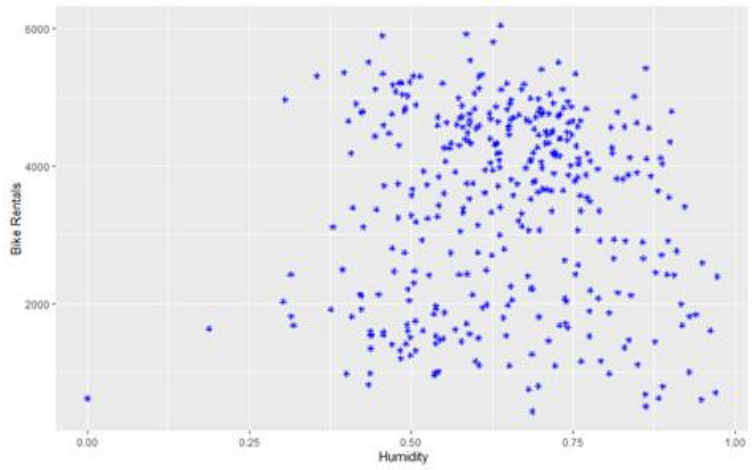

Figure 6:Relationship between bike rental and humidity

The data don't seem to resemble any pattern. Then no relationship exists between humidity and the number of total rented bikes.

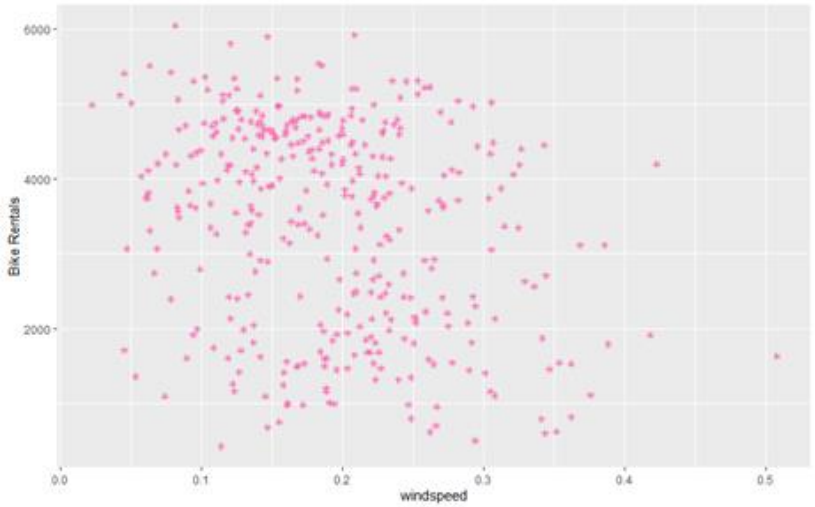

Figure 7:Relationship between bike rental and wind speed

This scatter plot does not seem to resemble any kind of pattern,then no relationship exists between wind speed and the number of total rented bikes.

Exploring qualitative variables vs count of bike rental

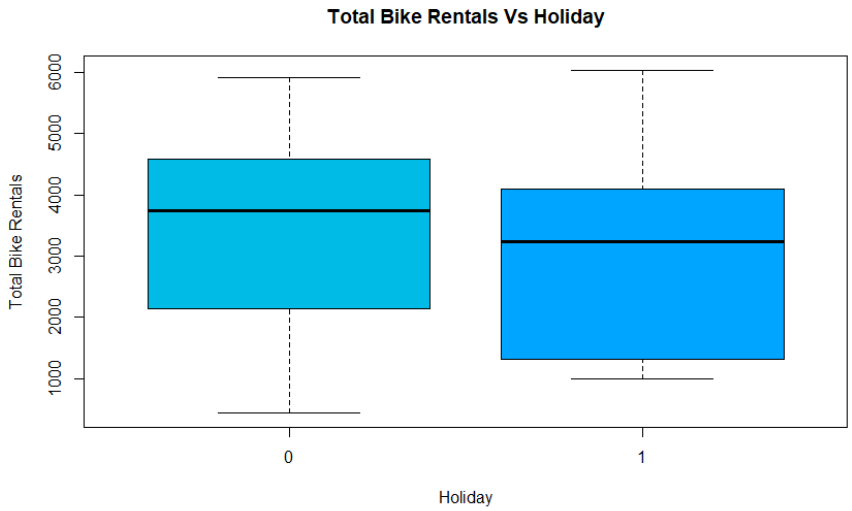

Figure 8:Distribution of total bike rentals according to the holiday

The plot shows the relationship between the Total Bike Rentals(cnt) variable and holiday. The average number of bike rentals on a working day is higher than on holiday. And there are no outliers present on holiday/working days.

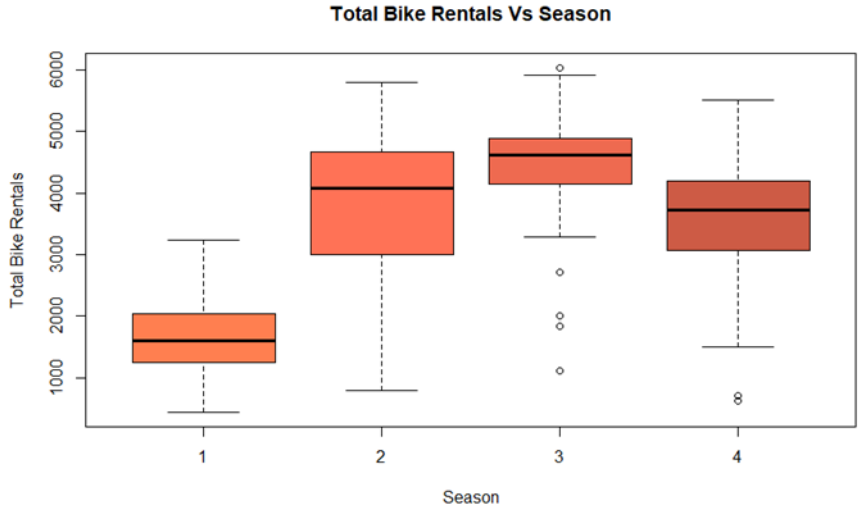

Figure 9:Distribution of total bike rentals according to the season

The plot shows the relationship between the Total Bike Rentals(cnt) variable and the season. The average bike rentals are the highest during summer(2) and lowest during spring(1). Moreover, outliers are higher in the fall season(3).

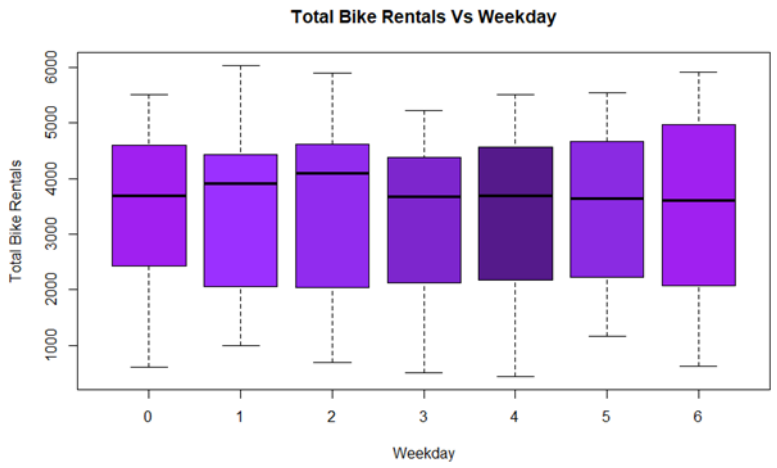

Figure 10:Distribution of total bike rentals according to the weekday

Figure 10 illustrates the relationship between the Total Bike Rentals(cnt) variable and weekdays. The average numbers of bike rentals are the highest on Monday, Tuesday, and Saturday. Except for these days, other days have approximately equal average numbers of bike rentals.

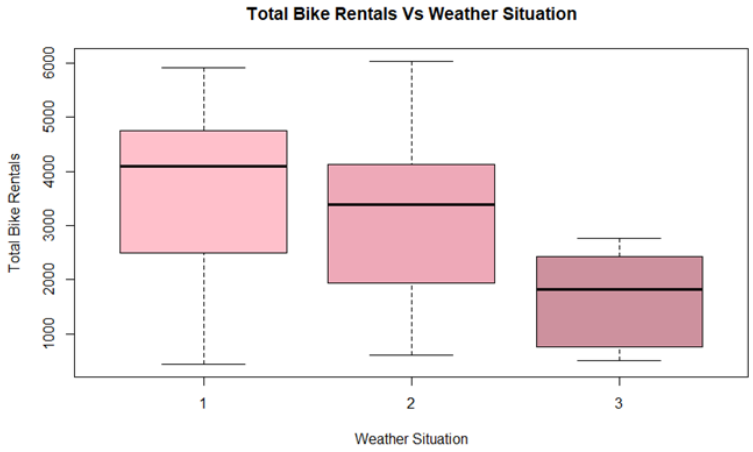

Figure 11:Distribution of total bike rentals according to the weather situation

The plot shows the relationship between the Total Bike Rentals(cnt) variable and the weather situation. The average numbers of bike rentals are the highest when there is a better weather situation and lowest when the situation is bad(3). There is a clear decreasing trend of bike rentals when the weather is bad.

Data analysis

Since the data set consist of qualitative variables using dummy variables are essential when building the model. There are four important qualitative variables such as season, weathersit, holiday and weekday. And parameter interpretation is based on the reference level. In here, season 1 (spring), holiday 0 (weather day is a holiday),weathersit 1 (Clear, few clouds, Partly cloudy), and weekday 0 (Sunday) consider as reference levels when creating the model.

Let Y be the count of total rental bikes.

Let's take X1, X2 and X3 as seasonal variables.

$$X_1 = \begin{cases} 1 \text{ if season 2} \\ 0 \text{ otherwise} \end{cases} \quad X_2 = \begin{cases} 1 \text{ if season 3} \\ 0 \text{ otherwise} \end{cases} \quad X_3 = \begin{cases} 1 \text{ if season 4} \\ 0 \text{ otherwise} \end{cases}$$

Let's take W1 and W2 as weather situation variables.

$$W_1 = \begin{cases} 1 \text{ if weathersit 2} \\ 0 \text{ otherwise} \end{cases} \quad W_2 = \begin{cases} 1 \text{ if weathersit 3} \\ 0 \text{ otherwise} \end{cases}$$

Let's take Z1 as holiday variable.

$$Z_1 = \begin{cases} 1 \text{ if holiday 1} \\ 0 \text{ otherwise} \end{cases}$$

Correlation Analysis and Multicollinearity

Effects of multicollinearity with qualitative variables

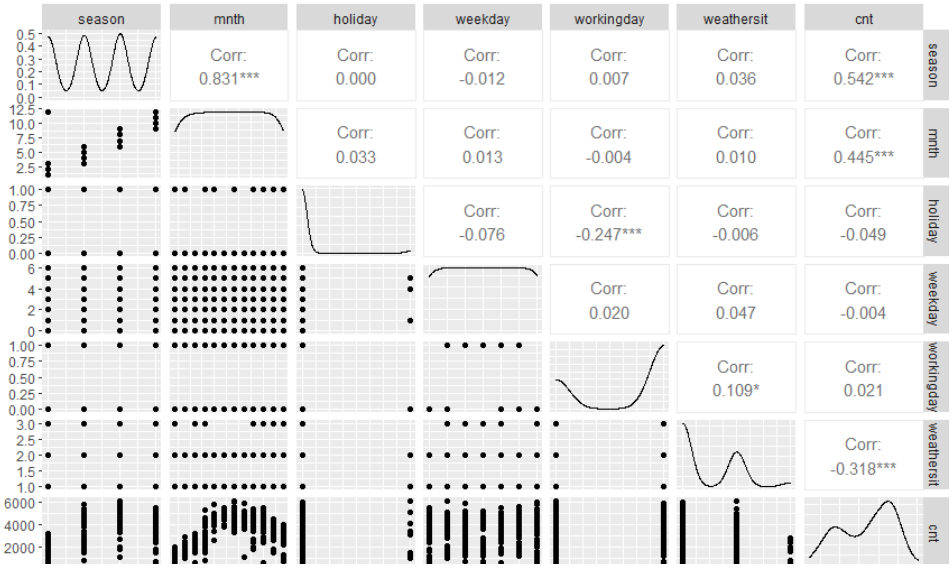

Figure 12: Explore correlation between qualitative variables

According to the correlation matrix, the season is moderately correlated with the count of total rental bikes; also, the month is strongly associated with the season. Other variables are slightly correlated with bike rental. Therefore, months were discarded from our MLR model due to the multicollinearity effect.

Effects of multicollinearity with quantitative variables

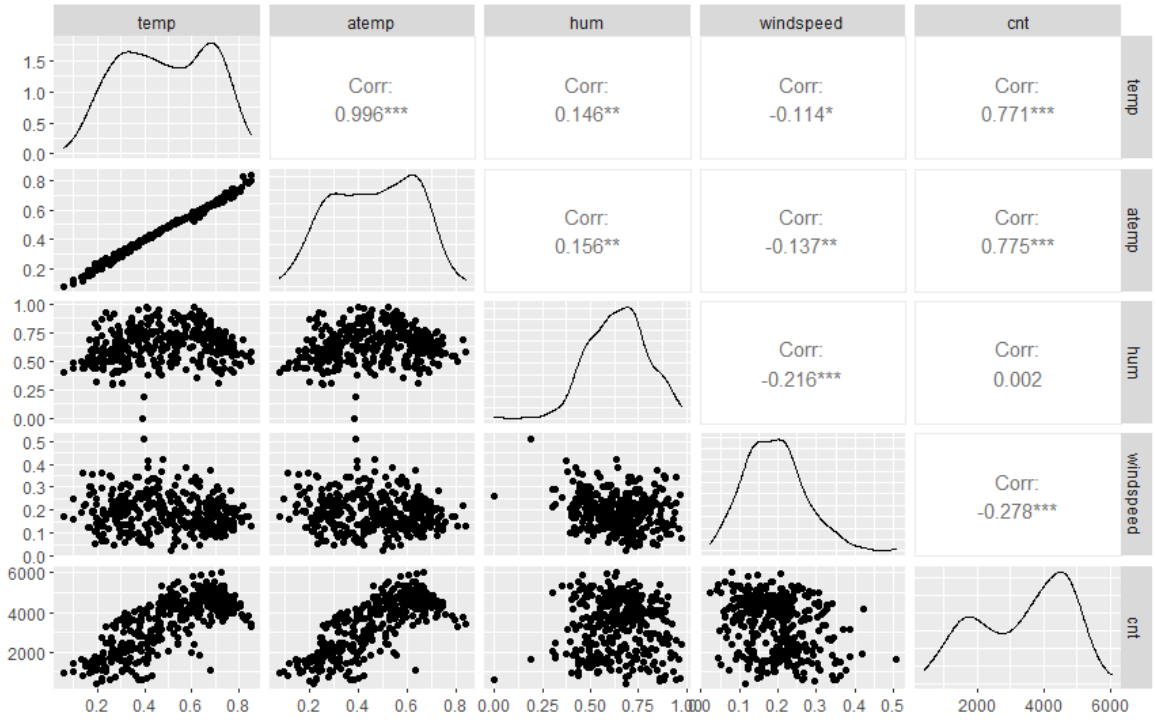

Figure 13:Explore correlation between quantitative variables

From the above correlation matrix, temperature and feeling temperature are more correlated with total rental bikes. Temperature is highly correlated with feeling temperature, but total rental bikes and other variables are slightly correlated. Due to the Multicollinearity effect, feeling temperature was removed from our MLR model.

Multi Linear Regression (MLR) Model

Backward selection method

Model 1
LM formula = Cnt~factor(season) + factor(holiday) + factor(weekday) + factor(workingday) + factor(weathersit) + temp + hum + windspeed

Table 2:Summary of model 1

Coefficient	P-value
Intercept	2.42e-10
Factor (season 2)	< 2e-16
Factor (season 3)	1.37e-08
Factor (season 4)	< 2e-16
Factor (holiday 1)	0.08254
Factor (weekday 1)	0.55322
Factor (weekday 2)	0.30934
Factor (weekday 3)	0.52881
Factor (weekday 4)	0.42484
Factor (weekday 5)	0.17667
Factor (weekday 6)	0.18705
Factor (working day 1)	NA
Factor (weathersit 2)	0.00032
Factor (weathersit 3)	< 2e-16
Temp	< 2e-16
Hum	0.00326
windspeed	9.98e-07

According to the table 2, p-values of weekdays are the largest and greater than 0.15(α_R). Therefore, we drop weekdays from the model.

Model 2
Lm formula = cnt ~ factor(season) + factor(holiday) + factor(weathersit) + temp + hum + windspeed

Table 3:Summary of model 2

Coefficients	P-value
Intercept	1.87e-13
Factor (season 2)	< 2e-16
Factor (season 3)	8.46e-09
Factor (season 4)	< 2e-16
Factor (holiday 1)	0.063952
Factor (weathersit 2)	0.000467
Factor (weathersit 3)	< 2e-16
Temp	< 2e-16
Hum	0.001611
windspeed	7.94e-07

According to the table, the holiday p-value is the largest and greater than 0.05(α_R). Therefore, we drop holiday from the model.

Residual analysis for model 2

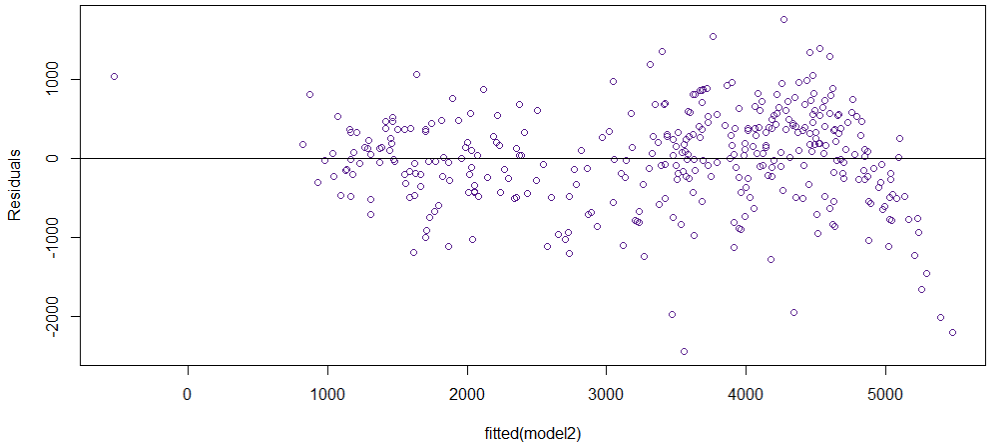

Figure 14:Residual Vs Fitted Values

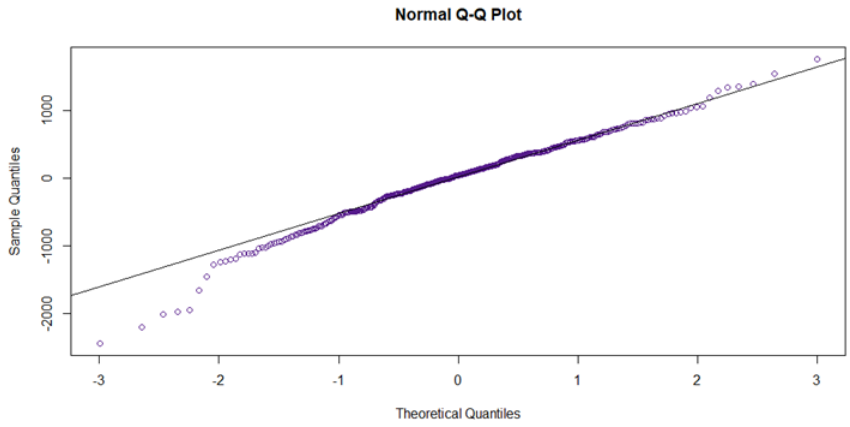

Figure 16:Q-Q plot for residuals

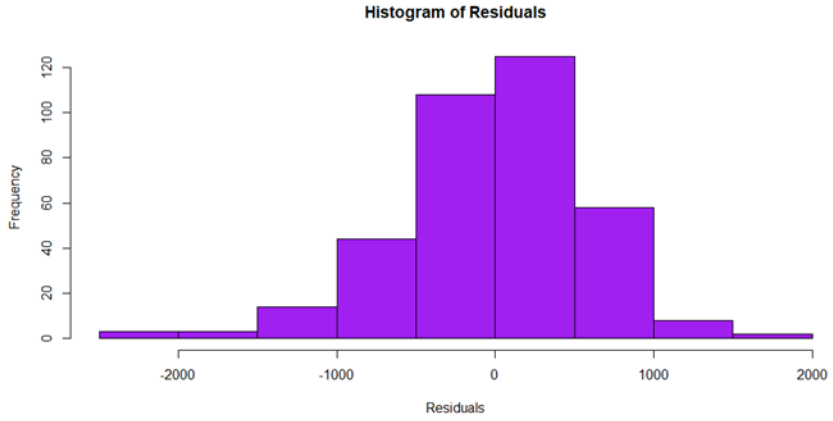

Figure 15:Histogram of residuals

- **Shapiro-Wilk normality test**

  W = 0.97934, p-value = 4.343e-05

- According to the histogram, there is a kind of symmetric pattern, and the normal probability plot shows the pattern symmetrical with heavy tails. But P-value of the Shapiro Wilcoxon test is 1.155e-05 .It is less than 0.05. Therefore, residuals are not normally distributed.
- According to the histogram, it is distributed around 0. Also figure 14 shows that, these values are also distributed about 0, so conclusion can be made that the error term's expected value is 0.
- According to the figure14 , no special pattern can be observed that means the errors have constant variance.

Detecting influential observations

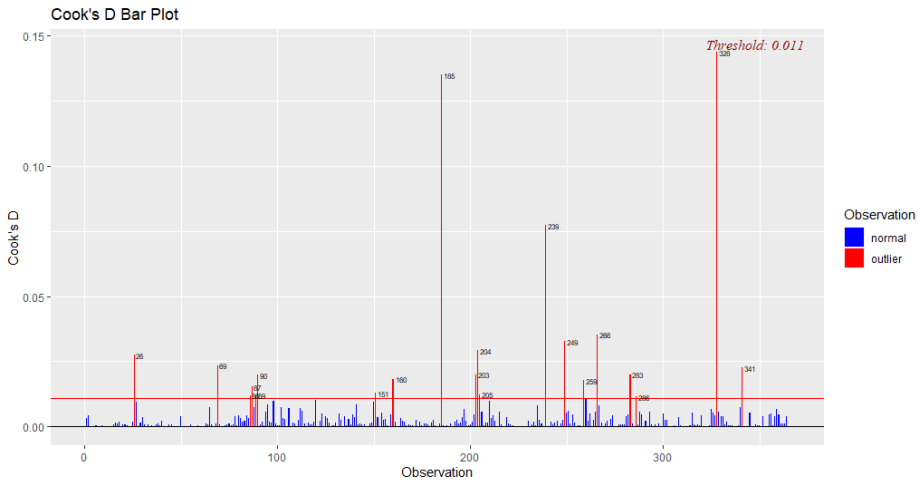

Figure 17:Detecting influential cases

This cook's distance plot illustrates observations with larger Cook's distance (points above the 0.011 line) than other data points. This observation doesn't stand out in other plots. Then we remove these influence observations from our data set and re-fit the model.

Re-fit model after removing influential cases

Model 3

Lm formula = cnt ~ factor(season) + factor(holiday) + factor(weathersit) + temp + hum + windspeed

- R-squared: **0.8578**

Residual analysis for model 3

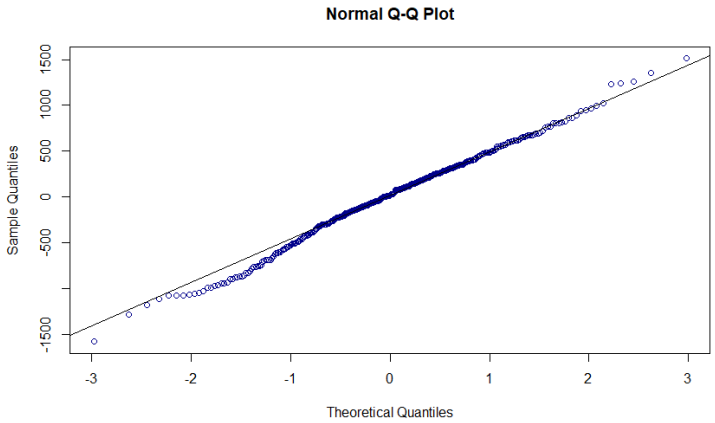

Figure 18:Q-Q plot for residuals

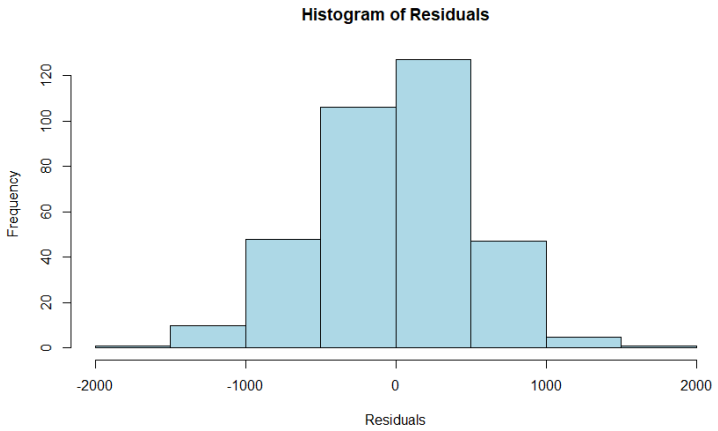

Figure 19:Histogram for residuals

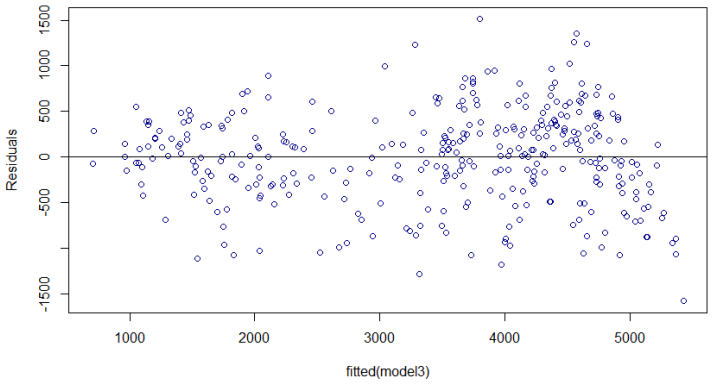

Figure 20:Residual Vs Fitted Values

- **Shapiro-Wilk normality test**

  W = 0.99438, p-value = 0.2356

- According to the histogram, there is a kind of symmetric pattern and in normal probability plot, most of the points have lied along the straight line. Moreover, the P-value of the Shapiro Wilcoxon test is 0. 2356.it is more significant than 0.05. Therefore, residuals are normally distributed.
- According to the histogram, it is distributed around 0. Figure 20 shows that, these values are also distributed around 0, so conclusion can be made that the expected value of the error term is 0.
- According to the figure 20, no particular pattern can be observed that means the errors have constant variance.

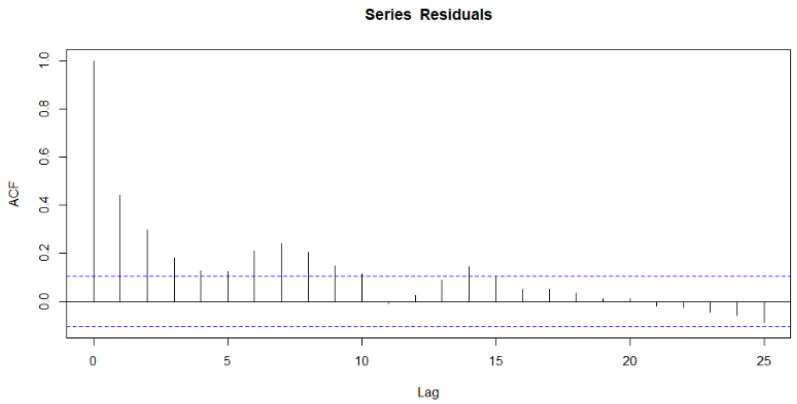

Figure 21:Autocorrelation Function-Correlogram

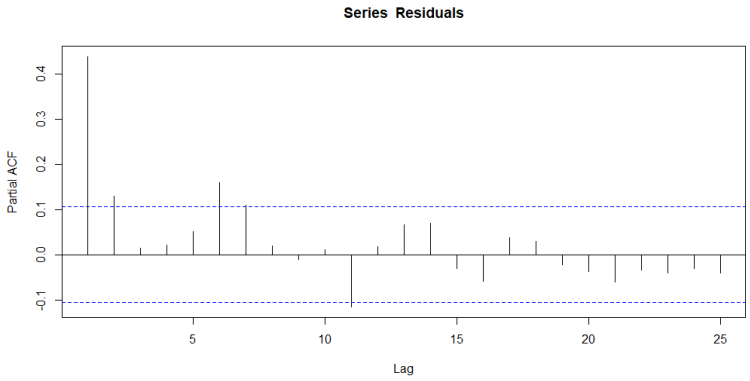

Figure 22:Partial Autocorrelation Function-Correlogram

ACF shows exponential decaying behavior and according to the PACF, there is a lag of 1 since the lag-1 partial autocorrelation is so large and way beyond the "5% significance limits" shown by the blue lines.

Testing for Autocorrelation

Breusch-Godfrey Test

For model 3

LM test = 68.724, df = 1, p-value < 2.2e-16

According to the Breusch-Godfrey test, p- the value is less than 0.05(α). Therefore, the conclusion can be made that there is significant autocorrelation in our regression model.

Cochrane-Orcutt estimation for first-order autocorrelation

.

The correlation was adjusted using the Cochrane-Orcutt procedure. Then following model was gotten as our best model.

Model 4 (Final model)
Count of total rental bikes = 1966.91+975.92 X_1 +840.05 X_2 + 1475.28 X_3 – 204.53 W_1 -1606.93 W_2 -285.86 Z_1+4687.12 temp -1696.31 hum -2003.46 windspeed

- R-squared :**0.7312**

This implies that our fitted model explains around 73% of the total variability of the total number of rental bikes.

Residual analysis for model 4

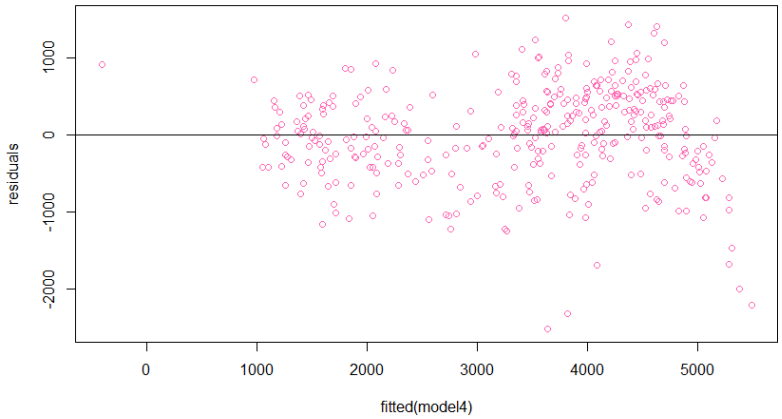

Figure 23:Residual Vs Fitted Values

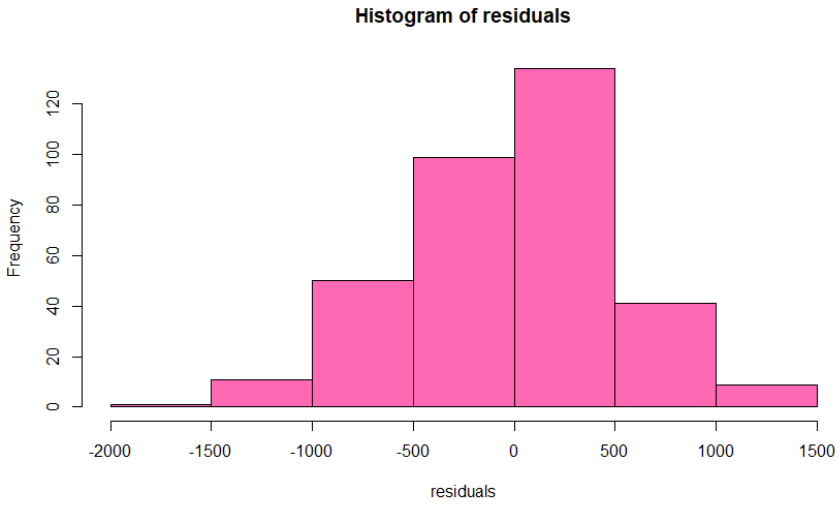

Figure 24:Histogram of residuals

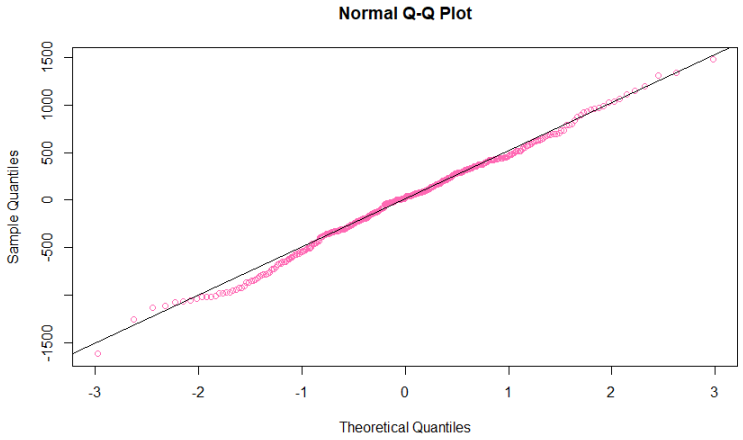

Figure 25:Q-Q plot for residuals

- **Shapiro-Wilk normality test**

W = 0.99517, p-value = 0.3587

- According to the histogram, there is a kind of symmetric pattern and in normal probability plot most of the points have lied along the straight line. Moreover, the P-value of the Shapiro Wilcoxon test is 0. 3587.It is more significant than 0.05. Therefore, the conclusion can be made that residuals are normally distributed.
- According to the histogram, it is distributed around 0 and figure 23 shows that, these values are also distributed around 0. Hence the expected value of the error term is 0.
- Figure 23 obtained that no special pattern can be observed that means the errors have constant variance.

Since assumptions of residuals are satisfied above, model 4 can take as the best-fitted model for this data set.

Model validation

MSE of train data set (2011) = 268912.1

MSPR of test data set (2012) = 4939138.139

MSPR is more significant than the MSE of the fitted model. The MSE of the fitted model is seriously biased and does not give an appropriate indication of the model's predictive ability.

Conclusion and Discussion

The research project's goal was to fit a multiple linear regression model that could be used to predict the number of bike rental users. We anticipated a multilinear relationship between the response variable and the predictors selected for this task.

The effects of environmental and seasonal factors on a rental bike system were investigated in this research work. The relationship between those factors was studied using a variety of approaches. When people choose this mode of transportation, they consider the weather and seasonal effects the most important variables.

This study shows that bike users increase in warmer weather but decrease when humidity levels are high. The number drops significantly during the winter when it is at its peak and temperatures are at their lowest. Windy days do not have the same impact as temperature, but rain and snow cause fewer people to use the system. More outliers were detected in humidity and wind speed. Also, in the fall season, there is a significant number of outliers can be seen.

The study shows that the first two models do not satisfy assumptions of residual analysis; hence, it is not fitted to the data set. After adjusting autocorrelation and removing influential cases, the final model was created. That model was satisfied the assumptions of the error term.

The final model developed in this research could explain nearly 73% of variations in the target variable (number of bike rentals) based on weather conditions, seasons and whether a day is a weekday, weekend, or holiday.

But after validating the model for the 2012 data set, the fitted model is seriously biased and does not give an appropriate indication of the model's predictive ability. Although we found the best-fitted model, it is not suitable for proper model validation. Therefore, we can conclude that multiple linear regression was not ideal for this research data.

For future research, we recommend using time series analysis methods to predict the number of bike rentals since the Highly significant autocorrelation structure in the residuals from regression models. For model validation, using 80% and 20% data might be led to unbiased validation.