STA 319 2.0 Advanced Regression Analysis

Project on

Predication of bike rental count daily based on the environmental and seasonal settings

Prepared by:

AS2018346

Content

1. Introduction

Bike-sharing systems are an alternate mode of public transportation service. It allows people to rent or borrow bicycles for short or long periods. Because of bike-sharing systems, thousands of people throughout the world enjoy cycling without needing to own a bike every day. Users can borrow a bike from a station near them and return it to another station in the same network near their destination. Bike-sharing programs have many advantages. These systems are helpful to control traffic. Also, it is a healthy and cheap mode of transportation. Furthermore, bike-sharing systems have a positive impact on the environment. However, different environmental situations play a significant role in their frequent usage. Poor weather conditions and seasonal effects can cause a drop in the use of these systems while great weather conditions can cause frequent use of these systems which may lead to problems in maintaining the bikes.

The objective of this report is to investigate how environmental factors such as weather conditions, rainfall, day of the week, season would affect the number of bicycle rentals per day. And to introduce a multiple linear regression model for predict the bike rental count daily based on environmental and seasonal factors.

1.1 Description of data

The dataset shows data corresponding to daily bike rental count for years 2011 and 2012.

Variable name	Description	Type
instant	Record index	
dteday	Date	
season	Season 1 – Spring 2 – Summer 3 – Fall 4 – Winter	Qualitative
yr	Year 0 – 2011 1 – 2012	Qualitative
mnth	Month 1 – January 2 – February . . . 12 – December	Qualitative
holiday	Whether day is holiday or not 0 – Not a holiday 1 – Holiday	Qualitative

weekday	Day of the week 0 – Sunday 1 – Monday 2 – Tuesday 3 – Wednesday 4 – Thursday 5 – Friday 6 – Saturday	Qualitative
workingday	Whether day is working day or not 0 – Not a working day 1 – Working day	Qualitative
weathersit	Weather situation 1 – Clear, Few clouds, 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	Qualitative
temp	Normalized temperature in Celsius	Quantitative
atemp	Normalized feeling temperature in Celsius	Quantitative
hum	Normalized humidity	Quantitative
windspeed	Normalized wind speed	Quantitative
cnt	Count of total rental bikes	Quantitative

2. Methodology

To predict the daily usage of bike rental count based on environmental and seasonal settings a multiple linear regression model was fixed. Count of total rental bikes(cnt) was considered as the response variable. Data set was split in to two parts as training set and testing set. Year 2011 data was used as training data set which then used to build the model. Testing set which contain data for year 2012 was used to validate the model. R software was used to analyse the data.

As the first step exploratory data analysis was done using the training data set. Bar charts were drawn to find how total rental bike count differ according to variables 'season', 'month', 'holiday', 'working day'and 'weather situation'. Then to find the relationship between total rental bike count(cnt) with the quantitative variables 'temp', 'atemp', 'hum', 'windspeed', scatter plots were drawn. Boxplots were drawn for each quantitative variables to identify the possible outliers. Finally, correlation analysis was done for quantitative variables and qualitative variables separately to detect multicollinearity. Pearson correlation coefficient was taken between each quantitative variables and Cramer's V correlation was taken between qualitative variables.

The multiple linear regression model was developed after analysing the possible association among the variables and removing few that could cause multicollinearity. The lm() function in R was used to create the regression model and function factor() was used to convert the categorical variables to factors. For each categorical variable, R considered the following as reference level.

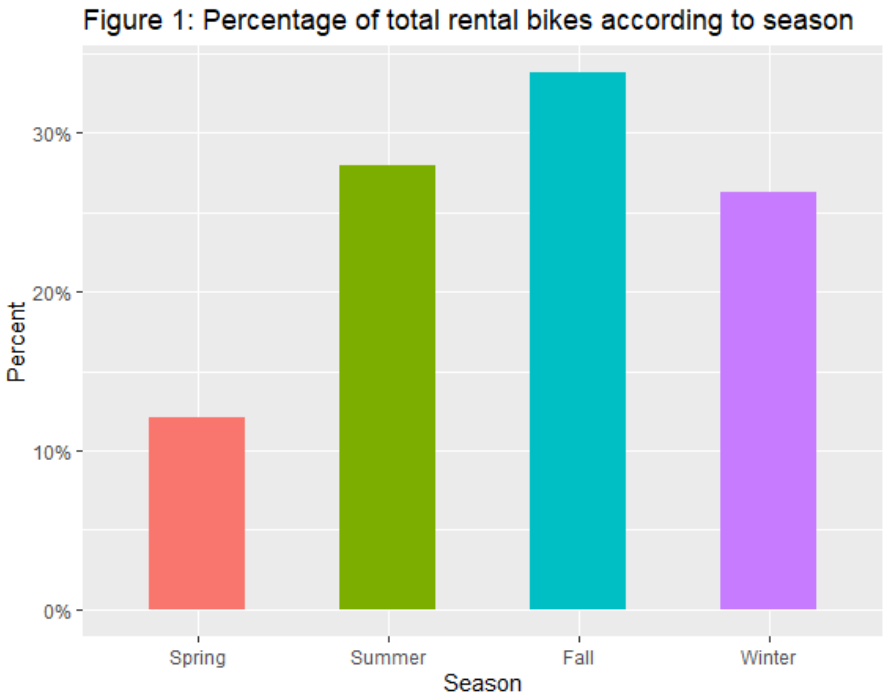
- season – season_1
- holiday – holiday_0
- workingday – workingday_0
- weathersit – weathersit_1

Then summary() function was used to get the outputs. The best model was chosen using the backward elimination method. Variables were dropped from the model if the p-value is greater than the alpha to remove that is 0.15 ($\alpha_R = 0.15$). Using Cook's distance influential cases was identified. After removing the influential cases model was refitted using the backward elimination method. Residual analysis was done using residual plot against the fitted values, histogram of residuals and normal probability plot of residuals to identify whether the model assumptions were satisfied. Further to test the normality of the residuals Shapiro-Wilk normality test was considered. Lastly the model was tested for the occurrence of autocorrelation using Breusch-Godfrey test. If the p-value is greater than the significance level 0.05, it was considered that there is no autocorrelation.

As the final step model validation was done by using testing data set. Each case in the testing set was predicted by using the fitted model. Mean of squared prediction error (MSPR) was calculated for the testing data. Then MSPR value was compared with the MSE of the fitted model. If MSPR value is close to the MSE of fitted model it was concluded that fitted model was appropriate for predict the bike rental count.

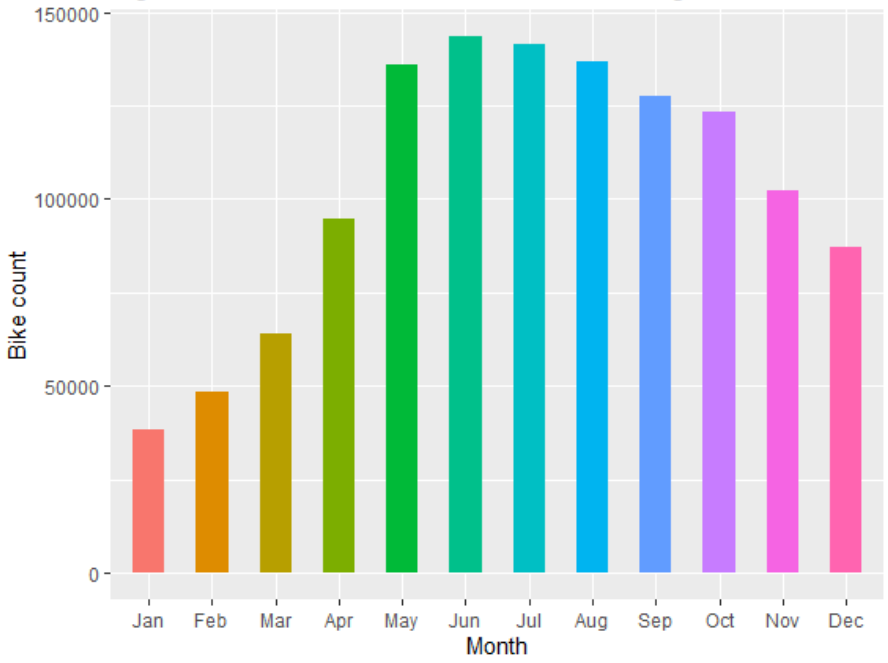
3. Data Exploration

3.1 Total rental bike count with qualitative variables



More than 30% of total bikes were rented in fall season while more than 25% of bikes were rented in summer and winter season. However lower number of bikes were rented in spring. According to figure 1 there is a high usage of rental bikes in summer, fall and winter.

Figure 2: Total count of rental bikes according to month



Higher number of bikes were rented in the middle of the year during May to September. Compared to the latter part of the year there were lower number of rental bikes in the beginning of the year.

Figure 3: Total count of rental bikes according to holiday and working day

Usage of rental bikes were highest when the day is not a holiday and a working day. According to figure people have been using bike sharing systems on high traffic days.

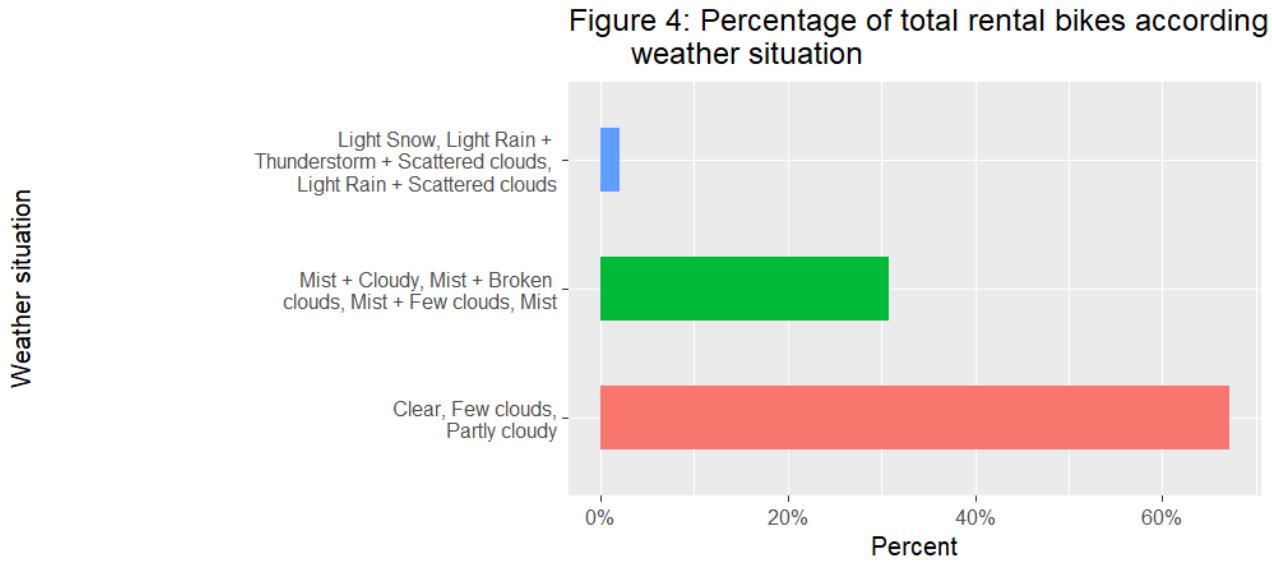


Figure 4 shows that above 60% of total rental bikes were used when there was a clear weather situation. However least number of bikes were used in snowy and rainy weather.

3.2 Total rental bike count with quantitative variables

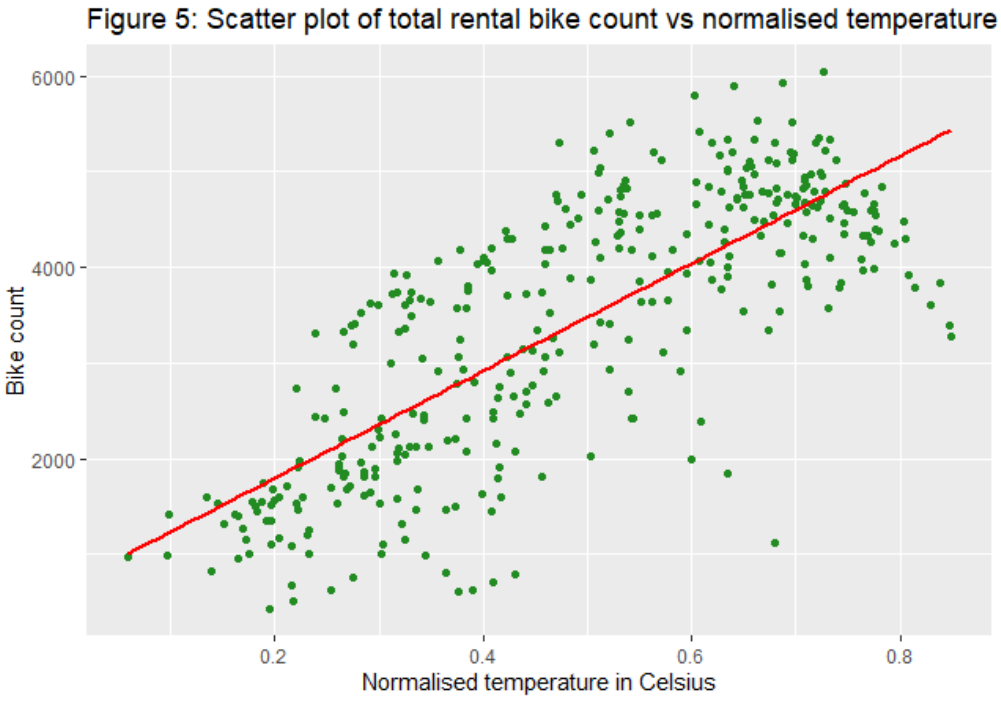
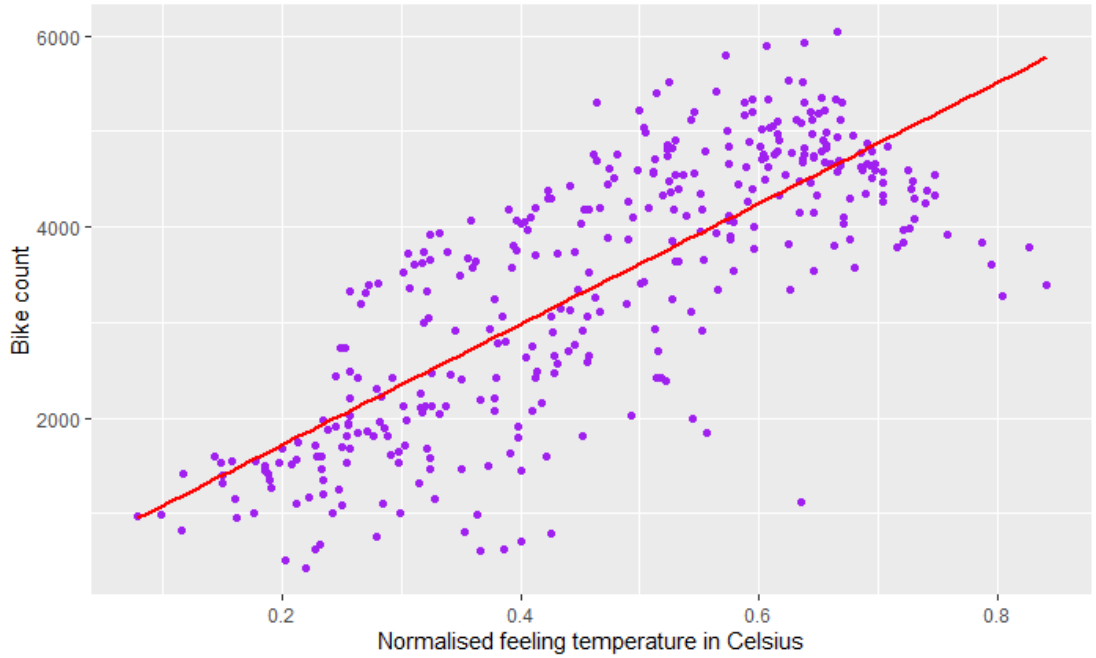
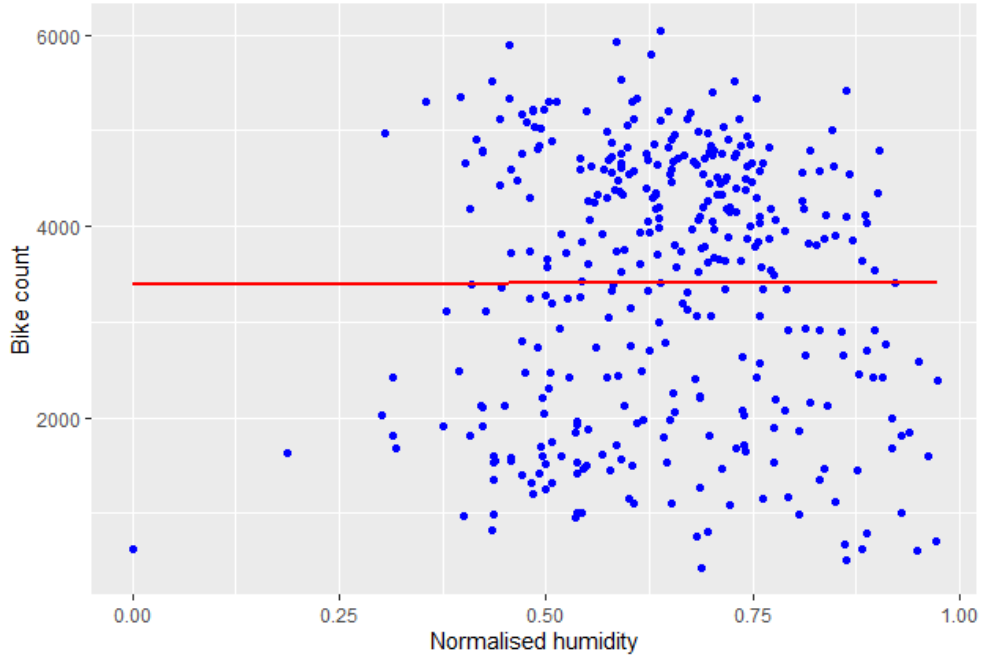


Figure 6: Scatter plot of total rental bike count vs normalised feeling temperature



According to figure 5 and figure 6 normalised temperature and normalised feeling temperature has a positive linear relationship with total rental bike count.

Figure 7: Scatter plot of total rental bike count vs normalised humidity



There is a weak linear relationship between normalised humidity and total rental bike count.

Figure 8: Scatter plot of total rental bike count vs normalised wind speed

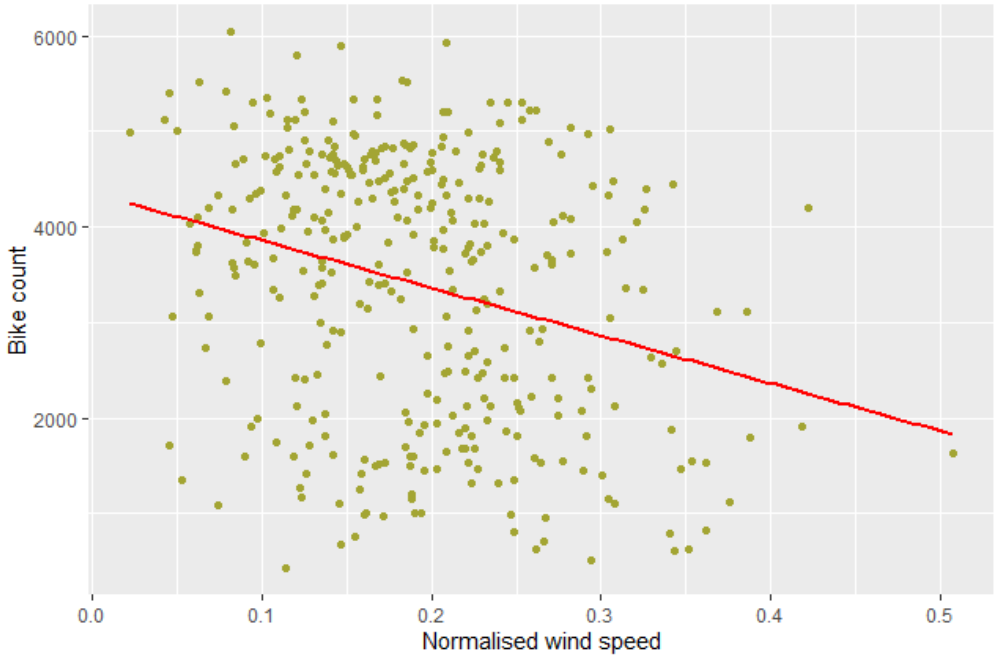
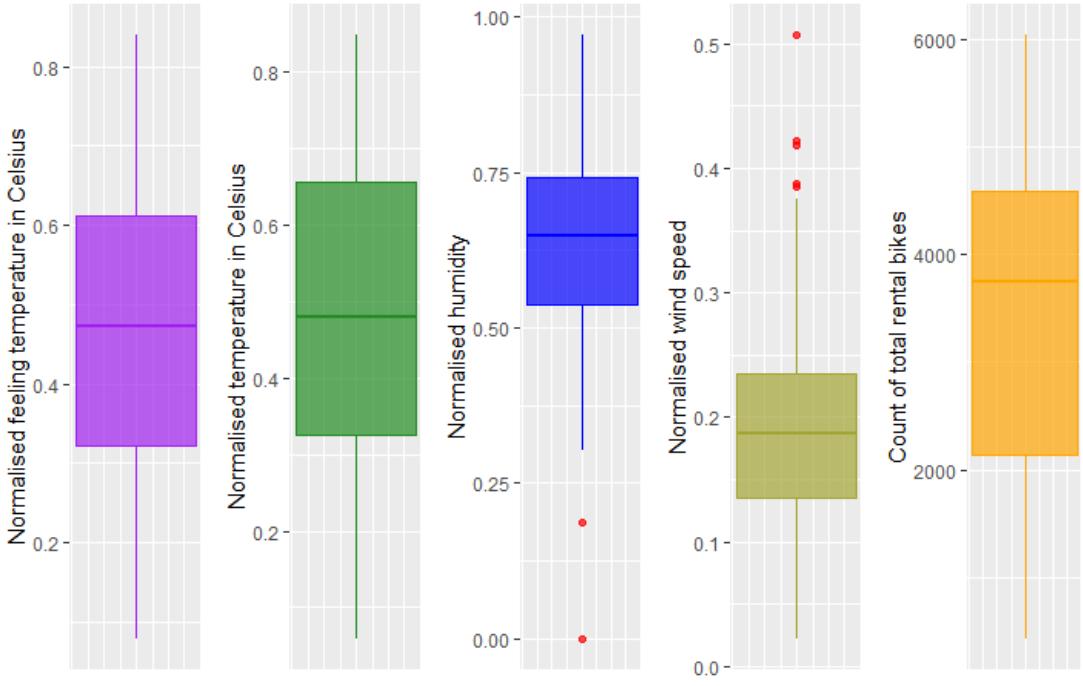


Figure 8 shows that there is a negative linear relationship between normalised wind speed and total rental bike count.

3.3 Identifying Outliers

Figure 9: Outliers



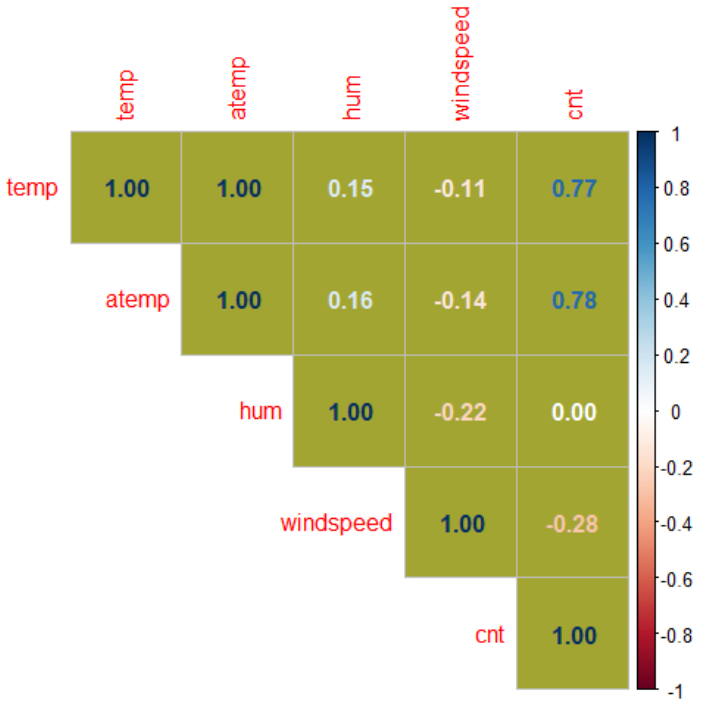
According to figure 9 normalised humidity and normalised wind speed contain extreme values. However, both variables are describing the environmental factors.

Since these factors can take extreme values depending on the day these outlier values are not removed when building the model.

3.4 Correlation Analysis

3.4.1 Correlation analysis for quantitative variables

Figure 10: Correlation Matrix



According to figure 10 variables temp and atemp are positively correlated with cnt while windspeed is negatively correlated with the cnt. Variable atemp is highly correlated with variable temp. Hence variable atemp was removed from the model to prevent multicollinearity.

3.4.2 Correlation analysis for qualitative variables

	season	**mnth**	**holiday**	**weekday**	**workingday**	**weathersit**
season	1	0.897	0.03682	0.02159	0.03161	0.1266
	mnth	1	0.1125	0.04288	0.07561	0.2237
		holiday	1	0.2804	0.2475	0.03841
			weekday	1	0.9414	0.1538
				workingday	1	0.1111
					weathersit	1

According to Cramer's V correlation variables season and month, weekday and workingday are highly correlated. Variables mnth and weekday have more categories than variables

season and workingday. Therefore, variables mnth and weekday were removed when building the model to prevent mulicollinearity.

4. Data Analysis and Results

Multiple linear regression model was created using variables season, holiday, workingday, weathersit, temp, hum, windspeed as predictor variables and variable cnt(total rental bike count) as response variable.

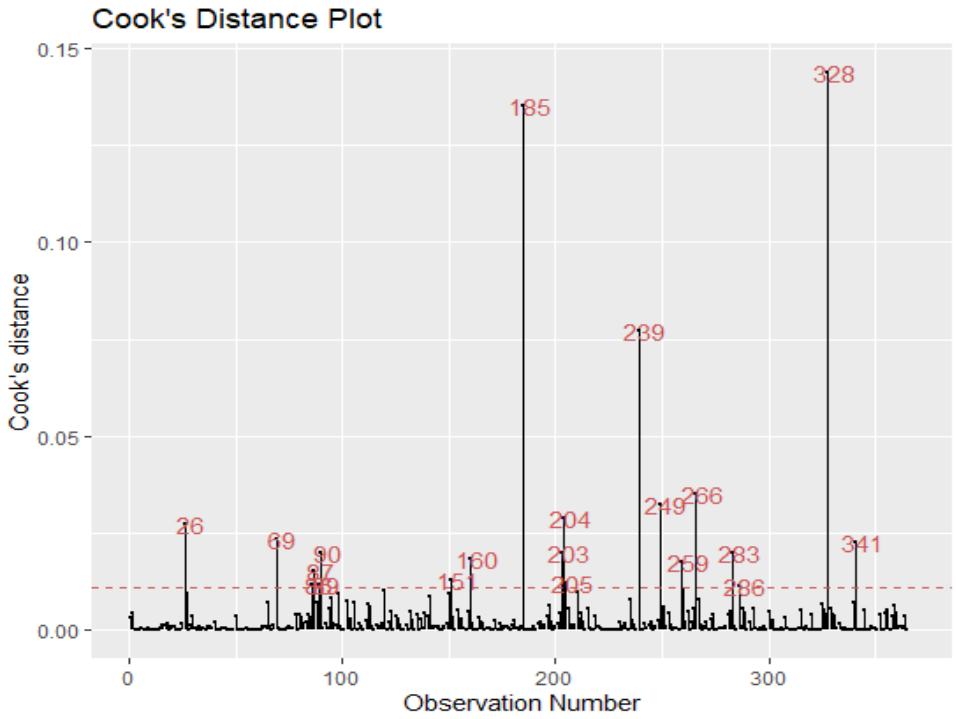
Backward elimination method:

Variables in the model	F^*/t^2	P-value	Conclusion
factor(season)2	$(8.782)^2$	< 2e-16	Variable workingday has the smallest F^* value and largest p-value. Since, p-value > $\alpha_R = 0.15$ variable workingday can be dropped from the model when all other variables are already in the model.
factor(season)3	$(5.902)^2$	8.41e-09	
factor(season)4	$(13.496)^2$	< 2e-16	
factor(holiday)1	$(-1.704)^2$	0.089351	
factor(workingday)1	$(0.376)^2$	0.707511 > $\alpha_R = 0.15$	
factor(weathersit)2	$(-3.547)^2$	0.000441	
factor(weathersit)3	$(-8.736)^2$	< 2e-16	
temp	$(13.724)^2$	< 2e-16	
hum	$(-3.146)^2$	0.001798	
windspeed	$(-5.013)^2$	8.49e-07	
factor(season)2	$(8.786)^2$	< 2e-16	All the variables p-value < $\alpha_R = 0.15$. Hence all the variables should be in the model.
factor(season)3	$(5.901)^2$	8.46e-09	
factor(season)4	$(13.508)^2$	< 2e-16	
factor(holiday)1	$(-1.858)^2$	0.063952	
factor(weathersit)2	$(-3.532)^2$	0.000467	
factor(weathersit)3	$(-8.758)^2$	< 2e-16	
temp	$(13.791)^2$	< 2e-16	
hum	$(-3.178)^2$	0.001611	
windspeed	$(-5.026)^2$	7.94e-07	

Best model:

$$\widehat{\text{cnt}} = 1706.54 + 1068.44(\text{season\_2}) + 941.20(\text{season\_3}) + 1454.63(\text{season\_4}) - 368.47(\text{holiday\_1}) - 302.24(\text{weathersit\_2}) - 1662.29(\text{weathersit\_3}) + 4253.02\text{temp} - 952.11\text{hum} - 2314.68\text{windspeed}$$

Identifying influential cases:



According to Cook's distance plot there are influential cases in the model. Hence model was refitted by removing influential observations.

Refitting the model – Backward elimination method:

Variables in the model	F^*/t^2	P-value	Conclusion
factor(season)2 factor(season)3 factor(season)4 factor(holiday)1 factor(workingday)1 factor(weathersit)2 factor(weathersit)3 temp hum windspeed	$(9.275)^2$ $(5.506)^2$ $(15.916)^2$ $(-2.406)^2$ $(-0.355)^2$ $(-3.148)^2$ $(-9.019)^2$ $(16.796)^2$ $(-4.152)^2$ $(-5.942)^2$	< 2e-16 7.32e-08 < 2e-16 0.01668 0.72275 > $\alpha_R = 0.15$ 0.00179 < 2e-16 < 2e-16 4.19e-05 7.06e-09	Variable workingday has the smallest F^* value and largest p-value. Since, p-value > $\alpha_R = 0.15$ variable workingday can be dropped from the model when all other variables are already in the model.
factor(season)2 factor(season)3 factor(season)4 factor(holiday)1 factor(weathersit)2 factor(weathersit)3 temp hum windspeed	$(9.299)^2$ $(5.523)^2$ $(15.945)^2$ $(-2.388)^2$ $(-3.211)^2$ $(-9.085)^2$ $(16.830)^2$ $(-4.143)^2$ $(-5.947)^2$	< 2e-16 6.71e-08 < 2e-16 0.01748 0.00145 < 2e-16 < 2e-16 4.35e-05 6.88e-09	All the variables p-value < $\alpha_R = 0.15$. Hence all the variables should be in the model.

Best model:

$$\widehat{\text{cnt}} = 1710.84 + 1011.03(\text{season\_2}) + 793.83\ (\text{season\_3}) + 1482.29(\text{season\_4}) - 475.30(\text{holiday\_1}) - 248.08(\text{weathersit\_2}) - 1791.61(\text{weathersit\_3}) + 4764.11\text{temp} - 1197.44\text{hum} - 2425.86\text{windspeed}$$

Summary:

Residual standard error: 518.7 on 335 degrees of freedom
Multiple R-squared: 0.8578, Adjusted R-squared: 0.8539
F-statistic: 224.5 on 9 and 335 DF, p-value: < 2.2e-16

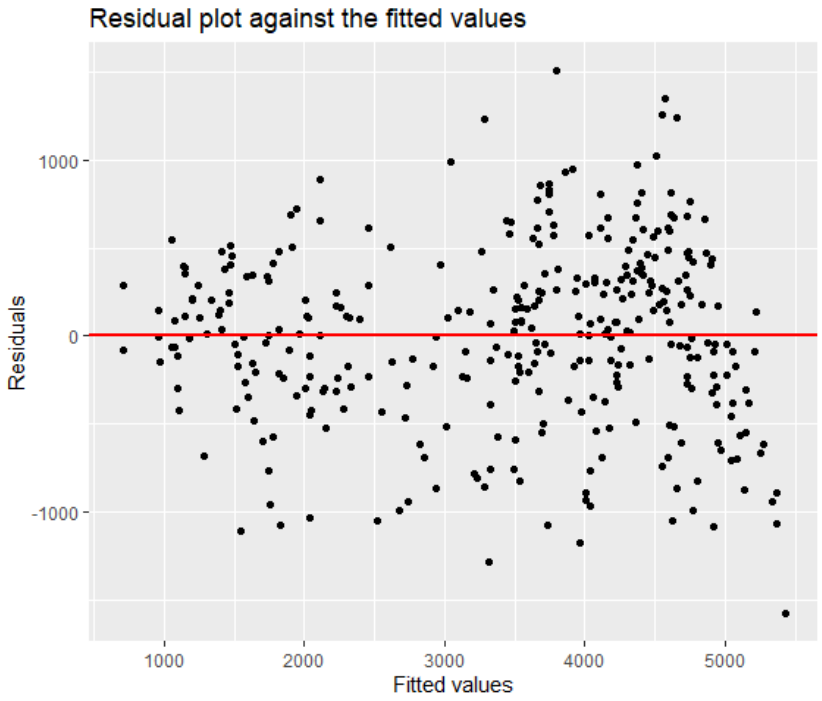
Adjusted R-squared value of the model is 0.8539. This means about 85% of the total variability of count of rental bikes can be explained by the fitted model.

Analysis of Variance Table:

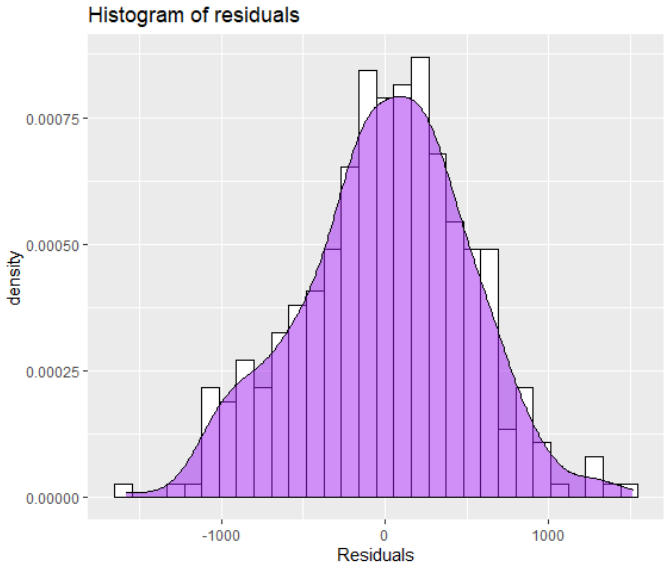
Response: cnt

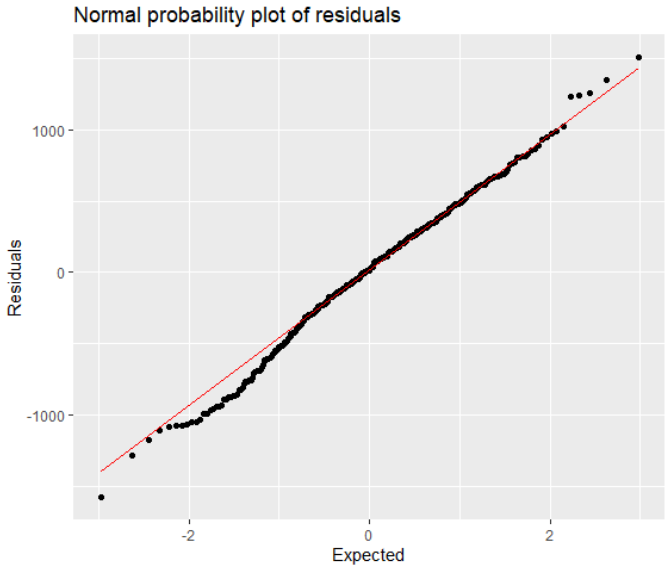
Source of variation	Df	Sum Sq	Mean Sq	F value	Pr(>F)
season	3	396484397	132161466	491.3023	< 2.2e-16
holiday	1	2015608	2015608	7.4929	0.006525
weathersit	2	63180060	31590030	117.4340	< 2.2e-16
temp	1	71029141	71029141	264.0466	< 2.2e-16
hum	1	1247062	1247062	4.6359	0.032024
windspeed	1	9512608	9512608	35.3626	6.881e-09
Residuals	335	90115774	269002		

Residual Analysis:



Residual plot against the fitted values does not show any systematic pattern. Therefore, variance of the error term is a constant.





Histogram of residuals shows a nearly symmetric shape. In normal probability plot of residuals the points are scattered around the straight line. Hence normality assumption of the error term is satisfied.

Test for normality:

H_0: Residuals follow normal distribution

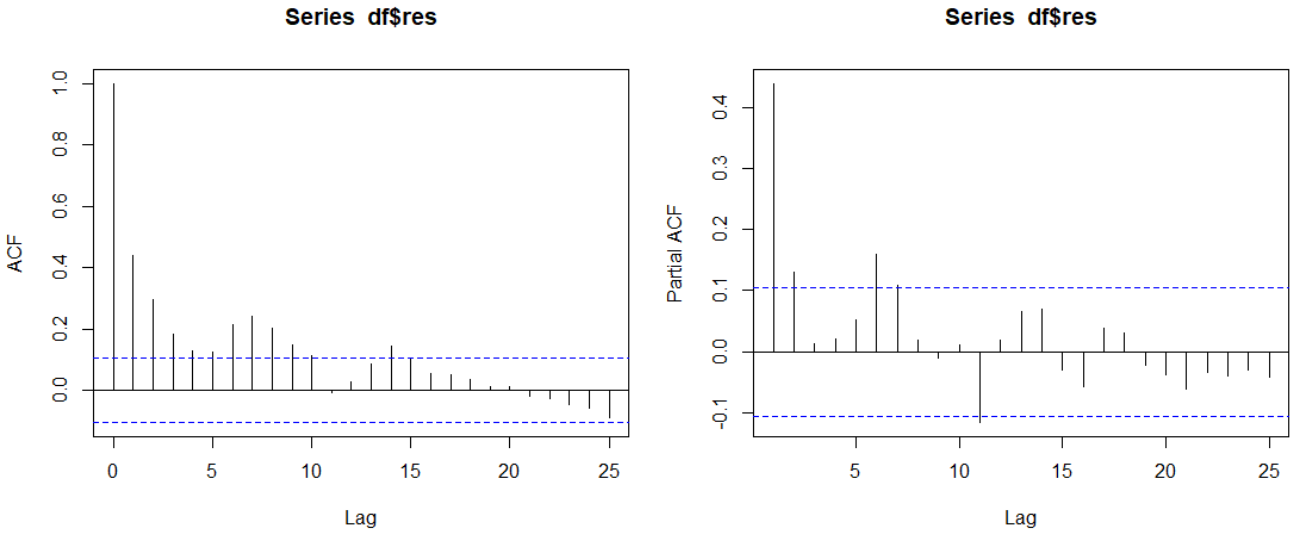
Vs.

H_1: Residuals do not follow normal distribution

Shapiro-Wilk normality test
W = 0.99438, p-value = 0.2356

Since p-value is greater than 0.05, there is no sufficient evidence to reject the null hypothesis. Hence residuals follow a normal distribution.

Testing for autocorrelation:



The ACF shows an exponentially decaying pattern and PACF cut off after lag 2. Therefore error terms are generated by AR(2) process.

Breusch-Godfrey test was performed to test the null hypothesis that there is no autocorrelation.

Breusch-Godfrey test for serial correlation of order up to 2
LM test = 74.571, df = 2, p-value < 2.2e-16

The p-value for the test is less than 0.05 which rejects the null hypothesis of the test. Hence There is a significant autocorrelation.

Model validation:

The fitted model is used to predict each case in testing data set and mean of squared prediction error (MSPR) was calculated.

$$MSPR = \frac{\sum_{i=1}^{366}(Y_i - \widehat{Y}_i)^2}{366}$$

$$MSPR = 4955864$$

Where Y_i is the ith value of the total rental bike count in the testing data set. $\widehat{Y}_i$ is the ith predicted value based on the fitted model.

From ANOVA table the MSE of the fitted model is 269002. However, MSPR value for testing data is 4955864. There is a significant difference between the MSPR and MSE of the fitted model. Hence the model does not show an appropriate indication of the predictive ability.

5. Conclusion and Discussion

The objective of this project was to build a regression model to predict the bike rental count daily based on the environment and seasonal settings and to find how different environment factors affect the number of bicycle rentals per day. Most of the bikes were rented in fall season and when the weather was clear with few clouds or partly cloudy. The data set did not contain any data on bike rental when the weather situation is heavy rain, thunderstorm, mist, and snow. According to the study the number of bike users were increased in warmer temperature however with higher wind speed this number decreases. Moreover, majority of the bikes were rented in the middle of the year and most of the days are working days which means the day is not a weekend or a holiday.

After building the multiple linear regression model it was found that variables season, holiday, weather situation, normalised temperature, normalised humidity, and normalised wind speed were the important variables to predict the count of total rental bikes. The model developed in this project, explained around 85% of the total variability of count of rental bikes. The model assumption that the errors are normally distributed with the constant variance was satisfied in residual analysis.

However, in the model validation it was found that there is a significant difference between the MSPR and MSE of the fitted model. Hence the fitted model is not useful to predict the daily bike rental count based on the environmental and seasonal factors. It is a one draw back of this project. Another limitation of the study is there was a significant autocorrelation in the model. The model developed for predict the daily rental bike count can be improved further by removing the autocorrelation of the residuals using transformed variables. Moreover, the model can be improved by considering the interaction effect of two or more independent variables.

6. References

1. Cycle Simply. 2021. *What Are Bike-sharing Systems? - Cycle Simply*. [online] Available at: <https://cyclesimply.com/bike-sharing-systems/>
2. Mangiafico, S., 2016. *R Handbook: Measures of Association for Nominal Variables*. [online] Rcompanion.org. Available at: <https://rcompanion.org/handbook/H_10.html>