Final Project

STAT 5543, Spring 2023

Regression Analysis of Bike Rental Data

Name: Debanik Chakraborty

Rental bikes are becoming widely popular in many urban cities. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. The prediction of bike count required at each hour has become vital for ensuring the stable supply of rental bikes. The given dataset called “Bikerental.csv” contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information. The target is to make an explanatory model to understand what and how much effect it has on the demand of bikes.

Response of the Analysis will be **RentedBikeCount**, which is the count of bike rented every hour. There are 12 variables in the dataset which we are intended to use as Predictors in the data analysis. These are,

Hour: Hour of the day

Temperature: Temperature in Celsius

Humidity: Humidity in %

Windspeed: - Wind speed in m/s

Visibility: A measurement of visibility at 10m

Dewpoint: Dew point temperature(°C)

SolarRadiation: Solar radiation in MJ/m2

Rainfall: Rainfall in mm

Snowfall: Snowfall in cm

Seasons: Winter, Spring, Summer, Autumn

Holiday: Holiday or No holiday

FunctioningDay: No (Non-Functional Hours) or Yes (Functional hours)

There are **8760 observations** or rows in the dataset, and data are given for every hour in a day (0-23 in Hour column). So, the data is given for a total of 365 days or whole one year.

Data Visualization:

The RentedBikeCount data ranges from 0 to 3556. That means, at some hour there were no bike rented and there was time when rented bike counts went up to 3556 as well per hour, which is huge. The distribution of the Rented Bike counts per hour are not evenly spread. From the histogram in Figure 1 and box plot in Figure 2, we get to know that the distribution is positively skewed, most of the observation are on left tail, many outliers on right tail, the mean seems around 500 visually, 705 to be precise. When I check the box plot for each of the predictors, looking at Figure 3 and Figure 4 I find that SolarRadiation is positively skewed. Rainfall and Snowfall are also a bit skewed to the right. Outliers effect might be prominent. These variables can be a problem if included in the actual model.

Figure 1

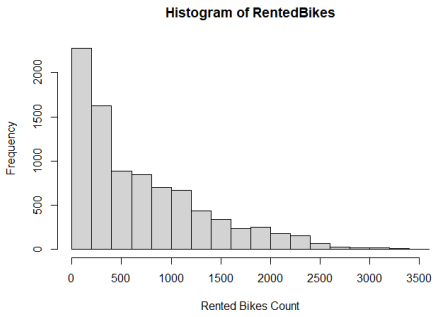


Figure 2

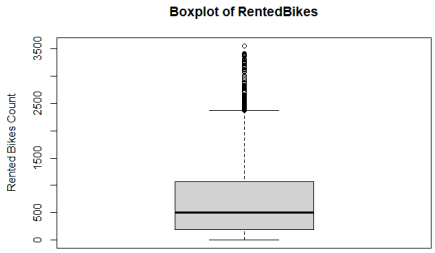


Figure 3

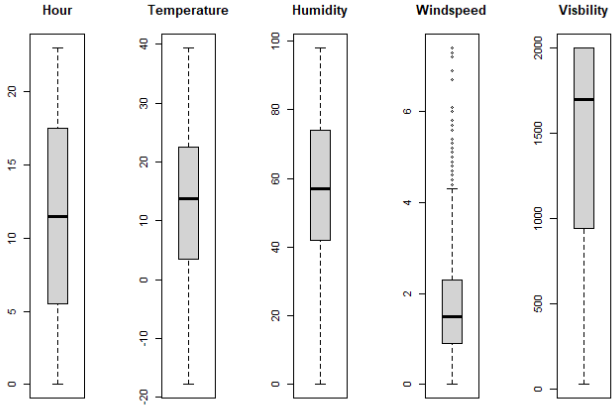
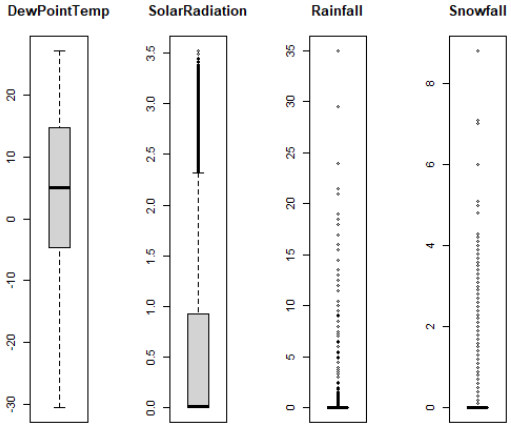


Figure 4



From Figure 5, we can see that Seasons do not seem to have any linear pattern, however, Holiday and FunctioningDay exhibit linear pattern to some extent. Still, I considered all these three as categorical variables and further checked their number of observations and all their possible combinations in tables. Here I find that the two combinations of Seasons and FunctioningDay are exhibiting zero observation.

	FunctioningDay	
Seasons	No	Yes
Autumn	247	1937
Spring	48	2160
Summer	0	2208
Winter	0	2160

That means that there is not sufficient sample for No FunctioningDay of Summer and Winter Seasons. I should not have considered any interaction between Seasons and FunctioningDay, even if the step function suggests.

From the Pairs plot in Figure 6 it seems like many predictors(non-categorical) are positively inter-related and some response vs predictor plot shows non-linear pattern. Even the inter correlation between them shows that there is strong positive Correlation between Dewpoint-Temperature (0.91), Humidity-Dewpoint (0.53), strong negative correlation between Humidity-Visibility(-0.54) and medium negative correlation between Humidity-SolarRadiation(-0.46).

Figure 5

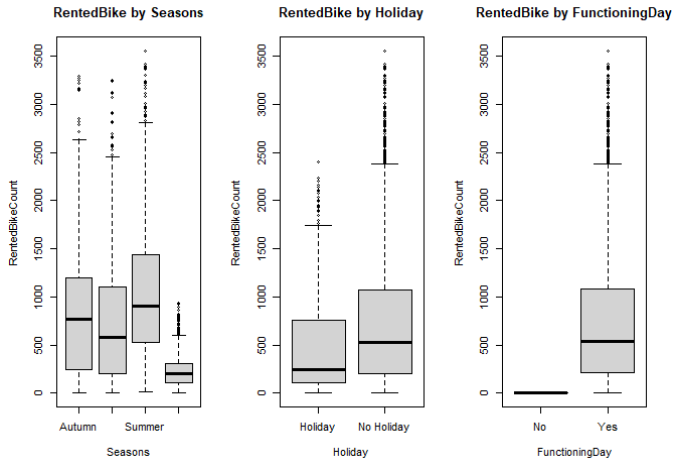
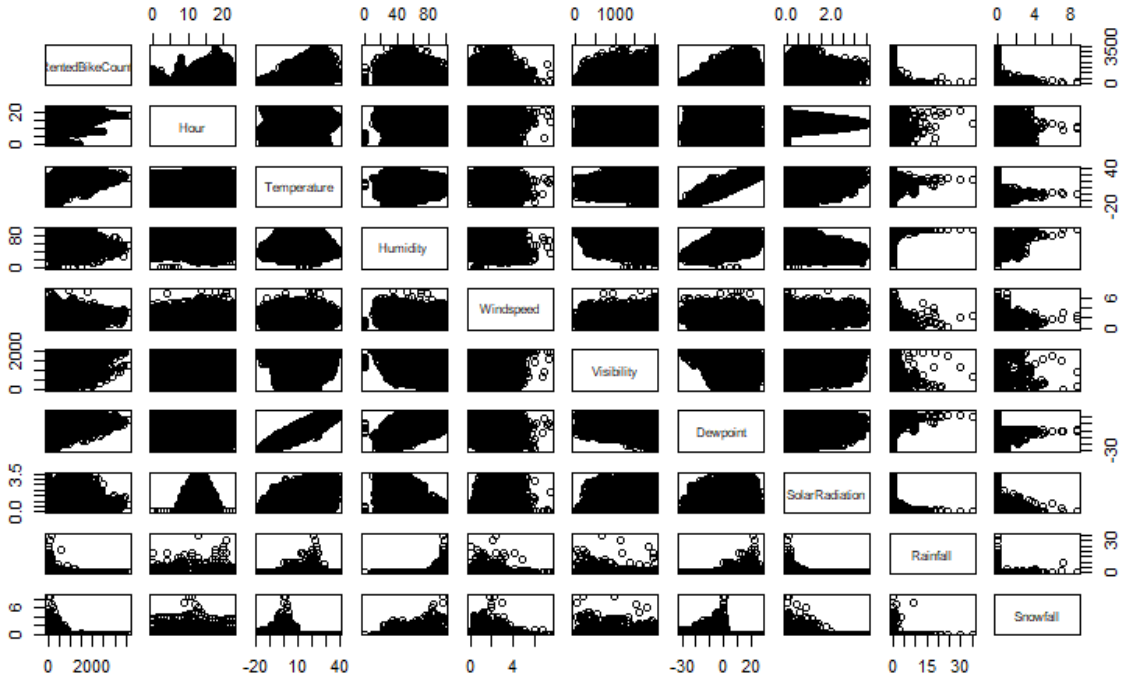


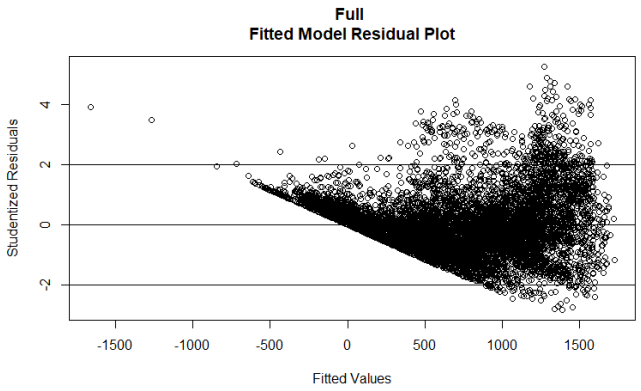
Figure 6



Fitting the Initial full Model and Model Diagnostic:

I started fitting the initial model with all 12 predictors and none of the beta shows nan value, which means there are no two perfectly correlated predictors, and R squared is 0.5504, which is not so high, indicating that the model needs improvement. Fitted values vs studentized residuals plot (Figure 7) shows the presence of Funnel Shape as well as a pattern in fitted value vs Error, which means there is presence of Non-linearity & Heteroscedasticity.

Figure 7



In the individual predictors vs studentized residuals plots (Figure 8 and 9), Temperature and Dewpoint are exhibiting funnel shape, SolarRadiation is exhibiting a reverse funnel shape, so these three are heteroscedastic. Rest is showing either unclear pattern or very weak heteroscedastic pattern which is negligible (like Windspeed). For Rainfall and Snowfall, predictor vs residual plots show that the outlier effect is very prominent, which created heteroscedasticity. However, in Figure 10 we can see that the categorical variables show slight but negligible heteroscedasticity, so these predictors are all right. In short, most of these heteroscedasticities, whichever predictors have, can be due to non-linearity as we saw in pairs plot.

Figure 8

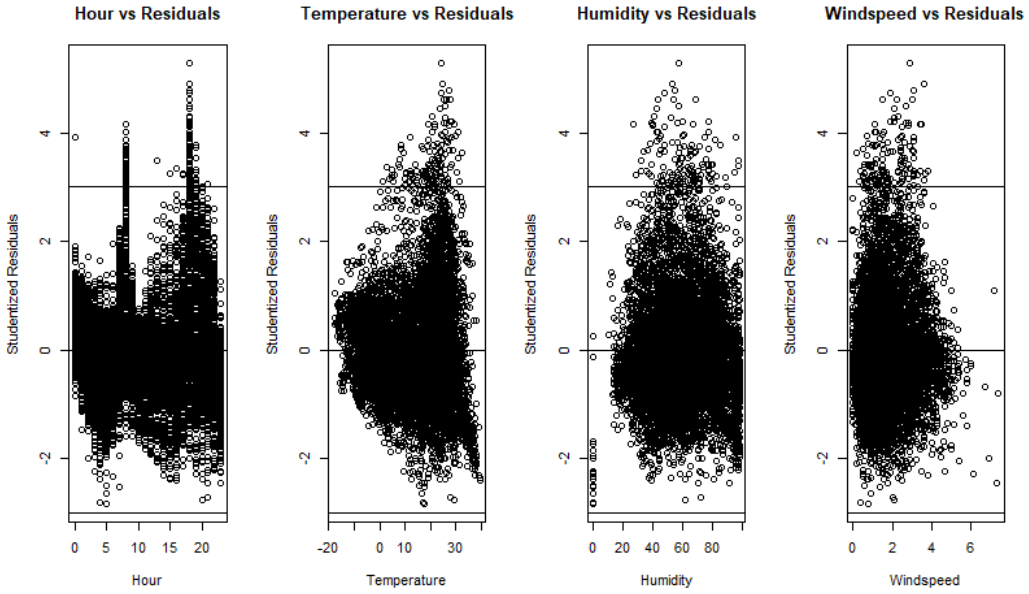


Figure 9

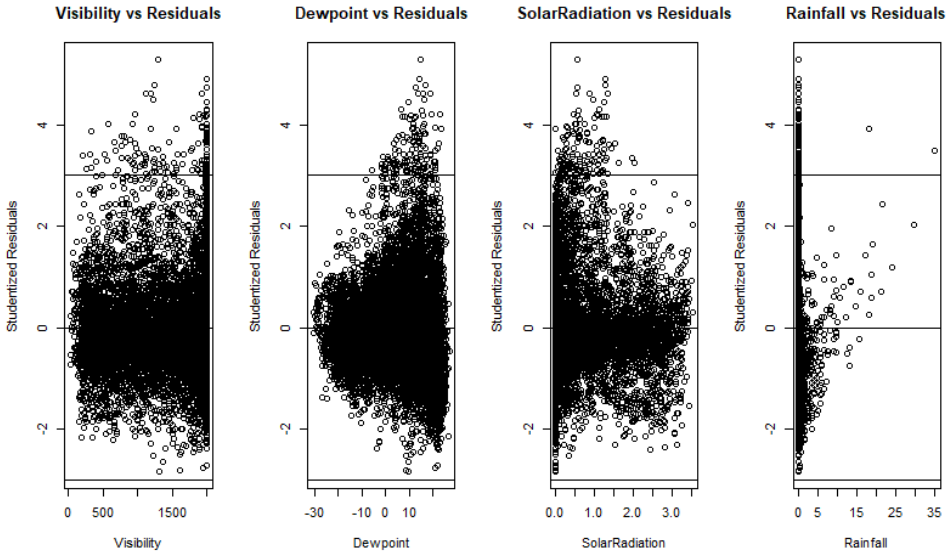
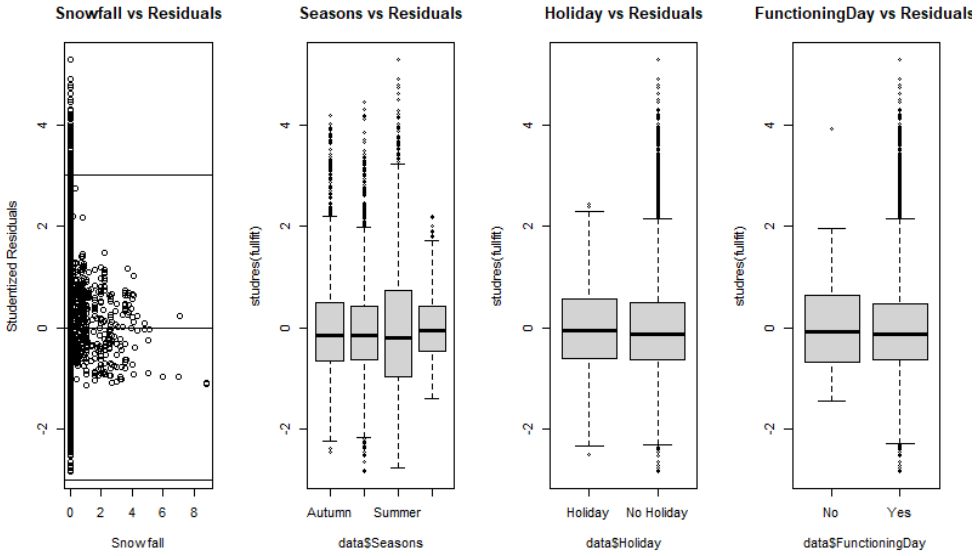


Figure 10



When I did QQplot(Figure 11), there was departure in right tail almost from the midway which indicates the mode is having non-normality of errors. I determine the correlation of two axis of this QQplot, and that is below my threshold of declaring normality 0.9751< 0.9850. So residuals are not normal enough.

As these data were collected every hour in a year, so error checking by time sequences seemed important to me, so I did residuals vs data sequence (1,2,3…n) plot (Figure 12) and saw that residuals are independent of each other, though an unclear pattern was observed. Then I plot the cook's distance (Figure 13) and find that even the highest value is 0.1055, and most of the values are way below that maximum value. There is no influential data point to worry about. Then I find the variance inflation

factor (VIF) and find that 3 values are outside our threshold 10, and among them Dewpoint has the highest VIF/GVIF of 117.30. We must drop this variable from the predictor list as it is showing high multicollinearity.

Figure 11

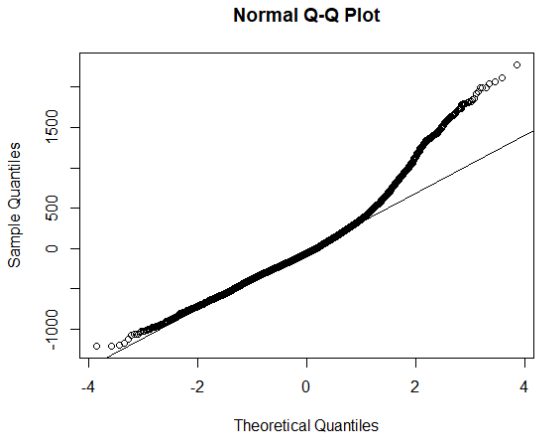


Figure 12

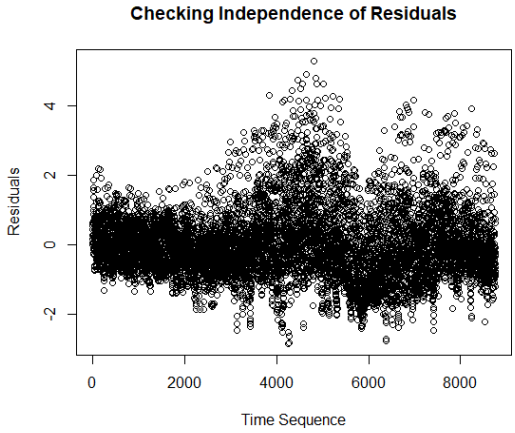
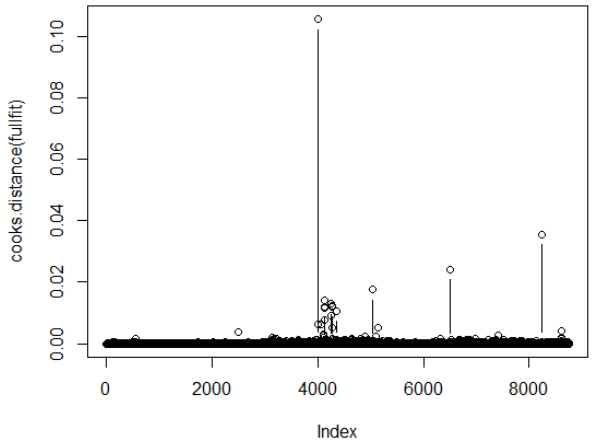


Figure 13



VIF Table

	GVIF	Df	GVIF^(1/(2*Df))
Hour	1.209577	1	1.099808
Temperature	89.477069	1	9.459232
Humidity	20.553911	1	4.533642
Windspeed	1.303644	1	1.141772
Visibility	1.689144	1	1.299671
Dewpoint	117.298694	1	10.830452
SolarRadiation	2.034617	1	1.426400
Rainfall	1.085306	1	1.041780
Snowfall	1.119845	1	1.058227
factor(Seasons)	5.526992	3	1.329683
factor(Holiday)	1.023340	1	1.011603
factor(FunctioningDay)	1.080974	1	1.039699

I dropped the Dewpoint from the Bikerental dataset and fitted the model with response and rest of the predictors once again. R square remains almost the same (0.55) in this new model, indicating that model did not improve much in terms of explaining the variability of response. I followed all the diagnostic again, and the Fitted Value vs Residuals and Predictors vs Residual plots remain same as before, indicating that nonlinearity and heteroscedasticity still remains in errors both against the fitted values and some predictors. QQplot remains the same, highest value of cook's distance (Appendix B) is showing (0.118) which is almost same as previous value (0.1055). When I check the VIF (Appendix B), all the GVIF values are below 10, so seems fine.

To fix heteroscedasticity due to non-linearity and non-normality of error, first I did log transformation of response, which was positively skewed. However, while fitting the model I got an error due to presence of negative infinite value in 'logRentedBike' column (which is the newly added log transformed column of response), which directed to the fact that actual response is having many values (295) equal to 0. So, it was clear that direct log transformation will not work for the dataset. However, I cannot drop those 0 values, because those values can also be useful, as we need to know and include when the demand for bike is absent as well. So, I tried the square root transformation of response. We cannot have any negative value as response (as it is count of bikes), so square root transformation is okay in here. I added a column named 'SqrtRentedBike' in new dataset. Figure 14 shows the histogram and box plot of the square root transformation of response, which seems more symmetric than actual response, and presence of outliers are almost zero in the transformed distribution.

Figure 14

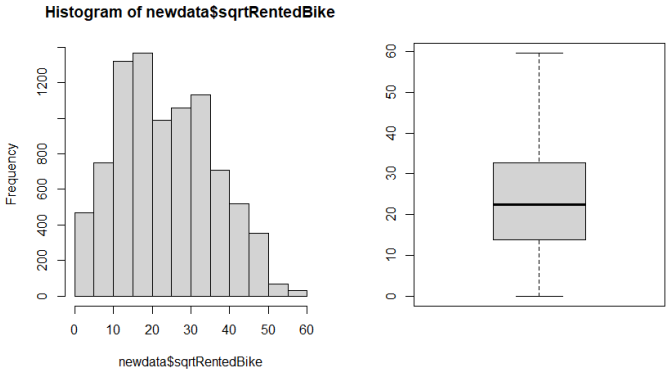
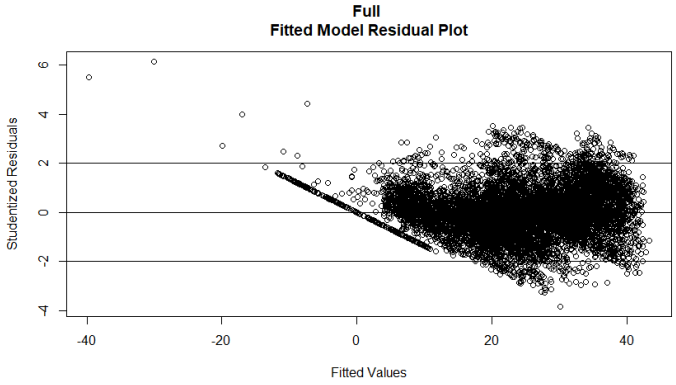
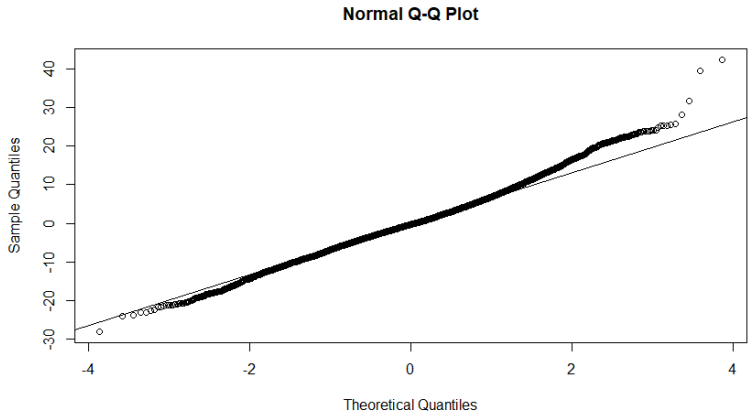


Figure 15



I fitted the model, using SqrtRentedBike column as response and 11 predictors (without Dewpoint) and found the R squared to be 0.65, which indicates the model has improved in terms of expressing variability in response. I did Fitted Values vs Studentized Residuals (Figure 15), and it shows that for a few data points there is a linear pattern (negative correlation) is still there, however, for most of the dataset, heteroscedasticity is resolved. The individual predictors vs residual plots (Appendix B) are all slightly improved than before (very negligible), however, the heteroscedasticity in Temperature is resolved. The outlier effect is still prominent in Rainfall and Snowfall. For other predictors only SolarRadiation remains a bit heteroscedastic as before. Categorical variables are better than before and seem fine (Appendix B).

Figure 16



Q-Q plot (Figure 16) is much better than before, some systematic departure in right tail, due to outliers probably, otherwise fine. The correlation of two axis of Q-Q plot is 0.9960, far above my threshold, indicates that it residuals are normal now. Figure 17 shows that Residuals vs Time Sequence plots are much better than before and in cook's distance(appendix B) highest value seems like 0.3453<1, so nothing to

worry about from influential aspect. In VIF all seems fine, as all below 10. When I check correlation among response and non-categorical predictors, Visibility-Humidity (-0.54) & Humidity-SolarRadiation (-0.46) are concerning, everything else is fine. From categorical variables, Seasons vs FunctioningDay table is a concern, everything else is fine (already mentioned before).

Presence of some outlier effects and slight but negligible non-linearity in some predictors vs residuals plots, negligible non-linearity in fitted value vs residual plot are noticed. Having said that, I think we can proceed to model building with current updates, regarding only Outlier effects as the Limitations of the study.

Model Building and Final Model Diagnostic

Using step wise function based on AIC (Appendix B- model building), we finally get a Main effects model with the 'square transformation of RentedBikeCount' as response and 9 variables as predictors. The step fucntion dropped Snowfall and Visibility from the last model (where Dewpoint was already removed). The main effect model is as follows,

$$sqrtRentedBike = \beta_0 + \beta_1 \text{Temperature} + \beta_2 \text{FunctioningDay} + \beta_3 \text{Hour} + \beta_4 \text{Humidity} + \beta_5 \text{Seasons} + \beta_6 \text{Rainfall} + \beta_7 \text{Holiday} + \beta_8 \text{SolarRadiation} + \beta_9 \text{Windspeed} + \epsilon$$

$$\text{where } \epsilon \sim N(0, \sigma^2)$$

Then I used stepwise function based on AIC again, using Main Effect model as initial model and including two-way interactions (model building in Appendix B). Here I should mention that the model suggested by Stepwise function was showing error message when I tried to see its VIF, as two predictors are having almost perfect linear relation (aliased coefficients). Removing Seasons:FunctioningDay interaction, this issue was resolved. May be there was strong correlation between these two categorical variables as we found before. I kept the main effects of them as both are important for our study.

As it includes so many interactions, I thought it is better to show in a picture than to show in a formal structure like Main effects. The picture of Final Model is as follows.

```
fitfinal<-lm(formula = sqrtRentedBike ~ Temperature + factor(FunctioningDay) +
               Hour + Humidity + factor(Seasons) + Rainfall + factor(Holiday) +
               SolarRadiation + Windspeed + Temperature:factor(Seasons) +
               Temperature:Hour + factor(FunctioningDay):Hour + Humidity:factor(Seasons) +
               Hour:Humidity + Humidity:SolarRadiation + Temperature:SolarRadiation +
               factor(FunctioningDay):Humidity + Humidity:Rainfall + Temperature:Humidity +
               factor(Seasons):SolarRadiation + Temperature:factor(FunctioningDay) +
               factor(FunctioningDay):Rainfall + factor(Seasons):Rainfall +
               factor(Seasons):Windspeed + Rainfall:Windspeed + SolarRadiation:Windspeed +
               Humidity:Windspeed + factor(Holiday):SolarRadiation + Humidity:factor(Holiday) +
               Temperature:Rainfall +
               Hour:factor(Holiday) + Rainfall:SolarRadiation + Hour:factor(Seasons) +
               Temperature:Windspeed + Hour:Rainfall + Temperature:factor(Holiday),
             data = newdata)
```

Many big GVIF Values are present in vif of final model, however, those same variables seem fine in main effects model. For example- Rainfall (while interactions are in the model) has the highest GVIF = 48^2 (2304), which is due to structural multicollinearity. Some Interaction terms also show big figures (less than 2304); however, all these can be ignored. We can accept structural multicollinearity if it's not 6,000

or 10,000. So, no need to worry about multicollinearity. R squared of the Final Model is biggest so far (0.7426).

I did a final diagnostic of the final model. Fitted Value vs Residuals plot (Figure 18) reports that for few data points: The length of the previously found linear negative correlation line has declined. Though, for most of the dataset, heteroscedasticity was resolved in the previous model, now it's almost homoscedastic. Overall, this graph is much better than the previous diagnostic graph and indicates that it is so far the best model. Figure 19,20,21 shows the individual predictors vs residual plot of the final model. We find that only SolarRadiation and is showing a clear heteroscedasticity, but for others that pattern is not too clear to declare them to be heteroscedastic. Rainfall is however improved than previous diagnostic plot, but outlier effect is still dominating.

Figure 18

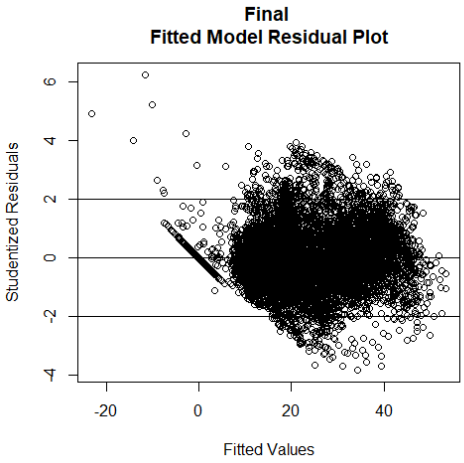


Figure 19

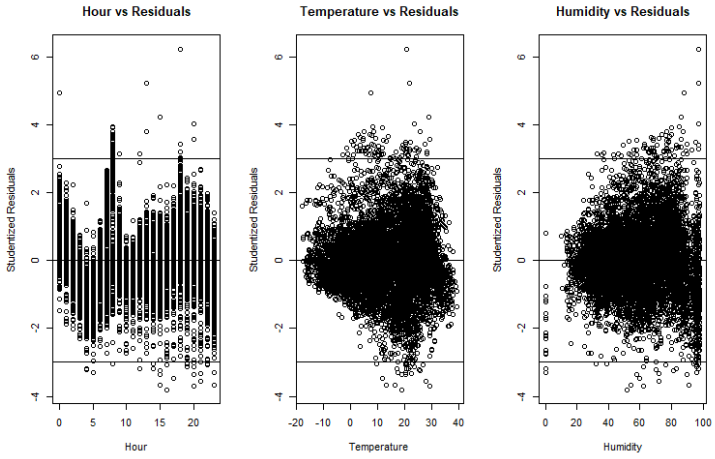


Figure 20

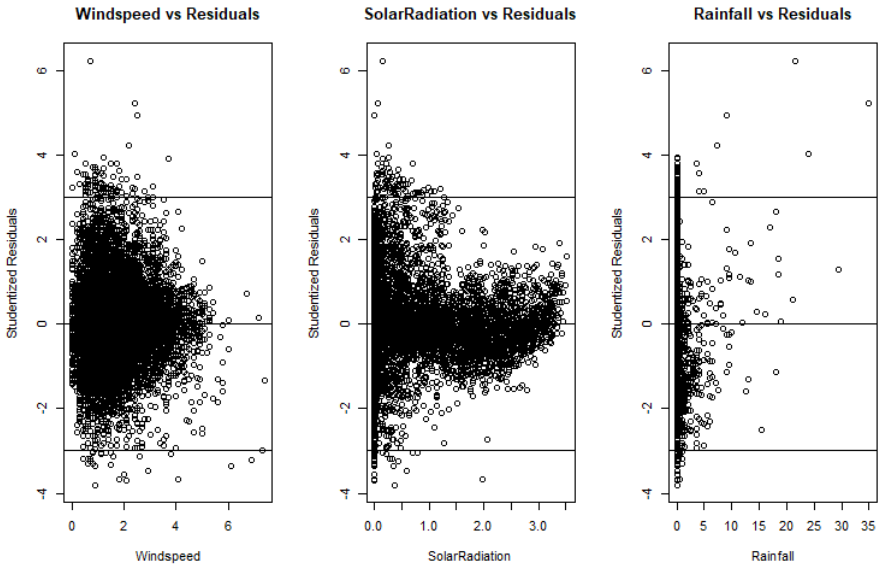
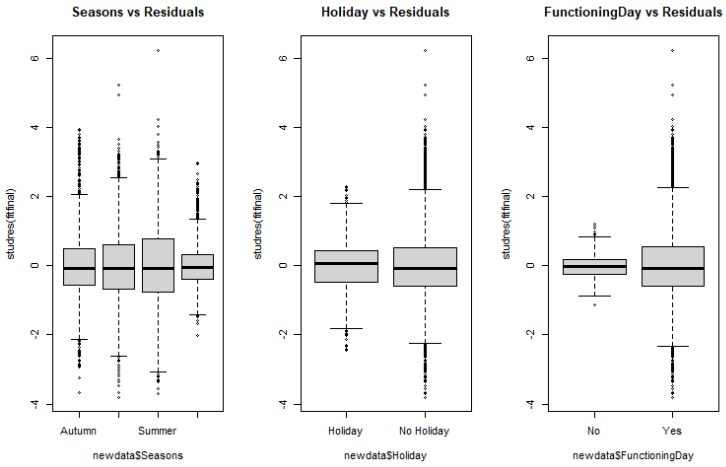


Figure 21



Q-Q plot (Figure 22) is worse than previous diagnostic due to including interaction perhaps, some departure in both tail but not systematic, due to outliers probably. Correlation between two axis of Q-Q plot is 0.9893>0.9850 indicates that residuals are almost normal. The independence of residuals was also checked, and it is same as the last diagnostic, it is fine. Cooks Distance (Figure 23) is showing the maximum value of 0.63<1, it is not influential. However, points got more influential than main effect model in final model. So, this increment of influentiality and decrement of normality can be two notable factors while considering interactions. The summary of main effect model shows 0.65 R squared, while models with interactions (final model) shows 0.74 R squared. The final model is better in terms of explaining more variability of response, and heteroscedasticity and non-linearity of fitted value vs residuals.

Figure 22

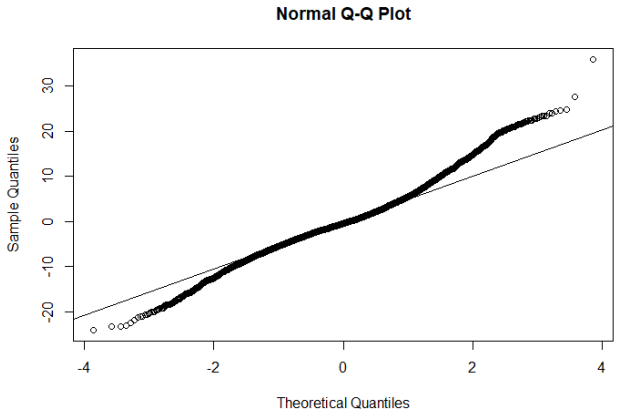
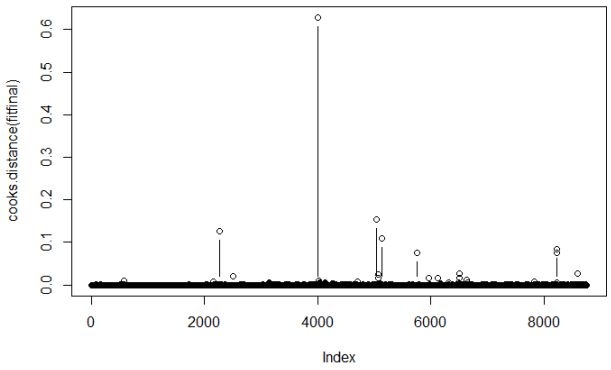


Figure 23



Interpretations, Findings and Conclusions

Checking the summary of both Main Effect model and Final fit model (See appendix), I find some predictor's regression coefficient or beta is significantly different from 0 in one model, however, other model has less significant p-value. So, I did ANOVA test for Functionality, Seasons, Rainfall and Temperature, which were my variable of interest based on the Final Model. I tested their effect on the relationship between response (sqrtRentedBike) and other predictors. I removed their main effect as well interactions to do the anova(fitreduced, finalfit), and all of them are found to be statistically significant. So, in short, I can say about interpretations from the summary that,

1. The average difference between the square roots of the rented bike counts between a Functioning Day and no Functioning Day is 24.34, at 0 Hour, 0 Temperature and 0 Rainfall, keeping all other variables unchanged.
2. The average decrease in the square roots of the rented bike counts is 28.89 for every one mm increase in Rainfall, for No Functioning Day, Autumn Season, 0 Humidity, 0 Hour, 0 Windspeed, 0 Temperature, 0 Solar Radiation, keeping all other variables constant.
3. The average decrease in the square roots of the rented bike counts from Autumn Season to and Winter Seasons is 8 at 0 Hour, 0 Temperature, 0 Humidity, 0 Windspeed, 0 Solar Radiation,0 Rainfall keeping all other variables unchanged.

I also did check some interesting and significant interactions, the results are as follows,

1. While testing the significance of the factor (Seasons): Rainfall interactions (there are 3 interactions as such) by anova, beta of the either one of them seems significant, and most probably the Winter Season: Rainfall. So, the interpretation is that the difference between the average change in square root of Rented Bike counts for one mm increment in Rainfall for Autumn and one mm increment in Rainfall for Winter is 3, keeping all other variables constant.
2. Did the same anova test for beta of Temperature:factor(FunctioningDay), and found it to be significant as mentioned in the summary(finalfit). It means that the difference between the average change in square root of Rented Bike counts for one-degree Celsius increment in Temperature for No FunctioningDay and one-degree Celcius increment in Temperature for a FunctioningDay is 5.5, keeping all other variables constant.

In short, from my understanding, there is a significant effect of temperature, functioning Day, Seasons and Rainfall on the prediction of rented bike counts based on other predictors. The rented bike count increases usually with presence of a Functioning Day and decreases with increment of Rainfall and presence of Winter Season (however, there are many other factors to be considered here). Finally, the winter season exhibits different change in bike counts than Autumn season due to increment in Rainfall. A Functioning Day exhibit different change in Bike counts due to increment in Temperature than A non-Functioning Day.

Appendix A: Codes of the overall Data Analysis:

```
library("car")

library("MASS")

data <- read.csv("Bikerental.csv", header=T)

sum(is.na(data)) #There is no missing value in this dataset, so no need to take
subset.

#Data Visualization

range(data$RentedBikeCount) #0 to 3556

hist(data$RentedBikeCount, main = 'Histogram of RentedBikes', xlab="Rented Bikes
Count")

#Positively Skewed

boxplot(data$RentedBikeCount, main="Boxplot of RentedBikes", ylab="Rented Bikes
Count")

#Significant no. of outliers on right tail

mean(data$RentedBikeCount) #705

par(mfcol=c(1,5))

boxplot(data$Hour, main="Hour")

boxplot(data$Temperature, main="Temperature")

boxplot(data$Humidity, main="Humidity")

boxplot(data$Windspeed, main="Windspeed")

boxplot(data$Visibility, main="Visbility")

par(mfcol=c(1,4))

boxplot(data$Dewpoint, main="DewPointTemp")

boxplot(data$SolarRadiation, main="SolarRadiation")

boxplot(data$Rainfall, main="Rainfall")

boxplot(data$Snowfall, main="Snowfall")

#SolarRadiation is positively skewed.

#However, Rainfall is a bit skewed to right. Outliers effect might be prominent

par(mfcol=c(1,3))

boxplot(RentedBikeCount ~ Seasons, data=data, main="RentedBike by Seasons") #No
Pattern

boxplot(RentedBikeCount ~ Holiday, data=data, main="RentedBike by Holiday") #Linear
Pattern

boxplot(RentedBikeCount ~ FunctioningDay, data=data, main="RentedBike by
FunctioningDay")

table(data$Seasons)

table(data$Holiday)

table(data$FunctioningDay)
```

```
#All of these catagorical variables have enough sample size

table(data[, c(11,12)])

table(data[, c(11,13)])

table(data[, c(12,13)])

#Seems good apart from table between Seasons and Functioning Day.

#Two of the combinations of these variables do not have any sample

#(No FunctioningDay of Summer and Winter )

#I would not have considered any interaction between these two even if my sample size
was smaller.

#However my overall sample size is large, so I will not exclude this interaction if
the step function suggest.

par(mfcol=c(1,1))

pairs(data[, -c(11,12,13)])

#Many predictors are positively related and some response vs predictor plot shows non-
linear pattern

cor(data[, -c(11,12,13)])#Strong Colrel betwn Dewpoint-Temp(0.91),Humidity-
Dewpoint(0.53),

#Humidity-Visibility(-0.54), Humidity-SolarRadiation(-0.46)

#Model Diagnostic

fullfit <- lm(RentedBikeCount ~ Hour + Temperature + Humidity + Windspeed + Visibility
+ Dewpoint + SolarRadiation

              + Rainfall + Snowfall+ factor(Seasons) + factor(Holiday) +
factor(FunctioningDay), data = data)

summary(fullfit) #I don't see any Nan value, which means that no two predictors are
perfectly inter-correlated

#R squared is 0.5504, 0.5497

par(mfcol=c(1,1))

plot(fullfit$fitted.values, studres(fullfit), xlab="Fitted Values", ylab="Studentized
Residuals", main="Full

Fitted Model Residual Plot")

abline(h=c(0, 2, -2))

#Presence of Funnel Shape as well as a pattern in fitted value vs Error,

#seems like presence of Non-linearity & Heteroscedasticity

#Individual Predictors vs Residuals Analysis

par(mfcol=c(1,4))

plot(data$Hour, studres(fullfit), xlab="Hour", ylab="Studentized Residuals",
main="Hour vs Residuals")

abline(h=c(0, 3, -3)) #unclear pattern

plot(data$Temperature, studres(fullfit), xlab="Temperature", ylab="Studentized
Residuals", main="Temperature vs Residuals")
```

```
abline(h=c(0, 3, -3))#Funnel shaped, Heteroscedastic

plot(data$Humidity, studres(fullfit), xlab="Humidity", ylab="Studentized Residuals",
main="Humidity vs Residuals")

abline(h=c(0, 3, -3))

plot(data$Windspeed, studres(fullfit), xlab="Windspeed", ylab="Studentized Residuals",
main="Windspeed vs Residuals")

abline(h=c(0, 3, -3)) #Reverse Funnel Shape, Heteroscedastic, but not strong

plot(data$Visibility, studres(fullfit), xlab="Visibility", ylab="Studentized
Residuals", main="Visibility vs Residuals")

abline(h=c(0, 3, -3))#There is a pattern but not clear

plot(data$Dewpoint, studres(fullfit), xlab="Dewpoint", ylab="Studentized Residuals",
main="Dewpoint vs Residuals")

abline(h=c(0, 3, -3)) #Funnel Shaped, Heteroscedastic

plot(data$SolarRadiation, studres(fullfit), xlab="SolarRadiation", ylab="Studentized
Residuals", main="SolarRadiation vs Residuals")

abline(h=c(0, 3, -3)) #Heteroscedastic, a reverse funnel shape has formed

plot(data$Rainfall, studres(fullfit), xlab="Rainfall", ylab="Studentized Residuals",
main="Rainfall vs Residuals")

abline(h=c(0, 3, -3)) #Outlier effect is clearly visible and highly heteroscedastic

plot(data$Snowfall, studres(fullfit), xlab="Snowfall", ylab="Studentized Residuals",
main="Snowfall vs Residuals")

abline(h=c(0, 3, -3)) #Highly heteroscedastic and has outlier effect

boxplot(studres(fullfit)~data$Seasons, main='Seasons vs Residuals') #Seems fine,
slight heteroscedasticity

boxplot(studres(fullfit)~data$Holiday, main='Holiday vs Residuals') #Seems fine,
slight heteroscedasticity

boxplot(studres(fullfit)~data$FunctioningDay, main='FunctioningDay vs Residuals')
#Seems fine, slight heteroscedasticity

#Most of these heterscedasticity can be due to non-linearity as we saw in pairs plot

#Normality of Residual Analysis

par(mfcol=c(1,1))

qqnorm(fullfit$residuals)

qqline(fullfit$residuals) #departure in right tail almost from the midway,

#indicating non-normality of errors

norm <- qqnorm(fullfit$residuals)

cor(norm$x, norm$y) #97.51< 98.5%)

dataseq<-c(1:nrow(data))

plot(dataseq, studres(fullfit), xlab='Time Sequence', ylab='Residuals', main='Checking
Independence of Residuals')

#No clear pattern, we can say residuals are independent from one another
```

```
plot(cooks.distance(fullfit), type='b')

cooks.distance(fullfit)[which.max(cooks.distance(fullfit))]

#the highest value is 0.1055, so nothing to worry about influentiality

vif(fullfit)

# Have to drop DewPoint as it has 100+ VIF value, vif=117.30

#Drop Dewpoint from Dataset

newdata<- data[,-7]


#Model Diagnostic 2

fullfit <- lm(RentedBikeCount ~ Hour + Temperature + Humidity + Windspeed + Visibility
+ SolarRadiation

              + Rainfall + Snowfall+ factor(Seasons) + factor(Holiday) +
factor(FunctioningDay), data = newdata)

summary(fullfit) #R square remains almost same

plot(fullfit$fitted.values, studres(fullfit), xlab="Fitted Values", ylab="Studentized
Residuals", main="Full

Fitted Model Residual Plot")

abline(h=c(0, 2, -2))

#Heteroscedasticity still remains as the plot remains same as before

#Individual Predictors vs Residuals Analysis

par(mfcol=c(1,4))

plot(newdata$Hour, studres(fullfit), xlab="Hour", ylab="Studentized Residuals",
main="Hour vs Residuals")

abline(h=c(0, 3, -3)) #Unclear Pattern

plot(newdata$Temperature, studres(fullfit), xlab="Temperature", ylab="Studentized
Residuals", main="Temperature vs Residuals")

abline(h=c(0, 3, -3))#Heteroscedastic, funnel shaped

plot(newdata$Humidity, studres(fullfit), xlab="Humidity", ylab="Studentized
Residuals", main="Humidity vs Residuals")

abline(h=c(0, 3, -3))

plot(newdata$Windspeed, studres(fullfit), xlab="Windspeed", ylab="Studentized
Residuals", main="Windspeed vs Residuals")

abline(h=c(0, 3, -3)) #Heteroscedastic, but not strong

plot(newdata$Visibility, studres(fullfit), xlab="Visibility", ylab="Studentized
Residuals", main="Visibility vs Residuals")

abline(h=c(0, 3, -3))#There is a pattern but not clear,cannot decide it to be heterosc

plot(newdata$SolarRadiation, studres(fullfit), xlab="SolarRadiation",
ylab="Studentized Residuals", main="SolarRadiation vs Residuals")

abline(h=c(0, 3, -3)) #Funnel Shapped, Heteroscedastic
```

```
plot(newdata$Rainfall, studres(fullfit), xlab="Rainfall", ylab="Studentized
Residuals", main="Rainfall vs Residuals")

abline(h=c(0, 3, -3)) #Outlier effect is clearly visible and highly heteroscedastic

plot(newdata$Snowfall, studres(fullfit), xlab="Snowfall", ylab="Studentized
Residuals", main="Snowfall vs Residuals")

abline(h=c(0, 3, -3)) #Highly heteroscedastic and has outlier effect

boxplot(studres(fullfit)~newdata$Seasons, main='Seasons vs Residuals') #Seems fine

boxplot(studres(fullfit)~newdata$Holiday, main='Holiday vs Residuals') #Seems fine

boxplot(studres(fullfit)~newdata$FunctioningDay, main='FunctioningDay vs Residuals')
#Seems fine

#All plots are almost same as before

#Nonrmality of residuals

par(mfcol=c(1,1))

qqnorm(fullfit$residuals)

qqline(fullfit$residuals) #departure in right tail almost from the midway,

#indicating non-normality of errors

norm <- qqnorm(fullfit$residuals)

cor(norm$x, norm$y) #97.50< 99.5%)

dataseq<-c(1:nrow(newdata))

plot(dataseq, studres(fullfit)) #Same plot as before, independent errors

plot(cooks.distance(fullfit), type='b')

cooks.distance(fullfit)[which.max(cooks.distance(fullfit))]

#highest value is 0.118, so nothing to worry about influentiality

vif(fullfit) #Seems okay now


#To fix heteroscedasticity due to non-linearity and non-normality of error, lets do
log transformation of response

newdata$logRentedBike <- log(newdata$RentedBikeCount)

fullfit <- lm(logRentedBike ~ Temperature + Humidity + Windspeed + Visibility +
SolarRadiation

              + Rainfall + Snowfall+ factor(Seasons) + factor(Holiday) +
factor(FunctioningDay), data = data)

#Got error due to presence of negative inf value in logRentedBike col,

#which directed to the fact that actual response is having many values equal to 0

nrow(data[data$RentedBikeCount == 0,])  #295 rows with response=0


#Square Transformation of RentedBikeCount

newdata$sqrtRentedBike <- sqrt(newdata$RentedBikeCount)
```

```
hist(newdata$RentedBikeCount)

hist(newdata$sqrtRentedBike) #much more symmetric and better than actual values
distribution

boxplot(newdata$sqrtRentedBike) #Box plot also seems perfect, almost 0 outliers


#Model Diagnostic 3

fullfit <- lm(sqrtRentedBike ~ Hour + Temperature + Humidity + Windspeed + Visibility
+ SolarRadiation

              + Rainfall + Snowfall+ factor(Seasons) + factor(Holiday) +
factor(FunctioningDay), data = newdata)

summary(fullfit) #Much improved R square 0.65

par(mfcol=c(1,1))

plot(fullfit$fitted.values, studres(fullfit), xlab="Fitted Values", ylab="Studentized
Residuals", main="Full

Fitted Model Residual Plot")

abline(h=c(0, 2, -2))

#For a few data points the a linear negative correlation pattern is still there,

#however for most of the dataset, heteroscedasticity is resolved

#Individual Predictors vs Residuals Analysis

par(mfcol=c(1,4))

plot(newdata$Hour, studres(fullfit), xlab="Hour", ylab="Studentized Residuals",
main="Hour vs Residuals")

abline(h=c(0, 3, -3))

plot(newdata$Temperature, studres(fullfit), xlab="Temperature", ylab="Studentized
Residuals", main="Temperature vs Residuals")

abline(h=c(0, 3, -3))#No pattern

plot(newdata$Humidity, studres(fullfit), xlab="Humidity", ylab="Studentized
Residuals", main="Humidity vs Residuals")

abline(h=c(0, 3, -3))

plot(newdata$Windspeed, studres(fullfit), xlab="Windspeed", ylab="Studentized
Residuals", main="Windspeed vs Residuals")

abline(h=c(0, 3, -3)) #Heteroscedastic, but not strong

plot(newdata$Visibility, studres(fullfit), xlab="Visibility", ylab="Studentized
Residuals", main="Visibility vs Residuals")

abline(h=c(0, 3, -3))#Unclear Pattern

plot(newdata$SolarRadiation, studres(fullfit), xlab="SolarRadiation",
ylab="Studentized Residuals", main="SolarRadiation vs Residuals")

abline(h=c(0, 3, -3)) #Heteroscedastic, funnel shape

plot(newdata$Rainfall, studres(fullfit), xlab="Rainfall", ylab="Studentized
Residuals", main="Rainfall vs Residuals")
```

```
abline(h=c(0, 3, -3)) #Outlier effect still remains-heteroscedastic, but pattern is
weaker than before

plot(newdata$Snowfall, studres(fullfit), xlab="Snowfall", ylab="Studentized
Residuals", main="Snowfall vs Residuals")

abline(h=c(0, 3, -3)) #Outlier effect still remains- heteroscedastic

boxplot(studres(fullfit)~newdata$Seasons, main='Seasons vs Residuals') #Seems fine

boxplot(studres(fullfit)~newdata$Holiday, main='Holiday vs Residuals') #Seems fine

boxplot(studres(fullfit)~newdata$FunctioningDay, main='FunctioningDay vs Residuals')
#Seems fine, better than before


#Checking Normality of residuals

par(mfcol=c(1,1))

qqnorm(fullfit$residuals)

qqline(fullfit$residuals) #much better than before,

#some systematic depurture in right tail, due to outliers probably

norm <- qqnorm(fullfit$residuals)

cor(norm$x, norm$y) #0.9960 indicates that it residuals are normal now

dataseq<-c(1:nrow(newdata))

plot(dataseq, studres(fullfit)) #Independence of residuals over time, better than
before

plot(cooks.distance(fullfit), type='b')

cooks.distance(fullfit)[which.max(cooks.distance(fullfit))]

#highest value seems like 0.3453, so nothing to worry about influentiality

vif(fullfit) #Seems fine, all bellow 10

cor(newdata[, -c(1,10,11,12,13)]) #Visibility:Humidity & Humidity:SolarRadiation

#are concerning, everything else is fine

table(newdata[, c(10,11)])

table(newdata[, c(11,12)])

table(newdata[, c(10,12)]) #Seasons: FunctioningDay is a concern, everyhting else is
fine


#Model Building

fit0 <- lm(sqrtRentedBike ~ 1, data = newdata)

step(fit0, sqrtRentedBike ~ Hour + Temperature + Humidity + Windspeed + Visibility +
SolarRadiation

     + Rainfall + Snowfall+ factor(Seasons) + factor(Holiday) +
factor(FunctioningDay),

     direction = "both", trace = 1)

#Dropped Snowfall and Visibility
```

```
fitmain<-lm(formula = sqrtRentedBike ~ Temperature + factor(FunctioningDay) +
              Hour + Humidity + factor(Seasons) + Rainfall + factor(Holiday) +
              SolarRadiation + Windspeed, data = newdata)
#Let go forward and include possible two way interactions of these predictors
step(fitmain, scope = .~.^2,
     direction = "both", trace = 0)
fitfinal<-lm(formula = sqrtRentedBike ~ Temperature + factor(FunctioningDay) +
               Hour + Humidity + factor(Seasons) + Rainfall + factor(Holiday) +
               SolarRadiation + Windspeed + Temperature:factor(Seasons) +
               Temperature:Hour + factor(FunctioningDay):Hour +
Humidity:factor(Seasons) +
               Hour:Humidity + Humidity:SolarRadiation + Temperature:SolarRadiation +
               factor(FunctioningDay):Humidity + Humidity:Rainfall +
Temperature:Humidity +
               factor(Seasons):SolarRadiation + Temperature:factor(FunctioningDay) +
               factor(FunctioningDay):Rainfall + factor(Seasons):Rainfall +
               factor(Seasons):Windspeed + Rainfall:Windspeed +
SolarRadiation:Windspeed +
               Humidity:Windspeed + factor(Holiday):SolarRadiation +
Humidity:factor(Holiday) +
               Temperature:Rainfall + factor(FunctioningDay):factor(Seasons) +
               Hour:factor(Holiday) + Rainfall:SolarRadiation + Hour:factor(Seasons) +
               Temperature:Windspeed + Hour:Rainfall + Temperature:factor(Holiday),
             data = newdata)
vif(fitfinal)
#Showing error message as two predictors are having almost perfect linear relation
(aliased coefficients)
#(proabably Seasons:FunctioningDay), making it impossible to estimate their individual
effects
#However, this message was not shown while fitting fitmain
#I decided to drop their interaction, but kept the main effect
fitfinal<-lm(formula = sqrtRentedBike ~ Temperature + factor(FunctioningDay) +
               Hour + Humidity + factor(Seasons) + Rainfall + factor(Holiday) +
               SolarRadiation + Windspeed + Temperature:factor(Seasons) +
               Temperature:Hour + factor(FunctioningDay):Hour +
Humidity:factor(Seasons) +
               Hour:Humidity + Humidity:SolarRadiation + Temperature:SolarRadiation +
               factor(FunctioningDay):Humidity + Humidity:Rainfall +
Temperature:Humidity +
```

```
                factor(Seasons):SolarRadiation + Temperature:factor(FunctioningDay) +

                factor(FunctioningDay):Rainfall + factor(Seasons):Rainfall +

                factor(Seasons):Windspeed + Rainfall:Windspeed +
SolarRadiation:Windspeed +

                Humidity:Windspeed + factor(Holiday):SolarRadiation +
Humidity:factor(Holiday) +

                Temperature:Rainfall +

                Hour:factor(Holiday) + Rainfall:SolarRadiation + Hour:factor(Seasons) +

                Temperature:Windspeed + Hour:Rainfall + Temperature:factor(Holiday),

              data = newdata)


#Final Model Diagnostic

vif(fitfinal) #Now it is running without error.

#However, many big GVIF values

vif(fitmain) #vif of only main effects seems fine

#Rainfall (while interactions are in the model) is having the highest GVIF = 48^2

#which is due to structural multicolinearity, can be ignored.

summary(fitfinal) #R squared is much better with 0.7426,0.7411 value

par(mfcol=c(1,1))

plot(fitfinal$fitted.values, studres(fitfinal), xlab="Fitted Values",
ylab="Studentized Residuals", main="Final

Fitted Model Residual Plot")

abline(h=c(0, 2, -2))

#For few data points: The length of the previously found linear negative correlation
line has declined,

#However for most of the dataset, heteroscedasticity was resolved before, now its
almost homoscedastic

#Overall, this graph is much better than previous diagnostic graph

#Individual Predictors vs Residuals Analysis

par(mfcol=c(1,3))

plot(newdata$Hour, studres(fitfinal), xlab="Hour", ylab="Studentized Residuals",
main="Hour vs Residuals")

abline(h=c(0, 3, -3))

plot(newdata$Temperature, studres(fitfinal), xlab="Temperature", ylab="Studentized
Residuals", main="Temperature vs Residuals")

abline(h=c(0, 3, -3))#No pattern

plot(newdata$Humidity, studres(fitfinal), xlab="Humidity", ylab="Studentized
Residuals", main="Humidity vs Residuals")

abline(h=c(0, 3, -3))
```

```
plot(newdata$Windspeed, studres(fitfinal), xlab="Windspeed", ylab="Studentized
Residuals", main="Windspeed vs Residuals")

abline(h=c(0, 3, -3)) #Heteroscedastic, but not strong

plot(newdata$SolarRadiation, studres(fitfinal), xlab="SolarRadiation",
ylab="Studentized Residuals", main="SolarRadiation vs Residuals")

abline(h=c(0, 3, -3)) #Heteroscedastic, funnel shape

plot(newdata$Rainfall, studres(fitfinal), xlab="Rainfall", ylab="Studentized
Residuals", main="Rainfall vs Residuals")

abline(h=c(0, 3, -3)) #Outlier effect still remains-heteroscedastic, but pattern is
weaker than before

boxplot(studres(fitfinal)~newdata$Seasons, main='Seasons vs Residuals') #Seems fine

boxplot(studres(fitfinal)~newdata$Holiday, main='Holiday vs Residuals') #Seems fine

boxplot(studres(fitfinal)~newdata$FunctioningDay, main='FunctioningDay vs Residuals')
#Seems fine

#Checking Normality of residuals

par(mfcol=c(1,1))

qqnorm(fitfinal$residuals)

qqline(fitfinal$residuals) #worse than previous diagnostic due to including
interaction probably,

#some departure in both tail but not systematic, due to outliers probably

norm <- qqnorm(fitfinal$residuals)

cor(norm$x, norm$y) #0.9893 indicates that residuals are almost normal

dataseq<-c(1:nrow(newdata))

plot(dataseq, studres(fitfinal)) #Independence of residuals over time: Seems fine

plot(cooks.distance(fitfinal), type='b')

cooks.distance(fitfinal)[which.max(cooks.distance(fitfinal))]

#Max cooks distance is 0.63 < 1, it is not a matter of concern

summary(fitfinal) #Open for interpretation

summary(fitmain) #All predictors beta seems significantly different from 0

#in main effect except for Windspeed #0.65 R squared

#So model with both interaction and main effects seems to be better model

#in terms of explanation of the variability.

#FunctioningDay, Humidity, Seasons, Ranifall and Holiday

#seems significantly associated to square root of rented bike counts

#in both main effect and interaction models

#However we can only be sure after checking both main effect and interactions

#related to a specific variable

#There is no term which only has a main effect but no interactions
```

```
#Testing for FunctioningDay predictability
fit3<-lm(formula = sqrtRentedBike ~ Temperature +
           Hour + Humidity + factor(Seasons) + Rainfall + factor(Holiday) +
           SolarRadiation + Windspeed + Temperature:factor(Seasons) +
           Temperature:Hour + Humidity:factor(Seasons) +
           Hour:Humidity + Humidity:SolarRadiation + Temperature:SolarRadiation +
           + Humidity:Rainfall + Temperature:Humidity +
           factor(Seasons):SolarRadiation +
           + factor(Seasons):Rainfall +
           factor(Seasons):Windspeed + Rainfall:Windspeed + SolarRadiation:Windspeed +
           Humidity:Windspeed + factor(Holiday):SolarRadiation +
Humidity:factor(Holiday) +
           Temperature:Rainfall +
           Hour:factor(Holiday) + Rainfall:SolarRadiation + Hour:factor(Seasons) +
           Temperature:Windspeed + Hour:Rainfall + Temperature:factor(Holiday),
         data = newdata)

anova(fit3, fitfinal) #P-value: 2.2e-16 *** (FunctioningDay is Significantly
associated to rented bike counts)

#Testing for Seasons predictability
fit5<-lm(formula = sqrtRentedBike ~ Temperature + factor(FunctioningDay) +
           Hour + Humidity + Rainfall + factor(Holiday) +
           SolarRadiation + Windspeed +
           Temperature:Hour + factor(FunctioningDay):Hour+
           Hour:Humidity + Humidity:SolarRadiation + Temperature:SolarRadiation +
           factor(FunctioningDay):Humidity + Humidity:Rainfall + Temperature:Humidity
+
           Temperature:factor(FunctioningDay) +
           factor(FunctioningDay):Rainfall
         + Rainfall:Windspeed + SolarRadiation:Windspeed +
           Humidity:Windspeed + factor(Holiday):SolarRadiation +
Humidity:factor(Holiday) +
           Temperature:Rainfall +
           Hour:factor(Holiday) + Rainfall:SolarRadiation +
           Temperature:Windspeed + Hour:Rainfall + Temperature:factor(Holiday),
```

```
         data = newdata)

anova(fit5, fitfinal) #P-value< 2.2e-16 *** (Seasons is Significantly associated to
rented bike counts)


#Testing for Rainfall predictability

fit6<-lm(formula = sqrtRentedBike ~ Temperature + factor(FunctioningDay) +

           Hour + Humidity + factor(Seasons) + factor(Holiday) +

           SolarRadiation + Windspeed + Temperature:factor(Seasons) +

           Temperature:Hour + factor(FunctioningDay):Hour + Humidity:factor(Seasons) +

           Hour:Humidity + Humidity:SolarRadiation + Temperature:SolarRadiation +

           factor(FunctioningDay):Humidity + Temperature:Humidity +

           factor(Seasons):SolarRadiation + Temperature:factor(FunctioningDay)+

           factor(Seasons):Windspeed + SolarRadiation:Windspeed +

           Humidity:Windspeed + factor(Holiday):SolarRadiation +
Humidity:factor(Holiday) +

           Hour:factor(Holiday) + Hour:factor(Seasons) +

           Temperature:Windspeed + Temperature:factor(Holiday),

         data = newdata)

anova(fit6, fitfinal) #P-value< 2.2e-16 *** (Rainfall is Significantly associated to
rented bike counts)


#Testing for Temperature predictability

fit8<-lm(formula = sqrtRentedBike ~ factor(FunctioningDay) +

           Hour + Humidity + factor(Seasons) + Rainfall + factor(Holiday) +

           SolarRadiation + Windspeed + factor(FunctioningDay):Hour +
Humidity:factor(Seasons) +

           Hour:Humidity + Humidity:SolarRadiation +

           factor(FunctioningDay):Humidity + Humidity:Rainfall +

           factor(Seasons):SolarRadiation +

           factor(FunctioningDay):Rainfall + factor(Seasons):Rainfall +

           factor(Seasons):Windspeed + Rainfall:Windspeed + SolarRadiation:Windspeed +

           Humidity:Windspeed + factor(Holiday):SolarRadiation +
Humidity:factor(Holiday) +

           Hour:factor(Holiday) + Rainfall:SolarRadiation + Hour:factor(Seasons) +

           Hour:Rainfall,

         data = newdata)

anova(fit8,fitfinal) #P-value< 2.2e-16 *** (Temperature also is Significantly
associated to rented bike counts)
```

```
#Indicate to the fact that interactions in this model has significant effect. Even if
the main effect does not seem significant,

#considering its associated interactions can make the overall predictor significant.


#Testing for some significant interactions -Difference of the difference exists or
not.

#factor(Seasons)Winter:Rainfall (Autumn vs Winter)

#Temperature:factor(FunctioningDay)Yes (No vs Yes)


fit9<-lm(formula = sqrtRentedBike ~ Temperature + factor(FunctioningDay) +

            Hour + Humidity + factor(Seasons) + Rainfall + factor(Holiday) +

            SolarRadiation + Windspeed + Temperature:factor(Seasons) +

            Temperature:Hour + factor(FunctioningDay):Hour + Humidity:factor(Seasons) +

            Hour:Humidity + Humidity:SolarRadiation + Temperature:SolarRadiation +

            factor(FunctioningDay):Humidity + Humidity:Rainfall + Temperature:Humidity
+

            factor(Seasons):SolarRadiation + Temperature:factor(FunctioningDay) +

            factor(FunctioningDay):Rainfall +

            factor(Seasons):Windspeed + Rainfall:Windspeed + SolarRadiation:Windspeed +

            Humidity:Windspeed + factor(Holiday):SolarRadiation +
Humidity:factor(Holiday) +

            Temperature:Rainfall +

            Hour:factor(Holiday) + Rainfall:SolarRadiation + Hour:factor(Seasons) +

            Temperature:Windspeed + Hour:Rainfall + Temperature:factor(Holiday),

         data = newdata)

anova(fit9,fitfinal) #P-Value = 5.753e-10 ***   Either of seasons: rainfall
significant.

#When testing for the seasons:rainfall either one to be significant, it is becoming
significant and most probably the term is winter: rainfall

#Lets interpret Autumn vs Winter for Rainfall.


fit11<-lm(formula = sqrtRentedBike ~ Temperature + factor(FunctioningDay) +

             Hour + Humidity + factor(Seasons) + Rainfall + factor(Holiday) +

             SolarRadiation + Windspeed + Temperature:factor(Seasons) +

             Temperature:Hour + factor(FunctioningDay):Hour + Humidity:factor(Seasons)
+

             Hour:Humidity + Humidity:SolarRadiation + Temperature:SolarRadiation +

             factor(FunctioningDay):Humidity + Humidity:Rainfall + Temperature:Humidity
+
```

```
            factor(Seasons):SolarRadiation+
            factor(FunctioningDay):Rainfall + factor(Seasons):Rainfall +
            factor(Seasons):Windspeed + Rainfall:Windspeed + SolarRadiation:Windspeed
+
            Humidity:Windspeed + factor(Holiday):SolarRadiation +
Humidity:factor(Holiday) +
            Temperature:Rainfall +
            Hour:factor(Holiday) + Rainfall:SolarRadiation + Hour:factor(Seasons) +
            Temperature:Windspeed + Hour:Rainfall + Temperature:factor(Holiday),
          data = newdata)
anova(fit11,fitfinal) #P-value= 1.202e-08 *** Temparature: factor(FuctioningDay) Yes.
#Its same as summary table. Significant.
```

Appendix B- Model Building

Main Effects Model:

```
Start:  AIC=44182.93
sqrtRentedBike ~ 1

                         Df Sum of Sq     RSS   AIC
+ Temperature             1    395210  962672 41172
+ factor(Seasons)         3    292362 1065520 42065
+ Hour                    1    206643 1151239 42739
+ factor(FunctioningDay)  1    167780 1190102 43030
+ SolarRadiation          1    124705 1233177 43341
+ Humidity                1     70163 1287718 43720
+ Visibility              1     56487 1301395 43813
+ Rainfall                1     36315 1321567 43947
+ Snowfall                1     32659 1325222 43972
+ Windspeed               1     17087 1340795 44074
+ factor(Holiday)         1     10103 1347778 44120
<none>                               1357882 44183

Step:  AIC=41171.77
sqrtRentedBike ~ Temperature

                          Df Sum of Sq     RSS   AIC
+ factor(Functioning Day)  1    195104  767568 39190
+ Hour                     1    144011  818661 39754
+ Humidity                 1    136752  825920 39832
+ Rainfall                 1     49487  913185 40711
+ Visibility               1     46624  916048 40739
+ Windspeed                1     23595  939077 40956
+ SolarRadiation           1     19582  943090 40994
+ factor(Seasons)          3      6461  956211 41119
+ factor(Holiday)          1      4285  958387 41135
+ Snowfall                 1      1979  960692 41156
<none>                               962672 41172
- Temperature              1    395210 1357882 44183

Step:  AIC=39189.76
sqrtRentedBike ~ Temperature + factor(FunctioningDay)

                         Df Sum of Sq     RSS   AIC
+ Hour                    1    140111  627457 37426
+ Humidity                1    132565  635004 37531
+ Visibility              1     51405  716163 38585
+ Rainfall                1     50394  717175 38597
+ factor(Seasons)         3     38305  729263 38747
+ Windspeed               1     23160  744408 38923
+ SolarRadiation          1     18275  749294 38981
+ Snowfall                1      2925  764644 39158
+ factor(Holiday)         1      2704  764864 39161
<none>                               767568 39190
- factor(FunctioningDay)  1    195104  962672 41172
- Temperature             1    422533 1190102 43030

Step:  AIC=37426.16
sqrtRentedBike ~ Temperature + factor(FunctioningDay) + Hour
```

```
                          Df Sum of Sq    RSS   AIC
+ Humidity                 1     75173 552284 36310
+ Rainfall                 1     50805 576652 36688
+ factor(Seasons)          3     38116 589341 36883
+ Visibility               1     36802 590655 36899
+ SolarRadiation           1      9017 618440 37301
+ Snowfall                 1      3153 624304 37384
+ factor(Holiday)          1      2999 624458 37386
+ Windspeed                1      2006 625451 37400
<none>                                627457 37426
- Hour                     1    140111 767568 39190
- factor(FunctioningDay)   1    191204 818661 39754
- Temperature              1    358017 985474 41379

Step:  AIC=36310.27
sqrtRentedBike ~ Temperature + factor(FunctioningDay) + Hour +
    Humidity

                          Df Sum of Sq    RSS   AIC
+ factor(Seasons)          3     48750 503534 35507
+ Rainfall                 1     26941 525343 35874
+ SolarRadiation           1      4750 547533 36237
+ factor(Holiday)          1      4398 547886 36242
+ Visibility               1      2313 549971 36276
+ Windspeed                1      1097 551187 36295
+ Snowfall                 1       184 552100 36309
<none>                                552284 36310
- Humidity                 1     75173 627457 37426
- Hour                     1     82720 635004 37531
- factor(FunctioningDay)   1    188742 741026 38883
- Temperature              1    410199 962483 41174

Step:  AIC=35506.74
sqrtRentedBike ~ Temperature + factor(FunctioningDay) + Hour +
    Humidity + factor(Seasons)

                          Df Sum of Sq    RSS   AIC
+ Rainfall                 1     25994 477539 35044
+ SolarRadiation           1      2976 500558 35457
+ factor(Holiday)          1      2941 500592 35457
+ Visibility               1       390 503143 35502
<none>                                503534 35507
+ Windspeed                1        83 503451 35507
+ Snowfall                 1        20 503514 35508
- factor(Seasons)          3     48750 552284 36310
- Temperature              1     53634 557167 36391
- Humidity                 1     85808 589341 36883
- Hour                     1     89405 592938 36936
- factor(FunctioningDay)   1    213407 716940 38600

Step:  AIC=35044.43
sqrtRentedBike ~ Temperature + factor(FunctioningDay) + Hour +
    Humidity + factor(Seasons) + Rainfall

                          Df Sum of Sq    RSS   AIC
+ factor(Holiday)          1      2954 474585 34992
+ SolarRadiation           1      2294 475245 35004
+ Visibility               1       162 477377 35043
<none>                                477539 35044
+ Snowfall                 1         2 477537 35046
```

```
+ Windspeed                1         1 477538 35046
- Rainfall                 1     25994 503534 35507
- factor(Seasons)          3     47804 525343 35874
- Temperature              1     53709 531248 35976
- Humidity                 1     60825 538364 36093
- Hour                     1     95628 573167 36641
- factor(FunctioningDay)   1    213752 691291 38283

Step:  AIC=34992.07
sqrtRentedBike ~ Temperature + factor(FunctioningDay) + Hour +
    Humidity + factor(Seasons) + Rainfall + factor(Holiday)

                          Df Sum of Sq    RSS   AIC
+ SolarRadiation           1      2408 472177 34950
+ Visibility               1       198 474388 34990
<none>                                 474585 34992
+ Windspeed                1         1 474584 34994
+ Snowfall                 1         0 474585 34994
- factor(Holiday)          1      2954 477539 35044
- Rainfall                 1     26007 500592 35457
- factor(Seasons)          3     46331 520916 35802
- Temperature              1     55328 529914 35956
- Humidity                 1     61217 535802 36053
- Hour                     1     94840 569426 36586
- factor(FunctioningDay)   1    212566 687151 38232

Step:  AIC=34949.51
sqrtRentedBike ~ Temperature + factor(FunctioningDay) + Hour +
    Humidity + factor(Seasons) + Rainfall + factor(Holiday) +
    SolarRadiation

                          Df Sum of Sq    RSS   AIC
+ Windspeed                1       123 472054 34949
<none>                                 472177 34950
+ Snowfall                 1        30 472147 34951
+ Visibility               1        17 472160 34951
- SolarRadiation           1      2408 474585 34992
- factor(Holiday)          1      3068 475245 35004
- Rainfall                 1     25310 497487 35405
- factor(Seasons)          3     44819 516996 35738
- Humidity                 1     55559 527736 35922
- Temperature              1     55863 528040 35927
- Hour                     1     91988 564166 36507
- factor(FunctioningDay)   1    212741 684918 38206

Step:  AIC=34949.23
sqrtRentedBike ~ Temperature + factor(FunctioningDay) + Hour +
    Humidity + factor(Seasons) + Rainfall + factor(Holiday) +
    SolarRadiation + Windspeed

                          Df Sum of Sq    RSS   AIC
<none>                                 472054 34949
- Windspeed                1       123 472177 34950
+ Snowfall                 1        29 472025 34951
+ Visibility               1        10 472045 34951
- SolarRadiation           1      2530 474584 34994
- factor(Holiday)          1      3083 475138 35004
- Rainfall                 1     25411 497465 35407
- factor(Seasons)          3     44863 516918 35739
- Humidity                 1     54232 526287 35900
```

```
- Temperature              1     55974 528029 35929
- Hour                     1     85362 557417 36403
- factor(FunctioningDay)   1    212845 684899 38208

Call:
lm(formula = sqrtRentedBike ~ Temperature + factor(FunctioningDay) +
    Hour + Humidity + factor(Seasons) + Rainfall + factor(Holiday) +
    SolarRadiation + Windspeed, data = newdata)

Coefficients:
               (Intercept)                 Temperature  factor(FunctioningDay)Yes
                   -4.9981                      0.4727                    28.3888
                      Hour                    Humidity      factor(Seasons)Spring
                    0.4940                     -0.1608                    -3.1176
    factor(Seasons)Summer      factor(Seasons)Winter                    Rainfall
                   -2.9487                     -8.1827                    -1.5606
factor(Holiday)No Holiday             SolarRadiation                   Windspeed
                    2.7699                     -0.8376                     0.1298
```

Interaction Effects Model: Too big, just output given.

```
Call:
lm(formula = sqrtRentedBike ~ Temperature + factor(FunctioningDay) +
    Hour + Humidity + factor(Seasons) + Rainfall + factor(Holiday) +
    SolarRadiation + Windspeed + Temperature:factor(Seasons) +
    Temperature:Hour + factor(FunctioningDay):Hour + Humidity:factor(Seasons) +
    Hour:Humidity + Humidity:SolarRadiation + Temperature:SolarRadiation +
    factor(FunctioningDay):Humidity + Humidity:Rainfall + Temperature:Humidity +
    factor(Seasons):SolarRadiation + Temperature:factor(FunctioningDay) +
    factor(FunctioningDay):Rainfall + factor(Seasons):Rainfall +
    factor(Seasons):Windspeed + Rainfall:Windspeed + SolarRadiation:Windspeed +
    Humidity:Windspeed + factor(Holiday):SolarRadiation + Humidity:factor(Holiday)
+
    Temperature:Rainfall + factor(FunctioningDay):factor(Seasons) +
    Hour:factor(Holiday) + Rainfall:SolarRadiation + Hour:factor(Seasons) +
    Temperature:Windspeed + Hour:Rainfall + Temperature:factor(Holiday),
    data = newdata)
```

Summary (fitmain)

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -4.998096   0.700424  -7.136 1.04e-12 ***
Temperature                0.472710   0.014677  32.207  < 2e-16 ***
factor(FunctioningDay)Yes 28.388844   0.452020  62.804  < 2e-16 ***
Hour                       0.494004   0.012421  39.773  < 2e-16 ***
Humidity                  -0.160821   0.005073 -31.702  < 2e-16 ***
factor(Seasons)Spring     -3.117559   0.228905 -13.619  < 2e-16 ***
factor(Seasons)Summer     -2.948696   0.290487 -10.151  < 2e-16 ***
factor(Seasons)Winter     -8.182704   0.328078 -24.941  < 2e-16 ***
Rainfall                  -1.560566   0.071915 -21.700  < 2e-16 ***
factor(Holiday)No Holiday  2.769880   0.366447   7.559 4.48e-14 ***
SolarRadiation            -0.837633   0.122331  -6.847 8.04e-12 ***
Windspeed                  0.129788   0.086046   1.508    0.131
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.346 on 8748 degrees of freedom
Multiple R-squared:  0.6524,    Adjusted R-squared:  0.6519
F-statistic:  1492 on 11 and 8748 DF,  p-value: < 2.2e-16
```

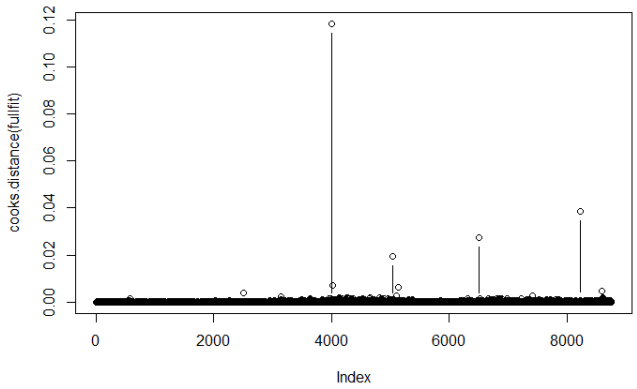
Summary(finalfit)

```
Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     -1.173e+01  3.474e+00  -3.375 0.000741 ***
Temperature                                      2.499e-01  1.191e-01   2.097 0.035991 *
factor(FunctioningDay)Yes                        2.434e+01  3.083e+00   7.895 3.26e-15 ***
Hour                                             7.959e-02  8.922e-02   0.892 0.372341
Humidity                                         2.183e-01  3.800e-02   5.743 9.59e-09 ***
factor(Seasons)Spring                           -8.126e+00  1.024e+00  -7.937 2.33e-15 ***
factor(Seasons)Summer                            1.618e+01  1.865e+00   8.680  < 2e-16 ***
factor(Seasons)Winter                           -8.045e+00  1.346e+00  -5.979 2.34e-09 ***
Rainfall                                        -2.888e+01  2.881e+00 -10.022  < 2e-16 ***
factor(Holiday)No Holiday                        5.052e+00  1.339e+00   3.773 0.000162 ***
SolarRadiation                                   3.645e-01  6.934e-01   0.526 0.599076
Windspeed                                        9.910e-01  3.709e-01   2.672 0.007559 **
Temperature:factor(Seasons)Spring                2.302e-01  3.313e-02   6.950 3.90e-12 ***
Temperature:factor(Seasons)Summer               -7.676e-01  4.357e-02 -17.619  < 2e-16 ***
Temperature:factor(Seasons)Winter               -2.842e-01  3.948e-02  -7.198 6.63e-13 ***
Temperature:Hour                                 2.116e-02  1.964e-03  10.773  < 2e-16 ***
factor(FunctioningDay)Yes:Hour                   6.202e-01  6.289e-02   9.862  < 2e-16 ***
Humidity:factor(Seasons)Spring                   2.007e-02  1.238e-02   1.621 0.104966
Humidity:factor(Seasons)Summer                   2.477e-03  1.832e-02   0.135 0.892468
Humidity:factor(Seasons)Winter                   4.894e-02  1.700e-02   2.879 0.003998 **
Hour:Humidity                                   -7.770e-03  6.519e-04 -11.920  < 2e-16 ***
Humidity:SolarRadiation                          1.211e-01  6.578e-03  18.409  < 2e-16 ***
Temperature:SolarRadiation                      -2.277e-01  1.723e-02 -13.218  < 2e-16 ***
factor(FunctioningDay)Yes:Humidity              -1.851e-01  3.015e-02  -6.141 8.56e-10 ***
Humidity:Rainfall                                2.844e-01  2.976e-02   9.557  < 2e-16 ***
Temperature:Humidity                            -7.497e-03  8.204e-04  -9.138  < 2e-16 ***
factor(Seasons)Spring:SolarRadiation             1.044e+00  3.015e-01   3.462 0.000539 ***
factor(Seasons)Summer:SolarRadiation             7.754e-01  3.566e-01   2.175 0.029682 *
factor(Seasons)Winter:SolarRadiation            -1.415e+00  5.280e-01  -2.679 0.007390 **
Temperature:factor(FunctioningDay)Yes            5.514e-01  9.666e-02   5.705 1.20e-08 ***
factor(FunctioningDay)Yes:Rainfall              -1.809e+00  3.970e-01  -4.557 5.27e-06 ***
factor(Seasons)Spring:Rainfall                   4.677e-01  2.235e-01   2.093 0.036394 *
factor(Seasons)Summer:Rainfall                   4.162e-01  2.512e-01   1.657 0.097595 .
factor(Seasons)Winter:Rainfall                   3.036e+00  4.561e-01   6.657 2.97e-11 ***
factor(Seasons)Spring:Windspeed                  6.537e-01  2.222e-01   2.942 0.003267 **
factor(Seasons)Summer:Windspeed                  9.328e-01  2.895e-01   3.221 0.001280 **
factor(Seasons)Winter:Windspeed                  1.171e-01  3.102e-01   0.377 0.705821
Rainfall:Windspeed                               3.094e-01  7.220e-02   4.286 1.84e-05 ***
SolarRadiation:Windspeed                        -6.125e-01  1.079e-01  -5.676 1.42e-08 ***
Humidity:Windspeed                              -2.191e-02  4.685e-03  -4.677 2.95e-06 ***
factor(Holiday)No Holiday:SolarRadiation        -1.667e+00  4.669e-01  -3.570 0.000359 ***
Humidity:factor(Holiday)No Holiday              -4.392e-02  1.886e-02  -2.329 0.019889 *
Temperature:Rainfall                             6.120e-02  1.893e-02   3.233 0.001230 **
Hour:factor(Holiday)No Holiday                   1.056e-01  4.782e-02   2.209 0.027196 *
Rainfall:SolarRadiation                         -1.630e+00  7.468e-01  -2.183 0.029098 *
Hour:factor(Seasons)Spring                      -8.069e-02  3.135e-02  -2.574 0.010064 *
Hour:factor(Seasons)Summer                      -2.036e-02  3.990e-02  -0.510 0.609818
Hour:factor(Seasons)Winter                      -1.040e-01  4.386e-02  -2.370 0.017814 *
Temperature:Windspeed                            2.560e-02  1.357e-02   1.886 0.059290 .
Hour:Rainfall                                   -2.023e-02  1.266e-02  -1.597 0.110219
Temperature:factor(Holiday)No Holiday           -5.090e-02  3.303e-02  -1.541 0.123349
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.335 on 8709 degrees of freedom
Multiple R-squared:  0.7426,    Adjusted R-squared:  0.7411
F-statistic: 502.4 on 50 and 8709 DF,  p-value: < 2.2e-16
```

2nd Diagnostic- After Dropping Dewpoint:

Cook's Distance

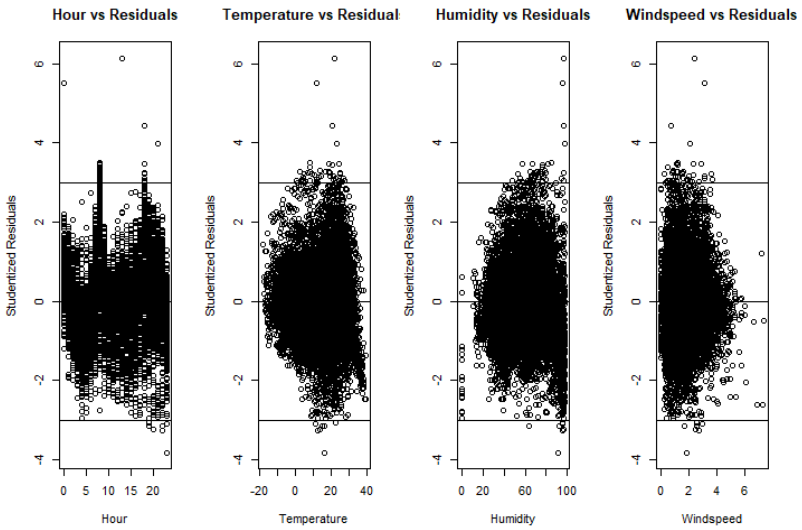


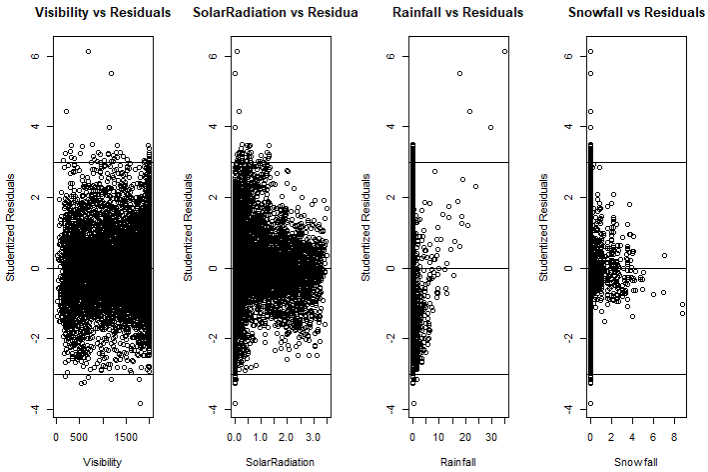
VIF for 2nd Diagnostic

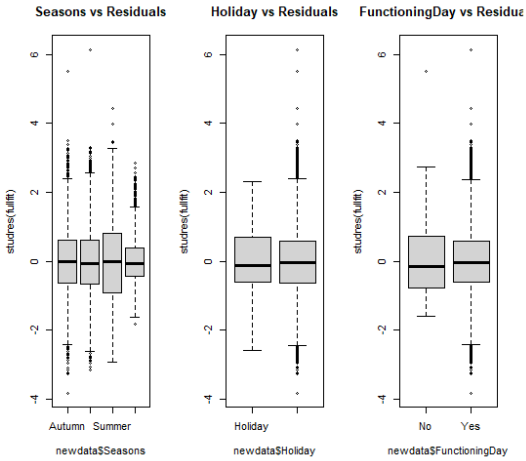
	GVIF	Df	GVIF^(1/(2*Df))
Hour	1.208225	1	1.099193
Temperature	5.040629	1	2.245134
Humidity	2.628551	1	1.621281
Windspeed	1.301764	1	1.140949
Visibility	1.683565	1	1.297522
SolarRadiation	1.946561	1	1.395192
Rainfall	1.070812	1	1.034800
Snowfall	1.113440	1	1.055197
factor(Seasons)	5.452335	3	1.326673
factor(Holiday)	1.023261	1	1.011564
factor(FunctioningDay)	1.080046	1	1.039253

3rd Diagnostic- After taking square root of rented bike count as response:

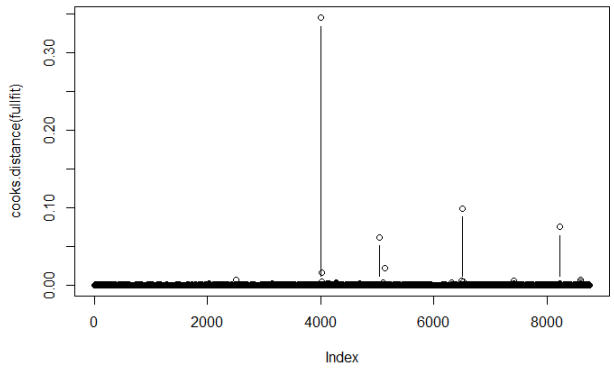
Individual Predictors vs Residuals







Cook's Distance



Anova test: Two are given out of many tests which were done.

Effect of FunctioningDay on overall model

```
Analysis of Variance Table

Model 1: sqrtRentedBike ~ Temperature + Hour + Humidity + factor(Seasons) +
    Rainfall + factor(Holiday) + SolarRadiation + Windspeed +
    Temperature:factor(Seasons) + Temperature:Hour + Humidity:factor(Seasons) +
    Hour:Humidity + Humidity:SolarRadiation + Temperature:SolarRadiation +
    +Humidity:Rainfall + Temperature:Humidity + factor(Seasons):SolarRadiation +
    +factor(Seasons):Rainfall + factor(Seasons):Windspeed + Rainfall:Windspeed +
    SolarRadiation:Windspeed + Humidity:Windspeed + factor(Holiday):SolarRadiation +
    Humidity:factor(Holiday) + Temperature:Rainfall + Hour:factor(Holiday) +
    Rainfall:SolarRadiation + Hour:factor(Seasons) + Temperature:Windspeed +
    Hour:Rainfall + Temperature:factor(Holiday)
Model 2: sqrtRentedBike ~ Temperature + factor(FunctioningDay) + Hour +
    Humidity + factor(Seasons) + Rainfall + factor(Holiday) +
    SolarRadiation + Windspeed + Temperature:factor(Seasons) +
    Temperature:Hour + factor(FunctioningDay):Hour + Humidity:factor(Seasons) +
    Hour:Humidity + Humidity:SolarRadiation + Temperature:SolarRadiation +
    factor(FunctioningDay):Humidity + Humidity:Rainfall + Temperature:Humidity +
    factor(Seasons):SolarRadiation + Temperature:factor(FunctioningDay) +
    factor(FunctioningDay):Rainfall + factor(Seasons):Rainfall +
    factor(Seasons):Windspeed + Rainfall:Windspeed + SolarRadiation:Windspeed +
    Humidity:Windspeed + factor(Holiday):SolarRadiation + Humidity:factor(Holiday) +
    Temperature:Rainfall + Hour:factor(Holiday) + Rainfall:SolarRadiation +
    Hour:factor(Seasons) + Temperature:Windspeed + Hour:Rainfall +
    Temperature:factor(Holiday)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   8714 582360
2   8709 349554  5    232806 1160.1 < 2.2e-16 ***
```

Significance of Seasons-Rainfall interaction:

```
Analysis of Variance Table

Model 1: sqrtRentedBike ~ Temperature + factor(FunctioningDay) + Hour +
    Humidity + factor(Seasons) + Rainfall + factor(Holiday) +
    SolarRadiation + Windspeed + Temperature:factor(Seasons) +
    Temperature:Hour + factor(FunctioningDay):Hour + Humidity:factor(Seasons) +
    Hour:Humidity + Humidity:SolarRadiation + Temperature:SolarRadiation +
    factor(FunctioningDay):Humidity + Humidity:Rainfall + Temperature:Humidity +
    factor(Seasons):SolarRadiation + Temperature:factor(FunctioningDay) +
    factor(FunctioningDay):Rainfall + factor(Seasons):Windspeed +
    Rainfall:Windspeed + SolarRadiation:Windspeed + Humidity:Windspeed +
    factor(Holiday):SolarRadiation + Humidity:factor(Holiday) +
    Temperature:Rainfall + Hour:factor(Holiday) + Rainfall:SolarRadiation +
    Hour:factor(Seasons) + Temperature:Windspeed + Hour:Rainfall +
    Temperature:factor(Holiday)
Model 2: sqrtRentedBike ~ Temperature + factor(FunctioningDay) + Hour +
    Humidity + factor(Seasons) + Rainfall + factor(Holiday) +
    SolarRadiation + Windspeed + Temperature:factor(Seasons) +
    Temperature:Hour + factor(FunctioningDay):Hour + Humidity:factor(Seasons) +
    Hour:Humidity + Humidity:SolarRadiation + Temperature:SolarRadiation +
    factor(FunctioningDay):Humidity + Humidity:Rainfall + Temperature:Humidity +
    factor(Seasons):SolarRadiation + Temperature:factor(FunctioningDay) +
    factor(FunctioningDay):Rainfall + factor(Seasons):Rainfall +
    factor(Seasons):Windspeed + Rainfall:Windspeed + SolarRadiation:Windspeed +
    Humidity:Windspeed + factor(Holiday):SolarRadiation + Humidity:factor(Holiday) +
    Temperature:Rainfall + Hour:factor(Holiday) + Rainfall:SolarRadiation +
    Hour:factor(Seasons) + Temperature:Windspeed + Hour:Rainfall +
    Temperature:factor(Holiday)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   8712 351404
2   8709 349554  3    1849.9 15.363 5.753e-10 ***
---
```