Objective

To analyze the data set of rented bicycles in Washington D.C. to make statistical inferences regarding how many bikes the company rents related to factors such as temperature, humidity, and wind speed. Through the use of statistical inferences, we intend to gain insight into consumer behavior regarding whether or not to rent a bicycle.

Introduction

Renting a bike is a reality of our time. However, surprisingly the first shared bike service was created in Amsterdam in 1965 (Readers Digest Deutschland - 2011)[1] since then bike-sharing has become increasingly popular. There are many advantages when choosing to cycle rather than driving: it's environmentally friendly, very economical, and is a great way to stay healthy.

According to the Earth Policy Institute (2013)[2], "more than 500 cities in 49 countries host advanced bike-sharing programs, with a combined fleet of over 500,000 bicycles. Capital Bike Share (CaBi) is the bicycle rental system that serves Washington D.C according to the CaBi website through Washington, Alexandria, and Arlington, there are over 200 bike stations."[3] Founded in September 2010, the system is the second-largest bike sharing in the United States behind New York City's Citi Bike.[4]

The price for a 30 minutes ride is $2 dollars[5], this is less than the metro ticket ($2.25 to $6 per trip during peak hours)[6]. Since the service has competitive prices, we would like to analyze what influences the consumer's decision to rent a bike other than price. Given the growing competition of bicycle rental services in the world, It is important to analyze how external factors influence the demand for rented bike service.

In the present project, we will analyze the data set of the Washington DC bicycle rental service through the year of 2011 to make inferences of the user's behavior according to the weather, month, working days and holidays throughout the year.

Sampling Design

Our data set was taken from the UCI Machine Learning Repository website. The data was facilitated by Capital Bikeshare, a program jointly owned and sponsored by the District of

[1] "Runde Sache". *Readers Digest Deutschland* (in German). 06/11: 74–75. June 2011
[2] Bike-Sharing Programs Hit the Streets in Over 500 Cities Worldwide; Earth Policy Institute; Larsen, Janet; 25 April 2013
[3] Larsen, Janet. 2013. "Bike Sharing Goes Global". *Grist*. Accessed November 26 2019. https://grist.org/cities/bike-sharing-programs-hit-the-streets-in-over-500-cities-worldwide/.
[4] Motivate International, Inc. "How Metro DC's Bikeshare System Works." Capital Bikeshare. Accessed November 20, 2019. https://www.capitalbikeshare.com/how-it-works.
[5] Motivate International, Inc. "How Metro DC's Bikeshare System Works." Capital Bikeshare. Accessed November 20, 2019. https://www.capitalbikeshare.com/how-it-works.
[6] "Navigating Washington, DC with Metro." Washington.org, July 3, 2019. https://washington.org/navigating-dc-metro.

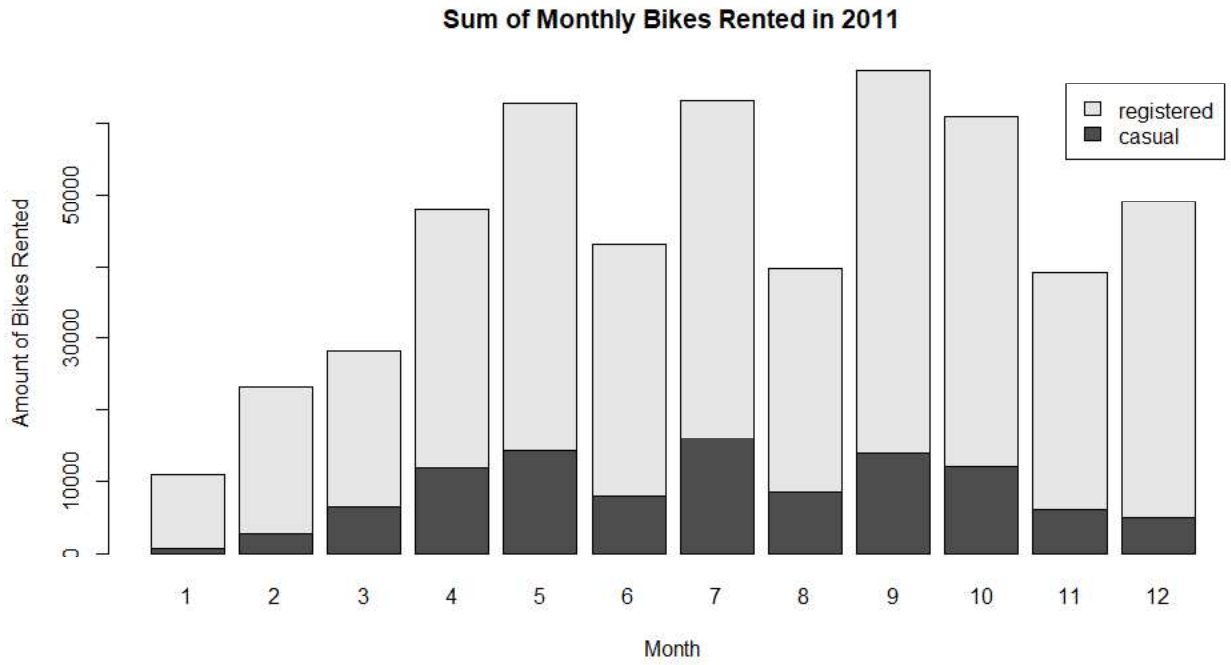
Columbia and Arlington County, VA. The population includes bike rentals of registered and casual users from the year 2011.

A sample size of 160 bike rentals records where selected through a Stratified Random Sampling. The data was divided into four strata based on quarter (seasons) namely Q1, Q2, Q3, Q4. In each stratum, 40 samples were randomly selected to represent each quarter. This sampling method is the most appropriate for accuracy purposes as there might be huge variations in bike usage between the different seasons due to summer breaks for school children, summer vacations etc. Taking samples from different seasons (quarter) will provide the most accurate representation of results on how the variables influence the bike rentals.

Analysis & Commentary

Univariate Analysis

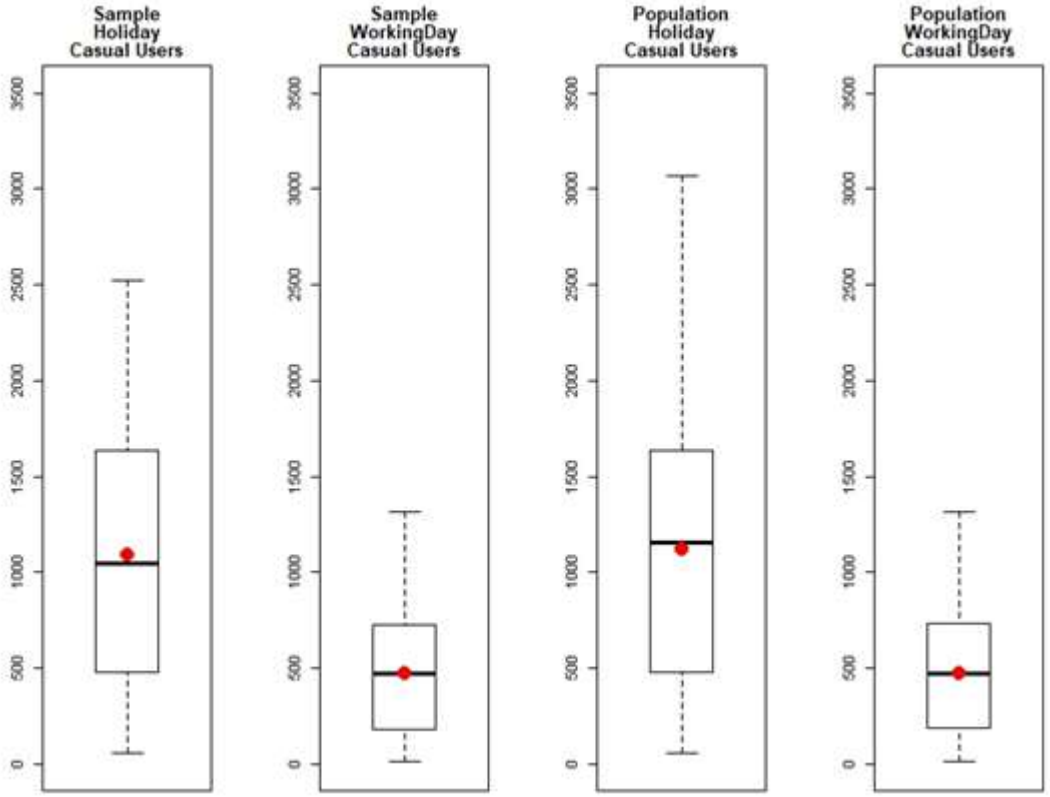
Average Number Of Total Bike Rentals During Different Months Of The Year



Our bike counts throughout the year 2011, shows a minimum of 10,992 in January, and a maximum of 67,358 in September. On average, the company will rent 44,631 bikes in a given month throughout the year.

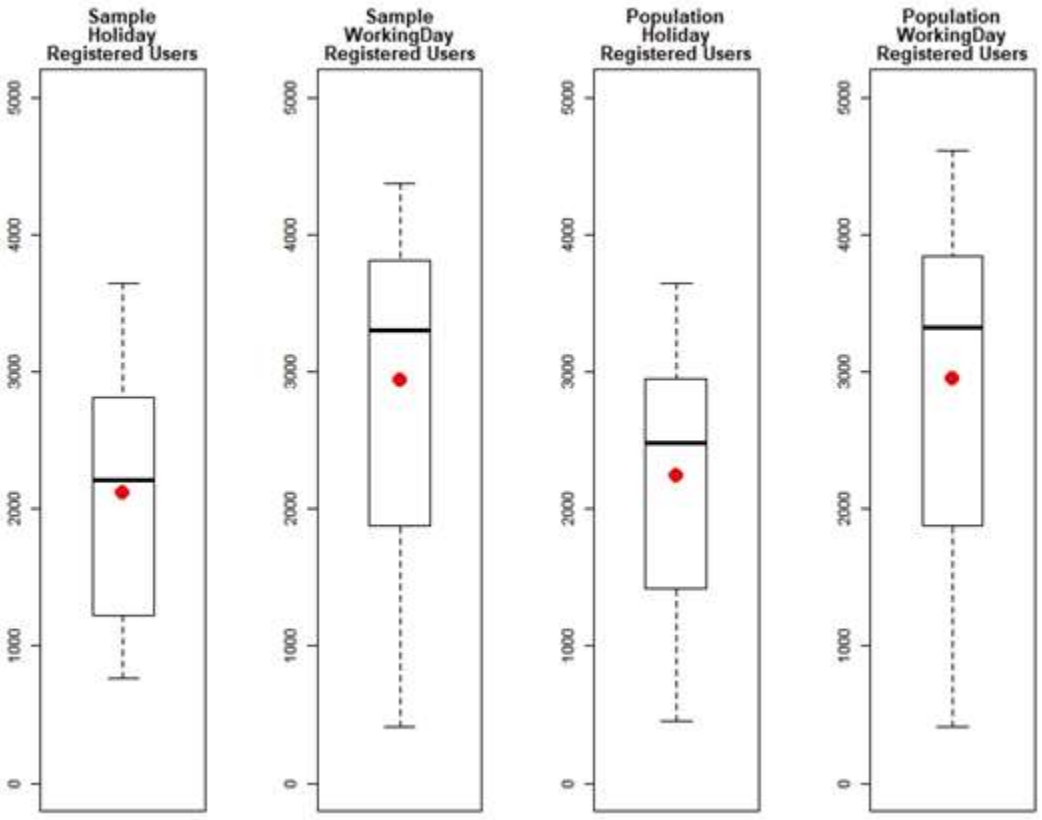
Casual Users In Holiday And Working Day

Casual User



	Sample of Casual Users in Holiday	Sample Casual Users in Working Day	Population Casual Users in Holiday	Population Casual Users in Working Day
Q_1	54	15	54	9
Q_2	480	182	476	186
Q_3	1047	470	1156	474
Q_4	1633	726	1636	729
Q_5	2521	1318	3065	1318
Range	2467	1303	3011	1309
IQR	1153	544	1160	543
Mean (red point)	1087.857	470.6036	1120.852	473.416

The sample and population's median mean and larger range and interquartile range have very similar characteristics. The sample is representative of the population. It is important to notice that the larger range for casual user is larger during holidays than working days. The sample cannot fully show the left skewness of the population of casual users in holiday.



	Sample of Registered Users in Holiday	Sample Registered Users in Working Day	Population Registered Users in Holiday	Population Registered Users in Working Day
Q_1	768	416	451	416
Q_2	1221	1875.5	1416	1878
Q_3	2211	3307	2484	3322
Q_4	2809	3808.5	2951.5	3840
Q_5	3647	4372	3647	4614
Range	2879	3956	3196	4198
IQR	1588	1933	1535.5	1962
Mean (red point)	2117.796	2941.252	2242.965	2951.64

For sample data of the registered users, the median and mean is higher, and the range and interquartile range is larger in working days compared during the holidays. Also, there is a left-skewed distribution in holiday and working day. The level of skewness is larger in the working day. The sample data shows similar characteristic of the population data.

Sampling Distribution

To show the confidence interval of casual, registered, and total users, we will do 1000 times of stratified sampling with sample size 160 without replacement from the 2011 population to create the sampling distribution.

Casual Users

Mean of sample mean = 672.7709
95% Confidence interval = [586.5777, 758.9641]
Population mean = 677.4027

Registered Users

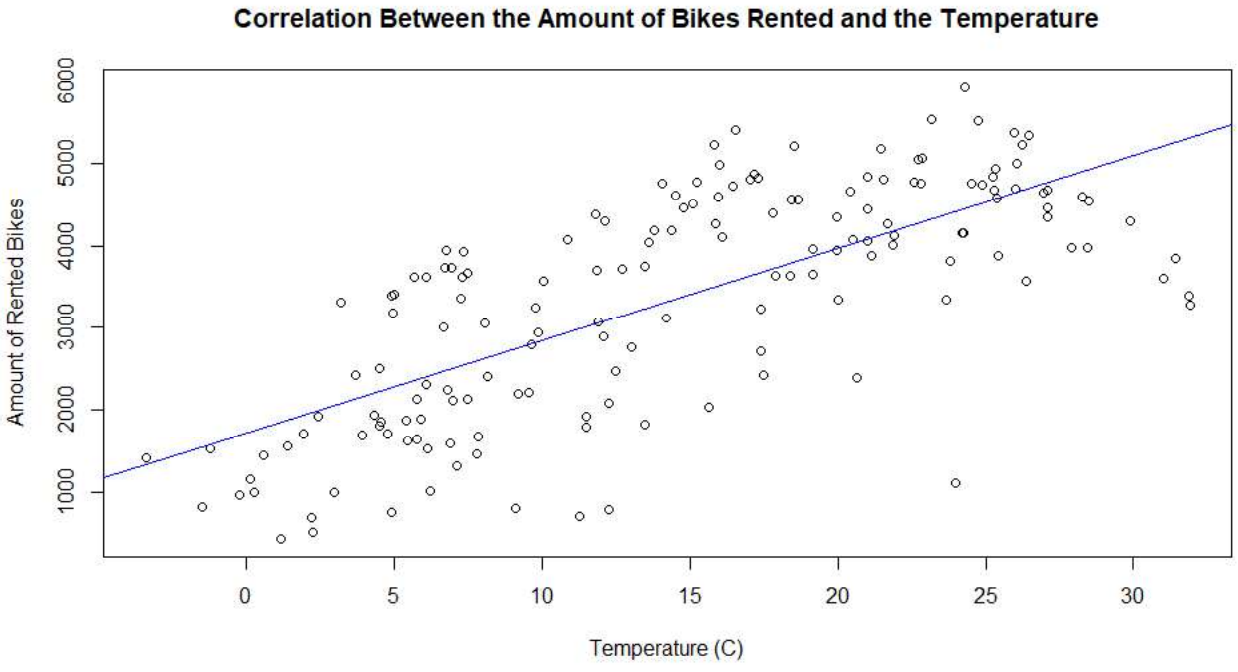
Mean of sample mean = 2720.115
95% Confidence interval = [2555.852, 2884.377]
Population mean = 2728.359

Total Users

Mean of sample mean = 3392.885
95% Confidence interval = [2179.249, 3606.522]
Population mean = 3405.762

Bivariate Analysis:

The Relationship Between The Number Of Registered And Casual Users Bike Rentals And Correlated With The Temperature.



y= 1715.48 + 112.42x

H0: there is no relationship between the temperature and the amount of rented bikes.

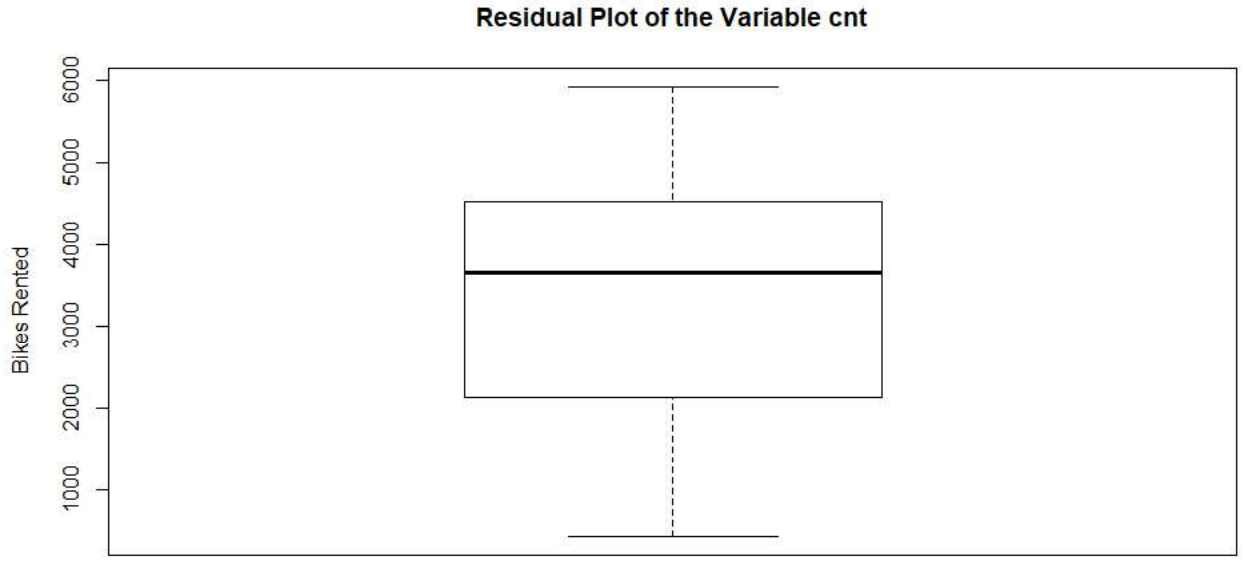
H1: there is a relationship between the temperature and the amount of rented bikes.

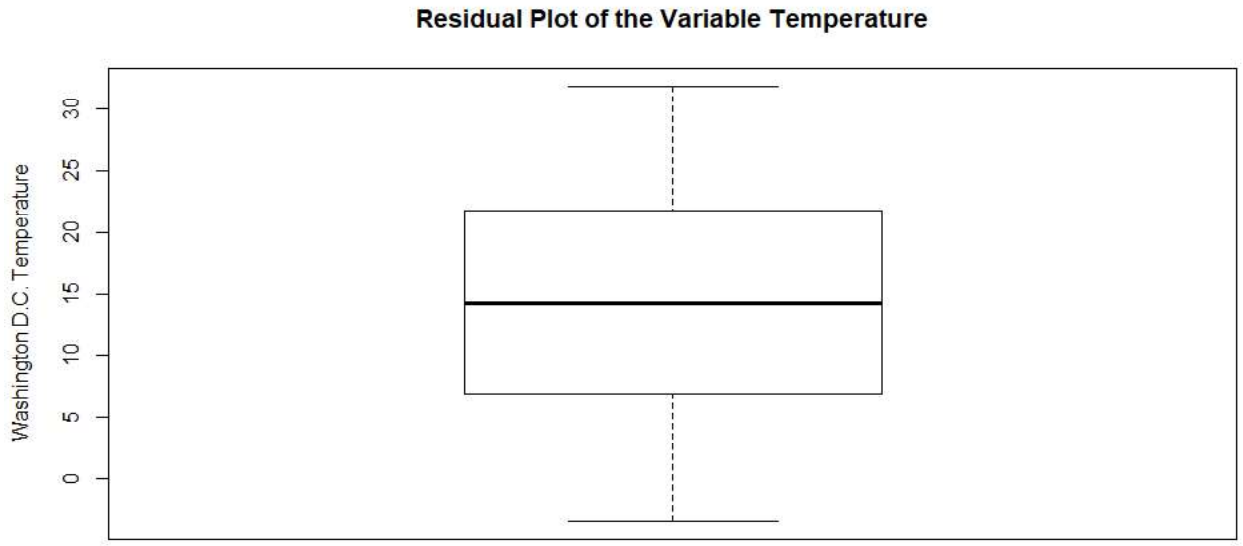
Decision Point n= 160 → DP=0.170

r=0.71563

0.170 < r → We reject H0 and conclude H1, there is a linear correlation between the temperature and the amount of rented bikes.

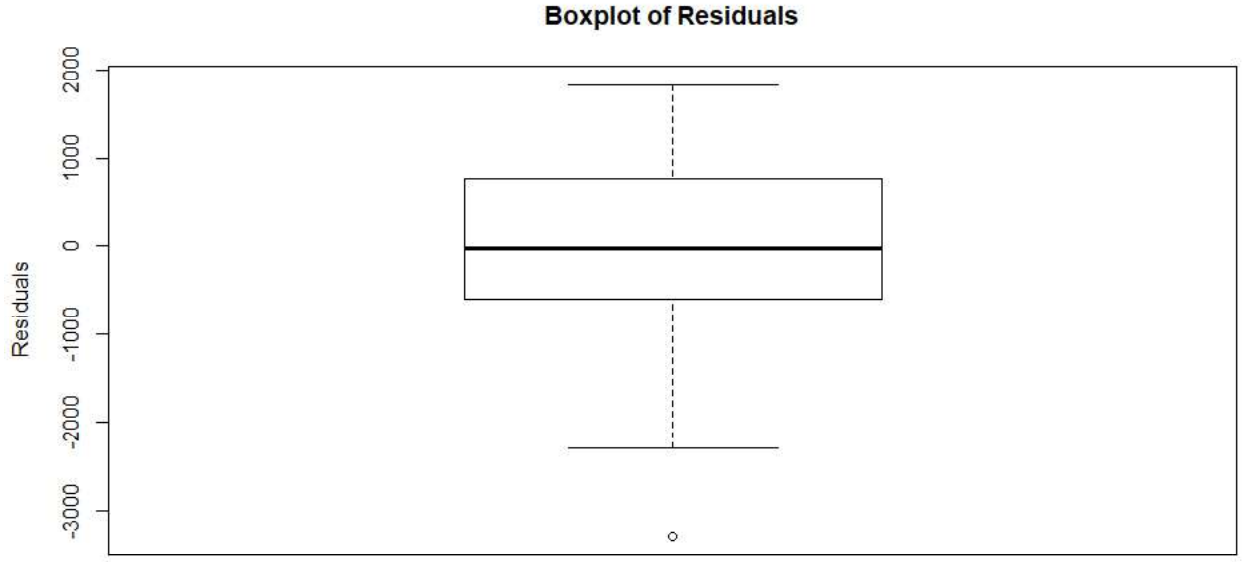
Relationship between number of bikes rented and weather temperature seem to have a linear relationship. Correlation is moderately strong (r=0.71563). The regression equation means that for every additional unit in the temperature (x), the company should expect to rent 112.42 additional bikes on a given day. Also, the variables "cnt" (number of bikes rented per day) and "temp" (Washington D.C.'s Temperature) have a normal distribution with 0 outliers, as seen below:

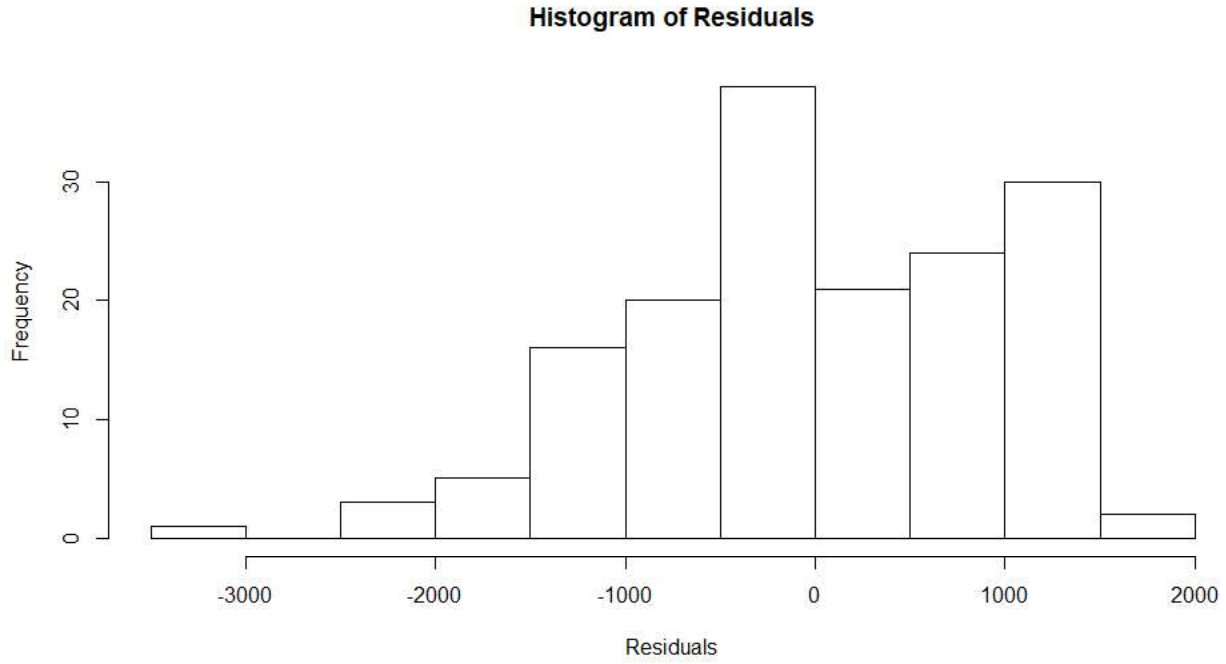




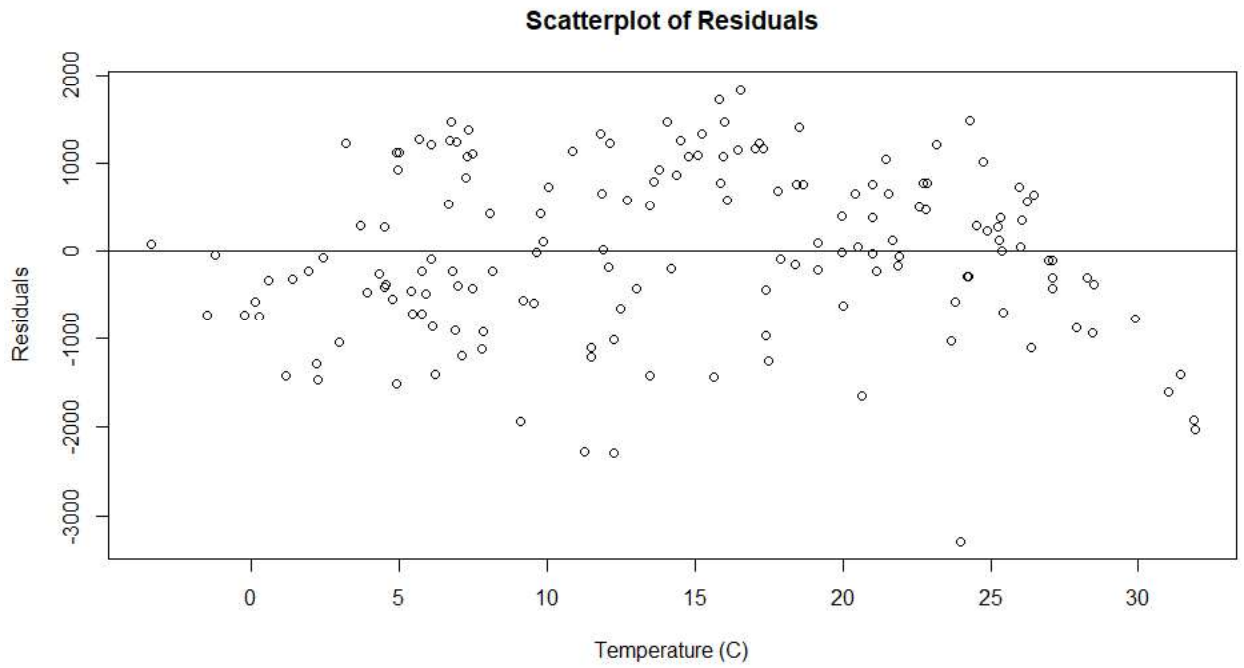
Validating Regression Assumptions:

Normality:





Both the histogram and residuals show a normal distribution with a slight skewness to the left. The histogram shows 3 outliers, but compared to the total sample size, they are not significant to influence our regression analysis.



Linearity:

The temp~cnt[7] graph shows a linear relationship with a moderately strong correlation coefficient of 0.71563.

Homoscedasticity:

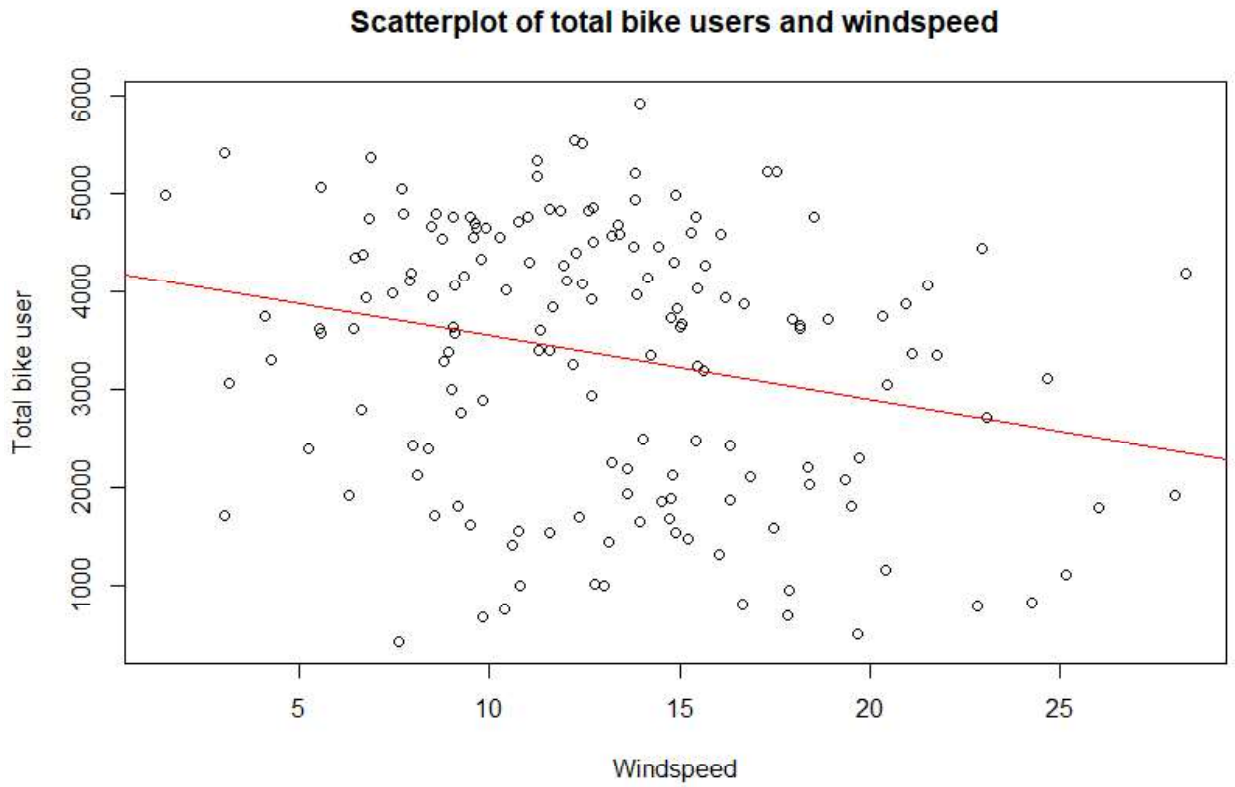
Our plot shows a homoscedastic relationship because residuals are scattered randomly and do not show a fan-shaped pattern.

Conclusion

Our assumptions test tells us that a linear regression is not valid as our normality did not pass the test. Although, logically it makes sense that a moderately warm temperature (15 to 30 C) will call for higher number of bikes rented, a non-linear regression analysis needs to be made to fully validate this assumption.

[7] Refer to page #

The Relationship Between the Number of Rented Bikes And Wind Speed



$$y=4199.73 - 65.25x$$

H0: There is no linear relationship between wind speed and the total number of rented bikes.

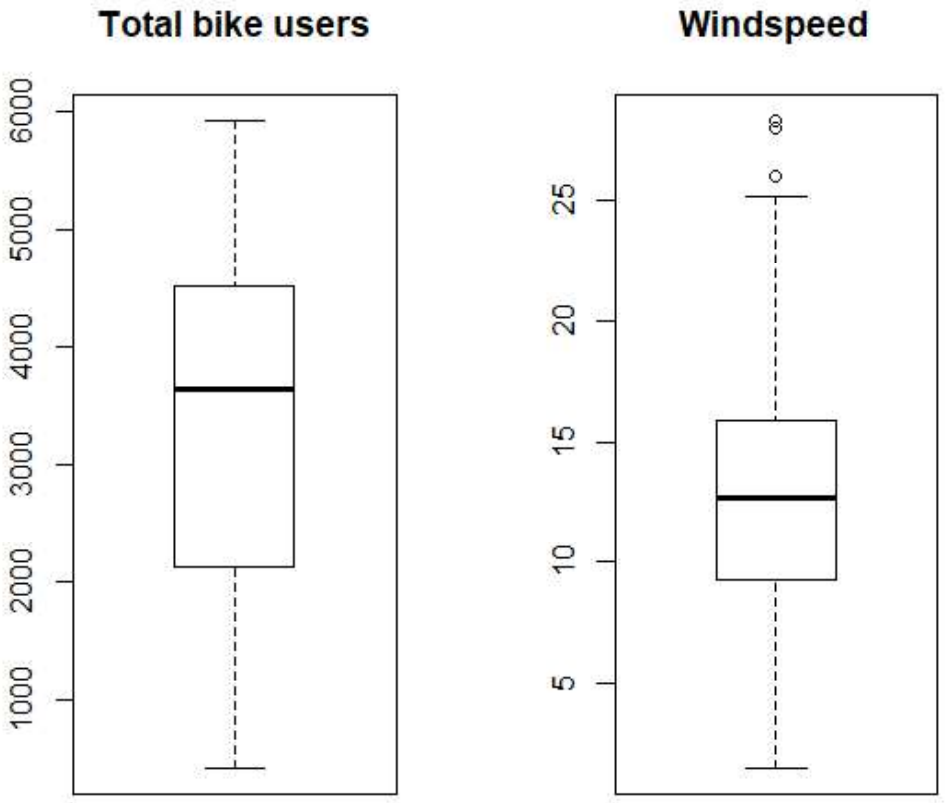
Ha: There is a linear relationship between wind speed and the total number of rented bikes.

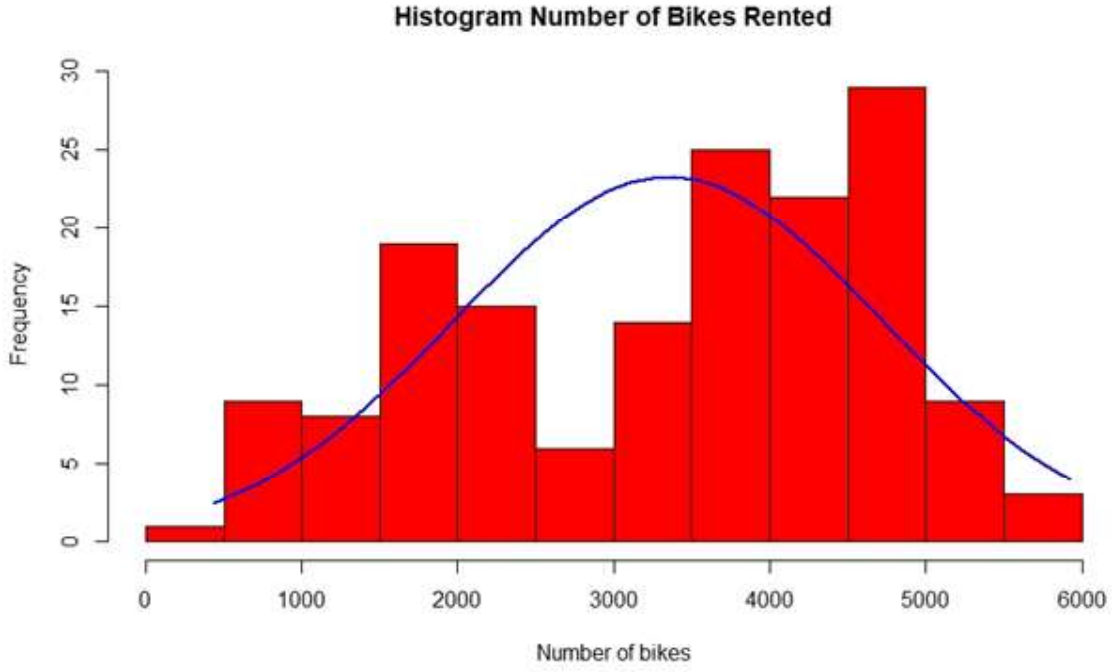
Decision Point n= 160 → DP=0.170

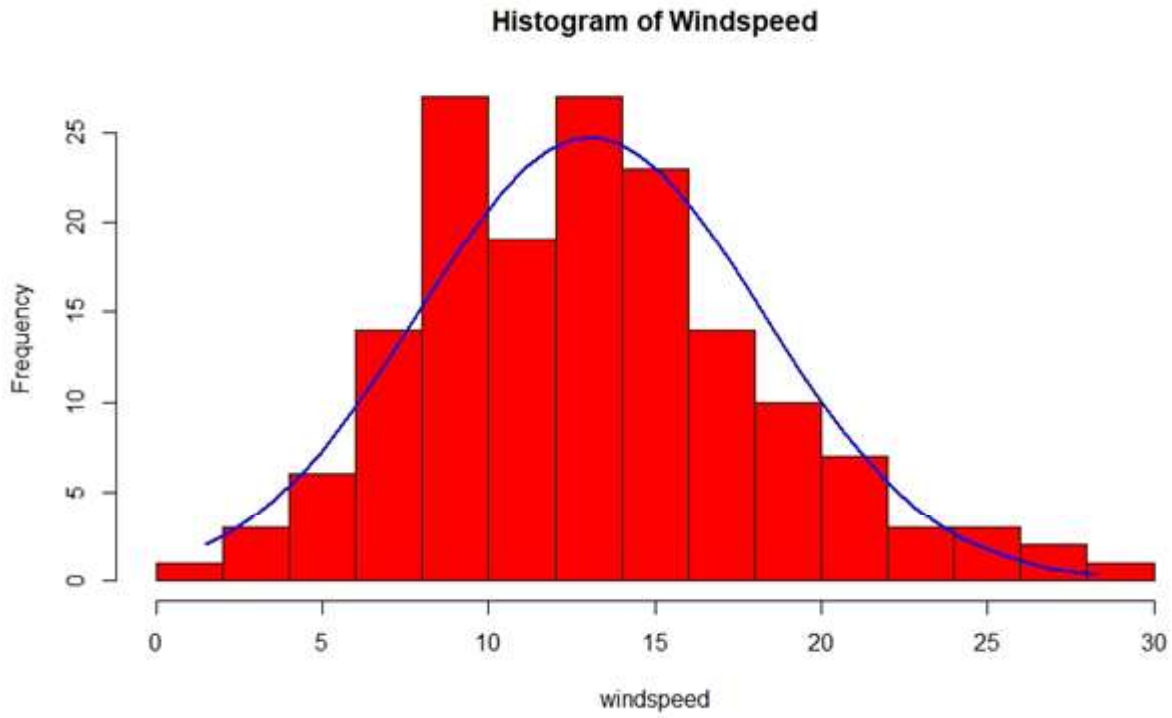
r= -0.246061

0.170 <|r| → We reject H null , and accept the alternative hypothesis.

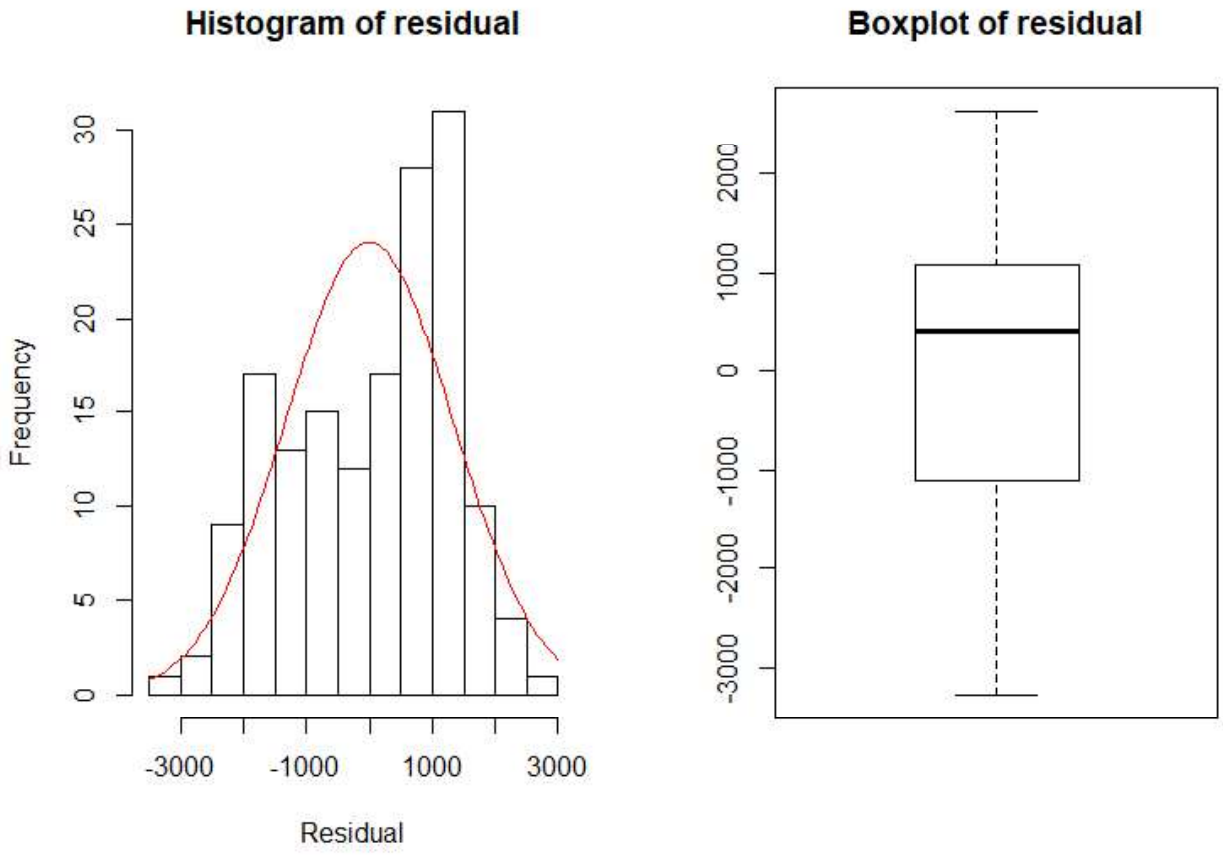
Relationship between the total number of bikes rented and wind speed seem to have a linear relationship. However, correlation is low (r=-0.246061). The regression equation means that for every square meter in the wind speed (x), the company should expect to rent minus 65.25 bikes on a given day. The variable “cnt” (number of bikes rented per day) has a normal distribution with zero outliers, however wind speed has outliers, as seen below:


Total bike users
Windspeed
1000
2000
3000
4000
5000
6000
5
10
15
20
25


Histogram Number of Bikes Rented
Frequency
0
5
10
15
20
25
30
0
1000
2000
3000
4000
5000
6000
Number of bikes



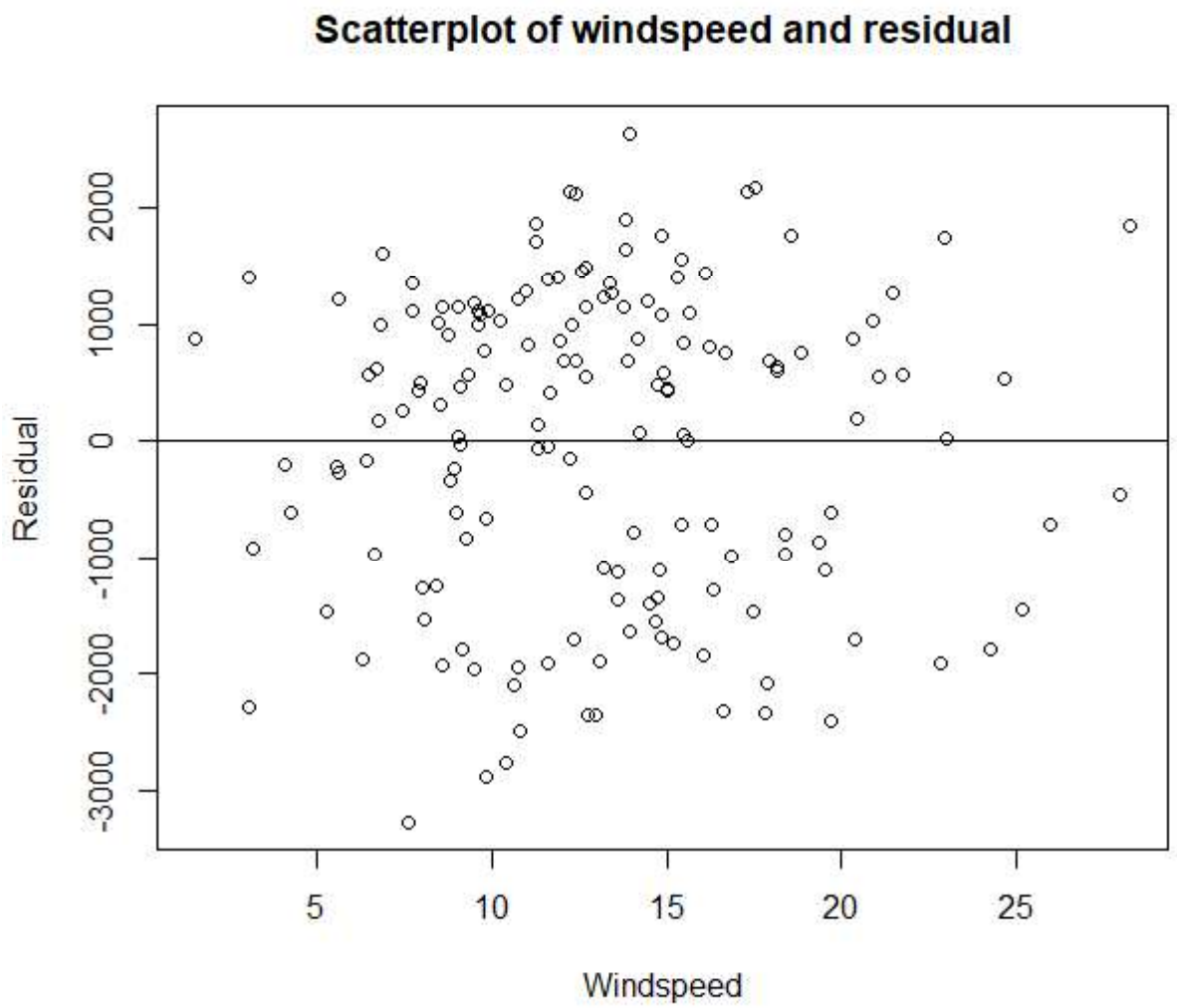
Validating Regression Assumptions:



Normality:

Both the histogram and boxplot, they show an approximately normal distribution.

Linearity: there is a linear relationship between wind speed and total number of bikes rented (p-value = 0.00171 < 0.05); however, the linear relationship is weak with the correlation coefficient is – 0.246061.



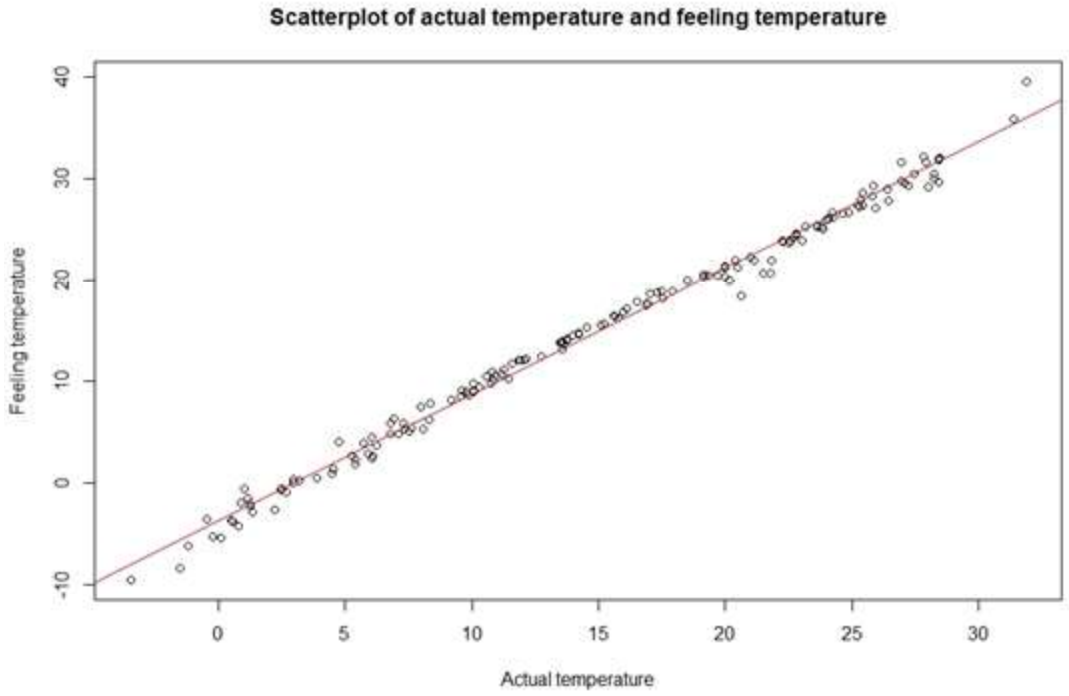
Homoscedasticity:

Our plot shows a homoscedasticity - there are not any obvious patterns.

Conclusion

Our assumptions test tells us that a linear regression is valid, since the three assumptions for linear regression passed the test. A low correlation probably suggests that much of the variation of the response variable (total number of rented bikes) is unexplained by the predictor (wind speed). Since the p-Value is smaller than the significance level (< 0.05), we reject the null hypothesis that the coefficient β of the predictor is zero.

The Relationship between actual temperature and the feeling temperature



y=-3.678878 + 1.240982x

H0: there is no relationship between feeling temperature and actual temperature in the population.

H1: there is a relationship between feeling temperature and actual temperature in the population.

The correlation coefficient is 0.9962529, the decision point with 160 sample size is 0.170, we reject H0, and conclude that there is a relationship between feeling temperature and actual temperature in the population.

Refer to the above regression, it is expected feeling temperature will be increased in 1.240982 when actual temperature increases by one degree celsius.

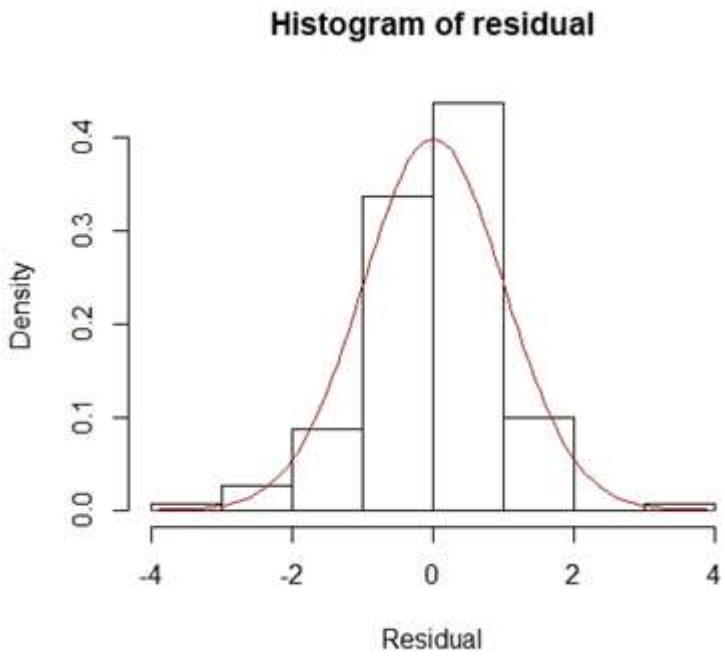
Also, we can expect that the feeling temperature equal to actual temperature when the actual temperature is 15.26619. If the actual temperature is smaller than temperature, feeling temperature is lower than the actual temperature, and vice versa.

Assumption test:

Linearity:

Refer to the scatterplot, there is a highly visible linear relationship between feeling temperature and actual temperature. The assumption of linearity is valid.

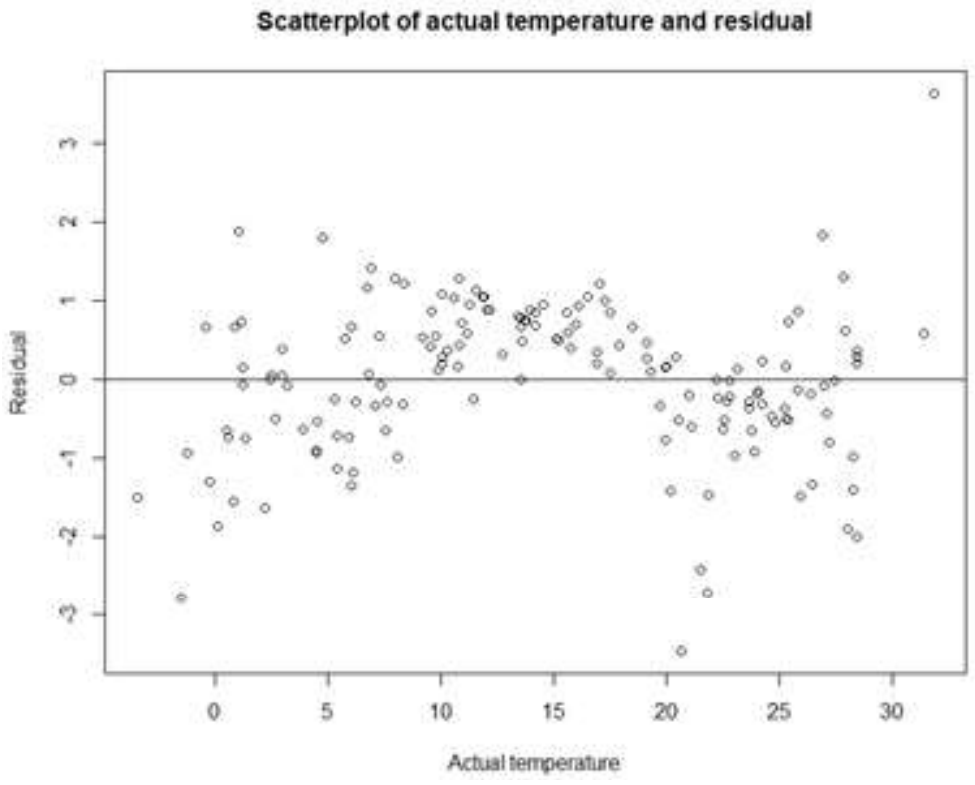
Normality



22

The distribution of the residual is close to normal distribution. Assumption of normality is valid.

Homoscedastic

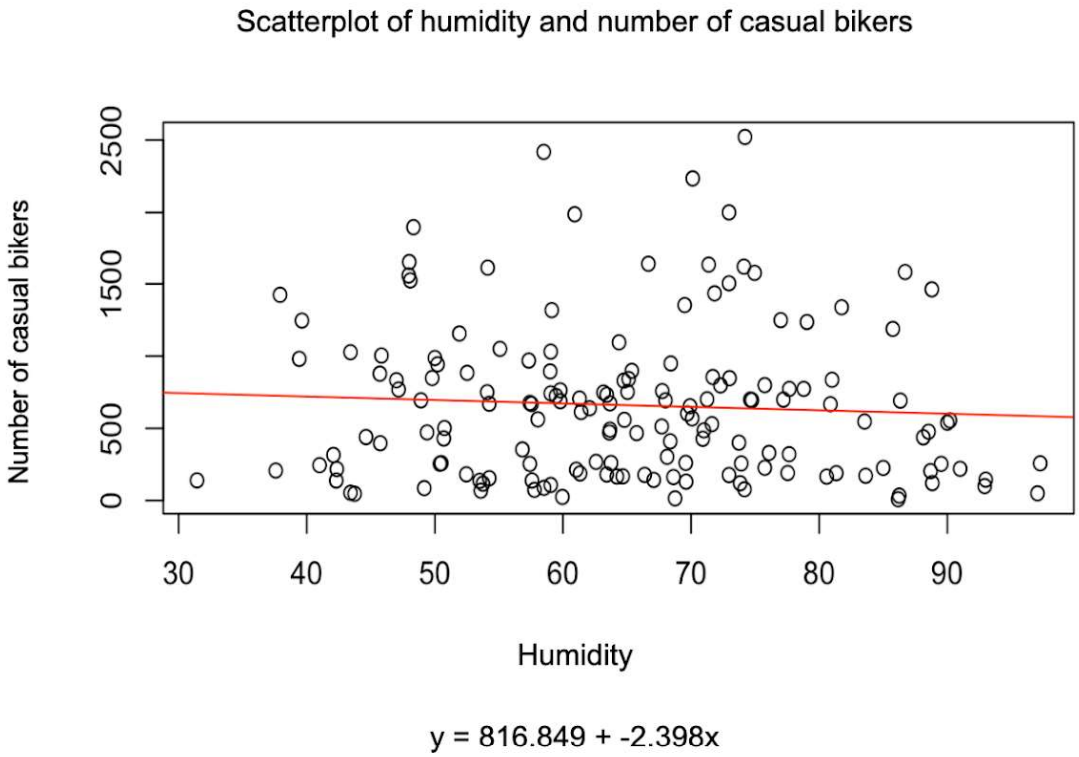


Refer to the residual plot above, there is a parabola pattern between residual and the independent variable, so the assumption of homoscedastic is invalid.

Conclusion

Overall, we can deduce that this actual temperature is related to the feeling temperature, but it is not a linear relationship.

The Relationship Between Humidity And The Number Of Casual Bike Rentals For That Day



The regression equation means that we expected that the number of casual bikers will decrease by 2.398 when humidity increases by 1%. This is perhaps because the casual bikers tend to be those that are cycling for lifestyle/entertainment purposes and not for commuting reasons. Casual bikers may be drawn to lower humilities.

H0: there is no relationship between humidity and the number of casual bike rentals in the population.

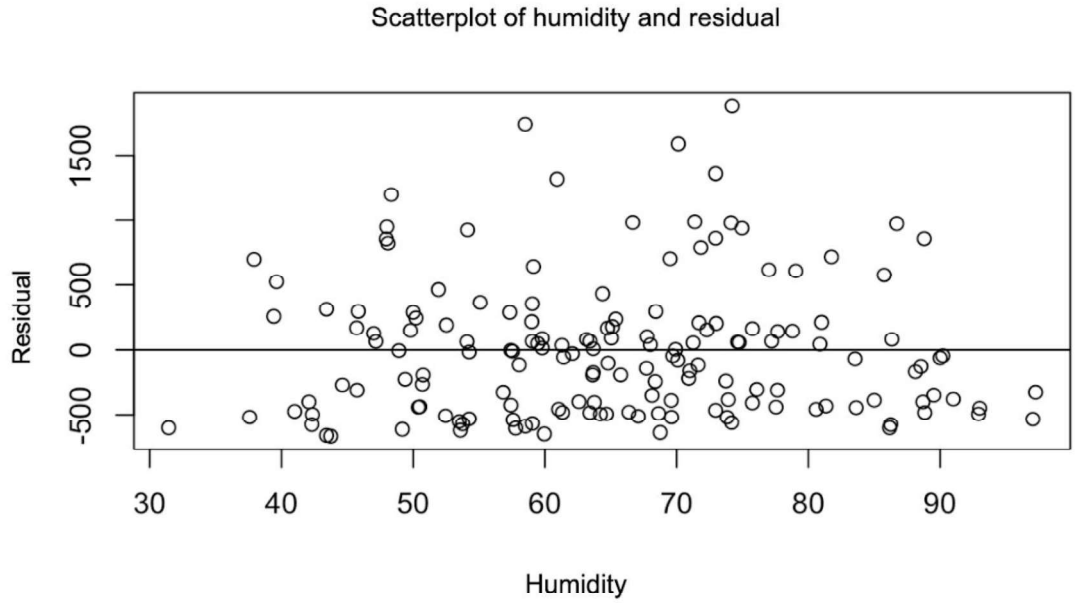
Ha: there is a relationship between humidity and number of casual bike rentals in the population.

r= -0.06437898
N → 160 → DP=0.17

We fail to reject H0, there is no linear correlation between humidity and number of casual bike rentals in the population.

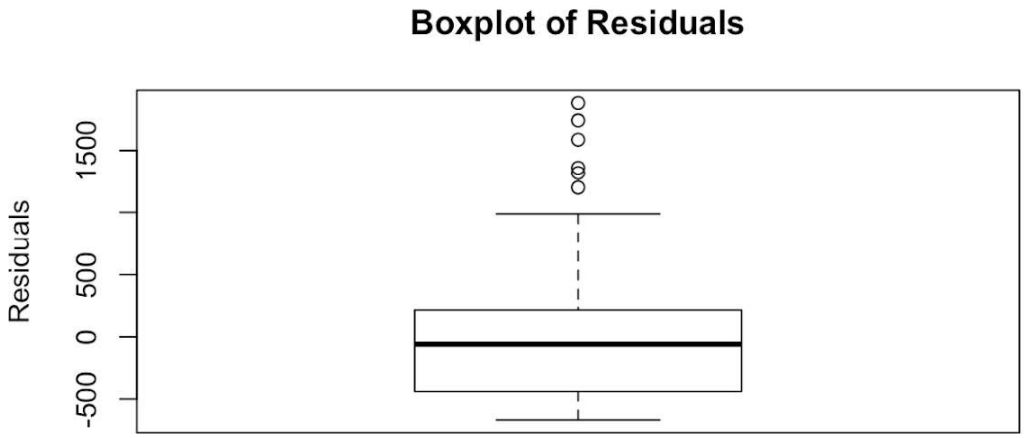
Assumption test:

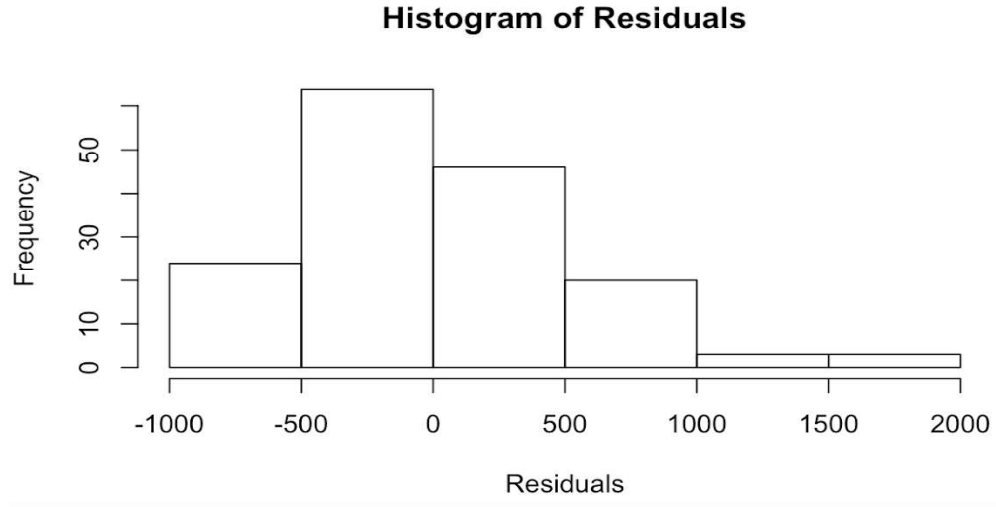
Homoscedasticity & Linearity



Refer to the residual plot, there is linearity as there is no apparent curves. The residual scatter plot also shows homoscedasticity as there is no fan shape.

Normality:





The boxplot and histogram of residuals shows normality. The regression model is therefore valid. We can say there is no correlation between humidity and number of casual bike users.

Dependency Of The Wind Speed And The Cold Temperature Feeling

The minimum and the maximum windspeed are 1.500244 and 28.29243 respectively, and the range of wind speed is 26.79218, which is rounded to 27. The wind speed is divided into 3 equal classes:

	Wind Speed 1	Wind Speed 2	Wind Speed 3
'Cold' Feeling	22	41	11
No 'Cold' Feeling	32	46	8

H0: The event of cool feeling and wind speed are independent in population

H1: The event of cool feeling and wind speed are dependent in population

X2 → 1.7226

Decision point with Degree of freedom 1 = 5.99

P-value → 0.4226

We fail to reject H0, we do not have enough evidence to conclude that the "cold" temperature and the wind speed are independent.

Limitations

Undoubtedly, it is a time series data. Referred to the univariate analysis, there is a seasonality in casual, registered and total number of rental bikes obviously. Some errors exist as we did not separate the data into seasonality and trend, the regression analysis should be focus on the trend. So, it is very easy to find out some non-linear regression model now.

Moreover, the sample size is not enough, there are only 160 samples out of 365. The accuracy of the analysis can be significantly improved. However, as we did not decompose the seasonality and trend of the data, the variance of residual will be too high if we contain too many data from 2011 dataset.

Final Conclusion

Analysis shows that there is a relationship between the number of registered and casual users bike rentals and temperature and working or holidays in Washington D.C. of the year 2011. For casual users, an increasing number of bike rental is evident during warmer temperatures (late spring and summer) and during national holidays. In contrast, registered users show constant bike usage during workdays and in colder temperatures. Also, our study shows that bike rental is independent from humidity and wind speed.

Appendix:

Sampling

Code:

Edward- Stratified Sampling

```
#draw random sample for analysis
set.seed(123)
data2011<-subset(data, year==2011)
data2012<-subset(data, year==2012)
sample<-numeric()
for (i in 1:4)
{
  sample<-rbind(sample,subset(data, year==2011 & season == i)[sample(1:nrow(subset(data, year==2011 & season == i)),40,replace = FALSE),])
}
```

Univariate Analysis:

Code:

DAVID - Average number of total bike rentals during different months of the year

```
##Libraries Used
library(dplyr)
library(tidyr)

##R-Queries for Pivot Table
pivot <- day_sample %>%
select(month,cnt)%>%
group_by(month)%>%
summarise(cntsum = ((sum(cnt))))

head(pivot)

##Bar Chart of month~cntsum
cnt_month_graph <- barplot(pivot$cntsum ~pivot$month, xlab = "Month", ylab = "Bike Count")
text(cnt_month_graph, day_sample$cnt, labels=day_sample$cnt, pos = 3, offset = 9)

##Stats on Pivot$cntsum
summary(pivot$cntsum)
sd(pivot$cntsum)
```

Edward - Sampling distribution and the confidence interval of casual, registered and total users

```
#Sampling distribution for causal
n_sampling<-1000
dist<-numeric()
set.seed(123)
for (j in 1:n_sampling)
{
  sam_list<-numeric()
  for (i in 1:4)
  {
    draw<-sample(nrow(subset(data2011, season == i)),40,replace = FALSE)
    sam_list<-c(sam_list,subset(data2011, season == i)[draw,]$casual)
  }
  dist<-c(dist,mean(sam_list))
}
print(c(mean(dist)-qnorm(0.975)*(sd(data2011$casual)/sqrt(160)),
mean(dist)+qnorm(0.975)*(sd(data2011$casual)/sqrt(160))))
mean(data2011$casual)
mean(dist)

#Sampling distribution for registered
n_sampling<-1000
dist<-numeric()
set.seed(123)
for (j in 1:n_sampling)
{
  sam_list<-numeric()
  for (i in 1:4)
  {
    draw<-sample(nrow(subset(data2011, season == i)),40,replace = FALSE)
    sam_list<-c(sam_list,subset(data2011, season == i)[draw,]$registered)
  }
  dist<-c(dist,mean(sam_list))
}
print(c(mean(dist)-qnorm(0.975)*(sd(data2011$registered)/sqrt(160)),
mean(dist)+qnorm(0.975)*(sd(data2011$registered)/sqrt(160))))
mean(data2011$registered)
mean(dist)

#Sampling distribution for total
n_sampling<-1000
dist<-numeric()
set.seed(123)
for (j in 1:n_sampling)
{
  sam_list<-numeric()
  for (i in 1:4)
  {
    draw<-sample(nrow(subset(data2011, season == i)),40,replace = FALSE)
    sam_list<-c(sam_list,subset(data2011, season == i)[draw,]$cnt)
  }
  dist<-c(dist,mean(sam_list))
}
```

```
print(c(mean(dist)-qnorm(0.975)*(sd(data2011$cnt)/sqrt(160)),
mean(dist)+qnorm(0.975)*(sd(data2011$cnt)/sqrt(160))))
mean(data2011$cnt)
mean(dist)
```

Edward - Univariate analysis of casual, registered and total users based on working day

```
# casual by workingday
workingday.name<-c('Holiday','WorkingDay')
par(mfrow=c(1,4))
for (i in 0:1)
{
 boxplot(subset(sample, workingday == i)$casual, outline = TRUE, main =
c('Sample',workingday.name[i+1],'Casual Users'), ylim = c(0,3500))
 points(mean(subset(sample, workingday == i)$casual), col= "red", pch = 19, cex = 2)
}
for (i in 0:1)
{
 boxplot(subset(data2011, workingday == i)$casual, outline = TRUE, main =
c('Population',workingday.name[i+1],'Casual Users'), ylim = c(0,3500))
 points(mean(subset(data2011, workingday == i)$casual), col= "red", pch = 19, cex = 2)
}
par(mfrow=c(1,1))
fivenum(subset(sample, workingday == 0)$casual)
fivenum(subset(sample, workingday == 1)$casual)
mean(subset(sample, workingday == 0)$casual)
mean(subset(sample, workingday == 1)$casual)
fivenum(subset(data2011, workingday == 0)$casual)
fivenum(subset(data2011, workingday == 1)$casual)
mean(subset(data2011, workingday == 0)$casual)
mean(subset(data2011, workingday == 1)$casual)

# register by workingday
par(mfrow=c(1,4))
for (i in 0:1)
{
 boxplot(subset(sample, workingday == i)$registered, outline = TRUE, main =
c('Sample',workingday.name[i+1],'Registered Users'), ylim = c(0,5000))
 points(mean(subset(sample, workingday == i)$registered), col= "red", pch = 19, cex = 2)
}
for (i in 0:1)
{
 boxplot(subset(data2011, workingday == i)$registered, outline = TRUE, main =
c('Population',workingday.name[i+1],'Registered Users'), ylim = c(0,5000))
 points(mean(subset(data2011, workingday == i)$registered), col= "red", pch = 19, cex = 2)
}
```

```
par(mfrow=c(1,1))
fivenum(subset(sample, workingday == 0)$registered)
fivenum(subset(sample, workingday == 1)$registered)
mean(subset(sample, workingday == 0)$registered)
mean(subset(sample, workingday == 1)$registered)
fivenum(subset(data2011, workingday == 0)$registered)
fivenum(subset(data2011, workingday == 1)$registered)
mean(subset(data2011, workingday == 0)$registered)
mean(subset(data2011, workingday == 1)$registered)

# total by workingday
par(mfrow=c(1,4))
for (i in 0:1)
{
  boxplot(subset(sample, workingday == i)$cnt, outline = TRUE, main =
c('Sample',workingday.name[i+1],'Total Users'), ylim = c(0,6500))
  points(mean(subset(sample, workingday == i)$cnt), col= "red", pch = 19, cex = 2)
}
for (i in 0:1)
{
  boxplot(subset(data2011, workingday == i)$cnt, outline = TRUE, main =
c('Population',workingday.name[i+1],'Total Users'), ylim = c(0,6500))
  points(mean(subset(data2011, workingday == i)$cnt), col= "red", pch = 19, cex = 2)
}
par(mfrow=c(1,1))
fivenum(subset(sample, workingday == 0)$cnt)
fivenum(subset(sample, workingday == 1)$cnt)
mean(subset(sample, workingday == 0)$cnt)
mean(subset(sample, workingday == 1)$cnt)
fivenum(subset(data2011, workingday == 0)$cnt)
fivenum(subset(data2011, workingday == 1)$cnt)
mean(subset(data2011, workingday == 0)$cnt)
mean(subset(data2011, workingday == 1)$cnt)
```

Bivariate Analysis:

Code:

DAVID - The relationship between the number of registered and casual users bike rentals and correlated with the temperature.

```
#REGRESSION ANALYSIS

##Scatterplot - Looks non-linear
users_temp <- plot(day_sample$temp,day_sample$cnt)
```

```
abline(regression_analysis, col="blue")

##1. Boxplots - no outliers
cnt_boxplot <- boxplot(day_sample$cnt, ylab = "Bikes Rented")
temp_boxplot <- boxplot(day_sample$temp, ylab = "Washington D.C. Temperature")

##2. Correlation -
cor(day_sample$temp,day_sample$cnt)

## 3. Regression Analysis
regression_analysis <- lm(day_sample$temp ~ day_sample$cnt)
summary(regression_analysis)
abline(regression_analysis, col='red')

## Regression Equation: y=-0.7329243 + 0.0045553x

#CHECKIN ASSUMPTIONS FOR LINEAR REGRESSION

##Normality (Looks normal)
hist(regression_analysis$residuals, main = "Histogram of Residuals",xlab = "Residuals")
boxplot(regression_analysis$residuals, main = "Boxplot of Residuals",ylab = "Residuals")

##Linearity & Homoscdacity  (No curves, but shows a fan pattern)
plot(day_sample$temp, regression_analysis$residuals)
abline(0,0)

Edward - The relationship between the actual temperature and the feeling temperature.
#atemp~temp
temp.atemp<-lm(atemp~temp, sample)
summary(temp.atemp)
plot(sample$temp,sample$atemp, main = "Scatterplot of actual temperature and feeling temperature"
  , ylab = "Feeling temperature", xlab = "Actual temperature")
abline(temp.atemp, col = "red")
cor.test(sample$temp,sample$atemp)

plot(sample$temp,temp.atemp$residuals, main = "Scatterplot of actual temperature and residual"
     , ylab = "Residual", xlab = "Actual temperature")
abline(h=0)

hist(temp.atemp$residuals,xlab = "Residual", main = "Histogram of residual",freq = FALSE)
curve(dnorm,add = TRUE,col='red')

Edward - Chi-Square test between the windspeed and the cold feeling
min(sample$windspeed)
max(sample$windspeed)
max(sample$windspeed)-min(sample$windspeed)

q<-numeric()
for (i in (1:nrow(sample)))
{
```

```
  if (sample$windspeed[i] >=1.5 && sample$windspeed[i] < 10.5)
  {
    q<-rbind(q,1)
  } else
  if (sample$windspeed[i] >=10.5 && sample$windspeed[i] < 19.5)
  {
    q<-rbind(q,2)
  }  else
  {
    q<-rbind(q,3)
  }
}
sample<-cbind(sample,q)
chisq.test(table(sample$feelcool,sample$q))
```

The Relationship Between the Number of Rented Bikes And Wind Speed

```
##Boxplot

boxplot(totalbikeswindspeed$numberofbikes, ylab = "Bikes Rented", main="Number of Bikes Rented",

        sub=paste("Outlier rows: ", boxplot.stats(totalbikeswindspeed$numberofbikes)$out))

boxplot(totalbikeswindspeed$windspeed, ylab = "Windspeed", main="Windspeed",

        sub=paste("Outlier rows: ", boxplot.stats(totalbikeswindspeed$windspeed)$out))

 ##Correlation

cor(totalbikeswindspeed$windspeed,totalbikeswindspeed$numberofbikes)

 ##Regression_analysiz

regression_analysis <- lm(totalbikeswindspeed$numberofbikes ~ totalbikeswindspeed$windspeed)

summary(regression_analysis)

abline(regression_analysis, col='red')

 ##Histogram of residuals

hist(regression_analysis$residuals, main = "Histogram of Residuals",xlab = "Residuals")

## Add a normal curve

hist(regression_analysis$residuals, main = "Histogram of Residuals",xlab = "Residuals")

x <- regression_analysis$residuals
```

```
h<-hist(x, breaks=10, col="red", xlab="Residuals",

        main="Histogram of Residuals")

xfit<-seq(min(x),max(x),length=40)

yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))

yfit <- yfit*diff(h$mids[1:2])*length(x)

lines(xfit, yfit, col="blue", lwd =2)

 ##Boxplot of residuals

boxplot(regression_analysis$residuals, main = "Boxplot of Residuals",ylab = "Residuals")

 ##Scatterplot of residuals

plot(totalbikeswindspeed$windspeed, regression_analysis$residuals, ylab = "Residuals", xlab = "Windspeed",main="Scatterplot of Residuals")

abline(0,0)

 ##Regression

linearMod <- lm(numberofbikes ~ windspeed, data= totalbikeswindspeed)

summary(linearMod)

modelCoeffs <- modelSummary$coefficients

  # scatterplot between wind speed and total number of bikes rented

 plot(sample$windspeed,sample$cnt, main = "Scatterplot of total bike users and windspeed",

     ylab = "Total bike user", xlab = "Windspeed")

abline(cnt.windspeed,col = 'red')

#Histogram of number of bikes

hist(totalbikeswindspeed$numberofbikes, breaks=12, col="red", main = "Histogram of Number of Bikes Rented")

# Add a Normal Curve

x <- totalbikeswindspeed$numberofbikes

h<-hist(x, breaks=12, col="red", xlab="Number of bikes",
```

```
        main="Histogram Number of Bikes Rented")

xfit<-seq(min(x),max(x),length=160)

yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))

yfit <- yfit*diff(h$mids[1:2])*length(x)

lines(xfit, yfit, col="blue", lwd =2)

 ##Histogram of windspeed

hist(totalbikeswindspeed$windspeed, breaks=12, col="red", main = "Histogram of Windspeed")

# Add a Normal Curve

x <- totalbikeswindspeed$windspeed

h<-hist(x, breaks=12, col="red", xlab="windspeed",

        main="Histogram of Windspeed")

xfit<-seq(min(x),max(x),length=40)

yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))

yfit <- yfit*diff(h$mids[1:2])*length(x)

lines(xfit, yfit, col="blue", lwd =2)
```

Variables used

Variable Name	Variable Description	Variable Class	Possible Values of Variable
day	Number of month day	Categorical	1, 2, …, 30, 31
month	Month in the year 2011 & 2012	Categorical	1, 2, …, 11, 12
year	Year of the date of recording this observation	Categorical	2011, 2012
season	Spring, Fall, Summer, Winter	Categorical	1, 2, 3, 4
holiday	The recording day is a public holiday or not	Categorical	0, 1
weekday	Day of the week	Categorical	0, 1, 2, 3, 4, 5, 6
working day	The recording day is a working day or not (Monday to Friday)	Categorical	0, 1
weathersit	The weather situation of the recording day	Categorical	1, 2, 3, 4
temp	Actual temperature	Quantitative	[-10, -9, …, 37, 38, 39]
atemp	"Feels-like" Temperature	Quantitative	[-10, -9, …, 37, 38, 39]
feelcool	The "Feels-like" temperature is lower than the actual temperature or not	Categorical	0, 1
hum	The humidity of the recorded day	Quantitative	[0, 0.1, …, 99.9, 100]
windspeed	The wind speed of the recording day	Quantitative	[0, 0.1, …, 99.9, 100]
casual	The number of bike rented by walking users (have no continuous membership)	Quantitative	[0,∞)
registered	The number of bike rented by registered members	Quantitative	[0,∞)
cnt	The total number of rented bikes	Quantitative	[0,∞)