Bike Sharing Data Analysis and Prediction

Abstract

Introduction

Background

Objective

Data understanding

Preparation

To perform the analysis, certain R libraries were used. The code below was used to load and initialize the libraries.

```
# Will use the following packages for this analysis:

library(ggplot2)
library(dplyr)
library(GGally)
library(magrittr)
library(ggpubr)
library(reshape2)
library(weathermetrics)
library(hexbin)
library(Hmisc)
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(corrgram)
```

```
## Registered S3 method overwritten by 'seriation':
##   method         from
##   reorder.hclust gclus
```

```
##
## Attaching package: 'corrgram'
```

```
## The following object is masked from 'package:lattice':
##
##     panel.fill
```

```
library(caTools)
```

Reading the Bikedatasets "hour.csv". There are two csv file in the Bike Datasets: hour.csv, day.csv. The different between these two data is that hour.csv has hr attibution. We will only use hour.csv for this study.

The data source is from UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/machine-learning-databases/00275/

```
bike <- read.csv("hour.csv", header = TRUE)
```

Preview of the data

```
head(bike)
```

```
##   instant     dteday season yr mnth hr holiday weekday workingday weathersit
## 1       1 2011-01-01      1  0    1  0       0       6          0          1
## 2       2 2011-01-01      1  0    1  1       0       6          0          1
## 3       3 2011-01-01      1  0    1  2       0       6          0          1
## 4       4 2011-01-01      1  0    1  3       0       6          0          1
## 5       5 2011-01-01      1  0    1  4       0       6          0          1
## 6       6 2011-01-01      1  0    1  5       0       6          0          2
##   temp  atemp  hum windspeed casual registered cnt
## 1 0.24 0.2879 0.81    0.0000      3         13  16
## 2 0.22 0.2727 0.80    0.0000      8         32  40
## 3 0.22 0.2727 0.80    0.0000      5         27  32
## 4 0.24 0.2879 0.75    0.0000      3         10  13
## 5 0.24 0.2879 0.75    0.0000      0          1   1
## 6 0.24 0.2576 0.75    0.0896      0          1   1
```

Data attibutes summary

- instant: record index
- dteday : date
- season : season (1:winter, 2:spring, 3:summer, 4:fall)
- yr : year (0: 2011, 1:2012)
- mnth : month (1 to 12)
- hr : hour (0 to 23)
- holiday : weather day is holiday or not (extracted from [Web Link])
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit :
- 1: Clear, Few clouds, Partly cloudy, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)
- atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users
- registered: count of registered users

- cnt: count of total rental bikes including both casual and registered

```
str(bike)
```

```
## 'data.frame':    17379 obs. of  17 variables:
##  $ instant   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ dteday    : Factor w/ 731 levels "2011-01-01","2011-01-02",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ season    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ yr        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mnth      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hr        : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ weekday   : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ workingday: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ weathersit: int  1 1 1 1 1 2 1 1 1 1 ...
##  $ temp      : num  0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24 0.32 ...
##  $ atemp     : num  0.288 0.273 0.273 0.288 0.288 ...
##  $ hum       : num  0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75 0.76 ...
##  $ windspeed : num  0 0 0 0 0 0.0896 0 0 0 0 ...
##  $ casual    : int  3 8 5 3 0 0 2 1 1 8 ...
##  $ registered: int  13 32 27 10 1 1 0 2 7 6 ...
##  $ cnt       : int  16 40 32 13 1 1 2 3 8 14 ...
```

Check for missing values and Data preparation

```
sum(is.na(bike))
```

```
## [1] 0
```

```
bike$instant <- NULL
```

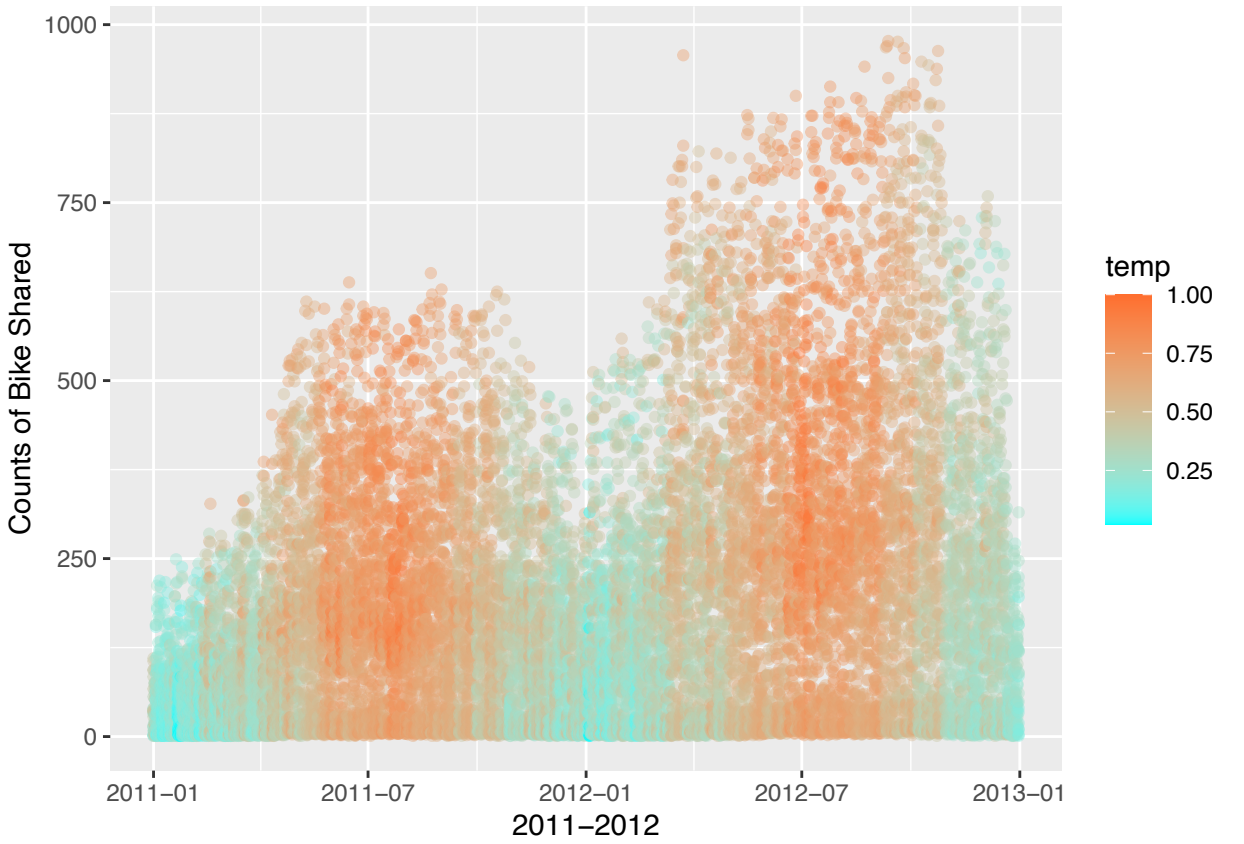
Change Date to datetime format.

```
bike$dteday <- as.POSIXct(bike$dteday)
```

Data Analysis and Visualization

The number of bike shared in the year of 2011 and 2012.

```
ggplot(bike,aes(dteday,cnt)) + geom_point(aes(color=temp),alpha=0.4)+ scale_color_continuous(low="85d2d4
```



There is a curve in each year: Summer and Autums when the temperature is high, there are more bikes were tent out.

Explore the data base on one time cycle(one year)

```
which(bike$dteday== "2012-01-01")
```

```
##  [1] 8646 8647 8648 8649 8650 8651 8652 8653 8654 8655 8656 8657 8658 8659 8660
## [16] 8661 8662 8663 8664 8665 8666 8667 8668 8669
```

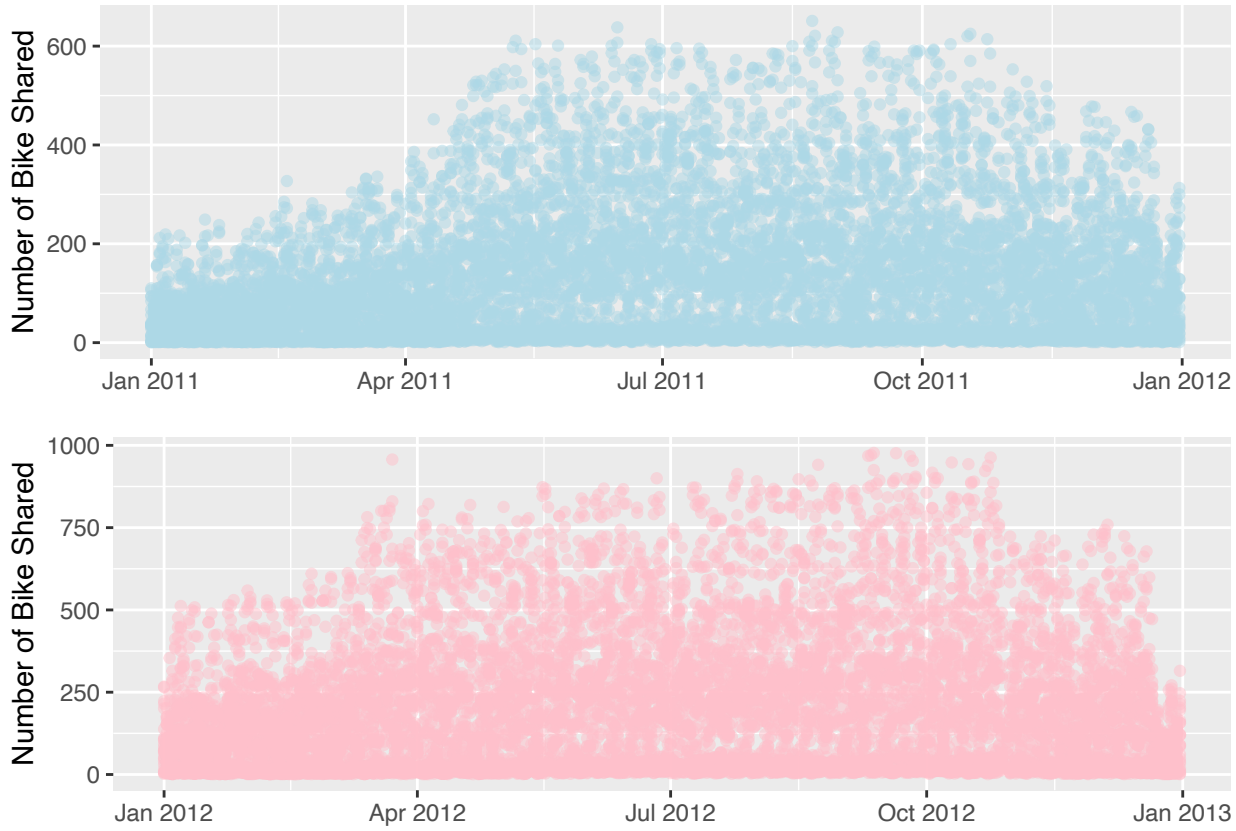
```
bike2011 <- bike[1:8645,]
bike2012 <- bike[8646:17379,]
```

```
head(bike2011)
```

```
##       dteday season yr mnth hr holiday weekday workingday weathersit temp
## 1 2011-01-01      1  0    1  0       0       6          0          1 0.24
## 2 2011-01-01      1  0    1  1       0       6          0          1 0.22
## 3 2011-01-01      1  0    1  2       0       6          0          1 0.22
## 4 2011-01-01      1  0    1  3       0       6          0          1 0.24
## 5 2011-01-01      1  0    1  4       0       6          0          1 0.24
## 6 2011-01-01      1  0    1  5       0       6          0          2 0.24
##    atemp  hum windspeed casual registered cnt
## 1 0.2879 0.81    0.0000      3         13  16
## 2 0.2727 0.80    0.0000      8         32  40
## 3 0.2727 0.80    0.0000      5         27  32
## 4 0.2879 0.75    0.0000      3         10  13
## 5 0.2879 0.75    0.0000      0          1   1
```

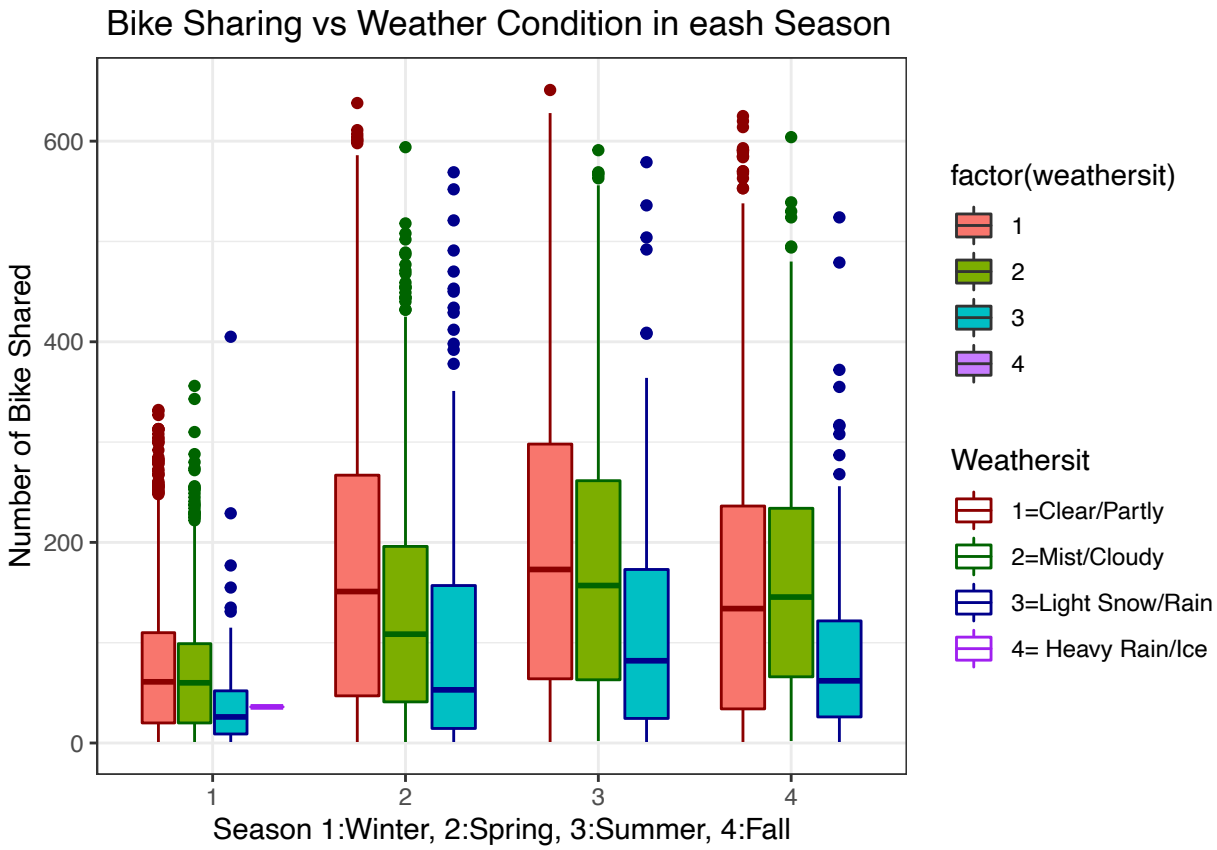
```
## 6 0.2576 0.75    0.0896      0          1   1
```

```
b2011<- ggplot(bike2011, aes(x= dteday, y = cnt)) + geom_point(color="light blue",alpha= 0.5) + xlab(el

b2012<- ggplot(bike2012, aes(x=dteday, y = cnt)) + geom_point(color="pink",alpha= 0.5) + xlab(element_bl

ggarrange(b2011,b2012 ,
ncol = 1, nrow = 2)
```



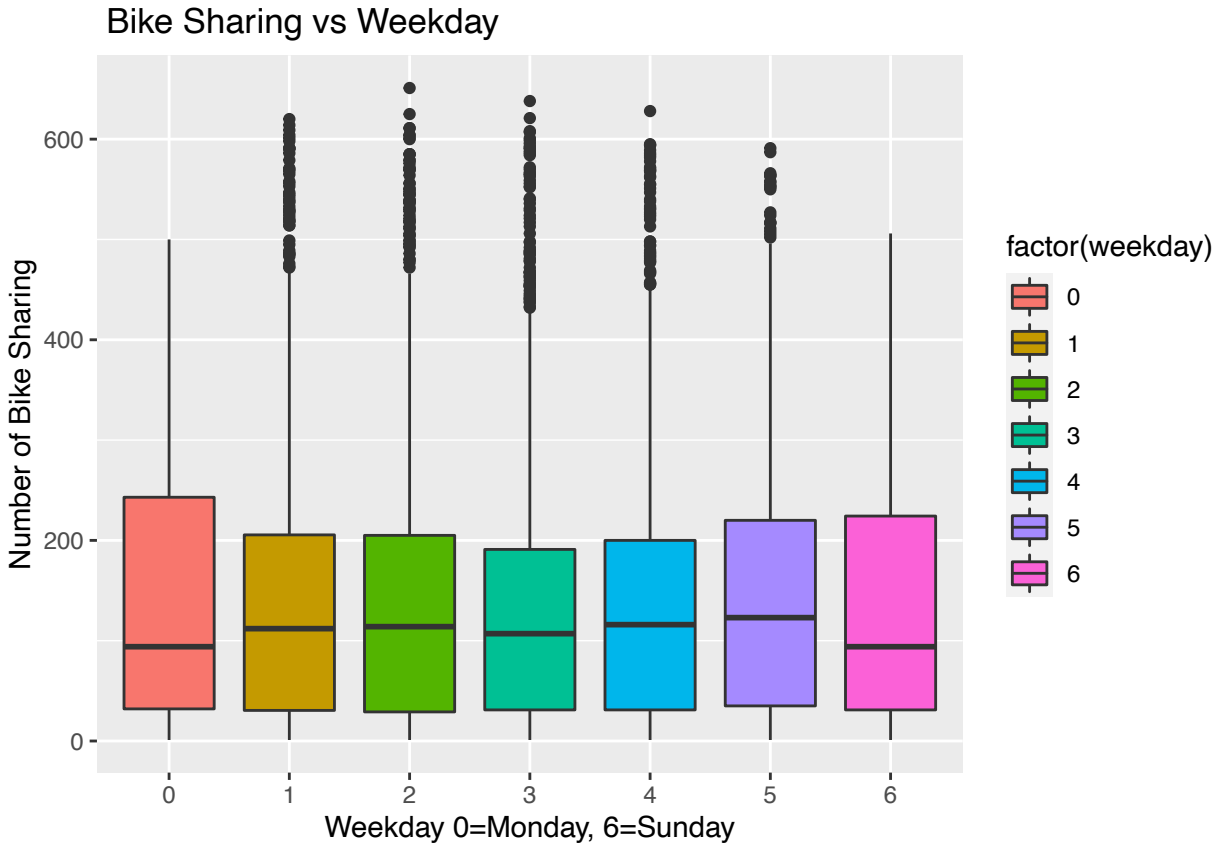
We see there is linear increase of the number of bike shared from Jan to middle Oct. Then there is a decrease from middle Oct to end of Dec in both 2011 and 2012.

```
ggplot(bike2011, aes(x=factor(season), y = cnt, color= factor(weathersit))) + geom_boxplot(aes(fill=fac
labs(title =" Bike Sharing vs Weather Condition in eash Season",
     x = "Season 1:Winter, 2:Spring, 3:Summer, 4:Fall",
     y = "Number of Bike Shared") +
     scale_color_manual(labels = c("1=Clear/Partly", "2=Mist/Cloudy",
                                   "3=Light Snow/Rain", "4= Heavy Rain/Ice"),
                        values = c( "dark red", "dark green","dark blue",
                                    "purple"))+
                        theme_bw()+
  guides(color=guide_legend("Weathersit"))
```



We can see Winter time had the lowest number of bike rent, Spring, Summer had almost the same counts while there is a decrese in Fall. When the weather situation is "1", meaning clear/partly cloud, there were more number of bike shared. Weathersit 4, heavy rain/ice/storm had the least of numbers of shared.

```
ggplot(bike2011, aes(x=factor(weekday), y = cnt)) + geom_boxplot(aes(fill = factor(weekday))) +
xlab("Weekday 0=Monday, 6=Sunday") + ylab("Number of Bike Sharing") + ggtitle(" Bike Sharing vs Weekday
```



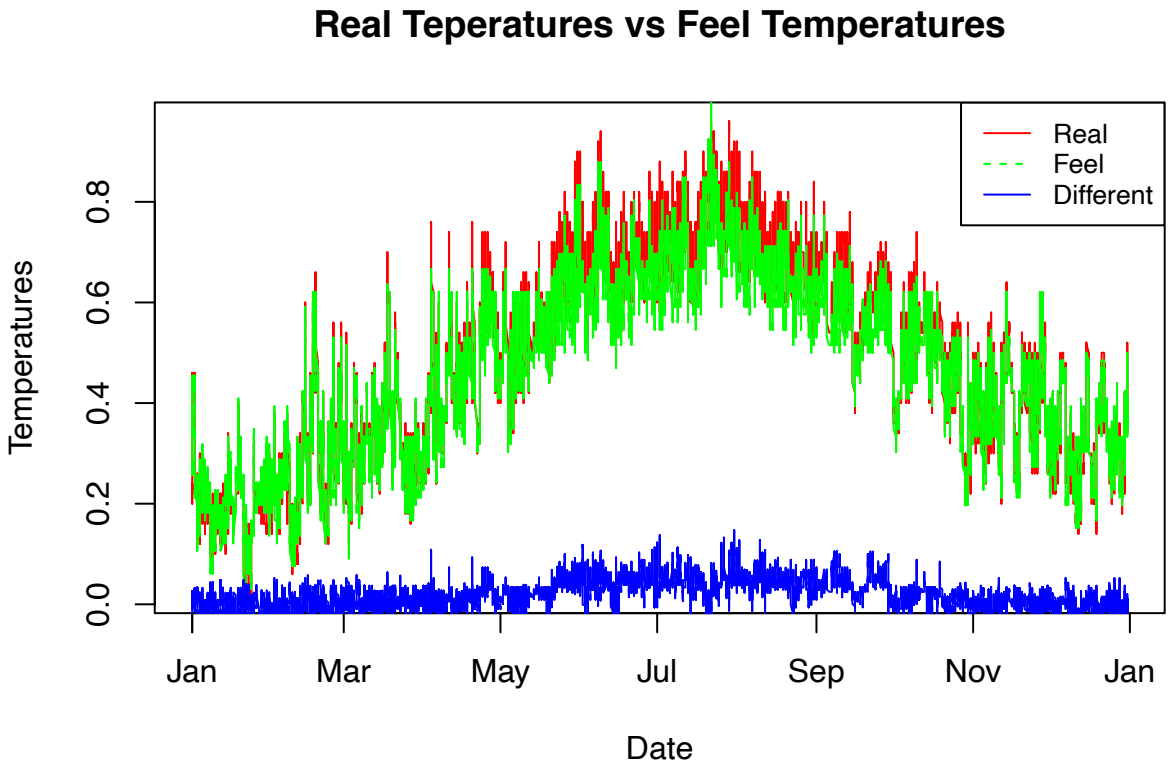
Monday and Sunday had the most number of bike shared. There is a slike up linear curve from Tuesday to Sunday.

Comparing the different between real temperature and feel temperature

```
bike2011$dtemp <- (bike2011$temp - bike2011$atemp)
mean(bike2011$dtemp)
```

```
## [1] 0.02006837
```

```
plot(bike2011$dteday, bike2011$temp,type="l",col="red", xlab="Date", ylab="Temperatures",
     main= "Real Teperatures vs Feel Temperatures" )
lines(bike2011$dteday, bike2011$atemp,type="l",col="green")
lines(bike2011$dteday, bike2011$dtemp,type="l",col="blue")

legend("topright",legend=c("Real","Feel","Different"),col=c("red","green", "blue"),lty = 1:2,
                                                                    cex=0.8)
```

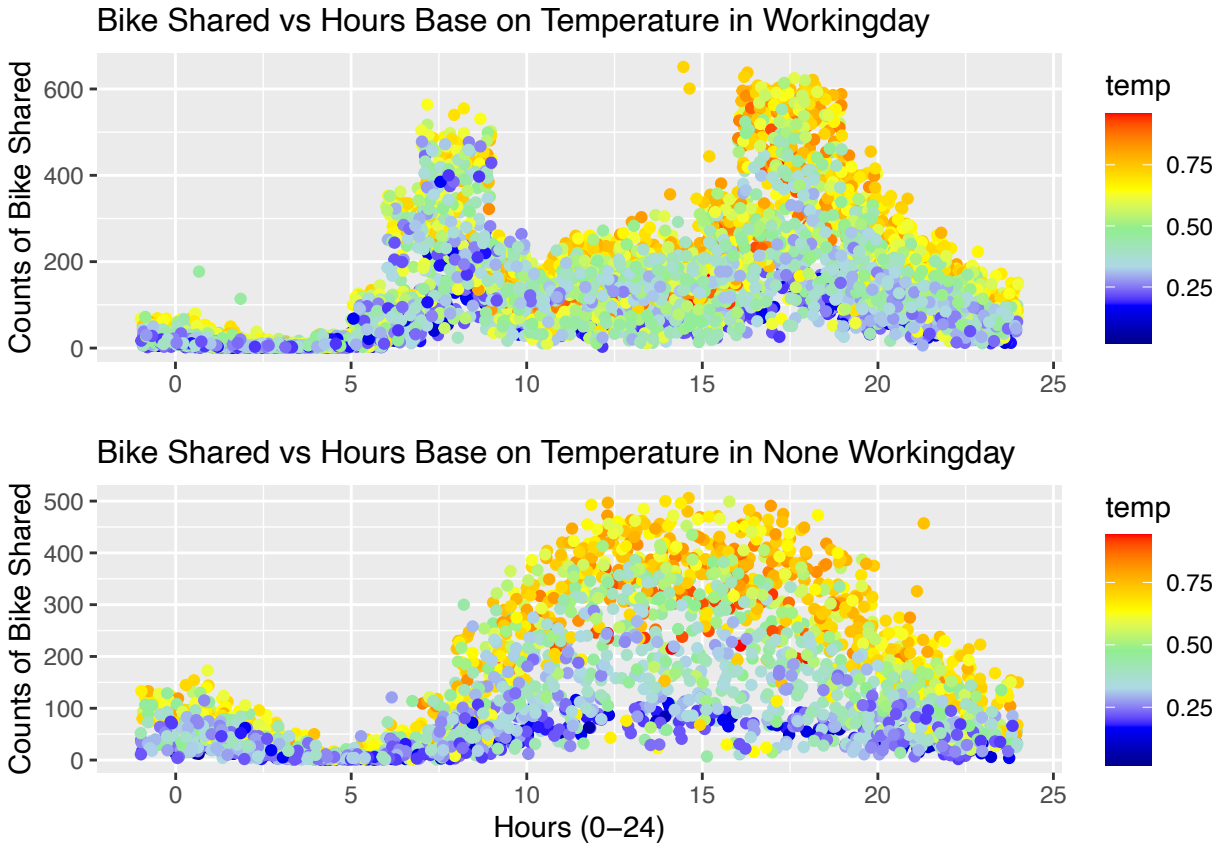


###

There is no much different between real temperature and feel temperature. Will only use the temp to do models. Remove the new created column "dtemp"

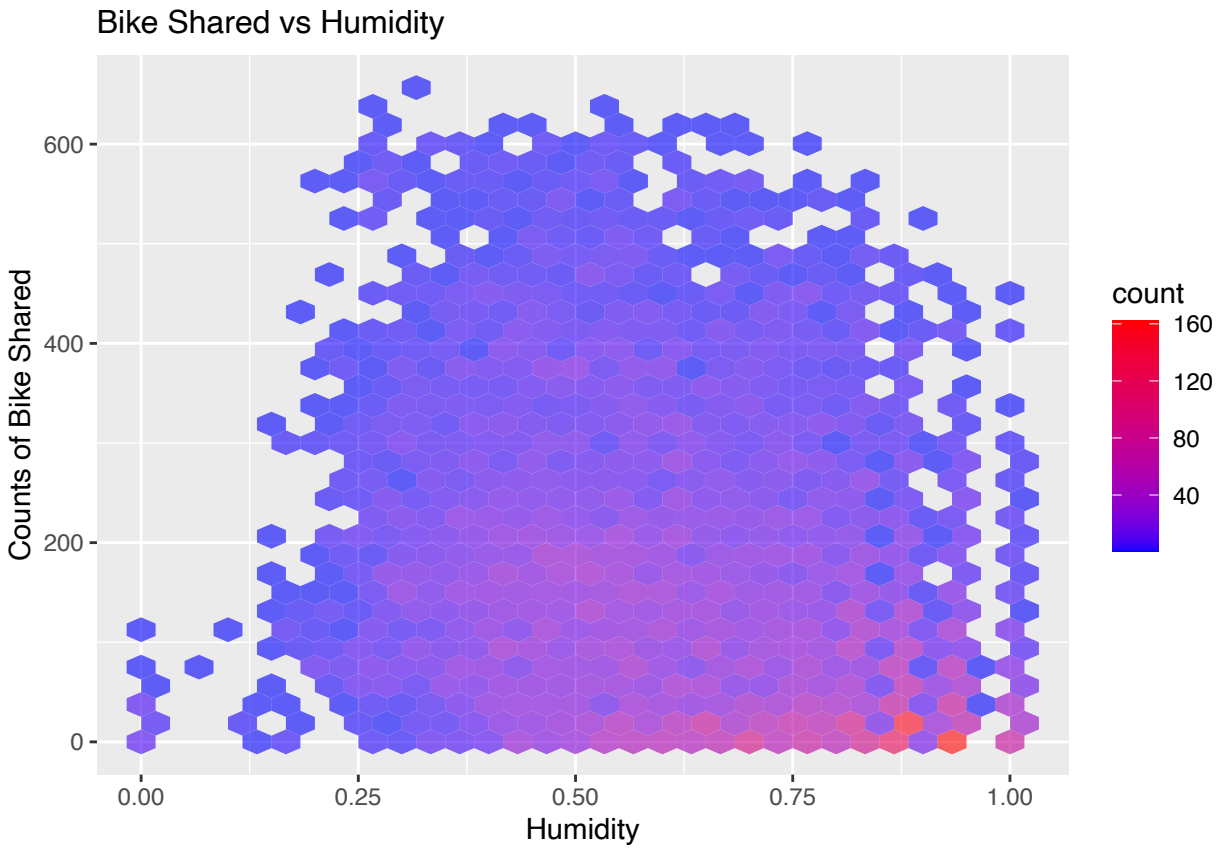
```
bike2011$dtemp <- NULL
```

```
weekd <- ggplot(filter(bike2011, workingday== 1), aes(hr,cnt)) +
  geom_point(position = position_jitter(w=1,h=0),aes(color=temp)) +
  scale_colour_gradientn(colours = c("dark blue","blue", "light blue", "light green",
                                     "yellow","orange","red")) +
  xlab(element_blank())+ ylab("Counts of Bike Shared") +
  ggtitle("Bike Shared vs Hours Base on Temperature in Workingday") +
  theme(plot.title = element_text(size = 12))

wend<- ggplot(filter(bike2011, workingday== 0), aes(hr,cnt)) +
  geom_point(position = position_jitter(w=1,h=0),aes(color=temp)) +
  scale_colour_gradientn(colours = c("dark blue","blue", "light blue", "light green",
                                     "yellow","orange", "red")) +
  xlab("Hours (0-24)")+ ylab("Counts of Bike Shared") +
  ggtitle("Bike Shared vs Hours Base on Temperature in None Workingday") +
  theme(plot.title = element_text(size = 12))

ggarrange(weekd,wend,ncol = 1, nrow = 2)
```



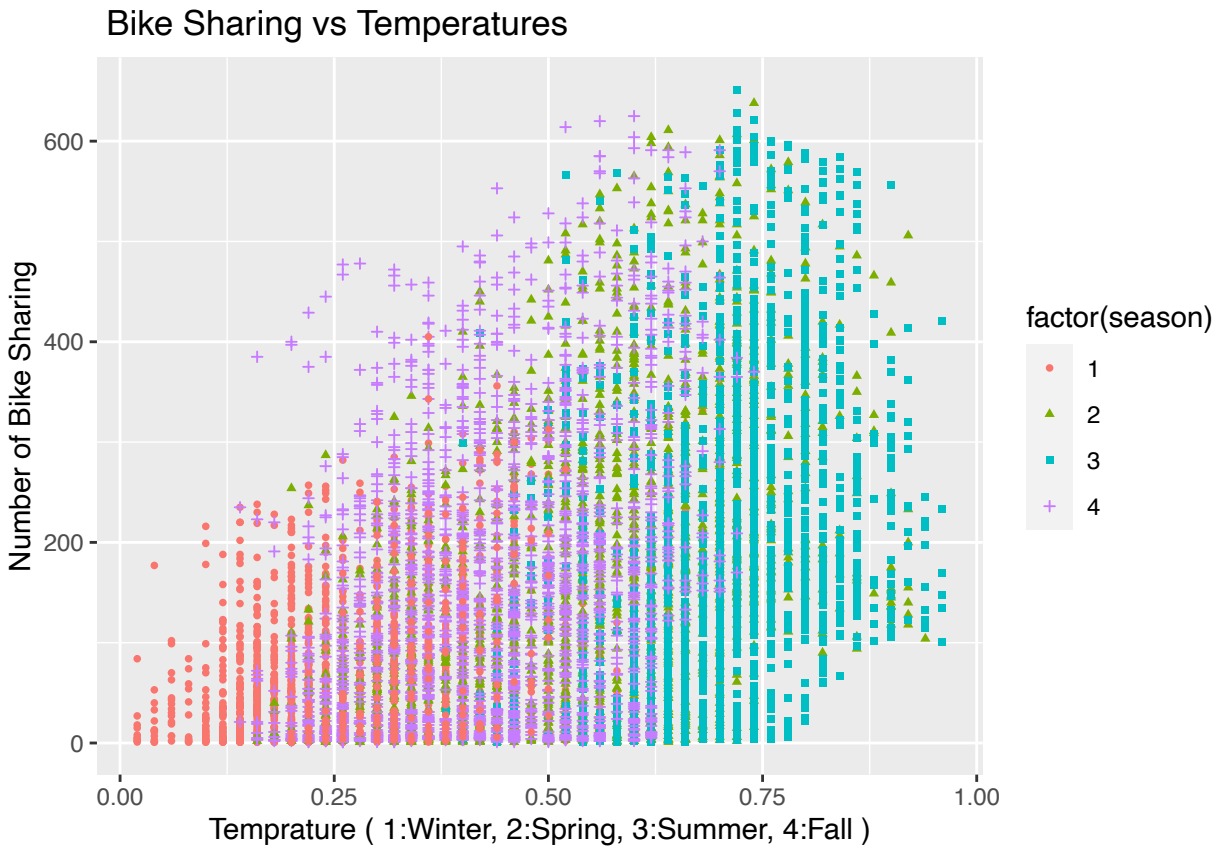
In workingday, there were two peak time that the number of bike shared had the highest counts. One was the time during 7am to 9am; another time was the period of 16pm to 19pm. Seems like people rent bike going to work or back home. ### In the none workingdays, The highest numbers of bike rent was in the period of 10am to 19pm.

```
ggplot(bike2011, aes(x=hum, y = cnt)) +
geom_hex(alpha=0.6) + scale_fill_gradient(high = "red", low = "blue") +
  xlab("Humidity")+ ylab("Counts of Bike Shared") +
ggtitle("Bike Shared vs Humidity") +
theme(plot.title = element_text(size = 12))
```



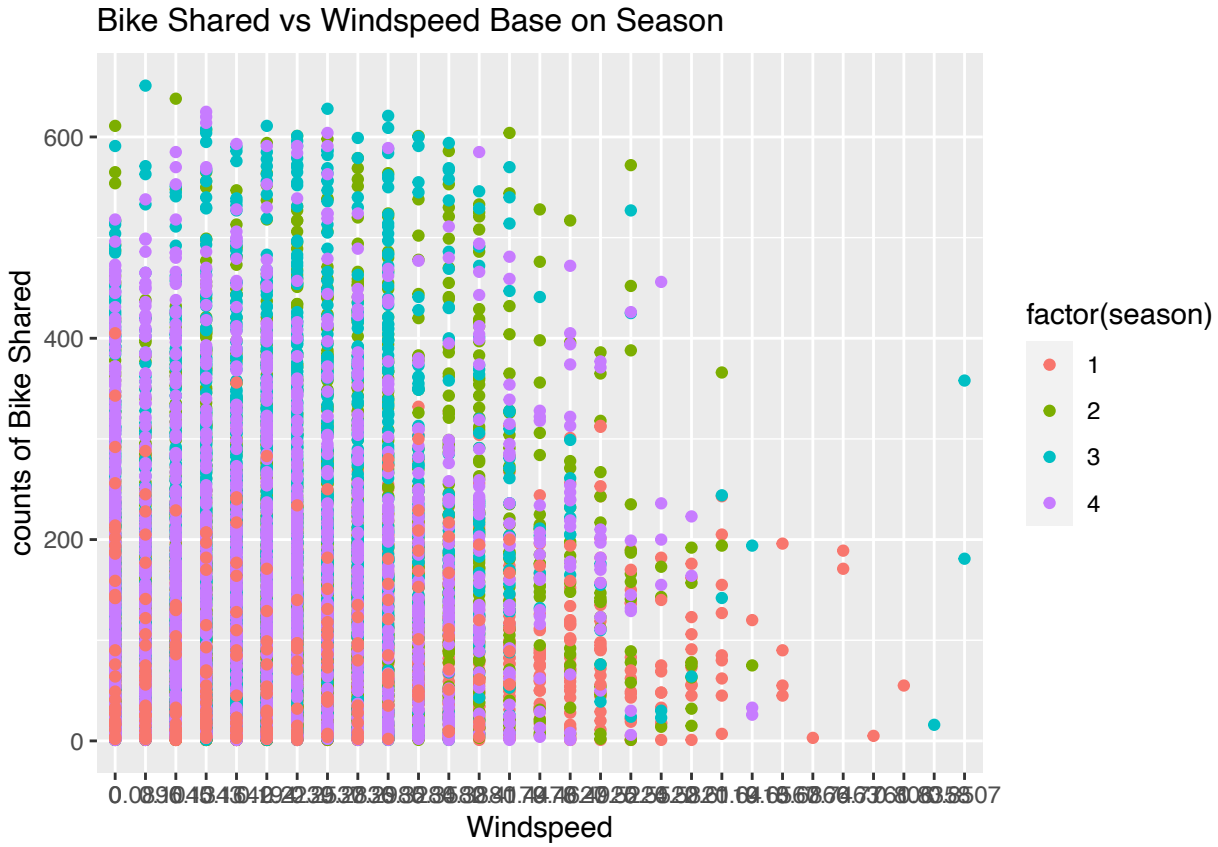
We can see a nagetive corrlation between humidity and Number of bike rented.

```
ggplot(bike2011, aes(x= temp, y = cnt)) + geom_point(aes(shape=factor(season),color=factor(season)), si
  xlab("Temprature ( 1:Winter, 2:Spring, 3:Summer, 4:Fall )")+
     ylab("Number of Bike Sharing") +
ggtitle(" Bike Sharing vs Temperatures")
```



We see a positive correlation between temperature and number of bike rented.

```
ggplot(bike2011, aes(x=factor(windspeed), y= cnt)) +
  geom_point(aes(color=factor(season))) +
  xlab("Windspeed") + ylab("counts of Bike Shared") +
  ggtitle("Bike Shared vs Windspeed Base on Season") +
  theme(plot.title = element_text(size = 12))
```



There is a nagetive correlation between windspeed and number of bike shared.

Models selection:

Linear Regression

The dataset containes both continues inteval numerics (temp, hum, windspeed) and norminal numbers(months, weeks, hr, seasons, holidays, workingdays). The prediction number of bike shared is a count data. We will use several regression models to see which one fit most.

We first will use 2011 bike data to do a Linear Regression as a train dataset, then we will use the 2012 data as the test dataset.

We also will check if use the whole dataset which is including both 2011 and 2012 dateset to see what will happen.

```
colnames(bike)
```

```
##  [1] "dteday"     "season"     "yr"         "mnth"       "hr"
##  [6] "holiday"    "weekday"    "workingday" "weathersit" "temp"
## [11] "atemp"      "hum"        "windspeed"  "casual"     "registered"
## [16] "cnt"
```

remove "dteday","atemp", "casual", "registered". Rename column names for convenience.

```
bike_sub <- select(bike,-"dteday",-"atemp", -"casual", -"registered")
```

```
head(bike_sub)
```

```
##   season yr mnth hr holiday weekday workingday weathersit temp  hum windspeed
## 1      1  0    1  0       0       6          0          1 0.24 0.81    0.0000
## 2      1  0    1  1       0       6          0          1 0.22 0.80    0.0000
## 3      1  0    1  2       0       6          0          1 0.22 0.80    0.0000
## 4      1  0    1  3       0       6          0          1 0.24 0.75    0.0000
## 5      1  0    1  4       0       6          0          1 0.24 0.75    0.0000
## 6      1  0    1  5       0       6          0          2 0.24 0.75    0.0896
##   cnt
## 1  16
## 2  40
## 3  32
## 4  13
## 5   1
## 6   1
```

Seperate 2011 and 2012 data. 2011 will be use as train dataset.

```
bike2011_sub <- bike_sub[1:8645,]
bike2012_sub <- bike_sub[8646:17379,]
```

```
head(bike2011_sub)
```

```
##   season yr mnth hr holiday weekday workingday weathersit temp  hum windspeed
## 1      1  0    1  0       0       6          0          1 0.24 0.81    0.0000
## 2      1  0    1  1       0       6          0          1 0.22 0.80    0.0000
## 3      1  0    1  2       0       6          0          1 0.22 0.80    0.0000
## 4      1  0    1  3       0       6          0          1 0.24 0.75    0.0000
## 5      1  0    1  4       0       6          0          1 0.24 0.75    0.0000
## 6      1  0    1  5       0       6          0          2 0.24 0.75    0.0896
##   cnt
## 1  16
## 2  40
## 3  32
## 4  13
## 5   1
## 6   1
```

```
cor(bike2011_sub)
```

```
## Warning in cor(bike2011_sub): the standard deviation is zero
```

```
##                  season yr         mnth            hr       holiday
## season      1.000000000 NA  0.829054032 -0.0121785916 -0.0011158991
## yr                   NA  1           NA            NA            NA
## mnth        0.829054032 NA  1.000000000 -0.0118141095  0.0298058604
## hr         -0.012178592 NA -0.011814109  1.0000000000 -0.0009291255
## holiday    -0.001115899 NA  0.029805860 -0.0009291255  1.0000000000
## weekday    -0.013639174 NA  0.012023684 -0.0056183913 -0.0763242171
## workingday  0.013762919 NA  0.003573409  0.0037919271 -0.2479336547
## weathersit -0.015351416 NA -0.020661994 -0.0165211031  0.0080013044
## temp        0.343534543 NA  0.260442310  0.1205477924 -0.0200586618
## hum         0.191793265 NA  0.188060971 -0.2511876825 -0.0228952488
```

```
## windspeed  -0.154735348 NA -0.155644730  0.1251531846  0.0005018825
## cnt         0.221719405 NA  0.179273000  0.4074862802 -0.0229118108
##                 weekday   workingday   weathersit         temp         hum
## season     -0.013639174  0.013762919 -0.015351416  0.343534543  0.19179327
## yr                   NA           NA           NA           NA          NA
## mnth        0.012023684  0.003573409 -0.020661994  0.260442310  0.18806097
## hr         -0.005618391  0.003791927 -0.016521103  0.120547792 -0.25118768
## holiday    -0.076324217 -0.247933655  0.008001304 -0.020058662 -0.02289525
## weekday     1.000000000  0.018572083  0.027766050 -0.038968743 -0.05130744
## workingday  0.018572083  1.000000000  0.068628489  0.053405209  0.02530464
## weathersit  0.027766050  0.068628489  1.000000000 -0.092036416  0.40763350
## temp       -0.038968743  0.053405209 -0.092036416  1.000000000 -0.03952480
## hum        -0.051307436  0.025304636  0.407633503 -0.039524797  1.00000000
## windspeed   0.039585548  0.010509122  0.048413636 -0.005989444 -0.26319623
## cnt        -0.004320854  0.011704243 -0.143288261  0.451232536 -0.28861522
##                windspeed          cnt
## season     -0.1547353480  0.221719405
## yr                   NA           NA
## mnth       -0.1556447303  0.179273000
## hr          0.1251531846  0.407486280
## holiday     0.0005018825 -0.022911811
## weekday     0.0395855477 -0.004320854
## workingday  0.0105091219  0.011704243
## weathersit  0.0484136357 -0.143288261
## temp       -0.0059894439  0.451232536
## hum        -0.2631962300 -0.288615224
## windspeed   1.0000000000  0.085355569
## cnt         0.0853555688  1.000000000
```

"yr has zeor standard deviation. Drop"yr

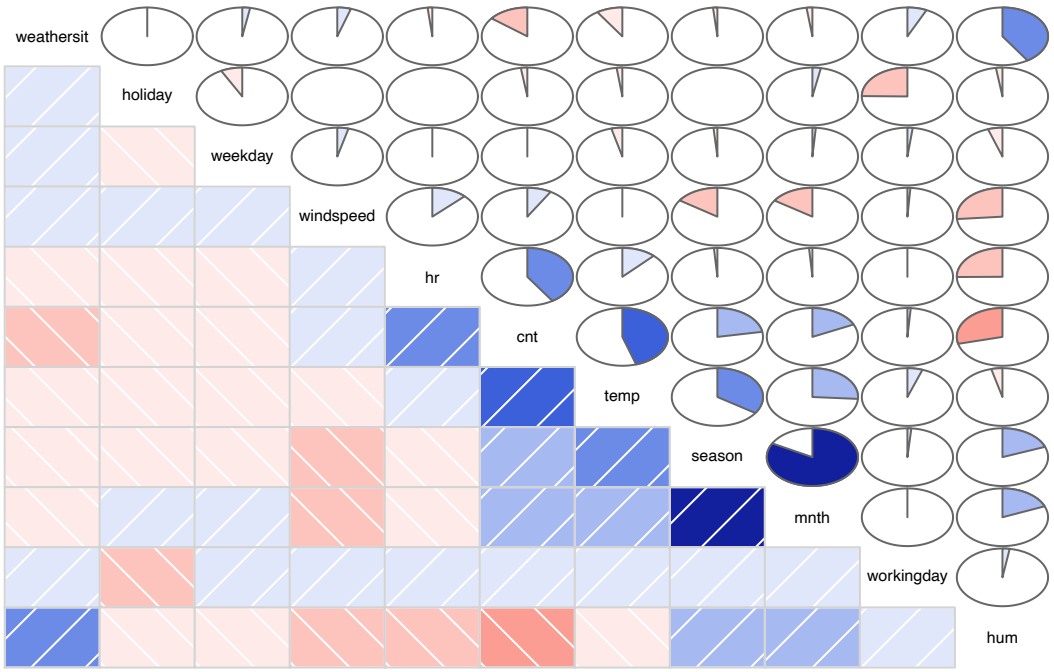
```
bike2011_sub$yr<-NULL
```

```
cor(bike2011_sub)
```

```
##                  season         mnth            hr       holiday      weekday
## season      1.000000000  0.829054032 -0.0121785916 -0.0011158991 -0.013639174
## mnth        0.829054032  1.000000000 -0.0118141095  0.0298058604  0.012023684
## hr         -0.012178592 -0.011814109  1.0000000000 -0.0009291255 -0.005618391
## holiday    -0.001115899  0.029805860 -0.0009291255  1.0000000000 -0.076324217
## weekday    -0.013639174  0.012023684 -0.0056183913 -0.0763242171  1.000000000
## workingday  0.013762919  0.003573409  0.0037919271 -0.2479336547  0.018572083
## weathersit -0.015351416 -0.020661994 -0.0165211031  0.0080013044  0.027766050
## temp        0.343534543  0.260442310  0.1205477924 -0.0200586618 -0.038968743
## hum         0.191793265  0.188060971 -0.2511876825 -0.0228952488 -0.051307436
## windspeed  -0.154735348 -0.155644730  0.1251531846  0.0005018825  0.039585548
## cnt         0.221719405  0.179273000  0.4074862802 -0.0229118108 -0.004320854
##                workingday   weathersit         temp         hum     windspeed
## season      0.013762919 -0.015351416  0.343534543  0.19179327 -0.1547353480
## mnth        0.003573409 -0.020661994  0.260442310  0.18806097 -0.1556447303
## hr          0.003791927 -0.016521103  0.120547792 -0.25118768  0.1251531846
## holiday    -0.247933655  0.008001304 -0.020058662 -0.02289525  0.0005018825
## weekday     0.018572083  0.027766050 -0.038968743 -0.05130744  0.0395855477
```

```
## workingday  1.000000000  0.068628489  0.053405209  0.02530464  0.0105091219
## weathersit  0.068628489  1.000000000 -0.092036416  0.40763350  0.0484136357
## temp        0.053405209 -0.092036416  1.000000000 -0.03952480 -0.0059894439
## hum         0.025304636  0.407633503 -0.039524797  1.00000000 -0.2631962300
## windspeed   0.010509122  0.048413636 -0.005989444 -0.26319623  1.0000000000
## cnt         0.011704243 -0.143288261  0.451232536 -0.28861522  0.0853555688
##                     cnt
## season      0.221719405
## mnth        0.179273000
## hr          0.407486280
## holiday    -0.022911811
## weekday    -0.004320854
## workingday  0.011704243
## weathersit -0.143288261
## temp        0.451232536
## hum        -0.288615224
## windspeed   0.085355569
## cnt         1.000000000
```

```
corrgram(bike2011_sub,order=TRUE, lower.panel=panel.shade, upper.panel=panel.pie,
         text.panel=panel.txt)
```




```
cor_result = rcorr(as.matrix(bike2011_sub))
cor_result
```

```
##            season  mnth    hr holiday weekday workingday weathersit  temp   hum
## season       1.00  0.83 -0.01    0.00   -0.01       0.01      -0.02  0.34  0.19
## mnth         0.83  1.00 -0.01    0.03    0.01       0.00      -0.02  0.26  0.19
## hr          -0.01 -0.01  1.00    0.00   -0.01       0.00      -0.02  0.12 -0.25
## holiday      0.00  0.03  0.00    1.00   -0.08      -0.25       0.01 -0.02 -0.02
## weekday     -0.01  0.01 -0.01   -0.08    1.00       0.02       0.03 -0.04 -0.05
## workingday   0.01  0.00  0.00   -0.25    0.02       1.00       0.07  0.05  0.03
## weathersit  -0.02 -0.02 -0.02    0.01    0.03       0.07       1.00 -0.09  0.41
```

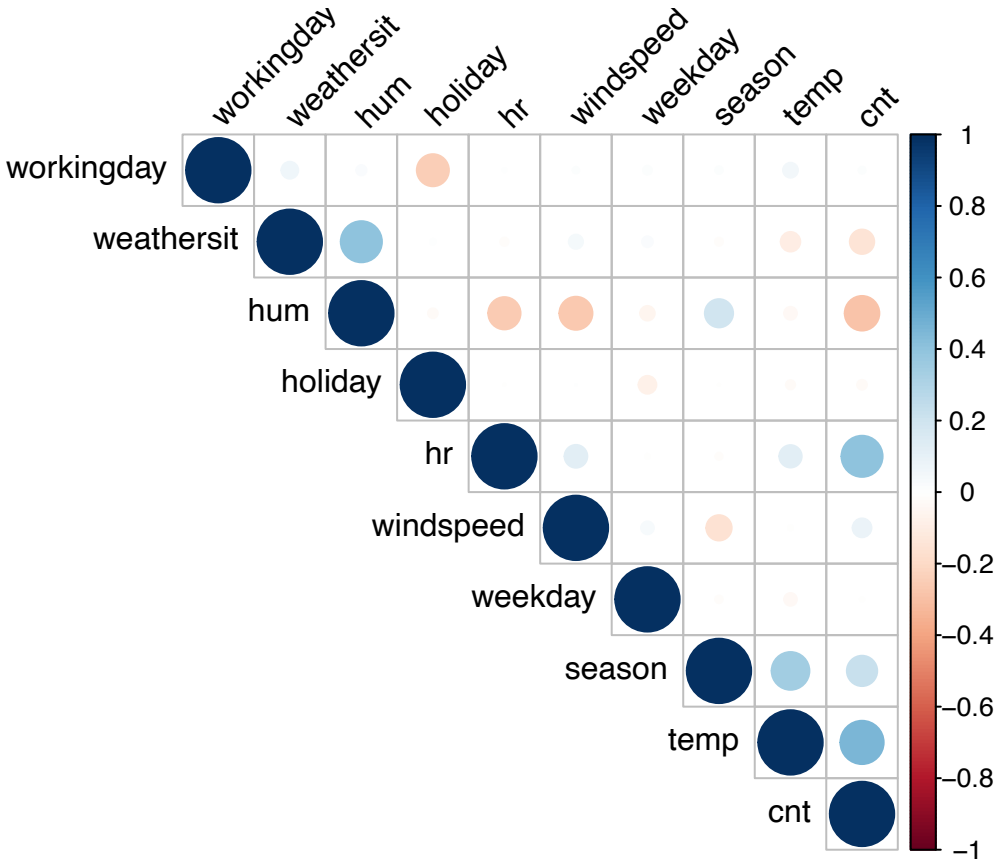
```
## temp         0.34  0.26  0.12   -0.02   -0.04       0.05      -0.09  1.00 -0.04
## hum          0.19  0.19 -0.25   -0.02   -0.05       0.03       0.41 -0.04  1.00
## windspeed   -0.15 -0.16  0.13    0.00    0.04       0.01       0.05 -0.01 -0.26
## cnt          0.22  0.18  0.41   -0.02    0.00       0.01      -0.14  0.45 -0.29
##            windspeed   cnt
## season         -0.15  0.22
## mnth           -0.16  0.18
## hr              0.13  0.41
## holiday         0.00 -0.02
## weekday         0.04  0.00
## workingday      0.01  0.01
## weathersit      0.05 -0.14
## temp           -0.01  0.45
## hum            -0.26 -0.29
## windspeed       1.00  0.09
## cnt             0.09  1.00
##
## n= 8645
##
##
## P
##            season mnth   hr     holiday weekday workingday weathersit temp
## season            0.0000 0.2575 0.9174  0.2048  0.2007     0.1535     0.0000
## mnth       0.0000        0.2721 0.0056  0.2636  0.7397     0.0547     0.0000
## hr         0.2575 0.2721        0.9312  0.6014  0.7244     0.1245     0.0000
## holiday    0.9174 0.0056 0.9312         0.0000  0.0000     0.4570     0.0622
## weekday    0.2048 0.2636 0.6014 0.0000          0.0842     0.0098     0.0003
## workingday 0.2007 0.7397 0.7244 0.0000  0.0842             0.0000     0.0000
## weathersit 0.1535 0.0547 0.1245 0.4570  0.0098  0.0000                0.0000
## temp       0.0000 0.0000 0.0000 0.0622  0.0003  0.0000     0.0000
## hum        0.0000 0.0000 0.0000 0.0333  0.0000  0.0186     0.0000     0.0002
## windspeed  0.0000 0.0000 0.0000 0.9628  0.0002  0.3286     0.0000     0.5777
## cnt        0.0000 0.0000 0.0000 0.0331  0.6879  0.2765     0.0000     0.0000
##            hum    windspeed cnt
## season     0.0000 0.0000    0.0000
## mnth       0.0000 0.0000    0.0000
## hr         0.0000 0.0000    0.0000
## holiday    0.0333 0.9628    0.0331
## weekday    0.0000 0.0002    0.6879
## workingday 0.0186 0.3286    0.2765
## weathersit 0.0000 0.0000    0.0000
## temp       0.0002 0.5777    0.0000
## hum               0.0000    0.0000
## windspeed  0.0000           0.0000
## cnt        0.0000 0.0000
```

mnth(months) and sns(season) has 0.83 correlatin value which means these tow variables have noticeable correlation.

```
bike2011_sub$mnth <- NULL
```

```
cor_result = rcorr(as.matrix(bike2011_sub))
```

```r
corrplot(cor_result$r, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
```



Create Linear Regression Model

Step one : using 2011 bike sharing data as train dataset.

```r
# set.seed(99)
# split =  sample.split(bike2012_sub$cnt, SplitRatio = 0.7)
# train = subset(bike2012_sub, split==TRUE)
# test= subset(bike2012_sub, split==FALSE)
```

```r
bike2012_sub$yr <- NULL
bike2012_sub$mnth <- NULL
```

```r
train= bike2011_sub
test= bike2012_sub
```

```r
summary(bike2011.model <- lm(cnt ~ ., data= train))
```

```
##
## Call:
## lm(formula = cnt ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -237.44  -69.50  -20.59   44.42  411.47
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.25392    6.76118   1.517   0.1294
## season        17.86362    1.12443  15.887   <2e-16 ***
## hr             6.01543    0.17128  35.120   <2e-16 ***
## holiday      -18.37062    7.13857  -2.573   0.0101 *
## weekday        0.03138    0.56698   0.055   0.9559
## workingday    -2.76602    2.51804  -1.098   0.2720
## weathersit    -3.24267    1.94963  -1.663   0.0963 .
## temp         238.82713    6.18119  38.638   <2e-16 ***
## hum         -146.25495    6.98129 -20.950   <2e-16 ***
## windspeed     17.19107    9.71713   1.769   0.0769 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105 on 8635 degrees of freedom
## Multiple R-squared:  0.3847, Adjusted R-squared:  0.384
## F-statistic: 599.8 on 9 and 8635 DF,  p-value: < 2.2e-16
```

Adjusted R-Squared is 0.384. Weekday shows none significant in P-valuse < 0.05. Remove weekday to do a new model.

```
summary(bike2011.model1 <- lm(cnt ~ .-weekday, data= train))
```

```
##
## Call:
## lm(formula = cnt ~ . - weekday, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -237.47  -69.48  -20.57   44.40  411.47
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.3625     6.4699   1.602  0.10927
## season        17.8648     1.1242  15.892  < 2e-16 ***
## hr             6.0152     0.1712  35.129  < 2e-16 ***
## holiday      -18.4013     7.1166  -2.586  0.00974 **
## workingday    -2.7662     2.5179  -1.099  0.27197
## weathersit    -3.2372     1.9470  -1.663  0.09642 .
## temp         238.8135     6.1760  38.668  < 2e-16 ***
## hum         -146.2813     6.9647 -21.003  < 2e-16 ***
## windspeed     17.2018     9.7146   1.771  0.07664 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105 on 8636 degrees of freedom
## Multiple R-squared:  0.3847, Adjusted R-squared:  0.3841
## F-statistic: 674.8 on 8 and 8636 DF,  p-value: < 2.2e-16
```

Adjusted R-squared remains 0.3841. Workingday shows none significant P-values <0.05. remove workingday as feature of model.

```
summary(bike2011.model2 <- lm(cnt ~ .-weekday -workingday, data= train))
```

```
##
## Call:
## lm(formula = cnt ~ . - weekday - workingday, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -238.25  -69.65  -20.83   44.91  410.63
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.7720     6.3059   1.391   0.1642
## season        17.8673     1.1242  15.894   <2e-16 ***
## hr             6.0162     0.1712  35.135   <2e-16 ***
## holiday      -16.4581     6.8934  -2.388   0.0170 *
## weathersit    -3.3939     1.9418  -1.748   0.0805 .
## temp         238.4447     6.1669  38.665   <2e-16 ***
## hum         -146.2143     6.9645 -20.994   <2e-16 ***
## windspeed     17.1524     9.7146   1.766   0.0775 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105 on 8637 degrees of freedom
## Multiple R-squared:  0.3846, Adjusted R-squared:  0.3841
## F-statistic:   771 on 7 and 8637 DF,  p-value: < 2.2e-16
```

Adjusted R-squared remain 0.3841. weathersit and windspeed show less significant P-value <0.05.Since windspeed have many varies of degree. They might mislead the model. So that remove windspeed.

```
summary(bike2011.model3 <- lm(cnt ~ .-weekday -workingday- windspeed, data= train))
```

```
##
## Call:
## lm(formula = cnt ~ . - weekday - workingday - windspeed, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -237.10  -69.86  -21.06   44.69  408.18
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.5175     5.7051   2.369   0.0178 *
## season        17.6699     1.1187  15.794   <2e-16 ***
## hr             6.0308     0.1711  35.257   <2e-16 ***
## holiday      -16.5521     6.8941  -2.401   0.0164 *
## weathersit    -2.8352     1.9161  -1.480   0.1390
## temp         238.7393     6.1654  38.722   <2e-16 ***
## hum         -149.4536     6.7193 -22.242   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105 on 8638 degrees of freedom
## Multiple R-squared:  0.3844, Adjusted R-squared:  0.3839
## F-statistic: 898.8 on 6 and 8638 DF,  p-value: < 2.2e-16
```

Adjusted R-squared decrese to 0.3839. weathersit P-value lesser significant. Let's see if we keep windspeed and remove weathersit.

```
summary(bike2011.model4 <- lm(cnt ~ .-weekday -workingday-weathersit, data= train))
```

```
##
## Call:
## lm(formula = cnt ~ . - weekday - workingday - weathersit, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -237.32  -69.43  -20.50   44.69  409.48
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.6190     6.2720   1.215   0.2245
## season        17.9907     1.1221  16.033   <2e-16 ***
## hr             5.9867     0.1704  35.129   <2e-16 ***
## holiday      -16.6979     6.8929  -2.422   0.0154 *
## temp         239.1364     6.1549  38.853   <2e-16 ***
## hum         -151.6505     6.2321 -24.334   <2e-16 ***
## windspeed     14.3854     9.5859   1.501   0.1335
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105 on 8638 degrees of freedom
## Multiple R-squared:  0.3844, Adjusted R-squared:  0.3839
## F-statistic: 898.8 on 6 and 8638 DF,  p-value: < 2.2e-16
```

We get the same result that of if we select weathersit instead of windspeed. Let's remove both.

```
summary(bike2011.model <- lm(cnt ~ .-weekday -workingday- windspeed-weathersit, data= train))
```

```
##
## Call:
## lm(formula = cnt ~ . - weekday - workingday - windspeed - weathersit,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -236.46  -69.55  -20.71   44.37  408.92
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   11.8712     5.5960   2.121   0.0339 *
## season        17.8031     1.1152  15.964   <2e-16 ***
## hr             6.0035     0.1701  35.301   <2e-16 ***
## holiday      -16.7448     6.8933  -2.429   0.0152 *
## temp         239.2920     6.1545  38.881   <2e-16 ***
## hum         -153.6704     6.0855 -25.252   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105 on 8639 degrees of freedom
## Multiple R-squared:  0.3842, Adjusted R-squared:  0.3838
## F-statistic:  1078 on 5 and 8639 DF,  p-value: < 2.2e-16
```

We get decresed Adjusted R-Sqaure. Let's just keep both weathersit and windspeed for modeling.

```
summary(bike2011.model5<- lm(cnt ~ .-weekday -workingday, data= train))
```

```
##
## Call:
## lm(formula = cnt ~ . - weekday - workingday, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -238.25  -69.65  -20.83   44.91  410.63
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.7720     6.3059   1.391   0.1642
## season        17.8673     1.1242  15.894   <2e-16 ***
## hr             6.0162     0.1712  35.135   <2e-16 ***
## holiday      -16.4581     6.8934  -2.388   0.0170 *
## weathersit    -3.3939     1.9418  -1.748   0.0805 .
## temp         238.4447     6.1669  38.665   <2e-16 ***
## hum         -146.2143     6.9645 -20.994   <2e-16 ***
## windspeed     17.1524     9.7146   1.766   0.0775 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 105 on 8637 degrees of freedom
## Multiple R-squared:  0.3846, Adjusted R-squared:  0.3841
## F-statistic:   771 on 7 and 8637 DF,  p-value: < 2.2e-16
```

```
coef(bike2011.model5)
```

```
## (Intercept)      season          hr     holiday  weathersit        temp
##    8.772001   17.867298    6.016185  -16.458139   -3.393894  238.444708
##         hum   windspeed
## -146.214259   17.152361
```

###Explaine the model manually - temp : Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)

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

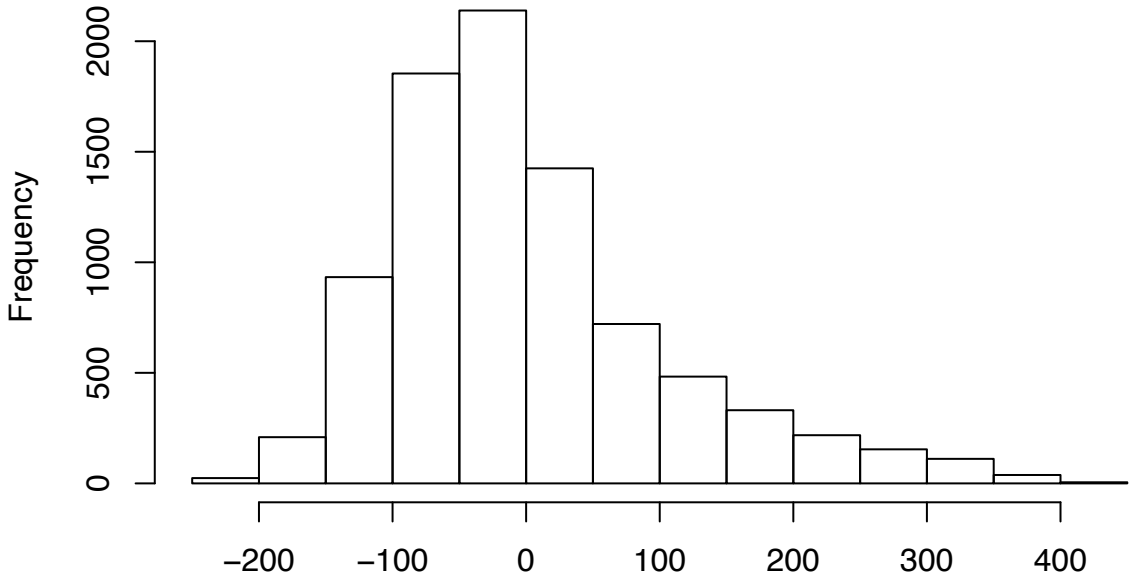
- windspeed: Normalized wind speed. The values are divided to 67 (max)

Evaluate the model

```
res <- residuals(bike2011.model5)
head(as.data.frame(res))
```

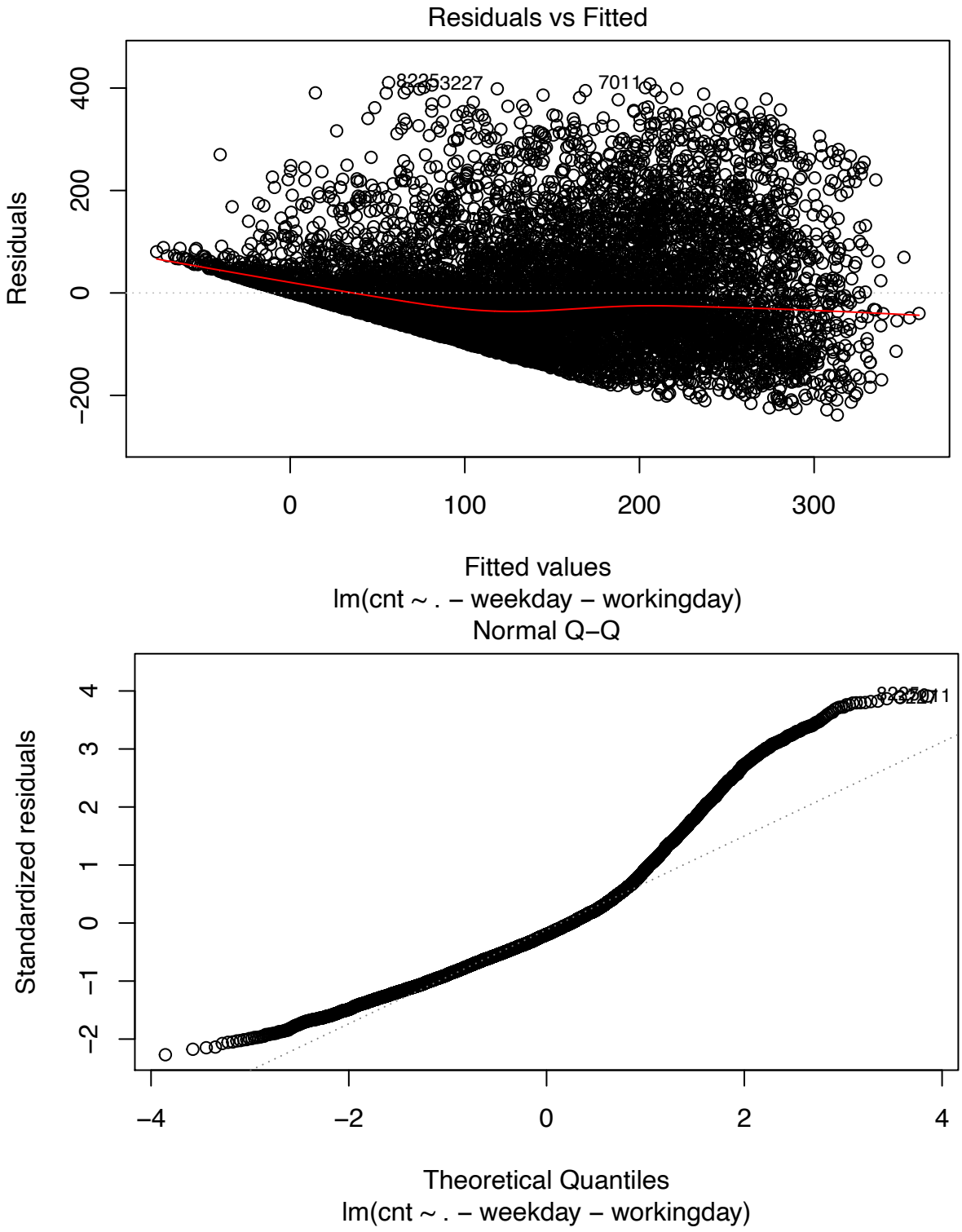
```
##         res
## 1 53.961415
## 2 75.251982
## 3 61.235798
## 4 24.140006
## 5  6.123821
## 6  1.964679
```

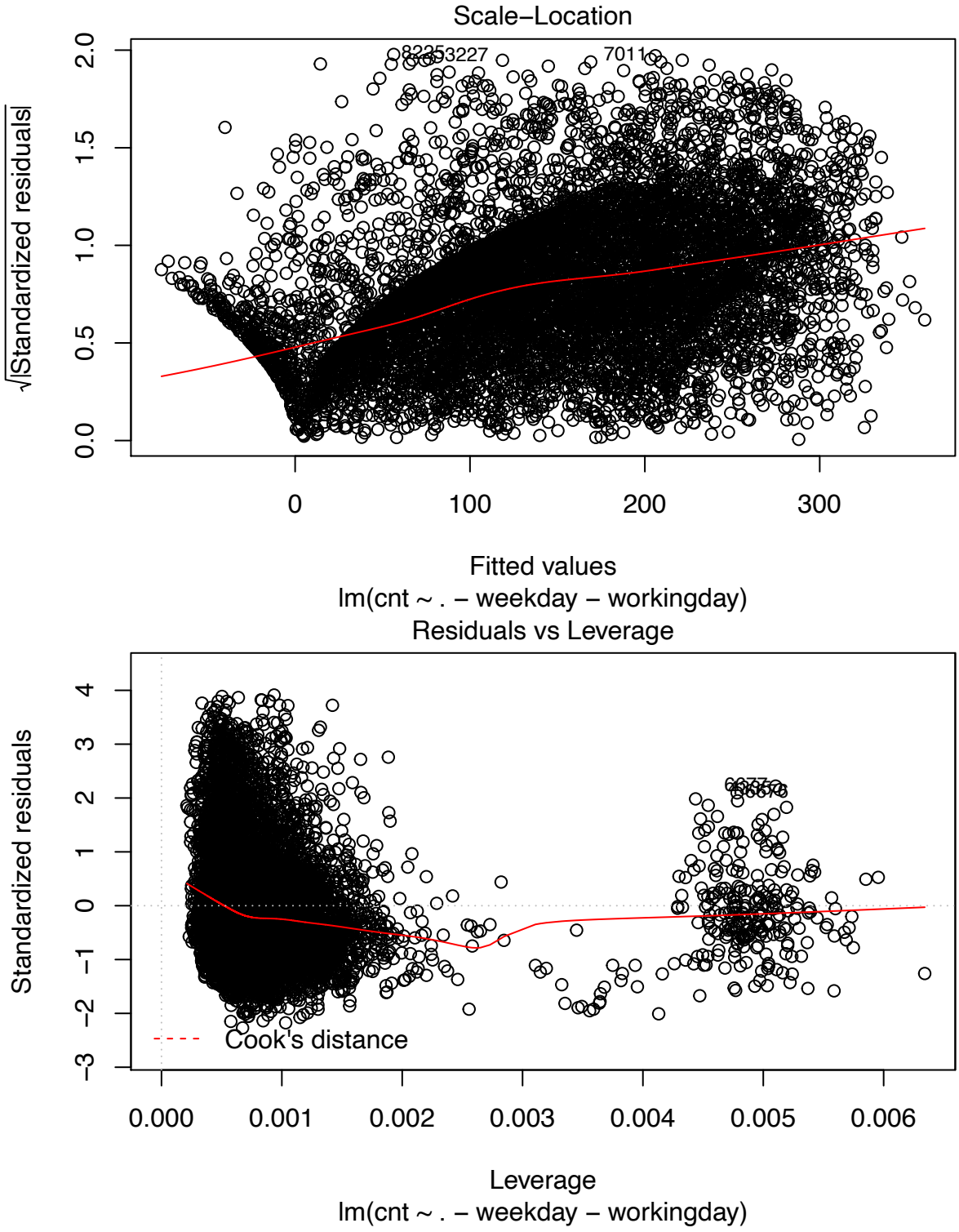
```
hist(residuals(bike2011.model5), xlab = "",main = "")
```



The residuals shows mormally distribution.

```
plot(bike2011.model5)
```


Residuals vs Fitted
Residuals
Fitted values
lm(cnt ~ . – weekday – workingday)
Normal Q–Q
Standardized residuals
Theoretical Quantiles
lm(cnt ~ . – weekday – workingday)



The first plot (residuals vs fitted values) is a simple scatterplot betwenn residuals and predicted values. It should look more or less random.

The second plot (normal Q-Q) is a normal probability plot. It will give a straight line if the errors are distributed mormally.

The third plot (Scale_location), should look random, no patterns.

The last plot(Cook's distance) tells us which points have the greatest influence on the regression (lebverage points). We see row number 13084, 16441 and 13165 have the great influence on the regression.

Prediction future bike rents by the bike2012.model.

```
cnt.pred <- predict(bike2011.model5,test)
```

Determine the increasing ratio of the counts of bike rented from 2011 to 2012.

```
sum(bike2012$cnt)
```

```
## [1] 2049576
```

```
sum(bike2011$cnt)
```

```
## [1] 1243103
```

```
sum(bike2012$cnt)/sum(bike2011$cnt)
```

```
## [1] 1.648758
```

There is 1.648758 marketing incresing from 2011 to 2012. We will add the incresing rate to the prediction.

```
results <- cbind(cnt.pred*1.648758,test$cnt)
colnames(results) <- c("predicted", "actual")
results <- as.data.frame(results)

head(results)
```

```
##       predicted actual
## 8646 20.7481170     48
## 8647 34.4653658     93
## 8648  0.7538841     75
## 8649 -9.2432324     52
## 8650 -4.6528562      8
## 8651  5.6877495      5
```

It is no sense if there is any negative number of bike rented. Adjust negative number to zero.

```
to_zero <- function(x){
  if (x<0){
    return(0)
  }else{
    return(x)
  }
}
```

```
results$predicted <- sapply(results$predicted,to_zero)

head(results)
```

```
##       predicted actual
## 8646 20.7481170     48
## 8647 34.4653658     93
## 8648  0.7538841     75
## 8649  0.0000000     52
## 8650  0.0000000      8
## 8651  5.6877495      5
```

Let's predict how many bike will be rent on April 22 2012, Wednsday base on the weather information brocasted in the www.freemeteo.com.

The original dataset had been normalized as follows. - temp : Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)

we need to convert the real teperature to normalized temperature for prediction.

The formular is normed_temp= (real_temp - min)/(max-min). For example the real temp is 15 Celsius, the normalized temp should be 0.38.

```
(15+8)/(39+8) # if the real temp is 15 Celsuis
```

```
## [1] 0.4893617
```

We also need to convert real humidity and windspeed to nomalized data .

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

```
23/100 # if the real humidity is 23%
```

```
## [1] 0.23
```

###- windspeed: Normalized wind speed. The values are divided to 67 (max). The convert formular is real_windspeed * 67

```
21/67 # if the real windspeed is 21Km/H
```

```
## [1] 0.3134328
```

Let's predict on April 22, Spring, at 8am, non_holiday,weather condision is clear, temperature is 15 Celsius, humidity is 23%, windspeed is 21Km/H.

```
predict(bike2011.model, data.frame(season=2, hr=8 , holiday=0, weathersit = 1, temp=0.489 , hum =0.23, w
```

```
##       1
## 177.175
```

```
sum(bike2011$cnt)/(360*24)# average number of bike rend hourly in 2011.
```

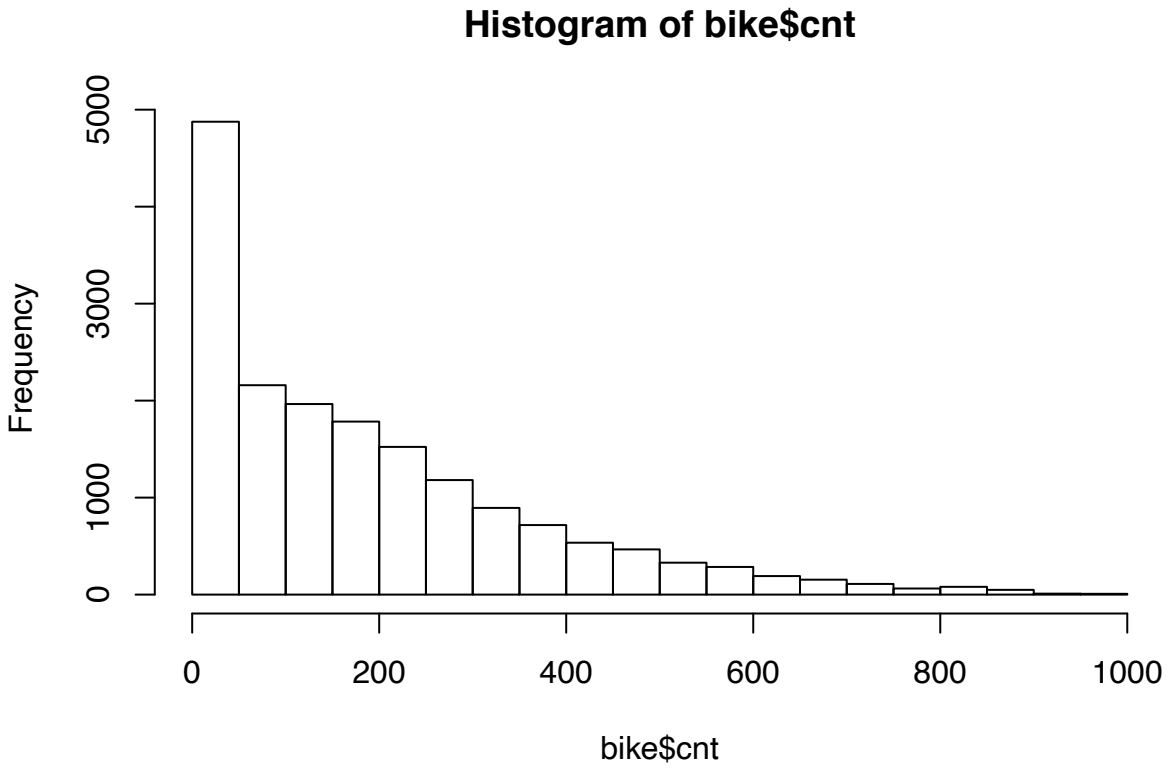
```
## [1] 143.8777
```

```
143.8777*1.64875
```

```
## [1] 237.2184
```

The estimate number of bike could be rent out is in the range of 143 to 237 on April 22, at 8am.

```
hist(bike$cnt)
```



###

The tount of bike shared is not normal distributed. Linear Regression might not be the best model for this time of dataset. Maybe might be one of the choice. Base on current knowledge, we will not go further to do the Poisson model.