ITMD 514 Final Project

Jui Dalal, Vikas Sanil

Submitted on 4/27/2022

```
-- Attaching packages --------------------------------------- tidyverse 1.3.1 --

v ggplot2 3.3.5     v purrr   0.3.4
v tibble  3.1.6     v dplyr   1.0.7
v tidyr   1.1.4     v stringr 1.4.0
v readr   2.1.1     v forcats 0.5.1

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()

Warning: package 'yarrr' was built under R version 4.1.3

Loading required package: jpeg

Loading required package: BayesFactor

Warning: package 'BayesFactor' was built under R version 4.1.3

Loading required package: coda

Warning: package 'coda' was built under R version 4.1.3

Loading required package: Matrix

Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack

************
Welcome to BayesFactor 0.9.12-4.3. If you have questions, please contact Richard Morey (richarddmorey@g

Type BFManual() to open the manual.
************

Loading required package: circlize

Warning: package 'circlize' was built under R version 4.1.3

========================================
circlize version 0.4.14
CRAN page: https://cran.r-project.org/package=circlize
Github page: https://github.com/jokergoo/circlize
Documentation: https://jokergoo.github.io/circlize_book/book/

If you use it in published research, please cite:
Gu, Z. circlize implements and enhances circular visualization
  in R. Bioinformatics 2014.

This message can be suppressed by:
  suppressPackageStartupMessages(library(circlize))
========================================
```

```
yarrr v0.1.5. Citation info at citation('yarrr'). Package guide at yarrr.guide()

Email me at Nathaniel.D.Phillips.is@gmail.com

Attaching package: 'yarrr'

The following object is masked from 'package:ggplot2':

    diamonds

Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine

Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select

Warning: package 'MLmetrics' was built under R version 4.1.3

Attaching package: 'MLmetrics'

The following object is masked from 'package:base':

    Recall

Loading required package: ISLR

Warning: package 'ISLR' was built under R version 4.1.3
```

Data Preperation.

We have selected Seoul Bike Sharing Demand Data from UCI Machine Learning Repository for our final Project.

Why?!

Source:
URL: https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

Abstract: The dataset contains count of public bikes rented at each hour in Seoul Bike sharing System with the corresponding Weather data and Holidays information.

```
'data.frame':   8760 obs. of  14 variables:
 $ Date                     : chr  "01/12/2017" "01/12/2017" "01/12/2017" "01/12/2017" ...
 $ Rented.Bike.Count        : int  254 204 173 107 78 100 181 460 930 490 ...
 $ Hour                     : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Temperature..C.          : num  -5.2 -5.5 -6 -6.2 -6 -6.4 -6.6 -7.4 -7.6 -6.5 ...
 $ Humidity...              : int  37 38 39 40 36 37 35 38 37 27 ...
 $ Wind.speed..m.s.         : num  2.2 0.8 1 0.9 2.3 1.5 1.3 0.9 1.1 0.5 ...
 $ Visibility..10m.         : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 1928 ...
 $ Dew.point.temperature..C.: num  -17.6 -17.6 -17.7 -17.6 -18.6 -18.7 -19.5 -19.3 -19.8 -22.4 ...
 $ Solar.Radiation..MJ.m2.  : num  0 0 0 0 0 0 0 0 0.01 0.23 ...
 $ Rainfall.mm.             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Snowfall..cm.            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Seasons                  : chr  "Winter" "Winter" "Winter" "Winter" ...
 $ Holiday                  : chr  "No Holiday" "No Holiday" "No Holiday" "No Holiday" ...
 $ Functioning.Day          : chr  "Yes" "Yes" "Yes" "Yes" ...

'data.frame':   8760 obs. of  15 variables:
 $ Date               : Date, format: "2017-12-01" "2017-12-01" ...
 $ RentedBikeCount    : int  254 204 173 107 78 100 181 460 930 490 ...
 $ Hour               : int  0 1 2 3 4 5 6 7 8 9 ...
 $ TemperatureC       : num  -5.2 -5.5 -6 -6.2 -6 -6.4 -6.6 -7.4 -7.6 -6.5 ...
 $ Humidity           : int  37 38 39 40 36 37 35 38 37 27 ...
 $ WindSpeedms        : num  2.2 0.8 1 0.9 2.3 1.5 1.3 0.9 1.1 0.5 ...
 $ Visibility10m      : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 1928 ...
 $ DewPointTempC      : num  -17.6 -17.6 -17.7 -17.6 -18.6 -18.7 -19.5 -19.3 -19.8 -22.4 ...
 $ SolarRadiationMJm2: num  0 0 0 0 0 0 0 0 0.01 0.23 ...
 $ Rainfallmm         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Snowfallcm         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Seasons            : chr  "Winter" "Winter" "Winter" "Winter" ...
 $ Holiday            : chr  "No Holiday" "No Holiday" "No Holiday" "No Holiday" ...
 $ FunctioningDay     : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Weekday            : chr  "Thursday" "Thursday" "Thursday" "Thursday" ...

     [,1]          [,2]
[1,] "Observation" "8760"
[2,] "Attributes"  "15"

      Date             RentedBikeCount       Hour        TemperatureC
 Min.   :2017-12-01   Min.   :   0.0   Min.   : 0.00   Min.   :-17.80
 1st Qu.:2018-03-02   1st Qu.: 191.0   1st Qu.: 5.75   1st Qu.:  3.50
 Median :2018-06-01   Median : 504.5   Median :11.50   Median : 13.70
 Mean   :2018-06-01   Mean   : 704.6   Mean   :11.50   Mean   : 12.88
 3rd Qu.:2018-08-31   3rd Qu.:1065.2   3rd Qu.:17.25   3rd Qu.: 22.50
 Max.   :2018-11-30   Max.   :3556.0   Max.   :23.00   Max.   : 39.40
    Humidity      WindSpeedms    Visibility10m  DewPointTempC
 Min.   : 0.00   Min.   :0.000   Min.   :  27   Min.   :-30.600
 1st Qu.:42.00   1st Qu.:0.900   1st Qu.: 940   1st Qu.: -4.700
 Median :57.00   Median :1.500   Median :1698   Median :  5.100
 Mean   :58.23   Mean   :1.725   Mean   :1437   Mean   :  4.074
 3rd Qu.:74.00   3rd Qu.:2.300   3rd Qu.:2000   3rd Qu.: 14.800
 Max.   :98.00   Max.   :7.400   Max.   :2000   Max.   : 27.200
 SolarRadiationMJm2   Rainfallmm        Snowfallcm        Seasons
 Min.   :0.0000     Min.   : 0.0000   Min.   :0.00000   Length:8760
```

```
1st Qu.:0.0000     1st Qu.: 0.0000   1st Qu.:0.00000   Class :character
Median :0.0100     Median : 0.0000   Median :0.00000   Mode  :character
Mean   :0.5691     Mean   : 0.1487   Mean   :0.07507
3rd Qu.:0.9300     3rd Qu.: 0.0000   3rd Qu.:0.00000
Max.   :3.5200     Max.   :35.0000   Max.   :8.80000
  Holiday          FunctioningDay       Weekday
Length:8760        Length:8760        Length:8760
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character
```

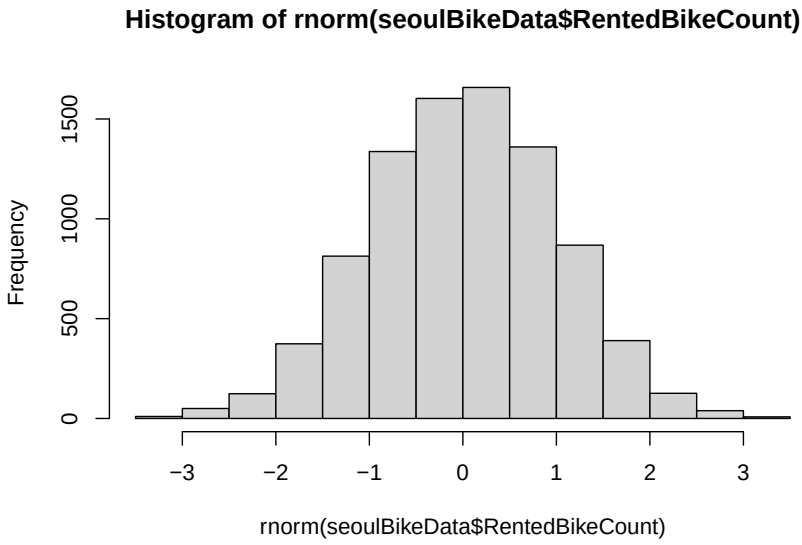
Output of data preperation

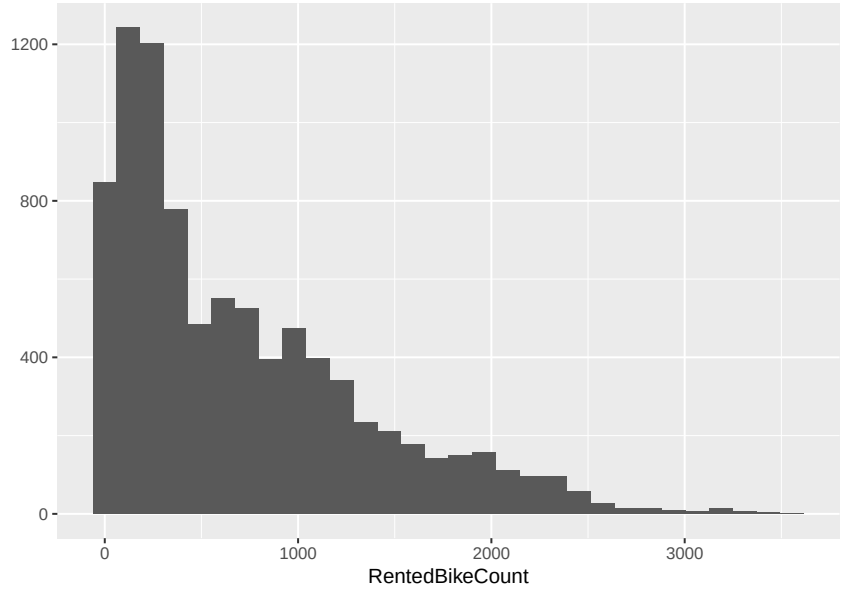
Observation	8760
Attributes	15

Number of Missing values 0.
Number of duplicated value 0.

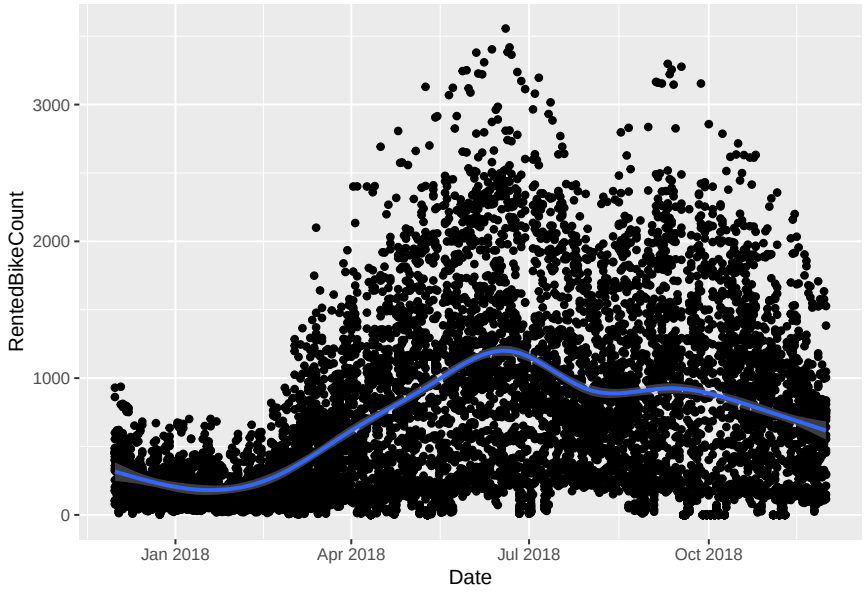
Attribute Information:
Date : year-month-day(Data was in Character type, we have converted it to Date type for processing purpose)
Rented Bike count - Count of bikes rented at each hour
Hour - Hour of the day
Temperature-Temperature in Celsius
Humidity - %
Windspeed - m/s
Visibility - 10m
Dew point temperature - Celsius
Solar radiation - MJ/m2
Rainfall - mm
Snowfall - cm
Seasons - Winter, Spring, Summer, Autumn
Holiday - Holiday/No holiday
Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

We are considering *RentedBikeCount* as our *response/output* in Seoul Bike Sharing Demand data set.

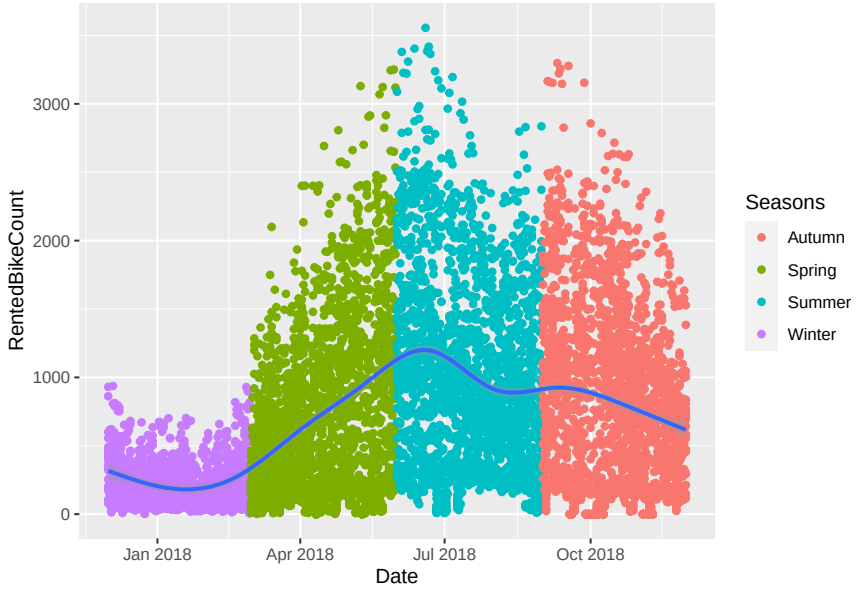




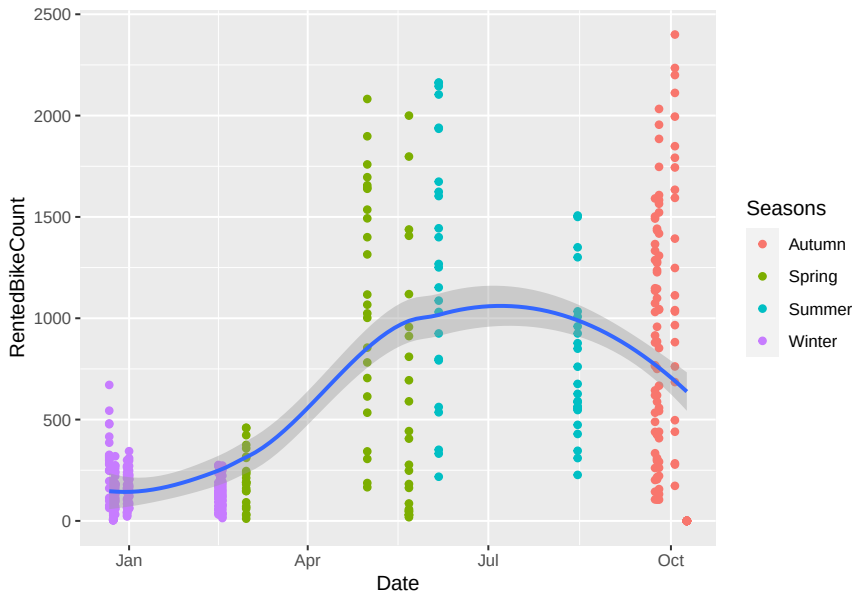

```
‘geom_smooth()‘ using method = ’gam’ and formula ’y ~ s(x, bs = "cs")’
```



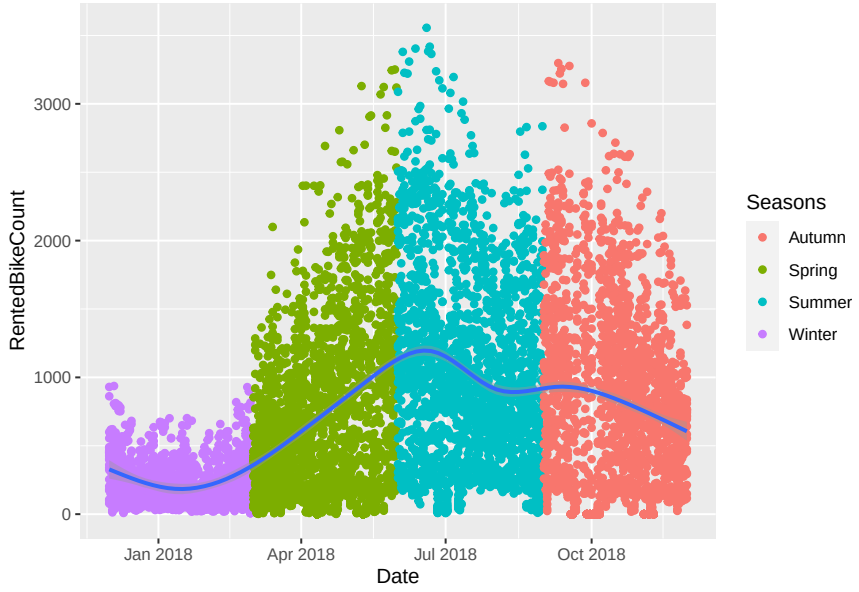

```
`geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

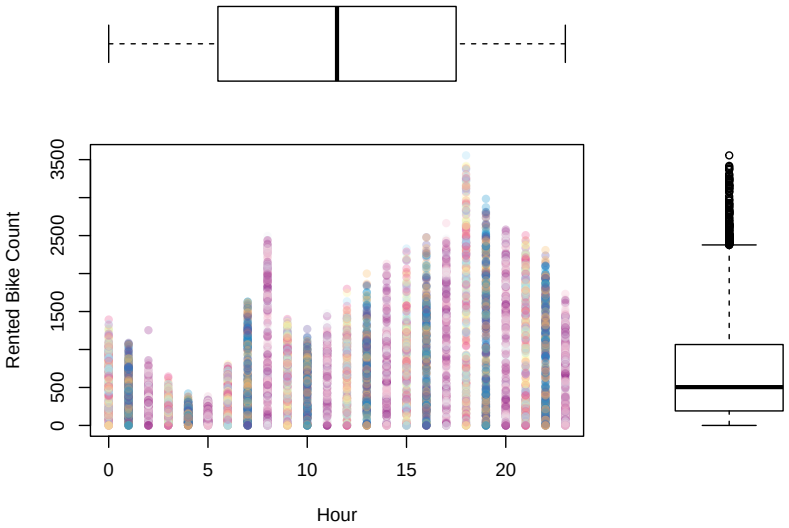



```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

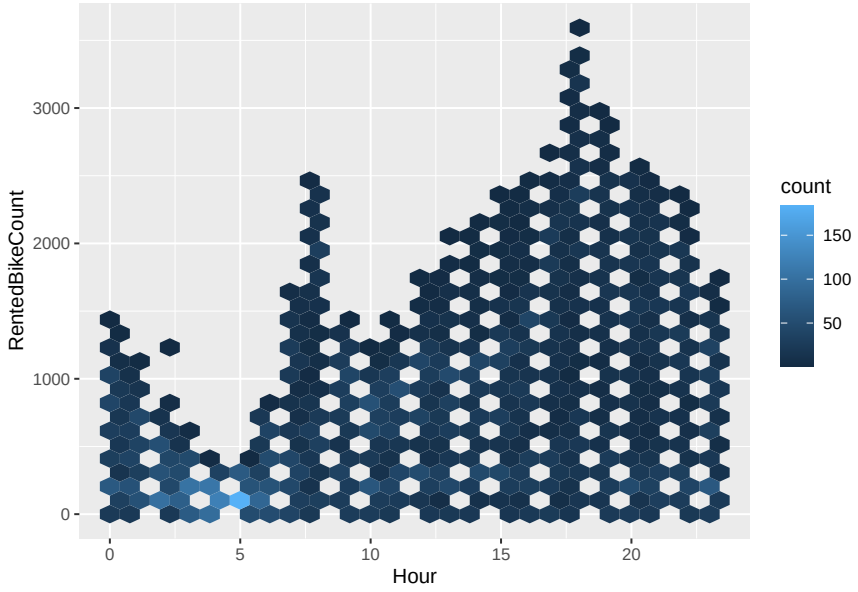


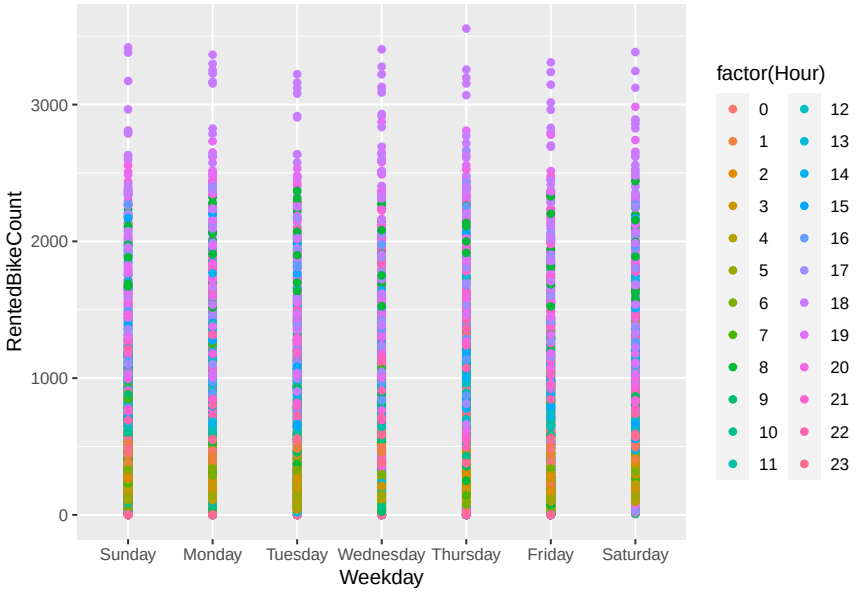

```
`geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

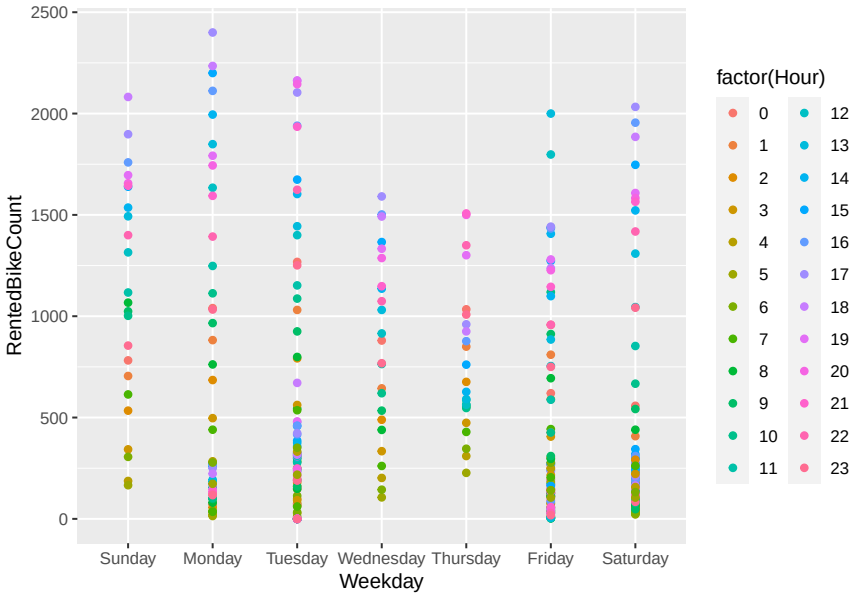


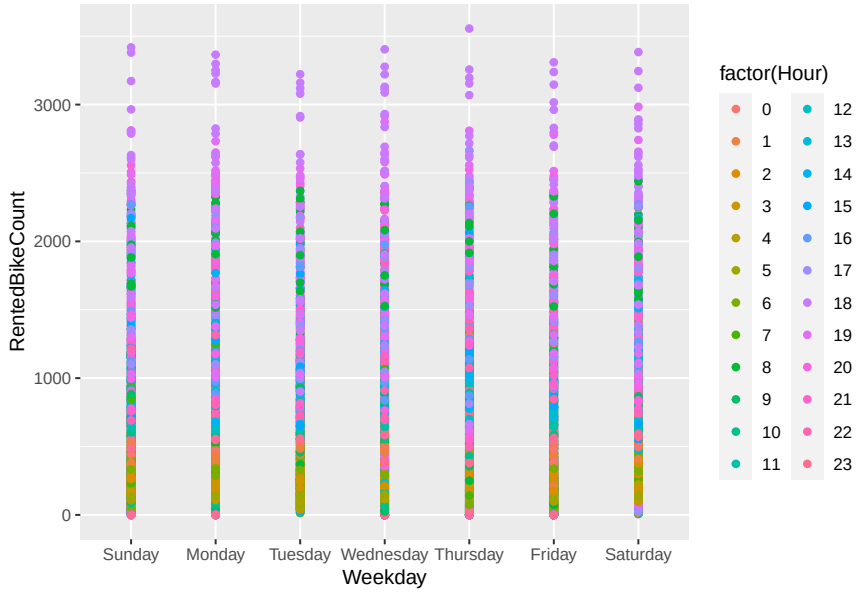



```
[1] "Source: https://bookdown.org/ndphillips/YaRrr/arranging-plots-with-parmfrow-and-layout.html"
```

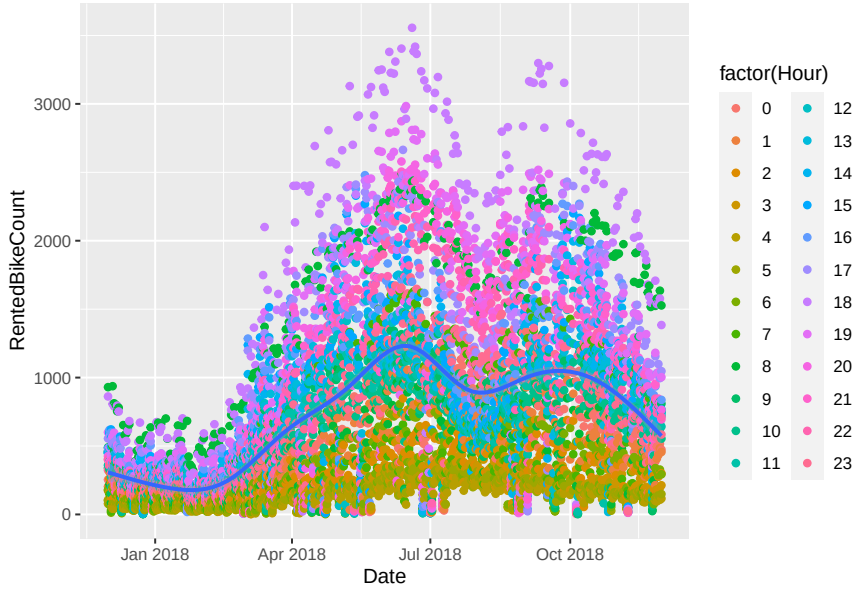



RentedBikeCount
3000
2000
1000
0
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Weekday
factor(Hour)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

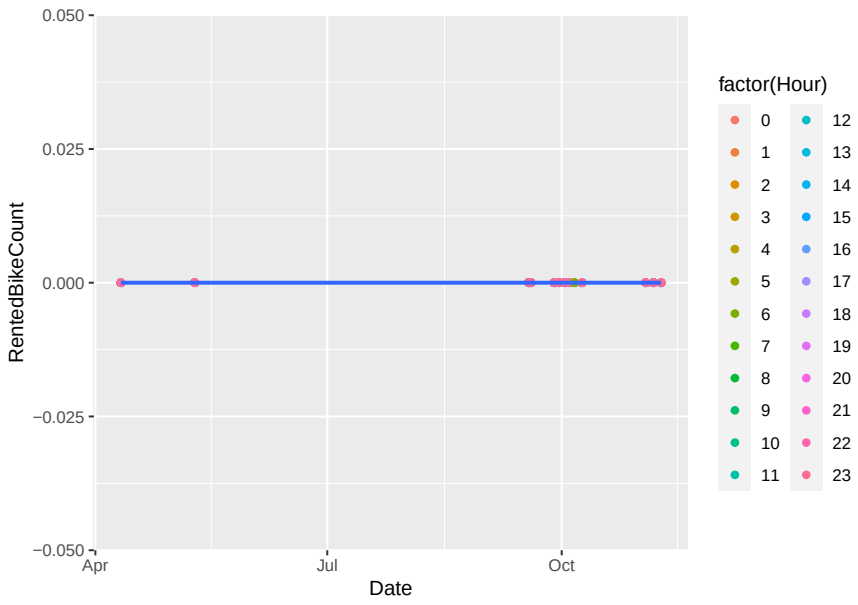

RentedBikeCount
2500
2000
1500
1000
500
0
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Weekday
factor(Hour)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

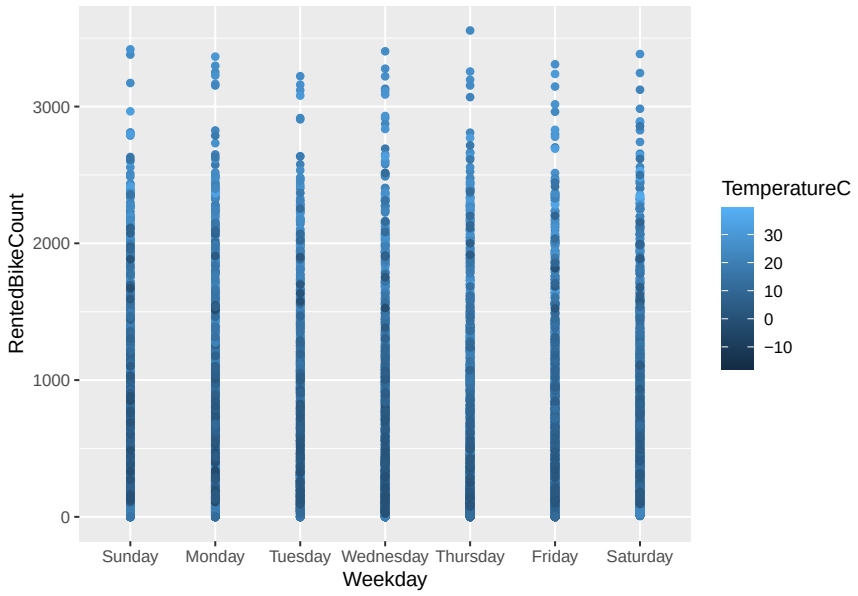


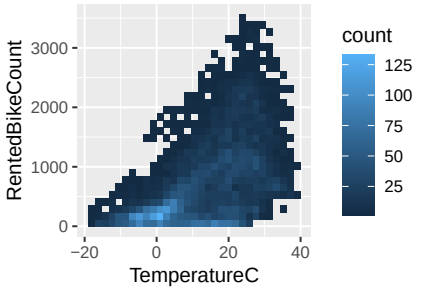

```
`geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

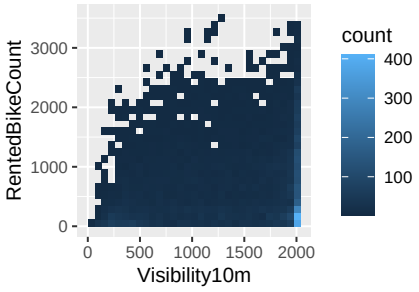


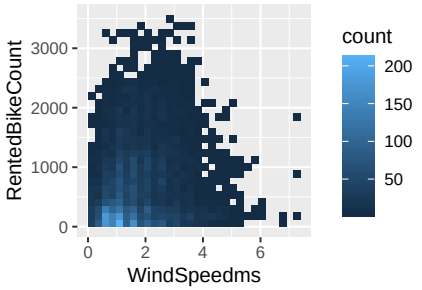

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

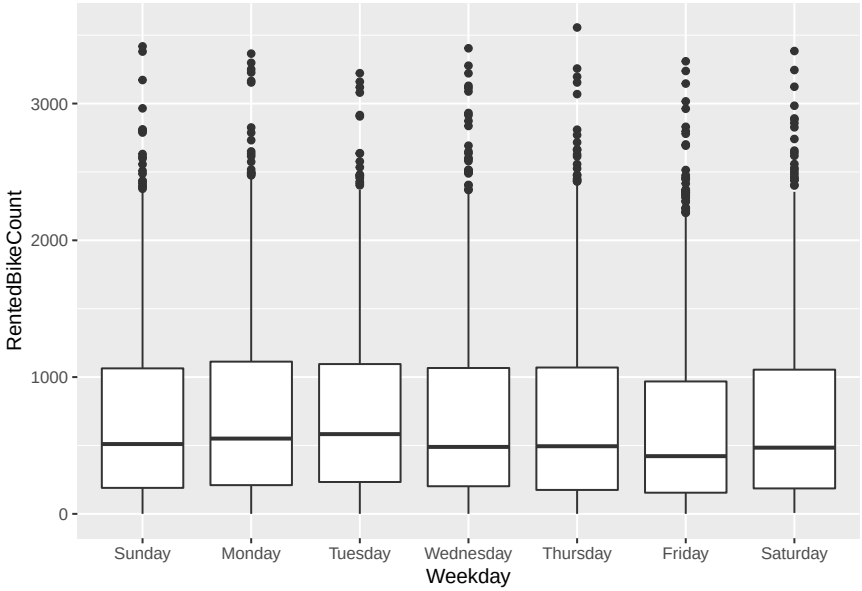

0.050
0.025
0.000
−0.025
−0.050
RentedBikeCount
Apr
Jul
Oct
Date
factor(Hour)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

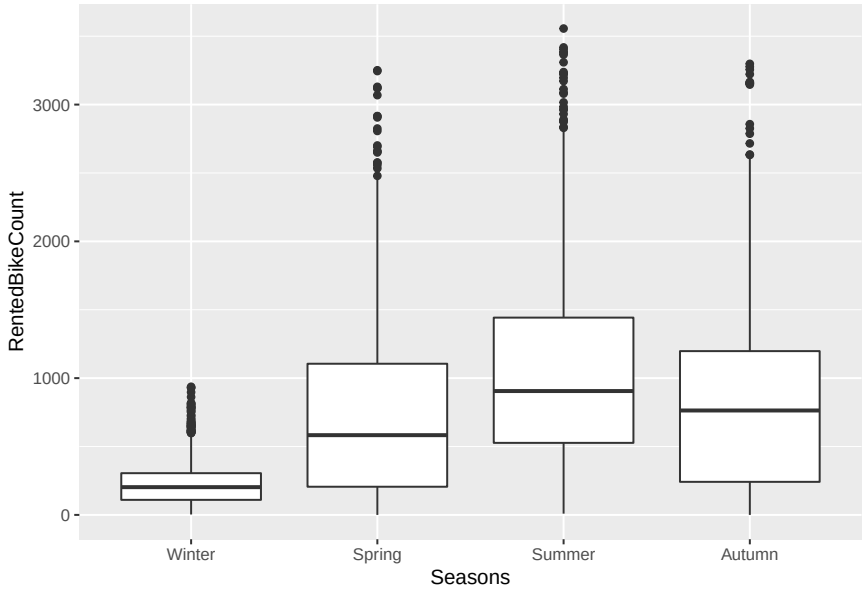

3000
2000
1000
0
RentedBikeCount
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Weekday
TemperatureC
30
20
10
0
−10

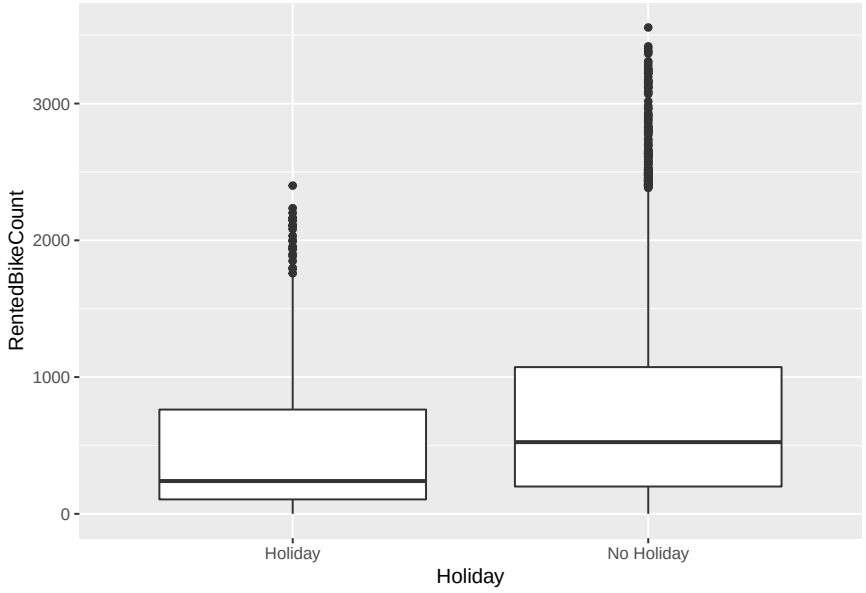

RentedBikeCount
3000
2000
1000
0
−20
0
20
40
TemperatureC
count
125
100
75
50
25

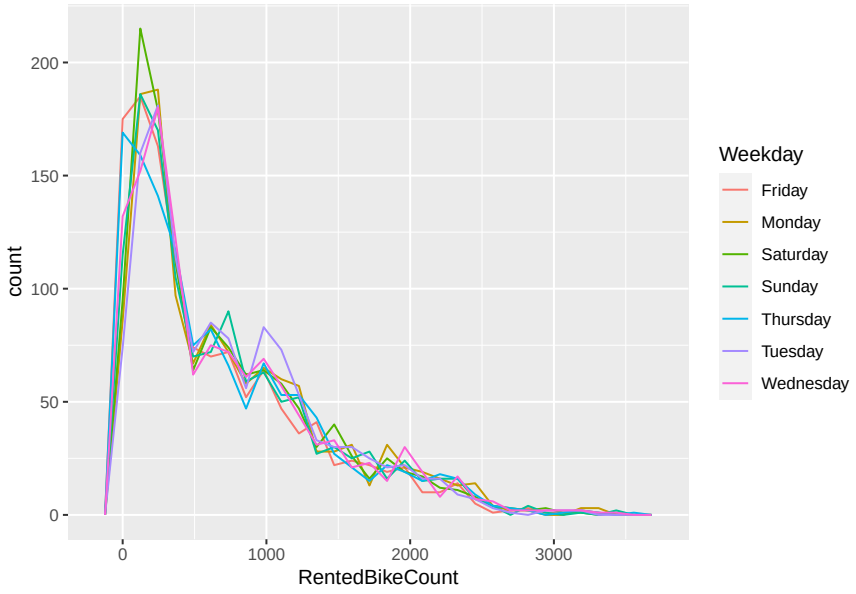

RentedBikeCount
3000
2000
1000
0
0
500
1000
1500
2000
Visibility10m
count
400
300
200
100

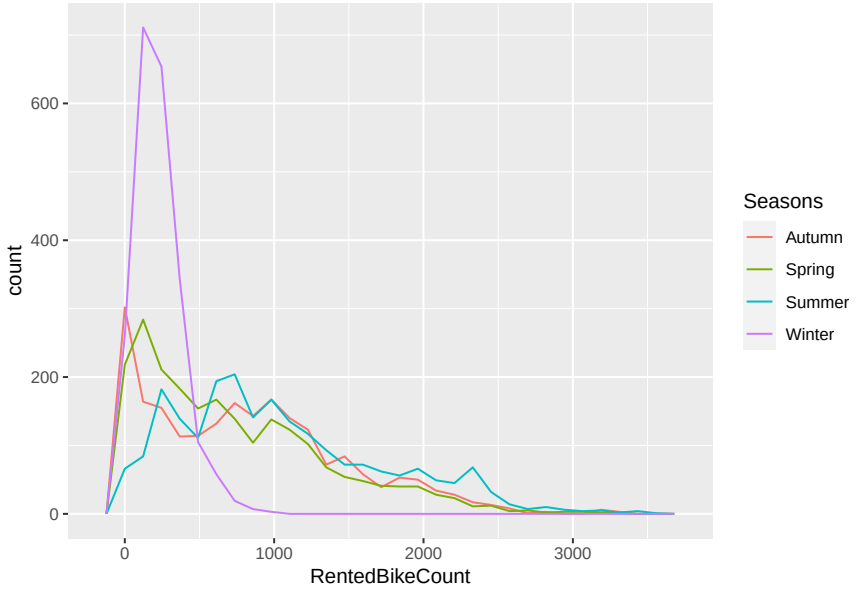

RentedBikeCount
3000
2000
1000
0
0
2
4
6
WindSpeedms
count
200
150
100
50

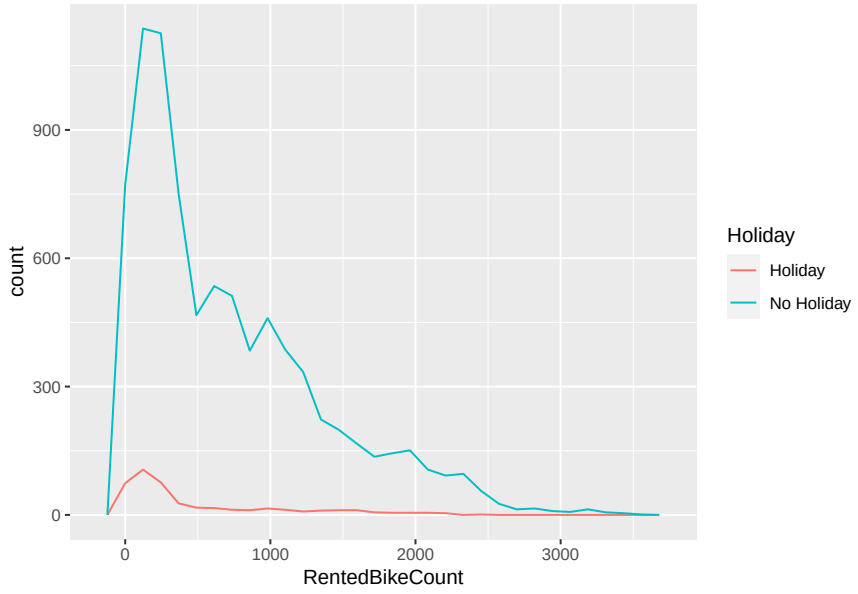

RentedBikeCount
3000
2000
1000
0
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Weekday

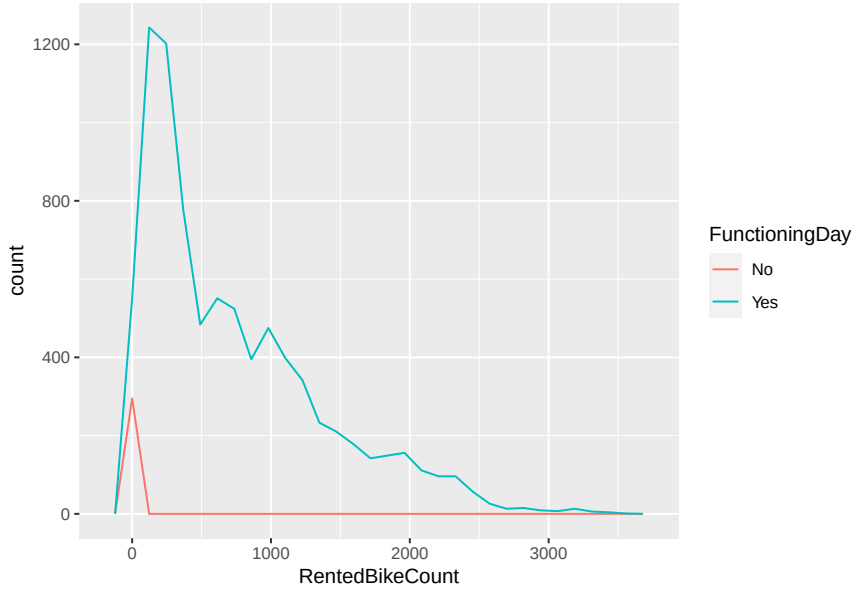

RentedBikeCount
3000
2000
1000
0
Winter
Spring
Summer
Autumn
Seasons


RentedBikeCount
3000
2000
1000
0
Holiday
No Holiday
Holiday


count
200
150
100
50
0
0
1000
2000
3000
RentedBikeCount
Weekday
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday


count
600
400
200
0
0
1000
2000
3000
RentedBikeCount
Seasons
Autumn
Spring
Summer
Winter





We should use t test to compare mean, as we don't know the variance. We can conduct an F test first to see if they have the same variance.

The hypothesis tested:
$H_0 : \sigma^2_{Holiday}/\sigma^2_{NoHoliday} \neq 1$
$H_1 : \sigma^2_{Holiday}/\sigma^2_{NoHoliday} = 1$

```
var.test(holidaySeoulBikeRented$RentedBikeCount, noHolidaySeoulBikeRented$RentedBikeCount)
```

```
    F test to compare two variances

data:  holidaySeoulBikeRented$RentedBikeCount and noHolidaySeoulBikeRented$RentedBikeCount
F = 0.77854, num df = 431, denom df = 8327, p-value = 0.000601
```

```
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6817284 0.8967561
sample estimates:
ratio of variances
         0.7785387
```

Based on the results, we know the variance are quite different, we are doing t-test with different variances. The hypotheis test is \begin{align*} $H_0 : \mu_{Holiday} - \mu_{NoHoliday} \neq 0$,
$H_1 : \mu_{Holiday} - \mu_{NoHoliday} = 0$

```
t.test(holidaySeoulBikeRented$RentedBikeCount, noHolidaySeoulBikeRented$RentedBikeCount)
```

```
    Welch Two Sample t-test

data:  holidaySeoulBikeRented$RentedBikeCount and noHolidaySeoulBikeRented$RentedBikeCount
t = -7.5973, df = 490.23, p-value = 1.545e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -271.1960 -159.7461
sample estimates:
mean of x mean of y
 499.7569  715.2280
```

```
#Simple Linear regression test
fitlm1<-lm(RentedBikeCount~Hour, data=funcDaySeoulBikeRented)
summary(fitlm1)
```

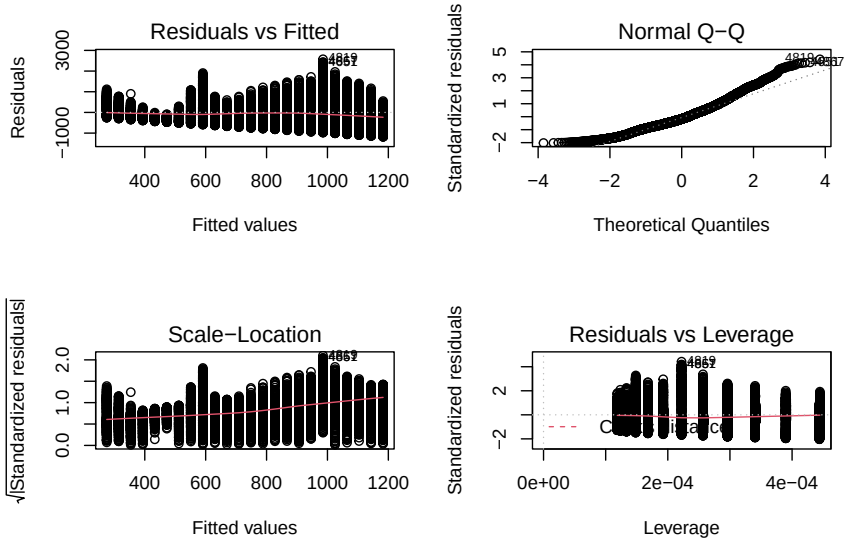
```
Call:
lm(formula = RentedBikeCount ~ Hour, data = funcDaySeoulBikeRented)

Residuals:
     Min       1Q   Median       3Q      Max
-1178.78  -397.80   -99.98   321.63  2570.57

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 274.9817    12.2613   22.43   <2e-16 ***
Hour         39.4694     0.9131   43.22   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 581.4 on 8463 degrees of freedom
Multiple R-squared:  0.1808,    Adjusted R-squared:  0.1807
F-statistic:  1868 on 1 and 8463 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fitlm1)
```




```
#Predicted Rented Bike Count when Hour of the day is 18:00
newHour<-data.frame(Hour = 18)
predict(fitlm1,newHour, interval="confidence")
```

```
       fit      lwr      upr
1 985.4306 968.4447 1002.416
```

```
predict(fitlm1,newHour, interval="prediction")
```

```
       fit    lwr      upr
1 985.4306 -154.4 2125.261
```

```
fitlm2<-lm(RentedBikeCount~Date, data=funcDaySeoulBikeRented)
summary(fitlm2)
```

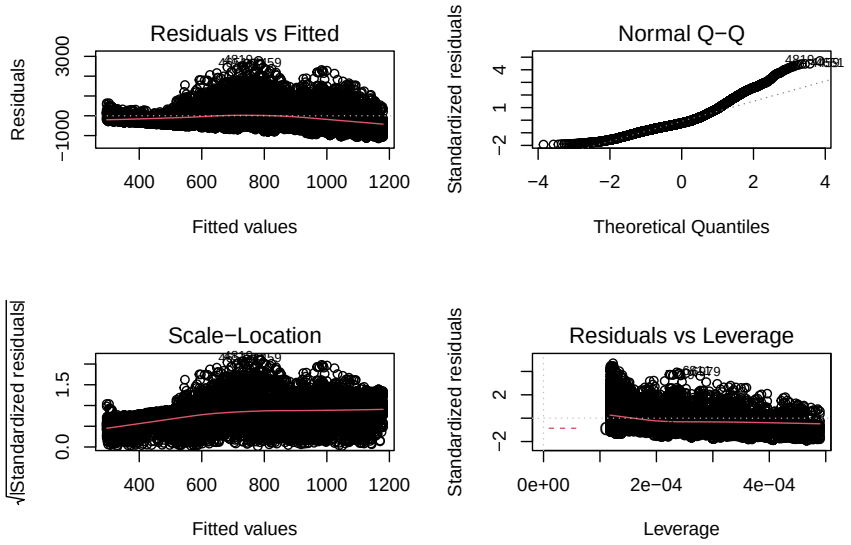
```
Call:
lm(formula = RentedBikeCount ~ Date, data = funcDaySeoulBikeRented)

Residuals:
    Min      1Q  Median      3Q     Max
-1141.2  -367.9  -134.7   265.8  2774.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.219e+04  1.083e+03  -38.96   <2e-16 ***
Date         2.428e+00  6.126e-02   39.63   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 590 on 8463 degrees of freedom
Multiple R-squared:  0.1565,    Adjusted R-squared:  0.1564
F-statistic:  1571 on 1 and 8463 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fitlm2)
```




```
#Predicted Rented Bike Count when Date of the day is 05/05/2018
newDate<-data.frame(Date = as.Date("05/05/2018", "%d/%m/%Y"))
predict(fitlm2,newDate, interval="confidence")
```

```
       fit      lwr      upr
1 672.3822 659.5026 685.2619
```

```
predict(fitlm2,newDate, interval="prediction")
```

```
       fit       lwr      upr
1 672.3822 -484.1779 1828.942
```

```
#res<-cor(funcDaySeoulBikeRented)
#round(res,2)

#i<-sample(2, nrow(funcDaySeoulBikeRented), replace = TRUE, prob = c(0.8,0.2))
#funcDaySeoulBikeRentedTraining<-funcDaySeoulBikeRented[i==1,]
#funcDaySeoulBikeRentedTest<-funcDaySeoulBikeRented[i==2,]

#Forward Stepwise regression model
#null model
#intercept_only<-lm(RentedBikeCount~ 1, data=funcDaySeoulBikeRentedTraining)
#  full model
#all<-lm(RentedBikeCount~., data = funcDaySeoulBikeRentedTraining)
# forward set-wise regression
#forward<- stepAIC(intercept_only, direction='forward', scope=formula(all))
#results
#forward$anova
```

```
#summary(forward)
#MAE and MSE calculation
#ypred_forward<-predict(object = forward, newdata=funcDaySeoulBikeRentedTest)
#MAE(y_pred = ypred_forward, y_true = funcDaySeoulBikeRentedTest$RentedBikeCount)
#MSE(y_pred = ypred_forward, y_true = funcDaySeoulBikeRentedTest$RentedBikeCount)

#Backward Stepwise regression model
#backward<- stepAIC(all, direction='backward')
#results
#backward$anova
#summary(backward)
#MAE and MSE calculation
#ypred_backward<-predict(object = backward, newdata=funcDaySeoulBikeRentedTest)
#MAE(y_pred = ypred_backward, y_true = funcDaySeoulBikeRentedTest$RentedBikeCount)
#MSE(y_pred = ypred_backward, y_true = funcDaySeoulBikeRentedTest$RentedBikeCount)

#Compare Regression Model
#par(mfrow=c(2,2))
#plot(forward)
#par(mfrow=c(2,2))
#plot(backward)
```