R Notebook

The following is your first chunk to start with. Remember, you can add chunks using the menu above (Insert -> R) or using the keyboard shortcut Ctrl+Alt+I. A good practice is to use different code chunks to answer different questions. You can delete this comment if you like.

Other useful keyboard shortcuts include Alt- for the assignment operator, and Ctrl+Shift+M for the pipe operator. You can delete these reminders if you don't want them in your report.

```
setwd("C:/") #Don't forget to set your working directory before you start!

library("tidyverse")
```

```
## -- Attaching packages ----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr 1.3.1 v forcats 0.4.0

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

```
library("tidymodels")
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo

## -- Attaching packages --------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------- tidymodels 0.0.3 --

## v broom     0.5.3     v recipes   0.1.9
## v dials     0.0.4     v rsample   0.0.5
## v infer     0.5.1     v yardstick 0.0.4
## v parsnip   0.0.5

## -- Conflicts ---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------- tidymodels_conflicts() --
## x scales::discard()   masks purrr::discard()
```

```
## x dplyr::filter()     masks stats::filter()
## x recipes::fixed()    masks stringr::fixed()
## x dplyr::lag()        masks stats::lag()
## x dials::margin()     masks ggplot2::margin()
## x yardstick::spec()   masks readr::spec()
## x recipes::step()     masks stats::step()
## x recipes::yj_trans() masks scales::yj_trans()
```

```r
library("plotly")
```

```
##
## Attaching package: 'plotly'

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

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```r
library("skimr")
library("lubridate")
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```r
library('car')
```

```
## Loading required package: carData

## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##  cooks.distance.influence.merMod  lme4
##    dfbeta.influence.merMod        lme4
##   dfbetas.influence.merMod        lme4

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```r
library("caTools")
```

```r
dfbOrg <-
  read_csv("assignment2BikeShare.csv")
```

```
## Parsed with column specification:
## cols(
##   DATE = col_date(format = ""),
##   HOLIDAY =  col_character(),
##   WEEKDAY =  col_character(),
##   WEATHERSIT =  col_double(),
##   TEMP = col_double(),
##   ATEMP =  col_double(),
##   HUMIDITY = col_double(),
##   WINDSPEED = col_double(),
##   CASUAL =  col_double(),
##   REGISTERED = col_double()
## )
```

```r
skim(dfbOrg)
```

Data summary

Name	dfbOrg
Number of rows	731
Number of columns	10

Column type frequency:

character	2
Date	1
numeric	7

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
HOLIDAY	0	1	2	3	0	2	0
WEEKDAY	0	1	2	3	0	2	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
DATE	0	1	2011-01-01	2012-12-31	2012-01-01	731

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
WEATHERSIT	0	1	1.40	0.54	1	1.0	1	2.00	3.00	
TEMP	0	1	15.87	8.83	1	8.0	16	23.15	34.00	
ATEMP	0	1	16.00	9.67	1	6.6	16	23.95	41.00	
HUMIDITY	0	1	63.17	15.47	17	51.0	62	74.00	100.00	
WINDSPEED	0	1	12.82	5.54	0	9.0	12	16.00	40.16	
CASUAL	0	1	848.18	686.62	2	315.5	713	1096.00	3410.00	
REGISTERED	0	1	3656.17	1560.26	20	2497.0	3662	4776.50	6946.00	

```r
head(dfbOrg)
```

```
## # A tibble: 6 x 10
##   DATE       HOLIDAY WEEKDAY WEATHERSIT  TEMP ATEMP HUMIDITY WINDSPEED
CASUAL
##   <date>     <chr>   <chr>        <dbl> <dbl> <dbl>    <dbl>     <dbl>
<dbl>
## 1 2011-01-01 NO      NO               2  11    11       81          17
331
## 2 2011-01-02 NO      NO               2   9     6.5     71.5        17
131
## 3 2011-01-03 NO      YES              1   1     4       44          18
120
## 4 2011-01-04 NO      YES              1   2     2.5     64           9
108
## 5 2011-01-05 NO      YES              1   2.5   1       42.5        13
82
## 6 2011-01-06 NO      YES              1   2     2       52           6
88
## # ... with 1 more variable: REGISTERED <dbl>
```

Question 1)(a) (i)

```
dfbOrg <-dfbOrg %>%
  mutate(COUNT= CASUAL+ REGISTERED)
dfbOrg
```

```
## # A tibble: 731 x 11
##    DATE       HOLIDAY WEEKDAY WEATHERSIT  TEMP ATEMP HUMIDITY WINDSPEED
CASUAL

##    <date>     <chr>   <chr>        <dbl> <dbl> <dbl>    <dbl>     <dbl>

##  1 2011-01-01 NO      NO               2  11    11       81          17
331
##  2 2011-01-02 NO      NO               2   9     6.5     71.5        17
131
##  3 2011-01-03 NO      YES              1   1     4       44          18
120
##  4 2011-01-04 NO      YES              1   2     2.5     64           9
108
##  5 2011-01-05 NO      YES              1   2.5   1       42.5        13
82
##  6 2011-01-06 NO      YES              1   2     2       52           6
88
##  7 2011-01-07 NO      YES              2   1     3       47.5        11
148
##  8 2011-01-08 NO      NO               2   1     5       51          17
68
##  9 2011-01-09 NO      NO               1   2     8.5     46          25
54
## 10 2011-01-10 NO      YES              1   2     6       50          15
41
## # ... with 721 more rows, and 2 more variables: REGISTERED <dbl>, COUNT
<dbl>
```

Question 1)(a) (ii)

```
y <- months(dfbOrg$DATE,abbr =TRUE)

dfbOrg <-dfbOrg %>%
  mutate(MONTH= y)
dfbOrg
```

```
## # A tibble: 731 x 12
##    DATE       HOLIDAY WEEKDAY WEATHERSIT  TEMP ATEMP HUMIDITY WINDSPEED
CASUAL
##    <date>     <chr>   <chr>        <dbl> <dbl> <dbl>    <dbl>     <dbl>

##  1 2011-01-01 NO      NO               2  11    11       81          17
331
##  2 2011-01-02 NO      NO               2   9     6.5     71.5        17
131
##  3 2011-01-03 NO      YES              1   1     4       44          18
120
```

```
##  4 2011-01-04 NO      YES              1   2     2.5     64           9
108
##  5 2011-01-05 NO      YES              1   2.5   1       42.5        13
82
##  6 2011-01-06 NO      YES              1   2     2       52           6
88
##  7 2011-01-07 NO      YES              2   1     3       47.5        11
148
##  8 2011-01-08 NO      NO               2   1     5       51          17
68
##  9 2011-01-09 NO      NO               1   2     8.5     46          25
54
## 10 2011-01-10 NO      YES              1   2     6       50          15
41
## # ... with 721 more rows, and 3 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>
```

Question 1)(b)

```
dfbStd <-dfbOrg %>%
  mutate(TEMP= scale(dfbOrg$TEMP, center = TRUE, scale = TRUE)) %>%
  mutate(ATEMP= scale(dfbOrg$ATEMP, center = TRUE, scale = TRUE)) %>%
  mutate(HUMIDITY= scale(dfbOrg$HUMIDITY, center = TRUE, scale = TRUE)) %>%
  mutate(WINDSPEED= scale(dfbOrg$WINDSPEED, center = TRUE, scale = TRUE))

dfbStd

## # A tibble: 731 x 12
##    DATE       HOLIDAY WEEKDAY WEATHERSIT TEMP[,1] ATEMP[,1] HUMIDITY[,1]
##    <date>     <chr>   <chr>        <dbl>    <dbl>     <dbl>        <dbl>
##  1 2011-01-01 NO      NO               2   -0.552    -0.517       1.15
##  2 2011-01-02 NO      NO               2   -0.779    -0.982       0.538
##  3 2011-01-03 NO      YES              1   -1.68     -1.24       -1.24
##  4 2011-01-04 NO      YES              1   -1.57     -1.40        0.0536
##  5 2011-01-05 NO      YES              1   -1.51     -1.55       -1.34
##  6 2011-01-06 NO      YES              1   -1.57     -1.45       -0.722
##  7 2011-01-07 NO      YES              2   -1.68     -1.34       -1.01
##  8 2011-01-08 NO      NO               2   -1.68     -1.14       -0.787
##  9 2011-01-09 NO      NO               1   -1.57     -0.775      -1.11
## 10 2011-01-10 NO      YES              1   -1.57     -1.03       -0.852
## # ... with 721 more rows, and 5 more variables: WINDSPEED[,1] <dbl>,
## #   CASUAL <dbl>, REGISTERED <dbl>, COUNT <dbl>, MONTH <chr>
```

Question 2)(a)

```
fitAll <- lm(COUNT~., data=dfbStd)
summary(fitAll)

## Warning in summary.lm(fitAll): essentially perfect fit: summary may be
## unreliable
```

```
##
## Call:
## lm(formula = COUNT ~ ., data = dfbStd)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.130e-11 -1.608e-13  1.820e-14  1.972e-13  2.883e-11
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept) -4.289e-11  7.537e-12 -5.691e+00 1.85e-08 ***
## DATE         2.909e-15  5.104e-16  5.698e+00 1.77e-08 ***
## HOLIDAYYES  -4.205e-14  3.764e-13 -1.120e-01   0.9111
## WEEKDAYYES  -8.479e-13  2.125e-13 -3.990e+00 7.29e-05 ***
## WEATHERSIT   3.566e-13  1.447e-13  2.465e+00   0.0140 *
## TEMP         3.776e-13  4.324e-13  8.730e-01   0.3828
## ATEMP        4.367e-13  4.049e-13  1.079e+00   0.2812
## HUMIDITY     1.400e-13  8.356e-14  1.676e+00   0.0942 .
## WINDSPEED    7.337e-14  6.537e-14  1.122e+00   0.2621
## CASUAL       1.000e+00  1.612e-16  6.204e+15  < 2e-16 ***
## REGISTERED   1.000e+00  8.696e-17  1.150e+16  < 2e-16 ***
## MONTHAug    -1.965e-13  3.362e-13 -5.840e-01   0.5591
## MONTHDec     1.561e-13  3.439e-13  4.540e-01   0.6501
## MONTHFeb     2.302e-13  3.202e-13  7.190e-01   0.4724
## MONTHJan    -7.314e-14  3.410e-13 -2.150e-01   0.8302
## MONTHJul    -2.267e-13  3.643e-13 -6.220e-01   0.5339
## MONTHJun    -2.030e-13  3.283e-13 -6.180e-01   0.5366
## MONTHMar     1.247e-13  2.839e-13  4.390e-01   0.6607
## MONTHMay    -6.726e-14  2.953e-13 -2.280e-01   0.8199
## MONTHNov     1.349e-13  3.157e-13  4.270e-01   0.6694
## MONTHOct    -2.730e-15  2.900e-13 -9.000e-03   0.9925
## MONTHSep    -1.123e-13  3.088e-13 -3.640e-01   0.7162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.52e-12 on 709 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 5.648e+31 on 21 and 709 DF,  p-value: < 2.2e-16
```

Question 3) (a)

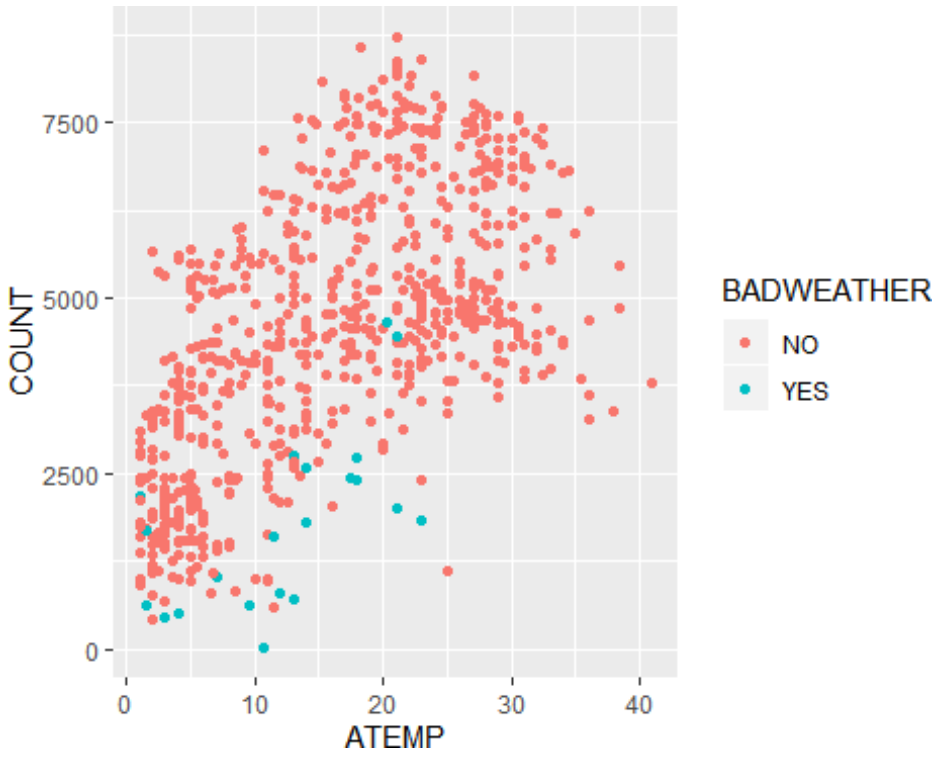
```
dfbOrg <- dfbOrg %>%
   mutate(BADWEATHER = ifelse(WEATHERSIT==3| WEATHERSIT==4 , 'YES', 'NO'))
dfbOrg
```

```
## # A tibble: 731 x 13
##    DATE       HOLIDAY WEEKDAY WEATHERSIT  TEMP ATEMP HUMIDITY WINDSPEED
CASUAL
##    <date>     <chr>   <chr>        <dbl> <dbl> <dbl>    <dbl>     <dbl>
<dbl>
```

```
##  1 2011-01-01 NO      NO                2  11    11       81          17
331
##  2 2011-01-02 NO      NO                2   9     6.5     71.5        17
131
##  3 2011-01-03 NO      YES               1   1     4       44          18
120
##  4 2011-01-04 NO      YES               1   2     2.5     64           9
108
##  5 2011-01-05 NO      YES               1   2.5   1       42.5        13
82
##  6 2011-01-06 NO      YES               1   2     2       52           6
88
##  7 2011-01-07 NO      YES               2   1     3       47.5        11
148
##  8 2011-01-08 NO      NO                2   1     5       51          17
68
##  9 2011-01-09 NO      NO                1   2     8.5     46          25
54
## 10 2011-01-10 NO      YES               1   2     6       50          15
41
## # ... with 721 more rows, and 4 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>, BADWEATHER <chr>
```

Question 3) (b)

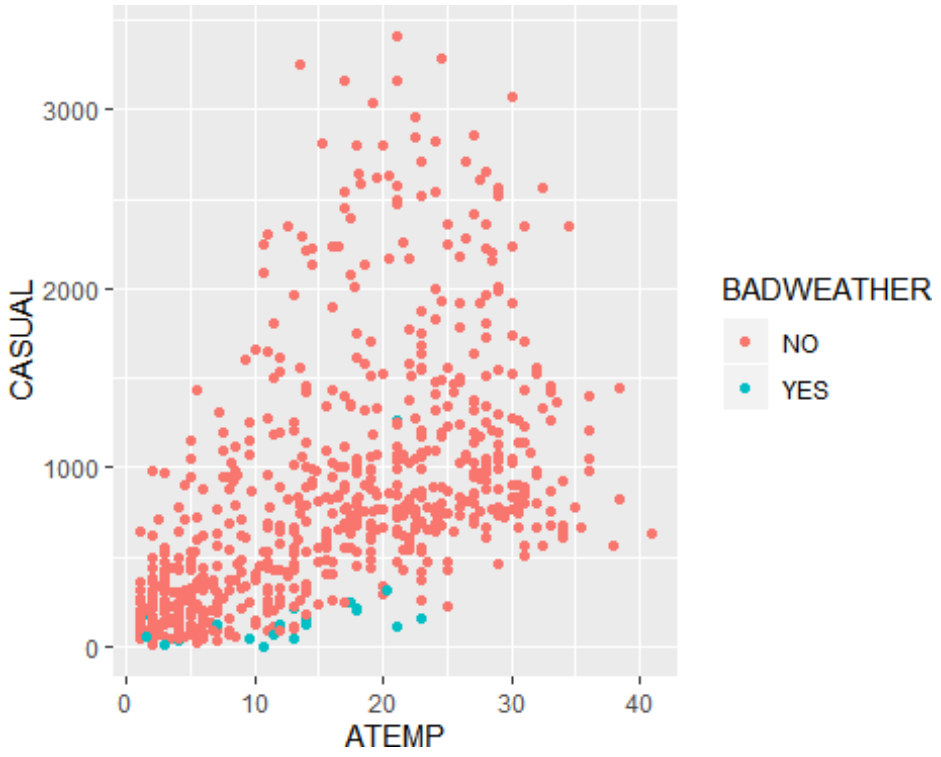
```
plot1 <- dfbOrg %>%
ggplot(mapping= aes(x=ATEMP,y=COUNT,color= BADWEATHER ))+geom_point()
plot1
```



Question 3) (c)

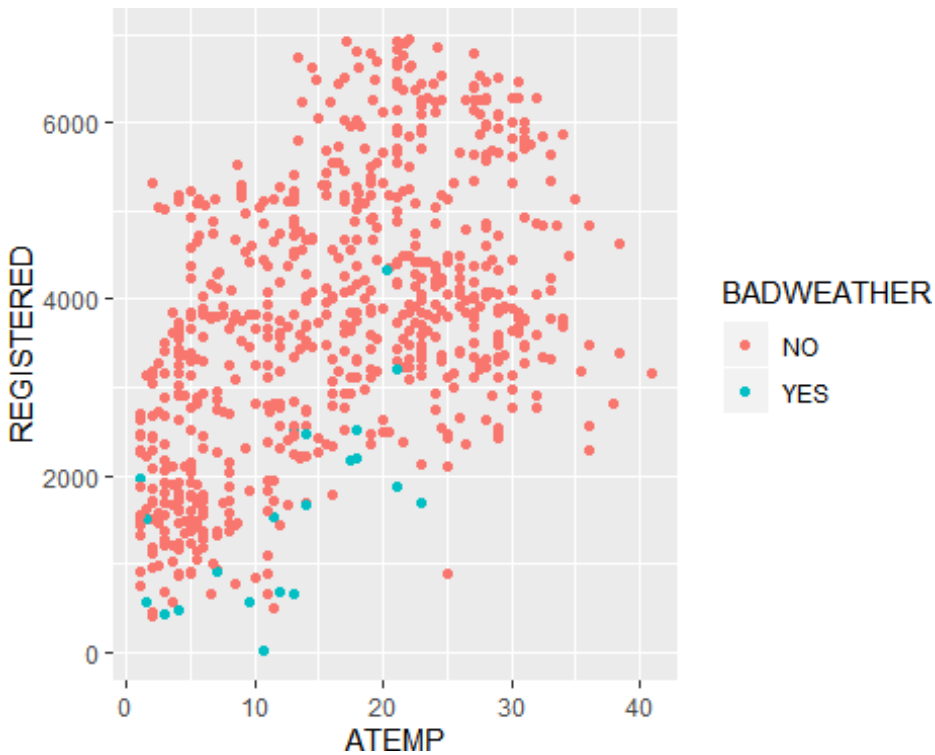
#ATEMP vs. CASUAL

```
plot2 <- dfbOrg %>%
ggplot(mapping= aes(x=ATEMP,y=CASUAL,color= BADWEATHER ))+geom_point()
plot2
```



#ATEMP vs.REGISTERED

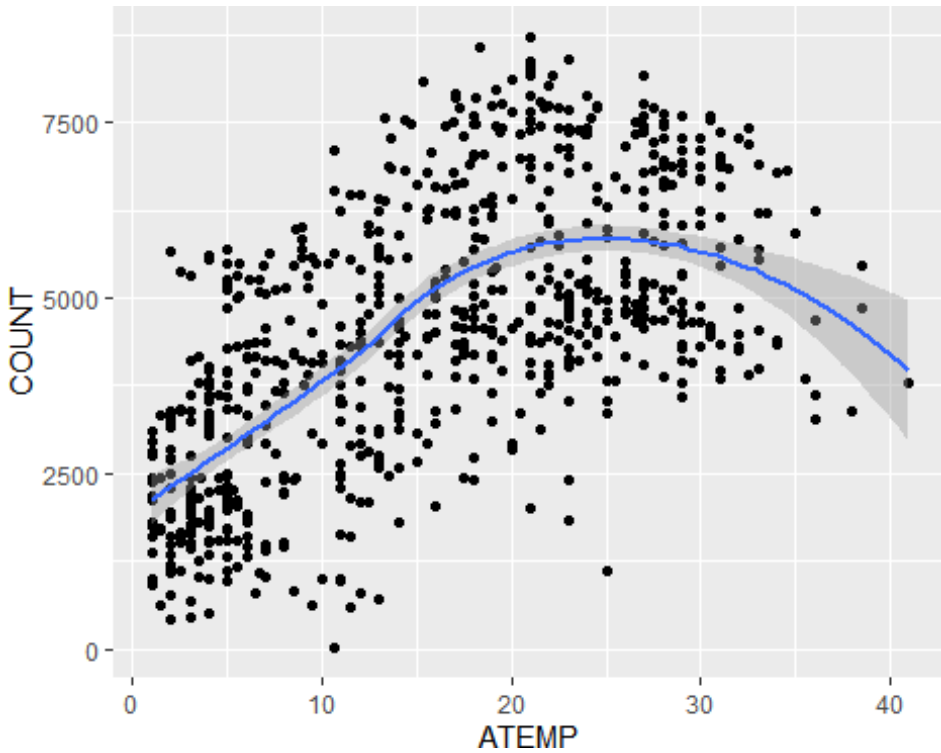
```
plot3 <- dfbOrg %>%
ggplot(mapping= aes(x=ATEMP,y=REGISTERED,color= BADWEATHER ))+geom_point()
plot3
```



Question 3) (c) (iv)

```
plot4 <- dfbOrg %>%
ggplot(mapping= aes(x=ATEMP,y=COUNT ))+geom_point()+geom_smooth()
plot4

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



Question 4)

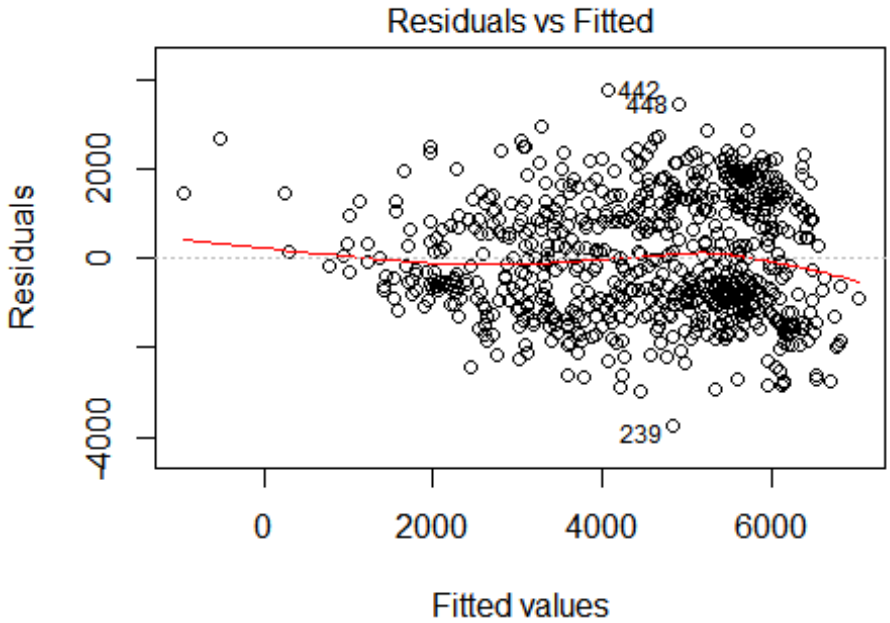
```
fitNew<- lm(formula=COUNT~ MONTH+WEEKDAY+BADWEATHER+TEMP+ATEMP+HUMIDITY,
data=dfbOrg)
summary(fitNew)

##
## Call:
## lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + TEMP + ATEMP +
##     HUMIDITY, data = dfbOrg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -3729.0  -1005.1  -190.3  1115.0  3750.1
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3967.981    335.628  11.823  < 2e-16 ***
## MONTHAug       -209.660    291.004  -0.720  0.47147
## MONTHDec        105.664    265.660   0.398  0.69094
## MONTHFeb       -802.319    273.000  -2.939  0.00340 **
## MONTHJan       -858.334    293.371  -2.926  0.00355 **
## MONTHJul       -676.644    312.956  -2.162  0.03094 *
## MONTHJun       -189.229    286.067  -0.661  0.50851
## MONTHMar       -242.020    249.333  -0.971  0.33204
## MONTHMay        279.730    259.634   1.077  0.28166
## MONTHNov        651.966    257.460   2.532  0.01154 *
```

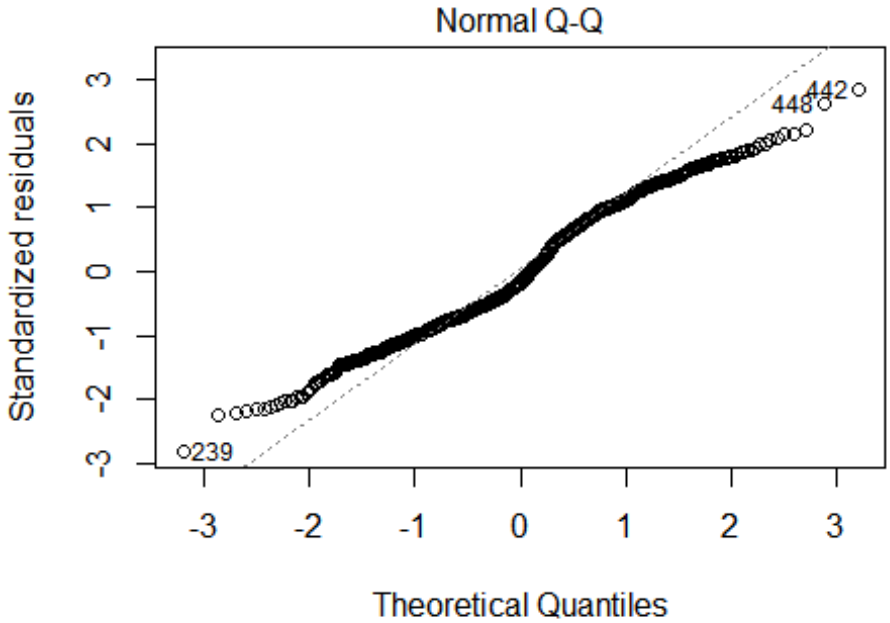
```
## MONTHOct       1072.312    246.970   4.342 1.62e-05 ***
## MONTHSep        742.473    267.293   2.778  0.00562 **
## WEEKDAYYES       69.745    110.118   0.633  0.52670
## BADWEATHERYES -1954.835    316.601  -6.174 1.11e-09 ***
## TEMP            184.596     42.011   4.394 1.28e-05 ***
## ATEMP           -48.640     36.621  -1.328  0.18454
## HUMIDITY        -25.341      3.623  -6.995 6.09e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1341 on 714 degrees of freedom
## Multiple R-squared:  0.5315, Adjusted R-squared:  0.521
## F-statistic: 50.64 on 16 and 714 DF,  p-value: < 2.2e-16
```

Question 5)

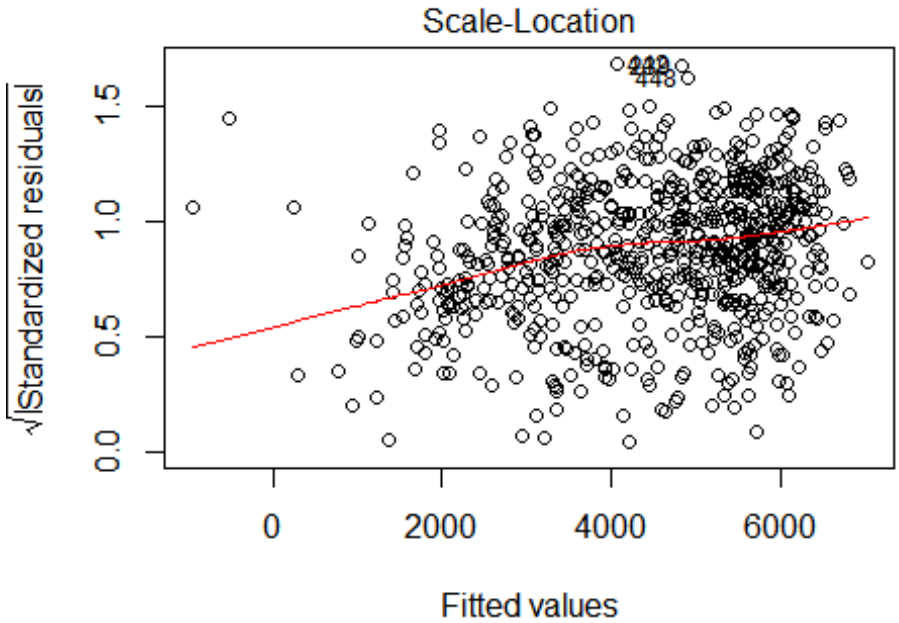
```r
plot(fitNew)
```



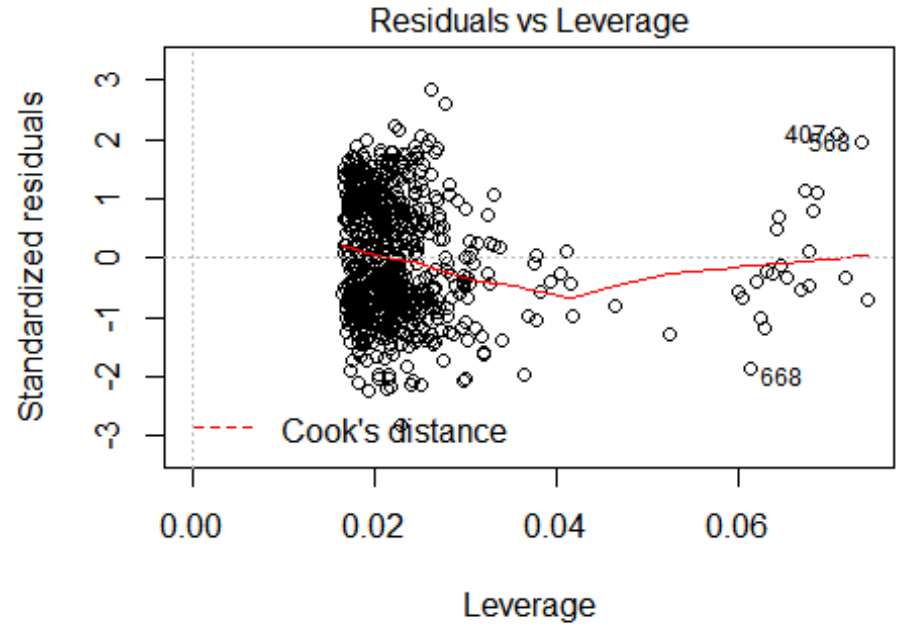
)UNT ~ MONTH + WEEKDAY + BADWEATHER + TEMP + ATEMP +



)UNT ~ MONTH + WEEKDAY + BADWEATHER + TEMP + ATEMP +



UNT ~ MONTH + WEEKDAY + BADWEATHER + TEMP + ATEMP +

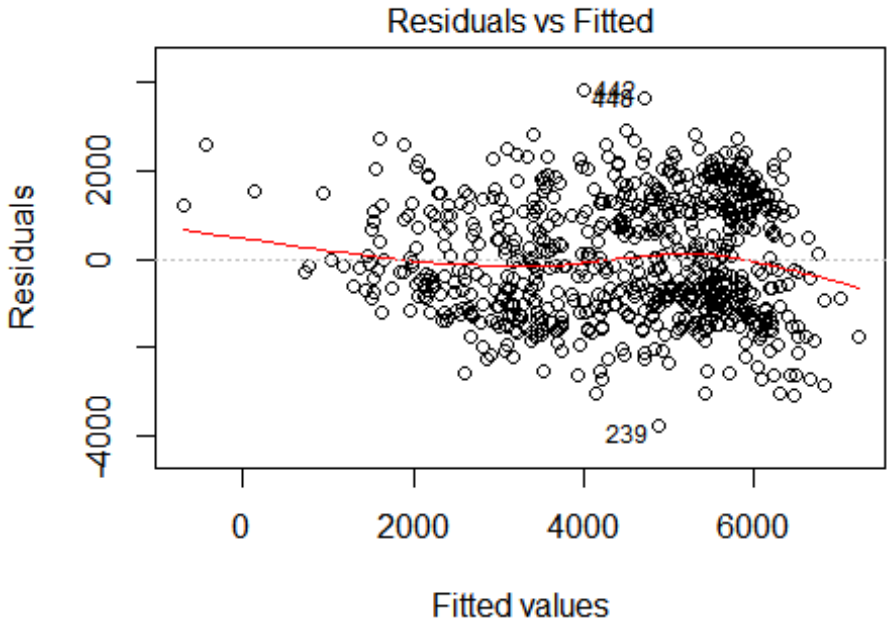


UNT ~ MONTH + WEEKDAY + BADWEATHER + TEMP + ATEMP +

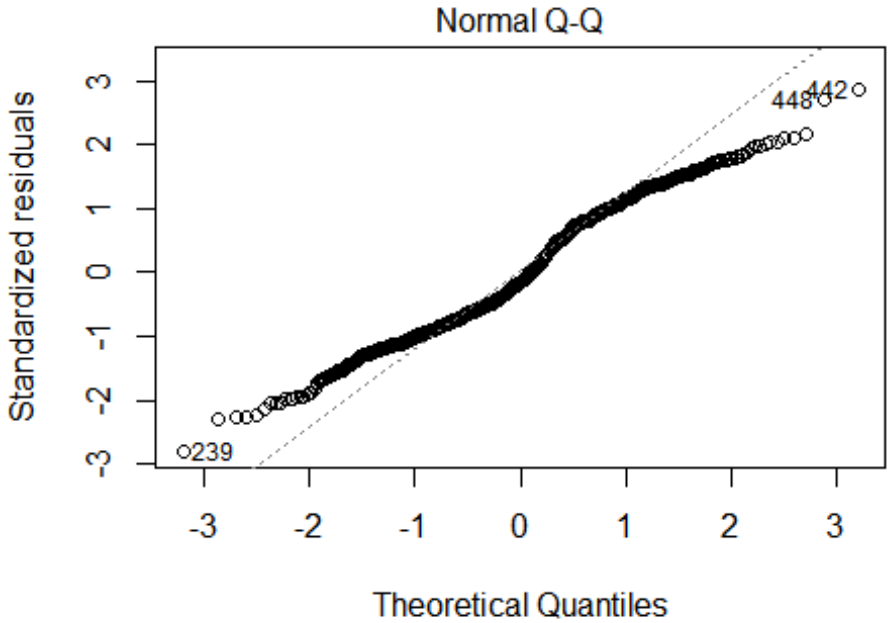
```
fitNew2<- lm(formula=COUNT~ MONTH+WEEKDAY+BADWEATHER+ATEMP+HUMIDITY,
data=dfbOrg)
summary(fitNew2)
```

```
##
## Call:
## lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HUMIDITY,
##     data = dfbOrg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3760.9 -1058.5  -207.5  1154.8  3822.9
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4503.4952   316.6962  14.220  < 2e-16 ***
## MONTHAug        -70.1865   292.9479  -0.240  0.81072
## MONTHDec          0.6468   267.9485   0.002  0.99807
## MONTHFeb      -1016.9096   272.0127  -3.738  0.00020 ***
## MONTHJan      -1386.5736   271.0121  -5.116 4.01e-07 ***
## MONTHJul       -585.3680   316.2385  -1.851  0.06458 .
## MONTHJun        -17.4214   286.9867  -0.061  0.95161
## MONTHMar       -285.6783   252.3046  -1.132  0.25790
## MONTHMay        378.1598   261.9562   1.444  0.14929
## MONTHNov        462.3246   257.0456   1.799  0.07250 .
## MONTHOct       1033.8276   249.9540   4.136 3.95e-05 ***
## MONTHSep        841.6233   269.7273   3.120  0.00188 **
## WEEKDAYYES       91.4446   111.4065   0.821  0.41202
## BADWEATHERYES -1961.8521   320.6243  -6.119 1.55e-09 ***
## ATEMP           103.1721    12.2943   8.392 2.55e-16 ***
## HUMIDITY        -25.4375     3.6686  -6.934 9.16e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1358 on 715 degrees of freedom
## Multiple R-squared:  0.5189, Adjusted R-squared:  0.5088
## F-statistic: 51.41 on 15 and 715 DF,  p-value: < 2.2e-16
```

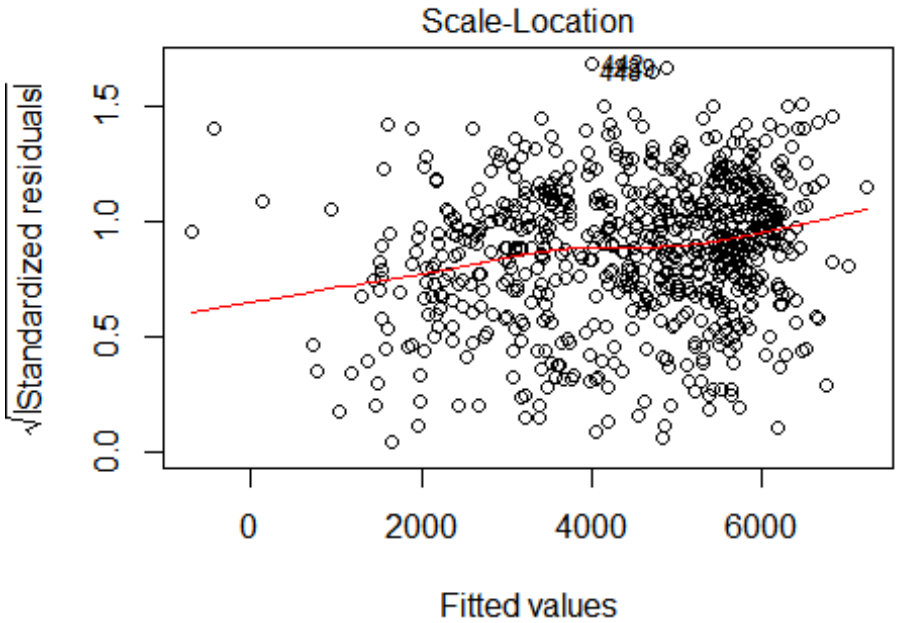
```r
plot(fitNew2)
```



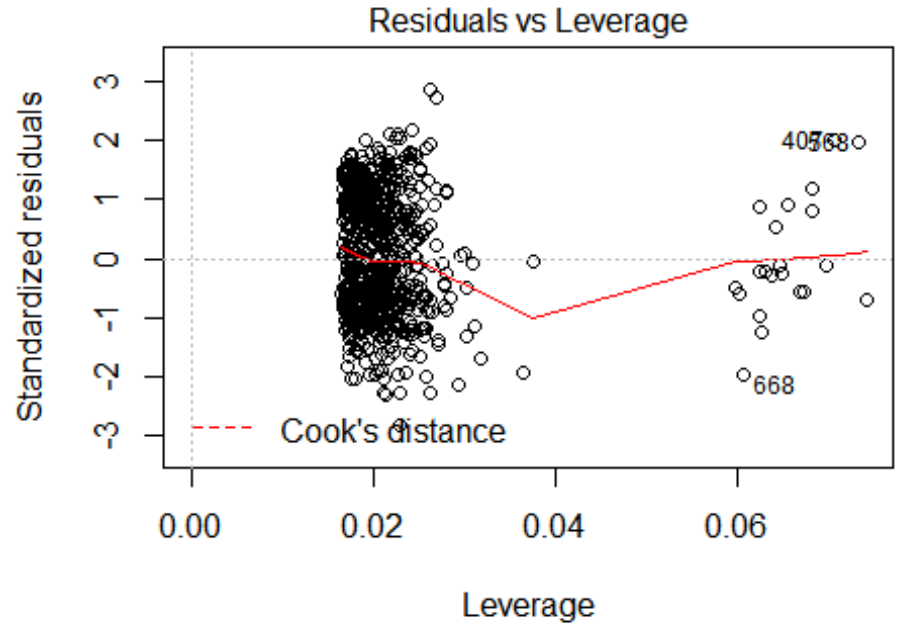
ı(COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HUI



ı(COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HUI



ı(COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HUI



ı(COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HUI

Question 6)

```
fitBadWr<- lm(formula=COUNT~BADWEATHER, data=dfbOrg)
summary(fitBadWr)
```

```
##
## Call:
## lm(formula = COUNT ~ BADWEATHER, data = dfbOrg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4153.2 -1257.7     1.8  1404.8  4129.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4584.24      70.63  64.908  < 2e-16 ***
## BADWEATHERYES -2780.95     416.69  -6.674 4.93e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1882 on 729 degrees of freedom
## Multiple R-squared:  0.05758,    Adjusted R-squared:  0.05629
## F-statistic: 44.54 on 1 and 729 DF,  p-value: 4.934e-11
```

Question 6) c)

```r
fitBadWrWd<- lm(formula=COUNT~ BADWEATHER*WEEKDAY , data=dfbOrg)
summary(fitBadWrWd)
```

```
##
## Call:
## lm(formula = COUNT ~ BADWEATHER * WEEKDAY, data = dfbOrg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4206.7 -1262.1   -3.7   1405.3   4261.5
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4452.5      131.5  33.861  < 2e-16 ***
## BADWEATHERYES             -2637.1      852.2  -3.095  0.00205 **
## WEEKDAYYES                  185.3      155.9   1.188  0.23514
## BADWEATHERYES:WEEKDAYYES   -201.2      977.1  -0.206  0.83695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1883 on 727 degrees of freedom
## Multiple R-squared:  0.05941,    Adjusted R-squared:  0.05553
## F-statistic: 15.31 on 3 and 727 DF,  p-value: 1.15e-09
```

```r
anova(fitBadWr, fitBadWrWd)
```

```
## Analysis of Variance Table
##
## Model 1: COUNT ~ BADWEATHER
## Model 2: COUNT ~ BADWEATHER * WEEKDAY
```

```
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    729 2581793230
## 2    727 2576788128  2   5005101 0.7061 0.4939
```

Question 7) (a), (b)

```r
set.seed(333)
dfbTrain <- dfbOrg %>%
sample_frac(0.8)
dfbTest <- dplyr::setdiff(dfbOrg, dfbTrain)


#First Model
fitOrg <-lm(formula=COUNT~ MONTH+WEEKDAY+BADWEATHER+ATEMP+HUMIDITY,
data=dfbTrain)
summary(fitOrg)
```

```
##
## Call:
## lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HUMIDITY,
##    data = dfbTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -3730.4  -1059.6  -123.3  1136.4  3935.6
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4682.429    349.954  13.380  < 2e-16 ***
## MONTHAug       -180.796    325.897  -0.555 0.579273
## MONTHDec        -66.799    295.882  -0.226 0.821467
## MONTHFeb      -1120.863    303.118  -3.698 0.000239 ***
## MONTHJan      -1437.306    303.674  -4.733 2.79e-06 ***
## MONTHJul       -526.826    347.187  -1.517 0.129718
## MONTHJun        -71.630    310.819  -0.230 0.817820
## MONTHMar       -494.433    280.474  -1.763 0.078463 .
## MONTHMay        330.771    288.889   1.145 0.252700
## MONTHNov        423.187    290.993   1.454 0.146419
## MONTHOct        988.645    281.837   3.508 0.000487 ***
## MONTHSep        663.921    302.925   2.192 0.028806 *
## WEEKDAYYES       88.645    124.513   0.712 0.476797
## BADWEATHERYES -2141.259    368.143  -5.816 1.00e-08 ***
## ATEMP           101.880     13.638   7.470 3.03e-13 ***
## HUMIDITY        -26.229      4.101  -6.396 3.32e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1354 on 569 degrees of freedom
## Multiple R-squared:  0.5219, Adjusted R-squared:  0.5093
## F-statistic:  41.4 on 15 and 569 DF,  p-value: < 2.2e-16
```

```r
resultsOrg <-dfbTest %>%
mutate(predictedCount = predict(fitOrg, dfbTest))
resultsOrg
```

```
## # A tibble: 146 x 14
##    DATE       HOLIDAY WEEKDAY WEATHERSIT  TEMP ATEMP HUMIDITY WINDSPEED
CASUAL
##    <date>     <chr>   <chr>        <dbl> <dbl> <dbl>    <dbl>     <dbl>
<dbl>
##  1 2011-01-10 NO      YES              1   2     6       50          15
41
##  2 2011-01-11 NO      YES              2   1     3.5     57           7
43
##  3 2011-01-13 NO      YES              1   2     7       48.5        20
38
##  4 2011-01-16 NO      NO               1   2.5   2       49.5        15
251
##  5 2011-01-19 NO      YES              2   5.5   2.5     71.5        10
78
##  6 2011-01-20 NO      YES              2   4     2       56          15
83
##  7 2011-01-23 NO      NO               1   4    10       42          15
150
##  8 2011-01-25 NO      YES              2   2     4       65           9
186
##  9 2011-02-13 NO      NO               1   9.5   6       36          20
397
## 10 2011-02-15 NO      YES              1   4     3.5     32          17
140
## # ... with 136 more rows, and 5 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>, BADWEATHER <chr>, predictedCount <dbl>
```

```r
performance <- metric_set(rmse, mae)
performance(resultsOrg, truth= COUNT, estimate=predictedCount)
```

```
## # A tibble: 2 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       1386.
## 2 mae     standard       1175.
```

```r
#Splitting for Second Model
set.seed(333)
dfnwTrain <- dfbOrg %>%
sample_frac(0.8)
dfnwTest <- dplyr::setdiff(dfbOrg, dfnwTrain)

#Second Model
fitNew <-lm(formula=COUNT~ MONTH+WEEKDAY+BADWEATHER+ATEMP+HUMIDITY+WINDSPEED,
data=dfnwTrain)
summary(fitNew)
```

```
##
## Call:
## lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HUMIDITY +
##     WINDSPEED, data = dfnwTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3424.6 -1039.4  -130.3  1131.8  3608.3
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5964.943    420.564  14.183  < 2e-16 ***
## MONTHAug       -278.536    319.113  -0.873  0.38312
## MONTHDec       -290.099    292.348  -0.992  0.32147
## MONTHFeb      -1236.981    297.128  -4.163 3.63e-05 ***
## MONTHJan      -1544.870    297.554  -5.192 2.90e-07 ***
## MONTHJul       -722.754    341.431  -2.117  0.03471 *
## MONTHJun       -191.150    304.683  -0.627  0.53067
## MONTHMar       -536.748    274.285  -1.957  0.05085 .
## MONTHMay        236.844    282.960   0.837  0.40293
## MONTHNov        267.195    286.002   0.934  0.35058
## MONTHOct        831.575    277.124   3.001  0.00281 **
## MONTHSep        508.483    297.594   1.709  0.08806 .
## WEEKDAYYES       70.180    121.764   0.576  0.56460
## BADWEATHERYES -1809.910    365.373  -4.954 9.62e-07 ***
## ATEMP            98.665     13.345   7.393 5.16e-13 ***
## HUMIDITY        -32.191      4.167  -7.726 5.06e-14 ***
## WINDSPEED       -57.016     10.876  -5.242 2.24e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1323 on 568 degrees of freedom
## Multiple R-squared:  0.5439, Adjusted R-squared:  0.5311
## F-statistic: 42.34 on 16 and 568 DF,  p-value: < 2.2e-16
```

```
resultsNew <-dfnwTest %>%
mutate(predictedCount = predict(fitNew, dfnwTest))
resultsNew
```

```
## # A tibble: 146 x 14
##    DATE       HOLIDAY WEEKDAY WEATHERSIT  TEMP ATEMP HUMIDITY WINDSPEED
CASUAL
##    <date>     <chr>   <chr>        <dbl> <dbl> <dbl>    <dbl>     <dbl>
<dbl>
##  1 2011-01-10 NO      YES              1   2     6       50          15
41
##  2 2011-01-11 NO      YES              2   1     3.5     57           7
43
##  3 2011-01-13 NO      YES              1   2     7       48.5        20
38
```

```
##  4 2011-01-16 NO      NO                1   2.5   2       49.5        15
251
##  5 2011-01-19 NO      YES               2   5.5   2.5     71.5        10
78
##  6 2011-01-20 NO      YES               2   4     2       56          15
83
##  7 2011-01-23 NO      NO                1   4    10       42          15
150
##  8 2011-01-25 NO      YES               2   2     4       65           9

##  9 2011-02-13 NO      NO                1   9.5   6       36          20

## 10 2011-02-15 NO      YES               1   4     3.5     32          17
140
## # ... with 136 more rows, and 5 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>, BADWEATHER <chr>, predictedCount <dbl>
```

```
performance <- metric_set(rmse, mae)
performance(resultsNew, truth= COUNT, estimate=predictedCount)
```

```
## # A tibble: 2 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       1340.
## 2 mae     standard       1150.
```

```
#Question 8

dftrainY2011 <- dfbOrg %>%
##    DATE       HOLIDAY WEEKDAY WEATHERSIT  TEMP ATEMP HUMIDITY WINDSPEED
CASUAL
##    <date>     <chr>   <chr>        <dbl> <dbl> <dbl>    <dbl>     <dbl>

##  1 2011-01-01 NO      NO                2  11    11       81          17
331
##  2 2011-01-02 NO      NO                2   9     6.5     71.5        17
131
##  3 2011-01-03 NO      YES               1   1     4       44          18
120
##  4 2011-01-04 NO      YES               1   2     2.5     64           9
108
##  5 2011-01-05 NO      YES               1   2.5   1       42.5        13
82
##  6 2011-01-06 NO      YES               1   2     2       52           6
88
##  7 2011-01-07 NO      YES               2   1     3       47.5        11
148
##  8 2011-01-08 NO      NO                2   1     5       51          17
```

```
68
##   9 2011-01-09 NO      NO                1   2     8.5     46          25
54
## 10 2011-01-10 NO      YES               1   2     6       50          15
41
## # ... with 355 more rows, and 4 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>, BADWEATHER <chr>
```

```
dfTestY2011 <- dfbOrg %>%
  filter(year(DATE)==2012)
dfTestY2011
```

```
## # A tibble: 366 x 13
##    DATE       HOLIDAY WEEKDAY WEATHERSIT  TEMP ATEMP HUMIDITY WINDSPEED
CASUAL
##    <date>     <chr>   <chr>        <dbl> <dbl> <dbl>    <dbl>     <dbl>
<dbl>
##  1 2012-01-01 NO      NO                1  11    11       65          17
686
##  2 2012-01-02 YES     YES               1   4     2       36.5        21
244
##  3 2012-01-03 NO      YES               1   2     8       42.5        24
89
##  4 2012-01-04 NO      YES               2   2     7       42.5        13
95
##  5 2012-01-05 NO      YES               1   3.5   2       56           6
140
##  6 2012-01-06 NO      YES               1   9     7       50          12
307
##  7 2012-01-07 NO      NO                1  10.5   9.5     45          13
1070
##  8 2012-01-08 NO      NO                1   7     5.5     49          14
599
##  9 2012-01-09 NO      YES               2   2     1       70           7
106
## 10 2012-01-10 NO      YES               1   4     4       81          11
173
## # ... with 356 more rows, and 4 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>, BADWEATHER <chr>
```

```
fitYear2011 <-lm(formula=COUNT~ MONTH+WEEKDAY+BADWEATHER+ATEMP+HUMIDITY,
data=dftrainY2011)
summary(fitYear2011)
```

```
##
## Call:
## lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HUMIDITY,
##     data = dftrainY1)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2934.25  -312.97    31.75   367.72  1998.44
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3440.760    231.390  14.870  < 2e-16 ***
## MONTHAug        595.712    199.516   2.986  0.00303 **
## MONTHDec         36.819    178.838   0.206  0.83701
## MONTHFeb      -1233.561    186.054  -6.630 1.28e-10 ***
## MONTHJan      -1613.793    185.158  -8.716  < 2e-16 ***
## MONTHJul        514.856    222.028   2.319  0.02098 *
## MONTHJun        938.944    199.487   4.707 3.63e-06 ***
## MONTHMar       -800.726    178.705  -4.481 1.01e-05 ***
## MONTHMay        969.720    173.973   5.574 4.99e-08 ***
## MONTHNov        548.346    170.652   3.213  0.00143 **
## MONTHOct        999.192    166.284   6.009 4.70e-09 ***
## MONTHSep        996.268    181.094   5.501 7.30e-08 ***
## WEEKDAYYES       11.717     75.181   0.156  0.87624
## BADWEATHERYES -1425.047    186.568  -7.638 2.14e-13 ***
## ATEMP            44.087      8.669   5.086 5.99e-07 ***
## HUMIDITY        -12.969      2.503  -5.182 3.72e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 645.9 on 349 degrees of freedom
## Multiple R-squared:  0.7896, Adjusted R-squared:  0.7806
## F-statistic: 87.32 on 15 and 349 DF,  p-value: < 2.2e-16
```

```
resultsY2011 <-dfTestY2011 %>%
mutate(predictedCount = predict(fitYear2011,
dfTestY2011))
resultsY2011
```

```
## # A tibble: 366 x 14
##    DATE       HOLIDAY WEEKDAY WEATHERSIT  TEMP ATEMP HUMIDITY WINDSPEED
CASUAL
##    <date>     <chr>   <chr>        <dbl> <dbl> <dbl>    <dbl>     <dbl>

##  1 2012-01-01 NO      NO               1  11    11       65          17
686
##  2 2012-01-02 YES     YES              1   4     2       36.5        21
244
##  3 2012-01-03 NO      YES              1   2     8       42.5        24
89
##  4 2012-01-04 NO      YES              2   2     7       42.5        13
95
##  5 2012-01-05 NO      YES              1   3.5   2       56           6
140
##  6 2012-01-06 NO      YES              1   9     7       50          12
```

```
##  7 2012-01-07 NO      NO                1  10.5   9.5     45          13
1070
##  8 2012-01-08 NO      NO                1   7     5.5     49          14
599
##  9 2012-01-09 NO      YES               2   2     1       70           7
106
## 10 2012-01-10 NO      YES               1   4     4       81          11
173
## # ... with 356 more rows, and 5 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>, BADWEATHER <chr>, predictedCount <dbl>
```

```
performance(resultsY2011, truth=COUNT, estimate=predictedCount)
```

```
## # A tibble: 2 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2388.
## 2 mae     standard       2200.
```

```
dfTestY2012 <- dfbOrg %>%
  filter(year(DATE)==2012) %>%
  filter(month(DATE)>06)
```

```
dfTestY2012
```

```
## # A tibble: 184 x 13
##    DATE       HOLIDAY WEEKDAY WEATHERSIT  TEMP ATEMP HUMIDITY WINDSPEED
CASUAL
##    <date>     <chr>   <chr>        <dbl> <dbl> <dbl>    <dbl>     <dbl>
<dbl>
##  1 2012-07-01 NO      NO               1  32    33       44           9
1421
##  2 2012-07-02 NO      YES              1  29    30       51          13
904
##  3 2012-07-03 NO      YES              1  28.5  30       54.5         9
1052
##  4 2012-07-04 YES     YES              1  31.5  32.5     51.5         9
2562
##  5 2012-07-05 NO      YES              1  33    36       47.5        14
1405
##  6 2012-07-06 NO      YES              1  32    33.5     39.5         9
1366
##  7 2012-07-07 NO      NO               1  34    38.5     46.5        11
1448
##  8 2012-07-08 NO      NO               1  31    36       59           7
1203
##  9 2012-07-09 NO      YES              2  26    28       65          11
998
## 10 2012-07-10 NO      YES              2  26    27       74           9
954
## # ... with 174 more rows, and 4 more variables: REGISTERED <dbl>, COUNT
```

```
<dbl>,
## #   MONTH <chr>, BADWEATHER <chr>
```

```
dfTrainY2012 <- dplyr::setdiff(dfbOrg, dfTestY2012)

fitYear2012 <-lm(formula=COUNT~ MONTH+WEEKDAY+BADWEATHER+ATEMP+HUMIDITY,
data= dfTrainY2012)
summary(fitYear2012)
```

```
##
## Call:
## lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HUMIDITY,
##    data  =  dfTrainY2)
##
## Residuals:
##    Min      1Q  Median       3Q     Max
##  -2746.4  -844.2   -59.2   816.3  3727.6
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4430.880    290.817  15.236  < 2e-16 ***
## MONTHAug      -1268.148    292.656  -4.333 1.76e-05 ***
## MONTHDec       -669.439    271.080  -2.470   0.0138 *
## MONTHFeb      -1037.452    233.571  -4.442 1.09e-05 ***
## MONTHJan      -1416.846    233.575  -6.066 2.49e-09 ***
## MONTHJul      -1609.413    319.011  -5.045 6.24e-07 ***
## MONTHJun         52.137    250.054   0.209   0.8349
## MONTHMar       -307.075    213.766  -1.436   0.1514
## MONTHMay        387.779    224.251   1.729   0.0844 .
## MONTHNov       -341.167    262.914  -1.298   0.1950
## MONTHOct        -72.375    257.434  -0.281   0.7787
## MONTHSep       -449.506    277.056  -1.622   0.1053
## WEEKDAYYES      -24.604    108.434  -0.227   0.8206
## BADWEATHERYES -1463.182    301.821  -4.848 1.64e-06 ***
## ATEMP            99.093     11.707   8.464 2.53e-16 ***
## HUMIDITY        -22.078      3.422  -6.451 2.50e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 531 degrees of freedom
## Multiple R-squared:  0.5667, Adjusted R-squared:  0.5545
## F-statistic:  46.3 on 15 and 531 DF,  p-value: < 2.2e-16
```

```
resultsY2012 <- dfTestY2012 %>%
mutate(predictedCount1 = predict(fitYear2012,
dfTestY2012))
resultsY2012
```

```
## # A tibble: 184 x 14
##    DATE       HOLIDAY WEEKDAY WEATHERSIT  TEMP ATEMP HUMIDITY WINDSPEED
CASUAL
##    <date>     <chr>   <chr>        <dbl> <dbl> <dbl>    <dbl>     <dbl>
```

```
<dbl>
##  1 2012-07-01 NO      NO                1  32    33       44           9
1421
##  2 2012-07-02 NO      YES               1  29    30       51          13
904
##  3 2012-07-03 NO      YES               1  28.5  30       54.5         9
1052
##  4 2012-07-04 YES     YES               1  31.5  32.5     51.5         9
2562
##  5 2012-07-05 NO      YES               1  33    36       47.5        14
1405
##  6 2012-07-06 NO      YES               1  32    33.5     39.5         9
1366
##  7 2012-07-07 NO      NO                1  34    38.5     46.5        11
1448
##  8 2012-07-08 NO      NO                1  31    36       59           7
1203
##  9 2012-07-09 NO      YES               2  26    28       65          11
998
## 10 2012-07-10 NO      YES               2  26    27       74           9
954
## # ... with 174 more rows, and 5 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>, BADWEATHER <chr>, predictedCount1 <dbl>
```

```
performance <- metric_set(rmse, mae)
performance(resultsY2012, truth= COUNT, estimate=predictedCount1)
```

```
## # A tibble: 2 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2363.
## 2 mae     standard       2186.
```