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("/Users/shruthinair/Desktop/Lumos/DM") #Don't forget to set your
working directory before you start!

library("tidyverse")

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.0 ──

## ✓ggplot2 3.2.1     ✓purrr   0.3.3
## ✓tibble  2.1.3     ✓dplyr   0.8.3
## ✓tidyr   1.0.2     ✓stringr 1.4.0
## ✓readr   1.3.1     ✓forcats 0.4.0

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

library("tidymodels")

## ── Attaching packages ────────────────────────────────────── tidymodels
0.0.3 ──

## ✓broom     0.5.4     ✓recipes   0.1.9
## ✓dials     0.0.4     ✓rsample   0.0.5
## ✓infer     0.5.1     ✓yardstick 0.0.5
## ✓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

library("skimr")

dfbOrg <-
read_csv("/Users/shruthinair/Desktop/Lumos/DM/Data/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()
## )

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	▂▃▅▃▂

Question 1: a. Create additional variables:

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

b. Scale:

```
dfbStd <- dfbOrg %>%
   mutate_at(c(5:8), funs(c(scale(.))))

## Warning: funs() is soft deprecated as of dplyr 0.8.0
## Please use a list of either functions or lambdas:
##
```

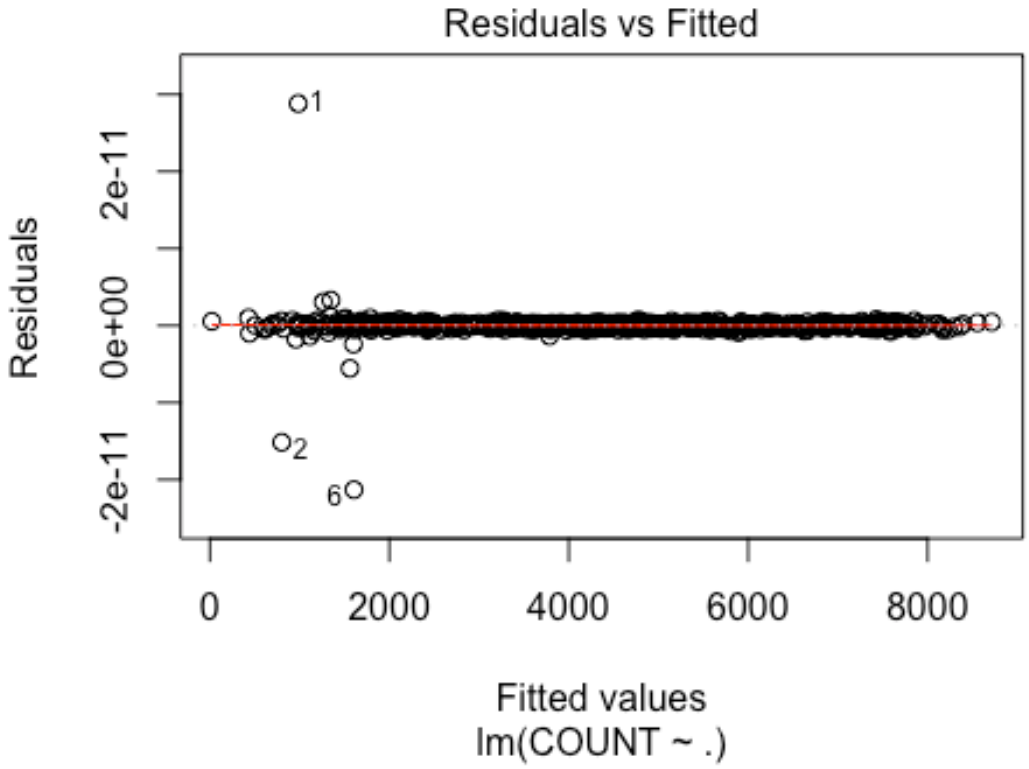
```
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once per session.
```

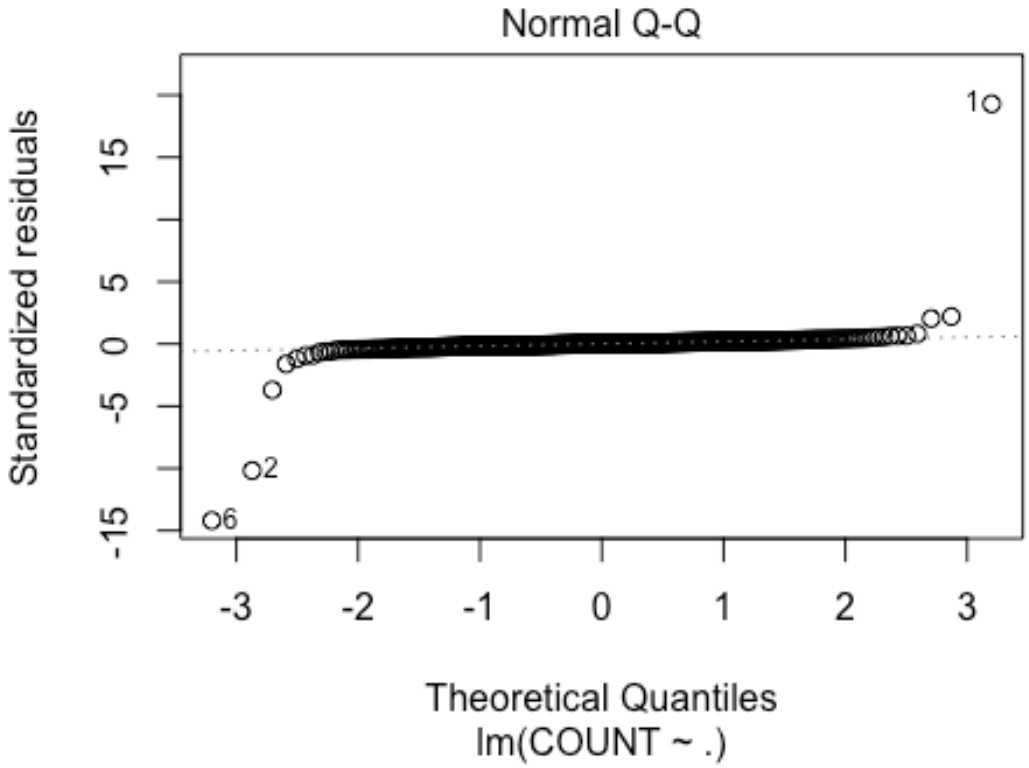
Question 2:

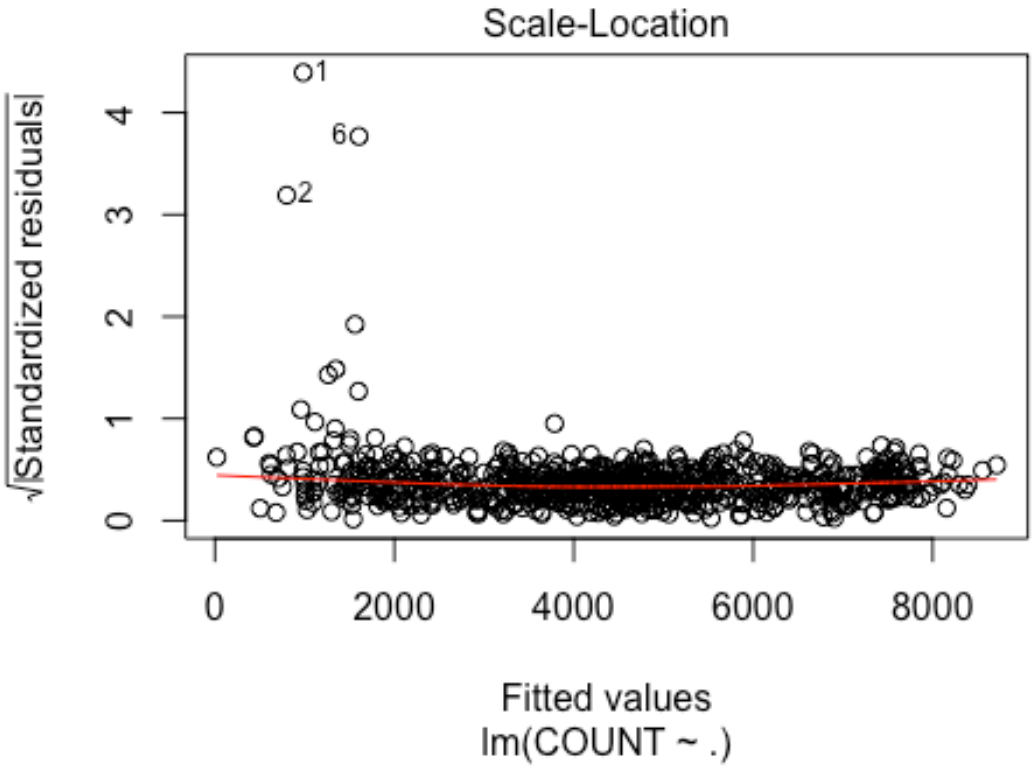
```
fitAll <-
  lm(formula = COUNT ~ ., data = dfbStd)
summary(fitAll)

## 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 ***
## MONTHAugust    -1.965e-13  3.362e-13 -5.840e-01   0.5591
## MONTHDecember   1.561e-13  3.439e-13  4.540e-01   0.6501
## MONTHFebruary   2.302e-13  3.202e-13  7.190e-01   0.4724
## MONTHJanuary   -7.314e-14  3.410e-13 -2.150e-01   0.8302
## MONTHJuly      -2.267e-13  3.643e-13 -6.220e-01   0.5339
## MONTHJune      -2.030e-13  3.283e-13 -6.180e-01   0.5366
## MONTHMarch      1.247e-13  2.839e-13  4.390e-01   0.6607
## MONTHMay       -6.726e-14  2.953e-13 -2.280e-01   0.8199
## MONTHNovember   1.349e-13  3.157e-13  4.270e-01   0.6694
## MONTHOctober   -2.730e-15  2.900e-13 -9.000e-03   0.9925
## MONTHSeptember -1.123e-13  3.088e-13 -3.640e-01   0.7162
```

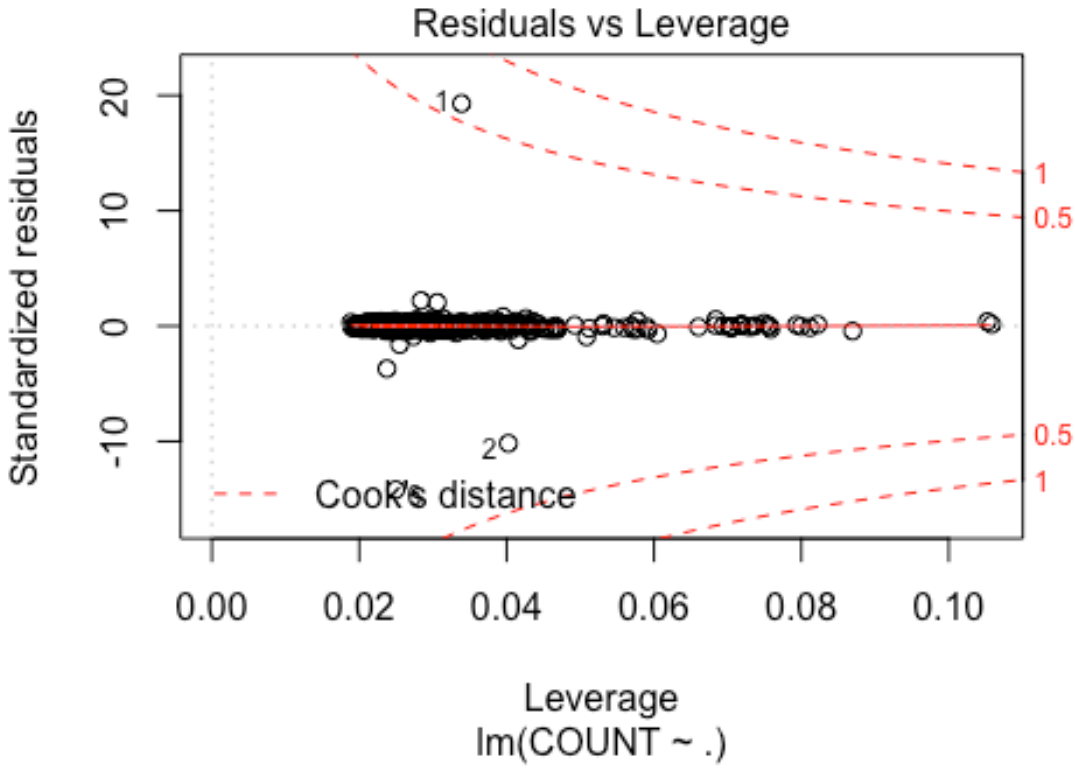
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.52e-12 on 709 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 5.648e+31 on 21 and 709 DF,  p-value: < 2.2e-16
```

```r
plot(fitAll)
```


Residuals vs Fitted
Residuals
2e-11
0e+00
-2e-11
1
2
6
0
2000
4000
6000
8000
Fitted values
lm(COUNT ~ .)


Normal Q-Q
Standardized residuals
15
5
0
-5
-15
1
2
6
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(COUNT ~ .)


Scale-Location
√|Standardized residuals|
0
1
2
3
4
1
6
2
0
2000
4000
6000
8000
Fitted values
lm(COUNT ~ .)



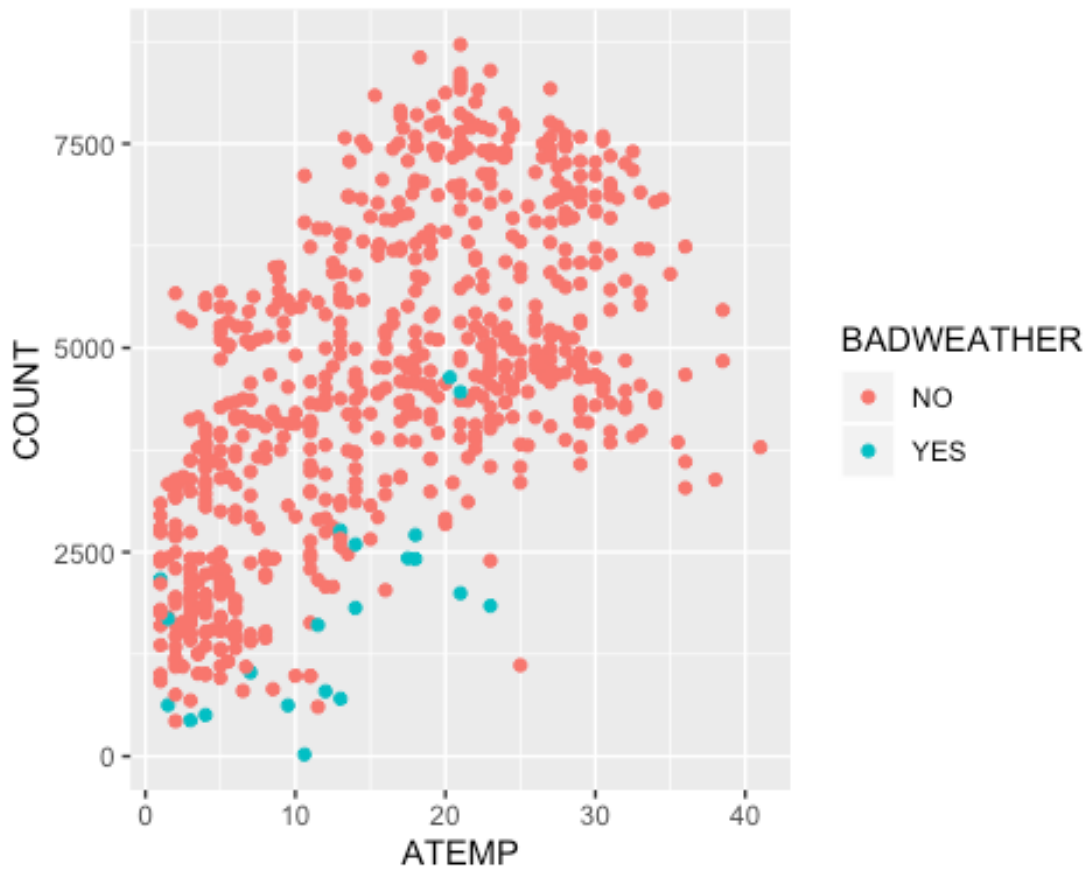
Question 3:

a. Adding BADWEATHER:

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

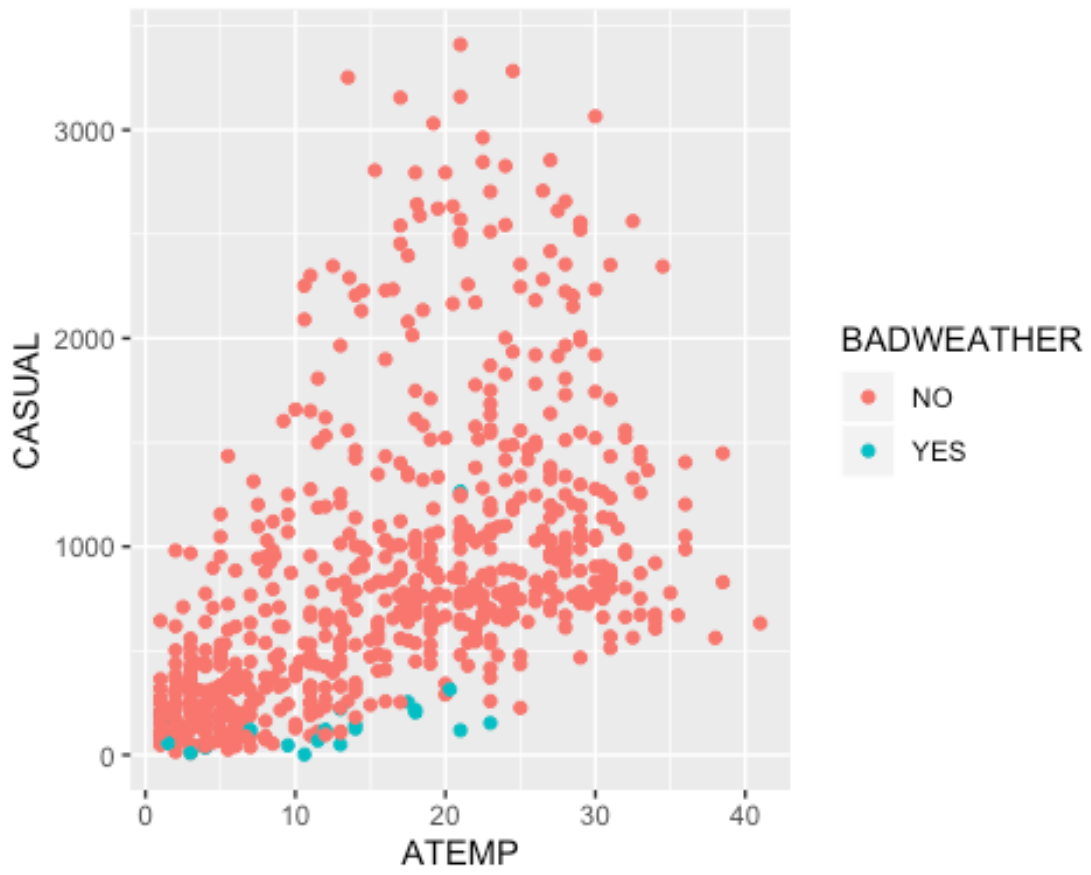
b. Scatterplot:

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

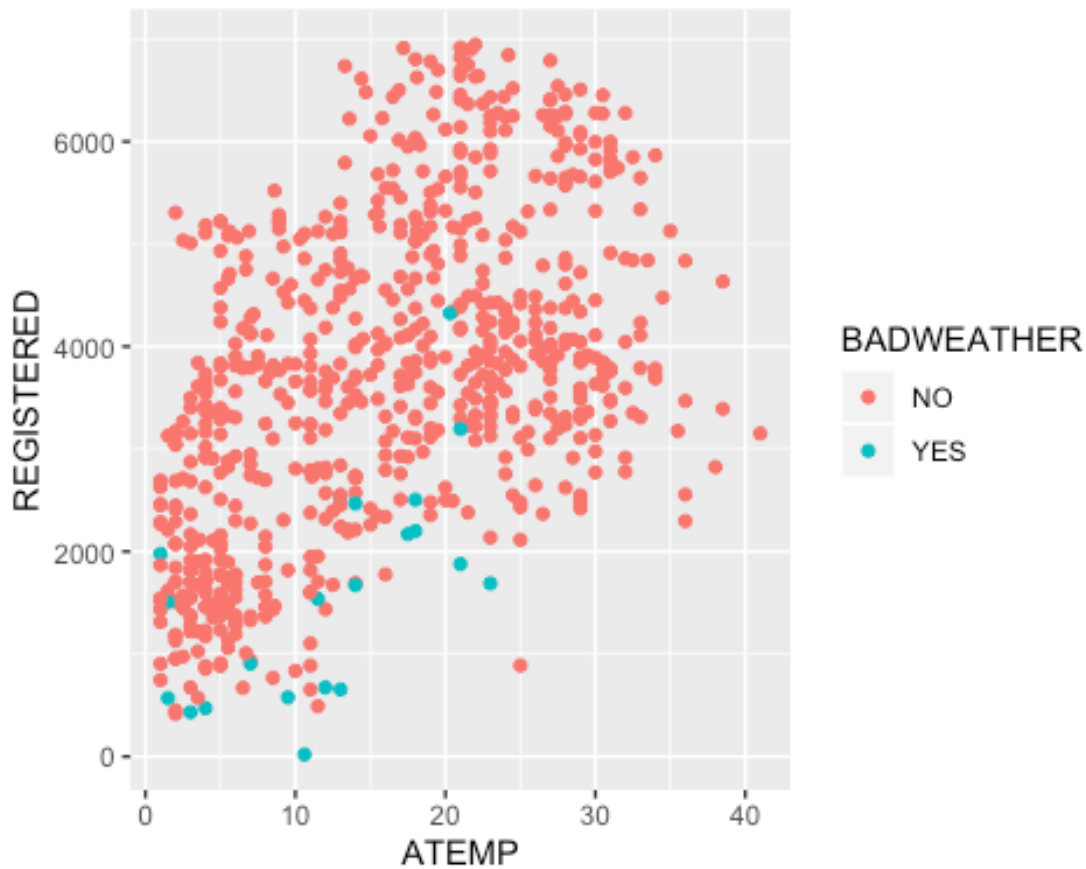


c: Scatterplots:

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

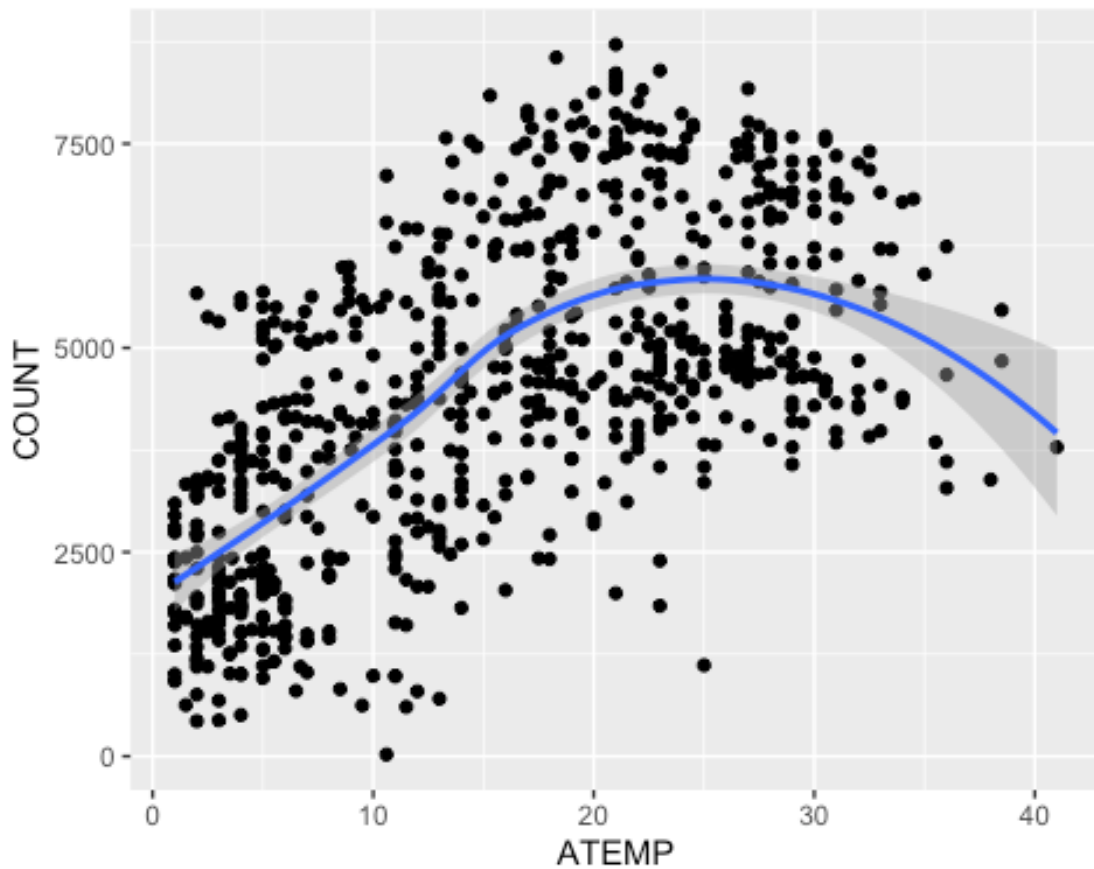



```
dfbOrg %>%
ggplot(mapping = aes(x=ATEMP,y=REGISTERED, color =BADWEATHER)) + geom_point()
```



Question 3 (iv):

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

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



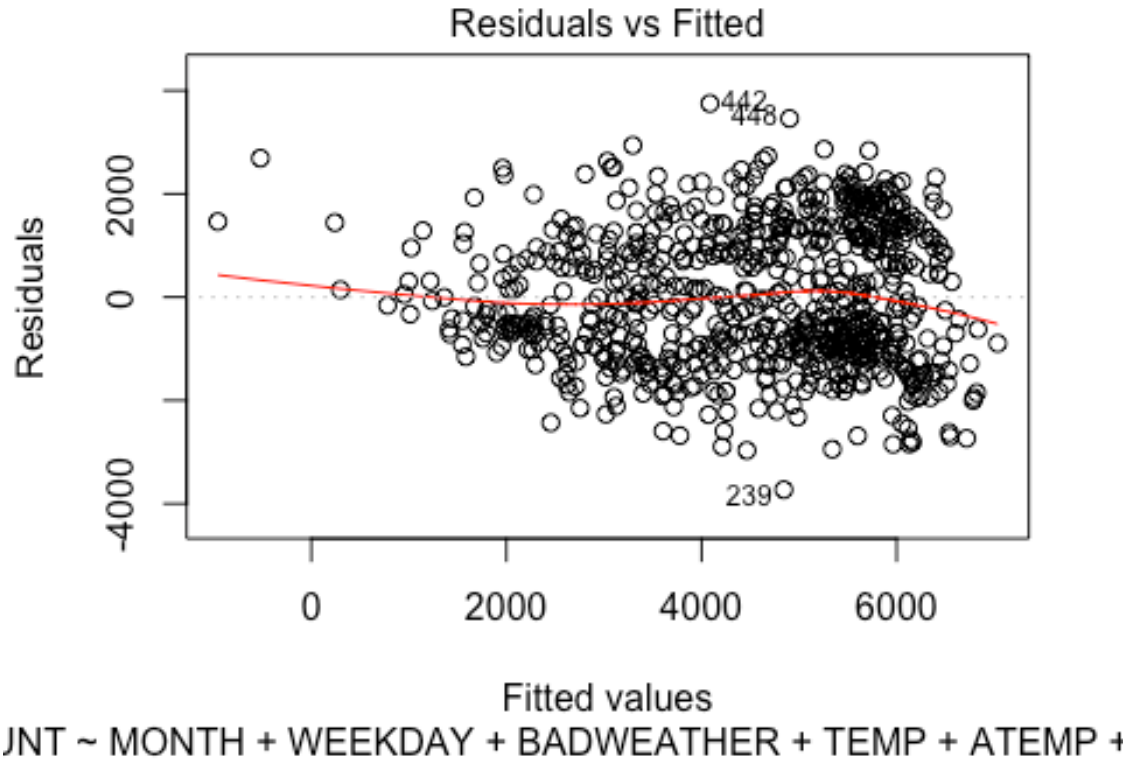
Question 4:

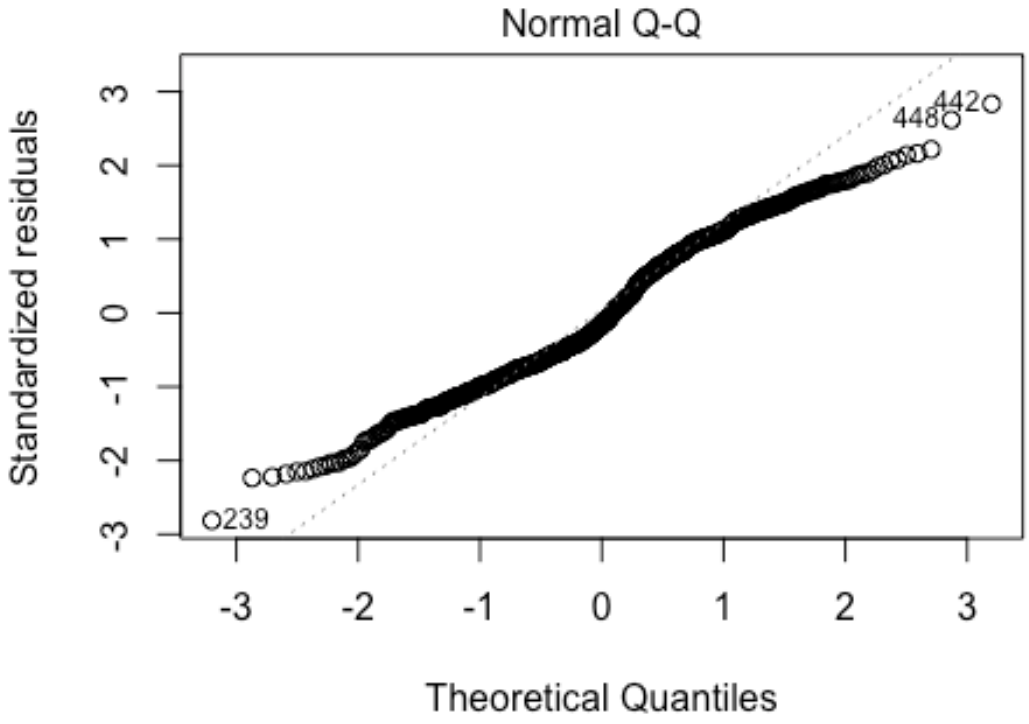
```
fitCount <-
  lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + TEMP + ATEMP +
HUMIDITY, data = dfbOrg)
summary(fitCount)

##
## 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 ***
## MONTHAugust     -209.660    291.004  -0.720  0.47147
## MONTHDecember    105.664    265.660   0.398  0.69094
## MONTHFebruary   -802.319    273.000  -2.939  0.00340 **
## MONTHJanuary    -858.334    293.371  -2.926  0.00355 **
```

```
## MONTHJuly       -676.644    312.956  -2.162  0.03094 *
## MONTHJune       -189.229    286.067  -0.661  0.50851
## MONTHMarch      -242.020    249.333  -0.971  0.33204
## MONTHMay         279.730    259.634   1.077  0.28166
## MONTHNovember    651.966    257.460   2.532  0.01154 *
## MONTHOctober    1072.312    246.970   4.342 1.62e-05 ***
## MONTHSeptember   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
```

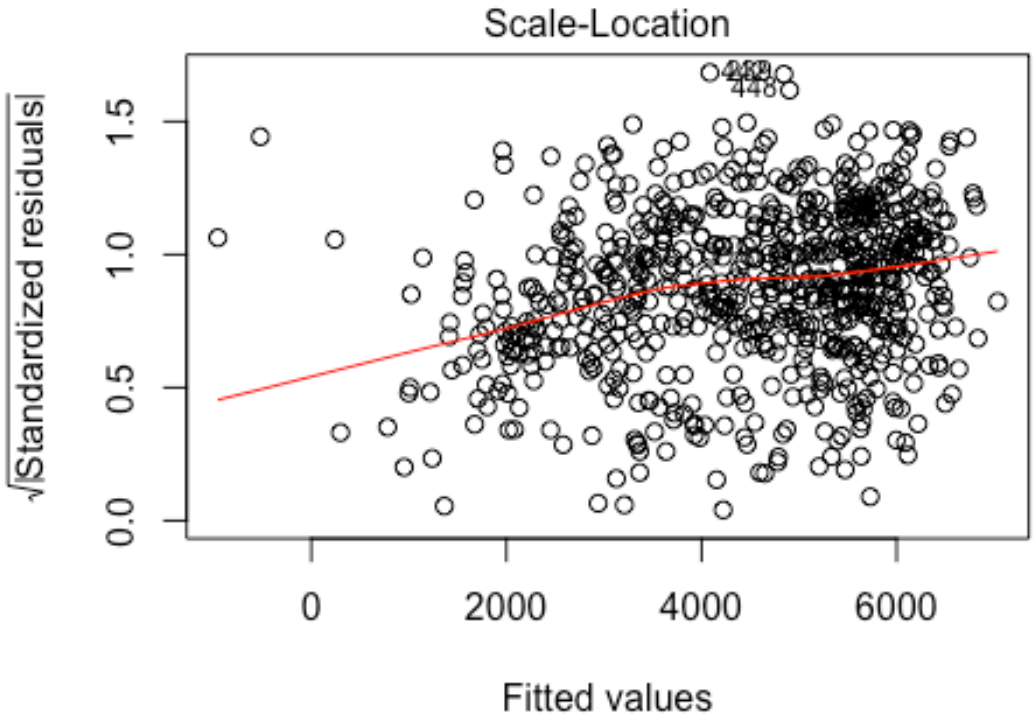
Question 5:

```r
plot(fitCount)
```

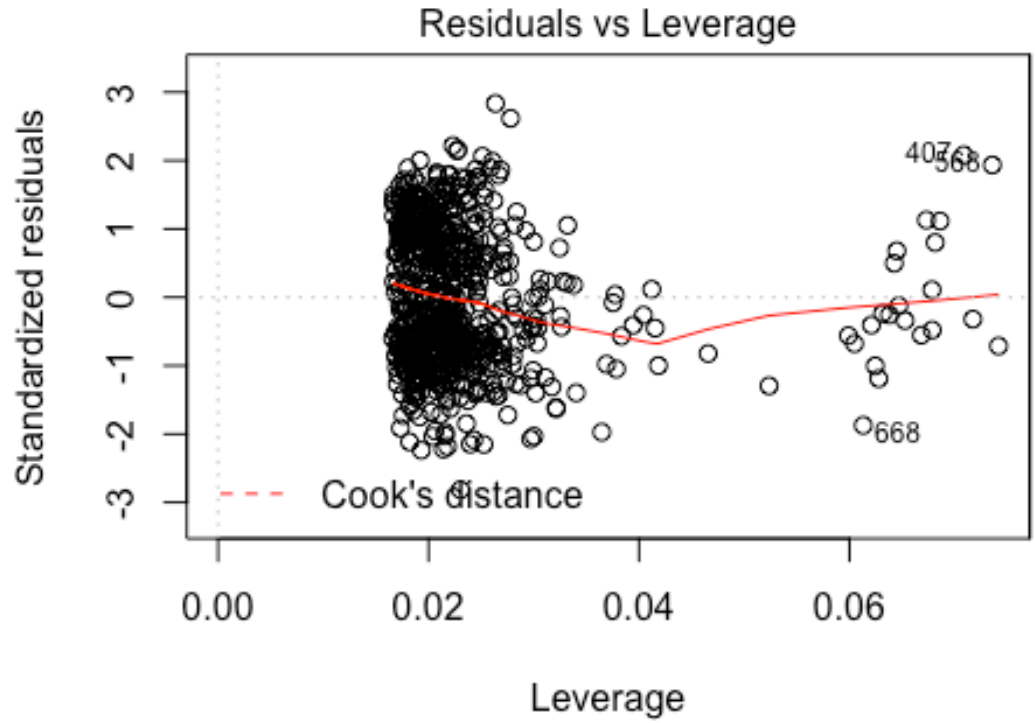

Residuals vs Fitted
Residuals
Fitted values
442
448
239
2000
0
-4000
0
2000
4000
6000
JNT ~ MONTH + WEEKDAY + BADWEATHER + TEMP + ATEMP +



JNT ~ MONTH + WEEKDAY + BADWEATHER + TEMP + ATEMP +



JNT ~ MONTH + WEEKDAY + BADWEATHER + TEMP + ATEMP +



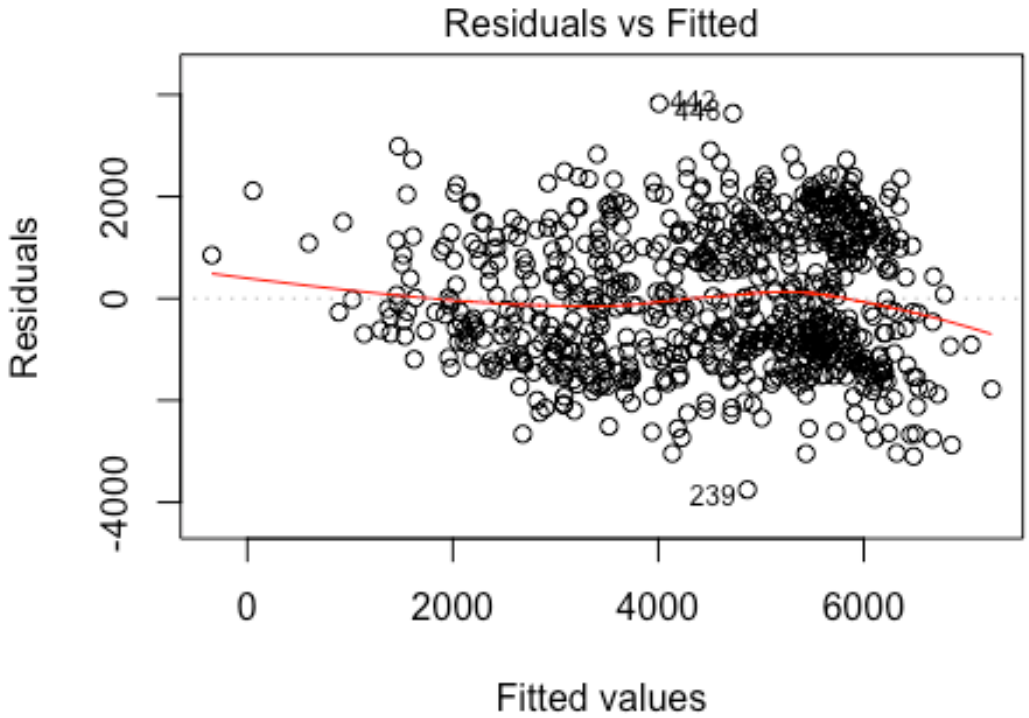
JNT ~ MONTH + WEEKDAY + BADWEATHER + TEMP + ATEMP +

Heteroskedasticity found. (Plot)

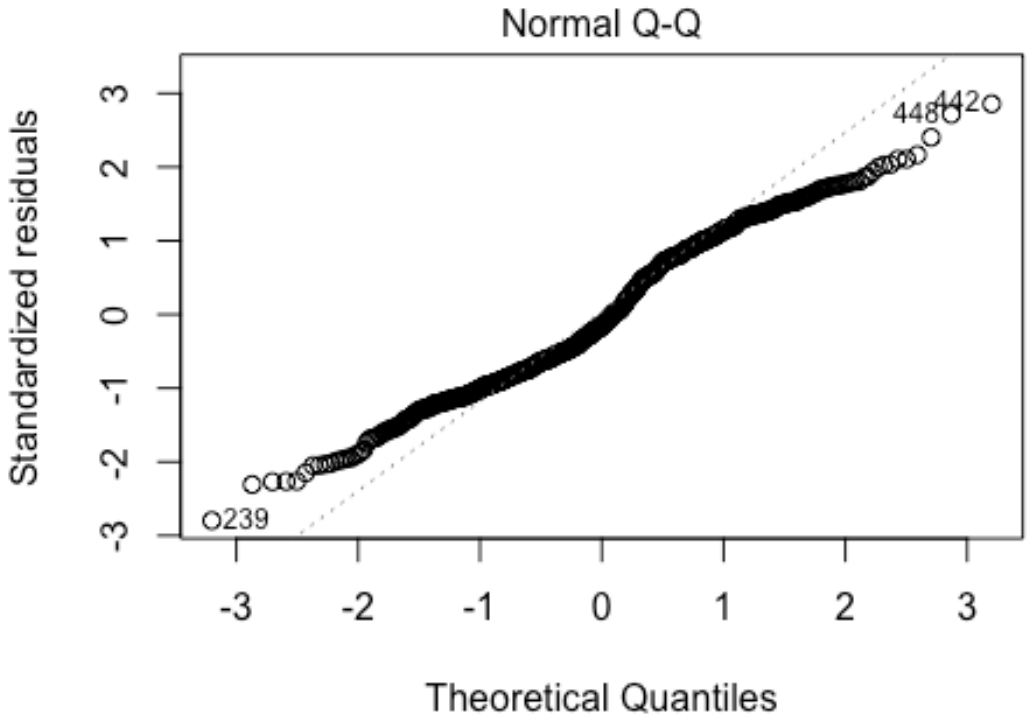
```
car::vif(fitCount)

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

##                 GVIF Df GVIF^(1/(2*Df))
## MONTH       8.480466 11        1.102049
## WEEKDAY     1.009743  1        1.004859
## BADWEATHER  1.137470  1        1.066522
## TEMP       55.856782  1        7.473739
## ATEMP      50.923158  1        7.136046
## HUMIDITY    1.275120  1        1.129212

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

```
##
## Call:
## lm(formula = COUNT ~ WEEKDAY + MONTH + ATEMP + BADWEATHER + HUMIDITY +
##     ATEMP * BADWEATHER, data = dfbOrg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3752.6 -1050.5  -207.3  1130.8  3828.3
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4483.55     317.27  14.132  < 2e-16 ***
## WEEKDAYYES             96.17     111.50   0.863 0.388675
## MONTHAugust           -82.69     293.19  -0.282 0.777995
## MONTHDecember          11.23     268.13   0.042 0.966613
## MONTHFebruary       -1010.97     272.06  -3.716 0.000218 ***
## MONTHJanuary        -1376.50     271.18  -5.076 4.92e-07 ***
## MONTHJuly            -595.77     316.38  -1.883 0.060098 .
## MONTHJune             -28.88     287.19  -0.101 0.919937
## MONTHMarch           -285.54     252.29  -1.132 0.258113
## MONTHMay              374.29     261.97   1.429 0.153518
## MONTHNovember         474.91     257.32   1.846 0.065367 .
## MONTHOctober         1043.06     250.10   4.171 3.41e-05 ***
## MONTHSeptember        855.88     270.07   3.169 0.001594 **
## ATEMP                 104.47      12.36   8.454  < 2e-16 ***
## BADWEATHERYES       -1409.43     623.80  -2.259 0.024157 *
## HUMIDITY              -25.54       3.67  -6.959 7.77e-12 ***
## ATEMP:BADWEATHERYES   -45.45      44.02  -1.032 0.302245
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1358 on 714 degrees of freedom
## Multiple R-squared:  0.5196, Adjusted R-squared:  0.5088
## F-statistic: 48.27 on 16 and 714 DF,  p-value: < 2.2e-16
```

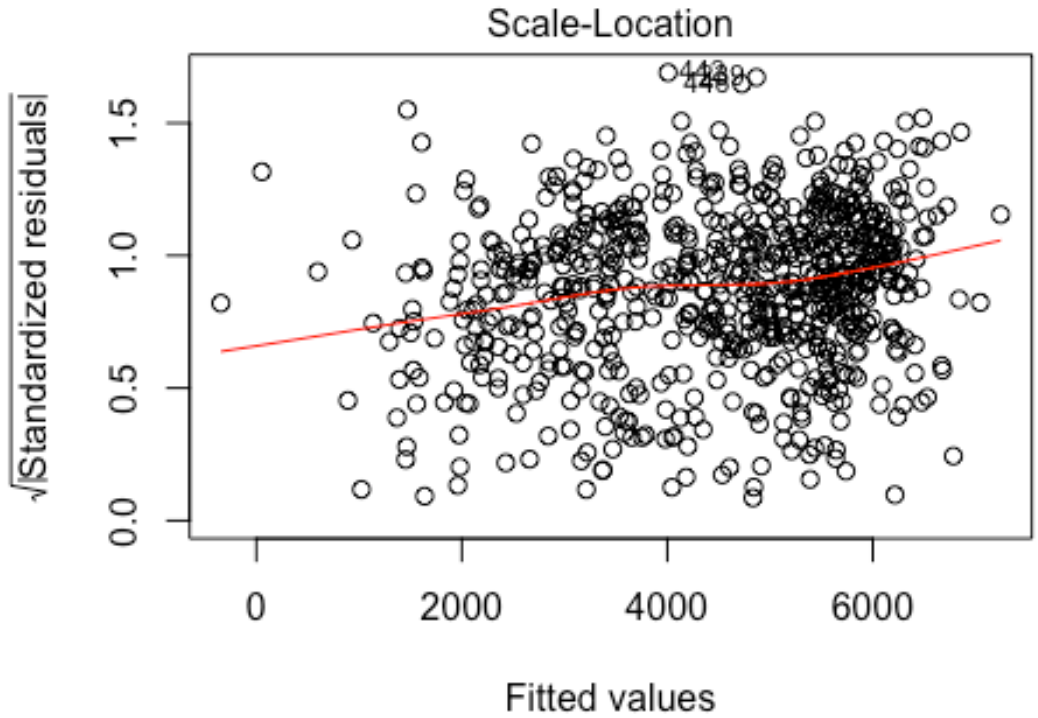
```r
plot(fitCountMod)
```



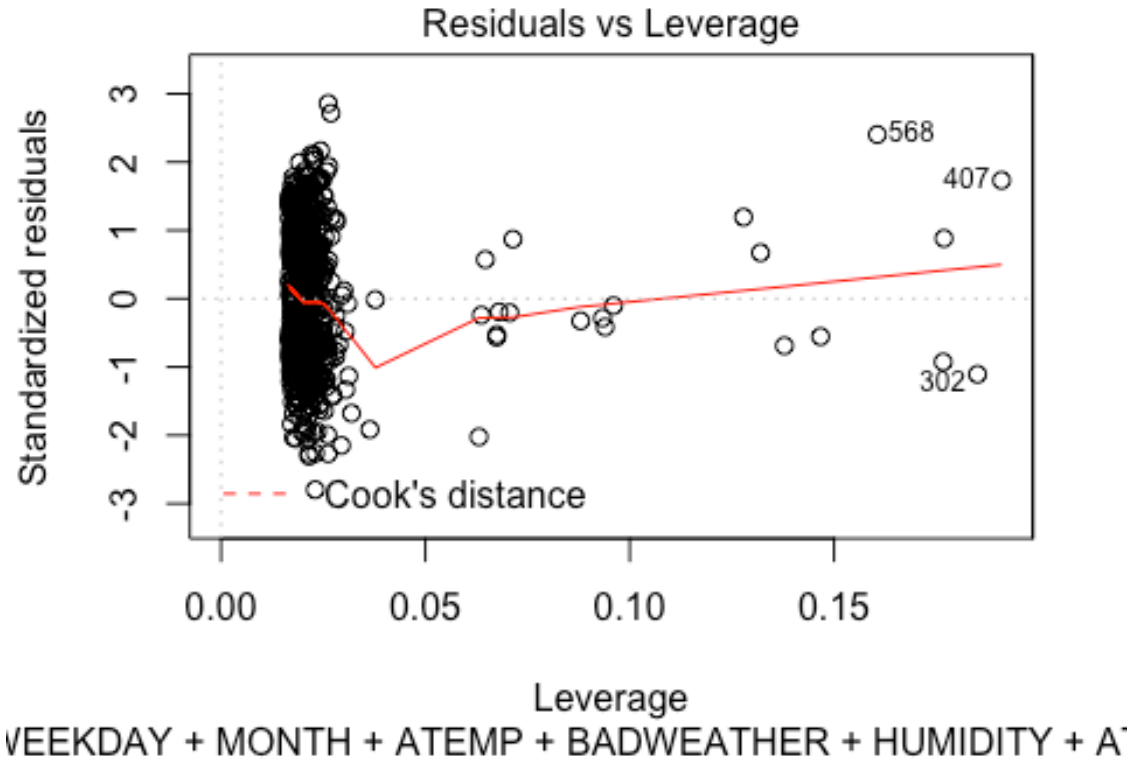
VEEKDAY + MONTH + ATEMP + BADWEATHER + HUMIDITY + A



VEEKDAY + MONTH + ATEMP + BADWEATHER + HUMIDITY + A



VEEKDAY + MONTH + ATEMP + BADWEATHER + HUMIDITY + A




```
car::vif(fitCountMod)

##                       GVIF Df GVIF^(1/(2*Df))
## WEEKDAY           1.009413  1        1.004696
## MONTH             6.465577 11        1.088543
## ATEMP             5.654853  1        2.377994
## BADWEATHER        4.305883  1        2.075062
## HUMIDITY          1.275947  1        1.129578
## ATEMP:BADWEATHER 4.182084  1        2.045014
```

Question 6: a.

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

##
## 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
```

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

```
##
## Call:
## lm(formula = COUNT ~ BADWEATHER + WEEKDAY + 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
```

Question 7: a

```
set.seed(333)
```

b

```
dfbTrain <- dfbOrg %>% sample_frac(0.8)
dfbTest <- setdiff(dfbOrg, dfbTrain)
```

c

```
fitOrg <-
  lm(formula = COUNT ~ WEEKDAY + MONTH + ATEMP + BADWEATHER + HUMIDITY +
```

```r
ATEMP*BADWEATHER, data = dfbTrain)
summary(fitOrg)
```

```
##
## Call:
## lm(formula = COUNT ~ WEEKDAY + MONTH + ATEMP + BADWEATHER + HUMIDITY +
##     ATEMP * BADWEATHER, data = dfbTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3718.1 -1074.1  -117.8  1123.3  3943.3
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4654.728    350.786  13.269  < 2e-16 ***
## WEEKDAYYES             94.213    124.591   0.756 0.449859
## MONTHAugust          -199.470    326.274  -0.611 0.541208
## MONTHDecember         -50.779    296.181  -0.171 0.863935
## MONTHFebruary       -1097.340    303.809  -3.612 0.000331 ***
## MONTHJanuary        -1421.648    303.948  -4.677 3.64e-06 ***
## MONTHJuly            -552.096    347.875  -1.587 0.113057
## MONTHJune             -89.148    311.164  -0.286 0.774600
## MONTHMarch           -493.648    280.421  -1.760 0.078881 .
## MONTHMay              325.171    288.878   1.126 0.260796
## MONTHNovember         443.467    291.517   1.521 0.128757
## MONTHOctober         1005.415    282.193   3.563 0.000398 ***
## MONTHSeptember        686.920    303.584   2.263 0.024031 *
## ATEMP                 103.895     13.757   7.552 1.72e-13 ***
## BADWEATHERYES       -1467.115    713.164  -2.057 0.040124 *
## HUMIDITY              -26.423      4.104  -6.439 2.56e-10 ***
## ATEMP:BADWEATHERYES   -55.031     49.863  -1.104 0.270219
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1354 on 568 degrees of freedom
## Multiple R-squared:  0.5229, Adjusted R-squared:  0.5094
## F-statistic: 38.91 on 16 and 568 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>
```

```
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.
```

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

```
##
## Call:
## lm(formula = COUNT ~ WEEKDAY + MONTH + ATEMP + BADWEATHER + HUMIDITY +
##     WINDSPEED + ATEMP * BADWEATHER, data = dfbTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3455.8 -1039.3  -139.6  1122.5  3610.2
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5960.61     419.99  14.192  < 2e-16 ***
## WEEKDAYYES             77.61     121.68   0.638  0.52384
## MONTHAugust          -307.85     319.19  -0.964  0.33523
## MONTHDecember        -273.38     292.13  -0.936  0.34977
## MONTHFebruary       -1206.62     297.32  -4.058 5.64e-05 ***
```

```
## MONTHJanuary        -1525.49     297.39  -5.130 3.99e-07 ***
## MONTHJuly            -764.16     341.93  -2.235  0.02582 *
## MONTHJune            -219.41     304.77  -0.720  0.47187
## MONTHMarch           -536.79     273.90  -1.960  0.05051 .
## MONTHMay              226.28     282.64   0.801  0.42370
## MONTHNovember         291.84     286.02   1.020  0.30800
## MONTHOctober          851.18     277.01   3.073  0.00222 **
## MONTHSeptember        537.02     297.71   1.804  0.07179 .
## ATEMP                 101.45      13.44   7.549 1.76e-13 ***
## BADWEATHERYES        -839.46     706.03  -1.189  0.23494
## HUMIDITY              -32.63       4.17  -7.826 2.49e-14 ***
## WINDSPEED             -58.58      10.90  -5.372 1.14e-07 ***
## ATEMP:BADWEATHERYES   -78.48      48.88  -1.606  0.10893
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1322 on 567 degrees of freedom
## Multiple R-squared:  0.546,  Adjusted R-squared:  0.5324
## F-statistic: 40.11 on 17 and 567 DF,  p-value: < 2.2e-16
```

```
resultsNew <- dfbTest %>%
 mutate(PREDICTEDCOUNT = predict(fitNew, dfbTest))
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
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(resultsNew, truth=COUNT, estimate=PREDICTEDCOUNT)
```

```
## # A tibble: 2 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       1341.
## 2 mae     standard       1150.
```

Question 8: Model 1:

```r
dfbOrgTs <- dfbOrg %>%
  mutate(YEAR = lubridate::year(DATE))

dfbTrainTs <- dfbOrgTs %>% filter( YEAR == 2011)
dfbTestTs <- setdiff(dfbOrgTs, dfbTrainTs)

fitOrgTs <-
  lm(formula = COUNT ~ WEEKDAY + MONTH + ATEMP + BADWEATHER + HUMIDITY +
ATEMP*BADWEATHER, data = dfbTrainTs)
summary(fitOrgTs)
```

```
##
## Call:
## lm(formula = COUNT ~ WEEKDAY + MONTH + ATEMP + BADWEATHER + HUMIDITY +
##     ATEMP * BADWEATHER, data = dfbTrainTs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2917.56  -315.57    49.21   369.71  2002.70
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3426.843    230.586  14.861  < 2e-16 ***
## WEEKDAYYES             18.645     74.969   0.249  0.80373
## MONTHAugust           567.273    199.265   2.847  0.00468 **
## MONTHDecember          57.794    178.457   0.324  0.74625
## MONTHFebruary       -1216.425    185.528  -6.557 1.99e-10 ***
## MONTHJanuary        -1608.827    184.444  -8.723  < 2e-16 ***
## MONTHJuly             476.451    222.032   2.146  0.03257 *
## MONTHJune             910.180    199.249   4.568 6.84e-06 ***
## MONTHMarch           -809.937    178.061  -4.549 7.47e-06 ***
## MONTHMay              959.356    173.367   5.534 6.18e-08 ***
## MONTHNovember         559.348    170.071   3.289  0.00111 **
## MONTHOctober         1000.676    165.628   6.042 3.92e-09 ***
## MONTHSeptember       1026.205    181.035   5.669 3.03e-08 ***
## ATEMP                  46.317      8.711   5.317 1.89e-07 ***
## BADWEATHERYES        -744.681    396.370  -1.879  0.06111 .
```

```
## HUMIDITY              -13.327      2.500  -5.332 1.75e-07 ***
## ATEMP:BADWEATHERYES   -53.091     27.320  -1.943  0.05279 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 643.3 on 348 degrees of freedom
## Multiple R-squared:  0.7919, Adjusted R-squared:  0.7823
## F-statistic: 82.75 on 16 and 348 DF,  p-value: < 2.2e-16
```

```
resultsTs <- dfbTestTs %>%
            mutate(PREDICTEDCOUNT = predict(fitOrg, dfbTestTs))
resultsTs
```

```
## # A tibble: 366 x 15
##    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 6 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>, BADWEATHER <chr>, YEAR <dbl>, PREDICTEDCOUNT <dbl>
```

```
performance <- metric_set(rmse, mae)
performance(resultsTs, truth=COUNT, estimate=PREDICTEDCOUNT)
```

```
## # A tibble: 2 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       1426.
## 2 mae     standard       1239.
```

Model 2:

```
dfbTrainTs1 <- dfbOrgTs %>%
  filter("2011-01-01" <= DATE & DATE < "2012-06-01")
dfbTestTs1 <- dplyr::setdiff(dfbOrgTs, dfbTrainTs1)

fitNewTs1 <- lm(COUNT ~ MONTH + WEEKDAY + BADWEATHER*ATEMP + HUMIDITY +
HOLIDAY, data = dfbTrainTs1)
summary(fitNewTs1)

##
## Call:
## lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER * ATEMP + HUMIDITY +
##     HOLIDAY, data = dfbTrainTs1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3019.3  -767.4     5.3   754.8  3616.2
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4053.546    282.332  14.357  < 2e-16 ***
## MONTHAugust         -1552.320    281.747  -5.510 5.77e-08 ***
## MONTHDecember        -503.380    259.796  -1.938 0.053237 .
## MONTHFebruary        -870.757    223.698  -3.893 0.000113 ***
## MONTHJanuary        -1223.118    224.387  -5.451 7.89e-08 ***
## MONTHJuly           -1947.321    308.884  -6.304 6.37e-10 ***
## MONTHJune           -1205.547    281.411  -4.284 2.20e-05 ***
## MONTHMarch           -246.938    203.630  -1.213 0.225825
## MONTHMay              236.644    214.061   1.105 0.269477
## MONTHNovember        -240.495    250.908  -0.959 0.338275
## MONTHOctober          -74.583    244.633  -0.305 0.760588
## MONTHSeptember       -566.657    265.896  -2.131 0.033567 *
## WEEKDAYYES             37.990    106.996   0.355 0.722696
## BADWEATHERYES        -775.461    559.661  -1.386 0.166491
## ATEMP                 119.559     11.735  10.188  < 2e-16 ***
## HUMIDITY              -21.030      3.326  -6.324 5.67e-10 ***
## HOLIDAYYES           -584.304    288.469  -2.026 0.043344 *
## BADWEATHERYES:ATEMP   -63.017     41.062  -1.535 0.125494
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1084 on 499 degrees of freedom
## Multiple R-squared:  0.5586, Adjusted R-squared:  0.5435
## F-statistic: 37.14 on 17 and 499 DF,  p-value: < 2.2e-16

resultsNewTs1 <- dfbTestTs1 %>%
            mutate(predictedCount = predict(fitNewTs1, dfbTestTs1))

resultsNewTs1
```

```
## # A tibble: 214 x 15
##    DATE       HOLIDAY WEEKDAY WEATHERSIT  TEMP ATEMP HUMIDITY WINDSPEED
CASUAL
##    <date>     <chr>   <chr>        <dbl> <dbl> <dbl>    <dbl>     <dbl>
<dbl>
##  1 2012-06-01 NO      YES              2  23    23       78          15
533
##  2 2012-06-02 NO      NO               1  20    20       49          13
2795
##  3 2012-06-03 NO      NO               1  22    21       45          12
2494
##  4 2012-06-04 NO      YES              1  21    21       46.5        20
1071
##  5 2012-06-05 NO      YES              2  18    18       56          13
968
##  6 2012-06-06 NO      YES              1  18    18       68           6
1027
##  7 2012-06-07 NO      YES              1  21    21       49.5        11
1038
##  8 2012-06-08 NO      YES              1  24.5  24.5     44.5        11
1488
##  9 2012-06-09 NO      NO               1  26    26.5     50.5        11
2708
## 10 2012-06-10 NO      NO               1  27    28       58           8
2224
## # … with 204 more rows, and 6 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>, BADWEATHER <chr>, YEAR <dbl>, predictedCount <dbl>
```

```
performanceB <- metric_set(rmse, mae)
performanceB(resultsNewTs1, truth = COUNT, estimate = predictedCount)
```

```
## # A tibble: 2 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2347.
## 2 mae     standard       2149.
```