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.

```r
#setwd("C:/") #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   1.0.0     ✔ stringr 1.4.0
## ✔ readr   1.3.1     ✔ forcats 0.4.0

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

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

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

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

## ── Conflicts
────────────────────────────────────────────────────────────────────────────
─────────── tidymodels_conflicts() ──
## ✖ scales::discard()   masks purrr::discard()
## ✖ dplyr::filter()     masks stats::filter()
## ✖ recipes::fixed()    masks stringr::fixed()
## ✖ dplyr::lag()        masks stats::lag()
## ✖ dials::margin()     masks ggplot2::margin()
```

```
## ✖ yardstick::spec()   masks readr::spec()
## ✖ recipes::step()     masks stats::step()
## ✖ 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")
```

```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
dfbOrg
```

```
## # A tibble: 731 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
##  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 1 more variable: REGISTERED <dbl>
```

```
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	▂▃▅▃▂

```
summary(dfbOrg)

##       DATE               HOLIDAY            WEEKDAY            WEATHERSIT
##  Min.   :2011-01-01   Length:731         Length:731         Min.   :1.000
##  1st Qu.:2011-07-02   Class :character   Class :character   1st Qu.:1.000
##  Median :2012-01-01   Mode  :character   Mode  :character   Median :1.000
##  Mean   :2012-01-01                                         Mean   :1.395
##  3rd Qu.:2012-07-01                                         3rd Qu.:2.000
##  Max.   :2012-12-31                                         Max.   :3.000
##       TEMP           ATEMP          HUMIDITY        WINDSPEED
##  Min.   : 1.00   Min.   : 1.00   Min.   : 17.00   Min.   : 0.00
##  1st Qu.: 8.00   1st Qu.: 6.60   1st Qu.: 51.00   1st Qu.: 9.00
##  Median :16.00   Median :16.00   Median : 62.00   Median :12.00
##  Mean   :15.87   Mean   :16.00   Mean   : 63.17   Mean   :12.82
##  3rd Qu.:23.15   3rd Qu.:23.95   3rd Qu.: 74.00   3rd Qu.:16.00
##  Max.   :34.00   Max.   :41.00   Max.   :100.00   Max.   :40.16
##      CASUAL         REGISTERED
##  Min.   :   2.0   Min.   :  20
##  1st Qu.: 315.5   1st Qu.:2497
##  Median : 713.0   Median :3662
##  Mean   : 848.2   Mean   :3656
##  3rd Qu.:1096.0   3rd Qu.:4776
##  Max.   :3410.0   Max.   :6946
```

#Data preparation #Create the additional variables: #Create the COUNT variable and add it to the data frame. #Extract MONTH from the DATE variable and add it to the data frame. This time, do NOT use lubridate. Use the base months() function instead.

```
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>
<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>
```

```
dfbOrg$MONTH <- months(dfbOrg$DATE)

dfbOrg
```

```
## # A tibble: 731 x 12
##    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 3 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>
```

#Scale the data (and save it as dfbStd): Start by standardizing the four variables, TEMP, ATEMP, HUMIDITY, WINDSPEED. If you don't remember what it means to standardize a variable, see the link. Surely, you don't need to do this manually!

```
dfbStd <- dfbOrg %>% mutate_at(c("TEMP" , "ATEMP", "HUMIDITY", "WINDSPEED"),
~scale(.) %>% as.vector())
dfbStd

## # A tibble: 731 x 12
##    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 -0.552 -0.517   1.15      0.756
331
##  2 2011-01-02 NO      NO               2 -0.779 -0.982   0.538     0.756
131
##  3 2011-01-03 NO      YES              1 -1.68  -1.24   -1.24      0.936
120
##  4 2011-01-04 NO      YES              1 -1.57  -1.40    0.0536   -0.689
108
##  5 2011-01-05 NO      YES              1 -1.51  -1.55   -1.34      0.0332
82
##  6 2011-01-06 NO      YES              1 -1.57  -1.45   -0.722    -1.23
88
##  7 2011-01-07 NO      YES              2 -1.68  -1.34   -1.01     -0.328
148
##  8 2011-01-08 NO      NO               2 -1.68  -1.14   -0.787     0.756
68
##  9 2011-01-09 NO      NO               1 -1.57  -0.775  -1.11      2.20
54
## 10 2011-01-10 NO      YES              1 -1.57  -1.03   -0.852     0.394
41
## # … with 721 more rows, and 3 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>
```

#Basic regression in R: In dfbStd, run a regression model fitAll using COUNT as the DV, and all the variables as independent variables. [Don't forget to use summary(fitAll)] #Does this

appear to be a good model? Why or why not? #According to your model, what is the effect of humidity on the total bike count in a formal interpretation? Does this finding align with your answer to Part (a)?

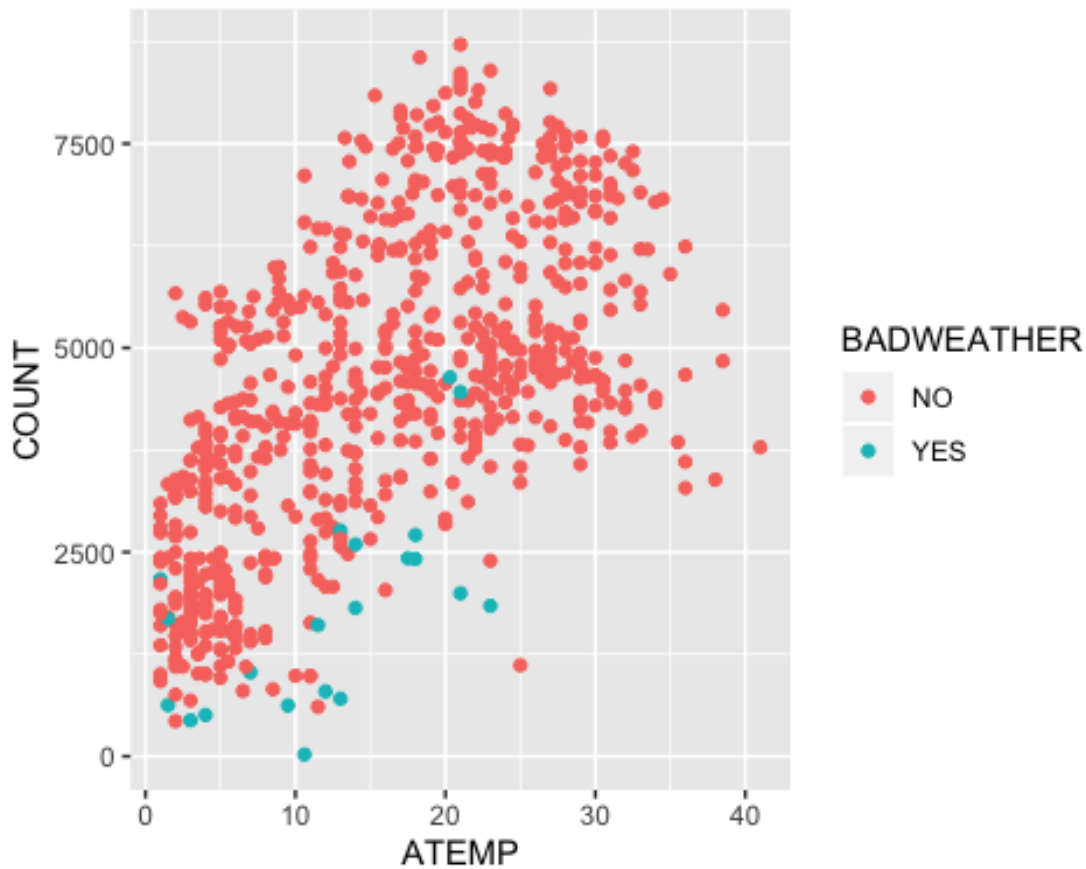
```
fitAll <- lm(formula = COUNT ~ ., data = dfbStd)
summary(fitAll)

## 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
```

#3.Working with data and exploratory analysis: #Add a new variable and call it BADWEATHER, which is "YES" if there is light or heavy rain or snow (if WEATHERSIT is 3 or 4), and "NO" otherwise (if WEATHERSIT is 1 or 2). You know what functions to use at this step.

```
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>
```

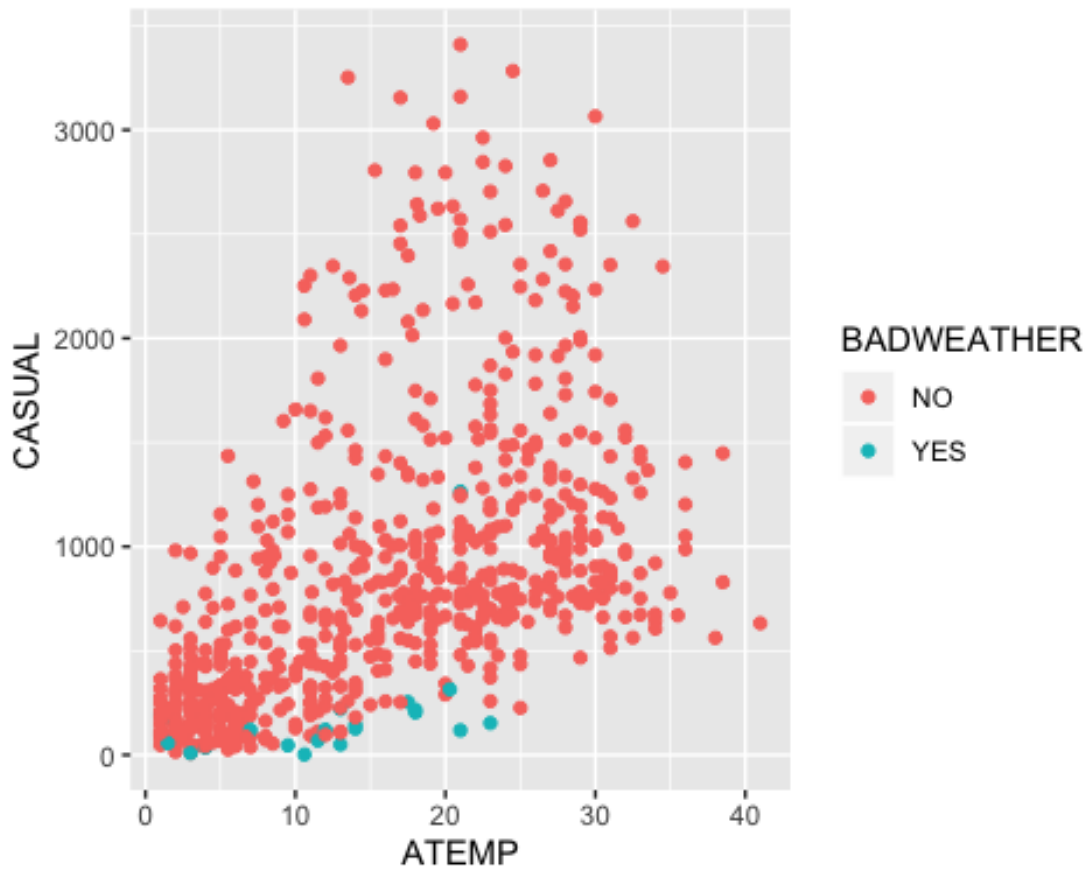
#Present a scatterplot of COUNT (y-axis) and ATEMP (x-axis). Use different colors or symbols to distinguish "bad weather" days. Briefly describe what you observe.

```
plot <- ggplot(data = dfbOrg, aes(x = ATEMP, y = COUNT, color= BADWEATHER)) +
geom_point()
ggplotly(plot)

plot
```

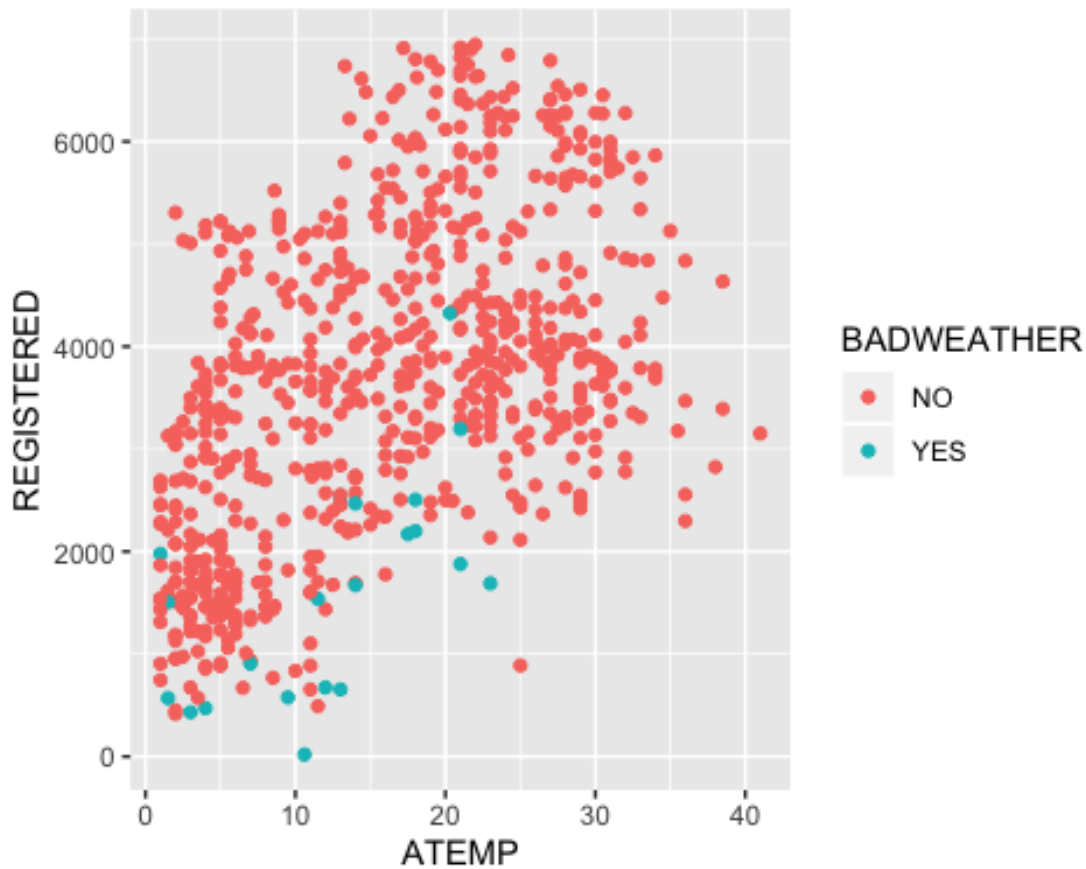


#Make two more scatterplots (and continue using the differentiated coloring for BADWEATHER) by keeping ATEMP on the x-axis and changing the variable on the y-axis: One plot for CASUAL and another for REGISTERED. #How is temperature associated with casual usage? Is that different from how it is associated with registered usage? #How is bad weather associated with casual usage? Is that different from how it is associated with registered usage? #Do your answers in (i) and (ii) make logical sense? Why or why not? #Keep ATEMP in the x-axis, but change the y-axis to COUNT. Remove the color variable and add a geom_smooth() without any parameters. How does the overall relationship between temperature and bike usage look? Does this remind you of Lab 2? Why do you think the effects are similar?

```
plot1 <- ggplot(data = dfbOrg, aes(x = ATEMP, y = CASUAL, color= BADWEATHER))
+ geom_point()
ggplotly(plot1)

plot1
```

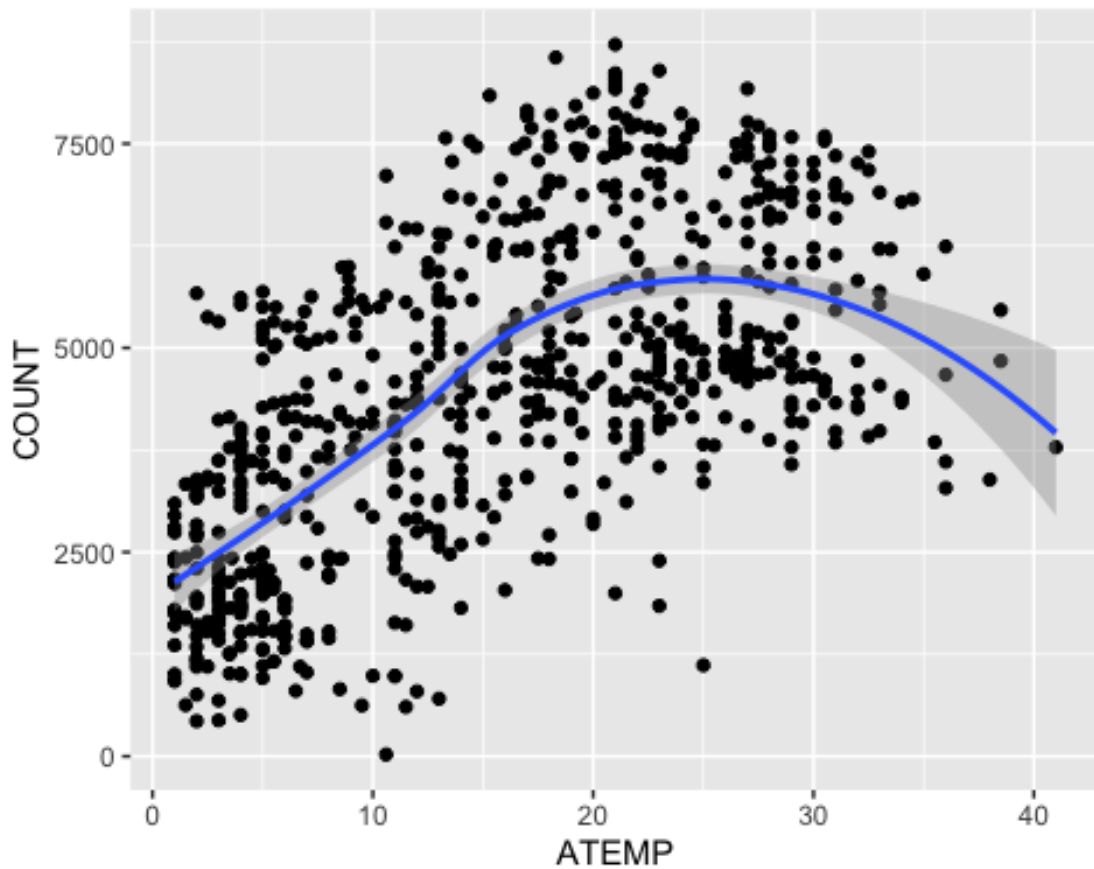



```
plot3 <- ggplot(data = dfbOrg, aes(x = ATEMP, y = REGISTERED, color=
BADWEATHER)) + geom_point()
ggplotly(plot3)

plot3
```



#Keep ATEMP in the x-axis, but change the y-axis to COUNT. Remove the color variable and add a geom_smooth() without any parameters. How does the overall relationship between temperature and bike usage look? Does this remind you of Lab 2? Why do you think the effects are similar?

```
plot4 <- ggplot(data = dfbOrg, aes(x = ATEMP, y = COUNT)) + geom_point() +
geom_smooth()
ggplotly(plot4)

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

plot4

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



#4.

More linear regression: Using dfbOrg, run another regression for COUNT using the variables MONTH, WEEKDAY, BADWEATHER, TEMP, ATEMP, and HUMIDITY. #What is the resulting adjusted R2? What does it mean?

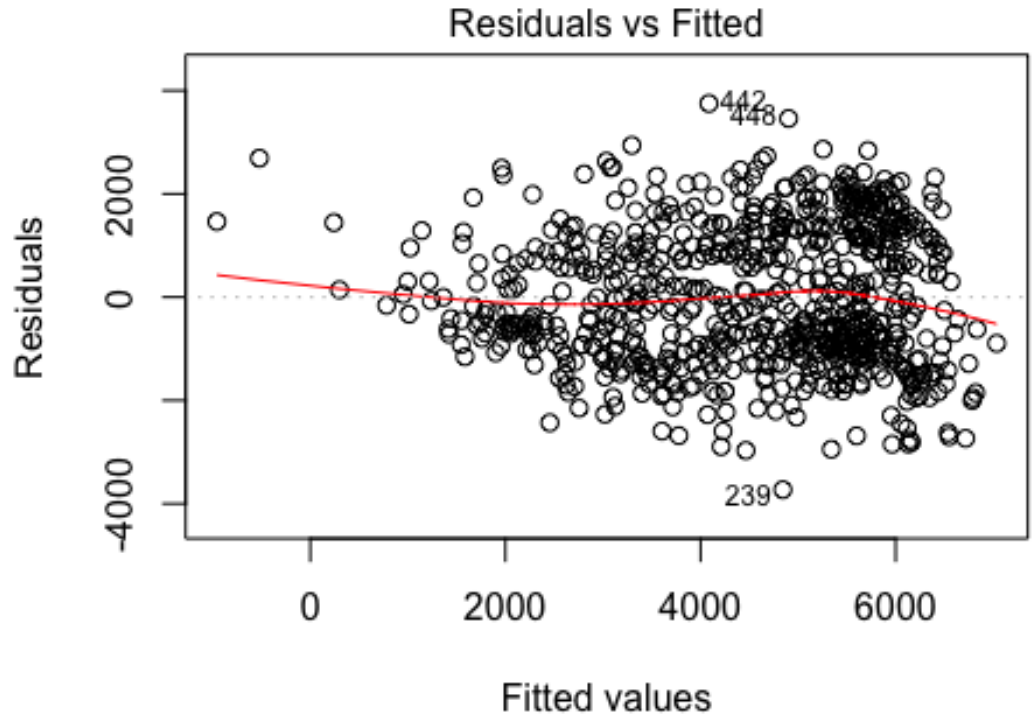
```
dfbReg <- lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + TEMP + ATEMP +
HUMIDITY, data = dfbOrg)
summary(dfbReg)
```

```
##
## 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 **
```

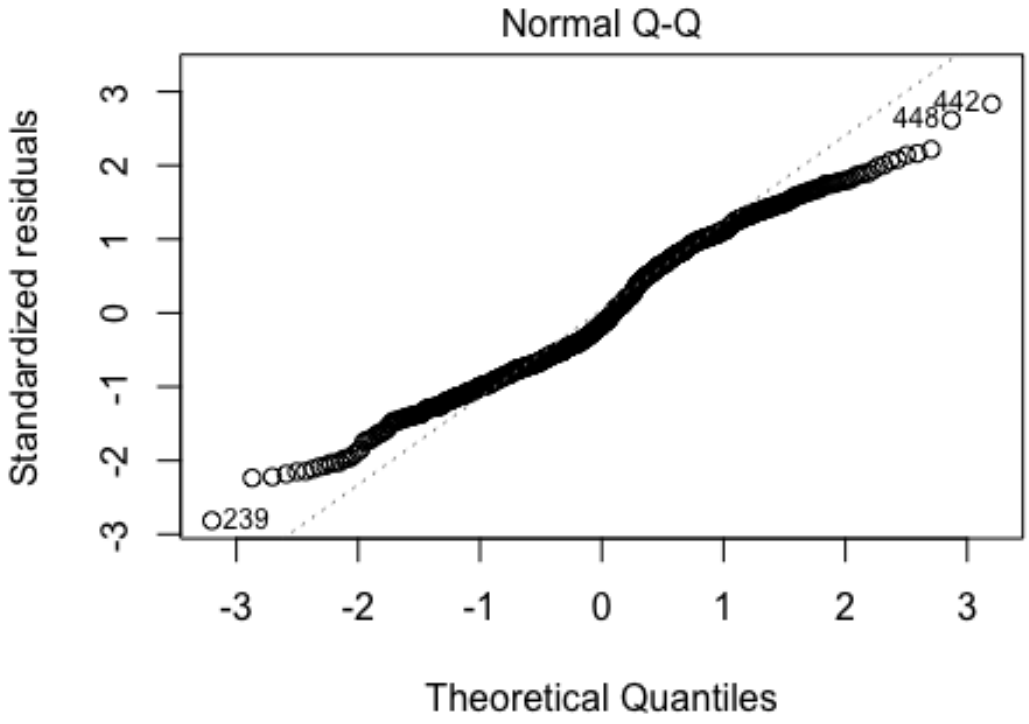
```
## MONTHJuly       -676.644    312.956  -2.162  0.03094 *
## MONTHJune       -189.229    286.067  -0.661  0.50851
## MONTHMarch      -242.020    249.333  -0.971  0.33204
## MONTHMay         279.730    259.634   1.077  0.28166
## MONTHNovember    651.966    257.460   2.532  0.01154 *
## MONTHOctober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

#5. Regression diagnostics: Run the regression diagnostics for the model developed in Q4. Discuss whether the model complies with the assumptions of multiple linear regression. If you think you can mitigate a violation, take action, and check the diagnostics again. Hint: The Q-Q plot and the other diagnostics from the plot() function look fine to me!

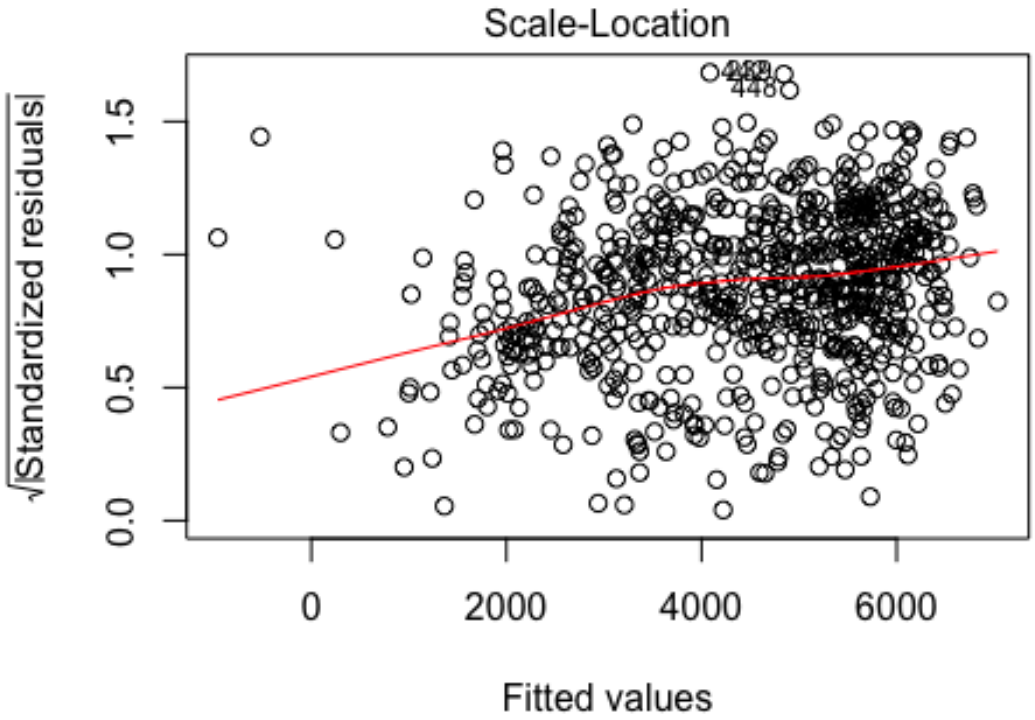
```r
plot(dfbReg)
```



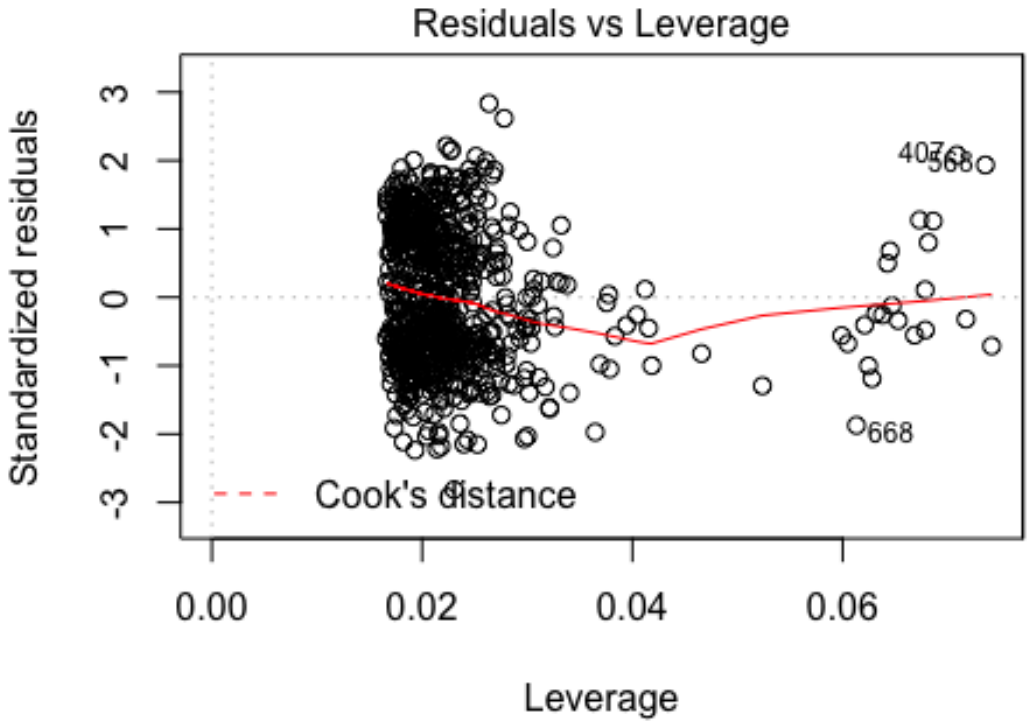
JNT ~ MONTH + WEEKDAY + BADWEATHER + TEMP + ATEMP +



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



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



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

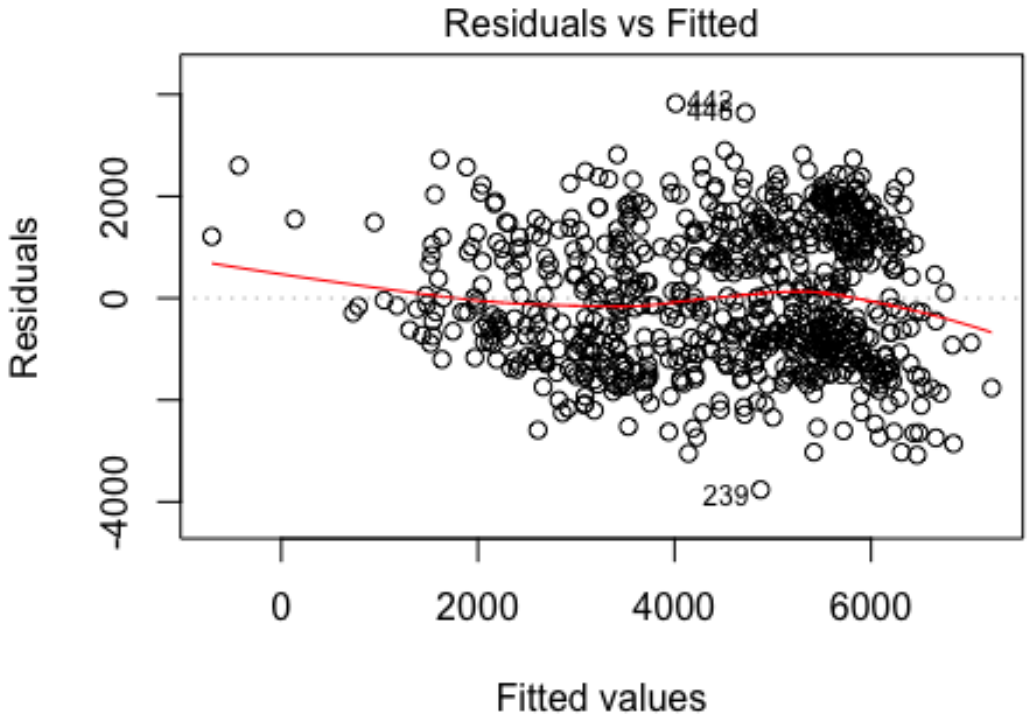
```
#install.packages("car")
#library(car)
#cor(dfbOrg[,c(5,6,7)])
#vif(dfbReg)
```

To mitigate the risk I removed TEMP

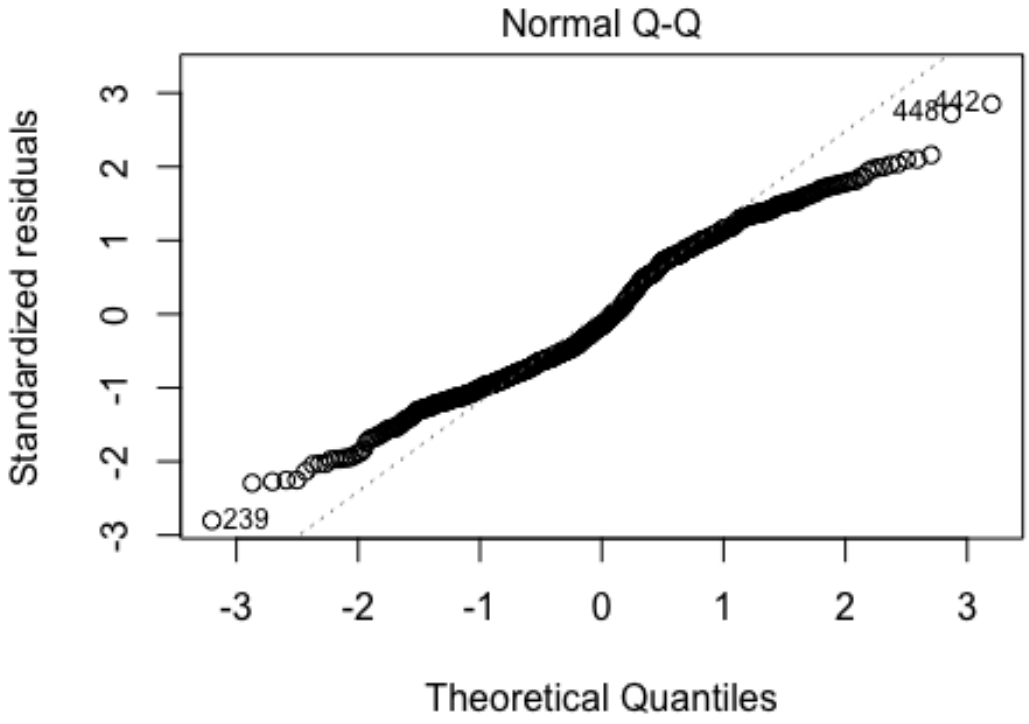
```
dfbReg1 <- lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP +
HUMIDITY, data = dfbOrg)
summary(dfbReg1)

##
## 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 ***
```

```
## MONTHAugust      -70.1865   292.9479  -0.240  0.81072
## MONTHDecember      0.6468   267.9485   0.002  0.99807
## MONTHFebruary  -1016.9096   272.0127  -3.738  0.00020 ***
## MONTHJanuary   -1386.5736   271.0121  -5.116 4.01e-07 ***
## MONTHJuly       -585.3680   316.2385  -1.851  0.06458 .
## MONTHJune        -17.4214   286.9867  -0.061  0.95161
## MONTHMarch      -285.6783   252.3046  -1.132  0.25790
## MONTHMay         378.1598   261.9562   1.444  0.14929
## MONTHNovember    462.3246   257.0456   1.799  0.07250 .
## MONTHOctober    1033.8276   249.9540   4.136 3.95e-05 ***
## MONTHSeptember   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
```

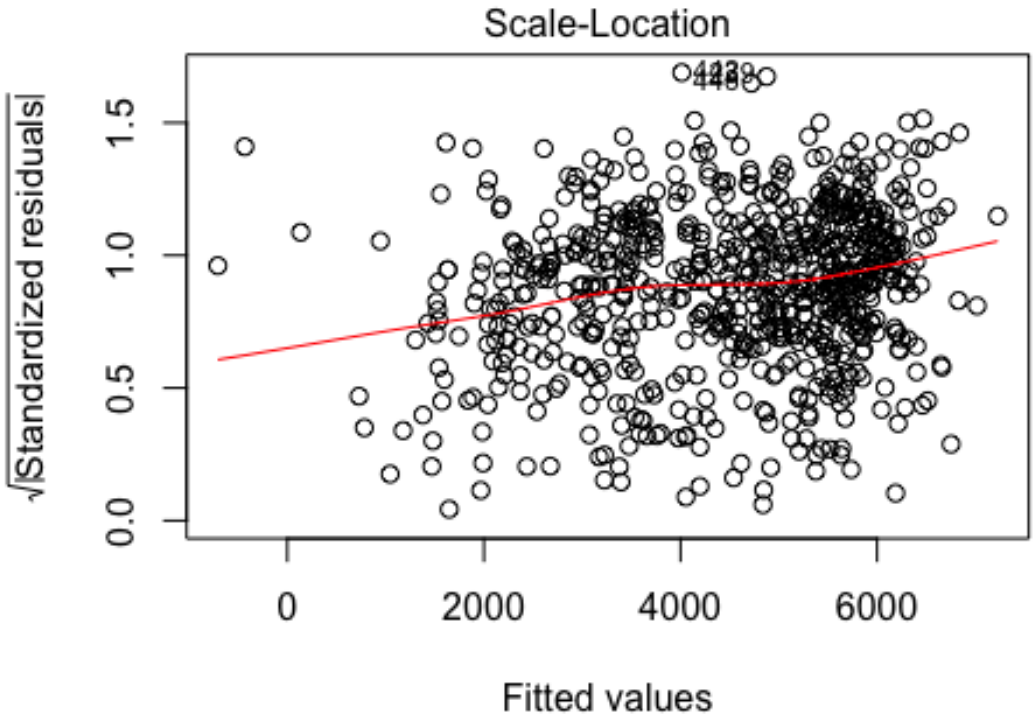
```r
plot(dfbReg1)
```



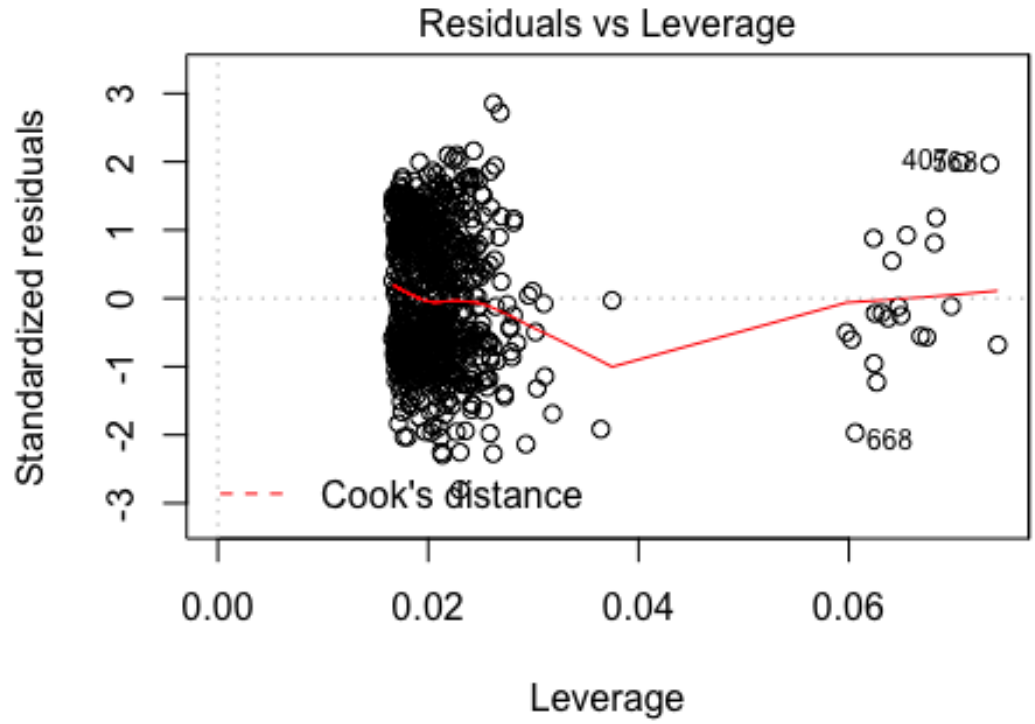
COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HU



COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HU



COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HU



COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HU

#6.Even more regression: Run a simple linear regression to determine the effect of bad weather on COUNT when none of the other variables is included in the model.

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

##
## 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
```

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

```
##
## 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
```

#7.Predictive analytics: Follow the steps below to build two predictive models. Which model is a better choice for predictive analytics purposes? Why? Does your conclusion remain the same for explanatory analytics purposes? Please copy and paste the predictive and explanatory performance levels of both models into your response. #Set the seed to 333 (Always set the seed and split your data in the same chunk!). #Split your data into two: 80% for the training set, and 20% for the test set #Call the training set dfbTrain and the test set dfbTest #Build two different models, calculate, and compare performance. #The first model will include the variables in Q4 with any adjustments you may have made during the diagnostics tests in Q5 (call this one fitOrg). The second model will add WINDSPEED to this model -Call it fitNew.

```
library(modelr)
```

```
##
## Attaching package: 'modelr'

## The following objects are masked from 'package:yardstick':
##
##     mae, mape, rmse

## The following object is masked from 'package:broom':
##
##     bootstrap
```

```r
detach('package:modelr', unload=TRUE)
```

```
## Warning: 'modelr' namespace cannot be unloaded:
##   namespace 'modelr' is imported by 'tidyverse' so cannot be unloaded
```

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

#Model1
fitOrg <-lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP +
HUMIDITY, data = dfbOrg)
fitOrg
```

```
##
## Call:
## lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HUMIDITY,
##     data = dfbOrg)
##
## Coefficients:
##    (Intercept)     MONTHAugust   MONTHDecember   MONTHFebruary
MONTHJanuary
##      4503.4952        -70.1865          0.6468      -1016.9096      -
1386.5736
##      MONTHJuly       MONTHJune      MONTHMarch        MONTHMay
MONTHNovember
##      -585.3680        -17.4214       -285.6783        378.1598
462.3246
##   MONTHOctober  MONTHSeptember      WEEKDAYYES   BADWEATHERYES
ATEMP
##      1033.8276        841.6233         91.4446      -1961.8521
103.1721
##       HUMIDITY
##       -25.4375
```

```r
summary(fitOrg)
```

```
##
## 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 ***
## MONTHAugust      -70.1865   292.9479  -0.240  0.81072
## MONTHDecember      0.6468   267.9485   0.002  0.99807
```

```
## MONTHFebruary  -1016.9096   272.0127  -3.738  0.00020 ***
## MONTHJanuary   -1386.5736   271.0121  -5.116 4.01e-07 ***
## MONTHJuly       -585.3680   316.2385  -1.851  0.06458 .
## MONTHJune        -17.4214   286.9867  -0.061  0.95161
## MONTHMarch      -285.6783   252.3046  -1.132  0.25790
## MONTHMay         378.1598   261.9562   1.444  0.14929
## MONTHNovember    462.3246   257.0456   1.799  0.07250 .
## MONTHOctober    1033.8276   249.9540   4.136 3.95e-05 ***
## MONTHSeptember   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
```

```
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(data= resultsOrg, truth= COUNT, estimate= predictedCOUNT)
```

```
## # A tibble: 2 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       1362.
## 2 mae     standard       1152.
```

```r
#Model2
fitNew <- lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP +
HUMIDITY + WINDSPEED , data = dfbOrg)
fitNew
```

```
##
## Call:
## lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HUMIDITY +
##     WINDSPEED, data = dfbOrg)
##
## Coefficients:
##    (Intercept)     MONTHAugust   MONTHDecember   MONTHFebruary
MONTHJanuary
##        5877.66         -203.44         -218.08        -1146.67        -
1496.01
##      MONTHJuly       MONTHJune      MONTHMarch        MONTHMay
MONTHNovember
##        -821.87         -178.97         -325.37          263.63
292.05
##   MONTHOctober  MONTHSeptember      WEEKDAYYES   BADWEATHERYES
ATEMP
##         869.24          668.69           76.98        -1509.72
100.87
##       HUMIDITY       WINDSPEED
##         -32.23          -60.32
```

```r
summary(fitNew)
```

```
##
## Call:
## lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HUMIDITY +
##     WINDSPEED, data = dfbOrg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3303.3 -1032.9  -161.9  1142.4  3473.0
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     5877.66     380.06  15.465  < 2e-16 ***
## MONTHAugust     -203.44     286.38  -0.710 0.477687
## MONTHDecember   -218.08     263.57  -0.827 0.408276
## MONTHFebruary  -1146.67     265.99  -4.311 1.85e-05 ***
## MONTHJanuary   -1496.02     264.77  -5.650 2.32e-08 ***
## MONTHJuly       -821.87     310.62  -2.646 0.008327 **
## MONTHJune       -178.97     280.97  -0.637 0.524351
## MONTHMarch      -325.37     246.03  -1.322 0.186439
## MONTHMay         263.63     256.03   1.030 0.303499
## MONTHNovember    292.05     252.07   1.159 0.247001
## MONTHOctober     869.24     245.10   3.546 0.000416 ***
## MONTHSeptember   668.69     264.41   2.529 0.011653 *
## WEEKDAYYES        76.98     108.62   0.709 0.478764
## BADWEATHERYES  -1509.72     320.95  -4.704 3.06e-06 ***
## ATEMP            100.87      11.99   8.413  < 2e-16 ***
## HUMIDITY         -32.23       3.74  -8.617  < 2e-16 ***
## WINDSPEED        -60.32       9.73  -6.199 9.61e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1324 on 714 degrees of freedom
## Multiple R-squared:  0.5435, Adjusted R-squared:  0.5332
## F-statistic: 53.12 on 16 and 714 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>
```

```
performance(data= resultsNew, truth= COUNT, estimate= predictedCOUNT)
```

```
## # A tibble: 2 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       1311.
## 2 mae     standard       1124.
```

#8.More predictive analytics: In this final question, experiment with the time component. In a way, you will almost treat the data as a time series. We will cover time series data later, so this is just a little experiment. Taking into account date, you can't split your data randomly (well, evidently, you would not want to use future data to predict the past). Instead, you have to split your data by time. Start with dfbOrg and use the variables you used in fitOrg from Q7c. Split your data into training using the year "2011" data, and test using the "2012" data. Has the performance improved over the random split that assumed cross-sectional data (which you did in the previous questions)? Why do you think so? Split again by assigning 1.5 years of data starting from January 1st, 2011 to the training set and the remaining six months of data (the last six months) to the test set. Does this look any better? Discuss your findings.

```
library(lubridate)
```

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

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

```
dfbOrg2011 <- dfbOrg %>% filter(year(DATE) == 2011)
dfbOrg2011
```

```
## # A tibble: 365 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 355 more rows, and 4 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>, BADWEATHER <chr>
```

```
dfbOrg2012 <- dfbOrg %>% filter(year(DATE) == 2012)
dfbOrg2012
```

```
## # 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>
```

```r
set.seed(333)
dfbTrainTime <- dfbOrg2011 %>% sample_frac(0.8)
dfbTestTime <- dplyr::setdiff(dfbOrg2012, dfbTrainTime)

fitOrg2012 <-lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP +
HUMIDITY, data = dfbOrg2011)
summary(fitOrg2012)
```

```
##
## Call:
## lm(formula = COUNT ~ MONTH + WEEKDAY + BADWEATHER + ATEMP + HUMIDITY,
##     data = dfbOrg2011)
##
## 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 ***
## MONTHAugust      595.712    199.516   2.986  0.00303 **
## MONTHDecember     36.819    178.838   0.206  0.83701
## MONTHFebruary  -1233.561    186.054  -6.630 1.28e-10 ***
## MONTHJanuary   -1613.793    185.158  -8.716  < 2e-16 ***
## MONTHJuly        514.856    222.028   2.319  0.02098 *
## MONTHJune        938.944    199.487   4.707 3.63e-06 ***
## MONTHMarch      -800.726    178.705  -4.481 1.01e-05 ***
## MONTHMay         969.720    173.973   5.574 4.99e-08 ***
## MONTHNovember    548.346    170.652   3.213  0.00143 **
## MONTHOctober     999.192    166.284   6.009 4.70e-09 ***
## MONTHSeptember   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
```

```r
resultsOrg2012 <- dfbTestTime %>%
            mutate(predictedCOUNT = predict(fitOrg2012, dfbTestTime))
resultsOrg2012
```

```
## # A tibble: 366 x 14
##    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 5 more variables: REGISTERED <dbl>, COUNT
<dbl>,
## #   MONTH <chr>, BADWEATHER <chr>, predictedCOUNT <dbl>
```

```r
performance(data= resultsOrg2012, truth= COUNT, estimate= predictedCOUNT)
```

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