City Bike Data Analysis

Robin Lee

December 18, 2015

Introduction

I am studying how weather (temperature and preciptation) affects the number of trips taken in City Bike. Riders might be less inclined to ride bicycles when there's rain or snow. Also, a sudden drop

Hypothesis -
1. Higher temperature is associated with more trips taken at a given day.
2. Rain or snow would reduce the number of trips
3. Short-term riders are more subject to the influence of weather.

Indendent variables include mean temperature, precipitation, membership type and weekend indicator.

Dependent variable is Daily total of trips.

Description of Data Set and Variables

I obtain my data from two sources. One source gives me city bike trip data. The other gives me daily weather data. City bike trip data is obtained from City Bike System Data (https://www.citibikenyc.com/system-data). The website contains trip level data for each month from July 2013 to December 2015. I analyzed the data from Jan 2014 to December 2014. Because the dataset is trip-level, I then aggregate the number of trips, total time of trips by membership for each day.

Below is the code that shows how I obtain the city bike data.

```
library(rvest)
base.url <- html("https://www.citibikenyc.com/system-data")
data <- base.url %>%
  html_nodes("#system-data li a")

# choose again, specify hyperlink
links <- base.url %>%
  html_nodes("#system-data li a") %>%
  html_attr("href")

# i want trip data
trip_links <- links[1:27]

# Step 2: Download zip files
# the last link is google drive. that might be tricky
for(i in 1:length(trip_links)){
  time = substr(basename(trip_links[i]),1,6)
  download.file(trip_links[i],paste0("bikedata/",time,".zip"), method = "libcurl")
}
```

This is a sample dataset for trip level data

```
load("preview1.RData")
triplevel
```

```
##   tripduration           starttime            stoptime start station id
## 1          471 2014-01-01 00:00:06 2014-01-01 00:07:57             2009
## 2         1494 2014-01-01 00:00:38 2014-01-01 00:25:32              536
## 3          464 2014-01-01 00:03:59 2014-01-01 00:11:43              228
## 4          373 2014-01-01 00:05:15 2014-01-01 00:11:28              519
## 5          660 2014-01-01 00:05:18 2014-01-01 00:16:18               83
## 6          330 2014-01-01 00:05:55 2014-01-01 00:11:25              422
##              start station name start station latitude
## 1      Catherine St & Monroe St               40.71117
## 2                1 Ave & E 30 St               40.74144
## 3                E 48 St & 3 Ave               40.75460
## 4             Pershing Square N                40.75188
## 5 Atlantic Ave & Fort Greene Pl                40.68383
## 6               W 59 St & 10 Ave               40.77051
##   start station longitude end station id         end station name
## 1               -73.99683            263 Elizabeth St & Hester St
## 2               -73.97536            259  South St & Whitehall St
## 3               -73.97188           2022      E 59 St & Sutton Pl
## 4               -73.97770            526          E 33 St & 5 Ave
## 5               -73.97632            436 Hancock St & Bedford Ave
## 6               -73.98804            526          E 33 St & 5 Ave
##   end station latitude end station longitude bikeid   usertype birth year
## 1             40.71729             -73.99638  16379 Subscriber       1986
## 2             40.70122             -74.01234  15611 Subscriber       1963
## 3             40.75849             -73.95921  16613 Subscriber       1991
## 4             40.74766             -73.98491  15938 Subscriber       1989
## 5             40.68217             -73.95399  19830 Subscriber       1990
## 6             40.74766             -73.98491  17343 Subscriber       1987
##   gender
## 1      1
## 2      1
## 3      1
## 4      1
## 5      1
## 6      1
```

Below is the code showing how I aggregate it into daily level

```
library(readr)
library(dplyr)

aggrdaily <- function(x){
  x$date <- format(x$starttime, "%Y%m%d")
  summary <- x %>% group_by(date, usertype) %>%
    summarize(trips = n(), totaltime=sum(tripduration))
  return(summary)
}

temp = list.files("bikedata/",pattern="*.csv")
myfiles = lapply(paste("bikedata/",temp, sep=""), read_csv)
```

```
for(i in 9:12){
myfiles[[i]]$starttime <- as.POSIXct(myfiles[[i]]$starttime, format="%m/%d/%Y %H:%M:%S")
}

summary <- list()
for(i in 1:12){
  summary[[i]] <- aggrdaily(myfiles[[i]])
}

summary2014 <- data.frame()
for(i in 1:12){
  summary2014 <- rbind(summary2014,summary[[i]])
}
```

This is a sample dataset for daily aggregate data. I will then combine this with weather data.

```
load("summary2014.RData")
head(summary2014)
```

```
##         date   usertype trips totaltime
## 1 2014-01-01   Customer   652   1809131
## 2 2014-01-01 Subscriber  5407   3781185
## 3 2014-01-02   Customer   181    237960
## 4 2014-01-02 Subscriber  8419   6459889
## 5 2014-01-03   Customer    21     20619
## 6 2014-01-03 Subscriber  1123    898393
```

I obtain the weather data from WeatherUnderground.com (http://www.wunderground.com/history/airport/KNYC/2014/1/1/CustomHistory.html?dayend=31&monthend=12&yearend=2014&req_city=&req_state=&req_statename=&reqdb.zip=&reqdb.magic=&reqdb.wmo=&MR=1). I look at the weather data from Jan 1, 2014 to Dec 31, 2014.

The variables of the weather data include date, temperature, dew point, humidty, sea level pressure, visibility, windspeed, precipitation, max gust speed, cloudcover and events. Variables other than date, precipitation, max gust speed, cloudcover and event all have three measurements - max, mean and min.

Then I combined the two data sources into one dataset

Here's the variables list of the combined dataset. The coding principle is quite simple.

```
names(comb)
```

```
##  [1] "date"                      "usertype"
##  [3] "trips"                     "totaltime"
##  [5] "Max TemperatureF"          "Mean TemperatureF"
##  [7] "Min TemperatureF"          "Max Dew PointF"
##  [9] "MeanDew PointF"            "Min DewpointF"
## [11] "Max Humidity"              "Mean Humidity"
## [13] "Min Humidity"              "Max Sea Level PressureIn"
## [15] "Mean Sea Level PressureIn" "Min Sea Level PressureIn"
## [17] "Max VisibilityMiles"       "Mean VisibilityMiles"
## [19] "Min VisibilityMiles"       "Max Wind SpeedMPH"
## [21] "Mean Wind SpeedMPH"        "Max Gust SpeedMPH"
```

```
## [23] "PrecipitationIn"           "CloudCover"
## [25] "Events"                    "WindDirDegrees"
## [27] "tempdiff"
```

Descriptive Statistics

```
summary(comb)
```

```
##       date              usertype            trips
##  Min.   :2014-01-01   Length:730         Min.   :    3
##  1st Qu.:2014-04-02   Class :character   1st Qu.: 1695
##  Median :2014-07-02   Mode  :character   Median : 6174
##  Mean   :2014-07-02                      Mean   :11070
##  3rd Qu.:2014-10-01                      3rd Qu.:20429
##  Max.   :2014-12-31                      Max.   :35377
##
##    totaltime        Max TemperatureF Mean TemperatureF Min TemperatureF
##  Min.   :    4289   Min.   :18.00    Min.   :12.00     Min.   : 4.00
##  1st Qu.: 2582154   1st Qu.:45.00    1st Qu.:40.00     1st Qu.:34.00
##  Median : 6401432   Median :65.00    Median :57.00     Median :50.00
##  Mean   : 9422645   Mean   :61.65    Mean   :54.74     Mean   :47.35
##  3rd Qu.:15878266   3rd Qu.:78.00    3rd Qu.:71.00     3rd Qu.:63.00
##  Max.   :32830869   Max.   :92.00    Max.   :85.00     Max.   :77.00
##
##  Max Dew PointF  MeanDew PointF   Min DewpointF     Max Humidity
##  Min.   :-8.00   Min.   :-12.00   Min.   :-16.00   Min.   : 39.00
##  1st Qu.:32.00   1st Qu.: 25.00   1st Qu.: 18.00   1st Qu.: 64.00
##  Median :48.00   Median : 41.00   Median : 34.00   Median : 77.00
##  Mean   :45.19   Mean   : 39.11   Mean   : 32.56   Mean   : 75.39
##  3rd Qu.:61.00   3rd Qu.: 56.00   3rd Qu.: 50.00   3rd Qu.: 90.00
##  Max.   :75.00   Max.   : 71.00   Max.   : 70.00   Max.   :100.00
##
##  Mean Humidity    Min Humidity   Max Sea Level PressureIn
##  Min.   :30.00   Min.   :12.00   Min.   :29.63
##  1st Qu.:49.00   1st Qu.:32.00   1st Qu.:30.01
##  Median :59.00   Median :40.00   Median :30.12
##  Mean   :59.22   Mean   :42.64   Mean   :30.14
##  3rd Qu.:70.00   3rd Qu.:51.00   3rd Qu.:30.25
##  Max.   :96.00   Max.   :92.00   Max.   :30.74
##                                  NA's   :16
##  Mean Sea Level PressureIn Min Sea Level PressureIn Max VisibilityMiles
##  Min.   :29.24             Min.   :29.02            Min.   : 6.00
##  1st Qu.:29.89             1st Qu.:29.80            1st Qu.:10.00
##  Median :30.02             Median :29.93            Median :10.00
##  Mean   :30.03             Mean   :29.92            Mean   : 9.93
##  3rd Qu.:30.16             3rd Qu.:30.07            3rd Qu.:10.00
##  Max.   :30.66             Max.   :30.59            Max.   :10.00
##  NA's   :16                NA's   :16               NA's   :16
##  Mean VisibilityMiles Min VisibilityMiles Max Wind SpeedMPH
##  Min.   : 2.000       Min.   : 0.000      Min.   : 4.00
##  1st Qu.: 8.000       1st Qu.: 2.000      1st Qu.:12.00
##  Median :10.000       Median : 9.000      Median :13.00
```

```
##  Mean   : 8.725       Mean   : 6.619      Mean   :13.94
##  3rd Qu.:10.000       3rd Qu.:10.000      3rd Qu.:16.00
##  Max.   :10.000       Max.   :10.000      Max.   :99.00
##  NA's   :16           NA's   :16          NA's   :16
##  Mean Wind SpeedMPH Max Gust SpeedMPH PrecipitationIn      CloudCover
##  Min.   : 1.000     Min.   :11.00     Length:730         Min.   :0.000
##  1st Qu.: 4.000     1st Qu.:18.00     Class :character   1st Qu.:1.000
##  Median : 5.000     Median :22.00     Mode  :character   Median :3.000
##  Mean   : 5.644     Mean   :22.55                        Mean   :3.415
##  3rd Qu.: 7.000     3rd Qu.:26.00                        3rd Qu.:6.000
##  Max.   :99.000     Max.   :99.00                        Max.   :8.000
##  NA's   :16         NA's   :28                           NA's   :16
##     Events          WindDirDegrees     tempdiff
##  Length:730         Min.   : -1.0   Min.   :-25.000000
##  Class :character   1st Qu.: 70.0   1st Qu.: -3.000000
##  Mode  :character   Median :237.0   Median :  0.000000
##                     Mean   :191.6   Mean   :  0.008219
##                     3rd Qu.:286.0   3rd Qu.:  4.000000
##                     Max.   :355.0   Max.   : 18.000000
##
```

Initial Model

. Tell me what model you are using and why (logit, probit, LPM, fixed effects, etc.). Start off with a simple model relating yo u main IV to your main DV. Explain the relationship and why this initial model is insufficient. Maybe you need to make a scale/index of variables. Maybe you need to control for additional factors. Maybe you want to include interaction terms. Maybe you need to check for serial correlation. Etc. Interpret everything correctly (ceteris paribus, on the right scale, etc.)

My initial model is a OLS model. The Y variable is the number of trips among annual subscribers. The X variable is daily average temperature.

A scatter plot supports my intuition.

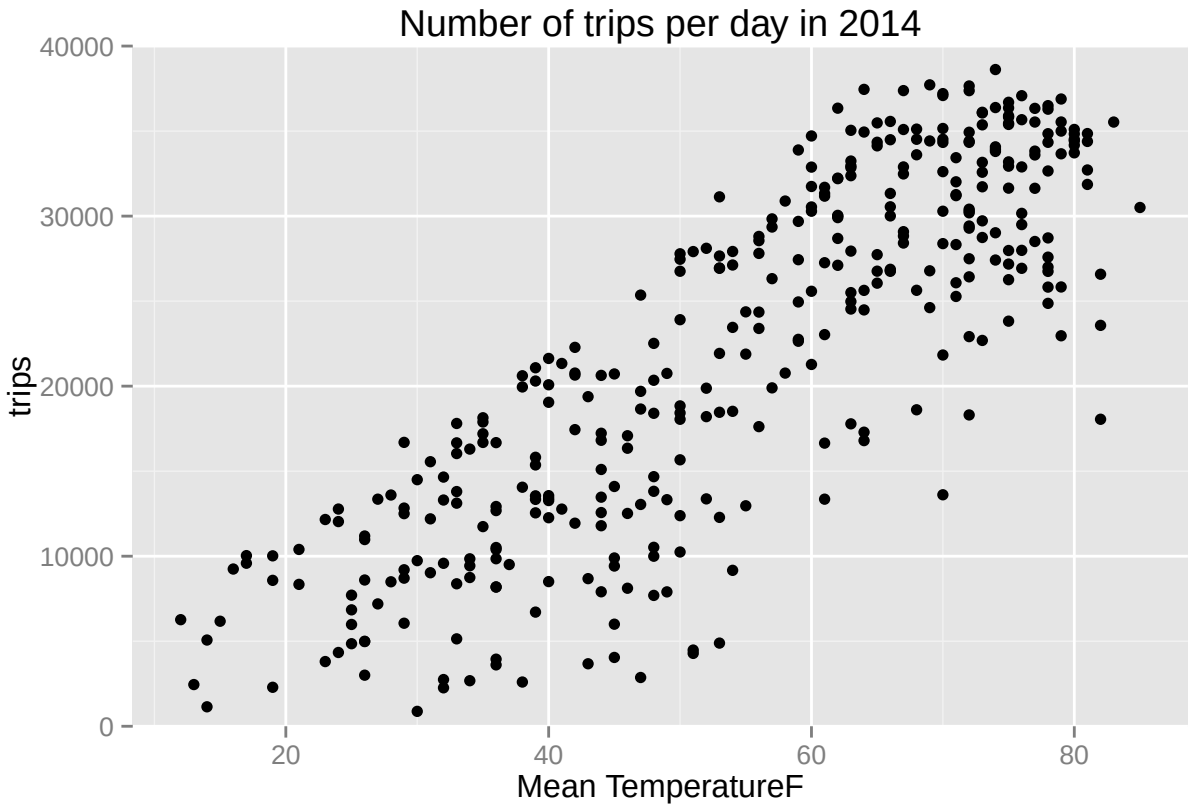
```
library(ggplot2)
library(dplyr)
load("combined2014.RData")
load("summary2014.RData")

total <- summary2014 %>% group_by(date) %>% summarise(trips=sum(trips), totaltime=sum(totaltime))

comb_total <- inner_join(total, weather)
```

```
## Joining by: "date"
```

```
ggplot(data=comb_total, aes(`Mean TemperatureF`, trips))+
  geom_point()+
  scale_y_continuous(limits=c(0,40000), expand=c(0,0))+
  ggtitle("Number of trips per day in 2014")
```



The linear regression model is as follows

```
m1 <- lm(data=comb_total, trips~`Mean TemperatureF`)

summary(m1)
```

```
##
## Call:
## lm(formula = trips ~ `Mean TemperatureF`, data = comb_total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17213.3  -3263.8    515.1   4481.3  10854.7
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -4221.1      939.1  -4.495 9.38e-06 ***
## `Mean TemperatureF`    481.6       16.3  29.546  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5598 on 363 degrees of freedom
## Multiple R-squared:  0.7063, Adjusted R-squared:  0.7055
## F-statistic: 872.9 on 1 and 363 DF,  p-value: < 2.2e-16
```

The initial model suggests that temperature explains the number of city bike trips well. The adjusted R-squared is 0.7. Net of other factors, a degree Farenheit increase in mean temperature leads to an increase of 482 trips per day.

But there's room for improvement

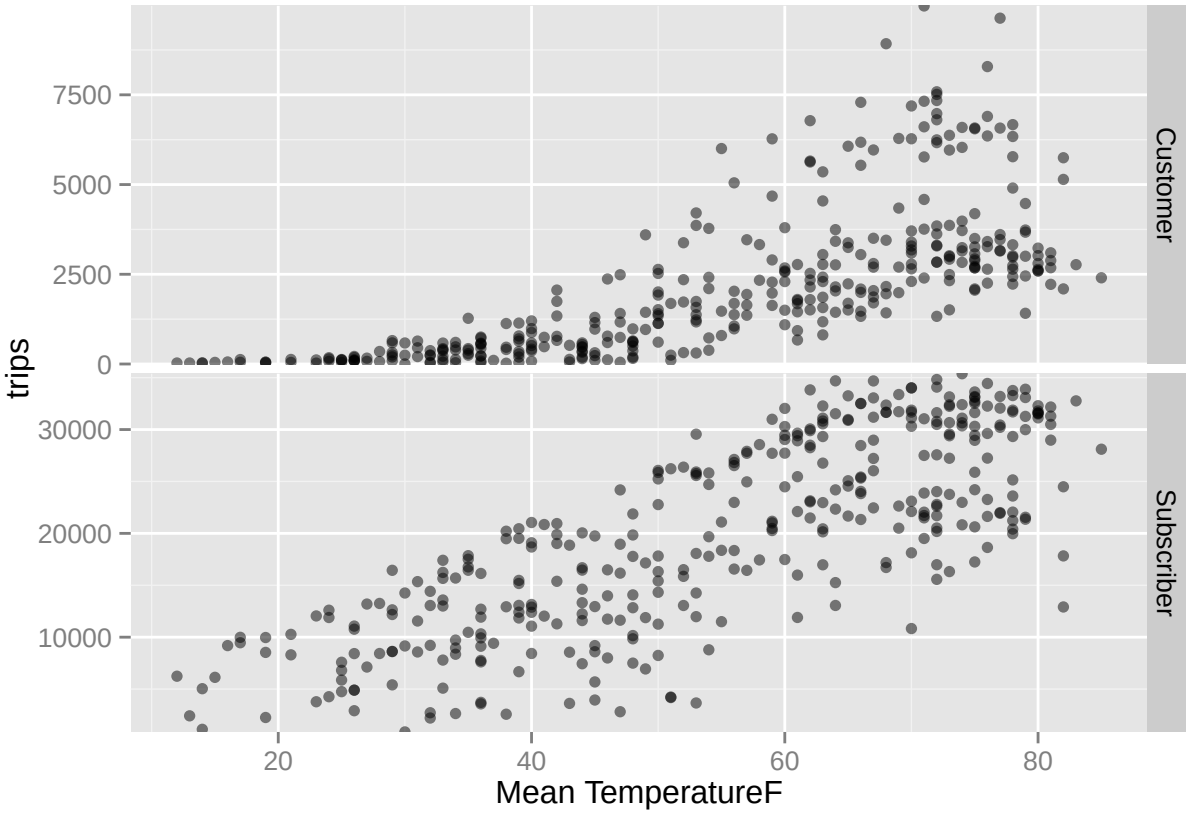
1. Membership type and interaction

Weather change might affect short-term and annual rider differently. Annual riders are likely to be commuters and would be less affected by the change in weather. Short-term riders may be tourists or people who are trying out city bike. I would expect them to be affected by the change in weather more.

I made scatterplot by usertype (short-term user or annual subscribers). The graph suggested the regression lines might have different slopes across the two graphs. It motivated me to include user type as well as its interaction term with the temperature into the model.

```
d <- ggplot(data=comb, aes(`Mean TemperatureF`, trips))+
  geom_point(alpha=.5)+
  scale_y_continuous(expand = c(0,5))

d + facet_grid(usertype~., scales="free_y")
```



2. Temperature is relative?

To riders, change in temperature could be a relative term. A 60 degree weather may be considered warm if the previous day's temperature is 50, but would be considered cold to the rider if the previous day's temperature is 70.

```
m3 <-lm(data=comb, trips~`Mean TemperatureF`*usertype+tempdiff)
summary(m3)
```

```
##
## Call:
## lm(formula = trips ~ `Mean TemperatureF` * usertype + tempdiff,
##     data = comb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17680.3  -1362.2   -171.8   2332.6  10991.3
##
## Coefficients:
##                                      Estimate Std. Error t value
## (Intercept)                          -2388.56     701.38  -3.406
## `Mean TemperatureF`                     83.35      12.18   6.842
## usertypeSubscriber                     287.05     985.06   0.291
## tempdiff                               -40.52      24.76  -1.637
## `Mean TemperatureF`:usertypeSubscriber   319.78      17.10  18.704
##                                      Pr(>|t|)
## (Intercept)                          0.000697 ***
## `Mean TemperatureF`                  1.66e-11 ***
## usertypeSubscriber                   0.770823
## tempdiff                             0.102143
## `Mean TemperatureF`:usertypeSubscriber  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4152 on 725 degrees of freedom
## Multiple R-squared:  0.8612, Adjusted R-squared:  0.8604
## F-statistic:  1124 on 4 and 725 DF,  p-value: < 2.2e-16
```

The regression result shows that adding temperature difference doesn't improve the model. This suggests that riders behavior doesn't depend on the change in temperature much. I am not including this variable to the model.

3. What if it rains or snows?

Precipitation would also affect riders' behavior. I compared two models - one with precipitation level as continuous variable, the other with precipitation event as a boolean variable. Performance is similar, but I go with the one with boolean variable. The adjusted R sqaure is slightly better and interpretation is easier with precipitation as a dummy variable.

```
m4 <- lm(data=comb, trips~`Mean TemperatureF`*usertype+PrecipitationIn)
summary(m4)
```

```
##
## Call:
## lm(formula = trips ~ `Mean TemperatureF` * usertype + PrecipitationIn,
##     data = comb)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16228  -1503   -367   2245  14318
##
## Coefficients:
```

```
##                                       Estimate Std. Error t value
## (Intercept)                           -1696.65     672.75  -2.522
## `Mean TemperatureF`                      80.00      11.62   6.886
## usertypeSubscriber                      417.37     948.18   0.440
## PrecipitationIn                       -3286.13     340.73  -9.644
## `Mean TemperatureF`:usertypeSubscriber   315.79      16.43  19.221
##                                       Pr(>|t|)
## (Intercept)                             0.0119 *
## `Mean TemperatureF`                   1.29e-11 ***
## usertypeSubscriber                      0.6599
## PrecipitationIn                        < 2e-16 ***
## `Mean TemperatureF`:usertypeSubscriber  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3919 on 689 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.8757, Adjusted R-squared:  0.8749
## F-statistic:  1213 on 4 and 689 DF,  p-value: < 2.2e-16
```

```
#recode rain as a dummy
comb$precipitat <- comb$PrecipitationIn!=0
m5 <- lm(data=comb, trips~`Mean TemperatureF`*usertype+precipitat)
summary(m5)
```

```
##
## Call:
## lm(formula = trips ~ `Mean TemperatureF` * usertype + precipitat,
##     data = comb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15813.7  -1750.6    127.8   2257.1  10013.4
##
## Coefficients:
##                                       Estimate Std. Error t value
## (Intercept)                           -1054.69     677.67  -1.556
## `Mean TemperatureF`                      79.74      11.56   6.895
## usertypeSubscriber                      417.37     943.85   0.442
## precipitatTRUE                        -3072.50     306.94 -10.010
## `Mean TemperatureF`:usertypeSubscriber   315.79      16.35  19.309
##                                       Pr(>|t|)
## (Intercept)                              0.120
## `Mean TemperatureF`                   1.22e-11 ***
## usertypeSubscriber                       0.658
## precipitatTRUE                         < 2e-16 ***
## `Mean TemperatureF`:usertypeSubscriber  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3901 on 689 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.8768, Adjusted R-squared:  0.8761
## F-statistic:  1226 on 4 and 689 DF,  p-value: < 2.2e-16
```

4. Weekday vs weekend?

I think days of week would affect the riders' behaviors. Subscribers tend to be commuters and would not need to work during weekends. Short-term users acutally would behave on the opposite.

I then constructed a dummy variable weekend. I build the following models to validate my hypothesis. The result suggest that it's better to include both weekend and its interaction with user type. The model with interaction has a higher adjusted R sqaured at 0.92.

```
comb$weekday <- weekdays(comb$date)
comb$weekend <- comb$weekday=="Saturday"|comb$weekday=="Sunday"

m6 <- lm(data=comb, trips~`Mean TemperatureF`*usertype+precipitat+weekend)
summary(m6)
```

```
##
## Call:
## lm(formula = trips ~ `Mean TemperatureF` * usertype + precipitat +
##     weekend, data = comb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14734.9  -2192.8    -29.1   2529.7   9293.3
##
## Coefficients:
##                                      Estimate Std. Error t value
## (Intercept)                          -500.62     655.24  -0.764
## `Mean TemperatureF`                    82.34      11.12   7.406
## usertypeSubscriber                    417.37     906.97   0.460
## precipitatTRUE                      -3124.94     295.03 -10.592
## weekendTRUE                         -2421.41     317.49  -7.627
## `Mean TemperatureF`:usertypeSubscriber   315.79      15.72  20.095
##                                     Pr(>|t|)
## (Intercept)                            0.445
## `Mean TemperatureF`                 3.82e-13 ***
## usertypeSubscriber                     0.646
## precipitatTRUE                      < 2e-16 ***
## weekendTRUE                         8.03e-14 ***
## `Mean TemperatureF`:usertypeSubscriber  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3749 on 688 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.8864, Adjusted R-squared:  0.8856
## F-statistic:  1074 on 5 and 688 DF,  p-value: < 2.2e-16
```

```
m7 <- lm(data=comb, trips~`Mean TemperatureF`*usertype+precipitat+weekend+weekend:usertype)
summary(m7)
```

```
##
## Call:
## lm(formula = trips ~ `Mean TemperatureF` * usertype + precipitat +
```

```
##     weekend + weekend:usertype, data = comb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16074.1  -1299.6    144.5   1745.7   8038.2
##
## Coefficients:
##                                      Estimate Std. Error t value
## (Intercept)                         -1456.849    562.944  -2.588
## `Mean TemperatureF`                    77.666      9.502   8.174
## usertypeSubscriber                   2329.828    783.990   2.972
## precipitatTRUE                      -3124.943    252.041 -12.399
## weekendTRUE                          1914.677    383.520   4.992
## `Mean TemperatureF`:usertypeSubscriber  325.128     13.438  24.195
## usertypeSubscriber:weekendTRUE      -8672.176    542.306 -15.991
##                                     Pr(>|t|)
## (Intercept)                          0.00986 **
## `Mean TemperatureF`                 1.44e-15 ***
## usertypeSubscriber                   0.00306 **
## precipitatTRUE                       < 2e-16 ***
## weekendTRUE                         7.57e-07 ***
## `Mean TemperatureF`:usertypeSubscriber  < 2e-16 ***
## usertypeSubscriber:weekendTRUE       < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3203 on 687 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.9172, Adjusted R-squared:  0.9165
## F-statistic:  1269 on 6 and 687 DF,  p-value: < 2.2e-16
```

5. Ever too hot to ride a bike?

Before doing analysis, I was suspecting there be a quadratic function on temperature, because people don't want to ride in the hot days. But the scatterplot doesn't suggest including temperature as a quadratic term.

I built a model with quadratic term to confirm my informed guess. The regression suggests that I should not include temperature as a quadratic term.

```
m8<- lm(data=comb_total, trips~poly(`Mean TemperatureF`,2))
summary(m8)
```

```
##
## Call:
## lm(formula = trips ~ poly(`Mean TemperatureF`, 2), data = comb_total)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17297  -3218    461   4450  10880
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   22140.3      293.4  75.463   <2e-16 ***
```

```
## poly(`Mean TemperatureF`, 2)1 165389.3     5605.3  29.506   <2e-16 ***
## poly(`Mean TemperatureF`, 2)2    830.8     5605.3   0.148    0.882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5605 on 362 degrees of freedom
## Multiple R-squared:  0.7063, Adjusted R-squared:  0.7047
## F-statistic: 435.3 on 2 and 362 DF,  p-value: < 2.2e-16
```

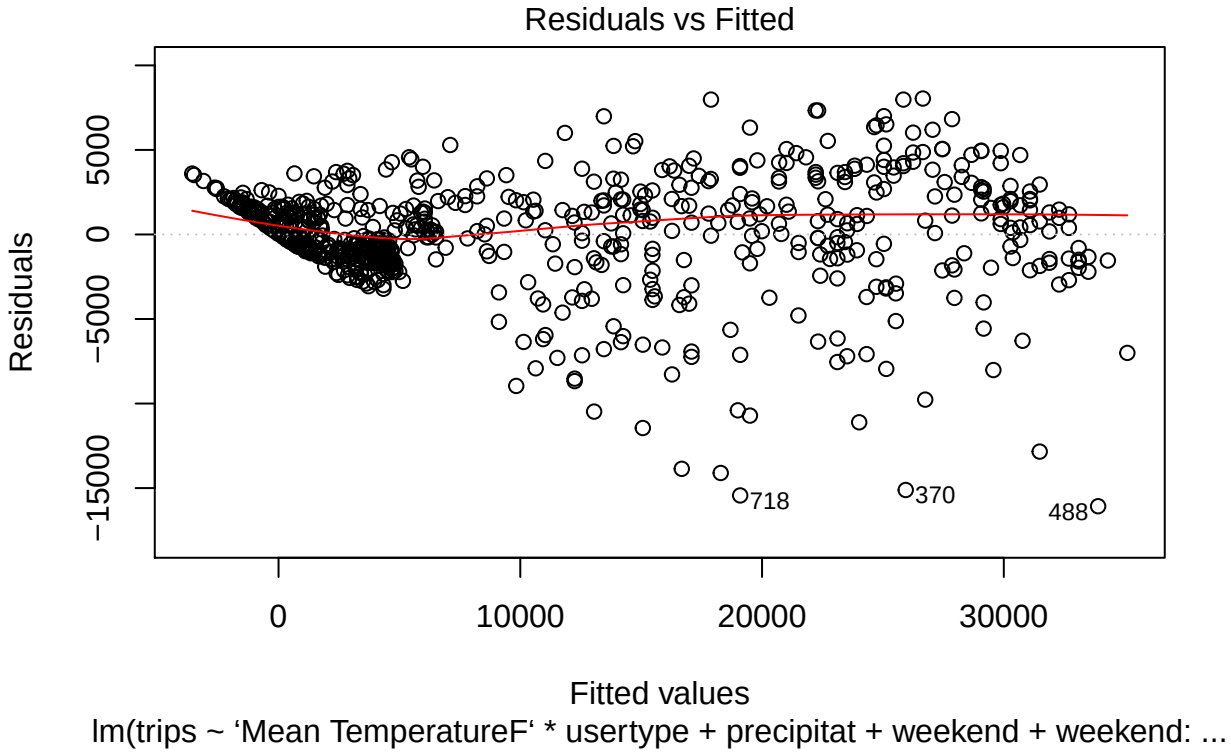
Final Models & Conclusion

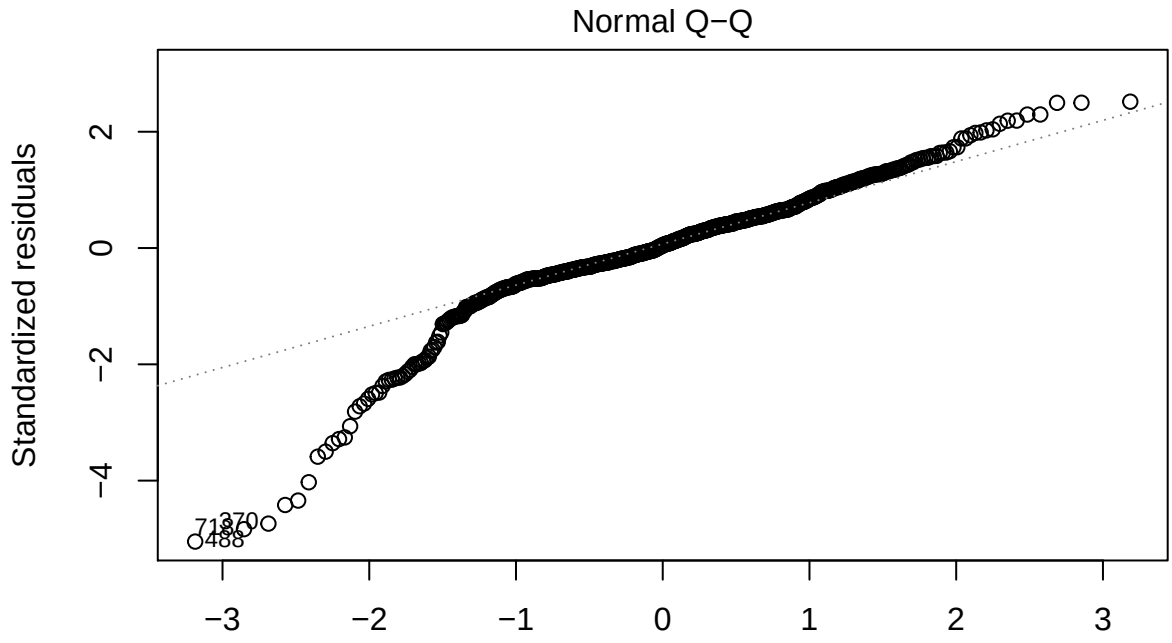
After going through these intermediate steps to improve my model, my final model is `m7 <- lm(data=comb, trips~`Mean TemperatureF`*usertype+precipitat+weekend+weekend:usertype)`. The model performs quite well. The adjusted R squared is 0.92.

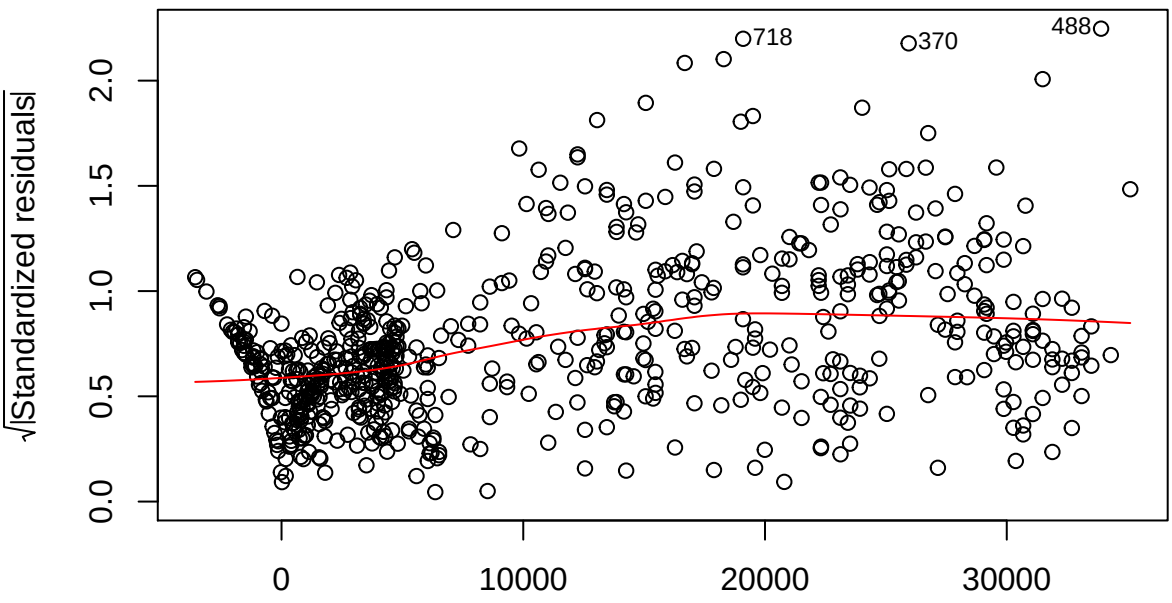
Under this model, my explanatory variables are mean temperature, usertype, precipitation (event/dummy), weekend (dummy) and interaction terms between usertype and weekend, as well as that between usertype and mean temperature.

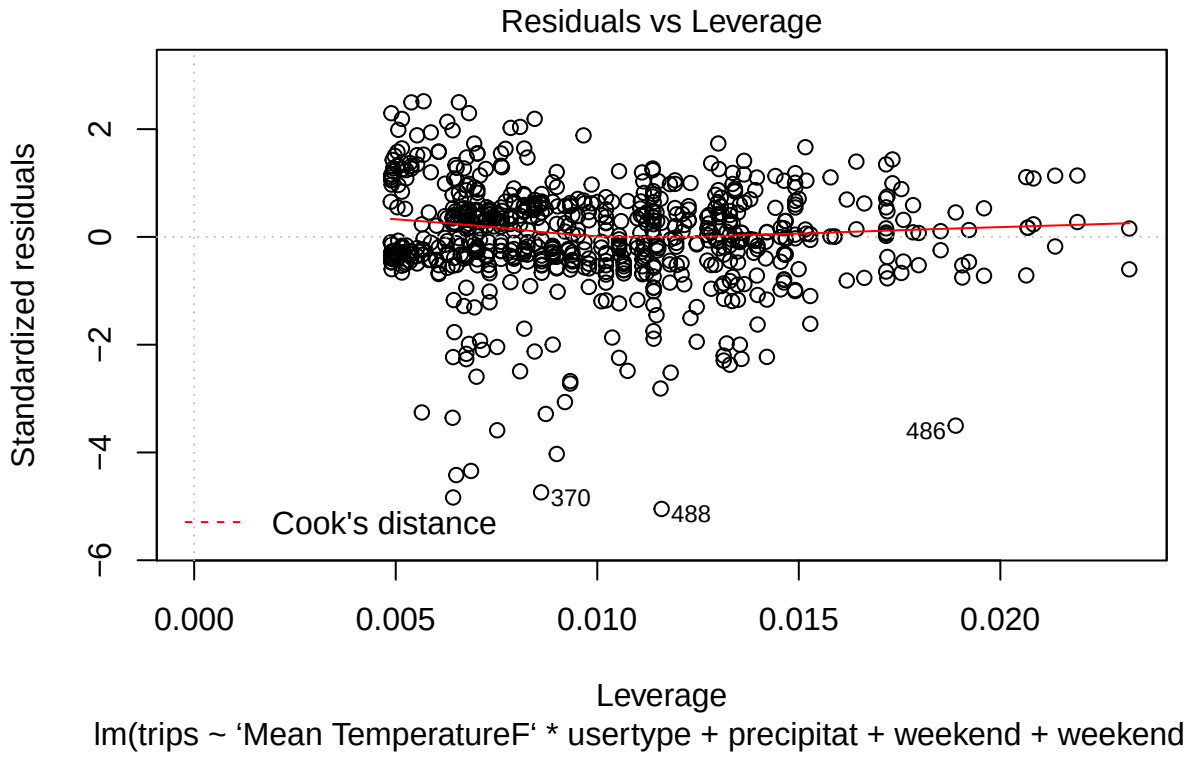
Based on the following diagnostic plots, there are some violation of the regression assumption at the extreme ends. But I think the violation is not serious. It is reasonable to use regression model.

```
plot(m7)
```




Normal Q−Q
Standardized residuals
2
0
−2
−4
718
370
488
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(trips ~ ‘Mean TemperatureF‘ * usertype + precipitat + weekend + weekend: ...


Scale−Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
718
370
488
0
10000
20000
30000
Fitted values
lm(trips ~ ‘Mean TemperatureF‘ * usertype + precipitat + weekend + weekend: ...



My initial hypothesis was supported. The only expectation not met was the relative change in temperature. I guess New Yorkers check their weather app before they make a decision on using city bike. It's not driven by "Hmm.. it's colder than yesterday. I will not ride a bike."

I found that higher temperature is associated with higher rides per day. Short-term users are more likely to use city bikes on weekends, but not annual subscribers. Rain or snow reduces the number rides per day.

But I didn't know how to determine whether subscirbers are more subject to weather change or customers. I wish I could have more time and use another model to detect that. I suspect I could accomplish this by scaling the trips in two groups. Also, I wanted to do somewith in spatial relationship, such as identifying popular routes, but didn't have enough skills to accomplish it.

There's one limitation with my analysis. City bike station might increase over time. Changes in number of trips could be affected by expansion of the city bike system. But that was not a poor decision not to consider this in the beginning. I later found that since the start of the operation in 2013, there's no expansion until late 2015.

Reflecting back to this project, there's quite little surprise to my finding. But I was able to learn more web-scraping with rvest package, get more familiar with dplyr, and practice plotting multiple plots with facet wrap in ggplot2. I also implemented some workflow ideas where I have 4 R scripts to scrape, clean, combine, and analyze my data respectively. Sadly, my analysis is not relevant to my thesis.