Introduction to Statistics for Data Scientists

BIKESHARING DEMAND ESTIMATION

Eastern Market Metro
8th St & Pennsylvania Ave SE
capital bikeshare
Local Capital Bikeshare Stations
bikeshare

Submitted by:

Arpit Sidana

Chloe Ni

Gunreet Thind

Saurabh Kumar

Contents

Introduction

Bike sharing systems are a new generation of traditional bike rentals where all the touchpoints in the whole process from membership, rental to return have become automated. Through these systems, user is able to easily rent a bike from a particular position and return it back at another position.

Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in a city.

Data Gathering

Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day etc. can affect the rental behaviors. The data set we are using is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available in UCI Machine Learning Repository. The dataset contains the daily count of rental bikes between years 2011 and 2012 in Capital Bikeshare system with the corresponding weather and seasonal information.

Objective

The business objective is to determine factors that influence people's decision to avail bikeshare rentals from Capital Bikeshare system Washington D.C to estimate demand and manage inventory on a daily basis.

We chose casual users from total users, casual and registered users to define our business problem for high business relevance such that we can increase conversion rate for the average casual user to drive increased profits. We built a multiple regression model using the statistical programming language R to conclude about the relationship between the count of casual users who rented bikes and corresponding weather and seasonal information.

Attributes used

Weather: it is one of the most important factors when people choose their daily transportation. We prefer not to ride bike to school if it's raining or it is windy. In order to measure the impact of weather comprehensively, we include several variables in our model:

- Temperature (Interval, normalized temperature in Celsius, divided to 41 (max))

- Humidity (Interval, normalized humidity, divided to 100 (max))

- Wind speed (Interval, normalized wind speed, divided to 67 (max))

- Weather description (such as cloudy, misty, snowy, rainy, thunderstorm, ice pallet. We use four levels(nominal) to evaluate the weather.

1: Clear, Few clouds, Partly cloudy, Partly cloudy

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

4: Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog)

Time Period: People’s willingness to rent bikes may also change among different time periods. Winter might be too cold to ride a bike and there might be larger demand for renting bikes on weekdays. So we add the following variables into consideration:

- Year (0: 2011, 1: 2012)
- Season (1 to 4, 1 is spring, 2 is summer, 3 is autumn, 4 is winter)
- Month (1 to 12)
- Weekday (0 to 6)
- Working day (1: weekend or holiday; 0: otherwise)
- Hour (0 to 23)
- Holiday (1: Holiday; 0: not holiday)

Cleaning Data

A preliminary analysis for all the attributes shows that there are no missing values in the data. Also there were no outliers found in the data which has been explained to a greater detail under Note 1 of the Appendix.

Descriptive Statistics

We used visualizations of the likes of histograms and scatter plots and found significant fluctuation in weather variables such as Windspeed and fluctuations in trends of casual user’s rental over various hours of the day which led us to conclude that the relationship is worth investigation in a statistical model.

Generally, the variable characteristics observed for the various predictor variables have been explained in the Appendix under Note 2.

Some of the observations are that more people (casual users) rented bike in 2012 as compared to 2011. More people rented bikes during the working day as compared to the weekend. Another observation is that more bikes are rented in spring as compared to the other season where winter has minimal renting.

Assumptions

Before we build the analysis model, following are the assumptions:

1. We assume that the data collecting process is stable over time and has no bias.
2. Also the residuals should be normally distributed. However, based on the statistical test, the residuals of our results do not satisfy the normal distribution assumption (for more details please refer the Appendix Note 5. Therefore, to further improve the model in the future, we need to pay attention to it and modify the model if necessary.

Analysis

Model Building

The best-fit model suggests that the following variables as the most crucial predictors for determining the number of casual users taking the bikes which are:

- Hour of the day
- Year

- Month
- Type of Weather
- Whether day is a Working Day or not
- Temperature
- Wind speed
- Humidity

We observed that the 45% of the variability in the response variable is explained by the above mentioned factors. This might be improved by adding few external factors like regional distribution of bike usage in Washington for which we didn’t have the data. It is noteworthy that we have considered the possible tradeoffs on errors prediction (Residuals), the variability accounted (Multiple R-squared) and statistical significance(p-values).

We have considered some changes in the predictor variables selection:

- We dropped ‘Week Day’ and ‘Holiday’ and kept ‘Working Day’ due to correlation among them.
- We dropped ‘Feeling Temperature’ due to its lower statistical significance and correlation with ‘Temperature’

Interpretation

The final model suggests the following conclusions:

Monthly Trends:

- The number of casual users decreased by 1 on average with the change in the month from January to February keeping other predictors constant.
- Similarly, we checked increase for subsequent months relative to January through other Month estimates
- We noted most significant increases in October of 13 casual users against January and other significant (>10 increase) during months of September to October

Hourly Trends:

- To interpret hourly trends, base was relevelled to 4-5 am slot since it had lowest number of bikes rented
- The “Hour” estimate (beta-hat for level factor(Hour)5) indicated increase of 2 bikes on average from 4-5 am to 5-6 am, keeping other predictors constant
- Similarly, we checked increase for subsequent hours relative to 4-5 am through other Hour estimates (Hour1: Hour12) except Hour4
- Most significant increase during 5-6 pm of 60 bikes and other significant (>40 increase) from 12 noon to 4 pm
- For every 1-degree Celsius increase in temperature the no. of bikes rented increases by approx. 2 keeping all other predictor variables fixed
- For every 5 mph increase in wind speed the demand for renting bikes falls by about 1 unit keeping all other predictor variables fixed
- For every 1% increase in humidity, the no. of rented bikes falls by 0.285, implying 5% increase in humidity would cause the number of bikes to decrease by 1 unit

Conclusion

We found statistical significance of weather (Humidity, Temperature, Weather Description and Windspeed) and time based indicators (Hour, Year, Month) on the decision of casual users as to whether or not they opt for a bike rental from Capital Bikeshare system Washington D.C.

Our insights from hourly trends can suggest significant savings for Capital Bikeshare system Washington D.C. if they can optimize their supply chain and logistics in sync with hourly demand.

The weather indicators provide opportunities for Capital Bikeshare system Washington D.C. to bundle products to increase revenue by providing additional services such as helmets, raincoats etc.

Future Prospects

Model accuracy improvement

We propose to improve our model accuracy by increasing its mathematical complexity by the introduction of higher order terms.

We also propose to introduce lag variables to check the dependency of the model on its previous time periods for various units of time. While our model allowed for us to study the change in rental demand for months relative to the month of January, the introduction of lag would allow us to compare it with previous month.

Addition of external factors

Population of the state at a city and county level can help increase accuracy of prediction thereby improving inventory management.

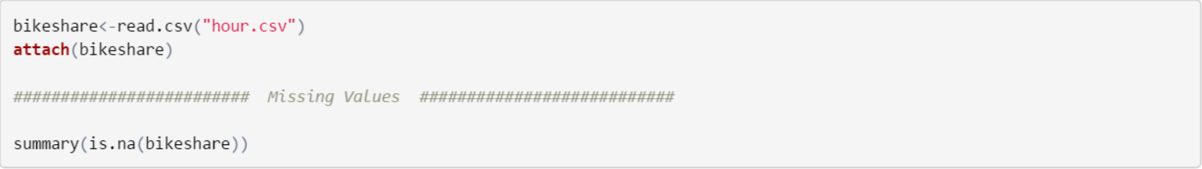
Introduction of points of interest such as new biking trails or parks in a locality can cause a spike in bike rental while data on construction sites can help identify bottlenecks to the demand.

Also, an attempt to improve the model could possibly be achieved by removing the less predictive features in the model and adding additional predictive features not currently in this data set. Factors like population, distance from nearest grocery store, availability of biking lanes per unit area can also affect the count of the bikes. There is scope for gathering this data and including them in the model to make it more efficient and robust. The data we used was only over a two-year period, as these schemes have not been in existence for a long period. Collection of more data over several years combined with the accuracy in which the data is currently being collected could help build a more efficient predictive model in the future.

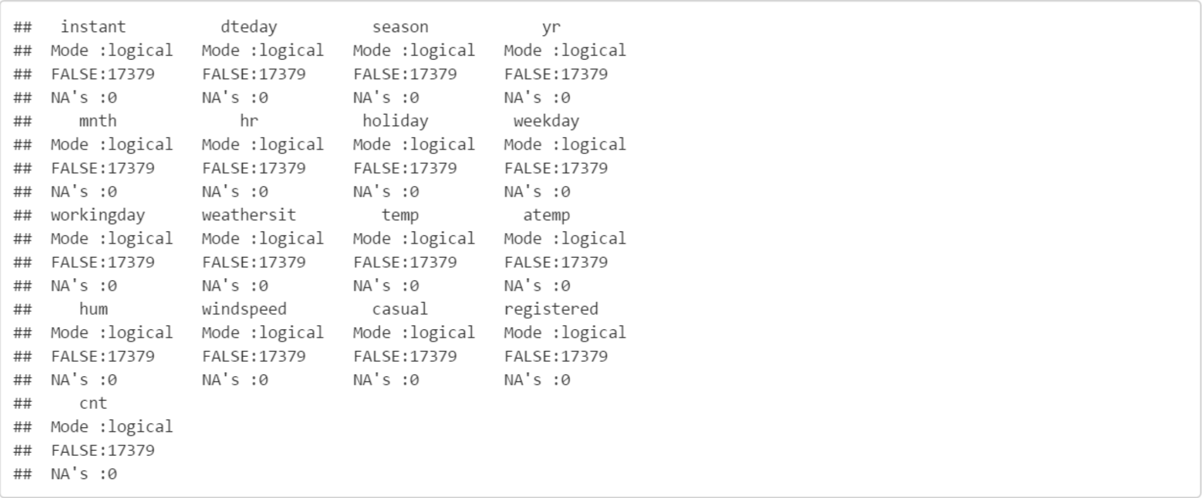
Technical Appendix

Note 1: Missing values analysis

From viewing the data and running the summary R code any missing values within the dataset, we can establish in R by using the function below:



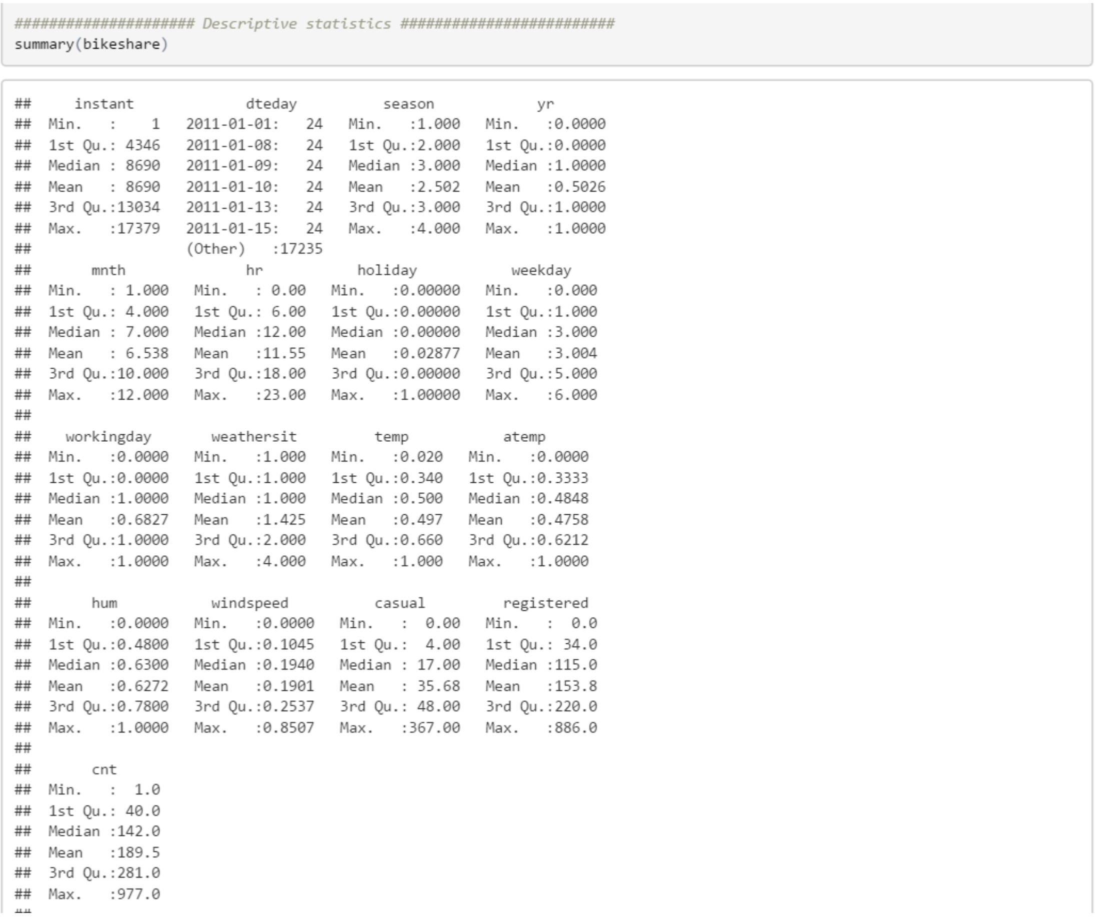
```
bikeshare<-read.csv("hour.csv")
attach(bikeshare)

#########################  Missing Values  ###########################

summary(is.na(bikeshare))
```



```
##   instant         dteday          season           yr
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:17379     FALSE:17379     FALSE:17379     FALSE:17379
##  NA's :0         NA's :0         NA's :0         NA's :0
##     mnth            hr           holiday         weekday
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:17379     FALSE:17379     FALSE:17379     FALSE:17379
##  NA's :0         NA's :0         NA's :0         NA's :0
##  workingday      weathersit        temp           atemp
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:17379     FALSE:17379     FALSE:17379     FALSE:17379
##  NA's :0         NA's :0         NA's :0         NA's :0
##     hum          windspeed         casual        registered
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:17379     FALSE:17379     FALSE:17379     FALSE:17379
##  NA's :0         NA's :0         NA's :0         NA's :0
##     cnt
##  Mode :logical
##  FALSE:17379
##  NA's :0
```

Note 2: Descriptive Characteristics

The summary characteristics show that there are no outliers in the data as depicted in the R code below:

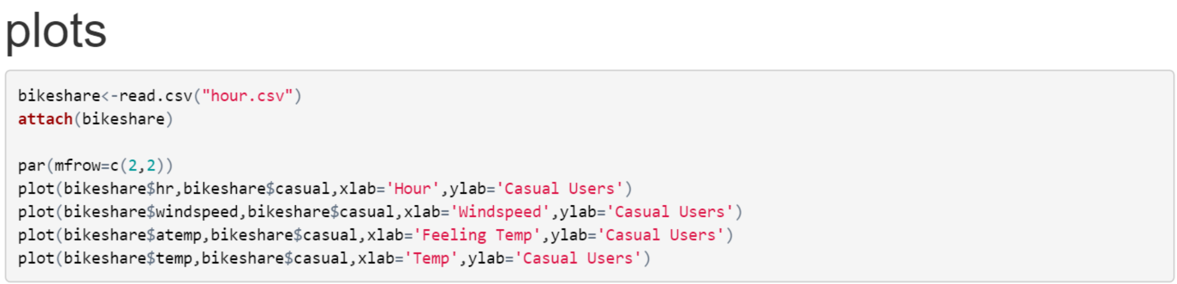


```
#################### Descriptive statistics ########################
summary(bikeshare)
```

```
##     instant            dteday          season            yr        
##  Min.   :    1   2011-01-01:   24   Min.   :1.000   Min.   :0.0000  
##  1st Qu.: 4346   2011-01-08:   24   1st Qu.:2.000   1st Qu.:0.0000  
##  Median : 8690   2011-01-09:   24   Median :3.000   Median :1.0000  
##  Mean   : 8690   2011-01-10:   24   Mean   :2.502   Mean   :0.5026  
##  3rd Qu.:13034   2011-01-13:   24   3rd Qu.:3.000   3rd Qu.:1.0000  
##  Max.   :17379   2011-01-15:   24   Max.   :4.000   Max.   :1.0000  
##                  (Other)   :17235                                   
##       mnth              hr           holiday           weekday     
##  Min.   : 1.000   Min.   : 0.00   Min.   :0.00000   Min.   :0.000  
##  1st Qu.: 4.000   1st Qu.: 6.00   1st Qu.:0.00000   1st Qu.:1.000  
##  Median : 7.000   Median :12.00   Median :0.00000   Median :3.000  
##  Mean   : 6.538   Mean   :11.55   Mean   :0.02877   Mean   :3.004  
##  3rd Qu.:10.000   3rd Qu.:18.00   3rd Qu.:0.00000   3rd Qu.:5.000  
##  Max.   :12.000   Max.   :23.00   Max.   :1.00000   Max.   :6.000  
##                                                                    
##    workingday       weathersit         temp           atemp       
##  Min.   :0.0000   Min.   :1.000   Min.   :0.020   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.340   1st Qu.:0.3333  
##  Median :1.0000   Median :1.000   Median :0.500   Median :0.4848  
##  Mean   :0.6827   Mean   :1.425   Mean   :0.497   Mean   :0.4758  
##  3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:0.660   3rd Qu.:0.6212  
##  Max.   :1.0000   Max.   :4.000   Max.   :1.000   Max.   :1.0000  
##                                                                   
##       hum           windspeed          casual          registered   
##  Min.   :0.0000   Min.   :0.0000   Min.   :  0.00   Min.   :  0.0  
##  1st Qu.:0.4800   1st Qu.:0.1045   1st Qu.:  4.00   1st Qu.: 34.0  
##  Median :0.6300   Median :0.1940   Median : 17.00   Median :115.0  
##  Mean   :0.6272   Mean   :0.1901   Mean   : 35.68   Mean   :153.8  
##  3rd Qu.:0.7800   3rd Qu.:0.2537   3rd Qu.: 48.00   3rd Qu.:220.0  
##  Max.   :1.0000   Max.   :0.8507   Max.   :367.00   Max.   :886.0  
##                                                                    
##       cnt       
##  Min.   :  1.0  
##  1st Qu.: 40.0  
##  Median :142.0  
##  Mean   :189.5  
##  3rd Qu.:281.0  
##  Max.   :977.0  
```

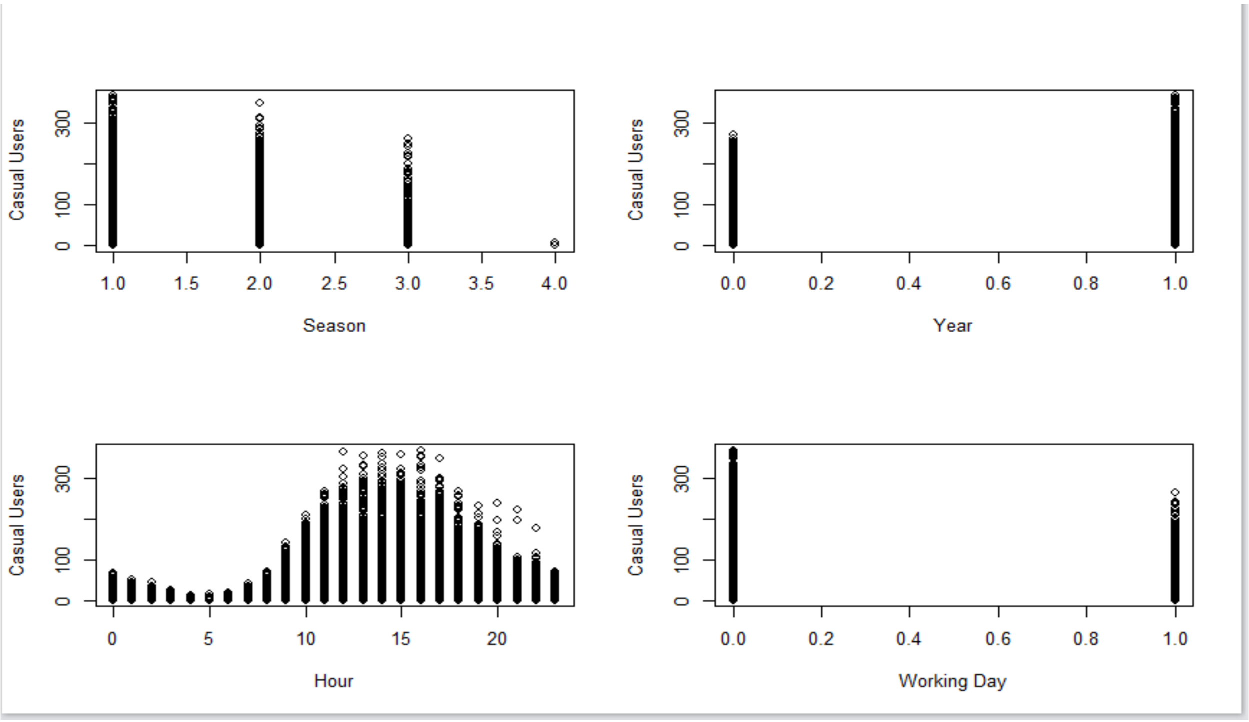
Note 3: Scatter plots

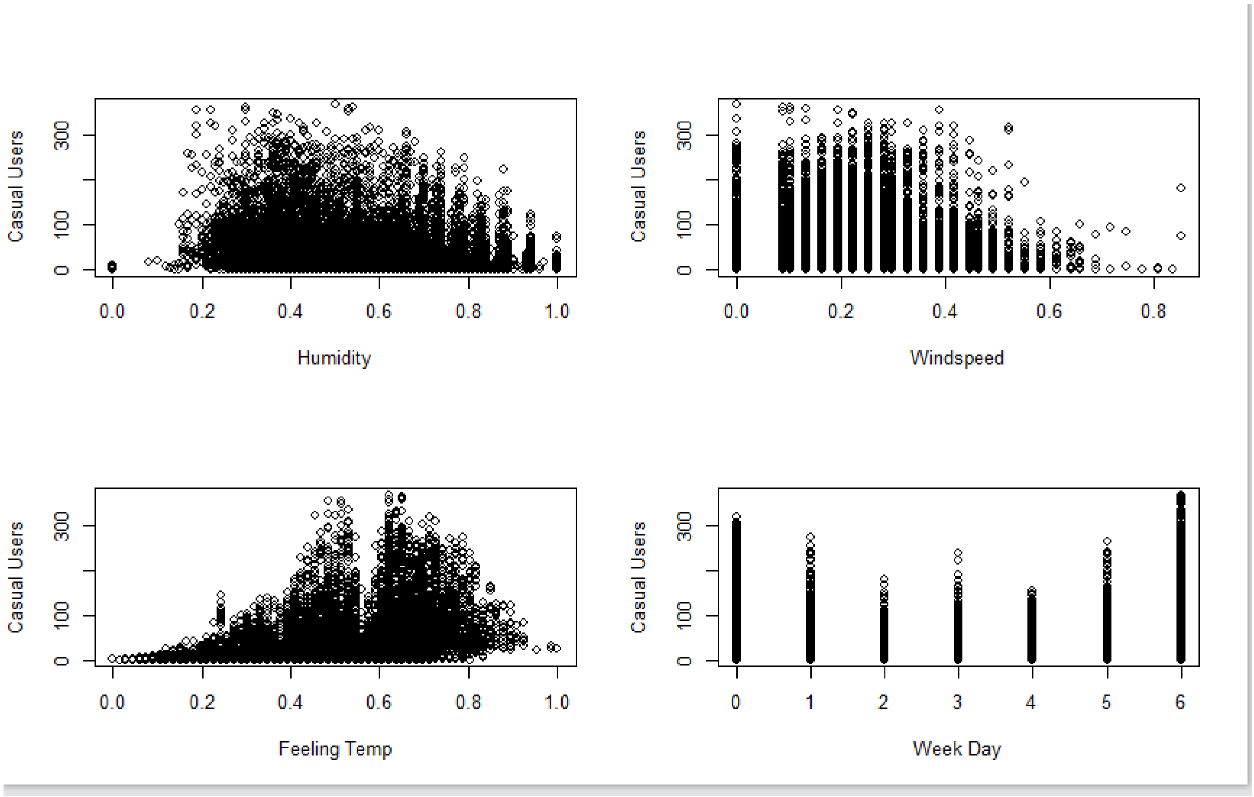
Below are the scatter plots of the response variable (casual users count) with respect to various predictor variables along with the R code



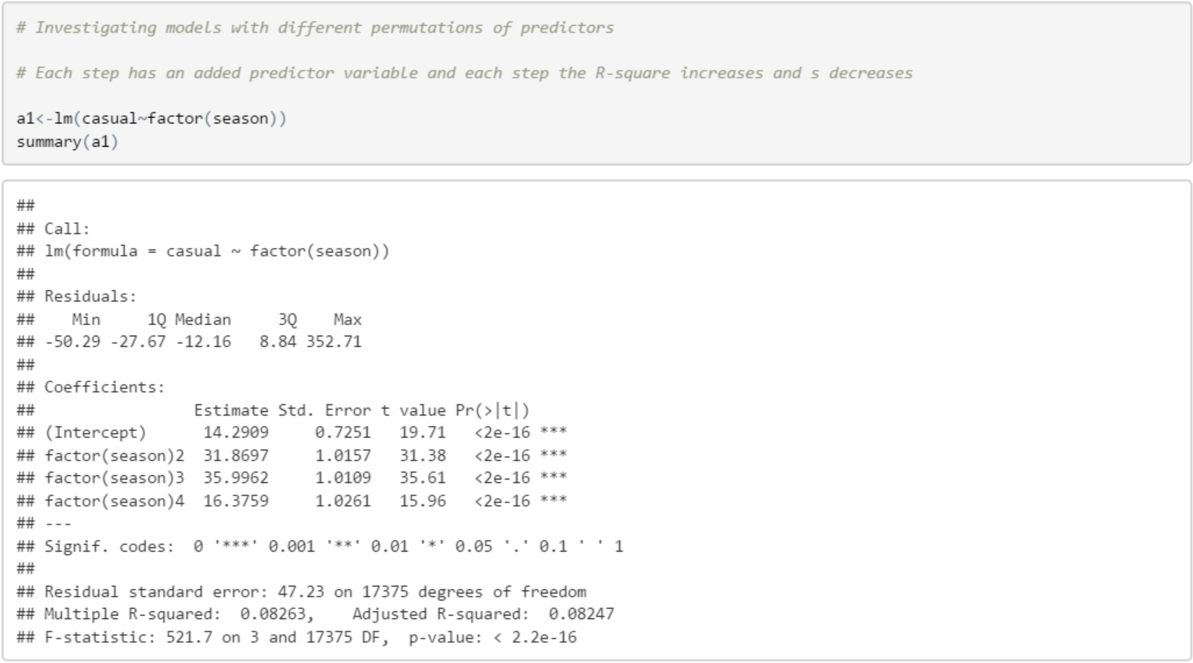
plots

```
bikeshare<-read.csv("hour.csv")
attach(bikeshare)

par(mfrow=c(2,2))
plot(bikeshare$hr,bikeshare$casual,xlab='Hour',ylab='Casual Users')
plot(bikeshare$windspeed,bikeshare$casual,xlab='Windspeed',ylab='Casual Users')
plot(bikeshare$atemp,bikeshare$casual,xlab='Feeling Temp',ylab='Casual Users')
plot(bikeshare$temp,bikeshare$casual,xlab='Temp',ylab='Casual Users')
```


Casual Users
0
100
300
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Season
Casual Users
0
100
300
0.0
0.2
0.4
0.6
0.8
1.0
Year
Casual Users
0
100
300
0
5
10
15
20
Hour
Casual Users
0
100
300
0.0
0.2
0.4
0.6
0.8
1.0
Working Day

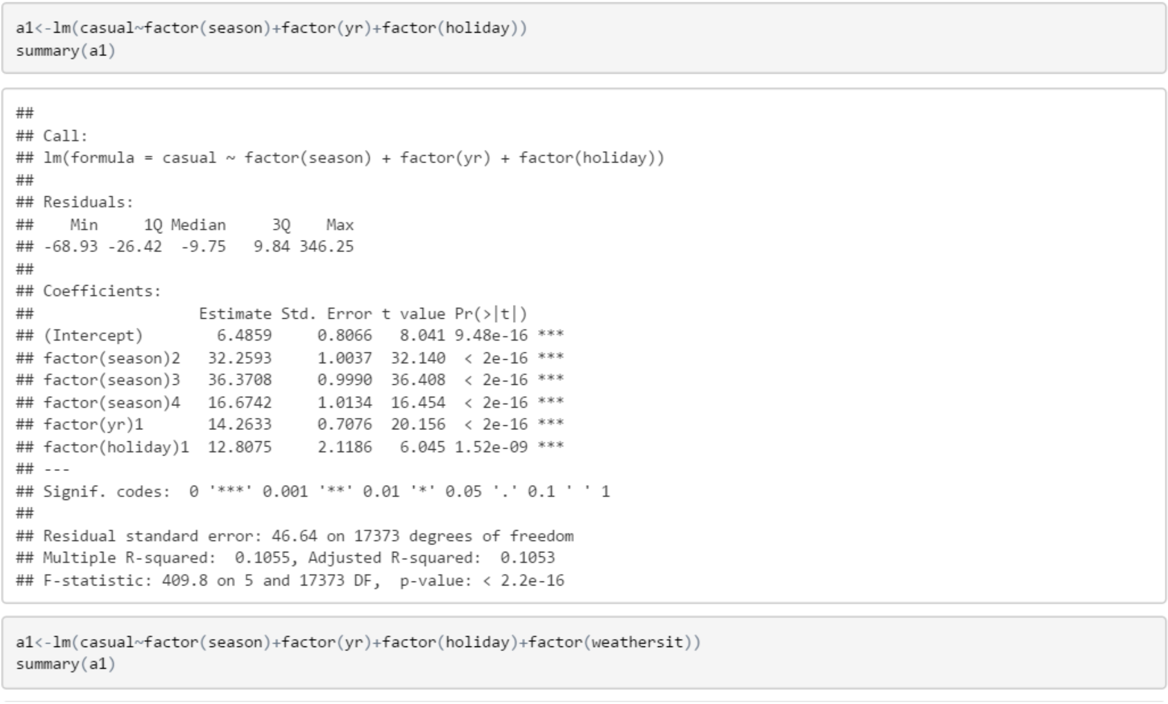

Casual Users
0
100
300
0.0
0.2
0.4
0.6
0.8
1.0
Humidity
Casual Users
0
100
300
0.0
0.2
0.4
0.6
0.8
Windspeed
Casual Users
0
100
300
0.0
0.2
0.4
0.6
0.8
1.0
Feeling Temp
Casual Users
0
100
300
0
1
2
3
4
5
6
Week Day

Note 4: Model building



```
# Investigating models with different permutations of predictors

# Each step has an added predictor variable and each step the R-square increases and s decreases

a1<-lm(casual~factor(season))
summary(a1)
```

```
##
## Call:
## lm(formula = casual ~ factor(season))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50.29 -27.67 -12.16   8.84 352.71
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      14.2909     0.7251   19.71   <2e-16 ***
## factor(season)2  31.8697     1.0157   31.38   <2e-16 ***
## factor(season)3  35.9962     1.0109   35.61   <2e-16 ***
## factor(season)4  16.3759     1.0261   15.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.23 on 17375 degrees of freedom
## Multiple R-squared:  0.08263,    Adjusted R-squared:  0.08247
## F-statistic: 521.7 on 3 and 17375 DF,  p-value: < 2.2e-16
```



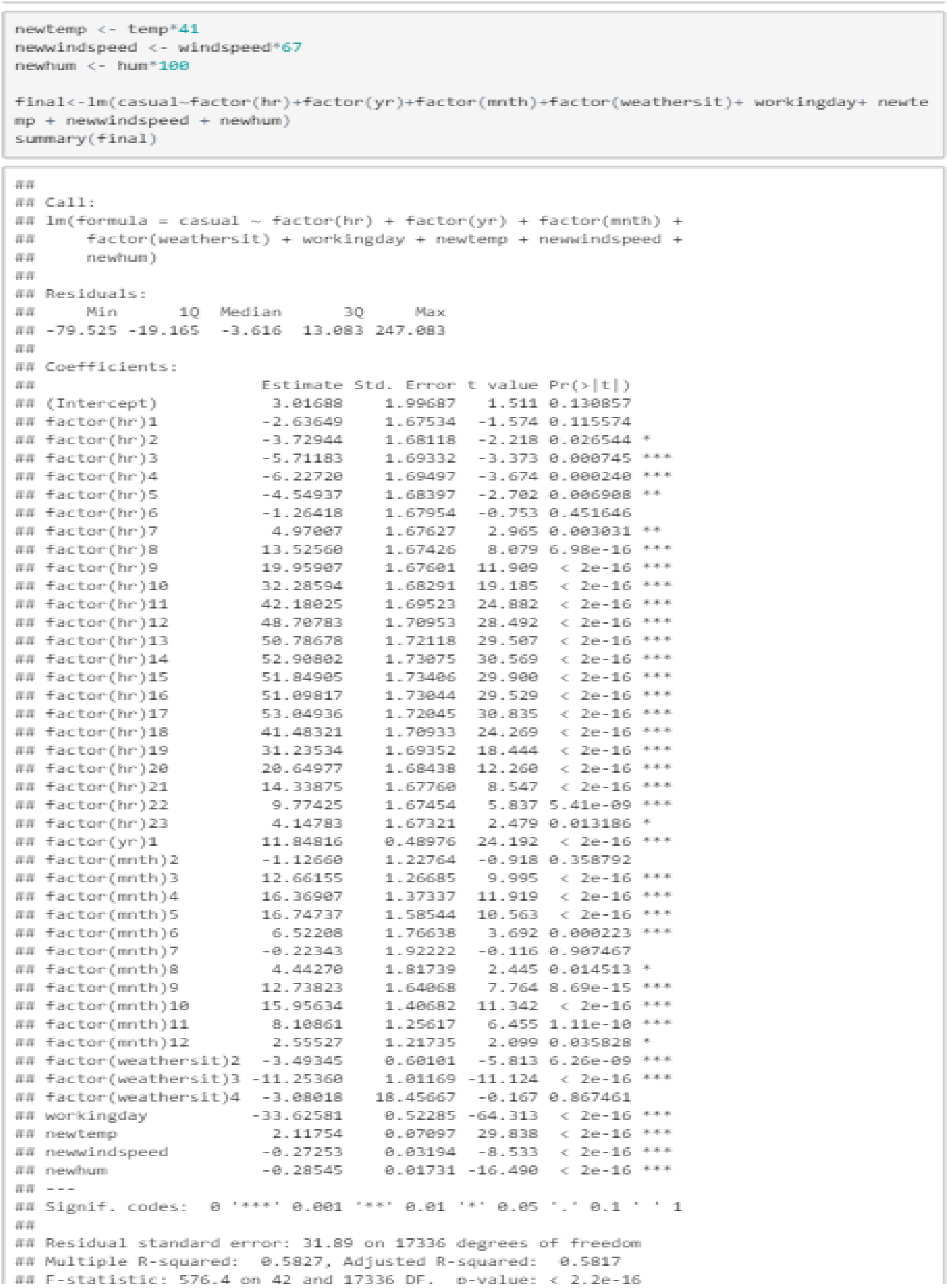
```
a1<-lm(casual~factor(season)+factor(yr)+factor(holiday))
summary(a1)
```

```
##
## Call:
## lm(formula = casual ~ factor(season) + factor(yr) + factor(holiday))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68.93 -26.42  -9.75   9.84 346.25
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.4859     0.8066   8.041 9.48e-16 ***
## factor(season)2   32.2593     1.0037  32.140  < 2e-16 ***
## factor(season)3   36.3708     0.9990  36.408  < 2e-16 ***
## factor(season)4   16.6742     1.0134  16.454  < 2e-16 ***
## factor(yr)1       14.2633     0.7076  20.156  < 2e-16 ***
## factor(holiday)1  12.8075     2.1186   6.045 1.52e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.64 on 17373 degrees of freedom
## Multiple R-squared:  0.1055, Adjusted R-squared:  0.1053
## F-statistic: 409.8 on 5 and 17373 DF,  p-value: < 2.2e-16
```

```
a1<-lm(casual~factor(season)+factor(yr)+factor(holiday)+factor(weathersit))
summary(a1)
```

```r
newtemp <- temp*41
newwindspeed <- windspeed*67
newhum <- hum*100

final<-lm(casual~factor(hr)+factor(yr)+factor(mnth)+factor(weathersit)+ workingday+ newtemp + newwindspeed + newhum)
summary(final)
```

```
##
## Call:
## lm(formula = casual ~ factor(hr) + factor(yr) + factor(mnth) +
##     factor(weathersit) + workingday + newtemp + newwindspeed +
##     newhum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.525 -19.165  -3.616  13.083 247.083
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.01688    1.99687   1.511 0.130857
## factor(hr)1          -2.63649    1.67534  -1.574 0.115574
## factor(hr)2          -3.72944    1.68118  -2.218 0.026544 *
## factor(hr)3          -5.71183    1.69332  -3.373 0.000745 ***
## factor(hr)4          -6.22720    1.69497  -3.674 0.000240 ***
## factor(hr)5          -4.54937    1.68397  -2.702 0.006908 **
## factor(hr)6          -1.26418    1.67954  -0.753 0.451646
## factor(hr)7           4.97007    1.67627   2.965 0.003031 **
## factor(hr)8          13.52560    1.67426   8.079 6.98e-16 ***
## factor(hr)9          19.95907    1.67601  11.909  < 2e-16 ***
## factor(hr)10         32.28594    1.68291  19.185  < 2e-16 ***
## factor(hr)11         42.18025    1.69523  24.882  < 2e-16 ***
## factor(hr)12         48.70783    1.70953  28.492  < 2e-16 ***
## factor(hr)13         50.78678    1.72118  29.507  < 2e-16 ***
## factor(hr)14         52.90802    1.73075  30.569  < 2e-16 ***
## factor(hr)15         51.84905    1.73406  29.900  < 2e-16 ***
## factor(hr)16         51.09817    1.73044  29.529  < 2e-16 ***
## factor(hr)17         53.04936    1.72045  30.835  < 2e-16 ***
## factor(hr)18         41.48321    1.70933  24.269  < 2e-16 ***
## factor(hr)19         31.23534    1.69352  18.444  < 2e-16 ***
## factor(hr)20         20.64977    1.68438  12.260  < 2e-16 ***
## factor(hr)21         14.33875    1.67760   8.547  < 2e-16 ***
## factor(hr)22          9.77425    1.67454   5.837 5.41e-09 ***
## factor(hr)23          4.14783    1.67321   2.479 0.013186 *
## factor(yr)1          11.84816    0.48976  24.192  < 2e-16 ***
## factor(mnth)2        -1.12660    1.22764  -0.918 0.358792
## factor(mnth)3        12.66155    1.26685   9.995  < 2e-16 ***
## factor(mnth)4        16.36907    1.37337  11.919  < 2e-16 ***
## factor(mnth)5        16.74737    1.58544  10.563  < 2e-16 ***
## factor(mnth)6         6.52208    1.76638   3.692 0.000223 ***
```



```r
newtemp <- temp*41
newwindspeed <- windspeed*67
newhum <- hum*100

final<-lm(casual~factor(hr)+factor(yr)+factor(mnth)+factor(weathersit)+ workingday+ newte
mp + newwindspeed + newhum)
summary(final)
```

```
##
## Call:
## lm(formula = casual ~ factor(hr) + factor(yr) + factor(mnth) +
##     factor(weathersit) + workingday + newtemp + newwindspeed +
##     newhum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.525 -19.165  -3.616  13.083 247.083
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.01688    1.99687   1.511 0.130857
## factor(hr)1          -2.63649    1.67534  -1.574 0.115574
## factor(hr)2          -3.72944    1.68118  -2.218 0.026544 *
## factor(hr)3          -5.71183    1.69332  -3.373 0.000745 ***
## factor(hr)4          -6.22720    1.69497  -3.674 0.000240 ***
## factor(hr)5          -4.54937    1.68397  -2.702 0.006908 **
## factor(hr)6          -1.26418    1.67954  -0.753 0.451646
## factor(hr)7           4.97007    1.67627   2.965 0.003031 **
## factor(hr)8          13.52560    1.67426   8.079 6.98e-16 ***
## factor(hr)9          19.95907    1.67601  11.909  < 2e-16 ***
## factor(hr)10         32.28594    1.68291  19.185  < 2e-16 ***
## factor(hr)11         42.18025    1.69523  24.882  < 2e-16 ***
## factor(hr)12         48.70783    1.70953  28.492  < 2e-16 ***
## factor(hr)13         50.78678    1.72118  29.507  < 2e-16 ***
## factor(hr)14         52.90802    1.73075  30.569  < 2e-16 ***
## factor(hr)15         51.84905    1.73406  29.900  < 2e-16 ***
## factor(hr)16         51.09817    1.73044  29.529  < 2e-16 ***
## factor(hr)17         53.04936    1.72045  30.835  < 2e-16 ***
## factor(hr)18         41.48321    1.70933  24.269  < 2e-16 ***
## factor(hr)19         31.23534    1.69352  18.444  < 2e-16 ***
## factor(hr)20         20.64977    1.68438  12.260  < 2e-16 ***
## factor(hr)21         14.33875    1.67760   8.547  < 2e-16 ***
## factor(hr)22          9.77425    1.67454   5.837 5.41e-09 ***
## factor(hr)23          4.14783    1.67321   2.479 0.013186 *
## factor(yr)1          11.84816    0.48976  24.192  < 2e-16 ***
## factor(mnth)2        -1.12660    1.22764  -0.918 0.358792
## factor(mnth)3        12.66155    1.26685   9.995  < 2e-16 ***
## factor(mnth)4        16.36907    1.37337  11.919  < 2e-16 ***
## factor(mnth)5        16.74737    1.58544  10.563  < 2e-16 ***
## factor(mnth)6         6.52208    1.76638   3.692 0.000223 ***
## factor(mnth)7        -0.22343    1.92222  -0.116 0.907467
## factor(mnth)8         4.44270    1.81739   2.445 0.014513 *
## factor(mnth)9        12.73823    1.64068   7.764 8.69e-15 ***
## factor(mnth)10       15.95634    1.40682  11.342  < 2e-16 ***
## factor(mnth)11        8.10861    1.25617   6.455 1.11e-10 ***
## factor(mnth)12        2.55527    1.21735   2.099 0.035828 *
## factor(weathersit)2  -3.49345    0.60101  -5.813 6.26e-09 ***
## factor(weathersit)3 -11.25360    1.01169 -11.124  < 2e-16 ***
## factor(weathersit)4  -3.08018   18.45667  -0.167 0.867461
## workingday          -33.62581    0.52285 -64.313  < 2e-16 ***
## newtemp               2.11754    0.07097  29.838  < 2e-16 ***
## newwindspeed         -0.27253    0.03194  -8.533  < 2e-16 ***
## newhum               -0.28545    0.01731 -16.490  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.89 on 17336 degrees of freedom
## Multiple R-squared:  0.5827, Adjusted R-squared:  0.5817
## F-statistic: 576.4 on 42 and 17336 DF,  p-value: < 2.2e-16
```

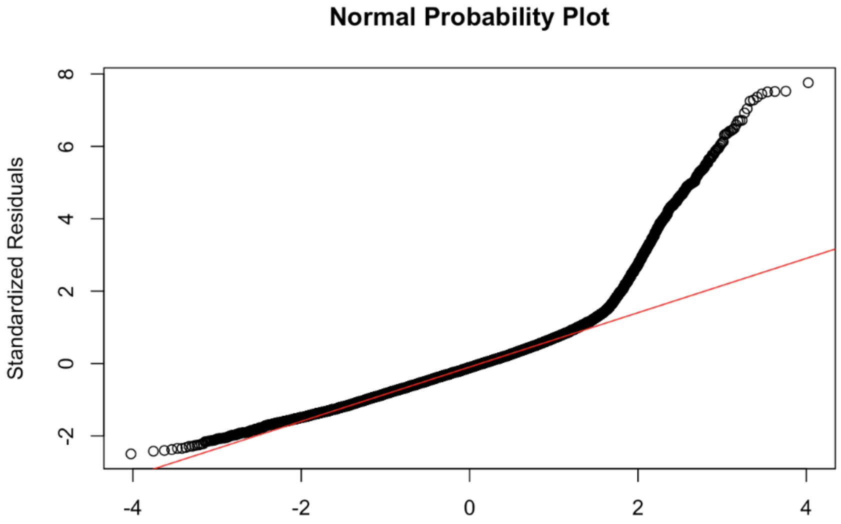
correlation

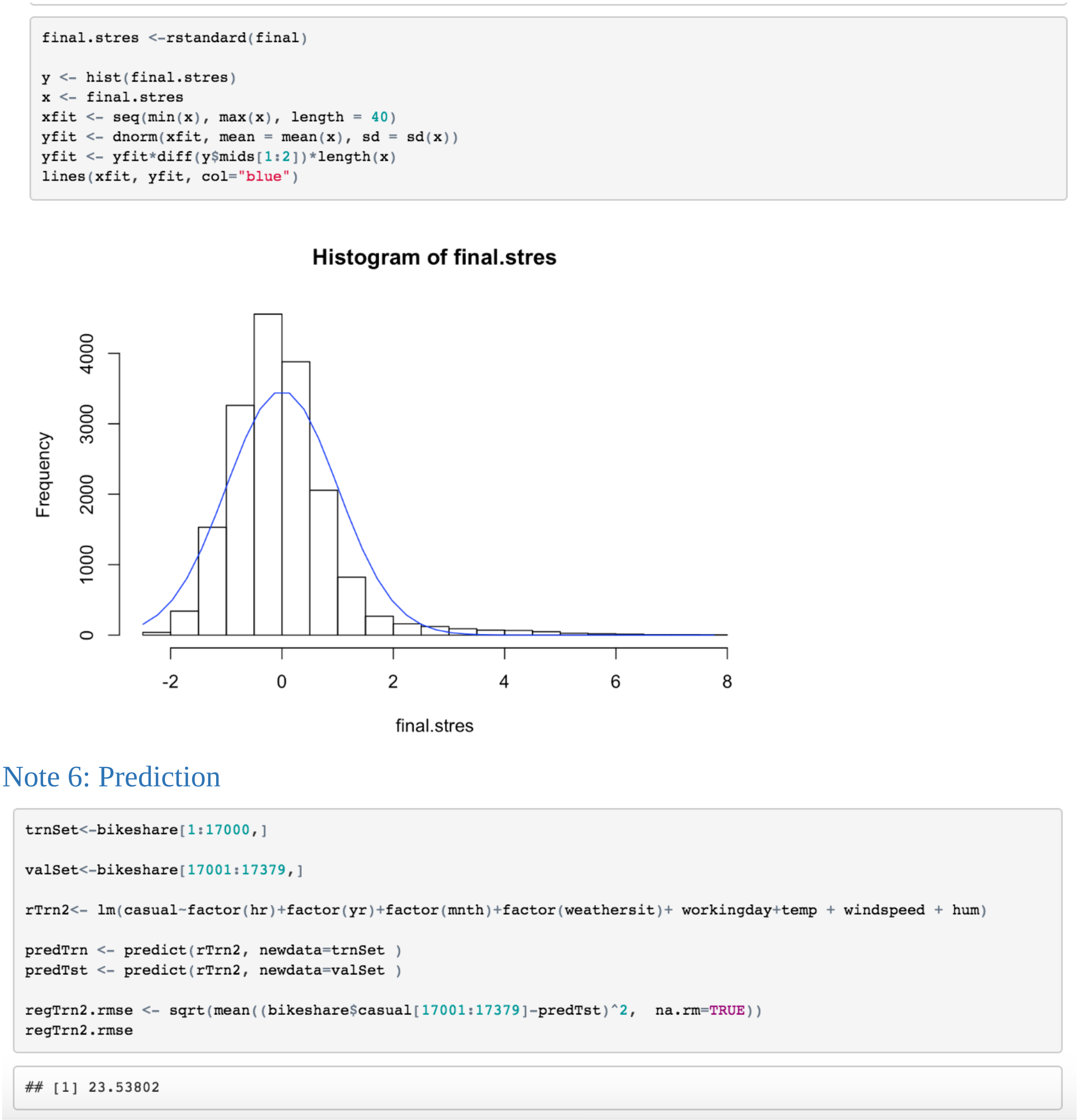
```
bikeshare<-read.csv("hour.csv")
attach(bikeshare)

cor(bikeshare[,11:17])
```

```
##                   temp       atemp         hum   windspeed      casual
## temp        1.00000000  0.98767214 -0.06988139 -0.02312526  0.45961565
## atemp       0.98767214  1.00000000 -0.05191770 -0.06233604  0.45408007
## hum        -0.06988139 -0.05191770  1.00000000 -0.29010490 -0.34702809
## windspeed  -0.02312526 -0.06233604 -0.29010490  1.00000000  0.09028678
## casual      0.45961565  0.45408007 -0.34702809  0.09028678  1.00000000
## registered  0.33536085  0.33255864 -0.27393312  0.08232085  0.50661770
## cnt         0.40477228  0.40092930 -0.32291074  0.09323378  0.69456408
##             registered         cnt
## temp        0.33536085  0.40477228
## atemp       0.33255864  0.40092930
## hum        -0.27393312 -0.32291074
## windspeed   0.08232085  0.09323378
## casual      0.50661770  0.69456408
## registered  1.00000000  0.97215073
## cnt         0.97215073  1.00000000
```

Note 5: Checking assumptions

```
qqnorm(final.stres, main = "Normal Probability Plot", xlab = "Normal Scores", ylab = "Standardized Residuals")
qqline(final.stres, col = "red")
```





```
final.stres <-rstandard(final)

y <- hist(final.stres)
x <- final.stres
xfit <- seq(min(x), max(x), length = 40)
yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))
yfit <- yfit*diff(y$mids[1:2])*length(x)
lines(xfit, yfit, col="blue")
```

Note 6: Prediction

```
trnSet<-bikeshare[1:17000,]

valSet<-bikeshare[17001:17379,]

rTrn2<- lm(casual~factor(hr)+factor(yr)+factor(mnth)+factor(weathersit)+ workingday+temp + windspeed + hum)

predTrn <- predict(rTrn2, newdata=trnSet )
predTst <- predict(rTrn2, newdata=valSet )

regTrn2.rmse <- sqrt(mean((bikeshare$casual[17001:17379]-predTst)^2,  na.rm=TRUE))
regTrn2.rmse
```

```
## [1] 23.53802
```