This report is about exploring the data set about investigate the relationship between power consumption, temperature in John Loane House and the weather in Dublin at 2018.

The dataset contain 8718 record and 16 variable.

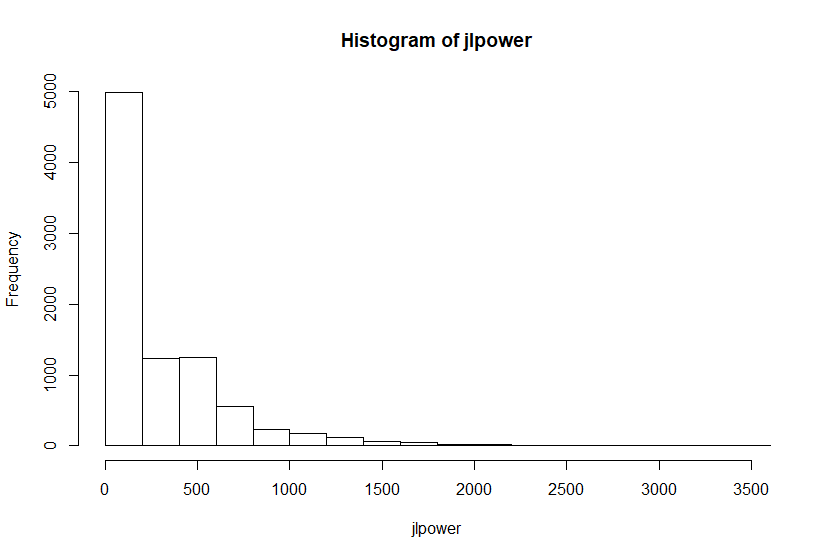
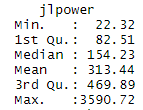
|  |  |  |
| --- | --- | --- |
| Jlpower | Numeric continoues | Power comsumption in John house |
| Jltemperature | Numeric continoues | Temperature record in john house |
| Rain | Numeric continoues | Precipitation Amount (mm) |
| Temp | Numeric continoues | Air Temperature ( ℃ ) |
| Wetb | Numeric continoues | Wet Bulb Air Temperature ( ℃ ) |
| Dewpt | Numeric continoues | Dew Point Air Temperature ( ℃ ) |
| Vappr | Numeric continoues | Vapour Pressure (hPa) |
| Rhum | Numeric continoues | Relative Humidity (%) |
| Msl | Numeric continoues | Mean Sea Level Pressure (hPa) |
| Wdsp | Numeric continoues | Mean Hourly Wind Speed (kt) |
| Wddir | Numeric continoues | Predominant Hourly wind Direction (kt) |
| Ww | Categorical Data | Synop Code Present Weather (0-99) |
| Sun | Numeric continoues | Sunshine duration (hour) |
| Vis | Numeric continoues | Visibility (m) |
| Clht | Numeric continoues | Cloud Ceiling Height (100s feet) |
| Clamt | Categorical Data | Cloud Amount (okta) |

In this dataset, got 42 missing data.

In this investigate report, I had remove all the indicator variable in the database and ‘w’ data set, because when exploring the dataset, the most of the data in 2018 is having the same indicator, and the Synop Code Past Weather should be 0-9, but in the dataset it is from 0-99.

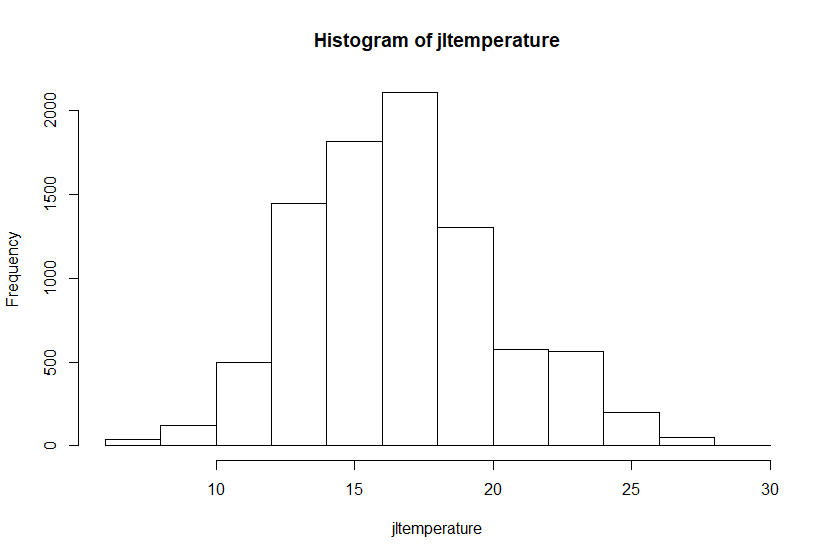
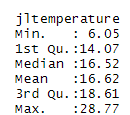
Following is go through every single variable.

Power:



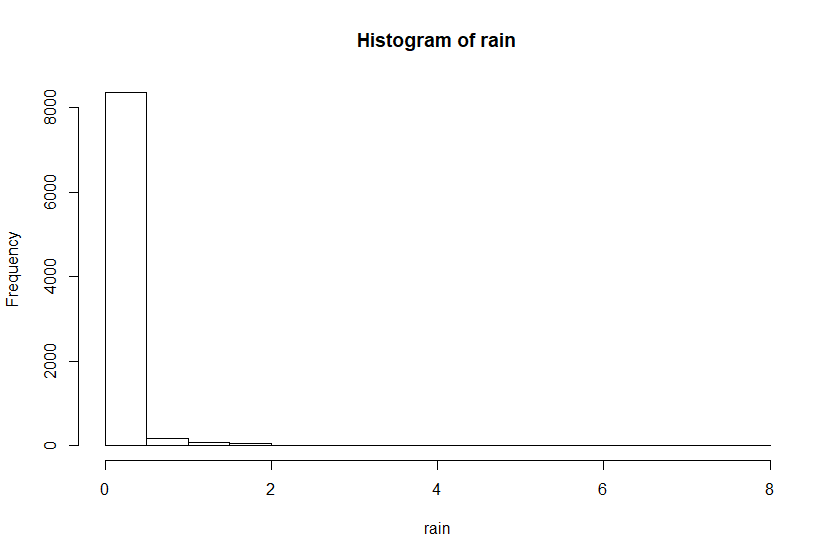
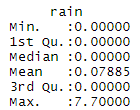
The power is strongly skewed right and have 408 outliers. The minimum is 22.32, median is 154.23 and mean is 313.44. The IQR is 315.66.

JLTemperature :



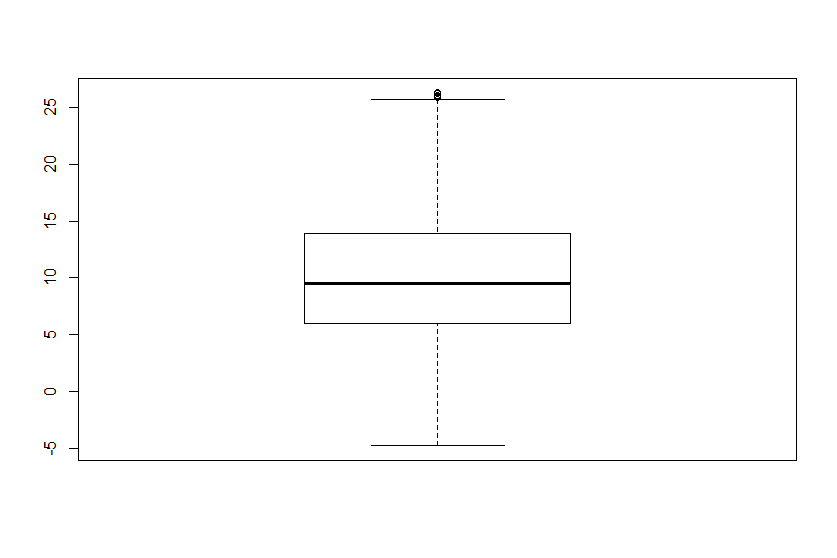
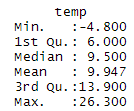
The temperature is roughly normal distributed and have 93 outliers. The minimum is 6.05, maximum is 28.77, median is 16.52 and mean 16.62. The IQR is 4.54.

Precipitation Amount :

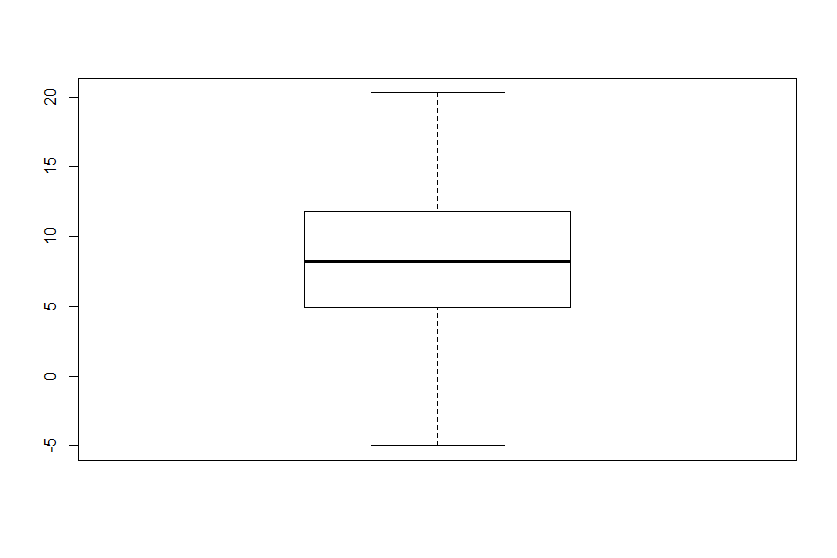
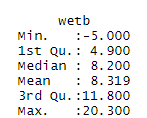


The rain amount is strongly skewed right with 1092 outliers, the min is 0, median is 0 , mean is 0.07885, max is 7.7 , the IQR is 0.

Temperature in weather station:

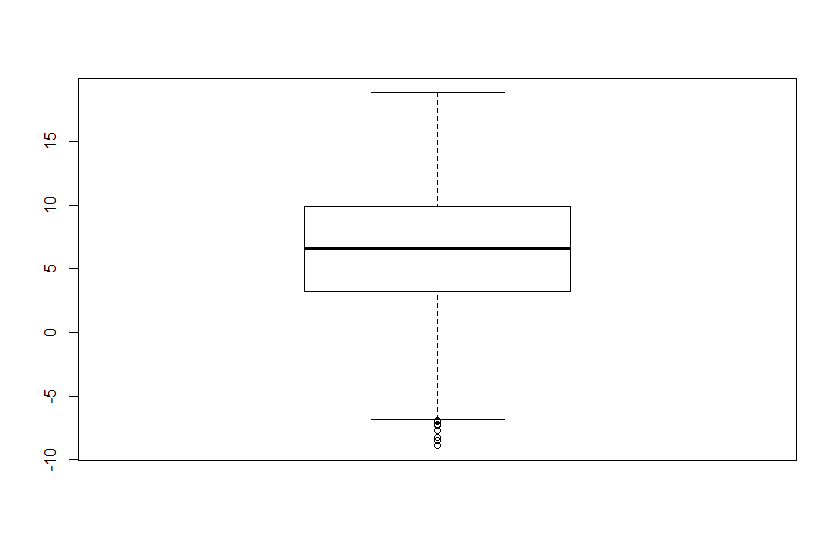
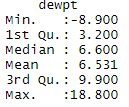


The temperature is normal distributed and 6 outliers. Min is -4.8, median is 9.5, mean is 9.947, max is 26.3. Standard deviation is 5.51.

Wet Bulb Air Temperature ( ℃ ): 

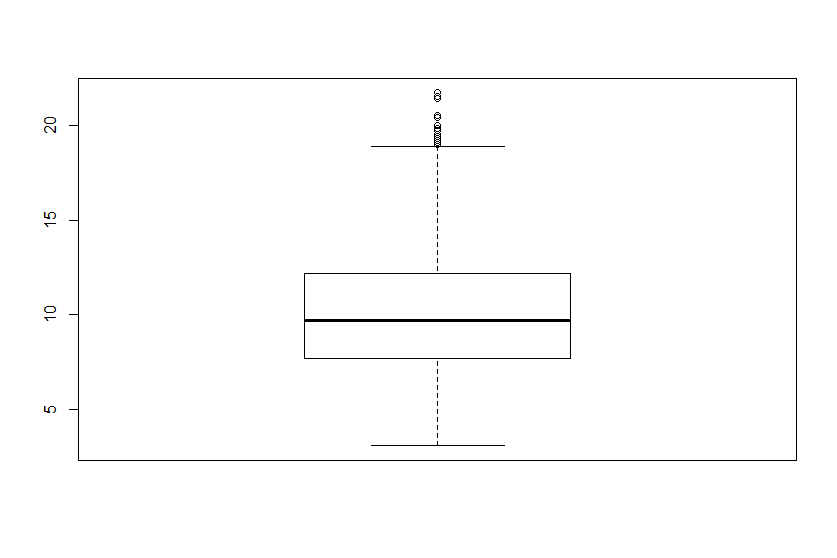
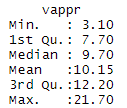
The wet bulb air temperature is normal distributed and no outliers. The min is -5 , median is 8.2, mean is 8.3 , max is 20.3. Standard Deviation is 4.72.

Dew Point Air Temperature ( ℃ ):



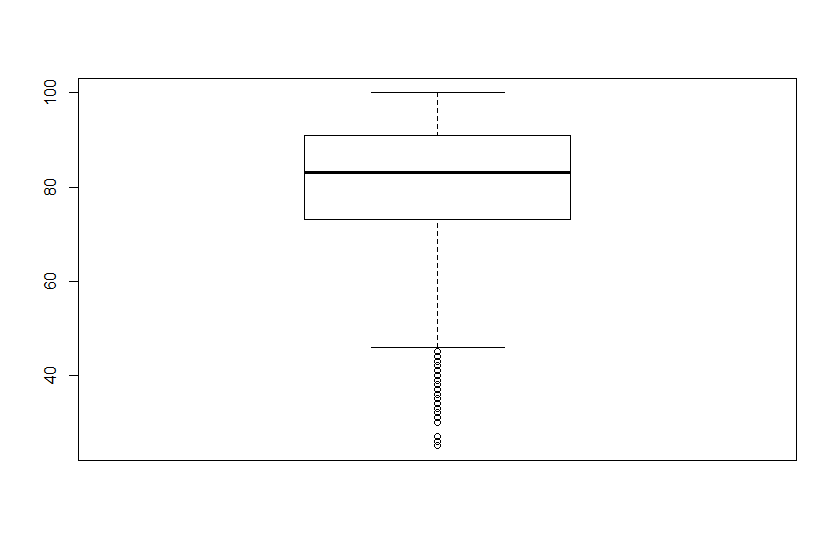
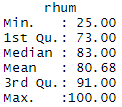
The dew point air temperature is normal distributed and 10 outliers. The min is -8.9, median is 6.6, mean is 6.531, max is 18.8. Standard deviation is 4.71.

Vapour Pressure (hPa):



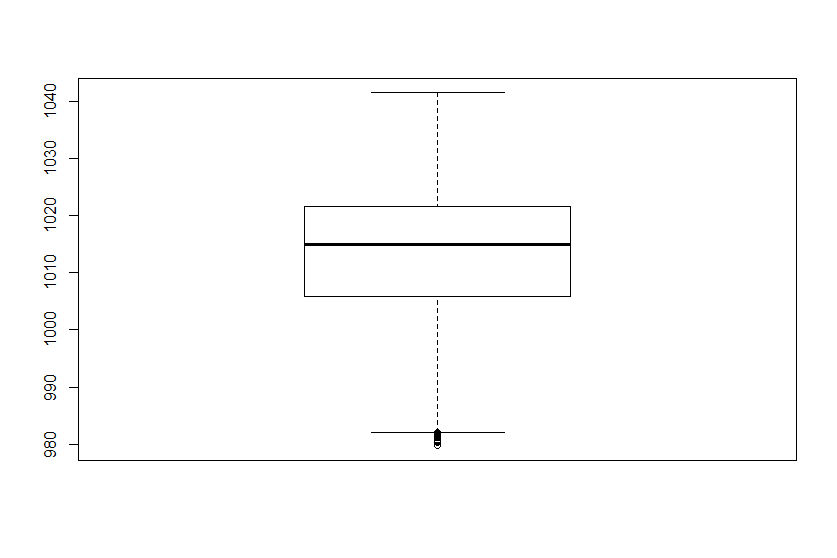
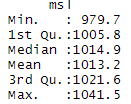
The vapour Pressure is normal distributed and contain 52 outliers. The min is 3.1 , median is 9.7, mean is 10.15, max is 21.7 and IQR is 4.5.

Relative Humidity (%)



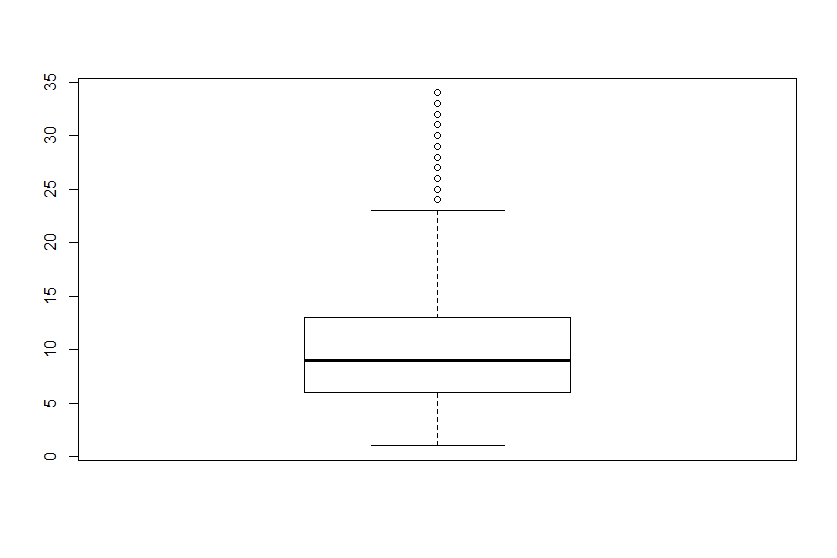
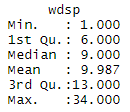
The relatiive humidity is normal distributed and contain 100 outliers. The min is 25% , median is 83% , mean is 80.68% and max is 100%. IQR is 18%.

Mean Sea Level Pressure (hPa):



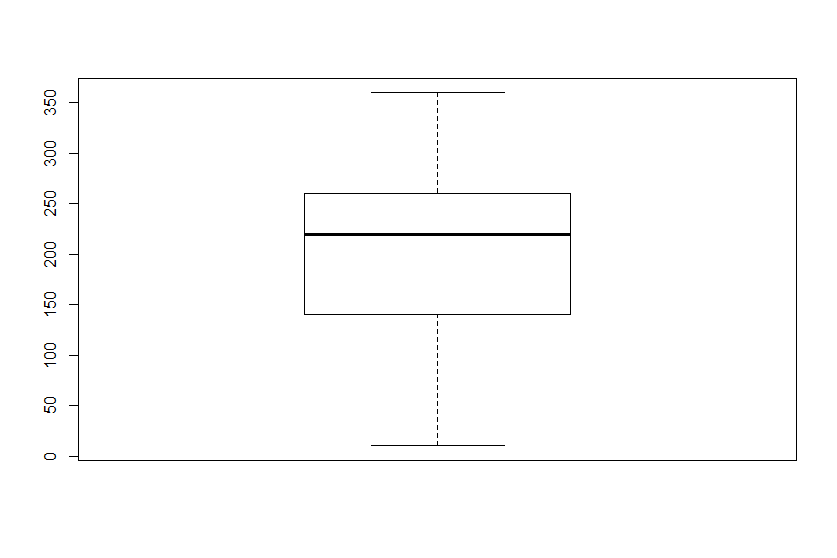
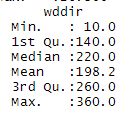
The mean sea level pressure is normal distributed and contain 19 outliers. The min is 979.7, median is 1014.9, mean is 1013.2, max is 1041.5. Standard deviation is 11.78.

Mean Hourly Wind Speed (kt):



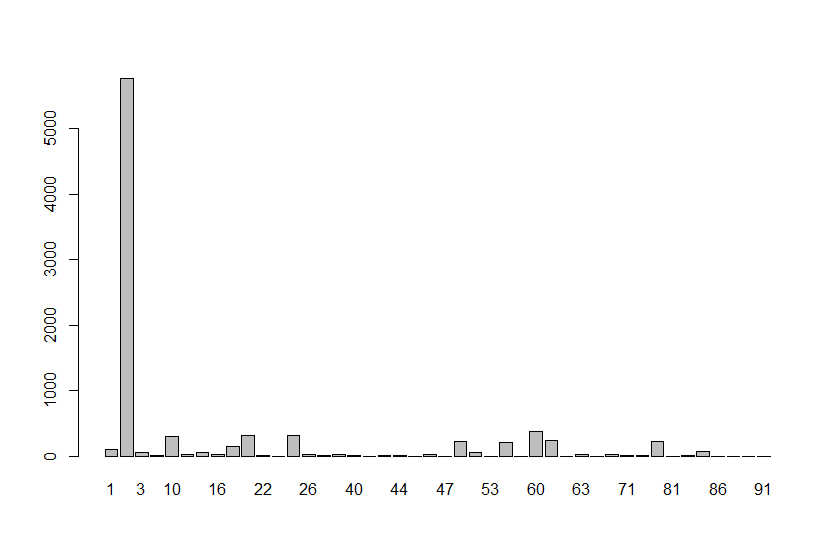
The mean hourly wind speed is normal distributed and contain 164 outliers. The min is 1, median Is 9 , mean is 9.987, and max is 34, IQR is 7.

Predominant Hourly wind Direction (kt):



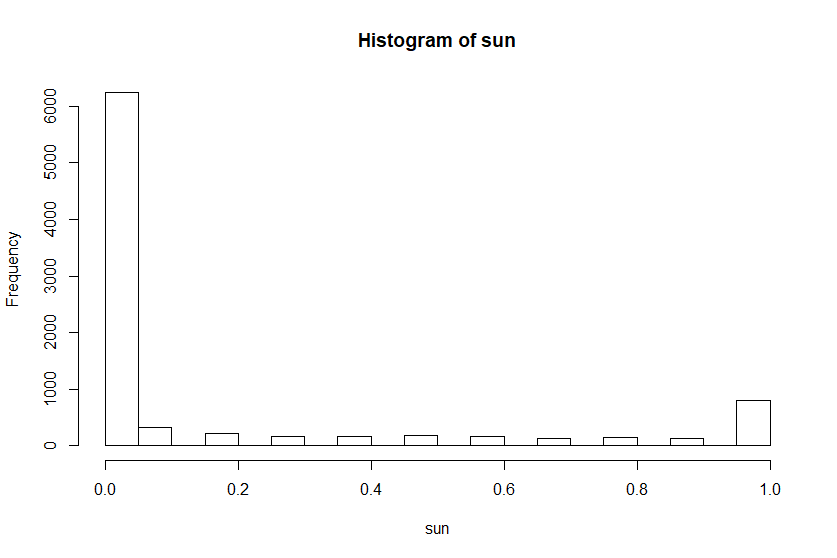
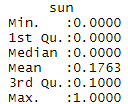
The predominant hourly wind direction is abit skewed left and don’t have outliers. The min is 10, median is 220, mean is 198.2, max is 360, standard deviation is 83.87.

Synop Code Present Weather (0-99)



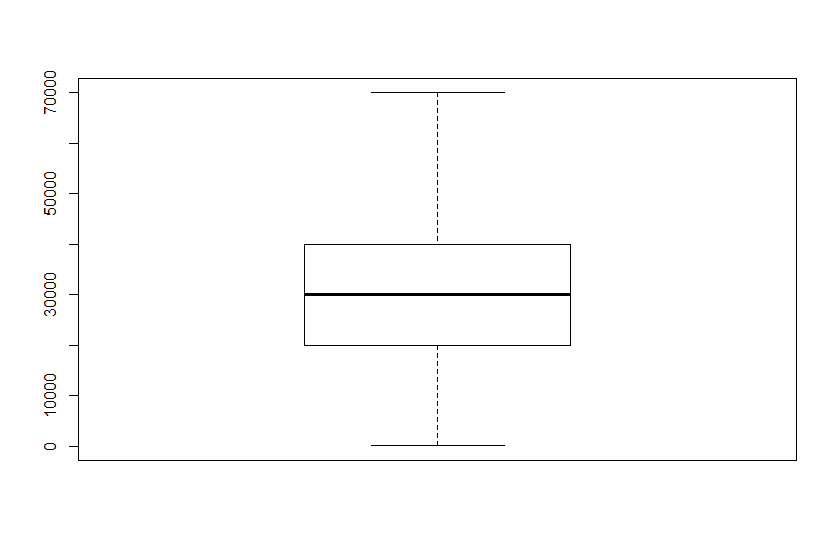
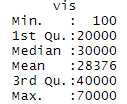
By the barplot , most of the record is with synop code 2, it mean State of sky on the whole unchanged.

Sunshine duration (hour)



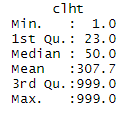
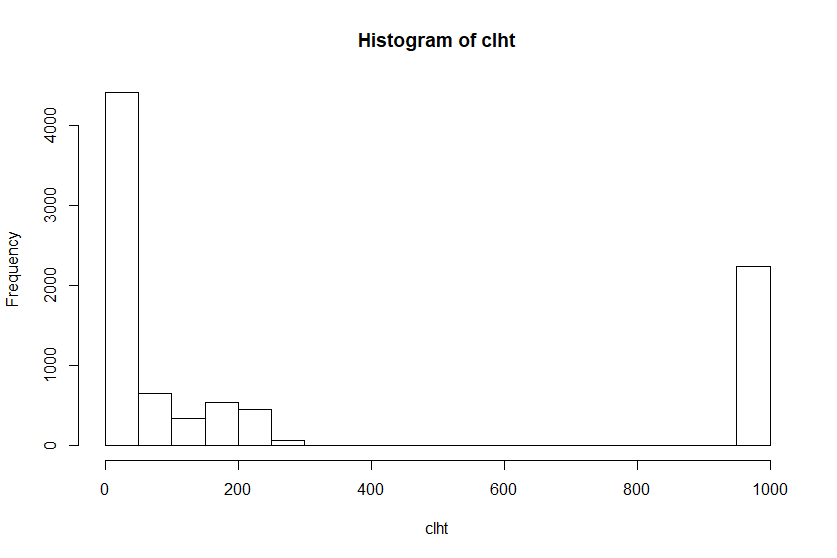
The sunshine duration is strongly skewed right and with 1923 outliers. The min is 0, median is 0, mean is 0.1763, max is 1. The IQR is 0.1.

Visibility (m):



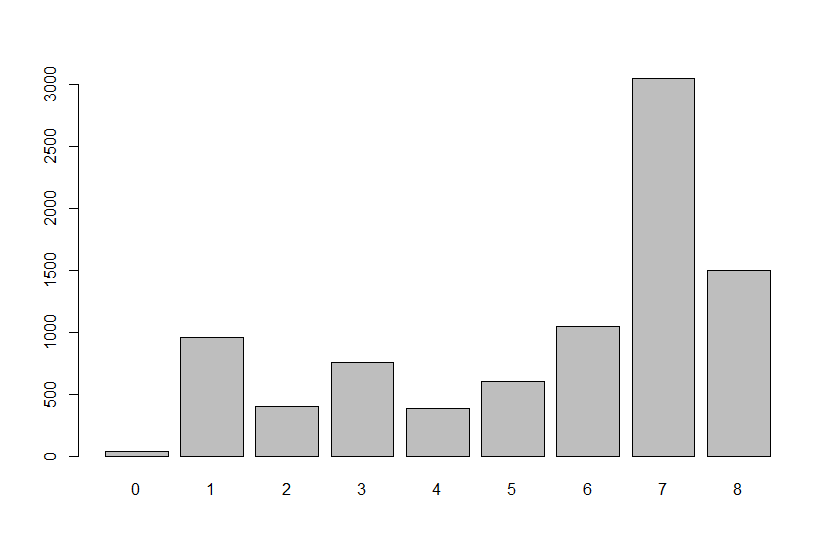
The visibility is normal distributed and don’t have outliers. The min is 100, median is 30000, mean is 28376, max is 70000, IQR is 20000.

Cloud Ceiling Height (100s feet):

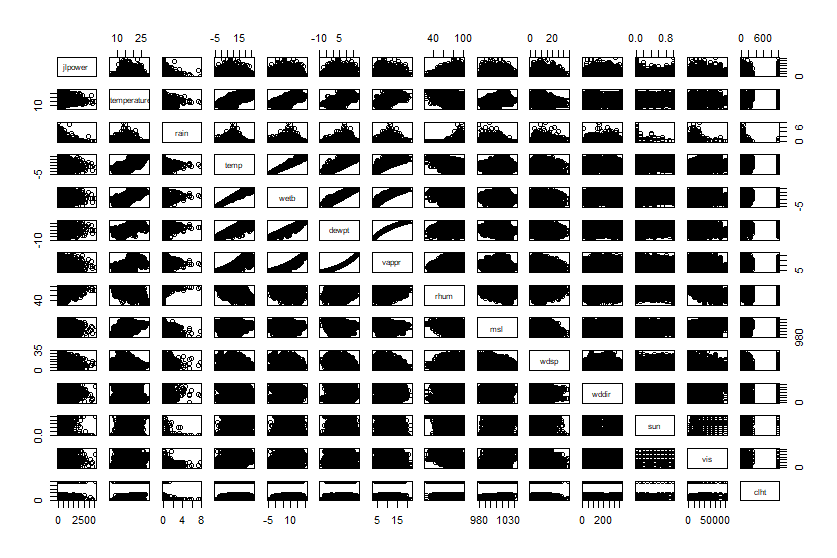
The cloud celling height is skewed right and no outliers if we condisering keep the 999 max height there. The min is 1, median is 50, mean is 307, and max is 999. Standard Deviation is 412.4403.

Cloud Amount (okta):

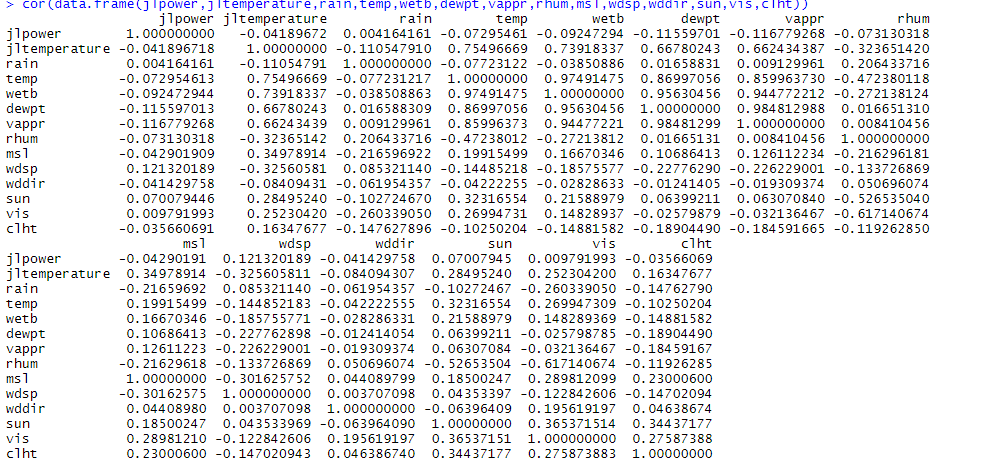


Most frequence status is 7okta ( the cloud almost cover the sky) and the less frequence status is 0 ( there is no cloud totally).

Pairs Plot:



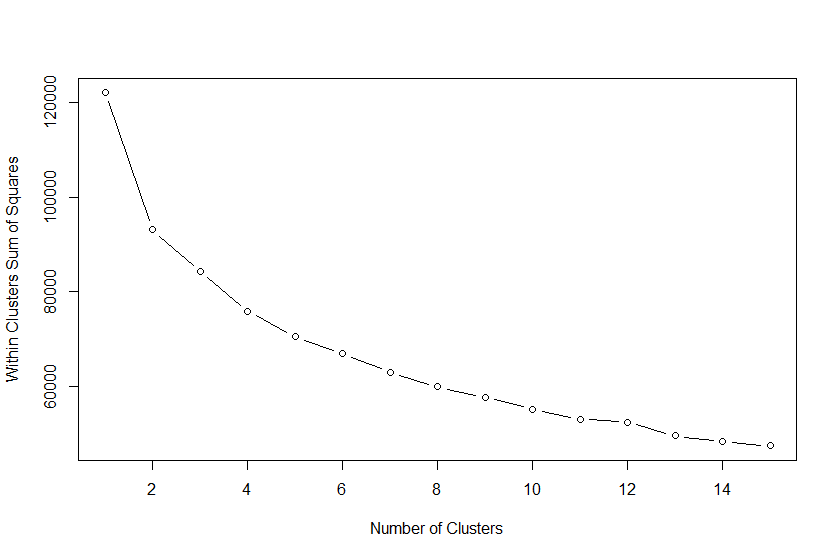
Correlation :



From the correlation resullt. The relationship between variable is

* JlTemperature and Air Temperature got strong positive relationship the correlation is 0.754967
* JlTemperature and webt got strong positive relationship the correlation is 0.7391834
* JlTemperature and dewpt got strong positive relationship the correlation is 0.6678024
* JlTemperature and vappr got strong positive relationship the correlation is 0.6624344
* JlTemperature and rhum got negative relationship the correlation is -0.3236514
* Jl temperature and msl got positive relationship, the correlation is 0.3497891
* JlTemperature and wdsp got negative relationship, the correlation is -0.3256058
* Temp and wetb got strong positive relationship, the correlation is 0.9749147
* Temp and dewpt got strong positive relationship, the correlation is 0.8699706
* Temp and vappr got strong positive relationship, the correlation is 0.8874733
* Temp and rhum got negative relationship, the correlation is -0.4723801
* Dewpt and vappr got strong positive relationship, the correlation is 0.9998436
* Rhum and sun got negative relationship,the correlation is -0.526535
* Rhun and vis got negative relationship, the correlation is -0.6171407

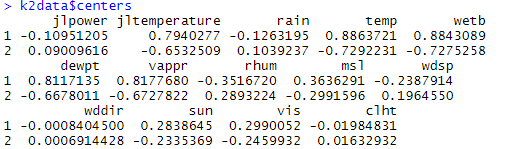




From the graph, I think there is 2 cluster in this data set.

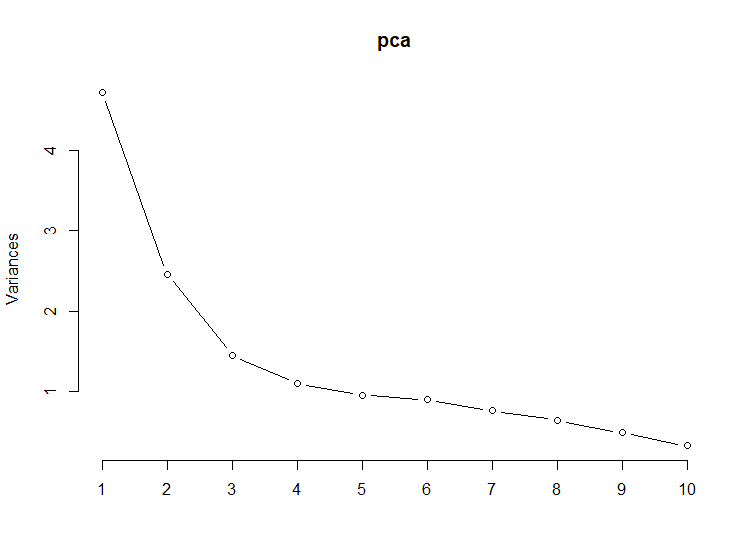
By running the k-mean analysis when k =2 , the result is

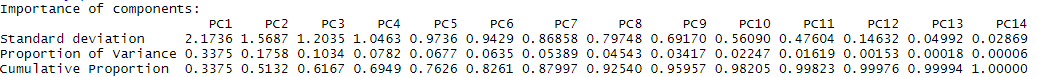


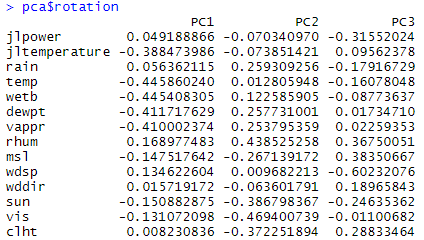
The cluster one got slightly low power from average, high jltemperature from average, slightly low rain from average, high temperature from average, high wetb from average, high dewpt from average, high vappr from average, low rhum from average, high msl from average, low wdsp from average, around average wddir, high sun from average, high vis from average, around average clht.

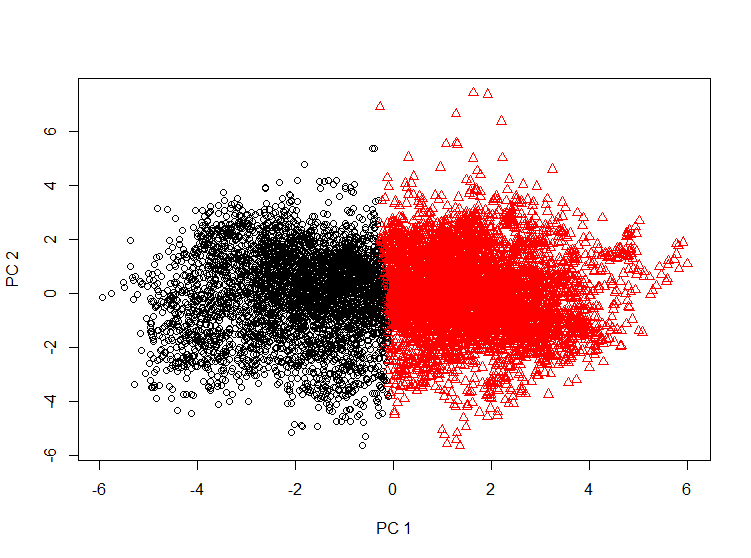
Cluster 2 got around average jlpower,wddir,clht. Higher than average rain,rhum,wdsp, Lower than average jltemperature,temp,wetb,dewpt,vappr,msl,sun and vis.

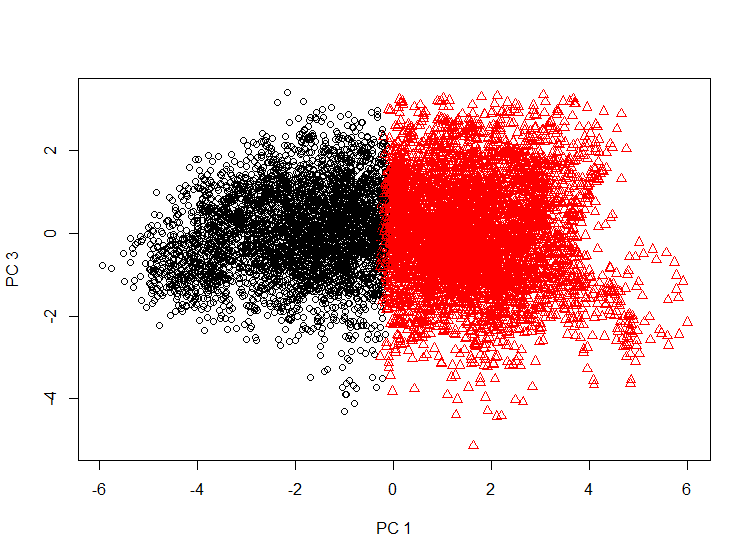


From the scree plot, I pick to reduce the dataset to 3 pca.

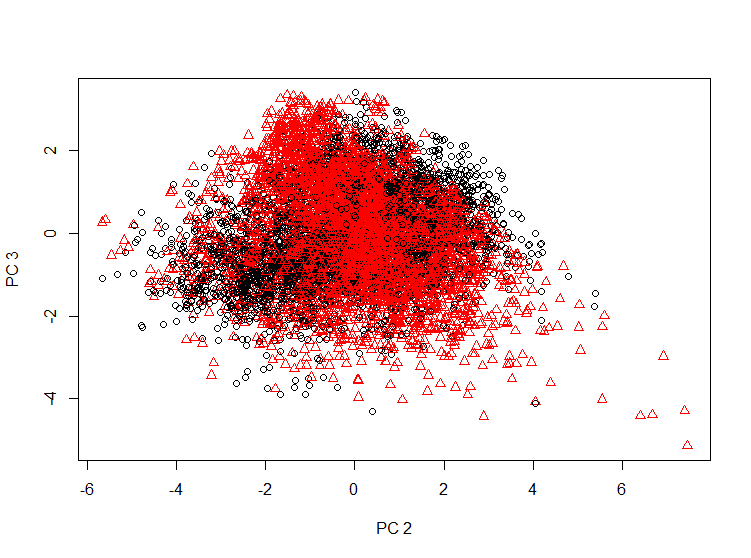
3 PCA explain 61.67% of variance.



From the pca rotation, we can see the difference from PC 1 and PC 2 is PC 1 got higher jlpower,sun, msl, vis and clht, lower jltemperature, rain, temp, wetb, dewpt, vappr, rhum.

Difference between PC 1 and PC 3 is PC 1 got higher power,rain,wdsp and sun, lower in jltemperature,temp,wetb,dewpt,vappr,rhum,msl,wddir,vis and clht.

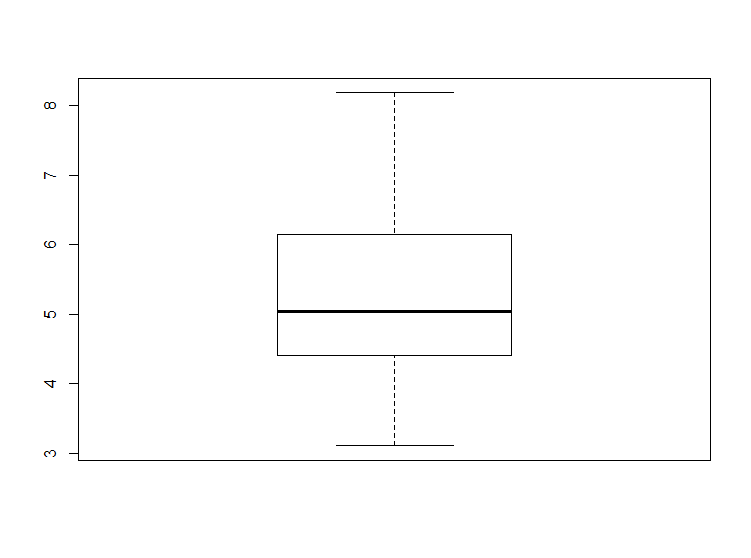
Difference between PC 2 and PC 3 is PC2 got higher power,jltemperature, rain,temp,wetb,dewpt,vappr,rhum,wdsp, lower wddir,sun,vis and clht.



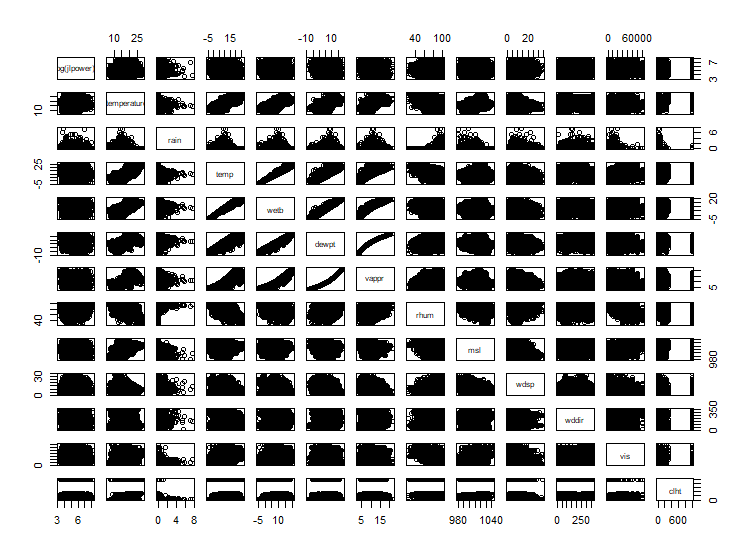
Simple Linear Regression

To find is there any variable will affect the power comsumption.

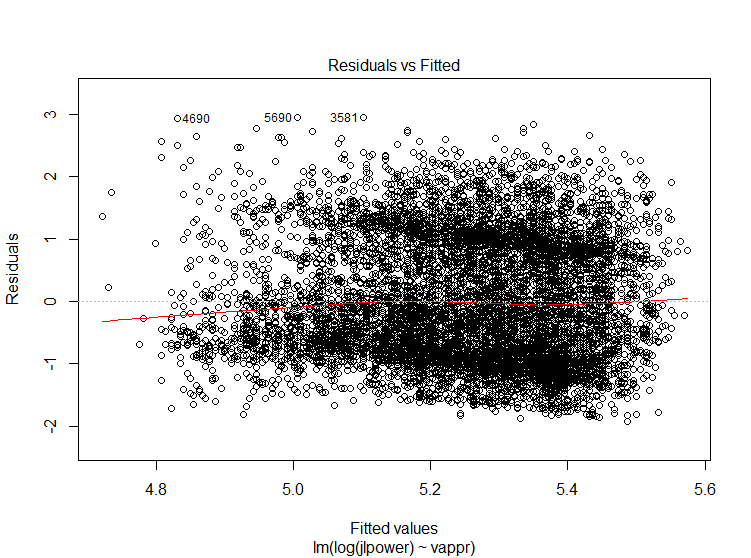
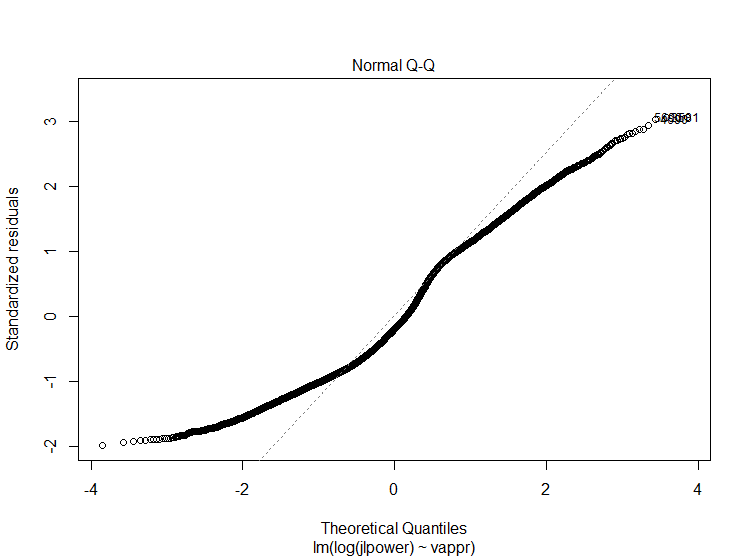
Since the power variable is not normal distributed, I log the power variable and make it be normal distributed.



Boxplot of Power After Log

Pair Plot for log(power) with other variable

The variable I pick for simple linear regression is vappr.

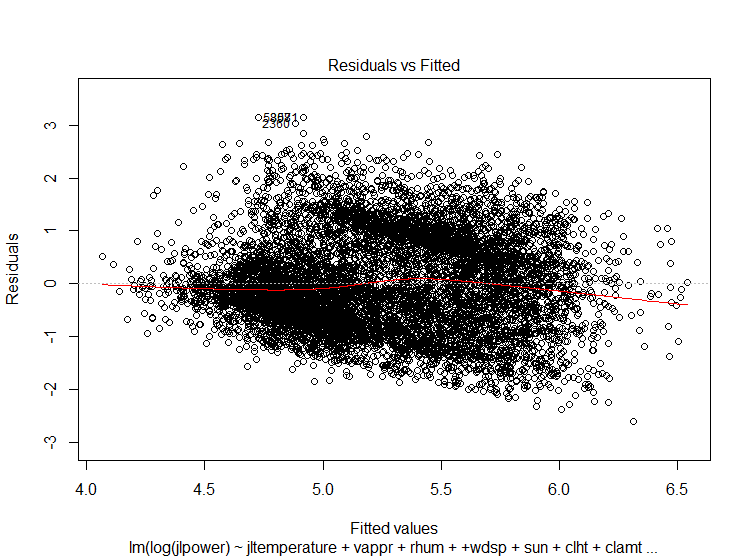
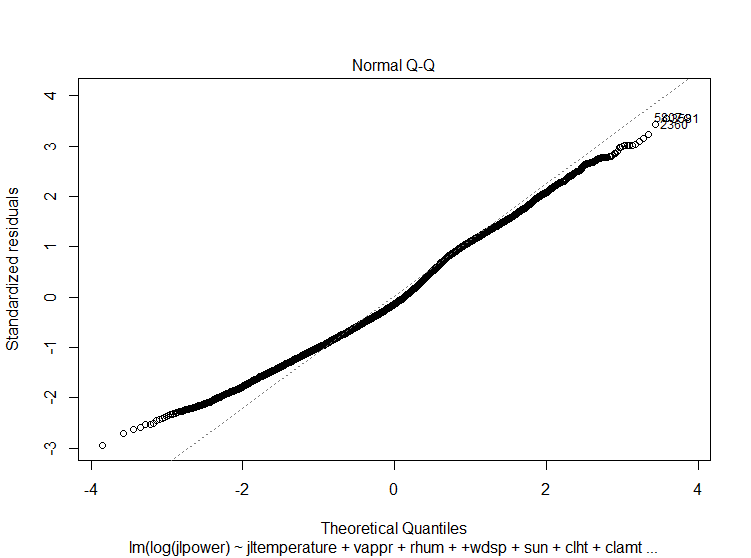
From the fitted vs rediduels plot, the iid and random scatter assumption is met. 

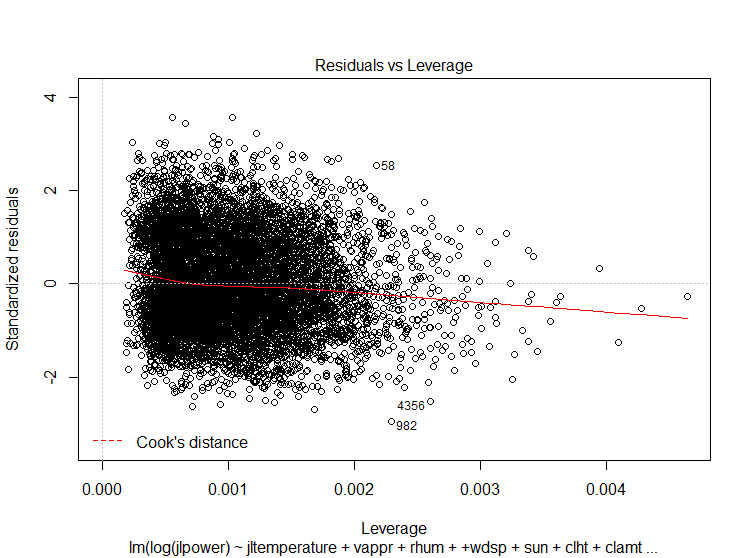
From the Q-Q plot, the normality assumption is violate. So I stop the simple linear regression analysis here.

Multiple Linear Regression.



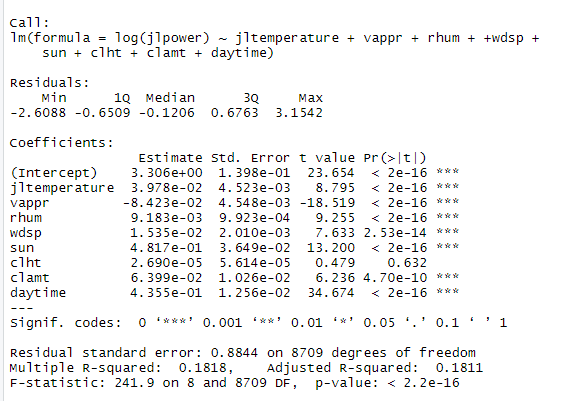
\*The daytime variable here is define by the hour of the record.

 \*Define as 0-9 = 1 , 9-18 = 2 , 18-24 = 3



The residual and fitted plot fix the iid assumption, the q-q plot meet the normality assumption, and the residuals and leverage plot show there don’t have outliers will heavily affect the result.

The result is



From this model, the residual standard error is 0.8844, and it explain 18.18% of variance with P value less than 0.05.

The formula from this model is

log(jlpower) = 3.306 + 0.03978\*jltemperature -0.08423 \*vappr + 0.009183 \* rhum +0.01535 \* wdsp + 0.04817 \* sun + 0.0000269 \* clht + 0.06399 \* clamt + 0.4355 \* daytime

So the power will be 27.2758038 while all variable is 0.

The power will increase by 1.48854629 for each unit jltemperature increase when all other variable stay constant.

The power will decrease by 1.08787908 for each unit increase when all other variable stay constant.

The power will increase by 1.00922529 for each unit rhum increase when all other variable stay constant.

The power will increase by 1.01546842 for each unit wdsp increase when all other variable stay constant.

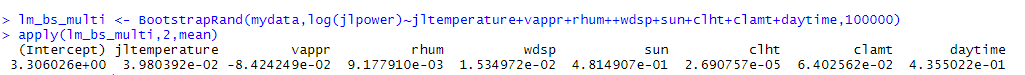
The power will increase by 1.04934903 for each unit sun increase when all other variable stay constant.

The power will increase by 1.06608174 for each unit clamt increase when all other variable stay constant.

The power will increse by 1.0000269 for each unit clht increase when all other variable stay constant.

The power will increase by 1.54573573 for each unit daytime increase when all other variable stay constant.

Althought the normality assumption is met, but the Q-Q plot head and tail is not actual on the line, so I do a 100000 sample bootstrapping to check the result I get.



And the result is roughly same with the model I get.

The model explain only 18% of variance, so I believe the dataset is not big enough to explain the power consumption, there must be other variable that will affect the power.