As explained in “instruction” file you have to run first this files:

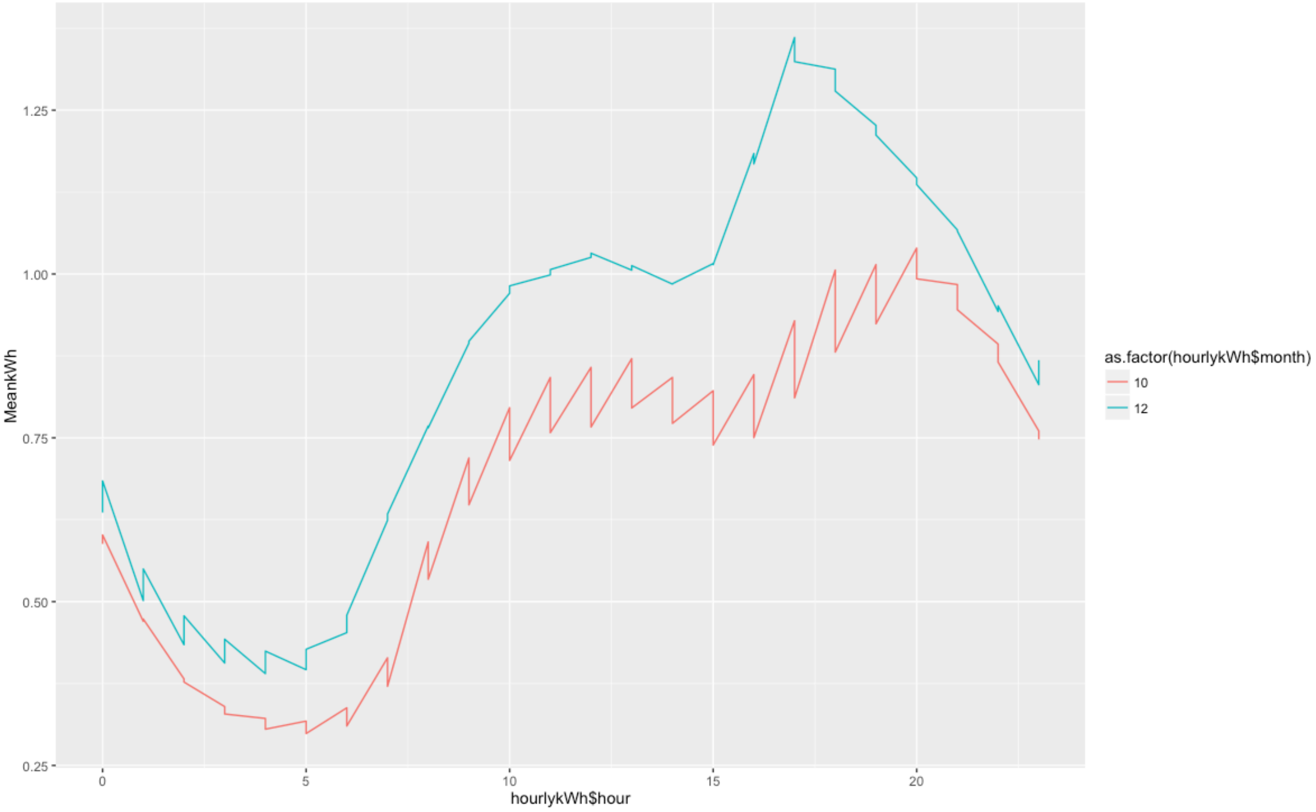
1- Run "Packages.R" -
2- Split.R- will do split all data file to csv file according to the your input
3- ReFormatingData.R - will be transform all data into readable format (data-time,...)
after executing this file program will be create *“FinalSample_01.csv”.*
we created this file because reading and reformatting data every time it’s difficult and costly so we decided to reformatting file just one time!
After running this file be sure you have *“FinalSample_01.csv” in directory which you defined in this file.*
If you have *FinalSample_01.csv file do not run FinalSample_01.csv again!!!*

Then run this file: *AnalyseData_01.R* , at first lines of this file we importing data to our main data frame!
mainData <- fread("/Volumes/DISK_IN/Projects/FinalSample_01.csv")

mainData is our main data set, this will be use whole program!

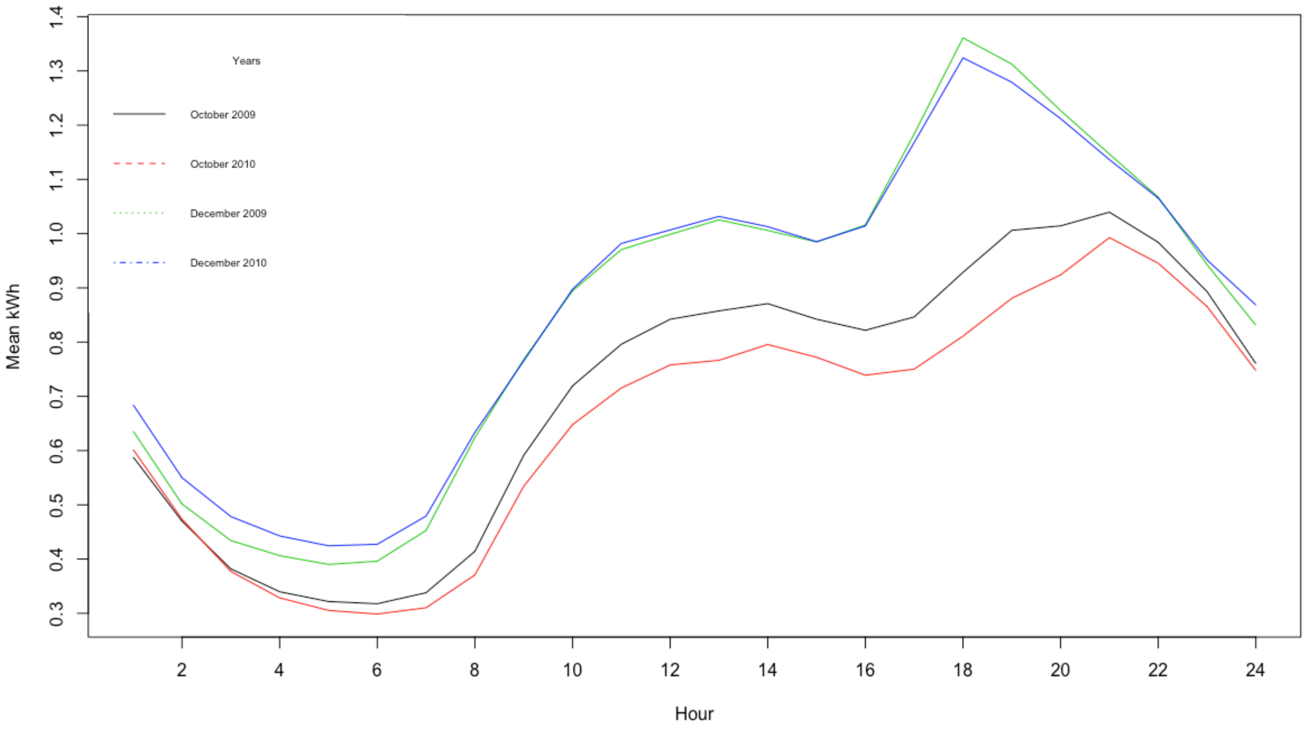
Tip: of course you have change directory which you saved file!

So, *FinalSample_01.csv* giving us information about means of KWh according to the hour, month, year (mean per half hour by month & year)
After I run this script my result is like this:



as you see it’s based on month(10,12) and mean Kwh

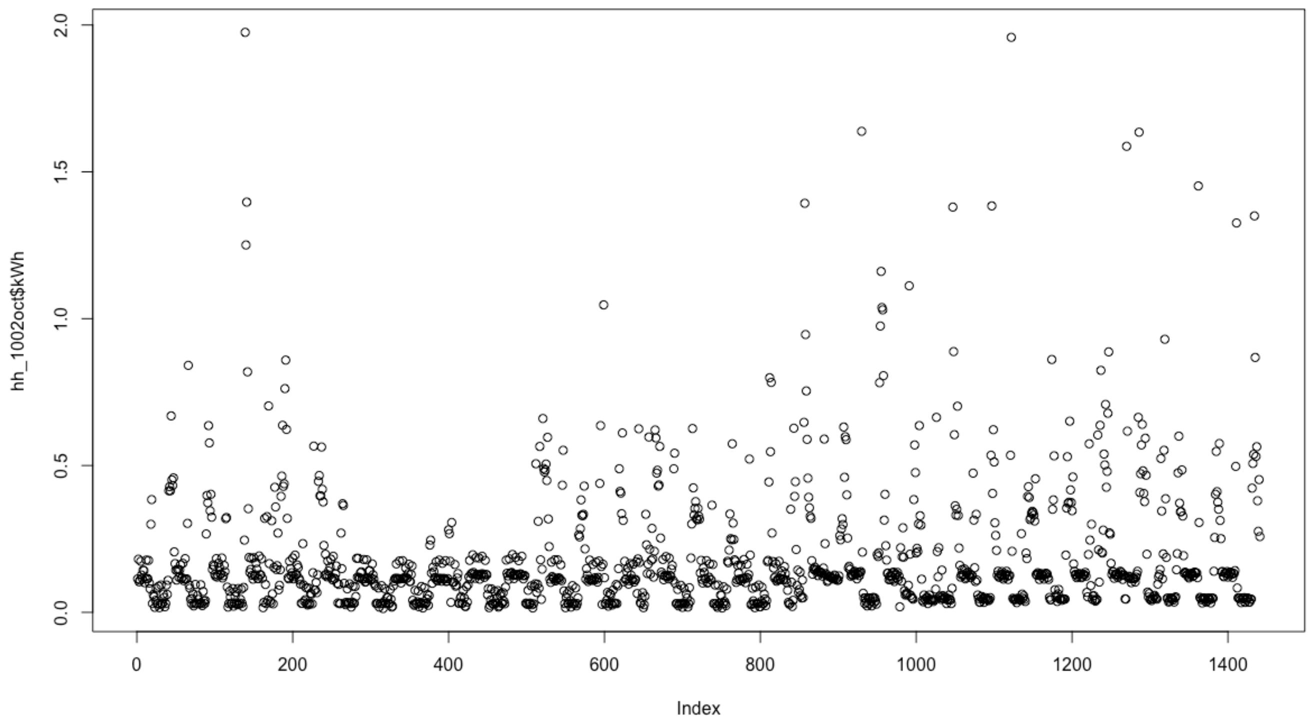
and I have this:



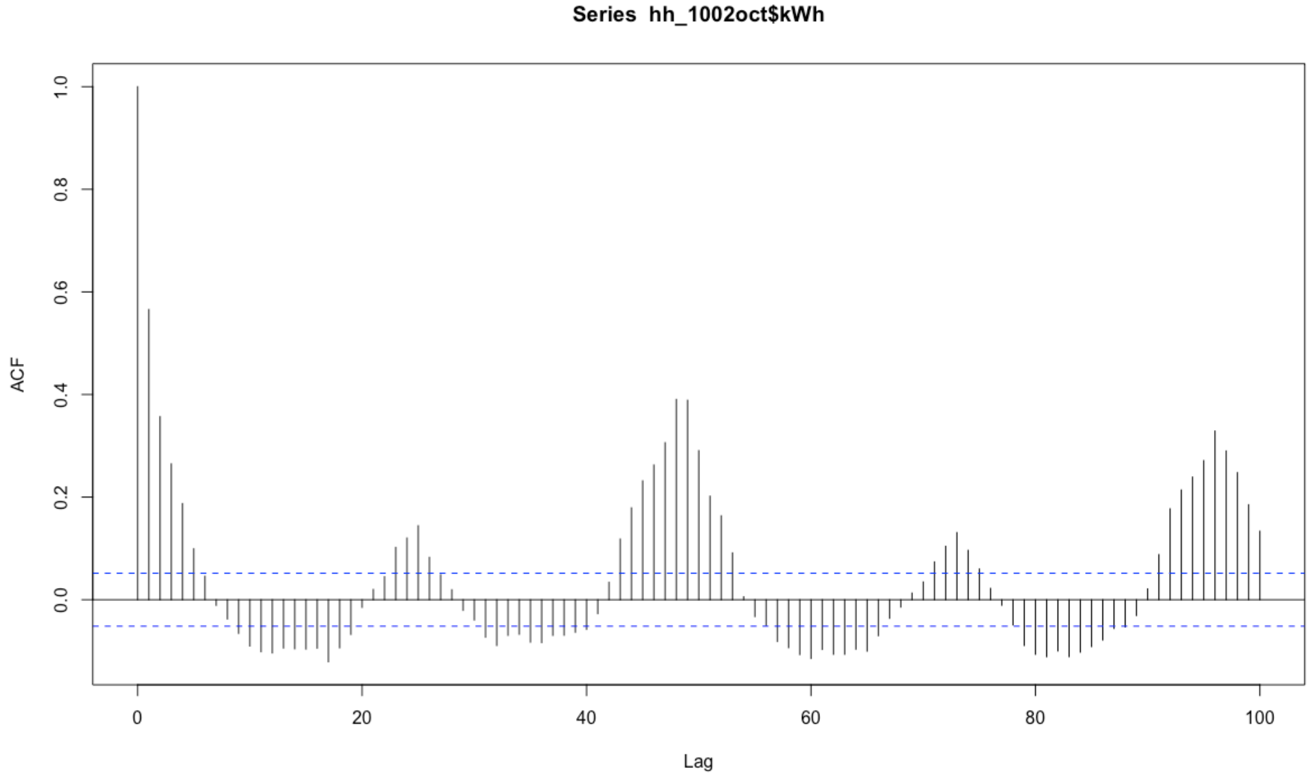
```
# calculate mean kWh consumption for evening peak 16:00 - 19:00
# October 2009 (pre trial)

  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0       0.1690  0.4220  0.9489  1.0740 37.5000

# sample size calculation
# see ?power.t.test
# let us assume we see a 10% change in the mean due to higher evening prices
# we want to test for a reduction (1 sided)
# for now assume this is a two sample test

Results :

Two-sample t test power calculation

              n = 3987.655
          delta = 0.0948859
             sd = 1.703823
      sig.level = 0.05
          power = 0.8
    alternative = one.sided

NOTE: n is number in *each* group
```

```
# Time series analysis ----
# create a subset for the household number 7443 only ( it’s example)
# select just October - you'll see why in a minute
##hh_1024oct <- subset(hh_1024, hh_1024$oct == 1)
```

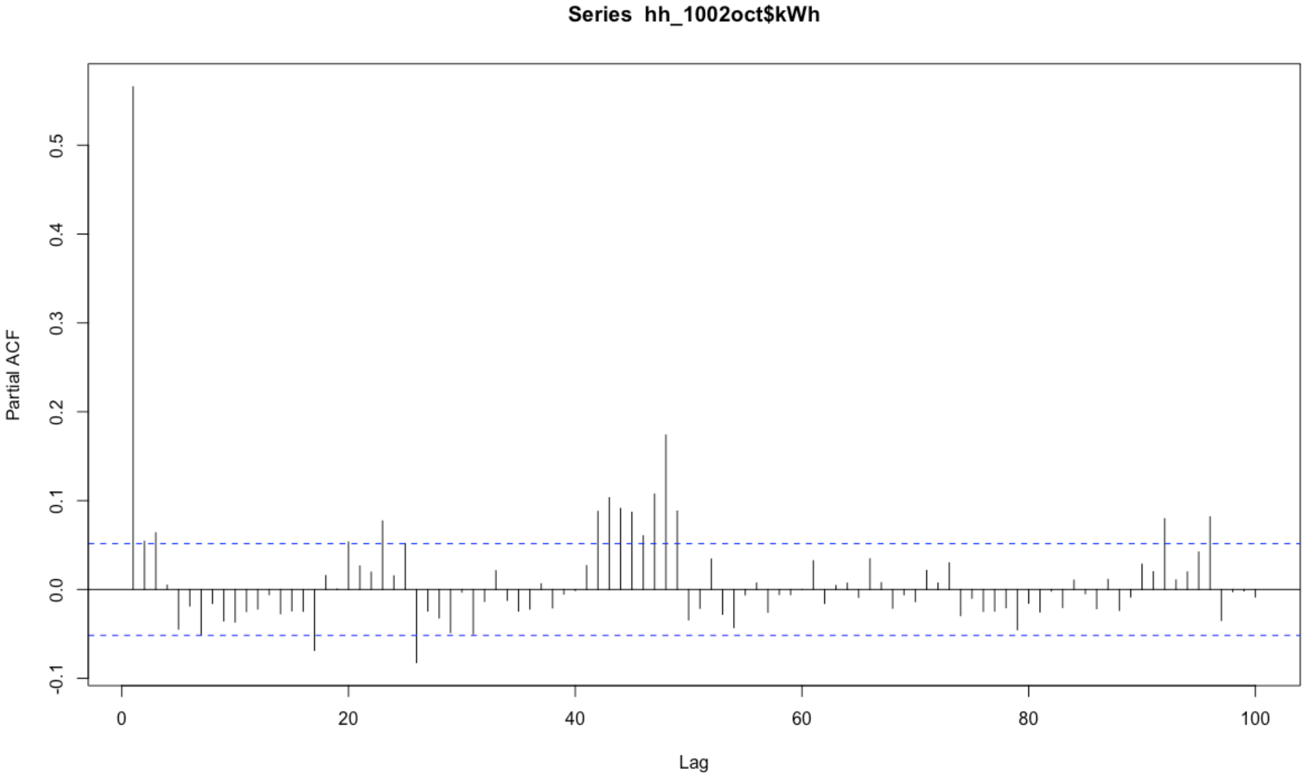
```
#L: we can select data from the dataframe by date range
# need to check is sorted by datetime (always increasing) and evenly spaced
# create zoo (time series) object to do this
```



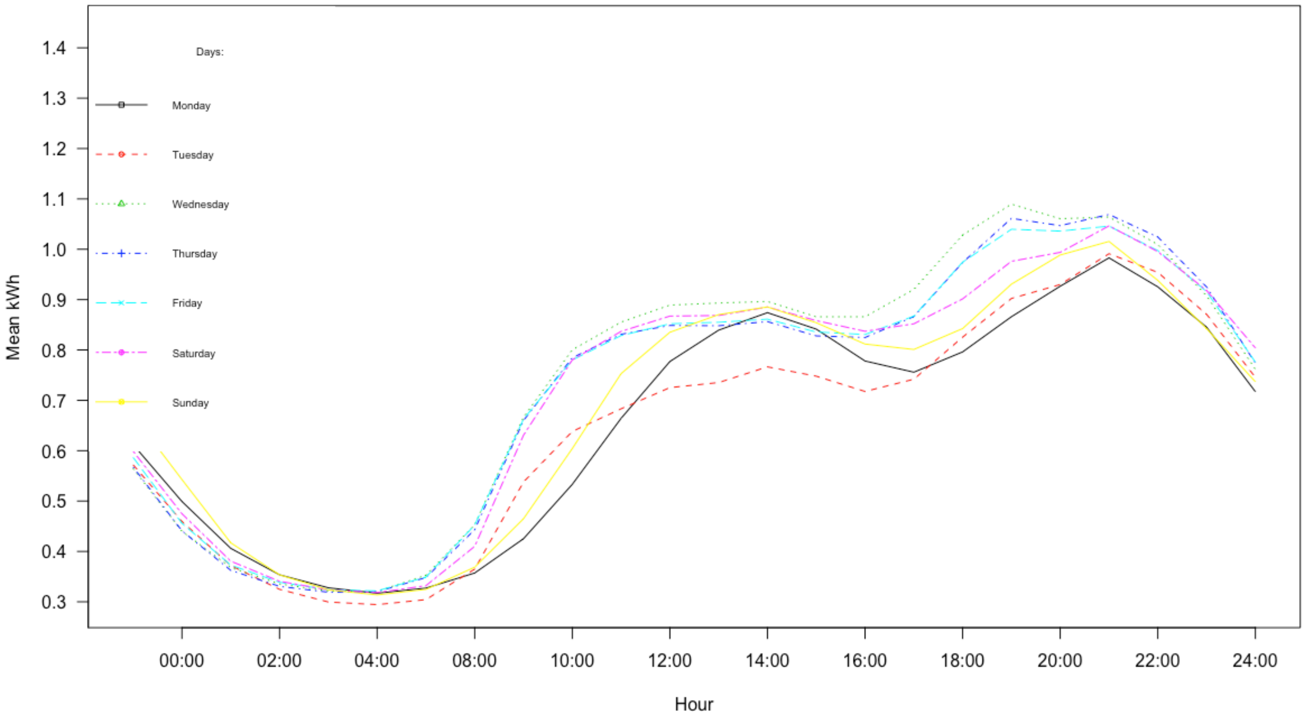
```
# run acf with the first household only up to just over 48 hours (96 half hours)
acf(hh_1002oct$kWh, lag.max = 100)
```



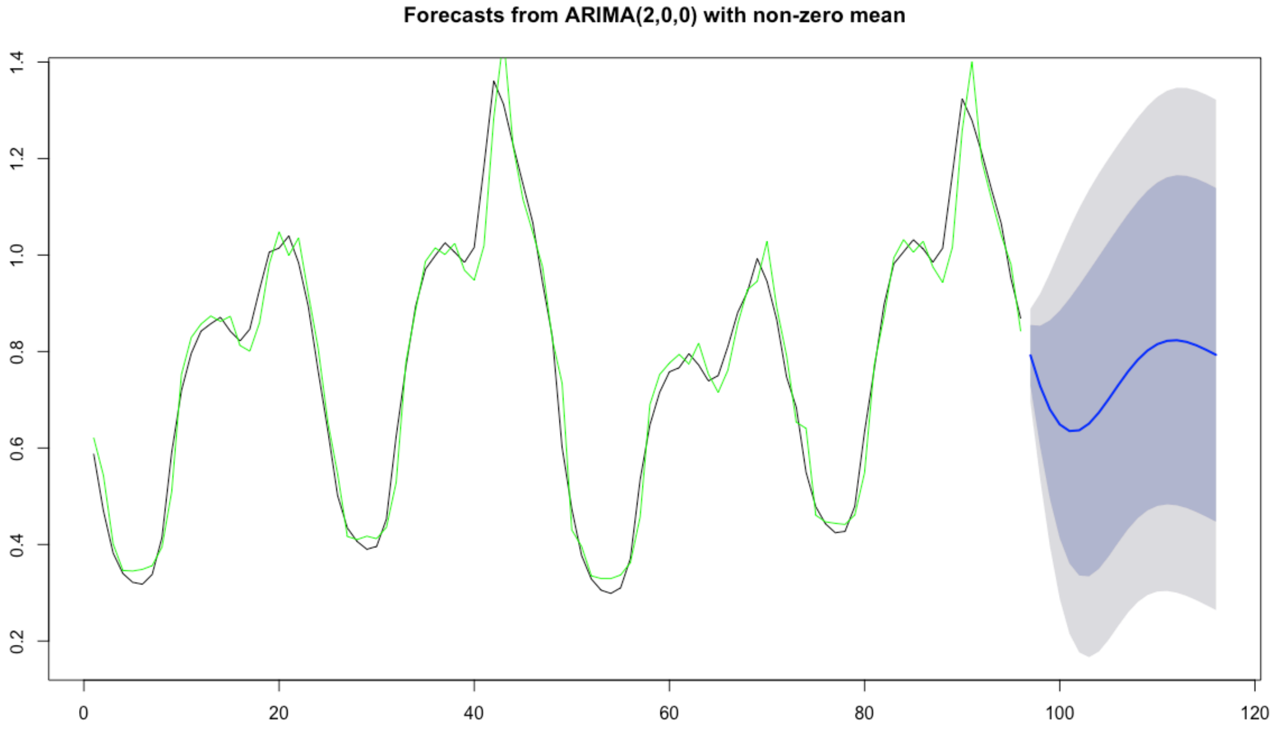
# let's find the *partial autocorrelation function* (pacf)
# this is the effect of successive lags with the effect of previous lags removed
# It shows more clearly how the random variation depends on the previous lags
# see https://www.youtube.com/watch?v=R-oWTWdS1Jg

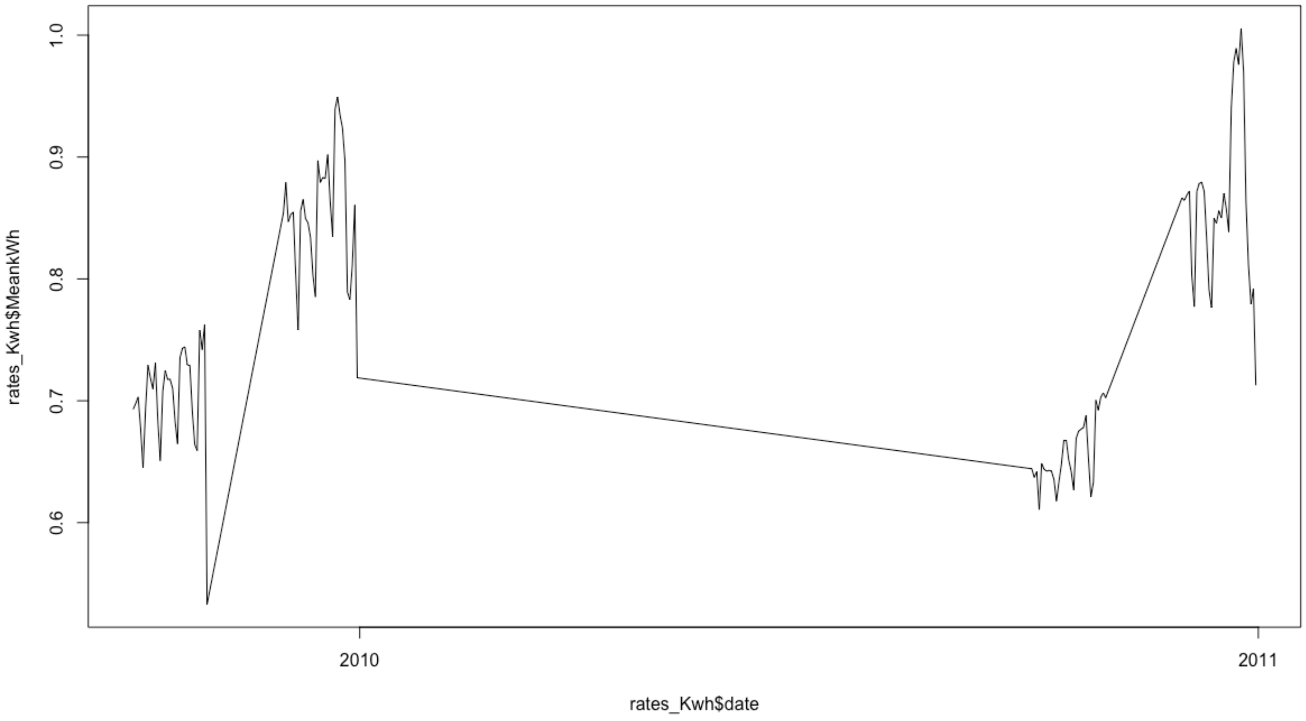


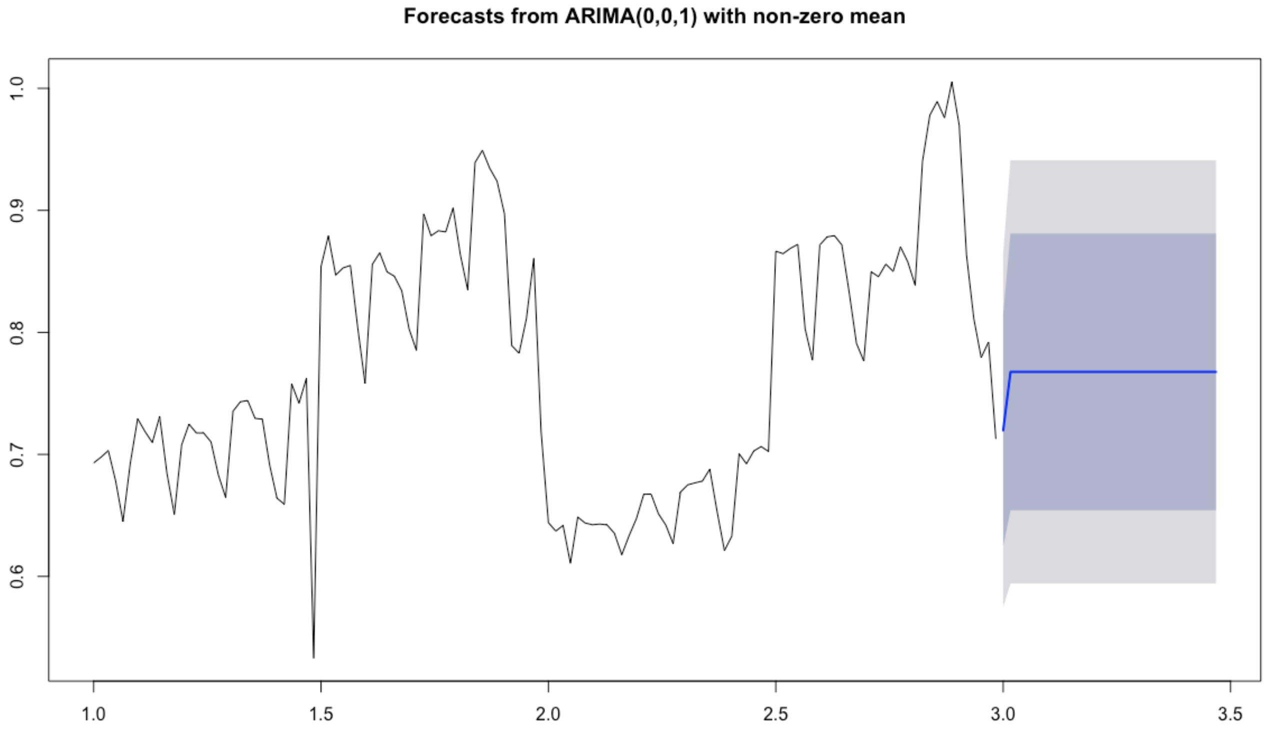
in *AnalyseData_02.R we are plotting data with special date-time to analyzing.*
as you see this is based on week days and means of hour KWh (24 h)



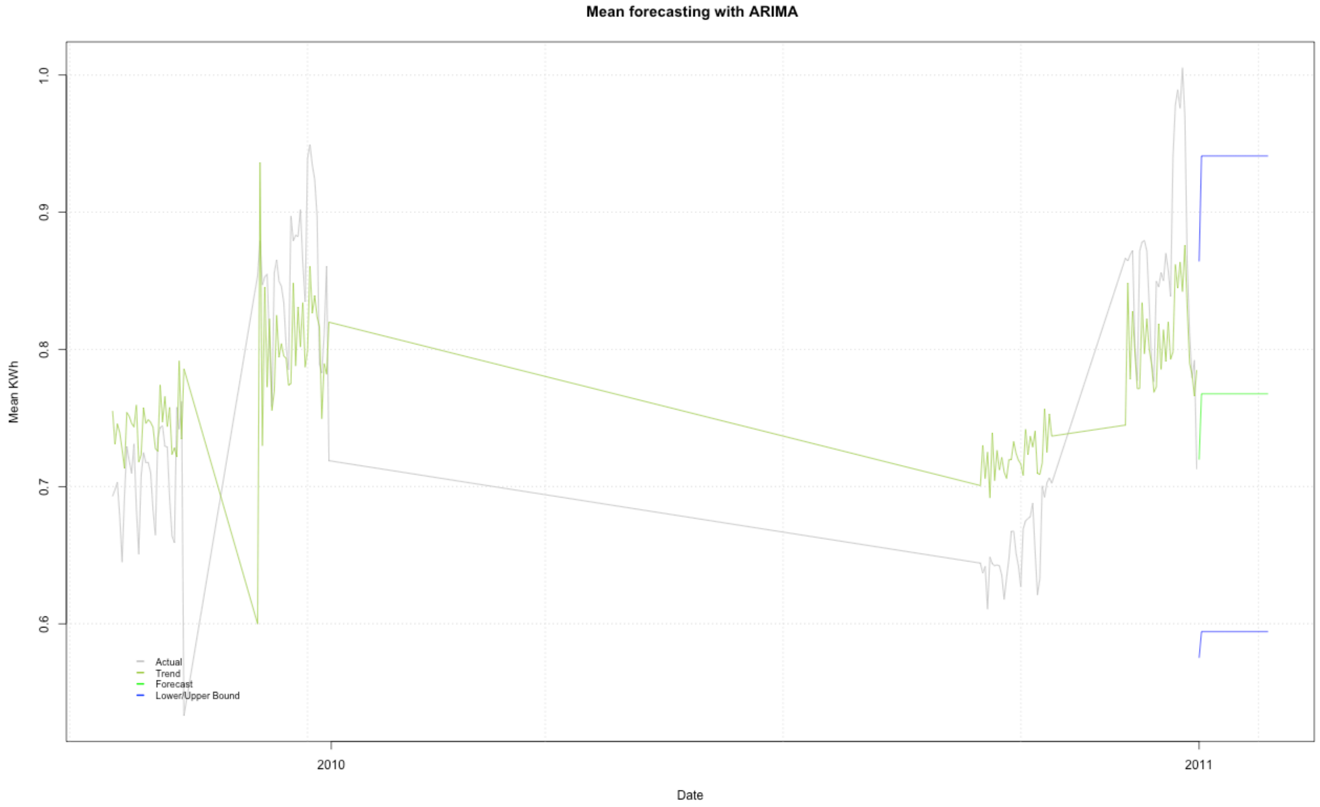
ARIMA forecasts



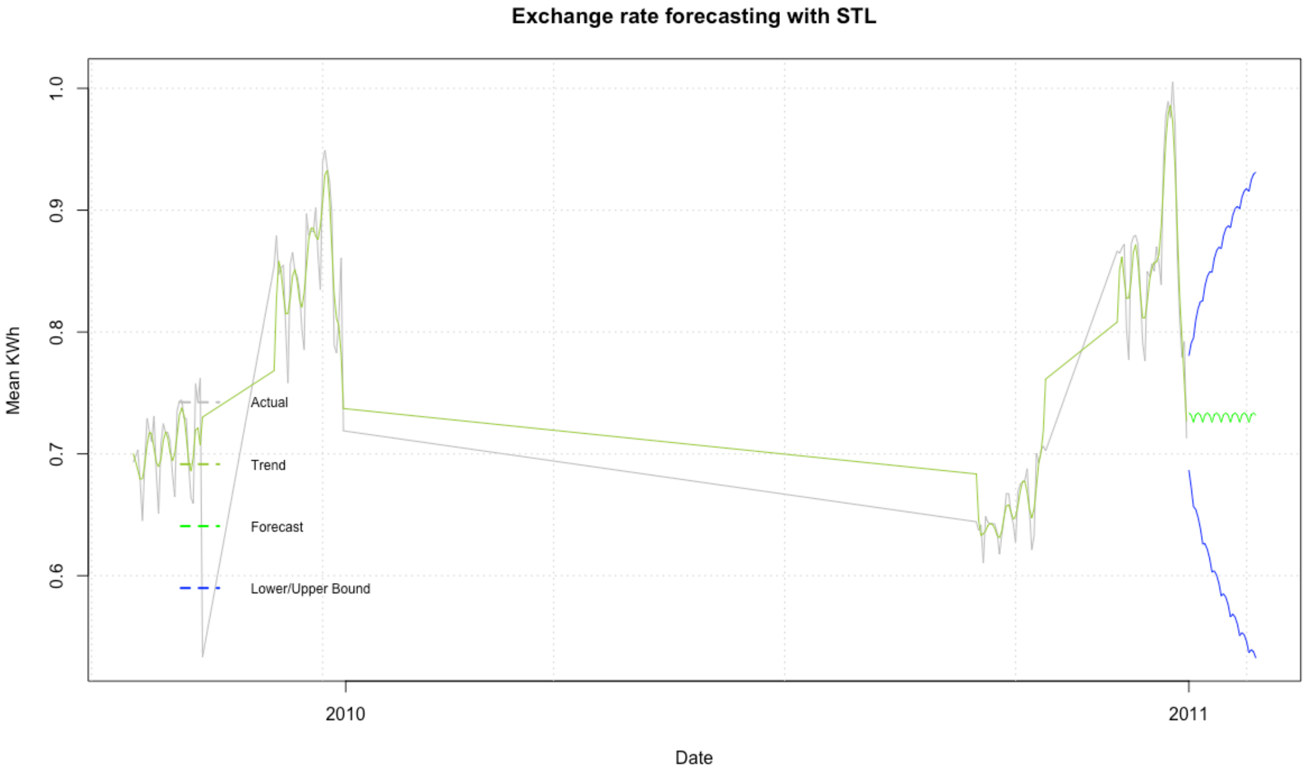


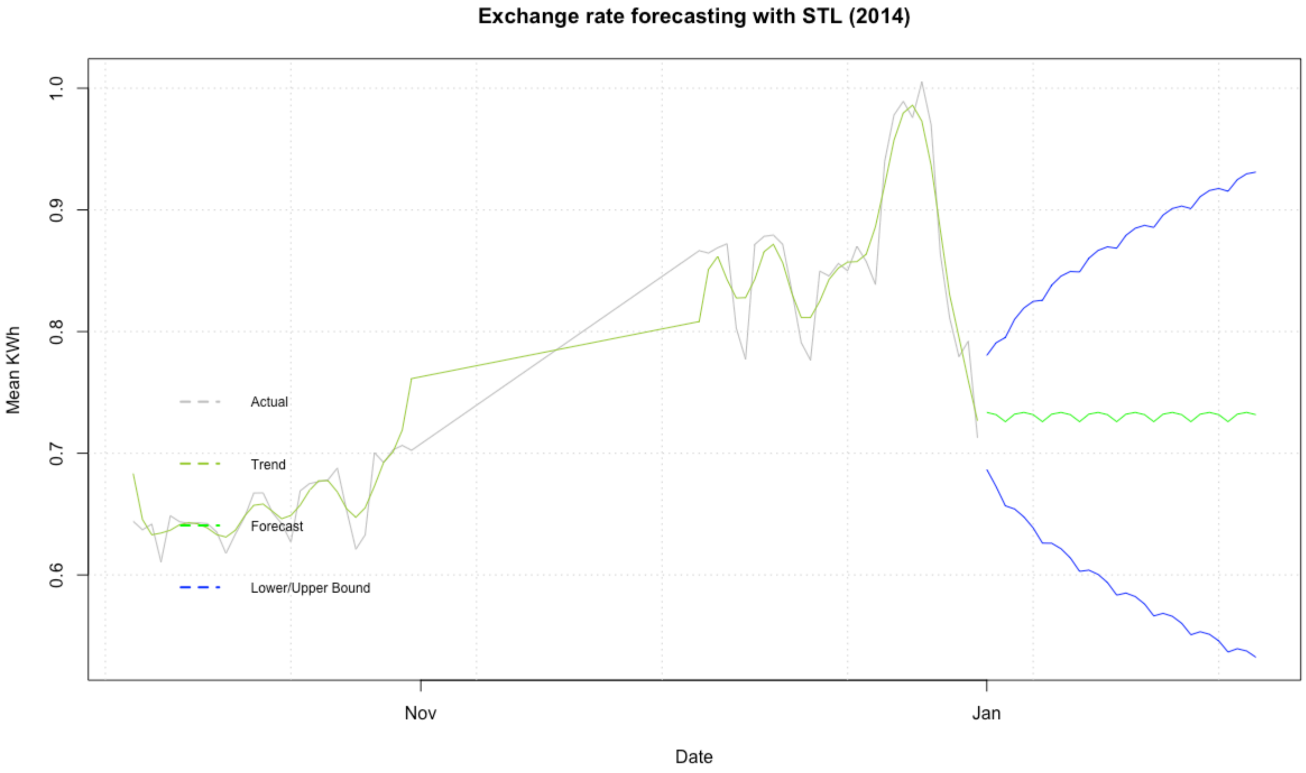


Mean Forecasting With ARIMA

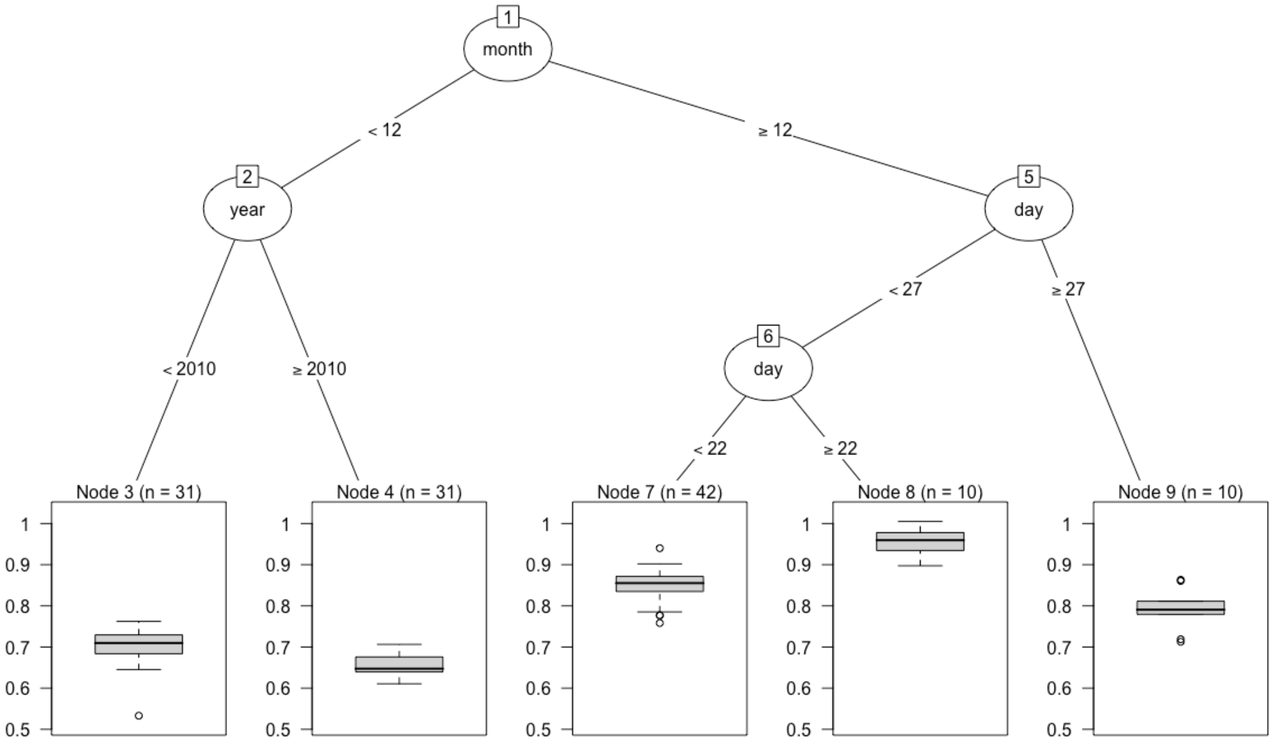


# Forecasting with STL

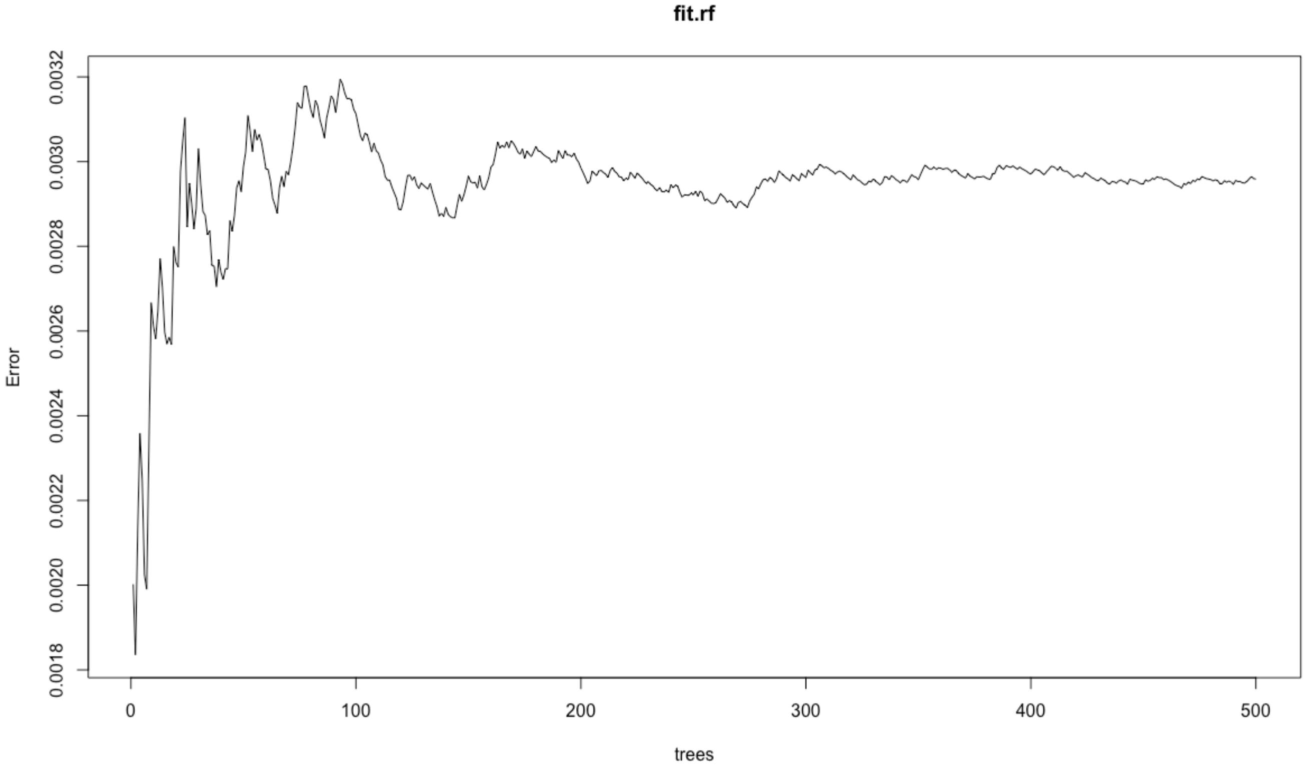


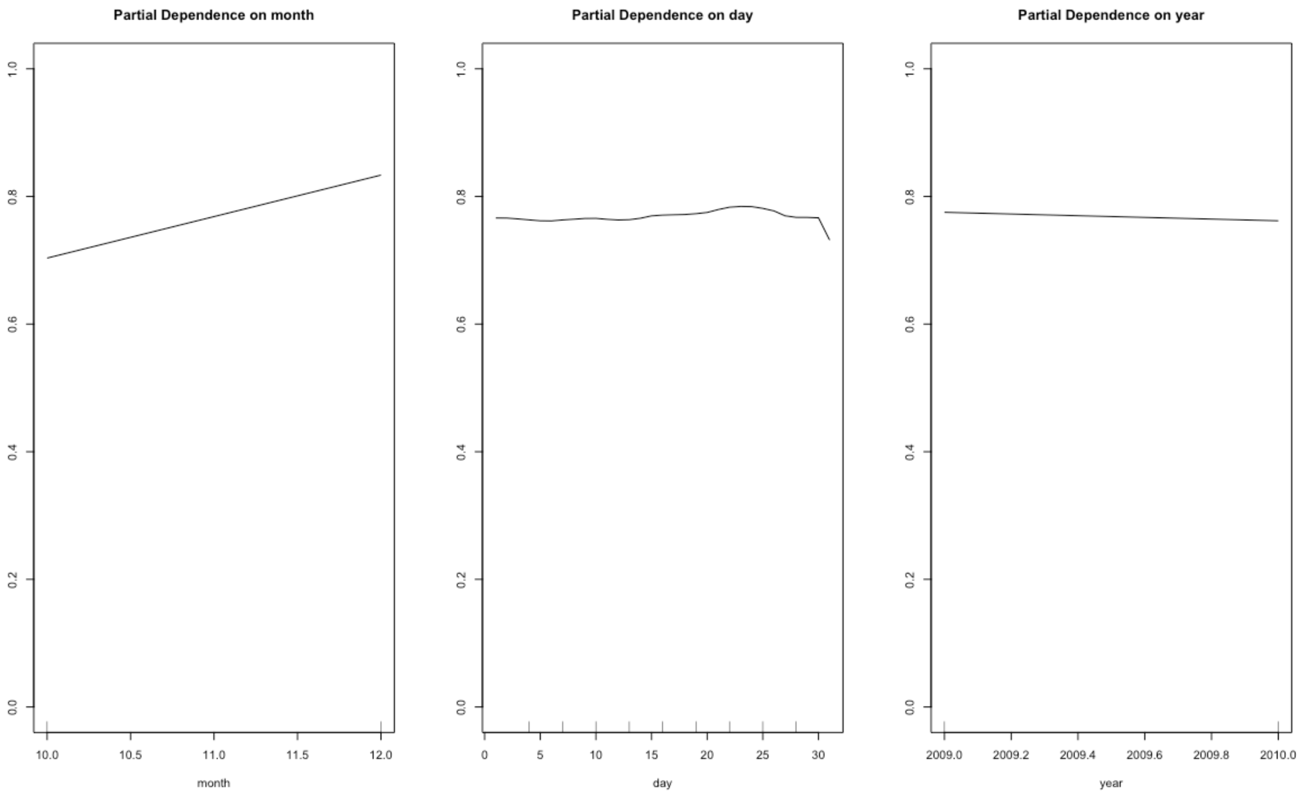


RandomForest.R
Creating random forest according to formula
EVTREE (Evoluationary Learning)



```
Call:
 randomForest(formula = frmla, data = RM_kWh)
            Type of random forest: regression
                  Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 0.002958645
                    % Var explained: 72.5
```

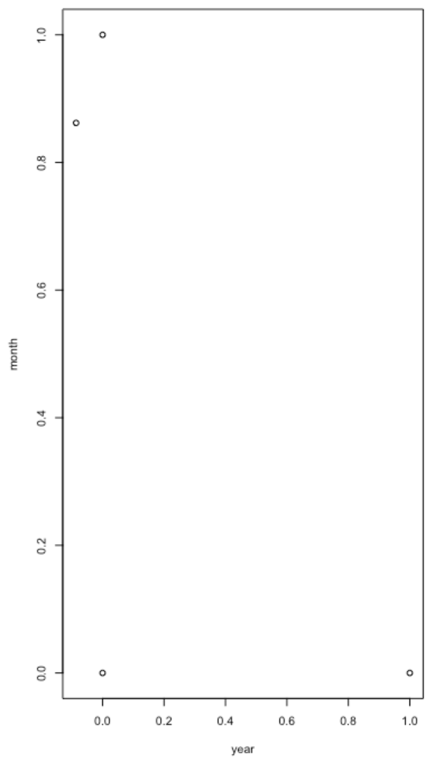




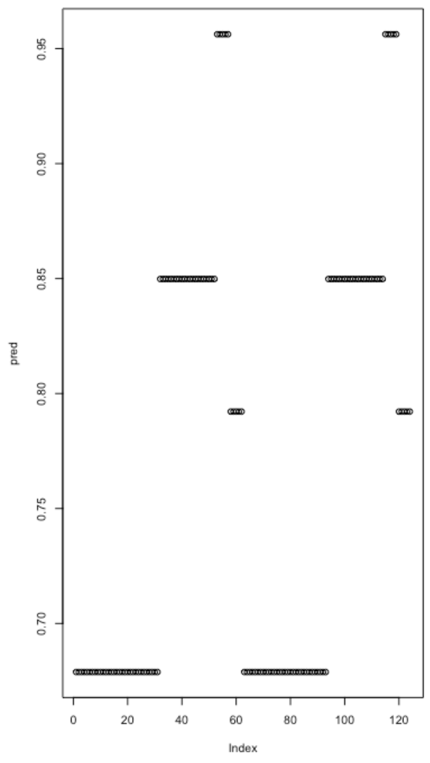
## varSelRF package

## Example of importance function show that forcing x1 to be the most important
## while create secondary variables that is related to x1.

```
        %IncMSE IncNodePurity
year     0.6790292      70.21114
month    2.1291241      67.38174
day     34.2885302    8761.00829
MeankWh  2.9829618     962.25909
```



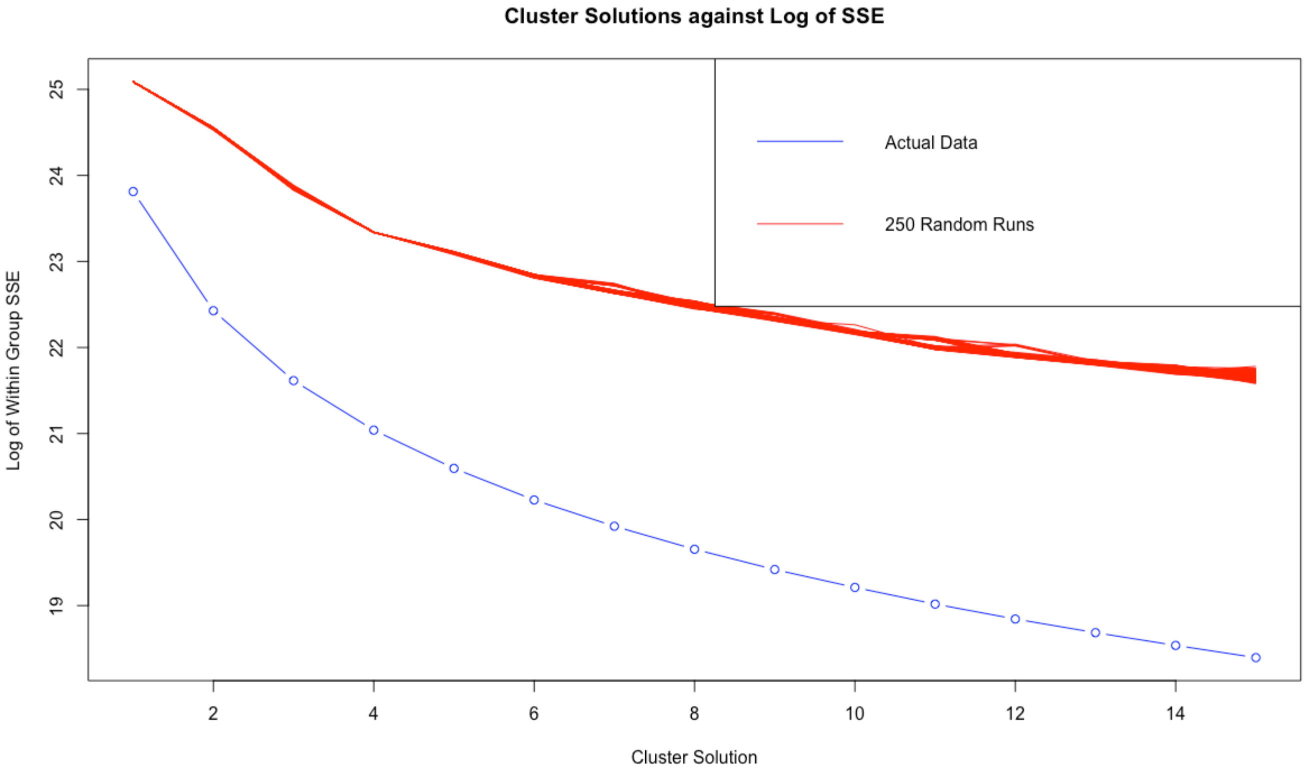
COREIearn

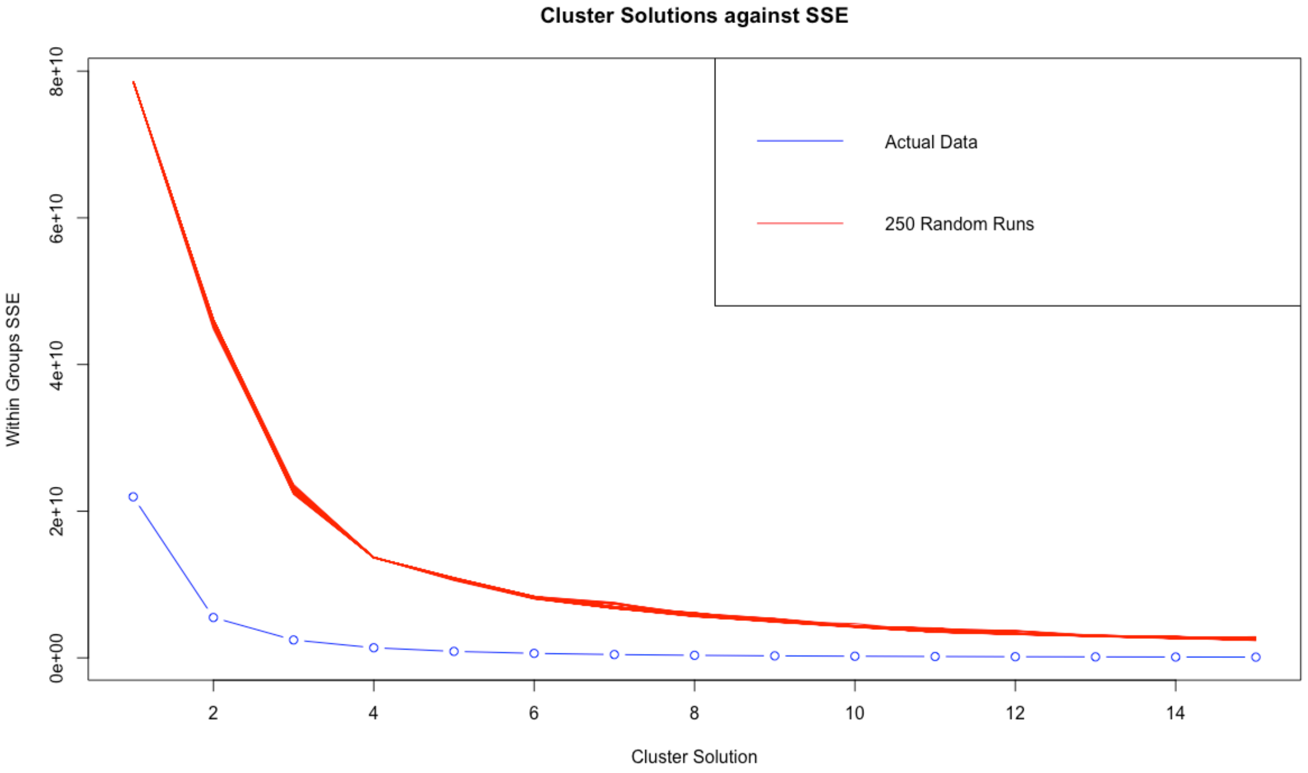


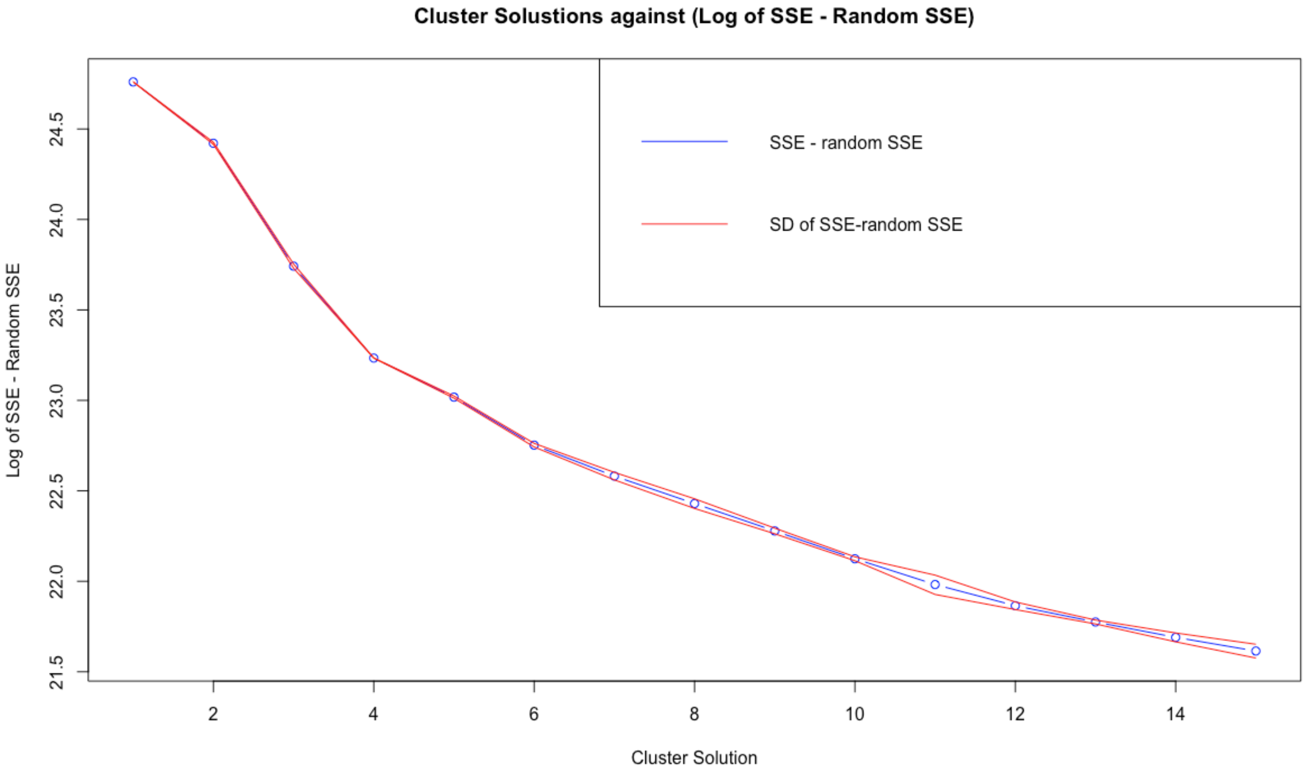
Kmeans.R, when you running this code enter input like below
Covert data to percents? 1=yes, 2=no : 2

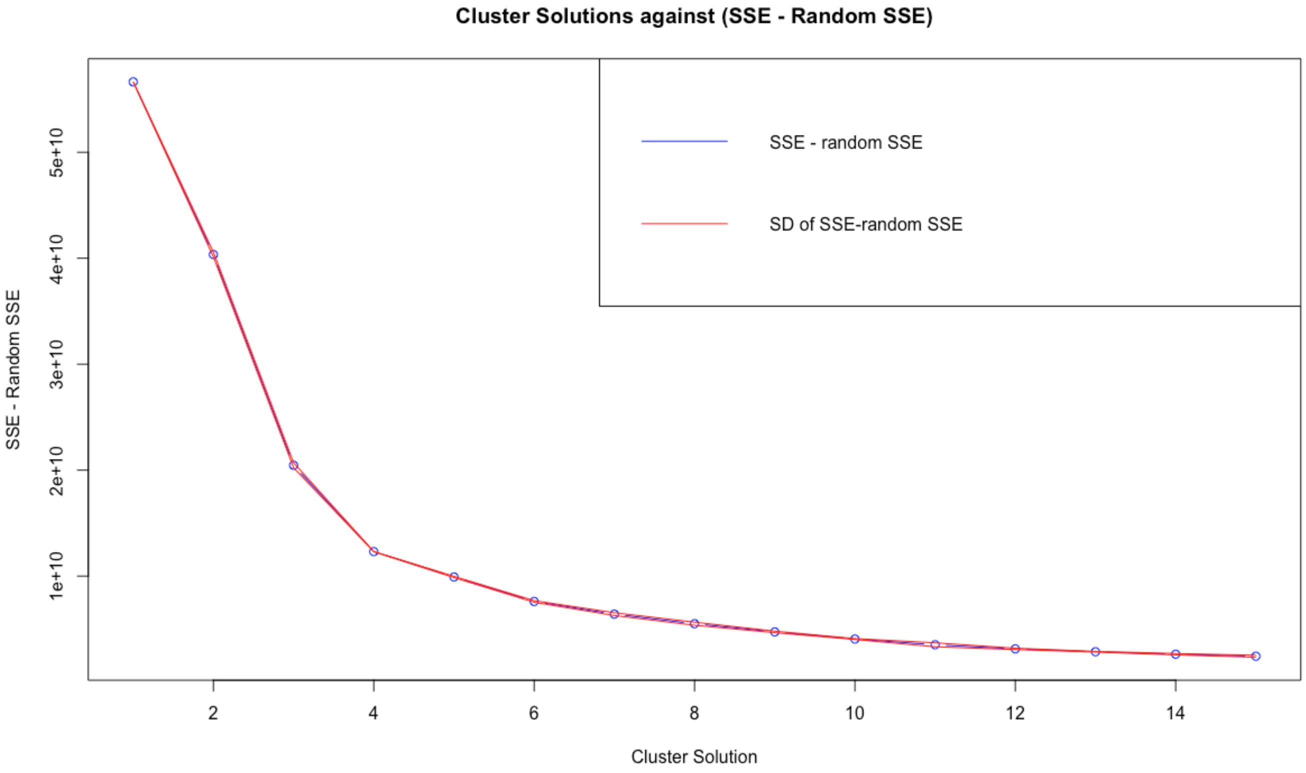
Z-score standardize data? 1=yes, 2=no : 2

How many clustering solutions to test (> row numbers)? 15

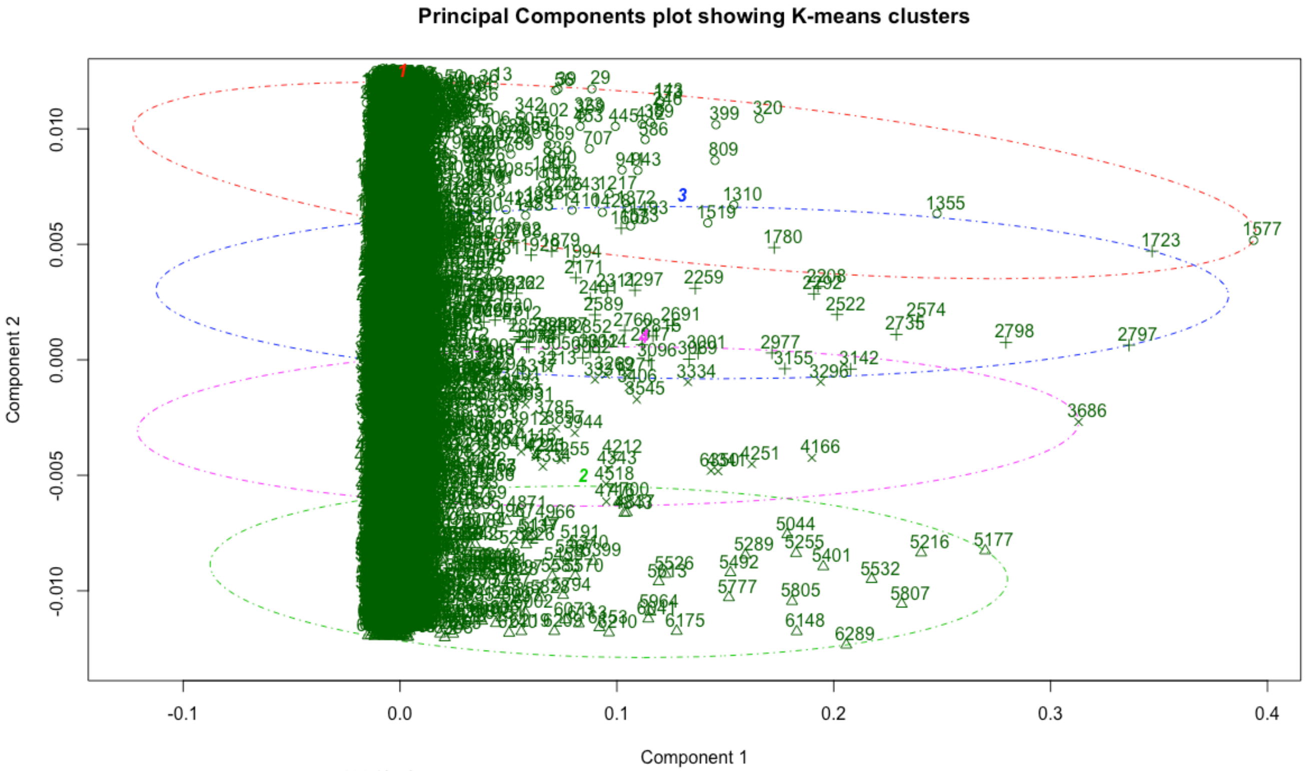








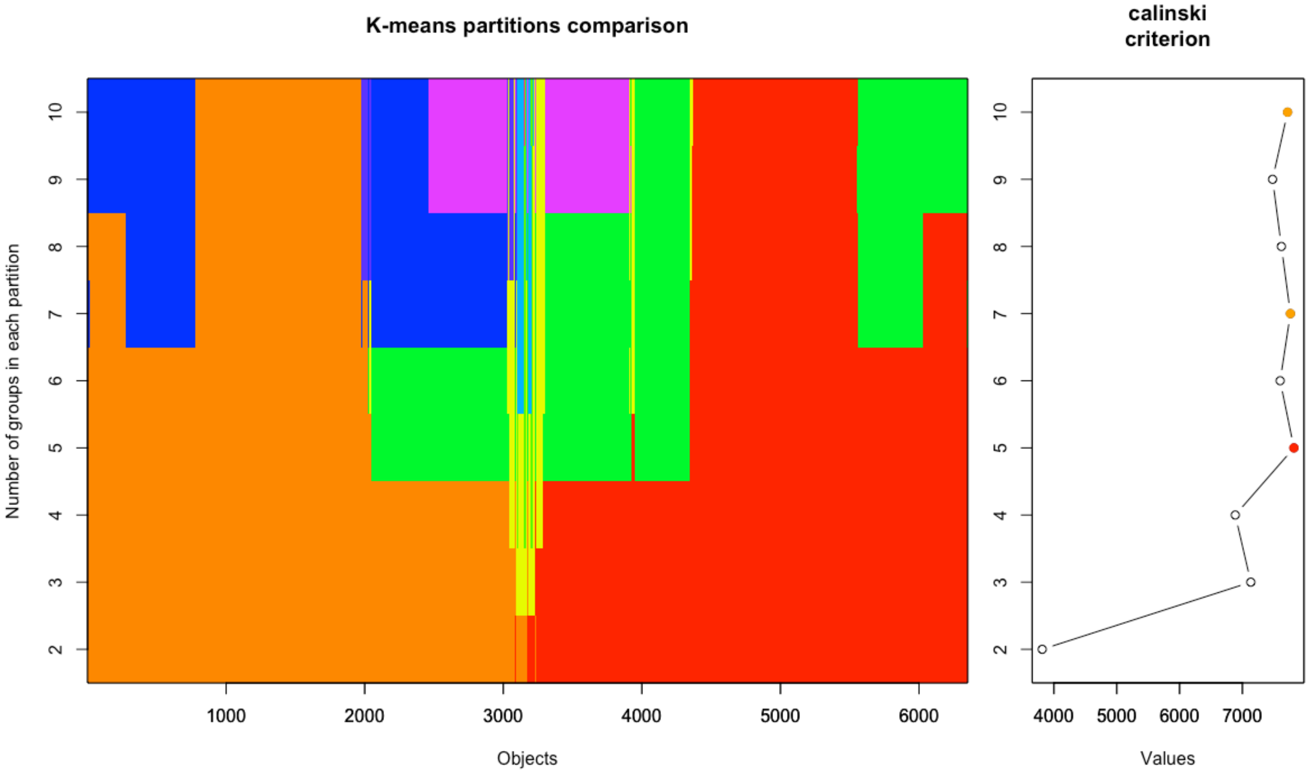
What clustering solution would you like to use? 4
After that script will be save result in file.
And this is graphic result:



These two components explain 100 % of the point variability.

cluster.R

# Calinsky criterion: approach to diagnosing how many clusters suit the data.



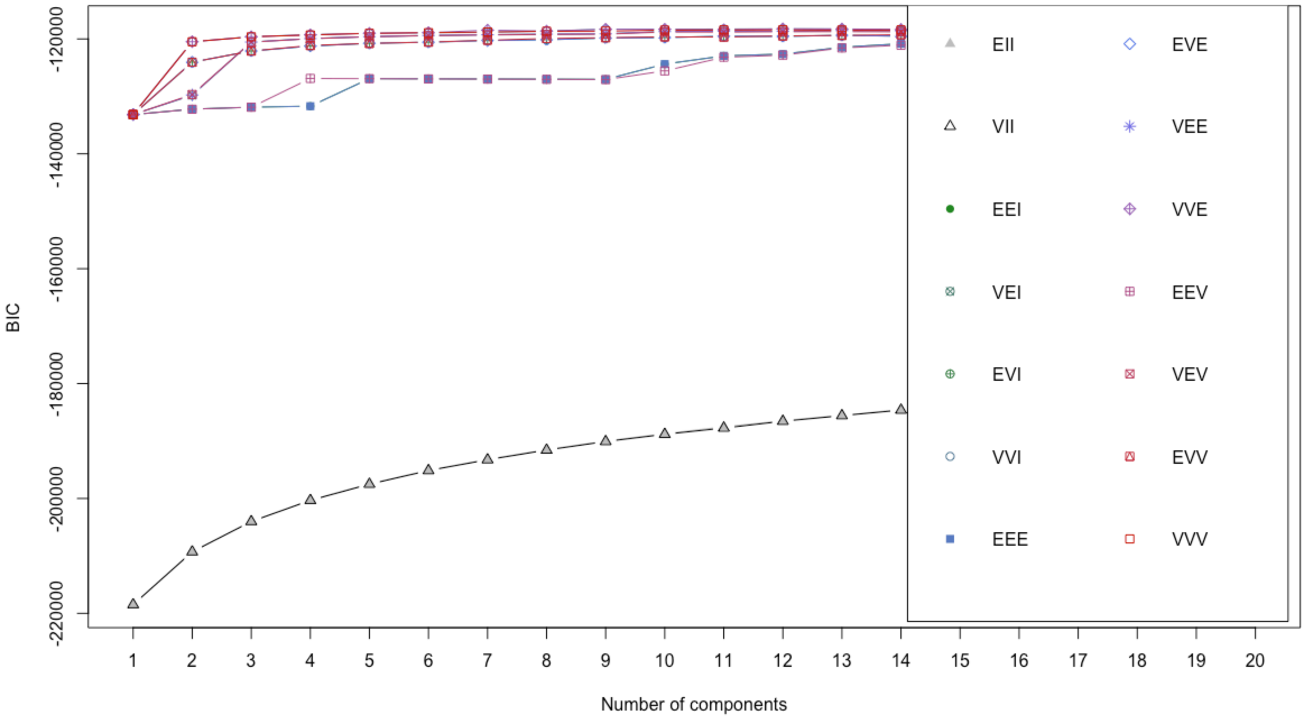
Model-based clustering plots:

1: BIC

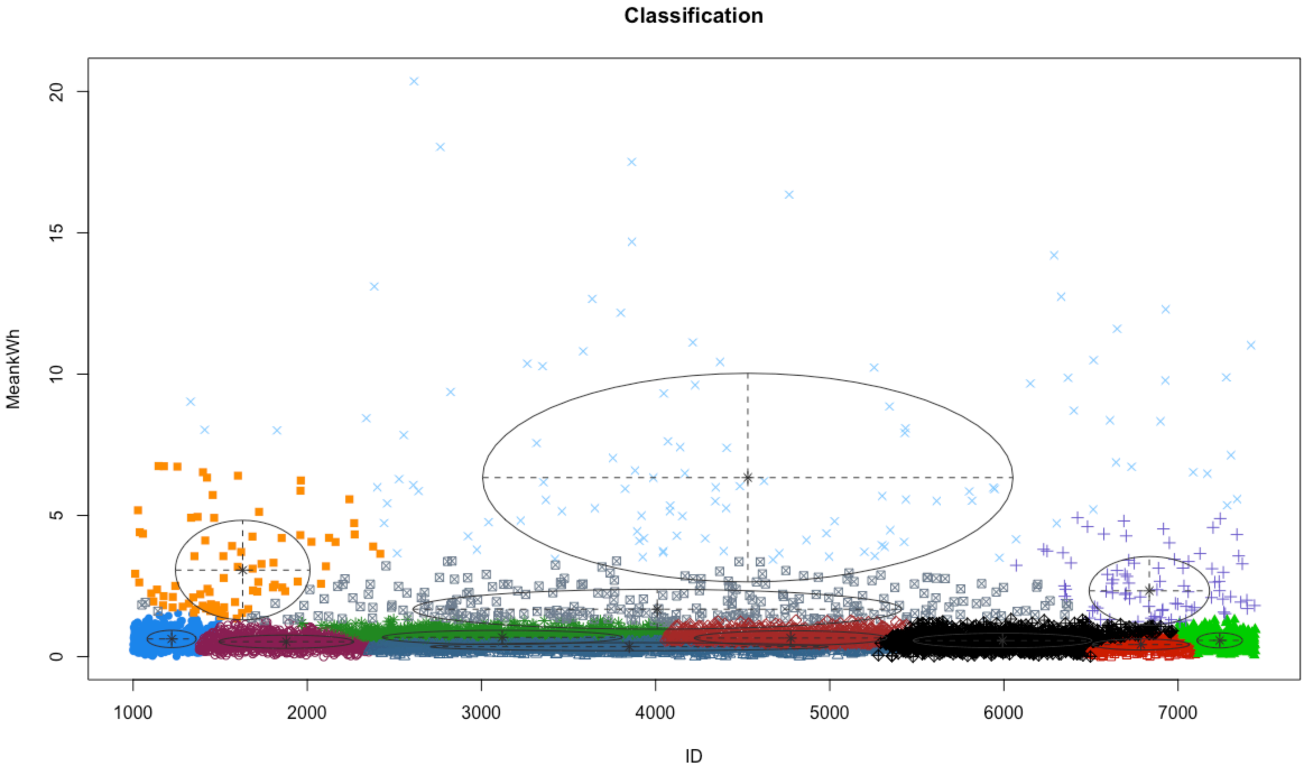
2: classification

3: uncertainty
4: density

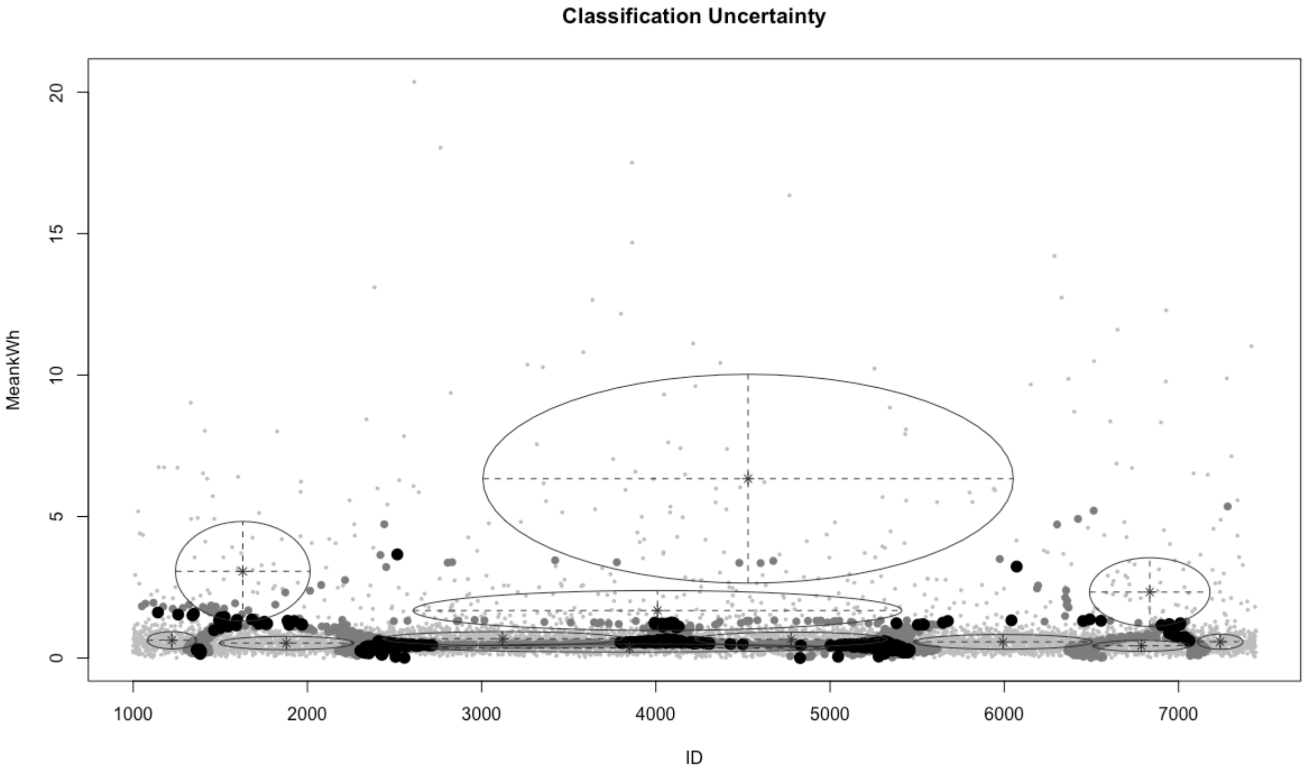
Selection: 1



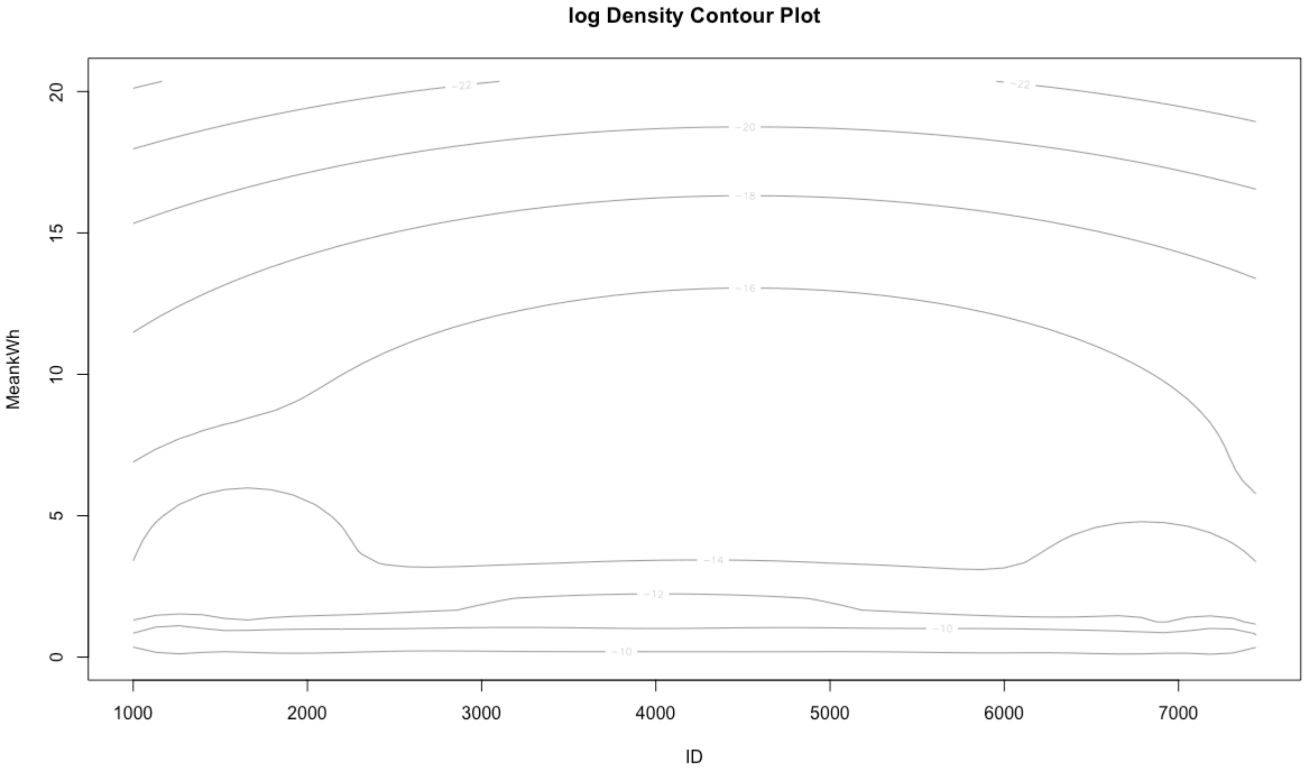
Selection: 2



Selection : 3


Classification Uncertainty
MeankWh
0
5
10
15
20
1000
2000
3000
4000
5000
6000
7000
ID

Selection 4:


log Density Contour Plot
MeankWh
0
5
10
15
20
1000
2000
3000
4000
5000
6000
7000
ID
-22
-22
-20
-18
-16
-14
-12
-10
-10