ENERGY USAGE OF APPLIANCES

Yosef Woldeamanuel
Jonathan De Los Santos

Introduction

Appliances

In most households around the world, there are multiple electrical appliances that consume energy either throughout the day or when used. Modern appliances have the options to be in standby mode to save energy when they are not used. Some appliances such as refrigerator, internet modems and security cameras have to be powered all the time and consume energy continuously.

In this dataset, we examine a low-energy home outfitted with multiple IoT devices to measure temperature (T) and relative humidity (RH) in multiple rooms along with the energy expended by lights and appliances in watt-hours (Wh). Weather data is also provided from a nearby weather station to improve the prediction modeling.

Energy Consumption

The amount of energy needed for a specific house will depend on many factors such as the number of appliances, frequency of use, geographical location and climate, number of occupants and efficiency of the house itself. The overall energy usage of a house can easily be aggregated but knowing individual factors which contributes in higher usage is not an easy task.

“Data driven prediction models of energy use of appliances in a low-energy house” is a paper by Luis M.Candanedo, Véronique Feldheim and Dominique Deramaix which presented and discussed data driven prediction models of energy use of appliances. Temperature and humidity data collected at different part of house for this paper will be used to apply multivariate analysis, data cleaning techniques, and other data visualization skills we learned in this course. We are interested on knowing how the inside and outside temperature, humidity and other parameters affect the energy consumption of appliances in the house and how variables are related to each other.

Methodology

To determine those relationships, we examine the correlations, perform dimension reduction analysis graphically and using principal component analysis, and cluster analysis. The correlation matrices reveal that there is a correlation between inside temperature, humidity, and appliance energy use. Additionally, we find that specific rooms have stronger correlations likely due to the specific appliances they contain.

PCA tells us that temperature is responsible for the highest variance of our dataset, followed by humidity. This may be due to the multiple variables for each that represent the temperature and humidity in different rooms. However, the correlation may imply that small differences in appliance and light energy use affect large variations in temperature and humidity. Finally, the cluster analysis performed on the first two principal components shows obvious clusters of temperatures and humidities. This is presented to show the strength of PCA in identifying variable groups, even if they were understood beforehand.

Attribute Information:

- date, time year-month-day hour:minute:second
- Appliances, energy use in Wh
- lights, energy use of light fixtures in the house in Wh
- T1, Temperature in kitchen area, in Celsius
- RH_1, Humidity in kitchen area, in %
- T2, Temperature in living room area, in Celsius
- RH_2, Humidity in living room area, in %
- T3, Temperature in laundry room area
- RH_3, Humidity in laundry room area, in %
- T4, Temperature in office room, in Celsius
- RH_4, Humidity in office room, in %
- T5, Temperature in bathroom, in Celsius
- RH_5, Humidity in bathroom, in %
- T6, Temperature outside the building (north side), in Celsius
- RH_6, Humidity outside the building (north side), in %
- T7, Temperature in ironing room , in Celsius
- RH_7, Humidity in ironing room, in %
- T8, Temperature in teenager room 2, in Celsius
- RH_8, Humidity in teenager room 2, in %
- T9, Temperature in parents room, in Celsius
- RH_9, Humidity in parents room, in %
- To, Temperature outside (from Chievres weather station), in Celsius
- Pressure (from Chievres weather station), in mm Hg
- RH_out, Humidity outside (from Chievres weather station), in %
- Wind speed, (from Chievres weather station), in m/s
- Visibility, (from Chievres weather station), in km

- Tdewpoint, (from Chievres weather station), Â°C

Data Cleaning and Visualization

Import data energy “csv” data.

```
options(digits = 2)
library(readr)

library(mclust)

#install.packages("CCA")

head(energydata_complete[,c(1:5,26:29)])
```

```
## # A tibble: 6 x 9
##   date                Appliances lights    T1  RH_1 Visibility Tdewpoint   rv1
##   <dttm>                   <dbl>  <dbl> <dbl> <dbl>      <dbl>     <dbl> <dbl>
## 1 2016-01-11 17:00:00         60     30  19.9  47.6       63         5.3  13.3
## 2 2016-01-11 17:10:00         60     30  19.9  46.7       59.2       5.2  18.6
## 3 2016-01-11 17:20:00         50     30  19.9  46.3       55.3       5.1  28.6
## 4 2016-01-11 17:30:00         50     40  19.9  46.1       51.5       5    45.4
## 5 2016-01-11 17:40:00         60     40  19.9  46.3       47.7       4.9  10.1
## 6 2016-01-11 17:50:00         50     40  19.9  46.0       43.8       4.8  44.9
## # … with 1 more variable: rv2 <dbl>
```

```
#dim(energydata_complete)
```

There are two columns with random values which are not part of the real data, these will be dropped from the data.

```
randomv =c("rv1","rv2")
energydata_complete = energydata_complete[ , !(names(energydata_complete) %in
% randomv)]
head(energydata_complete[,c(1:5,24:27)])
```

```
## # A tibble: 6 x 9
##   date                Appliances lights    T1  RH_1 RH_out Windspeed Visibility
##   <dttm>                   <dbl>  <dbl> <dbl> <dbl>  <dbl>     <dbl>      <dbl>
## 1 2016-01-11 17:00:00         60     30  19.9  47.6     92      7          63
## 2 2016-01-11 17:10:00         60     30  19.9  46.7     92      6.67       59.2
## 3 2016-01-11 17:20:00         50     30  19.9  46.3     92      6.33       55.3
## 4 2016-01-11 17:30:00         50     40  19.9  46.1     92      6          51.5
## 5 2016-01-11 17:40:00         60     40  19.9  46.3     92      5.67       47.7
## 6 2016-01-11 17:50:00         50     40  19.9  46.0     92      5.33       43.8
## # … with 1 more variable: Tdewpoint <dbl>
```

The energy data has values for every 10 minutes of the hour for over five months with 19735 rows. In order to reduce the number of rows, “Zoo” library will be used to get the aggregate average daily numbers:

```
library(zoo)
```

```
energydata_daily <-aggregate(read.zoo(energydata_complete, header = TRUE, tz
= "GMT"), as.Date, mean)

energydata_daily = as.data.frame(energydata_daily)
#dim(energydata_daily)
```

Let's use "PerformanceAnalytics" library to check for scatterplots between variables, correlation, and histogram of some of the variables which will help to see if there are extreme outliers.

```
library("PerformanceAnalytics")

chart.Correlation(energydata_daily[,1:10], histogram=TRUE, pch=19, col="")
```

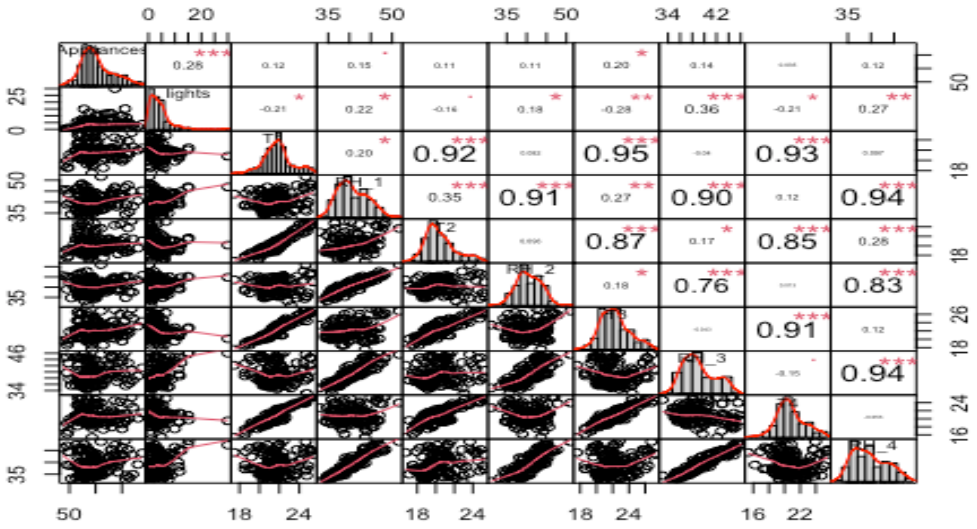

Figure 1

The only variable with a few outliers is lights as shown in figure 1. This may reflect a time where lights were left on while the owners were away, or a malfunction in the IoT equipment. The MVA library will be used to check which rows are with exterme outliers and clean them up.

```
#check outliers
library(MVA)

plot(energydata_daily$Appliances, energydata_daily$lights)
text(energydata_daily$Appliances, energydata_daily$lights, cex = 0.6,labels =
row.names(energydata_daily))
```

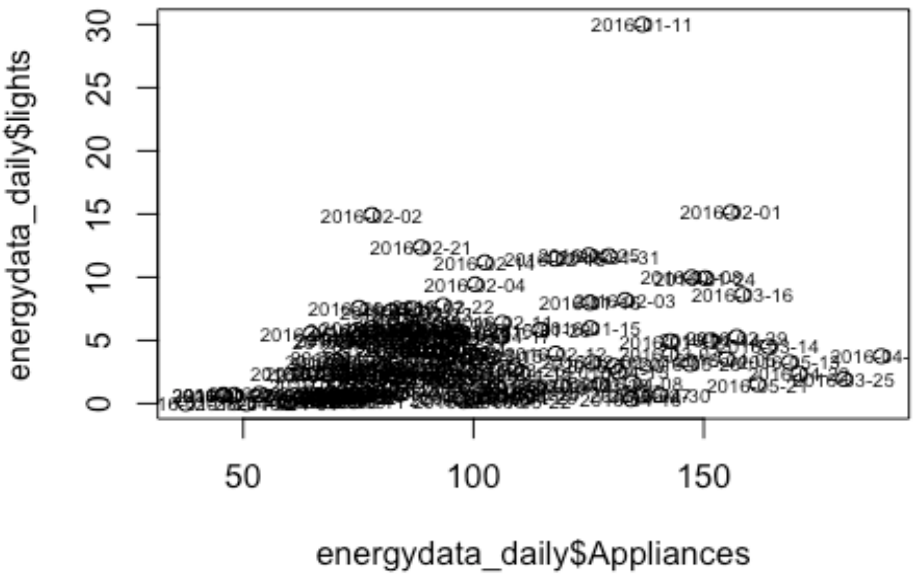

Figure 2

The scatterplot between appliances and lights show days such as 01/11/2016 that appear to be outliers as shown in figure 2. Bvbox will be used to confirm these as outliers.

```
Bvbox(energydata_daily[,c("Appliances","lights")])
text(energydata_daily[,c("Appliances","lights")]$Appliances, energydata_daily
[,c("Appliances","lights")]$lights)
```

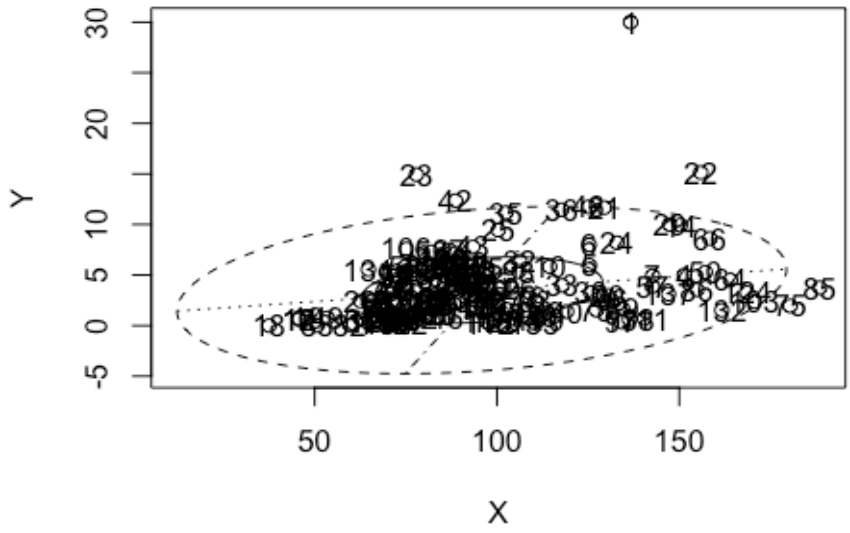

Figure 3

There are 6 rows `(1,28,22,42, 75,85)` outside of the ellipse which will be dropped to reduce the variation on the data analysis.

```
Outliers = c(1,28,22,42, 75,85)
energydata_daily = energydata_daily[-outliers,]
plot(energydata_daily$Appliances, energydata_daily$lights)
```

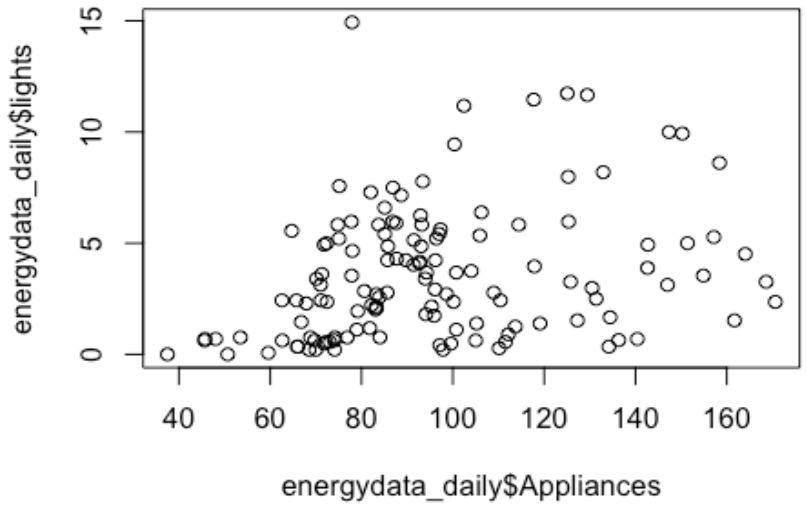

Figure 4

This are the resulting histograms and correlations of the aggregated energy data after cleaning.

```
chart.Correlation(energydata_daily[,1:8], histogram=TRUE, pch=19, col="")
```

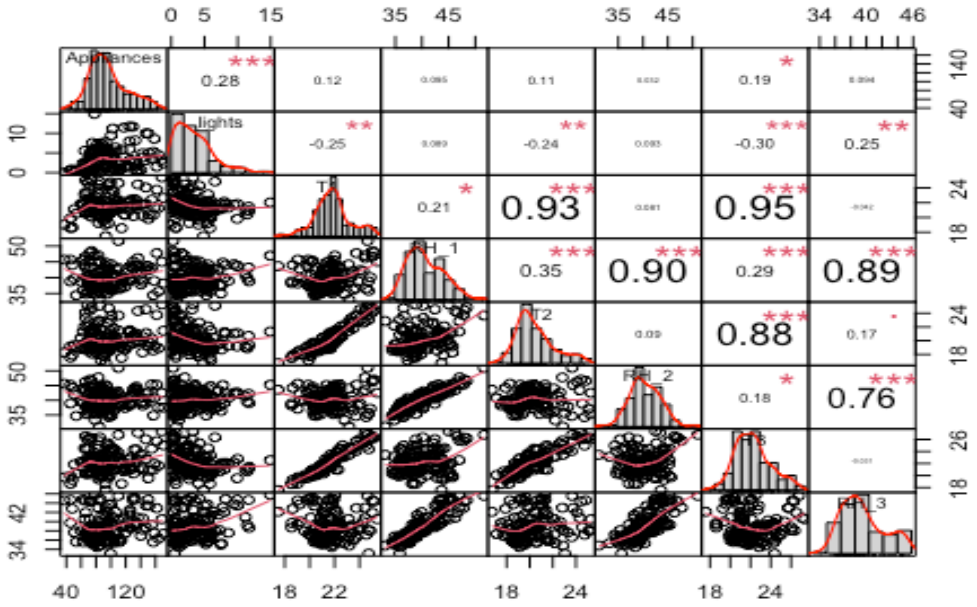

Figure 5

Now the data will be scaled to minimize variation due to the unity measurements difference between the variables. Since T6 and RH_6 are values measured outside of the

house, the two variables has been excluded from the correlation. Scaled appliances energy consumption correlation with all in-house variables is generated below.

```
daily.s = scale(energydata_daily)

round(cor(daily.s[c(1:20)]),3)
```

```
##             Appliances lights     T1   RH_1     T2   RH_2     T3   RH_3     T4
## Appliances      1.000  0.284  0.118  0.095  0.111  0.032  0.185  0.094  0.054
## lights          0.284  1.000 -0.252  0.089 -0.236  0.093 -0.299  0.249 -0.225
## T1              0.118 -0.252  1.000  0.211  0.927  0.081  0.953 -0.042  0.939
## RH_1            0.095  0.089  0.211  1.000  0.351  0.905  0.287  0.894  0.149
## T2              0.111 -0.236  0.927  0.351  1.000  0.090  0.879  0.165  0.865
## RH_2            0.032  0.093  0.081  0.905  0.090  1.000  0.184  0.760  0.029
## T3              0.185 -0.299  0.953  0.287  0.879  0.184  1.000 -0.031  0.917
## RH_3            0.094  0.249 -0.042  0.894  0.165  0.760 -0.031  1.000 -0.135
## T4              0.054 -0.225  0.939  0.149  0.865  0.029  0.917 -0.135  1.000
## RH_4            0.058  0.135  0.100  0.937  0.277  0.826  0.134  0.940 -0.032
## T5              0.056 -0.288  0.942  0.252  0.858  0.161  0.943 -0.055  0.929
## RH_5            0.137  0.291 -0.122  0.571  0.014  0.480 -0.128  0.694 -0.188
## T6             -0.011 -0.359  0.796  0.403  0.838  0.262  0.841  0.126  0.764
## RH_6            0.006  0.465 -0.696  0.249 -0.575  0.264 -0.740  0.548 -0.771
## T7              0.045 -0.354  0.901  0.071  0.792  0.002  0.893 -0.236  0.926
## RH_7            0.019  0.069  0.212  0.893  0.402  0.742  0.226  0.886  0.107
## T8              0.101 -0.274  0.870 -0.014  0.702 -0.028  0.854 -0.316  0.853
## RH_8            0.088  0.161  0.067  0.884  0.271  0.729  0.084  0.928 -0.044
## T9              0.022 -0.364  0.929  0.156  0.815  0.099  0.942 -0.186  0.935
## RH_9            0.094  0.149  0.163  0.896  0.343  0.752  0.168  0.924  0.040
##               RH_4     T5   RH_5     T6   RH_6     T7   RH_7     T8   RH_8
## Appliances  0.058  0.056  0.137 -0.011  0.006  0.045  0.019  0.101  0.088
## lights      0.135 -0.288  0.291 -0.359  0.465 -0.354  0.069 -0.274  0.161
## T1          0.100  0.942 -0.122  0.796 -0.696  0.901  0.212  0.870  0.067
## RH_1        0.937  0.252  0.571  0.403  0.249  0.071  0.893 -0.014  0.884
## T2          0.277  0.858  0.014  0.838 -0.575  0.792  0.402  0.702  0.271
## RH_2        0.826  0.161  0.480  0.262  0.264  0.002  0.742 -0.028  0.729
## T3          0.134  0.943 -0.128  0.841 -0.740  0.893  0.226  0.854  0.084
## RH_3        0.940 -0.055  0.694  0.126  0.548 -0.236  0.886 -0.316  0.928
## T4         -0.032  0.929 -0.188  0.764 -0.771  0.926  0.107  0.853 -0.044
## RH_4        1.000  0.099  0.625  0.311  0.402 -0.101  0.935 -0.172  0.942
## T5          0.099  1.000 -0.163  0.784 -0.727  0.919  0.191  0.879  0.040
## RH_5        0.625 -0.163  1.000 -0.123  0.519 -0.252  0.592 -0.252  0.662
## T6          0.311  0.784 -0.123  1.000 -0.678  0.730  0.423  0.624  0.263
## RH_6        0.402 -0.727  0.519 -0.678  1.000 -0.816  0.290 -0.762  0.452
## T7         -0.101  0.919 -0.252  0.730 -0.816  1.000  0.030  0.916 -0.171
## RH_7        0.935  0.191  0.592  0.423  0.290  0.030  1.000 -0.096  0.922
## T8         -0.172  0.879 -0.252  0.624 -0.762  0.916 -0.096  1.000 -0.222
## RH_8        0.942  0.040  0.662  0.263  0.452 -0.171  0.922 -0.222  1.000
## T9         -0.019  0.955 -0.246  0.796 -0.816  0.965  0.074  0.924 -0.098
## RH_9        0.943  0.124  0.625  0.340  0.363 -0.043  0.917 -0.137  0.916
##                T9   RH_9
## Appliances  0.022  0.094
## lights     -0.364  0.149
## T1          0.929  0.163
```

```
## RH_1         0.156  0.896
## T2           0.815  0.343
## RH_2         0.099  0.752
## T3           0.942  0.168
## RH_3        -0.186  0.924
## T4           0.935  0.040
## RH_4        -0.019  0.943
## T5           0.955  0.124
## RH_5        -0.246  0.625
## T6           0.796  0.340
## RH_6        -0.816  0.363
## T7           0.965 -0.043
## RH_7         0.074  0.917
## T8           0.924 -0.137
## RH_8        -0.098  0.916
## T9           1.000  0.008
## RH_9         0.008  1.000
```

There is a positive correlation between the energy consumption of appliances and lights. Light energy consumption usually increases when there is occupant in the house, and the positive correlation with appliances indicates that more appliances are used when the house is occupied.

Temperature and humidity at different areas of the house had positive correlations with appliance energy consumption. Kitchen and living area temperatures have a slightly higher correlation with appliances, which indicates occupant's usage of cooking and living room appliances or that these appliances cause the largest temperature change in their respective rooms.

Higher positive correlation numbers between temperature and humidity reading at different part of the house is expected since the overall inside house environment at different part of the house are related.

Dimension Reduction Analysis

Graphical MDA

Graphical MDA is used to check if some days share similar variables. Days with high similarities are located closer to each other in figure 6. Unsurprisingly, we see many clusters of days that are temporally "near" each other such as 2016-05-08 and 2016-05-09.

```
energy_dist = dist(daily.s[,c(1:20)])
energy.mds = cmdscale(energy_dist, eig = T)

plot(energy.mds$points[,1:2], pch = ".",col="red", cex= 5)
text(energy.mds$points[,1:2], labels = rownames(energydata_daily), cex = 0.6,
pos=2)
```

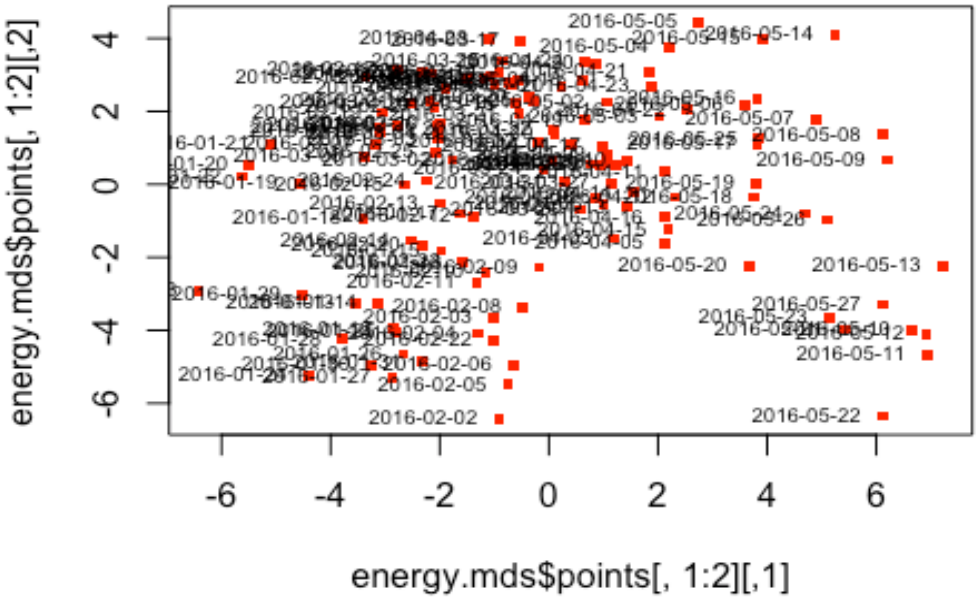

Figure 6

In addition to the days, graphical MDA for the variables shows that Temperature and humidity variables are grouped together as expected whereas some of the outside variables have their own properties and located on the middle and bottom part of figure 7. One interesting note is that dew point, which is a factor of humidity, almost evenly bisects the origin angle between the external temperature and humidity.

```
energy_dist.v = 1-cor(energydata_daily)
energy.v.mds = cmdscale(energy_dist.v, eig=T)
plot(energy.v.mds$points[,1:2], pch= ".", col="red", cex= 5)
text(energy.v.mds$points[,1:2], labels= colnames(energydata_daily), pos=4)
```

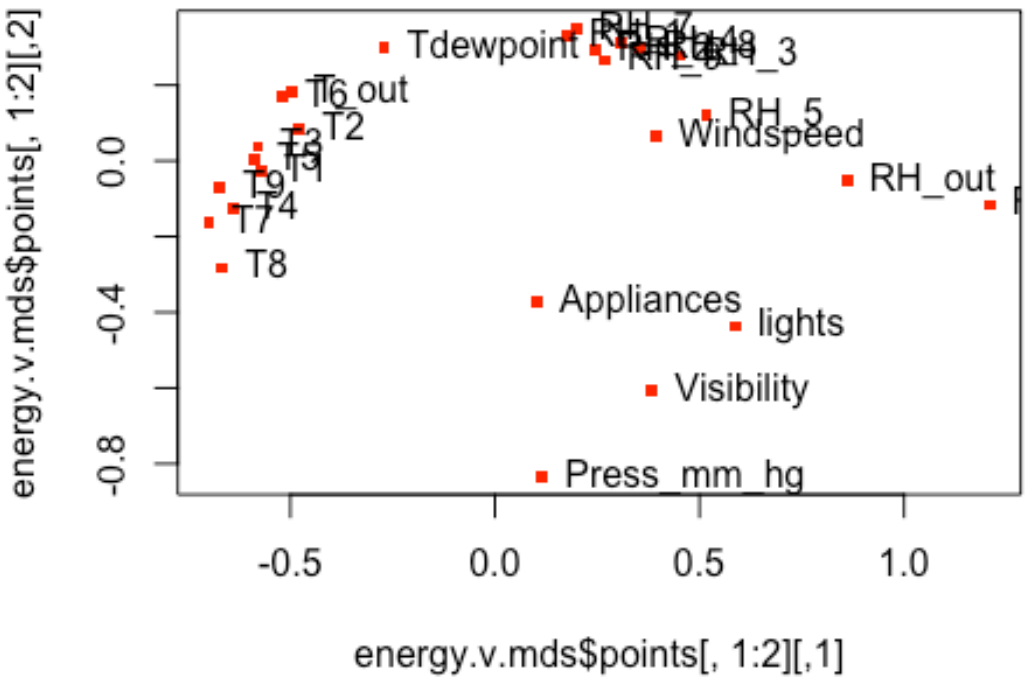

Figure 7

Principal component analysis

Principal component analysis is used to reduce the many variables in the energy data and get a few variables that explain most of the variation in the data.

Summary of the PCA shows that the first 2 PCs cover over 72% of the variation in the variables. The first PC covers the higher in temperature and Tdewpoint. The second PC covers the higher humidity values both inside and outside the house.

```
energydata_daily.pca <- princomp(energydata_daily, cor = T)
summary(energydata_daily.pca)
```

```
## Importance of components:
##                          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## Standard deviation         3.26   2.82  1.206  1.089   1.03  0.926  0.827  0.772
## Proportion of Variance     0.41   0.31  0.056  0.046   0.04  0.033  0.026  0.023
## Cumulative Proportion      0.41   0.72  0.771  0.817   0.86  0.890  0.917  0.940
##                          Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15
## Standard deviation       0.674   0.579  0.4918  0.3349  0.2915  0.2655  0.2440
## Proportion of Variance   0.017   0.013  0.0093  0.0043  0.0033  0.0027  0.0023
## Cumulative Proportion    0.957   0.970  0.9792  0.9835  0.9868  0.9895  0.9918
##                          Comp.16 Comp.17 Comp.18 Comp.19 Comp.20 Comp.21 Comp.22
## Standard deviation       0.2385  0.1985  0.1872   0.163 0.12823 0.12708   1e-01
## Proportion of Variance   0.0022  0.0015  0.0013   0.001 0.00063 0.00062   4e-04
## Cumulative Proportion    0.9939  0.9955  0.9968   0.998 0.99846 0.99908   1e+00
##                          Comp.23 Comp.24 Comp.25 Comp.26
```

```
## Standard deviation     0.08741 0.05965 4.2e-02 2.5e-02
## Proportion of Variance 0.00029 0.00014 6.9e-05 2.4e-05
## Cumulative Proportion  0.99977 0.99991 1.0e+00 1.0e+00

pPC1_2 = energydata_daily.pca$loadings[,c(1:2)]
pPC1_2

##             Comp.1  Comp.2
## Appliances   0.019  0.0319
## lights      -0.108  0.0935
## T1           0.291 -0.0325
## RH_1         0.107  0.3185
## T2           0.282  0.0305
## RH_2         0.068  0.2886
## T3           0.296 -0.0210
## RH_3         0.013  0.3419
## T4           0.283 -0.0718
## RH_4         0.063  0.3398
## T5           0.291 -0.0358
## RH_5        -0.035  0.2549
## T6           0.282  0.0394
## RH_6        -0.228  0.2079
## T7           0.278 -0.1029
## RH_7         0.098  0.3182
## T8           0.255 -0.1210
## RH_8         0.047  0.3359
## T9           0.291 -0.0782
## RH_9         0.076  0.3262
## T_out        0.280  0.0503
## Press_mm_hg -0.063 -0.1011
## RH_out      -0.148  0.2096
## Windspeed   -0.034  0.1425
## Visibility  -0.085  0.0061
## Tdewpoint    0.248  0.1602
```

Below are six days with highest PC1 score and another six with lowest PC1 scores. Temperature values comparison between the two groups confirms that higher PC1 means higher temperature.

```
#pca scores to data frame
epcascores= data.frame(energydata_daily.pca$scores)
#sort by Comp.1 in ascending order
epcascores = epcascores[order(-epcascores$Comp.1),]

#check the energy data for six of the highest Comp1
energydata_daily[rownames(head(epcascores)),rownames(pPC1_2)[abs(pPC1_2[,1])>
0.2]]

##            T1 T2 T3 T4 T5 T6 RH_6 T7 T8 T9 T_out Tdewpoint
## 2016-05-11 25 25 26 25 24 18   37 25 26 24    18        13
## 2016-05-13 25 25 27 25 24 17   27 25 26 24    17        11
## 2016-05-12 25 24 26 25 24 18   33 25 26 24    17        12
```

```
## 2016-05-10 25 24 27 25 24 18   37 24 25 24    17        14
## 2016-05-22 25 24 27 24 23 18   40 24 25 23    16        14
## 2016-05-27 24 24 27 25 23 20   17 24 24 23    17        12

#check energy data for six of the lowest Comp1
energydata_daily[rownames(tail(epcascores)),rownames(pPC1_2)[abs(pPC1_2[,1])>
0.2]]

##            T1 T2 T3 T4 T5    T6 RH_6 T7 T8 T9 T_out Tdewpoint
## 2016-02-18 20 19 20 19 18  1.94   72 18 20 18  1.17     -1.30
## 2016-01-19 19 18 19 19 17 -4.25   86 18 20 17 -2.93     -5.48
## 2016-01-23 17 17 18 15 15  6.41  100 16 17 15  5.86      5.42
## 2016-01-21 19 18 19 18 17 -0.38   94 17 19 16  0.18     -3.06
## 2016-01-22 19 18 19 16 16  1.43   95 16 18 16  1.60     -0.58
## 2016-01-20 19 18 19 17 17 -1.28   93 17 20 16 -1.62     -3.33
```

Below are six days with highest PC2 score and another six with lowest PC2 score. Humidity values comparison between the two groups confirms that higher PC2 means higher humidity.

```
#sort by Comp.2 in ascending order
epcascores = epcascores[order(-epcascores$Comp.2),]

#check the energy data for six of the highest Comp2
energydata_daily[rownames(head(epcascores)),rownames(pPC1_2)[abs(pPC1_2[,2])>
0.2]]

##            RH_1 RH_2 RH_3 RH_4 RH_5 RH_6 RH_7 RH_8 RH_9 RH_out
## 2016-02-02   47   45   46   48   61   99   44   52   50     91
## 2016-01-27   46   45   45   47   56   91   46   50   49     86
## 2016-05-22   52   51   46   49   54   40   46   53   50     91
## 2016-01-24   44   43   45   46   58  100   42   52   48     95
## 2016-02-05   46   44   45   47   61   95   43   52   49     94
## 2016-01-30   45   44   45   46   58   96   42   52   47     87

#check energy data for six of the lowest Comp2
energydata_daily[rownames(tail(epcascores)),rownames(pPC1_2)[abs(pPC1_2[,2])>
0.2]]

##            RH_1 RH_2 RH_3 RH_4 RH_5 RH_6 RH_7 RH_8 RH_9 RH_out
## 2016-04-28   35   35   35   34   47 32.0   28   38   34     77
## 2016-03-17   35   36   35   33   47 37.6   26   36   38     71
## 2016-05-15   36   36   36   34   40 21.8   30   36   36     73
## 2016-05-04   35   33   35   33   46 15.4   32   39   37     62
## 2016-05-14   36   36   33   34   45  9.9   31   37   37     64
## 2016-05-05   33   31   35   32   46  9.9   31   38   37     52
```

Cluster Analysis

The "plot.wgss" function will be used to check the number of clusters appropriate for the energy data for Kmeans clustering.

```
plot.wgss = function(mydata, maxc) {
  wss = numeric(maxc)
  for (i in 1:maxc)
    wss[i] = kmeans(mydata,centers=i, nstart = 10)$tot.withinss
  plot(1:maxc, wss, type="b", xlab="Number of Clusters",
  ylab="Within groups sum of squares", main="Scree Plot")
}
plot.wgss(energydata_daily, 10)
```

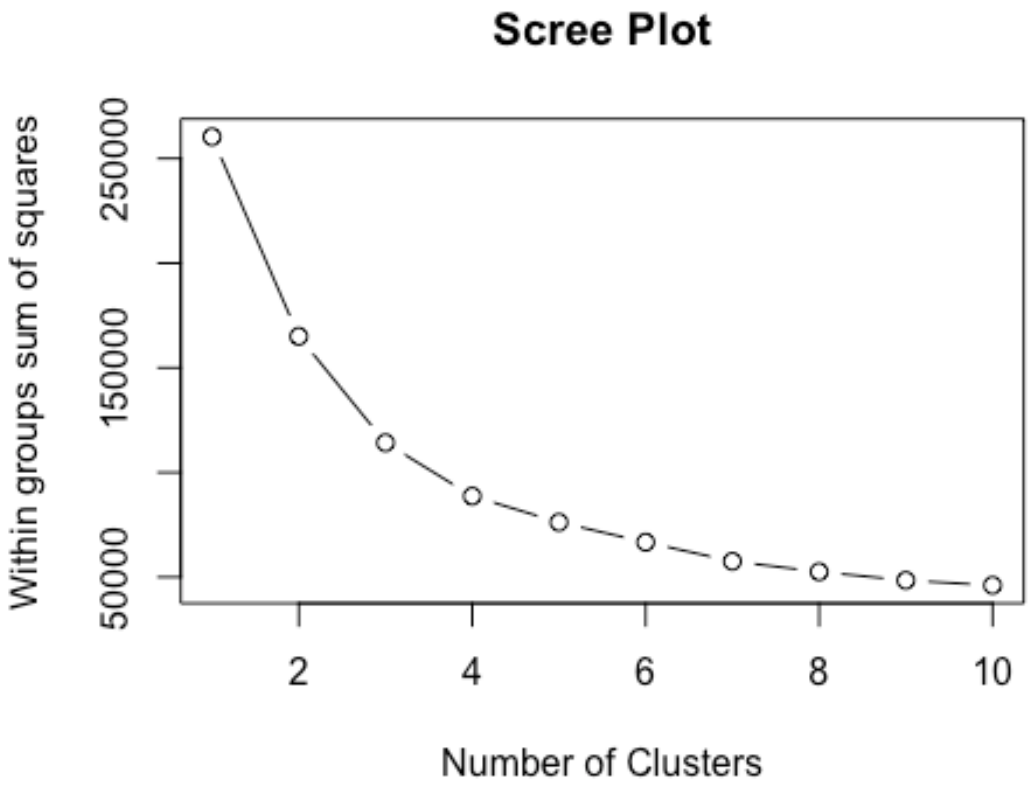

Figure 8

The "elbow Test" suggests that 3 clusters are appropriate for the energy data. The three cluster are identified in the plot below between the two PCs.

```
energy_km <- kmeans(energydata_daily[,1:20],3)
plot(energydata_daily.pca$scores[, 1:2], pch=".", cex= 2.5)
text(energydata_daily.pca$scores[, 1:2],labels=substr(row.names(energydata_da
ily),6,10), col = energy_km$cluster)
```

The three clusters for the whole energy data have overlaps when the two PCs value are used as shown in figure 9. It's possible that the third cluster would be more apparently along the third dimension.

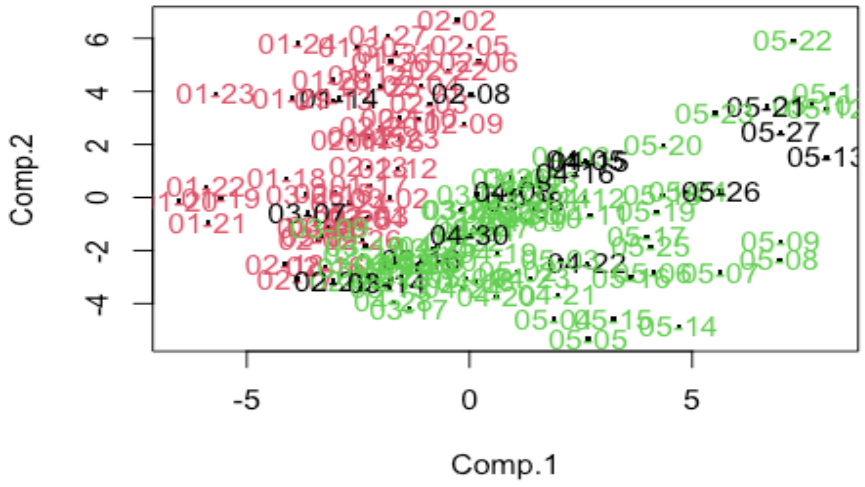

Figure 9

However, clustering based only on temperature or humidity data shows that the three clusters are well defined. Temperature values aligned with PC1 fall into distinct low, medium, and high clusters along the x-axis as shown on figure 10.

```
tempZone = c("T1","T2", "T3", "T4","T5", "T7","T8","T9")
energy_km.t <- kmeans(energydata_daily[,tempZone],3)
plot(energydata_daily.pca$scores[, 1:2], pch=".", cex= 2.5)
text(energydata_daily.pca$scores[, 1:2],labels= substr(row.names(energydata_d
aily),6,10), col = energy_km.t$cluster)
```

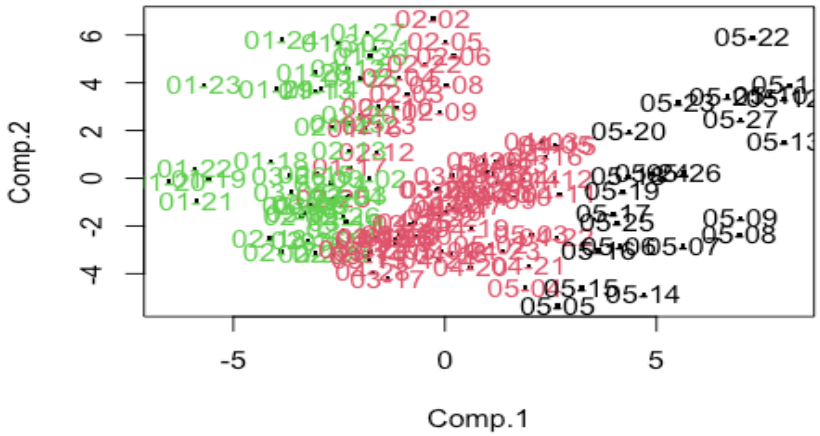

Figure 10

Humidity values aligned with PC2 fall into similar categories along the y-axis as showing in Figure 11.

```
humZone = c("RH_1","RH_2", "RH_3", "RH_4","RH_5","RH_7","RH_8","RH_9")
energy_km.h <- kmeans(energydata_daily[,humZone],3)
plot(energydata_daily.pca$scores[, 1:2], pch=".", cex= 2.5)
text(energydata_daily.pca$scores[, 1:2],labels=substr(row.names(energydata_da
ily),6,10), col = energy_km.h$cluster)
```

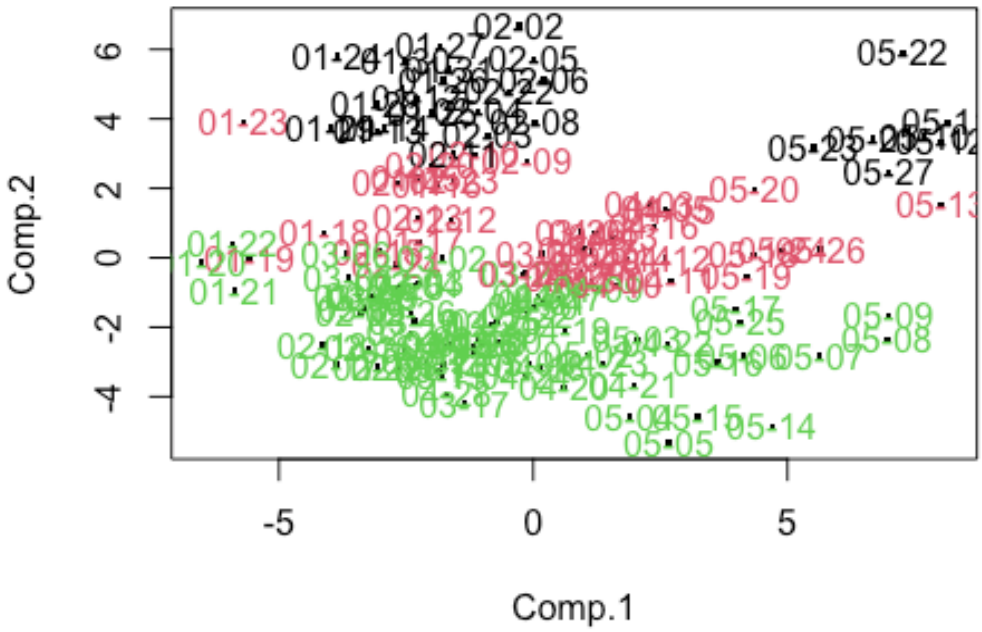

Figure 11