ENERGY USE OF APPLIANCES

Yosef WoldeamanuelJonathan 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 refrigerators, modems ,and security cameras are required 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, cluster analysis, and confirmatory factor 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. Our model for confirmatory factor analysis was unfortunately not supported by the data, but some approaches for future model attempts are discussed below.

Attribute Information:

Date	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 north side of building in Celsius
RH_6	Humidity outside north side of building 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 in Celsius - Chievres weather station
Pressure	In mm Hg - Chievres weather station
RH_out	Humidity outside in % - Chievres weather station
Wind speed	In m/s - Chievres weather station

Visibility	In km - Chievres weather station
Tdewpoint	Â°C - Chievres weather station
rv1	Random variable 1 nondimensional
rv2	Random variable 2 nondimensional

Data Cleaning

We begin by importing the raw "data energy" csv and examining the first few rows.

```
##
## The downloaded binary packages are in
##
/var/folders/1n/nvr79nb55tz9j4lsbrw6hsf80000gn/T//RtmpK5hwP6/downloaded_packa
ges

## # 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>
```

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

```
## # A tibble: 6 x 25
##   date                Appliances lights    T1  RH_1    T2  RH_2    T3
RH_3
##   <dttm>                   <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
<dbl>
## 1 2016-01-11 17:00:00         60     30  19.9  47.6  19.2  44.8  19.8
44.7
## 2 2016-01-11 17:10:00         60     30  19.9  46.7  19.2  44.7  19.8
44.8
```

```
## 3 2016-01-11 17:20:00         50     30  19.9  46.3  19.2  44.6  19.8
44.9
## 4 2016-01-11 17:30:00         50     40  19.9  46.1  19.2  44.6  19.8  45
## 5 2016-01-11 17:40:00         60     40  19.9  46.3  19.2  44.5  19.8  45
## 6 2016-01-11 17:50:00         50     40  19.9  46.0  19.2  44.5  19.8
44.9
## # … with 16 more variables: T4 <dbl>, RH_4 <dbl>, T5 <dbl>, RH_5 <dbl>,
## #   T7 <dbl>, RH_7 <dbl>, T8 <dbl>, RH_8 <dbl>, T9 <dbl>, RH_9 <dbl>,
## #   T_out <dbl>, Press_mm_hg <dbl>, RH_out <dbl>, Windspeed <dbl>,
## #   Visibility <dbl>, Tdewpoint <dbl>
```

The energy data have 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. Now we can see a single row for each day.

```
##            Appliances lights T1 RH_1 T2 RH_2 T3 RH_3 T4 RH_4 T5 RH_5 T7
RH_7 T8
## 2016-01-11        137   30.0 21   47 20   45 20   46 19   47 18   59 18
43 19
## 2016-01-12         86    4.2 20   45 19   44 20   45 20   45 18   51 18
42 19
## 2016-01-13         97    5.4 19   43 19   42 20   44 19   42 18   58 18
40 19
## 2016-01-14        151    5.0 20   42 20   41 21   43 19   43 18   58 18
40 19
## 2016-01-15        125    6.0 22   39 22   38 21   41 20   42 19   52 18
39 19
## 2016-01-16        125    8.0 22   40 21   39 21   42 21   42 19   53 19
40 20
##            RH_8 T9 RH_9 T_out Press_mm_hg RH_out Windspeed Visibility
Tdewpoint
## 2016-01-11   49 17   45   5.9         735     89       6.1         41
4.23
## 2016-01-12   50 17   46   5.6         743     88       5.8         37
3.60
## 2016-01-13   49 17   45   4.9         755     83       5.6         32
2.17
## 2016-01-14   47 17   45   3.4         750     86       6.3         35
1.32
## 2016-01-15   46 17   45   2.7         755     88       7.8         40
0.85
## 2016-01-16   47 18   44   2.2         763     90       3.5         35
0.54
```

Let’s use “PerformanceAnalytics” library to check for scatter plots between variables, correlation and histogram of some of the variables which will help to see if there extreme outliers.

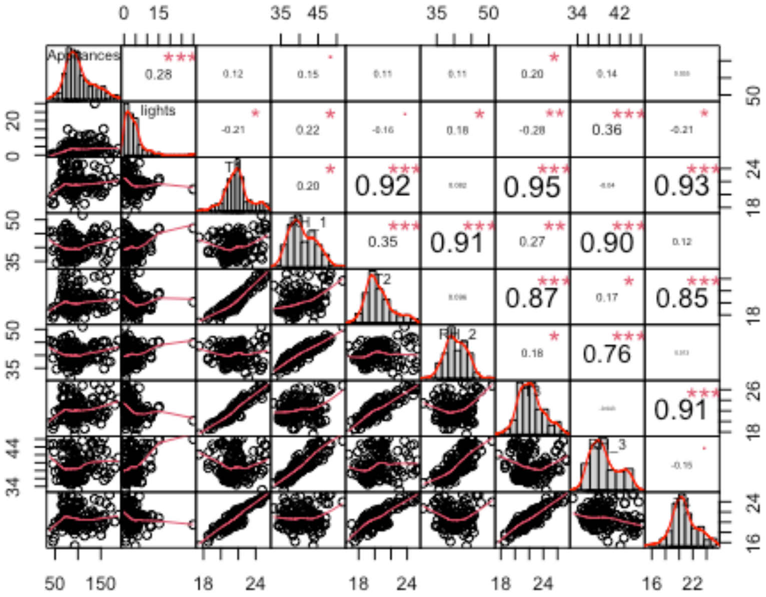


Figure 1: Variable Scatter Plots

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 internet-of-things (IoT) equipment. The MVA library will be used to check which rows are with extreme outliers and clean them up.

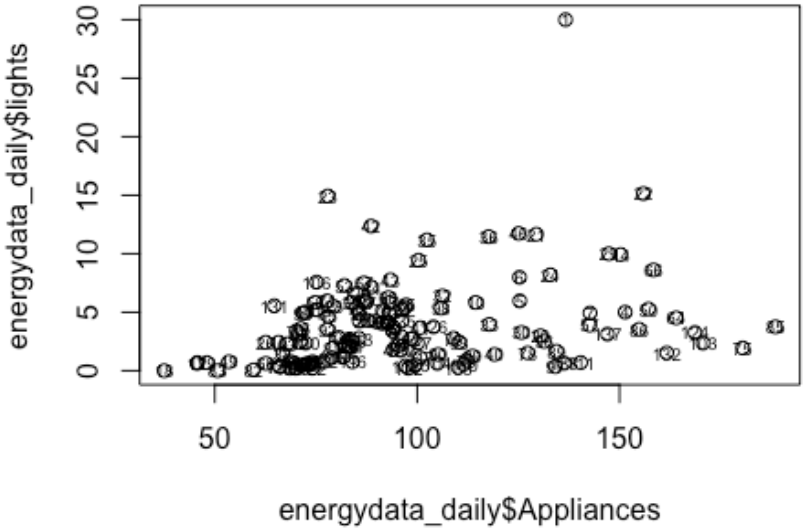


Figure 2: Appliance vs Lights

The scatter plot 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.

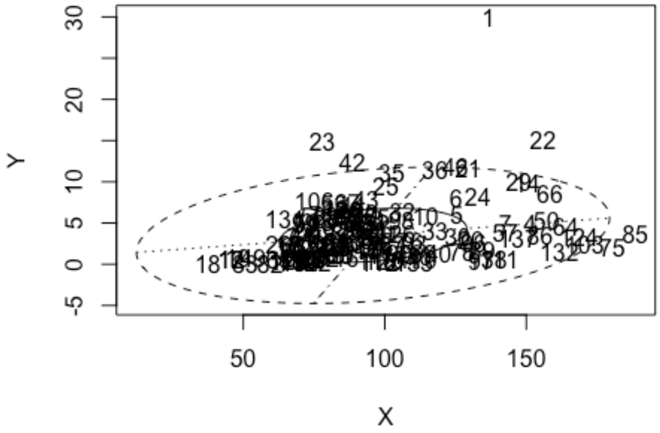


Figure 3: BV Box

There are 6 rows outside of the ellipse which will be dropped to reduce the variation on the data analysis.

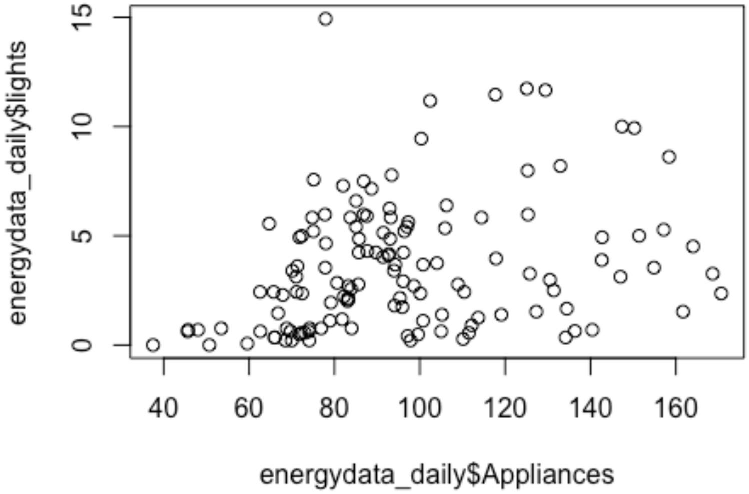


Figure 4: Outliers Removed

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

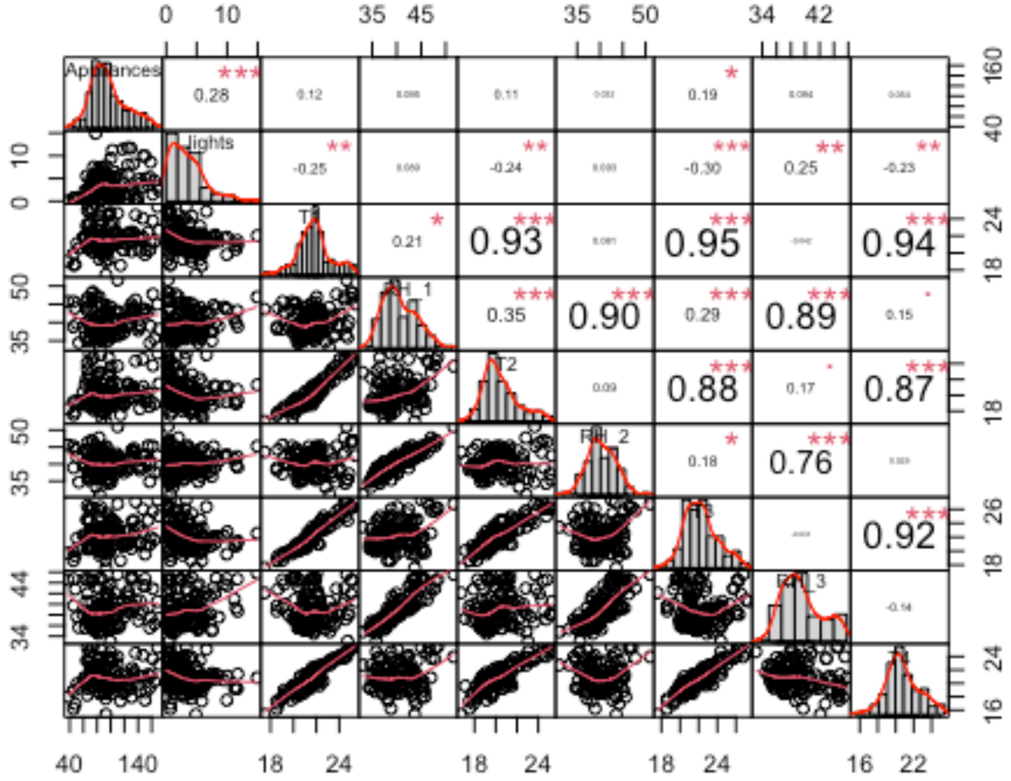


Figure 5: Cleaned Scatter Plots

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

Correlation Results

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

Temperature and humidity at different area of the house has a positive correlation with appliances 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.

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.

Graphical Dimension Reduction Analysis (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.

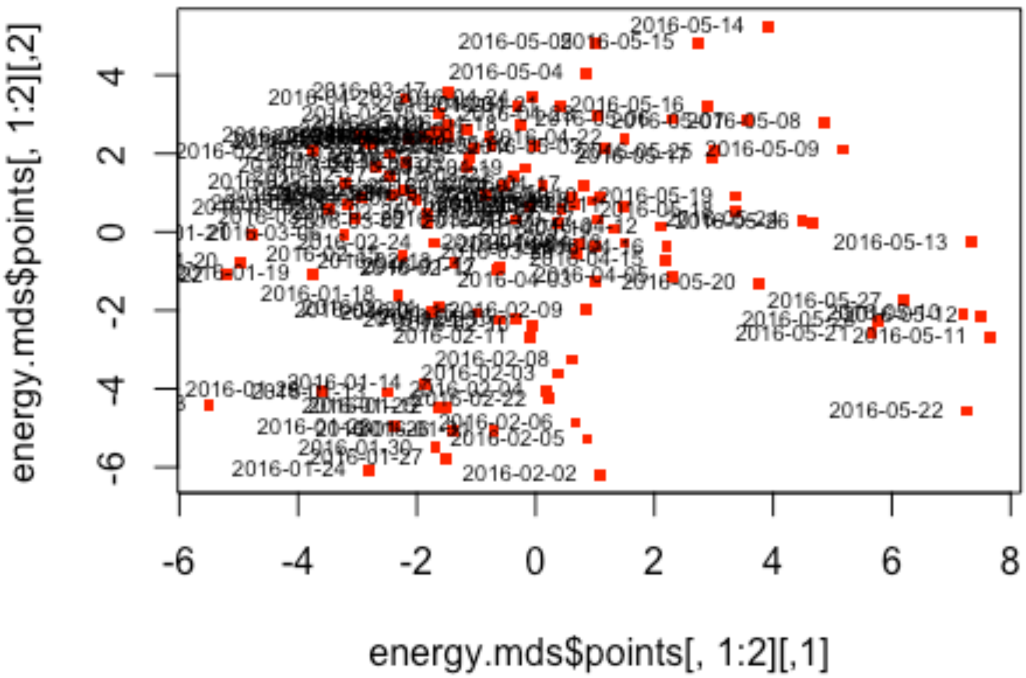


Figure 6: Graphical MDA Rows

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.

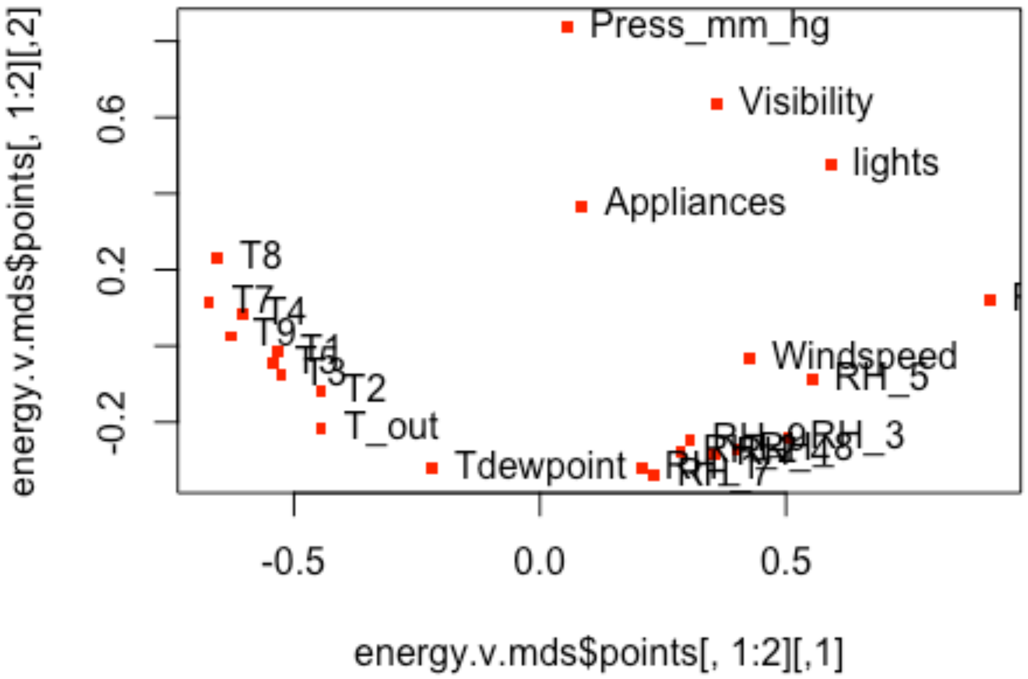


Figure 7: Graphical MDA Variables

Principle 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.

The 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.

```
## Importance of components:
##                          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
Comp.8
## Standard deviation          3.05   2.75  1.181  1.087  1.009  0.924  0.812
0.771
## Proportion of Variance      0.39   0.31  0.058  0.049  0.042  0.036  0.027
0.025
## Cumulative Proportion       0.39   0.70  0.761  0.811  0.853  0.889  0.916
0.941
##                          Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14
Comp.15
## Standard deviation         0.655   0.533  0.4581  0.3274  0.2911  0.2636
```

```
0.2420
## Proportion of Variance  0.018   0.012  0.0087  0.0045  0.0035  0.0029
0.0024
## Cumulative Proportion   0.959   0.971  0.9793  0.9838  0.9873  0.9902
0.9927
##                        Comp.16 Comp.17 Comp.18 Comp.19 Comp.20 Comp.21
Comp.22
## Standard deviation      0.2374  0.1977  0.1654 0.12810 0.12707 0.10190
0.08774
## Proportion of Variance  0.0023  0.0016  0.0011 0.00068 0.00067 0.00043
0.00032
## Cumulative Proportion   0.9950  0.9966  0.9978 0.99847 0.99914 0.99957
0.99989
##                        Comp.23 Comp.24
## Standard deviation     4.3e-02 2.8e-02
## Proportion of Variance 7.6e-05 3.2e-05
## Cumulative Proportion  1.0e+00 1.0e+00
```

```
##             Comp.1  Comp.2
## Appliances   0.028  0.0275
## lights      -0.093  0.1158
## T1           0.303 -0.0998
## RH_1         0.167  0.2982
## T2           0.302 -0.0337
## RH_2         0.121  0.2777
## T3           0.308 -0.0875
## RH_3         0.073  0.3415
## T4           0.287 -0.1367
## RH_4         0.125  0.3290
## T5           0.302 -0.1022
## RH_5         0.011  0.2640
## T7           0.277 -0.1668
## RH_7         0.157  0.2992
## T8           0.251 -0.1810
## RH_8         0.107  0.3281
## T9           0.293 -0.1441
## RH_9         0.136  0.3121
## T_out        0.295 -0.0086
## Press_mm_hg -0.088 -0.0857
## RH_out      -0.113  0.2409
## Windspeed   -0.014  0.1537
## Visibility  -0.088  0.0239
## Tdewpoint    0.283  0.1085
```

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.

```
##            T1 T2 T3 T4 T5 T7 T8 T9 T_out Tdewpoint
## 2016-05-11 25 25 26 25 24 25 26 24    18        13
## 2016-05-12 25 24 26 25 24 25 26 24    17        12
## 2016-05-10 25 24 27 25 24 24 25 24    17        14
## 2016-05-13 25 25 27 25 24 25 26 24    17        11
## 2016-05-22 25 24 27 24 23 24 25 23    16        14
## 2016-05-27 24 24 27 25 23 24 24 23    17        12

##            T1 T2 T3 T4 T5 T7 T8 T9 T_out Tdewpoint
## 2016-02-18 20 19 20 19 18 18 20 18  1.17     -1.30
## 2016-01-19 19 18 19 19 17 18 20 17 -2.93     -5.48
## 2016-01-23 17 17 18 15 15 16 17 15  5.86      5.42
## 2016-01-22 19 18 19 16 16 16 18 16  1.60     -0.58
## 2016-01-21 19 18 19 18 17 17 19 16  0.18     -3.06
## 2016-01-20 19 18 19 17 17 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.

```
##            RH_1 RH_2 RH_3 RH_4 RH_5 RH_7 RH_8 RH_9 RH_out
## 2016-02-02   47   45   46   48   61   44   52   50     91
## 2016-01-24   44   43   45   46   58   42   52   48     95
## 2016-01-27   46   45   45   47   56   46   50   49     86
## 2016-01-30   45   44   45   46   58   42   52   47     87
## 2016-01-31   43   42   45   45   55   43   52   49     93
## 2016-02-05   46   44   45   47   61   43   52   49     94

##            RH_1 RH_2 RH_3 RH_4 RH_5 RH_7 RH_8 RH_9 RH_out
## 2016-04-21   36   36   36   34   44   32   40   38     60
## 2016-03-17   35   36   35   33   47   26   36   38     71
## 2016-05-04   35   33   35   33   46   32   39   37     62
## 2016-05-15   36   36   36   34   40   30   36   36     73
## 2016-05-05   33   31   35   32   46   31   38   37     52
## 2016-05-14   36   36   33   34   45   31   37   37     64
```

Cluster Analysis

The “plot.wgss” function uses the within-group sum-of-squares to determine the number of clusters appropriate for the energy data for K-means clustering.

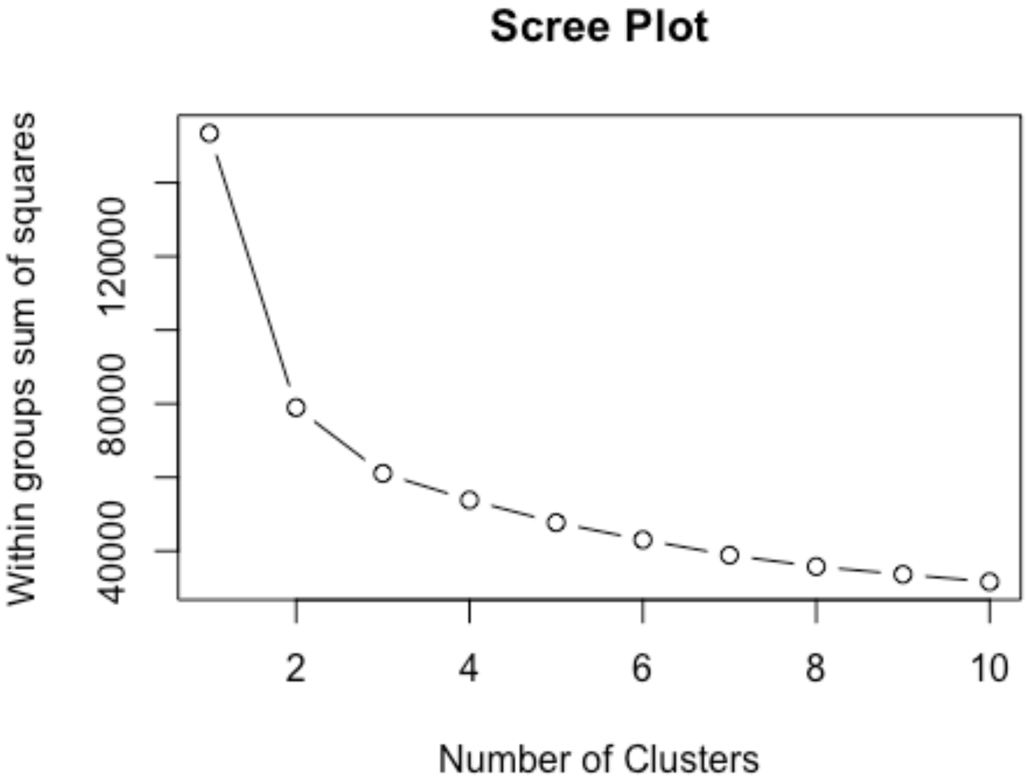


Figure 8: Scree Plot

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.

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.

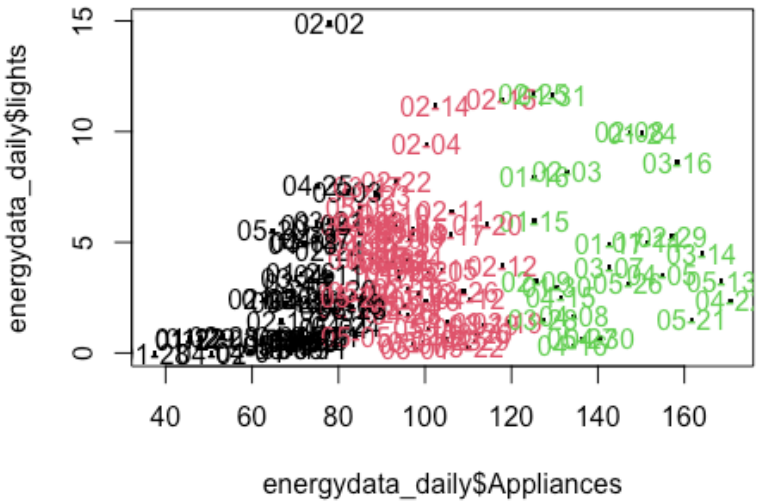


Figure 9: K-Means PC Values

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.

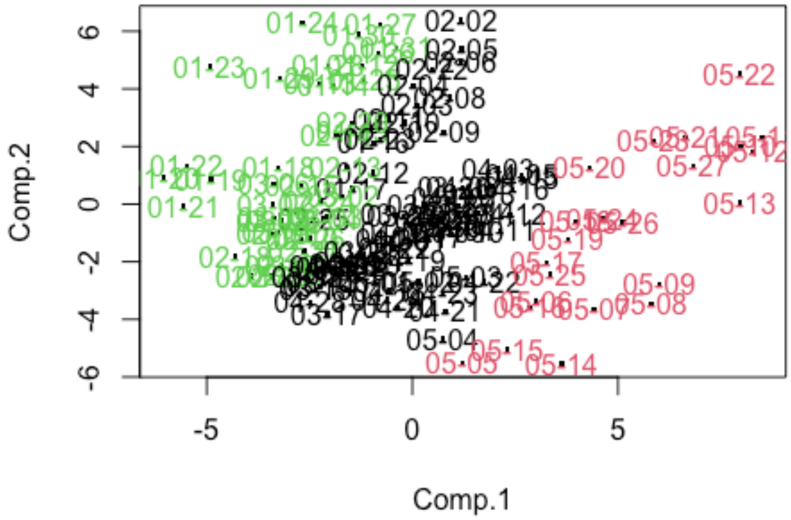


Figure 10: K-Means PC1

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

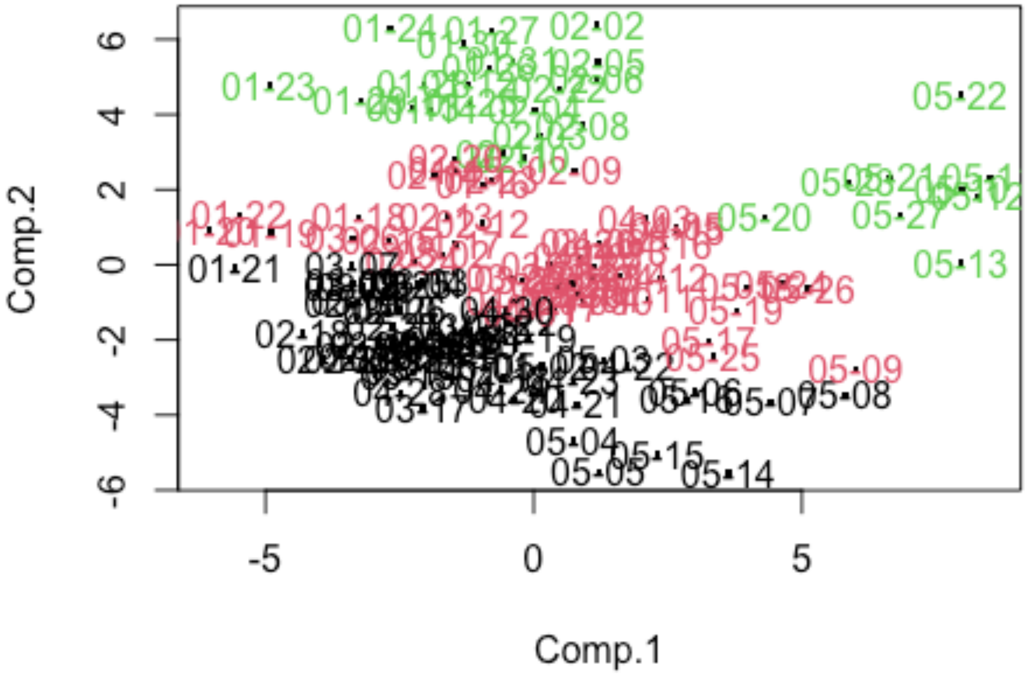
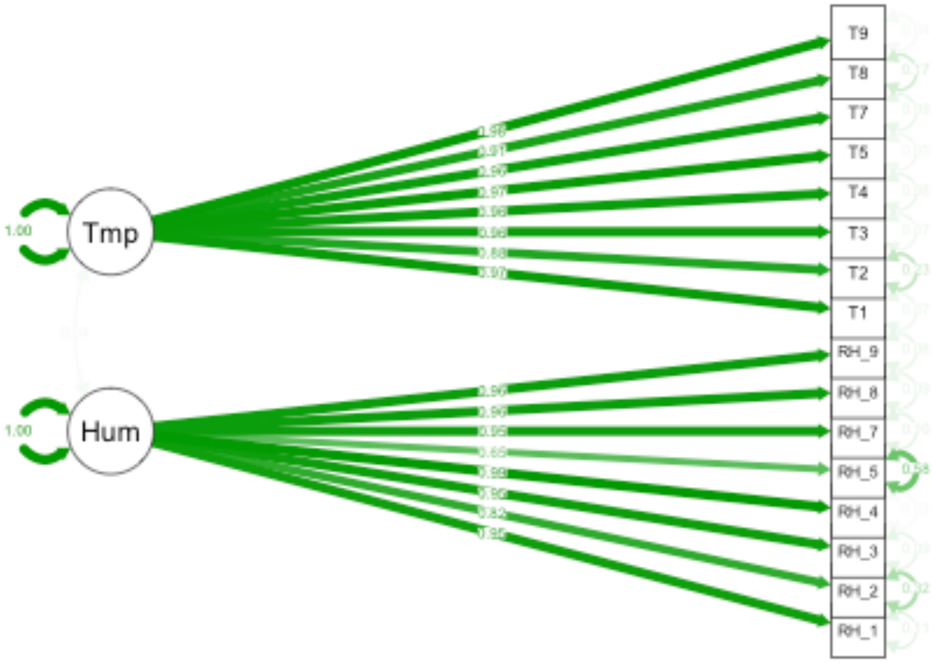


Figure 11: K-Means PC2

Confirmatory Factor Analysis

In confirmatory factor analysis (CFA), we want to test whether factors correlate to specific manifest variables. Our model tests whether that is the case with the temperature and humidity variables. Remember in our case that we have several instances of both in different locations.

The plot can help us visualize the model we've created with the corresponding correlations.The result shows extremely high correlations across the board, which we would expect since we have grouped variables that measure the same phenomena in close



areas.

This can be inspected more closely in the summary. Additionally, we can add some fit indices to test whether the covariance matrix produced by our model fits the non-retricted covariance matrix of the data.

For the data to support our model, we need a root-mean-square-error (SRMR) of less than 0.05, and both goodness-of-fits to be greater than 0.95. The fits are printed below and we conclude that our data doesn't support our CFA model.

```
## [1] "RMSE: 0.115697075478007"

## [1] "Goodness-of-Fit: 0.490514948020036"

## [1] "Adjusted-Goodness-of-Fit: 0.327281873113834"

## [1] "Data does not support the designed CFA model. MODEL IS NOT
CONFIRMED!"
```

Conclusion

From the very first correlation test, we could see that energy use of appliances and lights had a strong relationship with measures of temperature and humidity inside the house. However, this is a difficult data set for the application of multivariate analysis. Not because the data is time series, but because the conceptual variables of “temperature” and “humidity” are actually represented as several variables corresponding to different locations both inside and outside the house. This distorts MVA techniques because the correlation and distance between similar items makes it difficult to read any relationships with the energy-use variables.

The graphical dimensional reduction analysis did show that temperatures from room to room differed in their relationship to appliance energy use. Future analysis should consider aggregating the temperature variables even further, perhaps by section of the house or proximity to specific appliances, so these differences can be examined further.

Both the principal component and cluster analyses demonstrated how useful these tools can be in grouping variables because in our case, we knew which variables should exhibit clusters in the first place. PCA demonstrated this by finding principal components of temperature and humidity as PC1 and PC2 respectively. These were highlighted by K-means clustering when we colored the results separately based on temperature and humidity vectors in the data.

Finally, our confirmatory factor analysis model did not prove to be supported by our original covariance matrix. This may be due to how close the individual temperature and humidity variables were to each other already, which may create some tight tolerances that aren’t useful for CFA. Again, an interesting approach may be to try a model where sections of the house are used as the factors corresponding to the climate readings in specific rooms.

Ultimately, this was an extremely useful data set for understanding both the strengths and limitations of multivariate analysis. While we may not have been able to find many new relationships, it was informative to see how pre-existing clusters revealed themselves as we learned new tools and techniques. In addition, while there were several rich data points collected from this house, the house itself only represents an n of 1 for any population of households we would want to examine to make conclusive statements about energy use and in-house climate more generally. With a larger sample of homes, the findings of multivariate methods applied here could reveal far more interesting relationships.

Appendix

Tables

Principal Component Analysis

```
## Importance of components:
##                         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
Comp.8
## Standard deviation        3.05   2.75  1.181  1.087  1.009  0.924  0.812
0.771
## Proportion of Variance    0.39   0.31  0.058  0.049  0.042  0.036  0.027
0.025
## Cumulative Proportion     0.39   0.70  0.761  0.811  0.853  0.889  0.916
0.941
```

```
##                         Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14
Comp.15
## Standard deviation      0.655   0.533  0.4581  0.3274  0.2911  0.2636
0.2420
## Proportion of Variance  0.018   0.012  0.0087  0.0045  0.0035  0.0029
0.0024
## Cumulative Proportion   0.959   0.971  0.9793  0.9838  0.9873  0.9902
0.9927
##                         Comp.16 Comp.17 Comp.18 Comp.19 Comp.20 Comp.21
Comp.22
## Standard deviation      0.2374  0.1977  0.1654 0.12810 0.12707 0.10190
0.08774
## Proportion of Variance  0.0023  0.0016  0.0011 0.00068 0.00067 0.00043
0.00032
## Cumulative Proportion   0.9950  0.9966  0.9978 0.99847 0.99914 0.99957
0.99989
##                         Comp.23 Comp.24
## Standard deviation     4.3e-02 2.8e-02
## Proportion of Variance 7.6e-05 3.2e-05
## Cumulative Proportion  1.0e+00 1.0e+00
```

```
##              Comp.1  Comp.2
## Appliances   0.028  0.0275
## lights      -0.093  0.1158
## T1           0.303 -0.0998
## RH_1         0.167  0.2982
## T2           0.302 -0.0337
## RH_2         0.121  0.2777
## T3           0.308 -0.0875
## RH_3         0.073  0.3415
## T4           0.287 -0.1367
## RH_4         0.125  0.3290
## T5           0.302 -0.1022
## RH_5         0.011  0.2640
## T7           0.277 -0.1668
## RH_7         0.157  0.2992
## T8           0.251 -0.1810
## RH_8         0.107  0.3281
## T9           0.293 -0.1441
## RH_9         0.136  0.3121
## T_out        0.295 -0.0086
## Press_mm_hg -0.088 -0.0857
## RH_out      -0.113  0.2409
## Windspeed   -0.014  0.1537
## Visibility  -0.088  0.0239
## Tdewpoint    0.283  0.1085
```

Source Code

Data Cleaning

Library loadings and imports:

```
library(readr)
library(mclust)
options(digits = 2)
install.packages("CCA", repos = "http://cran.us.r-project.org")

energydata_complete <- read_csv(here("energydata_complete.csv"))
head(energydata_complete[,c(1:5,26:29)])
```

Dropping extraneous data:

```
randomv =c("T6","RH_6", "rv1","rv2")
energydata_complete = energydata_complete[ , !(names(energydata_complete)
%in% randomv)]
head(energydata_complete)
```

Aggregation with Zoo:

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

energydata_daily = as.data.frame(energydata_daily)
head(energydata_daily)
```

PerformanceAnalytics scatter plots to identify outliers:

```
library("PerformanceAnalytics")
chart.Correlation(energydata_daily[,1:9], histogram=TRUE, pch=19, col="")
```

Checking plot for outliers:

```
#check outliers
library(MVA)
plot(energydata_daily$Appliances, energydata_daily$lights)
text(energydata_daily$Appliances, energydata_daily$lights, cex = 0.6)
```

Confirming with bvbox:

```
bvbox(energydata_daily[,c("Appliances","lights")], type = "n")
text(energydata_daily[,c("Appliances","lights")]$Appliances,
energydata_daily[,c("Appliances","lights")]$lights)
```

Dropping outliers and final plot:

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

```
chart.Correlation(energydata_daily[,1:9], histogram=TRUE, pch=19, col="")
```

Scaling data:

```
daily.s = scale(energydata_daily)
head(daily.s)
daily_corr = cor(daily.s)
daily_corr[1:9,1:9]
```

Graphical MDA

MDA with distance matrix:

```
energy_dist = dist(daily.s[,c(1:18)])
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)
```

MDA with correlation matrix:

```
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)
```

Principal Component Analysis

PCA setup:

```
energydata_daily.pca <- princomp(energydata_daily, cor = T)
summary(energydata_daily.pca)
pPC1_2 = energydata_daily.pca$loadings[,c(1:2)]
pPC1_2
```

PC1 (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]]
#check energy data for six of the lowest Comp1
energydata_daily[rownames(tail(epcascores)),rownames(pPC1_2)[abs(pPC1_2[,1])>
0.2]]
```

PC2 (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]]
#check energy data for six of the lowest Comp2
energydata_daily[rownames(tail(epcascores)),rownames(pPC1_2)[abs(pPC1_2[,2])>
0.2]]
```

Cluster Analysis

Using WGSS to find number of kmeans clusters:

```
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)
```

K-means plot:

```
energy_km <- kmeans(energydata_daily[,1:18],3)

plot(energydata_daily$Appliances, energydata_daily$lights, pch=".", cex= 2.5)
text(energydata_daily$Appliances,
energydata_daily$lights,labels=substr(row.names(energydata_daily),6,10), col
= energy_km$cluster)
```

K-means colored for temperature:

```
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_daily),6,10), col = energy_km.t$cluster)
```

K-means colored for humidity:

```
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_daily),6,10), col =
energy_km.h$cluster)
```

Confirmatory Factor Analysis

Setting up CFA model with sem package:

```
library(sem)
energy_model <- specifyModel(file="energy_model.txt")
energy_sem <- sem(energy_model, cor(daily.s), nrow(daily.s))
#summary(energy_sem)
```

CFA Model (energy_consumption.txt):

```
Temp     -> T1, lambda1, NA
Temp     -> T2, lambda2, NA
Temp     -> T3, lambda3, NA
Temp     -> T4, lambda4, NA
Temp     -> T5, lambda5, NA
Temp     -> T7, lambda7, NA
Temp     -> T8, lambda8, NA
Temp     -> T9, lambda9, NA
Hum     -> RH_1, lambda11, NA
Hum     -> RH_2, lambda12, NA
Hum     -> RH_3, lambda13, NA
Hum     -> RH_4, lambda14, NA
Hum     -> RH_5, lambda15, NA
Hum     -> RH_7, lambda17, NA
Hum     -> RH_8, lambda18, NA
Hum     -> RH_9, lambda19, NA
Temp     <-> Hum, rho, NA
T1     <-> T1, theta1, NA
T2     <-> T2, theta2, NA
T3     <-> T3, theta3, NA
T4     <-> T4, theta4, NA
T5     <-> T5, theta5, NA
T7     <-> T7, theta7, NA
T8     <-> T8, theta8, NA
T9     <-> T9, theta9, NA
RH_1     <-> RH_1, theta11, NA
RH_2     <-> RH_2, theta12, NA
RH_3     <-> RH_3, theta13, NA
RH_4     <-> RH_4, theta14, NA
RH_5     <-> RH_5, theta15, NA
RH_7     <-> RH_7, theta17, NA
RH_8     <-> RH_8, theta18, NA
RH_9     <-> RH_9, theta19, NA
Temp <-> Temp, NA, 1
Hum <-> Hum, NA, 1
```

Plot of the model:

```
library(semPlot)
semPaths(energy_sem, rotation = 2, 'std', 'est')
```

Testing if data supports the model:

```r
options(fit.indices = c("GFI", "AGFI", "SRMR"))
energy.crtr = summary(energy_sem)

#print summary
#summary(energy_sem)

print(paste0("RMSE: ", energy.crtr$SRMR))

print(paste0("Goodness-of-Fit: ", energy.crtr$GFI))

print(paste0("Adjusted-Goodness-of-Fit: ", energy.crtr$AGFI))

#test case for whether the data supports our model
if(energy.crtr$SRMR < 0.05 && energy.crtr$GFI > 0.95 && energy.crtr$AGFI >
0.95){
  print("Data support the designed CFA model. MODEL is CONFIRMED!")
}else{
  print("Data does not support the designed CFA model. MODEL IS NOT
CONFIRMED!")
}
```