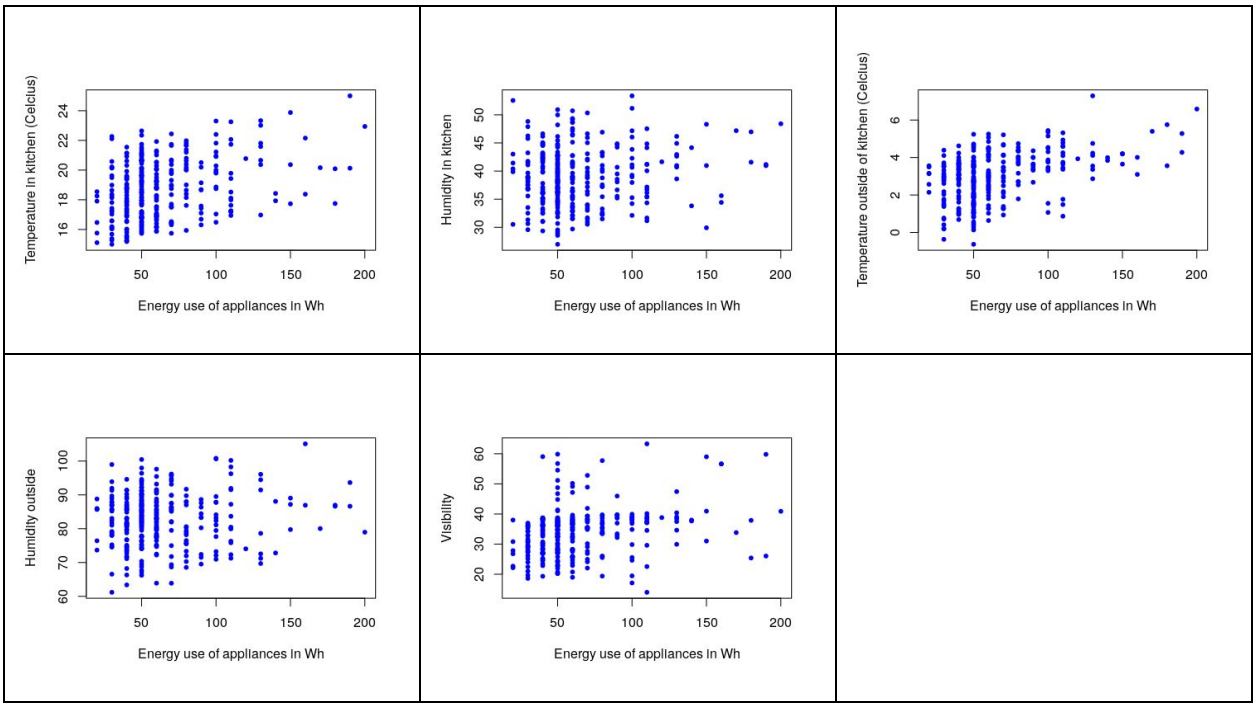
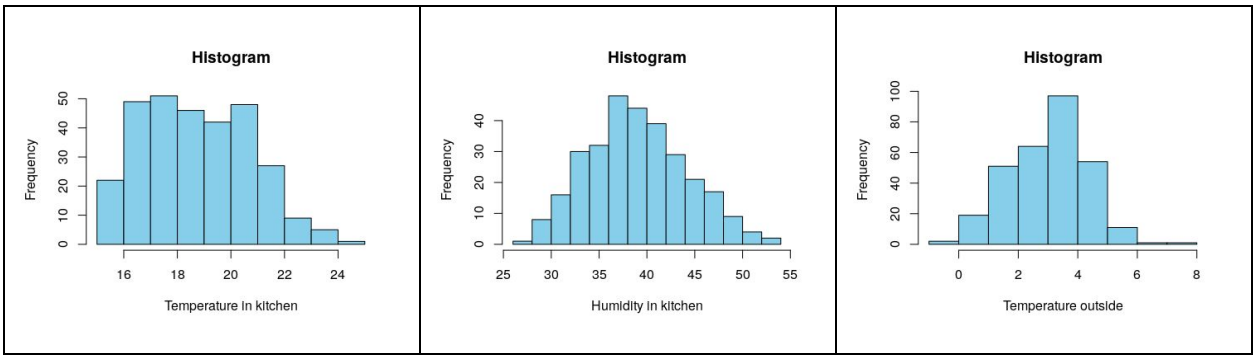
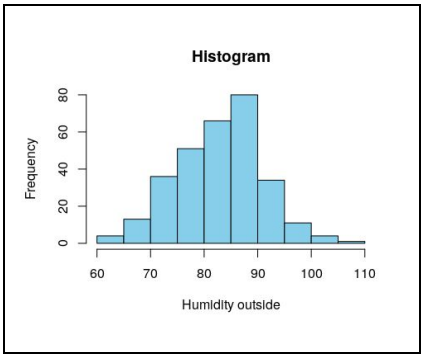
Scatter plots

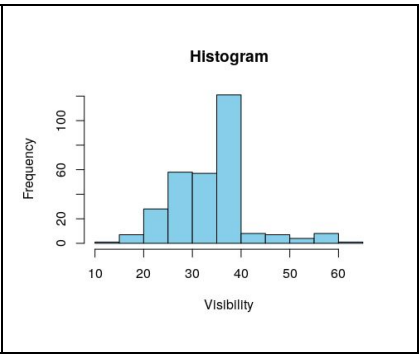
Scatter plots indicates that first four variables have low positive correlation with Y (Energy use of appliances) and among them second variable has lowest with many outliers than others. Visibility from the station seems to have no correlation upon visual inspection and has many outliers. This variable can be ignored while creating models for prediction. But, this is just assumption made visually and might not be very accurate.

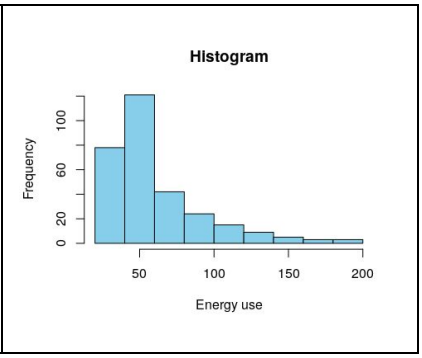


Histograms









Histograms shows that almost all variables are positively skewed except Humidity outside which looks slightly negatively skewed. Humidity in kitchen is least skewed where Energy use of appliances (Y) has the most skewed data. Variable 1, 3 and 4 are slightly skewed so we might be able use less extreme transformation technique such as cube root or square root on them. But, for Y we might have to use log or log + 1 transformation as it is heavily skewed towards on right side i.e. positively skewed (Corey Wade. 2019).

Transformation

I selected temperature inside and outside kitchen and humidity inside and outside of the kitchen ignoring visibility for further operations in the assignments.

Before transforming the variables I used skewness function from e1071 (CRAN - Package e1071. 2019) library for each variables. Negative value represents left skew.

V1	V2	V3	V4	V6
0.2841859	0.2500231	-0.1818761	-0.1063660	1.6146097

And as I mentioned before, I applied cube root transformation on each variables first. But, third variable (temperature in kitchen) was not not properly transformed as some value returned Nan. Since, this variable has skewness value close to 0 which means it is very close to normal distribution I used original value for this variable. For 4th variable (humidity outside kitchen) applying cube root, square root, log and log + 1 transformation made skewness even worse (more skewed towards left). Also, this variable originally has skewness close to 0 so I used original values for this variable as well. For Y using log transformation improved the distribution.

V1	Cube root transformation
V2	Cube root transformation
V3	No transformation
V4	No transformation
V6	Log transformation

After applying these transformation, I scaled the data to range of 0 - 1 using formula. This process is also known as min-max scaling.

$$V_{new} = \frac{V - min(v)}{max(v) - min(v)}$$

As we are dealing with 3 different unit of data with different magnitudes and range, This step is important as variable with higher values might dominate other variables while generating models and prediction of Y using such model.

Models

Error measures and correlations

Weighted arithmetic mean	RMSE 0.15728679870551 Av. abs error 0.124345603273659 Pearson correlation 0.60982375334709 Spearman correlation 0.57386585845592
Weighted power mean (p = .5)	RMSE 0.159288321146913 Av. abs error 0.125103641330662 Pearson correlation 0.596965614442919 Spearman correlation 0.559245925176568
Weighted power mean (p=2)	RMSE 0.159047539460785 Av. abs error 0.12611968744577 Pearson correlation 0.601472888696089 Spearman correlation 0.561276165137669
Ordered weighted averaging function	RMSE 0.173388031016167 Av. abs error 0.136428128970262 Pearson correlation 0.462320624544056 Spearman correlation 0.399730639852461 Orness 0.478798660749174
Choquet integral	RMSE 0.156414285539891 Av. abs error 0.121910777963189 Pearson Correlation 0.614628210643173 Spearman Correlation 0.567426975262585 Orness 0.511969991737955

Weight parameters

Weighted arithmetic mean	**1** 0.289245845393582 **2** 0 **3** 0.523334126399047 **4** 0.187420028207371
Weighted power mean (p = .5)	**1** 0.262037326234023 **2** 0 **3** 0.478413709043688

	4 0.259548964722289
Weighted power mean (p=2)	**1** 0.323254040611093 **2** 0 **3** 0.547889397496558 **4** 0.128856561892349
Ordered weighted averaging function	**1** 0.356699770255143 **2** 0.229643494795086 **3** 0.0342177173968876 **4** 0.379439017552887
Choquet integral	**shapely** **1** 0.320133475271193 **2** 1.08927219096463e-12 **3** 0.511383218157806 **4** 0.168483306573479 **binary number fm.weights** **1** 0.236800623481652 **2** 0 **3** 0.236800623481652 **4** 0.445574328650061 **5** 0.999999999999994 **6** 0.445574328650061 **7** 0.999999999999994 **8** 0.289465803186014 **9** 0.470308702293209 **10** 0.289465803186014 **11** 0.470308702296476 **12** 0.644034482900204 **13** 1.0000000000003 **14** 0.644034482900204 **15** 1.00000000000357

In my opinion choquet integral model is much better as it has Pearson and spearman correlation closest to 1. This model also has low RMSE and Average absolute error. Models other than Ordered weighted averaging function (OWA) has given no weight to second variable (Humidity in kitchen) but, OWA model has highest error measures and lowest correlation among all the models.

Since, I am going to use choquet model for predicting Y, Humidity in kitchen won't have effect on output value and temperature outside kitchen will have the most effect.

1 = 001= v({1}) = 0.236800623481652

2 = 010 = v({2}) = 0

4 = 100 = v({3}) = 0.445574328650061

8 = 1000 = v({4}) = 0.289465803186014

Individually, third variable (temperature outside kitchen) has most weight hence is the most defining variable among others (FutureLearn. 2019).

If we look at this interaction
v({1,2}) = v({1}) + v({2})
0.2368 = 0.2368 + 0 = 0.2368

It looks like we have a additive interaction. But, since second variable does not have any weight and looking at other weight values
v({1}) + v({3}) = 0.2368 + 0.4455 = 0.6823

Which is less than v({1,3}) = 0.9999 i.e. v({1,3}) > v({1}) + v({3}). Which suggests that we have complimentary interactions between these variables (FutureLearn. 2019).

Our selected model will predict higher value in case of higher input values and lower in cause of lower input values. For better energy usage our model favours lower input values.

Prediction

Using weights obtained from choquet function and provided inputs. I got the predicted value for Y to be 16.27918.

Upon inspection of the provided data, I found that with given input the predicted output value has to be around 50. But, Since our model is neglecting 2nd variable altogether this low value prediction is expected.

We might have to inspect correlation or use some effective tools that will suggest us to neglect variable on our second step. So that, we can include variables which does not have 0 weight in generated model and increase the accuracy of predicted value.

According to my chosen variables and best condition for low energy usage is low outside temperature and humidity. The temperature inside the kitchen has less effect on efficiency whereas humidity inside the kitchen does not have any effect on energy usage at all.

References

CRAN - Package e1071. 2019. CRAN - Package e1071. [ONLINE] Available at: https://cran.r-project.org/web/packages/e1071/index.html. [Accessed 24 April 2019].

Corey Wade. 2019. Transforming Skewed Data – Towards Data Science. [ONLINE] Available at: https://towardsdatascience.com/transforming-skewed-data-73da4c2d0d16. [Accessed 03 May 2019].

FutureLearn. 2019. Page from SIT718.3 Data Analysis with R - Deakin University. [ONLINE] Available at: https://www.futurelearn.com/courses/sit718-3/2/steps/489831. [Accessed 03 May 2019].

FutureLearn. 2019. Page from SIT718.3 Data Analysis with R - Deakin University. [ONLINE] Available at: https://www.futurelearn.com/courses/sit718-3/2/steps/489823. [Accessed 03 May 2019].