SIT718 Real World Analytics

Assessment 2

Yash Ahuja

StudentID: 219608443

1 (iv) Understanding the distribution of all variables and their relationship with the variable of interest

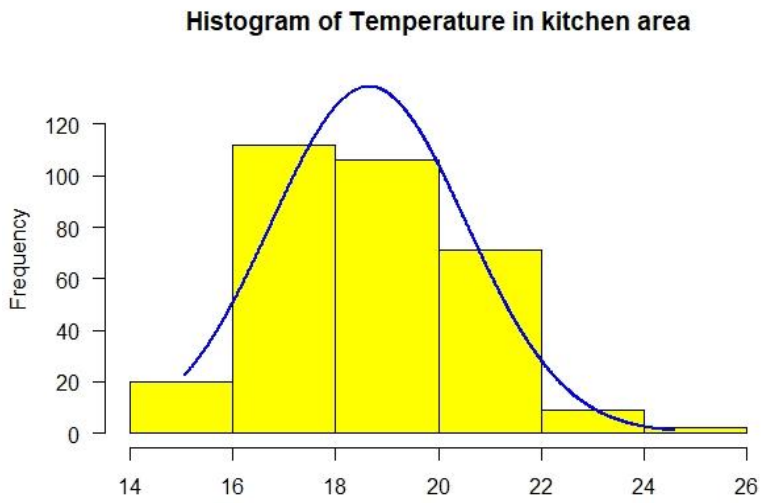


Figure 1

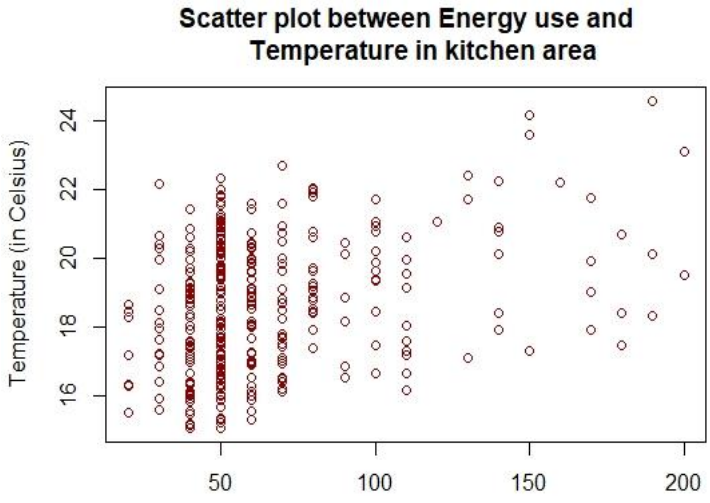


Figure 2

The normal histogram plot of the temperature in kitchen area variable indicates that the distribution of the variable is approximately normal as it is symmetric (see figure 1). Moreover, further K-S test (as sample size >50) was performed on the variable and the p-value of the test was greater than 0.05 which concluded that the distribution of the variable is approximately normal. Also, the scatter plot between energy use and temperature in kitchen area shows a positive relationship between the two variables (see figure 2).

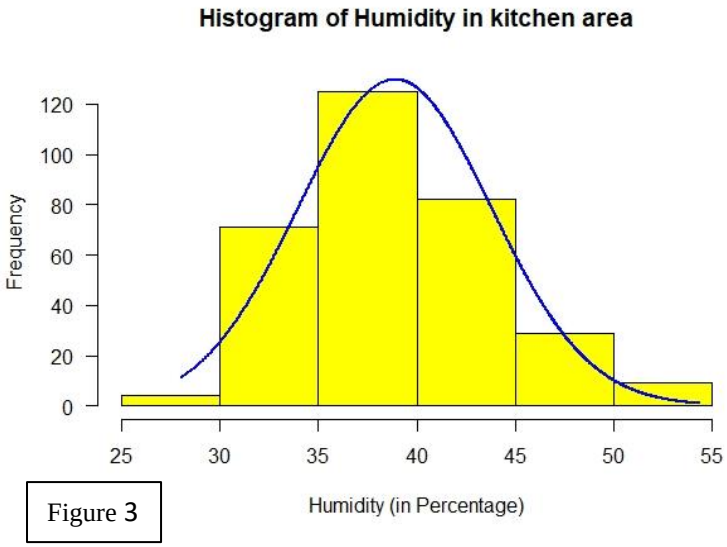


Figure 3

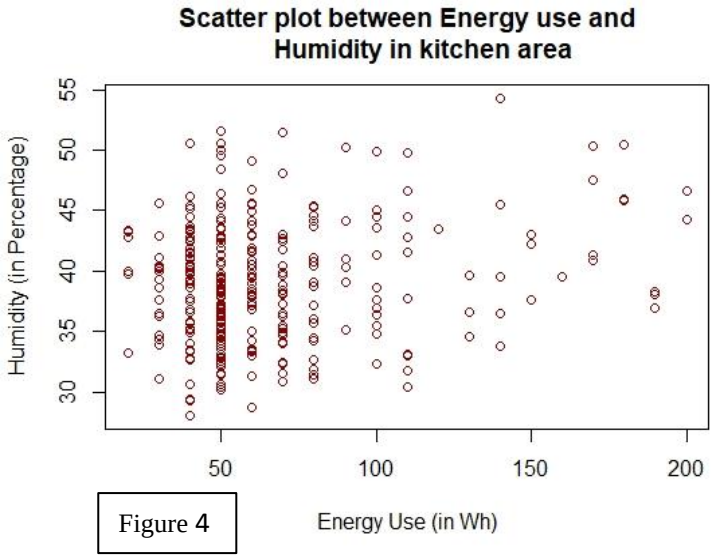


Figure 4

The normal histogram plot of the humidity in kitchen area variable indicates that the distribution of the variable is approximately normal as it is symmetric (see figure 3). Moreover, further K-S test was performed on the variable and the p-value of the test was greater than 0.05 which concluded that the distribution of the variable is

approximately normal. Also, the scatter plot between energy use and humidity in kitchen area shows a positive relationship between the two variables (see figure 4).

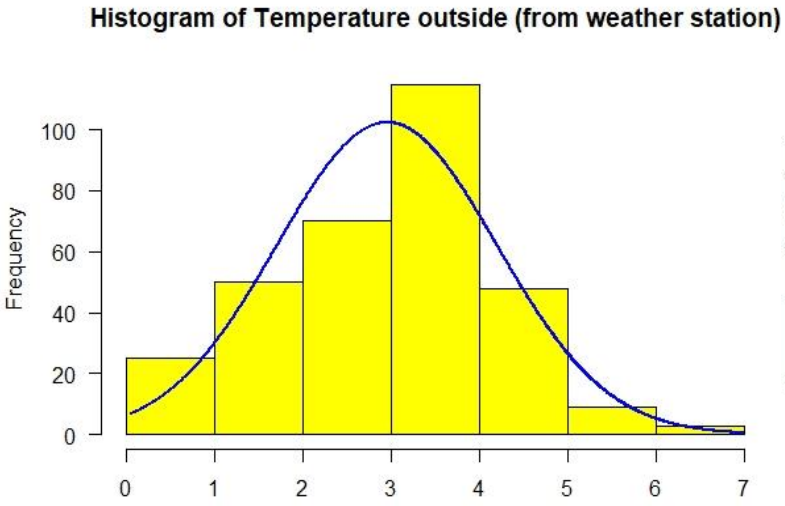


Figure 5

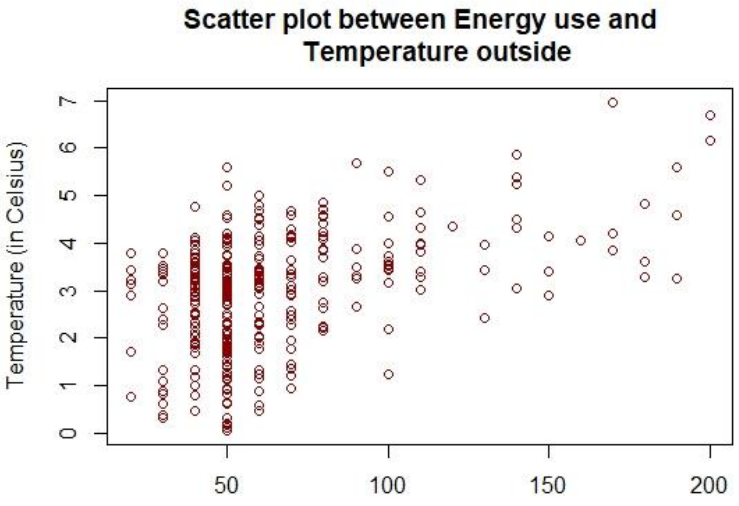


Figure 6

The normal histogram plot of the temperature outside variable indicates that the distribution of the variable is approximately normal as it is symmetric (see figure 5). Moreover, further K-S test was performed on the variable and the p-value of the test was greater than 0.05 which concluded that the distribution of the variable is approximately normal. Also, the scatter plot between energy use and temperature outside shows a positive relationship between the two variables (see figure 6).

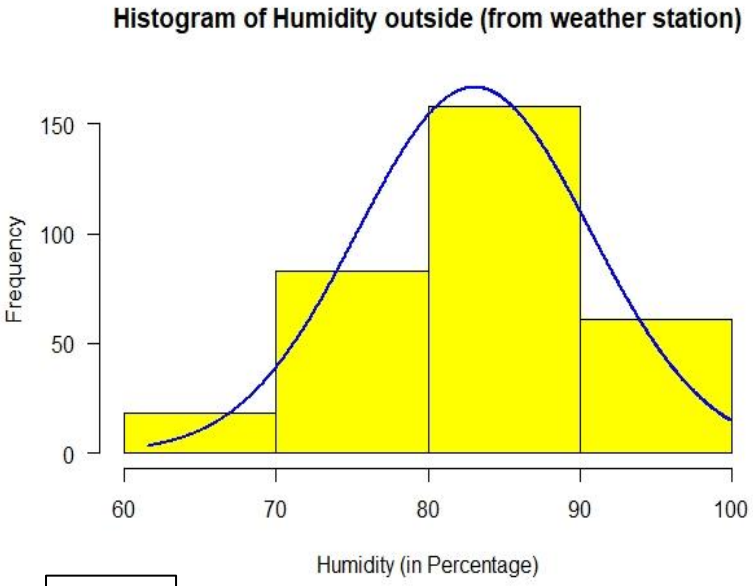


Figure 7

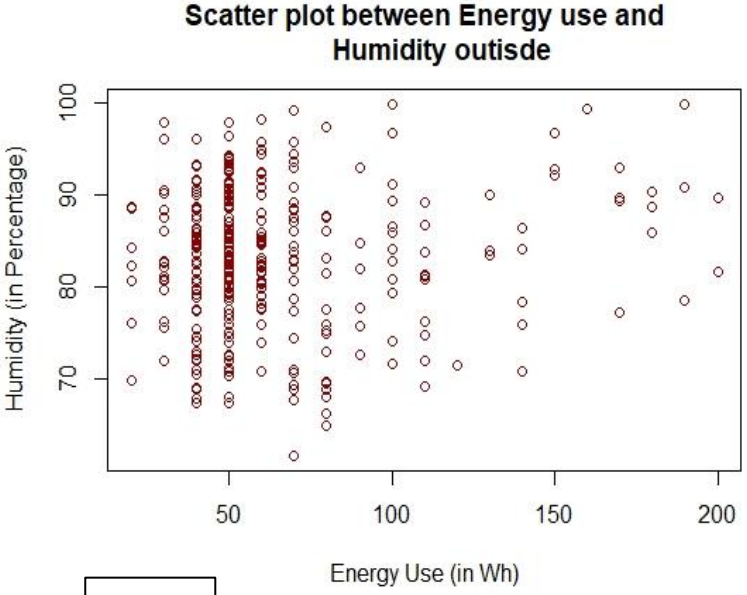


Figure 8

The normal histogram plot of the humidity outside variable indicates that the distribution of the variable is approximately normal as it is symmetric (see figure 7). Moreover, further K-S test was performed on the variable and the p-value of the test was greater than 0.05 which concluded that the distribution of the variable is

approximately normal. Also, the scatter plot between energy use and humidity outside shows a positive relationship between the two variables (see figure 8).

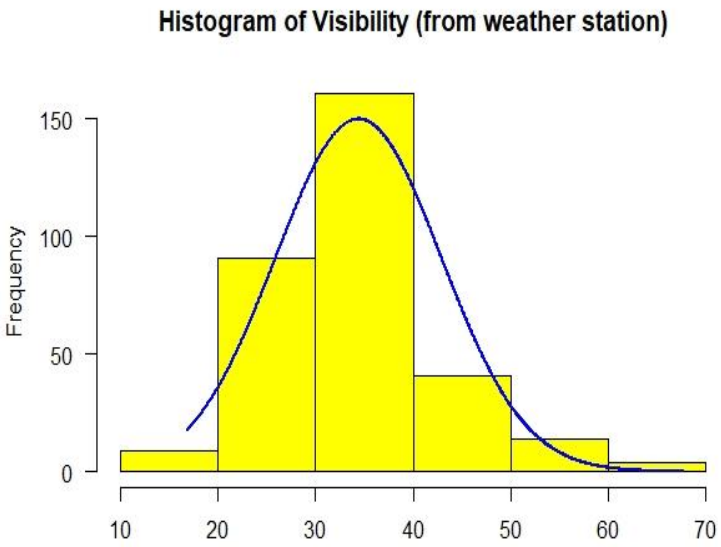


Figure 9

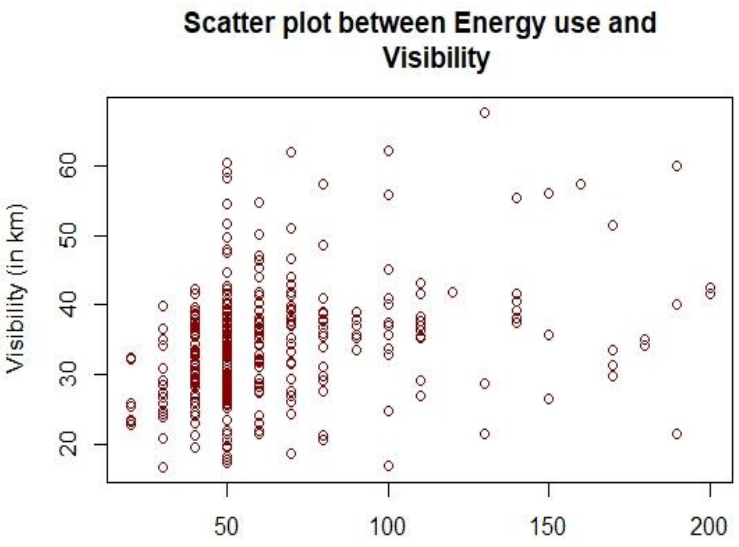


Figure 10

The normal histogram plot of the visibility variable indicates that the distribution of the variable is approximately normal as it is symmetric (see figure 9). However, further K-S test was performed on the variable and the p-value of the test was less than 0.05 which concluded that the distribution of the variable is not approximately normal. The skewness calculated by the skewness function was greater than 0.5 and indicated that it has moderately skewed (positive) distribution. Also, the scatter plot between energy use and temperature outside shows a positive relationship between the two variables (see figure 10).

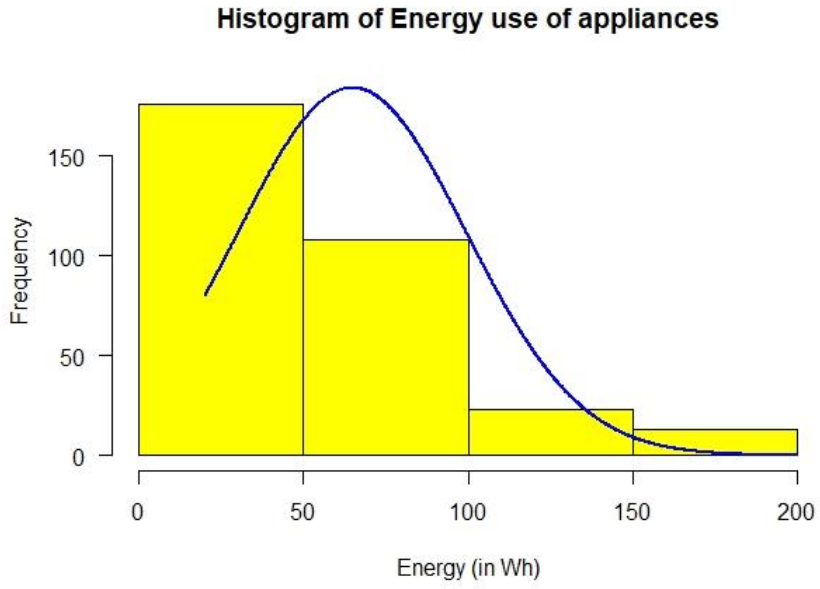


Figure 11

The normal histogram plot of the energy use variable indicates that the distribution of the variable is not approximately normal as it is not symmetric (see figure 11). Moreover, further K-S test was performed on variable and the p-value of the test was less than 0.05 which concluded that the distribution of the variable is not normal.

The skewness calculated by the skewness function was greater than 1 and indicated that it has highly skewed (positive) distribution.

2 (ii) Selecting the four variables and applying the necessary transformations on the selected four variables and the variable of interest

After performing the pearson test for analysing the strength of the relationship between each of the variables and the variable of interest, it was found that the variable X4 i.e. humidity outside (from weather station) has the weakest relationship with variable of interest i.e. energy consumption out of all the five independent variables based on pearson correlation coefficient (see figure 13). Therefore, the selected variables for transformation are temperature in kitchen area, humidity in kitchen area, temperature outside (from weather station), visibility and the variable of interest i.e. energy use of appliances (X1, X2, X3, X5 and Y).

Results: K-S test for Normality	
Variables	**p-value**
Temperature in kitchen area	0.3659
Humidity in kitchen area	0.261
Temperature outside (from weather station)	0.06447
Humidity outside (from weather station)	0.6452
Visibility (from weather station)	0.04292
Energy use of appliances	~0.00

Figure 12

Pearson Test Results	
Variables	**Pearson Correlation Coefficient (with Energy use variable)**
Temperature in kitchen area	0.29786526
Humidity in kitchen area	0.14526831
Temperature outside (from weather station)	0.43124855
Humidity outside (from weather station)	0.05751533
Visibility (from weather station)	0.34086755

Figure 13

Transformation on variables namely temperature in kitchen area (X1), humidity in kitchen area (X2) and temperature outside (X3)

Since, all the three variables namely temperature in kitchen area, humidity in kitchen area, temperature outside have approximately normal distribution as per the normal histogram plots and the K-S test results above (figure 12), there is no need of log or polynomial transformation as the distribution of these variables is not skewed (p-value > 0.05). Moreover, pearson test results indicated that all the three variables have positive relationship with the variable of interest. Hence, there is no need of negation transformation on these variables. Thus, only Min-max transformation is applied on these three variables to ensure that these variables taking values over different ranges can be transformed to the same unit interval [0,1] in order to be aggregated.

Min-max transformation formula: $x(new) = x - \min(x) \ / \max(x) - \min(x)$

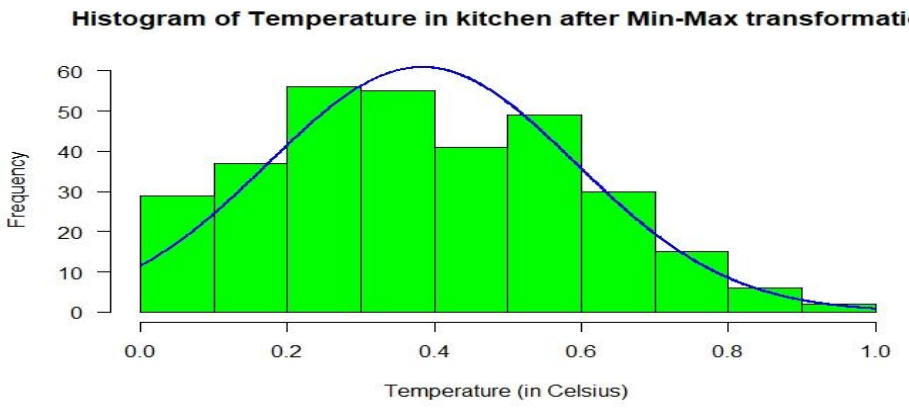


Figure 14

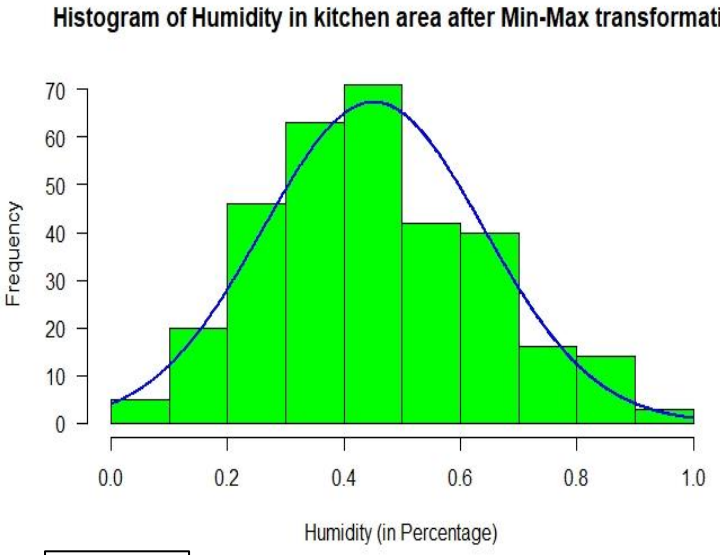


Figure 15

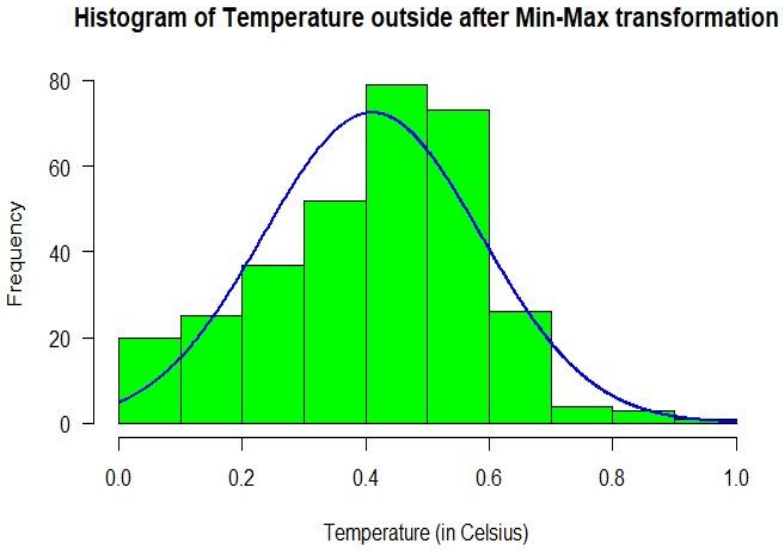


Figure 16

Transformation on variables namely visibility (X5) and energy use of appliances (Y)

Since, the distribution of the variables visibility and energy use is not approximately normal distribution as the p-values for their K-S test are less than 0.5, there is need of log or polynomial transformation (figure 12). After calculating their skewness through skewness function, it was found that the skewness of both the variables is greater than 0.5. Hence, both the variables are positively skewed, therefore, we will apply log transformation on both the variables first by taking log with base 10 and then Min-max transformation to these log transformed variables taking values over different ranges to ensure that all the variables are transformed to the same unit interval [0,1] in order to be aggregated.

1) **Log transformation formula:** $\boldsymbol{x(new) = \log 10(x)}$
2) **Final transformation:** $\boldsymbol{x(final) = x(new) - min(x\ (new))\ /\ max(x\ (new)) - min(x(new))}$

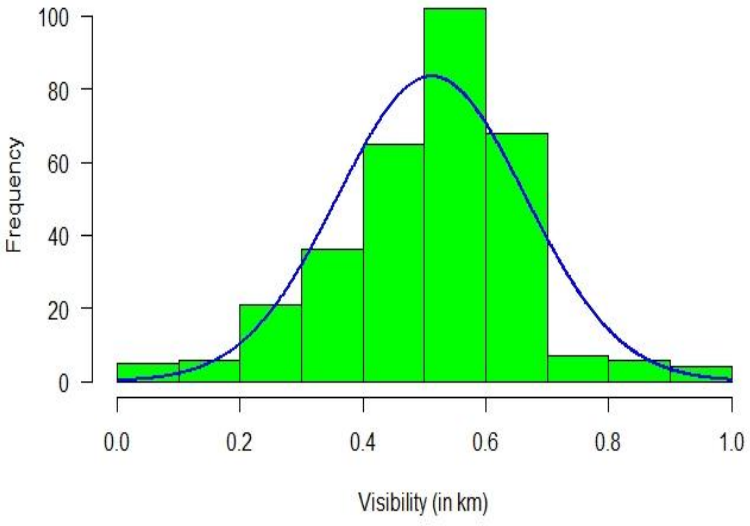


Figure 17

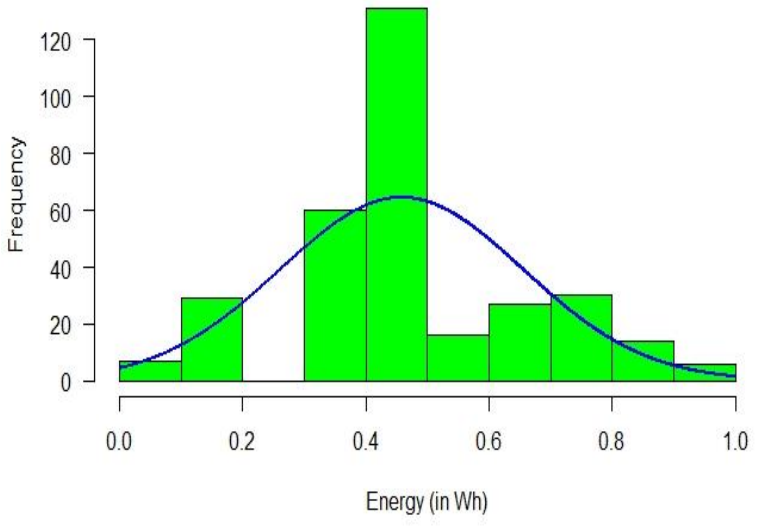


Figure 18

3 (iii) Error measures, correlation coefficients, weights/parameters for all the four fitting models used

Error Measures and Correlation Coefficients	Models				
	WAM	WPM (p=0.1)	WPM (p=10)	OWA	Choquet Integral
RMSE (Root Mean Square Error)	0.17126	0.181596069	0.179386462	0.178582	0.168188422
Average Absolute Error	0.13509	0.143506723	0.144321737	0.140434	0.132730637
Pearson Correlation	0.52013	0.46851221	0.500663385	0.453145	0.533453681
Spearman Correlation	0.50861	0.476145636	0.47069182	0.433734	0.5092572

Figure 19

Weights and Parameters	Models				
	WAM	WPM (p=0.1)	WPM (p=10)	OWA	Choquet Integral
Weight/Shapley (for Choquet) (w1)	0.25664	0.197635819	0.326392797	0.332539	0.280849046
Weight/Shapley (for Choquet) (w2)	0	0.101382016	0	0	0.031658159
Weight/Shapley (for Choquet) (w3)	0.4064	0.191240418	0.61183356	0.411059	0.448654864
Weight/Shapley (for Choquet) (w4)	0.33696	0.509741746	0.061773643	0.256402	0.23883793
Orness (for OWA & Choquet)	N/A	N/A	N/A	0.530441	0.578517053

(iv) Figure 20

a) Based on the figure 19, we can conclude that Choquet Integral model has the lowest values of Root Mean Square Error and Average Absolute Error. It means that the average difference between each predicted value and observed value is minimum for Choquet Model. Moreover, Pearson and Spearman correlation coefficients are highest for the choquet integral model. These coefficients give an idea of the strength of the relationship between the two variables. Higher values indicate strong relationship between the dependent and the independent variable and it is evident that the choquet model has highest values of Pearson and Spearman coefficients. Thus, the choquet integral model is performing well.

b) Based on the figure 13, it is evident that the third variable X3 i.e. temperature outside has the strongest relationship with the variable energy use as its Pearson Correlation coefficient was highest out of all the four selected variables. Therefore, more importance or in other words, more weight must be given to the third variable for better performance of the model and more accurate prediction of the values. It can be seen from figure 20 that all the models except Weighted Power Mean (p=0.1) model have given highest weight to the third variable while predicting the output. It may be the reason for the WPM (p=0.1) model to have the highest RMSE value (lowest performance) as compared to all the other models.

c) Choquet Integral Weight Results:

$$v(\{1,2,3,4\}) = 1,$$
$$v(\{1,2,3\}) = 0.7639,\quad v(\{1,2,4\}) = 0.5912,\ v(\{1,3,4\}) = 1,$$
$$v(\{2,3,4\}) = 0.6717,$$
$$v(\{1,2\}) = 0.5912,\ v(\{1,3\}) = 0.7639,\ v(\{1,4\}) = 0.5912,$$
$$v(\{2,3\}) = 0.6717,\ v(\{2,4\}) = 0.5033,\ v(\{3,4\}) = 0.6717,$$
$$v(\{1\}) = 0.3291,\ v(\{2\}) = 0,\ v(\{3\}) = 0.6717,\ v(\{4\}) = 0.3855,\ v(\{\phi\}) = 1$$

Interactions:

- For variables 1 and 2: $v(\{1,2\}) > v(\{1\}) + v(\{2\})$
  Therefore, these variables interact in a complementary way.
- For variables 1 and 3: $v(\{1,3\}) < v(\{1\}) + v(\{3\})$
  Therefore, these variables interact in a redundant way.

- For variables 1 and 4: v ({1,4}) < v ({1}) + v ({4})
  Therefore, these variables interact in a redundant way.
- For variables 2 and 3: v ({2,3}) = v ({2}) + v ({3})
  Therefore, these variables have no interaction.
- For variables 2 and 4: v ({2,4}) > v ({2}) + v ({4})
  Therefore, these variables interact in a complementary way.
- For variables 3 and 4: v ({3,4}) < v ({3}) + v ({4})
  Therefore, these variables interact in a redundant way.

d) Based on the figure 19, it is evident that Choquet Integral is the best model out of all the four models, and it has an orness measure approximately equal to 0.58. Orness is a concept that applies to all the averaging functions which gives an indication of how much the averaging function favours high inputs. In other words, it implies that how similar an aggregation function is to the maximum function which produces high outputs for high inputs. Since, orness of Choquet Integral model is greater than 0.5, we can conclude that it favours higher inputs.

4 (i) Choosing the best fitting model and predicting the energy use

The best fitting model is Choquet Integral as it has the best performance because of lowest values of Root Mean Square Error and Average Absolute Error measures (see figure 19). Moreover, it has highest values of Pearson and Spearman Correlation coefficients (see figure 19) attributing to its best performance out of all the four models.

Now using the best fitting model i.e. Choquet Integral model to predict the energy use for the new data. On applying the model, the predicted energy use for the given input is Y = 51.21951. Thus, the predicted energy use for the given input is approximately 51.21 Wh.

(ii) While, training the model, we concluded that the third variable X3 has the highest importance out of all the other variables (see figure 13). Since, the range of the third variable X3 i.e. temperature outside in the training dataset was [0,7], we can say that the predicted energy use of 51.21 Wh is reasonable. If the input X3 was higher, then the predicted energy use would have been higher than 51.21 Wh. Thus, both these factors of having lowest values of error measures and highest values of correlation coefficients are contributing to the strong predictive power of the Choquet Integral model in order to make a reasonable prediction of the energy use.

(iii) Based on the figure 20, we can conclude that the variable X2 has least average importance out of all the four selected variables (X1,X2,X3,X5) and X3 has the highest average importance while X1,X5 have moderate average importance as per their shapley values. Therefore, in order to get a low energy use of appliances, the input should be lowest value of X3 and lower values of X1 and X5. Thus, a low temperature outside of value (less than 4 degree in Celsius), a low temperature in kitchen area of value (less than 16 degree in Celsius), a low value of visibility (less than 31km) and approximately same humidity in kitchen area (38%) will yield a low energy use of appliances (lower than 51.21 Wh).

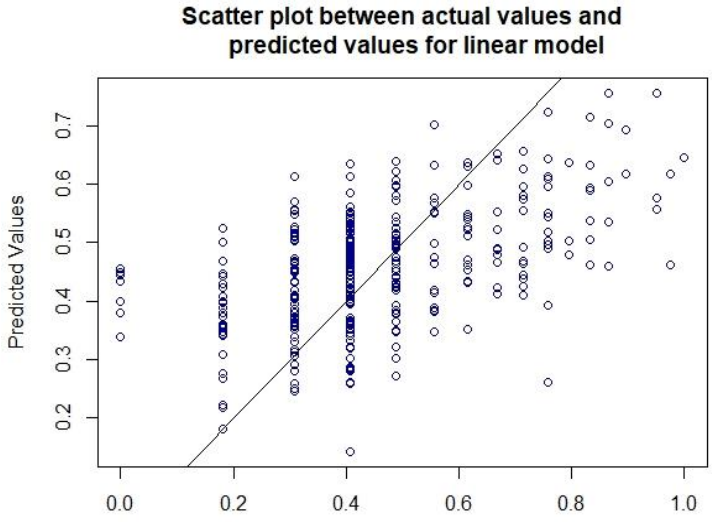
5 (ii) Comparing with linear model

Comparison between Linear Model and Choquet Integral		
Models	**Performance Measures**	**Values**
Choquet Integral	RMSE	0.168188422
Linear Model	Residual Standard Error	0.1692

Figure 21

Performance Measures of Linear Model	
Residual Standard Error	0.1692
R-Square (Multiple Regression)	0.2808
Adjusted R-square (Multiple Regression)	0.2716
p-value	~0.00

Figure 22



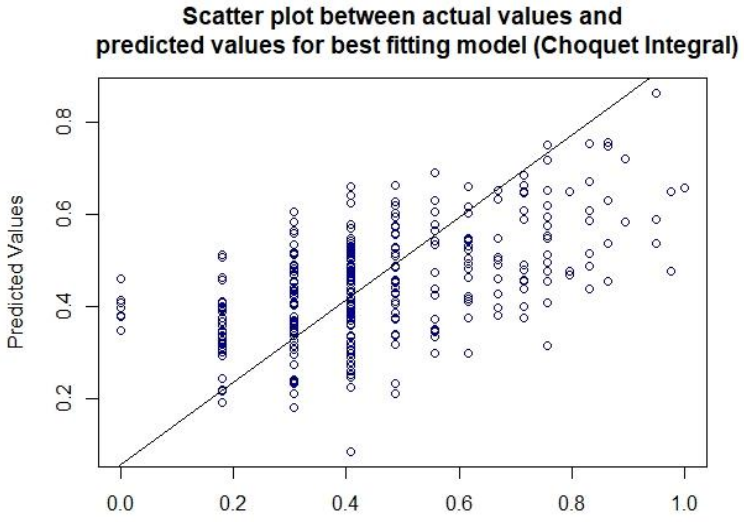


Figure 23

(iii) Based on figure 22, the residual standard error of the linear model is very small indicating that model fits the data well. However, only 27% (approximately) of variance in Y can be explained by all the independent variables, considering the number of independent variables used. This Adjusted R-square measure is widely used for evaluating model performance and is one of the advantages of linear model over choquet model. Moreover, the p-value for model is less than 0.05 implicating that the model is statistically significant and has some predictive power. However, linear model also has various assumptions including linearity, independence of errors, normality of errors and homoscedasticity. On comparing the RMSE of choquet integral model which is almost the same as residual standard error of linear model (instead of dividing by the sample size n, here we divide by n-1 degrees of freedom to get an unbiased estimation of the standard deviation of the residuals), it is evident that both the measures are approximately equal (figure 21) and the relationship between the predicted values of both the models is strong and positive (figure 24) implicating that both the models have similar performance. But, the performance measures for linear model tend to give an unbiased estimation of the model performance as compared to choquet model and thus the choquet model has certain limitations (George Choueiry n.d.). Moreover, the scatter plot between predicted values and actual values for both the models indicate that the relationship between predicted values and actual values is stronger for linear model as compared to the choquet integral model as the slope for regression line for linear model is higher than the choquet model (see figure 23). Therefore, it can be concluded that the linear model is performing better than the choquet model.

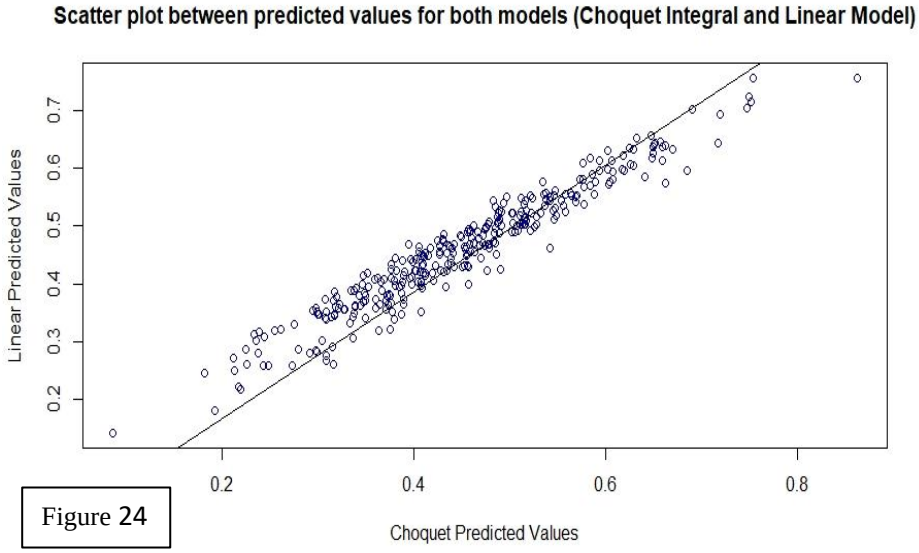


Figure 24

References:

- George Choueiry n.d., *Residual Standard Deviation/Error: Guide for Beginners,* Quantifying Health, retrieved 05 September 2020, <https://quantifyinghealth.com/residual-standard-deviation-error/>.