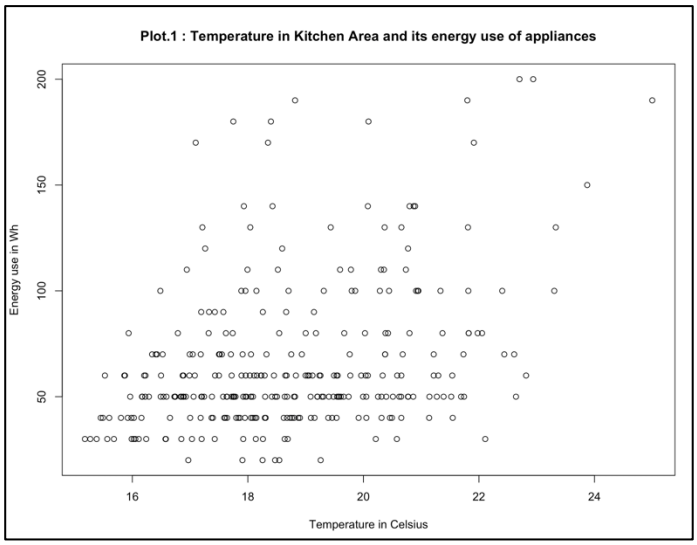
## SIT718 Real World Analytics

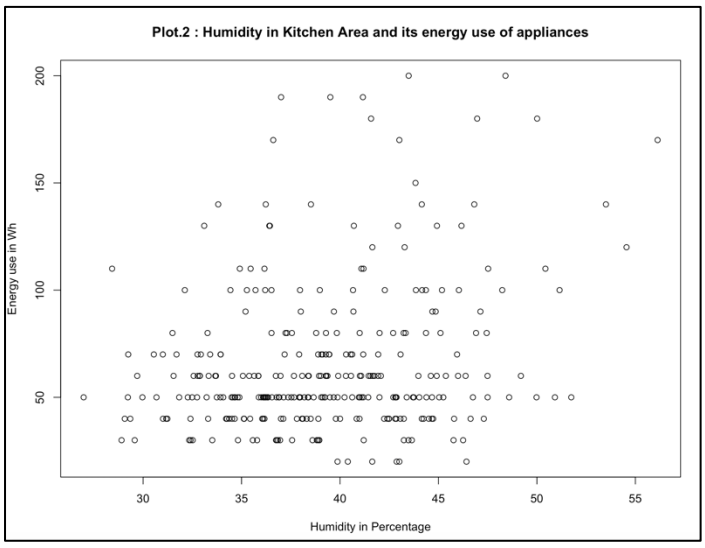
## Assessment Task 3: Problem Solving

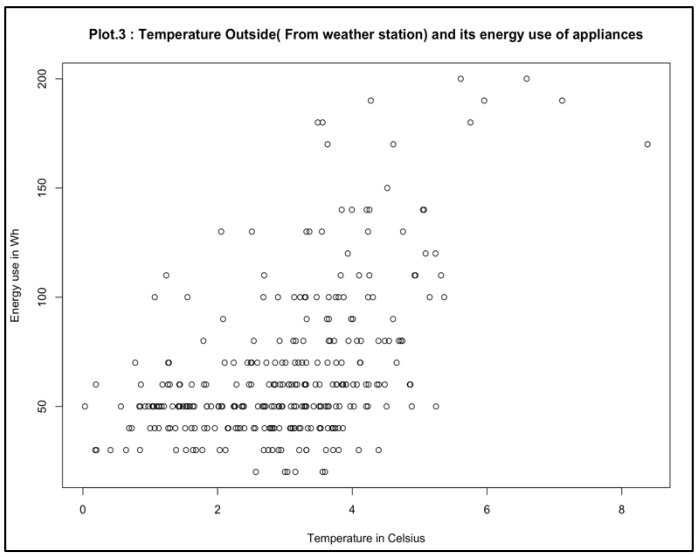
**Q.1)** Understand the data

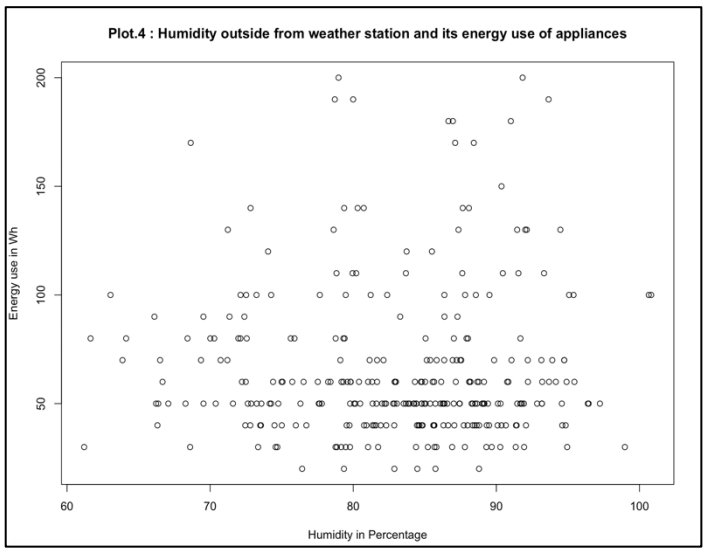
**(iv)**

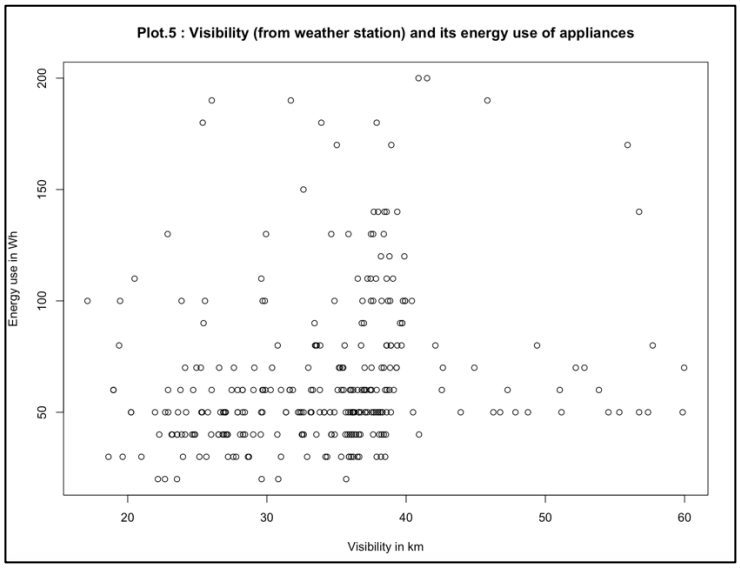
# Scatter Plots



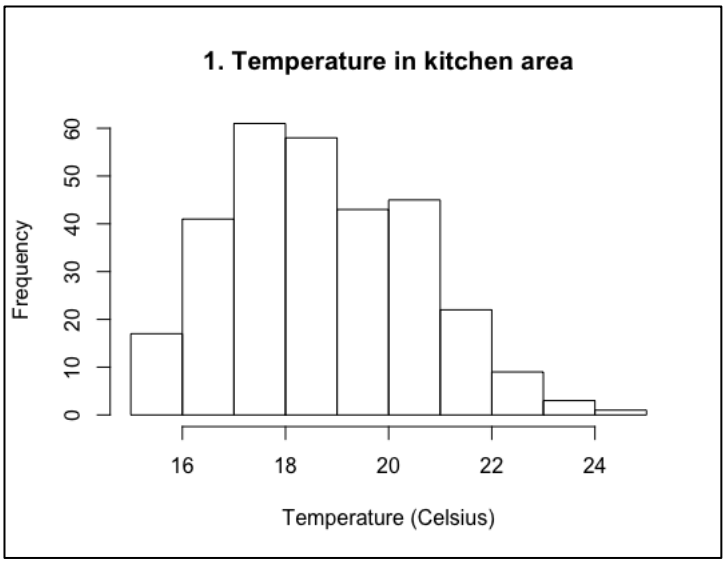



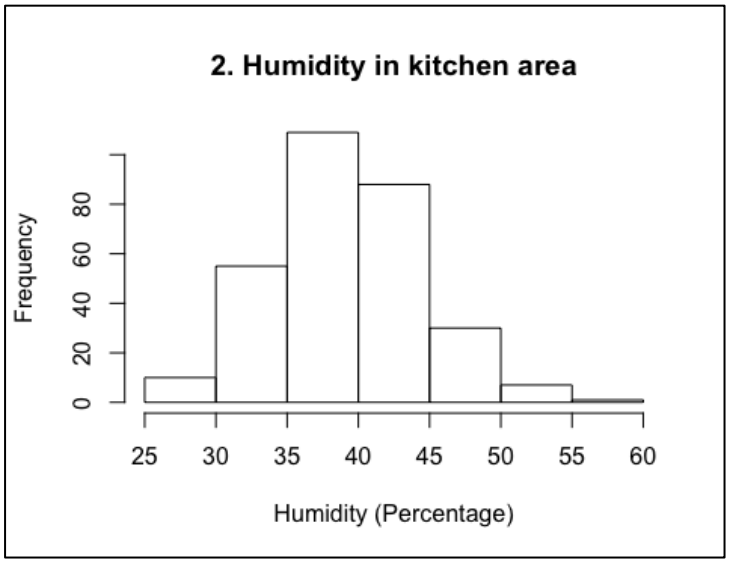



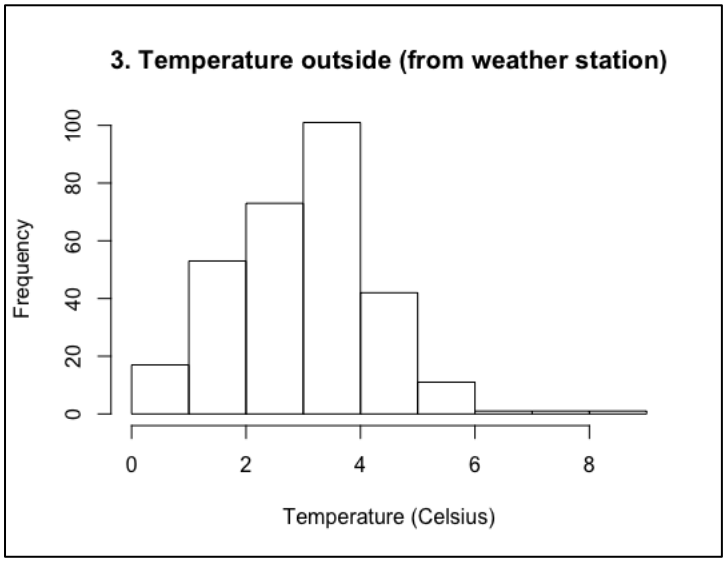



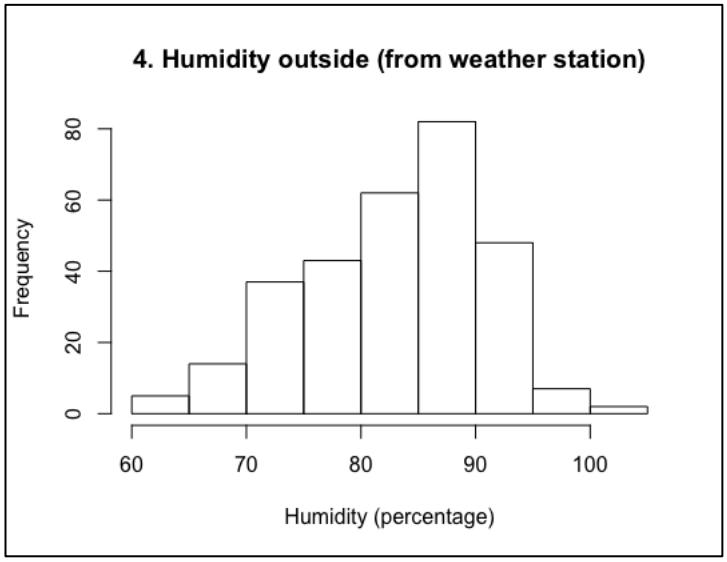

Plot.5 : Visibility (from weather station) and its energy use of appliances
Energy use in Wh
50
100
150
200
20
30
40
50
60
Visibility in km

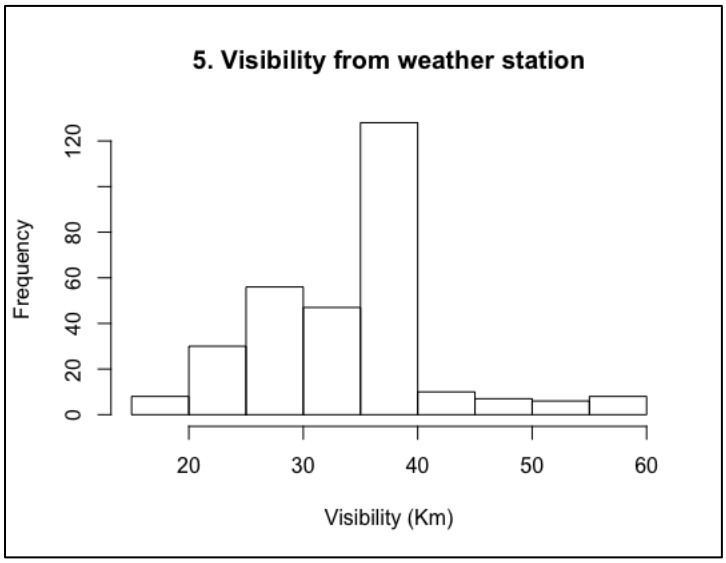
## Histograms

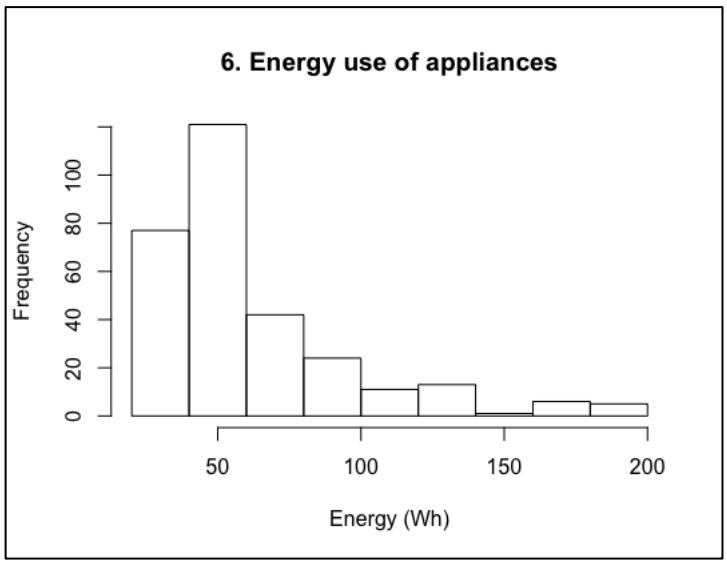

1. Temperature in kitchen area
Frequency
0
10
20
30
40
50
60
16
18
20
22
24
Temperature (Celsius)


2. Humidity in kitchen area
Frequency
0
20
40
60
80
25
30
35
40
45
50
55
60
Humidity (Percentage)


3. Temperature outside (from weather station)
Frequency
0
20
40
60
80
100
0
2
4
6
8
Temperature (Celsius)


4. Humidity outside (from weather station)
Frequency
0
20
40
60
80
60
70
80
90
100
Humidity (percentage)




**X1:** Temperature in the kitchen area was mostly observed between 16-21° C. The most frequent temperature in the kitchen was 18°C. The energy wasn't being used that much in the kitchen.

**X2:** Humidity in the kitchen area usually revolved around 30 – 50%. Mostly it was humid between 35-40%, since the energy use of its appliances wasn't that much.

**X3:** Temperature outside from the weather station can be observed to be less that the temperature observed inside the kitchen. The temperature revolved around 0 – 8 °C, being 3°C as the most frequent temperature observed outside. It was cold outside.

**X4:** It was a lot humid outside as observed by the weather station. The humidity was between 60 – 100%, and mostly it used to be between 85-90%. Weather outside is quite humid if compared to the weather inside kitchen area.

**X5:** Visibility from the weather station can be seen till 40 km max, as its observed that after 40km the visibility tends to decrease.

**Y:** Appliances aren't being used that much as their energy used is observed to be fluctuating between 0 to 100 Wh and then very less energy is used from 100-200 Wh.

## Q.2) Transform the Data

## (ii)

**X1:** Linear feature Scaling is used to scale the data between 0 to 1.

**X2:** Standardization is used as the data follows a normal distribution if seen in the histogram and then linear feature scaling transformations used to scale the data between 0 to 1.

**X3:** Linear feature scaling is used to scale the data between 0 to 1 as the data is right skewed.

**X4:** Log transformation is used and then linear feature scaling is used to scale the data between 0 to 1, so that the lower values are spread and the outliers are brought close to the group.

**Y:** Log transformation is used and then linear feature scaling is used since the variable of interest is right skewed and has bigger values close to 0, and to bring the outliers close to the group.

## Q.3) Build Models and investigate the importance of each variable

## iii)

## Table 1:

| **FITTING MODELS** | **ERROR MEASURES** | | **CORRELATIONS** | |
|---|---|---|---|---|
| Weighted Average Mean (WAM) | **RMSE:** 0.170985216122321 (17%) | **Av. Abs error:** 0.13312962200183 (13%) | **Pearson:** 0.602267039592576 | **Spearman:** 0.528749500577239 |
| Weighted Power Mean (WPM) p = 0.5 | **RMSE:** 0.176170019678336 (17%) | **Av. Abs error:** 0.137205966152505 (13%) | **Pearson:** 0.564396949506935 | **Spearman:** 0.489040968029164 |
| Weighted Power Mean (WPM) p = 2 | **RMSE:** 0.16695488164311 (16%) | **Av. Abs error:** 0.129133679266691 (12%) | **Pearson:** 0.623836868876823 | **Spearman:** 0.567166175600223 |
| Ordered Weighted Average (OWA) | **RMSE:** 0.176800398940278 (17%) | **Av. Abs error:** 0.129133679266691 (12%) | **Pearson:** 0.479553947199564 | **Spearman:** 0.479553947199564 |
| Choquet Integral | **RMSE:** 0.159797569654747 (15%) | **Av. Abs error:** 0.123371058231135 (12%) | **Pearson:** 0.627245228810622 | **Spearman:** 0.585523286544981 |

## Table 2:

| Fitting Models | Weights | Orness |
|---|---|---|
| Weighted Average Mean (WAM) | **1)** 0.253899578472887<br>**2)** 0<br>**3)** 0.476613729999503<br>**4)** 0.269486691527604 | |
| Weighted Power Mean (WPM) p = 0.5 | **1)** 0.249280863907075<br>**2)** 0<br>**3)** 0.413573003366405<br>**4)** 0.337146132726521 | |
| Weighted Power Mean (WPM) p = 2 | **1)** 0.219813515040448<br>**2)** 0<br>**3)** 0.612976074855674<br>**4)** 0.167210410103869 | |
| Ordered Weighted Average (OWA) | **1)** 0.0389640277014983<br>**2)** 0.468979857254385<br>**3)** 0.23863548545625<br>**4)** 0.253420629587866 | 0.568837572310162 |
| Choquet Integral | **Shapley**<br>**1)** 0.264914336731296<br>**2)** 0.0699335779312172<br>**3)** 0.595218507407297<br>**4)** 0.0699335779309052<br><br>**binary number fm.weights**<br>**1)** 0.556001867405153<br>**2)** 0.279734311723482<br>**3)** 0.556001867406037<br>**4)** 0.886306038081674<br>**5)** 0.999999999999959<br>**6)** 0.886306038082095<br>**7)** 1.00000000000122<br>**8)** 0.279734311723482<br>**9)** 0.556001867406019<br>**10)** 0.279734311723486<br>**11)** 0.556001867406507<br>**12)** 0.886306038081649<br>**13)** 1.00000000000051<br>**14)** 0.886306038082147<br>**15)** 1.00000000000072 | 0.685026487463083 |

## iv)

The choquet fitting function performed well as the lowest RMSE can be seen if compared to other fitting functions so that's the best fitting function. This means that the average difference between the prediction and output wasn't much. OWA, WPM,WAM gave a higher number of RMSE which means they have a higher margin of error. Moreover, seeing the pearson correlations as a goodness of fit measure, the value in choquet is more similar to the pearson of WAM and WPM. The value is close to 1, which means the relationship is linear. Choquet function gave the pearson and spearman closest to 1 if compared by other fitting functions.

| Decimal | Binary | Variable |
|---|---|---|
| 1 | 1 | {x1} = 0.56 |
| 2 | 10 | {x2} =0.28 |
| 3 | 11 | {x1,x2} =0.56 |
| 4 | 100 | {x3} =0.89 |
| 5 | 101 | {x1,x3} =0.99 |
| 6 | 110 | {x2,x3} =0.89 |
| 7 | 111 | {x1,x2,x3} = 1 |
| 8 | 1000 | {x4} =0.28 |
| 9 | 1001 | {x1,x4} =0.56 |
| 10 | 1010 | {x2,x4} =0.28 |
| 11 | 1011 | {x1,x2,x4} =0.56 |
| 12 | 1100 | {x3,x4} = 0.89 |
| 13 | 1101 | {x1,x3,x4} = 1 |
| 14 | 1110 | {x2,x3,x4} =0.88 |
| 15 | 1111 | {x1,x2,x3,x4}. =1 |

**Table 3: Fuzzy Measures Weights**

The variables I selected were temperature in kitchen area and outside from weather station, and humidity inside the kitchen area and outside from weather station. From the weights observed by the weighted power mean we got to know that the temperature inside the kitchen was by far the important variable in considering the prediction values, followed by humidity from weather station and then humidity from kitchen and then temperature from weather station. The temperature would be the best predictor given all the variables give a positive relation. After seeing the orness of OWA and choquet, the weights close to 0 can be seen in Choquet. This could be taken as an indication of over predicting the outputs, probably because of the skewed distribution of the outputs if compared to the input results of {X1,X2,X3,X4,Y}.

The most important point proved from the table 3 is that the temperature outside from the weather stations has the highest weight and the humidity inside (v({2})) and humidity outside from weather station (v({4})) proved to have the lowest weight. Whereas, if the lowest inputs are combined together with other inputs, they tend to have a subset weight higher than 0.5 and they are combined with the same input they give the same weight. Temperature from inside the kitchen area does not making any significant changes to the weight as the increase is mostly seen by the other input combining them. Thus, to conclude the fuzzy measures we need suitable data for temperature inside kitchen to have an appropriate prediction of the energy appliances, making the variables redundant.

As seen in OWA weights, the weight assigned to humidity inside in the kitchen area was the highest, as compared to the weight given to other. Whereas, in Choquet the weight assigned to temperature

outside from weather station was highest which was 0.595, which indicates that more than 50% of the weight was allocated to this input. The orness value was close to 0.7 for both the OWA and choquet concludes the tendency of a better model towards higher input. This also could account for that the high orness would need less factors and less high inputs to have an overall high output, as observed in Choquet for e.g the temperature inside the kitchen could correspond to the energy use of appliances, even if the humidity inside and outside was low.

## Q.4)

ii)

Choquet Integral.
Predicted Value of Energy use of appliances: **15.5**
Thus, the prediction according to me is not reasonable as the fitting model didn't produce the output closer to original values.

iii)

| | | | | |
|---|---|---|---|---|
| X1 = 18 | X2 = 44 | X3 = 4 | X4 = 74.8 | **(Predicted) Y = 15.5** |
| X1 = 16.470 | X2 = 41.247 | X3 = 2.14260 | X4 = 85.979 | **Y = 20** |

**Table 4 comparing the inputs**

Therefore, using this fitting model if the temperature inside kitchen is lower, the humidity inside the kitchen is lower, the temperature outside from weather station is lower, and the humidity outside from the weather station is higher only then it's probable to see the low Energy use of appliances. Therefore, to conclude the best fitting function I chose wasn't desirable.