1. Understand the data:

1.4 Histograms and Scatterplots:

1.4.1 Kitchen temperature (X1): Data is slightly positively skewed with skewness coefficient of 0.45. Almost 46% of the data lies towards the left of the mean (between 15° C and 18° C) and 54% of the data lies towards the right of the mean (between 18° C and 25° C). (Refer Fig-1 and Table-2).

1.4.2 Kitchen humidity (X2): Data is approximately symmetrical with skewness coefficient of 0.18. Almost 52% of the data lies towards the left of the mean (between 26 and 39) and 48% of the data lies towards right of the data (between 39 and 54). (Refer Fig-2 and Table-2).

1.4.3 Outside temperature (X3): Data is approximately symmetrical with skewness coefficient of -0.14. Almost 49% of the data lies towards the left of the mean (between -0.6° C and 2.96° C) and 51% of the data lies towards right of the mean (between 2.96° C and 7.5° C). (Refer Fig-3 and Table-2).

1.4.4 Outside humidity (X4): Data is slightly negatively skewed with skewness coefficient of -0.31. Almost 47% of the data lies towards the left of the mean (between 61 and 83) and 53% of the lies towards right of the mean (between 83 and 101). (Refer Fig-4 and Table-2).

1.4.5 Visibility (X5): Data is slightly positively skewed with skewness coefficient of 0.43. Almost 43% of the data lies towards the left of the mean (between 14 and 34) and 57% of the data lies towards right of the data (between 34 and 60). (Refer Fig-5 and Table-2).

1.4.6 Energy use (Y): Data is highly positively skewed with skewness coefficient of 1.76. Almost 89% of the data lies between (20 and 100) and 11% of the data lies between (100 to 200). (Refer Fig-6 and Table-2).

Fig-1

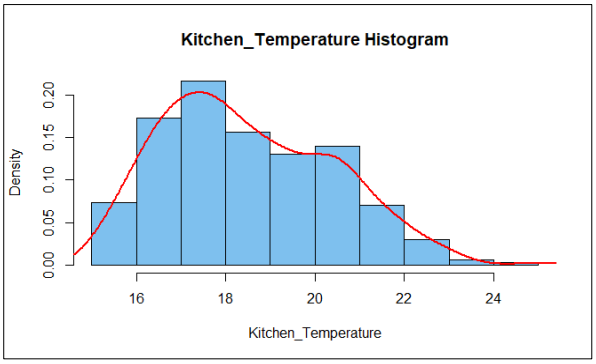


Fig-2

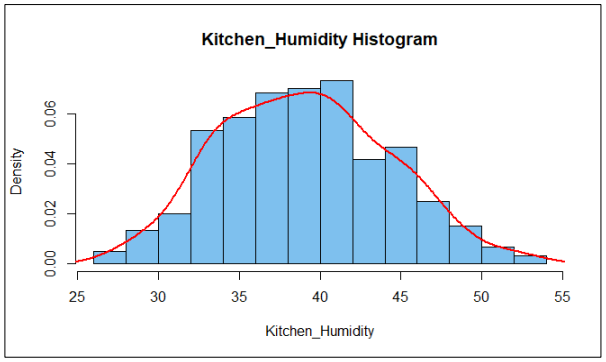


Fig-3

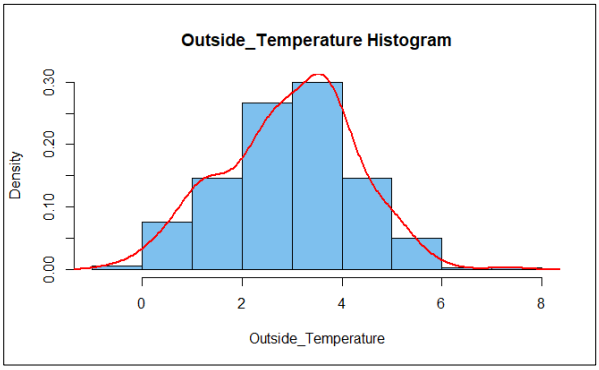


Fig-4

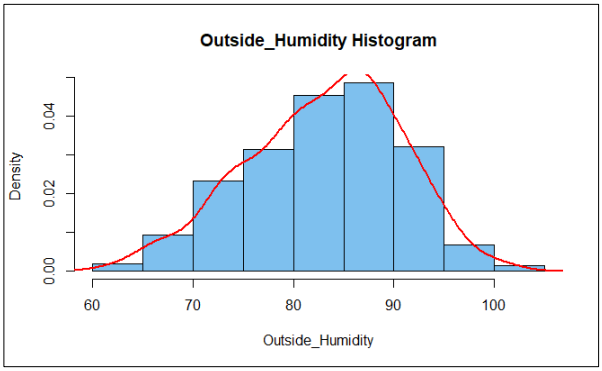


Fig-5

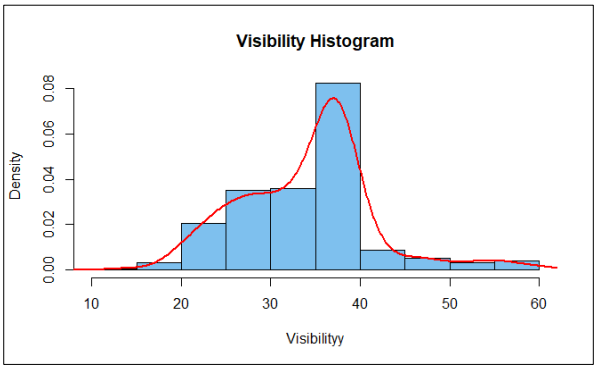


Fig-6

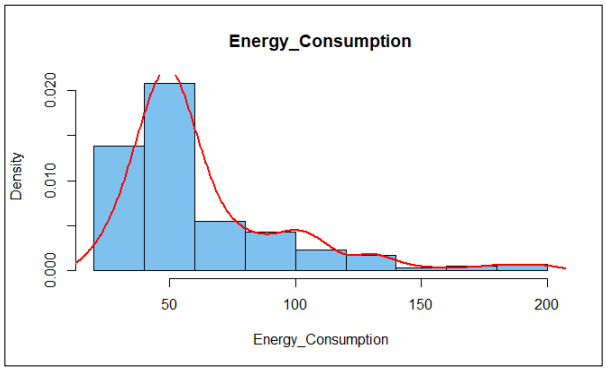


Table-1: Pearson correlation coefficient

Kitchen_Temperature	Kitchen_Humidity	Outside_Temperature	Outside_Humidity	Visibility
0.363	0.158	0.447	0.050	0.287

Table-2: Skewness of original sample data

Kitchen Temperature	Kitchen Humidity	Outside Temperature	Outside Humidity	Visibility	Energy used
0.45	0.18	-0.14	-0.31	0.43	1.76

Scatterplots between X variables and Y variable: (Pearson correlation)

1.4.7 Kitchen temperature scatterplot: There is a weak positive correlation between kitchen temperature and energy used with correlation coefficient 0.36. When kitchen temperature is up to 23° C, maximum energy consumed is up to 100. Almost 89% of data has a combination of kitchen temperature below 23° C and energy used up to 100. (Refer Fig-7 and Table-1)

1.4.8 Kitchen humidity scatterplot: There is very weak positive correlation between kitchen humidity and energy used with correlation coefficient 0.16. It seems kitchen humidity has an impact when the outside temperature is high with correlation of 0.4 between kitchen humidity and outside temperature. (Refer Fig-8 and Table-1)

1.4.9 Outside temperature scatterplot: There is a weak positive correlation between outside temperature and energy used with correlation coefficient 0.4. When outside temperature is up to 4° C, maximum energy consumed is up to 100. Almost 76% of data has a combination of outside temperature below 4° C and energy used up to 100. (Refer Fig-9 and Table-1)

1.4.10 Outside humidity: There is no correlation between outside humidity and energy used with correlation 0.05. However, it has moderate negative correlation of -0.51 with outside temperature, i.e. when outside temperature rises, outside humidity decreases. (Refer Fig-10 and Table-1)

1.4.11 Visibility: There is a weak positive correlation between visibility and energy used with correlation coefficient 0.29. (Refer Fig-11 and Table-1). Also, this variable has weak correlation with other X variables.

Fig-7

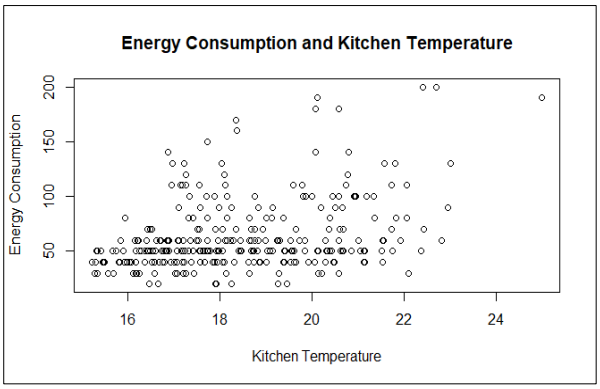


Fig-8

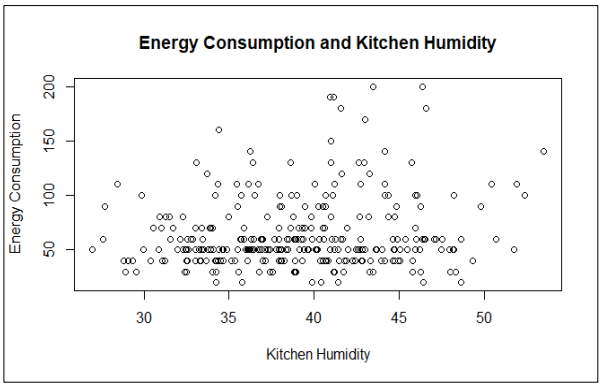


Fig-9

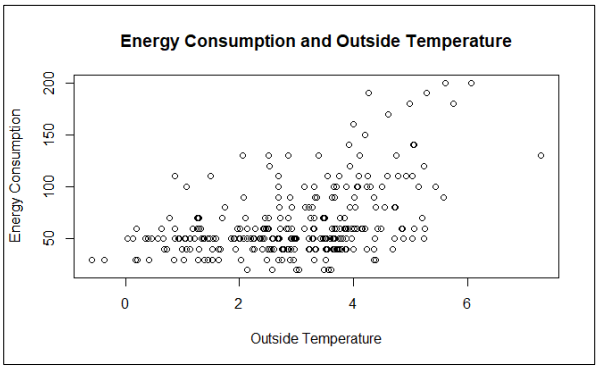


Fig-10

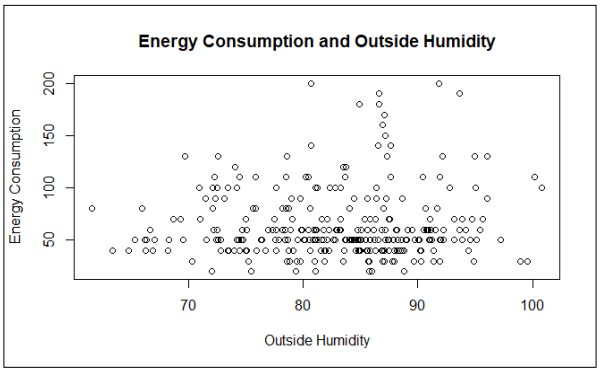
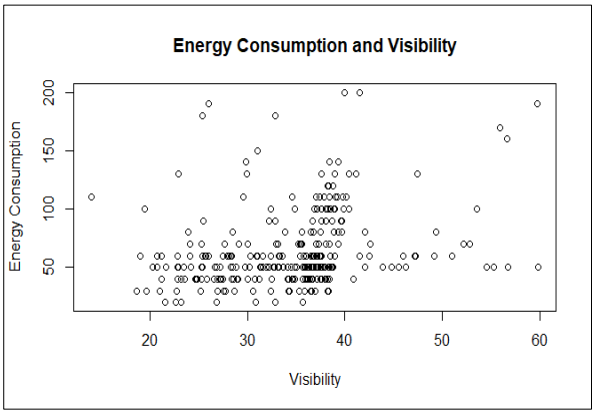


Fig-11



2.2 Transformation of the data:

Dropping a variable - Although, it is intuitive to drop Outside humidity (X4) as it has weakest correlation with Y. But RMSE of the models increased by dropping Outside humidity (X4). RMSE decreased when Kitchen humidity was dropped (X2) instead of any other variable. Since correlation does not mean causation, I crosschecked this result by applying linear regression (lm function in R) on standardised variables (Xs and Y). Kitchen humidity (X2) has lowest coefficient of -0.07 in regression, thus it was dropped.

Transformation - Standardisation was not used as no variable was approximately normally distributed. Natural log was not used as no variable has exponential distribution. Thus, first all the variables were rescaled to a unit interval using "Linear feature scaling" formula [(t-min(x))/range(x)]. Then, rescaled variables were transformed using polynomial function t^p instead of log function as rescaled data does not have large inputs. Power "p" for each variable is mentioned in Table-3.

Skewness dropped to 0 for variables except for “Energy used” which has skewness of 0.02 after transformation. Although, transformed variables does not follow normal distribution, but we can assume it will follow normal distribution asymptotically considering ‘Central Limit Theorem’ as our sample size is greater than 30. Refer below histograms.

Table-3: Power of polynomial function

X1 - Kitchen Temperature	X3 - Outside_Temperature	X4 - Outside_Humidity	X5 - Visibility	Y - Energy used
0.67	1.13	1.33	0.74	0.41

Histogram of transformed variables:

Fig-12

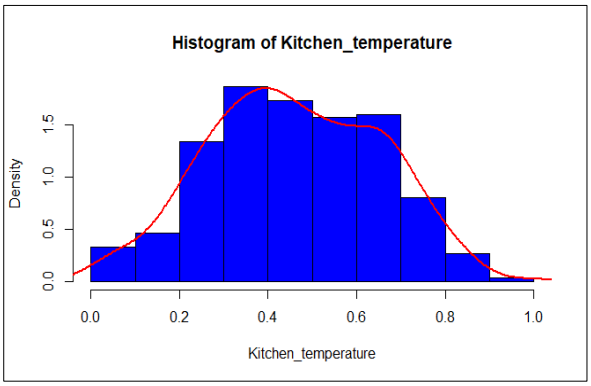


Fig-13

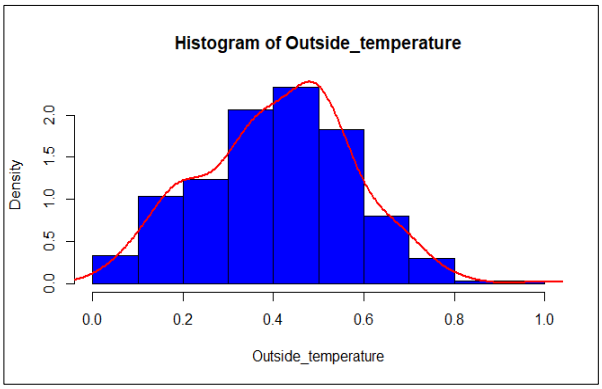


Fig-14

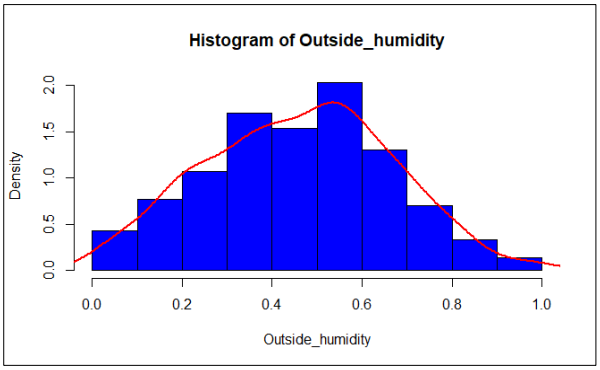


Fig-15

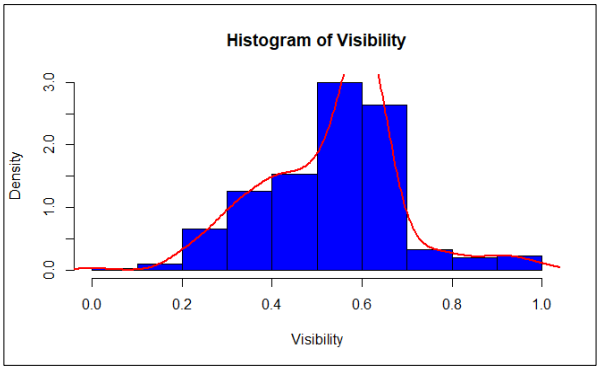
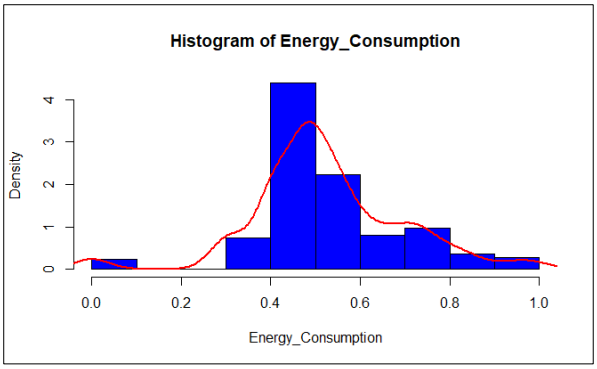


Fig-16



3.2 Output Tables:

Table-4: Error measures and correlation coefficient of fitting functions

Parameter	WAM	WPM (p = 0.5)	WPM (p = 2)	OWA	Choquet integral
RMSE	0.148	0.155	0.140	0.141	0.132
Av. abs error	0.109	0.114	0.103	0.105	0.094
Pearson correlation	0.549	0.498	0.591	0.541	0.621
Spearman correlation	0.536	0.514	0.558	0.506	0.585

Table-5: Weights and Orness of fitting functions

Parameter	WAM	WPM (p = 0.5)	WPM (p = 2)	OWA	Choquet integral
Orness	-	-	-	0.628	0.696
Kitchen Temperature	0.159	0.135	0.161	0.191	0.182
Outside Temperature	0.284	0.224	0.398	0.192	0.484
Outside Humidity	0.154	0.140	0.169	0.159	0.172
Visibility	0.402	0.501	0.272	0.458	0.162

3.4 Interpretation of the results:

a) Choquet integral is the best fitting function among five functions with lowest RMSE of 0.132 and highest Pearson correlation of 0.621 followed by WPM (p=2) with RMSE 0.14 and Pearson correlation of 0.59. Average of the predicted values of Energy used by Choquet integral is 0.52 (transformed) is closest to the average of the original sample data (0.54). WPM (p=0.5) has highest RMSE of 0.155 and lowest Pearson correlation of 0.498. OWA is slightly better than WAM in accuracy, but at par in Pearson correlation.

b) Choquet integral is the best fitting function followed by WPM (p=2). Both the functions have allocated more weight to Outside temperature. However, rest of the functions OWA, WAM, WPM (p=0.5) have allocated more weight to Visibility. Thus, as per Choquet integral and WPM (p=2), Outside Temperature is the most important variable and for rest of the functions, Visibility is the most important variable to predict Energy utilization. Weight allocated for Kitchen temperature and Outside humidity is between 0.13 and 0.19 by all the fitting functions. Thus, Outside temperature and Visibility are important variables comparatively.

c) Fuzzy measure weight of Outside temperature variable is maximum at 0.651 followed by Visibility at 0.462. Correlation of Kitchen temperature and Outside humidity with Energy used is 0.36 and 0.05 respectively, however their Fuzzy measure weights and fitting weights are similar. This shows correlation does not reflect causation. All variables interact in sub-additive/redundant way. No combination was greater as whole than addition of subsets. V({1,3,4}) has lowest Fuzzy measure as compared to the its sum, followed by V({1,2,3,4}). V({1,2}) Fuzzy measure was closed to its sum followed by V({2,3}). Fuzzy measure of Visibility V({4}) adds no value when combined with two or more variables. For eg. V({1,2,3,4}) ~ V({1,2,3}) = 1; V({2,3,4}) ~ V({2,3}) = 0.952; V({1,3,4}) ~ V({1,3}); and V({1,2,4}) ~ V({1,2}). (Refer Table-6).

d) Regarding OWA, maximum weight is allocated to W4 = 0.458 which is the highest input. Also, OWA has Orness of 0.628. This suggests that OWA favours higher inputs. Even, Choquet integral favours higher inputs as per Orness of 0.696. WPM with p=0.5 has less accuracy than WPM with p=2. This suggests that models favour higher inputs than lower inputs as power mean with p>1 is more affected by higher inputs.

Table-6: Fuzzy measure

Binary	V (Sets)	Fuzzy measure	Addition	Relationship	Fuzzy measure/Addition
0001	1	0.366			
0010	2	0.651			
0100	3	0.354			
1000	4	0.462			
0011	12	0.976	1.02	Redundant	96%
0101	13	0.612	0.72	Redundant	85%
1001	14	0.558	0.83	Redundant	67%
0110	23	0.952	1.01	Redundant	95%
1010	24	0.779	1.11	Redundant	70%
1100	34	0.596	0.82	Redundant	73%
0111	123	1.000	1.37	Redundant	73%
1011	124	0.976	1.48	Redundant	66%
1101	134	0.612	1.18	Redundant	52%
1110	234	0.952	1.47	Redundant	65%
1111	1234	1.000	1.83	Redundant	55%

4. Prediction using the best fitting model:

4.2 Since Choquet integral is the best function, Choquet(x,v) function is used to predict the value based on transformed new X variables. Predicted Energy used is 0.5060715 which gives the value of 54.19 after reverse transformation. Since predicted Energy used value is 54.19 (approximately average of 50 and 60), average of X variables of the observations with Energy used 50 and 60 is closed to the input (Refer Table-7). Thus, I think although correlation of the fitted values is 0.62, but still predicted value is reasonable.

Table-7: Comparison between new X variables and average value of 50 & 60 units of Energy

	Kitchen temperature	Outside temperature	Outside humidity	Visibility
Avg of original dataset	18.50	2.65	83.54	33.95
New X variables	18	4	4.8	31.4

4.3 Outside temperature is the key variable as humidity and visibility logically will depend on the temperature. Anybody would like to have normal temperature in kitchen, thus use of appliances (A/C or heater) will also depend on the outside temperature. I think if outside temperature is approximately 2.5 to 2.6, that will keep energy utilisation at 60 on an average, if outside temperature is greater than 3.5, then energy utilisation will be 80 on an average. However, if we have more information (Variables) like geographical location, appliances used in house, season of that region when data was captured, number of rooms in house, number of family members bifurcated in kids and adults, then accuracy of the prediction will increase, and we would be able to identify key variables to reduce electricity consumption.

Reference:

James, S. (2016). *An Introduction to Data Analysis using Aggregation Functions in R*. Cham: Springer International Publishing.