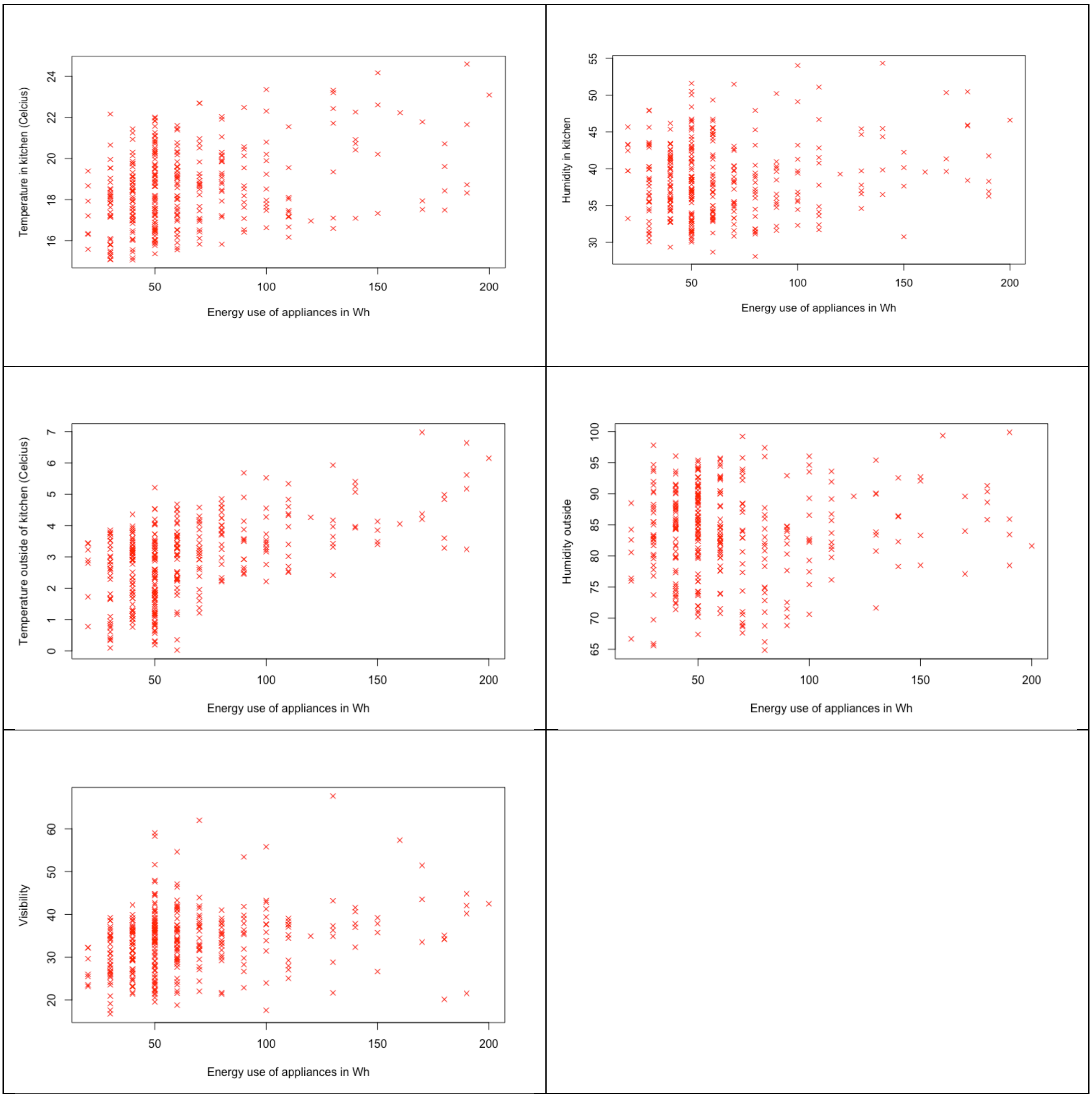
SIT718: Real World Analytics

Assignment 2

Energy Prediction of Domestic Appliances

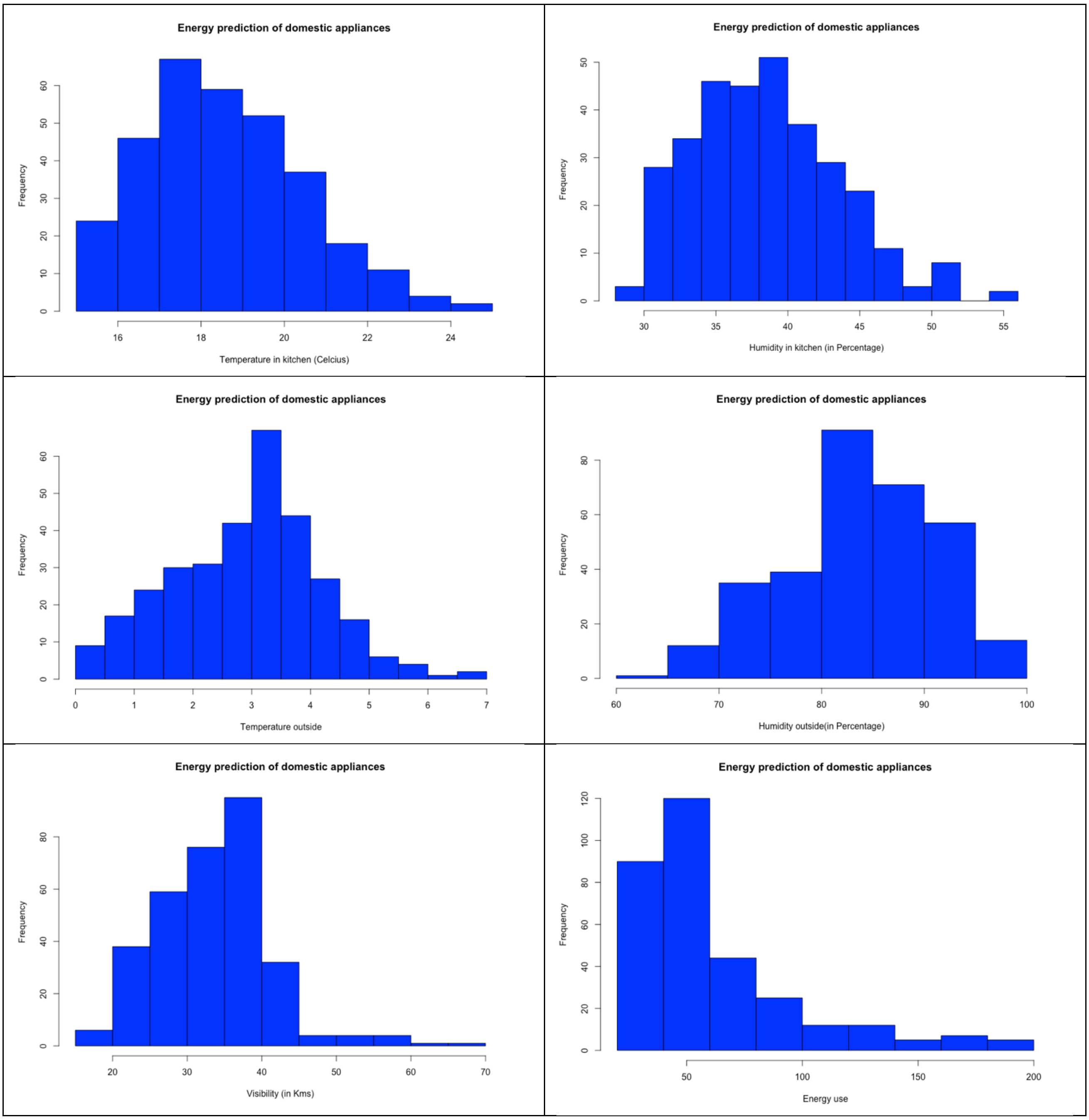
Sanket Anil Thakur
Student ID: 220617318
Graduate Certificate of Data Analytics

Task 1:
Visualization: Scatter plots



1) We can see that the correlation between the temperature in the kitchen and use of appliances is weak as the Correlation coefficient= 0.3292527.
2) We can see that the correlation between the humidity in the kitchen and use of appliances is weak as the Correlation coefficient= 0.1449481
3) We can see that the correlation between the temperature outside the kitchen and use of appliances is weak, as the Correlation coefficient= 0.5280358
4) We can see that there is no correlation between the humidity outside and use of appliances, as the Correlation coefficient= 0.06686507
5) We can see that the correlation between the visibility and use of appliances is weak as the Correlation coefficient= 0.2470743
Scatter plot might not be the best visualisation to create predictions as the assumptions are made visually and they might not be the most accurate.

Visualization: Histograms



To calculate the skewness of the given variables the E1071 package from the CRAN repository was used.

1) The first histogram indicates that the temperature inside the kitchen is between 17ºC to 20ºC
The skewness of the histogram is 0.4585775, hence is it positively skewed.
2) The second histogram shows the humidity in the kitchen. This is a positively skewed histogram as the skewness is 0.4976795. It shows the average humidity in the kitchen is around 38%
3) The histogram showing temperature outside the kitchen is negatively skewed as the skewness is -0.009929214. The average temperature outside lies between 2ºc to 4ºc
4) The fourth histogram show the humidity outside the kitchen lies between 60% to 70%. This is a negatively skewed histogram and the skewness is -0.3433646
5) This histogram shows the visibility from the weather station in kms. This is a positively skewed histogram and the skewness is indicated at 0.6847032
6) This histogram shows the use of energy in Wh. It is a positively skewed histogram and the skewness is 1.627407. It also shows that the average energy consumption lies between 40Wh to 50Wh

Task 2:

Transformation of the Data

The given data from "Energy20.txt" is transformed as "Sanket-transformed.txt" by using five variables and 320 rows. I have skipped the variable X5: Visibility (from weather station), because I think it doesn't have a direct relation with the other variables given in the data set. Hence, I am using the following variables to work on my assignment.

X1: Temperature in kitchen area	Normalization
X2: Humidity in kitchen area	Standardization
X3: Temperature outside	Normalization
X4: Humidity outside	Standardization
Y: Energy use of appliances	Standardization

I have used linear feature scaling in transforming the data as it helps us use the data ranges instead of the unit interval. Linear feature scaling also helps improve use of functions and algorithms.

I have used the Normalization or Min-Max scaling to transform variables X1 and X3. In this technique the value of the variables are scaled between a range of 0 and 1. (Bhandari, 2020)

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

I have also used standardization as another technique for transformation of the X2,X4 and Y. Standardization is a type of scaling where the values in the variable are concentrated around the mean, with the use of unit standard deviation. It also helps to interpret the abnormality in the data and bring all the values on a standard scale. (James, 2016)

$$X' = \frac{X - \mu}{\sigma}$$

Task 3
Build models and investigate the importance of each variable

Table of Correlation coeffects and Error measures

Functions	RMSE	Av. Abs Error	Person Correlation	Spearman Correlation	Orness
weighted arithmetic mean	0.985684139	0.791156435	0.287380935	0.175453478	
Weighted power means, p=0.1	abs error NaN	abs error NaN	abs error NaN	abs error NaN	
Weighted power means, p=10	1.047602125	0.910742687	0.537015773	0.494811502	
ordered weighted averaging function	0.976793233	0.748172122	0.235354091	0.121693312	0.379739926
Choquet integral	0.953716611	0.716522536	0.299481832	0.173015102	0.154071291

Table of Weights after prediction

Functions	Weight 1	Weight 2	Weight 3	Weight 4
weighted arithmetic mean	0	0.327722872	0.535861004	0.136416125
Weighted power means, p=0.1	0.072188792	0.001645833	0.924792655	0.00137272
Weighted power means, p=10	0.285839876	0.045251301	0.641627111	0.027281712
ordered weighted averaging function	0.372908553	0.114963116	0.512128331	0

Choquet Intergral

binary number	fm.weights
1	0
2	0
3	0
4	0.033841074
5	0.63471091
6	0.033841074
7	0.678200398
8	0
9	0
10	0
11	0
12	0.033841074
13	0.63471091
14	0.033841074
15	1

i Shapley	I
1	0.39538131
2	0.0949464
3	0.42922239
4	0.0804499

We have used five fitting functions to study the correlation coeffect and the error of measure. As per my opinion, the Choquet integral is the best fitting function, as it gives the least percentage of error compared to other funtions. This also has a relatively low Avg abs Error and RSME, compared to the other functions. The Weighted arithmetic mean has given no weight to X1: Temperature in kitchen area and the Ordered weighted averaging function has given no weight to X4: Humidity outside

We can notice that variable the most important variable for forecasting use of energy is X3: Temperature outside the kitchen. It the best correlation with variable Y compared to rest of the variables in the data set. X1 and X2 have a weak correlation with the

If we look at the Orness value in Choquet integral and Ordered weighted mean, it is less them 0.5. Hence it can be said that they are more affected by the lower input values.

Task 4

Here I have used the Choquet integral as the best fitting function. I have used the following variables in the data set

Given Data	X1=16; X2=38; X3=4; X4=77
After applying Choquet intergral	X1=0.09695058; X2=0.3822373; X3=0.58253329; X4=0.40284051

The given data set has been compared to the original data set (my.data). Then the linear regression was revered to get the correct answer
The energy use of the given data set can be determined as Y= 25.45296

It is reasonable to say that the choquet integral is the best fitting function since it has a low RMSE and Avg abs error.

As per my selected variable, the best condition for low usage of energy will be outside temperature. The temperature inside the kitchen has less to no effect on the usage of energy of the appliances.

Task 5

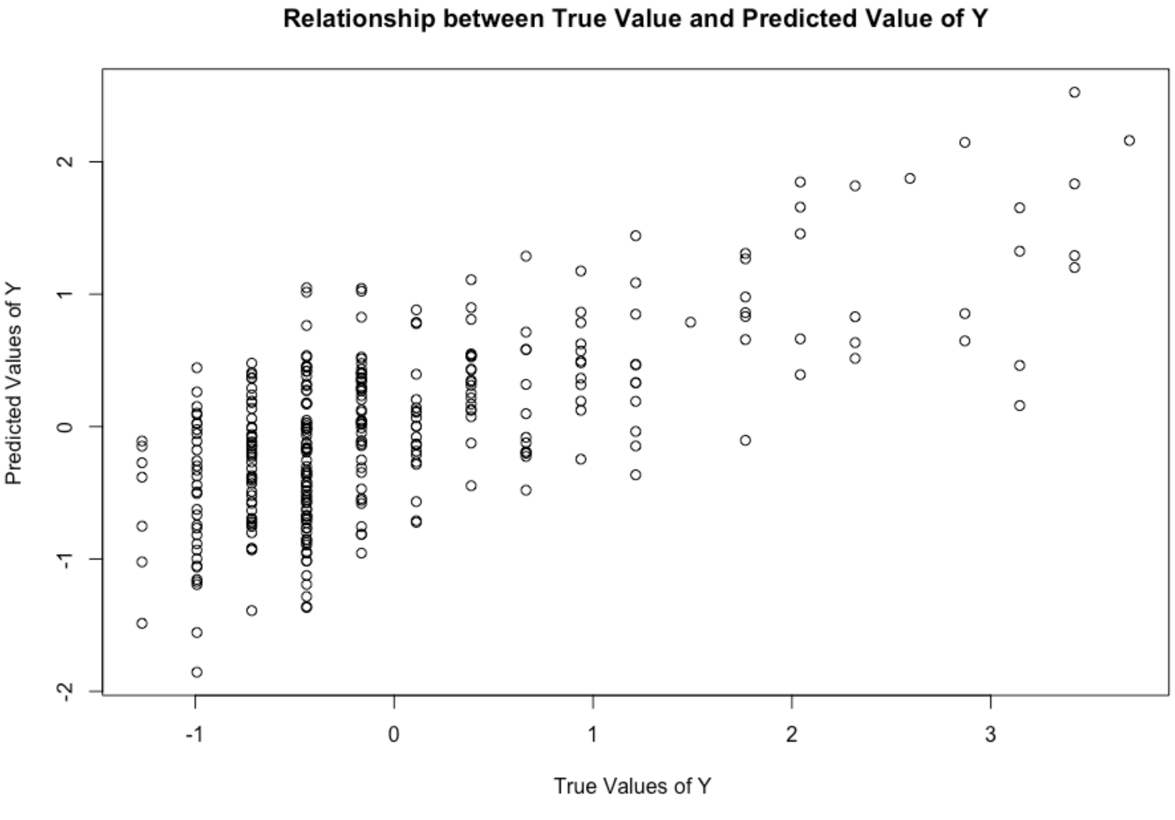
I have used my transformed data (sorted.data) as the data set in this section.

Summary of the Functions (fit.linear.model)

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.48904 -0.51800 -0.04382  0.38038  2.98660

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.07298    0.13702 -15.130  < 2e-16 ***
X1           1.32997    0.20713   6.421 4.99e-10 ***
X2          -0.09217    0.04420  -2.085   0.0379 *
X3           3.82065    0.27215  14.039  < 2e-16 ***
X4           0.35565    0.04623   7.694 1.85e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7378 on 315 degrees of freedom
Multiple R-squared:  0.4624,      Adjusted R-squared:  0.4556
F-statistic: 67.74 on 4 and 315 DF,  p-value: < 2.2e-16
```



We can say that the predicted values and the true values have a 38% variance in them.

Bibliography

Bhandari, A., 2020. *Analytics Vidhya.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/
[Accessed 28 August 2020].
James, S., 2016. *An Introduction to Data Analysis using Aggregation Functions in R.* Melbourne: Springer, Cham .