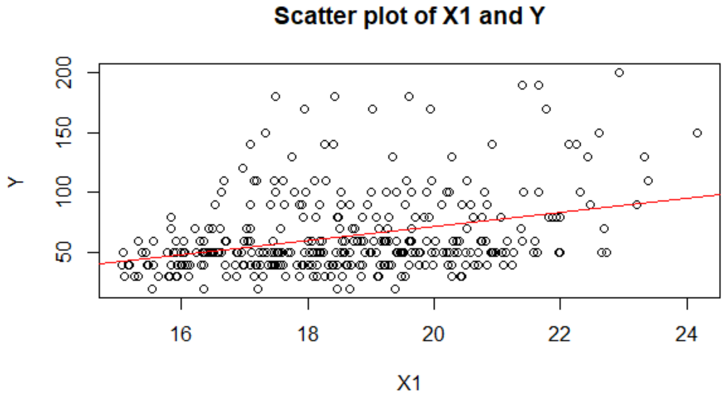
SIT718 - ASSESSMENT 2

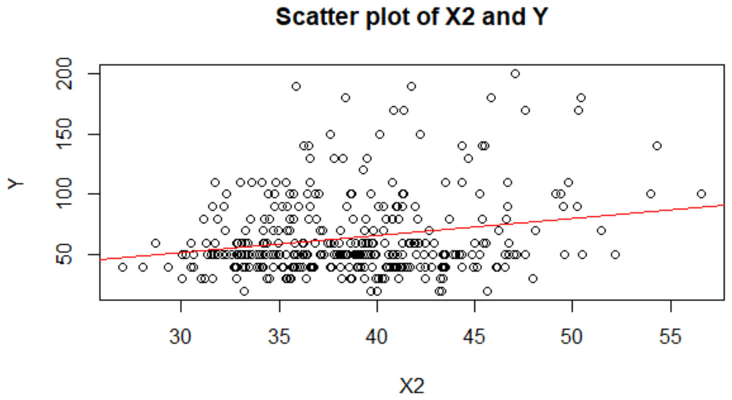
Problem Solving

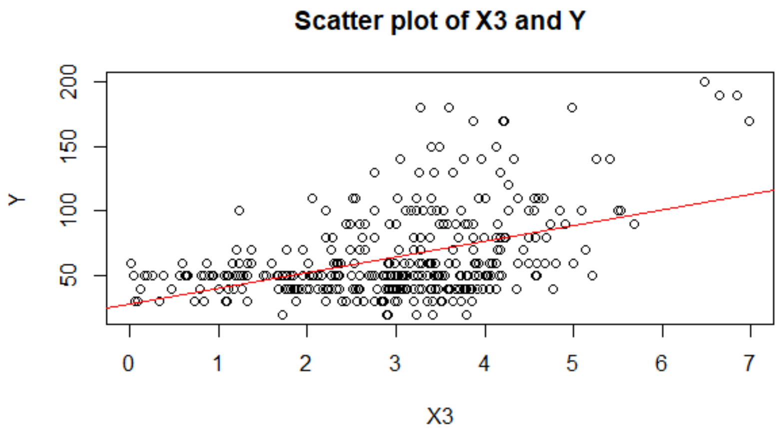
RAJESHKUMAR RAMPRASAD MOURYA
rmourya@deakin.edu.au
218615876

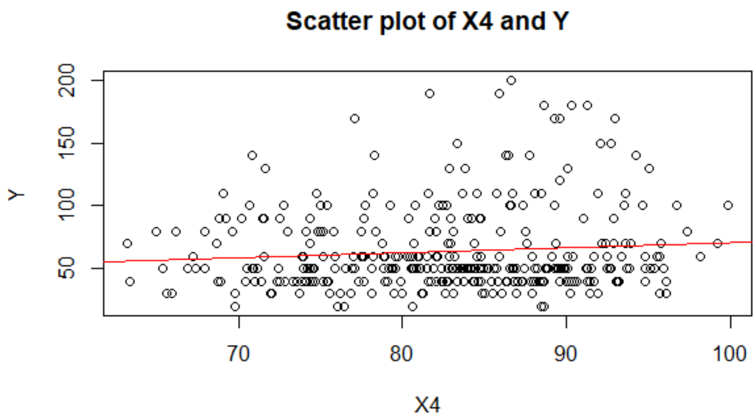
PART 1: Understand the data

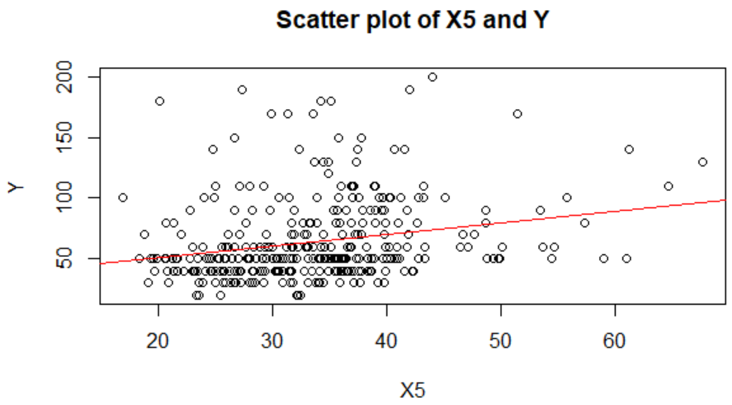
1. **Scatter plots**



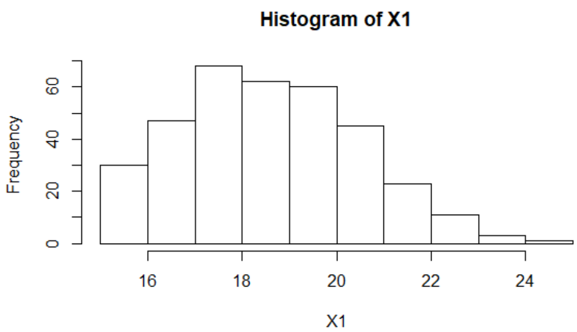


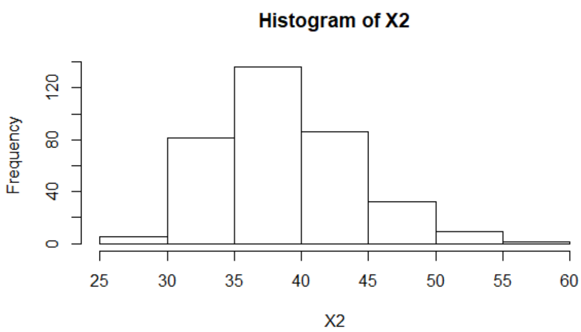


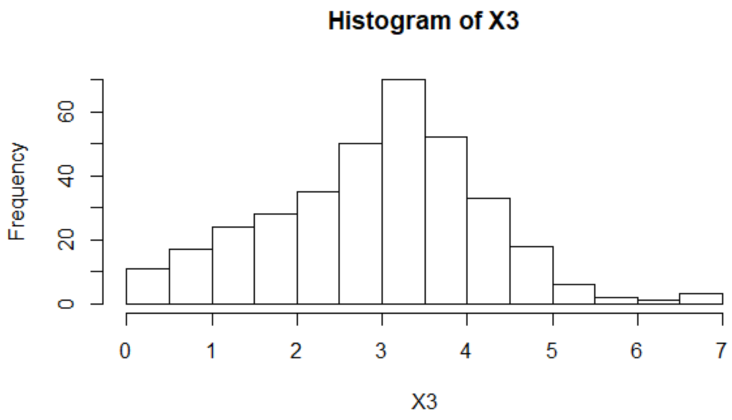


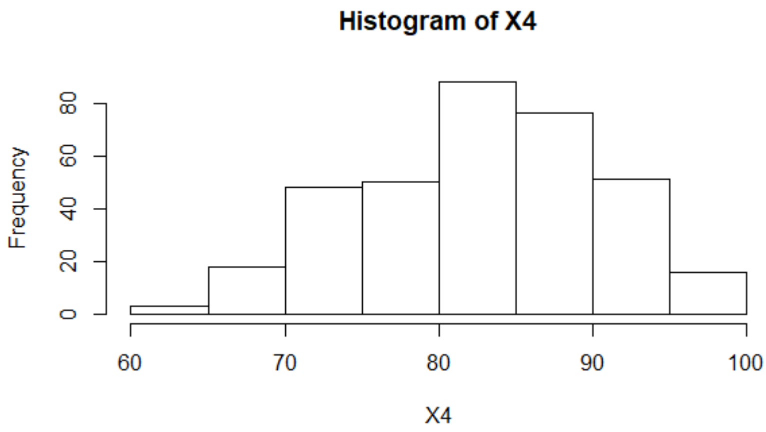


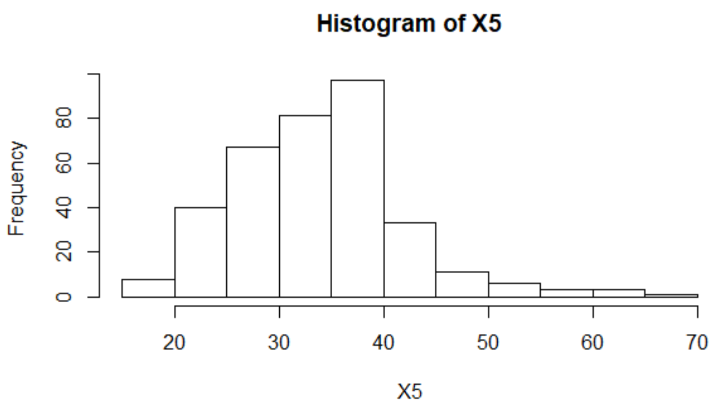
2. Histograms

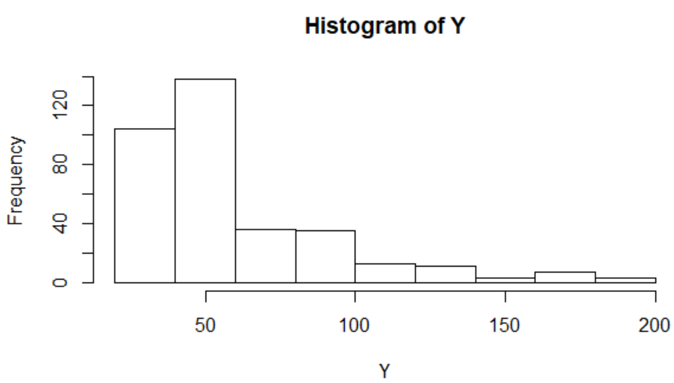





Histogram of X3
Frequency
0
20
40
60
0
1
2
3
4
5
6
7
X3


Histogram of X4
Frequency
0
20
40
60
80
60
70
80
90
100
X4


Histogram of X5
Frequency
0
20
40
60
80
20
30
40
50
60
70
X5


Histogram of Y
Frequency
0
40
80
120
50
100
150
200
Y

3. **Summary:**
   a. **Relationship between variables:**
   Observation of scatterplots suggest Y value increases with increase in X1, X2,X3, X5 but there is low positive correlation because of large number of outliers and for Y and X4, there is no correlation.
   b. **Independent Variables:**
   X1 : No normal distribution and Right skewness in the data
   X2: Right skewness is observed
   X3: No skewness but there is no symmetry in distribution
   X4: Low skewness is observed
   X5: Range of data show most of the data falls between range 20-50 and several outliers disturb symmetry of distribution
   Y: No normal distribution and Right skewness in the data

PART 2: Transform the data

1. **Selected variables with Y : X1, X2, X3 and X4**
   a. **Log transformations** are applied on X1,X2 and Y to reduce right skewness
   b. **Linear scaling** is applied for all variables from X1 to X4 and Y as the data ranges are in different, it is suitable to scale them back to unit interval before applying aggregation.

PART 3: Build models and investigate

1. **Tables**
   *a. **Error measure and correlation coefficient***

	WAM	**WPM**		**OWA**	**Choquet integral**
		p = 0.5	**p = 5**		
RMSE	0.155675870022490	0.160619849806489	0.167129679417582	0.166330414967351	0.154629861073782
Av. abs error	0.124047952680523	0.128892886268474	0.132845365178088	0.131486850614138	0.121400848324013
Pearson correlation	0.613267596315504	0.577361952650783	0.551363200085280	0.525227883323086	0.619912042355637
Spearman correlation	0.550647862202653	0.504085984458842	0.502634684105638	0.450031905469567	0.556160216252595
Orness				0.472419984886939	0.512531936724884

b. Summary of weights/parameters

Weights(w_i) and Shapley values					
	WAM	**WPM**		**OWA**	**Choquet integral**
		p = 0.5	**p = 5**		
i	w_i	w_i	w_i	w_i	Shapley i
1	0.271909385856564	0.241206211890278	0.133118032662503	0.350626369968100	0.282639018790293
2	0	0.037456701930855	0	0.206990084491559	0.019241062401148
3	0.469949427811505	0.401007431919290	0.835653414907011	0.116880766451717	0.478311496240708
4	0.258141186331932	0.320329654259576	0.031228552430486	0.325502779088609	0.219808422578908

Choquet integral fm weights

Binary	fm weights
1	0.219175498128903
2	0
3	0.219175498128903
4	0.315080630847907
5	0.947498408799821
6	0.545973379661499
7	0.947498408799875
8	0.347431549324234
9	0.475105592604460
10	0.347431549324234
11	0.475105592604460
12	0.693581165181730
13	1.000000000011030
14	0.693581165181768
15	1.000000000011060

2. Comparison of models

a. Fitness of Models

Compared to OWA, WPM models, WAM and Choquet integral perform better in terms error measures RMSE and Average absolute error as well as Spearman and Pearson coefficients. Compared to WAM, Choquet integral performs slightly better in these aspects. Hence, we will be using Choquet integral as the best fitting model for prediction.

b. Importance of variables

For all models expect OWA, highest weight is given to X3 followed by X1. Given the nature of OWA, which strengthen lower inputs, it suggests order of weights to be X1, X4, X2 followed by X3. Provided Choquet and OWA are better performing models, X3 is of most importance from X1, X2 and X3 to predict Y.

c. ***Interaction between variables***
Spearman and Pearson coefficients lying around 0.5 suggest there is a moderate positive correlation between variables.

d. **Choquet integral suggests Orness value higher than 0.5 which means better model tends towards higher inputs.**

PART 4: Use your model for prediction

1. **Best fitting model:** based on the comparison, Choquet integral is the best fitting model for this dataset
2. **Result of prediction:**

Predicted Y value for based on X1 = 17, X2 = 39, X3 = 4, X4 = 77 is **53.67676**

Provided the range of Y values in original dataset and the dataset observations against selected variables, I think this is a reasonable prediction.

3. **Conditions for high energy use of appliances**
Increase in the temperature of kitchen area (X1) and increase in outside temperature (X3) will result in increase in usage of appliances , i.e. higher temperature will result in higher energy consumptions.

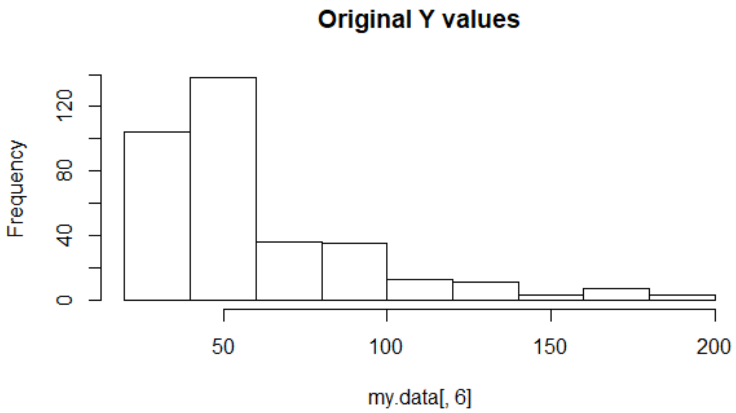
PART 5: Compare with a linear regression model

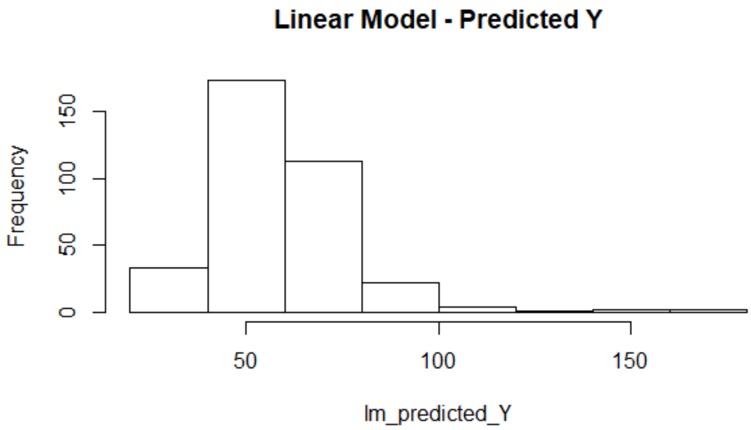
1. **Summary of Linear Model**

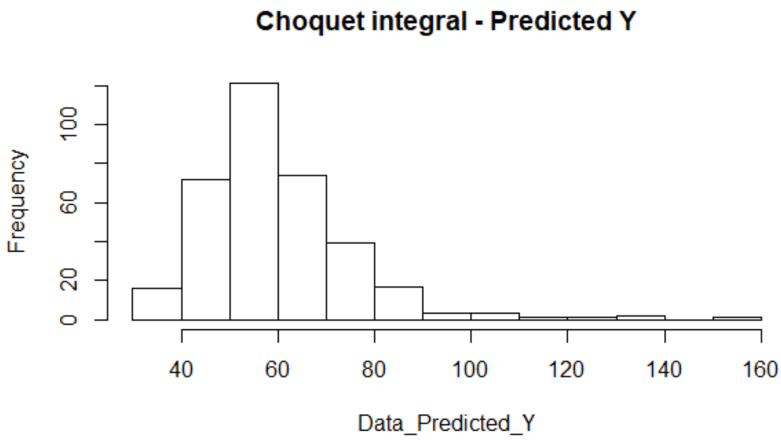
```
Residuals:
    Min       1Q   Median       3Q      Max
-0.50676 -0.10134  0.00279  0.10191  0.45115

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.10889    0.04496  -2.422    0.016 *
lm_X1        0.28392    0.03899   7.282 2.25e-12 ***
lm_X2       -0.03222    0.05070  -0.635    0.526
lm_X3        0.63385    0.05551  11.418  < 2e-16 ***
lm_X4        0.35771    0.04482   7.980 2.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1546 on 345 degrees of freedom
Multiple R-squared:  0.3818,	Adjusted R-squared:  0.3746
F-statistic: 53.27 on 4 and 345 DF,  p-value: < 2.2e-16
```

2. **Performance of the models:**
Comparison of predicted values in linear model and Choquet integral with original Y values suggest that maximum data resides in the range between 40 to 80 across all

these datasets as evident in the histogram below suggesting these models are relatively accurate.







3. **Difference between the linear model and Choquet integral**
   Linear model shows relatively better results in terms of predictions as it includes relatively distant outliers as well (ex. Y = 200), whereas Choquet integral model doesn't deal very well with such outliers and reduces to the data near more frequent data. In terms of computations Choquet integral requires more iteration (2^n computations for n variables) while Linear model builds relatively quicker. Given small data size, it is not an

issue now, but when higher number of variables are considered, linear model will perform better.

PART 6: References

James, Simon 2016, An introduction to data analysis using aggregation functions in R, Springer International Publishing, Cham, Switzerland, doi: 10.1007/978-3-319-46762-7.

Rcompanion.org 2016, R Handbook: Transforming Data, retrieved 25 April 2020 <https://rcompanion.org/handbook/I_12.html>.

Quick, John 2009, Simple linear regression example, retrieved 25 April 2020,

<https://www.r-bloggers.com/wpcontent/uploads/2009/11/simpleLinRegExample1.txt>.

www.rdocumentation.org. (n.d.). stats package | R Documentation, retrieved 29 April 2020,<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm>.

Chapter 2 Multiple Regression I (Part 1) 1 Regression several predictor variables. (n.d.), retrieved 29 April 2020,
<https://web.njit.edu/~wguo/Math644_2012/Math644_Chapter%202_part1.pdf>.