

Assignment Submission Coversheet

Faculty of Science, Engineering and
Built Environment



Student ID: 220129254

Student Name: Alina Kriutchenko

Campus: ☒ Burwood ☐ Waterfront ☐ Waurin Ponds ☐ Warrnambool ☐ Cloud

Assignment Title: Using aggregation functions for data analysis (Assignment 2)

Due Date: Thursday 10th May 2020, 11:30 pm

Course Code/Name: Master of Data Science

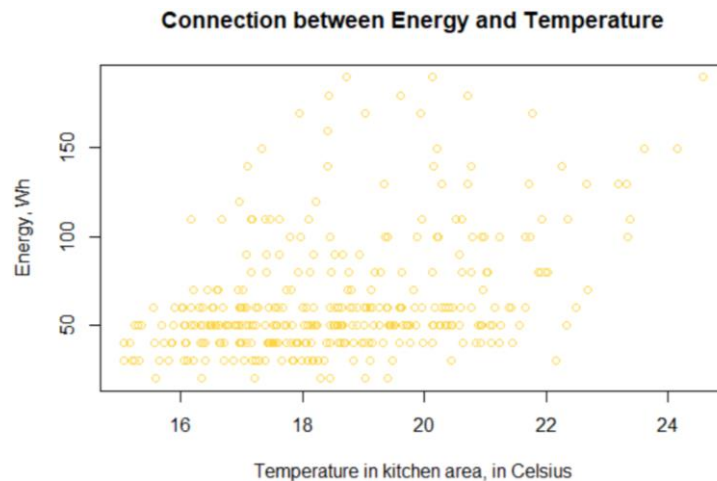
Unit Code/Name: MIS770

Unit Chair/Campus Coordinator: Dr. Ye Zhu

1. Understanding the data

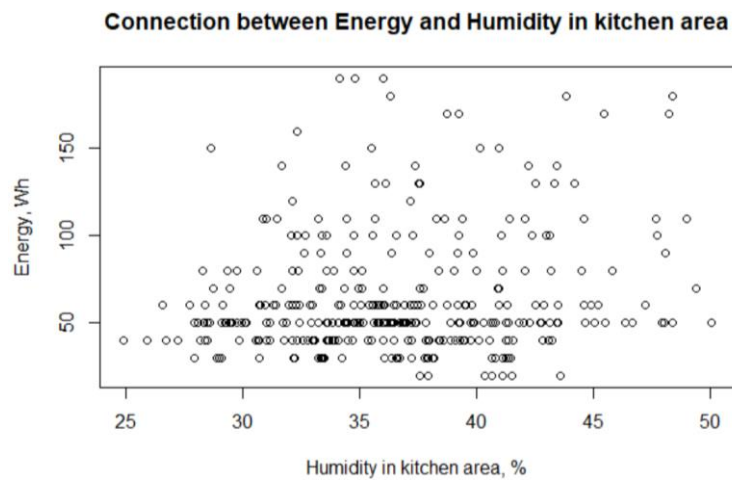
Using scatter plots and histograms, report on the general relationship between each of the variables X1, X2, X3, X4, X5 and the variable of interest Y.

Scatter plot 1



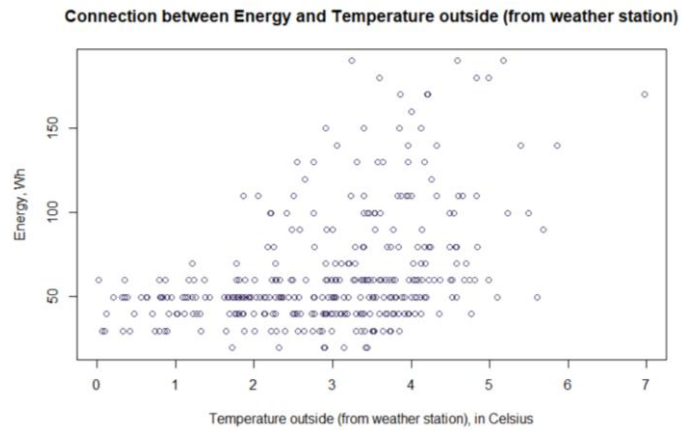
The above scatter plot shows the on average when the temperature in kitchen area is lower than 21 degrees of Celsius, the energy used more often.

Scatter plot 2



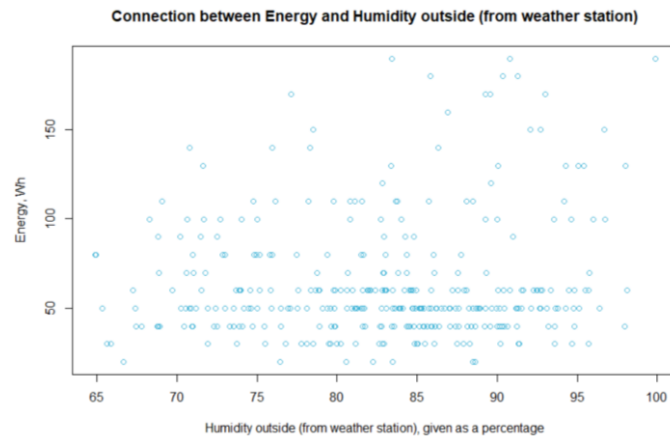
On average the energy used more often as a Humidity index is 30-40 percentage.

Scatter plot 3



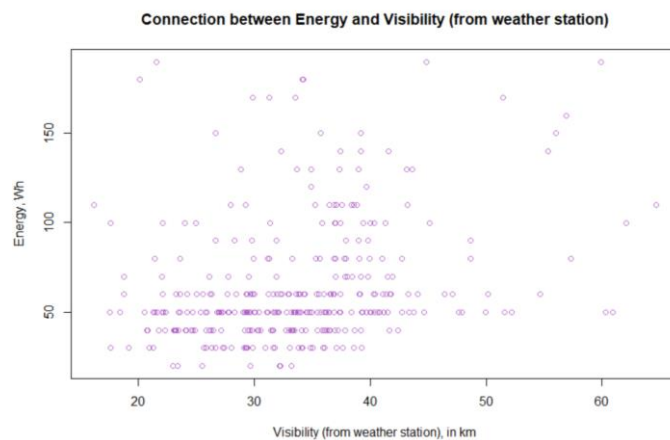
In general, energy use is approximately from 20 to 55 Wh. During these usual times, the temperature outside goes below 4 degrees of Celsius.

Scatter plot 4



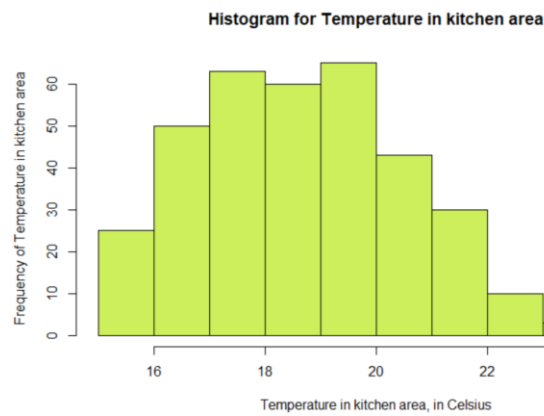
Energy used the most often during 75-90 percentage of humidity outside according to the weather station.

Scatter plot 5



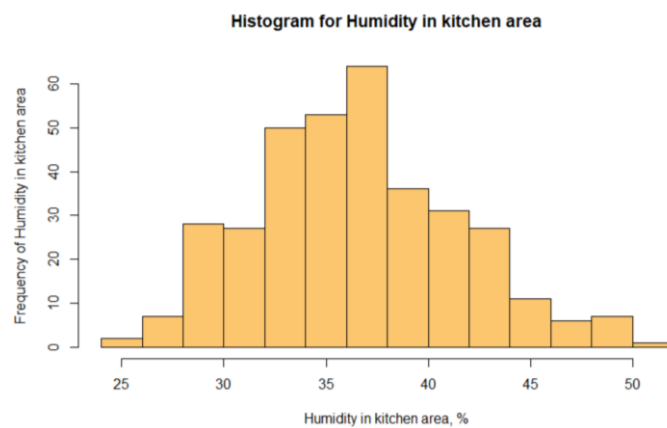
During lower visibility, the energy being used more often.

Histogram 1



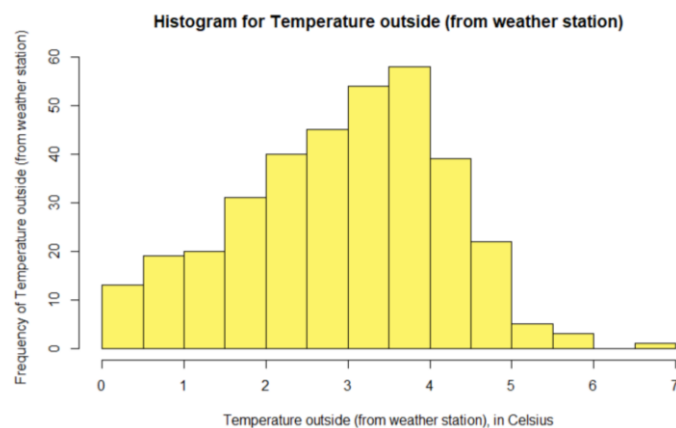
There is a small positive skewedness for the visualisation for the temperature in the kitchen area.

Histogram 2



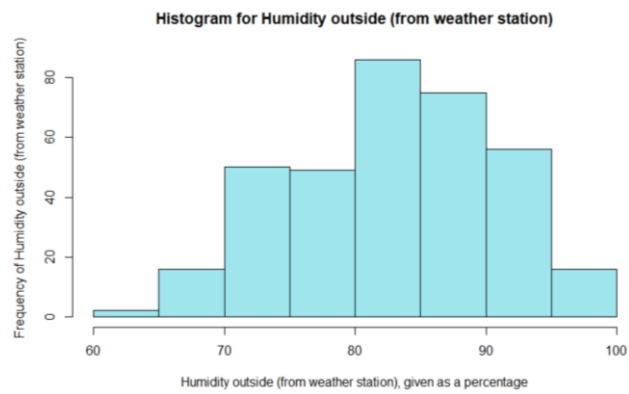
The humidity histogram's visualisation is symmetrical.

Histogram 3



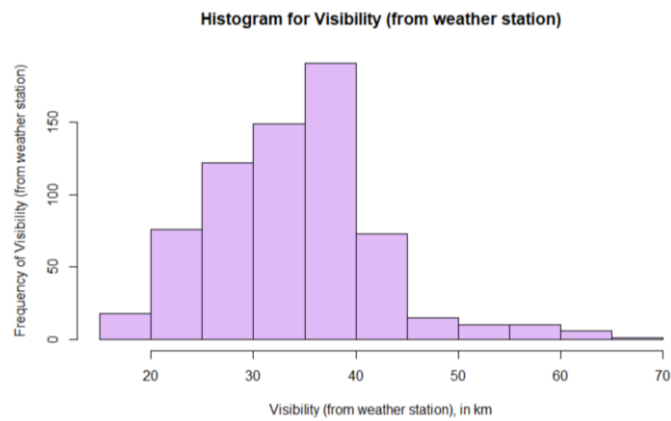
This histogram is relatively symmetric, but some skewedness still takes place.

Histogram 4



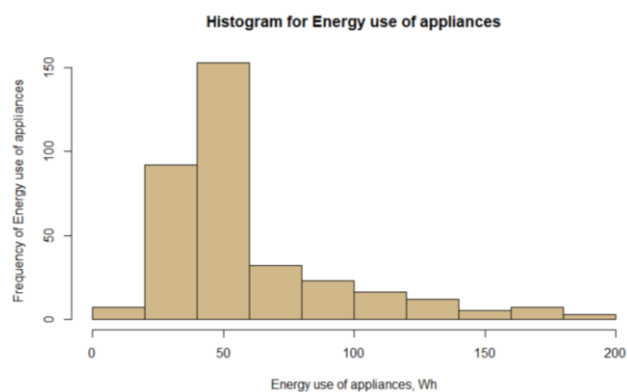
The histogram for the humidity outside has some negative skew.

Histogram 5



The distribution has a noticeable positive skew.

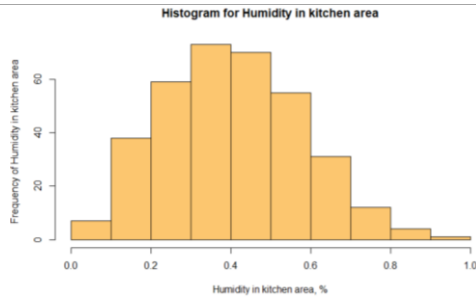
Histogram 6



The histogram for the energy use of appliances shows a significant positive skewedness that requires necessary transformations.

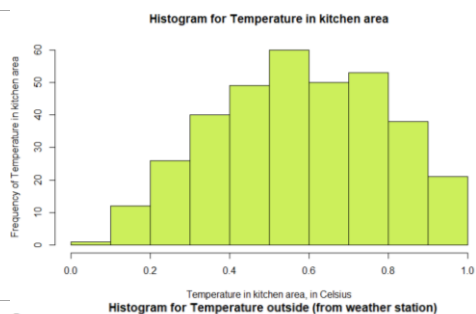
2. Data transformation

Linear Feature Scaling

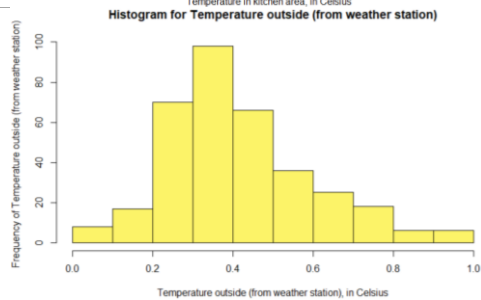


The Linear feature Scaling changed the x-axis to the diapason from 0 to 1.

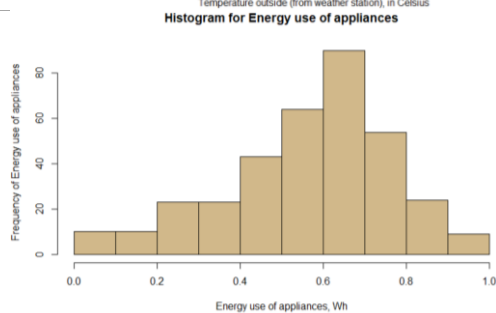
Polynomial transformation and scaling



It improved the skewedness of the histogram.

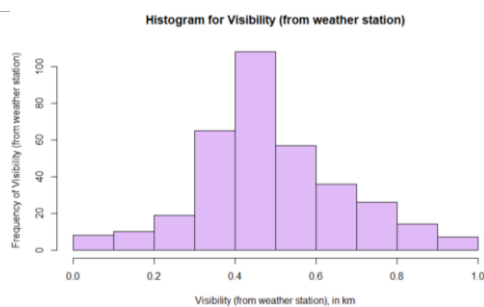


The histogram becomes less exponentially distributed, more symmetric.



As the previous ones the histogram becomes more visually balanced and changed the diapason of the x-axis.

Log transformation and Scaling



This function helped archived less skewed visualisation of visibility and changed the diapason of the x-axis.

3. Building models and investigating the importance of each variable

Table 1. Error measures and correlation coefficients

	Weighted arithmetic mean	Weighted power mean, $p=0.5$	Weighted power mean, $p=5$	Ordered weighted averaging function	Choquet integral
RMSE	0.1810538	0.4193281	0.1910299	0.1897017	0.1761856
Av. abs error	0.1485632	0.3753028	0.1495677	0.1481458	0.1390750
Pearson correlation	0.5609349	0.5985638	0.4203941	0.4228137	0.5445257
Spearman correlation	0.5284806	0.5282223	0.4206171	0.4164052	0.5234348
Orness				0.7344568	0.6596508

a. The WAM, WPM (0.5) and Choquet integral have a high Pearson correlation and similar spearman correlation. But WPM (0.5) has the highest number for RSME and Av. abs error, while Choquet has the lowest RMSE and Av. abs error amongst all of them. Therefore, Choquet integral is the best fitting function. The ordered weighted averaging function and Choquet integral are both close to the maximum function, but OWA has a slightly higher result. The second model is weighted arithmetic mean. The correlation result can be higher for this model, but the model is acceptable. The orness of the OWA function and choquet integral is neither maximum nor minimum function. However, it is closer to the maximum, especially OWA function.

Table 2. Fuzzy measures

	WAM	WPM, $p=0.5$	WPM, $p=5$	OWA function	Choquet integral
1	0.457106306	0.502347997	0.231839791	0.120925575	0.265393184
2	0	0	0.113631252	0.020440943	0.031326135
3	0.342947720	0.320034911	0.355177629	0.392971961	0.438676217
4	0.199945975	0.177617093	0.299351329	0.465661521	0.264604464

b. The importance of each of the variables

The most important variable for Choquet integral is the number three, which X3, the temperature outside. There are two equally important variables: 1 (X1, temperature in the kitchen area) and 4 (X5, Visibility). The least important is number 2 (X2, humidity in kitchen area).

In general humidity in the kitchen area is the most unimportant variable for each function. For WAM and WPM (0.5) the temperature in the kitchen area is the most essential parameter. The WPM function has the same the most important parameter as Choquet integral. Overall, the prediction of the Energy requires temperature in the kitchen area, the temperature outside and visibility.

c. Interaction between variables

The choquel integral the fuzzy measure demonstrates the good and bad variables. The $V(\{1,2,3,4\})$ and $V(\{1,3,4\})$ are complementary since equal to one. The $V(\{2\})$ is not the best representation.

$V(\{1,2\})$ - redundant	$V(\{1,2,4\})$ – redundant
$V(\{1,3\})$ - redundant	$V(\{1,3,4\})$ - complementary
$V(\{1,4\})$ – redundant	$V(\{2,3,4\})$ – redundant
$V(\{1,2,3\})$ - redundant	$V(\{1,2,3,4\})$ – complementary

d. Better model is favour a higher output.

Table 3. Shapley values

1	0.510215876076444	6	0.851078618772128	11	0.770622306991294
2	0	7	0.851078618771404	12	0.973065289002763
3	0.510215876076444	8	0.386651764298969	13	0.973065289001921
4	0.555969128314925	9	0.770622306991294	14	0.973065289000202
5	0.851078618772318	10	0.386651764298969	15	0.999999999996903

Shapley value shows the importance of each variable.

4. Use your model for prediction

The best fitting model is Choquet integral because it has the lowest RMSE, Av. abs error and a high correlation.

Choquet Integral

The new input: X1=17; X2=39; X3=4; X4=77; X5=32

Transformation of X1, X2, X3, X5:

```
new_row <- c(17,39,4,77,32, 10)
my.data <- rbind(my.data, new_row)
View(my.data):
```

X.349	0.4718743	0.45359013	0.18474910	0.46171372	0.45386000
new_row	0.7607652	0.43854125	0.26392291	0.48727220	1.00000000


```
> choquet(c(0.264, 0.761),c(0.245, 0.116, 0.410, 0.228))
[1] 0.776407
```

Data transformation:

After transformation of the data, the variable of interest $Y = 80$ Wh.

Comments on the result:

This result is expected based on the existing dataset. There are similar results in my.data.txt like number 163 and 261, which have similar parameters.

The highest Energy use of appliances will occur when:

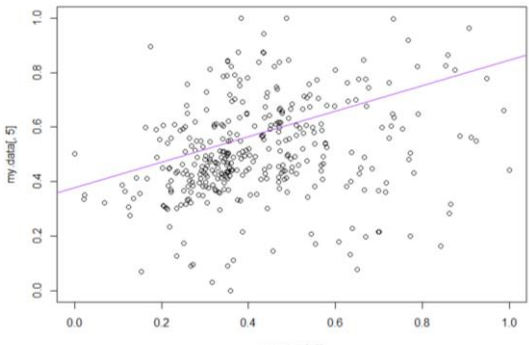
X1: Temperature in kitchen area is 15.064 °C

X2: Humidity in kitchen area is 56.581%

X3: Temperature outside (from weather station) is 0.044737 °C

X5: Visibility (from weather station) is 15 km

5. Comparing with a linear regression model

<pre>Call: lm(formula = your.data[, 5] ~ your.data[, 3]) Residuals: Min 1Q Median 3Q Max -0.48917 -0.10610 0.01361 0.11283 0.48387 Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 0.37790 0.02399 15.755 <2e-16 *** your.data[, 3] 0.46728 0.05216 8.959 <2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 0.1743 on 348 degrees of freedom Multiple R-squared: 0.1874, Adjusted R-squared: 0.1851 F-statistic: 80.27 on 1 and 348 DF, p-value: < 2.2e-16</pre>	<p>(i) The summary statistics:</p> <p>Based on residuals the distance from the data to fitted line can be called ideal since $-0.48917 \approx 0.48387$. The 1Q and 3Q are close to each other as well. The median is acceptable since it close to zero.</p> <p>There is 80% of size variation, which is good.</p>
	<p>(ii) Comparison with Choquel model</p> <p>Multiple R-squared in Linear regression is 0.1874. Square error in Choquel integral is 0.1762. The founded numbers are close.</p>

References:

1. S. James, 2010, The Use of Aggregation Functions in Decision Making, Deakin University, pp. 17-74, <<https://dro.deakin.edu.au/eserv/DU:30035914/james-useofaggregation-2010.pdf>>.
2. S. James, 2016, An Introduction to Data Analysis using Aggregation Functions in R, Springer International Publishing, pp. 42-125, <<https://link.springer.com/book/10.1007%2F978-3-319-46762-7>>.
3. 'StatQuest: Linear Regression in R', 2017, YouTube, StatQuest with Josh Starmer, 25 July, retrieved 10 May 2020, <https://www.youtube.com/watch?v=u1cc1r_Y7M0>.