

ASSESSMENT DETAILS

Plagiarism occurs when a student passes off as the student's own work, or copies without acknowledgement as to its authorship, the work of any other person or resubmits their own work from a previous assessment task.

Collusion occurs when a student obtains the agreement of another person for a fraudulent purpose, with the intent of obtaining an advantage in submitting an assignment or other work.

Work submitted may be reproduced and/or communicated by the university for the purpose of assuring academic integrity of submissions: <https://www.deakin.edu.au/students/study-support/referencing/academic-integrity>

Using aggregation functions for data analysis

The provided zip file contains the data file [*Energy20.txt*] and the R code [*AggWaFit718.R*] to use with the following tasks, include these in your R working directory.

Total Marks 100, Weighting 20%

Energy Prediction of Domestic Appliances Dataset

The given dataset, "Energy20.txt", can be used to create models of energy use of appliances in a energy-efficient house. The dataset provides the Energy use of appliances (denoted as Y) using 671 samples. It is a modified version of data used in the study [1]. The dataset includes 5 variables, denoted as X1, X2, X3, X4, X5, and Y, described as follows:

X1: Temperature in kitchen area, in Celsius

X2: Humidity in kitchen area, given as a percentage

X3: Temperature outside (from weather station), in Celsius

X4: Humidity outside (from weather station), given as a percentage

X5: Visibility (from weather station), in km

Y: Energy use of appliances, in Wh

Assignment Tasks

1. Understand the data [20 marks]

- (i) Download the txt file (Energy20.txt) from Future Learn and save it to your R working directory.
- (ii) Assign the data to a matrix, e.g. using
`the.data <- as.matrix(read.table("Energy20.txt"))`
- (iii) The *variable of interest* is Energy use of appliances (Y). To investigate Y, generate a subset of 350 data, e.g. using:

```
my.data <- the.data[sample(1:671,350),c(1:6)]
```

- (iv) 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. Include 5 scatter plots, 6 histograms, and 1 or 2 sentences for each of the variables, including the variable of interest Y.

2. Transform the data [10 marks]

- (i) Choose any **four** from the five variables (X1, X2,...,X5). Make appropriate transformations to the chosen four variables and the variable of interest Y so that the values can be aggregated in order to predict the *variable of interest*. Assign your *transformed* data along with your *transformed* variable of interest to an array (it should be 350 rows and 5 columns). Save it to a txt file titled "name-transformed.txt" using
- ```
write.table(your.data,"name-transformed.txt")
```
- where "name" is replaced with your name - you can use your surname or first name.
- (ii) Briefly explain the transformations applied for the selected four variables and the *variable of interest*. (1- 2 sentences each)

## 3. Build models and investigate the importance of each variable [30 marks]

- (i) Download the AggWaFit718.R file to your working directory and load into the R workspace using,
- ```
source("AggWaFit718.R")
```
- (ii) Use the fitting functions to learn the parameters for
- A weighted arithmetic mean (WAM)
 - Weighted power means (WPM) with $p = 0.5$, and $p = 5$,
 - An ordered weighted averaging function (OWA), and
 - A Choquet integral.
- (iii) Include two tables in your report - one with the error measures and correlation coefficients, and one summarising the weights/parameters and any other useful information learned for your data.
- (iv) Compare and interpret the data in your tables. Comment on
- a. How good the model is,
 - b. The importance of each of the variables (the four variables that you have selected),
 - c. Any interaction between any of those variables (are they complementary or redundant?) and
 - d. Better models favour higher or lower inputs.
- (1-3 paragraphs for part 3(iv))

4. Use your model for prediction [20 marks]

- (i) Choose your best fitting model.
Using your best fitting model, predict the Energy use of appliances for the following input $X_1=17$; $X_2=39$; $X_3=4$; $X_4=77$; $X_5=32$.
- (ii) Give your result and comment on whether you think it is reasonable. (1-2 sentences).
- (iii) Comment on the best conditions (in terms of your chosen four variables) under which a high Energy use of appliances will occur. (1-2 sentences).

5. Comparing with a linear regression model [20 marks]

Linear regression is used to predict the value of an outcome variable Y based on one or more input predictor variables X . The equation is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$. The built-in function `lm()` is used to fit linear models in R.

- (i) Build your linear model using the same dataset in Question 3 and describe the summary statistics for your model using the function `summary()`.
- (ii) Compare the performance of the linear model you got with your best fitting model in Question 4. Visualise the predicted Y values of both models on the 350 data and compare them with the true Y values.
- (iii) Give your comment on the differences between the linear model and your best fitting model. (2-4 sentences).

All supporting information should be presented in the pdf report. It will be assessed for style and grammar, professional presentation of figures, tables and references. List and quote in the text the references used, including books, articles and web resources.

Use the Harvard style: <https://www.deakin.edu.au/students/studying/study-support/referencing/harvard>

Submit to the **SIT718 CloudDeakin Dropbox**. Your final submission must include the following **three** files:

1. A report, "**name-report.pdf**", in pdf format (created in any word processor), covering all of the items in above (where "name" is replaced with your name -you can use your surname or first name). The total report must be **up to 8 pages**.
2. A data file named "**name-transformed.txt**" - just to help us distinguish them!).
3. The R code file (that you have written to produce your results) named "**name-code.R**" (where "name" is replaced with your name - you can use your surname or first name).

References:

1. Luis M. Candanedo, Veronique Feldheim, Dominique Deramaix. Data driven prediction models of energy use of appliances in a low-energy house, Energy and Buildings, Volume 140, 1 April 2017, Pages 81-97, ISSN 0378-7788.
<http://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction>