ASSESSMENT TASK 2 (PROBLEM SOLVING) in 2024T3

Using aggregation functions for data analysis

The provided zip file contains the data file [*ENB.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 Appliances Dataset

The Dataset for this assignment is modified version of a subset of data used in Candanedo et al, 2017.

The experimental data have been used to create models of energy use of appliances in a low-energy house.

The modified Dataset provides the energy use of Appliances (denoted as Y).

The Dataset comprises 5 features (variables), which are denoted as X1, X2, X3, X4 and X5.

The details about these variables are given below:

X1: Temperature in living room area (Celsius degrees)

X2: Humidity in living room area (percentage)

X3: Temperature in office room (Celsius degrees)

X4: Humidity in office room (percentage)

X5: Pressure (millimeter of mercury)

Y: Appliances energy consumption (Wh)

For more information about the variables see Candanedo et al, 2017.

Assignment tasks

T1. Understand the data

(i) Download the txt file (ENB.txt) from CloudDeakin and save it to your R working directory.

(ii) Assign the data to a matrix, e.g. using

```
the.data <- as.matrix(read.table("ENB.txt"))
```

(iii) The variable of interest is **Y**. To investigate **Y**, generate a subset of num_row=450 (use the same setting for the following tasks as well) with numerical data e.g. using:

```
my.data <- the.data[sample(1:num_samples,num_row) c(1:num_col)]
```

This would give you a new dataset with num_row rows and num_col columns. Values of num_sample and num_col have to be determined from the data provided.

(iv)Use scatter plots and histograms to understand the relationship between each of the variables **X1 X2, X3, X4, X5,** and your variable of interest **Y**, i.e., scatter plots of (**X1, Y**), (**X2, Y**), …, (**X5, Y**), and histograms of **X1 X2, X3, X4, X5, Y**.

T2. Transform the data

Choose **any FOUR** variables from **X1**, **X2, X3, X4, X5.**

Make appropriate transformations so that the values can be aggregated in order to predict the *variable of interest* **Y**.

Assign your *transformed* data along with your *transformed* variable of interest to an array (it should be ``num_row'' rows and 5 columns). Save it to a txt file titled "name-transformed.txt".

```
write.table(your.data,"name-transformed.txt")
```

The following tasks are based on the saved transformed data.

T3. Build models and investigate the importance of each variable.

(i) Download the AggWaFit.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. A weighted arithmetic mean (WAM),

b. Weighted power means (WPM) with $p = 0.5$,

c. Weighted power means (WPM) with $p = 2$,

d. An ordered weighted averaging function (OWA).

T4. Use your model for prediction.

Using your best fitting model from T3, i.e., WAM, WPM(0.5), WPM(2), or OWA, predict **Y** (Appliances) for the following inputs:

X1= 19.1, X2=43.29, X3=19.7, X4=43.4, X5=743.6

You should use the same pre-processing as in Task 2.

Compare your prediction with the measured Y=60.

T5. Summarise your data analysis in up to 20 slides for a 5-minute video presentation

The slides should include the following content:

- **Correlations between the variables**;
- What kinds of **data distributions** you have identified in the raw data, use the histograms you have produced;
- List and explain the **transformations** applied for the selected four variables and the variable of interest;
- Explain the **importance of the variables** you have selected;
- The **best fitting model** on your selected data; include two tables:
  one with the error measures and correlation coefficients, and one summarizing the weights/parameters and any other useful information learned for your data;
- Your **prediction result** and comment on wheather you think it is reasonable;
- Discuss the **conditions** (in terms of your chosen variables) under which low energy use of

appliances will occur.

- Comment on the **implications and limitations** of the fitting model you used for prediction.

The slides should contain all necessary information to prove your findings. All the **bold** terms above must appear in slide titles. Explanations and reasoning can be given verbally or in a written format.

For the 5-minute video presentation, you may provide a link to YouTube or upload a mp4 video.

SUBMISSION:

Submit to the **SIT718 CloudDeakin Dropbox**.

Your submissions must contain the following **Three** files (pay attention to file types):

1. The presentation slides, "name-slides.pdf", covering all of the items in above
 (where “name” is replaced with your name -you can use your surname or first name);
2. The 5-min video based on the slides (a link to YouTube or uploading a mp4 file): ``name-video.mp4’’
3. The R code file (that you have written to produce your results) named "name-code.R"
(where “name” is replaced with your surname or first name).

Additional Rules and Clarification:

* Showing your face in the video is not required (you could if you want).

* Any content beyond 5 minutes or exceeding 20 slides will not be graded.

* To receive marks for Task 3, you must follow the provided source code "**AggWaFit718.R**". Using alternative methods outside those taught in SIT718 will not be awarded any marks.

* Apply set.seed() with your student ID as the seed so that we can reproduce your results; otherwise, 15 marks will be deduced from your final score.

* 15 marks will be deduced from your final score if the pdf file for the presentation slides is missing.

* Your assignment will not be assessed (zero mark for this assessment) if any of the following conditions occurs:

** The R code is missing (other codes are not allowed, such as .RMD, .RData, .Rproj and .ipynb)

** The outputs of the code are inconsistent with the content of the video/slides

** Academic misconduct is substantiated by Academic Integrity Committee.

* For **referencing**, follow the Harvard style:

https://www.deakin.edu.au/students/studying/study-support/referencing/harvard

You **must cite** all the datasets, packages and literature you used for this assessment.

You will loose 5 marks for lack of or inappropriate citations/references.

References

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.

The original data are available in:

http://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction