ASSESSMENT TASK 2 (PROBLEM SOLVING)

Using aggregation functions for data analysis

The provided zip file contains the data file [*ENB_2022.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 Appliences 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) using 671 samples.

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 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: Appliances, energy use, in Wh

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

Assignment tasks

T1. Understand the data

(i) Download the txt file (ENB_2022.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_2022.txt"))
```

(iii) The variable of interest is **Y** (Appliences). To investigate **Y**, generate a subset of 350 with numerical data e.g. using:

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

This would give you a new dataset with 350 rows and 6 columns.

The following tasks are based on the 350 sample data.

(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**. (You should build 5 scatter plots and 6 histograms).

T2. Transform the data

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

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

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".

```
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 (from CloudDeakin) 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).

You can also use the Choquet integral - this is **Optional**.

T4. Use your model for prediction.

Using your best fitting model from T3, predict **Y** (the area) for the following input

X1=22; **X2**=38; **X3**=4; **X4**=88, **X5**=35.

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

Compare your prediction with the measured value of **Y**, Y=110.

T5. Summarise your data analysis in up to **20 slides** for a **5-minutes** 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;
- 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;
- Explain the importance of each of the variables (the four variables that you have selected);
- The best fitting model on your selected data;
- Your prediction result and comment on wheather you think it is reasonable;
- Discuss the best conditions (in terms of your chosen **FOUR variables**) under which a low energy use of Appliences will occur.
- Comment on the implications and the limitations of the fitting model you used for prediction.

The slides should contain all necessary information to prove your findings.

For the 5-minutes presentation, use a simple and accessible platform such as YouTube or PowerPoint Audio.

SUBMISSION:

Submit to the **SIT718 CloudDeakin Dropbox**.

Your final submission must include the following **TWO** files:

1. The presentation slides with audio, "**name-slides**" (pdf, pptx), covering all of the items in above (where "name" is replaced with your name -you can use your surname or first name) (a link to YouTube/Dropbox is acceptable).
2. 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).

Your assignment will not be assessed if the code is missing, or the outputs of the code are inconsistent with the content of the slides.

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 some 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