ASSESSMENT TASK 2 (PROBLEM SOLVING)

Using aggregation functions for data analysis

The provided zip file contains the data file [*Forest_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%

Forest Fires Dataset

The given dataset, **Forest_2022.txt**, is to model the burned area as a result of forest fires based on various metereological and other data. It is a **modified** version of an original dataset, used to predict the burned area of forest fires in the northeast region of Portugal (Cortes and Morais, 2009). The original dataset can be found in "UCI Machine Learning Repository: Forest Fires Data Set", 2017.

The modified dataset, **Forest_2022.txt**, contains the following numerical variables:

X1: x-axis special coordinate within map (1 to 9)

X2: y-axis special coordinate within map (2 to 9)

X3: month – month of the year (Jan=1 to Dec=12)

X4: day – day of the week (Mon=1 to Sun=7)

X5: FFMC - FFMC index from the FWI system

X6: DMC - DMC index from the FWI system

X7: DC - DC index from the FWI system

X8: ISI - ISI index from the FWI system

X9: temp - temperature in Celsius degrees

X10: RH - relative humidity in %

X11: wind - wind speed in km/h

X12: rain - outside rain in mm/m2

Y: area - the burned area of the forest (in ha)

For more information about the variables see (Cortez and Morais, 2007).

Assignment tasks

T1. Understand the data

(i) Download the txt file (Forest_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("Forest_2022.txt"))
```

(iii) The variable of interest is **Y** (area). To investigate **Y**, generate a subset of 330 with numerical data e.g. using:

```
my.data <- the.data[sample(1:517,330), c(1:13)]
```

This would give you a new dataset with 330 rows and 13 columns.

The following tasks are based on the 330 sample data.

(iv)Use scatter plots and histograms to understand the relationship between each of the variables **X5, X6, X7, X8, X9, X10, X11, X12,** and your variable of interest **Y**. (You should build 8 scatter plots and 9 histograms).

T2. Transform the data

Choose **any FOUR** variables from **X5**, **X6, X7, X8, X9, X10, X11, X12.**

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

Assign your *transformed* data along with your *transformed* variable of interest to an array (it should be 330 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

X5=96.1; **X6**=181.1; **X7**=671.2; **X8**=14.3; **X9**=20.7; **X10**=69; **X11**=4.9; **X12**=0.4

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

Compare your prediction with the measured value of **Y**, Y=0.0146.

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 large burned area 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

Cortez, P. and Morais, A. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 1 3th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9, available at http://www3.dsi.uminho.pt/pcortez/fires.pdf.

"UCI Machine Learning Repository: Forest Fires Data Set". *Archive.ics.uci.edu*. N.p., 2017, http://archive.ics.uci.edu/ml/datasets/forest+fires.