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 [*RedWine.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%

Red wine quality Dataset
The given dataset, "RedWine.txt", is used to model wine quality based on physicochemical tests. The dataset provides the 1,599 red wine samples from the north of Portugal. It is a modified version of the data used in the study [1]. This dataset includes 5 variables, denoted as X1, X2, X3, X4, X5, and Y, described as follows:

X1 - citric acid
X2 - chlorides
X3 - total sulfur dioxide
X4 - pH
X5 – alcohol
Y - quality (score between 0 and 10)

Assignment Tasks

***** *Q4 and Q5 are for students who are aiming for HD.*

1. Understand the data [10 marks]

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

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

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

(iii) The ***variable of interest*** is quality (Y). To investigate Y, generate a subset of 500 data, e.g. using:

```
my.data <- the.data[sample(1:1599,500),c(1:6)]
```

References:

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

(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 individually, 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 500 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. [All the following tasks are based on the saved transformed data]

(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 $p = 0.1$, and $p = 8$ [define your own generator]
- 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 [15 marks]

(i) Choose your best fitting model based on Q3(iv). Using your best fitting model, predict the wine quality for the following input X1=0; X2= 0.075; X3=41; X4=3.53; X5=9.3. [Use the same pre-process as Q2]

(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 higher quality wine will occur. (1-2 sentences).

5. Comparing with a linear regression model [15 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 $\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n + \varepsilon}$. 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 500 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).

6. Summarising your data analysis in a 3-minutes presentation [20 marks]

Using a simple and accessible platform such as YouTube or PowerPoint, create a 3 minute presentation that summarises your data analysis procedures, findings, implications, and the limitations of the model you used.

Following 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 **FOUR** files:

1. A PDF report, "**name-report.pdf**", 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** (for everything) including a cover page which contains your full name and student ID.
2. A data file named "**name-transformed.txt**" - just to help us distinguish them!).
3. Presentation recordings or slides with audio (a link to YouTube/Dropbox is acceptable)
4. 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).

The assessment will test your knowledge and your programming skills, so you need to provide sufficient and clear details for each question to award the full marks. We cannot evaluate the submission without the R-code or if the R-code is irrelevant or not working.