# ASSESSMENT DETAILS

# SIT718 Real World Analytics

## Assessment Task 3: Problem solving task 2



## Using aggregation functions for data analysis

This document supplies detailed information on assessment tasks for this unit.

#### **Key information**

Due: 9<sup>th</sup> January 2019 11.30pm AEDT

• Weighting: 30%

Reference style: Harvard

#### **Learning Outcomes**

This assessment assesses the following Unit Learning Outcomes (ULO) and related Graduate Learning Outcomes (GLO):

Unit Learning Outcome (ULO)	Graduate Learning Outcome (GLO)
<b>ULO1</b> – assessed through student ability to apply knowledge of multivariate functions, data transformations and data distributions to summarise data sets.	<b>GLO1</b> - Discipline knowledge and capabilities
	GLO4 – Critical thinking
	GLO5 – Problem solving
<b>ULO2</b> – assessed through the student ability to analyse datasets by interpreting summary statistics, model and function parameters.	
<b>ULO4</b> - assessed through student ability to develop software codes to solve computational problems for real world analytics.	

#### **Purpose**

This assignment will test your knowledge and understanding of the aggregation functions and their applications for data summarization and prediction. This assignment will also test your ability in R programming, in using specific R commands as well as R packages.

#### Instructions

The work is individual. Solutions and answers to the assignment must be explained carefully in a concise manner and presented carefully. Use of books, articles and/or online resources on share price related to SIT718 Real World Analytics is allowed. Students are expected to refer to the suitable literature where appropriate.

The assessment consists of **FOUR tasks**. Students must attempt all tasks and provide an individual written report in appropriate word processor.

The detailed problem description and data set will be released to students on Wednesday 5<sup>th</sup> December 2018.

#### **Submission details**

**No more than 7 A4 sides**, including Figures, Tables, Appendices and References. The report should be typed. Use minimal font 11pt and 2.5cm side margins. If the page limit is exceeded only the first 7 pages will be marked.

Assignment (a report in pdf format, software code and/or data) must be submitted via the assignment dropbox in the unit site (accessed in DeakinSync)

No e-mail or hardcopy submissions are accepted.

# ASSESSMENT DETAILS SIT718 Real World Analytics



#### Assessment Task 3: Problem solving task 2

#### **Extension requests**

Requests for extensions should be made to Unit/Campus Chairs well in advance of the assessment due date. If you wish to seek an extension for an assignment, you will need to apply by email directly to **Prof. Maia Angelova**(maia.a@deakin.edu.au), as soon as you become aware that you will have difficulty in meeting the scheduled deadline, but at least 3 days before the due date. When you make your request, you must include appropriate documentation (medical certificate, death notice) and a copy of your draft assignment.

Conditions under which an extension will normally be approved include:

**Medical** To cover medical conditions of a serious nature, e.g. hospitalisation, serious injury or chronic illness. Note: Temporary minor ailments such as headaches, colds and minor gastric upsets are not serious medical conditions and are unlikely to be accepted. However, serious cases of these may be considered.

Compassionate e.g. death of close family member, significant family and relationship problems.

**Hardship/Trauma** e.g. sudden loss or gain of employment, severe disruption to domestic arrangements, victim of crime. Note: Misreading the timetable, exam anxiety or returning home will not be accepted as grounds for consideration.

## Special consideration

You may be eligible for special consideration if circumstances beyond your control prevent you from undertaking or completing an assessment task at the scheduled time.

See the following link for advice on the application process:

http://www.deakin.edu.au/students/studying/assessment-and-results/special-consideration

#### Assessment feedback

Students will receive written feedback and model solutions to aid reflection and analysis of problem strategies and solutions for consideration in the upcoming problem-solving task.

#### Referencing

You must correctly use the Harvard method in this assessment. See the Deakin referencing guide.

#### Academic integrity, plagiarism and collusion

Plagiarism and collusion constitute extremely serious breaches of academic integrity. They are forms of cheating, and severe penalties are associated with them, including cancellation of marks for a specific assignment, for a specific unit or even exclusion from the course. If you are ever in doubt about how to properly use and cite a source of information refer to the referencing site above.

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: <a href="https://www.deakin.edu.au/students/study-support/referencing/academic-integrity">https://www.deakin.edu.au/students/study-support/referencing/academic-integrity</a>

# ASSESSMENT DETAILS

# **SIT718 Real World Analytics**

Assessment Task 3: Problem solving task 2

Using aggregation functions for data analysis

#### Forest Fires Data Set

In order to predict the burned area of forest fires ("UCI Machine Learning Repository: Forest Fires Data Set", 2017), in the northeast region of Portugal ("Montesinho.Com - Nature Tourism In Montesinho Natural Park", 2017), analysis of the meteorological and other data is required (see details at "Forest Fires Dataset", 2017), also consider the information given in http://cwfis.cfs.nrcan.gc.ca/background/summary/fwi . For this assignment you are provided with a modified dataset "Forest718.txt".

## **Attribute Information:**

X1: x-axis spatial coordinate within the Montesinho park map: 1 to 9 ("Montesinho.Com - Nature Tourism In Montesinho Natural Park", 2017)

X2: y-axis spatial coordinate within the Montesinho park map: 2 to 9 ("Montesinho.Com - Nature Tourism In Montesinho Natural Park", 2017)

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: 18.7 to 96.20 (Happe, 2017)

X6: DMC - DMC index from the FWI system: 1.1 to 291.3 (Happe, 2017)

X7: DC - DC index from the FWI system: 7.9 to 860.6 (Happe, 2017)

X8: ISI - ISI index from the FWI system: 0.0 to 56.10 (Happe, 2017)

X9: temp - temperature in Celsius degrees: 2.2 to 33.30

**X10: RH** - relative humidity in %: 15.0 to 100

**X11: wind** - wind speed in km/h: 0.40 to 9.40

X12: rain - outside rain in mm/m2 : 0.0 to 6.4

X13=Y: area - the burned area of the forest (in ha): 0.00 to 1090.84

# Assignment tasks

## 1. Understand the data

[20 marks]

- (i) Download the txt file (Forest718.txt) from Future Learn and save it to your R working directory [2 marks]
- (ii) Assign the data to a matrix, e.g. using

the.data <- as.matrix(read.table("Forest718.txt"))

Your variable of interest is X13=Y: area - the burned area of the forest (in ha): 0.00 to 1090.84 (the thirteenth column in the dataset). Generate a subset of 200 data e.g. using:

my.data <- the.data[sample(1:517,200),c(1:13)]

[3 marks]

(iii) Choose any FOUR variables from X5 to X11. Using scatter plots and histograms, report on the general relationship between each of the variables and your variable of interest Y. Include 4 scatter plots, 5 histograms and 1 or 2 sentences for each of the variables

[15 marks]

#### 2. Transform the data

[20 marks]

(i) For the chosen four variables and the variable of interest Y make appropriate transformations so that the values can be aggregated in order to predict the variable of interest (the area). Assign your transformed data along with your transformed variable of interest X13=Y to an array (it should be 200 rows and 5 columns). Save it to a txt file titled "name-transformed.txt".

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

[10 marks]

(iii) Briefly explain the general relationship between each of your transformed variables and your variable of interest (the area). (2-3 sentences each)

[10 marks]

# 3. Build models and investigate the importance of each variable

[35 marks]

(i) Download the AggWaFit.R file (from CloudDeakin) to your working directory and load into the R workspace using,

source ("AggWaFit718.R")

[5 marks]

(ii) Using the **fitting** functions to learn the parameters for:

- A weighted arithmetic mean,
- Weighted power means with p = 0.5, and p = 2,
- An ordered weighted averaging function, and
- A Choquet integral.

[10 marks]

(iii) Include two tables in your report - one on the error measures, and one summarising the weights/parameters that were learned for your data.

[10 marks]

- (iv) Compare and interpret the data in your tables. Be sure to 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)

[10 marks]

# 4. Use your model for prediction

[25 marks]

- (i) Using your best fitting model, predict the area for the following input:
  - X5=91.6; X6=181.3; X7=613; X8=7.6; X9=24.6; X10=44; X11=4; X12=0.
- (ii) Give your result and comment on whether you think it is reasonable. (1-2) sentences) [15 marks]
- (iii) Comment generally on the ideal conditions (in terms of your chosen four variables) under which an area will result. (1-2 sentences) [10 marks]

Your final submission, which should be submitted to the SIT718 CloudDeakin Dropbox, should include the following three files. Please follow the instructions below and do not compress your files.

- 1. A "name-report.pdf" report (created in any word processor), covering all of the items in above (items coloured blue usually have explicit instructions about what should be included). With plots and tables it should only be 3 5 pages.
- 2. A data file named "name-transformed.txt" (where `name' is replaced with your name you can use your surname or first name just to help me 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

"UCI Machine Learning Repository: Forest Fires Data Set". Archive.ics.uci.edu. N.p., 2017. Web. (<a href="http://archive.ics.uci.edu/ml/datasets/Forest+Fires">http://archive.ics.uci.edu/ml/datasets/Forest+Fires</a>), 29 Apr. 2017.

"Forest Fires Dataset". Dsi.uminho.pt. N.p., 2017. Web. (<u>www.dsi.uminho.pt/~pcortez/forestfires</u>), 29 Apr. 2017.

Cortez, P. and Morais, A.D.J.R., 2007. A data mining approach to predict forest fires using meteorological data (<a href="http://www3.dsi.uminho.pt/pcortez/fires.pdf">http://www3.dsi.uminho.pt/pcortez/fires.pdf</a>).

Happe, Harry. "Meteomalaga". <a href="https://Malagaweather.com">https://Malagaweather.com</a>. N.p., 2017. Web. 29 Apr. 2017.

"Montesinho.Com - Nature Tourism In Montesinho Natural Park". montesinho.com. N.p., 2017 (https://www.montesinho.com/en), 29 Apr. 2017.

# **Appendix - FWI System**

(Happe, 2017)

The FWI is based on weather readings taken at noon standard time and rates fire danger at the mid-afternoon peak from 2:00 - 4:00 pm. Weather readings required are:

- Air temperature (in the shade)
- Relative Humidity (in the shade)
- Wind speed (at 10 metres above ground level for an average over 10 minutes)
- Rainfall (For the previous 24 hours)

The Fire Weather Index has six components:

Three Fuel Moisture Codes are:

- 1. Fine Fuel Moisture Code
- 2. Duff Moisture Code
- 3. Drought Code

Three Fire Behaviour Indices

- 1. Initial Spread index
- 2. Build Up Index
- 3. Fire Weather Index

#### Three Fuel Moisture Codes

The FWI System evaluates fuel moisture content and relative fire behaviour using past and present weather effects on ground level fuels. The moisture codes reflect the net effects of daily moisture gains and losses.

## Fine Fuel Moisture Code - FFMC

This is a numerical rating of the moisture content of surface litter and other cured fine fuels. It shows the relative ease of ignition and flammability of fine fuels. The moisture content of fine fuels is very sensitive to the weather. Even a day of rain, or of fine and windy weather, will significantly affect the FFMC rating. The system uses a time lag of two-thirds of a day to accurately measure the moisture content in fine fuels. The FFMC rating is on a scale of 0 to 99. Any figure above 70 is high, and above 90 is extreme.

#### **Duff Moisture Code - DMC**

DMC is a numerical rating of the average moisture content of loosely compacted organic layers of moderate depth. The code indicates the depth that fire will burn in moderate duff layers and medium size woody material. Duff layers take longer than surface fuels to dry out but weather conditions over the past couple of weeks will significantly affect the DMC. The system applies a time lag of 12 days to calculate the DMC. A DMC rating of more than 30 is dry, and above 40 indicates that intensive burning will occur in the duff and medium fuels. Burning off operations should not be carried out when the DMC rating is above 40.

#### Drought Code - DC

The DC is a numerical rating of the moisture content of deep, compact, organic layers. It is a useful indicator of seasonal drought and shows the likelihood of fire involving the deep duff layers and large logs. A long period of dry weather (the system uses 52 days) is needed to dry out these fuels and affect the Drought Code. A DC rating of 200 is high, and 300 or more is extreme indicating that fire will involve deep sub-surface and heavy fuels. Burning off should not be permitted when the DC rating is above 300.

## Fire Behaviour Indices

The three behaviour indices are relative to the fuel moisture content. They indicate what a fire is likely to do. The lower the moisture content, the higher the Fuel Moisture Codes, and the higher the Fire Behaviour Indices – and the more active the fire will be.

# Initial Spread Index - ISI

This indicates the rate fire will spread in its early stages. It is calculated from the FFMC rating and the wind factor. The open-ended ISI scale starts at zero and a rating of 10 indicates high rate of spread shortly after ignition. A rating of 16 or more indicates extremely rapid rate of spread.

#### Build -Up Index - BUI

This index shows the amount of fuel available for combustion, indicating how the fire will develop after initial spread. It is calculated from the Duff Moisture Code and the Drought Code. The BUI scale starts at zero and is open-ended. A rating above 40 is high, above 60 is extreme.

#### Fire Weather index - FWI

Information from the ISI and BUI is combined to provide a numerical rating of fire intensity – the Fire Weather Index. The FWI indicates the likely intensity of a fire.

The FWI is divided into four fire danger classes:

Low 0 - 7 Medium 8 - 16 High I7 - 31 Extreme 32+