SIT743 Multivariate and Categorical Data Analysis
Assignment-1

Total Marks = 120, Weighting - 25%

Due date: 14 April 2019 by 11.30 PM

INSTRUCTIONS:

- For this assignment, you need to submit the following **THREE** files.

  1. **A written document** (*A **single pdf only***) covering all of the items described in the questions. All answers to the questions must be written in this document, i.e, **not** in the other files (code files) that you will be submitting. ***All the relevant results (outputs, figures) obtained by executing your R code must be included in this document.***
     *For questions that involve mathematical formulas, you may write the answers manually (hand written answers), **scan it to pdf** and **combine with your answer document**. Submit a combined **single pdf** of your answer document.*

  2. A **separate** ".R" file or '.txt' file containing your code (R-code script) that you implemented to produce the results. Name the file as "*name-StudentID-Ass1-Code.R*" (where `*name*' is replaced with your name - you can use your surname or first name, and *StudentID* with your student ID).

  3. **A data file** named "*name-StudentID-GBRMyData.txt*" (where `*name*' is replaced with your name - you can use your surname or first name, and *StudentID* with your student ID).

- All the documents and files should be submitted (uploaded) via *SIT 743 Clouddeakin Assignment Dropbox* by the due date and time.
- **Zip files are NOT accepted**. All three files should be uploaded **separately** to the CloudDeakin.
- E-mail or manual submissions are **NOT** allowed. Photos of the document are **NOT** allowed.
- The questions Q2, Q3 and Q4 **do not** require any R programming.

==========

Some of the questions in this assignment require you to use the "**Great Barrier Reef (GBR)**" dataset. This dataset is given as a CSV file, named "GBRData.csv". You can download this from the Assignment folder in CloudDeakin. Below is the description of this dataset.

Great Barrier Reef data (GBR):
This dataset gives the sea water measurements collected at a depth of around 5m below the sea surface at two islands in the Great Barrier Reef region (North Queensland, Australia), which are 71 km apart, namely ***Heron Island (HI)*** and ***Lady Musgrave Island*** **(LMI).**
[http://apps.aims.gov.au/reef-monitoring/reef/23082S ,
http://weather.aims.gov.au/#/station/130].

The data gives daily measurements collected over a 20 month period between January 2017 and August 2018.

The variables include the following (4 variables):

HeronIsland-Salinity: Salinity measurements expressed in units of *practical salinity units,* measured at Heron Island.

HeronIsland-Water Temperature: Water temperature (at 5m below the surface of the sea water) in degrees Celsius, measured at Heron Island.

HeronIsland-Water Pressure: Water pressure (at 5m below the sea surface) in decibar, measured at Heron Island.

LadyMusgrave-Water Temperature (degrees celsius): Water temperature (at 5m below the surface of the sea water) in degrees Celsius, measured at Lady Musgrave Island.

Q1) [18 Marks]:

- Download the GBR data file "GBRData.csv" and save it to your R working directory.
- Assign the data to a matrix, e.g. using

```
the.data <- as.matrix(read.csv("GBRData.csv", header = TRUE, sep = ","))
```

- Generate a sample of 250 data using the following:

```
my.data <- the.data [sample(1:556,250),c(1:4)]
```

Save "`my.data`" to a text file titled "*name-StudentID-GBRMyData.txt*" using the following R code (**NOTE: you must upload this text file with your submission**).

```
write.table(my.data,"name-StudentID-GBRMyData.txt")
```

Use the sampled data ("my.data") to answer the following questions.

1.1) Draw histograms for 'HeronIsland-Salinity' and 'HeronIsland-Water Temperature' values, and comment on them. [4 Marks]

1.2) Draw a parallel Box plot using the two variables; 'HeronIsland-Water Temperature' and the 'LadyMusgrave-Water Temperature'.
Find five number summaries of these two variables.
Use both five number summaries and the Boxplots to compare and comment on them. [6 Marks]

1.3) Draw a scatterplot of 'HeronIsland-Water Temperature' (as x) and 'LadyMusgrave-Water Temperature' (as y) for the ***first 150 data vectors selected from the "my.data"*** (name the axes).
Fit a linear regression model to the above two variables.
Plot the (regression) line on the same scatter plot.
Write down the linear regression equation.
Compute the correlation coefficient and the coefficient of Determination.
Explain what these results reveal. [8 Marks]

Q2) [19 Marks]

The table shows the results of a survey conducted about the type of occupation (in hundreds) and the age group in Melbourne during the year 2011.

		Occupation		
		Professionals (P)	Sales workers (S)	Community service (C)
Age group (in years)	Below 30 (B)	15	40	60
	30-50 (M)	100	10	10
	Above 50 (A)	28	4	3

Suppose we select a person at random,

2.1) What is the probability that the person is a Professional (P)? [1 mark]

2.2) What is the probability that the person is below 30 years old (B)? [1 mark]

2.3) What is the probability that the person's occupation is community service (C) and the age is between 30 and 50 years old (M)? [1 Mark]

2.4) What is the probability that the person is a sales worker (S) given that he/she is above 50 years old (A)? [2 Marks]

2.5) What is the probability that the person, who is a professional (P), is below 30 years old (B)? [2 Marks]

2.6) What is the probability that the person is a sales worker (S) or between 30 to 50 years old (M)? [3 Marks]

2.7) find the marginal distribution of the occupation [3 marks]

2.8) find the marginal distribution of the age group [3 marks]

2.9) Are the variables 'occupation' and 'Age group' independent random variables? Explain why or why not. [3 marks]

Q3) [6 Marks]

In a factory, drones (unmanned aerial vehicle) are produced from two types of assembly lines, called A and B. 70% of the drones are produced from assembly line A and 30% of them are from assembly line B. 60% of the drones produced by assembly line A passed the quality test, and 80% of the drones produced by assembly line B passed the quality test.

a) What is the overall proportion of the drones produced pass the quality test? [3 Marks]

b) If a randomly selected drone passed the quality test, what is the probability that it was produced by assembly line B? [3 Marks]

Q4) Maximum Likelihood Estimation (MLE) [16 Marks]

A data centre houses several computer servers to provide data storage solutions for companies working on big data. Let x_i $(x_i > 0)$ denote the "time-to-failure" of a computer server (in months), i.e., the time a server takes to fail completely. Assume that the time-to-failure of the server can be modelled using a special form of Weibull distribution with an unknown parameter θ $(\theta > 0)$ as given by the following equation.

$$\boldsymbol{x_i \sim spWeibull(\theta)}$$

$$\boldsymbol{spWeibull(\theta) = p(x_i|\theta) = 2\theta^2 x_i\, e^{-\theta^2 x_i^2}}$$

Assume that there are N servers used in the data centre, and the time-to-failure of the servers are independently and identically distributed (iid).

a) Show that the joint distribution of the time-to-failure of N servers can be given by the below equation. In other words, show that the expression for the likelihood distribution (joint distribution) $p(\boldsymbol{X}|\theta)$ of the time-to-failure of N servers $(\boldsymbol{X} = \{x_1, x_2, \dots, x_N\})$ is given by

$$\boldsymbol{p(X|\theta) = C\, 2^N \theta^{2N}\, e^{-S\theta^2}},$$

where $\boldsymbol{C = \prod_{i=1}^{N} x_i}$ and $\boldsymbol{S = \sum_{i=1}^{N} x_i^2}$

Hint: Since the time to failure of N servers are iid, the likelihood can be written as
$p(X|\theta) = p(x_1|\theta) \times p(x_2|\theta) \times p(x_3|\theta) \times \dots \times p(x_N|\theta)$

write down the equation for $p(x_1|\theta)$, $p(x_2|\theta)$, … $p(x_N|\theta)$ and compute the $p(X|\theta)$. Use the exponential Laws such as $a^m \times a^t = a^{m+t}$.

[6 marks]

b) Find an expression for the logliklihood function $\boldsymbol{L(\theta) = \ln(p(X|\theta))}$ **[3** marks**]**

c) In order to find the Maximum likelihood Estimation (MLE) of the parameter θ, you need to maximize the $L(\theta)$.
Find the value of θ that maximises $L(\theta)$ by differentiating the log likelihood function $L(\theta)$ with respect to θ and equating it to zero. Show that the Maximum likelihood Estimate $\hat{\theta}$ (MLE) of parameter θ is given by:

$$\hat{\theta} = \sqrt{\frac{N}{S}} \quad , \quad where \; S = \sum_{i=1}^{N} x_i^2$$

[5 Marks]

d) Suppose that the data center has experienced the following time-to-failures for five servers:
$x_1 = 20,\ x_2 = 15,\ x_3 = 12,\ x_4 = 40\ and\ x_5 = 35$.
What is the Maximum likelihood Estimate $\hat{\theta}$ (MLE) of parameter θ given this data? [2 Marks]

Q5) Bayesian inference for Gaussians (unknown mean and known variance) [21 marks]

5.1) What is the meaning of conjugate prior? [1 mark]

5.2) Why conjugate priors are useful in Bayesian statistics? [2 mark]

5.3) Give three examples of Conjugate pairs (i.e., give three pairs of distributions that can be used for prior and likelihood) [3 marks]

5.4) The annual snowfall received at Bright city are measured for n years. The **average** snowfall observed over the n years is 15 cm. Assume that the annual snowfall is **normally** distributed with unknown mean θ and known standard deviation 5 cm. Suppose your prior distribution for θ is **normal** with mean 10 cm and standard deviation 3 cm.

a) State the posterior distribution for θ (this will be in terms of n. Do not derive the formulae) [3 Marks]
b) For n=5, find the mean and the standard deviation of the posterior distribution. Comment on the posterior variance [3 Marks]
c) For n=15, find the mean and the standard deviation of the posterior distribution. Compare with the results obtained for n=5 in the above question Q5.4(b) and comment. [3 Marks]
d) Assume that the **prior** distribution is **changed**, and now the prior is distributed as **normal** with **mean 10cm** and **standard deviation 1cm**. For **n=10**, find the mean and the standard deviation of the posterior distribution.
Sketch, on a single coordinate axes, the prior, likelihood and the posterior distributions obtained (use different colures to show the distributions). Use R-program to plot these. [6 Marks]

Q6) C**lustering: [14 marks]**

6.1) ***K-Means clustering:*** Use the data file "SITEdata2019.txt" provided in CloudDeakin for this question. Load the file "SITEdata2019.txt" using the following:

```
zz<-read.table("SITEdata2019.txt")

zz<-as.matrix(zz)
```

a) Draw a scatter plot of the data. [1 mark].

b) State the number of classes/clusters that can be found in the data (by visual examination of the scatter plot) [1 marks].

c) Use the above number of classes as the k value and perform the k-means clustering on that data. Show the results using a scatterplot (show the different clusters with different colours). Comment on the clusters obtained. [4 Marks]

d) Vary the number of clusters (k value) from 2 to 20 in increments of 1 and perform the k-means clustering for the above data. Record the *total within sum of squares (TOTWSS)* value for each k, and plot a graph of TOTWSS verses k. Explain how you can use this graph to find the correct number of classes/clusters in the data. [3 marks]

6.2) ***Spectral Clustering:*** Use the same dataset (`zz`) and run a spectral clustering (use the number of clusters/centers as 5) on it. Show the results on a scatter plot (with colour coding). Compare these clusters with the clusters obtained using the k-means above and comment on the results. [5 Marks]

Q7) [26 Marks]

For this question you will be using "**Great Barrier Reef (GBR)**" dataset. This dataset is given as a CSV file, named "GBRData.csv".
You can download this dataset from the Assignment folder in CloudDeakin.

For this question, we consider only the data from one of the variables, namely **"LadyMusgrave-Water Temperature"** (called as 'LMWT') from this dataset.

You can use the following R code to load the data for LMWT variable

```
the.data <- as.matrix(read.csv("GBRData.csv", header =
                        TRUE, sep = ","))


#extract the 'LadyMusgrave-Water Temperature'  values

LMWTdata <- the.data[,4]
```

7.1) Provide a time series plot of the LMWT data (use the index as the time (x-axis)) using R code. [2 Marks]

7.2) Plot the histogram for LMWT data. Comment on the shape. How many ***modes*** can be observed in the data? [4 Marks]

7.3) Fit a **single Gaussian** model $\mathcal{N}(\mu, \sigma^2)$ to the distribution of the data, where μ is the **mean** and σ is the **standard deviation** of the Gaussian distribution.

Find the maximum likelihood estimate (MLE) of the parameters, i.e., the **mean** μ and the **standard deviation** (σ). You can use the following code to perform the fitting:

```
library(MASS)
fit1<-fitdistr(LMWTdata,"normal")
```

Plot the obtained density distribution.

[4 Marks]

7.4) Fit a **mixture of Gaussians** model to the distribution of the data using **the number of Gaussians equal to the number of modes** found in the data (in Q7.2 above) . Write the R code to perform this. Provide the **mixing coefficients**, **mean and standard deviation for each of the Gaussians** found. [5 Marks]

7.5) Plot these Gaussians on top of the histogram plot. Include a plot of the combined density distribution as well (use different colors for the density plots in the same graph). [4 Marks]

7.6) Provide a plot of the **log likelihood values** obtained over the iterations and comment on them. [2 Marks]

7.7) Comment on the distribution models obtained in Q7.3 and Q7.4. Which one is better? [2 Marks]

7.8) What is the main problem that you might come across when performing a maximum likelihood estimation using mixture of Gaussians? How can you resolve that problem in practice? [3 Marks]