SIT743 Multivariate and Categorical Data Analysis
Assignment-1

Total Marks = 100, Weighting - 15%

Due date: 26 August 2018 by 11.30 PM

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

1. **A written document** (only pdf) 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 and data files) that you will be submitting.
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. **Two data files** named "*name-StudentID-BikeShareMyData.txt*" and "*name-StudentID-PCASelData.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 four files should be uploaded separately to CloudDeakin. E-mail or manual submissions are **NOT** allowed. Photos of the document are **NOT** allowed.

==

Some of the questions in this assignment require you to use the "BikeShare" dataset. This dataset is given as a text file, named "BikeShareTabSep.txt". You can download this from the Assignment folder in CloudDeakin. Below is the description of this dataset.

Bike sharing dataset (BikeShare)
This dataset gives the count of bikes rented between 11am - 12pm on different days and locations through the *Capital Bikeshare System* (operating in US cities) between 2011 and 2012. The variables include the following (9 variables):

Season: Categorical: 1 = Spring, 2 = Summer, 3 = Autumn (fall), 4 = Winter

Working day: 0 = Weekend, 1 = Workday.

Weather: Categorical variable

1: Clear, Few clouds, Partly cloudy, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered cloud
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

Temperature: Temperature in Celsius.

**`Feeling' Temperature:** `Feels like' temperature, reported in Celsius.

Humidity: Humidity (given as a percentage).

Windspeed: Windspeed (measured in km/h).

Casual users: Count of casual users that used a bike at that time.

Registered users: Count of registered users that used a bike at that time.
Assignment tasks

Q1) [16 Marks]:

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

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

- Generate a sample of 400 data using the following:

```
my.data <- the.data [sample(1:727,400),c(1:9)]
```

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

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

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

1.1) Draw histograms for 'Registered users' and 'Temperature' values, and comment on them. [3 Marks]

1.2) Give the **five number summary** and the **mean value** for the 'Casual users' and the 'Registered users' separately. [3 Marks]

1.3) Draw a parallel Box plot using the two variables; 'Casual users' and the 'Registered users'. Use the answers to Q1.2 and the Boxplots to compare and comment on them. [3 Marks]

1.4) Draw a scatterplot of 'Temperature' and 'Casual users' for the ***first 200 data vectors selected from the "my.data"*** (name the axes) and comment on them [2 Marks]

1.5) Fit a linear regression model to the ‘temperature’ (as x) and the ‘casual users’ (as y) using the ***first 200 data vectors selected from the “my.data”***. Write down the linear regression equation. Plot the line on the same scatter plot. Compute the correlation coefficient and the coefficient of Determination. Explain what these results reveal. [5 Marks]

Q2) [21 Marks]

The table shows results of a survey conducted about the type of vehicle people own (in thousands) in different states over a five year period between 2011 and 2016.

		State			
		New south Wales (N)	**Victoria (V)**	**Queeensland (Q)**	Total
Vehicle type	**Passenger (P)**	1360	1140	810	**3310**
	Light commercial (C)	260	190	240	**690**
	Total	**1620**	**1330**	**1050**	**4000**

Suppose we select a person at random,

2.1) What is the probability that the person is from Victoria (V)? [1 mark]

2.2) What is the probability that the person owns a light commercial vehicle (C)? [1 mark]

2.3) What is the probability that the person owns a passenger vehicle (P) and from New South Wales (N)? [1 Mark]

2.4) What is the probability that the person owns a light commercial vehicle (C) given that he/she is from Queensland (Q)? [2 Marks]

2.5) What is the probability that the person, who owns a passenger vehicle is from Queensland (Q)? [2 Marks]

2.6) What is the probability that the person is from Victoria (V) or owns a passenger vehicle (P)? [3 Marks]

2.7) find the marginal distribution of the vehicle type [2 marks]

2.8) find the marginal distribution of the state [3 marks]

2.9) find the conditional distribution of vehicle type within each state. [6 marks]

Q3) [4 Marks]

Suppose that 20% of the adults smoke cigarettes. It is known that 60% of smokers and 15% of non-smokers develop a certain lung condition. What is the probability that someone with the lung condition was a smoker? [4 Marks]

Q4) Maximum Likelihood Estimation (MLE) [10 Marks]

The number of cars x_i arrive at a shopping centre on a given day i is modelled by a Poisson distribution with unknown parameter θ as given by the following equation.

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

$$\boldsymbol{Poid(\theta) = p(x_i|\theta) = \frac{\theta^{x_i} e^{-\theta}}{x_i!}}$$

Assume that we consider N consecutive days, and the cars arrive at the shopping centre are independently and identically distributed (iid).

a) Show that the expression for the likelihood (joint distribution) $p(\boldsymbol{X}|\theta)$ of the arrival of cars for N days ($\boldsymbol{X} = \{x_1, x_2, \dots, x_N\}$) is given by

$$\boldsymbol{p(X|\theta) = \frac{\theta^{N\bar{x}} \; e^{-N\theta}}{x_1!\, x_2!\, x_3! \dots x_N!}} \quad \textbf{, where } \boldsymbol{\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

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

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

[3 marks]

b) Find an expression for the logliklihood function $\boldsymbol{L(\theta) = \ln(p(X|\theta))}$ **[**2 marks**]**

c) In order to find the Maximum likelihood Estimation (MLE) of parameter θ, we 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:

$$\hat{\theta} = \boldsymbol{\bar{x}}\ ,\quad \boldsymbol{where\ \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

[3 Marks]

d) Suppose that we observe the number of cars arrived on the three days as $x_1 = 100,\ x_2 = 60\ and\ x_3 = 70$ .What is the MLE given this data? [2 Marks]

Q5) Bayesian inference for Gaussians (unknown mean and known variance) [16 marks]

5.1) What is the meaning of conjugate prior? [2 marks]

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

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 rainfall received at the Murray basin are measured for n years. The **average** rainfall observed over the n years is 1100 mm. Assume that the annual rainfall are **normally** distributed with unknown mean θ and known standard deviation 200 mm. Suppose your prior distribution for θ is **normal** with mean 800 mm and standard deviation 100 mm.

a) State the posterior distribution for θ (this will be in terms of n. Do not derive the formulae) [3 Marks]
b) For n=3, 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=3 in the above question Q5.4(b) and comment. [3 Marks]

Q6) Dimensionality Reduction: [19 Marks]

Use the "BikeShare" data for this question. Use the following code to load randomly selected 200 (or 100) data points. Note that only features from 4 to 9 are used here.

```
the.data <- as.matrix(read.table("BikeShareTabSep.txt"))
selData <- the.data [sample(1:727,200),c(4:9)]
```

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

```
write.table(selData,"name-StudentID-PCASelData.txt")
```

6.1) Conduct a principal component analysis (PCA) on this data (`selData`). Use the below mentioned "biplot" code (in R) to produce a scatterplot using the first two principal components. Comment on the plot. [4 Marks]

```
pZ <- prcomp(selData, tol = 0.01, scale = TRUE)
pZ
summary(pZ)
biplot(pZ)
```

6.2) Draw a graph of variance verses the principal components, and explain how this can be used to determine the correct number of principal components. [3 Marks]

6.3) For the same data above (selData), compute the Euclidean distance matrix. Use the distance matrix to perform a classical multidimensional scaling (classical MDS or Metric MDS). You can use the following command

```
mds <- cmdscale(selData.dist)  # here 'selData.dist' is the distance matrix
```

Plot the results and comment on them [4 Marks]

6.4) For the same data above (selData), perform a non-metric MDS, called 'isoMDS' in R using number of **dimensions k set to 2**. Use the following command to do this:

```
library(MASS)
fit<-isoMDS(selData.dist, k=2)
```

Plot the results of this isoMDS [2 Marks]

6.5) Draw the Shepard plot for this isoMDS results and comment on them [3 Marks]

6.6) For the same data above (selData), perform a non-metric MDS, called 'isoMDS' in R using the number of **dimensions k set to 4**.

```
library(MASS)
fit<-isoMDS(selData.dist, k=4)
```

Draw the Shepard plot for this isoMDS results and compare the plot obtained for k=2 in Q6.6 above. Comment on them [3 Marks]

Q7) Clustering: [14 marks]

7.1) ***K-Means clustering:*** Use the data file "SITdata2018.txt" provided in CloudDeakin for this question. Load the file "SITdata2018.txt" using the following:

```
zz<-read.table("SITdata2018.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 "SITdata2018" (zz) [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. 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]

7.2) ***Spectral Clustering:*** Use the same dataset (`zz`) and run a spectral clustering (use the number of clusters/centers as 4) 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]