Assignment Submission Coversheet

Faculty of Science, Engineering and Built Environment

Student ID:	218545396
Student Name:	Pooja Bhat
Campus:	☐ Burwood ☐ Waterfront ☐ Waurn Ponds ☐ Warrnambool ☒ Cloud

Assignment Title:	Assessment 1: Multivariate and Categorical Data Analysis		
Due Date:	14 April 2019 by 11.30 PM	**Assessment Item:**	Report
Course Code/Name:	S777/ Master of Data Analytics		
Unit Code/Name:	SIT743/ Multivariate and Categorical Data Analysis	**Unit Chair / Campus Coordinator:**	Sutharshan Rajasegarar
Practical Group: (if applicable)	Not applicable		

If this assignment has been completed by a group or team:

1. Each student in the group must complete and sign a separate coversheet
2. The assignment will be returned to the student in the group nominated below

Assignment to be returned to: (Student name and Student ID number)	

PLAGIARISM AND COLLUSION

Plagiarism occurs when a student passes off as the students own work, or copies without acknowledgement as to its authorship, the work of another person. 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 for the purpose of detecting plagiarism and collusion.

DECLARATION

I certify that the attached work is entirely my own (or where submitted to meet the requirements of an approved group assignment, is the work of the group), except where work quoted or paraphrased is acknowledged in the text. I also certify that it has not been previously submitted for assessment in this or any other unit or course unless permissions for this has been granted by the Unit Chair of this unit. I agree that Deakin University may make and retain copies of this work for the purposes of marking and review, and may submit this work to an external plagiarism-detection service who may retain a copy for future plagiarism detection but will not release it or use it for any other purpose.

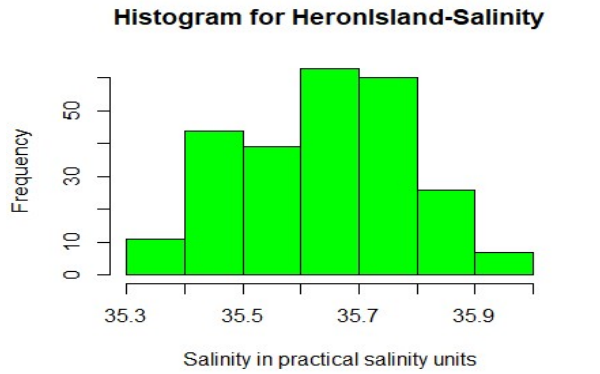
Signed:	POOJA BHAT	**Date:**	14.04.2019

An assignment will not be accepted for assessment if the declaration appearing above has not been signed by the author. If submitting electronically, print your full name in place of a signature.

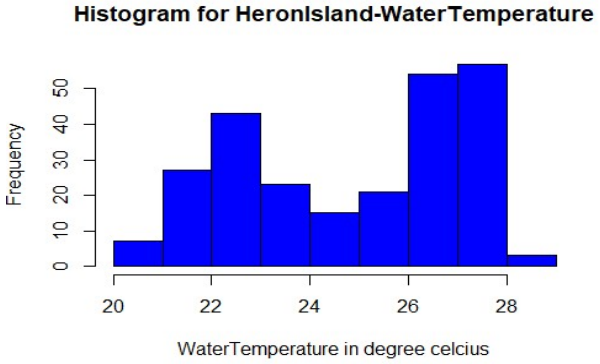
COMMENTS

Mark Awarded:		**Assessor's Signature:**		**Date:**	

1.1) Histograms for 'HeronIsland-Salinity' and 'HeronIsland-Water Temperature'



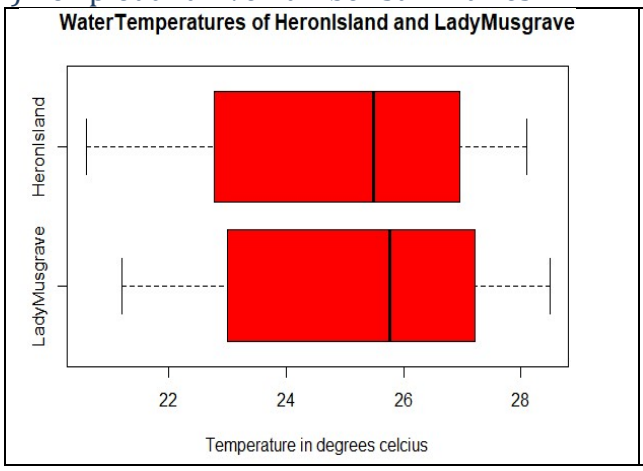
The histogram for the Salinity in HeronIsland shows that most of the samples showed salinity between 35.6 and 35.9 practical salinity units and the distribution looks mostly like a symmetric distribution. This is a unimodal histogram with a peak salinity between 35.6 and 35.8 practical salinity units. There were a few samples found to be below 35.4 and above 35.8 practical salinity units.



The histogram of HeronIsland – Water temperature shows 2 peaks at 27 degrees Celsius and 23 degrees Celsius, indicating this might be a bimodal distribution. There were a very few number of samples also found to be at lower temperatures of under 21 degrees Celsius an higher temperatures of above 28 degrees Celsius.

Figure 1: Histograms of HeronIsland's salinity and water temperatures.

1.2) Box plot and five number summaries



5 number summary for HeronIsland-WaterTemperature

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
20.60	22.77	25.49	24.86	26.96	28.09

5 number summary for LadyMusgrave-WaterTemperature

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
21.20	23.00	25.77	25.16	27.22	28.49

Figure 2: Boxplot and 5 number summary of water temperatures of the 2 islands

The Box plot of the water temperatures at both islands are negatively skewed. The median of temperatures at HeronIsland and LadyMusgrave were almost similar at about 25.49 and 25.77 respectively. The temperatures in HeronIsland range from 20.6 to 28.09 degrees Celcius. Whereas at LadyMusgrave range from 21.2 to 28.49 degrees Celicus. In HeronIsland, the lower quartile takes values from 20.6 to 22.77 degrees Celcius whereas the upper quartile takes value from 26.96 to 28.09 degrees Celicus. In LadyMusgrave, the lower quartile takes values from 21.2 to 23.0 wherease the higher quartile takes values from 27.22 to 28.49 degrees celcius. In both islands the middle 50% seems to be under similar

range of between 22.77 to 26.96 degrees celcius at HeronIsland and 23-27.22 degrees celcius at LadyMusgrave. The mean in both islands is lower than the median showing that the data is negatively skewed.

1.3) Scatterplot of 'HeronIsland-Water Temperature' and 'LadyMusgrave-Water Temperature'

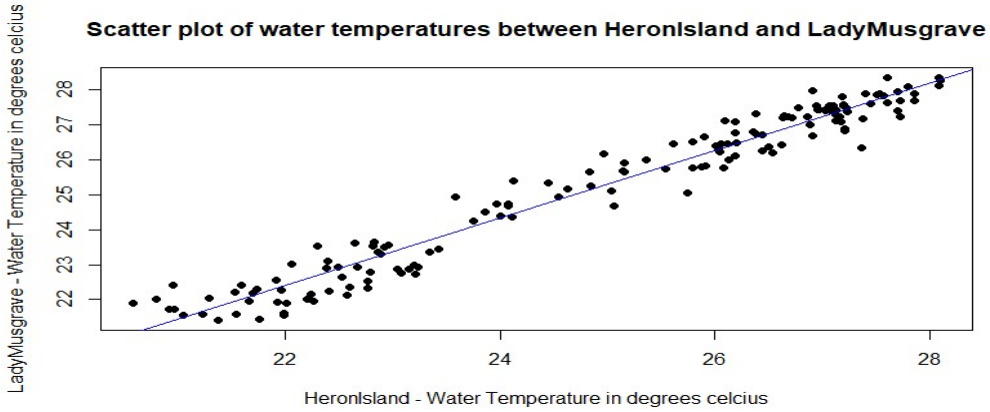


Figure 3: Scatterplot and regression line of water temperature between HeronIsland and LadyMusgrave

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.05437    0.41765   2.525   0.0126 *
HeronIsland  0.96978    0.01672  58.012   <2e-16 ***
Residual standard error: 0.4541 on 148 degrees of freedom
```

The Liner regression equation can be written as LadyMusgrave-Water Temperature in Degree Celcius = 1.05437 + .96978(HeronIsland-Water Temperature in Degree Celcius)

From the R output we can calculate the correlation coefficient and coefficient of determination

Correlation coefficient	0.9787113
Coefficient of Determination	0.9578759

From the plot above, we can see that there is a very strong relationship between the water temperatures of HeronIsland and LadyMusgrave. The regression line we plotted was a straight line and the coefficient of determination was 0.9578 which shows that the model is a very good model and that 95.78% of variation in water temperature at LadyMusgrave can be explained by variations of the water temperature at HeronIsland. About 4.22% of the variations in water temperature cannot be explained by variations of the water temperature at HeronIsland and are influenced by some other factors. The coefficient of correlation was 0.9787 showing a very strong positive relationship between the water temperatures at the two island.

2) Probabilty

		Occupation			Total
		Professionals (P)	**Sales Workers (S)**	**Community Service (C)**	
Age group in years	**Below 30 (B)**	15	40	60	**115**
	30-60 (M)	100	10	10	**120**
	Above 50 (A)	28	4	3	**35**
Total		**143**	**54**	**73**	**270**

Table 1: Contingency table for age and occupation

2.1) What is the probability that the person is a Professional (P)?
143/270 = 0.529629

2.2) What is the probability that the person is below 30 years old (B)?
115/270 = 0.425925

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)?
10/270 = 0.037037

2.4) What is the probability that the person is a sales worker (S) given that he/she is above 50 years old (A)?
4/35 = 0.114285

2.5) What is the probability that the person, who is a professional (P), is below 30 years old (B)?
15/143=0.104895

2.6) What is the probability that the person is a sales worker (S) or between 30 to 50 years old (M)?
P(S U M) = P(S) + P(M) – P(S∩M)
= (54/270) + (120/270) – (10/270) = 164/270 0.607407

2.7) Find the marginal distribution of the occupation

Professionals (P)	**Sales Workers (S)**	**Community Service (C)**
143/270 = 0.529629	54/270 = 0.2	73/270 = 0.270370

2.8) Find the marginal distribution of the age group

Below 30 (B)	**30-60 (M)**	**Above 50 (A)**
115/270 = 0.425926	120/270 = 0.444444	35/270 = 0.12963

2.9) Are the variables 'occupation' and 'Age group' independent random variables? Explain why or why not.
Events are independent when probability of one event is not affected by probability of another event.

When 2 event A and B are independent: $P(A \cap B) = P(A) \cdot P(B)$
In our case we will check with one of the occupation and one of the age group and see if the events are independent.

Taking
P(S) = 0.2
P(M) =0.444444
P(S ∩ M) = 0.037037

P(M) . P(S) = 0.088888

Thus P(M) . P(S) not equal to P(S ∩ M). Hence the 2 event are not independent, which means the age and occupation are related events.

3) Assembly Line

Let P(A) be probability of drones being manufactured on Assembly line A
Let P(B) be probability of drones being manufactured on Assembly line B
Let P(Passed) be the probability that drones passed the quality test
Let P(Failed) be the probability that drones failed the quality test

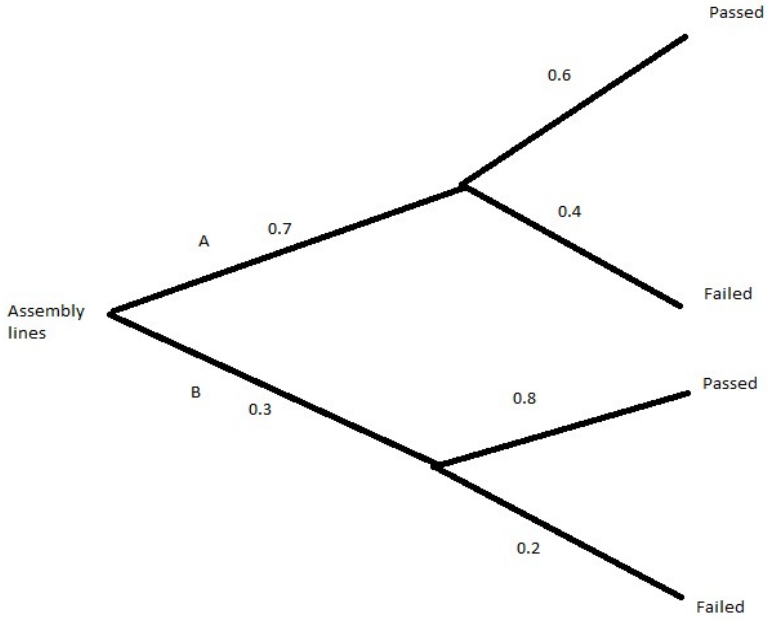


Figure 4: Tree diagram showing probabilities of drones.

a) What is the overall proportion of the drones produced pass the quality test?
Proportion of drones produced pass the quality test = P (A and Passed) + P(B and Passed)
= (0.70 x 0.60) + (0.30 x 0.80)
= 0.42 + 0.24
=.66
Therefore 66% of the total drones passed the quality test.

b) If a randomly selected drone passed the quality test, what is the probability that it was produced by assembly line B?
This can be written as : $\frac{P(B\ and\ Passed)}{((P(B\ and\ Passed)+P(A\ and\ Passed)}$
$\frac{P(0.3 \times 0.8)}{(P(0.3 \times 0.8)+P(0.7 \times 0.6))} = \frac{0.24}{0.66} = 0.363636$

The probability that the randomly selected drone was produced by assembly line B is 36.3636%

Q4) Maximum Likelihood Estimation

a) $x_i \sim sp\text{Weibull}(\theta)$

$sp\text{Weibull}(\theta) = p(x_i|\theta) = 2\theta^2 x_i e^{-\theta^2 x_i^2}$

The joint distribution $P(x|\theta)$ for $(x = \{x_1, x_2 \dots x_N\})$ can be written as

$P(x|\theta) = P(x_1|\theta) \times P(x_2|\theta) \times P(x_3|\theta) \times \dots P(x_N|\theta)$

Since the time of failures of N servers are iid.

Thus, $P(x|\theta)$ can be written as below.

$$P[x|\theta] = (2\theta^2 x_1 e^{-\theta^2 x_1^2}) \times (2\theta^2 x_2 e^{-\theta^2 x_2^2}) \dots (2\theta^2 x_3 e^{-\theta^2 x_3^2})$$

$$= C(2^{(1+1\dots1)}) \theta^{(2+2\dots2)^N} \prod_{i=1}^{N} x_i \, e^{-\theta^2 \sum_{i=1}^{N} x_i^2}$$

$$= C\, 2^N \theta^{2N} e^{-S\theta^2}$$

Where $C = \prod_{i=1}^{N} x_i$

$S = \sum_{i=1}^{N} x_i^2$

b) $L(\theta) = \ln\left[C\, 2^N \theta^{2N} e^{-S\theta^2}\right]$

$\ln(C) + N\ln(2) + 2N\ln\theta - S\theta^2$ $\quad \boxed{\ln(e^a) = a}$

c) To find the maximum likelihood estimate, we will differentiate the log likelihood function and equate it to zero.

$$\therefore \frac{dL(\theta)}{d\theta} = \frac{d}{d\theta}\left[\ln(C) + N\ln(2) + 2N\ln\theta - S\theta^2\right]$$

$$0 + 0 + \frac{2N}{\theta} - 2S\theta = 0$$

$$\therefore \frac{2N}{\theta} = 2S\theta$$

$$2N = 2S\theta^2$$

$$\theta^2 = \frac{N}{S}$$

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

d)

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

$$\hat{\theta} = \sqrt{\frac{N}{S}}$$

$$= \sqrt{\frac{5}{(20)^2 + (15)^2 + (12)^2 + (40)^2 + (35)^2}}$$

$$= \sqrt{\frac{5}{3594}}$$

$$= 0.037298$$

Q5) Bayesian inference for Gaussians (unknown mean and known variance)

1) If the posterior distributions $p(\mu/D)$ are in the same family as the prior probability distribution $p(\mu)$, the prior and posterior are then called conjugate distributions and the prior is called a conjugate prior for the likelihood function.
2) In Bayesian statistics, having a conjugate prior gives a closed form expression for the posterior Thus it becomes computationally efficient to derive the mean, variance of the posterior as well as it eliminates the need to use numerical integration when computing the posterior.
3) Examples of conjugate pairs:
   a) If likelihood is Gaussian, and the prior is Gaussian then the posterior is Gaussian
   b) If likelihood is a Multinomial model, and the prior is a Dirichlet model then the posterior is a Dirichlet model.
   c) If likelihood is a Bernoulli model, and the prior is a Beta model then the posterior is a Beta model.
4) a) The prior is a Gaussian with $P(\theta) \sim \mathcal{N}(m, \tau^2)$
   Likelihood is a Gaussian with $P(X/\theta) \sim (\theta, \sigma^2)$

   Therefore, the posterior is a Gaussian with $P(\theta|X) \sim \mathcal{N}(\boldsymbol{\mu_N}, \boldsymbol{\sigma_N}^2)$

   Where, $\boldsymbol{\mu_N}$ is the mean of the posterior and $\boldsymbol{\sigma_N}^2$ is the variance of the posterior.

$$\mu_N = \sigma_N^2 \left(\frac{N\bar{x}}{\sigma^2} + \frac{m}{\tau^2}\right); \quad \frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\tau^2} \quad \text{and } \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Where *N is the number of observations.*

Prior $P(\theta) \sim N(m, \tau^2)$
Likelihood $P(x|\theta) \sim N(\theta, \sigma^2)$
Posterior $P(\theta/x) \sim N(\mu_N, \sigma_N^2)$

$$\mu_N = \sigma_N^2\left(\frac{N\bar{x}}{\sigma^2} + \frac{m}{\tau^2}\right) = \sigma_N^2\left[\frac{N\bar{x}}{25} + \frac{10}{9}\right]$$

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\tau^2} = \frac{N}{25} + \frac{1}{9}$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = 15$$

$\bar{x} = 15\text{cm}$; $m = 10$, $\tau^2 = (3)^2 = 9$, $\sigma^2 = (5)^2 = 25$

b) $n = 5$ $\therefore \frac{1}{\sigma_N^2} = \frac{5}{25} + \frac{1}{9} = \frac{45+25}{225} = \frac{70}{225}$

$\therefore \sigma_N^2 = \frac{225}{70} = 3.214$

Std dev^2 $= \sqrt{\sigma_N^2} = 1.7928$

$$\mu_N = \sigma_N^2\left(\frac{N\bar{x}}{\sigma^2} + \frac{m}{\tau^2}\right) = \frac{225}{70}\left[\frac{5\times 15}{25} + \frac{10}{9}\right]$$

$$= \frac{225}{70}\left[\frac{675+250}{225}\right]$$

$$\mu_N = \frac{925}{70} = 13.2142$$

Posterior mean is 13.2142 cm and Posterior Standard deviation is 1.7928

The Posterior variance (3.214) is less than the prior variance (9) and the likelihood variance (25)

c) $n = 15$

$$\frac{1}{\sigma_N^2} = \frac{15}{25} + \frac{1}{9} = \frac{135+25}{225} = \frac{160}{225}$$

$$\sigma_N^2 = \frac{225}{160} = 1.40625$$

Std deviation $(\sqrt{\sigma_N^2}) = 1.1858$

$$\mu_N = \frac{225}{160}\left[\frac{15 \times 15}{25} + \frac{10}{9}\right]$$

$$\frac{225}{160}\left[\frac{2025+250}{225}\right]$$

$$\mu_N = \frac{2275}{160} = 14.21875$$

Posterior mean is 14.21875 and posterior Std deviation is 1.1858

From this we can see that as the value of N has increased, the posterior variance has decreased and the posterior mean has increased.

d) With new Prior values. $m = 10$ $(\tau)^2 = (1)^2 = 1$ $n = 10$

$\bar{x} = 15$ $\sigma^2 = 25$

$$\frac{1}{\sigma_N^2} = \frac{10}{25} + \frac{1}{1} = \frac{10+25}{25} = \frac{35}{25} \qquad \sigma_N^2 = \frac{25}{35} = 0.7142$$

SD = 0.8451

$$\mu_N = \frac{25}{35}\left[\frac{10 \times 15}{25} + \frac{10}{1}\right]$$

$$\mu_N \quad \frac{25}{35}\left[\frac{150+250}{25}\right] = \frac{400}{35} = 11.4285$$

The mean of the Posterior distribution is 11.4285 and the Standard deviation is 0.8451

Bayesian estimation

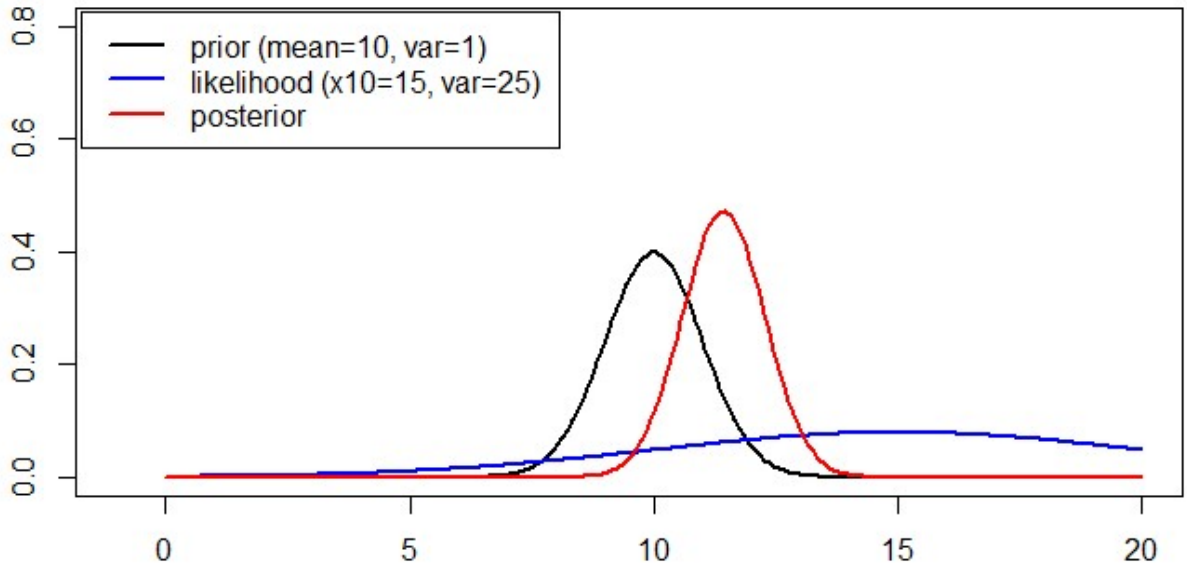


Figure 5: Output of R plots showing prior, likelihood and the posterior distributions

Q6.1) K-Means clustering:

1) Scatter plot of the data.

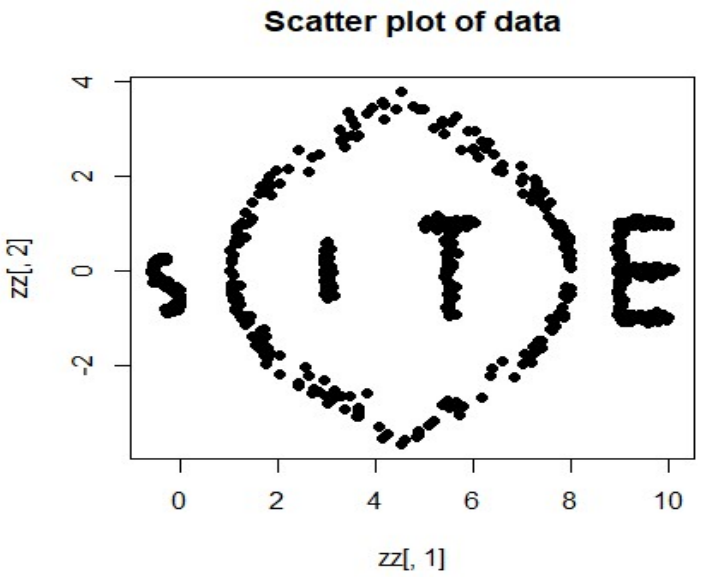


Figure 6: R output of scatter plot of provided data

2) From a visual examination of the scatter plot, there seem to be 5 distinct clusters, For S, I, T, E and the boundary that encircles the letters I and T
3) The scatter plot of the data after k-means clustering with 5 classes creates clusters by bunching the closest points together and does not take into account the pattern in the center around the letters I and T. This shows the shortcomings of K-means clustering.

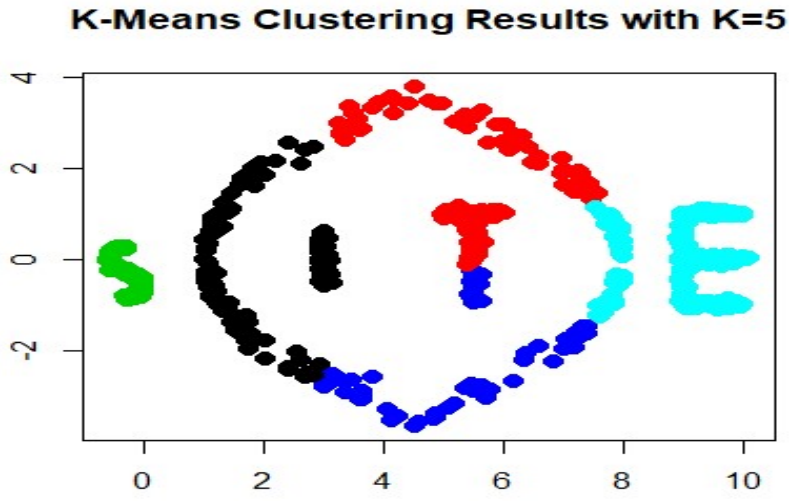


Figure 7: R output of K-means clustering with K-5

4) I have written R code to calculate the Total within sum of square values (TOTWSS) and plotted graph as below. The location where a bend (Knee) is apparent in the plot is considered as an indicator of the appropriate number of clusters.

From the below we plotted TOTWSS starting with k=2 to K=20. We can see 2 bends in the graph. One at K=3 and other at k=5

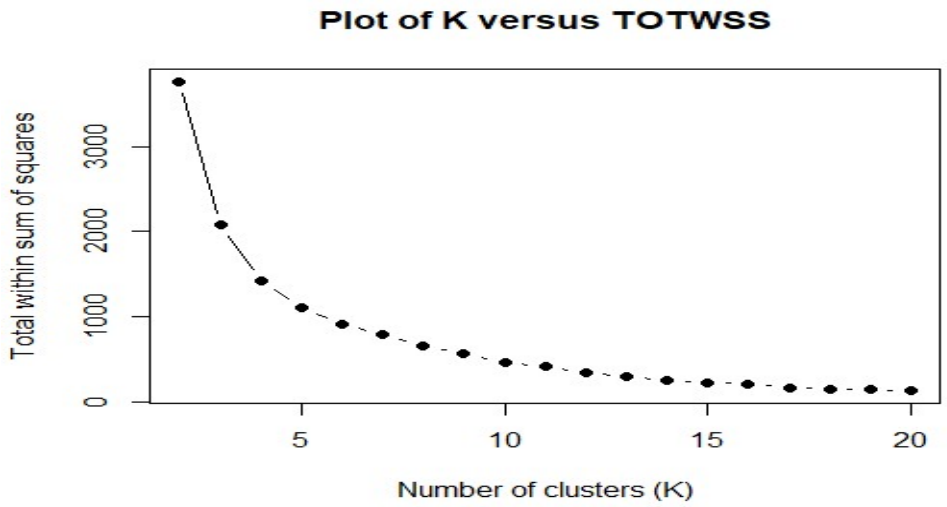


Figure 8: R output of Plot of K v/s TOTWSS

Q6.2) Spectral clustering:

The following plot was derived after trying various values for the nearest neighbors. After choosing 11 as the nearest neighbor, we get the below output. In comparison with the k-means clustering output, it is evident here that the circle around the letters S,I,T,E were correctly clustered by spectral clustering as opposed to k-means clustering which did not cluster them correctly. This shows that spectral clustering is more powerful in clustering when we have non-convex shapes in the dataset.

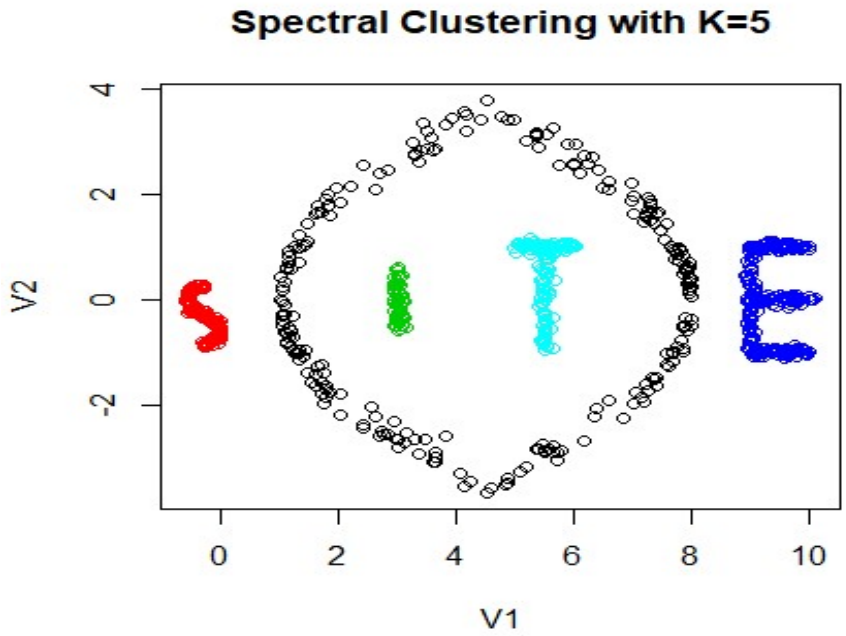


Figure 9: R output of Spectral clustering with K=5

Q7) LadyMusgrave Water temperature analysis

1) The time series plot of the LadyMusgrave Water temperature is as below

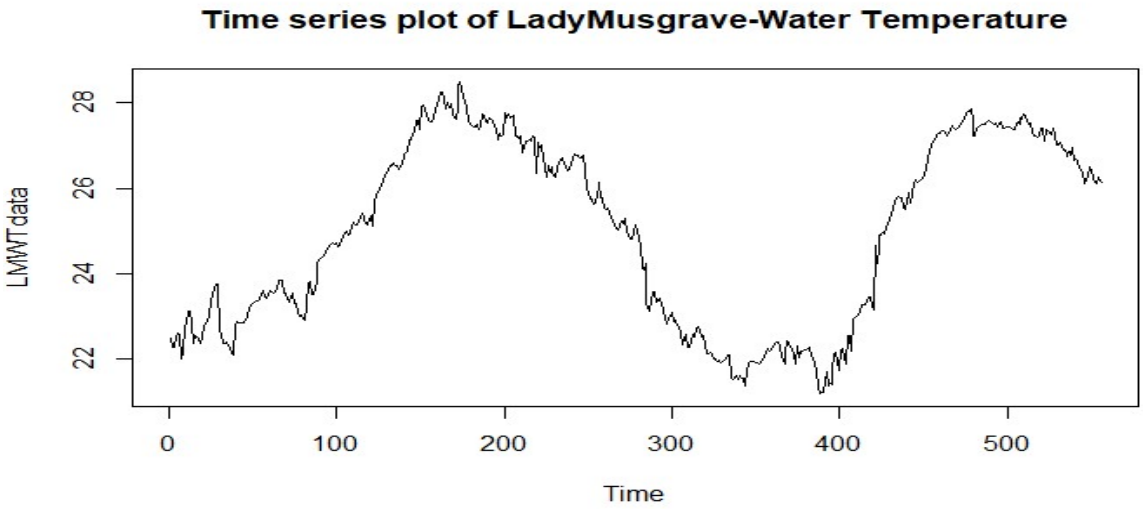


Figure 10: Time series plot of LadyMusgrave- Water Temperature

2) The Histogram of the LadyMusgrave Water temperature is plotted below. The histogram shows 2 modes, which may mean that there are 2 underlying distributions in this dataset and would need to be separated and analyzed separately.

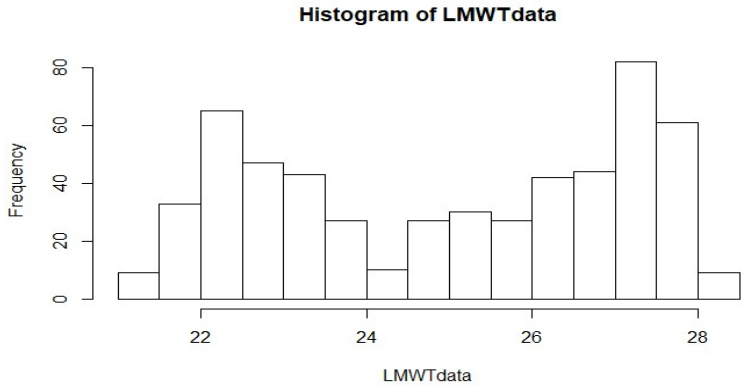


Figure 11: Histogram of LadyMusgrave- Water Temperature

3) After fitting a single Gaussian model to the distribution, we get the below output for Maximum likelihood estimates. The Mean μ is 24.98 and the standard deviation σ is 2.1489.

```
> fit1
      mean           sd
  24.98047914    2.14893739
 ( 0.09113525) ( 0.06444235)
```

Using the mean and the standard deviation derived, we will plot the density function of the distribution.

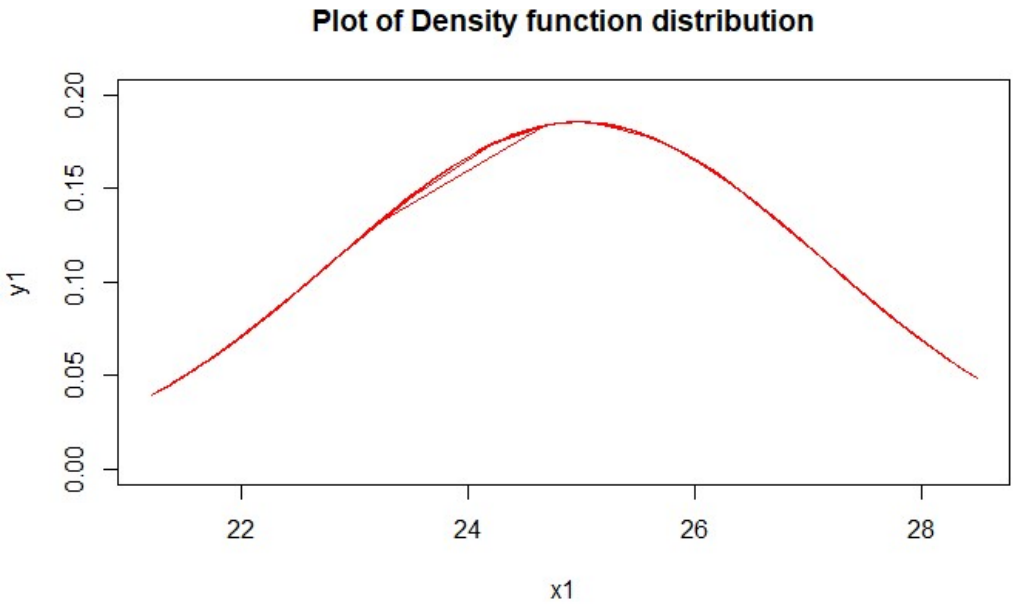


Figure 12: Plot of Density function distribution

4) Next we will try to fit a mixture of Gaussian model to the distribution, using 2 Gaussians since we saw 2 modes in the histograms. The derived values of mixing coefficients, mean and standard deviation for each of the Gaussians are as below.

	Component 1	Component 2
Mixing coefficient	0.411484	0.588516
mu	22.653813	26.607258
sigma	0.713408	1.029322

Table 2: Lambda, Mu and Sigma of the 2 Gaussians

5) Plot of these Gaussians on top of the histogram is as below, the combined density function is also mapped on this plot.

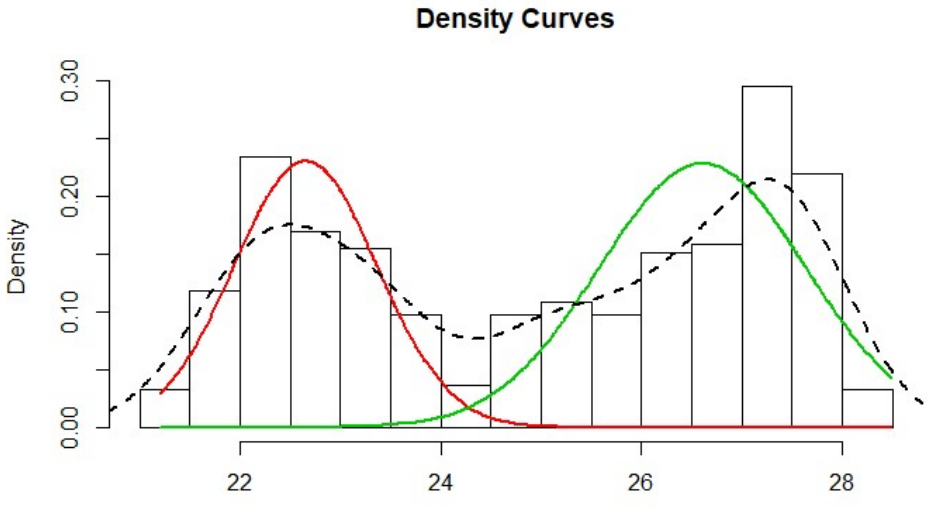


Figure 13: Plot of Density function for the 2 Gaussians and combined density function

6) Plot of the log likelihood values shows that the log likelihood converges to a stable state when the index is at 4 and stays constant thereafter.

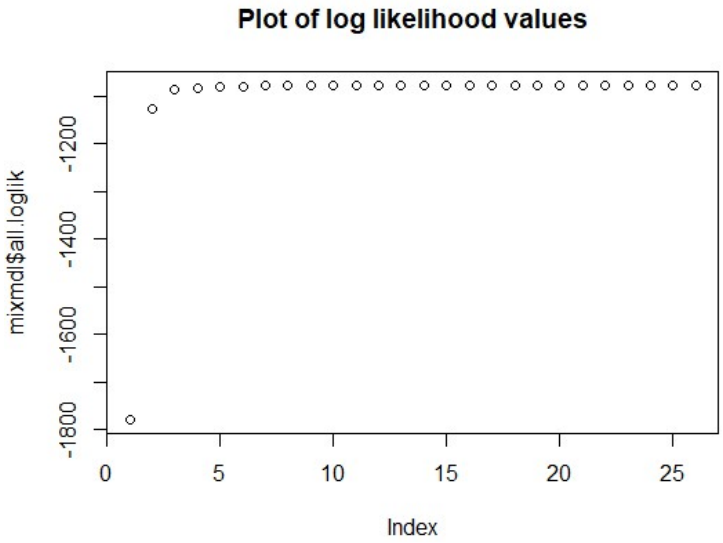


Figure 14: Plot of log likelihood function

7) In 7.3, we tried to fit a single Gaussian with normal distribution to the data, which resulted in a plot with a single peak at about 25. It was not a correct fit, since our data distribution was bimodal. In 7.4, we tried to fit 2 Gaussians which showed a much closer relationship to the 2 peaks in the histograms.
   If we compare the distribution models obtained in 7.3 and 7.4, we can see the 2 Gaussians fit the actual data distribution much closer than the single normal distribution.

8) There are 3 problems with performing maximum likelihood estimation using mixture of Gaussian. These are below:
   1) Presence of Singularities: If we have a mix of Gaussians with one of the Gaussian having a narrow spread with its mean exactly on one of the data points, when we compute the likelihood for this it results in likelihood of zero for all data points and one for the single point, causing severe overfitting. The way to overcome this is to use heuristics to identify if the variance is very small and reassign a large value to it and resetting the mean to a random value, and then continuing the optimization iteratively. This is the biggest problem of performing maximum likelihood function in Gaussian mixture models.
   2) Identifiability problem: A K component mixture has a total of K! possible solutions, for any given point in the space of parameter values there will be a further K!-1 additional points all giving exact same distribution – This is known as Identifiability problem. This needs to be considered when parameters discovered by a model are interpreted. However, this is not a major problem since in finding a good density model any of the assumed solution is as good as the other solutions.
   3) Complexity in maximizing mixture models: Maximizing the log likelihood for a Gaussian mixture model is more complex than for the case of a single Gaussian because the calculations require summation over k that appears inside the log. This problem can be overcome by using EM approach for finding maximum likelihood solutions.