SIT743 Multivariate and Categorical Data Analysis Assignment-1

Question:1)

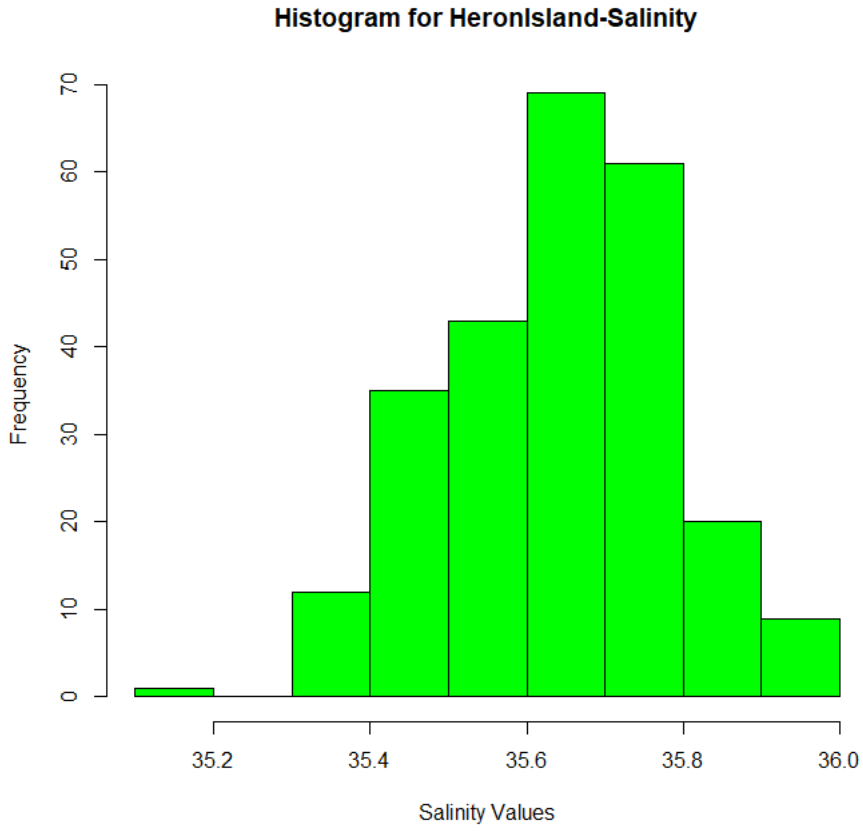
Code:

```
#Great Bareer Reef Data Question 1

the.data <-as.matrix(read.csv("GBRData.csv", header = TRUE, sep = ","))
my.data <- the.data [sample(1:556,250),c(1:4)]
write.table(my.data,"Ashutosh-217669865-GBRMyData.txt")
```

1.1) **Code:**

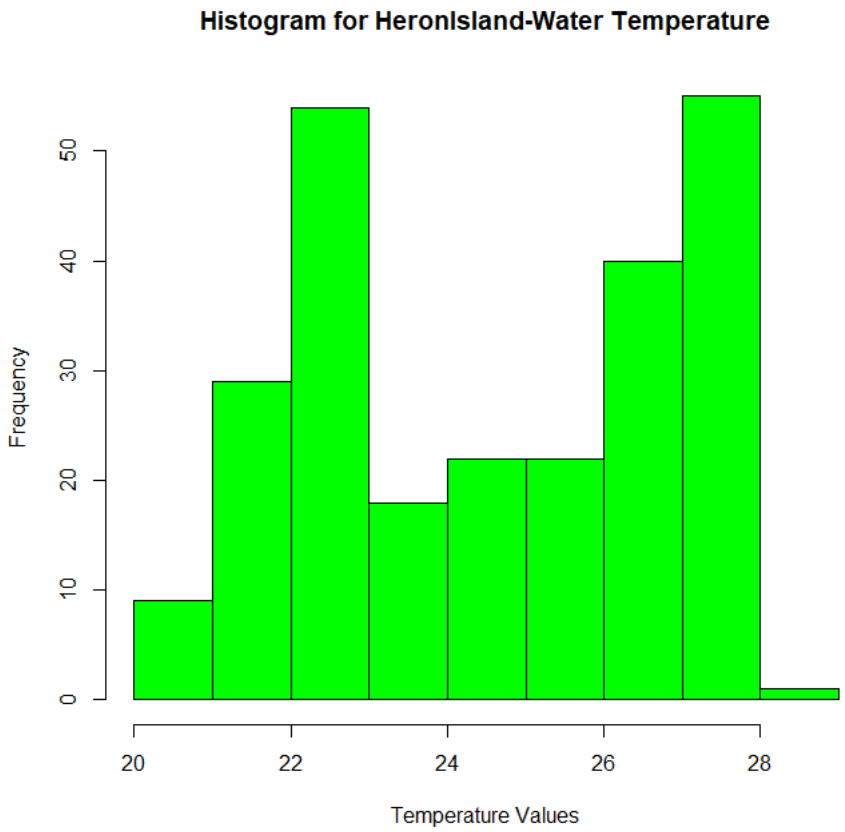
```
hist(my.data[,1],col="green", main="Histogram for HeronIsland-
Salinity",xlab="Salinity Values")
hist(my.data[,2],col="green", main="Histogram for HeronIsland-Water
Temperature",xlab = "Temperature Values")
```

Output:



As we can see that the histogram is Unimodal and left (negative)skewed as there are higher values on the right and it tails

towards left. It also has several outliers at its left i.e. the lower values of low frequency.

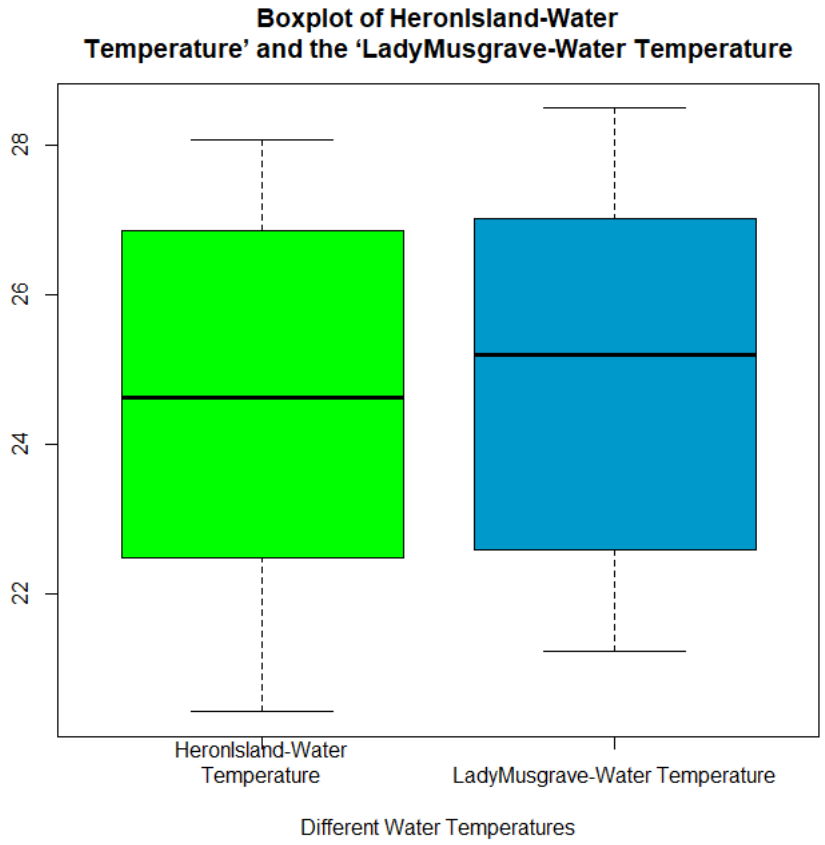


Here we can see that the histogram is Bimodal, and it is also negatively skewed a bot as higher values are accumulated at the right.

1.2) **Code:**

```
boxplot(my.data[,2],my.data[,4],main="Boxplot of HeronIsland-Water Temperature' and the 'LadyMusgrave-Water Temperature",xlab="Different Water Temperatures",names=c("HeronIsland-Water Temperature","LadyMusgrave-Water Temperature"),col=c("green","#0099CC"))

summary(my.data[,2])
summary(my.data[,4])
```

Output:




```
> summary(my.data[,2])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.42   22.49   24.63   24.58   26.86   28.08
> summary(my.data[,4])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.23   22.60   25.19   24.92   27.01   28.49
>
```

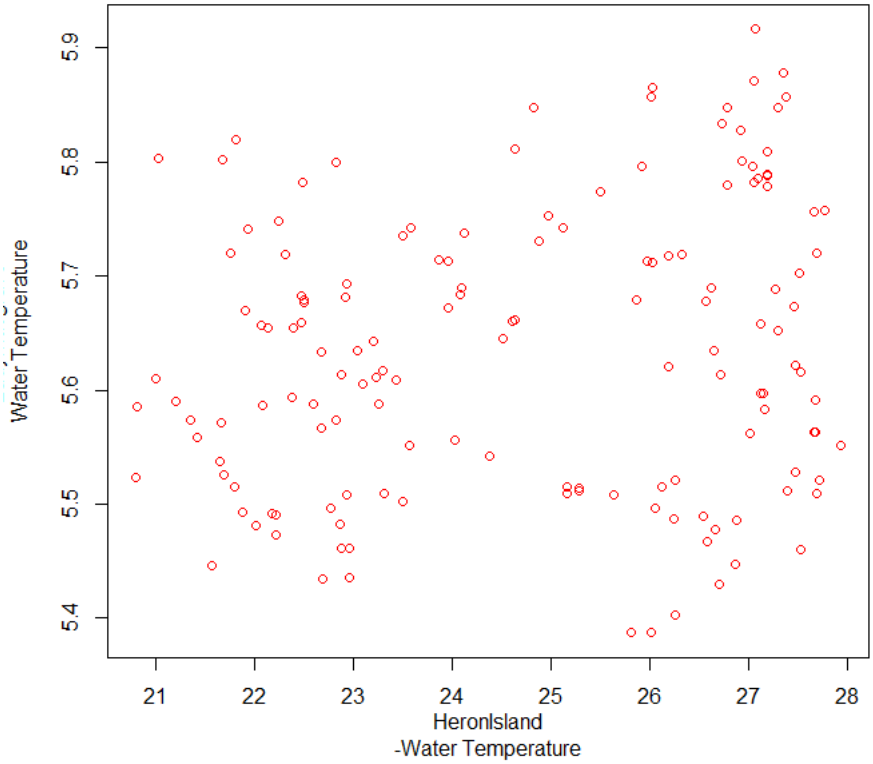
Comments:

- Lady Musgrave Water Temperature has higher interquartile range when compared to HeronIsland's which means that data of Lady Musgrave Water Temperature have higher variation in comparison to HeronIsland's Water Temperature.
- HeronIsland's Water Temperature is more balanced in terms of quartiles rather than Lady Musgrave Water Temperature around the median.
- The max and min value for HeronIsland's Water Temperature is 28.08 and 20.42 respectively whereas for Lady Musgrave Water Temperature it is slightly higher that of 21.23 and 28.49
- The means of both are 24.58 and 24.92 respectively.
- Now, the median is where both can be distinguished which is 24.63 and 25.19(larger to lady Musgrave)

1.3)**Code:**

```
e<-cbind(my.data[,2],my.data[,3])
plot(my.data[1:150,2],my.data[1:150,3],xlab = "HeronIsland-Water Temperature",
    ylab="LadyMusgrave-Water Temperature",col="red")
```

Output:



- Here we see all the points are scattered variably.
- From the above figure it can be said that it has weak relationship as there are huge variations as the data points are widely spread across the plot

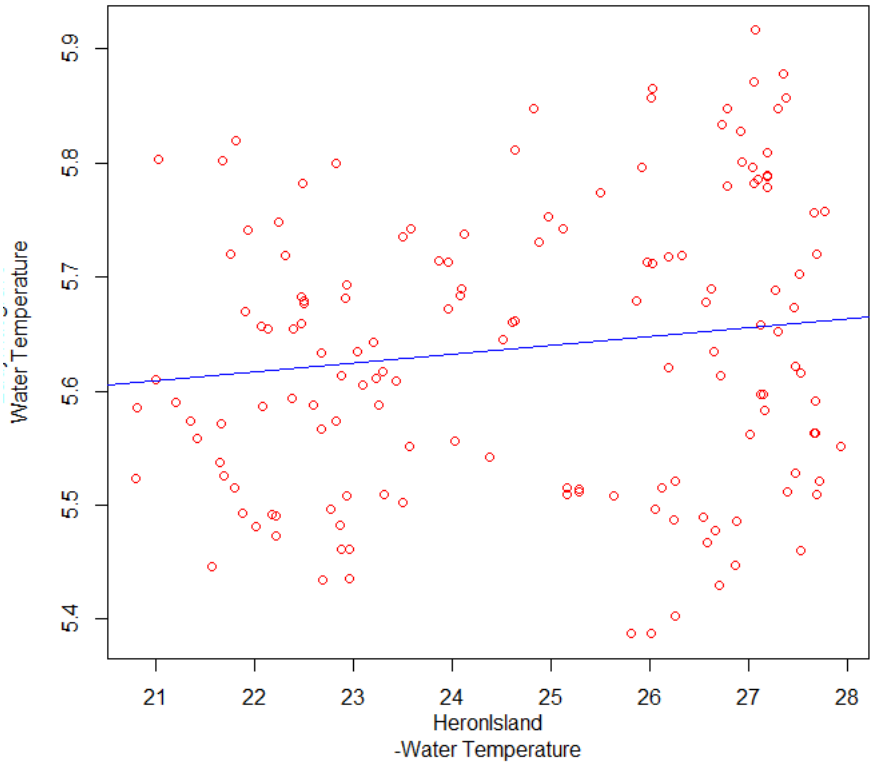
Equation of Linear regression line:
Y = 2.719*X +(9.256)
It's in the following form
$Y = a + bX$,
Where:
- X is the explanatory variable
- Y is the dependent variable.
- b is the slope of the line.
- a is the intercept (the value of y when $x = 0$).

Code:

```
lm(e[,1]~e[,2])#linear model fitting
abline(lm(e[,2]~e[,1]),col="blue")
```

Output:

```
> lm(e[,1]~e[,2])#linear model fitting

Call:
lm(formula = e[, 1] ~ e[, 2])

Coefficients:
(Intercept)        e[, 2]
      9.256         2.719

> abline(lm(e[,2]~e[,1]),col="blue")
>
```



Code:

```
correlation <- cor(e[,2],e[,1])
print(paste("Correlation Co-efficient = ",correlation))
coeffDet <- correlation^2
print(paste("Co-efficient of determination = ",coeffDet))
```

Output:

```
> correlation <- cor(e[,2],e[,1])
> print(paste("Correlation Co-efficient = ",correlation))
[1] "Correlation Co-efficient =  0.145512934511294"
> coeffDet <- correlation^2
> print(paste("Co-efficient of determination = ",coeffDet))
[1] "Co-efficient of determination =  0.0211740141100883"
>
```

→The correlation coefficient between the two is 0.145512934511294
And Co-efficient of determination = 0.0211740141100883.

- The result of "*Co-efficient of correlation*" is the strength of the linear relationship. The result is positive which means that temperatures are correlated.
- The result of "*Co-efficient of determination*" or the "squared correlation" gives the result of variance. In this case it is less than 0.2 which means that there is low variance.

Question:2)

2.1) the probability that the person is a Professional (P) is 52.96 %.
2.2) the probability that the person is below 30 years old is 42.59%.
2.3) person's occupation is community service (C) and the age is between 30 and 50 years old (M) is 0.037
2.4) Probability that the person is a sales worker (S) given that he/she is above 50 years old (A) is 0.114.

2.5) the probability that the person, who is a professional (P), is below 30 years old (B) is 0.104.

2.6) Probability that the person is a sales worker (S) or between 30 to 50 years old (M) is 0.644.

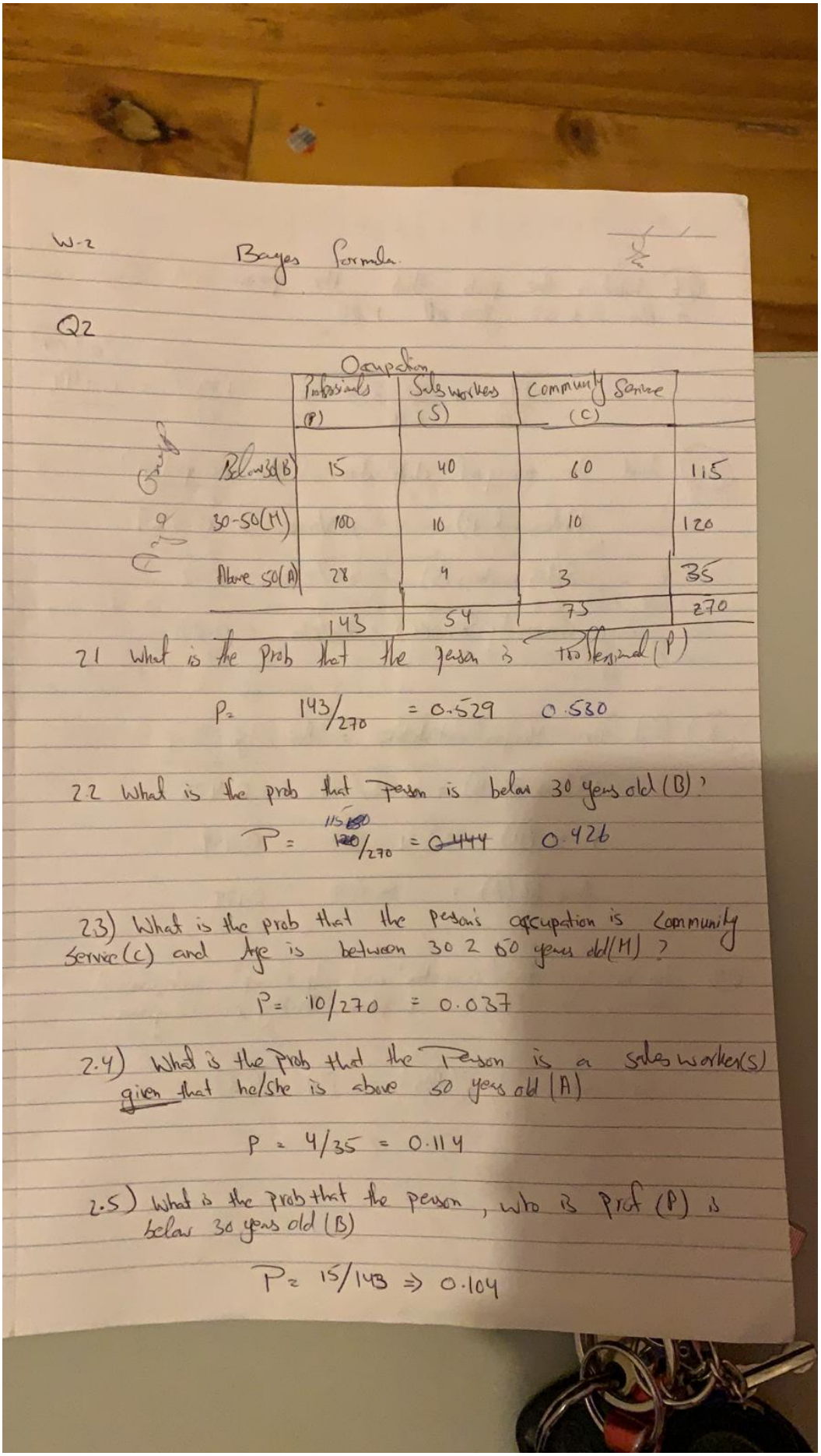
2.7) marginal distribution of the occupation:

- Professional = 0.529
- Sales Worker = 0.2
- Community Service = 0.270

2.8) marginal distribution of the Age Group:

- Professional = 0.425
- Sales Worker = 0.444
- Community Service = 0.129

2.9) Are the variables 'occupation' and 'Age group' independent random variables? Explain why or why not?!
Yes, they are as P(A intersection B) = P(A) X P(B).



W-2 Bayes Formula

Q2

	Occupation			
	Professionals (P)	Sales workers (S)	Community Service (C)	
Below 30 (B)	15	40	60	115
30-50 (M)	100	10	10	120
Above 50 (A)	28	4	3	35
	143	54	75	270

2.1 What is the prob that the person is professional (P)

$P = 143/270 = 0.529$ 0.580

2.2 What is the prob that person is below 30 years old (B)?

$P = 115/270 = 0.444$ 0.426

2.3) What is the prob that the person's occupation is Community Service (C) and age is between 30 & 50 years old (M)?

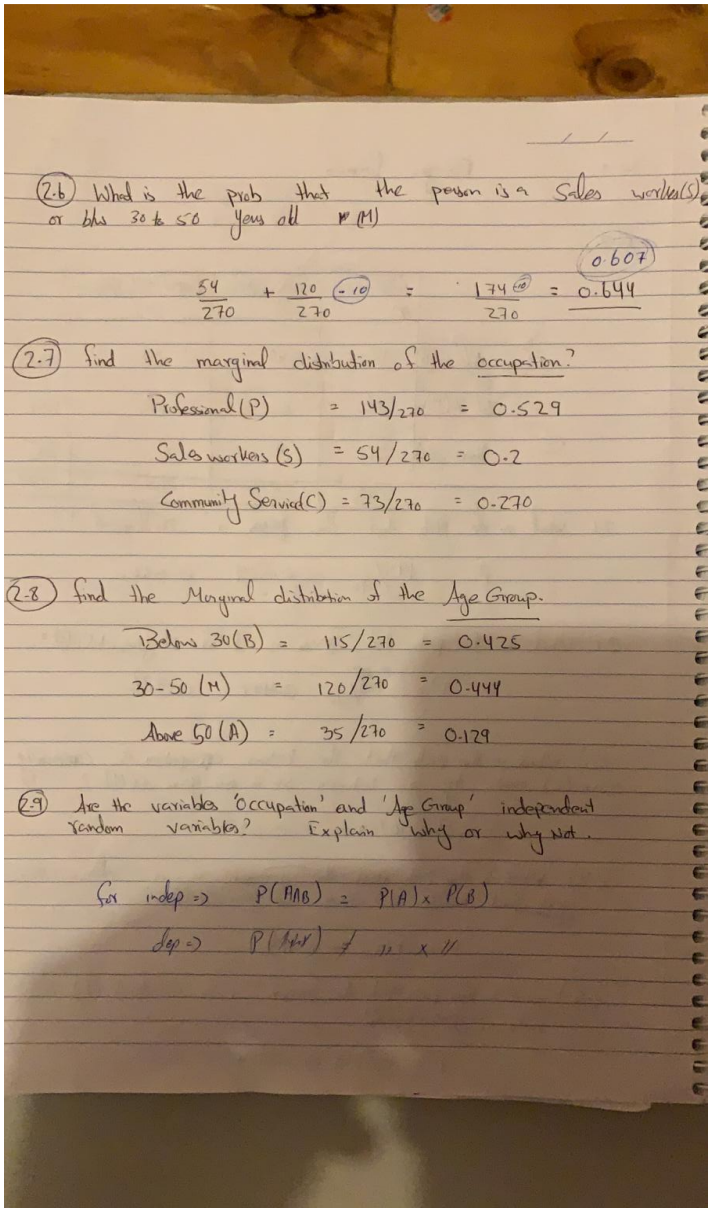
$P = 10/270 = 0.037$

2.4) What is the prob that the person is a sales worker (S) given that he/she is above 50 years old (A)

$P = 4/35 = 0.114$

2.5) What is the prob that the person, who is Prof (P) is below 30 years old (B)

$P = 15/143 \Rightarrow 0.104$



(2.6) What is the prob that the person is a Sales worker(S) or b/w 30 to 50 years old ∨ (M)

$$\frac{54}{270} + \frac{120}{270} - \frac{10}{270} = \frac{174}{270} = 0.644$$

0.607

(2.7) find the marginal distribution of the occupation?

Professional (P) = 143/270 = 0.529

Sales workers (S) = 54/270 = 0.2

Community Service(C) = 73/270 = 0.270

(2.8) find the Marginal distribution of the Age Group.

Below 30(B) = 115/270 = 0.425

30-50 (M) = 120/270 = 0.444

Above 50 (A) = 35/270 = 0.129

(2.9) Are the variables 'Occupation' and 'Age Group' independent random variables? Explain why or why not.

for indep => $P(A \cap B) = P(A) \times P(B)$

dep => $P(A \cap B) \neq P(A) \times P(B)$

Question:3)

a) overall proportion of the drones produced pass the quality test is 0.66 or 66%.
b) probability that it was produced by assembly line B is 0.364 or 36.46%.

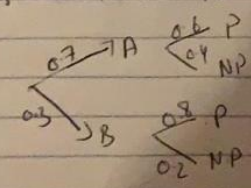
Q3 (6 Mark)

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 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 passed the quality test.?

0.42 + 0.24 = 0.660

b) If a randomly selected drone passed the quality test, what is the prob. that it was produced by assembly line B?



P(A,P) 0.7x0.6 = 0.42
P(A,NP) 0.7x0.4 = 0.28
P(B,P) 0.3x0.8 = 0.24
P(B,NP) 0.3x0.2 = 0.06

$= \frac{P(B/P)}{P(B)} = \frac{0.24}{0.42+0.24}$ = 0.364

Q3

(a) 70% x 60% + 30% x 80% = 66%

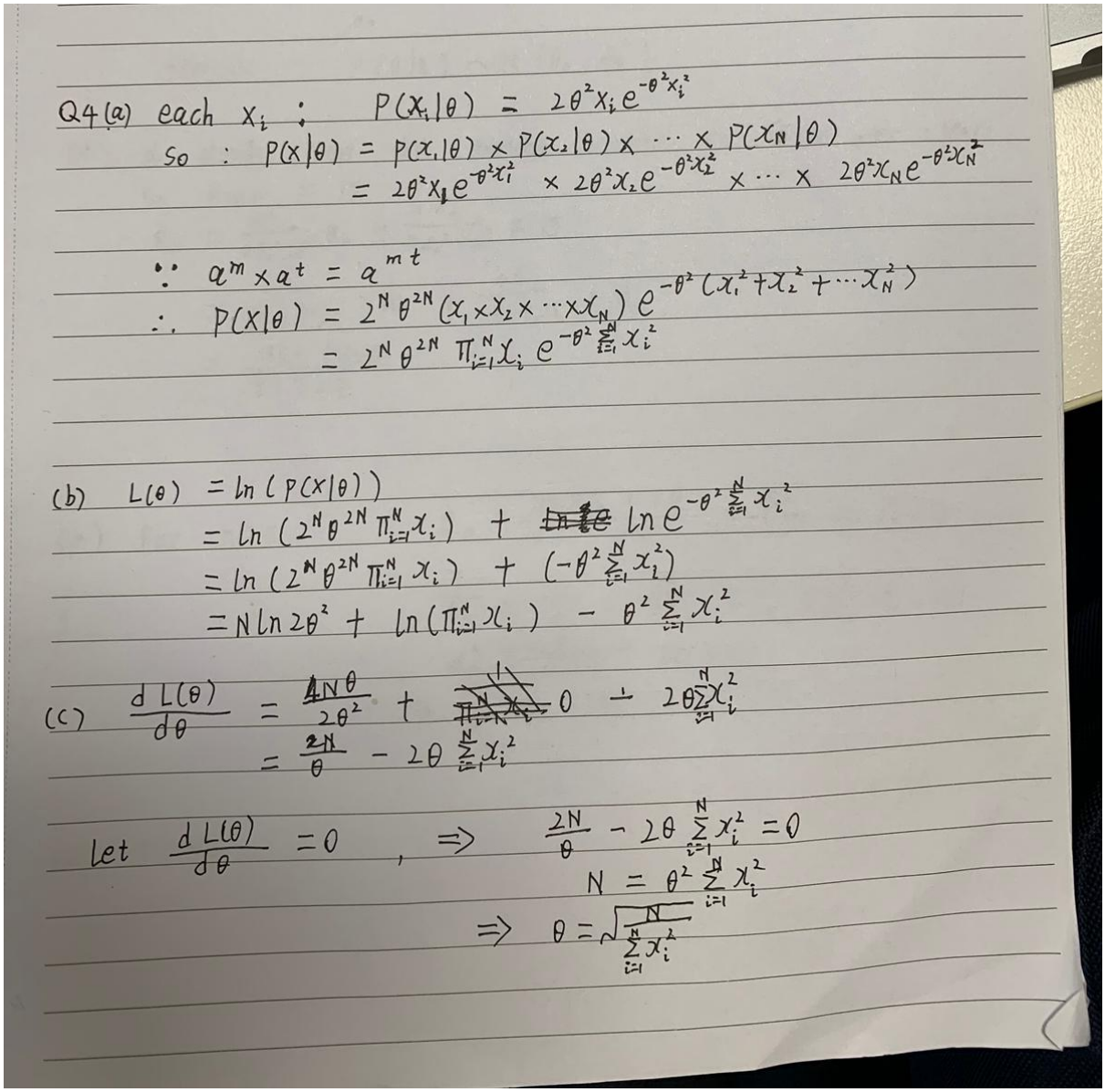
So there are 66% of the drones Produced pass the quality test.

(b) Let P represents ~~drone~~ the event of drone passed the quality test.

Let B represents drone that is produced by line B.

$P(B|P) = \frac{P(P|B) \times P(B)}{P(P)} = \frac{80\% \times 30\%}{66\%} \approx 0.364 = 36.4\%$.

Question:4)



Q4 (a) each x_i : $P(x_i|\theta) = 2\theta^2 x_i e^{-\theta^2 x_i^2}$

So : $P(x|\theta) = P(x_1|\theta) \times P(x_2|\theta) \times \cdots \times P(x_N|\theta)$

$= 2\theta^2 x_1 e^{-\theta^2 x_1^2} \times 2\theta^2 x_2 e^{-\theta^2 x_2^2} \times \cdots \times 2\theta^2 x_N e^{-\theta^2 x_N^2}$

$\because a^m \times a^t = a^{mt}$

$\therefore P(x|\theta) = 2^N \theta^{2N} (x_1 \times x_2 \times \cdots \times x_N) e^{-\theta^2 (x_1^2 + x_2^2 + \cdots x_N^2)}$

$= 2^N \theta^{2N} \prod_{i=1}^{N} x_i \, e^{-\theta^2 \sum_{i=1}^{N} x_i^2}$

(b) $L(\theta) = \ln(P(x|\theta))$

$= \ln(2^N \theta^{2N} \prod_{i=1}^{N} x_i) + \ln e^{-\theta^2 \sum_{i=1}^{N} x_i^2}$

$= \ln(2^N \theta^{2N} \prod_{i=1}^{N} x_i) + (-\theta^2 \sum_{i=1}^{N} x_i^2)$

$= N \ln 2\theta^2 + \ln(\prod_{i=1}^{N} x_i) - \theta^2 \sum_{i=1}^{N} x_i^2$

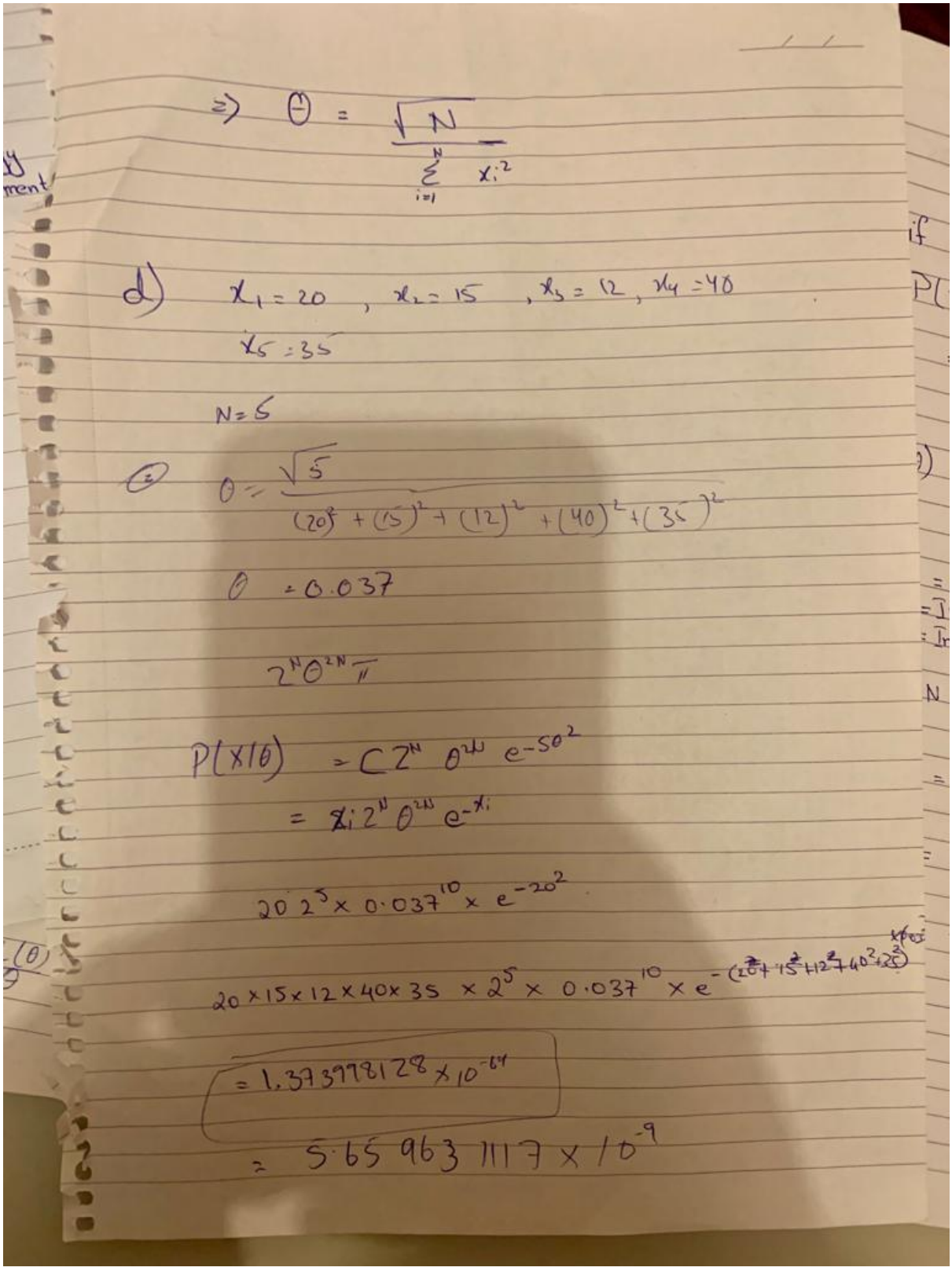
(c) $\frac{dL(\theta)}{d\theta} = \frac{4N\theta}{2\theta^2} + \frac{1}{\prod_{i=1}^{N} x_i} 0 - 2\theta \sum_{i=1}^{N} x_i^2$

$= \frac{2N}{\theta} - 2\theta \sum_{i=1}^{N} x_i^2$

Let $\frac{dL(\theta)}{d\theta} = 0$, $\Rightarrow \frac{2N}{\theta} - 2\theta \sum_{i=1}^{N} x_i^2 = 0$

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

$\Rightarrow \theta = \sqrt{\frac{N}{\sum_{i=1}^{N} x_i^2}}$



Question:5)

5.1:

A family ‘F’ of prior distributions **p(θ)** is conjugate to a likelihood **p(D|θ),** if the posterior

p(θ|D) is in F then it can be said as Conjugate prior.

5.2:
If prior belong to distribution(other than Bayesian) then we need to calculate integrals and update the variables where as if it's in the form of conjugate priors then we only need to update it.

5.3:
1. Gaussian – Gaussian Model
2. Beta is conjugate to Bernoulli Model
3. Dirichlet multinomial model

5.5)

In terms of mean and standard deviation we have the following values for both the sub questions:

(b) the mean is 13.214 and standard deviation is 1.793.

(c) the mean is 12.82 and standard deviation is 1.185.

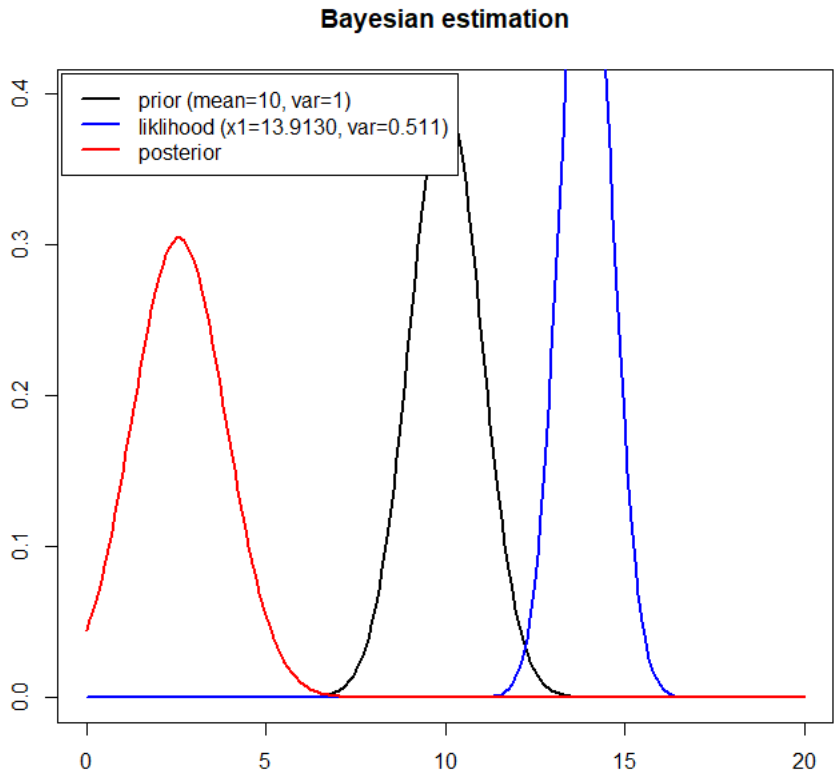
(d) for D we have the mean and standard deviation of posterior as 13.9130 and 0.715.

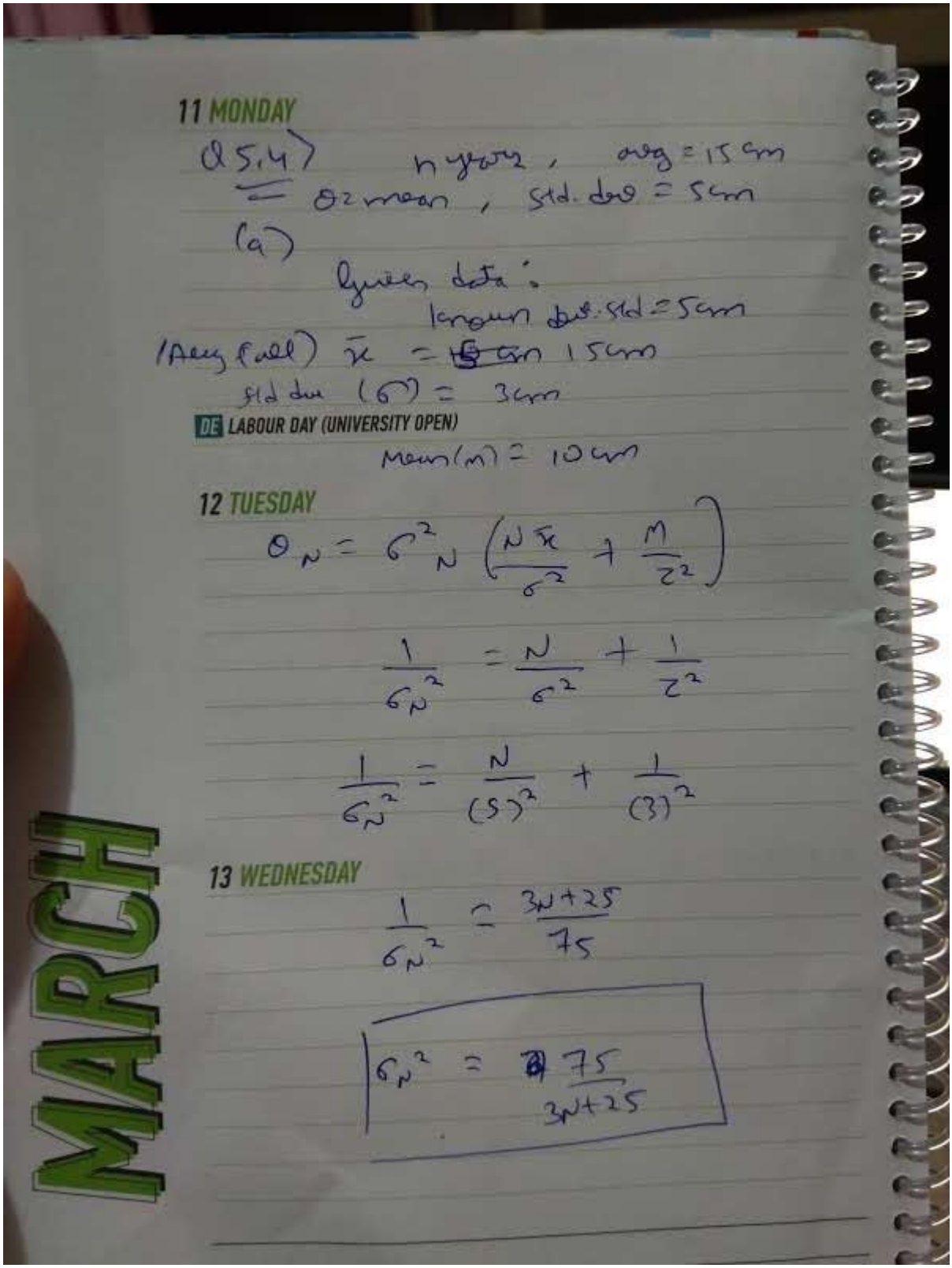
Code:

```
#Question 5

colors <- c("black", "blue", "red")

labels <- c("prior (mean=10, var=1)", "liklihood (x1=13.9130, var=0.511)", "posterior")

#prior

mean=10; sd=sqrt(1)

x <- seq(-10,10,length=200)*sd + mean

hx <- dnorm(x,mean,sd)

plot(x, hx, type="n", xlab="", ylab="", ylim=c(0, 0.4),main="Bayesian estimation",axes=TRUE)

lines(x, hx, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors[1])

#liklihood

mean1=13.9130; sd1=0.715

hx <- dnorm(x,mean1,sd1)

lines(x, hx,lwd=2, col=colors[2])

legend("topleft", inset=.005,
```

```
    labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors)
#posterior
mean2=18/7; sd2=sqrt(12/7)
hx <- dnorm(x,mean2,sd2)
lines(x, hx,lwd=2, col=colors[3])
legend("topleft", inset=.005,
    labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors)
```

Output:





11 MONDAY

(Q 5.4) n years, avg = 15 cm
= 0 z mean, Std. dev = 5 cm

(a)

Given data:

known dev. Std = 5 cm

Avg (all) $\bar{x}$ = ~~5 cm~~ 15 cm

Std dev (σ) = 3 cm

DE LABOUR DAY (UNIVERSITY OPEN)

Mean (μ) = 10 cm

12 TUESDAY

$$\theta_N = \sigma^2_N \left(\frac{N\bar{x}}{\sigma^2} + \frac{\mu}{z^2} \right)$$

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{z^2}$$

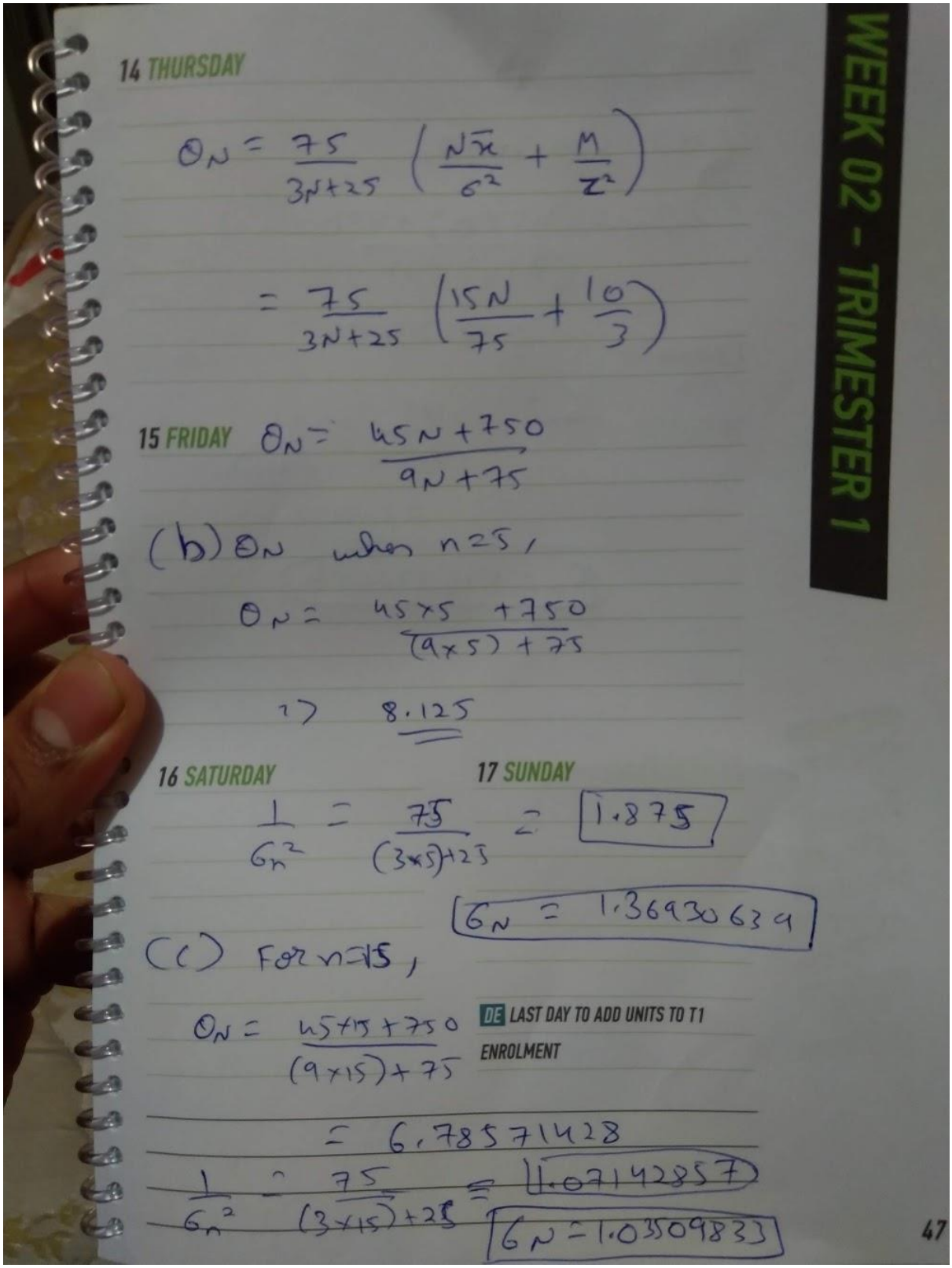
$$\frac{1}{\sigma_N^2} = \frac{N}{(5)^2} + \frac{1}{(3)^2}$$

13 WEDNESDAY

$$\frac{1}{\sigma_N^2} = \frac{3N+25}{75}$$

$$\sigma_N^2 = \frac{75}{3N+25}$$

MARCH



14 THURSDAY

$$\theta_N = \frac{75}{3N+25}\left(\frac{N\bar{x}}{\sigma^2} + \frac{M}{z^2}\right)$$

$$= \frac{75}{3N+25}\left(\frac{15N}{75} + \frac{10}{3}\right)$$

15 FRIDAY $\theta_N = \frac{45N+750}{9N+75}$

(b) θ_N when $n=5$,

$$\theta_N = \frac{45\times5+750}{(9\times5)+75}$$

$$\Rightarrow 8.125$$

16 SATURDAY

17 SUNDAY

$$\frac{1}{6_n^2} = \frac{75}{(3\times5)+25} = 1.875$$

$6_N = 1.36930634$

(c) For $n=15$,

$$\theta_N = \frac{45\times15+750}{(9\times15)+75}$$

DE LAST DAY TO ADD UNITS TO T1 ENROLMENT

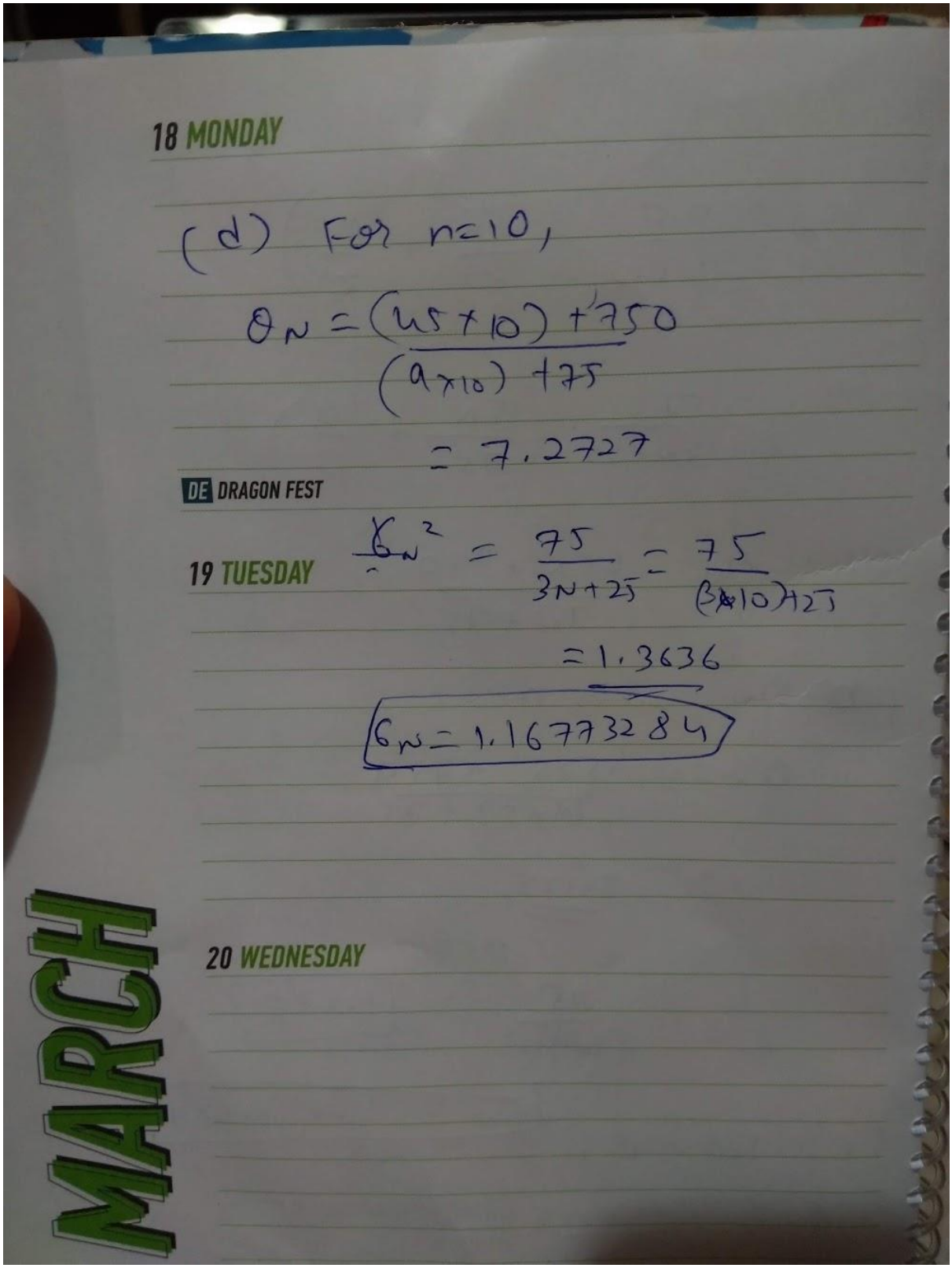
$$= 6.78571428$$

$$\frac{1}{6_n^2} = \frac{75}{(3\times15)+25} = 1.07142857$$

$6_N = 1.0350983$

WEEK 02 - TRIMESTER 1

47



18 MONDAY

(d) For n=10,

$$\theta_N = \frac{(45 \times 10) + 750}{(9 \times 10) + 75}$$

$$= 7.2727$$

DE DRAGON FEST

19 TUESDAY

$$\sigma_N^2 = \frac{75}{3N + 25} = \frac{75}{(3 \times 10) + 25}$$

$$= 1.3636$$

$$\sigma_N = 1.1677328 4$$

20 WEDNESDAY

MARCH

Question:6)

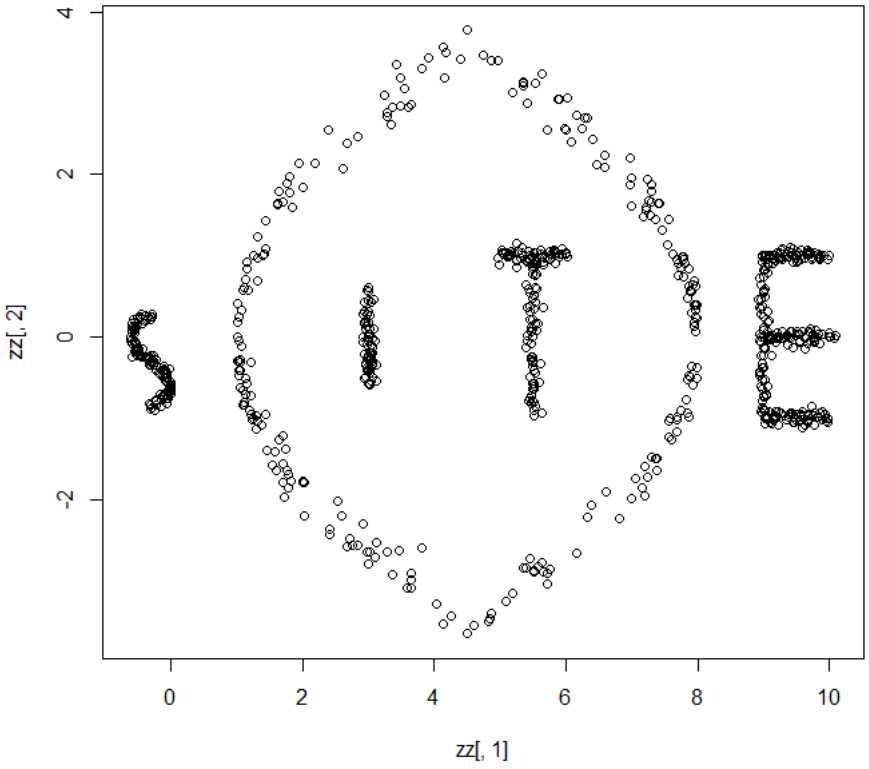
6.1) **Code:**

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

a) **Code:**

```
zz
plot(zz[,1],zz[,2])
```

Output:

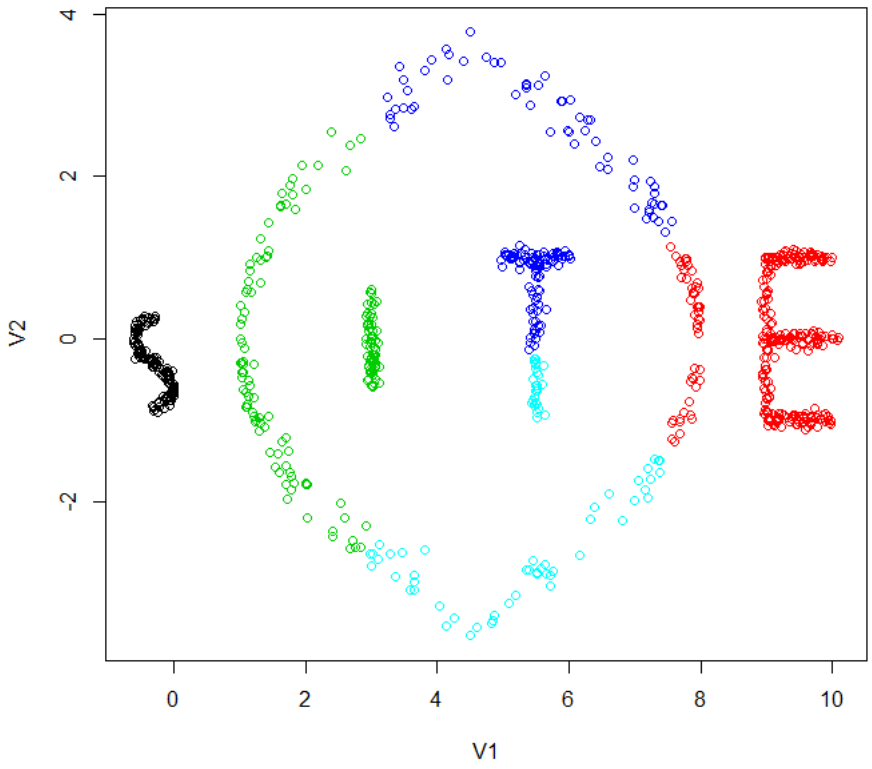


b) There are 5 clusters to be found in the plot by visual examination of the scatterplot.

c) **Code:**

```
#K means
library(stats)
km <- kmeans(zz, centers=5, nstart=5)
plot(zz, col=km$cluster)
```

Output:



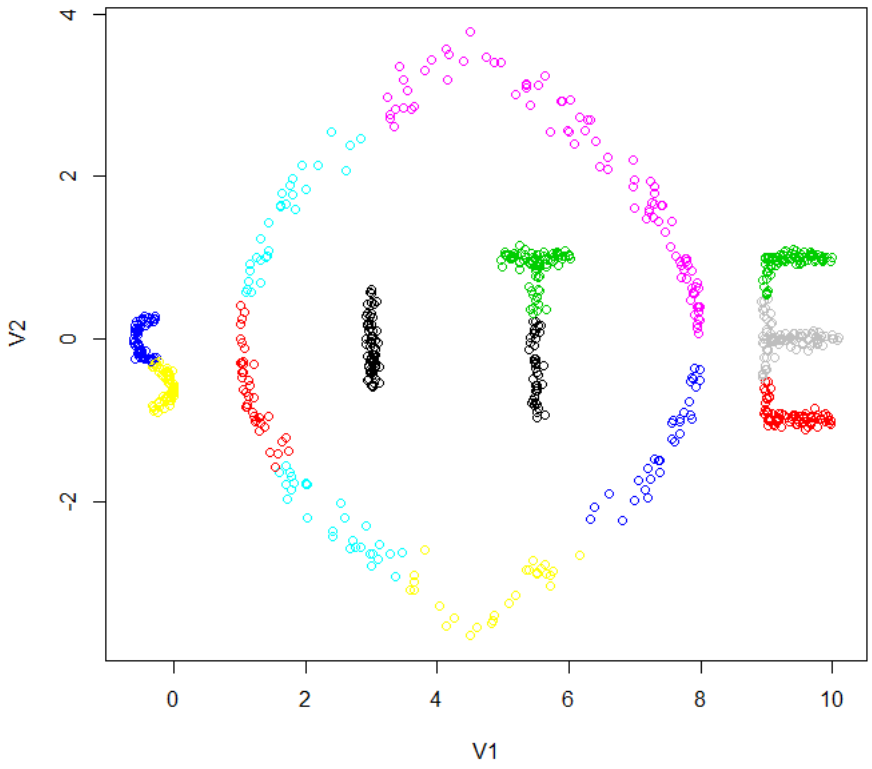
From this we can see that data points are arranged in terms of different colours each representing a cluster when we start with assumed centres to be 5.Here we see that points can be vivid in terms of final arrangement of different clusters according to black, green, blue, light blue and red colour.

d) Code:

```
#K means
library(stats)
km <- kmeans(zz, centers=15, nstart=5)
plot(zz, col=km$cluster)
```

Output:

For 15 clusters

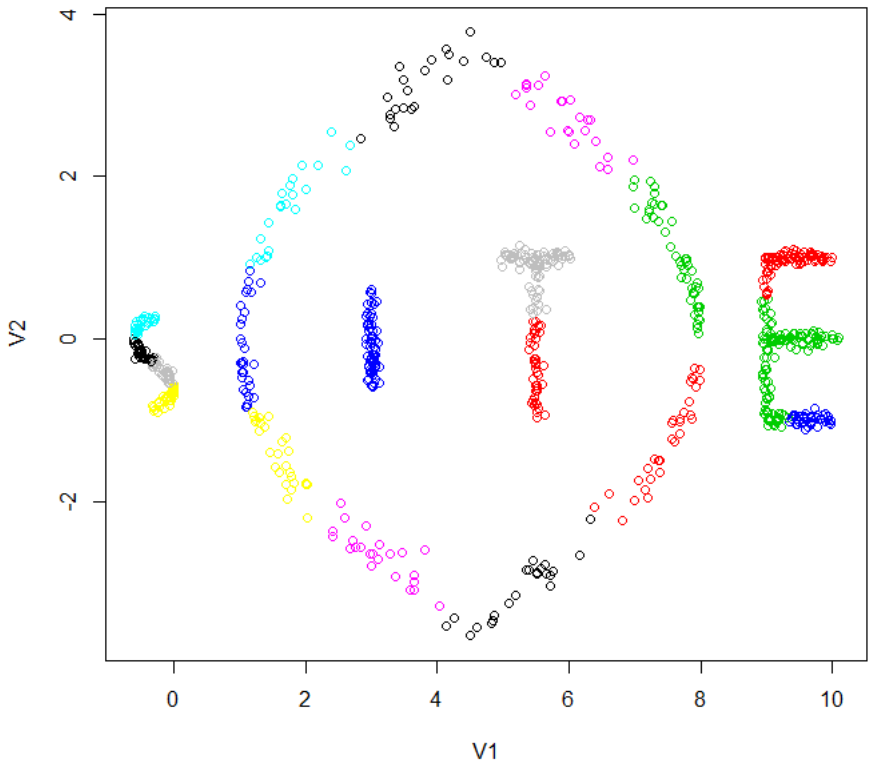


Code:

```
#K means
library(stats)
km <- kmeans(zz, centers=20, nstart=5)
plot(zz, col=km$cluster)
```

Output:

For 20 clusters

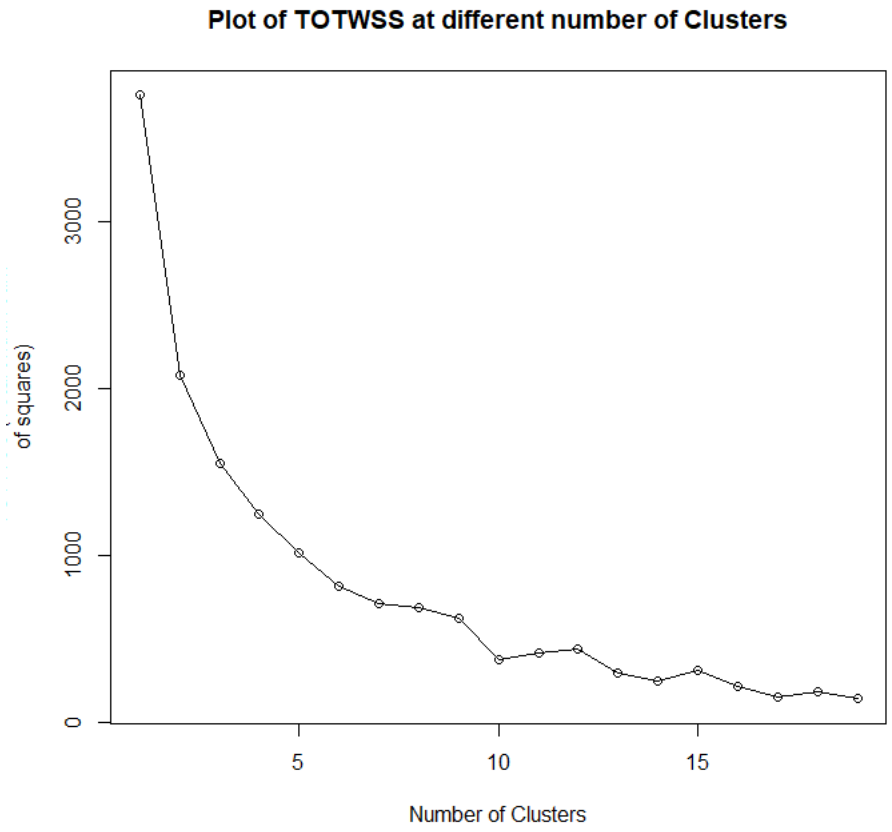


Here we see as we form more and more clusters the data becomes more specialised and more colours obtained on the graph represents more clusters formed.

Code for Total_within_sum_of_squares_(TOTWSS) value for each k:

```
r <- 2:20
gph <- c()
for (i in r) {
  cluster <- kmeans(zz,i)
  gph <- rbind(gph,c(cluster$tot.withinss))
}
plot(gph,xlab = "Number of Clusters",xlim=c(1,length(r)),ylab =
"TOTWSS (Total within sum
of squares)",main = "Plot of TOTWSS at different number of
Clusters",type="o")
```

Output:



→The within-cluster sum of squares is a measure of the variability of the observations within
each cluster and k=16 cluster formation is the most compact when compared to other K values.

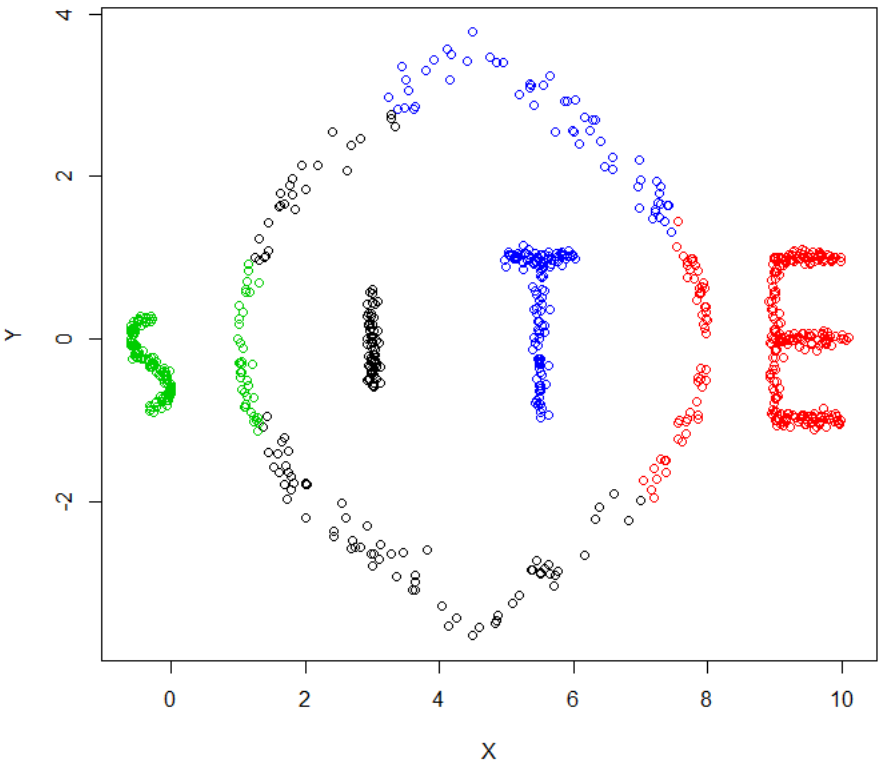
6.2)Spectral Clustering:-

Code:

```
#Task 6.2
#Spectral Clustering
install.packages("kernlab")
library(kernlab)
specClust<-specc(zz, centers=4)
plot(zz, col=specClust,xlab='X',ylab="Y")
```

{ or alternatively the long method given in comments in R

Output:



• **Spectral clustering:** data points as nodes of a connected graph and clusters are found by partitioning this graph, based on its spectral decomposition, into subgraphs.

• **K-means clustering:** divide the objects into k clusters such that some metric relative to the centroids of the clusters is minimized.

Here we can see that clusters are more accurately and significantly represented in spectral due to their spectral decomposition whereas k-means merely divides them in clusters of value k and then finalizes them to reap similar results.
I believe spectral clustering is more better arrangement to the data representation.
While KMeans uses Euclidian distance and split the data points such that it is relative to
centroid on the other hand Spectral Clustering divides the graph into subgraph. Which in
turn is more effective approach.

Question:7)

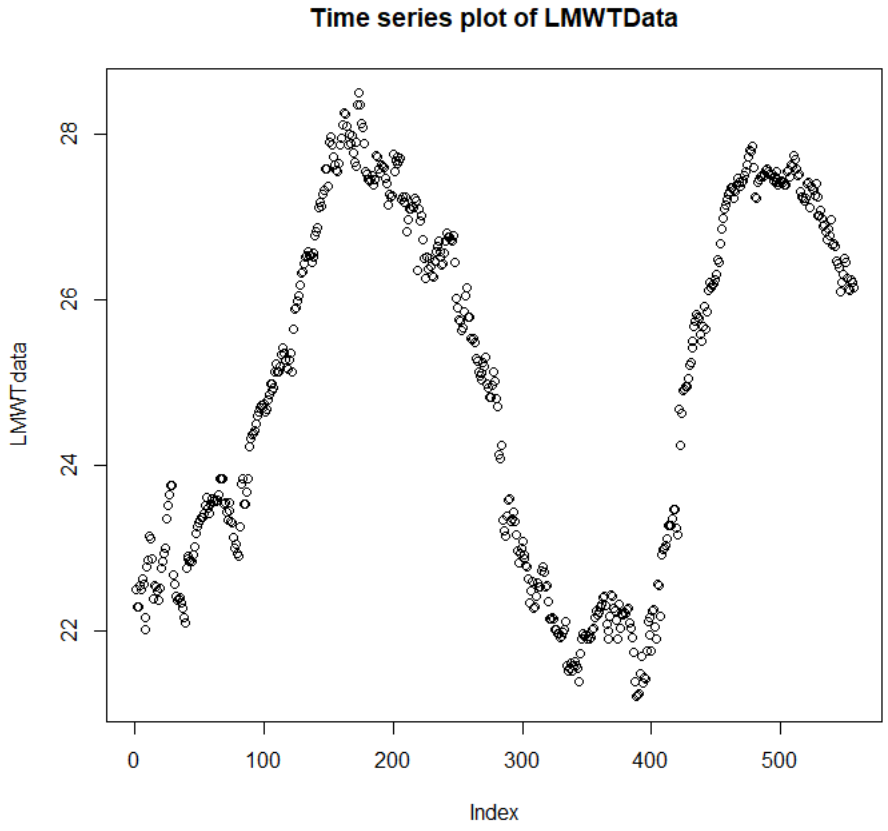
Code:

```
#Question 7
the.data <-as.matrix(read.csv("GBRData.csv", header =TRUE, sep = ","))
#extract the 'LadyMusgrave-Water Temperature'  values
LMWTdata <-the.data[,4]
```

7.1) Code:

```
plot(LMWTdata,main="Time series plot of LMWTData")
```

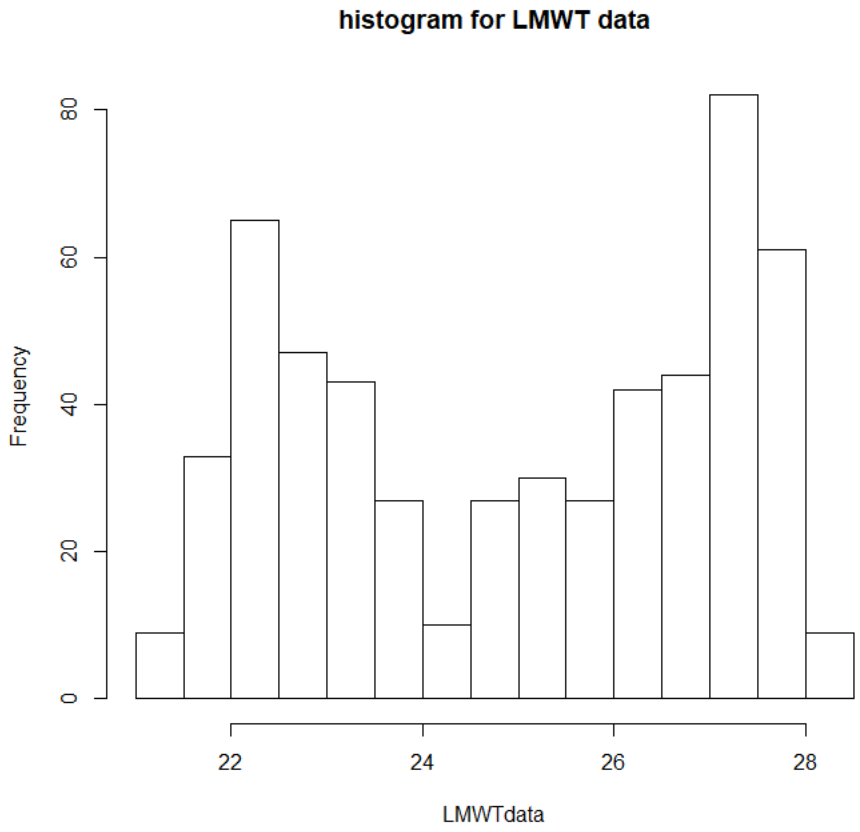
Output:



7.2) **Code:**

```
hist(LMWTdata,main = "histogram for LMWT data")
```

Output:

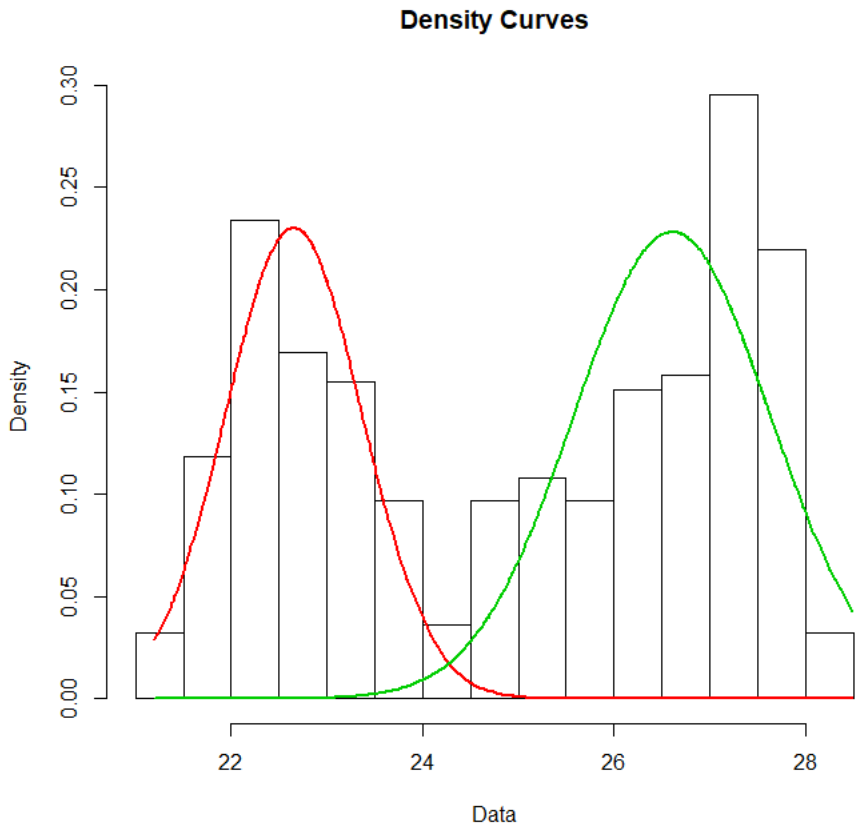


Here we can observe that it is bimodal as there are two peaks at different points in the graph at both ends left and right.

7.3) **Code:**

```
library(MASS)
fit1<-fitdistr(LMWTdata,"normal")
fit1
mixmdl = normalmixEM(LMWTdata)
mixmdl
summary(mixmdl)
plot(mixmdl,which=2)
```

Output:



7.4) **Code:**

```
lines(density(LMWTdata), lty=2, lwd=2)
mixmdl$lambda
mixmdl$mu
mixmdl$sigma
```

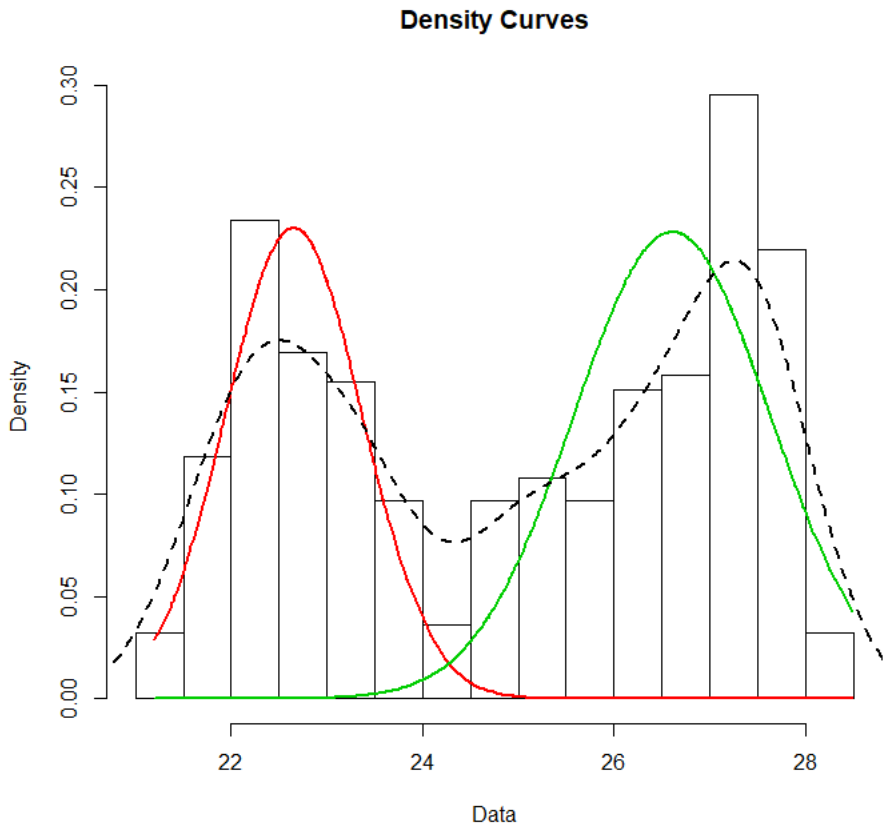
Output:

```
> mixmdl$lambda
[1] 0.4114853 0.5885147
> mixmdl$mu
[1] 22.65382 26.60726
> mixmdl$sigma
[1] 0.7134128 1.0293173
>
```

7.5) **Code:**

```
lines(density(LMWTdata), lty=2, lwd=2)
```

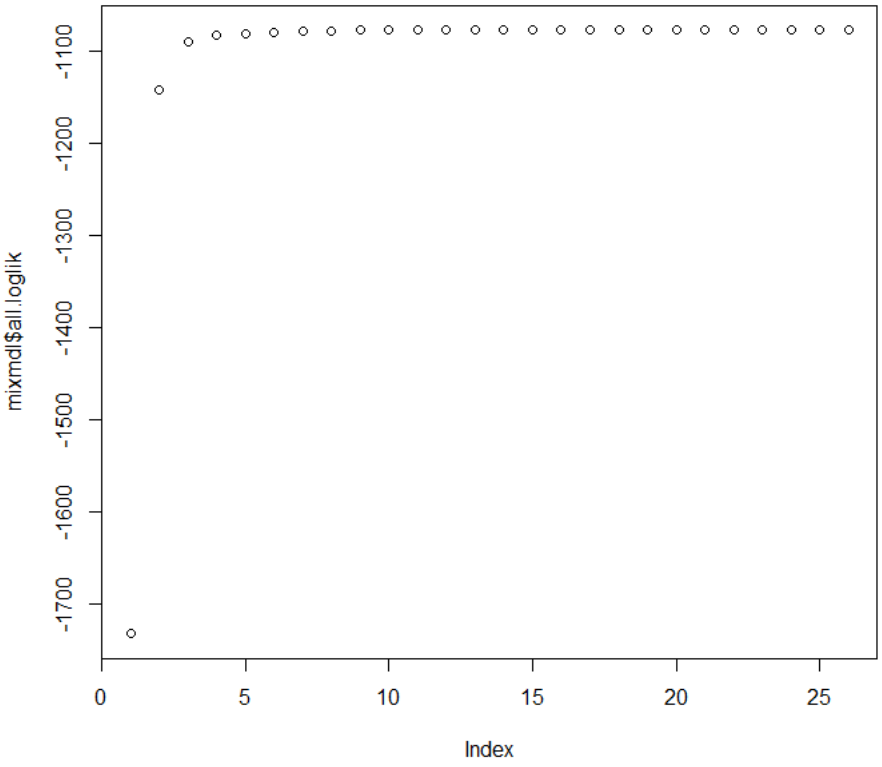
Output:



7.6) **Code:**

```
plot(mixmdl$all.loglik)
```

Output:

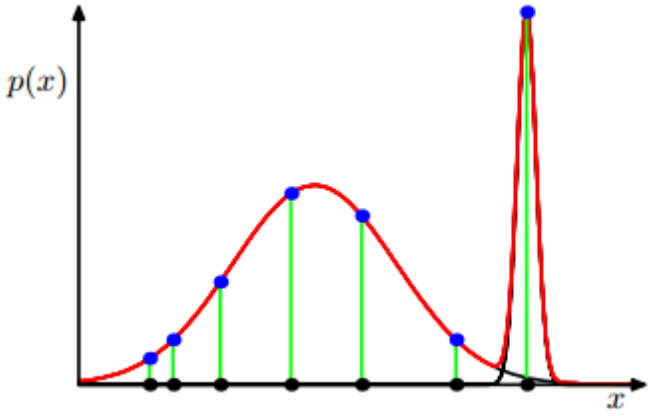


7.7) the distribution modes obtained in Q7.3 and Q7.4 show gaussian fitting of the graphical data in various representations.
We see one represents two curves whereas other one represents one curve with two peaks.
I think mixture of Gaussians models is more ideal method coz it fits the whole data and shows it in a streamlined and univariate manner which helps in seeing the entity as generalised and whole and makes predictions and analysis easy.

7.8)The main problem that you might come across when performing a maximum likelihood estimation using mixture of Gaussian is singularity.If we want to fit a Gaussian to a single data point using maximum likelihood, we will get a very spiky Gaussian that "collapses" to that point. The variance is zero when there's only one point, which in the multi-variate Gaussian case, leads to a singular covariance matrix, so it's called the singularity problem.

When the variance gets to zero, the likelihood of the Gaussian component goes to infinity and the model becomes overfitted. This doesn't occur when we fit only one Gaussian to a number of points since the variance cannot be zero. But it can happen when we have a mixture of Gaussians.

Illustration of how singularities in the likelihood function arise with mixtures of Gaussians. This should be compared with the case of a single Gaussian shown in Figure 1.14 for which no singularities arise.



The solutions for such a problem are:-

1) resetting the mean and variance when singularity occurs

seek local maxima of the likelihood function that are well behaved. We can hope to avoid the singularities by using suitable heuristics, for instance by detecting when a Gaussian component is collapsing and resetting its mean to a randomly chosen value while also resetting its covariance to some large value, and then continuing with the optimization.

2) using MAP instead of MLE by adding a prior.

The EM algorithm can also be used to find MAP (maximum posterior) solutions for models in which a prior $p(\boldsymbol{\theta})$ is defined over the parameters. In this case the E step remains the same as in the maximum likelihood case, whereas in the M step the quantity to be maximized is given by $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \ln p(\boldsymbol{\theta})$. Suitable choices for the prior will remove the singularities of the kind illustrated in Figure 9.7.