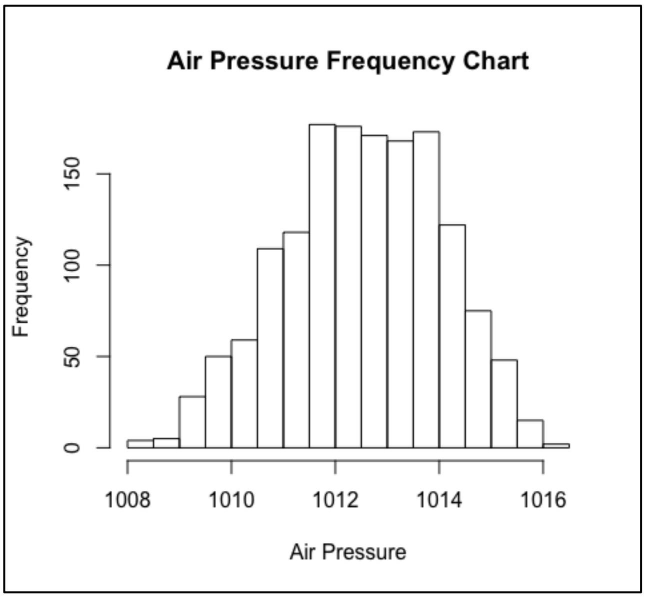
SIT743 Bayesian Learning and Graphical Models

Assignment-1

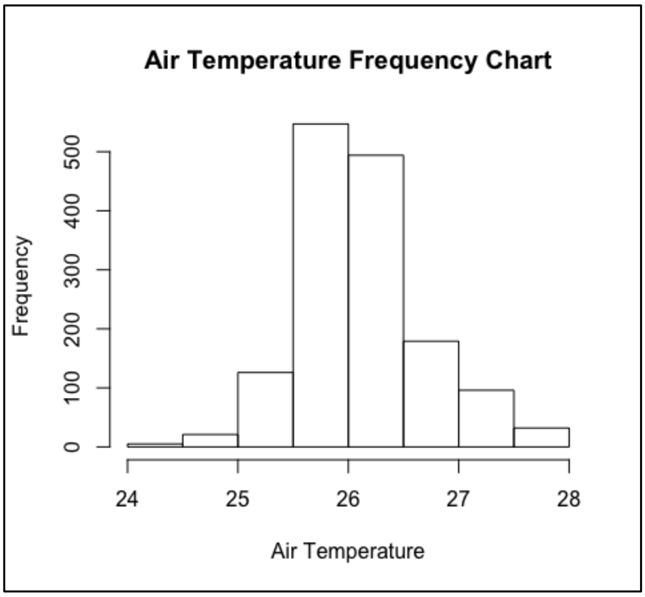
Q.1)

1.1)

This histogram shows a symmetric distribution with no outliers showing the air pressure from 1008 to 1016.



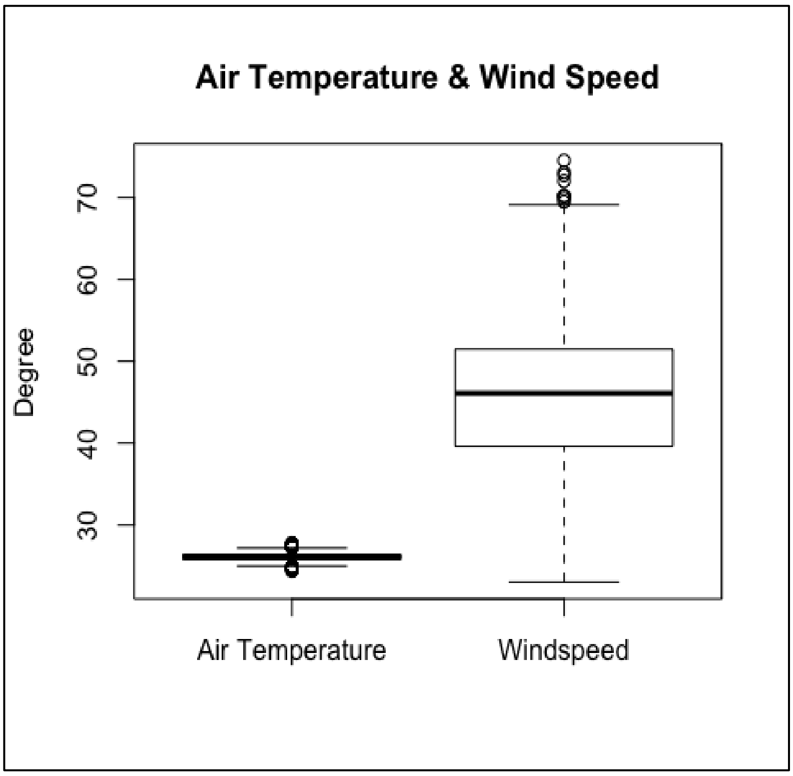
We could say that this histogram fairly shows a normal distribution with the minimum air temperature being 24 and max temp being between 25-26.



1.2)

Seeing the boxplot below it could be seen that air temperature was less than the windspeed, and the main windspeed was between 40-50 with having outliers as it goes above. Air temperature also had outliers, but that could be mainly because fitting both features in a single plot wouldn't have been efficient as they both use different units. We can see the summary measure for both of the features below. Also, we could see that the average air temperature was 26 whereas the average wind speed was 46. Windspeed can go a lot higher than the air temperature as seeing the max values go till 73. We could also see the interquartile range of windspeed is 12 whereas its less than 1 for Air temperature.

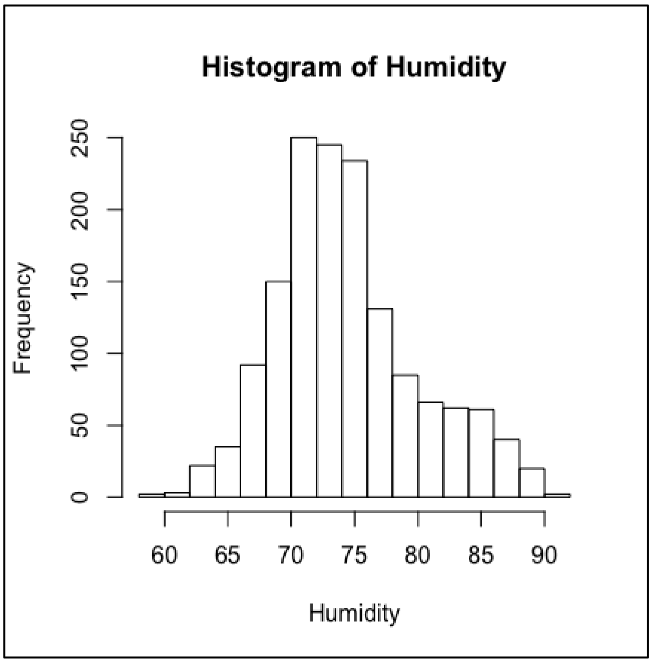
Summary Measure	**Air Temperature**	**Wind Speed**
Min	24.20	23.04
1st Quartile	25.80	40.32
Median	26.10	46.44
Mean	26.12	46.47
3rd Quartile	26.40	52.20
Max	27.80	73.08



1.3)

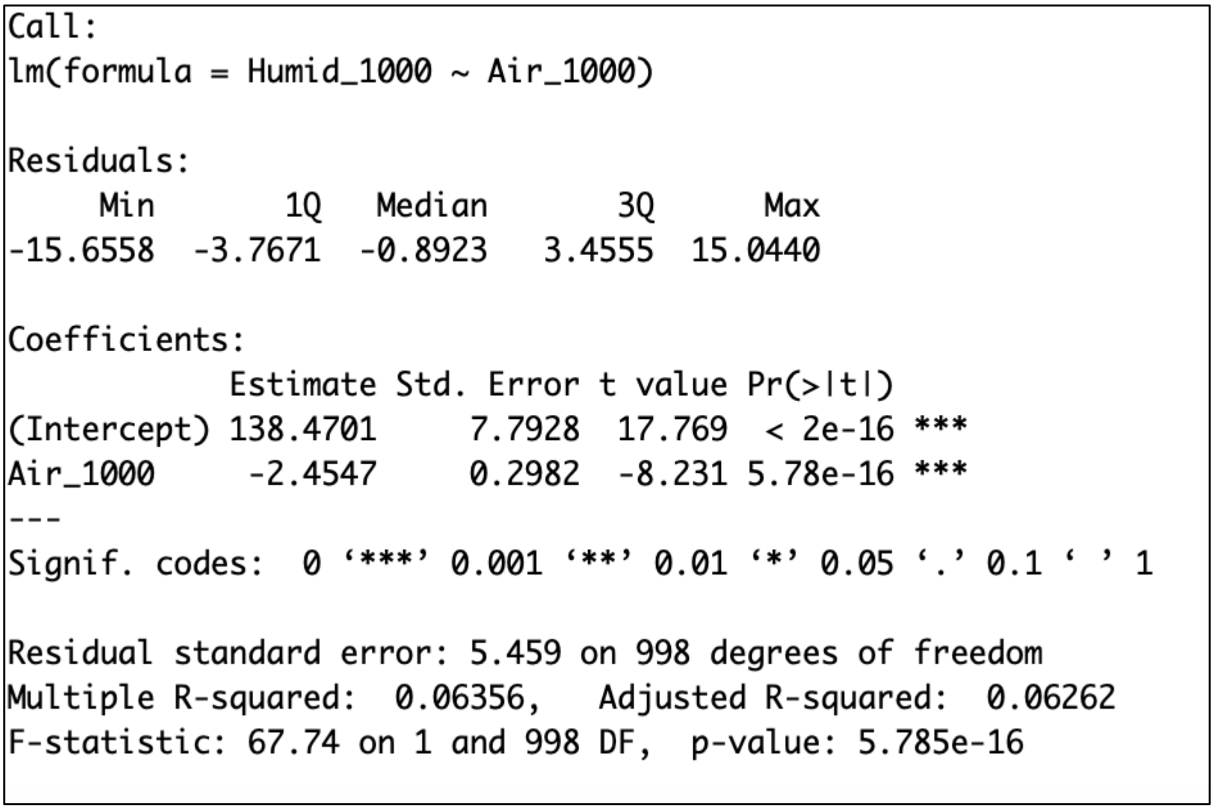
I would choose the *median* and *interquartile range* as the central tendency to interpret the centre and spread of the graph for Humidity. The reason to use the median was because we could see the centre of the humidity's distribution lies around the centre of the min and max values for humidity that is 70 – 80. Also, the reason we chose median for the spread was because 50 % of the data lies below the centre and the rest above it. The spread of the distribution as measured by the range is 6.62, from the summary measure statistics below.

Summary Measure	**Air Temperature**
Min	58.50
1st Quartile	70.60
Median	73.70
Mean	74.44
3rd Quartile	77.22
Max	91.90



1.4)

These were the results after fitting a lm model for Humidity and Air Temperature



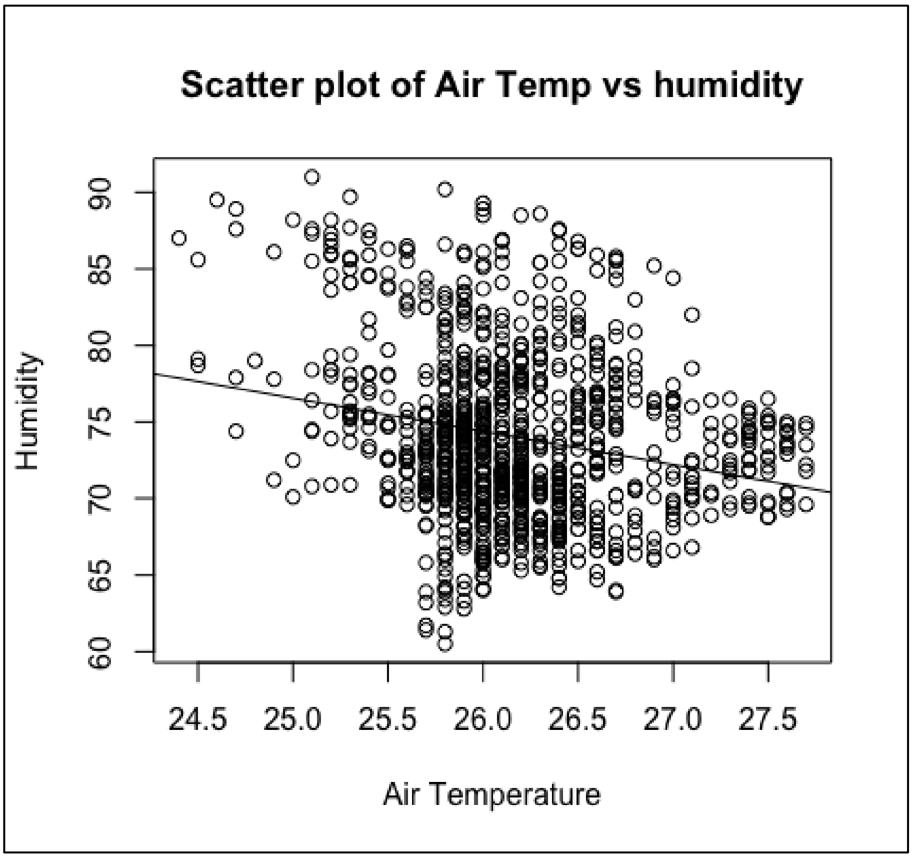
```
Call:
lm(formula = Humid_1000 ~ Air_1000)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.6558  -3.7671  -0.8923   3.4555  15.0440 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 138.4701     7.7928  17.769  < 2e-16 ***
Air_1000     -2.4547     0.2982  -8.231 5.78e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.459 on 998 degrees of freedom
Multiple R-squared:  0.06356,	Adjusted R-squared:  0.06262 
F-statistic: 67.74 on 1 and 998 DF,  p-value: 5.785e-16
```

The linear Regression equation could be as followed

$\hat{y}i$=138.5 – 2.4547xi + $\hat{\varepsilon}i$.   where $\varepsilon\sim$⍰(0, 5.459^2)

Looking at the output of the summary we can see the coefficients for the model fitted. Also, I computed the correlation for the Humidity and Air Temperature which resulted in giving a value of **-0.2521169,** telling us these both variables have a weak downhill (negative) linear relationship.

Also, two versions of r-squared tell us how much of the variation of the response variable is explained by our predictors, and not by error. In our case, the model explains around 6% of the variation of percent of Air Temperature. It also could be said that 6% of the variation can in the data can be explained by the linear relationship between Air Temperature and Humidity. The other 94% remains unexplained which is not a statistic.


Scatter plot of Air Temp vs humidity
Humidity
60
65
70
75
80
85
90
24.5
25.0
25.5
26.0
26.5
27.0
27.5
Air Temperature

Q.2)

2.1)

a) P(Victoria) = 3500/10000 => 0.35

b) P(Cricket & NSW) = 1400/10000 => 0.14

c) P(Footy give from Queensland) = 1300/2600 => 0.5

d) P(Basketball is from Victoria) = 500/3500 => 0.142

e) P(Victoria or Cricket) = P(Victoria) + P(cricket) - P(Victoria and Cricket) = 0.35 + 0.32 – 0.1 => 0.57

f) Marginal Distribution of Sports

		State			**Total**	**Marginal Distribution**
		NSW	**VIC**	**Queensland**		
Sports	**Footy**	1000	2000	1300	4300	4300/10000 = 0.43
	Basketball	1500	500	500	2500	2500 / 10000 = 0.25
	Cricket	1400	1000	800	3200	3200/10000 = 0.32
	Total	3900	3500	2600	**10000**	

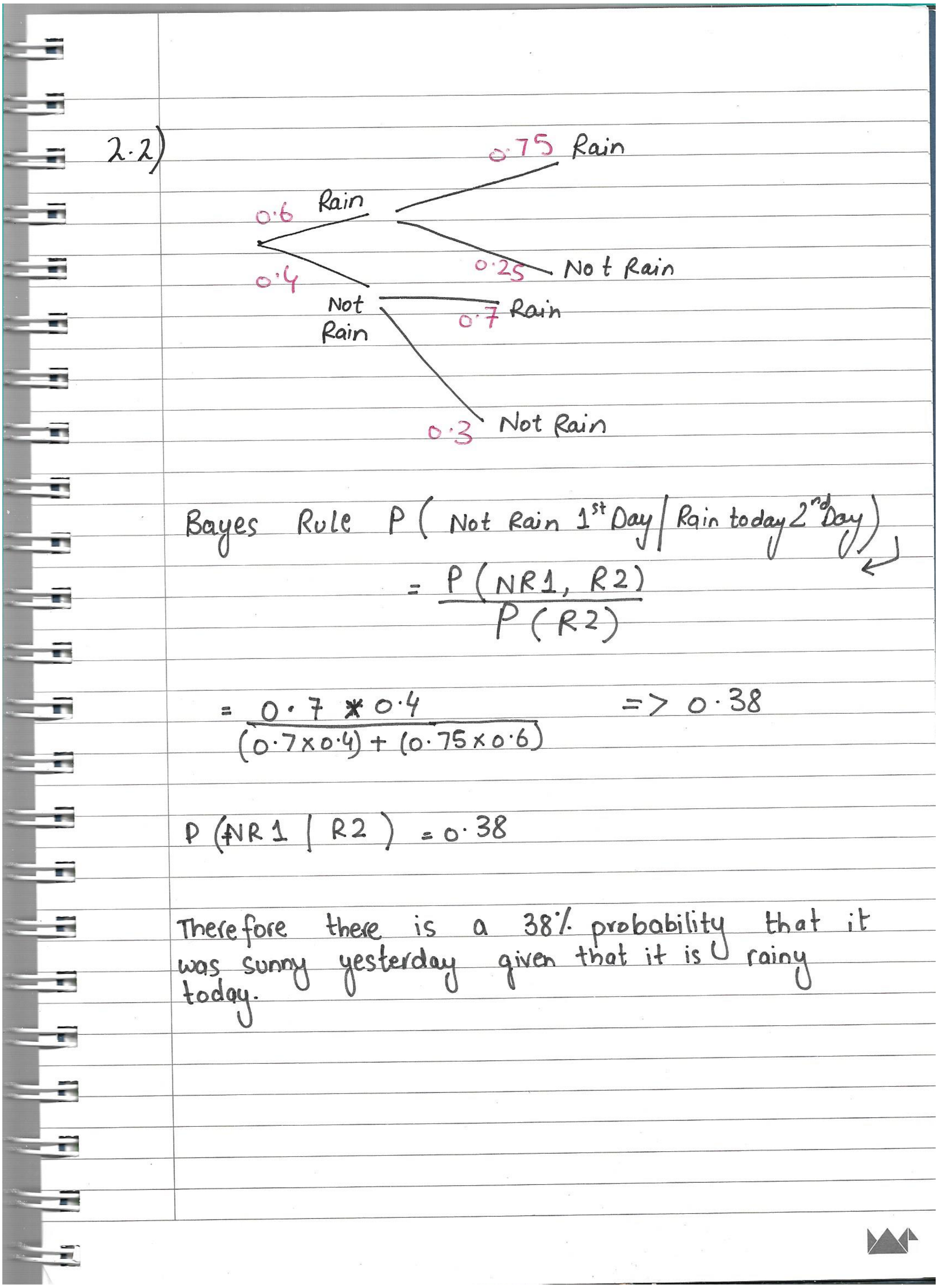
g) State and Sports are not mutually exclusive since they can happen both at the same time for example let's take the case of a person who likes to play footy and belongs to NSW = P(Footy and NSW) = 1000/3900 => 0.26 which is not equal to zero, thus stating they are not mutually exclusive.

h) Sate and Sports are independent events because of the outcome of State doesn't affects the outcome of sports. We can prove that by taking the example of VIC and Cricket. Let's compute the P(Victoria and Cricket) = P(Victoria)*P(Cricket)

P(V & C) = 1000/10000 => 0.1
P(V)*P(C) = (3500/10000) * (3200/10000) = 0.1

Therefore, since both of them are the same it could be said that State and sports are independent events.

2.2)



2.2)

0.6
Rain
0.4
Not Rain
0.75
Rain
0.25
Not Rain
0.7
Rain
0.3
Not Rain

Bayes Rule $P(\text{Not Rain } 1^{st}\text{ Day} \mid \text{Rain today } 2^{nd}\text{ Day})$

$$= \frac{P(NR1, R2)}{P(R2)}$$

$$= \frac{0.7 * 0.4}{(0.7 \times 0.4) + (0.75 \times 0.6)} \Rightarrow 0.38$$

$$P(NR1 \mid R2) = 0.38$$

Therefore there is a 38% probability that it was sunny yesterday given that it is rainy today.

Q.3)

3.1)

A frequentist inference way is considered to be a fixed parameter that is determined by an estimator. Frequentist use probability only to model certain processes. MLE is the most common used estimator for this inference. Whereas, Bayesian inference is a random variable described by a probability distribution, and they use probability more widely to model both uncertainty and sampling.

Moreover, frequentist inference considers that the sampling is infinite and therefore decision rules can be sharp, whereas for Bayesians unknown quantities are treated and can always be updated.

3.2)

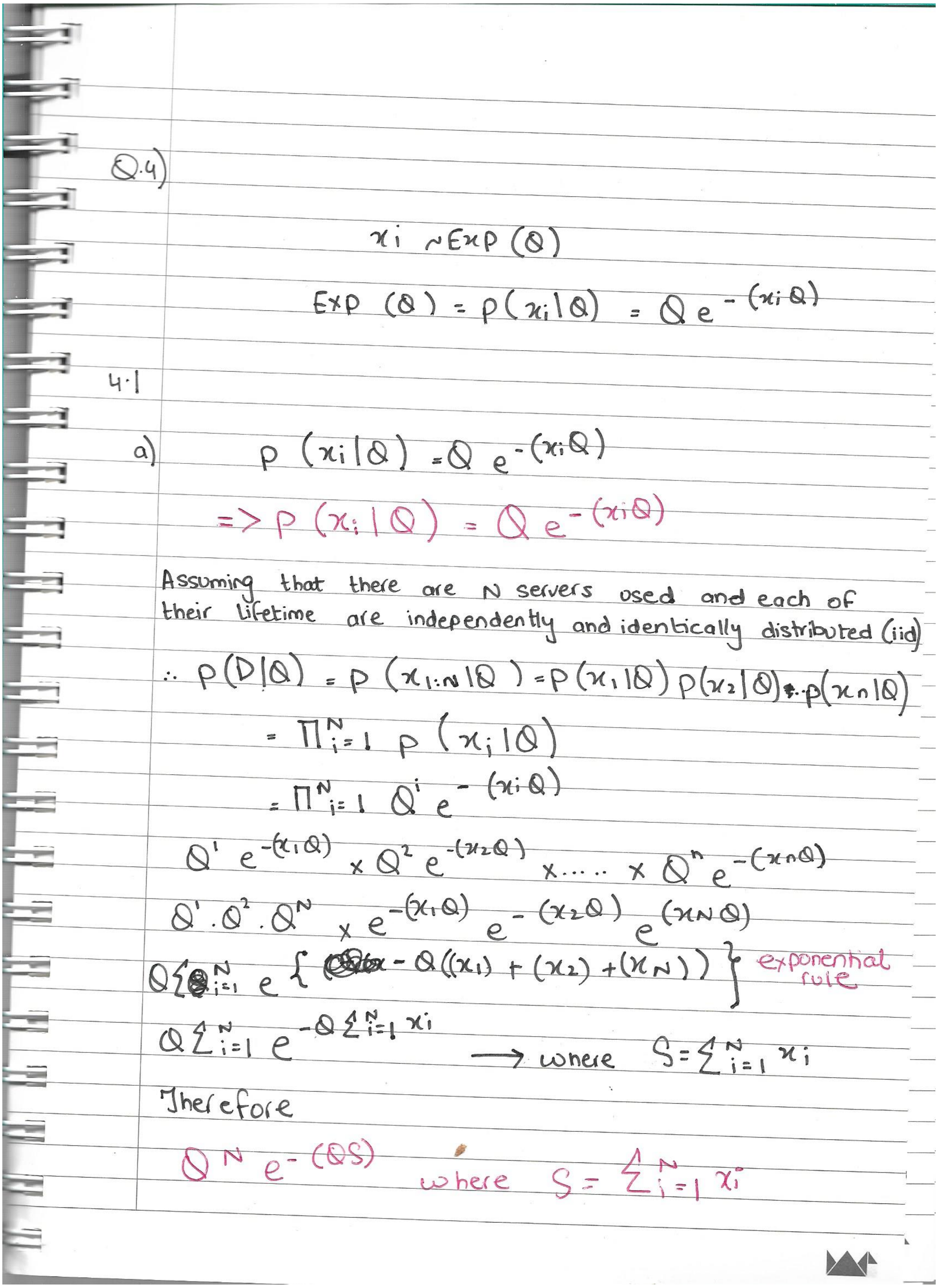
Conjugate priors reduce the complexity of calculations resulting in efficient and effective computations. This works well with Bayes as you can easily obtain the entire posterior from just point parameter estimates.

3.3)

a) A prior Dirichlet distribution followed by a multinomial likelihood could result in a conjugate distribution.

b) A prior beta distribution followed by binomial likelihood could result in a conjugate distribution

Q.4)



Q.4)

$$x_i \sim Exp\,(Q)$$

$$Exp\ (Q) = p(x_i|Q) = Q\,e^{-(x_i Q)}$$

4.1

a) $p\,(x_i|Q) = Q\,e^{-(x_i Q)}$

$\Rightarrow p\,(x_i\,|\,Q) = Q\,e^{-(x_i Q)}$

Assuming that there are N servers osed and each of their lifetime are independently and identically distributed (iid)

$\therefore\ p(D|Q) = p\,(x_{1:N}|Q) = p\,(x_1|Q)\,p(x_2|Q) \cdots p(x_n|Q)$

$= \prod_{i=1}^{N} p\,(x_i|Q)$

$= \prod_{i=1}^{N} Q^i\,e^{-(x_i Q)}$

$Q^1\,e^{-(x_1 Q)} \times Q^2\,e^{-(x_2 Q)} \times \dots\dots \times Q^n\,e^{-(x_n Q)}$

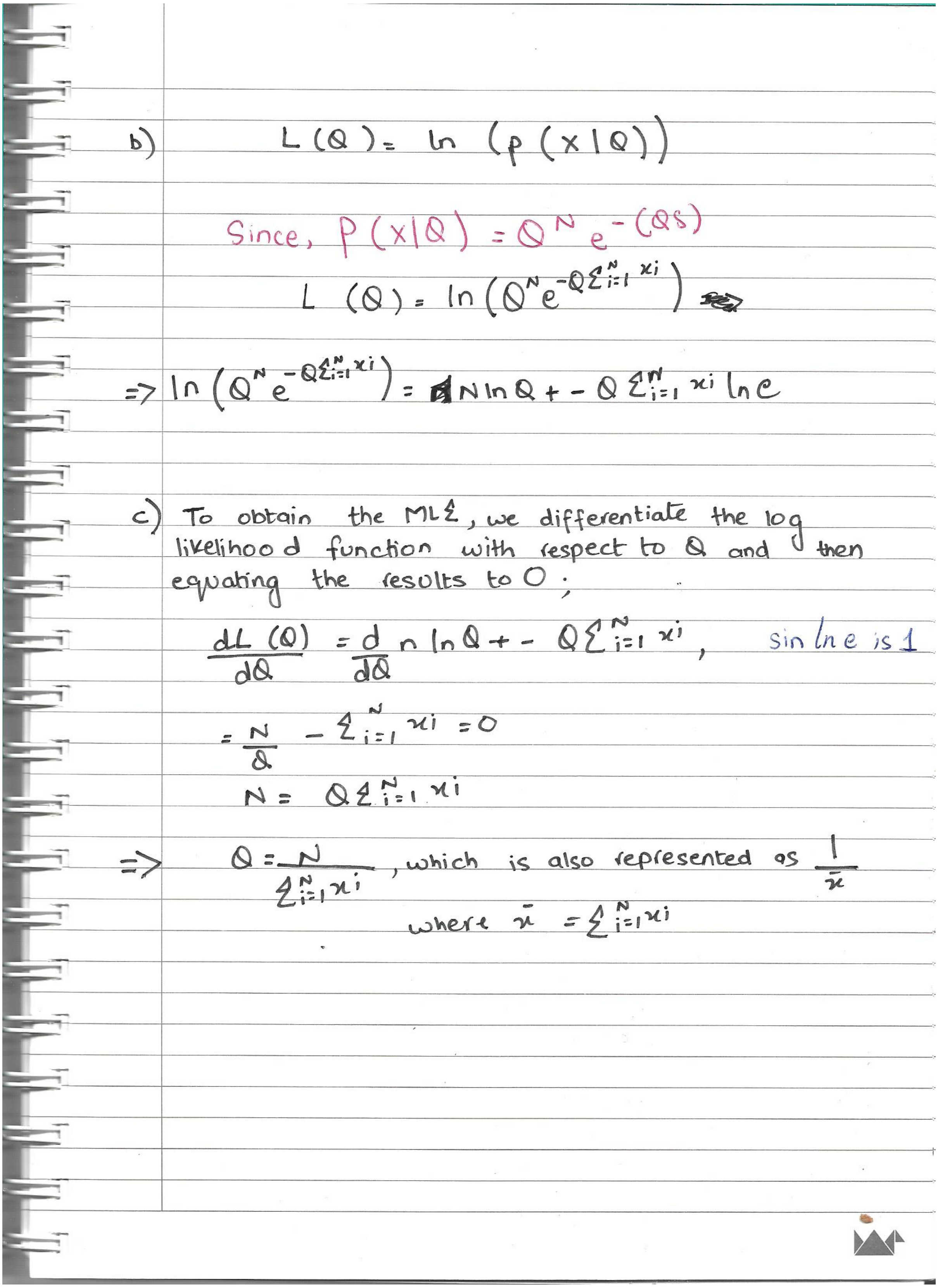
$Q^1 . Q^2 . Q^N \times e^{-(x_1 Q)}\,e^{-(x_2 Q)}\,e^{(x_N Q)}$

$Q \sum_{i=1}^{N} e^{\{ - Q((x_1) + (x_2) + (x_N))\}}$ exponential rule

$Q \sum_{i=1}^{N} e^{-Q \sum_{i=1}^{N} x_i} \longrightarrow$ where $S = \sum_{i=1}^{N} x_i$

Therefore

$Q^N\,e^{-(QS)}$ where $S = \sum_{i=1}^{N} x_i$



b) $L(Q) = \ln(p(x|Q))$

Since, $p(x|Q) = Q^N e^{-(Qs)}$

$L(Q) = \ln(Q^N e^{-Q\sum_{i=1}^N x_i})$

$\Rightarrow \ln(Q^N e^{-Q\sum_{i=1}^N x_i}) = N\ln Q + -Q\sum_{i=1}^N x_i \ln e$

c) To obtain the MLE, we differentiate the log likelihood function with respect to Q and then equating the results to 0;

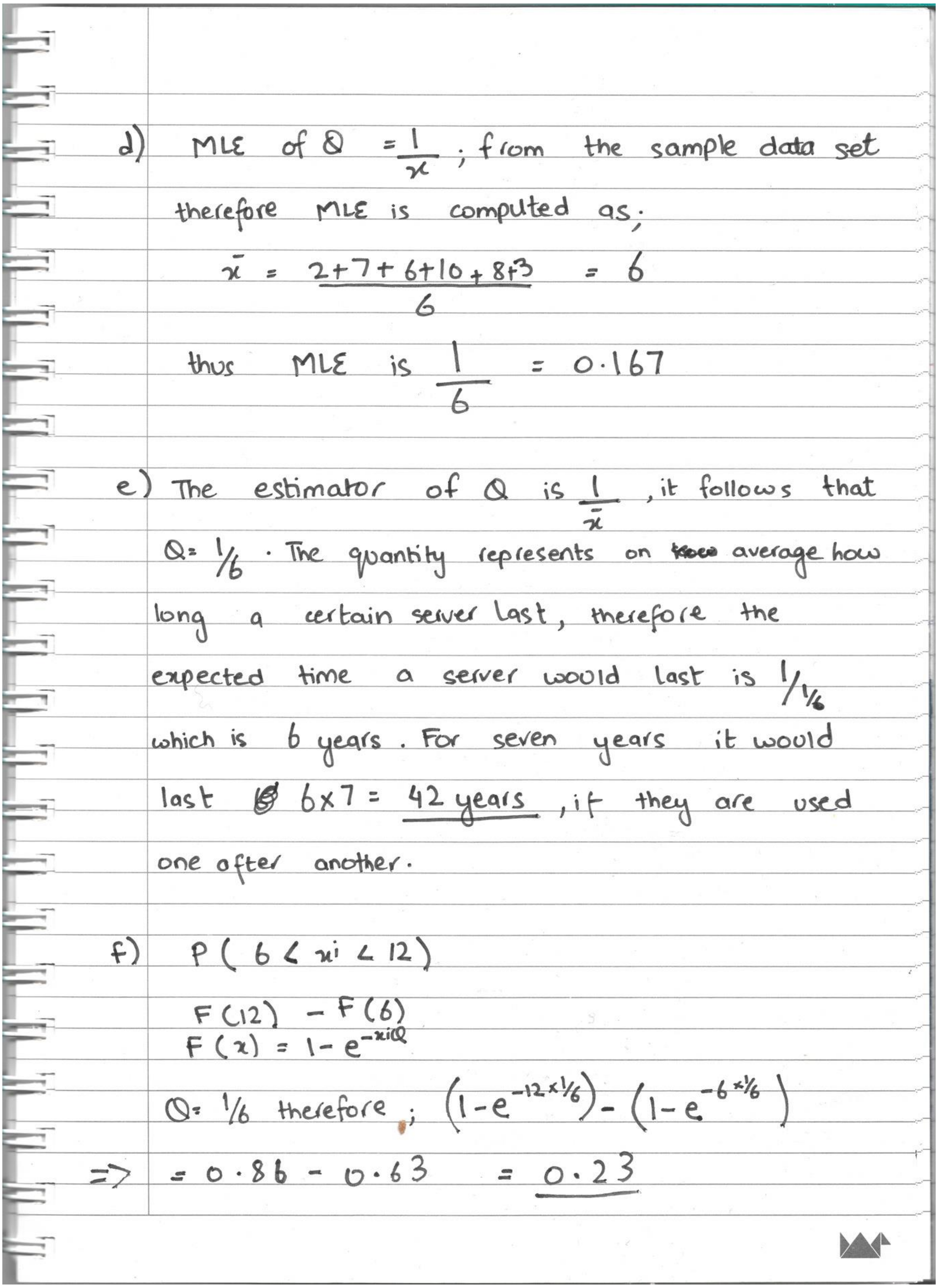
$\frac{dL(Q)}{dQ} = \frac{d}{dQ} n\ln Q + -Q\sum_{i=1}^N x_i$, sin $\ln e$ is 1

$= \frac{N}{Q} - \sum_{i=1}^N x_i = 0$

$N = Q\sum_{i=1}^N x_i$

$\Rightarrow Q = \frac{N}{\sum_{i=1}^N x_i}$, which is also represented as $\frac{1}{\bar{x}}$

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



d) MLE of $Q = \frac{1}{x}$; from the sample data set therefore MLE is computed as;

$$\bar{x} = \frac{2+7+6+10+8+3}{6} = 6$$

thus MLE is $\frac{1}{6} = 0.167$

e) The estimator of Q is $\frac{1}{\bar{x}}$, it follows that $Q = 1/6$. The quantity represents on average how long a certain server last, therefore the expected time a server would last is $1/_{1/6}$ which is 6 years. For seven years it would last $6 \times 7 = \underline{42 \text{ years}}$, if they are used one after another.

f) $P(6 < x_i < 12)$

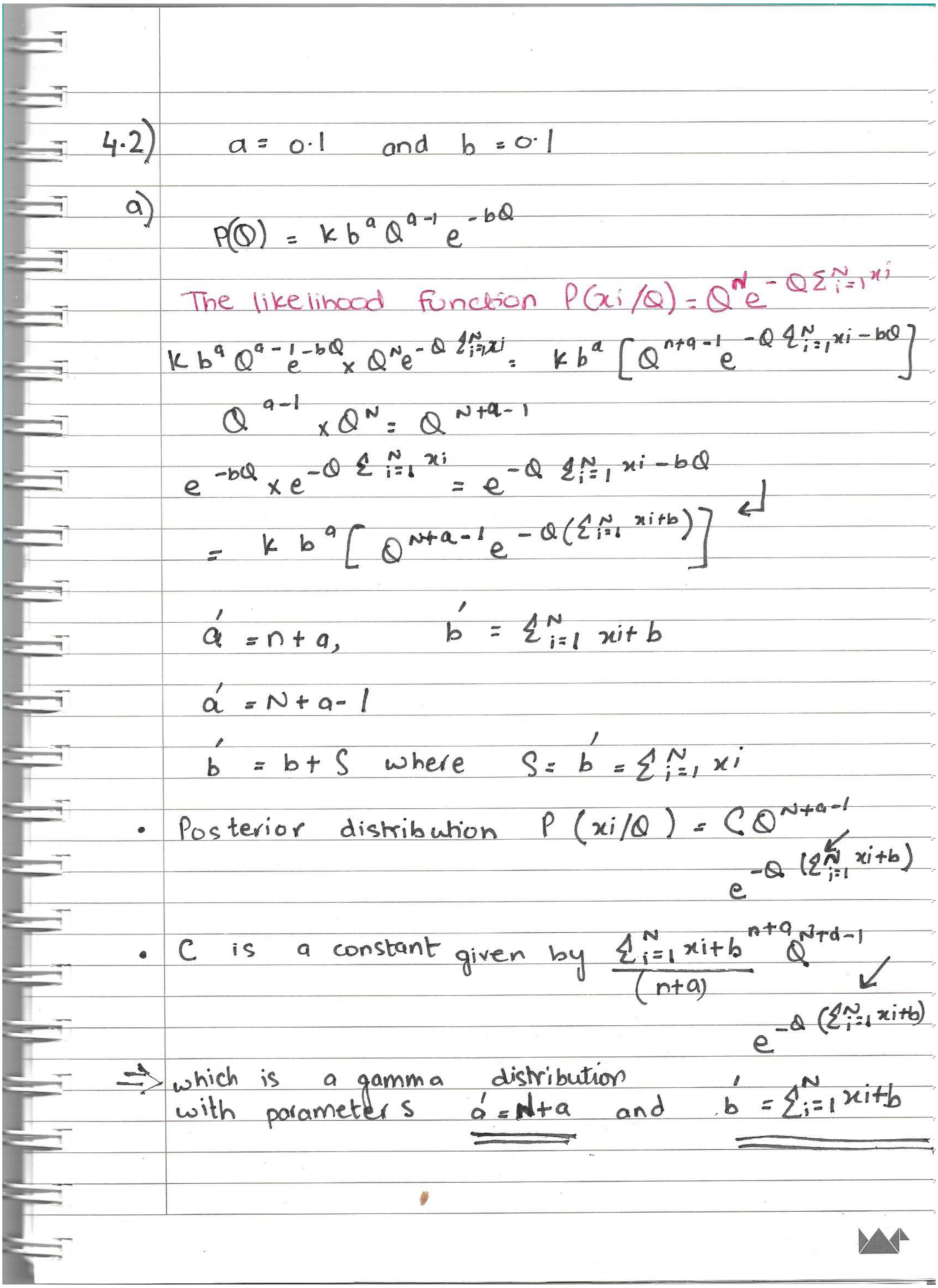
$F(12) - F(6)$

$F(x) = 1 - e^{-x_i Q}$

$Q = 1/6$ therefore; $(1 - e^{-12 \times 1/6}) - (1 - e^{-6 \times 1/6})$

$\Rightarrow = 0.86 - 0.63 = \underline{0.23}$

4.2)



4.2) $a = 0.1$ and $b = 0.1$

a) $P(Q) = k b^a Q^{a-1} e^{-bQ}$

The likelihood function $P(xi/Q) = Q^N e^{-Q \Sigma_{i=1}^N xi}$

$k b^a Q^{a-1} e^{-bQ} \times Q^N e^{-Q \Sigma_{i=1}^N xi} = k b^a \left[Q^{n+a-1} e^{-Q \Sigma_{i=1}^N xi - bQ} \right]$

$Q^{a-1} \times Q^N = Q^{N+a-1}$

$e^{-bQ} \times e^{-Q \Sigma_{i=1}^N xi} = e^{-Q \Sigma_{i=1}^N xi - bQ}$ ↲

$= k b^a \left[Q^{N+a-1} e^{-Q(\Sigma_{i=1}^N xi+b)} \right]$

$a' = n+a, \qquad b' = \Sigma_{i=1}^N xi + b$

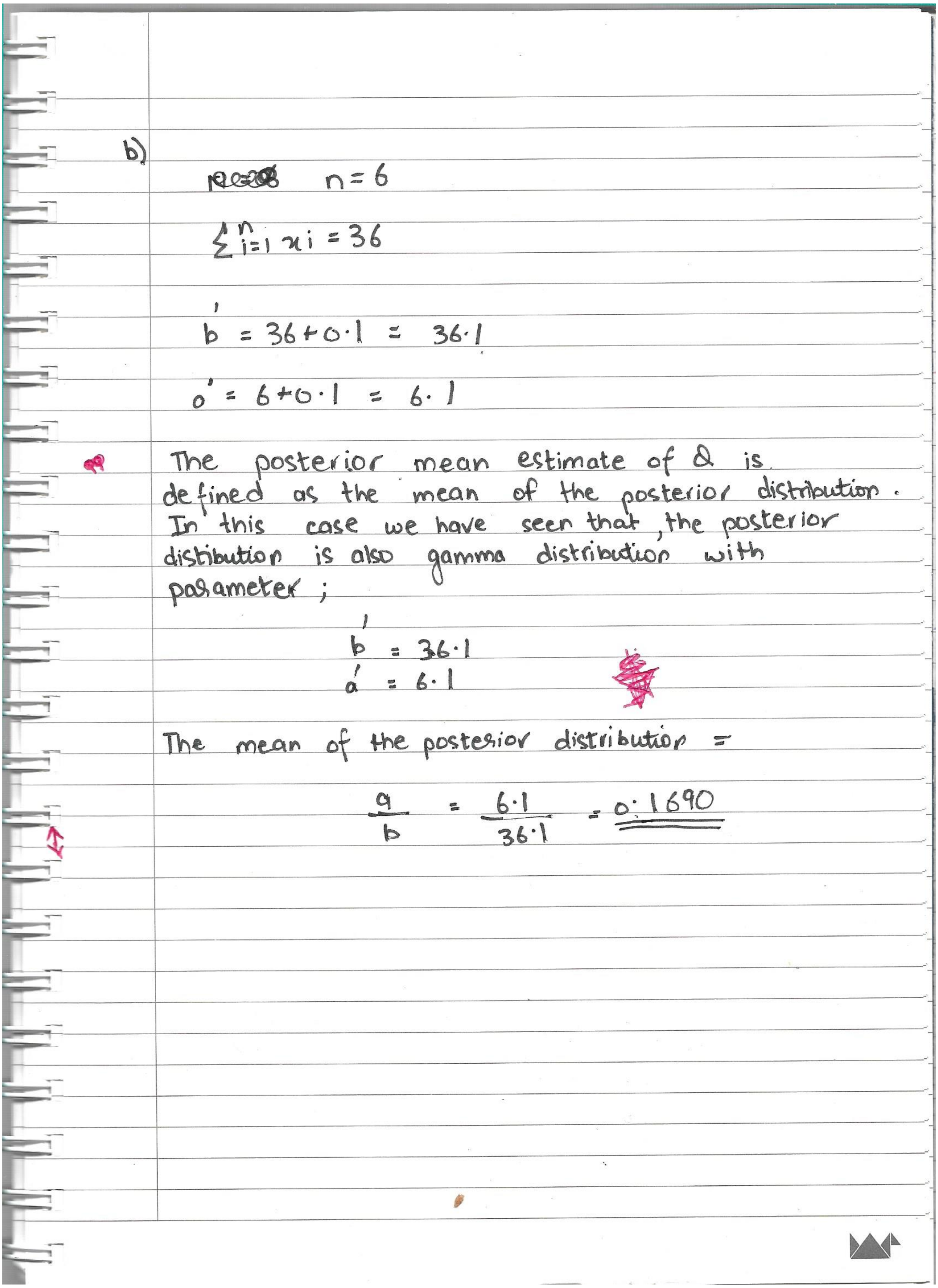
$a' = N+a-1$

$b' = b + S$ where $S = b' = \Sigma_{i=1}^N xi$

- Posterior distribution $P(xi/Q) = C Q^{N+a-1} e^{-Q(\Sigma_{i=1}^N xi+b)}$

- C is a constant given by $\dfrac{\Sigma_{i=1}^N xi+b^{n+a}}{(n+a)} Q^{N+d-1} e^{-Q(\Sigma_{i=1}^N xi+b)}$

⇒ which is a gamma distribution with parameters $\underline{\underline{a' = N+a}}$ and $\underline{\underline{b' = \Sigma_{i=1}^N xi+b}}$



b)

$n = 6$

$\sum_{i=1}^{n} x_i = 36$

$b' = 36 + 0.1 = 36.1$

$a' = 6 + 0.1 = 6.1$

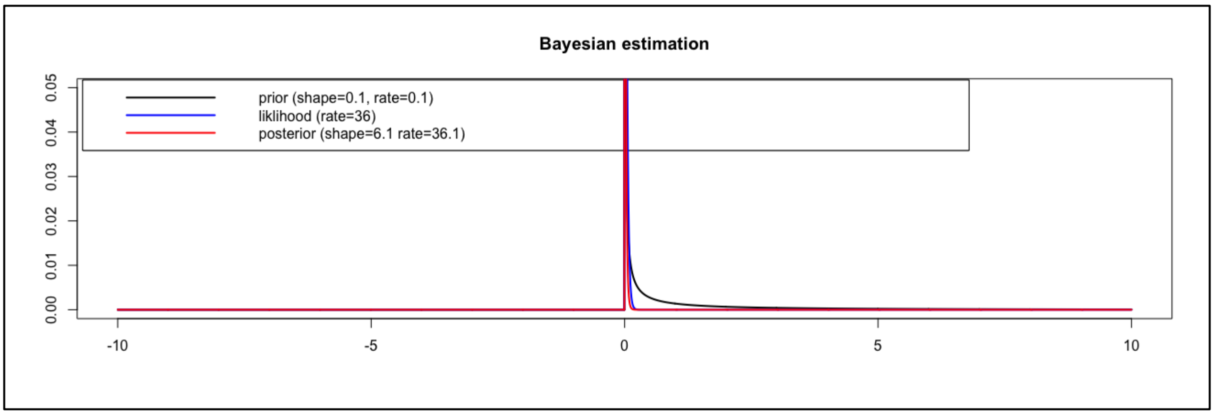
The posterior mean estimate of θ is defined as the mean of the posterior distribution. In this case we have seen that, the posterior distibution is also gamma distribution with parameter ;

$$b' = 36.1$$
$$a' = 6.1$$

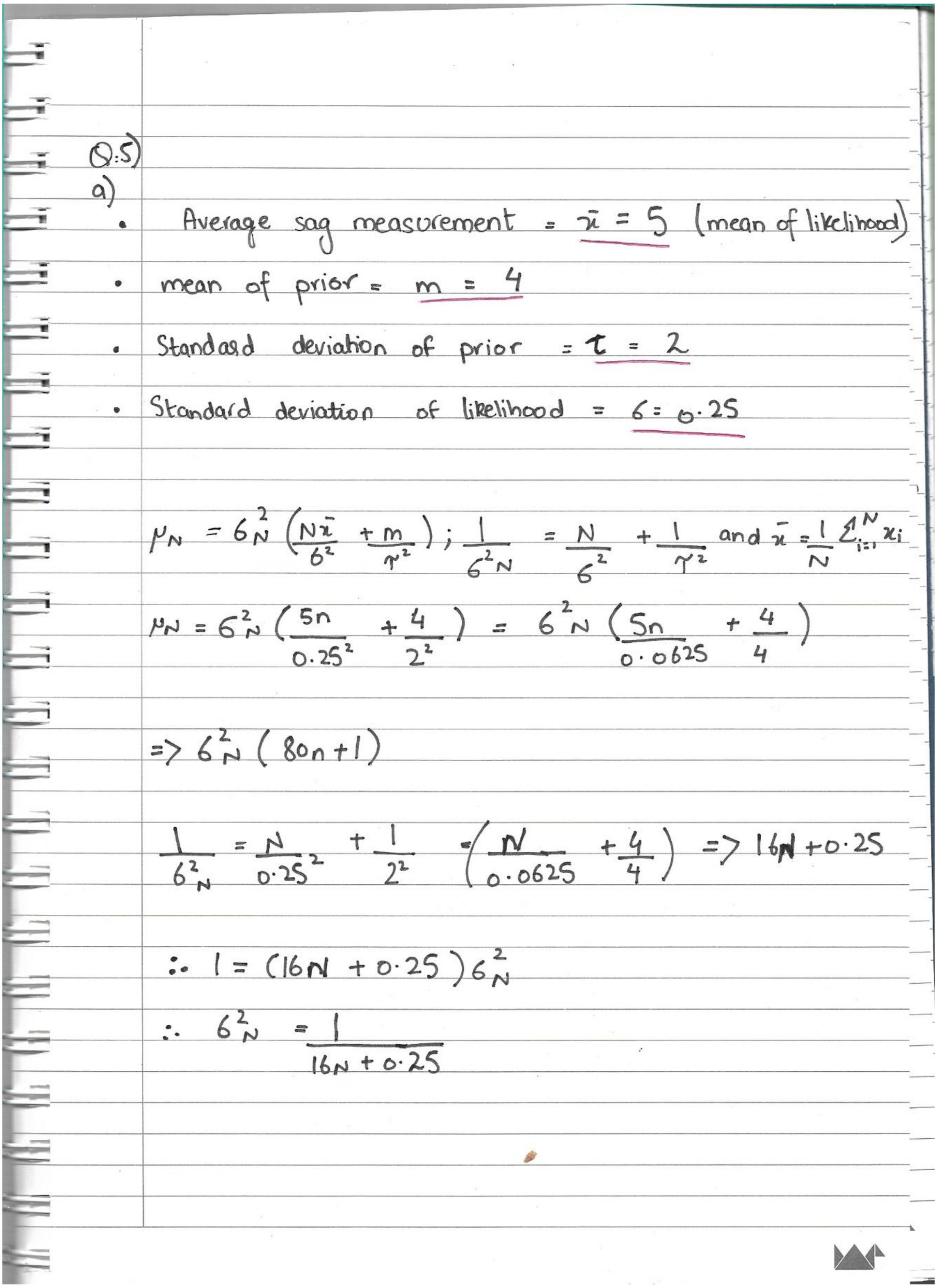
The mean of the posterior distribution =

$$\frac{a}{b} = \frac{6.1}{36.1} = \underline{\underline{0.1690}}$$

c)


Bayesian estimation
prior (shape=0.1, rate=0.1)
liklihood (rate=36)
posterior (shape=6.1 rate=36.1)
0.00
0.01
0.02
0.03
0.04
0.05
-10
-5
0
5
10

Q.5)



Q.5)

a)

- Average sag measurement = $\bar{x} = 5$ (mean of likelihood)
- mean of prior = $m = 4$
- Standard deviation of prior = $\tau = 2$
- Standard deviation of likelihood = $\sigma = 0.25$

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

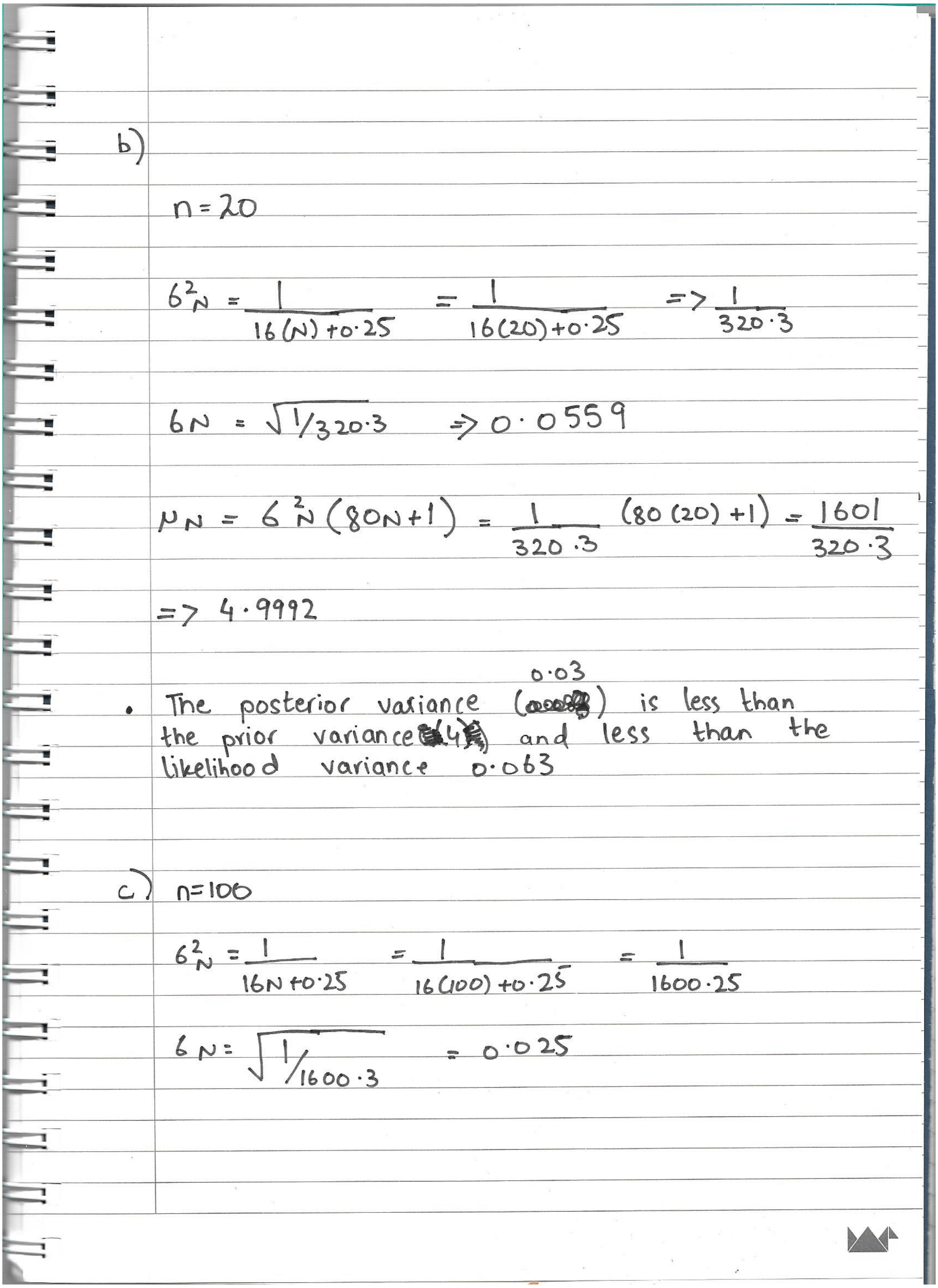
$\mu_N = \sigma_N^2\left(\frac{5n}{0.25^2} + \frac{4}{2^2}\right) = \sigma^2_N\left(\frac{5n}{0.0625} + \frac{4}{4}\right)$

$\Rightarrow \sigma_N^2(80n+1)$

$\frac{1}{\sigma_N^2} = \frac{N}{0.25^2} + \frac{1}{2^2} \left(\frac{N}{0.0625} + \frac{4}{4}\right) \Rightarrow 16N + 0.25$

$\therefore 1 = (16N + 0.25)\sigma_N^2$

$\therefore \sigma_N^2 = \frac{1}{16N + 0.25}$



b)

n=20

$$\sigma^2_N = \frac{1}{16(N)+0.25} = \frac{1}{16(20)+0.25} \Rightarrow \frac{1}{320.3}$$

$$\sigma_N = \sqrt{1/320.3} \Rightarrow 0.0559$$

$$\mu_N = \sigma^2_N(80N+1) = \frac{1}{320.3}(80(20)+1) = \frac{1601}{320.3}$$

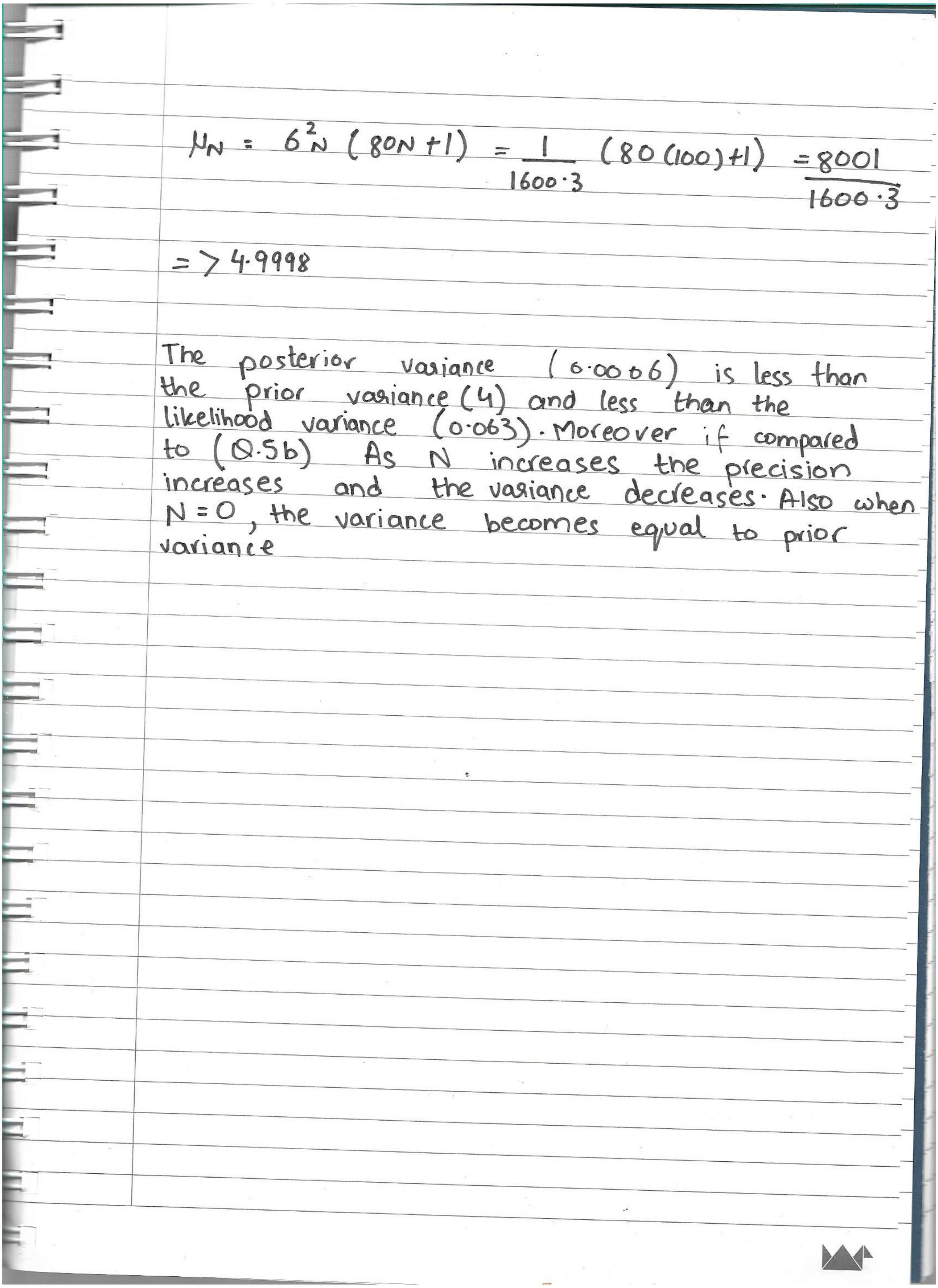
$$\Rightarrow 4.9992$$

- The posterior variance (0.03) is less than the prior variance (4) and less than the likelihood variance 0.063

c) n=100

$$\sigma^2_N = \frac{1}{16N+0.25} = \frac{1}{16(100)+0.25} = \frac{1}{1600.25}$$

$$\sigma_N = \sqrt{1/1600.3} = 0.025$$

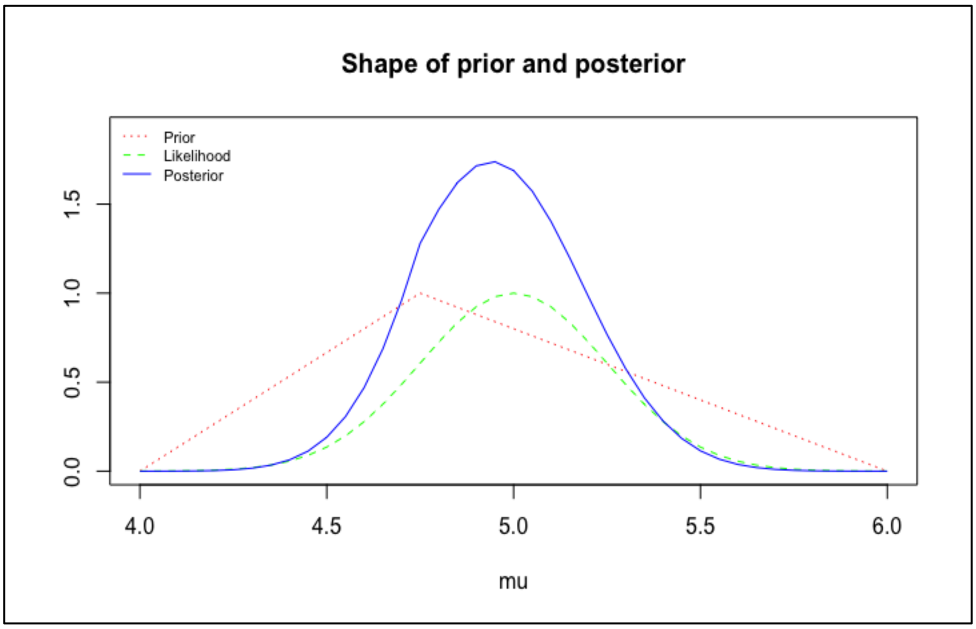


$$\mu_N = \sigma_N^2 (80N+1) = \frac{1}{1600\cdot 3}(80(100)+1) = \frac{8001}{1600\cdot 3}$$

$$=> 4.9998$$

The posterior variance (0.0006) is less than the prior variance (4) and less than the likelihood variance (0.063). Moreover if compared to (Q.5b) As N increases the precision increases and the variance decreases. Also when N = 0, the variance becomes equal to prior variance

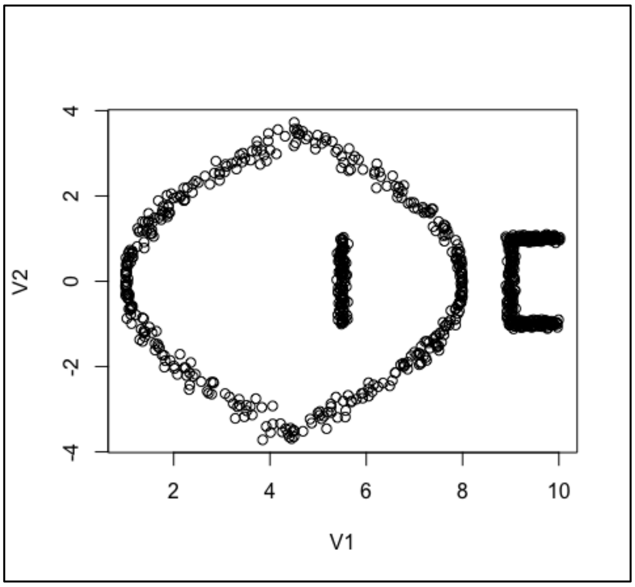
d)


Shape of prior and posterior
Prior
Likelihood
Posterior
0.0
0.5
1.0
1.5
4.0
4.5
5.0
5.5
6.0
mu

Q.6)

6.1)

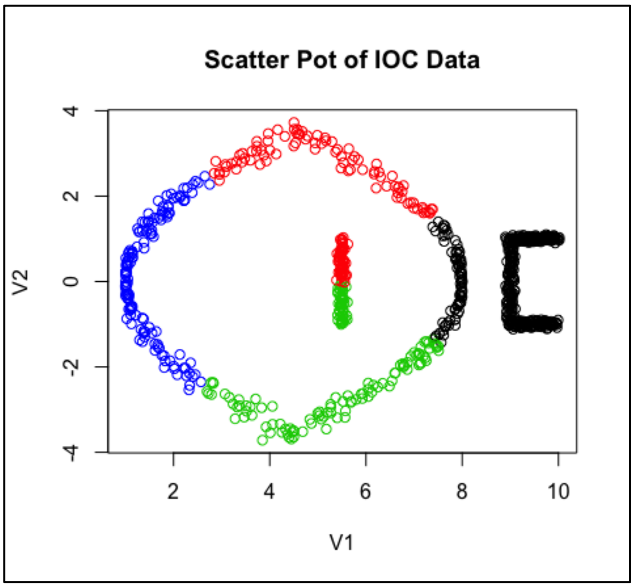
a)



b)

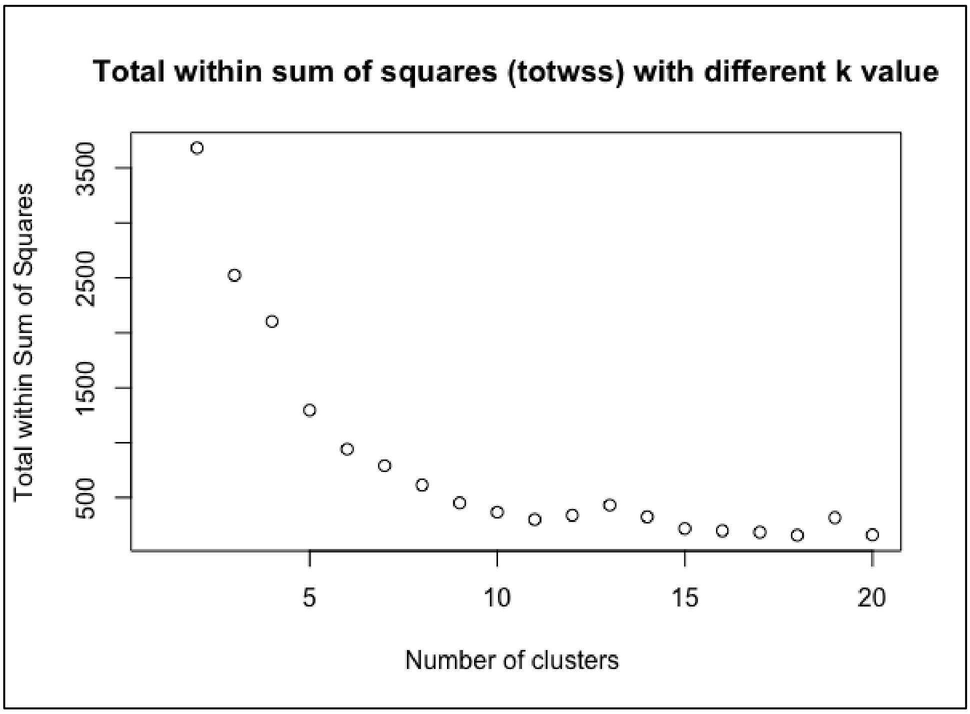
By visually looking at the scatter plot we could see there could be 4 clusters/classes that can be found in the data.

c)



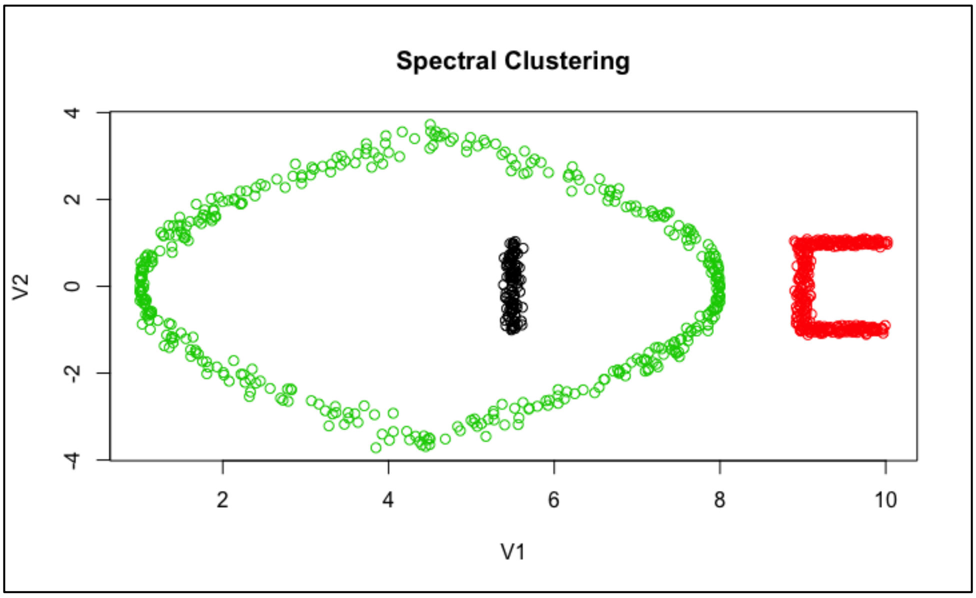
We can see the 4 different clusters plotted in the chart above meaning that there could be 4 different groups pf similar nature present in the data

d)



We can use this graph to incorporate the elbow method and find the correct number of clusters for this model and data and them use them to more efficient results. This method shows the diminishing reduction in total sum of squares. The number could be debateable but we can see by the looks of it that 7 or 8 clusters would be the best value for the number of clusters needed.

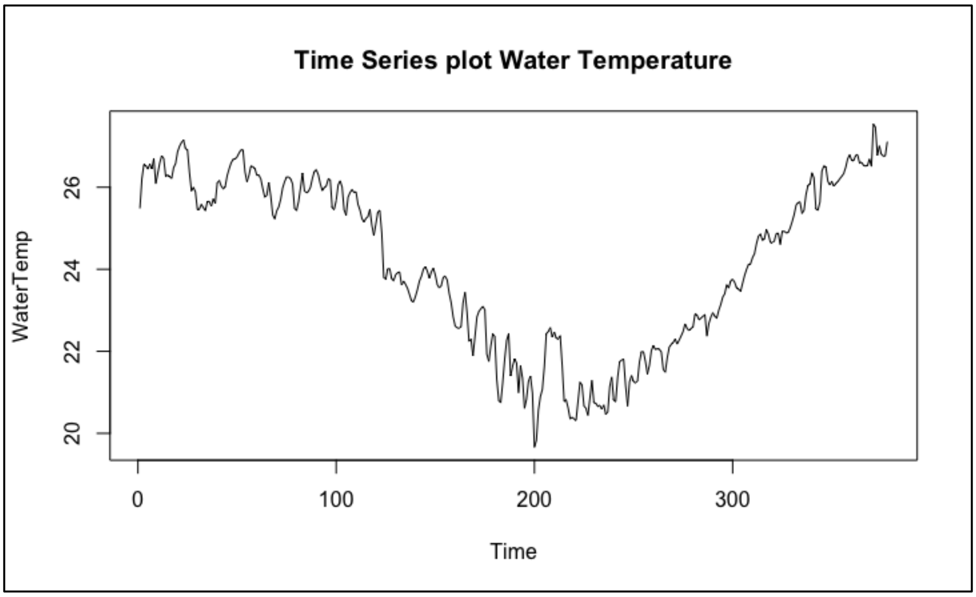
6.2)



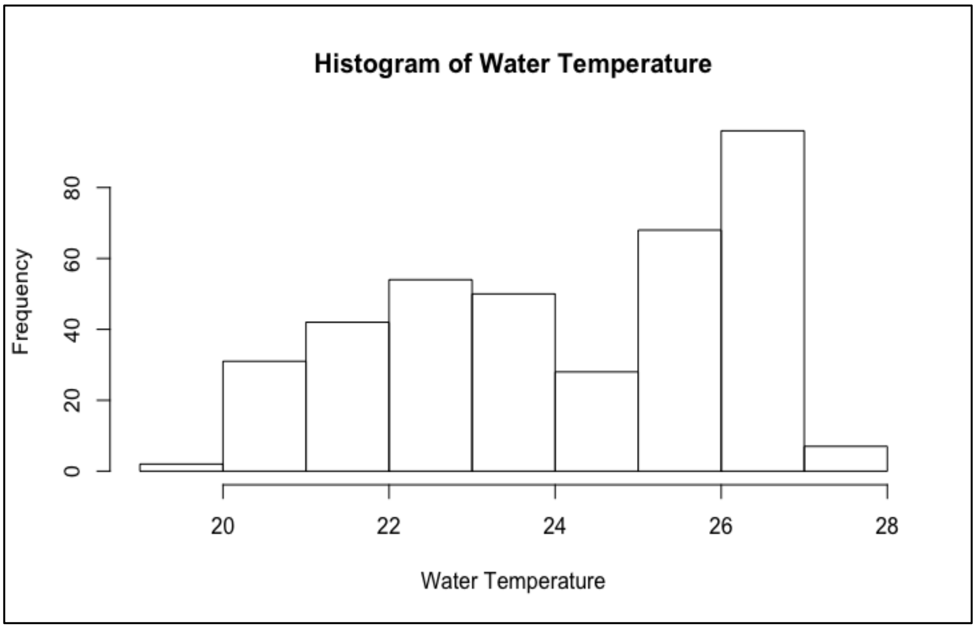
As we can see the clusters obtained through spectral clustering, we were able to identify the the groups of data on the graph with similar nature. Since the simple clustering we did before identified 4 different clusters of data that could have been possible groups, this spectral clustering embedding the vertices of a graph into a low dimensional space and identifying the nodes of the graph based on edging, gave a more clear picture of the similar coordinates.

Q.7)

7.1



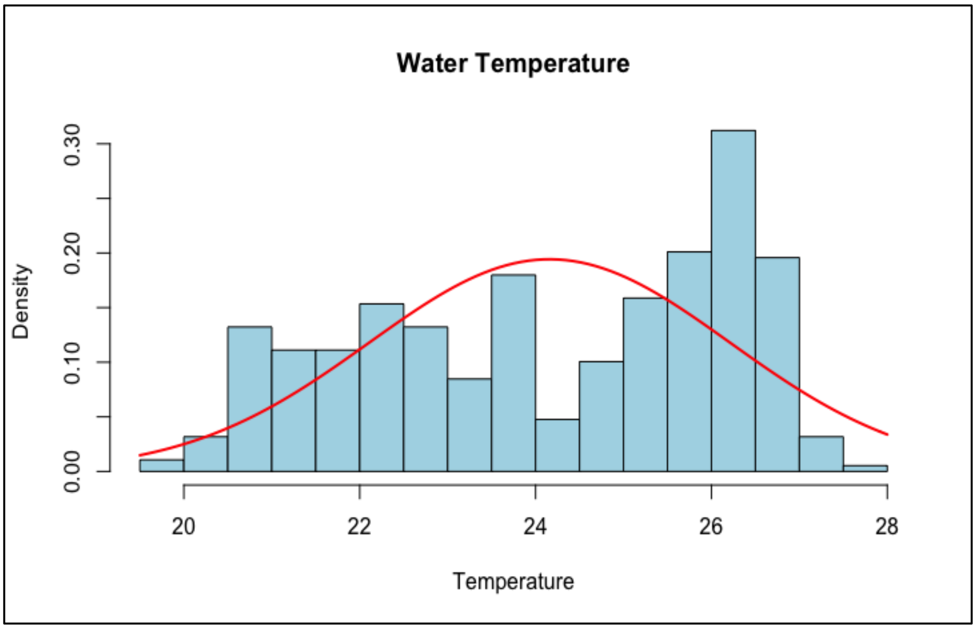
7.2



After looking at the data we could say that there are two modes in the histogram above, one mode could be observed at 22-23 and another one at 26 – 27.

7.3

```
       mean            sd
 24.15940726   2.05351532
( 0.10562143) ( 0.07468563)
```



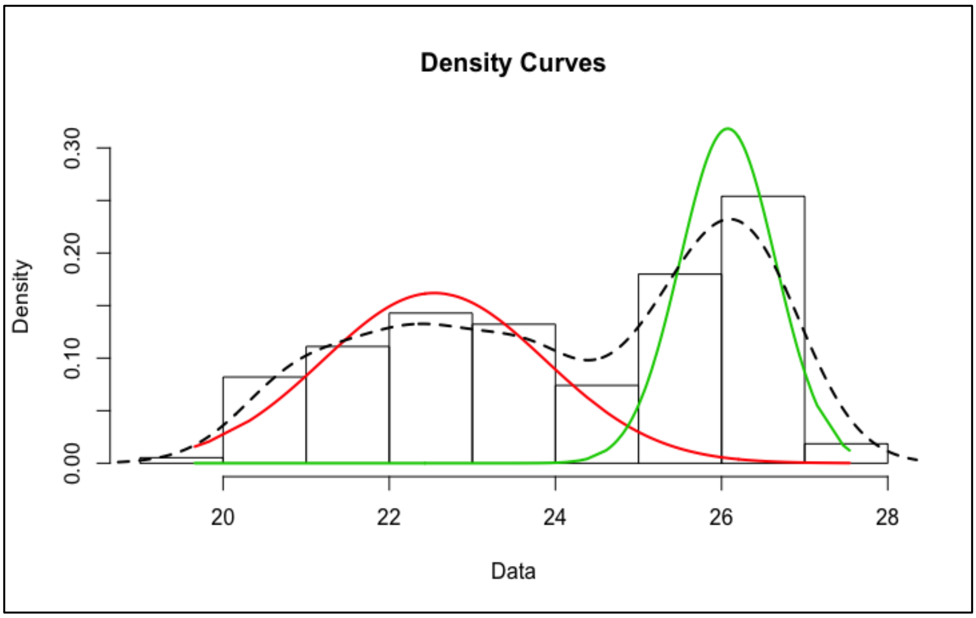
7.4

Mixing coefficients : 0.5420555 0.4579445

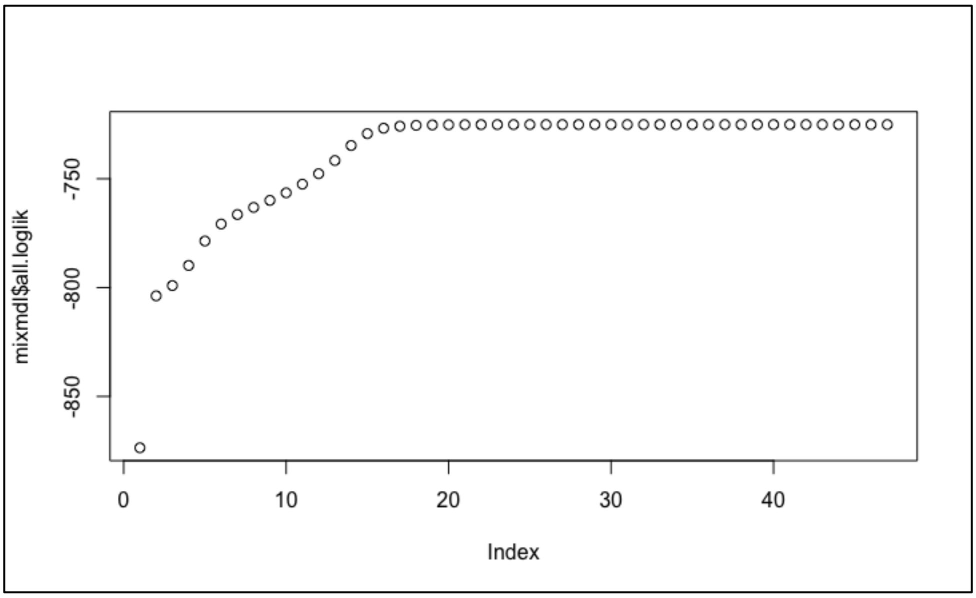
Mu : 22.54130 26.07471

Sigma : 1.3355762 0.5739647

7.5



7.6



After seeing the plot we can see the log likelihood values start to increase at a stable rate after the index starting to increase.

7.7

We made two models, **single gaussian model** and **mixture of gaussians model**. The single gaussian model had a wide spread as we can confirm it through the standard deviation, whereas the mixture of gaussians model gave a result of multiple distributions where the means of the distribution collided with others, but looking and comparing it to the initial model’s standard deviation we can say this model has smaller spreads. Therefore, I can conclude that the mixture of gaussians model (7.4) was more accurate in reflecting the data.

7.8

There is a presence of singularities , for some values if the variance goes to 0 then the term goes to infinity and so log likelihood. Hence, we won’t be able to maximize the log likelihood as its not a well posed problem, as severe over fitting occurs. To overcome this, we can use heuristics to detect this and reset the mean to a randomly chosen value while resetting its covariance to some large value, and then continuing with the optimisation.