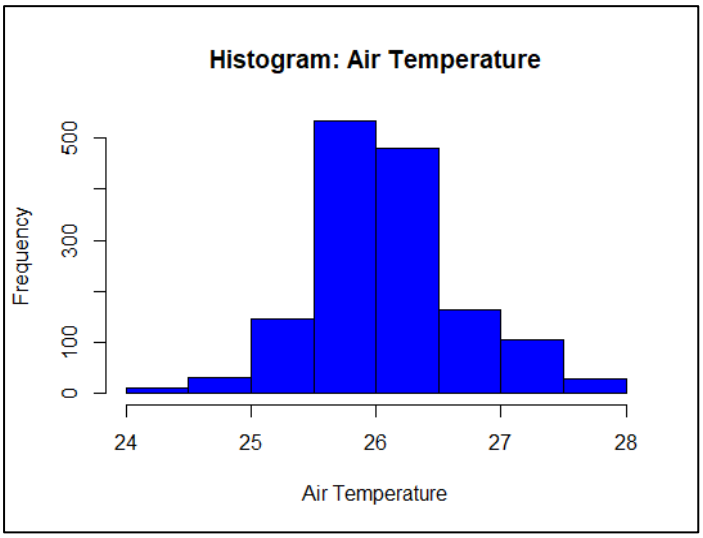
Table of Contents

Question 1

1.1 Draw histograms for 'Air temperature' and 'Air Pressure" values, and comment on them.

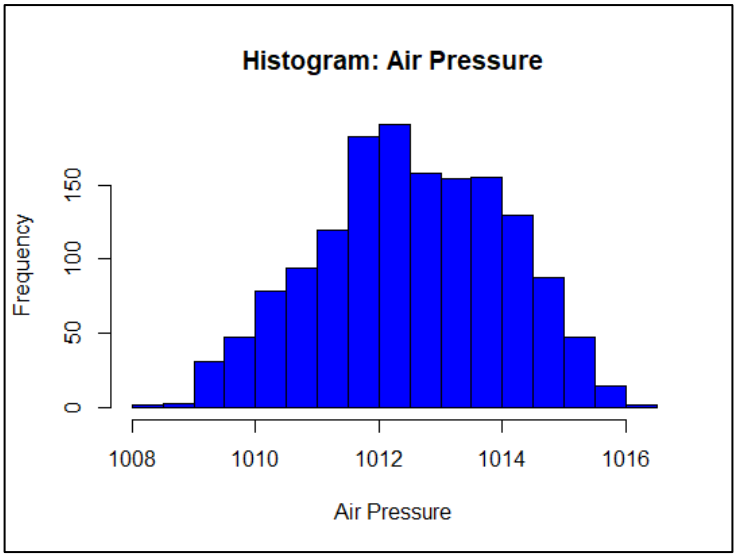
Air Temperature:

a) The peak of Air Temperature occurs between 25.5 to 26.5
b) Temperature spreads from 24 to 28
c) Data seems to be symmetrical as per histogram and also mean and median is similar (26.13 and 26.10 respectively)
d) Data is unimodal
e) Data seems to have outliers below 25 and above 27



Air Pressure:

a) The peak of Air Pressure occurs between 1011 to 1013
b) Air Pressure spreads from 1008 to 1016
c) Data seems to be symmetrical as per histogram and also mean and median is similar (1013 and 1013 respectively)
d) Data is unimodal
e) Data seems to have outliers near 1008



1.2 Draw a parallel Box plot using the two variables; 'Air Temperature' and the 'Wind Speed'. Find five number summaries of these two variables. Comment on boxplot and five number summaries.

Air Temperature:

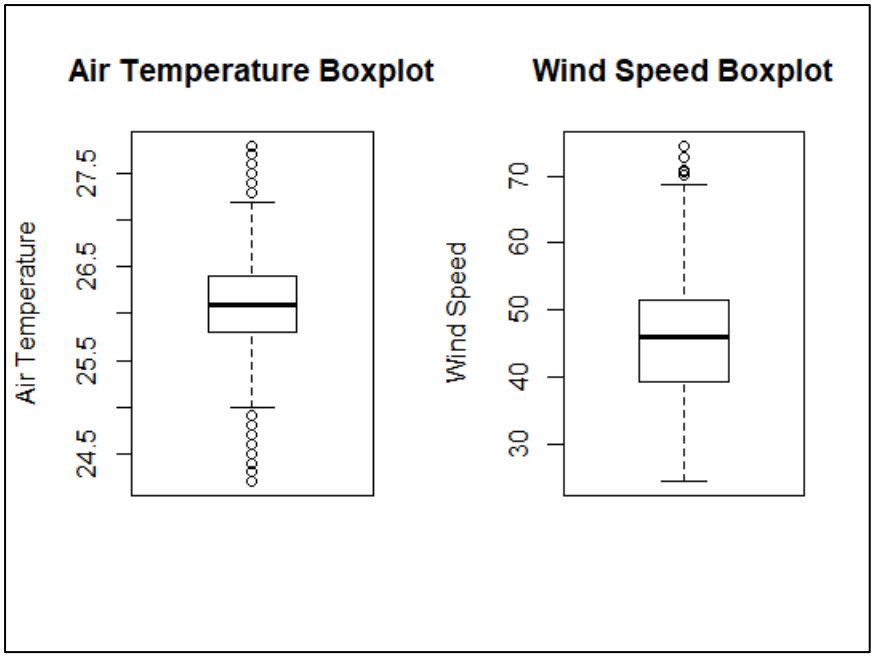
a) Data seems to be symmetrical as per boxplot and also mean and median is similar (26.13 and 26.10 respectively)
b) Data has outliers below 25 and above 27

Wind Speed:

a) Data seems to be symmetrical as per boxplot and also mean and median is similar (45.72 and 46.08 respectively)
b) Data has few outliers above 70.

Five number summaries:

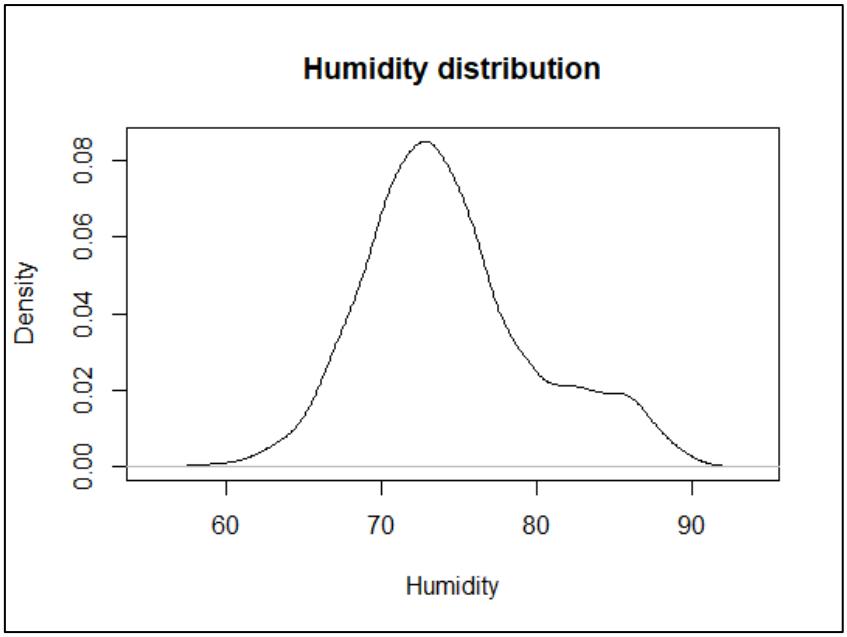
Variable	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Air Temperature	24.20	25.80	26.10	26.40	27.80
Wind Speed	24.48	39.24	46.08	51.48	74.52



1.3 Which summary statistics would you choose to summarize the center and spread for the 'Humidity' data? Why (support your answer with proper plot/s)? Find those summary statistics for the "Humidity" data.

As per statistical tests such as Shapiro test and JB test, Humidity data does not follow normal distribution, but as per density plot and similarity between mean and median. Thus, I would use mean and median for center and standard deviation for the spread to summarize the data.

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Humidity	58.20	70.60	73.55	74.36	77.20	91.00

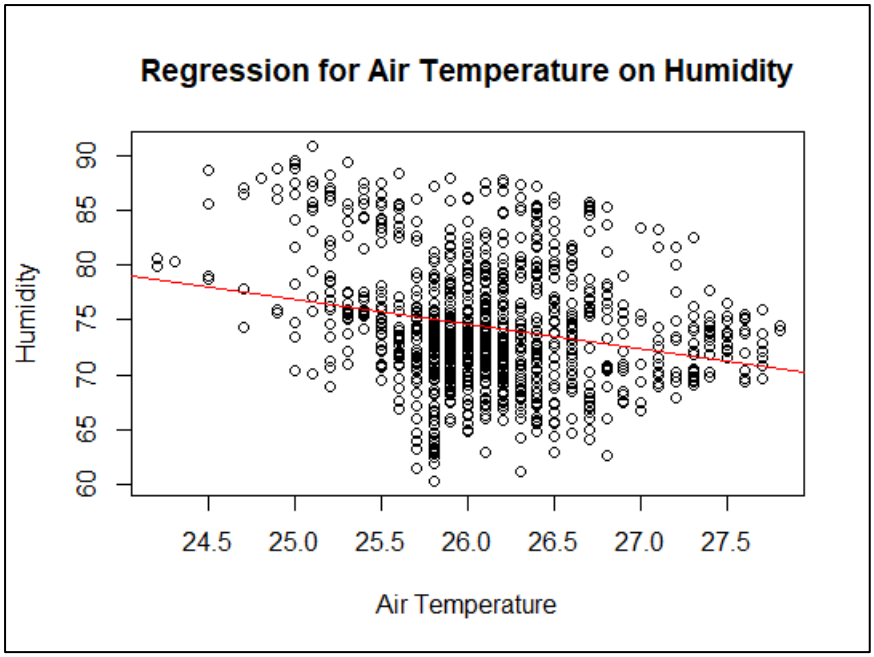


1.4 Draw a scatterplot of ''Air Temperature' (as x) and 'Humidity' (as y) for the first 1000 data vectors selected from the "my.data" (name the axes). Fit a linear regression model to the above two variables and plot the (regression) line on the same scatter plot. Write down the linear regression equation. Compute the correlation coefficient and the coefficient of Determination. Explain what these results reveal.

Linear Regression Equation:
$\hat{y} = 132.9877 - 2.2441\ X$

Correlation Coefficient between Air Temperature and humidity is -0.2345
Coefficient of determination (R^2) is approximately 0.055 (it is square of correlation coefficient)

Interpretation of above results:
Linear equation: In absence of Air Temperature, humidity will be approximately 133. For each unit increase in Air Temperature, humidity will decrease by 2.2441. Model is signification as p-value is less than 0.05 (alpha).
Correlation coefficient: There is weak negative correlation of 23.45% between these two variables. i.e. weak linear relationship between these two variables.
Coefficient of determination: Only 5.5% variation in humidity is explained by the variation in Air Temperature. This model is weak and need to consider other variables which affects humidity or to fit non-linear models.



Question 2

2.1 Computed manually

Q.2.

2.1

Sports	State	NSW (N)	Victoria (V)	Queensland (Q)	Total
	Footy (F)	1000	2000	1300	4300
	Basketball (B)	1500	500	500	2500
	Cricket (C)	1400	1000	800	3200
	Total	3900	3500	2600	10,000

Suppose we select a person at random;

a) $P(V) = \frac{3500}{10,000} = 35\%$

b) $P(C \cap N) = \frac{1400}{10,000} = 14\%$

c) $P(F|Q) = \frac{P(F \cap Q)}{P(Q)} = \frac{1300}{2600} = 50\%$

d) $P(V|B) = \frac{P(V \cap B)}{P(B)} = \frac{500}{2500} = 20\%$

e) $P(V \cup C) = P(V) + P(C) - P(V \cap C)$

$= \frac{3500 + 3200 - 1000}{10,000}$

$= 57\%$

f) Marginal probability of sports

Sports	Footy (F)	Basketball (B)	Cricket (C)
Marginal probability	43%	25%	32%

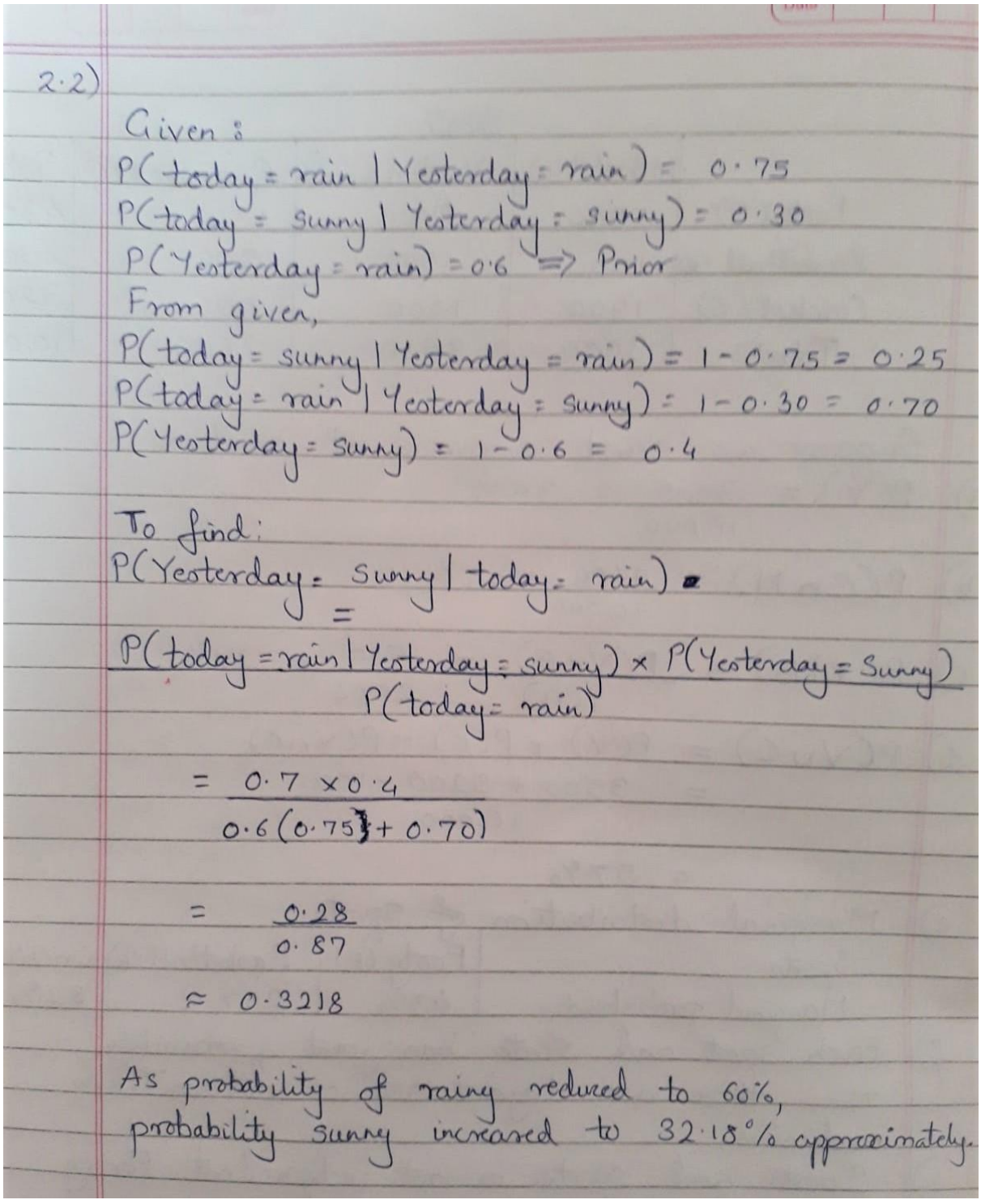
g) Each sport and state has joint probability, thus these are not disjoint or mutually exclusive events.

h) $P(F|Q) = 50\% \neq P(F) = 43\%$

Thus, sports and state are not independent.

2.2 Computed manually

Probability that it was sunny yesterday given that it is rainy today is ~32.18%



2.2)

Given :

$P(\text{today} = \text{rain} \mid \text{Yesterday} = \text{rain}) = 0.75$

$P(\text{today} = \text{sunny} \mid \text{Yesterday} = \text{sunny}) = 0.30$

$P(\text{Yesterday} = \text{rain}) = 0.6 \Rightarrow$ Prior

From given,

$P(\text{today} = \text{sunny} \mid \text{Yesterday} = \text{rain}) = 1 - 0.75 = 0.25$

$P(\text{today} = \text{rain} \mid \text{Yesterday} = \text{sunny}) = 1 - 0.30 = 0.70$

$P(\text{Yesterday} = \text{sunny}) = 1 - 0.6 = 0.4$

To find:

$P(\text{Yesterday} = \text{sunny} \mid \text{today} = \text{rain})$

$$= \frac{P(\text{today} = \text{rain} \mid \text{Yesterday} = \text{sunny}) \times P(\text{Yesterday} = \text{Sunny})}{P(\text{today} = \text{rain})}$$

$$= \frac{0.7 \times 0.4}{0.6(0.75) + 0.70)}$$

$$= \frac{0.28}{0.87}$$

$$\approx 0.3218$$

As probability of rainy reduced to 60%, probability sunny increased to 32.18% approximately.

Question 3

3.1 Two differences between frequentist way and the Bayesian way of estimating a parameter

Frequentist	Bayesian

θ is considered to be fixed parameter whose value is determined by some form of estimator which provides the point estimate.	θ is considered to be random variable and the uncertainty in the parameter is expressed through a probability distribution over prior θ.
Error bars on the estimate are obtained from sub-samples (distribution) of dataset D (Bootstrap method).	Single observed dataset D is used and the output is complete probability distribution (posterior) $p(\theta \mid D) = p(D \mid \theta) \times p(\theta)/p(D)$
Widely used estimator is Maximum likelihood Estimator. But there are chances of overfitting	Prior distribution over θ in the Bayesian setting avoids overfitting.

3.2 Why conjugate priors are useful in Bayesian statistics?

In Bayesian statistics, using a conjugate prior gives closed form expressions for posterior,

- No need to use numerical integration when computing the posterior
- Computationally efficient

Examples;

- Gaussian –Gaussian model (Gaussian prior, Gaussian likelihood => Gaussian posterior)
- Dirichlet-Multinomial model (Dirichlet prior, Multinomial likelihood => Dirichlet posterior)

3.3 Give two examples of Conjugate pairs (i.e., give two pairs of distributions that can be used for prior and likelihood)

- Exponential – Gamma model (Gamma prior, Exponential likelihood => Gamma posterior)
- Bernoulli - Beta model (Beta prior, Bernoulli likelihood => Beta posterior)

Question 4

4.1 Computed manually

4.1 $x_i \sim Exp(\theta)$

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

a) Joint distribution of lifetime of N servers

Sol^n $p(X|\theta) = p(x_{1:N}|\theta) = p(x_1|\theta) \cdot p(x_2|\theta) \cdot \dots \cdot p(x_n|\theta)$

Since $X \sim iid$

$p(x|\theta) = \theta e^{-\theta x_1} \times \theta e^{-\theta x_2} \times \dots \times \theta e^{-\theta x_N}$

$= \theta^N \times [e^{-\theta \sum_{i=1}^{N} x_i}]$

$= \theta^N \times e^{-\theta S}$

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

b) Simplified expression for the Log-likelihood $L(\theta) = \ln(p(X|\theta))$

Sol^n $L(\theta) = \ln(\theta^N \times [e^{-\theta S}])$

$\therefore L(\theta) = \ln \theta^N + \ln(e^{-\theta S})$

$\therefore L(\theta) = N \ln \theta - \theta S$; where $S = \sum_{i=1}^{N} x_i$

c) Show that MLE($\hat{\theta}$) of the parameter θ is given by:

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

Sol^n $\frac{d\,L(\theta)}{d\theta} = \frac{d}{d\theta}[n \ln\theta - \theta S] = 0$

$\frac{d}{d\theta} \frac{n}{\theta} \frac{d}{d\theta} n \ln\theta - \frac{d}{d\theta} \theta S = 0$

$\therefore \frac{n}{\theta} - S = 0$

$\therefore \frac{n}{\theta} = S$

$\therefore \theta = \frac{n}{S}$

$\therefore \theta = \frac{n}{\sum_{i=1}^{N} x_i}$

$\therefore \theta = \frac{1}{\bar{x}} \Rightarrow$ where $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$

d) Lifetime of 6 servers $\{2, 7, 6, 10, 8, 3\}$.
MLE $\hat{\theta}$ of parameter?

Soln.

$$\hat{\theta} = \frac{n}{S} = \frac{6}{2+7+6+10+8+3} = \frac{6}{36} = \frac{1}{6} \approx 0.167$$

e) On an average 7 servers should last for 42 years if they are used one after another.

$$\mu = \text{mean life of servers} = \frac{1}{\hat{\theta}} = \frac{1}{1/6} = 6 \text{ years}$$

$$7 \text{ servers} \times 6 \text{ years} = 42 \text{ years}$$

f) Probability that server lasts between six and twelve years;

$$P(6 \leq x \leq 12) = P(x \leq 12) - P(x \leq 6)$$
$$\therefore P(6 \leq x \leq 12) = 1 - e^{-\hat{\theta} x_1} - (1 - e^{-\hat{\theta} x_2})$$
$$\therefore P(6 \leq x \leq 12) = 1 - e^{-0.167 \times 12} - (1 - e^{-0.167 \times 6})$$
$$\therefore P(6 \leq x \leq 12) \approx 0.2323$$

4.2 Computed manually and b

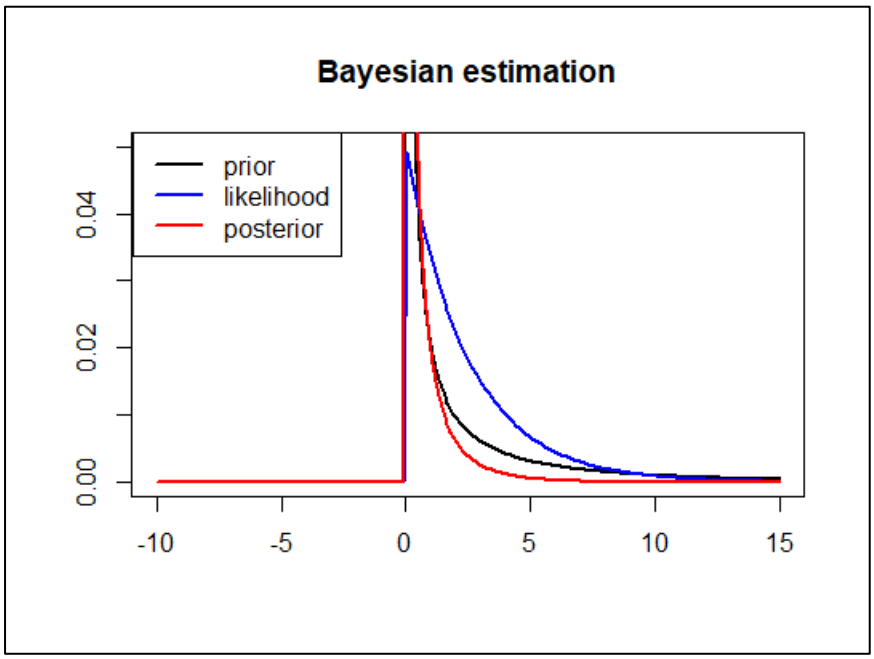
4.2 Gamma (a, b) prior
Exponential (θ) likelihood
Gamma - Exponential ⇒ Gamma (a', b') posterior

a = 0.1 ; b = 0.1
Gamma (a, b) = k $b^a \theta^{(a-1)} e^{-b\theta}$; where k is a constant
$k = \frac{1}{\Gamma(a)}$

a) Show that posterior distribution is also Gamma(a', b')
posterior ∝ likelihood × prior
$p(\theta|D) = p(D|\theta) \times p(\theta)$
$= \theta^N e^{-\theta S} \times b^a \theta^{(a-1)} \times e^{-b\theta}$
$= \theta^N \times \theta^{(a-1)} \times b^a \times e^{-\theta S} \times e^{-b\theta}$
$= \theta^{N+a-1} \times b^a \times e^{-\theta S + (-b\theta)}$
$= \theta^{N+a-1} \times b^a \times e^{-\theta(S+b)}$
$= \theta^{a'-1} \times b^a \times e^{-\theta b'}$
where $a' = N + a$ and $b' = (S + b)$
$\therefore p(\theta|D) = \frac{b^a}{\Gamma(a)} \times \theta^{a'-1} \times \theta^{-\theta b'}$
where $\frac{b^a}{\Gamma(a)}$ = constant
Here posterior follows Gamma (N+a, b+nx)
where $n\bar{x} = S$; i.e. Gamma (a', b')
Thus, prior and posterior are conjugate distribution

b) 6 Servers lifetime ⇒ {2, 7, 6, 10, 8, 3}. Find a', b' and posterior mean estimate of θ.
Sol^n:- N = 6, a = 0.1, b = 0.1, S = 36
a' = N + a = 6.1 , b' = S + b = 36.1
$\therefore \hat{\theta} = E(\theta|D) = \text{mean}[\text{Gamma}(\theta|a', b') = \frac{a'}{b'}$
$= \frac{6.1}{36.1} \approx 0.1689$

4.2.C Plot of prior, likelihood, and posterior distributions



Question 5

5 Computed manually

Q.5. Bayesian inference for Gaussians (unknown mean and known variance).

Given:-

Avg. sag measurement = 5 cm

Sag measurement $\sim N(\mu, \sigma) \Rightarrow N(\mu, 0.25)$

Prior $\sim N(4, 2) \Rightarrow N(\mu, \sigma)$

a) Find posterior distribution for θ in terms of n.

Soln:- $p(\mu|x) = N(\mu|\mu_N, \sigma_N^2)$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} \; ; \; \mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_n$$

where $\sigma_0^2 \Rightarrow$ prior variance

$\mu_0 \Rightarrow$ prior mean

$\sigma^2 \Rightarrow$ posterior variance

Posterior mean:

$$\mu_N = \frac{0.0625}{N4 + 0.0625} \times 4 + \frac{N4}{N4 + 0.0625} \times 5$$

Posterior variance

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

$$\therefore \frac{1}{\sigma_N^2} = \frac{1}{4} + \frac{N}{0.0625}$$

b) $n = 20$, find mean and standard deviation of posterior.

Soln:- $$\mu_{20} = \frac{0.0625}{20 \times 4 + 0.0625} \times 4 + \frac{20 \times 4}{20 \times 4 + 0.0625} \times 5$$

$\mu_{20} \approx 5$ cm

$$\frac{1}{\sigma_{\mu 20}^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} = \frac{1}{4} + \frac{20}{0.0625} = 320.25$$

$$\therefore \sigma^2_{\mu_{20}} = \frac{1}{320.25}$$

$$\therefore \sigma^2_{\mu_{20}} \approx 0.003122$$

Posterior variance is less than prior and likelihood variance.

c) $n = 100$, find mean and standard deviation of posterior.

Sol^n: $\mu_{100} = \frac{0.0625}{100 \times 4 + 0.0625} \times 4 + \frac{20 \times 4}{100 \times 4 + 0.0625} \times 5$

$\therefore \mu_{20} \approx 5$ cm

$$\frac{1}{\sigma^2_{100}} = \frac{1}{4} + \frac{100}{0.0625} = 1600.25$$

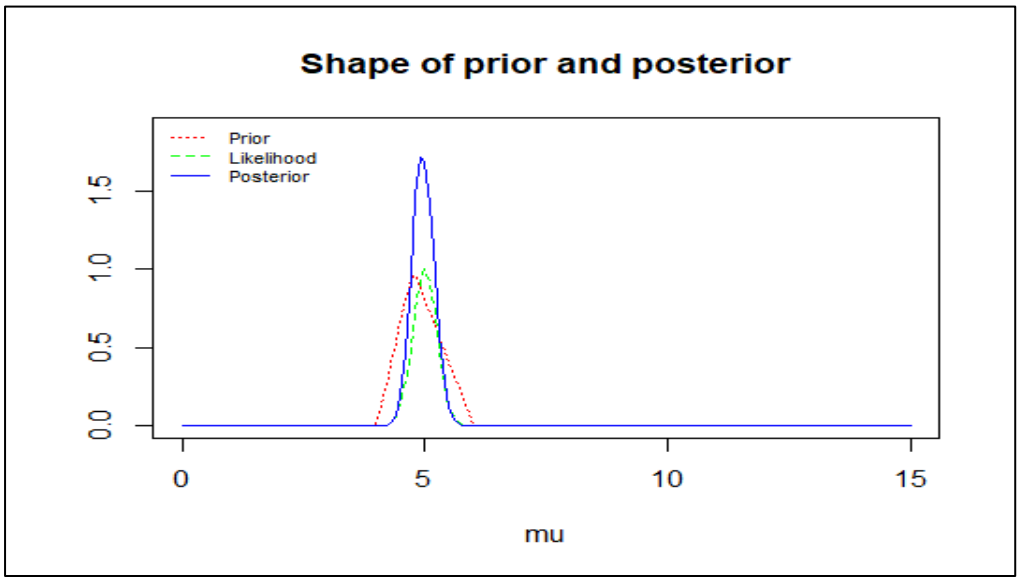
$$\therefore \frac{1}{\sigma^2_{100}} = 1600.25$$

$$\therefore \sigma^2_{100} \approx 0.0006249$$

As N increases, the precision $(1/\sigma^2_{\mu})$ increases, i.e. variance decreases.

5.d Triangle prior distribution

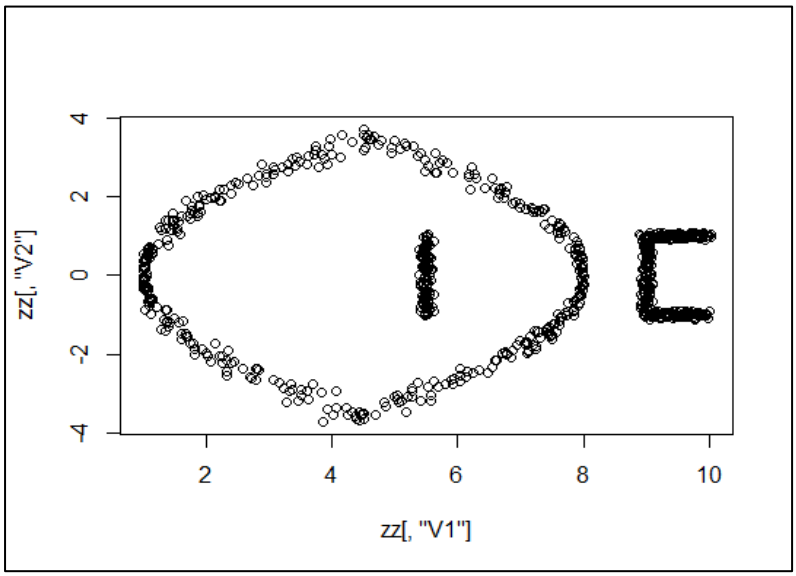
Posterior mean of theta is 4.96 cm



Question 6

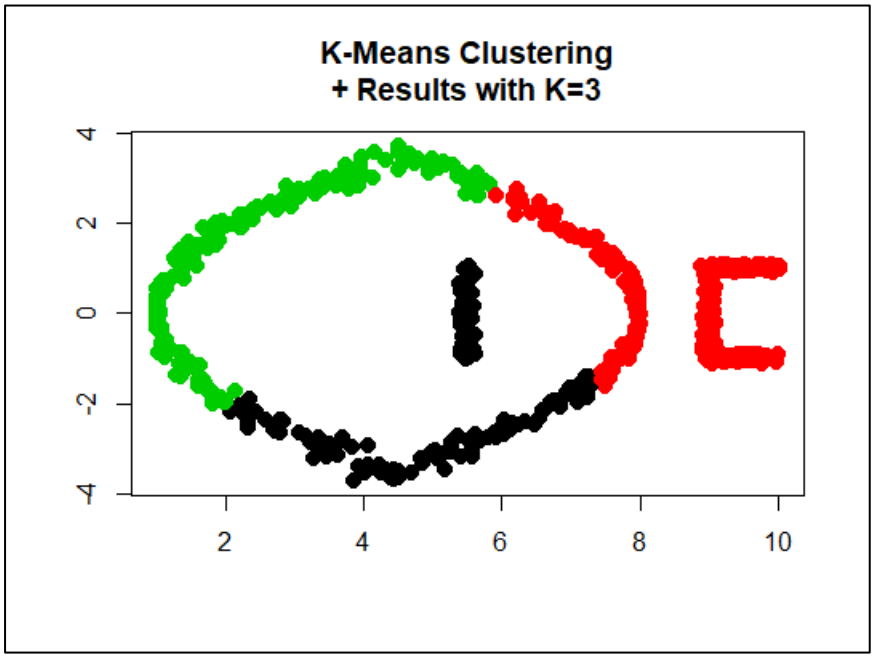
6.1 Kmeans clustering:

6.1.a Scatter plot of the data

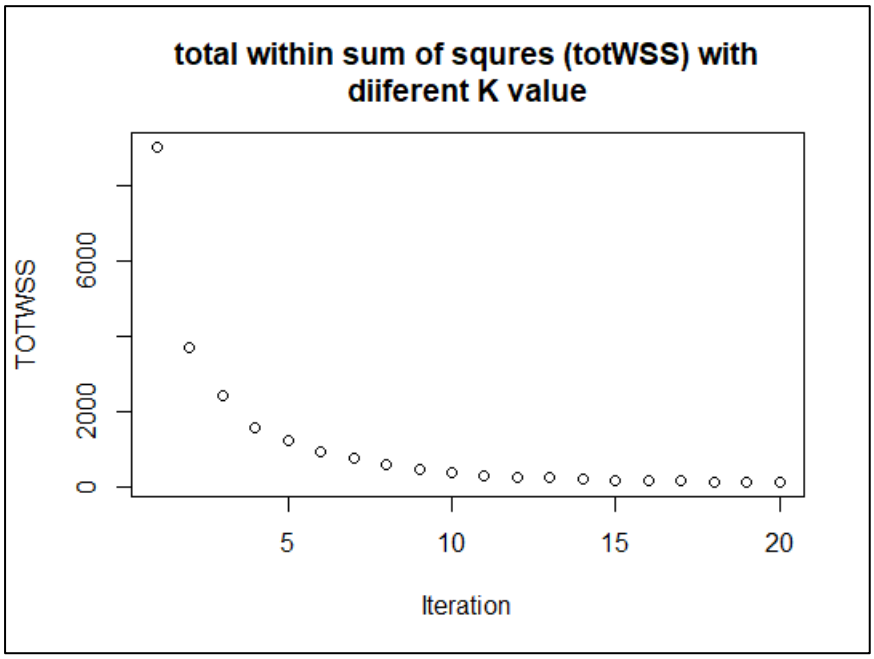


6.1.b Looking at the plot, 3 clusters can be found in the data.

6.1.c Scatter plot with 3 clusters. Since, shape of the scatter is spherical, it is difficult to identify which cluster a data point should go. For example, the line in the circle should be half green and half black.



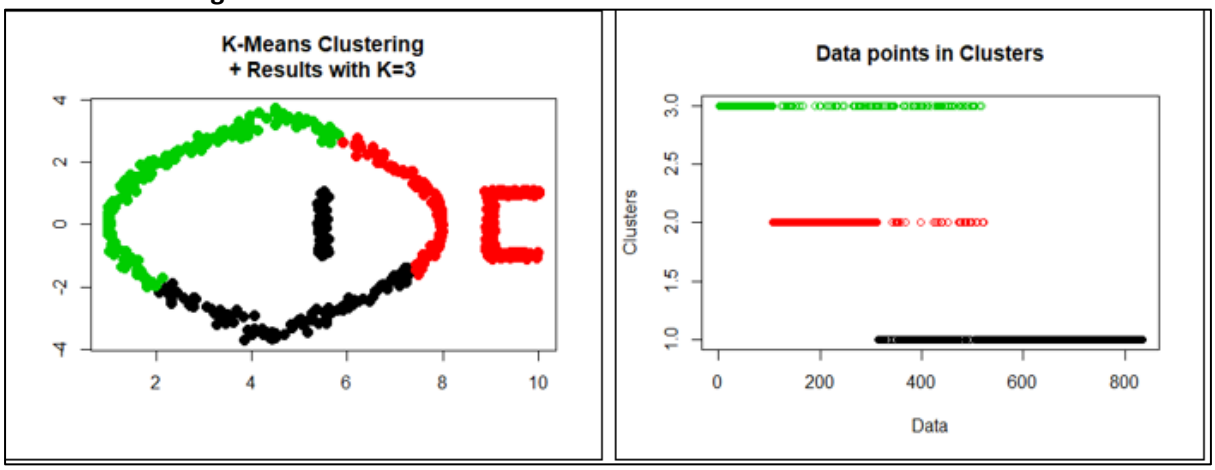
6.1.d As per graph, TOTWSS goes on decreasing as iterations increases. However, after 10 iterations, improvement in TOTWSS slows down. Thus, I would use 15 clusters where the curve becomes flat.



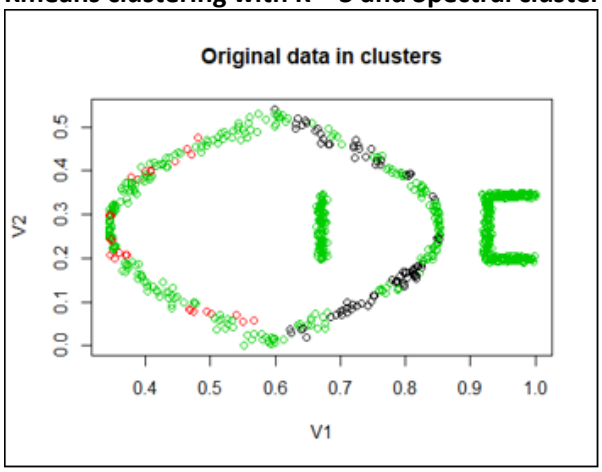
6.2 Spectral Clustering

First plot is from Kmeans clustering with K=3 and second plot is from Spectral clustering and Kmeans clustering with k=3. From both the plot we can infer, that Spectral clustering has more observations in cluster 3 (green color) whereas Kmeans clustering is evenly clustered with slightly more observations in cluster 1 (black). (Clusters might change due to randomness, but clustering pattern would be similar with one of the cluster having maximum data points assigned).

Kmeans clustering from 6.1 task

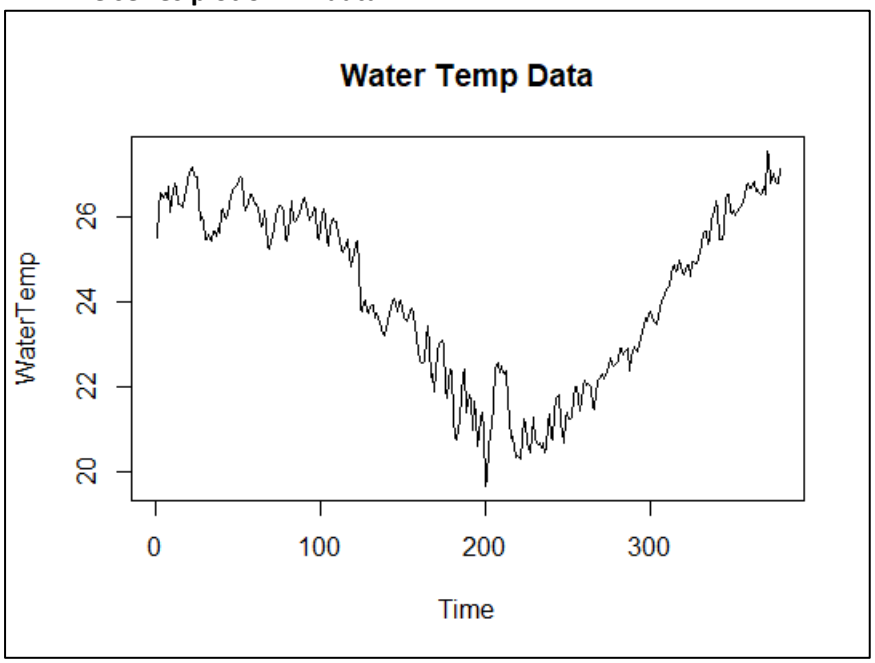


Kmeans clustering with K = 3 and Spectral clustering



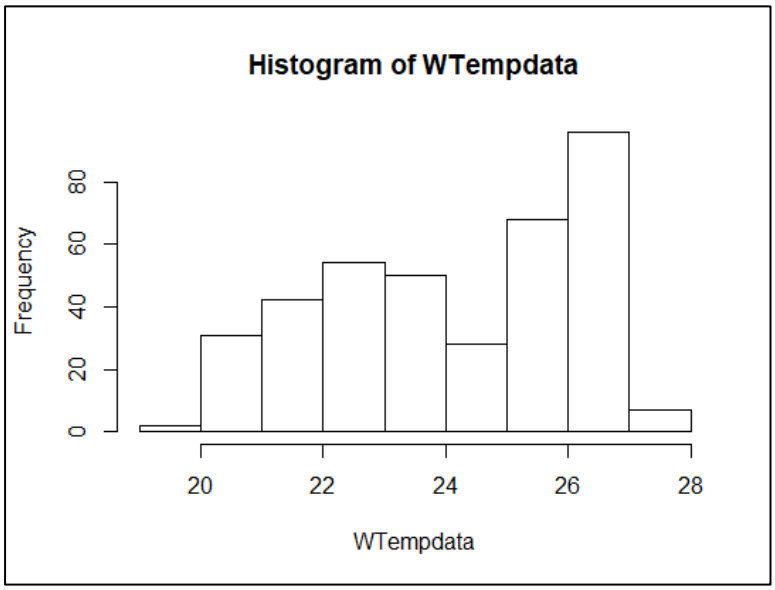
Question 7

7.1 Time series plot of WT data



7.2 Histogram of WT data

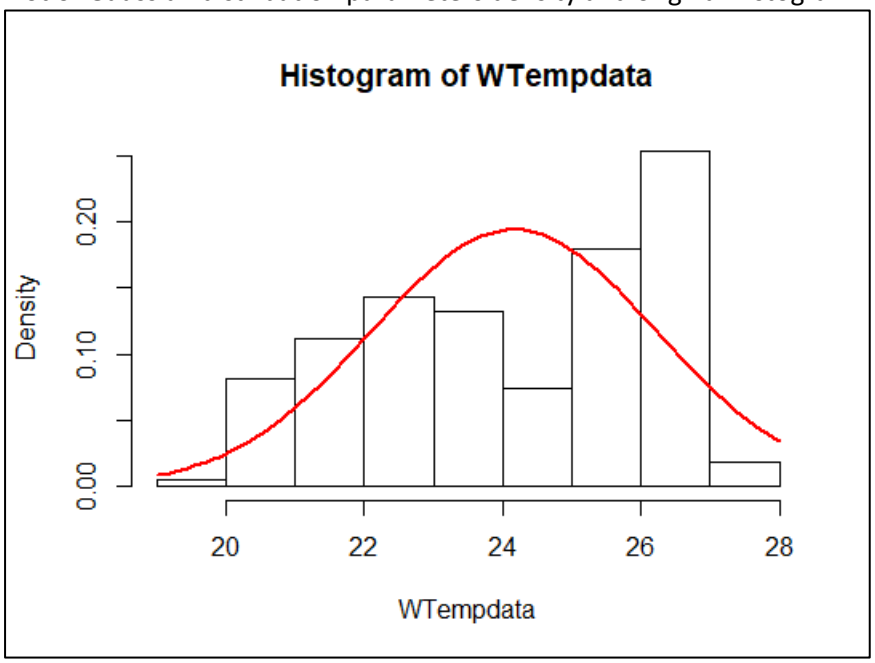
From histogram, we can infer that WT data is negatively skewed with two modes.



7.3 Fitting a single Gaussian model

MLE parameters of Gaussian distribution is same as its empirical mean and variance. Thus, MLE mean is 24.15 and standard deviation is 2.056

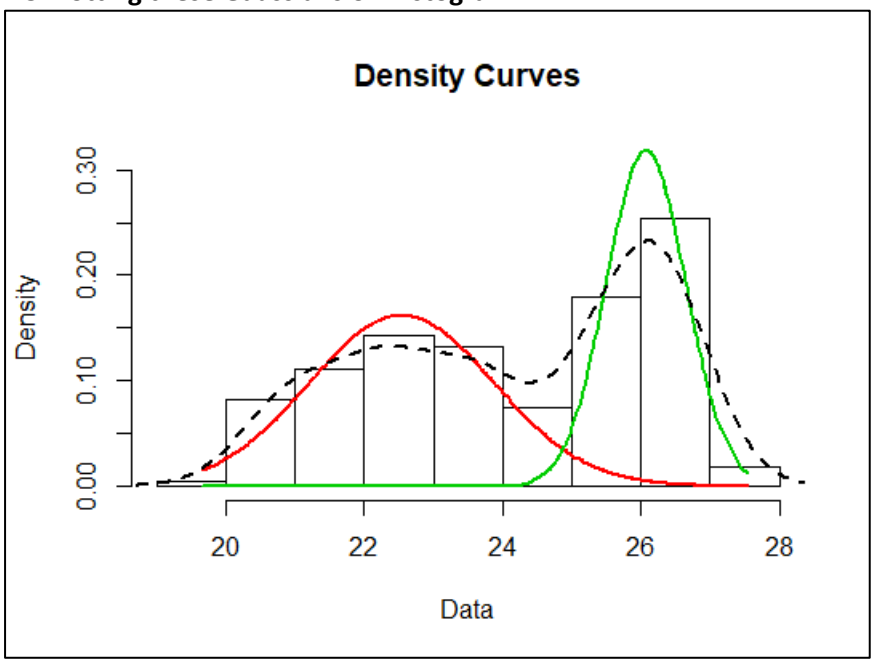
Plot of Gaussian distribution parameters density and original histogram



7.4 Fitting a mixture of two Gaussians as data is bimodal.

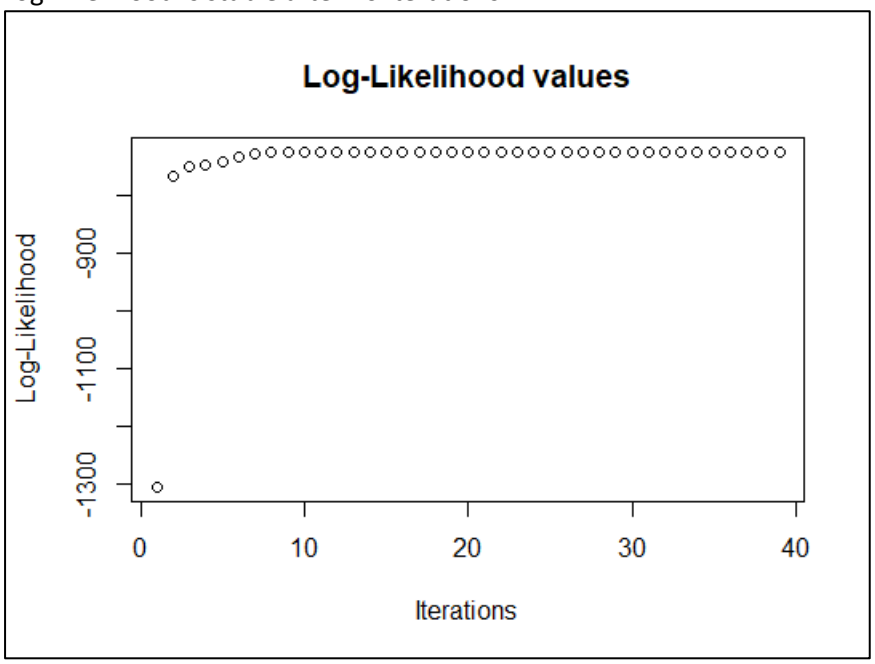
	Comp 1	**Comp 2**
Mixing coefficients	0.5420	0.4579
Mean	22.541	26.0747
Standard deviation	1.3357	0.5739

7.5 Plotting these Gaussians on histogram



7.6 Log-Likelihood values over iterations

Log-Likelihood is stable after 10 iterations.



7.7 Comment on the distribution models obtained in Q7.3 and Q7.4. Which one is better?

Data is bimodal and left skewed, thus, fitting a single Gaussian distribution will provide less accurate parameters. Distribution obtained in 7.4 is mixture of two Gaussians fits data better than single Gaussian and provides more accurate parameters. Data points are probabilistically assigned to either of the Gaussian distribution, thus we get an overall better fit and parameters. For eg. black density curve in 7.5 fits the data better than the red density curve in 7.3.

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?

Problems in maximizing likelihood using mixture of Gaussians:

1) **Presence of Singularities:**
   Consider a covariance matrix $\sum_K = \sigma^2 I$, If $\mu_K = x_n$ for some n value, i.e. mean is exactly equal to one of the data point.

$$N(x_n|x_n, \sigma_k^2 I) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma_j}$$

   If σ_j -> 0, then the term goes to infinity and so log-likelihood will also go to infinity which will pose a severe overfitting problem.
   To overcome this, use heuristics to detect this and reset the mean to a randomly chose value while resetting its covariance to some large value, and then continue with the optimization.

2) **Identifiability problem:**
   A K component mixture has a total of K! equivalent solutions
   - corresponding to the K! ways of assigning K sets of parameters to K solutions.
   - Example, if K = 3, then K! = 3! = 6. So, there are 6 possible ways to assign parameters for the 3 components.
   - i.e, for any given point in the space of parameter values there will be a further K!-1 additional points all giving exactly same distribution – This is known as Identifiability problem

   – Needs to be considered when parameters discovered by a model are interpreted.
   – However, for the purpose of finding a good density model, it is irrelevant because any of the equivalent solution is as good as any other.

3) **Complexity of maximizing the log likelihood in mixture of Gaussians:**
   Presence of summation over k that appears inside the log makes it harder. i.e., log function no longer directly acts on the Gaussian.

$$\ln p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k) \right\}$$

   Setting the derivative to zero will no longer results in a closed form solution and may have to use gradient based optimization.
   To overcome this, we can use Expectation Maximization algorithm. It is used to find maximum likelihood solutions for model with latent variables.