Wehrmuller-219369109-Ass1

Gabor Wehrmuller

24 08 2019

Q1)

Read in data,generate sample and write it to a .txt file:

```
thursday_data  <-  as.matrix(read.csv("C:/Users/gwehrm/Documents/Master/Deakin/Multivariate/Assignment_
                                      header  =  TRUE,
                                      sep = ","))

set.seed(18)# important, otherwise i get different results each time

my.data <- thursday_data[sample(1:15000,5000),c(1:6)]
setwd
```

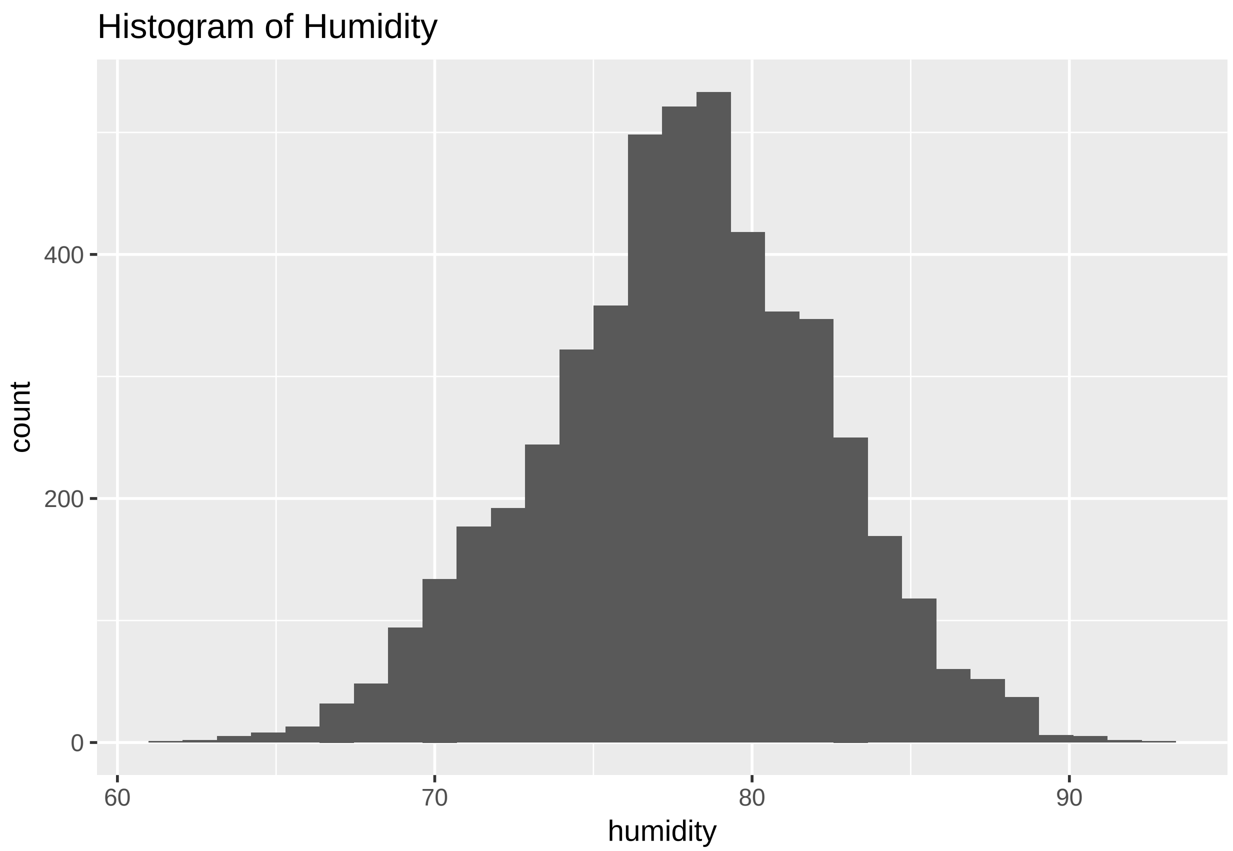
```
## function (dir)
## .Internal(setwd(dir))
## <bytecode: 0x0000000017844540>
## <environment: namespace:base>
```

```
write.table(my.data,"C:/Users/gwehrm/Documents/Master/Deakin/Multivariate/Assignment_1/Wehrmuller-21936
```

1.1)

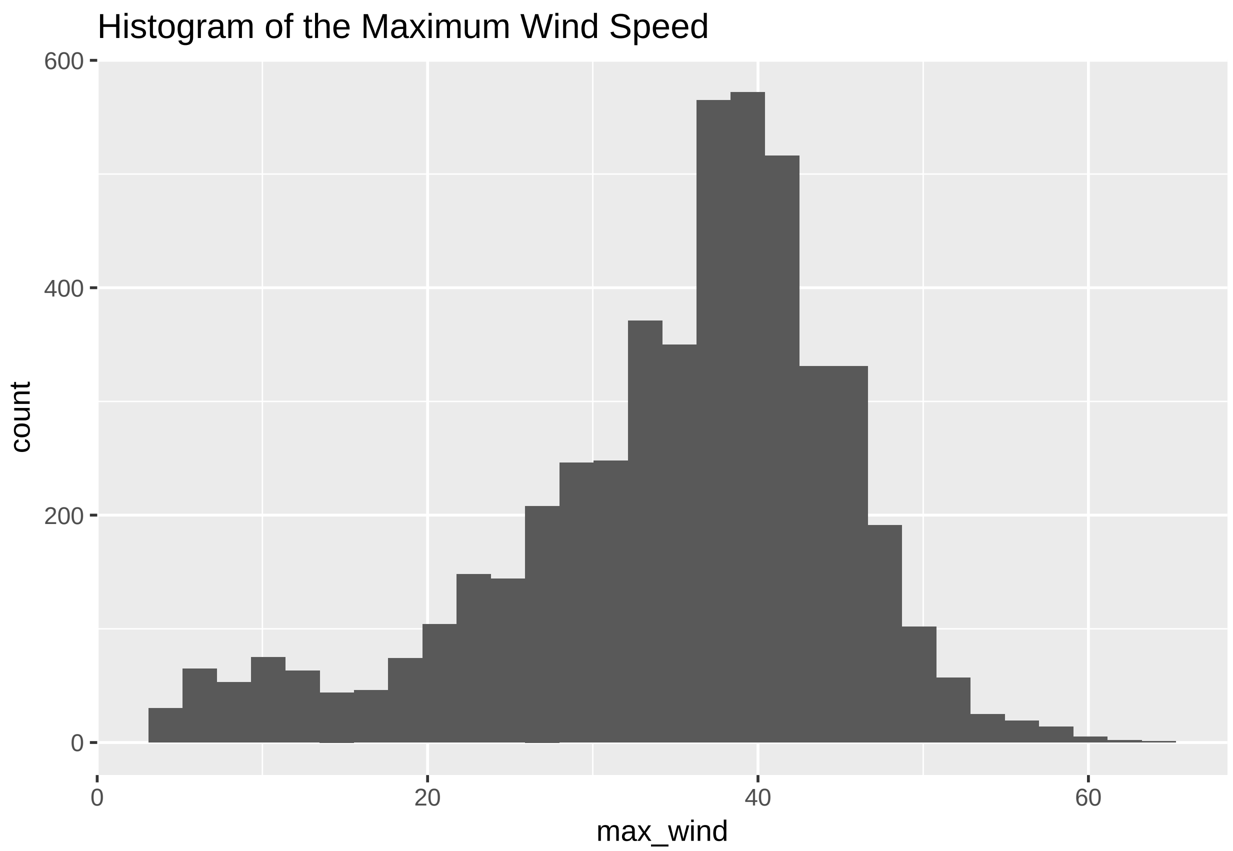
```
# convert the matrix to a dataframe - more convenient to draw plots with ggplot
my.data <- as.data.frame(my.data)
# change the column names to more readable ones
names(my.data) <- c("air_pressure","air_temp","humidity","max_wind","water_pressure","water_temp")
par(mfcol=c(1,2))# display both histograms on one line
ggplot(my.data, aes(x = humidity,y = ))+
  geom_histogram()+
  ggtitle("Histogram of Humidity")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(my.data, aes(x =max_wind, y =))+
  geom_histogram()+
  ggtitle("Histogram of the Maximum Wind Speed")
```

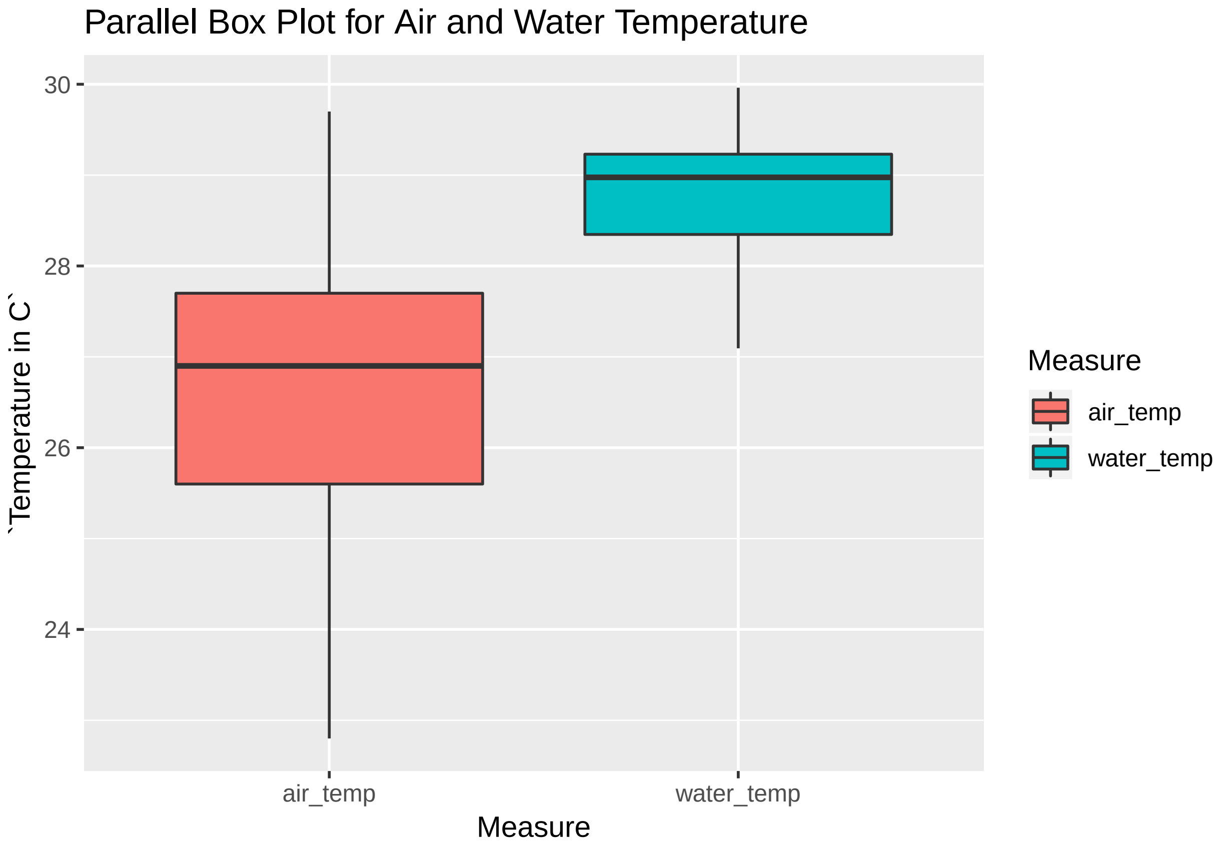
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The Humidity-histogram is unimodal and fairly symmetric . The tails are quite wide, but no specific outlier can be detected.
The Wind-Speed-histogram is less easily to interpret. With the hump around ten, the histogram is bimodal. The small hump around 35 seems more due to chance than it exist another mode. With this wide left tail, the distribution is not symmetric, but left skewed. But also here, nothing unusual or outliers can be detected.

1.2)

```r
# to properly plot, we need to reshape the data
par(mfcol=c(1,1))
my.data%>%
  dplyr::select(air_temp,water_temp)%>%
  gather(key = "Measure",
         value = "Temperature in C",
         air_temp:water_temp)%>%
  ggplot(aes(x = Measure,y = `Temperature in C`,fill = Measure))+
  geom_boxplot()+
  ggtitle("Parallel Box Plot for Air and Water Temperature")
```



```
par(mfcol=c(2,1))
summary(my.data$air_temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.80   25.60   26.90   26.76   27.70   29.70
```

```
summary(my.data$water_temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.09   28.35   28.98   28.78   29.23   29.96
```

From the boxplot and the five numbers summary the air temperature can be described as follows: The centre of the distribution is 26.76 (mean) with a median of 26.9. The IQR is 2.1 (27.7-25.6). The distribution ranges from 22.8 to 29.7. About the shape of the distributions can no comment be made with these plots. However, an indication, though not a reliable one, is that the median and mean are the same. It could be an indication the the distribution is symmetrical. From the boxplots there can not be any outlier detected

In the same way the water temperature can be described: The centre of the distribution is 28.78 (mean) with a median of 28.98 The IQR is 0.88 (29.23-28.35). The distribution ranges from 27.09 to 29.96. About the shape of the distributions can no comment be made with these plots. However, an indication, though not a reliable one, is that the median and mean are the same. It could be an indication the the distribution is symmetrical.From the boxplots there can not be any outlier detected

The difference between the two temperature measures in the range. The air temperature has a minimum of 22.8, whereas the water temperatures minimum is at 27.09. The maximum of them is nearly the same with 29.7 and 29.96. So, the water temperature has a more dense distribution. Both measures have also in common that the median is slightly bigger than the mean.

1.3)

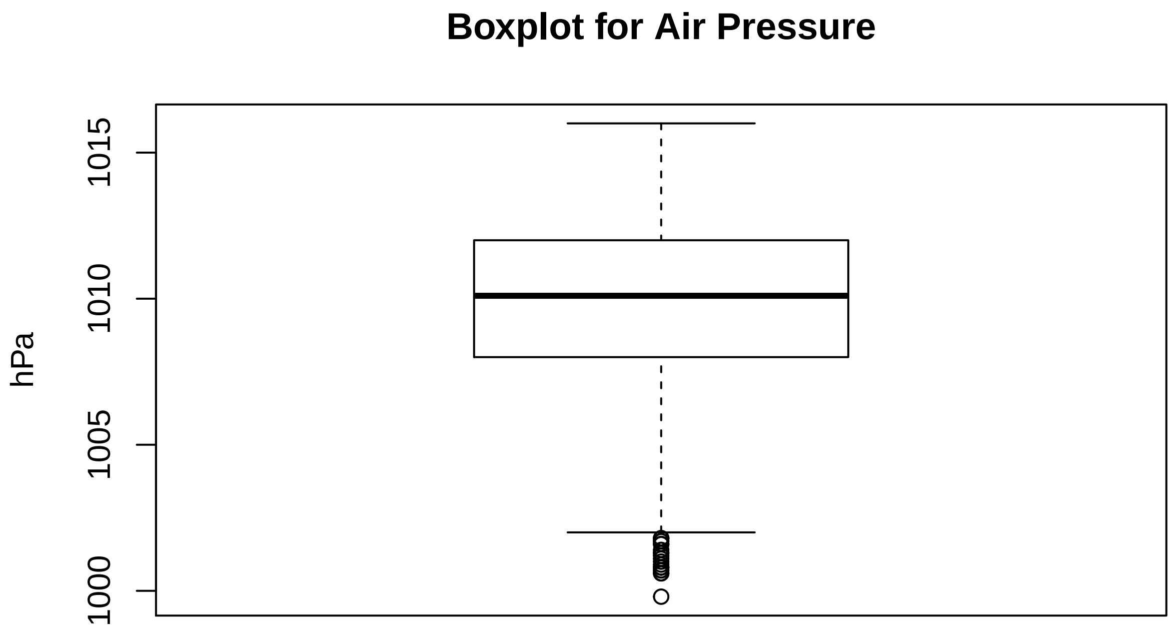
If only the center and spread are needed, then I would chose the five number summaries. The center of a distribution is described by the mean or median with the IQR and the range by the minimum and maximum. All these measures can be seen or derived by the five number summary.

```
summary(my.data$air_pressure)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   999.8  1008.0  1010.1  1009.8  1012.0  1016.0
```

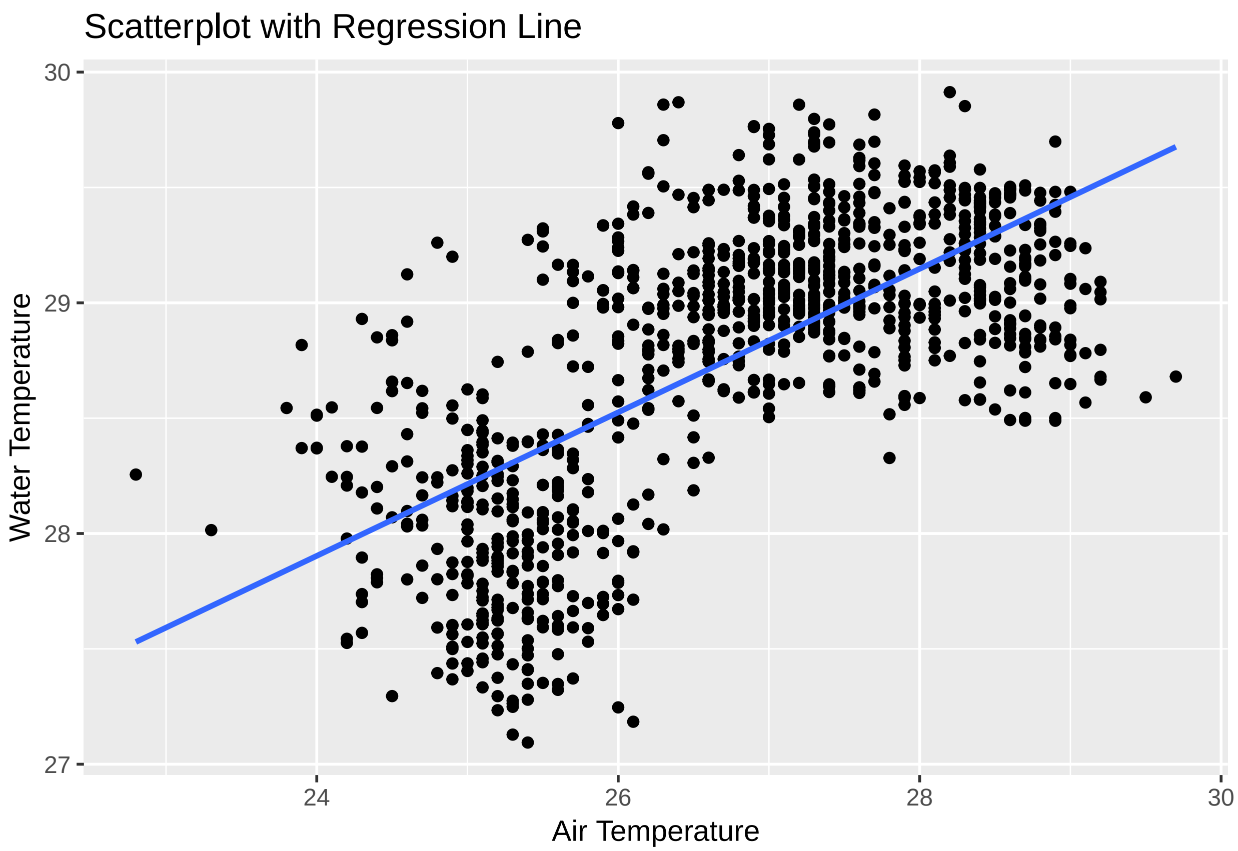
If a plot is asked, then I would propose the boxplot, since it shows easily the center and spread

```
boxplot(my.data$air_pressure, ylab = "hPa", main = "Boxplot for Air Pressure")
```



1.4)

```
top_1000 <- my.data[1:1000,]
top_1000 %>%
  ggplot(aes(x = air_temp,y=water_temp))+
  geom_point()+
  ggtitle("Scatterplot with Regression Line")+
  xlab("Air Temperature")+
  ylab("Water Temperature")+
  geom_smooth(method = "lm",se = F)
```



```
fit_1 <- lm(water_temp ~ air_temp, data = top_1000)
summary(fit_1)
```

```
##
## Call:
## lm(formula = water_temp ~ air_temp, data = top_1000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37223 -0.31661  0.03634  0.30756  1.25378
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.43782    0.29000   70.47   <2e-16 ***
## air_temp     0.31106    0.01082   28.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4524 on 998 degrees of freedom
## Multiple R-squared:  0.4531, Adjusted R-squared:  0.4526
## F-statistic:   827 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
cor(top_1000$water_temp,top_1000$air_temp)
```

```
## [1] 0.6731571
```

$$\hat{y} = 20.43782 + 0.31106x + \epsilon$$

The correlation between the two variables is 0.6731571. This coefficient is large (depending on discipline) and positiv, which indicates a stong positive realtionship between air and water temperature.

The coefficient of Determination, here R^2 is 0.4531. This value indicates how much of the variance of the dependent variable is explained by the variance of the independent variable. This means that by a total of 44.39% of the variance of the Water Temperature can be explained by the Air Temperature.

Q2)

2.1)

a)

P(J) = 4450/8000 = 0.55

b)

P(C,N) =800/8000 = 0.1

c)

P(J|V) = 1200/2200 = 0.55

d)

P(Q|C) =1800/3600 = 0.5

e)

P(J or V) =(2200+4400-1200)/8000 = 0.678

f)

marginal distribution of the poweboat type:

Type	Marginal
Cruiser	0.45
Jet Boat	0.55

g)

No, they are not mutually exclusive, since it is possible that they can occur at the same time. As seen in b):P(J,N) =800/8000 = 0.1. and $\neq 0$

h)

P(C|N) = 0.4

P(C) = 0.45

P(C|N) $\neq$ P(C), therefore they are not independent. The people from New south Wales are less likely to have a cruiser than overall.

2.2)

What is the posterior probability that the withe ball came from the black box?

B = Black box W = White ball

$P(B|W) = \frac{P(W|B)*P(B)}{P(W)}$

$P(W) = P(W|B) * P(B) + P(W|B') * P(B')$
$= 0.57 * 0.3 + 0.44 * 0.7$
$= 0.479$

Thus :

$P(B|W) = \frac{0.171*0.3}{0.479} \approx 0.11$

The posterior probability that a white ball came from the black box is 0.11.

Q3)

3.1)

The first difference is about the estimated Parameter. While in the Bayesian way, the parameter is considered a random variable, in the Frequentist way, it is seen as a fixed parameter. This leads also to the second difference. The Frequentists give an uncertainty for the obtained estimate. The Bayesian way however, delivers a way more informative result. It returns a complete probability distribution of the estimate and allows a more intuitive interpretation of the results.

3.2)

It is a good thing, since the computation is much more efficient. It is more efficient, because only the parameters of the distribution need to be updated. If no conjugate pars were involved, then the computations needs to be done numerically and would thus be more demanding.

3.3)

1) The normal distributions:

$$\begin{gathered} Prior\ \theta\ \mathcal{N}(\mu, \Gamma^2) \\ Likelihood:\quad X|\theta\ \mathcal{N}(\theta, \sigma^2) \end{gathered} \tag{1}$$

Where μ is the mean and Γ^2 the variance of the prior.

For the Likelihood is then θ the mean and σ^2 the variance of the likelihood.

2) Binomial and Beta distribution:

$$\begin{gathered} Prior - Beta:\ p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ Likelihood - Binomial:\ p(D|\theta) = p(x_{1:n}|\theta) = \theta^m(1-\theta)^{n-m} \end{gathered} \tag{2}$$

These two pairs ensure that the posterior follows the same distribution as the prior.

Q4)

4.1)

a)

$$\begin{aligned} Att(\theta) = p(x_i|\theta) = (1-\theta)^{x_i}\theta \\ p(X|\theta) = p(x_{i:N}|\theta) = p(x_1|\theta)p(x_2|\theta)p(x_3|\theta)...p(x_N|\theta), since\ iid. \\ = \Pi_{i=1}^{N} p(x_i|\theta) \\ = \Pi_{i=1}^{N} p(1-\theta)^{xi}\theta \\ = (1-\theta)^{x_1}\theta * (1-\theta)^{x_2}\theta * ...(1-\theta)^{x_N}\theta \\ = (1-\theta)^{(x_1+x_2+...+x_N)}\theta^{1+1..+1} \\ = \theta^N(1-\theta)^{\sum_{i=1}^{N} x_i} \\ = P(X|\theta) = \theta^N(1-\theta)^S,\ where\ S = \sum_{i=1}^{N} x_i \end{aligned} \tag{3}$$

b)

$$\begin{aligned} L(\theta) = ln(p(X|\theta)) \\ = ln(\theta^N(1-\theta)^S) \\ = ln(\theta^N) + ln((1-\theta)^S) \\ = N * ln(\theta) + S * ln(1-\theta) \end{aligned} \tag{4}$$

c)

$$\begin{aligned} \frac{dL(\theta)}{d\theta} = \frac{d}{d\theta}[Nln(\theta) + Sln(1-\theta)],\ where\ S = \sum_{i=1}^{N} x_i \\ = \frac{N}{\theta} + S\frac{-1}{1-\theta} \\ = \frac{N(1-\theta) - S\theta}{\theta(1-\theta)} \end{aligned} \tag{5}$$

$$\frac{dL(\theta)}{d\theta} = 0$$

only the numerator can equal to zero, therefore:

$$
\begin{gathered}
N(1-\theta) - S\theta = 0, \; where \, S = \sum_{i=1}^{N} x_i \\
N - N\theta - S\theta = 0 \\
N = \theta(N+S) \\
\frac{N}{N+S} = \theta \\
\frac{N}{N(1+\frac{1}{N}S)} = \theta \\
\frac{1}{(1+\frac{1}{N}S)} = \theta \\
\hat{\theta} = \frac{1}{1+\bar{X}}, \; where \, \bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i
\end{gathered} \tag{6}
$$

d)

plug in for $\hat{\theta}$: $\hat{\theta} = \frac{1}{1+\frac{1}{5}*10} \approx 0.33$

Given this data, the MLE is 0.33.

e)

Assuming that the data scientist needs to fix one bug, the question asked what is the likelihood that the data scientists successfully fixes the bug after zero or after the first,second or third try.

Given $\theta = 0.33$, plug in distribution for each in $x_i = \{0, 1, 2, 3\}$ and then add together:

$$
\begin{gathered}
p(x_i|\theta) = (1-\theta)^{x_i}\theta \\
= 0.33 + 0.2211 + 0.148137 + 0.09925179 \approx 0.79
\end{gathered} \tag{7}
$$

Therfore, the chance that the data scientist fixes the bug after one day is 0.79. Good job!

4.2)

a)

The steps involved:

1. Likelihood function
2. Prior
3. Compute posterior

Likelihood function:

From 4.1 a):

$$
p(x_i|\theta) = \theta^N (1-\theta)^S, \; where \, S = \sum_{i=1}^{N} x_i \tag{8}
$$

Prior:

$$\begin{gathered}
Beta(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\ is\ a\ constant \\
Beta(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}
\end{gathered} \tag{9}$$

Computation of the posterior:

$$\begin{gathered}
posterior: \\
p(\theta|D) \propto p(D|\theta)p(\theta) \\
= \theta^{N}(1-\theta)^{S} * \theta^{a-1}(1-\theta)^{b-1} \\
= \theta^{N+a-1}(1-\theta)^{S+b-1} \\
= \theta^{a'-1}(1-\theta)^{b'-1},\ where \\
a' = N + a\ and \\
b' = S + b \\
p(\theta|D) = C\theta^{a'-1}(1-\theta)^{b'-1}\ where\ C = \frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')}
\end{gathered} \tag{10}$$

b)

a' = N+a and b' = S+b

a = 1
b = 10
N = 5
$S = \sum_{i=1}^{N} x_i = 10$

Therefore:
a' = 5+1 = 6

b' = 10 + 10 = 20

Mean: $\hat{\theta} = E(\theta|D) = mean[Beta(\theta|a', b')] = \frac{a'}{a'+b'} = \frac{6}{6+20} = 0.2307692$

c)

The number of attempts made by five data scientists: 2,0,4,1,3

$$p(x_i|\theta) = \theta^{N}(1-\theta)^{S},\ where\ S = \sum_{i=1}^{N} x_i$$

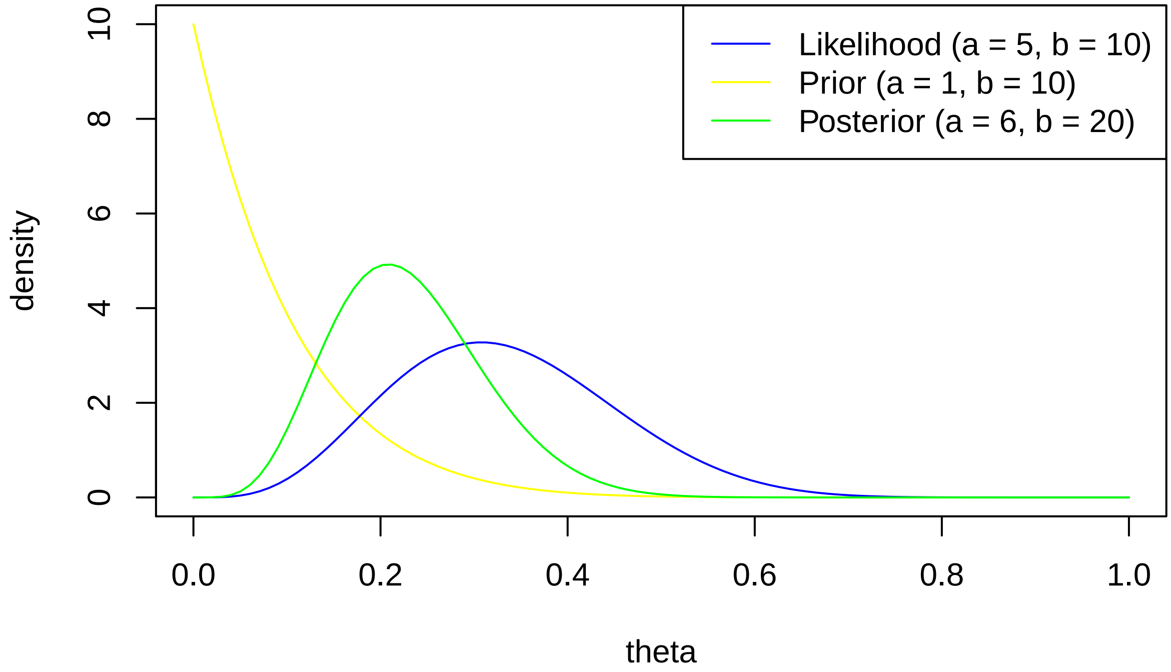
Therefore for the likelihood function:

N = 5 = a

S = 10 = b

```r
# set up sequence
theta <- seq(0,1, length=100)

# prior
plot(theta, dbeta(theta, 1, 10), ylab="density", type ="l", col="yellow")

# likelihood
```

```
lines(theta, dbeta(theta,5,10),ylab="density", type ="l", col="blue")

# posterior
lines(theta, dbeta(theta, 6, 20), ylab="density", type ="l", col="green")
#title(xlab="theta")

# legend
legend("topright",
  legend = c("Likelihood (a = 5, b = 10)", "Prior (a = 1, b = 10)","Posterior (a = 6, b = 20)"),
  col = c("blue","yellow","green"),
  lty = 1)
```



d)

uniform prior: Beta(1,1)
again the calculation of a' and b':

a' = N+a and b' = S+b

a = 1
b = 1 N = 5
$S = \sum_{i=1}^{N} x_i = 10$

Therefore: a' = 5+1 = 6

b' = 10 + 1 = 11

Mean: $\hat{\theta} = E(\theta|D) = mean[Beta(\theta|a', b')] = \frac{a'}{a'+b'} = \frac{6}{6+11} = 0.3529412$

The estimate of theta is slightly larger than the one obtained purely by the MLE. The posterior was slightly corrected. Also the estimated value is not the MAP, the estimate is still different from the MLE, even with an uniformed prior.

Q5)

a)

n = ?
$\mu_{ML} = 80$

$\mu = \theta$

$\sigma = 5$

prior for theta: $\mu_0 = 100$

$\sigma_0 = 10$

posterior: $p(\mu|x) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$

$\theta_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}$

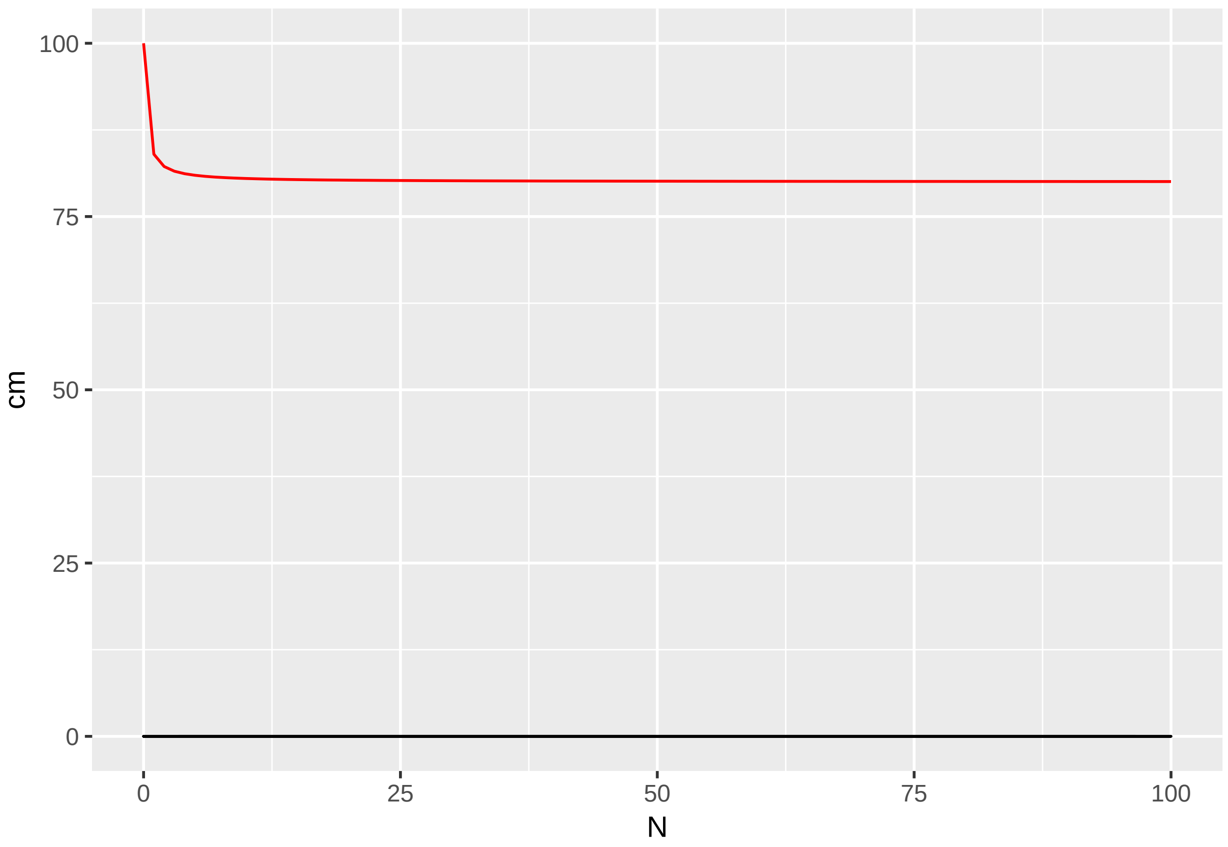
plug in values from above:

$\theta = \frac{2500+8000N}{25+100N}$

other Option would be to get the posterior variance first and then based on that calculate the mean:

$2.44(\frac{80*N}{25} + \frac{100}{100})$

```r
post_mean <- function(N){
  y = ( (25/(25+100*N))*100 + (100*N/(25+100*N))*80)
  return(y)
}
ggplot(data.frame(N =seq(0,100,by = 1)), aes(N)) +
  geom_density() +
  stat_function(fun = post_mean, colour = "red")+
  ylab("cm")
```



b)

n = 10:

$\theta = \frac{2500+8000*10}{25+100*10} = 80.4878$

$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$ plug in: $\frac{1}{\sigma_N^2} = \frac{1}{10^2} + \frac{10}{5^2} = 0.41$ $sigma_N^2 = \frac{1}{0.41} = 2.439024$

The posterior mean is 80.49 cm and the posterior variance 2.44. The posterior variance is smaller than the one of the prior or likelihood. With increasing N , an even smaller variance is expected.

c)

N = 200:

$\theta = \frac{2500+8000*200}{25+100*200} = 80.02497$ $\frac{1}{\sigma_N^2} = \frac{1}{10^2} + \frac{200}{5^2} = 8.01$ $sigma_N^2 = \frac{1}{8.01} = 0.1248439$

The posterior mean is 80.02 cm and the variance is 0.12

Compared to 5 b)'s result, the variance decreased, as expected. As N increases, the variance goes towards zero and the posterior distribution is only around the maximum likelihood estimation.
the more weight has the maximum likelihood solution.

d)

```
library(Bolstad)
set.seed(123)
```

```
# one observation
n = rnorm(1,80,5)

# mu values
mu = seq(65, 95, by = 0.1)

mu.prior = rep(0, length(mu))
mu.prior[mu <= 80] = (1/225)*mu-13/45
```

```
## Warning in mu.prior[mu <= 80] = (1/225) * mu - 13/45: number of items to
## replace is not a multiple of replacement length
```

```
mu.prior[mu > 80] = -(1/225)*mu+19/45
```

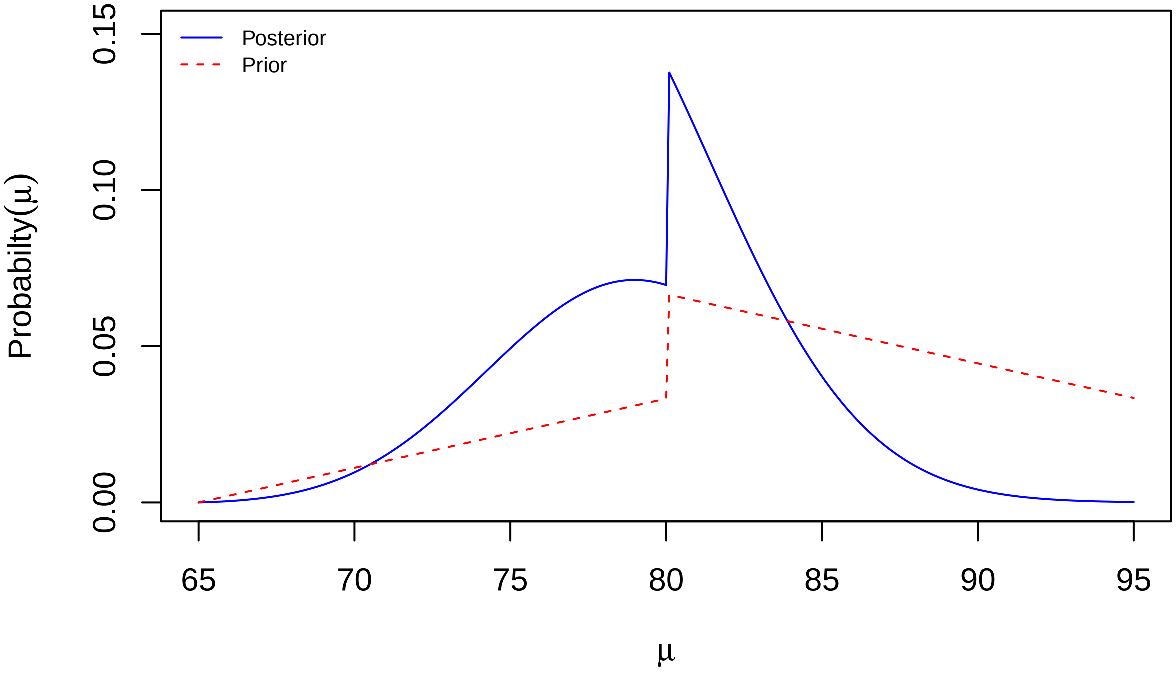
```
## Warning in mu.prior[mu > 80] = -(1/225) * mu + 19/45: number of items to
## replace is not a multiple of replacement length
```

```
results = normgcp(n, 5, density = "user", mu = mu, mu.prior = mu.prior)
```

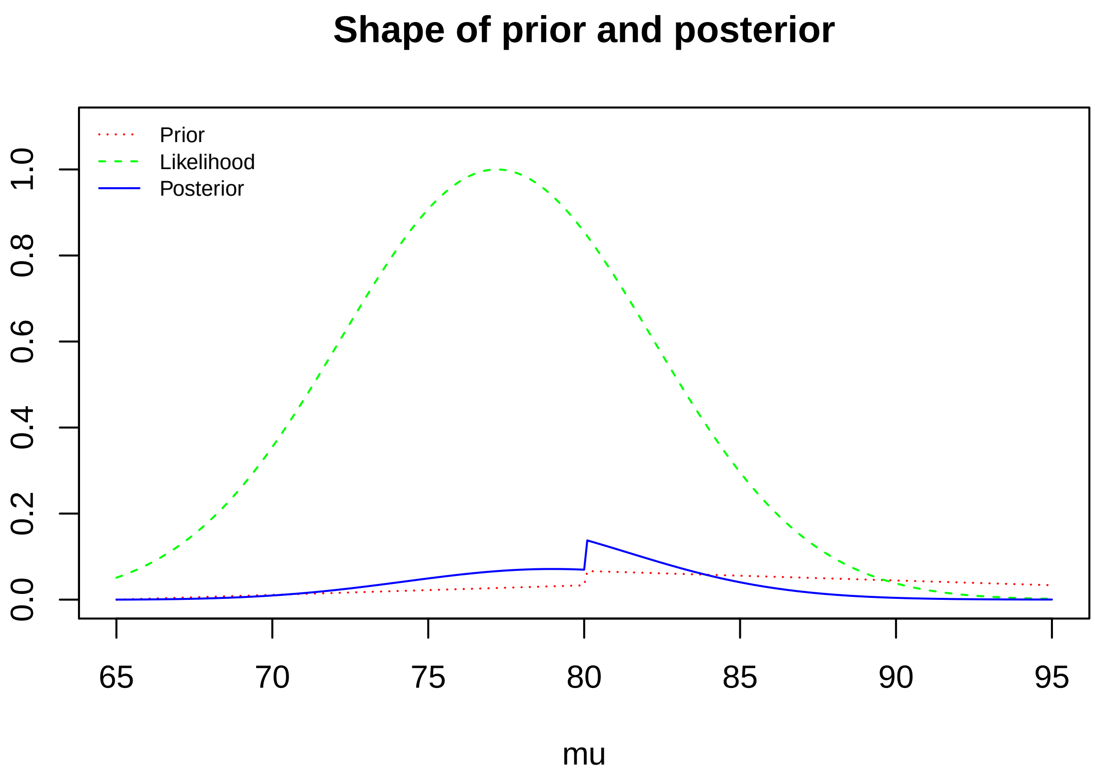
```
## Known standard deviation :5
```

```
## Warning in normgcp(n, 5, density = "user", mu = mu, mu.prior = mu.prior):
## The prior probabilities did not sum to 1, therefore the prior has been
## normalized
```

```
## [1] 2.006667
```



```
#
plot(results,overlay = TRUE, which = 1:3)
```



```
# on different axes
#decomp(results)

# posterior mean
print("posterior mean:")
```

```
## [1] "posterior mean:"
```

```
mean(results)
```

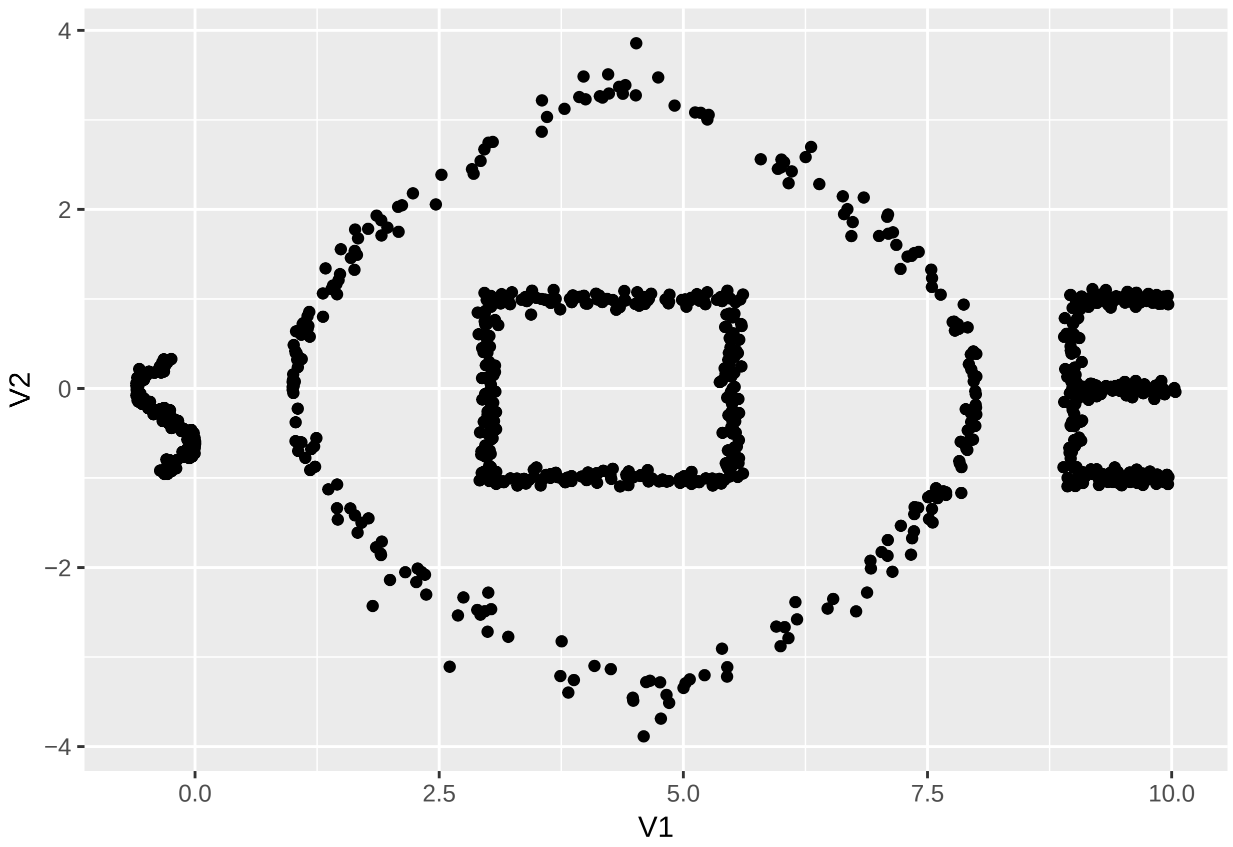
```
## [1] 79.656
```

Q6)

6.1)

a)

```
zz <- read.table("C:/Users/gwehrm/Documents/Master/Deakin/Multivariate/Assignment_1/Data/SITEdata2019Au
zz <- as.matrix(zz)
zz_df <- data.frame(zz)
ggplot(zz_df,aes(x  =V1,y=V2))+
  geom_point()
```

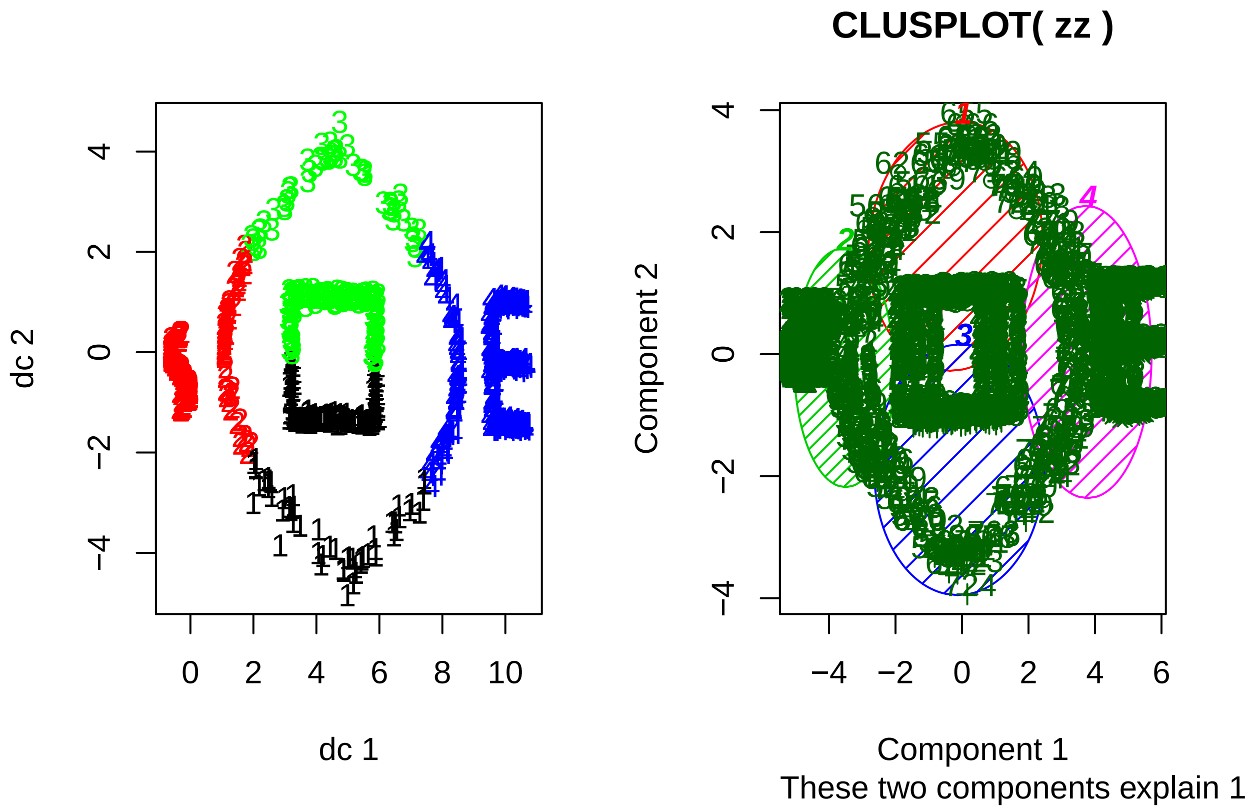


b)

By visual examination I see in total four clusters (S,square, circle and E).

c)

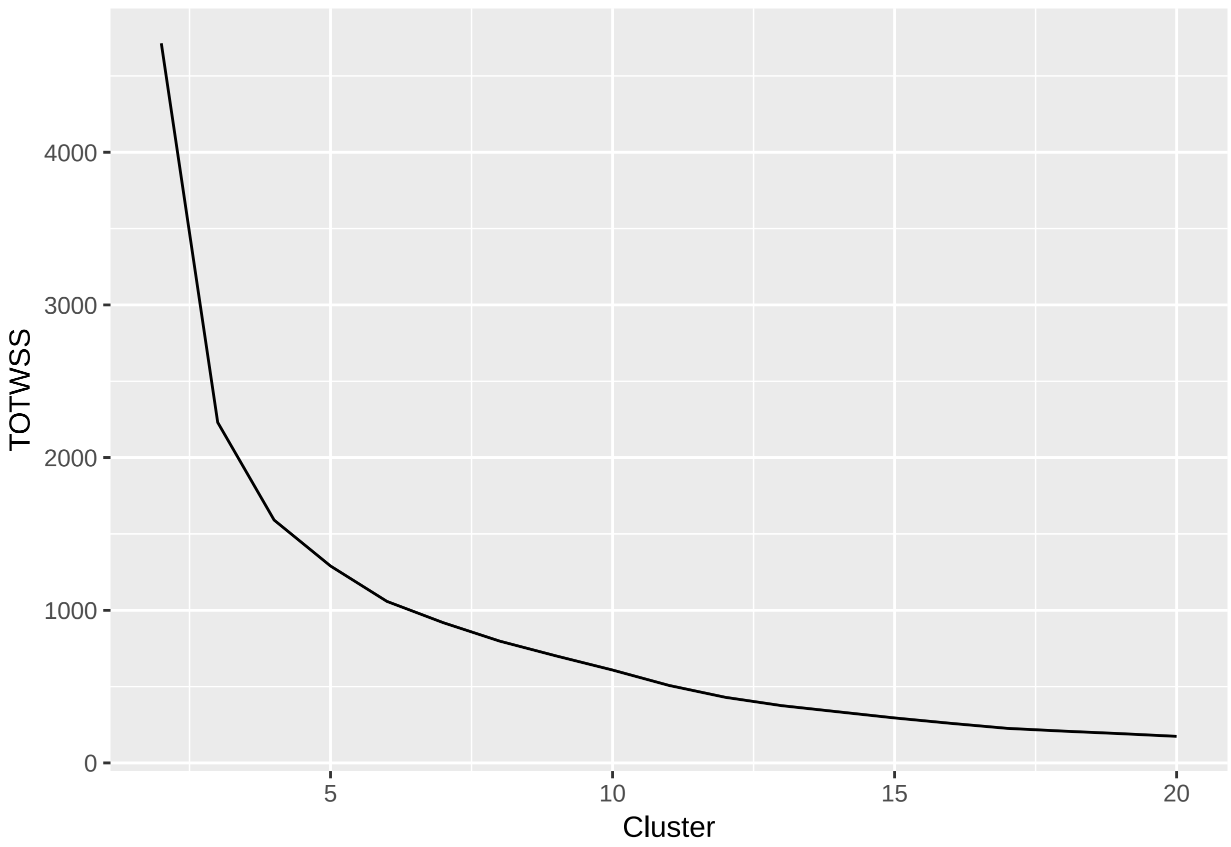
```
zz_clustered <- kmeans(zz,centers = 4, nstart = 100)
library(fpc)
library(cluster)
par(mfcol = c(1,2))
plotcluster(zz,zz_clustered$cluster)
clusplot(zz,zz_clustered$cluster,color = T,shade = T, labels = 2, lines = 0)
```



The clusters obtained, as shown by the two plots, do not statisfy the requirements. The kmeans fails to diverge the groups in the four found groups. This is caused due to the k-means algorithm which is not suitable for non-convex shapes

d)

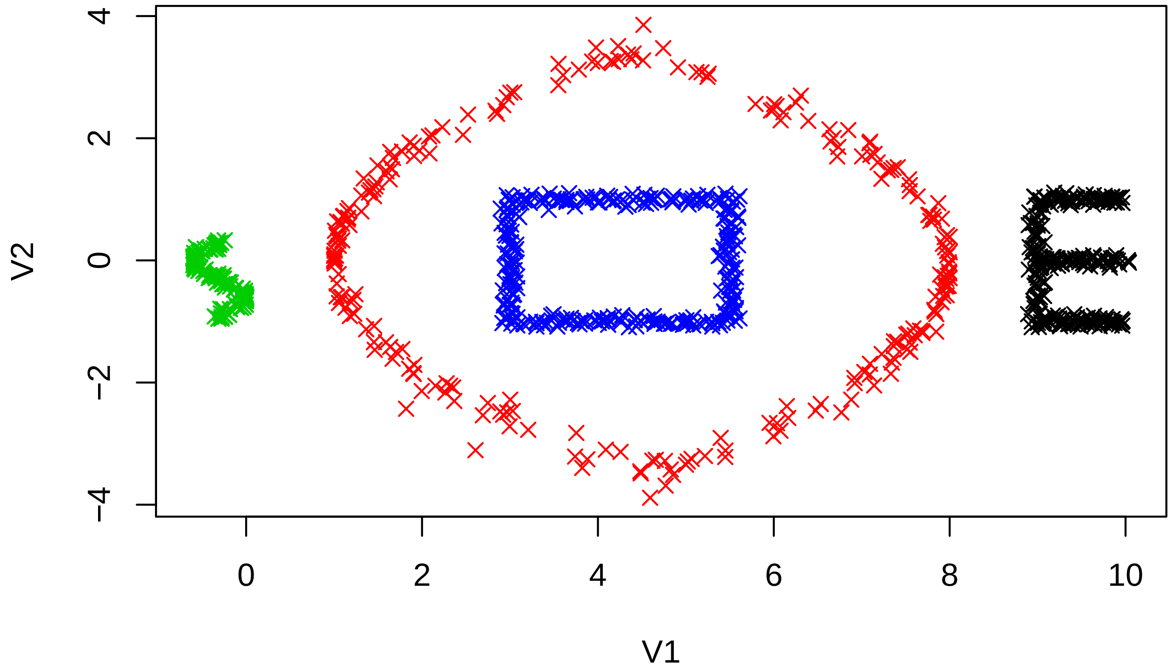
```
df_kmeans <- data.frame(Cluster = integer(),
                        TOTWSS = integer())

for (i in c(2:20)){
  clustered <- kmeans(zz,centers = i, nstart = 100) # a high nstart number results in a smooth(er) line
  df_kmeans[i-1,1]=i
  df_kmeans[i-1,2]=clustered$tot.withinss
}
ggplot(df_kmeans,aes(x=Cluster,y = TOTWSS))+
  geom_line()
```



Generally the smaller the total WSS the better. So the optimal cluster can be found when adding another, it doesnt reduce the TOTWSS that much. By the “elbow method”, the optimal k is where the corner is. In that case four clusters.

6.2)

```
set.seed(123)# set seed for rep
library(kernlab)
sc <- specc(zz, centers=4)
plot(zz, col=sc, pch=4)
```



```
centers(sc)
```

```
##           [,1]          [,2]
## [1,]  9.366113 -0.0008976253
## [2,]  4.500000 -0.0282465129
## [3,] -0.300000 -0.3099575815
## [4,]  4.250163 -0.0002305581
```

```
size(sc)
```

```
## [1] 249 248 248 282
```

```
withinss(sc)
```

```
## [1] 21928.70030  7753.86076    44.70159  5579.14745
```

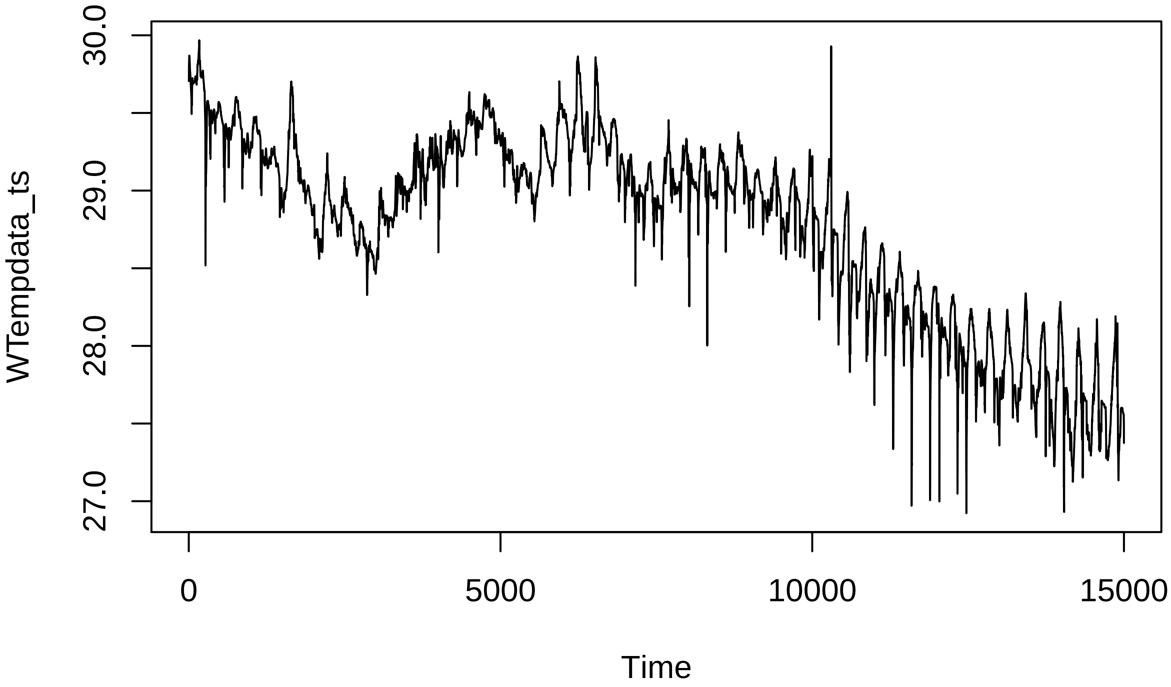
The spectral clustering delivers the desired results. This is possible, since the undirected graph concept is used to then finally cluster it. This is helpful, since it

Q7)

```
the.data <- as.matrix(read.csv("C:/Users/gwehrm/Documents/Master/Deakin/Multivariate/Assignment_1/Data/
                               header = TRUE,
                               sep = ","))
WTempdata <- the.data[,6]
```

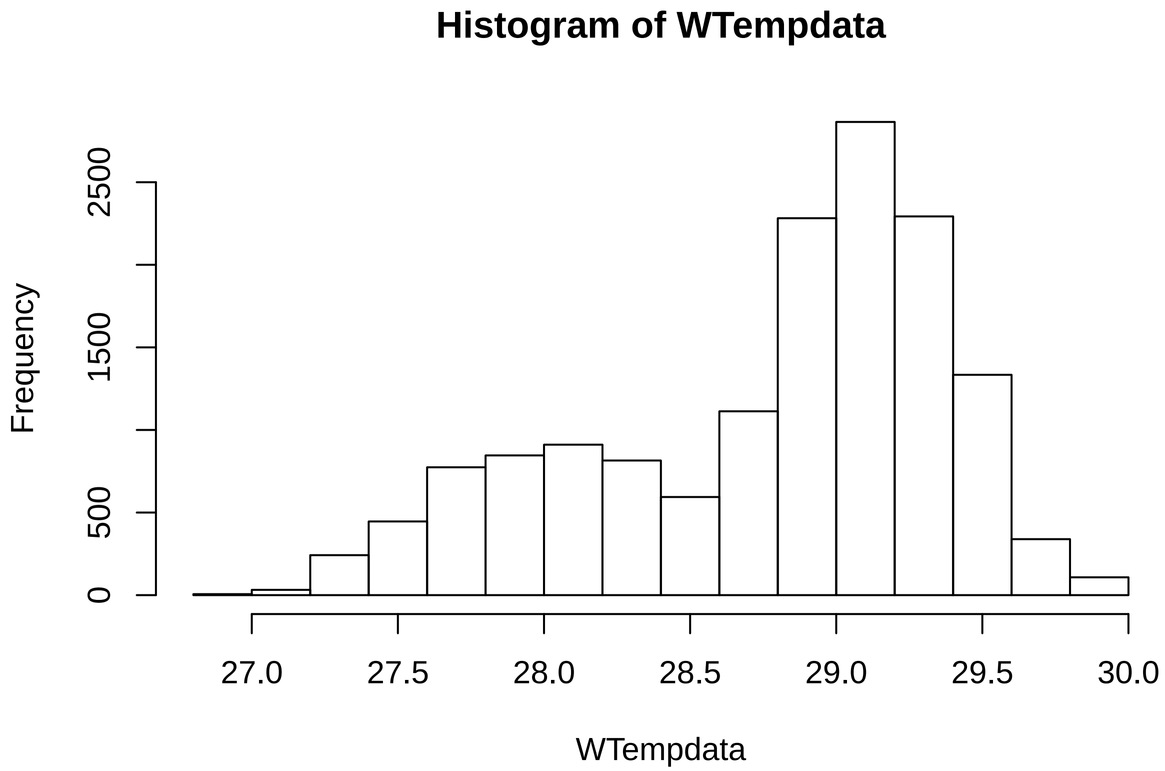
7.1)

```
WTempdata_ts <- ts(WTempdata, )
plot(WTempdata_ts)
```



7.2)

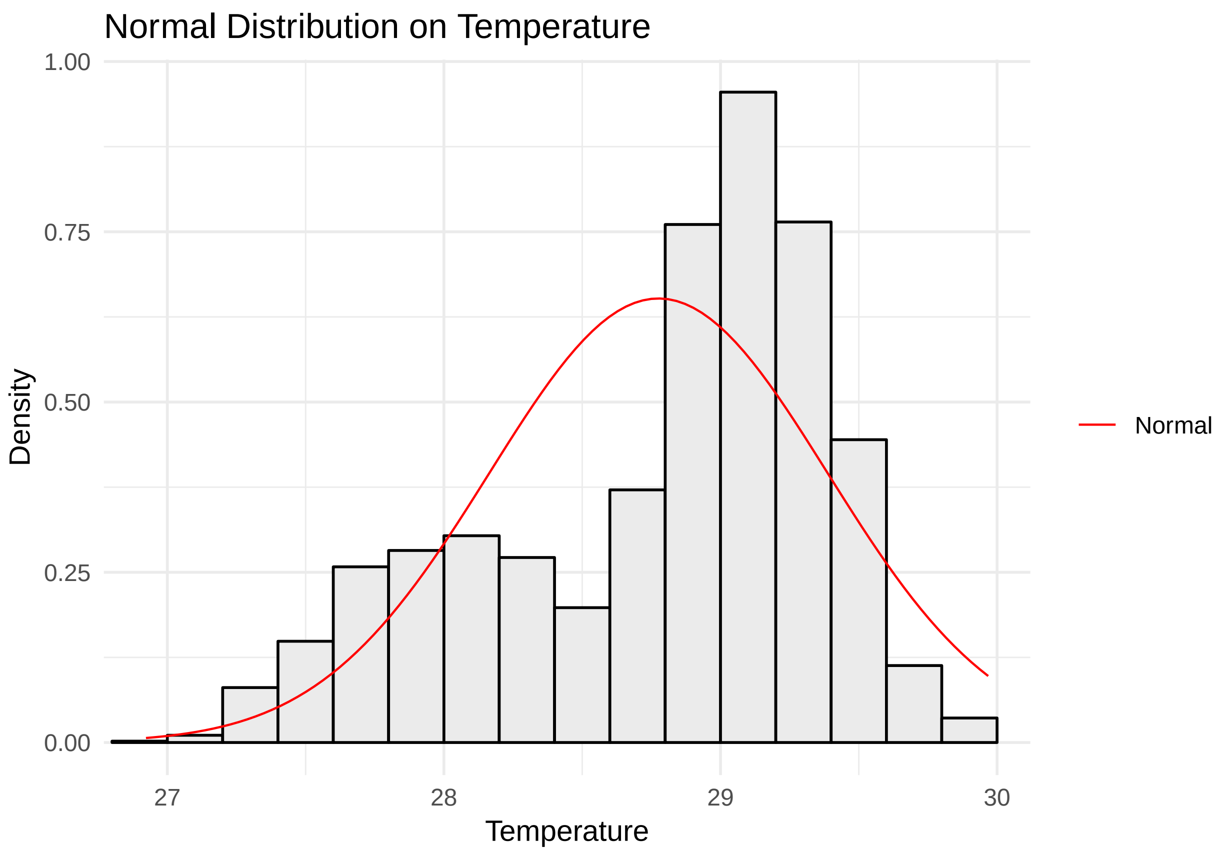
```
hist(WTempdata)
```



The histogram has two modes and the distribution is therefore bimodal. The distribution has a longer tail on the left and is not symmetrical, but skewed left. At last, anaything unusual or outliers cant be detected.

7.3)

```
# estimate the paramters for normal distribution
norm <- fitdist(WTempdata,distr = "norm")

dcomp <- denscomp(norm, legendtext = "Normal",
    xlab = "Temperature",
    fitlty = 1,
    xlegend = "topright",
    plotstyle = "ggplot",
    addlegend = FALSE)
dcomp + ggplot2::theme_minimal() + ggplot2::ggtitle("Normal Distribution on Temperature")
```



```
print(norm$estimate)
```

```
##       mean          sd
## 28.7757456  0.6118153
```

```
norm$loglik
```

```
## [1] -13914.2
```

7.4)

```
mixmdl = normalmixEM(WTempdata)  # default k=2 components
```

```
## number of iterations= 68
```

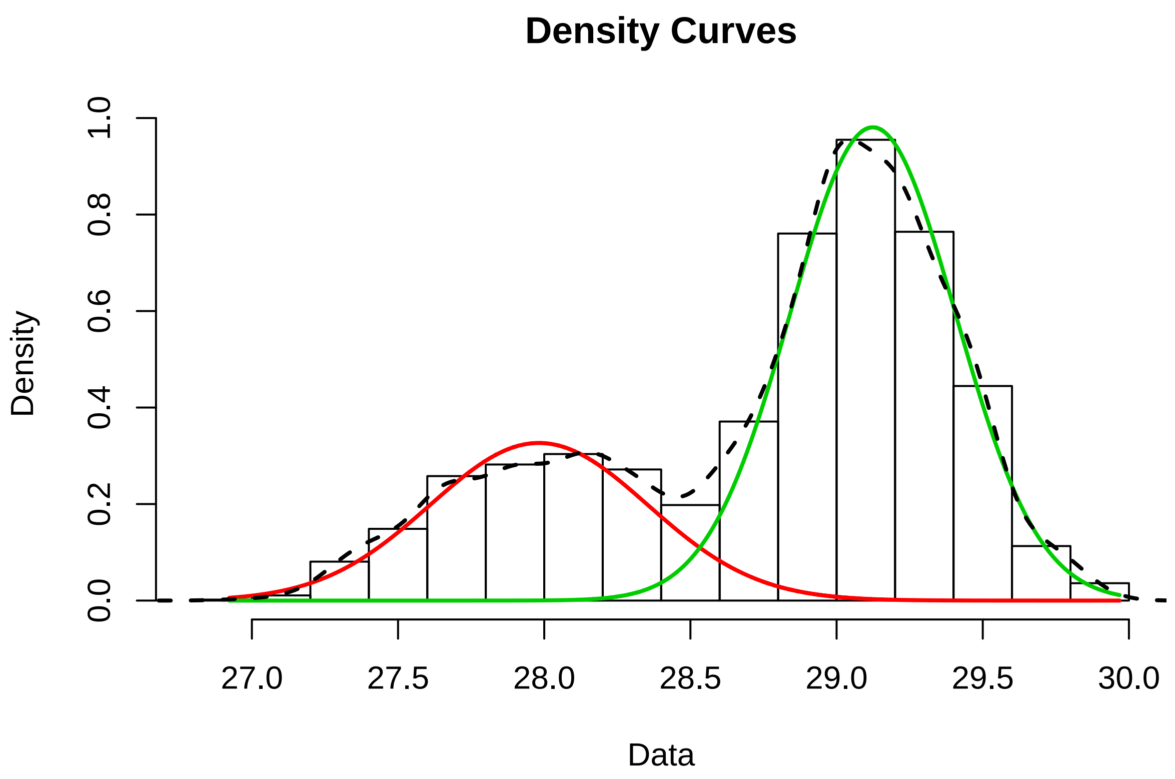
```
summary(mixmdl)
```

```
## summary of normalmixEM object:
##           comp 1    comp 2
## lambda  0.304882  0.695118
## mu     27.981830 29.123961
## sigma   0.372443  0.282750
## loglik at estimate:  -11441.47
```

```
# mixmdl$lambda
# mixmdl$mu
# mixmdl$sigma
#
```

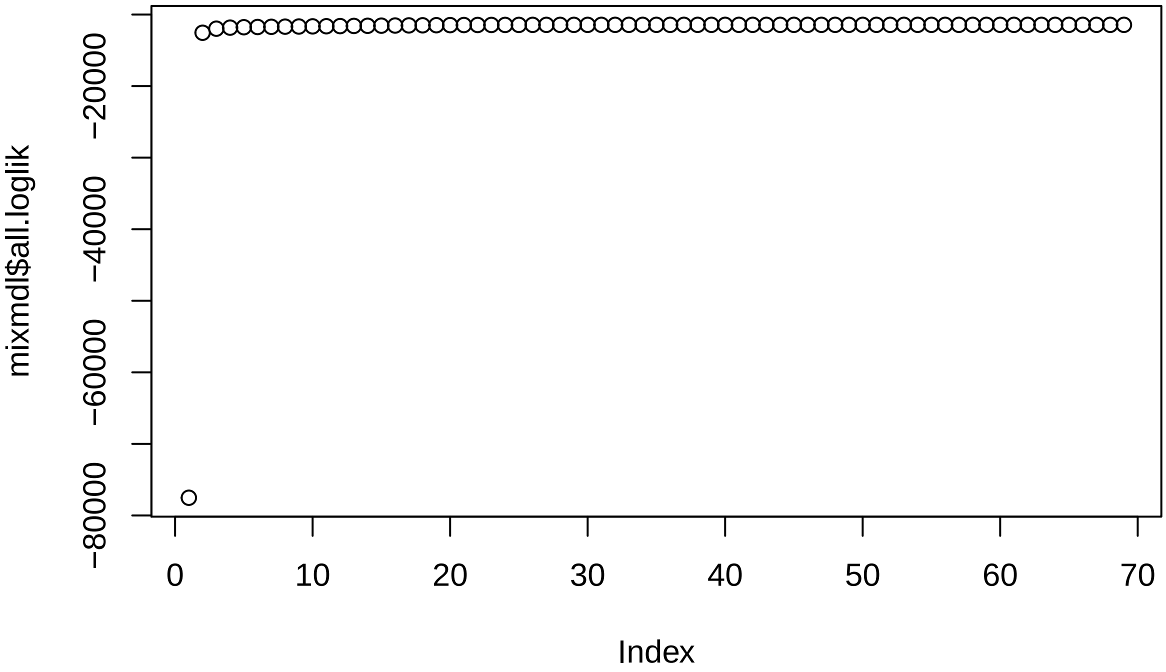
7.5)

```
plot(mixmdl,which=2)
lines(density(WTempdata), lty=2, lwd=2)
```



7.6)

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



After the first iteration, the log-likelihood increases drasticly and reaches shortly after convergence. This exact plot varies from execution to execution of the algorithm.

7.7)

The distribution obtained in 7.4 follows the data way better than the one obtained in 7.3. This can easily visually examined. Additionally it is visible, that the log-likelihood of 7.4 is higher and therefore it fits the data better.

```
# loglikelihood for 7.4
norm$loglik
```

```
## [1] -13914.2
```

```
# highest loglikelihood for 7.5
mixmdl$loglik
```

```
## [1] -11441.47
```

7.8)

Three main problems exist:

- Presence of Singularities
- Identifaiability Problem
- Maximising the log likelihood

But the main problem is the Presence of Singularities. This can be mitigated by observing the μ and σ while the algorithm is at work. If they get too small, it is adequate to intervene and increase them again.