FIT2086 Assignment 1

Name: Haw Xiao Ying
Student ID: 29797918
Laboratory: Group 04 (Friday 10a.m. - 12p.m.)

Question 1

a) Recommendation systems

The recommendation systems can find out the preference of the Netflix user according to various factors such as the categories and genre of the show. When the hidden patterns in the behaviour of Netflix users is discovered, Netflix can find shows and movies of interest to the users according to their preference from the viewed history.

b) Image recognition systems

Image recognition systems can identify objects, people and actions in the CCTV footage. Therefore, it can recognise which ones are considered as a human and then estimate the number of people at the platform of a train station.

c) Risk prediction

The risk prediction can predict the number of new cases of COVID-19 over the coming month by using the risk factor data of the people in the country collected. For example, if one had been in contact with a diagnosed patient of COVID-19 for the past 14 days and have symptoms such as coughing, then it will be predicted as a new case for the upcoming months.

d) Forecasting

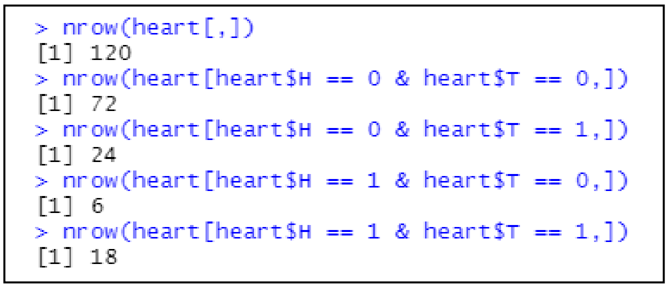
The forecasting system can analyse historical data to predict in determining the direction of the future. In this case, the forecasting system can use the historical data such as the recent usage habits or actions of the users of a mobile app, predicting whether or not the users will continue using the app in the future.

Question 2

a) Code for finding the total number of patients and the proportions of having or not having a heart attack given that an individual had normal or abnormal ECG result six months prior:

```
heart <- read.csv("heart.disease.csv", header = TRUE)
nrow(heart[,])
nrow(heart[heart$H == 0 & heart$T == 0,])
nrow(heart[heart$H == 0 & heart$T == 1,])
nrow(heart[heart$H == 1 & heart$T == 0,])
nrow(heart[heart$H == 1 & heart$T == 1,])
```

Results:



```
> nrow(heart[,])
[1] 120
> nrow(heart[heart$H == 0 & heart$T == 0,])
[1] 72
> nrow(heart[heart$H == 0 & heart$T == 1,])
[1] 24
> nrow(heart[heart$H == 1 & heart$T == 0,])
[1] 6
> nrow(heart[heart$H == 1 & heart$T == 1,])
[1] 18
```

Table of the proportions of having (H = 1)/not having (H = 0) a heart attack given that an individual had normal (T = 0)/abnormal (T = 1) ECG result six months prior:

	H=0	H=1
T=0	$\frac{72}{120} = 0.60$	$\frac{6}{120} = 0.05$
T=1	$\frac{24}{120} = 0.20$	$\frac{18}{120} = 0.15$

b) $P(H = 1) = 0.05 + 0.15 = 0.20$

c) $P(H = 1|T = 1) = \frac{0.05}{0.20+0.15} = 0.42857$

d) $P(H = 1|T = 0) = \frac{0.05}{0.60+0.05} = 0.076923$

e) No. The error rate of this ECG test is 25% $\left(\frac{False\ positive+false\ negative}{Positive+negative}\right)$ and the accuracy is only 75%. Since this is a predictor for medical use, which should be very accurate, therefore, in my opinion, this is not a good predictor.

Question 3

a)

$$P(X_1 = 1) = \frac{1}{6}$$
$$P(X_1 = 2) = \frac{1}{6}$$
$$P(X_1 = 3) = \frac{1}{6}$$
$$P(X_1 = 4) = \frac{1}{6}$$
$$P(X_1 = 5) = \frac{1}{6}$$
$$P(X_1 = 6) = \frac{1}{6}$$
$$P(Y_1 = 0) = \frac{1}{2}$$
$$P(Y_1 = 1) = \frac{1}{2}$$

$E[X_1] = 1 \times \left(\frac{1}{6}\right) + 2 \times \left(\frac{1}{6}\right) + 3 \times \left(\frac{1}{6}\right) + 4 \times \left(\frac{1}{6}\right) + 5 \times \left(\frac{1}{6}\right) + 6 \times \left(\frac{1}{6}\right) = \frac{7}{2}$

$E[Y_1] = 0 \times \left(\frac{1}{2}\right) + 1 \times \left(\frac{1}{2}\right) = \frac{1}{2}$

$E\left[{X_1}^2\right] = 1^2 \times \left(\frac{1}{6}\right) + 2^2 \times \left(\frac{1}{6}\right) + 3^2 \times \left(\frac{1}{6}\right) + 4^2 \times \left(\frac{1}{6}\right) + 5^2 \times \left(\frac{1}{6}\right) + 6^2 \times \left(\frac{1}{6}\right) = \frac{91}{6}$

$E\left[{Y_1}^2\right] = 0^2 \times \left(\frac{1}{2}\right) + 1^2 \times \left(\frac{1}{2}\right) = \frac{1}{2}$

$V[X_1] = E\left[{X_1}^2\right] - (E[X_1])^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$

$V[Y_1] = E\left[{Y_1}^2\right] - (E[Y_1])^2 = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4}$

$S = X_1 + 3Y_1$

$V[S] = V[X_1 + 3Y_1]$

$V[S] = V[X_1] + 3^2 V[Y_1]$

$V[S] = \frac{35}{12} + 9\left(\frac{1}{4}\right) = \frac{62}{12}$

$V[S] = \frac{31}{6}$

b) Since $P(Y_1 = 0)$ and $P(Y_1 = 1)$ are both $\frac{1}{2}$, so the probability is always $\frac{1}{2}$ for Y_1 and the probability obtaining either number of X_1 is always $\frac{1}{6}$. Hence, we can make a conclusion that every arrangement will have a probability of $\frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$. Therefore, I will state the arrangements for S = {1, 2, 3, 4, 5, 6, 7, 8, 9} and every arrangement will have a probability of $\frac{1}{12}$ which will need to be sum up.

P(S=1)

- ✓ Die: 1, Coin: 0

$$P(S=1)=\frac{1}{12}$$

P(S=2)

- ✓ Die: 2, Coin: 0

$$P(S=2)=\frac{1}{12}$$

P(S=3)

- ✓ Die: 3, Coin: 0

$$P(S=3)=\frac{1}{12}$$

P(S=4)

- ✓ Die: 1, Coin: 3
- ✓ Die: 4, Coin: 0

$$P(S=4)=\frac{1}{12}+\frac{1}{12}=\frac{1}{6}$$

P(S=5)

- ✓ Die: 2, Coin: 3
- ✓ Die: 5, Coin: 0

$$P(S=5)=\frac{1}{12}+\frac{1}{12}=\frac{1}{6}$$

P(S=6)

- ✓ Die: 3, Coin: 3
- ✓ Die: 6, Coin: 0

$$P(S=6)=\frac{1}{12}+\frac{1}{12}=\frac{1}{6}$$

P(S=7)

- ✓ Die: 4, Coin: 3

$$P(S=7)=\frac{1}{12}$$

P(S=8)

- ✓ Die: 5, Coin: 3

$$P(S=8)=\frac{1}{12}$$

P(S=9)

- ✓ Die: 6, Coin: 3

$$P(S=9)=\frac{1}{12}$$

s	1	2	3	4	5	6	7	8	9
P(S=s)	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

c) $E\left[\sqrt{S}\right] = \sqrt{1} \times \frac{1}{12} + \sqrt{2} \times \frac{1}{12} + \sqrt{3} \times \frac{1}{12} + \sqrt{4} \times \frac{1}{6} + \sqrt{5} \times \frac{1}{6} + \sqrt{6} \times \frac{1}{6} + \sqrt{7} \times \frac{1}{12} + \sqrt{8} \times \frac{1}{12} + \sqrt{9} \times \frac{1}{12} = \ 2.165963187$

$E\left[\sqrt{S}\right] \approx 2.1660$

d) $\mu_S = \ E[S] = 5$

$\sigma_S = V[S] = \frac{31}{6}$

$E[f(x)] \approx f(\mu_x) + \left[\frac{d^2 f(x)}{dx}\Big|_{x=\mu_x}\right] \times \frac{\sigma_x{}^2}{2}$

$E\left[\sqrt{S}\right] \approx \sqrt{\mu_S} + \left[\frac{d^2\sqrt{S}}{dS}\Big|_{S=\mu_S}\right] \times \frac{\sigma_S{}^2}{2}$

$E\left[\sqrt{S}\right] \approx \sqrt{5} + \left(-\frac{1}{4(5)^{\frac{3}{2}}} \times \frac{\frac{31}{6}}{2}\right)$

$E\left[\sqrt{S}\right] \approx 2.236067977 + (-0.05776508942)$

$E\left[\sqrt{S}\right] \approx 2.236067977 - 0.05776508942$

$E\left[\sqrt{S}\right] \approx 2.178302888$

$E\left[\sqrt{S}\right] \approx 2.1783$

e) Note that since X_1 and X_2 do the same action, they have the same expected value and variance.

There are two ways to find $V[X_1 + 3Y_1 + X_2]$.

Way 1:

$V[X_1 + 3Y_1 + X_2] = V[X_1] + 3^2 V[Y_1] + V[X_2]$

$V[X_1 + 3Y_1 + X_2] = \frac{35}{12} + 9\left(\frac{1}{4}\right) + \frac{35}{12} = \frac{97}{12}$

We have now obtained $V[X_1 + 3Y_1 + X_2]$ from way 1. To find $E[(X_1 + 3Y_1 + X_2)^2]$, we will now use way 2 and substitute the value of $V[X_1 + 3Y_1 + X_2]$ we just found into it.

$E[X_1 + 3Y_1 + X_2] = E[X_1] + 3E[Y_1] + E[X_2] = \frac{7}{2} + 3\left(\frac{1}{2}\right) + \frac{7}{2} = \ \frac{17}{2}$

Way 2:

$V[X_1 + 3Y_1 + X_2] = E[(X_1 + 3Y_1 + X_2)^2] - (E[X_1 + 3Y_1 + X_2])^2$

$\frac{97}{12} = E[(X_1 + 3Y_1 + X_2)^2] - \left(\frac{17}{2}\right)^2$

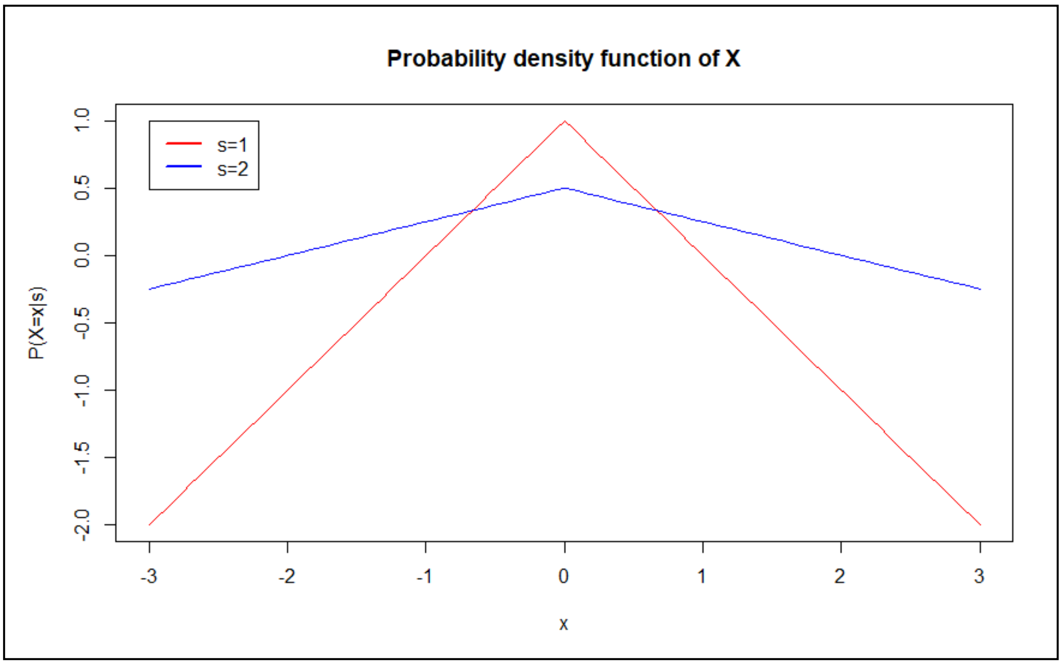
$E[(X_1 + 3Y_1 + X_2)^2] = \frac{97}{12} + \frac{289}{4} = \frac{241}{3}$

Question 4

a) Code to plot the probability density function of X:

```
f1 <- function(x){
   (1-abs(x))/(1^2)
}

f2 <- function(x){
   (2-abs(x))/(2^2)
}
curve(f1, from=-3, to=3, xlab="x", ylab="P(X=x|s)", main="Probability
density function of X", col="red")

curve(f2, from=-3, to=3, col="blue", add=TRUE)

legend(x=-3, y=1, legend=c("s=1","s=2"), col=c("red","blue"), lwd=2,
lty=c(1, 1))
```

Result (Graph):



b) $E[X] = \int_{-\infty}^{\infty} xf(x)dx$

$E[X] = \int_{-s}^{s} x\left(\frac{s-|x|}{s^2}\right)dx$

$E[X] = \frac{1}{s}\left(\frac{s^2}{2} - \frac{(-s)^2}{2}\right) - \frac{1}{s^2}\left(\frac{s^2|s|}{3} - \frac{(-s)^2|-s|}{3}\right)$

$E[X] = \frac{1}{s}(0) - \frac{1}{s^2}(0)$

$E[X] = 0$

c) $V[X] = \int_{-s}^{s} x^2 \left(\frac{s-|x|}{s^2}\right) dx - \mu^2$

$V[X] = \int_{-s}^{s} \frac{sx^2}{s^2}\, dx - \int_{-s}^{s} \frac{x^2|x|}{s^2}\, dx - \mu^2$

$V[X] = \frac{1}{s}\left(\frac{s^3}{3} - \frac{(-s)^3}{3}\right) - \frac{1}{s^2}\left(\frac{s^3|s|}{4} - \frac{(-s)^3|-s|}{4}\right) - \mu^2$

$V[X] = \frac{2s^2}{3} - \frac{1}{s^2}\left(\frac{2s^4}{4}\right) - 0$

$V[X] = \frac{2s^2}{3} - \frac{s^2}{2} = \frac{4s^2-3s^2}{6}$

$V[X] = \frac{s^2}{6}$

d) Cumulative distribution function:

$P(X \leq x) = \int_{-\infty}^{x} f(x)dx$

$P(X \leq x) = \int_{-s}^{x} \left(\frac{s-|x|}{s^2}\right) dx$

$P(X \leq x) = -\frac{x(|x|-2s)}{2s^2} - \left(-\frac{-s(|-s|-2s)}{2s^2}\right)$

$P(X \leq x) = \frac{-x|x|+2sx-s|s|+2s^2}{2s^2}$

e) $E[|X|] = \int_{-s}^{s} |x| f(|x|)dx$

$E[|X|] = \int_{-s}^{s} |x| \left(\frac{s-|x|}{s^2}\right) dx$

$E[|X|] = -\frac{(s)[2(s)^2-3s|(s)|]}{6s^2} - \left(-\frac{(-s)[2(-s)^2-3s|(-s)|]}{6s^2}\right)$

$E[|X|] = \frac{s}{6} + \frac{s}{6}$

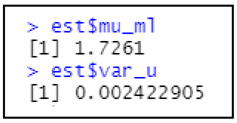
$E[|X|] = \frac{s}{3}$

Question 5

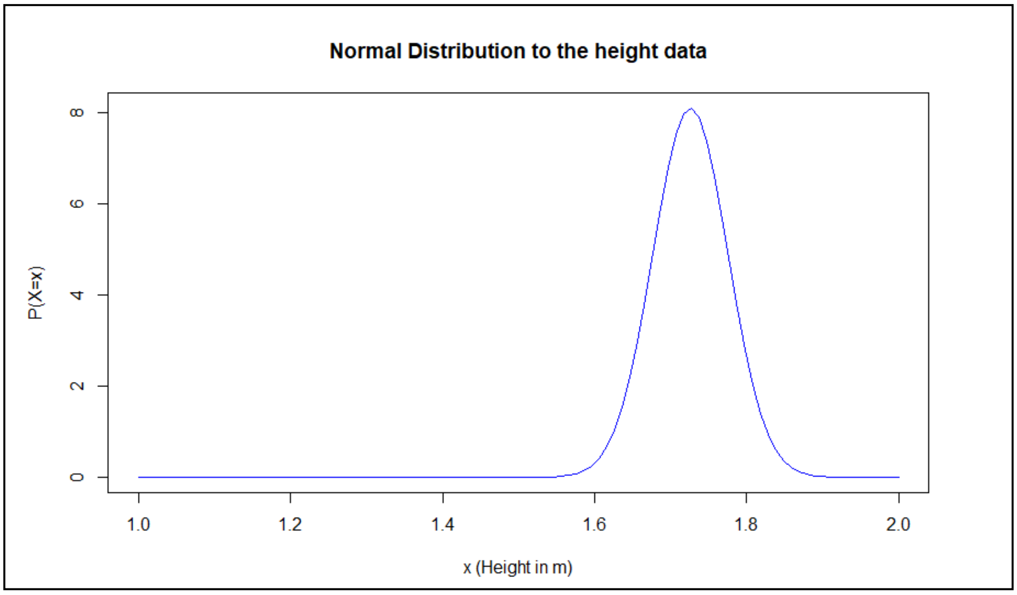
a) Code to fit a normal distribution to the height data using the maximum likelihood estimator and the unbiased estimator of variance:

```
height <- read.csv("heights.hk.csv", header = TRUE)

my_estimates <- function(X) {
  n = length(X)
  retval = list()

  # Calculate the sample mean
  retval$mu_ml = sum(X)/n

  # Calculate the squared deviations around the mean
  e2 = (X - retval$m)^2

  # Calculate the two estimates of variance
  retval$var_ml = sum(e2)/n
  retval$var_u  = sum(e2)/(n-1)

  return(retval)
}

est <- my_estimates(height$Height)

est$mu_ml
est$var_u

xseq <- seq(from=1, to=2, length.out=100)

plot(x=xseq, y=dnorm(x=xseq, mean=est$mu_ml, sd=sqrt(est$var_u)),
type = "l", xlab="x (Height in m)", ylab="P(X=x)", main="Normal
Distribution to the height data", col="blue")
```

Results:



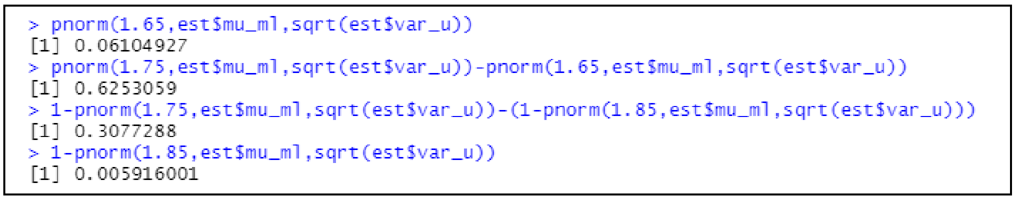
```
> est$mu_ml
[1] 1.7261
> est$var_u
[1] 0.002422905
```



Maximum likelihood estimator for μ = 1.7261
Unbiased estimator of variance for σ^2 = 0.002422905

$X \sim N(\mu, \sigma^2)$ where μ is the maximum likelihood estimator = 1.7261 and σ^2 is the unbiased estimator of variance = 0.002422905.

b) (i) Codes and results:



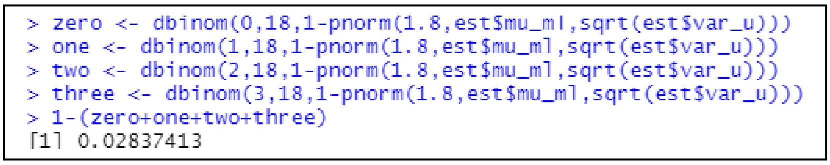
```
> pnorm(1.65,est$mu_ml,sqrt(est$var_u))
[1] 0.06104927
> pnorm(1.75,est$mu_ml,sqrt(est$var_u))-pnorm(1.65,est$mu_ml,sqrt(est$var_u))
[1] 0.6253059
> 1-pnorm(1.75,est$mu_ml,sqrt(est$var_u))-(1-pnorm(1.85,est$mu_ml,sqrt(est$var_u)))
[1] 0.3077288
> 1-pnorm(1.85,est$mu_ml,sqrt(est$var_u))
[1] 0.005916001
```

Proportions of candidate players that would fall into each of these height ranges:

- ✓ <1.65m: 0.06104927
- ✓ 1.65m – 1.75m: 0.6253059
- ✓ 1.75m – 1.85m: 0.3077288
- ✓ >1.85m: 0.005916001

(ii) If a new candidate player registers for the selection trial, the height range they are most likely to be in is 1.65m-1.75m since the proportions of candidate players of this height range is the highest.

(iii) Codes and result:



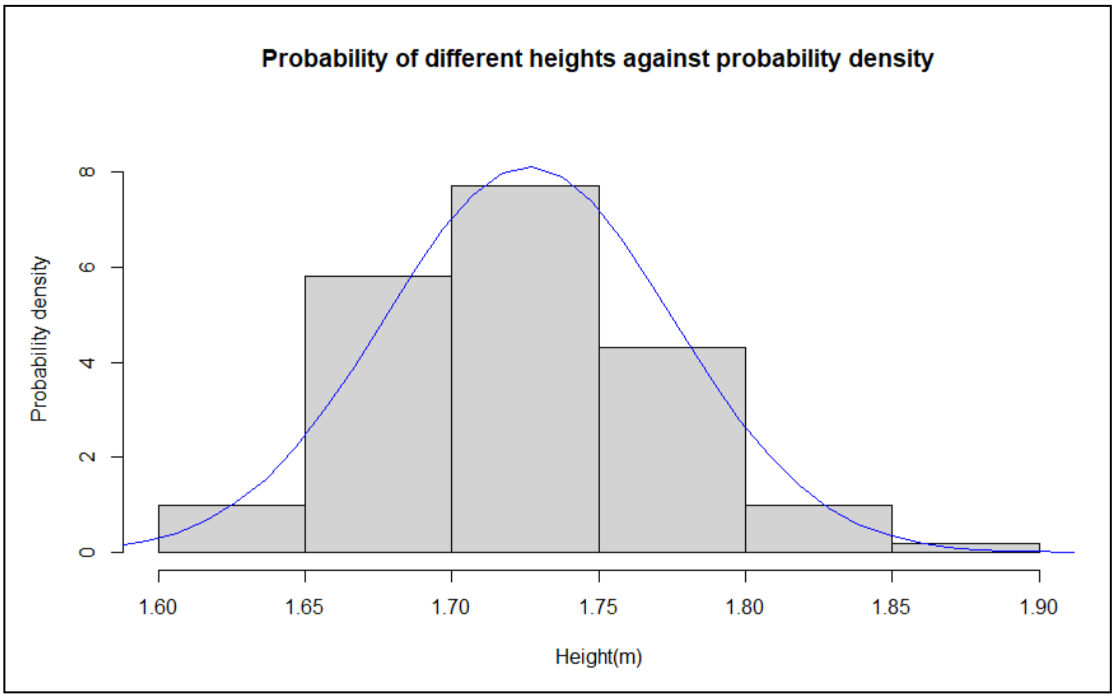
```
> zero <- dbinom(0,18,1-pnorm(1.8,est$mu_ml,sqrt(est$var_u)))
> one <- dbinom(1,18,1-pnorm(1.8,est$mu_ml,sqrt(est$var_u)))
> two <- dbinom(2,18,1-pnorm(1.8,est$mu_ml,sqrt(est$var_u)))
> three <- dbinom(3,18,1-pnorm(1.8,est$mu_ml,sqrt(est$var_u)))
> 1-(zero+one+two+three)
[1] 0.02837413
```

The probability that the coach will have at least 4 players taller than 1.80m available in their team is 0.02837413.

c) Codes to plot the observed probabilities of the different heights (using a histogram) against the probability density predicted by your Gaussian (normal distribution) model:

```
xseq <- seq(from=1, to=2, length.out=100)

hist(height$Height, prob=TRUE, ylim=c(0,9), xlab="Height(m)",
ylab="Probability density", main="Probability of different heights
against probability density")

lines(x=xseq, y=dnorm(x=xseq, mean=est$mu_ml, sd=sqrt(est$var_u)),
type = "l", col="blue")
```

Results:



I think that the normal distribution an appropriate model for this height data. As we can observed through the resulted graph, it can be clearly seen that the histogram and the curve of the normal distribution data fits perfectly. When the histogram reached its peak, the same goes to the normal distribution model. The shape of both the histogram and the Gaussian model is the same and therefore I believe that the distribution appears to be a good fit to this data.