STAT MODELING USING R AND SAS FINAL PROJECT

Michael Carabantes

mcaraban@iu.edu

12/13/2022

Commands

1)

```
f=file.choose()
ais=read.table(f,header = T)
library(dplyr)
ais_mc=sample_n(ais,size = 15)
dim(ais_mc)
```

2)

```
install.packages('fastDummies')
library('fastDummies')
ais_mc <- dummy_cols(ais_mc, select_columns = 'Sex')
ais_mc <- dummy_cols(ais_mc, select_columns = 'Sport')
```

3)

```
summary(ais_mc)
colSums(ais_mc[,14:24])
```

4)

```
library(tidyverse)
ggplot(ais_mc,aes(x=Sex,y=BMI))+geom_boxplot()
ggplot(ais_mc,aes(x=Sex,y=LBM))+geom_boxplot()
ggplot(ais_mc,aes(x=BMI,y=LBM,colour=Sex))+
  geom_point()
ggplot(ais_mc,aes(x=Ht,y=Wt,colour=Sex))+
  geom_point()
ggplot(ais_mc,aes(x=BMI))+geom_histogram(bins=10)
ggplot(ais_mc,aes(x=LBM))+geom_histogram(bins=10)
ggplot(ais_mc,aes(x=Ht))+geom_histogram(bins=10)
ggplot(ais_mc,aes(x=Wt))+geom_histogram(bins=10)
```

5)

```
MBMI=ais_mc[which(ais_mc$Sex=='male'),'BMI']
length(MBMI)
FBMI=ais_mc[which(ais_mc$Sex=='female'),'BMI']
length(FBMI)
mean(MBMI)
mean(FBMI)
mean(MBMI)-mean(FBMI)
t.test(MBMI,FBMI,conf.level = .95)
```

6)

```
Model=lm(LBM~BMI,data = ais_mc)
summary(Model)
ggplot(ais_mc,aes(x=BMI,y=LBM))+
  geom_point()+geom_smooth(method="lm")
```

7)

```
permutations <- function(n)
{if(n==1){
  return(matrix(1))
} else {
  sp <- permutations(n-1)
  p <- nrow(sp)
  A <- matrix(nrow=n*p,ncol=n)
  for(i in 1:n)
  { A[(i-1)*p+1:p,] <- cbind(i,sp+(sp>=i))
  }
  return(A)
}
}

MaleBMI=sample(MBMI,4)
FemaleBMI=sample(FBMI,4)
z=c(MaleBMI,FemaleBMI)
thetahat=mean(MaleBMI)-mean(FemaleBMI)
n=length(MaleBMI);m=length(FemaleBMI);N=m+n
perm=permutations(N)
thetahatstars=data.frame(t(perm))%>%map(~.x%in%(1:n))%>%bind_cols()%>%
  map_dbl(~mean(z[.x])-mean(z[!.x]))
#pvalue
mean(abs(thetahatstars)>=abs(thetahat))
```

8)

```
Actual.meandiff=abs(mean(MBMI)-mean(FBMI))
n=length(ais_mc$Sex)
B=1000
ais_mc = ais_mc %>% arrange(desc(Sex))
m=ais_mc$BMI

#Bootstrap Samples Matrix
BootStrap=matrix(sample(m,size = n*B,replace = T),nrow = n,ncol = B)
Boot.mean=rep(0,1000)
for (i in 1:B) {
  Boot.mean[i]=abs(mean(BootStrap[1:14,i])-mean(BootStrap[15:25,i]))

}

#p-value
mean(Boot.mean >= Actual.meandiff)
```

Project Summary

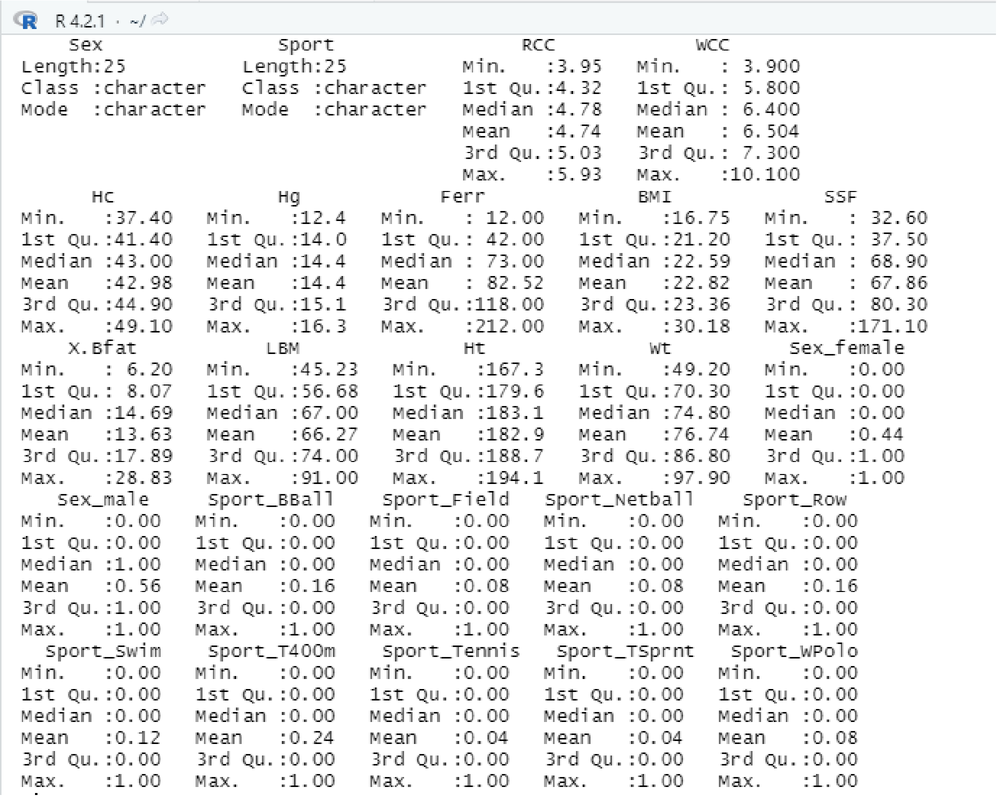
1) Read in data

[1] 25 24

After taking a random sample from the raw data, I have 25 observations with 24 variables

2) Descriptive Statistics

a)



```
     Sex                Sport                RCC             WCC        
 Length:25          Length:25          Min.   :3.95   Min.   : 3.900  
 Class :character   Class :character   1st Qu.:4.32   1st Qu.: 5.800  
 Mode  :character   Mode  :character   Median :4.78   Median : 6.400  
                                       Mean   :4.74   Mean   : 6.504  
                                       3rd Qu.:5.03   3rd Qu.: 7.300  
                                       Max.   :5.93   Max.   :10.100  
       HC              Hg             Ferr             BMI             SSF        
 Min.   :37.40   Min.   :12.4   Min.   : 12.00   Min.   :16.75   Min.   : 32.60  
 1st Qu.:41.40   1st Qu.:14.0   1st Qu.: 42.00   1st Qu.:21.20   1st Qu.: 37.50  
 Median :43.00   Median :14.4   Median : 73.00   Median :22.59   Median : 68.90  
 Mean   :42.98   Mean   :14.4   Mean   : 82.52   Mean   :22.82   Mean   : 67.86  
 3rd Qu.:44.90   3rd Qu.:15.1   3rd Qu.:118.00   3rd Qu.:23.36   3rd Qu.: 80.30  
 Max.   :49.10   Max.   :16.3   Max.   :212.00   Max.   :30.18   Max.   :171.10  
     X.Bfat           LBM              Ht              Wt          Sex_female  
 Min.   : 6.20   Min.   :45.23   Min.   :167.3   Min.   :49.20   Min.   :0.00  
 1st Qu.: 8.07   1st Qu.:56.68   1st Qu.:179.6   1st Qu.:70.30   1st Qu.:0.00  
 Median :14.69   Median :67.00   Median :183.1   Median :74.80   Median :0.00  
 Mean   :13.63   Mean   :66.27   Mean   :182.9   Mean   :76.74   Mean   :0.44  
 3rd Qu.:17.89   3rd Qu.:74.00   3rd Qu.:188.7   3rd Qu.:86.80   3rd Qu.:1.00  
 Max.   :28.83   Max.   :91.00   Max.   :194.1   Max.   :97.90   Max.   :1.00  
    Sex_male     Sport_BBall     Sport_Field    Sport_Netball     Sport_Row   
 Min.   :0.00   Min.   :0.00   Min.   :0.00   Min.   :0.00   Min.   :0.00  
 1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.00  
 Median :1.00   Median :0.00   Median :0.00   Median :0.00   Median :0.00  
 Mean   :0.56   Mean   :0.16   Mean   :0.08   Mean   :0.08   Mean   :0.16  
 3rd Qu.:1.00   3rd Qu.:0.00   3rd Qu.:0.00   3rd Qu.:0.00   3rd Qu.:0.00  
 Max.   :1.00   Max.   :1.00   Max.   :1.00   Max.   :1.00   Max.   :1.00  
   Sport_Swim     Sport_T400m    Sport_Tennis    Sport_TSprnt    Sport_WPolo  
 Min.   :0.00   Min.   :0.00   Min.   :0.00   Min.   :0.00   Min.   :0.00  
 1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.00  
 Median :0.00   Median :0.00   Median :0.00   Median :0.00   Median :0.00  
 Mean   :0.12   Mean   :0.24   Mean   :0.04   Mean   :0.04   Mean   :0.08  
 3rd Qu.:0.00   3rd Qu.:0.00   3rd Qu.:0.00   3rd Qu.:0.00   3rd Qu.:0.00  
 Max.   :1.00   Max.   :1.00   Max.   :1.00   Max.   :1.00   Max.   :1.00  
```

Some interesting things to note are that several variables have a higher median than mean. The variable Hg has the exact same median and mean implying there is no skewness. On the flip side, the Ferr variable has the biggest difference of mean and median, indicating skewness to the right The Sport dummy variables with the highest mean is T400m indicating that this sample has more track athletes than any other sport. Similarly, the mean for male dummy variable is higher than female indicating there are more males than females in this sample.

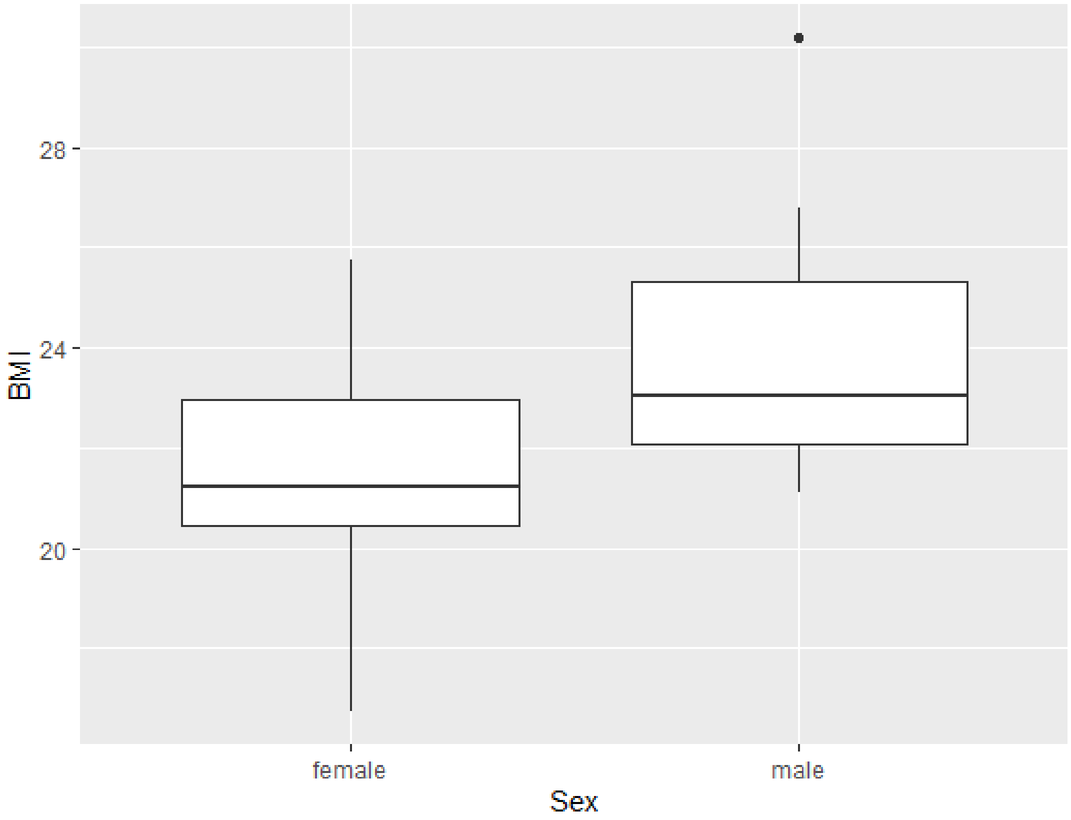
b)

```
 Sex_female     Sex_male   Sport_BBall   Sport_Field   Sport_Netball   Sport_Row
     11            14           4             2              2             4
 Sport_Swim   Sport_T400m  Sport_Tennis  Sport_TSprnt  Sport_WPolo
      3            6             1             1            2
```

Computing the frequencies, it is visible now that my earlier statement of there being more males than females and more track athletes than other atheist is correct.

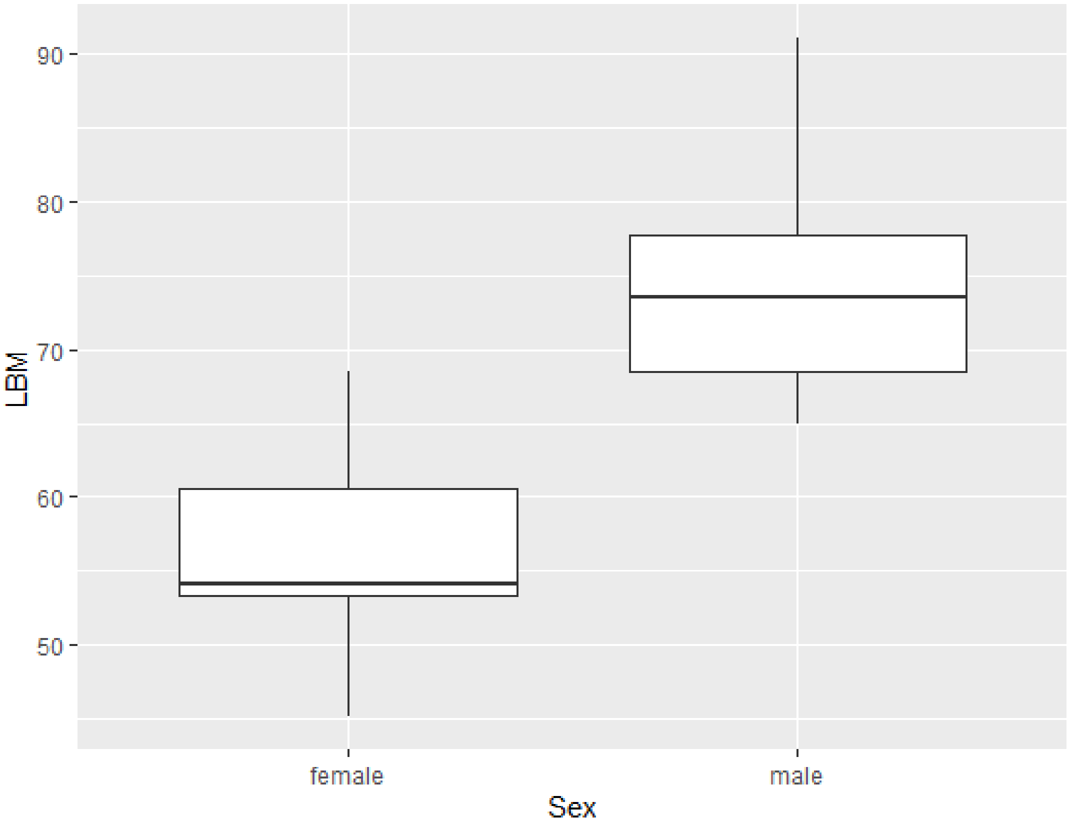
3) Graphical presentation

a)



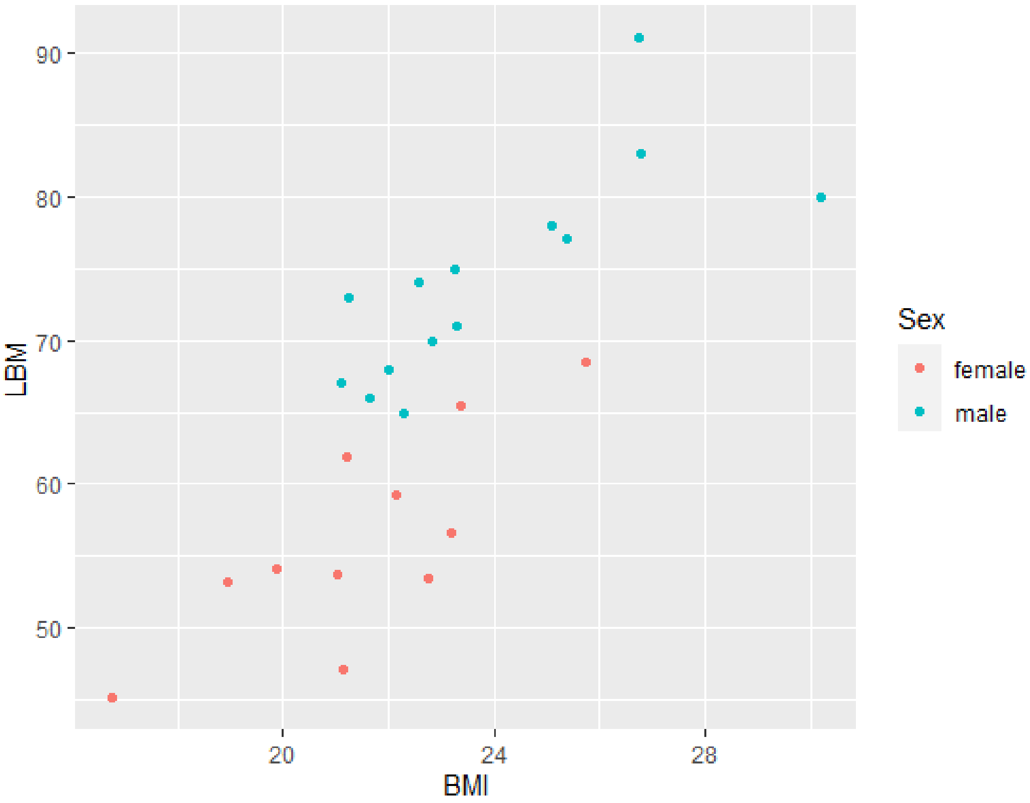
Comparing these two boxplots, one can see that the median for males is level if not past the top of boxplot for females which implies the difference between the two groups. The male group box is also bigger which means there is a bigger range. The overall minimum BMI belongs to a female and maximum belongs to a male. Both sets of data seem to be skewed to the left.

b)

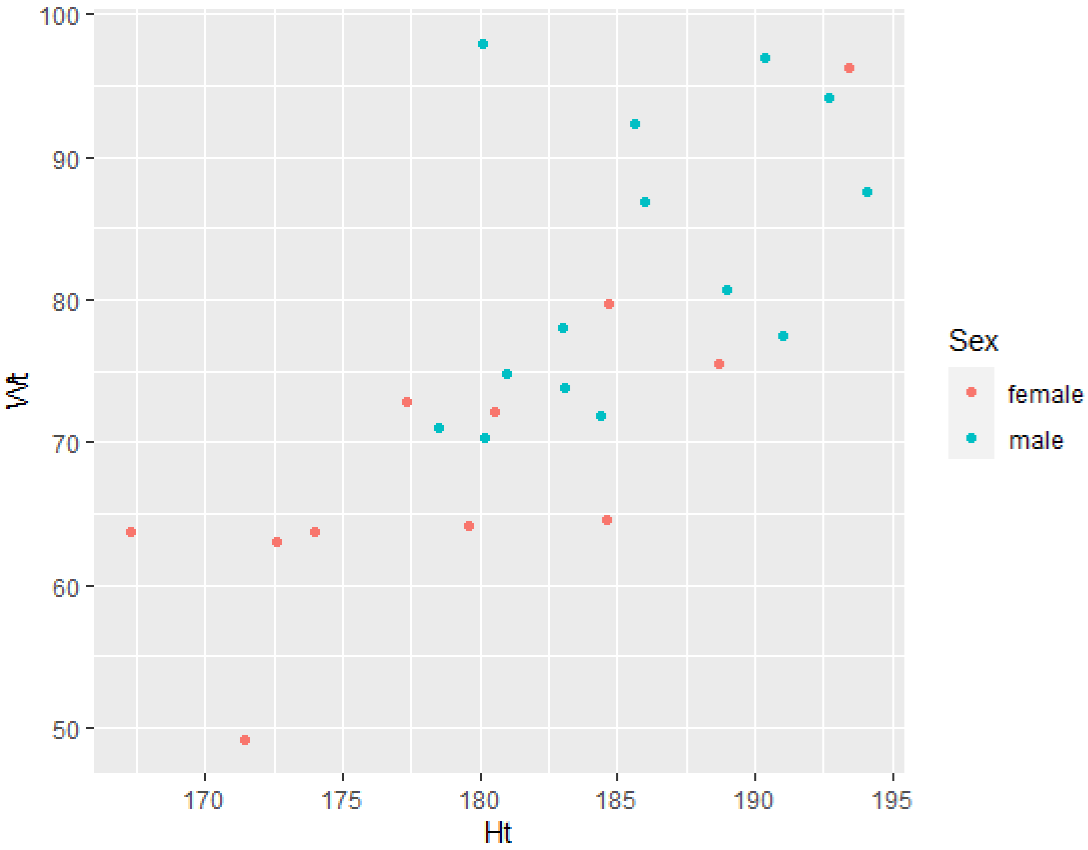


Comparing these two boxplots, one can see that the median for males is way past the top of boxplot for females which implies the large difference between the two groups. The size of the boxes are relatively the same so they have similar ranges. The overall minimum BMI belongs to a female and maximum belongs to a male once again. The male group seems to be slightly skewed to the right, while the female group is heavily skewed to the left.

c)

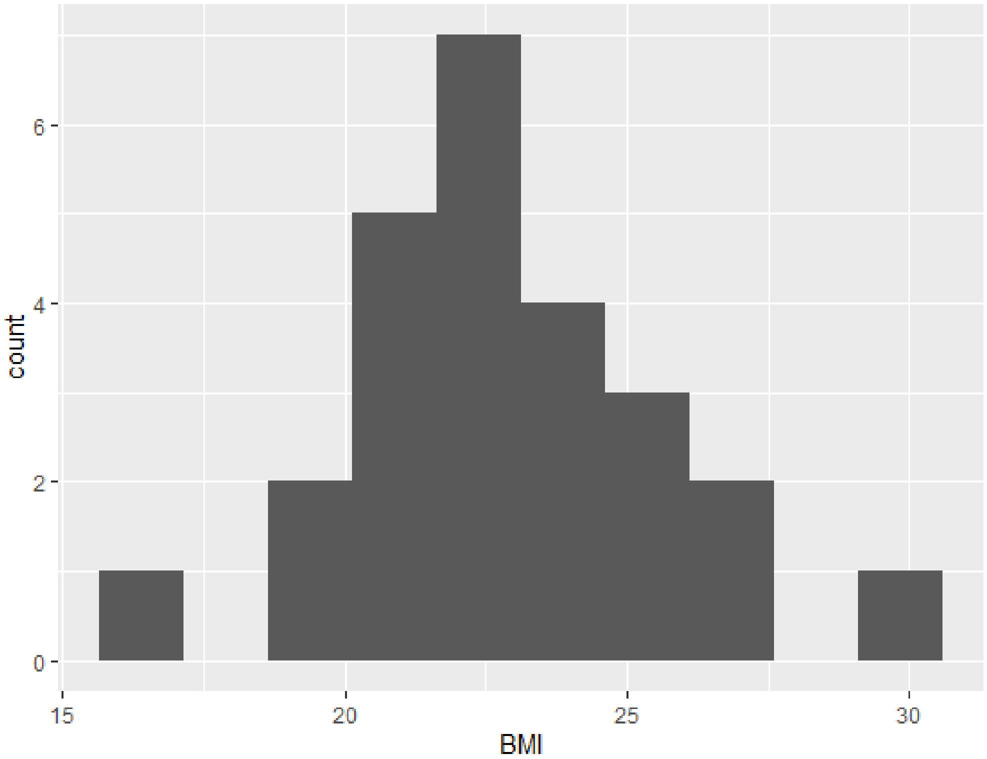


The scatterplot between LBM and BMI shows a positive linear relationship. As one can see, since most males have a higher value of LBM/BMI their data points end up being in the upper half of the scatterplot.

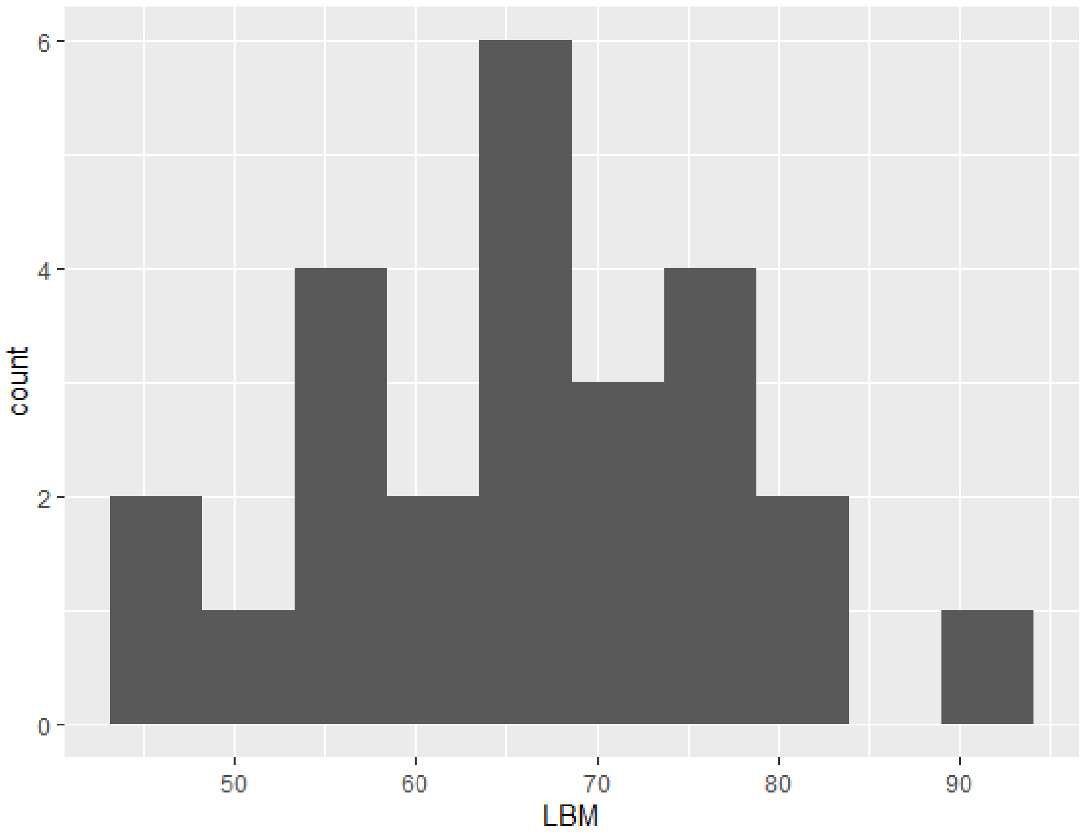


Similarly to the last scatter plot there is a positive linear relationship between Ht and Wt. However, this scatter plot isn't as segregated between males and females compared to the last one.

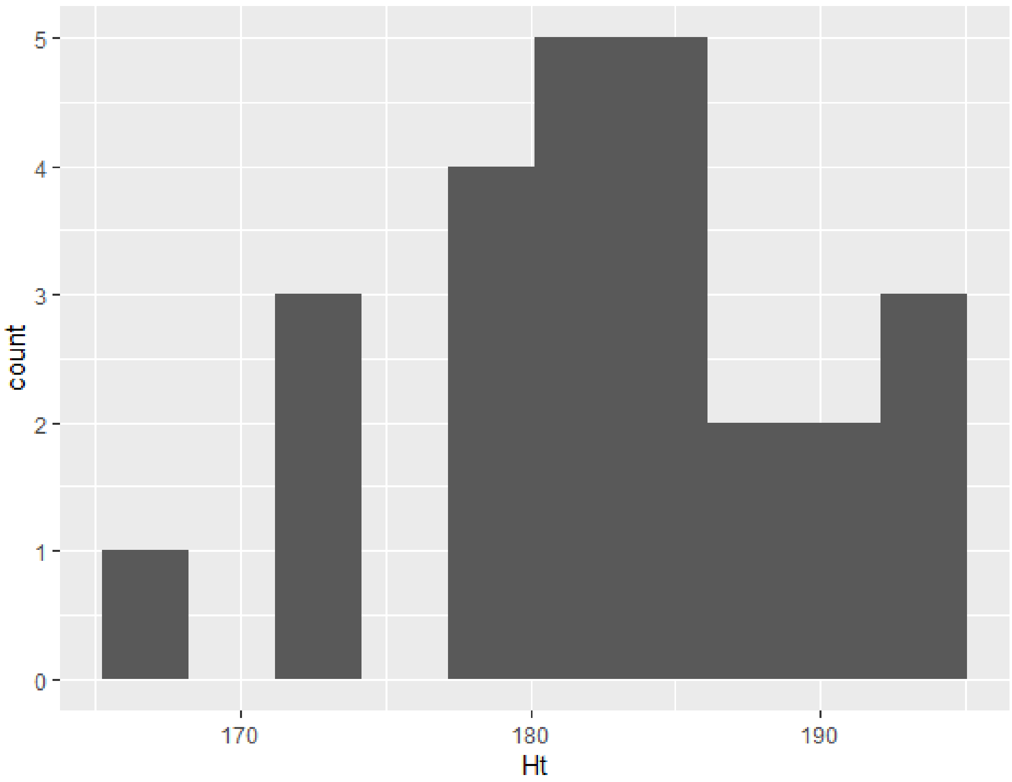
d)



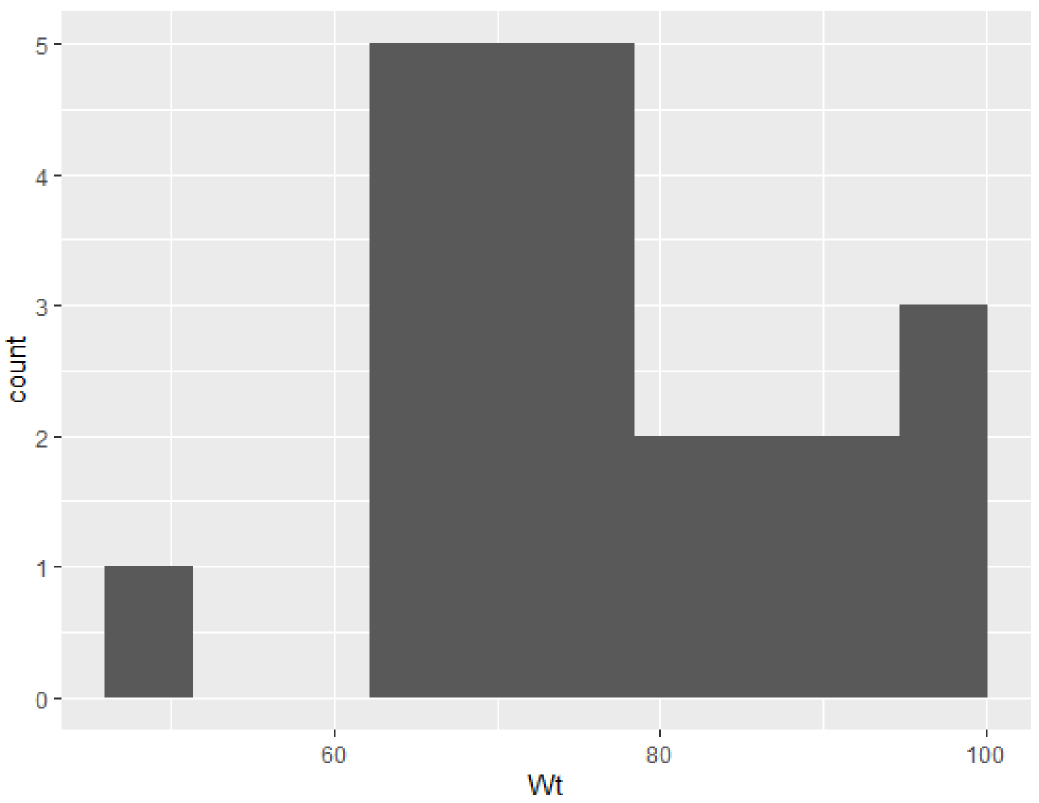
The histogram of the BMI is the most symmetrical of the 4 variables. They share a gap on both side and have a mode around 22.



The histogram of LBM is slightly right skewed. It contains a gap on the side implying there is a maximum outlier. Its mode lies somewhere around 65.



The height histogram seems to skewed to the left. There are two gaps to the left, one which provides a minimum. The histogram appears to be bimodal and lies somewhere between 180-185.



Similarly the histogram for weight is also bimodal with its mode lying between 63-78. Also, similarly to the height histogram, the weigh histogram contains a minimum outlier.

4) Statistical Test (Using .05 alpha level)

a)

H0=Means of BMI for males and females are equal

H1=Means of BMI for males and females are not equal

b)

The t-test is more appropriate here because my sampes contains 25 observations and variance of the population is similar.

c)

The sample size of males is,

```
[1] 14
```

and the sample size of females is,

```
[1] 11
```

while the mean of BMI for males is

```
[1] 23.88714
```

and the mean of BMI for females is

```
[1] 21.46909
```

Difference between means is

```
[1] 2.418052
```

d)

```
        Welch Two Sample t-test

data:  MBMI and FBMI
t = 2.3932, df = 22.333, p-value = 0.02553
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3244685 4.5116354
sample estimates:
mean of x mean of y
 23.88714  21.46909
```

Through the t-test, we get the p-value .02556. Using the significance level of .05, since the p-value is lower we reject the null hypothesis. This means that the means of male and female BMI are not equal. The test provides a 95 percent confidence interval as well.

6)

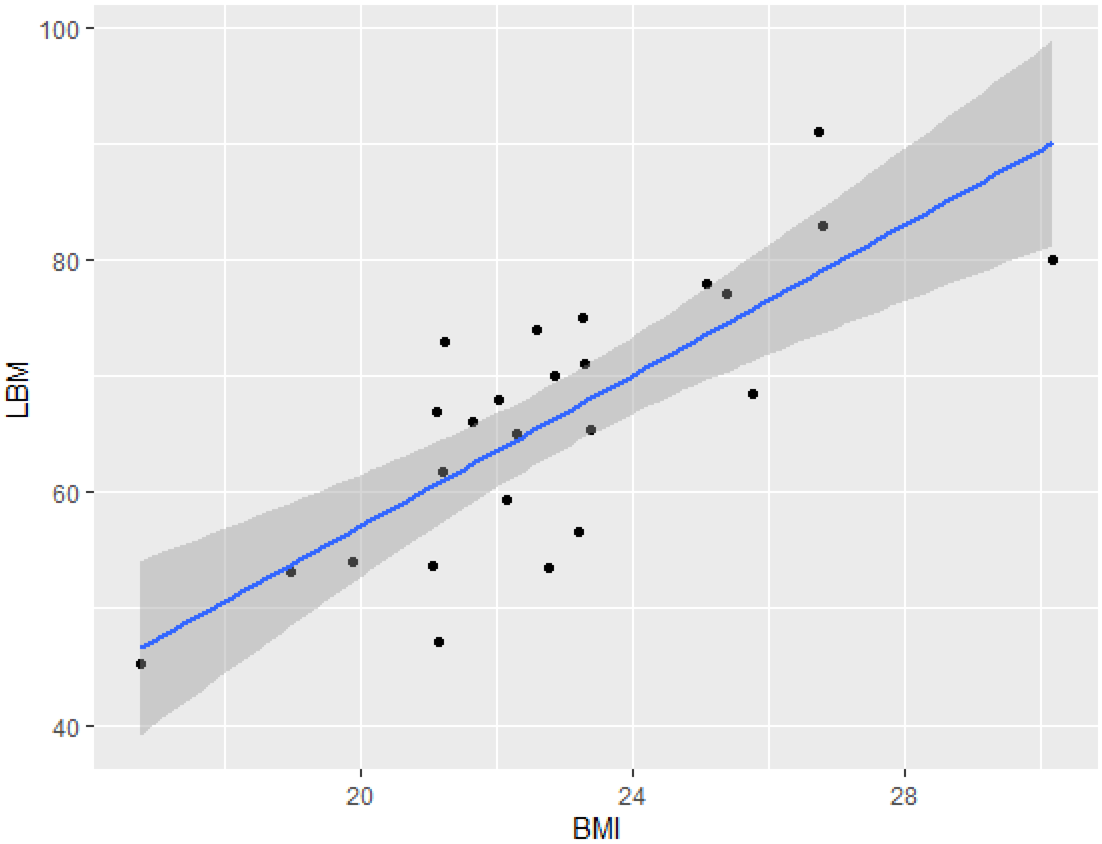
```
Call:
lm(formula = LBM ~ BMI, data = ais_mc)

Residuals:
     Min       1Q   Median       3Q      Max
-13.7687  -4.6988   0.8296   4.3594  12.0910

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.5578    12.4946  -0.605    0.551
BMI           3.2348     0.5436   5.950 4.57e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.365 on 23 degrees of freedom
Multiple R-squared:  0.6062,  Adjusted R-squared:  0.5891
F-statistic: 35.41 on 1 and 23 DF,  p-value: 4.57e-06
```

Y=-7.5578+3.2348*X



Like previously stated, there is a positive linear relationship. The slope has a small p-value meaning there is enough evidence to reject the null as previously stated as well.

7)

```
[1] 0.05714286
```

Choosing a significance level of .05 again, running the permutation test (with 4 samples from male and female) provides us with a p-value of .05714286. This is slightly higher than our significance level used so we would fail to reject the null.

8)

```
[1] 0.028
```

Again choosing a significance level of .05, running the bootstrap test with 1000 bootstrap samples provides us with a p-value of .028. This is lower than our significance level so we would reject the null.