Analysis of competitors at the London 2012 Olympic games

Summary

In this report we investigate characteristics of competitors who were involved in the London 2012 Summer Olympics games. The characteristics we will focus on are the weight, height and age of competitors. We aim to compare these characteristics between the populations of medallists/non-medallists and also medallists from different sports. The exploratory methods used are graphical analysis, hypothesis testing and summary statistics. Results indicate that medal winning swimmers are taller and older on average than non-medallists, and swimming medallists are taller and heavier than winning equestrian competitors.

Introduction

This report aims to analyse characteristics of competitors amongst different sports in the London 2012 Olympic Games. This event made London the first city to host the modern Olympics three times, and 26 sports were featured, including Women's boxing for the first time. The findings of this report are based on data from individual competitors'; gender, age(in years), height(in centimetres), weight(in kilograms), team, sport and the medal won (if any) are provided. The main focus of the investigation is on height, weight and age, paying particular attention to differences between medal winners and their counterparts. Analysis of these data is useful to investigate the physical requirements of various sports and thus the advantages of certain features to succeed within these disciplines.

Methods

The methods used to summarise the data include calculations of the mean, standard deviation, median and interquartile range for competitors' weight, height and age. To display these variables, histograms are used in order to observe distributions. Also, normal QQ plots were used to determine whether we can reasonably assume that each variable may be modelled by a normal distribution, supporting any observation made from the histograms. A matrix scatterplot for weight, height and age was produced, so that possible relationships could be seen and extreme values easily identified. Furthermore, adjacent boxplots for medal-winning/non medal-winning swimmers are found to aid comparison of the two populations and see if there might be significant differences between their mean values. Lastly, to investigate these differences, t-tests are applied to the swimming data and also a further investigation between medal-winning swimmers and equestrian competitors. All the calculations were performed in R, and the code to produce all the results is given in the Appendix.

Analysis of competitors' height, weight and age

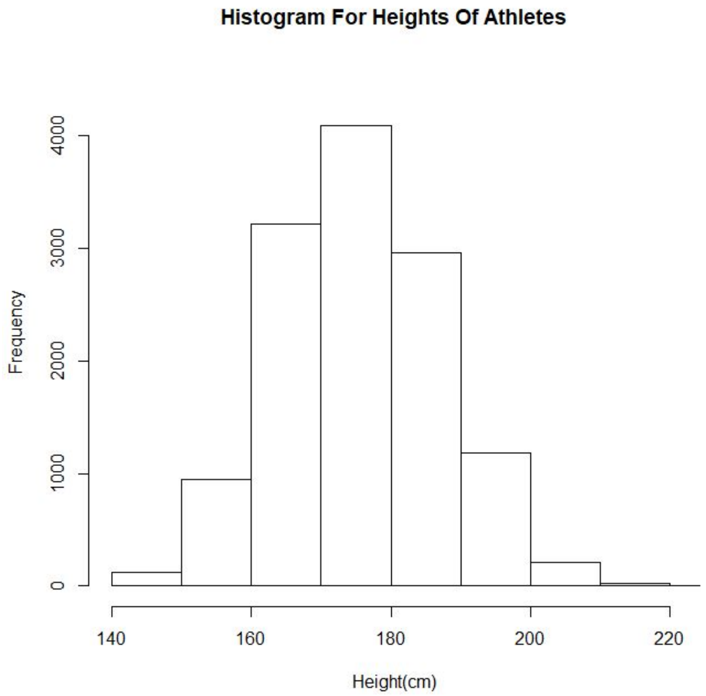
Variable	Mean	Standard deviation	Median	Interquartile Range
Height(cm)	176.3	11.45	176	16

Weight(kg)	71.32	15.86	69	20
Age(years)	25.96	5.68	25	7

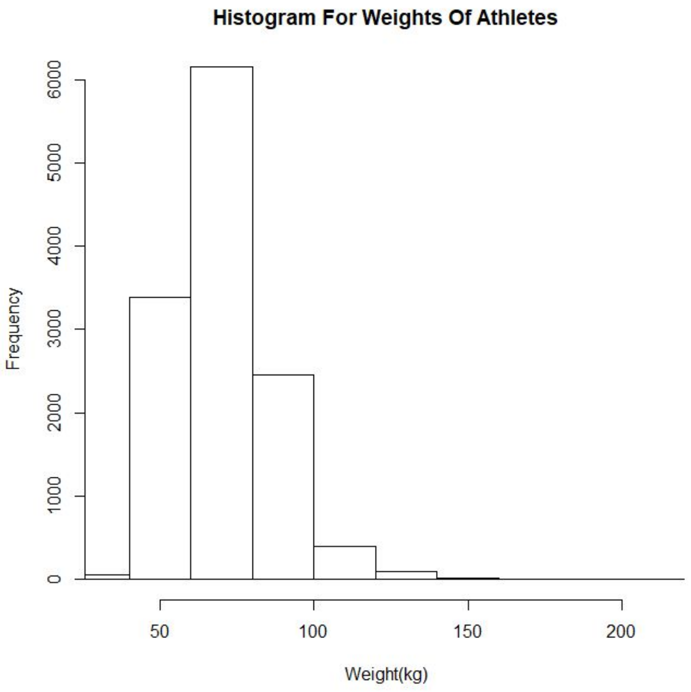
The weight of competitors is the most spread with largest standard deviation, there is a larger difference between mean and median values than the other variables which suggests that there is an extreme large value. Age is the least spread, the middle 50% of the data lying over a small difference of 7 years which seems to suggest that the majority of the competitors are between 20 and 30.

Histograms

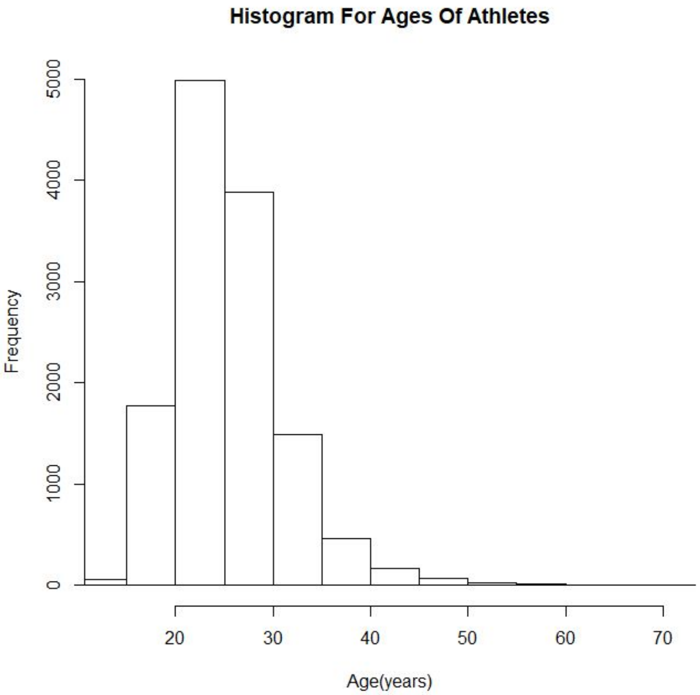
The histogram for height (right) shows that this data has a positive skew, with a modal class at 170-180cm, there are a few values beyond 200cm. Ignoring these, the distribution has some symmetry about the modal class, seeming to resemble a normal distribution.



The histogram for weight shows a highly positively skewed distribution. The modal class at approximately 65kg has a huge frequency at around 6000. The positive skew shows a large deviation, there are some extremes at high weight.

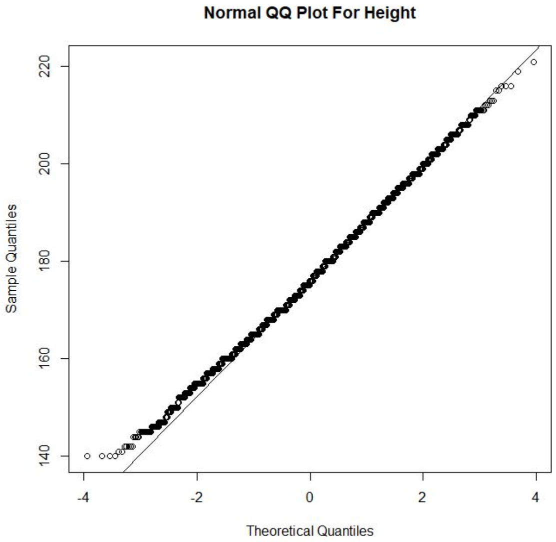


The histogram for age shows positive skew with most of the data at 20-30 years. It also shows there are competitors aged 50+ which is surprising.

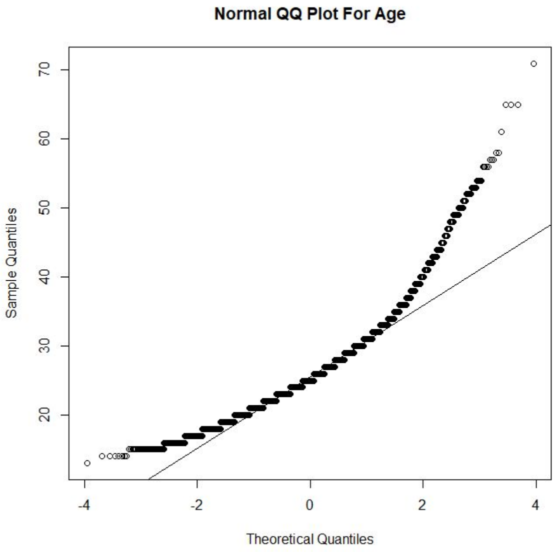


Normal QQ Plots

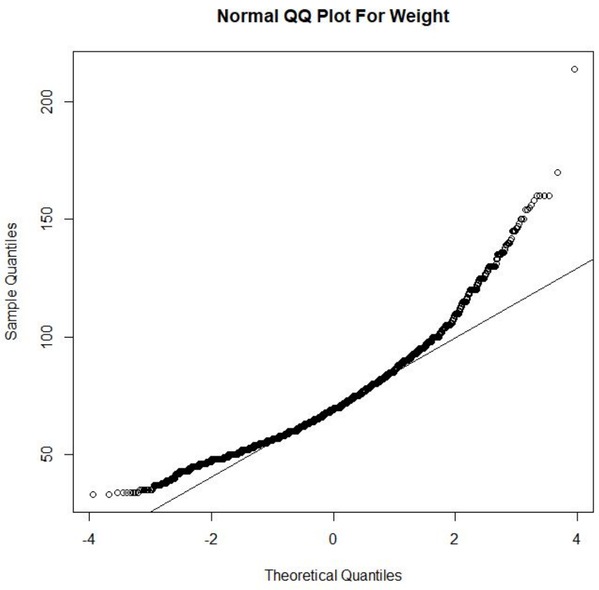
The normal QQ plot (right) for height supports the above claim that the distribution resembles a normal distribution, as the straight line plotted is aligned with the data.

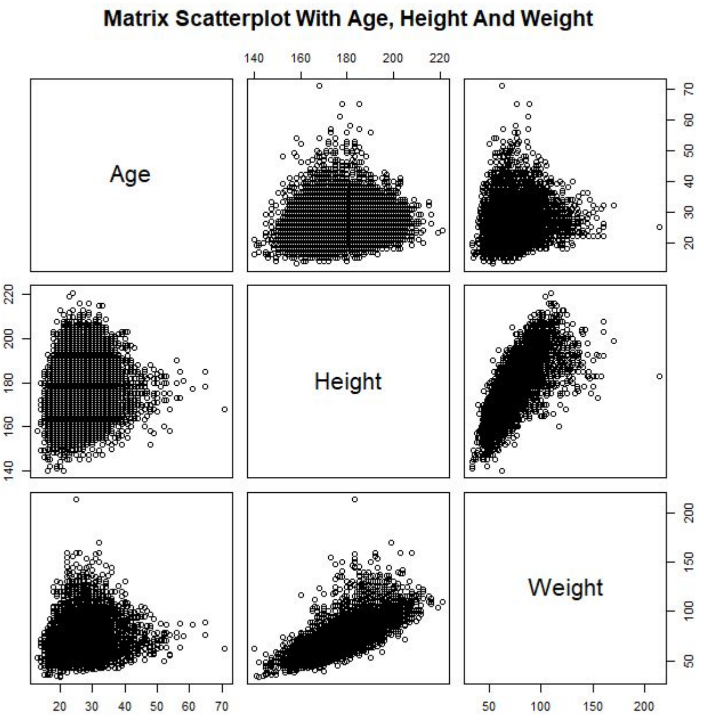


The QQ plot for age however shows this data does not resemble a normal distribution. It also shows the outlier at above 70 (top right of plot).



The plot for weight, similarly to age indicates a normal model is not reasonable. It also reveals the extreme (top right) beyond 200kg.





Scatterplot

The scatterplot shows that there is positive correlation between weight and height and no correlation between age and the other variables, since there is a whole range of results with age. The bottom middle plot shows that increasing weight is linked to increased height.

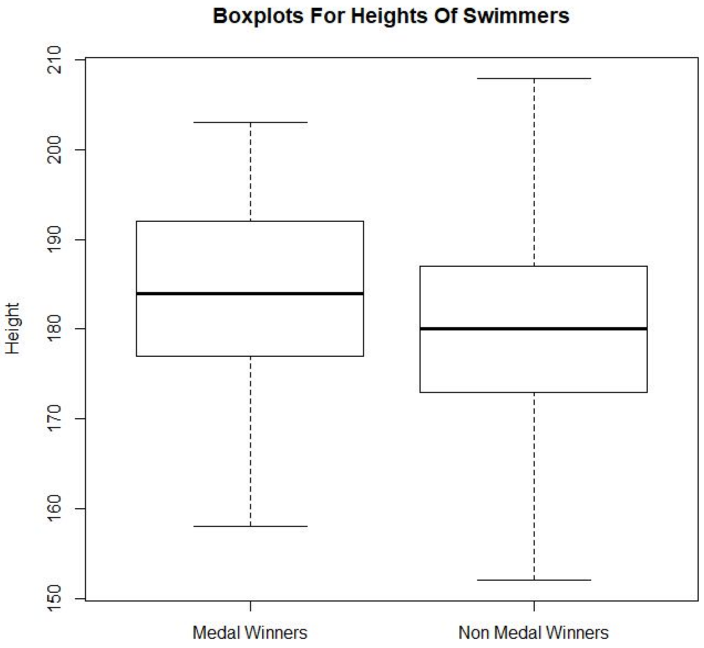
Outliers

Clearly there is a competitor with a larger weight than the rest of the data (high up point on bottom middle plot). This represents a competitor in Judo Men’s Heavyweight at 214kg. Note that the second heaviest competitor at 170kg competes in the same event thus it seems reasonable that large weight could aid a competitor in that event. Besides this, there is an equestrian competitor at the age of 71, and interestingly again the second oldest at 65 also competes in equestrianism, which seems to show experience may be an advantage in this event which involves horses.

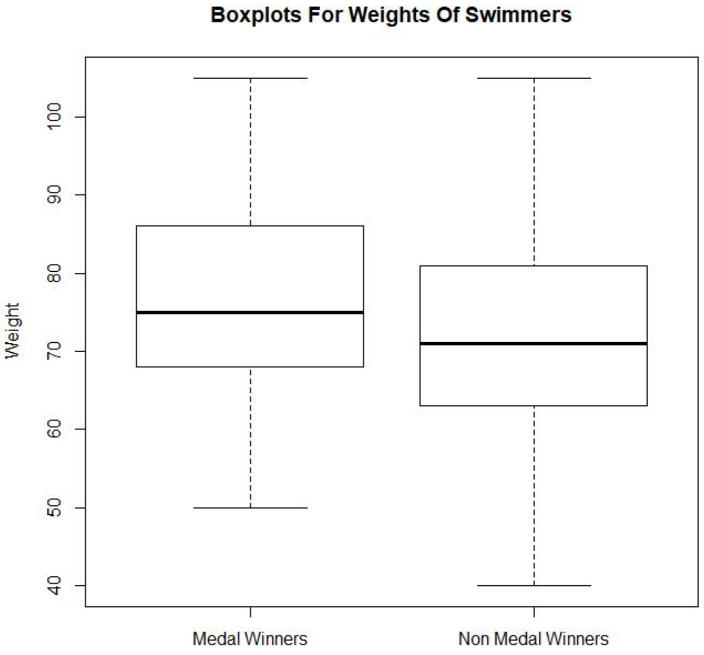
The association between medal-winning and height, weight and age amongst swimmers

Boxplots

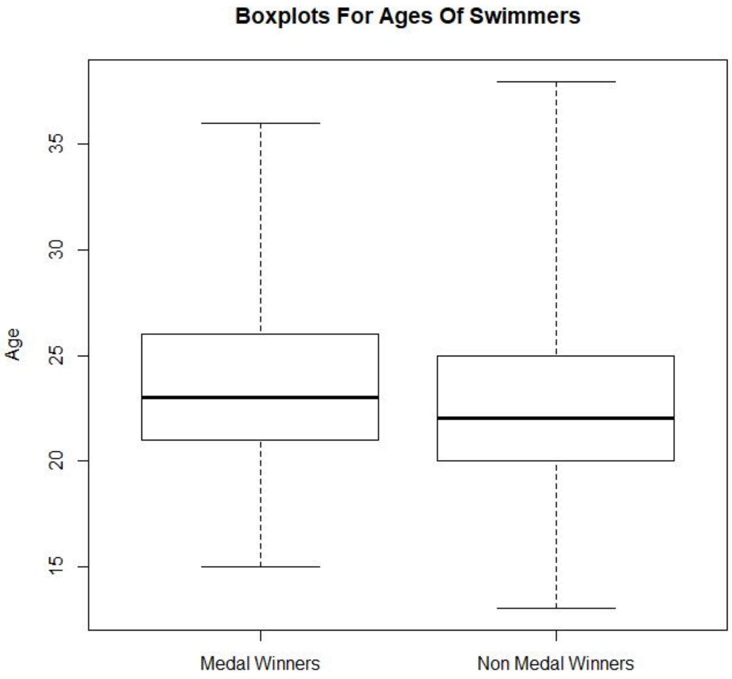
The boxplots for height show that the median height of medal winners is higher. The interquartile range (length of the boxes in height axis) are similar but the minimum and maximum heights are lower/higher respectively for non-medal winners.



The boxplots for weights show that the median weight of medal winners is greater. The maximum weight in each population is equal at just above 100kg but the minimum for non-medal winners is lower.



The boxplots for age show median age of medal winners slightly higher, again the non-medal winners have larger range in values.



We investigate this further using hypothesis tests. Suppose that μ_{medal} is the mean for medal-winning swimmers and that $\mu_{non\text{-}medal}$ is the mean of non-medal-winning swimmers. A two sample t-test (assuming equal variances) of H_0: $\mu_{medal}=\mu_{non\text{-}medal}$ versus two-sided H_1:$\mu_{medal}\neq\mu_{non\text{-}medal}$ was conducted, with results given by:

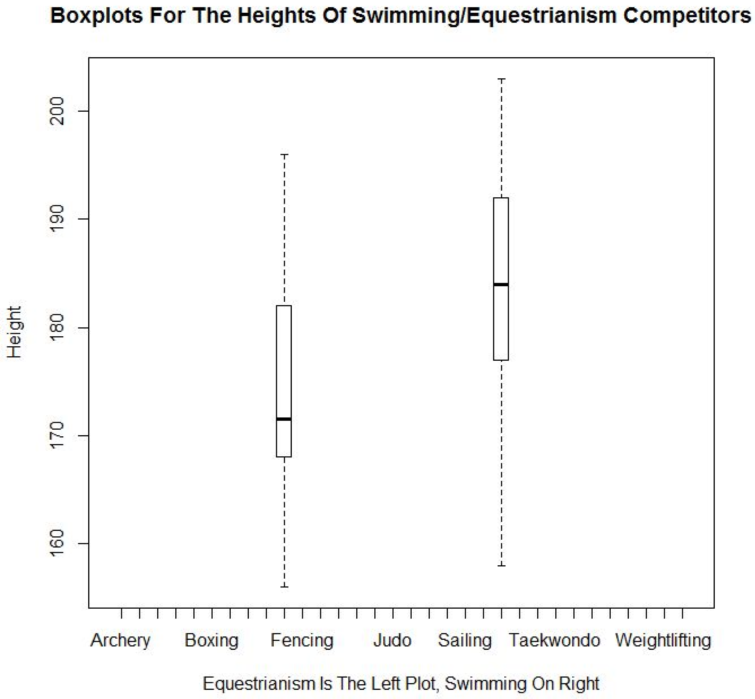
Variable	**Weight**	**Height**	**Age**
Test Statistic	5.6126	5.7301	3.4282
p-value	2.366e-08	1.208e-08	0.0006236
Acceptance Region	3.181513<t< 6.599997	2.746359<t<5.605257	0.4194069<t< 1.5412094

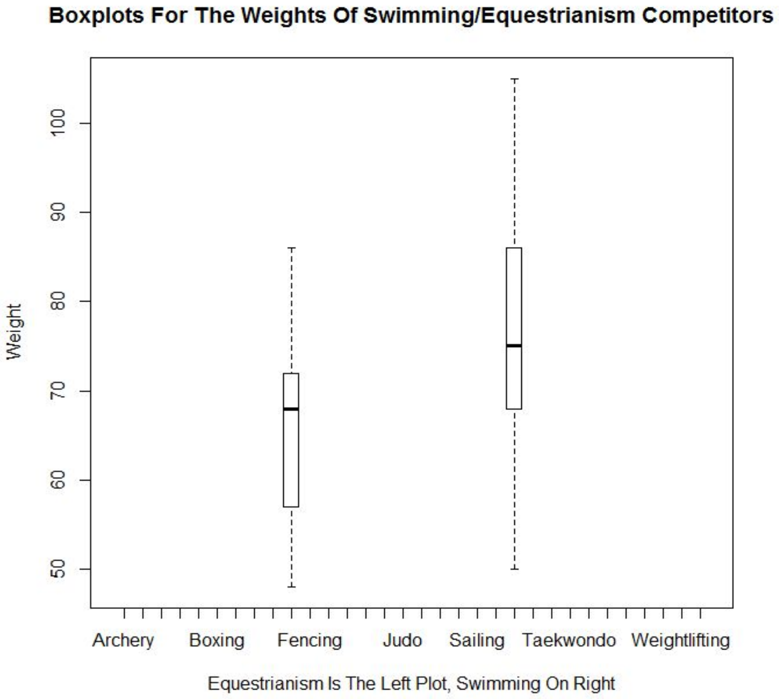
For weight, 3.181513<5.6126<6.599997 thus we do not reject H_0 , we conclude there is no significant difference between the weights of medal-winning swimmers and non-medal-winning swimmers.

For height, 5.7301>5.605257 thus we reject H_0 at the 95% level, there is strong evidence (p-value=1.208e-08<0.01) to suggest that the mean height of medal-winning swimmers is greater than non-medallists.
For age, 3.4282>1.5412094 thus we reject H_0 at 95% level, there is strong evidence(0.0006236<0.01) that medallists are older than non-medallists.

The association between height and weight for medal-winners in swimming and equestrian

Boxplots





These boxplots show that on average medal-winning equestrian competitors are both shorter and lighter than swimmers. The minimum values for equestrian are only slightly less than for swimming in the case of both variables. The maximum weight for swimmers is much heavier than the maximum for equestrian, these observations suggest that it may be an advantage for a swimmer to be heavier/taller than an equestrian competitor.

Hypothesis Tests

As above we take $\mu_{swimming}$ and $\mu_{equestrian}$ as the means for each sport, and test the same null/alternative hypotheses to see if there is a significant difference.

Variable	Height	Weight
Test statistic	6.5183	5.3812
p-value	4.237e-10	1.775e-07
Acceptance region	7.077418<t< 13.208363	6.564576<t< 14.146845

Thus there is strong evidence to reject the null for both variables, concluding that on average, swimmers height and weight is greater than that of the equestrian competitors. This shows in order to win a medal it may be an advantage to be lighter and shorter than a swimmer for an equestrian competitor. This is likely due to the fact that these features add ease for the horse to move. On the other hand a swimmer is aided by being taller and heavier as to swim quickly requires muscular mass, and being tall likely provides a slimmer frame aiding motion through water.

Conclusion

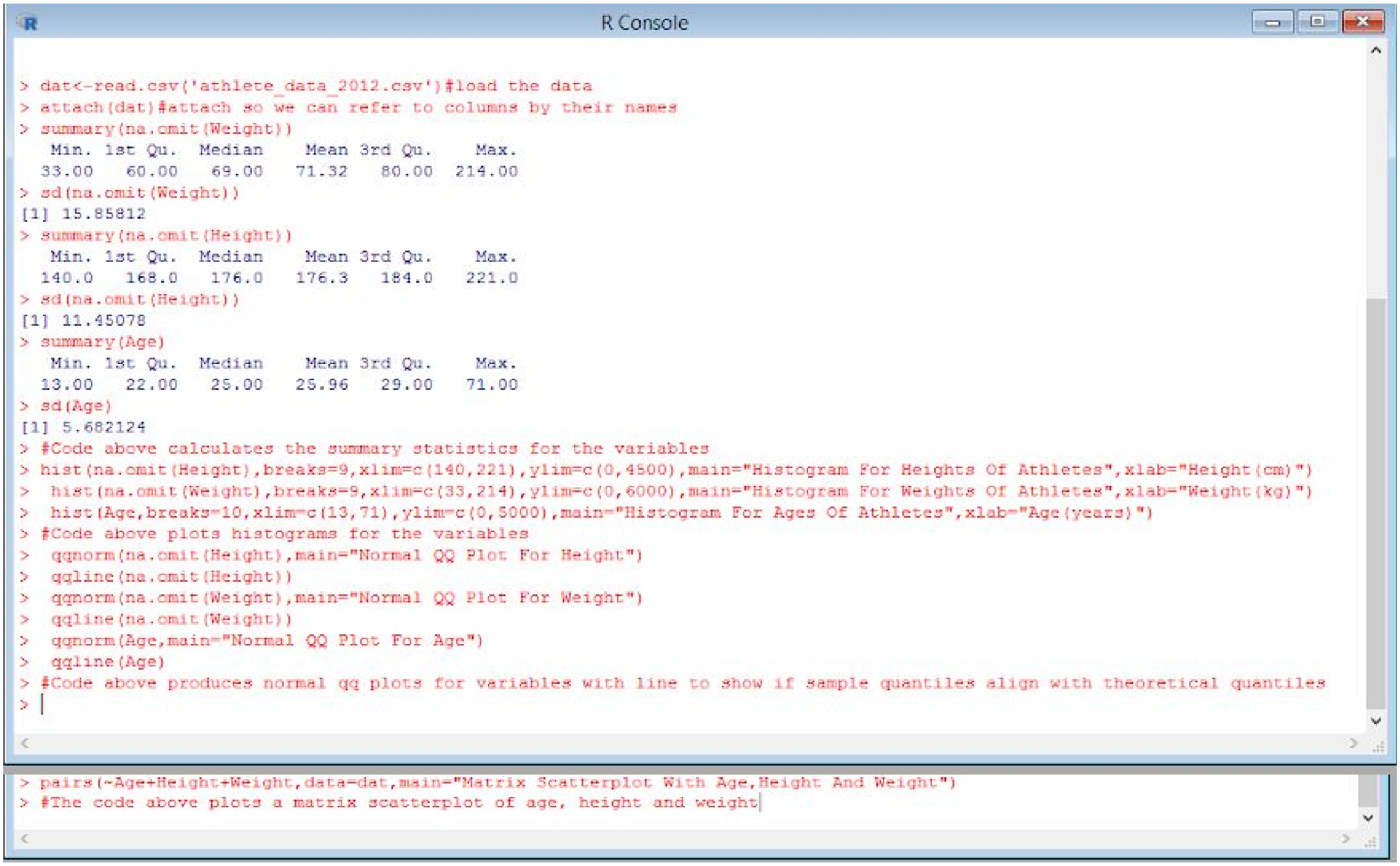
We conclude that the weight of London 2012 competitors has a larger variance than height or age. The distribution of their heights resembles a normal distribution. There seems to be some positive correlation between height and weight. Furthermore there are significant differences in the mean age and height of swimming medallists and non-medallists, with medallists being taller and older on average. Lastly, the mean heights/weights of swimming medallists is higher than that of equestrian medallists. Limitations of the data include many heights and weights not being measured (appear as NA in data), if they had been measured these may have altered the summary statistics especially for smaller subsets of the data that includes those competitors. To further analyse, data from older olympics could be compared with this data to see how the height/weight/age of the medallists changes over time for each sport.

Word count:1487

Appendix

The R code for producing the results and plots in this report is as follows.

Code For Analysis Of Competitors’ Height, Weight And Age:



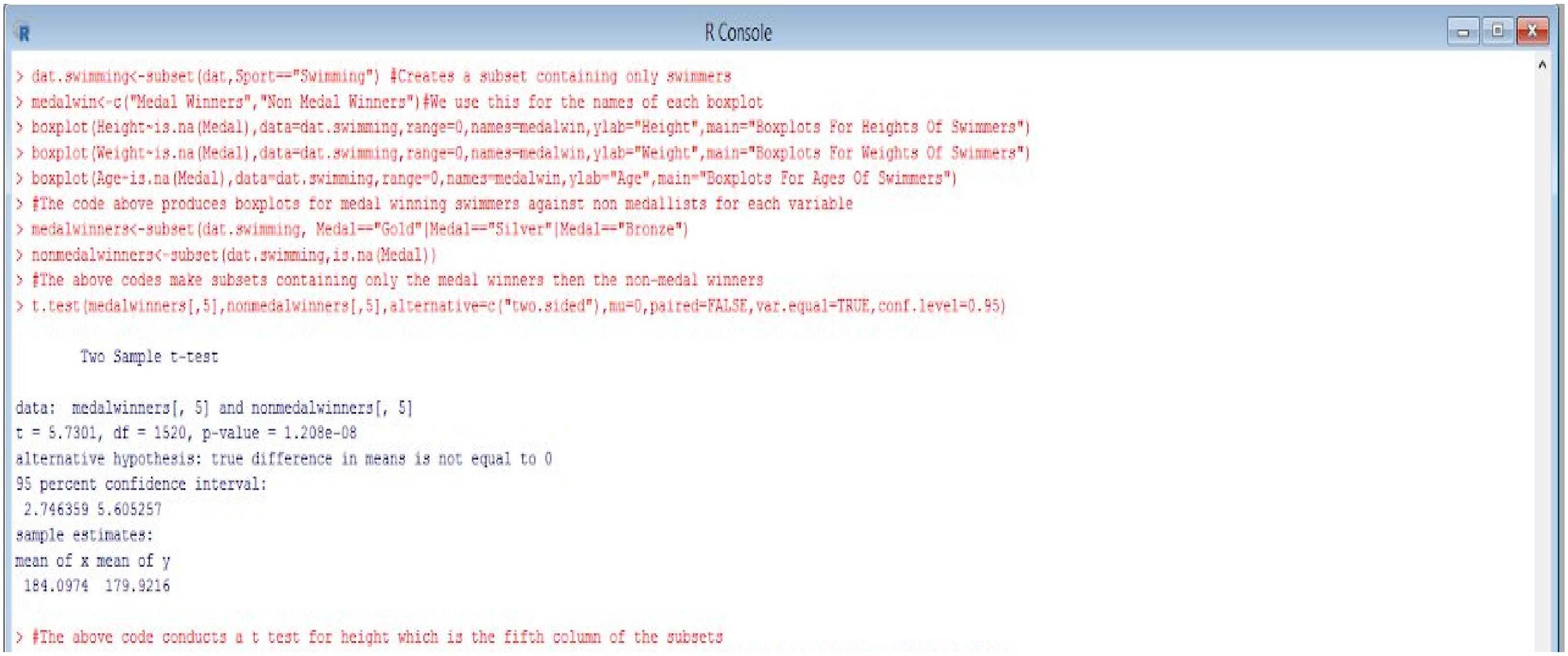
```
> dat<-read.csv('athlete_data_2012.csv')#load the data
> attach(dat)#attach so we can refer to columns by their names
> summary(na.omit(Weight))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  33.00   60.00   69.00   71.32   80.00  214.00 
> sd(na.omit(Weight))
[1] 15.85812
> summary(na.omit(Height))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  140.0   168.0   176.0   176.3   184.0   221.0 
> sd(na.omit(Height))
[1] 11.45078
> summary(Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  13.00   22.00   25.00   25.96   29.00   71.00 
> sd(Age)
[1] 5.682124
> #Code above calculates the summary statistics for the variables
> hist(na.omit(Height),breaks=9,xlim=c(140,221),ylim=c(0,4500),main="Histogram For Heights Of Athletes",xlab="Height(cm)")
>  hist(na.omit(Weight),breaks=9,xlim=c(33,214),ylim=c(0,6000),main="Histogram For Weights Of Athletes",xlab="Weight(kg)")
>  hist(Age,breaks=10,xlim=c(13,71),ylim=c(0,5000),main="Histogram For Ages Of Athletes",xlab="Age(years)")
> #Code above plots histograms for the variables
>  qqnorm(na.omit(Height),main="Normal QQ Plot For Height")
>  qqline(na.omit(Height))
>  qqnorm(na.omit(Weight),main="Normal QQ Plot For Weight")
>  qqline(na.omit(Weight))
>  qqnorm(Age,main="Normal QQ Plot For Age")
>  qqline(Age)
> #Code above produces normal qq plots for variables with line to show if sample quantiles align with theoretical quantiles
> 
```

```
> pairs(~Age+Height+Weight,data=dat,main="Matrix Scatterplot With Age,Height And Weight")
> #The code above plots a matrix scatterplot of age, height and weight
```

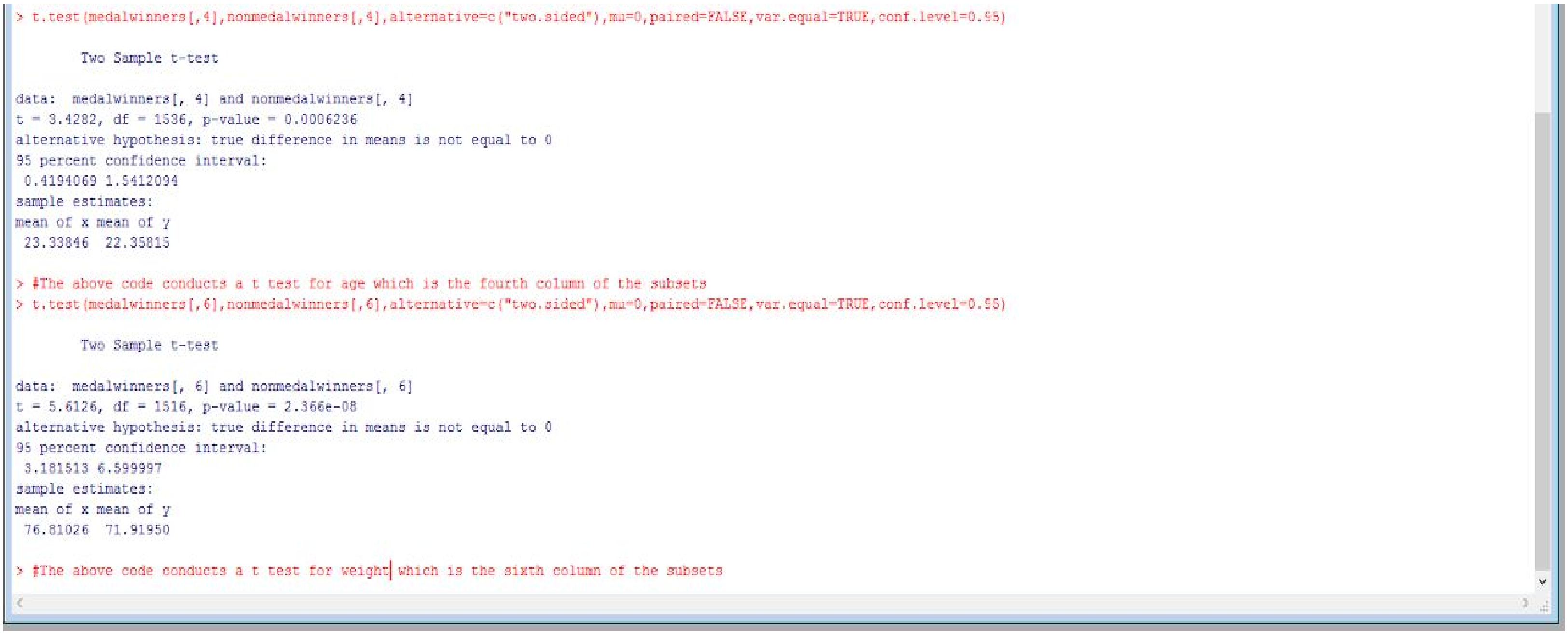
Code To Find Outliers:

```
subset(dat,Weight>=170)
subset(dat, Age>=60)
```

Code For The Association Between Medal-Winning And Height, Weight And Age Amongst Swimmers:

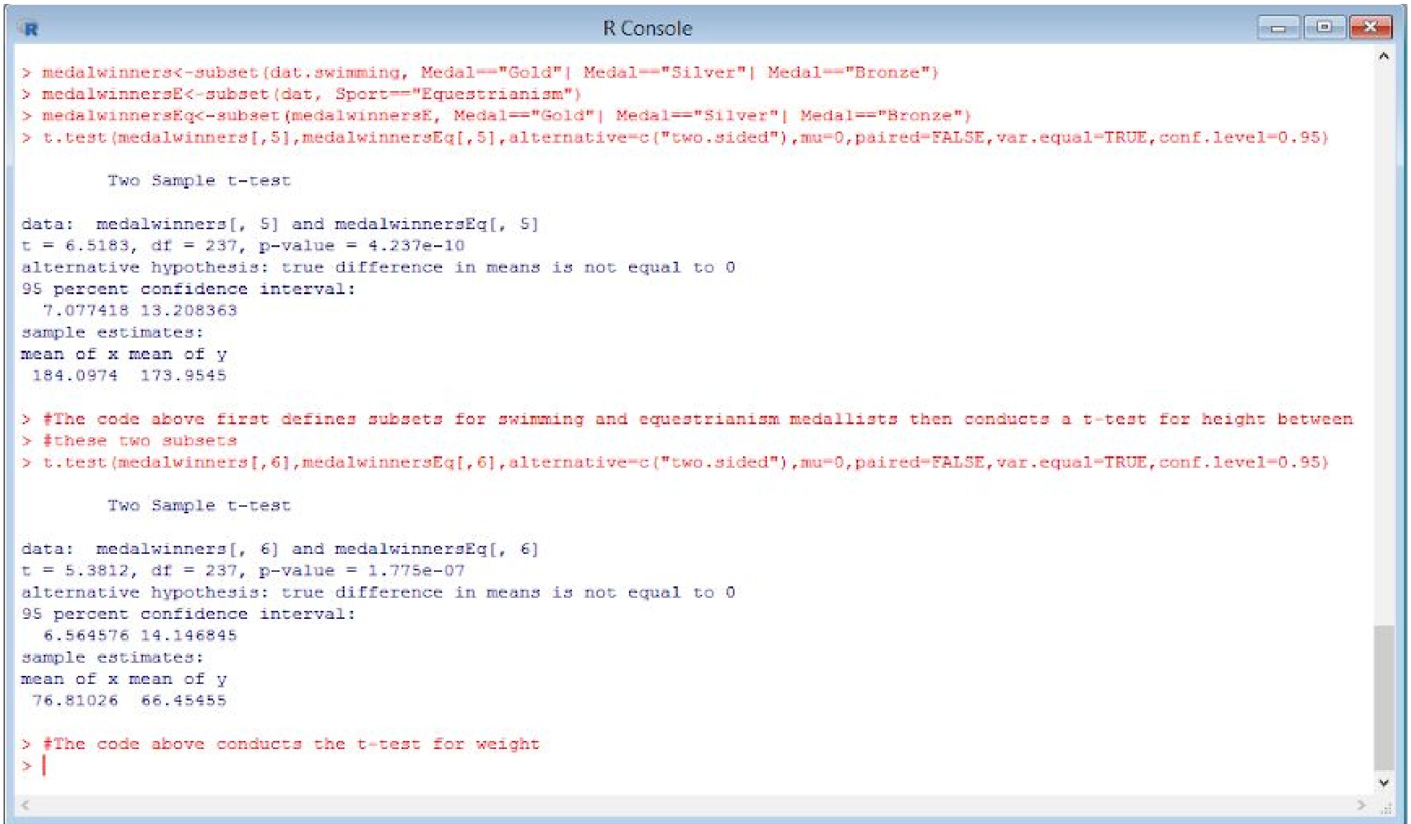


```
> dat.swimming<-subset(dat,Sport=="Swimming") #Creates a subset containing only swimmers
> medalwin<-c("Medal Winners","Non Medal Winners")#We use this for the names of each boxplot
> boxplot(Height~is.na(Medal),data=dat.swimming,range=0,names=medalwin,ylab="Height",main="Boxplots For Heights Of Swimmers")
> boxplot(Weight~is.na(Medal),data=dat.swimming,range=0,names=medalwin,ylab="Weight",main="Boxplots For Weights Of Swimmers")
> boxplot(Age~is.na(Medal),data=dat.swimming,range=0,names=medalwin,ylab="Age",main="Boxplots For Ages Of Swimmers")
> #The code above produces boxplots for medal winning swimmers against non medallists for each variable
> medalwinners<-subset(dat.swimming, Medal=="Gold"|Medal=="Silver"|Medal=="Bronze")
> nonmedalwinners<-subset(dat.swimming,is.na(Medal))
> #The above codes make subsets containing only the medal winners then the non-medal winners
> t.test(medalwinners[,5],nonmedalwinners[,5],alternative=c("two.sided"),mu=0,paired=FALSE,var.equal=TRUE,conf.level=0.95)

        Two Sample t-test

data:  medalwinners[, 5] and nonmedalwinners[, 5]
t = 5.7301, df = 1520, p-value = 1.208e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.746359 5.605257
sample estimates:
mean of x mean of y
 184.0974  179.9216

> #The above code conducts a t test for height which is the fifth column of the subsets
```

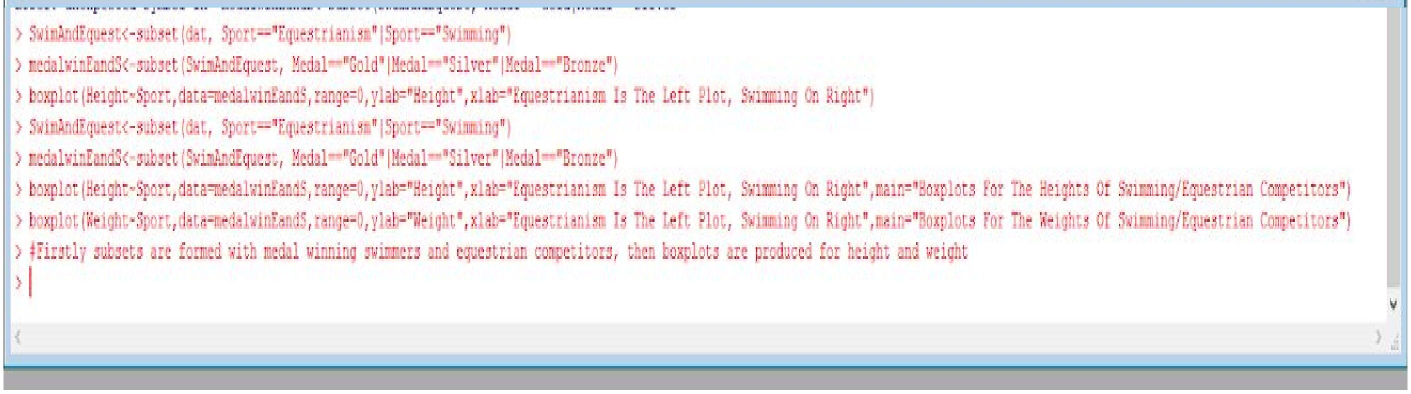


```
> t.test(medalwinners[,4],nonmedalwinners[,4],alternative=c("two.sided"),mu=0,paired=FALSE,var.equal=TRUE,conf.level=0.95)

        Two Sample t-test

data:  medalwinners[, 4] and nonmedalwinners[, 4]
t = 3.4282, df = 1536, p-value = 0.0006236
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4194069 1.5412094
sample estimates:
mean of x mean of y
 23.33846  22.35815

> #The above code conducts a t test for age which is the fourth column of the subsets
> t.test(medalwinners[,6],nonmedalwinners[,6],alternative=c("two.sided"),mu=0,paired=FALSE,var.equal=TRUE,conf.level=0.95)

        Two Sample t-test

data:  medalwinners[, 6] and nonmedalwinners[, 6]
t = 5.6126, df = 1516, p-value = 2.366e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.181513 6.599997
sample estimates:
mean of x mean of y
 76.81026  71.91950

> #The above code conducts a t test for weight which is the sixth column of the subsets
```

Code For The Association Between Height And Weight For Medal-Winners In Swimming And Equestrian:



```
> medalwinners<-subset(dat.swimming, Medal=="Gold"| Medal=="Silver"| Medal=="Bronze")
> medalwinnersE<-subset(dat, Sport=="Equestrianism")
> medalwinnersEq<-subset(medalwinnersE, Medal=="Gold"| Medal=="Silver"| Medal=="Bronze")
> t.test(medalwinners[,5],medalwinnersEq[,5],alternative=c("two.sided"),mu=0,paired=FALSE,var.equal=TRUE,conf.level=0.95)

        Two Sample t-test

data:  medalwinners[, 5] and medalwinnersEq[, 5]
t = 6.5183, df = 237, p-value = 4.237e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  7.077418 13.208363
sample estimates:
mean of x mean of y 
 184.0974  173.9545 

> #The code above first defines subsets for swimming and equestrianism medallists then conducts a t-test for height between
> #these two subsets
> t.test(medalwinners[,6],medalwinnersEq[,6],alternative=c("two.sided"),mu=0,paired=FALSE,var.equal=TRUE,conf.level=0.95)

        Two Sample t-test

data:  medalwinners[, 6] and medalwinnersEq[, 6]
t = 5.3812, df = 237, p-value = 1.775e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  6.564576 14.146845
sample estimates:
mean of x mean of y 
 76.81026  66.45455 

> #The code above conducts the t-test for weight
> |
```



```
> SwimAndEquest<-subset(dat, Sport=="Equestrianism"|Sport=="Swimming")
> medalwinEandS<-subset(SwimAndEquest, Medal=="Gold"|Medal=="Silver"|Medal=="Bronze")
> boxplot(Height~Sport,data=medalwinEandS,range=0,ylab="Height",xlab="Equestrianism Is The Left Plot, Swimming On Right")
> SwimAndEquest<-subset(dat, Sport=="Equestrianism"|Sport=="Swimming")
> medalwinEandS<-subset(SwimAndEquest, Medal=="Gold"|Medal=="Silver"|Medal=="Bronze")
> boxplot(Height~Sport,data=medalwinEandS,range=0,ylab="Height",xlab="Equestrianism Is The Left Plot, Swimming On Right",main="Boxplots For The Heights Of Swimming/Equestrian Competitors")
> boxplot(Weight~Sport,data=medalwinEandS,range=0,ylab="Weight",xlab="Equestrianism Is The Left Plot, Swimming On Right",main="Boxplots For The Weights Of Swimming/Equestrian Competitors")
> #Firstly subsets are formed with medal winning swimmers and equestrian competitors, then boxplots are produced for height and weight
> |
```