Assignment1

Data Overview (Univariate data)

Q: Univariate data. Get a univariate dataset from sources 1 or 2 and briefly describe it.

1078 measurements of a father's height and his son's height. The heights are measured in inch. To just have univariable data I just took the heights of the fathers and saved them into a seperate variable.

```
library(UsingR)
```

```
## Loading required package: MASS

## Loading required package: HistData

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units

##
## Attaching package: 'UsingR'

## The following object is masked from 'package:survival':
##
##     cancer
```

```
fatherHeight = round(father.son$fheight, 1)
length(fatherHeight)
```

```
## [1] 1078
```

```
fatherHeight=na.omit(fatherHeight)
length(fatherHeight)
```

```
## [1] 1078
```

```
tail(fatherHeight)
```

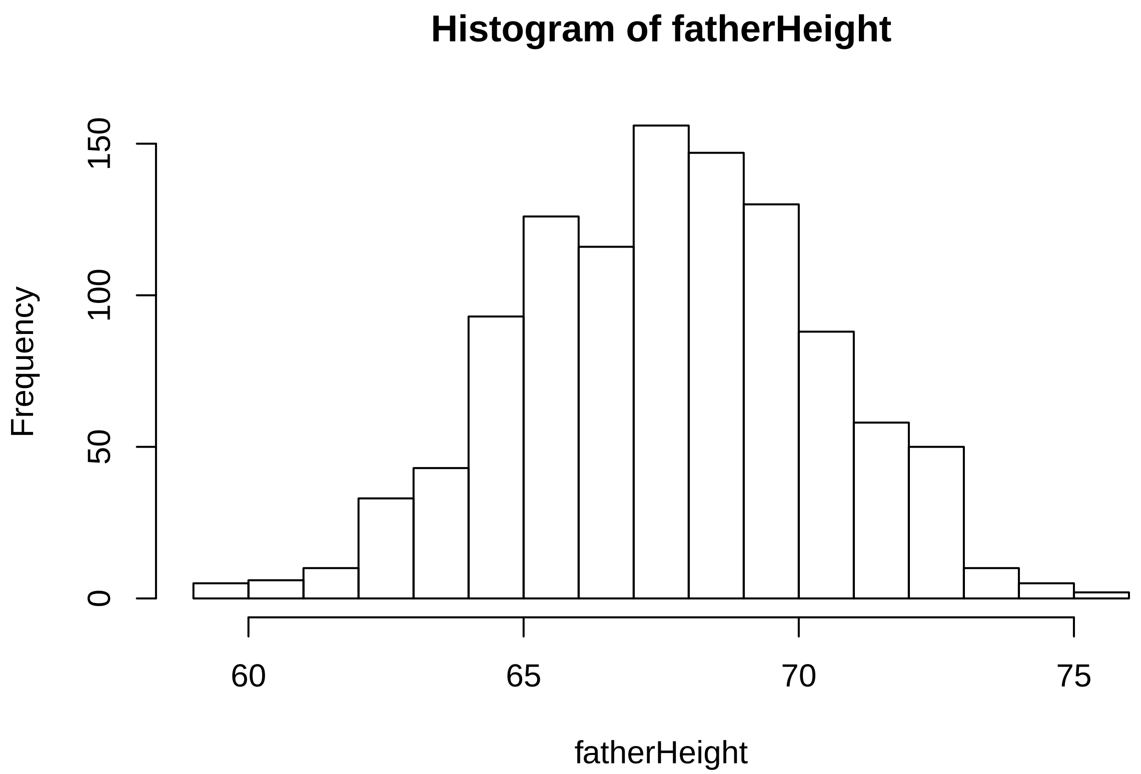
```
## [1] 67.7 67.0 71.3 71.8 70.7 70.3
```

We also see that there are no NA in the Dataset.And on top of that I am rounding the numerical values of the height. This is because otherwise some functions wont work properly.

Q: Construct a stem-and-leaf and histogram. Impose the empirical density estimate on the histogram. Discuss the results focusing on the shape of the plots and number of modes.

Histogram

```
hist(fatherHeight)
```



We see in the historgram that the height of the fathers is normally distributed.The empirical density of the historgram would be 1 inch.

Steam-and-leaf plot

```
stem(fatherHeight)
```

```
##
##   The decimal point is at the |
##
##   59 | 0556
##   60 | 02489
##   61 | 00111357889
##   62 | 0111234444555777778888999999
##   63 | 00000011222222333356666777777888889999999999999
##   64 | 0011111111133333333444444555555555555666666666666667777777777777788 8+4
##   65 | 00000000000111111122222222333333334444444444455555555555666666666666+39
##   66 | 000000000000000000111111111111111222222222222333333334444444455555555+45
##   67 | 00000000011111111111111111222222222222233333333333333344444444444444+70
##   68 | 00000000000000011111111111111222222222222223333333333333333333444444+67
##   69 | 00000000000000011111111111111111222222222233333333333333333344444444+51
##   70 | 00000000000000111111112222222222222222233333333344444444555555555555+16
##   71 | 0000001111122222222333333333334444444455555556667778888999
##   72 | 0000001112222233333334444444444455555666667778888899
##   73 | 00001223334599
##   74 | 4789
##   75 | 024
```

```r
which(table(fatherHeight) == max(table(fatherHeight)))
```

```
## 67.4 67.8 68.3
##   67   71   76
```

In the steam-and-leaf plot we can see that the mode is around 68.3 with 76 occurences. This function wont work unless it has only one diggit after the comma. We can also see the noramal distribution in this plot.

Q: Compute the mean and median. Based solely on that, conclude whether the distribution is skewed. Find the proportion of the data which are less than the mean value. Compute the 1st and 3rd quartiles, the 90th quantile and the mode. Explain the meaning of the obtained quantities. Find the value that cuts off the top 25% of the data.

Compute median and mean and quartil

```r
summary(fatherHeight)
```

Not nessesarily:
any symmetric
distribution

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   59.00   65.80   67.80   67.69   69.60   75.40
```

Here we cann see that the median and the mean are approximetally the same and therefore the data is not scewed -> normal distributed. The first quartil (25%) lies at 65.80. The third quartil (75%) lies at 69.60. Because I dont quite unsterstand what is meant by cutting of the "top" 25% of the data it could be either the first quartil cutting of the 25% of the data under 65.80 or the third quartil cutting of the last 25% above 69.60.

Correct

90th quartil

```r
quantile(fatherHeight, c(.90))
```

```
##  90%
## 71.3
```

This means that 90% of the data lies under 71.3.

Q: Compute the range, the sample standard deviation and the IQR. Construct the boxplot of the data. Comment on the boxplot including skewness, outliers etc.

Range, standart deviation and the IQR

```
range(fatherHeight)
```

```
## [1] 59.0 75.4
```

```
sd(fatherHeight)
```

```
## [1] 2.745827
```

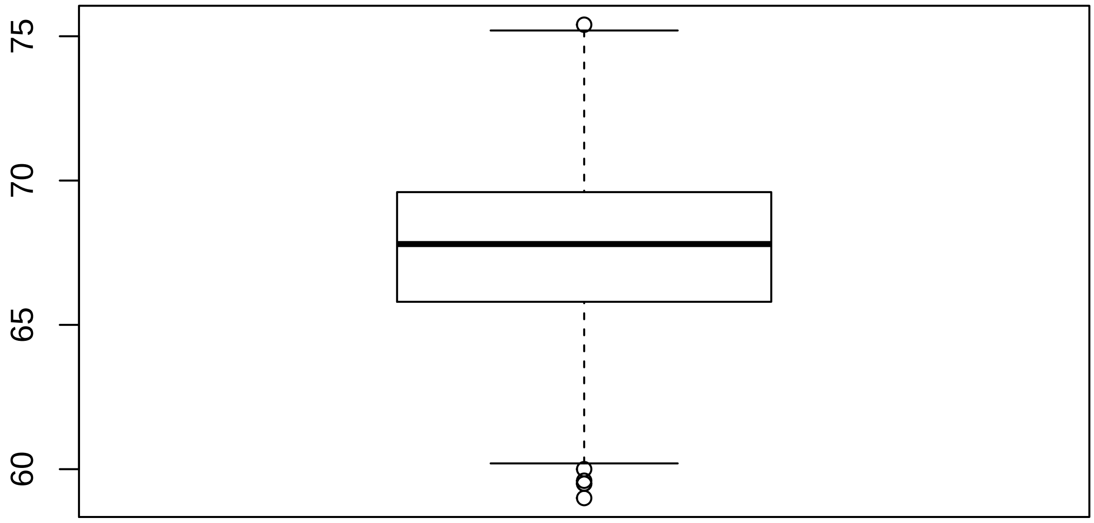
```
IQR(fatherHeight)
```

```
## [1] 3.8
```

The range would be from 59.0 to 75.4 and the standart deviation would be 2.74608. This means that most of my data lies in a range between 64,94392 and 70,43608. My IQR is 3.8. This means that 50% of my data lies in a range of 3.8.

Boxplot

```
boxplot(fatherHeight)
```

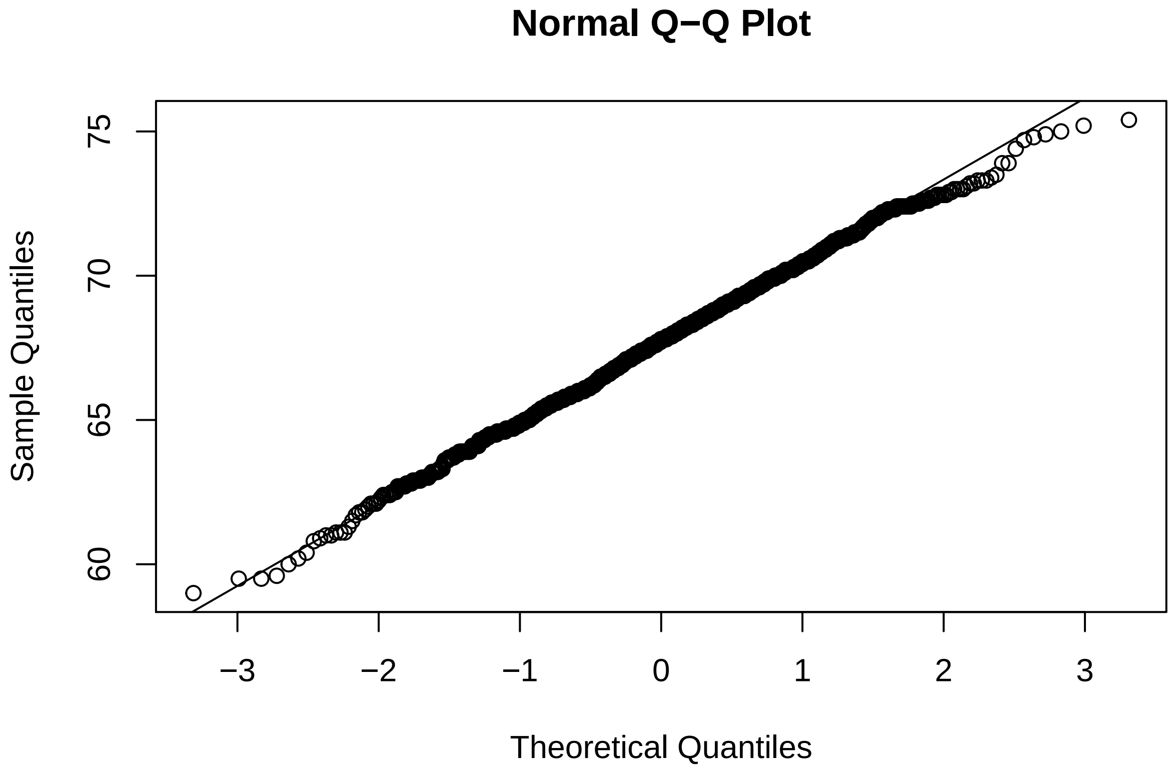


In the boxplot we also see the IQR visually.We also see that I have a noramal distribution. No scewness in my data. There are some outline which we see marked as circles above and below the plot.

Check whether the empirical distribution is normal by examining the QQ-plot.

Normal QQ-Plot

```
qqnorm(fatherHeight)
qqline(fatherHeight)
```



Here we see that the qqline and the data are in a high correlation. Therefore it normal distributed.

Data Overview (Bivariate data)

Because father.son has two variable ill use this dataset again for the bivariate tests.

```
fatherHeight = round(father.son$fheight, 1)
sonHeight = round(father.son$sheight, 1)
length(fatherHeight)
```

```
## [1] 1078
```

```
length(sonHeight)
```

```
## [1] 1078
```

```
fatherHeight=na.omit(fatherHeight)
sonHeight=na.omit(sonHeight)
length(fatherHeight)
```

```
## [1] 1078
```

```
length(sonHeight)
```

```
## [1] 1078
```

```
tail(fatherHeight)
```

```
## [1] 67.7 67.0 71.3 71.8 70.7 70.3
```

```
tail(sonHeight)
```

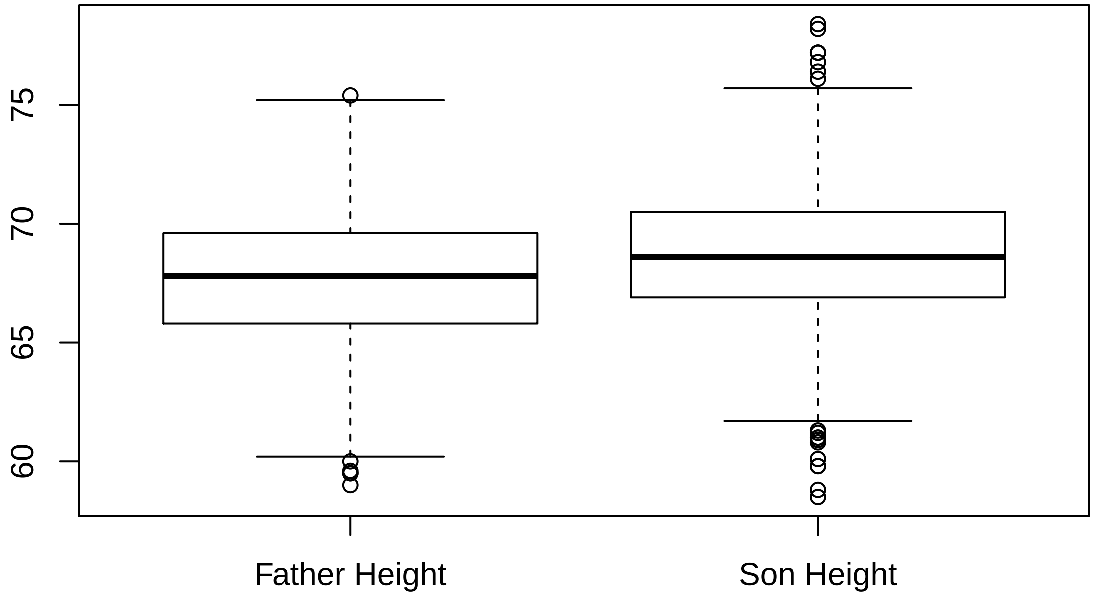
```
## [1] 59.8 70.8 68.3 69.3 69.3 67.0
```

There are on both Datasets 1078 dataentries and no NA.

Q: Create side-by-side boxplots. Compare the centers and spreads.

Side by Side Boxplot

```
boxplot(fatherHeight,sonHeight, names=c("Father Height","Son Height"))
```

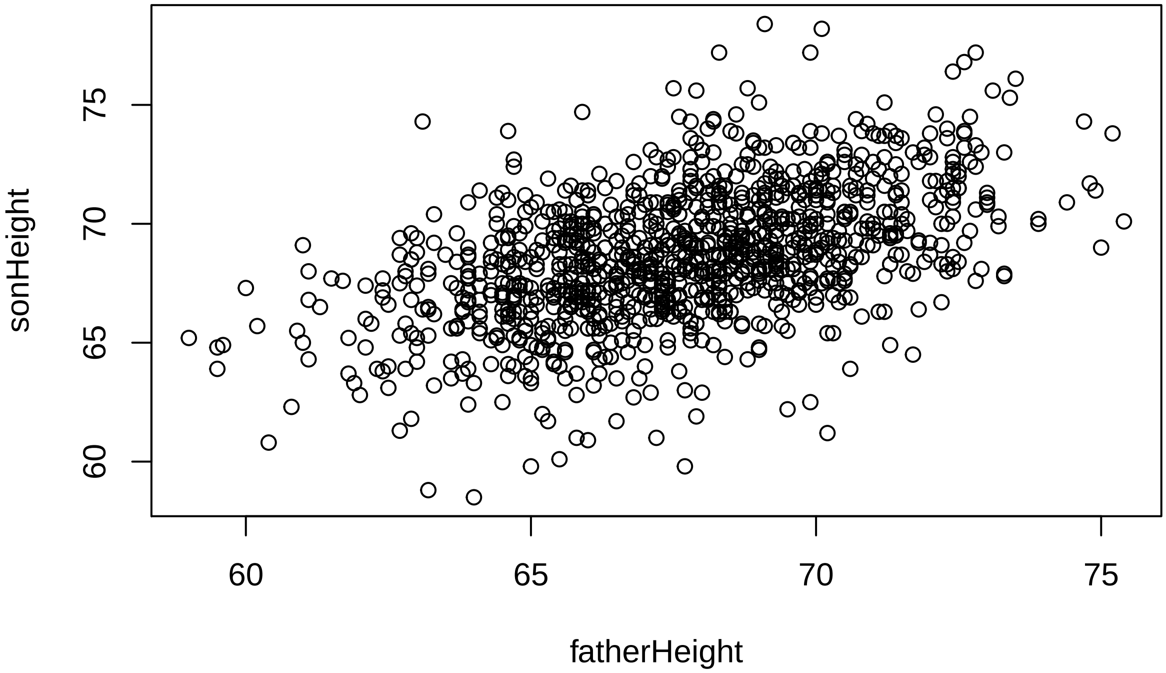


Here we see that The heights of the fathers are in dendency a bit lower that those of their sons. Aside from that the two boxplots look the same. There just seem to be more outliners on the sons height then on the side of the fatehrs height.

Q: Draw the scatter plot. Comment on the possible dependence and presence of outliers.

Scatter plot

```
plot(fatherHeight, sonHeight)
```



The data seems to be almost near together in a cycle. There are a few datapoint which are fzurther awa than the rest but They dont seem like really hard outliners. For me there seems to be a correlation between the two datasets. Also both are normal distributed.

Q: Compute Pearson's and Spearman's coeffcient of correlation. Interpret and compare their values. Are their values consistent with the scatter plot?

Spearman's Coefficient of Correlation

```
cor(fatherHeight, sonHeight, method="spearman")
```

```
## [1] 0.5056466
```

```
cor(fatherHeight, sonHeight, method="pearson")
```

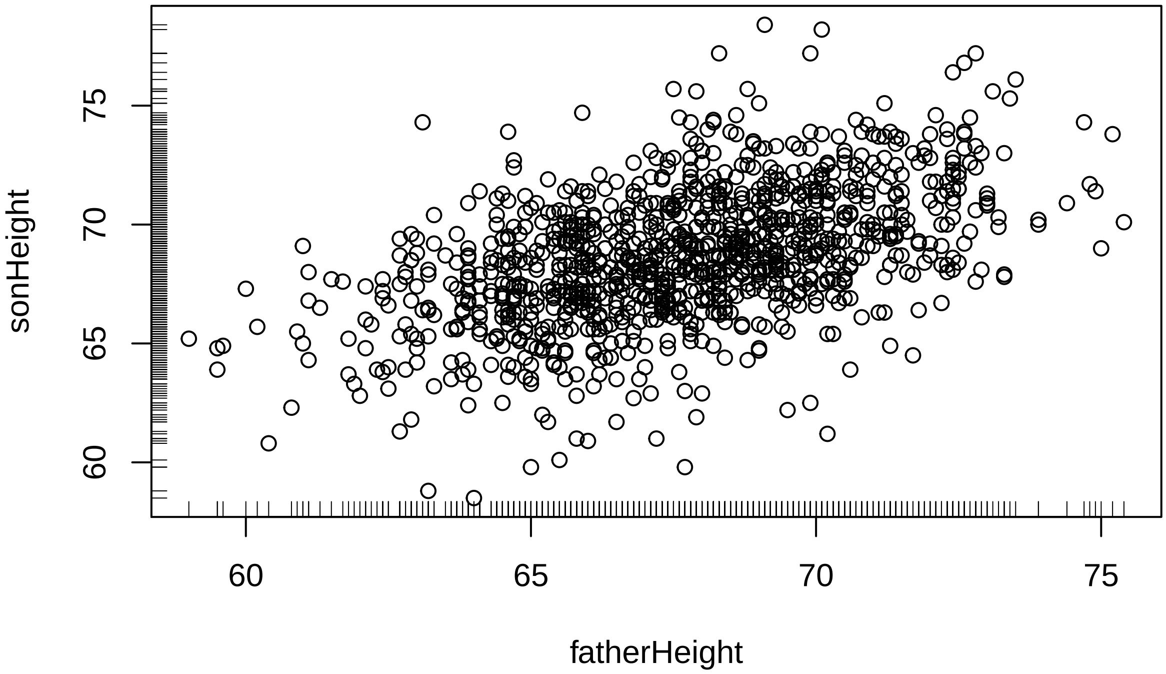
```
## [1] 0.5011627
```

A correlation of 0.5056466 for spearman and a correlation of 0.5011627 for spearman, means that there is a correlatoin between the heights of the father and the heights of their sons.

Q: Add the marginal distributions to the scatter plot. For that purpose, use histogram and box plot.

Marginal distribution

```
plot(fatherHeight, sonHeight)
rug(fatherHeight, side = 1)
rug(sonHeight, side = 2)
```



The "Plotting joint and marginal distributions together" is at the end of the pdf because of some problems

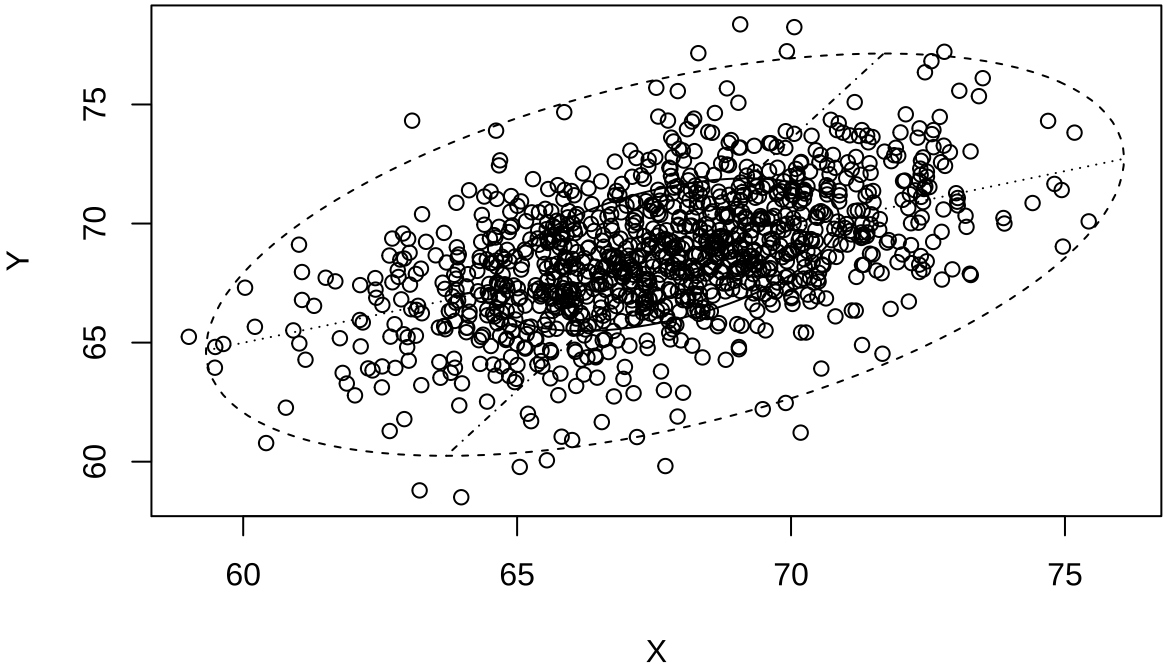
Q: Depict the bivariate box plot. Comment on the outliers. Remove the outliers, if any, and re-compute the Pearson correlation coeffcient.

Bivariate Boxplot

```
library(MVA)
```

```
## Loading required package: HSAUR2
## Loading required package: tools
```

```
bvbox(father.son)
```



```
upperwhiskerF = min(max(fatherHeight), 69.6 + 1.5 * (69.5-65.79))
lowerwhiskerF = max(min(fatherHeight), 65.79 - 1.5 * (69.5-65.79))
upperwhiskerS = min(max(sonHeight), 70.47 + 1.5 * (70.47-66.93))
lowerwhiskerS = max(min(sonHeight), 66.93 - 1.5 * (70.47-66.93))
fatherHeightO=subset(fatherHeight, fatherHeight<upperwhiskerF & fatherHeight>lowerwhiskerF)
sonHeightO=subset(sonHeight, sonHeight<upperwhiskerS & sonHeight>lowerwhiskerS)
cor(fatherHeight, sonHeight, method="pearson")
```

```
## [1] 0.5011627
```

Here we see some outliers liying outside.I havent found a way to remove the many outlires other than manually.

Because there are so many i didnt remove them

OK, got it

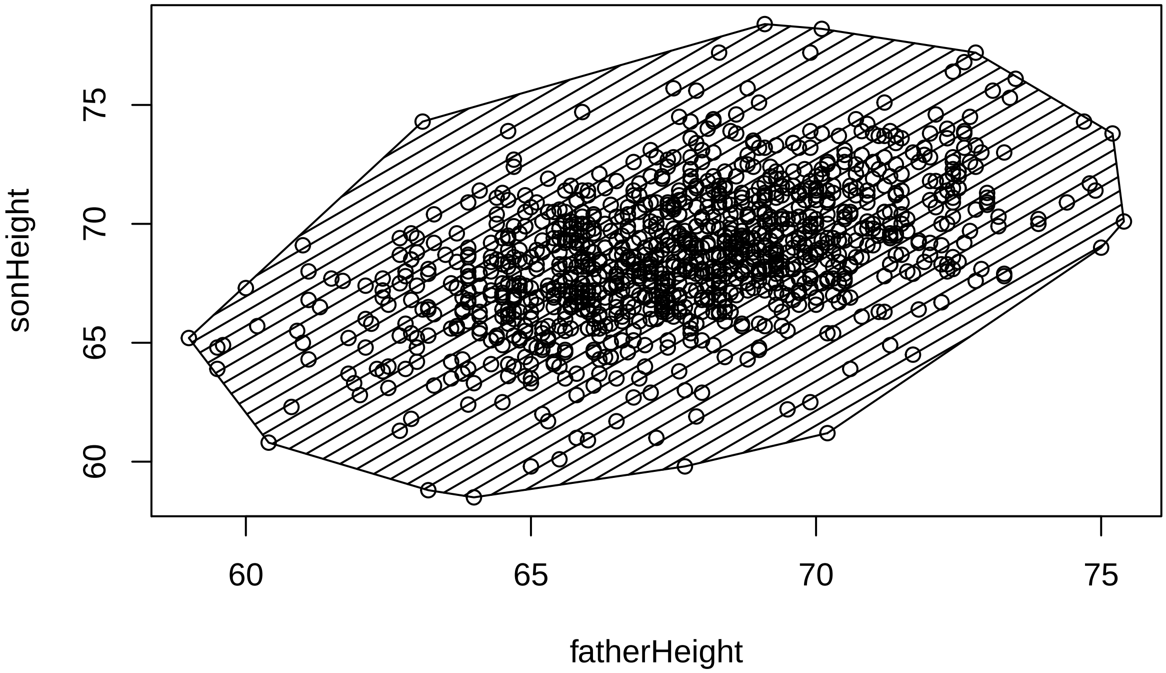
Q: Create the convex hull. Remove the observations lying on the hull and re-compute the correlation coeffcient.

Convex Hull

```
(hull <- with(father.son, chull(fatherHeight, sonHeight)))
```

```
##  [1]  852 1073  423 1070  420   11  197  851  635  204 1064  158  134
```

```
plot(fatherHeight, sonHeight, pch = 1)
with(father.son, polygon(fatherHeight[hull], sonHeight[hull], density = 15, angle = 30))
```



```
with(father.son, cor(fatherHeight[-hull],sonHeight[-hull]))
```

```
## [1] 0.5041026
```

After not considering the outliers I have a correlation of 0.5043638. But it is not much difference to 0.5011627 from pearson before.

Data Overview (Multivariate data)

Data set from UsingR: normtemp. A data set used to investigate the claim that “normal” temperature is 98.6 degrees. Gender 1 = male, 2 = female

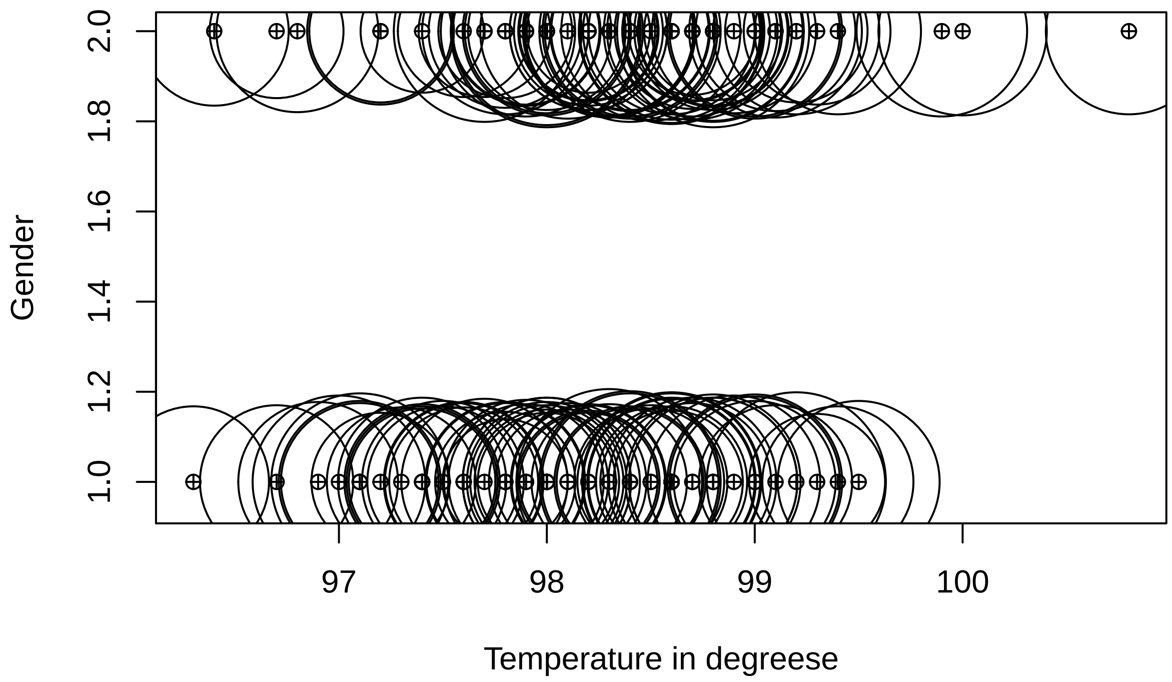
```
tail(normtemp)
```

```
##     temperature gender hr
## 125        99.2      2 66
## 126        99.3      2 68
## 127        99.4      2 77
## 128        99.9      2 79
## 129       100.0      2 78
## 130       100.8      2 77
```

Q: Pick up a dataset which has three variables (from source 2 or 3) and create the bubble plot. Interpret the result. See [2], Section 2.3.

Bubble Plot

```
ylim <- with(normtemp, range(normtemp$gender)) * c(0.95, 1)
plot(normtemp$gender ~ normtemp$temperature, data=normtemp,
     xlab = "Temperature in degreese",
     ylab = "Gender", pch = 10,
     ylim = ylim)
with(normtemp, symbols(normtemp$temperature, normtemp$gender, circles = normtemp$hr,
                       inches = 0.5, add = TRUE))
```



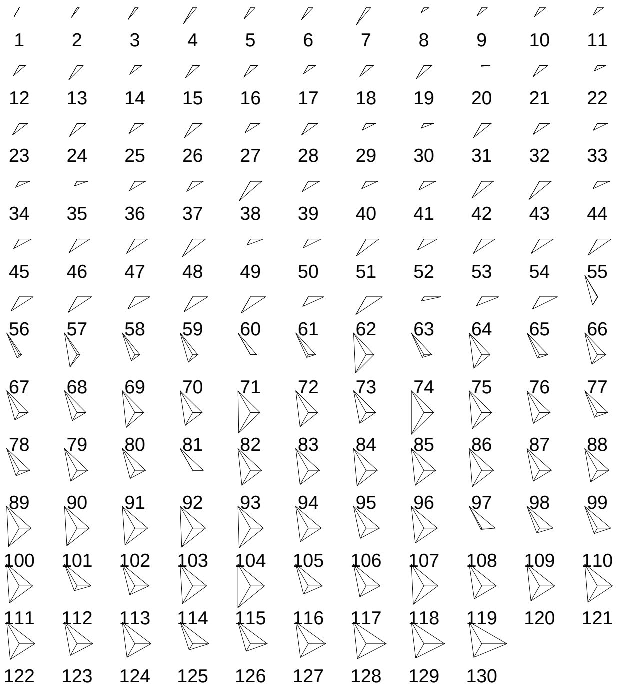
Here we see the two genders and their Temperature in degrees. The circles are the heart rate of the spicific points. here we see that for male = 1 the heart rate doesnt change that much despite a temperature change.

For females we see that there are changes in the heartrate while looking at the degrees.

Q: Use data source 2 or 3. Create the glyph plot of all observations, Section 2.3. Do any stars look alike?

Glyph plot

```
stars(normtemp, cex = 0.55)
```

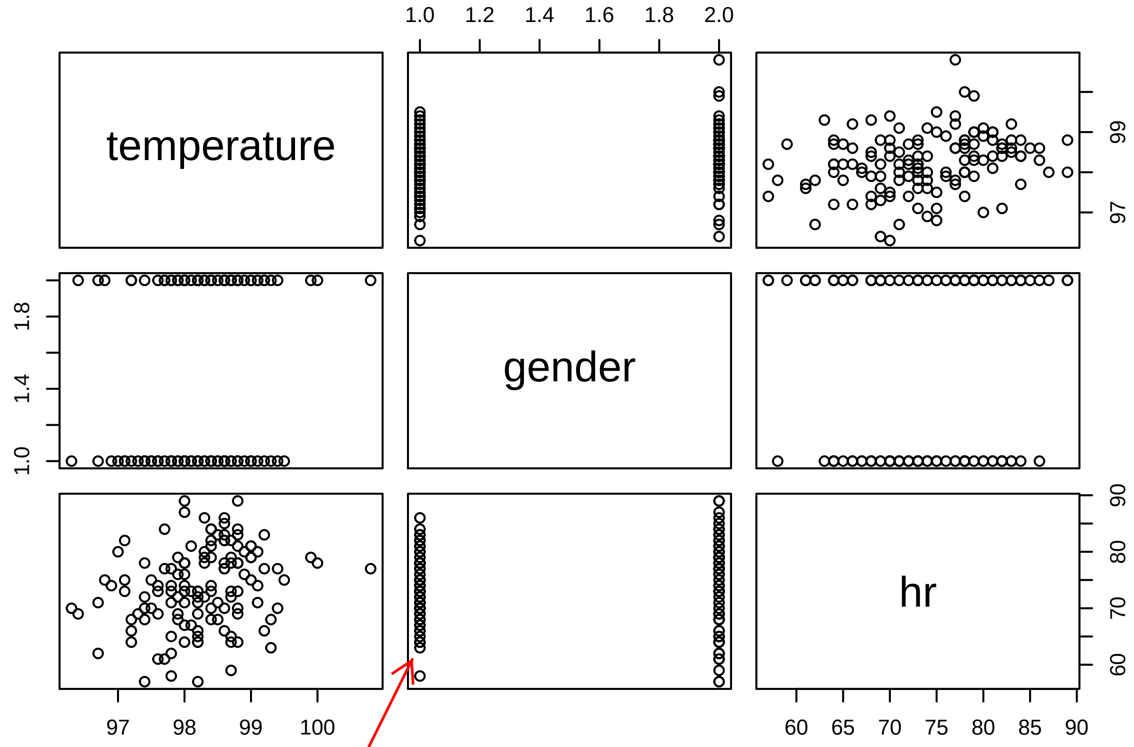


Here we can see that the symbols change trastically at around 62 this. Before and after the symbols almost look indentically. This could be the indicator that the gender is changed at aroung data 62.

Q: Use data source 2 or 3. Create the scatter plot matrix and analyze it. See [2], Section 2.4.

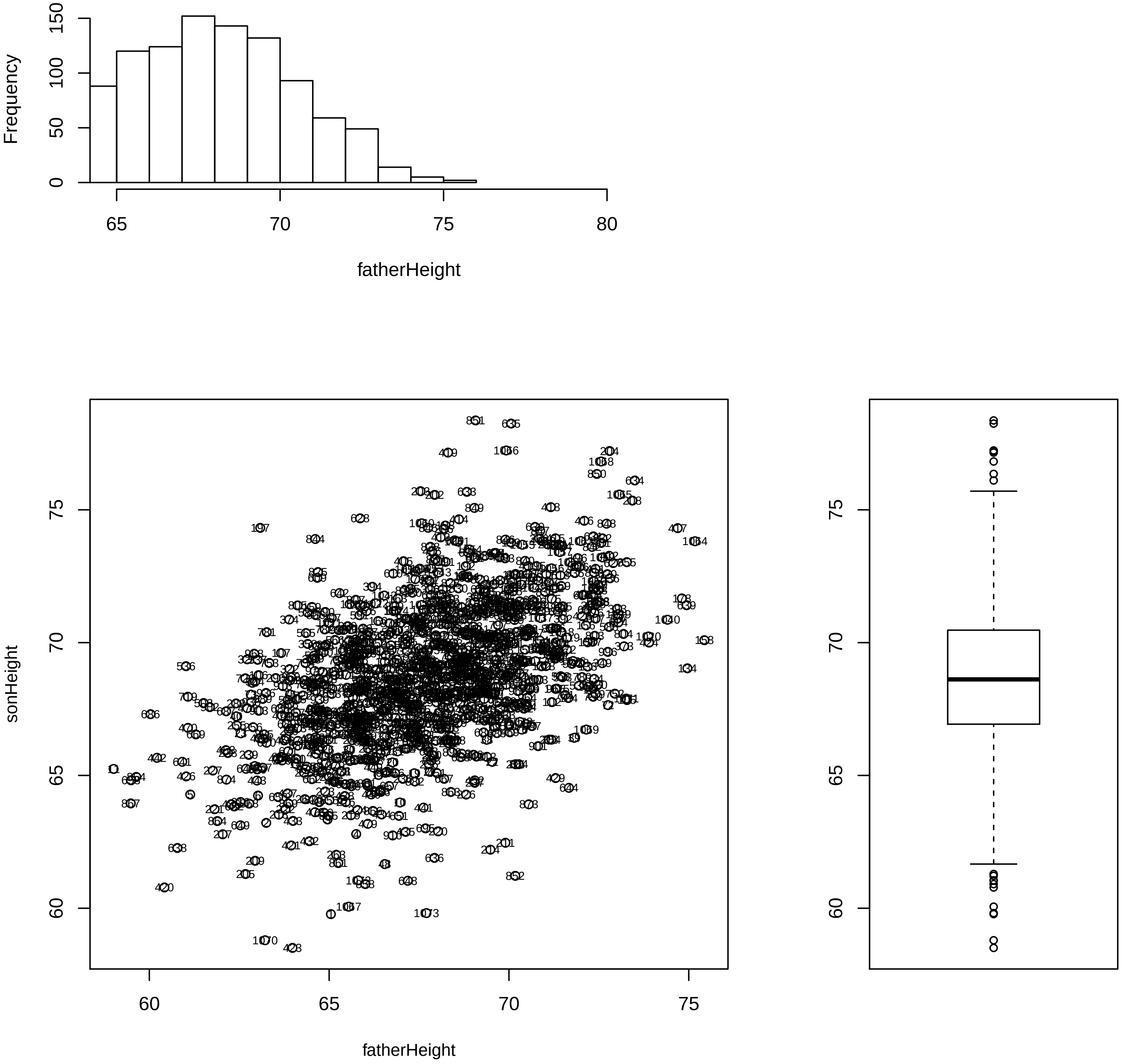
Scatter plot matrix

```
plot(normtemp)
```



Here we can see that the heartrate and the temperature are evenly scatter for both male and female. There might be a few "outliers" but I cant say that for sure. To me it looks like that the data is almost normally distributed.

Gender is a qualitative rather than quantitative variable. Therefore the concept of correlation cannot be applied, i.e., the wrong dataset for this task. A severe mistake.


Frequency
0
50
100
150
65
70
75
80
fatherHeight
sonHeight
60
65
70
75
60
65
70
75
fatherHeight
60
65
70
75