```
> myData <- read.table("Education.txt", header = T)
> # Let's see the first 10 rows of myData
> myData[1:10,]
   ID Education Gender   Wage
1   1        11      1 5313.0
2   2        15      2 5398.8
3   3        10      1 3764.4
4   4        11      2 3882.9
5   5        17      1 6849.9
6   6        16      1 6432.3
7   7        10      2 3414.3
8   8        17      1 6789.7
9   9        12      2 4113.7
10 10        13      1 5190.9
> # QUESTION 1
> # For each variable compute the mean, the median, the standard deviation,
> # the minimum and the maximum value.
> # Variable Education
> min(myData$Education)
[1] -12
> # -12 years of education makes no sense.
> # So we should remove this information from our table
> # We ask R which is the row where this value is located
> which(myData$Education == min(myData$Education))
[1] 234
> # We drop that line using negative subscript
> myData1 <- myData[-c(234),]
> # Let's check again the minimum value
> min(myData1$Education)
[1] 5
> max(myData1$Education)
[1] 22
> # Variable Gender
> max(myData1$Gender)
[1] 20
> # Gender can be 1 or 2 (male or female), so 20 makes no sense.
> # Therefore, we remove the line corresponding to this value
> which(myData1$Gender == max(myData1$Gender))
[1] 107
> myData2 <- myData1[-c(107),]
> # Check the max again
> max(myData2$Gender)
[1] 2
> min(myData2$Gender)
[1] 1
> # Variable Wage
> min(myData2$Wage)
```

```
[1] 41.8
> # 41.8 Chf a month is an odd income. It is better to remove the corresponding line from
the table
> which(myData2$Wage == min(myData2$Wage))
[1] 433
> myData3 <- myData2[-c(433),]
> min(myData3$Wage)
[1] 2047.2
> max(myData3$Wage)
[1] 8453.5
> # Now the table is clean (no values making no sense are there)
> # We also get rid of the Id column which is not useful
> usefuldata <- myData3[,-1]
> # So, the correct answers for min, max, mean, median and standard deviation are:
> attach(usefuldata)
> min(Education)
[1] 5
> max(Education)
[1] 22
> mean(Education)
[1] 14.25553
> median(Education)
[1] 14
> sd(Education)
[1] 2.914282
> min(Wage)
[1] 2047.2
> max(Wage)
[1] 8453.5
> mean(Wage)
[1] 5479.368
> median(Wage)
[1] 5520.5
> sd(Wage)
[1] 1200.513
> min(Gender)
[1] 1
> max(Gender)
[1] 2
> # Mean, median and sd of Gender make no sense.
> # The information that we may get from Gender is the percentage
> # of data relative to men and women in the table
> men <- Gender == "1"
> dataMen <- usefuldata[men,]
> d <- dim(dataMen)
> d
[1] 299   3
```

```
> # Then the percentage of men in the dataset is
> d[1]/497*100
[1] 60.16097
> # Then the percentage of women in the dataset is
> 100 - d[1]/497*100
[1] 39.83903
> # QUESTION 2
> # What do you think/infer from all these variables when
> # the main focus is to predict the value of Wage?
>
> # Wage can be explained by Education and Gender…
> # Do you expect positive or negative relationship between Education and
> # Wage? Why?
>
> # QUESTION 3
> # Select the wage and education values corresponding to Male (dataMen)
>
> # We remove from dataMen the information about the Gender, which is redundant now
> dMen <- dataMen[,-2]
> summary(dMen)
   Education          Wage
 Min.   : 5.00   Min.   :2047
 1st Qu.:12.00   1st Qu.:4913
 Median :14.00   Median :5730
 Mean   :14.26   Mean   :5730
 3rd Qu.:16.00   3rd Qu.:6534
 Max.   :21.00   Max.   :8454
> sd(dMen$Education)
[1] 2.941061
> sd(dMen$Wage)
[1] 1168.759
> # Select the wage and education values corresponding to Female
> # Gender 2=female
> women <- Gender == "2"
> dataWomen <- usefuldata[women,]
> dWomen <- dataWomen[,-2] # We remove Gender column
> d <- dim(dWomen)
> d
[1] 198   2
> # The percentage of women in the dataset is
> d[1]/497*100
[1] 39.83903
> summary(dWomen)
   Education          Wage
 Min.   : 8.00   Min.   :2640
 1st Qu.:12.00   1st Qu.:4242
 Median :14.00   Median :4967
 Mean   :14.25   Mean   :5100  
 3rd Qu.:16.00   3rd Qu.:5904  
 Max.   :22.00   Max.   :8329  
> sd(dWomen$Education)
[1] 2.880773
> sd(dWomen$Wage)
[1] 1149.944
> 
> 
```