Exercise 1

1 . For each of the variable, compute the mean, the median, the standard deviation, the minimum and maximum value.

Answer Let's first load the data:

```r
varTypes <- c('numeric','numeric','factor','numeric','character')
mydata = read.table('~/Documents/Education.txt',sep="\t",colClasses =varTypes,header=T)
```

Then we run the function *summary* on our data which contains many different function on itself and might help us to detect anomalies on our data:

```r
summary(mydata)
```

```
##        ID          Education     Gender       Wage
##  Min.   :  1.0   Min.   :-12.0   1 :299   Min.   :  41.8
##  1st Qu.:125.8   1st Qu.: 12.0   2 :200   1st Qu.:4693.0
##  Median :250.5   Median : 14.0   20:  1   Median :5510.4
##  Mean   :250.5   Mean   : 14.2            Mean   :5465.3
##  3rd Qu.:375.2   3rd Qu.: 16.0            3rd Qu.:6322.3
##  Max.   :500.0   Max.   : 22.0            Max.   :8453.5
##    Country
##  Length:500
##  Class :character
##  Mode  :character
##
##
##
```

Immediately *we detect some anomalies on our data*, like **for example**. there is a person with -12 years education and we know that there shouldn't be negative values on *Education* attribute, also in the *Gender* attribute (1=male,2=female) there is a person with gender 20 etc.

So, we need to clean the data a bit before we apply our functions required by the exercise:

```r
cleanedData<-droplevels(mydata[which(mydata$Gender==1 | mydata$Gender==2),])
cleanedData<-cleanedData[which(cleanedData$Education>0 & cleanedData$Wage>=1000),]
summary(cleanedData)
```

```
##        ID          Education     Gender      Wage        Country
##  Min.   :  1.0   Min.   : 5.00   1:299   Min.   :2047   Length:497
##  1st Qu.:126.0   1st Qu.:12.00   2:198   1st Qu.:4696   Class :character
##  Median :251.0   Median :14.00           Median :5520   Mode  :character
##  Mean   :250.5   Mean   :14.26           Mean   :5479
##  3rd Qu.:375.0   3rd Qu.:16.00           3rd Qu.:6332
##  Max.   :500.0   Max.   :22.00           Max.   :8454
```

Now, we procede to compute mean,median,max,min and standard deviation for variables, we see that most of them are already given by the summary function but anyway we will do it explicitly, we also see that for some variables it doesn't make sense to compute some of these functions, like for ID, Gender, Country.

Education variable:

```
mean(cleanedData$Education)
```

```
## [1] 14.25553
```

```
median(cleanedData$Education)
```

```
## [1] 14
```

```
sd(cleanedData$Education)
```

```
## [1] 2.914282
```

```
min(cleanedData$Education)
```

```
## [1] 5
```

```
max(cleanedData$Education)
```

```
## [1] 22
```

Wage variable:

```
mean(cleanedData$Wage)
```

```
## [1] 5479.368
```

```
median(cleanedData$Wage)
```

```
## [1] 5520.5
```

```
sd(cleanedData$Wage)
```

```
## [1] 1200.513
```

```
min(cleanedData$Wage)
```

```
## [1] 2047.2
```

```
max(cleanedData$Wage)
```

```
## [1] 8453.5
```

Gender variable: I will compute the mean for Gender just in case if we want to know the male/female ratio, because it can be helpful to analysis.

```
mean(as.numeric(cleanedData$Gender))
```

```
## [1] 1.39839
```

2 . Select the variables according to the underlying country. For each country, compute the mean, the median, the standard deviation, the minimum and maximum value. Do you see some differences between the countries?

Answer First, we split our data based on country:

```
us_data = cleanedData[which(cleanedData$Country=='US'),]
canada_data = cleanedData[which(cleanedData$Country=='Canada'),]
```

Now, to calculate the required functions I will just use the *summary()* function of R instead of doing it explicitly for each variable(and just add the mean for Gender variable and standard deviation for Wage and Education variables):

US-summary:

```
summary(us_data)
```

```
##        ID           Education      Gender       Wage
##  Min.   :  1.00   Min.   : 8.00   1:180   Min.   :2938
##  1st Qu.: 75.25   1st Qu.:12.00   2:118   1st Qu.:4586
##  Median :150.50   Median :14.00           Median :5447
##  Mean   :150.37   Mean   :14.19           Mean   :5459
##  3rd Qu.:224.75   3rd Qu.:16.00           3rd Qu.:6329
##  Max.   :300.00   Max.   :21.00           Max.   :8110
##    Country
##  Length:298
##  Class :character
##  Mode  :character
##
##
##
```

Canada-summary:

```
summary(canada_data)
```

```
##        ID          Education      Gender       Wage        Country
##  Min.   :301.0   Min.   : 5.00   1:119   Min.   :2047   Length:199
##  1st Qu.:350.5   1st Qu.:12.00   2: 80   1st Qu.:4794   Class :character
##  Median :400.0   Median :14.00           Median :5639   Mode  :character
##  Mean   :400.3   Mean   :14.35           Mean   :5509
##  3rd Qu.:450.5   3rd Qu.:17.00           3rd Qu.:6331
##  Max.   :500.0   Max.   :22.00           Max.   :8454
```

Gender variable(mean):

Canada male/female ratio:

```
mean(as.numeric(canada_data$Gender))
```

```
## [1] 1.40201
```

US male/female ratio:

```
mean(as.numeric(us_data$Gender))
```

```
## [1] 1.395973
```

Standard deviation(US,Canada):

```
sd(canada_data$Education)
```

```
## [1] 3.106995
```

```
sd(us_data$Education)
```

```
## [1] 2.781672
```

```
sd(canada_data$Wage)
```

```
## [1] 1247.675
```

```
sd(us_data$Wage)
```

```
## [1] 1169.657
```

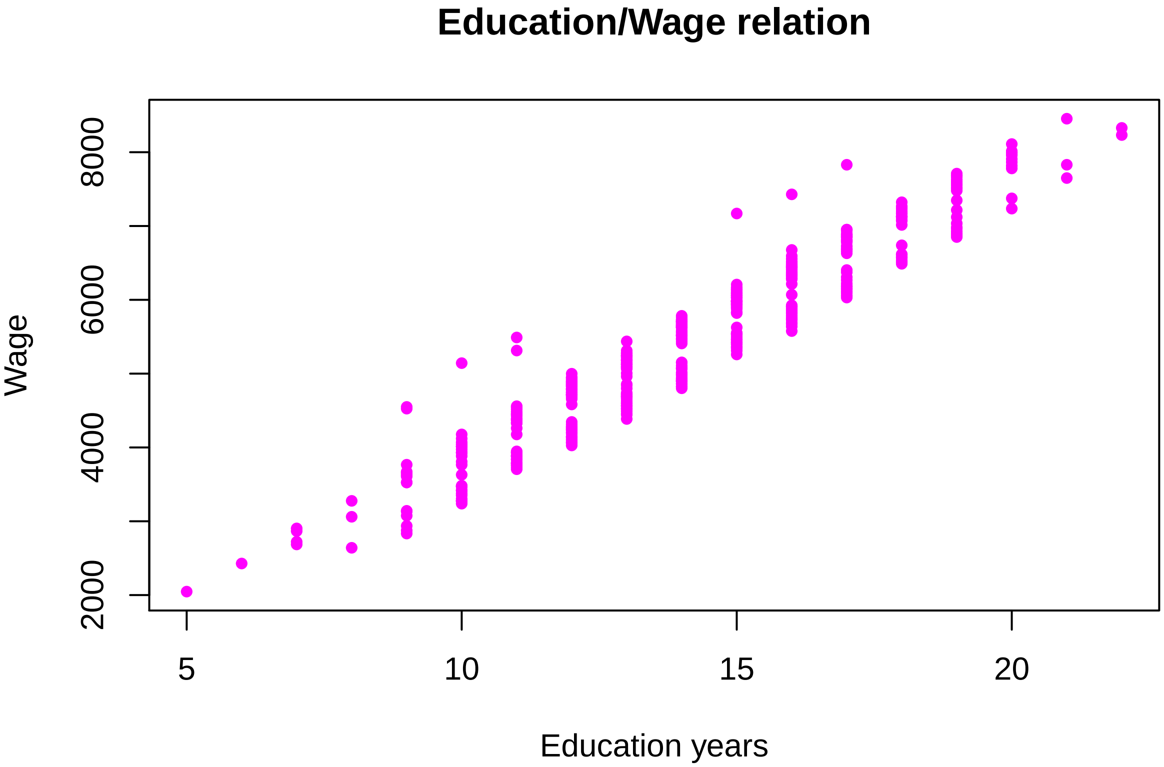
Yes, we can see that in our dataset there are more people from US, and that the ratio male/female in both countries is almost the same, also we can see that the **mean of wages** is actually a bit higher in Canada but also the **mean of education years** is also higher in Canada. Also from the standard deviation we can infer that there is more variability on Education data for people that are from Canada compared to that of US.

If we want to choose the **"best" number**(person) to represent our data, than we should choose it in relation to some *error metric*, such that to minimize the error as much as possible, if we choose the *absolute error metric(L1 error)* then the best number to choose is the **median**. Similarly, the **mean** minimizes squared error(that's why it's more affected by outliers than the median).

3 . (Ignore the difference between the countries). What do you think / infer from all these variables when the main focus is to predict the value of the variable Wage?

Answer We can see that there is some direct relation between the **Education** and **Wage** variables, possibly also the **Gender** but for now let's disscuss only the relation between the first two.

```
plot(cleanedData$Education,cleanedData$Wage,xlab="Education years",
     main="Education/Wage relation",ylab="Wage", pch=20, col='magenta')
```



So, the relation between years of education and wage is that *the higher the education is, the higher the wage* of the person will be(in general case).

4 . As the variable Gender is a categorical (binary data), select the wage and education values corresponding to each of the two possible gender values. Compute the mean, the median, the standard deviation, the minimum and maximum value for each gender separately. What can you infer from these values?

Answer Let's first select the Wage and Education attributes by genders:

```
male_education <- cleanedData$Education[which(cleanedData$Gender==1)]
male_wage <- cleanedData$Wage[which(cleanedData$Gender==1)]
female_education<- cleanedData$Education[which(cleanedData$Gender==2)]
female_wage <- cleanedData$Wage[which(cleanedData$Gender==2)]
```

Male data summary:

```
summary(male_education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   12.00   14.00   14.26   16.00   21.00
```

```
summary(female_education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   12.00   14.00   14.25   16.00   22.00
```

```
summary(male_wage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2047    4913    5730    5730    6534    8454
```

```
summary(female_wage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2640    4242    4967    5100    5904    8329
```

Standard deviation(male,female):

```
sd(male_education)
```

```
## [1] 2.941061
```

```
sd(male_wage)
```

```
## [1] 1168.759
```

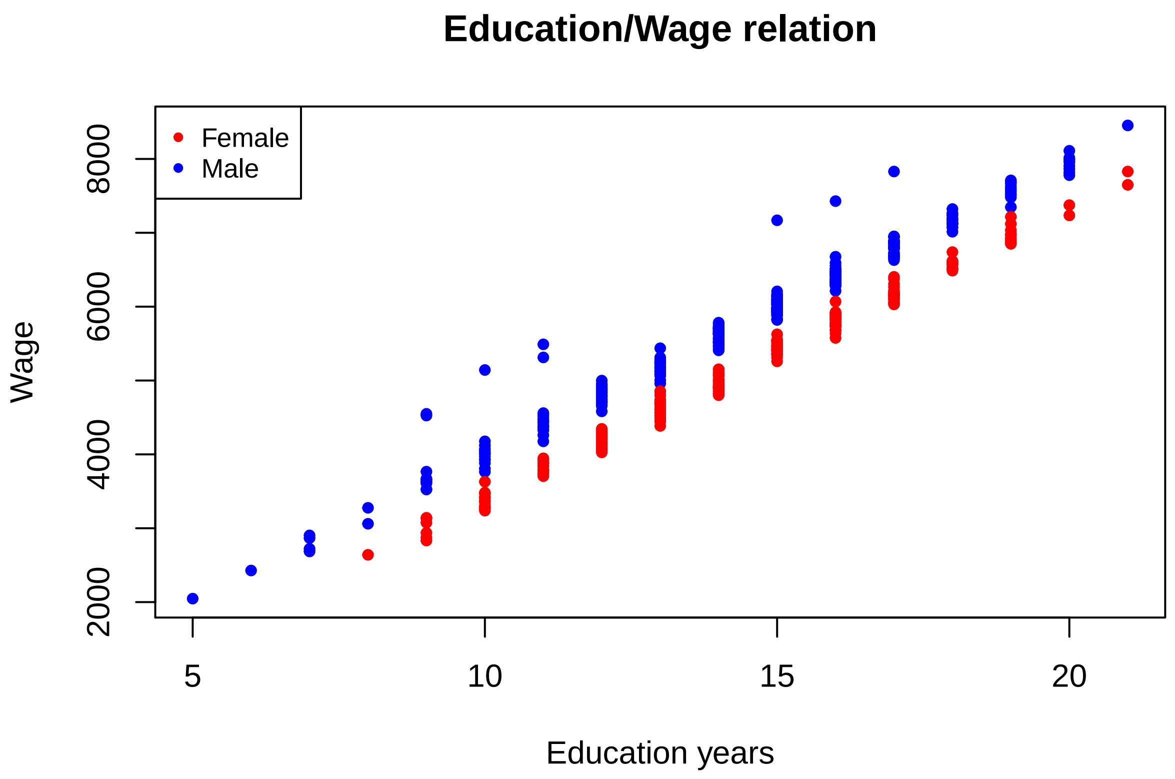
```
sd(female_education)
```

```
## [1] 2.880773
```

```
sd(female_wage)
```

```
## [1] 1149.944
```

```
plot(male_education,male_wage,xlab="Education years",ylab="Wage", main="Education/Wage relation", pch=2
points(female_education,female_wage,col='red1',pch=20)
legend("topleft", legend=c("Female", "Male"),
       col=c("red", "blue"), lty=points(1,2), cex = 0.8, pch=20)
```



We can see that there exists *a pay gap between genders*, male always get paid more for the same education. That we can see from the summary tables(mean and median are higher), but also in the graph we can clearly see the difference in wages for the same education years.