Exercise 2

Question 1. Download the car seats data set from the ILIAS website (Carseats.txt). Using the plot() function, show graphically the possible relationship between the dependent variable Sales and the independent variable Price. Can you change the title of the plot as well as the labels appearing in the two axes? It seems that there is an outlier, can you highlight it in red?

Answer First we will load the data(we did not select immediately the columns of Sales and Price, because as we will see there are some anomalies on other columns and in that way we would not be able to detect them at all):

```
pathToFile<- "~/Desktop/S2017/StatisticalLearning/Exercises/2/Carseats.txt"
mydata <- read.table(pathToFile, header=T)
```

We look at the summary of our data:

```
summary(mydata);
```

```
##      Sales          CompPrice       Income         Advertising
##  Min.   : 0.000   Min.   : 77   Min.   :  21.00   Min.   : 0.000
##  1st Qu.: 5.415   1st Qu.:115   1st Qu.:  42.75   1st Qu.: 0.000
##  Median : 7.495   Median :125   Median :  69.00   Median : 5.000
##  Mean   : 7.543   Mean   :125   Mean   :  92.67   Mean   : 6.635
##  3rd Qu.: 9.322   3rd Qu.:135   3rd Qu.:  91.00   3rd Qu.:12.000
##  Max.   :20.700   Max.   :175   Max.   :9700.00   Max.   :29.000
##    Population        Price        ShelveLoc        Age
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :-64
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.: 39
##  Median :272.0   Median :117.0   Medium:219   Median : 54
##  Mean   :264.8   Mean   :115.8                Mean   : 53
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.: 66
##  Max.   :509.0   Max.   :191.0                Max.   : 80
##    Education      Urban       US
##  Min.   :-15.00   No :118   No :142
##  1st Qu.: 12.00   Yes:282   Yes:258
##  Median : 14.00
##  Mean   : 13.82
##  3rd Qu.: 16.00
##  Max.   : 18.00
```

Then we clean the data a bit because there are some data with negative values in Education and Age attributes:

```
mydata <- mydata[-which(mydata$Education<0),]
mydata <- mydata[-which(mydata$Age<0),]
mydata <- mydata[,c("Sales", "Price")]
```

Now we can show graphically the relationship between Sales variable and Price variable:

```
outlier <- mydata[which(mydata$Sales>20),]
plot(mydata$Sales, mydata$Price,main = "Sales / Price relationship",
     xlab = "Sales", ylab = "Price")
points(outlier$Sales, outlier$Price, pch = 20, col = "red")
```

Question 2. Download the education data set from the ILIAS website (Education.txt). As the variable Country is a factor (categorical), select the wage and education values corresponding to each of the two possible country values. Show graphically the possible relationship between the independent variable Education and the dependent variable Wage but only when considering the observations corresponding to the USA and to the Canada. Can you plot on the same graphics the information for the US and for the Canada but with different colors? Can you add a legend with the corresponding colors?

Answer We load the data and then select only the columns that we need from the table, which are Education, Wage and Country:

```
pathToEdu<- "~/Desktop/S2017/StatisticalLearning/Exercises/2/Education.txt"
varTypes <- c('numeric','numeric','factor','numeric','factor')
eduData = read.table(pathToEdu, sep="\t", colClasses = varTypes, header=T)
eduData = eduData[,c('Education','Wage','Country')]
summary(eduData)
```

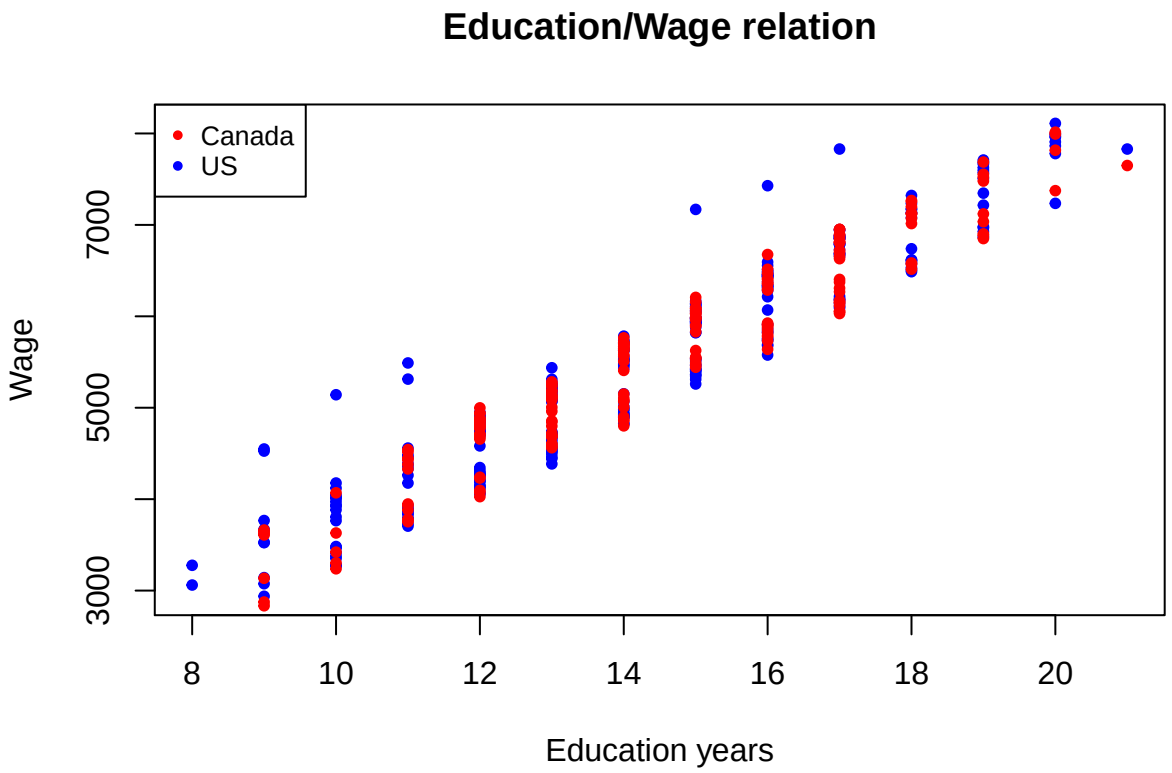
```
##    Education          Wage        Country
##  Min.   : 5.00   Min.   :2047   Canada:199
##  1st Qu.:12.00   1st Qu.:4696   US    :298
##  Median :14.00   Median :5520
##  Mean   :14.26   Mean   :5479
##  3rd Qu.:16.00   3rd Qu.:6332
##  Max.   :22.00   Max.   :8454
```

Then we split the data based on corresponding Country:

```
us_data = eduData[which(eduData$Country=='US'),]
ca_data = eduData[which(eduData$Country=='Canada'),]
```

We plot the data using the plot() and points() functions and we add a legend with the corresponding colors for each country:

```r
plot(us_data$Education,us_data$Wage,xlab="Education years",ylab="Wage",
     main="Education/Wage relation", pch=20, col="blue")
points(ca_data$Education,ca_data$Wage,col='red1',pch=20)
legend("topleft", legend=c("Canada", "US"), col=c("red", "blue"),
       lty=points(1,2), cex = 0.8, pch=20)
```



Question 3. Download from the ILIAS website the Mean20 data set (filename: Mean20.txt). This data set is composed by a single variable (time), the time delay in minutes between two calls in an info-center. Compute the mean, the median, the standard deviation, the minimum and maximum value of the variable time. Is there any outliers and possible invalid samples? Do you need to preprocess this list?

Answer Yes we need to preprocess, because after we load the data we can see that there are some NA values on the data, and some negative values on time delay, which don't make sense:

```r
pathToFile <- "~/Desktop/S2017/StatisticalLearning/Exercises/2/Mean20.txt"
timeDelay <- read.table(pathToFile, header=T, colClasses = c('numeric'))
summary(timeDelay)
```

```
##       time
##  Min.   :-7.01
##  1st Qu.: 6.96
##  Median : 7.01
##  Mean   : 6.34
##  3rd Qu.: 7.07
##  Max.   : 7.12
##  NA's   :1
```

We remove the NA values from our data, and also we remove data that contain negative time delay:

```
ctimeDelay <- na.omit(timeDelay)
ctimeDelay <- subset(ctimeDelay, ctimeDelay$time>0)
summary(ctimeDelay)
```

```
##       time
##  Min.   :6.850
##  1st Qu.:6.968
##  Median :7.010
##  Mean   :7.008
##  3rd Qu.:7.072
##  Max.   :7.120
```

```
sd(ctimeDelay$time)
```

```
## [1] 0.07515598
```

Question 4. We suppose that the mean delay between two calls is 7.05 minutes. Can you test this hypothesis with the data available? What is your conclusion? Do you see a difference when considering the original values and the preprocessed values?

Answer We already have the preprocessed data needed to test this hypothesis:

```
ttest <- t.test(ctimeDelay, alternative = 'two.sided', mu=7.05, conf.int = 0.95)
ttest
```

```
##
##  One Sample t-test
##
## data:  ctimeDelay
## t = -2.4992, df = 19, p-value = 0.02178
## alternative hypothesis: true mean is not equal to 7.05
## 95 percent confidence interval:
##  6.972826 7.043174
## sample estimates:
## mean of x
##     7.008
```

From the *p-value* (which is $<< 0.05$) we can conclude that there is strong evidence against null hypothesis H_0, *so we can reject it.* Our null hypothesis is that the mean delay is 7.05 minutes, for which we say that it's incorrect.

Let's consider the *t-test* on the original values and compare the result with that of preprocessed data:

```
ttest_unclean <- t.test(timeDelay, alternative = 'two.sided', mu=7.05, conf.int = 0.95)
ttest_unclean
```

```
##
##  One Sample t-test
##
## data:  timeDelay
## t = -1.0626, df = 20, p-value = 0.3006
## alternative hypothesis: true mean is not equal to 7.05
## 95 percent confidence interval:
##  4.947647 7.733306
## sample estimates:
## mean of x
##  6.340476
```

When we consider the t-test on the original values then we have a completely different result, now *we fail to reject the null hypothesis* $H_0 : \mu = 7.05$ (p-value=0.3006>0.05).

Question 5. For John, the average delay is greater than 7.05. Thus the only credible alternative hypothesis must take account of this fact. How can you test John's hypothesis?

Answer We can test John's hypothesis by doing a one-sided t-test on our preprocessed data with alternative option set to 'greater':

```
ttest <- t.test(ctimeDelay, alternative = 'greater', mu=7.05, conf.int = 0.95)
ttest
```

```
##
##  One Sample t-test
##
## data:  ctimeDelay
## t = -2.4992, df = 19, p-value = 0.9891
## alternative hypothesis: true mean is greater than 7.05
## 95 percent confidence interval:
##  6.978941      Inf
## sample estimates:
## mean of x
##     7.008
```

In the case of a one-sided alternative, the sign of the t-statistic matters A LOT. A negative sign implies that the sample mean is less than the hypothesized mean. This would be evidence against the null hypothesis IF (and only if) the alternative was that the true mean is LESS than the hypothesized value.

In this case, the **t-statistic is negative(t=-2.4992)**, that is NOT evidence against the null in favor of the alternative. So, *we fail to reject the null hypothesis* $H_0 : \mu = 7.05$(p-value=0.9891>>0.05 and sign of t-statistic is negative), so John's hypothesis that the average delay is greater than 7.05 is wrong.