RAZAVET Mael 24/01/2014

Data Mining – Exercise 3

Question 1:

```
par(mfrow=c(2,2))
hist(iris$Sepal.Length)
hist (iris$Sepal.Width)
hist (iris$Petal.Length)
hist (iris$Petal.Width)
```

The histograms show that the Sepal.Width attribute has a normal distribution

```
par(mfrow=c(1,4))
boxplot(iris$Sepal.Length, xlab="Sepal.Length")
boxplot(iris$Sepal.Width, xlab="Sepal.Width")
boxplot(iris$Petal.Length, xlab="Petal.Length")
boxplot(iris$Petal.Width, xlab="Petal.Width")
```

According to the boxplots, only the Sepal.Width attribute has 4 outliers. Indeed, one data point has 2.0 as a value and is outside the whiskers of the boxplot. Same for other values which are above the upper whisker (4.0). We can also observe that the data is very condensed between the lower and upper quartiles.

Concerning the Sepal.Length attribute, we can notice that the values are quite equally spread the median divides the box into 2 equally parts. And the data is a little bit more spread equally as the IQR is greater.

Concerning the Sepal.Length and Sepal.Width attribute, we can say that the IQR is even greater than the Sepal.Length, which means that the data even more spread equally.

```
#red is for setosa / green3 is for versicolor / blue is for virginica
pairs(iris[1:4], main = "Anderson's Iris Data -- 3 species", pch=21, bg = c("red", "green3",
"blue")[unclass(iris$Species)])
```

Concerning the matrix of scatter plots generated, we can notice that the instances with species labeled with “versicolor” and “virgina”, are within the same cluster, which indicates that these two classes have similiarities. Whereas, the class labeled “setosa” constitute a different cluster. Furthermore, the attributes Petal.Width and Petal.Length displayed in a scatter plot with any other variables, we can easily distinguish the three clusters, unlike for the other attributes. And finally, the Petal.Length and Petal.Width attributes seem to be correlated according to the positive skewness of the scatter plot. Besides, the Sepal.Length and Petal.Length seems to be correlated but less than the former two attributes, as well as the Sepal.Length with the Petal.Width attributes.

Question 2:

```
irisMissing = read.csv("irismissing.csv")
sum(is.na(irisMissing)) returns 16
```

Question 3:

```
irisDrop <- na.omit(irisMissing)
```

We could replace by the mean of the feature, or its median or its mode.

Question 4:

To replace by the mean value of the attribute:

```
replaceMissingValuesByMean=function(x){
 for(attribute in 1:length(names(x))){
  if(is.numeric(x[,attribute]))
  {
   z <- mean(x[,attribute], na.rm = TRUE)
   x[is.na(x[,attribute]), attribute] <- z
  }
 }
 return(x)
}
```

To replace by the median value of the attribute:

```
replaceMissingValuesByMedian=function(x){
 for(attribute in 1:length(names(x))){
  if(is.numeric(x[,attribute]))
  {
   z <- median(x[,attribute], na.rm = TRUE)
   x[is.na(x[,attribute]), attribute] <- z
  }
 }
 return(x)
}
```

To compute the mode value of the attribute:

```
getMode=function(x){
  z <- table(x)
  as.numeric(names(z)[z == max(z)])
}
replaceMissingValuesByMode=function(x){
 for(attribute in 1:length(names(x))){
  if(is.numeric(x[,attribute]))
  {
   z <- getMode(x[,attribute])
   x[is.na(x[,attribute]), attribute] <- z
  }
 }
 return(x)
}
```

Question 5:

```
foo=function(x, FUNC){
        d <- FUNC(x)
        return(d)
}
```

Then I call this function like this:

```
irisDropMean = foo(irisMissing, replaceMissingValuesByMean)
```

Question 6:

Histograms:
As we can see, if we replace the missing values by the mean, the histogram of the Sepal.Width attribute shows that the frequency of values around 3.0 (around the mean) increased.
Concerning the histogram displaying the data, which replaced the missing values by the median, shows the frequency of the value 3.0 increased as well. Previously, the frequency of this value was 35, now it is 50. We can have the same interpretation for the histogram displaying the data handling the missing values by replacing them by the mode of the attribute Sepal.Width, which is equalt to 3.0 like the median.

Boxplots:
In the Sepal.Width attribute, we have now 5 outliers (compared to 4 previously). The reason is that as we replaced the missing values by either the mean, the median or the mode value of this attribute, the IRQ decreased because we have more instances having these respective values. This leads to a less spread box between the lower and upper quartiles. Moreover, we can notice that the upper quartile has a lower value now, which is normal as the frequency of the values around the mean, median or mode has increased a lot due to replacing the missing values by these values. Then, the median line has only changed when we replace the missing values by the mean (mean = 3.061194). As the mean is greater than the previous median (3.0), this is normal to increase the median of the new data of this attribute.
N.B.: This can be a disadvantage compared to using the linear regression method for example, to replace missing value, which I chose not to implement here.