REPORT OF DATA MINING

Lab2

He NI

BI2

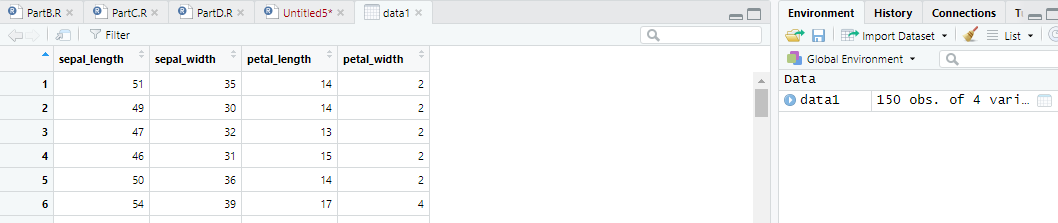
20150742

# A Multivariate data set : Fisher Iris

**In this exercice, we study the Iris data set.**

1. **Open the file “iris.csv” with a regular text editor to see what the data look like (how many rows, how many attributes, etc). Then, use the R command read.csv(· · ·) with the right parameters so that you can open this data set in R as a data matrix.**

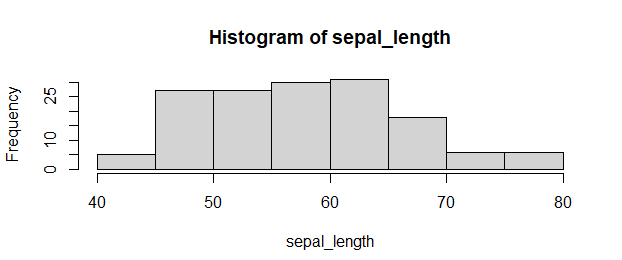
**data1=read.csv("C:\\Users\\Administrator\\Desktop\\Data Mining\\Lab2\\iris.csv")**



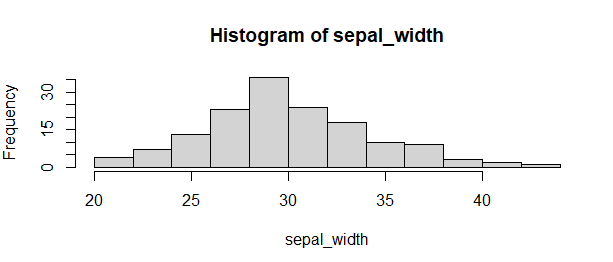
1. **Display the histograms of the different attributes. What can you say about their distributions ?**

**attach(data1)**

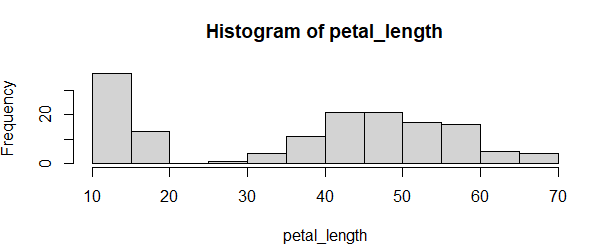
**hist(sepal\_length)**



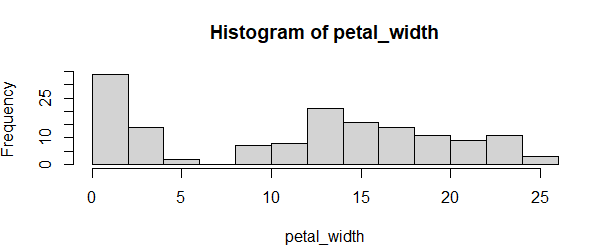
**hist(sepal\_width)**



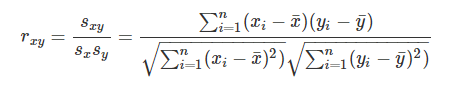
**hist(petal\_length)**



**hist(petal\_width)**



1. **Compute the coefficient of correlation between all attributes without using the dedicated R function.**



**cov(sepal\_length,sepal\_width)/(sd(sepal\_length)\*sd(sepal\_width))**

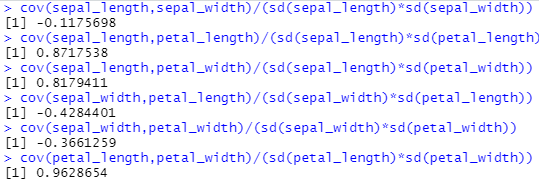
**cov(sepal\_length,petal\_length)/(sd(sepal\_length)\*sd(petal\_length))**

**cov(sepal\_length,petal\_width)/(sd(sepal\_length)\*sd(petal\_width))**

**cov(sepal\_width,petal\_length)/(sd(sepal\_width)\*sd(petal\_length))**

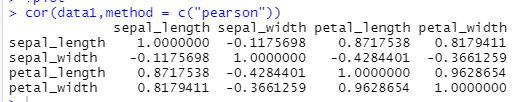
**cov(sepal\_width,petal\_width)/(sd(sepal\_width)\*sd(petal\_width))**

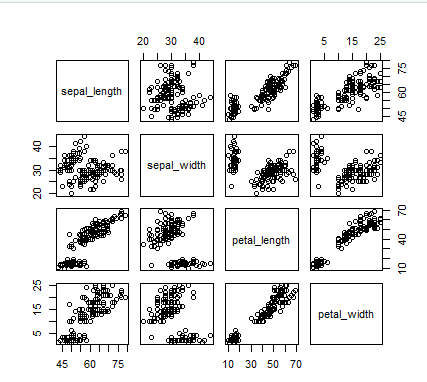
**cov(petal\_length,petal\_width)/(sd(petal\_length)\*sd(petal\_width))**



1. **Use the commands cor(data) and plot(data) to confirm your previous results and visualize the correlation between the different variables. Comment your results.**

**cor(data1,method = c("pearson"))**





1. **Compute the confidence intervals for the correlation coefficients (we will suppose that the attributes are following a normal distribution). Comment your results.**

**cor.test(sepal\_length,sepal\_width)**



**cor.test(sepal\_length,petal\_length)**



**cor.test(sepal\_length,petal\_width)**



**cor.test(sepal\_width,petal\_length)**



**cor.test(sepal\_width,petal\_width)**



**cor.test(petal\_length, petal\_width)**

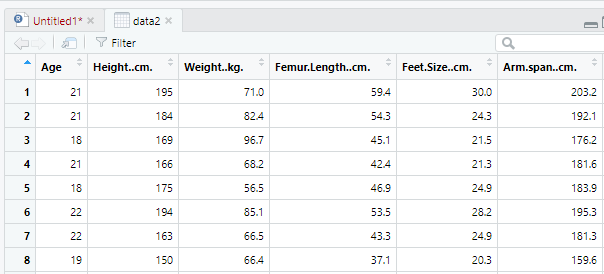


# B Multivariate data set : Anthropometric data

**In this exercice, we study the ”mansize” data set. These data described anthropometric features acquired in a famous medicine University based on a population of Bachelor students.**

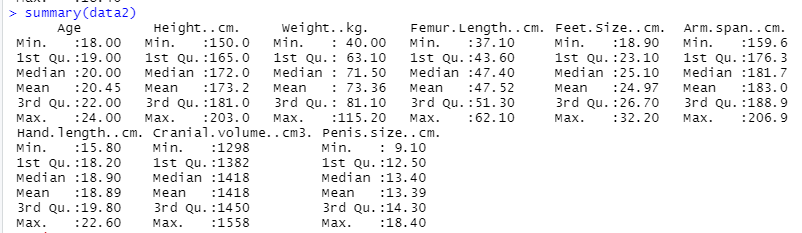
1. **Open the file “mansize.csv” with a regular text editor to see what the data look like (how many rows, how many attributes, etc). Then, use the R command read.csv(· · ·) with the right parameters so that you can open this data set in R as a data matrix.**

**data2=read.csv("C:\\Users\\Administrator\\Desktop\\DataMining\\Lab2\\mansize.csv",sep = ";")**

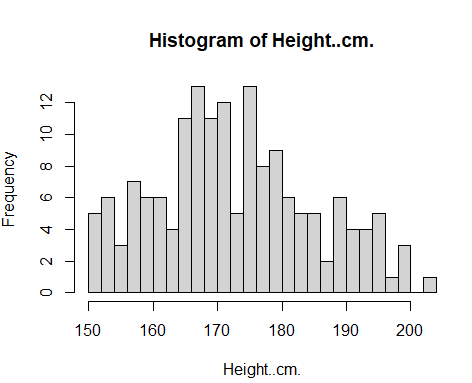
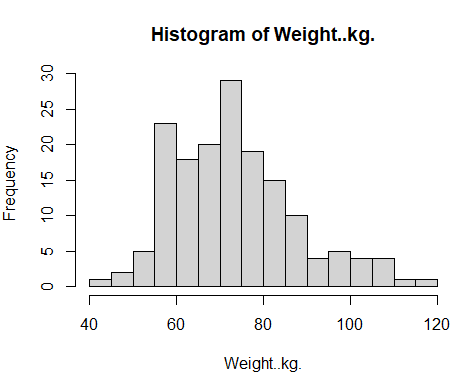
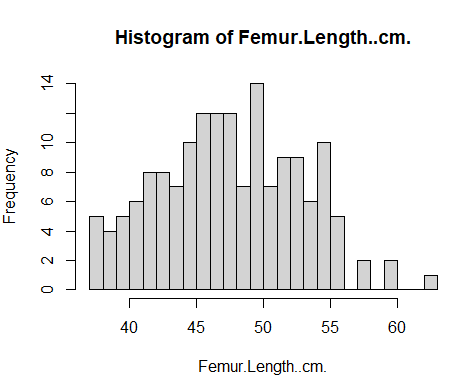
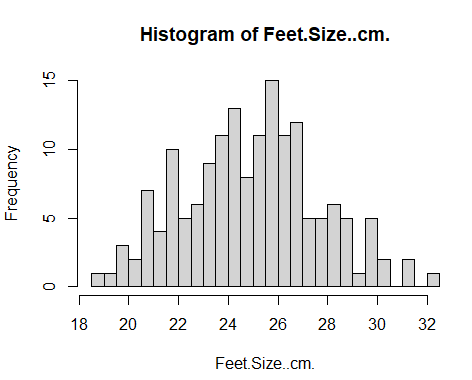
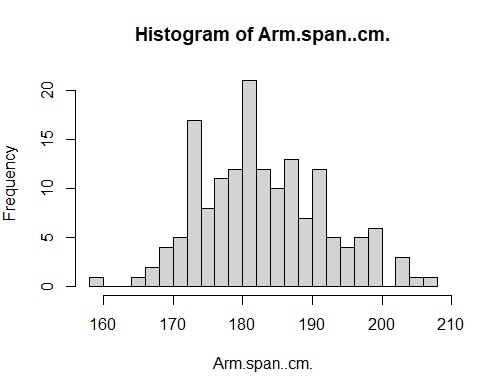
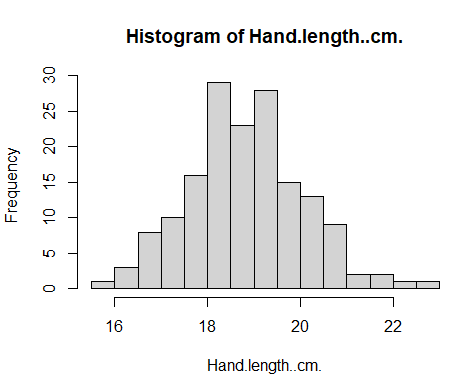
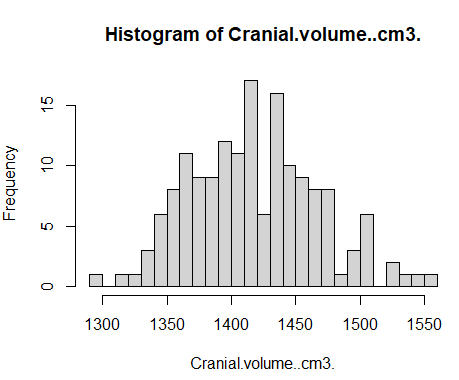
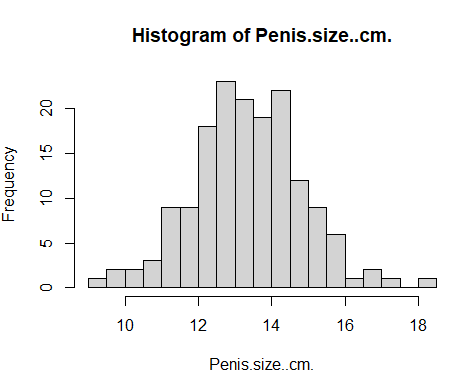


1. **Apply the function summary(·) to your data set. What does this function do ? Comment the results on your data.**

**summary is a generic function used to produce result summaries of the results of various model fitting functions. The function invokes particular methods which depend on the class of the first argument.**

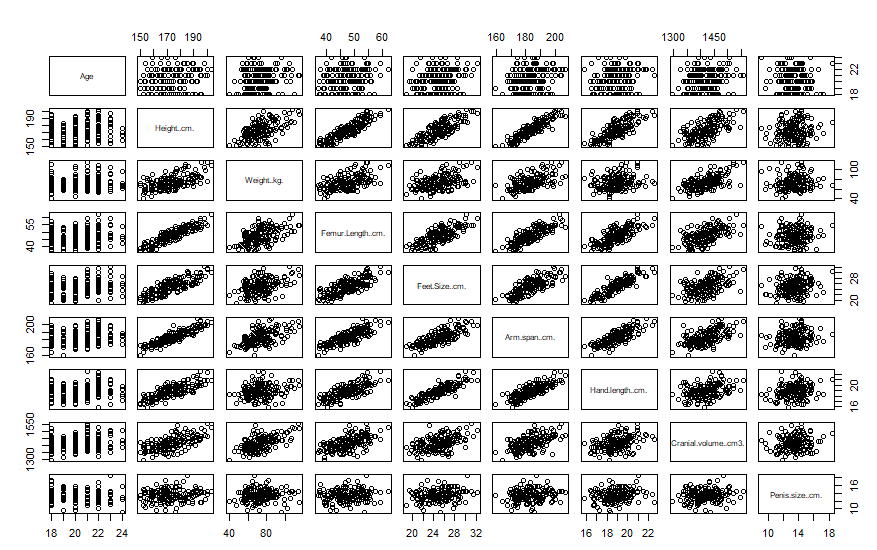
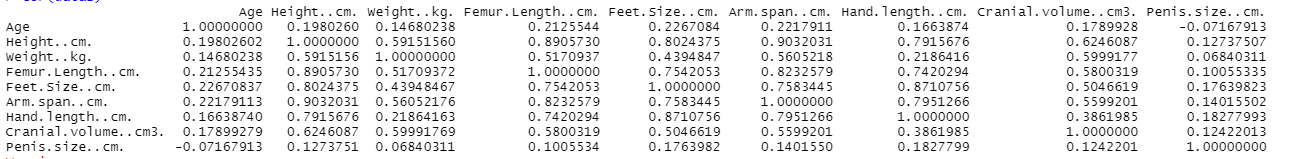


1. **Display the histograms of the different attributes. What can you say about their distributions ?**

**With the exception of Age, all other attributes are approximately normally distributed.**

1. **Use the commands cor(data) and plot(data) to visualize the correlation between the different variables. Comment your results. In particular, what can you say about the use in archaeology of the femur length to predict the height of an individual ?**



**The higher the correlation, the denser the scatter is, as can be seen from the plot scatter and the correlation cor.**

**So if we want to predict a person's height, then the two most relevant are Femur length and Arm span, which are 0.89 and 0.90 respectively.**

**So using these two characteristics to predict height will have a high success rate!**

1. **Compute the confidence intervals for the correlation coefficients (we will suppose that the attributes are following a normal distribution). Comment your results.**

*# Calculating Pearson's correlation coefficient*

**r = cor(data2)**

*#Calculating conversion values*

*The correlation coefficient between -1 and 1 is mapped to the whole number of real numbers.*

**Ztrans = function(r) 1/2\*log((1+r)/(1-r))**

**zr = Ztrans(r)**

*#Calculation of confidence intervals for conversion values*

**n = nrow(data2)**

**zr.sd = 1/sqrt(n-3)**

**leftzr = zr-1.96\*zr.sd**

**rightzr = zr+1.96\*zr.sd**

*#Reversing the confidence intervals of the transformation values into confidence intervals of the correlation coefficients*

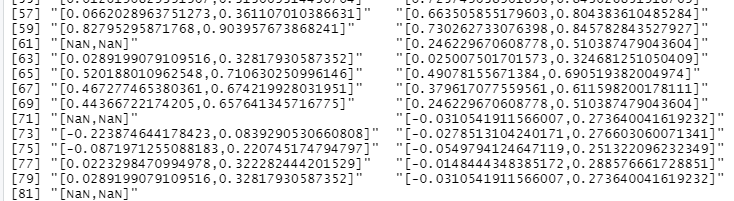
**revZ = function(z)((exp(2\*z)-1)/(exp(2\*z)+1))**

**lzr = revZ(leftzr)**

**rzr = revZ(rightzr)**

**msg = paste("123 [" ,lzr, "," ,rzr, "]", sep="")**

**print(msg)**



**Judging significance**

**If the lzr and rzr are both greater than 0, then lzr and rzr are significantly positively correlated.**

**If the If the lzr and rzr are both smaller than 0, then lzr and rzr are significantly negatively**

**correlated**

**The others situation is no correlated.**

1. **Based on the results of the previous questions as well as your analysis of the correlation and determination coefficients between the data, conclude on the links between the different variables in this dataset.**

**Conclusion: I think that the higher the correlation, the less likely the feature is to be true, so we calculated confidence intervals for the correlation coefficients, the narrower the confidence interval, the more likely it is that the correlation coefficients can be determined, and the more true they are.**

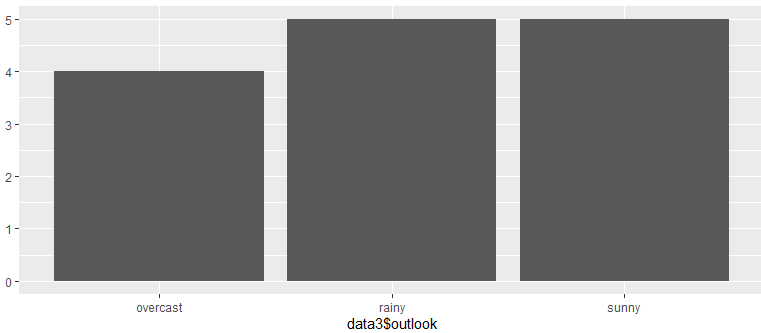
**So, we used features with a high correlation (correlation coefficient greater than 0.8) to make our predictions!**

# C Chi-squared test of independence and categorial variables

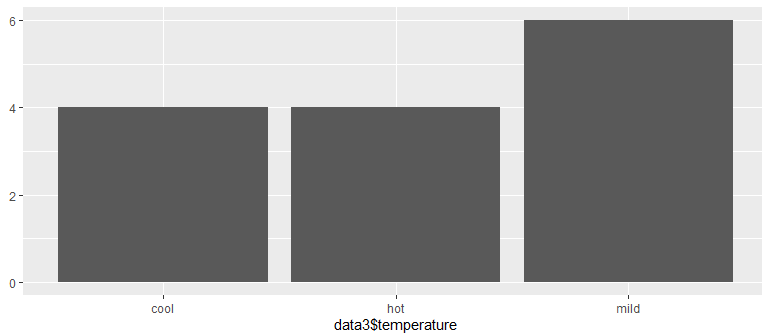
**In this exercice, we want to assess whether there is a link between different meteorological variables measured in different cities.**

1. **Open the “weather.csv” data set and describe the different variables and their values using histograms.**

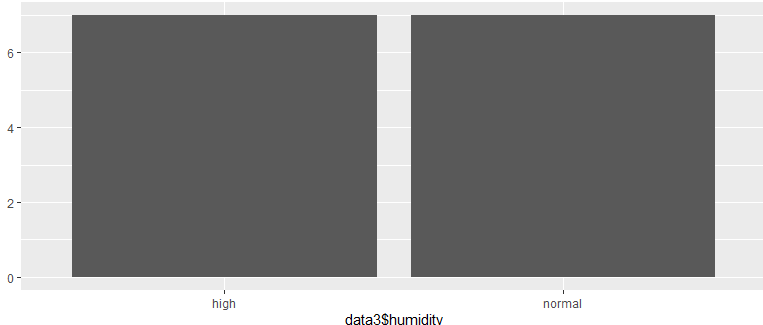
**ggplot2::qplot(data3$outlook)**



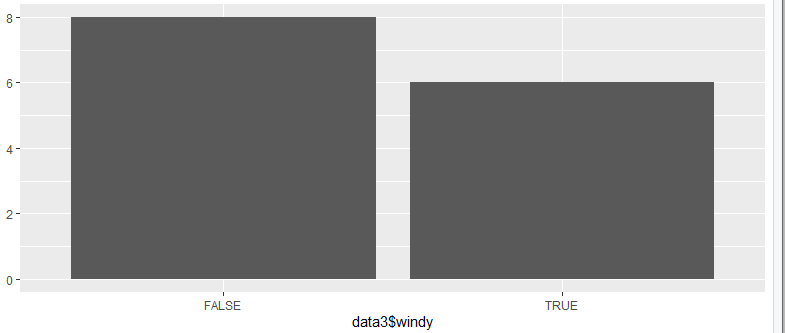
**ggplot2::qplot(data3$temperature)**



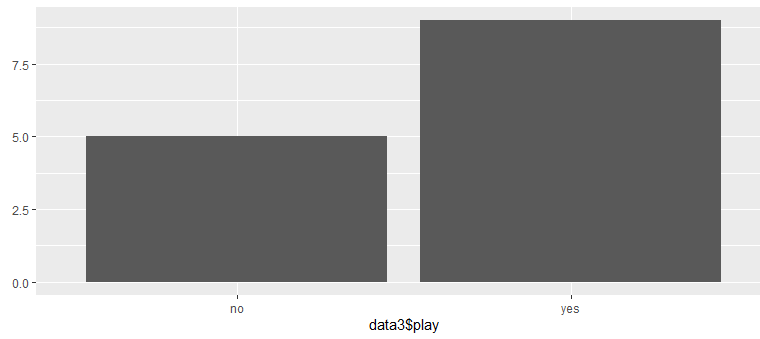
**ggplot2::qplot(data3$humidity)**



**ggplot2::qplot(data3$windy)**

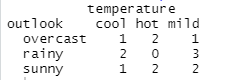


**ggplot2::qplot(data3$play)**



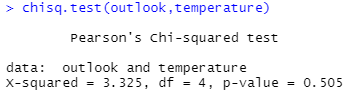
1. **Use the command table(·) to create the contingency table between the variables “outlook” and “temperature”. Comment the repartition of the variables in the resulting table. How many degrees of freedom do we have in this problem ?**

**table(outlook,temperature)**



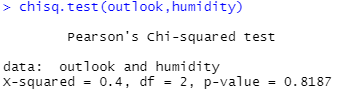
**degrees of freedom: (n-1)\* (c-1) = 4**

1. **Use the command chisq.test(·) on your table. From the result and if need be by computed other indexes, what can you conclude on the dependency between these two variables ?**



1. **Based on the methodology you used in the previous questions, assess whether there is a link between the other variables of your data set (outlook/humidity, temperature/humidity).**

**chisq.test(outlook,humidity)**

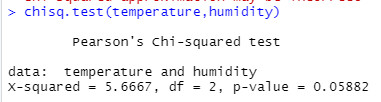


**X2(4, n=6) = 0.4,p>0.05**

**Because the p-value is more than 0.05**

**So there is no link between outlook and humidity**

**chisq.test(temperature,humidity)**



**X2(4, n=6) = 0.4, p>0.05882**

**Because the p-value is more than 0.05**

**So there is no link between temperature and humidity**