

# REPORT OF DATA MINING

Lab2

He NI

BI2

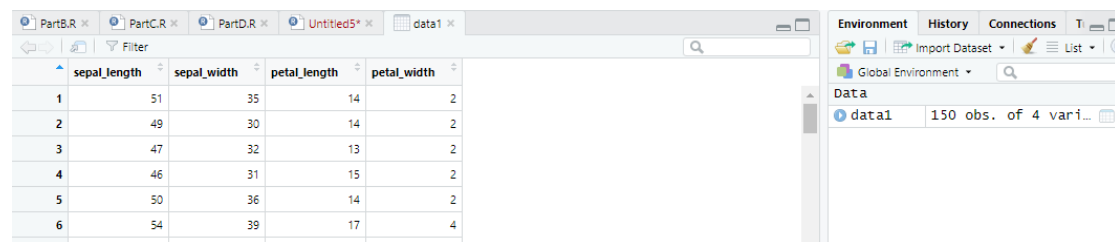
20150742

# A Multivariate data set : Fisher Iris

In this exercise, 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")
```

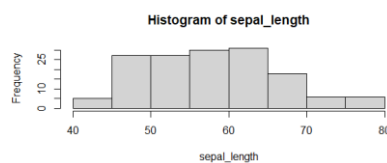


	sepal_length	sepal_width	petal_length	petal_width
1	51	35	14	2
2	49	30	14	2
3	47	32	13	2
4	46	31	15	2
5	50	36	14	2
6	54	39	17	4

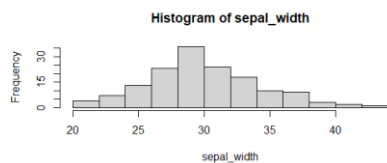
2. Display the histograms of the different attributes. What can you say about their distributions ?

```
attach(data1)
```

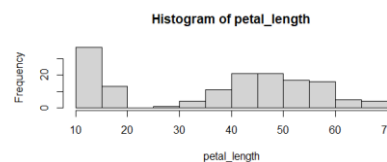
```
hist(sepal_length)
```



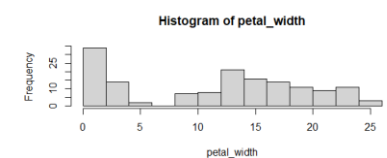
```
hist(sepal_width)
```



```
hist(petal_length)
```



```
hist(petal_width)
```



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

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

```

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))
> cov(sepal_length,sepal_width)/(sd(sepal_length)*sd(sepal_width))
[1] -0.1175698
> cov(sepal_length,petal_length)/(sd(sepal_length)*sd(petal_length))
[1] 0.8717538
> cov(sepal_length,petal_width)/(sd(sepal_length)*sd(petal_width))
[1] 0.8179411
> cov(sepal_width,petal_length)/(sd(sepal_width)*sd(petal_length))
[1] -0.4284401
> cov(sepal_width,petal_width)/(sd(sepal_width)*sd(petal_width))
[1] -0.3661259
> cov(petal_length,petal_width)/(sd(petal_length)*sd(petal_width))
[1] 0.9628654

```

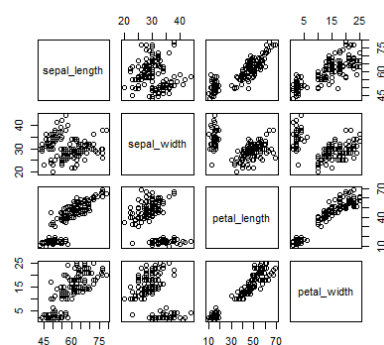
4. Use the commands `cor(data)` and `plot(data)` to confirm your previous results and visualize the correlation between the different variables. Comment your results.

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

```

> cor(data1,method = c("pearson"))
      sepal_length sepal_width petal_length petal_width
sepal_length  1.0000000 -0.1175698  0.8717538  0.8179411
sepal_width  -0.1175698  1.0000000 -0.4284401 -0.3661259
petal_length  0.8717538 -0.4284401  1.0000000  0.9628654
petal_width   0.8179411 -0.3661259  0.9628654  1.0000000

```



5. 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)
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.27269325  0.04351158

cor.test(sepal_length,petal_length)
95 percent confidence interval:
 0.8270363 0.9055080

cor.test(sepal_length,petal_width)
95 percent confidence interval:
 0.7568971 0.8648361

cor.test(sepal_width,petal_length)
95 percent confidence interval:
 -0.5508771 -0.2879499

cor.test(sepal_width,petal_width)
95 percent confidence interval:
 -0.4972130 -0.2186966

cor.test(petal_length, petal_width)
95 percent confidence interval:
 0.9490525 0.9729853
```

## B Multivariate data set : Anthropometric data

In this exercise, 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 = ";")
```

	Age	Height..cm.	Weight..kg.	Femur.Length..cm.	Feet.Size..cm.	Arm.span..cm.
1	21	195	71.0	59.4	30.0	203.2
2	21	184	82.4	54.3	24.3	192.1
3	18	169	96.7	45.1	21.5	176.2
4	21	166	68.2	42.4	21.3	181.6
5	18	175	56.5	46.9	24.9	183.9
6	22	194	85.1	53.5	28.2	195.3
7	22	163	66.5	43.3	24.9	181.3
8	19	150	66.4	37.1	20.3	159.6

2. 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.

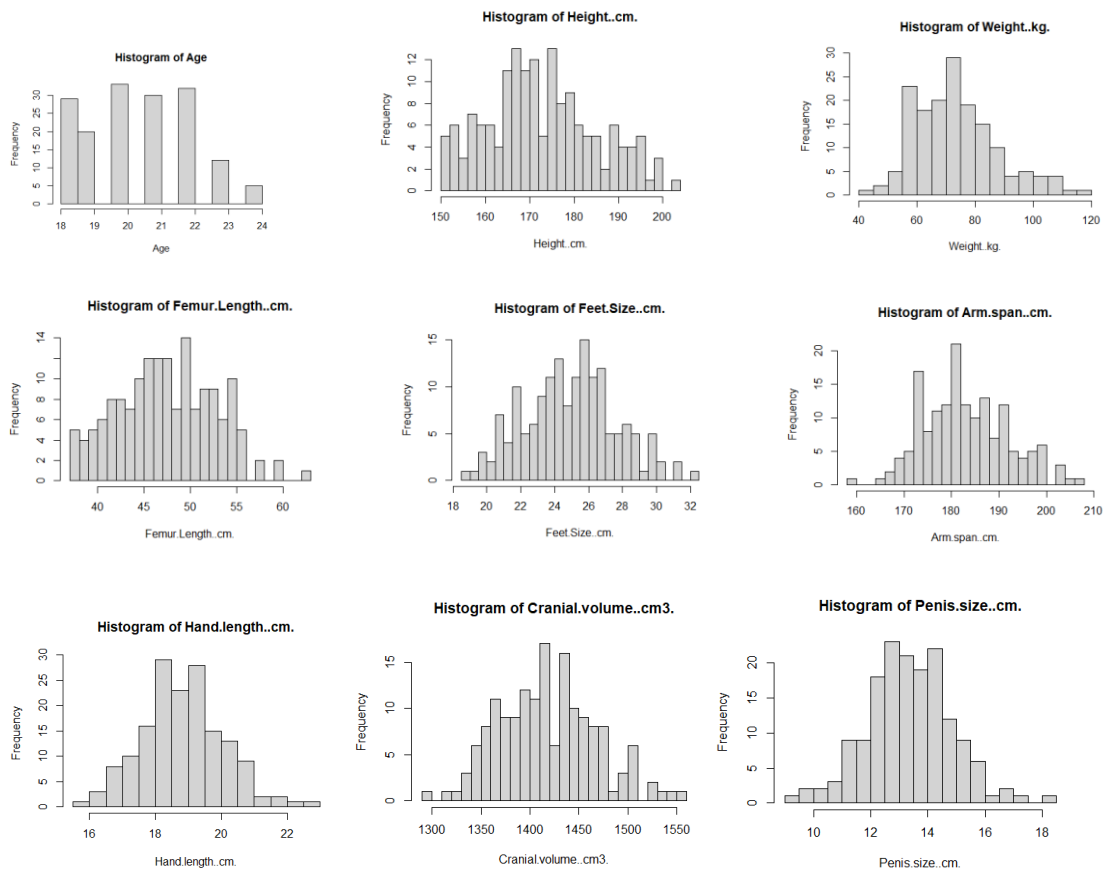
```
> summary(data2)
```

Age	Height..cm.	Weight..kg.	Femur.Length..cm.	Feet.Size..cm.	Arm.span..cm.
Min. :18.00	Min. :150.0	Min. : 40.00	Min. :37.10	Min. :18.90	Min. :159.6
1st Qu.:19.00	1st Qu.:165.0	1st Qu.: 63.10	1st Qu.:43.60	1st Qu.:23.10	1st Qu.:176.3
Median :20.00	Median :172.0	Median : 71.50	Median :47.40	Median :25.10	Median :181.7
Mean :20.45	Mean :173.2	Mean : 73.36	Mean :47.52	Mean :24.97	Mean :183.0
3rd Qu.:22.00	3rd Qu.:181.0	3rd Qu.: 81.10	3rd Qu.:51.30	3rd Qu.:26.70	3rd Qu.:188.9
Max. :24.00	Max. :203.0	Max. :115.20	Max. :62.10	Max. :32.20	Max. :206.9

Hand.length..cm.	Cranial.volume..cm3.	Penis.size..cm.
Min. :15.80	Min. :1298	Min. : 9.10
1st Qu.:18.20	1st Qu.:1382	1st Qu.:12.50
Median :18.90	Median :1418	Median :13.40
Mean :18.89	Mean :1418	Mean :13.39
3rd Qu.:19.80	3rd Qu.:1450	3rd Qu.:14.30
Max. :22.60	Max. :1558	Max. :18.40

3. 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.

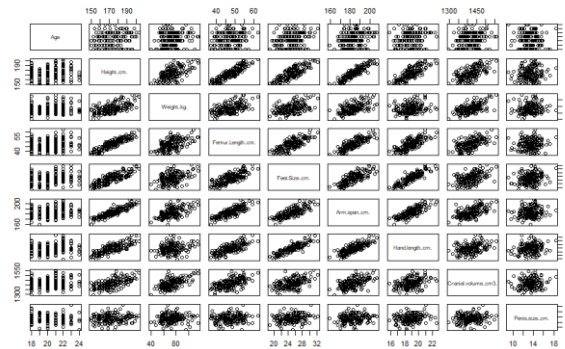
- Use the commands `cor(data)` and `plot(data)` to visualize the correlation between the different variables. Comment your results. In particular, what can you say about the use in archaeology of the femur length to predict the height of an individual ?

```

# Correlation matrix
cor(data)

```

	Age	Height..cm.	weight..kg.	Femur.Length..cm.	Feet.Size..cm.	Arm.span..cm.	Hand.length..cm.	Cranial.volume..cm3.	Penis.size..cm.
Age	1.0000000	0.1980260	0.14680238	0.2125544	0.2267084	0.2217911	0.1663874	0.1789928	-0.07167913
Height..cm.	0.19802602	1.0000000	0.5915156	0.8905730	0.8024375	0.9032031	0.7915676	0.6246087	0.12737507
weight..kg.	0.14680238	0.5915156	1.0000000	0.5170937	0.4394847	0.5605218	0.2186416	0.5999177	0.06840311
Femur.Length..cm.	0.21255435	0.8905730	0.51709372	1.0000000	0.7542053	0.8232579	0.7420294	0.5800319	0.10055335
Feet.Size..cm.	0.22670837	0.8024375	0.43948467	0.7542053	1.0000000	0.7583445	0.8710756	0.5046619	0.17639823
Arm.span..cm.	0.22179113	0.9032031	0.56052176	0.8232579	0.7583445	1.0000000	0.7951266	0.5599201	0.14015502
Hand.length..cm.	0.16638740	0.7915676	0.21864163	0.7420294	0.8710756	0.7951266	1.0000000	0.3861985	0.18277993
Cranial.volume..cm3.	0.17899279	0.6246087	0.59991769	0.5800319	0.5046619	0.5599201	0.3861985	1.0000000	0.12422013
Penis.size..cm.	-0.07167913	0.1273751	0.06840311	0.1005534	0.1763982	0.1401550	0.1827799	0.1242201	1.0000000



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!

- 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))

```

```

lzc = revZ(leftzr)

```

```

rzc = revZ(rightzr)

```

```

msg = paste("123 [" ,lzc , " , " ,rzc , "]", sep="")

```

```

print(msg)

```

```
[57] "[0.0662028963751273,0.361107010386631]" "[0.663505855179603,0.804383610485284]"
[59] "[0.82795295871768,0.903957673868241]" "[0.730262733076398,0.845782843527927]"
[61] "[NaN,NaN]" "[0.246229670608778,0.510387479043604]"
[63] "[0.0289199079109516,0.32817930587352]" "[0.025007501701573,0.324681251050409]"
[65] "[0.520188010962548,0.710630250996146]" "[0.49078155671384,0.690519382004974]"
[67] "[0.467277465380361,0.674219928031951]" "[0.379617077559561,0.611598200178111]"
[69] "[0.44366722174205,0.657641345716775]" "[0.246229670608778,0.510387479043604]"
[71] "[NaN,NaN]" "[-0.0310541911566007,0.273640041619232]"
[73] "[-0.223874644178423,0.0839290530660808]" "[-0.0278513104240171,0.276603060071341]"
[75] "[-0.0871971255088183,0.220745174794797]" "[-0.0549794124647119,0.251322096232349]"
[77] "[0.0223298470994978,0.322282444201529]" "[-0.0148444348385172,0.288576661728851]"
[79] "[0.0289199079109516,0.32817930587352]" "[-0.0310541911566007,0.273640041619232]"
[81] "[NaN,NaN]"
```

### Judging significance

If the lwr and rwr are both greater than 0, then lwr and rwr are significantly positively correlated.

If the lwr and rwr are both smaller than 0, then lwr and rwr are significantly negatively correlated

The others situation is no correlated.

- 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.

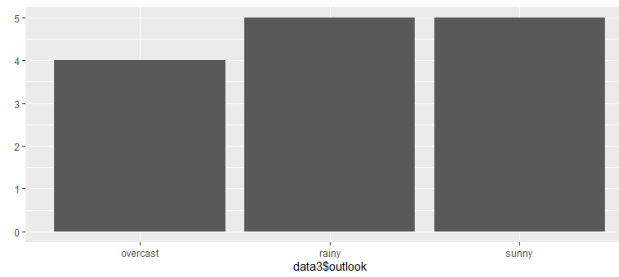
So, we used features with a high correlation (correlation coefficient greater than 0.8) to make our predictions!

## C Chi-squared test of independence and categorical variables

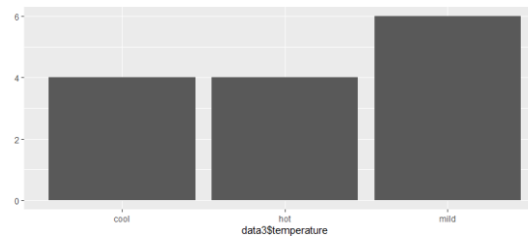
In this exercise, we want to assess whether there is a link between different meteorological variables measured in different cities.

- Open the "weather.csv" data set and describe the different variables and their values using histograms.

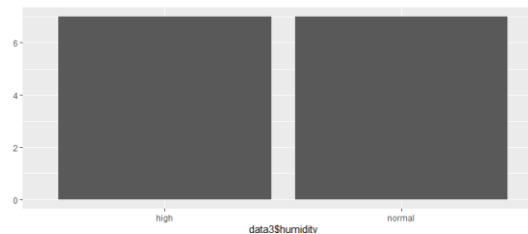
```
ggplot2::qplot(data3$outlook)
```



`ggplot2::qplot(data3$temperature)`



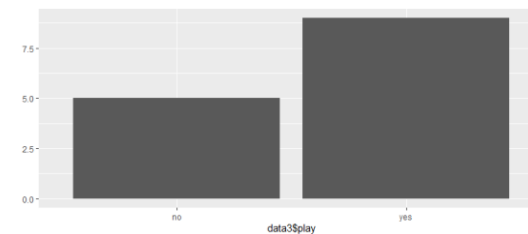
`ggplot2::qplot(data3$humidity)`



`ggplot2::qplot(data3$windy)`



`ggplot2::qplot(data3$play)`



- 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)`

outlook	temperature		
	cool	hot	mild
overcast	1	2	1
rainy	2	0	3
sunny	1	2	2

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



3. 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 ?

```
> chisq.test(outlook,temperature)

Pearson's Chi-squared test

data:  outlook and temperature
X-squared = 3.325, df = 4, p-value = 0.505
```

4. 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)

> chisq.test(outlook,humidity)

Pearson's Chi-squared test

data:  outlook and humidity
X-squared = 0.4, df = 2, p-value = 0.8187
```

$X^2(4, n=6) = 0.4, p > 0.05$

Because the p-value is more than 0.05

So there is no link between outlook and humidity

```
chisq.test(temperature,humidity)

> chisq.test(temperature,humidity)

Pearson's Chi-squared test

data:  temperature and humidity
X-squared = 5.6667, df = 2, p-value = 0.05882
```

$X^2(4, n=6) = 0.4, p > 0.05882$

Because the p-value is more than 0.05

So there is no link between temperature and humidity