



# مقدمة مبسطة في لغة R مع تطبيقات إحصائية

## Introduction to R Language with Statistical Applications

سبأ محمد علوان

# Summaries and Descriptive Statistics



mean, sd, var, min, max, median, range, quantile, summary

```
> x<-c(1,3,5,4,7,8,9,-4,-5,10)
> x
 [1]  1  3  5  4  7  8  9 -4 -5 10
> mean(x)
 [1] 3.8
> median(x)
 [1] 4.5
> max(x)
 [1] 10
> min(x)
 [1] -5
> range(x)
 [1] -5 10
> var(x)
 [1] 26.84444
> sd(x)
 [1] 5.181162
> quantile(x)
  0%   25%   50%   75%  100%
-5.00  1.50  4.50  7.75 10.00
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-5.00   1.50   4.50   3.80   7.75  10.00
```

# The apply function for loops



## Example

```
> x<-c(1,3,5,4,7,8,9,-4,-5,10)
> y<-c(2,3,5,6,8,7,7,9,-6,11)
> z<-matrix(c(x,y),nc=2)
> z
```

	[,1]	[,2]
[1,]	1	2
[2,]	3	3
[3,]	5	5
[4,]	4	6
[5,]	7	8
[6,]	8	7
[7,]	9	7
[8,]	-4	9
[9,]	-5	-6
[10,]	10	11

```
> apply(z,2,summary)
      [,1] [,2]
Min.   -5.00 -6.00
1st Qu.  1.50  3.50
Median  4.50  6.50
Mean    3.80  5.20
3rd Qu.  7.75  7.75
Max.   10.00 11.00
```



## The apply function for loops

The effect of calculating the mean of each column (dimension 2) of trees.

If we have used a 1 instead of a 2 , the mean will be calculated for every row.

Example: for the same data

```
> apply(z,2,summary)
```

```
      [,1]  [,2]  
Min.   -5.00 -6.00  
1st Qu.  1.50  3.50  
Median   4.50  6.50  
Mean     3.80  5.20  
3rd Qu.  7.75  7.75  
Max.    10.00 11.00
```

```
> apply(z,1,summary)
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]  
Min.   1.00   3    5  4.0  7.00  7.00  7.0 -4.00 -6.00 10.00  
1st Qu. 1.25   3    5  4.5  7.25  7.25  7.5 -0.75 -5.75 10.25  
Median 1.50   3    5  5.0  7.50  7.50  8.0  2.50 -5.50 10.50  
Mean   1.50   3    5  5.0  7.50  7.50  8.0  2.50 -5.50 10.50  
3rd Qu. 1.75   3    5  5.5  7.75  7.75  8.5  5.75 -5.25 10.75  
Max.   2.00   3    5  6.0  8.00  8.00  9.0  9.00 -5.00 11.00
```

```
> |
```

# Summaries and Descriptive Statistics



## Example

```
> x<-c(7.5,8.2,3.1,5.6,8.2,9.3,6.5,7.0,9.3,1.2,14.5,6.2)
```

```
> summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.200  6.050   7.250   7.217  8.475  14.500
```

```
> summary(x[1:6])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3.100  6.075   7.850   6.983  8.200   9.300
```

```
> summary(x[7:12])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.200  6.275   6.750   7.450  8.725  14.500
```

```
> summary(x[-(1:6)])# the same with x[7:12]
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.200  6.275   6.750   7.450  8.725  14.500
```

# Exercises



1. If `x<- c(5,9,2,3,4,6,7,0,8,12,2,9)` decide what each of the following is and use **R** to check your answers:
  - (a) `x[2]`
  - (b) `x[2:4]`
  - (c) `x[c(2,3,6)]`
  - (d) `x[c(1:5,10:12)]`
  - (e) `x[-(10:12)]`
2. The data `y<-c(33,44,29,16,25,45,33,19,54,22,21,49,11,24,56)` contain sales of milk in litres for 5 days in three different shops (the first 3 values are for shops 1,2 and 3 on Monday, etc.) Produce a statistical summary of the sales for each day of the week and also for each shop.

## Exercises:

## Attach (dataset)

Trees This data set provides measurements of the girth, height and volume of timber in 31 felled black cherry trees.

In actual fact, trees is an object called a dataframe, essentially a matrix with named columns (though a dataframe, unlike a matrix, may also include non-numerical variables, such as character names).

إذا قمت بكتابة الامر `mean` لأحد أعمدة `trees` فإنه لن يتعرف في أول مرة ما لم يتم استدعاء المجموعه أولاً.

```
> mean (Height)
Error in mean(Height) : object 'Height' not found
> |
```

```
> attach(trees)
> mean(Height)
[1] 76
```

```
> summary(Volume)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
10.20  19.40   24.20   30.17  37.30   77.00
\ |
```

```
> trees
  Girth Height Volume
1    8.3     70  10.3
2    8.6     65  10.3
3    8.8     63  10.2
4   10.5     72  16.4
5   10.7     81  18.8
6   10.8     83  19.7
7   11.0     66  15.6
8   11.0     75  18.2
9   11.1     80  22.6
10  11.2     75  19.9
11  11.3     79  24.2
12  11.4     76  21.0
13  11.4     76  21.4
14  11.7     69  21.3
15  12.0     75  19.1
16  12.9     74  22.2
17  12.9     85  33.8
18  13.3     86  27.4
19  13.7     71  25.7
20  13.8     64  24.9
21  14.0     78  34.5
22  14.2     80  31.7
23  14.5     74  36.3
24  16.0     72  38.3
25  16.3     77  42.6
26  17.3     81  55.4
27  17.5     82  55.7
28  17.9     80  58.3
29  18.0     80  51.5
30  18.0     80  51.0
31  20.6     87  77.0
```

```
> mean(Height)
Error in mean(Height) : object 'Height' not found
```



```
> Height
Error: object 'Height' not found
> trees$Height
[1] 70 65 63 72 81 83 66 75 80 75 79 76 76 69 75 74 85 86 71 64 78 80 74 72 77 81
> |
```

الامر `trees$Height`.. والذي يتضمن استدعاء و طلب كتابة الصيغة نلاحظ النتيجة كما سبق

### Another example

```
> quakes
      lat   long depth mag stations
1 -20.42 181.62   562 4.8        41
2 -20.62 181.03   650 4.2        15
3 -26.00 184.10    42 5.4        43
4 -17.97 181.66   626 4.1        19
5 -20.42 181.96   649 4.0        11
6 -19.68 184.31   195 4.0        12
7 -11.70 166.10    82 4.8        43
```

Note: quakes is a data set give the locations of 1000 seismic events of MB > 4.0. The events occurred in a cube near Fiji since 1964.

```
<
996 -25.93 179.54   470 4.4        22
997 -12.28 167.06   248 4.7        35
998 -20.13 184.20   244 4.5        34
999 -17.40 187.80    40 4.5        14
1000 -21.59 170.56   165 6.0       119
```





```
> summary(depth)
Error in summary(depth) : object 'depth' not found
> |
> summary(quakes$depth)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 40.0   99.0   247.0   311.4   543.0   680.0
> |
```

## Exercises

1. Attach to the dataset `quakes` and produce a statistical summary of the variables `depth` and `mag`.
2. Attach to the dataset `mtcars` and find the mean weight and mean fuel consumption for vehicles in the dataset (type `help(mtcars)` for a description of the variables available).



## Exercise

1. Repeat the analyses of the datasets `quakes` and `mtcars` using the function `apply` to simplify the calculations.
2. If

$$y = \begin{bmatrix} 1 & 4 & 1 \\ 0 & 2 & -1 \end{bmatrix}$$

what is the result of `apply(y[,2:3], 1, mean)`? Check your answer in R.

Also: `apply(y, 2, var)`



Many of the tedious statistical computations that would once have had to have been done from statistical tables can be easily carried out in R. This can be useful for finding confidence intervals etc. Let's take as an example the Normal distribution. There are functions in R to evaluate the density function, the distribution function and the quantile function (the inverse distribution function). These functions are, respectively, **dnorm**, **pnorm** and **qnorm**. Unlike with tables, there is no need to standardize the variables first. For example, suppose  $X \gg N(3; 2)$ , then

```
> dnorm(5, 3, 2)
[1] 0.1209854
```

will calculate the density function at points (5) contained (note, dnorm will assume mean 0 and standard **deviation 1 unless these are specified**. Note also that also, the function assumes you will give the **standard deviation** rather than the variance..

$$P(x < 4), x \sim N(3, 16)$$



## Probability Distributions in R

for example, the root name for the normal distribution

**pnorm, qnorm, dnorm, rnorm**

- **p**: for "probability", the cumulative distribution function (c. d. f.).
- **q**: for "quantile", the inverse c. d. f.
- **d**: for "density", the density function (p. f. or p. d. f.).
- **r**: for "random", a random variable having the specified distribution

In fact, there's not much use for the "d" function for any *continuous* distribution



## **pnorm and qnorm**

Are the distribution function and the quantile function (the inverse distribution function respectively) for the Normal distribution).

## **pt and qt**

Similar function, but it is necessary to give the degrees of freedom rather than the mean and standard deviation.

## **dbinom and pbinom**

Similar function for binomial distribution, but it is necessary to give the sample size and prop. Of success.



**pnorm** is the R function that calculates the c. d. f .

$$F(x) = P(X \leq x)$$

```
pnorm (27.4, mean=50, sd=20)
```

```
pnorm (27.4, 50, 20)
```

What is  $P(X > 19)$  when  $X$  has the  $N(17.46, 375.67)$  distribution?

```
> pnorm(19, mean=17.46, sd=sqrt(375.67))
```

```
[1] 0.5316644
```

**qnorm** is the R function that calculates the inverse c. d. f.  $F^{-1}$  of the normal distribution

$$x = F^{-1}(p)$$

What is  $F^{-1}(0.95)$  when  $X$  has the  $N(100, 15^2)$  distribution?

```
> qnorm(0.95, mean=100, sd=15)
```

```
[1] 124.6728
```



## “p”, “q” and “r” function for continuous distribution

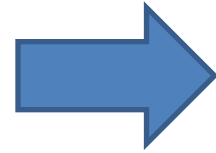
```
> pnorm(20, 10, 3)
[1] 0.9995709
> qnorm(0.2, 10, 3)
[1] 7.475136
> qnorm(0.95)
[1] 1.644854
> qnorm(0.995)
[1] 2.575829
> rnorm(15)
[1] 0.6809859 -1.4508194 -3.2667563 0.6034371 -1.2804231 0.7224153
[7] -0.3126890 1.2683494 0.0272767 -0.8733069 0.2140188 0.4437659
[13] 0.7921679 -2.7657378 -0.2212467
> rnorm(15, 2, 1)
[1] 3.2604238 1.5667078 2.5619750 1.4345865 2.2539238 4.1723859
[7] 2.2137455 1.4289406 4.0829603 2.3768709 -0.1225603 1.8184375
[13] 1.8880310 0.9062574 2.0850934
```

# The Binomial Distribution



Find:  $P(X = 27)$  when  $X$  is has the Bin(100, 0.25) distribution

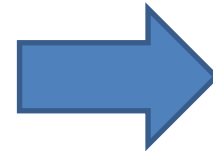
```
> dbinom(27, size=100, prob=0.25)
[1] 0.08064075
> dbinom(27, 100, 0.25)
[1] 0.08064075
```



“d”

Find:  $P(X \leq 27)$  when  $X$  is has the Bin(100, 0.25) distribution

```
> pbinom(27, size=100, prob=0.25)
[1] 0.7223805
> pbinom(27, 100, 0.25)
[1] 0.7223805
```



“p”

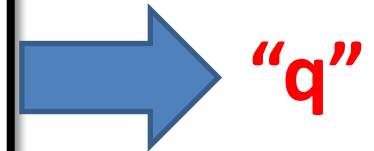




## Inverse function “q”

What are the 10th, 20th, and so forth quantiles of the Bin(10, 1/3) distribution?

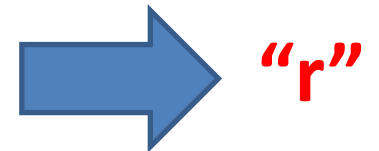
```
> qbinom(0.1, 10, 1/3)
[1] 1
> qbinom(0.2, 10, 1/3)
[1] 2
> # and so forth, or all at once with
> qbinom(seq(0.1, 0.9, 0.1), 10, 1/3)
[1] 1 2 3 3 3 4 4 5 5
```



“q”

Generate a random sample with size 20 from Bin(10, 1/3) distribution?

```
> rbinom(20, 10, 1/3)
[1] 7 3 2 3 5 3 5 0 3 4 2 4 1 5 1 6 1 0 3 4
```



“r”



Distribution	Functions			
<a href="#">Beta</a>	pbeta	qbeta	dbeta	rbeta
<a href="#">Binomial</a>	pbinom	qbinom	dbinom	rbinom
<a href="#">Cauchy</a>	pcauchy	qcauchy	dcauchy	rcauchy
<a href="#">Chi-Square</a>	pchisq	qchisq	dchisq	rchisq
<a href="#">Exponential</a>	pexp	qexp	dexp	rexp
<a href="#">F</a>	pf	qf	df	rf
<a href="#">Gamma</a>	pgamma	qgamma	dgamma	rgamma
<a href="#">Geometric</a>	pgeom	qgeom	dgeom	rgeom
<a href="#">Hypergeometric</a>	phyper	qhyper	dhyper	rhyper
<a href="#">Logistic</a>	plogis	qlogis	dlogis	rlogis
<a href="#">Log Normal</a>	plnorm	qlnorm	dlnorm	rlnorm
<a href="#">Negative Binomial</a>	pnbinom	qnbinom	dnbinom	rnbinom
<a href="#">Normal</a>	pnorm	qnorm	dnorm	rnorm
<a href="#">Poisson</a>	ppois	qpois	dpois	rpois
<a href="#">Student t</a>	pt	qt	dt	rt
<a href="#">Studentized Range</a>	ptukey	qtukey	dtukey	rtukey
<a href="#">Uniform</a>	punif	qunif	dunif	runif
<a href="#">Weibull</a>	pweibull	qweibull	dweibull	rweibull
<a href="#">Wilcoxon Rank Sum Statistic</a>	pwilcox	qwilcox	dwilcox	rwilcox
<a href="#">Wilcoxon Signed Rank Statistic</a>	psignrank	qsignrank	dsignrank	rsignrank



## Exercises

1. Suppose  $X \sim N(2, 0.25)$ . Denote by  $f$  and  $F$  the density and distribution functions of  $X$  respectively. Use **R** to calculate
  - (a)  $f(0.5)$
  - (b)  $F(2.5)$
  - (c)  $F^{-1}(0.95)$  (recall that  $F^{-1}$  is the quantile function)
  - (d)  $\Pr(1 \leq X \leq 3)$
2. Repeat question 1 in the case that  $X$  has a  $t$ -distribution with 5 degrees of freedom.
3. Use the function `rpois` to simulate 100 values from a Poisson distribution with a parameter of your own choice. Produce a statistical summary of the result and check that the mean and variance are in reasonable agreement with the true population values.
4. Repeat the previous question replacing `rpois` with `rexp`.



لغة R تعلية بسيطة بالصيغة لكنها في ذات الوقت تقدم خدمة عظيمة في إطار عرض ما لدينا من بيانات وتوضيح ما فيها من علاقات محتملة، تدعى هذه التعلية `pairs` وتقبل كدخل لها اسم إطار البيانات الذي لدينا كاملا، لتقوم بعدها برسم مصفوفة من المخططات البيانية لكل زوج ممكن من هذه البيانات على شكل مخطط مبعثر `scatter plot` بحيث يظهر كل زوج في مخططين بيانيين يتبادلان فيه مكان التمثيل على المحورين `x` و `y`، يظهر الشكل التالي مثالا عن ناتج تنفيذ هذه التعلية عند تطبيقها على إطار `trees` للبيانات السابق ذكرها .

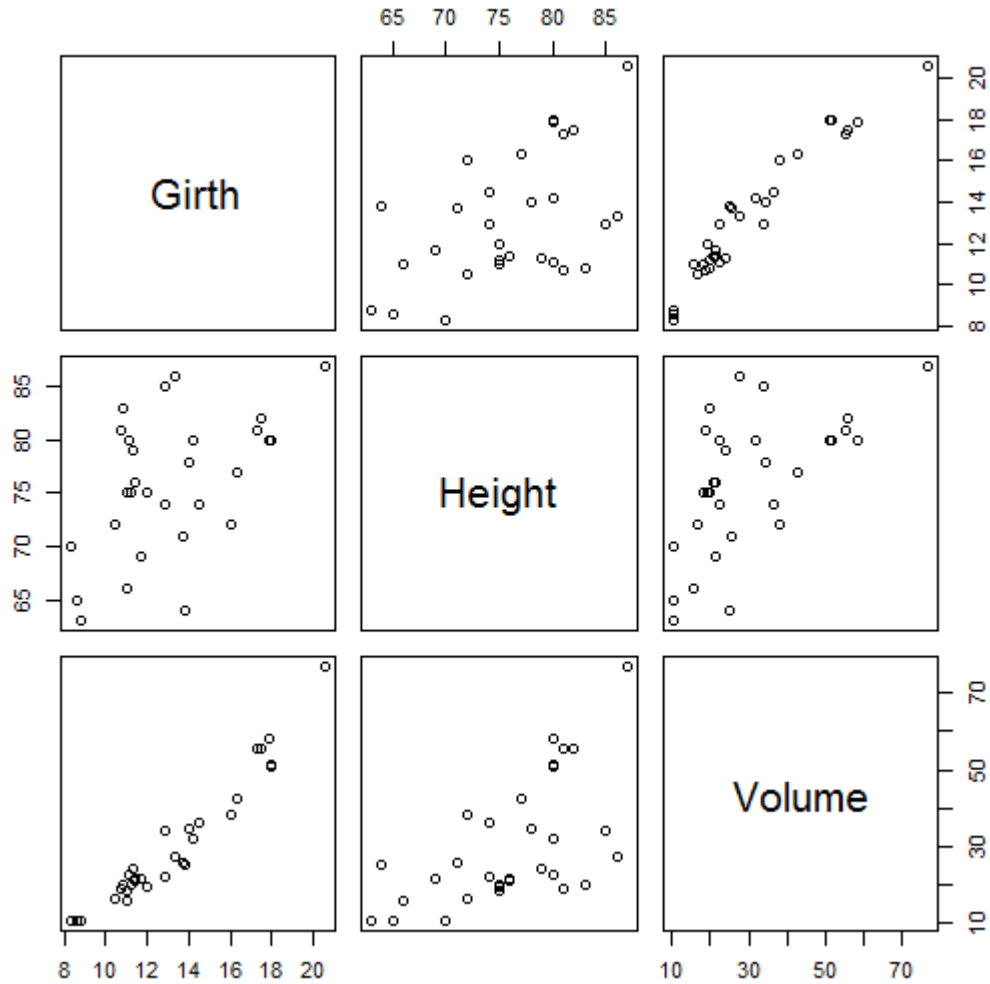
## > `pairs(trees)`

أما لرسم مخطط مبعثر `scatter` بين أي عمودين من البيانات نستطيع استخدام الدالة `plot` العامة الأغراض،  
فمثلا `plot(wt, mpg)`

## > `plot(wt, mpg)`



> pairs(trees)

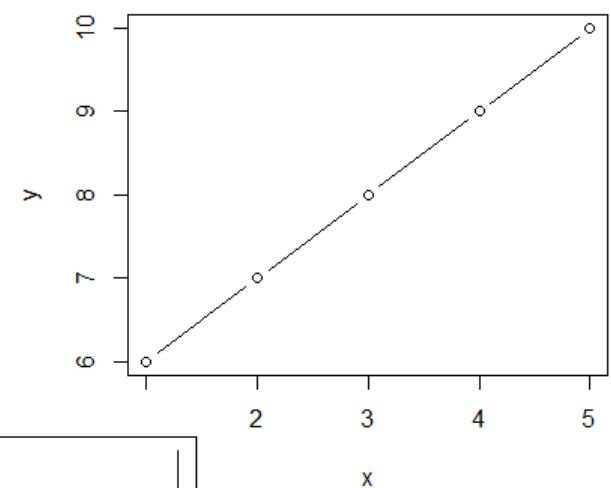
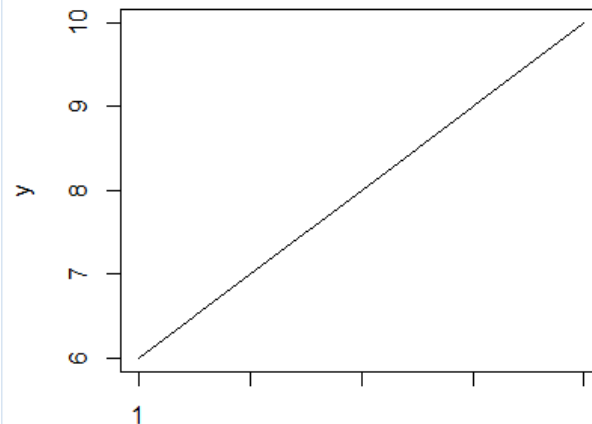
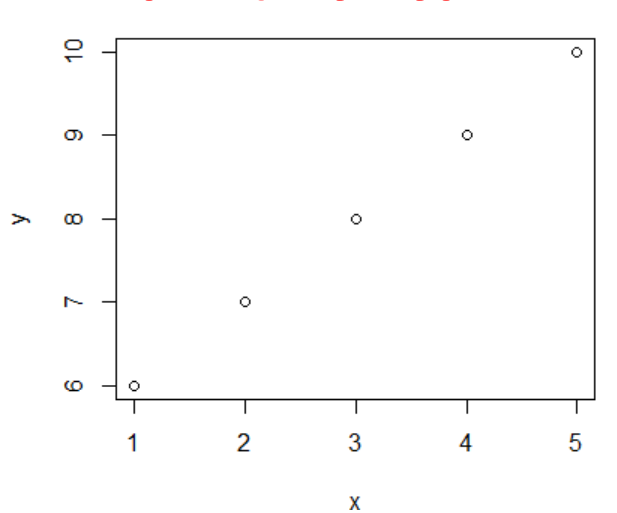




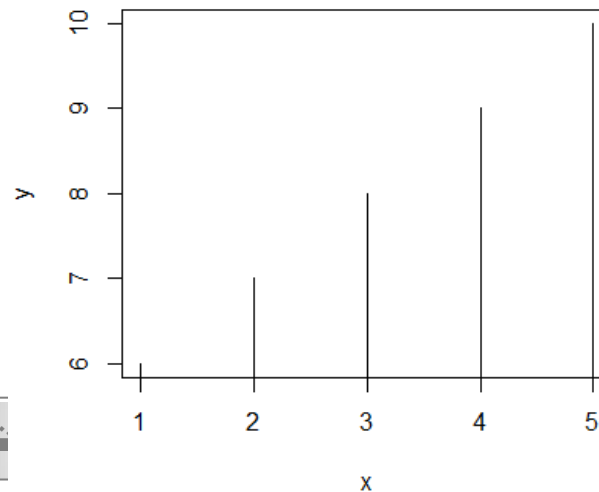
```
> x<- c(1,2,3,4,5)
> y<- c(6,7,8,9,10)
> plot(x,y)
```

```
> plot(x, y, type="l") # line
```

```
> plot(x, y, type="b") # both lines and points
```

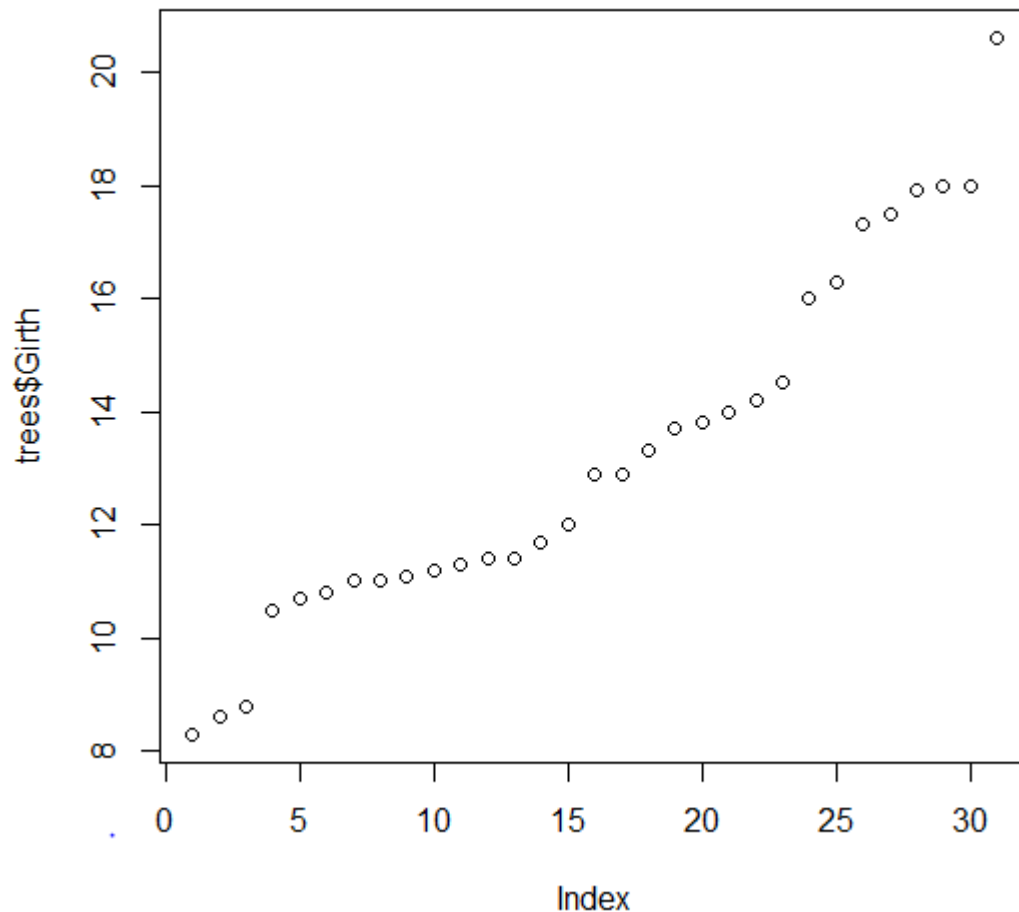


```
> plot(x, y, type="b")
> plot(x, y, type="l")
> plot(x, y, type="h")
```





```
> plot(trees$Girth, trees$Height)
```



هناك بعض المجموعات والتي تأتي محزومة مع اللغة بشكل افتراضي، مثل `trees` ، `mtcars` ، وللحصول على محتوى هذه المجموعات يمكن استخدام الاستدعاء



```
> mtcars
```

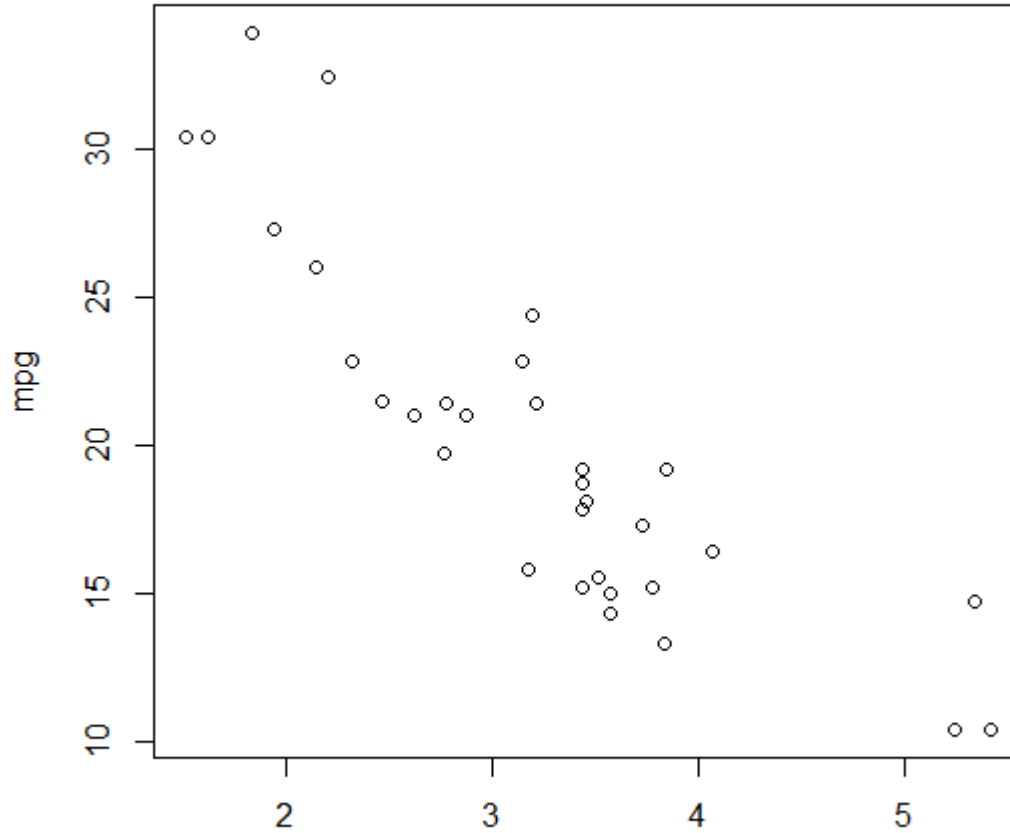
```
attach(datasetname)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2





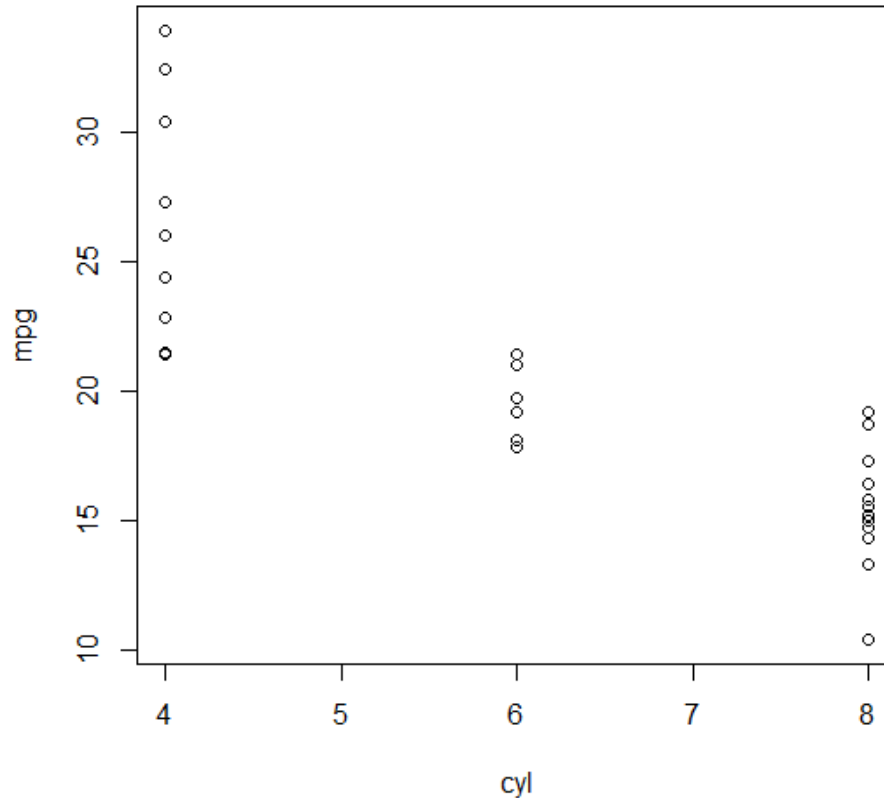
```
> plot(wt, mpg)
```



حيث سيمثل وزن السيارات بقيم wt على محور x فيما المسافة المقطوعة بغالون البنزين الواحد والتي تعطى بقيم mpg ستمثل على المحور y



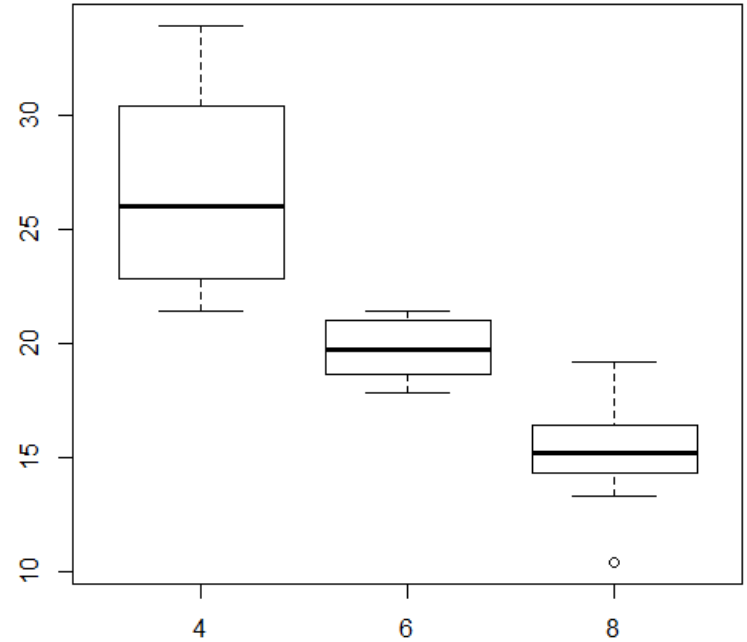
في بعض الأحيان قد لا يكون هذا النوع من المخططات البيانية هو الطريقة الأمثل لعرض ما لدينا من معلومات، خصوصا عندما تكون بيانات أحد طرفي العلاقة عبارة عن قيم محددة بعينها وليست قراءات تتوزع على طيف المحور المسندة إليه كما في حالة المخطط البياني الذي تولده التعليمة `plot(cyl, mpg)` حيث `cyl` تمثل عدد إسطوانات المحرك، حينها سيكون الشكل الناتج غريبا قليلا وأقل فائدة في التعبير عن ما يربط بين المقادير المرسومة كما هو موضح أدناه:





الدالة **plot** يتعدل بشكل آلي تبعا لطبيعة ونوع البيانات التي تمرر إليها، وما سنقوم به الآن هو تحويل نوع `cyl` إلى معاملة وذلك باستخدام الأمر `cyl <- factor(cyl)`، بمعنى أن لهذا المقدار قيم محددة لا يستطيع أن يأخذ غيرها، وسنلاحظ طبيعة هذا التغير في طريقة تعامل توابع لغة R المختلفة مع هذا المقدار الجديد بعد تغيير توصيفه (يمكن لك أن تجرب معه الدالة `summary` لترى أن ماتحصل عليه من ناتج يختلف عما سبق وأن رأيت، فعوضا عن القيمة الصغرى والعظمى والمتوسط والوسيط الخ... وهي المقادير التي توصف بها عادة أي مجموعة قيم عددية، أصبحنا نرى الآن عدد القيم المحددة التي يمتلكها هذا المعامل مقدار تكرر ظهور كل من تلك القيم).

ليس هذا فحسب بل إن سلوك الدالة `plot` سوف يتغير كذلك، فإن حاولت الآن إعادة تنفيذ ذات الأمر السابق `plot(cyl, mpg)` فسوف تحصل على المخطط البياني التالي:

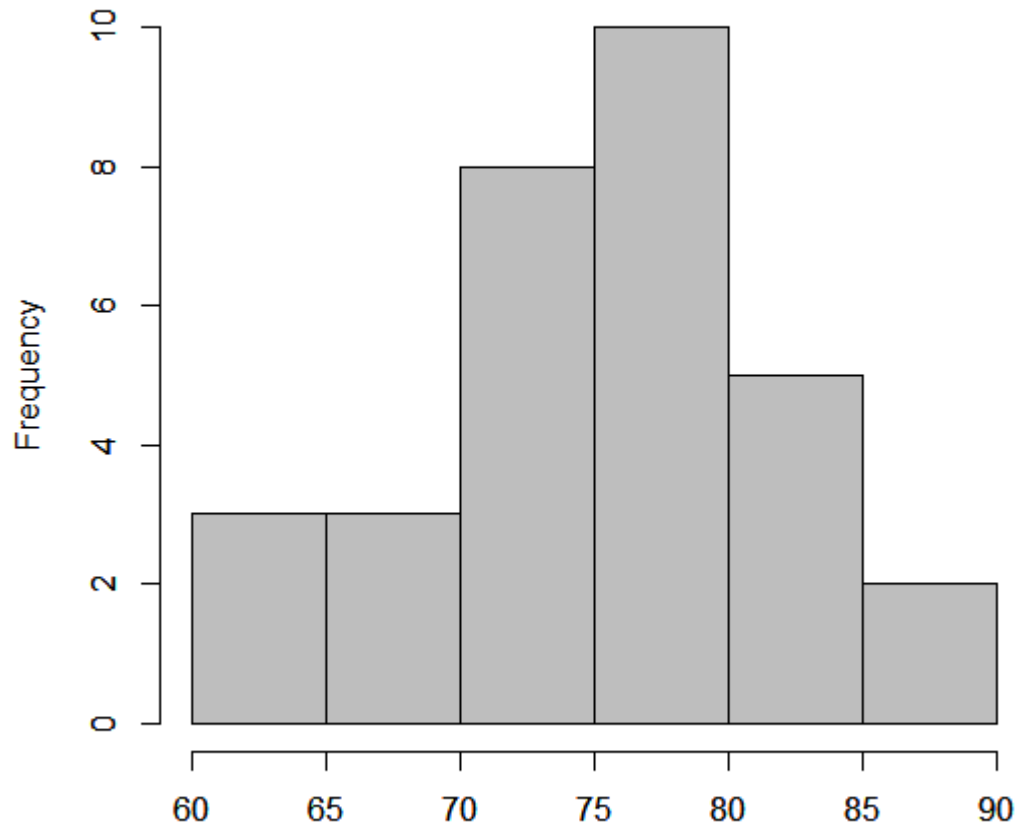


# Histogram



```
> hist(trees$Height, col="gray")
```

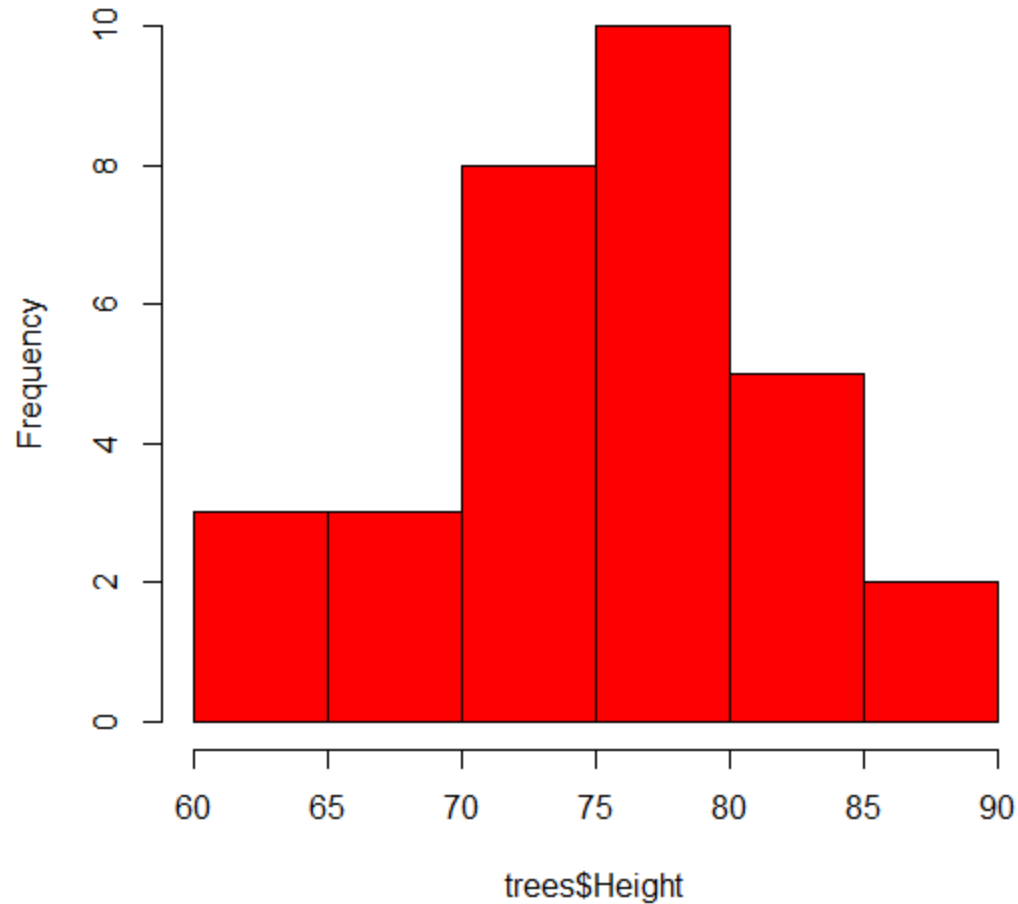
Histogram of trees\$Height



```
> hist(trees$Height, col="red")
```



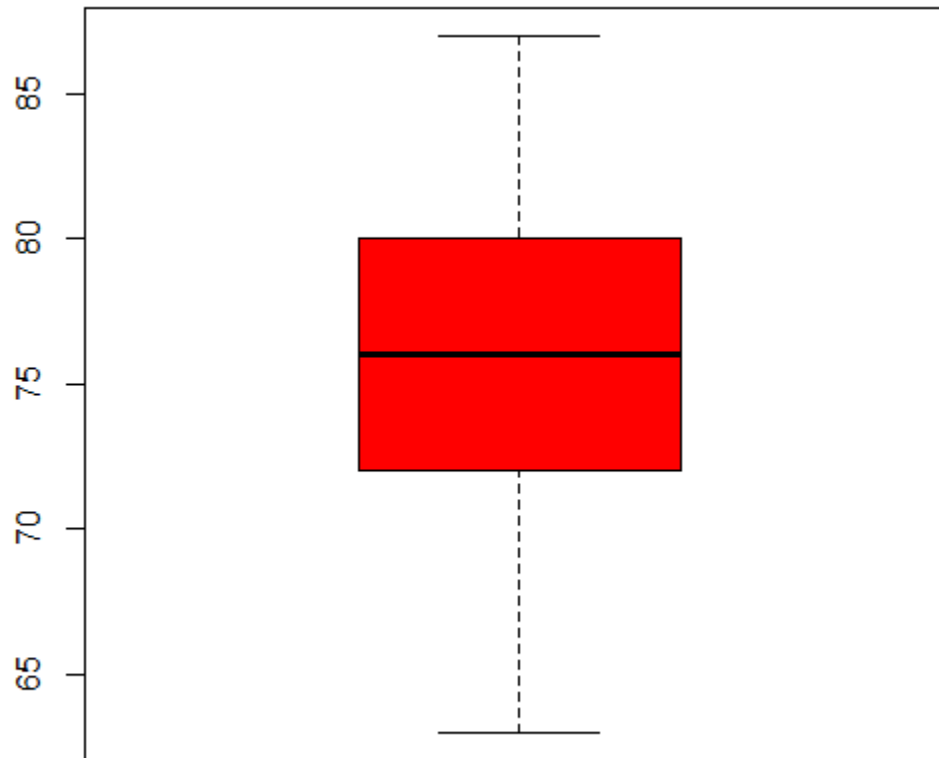
Histogram of trees\$Height





لدينا نوع آخر من المخططات البيانية ذات الصبغة الإحصائية متاح لنا وهو المخطط الصندوقي آنف الذكر، ويمكن طلب عرض بياناتنا من خلاله باستخدام الدالة `boxplot(qsec, col="gray")` حيث سنحصل بالنتيجة على الشكل التالي:

```
> boxplot(trees$Height, col="red")
```



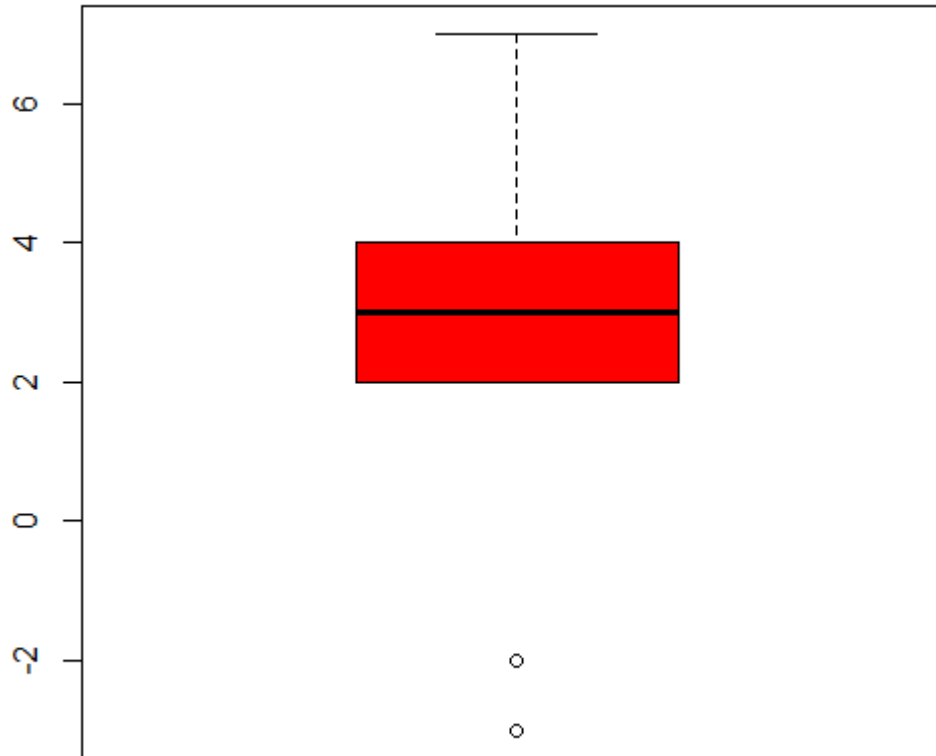


حيث يوضح الخطين الأفقيين على طرفي الرسم في الأعلى والأسفل كل من القيمة الصغرى (في الأسفل) والعظمى (في الأعلى)، أما الصندوق الموجود بينهما فتوضح بدايته من الأسفل ما ندعوه بحد الربع الأول (وهو ما كان يظهر ضمن خرج الدالة summary تحت التسمية (Q1)، وبالتالي يكون المجال المحدد ما بين القيمة الصغرى وطرف هذا الصندوق يتضمن ربع ما لدينا من قيم، أما المجال المحدد ما بين طرفي الصندوق الأسفل والأعلى فيتضمن بالضبط نصف ما لدينا من قيم حيث أن الحد الأعلى للصندوق هو الربع الثالث أي Q3، أما الخط الذي يقطع ذلك الصندوق بالعرض فهو الوسيط (وليس المتوسط الحسابي)، وهو يدل على الحد الذي يقسم كتلة البيانات التي لدينا إلى مجموعتين متساويتين في العدد إحداهما تتضمن القيم التي تعلو خط الوسيط والأخرى فيها القيم التي تقع أسفل خط هذا الوسيط.

في بعض الأحيان قد نرى دوائر أو نقاط تتجاوز حد القيمة العظمى أو تقل عن حد القيمة الصغرى، وهي في واقع الأمر من بياناتنا أيضا لكنها تعامل معاملة القيم الشاذة أو الغريبة وذلك حينما يتجاوز بعدها عن المتوسط ضعفي الانحراف المعياري ( standard deviation لمجموعة البيانات التي لدينا).



```
> boxplot(x, col="red")
```



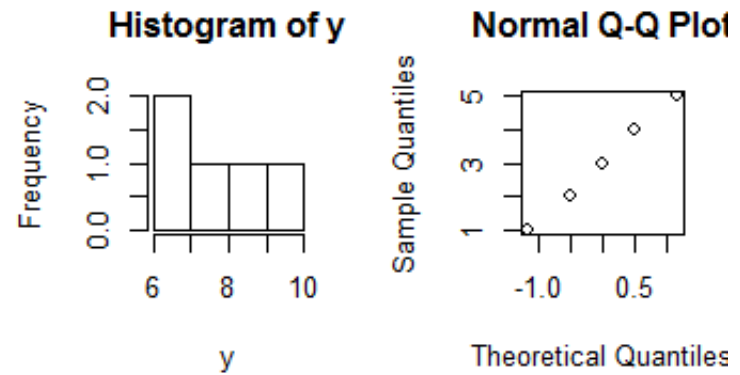
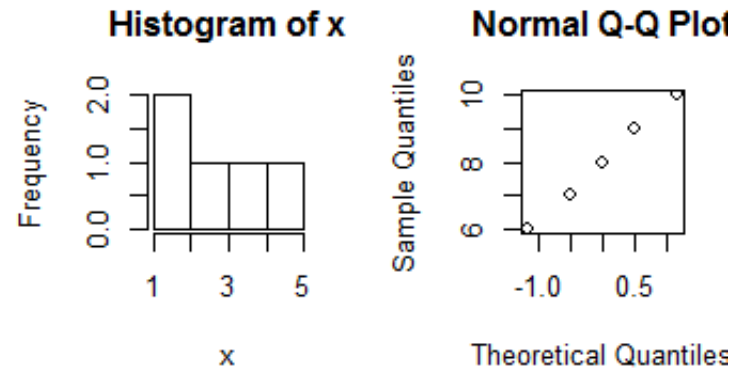
```
> plot(x)  
> hist(x)  
> boxplot(x)
```





في الإمكان وضع أكثر من رسم واحد في الصفحة وذلك بإستخدام المعلم `mfrow` والذي تكون قيمته لمتجة عددي صحيح ذا بعدين يعطي عدد الأسطر والأعمدة لاحظ نتيجة التالى:

```
> x<- c(1,2,3,4,5)
> y<- c(6,7,8,9,10)
> par(mfcol=c(2,2))
> hist(x)
> hist(y)
> qqnorm(y)
> qqnorm(x)
```



ما نتيجة الامر؟

```
> par(mfrow=c(2,2))
. . . . .
```



الدالة :  $cor(..., ...)$

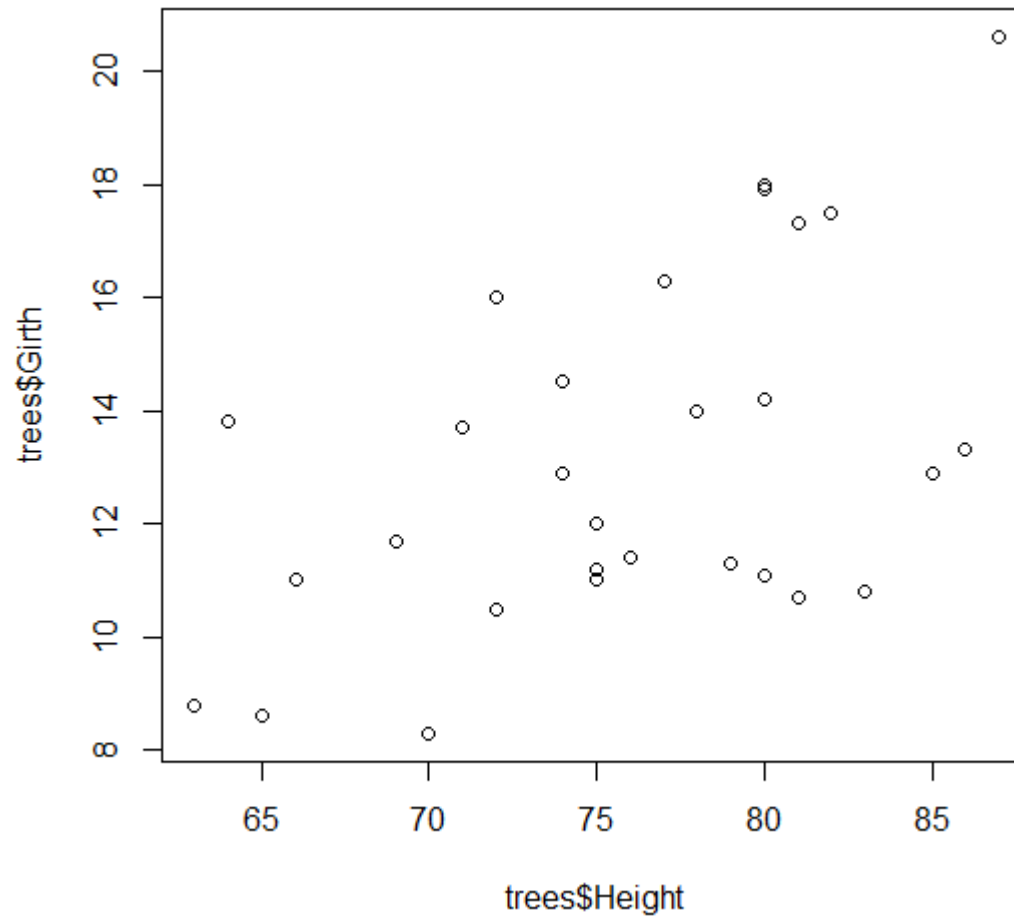
الارتباط : Correlation

You can use the **cor( )** function to produce correlations and **cov( )** function to produces covariances.

تحسب مقدار معامل الارتباط بين شعاعين من القيم هما في هذه الحالة  $x$  و  $y$  أو بين عمودين في حالة البيانات المستوردة من ملف ما)، حيث تتدرج قيمة معامل الارتباط من -1 حتى +1 وتشير القيم الموجبة إلى وجود ارتباط طردي يزداد وضوحا كلما اقترب من +1، أي كلما زادت قيمة  $x$  رافقها زيادة في قيمة  $y$ ، فيما تشير القيمة السالبة لمعامل الارتباط إلى وجود ارتباط أيضا لكنه في هذه الحالة عكسي ويزداد وضوحا كلما اقترب من -1، أي أن قيمة  $x$  تزداد بتناقص  $y$  والعكس صحيح، في حين تشير قيمة معامل الارتباط التي تقترب من الصفر إلى أن تغير قيمة  $x$  لا يظهر أي علاقة بتغير قيمة  $y$  المقابلة



```
> cor(trees$Height, trees$Girth)
[1] 0.5192801
> plot(trees$Height, trees$Girth)
.
```





## معنوية معامل الارتباط : الدالة ( cor.test ..., ... )

```
> cor.test(trees$Height, trees$Girth)
```

```
Pearson's product-moment correlation
```

```
data: trees$Height and trees$Girth
```

```
t = 3.2722, df = 29, p-value = 0.002758
```

```
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
```

```
0.2021327 0.7378538
```

```
sample estimates:
```

```
cor
```

```
0.5192801
```

قيمة معامل الارتباط هي: 0.519 أي أن احتمال المصادفة في الحصول على هذه القيمة لمعامل الارتباط بالنسبة لمجموعة البيانات المدروسة (أي (p-value) هو 0.002758 وهو هامش شك صغير يمكن تجاهله مقارنة بالحد الأقصى لهامش الشك المسموح به ألفا المحدد سبقا والذي يساوي 0.05، لذا من وجهة النظر الإحصائية يعد ذلك الارتباط معنويا و موجودا وليس عن طريق الصدفة. كما تشير النتائج إلى فترة ثقة للمعامل الارتباط بنسبة 95% هي (0.2021, 0.7378)



## تطبيق دالة الارتباط على كامل إطار البيانات trees

```
> cor(trees)
      Girth   Height   Volume
Girth 1.0000000 0.5192801 0.9671194
Height 0.5192801 1.0000000 0.5982497
Volume 0.9671194 0.5982497 1.0000000
```

تحفظ النتائج ضمن مصفوفة تعرض فيها قيم معامل الارتباط ما بين كل زوجين ممكنين من الأعمدة مثلى مثلى

## Example

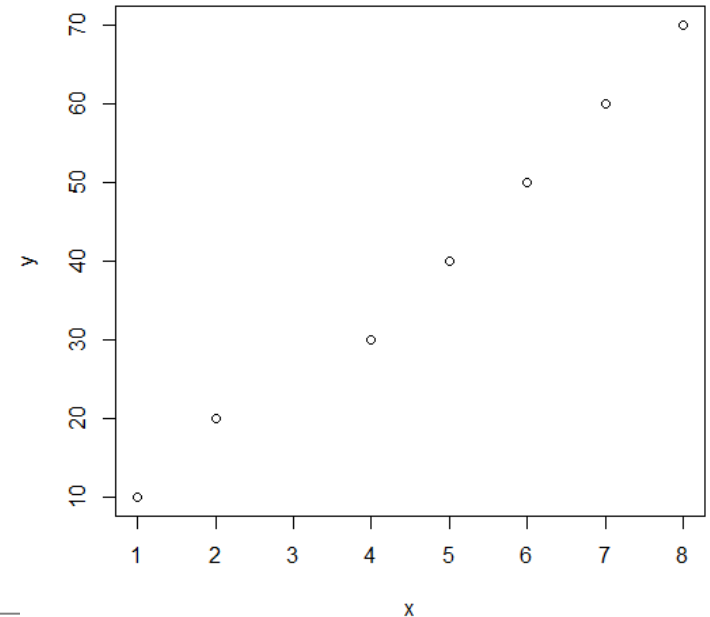


```
> x<-c(1,2,4,5,6,7,8)
> y<-c(10,20,30,40,50, 60,70)
> plot(x,y)
> cor(x,y)
[1] 0.9931833
> cor.test(x,y)
```

Pearson's product-moment correlation

```
data: x and y
t = 19.053, df = 5, p-value = 7.342e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9525932 0.9990370
sample estimates:
      cor
0.9931833

> cov(x,y)
[1] 55
. ■
```



## Example:

## الارتباط: correlatin تعميم



```
> x<-c(1,2,4,5,6,7,8)
> y<-c(10,20,30,40,50, 60,70)
> z<-c(1,3,5,7,9,10,18)
> S<-cbind(x,y,z)
> S
      x  y  z
[1,] 1 10  1
[2,] 2 20  3
[3,] 4 30  5
[4,] 5 40  7
[5,] 6 50  9
[6,] 7 60 10
[7,] 8 70 18
> cor(S)
      x          y          z
x 1.0000000 0.9931833 0.9315596
y 0.9931833 1.0000000 0.9517468
z 0.9315596 0.9517468 1.0000000
>
```



## Example

```
> y<-c(4,5,6)
> z<-c(2,5,8)
> x<-c(1,-2,-6)
> S<-cbind(x,y,z)
> S
      x y z
[1,]  1 4 2
[2,] -2 5 5
[3,] -6 6 8
> cor(S, use="complete.obs", method="kendall")
      x  y  z
x  1 -1 -1
y -1  1  1
z -1  1  1
> cor(S, use="complete.obs", method="pearson")
      x          y          z
x  1.0000000 -0.9966159 -0.9966159
y -0.9966159  1.0000000  1.0000000
z -0.9966159  1.0000000  1.0000000
```



## الانحدار : Regression



إن أشكال علاقات الانحدار يمكن لها أن تتدرج من البساطة (حيث قيمة  $y$  تحسب بدلالة متحول وحيد  $x_1$  إلى أشكال أكثر تعقيدا نستخدم فيها أكثر من متحول  $x$  لحساب قيمة  $y$  مثلا:

$$y = a + b x_1 + c x_2 + d x_3$$

كما أن علاقة الانحدار تلك يمكن أن لا تكون خطية فحسب (وفيها تظهر  $x$  من الدرجة الأولى فقط) بل يمكن لها أن تتعدى إلى صيغ تربيعية أو تكعيبية أو سواه.



بمجرد تحديد وجود ارتباط معنوي ما بين أي مقدارين، علينا عندها توصيف ذلك الارتباط بشكل كمي بعد أن حددنا وجوده بشكل وصفي، وهذا ما يتم توصيفه من خلال علاقة الانحدار ( Regression ) ومثالها العلاقة الخطية البسيطة والتي يعبر عنها بمعادلة خط مستقيم من الشكل  $y = a + b x$  حيث  $a$  هو الثابت الذي يمثل قيمة  $y$  حينما تكون قيمة  $x = 0$  في حين أن  $b$  تمثل ميل ذلك الخط المستقيم (أو نسبة تغير المقدار  $y$  من أجل تغير قيمة  $x$  بمقدار 1)، وتحسب هذه المقادير بحيث يمر الخط المستقيم من بين مجموعة النقاط (  $x, y$  ) بشكل يجعل مقدار الخطأ أقل ما يمكن، حيث أن قيمة الخطأ في كل نقطة  $x$  مقياسة هي عبارة عن الفرق ما بين قيمة  $y$  الفعلية المقابلة وبين قيمة  $y$  المحسوبة من خلال علاقة الخط المستقيم التي حصلنا عليها في علاقة الانحدار.



للقيام بحساب تابع الإنحدار الخطي البسيط بلغة R نستخدم التعليمة التالية

**fit <- lm(y ~ x)**

فمثلا في حال الرغبة في إيجاد تابع الإنحدار الخطي البسيط الذي يحسب ال Height بدلالة Girth نستخدم الأمر التالي:

**fit<-lm(Height, Girth)**

ثم استعراض النتائج من خلال الامر الرائع و العام `summay`

**Summary (fit)**



## Example

```
> fit <- lm(y ~ x)
> summary(fit)
```

Call:

```
lm(formula = y ~ x)
```

Residuals:

```
      1      2      3      4      5      6      7
1.0870  2.7174 -4.0217 -2.3913 -0.7609  0.8696  2.5000
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.5435	2.3186	0.234	0.824
x	8.3696	0.4393	19.053	7.34e-06 ***

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 2.758 on 5 degrees of freedom

Multiple R-squared: 0.9864, Adjusted R-squared: 0.9837

F-statistic: 363 on 1 and 5 DF, p-value: 7.342e-06

# Regression : الانحدار



```
> attach(trees)
> fit <- lm(Height ~ Girth )
> summary(fit)
```

```
Call:
lm(formula = Height ~ Girth)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-12.5816  -2.7686   0.3163   2.4728   9.9456
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.0313     4.3833  14.152 1.49e-14 ***
Girth         1.0544     0.3222   3.272 0.00276 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.538 on 29 degrees of freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:  0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

معادلة الانحدار :  $Height = 62.0313 + 1.0544 Girth$

إن كل من المعاملين  
معنويين عند مستوى  
دلالة 0.05 ، ونجد  
أن قيمة Adjusted  
= R-squared  
0.2445 أي أن  
المتغير Girth يفسر  
ما يقارب 24% من  
التغير في ال متغير  
التابع Height. كما  
أن هذا النموذج  
معنوي حيث أن P-  
value = 0.002 <  
 $\alpha = 0.05$

# الانحدار : Regression

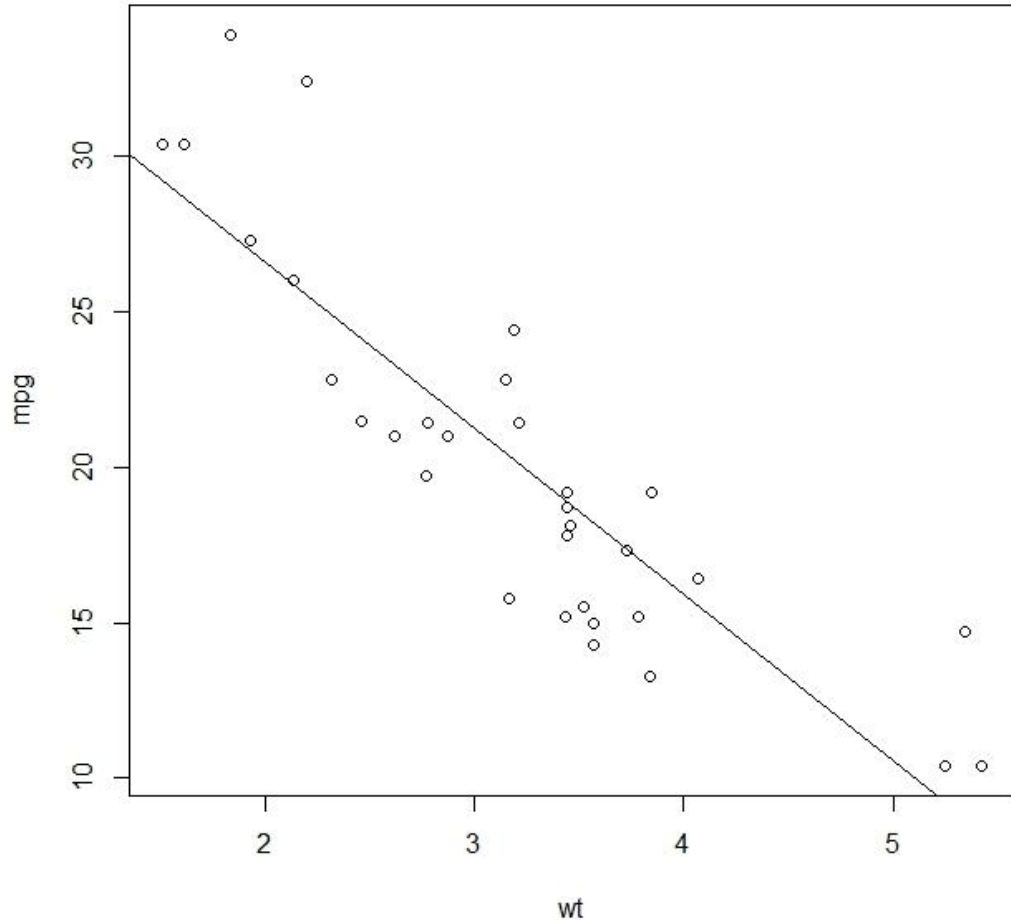


كذلك نستطيع عرض علاقة الانحدار الخطي البسيطة هذه بشكل رسومي أيضا وذلك من خلال التعليمتين التاليتين:

```
plot(wt, mpg)  
abline(fit)
```

Or

```
abline(plot(wt, mpg))
```



## Example:

## الانحدار : Regression



```
> x<-c(1,2,4,5,6,7,8)
> y<-c(10,20,30,40,50, 60,70)
> fit <- lm(y ~ x)
> summary(fit)
```

Call:

```
lm(formula = y ~ x)
```

Residuals:

```
      1      2      3      4      5      6      7
1.0870 2.7174 -4.0217 -2.3913 -0.7609  0.8696 2.5000
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5435      2.3186   0.234   0.824
x            8.3696      0.4393  19.053 7.34e-06 ***
```

---

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.758 on 5 degrees of freedom
```

```
Multiple R-squared:  0.9864,    Adjusted R-squared:  0.9837
```

```
F-statistic: 363 on 1 and 5 DF,  p-value: 7.342e-06
```

```
> abline(fit)
```

