Five

Describing Numerical Data

5.1 Introduction

There are two major ways of describing numerical data:

- Numerical descriptive measures
  - Location
  - Variability
  - Other measures
- Graphical methods
  - Histogram
  - Boxplot, Stem and leaf plot
  - Scatter plot for bivariate data

5.2 Exploring numerical data

- The number of observations in the data set.
- The "center" of the data –
  - Mean
  - Median
  - Mode
- The "variability" of the data –
  - Variance or standard deviation
  - Range
  - Interquartile range between the first and third quartile
  - Coefficient of Variation

- Other measures
  - Minimum and maximum
  - First quartile (25th percentile) and third quartile (75th percentile)
  - Percentiles
  - Skewness Standardised Third Moment
  - Kurtosis Standardised Fourth Moment

EXAMPLE 5.1 The data below describes the percentage returns achieved by 76 average-risk funds:

```
21.88 59.82 33.28 20.23 15.78 13.59 32.98 11.31 21.46
14.45 25.43 10.67 12.11 17.69 56.87  4.43 24.95 20.13
 8.75 12.45 19.89 37.71 49.97 19.67 55.4  27.16 39.44
19.45 15.42 55.9  28.25 31.03 31.15 41.86 49.76 24.89
11.37 18.19 32.5   7.57 29.18 31.7  67.69 10.09 61.88
24.08  7.97 24.22 18.17 24.21 31.96 45.39 56.63 14.36
10    45.37 27.16 -2.7  11.38 12.67 12.41 91.15  8.4
30.52 20.26  4.08 26.09 12.21 67.45 43.31 19.21 18.49
21.31 25.18 26.85 18.89
```

What can we say about the returns of these funds?

- Almost all the funds have a positive return.
- The spread of the returns ranges from -2.7% to 91.15%.
- There appears to be one very high return, 91.15%.

5.3 Descriptive statistics

We will now look at how we can obtain descriptive statistics using SAS, R and SPSS.

Descriptive statistics using SAS

We first look at how we can do so using SAS.

SAS: Proc Univariate

```
data ex5_1ar;
  infile "c:\ST2137\data\ex5_1ar.txt" firstobs=2;
  input preturn;
proc univariate data=ex5_1ar;
  title "Simple Descriptive Statistics";
  var preturn;
run;
```

Output from Proc Univariate

```
                           Simple Descriptive Statistics                    1

                           The UNIVARIATE Procedure
                           Variable:  preturn

                           Moments

N                               76    Sum Weights                 76
Mean                    27.0007895    Sum Observations       2052.06
Std Deviation           17.6647519    Variance            312.043458
Skewness                1.21608091    Kurtosis            1.54518833
Uncorrected SS          78810.4994    Corrected SS        23403.2594
Coeff Variation         65.4230939    Std Error Mean      2.02628601

                   Basic Statistical Measures

                 Location                    Variability

    Mean      27.00079     Std Deviation              17.66475
    Median    22.98000     Variance                  312.04346
    Mode      27.16000     Range                      93.85000
```

Interquartile Range	18.76500

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	13.32526	Pr > |t|	<.0001
Sign	M	37	Pr >= |M|	<.0001
Signed Rank	S	1462	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	91.150
99%	91.150
95%	61.880
90%	55.900
75% Q3	32.740
50% Median	22.980
25% Q1	13.975

10%	10.000
5%	7.570
1%	-2.700
0% Min	-2.700

Extreme Observations

----Lowest----		----Highest----	
Value	Obs	Value	Obs
-2.70	58	59.82	2
4.08	66	61.88	45
4.43	16	67.45	69
7.57	40	67.69	43
7.97	47	91.15	62

SAS: Proc Mean

```
proc means data=ex5_1ar n mean std min max stderr maxdec=4;
  title "Procedure Means";
  var preturn;
run;
```

Output from Proc Mean

```
                 The MEANS Procedure

                         Analysis Variable : preturn
```

N	Mean	Std Dev	Minimum	Maximum	Std Error
76	27.0008	17.6648	-2.7000	91.1500	2.0263

Descriptive statistics using R

```
> ex5.1ar <- read.table("c:/ST2137/data/ex5_1ar.txt", header=T)
> attach(ex5.1ar)
> summary(preturn)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -2.70   14.17   22.98   27.00   32.62   91.15

> mean(preturn)
[1] 27.00079

> median(preturn)
[1] 22.98

> min(preturn)
[1] -2.7

> max(preturn)
[1] 91.15

> range(preturn)
[1] -2.70 91.15
```

```
> quantile(preturn)
     0%      25%      50%      75%     100% 
-2.7000 14.1675 22.9800 32.6200 91.1500 

> var(preturn)
[1] 312.0435

> sd(preturn)
[1] 17.66475

> IQR(preturn)  # Interquartile range
[1] 18.4525

> preturn[order(preturn)[1:5]] # The smallest 5 observations
[1] -2.70  4.08  4.43  7.57  7.97

> nar <- length(preturn)
> preturn[order(preturn)[(nar-4):nar]] # The biggest 5 observations
[1] 59.82 61.88 67.45 67.69 91.15

> cv <- function(x) sd(x)/mean(x) # Compute coeff of variation
> cv(preturn)
[1] 0.6542309
```

Calculating sample skewness using R

Unbiased estimator of skewness is given by

$$\frac{\sqrt{n(n-1)}}{n-2}\left(\frac{m_3}{m_2^{3/2}}\right),$$

where m_i is the ith central moment.

```
> skew <- function(x){
+   n <- length(x)
+   m3 <- mean((x-mean(x))^3)
+   m2 <- mean((x-mean(x))^2)
+   sk=m3/m2^(3/2)*sqrt(n*(n-1))/(n-2)
+   return(sk)
+ }

> skew(preturn)
[1] 1.216081
```

Calculating sample kurtosis using R

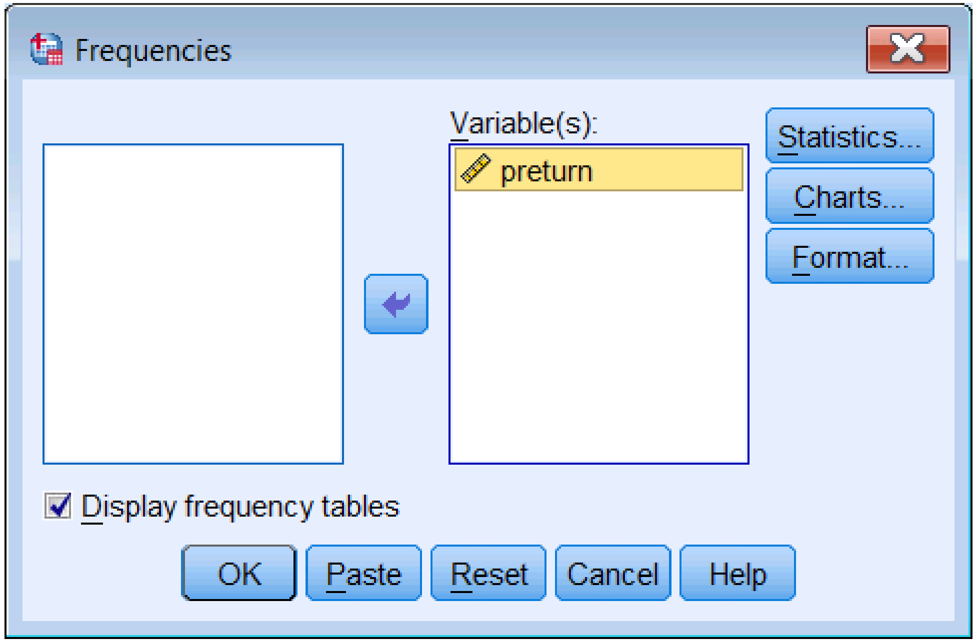
Unbiased estimator of kurtosis is given by

$$\frac{n-1}{(n-2)(n-3)}\left(\frac{(n+1)m_4}{m_2^2}-3(n-1)\right).$$

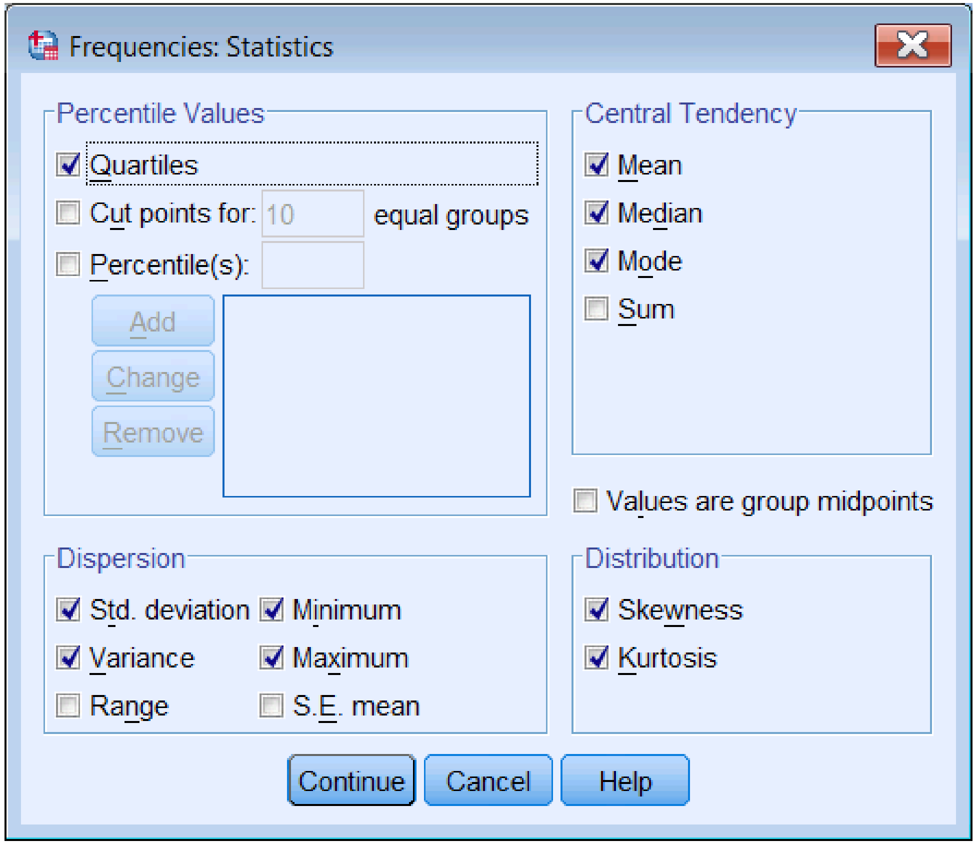
```
> kurt <- function(x){
+   n <- length(x)
+   m4 <- mean((x-mean(x))^4)
+   m2 <- mean((x-mean(x))^2)
+   kurt = (n-1)/((n-2)*(n-3))*((n+1)*m4/m2^2-3*(n-1))
+   return(kurt)
+ }

> kurt(preturn)
[1] 1.545188
```

Descriptive Statistics using SPSS

- Open the SPSS data file `ex5.1ar.sav`.
- "Analyze" → "Descriptive Statistics" → "Frequencies ...".
- Move "preturn" to the variable panel.



- Click "Statistics...".



- Choose the statistics you want.

- Click "Continue" → "OK".

- The output is show below:

Statistics

preturn

N	Valid	76
	Missing	0
Mean		27.0008
Median		22.9800
Mode		27.16
Std. Deviation		17.66475
Variance		312.043
Skewness		1.216
Std. Error of Skewness		.276
Kurtosis		1.545
Std. Error of Kurtosis		.545
Minimum		-2.70
Maximum		91.15
Percentiles	25	13.7825
	50	22.9800
	75	32.8600

5.4 Graphical Methods

Now we look at obtaining graphical summaries using those software.

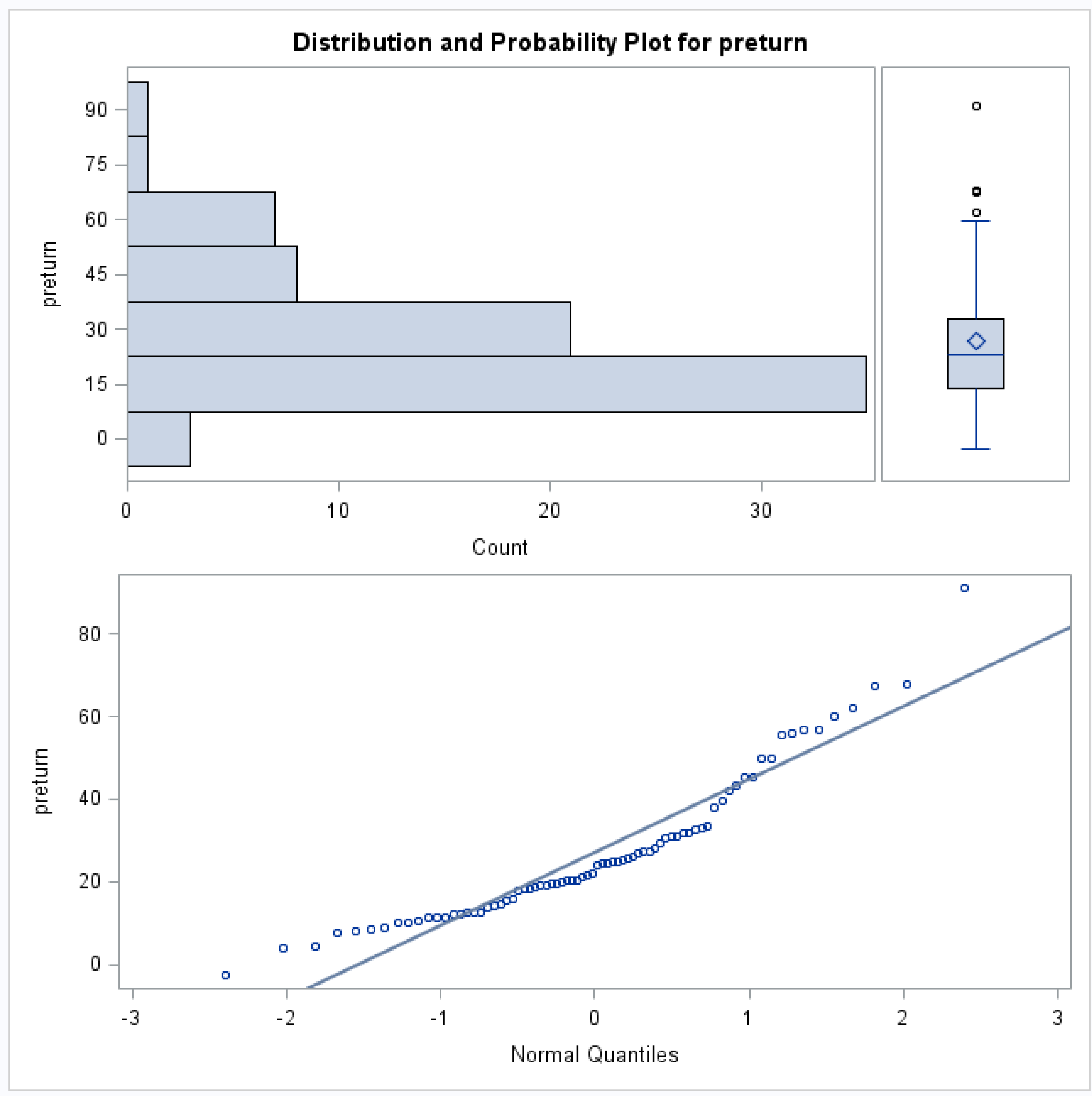
Graphical Methods using SAS

- Box plot
- “Distribution Plot”
- Histogram
- QQ plot

Plot in Proc Univariate: SAS

```
/* Boxplot, distribution plot, and qqplot */
proc univariate data=ex5_1ar plot;
  var preturn;
run;
```

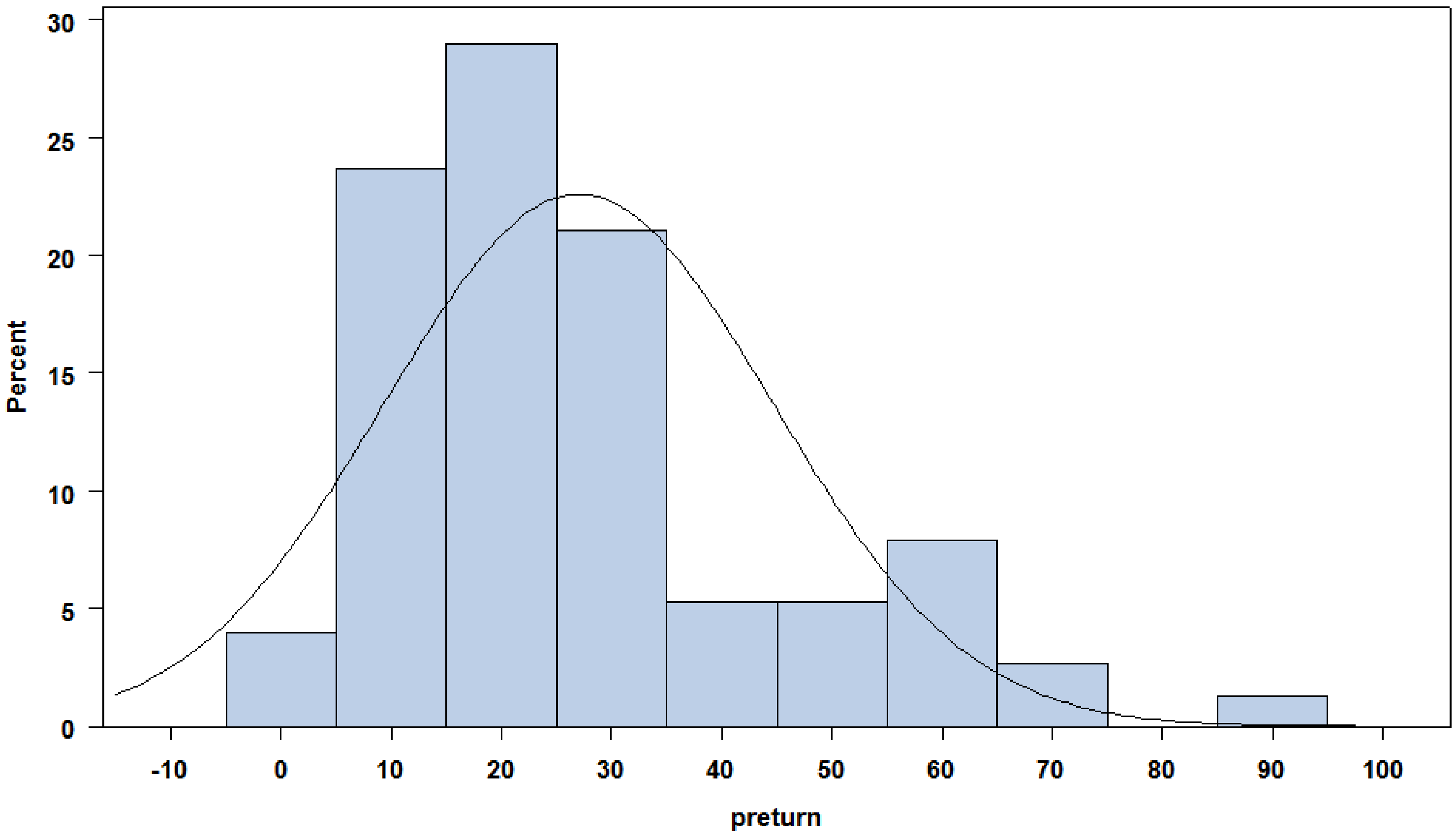
Output from Plot



Histogram: SAS

```
* Histogram;
proc univariate data=ex5_1ar;
  var preturn;          histogram preturn / normal
  histogram preturn / midpoints=-10 to 100 by 10 normal;
/* The option "normal" gives a superimposed normal plot */
run;
```

Output : histogram in SAS



Graphical Methods in R

R has very good graphical capabilities which we will explore here.

Histogram: R

```
> hist(preturn, freq=FALSE,
+  main = paste("Histogram of return"),
+  xlab = "return", ylab="density")
```

paste: join different string

```
> # Normal curve imposed on the histogram
> xpt <- seq(-10,100,0.1)
> ypt <- dnorm(seq(-10,100,0.1), mean(preturn), sd(preturn))
> lines(xpt,ypt)
```

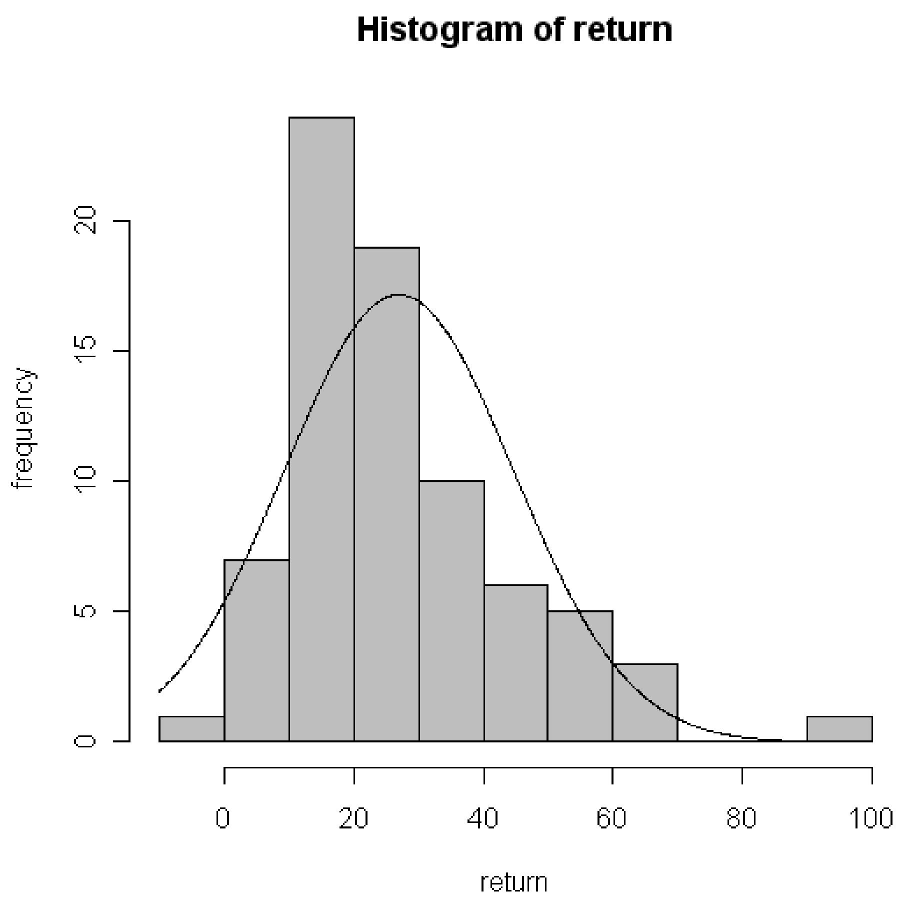
dnorn: value of the density curve

```
> # To get frequency histogram version
> hist(preturn, freq=TRUE,
+  main = paste("Histogram of return"),
+  xlab = "return", ylab="frequency")
```

the line is close to x-axis since the y scare is too big

```
> # Normal curve imposed on the histogram
> aypt <- ypt*length(preturn)*10
> # length(preturn)*10 is the area of the histogram
> lines(xpt,aypt)
```

Output: histogram in R

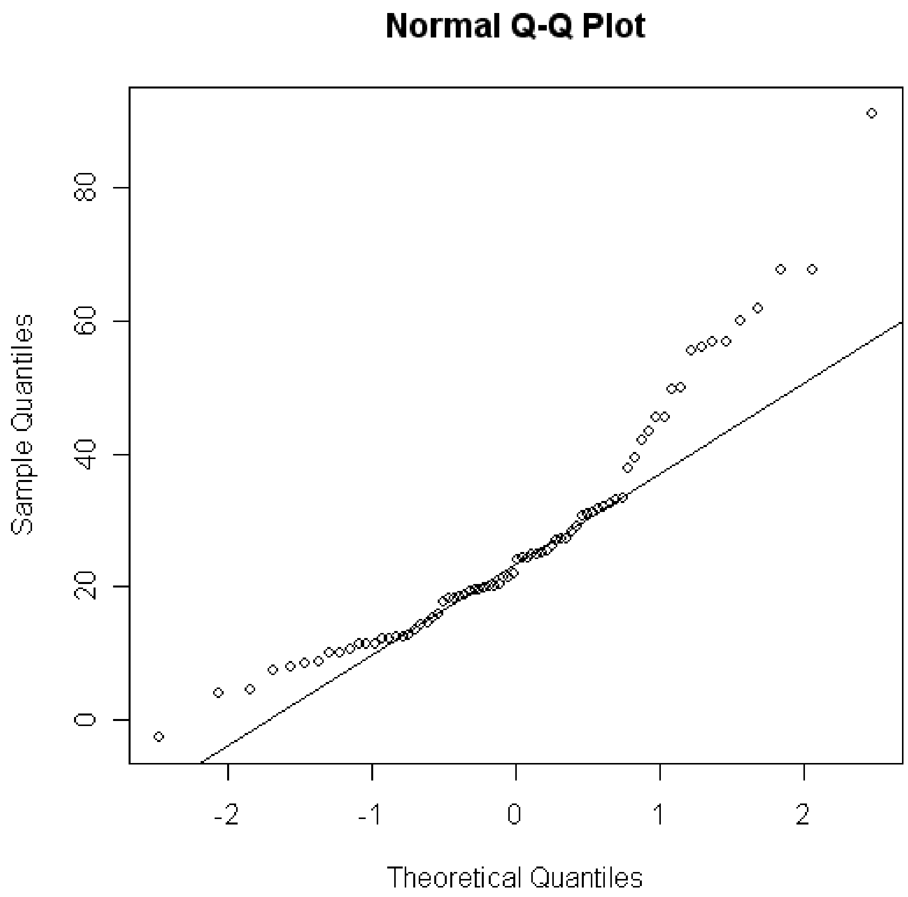


The shape of the histogram does not fit well with the normal curve.

QQplot: R

```
> # qqplot of normal distribution
> qqnorm(preturn)
> qqline(preturn)
```

qqline(preturn,,probe=c(0.01,0.99)
myqq=qqnorm(preturn)
abline(lm(myqq$y+myqq$x))



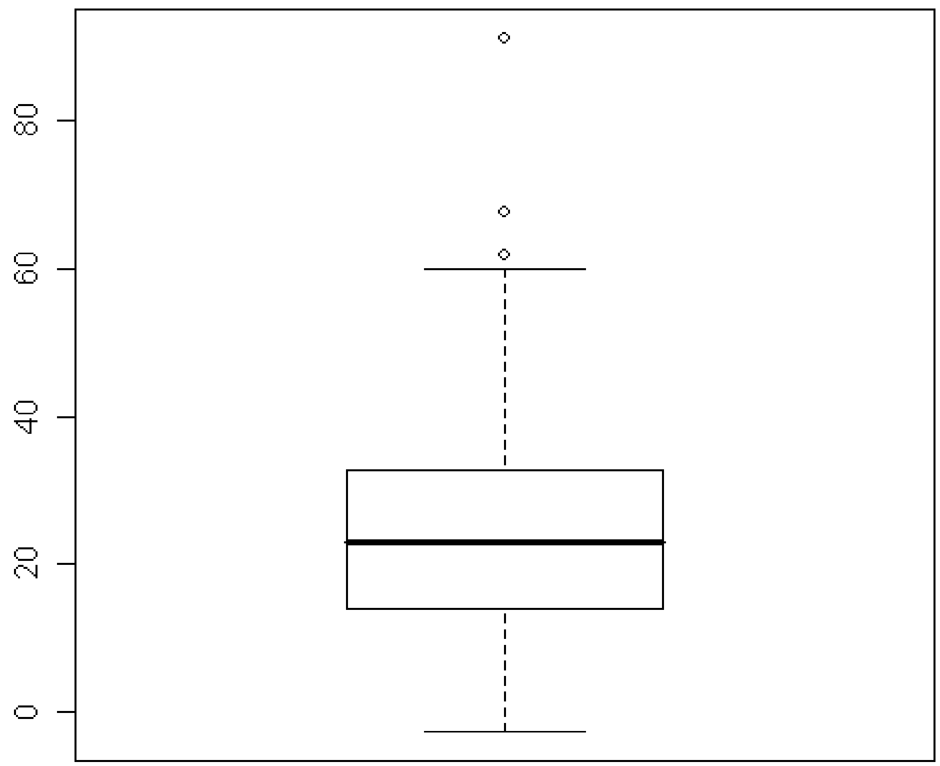
Stem and leaf plot: R sideway histogram

```
> stem(preturn)

  The decimal point is 1 digit(s) to the right of the |

  -0 | 3   -3
   0 | 448889
   1 | 001111222234444568888999
   2 | 00000112444555567777789
   3 | 1112233389
   4 | 2355
   5 | 005677
   6 | 0278   60,62,67,68
   7 |
   8 |
   9 | 1
```

The stem size is in the form of 10^x.

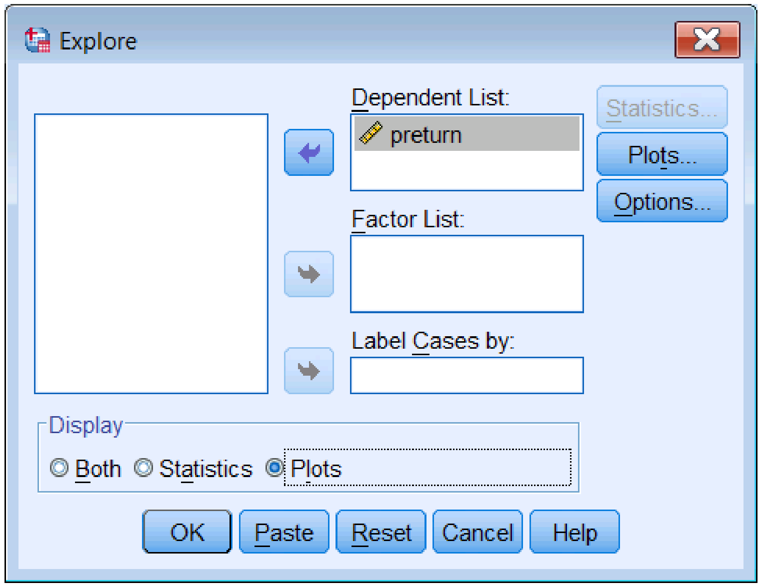
Boxplot: R

```
> boxplot(preturn)
```

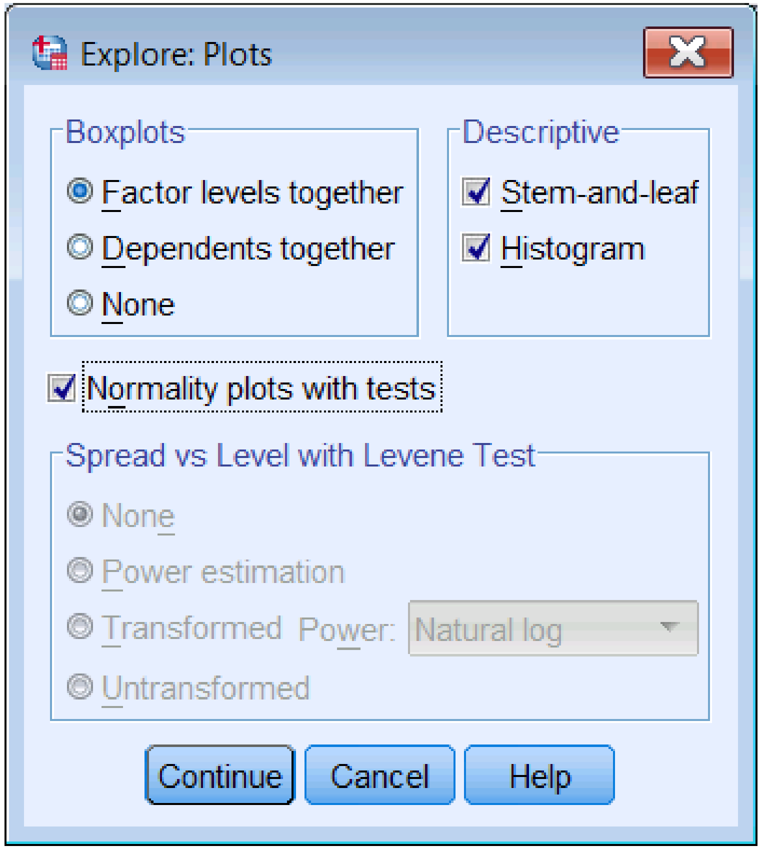


Plots using SPSS

- Open the SPSS data file `ex5.1ar.sav`.
- "Analyze" → "Descriptive Statistics" → "Explore ...".
- Move "preturn" to the variable panel.
- Click "Plots...".

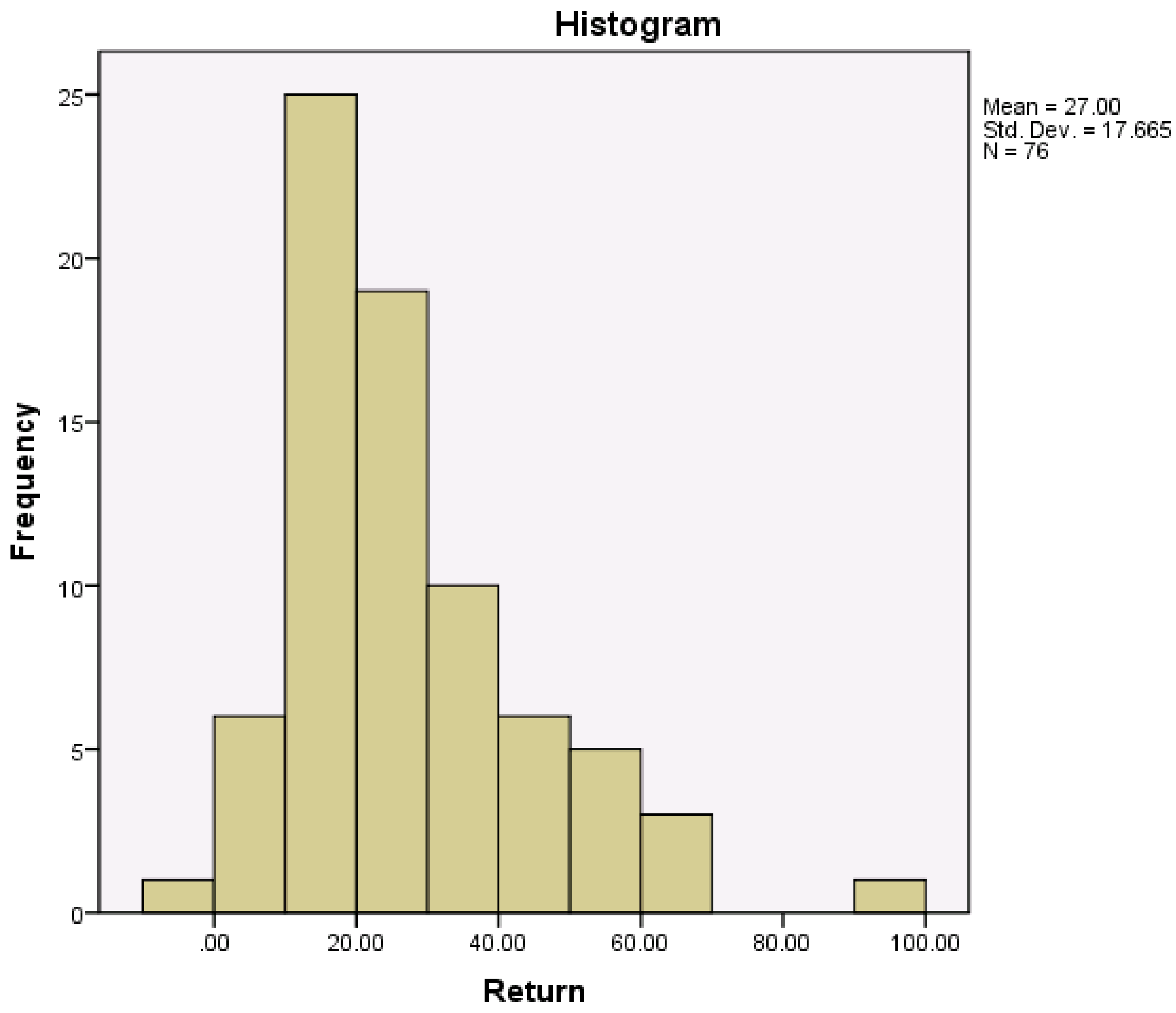


- Choose the plots you want (e.g. Stem-and-leaf, Histogram and Normality plot with tests).



- Click “Continue" → “OK".

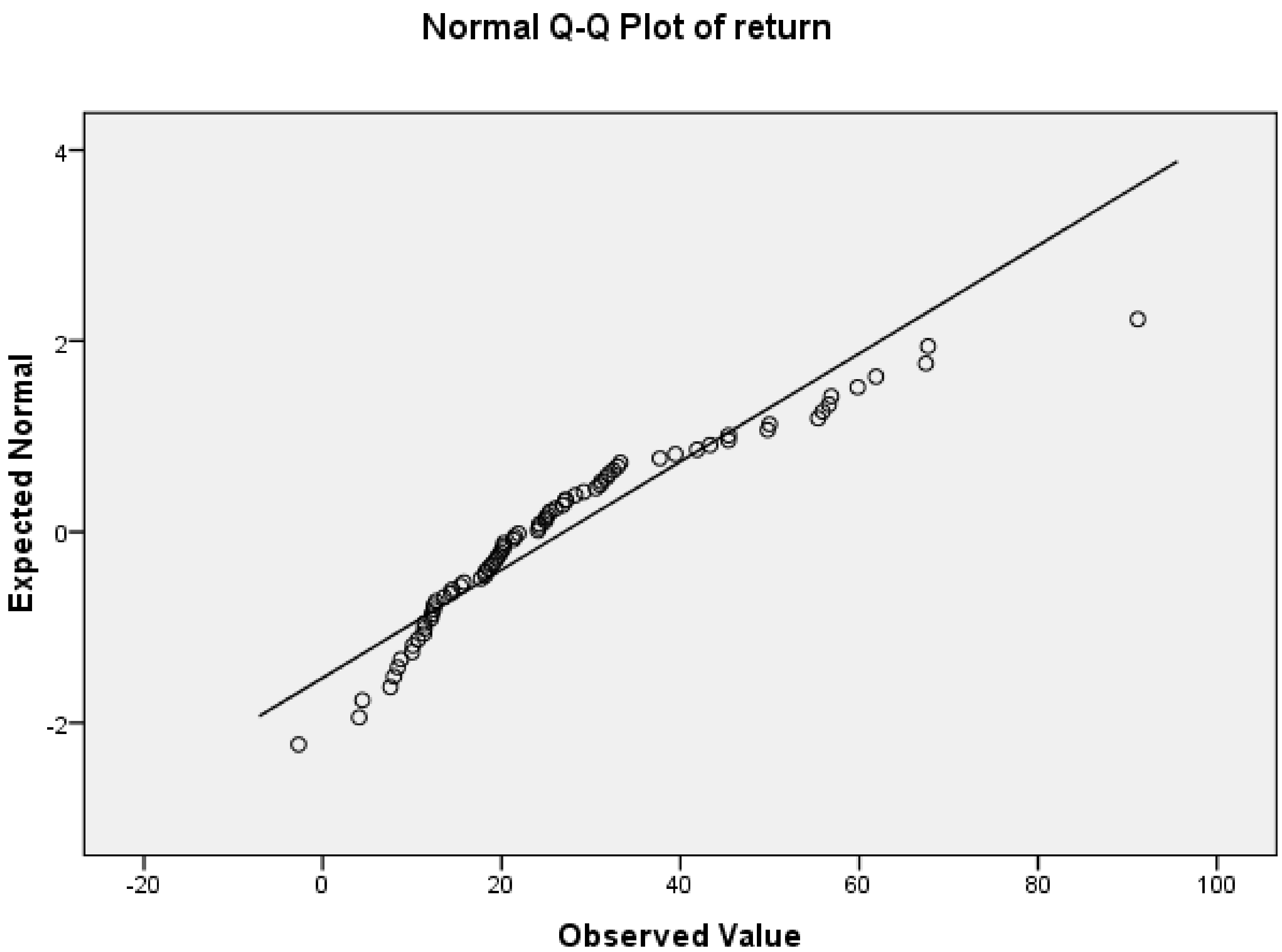
Output: Histogram in SPSS



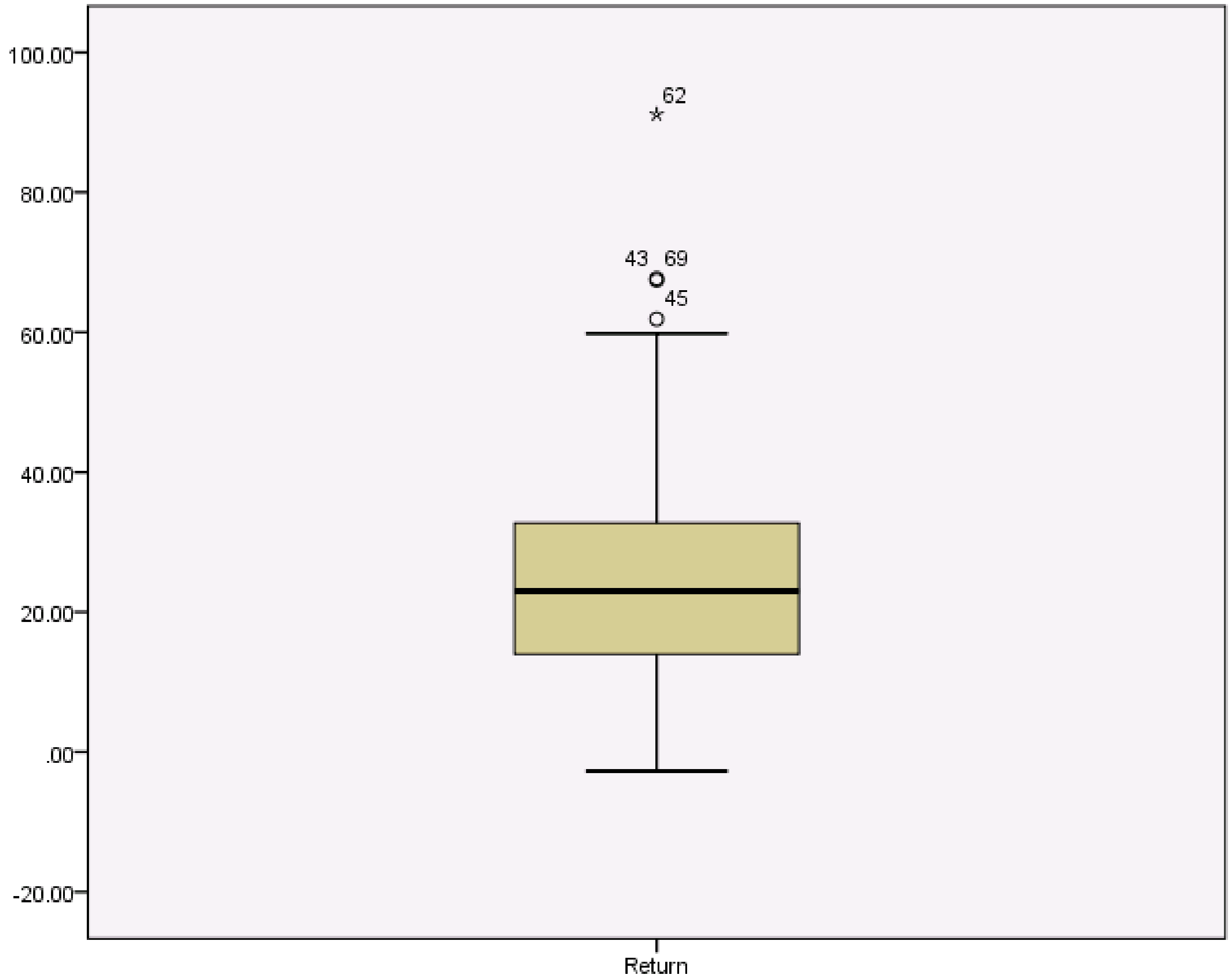
Output: Stem & Leaf Plot in SPSS

```
Return Stem-and-Leaf Plot
 Frequency      Stem &  Leaf
     1.00         -0 .  2
     6.00          0 .  447788
    25.00          1 .  0001112222234455788889999
    19.00          2 .  0001114444455667789
    10.00          3 .  0111122379
     6.00          4 .  135599
     5.00          5 .  55669
     4.00 Extremes     (>=62)
 Stem width:  10.00
 Each leaf:        1 case(s)
```

Output: QQ plot in SPSS



Output: Boxplot in SPSS

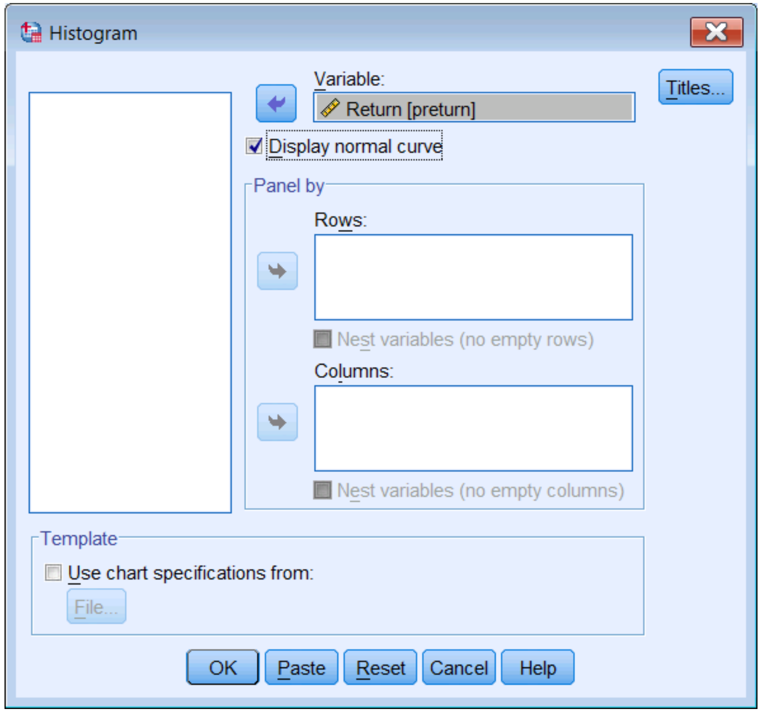


Alternative way to get histogram with normal curve in SPSS

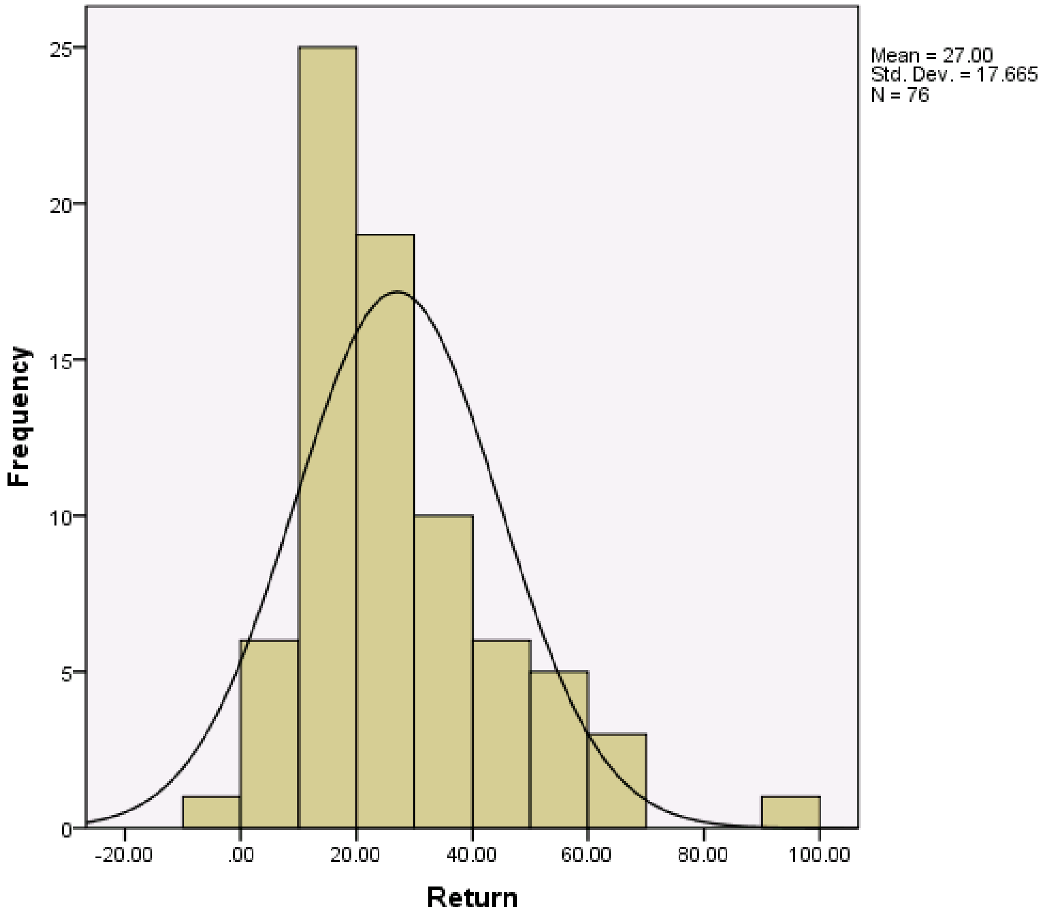
- Open the SPSS data file `ex5.1ar.sav`.

- "Graphs" → "Legacy Dialogs" → "Histogram ...".

- Move the variable "preturn" from the left panel to the right panel under "Variable".

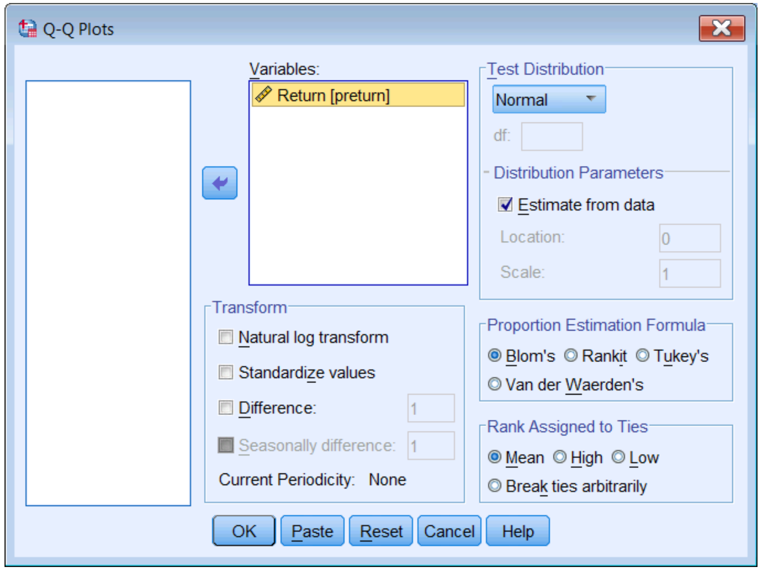


- Choose “Display normal curve", then click “OK".



Alternative way to get QQplot in SPSS

- Open the SPSS data file `ex5.1ar.sav`.
- "Analyze" → "Descriptive Statistics" → "QQ Plots ...".
- Move the variable "preturn" from the left panel to the right panel under "Variable".

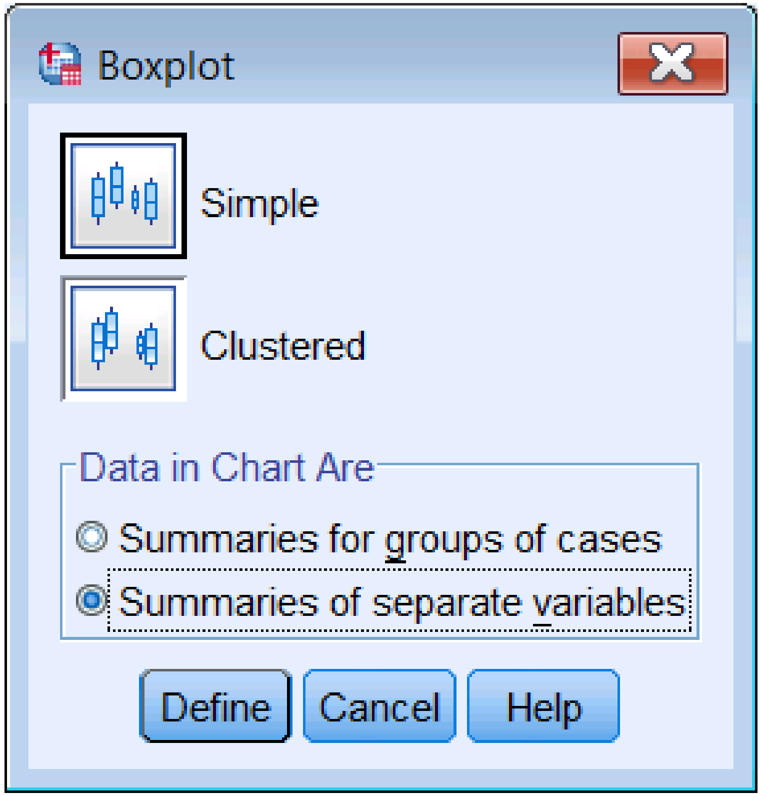


- Choose "Normal" in the "Test Distribution", then click "OK".

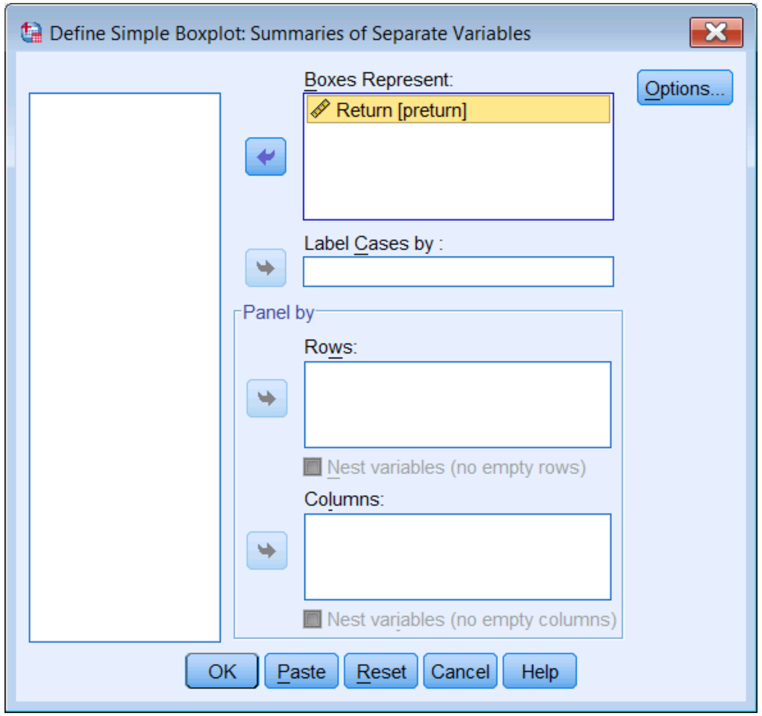
Alternative way to get Boxplot in SPSS

- Open the SPSS data file `ex5.1ar.sav`.

- "Graphs" → "Legacy Dialogs" → "Box plot ...".



- Choose “Simple" and “Summaries in separate variables", then click “Define".



- Move the variable “return" from the left panel to the right panel under “Variable".

- Click “OK".

EXAMPLE 5.2 In addition to the percentage returns of the 76 average-risk funds, we have the following percentage returns achieved by 47 high-risk funds:

24.47	8.76	58.71	35.07	-22.82
15.67	37.47	13.4	49.02	-3.16
43.97	13.91	23.84	-2.89	39.64
29.72	17.91	-10.55	47.38	68.58
86.13	-12.57	-0.33	-5.32	4
0.14	45.97	23.87	38.23	63.79
26.6	6.62	6.72	36.02	29.51
22.56	1.62	29.33	28.91	29.32
42.91	8.87	54.43	28.31	30.76
49.67	1.98			

Solution:

(a) **Preparing the dataset**

For a given fund, there are two variables:

- Return: Percentage return of the fund and
- Risk: 1 for average-risk fund and 2 for high-risk fund.

Notice that the variable “Risk" is a classification (categorical) variable.

(b) **How to enter the data**

- Two columns with the first column representing the return percentage while the second column identifies the type of funds.

```
Return Risk
21.88 1
59.89 1
..... .....
1.98 2
```

- There are 123 data in both columns.

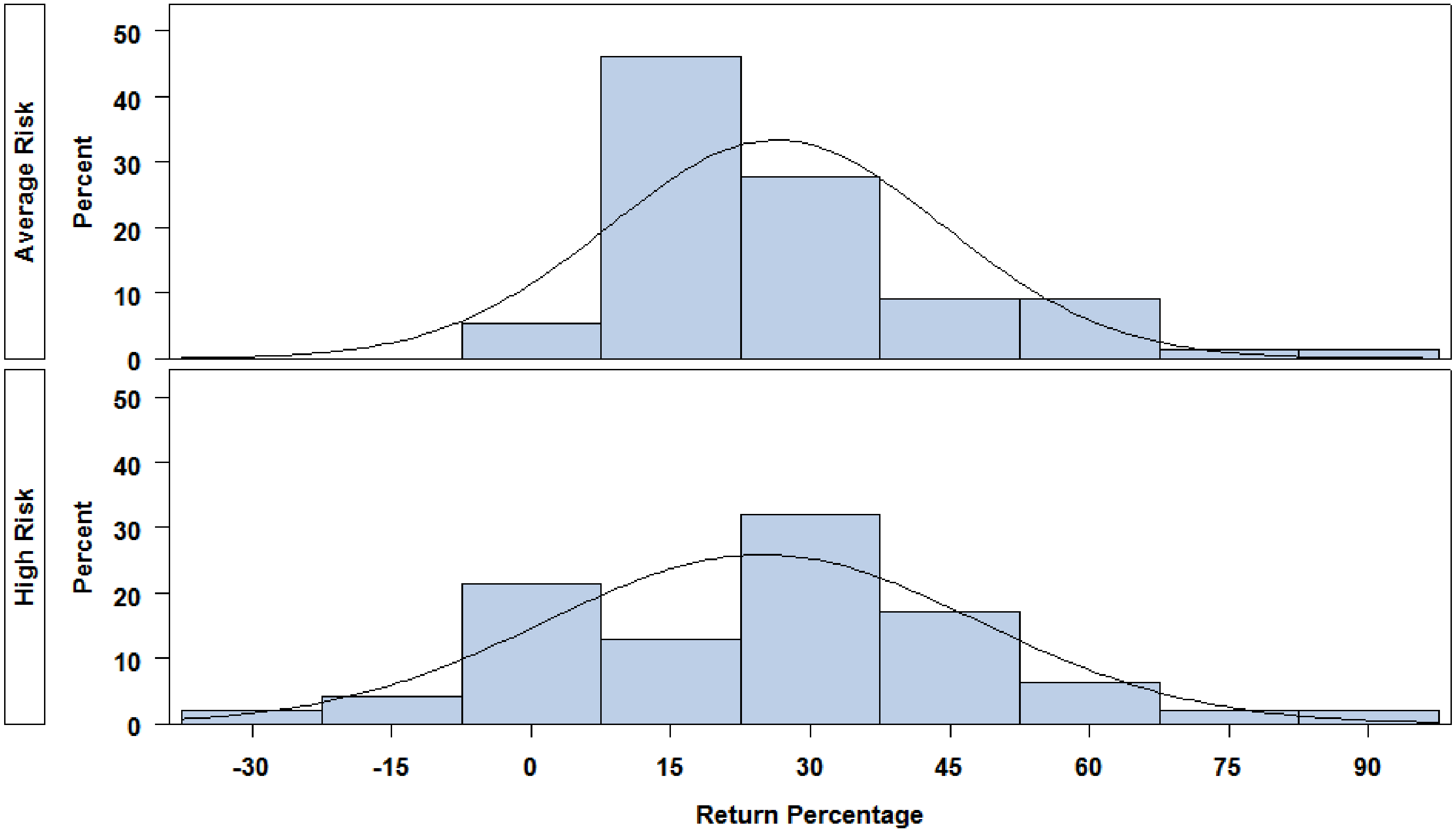
(c) Descriptive statistics by groups: SAS

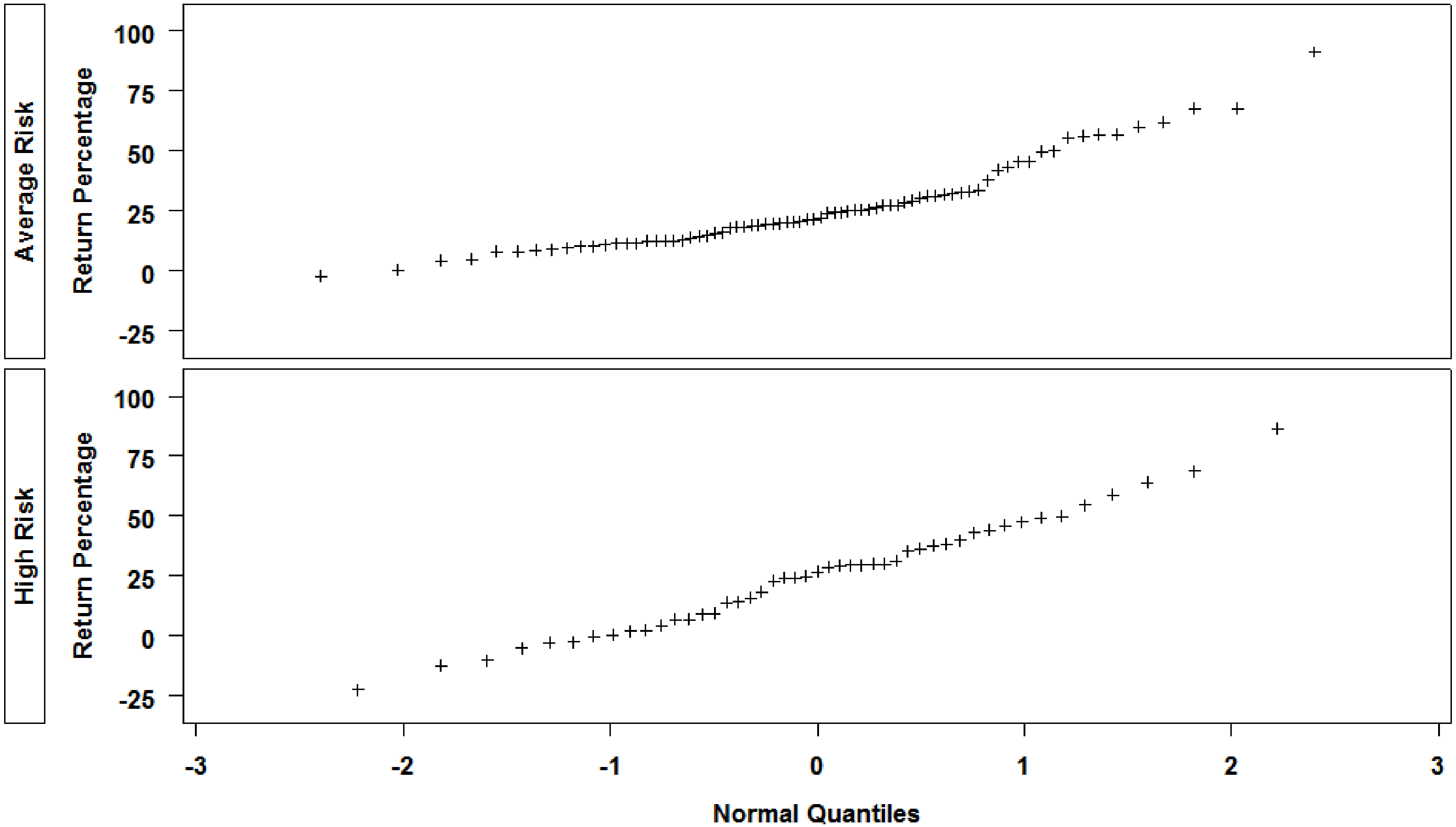
```
proc format;
  value $risk '1' = 'Average Risk'
              '2' = 'High Risk';
data ex5_2;
  infile "c:\ST2137\data\ex5_2.txt" firstobs=2;
  input preturn risk $;
  label preturn = 'Return Percentage';
 format risk $risk.;

proc univariate data=ex5_2;
  histogram preturn/ midpoints= -30 to 90 by 15  normal;
 class risk;
  qqplot preturn; /* you can also call qqplot this way,         /
                  /  allowing you to do side by side comparison */
run;
```

$表示把data当成string read in，如果要把data当作integer read in，要把所有¥去掉

这样就会根据这个class画两个图

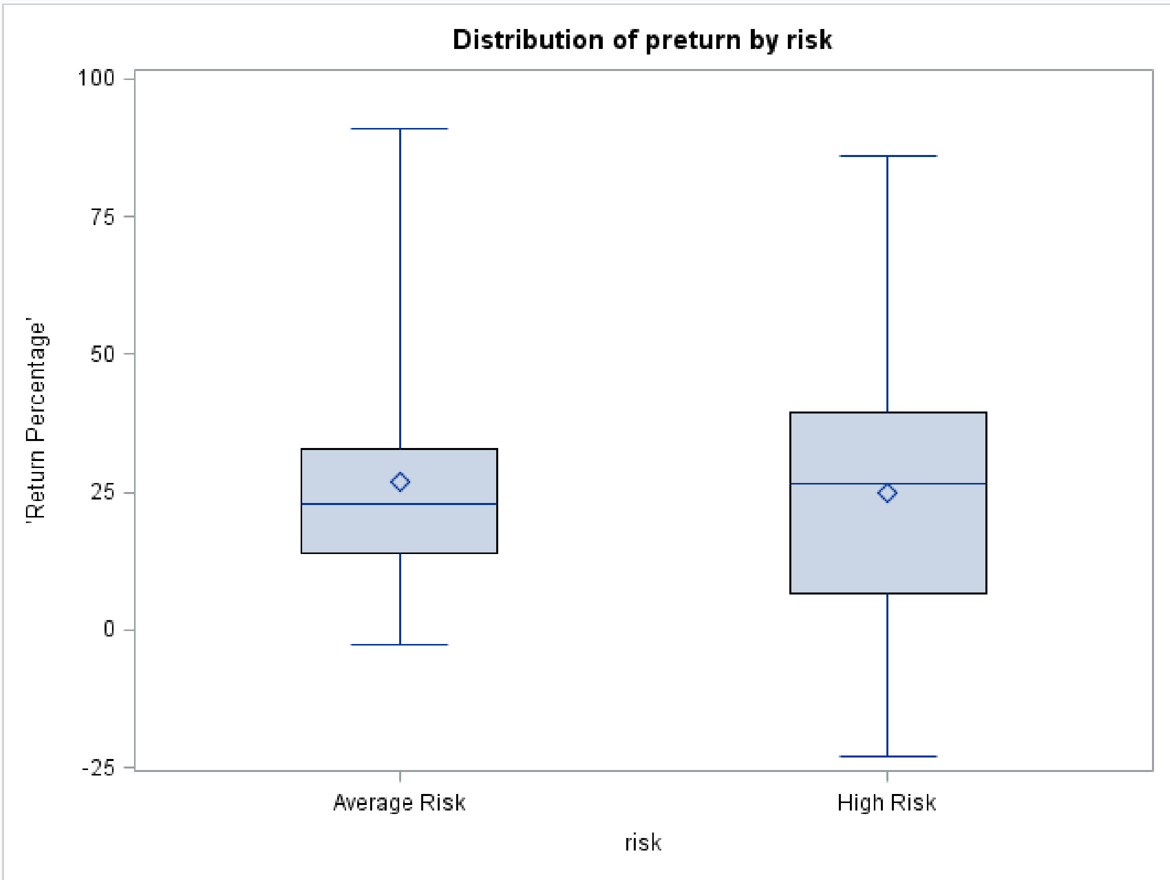
Output by groups: SAS




Average Risk
Return Percentage
100
75
50
25
0
-25
High Risk
Return Percentage
100
75
50
25
0
-25
-3
-2
-1
0
1
2
3
Normal Quantiles

(d) Boxplot by groups: SAS

```
proc boxplot data=ex5_2;
  plot preturn*risk;
run;
```



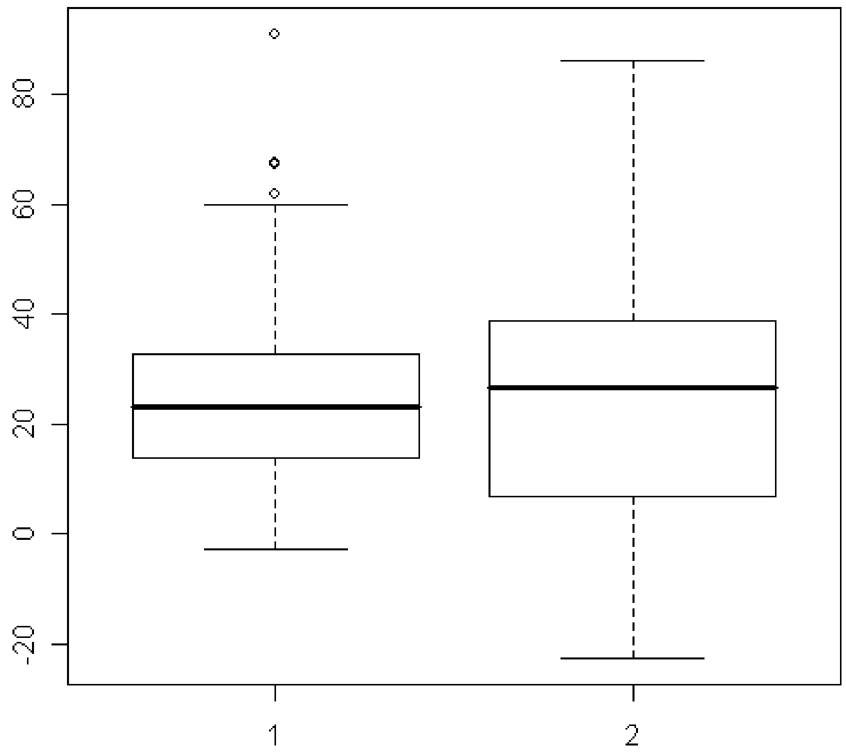
(e) Descriptive statistics by groups: R

```
> ex5.2 <- read.table("c:/ST2137/data/ex5_2.txt", header=T)
> attach(ex5.2)
> ex5.2ar <- ex5.2[risk==1,"preturn"]
> ex5.2hr <- ex5.2[risk==2,"preturn"]
> summary(ex5.2hr)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -22.82    6.67   26.60   24.81   38.94   86.13

> stem(ex5.2hr)

  The decimal point is 1 digit(s) to the right of the |

  -2 | 3
  -0 | 315330
   0 | 022477993468
   2 | 3444789990015678
   4 | 034679049
   6 | 49
   8 | 6
```

(f) Boxplot by groups: R

```
> boxplot(preturn~risk)
```

boxplot(ex5.2ar,ex5.2hr)

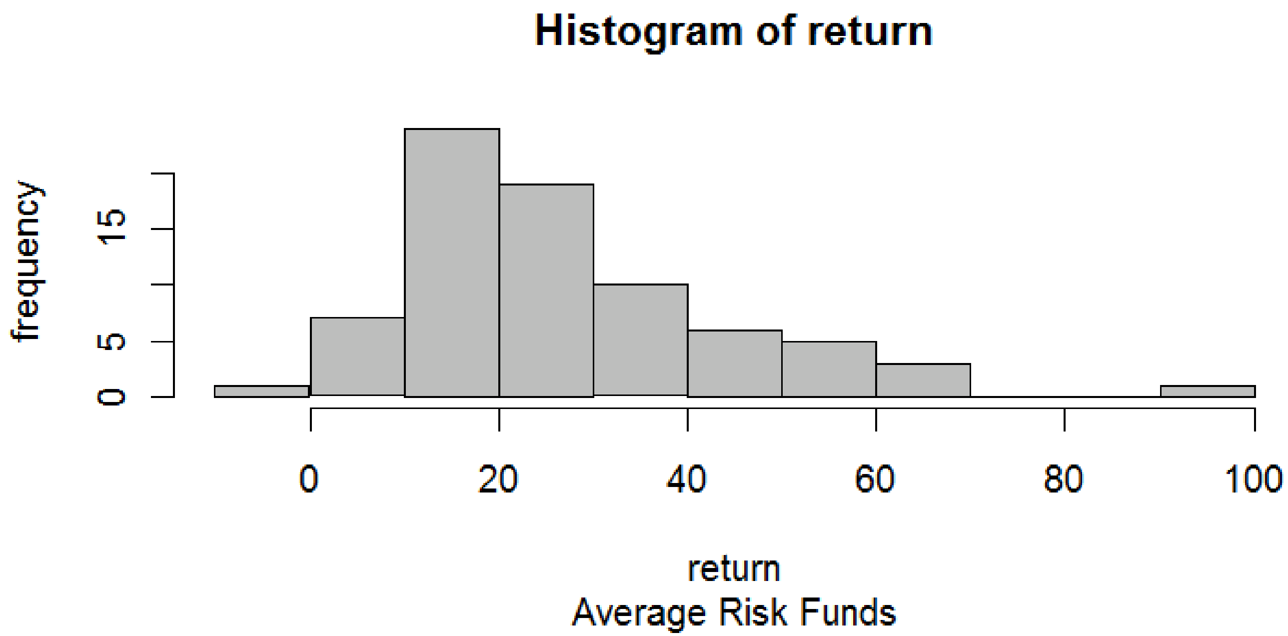


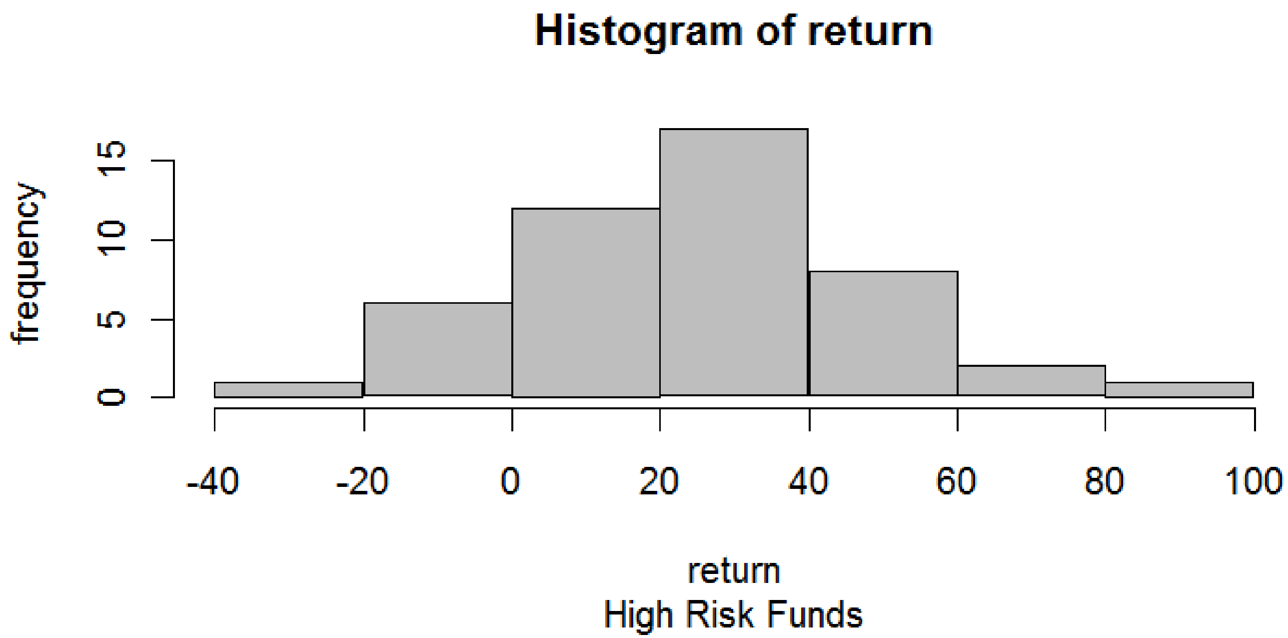
(g) Histogram by groups: R

```
# To specify that 2 graphs in one column in one page
par(mfrow=c(2,1))
```

want plotting device to have 2 rows and 1 column(split plotting device to two parts)

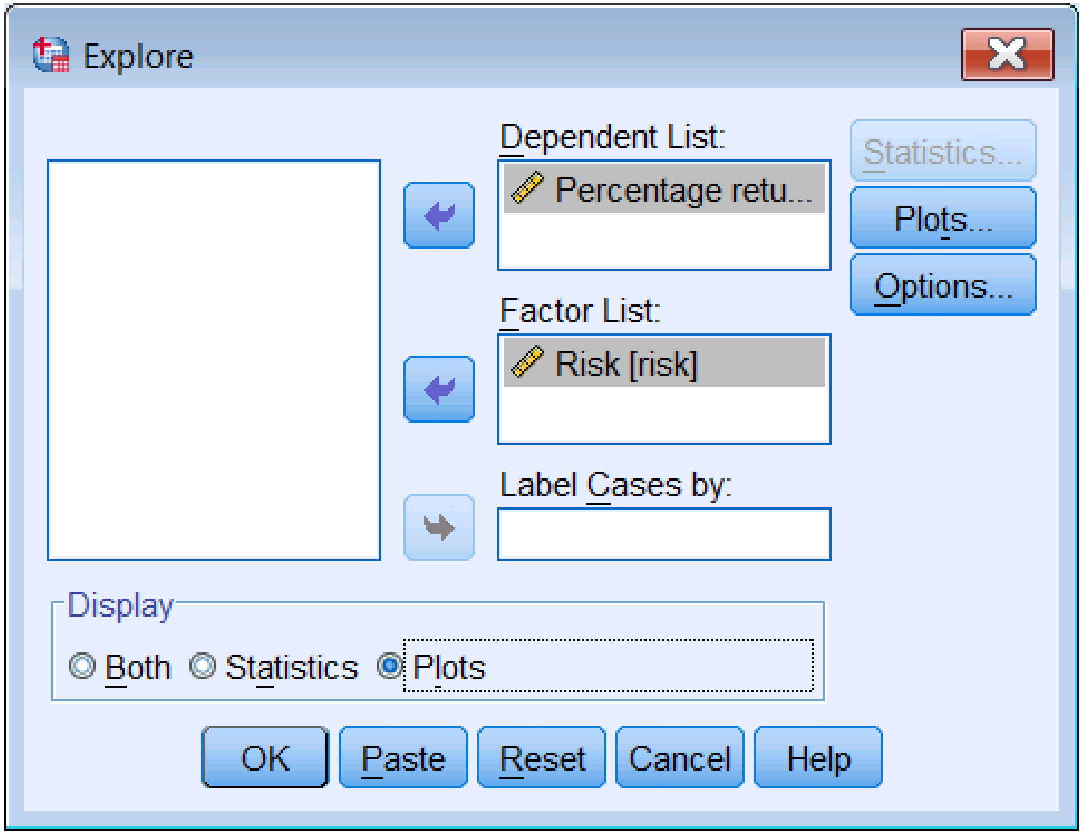
```
#Histogram for the return of the average risk funds
hist(ex5.2ar, include.lowest = TRUE, freq=TRUE,  col="grey",
main = paste("Histogram of Return"), sub=paste ("Average Risk Funds"),
xlab = "return", ylab="frequency", axes = TRUE)

#Histogram for the return of the high risk funds
hist(ex5.2hr, include.lowest = TRUE, freq=TRUE,  col="grey",
main = paste("Histogram of Return"), sub=paste ("High Risk Funds"),
xlab = "return", ylab="frequency", axes = TRUE)

#To get back to 1 graph in one page.
par(mfrow=c(1,1))
```


Histogram of return
frequency
0
5
15
0
20
40
60
80
100
return
Average Risk Funds

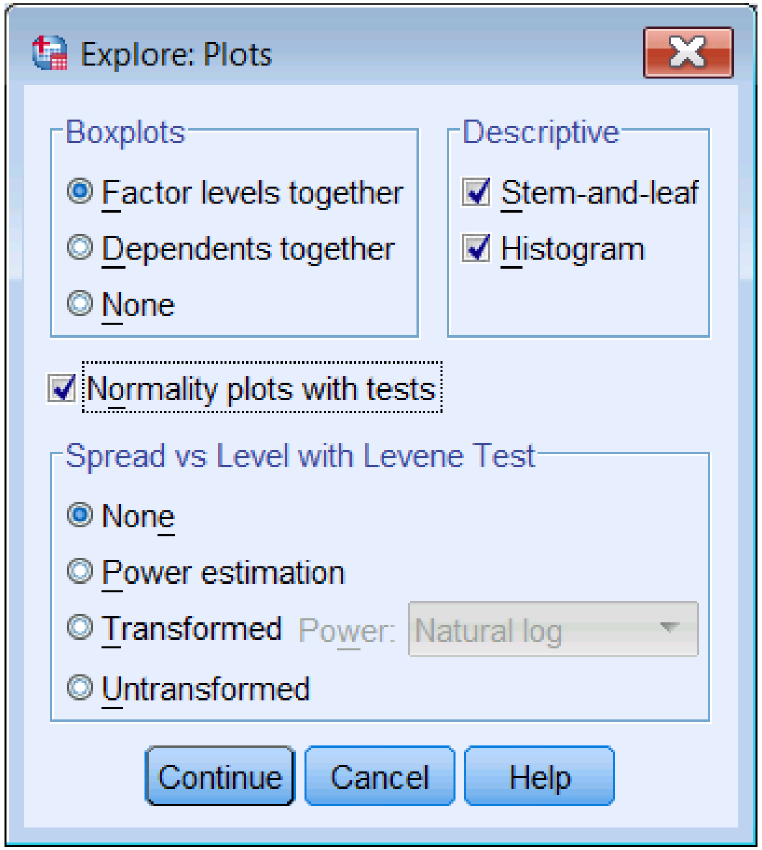

Histogram of return
frequency
0
5
10
15
-40
-20
0
20
40
60
80
100
return
High Risk Funds

(h) Descriptive statistics by groups: SPSS

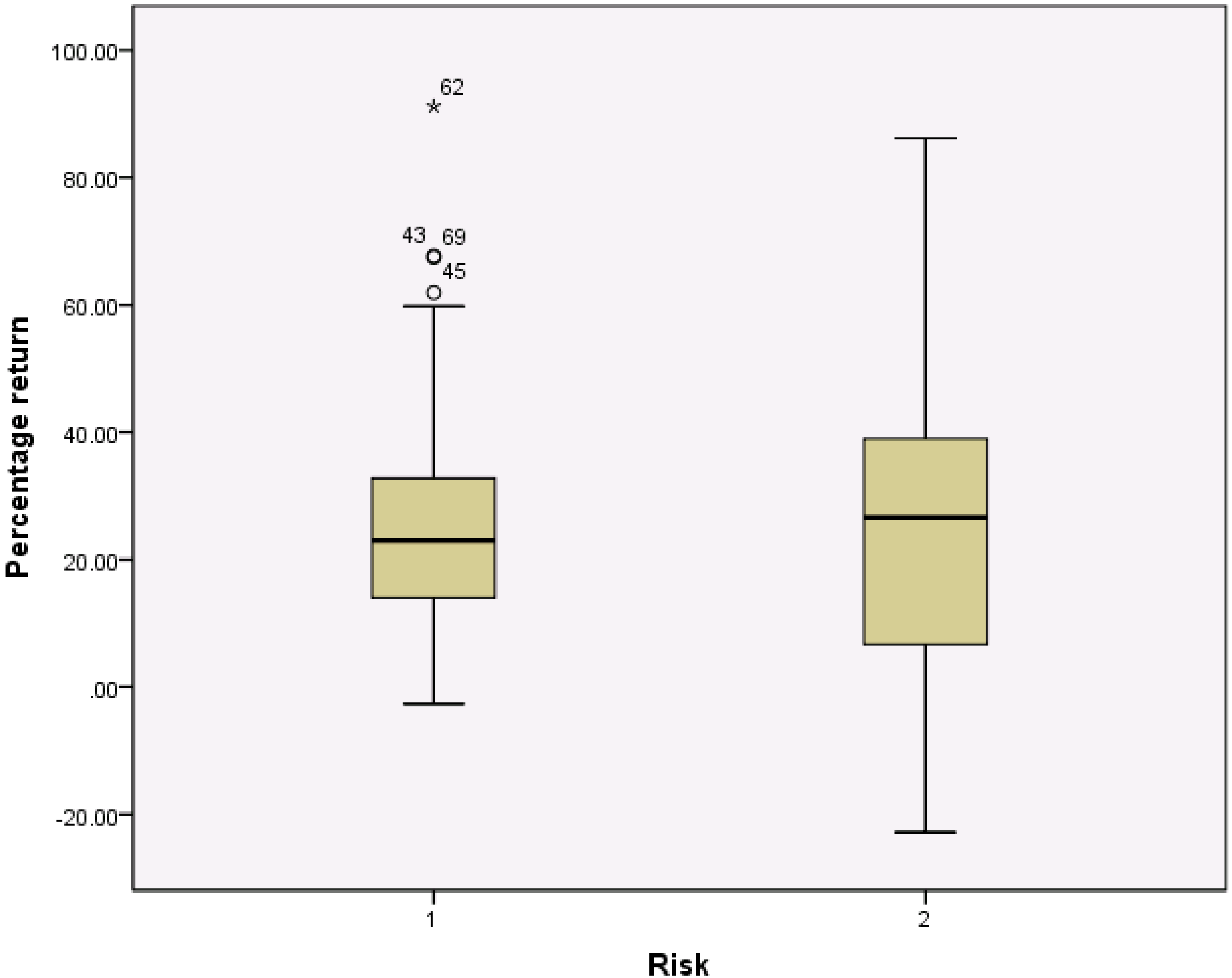
- Import `ex5.2.txt` into the SPSS and call it `ex5.2.sav`.
- "Analyze" → "Descriptive Statistics" → "Explore...".
- Move "preturn" from the left panel to the Dependent List panel.



- Move “risk" from the left panel to the Factor List panel.
- Click “Plots".
- Choose the plots that you want.



- Click “Continue" → “OK".



100.00
80.00
60.00
40.00
20.00
.00
-20.00
Percentage return
62
43
69
45
1
2
Risk

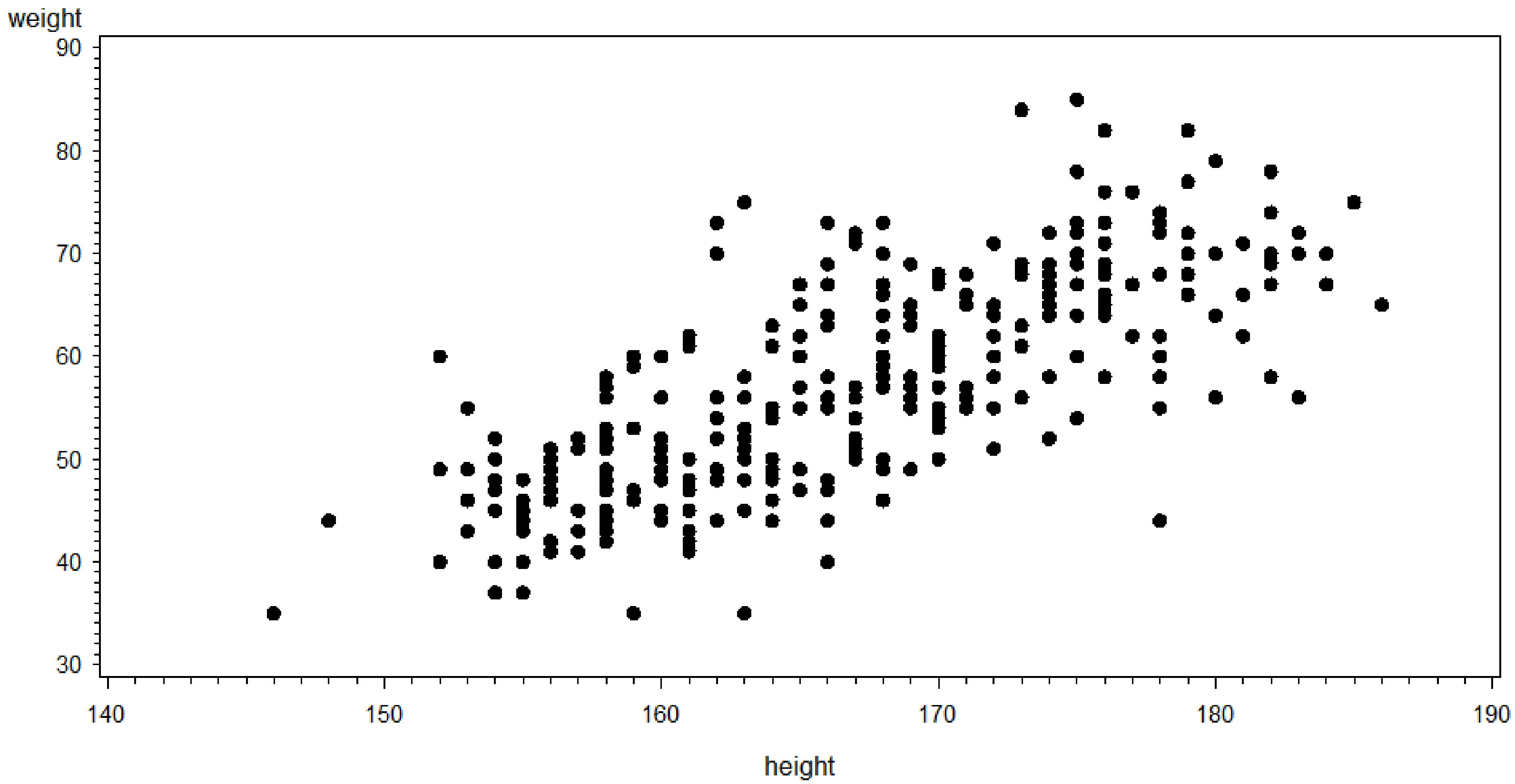
Plotting Bivariate Data

We now look at how we can handle bivariate data.

Plot of bivariate data: SAS

```
data htwt1;
  infile "c:/ST2137/data/htwt1.txt" firstobs=2;
  input id gender $ height weight siblings;
proc gplot data=htwt1;
  title "Scatter Plot of WEIGHT by HEIGHT";
  plot weight*height;
  symbol value = dot color = black;
run;
```

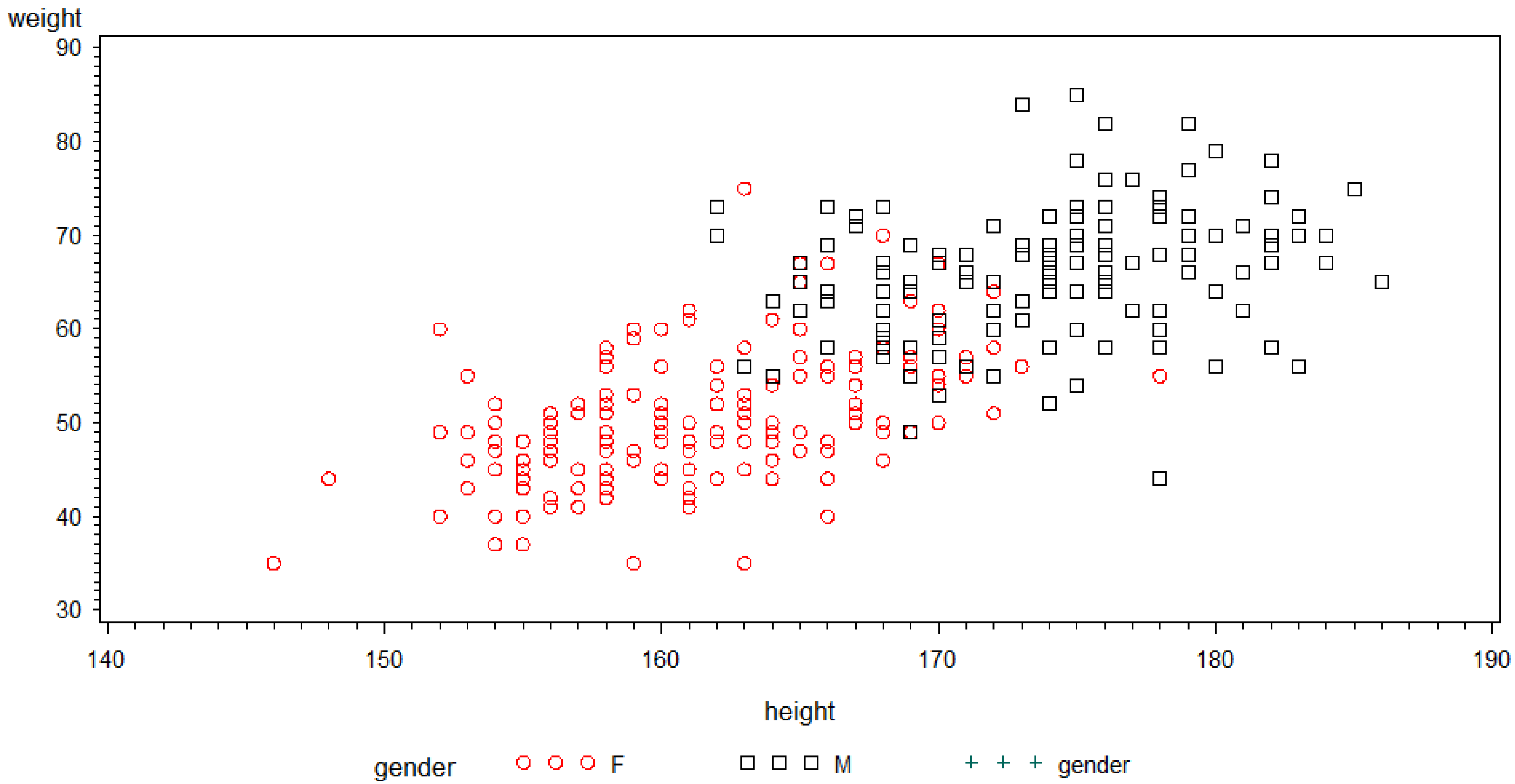
Scatter Plot of WEIGHT by HEIGHT



Plot of bivariate data for groups using different symbols : SAS

```
proc gplot data=htwt1;
  title "Using gender to generate the plotting symbols";
  plot weight*height=gender;
     symbol1 value = circle color=red;
     symbol2 value = square color = black ;
run;
```

Using gender to generate the plotting symbols

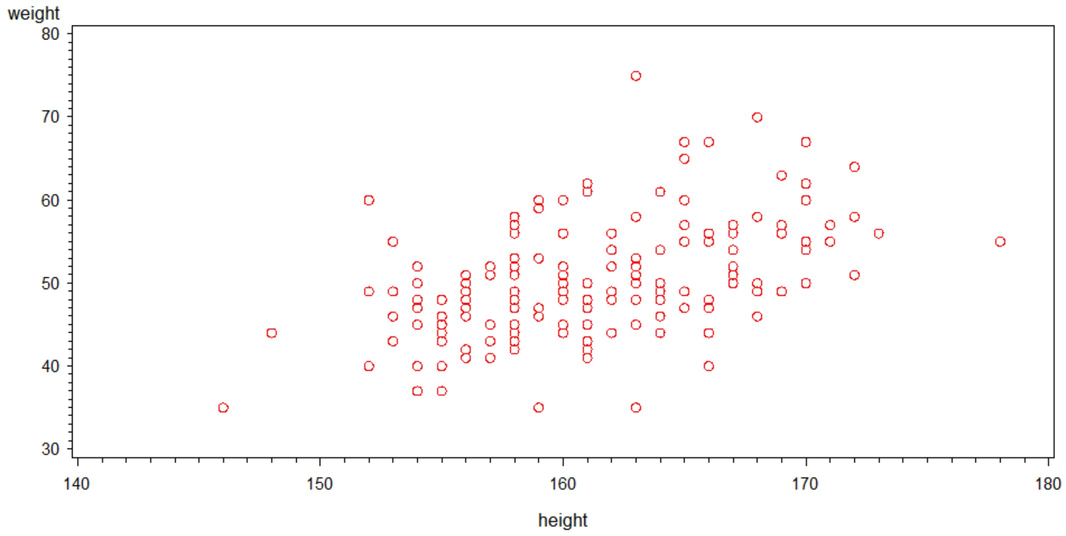


Separate plots by "gender": SAS

```
proc sort data=htwt1;
  by gender;
proc gplot data=htwt1;
  title "Separate Plots by GENDER";
  by gender;
  plot weight*height;
run;
```

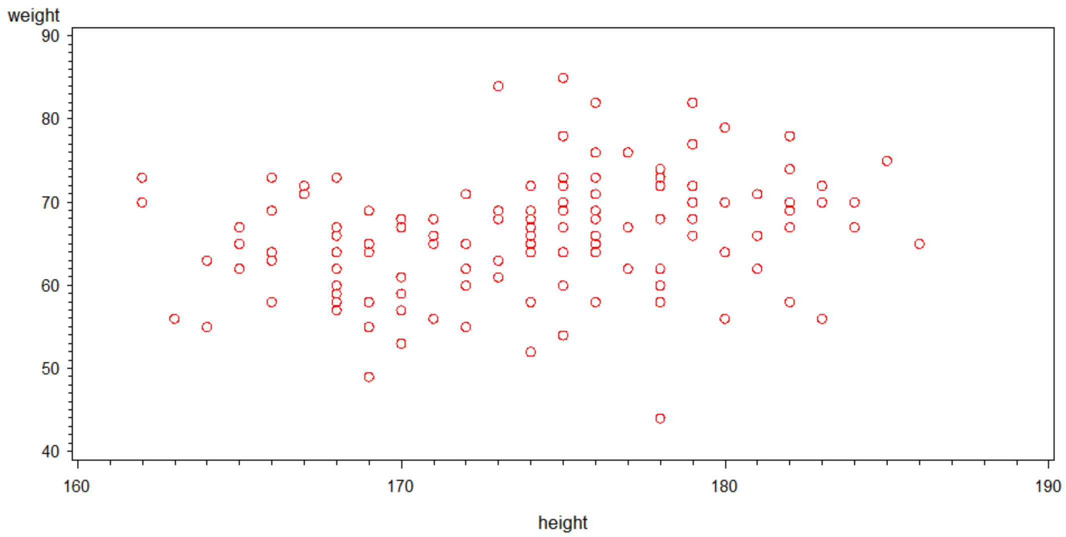
Separate Plots by GENDER

gender=F



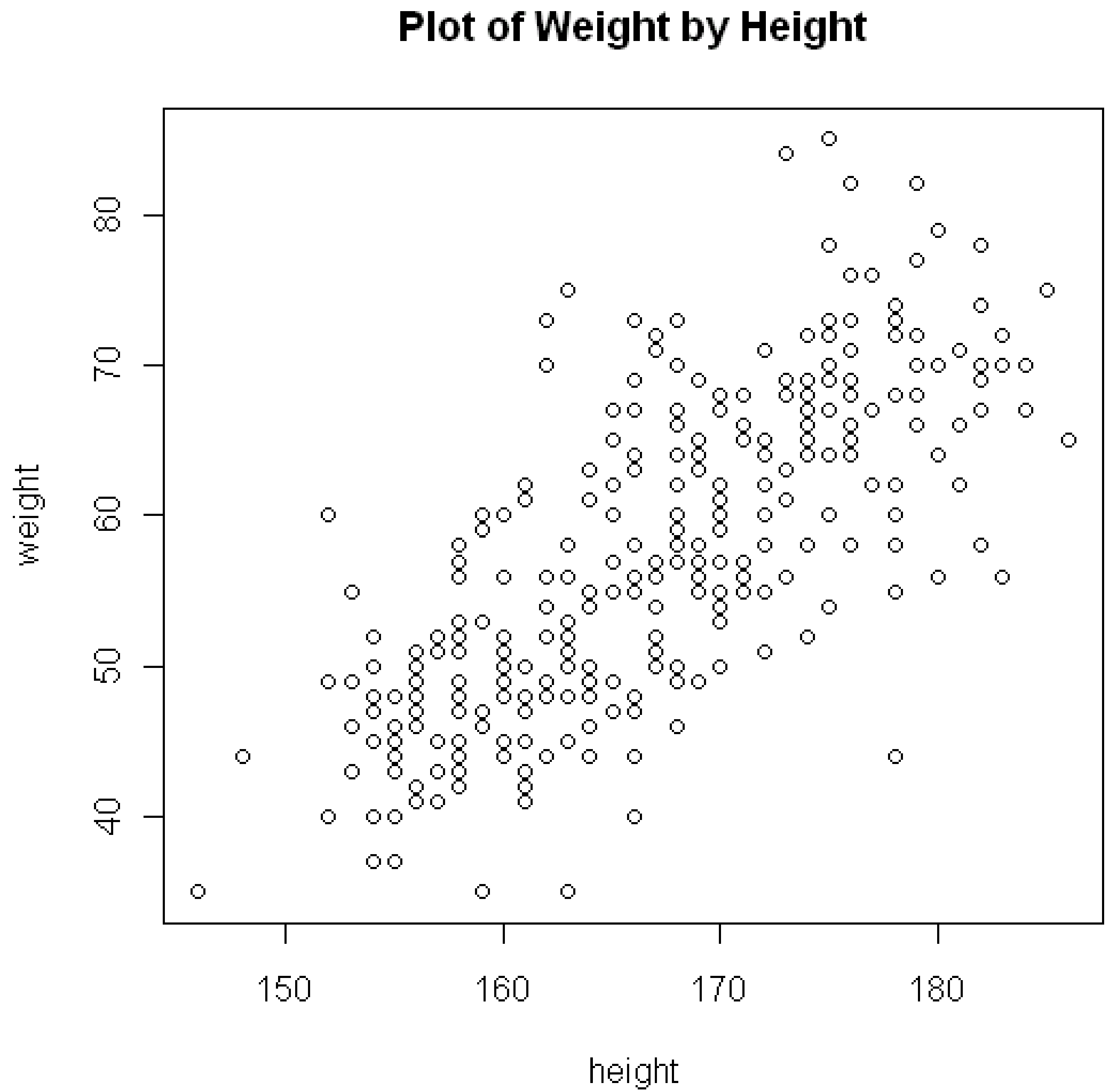
Separate Plots by GENDER

gender=M



Plot of bivariate data: R

```
> htwt1 <- read.table("c:/ST2137/data/htwt1.txt",header=T)
> attach(htwt1)
> plot(height, weight, main="Plot of Weight by Height")
```


Plot of Weight by Height
weight
80
70
60
50
40
150
160
170
180
height

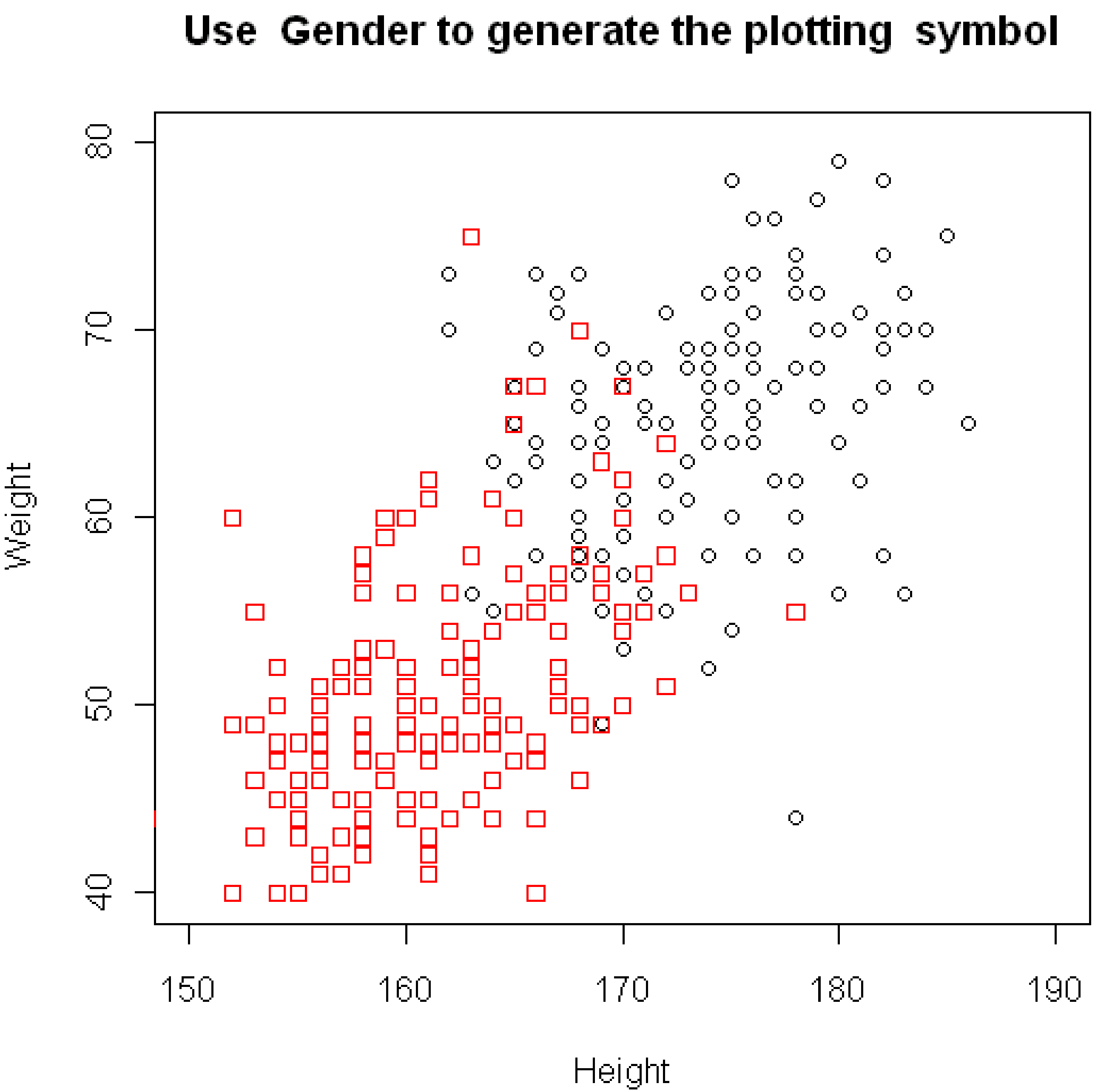
Plot of bivariate data for groups : R

```
> plot(height[gender=="M"],weight[gender=="M"],
+ main="Use Gender to generate the plotting symbol",
+ ylab="Weight",xlab="Height",
+ xlim=c(150,190),ylim=c(40,80))
> par(new=T)
> plot(height[gender=="F"],weight[gender=="F"],
+ main="",xlab="",ylab="",
+ xlim=c(150,190),ylim=c(40,80),
+ axes=F,pch=0,col=2)
```

skip the names since they were described before

```
# try points() for a more elegant solution!
```

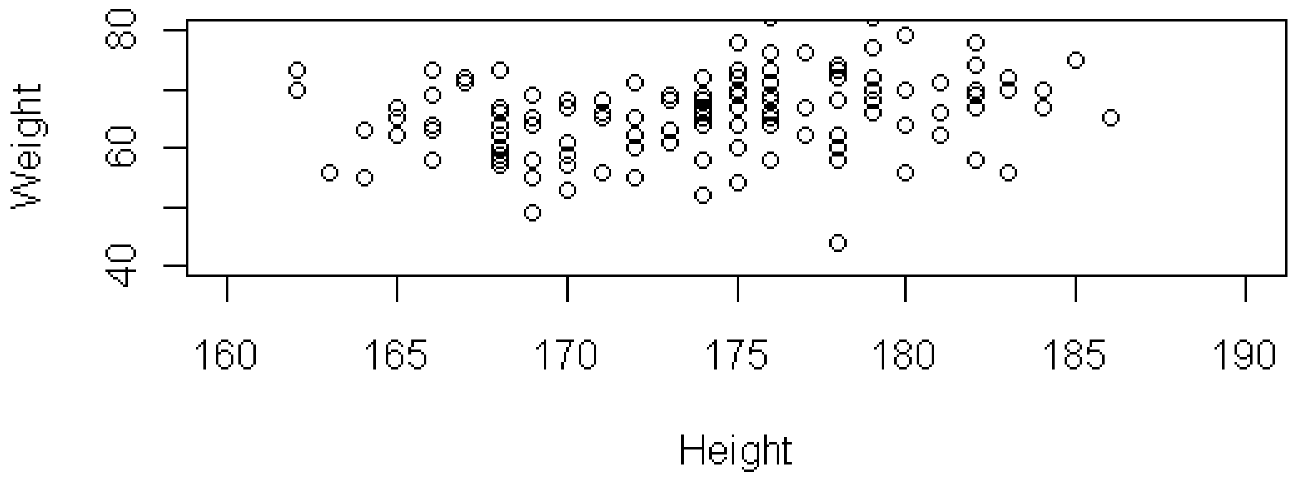
```
plot(height[gender=="M"],weight[gender=="M"],
+ main="Use Gender to generate the plotting symbol",
+ ylab="Weight",xlab="Height",
+ xlim=c(150,190),ylim=c(40,80))

points(height[gender==“F”],weight[gender==“F”],pch=0,col=2)
```


Use Gender to generate the plotting symbol
Weight
80
70
60
50
40
150
160
170
180
190
Height

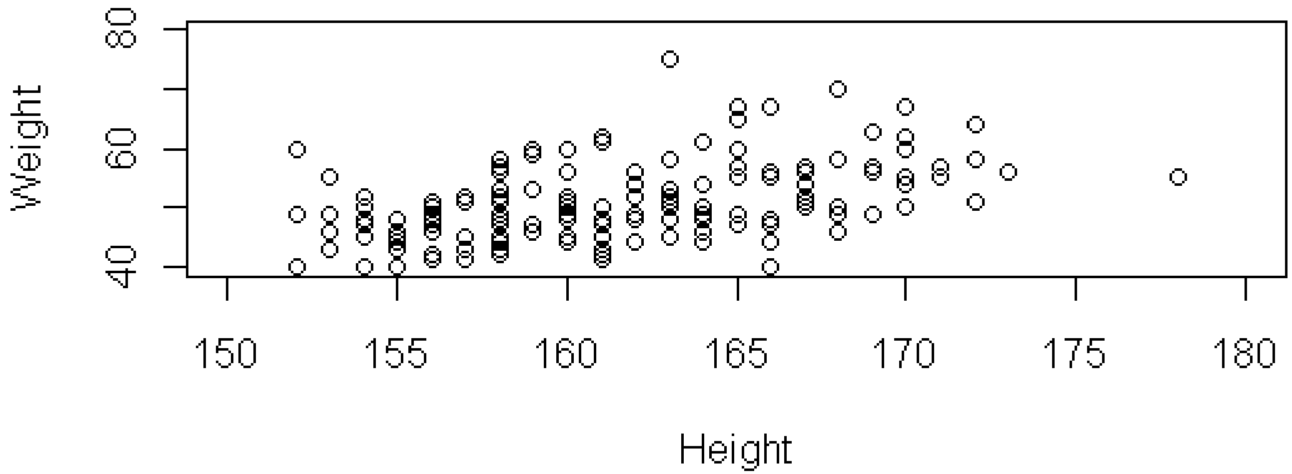
Separate plots by "gender" : R

```
> par(mfrow=c(2,1))
> plot(height[gender=="M"],weight[gender=="M"],
+ main="Plot of Weight by Height for Male",
+ ylab="Weight",xlab="Height",
+ xlim=c(160,190),ylim=c(40,80))
>
> plot(height[gender=="F"],weight[gender=="F"],
+ main="Plot of Weight by Height for Female",
+ ylab="Weight",xlab="Height",
+ xlim=c(140,180),ylim=c(40,80))
> par(mfrow=c(1,1))
```

Plot of Weight by Height for Male

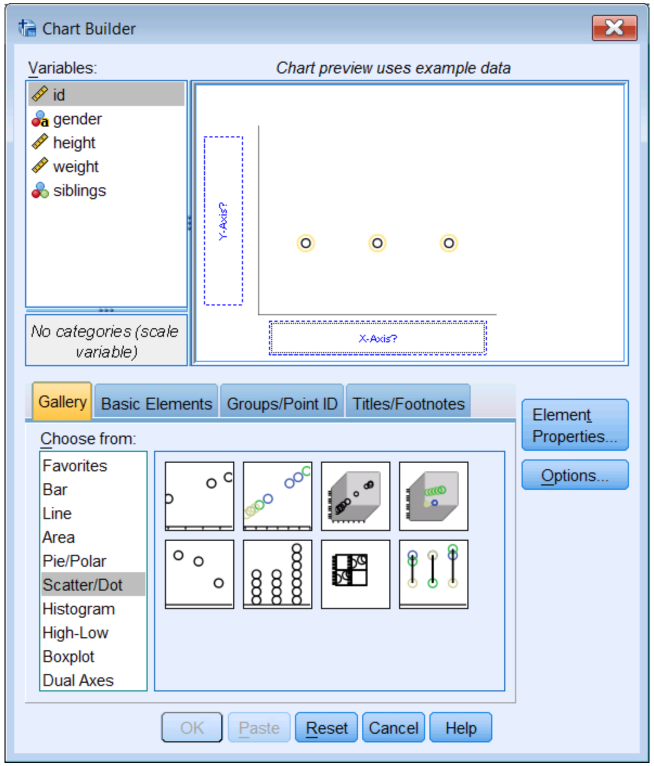


Plot of Weight by Height for Female

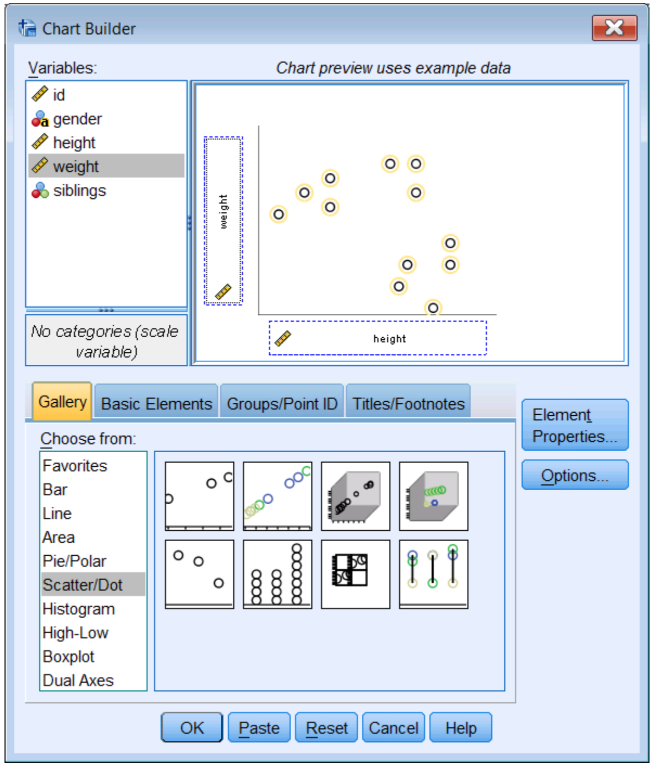


Plot of bivariate data: SPSS

- Open the data set `htwt1.sav`.
- "Graphs" → "Chart Builder".
- Highlight "Scatter/Dot" in the "Choose from" panel.

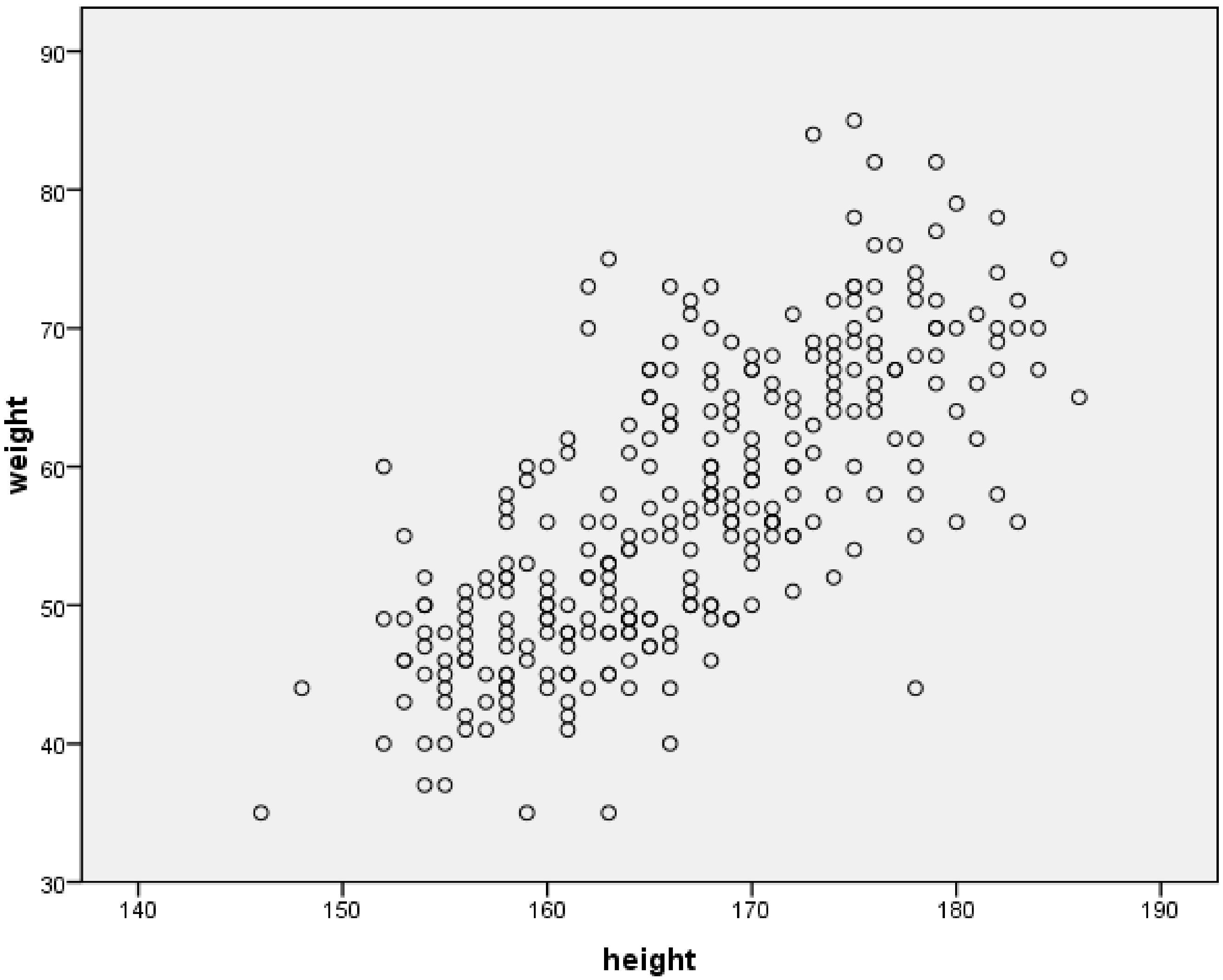


- Double click the simple scatter plot in the (1,1) position.
- Drag `height` in Variable panel to the "X-axis?" box in the chart preview.



- Drag `weight` in Variable panel to the "Y-axis?" box in the chart preview.

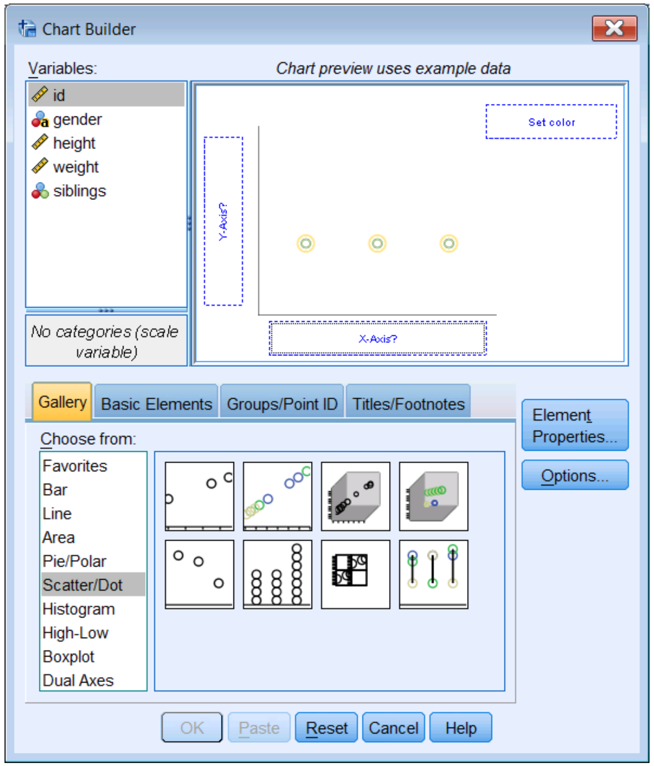
- Click “OK".



Plot of bivariate data for groups : SPSS

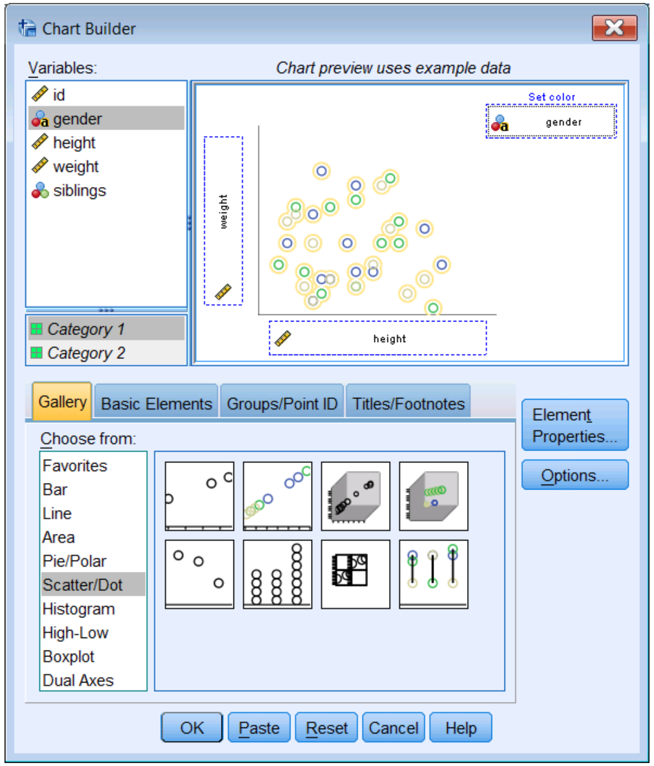
- "Graphs" → "Chart Builder".

- Highlight "Scatter/Dot" in the "Choose from" panel.



- Double click the simple scatter plot in the (1,2) position.

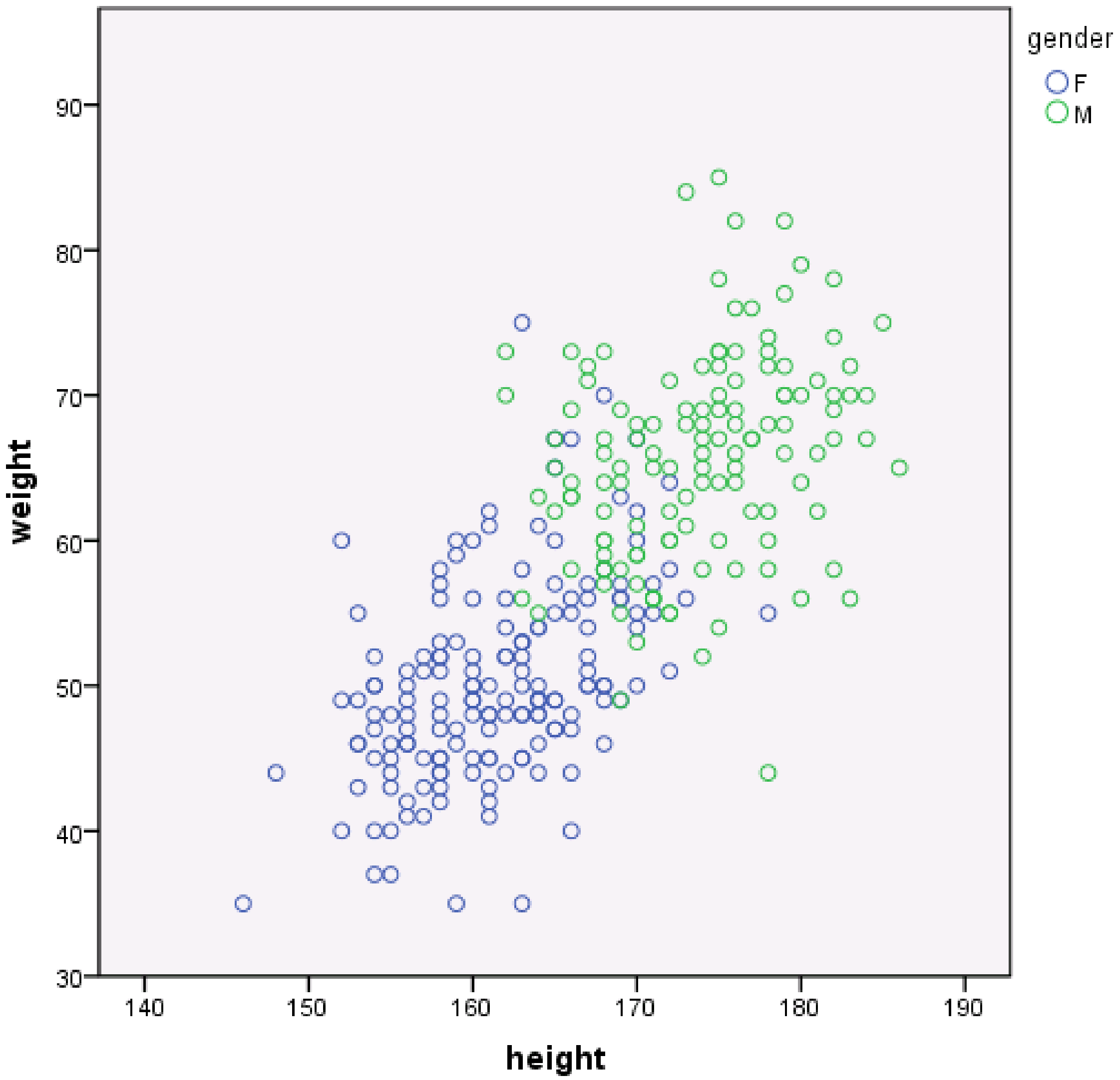
- Drag `height` in Variable panel to the "X-axis?" box in the chart preview.



- Drag `weight` in Variable panel to the "Y-axis?" box in the chart preview.

- Drag `gender` in Variables panel to the "Set color" box.

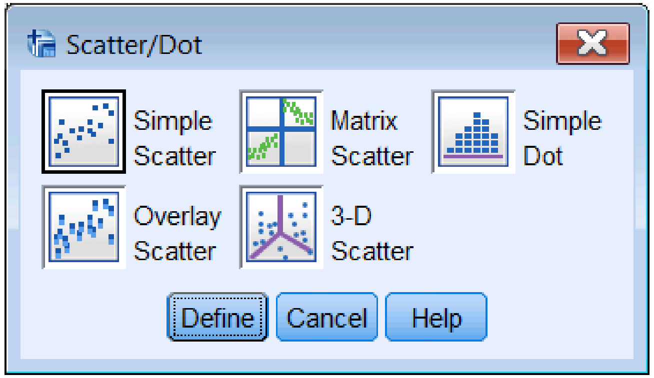
- Click “OK".



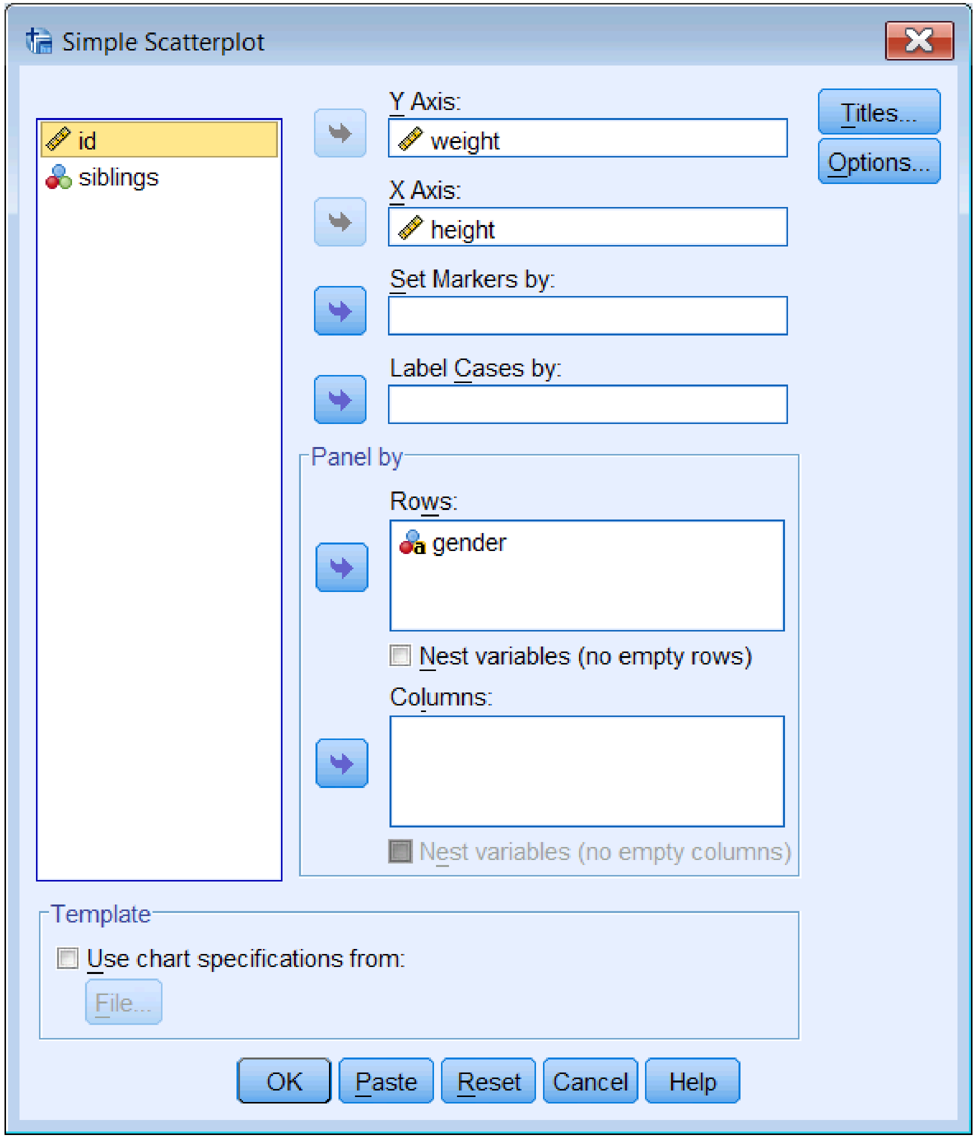
gender
F
M
weight
height
90
80
70
60
50
40
30
140
150
160
170
180
190

Separate plots by "gender": SPSS

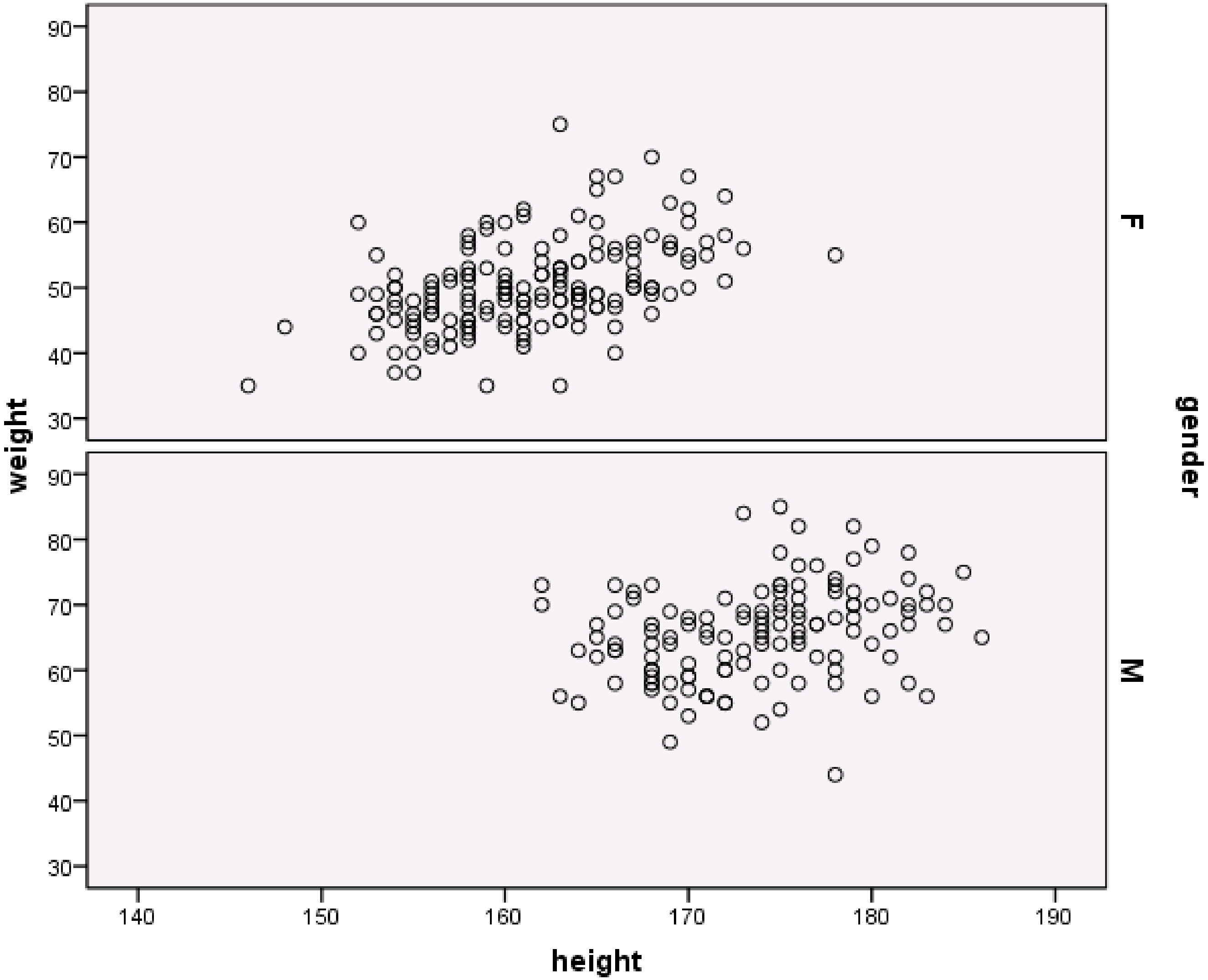
- "Graphs" → "Legacy Dialogs" → "Scatter/Dot".
- Choose "Simple Scatter".



- Click "Define".
- Drag the variable `height` on the left panel to "X-axis:" box.
- Drag the variable `weight` on the left panel to "Y-axis:" box.
- Drag the variable `gender` on the left panel to "Rows:" box in the "Panel by".


Simple Scatterplot
id
siblings
Y Axis:
weight
X Axis:
height
Set Markers by:
Label Cases by:
Titles...
Options...
Panel by
Rows:
gender
Nest variables (no empty rows)
Columns:
Nest variables (no empty columns)
Template
Use chart specifications from:
File...
OK
Paste
Reset
Cancel
Help

- Click “OK".



90
80
70
60
50
40
30
weight
F
gender
M
140
150
160
170
180
190
height