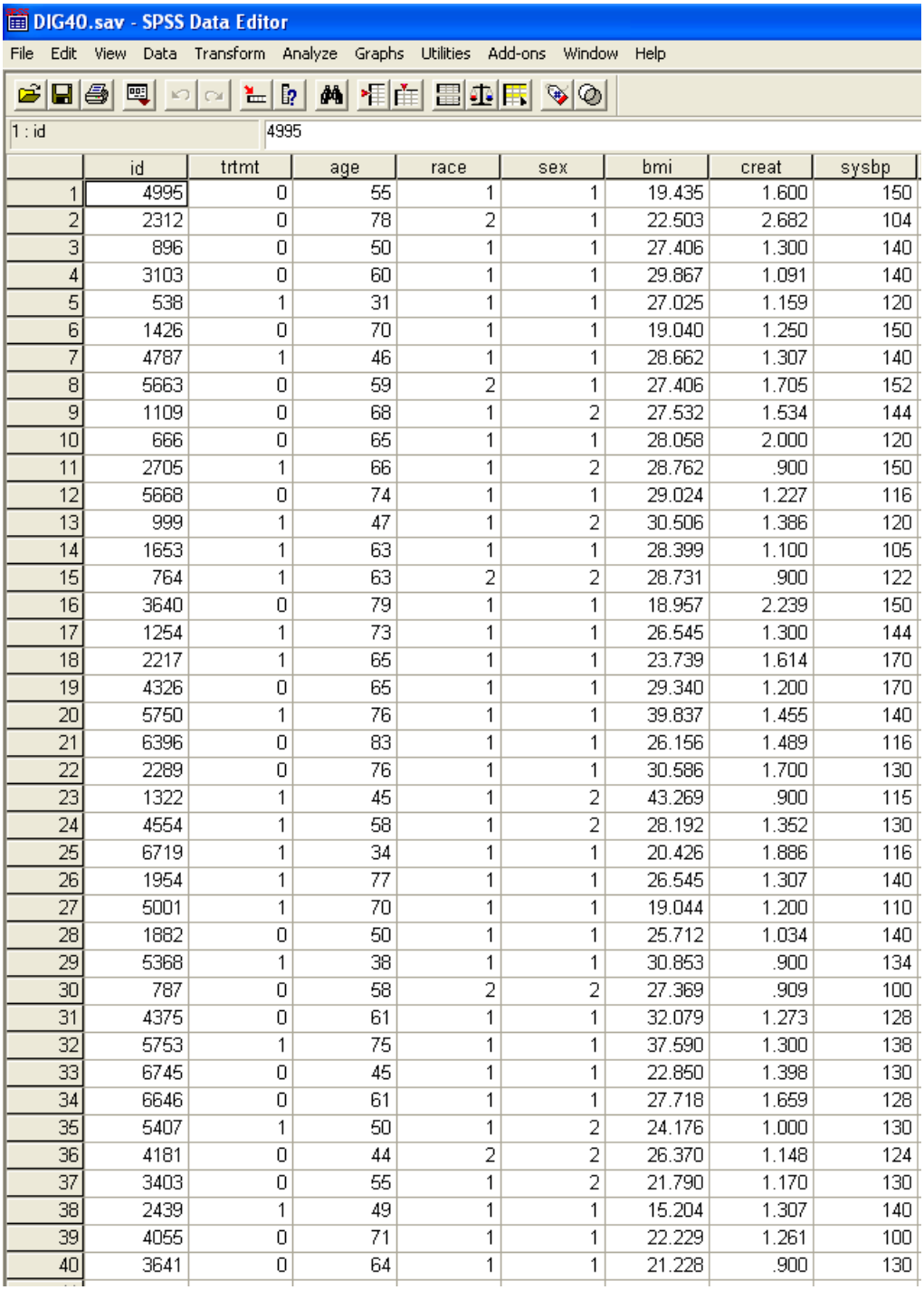
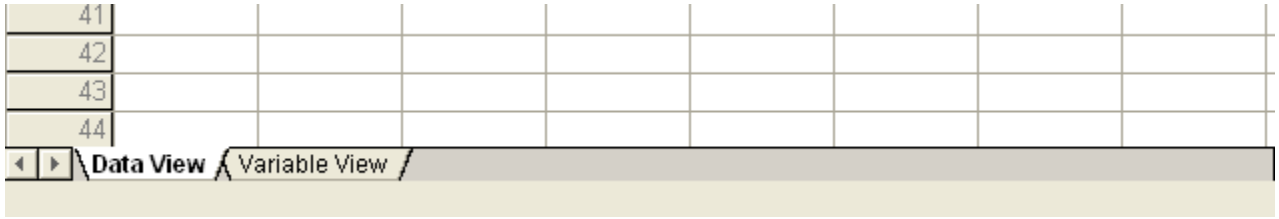
SPSS Program Notes
Biostatistics: A Guide to Design, Analysis, and Discovery Second Edition
by Ronald N. Forthofer, Eun Sul Lee, Mike Hernandez
Chapter 3: Descriptive Statistics

Before discussing the program notes for chapter 3, we present the data on 40 participants from the Digoxin clinical trial shown in Table 3.1. Notice that SPSS displays the data in a spreadsheet format.



	id	trtmt	age	race	sex	bmi	creat	sysbp
1	4995	0	55	1	1	19.435	1.600	150
2	2312	0	78	2	1	22.503	2.682	104
3	896	0	50	1	1	27.406	1.300	140
4	3103	0	60	1	1	29.867	1.091	140
5	538	1	31	1	1	27.025	1.159	120
6	1426	0	70	1	1	19.040	1.250	150
7	4787	1	46	1	1	28.662	1.307	140
8	5663	0	59	2	1	27.406	1.705	152
9	1109	0	68	1	2	27.532	1.534	144
10	666	0	65	1	1	28.058	2.000	120
11	2705	1	66	1	2	28.762	.900	150
12	5668	0	74	1	1	29.024	1.227	116
13	999	1	47	1	2	30.506	1.386	120
14	1653	1	63	1	1	28.399	1.100	105
15	764	1	63	2	2	28.731	.900	122
16	3640	0	79	1	1	18.957	2.239	150
17	1254	1	73	1	1	26.545	1.300	144
18	2217	1	65	1	1	23.739	1.614	170
19	4326	0	65	1	1	29.340	1.200	170
20	5750	1	76	1	1	39.837	1.455	140
21	6396	0	83	1	1	26.156	1.489	116
22	2289	0	76	1	1	30.586	1.700	130
23	1322	1	45	1	2	43.269	.900	115
24	4554	1	58	1	2	28.192	1.352	130
25	6719	1	34	1	1	20.426	1.886	116
26	1954	1	77	1	1	26.545	1.307	140
27	5001	1	70	1	1	19.044	1.200	110
28	1882	0	50	1	1	25.712	1.034	140
29	5368	1	38	1	1	30.853	.900	134
30	787	0	58	2	2	27.369	.909	100
31	4375	0	61	1	1	32.079	1.273	128
32	5753	1	75	1	1	37.590	1.300	138
33	6745	0	45	1	1	22.850	1.398	130
34	6646	0	61	1	1	27.718	1.659	128
35	5407	1	50	1	2	24.176	1.000	130
36	4181	0	44	2	2	26.370	1.148	124
37	3403	0	55	1	2	21.790	1.170	130
38	2439	1	49	1	1	15.204	1.307	140
39	4055	0	71	1	1	22.229	1.261	100
40	3641	0	64	1	1	21.228	.900	130

At the bottom the spreadsheet we find two tabs. The **Data View** tab allows you to view the data in the format shown above while the **Variable View** tab allows you to view the characteristics of each variable.



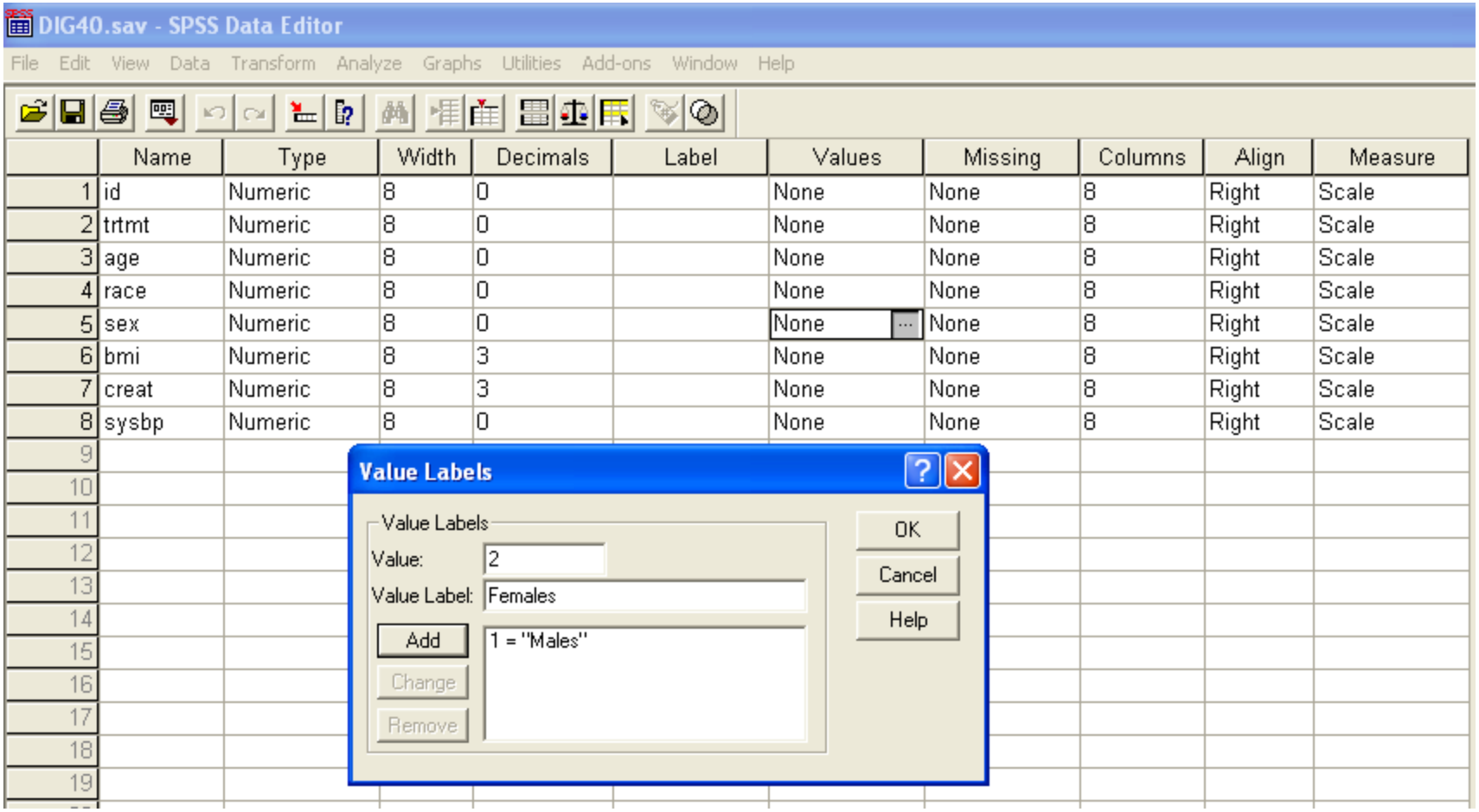
If we click on the **Variable View** tab, we get the following information regarding the eight variables displayed in Table 3.1 .

DIG40.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	8	0		None	None	8	Right	Scale
2	trtmt	Numeric	8	0		None	None	8	Right	Scale
3	age	Numeric	8	0		None	None	8	Right	Scale
4	race	Numeric	8	0		None	None	8	Right	Scale
5	sex	Numeric	8	0		None	None	8	Right	Scale
6	bmi	Numeric	8	3		None	None	8	Right	Scale
7	creat	Numeric	8	3		None	None	8	Right	Scale
8	sysbp	Numeric	8	0		None	None	8	Right	Scale

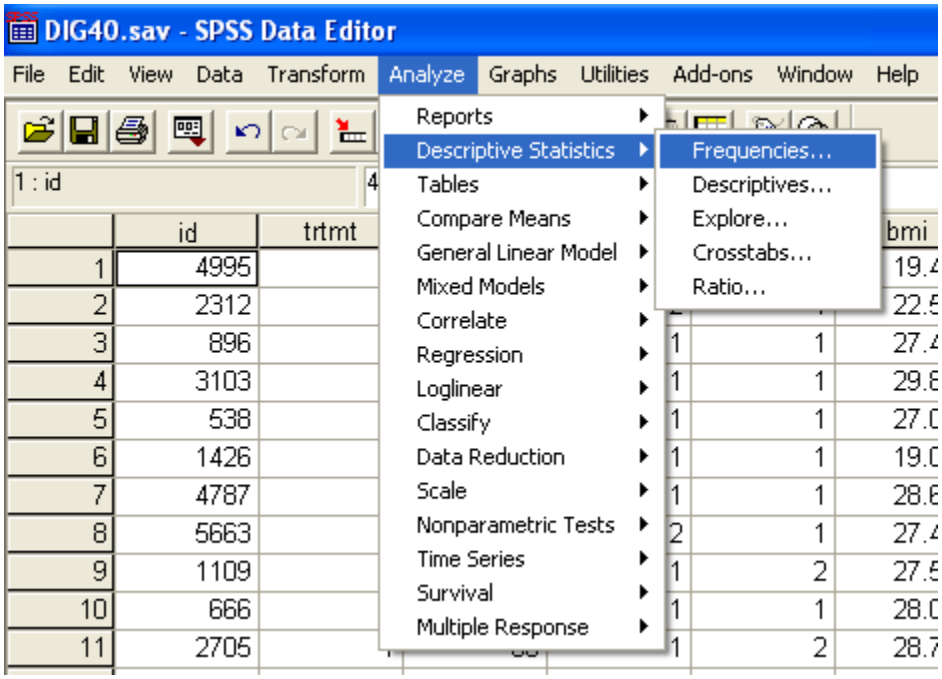
In the **Variable View** worksheet, we can click in the cell referring to the variable sex under **Values** column. The **Value Labels** window will allow you to add the labels “Males” and “Females” to the values 1 and 2 of the variable sex.

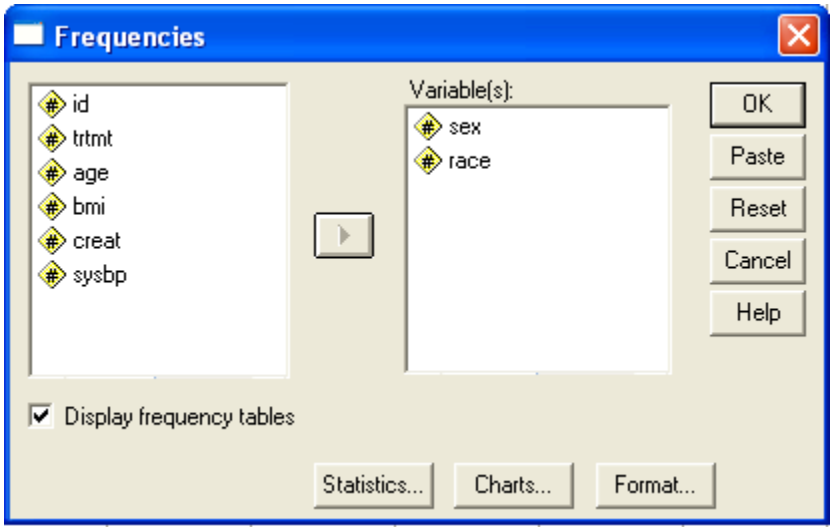


The same procedure can be used to add labels to the values of any other variable. For example, add the labels “White” and “Nonwhite” to the values 1 and 2 of variable race.

Program Note 3.1 – Tabulating data

By displaying the entire DIG40 data set, we are able to see the treatment, age, race, sex and other characteristics of the entire set of forty participants. However, there is also a need to summarize the information in the data set. For example, we may want to know how many males and females are in the DIG40 data set. This goal requires the creation of a one-way table displaying the frequency of males and females.





SPSS output is provided below.

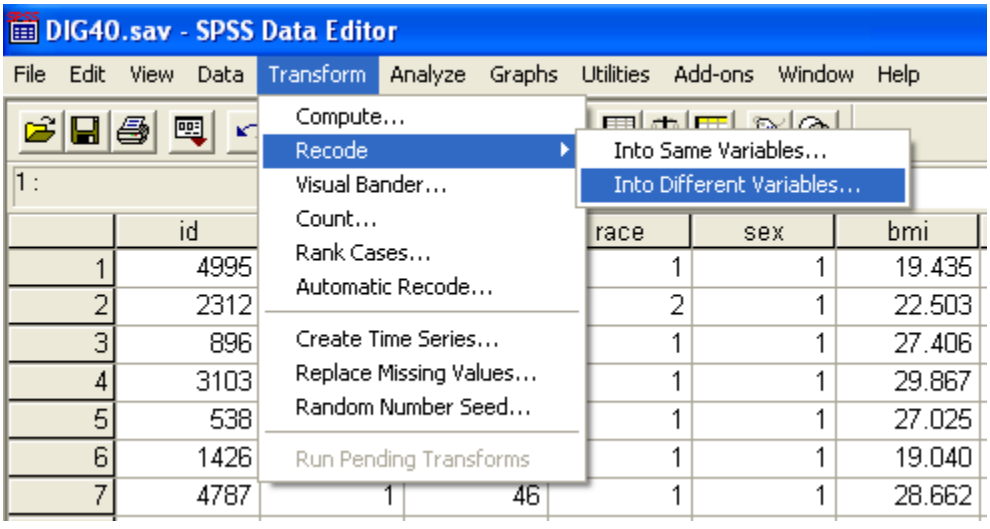
sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Males	30	75.0	75.0	75.0
	Females	10	25.0	25.0	100.0
	Total	40	100.0	100.0	

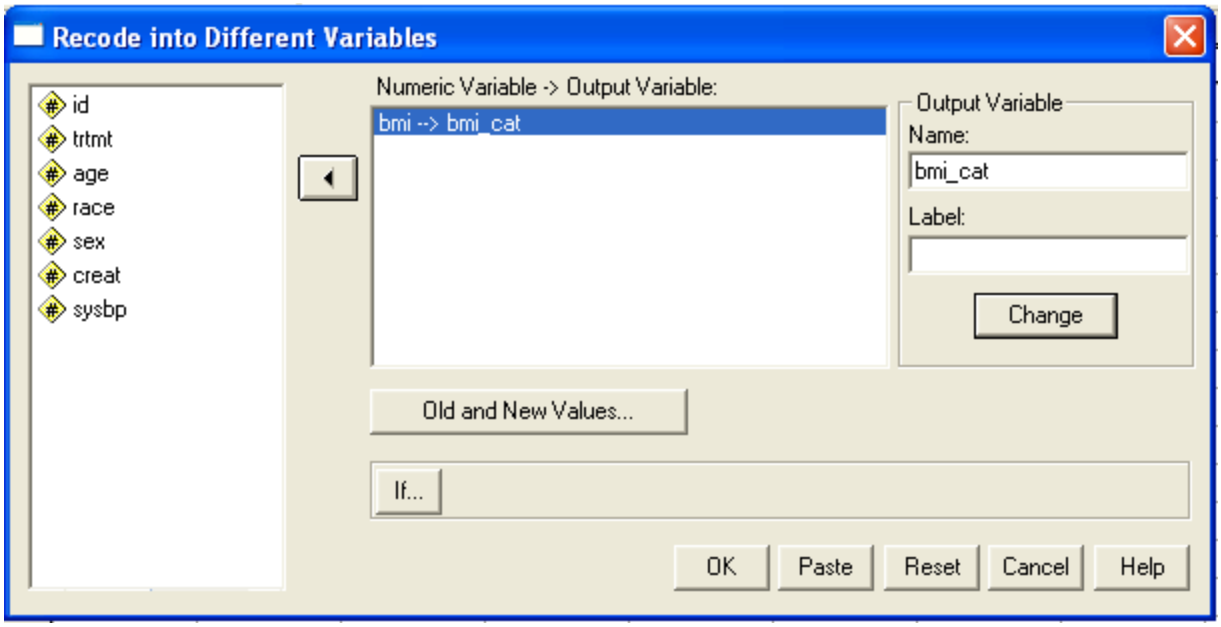
race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	White	35	87.5	87.5	87.5
	Nonwhite	5	12.5	12.5	100.0
	Total	40	100.0	100.0	

In some cases, we may need to create a categorical variable from a continuous variable. For example in Table 3.4, the continuous variable bmi is presented as a categorical variable. We will create a new variable named bmi_cat that is equal to 0 if bmi < 18.5 kg/m^2, 1 if 18.5 kg/m^2 ≤ bmi < 25 kg/m^2, 2 if 25 kg/m^2 ≤ bmi < 30 kg/m^2, and 3 if bmi ≥ 30. To do this use **Transform** > **Recode** > **Into Different Varaibles…** as shown below.

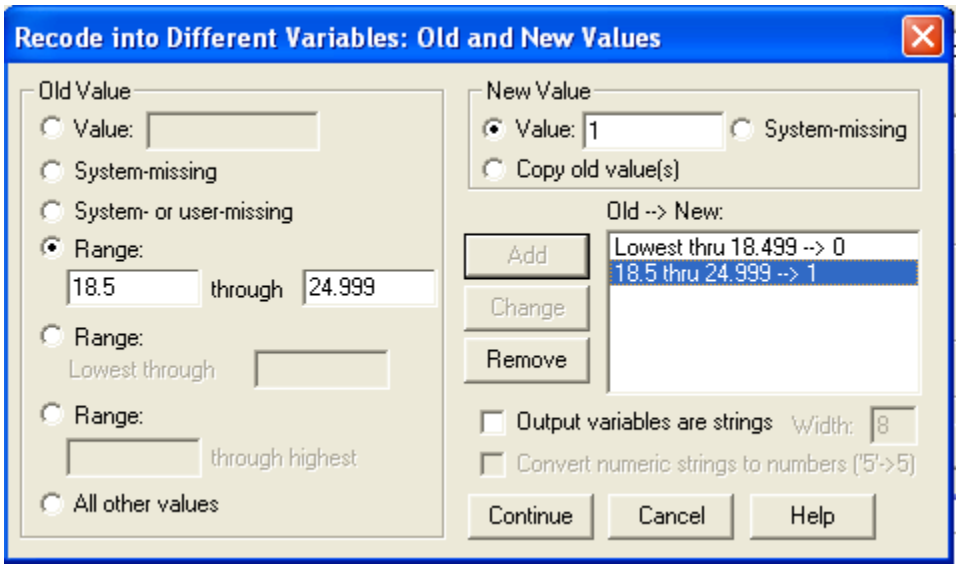


Once the **Recode into Different Variables** window appears type bmi_cat as the name of the new variable in the box under **Output Variable > Name:** then click on the change button below.

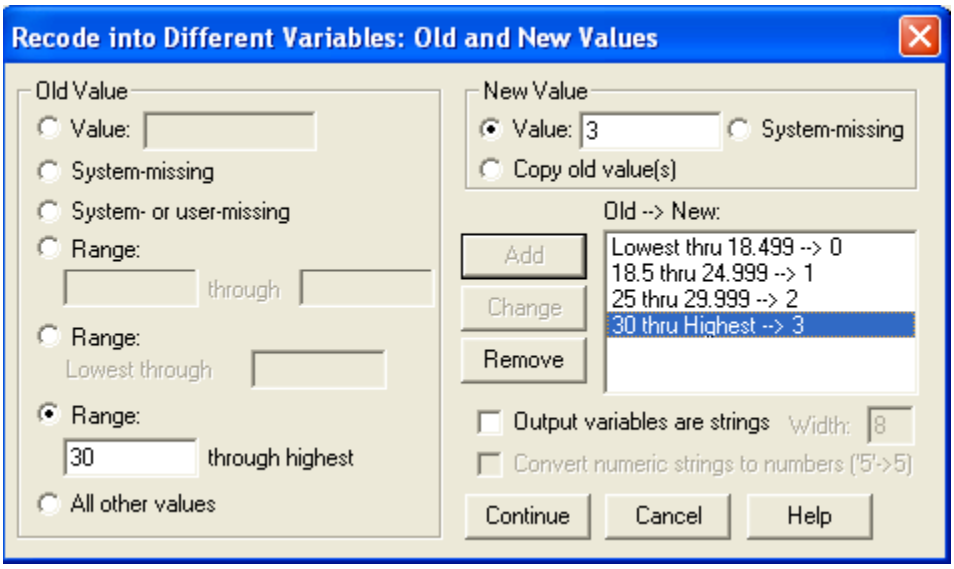


At this point just click on the **Old and New Values…** button and the widow below should appear. Recall that the objective was to create the new variable bmi_cat take on a value of 0 for bmi < 18.5 kg/m^2. Under **Old Value**, we bullet **Range: Lowest through** and then enter 18.499 in the box since we want to include all values less than 18.5. Under **New Value**, we bullet **Value:** and enter 0 in the box. After clicking on the **Add** button, we should get the widow below.

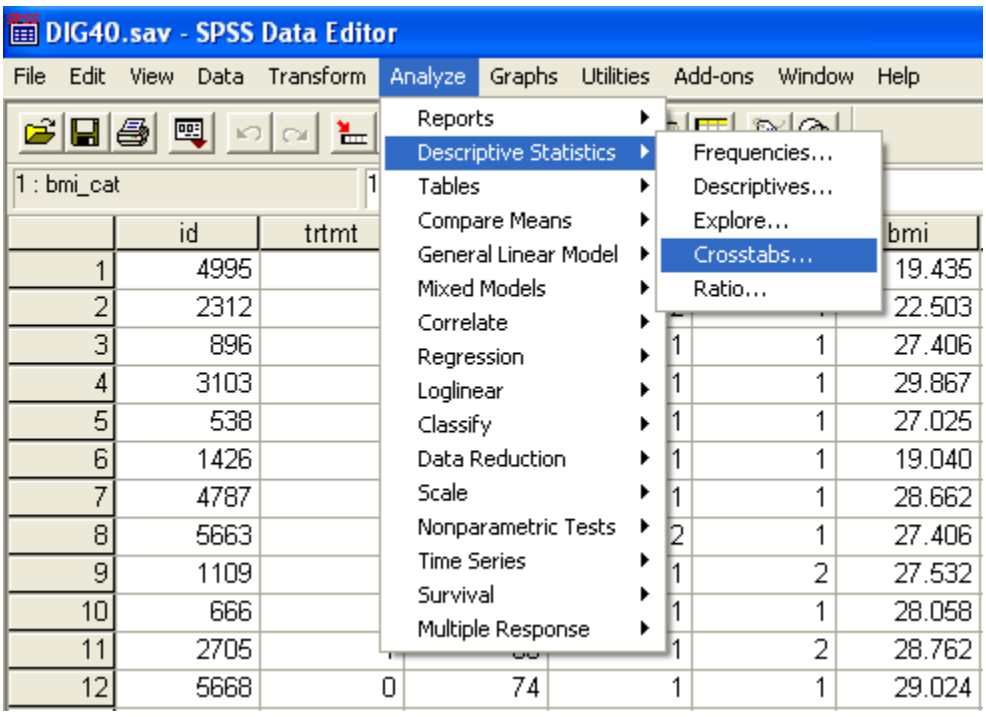
The next widow demonstrates how to assign bmi_cat a value of 1 when 18.5 kg/m^2 ≤ bmi < 25 kg/m^2.



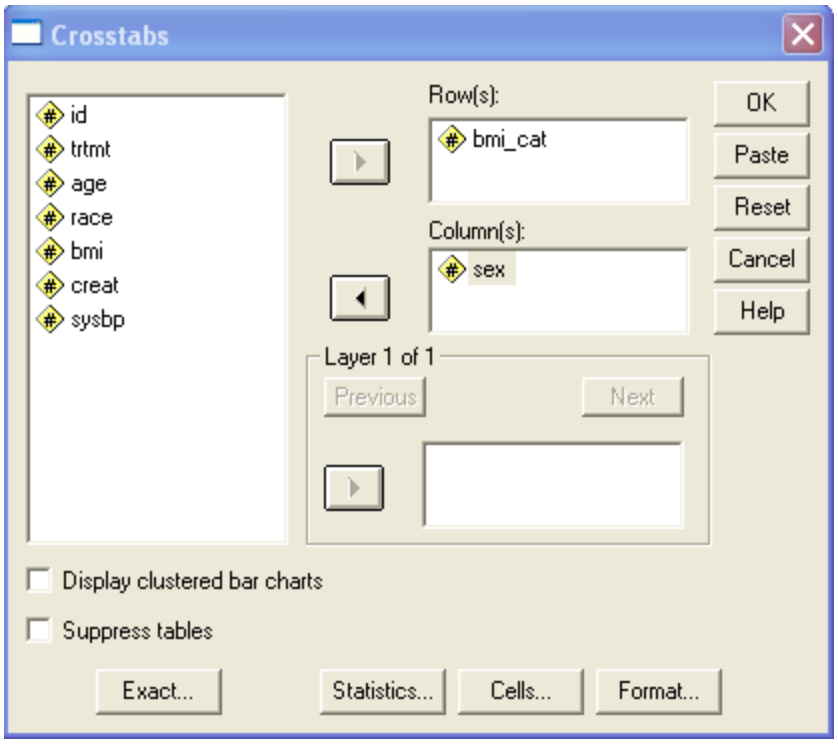
Finally, we assign bmi_cat a value of 3 when bmi ≥ 30.



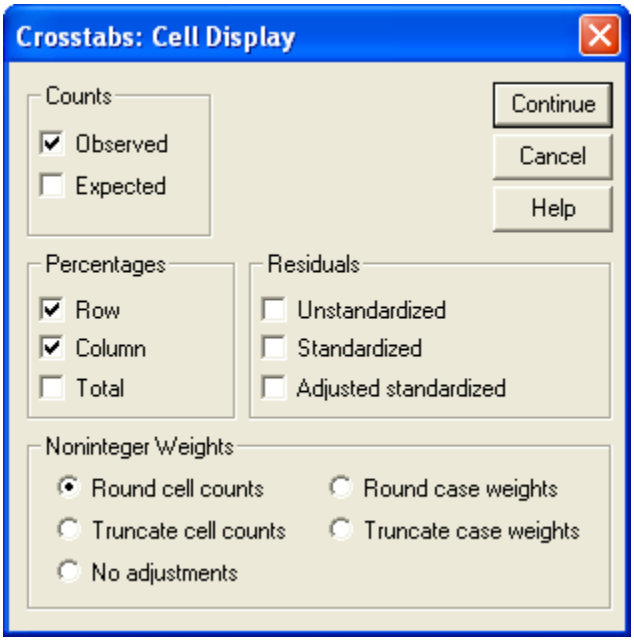
After clicking on **Continue** and **OK**, you should have a spreadsheet that contains the variable bmi_cat. Use Analyze > Descriptive Statistics > Crosstabs… to obtain a cross-tabulation of body mass index by sex.



Once the **Crosstabs** window appears, we want body mass index categories along the rows and the values for sex along the columns.



If you click on the **Cells…** box at the bottom of the **Crosstabs** window, the window below appears.



SPSS output is provided below.

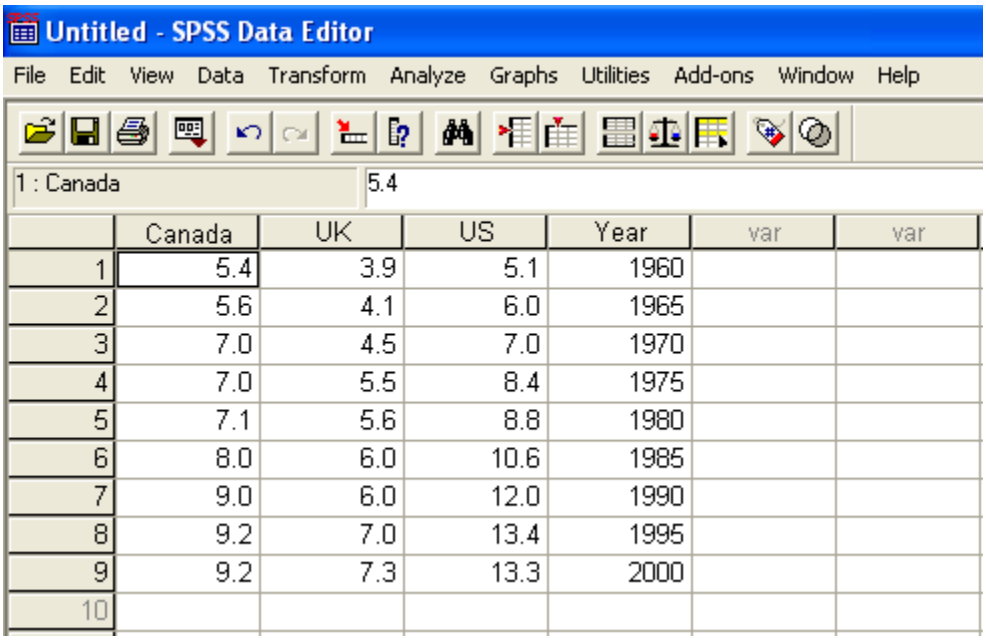
bmi_cat * sex Crosstabulation

			sex		
			Males	Females	Total
bmi_cat					
	.00	Count	1	0	1
		% within bmi_cat	100.0%	.0%	100.0%
		% within sex	3.3%	.0%	2.5%
	1.00	Count	10	2	12
		% within bmi_cat	83.3%	16.7%	100.0%
		% within sex	33.3%	20.0%	30.0%
	2.00	Count	14	6	20
		% within bmi_cat	70.0%	30.0%	100.0%
		% within sex	46.7%	60.0%	50.0%
	3.00	Count	5	2	7
		% within bmi_cat	71.4%	28.6%	100.0%
% within sex		16.7%	20.0%	17.5%	
Total		Count	30	10	40
		% within bmi_cat	75.0%	25.0%	100.0%
		% within sex	100.0%	100.0%	100.0%

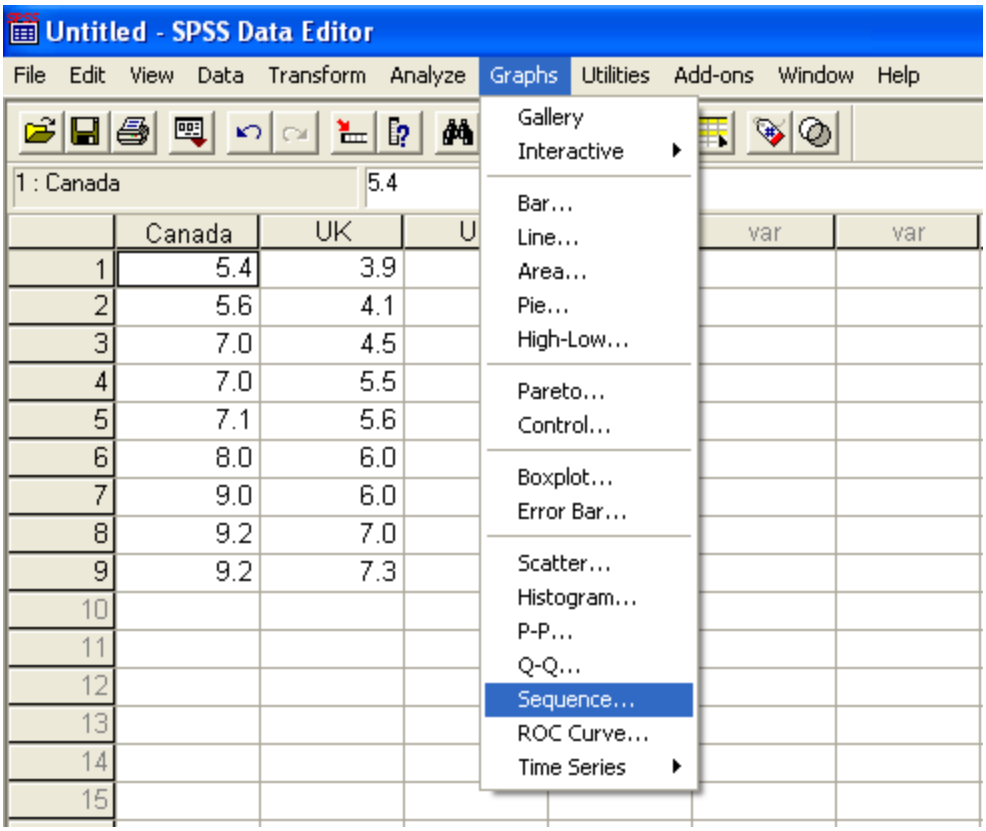
Program Note 3.2 – Creating Line graphs and Bar charts

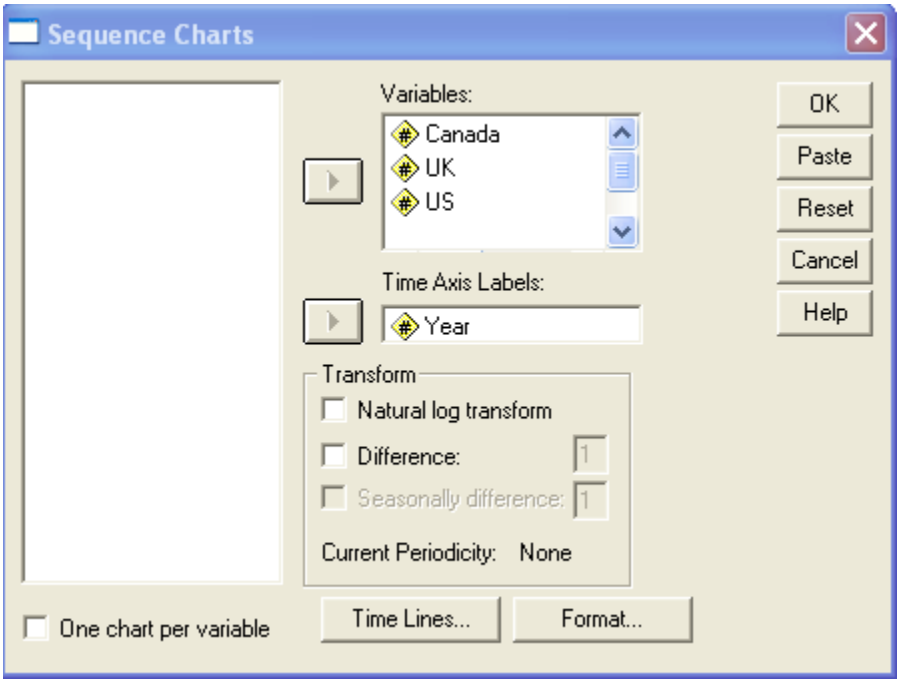
1. Line graphs

In Table 3.6, we present health expenditures data as a percentage of GDP by year for Canada, the United Kingdom, and the United States. The data are entered as shown below.

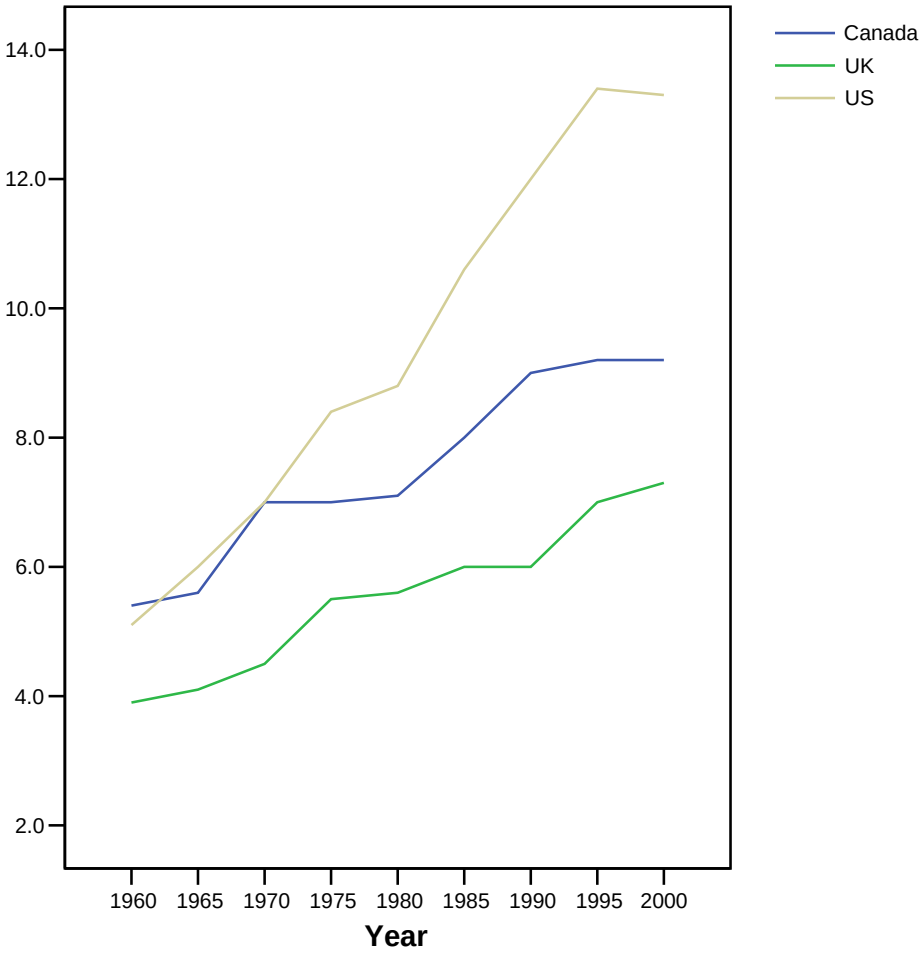


	Canada	UK	US	Year	var	var
1	5.4	3.9	5.1	1960		
2	5.6	4.1	6.0	1965		
3	7.0	4.5	7.0	1970		
4	7.0	5.5	8.4	1975		
5	7.1	5.6	8.8	1980		
6	8.0	6.0	10.6	1985		
7	9.0	6.0	12.0	1990		
8	9.2	7.0	13.4	1995		
9	9.2	7.3	13.3	2000		
10						


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Gallery
Interactive
Bar...
Line...
Area...
Pie...
High-Low...
Pareto...
Control...
Boxplot...
Error Bar...
Scatter...
Histogram...
P-P...
Q-Q...
Sequence...
ROC Curve...
Time Series
1 : Canada
5.4
Canada UK U var var
1 5.4 3.9
2 5.6 4.1
3 7.0 4.5
4 7.0 5.5
5 7.1 5.6
6 8.0 6.0
7 9.0 6.0
8 9.2 7.0
9 9.2 7.3
10
11
12
13
14
15

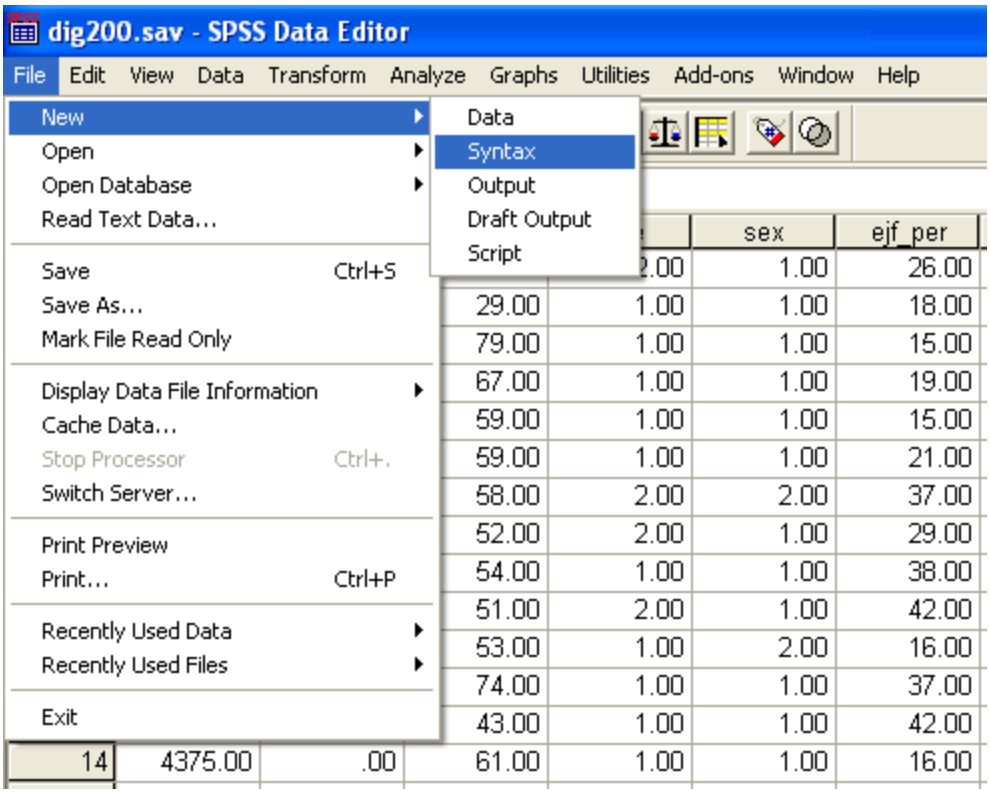

Sequence Charts
Variables:
Canada
UK
US
Time Axis Labels:
Year
OK
Paste
Reset
Cancel
Help
Transform
Natural log transform
Difference:
Seasonally difference:
Current Periodicity: None
One chart per variable
Time Lines...
Format...

SPSS output is provided below.

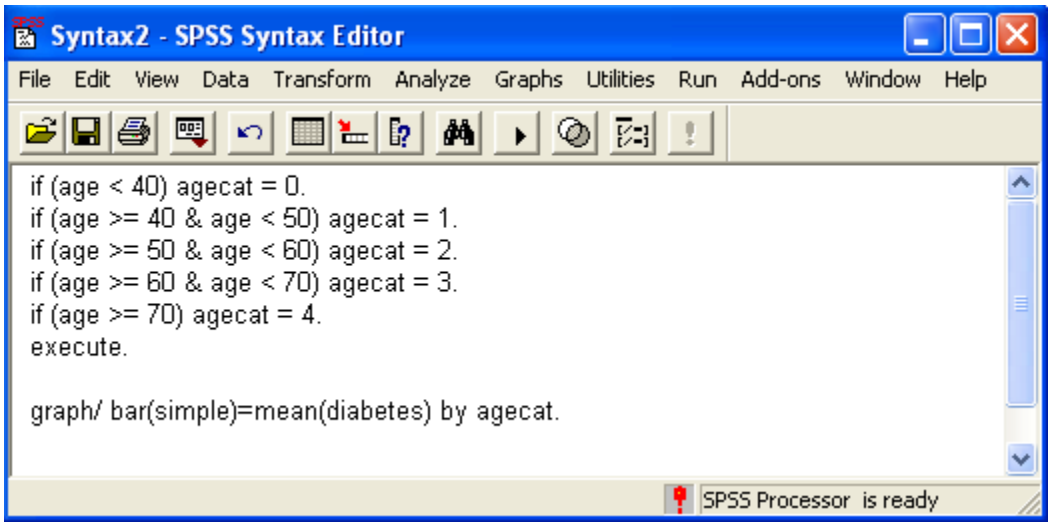


2. Bar charts

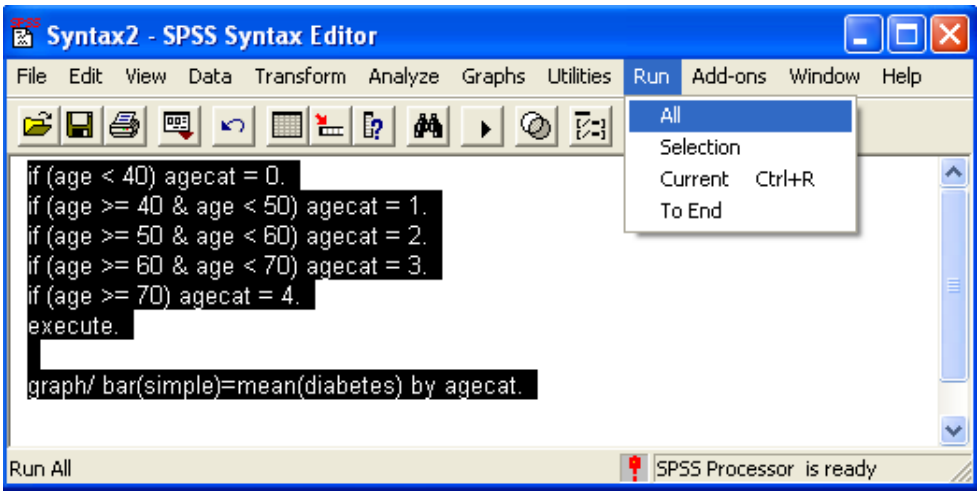
For example, the horizontal bar chart in Figure 3.5 displays the proportion of diabetes by age group for individuals in the DIG200 data set.



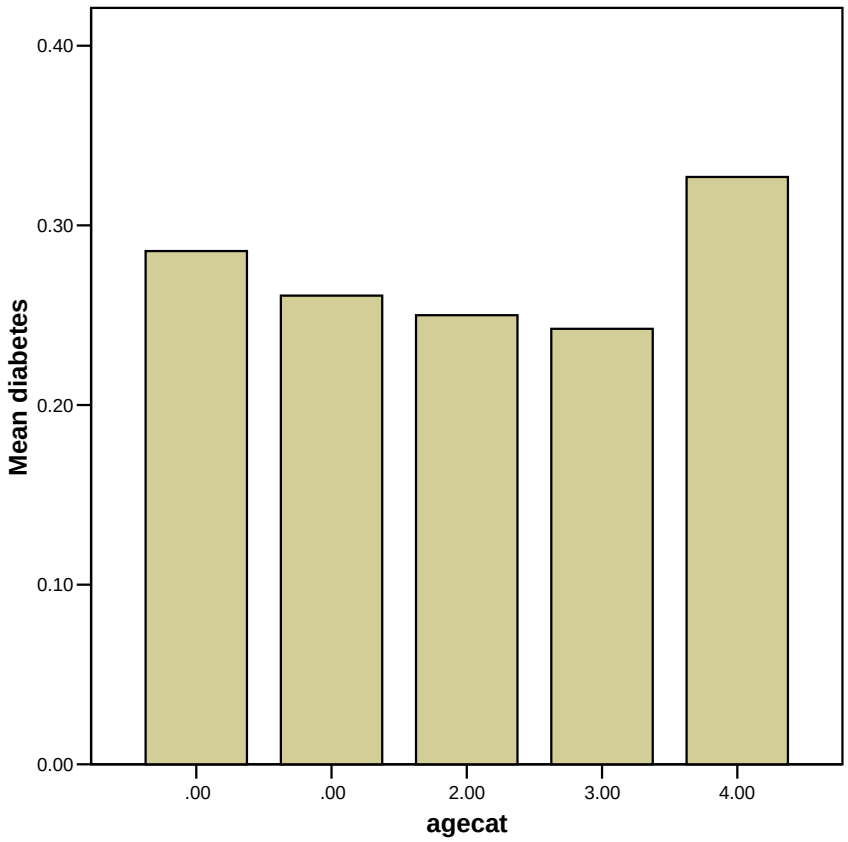
Notice that we can use the following SPSS syntax to create the new variable agecat.



We can choose **Run** > **All** to run the entire block of code or just select the block of code we would like to run and then use **Run** > **Selection**.



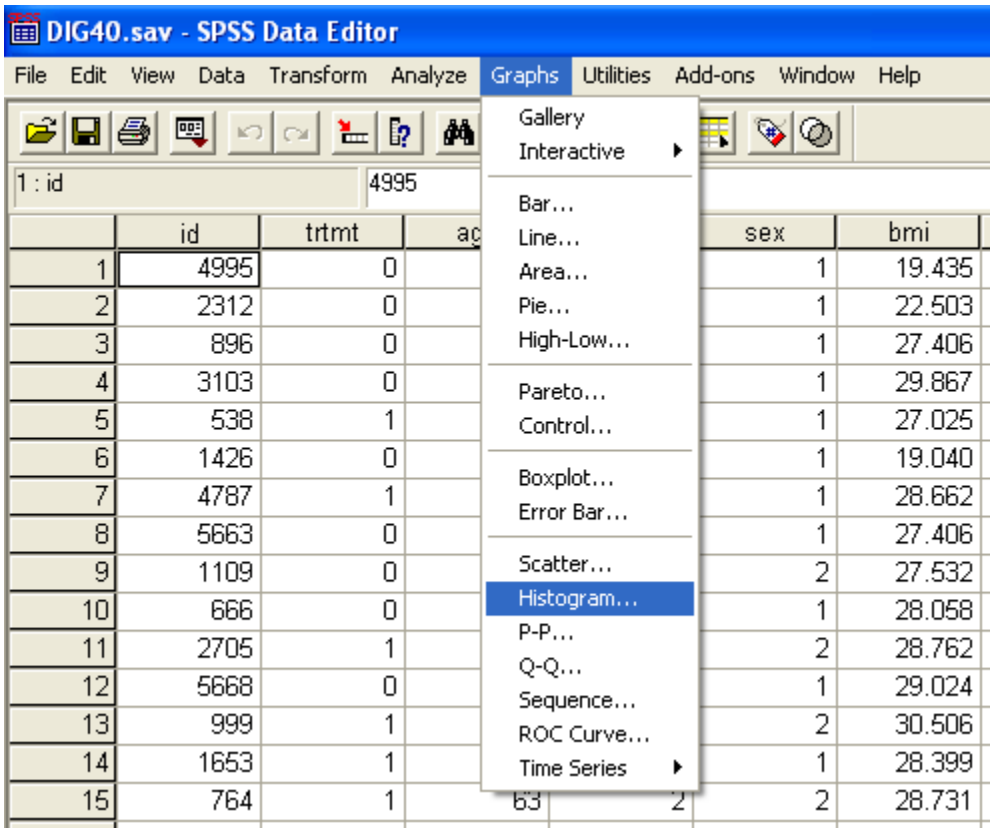
This produces the bar chart below.

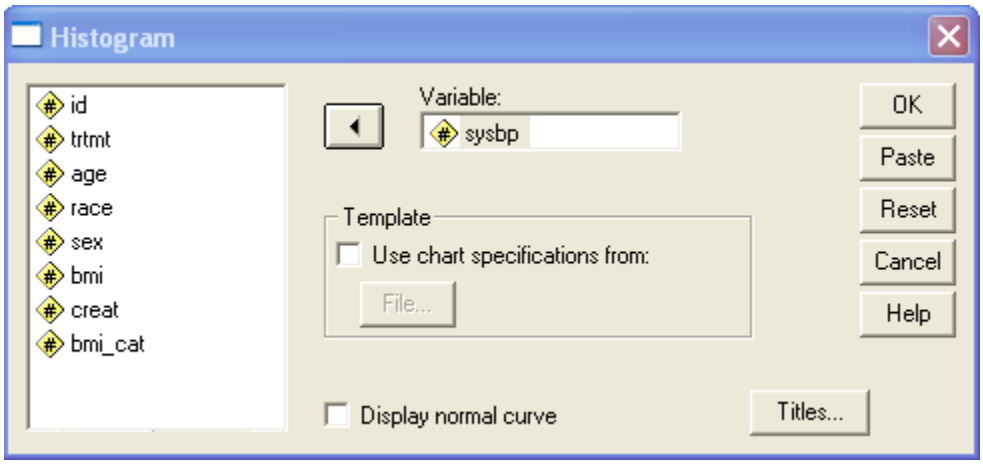


At this point, we strongly suggest using the SPSS syntax editor. In many of the examples that follow, we present the SPSS syntax along with window displays.

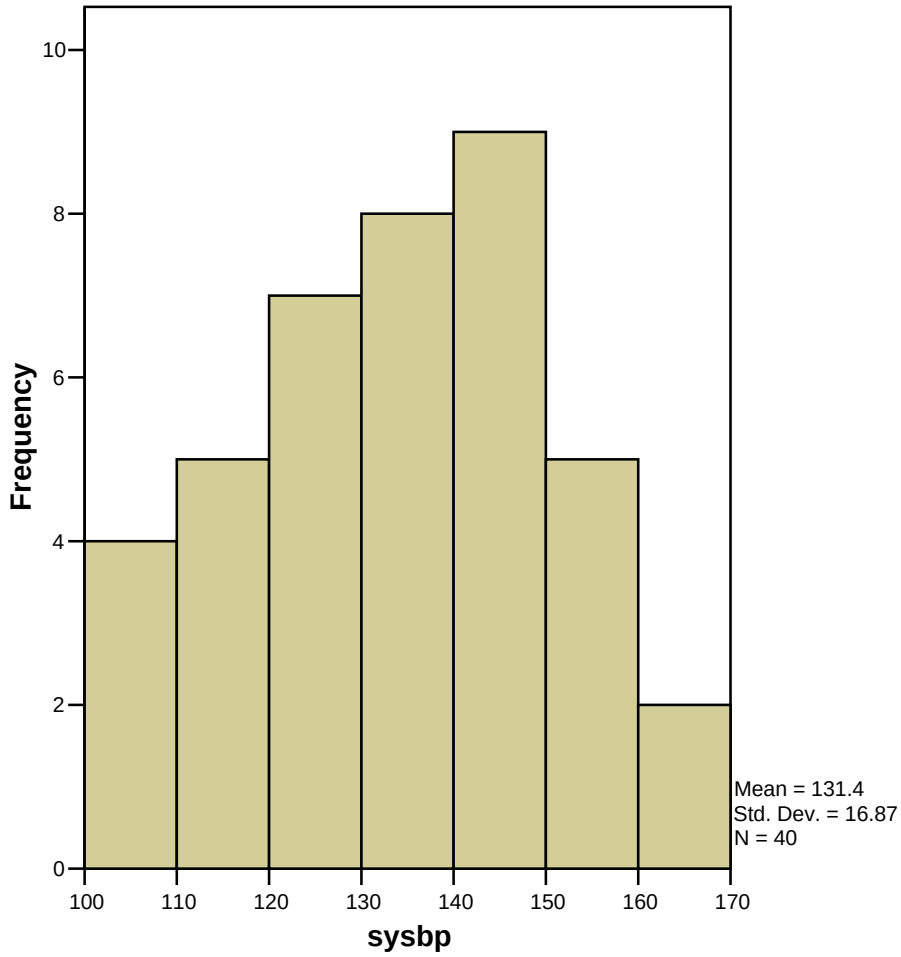
Program Note 3.3 – Creating histograms

As an example, we display a histogram of the systolic blood pressure readings of participants in the DIG40 data set.

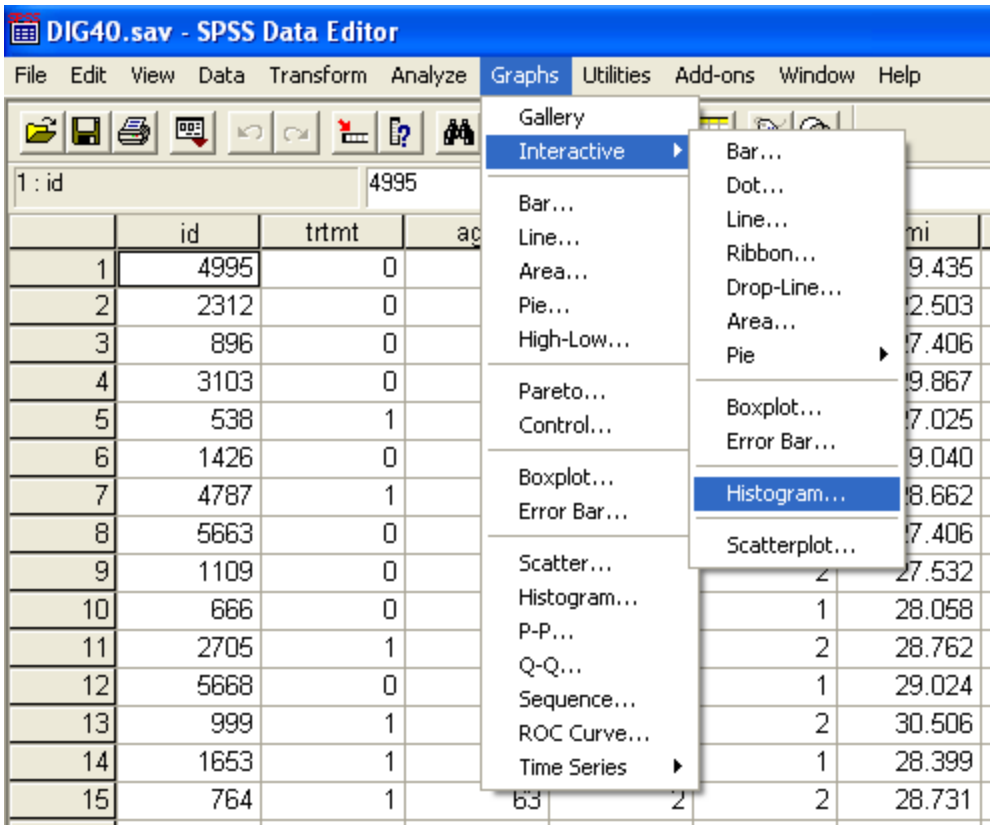




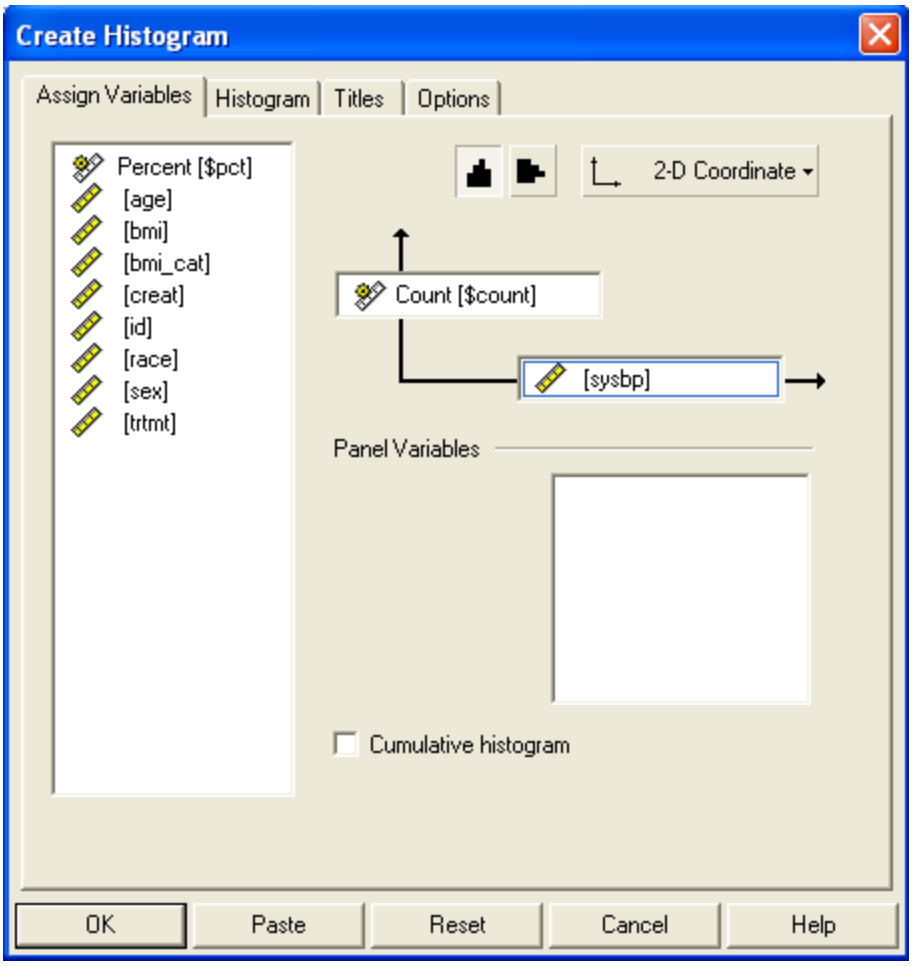
The SPSS output is shown below; however, it does not resemble the graph we produced. Below we will present the SPSS syntax that allows the user to specify the number of bins desired.



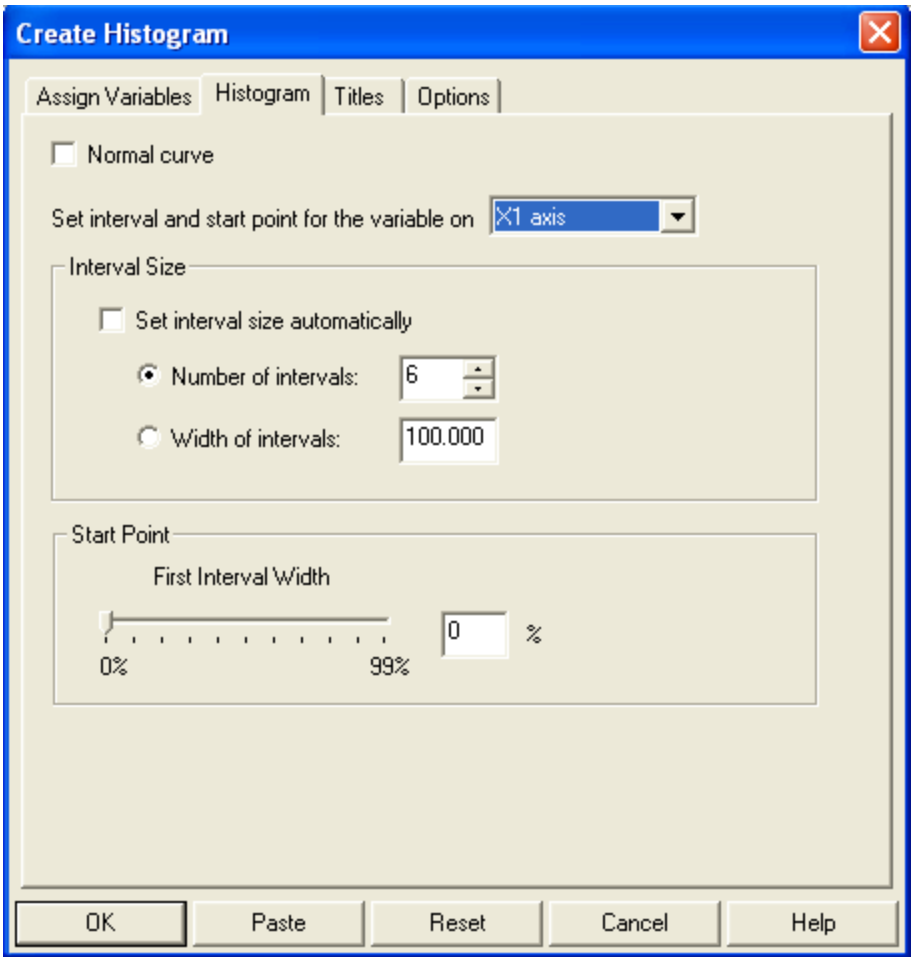
Another option is to use **Graphs** > **Interactive** > **Histogram…** which gives options to modify the histogram.



In the **Create Histogram** window, click and drag the variable of interest, sysbp, into the box overlapping the horizontal line.



Select the **Histogram** tab and under **Interval Size** bullet **Number of intervals:** then force the number of intervals to be 6.

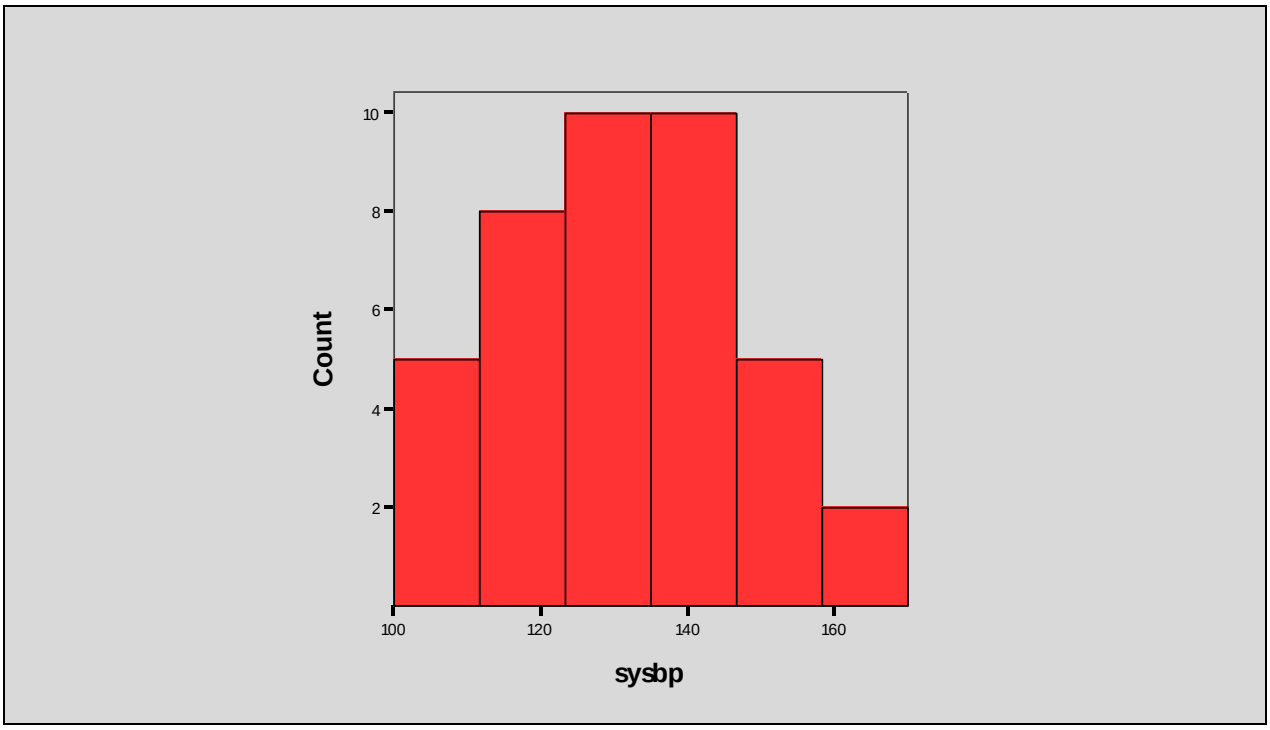


Before clicking **OK**, click on **Paste**. The **Paste** option allows you to obtain the SPSS syntax used to create the histogram with 6 intervals that starts at 100.

SPSS Syntax:

```
IGRAPH /VIEWNAME='Histogram' /X1 = VAR(sysbp) TYPE = SCALE /Y = $count /COORDINATE =
VERTICAL  /X1LENGTH=3.0 /YLENGTH=3.0
 /X2LENGTH=3.0 /CHARTLOOK='NONE' /Histogram  SHAPE = HISTOGRAM CURVE = OFF
X1INTERVAL NUM = 6 X1START = 0.
EXE.
```

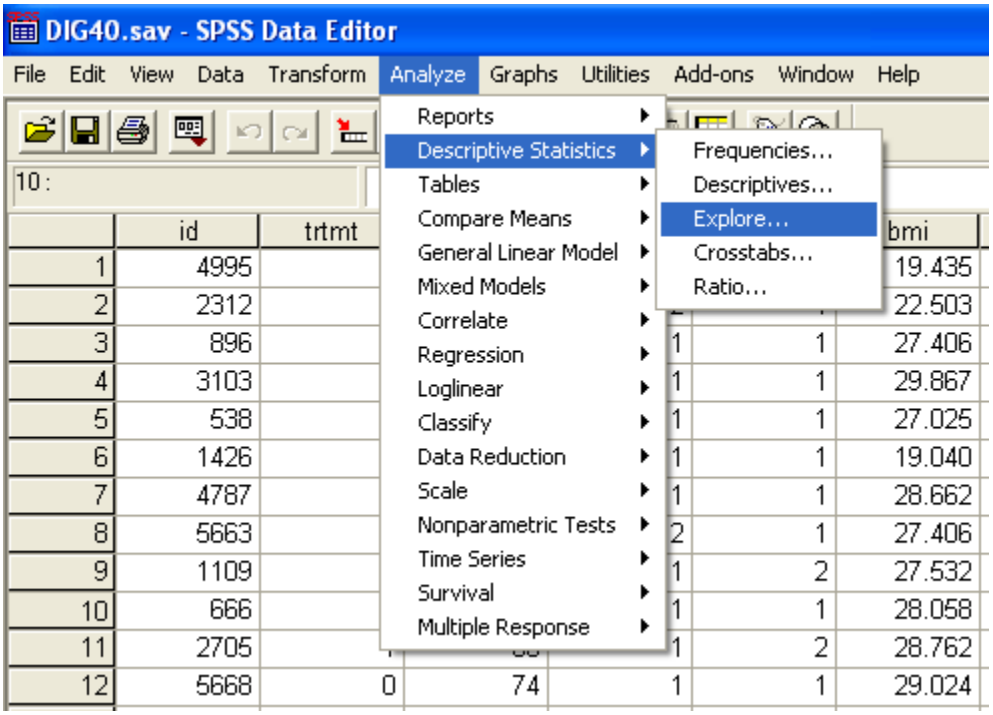
SPSS output:


Count
10
8
6
4
2
100
120
140
160
sysbp

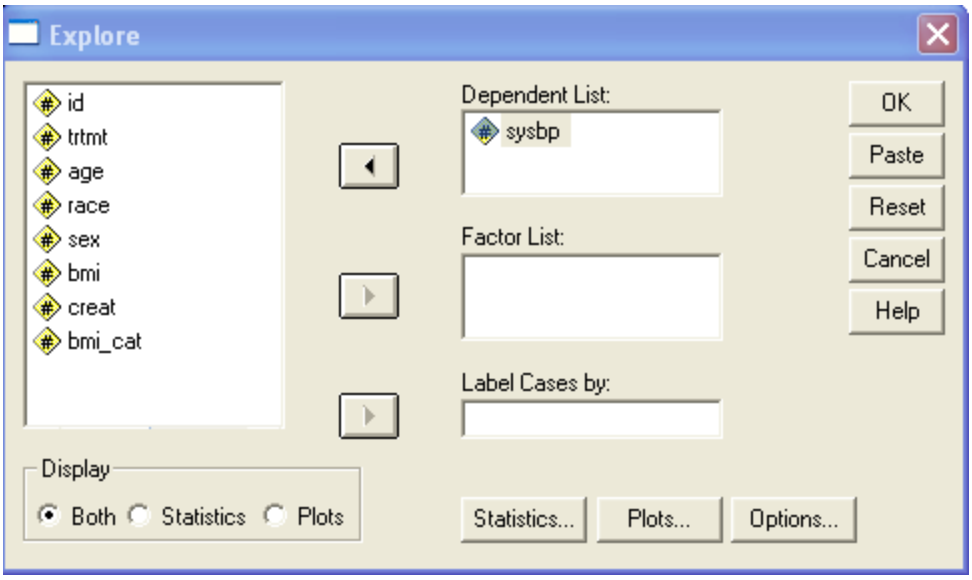
Program Note 3.4 – Creating stem and leaf plots and scatter plots

1. Stem and Leaf plots

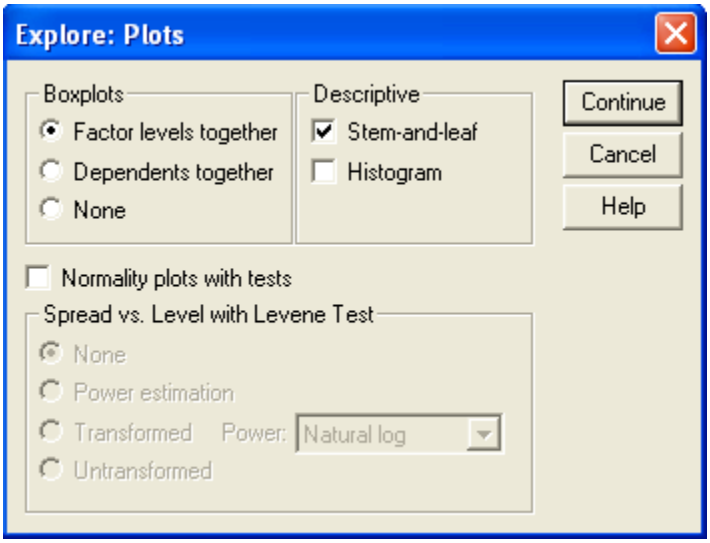
Using the DIG40 data set, a stem and leaf plot for systolic blood pressure readings can be created with the commands below:



Select the variable sysbp into the **Dependent List:** then click on **Plots…** which gives an option of available plots used by **Explore**.



In the **Explore: Plots** window under **Descriptive**, select **Stem-and-leaf**.



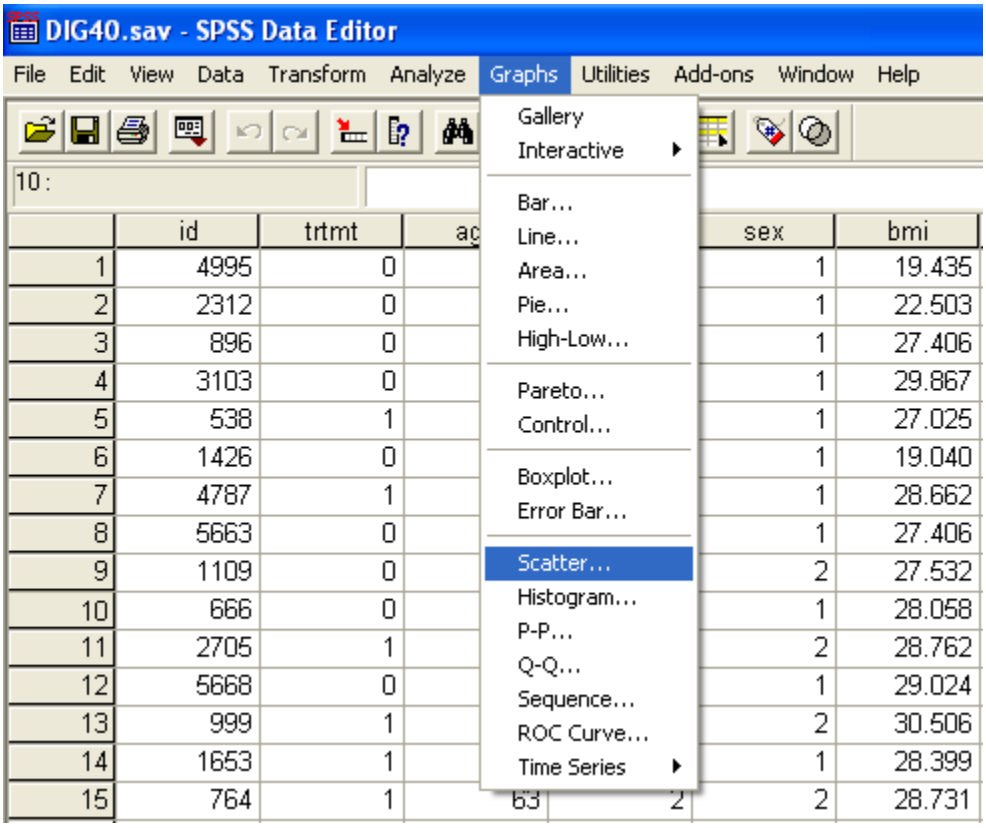
SPSS output is provided below.

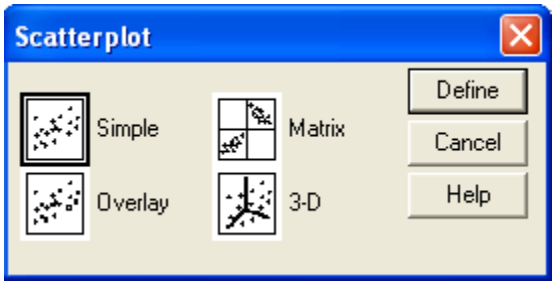
```
sysbp Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     3.00       10 .  004
     1.00       10 .  5
     1.00       11 .  0
     4.00       11 .  5666
     5.00       12 .  00024
     2.00       12 .  88
     7.00       13 .  0000004
     1.00       13 .  8
```

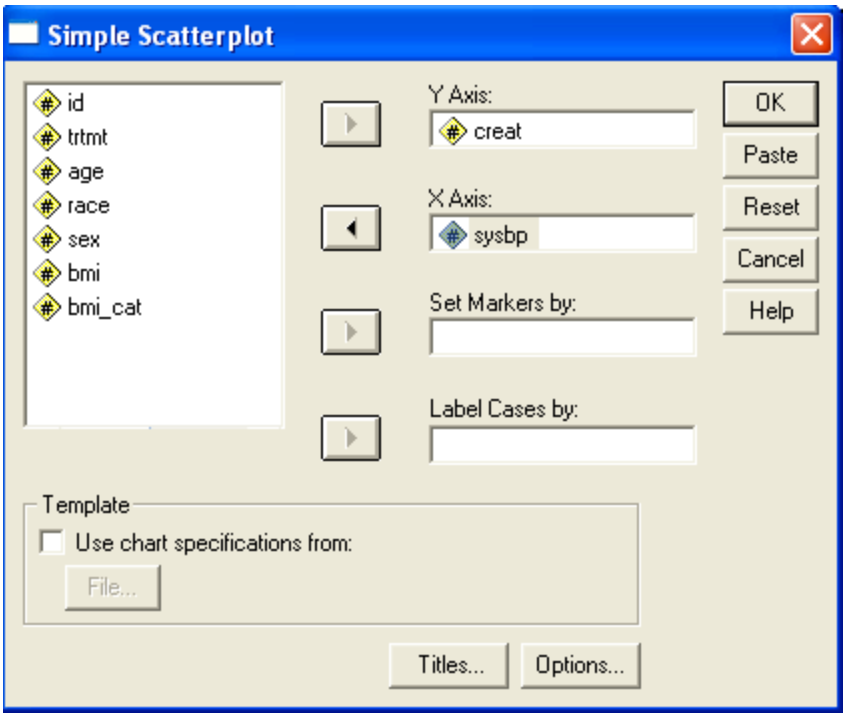
```
     9.00       14 .  000000044
      .00       14 .
     5.00       15 .  00002
     2.00 Extremes    (>=170)

 Stem width:        10
 Each leaf:       1 case(s)
```

2. Scatter plots

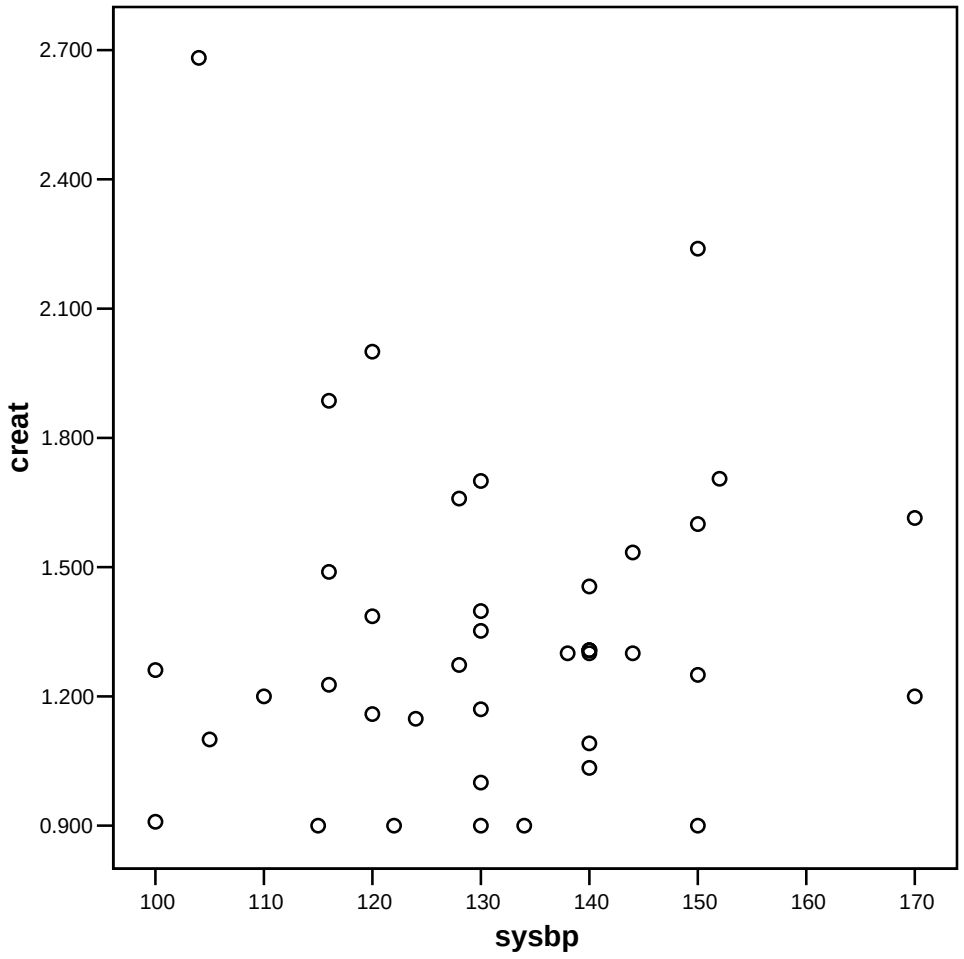
In Figure 3.12, we use a scatter plot to explore the relationship between serum creatinine (creat) and systolic blood pressure (sysbp) using the DIG40 data set.






Simple Scatterplot
id
trtmt
age
race
sex
bmi
bmi_cat
Y Axis:
creat
X Axis:
sysbp
Set Markers by:
Label Cases by:
OK
Paste
Reset
Cancel
Help
Template
Use chart specifications from:
File...
Titles...
Options...

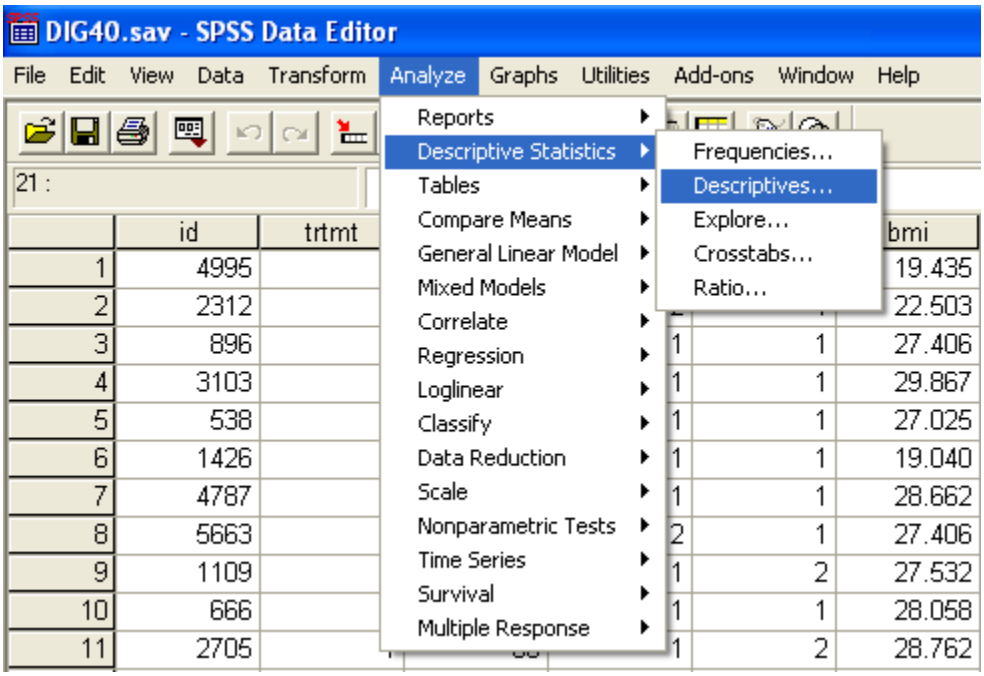
The SPSS output is provided below:

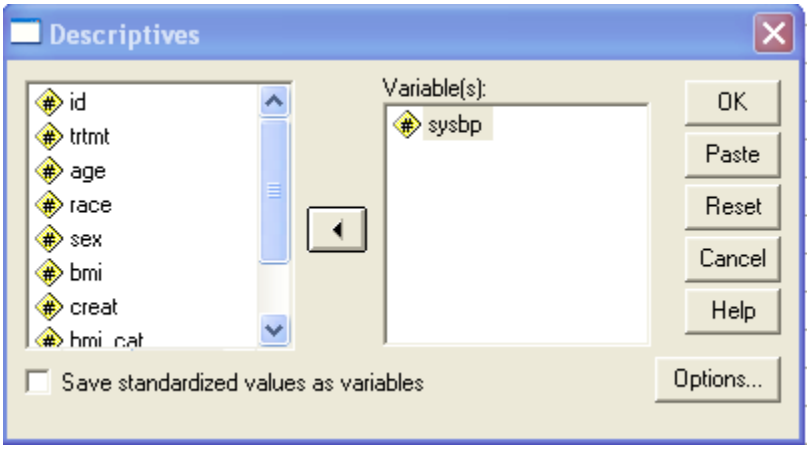


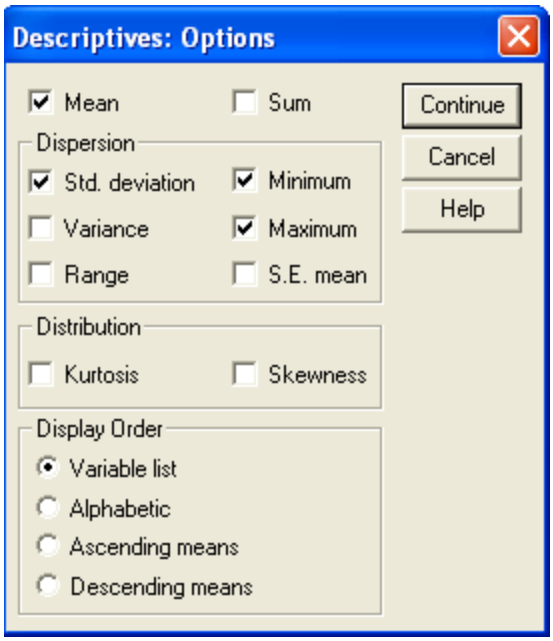
Program Note 3.5 – Descriptive statistics and creating box plots

1. Descriptive Statistics

SPSS can be used to get the mean, standard deviation, and range for systolic blood pressure for patients from the DIG40 data set.







SPSS Syntax:

```
DESCRIPTIVES
  VARIABLES=sysbp /STATISTICS=MEAN STDDEV MIN MAX .
```

SPSS output:

Descriptive Statistics

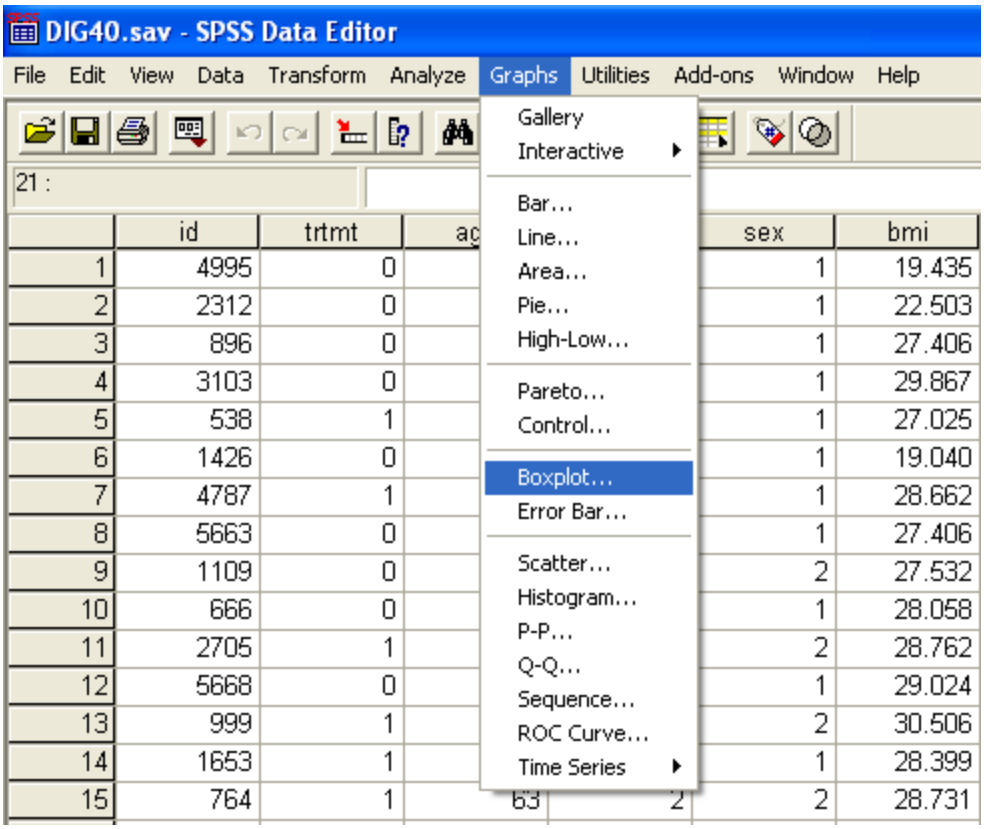
	N	Minimum	Maximum	Mean	Std. Deviation
sysbp	40	100	170	131.40	16.870
Valid N (listwise)	40				

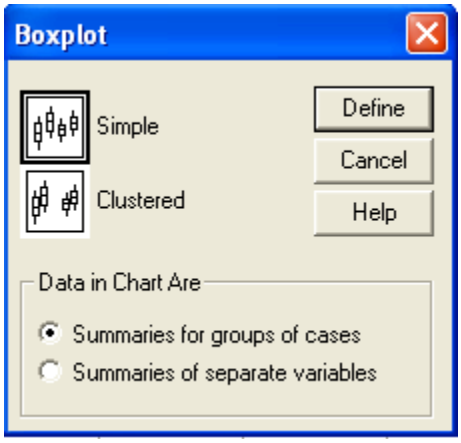
2. Box plots

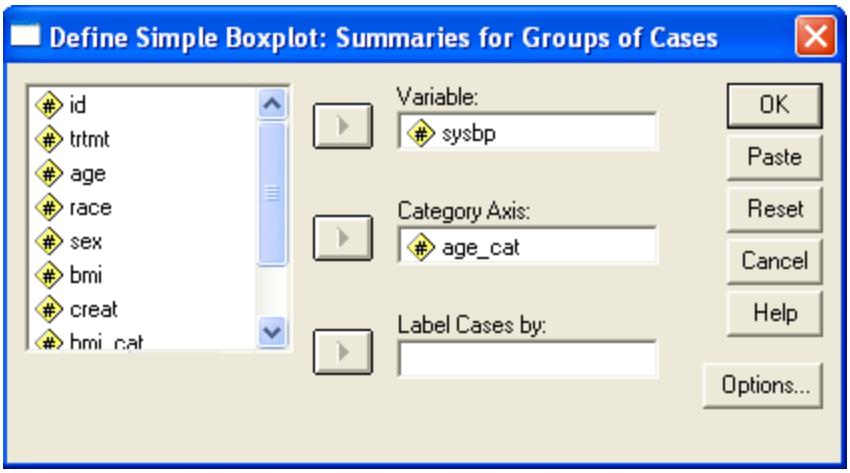
As an example, we show the commands that are used to create Figure 3.15 using the DIG40 data set.

SPSS Syntax:

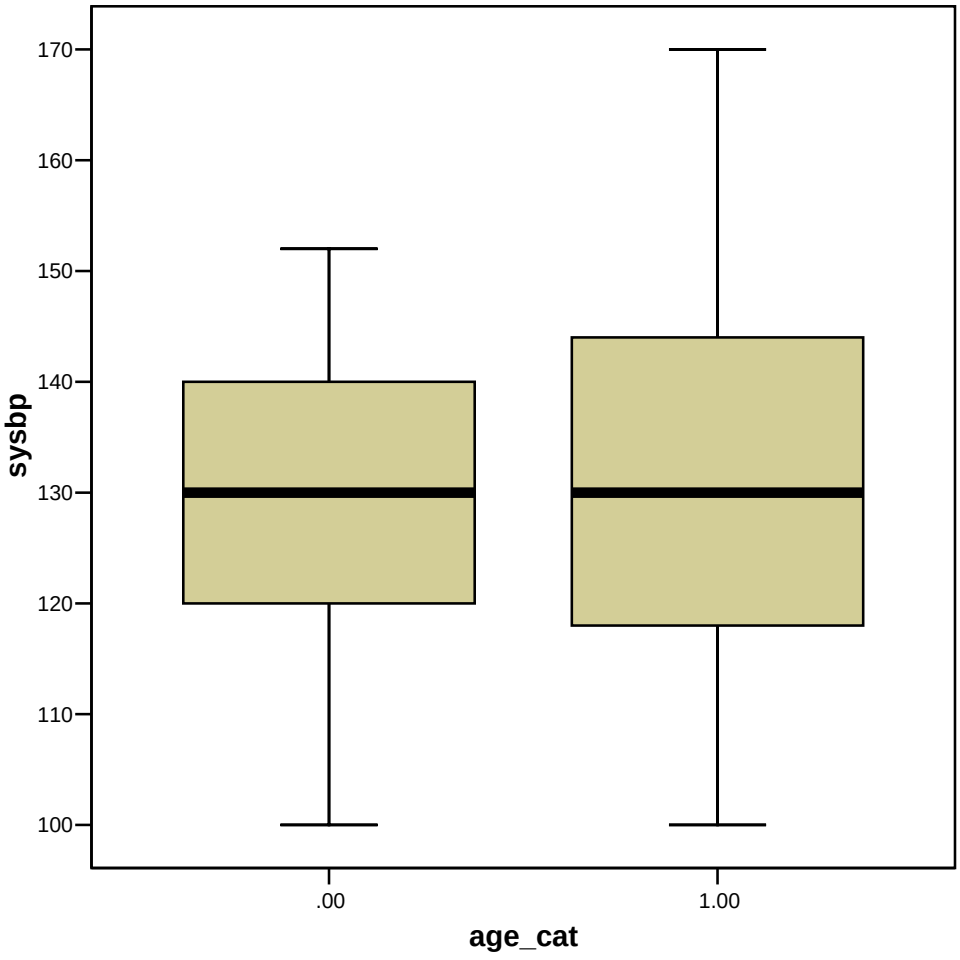
```
if (age<60) age_cat = 0.
if (age>=60) age_cat = 1.
execute.
```


DIG40.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Gallery
Interactive
Bar...
Line...
Area...
Pie...
High-Low...
Pareto...
Control...
Boxplot...
Error Bar...
Scatter...
Histogram...
P-P...
Q-Q...
Sequence...
ROC Curve...
Time Series
21 :
id trtmt sex bmi
1 4995 0 1 19.435
2 2312 0 1 22.503
3 896 0 1 27.406
4 3103 0 1 29.867
5 538 1 1 27.025
6 1426 0 1 19.040
7 4787 1 1 28.662
8 5663 0 1 27.406
9 1109 0 2 27.532
10 666 0 1 28.058
11 2705 1 2 28.762
12 5668 0 1 29.024
13 999 1 2 30.506
14 1653 1 1 28.399
15 764 1 2 28.731


Boxplot
Simple
Clustered
Define
Cancel
Help
Data in Chart Are
Summaries for groups of cases
Summaries of separate variables

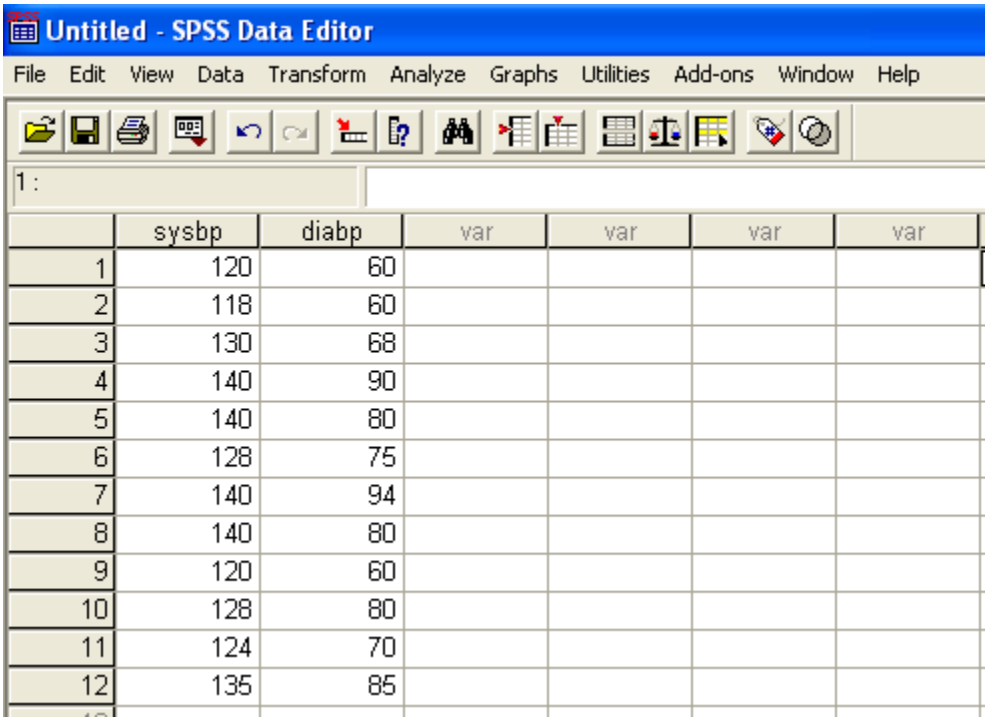

Define Simple Boxplot: Summaries for Groups of Cases
id
trtmt
age
race
sex
bmi
creat
bmi_cat
Variable:
sysbp
Category Axis:
age_cat
Label Cases by:
OK
Paste
Reset
Cancel
Help
Options...

SPSS output:

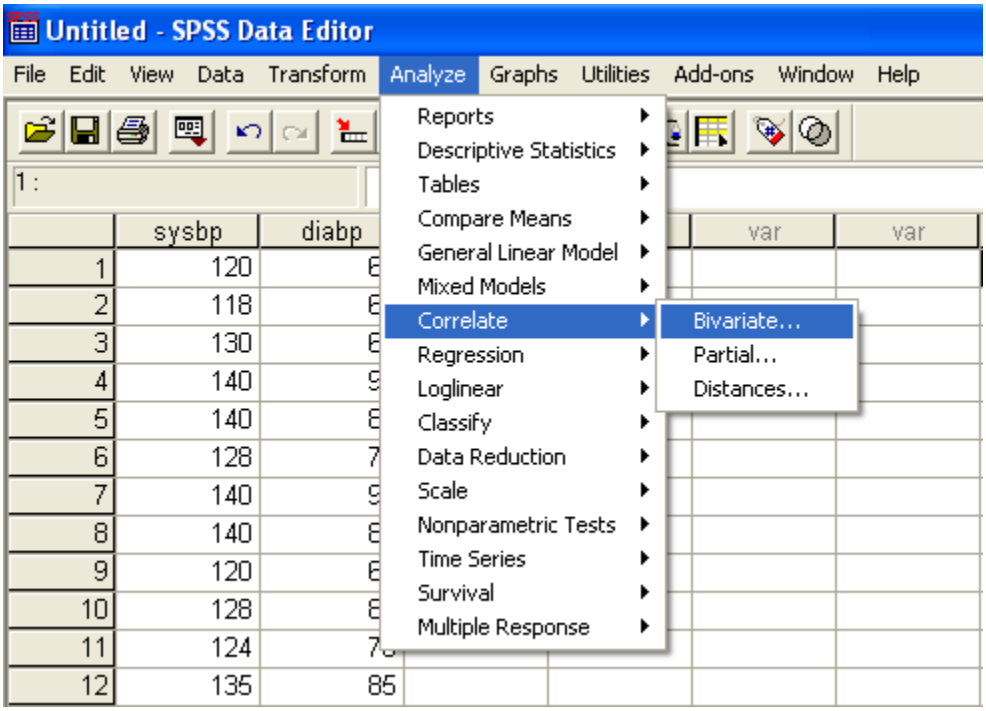

170
160
150
140
130
120
110
100
sysbp
.00
1.00
age_cat

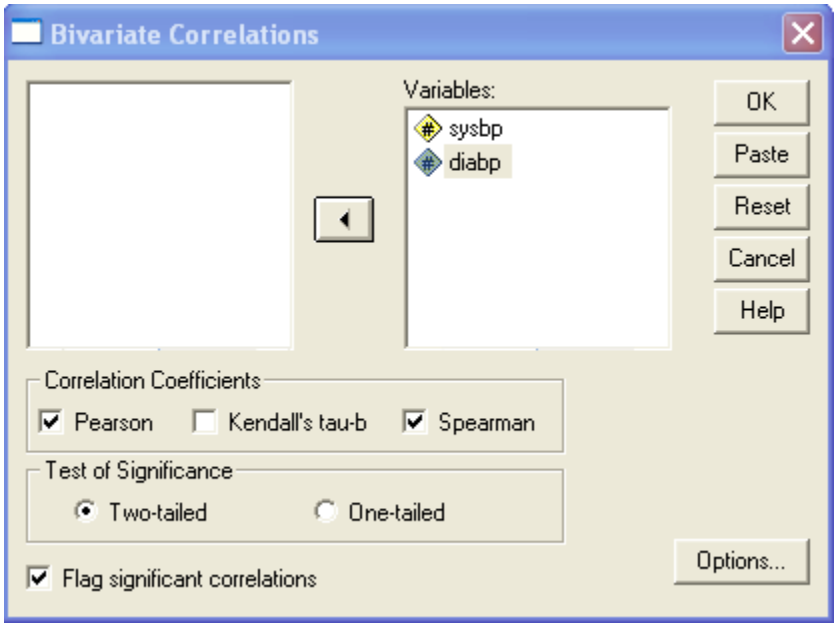
Program Note 3.6 – Calculating Pearson and Spearman correlation coefficients

Below is the data shown in Example 3.18.



	sysbp	diabp	var	var	var	var
1	120	60				
2	118	60				
3	130	68				
4	140	90				
5	140	80				
6	128	75				
7	140	94				
8	140	80				
9	120	60				
10	128	80				
11	124	70				
12	135	85				





In the SPSS output below, both the Peason and Spearman correlation coefficients are provided.

Correlations

		sysbp	diabp
sysbp	Pearson Correlation	1	.894(**)
	Sig. (2-tailed)	.	.000
	N	12	12
diabp	Pearson Correlation	.894(**)	1
	Sig. (2-tailed)	.000	.
	N	12	12

** Correlation is significant at the 0.01 level (2-tailed).

Correlations

			sysbp	diabp
Spearman's rho	sysbp	Correlation Coefficient	1.000	.866(**)
		Sig. (2-tailed)	.	.000
		N	12	12

	diabp	Correlation Coefficient	.866(**)	1.000
		Sig. (2-tailed)	.000	.
		N	12	12

** Correlation is significant at the 0.01 level (2-tailed).