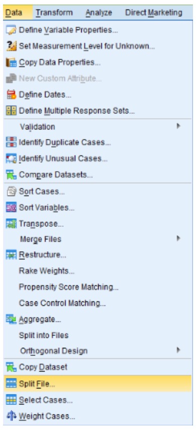
Summarizing Data: Relationships Between Variables in SPSS

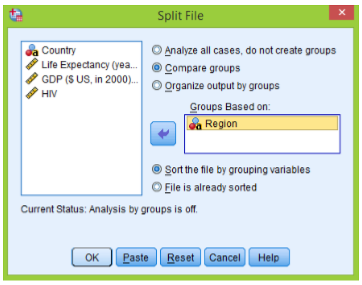
In this document we show how to find relationship between two variables. Since quantitative and categorical variables must be treated differently we show which plots, tables and statistics are appropriate in different cases.
For this document we will use ‘Skeleton’ and ‘Life Expectancy’ data sets. It is assumed that you have managed to upload all these data into SPSS (please refer to the ‘Data sets import in SPSS’ document for detailed explanation).

Relationship between quantitative and categorical variables

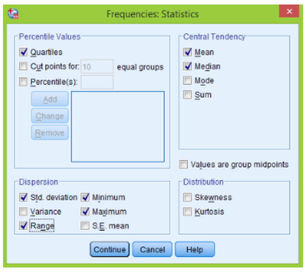
We start this section with the ‘Life Expectancy’ data set. We know how to find descriptive statistics and get a boxplot of one variable, for example ‘LifeExp’. Now we want to understand relationship between ‘LifeExp’ and ‘Region’. Hence we need ‘LifeExp’ statistics for every region. To do that we first need to split the data, go to **Data** > **Split File**



Here select ‘Compare groups’ then move ‘Region’ variable to the right window using arrow and also select ‘Sort the file by grouping variables’



Click **OK** and you will see that data set is sorted by ‘Region’. Next to get descriptive statistics (mean, median, quartiles etc.) by ‘Region’ just go to **Analyze** > **Descriptive Statistics** > **Frequencies**, move ‘LifeExp’ variable to the right, click on ‘Statistics’ button to select different statistics:



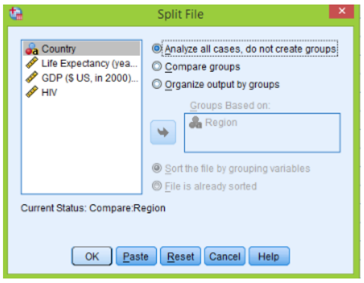
Click **Continue** > **OK** to produce the following tables:

Statistics

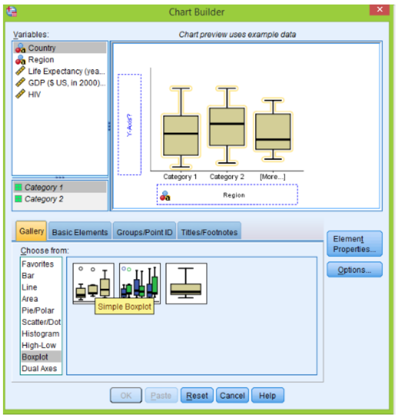
Life Expectancy (years)

Amer	N	Valid	39
		Missing	0
	Mean		74.91921
	Median		75.62000
	Std. Deviation		3.875471
	Range		18.917
	Minimum		62.095
	Maximum		81.012
	Percentiles	25	73.12600
		50	75.62000
		75	77.00500
EAP	N	Valid	30
		Missing	0
	Mean		73.08603
	Median		72.95000
	Std. Deviation		6.220434
	Range		20.919
	Minimum		62.475
	Maximum		83.394
	Percentiles	25	68.68625
		50	72.95000
		75	78.66425

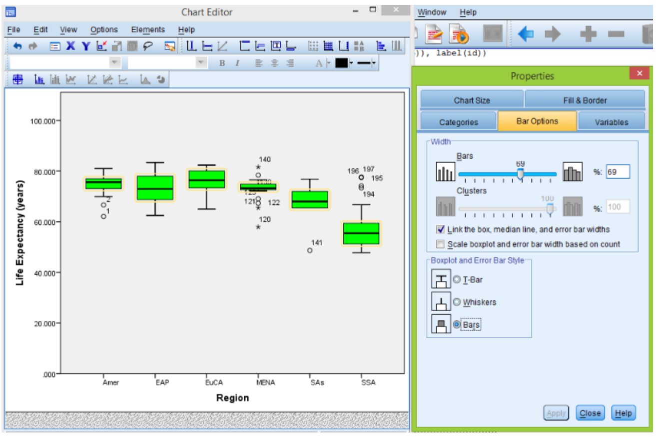
See that statistics are calculated for each region. To get boxplots for each region, just go to **Chart Builder** > **Boxplot** > **double click on 1-D Boxplot** then drag 'LifeExp' to the 'X-axis'. Six separate boxplots are printed in the output window corresponding to each region. If we want to see all the boxplots in one plot, we have to return to the un-split data file. Hence go to **Data** > **Split File** and select 'Analyze all cases, do not create groups':



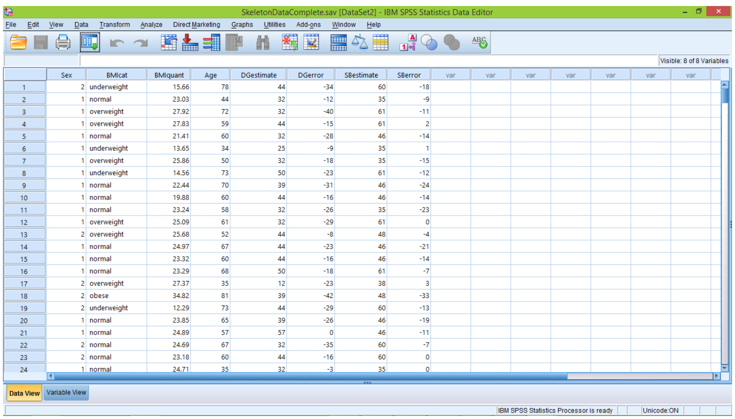
Now when the file is not split go to **Chart Builder** > **Boxplot** > **double click on Simple Boxplot**



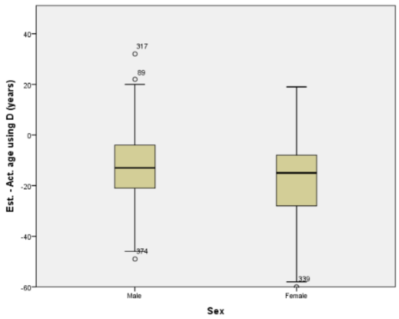
Drag 'Region' to the X-axis and 'LifeExp' to the Y-axis click **OK** and the plot is printed. Double click on the plot to open 'Chart Editor' to make some changes. Double click on any of the boxplots and 'Properties' window appears. To change color of the boxes click on 'Fill & Border', to change the width of the boxes click on 'Bar Options'



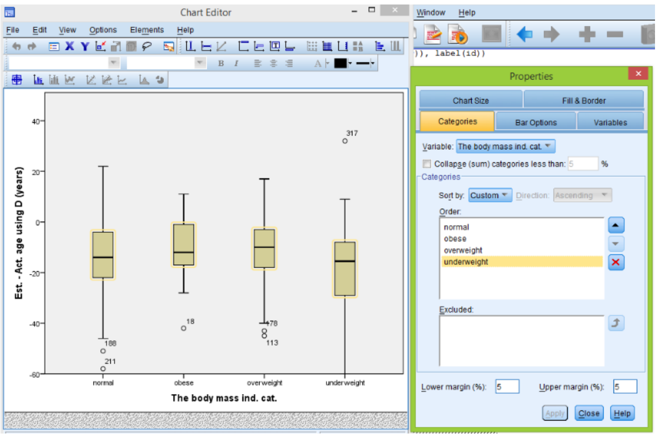
Next lets look at 'Skeleton' data set:



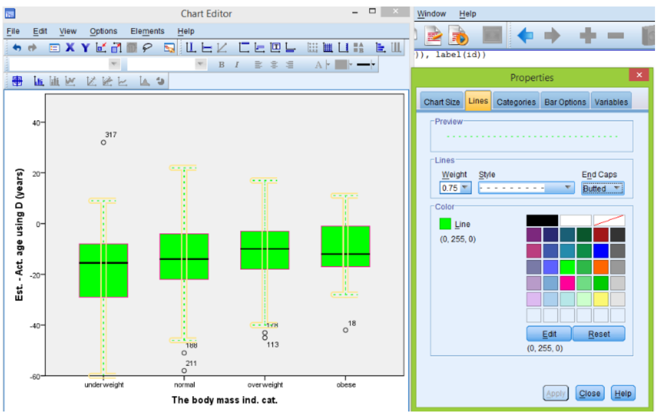
To get the relationship between 'DGerror' variable and 'Sex', we make the boxplots as described above: **Chart Builder > Boxplot > double click on Simple Boxplot**, drag 'Sex' variable to the X-axis and 'DGerror' to Y-axis, click **OK** and get:



Next we are interested in how ‘DGerror’ varies with ‘BMIcat’. We get the boxplots, but observe that the order of ‘BMIcat’ is not logical, we want it to be ‘underweight’, ‘normal’, ‘overweight’ and ‘obese’. It is very simple to change in SPSS. Double click on the plot then double click on any boxplot to open ‘Properties’ window and click on ‘Categories’ button

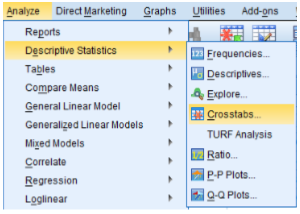


Now we put categories in the right order using arrows, to finish click **Apply** button. Also by double clicking on the ‘fence’ of any boxplot we can change its stile and color

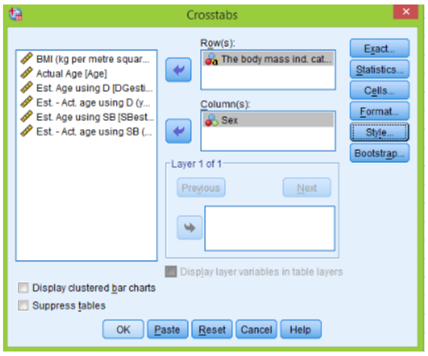


Relationship between two categorical variables

In this section we show how to work with two categorical variables. We start with the ‘Skeleton’ data. The goal is to investigate the relationship between ‘BMIcat’ and ‘Sex’ variables (both of them are categorical). First we create a table of counts, go to **Analyze** > **Descriptive Statistics** > **Crosstabs**



Then move 'BMIcat' variable to the 'Row(s)' window and 'Sex' to the 'Column(s)' section



Click **OK** and a simple table of counts is produced:

Crosstabs

Case Processing Summary

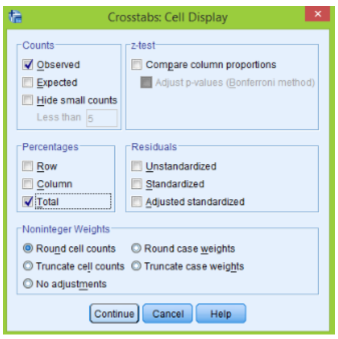
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
The body mass ind. cat. * Sex	400	100.0%	0	0.0%	400	100.0%

The body mass ind. cat. * Sex Crosstabulation

Count

		Sex		
		Male	Female	Total
The body mass ind. cat.	normal	166	59	225
	obese	10	10	20
	overweight	59	22	81
	underweight	46	28	74
Total		281	119	400

This table just shows how many observations are in each category. If we need joint or conditional distribution then as before **Analyze** > **Descriptive Statistics**> **Crosstabs** then click on 'Cells' button. If we need joint distribution then select 'Total' under 'Percentages' if we need for example conditional distribution given 'Sex' then select 'Column' option, here we need joint distribution:



Click **Continue** > **OK** and a new table is printed:

Crosstabs

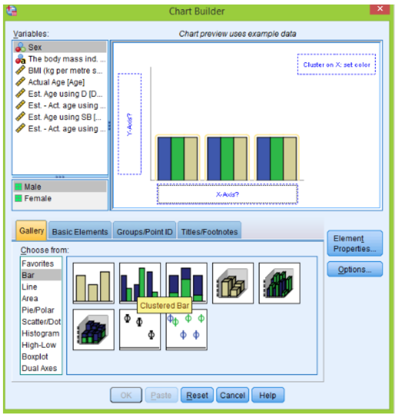
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
The body mass ind. cat. * Sex	400	100.0%	0	0.0%	400	100.0%

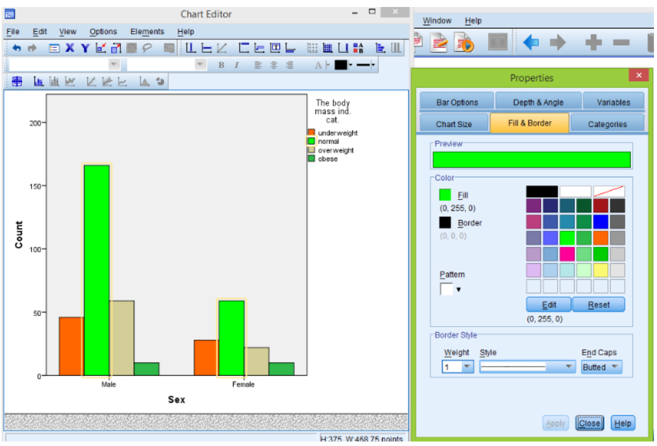
The body mass ind. cat. * Sex Crosstabulation

			Sex		Total
			Male	Female	
The body mass ind. cat.	normal	Count	166	59	225
		% of Total	41.5%	14.8%	56.3%
	obese	Count	10	10	20
		% of Total	2.5%	2.5%	5.0%
	overweight	Count	59	22	81
		% of Total	14.8%	5.5%	20.3%
	underweight	Count	46	28	74
		% of Total	11.5%	7.0%	18.5%
Total		Count	281	119	400
		% of Total	70.3%	29.8%	100.0%

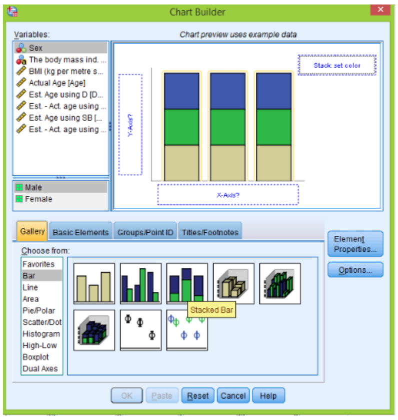
These percentages represent the joint distribution. To make a visual representation of the relationship we can use a bar-plot of counts for each 'BMIcat' and 'Sex' categories, go to **Chart Builder > Bar > double click on Clustered Bar**



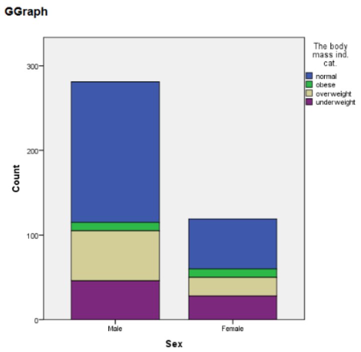
Drag 'Sex' variable to the X-axis and 'BMIcat' to the 'Cluster on X: set color', click **OK** and the bar-plot is produced. Once again the order of categories of the 'BMIcat' is not good, hence double click on the plot, double click on any bar and then as usual click on 'Categories' button. Also to change colors of the bars, double click on the colored squares in the legend



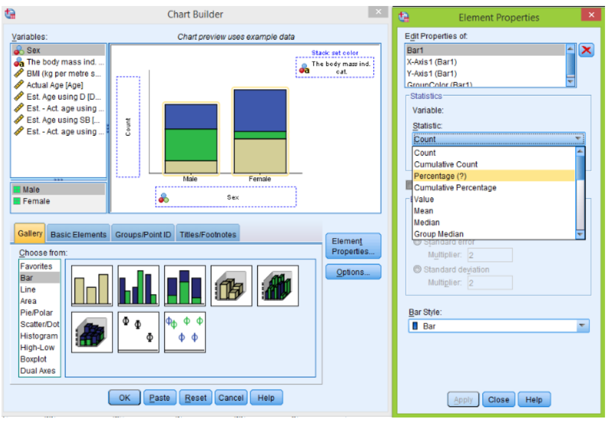
If we need stacked bars then **Chart Builder > Bar > double click on Stacked Bar**



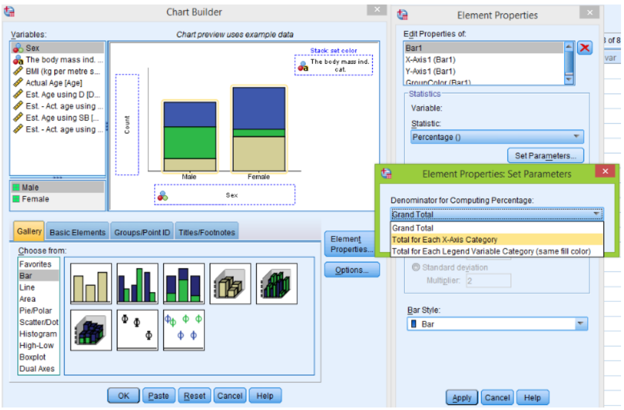
As before drag 'Sex' variable to X-axis and 'BMIcat' to the 'Stack: set color' and click **OK**



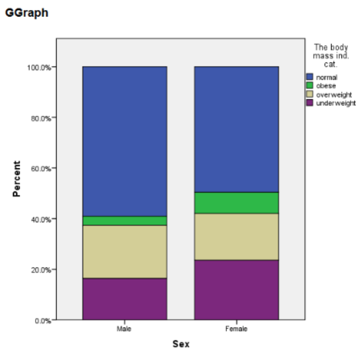
These were plots of counts, but next we want to plot conditional distributions of 'BMIcat' given 'Sex' factors. Lets make a stacked bar plot. In the 'Chart Builder' section there is an 'Element Properties' window and under 'Statistic' select 'Percentage':



Then we click on the 'Set Parameters' button and select 'Total for each X-Axis Category'



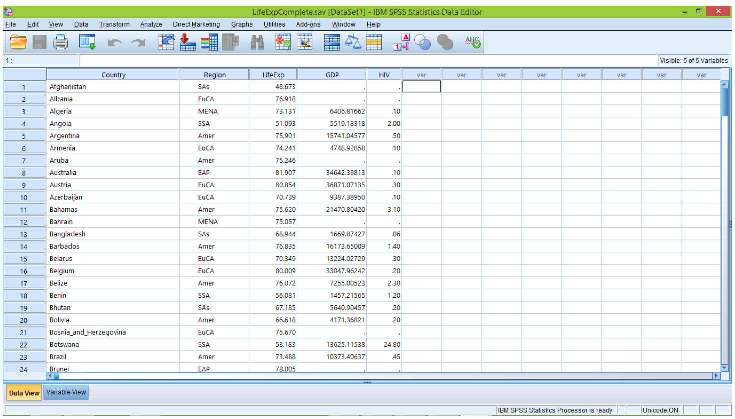
Click **Apply** > **OK** to finish, the following plot is produced:



See that now the heights of the stacked bars are the same and equal to 100%.

Relationship between two quantitative variables

Let's start with 'Life Expectancy' data set.



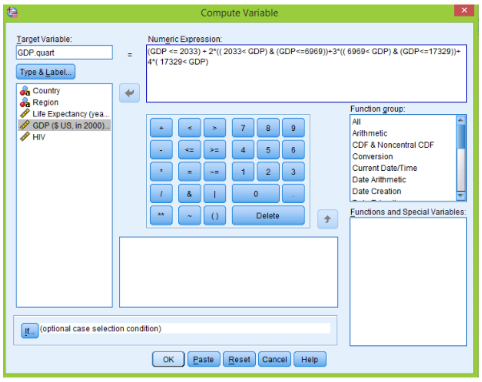
The goal is to investigate the relationship between 'GDP' and 'LifeExp' (both of them are quantitative). First we get basic statistics of 'GDP' variable in a usual way (**Analyze** > **Descriptive Statistics** > **Frequencies**):

Statistics

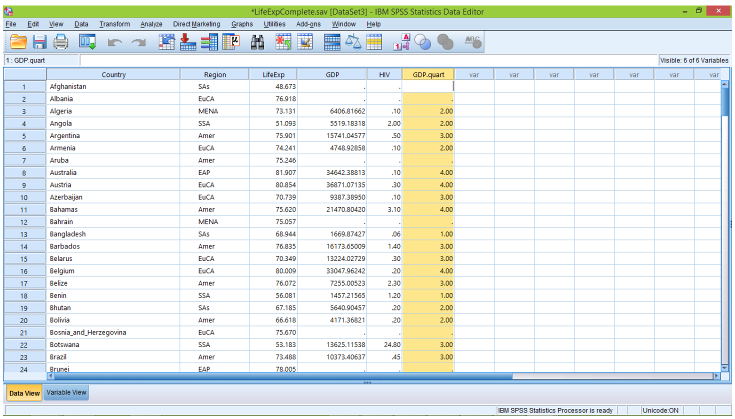
GDP ($ US, in 2000)

N	Valid	147
	Missing	50
Mean		12325.31865
Percentiles	25	2033.232880
	50	6969.563610
	75	17329.59545

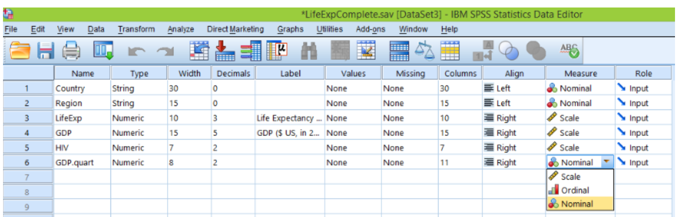
We can get boxplots of 'LifeExp' for four quarters of 'GDP' as we did in the first section of this document. Hence we need to construct a new categorical variable taking values 1,2,3,4 corresponding to which quarter, GDP observation belongs to. We will use the quantiles from the above table. Go to **Transform > Compute variable**, we call the target variable 'GDP.quart' and enter the following expression:



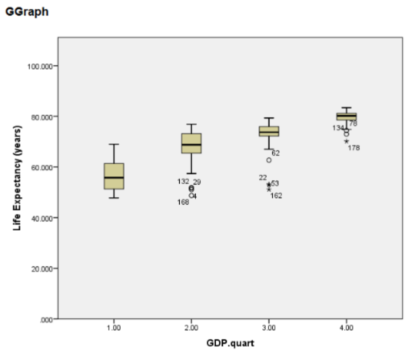
Note that & means 'and'. Therefore if GDP is less than the first quartile we get 1 if it is bigger than the first quartile but less than the median we get 2 and so on. Click **OK**, and a new column appears:



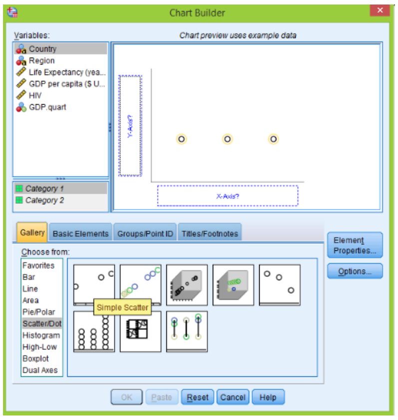
Click on the 'Variable View' button and make sure that the 'Measure' for this new variable is 'Nominal' since we want SPSS to treat this variables as categorical.



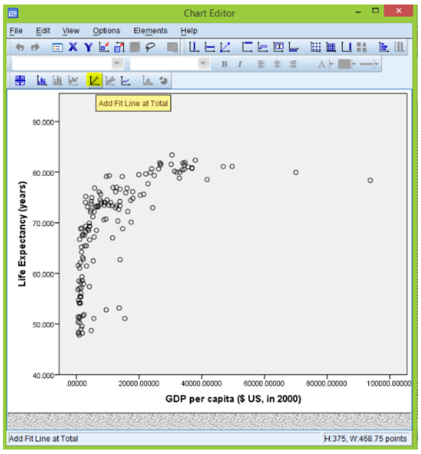
Now we can easily make a joint chart of boxplots for each quarter as explained in the first section.



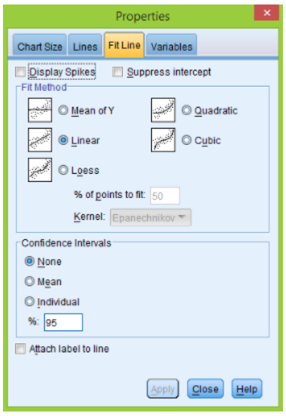
However we can also make a simple scatterplot of two quantitative variables, go to **Chart Builder > Scatter/Dot > double click on Simple Scatter**



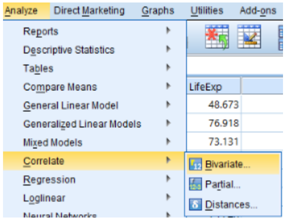
Afterwards drag 'GDP' to the X-axis and 'LifeExp' to Y-axis, click **OK** to produce the plot. If we want to add a line of best fit to the plot, then double click on the plot and click on 'Add Fit Line at Total'



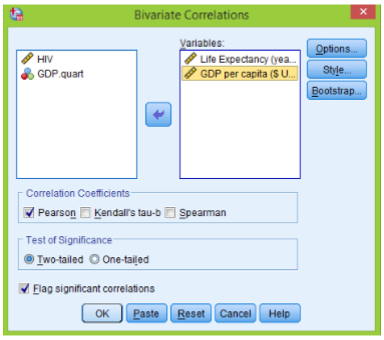
Immediately the line appears with the equation of this line, we can remove this equation if we deselect the 'Attach label to line' option



Also in the above 'Properties' window you can click on the 'Lines' button and make changes to this line (style and color). To get correlation statistic between 'GDP' and 'LifeExp' go to **Analyze** > **Correlate** > **Bivariate**



Send these two variables across; make sure 'Pearson' option is selected



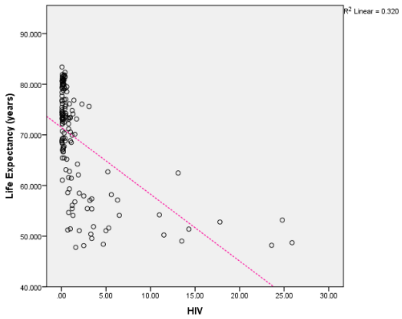
Click **OK** and the table is printed in the output window:

Correlations

		Life Expectancy (years)	GDP per capita ($ US, in 2000)
Life Expectancy (years)	Pearson Correlation	1	.635**
	Sig. (2-tailed)		.000
	N	197	147
GDP per capita ($ US, in 2000)	Pearson Correlation	.635**	1
	Sig. (2-tailed)	.000	
	N	147	147

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

We need only one number from this table, it is 0.635 which is correlation. Note: the correlation is positive and therefore the line of best fit increases. Similarly we make a scatterplot and find correlation for 'LifeExp' versus 'HIV':

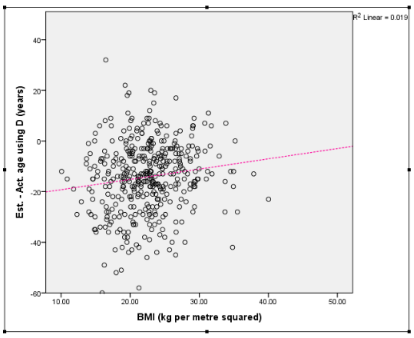


Correlations

		Life Expectancy (years)	HIV
Life Expectancy (years)	Pearson Correlation	1	-.566**
	Sig. (2-tailed)		.000
	N	197	147
HIV	Pearson Correlation	-.566**	1
	Sig. (2-tailed)	.000	
	N	147	147

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

In this case we observe negative association. Finally we move to the 'Skeleton' data set, as before we make a scatterplot (with the line of best fit) and find correlation for 'DGerror' versus 'BMIquant' variable:



Correlations

		Est. - Act. age using D (years)	BMI (kg per metre squared)
Est. - Act. age using D (years)	Pearson Correlation	1	.136**
	Sig. (2-tailed)		.006
	N	400	400
BMI (kg per metre squared)	Pearson Correlation	.136**	1
	Sig. (2-tailed)	.006	
	N	400	400

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

In this example correlation is positive and therefore there is a positive association between 'BMIquant' and 'DGerror'.