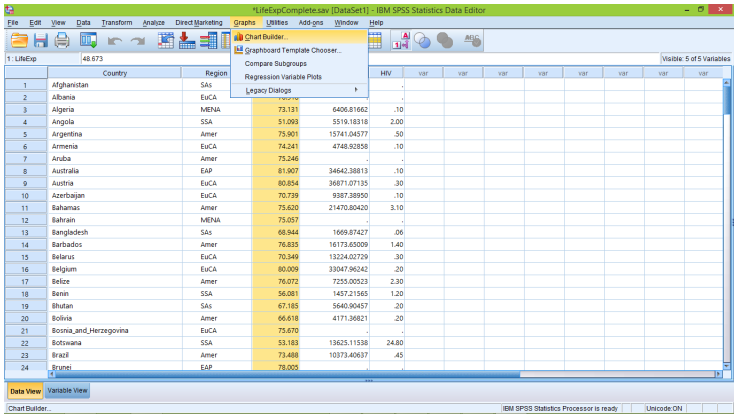
Summarizing Data in SPSS

This document describes how to find different descriptive statistics like mean, median, standard deviation and much more in SPSS. We also show how to make different kinds of plots for quantitative and categorical variables.
For this document we need 'Skeleton', 'Life Expectancy' and 'NY Red Bull Salaries' data sets. It is assumed that you have managed to upload all these data into SPSS (please refer to 'Data sets import in SPSS' document for detailed explanation).

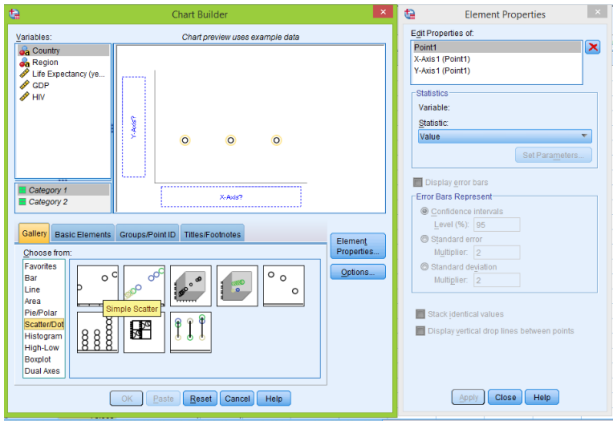
1 Five Number Summary

In this part we work with the 'Life Expectancy' data which shows for different countries the average life expectancy. These data contain five variables: Country, Region, LifeExp, GDP and HIV.

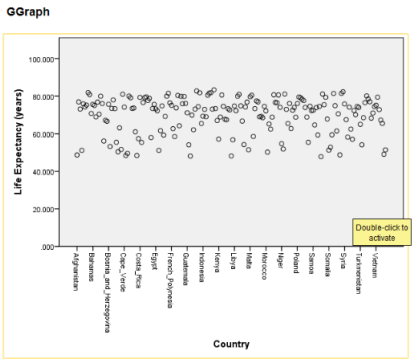
First we want to analyse 'LifeExp' variable. To make a very simple plot of this variable go to: **Graphs > Chart Builder**



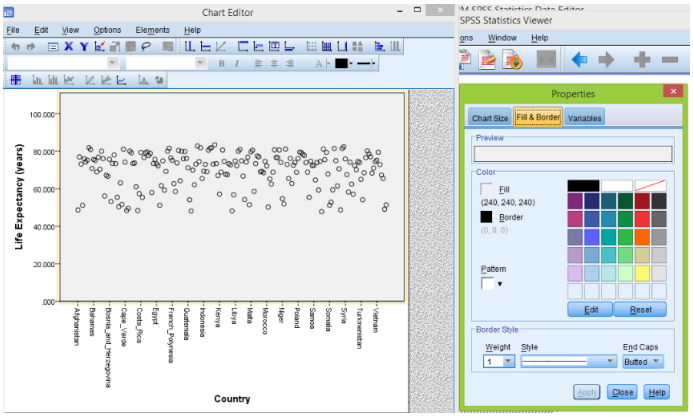
Then choose **Scatter/Dot > double click on Simple Scatter** to get:



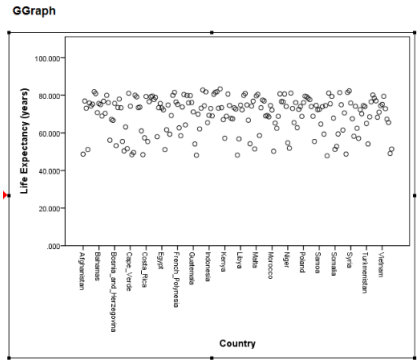
Then drag 'Life Expectancy' variable to the Y-axis and 'Country' to X-axis, click **OK** to get the following plot in the Output window:



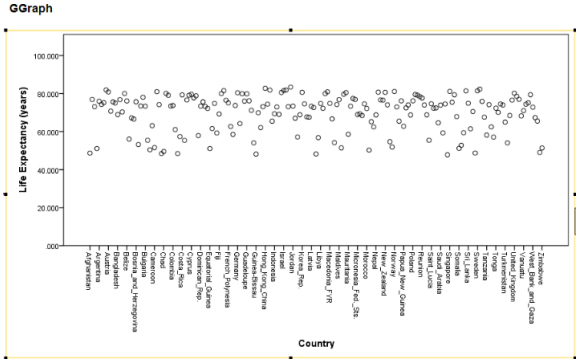
To make any changes to this plot like colors and other properties, **double click** on this plot leads to the 'Chart Editor'. To change the background color from grey to white (for example), **double click** on the background of the plot and get:



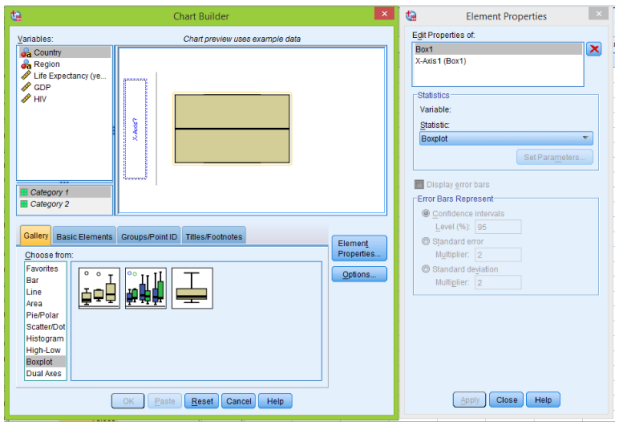
Choose white color, > **Apply** > **close Chart Editor** and get a new plot:



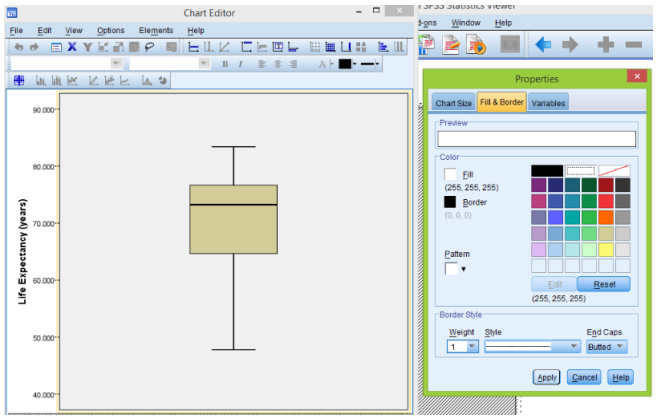
We can make this plot wider by dragging the right side of the plot to the right:



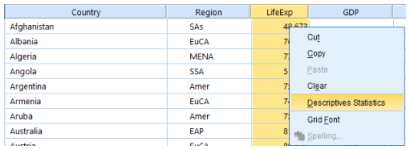
This was a very simple plot. Boxplots however produce more informative plots. From boxplot we can estimate the maximum, minimum, median, first quartile and third quartile. To draw the boxplot we start once again from 'Graphs' drop-down menu: **Graphs > Chart Builder > Boxplot > double click 1-D Boxplot**



As before drag 'Life Expectancy' to X-axis and click **OK**, the boxplot is produced. **Double click** on this plot opens 'Chart Editor'



If we want to find basic descriptive statistics of the 'LifeExp' variable (like minimum, maximum, mean, median and others), we can do that by **right click on the header of this variable > Descriptives Statistics**



Then get the following table in the 'Output' window:

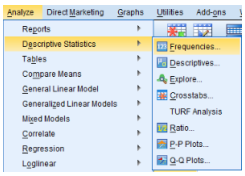
Frequencies

Statistics

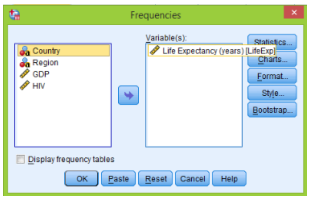
Life Expectancy (years)

N	Valid	197
	Missing	0
Mean		69.86282
Median		73.23500
Std. Deviation		9.668736
Range		35.600
Minimum		47.794
Maximum		83.394

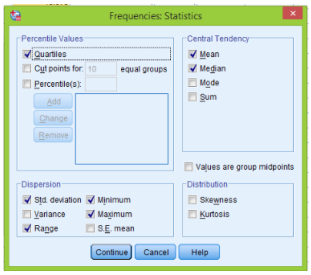
It shows that total number of observations is 197, no missing values and other different quantities. Note however, this table does not show first and third quartiles. To get these measures go to **Analyse** > **Descriptive Statistics** > **Frequencies**



Then select the 'Life Expectancy' variable and then click on the arrow to move this variable to the right window.



Then click on **Statistics**



Here select all the measures that we want. You see that now we have 'Quartiles' option for the first quartile, median and third quartile. To finish click **Continue** > **OK** to get the next table:

Frequencies

Statistics

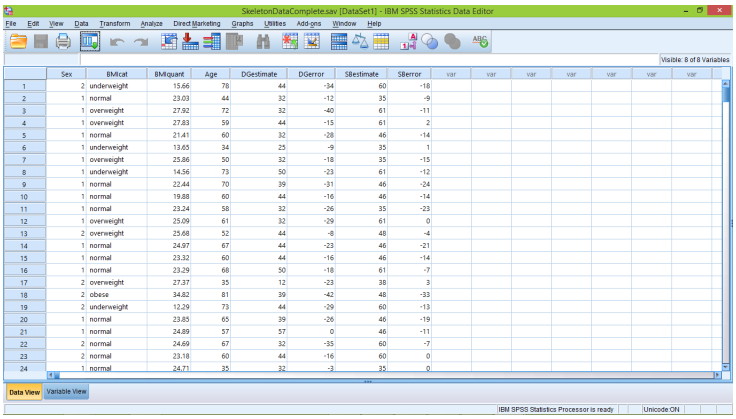
Life Expectancy (years)

N	Valid	197
	Missing	0
Mean		69.86282
Median		73.23500
Std. Deviation		9.668736
Range		35.600
Minimum		47.794
Maximum		83.394
Percentiles	25	64.44700
	50	73.23500
	75	76.74350

Here ‘Percentiles 25’ is the first quartile, ‘Percentiles 50’ and ‘Percentiles 75’ correspond to the median and third quartile respectively.

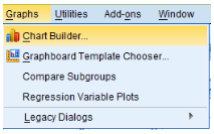
2 Center of Data

This section we are going to start with the Skeleton data.

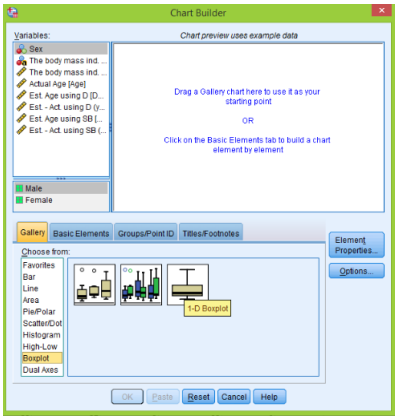


These data have 400 observations of skeletons, and we are interested in the error variable (‘DGerror’) which measures the difference between estimated and actual age using the method of Di Gangi.

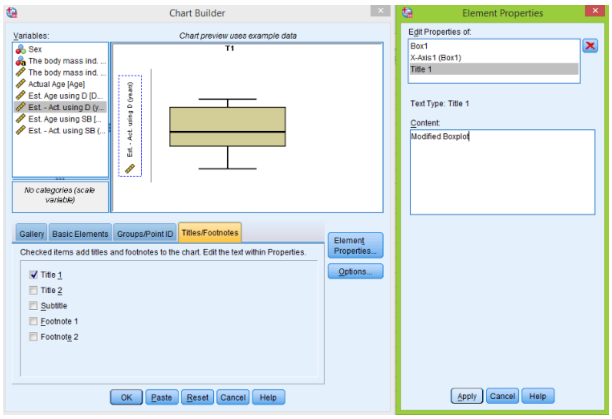
We want to plot the modified boxplot of the ‘DGerror’ variable (note that SPSS only produces modified boxplot that shows observations beyond fences), go to **Graphs** > **Chart Builder**



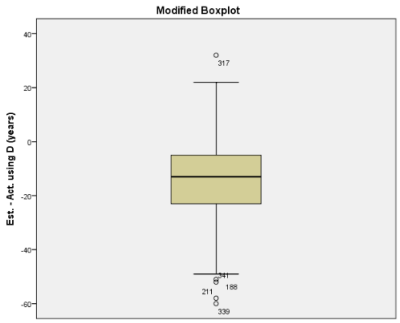
Then select **Boxplot** > **double click 1-D Boxplot**



As before drag the ‘DGerror’ variable into the X-axis of the boxplot. If you want to make a title, click on the **Titles/Footnotes** tab, then tick ‘Title 1’ and enter a title in the ‘Content’ window on the right side:

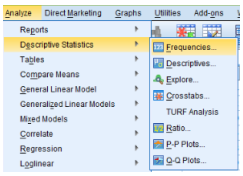


Click **Apply** > **OK** to get the plot:

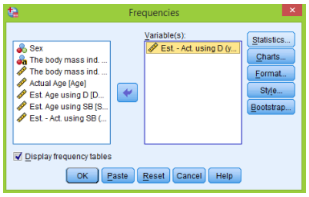


Note that now we have a title and some observations are beyond the fences.

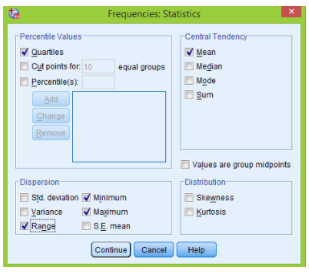
As before to get some statistics of this variable go to **Analyze** > **Descriptive Statistics** > **Frequencies**



Select 'Est. - Act. using D (years)' variable and click on the arrow to send this variable across:



Click on the 'Statistics' button and choose all the necessary quantities:



Continue > **OK** to produce the next table of statistics:

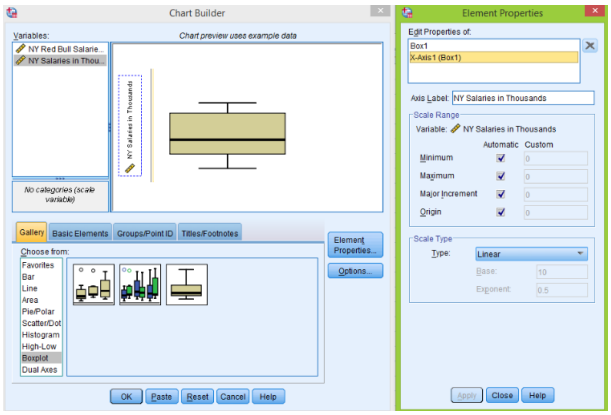
Frequencies

Statistics

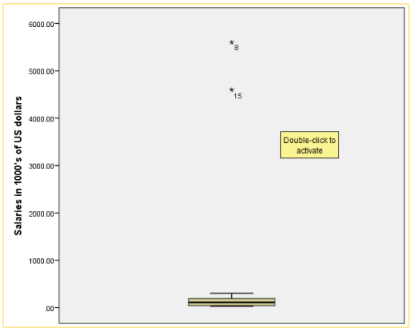
Est. - Act. using D (years)

N	Valid	400
	Missing	0
Mean		-14.15
Range		92
Minimum		-60
Maximum		32
Percentiles	25	-23.00
	50	-13.00
	75	-5.00

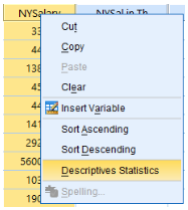
Next open another data set which consists of salaries for NY Red Bulls players. These data consist of only two variables. First variable is the original data of salaries, the second one is the salaries divided by thousand to make graphs scale better. To make a good visual representation of these data we use boxplot as before. To make a different label to the vertical axis, click on X-Axis1(Box1) in the 'Element Properties' window:



'Axis Label' now shows the usual label for this variable. You can easily change it to any other label; as usual **Apply** > **OK** to finish:



We immediately see two outliers that probably will affect the mean of this sample. To get some simple statistics of the 'NYSalary' variable **right click of the header of this variable** > **Descriptives Statistics**



And get the following table in the 'Output':

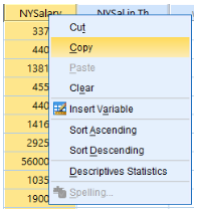
Frequencies

Statistics

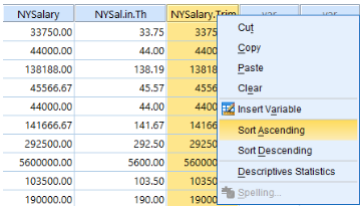
NY Red Bull Salaries

N	Valid	25
	Missing	0
Mean		518311.6352
Median		112495.5000
Std. Deviation		1388822.106
Range		5566250.00
Minimum		33750.00
Maximum		5600000.00

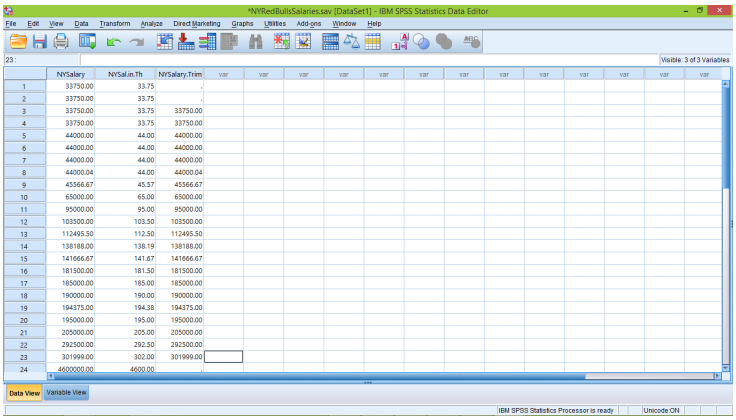
As expected there is a big difference between mean and median. Hence we want to trim the data and find a new mean. To do that copy the first column by **right click on the header > Copy**



Copy it to the third column and name this column 'NYSalary.Trim'. To trim the variable by 8% we need to delete 2 largest and 2 smallest observations (since 25 times 8% is 2). Hence we need to sort the variable first by **right click on the header > Sort Ascending**



Then just remove two smallest and two largest observations manually:



Now we get basic statistics for this trimmed variable:

Frequencies

Statistics

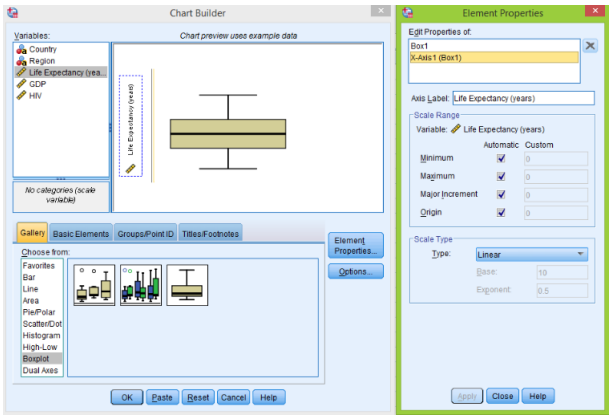
NY Red Bull Salaries

N	Valid	21
	Missing	4
Mean		128109.0895
Median		112495.5000
Std. Deviation		83990.81129
Range		268249.00
Minimum		33750.00
Maximum		301999.00

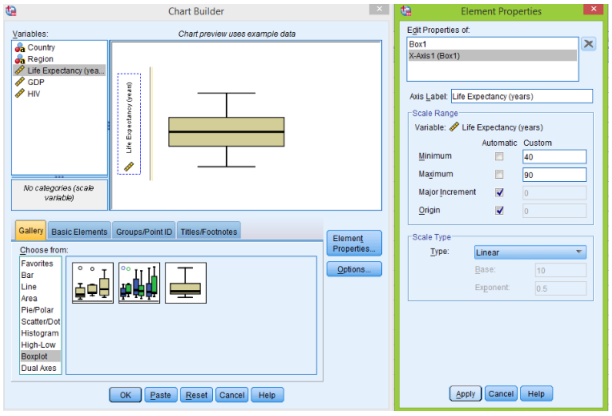
See how the mean changes when we trimmed the data.

3 Spread of the Data

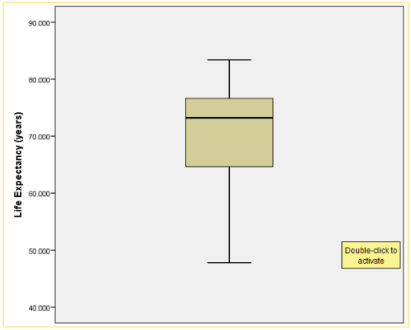
We start with the 'Life Expectancy' data set. We will analyse again the 'LifeExp' variable which contains the mean life expectancies values. Lets plot the boxplot of this variable again. As usual go to **Graphs** > **Chart Builder** > **Boxplot**. After selecting the 'LifeExp' variable, if you want to change the scale of the vertical axis of the boxplot, then under 'Element Properties' select 'X-Axis1 (Box1):



Then deselect the Automatic option of Maximum and Minimum and enter the appropriate values (we want the vertical axis to be from 40 to 90):



Finishing you get the usual plot:



To get basic statistics of this variable, as before **right click on the header of 'LifeExp'** > **Descriptives Statistics** and get the table:

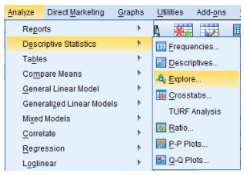
Frequencies

Statistics

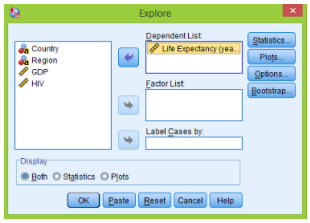
Life Expectancy (years)

N	Valid	197
	Missing	0
Mean		69.86282
Median		73.23500
Std. Deviation		9.668736
Range		35.600
Minimum		47.794
Maximum		83.394

As you know we have several measures of the spread of quantitative data: range, IQR and standard deviation. The above table shows only the range and standard deviation, to get also IQR (interquartile range) we can do the following: go to **Analyze** > **Descriptive Statistics** > **Explore**



Select 'Life Expectancy' variable from the left window and click on the arrow to move it to the right:



In the Display option below select 'Statistics' since we do not need plots here, click **OK** to finish the process and get this table in the Output window:

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Life Expectancy (years)	197	100.0%	0	0.0%	197	100.0%

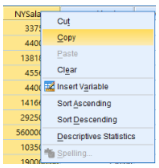
Descriptives

			Statistic	Std. Error
Life Expectancy (years)	Mean		69.86282	.688869
	95% Confidence Interval for Mean	Lower Bound	68.50427	
		Upper Bound	71.22136	
	5% Trimmed Mean		70.36401	
	Median		73.23500	
	Variance		93.484	
	Std. Deviation		9.668736	
	Minimum		47.794	
	Maximum		83.394	
	Range		35.600	
	Interquartile Range		12.296	
	Skewness		-.832	.173
	Kurtosis		-.399	.345

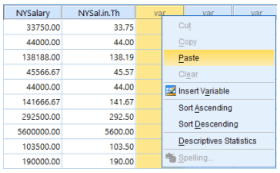
This table contains much more information than previous table. Here we have mean, 5% trimmed mean, median, range, standard deviation and IQR. Note that standard deviation is exactly the same as square root of the variance.

Next we want to compare different measures of data (statistics) for robustness using NY Red Bull salaries. Open this data file. From the last section we remember that this variable has two large outliers, so lets investigate which measures that we have learned change a lot in the presence of outliers and which are not (robust).

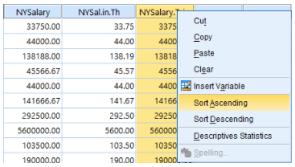
First copy the 'NYSalary' variable using **right click on header of 'NYSalary' > Copy**:



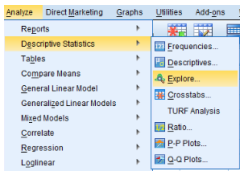
Then **right click on 'var' of the third column > Paste**:



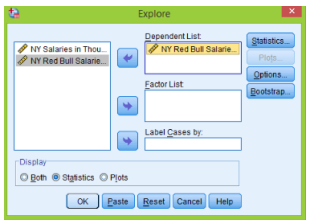
Then make a name for this column (we call it 'NYSalary.Trim'). Next sort this variable **right click on header > Sort Ascending**:



Then manually delete two largest and two smallest values. Now we have a column of the trimmed data. To find all the necessary statistics for the original variable, as before go **Analyse > Descriptive Statistics > Explore**:



Select 'NYSalary' variable and send it to the 'Dependent List' using arrow and select 'Statistics' in the display option:



Click **OK** to get the table for the original variable:

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
NY Red Bull Salaries	25	100.0%	0	0.0%	25	100.0%

Descriptives

			Statistic	Std. Error
NY Red Bull Salaries	Mean		518311.6352	277764.4211
	95% Confidence Interval for Mean	Lower Bound	-54965.9540	
		Upper Bound	1091589.224	
	5% Trimmed Mean		274026.8169	
	Median		112495.5000	
	Variance		1.929E+12	
	Std. Deviation		1388822.106	
	Minimum		33750.00	
	Maximum		5600000.00	
	Range		5566250.00	
	Interquartile Range		150687.50	
	Skewness		3.341	.464
	Kurtosis		10.182	.902

Now do exactly the same but for the trimmed variable, and get the following table:

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
NY Red Bull Salaries Trimmed	21	84.0%	4	16.0%	25	100.0%

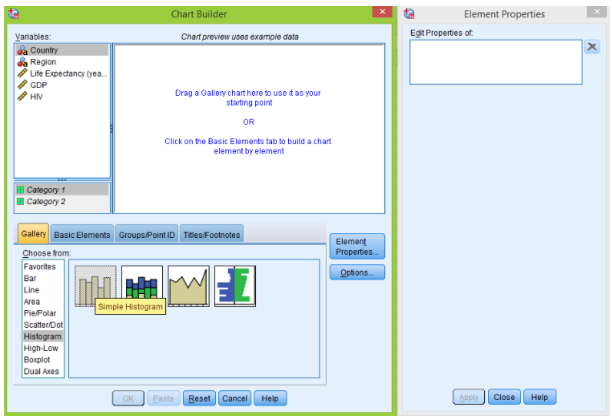
Descriptives

			Statistic	Std. Error
NY Red Bull Salaries Trimmed	Mean		128109.0895	18328.29764
	95% Confidence Interval for Mean	Lower Bound	89876.9306	
		Upper Bound	166341.2484	
	5% Trimmed Mean		123715.8402	
	Median		112495.5000	
	Variance		7054456381	
	Std. Deviation		83990.81129	
	Minimum		33750.00	
	Maximum		301999.00	
	Range		268249.00	
	Interquartile Range		148187.48	
	Skewness		.596	.501
	Kurtosis		-.557	.972

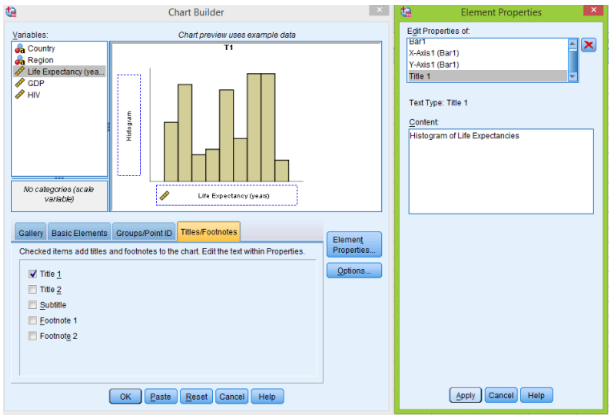
See that this variable has only 21 valid values and 4 are considered as missing because we have deleted 4 values. Based on these two tables we can conclude that median and IQR almost did not change after trimming the data and therefore are robust to outliers, on the other hand mean, range and standard deviation changed a lot by trimming and hence are not robust to outliers.

4 Shape of the Data

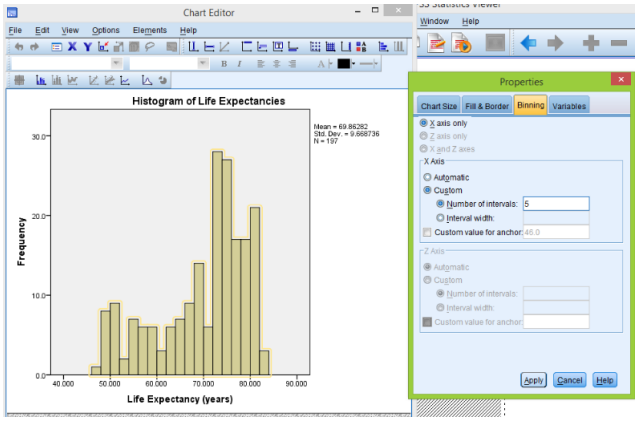
In this section we introduce histograms and investigate shapes of different variables. Lets start with plotting the histogram of the life expectancy data: go to **Graphs** > **Chart Builder** > **Histogram** > **double click on Simple Histogram**



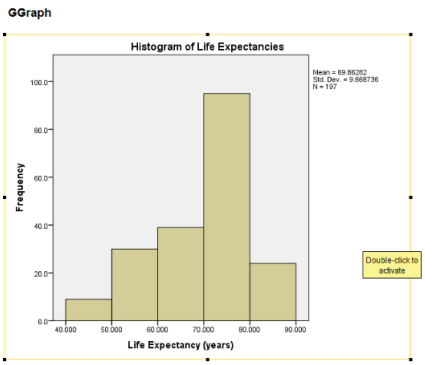
Then drag the 'Life Expectancy' variable to the horizontal axis. Also click on the 'Titles/Footnotes' tab to give a title for this plot (same as with boxplot):



Click **Apply** > **OK** to get the histogram of this variable; this histogram is noisy because it has many bins. To change the number of bins **double click on the plot** to open the 'Chart Editor', then **double click** on the bins opens the 'Properties' window, click on 'Binning'. Select the 'Custom' option and choose the number of intervals (for example 5):



Click **Apply** and close the 'Chart Editor' to get modified histogram:



Thus we can change the number of bins to make the plot more or less noisy. As before we can get the basic statistics of the Life Expectancy variable by **right click on the header of the column** > **Descriptives Statistics** and get the table:

Frequencies

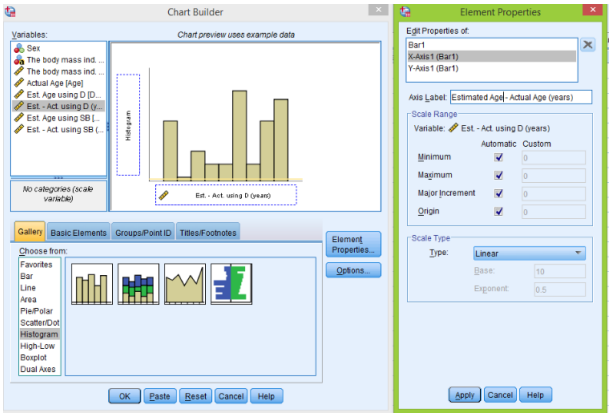
Statistics

Life Expectancy (years)

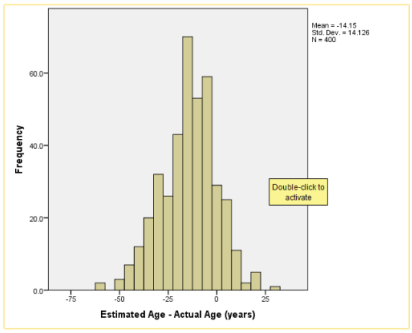
N	Valid	197
	Missing	0
Mean		69.86282
Median		73.23500
Std. Deviation		9.668736
Range		35.600
Minimum		47.794
Maximum		83.394

Since the mean is smaller than median and based on the histogram we can conclude that these data are skewed to the left (has longer left tail).

Now lets consider 'Skeleton' data set and make the histogram of 'DGerror' variable. As explained earlier go to **Graph** > **Chart Builder** > **Histogram** > **Simple Histogram**, then drag the 'Est. - Act. using D (years)' variable to the horizontal axis. The horizontal label of the histogram would be the label of this variable, if we want to change the label then under 'Elements Properties' select 'X-Axis1 (Bar1)' and change the 'Axis Label':



Do not forget to click **Apply** > **OK** to produce the histogram:



Next we find descriptive statistics of the 'DGerror' variable as usual and get:

Frequencies

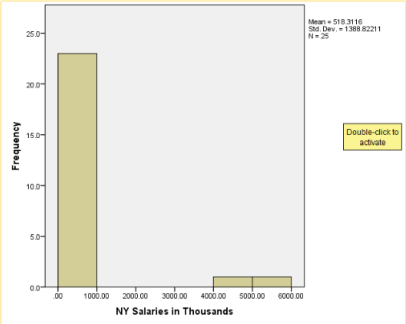
Statistics

Est. - Act. using D (years)

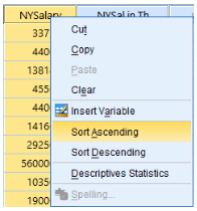
N	Valid	400
	Missing	0
Mean		-14.15
Median		-13.00
Std. Deviation		14.126
Range		92
Minimum		-60
Maximum		32

Based on the plot and since mean and median are almost the same we see that the distribution of differences is symmetric. The last data set that we will analyse in this section is 'NY Red Bull Salaries'. Plotting the histogram of these data we again notice two large outliers:

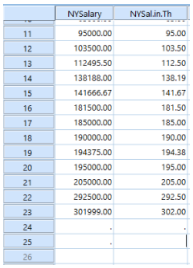
GGraph



To make a histogram without these two values, first sort the original data by **right click on the header > Sort Ascending**

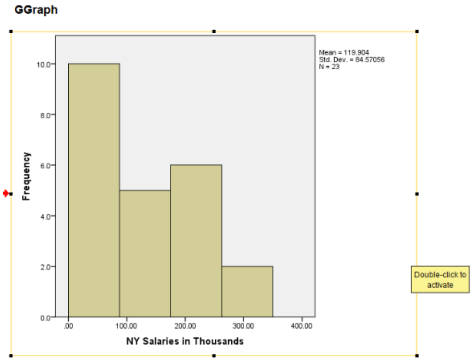


After the sorting, just delete two largest values:



	NYSalary	NYSal.in.Th
11	95000.00	95.00
12	103500.00	103.50
13	112495.50	112.50
14	138188.00	138.19
15	141666.67	141.67
16	181500.00	181.50
17	185000.00	185.00
18	190000.00	190.00
19	194375.00	194.38
20	195000.00	195.00
21	205000.00	205.00
22	292500.00	292.50
23	301999.00	302.00
24	.	.
25	.	.
26		

Now make a histogram of the second trimmed column (since it is in thousands) and get:



Then get descriptive statistics of the first column (trimmed NYSalaries):

Frequencies

Statistics

NY Red Bull Salaries

N	Valid	23
	Missing	2
Mean		119903.9513
Median		103500.0000
Std. Deviation		84570.55807
Range		268249.00
Minimum		33750.00
Maximum		301999.00

This data set has mean which is larger than median and looking on the plot we say that these data are skewed to the right (or has longer right tail).

Now lets return to the 'DGerror' variable, and we want to check that it follows the empirical rule. Since the distribution of this variable is unimodal and symmetric we expect to get close to theoretical results. First we find mean and standard deviation of the error variable using descriptive statistics:

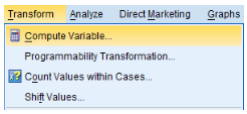
Frequencies

Statistics

Est. - Act. using D (years)

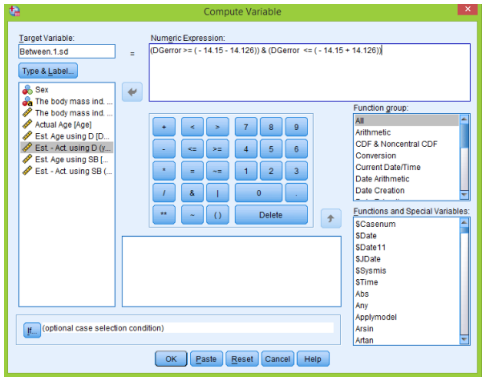
N	Valid	400
	Missing	0
Mean		-14.15
Median		-13.00
Std. Deviation		14.126
Range		92
Minimum		-60
Maximum		32

Note that the mean is -14.15 and standard deviation is 14.126. Next we want to see which observations are between mean plus/minus one standard deviation and which are not. To do that go to **Transform** > **Compute variable**:

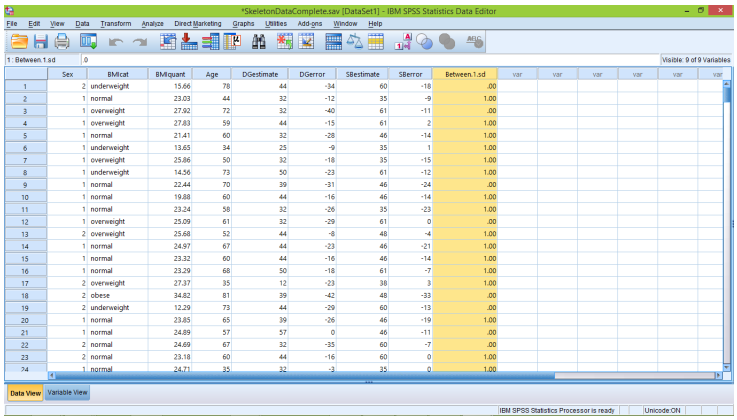


Here we construct a new 'Target variable' (which we name 'Between.1.sd'). In the numeric expression we enter the logical equation which takes two values: 1 if expression is true and 0 if it is false. We get 1 if 'DGerror' is greater than $-14.15 - 14.126$ (mean $-$ one standard deviation) and (& means and) smaller than $-14.15 + 14.126$ (mean $+$ one standard deviation) otherwise

zero. If both of these conditions hold then the value of the new variable is going to be 1 and these conditions are equivalent for 'DGerror' to be between mean and plus/minus one standard deviation.



Click on the **OK** button and get a new column of 1's and 0's:



Then to find proportion of values between mean plus/minus one standard deviation we just find descriptive statistics of the new variable ('Between.1.sd'):

Frequencies

Statistics

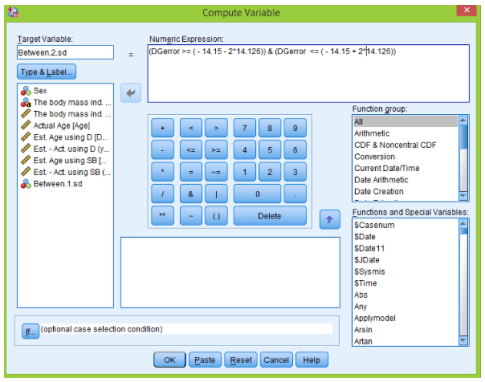
Between.1.sd

N	Valid	400
	Missing	0
Mode		1.00
Range		1.00
Minimum		.00
Maximum		1.00

Between.1.sd

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	127	31.8	31.8	31.8
	1.00	273	68.3	68.3	100.0
	Total	400	100.0	100.0	

We see that 68.3% of observations are within this range. Similarly we do it for two standard deviations. Construct a new variable:



Then get frequencies for this new variable:

Frequencies

Statistics

Between.2.sd

N	Valid	400
	Missing	0
Mode		1.00
Range		1.00
Minimum		.00
Maximum		1.00

Between.2.sd

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	20	5.0	5.0	5.0
	1.00	380	95.0	95.0	100.0
	Total	400	100.0	100.0	

We see that 95% are between mean and plus/minus two standard deviations. For the three standard deviations:

Frequencies

Statistics

Between.3.sd

N	Valid	400
	Missing	0
Mode		1.00
Range		1.00
Minimum		.00
Maximum		1.00

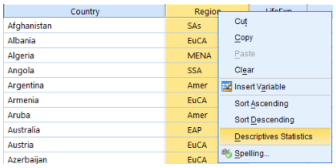
Between.3.sd

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	3	.8	.8	.8
	1.00	397	99.3	99.3	100.0
	Total	400	100.0	100.0	

Hence 99.3% are between mean plus/minus three standard deviations. The proportions that we get are very close to theoretical empirical rule!

5 Categorical Variables

In this section we focus on the visual representation of categorical variables. We start with the variable 'Region' from Life Expectancy data file that shows to which regions different countries belong. The variable of regions is of course categorical with six classes. First we make a table of counts and frequencies, **right click on Region header > Descriptives Statistics**



Then get the next table:

Frequencies

Statistics

Region

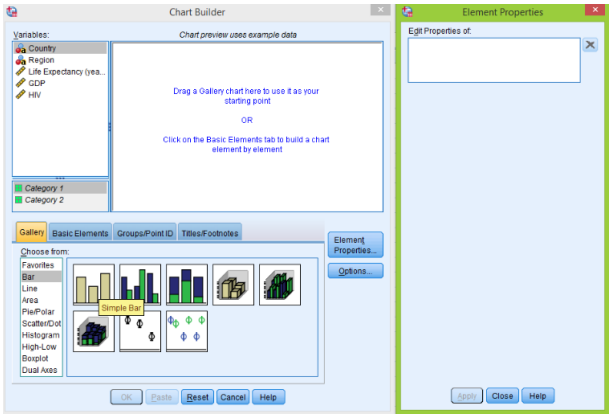
N	Valid	197
	Missing	0

Region

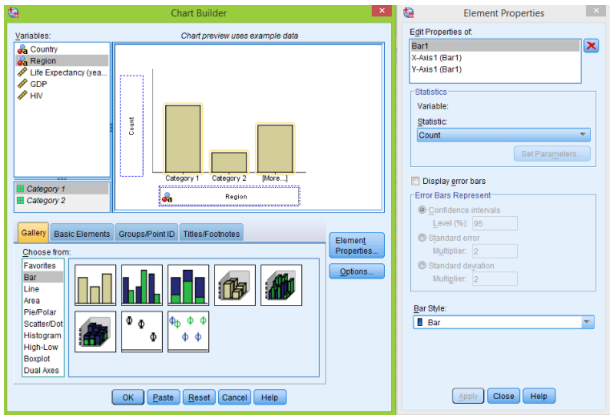
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Amer	39	19.8	19.8	19.8
	EAP	30	15.2	15.2	35.0
	EuCA	50	25.4	25.4	60.4
	MENA	21	10.7	10.7	71.1
	SAs	8	4.1	4.1	75.1
	SSA	49	24.9	24.9	100.0
	Total	197	100.0	100.0	

The first column of 'Frequencies' shows how many observations are in each category while the second 'Percent' column shows the relative frequency which is just count divided by total number of observations.

Next we want to make a visual representation of this variable. We start with a bar-plot of counts: **Graphs > Chart Builder > Bar > double click on Simple Bar**:

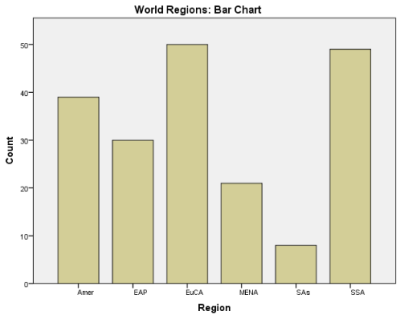


Then drag 'Region' variable to the horizontal axis, 'Count' in vertical axis appears automatically:

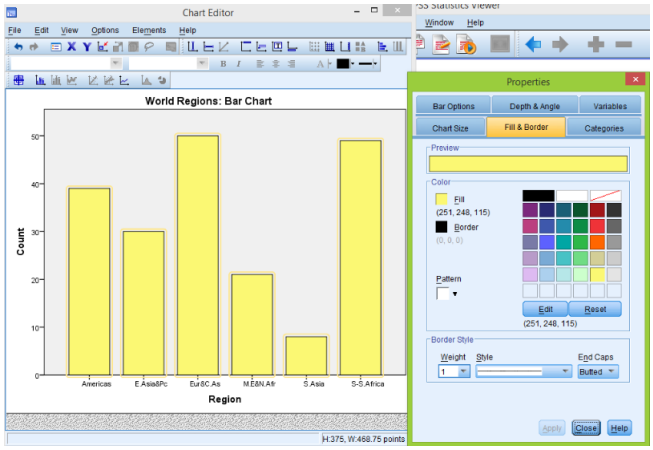


As before you can add a title to this plot and then click **OK** to get the bar-plot of counts:

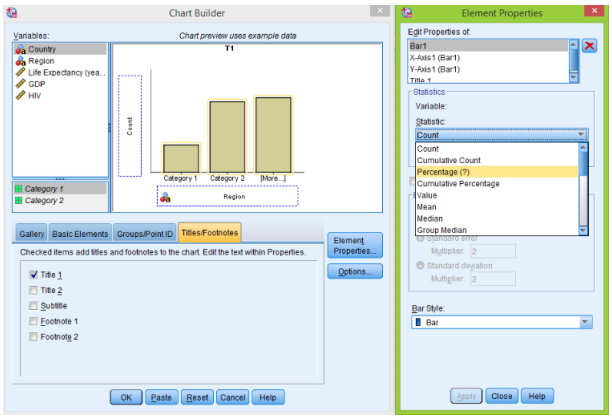
GGraph



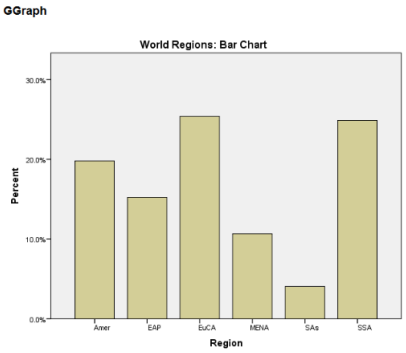
To make changes to the plot, **double click** on it to open 'Chart editor'. If you want to change the bar labels (now they are 'Amer', 'EAP',,) just **double click** on any of these labels and type other descriptions. To change color of the bars **double click** on any bar of the plot to open the 'Properties' window and select 'Fill & Border':



Now you can change the color then click **Apply** and close the 'Chart Editor'. To make a bar-plot of relative frequencies, as before **Graphs > Chart Builder > Bar > double click on Simple Bar**, drag 'Region' to the horizontal axis, open the 'Elements Properties' window and under 'Statistic' select 'Percentage' instead of 'Count':

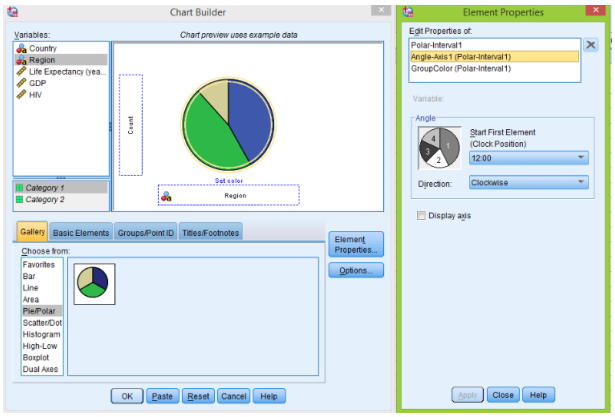


Finish with **Apply** > **OK** and get:

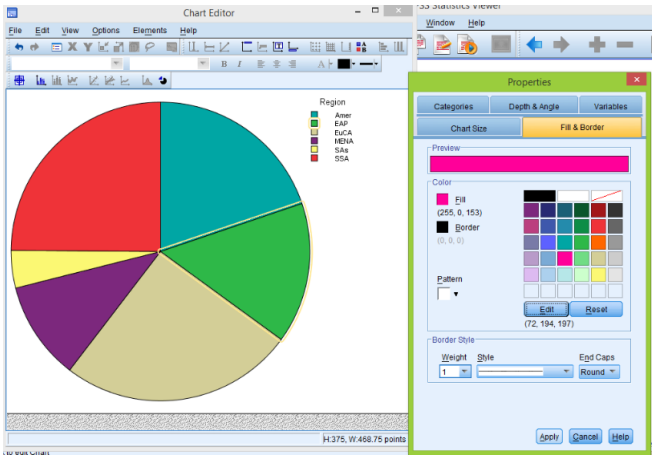


Now we see that vertical axis is 'Percent' which is just relative frequency.

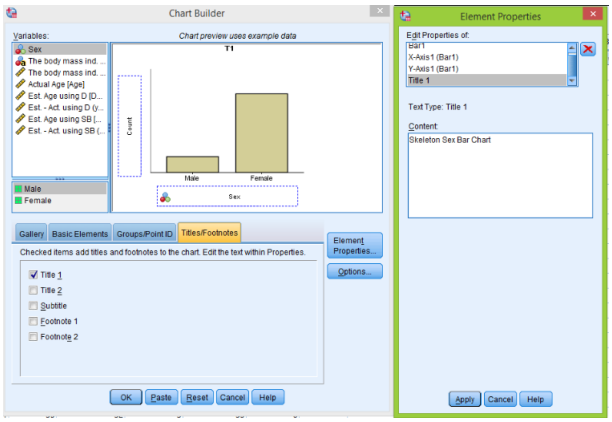
Another useful plot for categorical variables is a pie chart. Go to **Graphs** > **Chart Builder** > **Pie/Polar** > **double click on Pie Chart**, then drag 'Region' variable to the horizontal axis:



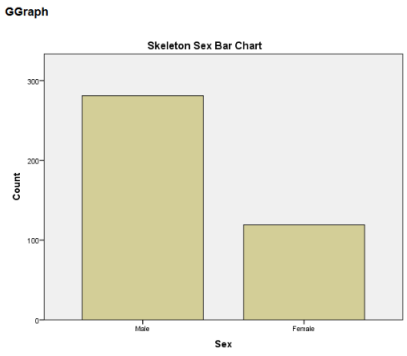
Click **OK** to get the Pie chart in the output window. To change the color of each individual slice, **double click** on the plot to open the 'Chart Editor' then **double click** on the slice we want change to open the 'Properties' window then select 'Fill & Border' and change the color:



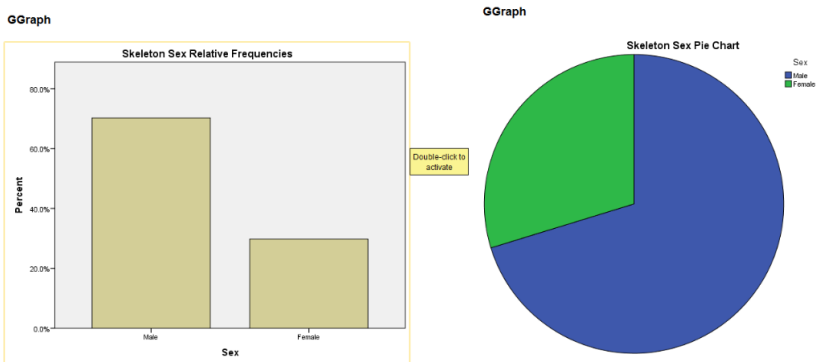
Now we return to the 'Skeleton' data and first analyse 'Sex' variable. Using the same procedure as before we make a bar-chart for counts. To make a title click on 'Titles/Footnotes' in the 'Chart Builder' window, select 'Title 1' and enter the title in the 'Content' window on the right:



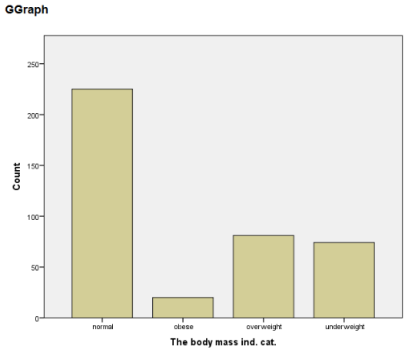
Click **Apply** > **OK** to get the bar-chart for the 'Sex' variable:



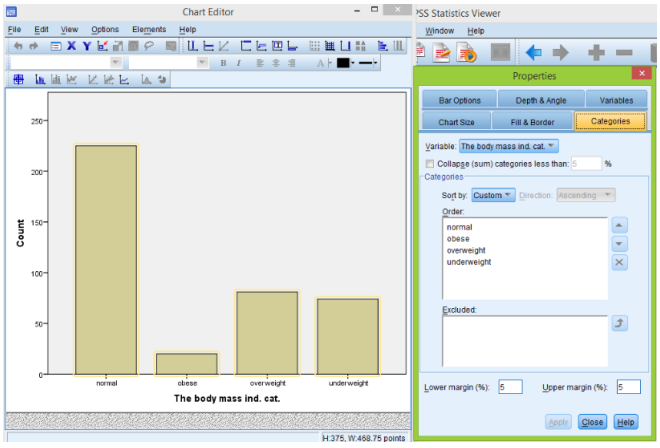
Similarly we make a bar-chart of relative frequencies and a Pie chart:



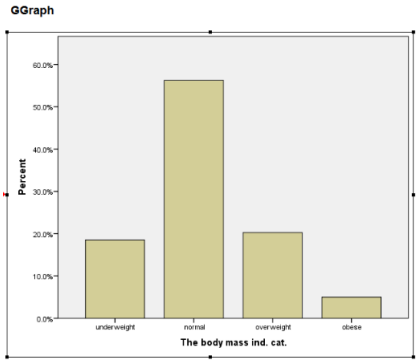
Similarly we do for the 'BMIcat' variable from 'Skeleton' file. First get the bar-chart for counts:



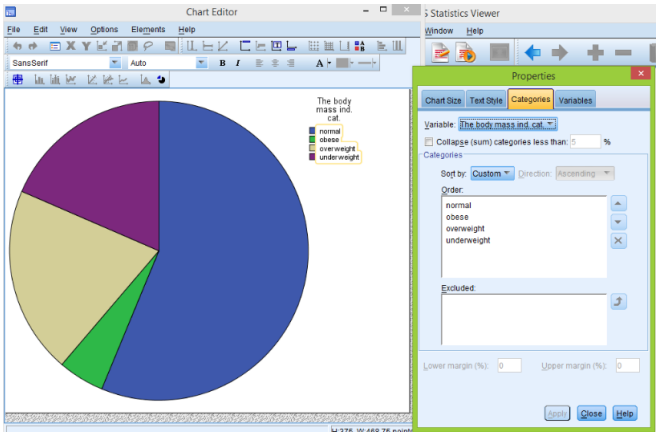
We see however that order of these bars is not logical. We probably want them to be from left to right: 'underweight', 'normal', 'overweight' and 'obese'. To change the order **double click** on the plot to open the 'Chart Builder', then **double click** on any bar to open the 'Properties' window and select 'Categories':



Now under 'Order' select labels and move them with the arrows to right positions. Then click **Apply**, close the 'Chart Editor' and get a modified bar chart. Similarly we can do for the relative frequencies



Finally we get the Pie chart for this variable; to change the order of slices **double click** on the chart, **double click** on any slice and select 'Categories':



Change the order and get the following pie chart with better ordering: