



Statistics 1 & 2

Practical 1: Introduction to SPSS and Descriptives

Assignment 1: Introduction to SPSS

Assignment 1 – Question 1

stressLAS.sav [DataSet1] - IBM SPSS Statistics Data Editor

	STRESS	SMOKE	RELATION	OPTIM	SATIS	NEGEMO	var	var
1	1	1	1	23	26	10		
2	1	1	1	20	26	15		
3	1	1	1	23	26	13		
4	1	1	1	19	22	15		
5	1	1	1	22	30	13		
6	1	1	1	25	30	8		
7	1	1	1	22	35	21		
8	1	1	1	27	36	16		
9	1	1	1	34	31	9		
10	1	1	1	19	36	21		
11	1	1	1	27	39	14		
12	1	1	1	25	34	17		
13	1	1	1	22	29	16		
14	1	1	1	21	27	13		
15	1	1	1	25	25	8		
16	1	1	1	22	28	16		
17	1	1	1	23	29	12		
18	1	1	1	21	28	13		
19	1	1	1	29	22	6		
20	1	1	1	26	29	14		

Data View Variable View

- The rows represent individual participants.
- The columns represent the variables that were measured. It shows the scores the participants scored on the variables.

Assignment 1 – Question 2

STRESSLAB.sav [DataSet1] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	STRESS	Numeric	1	0	Source of stress	{1, No Stres...	None	8	Right	Nominal	Input
2	SMOKE	Numeric	1	0	Smoking Behav...	{1, Non-smo...	None	8	Right	Nominal	Input
3	RELATION	Numeric	1	0	Relation	None	None	8	Right	Nominal	Input
4	OPTIM	Numeric	2	0	Optimism	None	None	8	Right	Scale	Input
5	SATIS	Numeric	2	0	Life satisfaction	None	None	8	Right	Scale	Input
6	NEGEMO	Numeric	2	0	Negative Emoti...	None	None	8	Right	Scale	Input

- Name: the technical name of the variable. Can max. be 8 characters and needs to be unique.
- Label: the description of the variable. The label is often shown in the output, thus a meaningful label helps you read the output.
- To understand the difference between name and label, think of our student administration. Each student has a unique SNR to be used in the information system, but also a family name, which may not be unique (Janssen). By using the SNR we can be sure the grade is assigned to the correct Janssen.
- Values: shows what the different scores on categorical variables indicate
- Measure: what type of variable it is: nominal, ordinal or scale (continuous).

Assignment 1 – Examining Value Labels & Question 3

- To examine the value labels, click on the relevant box in the 'Values' column. A blue box with 3 dots will then appear. Click on the box and the overview will appear.

stressLAS.sav [DataSet1] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	STRESS	Numeric	1	0	Source of stress	Stress}... ..	None	8	Right	Nominal	Input
2	SMOKE	Numeric	1	0	Smoking Behav...	{1, Non-smo...	None	8	Right	Nominal	Input

- Value labels of STRESS:

Value Labels

Value:

Label:

1 = "No Stress"
2 = "Work-related stress"
3 = "Personal-life related stress"

Add
Change
Remove

Spelling...

OK Cancel Help

- Value labels of SMOKE:

Value Labels

Value:

Label:

1 = "Non-smoker"
2 = "Smoker"

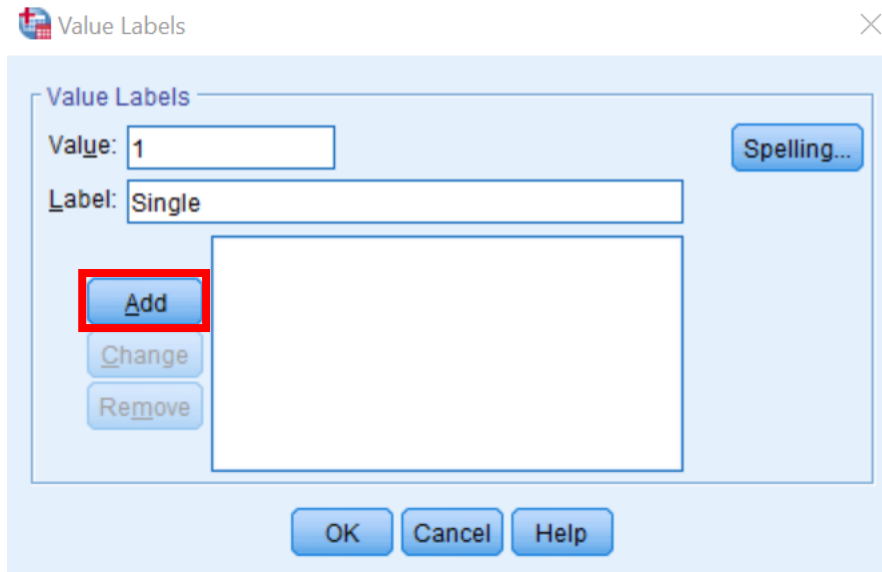
Add
Change
Remove

Spelling...

OK Cancel Help

Assignment 1 – Adding Value Labels + Question 3

- To add value labels, you also need to click on the blue box shown on the previous slide.
- In the menu that appears, fill in the relevant value and the corresponding value label:



The screenshot shows a 'Value Labels' dialog box. It has a title bar with a close button. The main area is titled 'Value Labels' and contains a 'Value:' text box with the number '1' and a 'Label:' text box with the word 'Single'. To the right of these text boxes is a 'Spelling...' button. Below the text boxes are three buttons: 'Add' (highlighted with a red rectangle), 'Change', and 'Remove'. At the bottom of the dialog are 'OK', 'Cancel', and 'Help' buttons.

- Then, click on add and repeat the steps until you have added all value labels. Finally, click on 'OK'.

Assignment 1 – Question 4

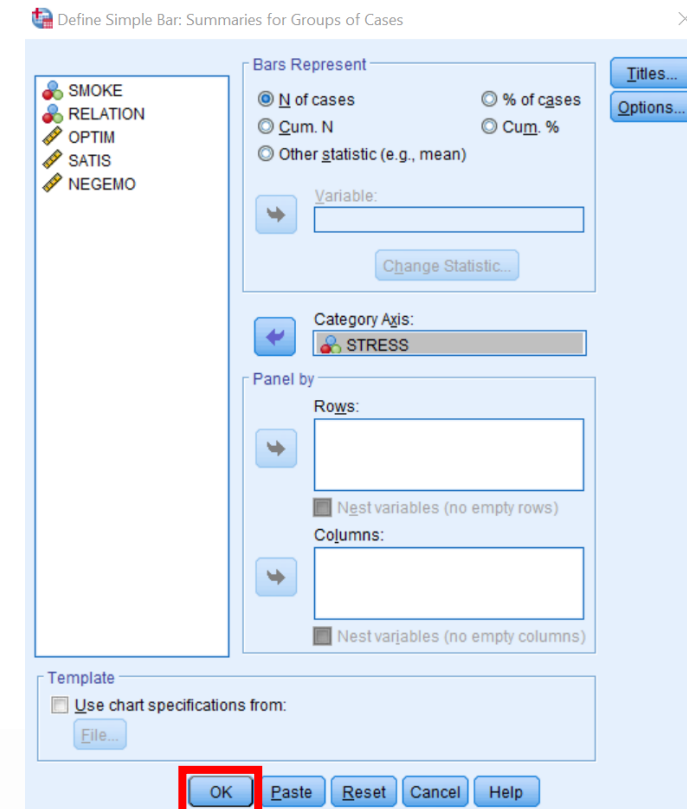
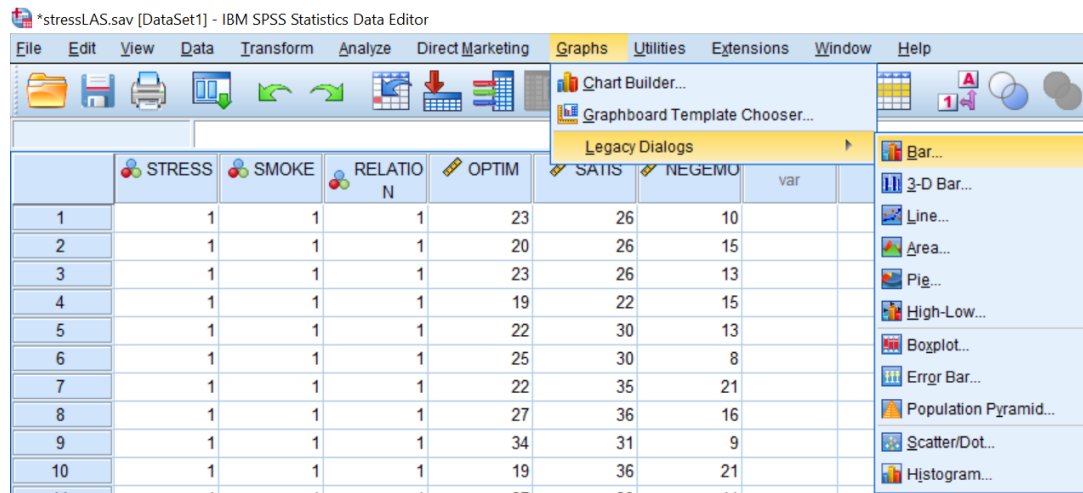
- There are 3 different measurement levels in SPSS:
 1. Nominal: this is a categorical variable, without an intrinsic ordering to the categories. An example is hair colour: 1 = blond, 2 = brown, 3 = black, 4 = red, 5 = other.
 2. Ordinal: this is a categorical variable, but here there is an intrinsic ordering to the categories. An example is education level: 1 = primary school, 2 = high school, 3 = some college, 4 = college degree.
 3. Scale: this is a continuous variable. An example is age.

In our current dataset, STRESS, SMOKE and RELATION are nominal variables. OPTIM, SATIS and NEGEMO are scale variables.

Assignment 2: Exploring Data - Graphical Summaries

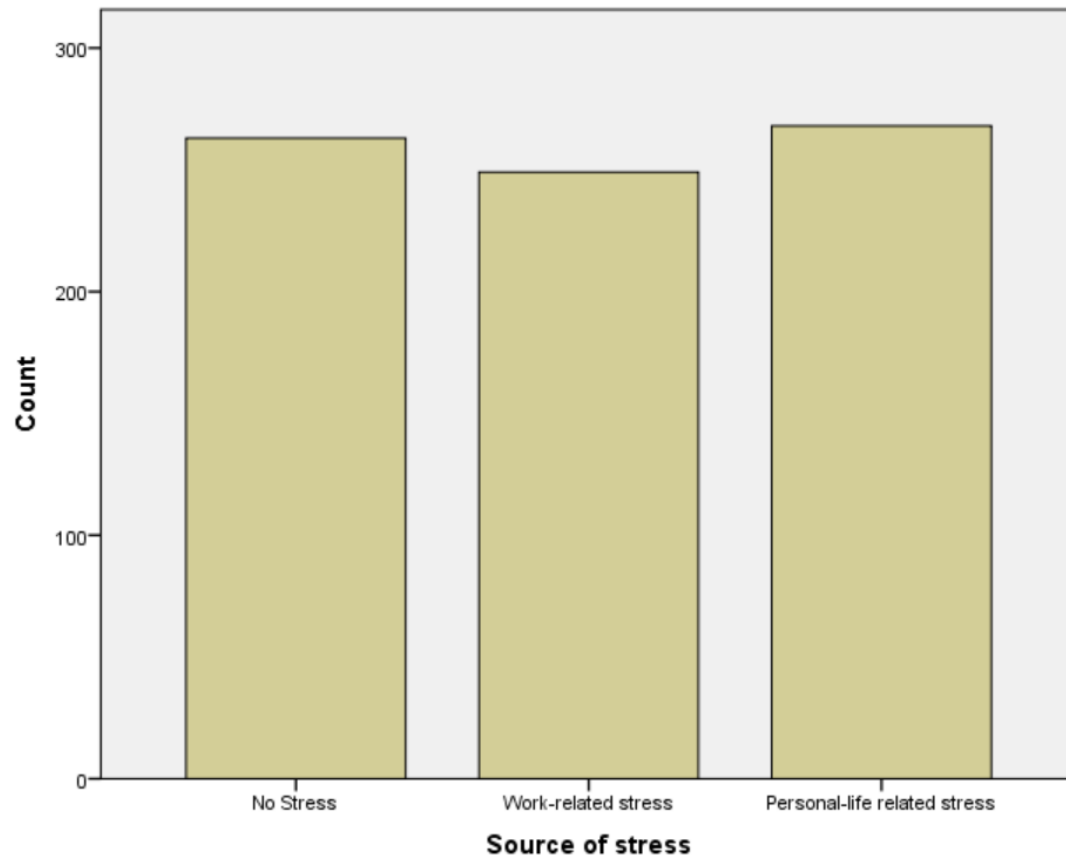
Assignment 2 – Creating a Bar Chart

- To create a bar chart, navigate to *Graphs > Legacy Dialogs > Bar*. Then, after choosing 'Simple' and clicking on 'Define' (not pictured), add the variable STRESS to the 'Category Axis'. Finally, click on OK.



Assignment 2 – Question 1

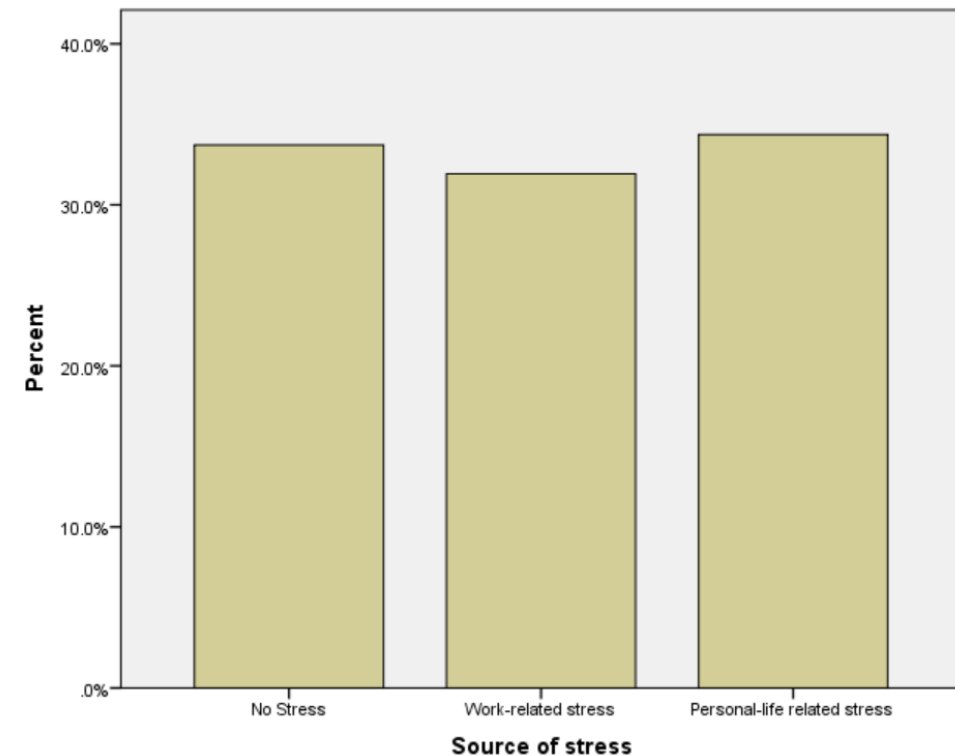
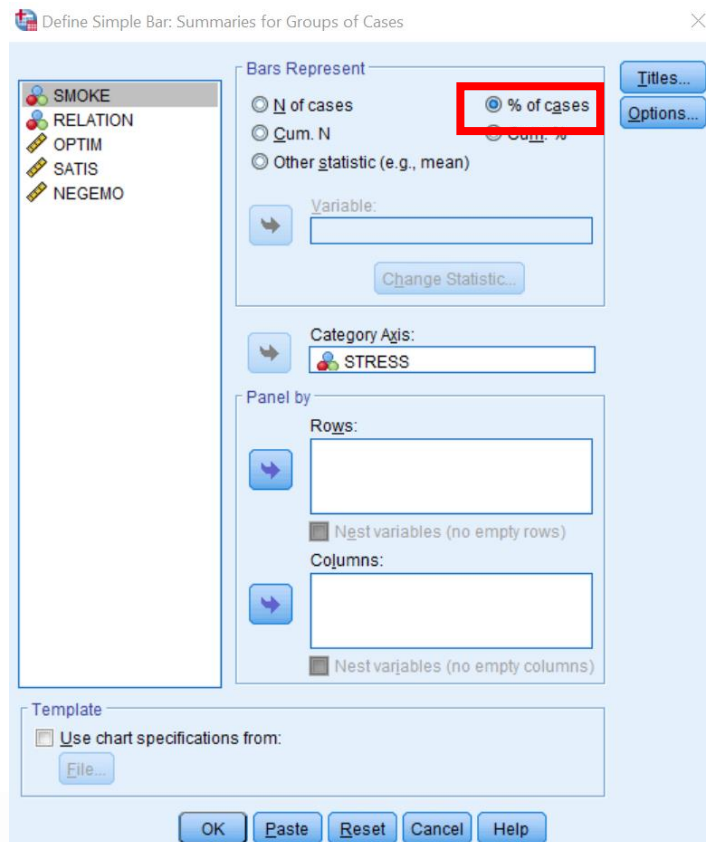
- In the output, you should see the following bar chart:



- The stress level seems almost evenly distributed; there are about as many respondents with no stress level as there are respondents with stress related to work or personal life.
- This does mean that there are about twice as many people who experience stress than there are people that experience no stress.

Assignment 2 – Bar Chart with % of Cases

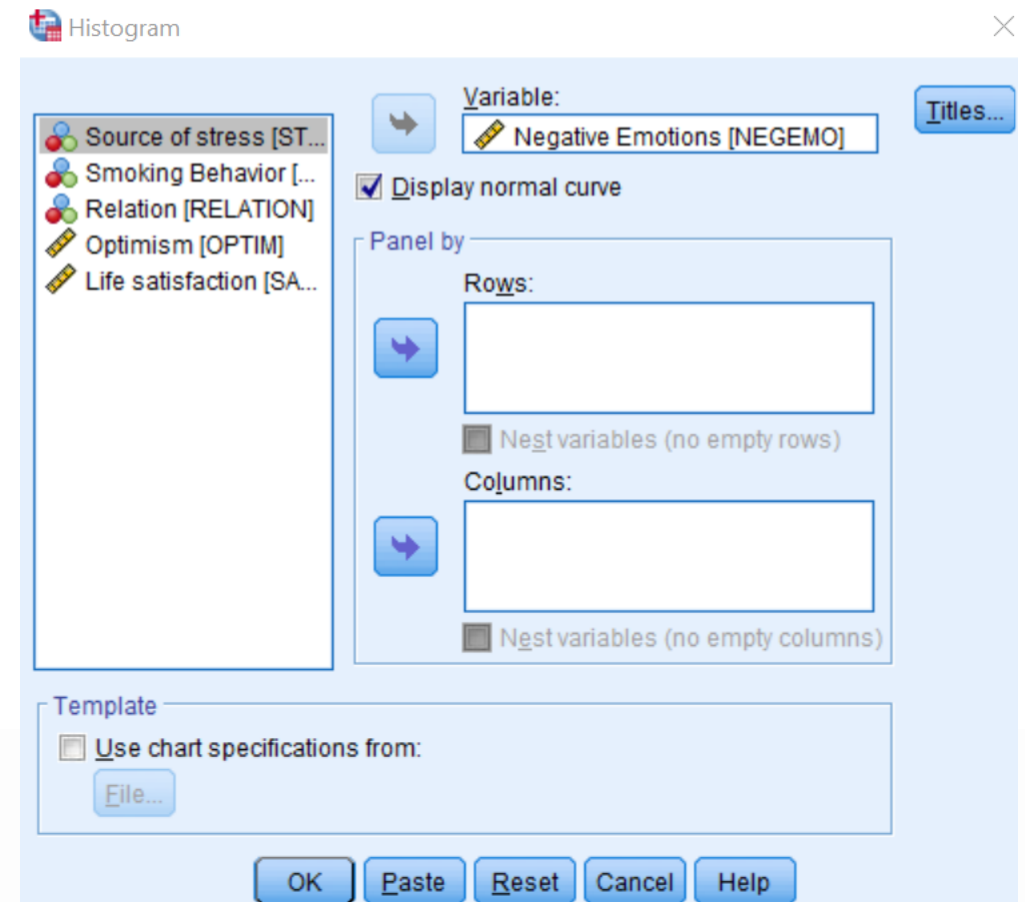
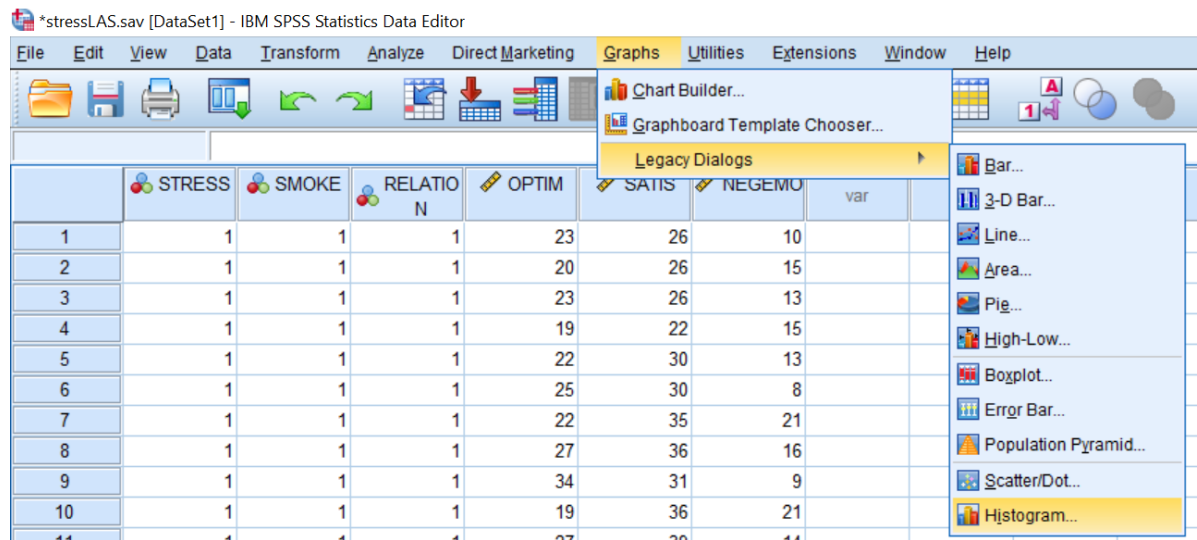
- To make a bar chart using percentages on the X-axis, make sure you check ' % of cases ' under ' Bars represent '. This creates the following bar chart:



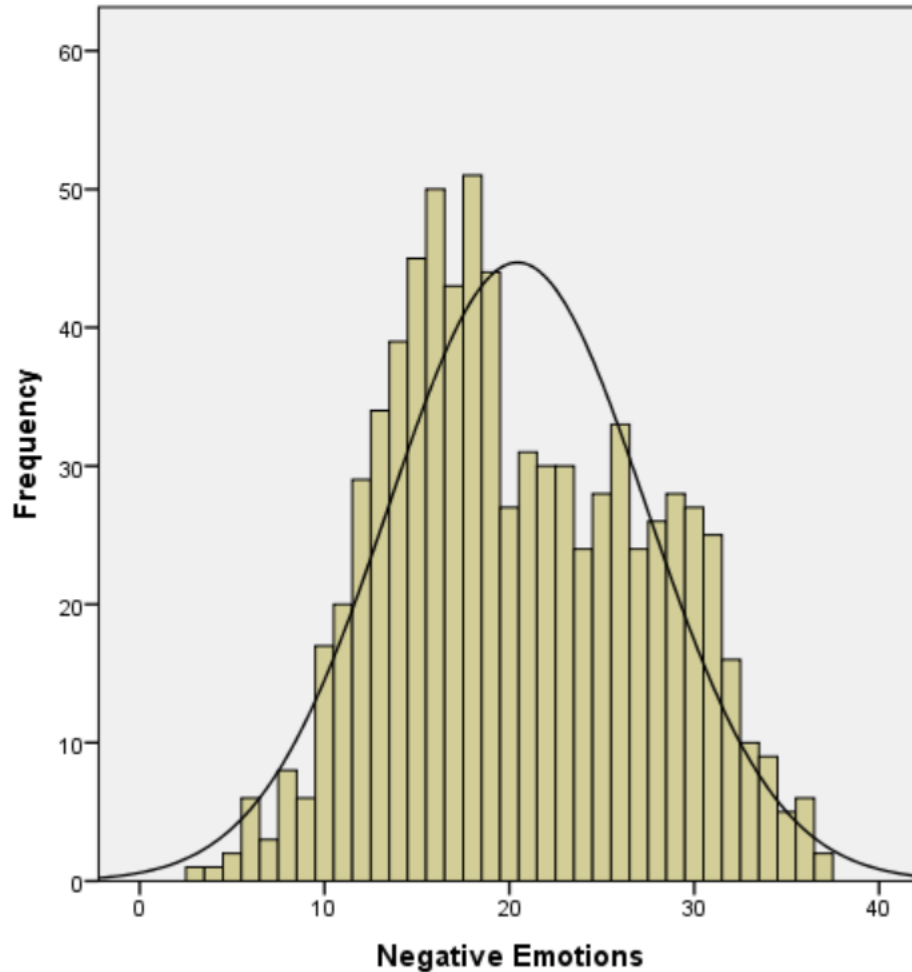
We can see that around 32% of the participants experiences work-related stress.

Assignment 2 – Creating a Histogram

- To create a histogram, navigate to *Graphs > Legacy Dialogs > Histogram*. Then, add the variable NEGEMO to 'variable'. Check the box before 'Display normal curve'. Finally, click on OK.



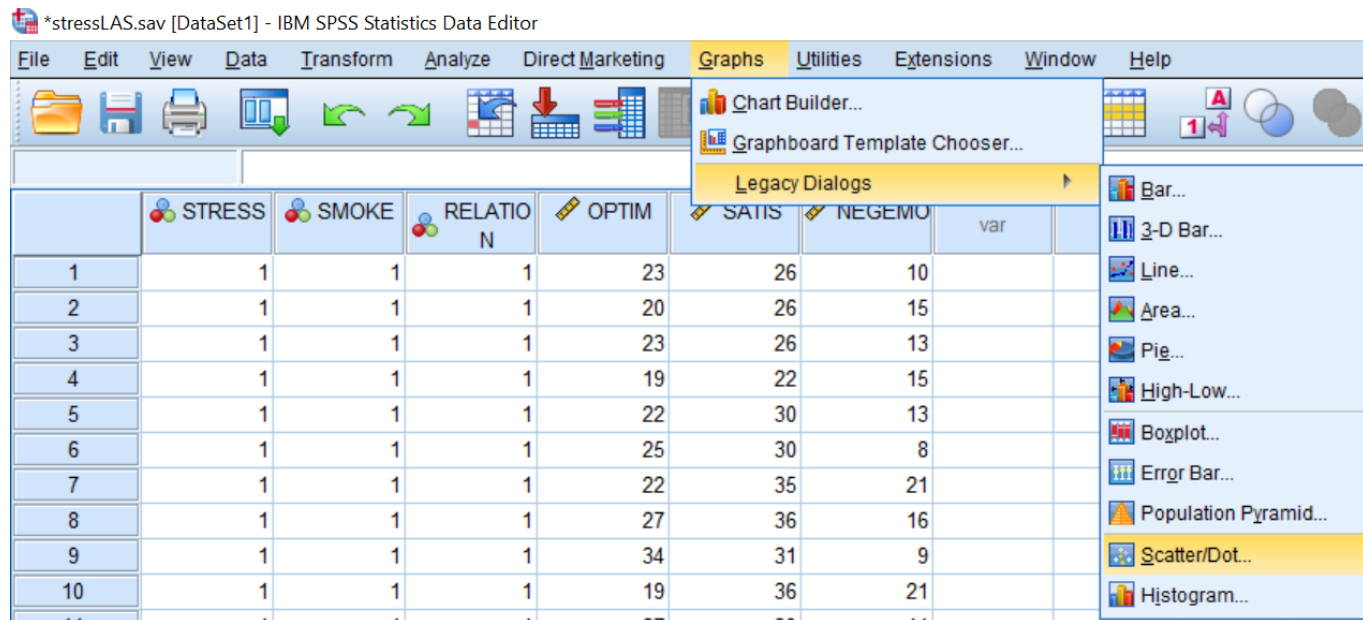
Assignment 2 – Question 3



- The x-axis shows the scores on the negative emotions (here numbers between 0 and 50). The bars represent score ranges; the more respondents with a score in that range, the higher the bar. The y-axis represents the observed number of respondents per score range.
- The distribution looks somewhat normal: there are fewer observations in the tails. However, in a perfect normal distribution we would expect more observations in the middle of the distribution (around 20-25 on negative emotions).

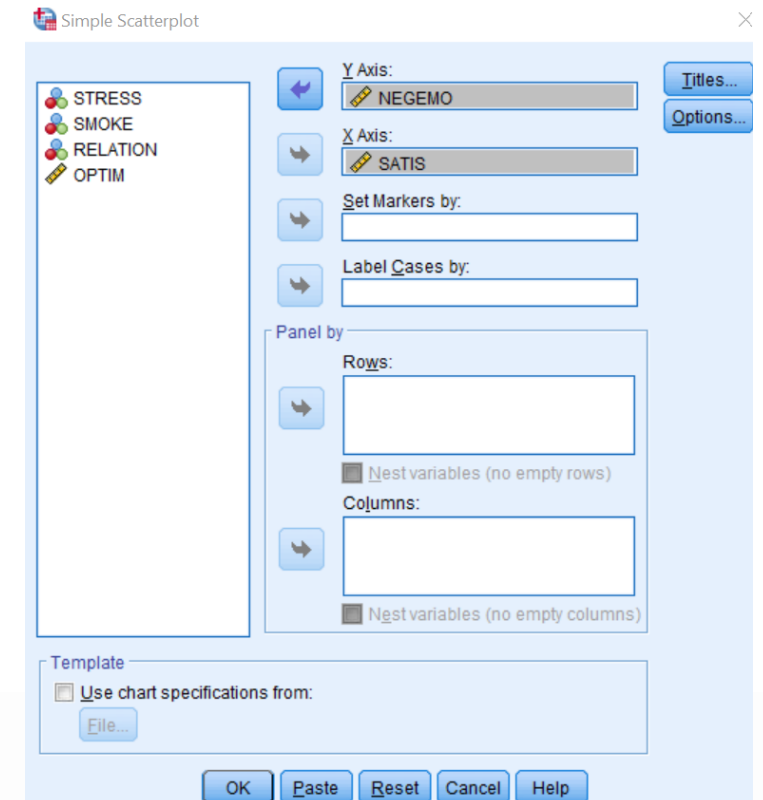
Assignment 2 – Creating a Scatter plot

- To create a scatter plot, navigate to *Graphs > Legacy Dialogs > Scatter/dot*. Choose 'Simple scatter' and click 'Define' (not pictured). Add 'SATIS' to the x-axis and 'NEGEMO' to the y-axis. Finally, click 'OK'.



*stressLAS.sav [DataSet1] - IBM SPSS Statistics Data Editor

	STRESS	SMOKE	RELATION	OPTIM	SATIS	NEGEMO	var
1	1	1	1	23	26	10	
2	1	1	1	20	26	15	
3	1	1	1	23	26	13	
4	1	1	1	19	22	15	
5	1	1	1	22	30	13	
6	1	1	1	25	30	8	
7	1	1	1	22	35	21	
8	1	1	1	27	36	16	
9	1	1	1	34	31	9	
10	1	1	1	19	36	21	



Simple Scatterplot

Y Axis: NEGEMO

X Axis: SATIS

Set Markers by:

Label Cases by:

Panel by

Rows:

Columns:

Template

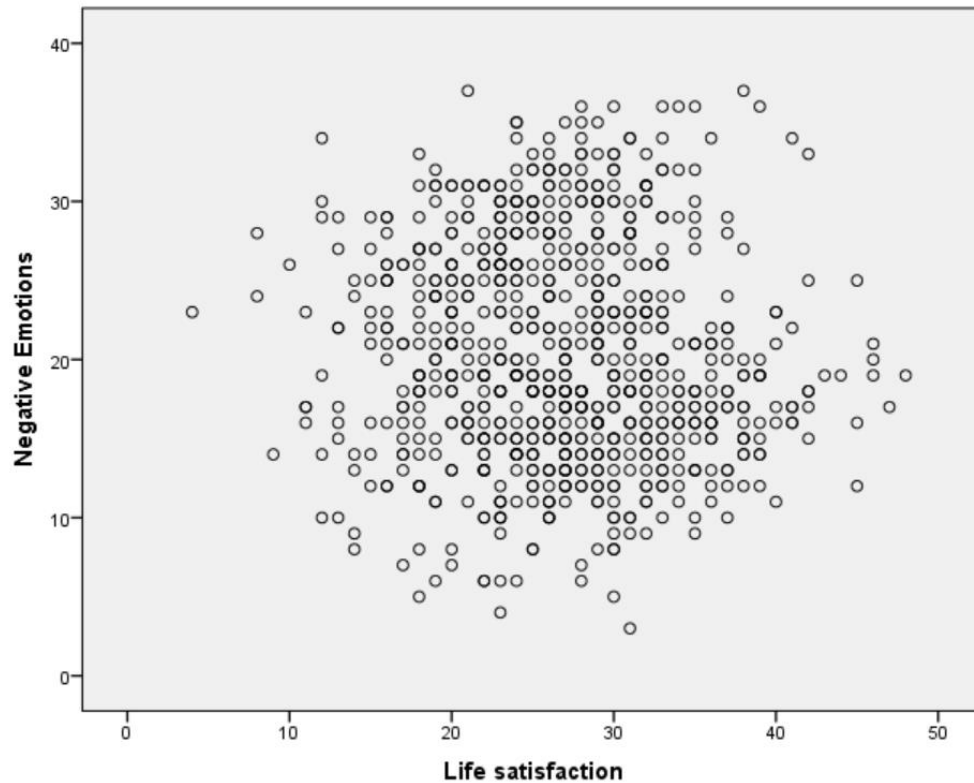
Use chart specifications from:

File...

OK Paste Reset Cancel Help

Assignment 2 – Question 4

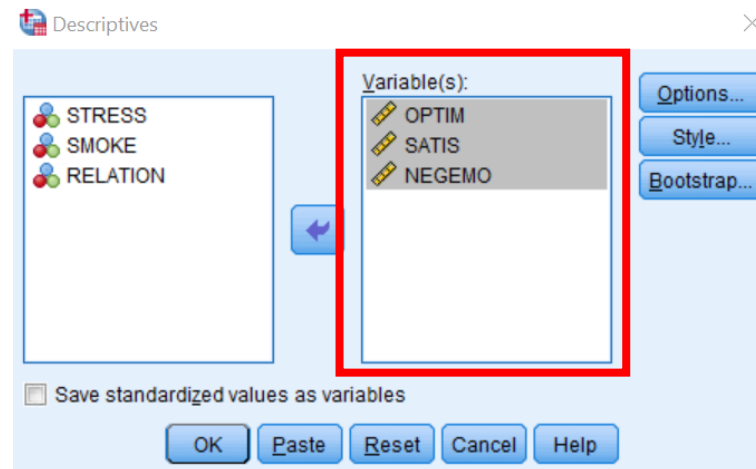
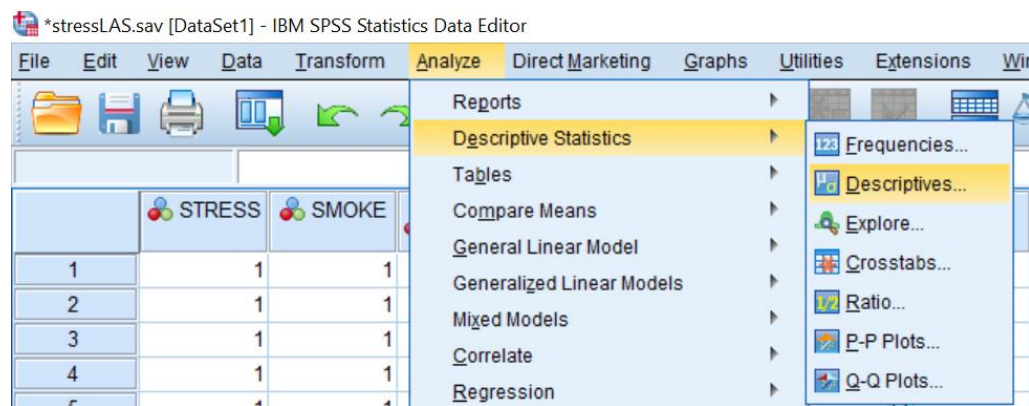
- There is no clear association; the level of Negative Emotions does not depend on Life Satisfaction, and vice versa (which may be a bit of a surprise).



Assignment 3: Exploring Data - Descriptive Statistics

Assignment 3 – Descriptive Statistics + Questions 1-4

- To compute descriptive statistics, navigate to *Analyze > Descriptive Statistics > Descriptives*. Add the relevant variables to the 'Variable(s)' box. Then, click on 'OK'.

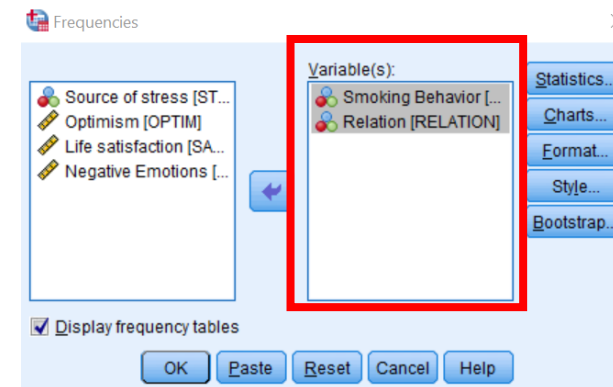
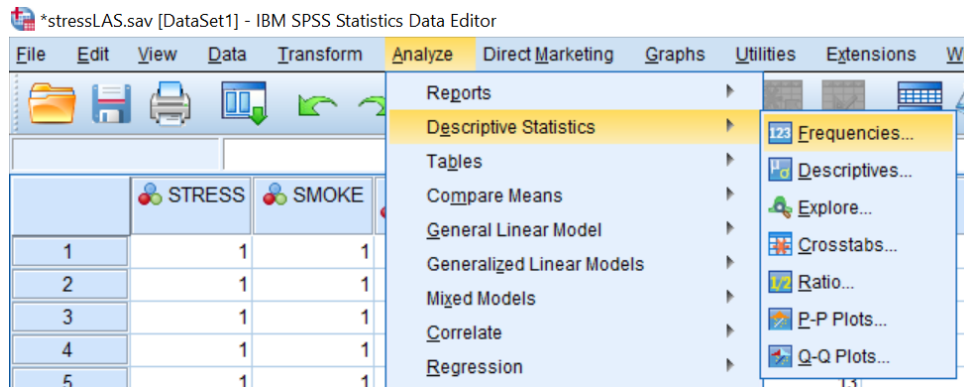


Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Optimism	780	3	220	19.13	8.846
Life satisfaction	780	4	48	26.81	6.989
Negative Emotions	780	3	37	20.43	6.960
Valid N (listwise)	780				

- There are 780 participants in the sample.
- The mean value of Optimism is 19.13.
- For Optimism the spread of the values is the largest.
- The minimum observed score for Negative Emotions is 3 and the maximum is 37.

Assignment 3 – Frequency Statistics + Question 5-6

- To compute frequency statistics, navigate to *Analyze > Descriptive Statistics > Frequencies*. Add the relevant variables to the 'Variable(s)' box. Then, click on 'OK'.



Smoking Behavior

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Non-smoker	375	48.1	48.1	48.1
	Smoker	405	51.9	51.9	100.0
	Total	780	100.0	100.0	

Relation

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Single	406	52.1	52.1	52.1
	Relationship	374	47.9	47.9	100.0
	Total	780	100.0	100.0	

- Non-smokers: 48.1% of participants.
- In a relationship: 47.9% of the participants.

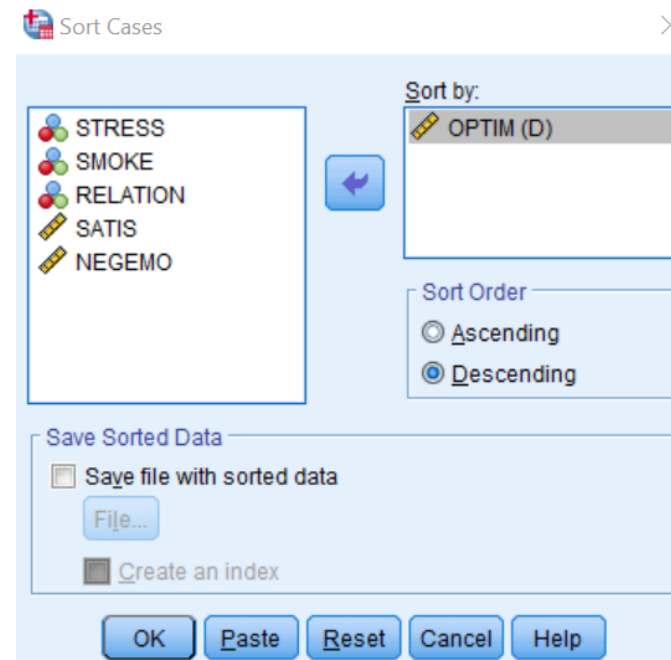
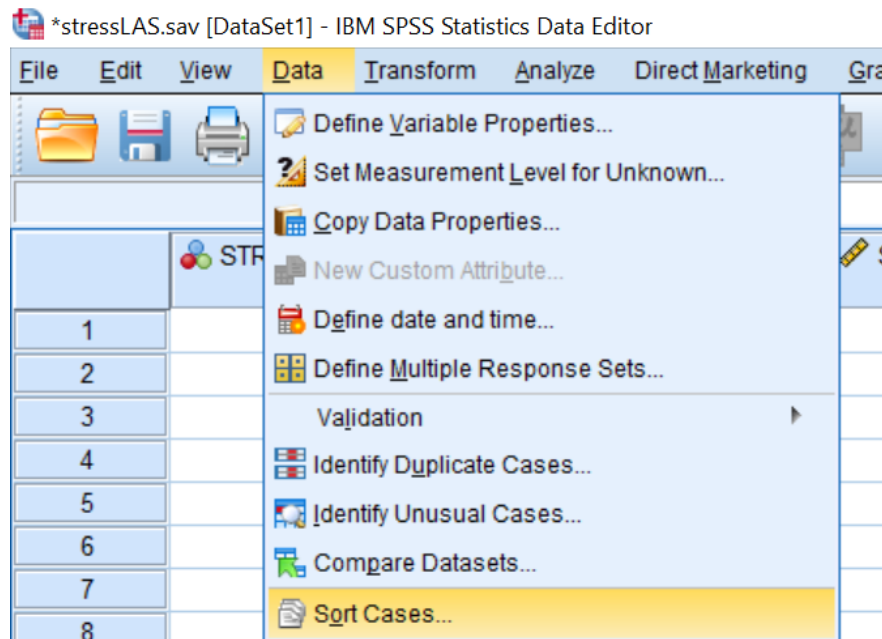
Assignment 3 – Erroneous variables + Question 7

- You can check if there are erroneous variables by first examining the theoretically possible range of values. In step 2 of this assignment we read:
 - source of stress (**stress**): 1 = no stress; 2 = work-related stress; 3 = personal-life stress.
 - smoking behavior (**smoke**) 1 = non-smoker; 2 = smoker;
 - being in a significant relationship (**relation**): 1 = single; 2 = relationship.
 - optimism (**optim**): scale scores between 1 and 50.
 - life satisfaction (**satis**): scale scores between 1 and 50.
 - negative emotions (**negemo**): scale scores between 1 and 50.
- If observed variables fall outside of that possible range, you know you have an erroneous variable.
- In the descriptives, we see that someone scored 220 on Optimism, whereas scores can only fall between 1 and 50.
- Thus, there is at least one erroneous variable on this variable.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Optimism	780	3	220	19.13	8.846
Life satisfaction	780	4	48	26.81	6.989
Negative Emotions	780	3	37	20.43	6.960
Valid N (listwise)	780				

Assignment 3 – Sorting Cases

- To sort cases, navigate to Data > Sort cases. Then, choose the variable you want to sort on. To find the erroneous cases, it is also easiest to sort on descending order – so the highest value are shown on top. Then, click on OK. In the data view, the cases are now sorted.



Assignment 3 – Question 8-9

- As we have just deleted a high value on Optimism, we can expect the mean and the standard deviation both to decrease. The average score will be lower and so will the deviations from the mean be (as that very high value was also very deviant from the mean).

- Before deletion:

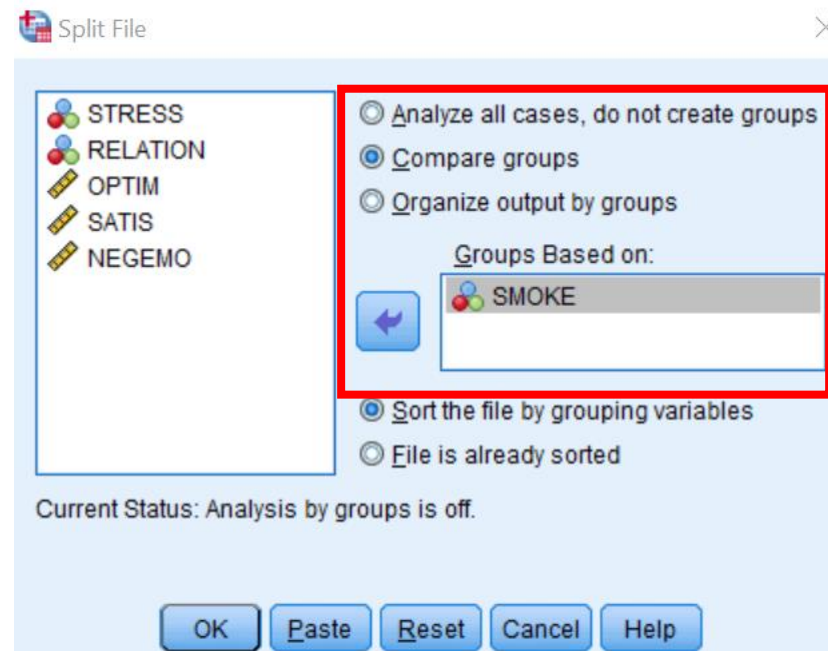
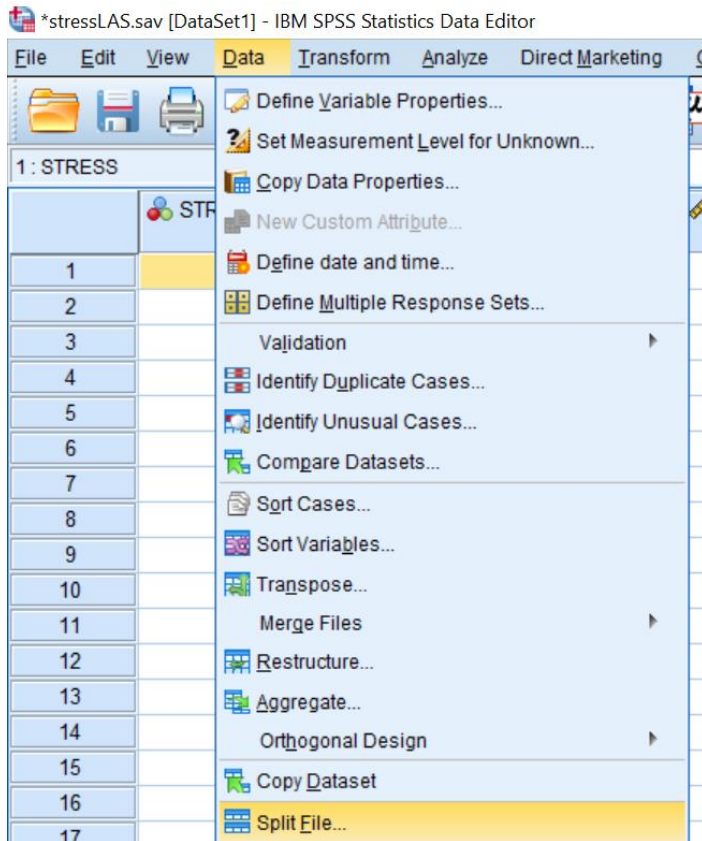
Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Optimism	780	3	220	19.13	8.846
Life satisfaction	780	4	48	26.81	6.989
Negative Emotions	780	3	37	20.43	6.960
Valid N (listwise)	780				

- After deletion:

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Optimism	779	3	34	18.87	5.139
Life satisfaction	779	4	48	26.81	6.990
Negative Emotions	779	3	37	20.44	6.963
Valid N (listwise)	779				

Assignment 3 – Splitting a File

- To split the file, navigate to Data > Split File > Compare groups. Choose the relevant variable and then click 'OK'.



Assignment 3 – Question 10-11

- On average, non-smokers score **26.71** on Life Satisfaction and smokers score **26.90**.

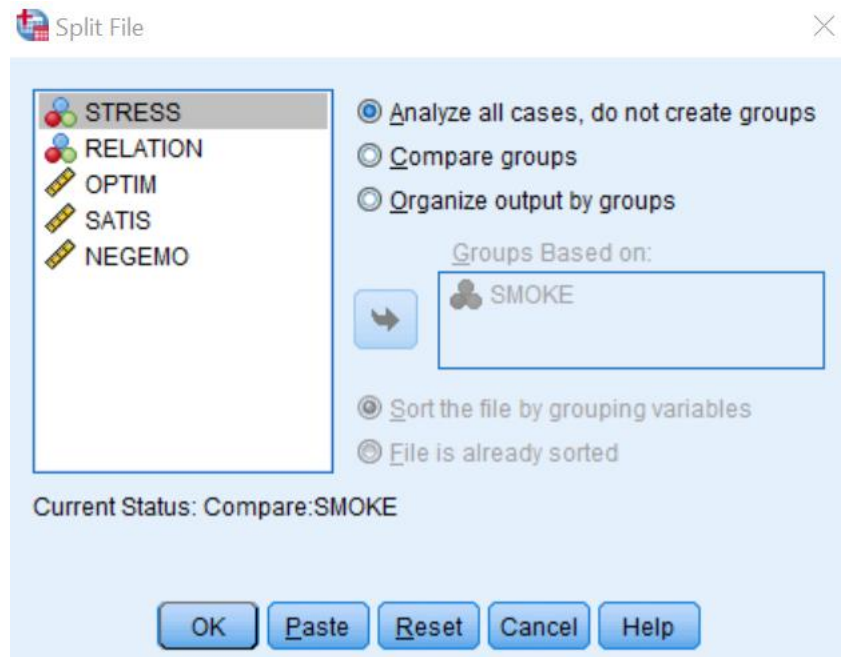
Descriptive Statistics						
Smoking Behavior		N	Minimum	Maximum	Mean	Std. Deviation
Non-smoker	Life satisfaction	374	10	48	26.71	6.925
	Valid N (listwise)	374				
Smoker	Life satisfaction	405	4	47	26.90	7.058
	Valid N (listwise)	405				

- We cannot draw conclusions about whether smokers are also more satisfied with their lives in the population. To answer that question, we will need to do more than only looking at the means in our sample. In a following practical you will learn how to answer this research question.

Assignment 4: Missing Values

Assignment 4 – Turning Off Split File

- Turning off split file:



Assignment 4 – Question 1 + 2 & Deleting Cases

- In the sample, 33.6% of participants experience no stress.
- In the sample, personal-life related stress is most common.

		Source of stress			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No Stress	262	33.6	33.6	33.6
	Work-related stress	249	32.0	32.0	65.6
	Personal-life related stress	268	34.4	34.4	100.0
	Total	779	100.0	100.0	

- After deleting the first 10 cases, your data view should look like this:

	STRESS	SMOKE	RELATIO N	OPTIM	SATIS	NEGEMO	var
1	.	1	1	34	31	9	
2	.	1	1	34	39	12	
3	.	1	1	33	42	15	
4	.	1	1	30	39	15	
5	.	1	1	30	38	17	
6	.	1	1	30	32	11	
7	.	1	2	30	33	12	
8	.	1	1	29	22	6	
9	.	1	1	29	34	13	
10	.	1	2	29	45	12	
11	1	1	2	29	36	15	
12	1	1	1	28	41	16	

Assignment 4 – Question 3 + 4

- Before deleting:

		Source of stress			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No Stress	262	33.6	33.6	33.6
	Work-related stress	249	32.0	32.0	65.6
	Personal-life related stress	268	34.4	34.4	100.0
	Total	779	100.0	100.0	

- After deleting:

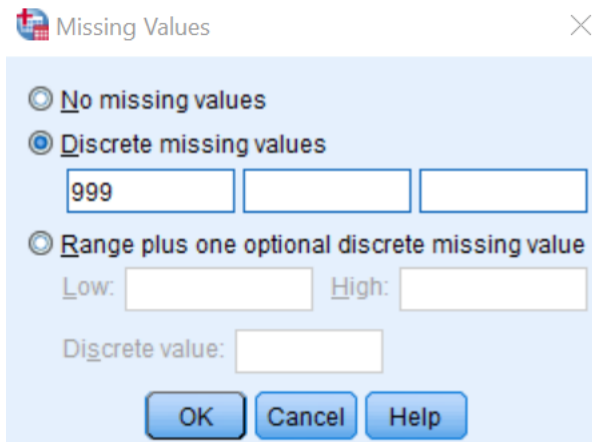
		Source of stress			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No Stress	252	32.3	32.8	32.8
	Work-related stress	249	32.0	32.4	65.1
	Personal-life related stress	268	34.4	34.9	100.0
	Total	769	98.7	100.0	
Missing	System	10	1.3		
Total		779	100.0		

- We can see that the values of Percent and Valid Percent have changed and that a 'missing' row has been added to the table. It makes sense that the percentages have changed, as there are now missing values.
- To calculate Percent by hand: $252/779 \times 100 = 32.3\%$
- To calculate Valid Percent by hand: $252/769 \times 100 = 32.8\%$
- So, Percent we divide by the total number of participants, including missings. For Valid Percent we divide by the total number of participants, excluding missings.

Assignment 4 – Indicating the Missing Values

- Indicating the missing values with '999':

We fill in '999' for the first 10 participants in the data set by hand. Then, go to variable view and click on Missing in the row of Stress. Click on the blue box with the three dots that appears. Then, in the menu that opens, indicate that 999 is a discrete missing value. Finally, click on 'OK'.



The image shows the 'Missing Values' dialog box in SPSS. It has a title bar with a red cross icon and the text 'Missing Values'. The dialog contains three radio button options: 'No missing values', 'Discrete missing values', and 'Range plus one optional discrete missing value'. The 'Discrete missing values' option is selected. Below this option, there are three input boxes; the first one contains the text '999'. Below these boxes are the 'Low:' and 'High:' labels followed by empty input boxes. Further down is the 'Discrete value:' label followed by an empty input box. At the bottom of the dialog are three buttons: 'OK', 'Cancel', and 'Help'.

Assignment 4 – Question 5

- When we calculate the frequencies of Stress again, the table looks as follows:

		Source of stress			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No Stress	252	32.3	32.8	32.8
	Work-related stress	249	32.0	32.4	65.1
	Personal-life related stress	268	34.4	34.9	100.0
	Total	769	98.7	100.0	
Missing	999	10	1.3		
Total		779	100.0		

- Now, the table shows us what the missing value is: 999.

Assignment 5: More on Descriptive Statistics

Assignment 5 – Question 1: Mean, Median, Mode

- Consider the following grades for 10 students: 6, 3, 4, 6, 7, 6, 8, 9, 10, 9.
Compute the mean, median, and mode.

- Mean = $\frac{\text{Sum of the observations}}{\text{number of observations}} = \frac{6+3+4+6+7+6+8+9+10+9}{10}$

Mean = 6.8

- Median = the middle number in a sorted list of numbers

Median = 3, 4, 6, 6, 6, 7, 8, 9, 9, 10

Here, the median falls between 6 and 7 → the median is 6.5

- Mode = the most common observation

Here, the mode is 6

Assignment 5 – Question 2: Variance & Standard Deviation

- Consider the grades of 6 students: 2, 7, 6, 7, 8, 9.
Compute by hand the variance and standard deviations.

- Variance =
$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

\bar{x} is the mean. $\bar{x} = (2 + 7 + 6 + 7 + 8 + 9)/6 = 6.5$

- Variance =
$$\frac{20.25 + 0.25 + 0.25 + 0.25 + 2.25 + 6.25}{5}$$

Variance = 5.9

$X_1 = 2$
In formula: $(2 - 6,5) = -4,5$
 $(-4,5)^2 = 20.25$

- Standard deviation = $\sqrt{\text{variance}} = \sqrt{5.9}$
Standard deviation = 2.429

Assignment 5 – Creating the Data Set

- Filling in the data in SPSS

Practical1_Grades.sav [DataSet1]

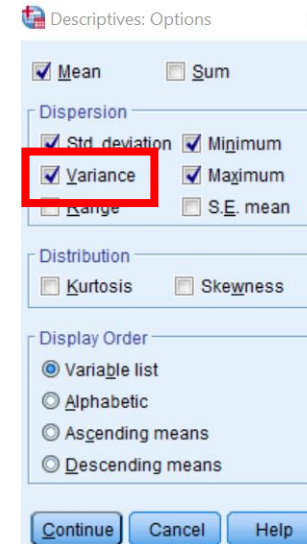
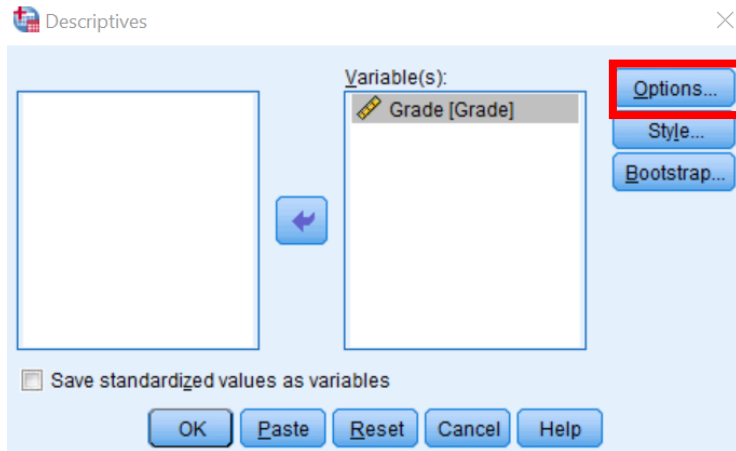
10 :	
	Grade
1	2
2	7
3	6
4	7
5	8
6	9
7	

Practical1_Grades.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Extensions Window Help											
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Grade	Numeric	8	0	Grade	None	None	8	Right	Scale	Input
2											

Assignment 5 – Checking Descriptives in SPSS

- Computing the variance and standard deviation

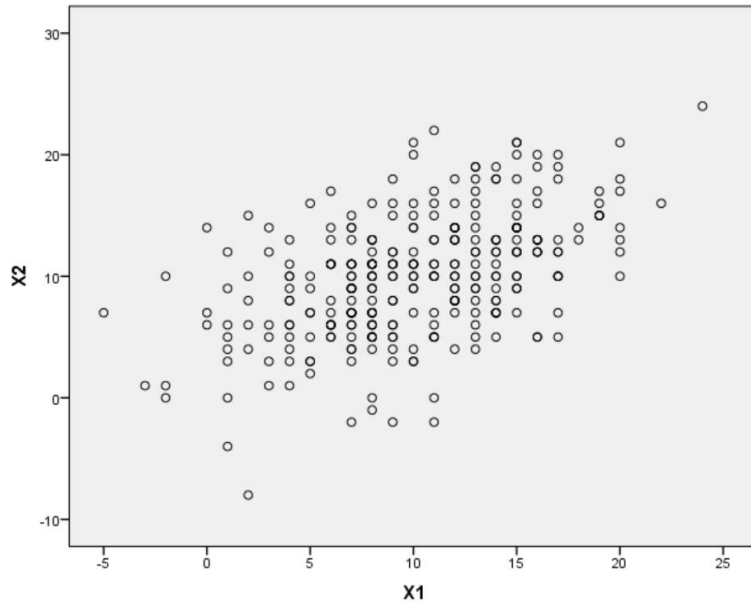


- The variance is indeed 5.9 and the standard deviation is indeed 2.429.

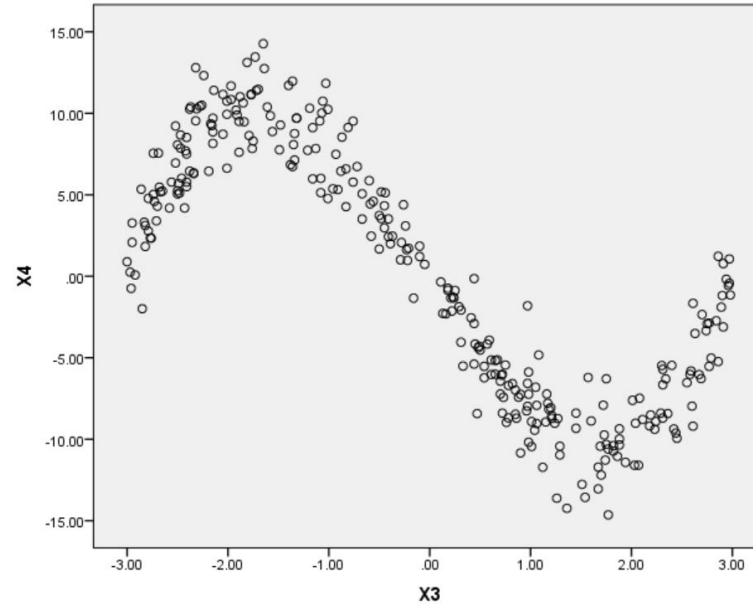
Descriptive Statistics						
	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Grade	6	2	9	6.50	2.429	5.900
Valid N (listwise)	6					

Assignment 5 – Scatter plots & Question 4 + 5

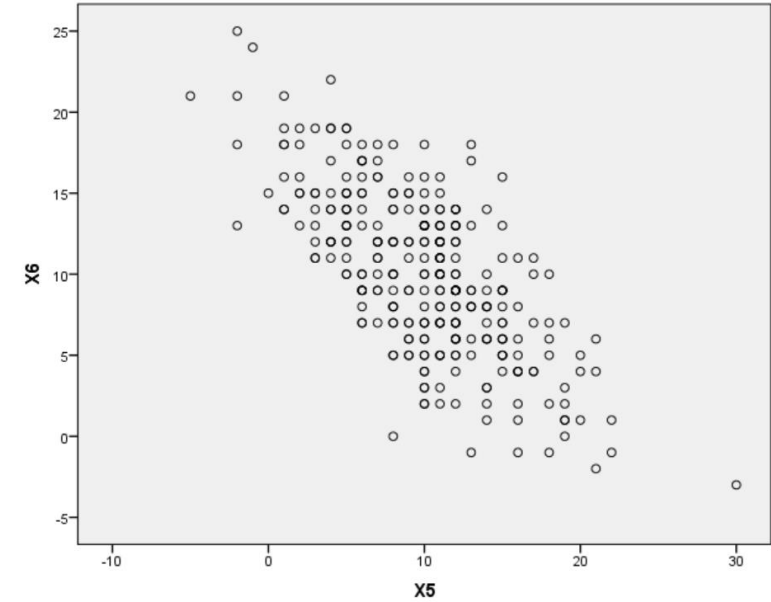
- For more information on how to create a scatterplot, see slide 14.



- A positive, linear relationship
Moderate in strength



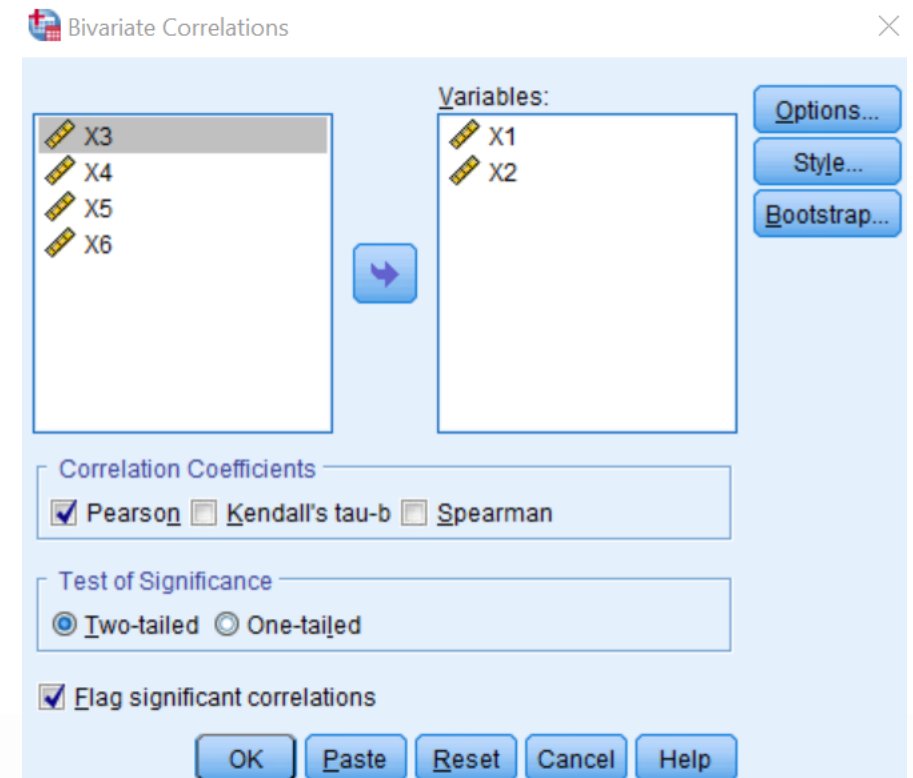
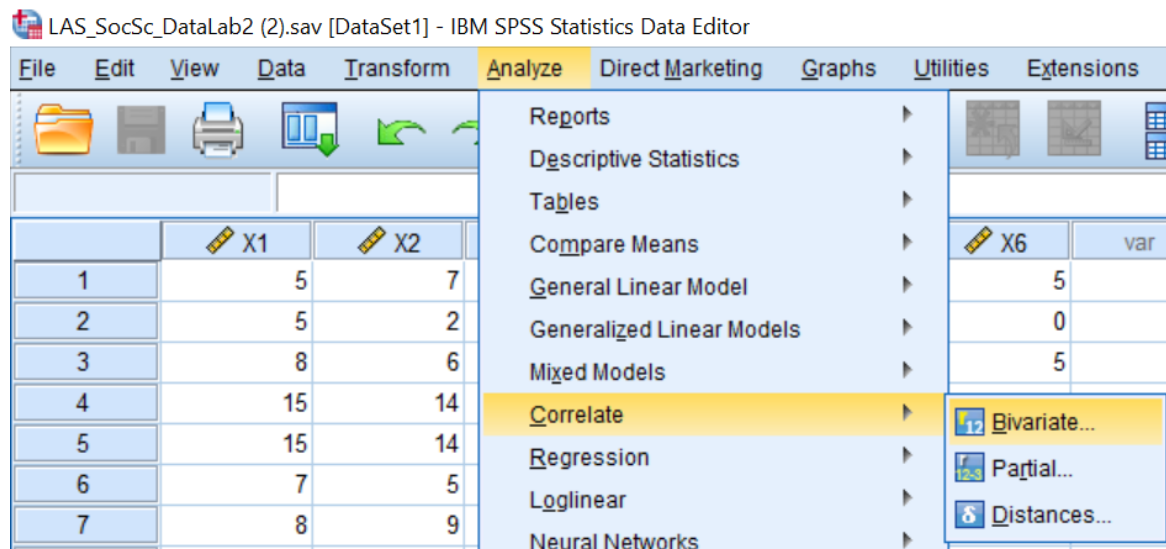
- A non-linear relationship
Strong in strength
- Example: time in the day and how far the water reaches up the beach (ebb and flow)



- A negative, linear relationship
Moderate to strong in strength

Assignment 5 – Calculating a Correlation Coefficient

- To calculate a correlation coefficient, navigate to *Analyze > Correlate > Bivariate*. Select the relevant variables and choose 'OK'.



Assignment 5 – Question 6 + 7 + 8

- The correlation between X1 and X2 is **.504**.
- The correlation between X2 and X6 is **.058**.
- The correlation between X3 and X4 is **-.798**. Even though the relationship between X3 and X4 is not linear, SPSS will still calculate a pearson correlation. This means that it is important you first check the scatterplot between two variables, before calculating the correlation coefficient, to check that the relationship is in fact linear.

Correlations

		X1	X2
X1	Pearson Correlation	1	.504**
	Sig. (2-tailed)		.000
	N	300	300
X2	Pearson Correlation	.504**	1
	Sig. (2-tailed)	.000	
	N	300	300

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

Correlations

		X2	X6
X2	Pearson Correlation	1	.058
	Sig. (2-tailed)		.315
	N	300	300
X6	Pearson Correlation	.058	1
	Sig. (2-tailed)	.315	
	N	300	300

Correlations

		X3	X4
X3	Pearson Correlation	1	-.798**
	Sig. (2-tailed)		.000
	N	300	300
X4	Pearson Correlation	-.798**	1
	Sig. (2-tailed)	.000	
	N	300	300

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

Assignment 5 – Question 9

- The correlation coefficient between X5 and X6 is **-.698**. Based on this coefficient, we can conclude that the relationship between X5 and X6 is negative and that the relationship is moderate to strong in strength.

Correlations			
		X5	X6
X5	Pearson Correlation	1	-.698**
	Sig. (2-tailed)		.000
	N	300	300
X6	Pearson Correlation	-.698**	1
	Sig. (2-tailed)	.000	
	N	300	300

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