INTRODUCTION TO SPSS

Anne Schad Bergsaker

20.-22. September 2022

Before we begin...

Learning goals

1. Become familiar with and know how to navigate between the different windows in SPSS
2. Know how to write a data file
3. Know how to import data
4. Identify errors and strange or even impossible measurements in a data set
5. Describe data using descriptive statistics that you calculate using SPSS (and know what these statistical measures can tell you)
6. Present data using graphs made in SPSS
7. Manipulate data (re-structure, re-code, calculate new variables from original variables)
8. Know how to easily and efficiently document everything you do in SPSS, and re-use your own methods

Technical prerequisites

If you have not got SPSS installed on your own device, use remote desktop or go to view.uio.no and log on with your UiO username and password.

The data files used for examples are either from the SPSS survival manual, or are fictitious data that I have made. These files can be downloaded as a single .zip file from the course website.

Try to do what I do, and follow the same steps. If you missed a step or have a question, don't hesitate to ask.

GETTING STARTED

What is SPSS?

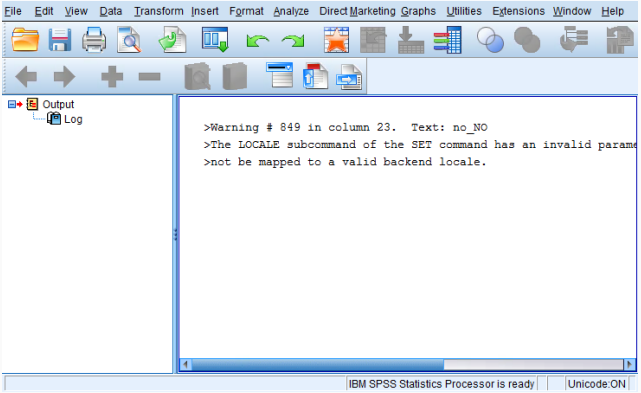
SPSS is short for for *Statistical Package for the Social Sciences*

The different windows

1. Output
2. Syntax
3. Data view
4. Variabel view

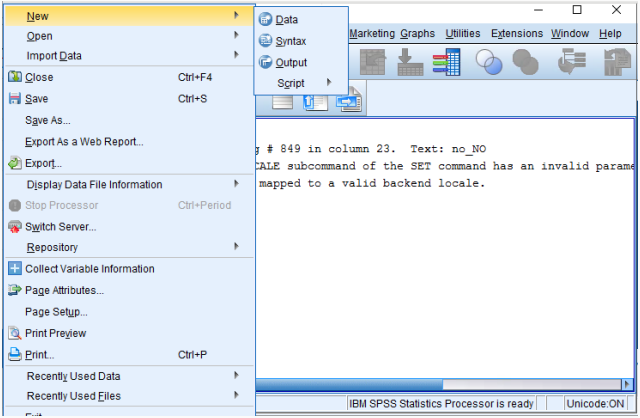
OUTPUT

Shows all commands that have been executed, and the corresponding results and outputs of those commands. This is where you will find the tables and graphs you produce.



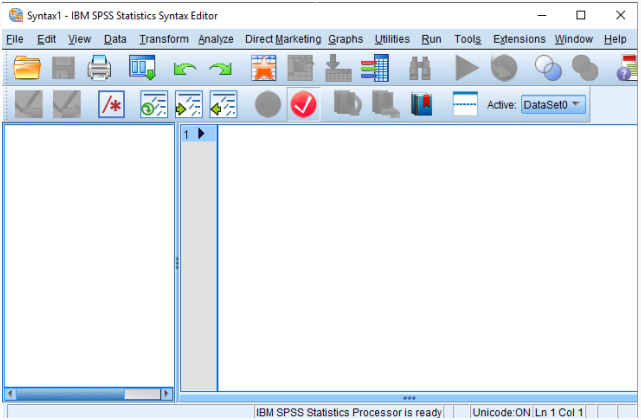
SYNTAX

The syntax file is the easiest way to keep track of everything you do, and it enables you to go back and either re-examine your analysis, or run the same analysis again.



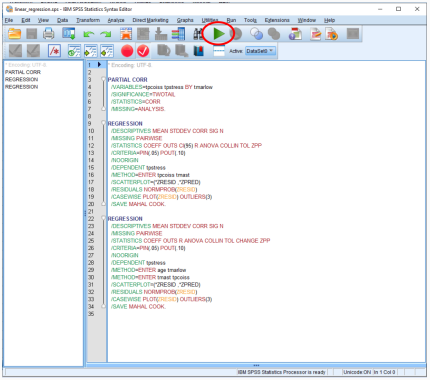
SYNTAX

SPSS allows for automatic pasting of every single procedure you run, or alternatively, you can type in procedures yourself.



SYNTAX

It also makes it easier to run the same procedure on different data sets and to run the same procedure multiple times.



You run syntax by clicking the green arrow button. If you only want to run part of the syntax, highlight that part and click the green arrow.

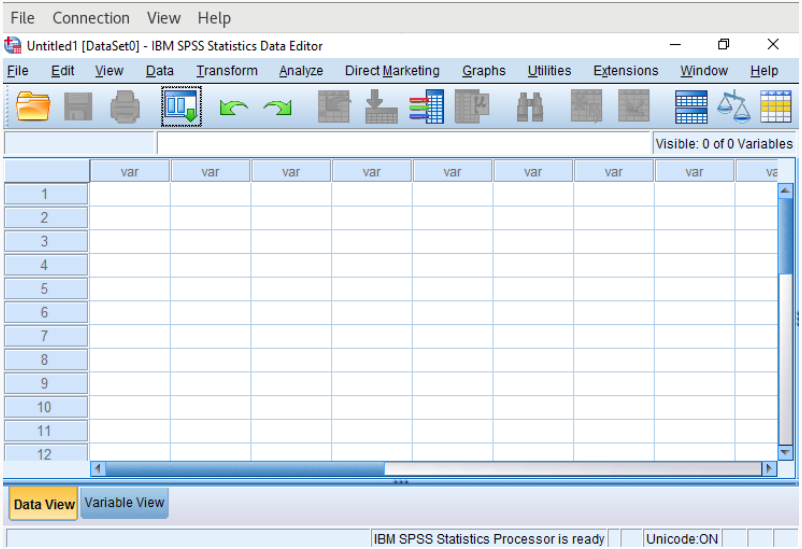
It can be saved as a .sps file, and can be opened by any simple text editor, like Notepad.

Syntax

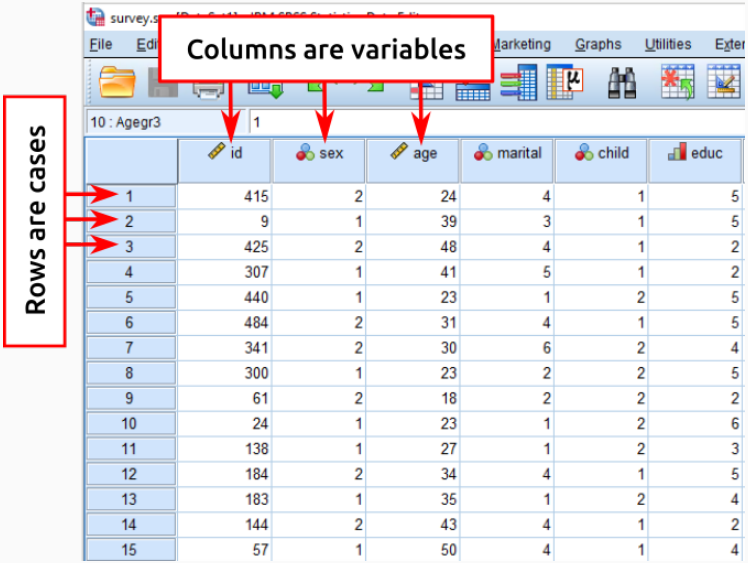
Colour coding

- Dark blue - Procedures
- Green - Sub-routines, connected with a procedure
- Red/Orange - Key words indicating choices you have made
- Grey - Comments (the line must begin with *)
- Black - Variable names

DATA VIEW

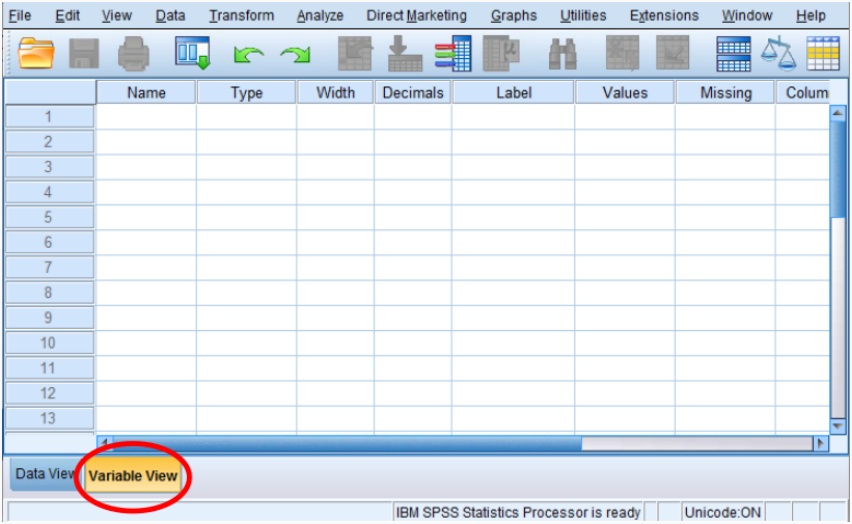

File Connection View Help
Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Extensions Window Help
Visible: 0 of 0 Variables
var
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

DATA VIEW

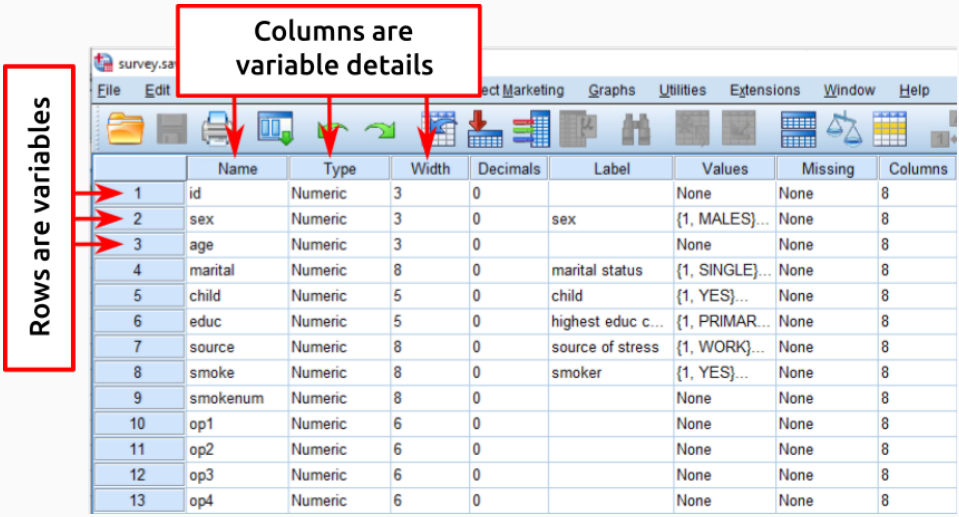


	id	sex	age	marital	child	educ
1	415	2	24	4	1	5
2	9	1	39	3	1	5
3	425	2	48	4	1	2
4	307	1	41	5	1	2
5	440	1	23	1	2	5
6	484	2	31	4	1	5
7	341	2	30	6	2	4
8	300	1	23	2	2	5
9	61	2	18	2	2	2
10	24	1	23	1	2	6
11	138	1	27	1	2	3
12	184	2	34	4	1	5
13	183	1	35	1	2	4
14	144	2	43	4	1	2
15	57	1	50	4	1	4

VARIABLE VIEW



VARIABLE VIEW



	Name	Type	Width	Decimals	Label	Values	Missing	Columns
1	id	Numeric	3	0		None	None	8
2	sex	Numeric	3	0	sex	{1, MALES}...	None	8
3	age	Numeric	3	0		None	None	8
4	marital	Numeric	8	0	marital status	{1, SINGLE}...	None	8
5	child	Numeric	5	0	child	{1, YES}...	None	8
6	educ	Numeric	5	0	highest educ c...	{1, PRIMAR...	None	8
7	source	Numeric	8	0	source of stress	{1, WORK}...	None	8
8	smoke	Numeric	8	0	smoker	{1, YES}...	None	8
9	smokenum	Numeric	8	0		None	None	8
10	op1	Numeric	6	0		None	None	8
11	op2	Numeric	6	0		None	None	8
12	op3	Numeric	6	0		None	None	8
13	op4	Numeric	6	0		None	None	8

The different types of variables in SPSS

- Nominal - This is a variable coded as a set of numbers, but where the numbers in themselves mean something else, e.g. gender (male: 1, female: 2), answer (yes: 1, no: 0).

The different types of variables in SPSS

- Nominal - This is a variable coded as a set of numbers, but where the numbers in themselves mean something else, e.g. gender (male: 1, female: 2), answer (yes: 1, no: 0).
- Ordinal - A variable where data is coded as numbers, like with nominal, but where the codes can be arranged in a meaningful order, e.g. How much do you agree with this statement, what is your highest achieved level of education, etc.

The different types of variables in SPSS

- Nominal - This is a variable coded as a set of numbers, but where the numbers in themselves mean something else, e.g. gender (male: 1, female: 2), answer (yes: 1, no: 0).
- Ordinal - A variable where data is coded as numbers, like with nominal, but where the codes can be arranged in a meaningful order, e.g. How much do you agree with this statement, what is your highest achieved level of education, etc.
- Scale - Numbers, where the number itself is the data, e.g. age, blood pressure, temperature, number of points achieved in a test, etc.

All the other values that define a variable

- Name - variable name (has to start with a letter (a-z))

All the other values that define a variable

- Name - variable name (has to start with a letter (a-z))
- Type - numeric, string, date, i.e. number, words, dates, etc.

All the other values that define a variable

- Name - variable name (has to start with a letter (a-z))
- Type - numeric, string, date, i.e. number, words, dates, etc.
- Width - how many characters will this variable need?

All the other values that define a variable

- Name - variable name (has to start with a letter (a-z))
- Type - numeric, string, date, i.e. number, words, dates, etc.
- Width - how many characters will this variable need?
- Decimals - how many decimals do you need?

All the other values that define a variable

- Name - variable name (has to start with a letter (a-z))
- Type - numeric, string, date, i.e. number, words, dates, etc.
- Width - how many characters will this variable need?
- Decimals - how many decimals do you need?
- Label - "longer" description of the variable

All the other values that define a variable

- Name - variable name (has to start with a letter (a-z))
- Type - numeric, string, date, i.e. number, words, dates, etc.
- Width - how many characters will this variable need?
- Decimals - how many decimals do you need?
- Label - "longer" description of the variable
- Values - for nominal and ordinal: this is where you code your variable

All the other values that define a variable

- Name - variable name (has to start with a letter (a-z))
- Type - numeric, string, date, i.e. number, words, dates, etc.
- Width - how many characters will this variable need?
- Decimals - how many decimals do you need?
- Label - "longer" description of the variable
- Values - for nominal and ordinal: this is where you code your variable
- Missing - if you want to give missing data a different code than the default (empty cell)

All the other values that define a variable

- Name - variable name (has to start with a letter (a-z))
- Type - numeric, string, date, i.e. number, words, dates, etc.
- Width - how many characters will this variable need?
- Decimals - how many decimals do you need?
- Label - "longer" description of the variable
- Values - for nominal and ordinal: this is where you code your variable
- Missing - if you want to give missing data a different code than the default (empty cell)
- Columns - how wide do you want the column to be

All the other values that define a variable

- Name - variable name (has to start with a letter (a-z))
- Type - numeric, string, date, i.e. number, words, dates, etc.
- Width - how many characters will this variable need?
- Decimals - how many decimals do you need?
- Label - "longer" description of the variable
- Values - for nominal and ordinal: this is where you code your variable
- Missing - if you want to give missing data a different code than the default (empty cell)
- Columns - how wide do you want the column to be
- Align - right or left alignment within the column

All the other values that define a variable

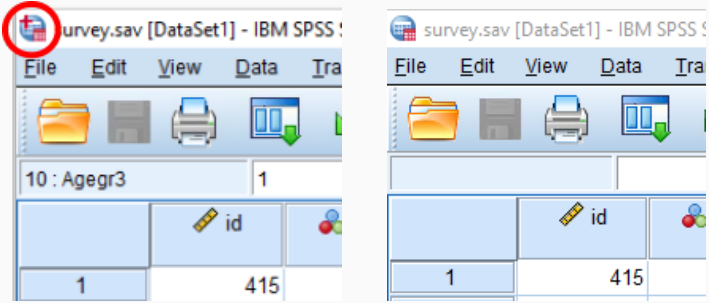
- Name - variable name (has to start with a letter (a-z))
- Type - numeric, string, date, i.e. number, words, dates, etc.
- Width - how many characters will this variable need?
- Decimals - how many decimals do you need?
- Label - "longer" description of the variable
- Values - for nominal and ordinal: this is where you code your variable
- Missing - if you want to give missing data a different code than the default (empty cell)
- Columns - how wide do you want the column to be
- Align - right or left alignment within the column
- Measure - this is where you choose scale, nominal or ordinal

All the other values that define a variable

- Name - variable name (has to start with a letter (a-z))
- Type - numeric, string, date, i.e. number, words, dates, etc.
- Width - how many characters will this variable need?
- Decimals - how many decimals do you need?
- Label - "longer" description of the variable
- Values - for nominal and ordinal: this is where you code your variable
- Missing - if you want to give missing data a different code than the default (empty cell)
- Columns - how wide do you want the column to be
- Align - right or left alignment within the column
- Measure - this is where you choose scale, nominal or ordinal
- Role - Give a predefined role from the start (if you want to), e.g. dependent or independent variable (not necessary)

Active window

Especially when you have several files/data sets open at the same time, it is important to know which spreadsheet is the active one, as this is the one that will be used in the analysis you do. SPSS indicates which is the *active* window with a small cross in the top left corner.



SHORTCUTS

Icon	Descrition
	Open file
	Save file
	Print
	Most recent procedures
	Undo
	Redo
	Go to case

Icon	Description
	Go to variable
	Variable information
	Descriptive statistics
	Find/search
	Insert case
	Insert variable
	Split file

Icon	Description
	Weight cases
	Select cases
	Choose if data or label is shown
	Use subset
	Show all variables
	Spell check

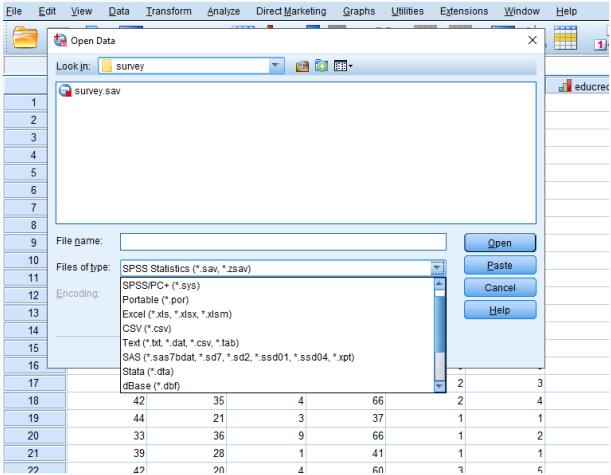
DATA MANAGEMENT IN SPSS

Open an existing data file

File > Open > Data

In SPSS, data files have the file extension .sav

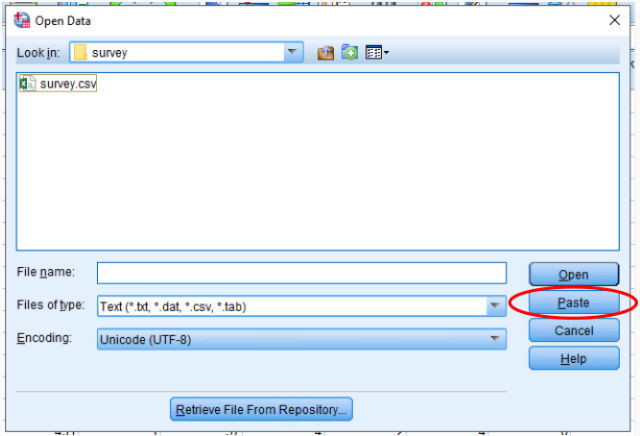
Choose *Paste*, not Open. This will paste the corresponding commands into the syntax file. Go to the syntax window, mark the text, and click on the *Run selection* button (green triangle).



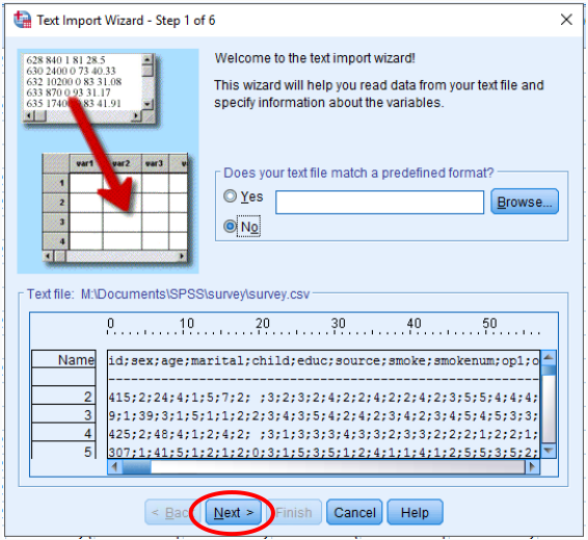
IMPORTING DATA - TEXT FILE

File > Import Data > Text Data

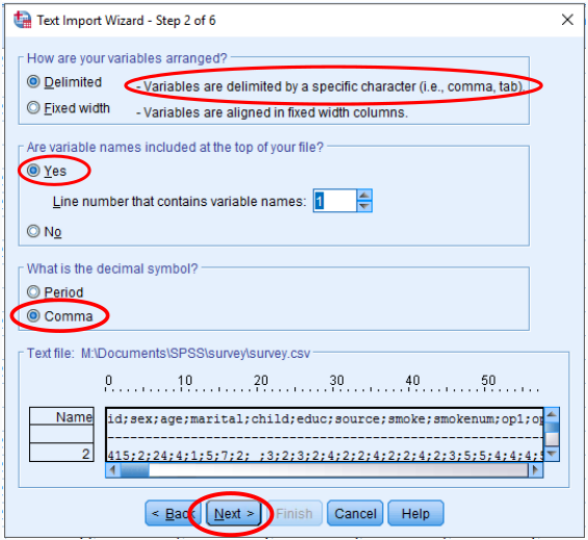
This dialogue box allows you to open .txt, .dat, .csv, and .tab files.



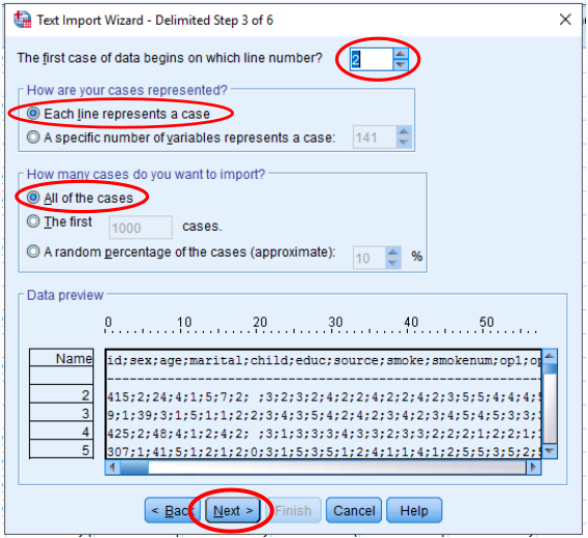
IMPORTING DATA - TEXT IMPORT WIZARD STEP 1 OF 6



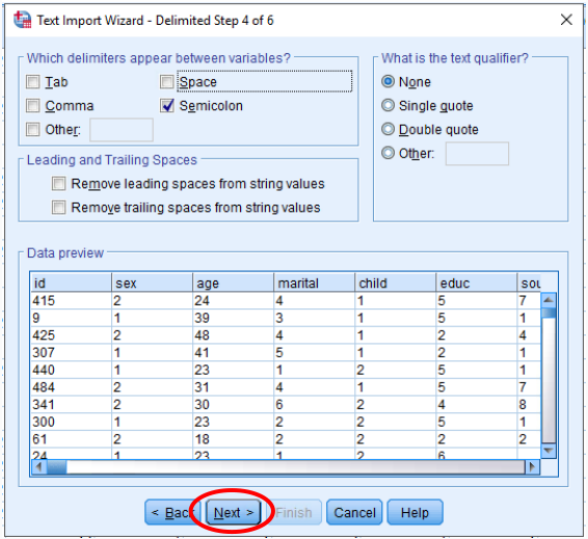
IMPORTING DATA - TEXT IMPORT WIZARD STEP 2 OF 6



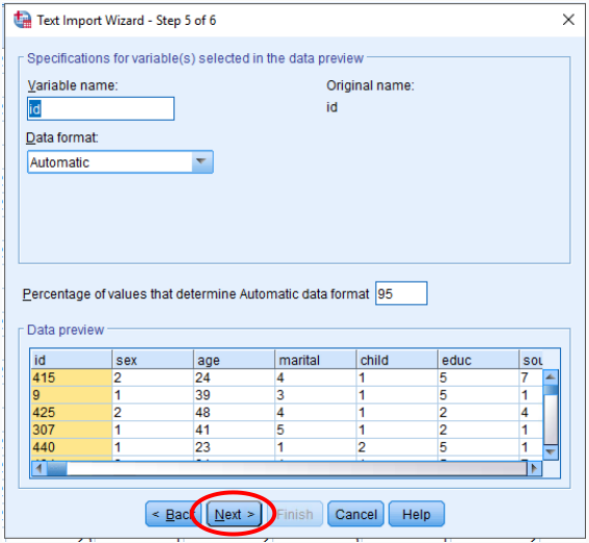
IMPORTING DATA - TEXT IMPORT WIZARD STEP 3 OF 6



IMPORTING DATA - TEXT IMPORT WIZARD STEP 4 OF 6

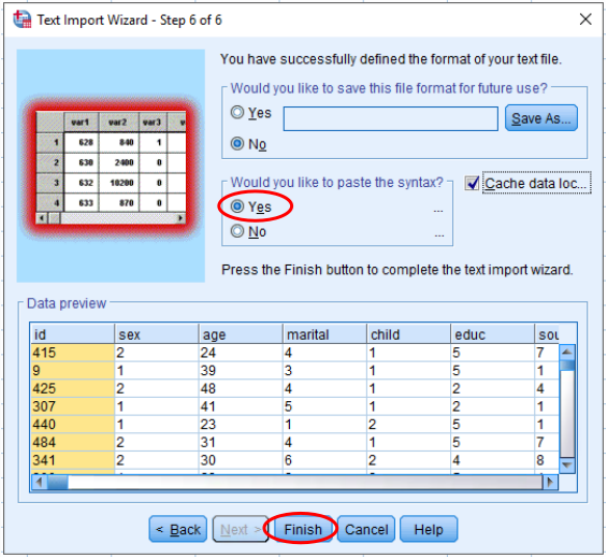


IMPORTING DATA - TEXT IMPORT WIZARD STEP 5 OF 6



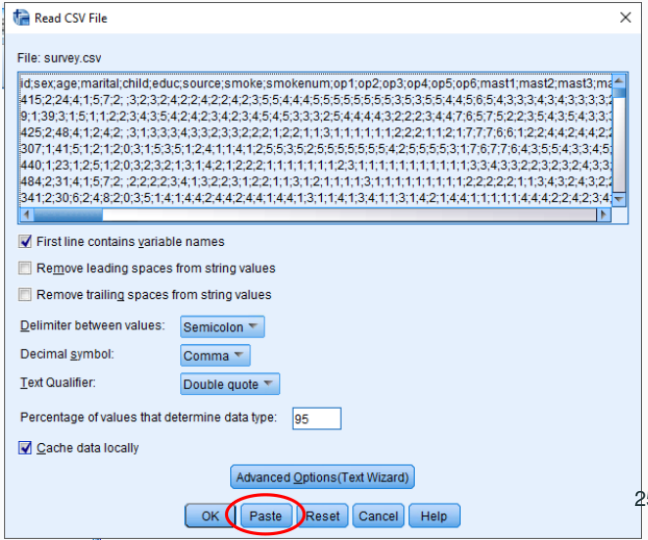
IMPORTING DATA - TEXT IMPORT WIZARD STEP 6 OF 6

Remember to go to your syntax file and run the latest bit of code added to it. This will actually import the data file.



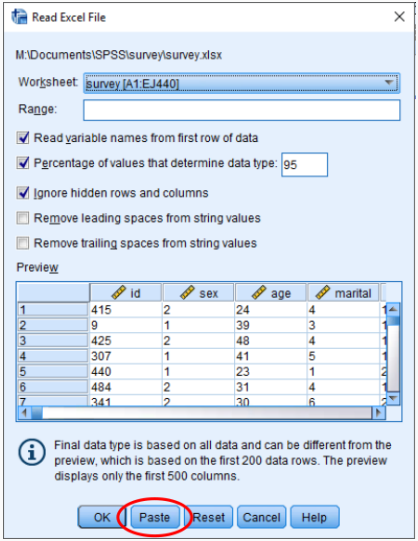
IMPORTING DATA - CSV

File > Import Data > CSV Data

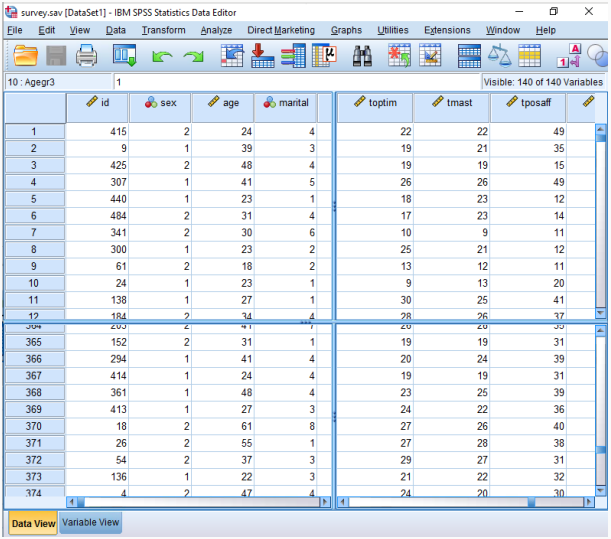


IMPORTING DATA - XLS/XLSX

File > Import Data > Excel



DATA VIEW - SPLIT VIEW

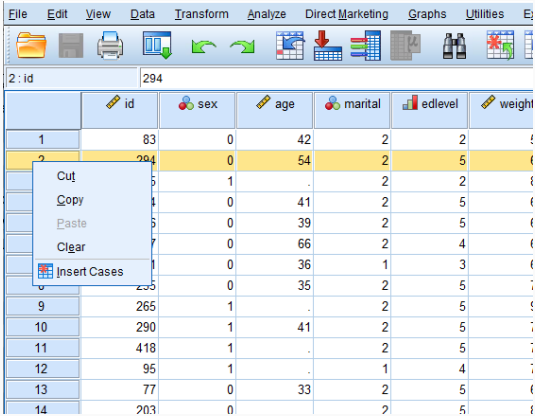


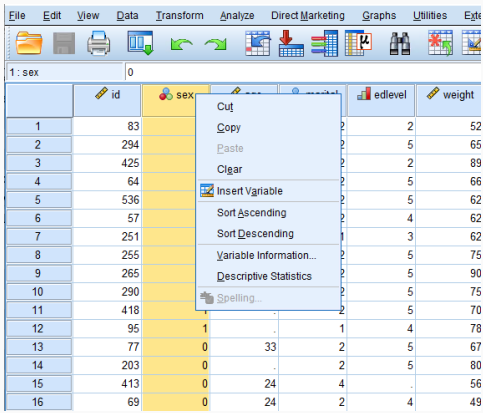
When you have very large data sets, it can be useful to 'split' the screen, so that you get four panels, all showing different sections of the spreadsheet.

Window > Split

EDITING A DATA FILE

Delete or insert cases or variables when you are in Data View

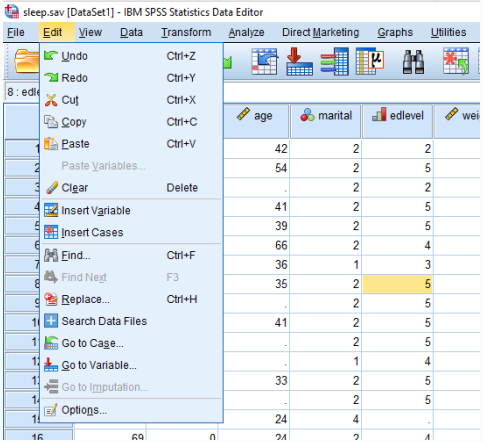


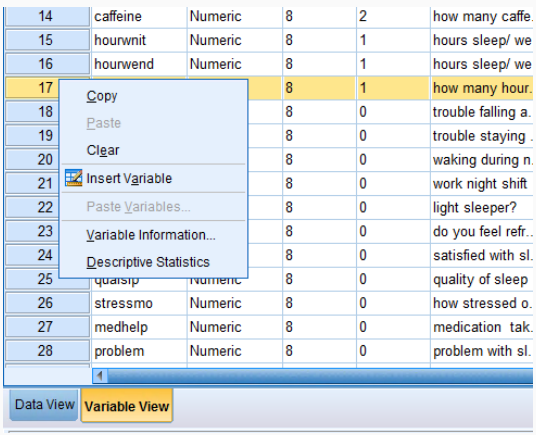


Remember to save the edited file with a new name, so the original data file is unaltered.

EDITING A DATA FILE

Other ways to delete or insert cases or variables

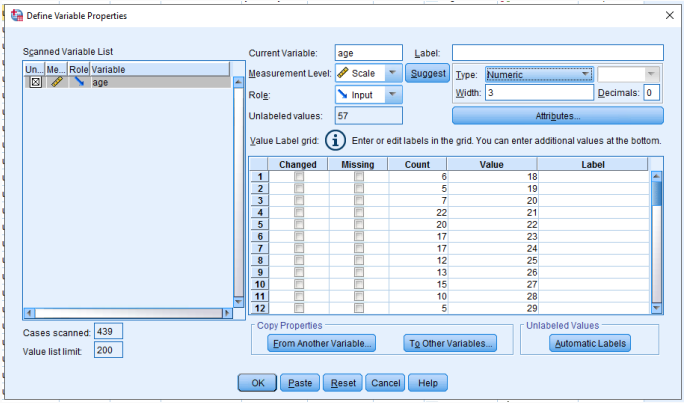




CHANGING VARIABLE PROPERTIES

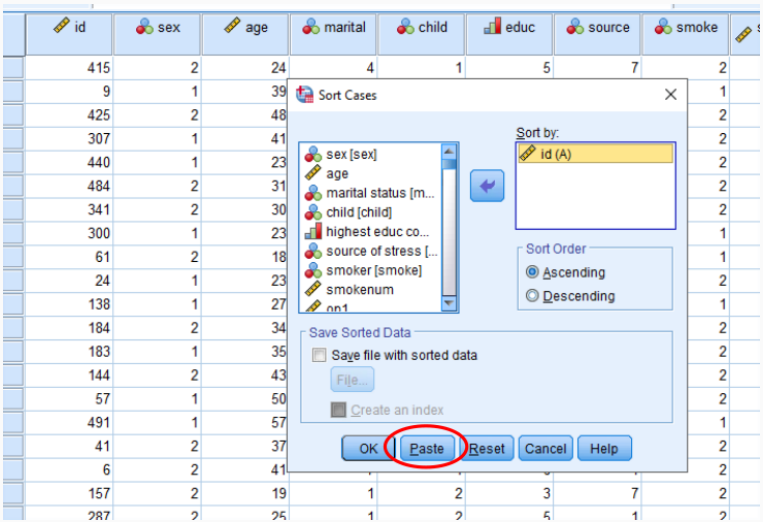
Data > Define Variable Properties

It is better to use this method than to do it directly in Variable View, as you have the possibility of documenting the changes you make, with the *Paste* function.



SORTING CASES

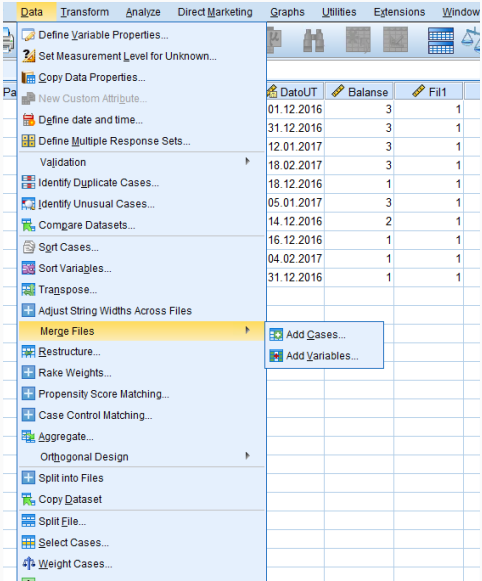
Data > Sort cases



MERGE TWO DATA FILES - ADDING CASES

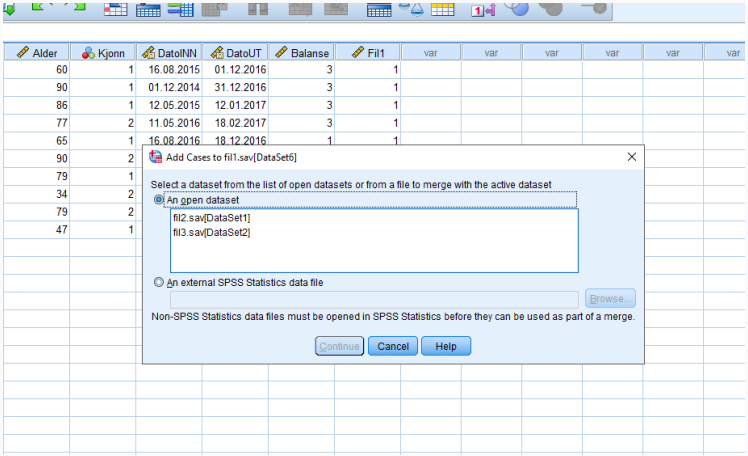
Data > Merge Files > Add Cases

We have two options when merging files; to add cases or to add variables. We will use files fil1, fil2 and fil3.



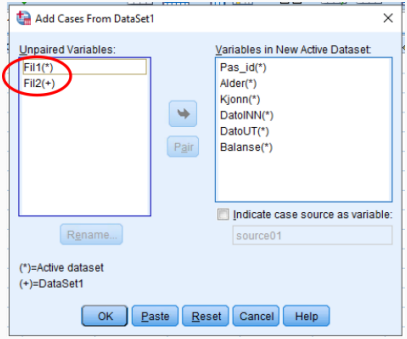
MERGE TWO DATA FILES - ADDING CASES

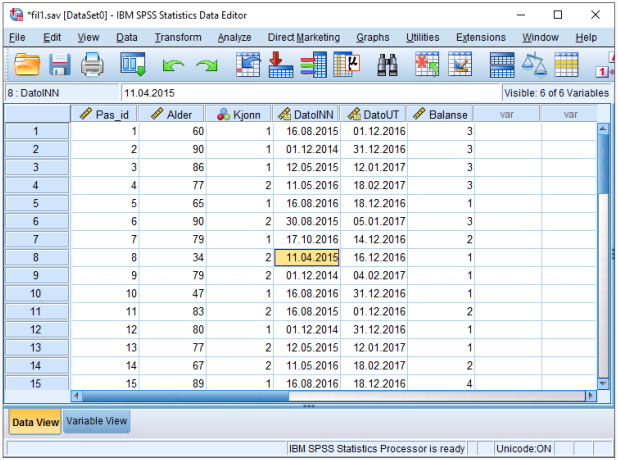
Choose either an open file (if the file you need is open), or select an external file (has to be .sav format)



MERGE TWO DATA FILES - ADDING CASES

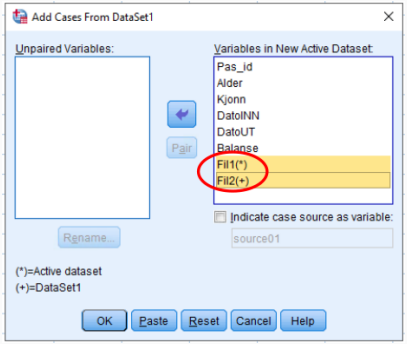
To merge files, but only with mutual variables. All other variables are excluded.

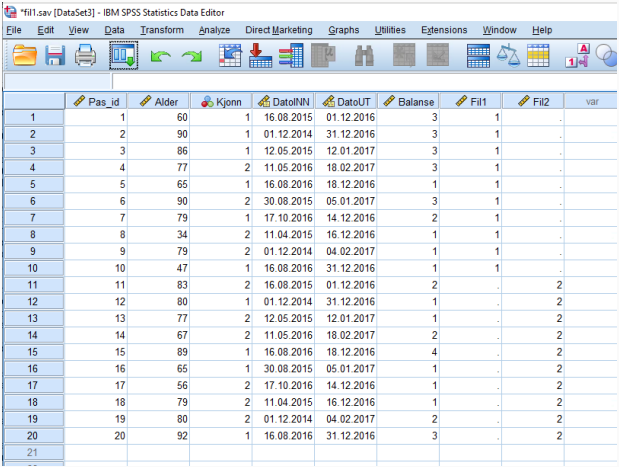




MERGE TWO DATA FILES - ADDING CASES

To merge files, with all variables, even if the variable is only found in one of the data sets

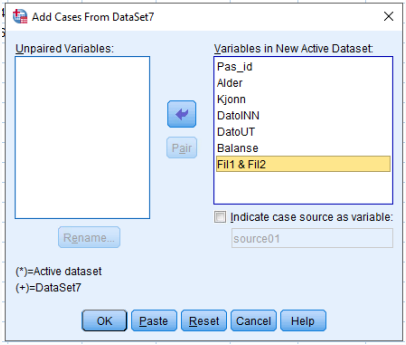


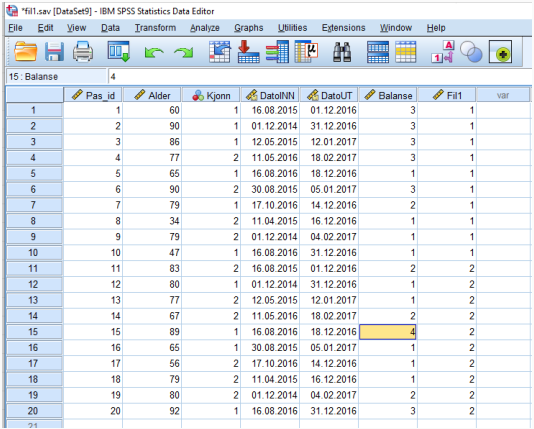


	Pas_id	Alder	Kjonn	DatoINN	DatoUT	Balanse	Fil1	Fil2	var
1	1	60	1	16.08.2015	01.12.2016	3	1	.	
2	2	90	1	01.12.2014	31.12.2016	3	1	.	
3	3	86	1	12.05.2015	12.01.2017	3	1	.	
4	4	77	2	11.05.2016	18.02.2017	3	1	.	
5	5	65	1	16.08.2016	18.12.2016	1	1	.	
6	6	90	2	30.08.2015	05.01.2017	3	1	.	
7	7	79	1	17.10.2016	14.12.2016	2	1	.	
8	8	34	2	11.04.2015	16.12.2016	1	1	.	
9	9	79	2	01.12.2014	04.02.2017	1	1	.	
10	10	47	1	16.08.2016	31.12.2016	1	1	.	
11	11	83	2	16.08.2015	01.12.2016	2	.	2	
12	12	80	1	01.12.2014	31.12.2016	1	.	2	
13	13	77	2	12.05.2015	12.01.2017	1	.	2	
14	14	67	2	11.05.2016	18.02.2017	2	.	2	
15	15	89	1	16.08.2016	18.12.2016	4	.	2	
16	16	65	1	30.08.2015	05.01.2017	1	.	2	
17	17	56	2	17.10.2016	14.12.2016	1	.	2	
18	18	79	2	11.04.2015	16.12.2016	1	.	2	
19	19	80	2	01.12.2014	04.02.2017	2	.	2	
20	20	92	1	16.08.2016	31.12.2016	3	.	2	
21									

MERGE TWO DATA FILES - ADDING CASES

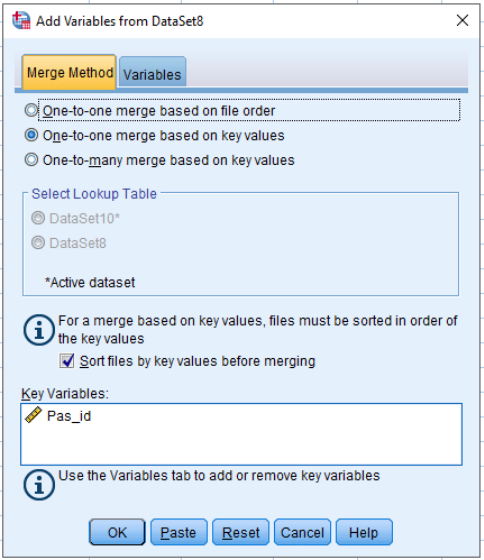
To merge files, with all variables, where variables that are only found in one of the sets can be merged with a variable only found in the other set. Use the *Pair* button rather than the arrow.





	Pas_id	Alder	Kjonn	DatoINN	DatoUT	Balanse	Fil1	var
1	1	60	1	16.08.2015	01.12.2016	3	1	
2	2	90	1	01.12.2014	31.12.2016	3	1	
3	3	86	1	12.05.2015	12.01.2017	3	1	
4	4	77	2	11.05.2016	18.02.2017	3	1	
5	5	65	1	16.08.2016	18.12.2016	1	1	
6	6	90	2	30.08.2015	05.01.2017	3	1	
7	7	79	1	17.10.2016	14.12.2016	2	1	
8	8	34	2	11.04.2015	16.12.2016	1	1	
9	9	79	2	01.12.2014	04.02.2017	1	1	
10	10	47	1	16.08.2016	31.12.2016	1	1	
11	11	83	2	16.08.2015	01.12.2016	2	2	
12	12	80	1	01.12.2014	31.12.2016	1	2	
13	13	77	2	12.05.2015	12.01.2017	1	2	
14	14	67	2	11.05.2016	18.02.2017	2	2	
15	15	89	1	16.08.2016	18.12.2016	4	2	
16	16	65	1	30.08.2015	05.01.2017	1	2	
17	17	56	2	17.10.2016	14.12.2016	1	2	
18	18	79	2	11.04.2015	16.12.2016	1	2	
19	19	80	2	01.12.2014	04.02.2017	2	2	
20	20	92	1	16.08.2016	31.12.2016	3	2	

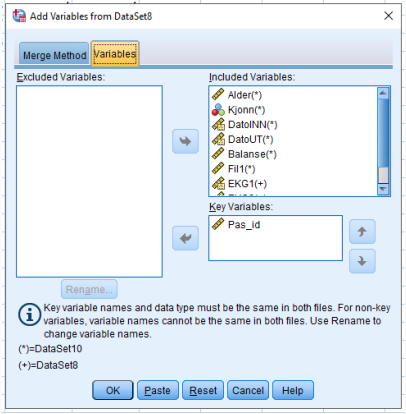
MERGE TWO DATA FILES - ADDING VARIABLES

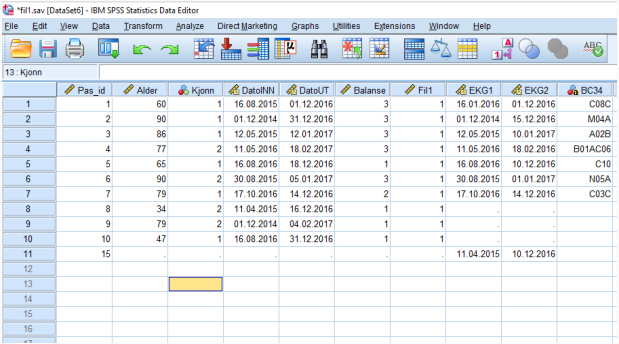


Data > Merge Files > Add Variables...

Here we use files fil1 and fil3. First the method of merging must be selected. *One-to-One merge based on key values* is default and gives all cases, from both files. A common variable is used as key, to link corresponding cases together.

MERGE TWO DATA FILES - ADDING VARIABLES

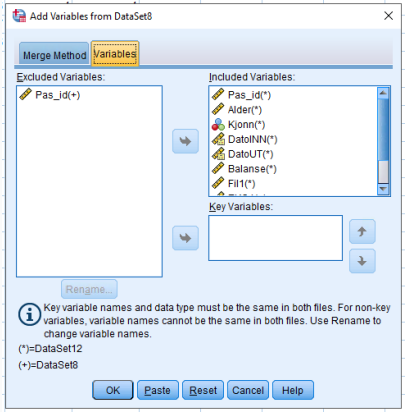


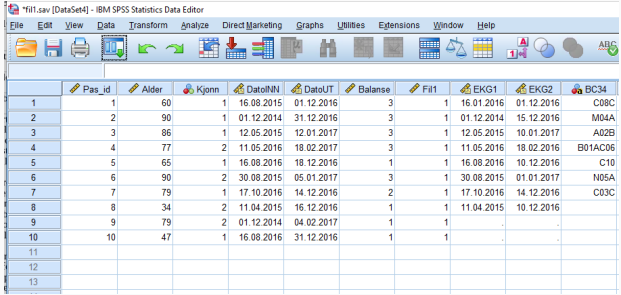


	Pas_id	Alder	Kjonn	DatoINN	DatoUT	Balanse	Fil1	EKG1	EKG2	BC34
1	1	60	1	16.08.2015	01.12.2016	3	1	16.01.2016	01.12.2016	C08C
2	2	90	1	01.12.2014	31.12.2016	3	1	01.12.2014	15.12.2016	M04A
3	3	86	1	12.05.2015	12.01.2017	3	1	12.05.2015	10.01.2017	A02B
4	4	77	2	11.05.2016	18.02.2017	3	1	11.05.2016	18.02.2016	B01AC06
5	5	65	1	16.08.2016	18.12.2016	1	1	16.08.2016	10.12.2016	C10
6	6	90	2	30.08.2015	05.01.2017	3	1	30.08.2015	01.01.2017	N05A
7	7	79	1	17.10.2016	14.12.2016	2	1	17.10.2016	14.12.2016	C03C
8	8	34	2	11.04.2015	16.12.2016	1	1	.	.	
9	9	79	2	01.12.2014	04.02.2017	1	1	.	.	
10	10	47	1	16.08.2016	31.12.2016	1	1			
11	15	.	.	.	.	.	.	11.04.2015	10.12.2016	
12										
13										
14										
15										
16										

MERGE TWO DATA FILES - ADDING VARIABLES

One-to-One merge based on file order results in a dataset containing all cases found in fil1, but excludes cases that are only found in fil3.

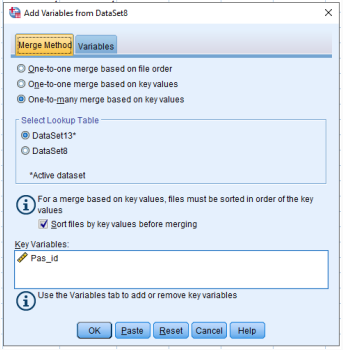


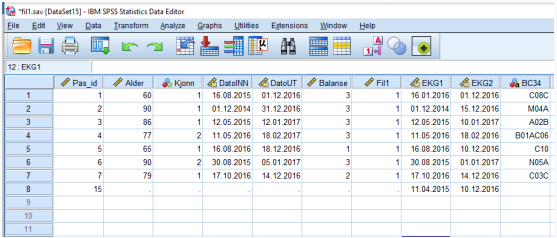


	Pas_id	Alder	Kjonn	DatoINN	DatoUT	Balanse	Fil1	EKG1	EKG2	BC34
1	1	60	1	16.08.2015	01.12.2016	3	1	16.01.2016	01.12.2016	C08C
2	2	90	1	01.12.2014	31.12.2016	3	1	01.12.2014	15.12.2016	M04A
3	3	86	1	12.05.2015	12.01.2017	3	1	12.05.2015	10.01.2017	A02B
4	4	77	2	11.05.2016	18.02.2017	3	1	11.05.2016	18.02.2016	B01AC06
5	5	65	1	16.08.2016	18.12.2016	1	1	16.08.2016	10.12.2016	C10
6	6	90	2	30.08.2015	05.01.2017	3	1	30.08.2015	01.01.2017	N05A
7	7	79	1	17.10.2016	14.12.2016	2	1	17.10.2016	14.12.2016	C03C
8	8	34	2	11.04.2015	16.12.2016	1	1	11.04.2015	10.12.2016	
9	9	79	2	01.12.2014	04.02.2017	1	1	.	.	
10	10	47	1	16.08.2016	31.12.2016	1	1	.	.	
11										
12										
13										

MERGE TWO DATA FILES - ADDING VARIABLES

One-to-Many merge based on key values has two options; where choosing the data set corresponding to fil3 will give the same result as *One-to-One merge based on file order*. If you choose the data set corresponding to fil1, you get only the cases found in fil3, with cases only found in fil1 excluded.

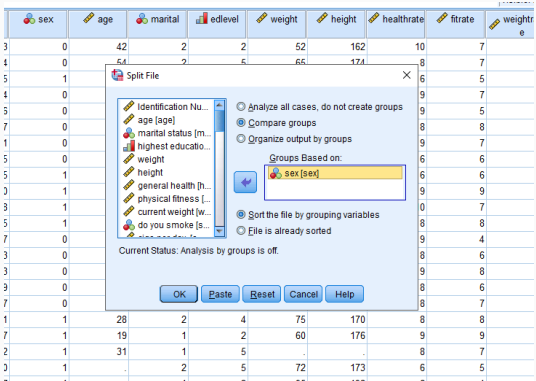




Split data set

Data > Split File

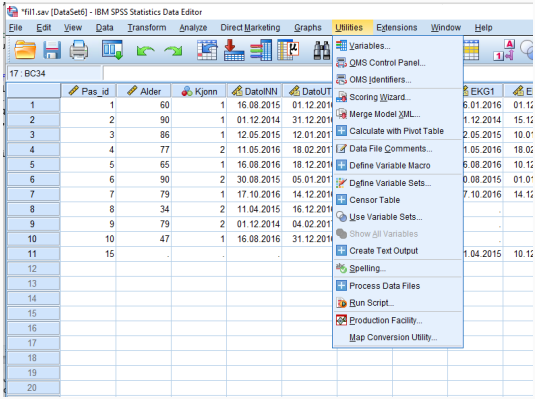
Choose *Compare groups* if you want data from different groups displayed in the same table. Choose *Organize output by groups* if you want results from different groups in separate tables.

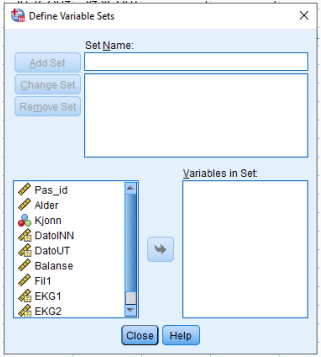


DEFINE A SUBSET FROM LARGE SETS

Utilities > Define Variable Sets

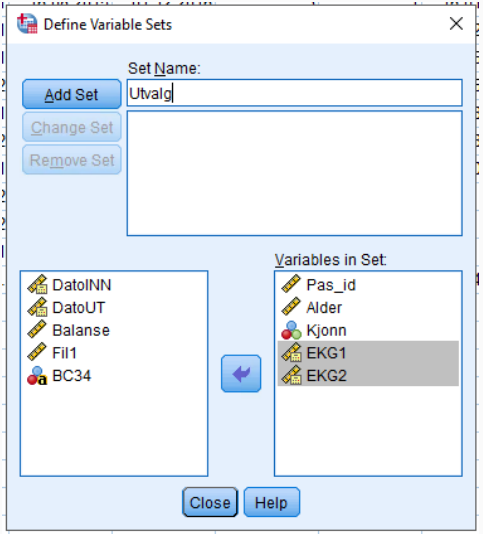
The new set can contain as many of the original variables as you want.





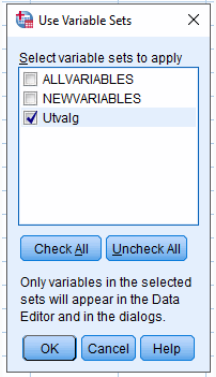
DEFINE A SUBSET FROM LARGE SETS

The new subset needs a name. Remember to click *Add set* before you close the dialogue box. If you don't, the new set you defined is not saved, and you have to define it again.



USING A SUBSET

Utilities > Use Variable Sets: In order to work with the new subset, select it from the list of available sets. When you wish to go back to working with the entire set, choose ALLVARIABLES in the same dialogue box.



*fil1.sav [DataSet6] - IBM SPSS Statistics Data Editor

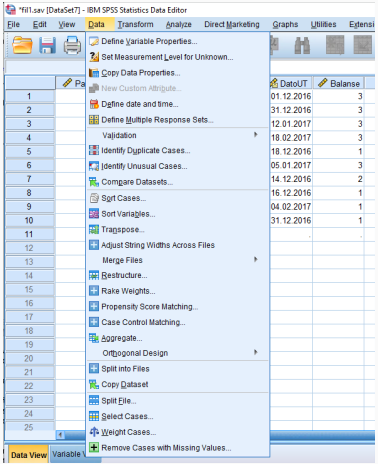
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Ex

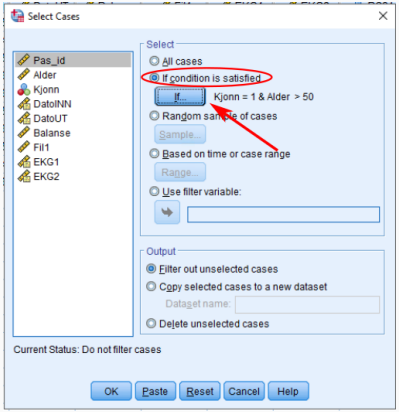
16 : Fil1

	Pas_id	Alder	Kjonn	EKG1	EKG2	var
1	1	60	1	16.01.2016	01.12.2016	
2	2	90	1	01.12.2014	15.12.2016	
3	3	86	1	12.05.2015	10.01.2017	
4	4	77	2	11.05.2016	18.02.2016	
5	5	65	1	16.08.2016	10.12.2016	
6	6	90	2	30.08.2015	01.01.2017	
7	7	79	1	17.10.2016	14.12.2016	
8	8	34	2	.	.	
9	9	79	2	.	.	
10	10	47	1	.	.	
11	15	.	.	11.04.2015	10.12.2016	
12						
13						

SELECTION OF DATA BASED ON A CONDITION

Data > Select cases: When you wish to make a more specific selection than you can achieve with the *split file* option.

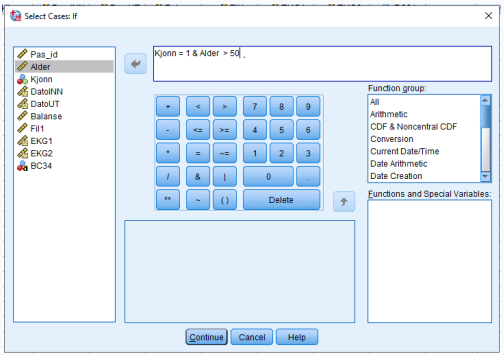


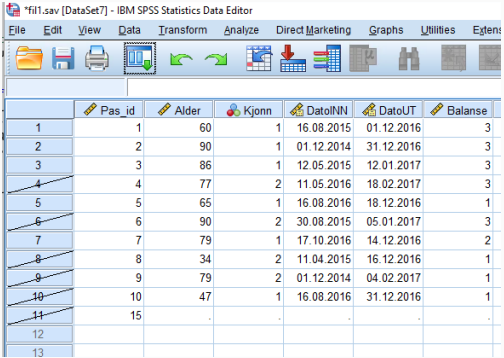


SELECTION OF DATA BASED ON A CONDITION

Select Cases: If

Choose which cases you want. As an example, here we select all men over 50.



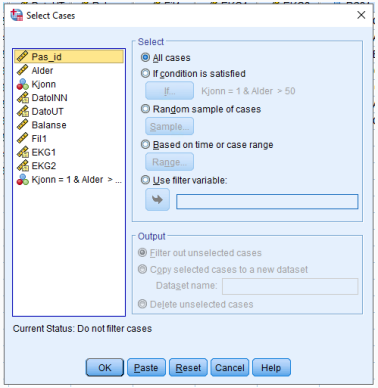


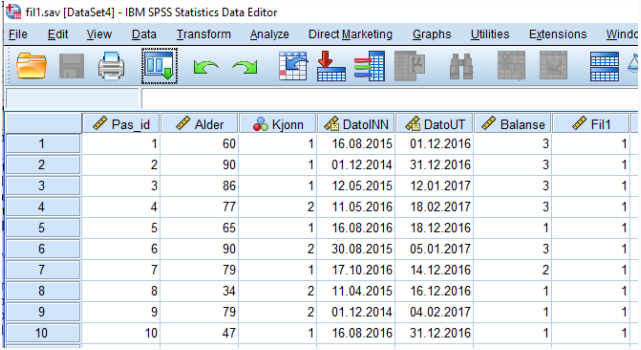
	Pas_id	Alder	Kjonn	DatoINN	DatoUT	Balanse
1	1	60	1	16.08.2015	01.12.2016	3
2	2	90	1	01.12.2014	31.12.2016	3
3	3	86	1	12.05.2015	12.01.2017	3
4	4	77	2	11.05.2016	18.02.2017	3
5	5	65	1	16.08.2016	18.12.2016	1
6	6	90	2	30.08.2015	05.01.2017	3
7	7	79	1	17.10.2016	14.12.2016	2
8	8	34	2	11.04.2015	16.12.2016	1
9	9	79	2	01.12.2014	04.02.2017	1
10	10	47	1	16.08.2016	31.12.2016	1
11	15	.	.	.	.	.
12						
13						

SELECTION OF DATA BASED ON A CONDITION

Select Cases: All cases

When you want to go back to looking at all cases again, remember to go through the same dialogue box, but select *All cases*.





	Pas_id	Alder	Kjonn	DatoINN	DatoUT	Balanse	Fil1
1	1	60	1	16.08.2015	01.12.2016	3	1
2	2	90	1	01.12.2014	31.12.2016	3	1
3	3	86	1	12.05.2015	12.01.2017	3	1
4	4	77	2	11.05.2016	18.02.2017	3	1
5	5	65	1	16.08.2016	18.12.2016	1	1
6	6	90	2	30.08.2015	05.01.2017	3	1
7	7	79	1	17.10.2016	14.12.2016	2	1
8	8	34	2	11.04.2015	16.12.2016	1	1
9	9	79	2	01.12.2014	04.02.2017	1	1
10	10	47	1	16.08.2016	31.12.2016	1	1

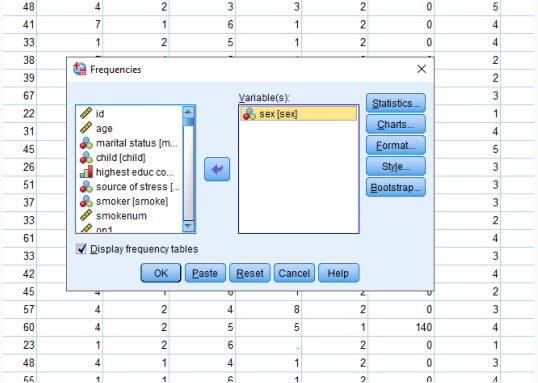
Before we begin to analyze our data...

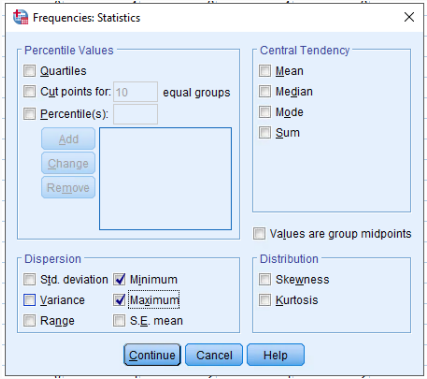
IS EVERYTHING CORRECT?

In case of data import, does it look correct? Are variables coded correctly?

For this part we use the file error.sav

Categorical variable: Analyze > Descriptive statistics > Frequencies





Is everything correct?

We get two tables. *Statistics* gives the number of valid data entries, number of missing values, maximum and minimum values. The other table tells us how many cases that have not been coded correctly.

Statistics

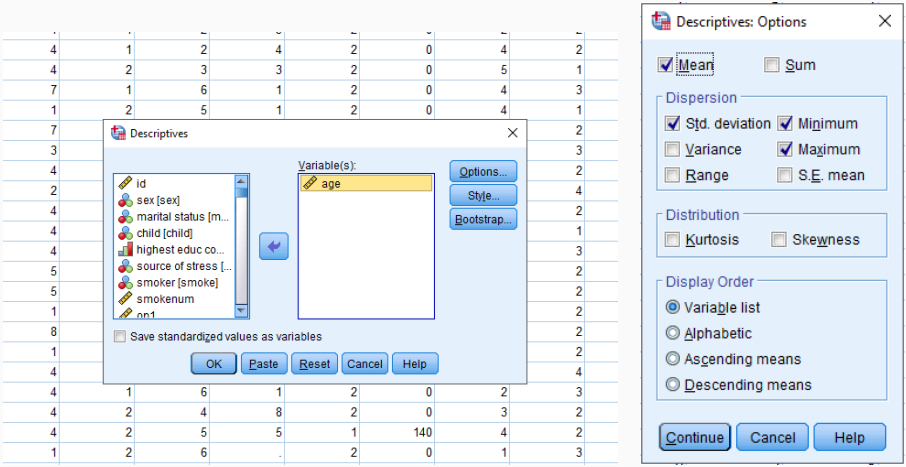
		sex	marital status	highest educ completed
N	Valid	438	438	438
	Missing	0	0	0
Minimum		1	1	1
Maximum		3	8	22

sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 MALES	185	42,2	42,2	42,2
	2 FEMALES	252	57,5	57,5	99,8
	3	1	,2	,2	100,0
	Total	438	100,0	100,0	

IS EVERYTHING CORRECT?

Continuous variable: Analyze > Descriptive statistics > Descriptives



Is everything correct?

The table gives amongst other things, maximum and minimum values, which can tell us if someone are outside of the possible range. In this case, at least one person has an age which is wrong, as all participants are over 18.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
age	438	2	82	37,42	13,290
Valid N (listwise)	438				

If you make any changes to the data, remember to document it

DESCRIPTIVE STATISTICS

The difference between a sample and a population

We use statistics to say something about a population as a whole, based on results from a sample.

The difference between a sample and a population

We use statistics to say something about a population as a whole, based on results from a sample.

Even though a sample is randomly chosen and representative of the population, there will always be small individual variations, which can give rise to slightly different results each time an experiment is carried out.

The difference between a sample and a population

We use statistics to say something about a population as a whole, based on results from a sample.

Even though a sample is randomly chosen and representative of the population, there will always be small individual variations, which can give rise to slightly different results each time an experiment is carried out.

This means that our results can be a bit different from the true values of the population.

The difference between a sample and a population

We use statistics to say something about a population as a whole, based on results from a sample.

Even though a sample is randomly chosen and representative of the population, there will always be small individual variations, which can give rise to slightly different results each time an experiment is carried out.

This means that our results can be a bit different from the true values of the population.

The greater the sample, the closer we are likely to be to the true values. Rule of thumb: have at least 30 cases, ideally many more. If there is much variation, or if you are looking for small differences or effects, you need more cases.

Important statistical measures

Important statistical measures

- Frequency and relative frequency - how many fall into each category, have answered yes or no, how many are men how many women, etc.

Important statistical measures

- Frequency and relative frequency - how many fall into each category, have answered yes or no, how many are men how many women, etc.
- Mode - the most common value or category you measure. There can be more than one mode.

Important statistical measures

- Frequency and relative frequency - how many fall into each category, have answered yes or no, how many are men how many women, etc.
- Mode - the most common value or category you measure. There can be more than one mode.
- Mean - the average. The mean does not have to be a value you have actually measured.

Important statistical measures

- Frequency and relative frequency - how many fall into each category, have answered yes or no, how many are men how many women, etc.
- Mode - the most common value or category you measure. There can be more than one mode.
- Mean - the average. The mean does not have to be a value you have actually measured.
- Median - the middle measurement when data are arranged in ascending order.

Important statistical measures

- Standard deviation and standard error - the variation or spread in the data, deviation from the mean.

Important statistical measures

- Standard deviation and standard error - the variation or spread in the data, deviation from the mean.
- Skewness - deviation from symmetry.

Important statistical measures

- Standard deviation and standard error - the variation or spread in the data, deviation from the mean.
- Skewness - deviation from symmetry.
- Kurtosis - long tails and pointy distribution, or short tails and data mostly found around or close to the mean.

Important statistical measures

- Standard deviation and standard error - the variation or spread in the data, deviation from the mean.
- Skewness - deviation from symmetry.
- Kurtosis - long tails and pointy distribution, or short tails and data mostly found around or close to the mean.
- Percentiles - value below which a certain percentage of the data are found, e.g. the 25% percentile indicates the value marking the upper limit of the lower 25% of the data.

Important statistical measures

- Standard deviation and standard error - the variation or spread in the data, deviation from the mean.
- Skewness - deviation from symmetry.
- Kurtosis - long tails and pointy distribution, or short tails and data mostly found around or close to the mean.
- Percentiles - value below which a certain percentage of the data are found, e.g. the 25% percentile indicates the value marking the upper limit of the lower 25% of the data.
- Interquartile Range (IQR) - the middle half of the data.

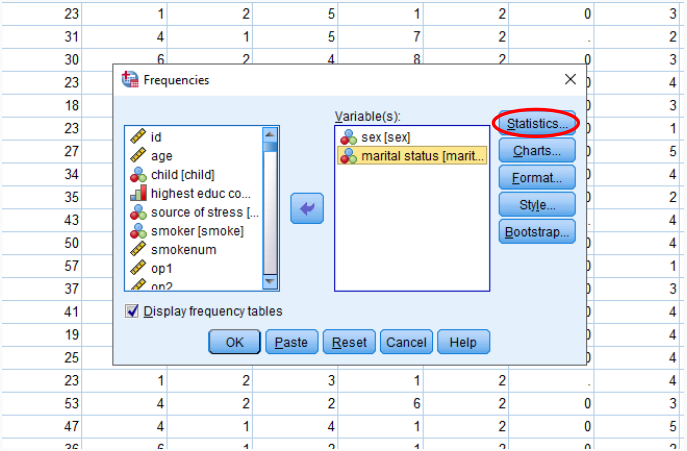
Important statistical measures

- Standard deviation and standard error - the variation or spread in the data, deviation from the mean.
- Skewness - deviation from symmetry.
- Kurtosis - long tails and pointy distribution, or short tails and data mostly found around or close to the mean.
- Percentiles - value below which a certain percentage of the data are found, e.g. the 25% percentile indicates the value marking the upper limit of the lower 25% of the data.
- Interquartile Range (IQR) - the middle half of the data.
- Confidence interval (CI) - interval estimate rather than a point estimate of a statistical measure. If we conduct the same experiment many times, then the 95% CI for the mean will in 95% of the experiments contain the true mean of the population.

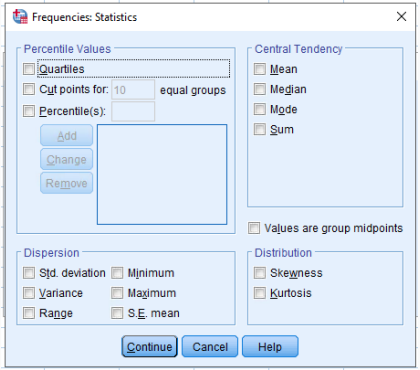
CATEGORICAL VARIABLES

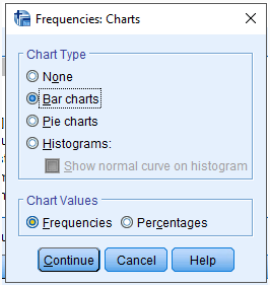
Analyze > Descriptive Statistics > Frequencies

- Frequencies gives a frequency table and useful plots for categorical variables
- The tick box *Display Frequency Tables* is ticked by default
- Several variables can be studied at the same time



Frequencies: Statistics og Frequencies: Charts





None of these statistical measures are appropriate for categorical variables. You can tick *Mode*, but this measure can just as easily be identified from a graph or a frequency table.

The most useful graphs for categorical variables are either bar or pie charts. Choose whichever you prefer.

CATEGORICAL VARIABLES - OUTPUT

Statistics

		sex	marital status
N	Valid	439	439
	Missing	0	0

Frequency Table

sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MALES	185	42,1	42,1	42,1
	FEMALES	254	57,9	57,9	100,0
	Total	439	100,0	100,0	

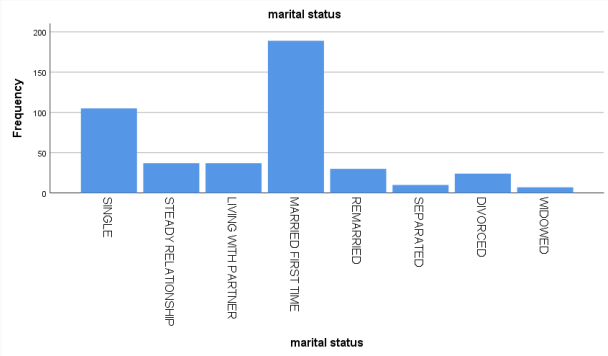
marital status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 SINGLE	105	23,9	23,9	23,9
	2 STEADY RELATIONSHIP	37	8,4	8,4	32,3
	3 LIVING WITH PARTNER	37	8,4	8,4	40,8
	4 MARRIED FIRST TIME	189	43,1	43,1	83,8
	5 REMARRIED	30	6,8	6,8	90,7
	6 SEPARATED	10	2,3	2,3	92,9
	7 DIVORCED	24	5,5	5,5	98,4
	8 WIDOWED	7	1,6	1,6	100,0
	Total	439	100,0	100,0	

The tables in the output show the number of cases included in total, and a frequency table for each of the variables included in the analysis. This provides you with both relative and regular frequency of each category.

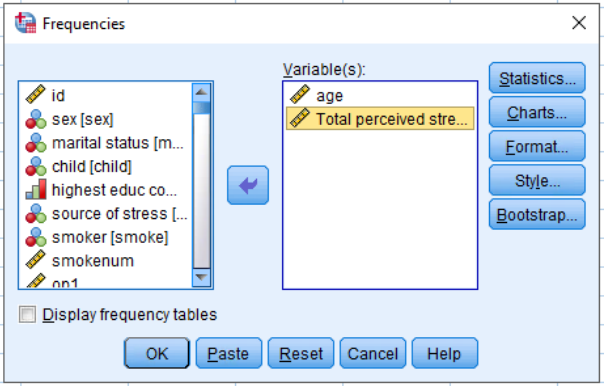
CATEGORICAL VARIABLE - OUTPUT

Bar chart of marital status. Here we can clearly see that *Married first time* is the mode. In addition we get a good overview of how the participants in this study are distributed over the different categories.



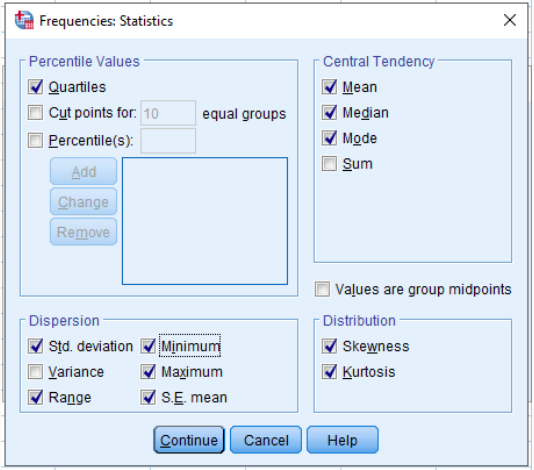
Analyze > Descriptive Statistics > Frequencies

Frequencies can also be used for continuous variables. Just remember to de-select *Display frequency tables*, otherwise your output will start with a very large table containing the frequencies for each of the measured values in the continuous variables.



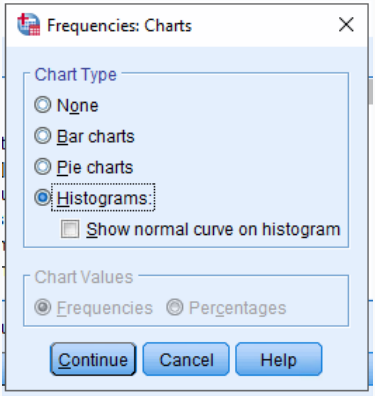
CONTINUOUS VARIABLES

With continuous variables you can select all the statistical measures that were ill suited for categorical variables, like mean, standard deviation, etc.



Frequencies: Charts

Bar and pie charts are less useful for continuous variables, but histograms are very useful. Tick the box that says *Show normal curve on histogram*, in order to get an indication of whether the variable is close to a normal distribution or not.



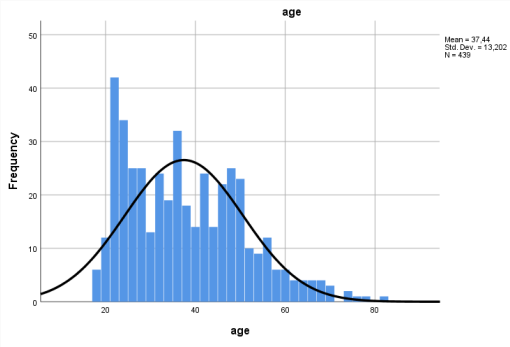
Continuous variables - Output

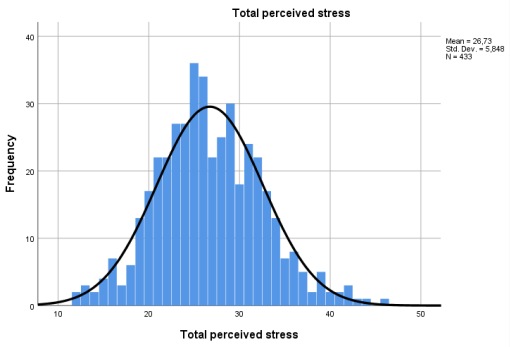
Statistics

		age	Total perceived stress
N	Valid	439	433
	Missing	0	6
Mean		37,44	26,73
Std. Error of Mean		,630	,281
Median		36,00	26,00
Mode		21	25
Std. Deviation		13,202	5,848
Skewness		,606	,245
Std. Error of Skewness		,117	,117
Kurtosis		-,203	,182
Std. Error of Kurtosis		,233	,234
Range		64	34
Minimum		18	12
Maximum		82	46
Percentiles	25	26,00	23,00
	50	36,00	26,00
	75	47,00	31,00

The table produced by the *Frequencies*-procedure gives a good overview of much descriptive statistics concerning these variables. This is very useful to go through before starting more complicated procedures and analyses.

The histogram with the normal curve indicates fairly easily how well the data fit with a normal distribution.

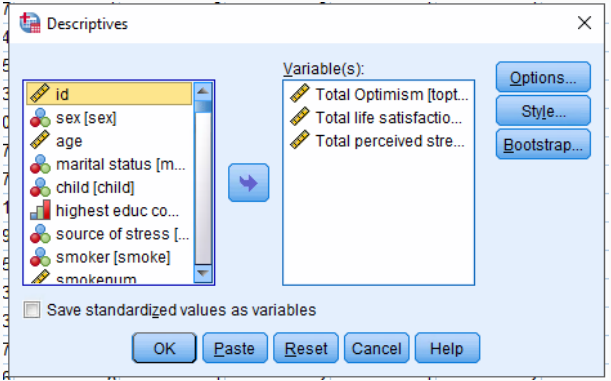




CONTINUOUS VARIABLE

For continuous variables we also have a different option.

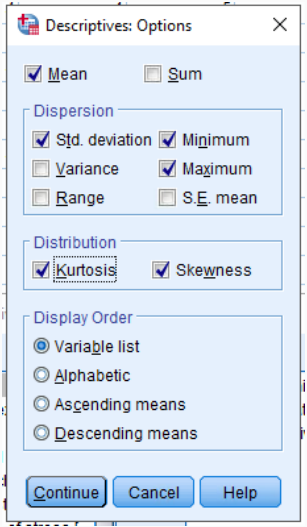
Analyze > Descriptive Statistics > Descriptives



This option is particularly useful if you wish to compare several continuous variables. In addition, it can be used to create new variables containing the data converted to a standard normal distribution (z-transform). This in turn can be useful when screening for outliers.

Descriptives: Options

Many of the same options as in *Frequencies* can also be selected here.



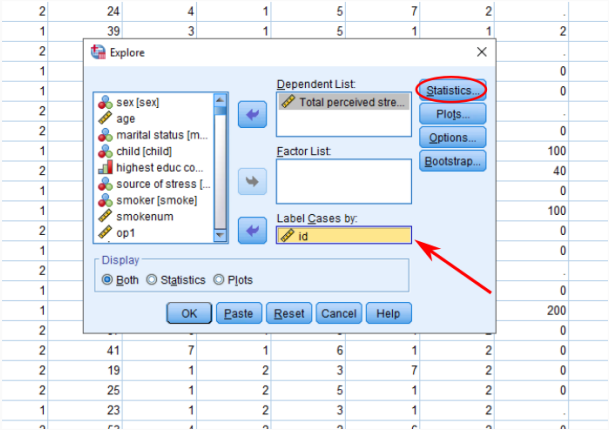
The benefit with this routine is that variables are listed below each other, with different statistical measures listed next to each other. This makes it easier to compare means, standard deviation, etc. for variables measured on the same scale.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Total Optimism	435	7	30	22,12	4,429	-,494	,117	,214	,234
Total life satisfaction	436	5	35	22,38	6,770	-,323	,117	-,450	,233
Total perceived stress	433	12	46	26,73	5,848	,245	,117	,182	,234
Valid N (listwise)	432								

For additional descriptives, and to check normality of continuous variables.

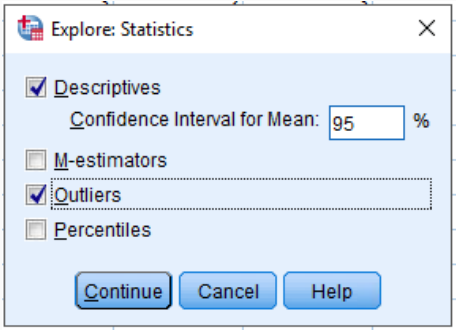
Analyze > Descriptive Statistics > Explore



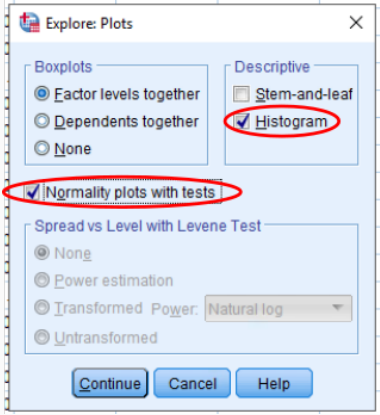
- Dependent list: The variable/variables that you wish to analyze
- Factor list: Categorical variable that will group the values in the continuous variable(s)
- Label cases by: Typically the ID variable, in order to identify extreme cases more easily

Explore: Statistics

- Descriptives: Calculates mean, confidence interval for the mean, median, standard deviation, etc.
- M-estimators: Robust alternatives to mean and standard deviation
- Outliers: Prints the five highest and lowest values measured
- Percentiles: Finds 5%, 10%, 25%, 75%, 90% and 95% percentiles



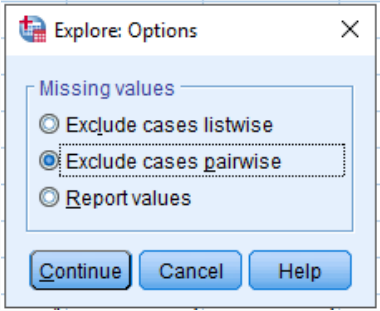
Explore: Plots



- Boxplots: Factor levels together - If you have provided a categorical grouping variable this option will plot the groups together
- Boxplots: Dependents together -If you have more than one dependent variable, these will be plotted together
- Descriptive: Histogram is usually the most useful choice
- Normality plots with tests: Plots and tables which can make it more clear if a variable is normally distributed or not

Explore: Options

- Exclude cases listwise: Excludes all cases where one or more data point is missing for any variable included in the analysis
- Exclude cases pairwise: Excludes only those cases that are missing one or more data point in the variable in question included in the specific analysis
- Report values: When a factor variable is chosen, empty cells will be treated as a separate category



Output: Case Processing Summary and Descriptives

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Total perceived stress	433	98,6%	6	1,4%	439	100,0%

Descriptives

			Statistic	Std. Error
Total perceived stress	Mean		26,73	,281
	95% Confidence Interval for Mean	Lower Bound	26,18	
		Upper Bound	27,28	
	5% Trimmed Mean		26,64	
	Median		26,00	
	Variance		34,194	
	Std. Deviation		5,848	
	Minimum		12	
	Maximum		46	
	Range		34	
	Interquartile Range		8	
	Skewness		,245	,117
	Kurtosis		,182	,234

- Case Processing Summary: number of cases, and how many valid
- Descriptives: Descriptive statistics. Look e.g. at the difference between mean and median to see if the mean seems like a good measure of central tendency. Skewness and kurtosis are also useful. The standard normal distribution has skewness and kurtosis equal to 0.

Output: Tests of Normality and Extreme Values

- Tests of Normality: Sigma greater than 0.05 indicates normality. HOWEVER, this is very rarely achieved with large data sets. It is usually better to look at the plots to determine normality.
- Extreme values: The five greatest and smallest values measured.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Total perceived stress	,069	433	,000	,992	433	,021

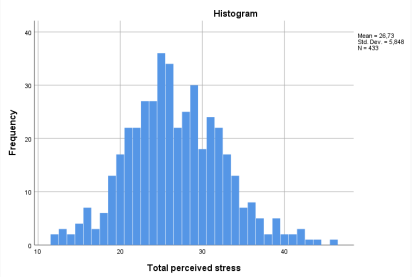
a. Lilliefors Significance Correction

Extreme Values

			Case Number	id	Value
Total perceived stress	Highest	1	10	24	46
		2	19	157	44
		3	9	61	43
		4	14	144	42
		5	18	6	42[a]
	Lowest	1	45	5	12
		2	42	404	12
		3	69	119	13
		4	56	301	13
		5	53	127	13

a. Only a partial list of cases with the value 42 are shown in the table of upper extremes.

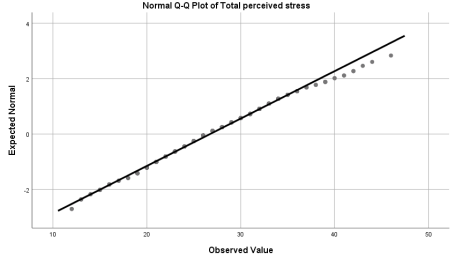
NORMALITY

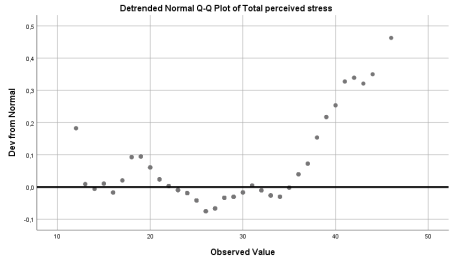


- Histogram: Here we see that the data distribution is a bit skewed, but the overall distribution is close to a normal distribution
- Boxplot: Shows much the same as the histogram

NORMALITY

- Normal Q-Q plot: The black line shows where the data should be if it is perfectly normal.Except for the right tail, the data lie fairly close to the line
- Detrended normal Q-Q plot: This shows the deviation between the data and the normal distribution more clearly. There is no clear trend in the deviation, which is a good sign, but we see even more clearly that the right tail is more heavy than expected compared to the normal distribution.





Useful procedures

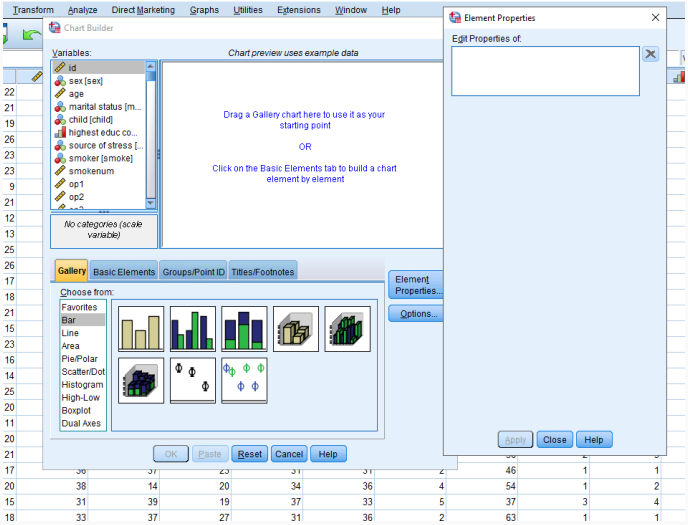
- Frequencies - distribution of answers for categorical variables, and much descriptive statistics for continuous variables
- Descriptives - descriptive statistics for continuous variables, especially useful for comparisons between variables
- Explore - All of the above, in addition to normality plots and tests, etc.

GRAPHS AND FIGURES

MAKING GRAPHS IN SPSS

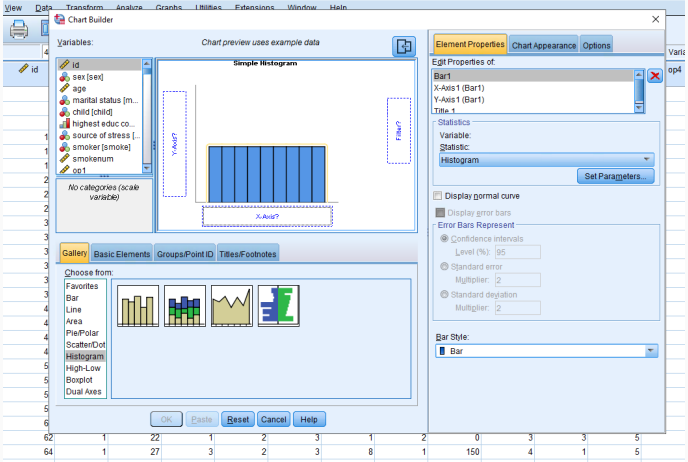
Graphs > Chart Builder

You choose which type of graph you want to make, by moving the selected type from the gallery to the preview window.



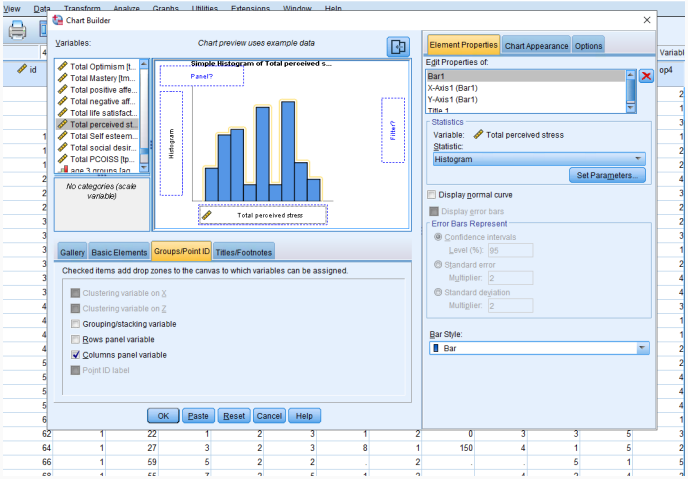
HISTOGRAM

This type of graph is usually used to present the distribution of a continuous variable.



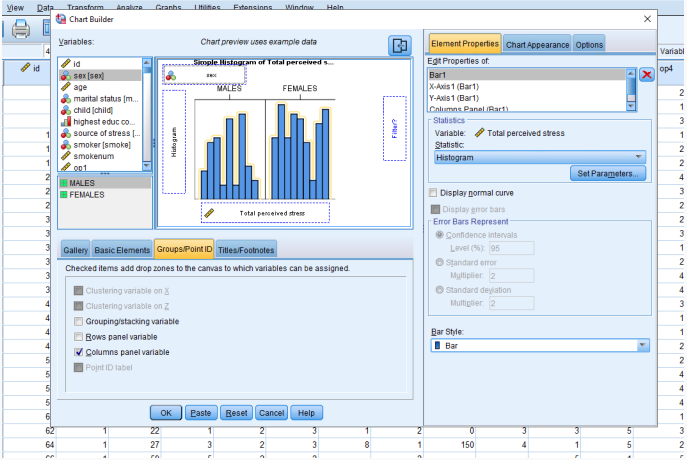
HISTOGRAM

Choose a continuous variable you want to plot. By also choosing *Columns panel variable* under the *Groups/Point ID* tab, you can group data according to groups in a categorical variable.



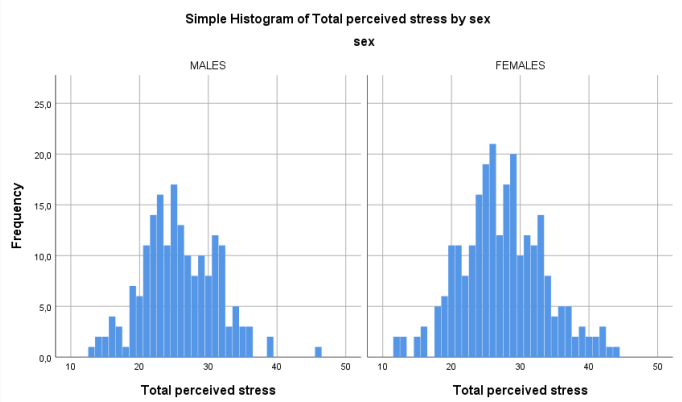
HISTOGRAM

This results in one histogram for each group in the categorical variable.



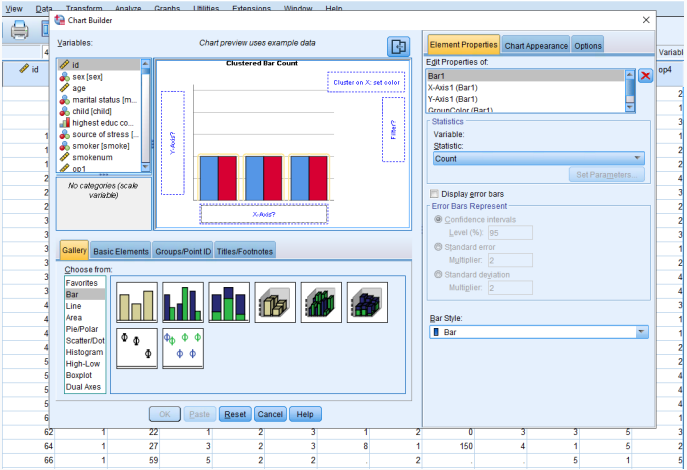
HISTOGRAM

The two graphs have the same scale on the y-axis, which shows the frequencies for intervals in the continuous variable.



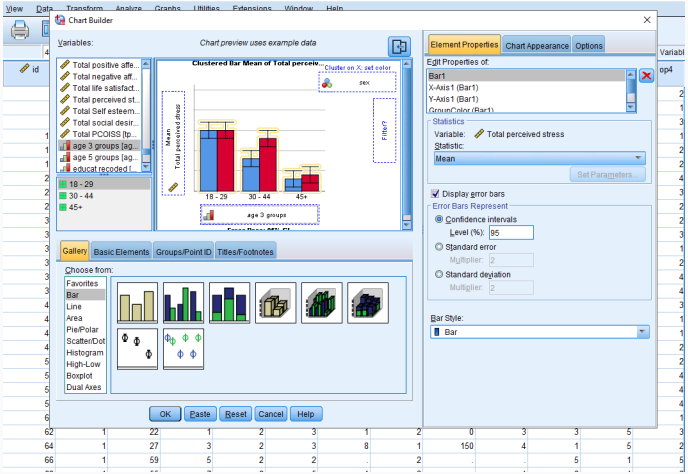
BAR CHART

Useful when you wish to show number of cases within different groups in a categorical variable, or to display the average in a continuous variable within different groups.



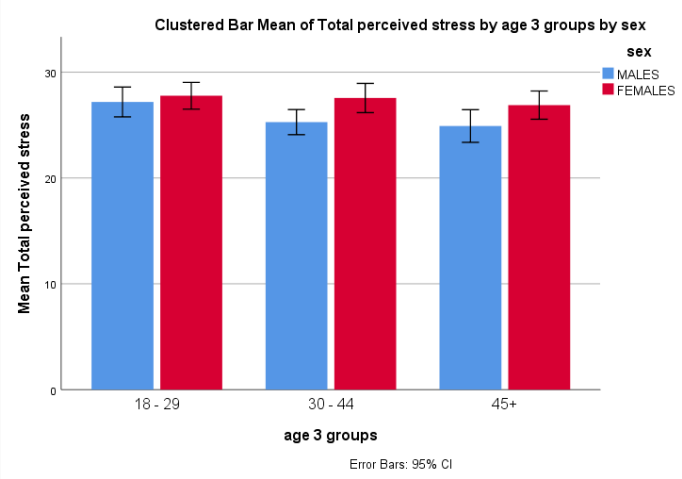
BAR CHART

By selecting two different categorical variables, you can make a clustered bar graph. Tick *Display error bars* in the *Element properties* tab, when plotting the mean of a continuous variable.



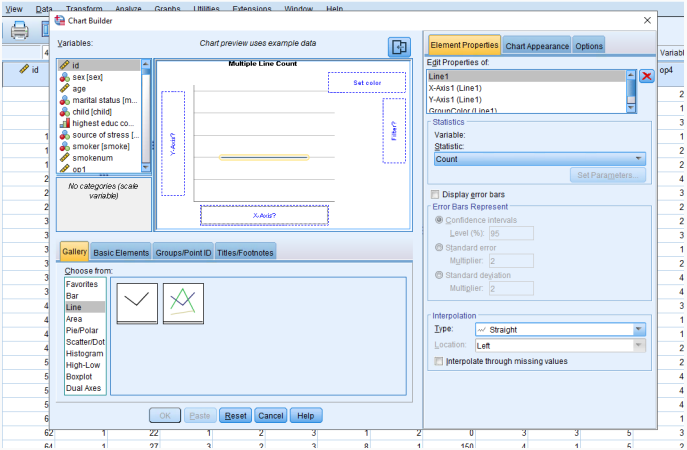
BAR CHART

The bars are separated into three main groups, according to the age variable, with two bars of different colours for each group, representing men and women.



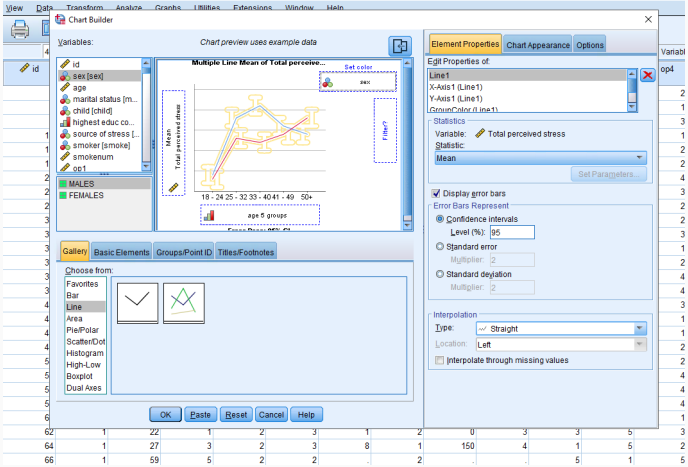
LINE CHART

Often used to display difference or development over time or at different experimental conditions, within a continuous variable.



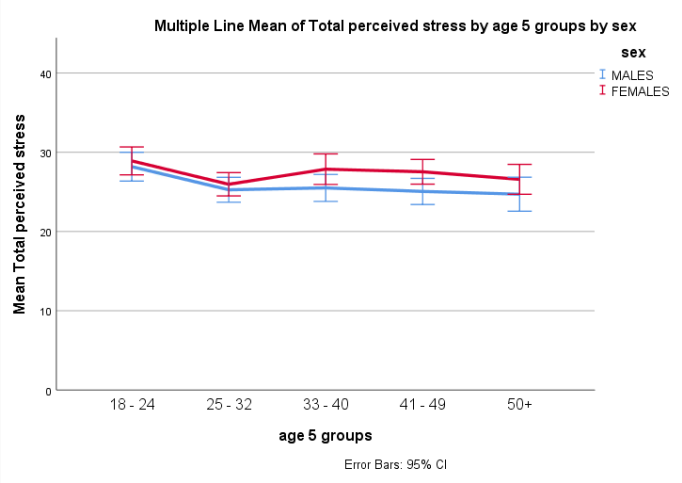
LINE CHART

By selecting two different categorical variables, we can both have one indicating the development (x-axis), and one giving groups that will be plotted as different lines. Select *Display error bars* in the *Element properties* tab.



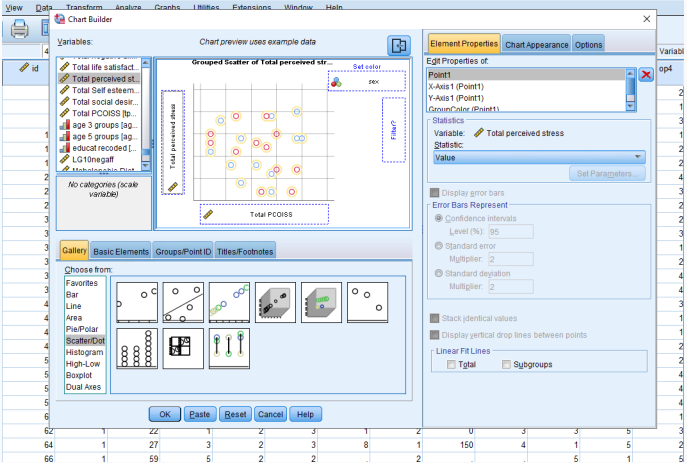
LINE CHART

In this case the different lines for men and women could help show if men and women have similar or different development patterns as age increases.



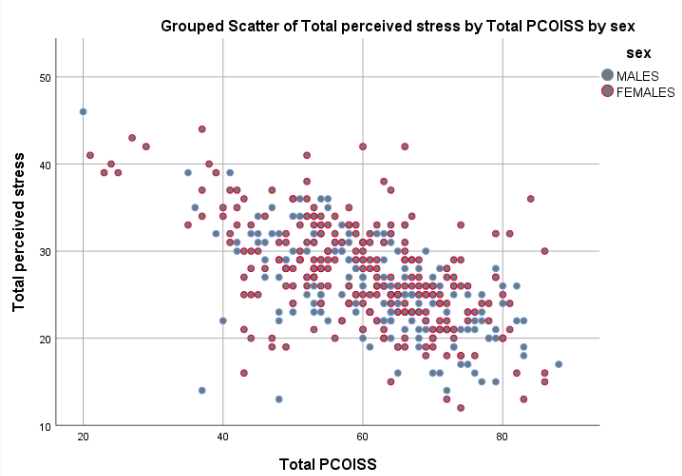
SCATTER PLOT

To show relation between two variables, where at least one of the variables is continuous. A third categorical variable can be used to plot cases from different groups as different colours.



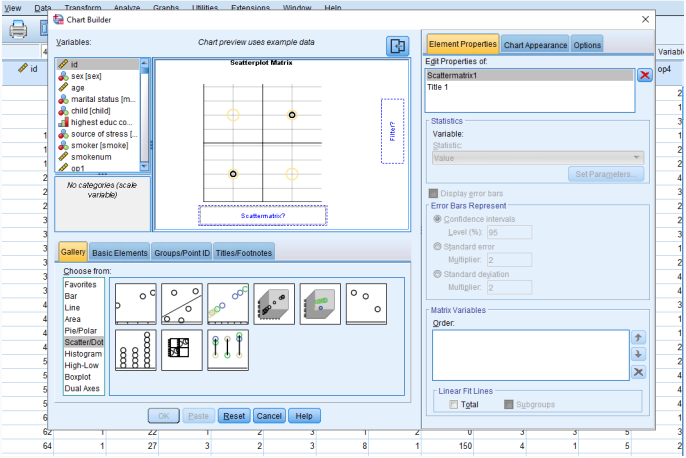
SCATTER PLOT

This type of graph is useful to look for correlation, as well as extreme outliers. In addition, it can show to some extent if there are large differences between groups.



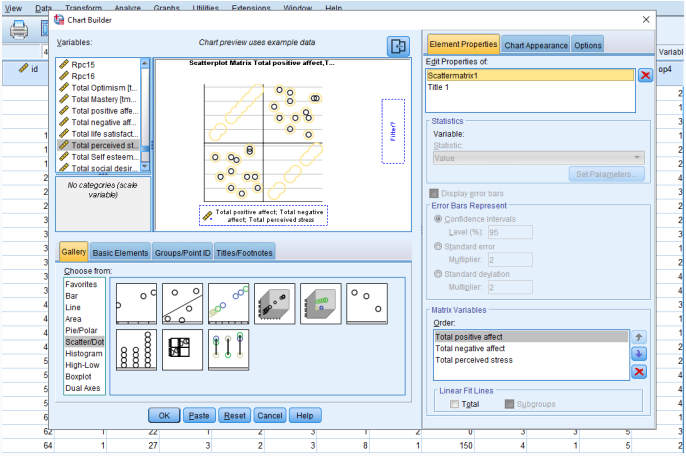
SCATTER MATRIX

If you wish to show relation between more than two continuous variables at the same time, you can create a scatter matrix.



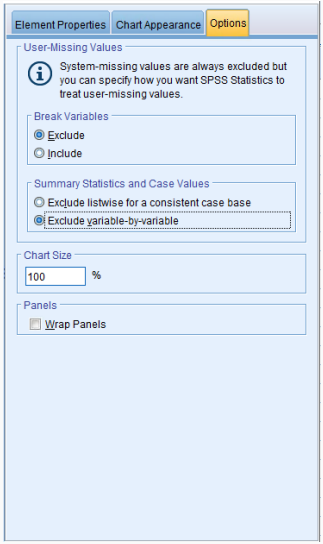
SCATTER MATRIX

Include all the variables you are interested in. The number of variables will decide the size of the matrix. With three variables, the matrix will contain 3×3 smaller boxes with charts.



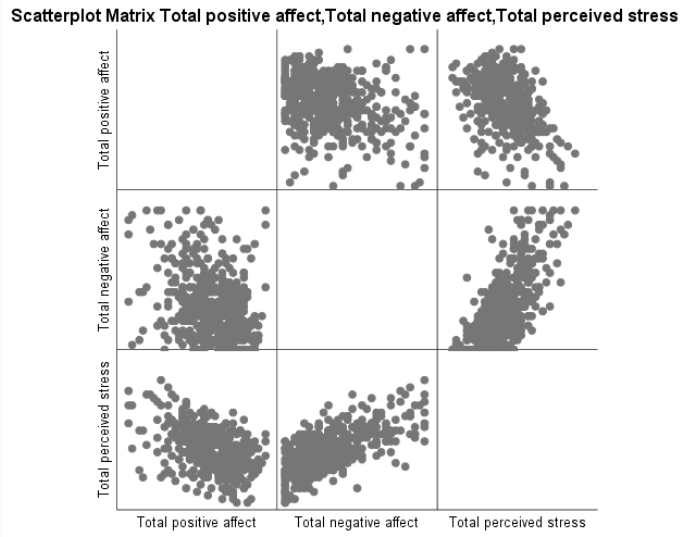
SCATTER MATRIX

Under *Options*, choose *Exclude variable-by-variable* to include as many data points as possible.



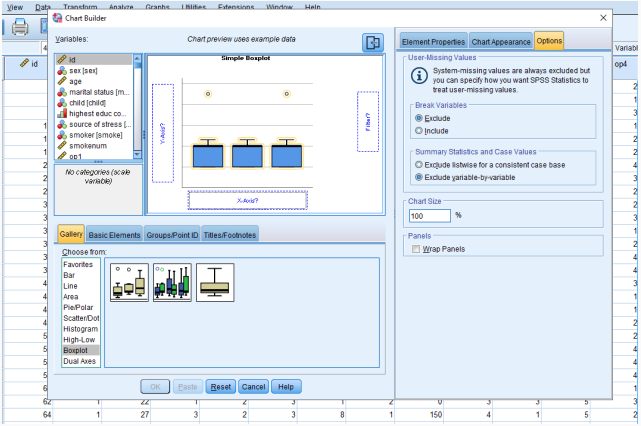
SCATTER MATRIX

Which variables are plotted is indicated by the axis labels. In the top right corner, *Total positive affect* is plotted against *Total perceived stress*. The graphs below the diagonal are the same as the graphs above it, only mirrored.



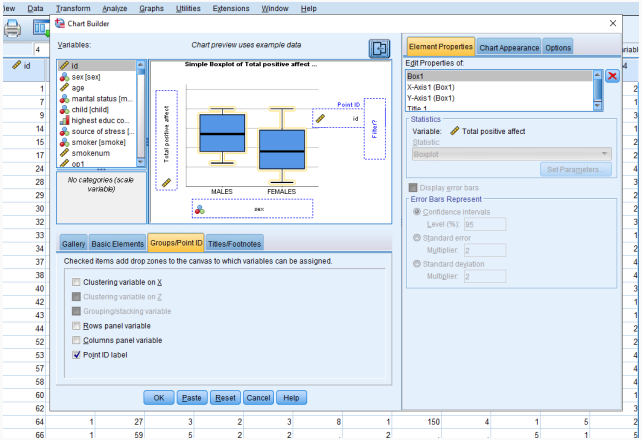
Box plot

Like the histogram, it shows the data distribution, but it also shows the median, interquartile range (IQR) as well as outliers.



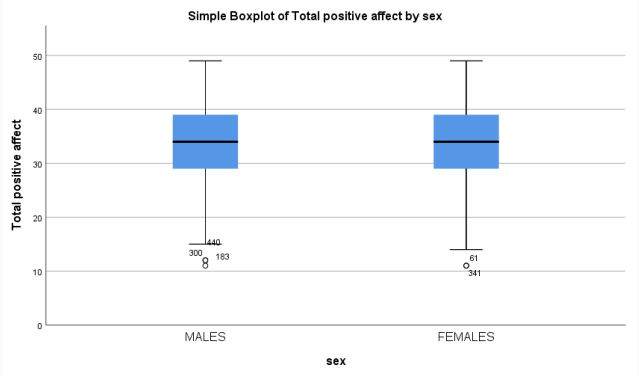
BOX PLOT

Useful to show similarities/differences between the distribution of a continuous variable within different groups/categories.



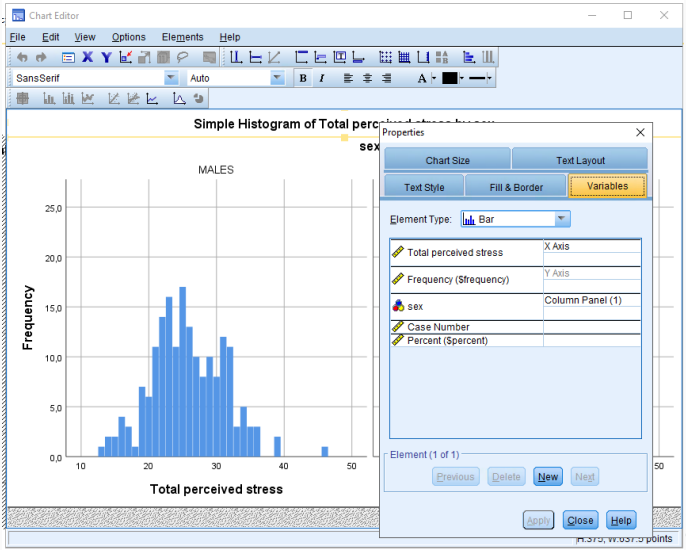
Box plot

Cases that are more than 1.5 times the IQR from the box edge, are marked by a circle, and are considered moderate outliers. Cases that are more than 3 times the IQR from the box edge are marked by a *, and are considered extreme outliers.



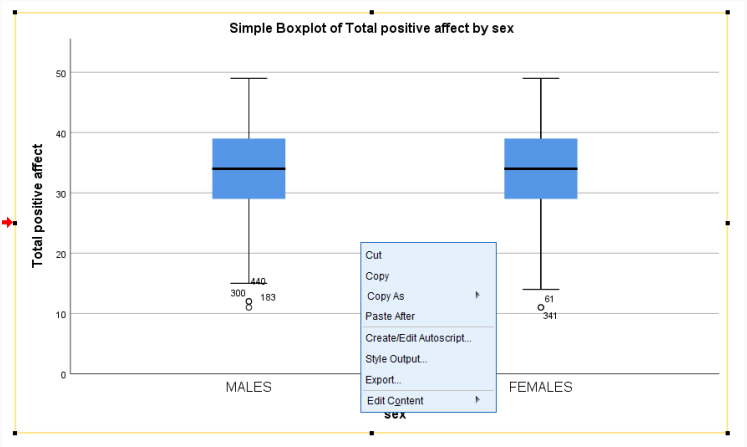
EDITING GRAPHS

Double click the graph in the Output window. To change font size, colours, direction of axis labels, width of the bars, change the y-axis, etc: Edit > Properties. To change direction of the bars, background grid, include text box, footnotes, etc: Options.



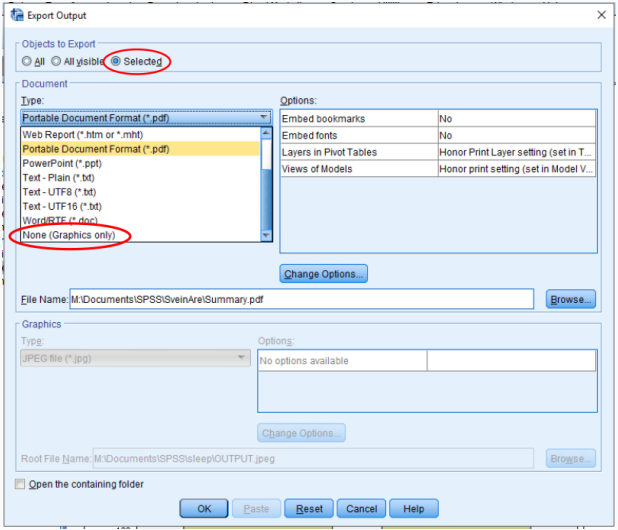
EXPORT GRAPHS AND TABLES FROM SPSS

Copy to Word document. Simply right click on the graph, and select copy to paste it directly into Word, or select *Export* to save as a separate file.



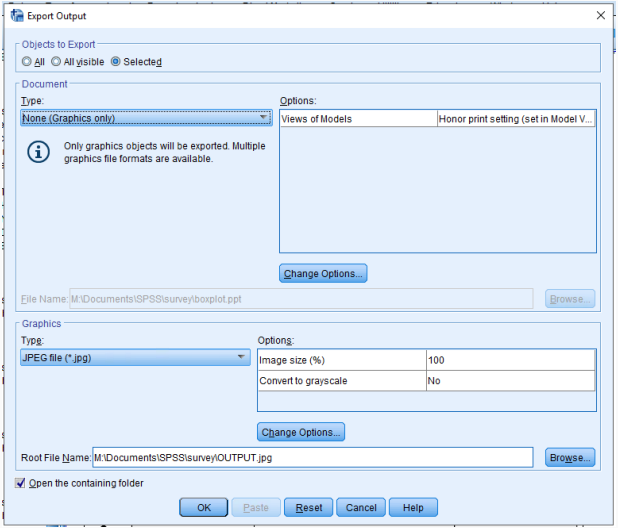
EXPORT GRAPHS AND TABLES FROM SPSS

In *Export Output* choose *Selected* and *Type: None (Graphics only)*.



EXPORT GRAPHS AND TABLES FROM SPSS

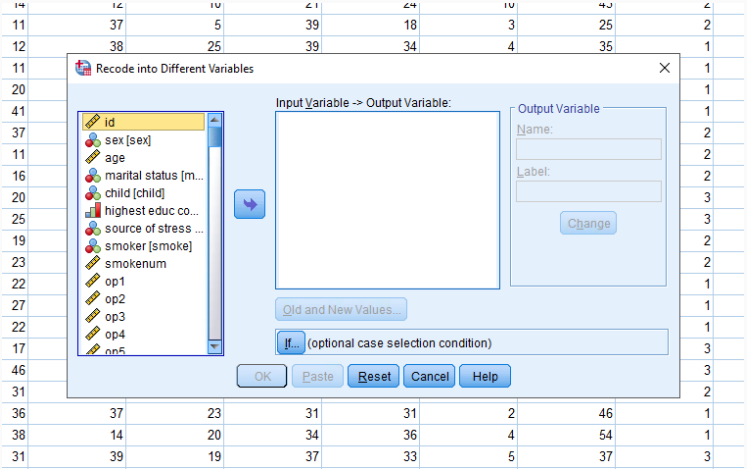
Choose the type of image file you prefer. You can export as .jpg/.png/.eps.



MANIPULATING DATA

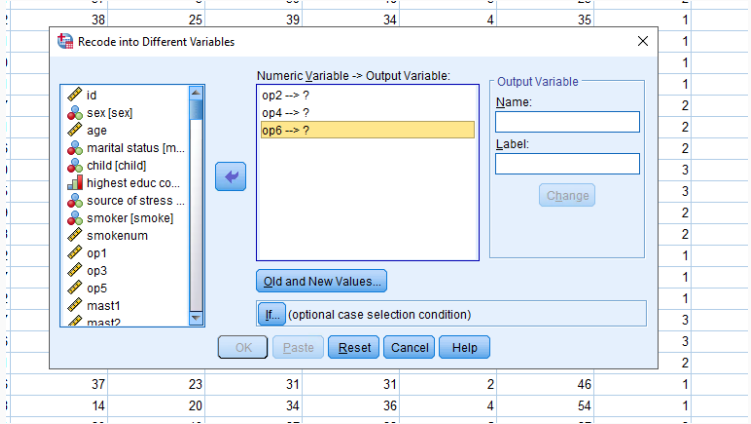
RECODE A VARIABLE

Transform > Recode into Different Variables: NEVER choose *Recode into Same Variable*, as you will lose the original information.



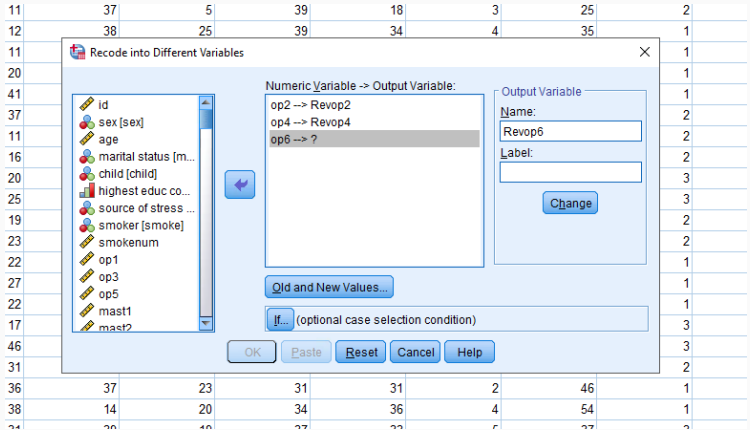
RECODE A VARIABLE

Add the variables you wish to recode by selecting them from the list on the left and clicking the blue arrow pointing to the right.

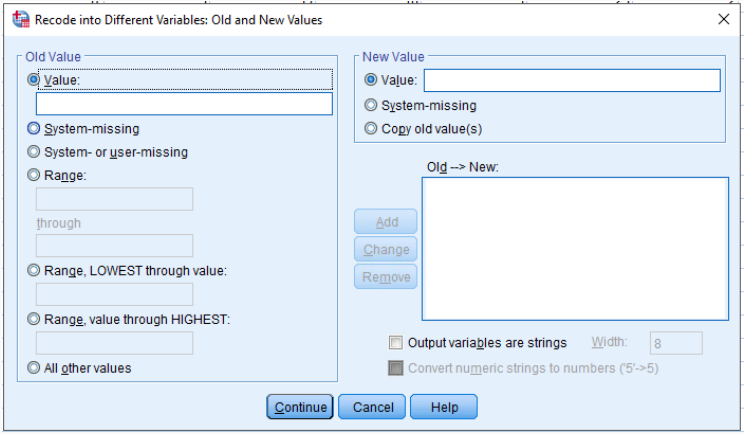


RECODE A VARIABLE

Give new names to the new versions of the variables, and click on *Change*. Click on *Old and New Values*.

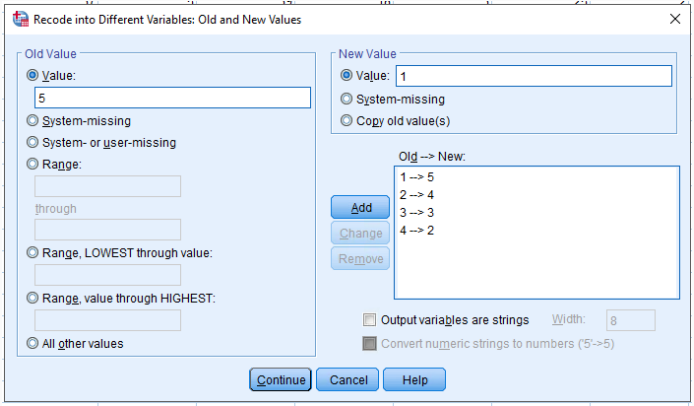


Recode into Different Variables: Old and New Values.



RECODE A VARIABLE

Type in the old value on the left hand side, and the new value on the right hand side, then click *Add*. The recoded values will appear as a list in the window *Old* $->$ *New*.



RECODE A VARIABLE

Check that the new variables are coded correctly, compared to the old ones.

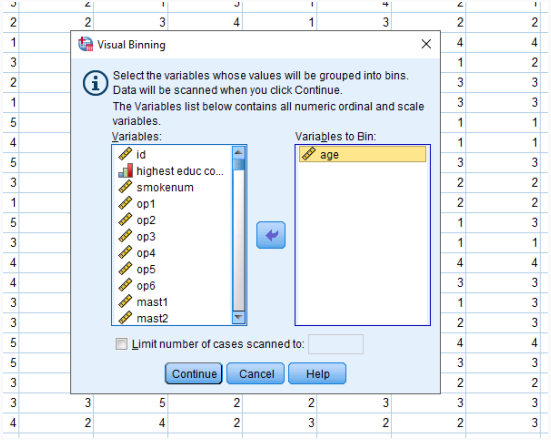
	_1	MAH_2	revop2	revop4	revop6
1	,00005	18,10064	4,00	4,00	4,00
2	,01585	14,48201	3,00	3,00	2,00
3	,01404	14,21429	5,00	3,00	2,00
4	,00540	14,16941	5,00	3,00	5,00
5	,00038	13,70297	4,00	4,00	3,00
6	,01361	13,66551	4,00	4,00	2,00

	mokenum	op1	op2	op3	op4	op5	op6
1	.	3	2	3	2	4	2
2	2	2	3	4	3	5	4
3	.	3	1	3	3	3	4
4	0	3	1	5	3	5	1
5	0	3	2	3	2	1	3
6	.	2	2	2	2	3	4

RECODE A VARIABLE: CONTINUOUS TO CATEGORICAL

How to recode a continuous variable into a categorical variable.

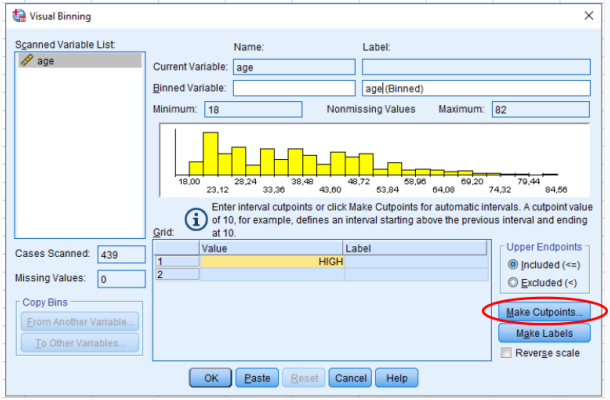
Transform > Visual Binning



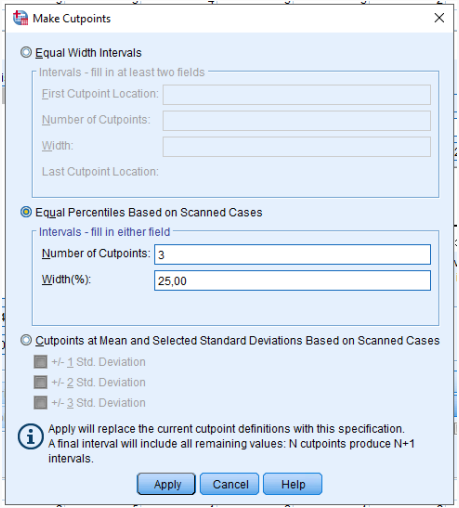
Move the variable you wish to recode into the window on the right hand side, by selecting it and pushing the blue arrow. Press continue.

RECODE A VARIABLE: CONTINUOUS TO CATEGORICAL

Here you can change the label of your new categorical variable if you wish. Press *Make Cutpoints* to indicate the number of categories, and how they should be divided.



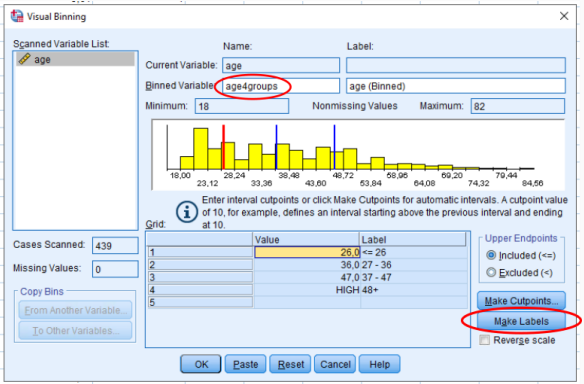
RECODE A VARIABLE: CONTINUOUS TO CATEGORICAL



Choose *Equal Width Intervals*, *Equal Percentiles Based on Scanned Cases* or *Cutpoints at Mean and Selected Standard...* Choose number of cutpoints. Remember that the number of categories equals the number of cutpoints plus one. With three cutpoints you get four categories.

RECODE A VARIABLE: CONTINUOUS TO CATEGORICAL

Press *Make Labels* to get labels for your categories. Give your new variable a name, and press *Paste*. Check that the new variable has values corresponding to the correct categories, by comparing with the old continuous variable.

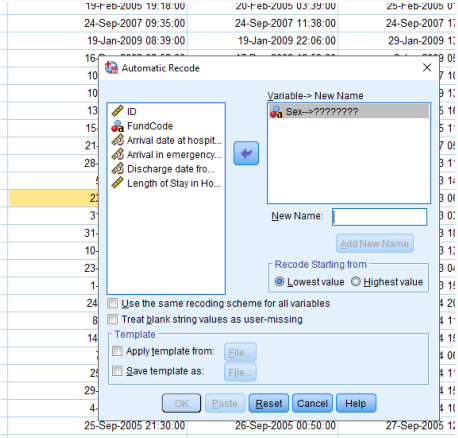


age4groups	age
1	24
3	39
4	48
3	41
1	23
2	31
2	30
1	23
1	18
1	23
2	27
2	34

Recode a variable: text to numerical variable

Automatic recoding from text to numbers. We will use the data set manipulate.sav

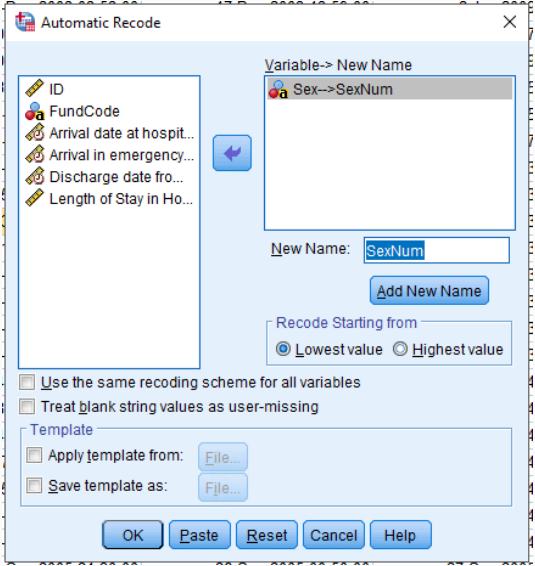
Transform > Automatic Recode



Choose the variable that should be recoded, and mover it to the window on the right hand side.

Recode a variable: text to numerical variable

Give the recoded variable a new name and click *Add New Name*. Choose if the recoding should start from the highest or lowest value. If there are missing values, it is useful to select *Treat blank string values as user missing*.



Recode a variable: text to numerical variable

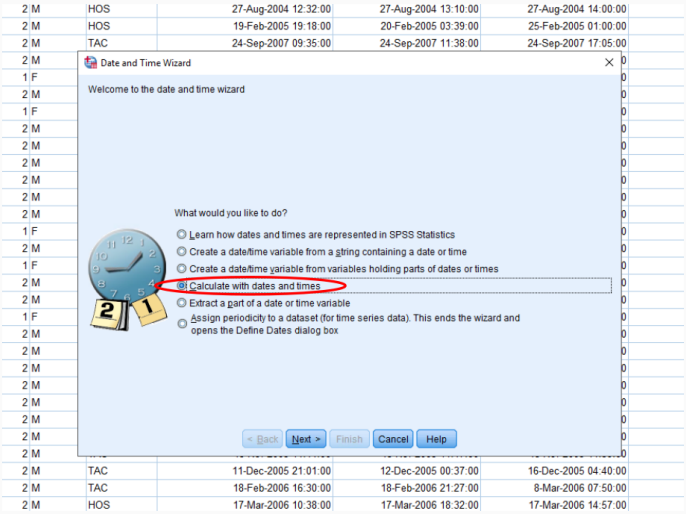
SexNum	Sex
2	M
2	M
2	M
2	M
2	M
1	F
2	M
1	F
2	M
2	M
2	M
2	M
2	M
2	M
2	M
1	F

Check that the recoding has worked properly. Here the values will be recoded according to their order in the alphabet, with *F* being coded as 1 and *M* as 2.

Recode a variable: Calculations with dates

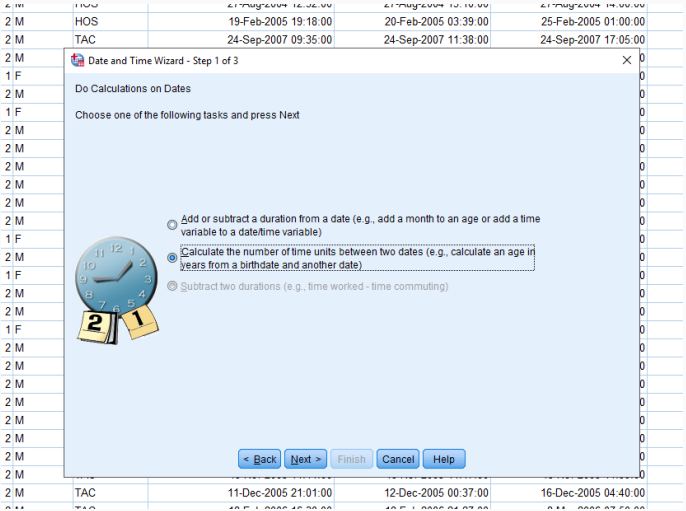
In order to work with date and time variables, we can use the *Date and Time Wizard*, and choose *Calculate with dates and times*.

Transform > Date and Time Wizard



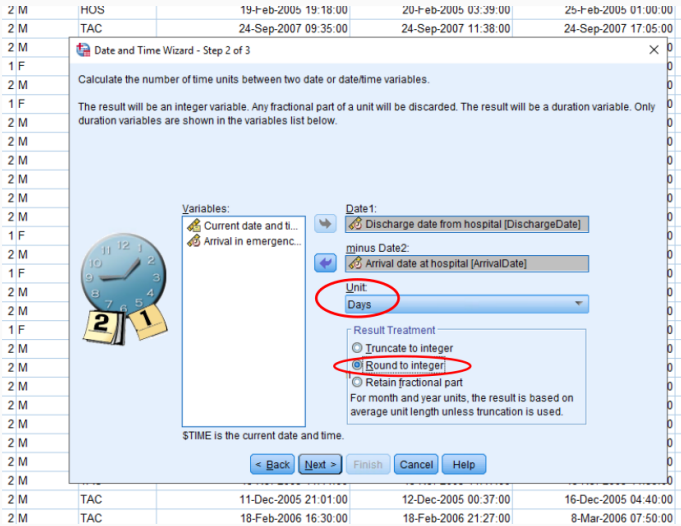
Recode a variable: Calculations with dates

Add or subtract a duration... can be used if you wish to calculate a new date based on one you already have/know. *Calculate the number of time units...* is used to find the amount of time between two dates.



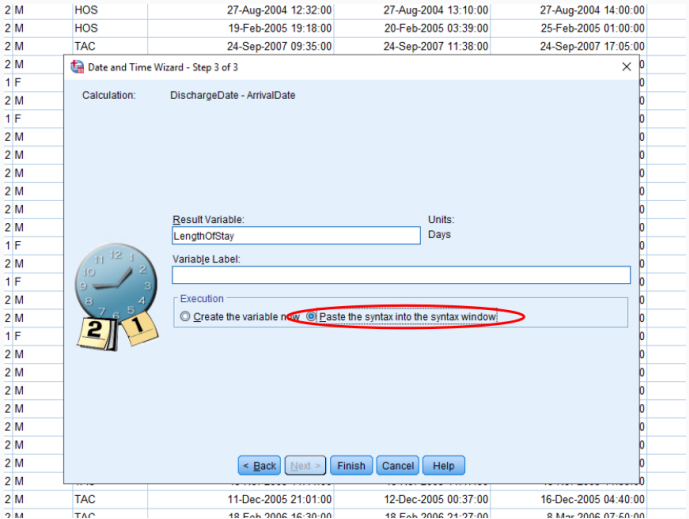
Recode a variable: Calculations with dates

Move the last date to *Date 1*. Move the first date to *Date 2*. This may seem counterintuitive, but the wizard will calculate Date 1-Date 2. Choose the appropriate time unit (e.g. days), and how you want to handle smaller units (e.g. round, truncate, decimal).



Recode a variable: Calculations with dates

Give the new variable a name, and if you wish to, a label. Choose *Paste the syntax into the syntax window*.



Recode a variable: Calculations with dates

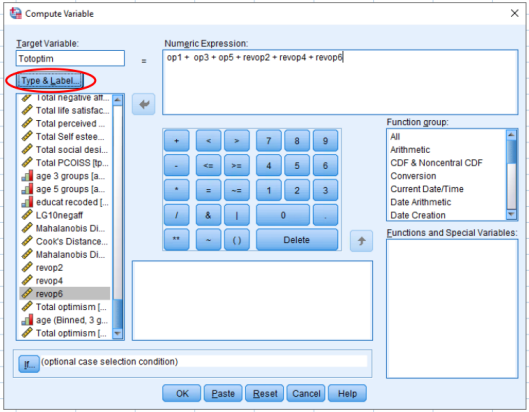
Check if the new variable seems to be correct. If you only get 0, you have probably just set the time unit to years or similar. If you get negative numbers, you have mixed up Date 1 and Date 2.

ArrivalDate	EmergDate	DischargeDate	LengthofStay
10-Mar-2005 03:04:00	10-Mar-2005 05:09:00	16-Mar-2005 13:00:00	6
27-Aug-2004 12:32:00	27-Aug-2004 13:10:00	27-Aug-2004 14:00:00	0
19-Feb-2005 19:18:00	20-Feb-2005 03:39:00	25-Feb-2005 01:00:00	5
24-Sep-2007 09:35:00	24-Sep-2007 11:38:00	24-Sep-2007 17:05:00	0
19-Jan-2009 08:39:00	19-Jan-2009 22:06:00	29-Jan-2009 13:01:00	10
16-Dec-2008 08:58:00	17-Dec-2008 18:59:00	2-Jan-2009 05:45:00	17
10-Jan-2007 15:00:00	10-Jan-2007 18:22:00	13-Jan-2007 10:00:00	3
10-Jun-2009 13:56:00	11-Jun-2009 13:05:00	11-Jun-2009 13:05:00	1
13-Jun-2005 15:10:00	13-Jun-2005 22:28:00	18-Jun-2005 16:55:00	5
15-Mar-2005 14:22:00	15-Mar-2005 16:53:00	23-Mar-2005 11:56:00	8
21-Feb-2007 06:20:00	22-Feb-2007 00:58:00	3-Mar-2007 05:00:00	10
28-Dec-2003 07:43:00	28-Dec-2003 09:25:00	31-Dec-2003 11:44:00	3
5-Jul-2003 08:13:00	5-Jul-2003 08:25:00	6-Jul-2003 14:45:00	1
23-Jul-2003 07:13:00	23-Jul-2003 07:40:00	29-Jul-2003 06:00:00	6
31-Jul-2003 05:06:00	31-Jul-2003 09:40:00	9-Aug-2003 03:30:00	9
31-Aug-2003 03:54:00	31-Aug-2003 05:20:00	31-Aug-2003 18:45:00	1
10-Sep-2003 01:25:00	10-Sep-2003 04:12:00	12-Sep-2003 13:45:00	3
23-Sep-2003 22:33:00	23-Sep-2003 23:30:00	24-Sep-2003 04:35:00	0

CALCULATE NEW VARIABLES

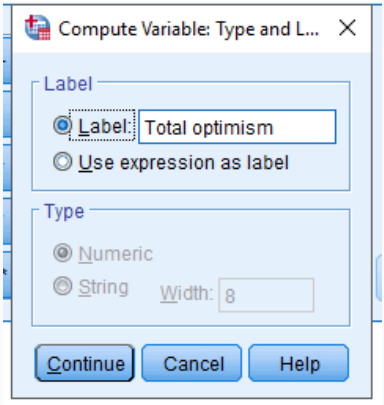
How to create a new variable from existing variables

Transform > Compute Variable



Give a new name to the new variable, and fill in the mathematical expression in the box *Numeric Expression*, using the calculator, the list of existing functions and the the list of existing variables.

Click *Type and Label* to provide a label for the new variable.



Check that the new variable looks correct.

op1	op2	op3	op4	op5	op6	revop2	revop4	revop6	Totoptim
3	2	3	2	4	2	4,00	4,00	4,00	22,00
2	3	4	3	5	4	3,00	3,00	2,00	19,00
3	1	3	3	3	4	5,00	3,00	2,00	19,00
3	1	5	3	5	1	5,00	3,00	5,00	26,00
3	2	3	2	1	3	4,00	4,00	3,00	18,00
2	2	2	2	3	4	4,00	4,00	2,00	17,00
3	5	1	4	1	4	1,00	2,00	2,00	10,00
4	1	3	1	4	2	5,00	5,00	4,00	25,00
3	4	2	4	2	4	2,00	2,00	2,00	13,00
1	3	1	4	1	5	3,00	2,00	1,00	9,00
5	1	5	1	5	1	5,00	5,00	5,00	30,00
4	1	4	1	5	1	5,00	5,00	5,00	28,00
2	4	5	5	2	2	2,00	1,00	4,00	16,00
4	3	3	3	4	3	3,00	3,00	3,00	20,00

Outliers?

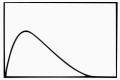
Are some cases strange, or very far away from the rest of the data?

There are generally four main reasons for outliers:

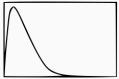
1. Wrong coding
2. Wrong code for missing data, i.e. missing data is being treated as real data
3. The case belongs to a different population than the one you meant to sample
4. **Greater dispersion/spread in the data than expected, and deviation from the normal distribution**

Transformation of 'abnormal' data

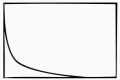
My data are so far from a normal distribution, what do I do?



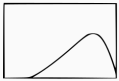
Kvadratrot
Formel: ny variabel = SQRT(gammel variabel)



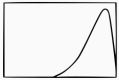
Logaritme
Formel: ny variabel = LG10(gammel variabel)



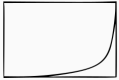
Invers
Formel: ny variabel = 1/(gammel variabel)



Speilet og kvadratrot
Formel: ny variabel = SQRT(K-gammel variabel)
K = største mulige verdi til variabel + 1



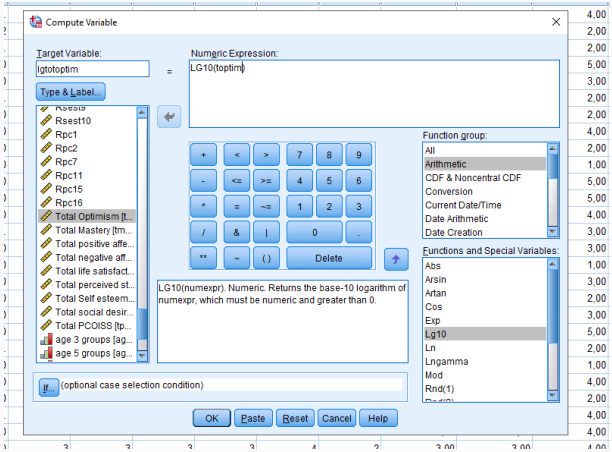
Speilet og logaritme
Formel: ny variabel = LG10(K-gammel variabel)
K = største mulige verdi til variabel + 1



Speilet og invers
Formel: ny variabel = 1/(K-gammel variabel)
K = største mulige verdi til variabel + 1

TRANSFORMATION OF 'ABNORMAL' DATA

Use the same method as we used to calculate a new variable. Use the mathematical expression that best fits with the distribution you already have.



Back to the study hall/office

Start using SPSS as quickly as possible on your own data (or someone else's for that matter)!

The only way to improve your understanding of statistics and SPSS, is to use it. Learning by doing.

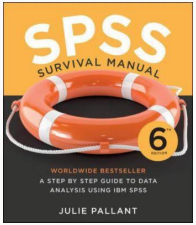
Make sure you have good textbooks and online resources when you need to check something.

A decent online tutorial can be found e.g. at https://libguides.library.kent.edu/SPSS

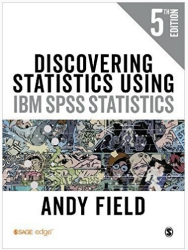
Ask google or youtube when you get stuck. If that doesn't help, ask us (statistikk@usit.uio.no) or attend our intermediate SPSS course, 27th - 29th of September

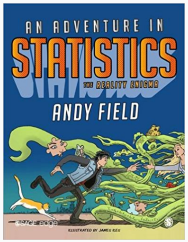
Suggested books

In SPSS: SPSS Survival manual by Julie Pallant



In statistics and SPSS: *Discovering statistics using IBM SPSS* and *An adventure in statistics* by Andy Field





Commercial break

Are you struggling with large, unwieldy data sets? Does your computer lag and struggle when you try to analyze them using SPSS? We have an introductory course in R/RStudio!

Benefits: R and RStudio is free, it can do more than SPSS can, without the need for a powerful computer. And also, that graphs look much better and more professional :D