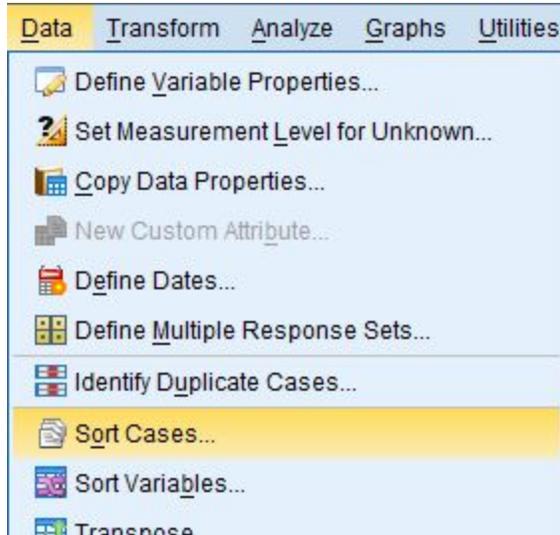# Data Transformation in SPSS

-Bijay Lal Pradhan

Sorting Data

Often you want the individuals/cases in a dataset to follow a certain logical sequence. Data can be sorted ascending, with the lowest values first, or descending, with the highest values first. You can sort the data in different ways, see below.

Use the Menu bar and choose Data >> Sort Cases This option enables you to sort by more than one variable. If you sort by e.g. gender first, and then by height, you will see one gender at the top, sorted by their heights, and at the bottom you will see the other gender, sorted by their heights.

Right click on the variable's name and choose Sort Ascending or Sort Descending.
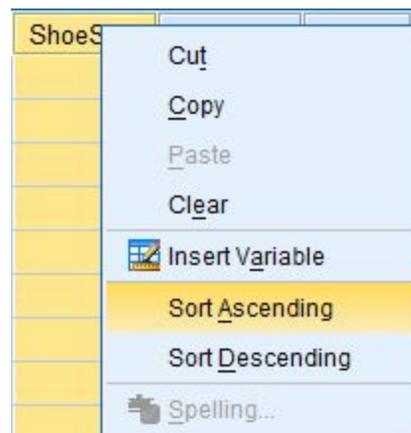
Creating new variables

Quite often we want to use existing variables to create new variables, e.g. when we want to use numbers instead of text for an ordered variable, or when you want to create classes or intervals for a continuous variable. This can be done in different ways;

The new variable is displayed in the Data Editor. Since the variable is added to the end of the file, it is displayed in the far right column in Data View and in the last row in Variable View.

Recode variables (e.g. creating classes or intervals, recoding text into numerical values, or creating dummy variables)

If you want to create classes or intervals you can choose Transform >> Recode into Different Variables from the Menu bar. The following dialog window will appear.
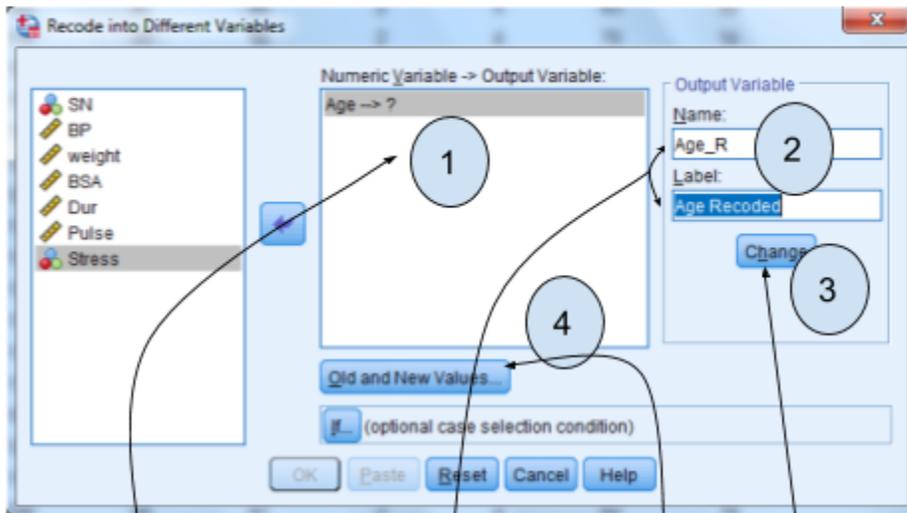
1. Choose which input variable you want to use, to create the new variable from (Age is being used in this example)
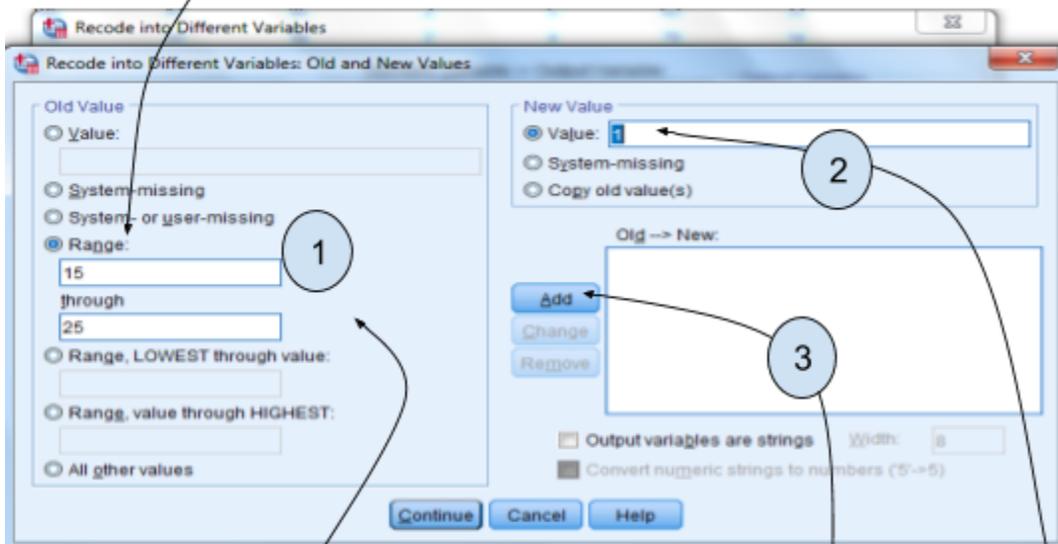
2. Type a name of the variable you are about to create and then label it.

4) Click "Old and New Values"

3) Click on Change

A new dialog window will appear,

1a) To e.g. recode a text variable into a numerical variable, select "Value" and type the value you want to recode.

1b) To create one of the intervals, type the range of the interval. Remember to use proper limits for the intervals/classes! Other alternatives can be chosen here, such as individual values (not creating intervals), missing values, etc. "Range, LOWEST through value:" creates an interval from the lowest existing value through the value that you type. "Range, value through HIGHEST:" creates an interval from the value that you type through the highest existing value.

3) Click "Add"

2) Type the value that you want the new variable to get. If you want the new variable to be a text variable, select "Output variables are strings".

Repeat steps 1 through 3 to add all values/intervals. Then click "Continue", and finally click "OK". Make sure to visually check that your new variable contains the values that you intended. To make this check easier you might want to place the new variable next to the original one.
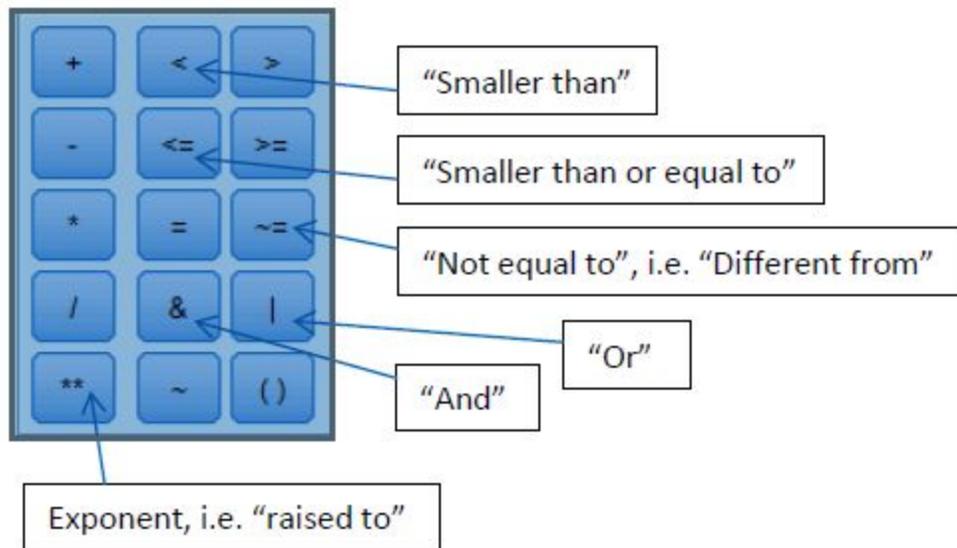
Calculate variables
Sometimes you want to create a new variable that is a function of another variable (or several other variables). Then you can choose Transform >> Compute Variable from the Menu bar. The dialog window below will appear.
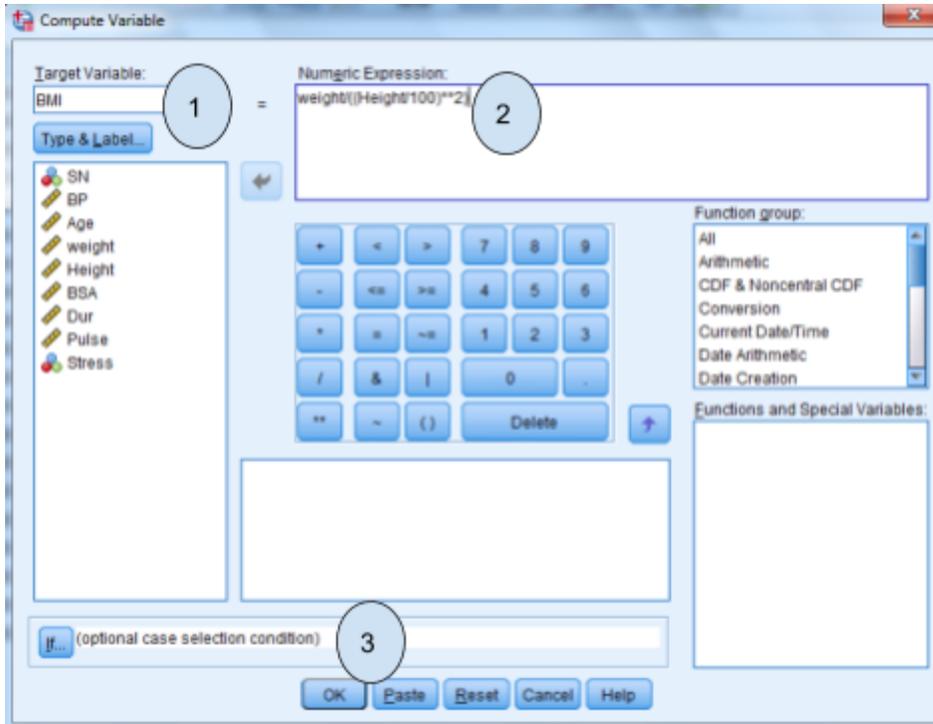
1) Type a name of the variable you are about to create.

2) Type the numeric expression. To e.g. calculate BMI = $Weight\ (kg)/Height\ (m)^2$, first choose variable "Weight" among the existing variables. Then type a division mark ("slash"), or click on the corresponding blue button and complete the expression.

3) Click "Ok" .

It will create a new variable BMI.

Numerical operations

When creating new variables we may want to use mathematical expressions of different kinds. It is very important to use correct expressions, and SPSS has some built in numerical operations that can be used. The most common ones are explained below.

"Smaller than"

"Smaller than or equal to"

"Not equal to", i.e. "Different from"

"Or"

"And"

Exponent, i.e. "raised to"

"if" function
When creating new variables, either by recoding or calculating them, we might want to perform the particular operation only for a subset of the individuals (e.g. for males only). Then we can use the "if" function.

Make calculations for selected individuals
Sometimes we want to perform the analyses only for a certain number of individuals/cases, or you want to perform the analyses separately for different groups (e.g. by gender).
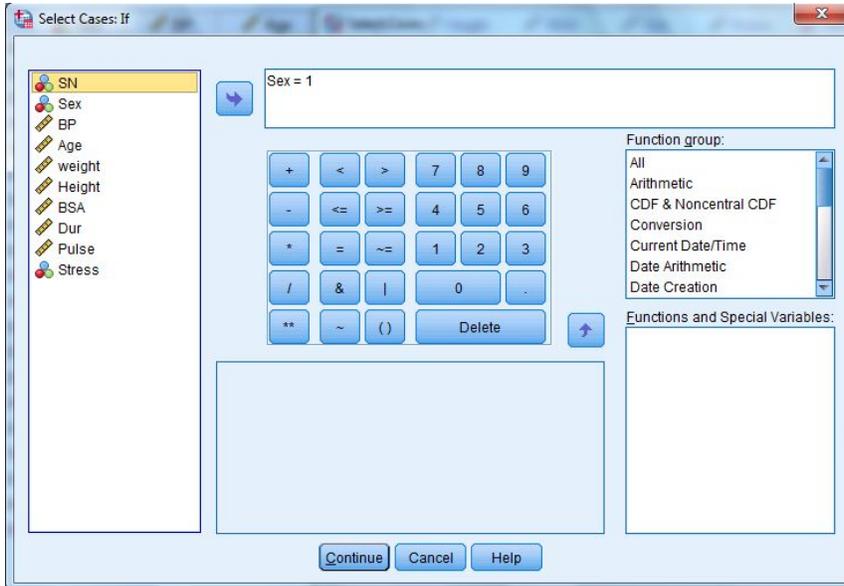
Make a subset of data (Select Cases)
Choose Data >> Select Cases from the Menu bar. This will open the dialog window below.
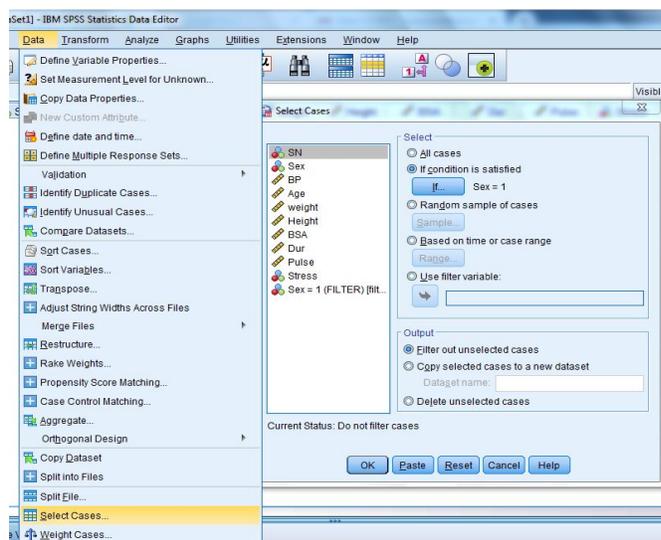
1) Select "If condition is satisfied"
2) Click "If…"
The dialog window below will appear.

1) Type the condition (numeric expression) in the white field. Individuals who fulfill the condition will be selected [in this example only individuals Male (code for male was 1) will be selected]. Different numerical conditions can be set up.

2) Click "Continue"

This will bring you back to the previous dialog window, then Click "Ok".
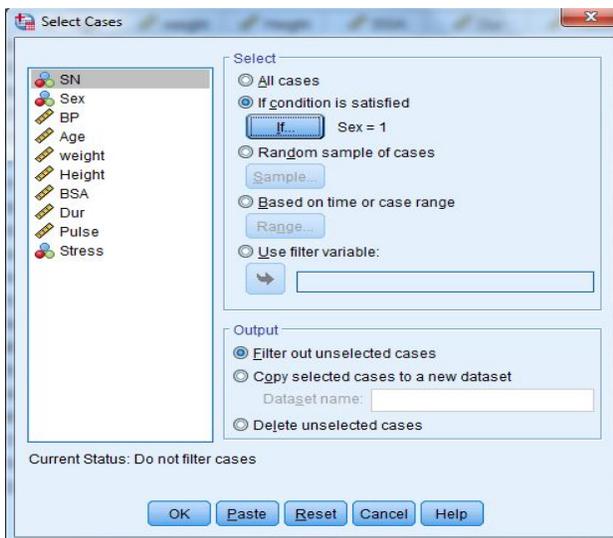
Choose what you want to do with the selected cases/individuals. Filter out unselected cases: A new variable named filter_$ will be created, where the value 1 denotes selected cases/individuals and 0 denotes unselected cases/individuals. Unselected cases will be marked in the Data Editor with a diagonal line through the row number.

Copy selected cases to a new dataset: A new dataset will be created, which contains only selected cases/individuals .

Delete unselected cases: Selected cases/individuals will be deleted.

IMPORTANT! Make sure to save the dataset under a new name! Deleted cases can be recovered only by exiting from the file without saving any changes and then reopening the file. The deletion of cases is permanent if you save the changes to the data file.

		Sex	filter_$	BP	Age	weight	
1	0	1.00	1	105	47	85	
2	0	1.00	1	115	49	94	
3	0	2.00	0	116	49	95	
4	0	1.00	1	117	50	95	
5	0	1.00	1	112	51	89	
6	0	1.00	1	121	48	100	
7	0	2.00	0	121	49	100	
8	0	2.00	0	110	47	91	
9	0	2.00	0	110	49	89	
10	0	2.00	0	114	48	93	
11	0	2.00	0	114	47	94	
12	0	1.00	1	115	49	94	
13	0	1.00	1	114	50	92	
14	0	2.00	0	106	45	87	
15	0	2.00	0	125	52	101	
16	0	1.00	1	114	46	95	
17	0	1.00	1	106	46	87	
18	0	1.00	1	113	46	95	
19	0	1.00	1	110	48	91	
20	0	2.00	0	122	56	96	

Automatic Recode

When writing down the observed values of a categorical variable, we can choose to write the values as words or as numeric codes. Either method of recording categorical variables is valid, but it is often easier to work with numeric codes in SPSS than it is to work with strings. This is because, when referring to the content of a string during a computation, the content must match exactly. If the content of the two strings is not an exact match, the computer will not recognize them as identical. This includes placing an extra space at the end of a string: the human eye won't detect the discrepancy, but the computer will. (Note that if your data was originally recorded in Excel, it is very easy for the values of string variables to accidentally be recorded with extra spaces at the end.)

If we have already recorded your categorical variables as strings, you can easily convert them to a numerically coded variable using the Automatic Recode procedure. This procedure assigns each unique category a numeric code, then saves the converted values as a new variable. It also automatically adds value labels: whatever the string value was before becomes the value label.

Choose Transform >> Select Automatic Recode from the Menu bar. This will open the dialog window below.
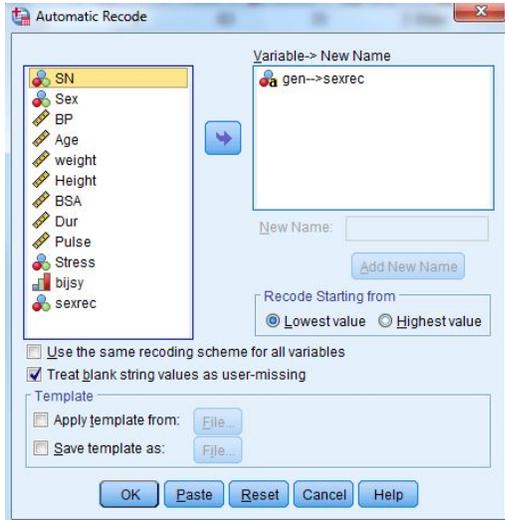
Additionally, if we have used blanks to indicate missing values for string variables, we may have noticed that SPSS doesn't automatically recognize those observations as missing. This is because SPSS, by default, recognizes "blank" strings as valid values. In this situation, we must use Automatic Recode in order for SPSS to recognize blank strings as missing values. Otherwise, SPSS will consider the "blank" category as a valid category.
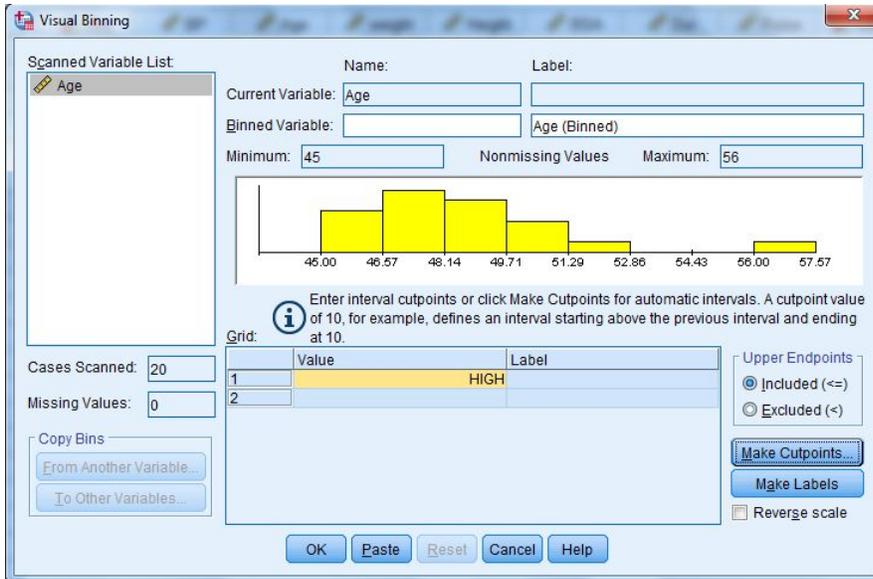
Visual Binning:

This function will create groups (bins, categories) from a continuous variable and visually control the process.

Choose Transform >> Select Visual Binning from the Menu bar. This will open the dialog window below.
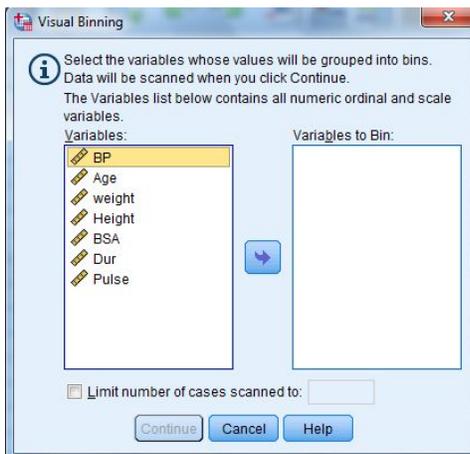
In the first dialogue specify the variable(s) you want to recode ("bin"). Then the following dialog comes up

- Specify a name for the Binned variable; SPSS already supplies a default Label, but you can of course change it.
- You now could enter (Grid) cutpoints, i.e. the (upper) limits that separate two intervals (bins) and the corresponding label. (HIGH means the highest value, in our case a value of 100.0).
- If you want to split the variable regularly, use Make Cutpoints
- Equal Width intervals defines evenly spaced intervals. We have to fill either the width or the number of intervals.
- Equal percentiles on scanned cases defines group/ classes with roughly the same number of observations. We have to either the number of cutpoints or the % width by entering the percentage of cases in each group. For example if you indicate 3 cutpoints or 25% of the cases, you will create four groups containing each one fourth of the observations.
- Cutpoints at mean/standard deviations: Define a cutpoint at the mean and at ± 1,2 or 3 standard deviations.
- Then Apply