CRJ 716 – Using Computers in Social Research

Modifying Data with SPSS

Lecture 5: Recoding Variables

Prof. Agron Kaci
John Jay College

Modifying Data with SPSS

In many cases in social research, you may need to modify given data so that you have a new perspective on what is going on. There are categorical variables that may have too many groups, and so the analysis becomes cumbersome. It is here that you may need to group together some values in order to answer your question. There are other cases where many categories simply do not contain significant values, as to produce a meaningful result. Again, you will need to "collapse" some categories in one to get a significant number of values. This procedure of transforming variables is called "recoding".

There are other ways to transform variables, such as *Compute*, *If*, and *Weight*, but these will be covered in the next lecture.

Recoding Variables

Recoding into Different Variables

Recoding into Same Variables

Recoding is used to merge categories of a variable into fewer groups or into different categories. A classical example is organizing age values into collapsed categories, such as "Young", "Middle aged", and "Older". We already know that age is a ratio variable, and its values are raw data, not categories. How can we, then, collapse these raw values into the three categories we set out? IBM SPSS has a command, called *Recode* that can be used to recode the data.

There are three general rules you may need to know regarding modifying data with *Recode*.

First, you need to choose cutting points at which to divide the categories. In the case of age, for example, which age value should start the "Middle age" category? Is it 30, 35, 40? Generally, start dividing categories at logical points. 18 years of age, for example is the traditional threshold of voting rights, 21 of adulthood. Sometimes, the categories will be divided arbitrarily, as done below.

The second rule is empirical, and that means that you may want to recode and combine small categories (those groups that have very few values) into bigger ones. This is done because few values in categories may result in insignificant analyses, therefore your report may suffer.

The third rule can be summarized in the need for type of analysis or test that you wish to conduct in SPSS. It has to do with the number of categories you want to utilize in your test. If, for example, you want to create a crosstabulation or a contingency table, you may want to limit the number of new categories to three or four. Or, in the case of dichotomous variables, two new categories need to be created.

Getting back to our variable "age", you may want to recode as following:

Age (in years)	The new, recoded categories
18 through 35 years old	Young
36 through 55 years old	Middle age
56 of age or over	Older

This will be the basis for our recode exercise in this lecture.

Let's launch the 2004GSS.sav file, to start the process of recoding the respondent's age into the three categories we set out above.

1. Recoding Into Different Variables

In the Data Editor view, follow the command:

Transform → **Recode into Different Variables**

The process should image Figure 5-1 below:

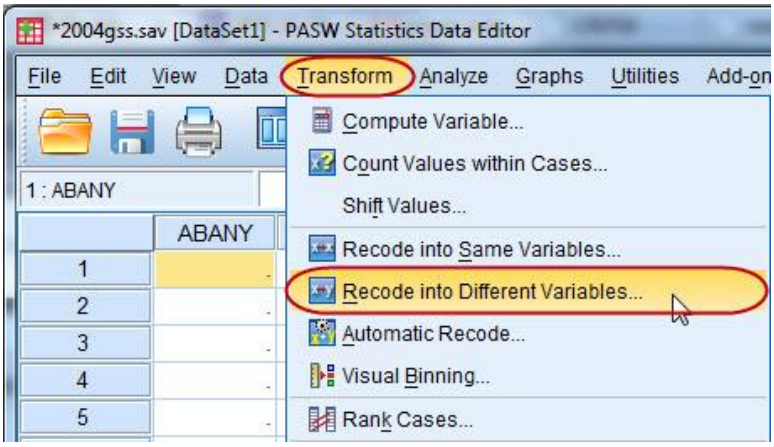


Figure 5- 1

As seen in the figure above, the *Transform* menu has another option, where you can "Recode into Same Variable". This allows you to change the original variable. If you use this command and save the file, and you did not have another copy of the file, you have just transformed the original variable(s) with no chance of recovering it. Briefly, this command will be covered towards the end of this lecture.

Once you click on the *Recode into Different Variables* option, a new dialog box opens that looks like figure 5.2 below.

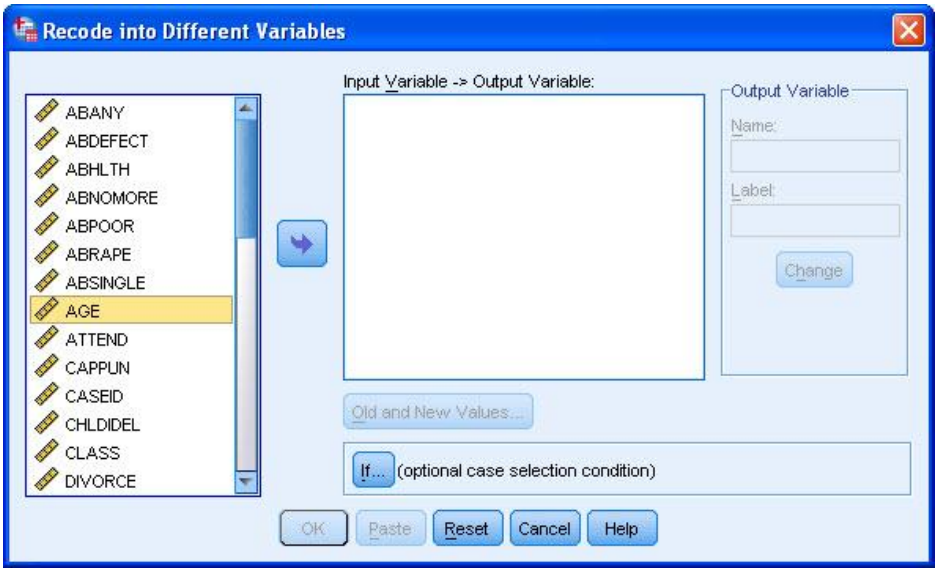


Figure 5- 2

Find and transfer the variable **AGE** into the *Input Variable…* field. Double-click the variable or click to select the variable and click the arrow button to transfer the variable appropriately.

The *Output Variable* section on the right allows you to create and enter a new name for the variable that will be created. Generally, enter a name that is associated with the original variable. For this exercise we will enter **AGE1** in the *Name* field. It is also advised that you enter a variable label now. You can leave this field empty for the time being, and still can enter a variable label in the Variable View editor later, after you create the new variable. However, if you wish to add it now, type *Age Recoded into Three Categories* as your label. Click on the *Change* button and observe that **AGE - -> AGE1** is entered in the big box in the middle, just like in Figure 5-3.

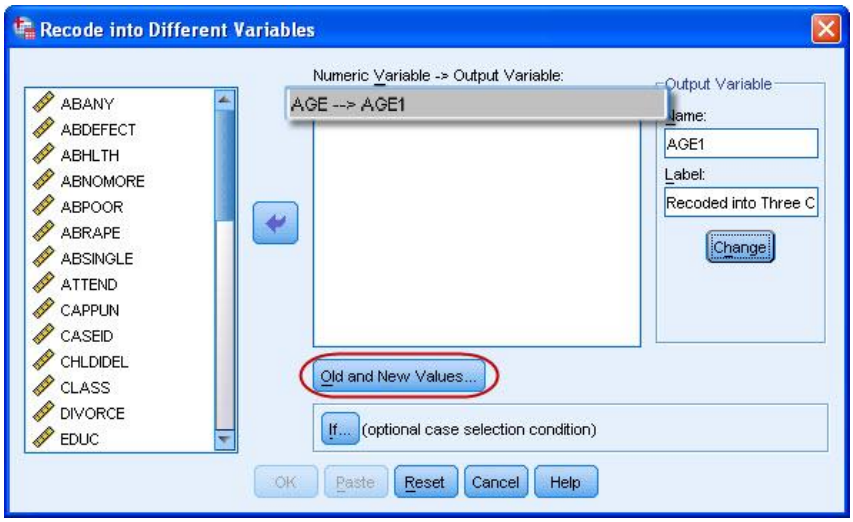


Figure 5- 3

Now, it's the time to create the new values based upon the old ones, so we need to tell SPSS which old values will create the new values. Click on the *Old and New Values* button to invoke the *Recode into Different Variables: Old and New Values*.

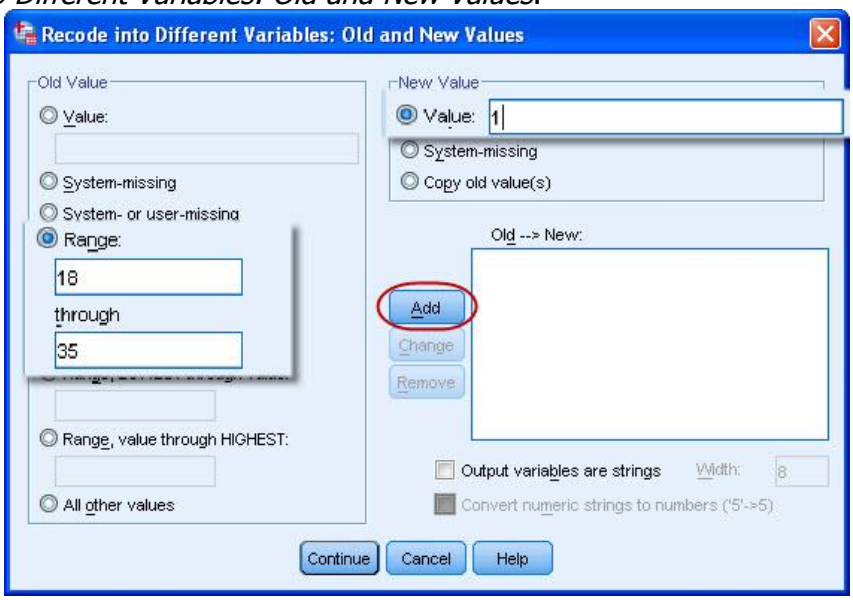


Figure 5- 4: **Shows a new value (1) created by a range of old values (18-35)**

This new window has two sides. The left side contains information about the old values, which are the original values and exist in the data set presently. The right side will contain the new values that we will create out of old values.

Note: The process of assigning new values (recoding process) needs some preliminary work done. This involves the gathering of all old values and a sketch of how will you categorize them. Another very important step is to have knowledge of missing values in the original variable.

The process of recoding entails typing of the old and new values in the respective boxes. It is here where we will practically change old values to new values.

The first new value that we set out to create is "Young". This is actually the label of the value (value label), not the actual value. Values in SPSS are entered as numbers. So, the new values that will be entered for the new variable **AGE1** will be 1, 2, and 3, that correspond to "Young", "Middle-aged", and "Old", respectively.

New Value	New Value *Label*	Old Values
1	Young	18-35
2	Middle Aged	36-55
3	Old	56-up

As we can see, all our new three categories are made up of ranges of old values; therefore we will need to use the "Range" options in the left side.

Select "Range" and type "**18**", which is the lowest value, and in the field below it type "**35**". See figure 5-4 above. Then click to activate the *Value* field in the *New Value* section and type "**1**", which will correspond to the "**Young**" age group. Then click on *Add* to add the new value in the *Old - -> New* box.

The same procedure should be followed for other groups. Type in **36** and **55** in the respective fields under *Range*. Enter "**2**" in the *New Value* and click *Add*. Type **56** and **89** (the highest value in the data set-you knew this because of the investigative work you've done before, right?), and give this a new value of "**3**". If your screen looks like Figure 5-5, you're in the right track.

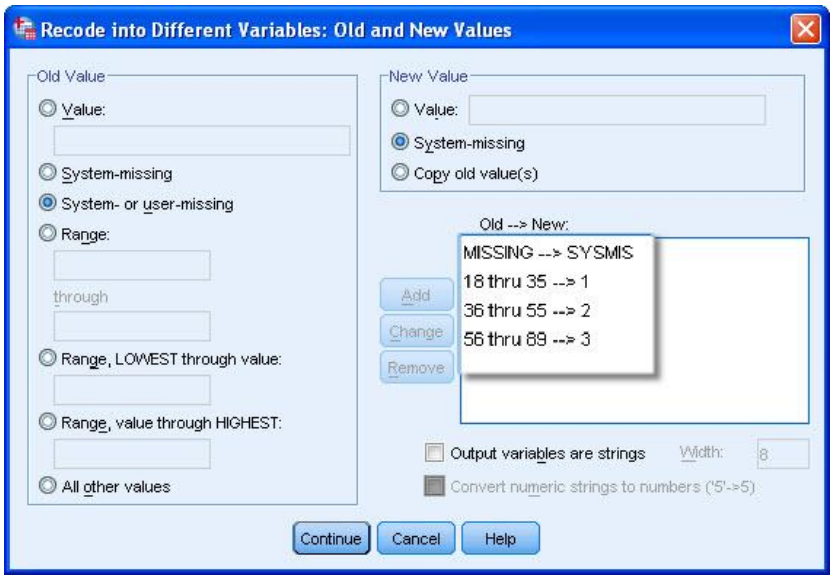


Figure 5- 5

If you made a mistake and you want to change one of your categories, click to select that category in the *Old --> New* box (step 1 in the Figure 5-6 below), make the changes in the Old value section (step 2) and then click on *Change* (step 3). If you need to delete a category, select it and click *Remove*.

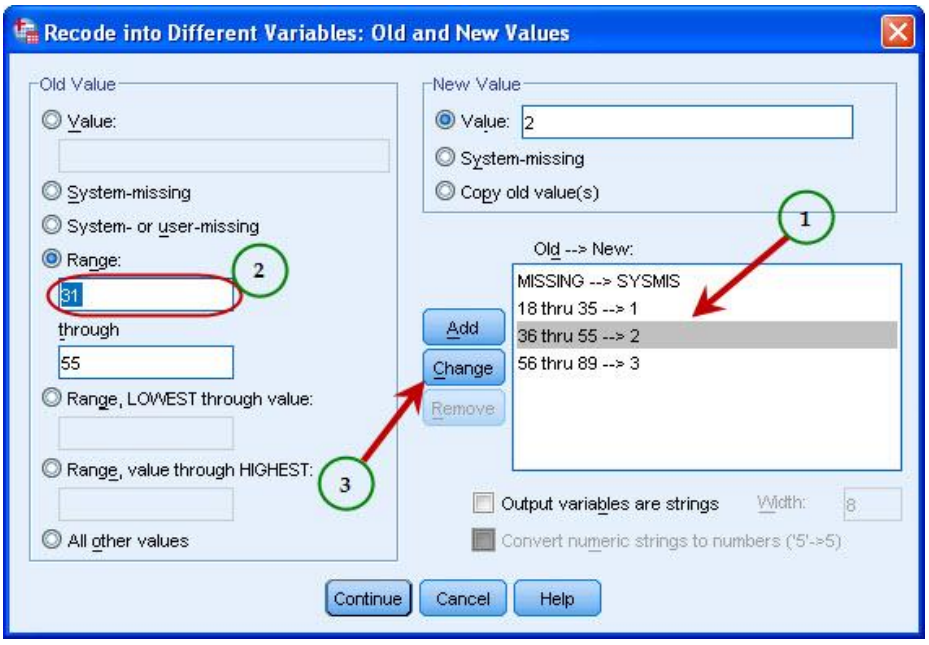


Figure 5-6

When you are ready, click *Continue*. Click **OK** in the *Recode into Different Variables* dialog box. The Data Editor window appears on the screen.

Now, we need to produce a simple frequency table to find out about our three categories. Follow the steps as shown below:

Analyze → **Descriptive Statistics** → **Frequencies**

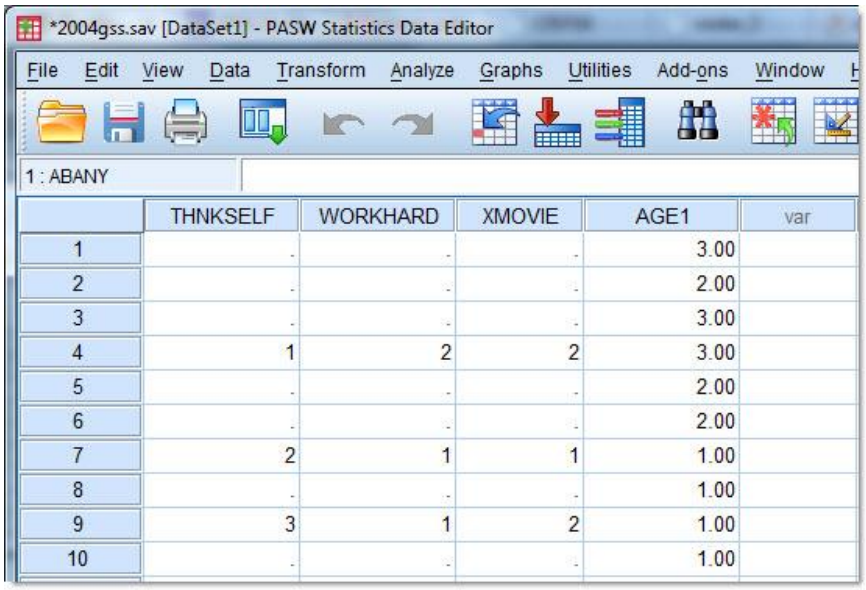
Observe that now you have a new variable, **AGE1** that appears in the list of variables on the left. Select it and move it to the *Variables* box, and click **OK**. The Output window will show the frequency table as exposed below in table 5-1.

Recoded into Three Categories

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	403	33.6	33.7	33.7
	2.00	478	39.8	40.0	73.7
	3.00	314	26.2	26.3	100.0
	Total	1195	99.6	100.0	
Missing	System	5	.4		
Total		1200	100.0		

Table 5- 1

Let's take a look at the data table in the Data Editor window in SPSS. Go to the **Data View** window, by clicking on the tab located in the lower-left corner of the window. Find the new variable, **AGE1**. Usually, new variables added to the data set are found on the rightmost side of the table, so scroll to the right until you see it. See Figure 5-7.

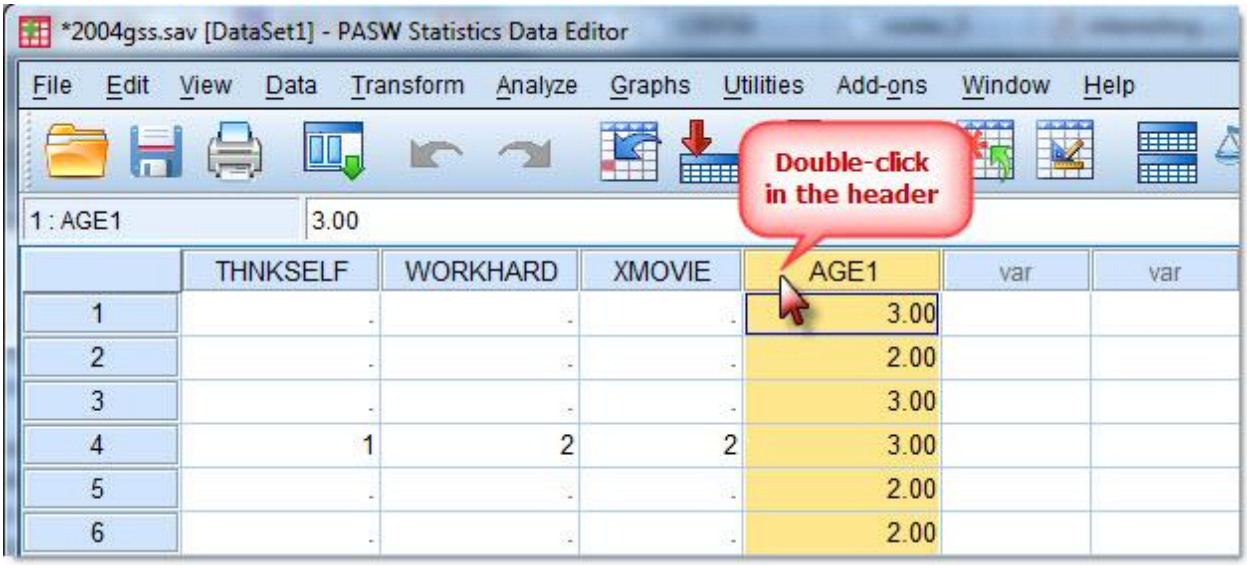


	THNKSELF	WORKHARD	XMOVIE	AGE1	var
1	.	.	.	3.00	
2	.	.	.	2.00	
3	.	.	.	3.00	
4	1	2	2	3.00	
5	.	.	.	2.00	
6	.	.	.	2.00	
7	2	1	1	1.00	
8	.	.	.	1.00	
9	3	1	2	1.00	
10	.	.	.	1.00	

Figure 5- 7

Here you see only numbers, which (we said early) are the new values of the new variable. You have an idea of what these values are, but o make the variable complete you need to add labels to these values.

To easily go to the properties of the new variable, double click the variable name (column header)-fig 5-8 below. Another way to get to the properties is through the *Variable View* tab (lower-left corner).



	THNKSELF	WORKHARD	XMOVIE	AGE1	var	var
1	.	.	.	3.00		
2	.	.	.	2.00		
3	.	.	.	3.00		
4	1	2	2	3.00		
5	.	.	.	2.00		
6	.	.	.	2.00		

Figure 5- 8

Double-clicking the variable header will open the "Variable View" tab in the Variable View window.

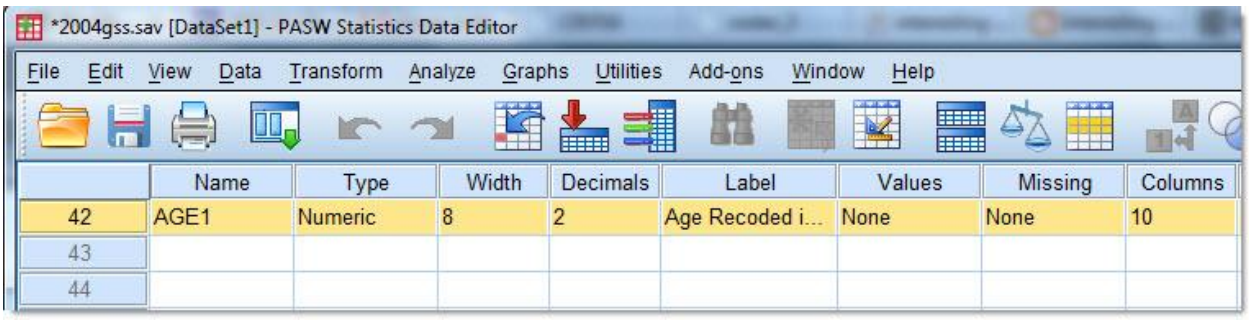


Figure 5- 9

Now is the time to enter labels, for the variable (remember at the beginning of the lecture?) and for the values. In the *Values* box click the little square button at the right side of the cell to go to the Value properties area. (Figure 5-10)

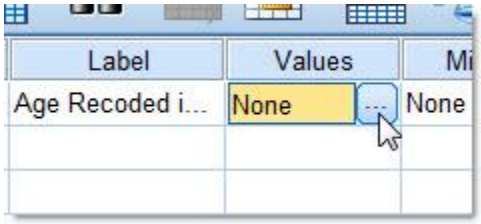


Figure 5- 10

The *Value Labels* dialog box opens. Remember: values are numbers, and labels are the alphanumeric descriptions of those values. Enter **1** in the *Value* field, and type **Young (18-35)** in the *Value Label* field. Click *Add* to enter the value label. Do the same procedure for values **2** and **3**, where you enter **Middle-aged (36-55)** and **Old (56 and up)** respectively. To correct an error in typing use the *Change* and *Remove*. If your screen looks like Figure 5-11 below, you're OK. Since we're at it, click **OK**.

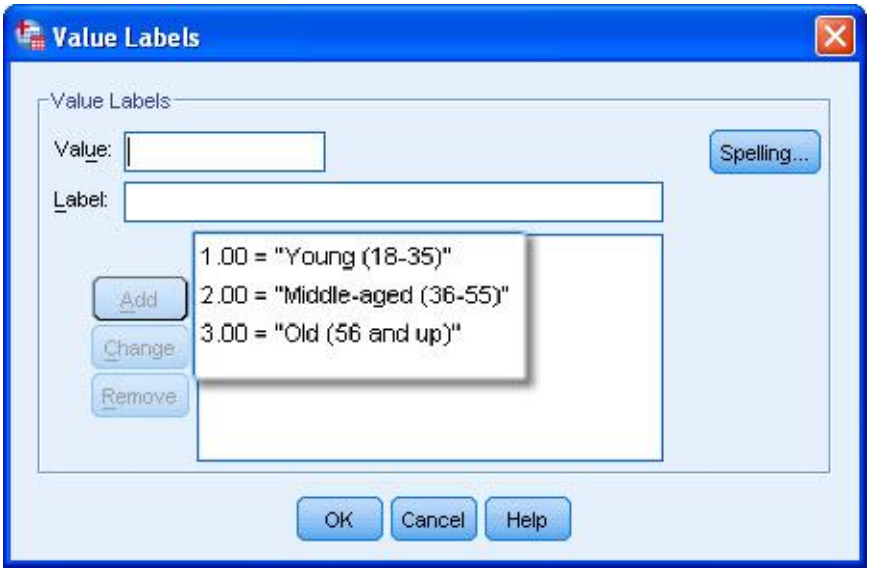


Figure 5- 11

One little stop here to learn how to reduce the decimal points. If you've carefully observed, new values are always created with two zeroes at the right of the decimal point. This is done by

default in SPSS. If these decimals annoy you, you can remove them by clicking the lower arrow in the *Decimals* cell in the Variable View window. See figure 5-12 for a visual.

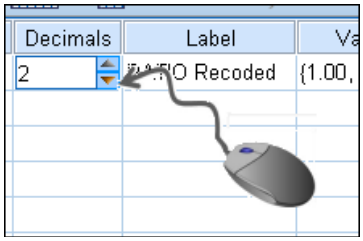


Figure 5-12

Rerun the frequencies distribution for **AGE1** again:

Analyze → **Descriptive Statistics** → **Frequencies**

This time, instead of values (1, 2, and 3), the table will display the value labels we just entered.

Recoded into Three Categories

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Young (18-35)	403	33.6	33.7	33.7
	Middle-aged (36-55)	478	39.8	40.0	73.7
	Old (56 and up)	314	26.2	26.3	100.0
	Total	1195	99.6	100.0	
Missing	System	5	.4		
Total		1200	100.0		

Table 5- 2

If you need to recode another variable into a new one, make sure to click on *Reset* button in the "Recode into Different Variables" dialog box to get rid of the previous recoding instructions.

We only covered the Range options in the recoding from old to new values. There are a few more options in the *Old Value* side of the dialog box.

- *Value* will help you recode just one value into a new one. For example, 3→1.
- *System-missing* and *System- or user-missing* will help you assign the old missing values in the new variable as they were in the old variable.
- *Range, LOWEST through value* is a shortcut to create a new value starting from the lowest old value to the cutoff you wish. For example, "lowest through 18 → Teenager".
- *Range, value through HIGHEST* is the same as the option above, but it includes all old values from the cutoff point and higher. For example, "56 through HIGHEST → Old". You have to be careful here, especially with the interval/ratio variables, since usually the highest values are assigned to the missing values; so, if you include value through HIGHEST, you may include the missing cases (such as 98, 99) into your analysis. Before recoding, look carefully at the old values, and note the missing values.
- Use *All other values* if you don't want to recode a particular value and want to leave that value in its original form. Recoding works like this: If the data set has a missing value, it retains its status as a missing value in the new variable. If it isn't any value (other than

a missing value) that is not recoded, it is changed into a system-missing value. *All other values* keep old values in their original form.

2. Recoding into the Same Variable

There are cases when you need to recode an old variable into a new one, and you don't need the old variable any more. This procedure is done through the *Recode into Same Variables* command. Basically, this command replaces the old variable. You have to be careful here: if you will need the old variable later again, make sure you create a copy of the variable, or even a copy of the dataset itself. This way, you may access the old variable again if needed.

To conduct this procedure, follow the steps in the diagram below:

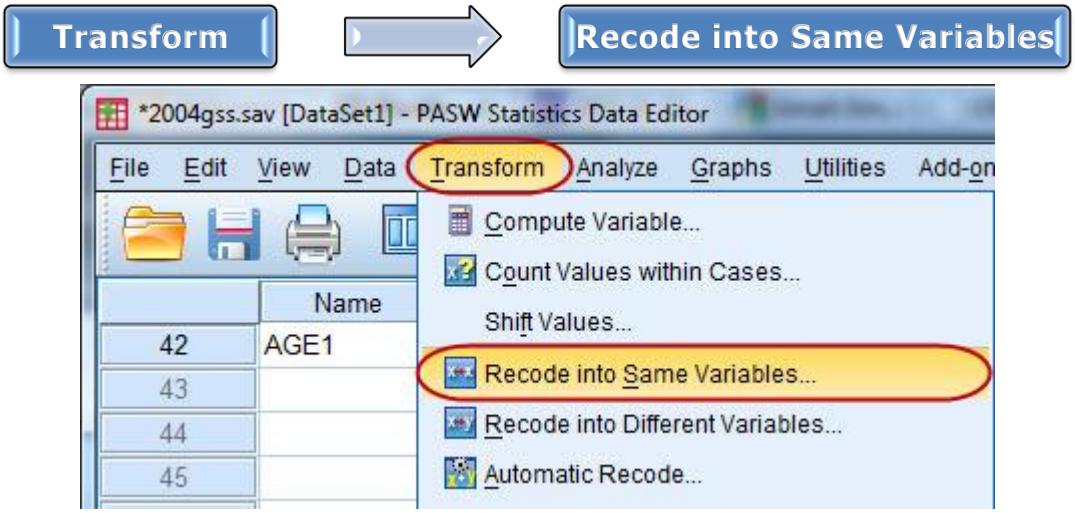


Figure 5- 13

Find and transfer the variable **PRAY** in the *Numeric Variables* field. Double-click the variable or click to select the variable and click the arrow button to transfer the variable appropriately.

Launch the *Recode into Same Variables: Old and New Values box* by clicking on the *Old and New Values* button. Figure 5-14 should reassure you that you are in the right track.

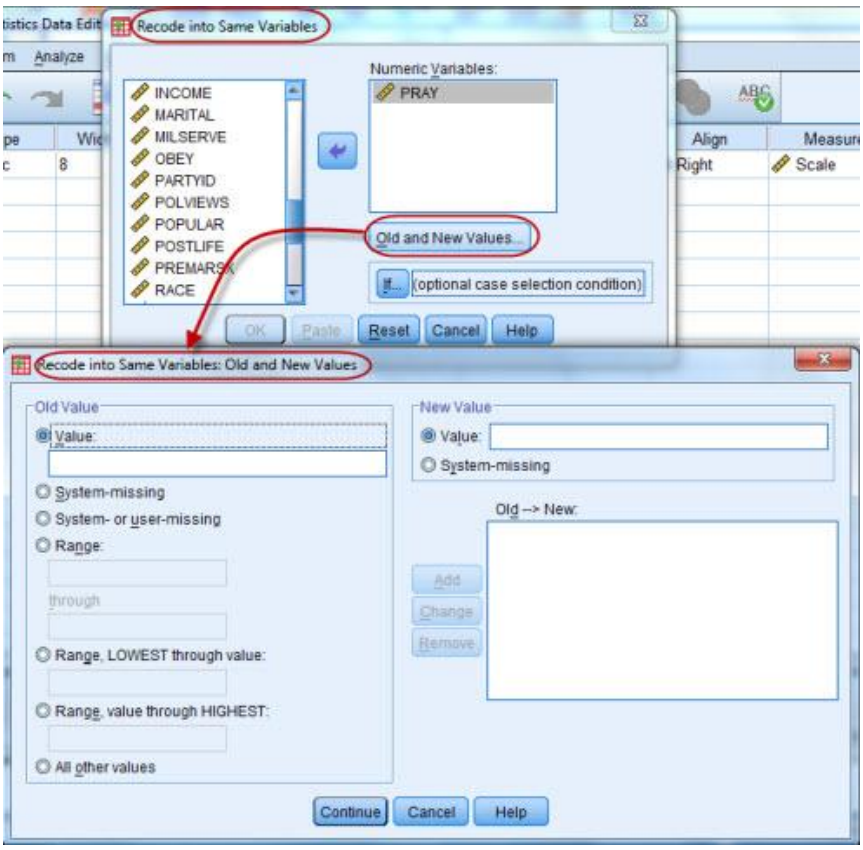


Figure 5- 14

This dialog box is very similar to the one we used to transform into different variables, (see Figure 5-4 above).

Recode as follows:

- 1 (several times a day) and 2 (once a day) will become new value 1 (prays a lot)
- 3 (several times a week) and 4 (once a week) will become new value 2 (prays moderately).
- 5 (less than once a week) and 6 (never) will become new value 3 (doesn't pray a lot).

Old Values	Old Value Labels	New Values	New Value Labels
1	Several times a day	1 (1-2)	Prays a lot
2	Once a day		
3	Several times a week	2 (3-4)	Prays moderately
4	Once a week		
5	Less than once a week	3 (5-6)	Doesn't pray a lot
6	Never		

Table 5-3

The variable will still be called **PRAY**.

The new value labels will be as depicted in the Table 5-2 above. To change the new value labels you conduct the same procedure as above, by double-clicking at the variable header/name (**PRAY**). In

the *Values* box (Variable View window) click the little square button at the right side of the cell to go to the Value properties area. There is no Add button, so you will have to use the *Change* and *Remove* buttons to change labels.

Run a frequencies procedure to get a frequency distribution for new **PRAY**.

Summary

Recoding is a very useful procedure that makes your data more manageable. Its value is especially evident in the case of interval and ratio variables, where they have many attributes, created by the answer categories. Also, make sure that you understand the different levels of measurement, since many times the recoded variable may have a different level than the old variable. Depending on what level of measurement the new variable has, you will need to tailor the difference statistical tests.