

How can string variables be automatically recoded into coded categories using the process of computing variables?

Authored by
stats writer

June 23, 2024

RECOMMENDED CITATION

stats writer (2024). *How can string variables be automatically recoded into coded categories using the process of computing variables?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=149309>

The process of computing variables allows for string variables to be automatically recoded into coded categories. This involves assigning numerical codes to different string values, based on predefined criteria or patterns. This method helps to organize and streamline data analysis, as it converts qualitative data into quantitative data that can be easily analyzed and compared. By utilizing this process, string variables can be efficiently transformed into coded categories, providing a more structured and organized approach to data analysis.

Recoding Categorical String Variables to Labeled Numeric Variables using Automatic Recode

When writing down the observed values of a categorical variable, you can choose to write the data 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.)

String 1	String 2	Match or mismatch?
Chicago	Chicago	Match
CHICAGO	CHICAGO	Match
Chicago	chicago	Mismatch (difference in capitalization)
chicago	chicago	Mismatch (notice the extra space after string 1)

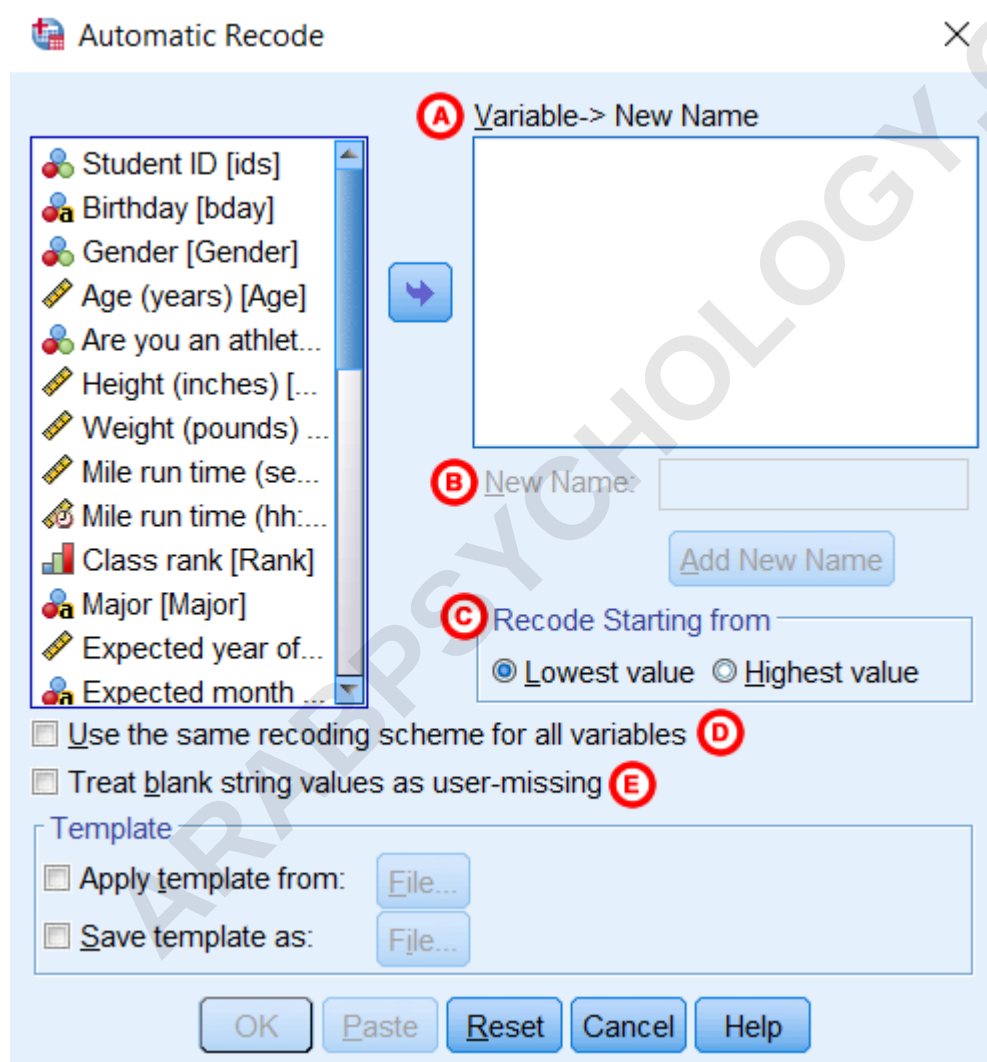
If you have already recorded your categorical variables as strings, you can easily convert them to a labeled, numerically coded variable using the Automatic Recode procedure. This procedure assigns each unique category a numeric code, then saves the converted values as a new variable. It also automatically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

Additionally, if you have used blanks to indicate missing values for string variables, you 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, you *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.

Note: Before using this procedure, you should resolve any issues with "mismatched" category strings. For example, if there are different capitalizations of the same word, you should "normalize" them to all use the same capitalization before you enter the variable into Automatic Recode. Functions like `UPCASE()` and `LOWCASE()` can perform these transformations (see the [Compute Variables](#) tutorial). A [frequency table](#) will help determine which categories, if any, are mismatched.

Running the Automatic Recode Procedure

To open the Automatic Recode procedure, click **Transform > Automatic Recode**.



A Variable -> New Name: The original variable(s) being transformed, and the name of the new variable(s) that the results will be saved as.

B New Name field and **Add New Name** button: These fields will activate after at least one variable has been added to the Variable -> New Name box. You will need to supply a new variable name

and click Add New Name for each variable being recoded.

CRecode Starting from: Should the new category numbering be in alphabetical order (Lowest value) or reverse alphabetical order (Highest value) with respect to the original values? This setting is applied to all of the variables being recoded.

DUse the same recoding scheme for all variables: When checked, the same numeric code is never re-used across variables, unless the category names are identical.

ETreat blank string values as user-missing: When checked, the numeric category assigned to blank strings will be set as a special missing value. This setting must be checked in order for missing values to be properly recognized.

To automatically recode variables:

Click **Transform > Automatic Recode**. Select the string variable of interest in the left column and move it to the right column. Enter a new name for the autorecoded variable in the **New Name** field, then click **Add New Name**. SPSS will assign numeric categories in alphabetical order. By default, this means that the lowest numeric categories will be assigned to category names coming first in the alphabet. You can change this so that categories coming later in the alphabet are given the lowest numeric category by clicking **Highest value**. If blanks were used to indicate missing values, select the **Treat blank string values as user-missing check box**. If you are converting multiple string variables and do not want the same number to be re-used as a category code across multiple variables, select the **Use same recoding scheme for all variables** check box. Click **OK** to finish.

Example: Recognizing Blank Strings as User-Missing Values

Problem Statement

In the sample data file, the variable State is a string variable representing whether the student is an in-state student or an out-of-state student. If you create a frequency table of this variable (**Analyze > Descriptive Statistics > Frequencies**), you will notice something strange:

```
FREQUENCIES VARIABLES=State
/ORDER=ANALYSIS.
```

Frequencies

Statistics

State Residency

N	Valid	435
	Missing	0

State Residency

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	27	6.2	6.2	6.2
In state	314	72.2	72.2	78.4
Out of state	94	21.6	21.6	100.0
Total	435	100.0	100.0	

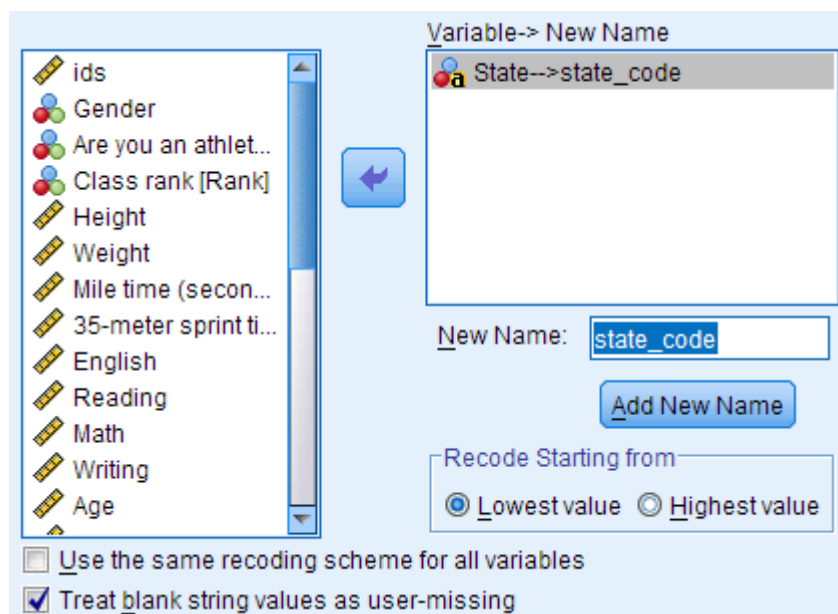
The dataset has 435 observations in all, and SPSS reports that there are zero missing values. But there is an apparently unlabeled category listed under the "Valid" categories in the frequency table that has 27 observations. **This is because SPSS does not automatically recognize blanks as missing values.** (Note: this behavior is different than SAS, which automatically recognizes blanks as missing values for string variables.)

In order for our analyses to be accurate, we'll need to fix this issue.

Running the Procedure

Using the Automatic Recode Dialog Window

Click **Transform > Automatic Recode**. Double-click variable State in the left column to move it to the Variable -> New Name box. Enter a name for the new, recoded variable in the **New Name** field, then click **Add New Name**. Check the box for **Treat blank string values as user-missing**. Click **OK** to finish.



Using Syntax

```
AUTORECODE VARIABLES=State  
/INTO state_code  
/BLANK=MISSING  
/PRINT.
```

Output

Running the procedure will produce the following message in the Output Viewer window:

```
State into state_code (State Residency)  
Old Value New Value Value Label
```

```
In state 1 In state  
Out of state 2 Out of state  
M 3M
```

This message tells us the mapping scheme that SPSS generated for the categories: "In state" became 1, "Out of state" became 2, and blanks became 3, which was set as a special missing value code. You can confirm In the Variable Viewwindow, you can see that in addition to copying the original string values to the category labels, SPSS also defined category 3 as "missing."

Name	Type	Width	Decimals	Label	Values	Missing
State	String	12	0		None	None
state_code	Numeric	1	0	State Residency	{1, In state}...	3

Now when we create a frequency table for the recoded variable, it should reflect the proportion of values that are missing:

```
FREQUENCIES VARIABLES=state_code
/ORDER=ANALYSIS.
```

Frequencies

Statistics

State Residency

N	Valid	408
	Missing	27

State Residency

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	In state	314	72.2	77.0	77.0
	Out of state	94	21.6	23.0	100.0
	Total	408	93.8	100.0	
Missing	3	27	6.2		
Total		435	100.0		

In the output, the recoded blank values are correctly counted as missing, but show their assigned numeric code in the Value Labels column in the output. You can improve the appearance of the missing value category by simply adding a value label (e.g. "Missing") for that particular code.

What if I use Automatic Recode when I've already defined special missing value strings?

If you recorded the missing values for a string variable using some kind of non-blank indicator (for example, 999 or -999) and have already defined that user-missing value in the Variable View window, Automatic Recode will preserve the 'missing' designation, but will still convert the category code to be in the range of the other categories.

Suppose that the below syntax was applied to a string variable with the valid categories "blue", "green", and "red", with missing values recorded using the code "999".

```
AUTORECODE VARIABLES=VAR00001  
/INTO v1  
/GROUP  
/BLANK=MISSING  
/PRINT.
```

Running that syntax will produce the following output:

```
User-missing values from VAR00001  
Old Value New Value Value Label
```

```
blue 1 blue  
green 2 green  
red 3 red  
999 M 4M 999
```

Here, you can see that observations originally coded as "999" have been recoded to the numeric indicator 4 with the value label "999". The letter M indicates that the label or code is a missing value indicator.

When would I use the same recoding scheme for all variables?

We have not yet discussed the option **Use the same recoding scheme for all variables**. What are some reasons to use this option?

You have many string categorical variables to recode, and do not want to have the same number re-used on unrelated categories. The variables have overlapping or identical categories, and you do not want categories with the same objective values to be assigned different numeric codes.

Here is a simple example of using the same recoding scheme for all variables.

VAR00001	VAR00002
green	orange
blue	violet
red	orange
blue	orange
green	999
green	violet
999	red
red	orange

Notice that both VAR00001 and VAR00002 have a category called "red". Since both variables represent colors, it makes sense to use a single coding scheme for all of the possible color categories (i.e., the category "red" will be represented by the numeric code 4 for both variables, rather than having a different code in VAR00001 versus VAR00002). Executing the recoding syntax produces the following output:

```
AUTORECODE VARIABLES=VAR00001 VAR00002
```

```
/INTO v1 v2
```

```
/GROUP
```

```
/PRINT.
```

```
User-missing values from VAR00001
```

```
Old Value New Value Value Label
```

```
blue 1 blue
```

```
green 2 green
```

```
orange 3 orange
```

```
red 4 red
```

```
violet 5 violet
```

```
999 M 6M 999
```

As you can see from the syntax, SPSS first alphabetizes all possible unique nonmissing category values across the two variables, then assigns numeric codes to each category. Notice that the "missing" category ("999") was recoded last, even though it is alphabetically before the category names.