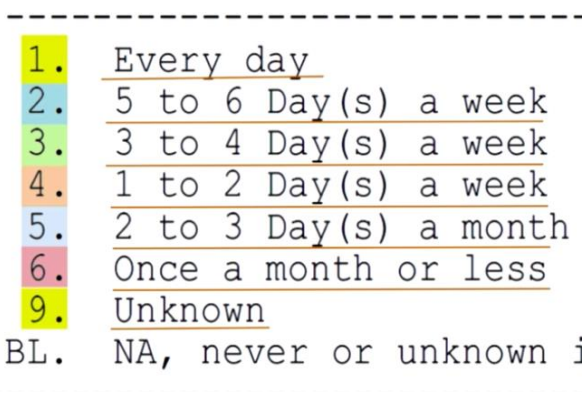
5. Data Management

Video Link:
https://www.youtube.com/watch?v=9KxSZEWodR0&index=5&list=PL2fQHGEDK7Yyl1W9tgIo8wpYFTDumgc_j

Section 5.1: Code Out Missing Data

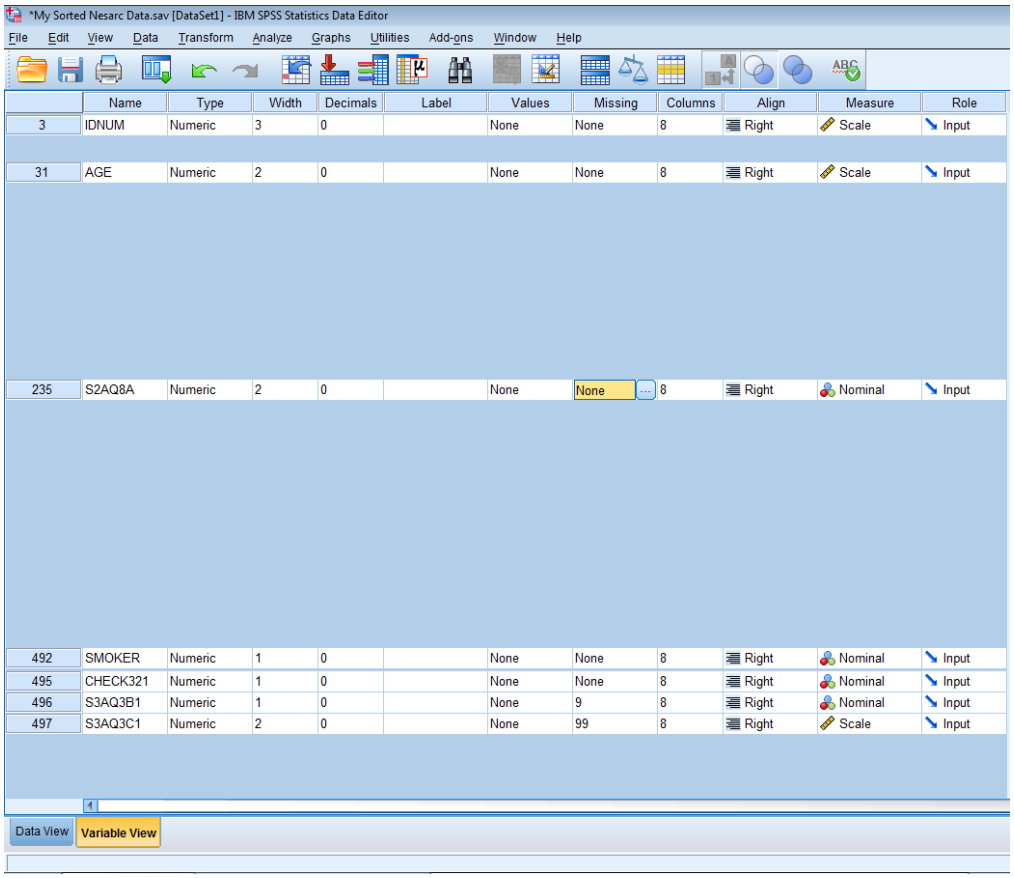
In this next step, we'll be making decisions about recoding data in ways that will help us answer our question. This process is also known as **data management**. A necessary first step to consider is whether or not you need to code out missing data. Some variables include values that represent uncertain or unknown information. Refer to your personal code book to determine if any of your variables have values that represent unknowns. In order to perform analyses on these variables, we have to set those values to missing. In the following frequency table, the value 9 represents an answer of unknown or not applicable.



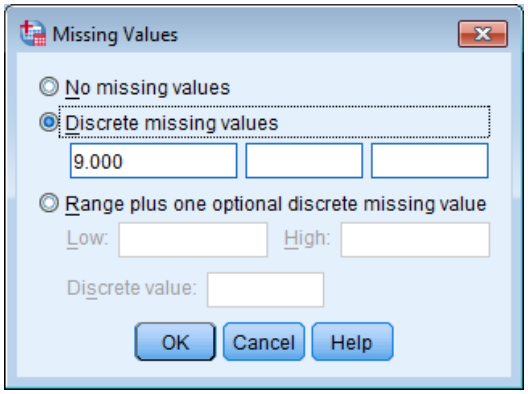
Usual Smoking Frequency

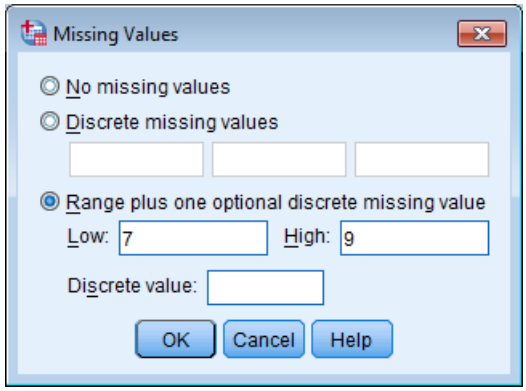
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1320	77.4	77.4	77.4
	2	68	4.0	4.0	81.4
	3	91	5.3	5.3	86.7
	4	88	5.2	5.2	91.9
	5	65	3.8	3.8	95.7
	6	71	4.2	4.2	99.8
	9	3	.2	.2	100.0
	Total	1706	100.0	100.0	

1. To set 9 to missing, click **Variable View** in the lower left hand corner. You will see the name of each variable you selected in the first column. Scroll down to the row corresponding to the variable you need to treat for missing data. Move your cursor to the seventh column titled **Missing**. Click in the box with the word **None**, then click the box with the three dots just to the right of **None**.



2. Click **Discrete** to list one value per box that you want to set to missing for the corresponding variable. To set multiple values that are in ascending order click **Range** plus one optional discrete missing value, then enter the Low: and High: numerical values you want to set as missing for that variable. Click **OK**.





If you run a frequency table on the variables you coded for missing you will see that at the end of the table the value(s) have been set to missing.

Usual Smoking Frequency

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1320	77.4	77.5	77.5
	2	68	4.0	4.0	81.5
	3	91	5.3	5.3	86.8
	4	88	5.2	5.2	92.0
	5	65	3.8	3.8	95.8
	6	71	4.2	4.2	100.0
	Total	1703	99.8	100.0	
Missing	9	3	.2		
Total		1706	100.0		

Return to Video - 05. Data Management

Section 5.2: Code in Valid Data

To avoid asking inappropriate questions, skip patterns are often created in surveys that allow participants to skip questions in which the answer can be logically determined. In this way, missing data on some questions might mean that we can reasonably recover valid information.

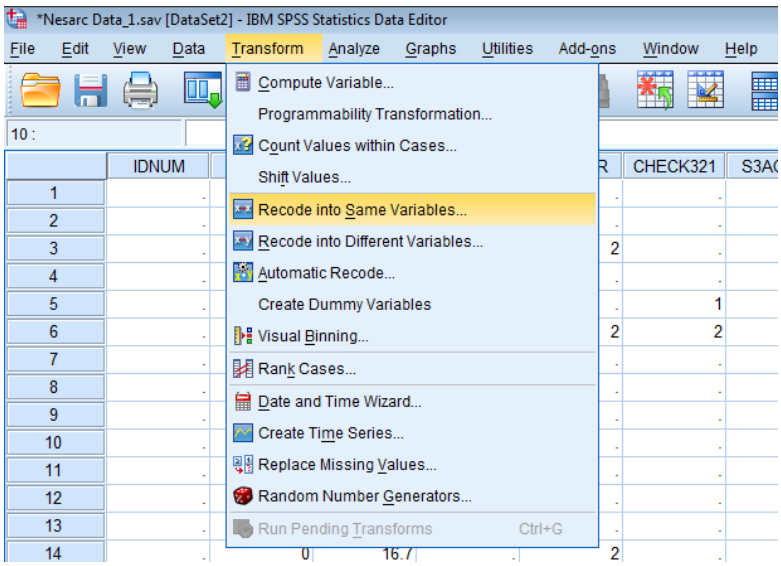
For the question, “Did you drink at least one alcoholic drink in the past 12 months?”, 16,116 participants said “No”. These individuals would not need to be asked the question about how often

they drank alcohol in the past 12 months. Instead, they will be set to blank, or in the case of a SPSS dataset, a dot or period. So, for the variable, how often did you drink alcohol in the past 12 months, it would be reasonable to code this as valid data rather than missing. In the variable below we see that **Missing System** frequency is 180. We can set those to valid responses.

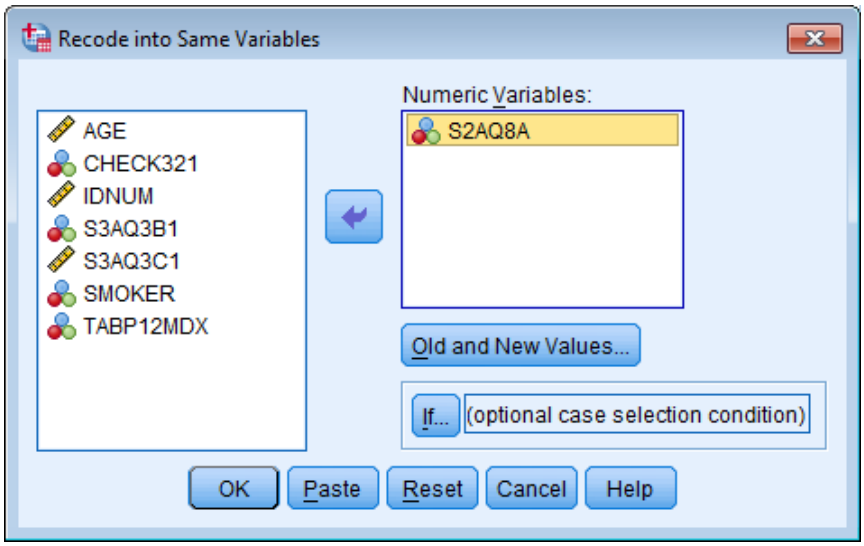
Frequency of Drinking past 12 months

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	76	4.5	5.0	5.0
	2	84	4.9	5.5	10.5
	3	194	11.4	12.7	23.2
	4	229	13.4	15.0	38.2
	5	216	12.7	14.2	52.4
	6	248	14.5	16.3	68.6
	7	134	7.9	8.8	77.4
	8	85	5.0	5.6	83.0
	9	134	7.9	8.8	91.7
	10	118	6.9	7.7	99.5
	99	8	.5	.5	100.0
	Total	1526	89.4	100.0	
Missing	System	180	10.6		
Total		1706	100.0		

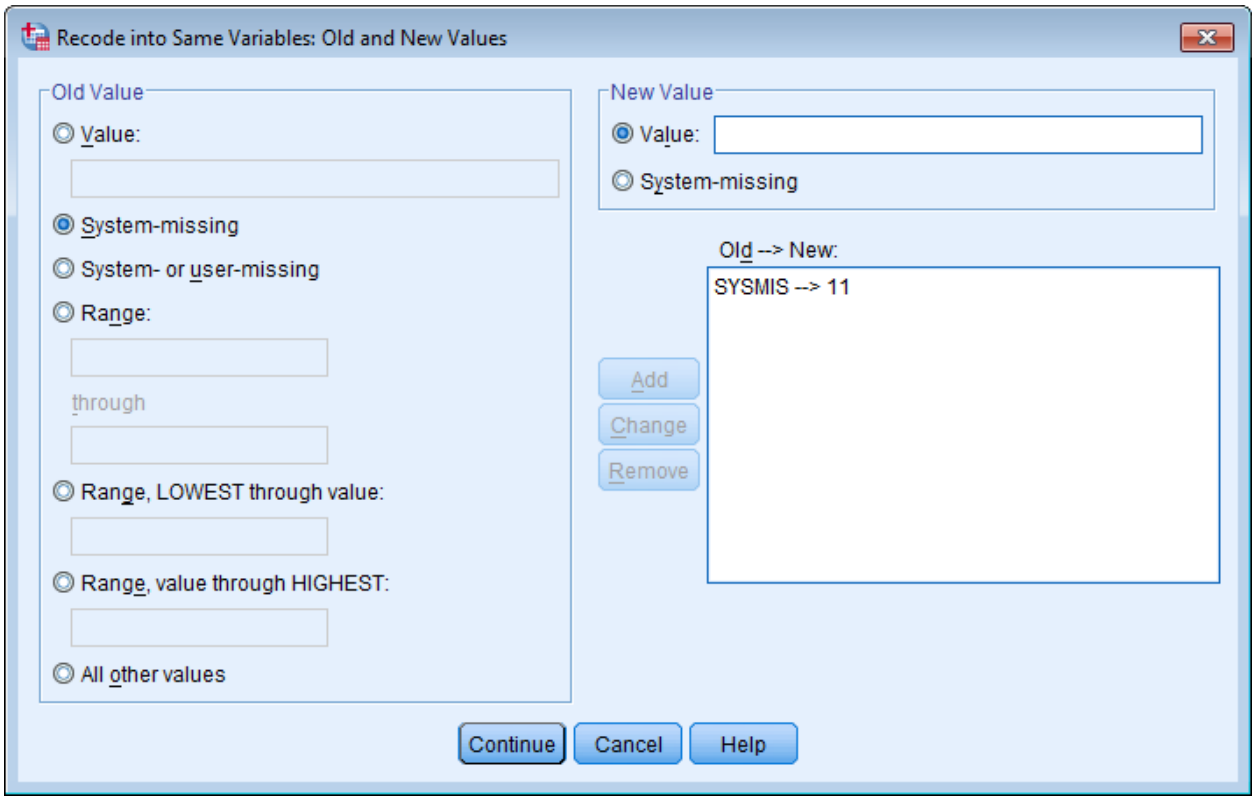
1. Click **Transform > Recode into Same Variables**.



2. Select the variable you want to recode from the list on the left, move it to the **Numeric Variables** column on the right using the arrow. Click **Old and New Values…**



3. In the **Old Value** column on the left hand side click **System-missing** since currently the value is treated as missing indicated by a "." in **Data View**. In the **New Value** column on the right hand side click **Value** then enter the numerical value (i.e., dummy code) you want it to now be. Refer to your personal code book to determine for each of your specific variables what would be an appropriate dummy code to use for each variable you intend to code in valid data. Click **Add** > **Continue.**



If you need to do this for more than one variable, move the variable currently listed under **Numeric Variables** you just 'coded in' as valid back to the left column using the arrow. Click on the name of the next variable and move it to the right column. Click **Old and New Values…** It is extremely important to click in the **Old --> New:** column in the lower right hand side, click on what you previously added, and then click **Remove**. It is unlikely that you would code in valid data on a different variable and use the same dummy code. Failure to do this will result in the variable you are working with now to be coded for valid data how you asked SPSS to code the previous variable. Repeat step #3 above.

Run a frequency table on the variable(s) on which you 'coded in' valid data to confirm that the new valid responses (i.e., dummy code of 11) shows in the table.

Frequency Drank Alcohol Last 12 Months

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	76	4.5	4.5	4.5
	2	84	4.9	4.9	9.4
	3	194	11.4	11.4	20.8
	4	229	13.4	13.5	34.3

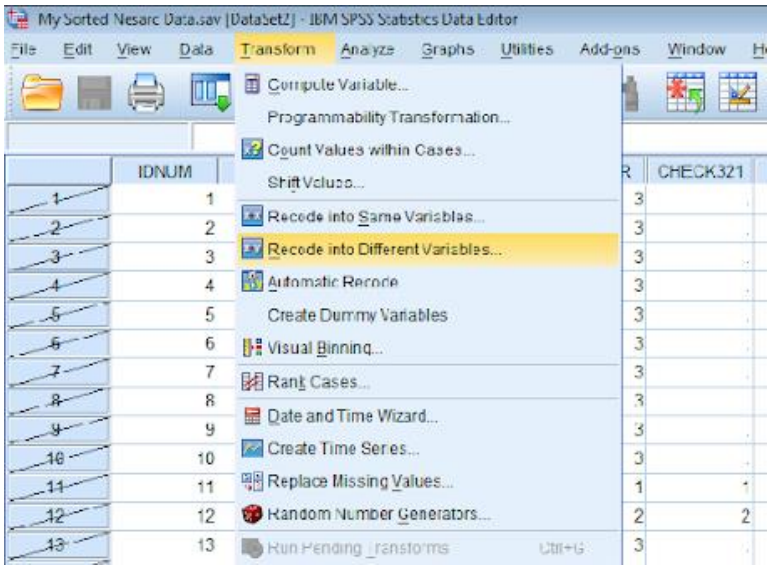
	5	216	12.7	12.7	47.1
	6	248	14.5	14.6	61.7
	7	134	7.9	7.9	69.6
	8	85	5.0	5.0	74.6
	9	134	7.9	7.9	82.4
	10	118	6.9	6.9	89.4
	11	180	10.6	10.6	100.0
	Total	1698	99.5	100.0	
Missing	99	8	.5		
Total		1706	100.0		

Return to Video - 05. Data Management

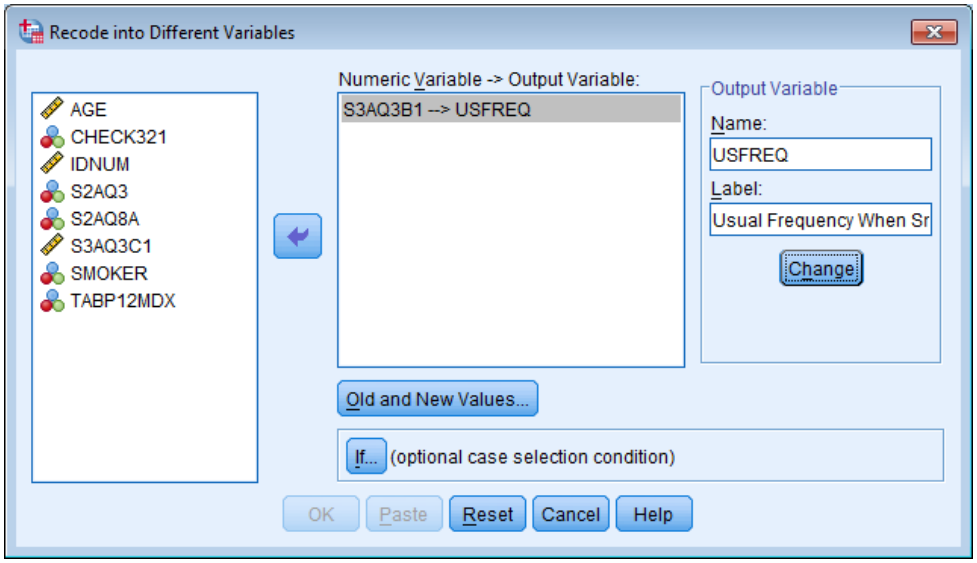
Section 5.3: Logical Response Codes

Another very useful step in data management is to give your variables response codes that may be more logical than those they were originally given. For the variable, usual smoking frequency, you can see from the code book entry that lower values mean that the respondents smoked more and higher values mean that the respondents smoked less. Perhaps this seems counterintuitive to you. We could chose to reverse code this variable to be more intuitive so that higher values mean more smoking and lower values mean less smoking.

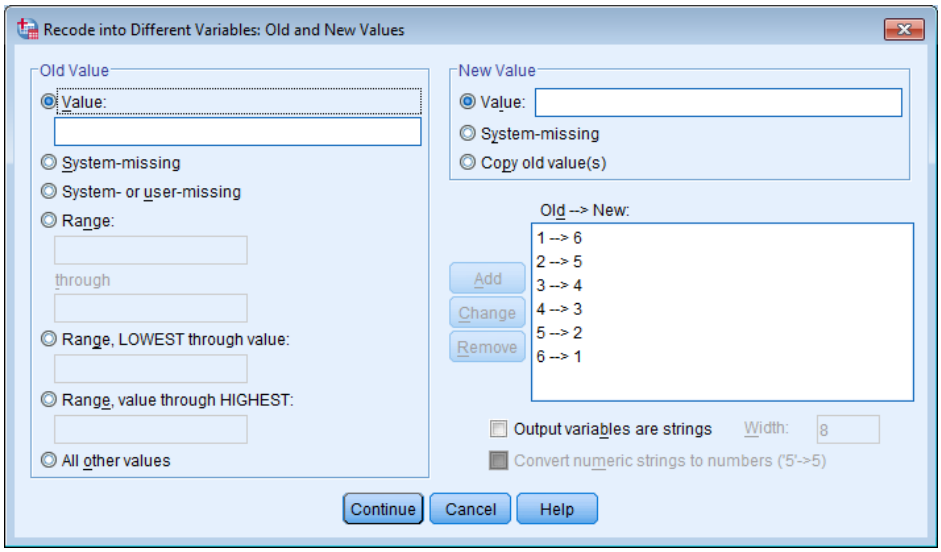
1. Click **Transform > Recode into Different Variables**.



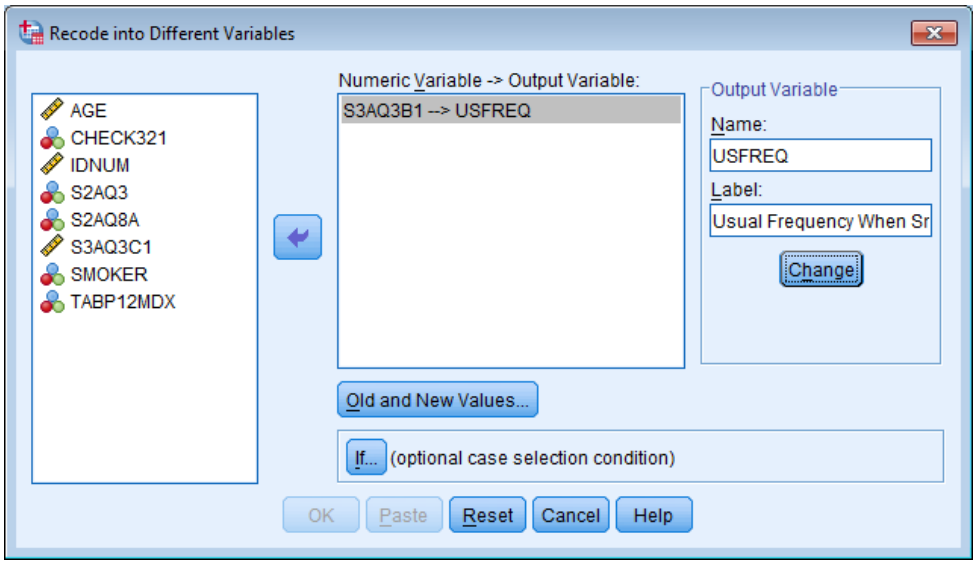
2. Choose the variable from the list in the far left column and move it to the middle column by clicking the arrow. In the far right column create a new variable name. This should be unique to the dataset, have no spaces, and be 12 characters or less. Directly below that, label this new variable. It can be the same label as the variable you are coding to be more logical or something new that appropriately describes the variable. Click **Change > Old and New Values…**



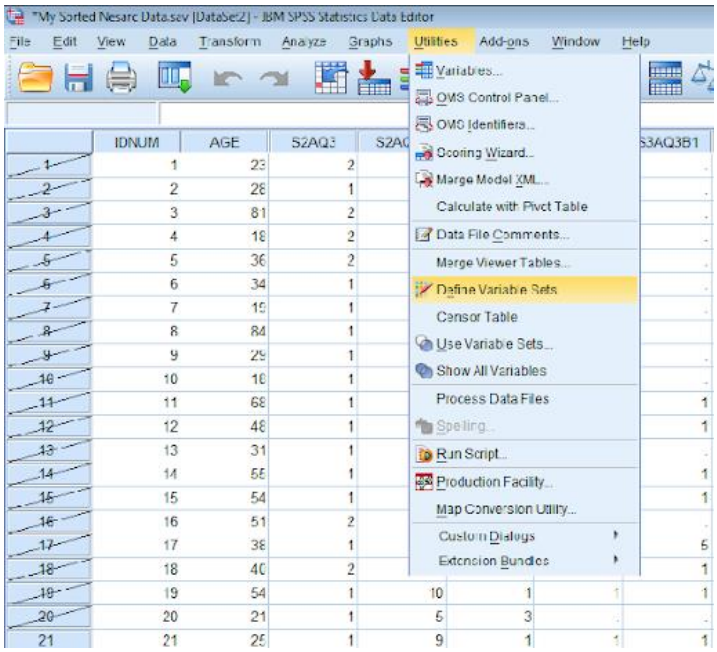
3. Refer to your personal code book to verify what the current dummy codes are for the variable you are recoding to be more logical and decide what the new dummy codes will be. Typically you will be reversing the current dummy codes so that lower dummy codes represent categories of lower value and higher dummy codes for higher value. In the **Old Value** column on the left side input the current dummy code under **Value**. In the **New Value** column on the ride side input the new dummy code that is more logical next to **Value**. Click **Add**. Repeat this for each dummy code for the variable you are recoding. Once you are done with this verify that in the lower right column under **Old-->New** you see the original dummy codes going down in ascending order and the new dummy codes that are more logical going down in descending order. Click **Continue**.



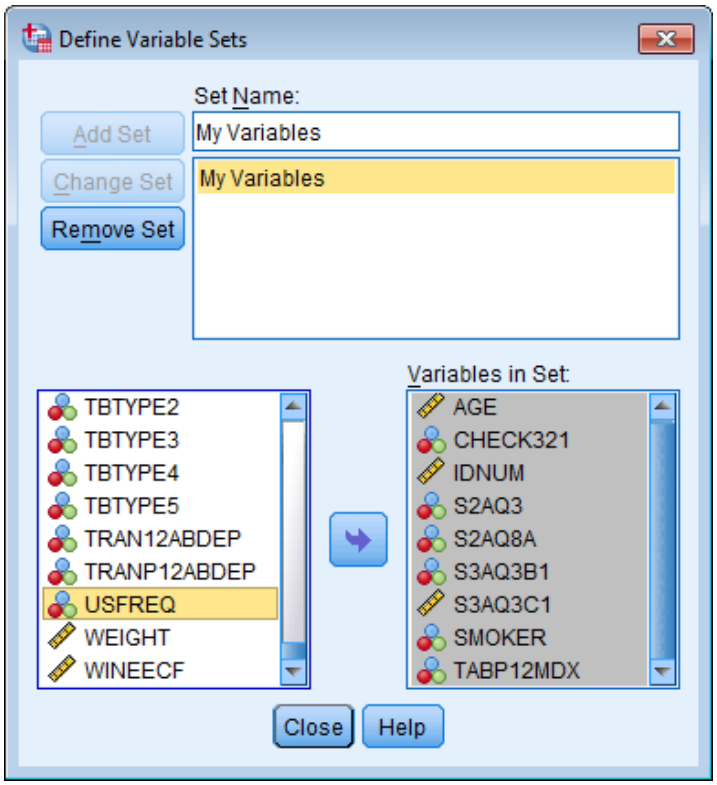
4. If you have more than one variable to recode click the variable name of old and new variable in the middle column and move it back to the left column using the arrow. Repeat step #2 through #3. Make sure to remove or change the Old-->New values in the lower right column otherwise the system will recode the current variable you are data managing the same way it did your previous variable.



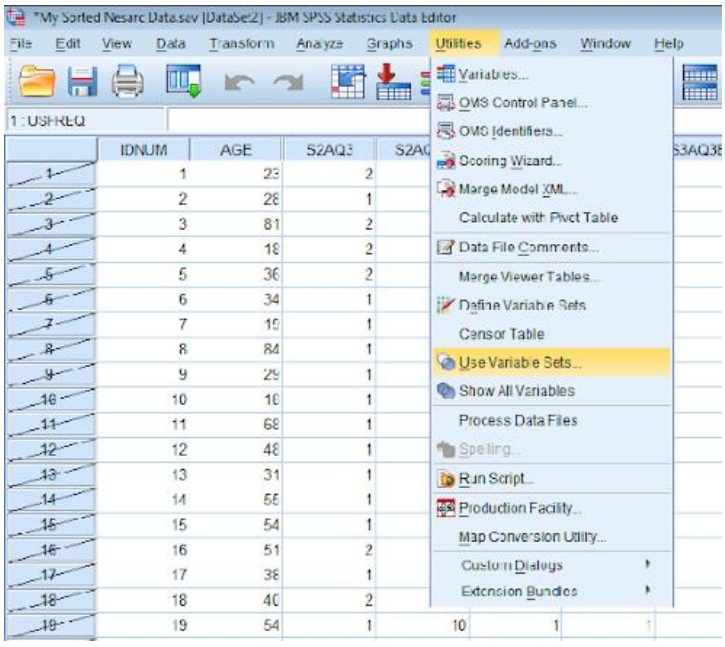
5. Once you have recoded to make more logical response categories for all of the variables you decided to recode, Click **OK**. Then at the top menu click **Utilities > Define Variable Sets**.



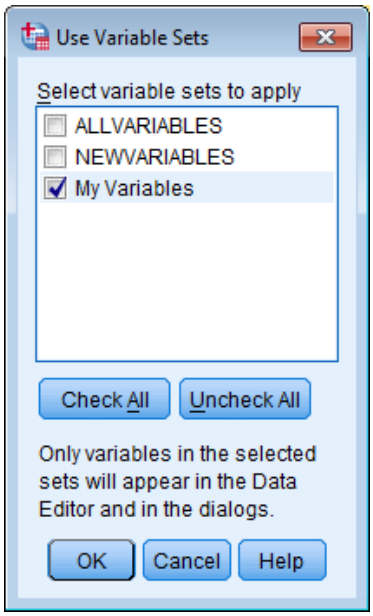
6. Click on **My Variables**, or whatever you previously named your Variable Set. Move the newly created variable(s) from the lower left column to the lower right column using the arrow. Click **Change Set** toward the top left. Click **Close**.



7. Click **Utilities > Use Variable Sets**.



8. Confirm that My Variables or whatever you named your Variable Set is still clicked and the others listed are unchecked. Click **OK**.



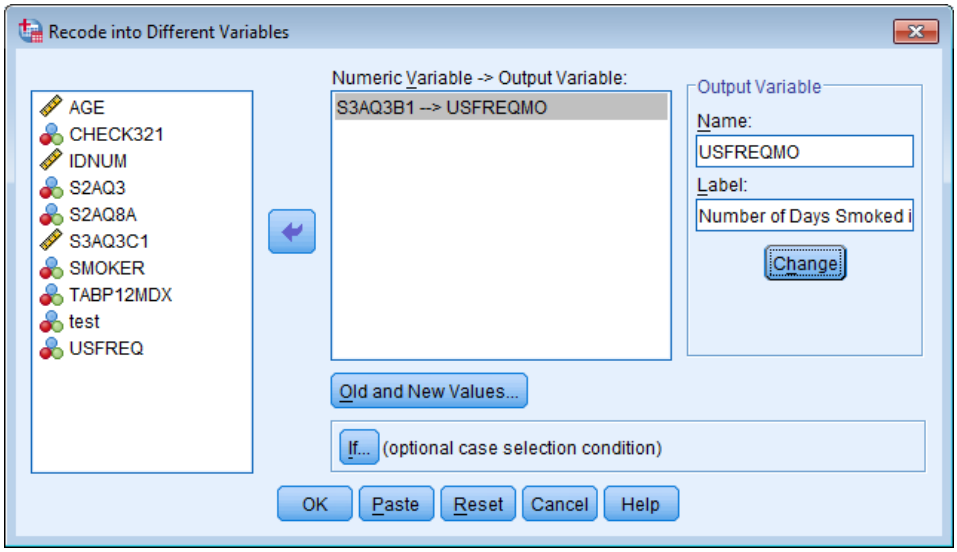
You will see your newly created variables in Data View and Variable View. **Now, rather than using the original variable, S3AQ3B1 in our analysis, we will use this new variable USFREQ**. It is important to add these details to the code book so there is good documentation of these changes as you move forward.

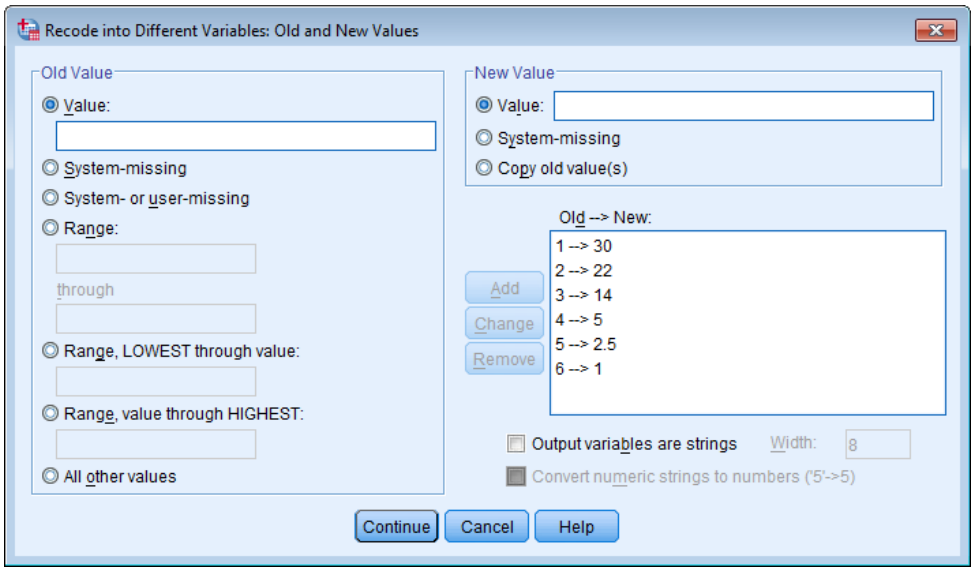
There is actually another option that may be even better for giving this variable values that are maximally informative. In a way, dummy codes 1 through 6 are unnecessarily categorical given that how often you smoke is generally a quantitative measurement. And it may be reasonable to recode the variable in order to capture more of its quantitative features.

To do this, we choose values that reasonably correspond to the number of times each individual smokes in a usual month. So, someone who reports smoking everyday could be said to smoke cigarettes on 30 days in a usual month. Someone who smokes 5 to 6 days a week could be said to smoke 22 days in a usual month. Someone who smokes 3 to 4 days a week could be said to smoke 14 days in a usual month.1 to 2 days a week, 5 days in a usual month. 2 to 3 days a month, 2.5 days in a usual month and once a month or less, 1.

Although these are estimates, they capture the quantitative nature of the measure and also keep individuals ordered in terms of the frequency with which they smoke. Our new variable is called USFREQMO, which stands for number of days smoked in a usual month.

9. Repeat steps #1 through #8.





So, this even makes more sense, doesn't it? What we've basically done is to spread the measurement of smoking frequency out in a meaningful way. So, that while it is a categorical variable, we are actually getting more information out of it than we had originally been given.

When we run a frequency table, the table now describes the sample of 1,706 young adult past year smokers, somewhat more clearly, by showing the numbers and percentages of individuals who smoked these approximate number of days in a usual month. You can see that the vast majority of the sample, 1,320 individuals or 77.4% smoked all 30 days.

Number of Days Smoked in a Usual Month

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	71	4.2	4.2	4.2
	2.50	65	3.8	3.8	8.0
	5.00	88	5.2	5.2	13.2
	14.00	91	5.3	5.3	18.5
	22.00	68	4.0	4.0	22.5
	30.00	1320	77.4	77.5	100.0
	Total	1703	99.8	100.0	
Missing	System	3	.2		
Total		1706	100.0		

Return to Video - 05. Data Management

Section 5.4: Create Secondary Variable from 2 Variables

Secondary variables are variables that include information from two or more primary variables. We can create secondary variables by using a mathematical or logical operation on two or more variables.

In this case we want to know the number of cigarettes smoked per month. We know the number of usual number of cigarettes smoked on days that an individual smokes (quantity), S3AQ3C1, and our new variable, USFREQMO, give us an estimate of the number of days smoked in a usual month. If we want to estimate the total number of cigarettes that participants smoked per month, it would make sense to multiply those two variables and get a product that represents the number of cigarettes per day times the number of days smoked per month.

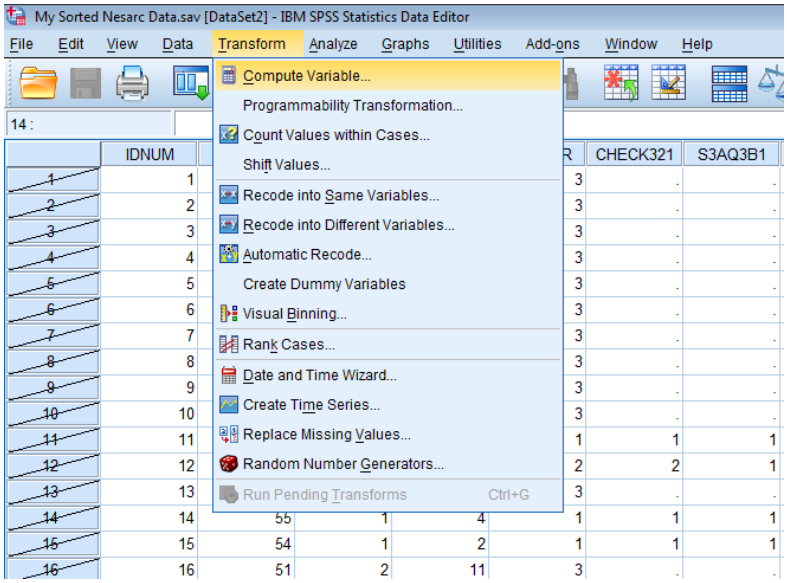
USFREQMO — Usual smoking days per month

1 = Once a month or less
2.5 = 2 – 3 Day(s) a month
6 = 1 – 2 Day(s) a week (1.5 x 4 wk)
14 = 3 – 4 Day(s) a week (3.5 x 4 wk)
22 = 5 – 6 Day(s) a week (5.5 x 4 wk)
30 = Every day

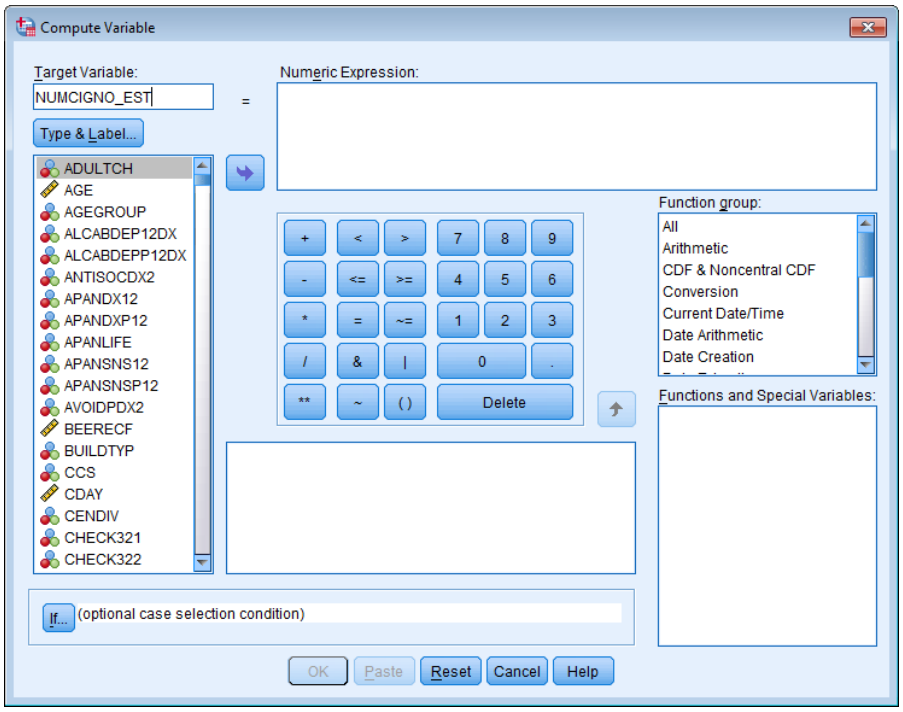
S3AQ3C1 — Usual quantity when smoked cigarettes

1 – 98 usual cigarettes/day

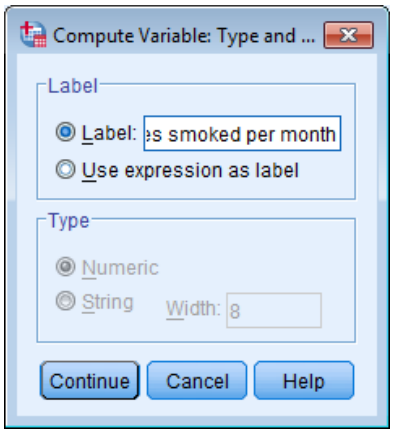
1. To do this, click **Transform > Compute Variable**.



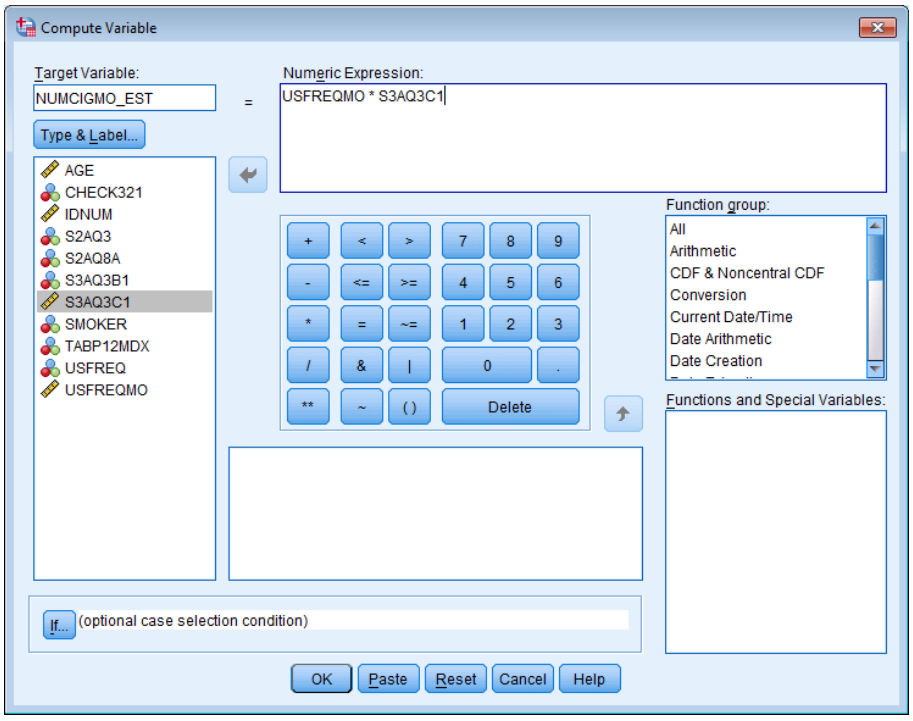
2. Enter the name of the new variable under **Target Variable** on the far upper left hand corner. For our example, let's call this new variable NUMCIGMO_EST, which stands for the estimate of the number of cigarettes smoked per month. Click **Type & Label...** directly below the variable name you just typed in.



3. Type in an appropriate label for the secondary variable you are creating. Click **Continue**.



4. In the lower left hand column find the name of the first variable you are using to create this secondary variable. Use the arrow to move it to the upper right column titled **Numeric Expression**. Since we are multiplying the two variables together we will insert an asterisk "*****" then the second variable name. Then click **OK**.



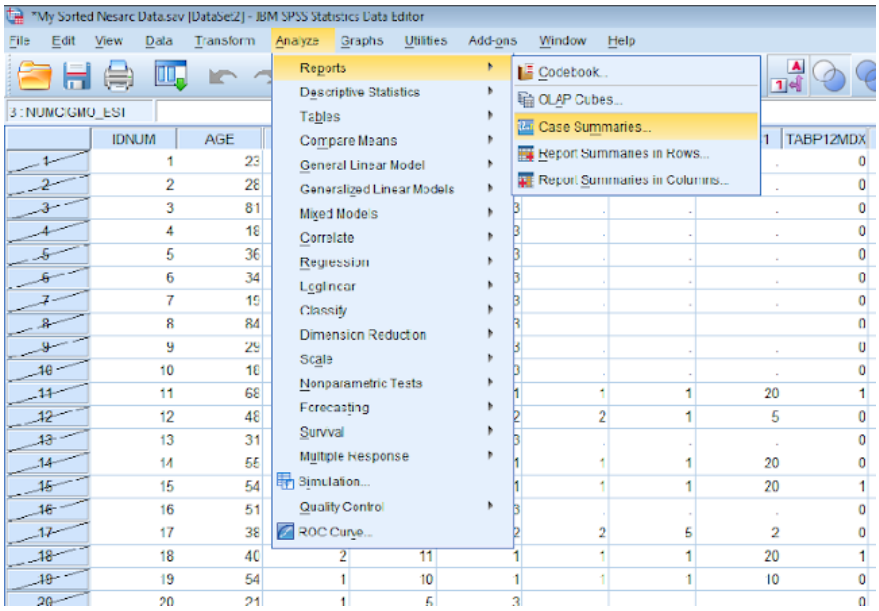
To create your secondary variable you may need to add "**+**", subtract "**-**", multiple "*****", divide "**/**" or a series of those steps.

5. Repeat steps 5.3 #5 through #8 to make sure that your secondary variable you just created to show up in Data View and Variable View.

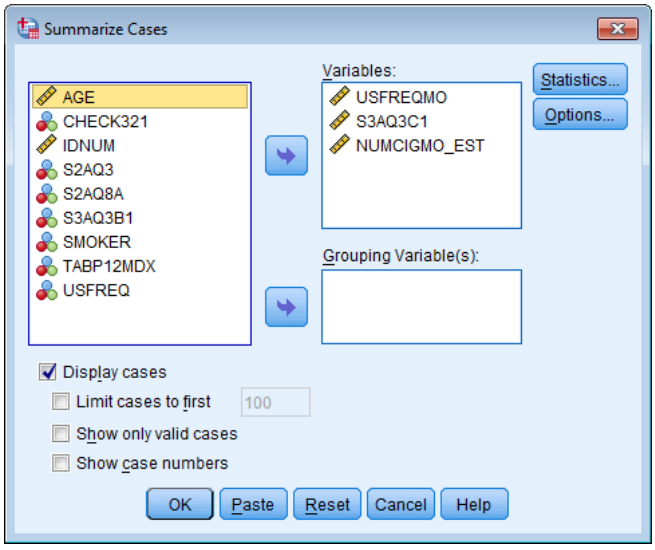
6. If you need to create another secondary variable repeat steps #1 through #5 making sure to change the **Target Variable** name, **Label,** and **Numeric Expression**.

If we generate a frequency distribution for this new variable, NUMCIGMO_EST, you see it is a quantitative variable that ranges from 1 to 2,940 with 9 missing observations. So, how can we check to make sure that this new secondary variable was created as we intended?

7. Click **Analyze > Reports > Case Summaries**.



8. Remembering the order you put your variables into the **Numeric Expression** when you created this secondary variable, move the variables along with the newly created secondary variable from the left column to the upper right column. In the lower left corner unclick **Limit cases to first** and **Show only valid cases** then click **OK**.



This will create a table in your output that will list every single participant id number in the first column, the values for each of the participant's responses for the variables you used to create the secondary variable and the secondary variable itself. (Note the middle of the table was

omitted to save space and we show only showing the beginning and the end.) The rows represent individual observations similar to the dataset itself. And the columns show the values for the specific variables. Looking over these values for a handful of individuals, we can see that, in fact, our new secondary variable does indeed show the value of number of days smoked per month times number of cigarettes smoked per day. Remember that whenever you are conducting data management, it is important to find a way to check for errors at each step of the process.

Case Summaries

	Number of Days Smoked in a Usual Month	Usual Quantity When Smoked Cigarettes	number of cigarettes smoked per month
1	30.00	3	90.00
2	22.00	3	66.00
3	30.00	10	300.00
4	30.00	10	300.00
5	30.00	20	600.00
6	30.00	5	150.00
7	30.00	8	240.00
8	30.00	1	30.00
9	30.00	10	300.00
10	30.00	20	600.00
11	5.00	2	10.00
12	30.00	3	90.00
13	14.00	5	70.00
14	30.00	1	30.00
15	30.00	98	2940.00
16	30.00	20	600.00
1689	5.00	4	20.00
1690	30.00	20	600.00
1691	30.00	40	1200.00
1692	30.00	10	300.00
1693	30.00	10	300.00
1694	30.00	20	600.00
1695	30.00	10	300.00
1696	30.00	20	600.00
1697	30.00	20	600.00
1698	30.00	20	600.00
1699	30.00	10	300.00
1700	5.00	2	10.00
1701	30.00	40	1200.00
1702	30.00	5	150.00
1703	1.00	3	3.00
Total N	1703	1697	1697

Return to Video - 05. Data Management

Section 5.5: Create Secondary Variable from more than 2 Variables

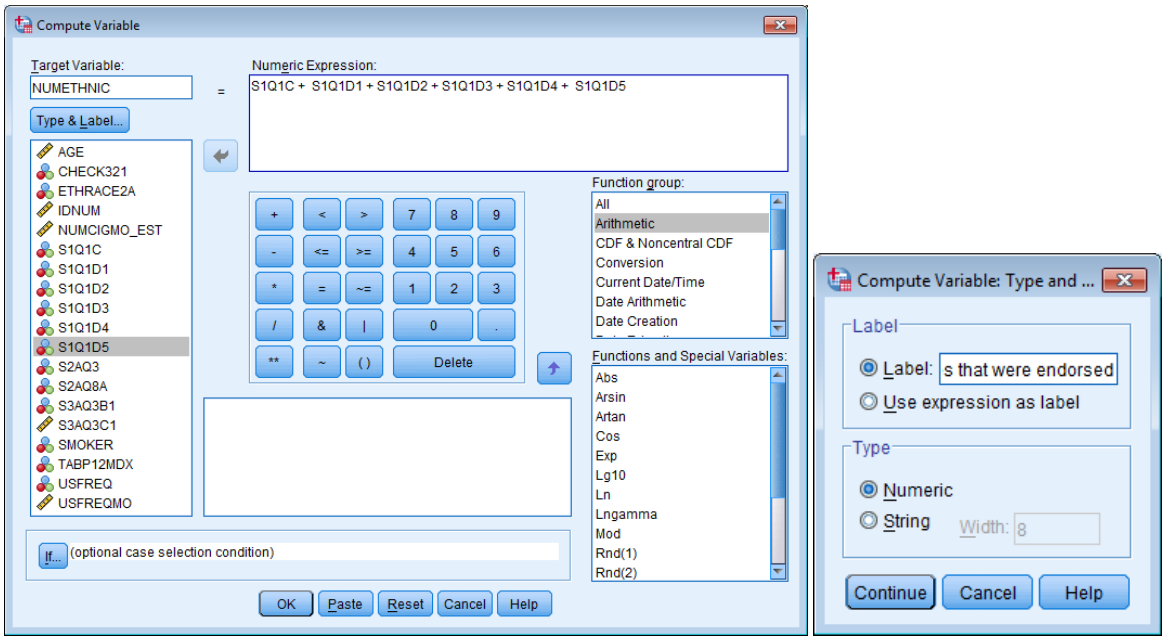
So what if you want to combine more than 2 variables? A good example of this would be creating a single secondary variable to characterize race/ethnicity from a number of separate primary variables available in the NESARC dataset. Race or ethnicity are measured by a series of questions, coded 1 if "Yes" and 2 if "No". Participants in this sample could have indicated more than one race or ethnicity, we could decide to characterize those participants who indicate multiple racial or ethnic groups separately from those who could be characterized with a single ethnicity. Below you can see the 6 variables we will need to use to create one secondary variable to represent race/ethnicity.

```
S1Q1C                HISPANIC OR LATINO ORIGIN
                     ------------------------
    8308              1.  Yes
   34785              2.  No

S1Q1D1               "AMERICAN INDIAN OR ALASKA NATIVE" CHECKED IN MULTIRACE CODE
                     ----------------------------------------------------------
    1304              1.  Yes
   41789              2.  No

S1Q1D2               "ASIAN" CHECKED IN MULTIRACE CODE
                     --------------------------------
    1334              1.  Yes
   41759              2.  No

S1Q1D3               "BLACK OR AFRICAN AMERICAN" CHECKED IN MULTIRACE CODE
                     ---------------------------------------------------
    8600              1.  Yes
   34493              2.  No

S1Q1D4               "NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER" CHECKED IN MULTIRACE CODE
                     -------------------------------------------------------------------
     363              1.  Yes
   42730              2.  No

S1Q1D5               "WHITE" CHECKED IN MULTIRACE CODE
                     --------------------------------
   32789              1.  Yes
   10304              2.  No
```

1. Add the 6 variables listed above to your Variable Set by repeating steps 5.3 #5 through #8.

2. For each of these 6 variables the response of "No" is dummy coded as a 2, but we need it to be dummy coded as a 0 so that when we sum the dummy codes for the 6 variables for each participant a sum of 2 or more would represent the participant marking "Yes" on 2 or more of the ethnicity questions. To recode these variables repeat steps 5.2 #1 through #3. Since you are doing the same recode to each of the 6 variables you can add all the variable names under **Numeric Variables** at one time then just once put in the **Old Value** for each of the 6 variables put a 2 and in the **New Value** put a 0. This will make the response of "No" dummy coded as a 0.

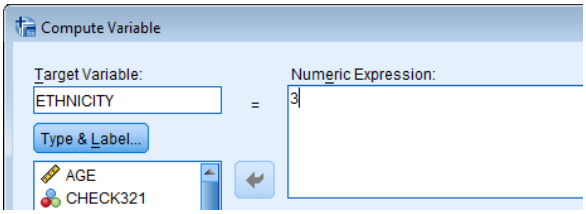
3. Next we need to sum the variables to get a new variable, which we'll call NUMETHNIC that indicates the number of race or ethnicity variables that were endorsed. Repeat steps 5.4 #1 through #4, but instead of multiplying you will add the 6 variables listed above using a "+" sign between each variable name.



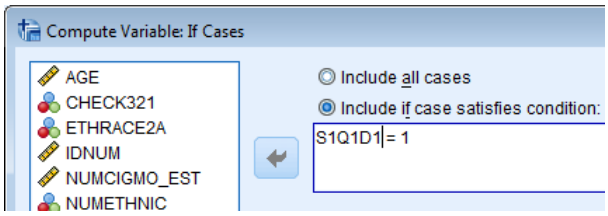
4. Add the NUMETHNIC variable to your Variable Set by repeating steps 5.3 #5 through #8. This could seem tedious, but throughout the rest of the semester this will save <u>you</u> a lot of time once we are analyzing the variables by eliminating searching through thousands of variables looking for your specific variable names.

5. Repeat steps 5.4 #1 through #3 naming the new variable ETHNICITY and properly labeling it.

6. Remove the variables currently listed under **Numeric Expression**. Move your NUMETHNIC variable from the left hand column to the **Numeric Expression** column in the upper right corner. Start with dummy code 3 by typing it into the **Numeric Expression**. Then click **If...** in the lower left hand corner.

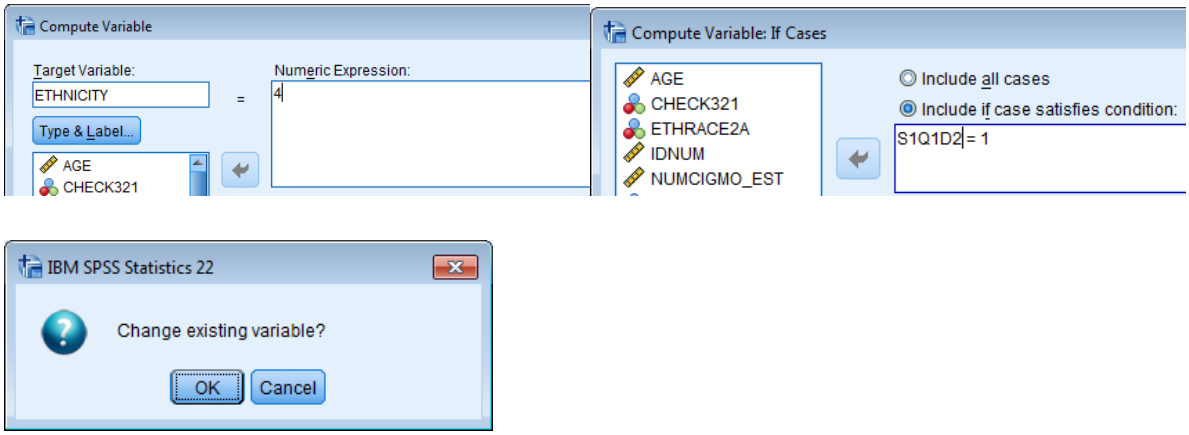


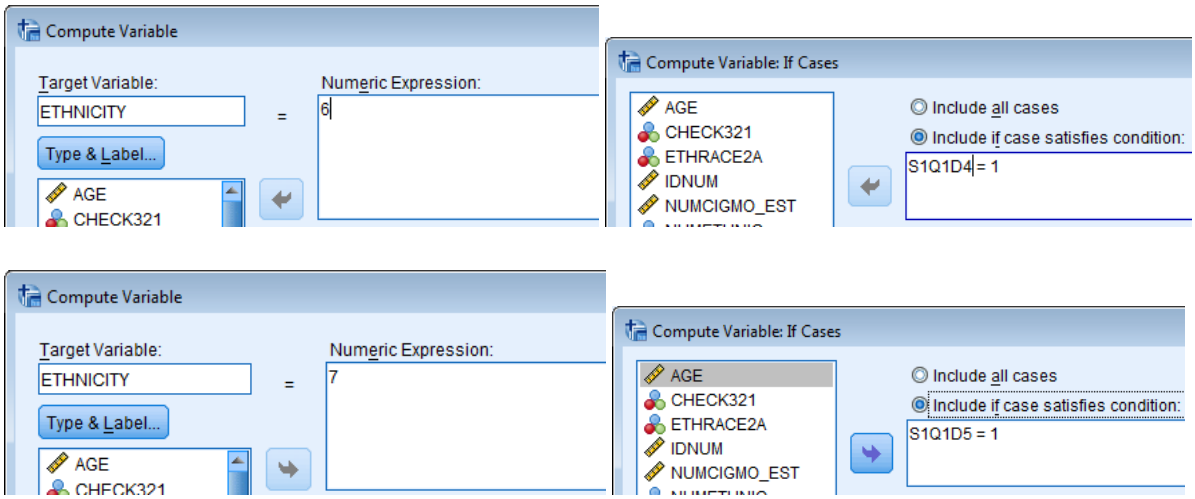
7. Click **Include if case satisfies condition:** Move the variable SC1Q1C to the upper right column from the left column using the arrow then type "**= 1**" click **Continue**.



8. Repeat steps #6 through #7 five more times, once for each of the 5 remaining variables used to create NUMETHNIC. Note it is important to start with dummy code 3, then go to 4, 5, 6, 7 then do step 9 for dummy codes 1 and 2 to work around SPSS order of operation. You will increase the dummy code value you enter in **Numeric Expression** by one with each repeat of step 5. If you refer back to the code book you will see that for each of the 6 variables we used to create NUMETHNIC a dummy code of 1 represents "Yes". We will need to tell SPSS for each of those variables dummy code of 1 to compute a different dummy code going in ascending order for the secondary variable ETHNICITY we are creating. We will start with 2 and end with 7 as 1 is already representing multiracial.

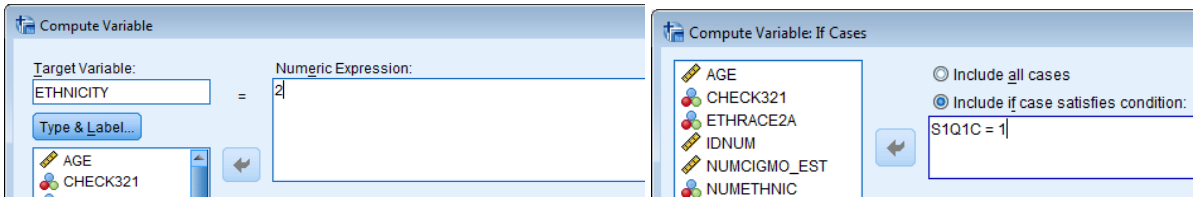
The second time you complete step #7 and #8 the below pop up window will appear each time you compute a new dummy code level for the Ethnicity variable. Click **OK** each time.





9. Repeat steps #6 & #7 using the dummy code of 1 for **Numeric Expression**. Click **Include if case satisfies condition:** type "NUMETHNIC >1" which tells SPSS that if the sum of the 6 variables we used to create NUMETHNIC is 2 or more than the new secondary variable I am creating called ETHNICITY should be dummy coded as a 1. This will represent a category of multiple racial or ethnic groups endorsed. **Click Continue > OK>OK**.

10. Repeat step 8 for the last variable used to create NUMETHNIC.



11. Add the ETHNICITY variable you just created to your Variable Set by repeating steps 5.3 #5 through #8.

12. Repeat steps 5.3 #7 through #8 using the 6 variables you used to create NUMETHNIC, as well as NUMETHNIC and ETHNICITY to check that how you intended the variable to be created is what actually happened. A portion of the table is below.

Case Summaries

	S1Q1C	S1Q1D1	S1Q1D2	S1Q1D3	S1Q1D4	S1Q1D5	Sum of Endorsed Ethnicities	Ethnicity of Participant
1	0	0	0	1	0	0	1.00	5.00
2	1	0	0	0	0	1	2.00	1.00
3	0	0	0	0	0	1	1.00	7.00
4	0	0	0	0	0	1	1.00	7.00
5	0	0	0	0	0	1	1.00	7.00
6	0	0	0	0	0	1	1.00	7.00
7	0	0	0	1	0	0	1.00	5.00
8	0	0	0	0	0	1	1.00	7.00
9	1	0	0	0	0	1	2.00	1.00
10	0	0	0	0	0	1	1.00	7.00
11	0	0	1	0	0	0	1.00	4.00
12	0	0	0	1	0	0	1.00	5.00
13	0	0	0	0	0	1	1.00	7.00
14	1	0	0	0	0	1	2.00	1.00
15	0	0	0	1	0	0	1.00	5.00
16	1	0	0	0	0	1	2.00	1.00
17	0	0	0	0	0	1	1.00	7.00
18	1	0	0	0	0	1	2.00	1.00
19	1	0	0	0	0	1	2.00	1.00
20	0	0	0	0	0	1	1.00	7.00
21	0	0	0	0	0	1	1.00	7.00
22	1	0	0	0	0	1	2.00	1.00

You can see in the table below we now have one variable that represents all ethnicities.

Ethnicity of Participant

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	49	2.9	2.9	2.9
	2.00	335	19.6	19.6	22.5
	3.00	16	.9	.9	23.4
	4.00	37	2.2	2.2	25.6
	5.00	199	11.7	11.7	37.3
	6.00	10	.6	.6	37.9
	7.00	1060	62.1	62.1	100.0
	Total	1706	100.0	100.0	

Return to Video - 05. Data Management

Section 5.6: Collapsing a Variable

Once you've created the secondary variables that you feel you will need, you can also consider whether any of your quantitative variables or categorical variables need to be further grouped or binned.

As an example, we're going to show you how we might choose to group age in the NESARC sample. Rather than looking at age as a quantitative variable, we want to compare age groups categorically. We could examine the number of observations at each age and make some decisions about how we could best make groups.

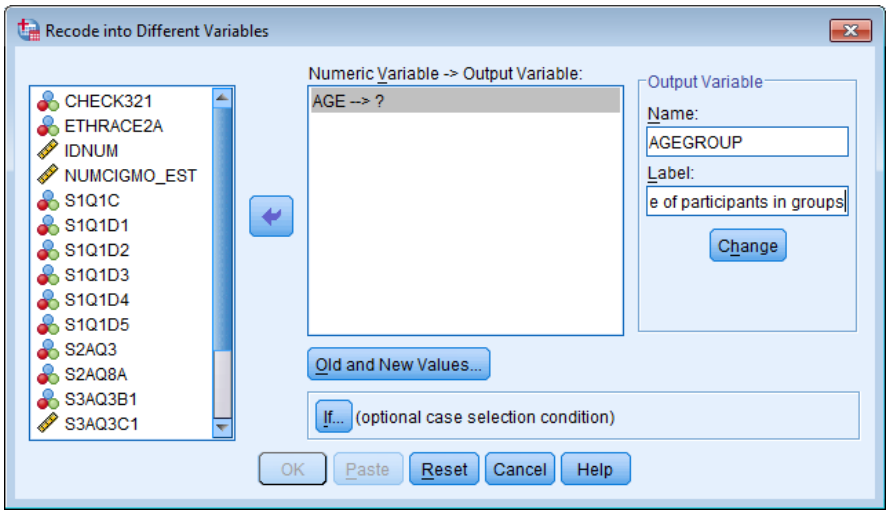
Here's the frequency table for age in our young adult past year smoking sample. The cumulative percent column displays percentiles that can help you divide the sample into a specific number of equal parts. If we wanted 4 age groups, we would look for cumulative percentages that are closest to 25%, 50%, and 75%. If we wanted three age groups, we would look for cumulative percentages that are closest to 33% of the sample, and 66% of the sample.

Age of Participant

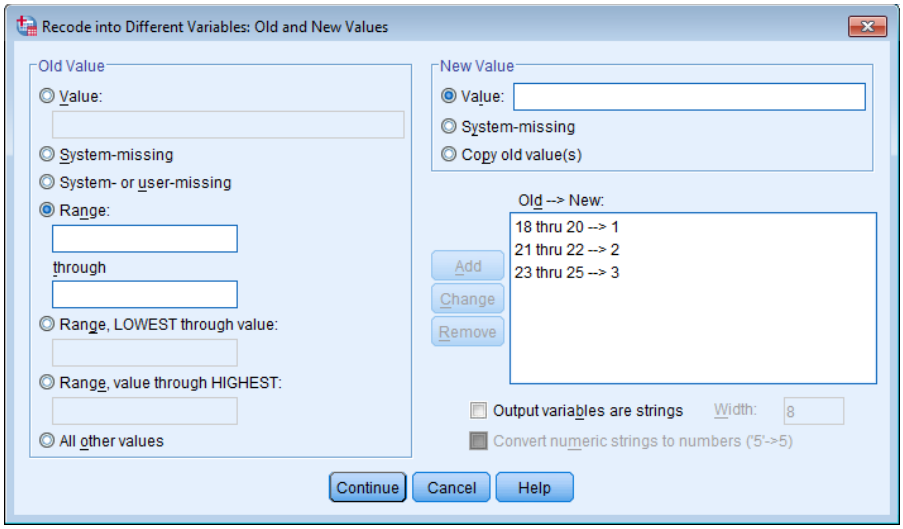
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18	161	9.4	9.4	9.4
	19	200	11.7	11.7	21.2
	20	221	13.0	13.0	34.1
	21	239	14.0	14.0	48.1
	22	228	13.4	13.4	61.5
	23	231	13.5	13.5	75.0
	24	241	14.1	14.1	89.2
	25	185	10.8	10.8	100.0
	Total	1706	100.0	100.0	

Let's create a categorical variable that divides the sample into roughly three equal size age groups. We chose the age cut points of 20 and 22 because these correspond roughly to the 33rd and 66th percentile.

1. Repeat steps 5.3 #1 and #2. We will name this new variable AGEGROUP. Remember to remove what is currently in the middle column, **Input Variable --> Output Variable**, using the arrow to move the variables back to the left column.



2. Repeat steps 5.3 #3 and #4. Remember to remove the **Old --> New** values in the lower right column before beginning. Click **Continue > OK**.



3. Repeat steps 5.3 #5 through #8 to add the AGEGROUP variable you just created to your Variable Set.

4. It is important to always verify that how you intended to collapse a variable worked properly. Run a frequency distribution on AGE and AGEGROUP. Confirm that the when you add the frequency of each level (i.e., Age) you collapsed into agroup (dummy codes 1, 2, and 3) match. When we add 161 + 200 + 221 = 582 which is category 1 for our new categorical variable AGEGROUP. You should verify each level this way as it is possible that one level collapsed correctly and another did not.

Age of Participant

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18	161	9.4	9.4	9.4
	19	200	11.7	11.7	21.2
	20	221	13.0	13.0	34.1
	21	239	14.0	14.0	48.1
	22	228	13.4	13.4	61.5
	23	231	13.5	13.5	75.0
	24	241	14.1	14.1	89.2
	25	185	10.8	10.8	100.0
	Total	1706	100.0	100.0	

Age of participants in groups

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	582	34.1	34.1	34.1
	2.00	467	27.4	27.4	61.5
	3.00	657	38.5	38.5	100.0
	Total	1706	100.0	100.0	

Now that you have seen a number of different data management decisions that could be made, I hope that you're ready to take on your own which you will do in the next assignment. How much data management you will need for your own variables will depend on the variables that you've selected and the decisions that you make about them. Some may choose to only code out missing data, while others may need to or want to do more. Data management is a part of the research process that you can and will return to again and again as you learn more and are able to make better decisions.