Introduction to Numeric Data Cleaning Using SPSS

McGill Library Data Lab
March 13, 2018
Berenica Vejvoda, Numeric Data Librarian, McGill University
berenica.vejvoda@mcgill.ca
514-398-3702

Objectives:

The main purpose of this session is to introduce some of the most basic data management functions using one of the most popular proprietary statistical software platforms: SPSS.

This session will also introduce the concept of micro-level data and the importance of using data documentation while working with micro-datasets.

By the end of the workshop you will be able to:

I. Use the SPSS environment

a) Navigate in SPSS using menus and dialogs
b) Use the SPSS Syntax Editor and SPSS Commands
c) Import data into SPSS

II. Manage your data

a) Remove Missing Values
b) Data aggregation
c) Transpose your data
d) Split file for further analysis (split file and select cases)

III. Transform your data

a) Recode variables
b) Create subset of a dataset
c) Combine 3 variables into 1

I. SPSS environment

Open SPSS (All programs -> IBM SPSS Statistics).

a) Navigate in SPSS using menus and dialogs

This will bring you to the Data Editor window. The Data Editor Window has two views. Data View and Variable View. You can switch between the two views by choosing the appropriate tabs in the left bottom corner of the screen.

The Data View displays the data and the Variable View contains information about the data (e.g., variable names, labels).

The Data Editor window contains the following menus that can be accessed on the top of the screen:

File – used to open existing files, read data files, save data files, and exit SPSS
Edit – copy, cut and paste functions
View – allows you to switch between data and variable view and hide or show toolbars/status bars
Data – used to perform various functions on your data (define variables, insert variables or cases, sort cases, transpose, merge files, aggregate, split files, select cases and replace missing values
Transform – used to perform computations on variables, to recode variables, count and rank cases, replace missing values, anonymize variables, etc.
Analyze – used to perform statistical analyses
Graphs – used to generate different graphs/charts

Most data entry, data manipulation and data analysis can be conducted in SPSS using pull-down menus.

A few scenarios exist where using the pull-down menu is **NOT** always the best strategy.

- You have to conduct the same analysis with similar datasets
- After running your analysis, you find a typo in the data and have to re-run analyses again
- Six months after submitting your study to a journal, you've got suggestions from the reviewers of your paper to modify some of the analyses of your data

b) Use the SPSS Syntax Editor and SPSS Commands

Using the **SPSS Syntax Editor** instead of the pull-down menus allows you to deal with these and other issues quickly and efficiently. Working with the Syntax Editor also allows you to keep a precise log of your work with SPSS and run the program as many times as needed. It also allows you to make minor changes for various analyses and then re-run the program.

To start using syntax, you simply need to **click "Paste" instead of OK** in any dialogue window. To run analyses from syntax, select syntax lines and click the "Run" button on the toolbar. You can

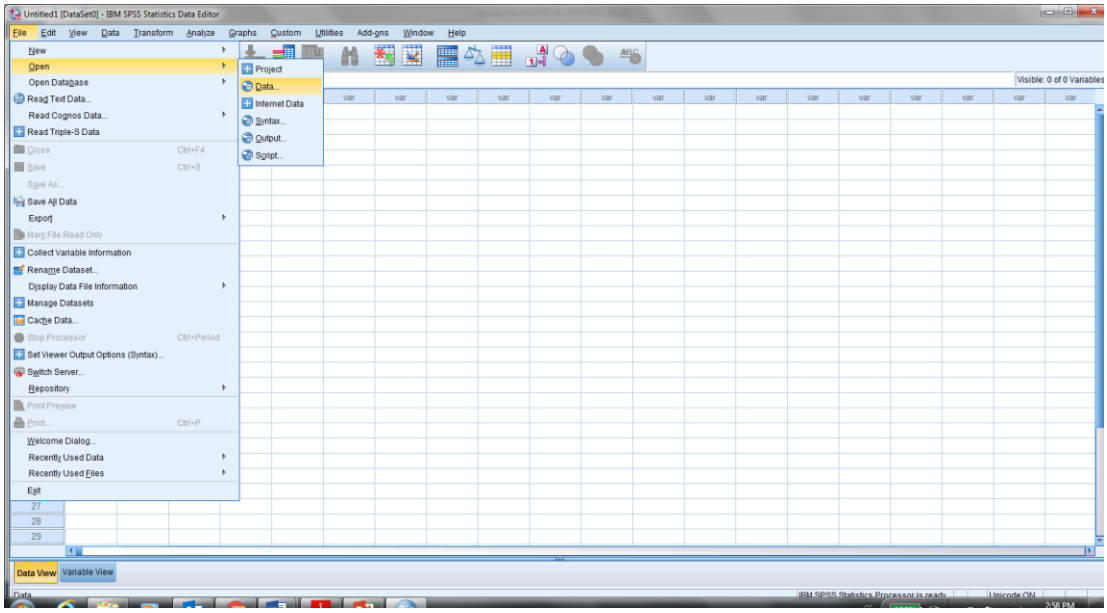
also customize your syntax by adding comments. Every comment should begin with * and end with a period.

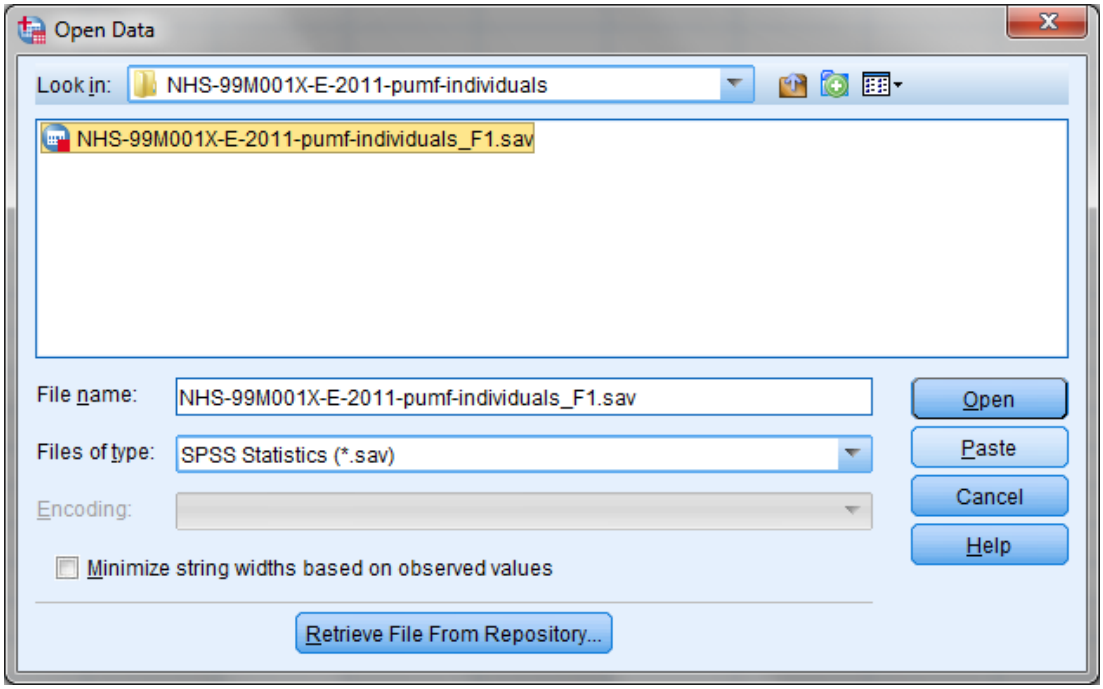
When working with Syntax, you will work in three windows in SPSS: Data Editor (.sav), Syntax (.sps), and Output (.spv). Each type of window saves a file with a different file extension.

The output window displays the results of your analyses. When you just manage and manipulate your data, you won't see any output except syntax commands.

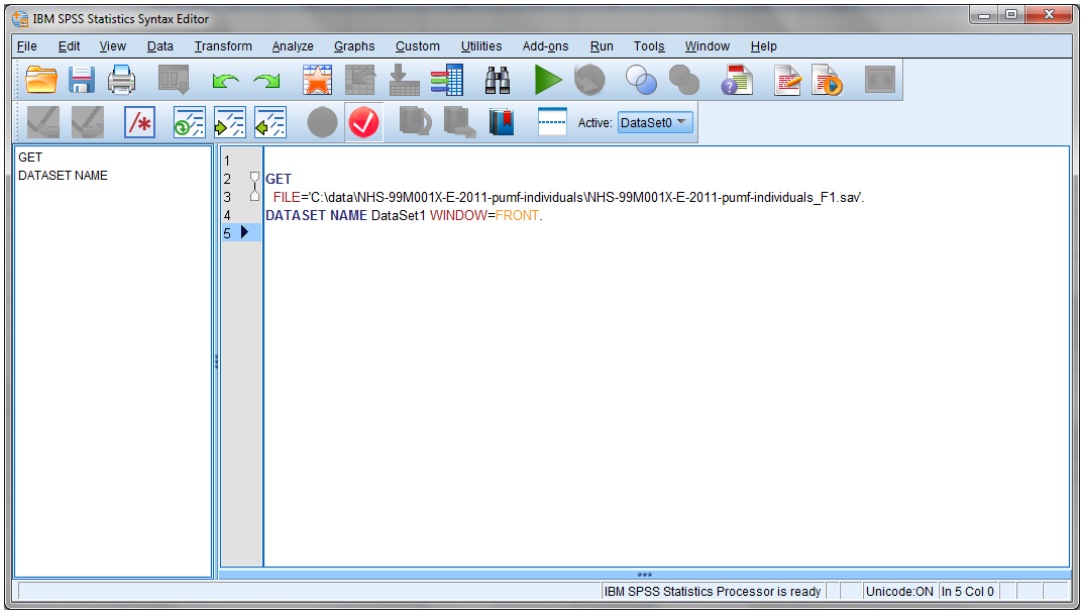
c) Import data into SPSS

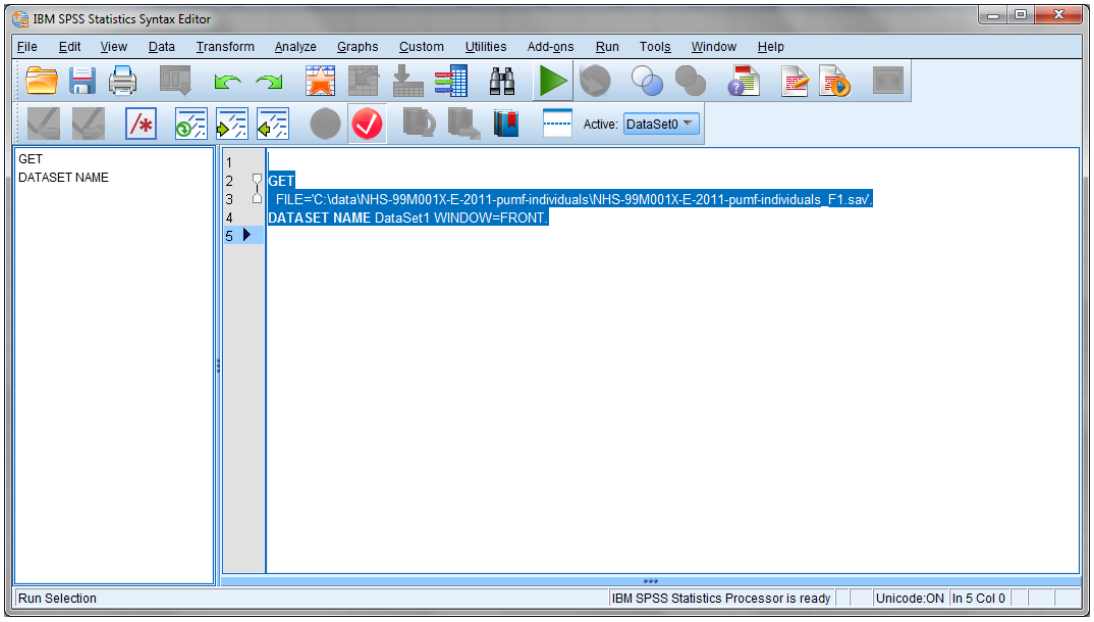
Open the SPSS data file (NHS-99M001X-E-2011-pumf-individuals_F1.sav) by going to **File -> Open -> Data…**, find the location of the file on your workstation and click "Paste". This will save the SPSS command for opening the file in the syntax.





In the Syntax Window, select the commands (CTRL-A) and Run them (Green arrow).

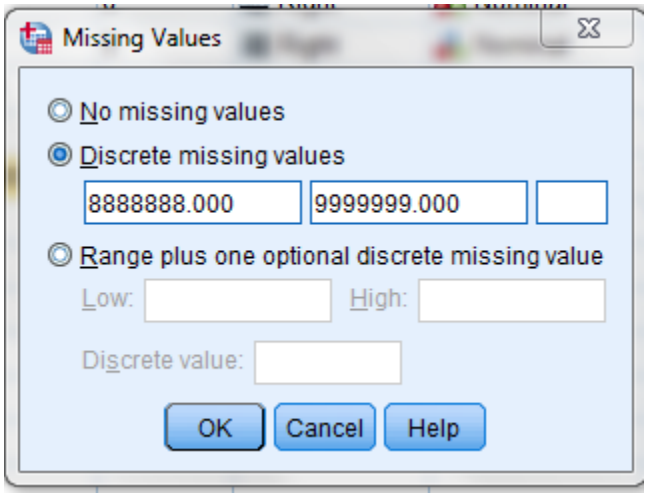




II) Manage your Data

a) Remove Missing Values

The data we are working with already have what are referred to as "user-defined" missing values. In other words, Statistics Canada has declared missing values in the dataset. When you **click on Variable View** and click on one of the missing cells for a particular variable you will see the assigned values for the missing variables.

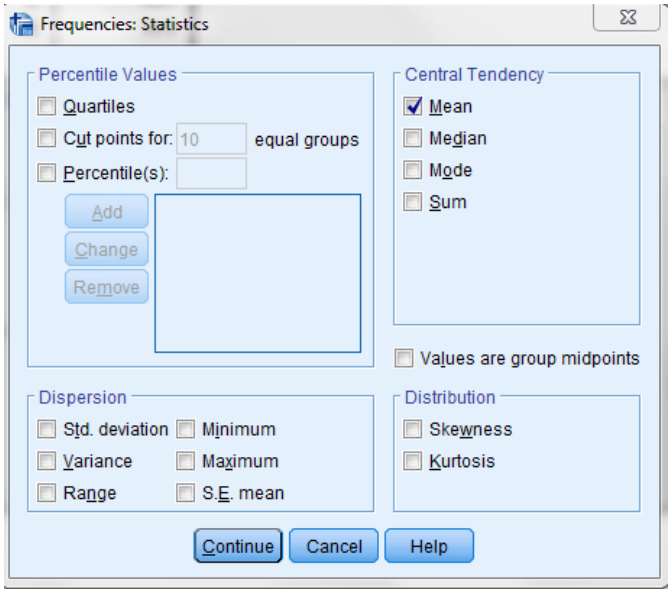


If the variables are explicitly declared (specified) as missing, SPSS automatically excludes them from your calculations.

Let's test that this is the case by running some descriptive analyses.

Click on Analyze -> Descriptive Statistics -> Frequencies…

Drag the Income: Total Income variable to the Variable(s) window. Click on **Statistics…** and select **mean**. Run your analysis.



You will see that the mean is 40341.61.

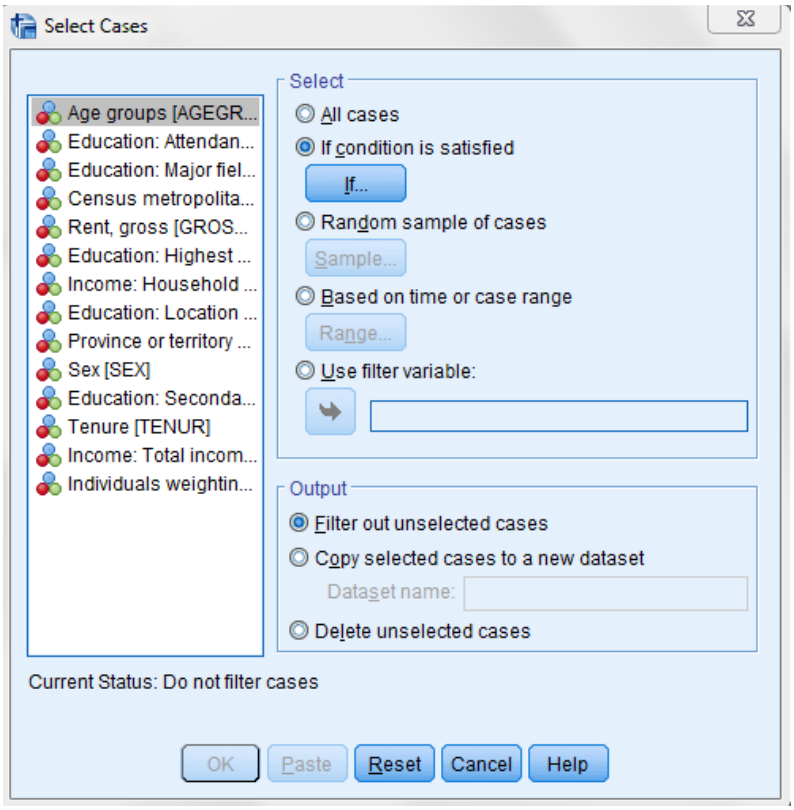
Frequencies

Statistics

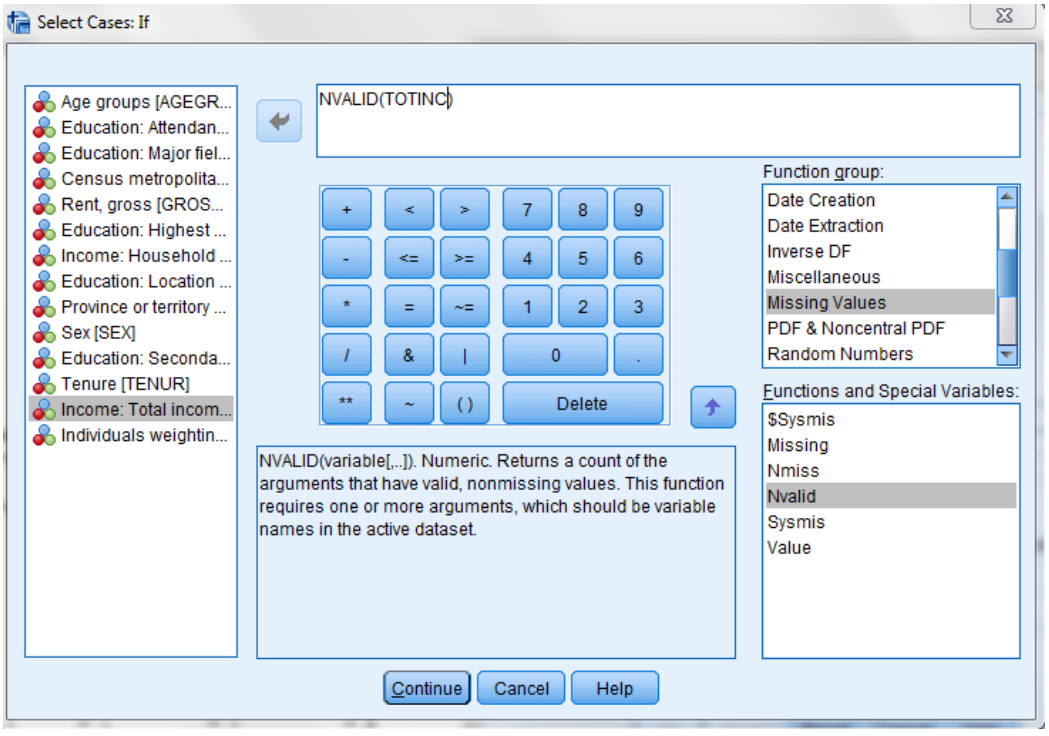
Income: Total income

N	Valid	734079
	Missing	152933
Mean		40341.61

Now let's test that this mean is the same if we select only the valid cases**. Click on Data -> Select Cases…** Select **If condition is satisfied** and click on **If…**



Under Function group double click on **Missing Values** and choose the **NValid** function by double-clicking on it.



Finally, double-click on the Income: Total Income variable and click on Continue and then OK to run the function.

You will see a new filter_$ column with 1's representing all the valid cases and excluding the missing values. And the left-hand side will show the filtered out cases.

	AGEGRP	ATTSCH	CIP2011	CMA	GROSRT	HDGREE	HHINC	LOCSTUD	PR	SEX	SSGRAD	TENUR	TOTINC	WEIGHT	filter_$
1	15	1	10	462	9999	9	1	5	24	1	8	1	1000	32.39361116	1
2	11	1	10	505	1100	9	18	6	35	2	8	2	68000	32.39361116	1
3	9	1	7	825	1300	8	22	16	48	1	7	2	20000	62.94483734	1
4	13	1	13	535	9999	2	21	99	35	1	4	1	84000	32.39361116	1
5	13	1	13	999	800	1	13	99	35	2	1	2	12000	32.39361116	1
6	13	1	5	999	9999	9	24	5	24	2	8	1	63000	32.39361116	1
7	10	1	88	999	9999	88	28	10	59	2	88	1	29000	32.39361116	1
8	14	1	13	999	9999	2	21	99	13	1	4	1	31000	32.39361116	1
9	12	8	4	505	900	9	11	13	35	2	8	2	34000	32.39361116	1
10	9	1	5	999	9999	8	10	10	59	2	7	1	28000	32.39361116	1
11	3	9	99	462	9999	99	22	99	24	2	99	1	9999999	32.39361116	0
12	9	1	11	462	9999	3	26	5	24	2	5	1	63000	32.39361116	1
13	12	1	2	535	800	6	1	6	35	2	6	2	1000	32.39361116	1
14	10	1	1	999	9999	9	32	9	48	1	8	1	110000	32.39361116	1
15	9	1	13	933	9999	2	27	99	59	1	4	1	30000	32.39361116	1
16	1	9	99	462	9999	99	20	99	24	2	99	1	9999999	116.7244234	0
17	9	2	10	999	9999	8	24	5	24	1	7	1	18000	32.39361116	1
18	10	1	8	933	2323	12	30	14	59	1	11	2	79000	197.1938674	1

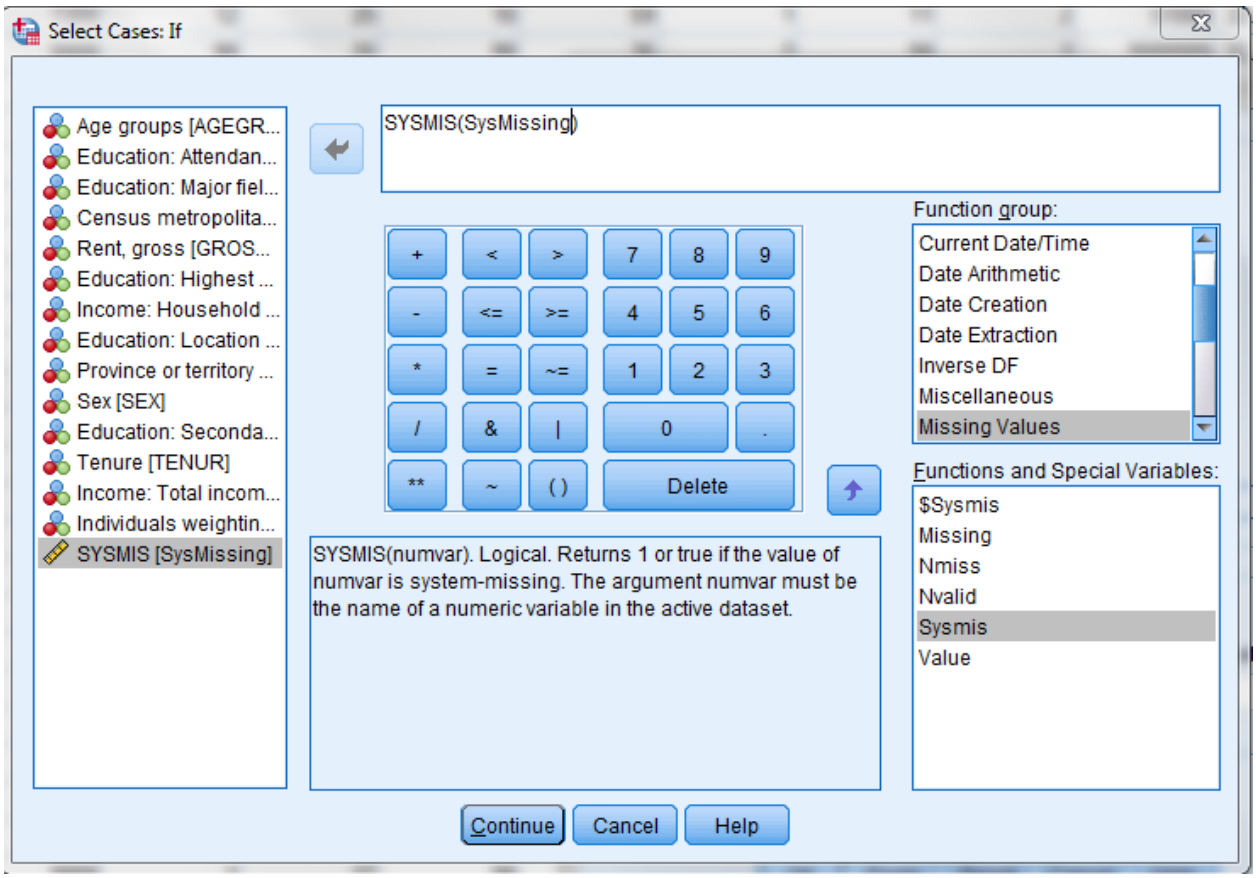
Re-run the descriptive analysis for Mean for TOTINC.

You should get the same Mean = 40341.61.

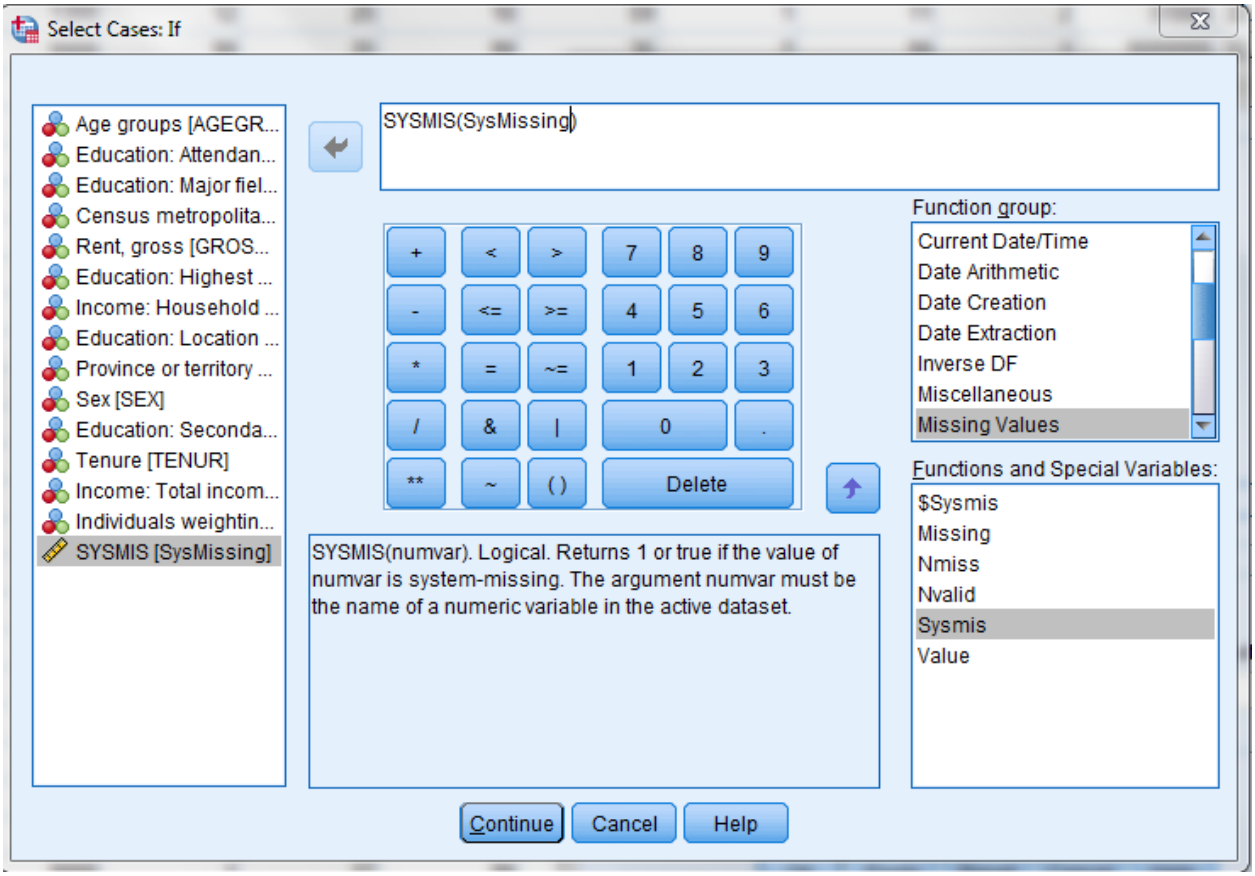
To turn the FILTER off go to **File -> New -> Syntax…** and type **FILTER OFF**. Select the code and Run it.

Finally, you will not always have missing values declared by data producers. Messier data will require that you declare what are called system-missing values.

To do this, go **to Data -> Select Cases…If…** and then again select **Missing Values** from the Function group window. Now double-click on **SysMis** and double-click the **SysMissing** variable.

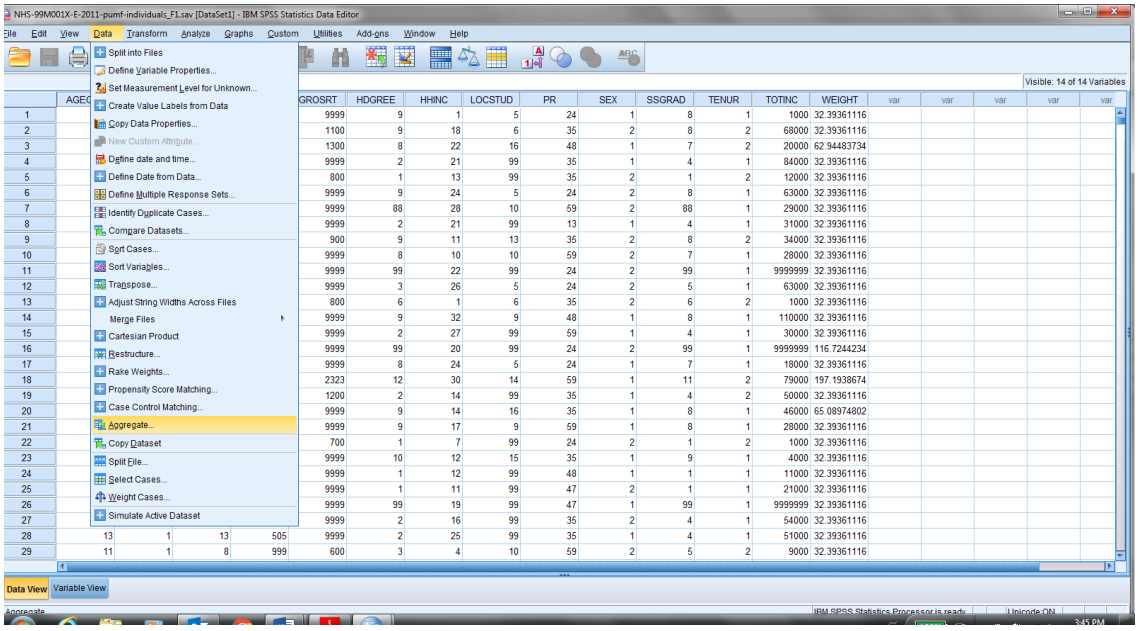


You will see now that SPSS has flagged all missing values as 1. You could then export those missing values into a new dataset or just have SPSS delete the cases that are missing from the current dataset. Alternatively, you could run NValid and just have SPSS flag the valid cases.



b) Data aggregation

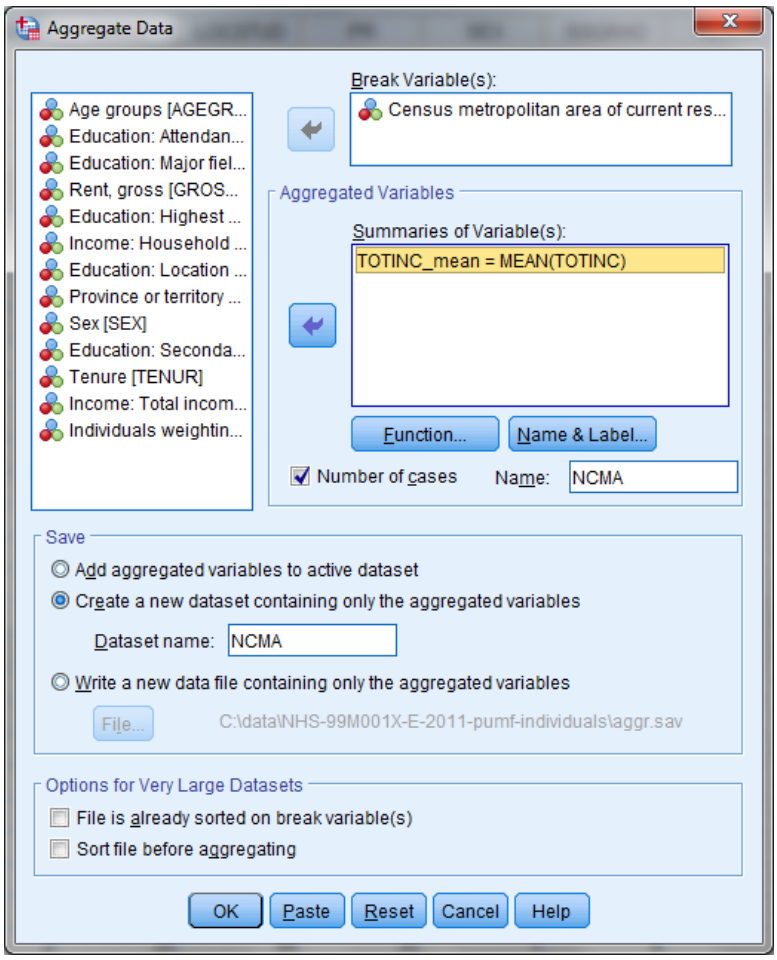
Click on **Data -> Aggregate**



Select **Census metropolitan area** and move it to the **Break Variable(s)** window. Select **Income: Total Income** and move it to the **Aggregate Variable(s)** window.

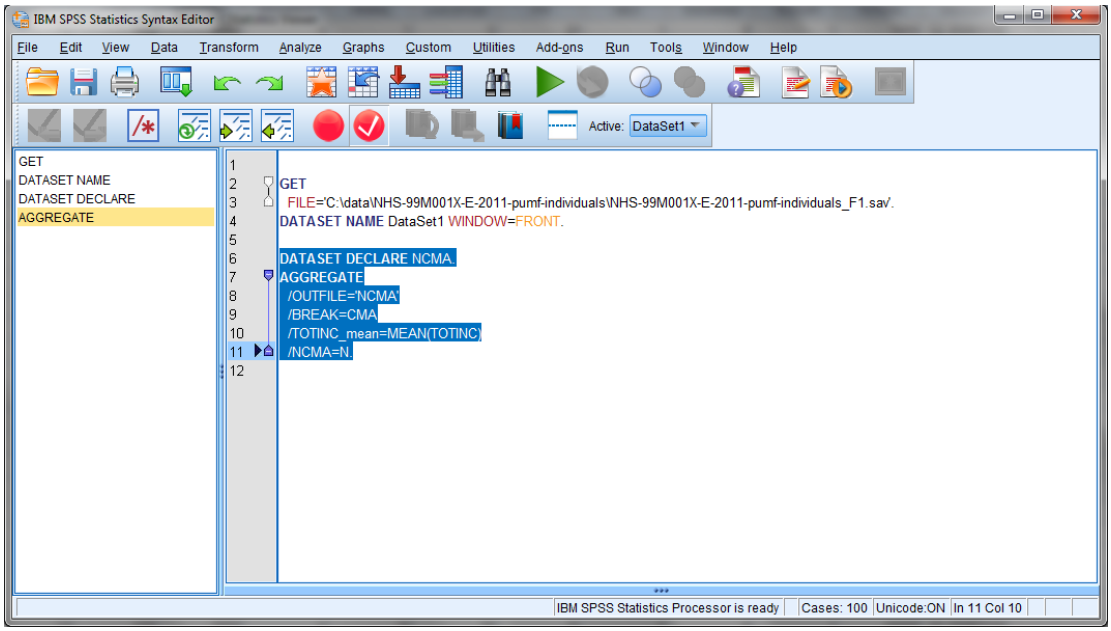
Click on **Number of Cases** and give this variable a name.

Click on the File button next to the **Write a new dataset containing only the aggregated variables** and give the name to the new file.

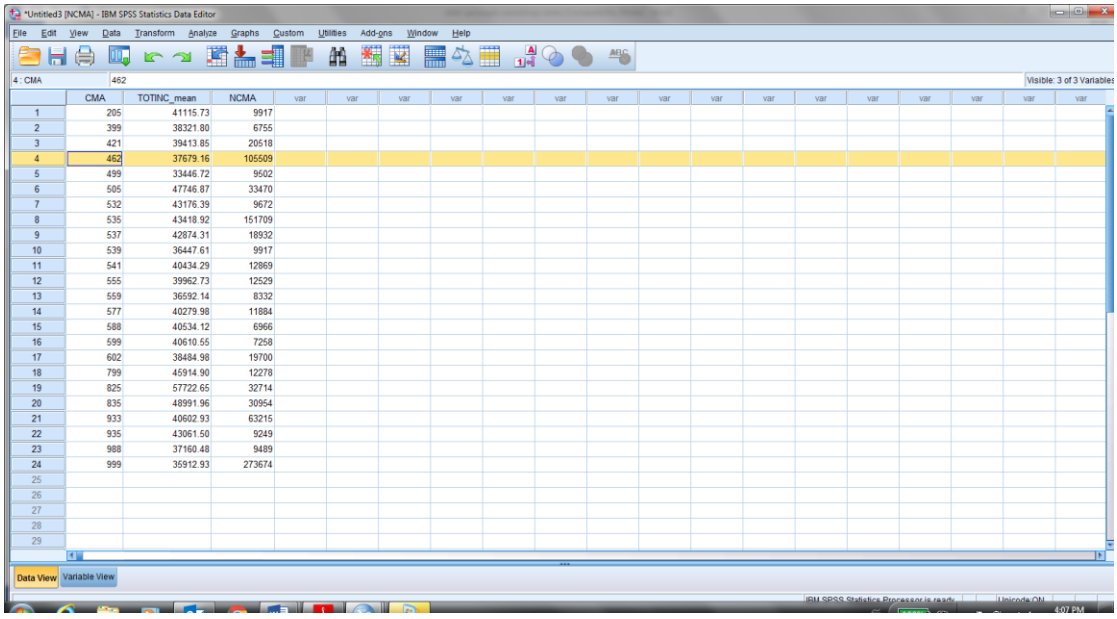


Click on **Paste**.

Go to the Syntax Editor Window, select the command and run it.



What is the average total individual income for Montreal?



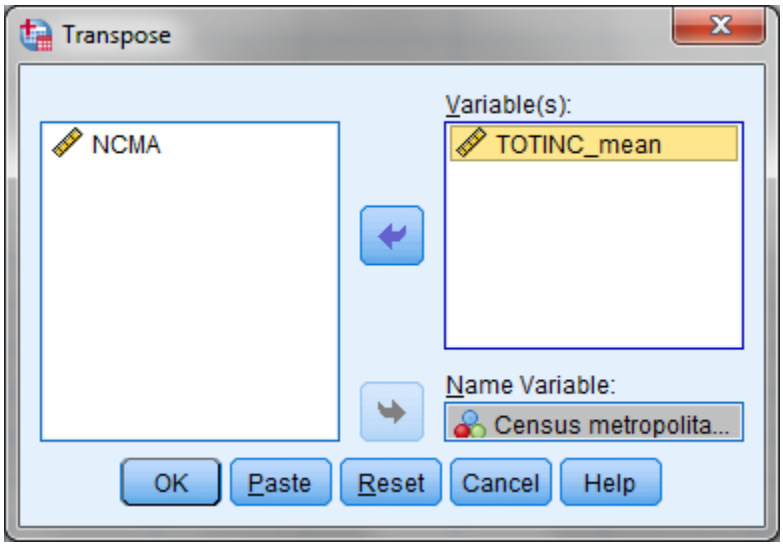
c) Transpose your data

Sometimes you may have the need to restructure your data from long to wide format or vice versa in order to run an analysis properly.
Transposing your data creates a new data file in which the rows and columns in the original data file are transposed so that cases (rows) become variables and variables (columns) become cases.

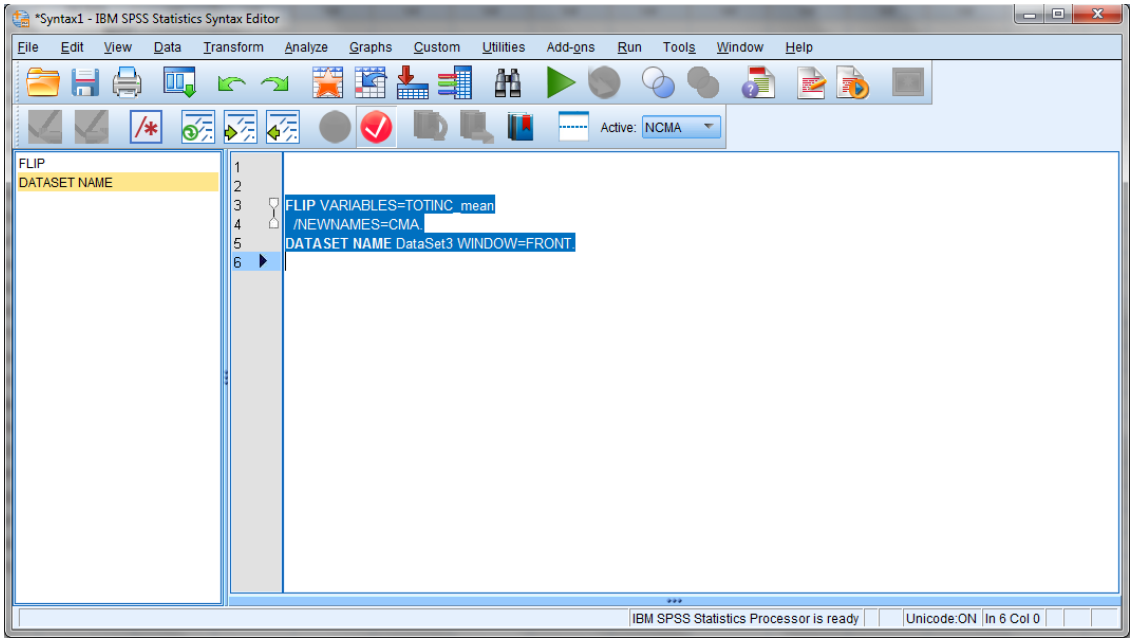
Person	Gender	Score
1	Male	5
2	Female	8
3	Female	10

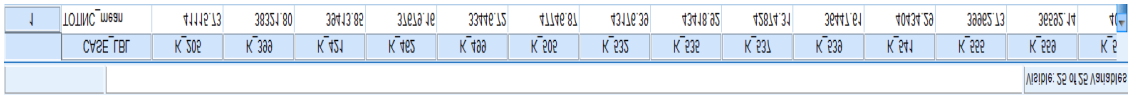
Person	1	2	3
Gender	Male	Female	Female
Score	5	8	10

Using the newly created aggregate dataset, click on **Data** then **Transpose**. Select the Mean individual income variable and move it to the **Variable(s)** window. Then move the Census Metropolitan Area variable to the **Name Variable** window.



Then click **Paste**.



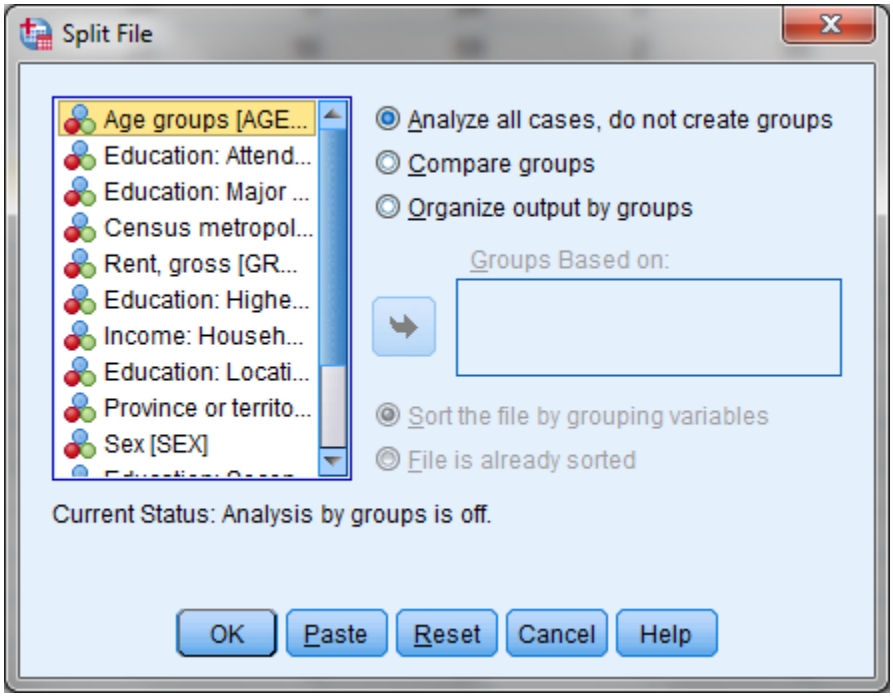


d) Split file for further analysis

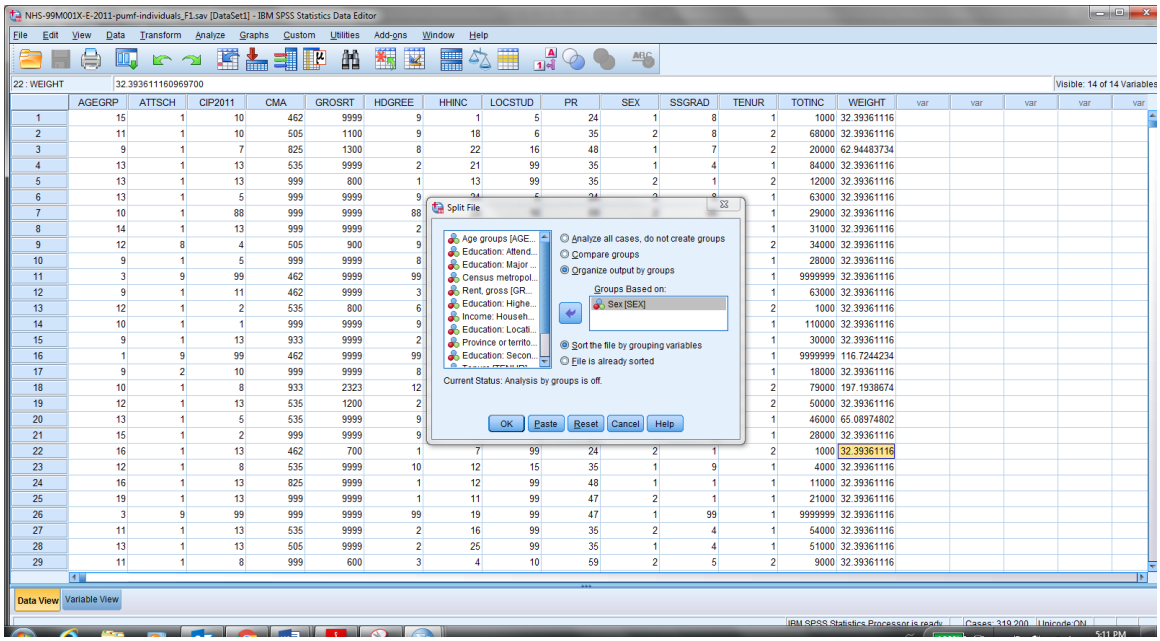
Split files into groups using gender (sex variable) as criterion

If your data file contains sub-samples of cases representing specific demographic groups (e.g. gender, age group, income group), you can run separate analyses to explore the characteristics of each subsample. You just need to tell SPSS to split the data file before running your analysis.

Go to the main NHS dataset and click on **Data** -> **Split File**. This brings up the following dialog box:



Choose SEX and move it to the "Groups Based on" box. Select the radio button for "Organize output by groups".

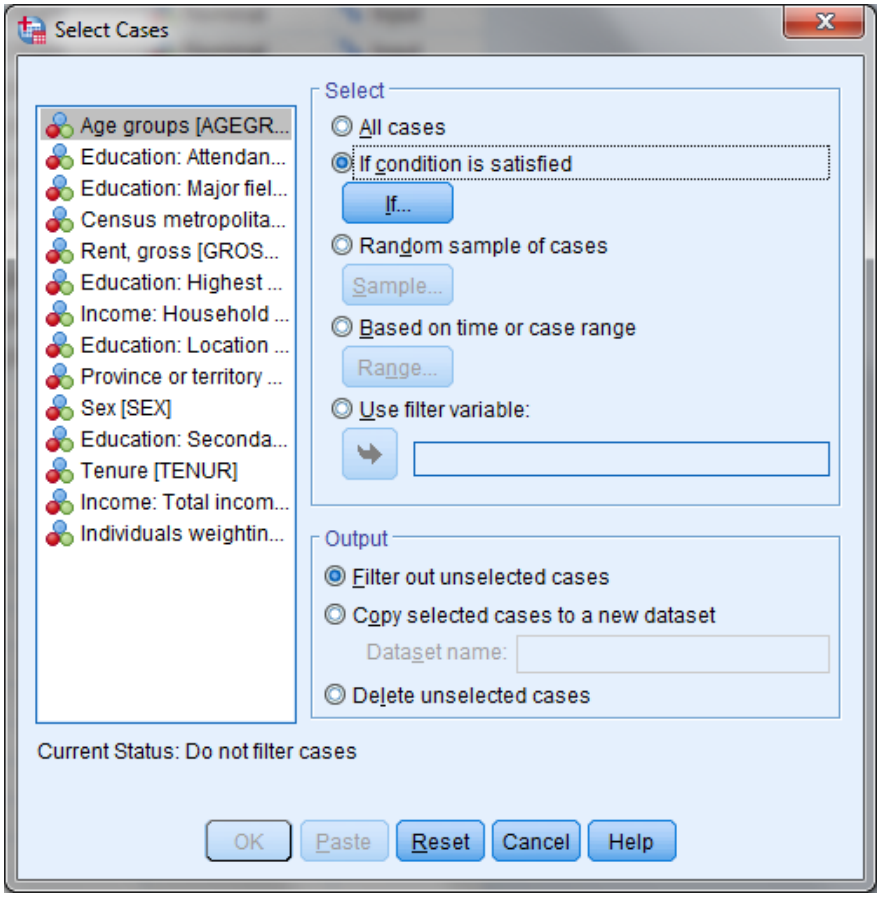


Running your analysis for a selected group of cases

You might also be interested in running some analyses only for a selected group of cases *with certain characteristics*.

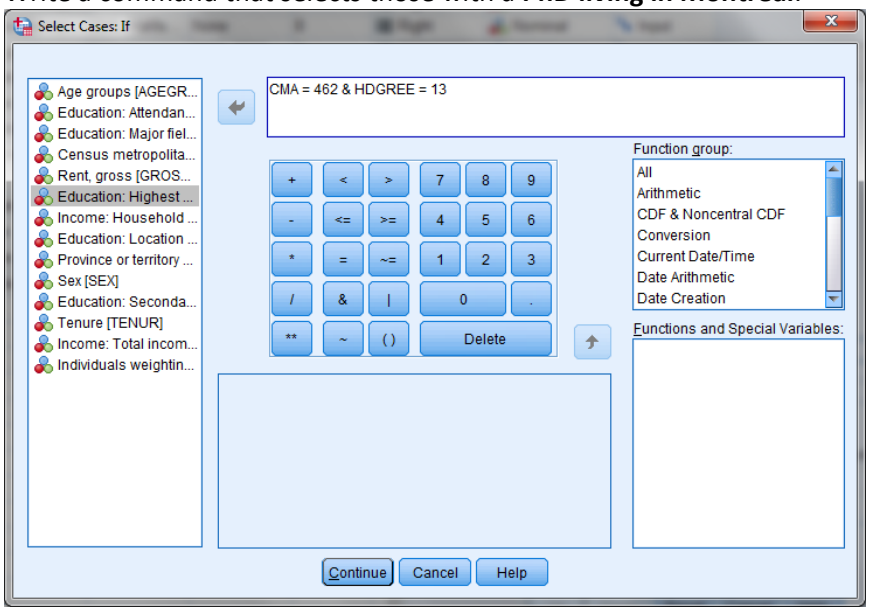
In this case, you will need to use the SPSS command called **select cases**.

To select cases, follow **Data -> Select cases**. Select **If condition is satisfied.**

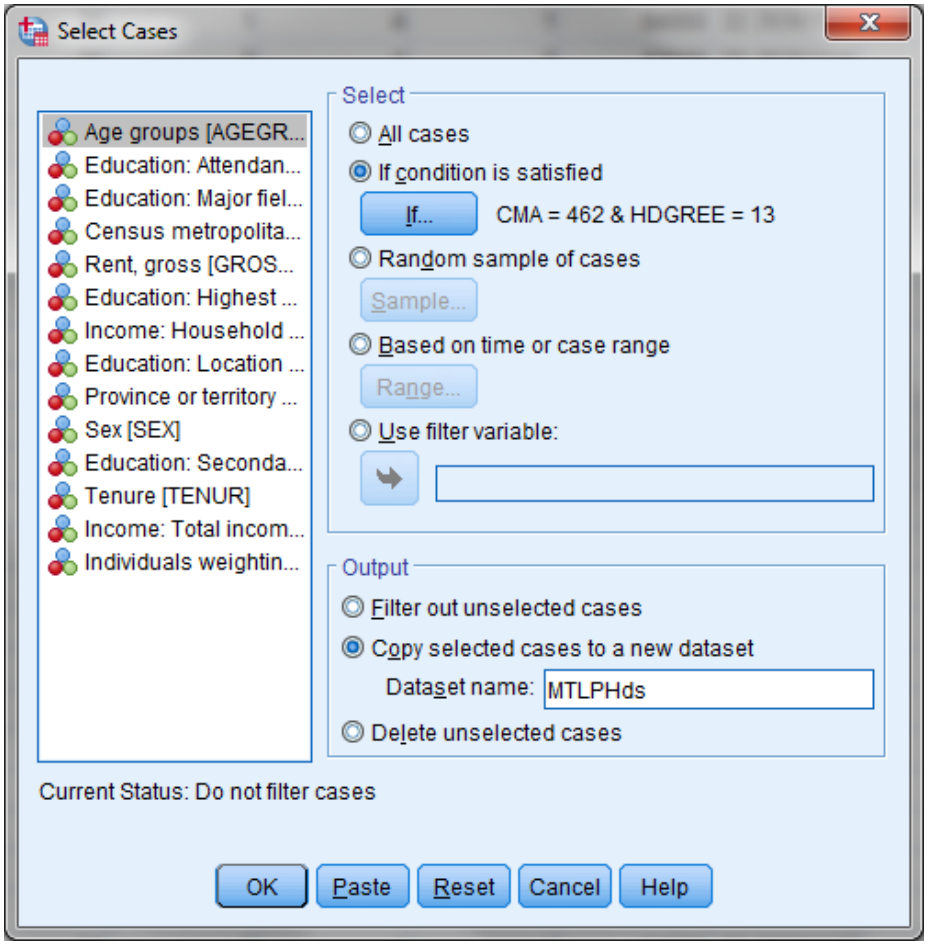


Then click on **If…**

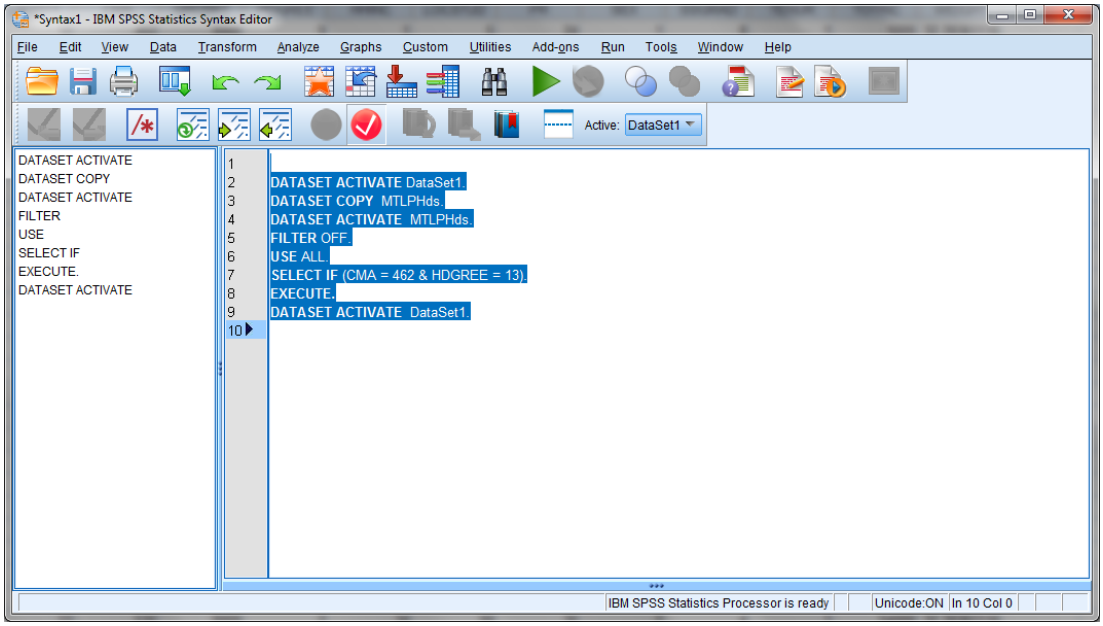
Write a command that selects those with a **PhD living in Montreal**.

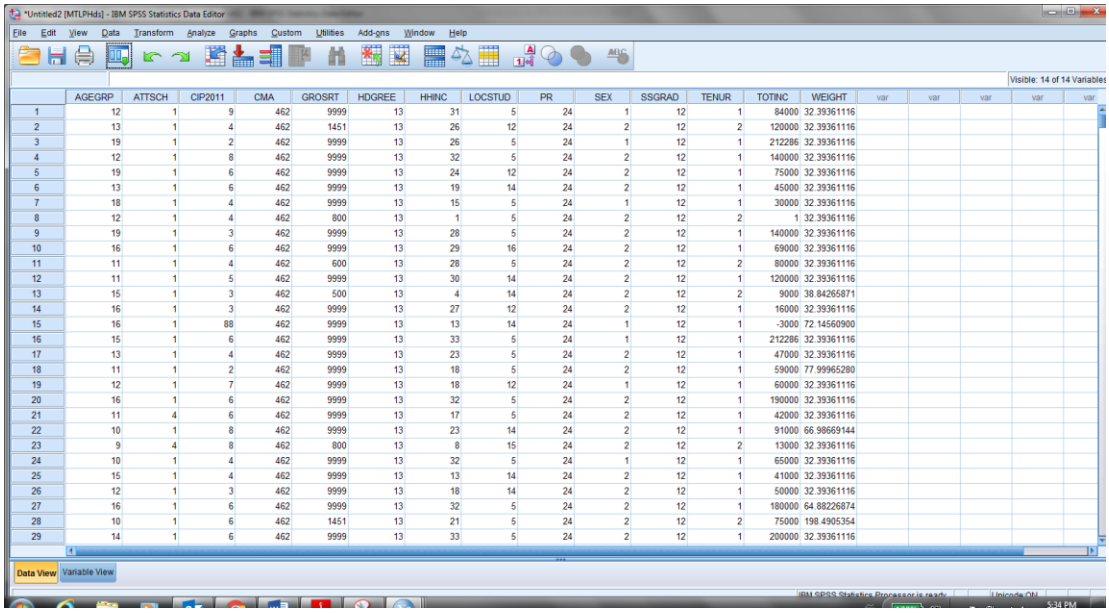


Click on **Continue**.



Choose **Copy selected cases to a new dataset**, give it a new name (PhDMTL), then click on **Paste** and run the syntax in the Syntax Window.


*Syntax1 - IBM SPSS Statistics Syntax Editor
File Edit View Data Transform Analyze Graphs Custom Utilities Add-ons Run Tools Window Help
Active: DataSet1
DATASET ACTIVATE
DATASET COPY
DATASET ACTIVATE
FILTER
USE
SELECT IF
EXECUTE.
DATASET ACTIVATE
1 DATASET ACTIVATE DataSet1.
2 DATASET COPY MTLPHds.
3 DATASET ACTIVATE MTLPHds.
4 FILTER OFF.
5 USE ALL.
6 SELECT IF (CMA = 462 & HDGREE = 13).
7 EXECUTE.
8 DATASET ACTIVATE DataSet1.
9
10
IBM SPSS Statistics Processor is ready
Unicode:ON In 10 Col 0


*Untitled2 [MTLPHds] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Custom Utilities Add-ons Window Help
Visible: 14 of 14 Variables
AGEGRP ATTSCH CIP2011 CMA GROSRT HDGREE HHINC LOCSTUD PR SEX SSGRAD TENUR TOTINC WEIGHT
1 12 1 9 462 9999 13 31 5 24 1 12 1 84000 32.39361116
2 13 1 4 462 1451 13 26 12 24 2 12 2 120000 32.39361116
3 19 1 2 462 9999 13 26 5 24 1 12 1 212286 32.39361116
4 12 1 8 462 9999 13 32 5 24 2 12 1 140000 32.39361116
5 19 1 6 462 9999 13 24 12 24 2 12 1 75000 32.39361116
6 13 1 6 462 9999 13 19 14 24 2 12 1 45000 32.39361116
7 18 1 4 462 9999 13 15 5 24 1 12 1 30000 32.39361116
8 12 1 4 462 800 13 1 5 24 2 12 2 1 32.39361116
9 19 1 3 462 9999 13 28 5 24 2 12 1 140000 32.39361116
10 16 1 6 462 9999 13 29 16 24 2 12 1 69000 32.39361116
11 11 1 4 462 600 13 28 5 24 2 12 2 80000 32.39361116
12 11 1 5 462 9999 13 30 14 24 2 12 1 120000 32.39361116
13 15 1 3 462 500 13 4 14 24 2 12 2 9000 38.84265871
14 16 1 3 462 9999 13 27 12 24 2 12 1 16000 32.39361116
15 16 1 88 462 9999 13 13 14 24 1 12 1 -3000 72.14560900
16 15 1 6 462 9999 13 33 5 24 1 12 1 212286 32.39361116
17 13 1 4 462 9999 13 23 5 24 2 12 1 47000 32.39361116
18 11 1 2 462 9999 13 18 5 24 2 12 1 59000 77.99965280
19 12 1 7 462 9999 13 18 12 24 1 12 1 60000 32.39361116
20 16 1 6 462 9999 13 32 5 24 2 12 1 190000 32.39361116
21 11 4 6 462 9999 13 17 5 24 2 12 1 42000 32.39361116
22 10 1 8 462 9999 13 23 14 24 2 12 1 91000 66.98669144
23 9 4 8 462 800 13 8 15 24 2 12 2 13000 32.39361116
24 10 1 4 462 9999 13 32 5 24 1 12 1 65000 32.39361116
25 15 1 4 462 9999 13 13 14 24 2 12 1 41000 32.39361116
26 12 1 3 462 9999 13 18 14 24 2 12 1 50000 32.39361116
27 16 1 6 462 9999 13 32 5 24 2 12 1 180000 64.88226874
28 10 1 6 462 1451 13 21 5 24 2 12 2 75000 198.4905354
29 14 1 6 462 9999 13 33 5 24 2 12 1 200000 32.39361116
Data View
Variable View

III) Transform your data

a) Recode variables (1 variable to another new variable)

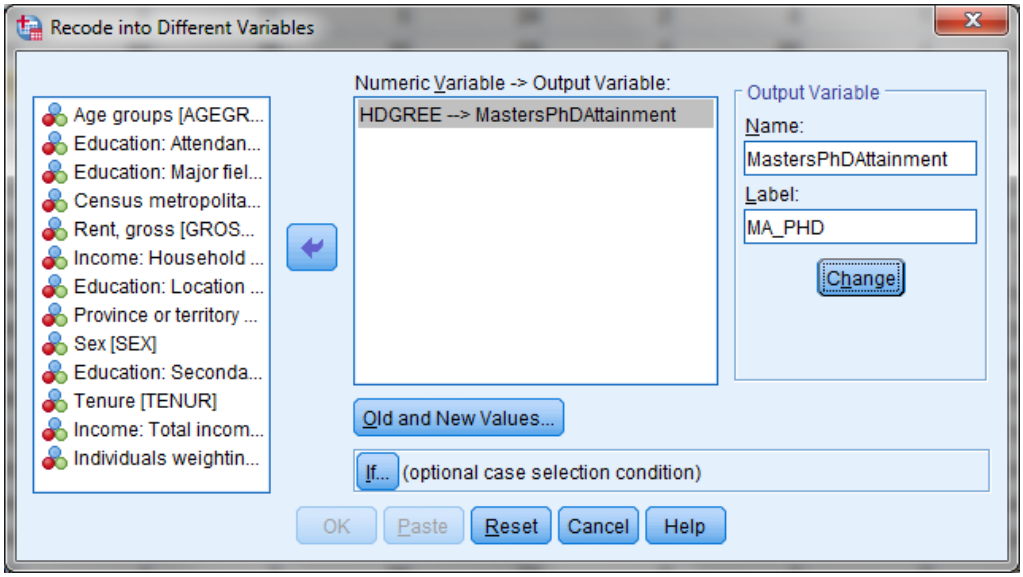
Sometimes you will need to recode a variable before running an analysis. For example, you may have a need to aggregate those holding a Master's degree and a PhD into one category instead of treating them as separate groups.

To recode a variable, click on **Transform**, then **Recode into different variables…**

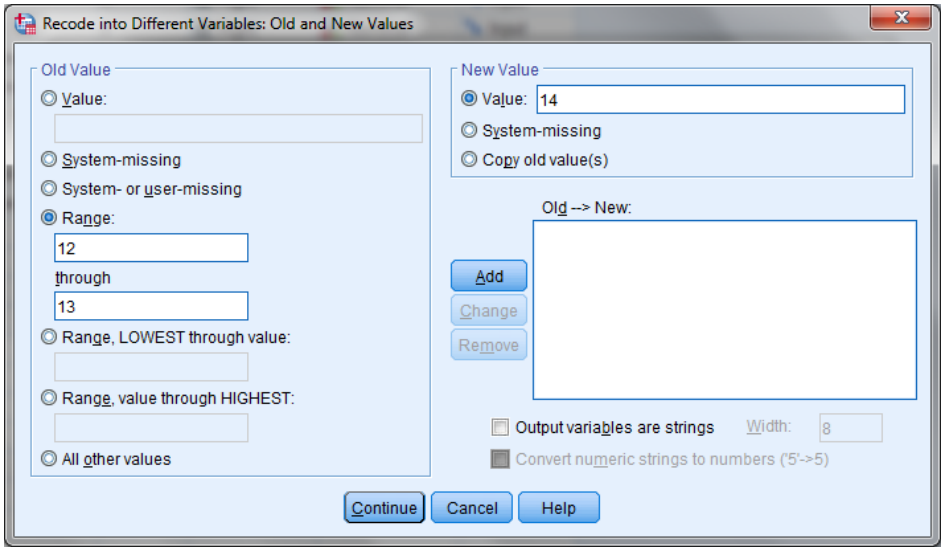
Select **Education: Highest certificate, diploma or degree** and move it to the **Input variable** window.

Change the **Output Variable** name and give it a label and then click on **Change**.

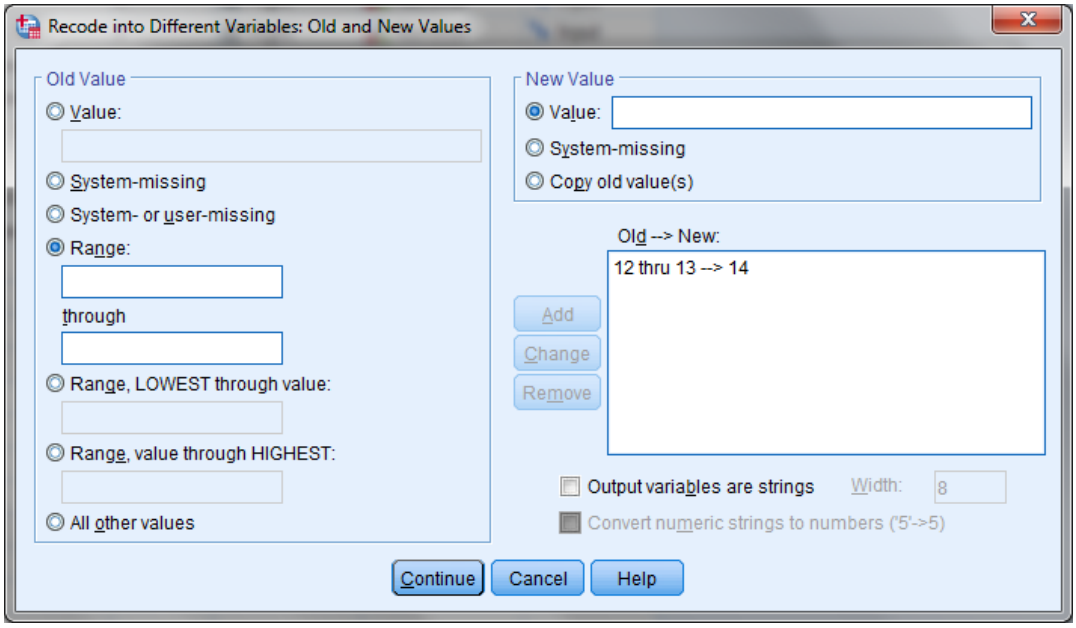
Then click on **Old and new values…**



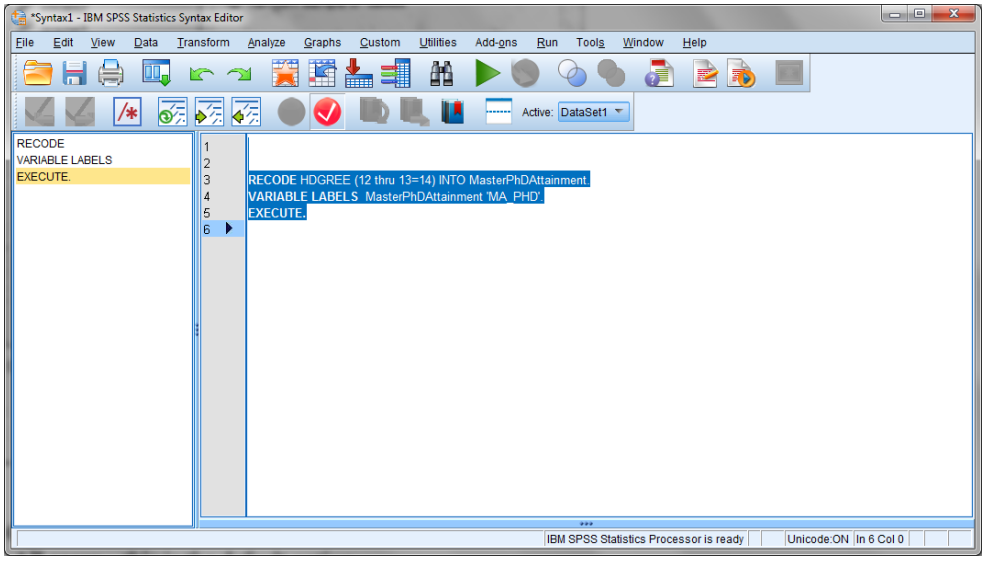
Choose **Range** and enter 12 through 13 in the **Old Value** options then under **New Value**, choose a new value – 14.

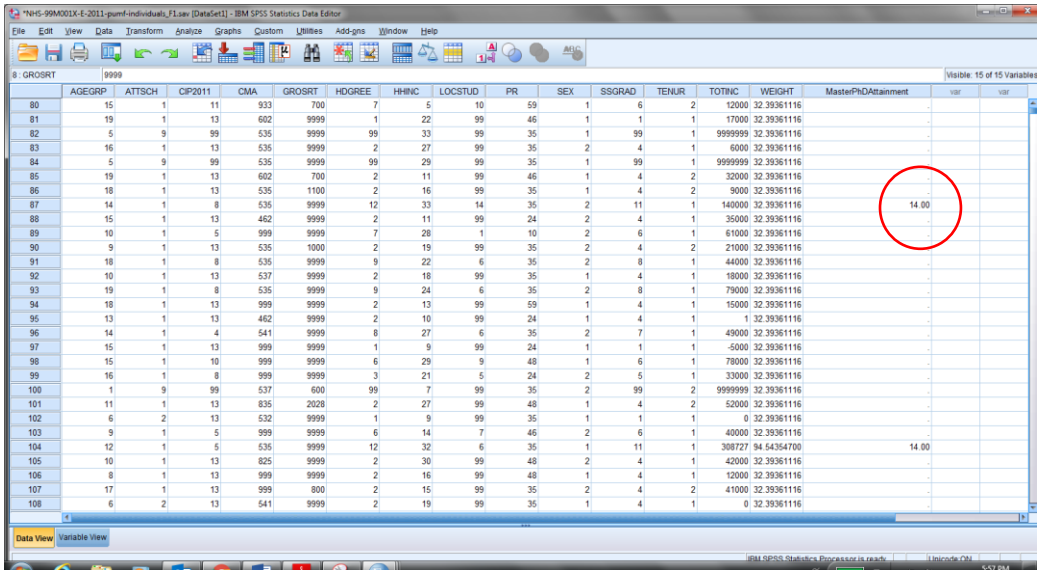


Click on **Add**, then **Continue**.



Click Paste. Select the code and click on the green run arrow.

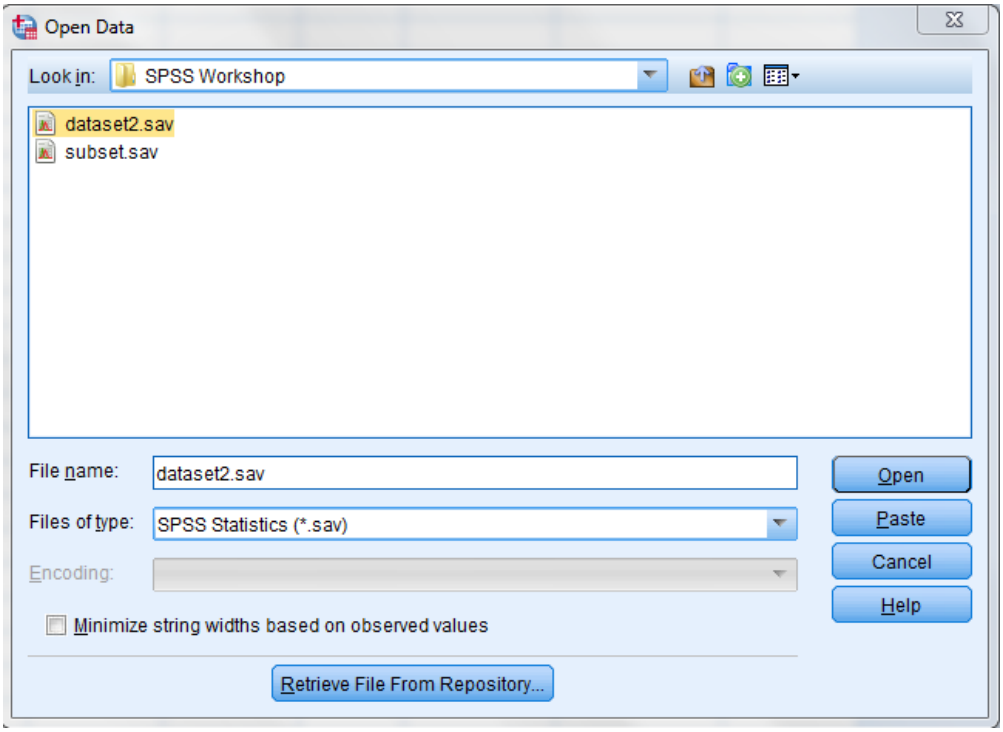




b) Create subset of a dataset

Using the second dataset that came in the SPSSWorkshop.zip file called "dataset2" let's now look at two more functions: creating a subset from a dataset and then another recode that combines three variables into one.

First, open **"dataset2.sav"** by clicking on **File -> Open -> Data…** and selecting "dataset2.sav"



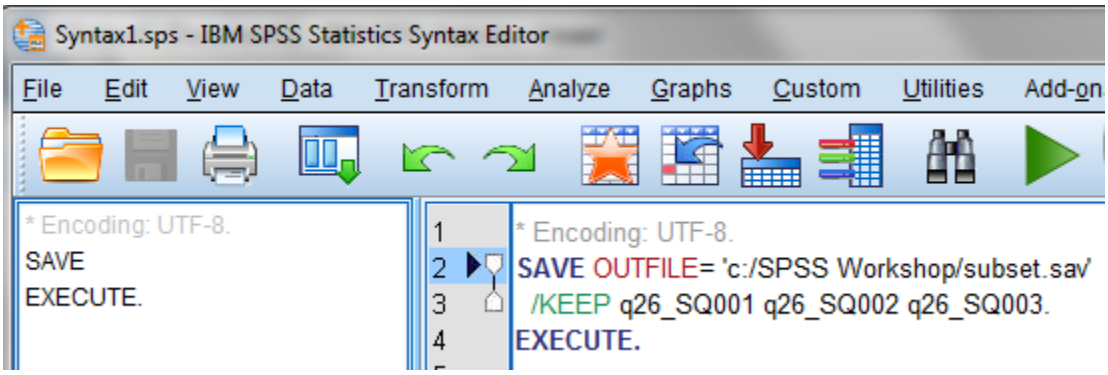
You will see in the SPSS Variable View window if you scroll down three variables: q26_SQ001 q26_SQ002 q26_SQ003. We will create a new dataset that consists of only these three variables.

315	TOPARC	Numeric	1	0	[Data archiving]...	{0, Not sele...	None	8	Right	Nominal	Input
316	TOPOTH	String	19	0	[Other] Do you ...	None	None	19	Left	Nominal	Input
317	q26_SQ001	Numeric	1	0	[Yes] Do you u...	{0, Not sele...	None	8	Right	Nominal	Input
318	q26_SQ002	Numeric	1	0	[No] Do you us...	{0, Not sele...	None	8	Right	Nominal	Input
319	q26_SQ003	Numeric	1	0	[Not applicable]...	{0, Not sele...	None	8	Right	Nominal	Input
320	q51	String	1750	0	Please provide ...	None	None	26	Left	Nominal	Input

You can do this using the SPSS Syntax Editor.

Click on **File -> Open -> Syntax…**

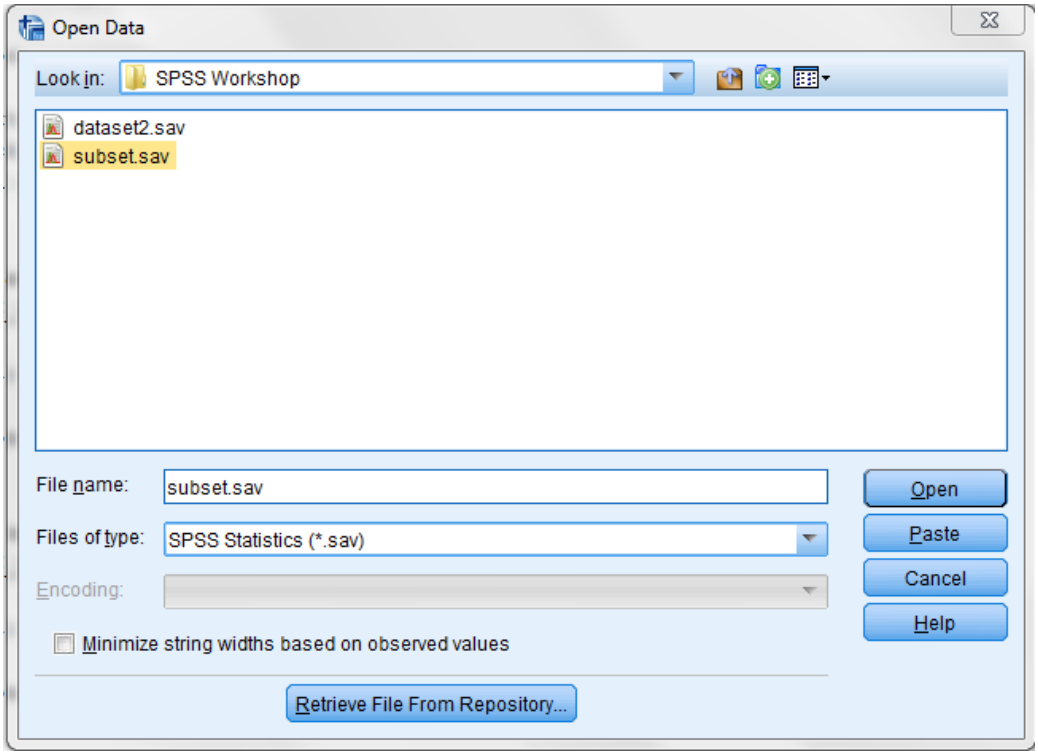
In the Syntax Editor create the following 3 statements:



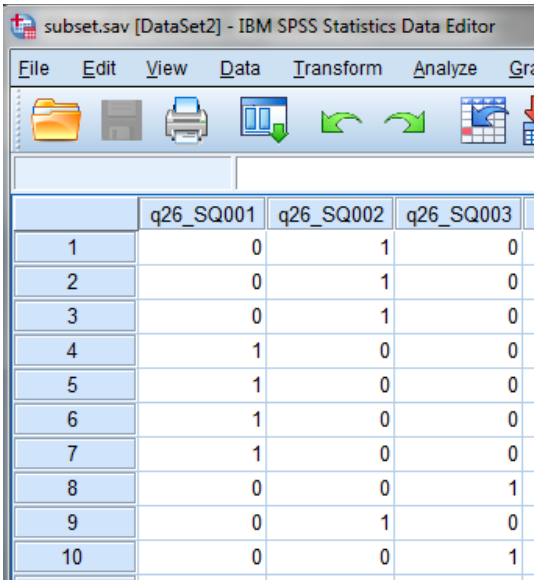
This creates a new subsetted dataset called “subset.sav” and the variables are chosen using the **KEEP command**. You always need to use an EXECUTE statement and to close the program with a full stop (.)

Next save your syntax (CTRL-S) and then select all the code and hit the green RUN SELECTION arrow.

To see the new “subset.sav” dataset, click File -> Open -> Data…



You will see the new dataset only has the variables you declared to keep.



	q26_SQ001	q26_SQ002	q26_SQ003
1	0	1	0
2	0	1	0
3	0	1	0
4	1	0	0
5	1	0	0
6	1	0	0
7	1	0	0
8	0	0	1
9	0	1	0
10	0	0	1

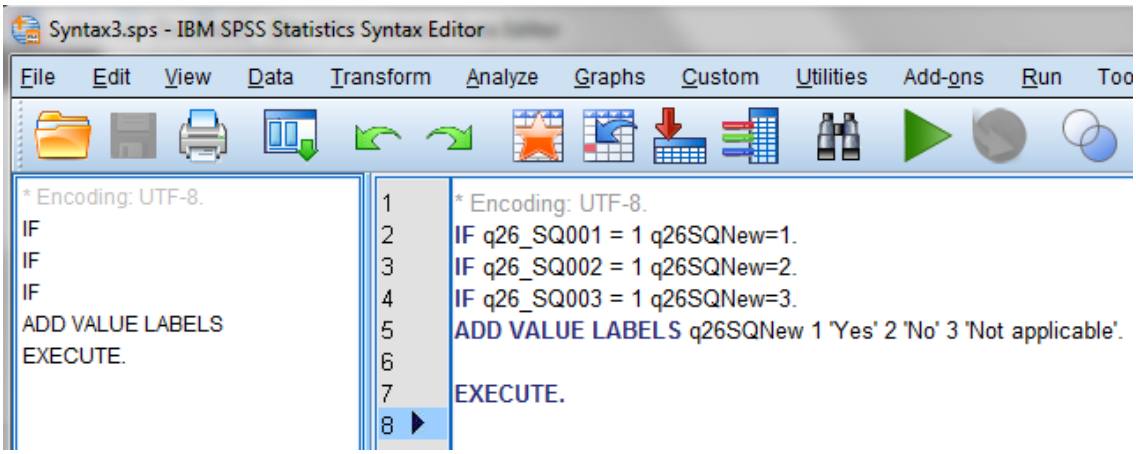
c) Combine 3 variables into 1

Finally, we will do another type of recode that will collapse these 3 variables into a single variable.

The reason for doing this is to reduce redundancy between these three measures since each variable is asking the same question but with one response option. We will create a single variable that combines all three response options for the single question into a single variable.

To get started, open a new syntax window (File -> New -> Syntax…)

And enter the following statements:



In essence, we are assigning the old values of "1" in each of the three variables to a new value in a single variable.

Next, save (CTRL-S) the syntax you have created and then select all the code and click on the green RUN SELECTION arrow.

You will now see the new variable "q26SQNew" in the Data Editor window and if you go to Variable View you will see the new values and labels for this new variable.