[Commentary by **John F Hall**] [**Draft only**: last updated 2 May 2018]

John MacInnes
An Introduction to Secondary Data Analysis with IBM SPSS Statistics
(Sage, Dec. 2017)

5.1 Chapter 5 video tutorials (direct link to companion website)
[**NB:** All video tutorials for chapter 5 are on the same web page and cannot (yet) be disaggregated.

Video 5.1.10: (8'40") **Creating and downloading a data extract from GSS**

Sections 5.6 (p126) and 5.7 (pp 127-131) in chapter 5 use data from the **General Social Survey** run annually by the **National Opinion Research Center** (University of Chicago**).** Data are also distributed by the **Roper Center**

Video 5.1.10 does a quick tour of the NORC/GSS site covering FAQs, GSS Data Explorer, creating an account and specifying a project page. The opening page has links to research staff, research reports and to two separate websites:

General Social Survey (not the same page as above) contains general information and a link:

Get the Data >>

GSS Data Explorer (below)

. . has buttons which allow you to Search Variables or Create an Account

It is very complicated for first time users who need to search the data page and locate variable(s) by keyword. As JM says, it can also be a bit "hit and miss" depending on who catalogued the data and the meaning of words: it takes a bit of lateral thinking. One problem is to find the variable relating to "*attitudes to women taking up employment when they have a pre-school age child at home*". He knows where to find it, but do we? (It's called **[wrkbaby]**) He then proceeds to find respondents' **sex**, **age**, **race** and **education level** (he guesses at **educ** and finds highest year of school completed) and adds a weighting variable.

Creating the extract takes 10 – 15 minutes. The download has variables in the following order:

YEAR ID_ AGE EDUC SEX RACE WRKBABY WTSSALL

	Name
1	YEAR
2	ID_
3	AGE
4	EDUC
5	SEX
6	RACE
7	WRKBABY
8	WTSSALL

JM says he already has an account and that users will need to create one if they want to use the GSS. In fact, you can download the entire GSS dataset **without an account**. Would it be quicker to download the whole file and extract variables from it?

GSS Data Set >>

Select download format:

SPSS STATA

offers a choice of download formats and a choice of either the entire series:

GSS 1972-2016 Cross-Sectional Cumulative Data (Release 3, March 9, 2018)

GSS 1972-2016 Cross-Sectional Cumulative Data (Release 3, March 9, 2018) - With GSS Codebook

or for one or more specific years:

1972	1973	1974	1975	1976	1977	1978		
1980	1982	1983	1984	1985	1986	1987	1988	1989
1990	1991	1993	1994	1996	1998			
2000	2002	2004	2006	2008				
2010	2012	2014	2016					

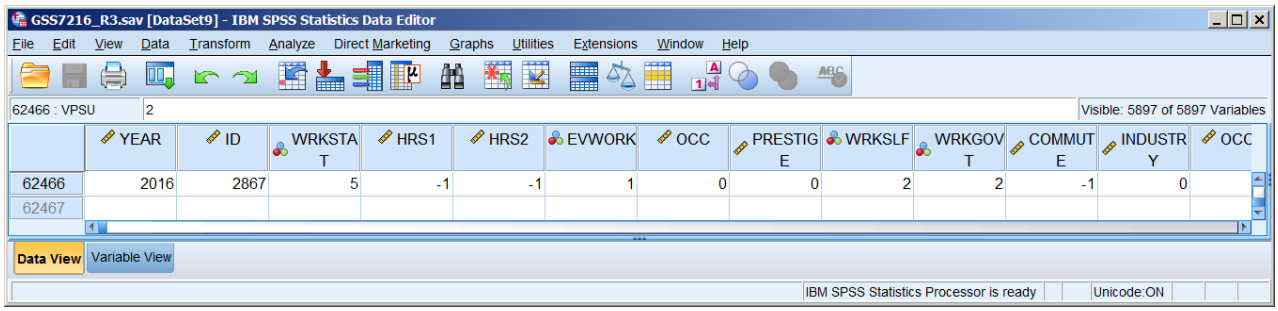
[**NB:** Only year-specific variables are included in the yearly data files.]

To follow JM's steps in creating his SPSS saved file, you need to watch the video with constant backward skips and frequent stop/starts, but I have replicated his full process step by step in the Appendix to this document.

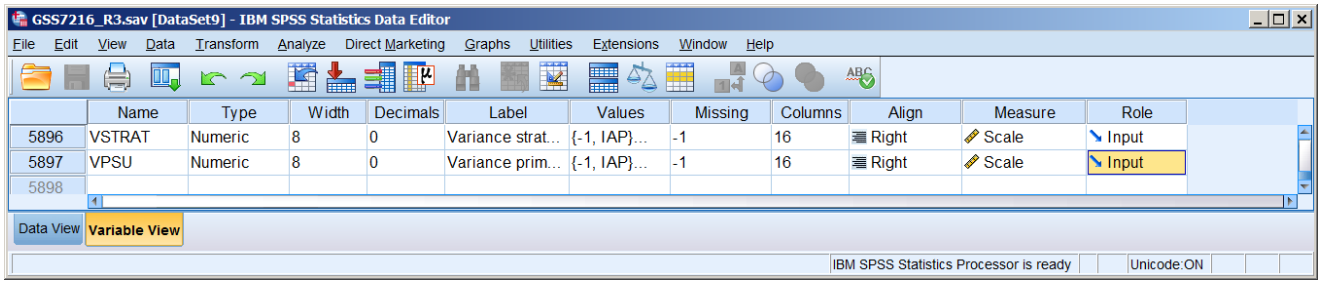
A more direct way is to download the entire cumulative data set[1]

SPSS saved file GSS7216_R3 is a cumulative data set containing the entire data for all years 1972 to 2016.

It contains 62,466 cases:



. . and 5,897 variables:

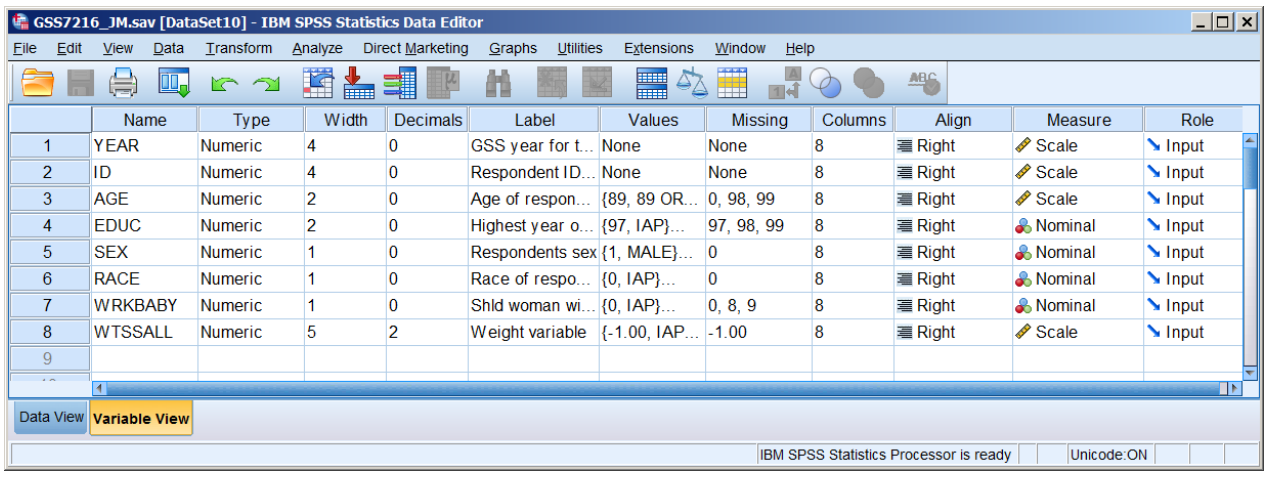


With file GSS7216_R3 open, I attempted to replicate JM's file by:

```
save outfile
    'F:\Desktop folders\Research 2018\MacInnes  2017\JFH SPSS files\Ch5_Replications\GSS7216_JM.sav'
   /keep YEAR  ID  AGE  EDUC  SEX  RACE  WRKBABY  WTSSALL.
```

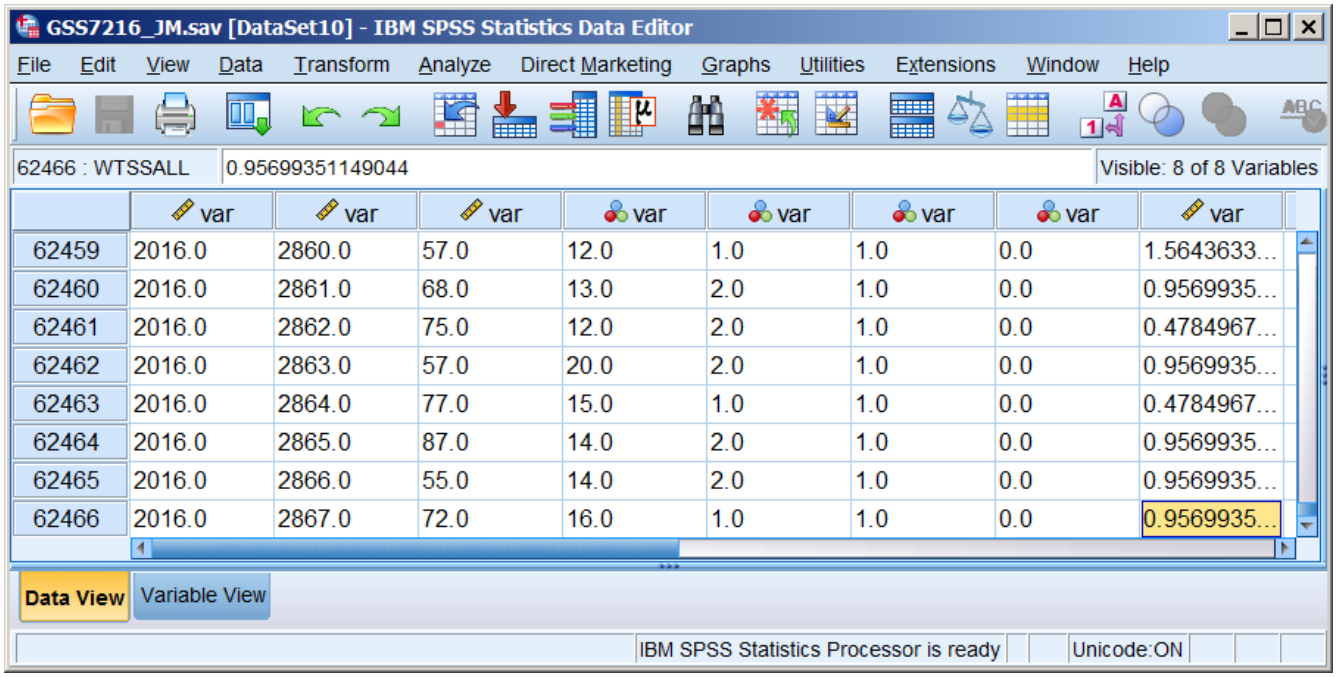
File GSS7216_JM is saved in folder **Ch5_Replications**

In Variable View it contains 8 variables:



[1] [**NB:** The SPSS student version, which can handle up to 50 variables and 1500 cases, **cannot open** the GSS data files.]

. . and 62,466 cases:

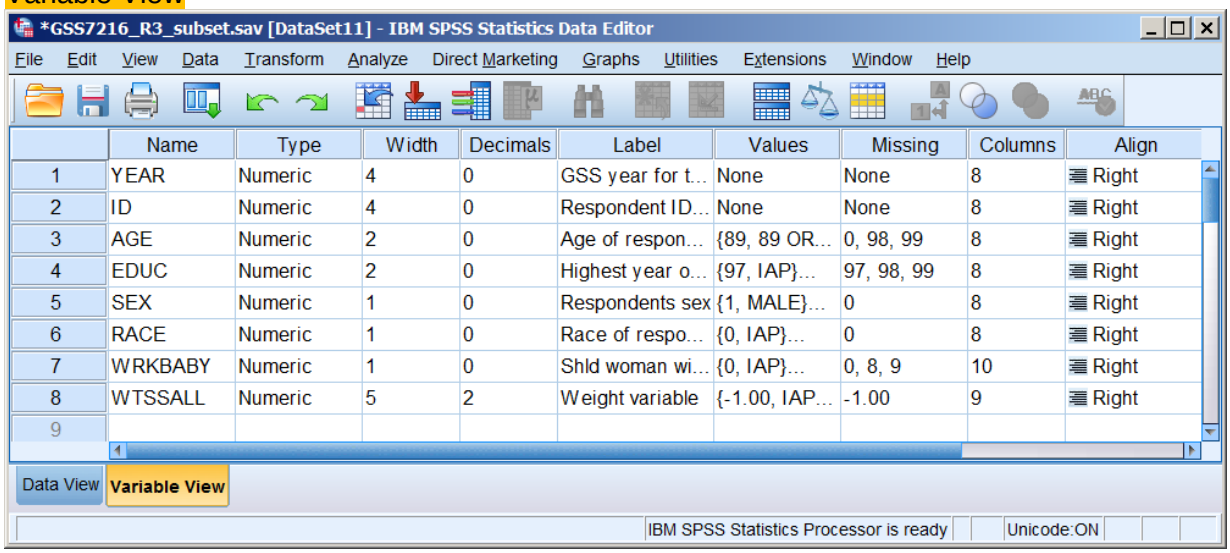


However, in Data View above the file doesn't display the original variable names and, despite **Decimals** being declared as 0, all integer values are displayed with one decimal place.

After an appeal to the gurus on SPSSX-L@LISTSERV.UGA.EDU, I was sent a new file[2] in the correct format, but as yet do not know why my syntax didn't work.

GSS7216_R3_subset

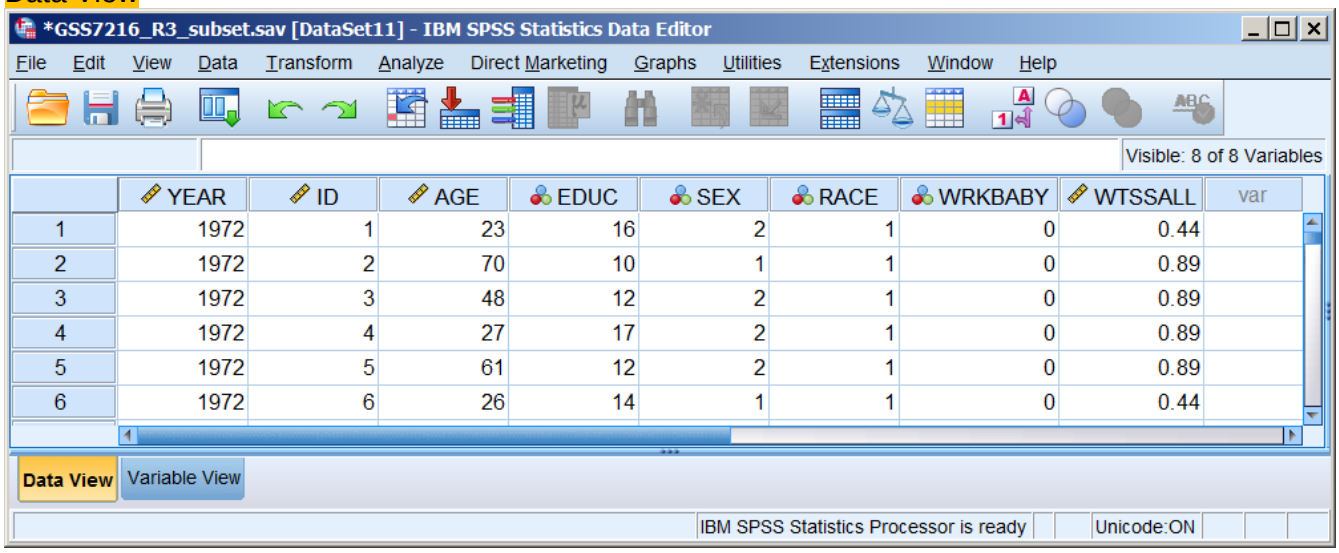
Variable View



[2] The file was kindly created and sent via Dropbox by Anthony Babinec, President of AB Analytics and co-author of Data Analysis with IBM SPSS Statistics: Implementing data modeling, descriptive statistics and ANOVA.
My request was sent on 01 May 2018 19:54 and he replied on 02 May 2018 03:57.

.

Data View



	YEAR	ID	AGE	EDUC	SEX	RACE	WRKBABY	WTSSALL	var
1	1972	1	23	16	2	1	0	0.44	
2	1972	2	70	10	1	1	0	0.89	
3	1972	3	48	12	2	1	0	0.89	
4	1972	4	27	17	2	1	0	0.89	
5	1972	5	61	12	2	1	0	0.89	
6	1972	6	26	14	1	1	0	0.44	

display labels.

Variable Labels

Variable	Position	Label
YEAR	1	GSS year for this respondent
ID	2	Respondent ID number
AGE	3	Age of respondent
EDUC	4	Highest year of school completed
SEX	5	Respondents sex
RACE	6	Race of respondent
WRKBABY	7	Shld woman with preschooler work?
WTSSALL	8	Weight variable

Variables in the working file

frequencies wrkbaby.

WRKBABY Shld woman with preschooler work?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 WORK FULL-TIME	581	0.9	12.9	12.9
	2 WORK PART-TIME	1763	2.8	39.0	51.9
	3 STAY HOME	2172	3.5	48.1	100.0
	Total	4516	7.2	100.0	
Missing	0 IAP	57132	91.5		
	8 CANT CHOOSE	744	1.2		
	9 NA	74	0.1		
	Total	57950	92.8		
Total		62466	100.0		

means wrkbaby by year. [Quick check to see which years use **wrkbaby**]

Report

WRKBABY Shld woman with preschooler work?

YEAR GSS year for this respondent	Mean	N	Std. Deviation
1988	2.44	1242	0.682
1994	2.44	1273	0.684
2002	2.32	1014	0.694
2012	2.17	987	0.697
Total	2.35	4516	0.697

wrkbaby was used in only four waves 1988, 1994, 2002 and 2012. The means don't tell you much: the changes are more marked if you compare percentages.

crosstabs year by wrkbaby /cells count row.

YEAR GSS year for this respondent * WRKBABY Shld woman with preschooler work? Crosstabulation

			WRKBABY Shld woman with preschooler work?			Total
			1 WORK FULL-TIME	2 WORK PART-TIME	3 STAY HOME	
YEAR GSS year for this respondent	1988	Count	136	425	681	1242
		% within YEAR GSS year for this respondent	11.0%	34.2%	54.8%	100.0%
	1994	Count	141	436	696	1273
		% within YEAR GSS year for this respondent	11.1%	34.2%	54.7%	100.0%
	2002	Count	133	421	460	1014
		% within YEAR GSS year for this respondent	13.1%	41.5%	45.4%	100.0%
	2012	Count	171	481	335	987
		% within YEAR GSS year for this respondent	17.3%	48.7%	33.9%	100.0%
Total		Count	581	1763	2172	4516
		% within YEAR GSS year for this respondent	12.9%	39.0%	48.1%	100.0%

GSS_72-16_R3_subset,sav contains all 62466 cases, but the key dependent variable **wrkbaby** occurs only in years 1988, 1994, 2002 and 2012 so I created a smaller version containing only those years (N = 4516). Users of the textbook will find it difficult or impossible to create this subset.

A file containing only the 4516 non-missing cases for **wrkbaby** can be created with:

select if (not (missing (wrkbaby))).
frequencies wrkbaby.

WRKBABY Shld woman with preschooler work?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 WORK FULL-TIME	581	12.9	12.9	12.9
	2 WORK PART-TIME	1763	39.0	39.0	51.9
	3 STAY HOME	2172	48.1	48.1	100.0
	Total	4516	100.0	100.0	

Make sure the file is saved with a different name.

If you want to keep cases with missing values for **wrkbaby**

count years = year (1988, 1994, 2002, 2012).
select if (years eq 1).
frequencies wrkbaby.

WRKBABY Shld woman with preschooler work?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 WORK FULL-TIME	581	6.3	12.9	12.9
	2 WORK PART-TIME	1763	19.1	39.0	51.9
	3 STAY HOME	2172	23.6	48.1	100.0
	Total	4516	49.0	100.0	
Missing	0 IAP	3878	42.1		
	8 CANT CHOOSE	744	8.1		
	9 NA	74	0.8		
	Total	4696	51.0		
Total		9212	100.0		

Clear the intermediate variable **years** and save the file with a different name.

To save users a great deal of frustration, and with authorisation from NORC, the SPSS saved file extracted from NORC General Social Survey 1972 - 2016 (needed for the examples in sections 5.6 and 5.7 and for exercises 6, 7 and 8 in chapter 5) has been authorised by NORC for uploading to my site. The link is GSS7216_R3_wrkbaby2.sav.

End of: **5.1.10 Creating and downloading a data extract from GSS**

Back to: MacInnes (2017)

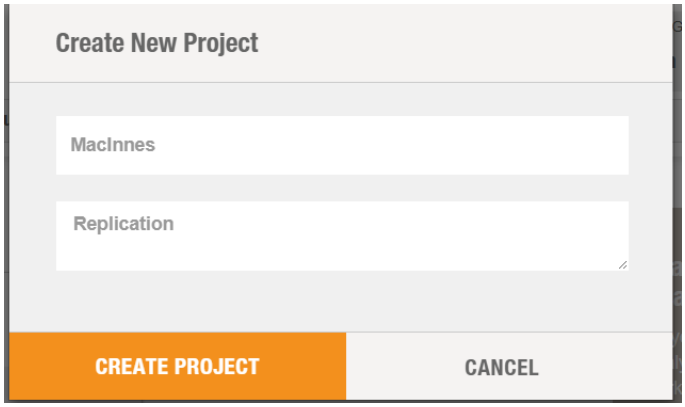
Back to: 5.1.9 Using syntax to repeat analyses on new data

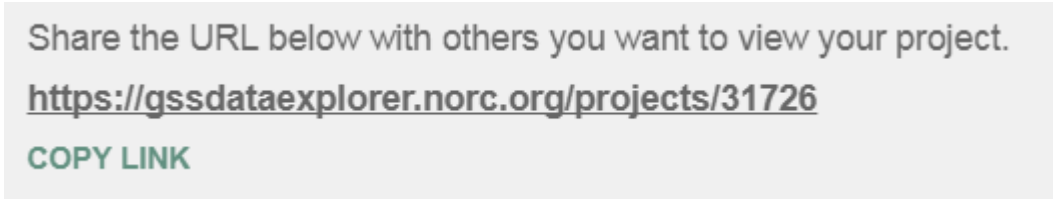
Forward to: Appendix

Forward to: **5.1.11 Creating a filepath to the GSS command file**

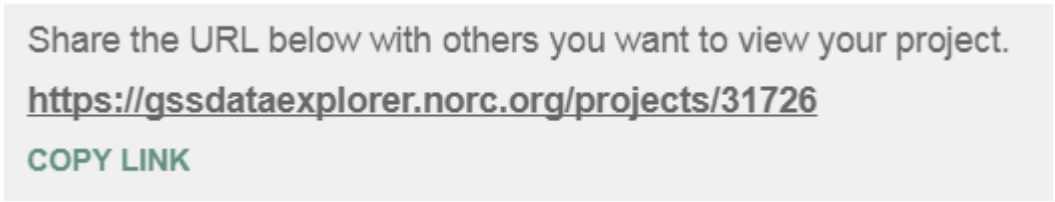
Appendix

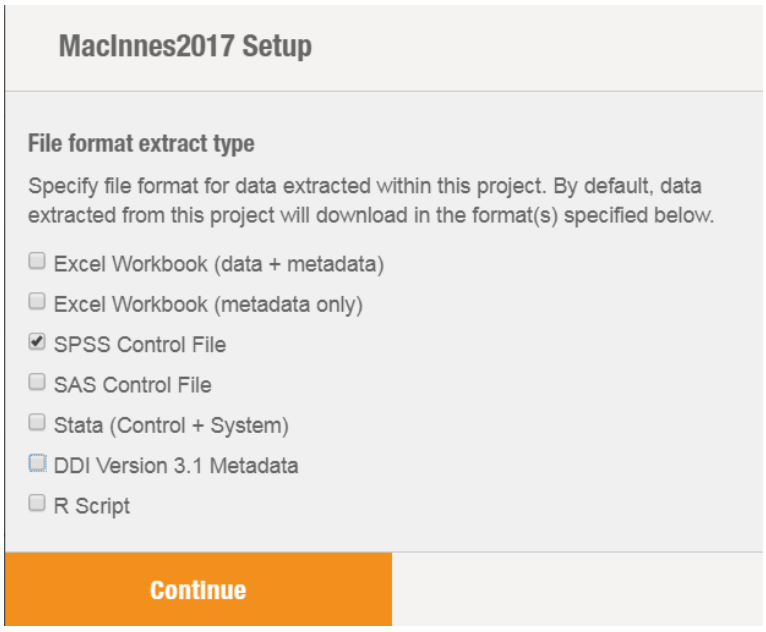
Below are the steps I followed in recreating JM's SPSS saved file whilst following the video.

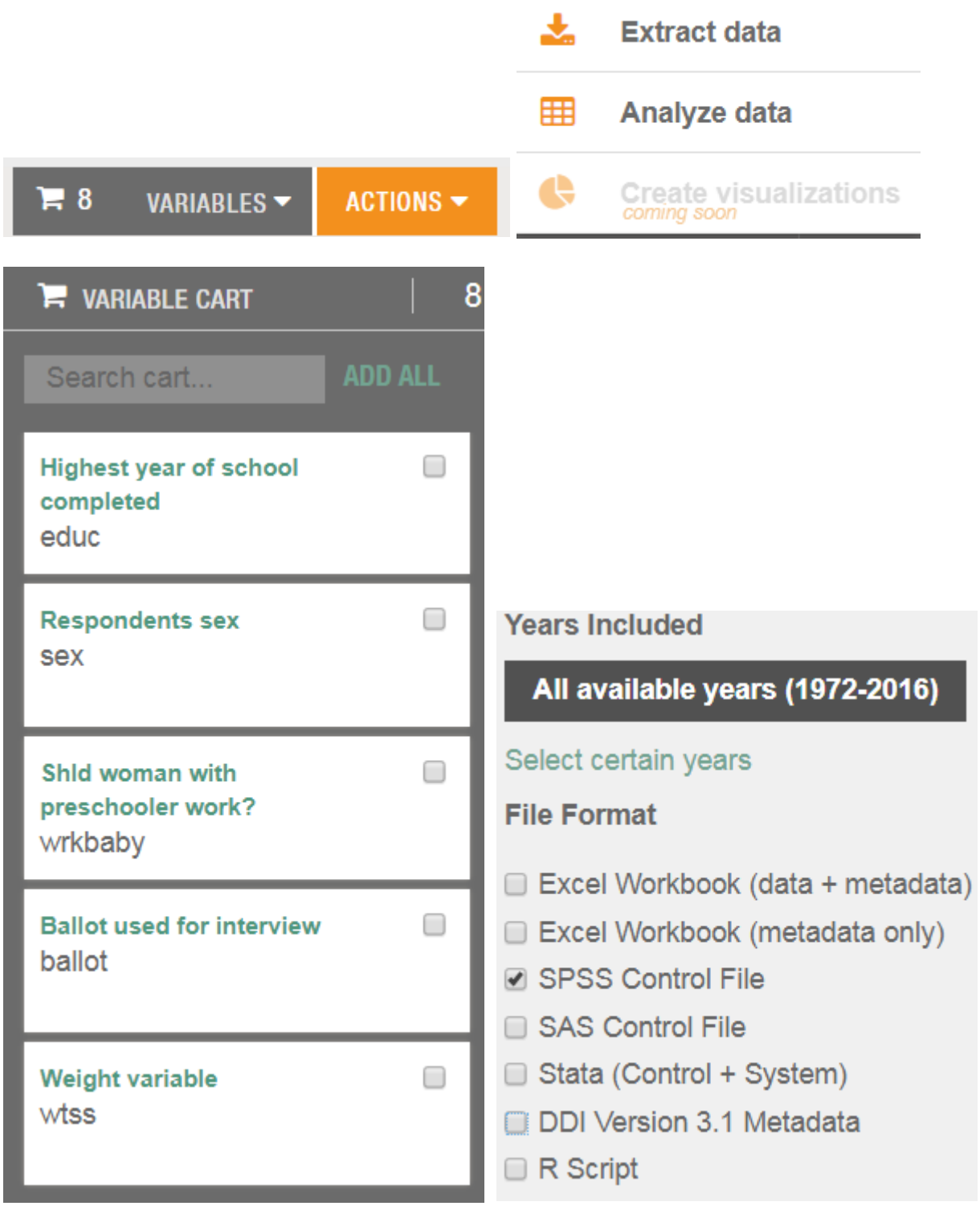




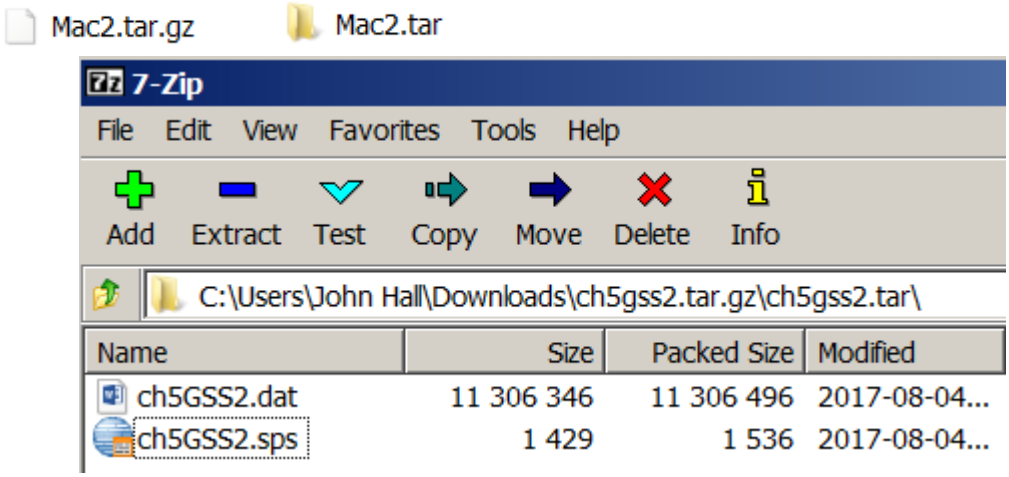
Project name: Gender
Description Variables about attitudes to working mothers



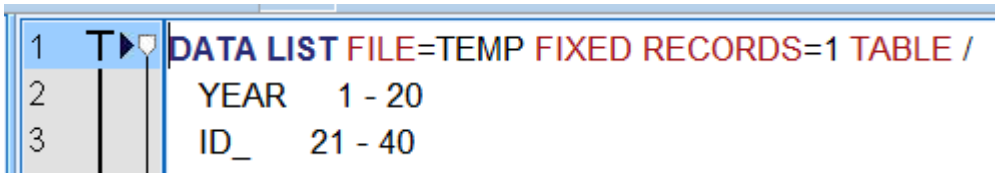




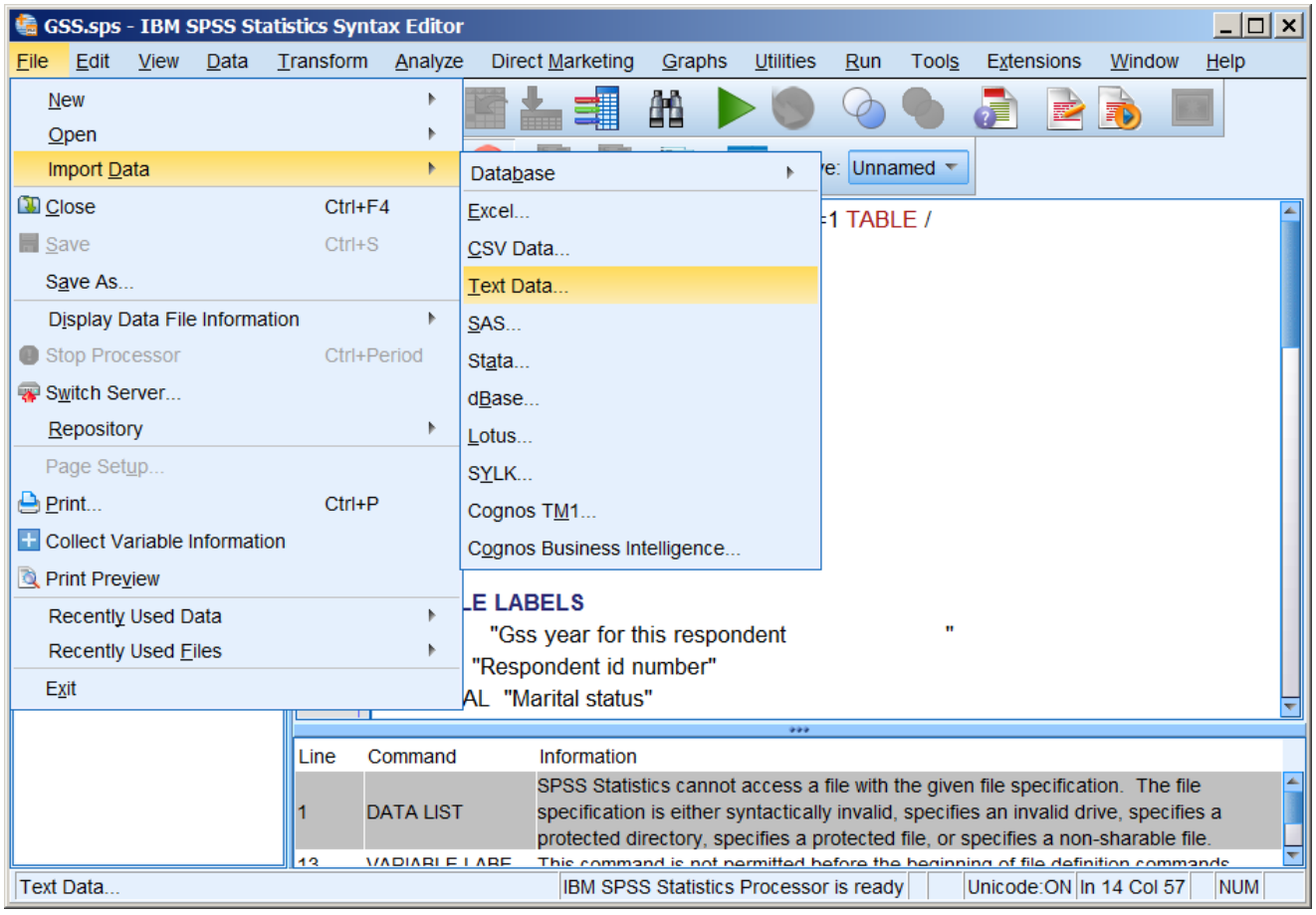
Name	Date	Created By	Revisions	Actions
Mac2	8/1/17 7:06 pm	John F Hall	3	



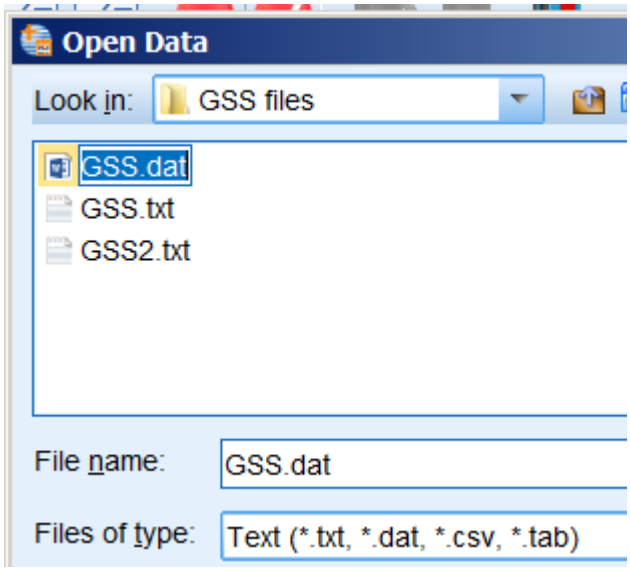
Replace TEMP with a path

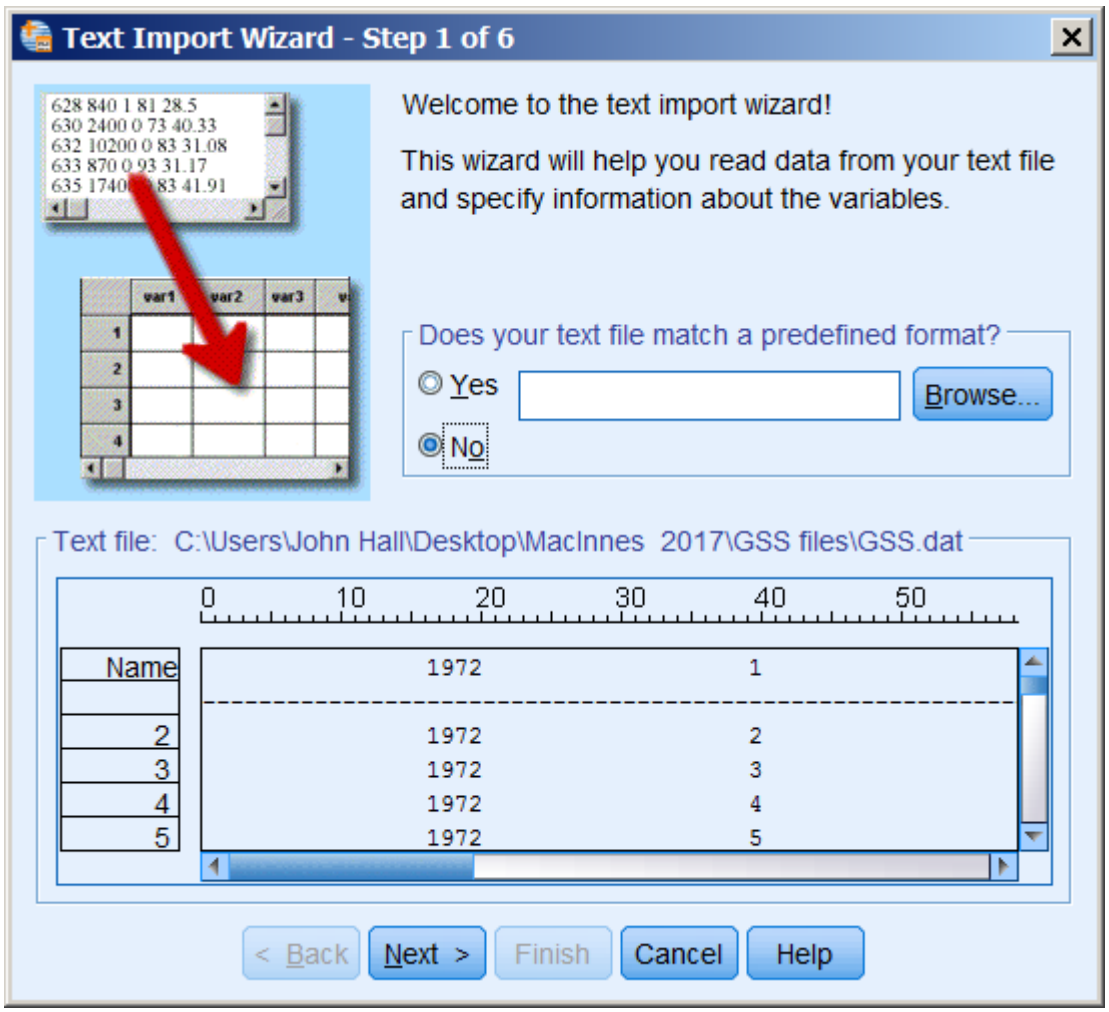


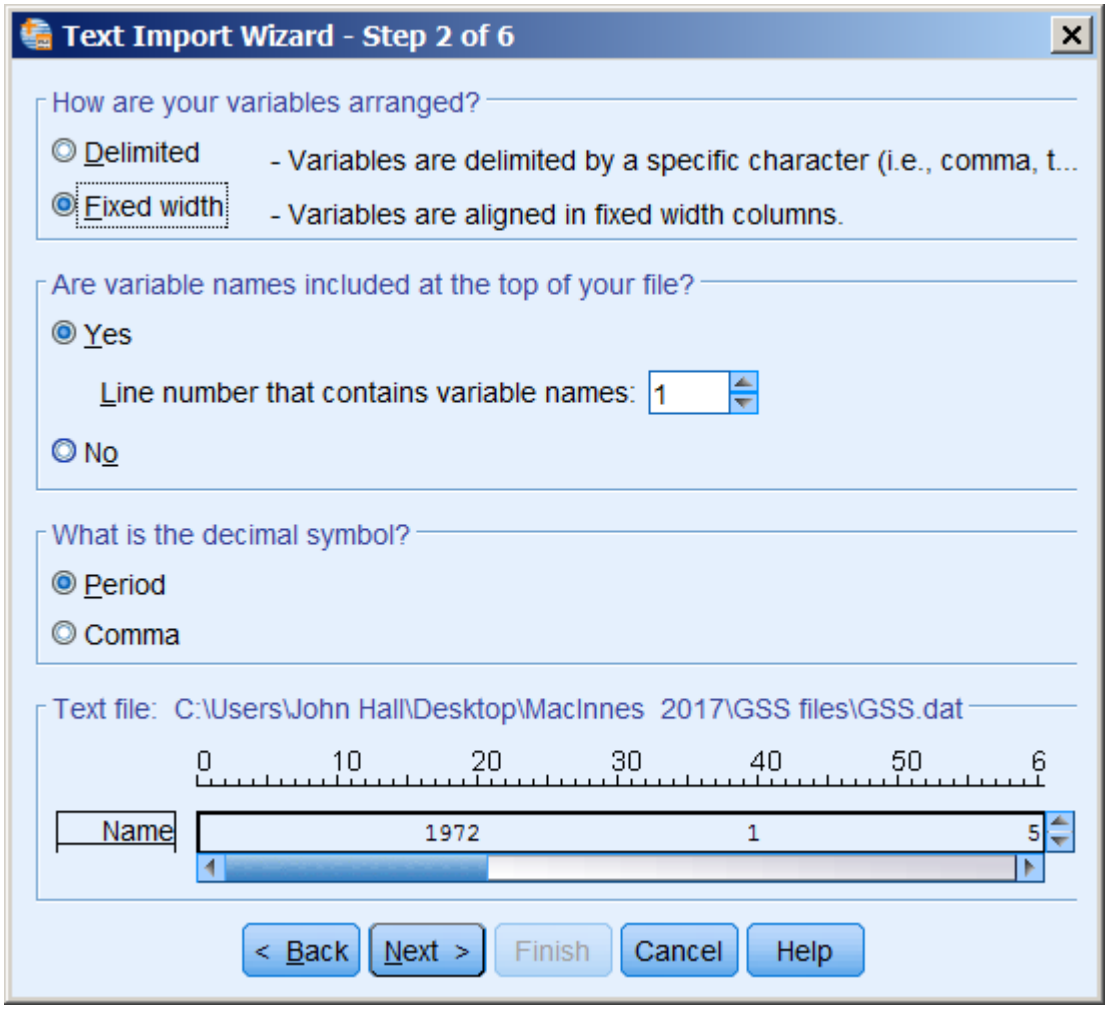
File >> Read Text Data >> Open



Text import wizard








```
1  DATA LIST FILE='C:\Users\John Hall\Desktop\MacInnes  2017\GSS files\gss.dat'
2  / FIXED RECORDS=1 TABLE / .
```

The extract arrives in a zip file containing two files: a data file and an SPSS syntax file. JM says the syntax generates an SPSS saved file, but it doesn't. The syntax uses the data file to generate a new active SPSS **Data Editor** named on his video **Untitled3**.

```
DATA LIST ~ ~ ~ .
VARIABLE LABELS ~ ~ ~ .
VALUE LABELS ~ ~ ~ .
EXECUTE .
CACHE .
EXECUTE .
DATASET NAME ~ ~ ~ .
```

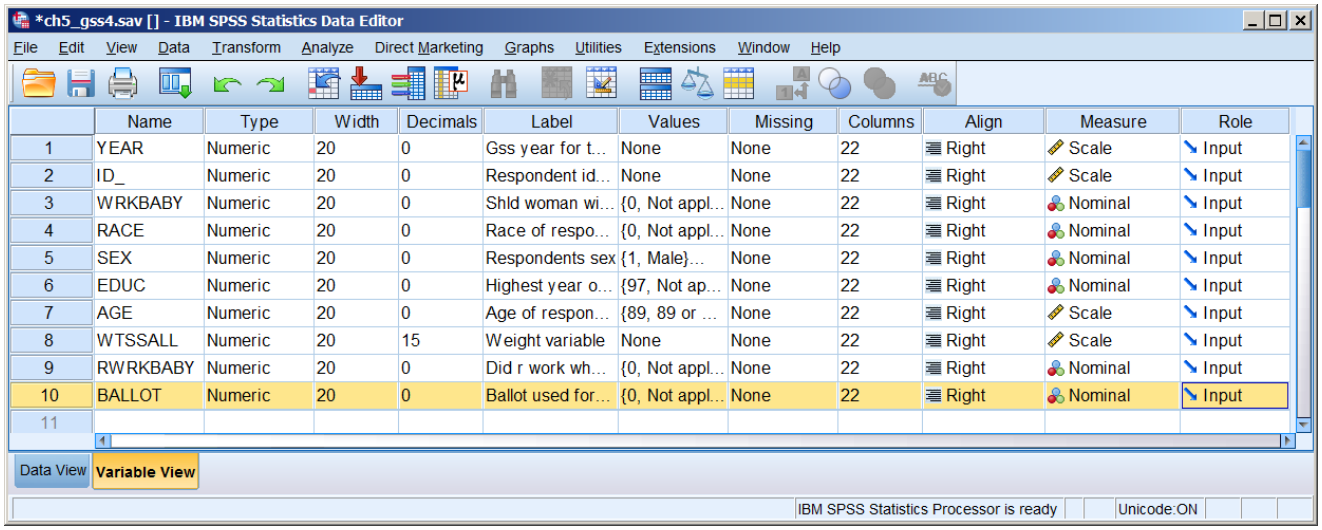
It allocates names for the variables, specifies their format and location in the data file and adds dictionary information[3]. No missing values are specified at this stage, but this is standard practice in many archives, caused by their regular use of automatic archiving software.

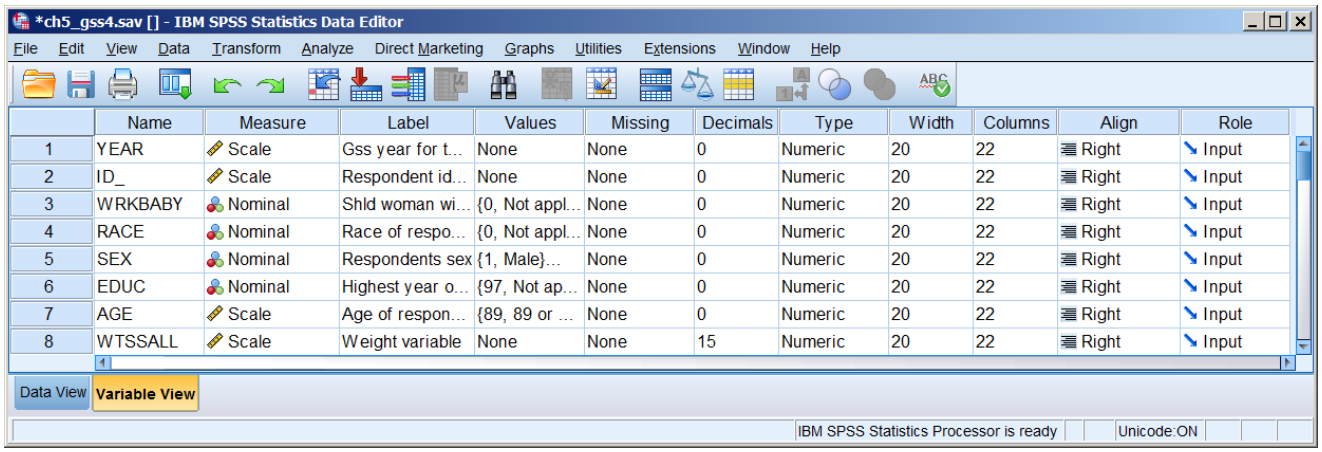
The **Data Editor** only becomes an SPSS *.sav file once it has been saved with a name of his choice.

"Give it any name that's appropriate." He calls his "GSS_mother_work"

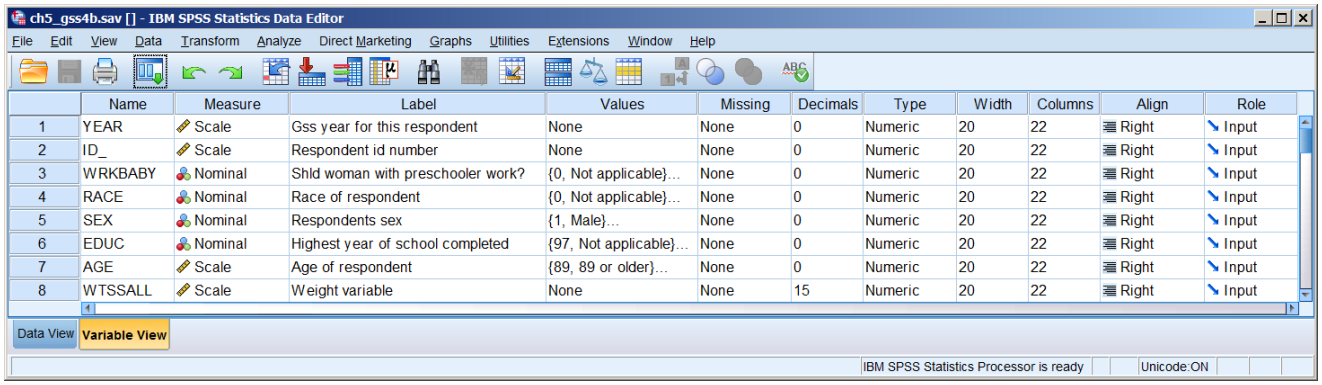
JM later adds **MISSING VALUES** (but not for all variables) and **WEIGHT** before doing any analysis.

```
missing values
   wrkbaby (0 8 9)
   /race (0)
   /educ (97 98 99)
   /age (98 99).
```





[3] For a detailed explanation of how data relate to questionnaires and how SPSS files are generated from these data, see Block 1: From questionnaire to SPSS saved file especially the sequence of tutorials in sections **1.3: Reading raw data into SPSS** and **1.4: Completing your data dictionary**



	Name	Measure	Label	Values	Missing	Decimals	Type	Width	Columns	Align	Role
1	YEAR	Scale	Gss year for this respondent	None	None	0	Numeric	20	22	Right	Input
2	ID_	Scale	Respondent id number	None	None	0	Numeric	20	22	Right	Input
3	WRKBABY	Nominal	Shld woman with preschooler work?	{0, Not applicable}...	None	0	Numeric	20	22	Right	Input
4	RACE	Nominal	Race of respondent	{0, Not applicable}...	None	0	Numeric	20	22	Right	Input
5	SEX	Nominal	Respondents sex	{1, Male}...	None	0	Numeric	20	22	Right	Input
6	EDUC	Nominal	Highest year of school completed	{97, Not applicable}...	None	0	Numeric	20	22	Right	Input
7	AGE	Scale	Age of respondent	{89, 89 or older}...	None	0	Numeric	20	22	Right	Input
8	WTSSALL	Scale	Weight variable	None	None	15	Numeric	20	22	Right	Input

My way is quicker and easier. The SPSS saved file needed for exercises 6, 7 and 8 in chapter 5 (extracted from NORC General Social Survey 1972 - 2016) has been authorised by NORC for uploading to my site. The link is GSS7216_R3_wrkbaby2.sav.