3.2: Three (or more) variables   **3.2.1 Elaboration**

3.2.1.2 Elaboration 2 (Income differences BSA 2009 - 2014)
(Replication of elaboration exercise 3.2.1.1 using data from BSA 2009 to 2014)

This set of tutorials uses data from the British Social Attitudes series to explore the following research questions.

  1: Is there a difference between the earnings (from paid work) of men and women?
  2: What other variables might account for differences in earnings?
  3: What effect do they have by themselves?
  4: What happens to any differences in earnings between men and women when controlling for these other variables?

Model

Logical model is **X → Y . T**   (the effect of **X** on **Y** controlling for **T**) where:

Y = Dependent variable
X = Independent variable
T = Test variable(s)

(See Elaboration (extract from Jim Ring's Statistical Notes specially written for this course)

Y (Dependent)	**X (Independent)**	T_n **(Test or control)**	
Gross earnings from paid work	Sex	T_1	Working full time or part time
		T_2	Employee or self employed
		T_3	Level of education
		T_4	Age
		T_5	Economic sector
		T_6	Socio-economic grade of work
		T_7	Geographical region

Data source: British Social Attitudes 1983 to 2014: Cumulative SPSS file,

Tutorial 3.2.4.1 Income differences - Elaboration used data from the 1989 wave of the British Social Attitudes Survey (BSAS) to analyse differences between the earnings of men and women from paid work. This new tutorial replicates that exercise on data from the 2009 to 2014 waves.

The cumulative mother file for 1983 to 2014 contains 95,630 cases and 10,773 variables. Not all variables are present in all waves. Even when they are, they may have different coding schemes.

For most variables in this and related exercises, I have retained the original metadata (variable names, variable labels, value labels, missing values, measurement levels, formats) but have added these where they were absent or modified them where they were confusing, incomplete or incorrect.

Positive missing values have been changed to negative, including some which, though undeclared, are clearly intended as missing. Where necessary I have created a new variable.

Dependent variable:	Description	Name
Gross earnings	Grouped as per showcard	**rearn**
from paid work	Grouped by deciles (2014 labels)	**rearnq**
(if working)	Grouped by quartiles (2014 labels)	**rearnq**
	Grouped by quartiles (new labels)	**rearngrp**

Variable **[rearn]** "Grouped gross earnings" is present in all waves from 1983 to 2010, with values ranging from **1** to **24** (Earnings groups) **0** (Skipped) and **97** to **99** (missing, but not declared) with category boundaries maintained, but upward movement of earnings over time has resulted in empty counts, depending on the year. In 1983 the codes range from **1** to **11**, in 2001 **3**, **5**, **7** to **24**, in 2008 **1 - 17**.   Value **0** (skipped) only appears from 1983 to 1991. Grouping categories for earnings codes have changed over the years and have different ranges: they are inconsistent and therefore too complex to use for the purpose of this exercise.

From 2011 onwards, variable **[rearn]** is not included in the distributed data: it has been replaced by derived variables **[rearnd**] "Decile groups of R's gross earnings" and **[rearnq]** "Quartile groups pf R's gross earnings". For this exercise I have selected **[rearnq]** which groups earnings into four categories, each containing approximately 25% of the sample, but is available only for years 2009 to 2014.   Because the quartile cutting points are different between waves, I have created a new variable **[earngrp]** from **[earnq]** and given it different labels indicating quartiles rather than ranges.

Independent variable:	Description	Name
Sex of respondent	Male/Female	**Rsex**

Variable **[rsex]** is present in all waves and has no missing values. Users may prefer to rename it as **[sex]** or **[gender]** according to their preferences.
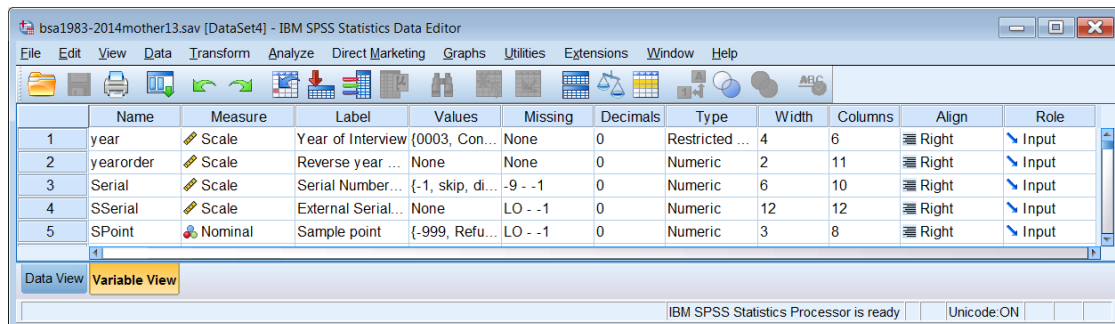
Which other variables to include?

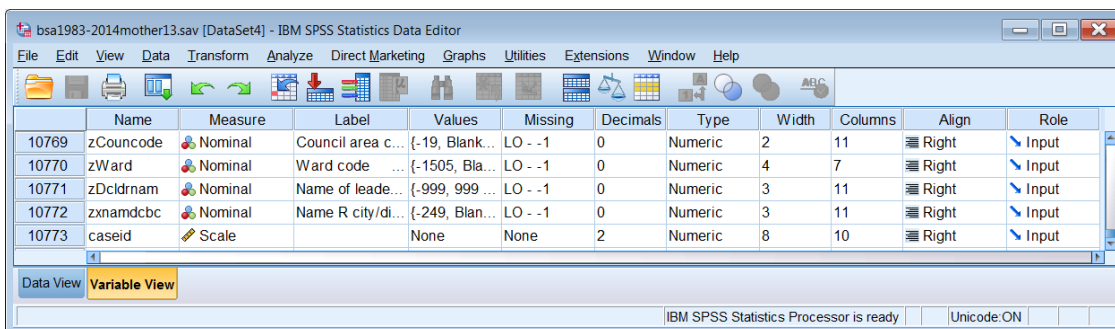| Test variables:	Description	Name
Work-related	Employee/Self-employed	**remploye**
	Hours worked	**ejbhrcax**
	Social class of job	**rnsoccl**
	Private / public sector	**rsector**
		(not used in 2009 – 2014)
		rocsect2 used instead.
Education	Age completed full time educ.	**tea**
	Examinations at school	**schqual**
	Qualifications post-school	**pschqual**
	Highest educational qualification	**hedqual3**
Age	Age last birthday	**rage**
	Age group	**ragecat3**
Location	Region	**gor2**
	Country	**country**

Stage 1: Extraction of variables from mother file[1]

1.1: Locate and open cumulative mother file: 🔲 bsa1983-2014mother13

[NB: This file is pass-word protected: access is via a Dropbox link and users need to be currently licenced at UKDS and registered for use of BSA data. However you can follow the tutorial without this or later files, and without access to SPSS]
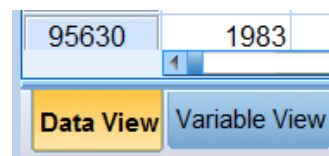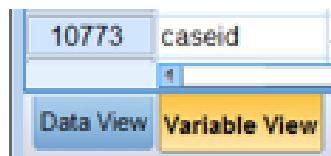

Mother file has 10773 variables   and 95630 cases

1.2: Create a new working file containing the dependent, independent and selected test variables, plus wave, weight and case ID number:

Admin:	**caseid year wtfactor**
Dependent:	**rearnq**
Independent:	**rsex**
Test:	**remploye rocsect2 ejbhrcax rnsoccl**
	tea schqual pschqual hedqual3
	ragecat3 gor2 country rage

 [NB: caseid is computed from the SPSS reserved variable name **$casenum** (the row number of the case in the mother file). It will be needed as a key for adding new variables from the mother file or for restoring the order of cases after any sorting.]

---

[1] I am seeking authorisation from Natcen and UKDS to have the small teaching file from this exercise stored on my site and downloadable on request.

1.3: Extract and save the selected variables

SPSS command **SAVE**

[General format}   **save**   outfile '<drive>:\<pathway>\<filename>.sav'
                                /keep <varlist>.

File >> New >> Syntax to open a new **Syntax Editor**:

For the selected variables (on my computer: yours may be different) type in the following lines:
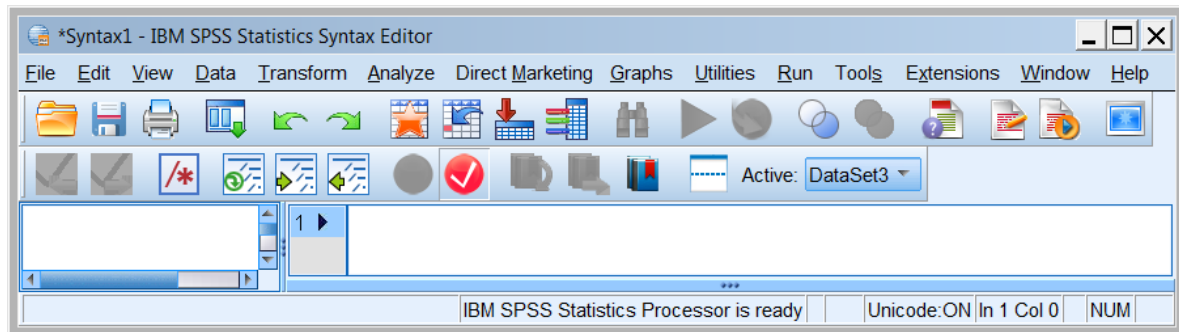
>   **title**   'Replication of exercise 3.2.4 (1989) on BSA 2009 to 2014'.
>   **save**   outfile = 'C: \Users\John\Desktop\Elaboration**bsa83-14_elab1.sav**'
>   /keep   caseid year WtFactor
>   REarnQ RSex
>   REmploye ROcSect2 EJbHrCaX RNSocCl
>   Tea SchQual PschQual HEdQual3
>   RAgecat3 gor2 country rage.

Don't forget the full stops (periods) to mark the end of each command !!

To run the job click on Run >> All

[File **bsa83-14_elab1.sav** has now been saved in folder **Elaboration** on my Desktop.]

| | bsa83-14_elab1 | 04/08/2016 18:30 | SPSS Statistics Data Document | 3,720 KB |

Close file   bsa1983-2014mother13

Keep the **Syntax Editor** open.

1.4 Navigate to folder **Elaboration** and find:
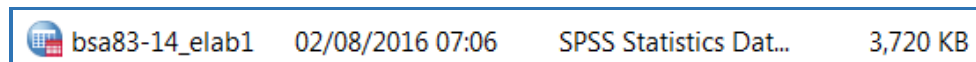
Double click on the icon to open the file: if it opens in Data View, switch to Variable View

|	Name	Measure	Label	Values	Missing	Decimals	Type	Width	Columns	Align	Role
1	caseid	Scale		None	None	2	Numeric	8	10	Right	Input
2	year	Scale	Year of ...	{0003, ...	None	0	Restricted ...	4	6	Right	Input
3	WtFactor	Scale	Final B...	{-99.00,...	LO - -1.00	2	Numeric	8	9	Right	Input
4	REarnQ	Ordinal	Respon...	{-99, Re...	LO - -1	0	Numeric	2	8	Right	Input
5	Rsex	Nominal	Person ...	{-9, Not ...	LO - -1	0	Numeric	2	8	Right	Input
6	Remploye	Nominal	Respon...	{-9, Not ...	LO - -1	0	Numeric	2	8	Right	Input
7	ROcSect2	Nominal	Which ...	{-99, Re...	LO - -1	0	Numeric	2	8	Right	Input
8	EJbHrCaX	Nominal	Hours R...	{-9, Ref...	LO - -1	0	Numeric	2	8	Right	Input
9	RNSocCl	Ordinal	Respon...	{-8, Ins...	LO - -1	0	Numeric	2	8	Right	Input
10	Tea	Ordinal	How old...	{-99, Re...	LO - -1	0	Numeric	2	8	Right	Input
11	SchQual	Nominal	Have y o...	{-9, Ref...	LO - -1	0	Numeric	2	8	Right	Input
12	PschQual	Nominal	Have y o...	{-9, Ref...	LO - -1	0	Numeric	2	8	Right	Input
13	HEdQual3	Nominal	Highest ...	{-8.00, ...	LO - -1.00	2	Numeric	8	10	Right	Input
14	RAgecat3	Ordinal	Age of r ...	{-9, DK/...	LO - -1	0	Numeric	2	10	Right	Input
15	GOR2	Nominal	Govern...	{-99, Re...	LO - -1	0	Numeric	2	8	Right	Input
16	Country	Nominal	England...	{-9, Not ...	LO - -1	0	Numeric	2	8	Right	Input
17	Rage	Scale	Person ...	{-99, Re...	LO - -1	0	Numeric	2	8	Right	Input
18											

The **Data Editor** contains 17 variables extracted from the mother file.

You can slide the column dividers to see more of the labels and hide the attributes you don't really need:

	Name	Measure	Label	Values	Missing	Decimals
1	caseid	Scale		None	None	2
2	year	Scale	Year of Interview	{0003, Constant}...	None	0
3	WtFactor	Scale	Final BSA weight	{-99.00, Refused}...	LO - -1.00	2
4	REarnQ	Ordinal	Respondent earnings quartiles (dv)	{-99, Refused}...	LO - -1	0
5	Rsex	Nominal	Person 1 SEX	{-9, Not answered}...	LO - -1	0
6	Remploye	Nominal	Respondent currently employee or self-employed dv	{-9, Not answered}...	LO - -1	0
7	ROcSect2	Nominal	Which of the types of organisation on this card do/did you work for?	{-99, Not answered}...	LO - -1	0
8	EJbHrCaX	Nominal	Hours R works per week, excluding overtime [employee]. DV	{-9, Refusal}...	LO - -1	0
9	RNSocCl	Ordinal	Respondent : social class [pre-SOC2000] best estimate dv	{-8, Insuff}...	LO - -1, 7	0
10	Tea	Ordinal	How old when completed your continuous full-time education?[compressed] dv	{-99, Ref}...	LO - -1	0
11	SchQual	Nominal	Have you passed any school examinations	{-9, Ref}...	LO - -1	0
12	PschQual	Nominal	Have you achieved any post-school qualifications	{-9, Ref}...	LO - -1	0
13	HEdQual3	Nominal	Highest educational qual obtained - dv	{-8.00, DK/Refusal/NA}...	LO - -1.00	2
14	RAgecat3	Ordinal	Age of respondent(grouped)<6 category> dv	{-9, DK/Ref}...	LO - -1	0
15	GOR2	Nominal	Government office region 2003 version	{-99, Refused}...	LO - -1	0
16	Country	Nominal	England, Scotland or Wales?	{-9, Not answered}...	LO - -1	0
17	Rage	Scale	Person 1 age last birthday	{-99, Refused}...	LO - -1	0

It is useful to have hard copies of file information.

1.5 Go back to your **Syntax Editor**
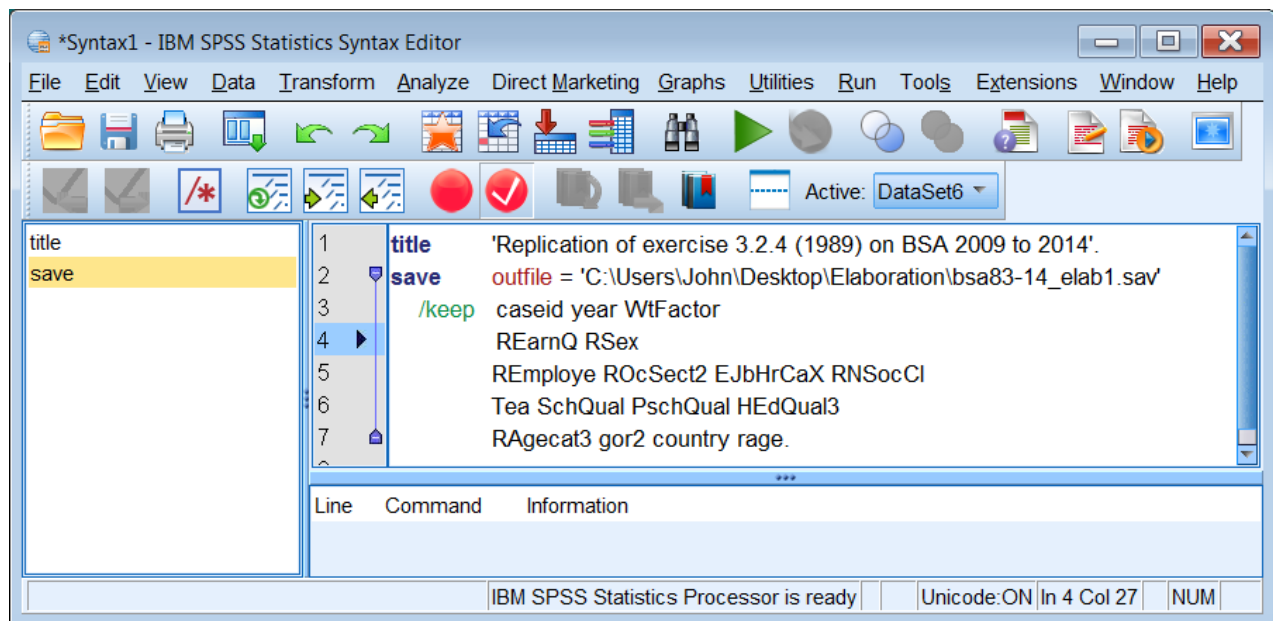
1.6: To see a list of variable labels, type (from line 8):

> **subtitle** 'Check contents of file'.
> **display** labels.   [NB: **display** command not available from the GUI]

To run the **display** command, place cursor in the line and use **Ctrl R** or press ▶.

Variable Labels

Variable	Position	Label
caseid	1	<none>
year	2	Year of Interview
WtFactor	3	Final BSA weight
REarnQ	4	Respondent earnings quartiles (dv)
Rsex	5	Person 1 SEX
Remploye	6	Respondent currently employee or self-employed dv
ROcSect2	7	Which of the types of organisation on this card do/did you work for?
EJbHrCaX	8	Hours R works per week, excluding overtime [employee]. DV
RNSocCl	9	Respondent : social class [pre-SOC2000] best estimate dv
Tea	10	How old when completed your continuous full-time education?[compressed] dv
SchQual	11	Have you passed any school examinations
PschQual	12	Have you achieved any post-school qualifications
HEdQual3	13	Highest educational qual obtained - dv
RAgecat3	14	Age of respondent(grouped)<6 category> dv
GOR2	15	Government office region 2003 version
Country	16	England, Scotland or Wales?
Rage	17	Person 1 age last birthday

Variables in the working file

[NB: The above table was edited in Pivot Tables to shorten inordinately long variable labels and drag the right hand margin out, extracted from the SPSS output with: [**right click] > Copy** then pasted into this tutorial with **Ctrl+V**]

1.7: To see a table of variable attributes, type:

display variables.   [NB: **display** command not available from the GUI]

To run the command, place cursor in the line and use **Ctrl** **R** or press ▶.

Variable Information

Variable		Label	Measurement Level	Role	Print	Write	Missing Values
caseid	1	<none>	Scale	Input	F8.2	F8.2	
year	2	Year of Interview	Scale	Input	N4	N4	
WtFactor	3	Final BSA weight	Scale	Input	F8.2	F8.2	Lowest thru -1.00
REarnQ	4	Respondent earnings quartiles	Ordinal	Input	F2	F2	Lowest thru -1
Rsex	5	Person 1 SEX	Nominal	Input	F2	F2	Lowest thru -1
Remploye	6	Employee or self-employed	Nominal	Input	F2	F2	Lowest thru -1
ROcSect2	7	Type of organisation work for?	Nominal	Input	F2	F2	Lowest thru -1
EJbHrCaX	8	Hours R works per week	Nominal	Input	F2	F2	Lowest thru -1
RNSocCl	9	R's Social class	Ordinal	Input	F2	F2	7, Lowest thru -1
Tea	10	Age completed full-time education	Ordinal	Input	F2	F2	Lowest thru -1
SchQual	11	Passed school examinations	Nominal	Input	F2	F2	Lowest thru -1
PschQual	12	Achieved post-school qualifications	Nominal	Input	F2	F2	Lowest thru -1
HEdQual3	13	Highest educational qual	Nominal	Input	F8.2	F8.2	Lowest thru -1.00
RAgecat3	14	Age of respondent (grouped)	Ordinal	Input	F2	F2	Lowest thru -1
GOR2	15	Government office region	Nominal	Input	F2	F2	Lowest thru -1
Country	16	England, Scotland or Wales?	Nominal	Input	F2	F2	Lowest thru -1
Rage	17	Person 1 age last birthday	Scale	Input	F2	F2	Lowest thru -1

Variables in the working file

[NB: Inordinately long variable labels edited down: **thru** replaces **through** in **Missing**]

1.8: To find out how many cases there are in the file (choose a variable with few categories: **rsex** has only two):

[**NB:** All tables in this and related exercises are from **unweighted** data]

frequencies rsex.

Statistics

Rsex Person 1 SEX

N	Valid	95630
	Missing	0

Rsex Person 1 SEX

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	42138	44.1	44.1	44.1
	2 Female	53492	55.9	55.9	100.0
	Total	95630	100.0	100.0	

The **Data Editor** contains all 95630 cases from all waves 1983 to 2014.

You can stop here and carry on later if you like, but make sure you **save the Syntax Editor** first.

To save the **Syntax Editor**:
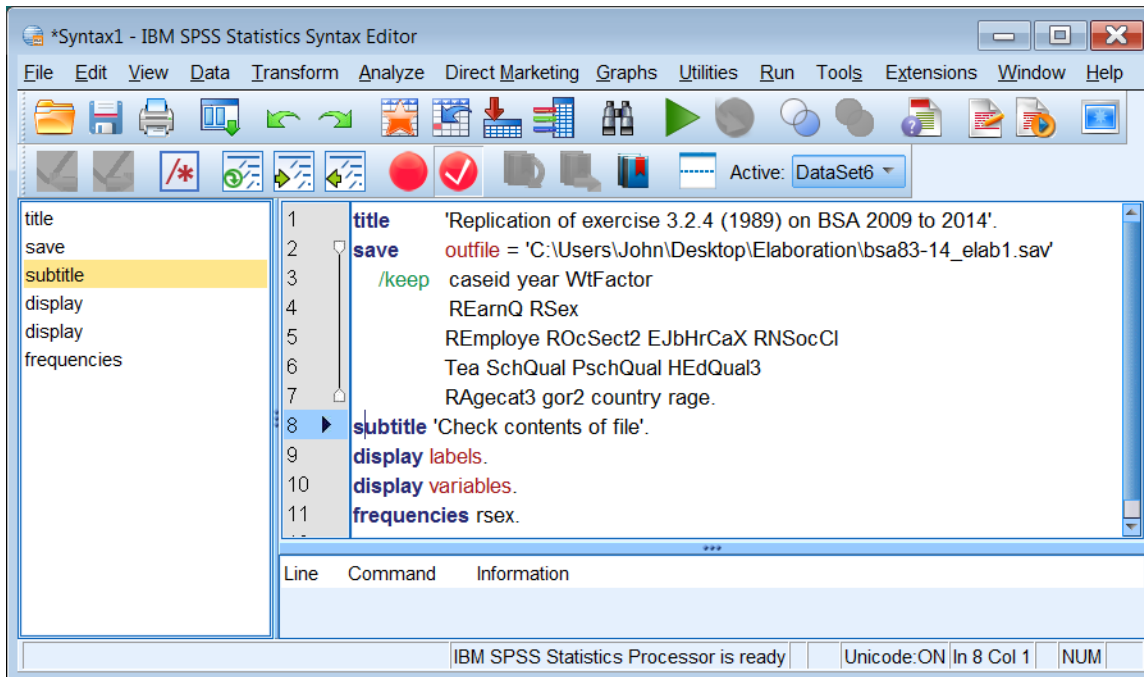
<mark>File</mark> >> <mark>Save As</mark>
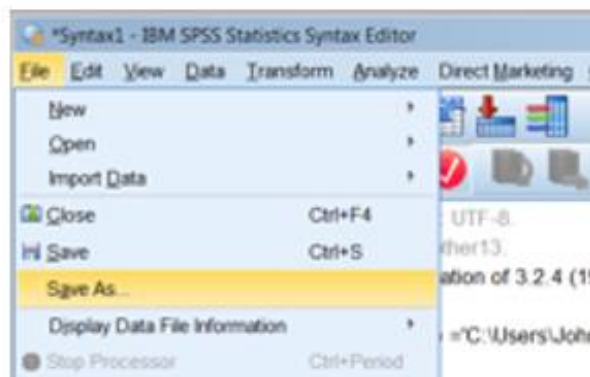
Change   **Syntax1**   to   **bsa83-14_elab1.sps**   and click on

The **Syntax Editor** has been saved in folder **Elaboration**:
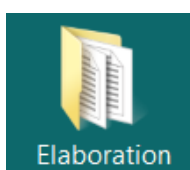
bsa83-14_elab1	02/08/2016 07:06	SPSS Statistics Data Docu...	3,720 KB
bsa83-14_elab1	02/08/2016 10:34	SPSS Statistics Syntax File	1 KB

Take a break.

Stage 2: Selecting years 2009 to 2014 only

Open file | 🗃 bsa83-14_elab1 | 02/08/2016 07:06 | SPSS Statistics Dat... | 3,720 KB

File >> New >> Syntax to open a new **Syntax Editor**:
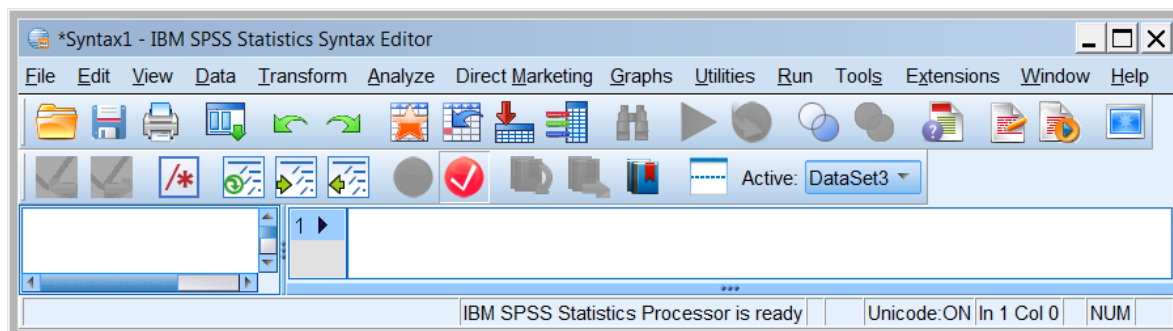
2.1: Select years 2009 to 2014 only.

> **title** 'Select years 2009 to 2014'.
> **select if** ((year ge 2009) and (year le 2014)).

[Note: **ge** means "greater than or equal to": **le** "less than or equal to": you can also use **>=** or **<=**]

This is a **permanent** change: cases from years 1983 to 2008 will be discarded.

> *check frequencies again.
> **frequencies** rsex year.

Statistics

		Rsex Person 1 SEX	year Year of Interview
N	Valid	19399	19399
	Missing	0	0

The number of case in the file has been reduced from 95,630 to 19,399

Rsex Person 1 SEX

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	8530	44.0	44.0	44.0
	2 Female	10869	56.0	56.0	100.0
	Total	19399	100.0	100.0	

year Year of Interview

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2009	3421	17.6	17.6	17.6
	2010	3297	17.0	17.0	34.6
	2011	3311	17.1	17.1	51.7
	2012	3248	16.7	16.7	68.4
	2013	3244	16.7	16.7	85.2
	2014	2878	14.8	14.8	100.0
	Total	19399	100.0	100.0	

2.2: As a prelude to choosing cutting points for dichotomies or trichotomies, or as a basis for selecting subsamples to create a more homogenous sample, check the initial frequencies of dependent, independent and test variables. Annotate tables with comments.

subtitle 'Check initial frequencies'.
*Dependent variable.
frequencies rearnq.

REarnQ Respondent earnings quartiles (dv)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Less than £930 p.m.	2157	11.1	22.0	22.0
	2 £931 - 1,600 p.m.	2430	12.5	24.8	46.8
	3 £1,601 - 2,500 p.m.	2196	11.3	22.4	69.2
	4 £2,501 or more p.m.	2006	10.3	20.5	89.7
	7	839	4.3	8.6	98.3
	8	168	0.9	1.7	100.0
	9	1	0.0	0.0	100.0
	Total	9797	50.5	100.0	
Missing	-97 Refused information	117	0.6		
	-1 Not applicable	9485	48.9		
	Total	9602	49.5		
Total		19399	100.0		

[Codes 7, 8 and 9 have no labels and should be missing: recode to negative.]

*Independent variable.
frequencies rsex.

Rsex Person 1 SEX

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	8530	44.0	44.0	44.0
	2 Female	10869	56.0	56.0	100.0
	Total	19399	100.0	100.0	

[OK as is]

*test variables.
frequencies remploye to country.

Work

Remploye Respondent currently employee or self-employed dv

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Employee	8438	43.5	85.1	85.1
	2 self-employed	1472	7.6	14.9	100.0
	Total	9910	51.1	100.0	
Missing	-9 Not answered	22	0.1		
	-8 Don't know	6	0.0		
	-3	4481	23.1		
	-1 Not applicable	4980	25.7		
	Total	9489	48.9		
Total		19399	100.0		

[Keep. Value label 2 needs capital S: -3 has no label]

Work (contd.)

ROcSect2 Which of the types of organisation on this card do/did you work for?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 PRIVATE SECTOR FIRM	10353	53.4	63.1	63.1
	2 NATIONALISED INDUSTRY	393	2.0	2.4	65.5
	3 OTHER PUBLIC SECTOR	4990	25.7	30.4	95.9
	4 CHARITY/ VOLUNTARY	520	2.7	3.2	99.0
	7 Other answer (WRITE IN)	89	0.5	0.5	99.6
	8 EDIT ONLY: Self-employed	2	0.0	0.0	99.6
	99	65	0.3	0.4	100.0
	Total	16412	84.6	100.0	
Missing	-98 Don't know	28	0.1		
	-3	1181	6.1		
	-1 Not applicable	1778	9.2		
	Total	2987	15.4		
Total		19399	100.0		

[Treat codes 7, 8 and 99 missing: no labels for -3, 99]

EJbHrCaX Hours R works per week, excluding overtime [employee]. DV

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 less than 10 hours a week	165	0.9	2.0	2.0
	1 10-15 hours a week	393	2.0	4.7	6.7
	2 16-23 hours a week	1009	5.2	12.2	18.9
	3 24-29 hours a week	550	2.8	6.6	25.5
	4 30 or more hours a week	6020	31.0	72.5	98.0
	5 varies too much to say	165	0.9	2.0	100.0
	Total	8302	42.8	100.0	
Missing	-9 Refusal	10	0.1		
	-8 Don't know	132	0.7		
	-1 Not applicable	10955	56.5		
	Total	11097	57.2		
Total		19399	100.0		

[Create new dichotomy: **workmode** 1 = Working full time 30 hours +, 2 = part-time]

RNSocCI Respondent : social class [pre-SOC2000] best estimate dv

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 I (SC=1)	1074	5.5	5.8	5.8
	2 II (SC=2)	6184	31.9	33.3	39.1
	3 III (non-manual) (SC=3)	4025	20.7	21.7	60.8
	4 III (manual) (SC=4)	3482	17.9	18.8	79.5
	5 IV (SC=5)	2933	15.1	15.8	95.3
	6 V (SC=6)	836	4.3	4.5	99.8
	7 Armed forces	34	0.2	0.2	100.0
	Total	18568	95.7	100.0	
Missing	-8 Insufficient information	212	1.1		
	-1 Not applicable	619	3.2		
	Total	831	4.3		
Total		19399	100.0		

[Will need to group into fewer categories: for purposes of this exercise, value 7 (Armed forces) will be treated as missing.]

Education

Tea How old when completed your continuous full-time education?[compressed] dv

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 15 or under	5105	26.3	26.4	26.4
	2 16	5305	27.3	27.5	53.9
	3 17	1537	7.9	8.0	61.9
	4 18	2215	11.4	11.5	73.3
	5 19 or over	4719	24.3	24.4	97.8
	6 Still at school	37	0.2	0.2	98.0
	7 Still at college/university	373	1.9	1.9	99.9
	97 Other answer (WRITE IN)	22	0.1	0.1	100.0
	Total	19313	99.6	100.0	
Missing	-99 Refusal	51	0.3		
	-98 Don't know	34	0.2		
	-1 Not applicable	1	0.0		
	Total	86	0.4		
Total		19399	100.0		

[Will need to group into fewer categories: code 97 clearly intended as missing. Discard categories 6 and 7 (still at school or college)]

SchQual Have you passed any school examinations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Yes	13176	67.9	68.2	68.2
	2 No	6130	31.6	31.8	100.0
	Total	19306	99.5	100.0	
Missing	-9 Not answered	54	0.3		
	-8 Don't know	38	0.2		
	-1 Not applicable	1	0.0		
	Total	93	0.5		
Total		19399	100.0		

[Already a dichotomy]

PschQual Have you achieved any post-school qualifications

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Yes	11606	59.8	60.1	60.1
	2 No	7704	39.7	39.9	100.0
	Total	19310	99.5	100.0	
Missing	-9 Not answered	58	0.3		
	-8 Don't know	30	0.2		
	-1 Not applicable	1	0.0		
	Total	89	0.5		
Total		19399	100.0		

[Already a dichotomy]

Education (contd.)

HEdQual3 Highest educational qual obtained - dv

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 Degree	1915	9.9	22.1	22.1
	2.00 Higher educ below degree/A level	2480	12.8	28.6	50.6
	3.00 O level or equiv/CSE	2192	11.3	25.3	75.9
	4.00 No qualification	2093	10.8	24.1	100.0
	Total	8680	44.7	100.0	
Missing	-8.00 DK/Refusal/NA	690	3.6		
	System	10029	51.7		
	Total	10719	55.3		
Total		19399	100.0		

[Superfluous decimal places: needs format (f2.0)]

Location

GOR2 Government office region 2003 version

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 North East	979	5.0	5.0	5.0
	2 North West	2408	12.4	12.4	17.5
	3 Yorkshire and Humberside	1660	8.6	8.6	26.0
	4 East Midlands	1624	8.4	8.4	34.4
	5 West Midlands	1817	9.4	9.4	43.8
	6 SW	1689	8.7	8.7	52.5
	7 Eastern	1936	10.0	10.0	62.4
	8 Inner London	712	3.7	3.7	66.1
	9 Outer London	1211	6.2	6.2	72.4
	10 South East	2576	13.3	13.3	85.6
	11 Wales	1054	5.4	5.4	91.1
	12 Scotland	1733	8.9	8.9	100.0
	Total	19399	100.0	100.0	

[Keep for now: may need fewer groups later, but could always compare two regions]

Country England, Scotland or Wales?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 England	16612	85.6	85.6	85.6
	2 Scotland	1733	8.9	8.9	94.6
	3 Wales	1054	5.4	5.4	100.0
	Total	19399	100.0	100.0	

[Keep for now]

2.3 Using **FREQUENCIES** to tabulate variables with many values (eg. Age last birthday) creates large tables: better to produce a summary with **DESCRIPTIVES**.

*rage has many values.
descriptives rage /sta min max.
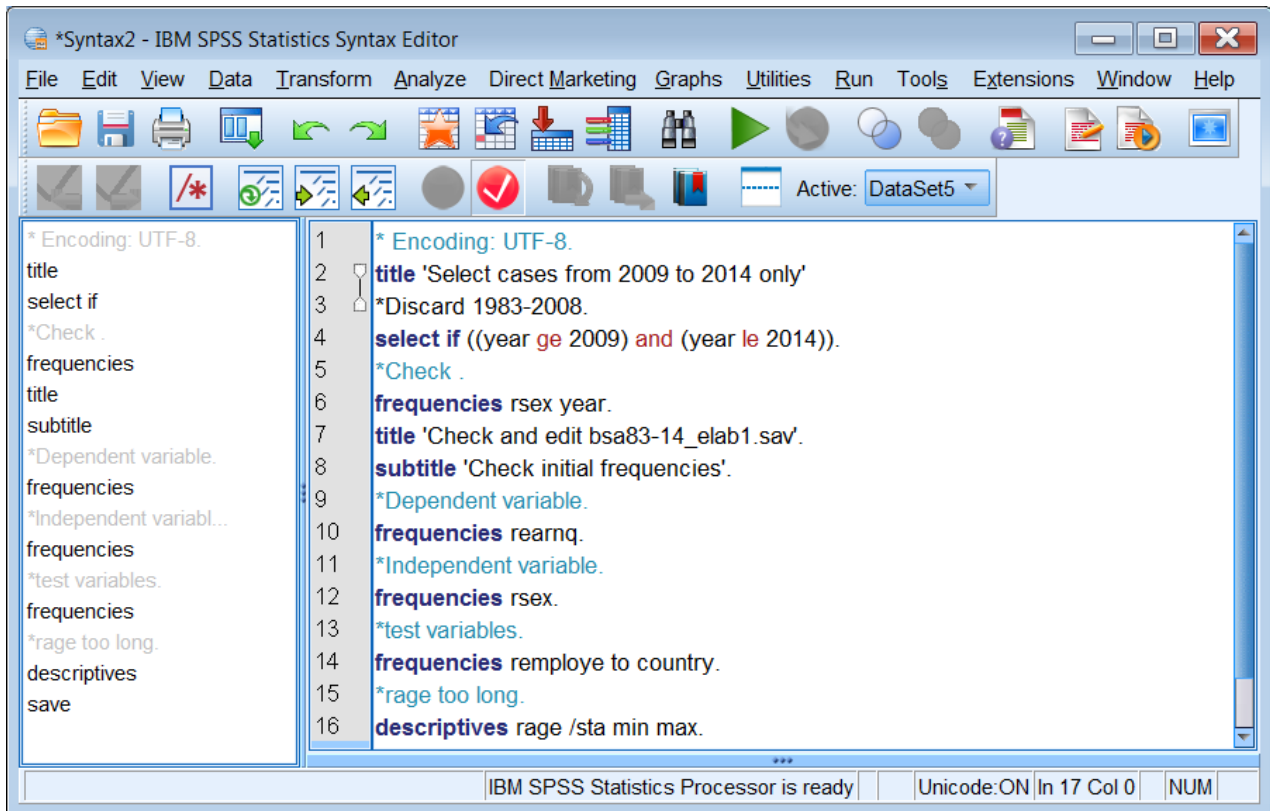
Descriptive Statistics

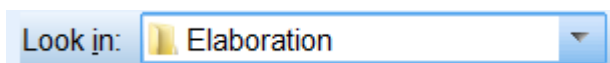|	N	Minimum	Maximum
Rage Person 1 age last birthday	19367	17	97
Valid N (listwise)	19367		

[97 = "97 or older", 98, "Don't know" and 99 "Refused" already missing]

2.4 Save your work again!

Syntax Editor ***Syntax2** should look something like this:

File >> **Save As**

Change   **Syntax2**   to   **bsa83-14_elab2.sps**   and click on

File **bsa83-14_elab1.sav** has been changed by selecting only cases from 2009 to 2014. You can save it with **Ctrl** **S** but best practice is to **save it with a new name**: **bsa83-14_elab2.sav**.

File >> **Save As**
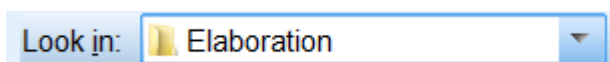
Change   **bsa83-14_elab1.sav**   to   **bsa83-14_elab2.sav**   and click on

Your new files are now both saved in folder **Elaboration**
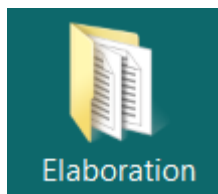
bsa83-14_elab2	02/08/2016 12:05	SPSS Statistics Data Document
bsa83-14_elab2	02/08/2016 11:20	SPSS Statistics Syntax File

As an alternative to <mark>File</mark> >> <mark>Save As</mark> you could also add a line of syntax:

      save outfile ='C:\Users\John\Desktop\Elaboration\bsa83-14_elab2.sav'.

but you would then have to navigate to the folder and open it. <mark>File</mark> >> <mark>Save As</mark> renames the working file and keeps it open, leaving the previous version unchanged.

You can stop here and carry on later if you like. **Take a break**.

Stage 3: Tidy up metadata:

3.1 For file **bsa83-14_elab2.sav** we now need to add or correct value labels, formats, measurement levels and missing values.

File >> New >> Syntax to open a new **Syntax Editor**:
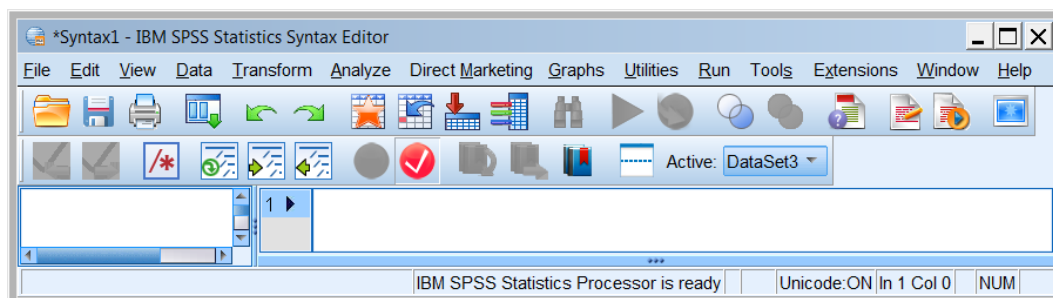
. . then type in the following syntax:

        title   'Tidy up metadata in bsa89-14_elab2.sav'.
        recode
                 rearnq (7=-7)(8=-8)(9=-9)(97=-97)
                /rocsect2 (7=-7)(8=-8)(99=-99)
                /EJbHrCaX (8=-8)(9=-9)
                /tea (97=-97)
                /hedqual3 (8=-8).
        formats caseid hedqual3 (f2.0).
        missing values
                REarnQ to EJbHrCaX tea to rage (lo thru -1)
                /RNSocCl (7, lo thru -1).
        add value labels
                rearnq -97 "Refused information"
                rocsect2 -7 'Other' -8 'Edit' -99 'Not answered'
                /EJbHrCaX -8 "Don't know" -9 'Refusal'
                /tea -97 "Other answer"
                /hedqual3 -8 'DK/Refusal/NA'.

[**NB: add value labels is** needed here: **value labels** would over-write all the other labels.]

To run the job click on Run >> All
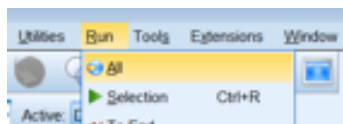
If it runs without error use **Ctrl** **S** to save the syntax. If there are errors keep correcting them until error-free, then **Ctrl** **S** again.

3.2 **Save your work**.

        Syntax Editor File >> Save As   **bsa83-14_elab3.sps**

        Data Editor   File >> Save As   **bsa83-14_elab3.sav**

Stage 4:   Derived variables

4.1 Create two derived variables **earngrp** and **workmode**:

earngrp (same as **rearnq**, but with categories relabelled to indicate quartiles:
approximately 25% of sample in each category)

workmode (hours worked reduced to two categories: 30 hours or more; under 30 hours)

Specify formats, measurement levels, missing values, variable and value labels (NB: new value labels needed for **earngrp** as income ranges for **rearnq** change between waves)

File >> New >> Syntax to open a new **Syntax Editor**:
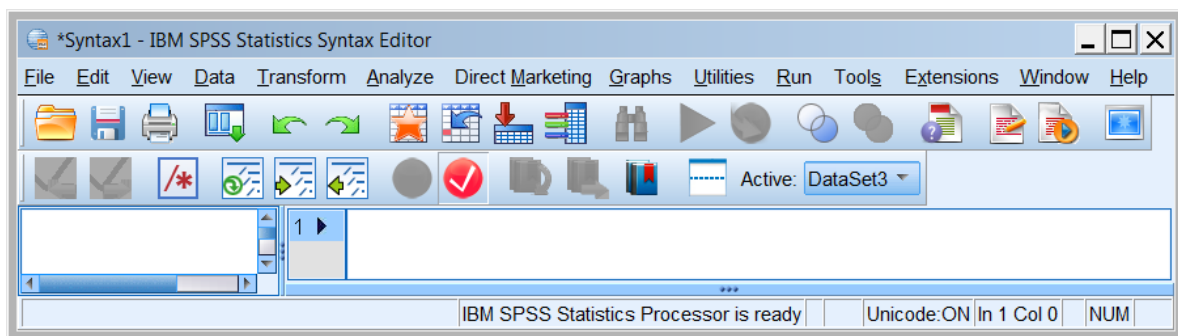
. . then type in the following syntax:

```
title      'Create derived variables'.
compute    earngrp = rearnq.
compute    workmode = EJbHrCaX.
recode     workmode (0 1 2 3 5 = 2)(4=1).
formats
           earngrp workmode (f2.0).
missing values
           earngrp workmode (lo thru -1).
variable level
           rearnq earngrp (ordinal)
           workmode (nominal) .
variable  labels
           earngrp "Quartile group of R's earnings from paid work"
           /workmode 'R working full- or part-time'.
value labels
           earngrp
                 1 'Q1' 2 'Q2' 3 'Q3' 4 'Q4'
           /workmode
                   1 'Full time'  2 'Part time'
                    -8 "Don't know" -9 'Refusal' -1 'Not employee'.
```
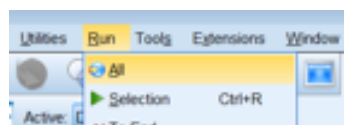

To run the job click on Run >> All

If it runs without error use **Ctrl** **S** to save the **Syntax Editor** as **bsa83-14_elab4.sps**. If there are errors keep correcting them until error-free, then **Ctrl** **S** again.

4.2: After modiflcations, always check your data!

Check frequencies for new variables.
frequencies earngrp workmode.

earngrp Quartile group of R's earnings from paid work

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Q1	2157	11.1	24.5	24.5
	2 Q2	2430	12.5	27.6	52.2
	3 Q3	2196	11.3	25.0	77.2
	4 Q4	2006	10.3	22.8	100.0
	Total	8789	45.3	100.0	
Missing	System	10610	54.7		
Total		19399	100.0		

workmode R working full- or part-time

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Full time	6020	31.0	72.5	72.5
	2 Part time	2282	11.8	27.5	100.0
	Total	8302	42.8	100.0	
Missing	System	11097	57.2		
Total		19399	100.0		

4.3: Select only cases with valid values for **earngrp** (and **check your data** again):

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

earngrp Quartile group of R's earnings from paid work

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Q1	2157	24.5	24.5	24.5
	2 Q2	2430	27.6	27.6	52.2
	3 Q3	2196	25.0	25.0	77.2
	4 Q4	2006	22.8	22.8	100.0
	Total	8789	100.0	100.0	

[NB: A quartile is **not a group**, it's a **cutting point**: value 1 = "Above top quartile", value 4 = "Below bottom quartile"]

[**NB:** This a **permanent** selection. All cases with no valid values for **earngrp** have been discarded and the sample size has now been reduced from 19,399 to 8,789]

4.4: Save your work again!

You could use ⌨Ctrl S again to save the current working file, but that would permanently lose cases from the earlier file(s). It is far preferable (and **good professional practice**) to **save a new file with a different name**: the earlier file(s) will be unaffected and cases will be retained.

File >> Save As to save your **Syntax Editor** as **bsa83-14_elab4.sps**

File >> Save As to save the **Data Editor** as **bsa83-14_elab4.sav**

Take a break. You can stop here and carry on later if you like.

Stage 5: Check contents of file

5.1: In file **bsa83-14_elab4.sav** check the frequencies of variables in the model.

File >> New >> Syntax

Dependent variable Quartile group for earnings from paid work: **rearnq**

> *check initial frequencies.
> *Dependent variable.
> **frequencies** rearnq.

REarnQ Respondent earnings quartiles (dv)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Less than £930 p.m.	2157	24.5	24.5	24.5
	2 £931 - 1,600 p.m.	2430	27.6	27.6	52.2
	3 £1,601 - 2,500 p.m.	2196	25.0	25.0	77.2
	4 £2,501 or more p.m.	2006	22.8	22.8	100.0
	Total	8789	100.0	100.0	

[OK, but labels confusing: use **earngrp** instead]

New dependent variable   Quartile group for earnings from paid work: **earngrp**

> *New dependent variable.
> **frequencies** earngrp.

earngrp Quartile group of R's earnings from paid work

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Q1	2157	24.5	24.5	24.5
	2 Q2	2430	27.6	27.6	52.2
	3 Q3	2196	25.0	25.0	77.2
	4 Q4	**2006**	22.8	22.8	100.0
	Total	8789	100.0	100.0	

[**My emphasis**: there are **2006** cases above the upper quartile. Later on we will use these cases as a criterion group.]

Independent variable   Sex   **rsex**

> *Independent variable.
> **frequencies** rsex.

Rsex Person 1 SEX

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	4299	48.9	48.9	48.9
	2 Female	4490	51.1	51.1	100.0
	Total	8789	100.0	100.0	

Test variables: Work

ROcSect2 Which of the types of organisation on this card do/did you work for?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 PRIVATE SECTOR	4648	52.9	61.7	61.7
	2 NATIONALISED	109	1.2	1.4	63.2
	3 OTHER PUBLIC	2455	27.9	32.6	95.8
	4 CHARITY/ VOLUNTARY	317	3.6	4.2	100.0
	Total	7529	85.7	100.0	
Missing	-99 Not answered	2	0.0		
	-98 Don't know	4	0.0		
	-7 Other	30	0.3		
	-3	633	7.2		
	-1 Not applicable	591	6.7		
	Total	1260	14.3		
Total		8789	100.0		

[Table edited to reduce inordinate length of value labels]

EJbHrCaX Hours R works per week, excluding overtime [employee]. DV

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 less than 10 hours a week	130	1.5	1.7	1.7
	1 10-15 hours a week	344	3.9	4.6	6.4
	2 16-23 hours a week	898	10.2	12.0	18.4
	3 24-29 hours a week	487	5.5	6.5	24.9
	4 30 or more hours a week	5459	62.1	73.2	98.1
	5 varies too much to say	140	1.6	1.9	100.0
	Total	7458	84.9	100.0	
Missing	-9 Refusal	3	0.0		
	-8 Don't know	104	1.2		
	-1 Not applicable	1224	13.9		
	Total	1331	15.1		
Total		8789	100.0		

[Used as source variable for **workmode**]

RNSocCl Respondent : social class [pre-SOC2000] best estimate dv

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 I (SC=1)	666	7.6	7.6	7.6
	2 II (SC=2)	3394	38.6	39.0	46.6
	3 III (non-manual) (SC=3)	1710	19.5	19.6	66.2
	4 III (manual) (SC=4)	1569	17.9	18.0	84.2
	5 IV (SC=5)	1107	12.6	12.7	97.0
	6 V (SC=6)	265	3.0	3.0	100.0
	Total	8711	99.1	100.0	
Missing	-8 Insuff	50	0.6		
	-1 Not applicable	12	0.1		
	7 Armed forces	16	0.2		
	Total	78	0.9		
Total		8789	100.0		

[Label for -8 should be "Insufficient information"]

Test variables: Education

Tea How old when completed your continuous full-time education?[compressed] dv

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 15 or under	1013	11.5	11.5	11.5
	2 16	2668	30.4	30.4	41.9
	3 17	784	8.9	8.9	50.9
	4 18	1301	14.8	14.8	65.7
	5 19 or over	2964	33.7	33.8	99.5
	6 Still at school	4	0.0	0.0	99.5
	7 Still at college/university	44	0.5	0.5	100.0
	Total	8778	99.9	100.0	
Missing	-99 Ref	1	0.0		
	-98 DK	3	0.0		
	-97 Other answer	7	0.1		
	Total	11	0.1		
Total		8789	100.0		

[Discard categories 6 and 7 (still at school, college or university)]

SchQual Have you passed any school examinations

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Yes	7518	85.5	85.6	85.6
	2 No	1265	14.4	14.4	100.0
	Total	8783	99.9	100.0	
Missing	-9 Ref	1	0.0		
	-8 DK	5	0.1		
	Total	6	0.1		
Total		8789	100.0		

PschQual Have you achieved any post-school qualifications

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Yes	6624	75.4	75.4	75.4
	2 No	2159	24.6	24.6	100.0
	Total	8783	99.9	100.0	
Missing	-9 Ref	1	0.0		
	-8 DK	5	0.1		
	Total	6	0.1		
Total		8789	100.0		

HEdQual3 Highest educational qual obtained - dv

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Degree	1185	13.5	30.1	30.1
	2 Higher educ below degree/A level	1389	15.8	35.3	65.5
	3 O level or equiv/CSE	1039	11.8	26.4	91.9
	4 No qualification	319	3.6	8.1	100.0
	Total	3932	44.7	100.0	
Missing	-8 DK/Refusal/NA	210	2.4		
	System	4647	52.9		
	Total	4857	55.3		
Total		8789	100.0		

[Will need to choose only one education variable]

Test variables: Age

RAgecat3 Age of respondent(grouped)<6 category> dv

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 18-25	424	4.8	7.5	7.5
	2 26-35	1193	13.6	21.2	28.7
	3 36-45	1551	17.6	27.6	56.3
	4 46-55	1433	16.3	25.5	81.8
	5 56-65	874	9.9	15.5	97.3
	6 >66	150	1.7	2.7	100.0
	Total	5625	64.0	100.0	
Missing	-9 DK/Ref	1	0.0		
	System	3163	36.0		
	Total	3164	36.0		
Total		8789	100.0		

[**RAgecat3 has t**oo many missing values: sample reduced to 5625. Better to derive new age groupings from **rage**]

Test variables: Location

GOR2 Government office region 2003 version

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 North East	444	5.1	5.1	5.1
	2 North West	1024	11.7	11.7	16.7
	3 Yorks and Humberside	725	8.2	8.2	25.0
	4 East Midlands	721	8.2	8.2	33.2
	5 West Midlands	801	9.1	9.1	42.3
	6 SW	772	8.8	8.8	51.1
	7 Eastern	958	10.9	10.9	62.0
	8 Inner London	346	3.9	3.9	65.9
	9 Outer London	563	6.4	6.4	72.3
	10 South East	1253	14.3	14.3	86.6
	11 Wales	416	4.7	4.7	91.3
	12 Scotland	766	8.7	8.7	100.0
	Total	8789	100.0	100.0	

[Can always derive different groupings, or, if sufficient cases, compare two regions]

Country England, Scotland or Wales?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 England	7607	86.6	86.6	86.6
	2 Scotland	766	8.7	8.7	95.3
	3 Wales	416	4.7	4.7	100.0
	Total	8789	100.0	100.0	

End of tutorial: 3.2.1.2 Elaboration 2 (Income differences BSA 2009 - 2014)
                           [Last updated: 11 August 2016]

Next tutorial: **3.2.1.3 Elaboration 3 (Income differences 2009 – 2014_CROSSTABS)**
                          (Zero-order tables to answer research questions 1 to 3)

    1: Is there a difference between the earnings (from paid work) of men and women?
    2: What other variables might account for differences in earnings?
    3: What effect do they have by themselves?

Appendix: SPSS syntax listings

Stage 1: Extraction of variables from mother file

bsa09-14_elab1.sps

```
* Encoding: UTF-8.
title 'Replication of 3.2.4 (1989) on BSA 2009 to 2014'.
*bsa83-14.
save outfile ='C:\Users\John\Desktop\Elaboration\bsa83-14_elab1.sav'
        /keep    caseid year WtFactor    REarnQ RSex
                 REmploye ROcSect2 EJbHrCaX RNSocCl
                 Tea SchQual PschQual HEdQual3
                 RAgecat3 gor2 country rage .
subtitle 'Check contents of file'.
display labels.
display variables.
*Check no of cases.
frequencies rsex.
```


Stage 2: Selecting years 2009 to 2014 only

bsa83-14_elab2.sps

```
title 'Select years 2009 to 2014 only'.
select if ((year ge 2009) and (year le 2014)).
*check frequencies.
frequencies rsex year.

subtitle 'Check initial frequencies'.
*Dependent variable.
frequencies rearnq.
*Independent variable.
frequencies rsex.
*test variables.
frequencies remploye to country.
*rage too long.
descriptives rage /sta min max.

*optional.
save outfile ='C:\Users\John\Desktop\Elaboration\bsa83-14_elab2.sav'.
```

Stage 3: Tidy up metadata:

bsa83-14_elab3.sps

```
title 'Tidy up metadata'.
formats
   hedqual3 (f2.0).
recode
   rearnq (7=-7)(8=-8)(9=-9)(97=-97)
   /rocsect2 (7=-7)(8=-8)(99=-99)
   /EJbHrCaX (8=-8)(9=-9)
   /tea (97=-97)
   /hedqual3 (8=-8).
missing values
   REarnQ to EJbHrCaX  tea to rage (lo thru -1)
   /RNSocCl (7, lo thru -1).
add value labels
   rearnq -97 "Refused information"
   /rocsect2 -7 'Other' -8 'Edit' -99 'Not answered'
   /EJbHrCaX -8 "Don't know" -9 'Refusal'
   /tea -97 "Other answer"
   /hedqual3 -8 'DK/Refusal/NA'.

*optional.
save outfile ='C:\Users\John\Desktop\Elaboration\bsa83-14_elab3.sav'.
```


Stage 4: Derived variables

bsa83-14_elab3.sps

```
title 'Derive new variables'.
compute    earngrp = rearnq.
compute    workmode = EJbHrCaX.
recode     workmode (0 1 2 3 5 = 2)(4=1).
formats    earngrp workmode (f2.0).
missing values    earngrp workmode (lo thru -1).
variable level  rearnq earngrp (ordinal) workmode (nominal) .
variable  labels
   earngrp "Quartile group of R's earnings from paid work"
   /workmode 'R working full- or part-time'.
value labels
   earngrp    1 'Q1' 2 'Q2' 3 'Q3' 4 'Q4'
   /workmode 1  'Full time'  2 'Part time'
             -8 "Don't know" -9 'Refusal' -1 'Not employee'.
*Check frequencies for new variables.
frequencies earngrp workmode.
*Select cases with valid values for earnings.
select if      not (missing (earngrp)).
frequencies    earngrp.

*optional.
save outfile ='C:\Users\John\Desktop\Elaboration\bsa83-14_elab4.sav'.
```

25