

**Block 3: Analysing two variables (and sometimes three)** [Draft only: 19 August 2013]

**3.1.4.4 Income differences - Choose test variables and cutting points**

**Research questions:**

Is there a difference between the earnings (from paid work) of men and women?

See previous sessions: [2.3.1.6.2: Specimen answer for tasks 3 and 4](#) and [3.1.4.1 Income differences work-through](#)

What other variables might account for differences in earnings?

What effect do they have by themselves?

What happens to any differences in earnings between men and women when controlling for these other variables?

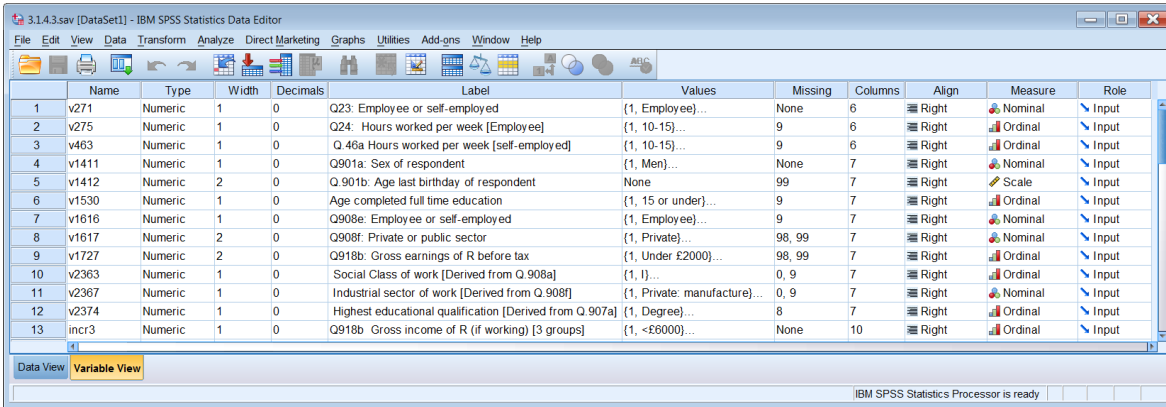
**Exemplar:** British Social Attitudes 1989

**Files:** [3.1.4.3.sav](#)  
[Created in session 3.1.4.3 and saved to **e:weebly downloads\bsa89\**]

	Question	record	column(s)	Name
<b>Dependent variable:</b>	Recoded personal gross earnings (three groups)			<b>incr3</b>
<b>Independent variable:</b>	Sex	Q.901a	14 11	<b>v1411</b>
<b>Test variables:</b>	<b>Work</b>			
	Employee or self-employed	Q.23	2 71	<b>v271</b>
	Hours worked, employee	Q.24	2 75	<b>v275</b>
	Hours worked, self-employed	Q.46a	4 61	<b>v461</b>
	Public or private sector	Q.908f	16 17-18	<b>v1617</b>
	Level of work	Q.908a	23 61	<b>v2361</b>
	<b>Education</b>			
	Terminal Education Age	Q.906a	15 30	<b>v1530</b>
	Level of education [derived]	Q.907b	23 74	<b>v2374</b>
	<b>Other</b>			
	Age last birthday	Q.901b	14 12-13	<b>v1412</b>

- Task**
- 1: Decide which test variables to use and choose cutting points
  - 2: Recode test variables into derived test variables with fewer categories
  - 3: Produce two-way contingency tables to show differences in earnings for categories within the derived test variables.

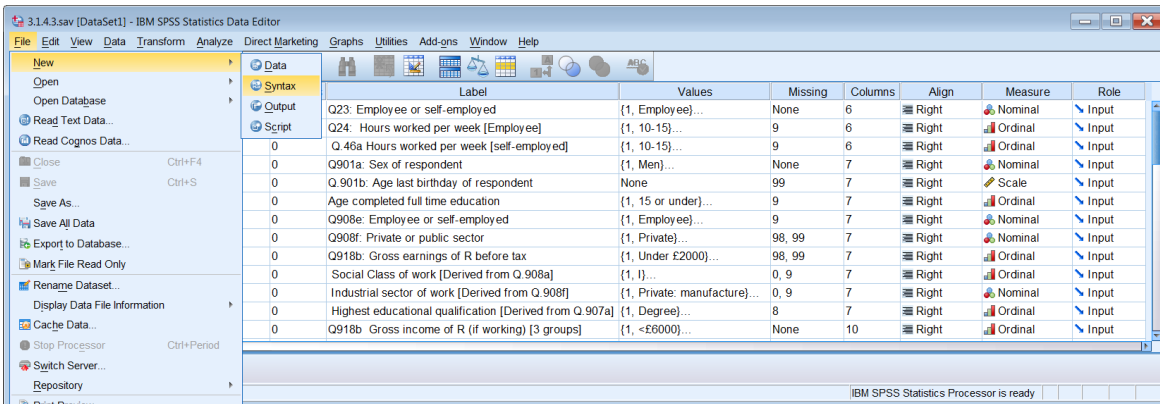
**Step 1:** Open file [3.1.4.3.sav](#) and adjust columns widths to see the labels clearly.



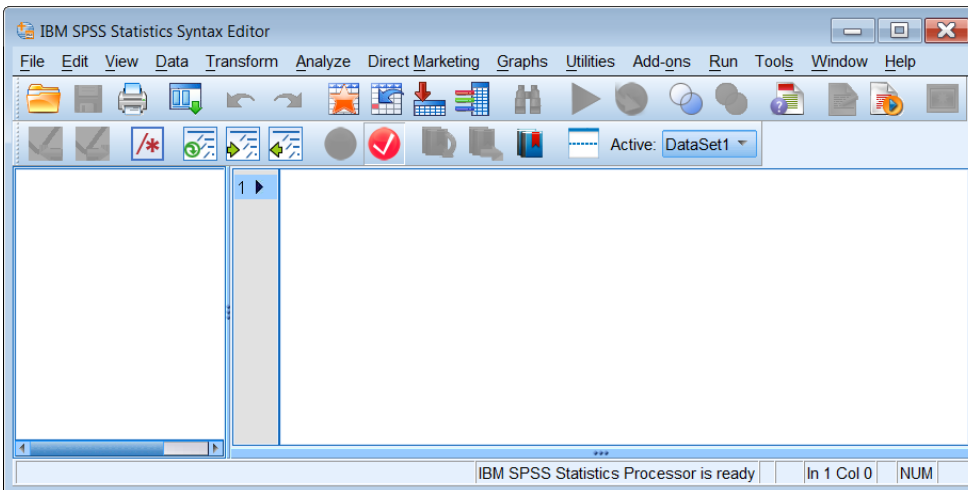
**Step 2: Rename demographic variables**

It's much easier to remember mnemonic names for commonly used demographics such as **sex** or **age** than to remember names like **v1411** or **v1412**, so we can now **RENAME** them.

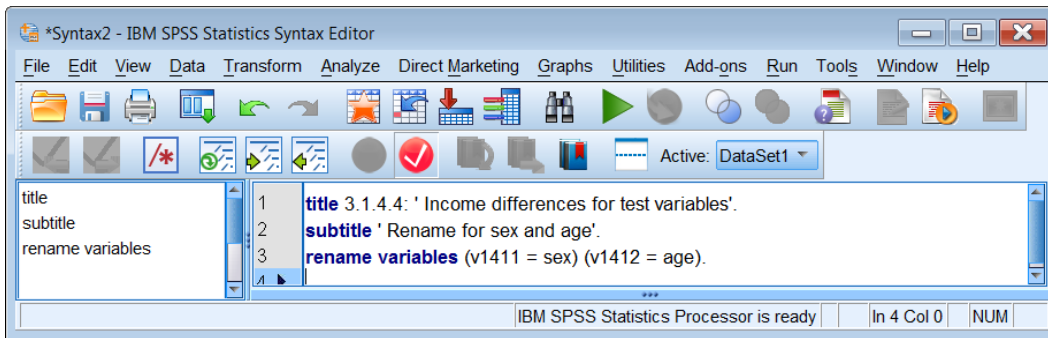
**File > New > Syntax**



to open a new **Syntax Editor**:



**title 3.1.4.4: ' Income differences for test variables'**  
**subtitle ' Rename for sex and age'**  
**rename variables (v1411 = sex) (v1412 = age).**



Place the cursor on the **title** command and press **Run > →| To End**. On rows 4 and 5 of the **Data Editor** the variable names have been changed:

4	sex	Numeric	1	0	Q901a: Sex of... {
5	age	Numeric	2	0	Q.901b: Age l... N

### Step 3: Choose cutting points for recoding test variables

In order to reduce the number of categories we are dealing with and also to optimise the individual cell counts used as a base for percentages, we need to decide on cutting points for the test variables. Apart from **sex** and **v271** which are already dichotomies, we could create fewer groups by simply recoding the original variables, but that would lose information from the original data that we might need later. It is safer to **RECODE** the other variables **INTO** new derived variables.

The table below summarises what we are about to do.

	Source	Target	Categories		
			1	2	3
	<b>v1411</b>	<b>sex</b>	Men	Women	
Employee or self-employed	<b>v271</b>		Employee	Self-employed	
Hours worked, employee	<b>v275</b>	<b>workmode</b>	Part time	Full time	
Hours worked, self-employed	<b>v461</b>				
Public or private sector	<b>v1617</b>	<b>sector</b>	Private	Public	
Social class of work	<b>v2361</b>	<b>class</b>	Non-Manual	Manual	
Terminal Education Age	<b>v1530</b>	<b>tea</b>	15 or under	16 and 17	18 or over
Level of education [derived]	<b>v2374</b>	<b>edlevel</b>	>=A-level	O, CSE	None
Age last birthday	<b>v1412</b>	<b>agegroup</b>	Problematic with different retirement ages		

Weekly hours worked requires a conditional transformation. Age is problematic because of different retirement ages for men (65) and women (60). To keep things manageable we shall need to reduce the data by restricting the sample to a more homogenous group of people below pensionable age working 30 hours or more per week, which we define as **working full-time**.

### Step 4: Recode test variables into derived variables with fewer categories.

First let's generate dichotomies for our test variables:

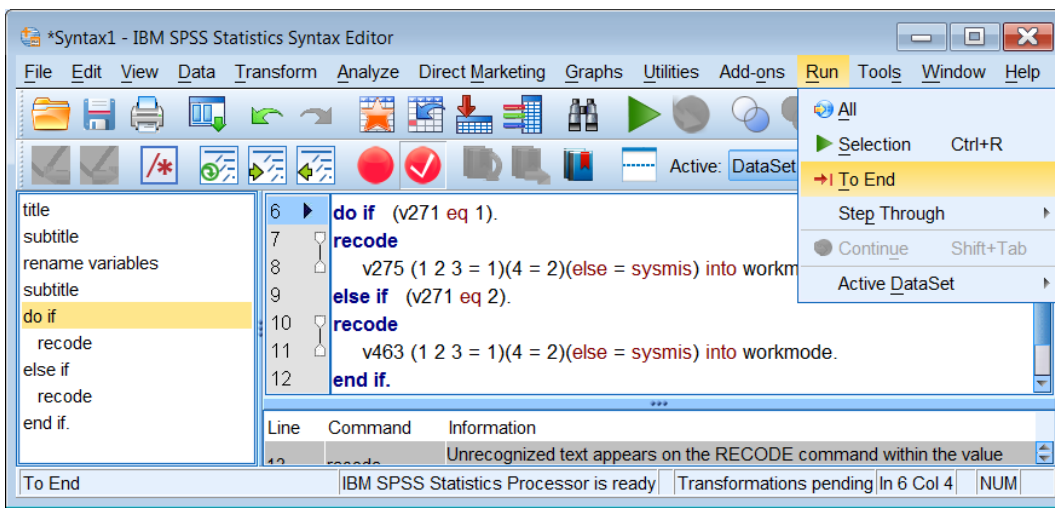
Being self-employed or an employee is already a dichotomy. Hours worked needs to be grouped into part-time and full-time, RG Social Class into manual and non-manual, industrial sector into private and public and educational qualifications into high and low. The cutting points can be arbitrary to divide into approximately equal-sized groups, or based on sociologically meaningful criteria.

Our first test variable **v271** is already dichotomised into self-employed or employee. The second one, **weekly hours worked**, needs to be recoded into a derived variable **workmode** with two categories, working **part-time** (under 30 hours) and **full-time** (30 or more hours). This requires a conditional transformation as the data are in different places for self-employed and employees:

```

subtitle 'Recode test vars into fewer categories'.
do if (v271 eq 1).
recode
    v275 (1 2 3 = 1)(4 = 2)(else = sysmis) into workmode.
else if (v271 eq 2).
recode
    v463 (1 2 3 = 1)(4 = 2)(else = sysmis) into workmode.
end if.

```



The new variable **workmode** has been appended to the file with **Decimals** set at **2** and **Measure** displayed as **Unknown**. There are more new variables to create, so we'll set **Decimals** and **Measure** for them all later.

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
v2374	Numeric	1	0	Highest educational qualification [Derived from Q.907a]	{1, Degree}...	8	7	Right	Ordinal	Input
incr3	Numeric	1	0	Q918b Gross income of R (if working) [3 groups]	{1, <£6000}...	None	10	Right	Ordinal	Input
workmode	Numeric	8	2		None	None	10	Right	Unknown	Input

**Social class of work** divides neatly into **manual** and **non-manual** in approximately equal numbers 52% / 48%.

**Industrial sector** [v2367] didn't look as if it made much difference to earnings on its own and can perhaps be left out at this stage.

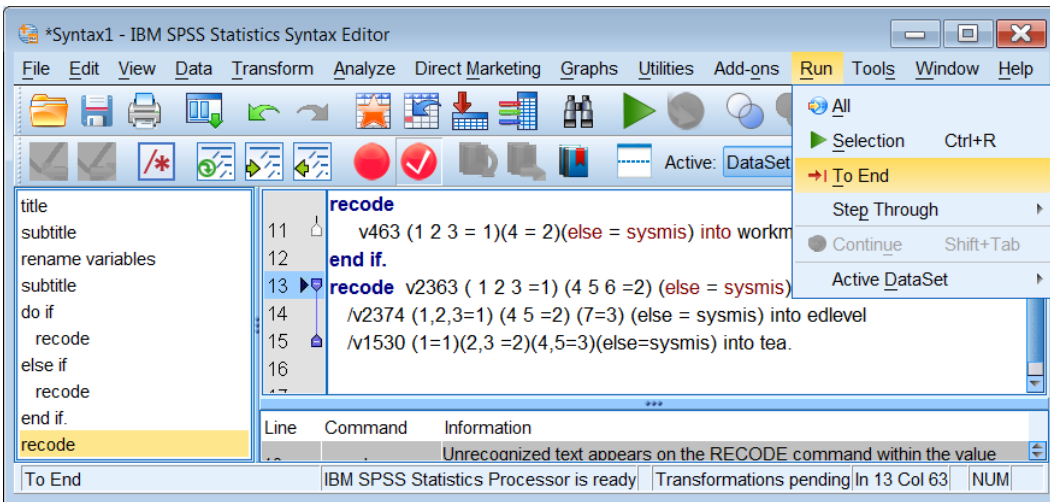
**Educational level** [v2374] splits 49% / 51% at O-level, but perhaps A-level at 31% / 69% is a better criterion to define "high" education? Also 42% have no qualifications at all, so perhaps three groups (A-level or above, O-level/CSE, and None) are better than two. The four people with foreign qualifications can be left out.

**Terminal Education Age** is better divided into three groups 15 or under, 16-17 and 18 or over

```

recode v2363 ( 1 2 3 =1) (4 5 6 =2) (else = sysmis) into class
/v2374 (1,2,3=1) (4 5 =2) (7=3) (else = sysmis) into edlevel
/v1530 (1=1)(2,3 =2)(4,5=3)(else=sysmis) into tea.

```

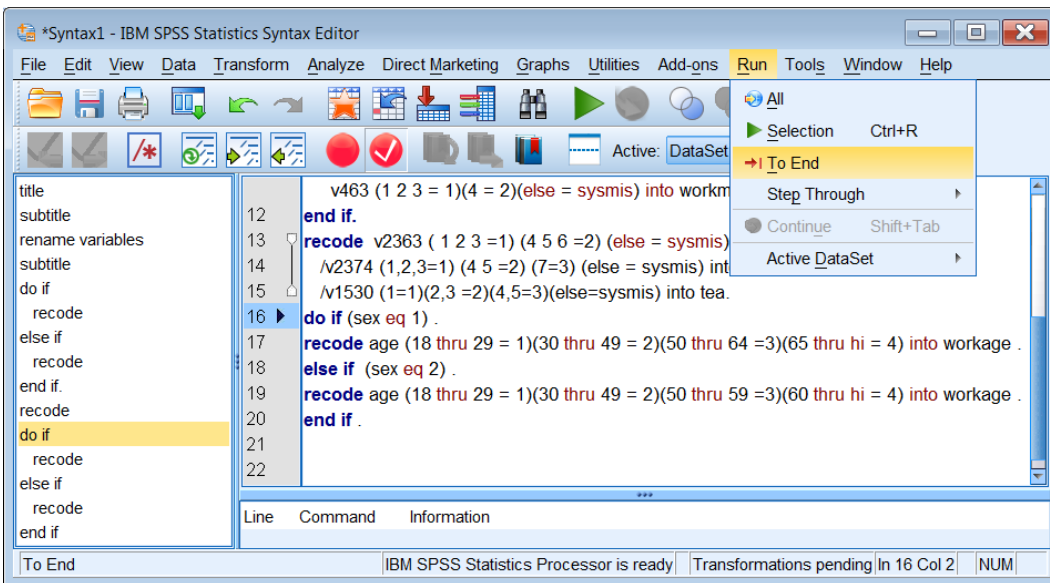


Age (last birthday) needs a conditional transformation to take account of the different pensionable ages for men (65) and women (60) in 1989.

```

do if (sex eq 1) .
recode age (18 thru 29 =1)(30 thru 49 = 2)(50 thru 64 =3)(65 thru hi = 4) into workage .
else if (sex eq 2) .
recode age (18 thru 29 =1)(30 thru 49 = 2)(50 thru 59 =3)(60 thru hi =4) into workage .
end if .

```



Derived variables **class**, **edlevel**, **tea** and **workage** have been appended to the file, again with **Decimals** set at **2** and **Measure** displayed as **Unknown**.

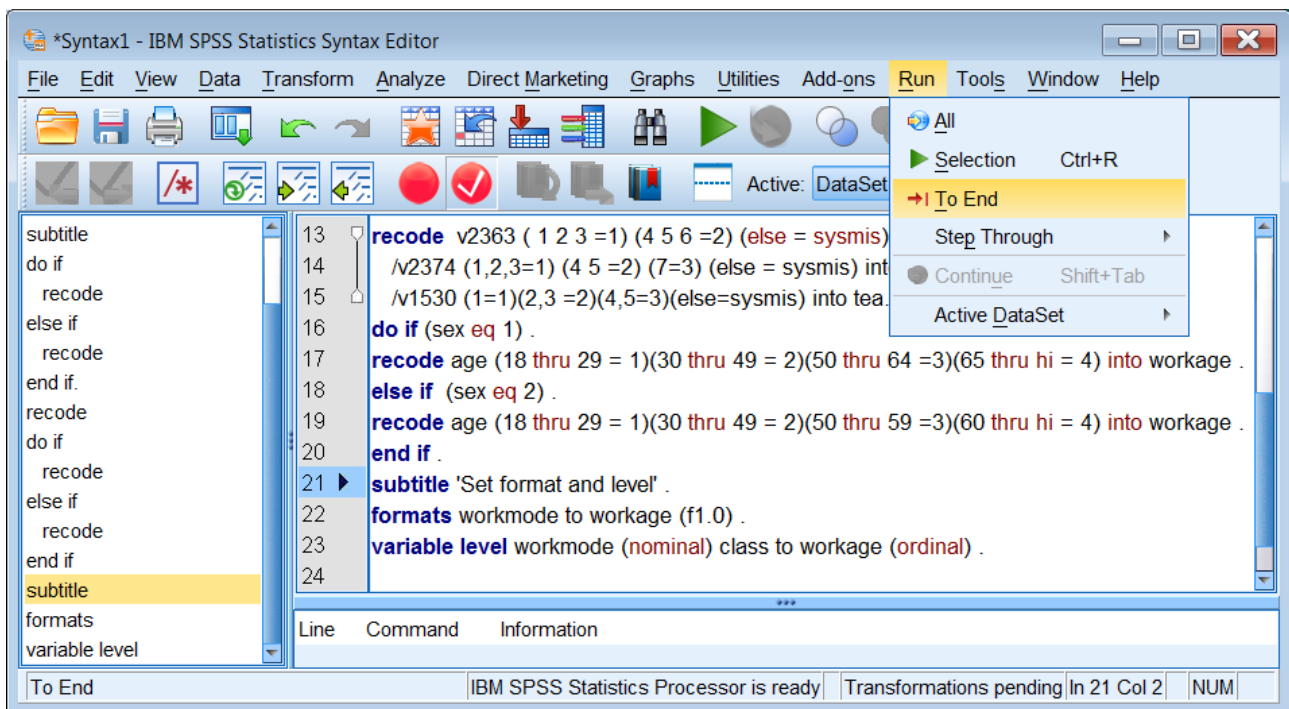
Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role	
13	incr3	Numeric	1	0	Q918b Gross income of R (if working) [3 groups]	{1, <£6000}...	None	10	Right	Ordinal	Input
14	workmode	Numeric	8	2		None	10	Right	Unknown	Input	
15	class	Numeric	8	2		None	10	Right	Unknown	Input	
16	edlevel	Numeric	8	2		None	10	Right	Unknown	Input	
17	tea	Numeric	8	2		None	10	Right	Unknown	Input	
18	workage	Numeric	8	2		None	10	Right	Unknown	Input	

## Step 5: Set Decimals and Measure for the new variables:

subtitle 'Set format and level' .

formats workmode to workage (f1.0) .

variable level workmode (nominal) class to workage (ordinal) .



Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
incr3	Numeric	1	0	Q918b Gross income of R (if working) [3 groups]	{1, <£6000}...	None	10	Right	Ordinal	Input
workmode	Numeric	1	0		None	None	10	Right	Nominal	Input
class	Numeric	1	0		None	None	10	Right	Ordinal	Input
edlevel	Numeric	1	0		None	None	10	Right	Ordinal	Input
tea	Numeric	1	0		None	None	10	Right	Ordinal	Input
workage	Numeric	1	0		None	None	10	Right	Ordinal	Input

It's a good idea to **save** our work regularly, so at this point we should save both the **Syntax Editor** and the **Data Editor** with the prefix **3.1.4.4** to tally with this tutorial (**3.1.4.4.sps** and **3.1.4.4.sav**).

## Step 6: Add dictionary information for the new variables

subtitle 'Add dictionary information'

variable labels

workmode 'Mode of work'

/class 'Social class of work'

/edlevel 'Highest qualification level'

/tea 'Age completed full-time education'

/workage 'Age group if working' .

value labels

workmode 1 'Parttime' 2 'Fulltime'

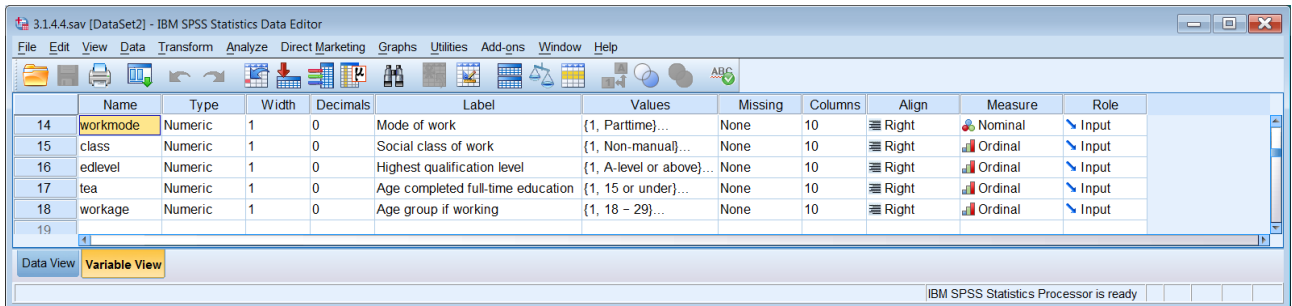
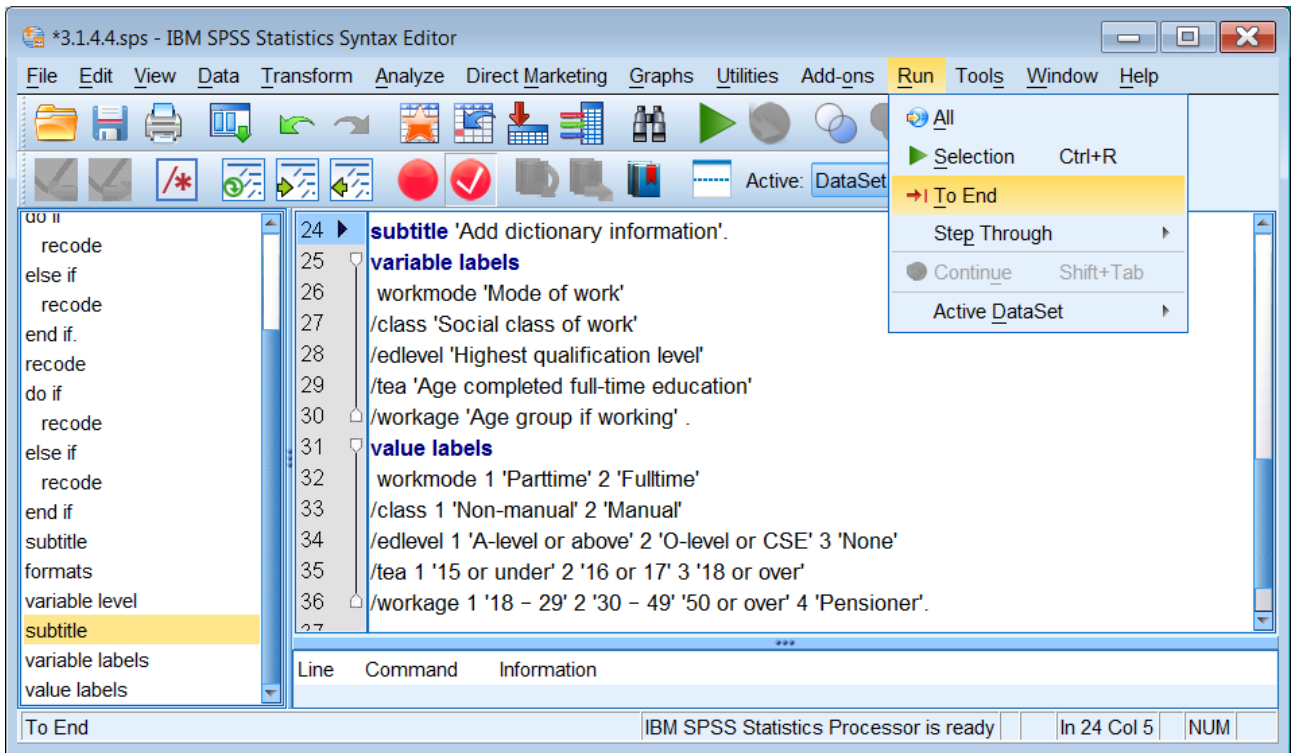
/class 1 'Non-manual' 2 'Manual'

/edlevel 1 'A-level or above' 2 'O-level or CSE' 3 'None'

/tea 1 '15 or under' 2 '16 or 17' 3 '18 or over'.

/workage 1 '18 – 29' 2 '30 – 49' 3 '50 or over' 4 'Pensioner' .

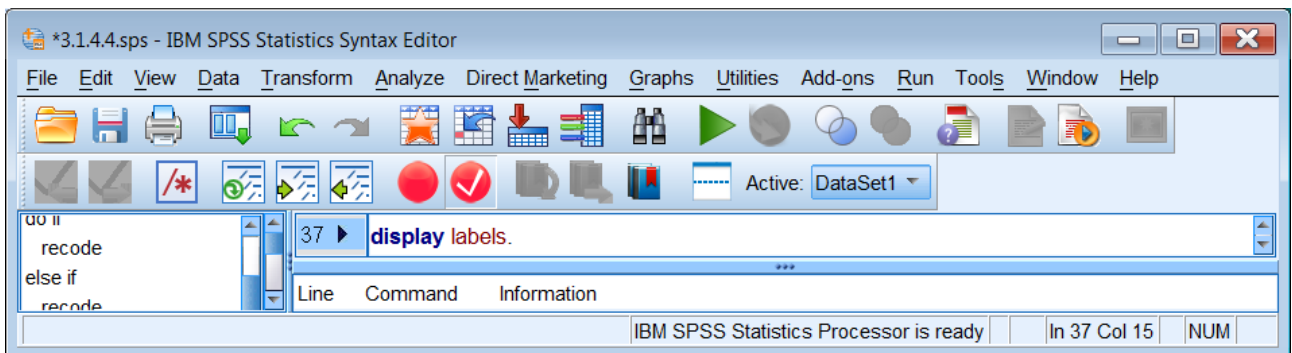




[CTRL]S to **save the file**, then check contents.

### Step 7: Check your file

**display labels .**

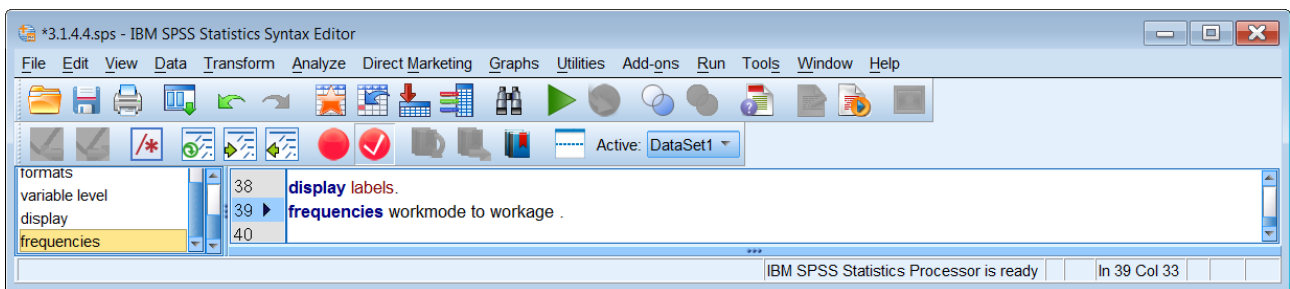


**Variable Labels**

Variable	Position	Label
v271	1	Q23: Employee or self-employed
v275	2	Q24: Hours worked per week [Employee]
v463	3	Q.46a Hours worked per week [self-employed]
sex	4	Q901a: Sex of respondent
age	5	Q.901b: Age last birthday of respondent
v1530	6	Age completed full time education
v1616	7	Q908e: Employee or self-employed
v1617	8	Q908f: Private or public sector
v1727	9	Q918b: Gross earnings of R before tax
v2363	10	Social Class of work [Derived from Q.908a]
v2367	11	Industrial sector of work [Derived from Q.908f]
v2374	12	Highest educational qualification [Derived from Q.907a]
incr3	13	Q918b Gross income of R (if working) [3 groups]
workmode	14	Mode of work
class	15	Social class of work
edlevel	16	Highest qualification level
tea	17	Age completed full-time education
workage	18	Age group if working

Variables in the working file

**frequencies workmode to workage .**



[CTRL]S to **save the file**, then run job to get:

**Statistics**

		Mode of work	Social class of work	Highest qualification level	Age completed full-time education	Age group if working
N	Valid	1682	2846	3005	2961	0
	Missing	1343	179	20	64	3025

**Mode of work**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Parttime	317	10.5	18.8	18.8
	Fulltime	1365	45.1	81.2	100.0
	Total	1682	55.6	100.0	
Missing	System	1343	44.4		
Total		3025	100.0		

**Social class of work**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Non-manual	1487	49.2	52.2	52.2
	Manual	1359	44.9	47.8	100.0
	Total	2846	94.1	100.0	
Missing	System	179	5.9		
Total		3025	100.0		



**Highest qualification level**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A-level or above	944	31.2	31.4	31.4
	O-level or CSE	778	25.7	25.9	57.3
	None	1283	42.4	42.7	100.0
	Total	3005	99.3	100.0	
Missing	System	20	.7		
Total		3025	100.0		

**Age completed full-time education**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	15 or under	1421	47.0	48.0	48.0
	16 or 17	972	32.1	32.8	80.8
	18 or over	568	18.8	19.2	100.0
	Total	2961	97.9	100.0	
Missing	System	64	2.1		
Total		3025	100.0		

**Age group if working**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18 – 29	668	22.1	22.1	22.1
	30 – 49	1124	37.2	37.2	59.2
	50 or over	538	17.8	17.8	77.0
	Pensioner	695	23.0	23.0	100.0
	Total	3025	100.0	100.0	

As a check on sex and (working) agegroups:

**Q901a: Sex of respondent \* Age group if working Crosstabulation**

Count		Age group if working				Total
		18 – 29	30 – 49	50 or over	Pensioner	
Q901a: Sex of respondent	Men	321	493	318	261	1393
	Women	347	631	220	434	1632
Total		668	1124	538	695	3025

These all seem to be OK.

[NB: Missing values on the original test variables have been set to system missing on these derived variables.]

**End of session: 3.1.4.4 Income differences - Choose test variables and cutting points**

**Back to:** [Block 3 Analysing two variables \(and sometimes three\)](#)  
[3.1 Two variables \(CROSSTABS\)](#)

[3.1.4.3 Income differences for test variables](#) [d]

**Forward to:** [3.1.4.5 Income differences for derived test variables](#) [b]