Analysing Quantitative Data Computer Lab 1

Welcome to the first quantitative data analysis lab for Social Research Methods. There are four computer lab sessions as part of this module. In these sessions you will learn how to perform some straightforward quantitative data analysis techniques using the computer program, SPSS.

This workbook will talk you through the activities you should undertake in the first lab session. Please work through these activities at your own pace. Please feel free to work in pairs and to discuss the activities with your classmates. Teaching staff will be on hand to answer any questions, please do not hesitate to ask for assistance.

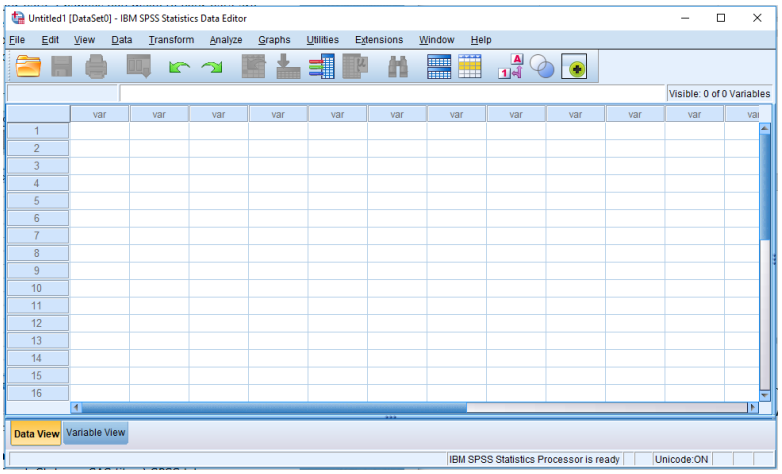
This first session will provide an introduction to SPSS and you will learn how to produce some simple univariate descriptive statistics (e.g. the mean, median and mode of a variable). You should take a little time to practice these activities outside class, as each week's lab will build on the activities from the week before. You can download SPSS to use on your own computer here. If you have any difficulty downloading SPSS you should seek assistance from the IT Services Help Desk.

Section 1: Getting Started in SPSS

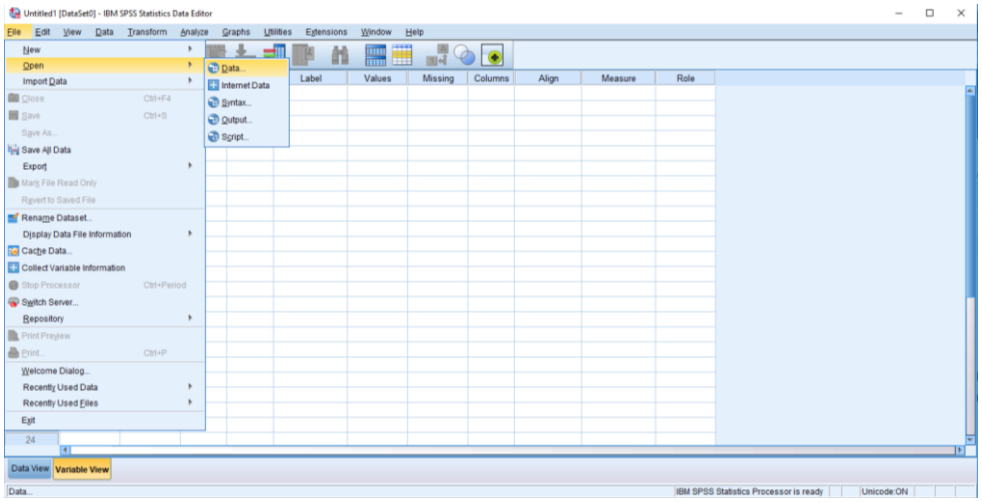
To open SPSS, click **START**, then navigate to '**IBM SPSS Statistics 25**'. You should ensure that you open version 25, there may be older versions of the software installed on the computer.

When SPSS is started for the first time you may be presented with a dialog box like the one below. Tick the '**Don't show this dialog in the future**' box (on the bottom left hand side, red arrow), then click **CLOSE** (on the bottom right hand side, green arrow). Do not worry if this dialog box does not show up in the lab, just move on to the next step.

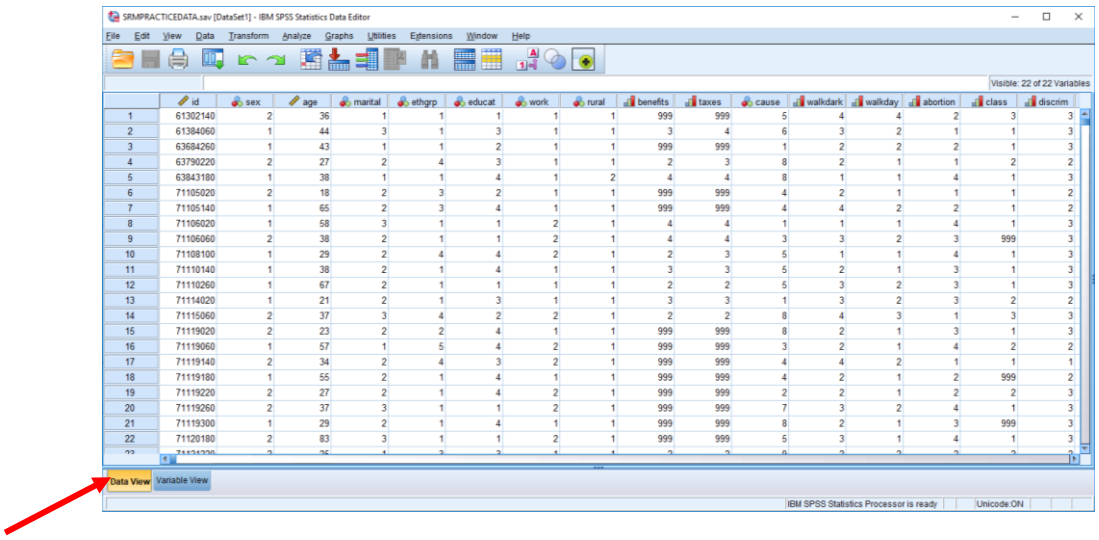
You should now be able to see a window that looks something like an excel spreadsheet (see below). This is your SPSS Data Editor window.



On the week 4 VLE page (where you got this workbook) you will see a link to the '**SRM Practice Data**'. Download this file onto your computer. Click on **FILE**, **OPEN**, **DATA** and open the data set from where you saved it. These data were collected from a sample of the UK population aged 16 and over.



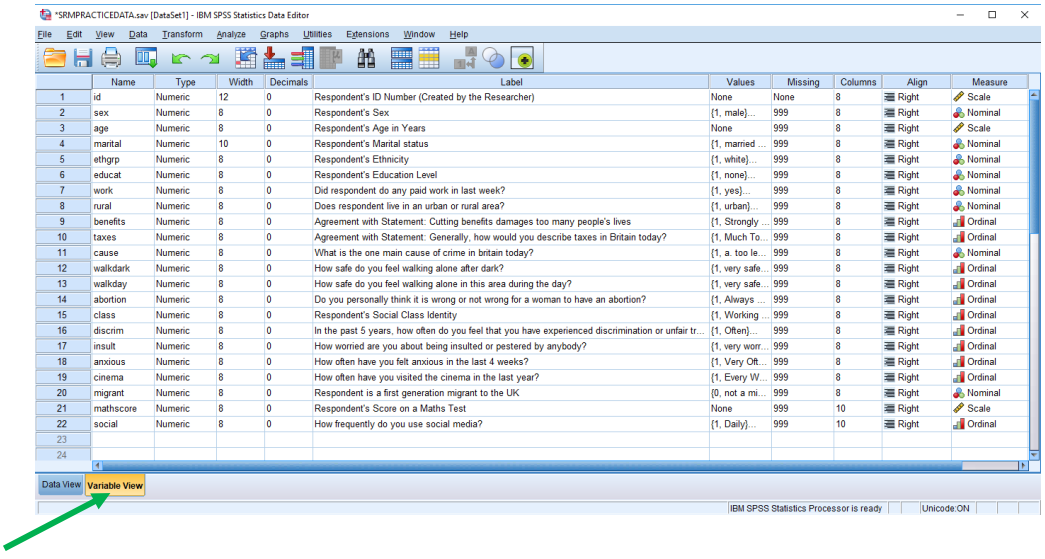
Now your data is open you can view it in two different ways. On the bottom right hand side of your Data Editor window you will see two tabs, 'Data View' and 'Variable View'. If you click on **Data View** (see red arrow below) your window will look like the image below.



Each row in this spread sheet contains information for each respondent who completed this survey. Each column in this spread sheet contains information for each piece of information that was collected in this survey.

Each piece of information is called a 'variable'. A variable is an entity that can take on different values, for example an individual's age is a variable. For a more detailed introduction to variables see here.

We can see the name of all the variables in the data set at the top of each column. However the variable's short names do not tell us very much about what has been measured and what the numbers shown actually mean. To get this information you can look at the Variable View tab. Click on the **Variable View** tab at the bottom of the window (see green arrow below).



In Variable View each row gives you information about each of the variables in the data set. There is more information on this screen than you need, you only need pay attention to elements highlighted below.

The first column labelled 'Name' tells you the name of each variable (e.g. age, sex, rural, class).

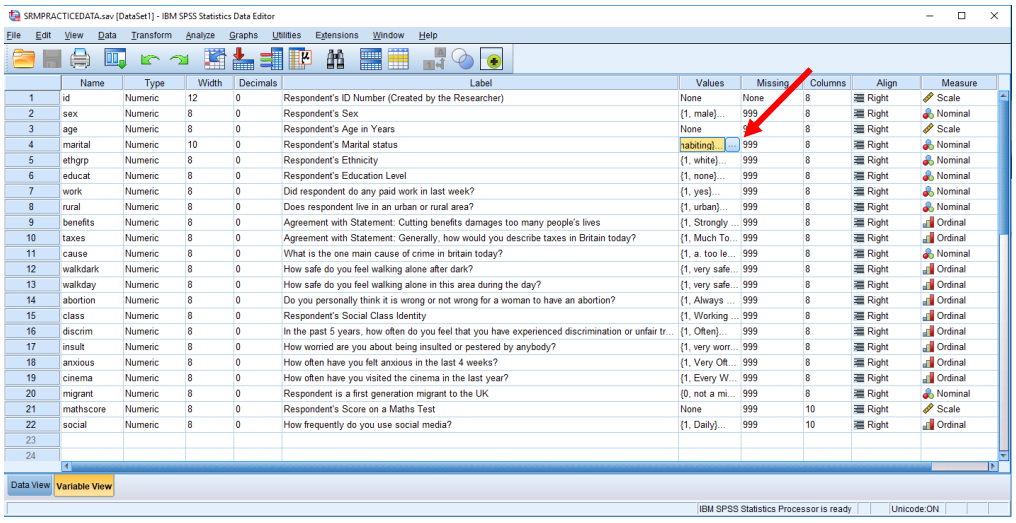
The fifth column labelled 'Label' tells you a little more about this variable and what it shows. For example we can see that the variable 'marital' contains information on the respondent's marital status, and the variable 'social' contains information on how frequently the respondent uses social media.

The sixth column labelled 'Values' tells you what each of the numbers in the spreadsheet mean. The information here will vary depending on the type of variable you have.

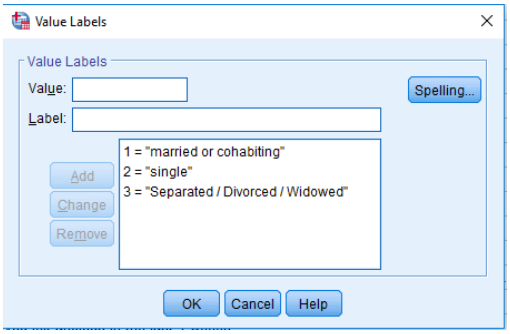
- With a scale variable the numbers represent meaningful metrics (i.e. the numbers naturally mean something). For example if someone is asked their age in years their response will be a number, and this will be intrinsically meaningful. This would also apply to annual income in pounds, weight in kg, or the respondent's score on a maths test.

- With a categorical variable individual's responses are placed into discrete categories. For example when asked whether they live in a rural or urban area the responses are recorded as '1' if they live in an urban area, and '2' if they live in a rural area. Categorical variables can be ordinal (i.e. they can have an order) or they can be nominal (i.e. they do not have an order). For example when asked 'how often have you felt anxious in the last 4 weeks?', the responses have an intrinsic order 'Very Often', 'Often', 'Not Often' or 'Not at All'. In comparison the 'marital' variable which indicates marital status is categorical but they categories have no intrinsic order ('married or cohabiting', 'single', 'Separated/Divorced/Widowed'), therefore this variable is nominal.

If a variable is categorical SPSS will need to know what each of the numbers actually mean. This is specified in sixth column labelled 'Values'. If you click on the cell for the marital variable under the sixth column a little blue square will appear, click on this (see red arrow below).

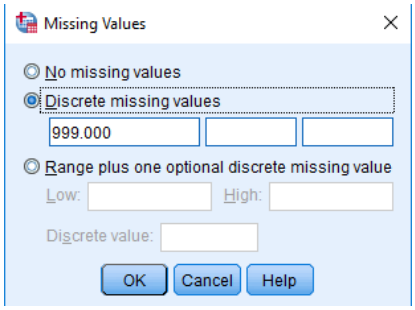


A dialog box will appear which will tell you what each of the numbers mean for this categorical variable. You can see that when the value '1' is shown for this variable that means that the respondent is 'married or cohabiting', the value '2' means that the respondent is 'single', and the value '3' means that the respondent is 'Separated /Divorced / Widowed'. Click **CANCEL** to close this this window.



The final column to look at in Variable View is the 'Missing' column. Missing data is a feature of nearly all quantitative data resources. Missing data occurs when people do not complete your survey or they do not complete some questions. If you don't have any information on a variable for a respondent you enter a value which means that this information is missing. Usually you enter an implausible value (i.e. a value that is very high or very low) to indicate the value is missing.

In this data set the value '999' was entered when information was missing. You do not want SPSS to use these missing values when you analyse your data so you have to tell SPSS that '999' means that this information is missing. Click next to one of the '999' values in the 'Missing' column and a little blue box will appear, click on this box and you will see the dialog box below. Here you can see that the number '999' has been entered to tell SPSS that this is a missing value. Click on **CANCEL** to close this window.

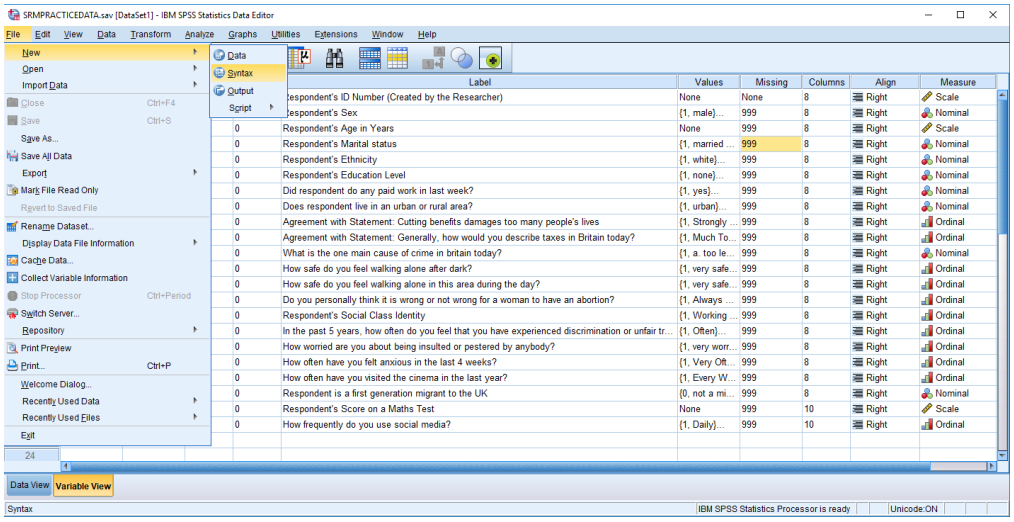


Take a few minutes to look at the data. Look at the Variable View tab and the Data View tab. Take a look at all the variables that are included in this data set. Often the best way to check whether you understand something is to try to explain it to someone else. Talk to person next to you and explain what information is provided in Variable View and Data View. If you are in any doubt please ask one of the teaching staff.

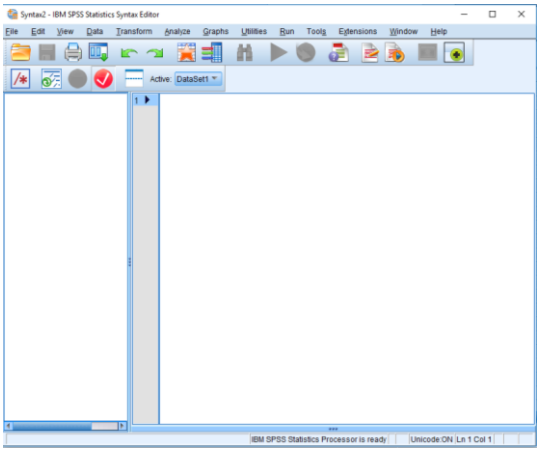
Section 2: Univariate Descriptive Statistics

Now you have explored the data we will undertake some basic data analysis tasks. You give SPSS instructions to undertake certain tasks using Syntax, this is simple computer code. You write your instructions to SPSS in the Syntax Editor window.

Open a Syntax Editor window by clicking on **FILE**, **NEW**, **SYNTAX** (see below).



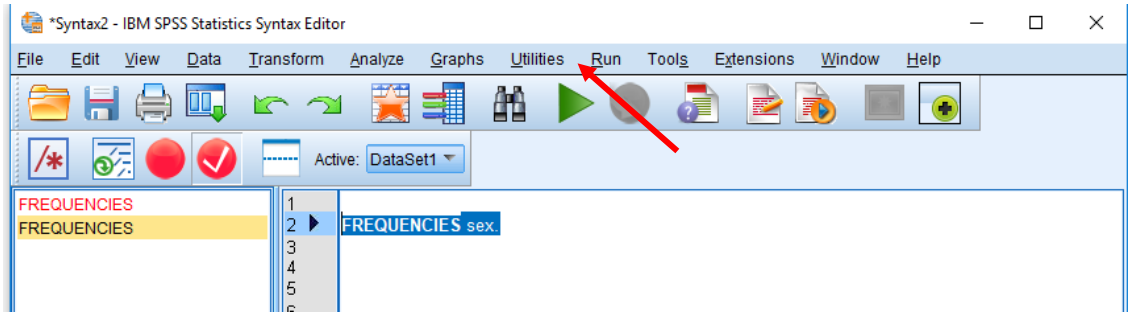
The Syntax Editor will look like the image below. You will write SPSS commands by typing into this window.



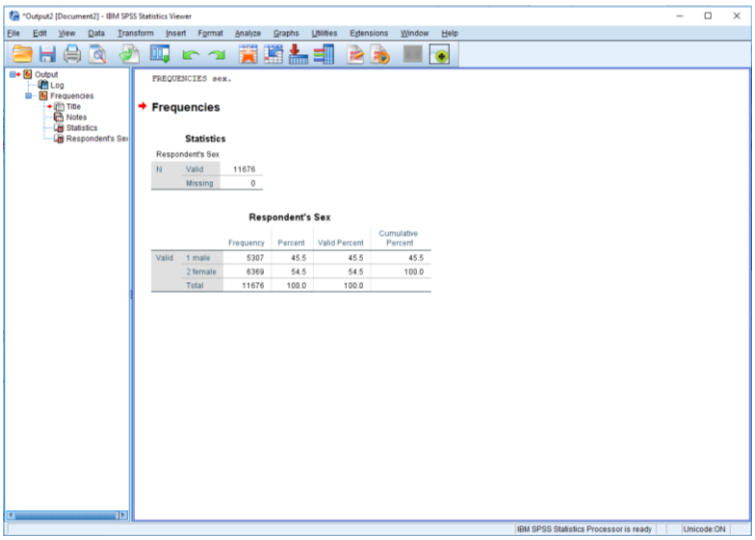
The first thing we will ask SPSS to do is to tell us how many men and women are in our sample. You can do this by typing the code below into your Syntax Editor. Your code has to look exactly like this with a full stop at the end. If you make any typos SPSS will not understand what you are asking it to do. Please note that 'sex' is the name of one of your variables, so you are asking SPSS to report the 'FREQUENCIES' for variable 'sex'.

```
FREQUENCIES sex.
```

After you have typed your code you need to 'run' it (i.e. tell SPSS to do this task). To do this you should highlight the code and then click on the green triangle at the top of the Syntax Editor Window (see red arrow).



After you have done this a third SPSS window will open on your computer, this new window is called the 'Output Viewer'. This is where you view the 'Output' or the results of your analyses.



You wrote some code which asked SPSS to tell you how many men and women are in your sample. You will see that SPSS has printed a table in your Output Viewer which tells you this. You can see that there are 5,307 men and 6,369 women. You can also see that men make up 45.5% of your sample and women make up 54.5% of your sample.

You now know how to undertake quantitative data analysis in SPSS!

You might be wondering, how will I know what code to write? You will be shown all the code you need to write and you only need to make small changes to the code to repeat the analysis with different variables.

The code we wrote to examine the 'sex' variable is shown below.

```
FREQUENCIES sex.
```

Now we want to examine the marital variable. All we need to do is change the variable name 'sex' to the variable name 'marital' in the code (see below). Write this in your Syntax Editor window.

```
FREQUENCIES marital.
```

Run this code by highlighting it and clicking on the green triangle. Now take a look in your Output Viewer window and you will see some tables describing marital status (see below).

Statistics

Respondent's Marital status

N	Valid	11668
	Missing	8

Respondent's Marital status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 married or cohabiting	5559	47.6	47.6	47.6
	2 single	3406	29.2	29.2	76.8
	3 Separated / Divorced / Widowed	2703	23.2	23.2	100.0
	Total	11668	99.9	100.0	
Missing	999	8	.1		
Total		11676	100.0		

The first table tells you how many people responded to the marital status question in the survey. The number next to 'valid' is the number of people who gave a response to the question (i.e. they completed this question). The number next to 'Missing' is the number of people who did not complete this question. These are the people who would have been given a value of 999 in the data set. We can see that 8 people did not respond to this question, and therefore 8 people had missing data.

In the next table we can see the number of people in each category of 'marital status'. Under the 'Frequency' column we can see that 5,559 people in the sample described themselves as 'married or cohabiting', 3,406 people in the sample described themselves as 'single', 2,703 described themselves as 'Separated/Divorced/Widowed'. We can also see again that 8 people had a missing response.

Next there are three columns which tell you about the percent of people in each category: 'Percent', 'Valid Percent' and 'Cumulative Percent'. What are the differences between these different 'percents'?

- 'Percent' is calculated including those people who did not give a response. This allows you to see that 0.1% of people have a missing response. This is useful when describing how much of your data is missing (i.e. how many people did not complete this question). However, this is not so helpful when you want to describe how many people are in each category, as it would make more sense to calculate this without including the missing responses.

- 'Valid Percent' is calculated without including the missing responses. This is the most useful percent figure to look at when trying to understand what percentage of people are in each category.
- The third percent is 'Cumulative Percent'. Cumulative percent is not always the most useful. This is the total percent in a category added to the percent in the category before. So the percent for single people includes the percent for single people (29.2%) added to the percent for married and cohabiting people (47.6%). You generally don't need to look at this percent. As you will learn SPSS shows you far more information than you actually need, and additional information such as this can be ignored for most purposes.

Now we will look at one of the scale variables in the data set (i.e. those which do not have categories). Simple initial analyses of scale variables include calculating the mean, median, mode and range of these variables. You have probably come across these concepts before and a quick reminder is provided below.

- The mean is the average number. To find the mean add all the values together and the divide by the number of numbers (e.g. $6 + 3 + 100 + 3 + 13 = 125 \div 5 = 25$).
- The median is the middle value. To find the median you put your values in order and find the middle value (e.g. 3, 3, 6, 13, 100 = 6).
- The mode is the number which appears the most. To find the mode you put your values in order and see which number appears most often (e.g. 3, 3, 6, 13, 100 = 3).
- The range is the distance between the highest and lowest numbers. To find the range you subtract the lowest number from the highest number (e.g. 100 – 3 = 97).

Here is the code you need to write in your Syntax Editor window to find the mean, median, mode and range for the age variable in the data set. Type this code into your Syntax Editor and then run it (i.e. highlight it and click on the green triangle).

```
FREQUENCIES age
  /FORMAT NOTABLE
  /STATISTICS=MEAN MEDIAN MODE RANGE.
```

Take a look in your Output Viewer window and you will see that SPSS has printed information on the mean, median, mode and range of the respondents' ages (see below).

Statistics

Respondent's Age in Years

N	Valid	11661
	Missing	15
Mean		50.42
Median		49.00
Mode		42
Range		85

You can see that 11,661 respondents answered the age question, and 15 respondents did not (i.e. they have missing data). The mean age was 50.42 years, the median age was 49 years, and the mode age was 42. The age range was 85 years.

We could ask SPSS to show us the highest and lowest ages by adding to our code a little. If you type the code below and run it (i.e. highlight it and click on the green triangle), SPSS will print a new table in your Output Viewer also showing the minimum and maximum age.

```
FREQUENCIES age
  /FORMAT NOTABLE
  /STATISTICS=MEAN MEDIAN MODE RANGE MIN MAX.
```

You can see that the youngest person in the sample is aged 16 years (i.e. minimum age) and the oldest person in the sample is aged 101 year (i.e. maximum age).

Statistics

Respondent's Age in Years

N	Valid	11661
	Missing	15
Mean		50.42
Median		49.00
Mode		42
Range		85
Minimum		16
Maximum		101

We will repeat this process to examine another scale variable, the respondent's score on a maths test. We can use the same code that we used above but you will need to change the name of the variable you are analysing from 'age' to 'mathscore'. Type the code below to find the mean, median, mode, range, min and max maths score. Run this code (i.e. by highlighting it and clicking on the green triangle).

```
FREQUENCIES mathscore
  /FORMAT NOTABLE
  /STATISTICS=MEAN MEDIAN MODE RANGE MIN MAX.
```

Take a look in your 'Output Viewer' window and you will see the table SPSS has printed in response to this command.

Statistics

Respondent's Score on a Maths Test

N	Valid	11661
	Missing	15
Mean		28.83
Median		24.00
Mode		18
Range		97
Minimum		3
Maximum		100

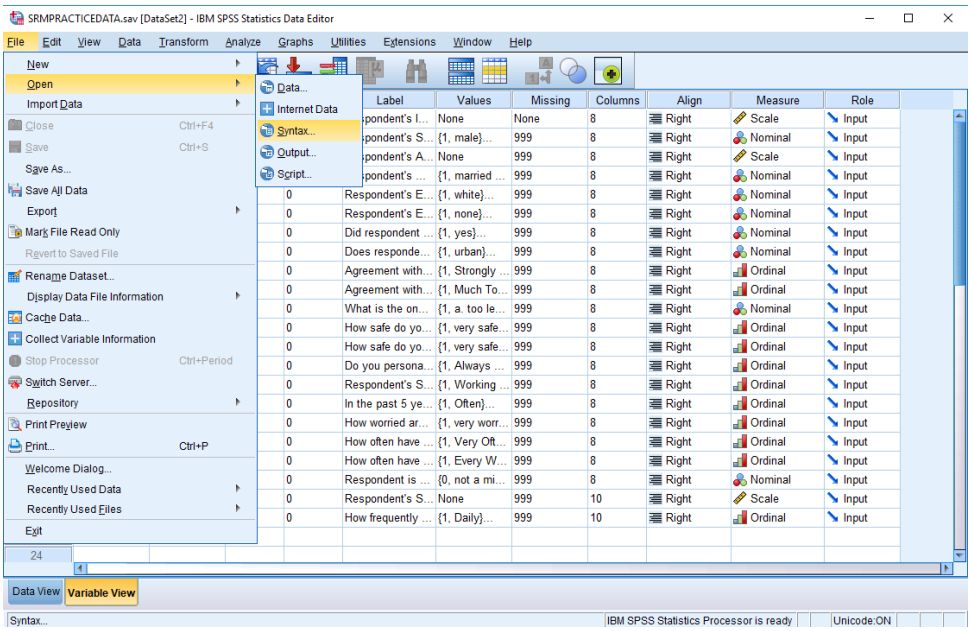
Your output shows that 11,661 respondents have information on their score on the maths test, and 15 respondents have missing data. The mean maths test score was 28.83, the median score

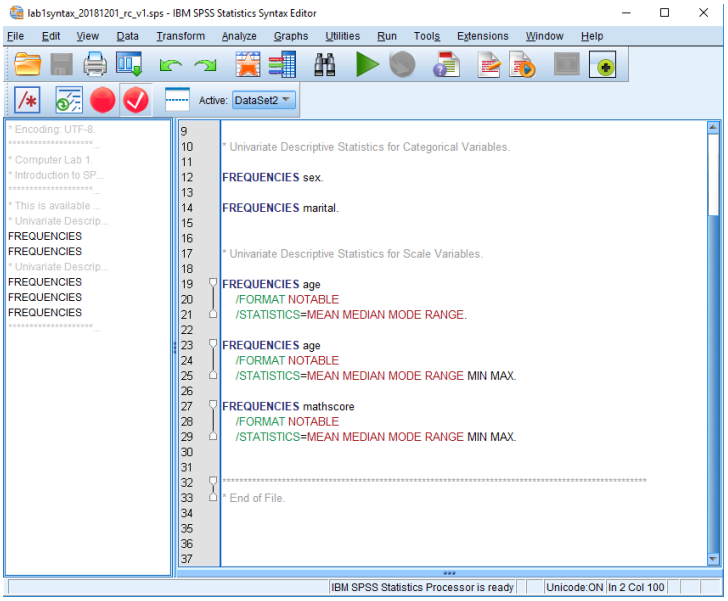
was 24.00, the mode score was 18. The range was 97. The minimum score was 3 and the maximum score was 100.

You now know how you can examine and describe categorical and scale variables, the analyses you have done so far can be described as 'univariate descriptive statistics', these are statistics which describe single variables. This is the first step you take when analysing quantitative data. Examining univariate descriptive statistics allows the researcher to get to know their dataset and to understand how people have responded to survey questions.

You may be wondering how you will remember the different pieces of code you need to write to analyse the data. Learning to write code is much like learning a new language, practice makes perfect. By the end of this module you will find that you remember little bits of code. However, we do not expect you to memorise the code you are introduced to. All the code introduced each week is shown in a syntax file available on the VLE. When you need to undertake a data analysis task you can look at these files, and copy and paste the required code.

On the week 4 VLE page (where you got this workbook) you will also see a link to the 'Lab 1 Syntax'. Download this file onto your computer. Once you have saved this syntax file, you should click on **FILE**, **OPEN**, **SYNTAX** and open the syntax file from where you saved it.

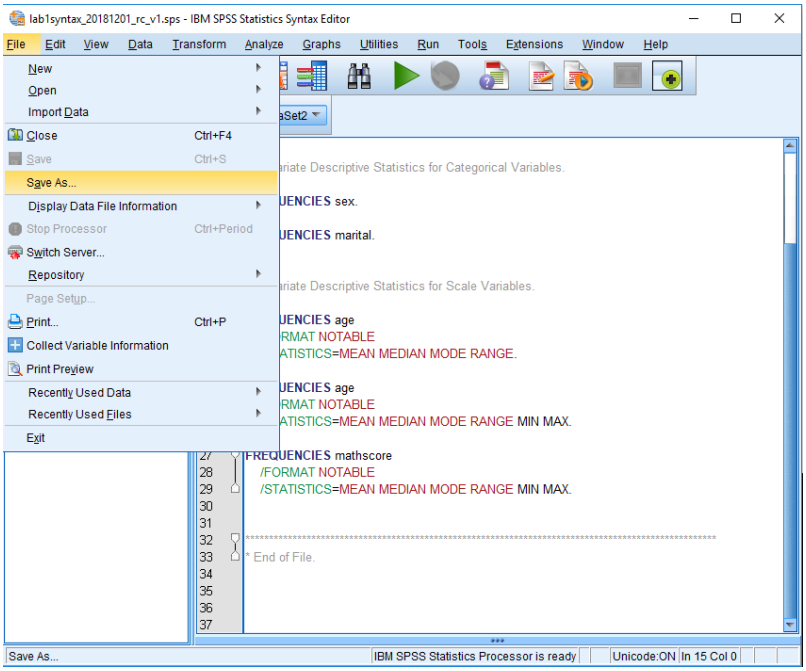




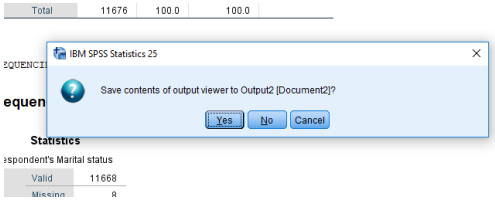
In the Syntax Editor window you have just opened (see above), you will see all the code you have run so far. You will also see that there is also some additional text that appears in grey. These are comments, they do not tell SPSS to do anything. When writing statistical code it is important to write lots of comments to remind yourself, and to inform others, what you have done. You may find it useful to write comments in your syntax files to remind yourself what each piece of code does. To write a comment start with an asterisk (*) and end your comment with a full stop. If you do this SPSS will 'grey out' the text and it will know that this is not an instruction to do something.

- Write a new comment in the syntax file, you will see that as long as it starts with a * and ends with a . it will become grey.

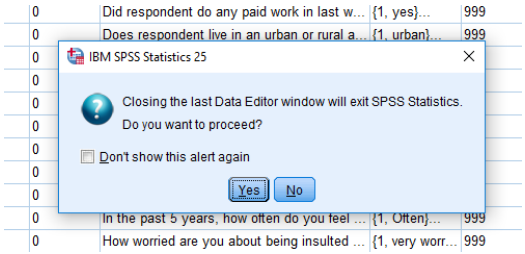
Practice the skills you have developed so far to explore the variables in the data set. Most of the variables in the data set are categorical so you should use the 'Frequencies' code. Repeat the code for different variables and note the number of valid and missing responses, and the percentage of people who fall in each response category. Don't delete code once you have used it, use your syntax file to keep a record of everything you have done. Save your syntax file regularly, just as you would with a word file. Click on **FILE** and then **SAVE** or **SAVE AS** in the top left hand corner of your syntax editor window.



Now we will close SPSS. Make sure you save your syntax file (as shown above), then you can close it by clicking on the cross on the top right hand side of the window. Next close your Output Viewer window by clicking on the cross on the top right hand side of the window below. There is no need to save your output unless it is important (i.e. you might save this file when you start working on your assessment). These are very large files that will take up a lot of space, also you can very easily and quickly recreate all of the tables you have made by re-running your code. When you click on the cross to close your Output Viewer window the warning below will pop up, you should click on 'No'.



Now close the Data Viewer window. You should not save this file. Click on the cross on the top right hand side of the Data Viewer window. The warning below will pop up, you should click 'Yes'. SPSS will now close.



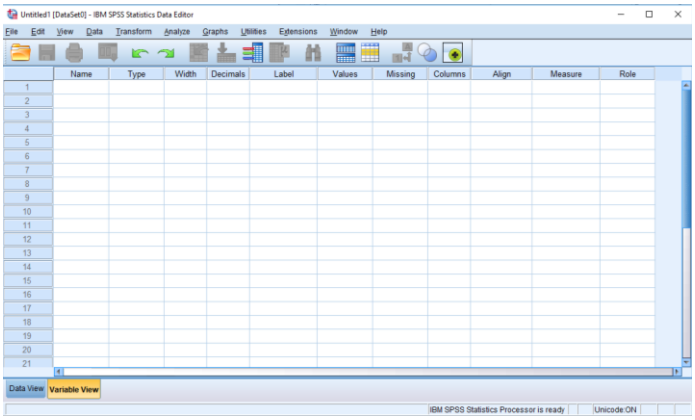
Section 3: Entering Data

- Reopen SPSS, if you cannot remember how to do this return to the start of the workbook.

Now you are going to enter some data into SPSS manually. Modern researchers very rarely have to enter quantitative data into computer systems themselves. Surveys are usually collected via computer assisted technology, even if interviewers collect data via face to face interviews they will usually do so with the aid of a tablet or computer which will create a data set automatically. Paper questionnaires are relatively rare, except for in certain circumstances (e.g. when asking very sensitive questions which would be strongly influenced by face to face interaction).

Social science researchers also have a vast array of existing data resources available to them from large government social surveys, administrative data (e.g. data collected via the running or government and organisations) and data collected via other means (e.g. data derived from online interactions).

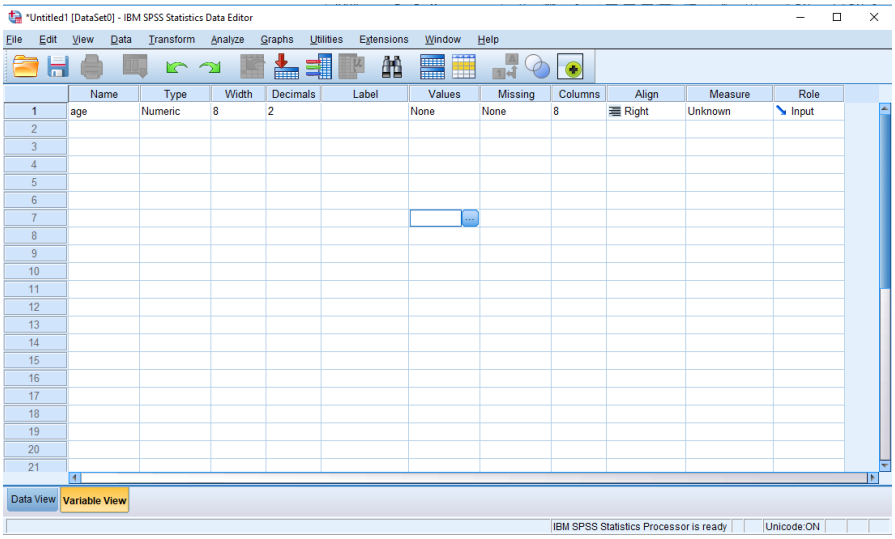
Nevertheless, it is important that you are aware of how to enter data into SPSS. **Navigate to the 'Variable View' of your Data Editor tab.** You should see a blank spreadsheet (see below).



Remember that each row in the spread sheet represents a single variable (e.g. age, hair colour, income, test score).

The first column is labelled 'Name'. This is where you enter the name of your variable. For example, if we had information on survey respondents' age we could call this variable 'age'. You can call your variables anything you like, with a few restrictions. In SPSS variable names must begin with a letter and be less than 64 characters long. Some punctuation symbols are not accepted, so these are best avoided.

In the first column enter the variable name 'age', then click on an empty cell below. SPSS will automatically fill in some additional information in the first row (see below).



The next column is labelled 'Type'. Type refers to the way in which the data is stored. This should be set automatically as 'Numeric' and you should not change this, we will be entering only numeric data (i.e. numbers).

The next column is called 'Width'. 'Width' is an option which you can just ignore for now.

The next column is called 'Decimals'. This defines how many decimal places will be shown in the data set. By default this is set to 2, so the data will be shown to two decimal places for this variable (i.e. 20.00). We are going to enter age as a whole number so there is no need to have two decimal places, you can therefore change 'Decimals' to 0.

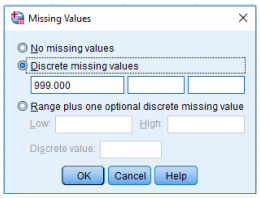
The next column is labelled 'Label'. This allows you to provide a slightly more detailed description of what the variable shows. It is very important to include variable labels to provide a little extra information about the data. For our age variable enter the text 'Age in years at last birthday'.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	age	Numeric	8	0	Age in years at...	None	None	8	Right	Unknown	Input

The 'Values' attribute allows you to create a list of value labels. Categorical variables will share a common set of value labels. For example you might ask which month someone was born in and the value labels would be 1 = January, 2 = February, 3 = March etc. We do not require value labels for our age variable as it is not a categorical variable, we can therefore leave this blank. We will enter a categorical variable later.

The 'Missing' column allows you to define values which are attributed to missing values. Missing data is feature of almost all quantitative data resources. For example when you administer a survey some people will refuse to participate or forget to return their survey and therefore all of their information will be missing. Some respondents will fail to complete certain questions, either by accident or by refusing to provide certain information. You should enter a value in your spreadsheet which indicates that this information is missing. For example, if someone failed to provide their age you could enter the value 999 in your spreadsheet for this respondent. It is impossible to have someone actually aged 999, this value indicates that their age is missing. But you have to tell SPSS which values actually mean missing, so SPSS does not include these cases when you undertake any analyses.

If you click on the little box in the 'Missing' cell you will see the dialog box below. You should click on 'Discrete missing values' and enter your missing value '999'. You can then click OK.



SPSS gives you the option to include multiple missing values, you might do this is you have different types of 'missingness'. For example you could give the value 999 to those respondents who refused to give their age, and 997 to those whose value is missing for another reason. This is sometimes useful information to record, but we will only define one value in this example.

The 'Columns' attribute refers to the width of each column in the spreadsheet. This is an option which you can just ignore for now.

The 'Align' attribute refers to the alignment of the cells in the spreadsheet. This is an option which you can just ignore for now.

The 'Measure' attribute refers to the level of measurement of the variable. If you click in this cell you can select 'Scale', 'Ordinal' or 'Nominal'. Age is a scale variable so you should select scale under the 'Measure' column.

The 'Role' attribute refers to how the variable will be used in analysis. This is an option which you can just ignore for now.

Your row for age should now look like this:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	age	Numeric	8	0	Age in years at...	None	999	8	Right	Scale	Input

Now you have told SPSS all the details of this variable you can enter the data. **Navigate to the 'Data View' tab of the Data Editor.** You should see that the first column now has the title 'age', the variable you have defined in the 'Variable View' tab.

Enter the information on the ages of these 10 respondents. Do this by typing their ages into the first column, make sure you enter the ages in the order shown here.

Respondent	age
1	21
2	23
3	25
4	21
5	22
6	26
7	33
8	20
9	42
10	22

Your Data Editor window should now look like this:

	age	var	var
1	21		
2	23		
3	25		
4	21		
5	22		
6	26		
7	33		
8	20		
9	42		
10	22		

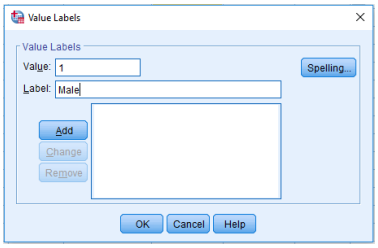
In the table below you will see information on height in cm for the same 10 respondents. Height is a scale variable that has been recorded to 0 decimal places. Return to variable view and enter the details for the variable 'height'. You should include a variable label (e.g. 'Height in cm'), and enter 999 as your missingness indicator.

Then enter the data in data view.

Respondent	Age	height
1	21	177
2	23	163
3	25	179
4	21	178
5	22	160
6	26	162
7	33	163
8	20	177
9	42	182
10	22	155

Now you are going to enter the data for a categorical variable, 'sex'. In our data this takes on two values 1 = Male and 2 = Female. You should enter the details for this variable in Variable View tab. You should select 'Nominal' under the 'Measure' column.

You will need to define some values under the 'Values' Column. Click on the little box in the values cell for this variable and the dialog box below will come up.

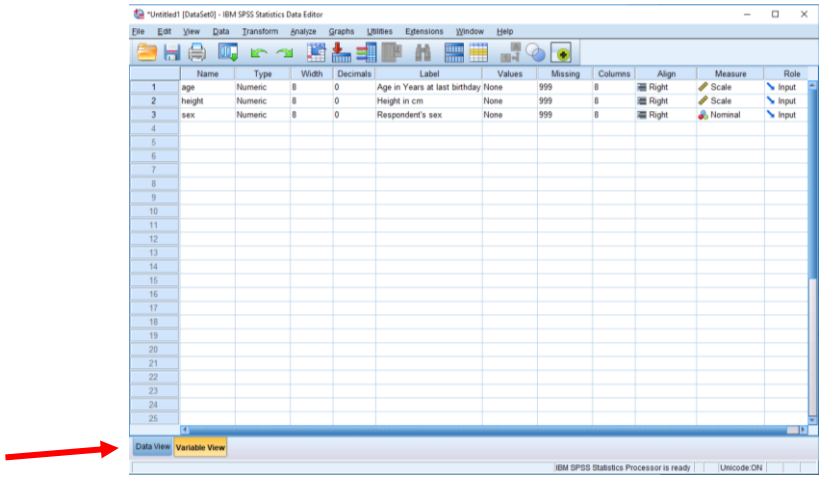


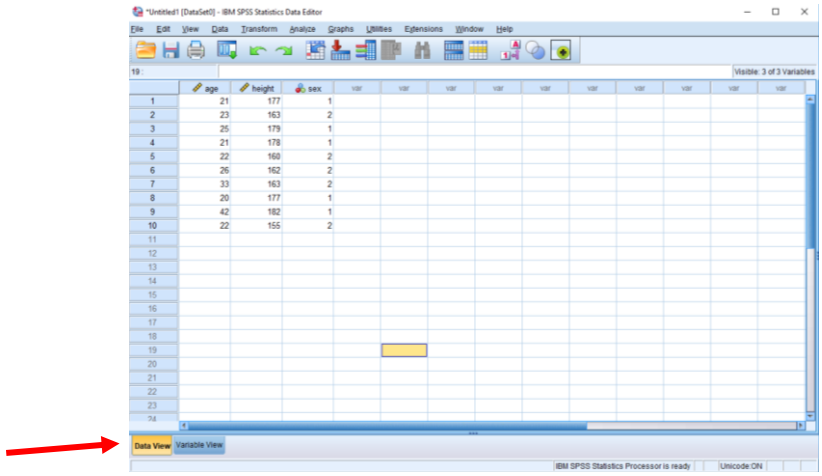
Enter the 1 in the 'Value' box and 'Male' in the 'Label' box then click '**Add**'. Now repeat this with 2 and Female then click '**Add**'. Then click **OK**. The values which this variable can take should now show in variable view.

Once you have entered all the information associated with the 'sex' variable navigate back to data view, and enter the values for sex.

Respondent	Age	height	sex
1	21	177	1
2	23	163	2
3	25	179	1
4	21	178	1
5	22	160	2
6	26	162	2
7	33	163	2
8	20	177	1
9	42	182	1
10	22	155	2

You should now have a data set entered in SPSS with three variables for 10 respondents. The view from your 'Variable View' tab and 'Data View' tab is shown below. You now know how to enter data in SPSS.





Now you have entered these data, you can produce some univariate descriptive statistics. Open a blank syntax file by clicking on **FILE**, **NEW**, **SYNTAX**.

Write the code to produce tables which tell you the answers to the following questions (if you wish to copy and paste the code you used above re-open the Lab 1 Syntax File by clicking on **FILE**, **OPEN**, **SYNTAX**):

- What percentage of the sample are male and female?
- What is the mean height of the sample?
- What is the mode age in the sample?

You have now completed the first SPSS computer lab, well done!

The computer lab sessions are cumulative (i.e. they build on each other). When you come to next week's lab we will assume that you are confident with the tasks covered here. You should make some time to practice these tasks before next week's lab. Additional supporting resources are available on the VLE. Please do not hesitate to ask if you require any additional support or assistance.