## Main Instructions

1. Read description of bfi dataset in "bfi-data-description.rtf"
2. Import the comma delimited text file "bfi.csv" into SPSS
3. Add the variable labels from "bfi-personality-meta-data.xlsx" to the SPSS personality variables
4. Add value labels for gender and education
5. Are all the values within the range of permissible values for all variables (e.g., all personality test items are on a 1 to 6 scale)? (Tip: Analyze - Descriptives - Frequencies)
6. To what extent is missing data a problem in the dataset? How much missing data is there? Which variables have more missing data? Why?
7. Create a variable that stores the number of missing responses a particular person has (Transform - compute; use the "nmiss" function with comma separated variables or just "nmiss(A1 to age)" Are there any cases with a huge amount of missing data? What threshold would you use for deleting the case? Why?
8. Compute scale scores for each big 5 factor (Tip: see this example: <http://jeromyanglim.blogspot.com.au/2009/10/scale-construction-item-reversal-scale.html> )
9. Compute cronbach's alpha reliability for each of the big 5 factors.

## Additional Exercises

1. Compare mean scores on the five factors of personality for the different genders. Are the results as you would expect?
2. Correlate the five factors of personality. Are they correlated? Is the pattern and general size of correlations what you would expect?

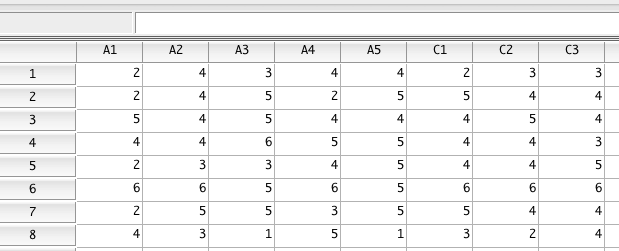
# Answers

## Import the comma delimited text file "bfi.csv" into SPSS

SPSS provides the Read Text Wizard for importing text data. In general, when you have your own data, you may have to try the options a few time until you get it right.

* Use the Read Text Wizard in SPSS to import the data
  + Go to File - Read Text Data or drag and drop the csv file into SPSS
  + How are your variables arranged? "delimited": The data is made up of one row per participant and a comma separates each variable. Thus, the data is delimited by a coma.
  + Are your variable names included at the top of the file? "yes"
  + Which delimiters appear between variables? just select "comma"

Here's a screen shot of the data after importing.



## Add the variable labels from "bfi-personality-meta-data.xlsx" to the SPSS personality variables

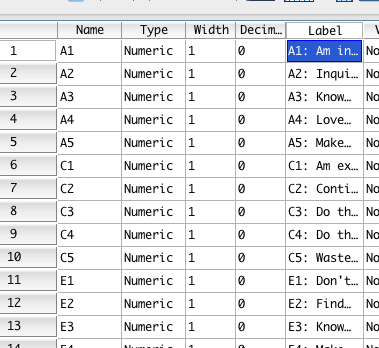
SPSS provides data view and variable view. Data view shows the raw data where rows are cases and columns are variables. Variable view shows the metadata where one row is a variable and columns correspond to aspects of metadata.

There are several ways to toggle between the two views, either using buttons at the bottom of the data window, or the menu, or a keyboard shortcut.

After importing new data into SPSS your first job is typically to configure the metadata. Some important aspects include:

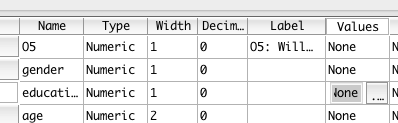
* Name: you can change the variable name
* Type: Two major types are "string" for textual data and "numeric" for numbers.
* Decimals: This influences how many decimals are displayed in the variable view and in the output tables.
* Label: This is the variable's label. This label is used in SPSS tables. Labels permit a more descripting and longer title.
* Values: Typically known as value labels, these are used to assign text to numbers. For example, it is common to enter categorical data as numbers and assign value labels: E.g., 1 = male, 2 = female.
* Missing: You can represent missing data simply by leaving a cell blank. Alternatively, some people use special missing value codes such as -99. By putting these values in the missing box, you tell SPSS, that such values are not actual values, they represent missing data.

Here's a snapshot of variable view after pasting in the labels. Note that it often requires some trial an error to work out how to paste in multiple rows of data all at once. On my Mac this involves selecting the top cell. In previous versions, you needed to highlight all cells that were going to be pasted over.

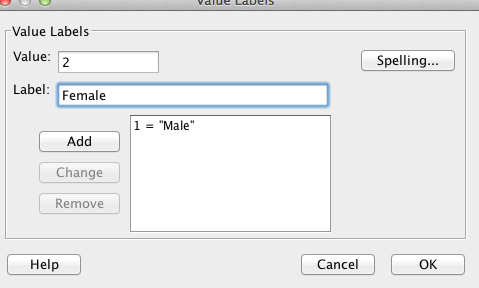
****

## Add value labels for gender and education

Click on the "values" column for gender and education respectively.



And enter in the value and label and click add for each value-label combination.



## Are all the values within the range of permissible values for all variables (e.g., all personality test items are on a 1 to 6 scale)? (Tip: Analyze - Descriptives - Frequencies)

Run analyze - descriptives - frequencies and select all variables.

Answer:

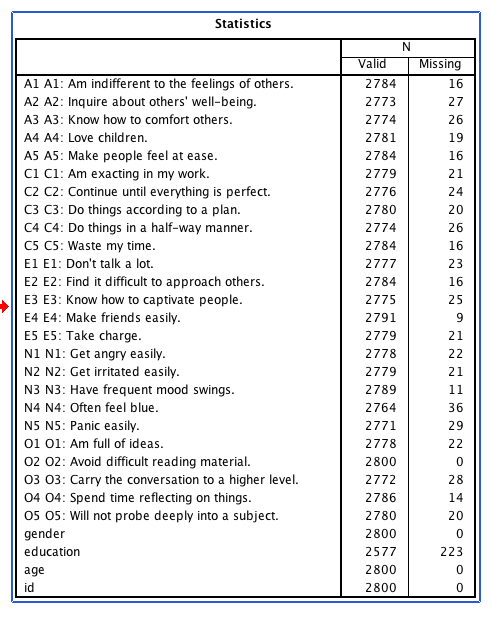
* All personality variables range from 1 to 6, which seems right.
* Gender and education only have appropriate categories
* Age ranges from 3 to 86. These are valid ages, although we might wonder about the validity of the "3" value.

## To what extent is missing data a problem in the dataset? How much missing data is there? Which variables have more missing data? Why?

There are many ways of getting missing data information. One option is to use the statistics table produced by analyze - descriptives - frequencies.

To make the table easier to use I double clicked on the table and selected "transpose rows and columns from the menu". I also adjusted the width of the label column. As a general aside, learning how to tweak tables to so that you can see what you need is a useful skill. In SPSS, pivot trays are a useful tool.

The end result looked like this.



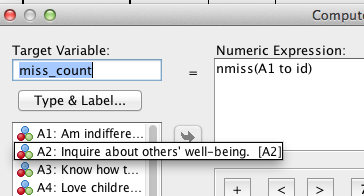
Answer:

* Most personality items has a small amount of missing data. I don't quite see the pattern behind this missing data. In fact, I almost think that something is a bit strange that there are no missing responses for item O2.
* There is quite a lot of missing data for education. Presumably this was not a required response.

Thus, for the most part, given the scale of the dataset, missing data is a nuisance in this data, but it is not a major issue.

## Create a variable that stores the number of missing responses a particular person has (Transform - compute; use the "nmiss" function with comma separated variables or just "nmiss(A1 to age)" Are there any cases with a huge amount of missing data? What threshold would you use for deleting the case? Why?

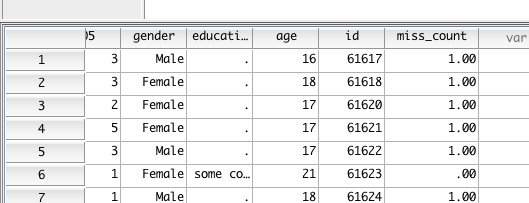
Run Transform - compute, then enter the following:



To learn more browse the functions ordered by "function group" and find the one labelled "missing values". You'll then find "nmiss" and this will give some basic information about what the function does.

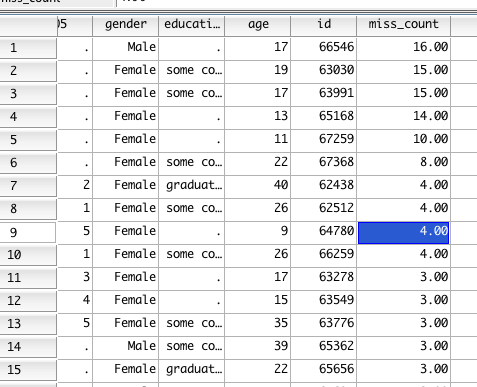
In general you would write nmiss(A1, A2, A3, and so on...) but a shortcut when you want to specify a range is to use the "to" keyword which automatically selects the variables from the first to the last.

I also strongly encourage you to paste anything you do with transform - compute into the syntax window for storage.

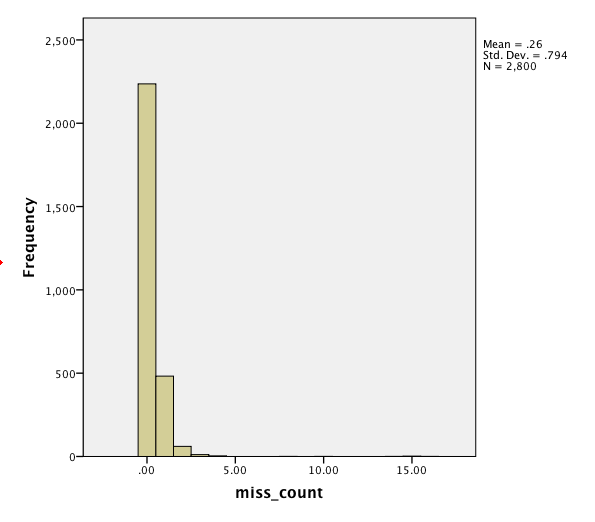


There are then several options for seeing cases with many missing values. One simple option is to sort the newly created variable miss\_count descendingly (just right click on the variable name in variable view).

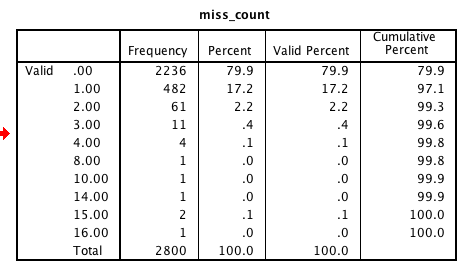
This looks like this:



Or you could look at it graphically using something like graphs - legacy -histogram (miss\_count), but in this case, the prevalence is so low you can barely see it:



Alternatively, a frequency table (analyze - descriptives - frequencies) is pretty helpful:



Given that there are 25 personality items and 3 demographics, 10 or more missing items suggests something went seriously wrong with data collection for those participants. In the case of online surveys, I would be inclined not to trust such participants' remaining data. It is makes other analyses more convenient if there is minimal missing data.

Thus, I'd probably delete the cases with 8 or missing data entirely.

## Compute scale scores for each big 5 factor (Tip: see this example: <http://jeromyanglim.blogspot.com.au/2009/10/scale-construction-item-reversal-scale.html> )

Computing scale scores requires that you reverse reversed-items and then create a mean or sum of items in the scale. It is particularly easy to make a mistake with this, so it is important to save your syntax so that you can double check what you've done. You should also double check manually a few scores or adopt some other procedure to check particularly when you are first learning the commands. Also, given how many items you are going to need to reverse, syntax very quickly becomes simpler and more reliable than menus.

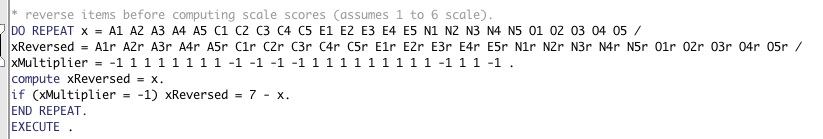
The basic syntax of item reversal of for example item "A1":

compute A1r = 7 - A1.

execute.

I.e., "compute" is an SPSS keyword; "A1r" is the name of a new variable where adding "r" indicates that the item is reversed; "7" is chosen because max + min of the current scale is 6 + 1; and "A1" is the original variable name. The full stop at the end indicates the end of the command. The "Execute." is required to actually run the command. You can have multiple compute statements and then a single execute statement at the end.

A more sophisticated approach is to use a loop to do the reversal.



The syntax is a little bit complicated at first, but it can make life easier if you need to compute many different psychological scales. If you're doing it yourself, you'll also probably get a few errors at first while you learn the syntax. But that's all part of learning to program.

Details of how the code works is described here: <http://jeromyanglim.blogspot.com.au/2009/10/scale-construction-item-reversal-scale.html>

In summary, you have three variables, the original (x) , the reversed (xReversed), and whether the item needs to be reversed (xMultiplier). For each one of these, a new variable is created (i.e., A1r, A2r, etc.) and if the variable needs to be reversed as indicated by whether it has a -1 or 1 associated with it (-1 for reverse, 1 for not reversed) it is reversed.

Note that all the information about whether an item is reversed is stored in a the excel item database. I strongly encourage you to copy and paste from the database rather than manually typing the data into SPSS in order to prevent errors.

Thus, you end up with a set of variables A1r, A2r, etc that are reversed if needed and are now suitable for creating scale scores and running reliability analysis.

**Scale scores:** Once you have reversed items, you can use several different approaches to create scale scores. A good approach is to use the "mean.x" functions where x is a number corresponding to the minimum number of items required to not return a missing value. Thus, we could choose "mean.3" because it is a five item scale and three seems like the minimum needed to get reasonable information. Alternatively, you could handle missing data before calculating scale scores.

Here's an example of what the syntax could look like.

\*compute scale scores.

COMPUTE agreeableness = mean.3(A1r, A2r, A3r, A4r, A5r).

COMPUTE conscientiousness = mean.3(C1r, C2r, C3r, C4r, C5r).

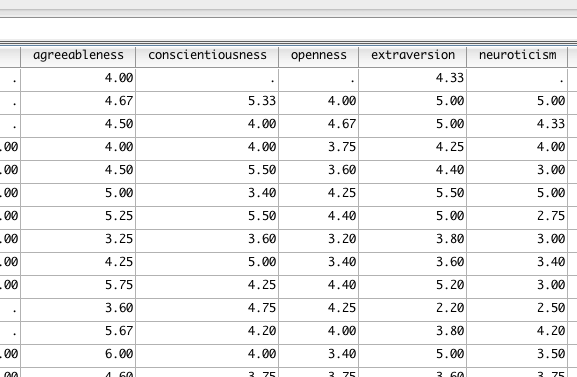
COMPUTE extraversion = mean.3(E1r, E2r, E3r, E4r, E5r).

COMPUTE neuroticism = mean.3(N1r, N2r, N3r, N4r, N5r).

COMPUTE openness = mean.3(O1r, O2r, O3r, O4r, O5r).

EXECUTE.

Resulting in new variables in the data file:



## Compute cronbach's alpha reliability for each of the big 5 factors.

* Run Analyze scale reliability for each scale
* I.e., for agreeableness put in A1r A2r A3r A4r A5r
* It is also useful to click "statistics - scale if item deleted"

This is also another example, where you'll quickly see the efficiency of clicking "paste" to get the syntax for agreeableness. Then you can just copy and paste a few times and modify the variable names to get reliability for the other big 5.

For example the code would look like:

RELIABILITY /VARIABLES=A1r A2r A3r A4r A5r /SCALE('ALL VARIABLES') ALL /MODEL=ALPHA /SUMMARY=TOTAL.

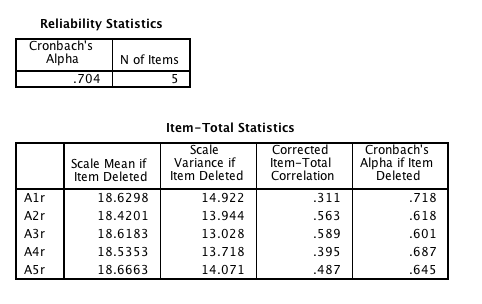
RELIABILITY /VARIABLES=c1r c2r c3r c4r c5r /SCALE('ALL VARIABLES') ALL /MODEL=ALPHA /SUMMARY=TOTAL.

RELIABILITY /VARIABLES=e1r e2r e3r e4r e5r /SCALE('ALL VARIABLES') ALL /MODEL=ALPHA /SUMMARY=TOTAL.

RELIABILITY /VARIABLES=n1r n2r n3r n4r n5r /SCALE('ALL VARIABLES') ALL /MODEL=ALPHA /SUMMARY=TOTAL.

RELIABILITY /VARIABLES=o1r o2r o3r o4r o5r /SCALE('ALL VARIABLES') ALL /MODEL=ALPHA /SUMMARY=TOTAL.

To take the agreeableness reliability output, we see reasonably reliable scale (cronbach's alpa of .704) and all items have reasonable corrected item total correlations. On other occasions, you may find it useful to add variable labels to the reversed items to see more clearly why some items have stronger or weaker correlations with the corrected total.



Other reliabilities were

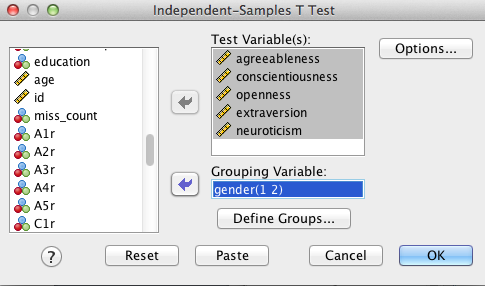
* Conscientiousness: .729
* Extraversion: .761
* Neuroticism: .813
* Openness: .603

If this was your thesis, this would be essential information to report.

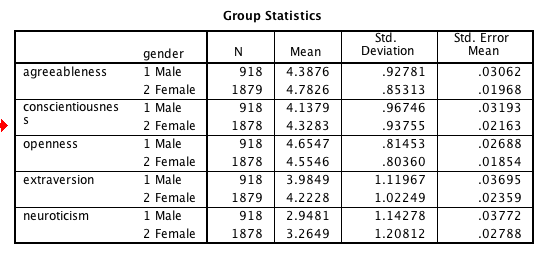
# Additional Exercises

## Compare mean scores on the five factors of personality for the different genders. Are the results as you would expect?

There are many ways to compare group means in SPSS. One option is analyze - compare means - independent samples t-test



All the effects are statistically significant, because of the huge sample.

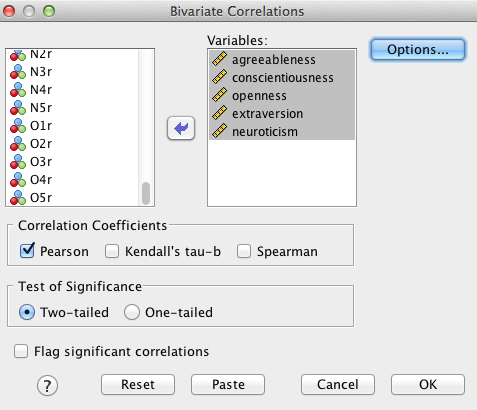


Thus, the main focus is on the means. Females appear to be about a third of a standard deviation higher on agreeableness, conscientiousness, extraversion, and neuroticism. This corresponds conceptually to a small to medium difference.

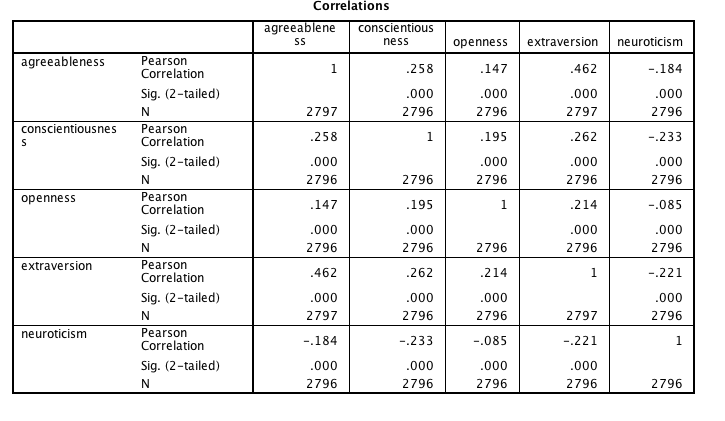
## Correlate the five factors of personality. Are they correlated? Is the pattern and general size of correlations what you would expect?

The main way to get correlations is analyze -correlate - bivariate.

Given that the sample size is huge, I've turned off "flag significant correlations".



The default output is fine

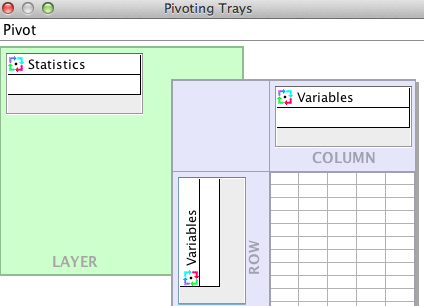


But I find it nicer just to see the correlations particularly as the number of variables gets large.

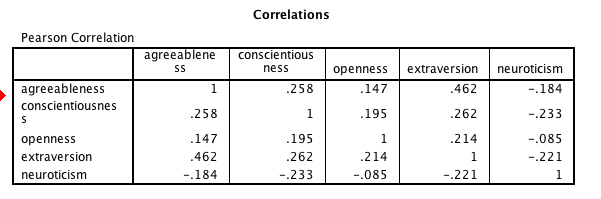
You can see more about formatting correlation matrices for psychological reports here:

<http://jeromyanglim.blogspot.com.au/2009/02/formatting-correlation-matrices-in.html>

But a general tip is to double click on the table and select pivot trays from the menu; you can then drag and drop statistics into the layer.



That gives you a much nicer correlation matrix to scan:

****

One theory on the Big 5 personality is that they uncorrelated factors. However, this matrix suggests that there are small to moderate correlations between the factors. In particular, extraversion has a fairly strong correlation with agreeableness. In general the "good" factors correlate positively with each other and neuroticism correlates negatively with the others. These results are broadly consistent with Big 1 and Big 2 higher level models of the big 5.