Micro-module D: Big Data Analytics

D2- Statistical Data Analysis

SPSS is a widely used program for statistical analysis in social science.[6] It is also used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations, data miners, and others. We usually use SPSS to input, manage, analysis a large amount of quantitative data in the urban studies field. This tutorial will show the basic concepts for beginners to manage their data in SPSS, including the introduction of SPSS interface, data input, and categories of variables, it will also include the descriptive statistics step-by-step guidelines.

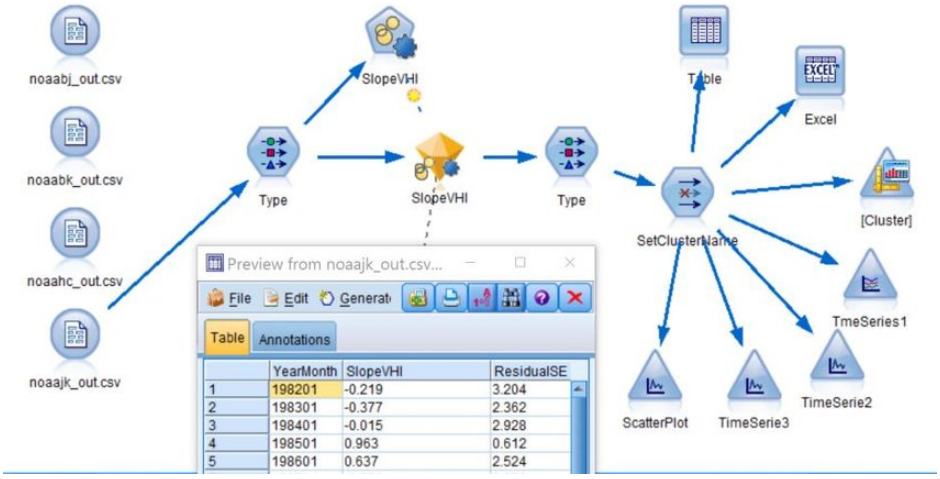
1. Introduction of SPSS

1.1 About SPSS

The program, originally called Statistical Package for the Social Sciences, was released in 1968 and quickly became one of the most widely used statistics programs in the social sciences, including in healthcare, government, market research and surveying.

SPSS is great for

1) Opening data files, either in SPSS' own file format or many others;
2) editing data such as computing sums and means over columns or rows of data. SPSS has outstanding options for more complex operations as well.
3) creating tables and charts containing frequency counts or summary statistics over (groups of) cases and variables.
4) running inferential statistics such as ANOVA, regression and factor analysis.
5) saving data and output in a wide variety of file formats.



Source:

https://www.researchgate.net/publication/322299718_Greenness_pattern_analysis_with_the_remote_sensing_index_clustering/

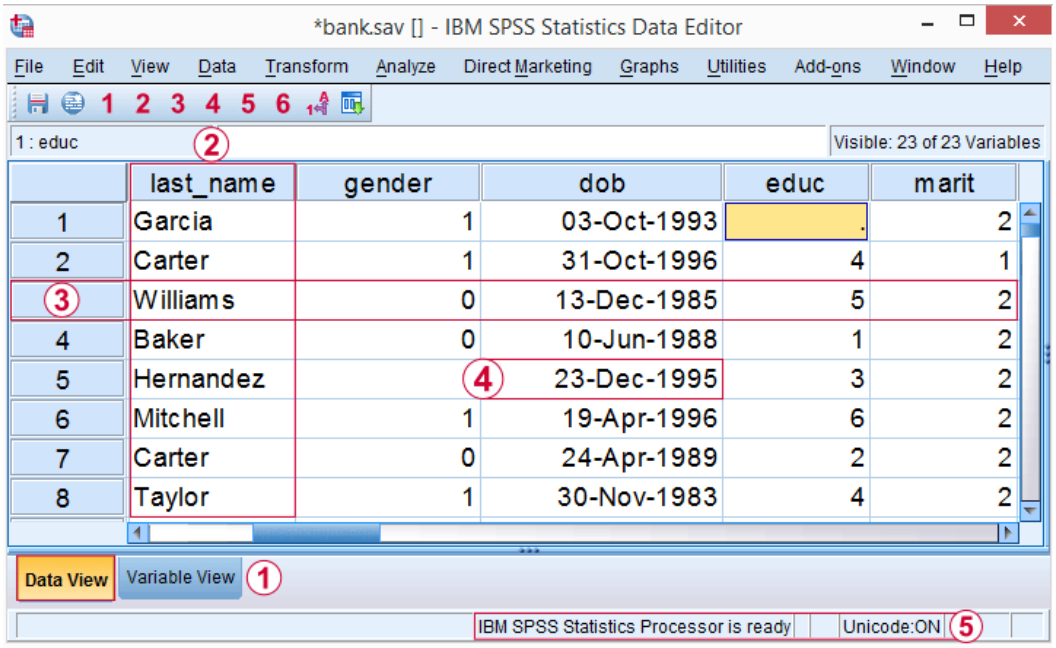
1.2 The SPSS Interface

It contains two main part

1) data values which we see in Data View and
2) dictionary information about our data in Variable View.

After opening data, SPSS displays them in a spreadsheet-like fashion as shown in the

screenshot below from freelancers.sav. This sheet -called data view- always displays our data values. For instance, our first record seems to contain a male respondent from 1979 and so on. A more detailed explanation on the exact meaning of our variables and data values is found in a second sheet shown below.



	last_name	gender	dob	educ	marit
1	Garcia	1	03-Oct-1993	.	2
2	Carter	1	31-Oct-1996	4	1
3	Williams	0	13-Dec-1985	5	2
4	Baker	0	10-Jun-1988	1	2
5	Hernandez	4	23-Dec-1995	3	2
6	Mitchell	1	19-Apr-1996	6	2
7	Carter	0	24-Apr-1989	2	2
8	Taylor	1	30-Nov-1983	4	2

1) The data editor has tabs for switching between Data View and Variable View. For now, make sure you're in Data View.

2) Columns of cells are called variables. Each variable has a unique name (“gender”) which is shown in the column header.

3) Rows of cells are called cases. Oftentimes, each respondent in a study is represented as a single case.

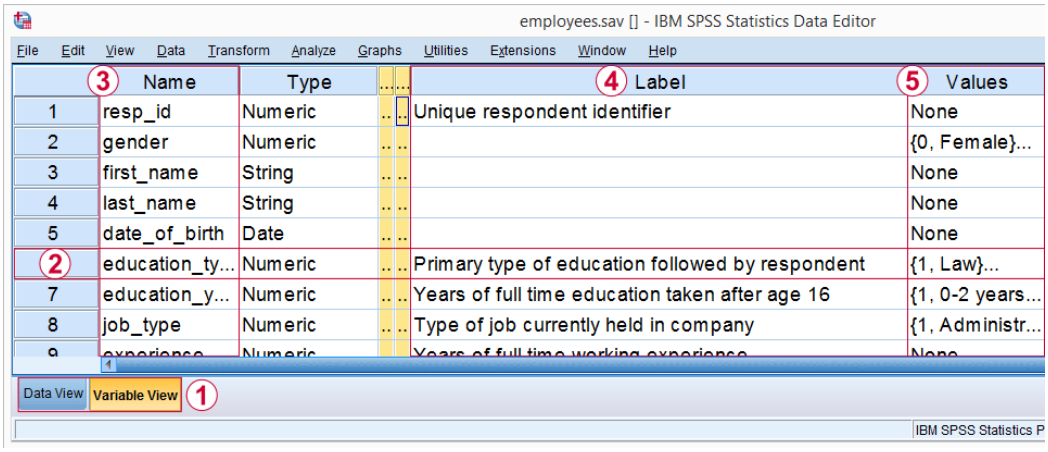
4) In SPSS, values refer to cell contents.

5) The status bar may give useful information on the data, for instance whether a WEIGHT, FILTER, SPLIT FILE or Unicode mode is in effect.

1.3 SPSS Variable View

An SPSS data file always has a second sheet called variable view. It shows the metadata associated with the data. Metadata is information about the meaning of variables and data values. This is generally known as the “codebook” but in SPSS it's called the dictionary.

For non SPSS users, the look and feel of SPSS’ Data Editor window probably come closest to an Excel workbook containing two different but strongly related sheets.



1) In the left bottom corner we find tabs for switching between Variable View and Data View. For now, select Variable View.

2) In Variable View, variables are shown as rows of cells.

3) The first column shows the variable name for each variable.

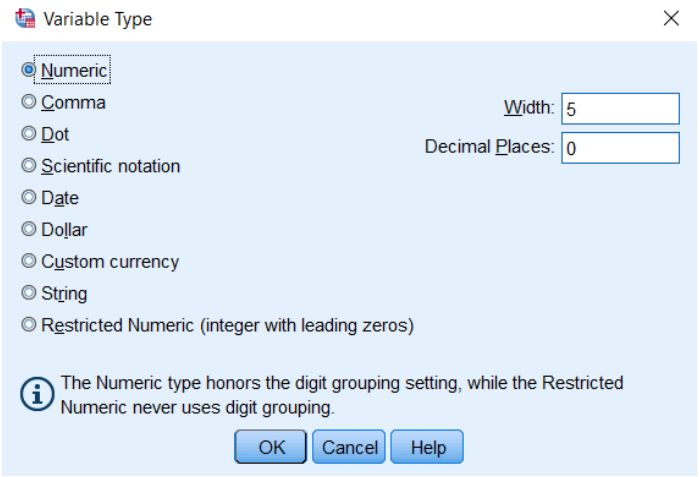
4) The fifth column may or may not contain a variable label. This describes the exact meaning of each variable.

5) The sixth column shows value labels: descriptions of the meaning of one, many or all values that a variable may contain.

In short, Variable View does not show the data itself but, rather, information about the data. This is sometimes called “metadata” or “the codebook”. In SPSS, however, it's called the dictionary.

1.4 SPSS Variable Types

What type of data you want to enter. Click the square to the right of the box to open a dialog box for options. Most of these variables are self-explanatory (like dollar, date, or scientific notation), but there are a couple that aren’t.



Numeric: Numeric variables, as you might expect, have data values that are recognized as numbers. This means that they can be sorted numerically or entered into arithmetic calculations. When viewed in the Data View window, system-missing values for numeric variables will appear as a dot (i.e., “.”). (Note that one should not type in a period character in a cell to specify a missing value. Simply leave the cell blank, and SPSS will

recognize it as system-missing.)

Continuous variables that can take on any number in a range (e.g., height in centimeters and weight in kilograms) should be treated as numeric variables.

Counts (e.g., number of people living in a household) should be treated as numeric variables with zero decimal places. In this situation, the Measure setting should be defined as Scale. Certain mathematical calculations are valid when applied to count variables (e.g., mean and standard deviation), but some statistical procedures requiring continuous numeric variables may not be (e.g., the dependent variable in a linear regression), depending on the distribution of the variable.

String variable: Use when you want to type letters. For example, peoples' names, breeds of dog, occupations. You can also include numbers or symbols, but they will be treated by SPSS as text. For example, zip codes are numeric but you may want to treat them as text (i.e. you don't actually want to perform calculations on them like 90210 * 10 !).

Comma: Numeric variables that are separated every three places by a comma. For example, 100,000.00 or 999,988,565.21.

Dot: Similar to comma, but the dot is used to separate the three places and a comma is used to indicate a decimal. For example. 100.000,00 and 999.988.565,21. Not used in the UK or USA, but common in some other countries.

2. Descriptive Statistics SPSS

If you are working in huge numbers of data, descriptive statistics help you to provide the summary and the characteristics of the data.

Descriptive statistics is a statistical analysis process that focuses on management, presentation, and classification which aims to describe the condition of the data.

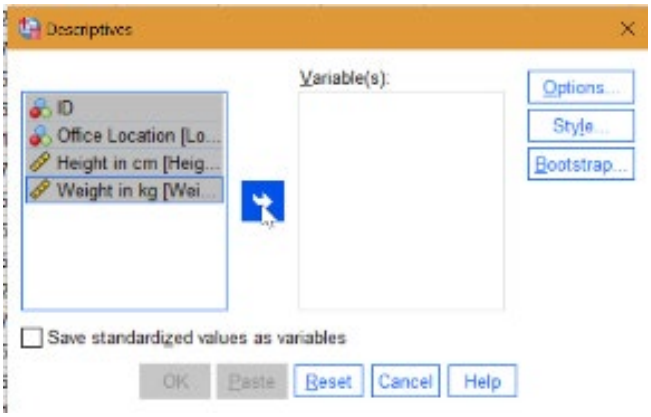
With this process, the data presented will be more attractive, easier to understand, and able to provide more meaning to data users.

In general, descriptive statistics must be able to give an idea of what information can be obtained from the data we use. Instead of just using numbers without a standard format, it would be more interesting if displayed in graphs and tables.

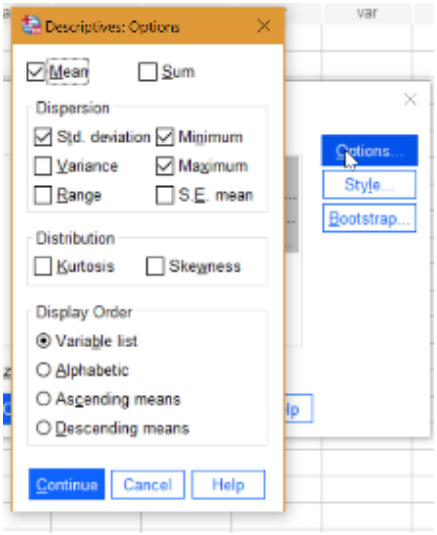
2.1 Descriptives Option

Step 1: Click Analyze, mouse over Descriptive Statistics, and then click Descriptives to open the Descriptives box.

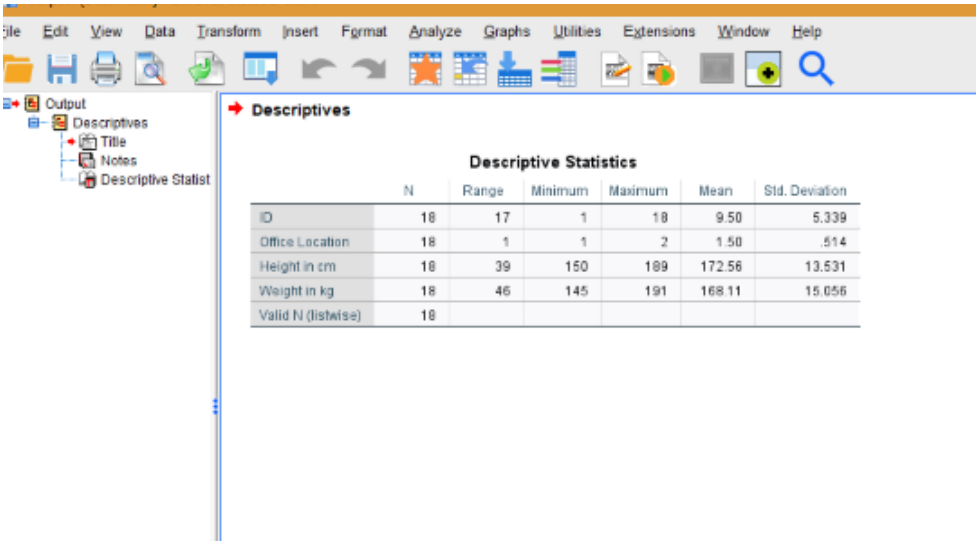
Step 2: Select the variables you want descriptive statistics in SPSS for. If you have multiple variables, you can select one at a time or hold down the Ctrl key and click to select all of your variables of interest. Click the blue arrow in the center to move the selected variables from the left to the right-hand Variables box.



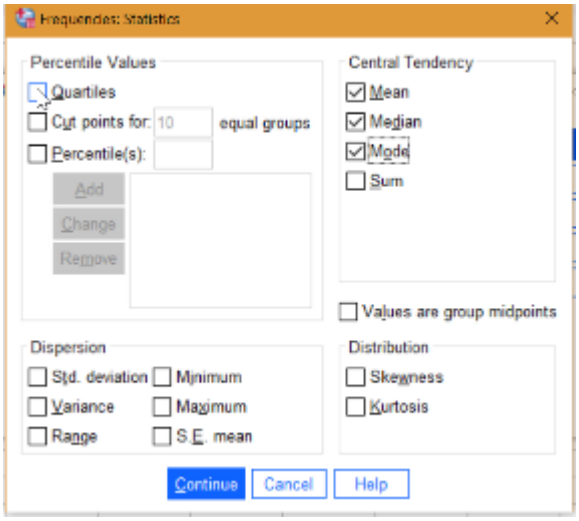
Step 3: Click Options in the right-hand column of the Descriptives box. This action opens the Options window.



Step 4: Click Continue, then click OK. The descriptive statistics SPSS calculates will be displayed in the output window. Unlike many SPSS tests and features, a single box will appear with your chosen statistics; There is no need to scroll down.



2.2 Frequencies Option



Step 1: Click Analyze, then mouse over Descriptive Statistics, and then click Frequencies to open the Descriptives box.

Step 2: Move the Variables you want to analyze. Click the variables one at a time (or all together), then click the blue arrow to move them over to the Variables box.

Step 3: Click Statistics in the right-hand column (the top blue button) to open the Frequencies: Statistics box.

Step 4: Click the box to check the statistics you would like. More options are available here, like Quartiles, which were not available in the Descriptives box in Part 1 above.

Step 5: Click Continue, then click OK. The Descriptive Statistics SPSS output window will display the requested results.