NDONG HENRY NDANG

Buea, SW Region, Cameroon

ndonghenryndang@gmail.com

Summary Guide To

Data Analysis

with IBM SPSS Statistics

By NDONG HENRY NDANG.

Chapter 1

Basics of Data Analysis

***

"Data are becoming the new raw material of business"

***

- **Data:** it simply refers to facts and statistics collected together for reference and or analysis….

  Collecting data is very important and crucial step in the process of data analysis, this is because whatever results are obtained and whatever their interpretations may be will depend solely on the data collected.

  Now after getting or collecting data, what do you do with? ...

  Remember the essence of every data collected is to get useful *information* which intent enable us get *insights*, hence the process of *data analysis* comes in….

- **Data Analysis:** it is simply the process of talking that data collected and turning it into information that can be used to make strategic decisions.

***

"If you do not know how to ask the right questions, you discover nothing"

***

Take note, ***information*** is simply data that make sense and getting information from our data is not all we desire to achieve….

Data analysis goes beyond that, it actually aims at making sense of all the information we have to find patterns and trends and correlations. It's a way to get insights that can help you make better decisions, identify opportunities and problems and track progress just to name a few.

***

"Torture the data, and it confess to anything"

***

And it's not just for businesses, the Government, non-profit organizations and individuals can benefit from data analysis. *The key is to figure out what questions you want answered and then use the right tools to find the answers…*

***

"The goal is to turn data into information, and information into insight."

***

- **Basic steps of data analysis:**
    1. *Defining your question*, what are you trying to find out?
    2. *Gather your data*, this can be anything from surveys, questionnaires, sales figures etc.
    3. *Analyze your data*, what are the trends or what's happening in different areas.
    4. *Interpret your findings*, what do your results signify
    5. *Take action*…

***

"You can have data without information, but you cannot have information without data"

***

- **How can data be used in decision making:** you can use data analysis in a lot of ways but one of the most important is to make informed decision. By understanding what your data is telling you, you can figure out what is working and what is not working and adjust your strategies accordingly….

  For example you are running a market campaign, you can use data analysis to figure out how well its performing, where your leads are coming from and what kind of return on investment you are getting. This information will help you make decisions about whether to continue with the campaign, tweak it or scrap it altogether.

  Data analysis is also a great tool for troubleshooting if your experiencing problems with your business, data analysis can help you identify the root cause and find solutions.

***

"Data is the new science. Big data holds the answers"

***

- **Nature of data:** data can be *Quantitative* or *Qualitative*.

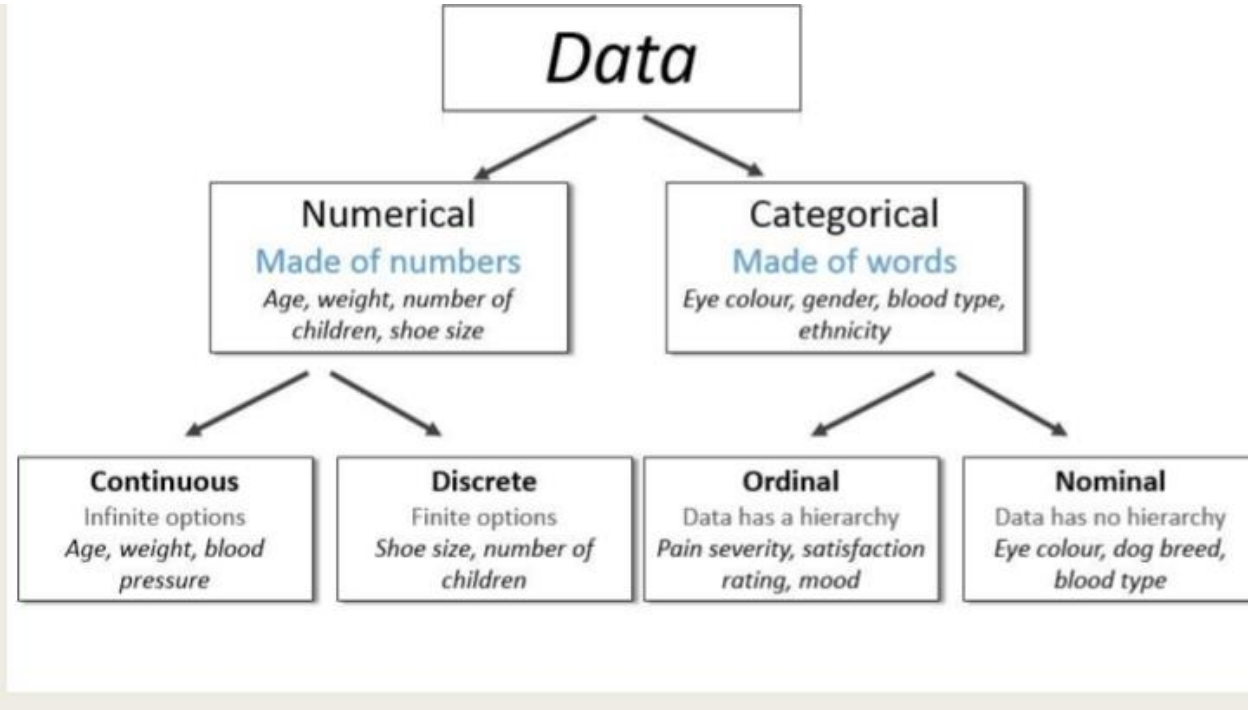
  Quantitative data is any data that can be counted or is data that are in form of values or counts and expressed as numbers for example age, weight or even height. *Qualitative data* is descriptive referring to things that can be observed but not measured. For example gender (male or female), yes or no…

  1. **Quantitative data:** can be further divided into two;
     a) **Discrete quantitative data:** takes on fixed numerical values and cannot be broken down further, it usually represent a whole (precise) number, for example the number of people in a hall or number of books on the table.
     b) **Continuous quantitative data:** can contain in between values such as decimals, for example, height in cm or temperature in degree.
  2. **Qualitative data:** which can be nominal or ordinal
     a) **Nominal data:** is used to label or categorize certain variables without any order or hierarchy, for example, gender is a nominal variable because there is no order whether male is to come before female and vice versa
     b) **Ordinal data:** is when the categories used to classify your qualitative data falls into a natural order, for example, level of education would be ordinal because to be in secondary school you must go through primary school or to be in class 4, you must have passed the lower classes which gives a certain hierarchy in naming the variable.

- **Pictorial Summary:**

Steps of data analysis

Classification of data:



Chapter 2

Statistics

- **Statistics** is the discipline that concerns the collection, organization, analysis,

  Interpretation and presentation of data.

Statistics can also be sent as a science concerned with developing and studying methods for collecting, analysis, interpreting and presenting data.

- **Relationship between statistics and data analysis.**

  Data analysis observes trends and patterns in data, statistics validates those theories using scientific processes, that is while analysis provide means for inference, statistics provide a methodology for data collection and computation hence statistics is at the core of data analysis.

- **Categories/Main types of statistics**

  a) **Descriptive statistics:** which describe the properties of the sample or population data, that is, it gives a general idea on the variables in a dataset and enables analysts the patterns in the dataset. **Descriptive statistics cannot be used to draw inference or conclusions.**

  Descriptive statistics mostly focus on the central tendency, variability and distribution of sample data.

Central tendency locates the distribution of data by various points and is used to show average or most commonly indicated responses in a data set. Measures of central tendency include the mean, median and mode.

Variability or dispersion refers to a set of statistics that show how much difference there is among the elements of a sample or population. It also denotes the range and width of distribution values in a dataset and determines how spread apart the data points are from the center and these measures include range, variance and standard deviation of scores in a sample.

The distribution refers to the overall sharp of the data which can be depicted in a chart such as the histogram and includes such as the skewness and kurtosis.

Descriptive statistics help us understand the collective properties of the elements in a dataset and forms the basis of testing hypothesis and making inferences.

Descriptive statistics for Qualitative data:

We use frequency tables and percentages to find out how many or probably what percent of the sample lie within a certain class or category. The charts used here are either pie or bar charts.

Descriptive statistics for Quantitative data analyses or study the mean, standard deviation, median, kurtosis and so on. The histogram is commonly used here.

b) Inferential statistics focus on making predictions about a large group of data based on a representative sample of the population. A random sample of data is considered from a population to describe and make inferences about the population.

Inferential statistics are used to make generalizations about large *groups*, such as estimating average demands for a product by surveying a sample of consumers' buying habits or to attempt to predict the future return of a security or asset class. Inferential statistics can be regression analysis, hypothesis testing etc.

Chapter 3

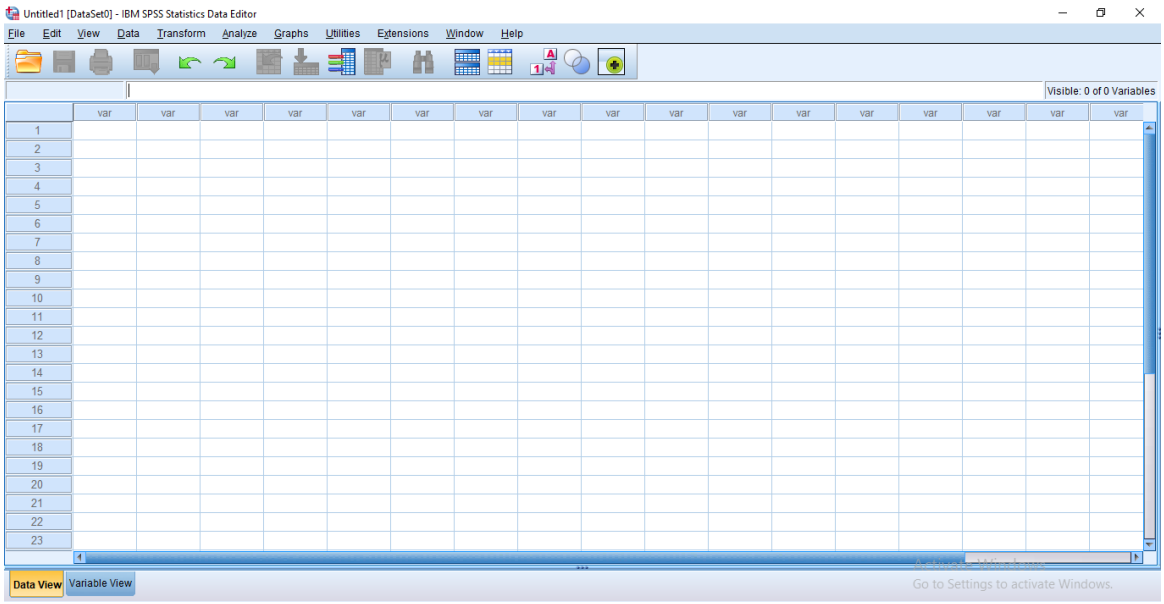
The SPSS Interface

Introduction to the SPSS software: SPSS is a data analysis software which can be used for statistical analysis. SPSS is a simple to use software due to its very friendly user interface with commands perform with just a few clicks. Nonetheless SPSS also has a syntax window where you can either write your own commands or code or simply paste each command you select, being able to paste more than a single command and run multiple queries at a time.

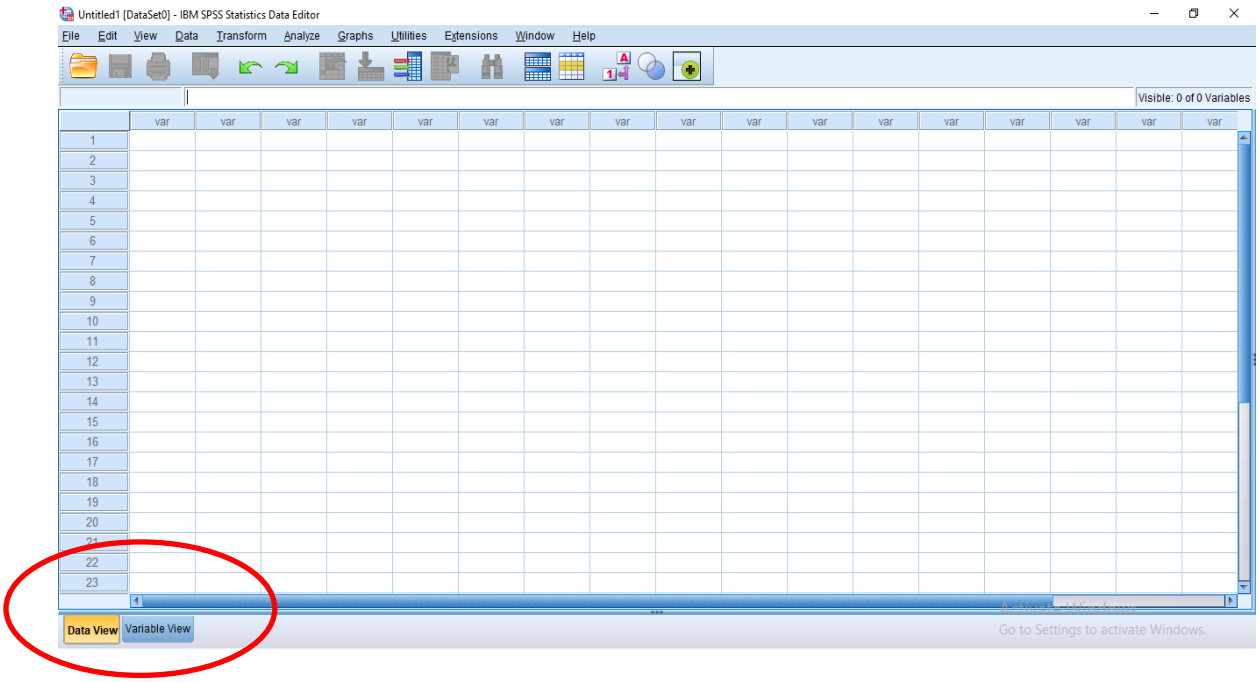
SPSS has a number of windows which are ***the syntax editor window, the data editor window and the output viewer window.***

- **Data editor window:**

  This is the window where you can see your data and the information about the variables in your data set. Below is the SPSS data editor….



The data editor window has 2 views,

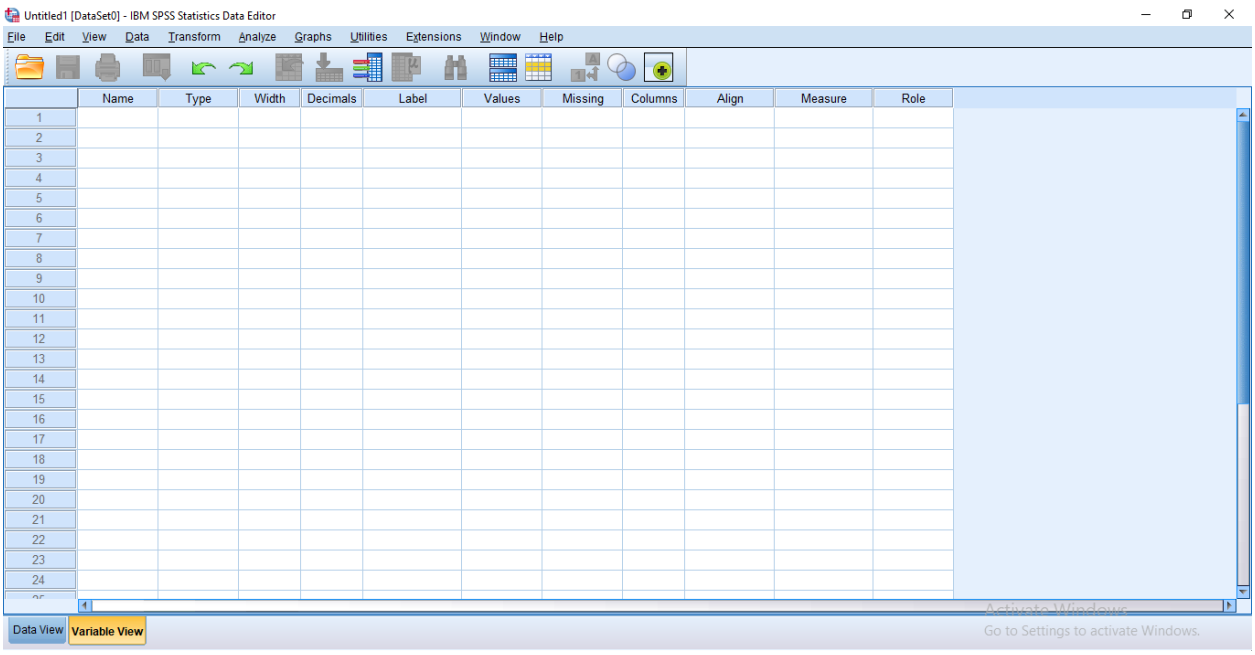


a) **Data view:** presently our data editor window is in the data view that is why the button is yellow. This is where you can actually see your dataset for each record and each variable.

   The data view has a column which holds data for one particular variable about all those involve in the sample and rows which holds all the data concerning each participant.

b) **Variable view:** is simply brought to live once you click on the icon 'variable view'. This is where you are given a summary for each variable in your dataset including the variable name, type, and other properties about the variable, that is, the variable view is all about information patterning to your variables.

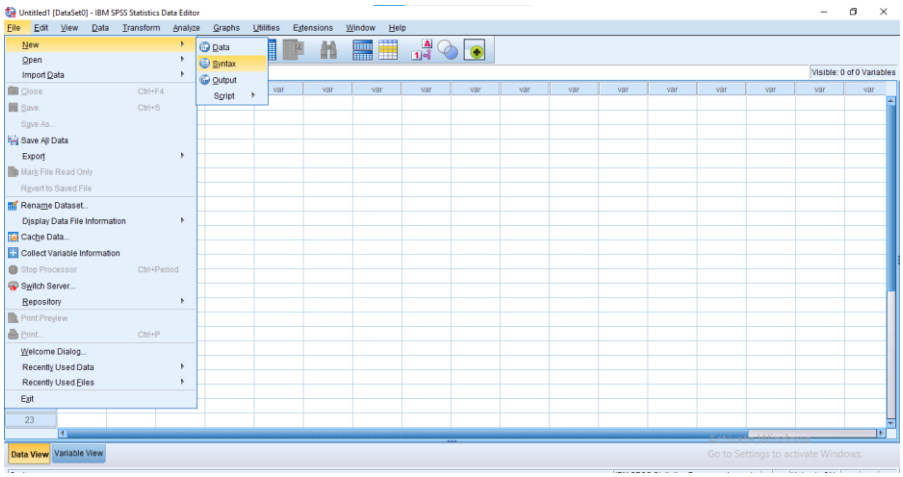
   Now let's explore the variable view,



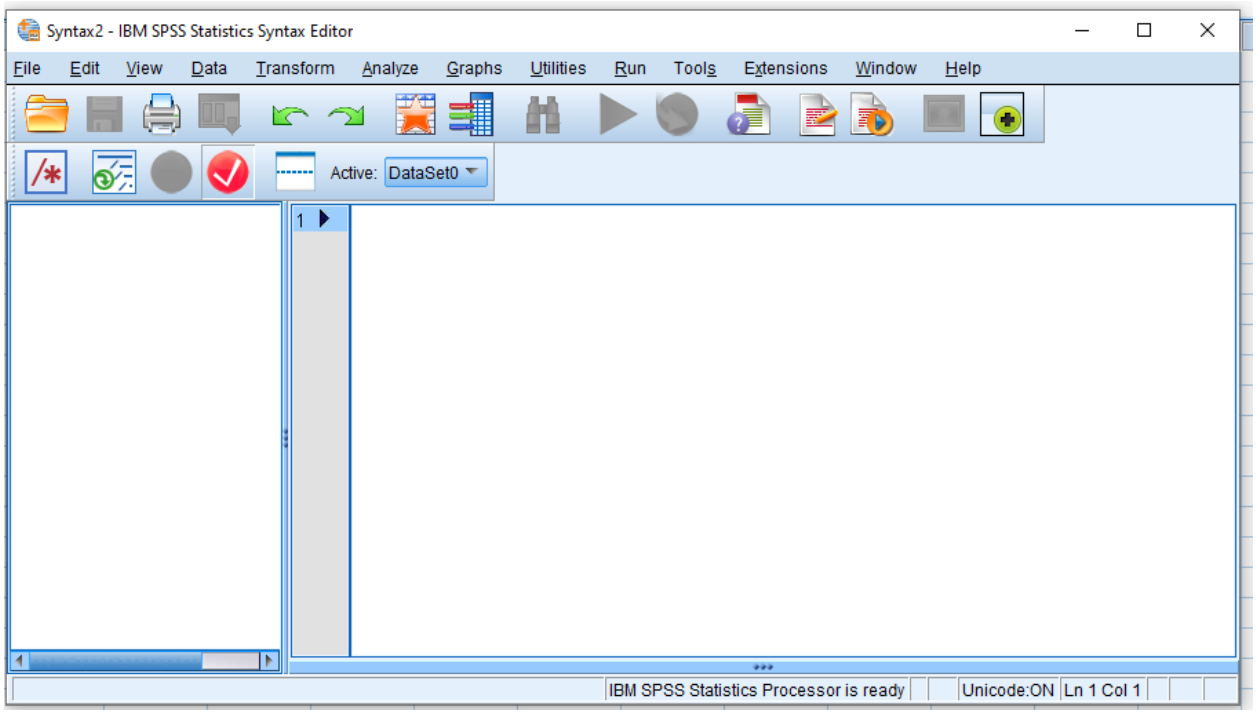
So this is the variable view which gives us information about our variable. Here, each column represent a single property about each variable like the name of the variable, type, number of decimal places contained in the variable etc.. Each row represents all the properties of a single variable.

- **Syntax editor window:** this window is not automatically opened in SPSS, so let's open one now,

  Go to file---New---syntax



After the above command, we obtain a blank file window as seen below,



- **Output viewer window:** any time you do something in SPSS, even just opening a file, SPSS will document it in the output window. It is opened automatically when SPSS is opened.

  The output window will keep a record of any and all commands you give SPSS. It is also the window in which you can view the results of any data/statistical procedures undertaken...

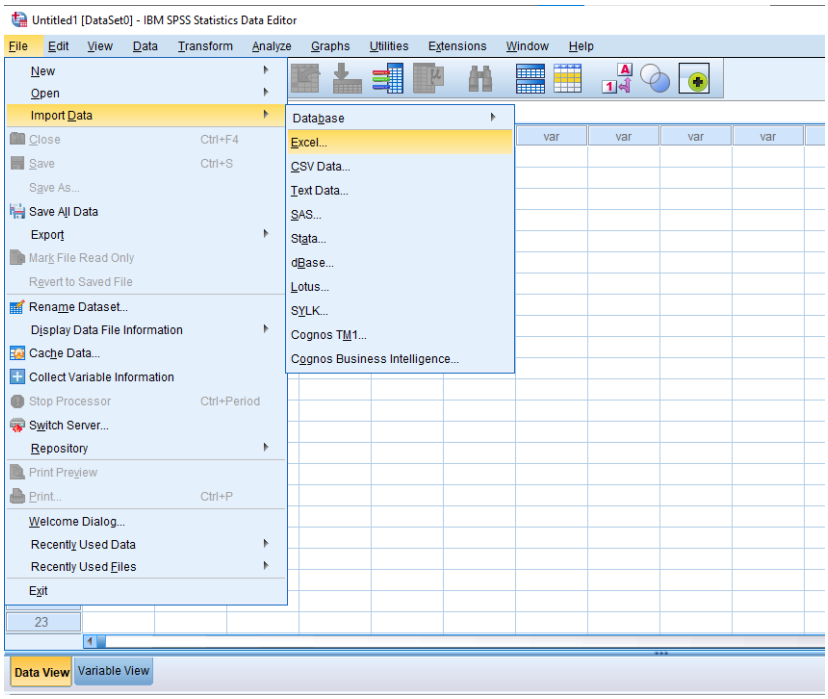
Chapter 4

Importing, Entering or modifying data

Importing data from Excel to SPSS: one of the most common formats for entering and storing raw data is MS excel. If you want to work with your data in SPSS, however, you will need to import the data from excel to SPSS. This can be relatively straight forward.

- **Task: Importing data from excel into SPSS**: to open an excel file in excel, follow the command below,

  Go to file--- import data---excel:



Next navigate and select the file where you stored it.

Next, check open----ok, now our data has been imported.

- **Inspecting our imported data:** Before you start analyzing data, you need to familiarize yourself with your data. I like to call "***getting to know your data***".

Now, the most important step is to go to variable view and ensure everything is in order.

Note: Analysis and tests in SPSS are best performed with data which is numeric, limitations might be encountered with string data, reason why it is important to covert or recode all your data (categorical data) into numeric values in which case their type will be changed to numeric. This is achieved by assigning values to categorical data, for example, 1=male and 2=female, here we have simply recode male and female into 1 and 2 respectively.

Next we would want to check and make sure our variables have the right measures selected (nominal, ordinal or scale). This is also important.

We also check and indicate if any missing values exist in our dataset and also label our variables (give our variables more descriptive names)

Once all this is done we can go ahead to do a codebook/data dictionary for our data or some simple commands to understand our data.

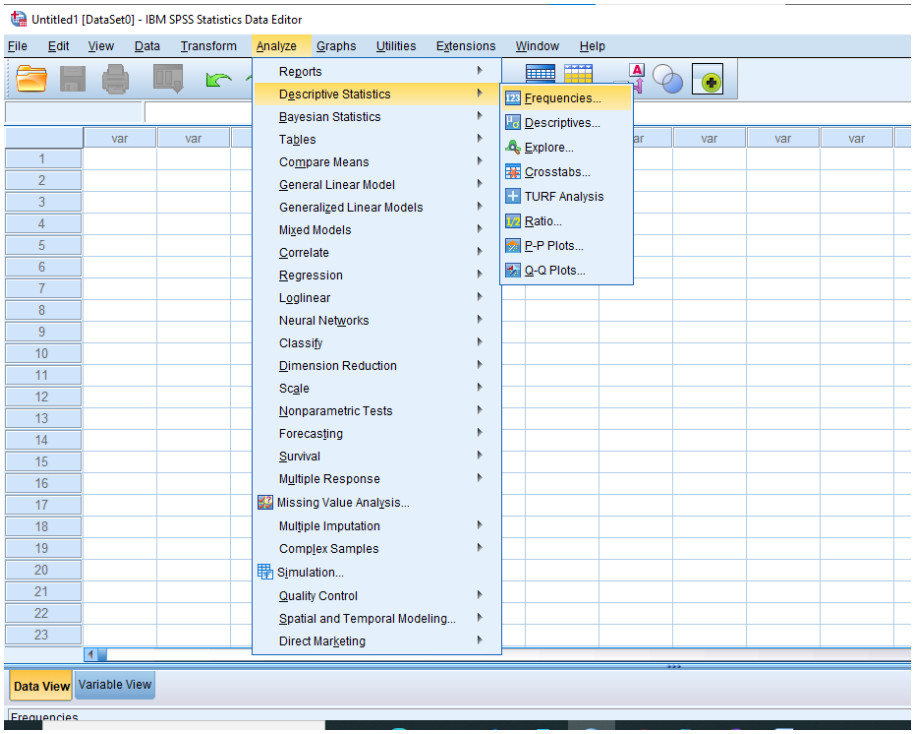
Chapter 5

Describing Data

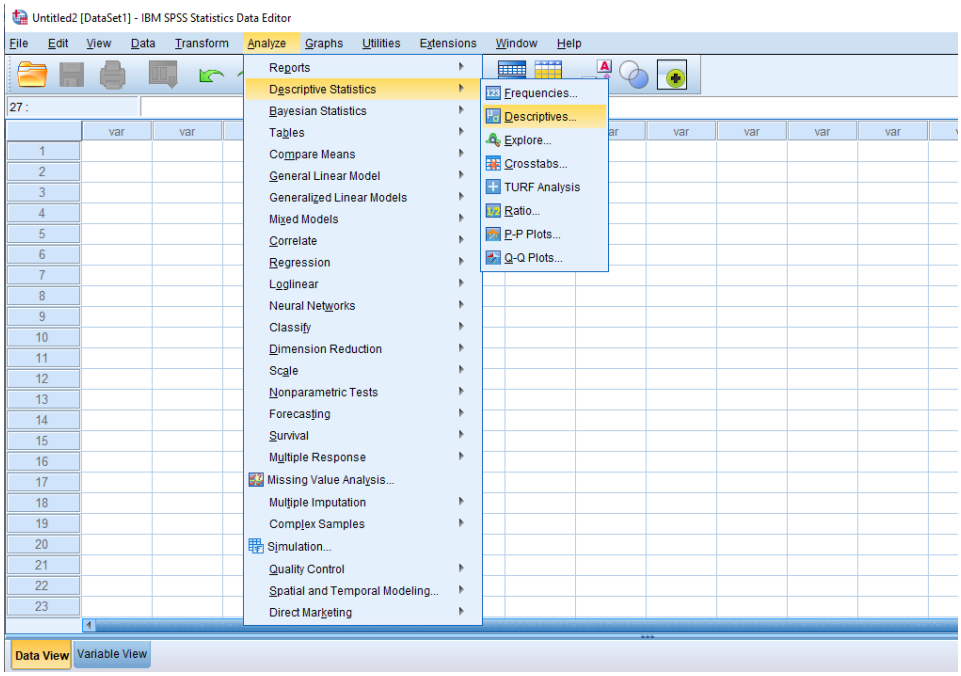
Before initiating data analysis, it's a good idea to check the frequencies of all variables, how categorical variables have been coded, the minimum and maximum values and number of missing observations. This is a good way to identify any outliers and potential mistakes in the dataset.

There are several commands available to describe your data;

a. **Frequencies command:** provides counts and proportions of all values within a variable. This is a good way to summarize categorical data. Here is the SPSS command;



b. **Descriptives command:** provides an overall, minimum, and maximum values, means and standard deviation. Good for summarizing continuous data.



c. **Explore:** Provides summary measures of a continuous variable against categories of another (categorical) variable.

d. **Crosstabs:** cross tabulates counts and or percentages of categories in one variable against categories in one or more variables.

Chapter 6

Inferential statistics

Inferential statistics is a way of making inferences or draw conclusions about populations based on samples.

It helps in making generalizations about the population by using various analytical test and tools. This is the main goal of inferential statistics.

Types of inferential statistics:

Inferential statistics can be classified into hypothesis testing and regression analysis.

Hypothesis testing: is a type of inferential statistics that is used to test assumption and draw conclusions about the population from available sample data. It involves setting up a null hypothesis and an alternative hypothesis followed by conducting a statistical test.

Regression Analysis: used to quantify how one variable will change with respect to another variable. Most common is linear regression. Linear regression checks the effect of a unit change of the independent variable in the dependent variable.

- **Hypothesis Testing:** is a formal procedure used to test whether a hypothesis can be accepted or not. A hypothesis is an assumption about something. For example, a hypothesis about height could be that the average height of adult is 1.7cm.

  We have 2 types of hypothesis

  a. **Null hypothesis:** it is the hypothesis of equality or no difference. The null hypothesis always says the two or more quantities (parameters) are equal.
  b. **Alternative hypothesis:** it is the negation or opposite of the null hypothesis.

Note that we always test the null hypothesis not the alternative. We either reject or fail to reject the null hypothesis, if we reject the null hypothesis, only then can we accept the alternative hypothesis

It is therefore very important to have a clear understanding about the null hypothesis.

We say the test is significant if we reject the null hypothesis.

Checking data for Normality: checking and ensuing data has come from a normally distributed population is an important assumption before a parametric test or method is used (which gives more credibility to your inferences or conclusions) otherwise a non-parametric test is used.

There are many ways to check for normality and three of these ways are;

a. Graphs: such histograms and Q-Q plots
b. Descriptive statistics: using skewness and kurtosis.
c. Formal statistical test such as *Kolmogorov-smirnov (k-s) test and shapiro-wilk test*.

Data transformation: if your data is not normally distributed, you can use data transformation to make the dataset normally distributed in order to use a parametric test. The types of data transformation methods are

- Square root transformation
- ln transformation
- log transformation
- inverse transformation

Parametric Test for hypothesis testing

- **T-test**

1) **One sample t-test:** is carried out when we want to compare the mean of a variable with a hypothetical (standard) value. For example we want to find out if the mean height of students in our sample is equivalent to a certain value say 1.7cm.

   Assumptions before a t-test is carried out

   -The distribution of the variable in the population is normal.

   -The sample is a random sample from the population

2) **Independent sample t-test:** involves one categorical variable with two levels (2-categories) and one quantitative variable. This test is done to compare the means of two categories of the categorical variable. For example we could be interested in finding out if the mean age (dependent variable) of diabetic and non-diabetic is same or not.

   Assumptions for this test to be used:

   -The dependent variable should be normally distributed at each level of the independent variable.

   -The variances of the dependent variable at each level of the independent variable are same.

   -Subjects represent random samples from the population.

3) **Paired sample t-test:** it is done to compare the difference between two means of related sample. Related samples indicate measurements taken from the same subjects in two or more different times or situations.

4) **ANOVA (Analysis of variance):** it is a statistical test used to analyze the difference between the means of more than two groups.

   Assumptions to use ANOVA:

   -Your independent variable should be continuous.

   -Your independent variable should consist of two or more categorical, independent groups

   -Your independent variable should be approximately normally distributed for each category of the independent variable.

5) **Chi-square test:** it is a statistical method commonly used in data analysis to determine if there is a significant association between two categorical variables.
6) **Correlation:** it is a statistical test used to find out whether a relationship exists between variables and then determining the magnitude and action of that relationship.
7) **Linear regression:** regression is a predictive analysis and used when we want to predict the value of a variable based on the value of another variable.