

Micro-module D: Big Data Analytic

Big data is an extremely large volume of data and datasets that come in diverse forms and from multiple sources. Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions. Module D ‘Big Data Analytic’ is consist of 2 micro modules, which aim to provide beginner-friendly tutorials for students about data filtering and cleaning and statistical analysis.

1. Big Data and Big Data Analytic

Big Data is the term describing large sets of diverse data – structured, unstructured, and semi-structured – that are continuously generated at a high speed and in high volumes. Not only does Big Data apply to the huge volumes of continuously growing data that come in different formats, but it also refers to the range of processes, tools, and approaches used to gain insights from that data.

Big Data analytics encompasses the processes of collecting, processing, filtering/cleansing, and analyzing extensive datasets.



2. Micro-module D-D1: Data Filter and Cleaning

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

In this module, you will learn the basic procedures and steps to do data cleaning in Excel, including editing data format, selecting and replacing the blank values, trimming the space, filtering data, and conditional formatting.

Introduction of Data Cleaning

What are the different types of data issues?

- **Duplicate data:** There are 2 or more identical records. This may cause misrepresentation of inventory counts/duplication of marketing collateral or unnecessary billing activities.
- **Conflicting Data:** When there are some records with different attributes, it means data is conflicting. For example, a company with different versions of addresses may cause delivery issues.
- **Incomplete Data:** The data that has missing attributes. Payrolls of employees may not be processed due to their missing social security numbers in the database.
- **Invalid Data:** Data attributes are not conforming to standardization. For example, 9-digit phone number records rather than 10 digits.

DATA CLEANING CHECKLIST



Tools for Data Cleaning

R offers a wide range of options for dealing with dirty data. The collection of packages known as the tidyverse, and adjacent packages that take a "tidy" approach, provide a range of functionality. From importing to cleaning to reshaping, these packages can help you quickly and efficiently clean messy data.

EXCEL quick data-cleaning tips
Created by Miranda Lee January 2020

This resource provides strategies for cleaning data in Microsoft Excel. Below is a brief overview of five situations you may find yourself in ("What") and corresponding solutions ("How"), followed by detailed instructions to implement the solutions.

What? **How?**

- Identify all cells that contain a specific word or (short) phrase in a column with open-ended text. **Use Conditional Formatting**
- Identify and remove duplicate data. **Use Remove Duplicates function or Conditional Formatting**
- Identify the outliers within a data set (e.g., dates or grades). **Use Data Validation function**
- Separate data from a single column into two or more columns (e.g., first and last names). **Use Flash Fill**
- Categorize data in a column, such as class assignments or subject groups. **Use Formulas to fill in the category column**

Remove Rows

One way to deal with empty cells is to remove rows that contain empty cells. This is usually OK, since data sets can be very big, and removing a few rows will not have a big impact on the result.

Example
Return a new Data Frame with no empty cells:

```
import pandas as pd
df = pd.read_csv('data.csv')
new_df = df.dropna()
print(new_df.to_string())
```

Try it Yourself

While you can certainly do data cleaning in Excel, switching to R enables you to make your work reproducible. Similarly, we can use Python's Pandas and NumPy libraries to deal with messy data, whether that means missing values, inconsistent formatting, malformed records, or nonsensical outliers.

Trim the Space

property_id	web_URL	address	date	Lat	Lng	room	sold price	SA1	SA2	GFA1	GFA2
5	https://hk.18	Yan King Road	2021/11/19	22.32414	114.2572	3	10,150,000	659 ft²	15,402/ft²	926	10,961/ft²
6	https://hk.11	Tong Chun Street	2021/11/19	22.3075	114.2624	NA	6,800,000	591 ft²	11,506/ft²	1,045	11,981/ft²
7	https://hk.31	Razor Hill Road	2021/11/19	22.33731	114.2492	NA	12,520,000	842 ft²	14,869/ft²	743	11,981/ft²
8	https://hk.21	Hiu Kwong Street	2021/11/19	22.32173	114.2289	NA	3,900,000	540 ft²	7,222/ft²	652	5,249/ft²
27	https://hk.1	Po Lam Road North	2021/11/18	22.31973	114.253	NA	5,180,000	588 ft²	8,810/ft²	617	7,945/ft²
29	https://hk.9	Tong Tak Street	2021/11/18	22.3072	114.2571	2	8,500,000	465 ft²	18,280/ft²	596	13,776/ft²
31	https://hk.21	Hiu Kwong Street	2021/11/18	22.32138	114.2283	NA	3,980,000	433 ft²	9,192/ft²	606	6,678/ft²
32	https://hk.1	Ngan O Road	2021/11/18	22.31386	114.2658	2	7,900,000	436 ft²	18,119/ft²	644	13,036/ft²
33	https://hk.13	Laguna Street	2021/11/18	22.30491	114.2268	2	8,200,000	517 ft²	15,861/ft²	939	12,733/ft²
34	https://hk.13	Laguna Street	2021/11/18	22.30491	114.2268	3	11,500,000	748 ft²	15,374/ft²	677	12,247/ft²
35	https://hk.11	Lei Yue Mun Road	2021/11/18	22.30491	114.2268	2	5,200,000	506 ft²	10,277/ft²	1,955	7,681/ft²
36	https://hk.20	Shan Kwong Road	2021/11/18	22.30491	114.2268	3	36,000,000	1,531 ft²	23,514/ft²	1,955	18,414/ft²

3. Micro-module D-D2: Statistical Data Analysis

SPSS is a widely used program for statistical analysis in social science. 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.

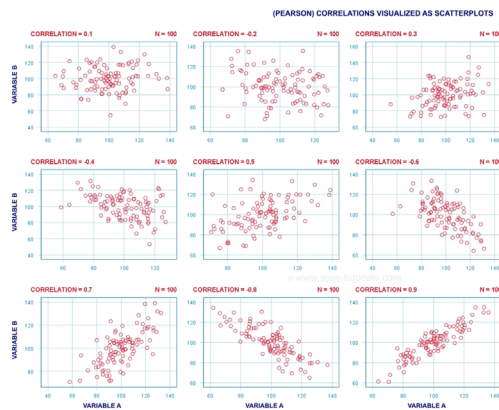
Introduction of SPSS

What is 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.



<https://www.spss-tutorials.com/spss-correlation-analysis/>

Variable Type

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., ".").

STRING

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.

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.

Frequencies Option

The screenshot shows the SPSS Statistics Data Editor interface. The main window displays a data table with 28 rows and 13 columns. The first column contains street names, and the subsequent columns contain numerical values. A dialog box titled 'Frequencies: Statistics' is open in the center of the screen, allowing users to select statistical options for the data.

Row	Street Name	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10	Col 11	Col 12	Col 13
1	Tai Po Road #1	11	5	6	2								
2	Tai Po Road #2	4	4	0	2								
3	Tai Po Road #3	4	4	0	3								
4	Tai Po Road #4	2											
5	Sha Tin Centre St #1	34											
6	Sha Tin Centre St #2	16											
7	Sha Tin Centre St #3	10											
8	Sha Tin Centre St #4	12											
9	Sha Tin Centre St #5	21											
10	Sha Tin Centre St #6	27											
11	Sha Tin Centre St #7	7											
12	Tam Koon Poon St #1	15											
13	Tam Koon Poon St #2	25											
14	Yuen Wo Road #1	5											
15	Yuen Wo Road #2	20											
16	Wam Pok St #1	21											
17	Wam Pok St #2	27											
18	Wam Pok St #3	24											
19	Pak Hok Tin St #1	8											
20	Pak Hok Tin St #2	4											
21	Pak Hok Tin St #3	3											
22	Tin Ho Road #1	9	9	0	1								
23	Tin Ho Road #2	14	13	1	2								
24	Tin Ho Road #3	13	13	0	1								
25	Tin Mai St #1	9	6	3	2								
26	Tin Mai St #2	12	12	0	2								
27	Tin Yui Road #1	28	16	12	11								
28	Tin Yui Road #2	23	18	5	7								

The 'Frequencies: Statistics' dialog box includes the following options:

- Percentile Values:**
 - Quartiles
 - Cgt points for: 10 equal groups
 - Percentiles
- Central Tendency:**
 - Mean
 - Median
 - Mode
 - Sum
- Dispersion:**
 - Std. deviation
 - Minimum
 - Variance
 - Maximum
 - Range
 - S.E. mean
- Distribution:**
 - Skewness
 - Kurtosis
- Values are group midpoints

Buttons: Add, Change, Remove, Continue, Cancel, Help.