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.

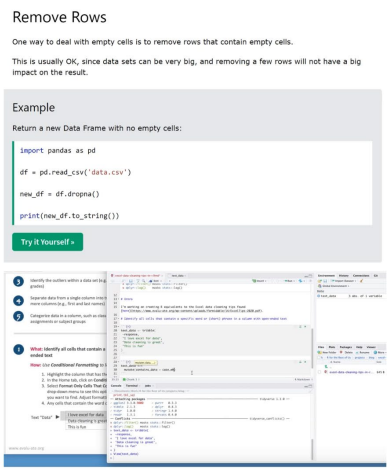
Introduction of Data Cleaning

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.

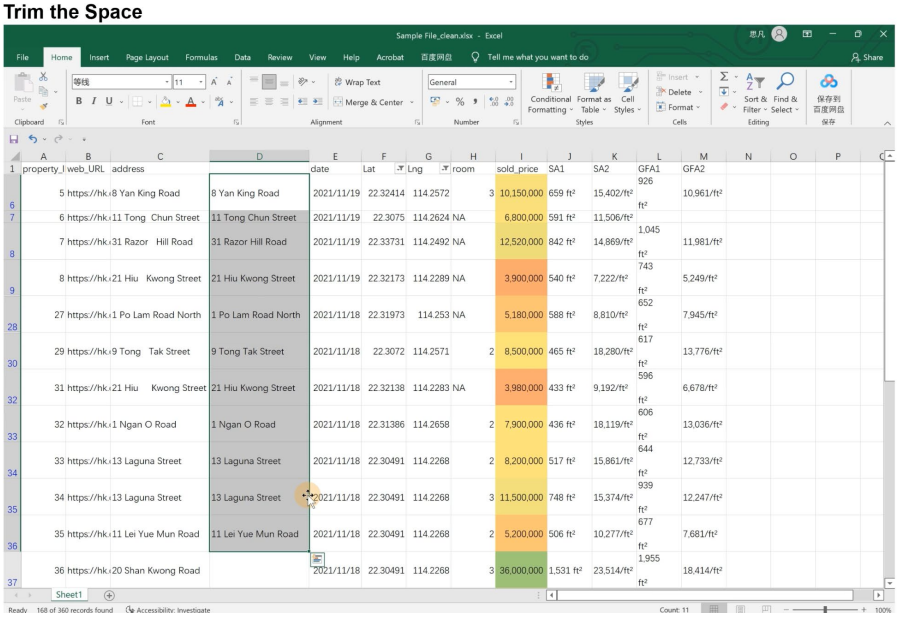
Tools for Data Cleaning

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.



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 to deal with messy data, whether that means missing values, inconsistent formatting, malformed records, or nonsensical outliers.

Trim the Space



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

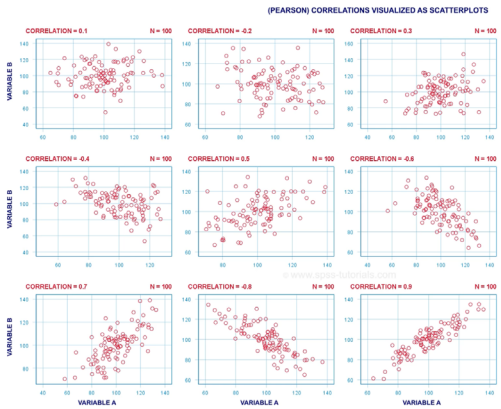
SPSS is a widely used program for statistical analysis in social science.It is also used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations, data miners, and others.

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