Data sources and access

An introduction to spreadsheets

What is a spreadsheet file?

A **spreadsheet** is a file that contains data arranged in rows and columns.

Data stored can either be in plain form, such as text, numbers, etc., or as functions.

Text

Number

Function

Microsoft Excel

Google Sheets

LibreOffice

The data in a spreadsheet file can be captured, viewed, or manipulated in various ways using a **spreadsheet application.**

Important spreadsheet terminology

01. Cell

Cells are the boxes we see on the grid of a worksheet. Each cell represents a single data point.

02. Cell reference

A cell reference is used to identify a cell. It is made up of the column letter and the row number in which the cell is located; for example, column A and row 1 become cell A1.

03. Column

Columns are the vertical series of cells in the spreadsheet. Each column is identified by a specific column letter located at the top of the worksheet.

04. Row

Rows are the horizontal series of cells in a spreadsheet. Each row is identified by a specific row number located on the far left of the worksheet.

05. Range

A range is a collection of two or more selected cells running across a column, row, or a combination of both.

06. Formula

A formula is a mathematical equation designed by the user to perform calculations on a set of data and return the result in a given cell.

Important spreadsheet terminology

07. Function

Functions are predefined formulas that are already available in the spreadsheet application, which makes it easier to perform calculations.

08. Argument

An argument refers to the specific values required by a function for it to perform its calculation and return a result.

09. Worksheet

A worksheet is a single page of rows and columns.

10. Workbook

A workbook refers to the entire spreadsheet file containing a collection of one or more worksheets.

Text file formats

> The two most commonly used text file formats in spreadsheets are **comma-separated values (.csv)** and **tab-separated values (.tsv)**.

.csv

The most common delimiter for a CSV file is a **comma**. However, it can also have a **semicolon, tab**, **space**, or any other **custom** delimiter characters.

It is also possible to have a CSV file with inconsistent delimiters.

.tsv

A TSV file is similar to a CSV, only it uses the **tab** delimiter to separate the values.

It is more suitable for text-heavy files, as it is less likely for there to be a tab character within the text.

Spreadsheet file formats

A spreadsheet file can be saved in **several various formats**, each having a different file extension. Spreadsheet file formats not only store data but also **other properties surrounding the data**, such as cell and font formatting, formulas, and functions.

Common spreadsheet file formats include:

.xls

- Excel **Binary File** format.
- It was the default format in Microsoft Excel 97-2003.

*__Binary file__ – a non-text file that stores data in the form of a series of bytes, each eight bits in length.

.xlsx

- The current default file format for Microsoft Excel.
- The **additional x** shows that the file is based on **the XML** standard.

*__XML__ – a self-describing markup language that defines a set of rules for encoding data and documents.

.xlsm

- XML-based and macro-enabled file format.
- The **additional m** shows that the file contains **macros.**

*__Macros__ – a set of instructions used for automating repeated tasks/processes.

Text file formats

> A text file is a digital file that contains **plain text only**. Text file formats **only store text strings** without any additional styling or formatting information.

Why use text file formats?

Spreadsheet files are **not suitable for storing data long-term**; they are software-specific and may become obsolete and unreadable with subsequent spreadsheet software upgrades.

Text files are more ideal for sharing and preserving original raw data.

- **Simple and lightweight—**they do not store any extra information.
- **Easily readable, even in the future—**they do not require special software.

In **text file formats:**

Each **line** represents a **row** and is separated from the next using a **line break**.

Each **value** in a **row** is separated from the other using a **delimiter character**, which eventually forms **columns.**

Text file formats—delimiters

A **delimiter** is one or more characters that are used to **split plain text** into separate data units.

Common delimiters

- **Comma (,)**
- **Semicolon (;)**
- **Tab (/t)**
- **Space ( )**
- **Pipe (|)**

Why use delimiters?

- They **organize plain text data** into rows and columns.
- They specify the exact location where **each data unit within a text string starts and ends—**we don't need fixed lengths within the text to be able to split it.
- By splitting the data, we can **perform certain operations** on it—such as sorting or filtering based on a particular component of the data.