**Raw Data**

Raw data typically refers to tables of data where each row contains an observation and each column represents a variable that describes some property of each observation. Data in this format is sometimes referred to as tidy data, flat data, primary data, atomic data, and unit record data. Sometimes raw data refers to data that has not yet been processed.

Raw data has the potential to become "information," it requires selective extraction, organization and sometimes analysis and formatting for presentation. Because of processing, raw data sometimes ends up in a database, which enables the data to become accessible for further processing and analysis in a number of different ways.

**How raw data works**

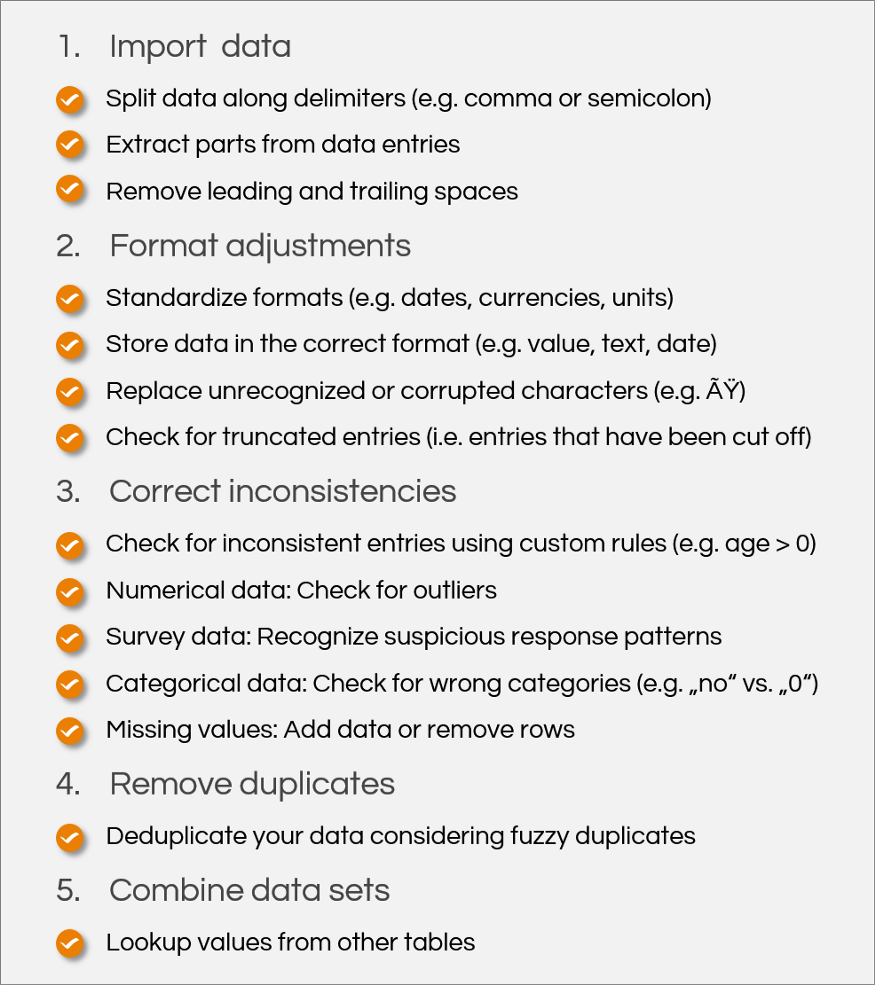
Tremendous amounts of raw data surround us and are produced every day. The human brain is incredibly good at taking in raw data, processing it and using it to make decisions.

For example, imagine you are trying to cross a busy road. The eyes capture raw data as flashes of light and dark. Then the brain takes these flashes and resolves them into objects such as street signs and cars. The working memory can tell you if that car is sitting still, getting bigger as it comes toward you, or getting smaller as it drives away. Meanwhile, the ears take in raw information in the form of vibrations in the air, which the brain translates into sounds that can be interpreted as the wind, voices or a car engine. Finally, all this processed data that came in through the eyes, ears and memory helps you make the informed decision to cross the street or not.

**Prepare your data for analysis using Excel**

Most data sets require preparation before analysis. Garbage in, garbage out — your analysis will only yield meaningful results if your data has high quality. Particularly when dealing with small or medium-size data sets, wrong entries and inconsistent values can have a tremendous impact on your analysis — they will introduce bias. The data preparation process can imply different tasks, depending on the type of data you analyse.

What should you pay attention to when getting your data ready for analysis? And how can you perform these steps efficiently in Microsoft Excel? Assuming that you already collected your data, you should go through the five steps of data preparation. Our checklist guides you through the data preparation process. The comprehensive explanations below provide you with some hints on how to implement these data cleaning tasks in Excel. Now, we show how to work with “pure” Excel functionalities.

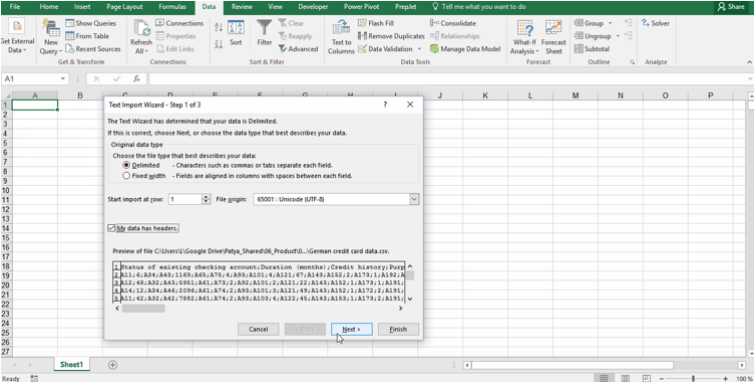


**Note**: Your raw data before going on with data preparation in case you want to lookup some information in the original data. One more tip beforehand: If you are dealing with large data sets that cause Excel to become slow when performing certain operations, simply try the operation with a small sample before applying it to the whole data set. You can grab some coffee while Excel is calculating. If your coffee consumption gets too high while preparing your data, you might want to prepare, which can perform certain operations more quickly than Excel formulas (e.g., when looking up data from different tables).

**1. Import data**

**Split data along delimiters**

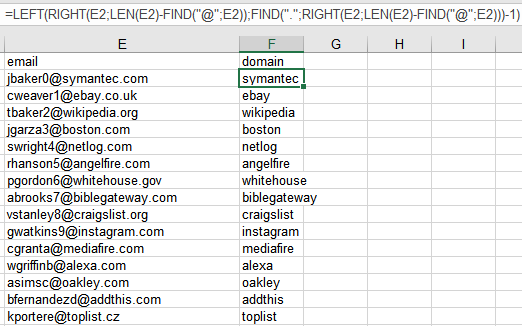
When you import your data, you should be aware of a clean delimitation of your entries. IT systems usually define a delimiter such as a semicolon or a comma. In Microsoft Excel, we recommend to use the import function to get your data into a tabular shape, also if you import a csv and not a txt file. Compared to simply opening the csv file, the import function has the advantage that you can define the character encoding. Our animation shows you how to use the import function.



If you need a more sophisticated separation of your columns, you can use “Separate Cell Content” function. It allows you to define the position of the delimiters.

**Extract parts from data entries**

In case your import did not work well or you need only a part of a certain attribute (such as a part from a nested ID), you have to perform more advanced split operations. For example, if you want to extract the domains of email addresses, you have to extract the part between the @ and the last dot. If you use original Excel functions, LEFT, RIGHT will help you to extract text parts. With FIND, you can specify the delimiters (in our case “@” and “.”), with LEN, you can define the length of the character sequences you want to extract. You will have to work around with a few combinations to finally get the domain:



**Remove leading and trailing spaces**

System exports or web-scraped data often cause unwanted leading and trailing spaces. You can remove those spaces by using the Excel TRIM function (do not use find and replace as this will also remove spaces between words!). The syntax is as follows:

=TRIM (text)

If you do not want to repeat this exercise for every single column, you can use Trim Spaces function which allows you to trim your whole sheet in a single click.

2. **Format adjustments**

Standardize formats

Before combining and analysing data, it is crucial to harmonize the formats of your data. If your data comes from different countries or IT systems with different languages, you should make sure to have consistent decimal separators (comma vs. dot). The same applies for date formats (e.g., DD.MM.YYYY vs. MM/DD/YYYY) currencies (e.g., EUR vs. USD) or measurement units (e.g., miles vs. kilometres). In Excel, the best solution is to transform data sets with different formats into one standard before combining them. For date types, you can use Excel’s “Number” functionality. Click on the lower right arrow to define a custom format:

If you want to unify the spelling of your data entries (e.g., UPPERCASE vs. Normal Case), you can use Change Case function.

**Store data in the correct format**

To make sure that your data is analysed appropriately, you should store it in the correct format. For example, your data might contain a numerical Identifier which has however no numerical meaning. Tell Excel that this is not a number by classifying it as text. Excel helps you to specify the date format in the Home Ribbon in the section “Number” (see above). To perform a final check on the format of your data.

**Replace unrecognized or corrupted characters**

In case your import failed for some reason, some symbols might be corrupted after an export as the encoding of characters might differ between IT systems. Characters like “ÃŸ” might introduce problems if you want to work with your data. We recommend to correct them using the find and replace function.

**Check for truncated entries**

When exporting data from IT systems, a lot of annoying accidents can happen. One of them is truncations, i.e., data entries are cut off at a certain position. Some manual screening should quickly lead you to suspicious data. How to fix it? The best solution is probably to request a new and healthy export.

**3. Correct inconsistencies**

Check for inconsistent entries using custom rules

The most valuable resource when preparing data for analysis is your own knowledge about the data. With custom rules (so-called Business Rules), you can detect wrong data. The whole process follows the principles of Boolean logic. You can define simple conditions. Before applying any rules to your data, it is always helpful to spell them in natural language, e.g.: “The entries in column ‘Age’ always have to be greater or equal zero.”

**Missing entries: Add data or remove rows**

Missing values are an issue that is hard to fix. If you have the chance to get the original data without inappropriate efforts, the best solution is to add it. Otherwise, you might want to ignore the rows that contain empty cells when performing certain analyses. A more advanced option is to impute the data (e.g. by the mean or median or by a logistic regression). This is only recommendable for users with advanced statistical knowledge as it might bias your results.

**4. Remove duplicates**

**Deduplicate your data considering fuzzy duplicates**

Another important step before you combine and analyse data sets is to remove duplicate entries. This is simple if you only want to find exact duplicates. You can directly remove them with Excel’s Remove Duplicates function in the Data Ribbon. If you first want to check the duplicate entries to examine where they come from, you can highlight and sort them.

**5. Combine data sets**

**Lookup data from other tables**

If you do not only want to analyse a single data set, the final step to get your data ready for analysis is to combine it. You could simply copy columns from one sheet to another if your data is sorted. However, this is not recommendable as you risk to copy the wrong data in case one data set misses a row. It is better to identify match criteria and pull your data based on these criteria from one table into another. If you have one match criterion that is unique for each row of your data set (a so-called unique identifier), you can use this match criterion and perform a VLOOKUP in Excel. If you have more than one match criteria, you have to use a combination of INDEX and MATCH