**Cleaning data in Excel and Reasons for Each Action:**

**Removing Duplicate or Inaccurate Data and Empty Rows**

* Spelling checking: Spelling errors could lead to mismatches when querying data. Correcting spelling ensures better data matching. We perform spelling checking by using Spelling button under Review tab. For example, replacing ‘American Expres’ into ‘American Express’ while leaving the ‘jcb’.
* Empty rows: It is important to delete empty rows because:
  + It causes issues in data analysis and interpretation. It affects the total numbers of data.
  + It affects the representation of data and lead to misleading visuals. Removing empty rows ensures more accurate and meaningful visualizations.
  + It complicates data manipulation tasks such as filtering or sorting. Removing them simplifies these tasks and improve data handling. In some tools or programming environments, empty rows can lead to errors or unexpected behavior.

There are 2 ways to delete the empty rows:

1. By selecting the dataset and using the Filter option under the Data tab and manually deleting them.
2. Or by using the “Find & Select” menu in Home tab and click “Go to Special”.

* Empty Column: It is common to find empty column if we have a large data set. Reducing the column will make it easier to navigate the spreadsheet and speed up the performance of Excel.
* Duplicated rows of data: The problems that will arise if duplicated data is not removed:

1. It could skew the data analysis as some is over represented. Removing the duplicates ensures no extra weight is given to the same information. Therefore, we need to ensure that each data point is counted only once. This process improves data accuracy.

Moreover, we could prevent biased results. The redundancy may influence statistical calculation or machine learning algorithms.

1. It may contain slight variation of error, for example:

Customer: John Doe, Address:123 First St.

Customer: John Due, Address:123 First Street.

By eliminating duplicates, we maintain data consistency

1. It improves efficiency as duplicated data will consume storage space and processing time during analysis.

**Dealing with Inconsistencies in Data**

Data inconsistencies could happen in different ways, including:

1. Human errors, such as misspelled names, incorrect dates, or numeric typos.
2. Data format differences, such as date formats, currency symbols, units of measurement.
3. The use of different abbreviation, acronyms, or shorthand notations. For example, “USA”, “U.S.” and “United States” refer to the same country.
4. Data migration from one system to another.

There are several ways to deal with inconsistencies in data as bellows:

* Changing the case of text: by using function like **UPPER, LOWER, and PROPER** to modify text case
* Formatting Dates: Adjust date format for consistency.
* Removing extra whitespace, reducing the to single spaces or excess of space by using the **TRIM** function.

**Additional Excel Features for Cleaning Data:**

* Flash Fill

Flash Fill could help us entering data that fits specific pattern such as Days of the Week, Months of the Year. It could also be useful as data cleaning tools.

We could combine two columns into one column such as name and addresses. It also can split one column of data into two.

* Text to Column

For example, it can split a column of a contact full name into two separate columns for the forename and surname.

The dataset comes from the following source:

<https://dataplatform.cloud.ibm.com/exchange/public/entry/view/f8ccaf607372882403a37d9019b3abf4>.

This dataset is published by **IBM**, and includes fictitious customer demographics and sales data.