Module 3: Collecting and Cleaning Data

Mission 3.1 - Data Sources

Data comes from multiple sources, and data analysts often gather data from multiple sources and combine these for data analysis. The process of gathering data from these sources and presenting it is called **data consolidation**. Data consolidation is a crucial step, as the accuracy of the insights from your data analysis depends heavily on the quality of data used.

Task

Download the data sheets below and consolidate them on Google Sheets.

**Note that some of the data provided has been manipulated for the purposes of this course*

- Data Sheet Udemy Courses - Business Courses.csv 218.5KB
- Data Sheet Udemy Courses - Design Courses.csv 109.0KB
- Data Sheet Udemy Courses - Music Courses.csv 126.9KB
- Data Sheet Udemy Courses - Web Development.csv 231.8KB

Uploading to Google Sheets

- Open Google Sheets
- Open or create a sheet and click File > Import. Import the first file above as 'Replace spreadsheet.'
- For following rows select the 'Append rows to current sheet' option when importing.

You will see that not all of the data is consistent or clean. We will be working on this in the next module.

Mission 3.2 - Data Cleaning

Good data is essential when using data to derive insights and make business decisions.

Garbage in, garbage out, is a concept common to computer science and mathematics that can be applied in Data Analysis - the quality of output (your insights generated) are determined by the quality of the input.

This is where Data cleaning comes in - Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted.

Data cleaning plays an important role in the analytical process and making sure that the answers we uncover are reliable and of a high quality.

We will be going through a few functions to clean our data in excel.

Task

Use the below functions to clean your data.

1) Remove duplicates

Select the entire data sheet data to remove duplicates from.

Data > Remove Duplicates

2) Removing blank cells

- Select the entire sheet or dataset and go to **Data > Create a filter**
- Click on the Filter icon at the top of any column, then click on Filter by condition and select 'is empty'.
- Blank cells will arise to the top of the sheet and can be removed.

3) Headers

Ensure you have clear and concise names for headers and use dashes or underscores in between words to make it easier to parse later on.

4) Find and replace

If you examine the data, you will see that the Web Development subject title is not the same as other subject titles. Use the Find and Replace function to make the Web Development subject consistent with other subjects.

Edit > Find and Replace

Submission

**You may be redirected to a separate Google Form Page, and will be required to sign into a Google account or create a new one.*

**Make sure to keep a local file for your Portfolio*

https://docs.google.com/forms/d/e/1FAIpQLSeevPfrP6OHuYneTCdWmjd2XLi3zElnFmXcE3CvBYa7klwwnQ/viewform