

Assignment 1 instructions

Task: In this assignment, you will collect, clean, and process large datasets using an ETL (Extract, Transform, Load) process. This means you'll gather data, clean it up, change it to the right format, and then store it for later use.

Data Collection - Gather data from different sources (APIs, databases, files, or websites). Data must be at least 20k records.

Data Transformation:

Data Cleaning: Removing duplicates, handling missing values, and correcting data types.

Data Aggregation: Summing, averaging, or grouping data as required.

Data Formatting: Standardizing date formats, converting currencies, and creating new derived columns.

Filtering: Selecting rows that meet certain criteria (e.g., date ranges, geographical regions).

Data Loading - Save the transformed data into a database by using MongoDB. Check for any errors while loading the data. **Write at least 3 SELECT's from your databases.**

Defense: You have to do the presentation, where you will show how your code is working. You have to demonstrate your code. At the same time, you have to explain the methods which you have done.

Result: .ipynb/.html files with analyzed data.