What is data wrangling?

Data wrangling is the process of cleaning, organizing, and transforming raw data into a usable format for analysis or other tasks. The process includes connecting to data sources, reformatting the information so it’s consistent, removing duplicates, and filtering out unnecessary noise

What are the 6 main tasks in data wrangling? What happens in each stage?

* Data discovery
  + Understand the data by exploring its structure, types, and contents to identify patterns
* Data structuring
  + Organising the data into a usable format (tables, rows, columns)
  + Potentially combining datasets for consistency
* Data cleaning
  + Fixing errors
  + Handle missing/inconsistent values
  + Remove duplicate/irrelevant entries
* Data enrichment
  + Add new information by merging datasets or creating new columns/rows to improve the dataset
* Data validation
  + Making sure the data is complete, accurate, and consistent
* Data exporting
  + Save/export the new dataset so that it can be used for analysis

What skills are required for data wrangling?

* Technical skills:
  + Programming is important as it allows for automated data manipulation
  + Must be able to work with different data types/formats (e.g., CSV, JSON, XML)
  + Must be able to understand how databases are structured and how to query them
* Analytical skills:
  + Problem solving to identify/fix problems with the datasets (e.g., missing entries)
  + Attention to detail to ensure accuracy during cleaning and transformation
  + Logical thinking to structure the data in a usable format
* Data knowledge:
  + Must be able to identify issues such as missing values or duplicates
  + Domain understanding – knowing the context of the data to make informed decisions
* Communication skills:
  + Clearly document cleaning steps and changes made
  + Work well in teams to handle shared datasets