Case Study – Working Student Data Engineering

Extract and Prepare Data from a Public Source

Task overview

- Your task is to identify a publicly accessible dataset, extract the data using Python, clean and transform it, and deliver a structured output.
- You are free to choose any suitable public source that provides company- or industry-related data.

Data extraction

- Use Python to extract the data from your chosen source.
- You can choose any extraction method you consider appropriate.

Data requirements

- Prepare a structured dataset that includes the following fields, where available:
  - Company name: Name of the company
  - Country: country in which the company has its main headquarters
  - Industry: Industry classification based on your source
  - Year(s): Year(s) associated with the financial value; please extract the most recent 3 years of the companies' financials, if available
  - Revenue: Revenue figure
  - Revenue unit: Unit or currency of the revenue
  - *(Optional: Add 3–5 additional KPIs of your choice in the same manner as for Revenue.)*
- Please ensure that the dataset contains at least 100 companies and no more than 500 companies.

Submission

- Share the GitHub repository link via email and ensure that it is publicly accessible, or
- Compress the output (e.g. ZIP folder) and share it via email to me (thu-huyen-my.nguyen@statista.com)
- Timeframe: 1 week

Note

If you cannot finish the entire task, it is completely fine to submit only parts of it. You can verbally explain where you struggled and why you omitted certain parts.

We are very excited to see your results.

Good luck!

statista