Case Study: Implementing data quality checks

Short description:

Ensure the accuracy and consistency of company financial data extracted from various sources by implementing automated quality checks. Furthermore, we ask you to prepare a PowerPoint presentation.

Detailed tasks description:

Quality checks of extracted company data with Python

The attached company data is extracted from multiple sources, such as APIs or web scraping, and is used for further modeling or publishing on the Statista platform. However, errors can occur due to incorrect extraction, missing values, inconsistencies, outliers etc.

Your task is to design and implement a set of automated plausibility checks to detect and flag potential data issues before storing the data in the database.

- Come up with at max. 5 quality measures (e.g. consistency, correctness…) and implement the belonging automated checks in a Python-based project
- Input: attached excel file
- Output: excel file plus additional columns flagging issues
- For at least 1 quality measure, use prompt engineering with an LLM (e.g., OpenAI's GPT) to assist in identifying inconsistencies, wrong extraction or anomalies in the data

PowerPoint presentation

- Please prepare a PowerPoint presentation of max. 5 slides in which you explain:
  - A summary / visual presentation of your code
  - Why did you choose these 5 quality measures and what other quality measures did you have in mind?
  - What was the main challenge of this task?

Submission:

- Share the GitHub repository link (ensure it is publicly accessible), as well as your PowerPoint presentation via email to us.
- Our emails: oxana.iakovleva@statista.com and godsfavour.ikwuka@statista.com
- Deadline: 24 hours prior to the interview

Note:

- If you cannot finish the whole task, it is totally fine to submit only parts of it. You can verbally tell us where you struggled and why you left out parts of the task.

We are very excited to see your results.

Good luck!

statista