PwC Shanghai AC—-Analytics & Insights

Take Home Test – Pandas

horizontal line

# Background：

You start working on a project that includes a financial business. The data engineer of this project regularly gets some financial data of the company from an excel file. Then you import the data into the database through the Data Pipeline.

Soon, there were some anomalies in the project's data, from which we selected some suspected anomalies and asked you to identify the parts that went wrong.

Here is an Excel file that contains two Sheets, Example\_Data Sheet and the Example\_DB Sheet. Example\_Data is the original file, and Example\_DB is the data in the database corresponding to that file.

Please observe this file, the Example\_Data Sheet includes Company Info information and 4 Company Metric/Value information, and the database (the Example\_DB Sheet) stores the 4 Metric information directly into the Metric Column.

**Example\_Data Sheet contains the following data cleanup logic:**

* Company ID and Company Name should be corresponding
  + if there is one ID corresponding to multiple Names or one Name corresponding to multiple IDs, the ID/Name is invalid and needs to be cleaned up.
* Fiscal Year is the year, between 1999 and 2021.
* SIC Code is a four-digit number.
* Trading Currency can only be one of two units, USD or GBP.
* SP / CDS / APD / ARD / ADA is an integer number.

**Example\_Data.xlsx:**

* [Download Link 1: Githubusercontent](https://raw.githubusercontent.com/czwstc/exampledata/main/Example_Data.xlsx)
* [Download Link 2: BaiduNetDisk](https://pan.baidu.com/s/1oheraePk6Nunna1CLoyDTA?pwd=yh5d)

# Task：

## **(Data Cleaning) You need to complete the cleaning of the Example\_Data Sheet data according to the cleaning rules**

## **(Data Processing) You need to refer to the format of the Example\_DB Sheet to complete the processing of the Example\_Data Sheet data.**

## **(Data comparison) After completing the data cleaning, refer to the Example Answer Sheet and find the problematic data in the Example\_DB Sheet and Example\_Data Sheet.（Print it out as a Dataframe.）**

## **Commit your code to your repo on Github(.py or .ipynb only)**

# NOTES

* **Minimal use of pandas loop operations (iterrows()) for efficiency.**
* **Please avoid having questions, data sets, company names, personal information, and any information other than interview-related answers in your repo.**
* **You can send us the link to your repo once you have completed the tasks.**
* **Please clean up your Repo after completing the interview process.**

# 

# Background：

你开始参与了一个包含金融业务的项目，这个项目的数据工程师会定期从excel文件中取得一些公司的财务数据。然后通过Data Pipeline将数据导入数据库。

很快，项目上有一些数据出现了一些异常，从中挑选了部分疑似异常的数据，来判断一下出现的问题。

这里有一份Excel文件，里面包含两个Sheet, Example\_Data 和 对应的的Example\_DB，其中Example\_Data是某一个原始文件。而Example\_DB是该文件对应的数据库中的数据。

请观察文件，Example\_Data Sheet中包括Company Info 信息和4个Company Metric/Value信息，数据库(的Example\_DB Sheet)中存储的时候直接将4个Metric信息合并到了Metric Column中。

**Example\_Data Sheet包含以下数据清理逻辑：**

* Company ID 和 Company Name 应该是对应的，如果出现一个ID 对应多个Name或者一个Name对应多个ID，该条ID/Name 失效，需要清理。
* Fiscal Year 是年份，在1999-2021之间。
* SIC Code 是四位数字
* Trading Currency 只能是 USD或者GBP两个单位之一。
* SP / CDS / APD / ARD / ADA 是整数数字。

**Example\_Data.xlsx:**

* [Download Link 1: Githubusercontent](https://raw.githubusercontent.com/czwstc/exampledata/main/Example_Data.xlsx)
* [Download Link 2: BaiduNetDisk](https://pan.baidu.com/s/1oheraePk6Nunna1CLoyDTA?pwd=yh5d)

# Task：

## （数据清洗）你需要依据清洗规则，完成Example\_Data Sheet数据的清洗

## （数据处理）你需要参考Example\_DB Sheet 的格式完成Example\_Data Sheet 数据的处理。

## （数据对比）完成数据清洗之后，参考Example Answer Sheet的样子，找到Example\_DB Sheet 和 Example\_Data Sheet 中有问题的数据。以Dataframe的形式打印出来。

## 创建一个Github repo并保存你的代码(.py or .ipynb only)

# NOTES

* **请尽可能减少使用Pandas的循环函数(iterrows())**
* **请避免在你的repo中出现题目本身、数据集、公司名字、个人信息或者任何与答案无关的信息。**
* **完成本任务之后，请将Repo的地址发送给我们**
* **请在完成面试流程之后清理你的Repo**