FINTECH SANDPIT

Data Engineer Interview Task

Background

The key to this task is to understand your approach to data analytics, and your comfort in Python data and automation tools. At Fintech Sandpit, a lot of our data analytics work involves building generators for synthetic data generation, identifying and visualising data lineage maps, automating data processing pipelines, and ethical web scraping.

There is potential that internships convert into a paid role at their completion. We find most of our engineers and data scientists this way.

Brief

Fintech Sandpit is a B2B Fintech Innovation Ecosystem. We provide industry innovation infrastructure to public sector institutions, communities, and financial institutions. Our primary offering is a Test-Environment-as-a-Service with a marketplace of pre-integrated Tech partners and Datasets.

The purpose of the Data Supermarket module is to help financial institutions find datasets to innovate with that solve their business problem. We do this by providing them categorised synthetic datasets on a single API along with a pre-integrated tech companies so they can access the tech and data they need to innovate all with one API key.

Our application is currently facilitating a global collaboration between the UK regulator and 800+ industry participants from tier-1 banks and data tech companies.

Task

A. **Source Data**
    a. Research the web and source 3 datasets of banking transactions - you can use any data source
    b. Download or scrape the data
    c. List the datasets and briefly describe how you found them and how you scraped them if you had to scrape them

B. **Exploratory Analysis**
    a. Conduct an Exploratory Data Analysis on the banking transaction dataset
    b. Briefly describe 5 key educational insights about that chosen dataset

C. Manipulating the Dataset

   a. Create a class that accepts the dataset file path as a parameter in the initialisation of the class.

   b. Implement the following functions:

      i. stdDev(columnName: string) → number // returns the standard deviation of the column

      ii. mean(columnName: string) → number // returns the mean of the column

      iii. outliers() → number[] // returns a list of rows that are outliers in the dataset

1. Questions: please submit your response with answers to these questions:

   a. What (if any) experience do you have in Cloud Services (eg. AWS, GCP)?

   b. What is your experience with building machine learning models? What packages are you familiar with?

2. Please submit your Jupyter notebook in PDF format to contact@fintechsandpit.com and **let us know how long the task took you**.