Data Analyst Assignment Brief

We understand that your time is valuable and want to thank you for working on this exercise.

A major aspect of the role is making data available for analysis by yourself, the rest of the data team and product managers. In this exercise you will be provided a credit risk dataset and work to make it available for others to analyze.

To help future analysis, you will also get familiar with the data and share some very basic insights about the dataset.

Please note that you entirely own anything you build as part of this assignment in perpetuity.

The Assignment

Please approach both parts of the exercise as if you are working in a real production environment.
Do not hesitate to ask any questions about the exercise, and please document any assumptions and decisions you make, and any shortcuts you take (for example, due to time constraints) in the process.

Please share the code, and documents you create or are going to present as part of this project with your interviewers. Hosting solutions like GitHub, Gitlab, Google Drive, BitBucket, etc… are all acceptable). Please restrict public access to your submission.

We might store/archive your solution and use it internally for discussion and evaluation.

Part 1 - Data Ingestion

The first part of this exercise requires you to build a PoC for a data ingestion system to make incoming CSV data easy to use / query. You can find the data attached:

- data_dictionary.pdf - A variable dictionary is provided with definitions for each variable.
- sample_data.csv - The dataset is provided to you in the form of a CSV file.

You have the freedom to choose how you go about building the PoC but here are a few guidelines:

- You should expect to receive files with data (assume the same format) so your solution should be able to ingest them on a regular basis (e.g. every hour or day).
- The data should be stored in a central place and accessible/readable by multiple data analysts (even in parallel).
- It is up to you to choose the underlying data storage/compute engine/database you use, but the data should also be accessible via SQL, Python and/or R.

Part 2 - Understanding the Data

A big part of our work is to provide insights and self serve dashboards for the business.
So it’s important to understand the data and build structures that make it easy to analyse it.

For the 2nd part of this exercise we’d like for you to give a short presentation (10-15 minutes) describing the data in a way that would be relevant for other data analysts and product managers.

A few guidelines for this part:

- Assume that the audeience is completely unfamiliar with the new data and haven’t heard of it before
- Please take this as far as you’d like, but note that you are not expected to train a Machine Learning model or come up with a credit policy.
- You can use whatever tools you prefer for this.
- Beautiful visualizations are great but descriptive summary tables are also great.
- Your exploration of the data does not have to be strictly about credit risk or even finance. If you find something interesting, we’d like to know about it.