ReadMe: DataHub and Data Explorer Evaluation Report (Databricks)

Overview

This script is designed to create a report that compares evaluation data between two sources: DataHub and Data Explorer. It merges data from both sources based on common indices and checks for discrepancies in various fields, such as project outcome ratings, quality at entry, quality at supervision, overall bank performance, and more. The resulting output is a CSV report that summarizes the evaluations and any inconsistencies found.

Required Libraries

The following libraries are required to run this script:

pandas: For data manipulation and handling Excel and CSV files.

datetime: For generating timestamped filenames.

Ensure these libraries are installed in your Databricks environment before running the script.

Running the Script:

1. Databricks Workspace Setup

If you do not have a Databricks account, create one.

Log in to your Databricks workspace.

2. Install Required Libraries

To ensure all dependencies are met, run the following command in your Databricks notebook:

%pip install pandas

3. Prepare the Data

* Place the required Excel files (IEG Data Hub - IEG Ratings.xlsx and Data Explorer.xlsx) in the same directory as your Databricks notebook.
* Make sure the files contain the expected data with the correct sheet names and column structures.

4. Execute the Script

* After preparing the data and ensuring the necessary libraries are installed, run the script in your notebook.
* Monitor the output for any error messages. Databricks provides detailed logs to aid in troubleshooting.

5. Access the Output

* Once the script completes successfully, a CSV report will be generated with a timestamped filename.
* You can download the CSV file for further analysis or work with the data within Databricks.

Script Functionality

* The script imports data from the DataHub and Data Explorer Excel files, focusing on specific columns like Project ID, Exit FY, Outcome Rating, and others.
* It merges the data on common indices (Project ID, Evaluation Date, and Evaluation Type) and checks for discrepancies in various ratings and metrics.
* If any discrepancies are found, the script generates a message indicating the differences and adds it to the report.
* The script also accounts for cases where some evaluations might not be present in one of the sources.

Output Report

* The output report contains the following information:
* Project ID: Unique identifier for the project.
* Evaluation Date: Date of the evaluation.
* Evaluation Type: Type of evaluation.
* Latest Eval, ICRR, PPAR: Flags indicating if these evaluations are the latest.
* Test Results: A summary of any discrepancies found.
* Instrument Type: The type of instrument used for the project.
* Various metrics for both DataHub and Data Explorer, such as Outcome, Quality at Entry, Quality at Supervision, Overall Bank Performance, Overall Borrower Performance, ME Quality, and Exit FY.

Troubleshooting

* If the script fails to find the specified Excel files, ensure they are in the same directory as your Databricks notebook.
* If there are missing values or inconsistencies, check the input Excel files to ensure they contain the expected data.
* If errors occur during execution, review the error messages to identify the source and adjust the script or data accordingly.