ReadMe: V3 (Databricks)

Overview

This script compares project-related data between multiple data sources to identify discrepancies in various ratings, such as outcomes, efficiency, quality at entry, quality at supervision, bank and borrower performance, and more. The script reads data from Excel files, processes the information, and generates a report in CSV format that summarizes the results of the comparisons.

Required Libraries

To run this script, the following Python libraries are required:

* pandas: For data manipulation and handling Excel and CSV files.
* datetime: To generate timestamps for output file names.
* 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 - ICRR Ratings.xlsx, PROJECT\_IEG\_RATING.xlsx, 1a.xlsx, 3a.xlsx, 5.xlsx, 8.xlsx, 9.xlsx, 12.xlsx) in the same directory as your Databricks notebook.
* Ensure these files have the correct sheet names and column structures.

4. Execute the Script

* Once you have prepared the data and ensured the necessary libraries are installed, run the script cells in your notebook.
* Monitor the output for any error messages. Databricks provides detailed logs to aid in troubleshooting.

5. Access the Output

* After 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 loads project data from various Excel files and creates a list of Project IDs (pid\_list).
* It initializes a DataFrame (output) with specific columns to store the results of the comparison.
* The script iterates over the pid\_list, checking for discrepancies in various ratings and attributes between the Data Explorer and DataHub sources.
* Discrepancies are noted with a message indicating the source and the nature of the difference.
* The script creates a final DataFrame with these results and exports it to a CSV file.

Output Report:

The output report contains the following columns:

* ProjectID: The unique identifier for each project.
* Test Results: A summary of any discrepancies found.
* Inst Type: The type of instrument used for the project.
* Various metrics from both Data Explorer and DataHub, such as Outcome, Relevance of Objective, Relevance of Design, Efficiency, Risk to Development Outcome (RDO), Bank Performance, Quality at Entry (QaE), Quality at Supervision (QoS), Borrower Performance, Implementing Agency Performance, Borrower Compliance, and ICR Quality.

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.