ReadMe: DPF\_Script (Databricks)

Overview

This script is designed to extract project data from multiple Excel files, process the information, and generate a detailed report that includes various ratings and metrics for each project. The script iterates over a list of Project IDs, retrieves specific data from different Excel files, and compares certain values to determine if a project is part of a Development Policy Financing (DPF) series, and if so, whether it is a "First," "Intermediate," or "Last" member. The final output is saved in a CSV file for further analysis.

Required Libraries

This script requires the following Python libraries:

pandas: For data manipulation and Excel/CSV file handling.

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

* Ensure the Excel files containing project data are in the same directory as your Databricks notebook or in an accessible DataHub.
* Update the script with the correct paths to the Excel files.

4. Execute the Script

* Once you have adjusted the paths 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, the output CSV file will be saved with a timestamped name.
* You can download the CSV file for further analysis or work with the data directly in Databricks.
* Understanding the Script
* The script processes a list of Project IDs and retrieves various ratings and information from multiple Excel files.
* It checks if a Project ID is part of a DPF series and determines its type (e.g., "First," "Intermediate," or "Last").
* It compares the ratings and metrics of the Project ID with its parent project, if applicable, and generates a summary of the comparisons.
* The final output DataFrame contains various fields related to project ratings and metrics, which is then exported to a CSV file.

Output Fields:

The output report contains the following fields:

* Project Id: The unique identifier for each project.
* Test Results: A summary of the comparison results.
* Prog DPF Member Type: Indicates whether the project is part of a DPF series.
* Series First Project Id: The Project ID of the first project in the DPF series.
* Various ratings and metrics, including Outcome, Relevance of Objectives, Efficiency, Risk to Development Outcome (RDO), Bank Performance, Quality at Entry (QaE), Quality at Supervision (QoS), Borrower Performance, Government Performance, Implementing Agency Performance, Monitoring & Evaluation (M&E) Quality, and ICR Quality.
* Troubleshooting
* If the script fails to load any Excel file, ensure the paths are correct and the files are accessible.
* 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 cause and make necessary adjustments to the script or data.