ASSIGNMENT: BANK STATEMENT ANALYSIS PROOF-OF-CONCEPT

CPC ANALYTICS| CONFIDENTIAL

Congratulations on moving forward in our hiring process. This assignment is based on a real-world case: analyzing financial data from various bank statements to identify key patterns and transactions.

Your task is to build a robust data extraction and analysis tool. We have provided two sample PDF bank statements from different banks. Your solution will be evaluated on its ability to correctly process these samples **and** four other unseen statements from the same banks.

We are most interested in the resilience of your code, your approach to handling messy data, and how you present your findings.

The Challenge

Financial investigators receive bank statements as PDFs from numerous banks, each with a unique layout. Manually extracting and analyzing this data is slow and error-prone. Your goal is to automate this process.

Part 1: The Dataset

You are provided with two sample bank statement PDF files:

1. HDFC – Click here
2. ICICI – Click here

Part 2: The Deliverables

Please complete the following tasks using Python. You may use any libraries/APIs you want to.

Task 1: Robust PDF Data Extraction

Write a script that can:

1. **Ingest a PDF bank statement from one of these banks** as an input.

2. **Automatically identify and extract the main transaction table**
3. **Parse the data into these structured formats**

   A single CSV file for account information:

   → account_number

   → account_holder_name

   → account_type (Savings or Current)

   → ifsc

   → micr

   → bank_name

   → address

   A single CSV file for transactions with the following standardized columns:

   → transaction_date

   → description

   → withdrawal_amount

   → deposit_amount

   → balance

4. **Demonstrate** that your script works on **both** of the provided PDF samples.

Task 2: Transaction Analysis & Flagging

Using the structured data from Task 1, write functions to analyze the transactions.

Implement functions to flag and count transactions that meet **at least one** of the following criteria:

- **DD Large Withdrawal:** Identify all transactions that are done by DD, and above INR 10,000
- **RTGS large Deposits:** Flag any deposit that's above INR 50,000 and done by RTGS

- **Transactions by specific entities:** Flag all transactions that appear to be done with any entity named “Guddu” or “ Prabhat” or “Arif” or “Coal India” (These names are not case sensitive)

Task 3: Summary Report and Visualization (Max 3 pages, font 11, Arial or Times New Roman only)

1. **Prepare a Brief Report:** Summarize your work in a short PDF file. Mention your name and email clearly in the PDF.
    - Explain your methodology for extracting data from the PDFs
    - You can include challenges you faced, if any
2. **Create a Visualization:** Generate one clear visualization for the ICICI Bank statement given to you, showing a **timeline plot** showing withdrawals and deposits over time.

Submission & Evaluation

- **Submission:** A single .zip file containing your source code (.py), a requirements.txt file, and your final report.
- **Evaluation:** We will evaluate your submission by testing on other bank statements from the same banks, that have a similar format.
    - **Accuracy & Robustness:** How well does your PDF extractor work on the provided files and our internal test files?
    - **Code Quality:** Is your code clean, modular, and well-commented?
    - **Analytical Thinking:** How did you approach the analysis and flagging task?
    - **Communication:** How clearly does your report explain your work and findings? How are your visualizations?

Submissions to be emailed to judy@cpc-analytics.com latest by **28th September 2025, 17:00 IST**. Evaluations would be done on a rolling basis, therefore earlier submissions are appreciated.

Good luck! We look forward to seeing your solution.