Methodology for Wallet Credit Scoring Analysis

1. Data Collection and Preparation

The first step in the methodology involves collecting and preparing the data for analysis. The dataset used consists of transaction data for various wallets interacting with a decentralized finance platform.

- **Data Sources**: The data comes from wallet transactions within the platform. Each transaction record contains the following key columns:
  - account_id: Unique identifier for the wallet.
  - amountUSD: The USD equivalent value of the transaction.
  - timestamp: The timestamp of the transaction in UNIX format.
  - asset_symbol: The type of asset (e.g., ETH, BTC) involved in the transaction.
- **Data Preprocessing**:
  - **Handling Missing Data**: Missing values in the dataset (such as missing wallet activity or asset types) are handled by filling them with default values or dropping rows, depending on the severity of the missing data.
  - **Datetime Conversion**: The timestamp column is converted from UNIX format to a readable datetime format using pd.to_datetime() in pandas to enable easier analysis.
- **Normalization**: All amounts are normalized to USD using the amountUSD column, which provides a consistent currency value for analysis. This ensures consistency when comparing wallets with different asset holdings.

2. Credit Scoring Calculation

The next step involves calculating a **credit score** for each wallet based on its transaction history. The scoring model considers multiple factors:

- **Total Deposit Amount**: This is the total USD value of funds deposited by a wallet over its history. Wallets with larger deposits are considered more reliable, as they show consistent participation in the platform.
- **Transaction Frequency**: The number of transactions performed by a wallet is an indicator of its activity level. Wallets with more frequent transactions may indicate active and engaged users.
- **Unique Assets**: The diversity of assets held by a wallet is another factor. A wallet with holdings in various assets demonstrates diversified activity, which may indicate a higher level of engagement and creditworthiness.

The credit score is calculated by assigning weights to these factors. For simplicity, we use the following general weights:

- **Total Deposit**: 40%
- **Transaction Count**: 30%
- **Unique Assets**: 30%

The overall credit score for each wallet is calculated as a weighted sum of these factors, scaled to a value between 0 and 100. For example:

Credit Score=0.4×Total Deposit Score+0.3×Transaction Count Score+0.3×Unique Assets Score\text{Credit Score} = 0.4 \times \text{Total Deposit Score} + 0.3 \times \text{Transaction Count Score} + 0.3 \times \text{Unique Assets Score}Credit Score=0.4×Total Deposit Score+0.3×Transaction Count Score+0.3×Unique Assets Score

Each factor is normalized on a scale of 0 to 100 before being multiplied by the respective weight.

3. Wallet Segmentation

Once the credit scores are calculated for all wallets, the wallets are divided into two main groups:

- **High-Scoring Wallets**: Wallets with the top N highest credit scores. These wallets demonstrate consistent and active participation in the platform and are considered to have a higher creditworthiness.
- **Low-Scoring Wallets**: Wallets with the bottom N lowest credit scores. These wallets may indicate minimal activity or low engagement with the platform and could be at higher risk.

The number of wallets to include in each group is specified by the num_wallets parameter, which is typically set to 5.

4. Analysis Report Generation

For both high-scoring and low-scoring wallets, a detailed analysis is generated, summarizing their transaction history and providing observations about their activity:

- **High-Scoring Wallets**:
  - **Total USD**: The sum of all USD transactions for the wallet.
  - **Transaction Count**: The number of transactions made by the wallet.
  - **First Transaction Date**: The timestamp of the first transaction.
  - **Last Transaction Date**: The timestamp of the most recent transaction.
  - **Unique Assets**: The number of distinct assets the wallet has interacted with.
- **Observation**: These wallets typically have a high deposit amount, frequent transactions, and diversified holdings, which contribute to their high credit score.
- **Low-Scoring Wallets**:
  - **Total USD**: The sum of all USD transactions for the wallet.
  - **Transaction Count**: The number of transactions made by the wallet.
  - **First Transaction Date**: The timestamp of the first transaction.
  - **Last Transaction Date**: The timestamp of the most recent transaction.
  - **Unique Assets**: The number of distinct assets the wallet has interacted with.
- **Observation**: Low-scoring wallets are characterized by low deposit amounts, fewer transactions, and minimal asset diversification. This behavior often results in lower credit scores.

Each wallet's analysis is accompanied by observations and justifications for its high or low credit score.

5. HTML and PDF Report Generation

After the wallet analysis is performed, the results are compiled into an HTML report:

- **HTML Structure**: The HTML content is divided into sections for high-scoring and low-scoring wallets. For each wallet, key transaction statistics are presented in an unordered list format.
- **PDF Conversion**: The HTML report is converted into a PDF file using pdfkit, which utilizes wkhtmltopdf to ensure proper formatting and styling during conversion.
- **Download**: Once the PDF report is generated, it is made available for download to the user via the files.download() function in Google Colab.

6. Final Output

The output of this methodology is a downloadable PDF report containing:

- A detailed analysis of the top N high-scoring wallets.
- A detailed analysis of the bottom N low-scoring wallets.
- Observations and justifications for each wallet’s credit score.

7. Potential Improvements

Future versions of this model could integrate additional data points, such as:

- **Wallet Interactions**: Analyzing interactions with smart contracts, decentralized applications, or lending/borrowing behavior.
- **Temporal Analysis**: Examining the time-based trends of wallet activity, such as periodic fluctuations in deposits or withdrawals.
- **External Risk Factors**: Integrating data from external sources to assess the overall risk level of a wallet, such as wallet address reputation or association with known malicious entities.

By expanding the scoring criteria and incorporating new features, the system can provide even more accurate assessments of wallet creditworthiness.