PAN Number Validation Project

Objective

The objective of this project is to clean and validate a dataset containing the Permanent Account Numbers (PAN) of Indian nationals. The goal is to ensure that each PAN number adheres to the official format and is categorised as either Valid or Invalid.

1. Data Cleaning and Preprocessing

- Identify and handle missing data (remove or impute).
- Check for duplicates and remove them.
- Handle leading/trailing spaces.
- Ensure all PAN numbers are in uppercase.

2. PAN Format Validation

A valid PAN must satisfy the following rules:

- It is exactly 10 characters long.
- Format: AAAAA1234A
  - First 5 characters should be alphabets (A–Z).
    - Adjacent characters cannot be the same (e.g., AABCD is invalid).
    - All five characters cannot form a sequence (e.g., ABCDE is invalid).
  - Next 4 characters should be digits (0–9).
    - Adjacent digits cannot be the same (e.g., 1123 is invalid).
    - All four digits cannot form a sequence (e.g., 1234 is invalid).
  - Last character should be an alphabet (A–Z).

Example of a valid PAN: AHGVE1276F
Example of an invalid PAN: ABCDE1234F

3. Categorisation

- Valid PAN: Matches the format and rules.
- Invalid PAN: Fails to meet format, length, or sequence rules.

4. Summary Report

The summary report must include:

- Total records processed
- Total valid PANs
- Total invalid PANs
- Total missing or incomplete PAN