PAN Number Validation Project

Data Cleaning and Validation

Objective

You are tasked with cleaning and validating 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. The dataset is given in a separate Excel file.

PAN Number Validation Dataset.xlsx

Instructions

1. Data Cleaning and Preprocessing:
    - Identify and handle missing data: PAN numbers may have missing values. These missing values need to be handled appropriately, either by removing rows or imputing values (depending on the context).
    - Check for duplicates: Ensure there are no duplicate PAN numbers. If duplicates exist, remove them.
    - Handle leading/trailing spaces: PAN numbers may have extra spaces before or after the actual number. Remove any such spaces.
    - Correct letter case: Ensure that the PAN numbers are in uppercase letters (if any lowercase letters are present).
2. PAN Format Validation: A valid PAN number follows the format:
    - It is exactly 10 characters long.
    - The format is as follows: AAAAA1234A
        - The first five characters should be alphabetic (uppercase letters).

        1. Adjacent characters(alphabets) cannot be the same (like AABCD is invalid; AXBCD is valid)
        2. All five characters cannot form a sequence (like: ABCDE, BCDEF is invalid; ABCDX is valid)
    - The next four characters should be numeric (digits).
        1. Adjacent characters(digits) cannot be the same (like 1123 is invalid; 1923 is valid)
        2. All four characters cannot form a sequence (like: 1234, 2345)
    - The last character should be alphabetic (uppercase letter).

  Example of a valid PAN: AHGVE1276F

3. Categorisation:
    - Valid PAN: If the PAN number matches the above format.
    - Invalid PAN: If the PAN number does not match the correct format, is incomplete, or contains any non-alphanumeric characters.
4. Tasks:
    - Validate the PAN numbers based on the format mentioned above.
    - Create two separate categories:
        - Valid PAN
        - Invalid PAN
    - Create a summary report that provides the following:
        - Total records processed
        - Total valid PANs
        - Total invalid PANs
        - Total missing or incomplete PANs (if applicable)

Note:

- Feel free to use either SQL or Python to complete this data cleaning and validation project.