Cleaning BANK MARKETING CAMPAIGN DATA

Overview

Banks generate substantial revenue from personal loans, with the average interest rate for a two-year loan in the UK sitting around 10%. While this may seem modest, UK consumers borrowed approximately £1.5 billion in September 2022 alone source, potentially generating around £300 million in interest for banks over two years.

Problem Statement

A bank has marketing campaign data stored in a CSV file called **"bank_marketing.csv"**. This data needs to be cleaned, formatted, and split into three separate CSV files following a specific structure suitable for import into a PostgreSQL database.

Goals

01

Clean the Data
Address inconsistencies and missing values in the data.

02

Format the Data
Convert data types and create new columns based on existing ones (e.g., last_contact_date).

03

Structure the Data
Split the original data into three distinct CSV files, each with specific columns and data types, aligned with the intended database tables.

Dataset

The dataset is a CSV file named **"bank_marketing.csv"** containing marketing campaign data for the bank. It includes information about clients (age, job, marital status, etc.) and details of the marketing campaign itself (number of contacts, outcomes, etc.). The source of this data is likely the bank's internal marketing campaign management system.

Data Findings (Implicit in the Code)

- The code identifies potential issues in several columns:
  - Dots (".") in "job" and "education" columns.
  - Missing values ("unknown") in the "education" column.
  - Inconsistent values ("yes", "no", "unknown") in "credit_default", "mortgage", "previous_outcome", and "campaign_outcome" columns.
  - Missing date information for the "last_contact" event.

- The code addresses these issues through cleaning and transformation steps.

Outputs

```
credit_default
--------------
no         32588
unknown     8597
yes            3
Name: credit_default, dtype: int64
mortgage
--------------
yes        21576
no         18622
unknown      990
Name: mortgage, dtype: int64
previous_outcome
--------------
nonexistent    35563
failure         4252
success         1373
Name: previous_outcome, dtype: int64
campaign_outcome
--------------
no     36548
yes     4640
Name: campaign_outcome, dtype: int64
```

The provided image shows the value counts for four columns in the **bank_marketing.csv** dataset: **credit_default**, **mortgage**, **previous_outcome**, and **campaign_outcome**. These columns represent categorical variables that indicate certain characteristics or outcomes related to the bank's clients and marketing campaigns.

Here's a breakdown of each column:

- **credit_default:**
  - This column indicates whether the client has defaulted on their credit.
  - The majority of clients have not defaulted (no), followed by a significant number of unknown cases. Only a very small number of clients have defaulted (yes).
- **mortgage:**
  - This column indicates whether the client has a mortgage.
  - The majority of clients have a mortgage (yes), followed by those who don't (no). A small number of cases are unknown.
- **previous_outcome:**
  - This column indicates the outcome of the previous marketing campaign for the client.
  - The most common outcome is "nonexistent," meaning there was no previous campaign.
  - A significant number of previous campaigns resulted in "failure," while a smaller number were "successes."
- **campaign_outcome:**
  - This column indicates the outcome of the current marketing campaign for the client.
  - The majority of campaigns were not successful (no), while a smaller number resulted in "yes," indicating success.

Data Results

- campaign.csv
- client.csv
- economics.csv

Insights

By cleaning and structuring the data, the bank can:

Improve data quality for further analysis.

Facilitate efficient storage and retrieval of data in a relational database like PostgreSQL.

Enable easier querying and manipulation of data for campaign analysis and future marketing.efforts.