Database Design and Implementation for Bank Marketing Analysis

# 1. Dataset Transformation into Relational Database Tables:

## Dataset Overview:

The dataset in focus, originating from the UCI Machine Learning Repository, encompasses comprehensive data from a Portuguese bank's direct marketing campaigns. It details interactions via phone calls with potential clients, aiming to predict the likelihood of subscribing to a term deposit.

## Methodology for Data Transformation:

The transformation process involved meticulously dissecting the dataset into a set of interrelated tables, each representing a unique aspect of the data. This involved:

* **Normalization**: Breaking down the structure of the CSV file into a series of normalized tables, thereby reducing data redundancy and enhancing query efficiency.
* **Data Alignment**: The CSV data was processed using Python and pandas to align with the relational model, ensuring each table accurately represented a specific domain such as Job, Education, Loan, etc.
* **Schema Design and Table Creation**: Utilizing SQLAlchemy, I crafted a schema that mirrored our analytical objectives, establishing a set of tables within a MySQL database, each defined with appropriate data types and keys.

# 2. Data Importation: Challenges and Resolutions:

## Column Name Alignment:

A primary challenge was ensuring the column names from the CSV file corresponded seamlessly with those in our MySQL tables. This was crucial for a smooth data import process and was achieved through careful renaming of DataFrame columns in Python to match the SQL schema.

## Managing 'Unknown' Values:

The dataset presented numerous instances of 'unknown' values across various fields. Recognizing the importance of these entries in providing a complete picture of the dataset, I opted to retain them as distinct categories, thereby preserving the authenticity of the data.

# 3. Comprehensive Data Dictionary:

## Job Table:

* **job\_id**: (INT) Unique identifier for the job category.
* **job\_title**: (VARCHAR) Descriptive title of the job (e.g., 'admin.', 'blue-collar').

## Education Table:

* **education\_id**: (INT) Unique identifier for the education level.
* **education\_level**: (VARCHAR) Description of the education level (e.g., 'basic.9y', 'university.degree').

## Loan Table:

* **loan\_id**: (INT) Unique identifier for each combination of housing and personal loan status.
* **housing**: (VARCHAR) Status of housing loan (e.g., 'yes', 'no').
* **personal\_loan**: (VARCHAR) Status of personal loan (e.g., 'yes', 'no').

## Contact Table:

* **contact\_id:** (INT) Unique identifier for each type of contact.
* **contact\_type**: (VARCHAR) Type of contact (e.g., 'cellular', 'telephone').
* **month**: (VARCHAR) Month of last contact.
* **day\_of\_week**: (VARCHAR) Day of the week of last contact.
* **duration**: (INT) Duration of the last contact in seconds.

## Campaign Table:

* **campaign\_id**: (INT) Unique identifier for each campaign attribute.
* **campaign**: (INT) Number of contacts performed during this campaign.
* **pdays**: (INT) Number of days passed after last contact from a previous campaign.
* **previous**: (INT) Number of contacts before this campaign.
* **poutcome**: (VARCHAR) Outcome of the previous marketing campaign.

## Economic Table:

* **economic\_id**: (INT) Unique identifier for economic factors.
* **emp\_var\_rate**: (FLOAT) Employment variation rate.
* **cons\_price\_idx**: (FLOAT) Consumer price index.
* **cons\_conf\_idx**: (FLOAT) Consumer confidence index.
* **euribor3m**: (FLOAT) Euribor 3 month rate.
* **nr\_employed**: (INT) Number of employees.

## Person Table:

* **person\_id**: (INT) Unique identifier for each client.
* **age**: (INT) Age of the client.
* **job\_id, education\_id, loan\_id, contact\_id, campaign\_id, economic\_id**: (INT) Links to respective tables.
* **marital\_status**: (VARCHAR) Marital status of the client.
* **subsribed**: (VARCHAR) Subscription status to a term deposit ('yes', 'no').

# 4. Targeted Business Questions for SQL-Based Analysis:

1. What is the distribution of term deposit subscriptions across different job categories and how does age factor into this?
2. How does the level of education influence a client's decision to subscribe to a term deposit?
3. What patterns emerge in term deposit subscriptions with respect to different months and contact methods?
4. Does the outcome of previous marketing campaigns correlate with the current campaign's success rate?
5. How do key economic indicators like the employment rate and consumer confidence index affect the propensity for subscribing to term deposits?
6. Is there a notable difference in subscription rates among different marital statuses?
7. What is the typical duration of successful marketing calls compared to those that do not result in a subscription?
8. How does the frequency of contact in the current campaign impact the likelihood of subscription?

# Conclusion:

This project entailed a detailed conversion of a comprehensive dataset into a structured, query-efficient database, enabling nuanced analysis of marketing strategies and client behaviors. The data dictionary serves as a foundational guide for navigating the database, while the outlined business questions provide a roadmap for extracting actionable insights through SQL queries.