# Mehtre_Term_Project

Vijaykumar Mehtre

2023-11-04

## Introduction

### Problem Statement :

### Reducing marketing resources by identifying customers who would subscribe to term deposit and thereby direct marketing efforts to them.

Bank marketing is known for its nature of developing a unique brand image, which is treated as the capital reputation of the financial academy. It is very important for a bank to develop good relationship with valued customers accompanied by innovative ideas which can be used as measures to meet their requirements.

Customers expect quality services and returns. There are good chances that the quality factor will be the sole determinant of successful banking corporations. Therefore, banks need to acknowledge the imperative of proactive Bank Marketing and Customer Relationship Management and also take systematic steps in this direction.

#### What is a Term Deposit ?

A time deposit or term deposit is a deposit in a financial institution with a specific maturity date or a period to maturity, commonly referred to as its "term". Time deposits differ from at call deposits, such as savings or checking accounts, which can be withdrawn at any time, without any notice or penalty. Deposits that require notice of withdrawal to be given are effectively time deposits, though they do not have a fixed maturity date.

A term deposit is a fixed-term investment that includes the deposit of money into an account at a financial institution. Term deposit investments usually carry short-term maturities ranging from one month to a few years and will have varying levels of required minimum deposits.

The investor must understand when buying a term deposit that they can withdraw their funds only after the term ends. In some cases, the account holder may allow the investor early termination or withdrawal if they give several days notification. Also, there will be a penalty assessed for early termination.

## Research questions

1. What is the relationship between marital status and the likelihood of subscribing to a term deposit?
2. How does the type of job held by a customer impact their decision to subscribe to a term deposit?
3. Are customers with a previous default more or less likely to subscribe to a term deposit, and how does this impact marketing strategies?

4. What is the effect of the number of times a customer was previously contacted (previous) on their likelihood to subscribe to a term deposit, and how should this influence future marketing efforts?

5. How does the duration of the marketing campaign (duration) affect the subscription rate, and can this be used to optimize future campaign strategies?

6. What are the most effective communication channels (contact) for encouraging term deposit subscriptions, and how can this knowledge inform marketing resource allocation?

## Approach

To address the problem statement of reducing marketing resources by identifying customers likely to subscribe to a term deposit in R, you can follow these general steps:

1. Data Import and Preprocessing:
    - Load the dataset into R using functions like `read.csv` or any relevant data import functions.
    - Preprocess the data, which may include handling missing values, data type conversions, and data cleaning.
2. Exploratory Data Analysis (EDA):
    - Conduct EDA to understand the dataset. Use summary statistics, visualizations (e.g., histograms, box plots, or scatter plots), and correlation analysis to gain insights into the data.
3. Feature Engineering:
    - Create new features or modify existing ones if needed. For example, you might want to convert categorical variables into numerical format using techniques like one-hot encoding.
4. Data Splitting:
    - Split the data into training and testing sets. The training set will be used to build predictive models, and the testing set will be used to evaluate their performance.
5. Model Building:
    - Choose an appropriate predictive modeling technique, such as logistic regression, decision trees, random forests, or gradient boosting.
    - Train the model using the training data.
6. Model Evaluation:
    - Assess the model's performance on the testing data using evaluation metrics like accuracy, precision, recall, F1-score, and ROC-AUC, depending on the nature of the problem (classification).
7. Feature Importance Analysis:
    - Determine the importance of each feature in predicting term deposit subscriptions. This can help identify which customer attributes are most influential.
8. Customer Segmentation:
    - Use clustering techniques like K-means or hierarchical clustering to segment customers based on their characteristics or behavior.
9. Resource Allocation:
    - Develop a strategy for allocating marketing resources based on the segmentation and predictive model results. For example, allocate more resources to segments with a higher likelihood of subscribing.

10. Ethical Considerations:
    - Consider ethical implications, especially with regards to data privacy and fairness when targeting customers.
11. Optimization:
    - Continuously monitor and optimize the marketing strategy based on the results and feedback. This may involve A/B testing and fine-tuning marketing efforts.
12. Reporting and Visualization:
    - Present your findings and strategies in a clear and concise manner using RMarkdown or other reporting tools. Visualizations, tables, and graphs can be helpful for conveying the results.

## Discuss how your proposed approach will address (fully or partially) this problem.

The proposed approach outlined earlier aims to address the problem of reducing marketing resources by identifying customers likely to subscribe to a term deposit through a systematic and data-driven process. Here's how each step in the approach helps to tackle this problem:

1. **Data Import and Preprocessing:**
    - This step ensures that the data is ready for analysis by handling missing values, converting data types, and cleaning the dataset. Clean data is essential for accurate modeling and decision-making.
2. **Exploratory Data Analysis (EDA):**
    - EDA helps in understanding the dataset, including the distribution of variables and potential patterns. This understanding can guide subsequent modeling and segmentation efforts.
3. **Feature Engineering:**
    - Creating or modifying features allows you to capture the most relevant information from the dataset, improving the model's ability to identify potential term deposit subscribers.
4. **Data Splitting:**
    - By splitting the data into training and testing sets, the approach ensures that the predictive models are evaluated on unseen data, providing a realistic estimate of their performance.
5. **Model Building:**
    - Building predictive models, such as logistic regression or decision trees, allows for the quantification of the relationship between customer attributes and the likelihood of subscribing to a term deposit.
6. **Model Evaluation:**
    - Model evaluation metrics (e.g., accuracy, precision, recall, F1-score) provide a quantitative measure of the model's performance, indicating its ability to predict term deposit subscriptions accurately.
7. **Feature Importance Analysis:**
    - Understanding the importance of each feature helps in identifying which customer attributes have the most influence on subscription decisions. This information is valuable for focusing marketing efforts.
8. **Customer Segmentation:**

- Clustering customers into segments based on their characteristics or behavior enables more targeted marketing. Customers with similar traits can be addressed with specific marketing strategies.

9. **Resource Allocation:**
    - Using the insights from customer segmentation and predictive modeling, the approach guides resource allocation. More resources can be allocated to customer segments with a higher likelihood of subscription, optimizing marketing efforts.
10. **Ethical Considerations:**
    - Addressing ethical considerations ensures that marketing strategies respect data privacy and promote fairness in customer targeting, which is essential for maintaining trust and compliance with regulations.
11. **Optimization:**
    - Continuous monitoring and optimization of marketing strategies based on results and feedback allow for adaptation to changing customer behavior and market conditions, ensuring continued effectiveness.
12. **Reporting and Visualization:**
    - Communicating the findings and strategies through reporting and visualization tools ensures that the insights are accessible and actionable for marketing and management teams.

Overall, this approach leverages data analysis and modeling to identify customer segments most likely to subscribe to a term deposit, allowing marketing resources to be directed more effectively, ultimately reducing costs and improving the success rate of marketing campaigns. It combines data science techniques with business strategy to address the problem comprehensively.

# Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

## Original source where the data was obtained is cited and, if possible, hyperlinked.

*Bank_Dataset_1*:https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset

*Bank_Dataset_2*:https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets

*Bank_Dataset_3*:Moro,S., Rita,P., and Cortez,P.. (2012). Bank Marketing. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.

## Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).

**Bank_Dataset_1:** Context Find the best strategies to improve for the next marketing campaign. How can the financial institution have a greater effectiveness for future marketing campaigns? In order to answer this, we have to analyze the last marketing campaign the bank performed and identify the patterns that will help us find conclusions in order to develop future strategies.

Source [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

This is the classic marketing bank dataset uploaded originally in the UCI Machine Learning Repository. The dataset gives you information about a marketing campaign of a financial institution in which you will have to analyze in order to find ways to look for future strategies in order to improve future marketing campaigns for the bank.

There are 17 Columns , having 7 Integer , 6 String and 4 Boolean.

## Column Non-Null Count Dtype

---

0 age 11162 non-null int64 1 job 11162 non-null object 2 marital 11162 non-null object 3 education 11162 non-null object 4 default 11162 non-null object 5 balance 11162 non-null int64 6 housing 11162 non-null object 7 loan 11162 non-null object 8 contact 11162 non-null object 9 day 11162 non-null int64 10 month 11162 non-null object 11 duration 11162 non-null int64 12 campaign 11162 non-null int64 13 pdays 11162 non-null int64 14 previous 11162 non-null int64 15 poutcome 11162 non-null object 16 deposit 11162 non-null object

```
# Specify the file path to your CSV file with forward slashes
DS_1_file_path <- "C:/Users/VJ/Desktop/Bruin/DSC 520/Project/bank_dataset_1.csv"

# Use read.csv() to import the CSV file into a data frame
Data_File_1 <- read.csv(DS_1_file_path)
head(Data_File_1)
```

```
##   age        job marital education default balance housing loan contact day
## 1  59     admin. married secondary      no    2343     yes   no unknown   5
## 2  56     admin. married secondary      no      45      no   no unknown   5
## 3  41 technician married secondary      no    1270     yes   no unknown   5
## 4  55   services married secondary      no    2476     yes   no unknown   5
## 5  54     admin. married  tertiary      no     184      no   no unknown   5
## 6  42 management  single  tertiary      no       0     yes  yes unknown   5
##   month duration campaign pdays previous poutcome deposit
## 1   may     1042        1    -1        0  unknown     yes
## 2   may     1467        1    -1        0  unknown     yes
## 3   may     1389        1    -1        0  unknown     yes
## 4   may      579        1    -1        0  unknown     yes
## 5   may      673        2    -1        0  unknown     yes
## 6   may      562        2    -1        0  unknown     yes
```

```
summary(Data_File_1)
```

```
##       age            job              marital           education
##  Min.   :18.00   Length:11162       Length:11162       Length:11162
##  1st Qu.:32.00   Class :character   Class :character   Class :character
##  Median :39.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :41.23
##  3rd Qu.:49.00
##  Max.   :95.00
##    default             balance        housing              loan
##  Length:11162       Min.   :-6847   Length:11162       Length:11162
```

```
##  Class :character   1st Qu.:  122   Class :character   Class :character
##  Mode  :character   Median :  550   Mode  :character   Mode  :character
##                     Mean   : 1529
##                     3rd Qu.: 1708
##                     Max.   :81204
##    contact               day           month              duration
##  Length:11162       Min.   : 1.00   Length:11162       Min.   :   2
##  Class :character   1st Qu.: 8.00   Class :character   1st Qu.: 138
##  Mode  :character   Median :15.00   Mode  :character   Median : 255
##                     Mean   :15.66                      Mean   : 372
##                     3rd Qu.:22.00                      3rd Qu.: 496
##                     Max.   :31.00                      Max.   :3881
##     campaign          pdays           previous         poutcome
##  Min.   : 1.000   Min.   : -1.00   Min.   : 0.0000   Length:11162
##  1st Qu.: 1.000   1st Qu.: -1.00   1st Qu.: 0.0000   Class :character
##  Median : 2.000   Median : -1.00   Median : 0.0000   Mode  :character
##  Mean   : 2.508   Mean   : 51.33   Mean   : 0.8326
##  3rd Qu.: 3.000   3rd Qu.: 20.75   3rd Qu.: 1.0000
##  Max.   :63.000   Max.   :854.00   Max.   :58.0000
##    deposit
##  Length:11162
##  Class :character
##  Mode  :character
##
##
##
```

```r
missing_values <- sum(is.na(Data_File_1))
print(missing_values)
```

```
## [1] 0
```

**Bank_Dataset_2:**

Term deposits are a major source of income for a bank. A term deposit is a cash investment held at a financial institution. Your money is invested for an agreed rate of interest over a fixed amount of time, or term. The bank has various outreach plans to sell term deposits to their customers such as email marketing, advertisements, telephonic marketing, and digital marketing.

Telephonic marketing campaigns still remain one of the most effective way to reach out to people. However, they require huge investment as large call centers are hired to actually execute these campaigns. Hence, it is crucial to identify the customers most likely to convert beforehand so that they can be specifically targeted via call.

The data is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe to a term deposit (variable y).

Content The data is related to the direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed by the customer or not. The data folder contains two datasets:-

45,211 rows and 18 columns ordered by date (from May 2008 to November 2010)

Detailed Column Descriptions bank client data:

1 - age (numeric) 2 - job : type of job (categorical: “admin.”,“unknown”,“unemployed”,“management”,“housemaid”,“entrepren
“blue-collar”,“self-employed”,“retired”,“technician”,“services”) 3 - marital : marital status (categorical: “married”,“divorced”,“single”; note: “divorced” means divorced or widowed) 4 - education (categorical: “unknown”,“secondary”,“primary”,“tertiary”) 5 - default: has credit in default? (binary: “yes”,“no”) 6 - balance: average yearly balance, in euros (numeric) 7 - housing: has housing loan? (binary: “yes”,“no”) 8 - loan: has personal loan? (binary: “yes”,“no”) # related with the last contact of the current campaign: 9 - contact: contact communication type (categorical: “unknown”,“telephone”,“cellular”) 10 - day: last contact day of the month (numeric) 11 - month: last contact month of year (categorical: “jan”, “feb”, “mar”, . . . , “nov”, “dec”) 12 - duration: last contact duration, in seconds (numeric) # other attributes: 13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact) 14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted) 15 - previous: number of contacts performed before this campaign and for this client (numeric) 16 - poutcome: outcome of the previous marketing campaign (categorical: “unknown”,“other”,“failure”,“success”)

Output variable (desired target): 17 - y - has the client subscribed a term deposit? (binary: “yes”,“no”)

Missing Attribute Values: None

Citation This dataset is publicly available for research. It has been picked up from the UCI Machine Learning with random sampling and a few additional columns.

Please add this citation if you use this dataset for any further analysis.

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

```
# Specify the file path to your CSV file with forward slashes
DS_2_file_path <- "C:/Users/VJ/Desktop/Bruin/DSC 520/Project/bank_dataset_2.csv"

# Use read.csv() to import the CSV file into a data frame
Data_File_2 <- read.csv(DS_2_file_path)
head(Data_File_2)
```

```
##   age.job.marital.education.default.balance.housing.loan.contact.day.month.duration.campaign.pdays.p
## 1                                    58;management;married;tertiary;no;2143;yes;no;unknown;5;may;261
## 2                                      44;technician;single;secondary;no;29;yes;no;unknown;5;may;151
## 3                                    33;entrepreneur;married;secondary;no;2;yes;yes;unknown;5;may;76
## 4                                     47;blue-collar;married;unknown;no;1506;yes;no;unknown;5;may;92
## 5                                             33;unknown;single;unknown;no;1;no;no;unknown;5;may;198
## 6                                     35;management;married;tertiary;no;231;yes;no;unknown;5;may;139
```

```
summary(Data_File_2)
```

```
##  age.job.marital.education.default.balance.housing.loan.contact.day.month.duration.campaign.pdays.pr
##  Length:45211
##  Class :character
##  Mode  :character
```

```
missing_values <- sum(is.na(Data_File_2))
print(missing_values)
```

```
## [1] 0
```

**Bank_Dataset_3:**

Abstract: The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Data Set Information:

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Attribute Information:

Bank client data:

Age (numeric) Job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown') Marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed) Education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate','professional.course','university.degree', 'unknown') Default: has credit in default? (categorical: 'no', 'yes', 'unknown') Housing: has housing loan? (categorical: 'no', 'yes', 'unknown') Loan: has personal loan? (categorical: 'no', 'yes', 'unknown') Related with the last contact of the current campaign: Contact: contact communication type (categorical:'cellular','telephone') Month: last contact month of year (categorical: 'jan', 'feb', 'mar',..., 'nov', 'dec') Day_of_week: last contact day of the week (categorical:'mon','tue','wed','thu','fri') Duration: last contact duration, in seconds (numeric). Importantnote: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known.Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

Campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact) Pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted) Previous: number of contacts performed before this campaign and for this client (numeric) Poutcome: outcome of the previous marketing campaign (categorical:'failure','nonexistent','success')

```
# Specify the file path to your CSV file with forward slashes
DS_3_file_path <- "C:/Users/VJ/Desktop/Bruin/DSC 520/Project/bank-additional-full_dataset3.csv"

# Use read.csv() to import the CSV file into a data frame
Data_File_3 <- read.csv(DS_3_file_path)
head(Data_File_3)
```

```
##   age.job.marital.education.default.housing.loan.contact.month.day_of_week.duration.campaign.pdays.p
## 1                                                                                  56;housemaid;married;basic.4y
## 2                                                                              57;services;married;high.school;unkn
## 3                                                                                  37;services;married;high.school;
## 4                                                                                       40;admin.;married;basic.6y
## 5                                                                                  56;services;married;high.school;
## 6                                                                                 45;services;married;basic.9y;unkn
```

```
summary(Data_File_3)
```

```
##  age.job.marital.education.default.housing.loan.contact.month.day_of_week.duration.campaign.pdays.pr
##  Length:41188
```

```
##  Class :character
##  Mode  :character
```

```
missing_values <- sum(is.na(Data_File_3))
print(missing_values)
```

```
## [1] 0
```

## Required Packages

R offers a wide range of packages and libraries for each of these steps. For example, `tidyverse` or ggplot2,dplyr can be useful for data manipulation and visualization, while `caret` and `randomForest` are popular for modeling. The choice of specific packages and functions will depend on your data and research needs.

## Plots and Table Needs

We are going to use different Charts . Bar Charts:

Use bar charts to visualize categorical data, such as the distribution of term deposit subscriptions by marital status, education level, job type, etc. Grouped bar charts can be used to compare the subscription rates between different categories.

Histograms and Density Plots:

Use histograms and density plots to show the distribution of continuous variables like age, balance, and campaign duration. These plots help you understand the spread and central tendencies of the data. Pie Charts:

Pie charts can be used to show the distribution of categorical variables, such as the proportion of different education levels or housing loan statuses.

Scatter Plots:

Scatter plots are useful for visualizing the relationship between two continuous variables. For example, you can use them to explore the relationship between balance and campaign duration.

Tables:

Tables can present summary statistics, including means, medians, standard deviations, and other relevant metrics for key variables. Cross-tabulation tables can be used to display the relationship between two categorical variables, showing counts and percentages.

## Project Step 2

### How to import and clean my data

```
# Specify the file path to your CSV file with forward slashes
DS_1_file_path <- "C:/Users/VJ/Desktop/Bruin/DSC 520/Project/bank_dataset_1.csv"

# Use read.csv() to import the CSV file into a data frame
Data_File_1 <- read.csv(DS_1_file_path)
```

### What does the final data set look like?

```
head(Data_File_1)
```

```
##   age        job marital education default balance housing loan contact day
## 1  59     admin. married secondary      no    2343     yes   no unknown   5
## 2  56     admin. married secondary      no      45      no   no unknown   5
## 3  41 technician married secondary      no    1270     yes   no unknown   5
## 4  55   services married secondary      no    2476     yes   no unknown   5
## 5  54     admin. married  tertiary      no     184      no   no unknown   5
## 6  42 management  single  tertiary      no       0     yes  yes unknown   5
##   month duration campaign pdays previous poutcome deposit
## 1   may     1042        1    -1        0  unknown     yes
## 2   may     1467        1    -1        0  unknown     yes
## 3   may     1389        1    -1        0  unknown     yes
## 4   may      579        1    -1        0  unknown     yes
## 5   may      673        2    -1        0  unknown     yes
## 6   may      562        2    -1        0  unknown     yes
```

```
summary(Data_File_1)
```

```
##       age            job              marital           education
##  Min.   :18.00   Length:11162       Length:11162       Length:11162
##  1st Qu.:32.00   Class :character   Class :character   Class :character
##  Median :39.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :41.23
##  3rd Qu.:49.00
##  Max.   :95.00
##    default             balance        housing              loan
##  Length:11162       Min.   :-6847   Length:11162       Length:11162
##  Class :character   1st Qu.:  122   Class :character   Class :character
##  Mode  :character   Median :  550   Mode  :character   Mode  :character
##                     Mean   : 1529
##                     3rd Qu.: 1708
##                     Max.   :81204
##    contact                day           month              duration
##  Length:11162       Min.   : 1.00   Length:11162       Min.   :   2
##  Class :character   1st Qu.: 8.00   Class :character   1st Qu.: 138
##  Mode  :character   Median :15.00   Mode  :character   Median : 255
##                     Mean   :15.66                      Mean   : 372
##                     3rd Qu.:22.00                      3rd Qu.: 496
##                     Max.   :31.00                      Max.   :3881
##     campaign          pdays           previous         poutcome
##  Min.   : 1.000   Min.   : -1.00   Min.   : 0.0000   Length:11162
##  1st Qu.: 1.000   1st Qu.: -1.00   1st Qu.: 0.0000   Class :character
##  Median : 2.000   Median : -1.00   Median : 0.0000   Mode  :character
##  Mean   : 2.508   Mean   : 51.33   Mean   : 0.8326
##  3rd Qu.: 3.000   3rd Qu.: 20.75   3rd Qu.: 1.0000
##  Max.   :63.000   Max.   :854.00   Max.   :58.0000
##    deposit
##  Length:11162
##  Class :character
```

```
##  Mode  :character
##
##
##
```

```
str(Data_File_1)
```

```
## 'data.frame':    11162 obs. of  17 variables:
##  $ age      : int  59 56 41 55 54 42 56 60 37 28 ...
##  $ job      : chr  "admin." "admin." "technician" "services" ...
##  $ marital  : chr  "married" "married" "married" "married" ...
##  $ education: chr  "secondary" "secondary" "secondary" "secondary" ...
##  $ default  : chr  "no" "no" "no" "no" ...
##  $ balance  : int  2343 45 1270 2476 184 0 830 545 1 5090 ...
##  $ housing  : chr  "yes" "no" "yes" "yes" ...
##  $ loan     : chr  "no" "no" "no" "no" ...
##  $ contact  : chr  "unknown" "unknown" "unknown" "unknown" ...
##  $ day      : int  5 5 5 5 5 5 6 6 6 6 ...
##  $ month    : chr  "may" "may" "may" "may" ...
##  $ duration : int  1042 1467 1389 579 673 562 1201 1030 608 1297 ...
##  $ campaign : int  1 1 1 1 2 2 1 1 1 3 ...
##  $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : chr  "unknown" "unknown" "unknown" "unknown" ...
##  $ deposit  : chr  "yes" "yes" "yes" "yes" ...
```

```
# Handle missing values
Bank_Dataset1 <- is.na(Data_File_1)
```

There is no Missing data in the given Dataset

```
Bank_Dataset1 <- unique(Data_File_1)
```

All the recods are unquie in the dataset

## Questions for future steps.

1. Right now don't know the outliers and relationship between Variables

## What information is not self-evident?

1. Understand the Data:

Begin by thoroughly understanding the dataset, its structure, and the context in which it was collected. This involves examining data documentation and metadata, if available.

2.Data Visualization:

Visualize the data using various plots and charts, such as histograms, box plots, scatter plots, bar charts, and heatmaps. Visualization can reveal patterns, trends, and potential outliers that may not be immediately obvious from the raw data.

3.Summary Statistics:

Calculate and analyze summary statistics like mean, median, standard deviation, variance, skewness, and kurtosis. These statistics can provide insights into the central tendency, spread, and shape of data distributions.

4.Correlation Analysis:

Explore correlations between variables to identify relationships and dependencies. Correlation matrices and scatter plots can reveal connections between variables, which may suggest causal relationships.

5.Hypothesis Testing:

Formulate hypotheses about the data and use statistical tests to test these hypotheses. Common tests include t-tests, chi-squared tests, and analysis of variance (ANOVA). The results of hypothesis tests can reveal significant differences or associations in the data.

## What are different ways you could look at this data?

We can employ various data analysis and machine learning techniques. Here are different approaches and methods you could use to answer this question:

*1.Descriptive Statistics and Data Visualization:*

Start with basic statistics and data visualization to understand the characteristics of customers who have subscribed to term deposits. Compare them to those who haven't. Use bar charts, pie charts, and histograms to visualize demographic information, such as age, education, and marital status, for both groups. Explore the distribution of numeric variables like balance and duration for subscribers and non-subscribers. ] *2.Feature Importance Analysis:*

Use feature importance techniques to identify which customer attributes have the most influence on subscription decisions. Techniques such as tree-based models (e.g., Random Forest) or statistical methods like chi-squared tests can help determine the most significant features.

*3.Predictive Modeling:*

Build predictive models to estimate the likelihood of a customer subscribing to a term deposit. Utilize classification algorithms like Logistic Regression, Decision Trees, Random Forest, or Gradient Boosting. Use a dataset split into training and testing sets to evaluate model performance. Assess model accuracy, precision, recall, F1-score, and ROC-AUC to understand the model's predictive power.

*4.Segmentation Analysis:*

Segment the customer base into groups with similar characteristics or behaviors. Cluster analysis or customer segmentation techniques can help identify distinct customer segments. Analyze the subscription rates within each segment and focus marketing efforts on high-potential segments.

*5.Customer Behavior Analysis:*

Analyze historical customer behavior and interactions with the bank. Explore how past behavior, such as the number of previous campaigns, affects the likelihood of subscribing to a term deposit. Utilize sequence analysis or association rule mining to find patterns in customer interactions.

*6.Time Series Analysis:*

If the data includes a time dimension, perform time series analysis to uncover trends in subscription rates over time. Identify seasonality and temporal patterns that may impact marketing efforts.

### How do you plan to slice and dice the data?

I am planing to slice the data based on predication variable i.e Deposit Variable = TRUE . So that i can see by historically who has subscribe for term deposit and then decide which variable is contributing most in customer positive decision.

Also , age is the big factor in the data and i am going to slice it by age as well to see which age group customers are likely has more term deposits .

### How could you summarize your data to answer key questions?

*1.Frequency Distributions:*

Create frequency distributions to summarize the number of occurrences of each category or value within a variable. This is particularly useful for understanding the distribution of categorical data.

*2.Grouping and Aggregation:*

Group data by specific attributes or categories and calculate aggregate statistics for each group. This is helpful for comparing subsets of data and identifying patterns or differences.

*3.Explore the Correlation between numerical features*

*4.Find Pair Plot*

*5.Check the Data set is balanced or not based on target values in classification*

### What types of plots and tables will help you to illustrate the findings to your questions?

1. Heatmap - Mostly going to use to see the correlation
2. Box Plot - going to use for Outlier detection
3. Bar and Line - Distribution of Variable

### What do you not know how to do right now that you need to learn to answer your questions?

I am not sure which Linear/Logistical = Model i am going to use for my data analysis

##Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

As of now i have not think about it but i will research it more and try to apply machine learning techniques as well if required.

## Questions for future steps

*1. Find out the Outliers in the data. 2. Analysis about correlation between the variables. 3. Apply the regression model to predict the Target variable 4. Implement the model*