Project: Bank Marketing (Campaign)

Name: Project Week 8 (Bank Marketing Campaign)
Week 7: Deliverables
Name: Priyanjali Patel
Email: priyanjalipatel@gmail.com
Country: India
Batch Code:LISUM45
Specialization: Data Science
Submission Date: 6 July 2025
Submitted to: Data Glacier
(Individual project)

Table of Contents

1. Problem Description

ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

2. Data Understanding

- Type of data you have got for analysis

Input variables:

The dataset consists of 21 columns representing both customer-specific and macroeconomic features collected during previous telemarketing campaigns. It includes a mix of categorical variables (e.g., job, marital, education, contact), numerical variables (e.g., age, campaign, pdays, emp.var.rate), and a binary target variable (y) indicating whether a customer subscribed to a term deposit.

Numerical Columns (10 total)

Column	Type	Description
age	Integer	Age of the client
duration	Integer	Duration of last contact (in seconds) – not usable for real-time prediction
campaign	Integer	Number of contacts in the current campaign
pdays	Integer	Days since the client was last contacted
previous	Integer	Number of contacts before this campaign
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	Float	Number of employees (macro indicator)

Categorical Columns (11 total)

Column	Type	Description
job	Object	Type of job (e.g., admin, technician)
marital	Object	Marital status
education	Object	Education level

default Object Has credit in default?

housing Object Has a housing loan?

Y Object Target variable: subscribed to term deposit? (yes/no)

This classification helps inform encoding strategies, missing value handling, and feature transformation techniques used during data preparation and modeling.

As instructed in the dataset, the duration feature is dropped from the dataset.

duration : Length of the last call (in seconds). **Important**: This strongly influences whether the client subscribed, but since it's only known *after* the call, it shouldn't be used in a real predictive model.

- What are the problems in the data (number of NA values, outliers , skewed etc)

  i) There are zero missing values in the dataset.

```
[5]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        41188 non-null  int64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
 14  poutcome        41188 non-null  object
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m       41188 non-null  float64
 19  nr.employed     41188 non-null  float64
 20  y               41188 non-null  object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```

There are zero missing values in the dataset.

  ii) There are 12 duplicate rows out of `41188 rows.`

df.drop_duplicates()

```
[7]: #dropping duplicates

     df.drop_duplicates()
     #df.drop_duplicates(keep = 'first', inp
     print(df.duplicated().sum())
     df= df.drop_duplicates()

     print(df.shape)
```

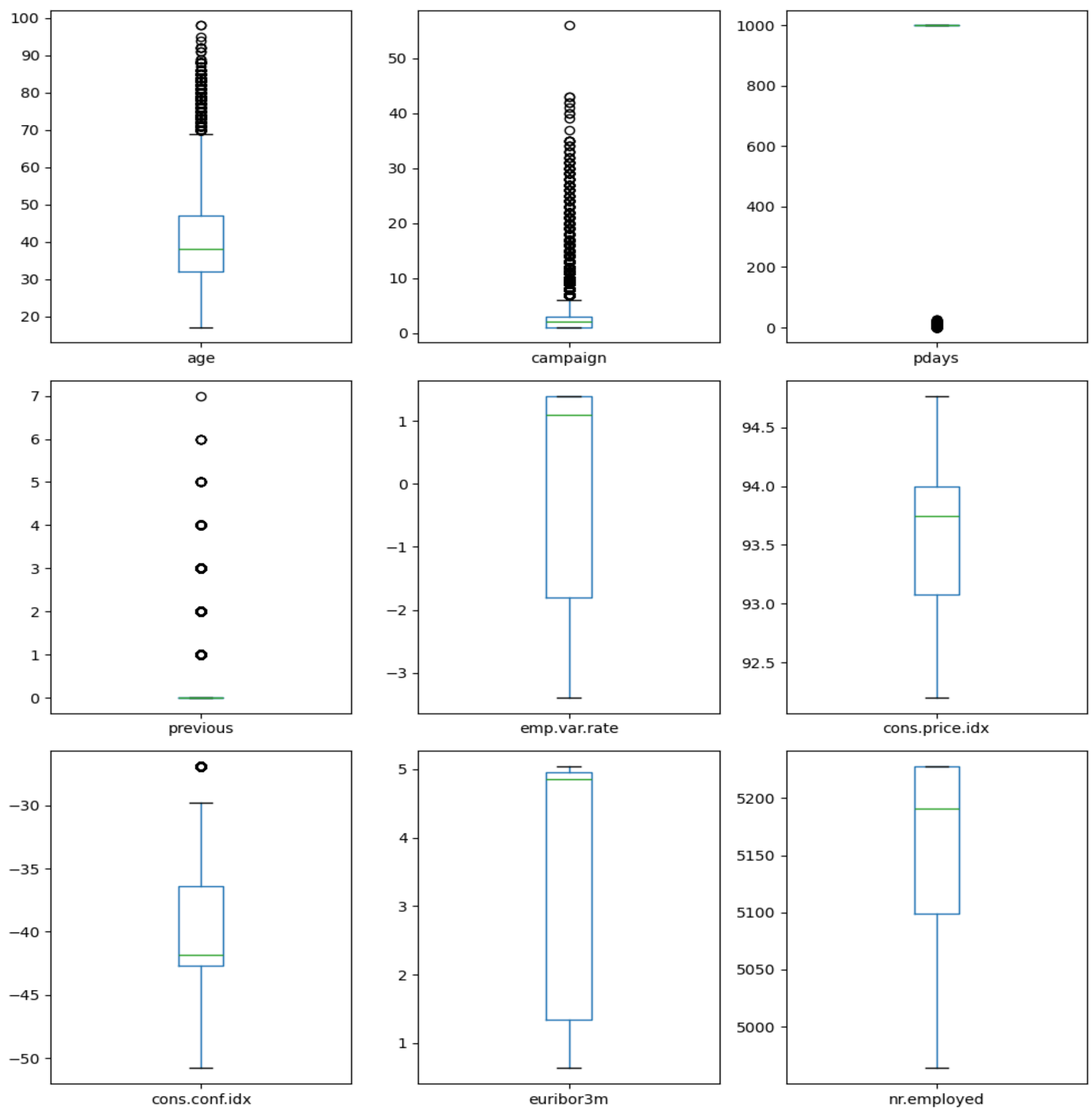
```
12
(41176, 21)
```

iii) No of unique values in each column

```
[8]: print(df.nunique().sort_values(ascending=False))
```

```
duration          1544
euribor3m          316
age                 78
campaign            42
pdays               27
cons.conf.idx       26
cons.price.idx      26
job                 12
nr.employed         11
month               10
emp.var.rate        10
previous             8
education            8
day_of_week          5
marital              4
default              3
poutcome             3
loan                 3
housing              3
contact              2
y                    2
dtype: int64
```

- Outliers can distort statistical measures like mean and standard deviation, leading to misleading analysis. They often reduce the accuracy of machine learning models—especially those sensitive to extreme values—by skewing predictions and harming generalization. Outliers also affect data visualizations by stretching chart scales, making it harder to spot real trends in the data. Removing or treating them helps improve model performance and data interpretation.



In the box plot, outliers are the points outside the **whiskers**. The outlier values are much higher or lower than the rest of the points.

These are the outliers in the dataset:

Age, pdays, campaign, Previous

- **Approaches to Handle Data Issues**

To ensure data quality and improve model performance, the following approaches are applied:

🔹 1. Handling Missing/Unknown Values (NA or 'unknown')

- **Categorical columns** like job, marital, education, default, etc., contain 'unknown' instead of actual NAs.
- **Approach**: Treat 'unknown' as a separate category or impute using the median, mean or mode when appropriate.
- **Why**: This preserves data volume and avoids biased removal of potentially informative rows.

🔹 2. Handling Outliers

- Columns like 'age', 'campaign', 'pdays','previous','emp.var.rate' have extreme values.
- **Approach**:
  - Cap values at the 95th percentile (Winsorizing) for campaign and previous. (The term "quantile(0.95)" refers to the 95th percentile of a dataset, which is the value below which 95% of the data points fall. In simpler terms, it's the point where 95% of the values are less than or equal to that specific value)
  - Treat pdays = 999 as a special case (e.g., create a binary feature: previously contacted or not).
  - Visualize and evaluate outliers in age for domain consistency.
- **Why**: Outliers can distort model learning, lead to overfitting, and affect interpretability.

🔹 3. Encoding Categorical Variables

- Features like job, education, marital, contact, etc., are categorical.

- **Approach**: Use **one-hot encoding** or **label encoding** depending on the model used.
- **Why**: ML algorithms require numerical inputs.

🔹 4. Class Imbalance in Target Variable

- Target (y) is imbalanced (no >> yes).
- **Approach**: Try **SMOTE**, **class weighting**, or **undersampling** techniques during model training.
- **Why**: To avoid biasing the model toward the majority class and improve recall for minority class.

These preprocessing steps ensure that the dataset is clean and suitable for training accurate and generalizable machine learning models.

5. Github Repo link

https://github.com/priyanjalipatel/Data_Glacier_Final_Project/tree/main