

Data Analysis Task: Data Preprocessing & Visualization

Dataset: Bank Marketing Dataset (`bank.csv`)

Objective:

Students will practice **data cleaning**, **handling missing values**, and **basic data visualization** using Python.

Part 1: Data Understanding

1. Load the dataset using **pandas**.
 2. Display:
 - First 5 rows
 - Dataset shape (rows, columns)
 - Column names
 3. Use `info()` to:
 - Identify **data types**
 - Check **non-null counts**
-

Part 2: Data Type Checking & Handling

1. Separate columns into:
 - **Numerical columns** (int, float)
 - **Categorical columns** (object)
 2. Check if any column needs **data type conversion** (e.g., numbers stored as objects).
 3. Convert data types if necessary.
- List of numerical columns
 - List of categorical columns
-

Part 3: Missing Values Analysis

1. Check for missing values using:
 - `isnull().sum()`
2. Calculate the **percentage of missing values** for each column.

A table showing:

- Column name
- Number of nulls

- Percentage of nulls

Part 4: Handling Missing Values (3 Methods)

◆ Method 1: Drop Columns

- Drop columns that have **more than 40% missing values**.
- Explain why dropping columns may be risky.

◆ Method 2: Drop Rows

- Drop rows containing null values.
- Compare dataset size **before and after** dropping rows.

Method 3: Imputation

Handle missing values as follows:

Numerical Columns:

- Replace nulls using:
 - Mean
 - Median

Categorical Columns:

- Replace nulls using:
 - Mode (most frequent value)

Deliverable:

- Code for each method
- Short explanation of **when to use each method**

Part 5: Basic Data Visualization

Create **at least 5 visualizations**:

Numerical Visualizations

1. Histogram of **age**
2. Boxplot of **balance**
3. Distribution of **duration**

Categorical Visualizations

4. Bar chart of **job**
5. Bar chart of **deposit (yes/no)**

Relationship Plot

6. Count plot of **deposit vs housing**

Deliverable:

- Clear titles
 - Labeled axes
 - Brief interpretation (1–2 sentences per plot)
-

Part 6: Final Questions (Short Answers)

1. Which preprocessing method was **most effective** for this dataset?
 2. Which feature seems most related to **deposit subscription**?
 3. What problems could appear if preprocessing is skipped?
-



Tools & Libraries

- Python
 - pandas
 - numpy
 - matplotlib / seaborn
-