WEEK 8: Data Understanding

Name: Sainad Reddy Naini (Individual)
Email: nainisainad@gmail.com
Country: India
College: University of Bath (Graduate -UK)
Specialization: Data Science

1: PROBLEM DESCRIPTION

ABC Bank is preparing to launch a new term deposit product and aims to maximize the effectiveness of its marketing efforts. To achieve this, they want to develop a machine learning (ML) model that can predict whether a customer will subscribe to the term deposit based on historical data from previous marketing campaigns. This predictive model will help the bank identify potential customers who are more likely to purchase the product, allowing them to focus their marketing efforts more effectively and efficiently.

2: DATA UNDERSTANDING

The dataset is related to marketing campaigns of a Portuguese banking institution, specifically focused on term deposit subscriptions. It is publicly available and can be found on the UCI Machine Learning Repository without needing special permissions. This dataset is named "bank marketing" and consists of two files: bank-full.csv and bank-full-additional.csv.

We will use bank-full-additional.csv because it includes five new features that enhance prediction accuracy, even though the duration feature is not considered. We are avoiding the duration feature as it has a strong impact on the outcome but is impractical for predictive models since it is only known after the call is made. Therefore, bank-full-additional.csv will be our dataset.

This dataset contains 21 features/attributes (20 input variables and 1 output variable) with 41,188 observations (rows). We will use the input variables to predict the future outcome variable (term deposit subscription) after training a machine learning model on the given input and output data. Here are the data attributes:

Data Attributes: The dataset contains both numerical and categorical data.

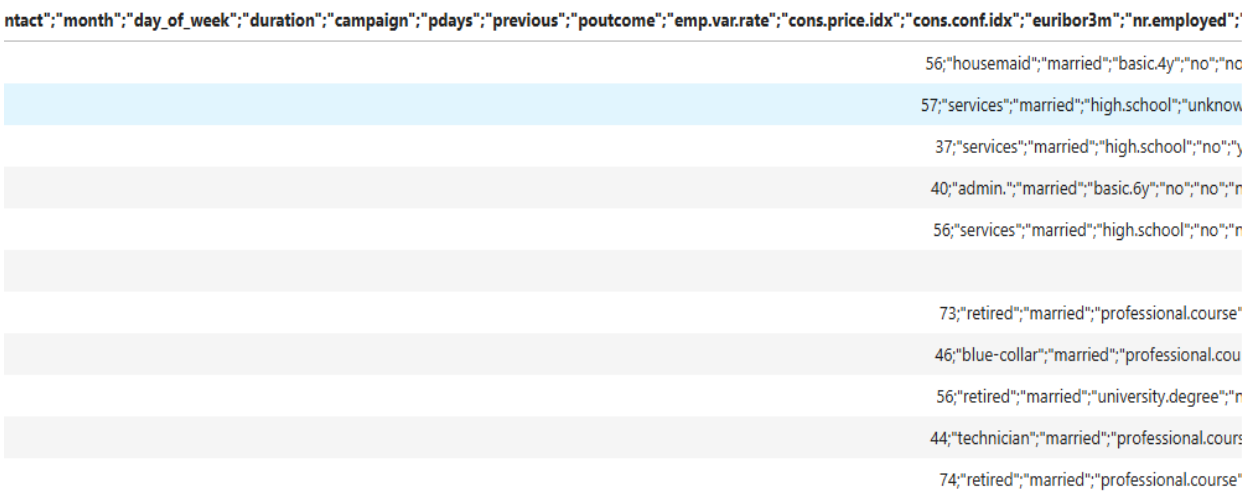
1. **Client Data**: Age, job, marital status, education, default, housing, loan.
2. **Campaign Data**: Contact, month, day_of_the_week, duration, campaign, pdays, previous, poutcome.
3. **Economic Indicators**: emp.var. rate (Employment variation rate), cons.price.idx (consumer price index), cons.conf.idx (consumer confidence index), euribor3m (euribor 3-month rate), nr. employed (number of employees)
4. **Target Variable**: y (Whether the client subscribed to a term deposit(yes/no))

Brief description of each attribute, sorted into numerical and categorical types:

VARIABLES	DESCRIPTION	TYPE
Age	Clients age	Numeric
Job	Type of job	Categorical
Marital	marital status	Categorical
Education	Education	Categorical
Default	Credit in Default	Categorical
Housing	Housing loan	Categorical
Loan	Personal loan	Categorical
Contact	communication type	Categorical
Month	last contact month of year	Categorical
Day_Of_Week	last contact day of the week	Categorical
Duration	last contact duration, in seconds	
Campaign	number of contacts performed during this campaign	Categorical
pdays	days since last contact from previous campaign (numeric; 999 means client was not previously contacted)	Numeric
Previous	number of contacts performed before this campaign and for this client	Numeric
poutcome	outcome of the previous marketing campaign	Categorical
emp.var. rate	employment variation rate - quarterly indicator	Numeric
cons.price.idx	consumer price index - monthly indicator	Numeric
cons.conf.idx	consumer confidence index - monthly indicator	Numeric
euribor3m	euribor 3-month rate - daily indicator	Numeric
nr. employed	number of employees - quarterly indicator	Numeric
y	Whether the client subscribed to a term deposit	Categorical

3: DATA PROBLEMS

Data structure: The data is structured in **semi-colon-separated values**.



NA Values: There are no missing values in the dataset, but there are unknown values in the **job**, **marital**, **education**, **default**, **housing**, and **loan** fields. While null value represents the absence of any value or a missing value in a dataset and unknown value, conceptually, refers to a value that exists but is not currently known.

Snapshot of null values:

```
<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
```

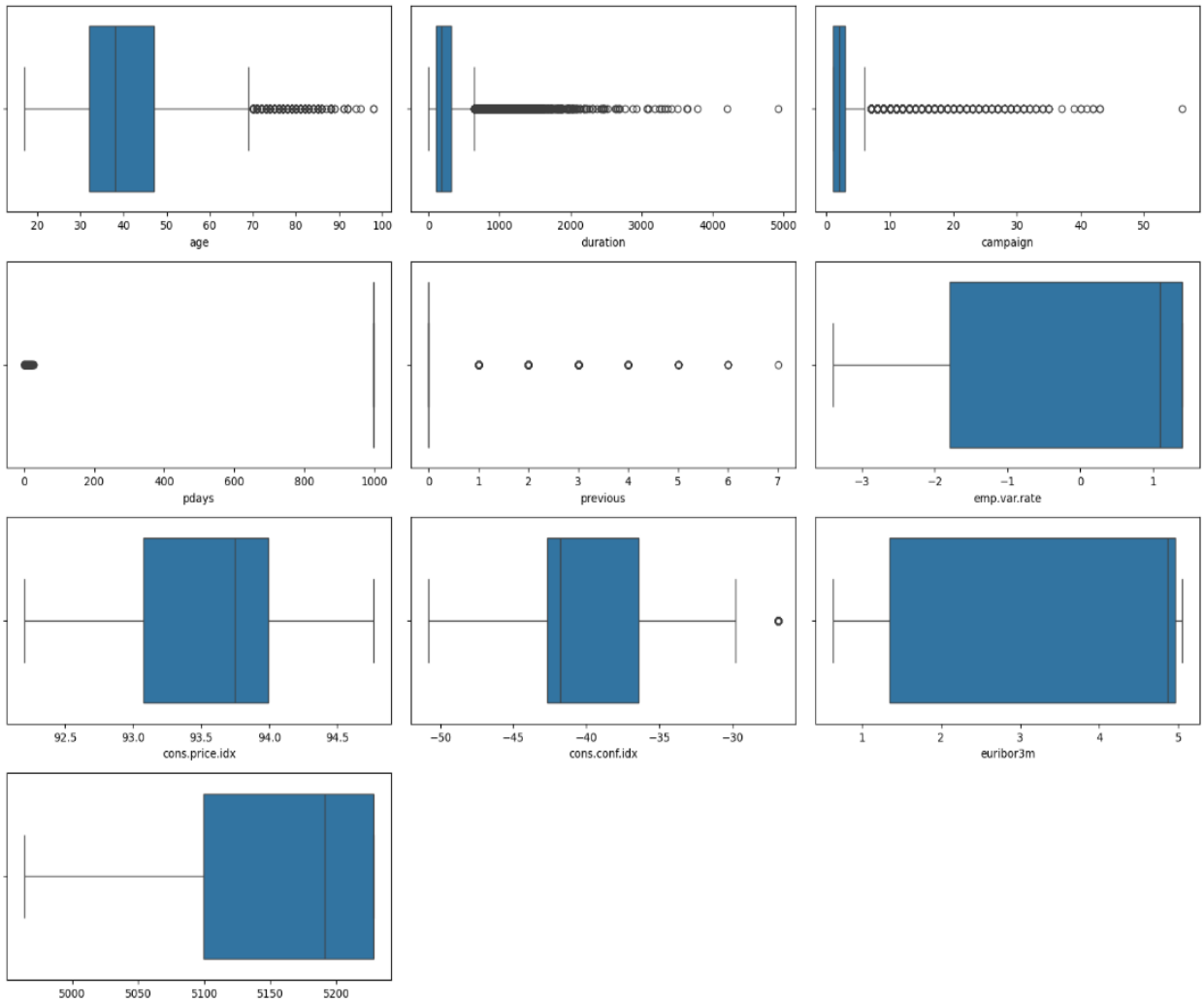
Snapshot of unknown values:

```
The unknown values in job column are 330
The unknown values in marital column are 80
The unknown values in education column are 1731
The unknown values in default column are 8597
The unknown values in housing column are 990
The unknown values in loan column are 990
```

Outliers: An outlier is a data point that differs significantly from other observations in a dataset. Identifying and analysing outliers is important because they can impact the results of statistical analyses and may provide insights into underlying processes or issues. The best practice to find outliers is through visualizations, such as using a boxplot for numeric data. For non-numeric (categorical) data, frequency distribution can be used.

Potential outliers exist in numerical features like **age**, **duration**, **campaign**, **pdays**, and **previous**.

Snapshot of boxplots:

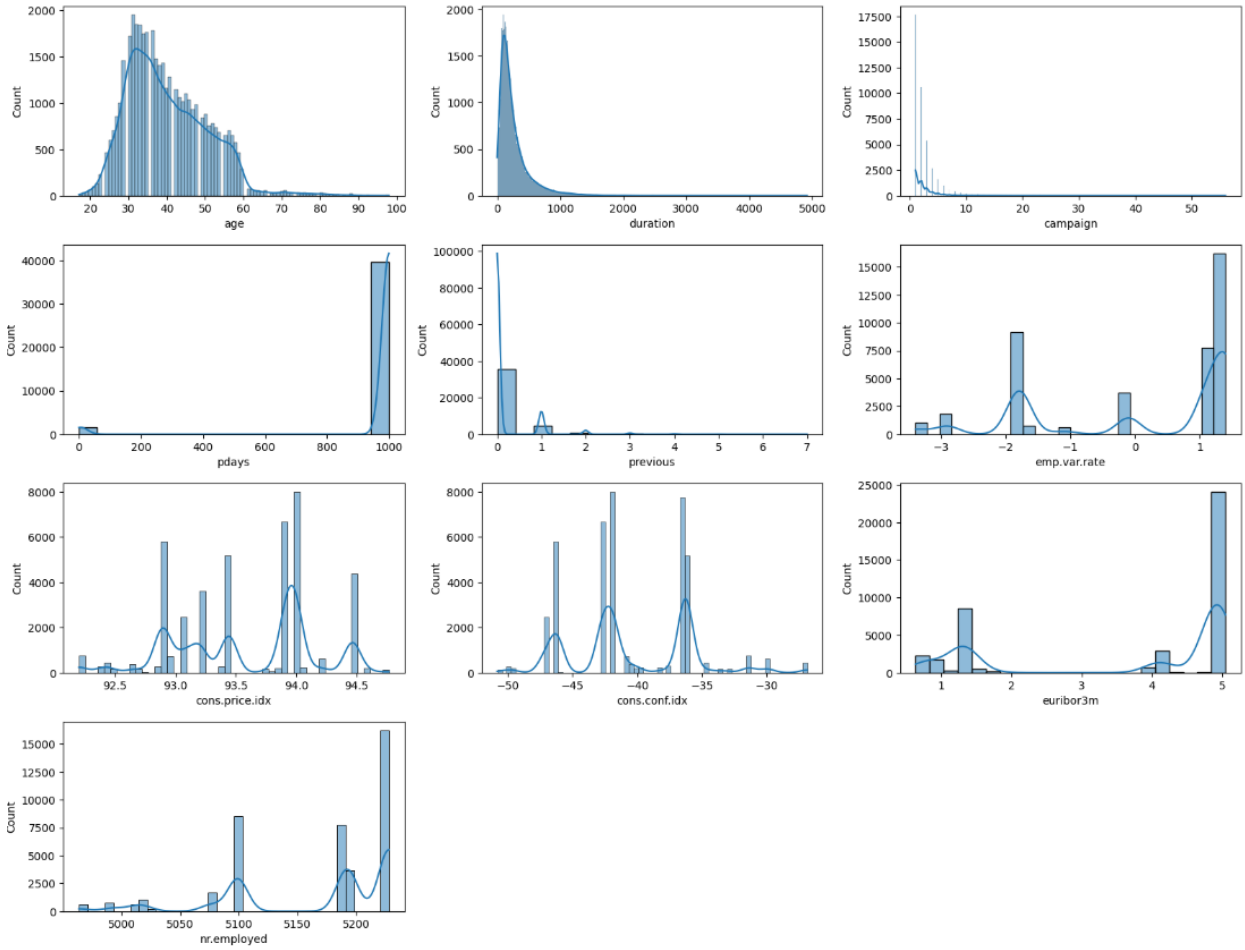


Skewness: Skewness indicates whether the data is spread out more to the left or the right of the mean, or if it is symmetric. Understanding skewness is crucial for accurate data analysis and interpretation. Addressing skewness can lead to more accurate models, better predictions, and clearer communication of results. Skewness can also be visually assessed using histogram

- **Right-Skewed**: The tail is longer on the right side.
- **Left-Skewed**: The tail is longer on the left side.

We can see skewness in the following plots:

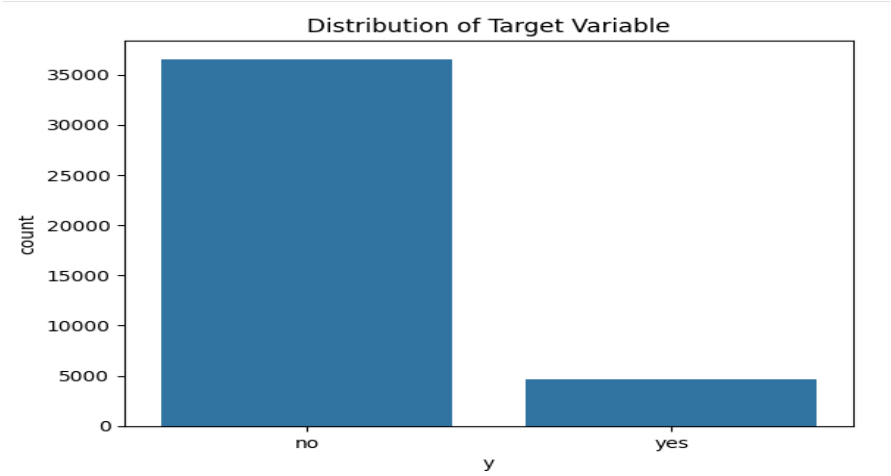
```
age 0.78 - Positive Skewness (Right-Skewed)
duration 3.26 - Positive Skewness (Right-Skewed)
campaign 4.76 - Positive Skewness (Right-Skewed)
pdays -4.92 - Negative Skewness (Left-Skewed)
previous 3.83 - Positive Skewness (Right-Skewed)
emp.var.rate -0.72 - Negative Skewness (Left-Skewed)
cons.price.idx -0.23 - Approximately Symmetric
cons.conf.idx 0.30 - Approximately Symmetric
euribor3m -0.71 - Negative Skewness (Left-Skewed)
nr.employed -1.04 - Negative Skewness (Left-Skewed)
```



Imbalance: Imbalance occurs when the distribution of classes is uneven in a dataset, i.e., the frequency of **one class** is much **higher** than the **other class**. This imbalance can affect model performance by introducing bias, leading the model to predict the majority class more frequently.

For the **output variable** ('y'), the count of class "**no**" is approximately **88.7%** (36,537), and class "**yes**" is **11.3%** (4,639), showing a significant imbalance in the data.

Bar graph of Target variable:



4: APPROACHES TO OVERCOME DATA PROBLEMS

1. **Data Loading:** To load the data correctly, specify the delimiter as a semi-colon when using Pandas to read the CSV file.
2. **Duplicates:** There are 12 duplicate rows in the dataset. We can use the drop_duplicates method to remove these rows, keeping only the first occurrence.
3. **Unknown Values:** Although there are no missing data, there are unknown values. We can handle these using methods like mean, median, mode, or KNN imputation. If the amount of unknown data is very small, it might be more practical to drop it.
4. **Outliers:** We can address outliers by removing them, capping them, or applying transformations.
5. **Skewness:** To address skewness, we can apply transformations such as log or square root for positive skewness and inverse or log of absolute values for negative skewness.
6. **Imbalance:** To handle class imbalance, we can use resampling techniques like under sampling or oversampling, and evaluate model performance using metrics such as the F1 score.

Note: The outlined data problems and proposed solutions represent an initial overview and general approaches. The most effective solution will be determined after conducting a more in-depth analysis. Additionally, the source code for this initial overview, along with the data cleaning and transformation processes, will be uploaded next week.

5: GITHUB REPO LINK

https://github.com/sainadreddy/Data_Glacier_Intership_2024