Group Name: Project on my own.
Name: Aditi Dadariya
Email: aditi.dadariya@gmail.com
Country: United Kingdom
College/Company: MSc (Data Science and Advanced Computing) from University of Reading
Specialization: Data Science
Project Name: Bank Marketing (Campaign)
Github Repo link: https://github.com/aditidadariya/BankMarketingCampaign

Problem description: ABC Bank wants to sell its 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).

Data Cleaning and Transformation: The dataset consists of 41188 records and 21 variables. The variables are 'age', 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays', 'previous', 'poutcome', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed', 'y'. The variable 'y' is the target variable, while the rest of the variables are considered as features.

The variables 'age', ''duration', 'campaign', 'pdays' and 'previous' are of int64 data type. The variables 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m' and 'nr.employed' are of float64 data type. The variables 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'poutcome' and 'y' are of object data type. Please refer Fig 1 below.

```
<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
None
```

Fig 1: Data info.

Age varies between 17 to 98. Duration varies between 0 to 4918. Campaign varies between 1 to 56. Pdays varies between 0 to 999. Previous varies between 1 to 7. Emp.var.rate varies between -3.4 to 1.4. Cons.price.idx varies between 92.2 to 94.77. Cons.conf.idx varies to -50.8 to -26.9. Euribor3m varies between 0.63 to 5.04. Nr.employed varies between 4963.6 to 5228.1. Please refer Fig 2 below for these details.

	age	duration	campaign	pdays	previous \
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963
std	10.42125	259.279249	2.770014	186.910907	0.494901
min	17.00000	0.000000	1.000000	0.000000	0.000000
25%	32.00000	102.000000	1.000000	999.000000	0.000000
50%	38.00000	180.000000	2.000000	999.000000	0.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000
max	98.00000	4918.000000	56.000000	999.000000	7.000000

	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	1.570960	0.578840	4.628198	1.734447	72.251528
min	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	1.400000	94.767000	-26.900000	5.045000	5228.100000

Fig 2: Data describe.

Steps executed to clean and transform the dataset are below:

1. <u>Missing Values:</u> The dataset does not contain any null (NaN) values. However, there are 'unknown' values present in 'job', 'marital', 'education', 'default', 'housing', 'loan' columns, which can be treated as missing values. The variable 'job' has 330 "unknown" values. The variable 'marital' has 80 "unknown" values. The variable 'education' has 1731 "unknown" values. The variable 'default' has 8597 "unknown" values. The variable 'housing' has 990 "unknown" values. The variable 'loan' has 990 "unknown" values. Please refer Fig 3. The "unknown" values in the dataset have been replaced with NaN values. Subsequently, the NaN values have been replaced with its most frequently used value by SimpleImpu-ter method. The decision was made not to drop the missing values due to their substantial presence in the dataset. Removing these values could potentially impact the integrity of the training data.

```
Variables having "unknown" value ['job' 'marital' 'education' 'default' 'housing' 'loan']


Variables with "unknown" counts:
age                  0
job                330
marital             80
education         1731
default           8597
housing            990
loan               990
contact              0
month                0
day_of_week          0
duration             0
campaign             0
pdays                0
previous             0
poutcome             0
emp.var.rate         0
cons.price.idx       0
cons.conf.idx        0
euribor3m            0
nr.employed          0
y                    0
dtype: int64
```

Fig 3: "unknown" values

2. Inconsistent datatypes: Once the "unknown" values were transformed into NaN values, the categorical variables ('job', 'marital', 'education', 'default', 'housing', 'loan') were examined for potential mixed datatypes. However, it was confirmed that these variables do not exhibit multiple datatypes, as they are consistently of the same datatype throughout the dataset.

3. Encoding the categorical value: The categorical variables ‘job’, ‘marital’, ‘education’, ‘default’, ‘housing’, ‘loan’, ‘contact’, ‘month’, ‘day_of_week’, ‘poutcome’ and ‘y’ are converted to numeric using Label Encoder to maintain the consistency of data.

4. Box Plot: A box plot was generated to visually examine the presence of outliers in the dataset. Please refer to Fig 4 for the plot. Upon analysis, it is evident that the variables contain outliers, as indicated by the data points beyond the whiskers of the box plot.

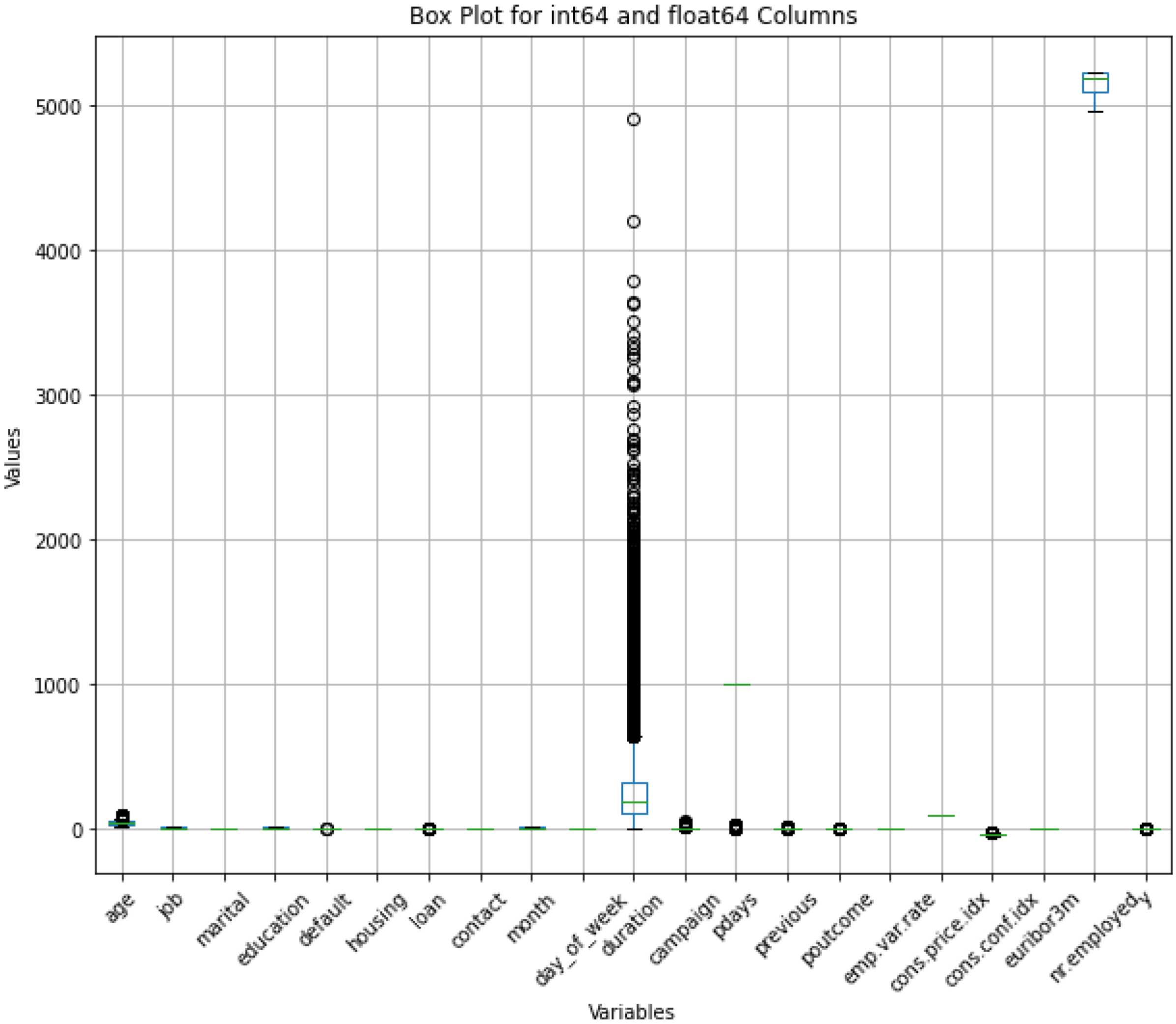

Fig 4: Box Plot of numeric variables

5. Outlier Detection by IQR method: To identify outliers in the dataset, the Interquartile Range (IQR) method was used. The analysis revealed the presence of outliers in multiple columns, including 'age', 'duration', 'campaign', 'pdays', 'previous', and 'cons.conf.idx'. Please refer to Fig 5 for a visual representation of the number of outliers in each column. The IQR method identified a significant number of outliers. To validate these findings, the ZScore method was also employed to detect outliers.

```
469 number of row are Outliers in column age
3 number of row are Outliers in column default
6248 number of row are Outliers in column loan
2963 number of row are Outliers in column duration
2406 number of row are Outliers in column campaign
1515 number of row are Outliers in column pdays
5625 number of row are Outliers in column previous
5625 number of row are Outliers in column poutcome
447 number of row are Outliers in column cons.conf.idx
4640 number of row are Outliers in column y
```

Fig 5: Number of outliers by IQR

6. Outlier Detection by ZScore: After detecting outliers using the IQR method, the dataset was further examined using the Z-score method to validate the findings. The analysis revealed the presence of outliers in the columns 'default', 'duration', 'campaign', 'pdays', and 'previous'. Please refer to Fig 6 for a visual representation of the outliers. To address this issue, the QuantileTransformer technique was employed to impute the outliers. The outliers were replaced with the corresponding quantile value, resulting in a more robust representation of the data. As the 'age' variable is assumed to be a critical factor and should not be imputed with any value, it was excluded from the outlier detection evaluation.

```
3 number of row are Outliers in column default
861 number of row are Outliers in column duration
869 number of row are Outliers in column campaign
1515 number of row are Outliers in column pdays
1064 number of row are Outliers in column previous
```

Fig 6: Number of outliers by ZScore

7. Imbalance data: The target variable 'y' in the dataset consists of 36,548 values for the category 'no' and 4,640 values for the category 'yes'. This reveals an imbalance in the dataset. To address this issue, the SMOTE (Synthetic Minority Over-sampling Technique) method was employed to balance the dataset, ensuring a more equitable distribution for subsequent analysis. Please refer Fig 7 and Fig 8.

```
The count of each category in the target variable is:
0    36548
1     4640
Name: y, dtype: int64
```

Fig 7: Number of categories in target variable

```
The count of each category in the target variable is:
y
0    36548
1    36548
dtype: int64
```

Fig 8: Number of categories in target variable after balancing