Data Glacier

Your Deep Learning Partner

Bank Marketing (Campaign)

Data Science Virtual Internship

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

Agenda

- Problem Description
- Approach
- EDA
- EDA Summary
- Recommended Models

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.

Approach

1. **Data Preparation**
    - Check for null values, missing data, data types, and duplicates to ensure clean data.
2. **Feature Engineering**
    - Enhance the dataset by adding relevant columns (features) to facilitate visualizations and trend analysis.
3. **Exploratory Data Analysis**
    - Explore the data to understand patterns, relationships, and key insights.
4. **Insights and recommendations**
    - Summarize key findings and provide actionable insights.
5. **Model building**
    - While this presentation focuses on recommendations, we suggest specific models for technical users to explore.

EDA
(Understanding the Data)

Before diving into the analysis, let's first understand the dataset. It is a publicly available dataset from the UCI Machine Learning Repository, containing information from a marketing campaign by a Portuguese banking institution. The dataset includes details about customer features and whether they subscribed to a term deposit.

- There are two files in the dataset, but I have used the **bank-full-additional.csv** file, which contains **21** columns and **41,188** rows. This file provides more data, which helps improve the accuracy of the predictions compared to the other file.
- The dataset contains both numeric and categorical columns, so we will proceed with exploratory data analysis (EDA) accordingly.

EDA

(Understanding the Data)

While understanding the data, I focused on two main questions:

- What are the problems in the dataset ?
- How can we solve these problems?

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	Numeric
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

Fig:- The dataset consists of 21 columns, including the target variable (y), along with some basic descriptions.

EDA

(Understanding the Data- What are the problems in the dataset ?)

What are the problems in the dataset ?

The dataset has no missing values, but it does contain some unknown values in the columns for **job**, **marital**, **education**, **default**, **housing**, and **loan**. There are also outliers present in the columns for **age**, **duration**, **pdays**, **previous**, and **campaign**. Additionally, the dataset is imbalanced, meaning some categories have far more data than others. Finally, there is some skewness in the data, which means that certain variables are not evenly distributed.

EDA
(Understanding the Data- How can we solve these problems?)

How can we solve these problems?

- Although there are no missing values, the dataset has some unknown values. We can handle these by replacing them with the most common value in that column, like using the most frequent answers for **housing**, **loan**, and **education**. If the unknown data is very small, it might be easier to just remove those entries, such as for **job** and **marital**.
- To deal with outliers, which are extreme values that can affect our results, we can remove them, adjust their values, or apply techniques to reduce their impact.
- For the problem of class imbalance, where some categories have less values compared to others, we can use methods to balance the data, such as undersampling(reducing examples for the overrepresented ones). Another option is to ensure that both training and testing datasets are balanced. We should also keep an eye on how well our model performs and adjust our methods as needed to improve accuracy.

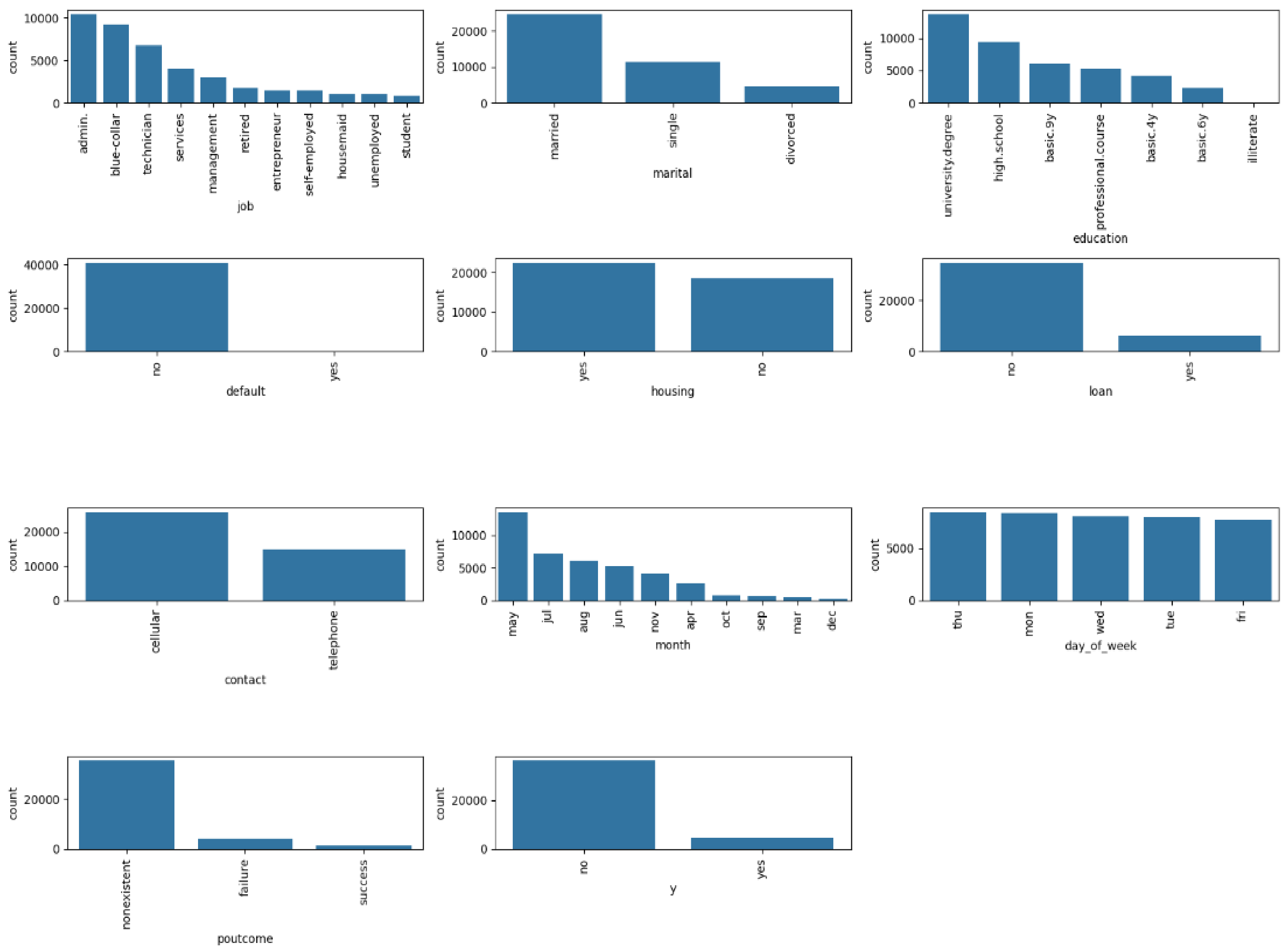
EDA
(Types of Analysis)

- Before moving forward, I have already dealt with the issues in the data, like unknown values, outliers, and duplicate entries, as explained in the previous slide. These are basic steps in data cleaning and preparation. Now, we are ready to move on to more detailed analysis.
- Since we have both numerical and categorical (non-numeric) values, we will explore univariate and bivariate analysis, which involves analyzing single and paired variables for both types of data.

EDA

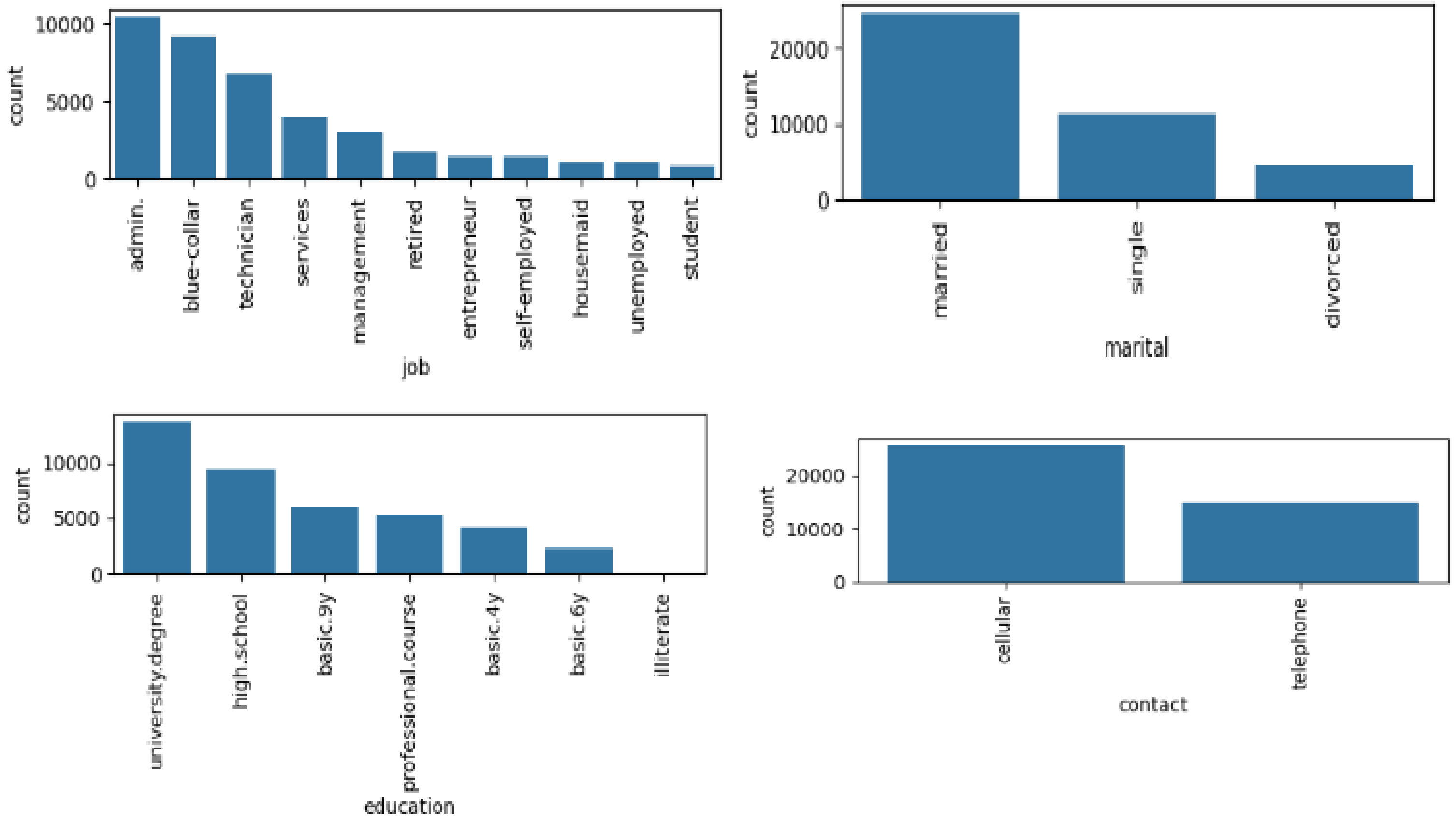
(Univariate Analysis of Categorical Features)

- There are 10 categorical features(columns) in this dataset excluding y(target variable).



EDA
(Univariate Analysis of Categorical Features)

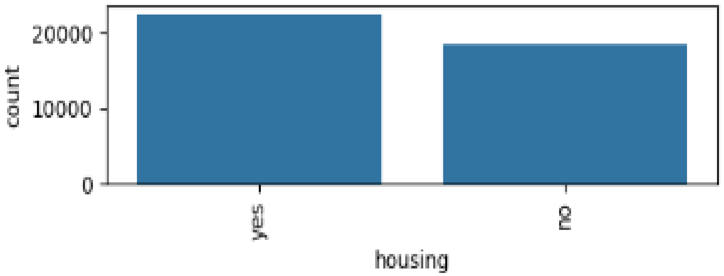
- Most people in the dataset have administrative jobs, followed by blue-collar workers and technicians. There are fewer students.
- More people in the dataset are married.
- Most of the clients in the dataset are educated.
- Most of the contacts were made using mobile phones.

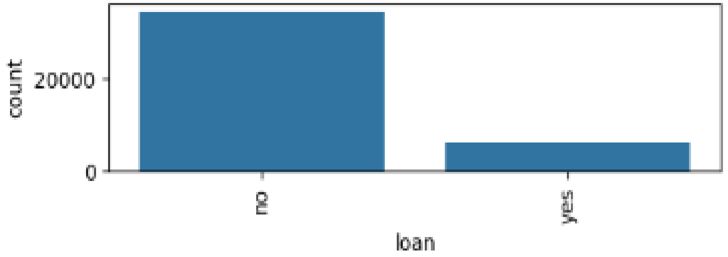


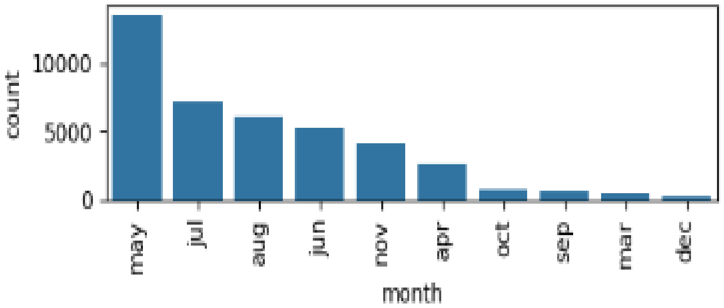
The conclusions from the bar graphs are shown on the left.

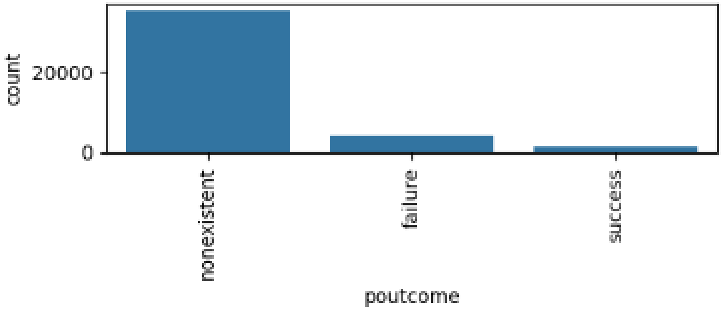
EDA
(Univariate Analysis of Categorical Features)

- There are more people with a house loan compared to those without a personal loan.
- We have more data from May and less from December.
- Most of the poutcome are failures.
- There are significantly more cases of "no" compared to "yes" in the default feature, so will drop this.







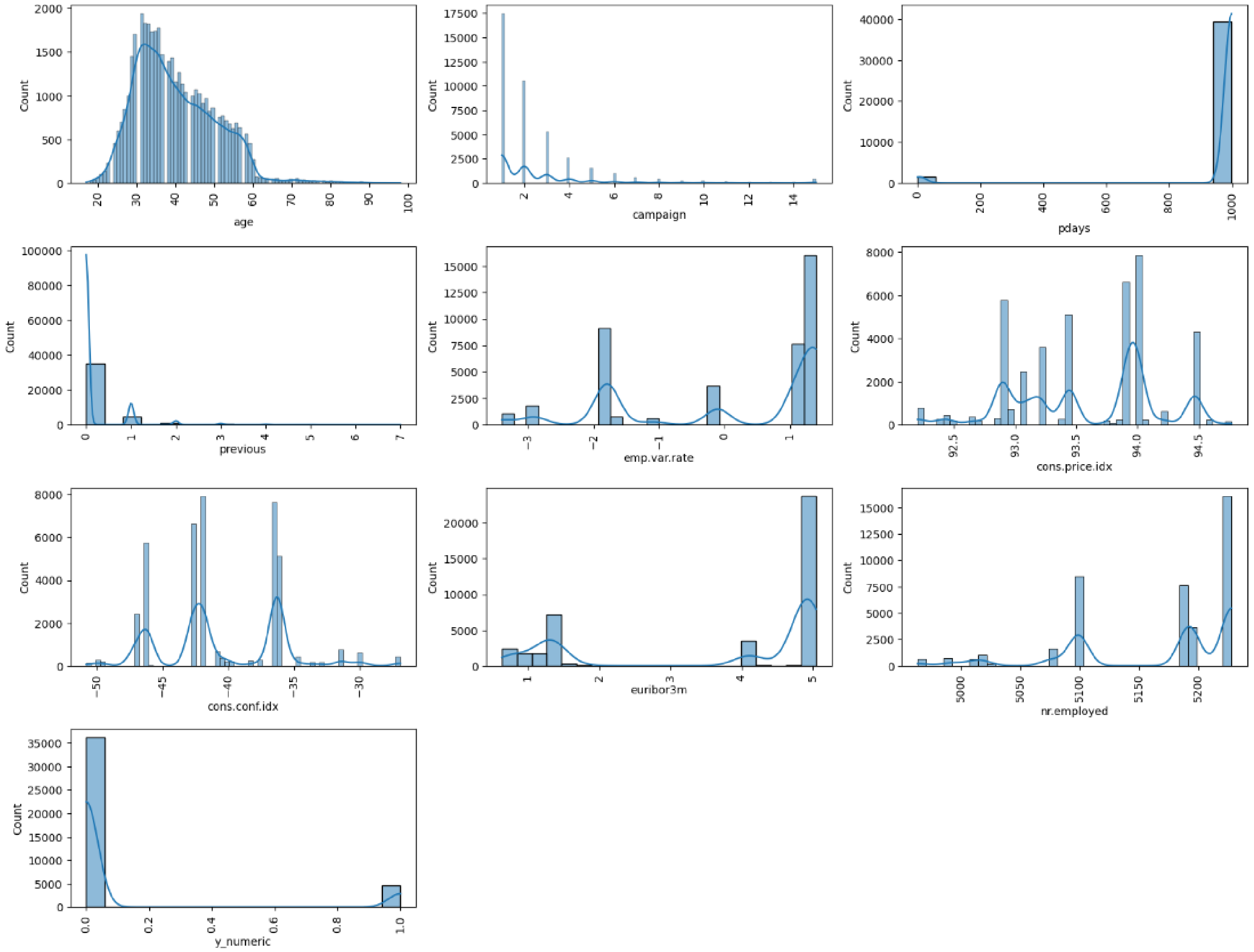


The conclusions from the bar graphs are shown on the left.

EDA

(Univariate Analysis of Numerical Features)

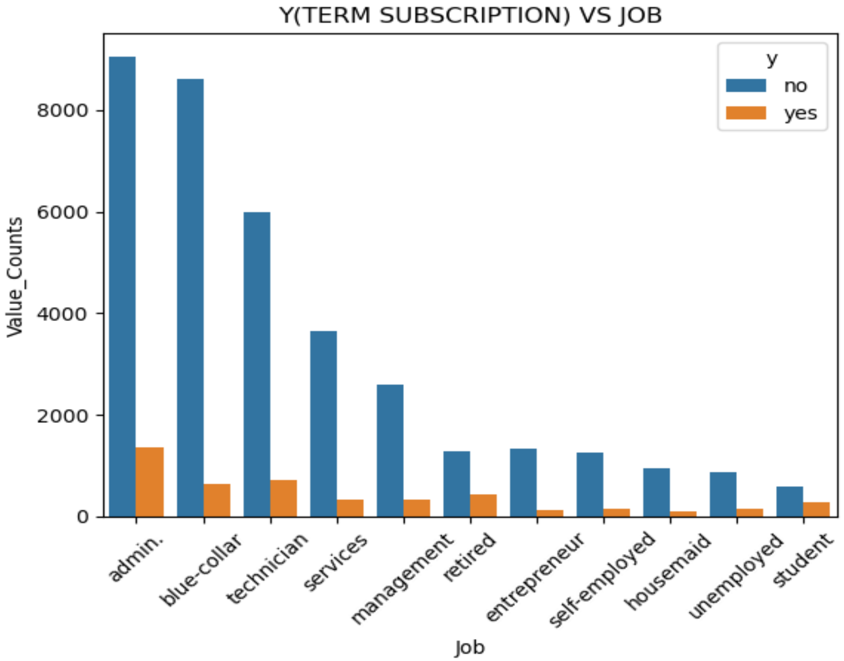
- There are 9 numerical features in the dataset, excluding the target variable (y_numeric)
- The histograms show how the data is distributed.
- Based on these histograms, I have addressed issues like outliers and skewness.
- Most of the people contacted are between the ages of 20 and 60.



EDA

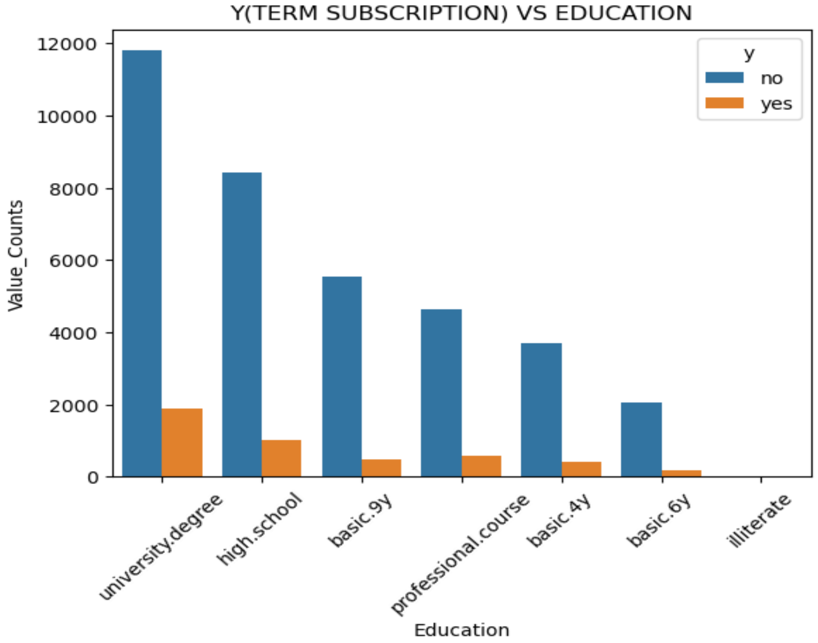
(Bi-variate Analysis of Categorical Features w.r.t term deposit)

- Even though many people work in administrative jobs, retirees are more interested in term deposits.
- Most clients are literate, but those with less education (like illiterates) actually sign up for term deposits more, followed by those with university degrees.




```
Percentage breakdown for job:
job  admin.  blue-collar  entrepreneur  housemaid  management  retired  \
y
no    87.03        93.13         91.53      89.97       88.77    74.72
yes   12.97         6.87          8.47      10.03       11.23    25.28

job  self-employed  services  student  technician  unemployed
y
no           89.48     91.85    68.54       89.18       85.73
yes          10.52      8.15    31.46       10.82       14.27
```



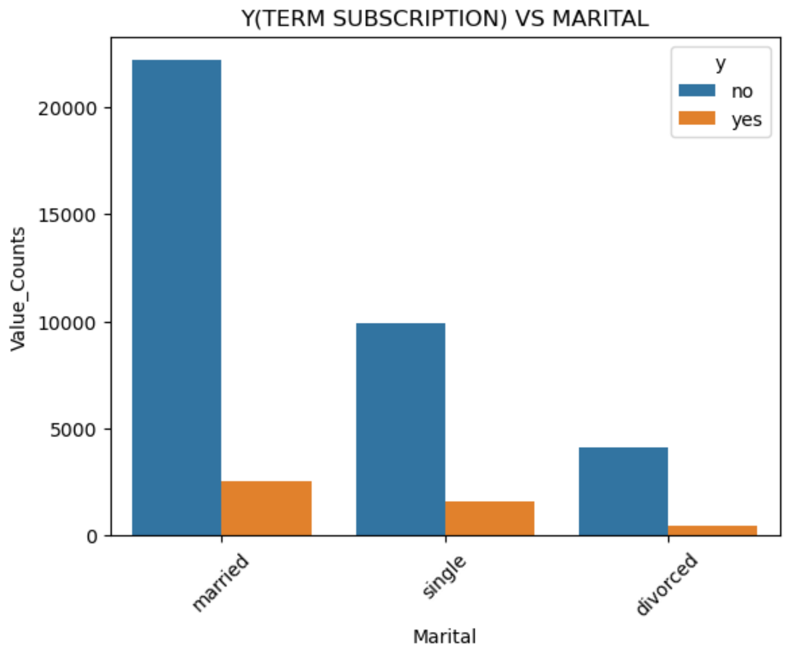

```
Percentage breakdown for education:
education  basic.4y  basic.6y  basic.9y  high.school  illiterate  \
y
no            89.73     91.74     92.17        89.13       77.78
yes           10.27      8.26      7.83        10.87       22.22

education  professional.course  university.degree
y
no                       88.63              86.21
yes                      11.37              13.79
```

Data Glacier
Your Deep Learning Partner

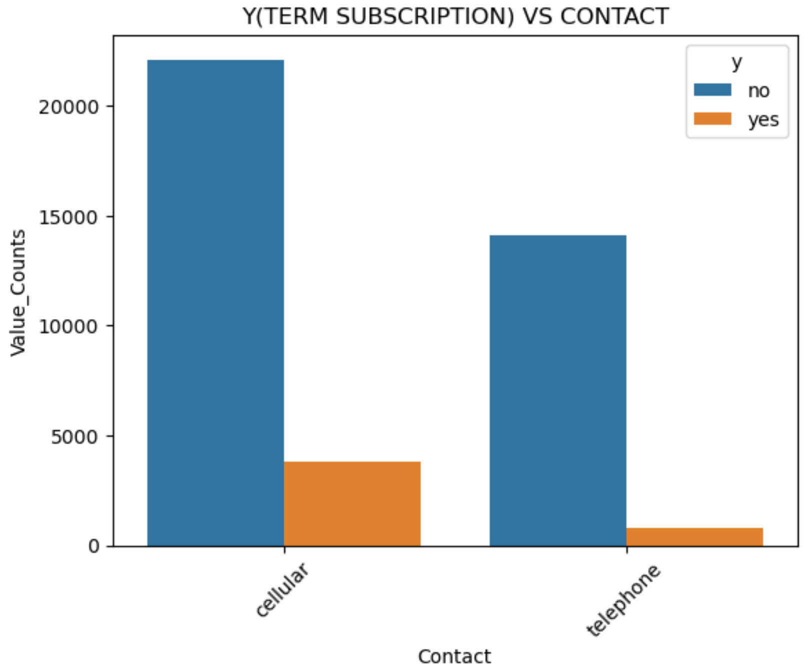
EDA
(Bi-variate Analysis of Categorical Features w.r.t term deposit)

- Single people are more likely to sign up for term deposits compared to those who are married.
- Clients contacted by cell phone have a better chance of subscribing to term deposits compared to those contacted by telephone.



Percentage breakdown for marital:

marital	divorced	married	single
y			
no	89.71	89.81	86.03
yes	10.29	10.19	13.97

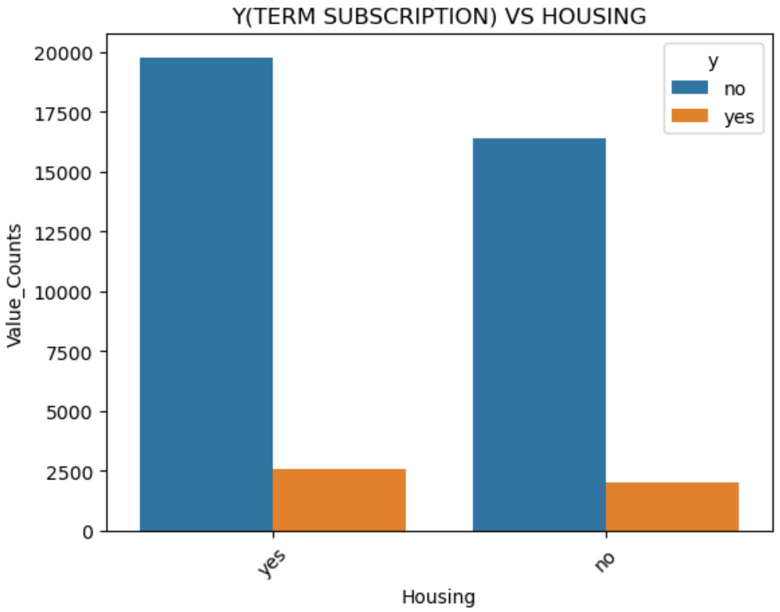


Percentage breakdown for contact:

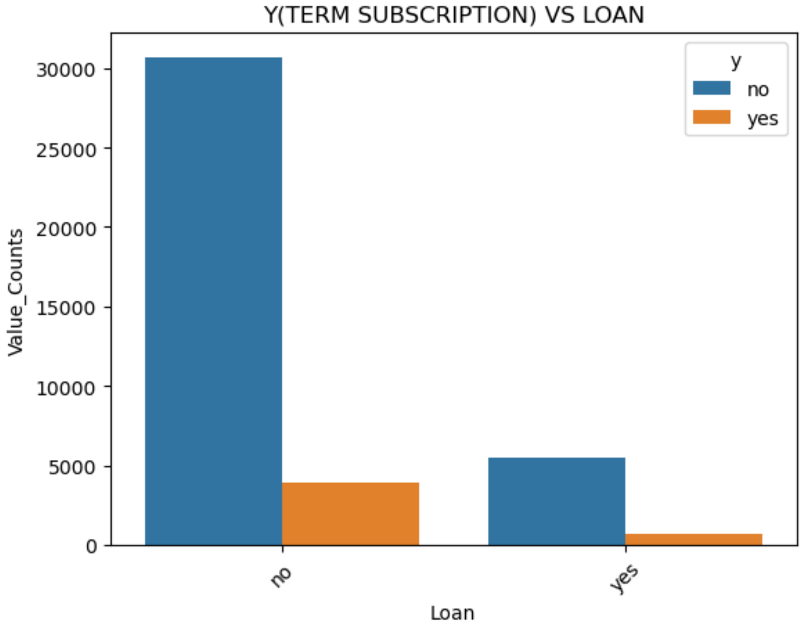
contact	cellular	telephone
y		
no	85.27	94.78
yes	14.73	5.22

EDA
(Bi-variate Analysis of Categorical Features w.r.t term deposit)

- People with personal loans are less likely to sign up for term deposits compared to those with house loans. Overall, having a loan doesn't seem to affect the likelihood of subscribing to term deposits much.




```
Percentage breakdown for housing:
housing     no    yes
y
no       89.13  88.41
yes      10.87  11.59
```

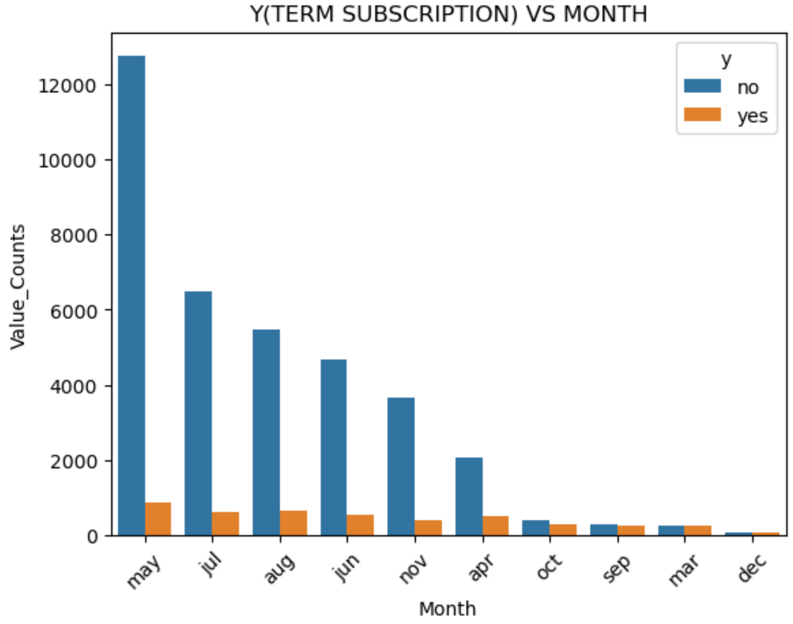



```
Percentage breakdown for loan:
loan     no    yes
y
no    88.68  89.03
yes   11.32  10.97
```

EDA

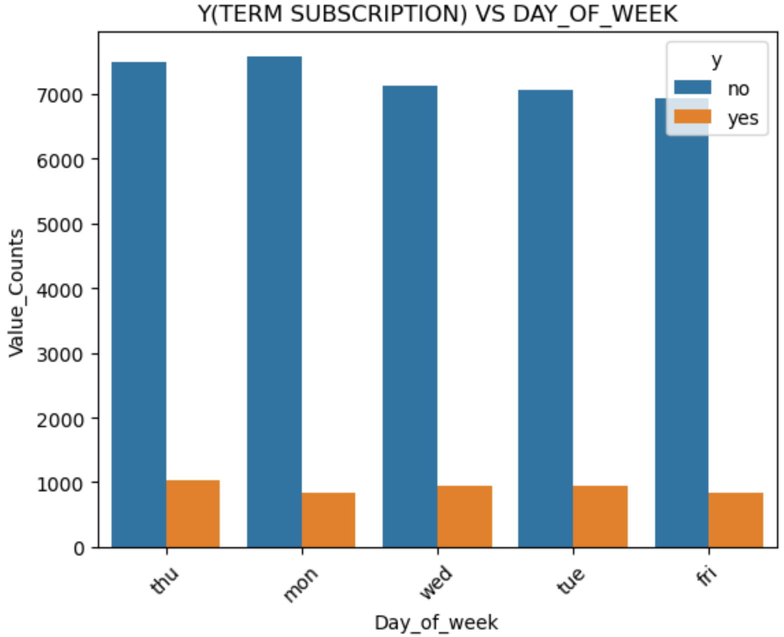
(Bi-variate Analysis of Categorical Features w.r.t term deposit)

- There is more interest in term deposits in December, March, October, and September. Subscriptions happen most often on Thursdays and Tuesdays.



Percentage breakdown for month:

month	apr	aug	dec	jul	jun	mar	may	nov	oct	sep
y										
no	79.5	89.42	51.11	90.99	89.55	49.35	93.52	89.89	55.95	54.98
yes	20.5	10.58	48.89	9.01	10.45	50.65	6.48	10.11	44.05	45.02



Percentage breakdown for day_of_week:

day_of_week	fri	mon	thu	tue	wed
y					
no	89.21	90.01	87.89	88.18	88.39
yes	10.79	9.99	12.11	11.82	11.61

EDA

(Bi-variate Analysis of Numerical Features w.r.t term deposit without duration feature)

- A correlation heatmap shows how closely different things in a group are connected. It helps to quickly spot patterns and decide which variables might be important for further Analysis.
- From the heatmap we can say that no two variables have strong positive or negative relationship.

	age	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y_numeric
age	1.00	0.00	-0.03	0.02	-0.00	0.00	0.13	0.01	-0.02	0.03
campaign	0.00	1.00	0.06	-0.08	0.16	0.13	-0.01	0.14	0.15	-0.07
pdays	-0.03	0.06	1.00	-0.59	0.27	0.08	-0.09	0.30	0.37	-0.33
previous	0.02	-0.08	-0.59	1.00	-0.42	-0.20	-0.05	-0.45	-0.50	0.23
emp.var.rate	-0.00	0.16	0.27	-0.42	1.00	0.78	0.20	0.97	0.91	-0.30
cons.price.idx	0.00	0.13	0.08	-0.20	0.78	1.00	0.06	0.69	0.52	-0.14
cons.conf.idx	0.13	-0.01	-0.09	-0.05	0.20	0.06	1.00	0.28	0.10	0.05
euribor3m	0.01	0.14	0.30	-0.45	0.97	0.69	0.28	1.00	0.95	-0.31
nr.employed	-0.02	0.15	0.37	-0.50	0.91	0.52	0.10	0.95	1.00	-0.35
y_numeric	0.03	-0.07	-0.33	0.23	-0.30	-0.14	0.05	-0.31	-0.35	1.00

Note:- correlation does not imply causation means that just because two things correlate does not necessarily mean that one causes the other

EDA

(Bi-variate Analysis of Numerical Features w.r.t term deposit with duration feature)

- We avoided using the duration feature in the initial analysis because it has a strong influence on whether a term deposit is subscribed. However, the duration is only known after a call is made, which makes it less practical for building a predictive model. Since our goal is to predict outcomes before a call, we will only use the duration feature to compare model performance, not for the actual predictions.

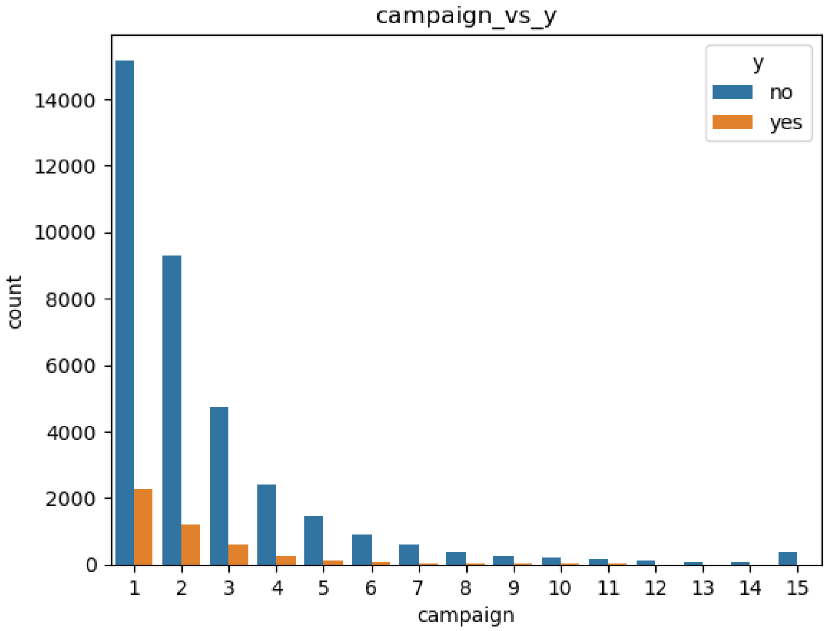
Note:- correlation does not imply causation means that just because two things correlate does not necessarily mean that one causes the other

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y_numeric
age	1.00	0.00	0.00	-0.03	0.02	-0.00	0.00	0.13	0.01	-0.02	0.03
duration	0.00	1.00	-0.07	-0.05	0.02	-0.03	0.01	-0.01	-0.03	-0.04	0.41
campaign	0.00	-0.07	1.00	0.06	-0.08	0.16	0.13	-0.01	0.14	0.15	-0.07
pdays	-0.03	-0.05	0.06	1.00	-0.59	0.27	0.08	-0.09	0.30	0.37	-0.33
previous	0.02	0.02	-0.08	-0.59	1.00	-0.42	-0.20	-0.05	-0.45	-0.50	0.23
emp.var.rate	-0.00	-0.03	0.16	0.27	-0.42	1.00	0.78	0.20	0.97	0.91	-0.30
cons.price.idx	0.00	0.01	0.13	0.08	-0.20	0.78	1.00	0.06	0.69	0.52	-0.14
cons.conf.idx	0.13	-0.01	-0.01	-0.09	-0.05	0.20	0.06	1.00	0.28	0.10	0.05
euribor3m	0.01	-0.03	0.14	0.30	-0.45	0.97	0.69	0.28	1.00	0.95	-0.31
nr.employed	-0.02	-0.04	0.15	0.37	-0.50	0.91	0.52	0.10	0.95	1.00	-0.35
y_numeric	0.03	0.41	-0.07	-0.33	0.23	-0.30	-0.14	0.05	-0.31	-0.35	1.00

EDA

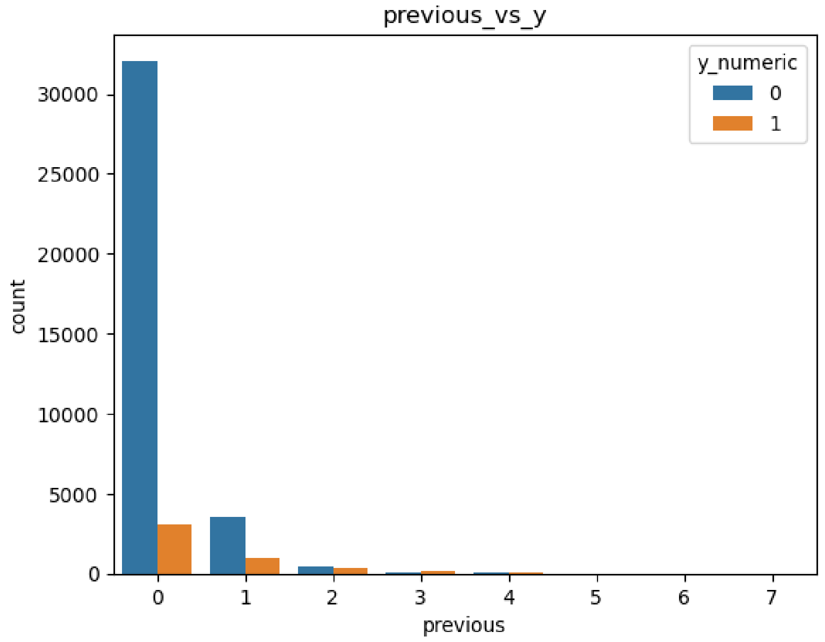
(Bi-variate Analysis of Categorical Features w.r.t term deposit)

- can say that more contacts do not necessarily lead to more term subscriptions.



percentage of term deposit acceptance according to number of calls

y_numeric campaign	0	1
1	86.991171	13.008829
2	88.508488	11.491512
3	89.214205	10.785795
4	90.635706	9.364294
5	92.433796	7.566204
6	92.339545	7.660455
7	93.890675	6.109325
8	95.939086	4.060914
9	93.862816	6.137184
10	95.045045	4.954955
11	93.181818	6.818182
12	97.580645	2.419355
13	95.505618	4.494382
14	98.550725	1.449275
15	98.496241	1.503759

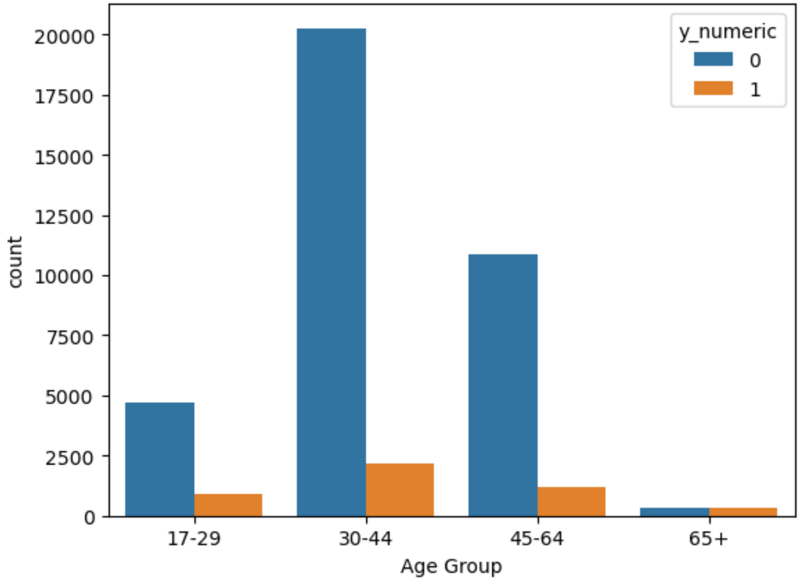


y_numeric previous	0	1
0	91.157004	8.842996
1	78.814684	21.185316
2	53.918919	46.081081
3	40.654206	59.345794
4	45.714286	54.285714
5	27.777778	72.222222
6	40.000000	60.000000
7	100.000000	NaN

EDA

(Bi-variate Analysis of Categorical Features w.r.t term deposit)

- We can say that people aged between 30 and 64 have fewer term deposits compared to other age groups.
- Most of the calls were made to individuals aged between 30 and 64.



y_numeric	0	1
Age Group		
17-29	4718	910
30-44	20268	2155
45-64	10847	1221
65+	344	305

Age Group	
17-29	13649
30-44	56348
45-64	30984
65+	1289

Name: campaign, dtype: int64

Total Number of calls made according to age brackets.

EDA Summary

Key Findings:

- None of the variables are strongly connected to the target (whether someone subscribed).
- Retirees and single people are more likely to subscribe to term deposits.
- People tend to subscribe more when contacted by phone.
- The highest number of subscriptions happened in December, March, October, and September, mostly on Thursdays and Tuesdays.

Also, making more calls doesn't guarantee more subscriptions. But focusing on these patterns could help increase term deposits.

Recommended Models

Since our problem is a binary classification (yes or no) with imbalanced data, I have chosen the following models:

- **Logistic Regression**: A simple, easy-to-understand model. However, it might not perform well with complex relationships, so we will use this as a baseline.
- **Decision Tree**: Handles both numeric and categorical values and helps identify important features, but it may overfit the data.
- **Random Forest (Ensemble Model)**: Solves the overfitting issue in decision trees and manages imbalanced data by adjusting class weights.
- **Gradient Boosting Machines (GBM) or XGBoost**: Like Random Forest but more accurate. These models handle non-linear relationships well and are effective for imbalanced data.

There are other models, but these are better suited for handling imbalanced data. As mentioned earlier, I will use resampling techniques to balance the data. We will evaluate model performance using metrics like F1 score, precision, recall, etc. Based on these, we will fine-tune the models and decide which one to use for the final predictions. We will also compare the performance of models with and without using the **duration** feature.

Thank You