Data Glacier

Data Science

Project Report : Bank Marketing(Campaign)

Group Name: Project Group 1

Members:

No	Name	Email	Country	Specialization
1	Preeti Verma	vermapreeti.dataanalyst@gmail.com	Canada	Data Science
2	Thanuja Modiboina	thanujayadav953@gmail.com	UK	Data Science
3	Abishek James	abishekjames1998@gmail.com	Ireland	Data Science

Report date: 08-05-2023
Internship Batch: LISUM19
Data intake by: Abishek James
Data intake reviewer: Data Glacier
Data storage location: https://github.com/abishekjames/Data-Glacier-project/tree/main/Week10

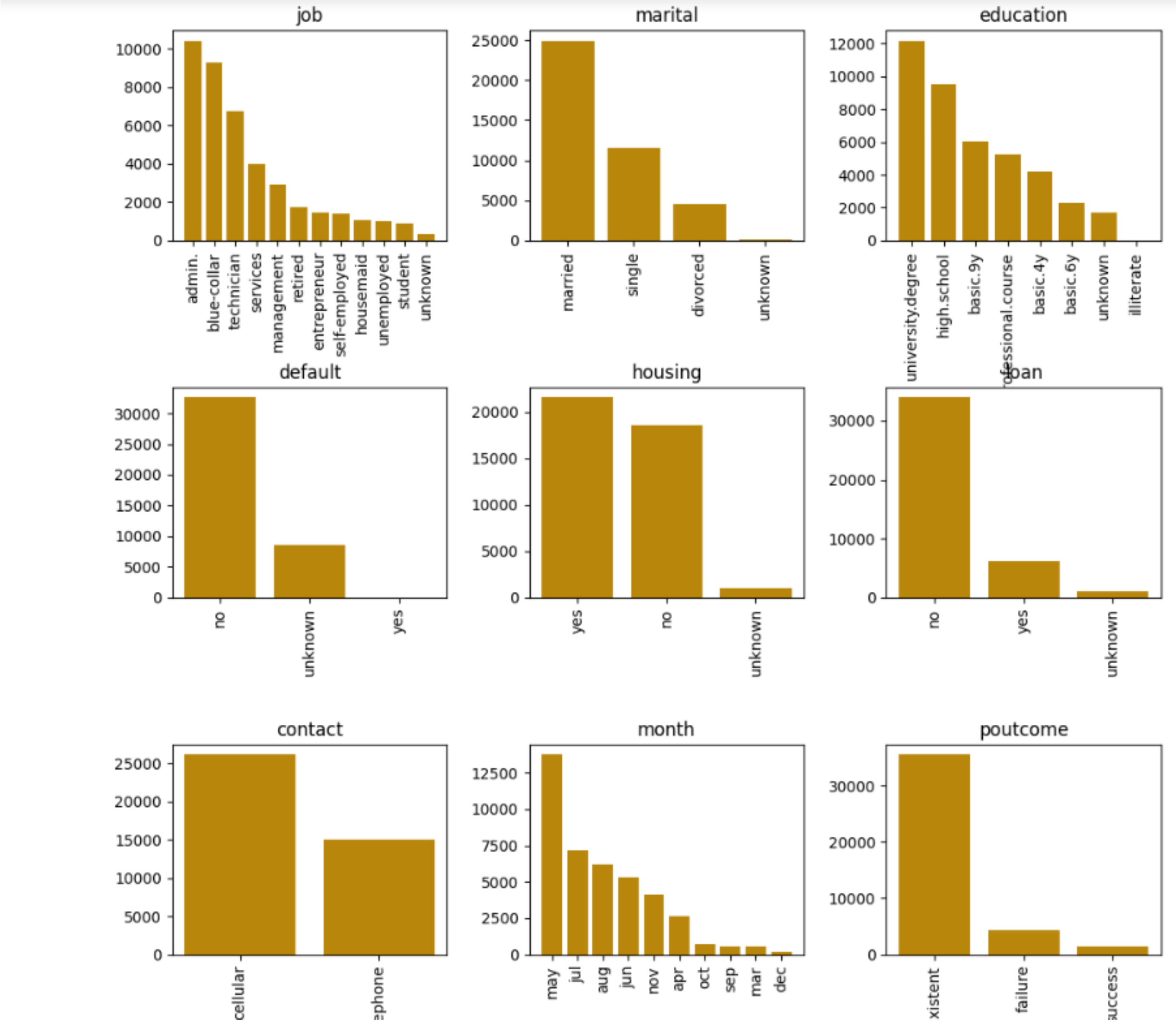
Problem Description:

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). This is an application of company's marketing data.

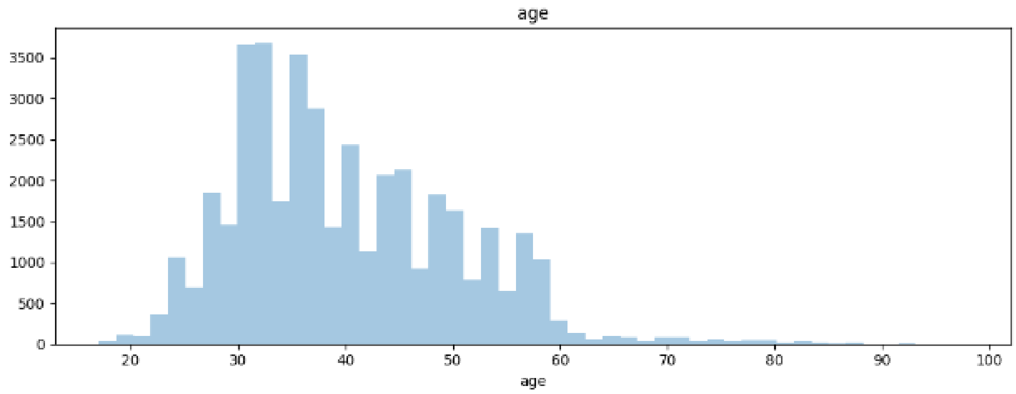
EDA

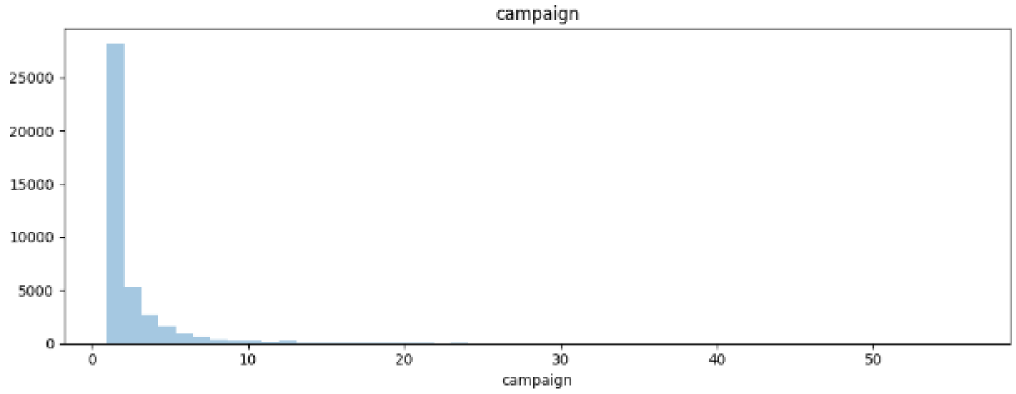
To obtain a better understanding of the dataset, the distribution of key variables and the relationships among them were plotted.

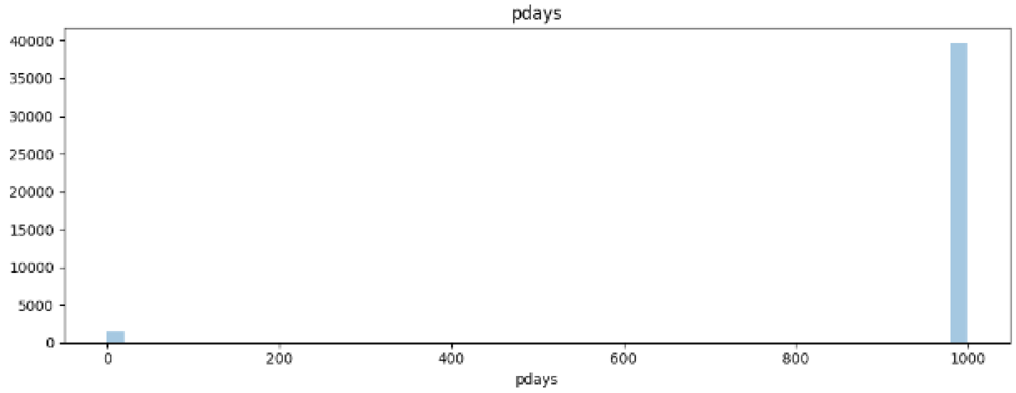
1.1 Exploring Categorical values

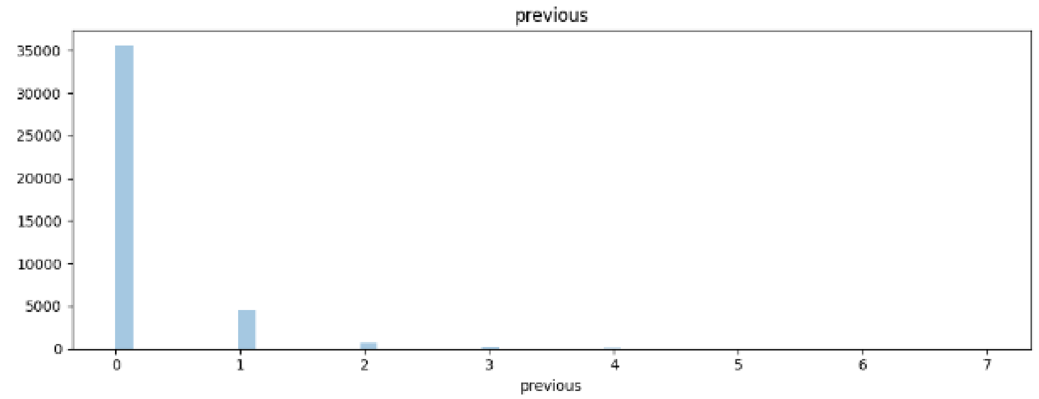


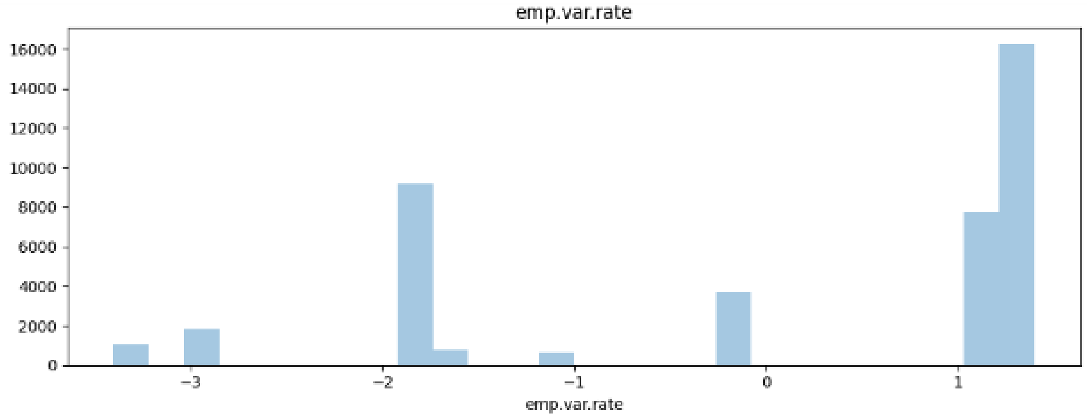
1.2 Exploring Numerical features

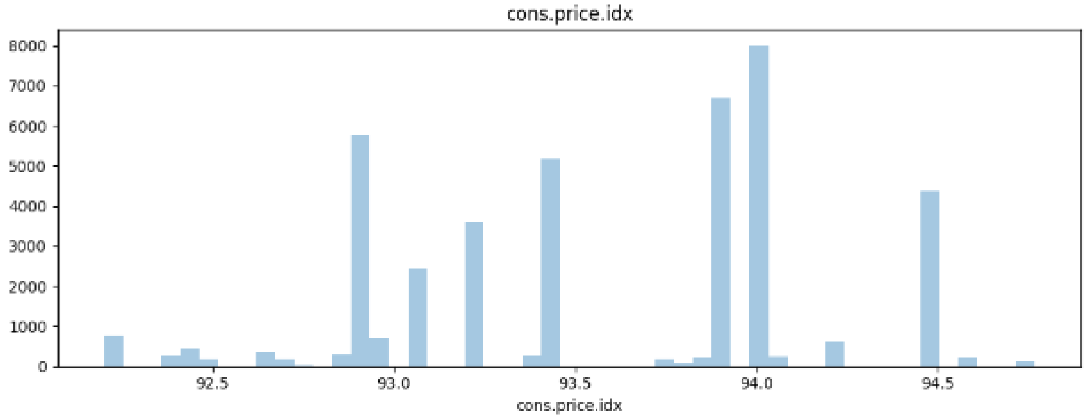


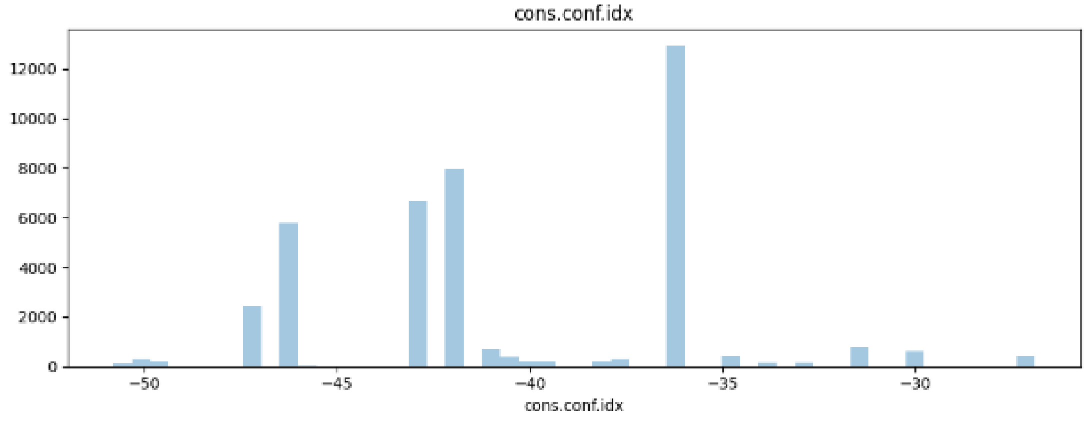


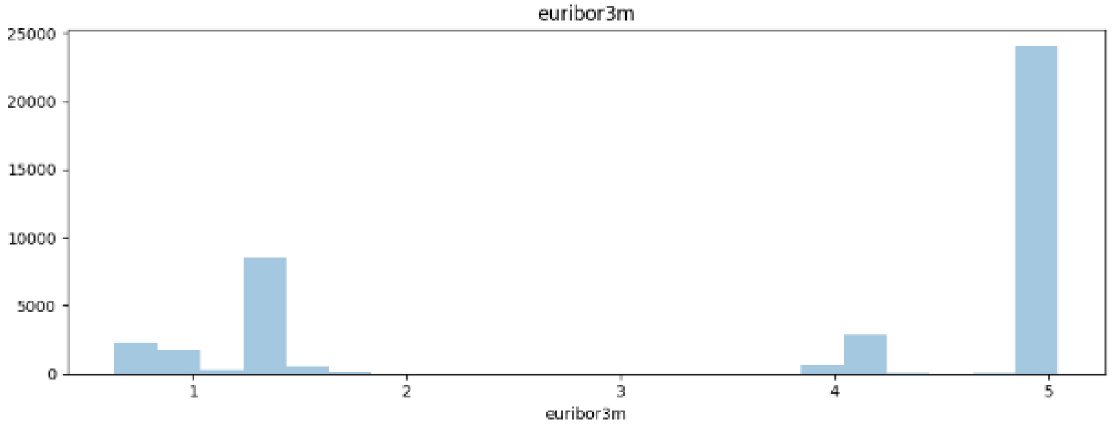


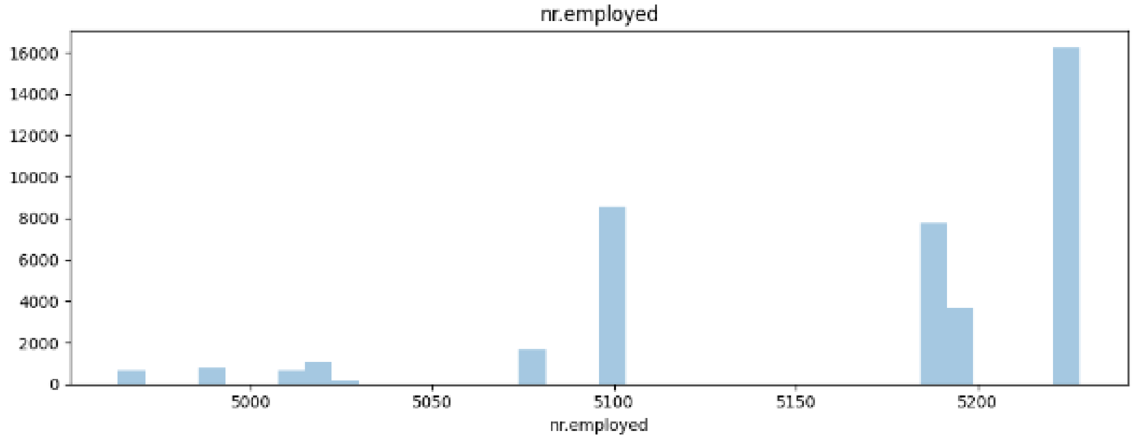

previous
35000
30000
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
6
7
previous


emp.var.rate
16000
14000
12000
10000
8000
6000
4000
2000
0
−3
−2
−1
0
1
emp.var.rate

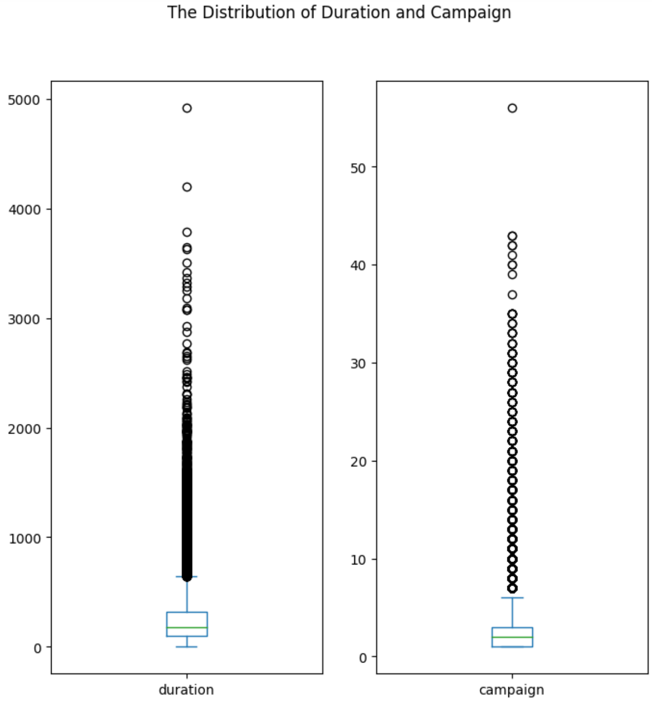

cons.price.idx
8000
7000
6000
5000
4000
3000
2000
1000
0
92.5
93.0
93.5
94.0
94.5
cons.price.idx


cons.conf.idx
12000
10000
8000
6000
4000
2000
0
−50
−45
−40
−35
−30
cons.conf.idx





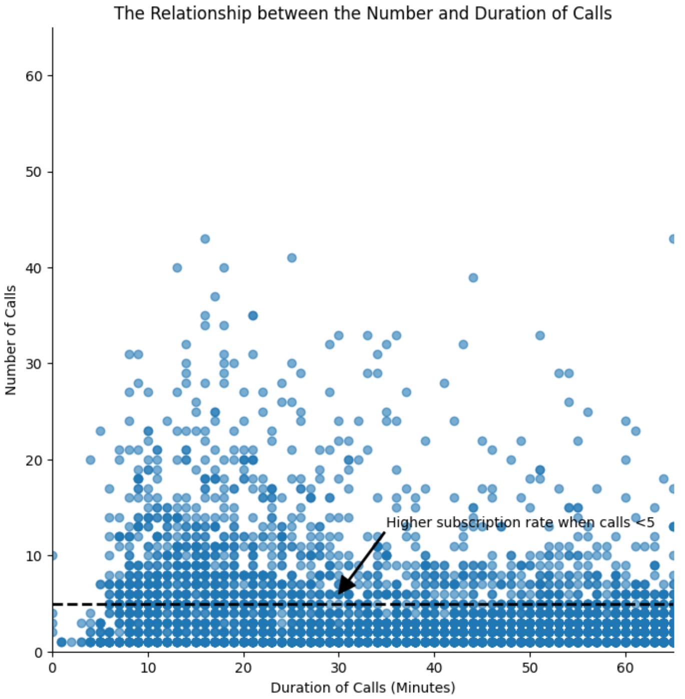
1.3 Visualize the distribution of 'duration' & 'campaign'



The distribution of duration: As observed from the box plot, the duration of contact has a median of 3 minutes, with an interquartile range of 1.73 minutes to 5.3 minutes. The left-skewed boxplot indicates that most calls are relatively short. Also, there are many outliers ranging from 10 minutes to 40 minutes, which are worth further study.

The distribution of campaign: About half of the clients have been contacted by the bank for the second time, while 25% was first introduced to the term deposit. Most clients have been reached by the bank for one to three times, which is reasonable. However, some clients have been contacted by as high as 58 times, which is not normal. These clients may have some special needs that require frequent contact.

1.4 Visualize the relationship between 'duration' & 'campaign': with response result

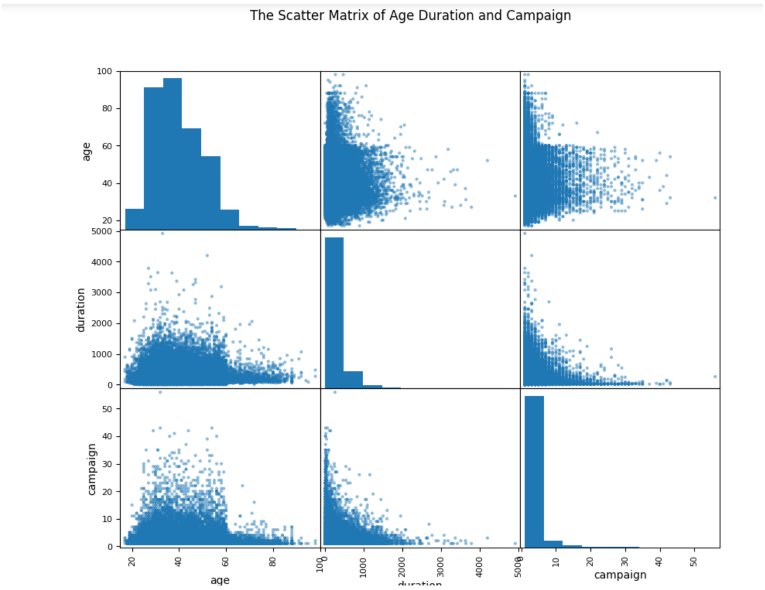


In this scatter plot, clients subscribed to term deposits are denoted as "yes" while those did not are denoted as "no".

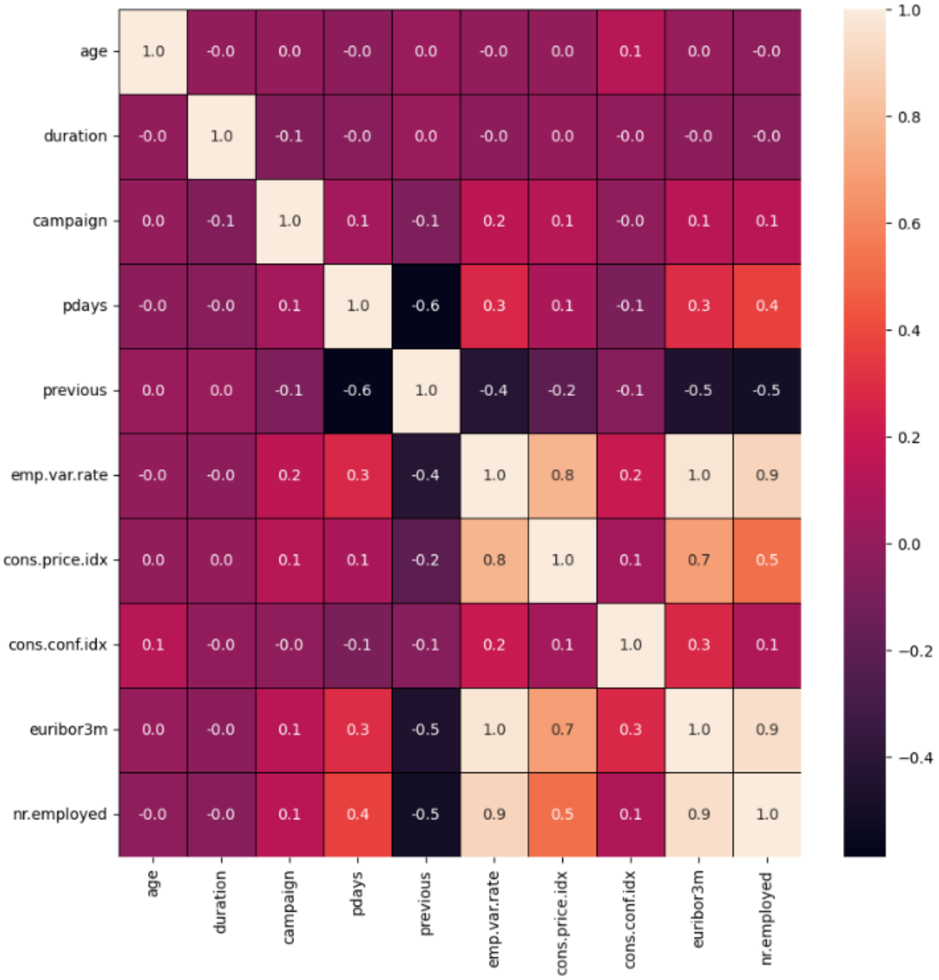
As we can see from the plot, “yes” clients and “no” clients are forming two relatively separate clusters. Compared to “no” clients”, “yes” clients were contacted by fewer times and had longer call duration. More importantly, after five campaign calls, clients are more likely to reject the term deposit unless the duration is high. Most “yes” clients were approached by less than 10 times.

This suggests that the bank should resist calling a client for more than five times, which can be disturbing and increase dissatisfaction.

1.5Scatter matrix of ‘age’ ,’duration’, ’campaign’



1.6 Correlation matrix



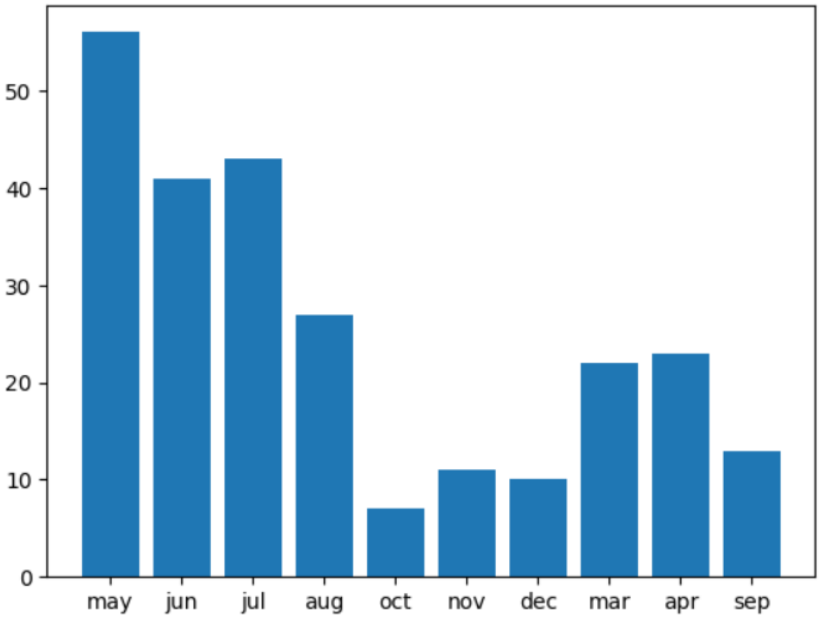
The scatter matrix does not reveal any clear relationship among age, duration and campaign.

To investigate more about correlation, a correlation matrix was plotted with all qualitative variables. Clearly, “campaign outcome” has a strong correlation with “duration”, a moderate correlation with “previous contacts”, and mild correlations between “month of contact” and “number of campaign”. Their influences on campaign outcome will be investigated further in the machine learning part.

1.7 Data Visualization

Since we have much numerical data, let's keep our plots much targeted towards our machine learning models. Also let's figure out which feature importance’s and prune away least important ones.

1. Campaign vs Month

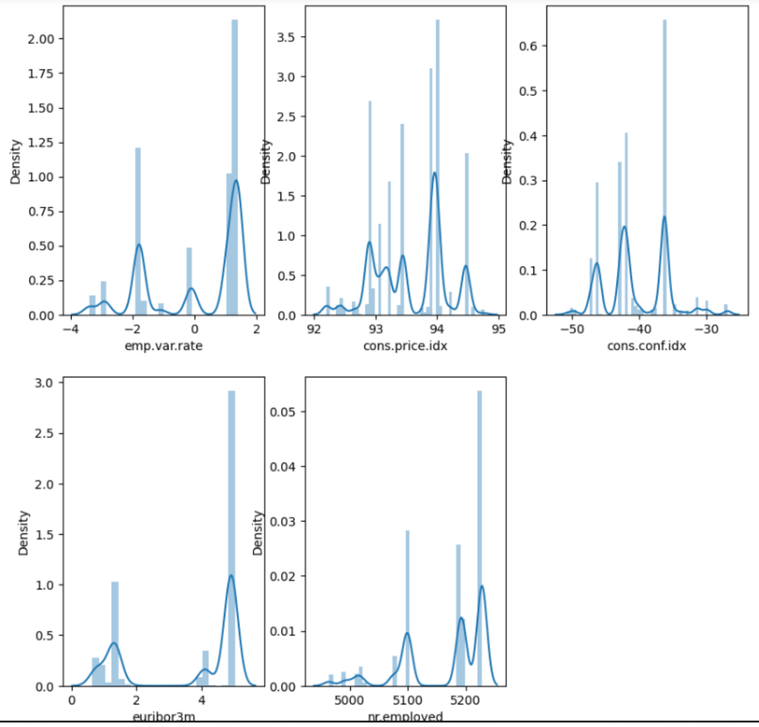


• We can see the campaign were mostly concentrated in the starting of the bank period (May, June and July)

• Usually education period starts during that time so there is a possibility that parents make deposits in the name of their children

• They also have made their campaign in the end of the bank period.

2. Distribution of Quarterly Indicators



Insights:

• We can see there is a high employee variation rate which signifies that they have made the campaign when there were high shifts in job due to conditions of economy

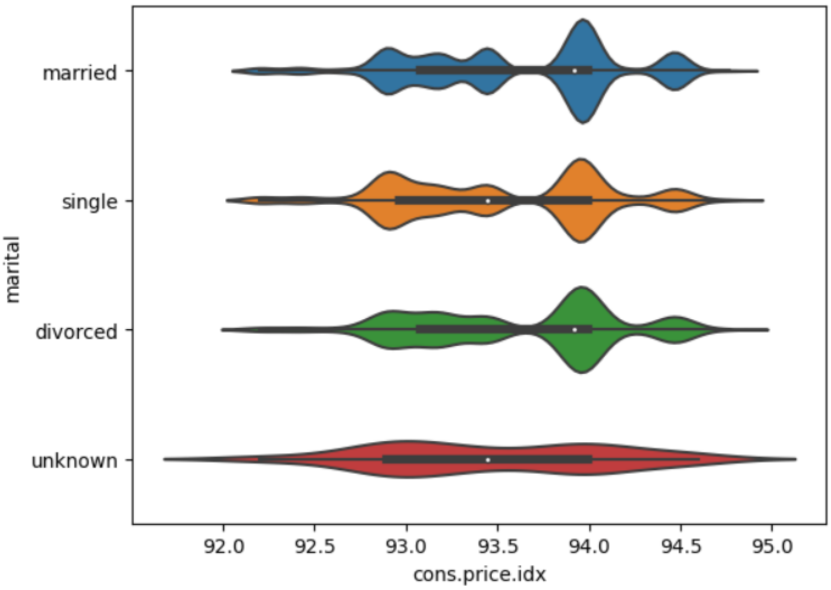
• The Consumer price index is also good which shows the leads were having good price to pay for goods and services may be that could be the reason to stimulate these leads into making a deposit and plant the idea of savings

• Consumer confidence index is low as they don't have much confidence on the fluctuating economy

• The 3-month Euribor interest rate is the interest rate at which a selection of European banks lends one another funds denominated in euros whereby the loans have a maturity of 3 months. In our case the interest rates are high for lending their loans

• The number of employees were also at peak which can increase their income index that could be the reason the campaign targeted the leads who were employed to make a deposit

3. Marital Status vs Price index

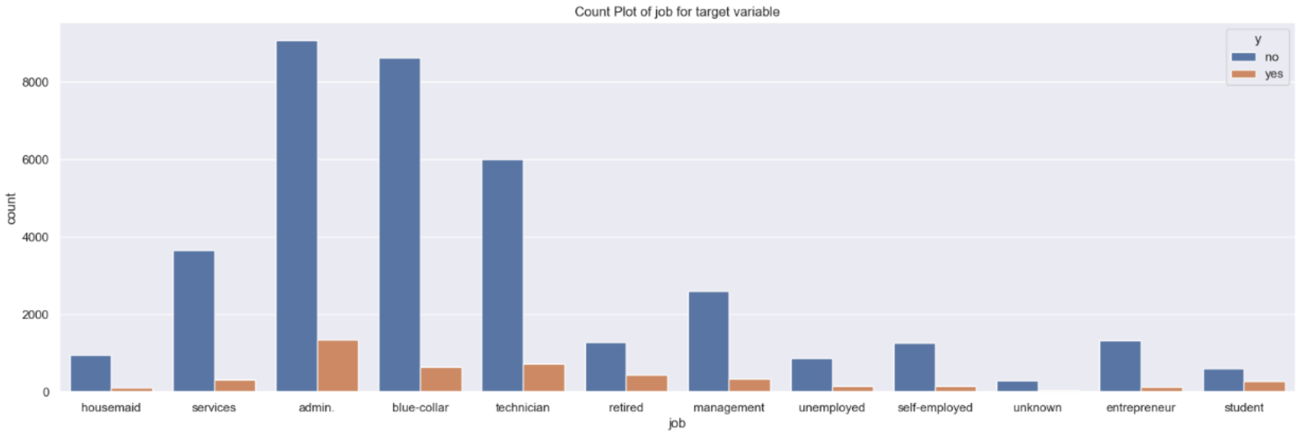


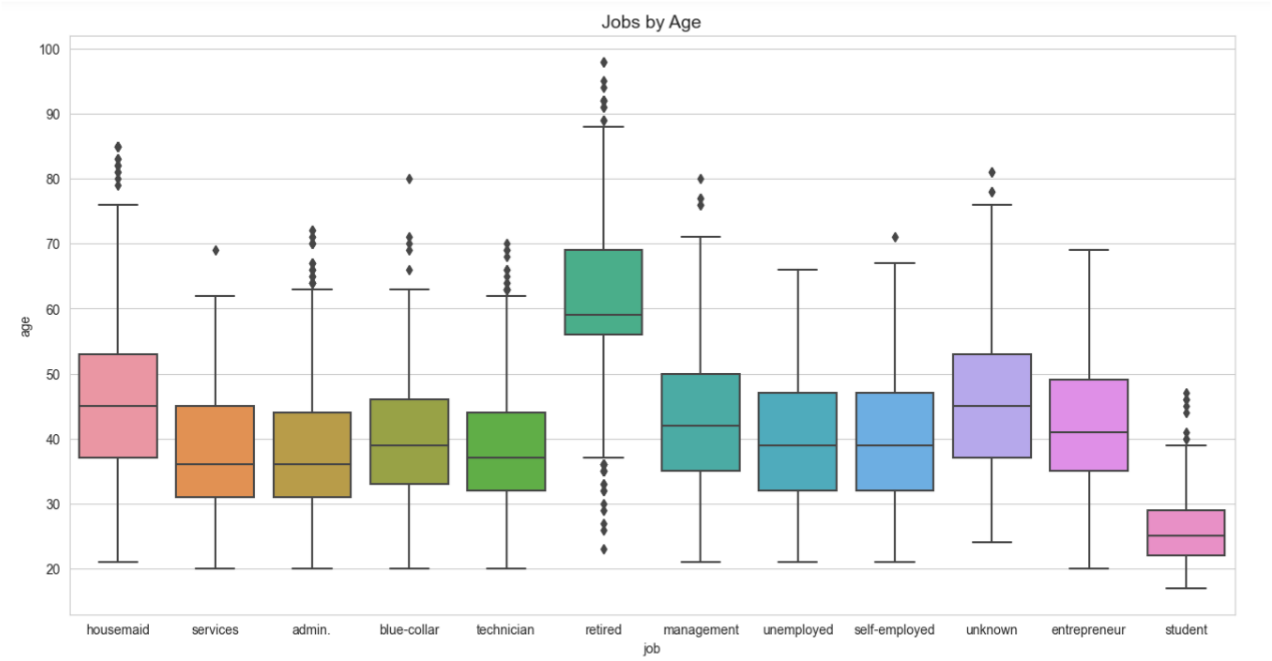
Insights:

- There are very minute differences among the price index
- Married leads have considerably have an upper hand as they have index contributing as couple

Customer Information

1.Job



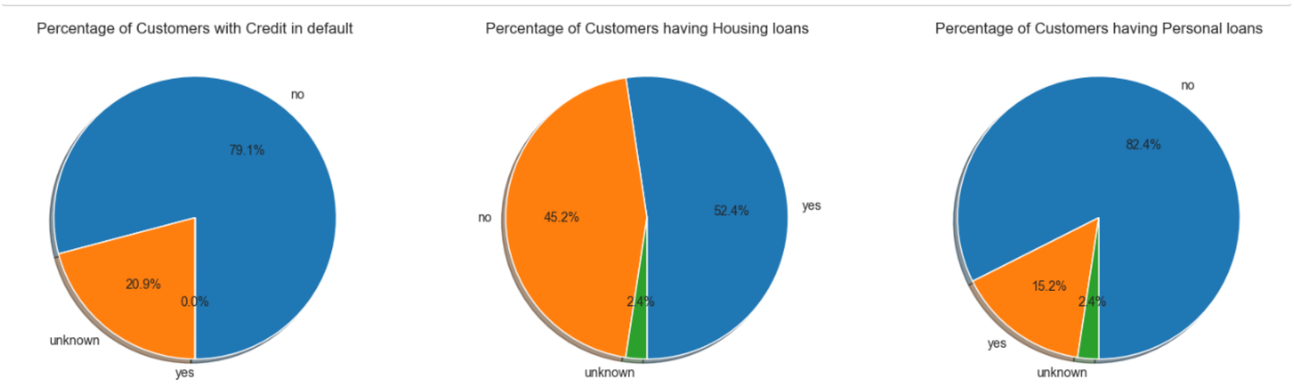

Jobs by Age

Observation:

- Top contacted clients are from job type: 'blue-collar', 'management' & 'technician'.

- Retired people are mostly at the age of above 55 and most students are at the age of 20s.

2.Default, Housing and Loan.


Percentage of Customers with Credit in default

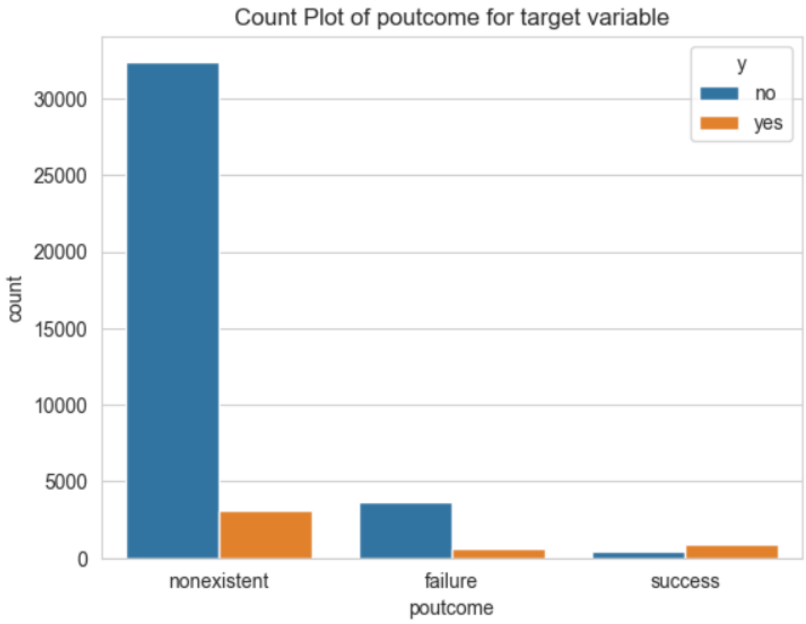
Percentage of Customers having Housing loans

Percentage of Customers having Personal loans

Observation:

- People with Credit in default account for the majority of the total customers (98.2%).

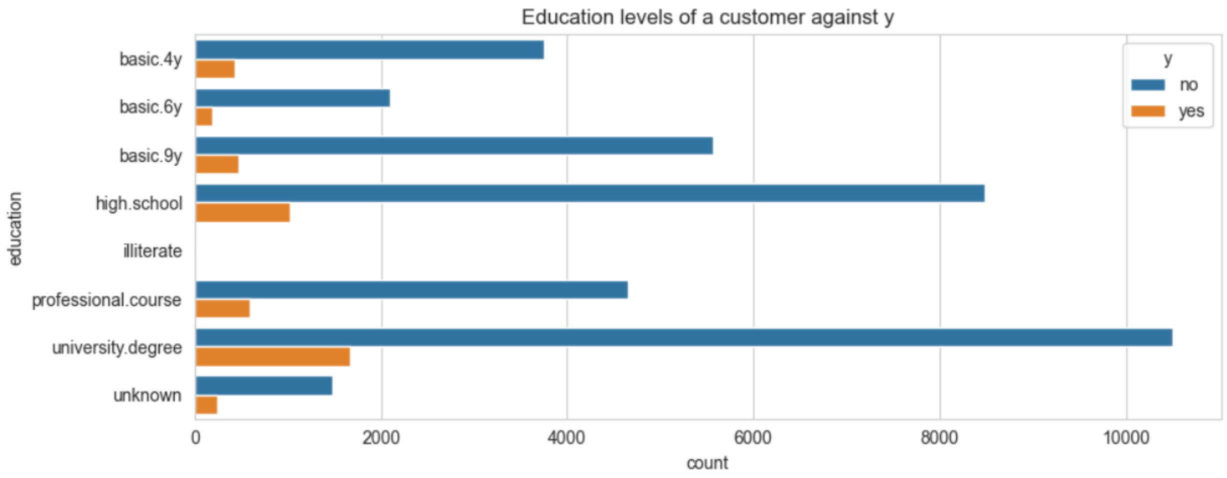
3.Poutcome



Observation:

- Most of the clients contacted have previous outcome as 'unknown'.

4.Education

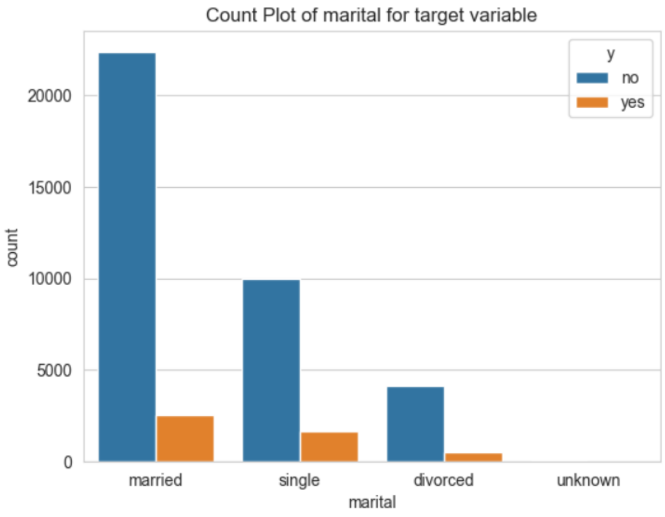


Observation:

- Most of the people who are contacted have tertiary or secondary education.

- Customers with university degree have subscribed to the term deposit more

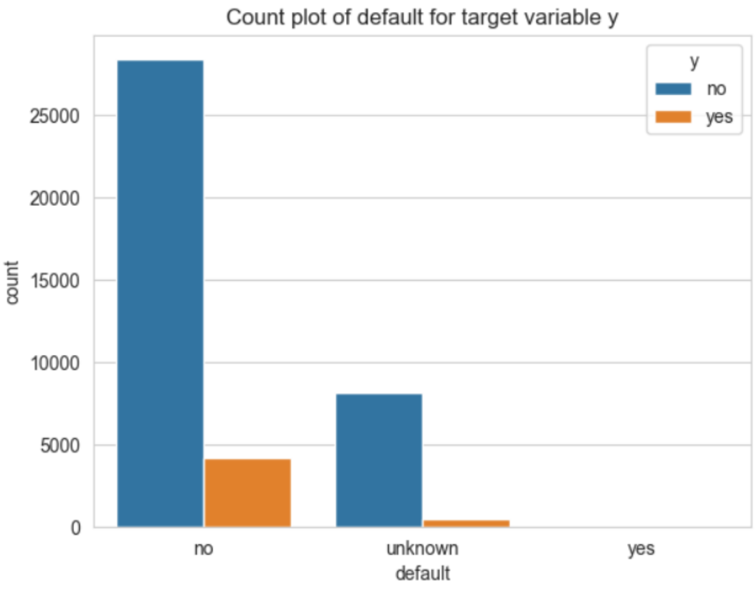
5.Marital



Observation:

married and single customers are the majority of the customer base and comparatively married customers have taken the term deposit

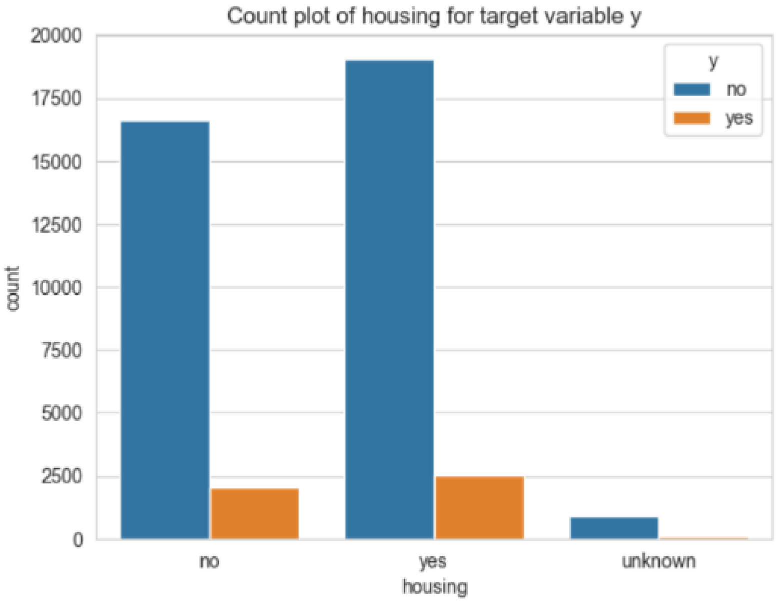
6.Default



Observation:

- Very few clients are contacted who are defaulter

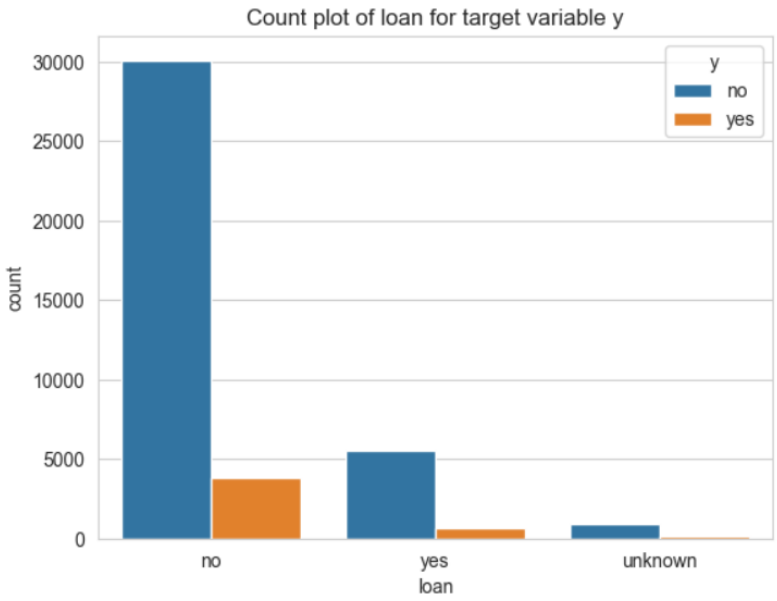
7.Housing



Observation:

- Number of customers who have subscribed to the term deposit is comparatively more for those with housing loan

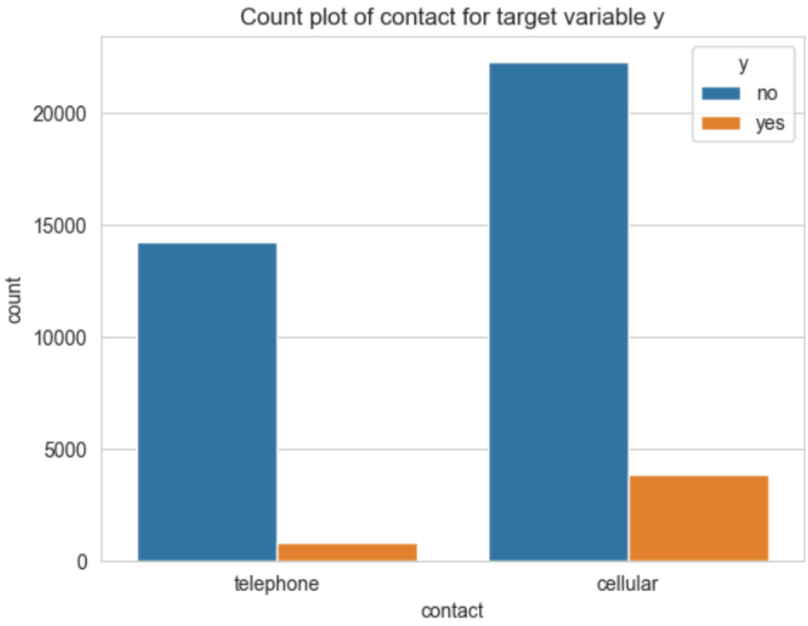
8.Loan



Observation:

- As seen for default variable, less client are contacted who have loan.

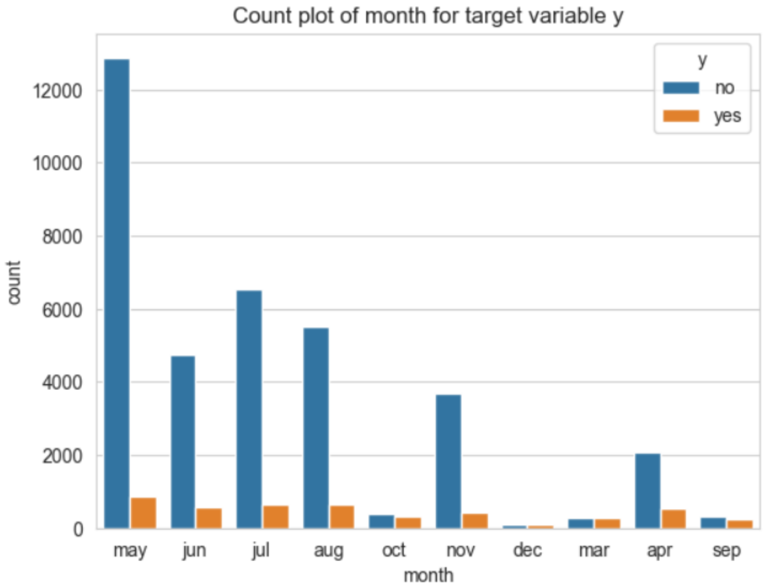
9.Contact



Observation:

- Most of the people are contacted through cellular

10.Month



Observation:

- Most of the people are contacted through cellular

- Regarding months, the highest volume of customers occurs during May. However, this month also saw to the lowest conversion rate, meaning the promoted customers choose to reject the subscription. Hence, the bank should reallocate resources to other months that effective rate is high, such as March, December, September, and October. Note that the month of December requires further investigation due to small sample size.

→ Contact time could be a potential variable to predict.

Recommendation

- First we did EDA and figured out that there is no null values for the data, and the data is imbalanced, where “no” is the majority class.

- The 'poutcome' - outcome of the previous marketing campaign has the greatest influence on the current campaign's outcome.

- Whether or not the customer has housing loan and personal loan could greatly affect the campaign's outcome as well.

- The customers occupation and 'maritial' status are also among the factors that significantly impact the results.

- The 'duration' of the last contact is also important for our prediction and it is the most important feature while 'education' is the least important feature.

- 'Month' of May have seen the highest number of clients contacted but have the least success rate. Highest success rate is observed for end month of the financial year as well as the calendar year. So one can say that our dataset have some kind of seasonality.

- When visualized 'age' in groups, it is found that clients with age less than 30 and greater than 60 are less contacted through the campaign but have a higher success rate.

- For each encoding of categorical data, we have used models to compare which encoding would work better.