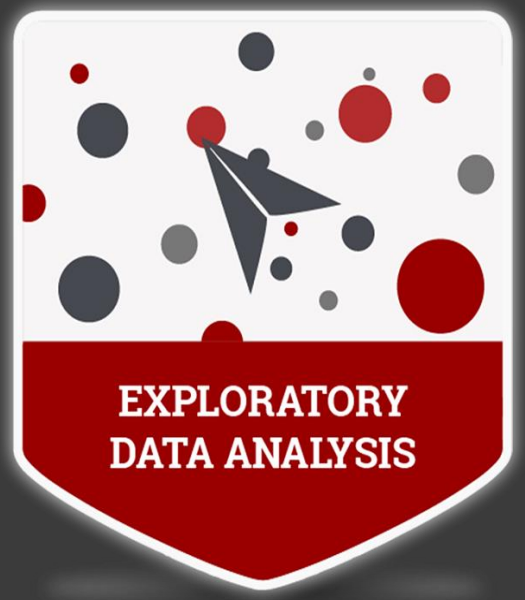
Data Glacier

Your Deep Learning Partner



Bank Marketing Campaign

Preeti Verma	vermapreeti.dataanalyst@gmail.com	Data Science
Thanuja	thanujayadav953@gmail.com	Data Science
Abishek James	abishekjames1998@gmail.com	Data Science

15 May 2023

Agenda

- Problem Statement
- Datasets Overview
- EDA
- EDA Summary
- Recommendations

Problem Statement

- Finding out the characteristics that are helping Bank to make customers successfully subscribe for deposits, which helps in increasing campaign efficiently and selecting high value customers.
- The purpose is increasing the campaign effectiveness by identifying the main characteristics that affect the success based on a handful of algorithms that we will test (e.g. Logistic Regression, Random Forests, Decision Trees and others).
- The aim of the marketing campaign was to get customers to subscribe to a bank term deposit product. Whether they did this or not is variable ‘y’ in the data set. The bank in question is considering how to optimize this campaign in future.

Dataset Overview

Datasets for the analysis:

- bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010).
- bank-additional.csv with 10% of the examples (4119), randomly selected 20 inputs(+1 target variable).

Assumptions:

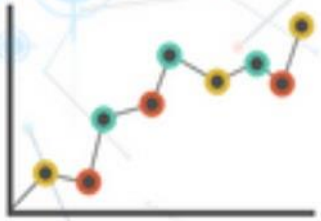
- For each observation, the dataset records have input variables that stand for both qualitative and quantitative attributes of the customer, such as age, job, housing and personal loan status, account balance, and the number of contacts.
- There is no missing value in this dataset. Nevertheless, there are values like “unknown” which are helpless just like missing values. Thus, these ambiguous values are changing to null values.





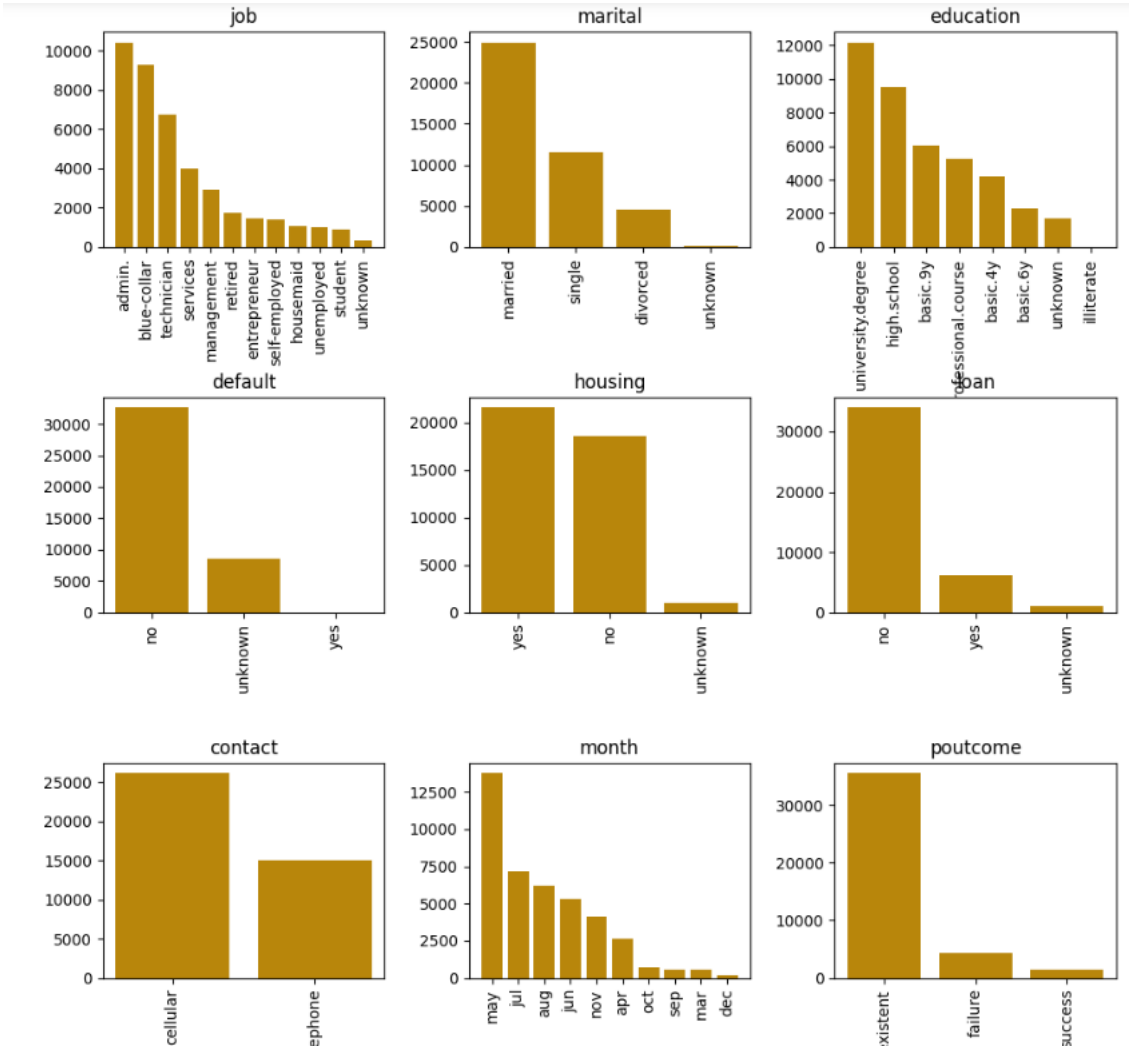

Exploratory Data Analysis




Categorical features

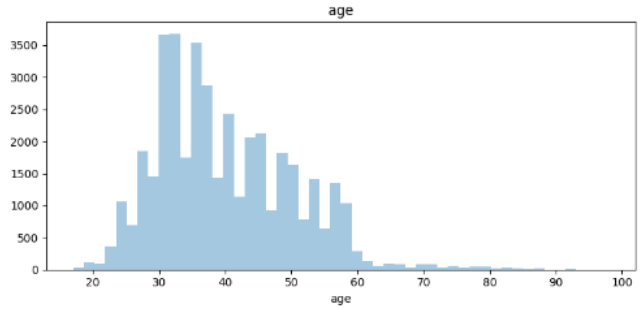
- To obtain a better understanding of the dataset, the distribution of key variables and the relationships among them were plotted. These variables are 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'poutcome'

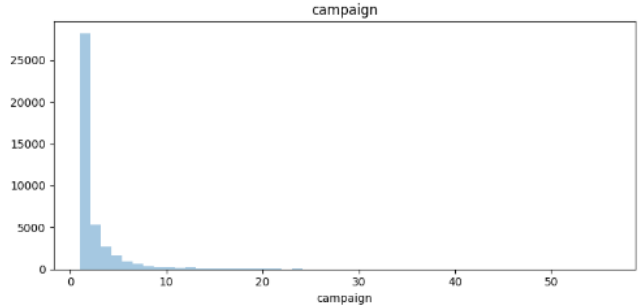


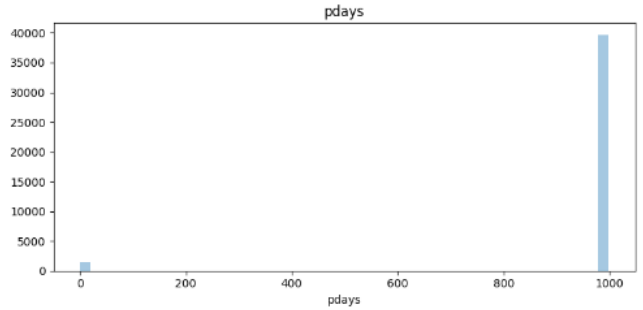
Data Glacier

Numerical features

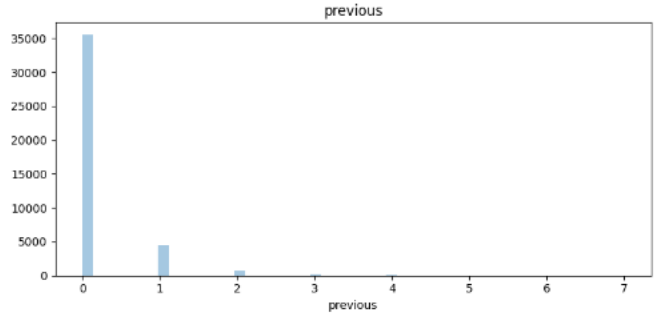
'age','campaign', 'pdays', 'previous', 'emp.var.rate', 'cons.price.idx','cons.conf.idx','euribor3m'

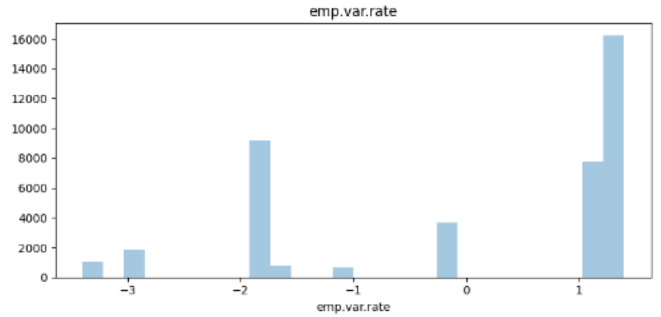


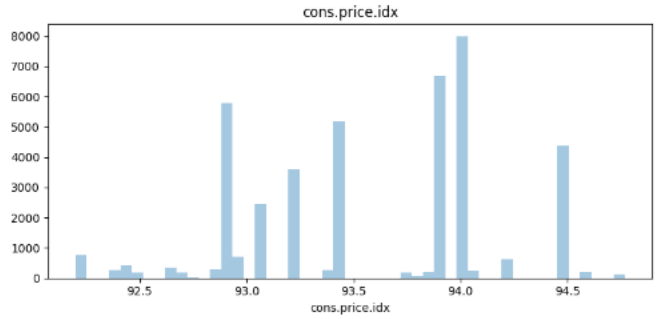


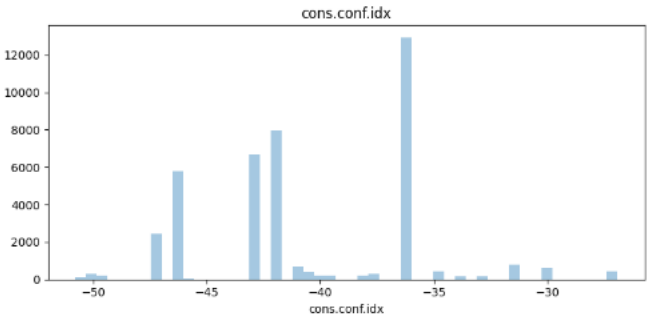


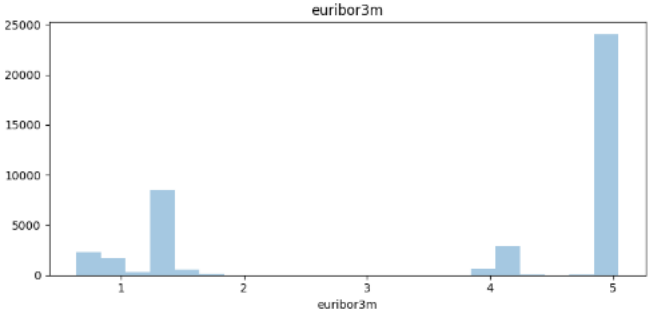
previous

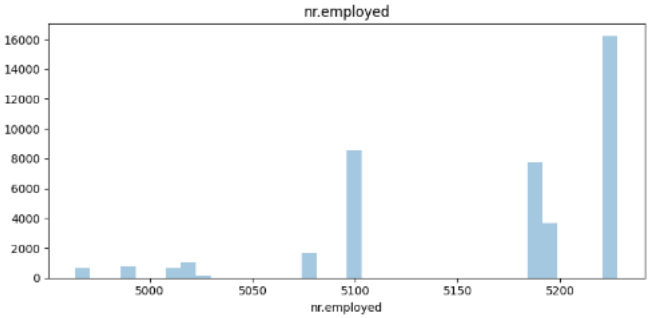








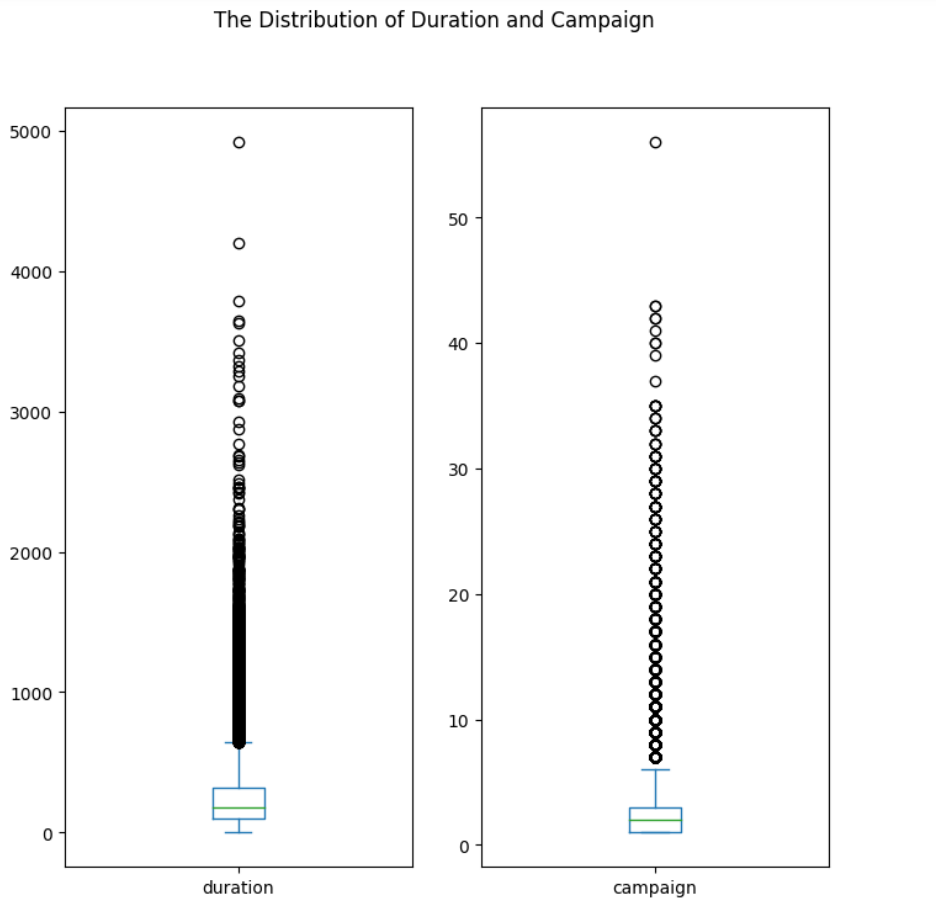




Visualize the distribution of ‘duration’ & ‘campaign

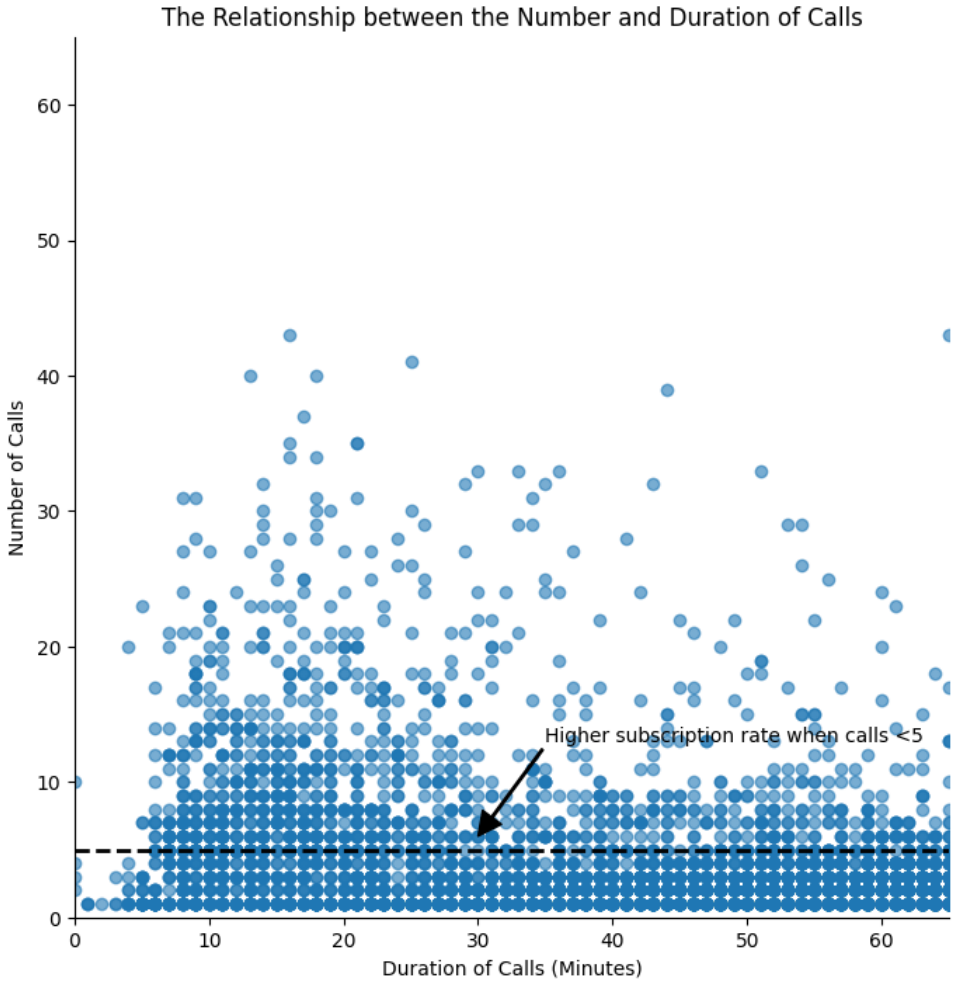
The distribution of duration: As observed from the box plot, the duration of contact has a median of 3 minutes, with an interquartile range of 1.73 minutes to 5.3 minutes. The left-skewed boxplot indicates that most calls are relatively short. Also, there is a large number of outliers ranging from 10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



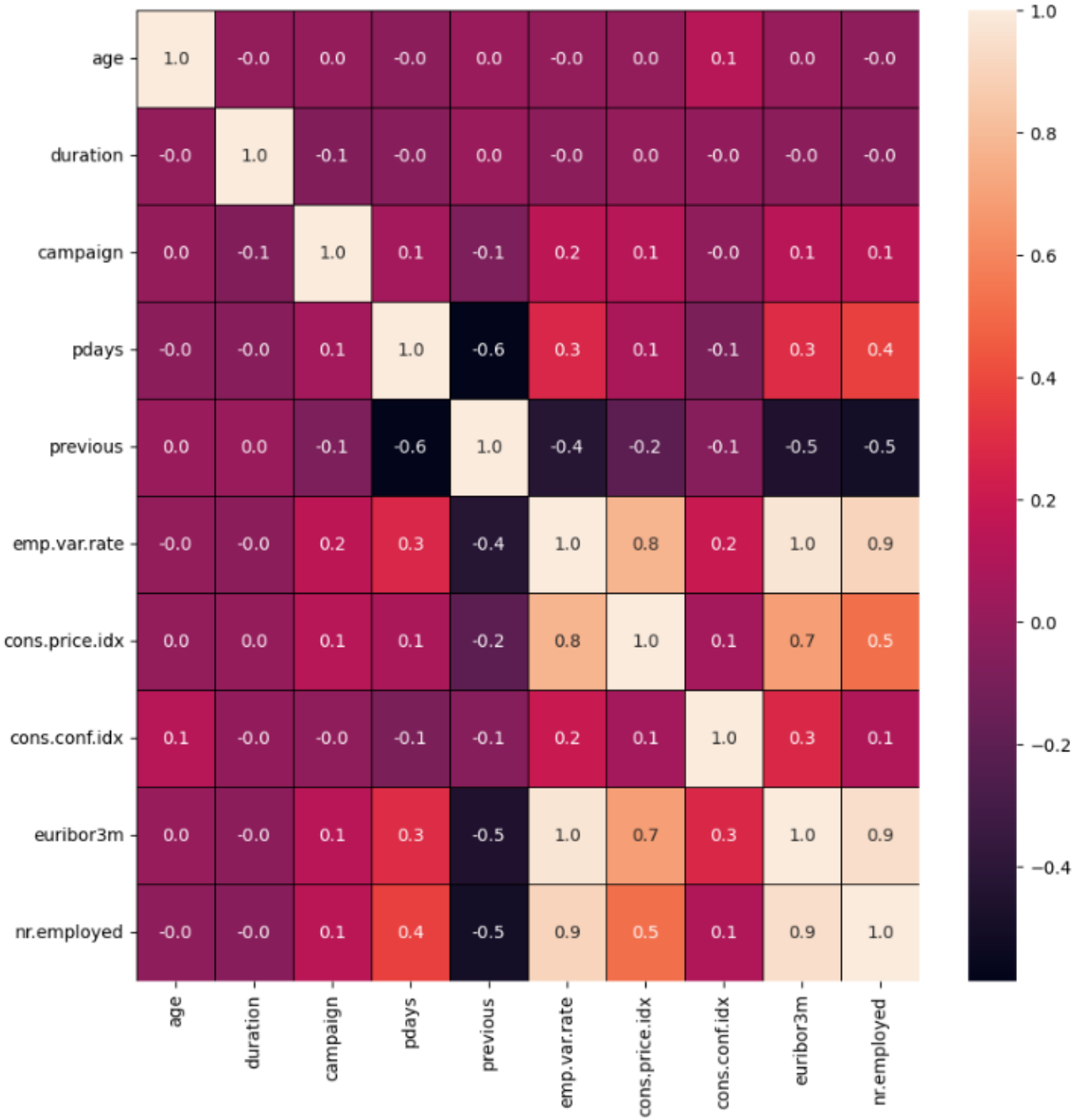
Visualize the relationship b/w ‘duration’ & ‘campaign with response result

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



Correlation between variables

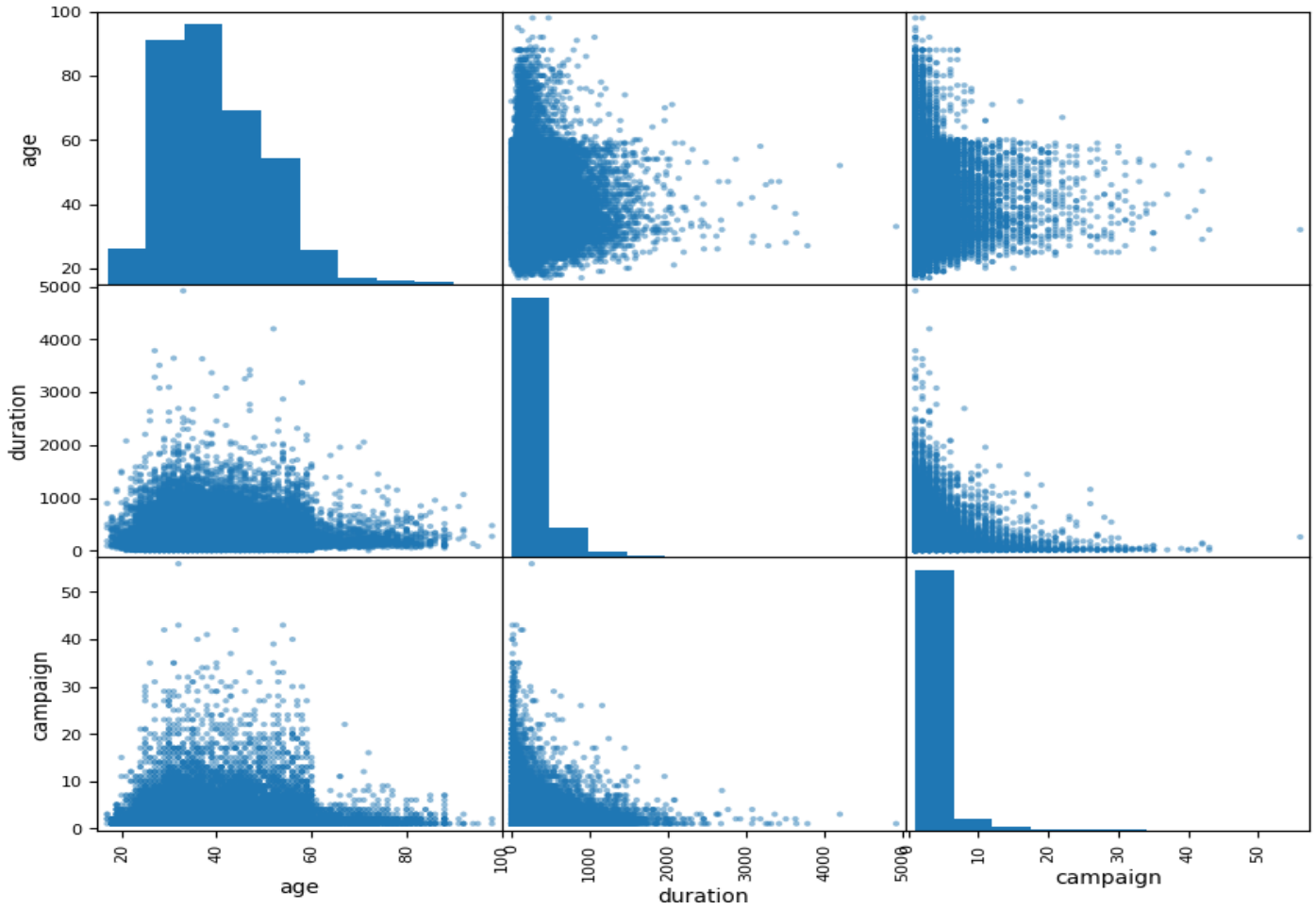
- No correlaiton r=0
- Very weak correlation: r<20
- Weak correlation: between 0.20-0.49
- Moderate correlation: between 0.5-0.79
- Strong correlation: between 0.8-0.99
- Perfect correlation: r=1



Data Glacier
Your Deep Learning Partner

Scatter Matrix

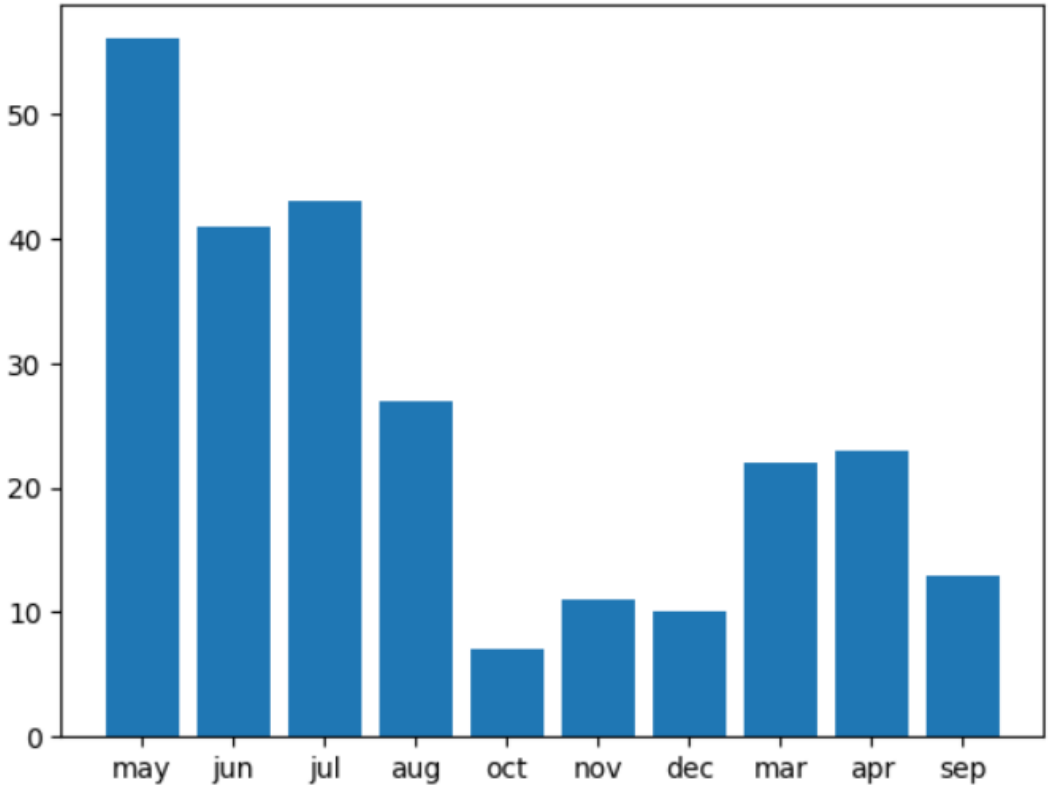
- The scatter matrix does not reveal any clear relationship among age,, duration and campaign.



Campaign v/s Month

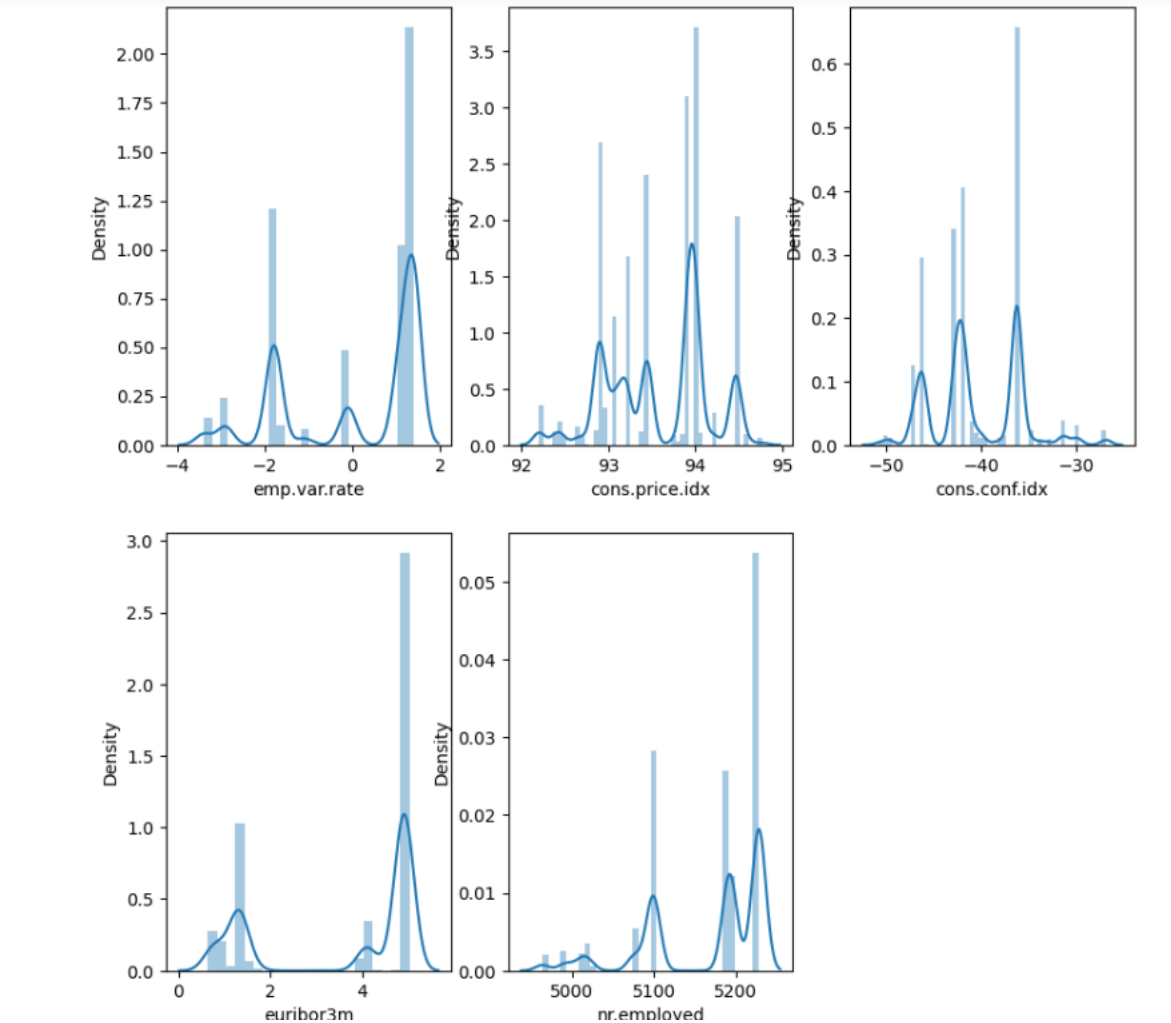
We can see the campaign were mostly concentrated in the starting of the bank period (May, June and July)

- 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.

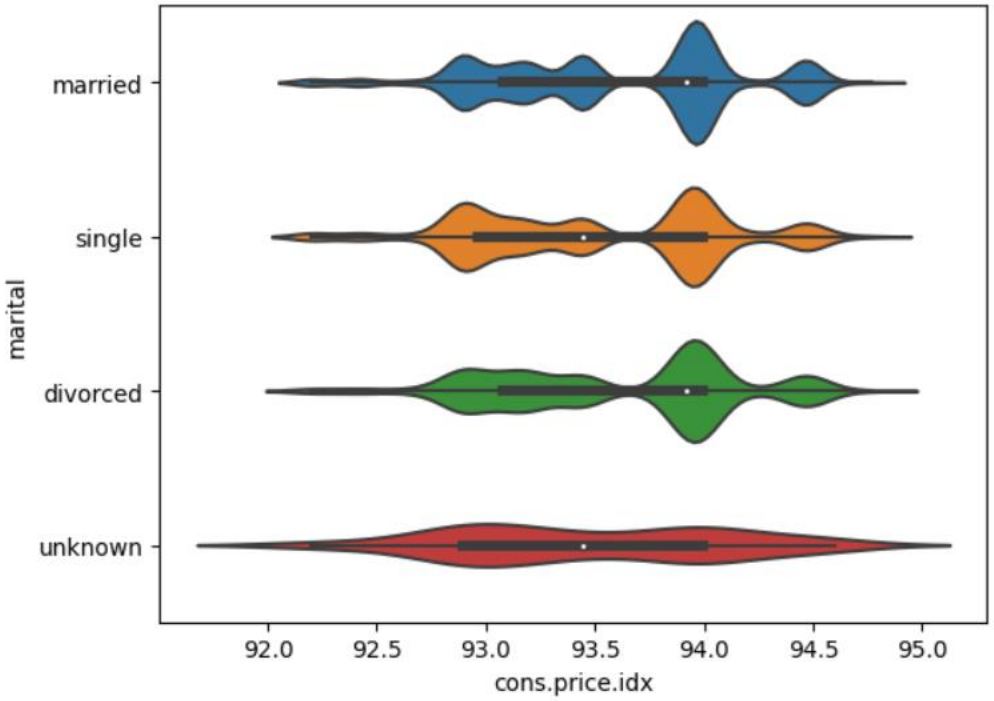


Distribution of Quarterly Indicators

- 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 where 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 pretty 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 lend 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



Marital Status v/s Price Index

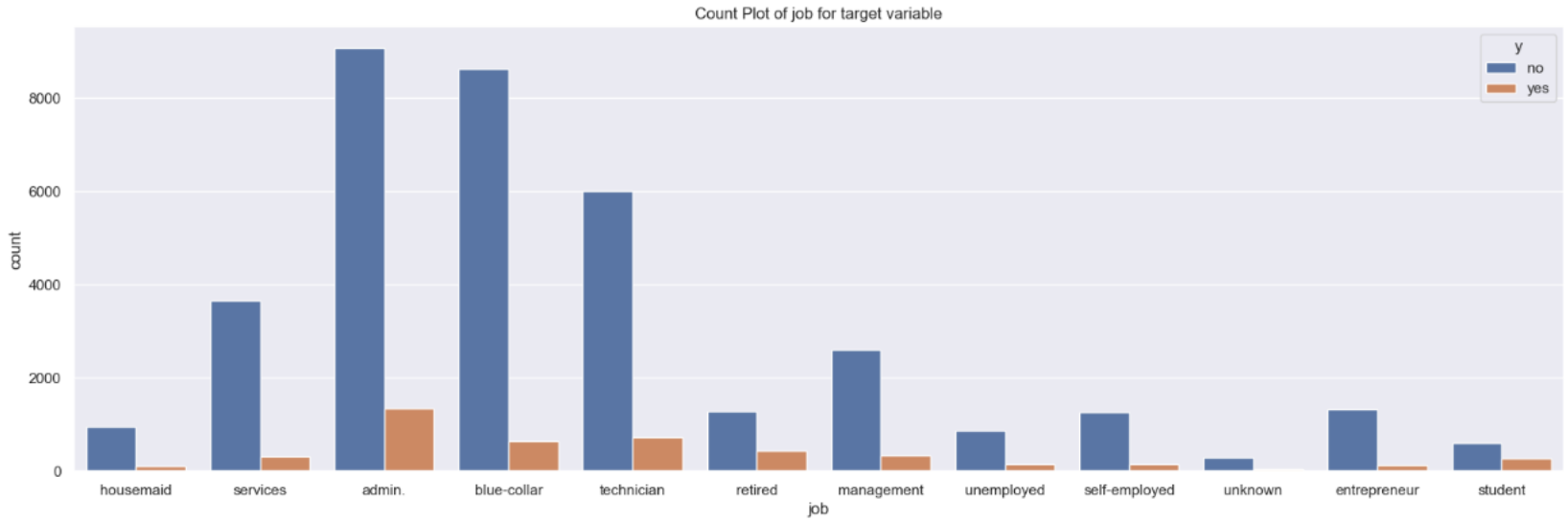


- There are very minute differences among the price index
- Married leads have considerably have an upper hand as they have index contributing as couple

Customer Information

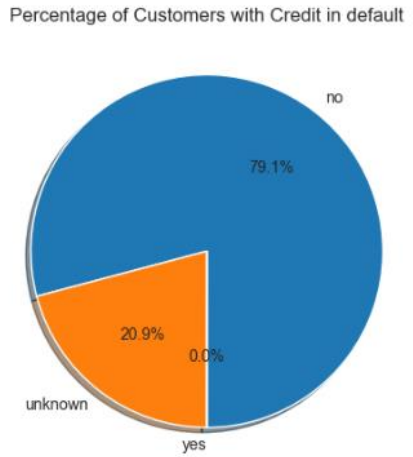
Job

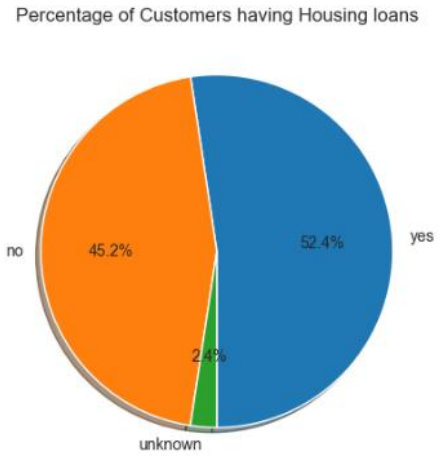
- Top contacted clients are from job type: 'blue-collar', 'management' & 'technician'
- Success rate is highest for student

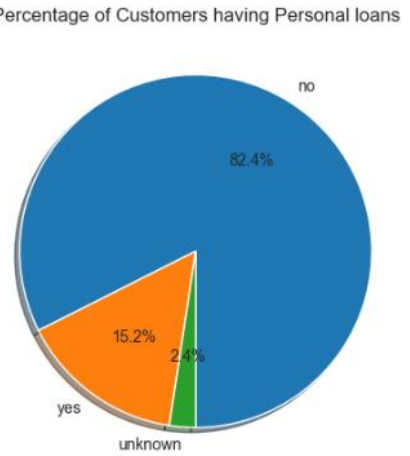


Default, Housing and Loan

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

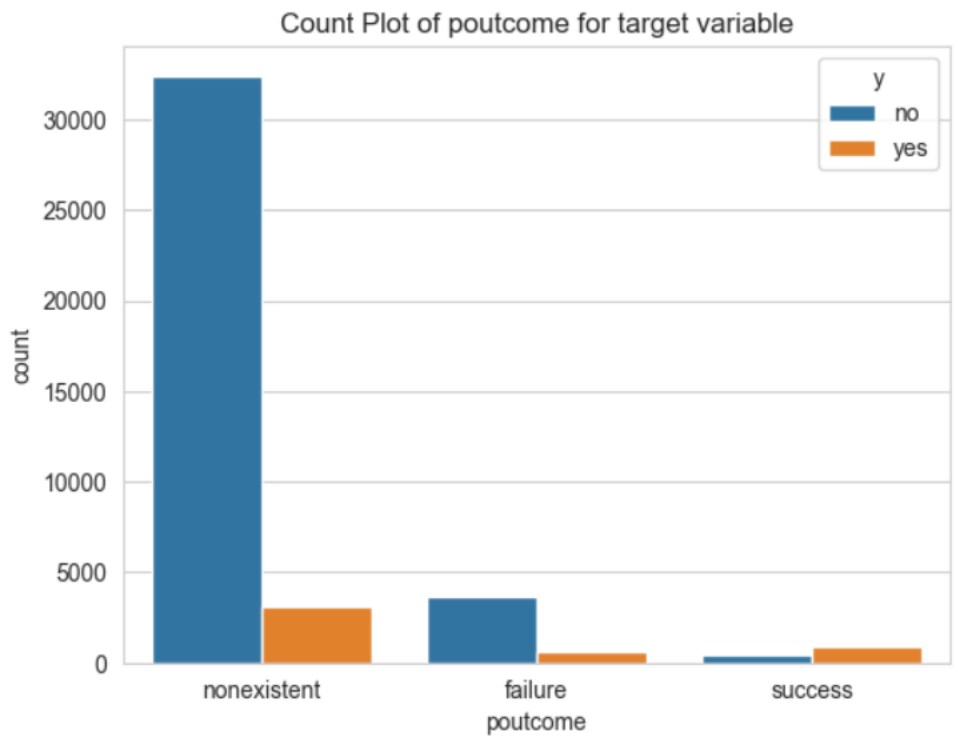






Data Glacier

Customer Information

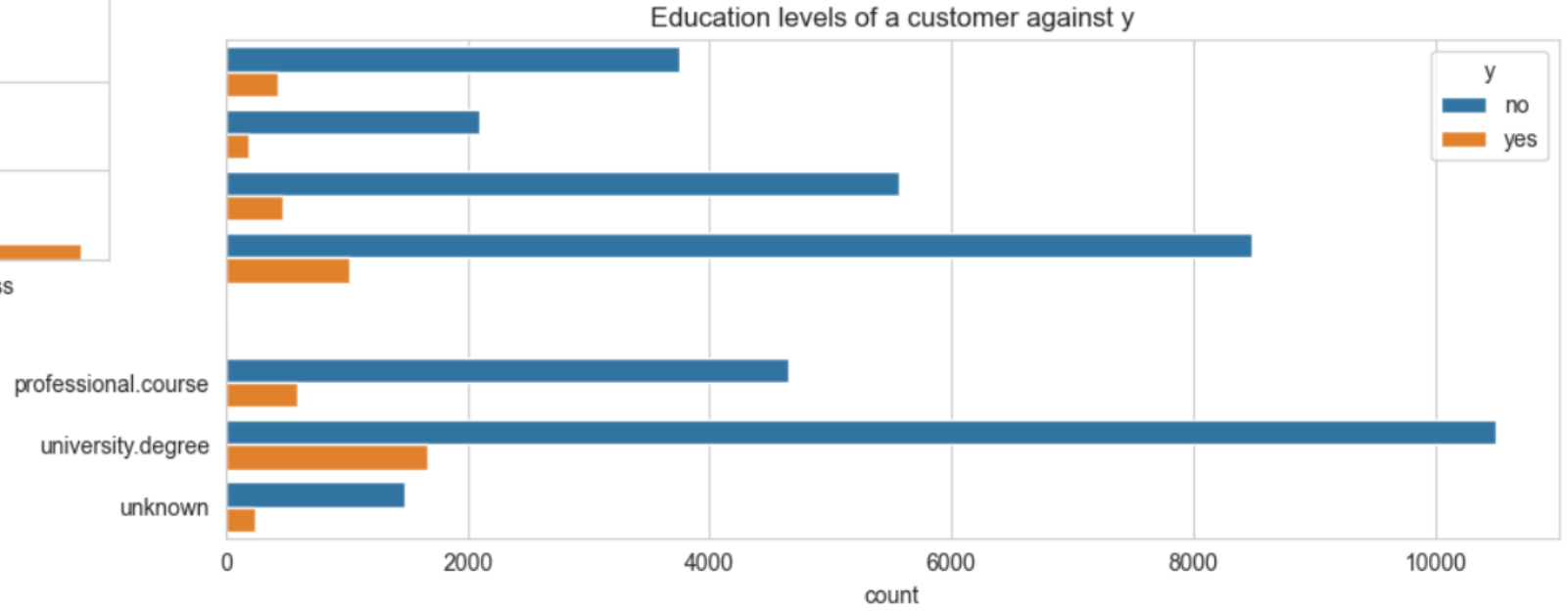


Poutcome

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

Education

- Most of the people who are contacted have tertiary or secondary education.
- Customers with university degree have subscribed to the term deposit more

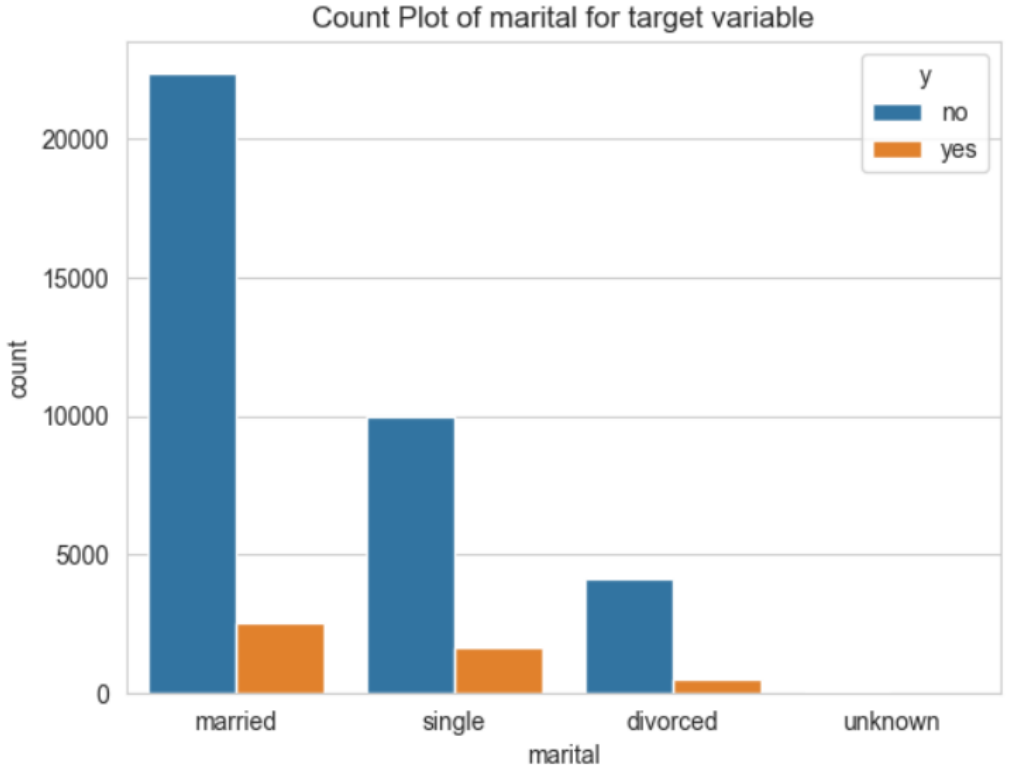


Data Glacier

Customer Information

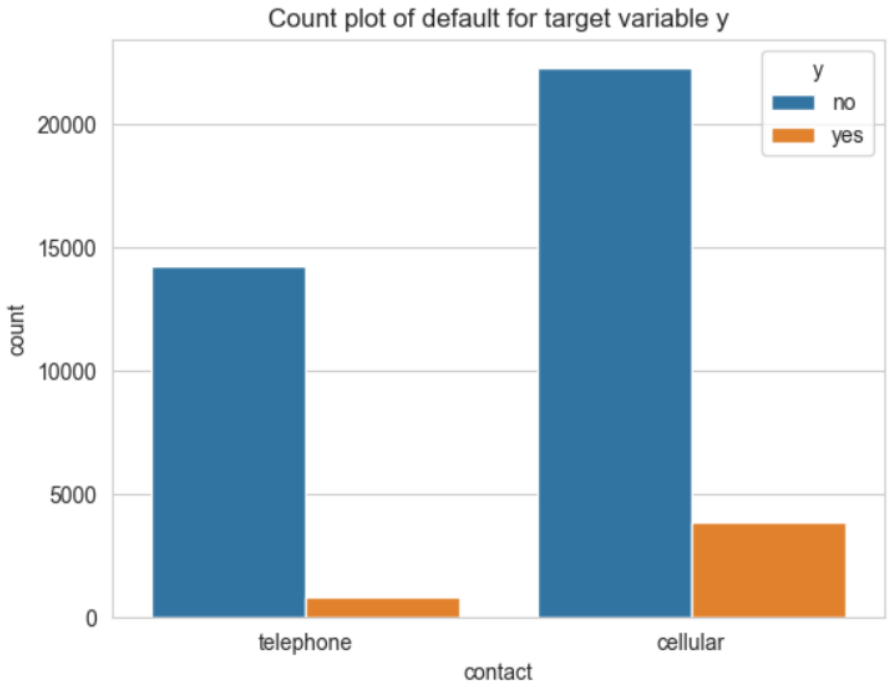
Month

- 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.



Contact

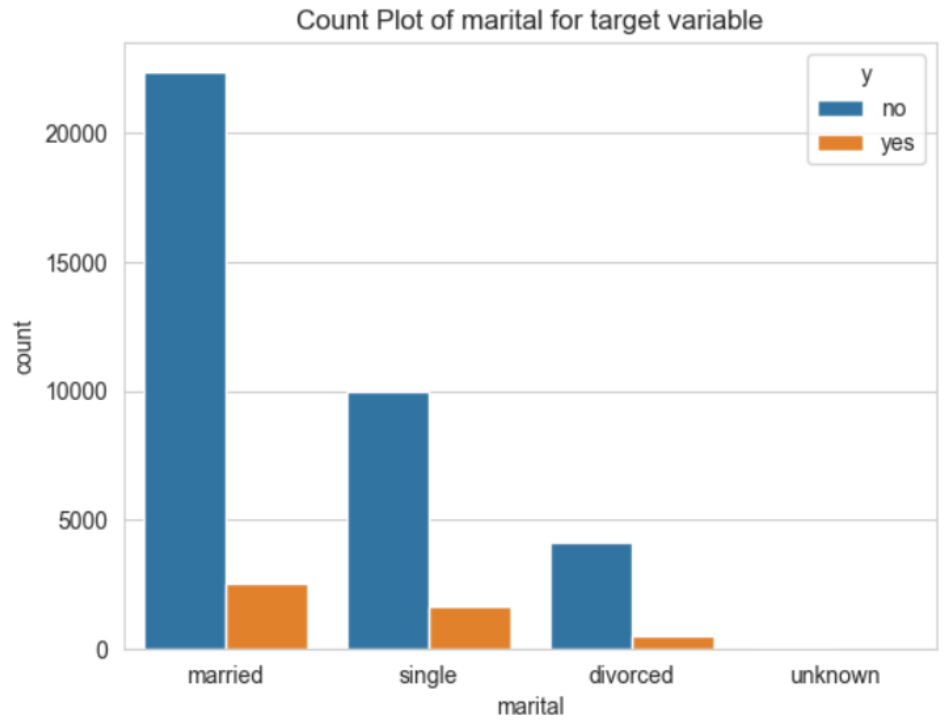
- Most of the people are contacted through cellular



Customer Information

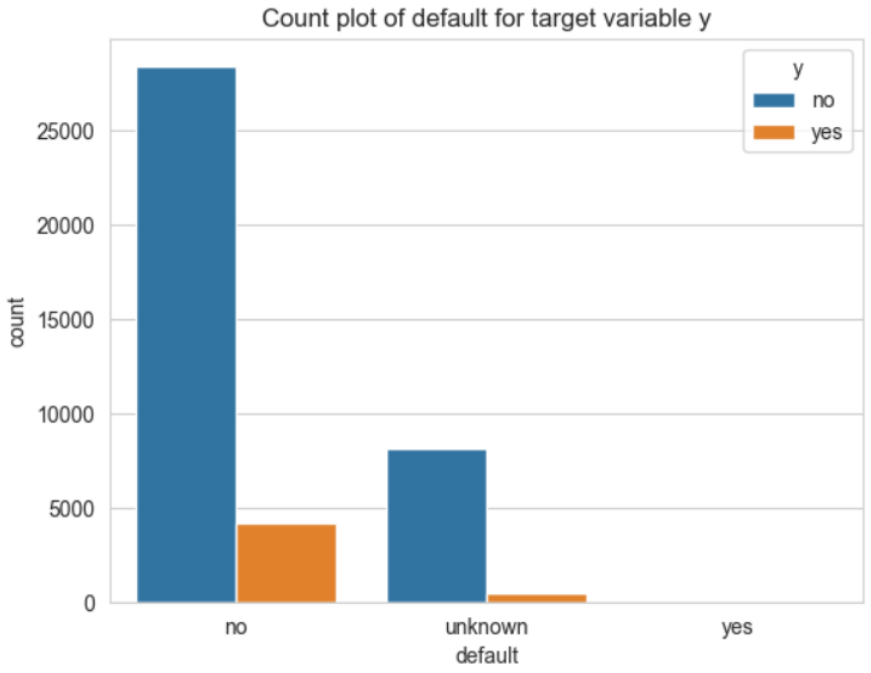
Marital

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



Default

- Very few clients are contacted who are defaulter



Recommended Models

- Logic Regression
- Decision Tree
- Random Forest
- Extreme Gradient Boosting

Recommendations

- May is the most effective month to contact customers.
- Give more focus on university graduate students and high school degree students.
- Age groups of 26-40 and 41-60 have a higher proportion among customers, therefore these groups present a profitable target for the marketing team.
- Target the admins and technicians for more subscriptions.
- The 'poutcome' - outcome of the previous marketing campaign has the greatest influence on the current campaign's outcome
- The customers occupation and 'maritial' status are also among the factors that significantly impact the results.
- '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.

Data Glacier