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).

Business understanding:

Bank wants to use ML model to shortlist customer whose chance of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc.) can focus only to those customers whose chance of buying the product is more.

This will save resource and their time (which is directly involved in the cost (resource billing)).

Data Intake Report

Name: Bank Marketing (campaign)
Report date: 08/25/2022
Internship Batch: LISUM11: 30
Version: 1.0
Data intake by: Priyadarshani Kamble
Data intake reviewer: NA
Data storage location: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing/bank-additional-full.csv

Tabular data details:

Total number of observations	41188
Total number of files	1
Total number of features	21
Base format of the file	csv
Size of the data	5699kb

Note: Replicate same table with file name if you have more than one file.

Data Understanding:

Step 1 - Import data and explore the data

- The data is available in csv file.
- The Key attributes include age, job, marital, education, default, balance, housing loan, contact, day, month, duration, campaign, pdays, previous, poutcome, y.
- There are 41188 records and 21 Features.
- There are 10 numeric columns and 11 Categorical Columns
- Checked the no of unique values in each columns. If the feature has constant value or 1 value then such columns can be dropped as they do add any value in model building. The dataset does not any such columns.

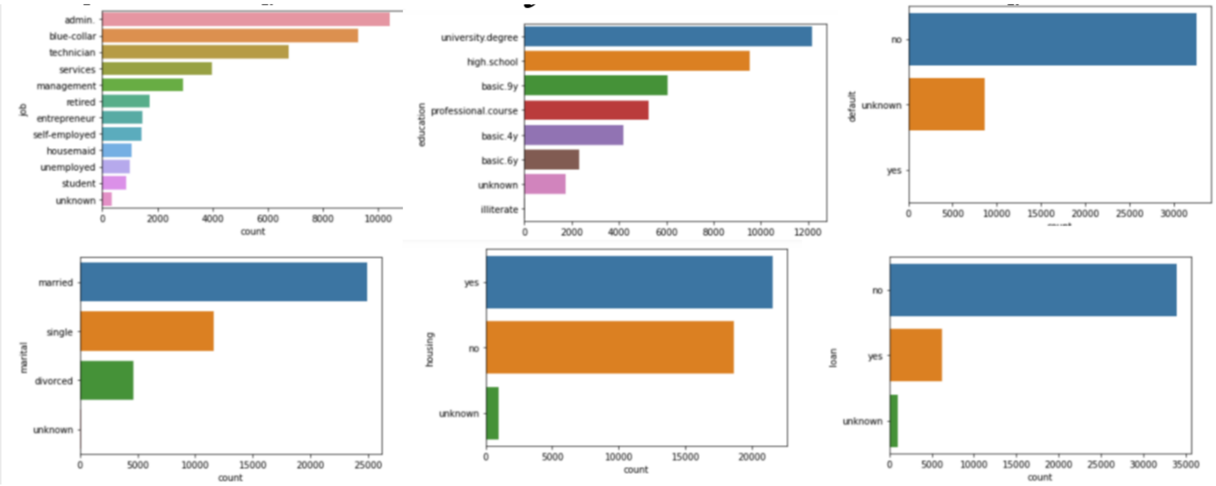
Step 2-Data Cleaning:

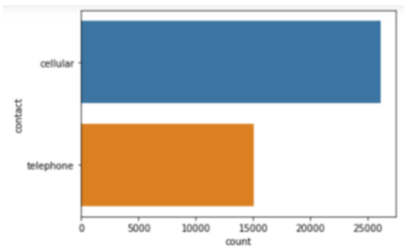
- Checked for missing values. No missing data found.
- 12 duplicate rows found. Deleted the duplicate records.

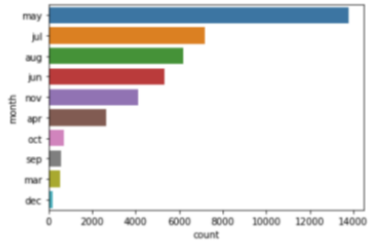
Step 3- Exploratory data analysis:

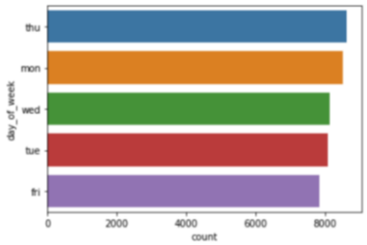
Univariate Analysis:

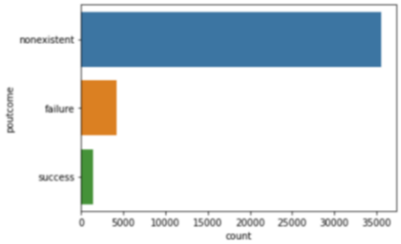
Check how Each categorical feature is distributed using Countplot. Below are the findings:

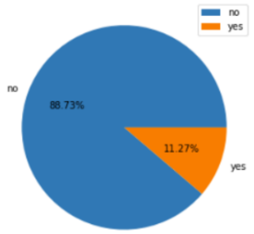






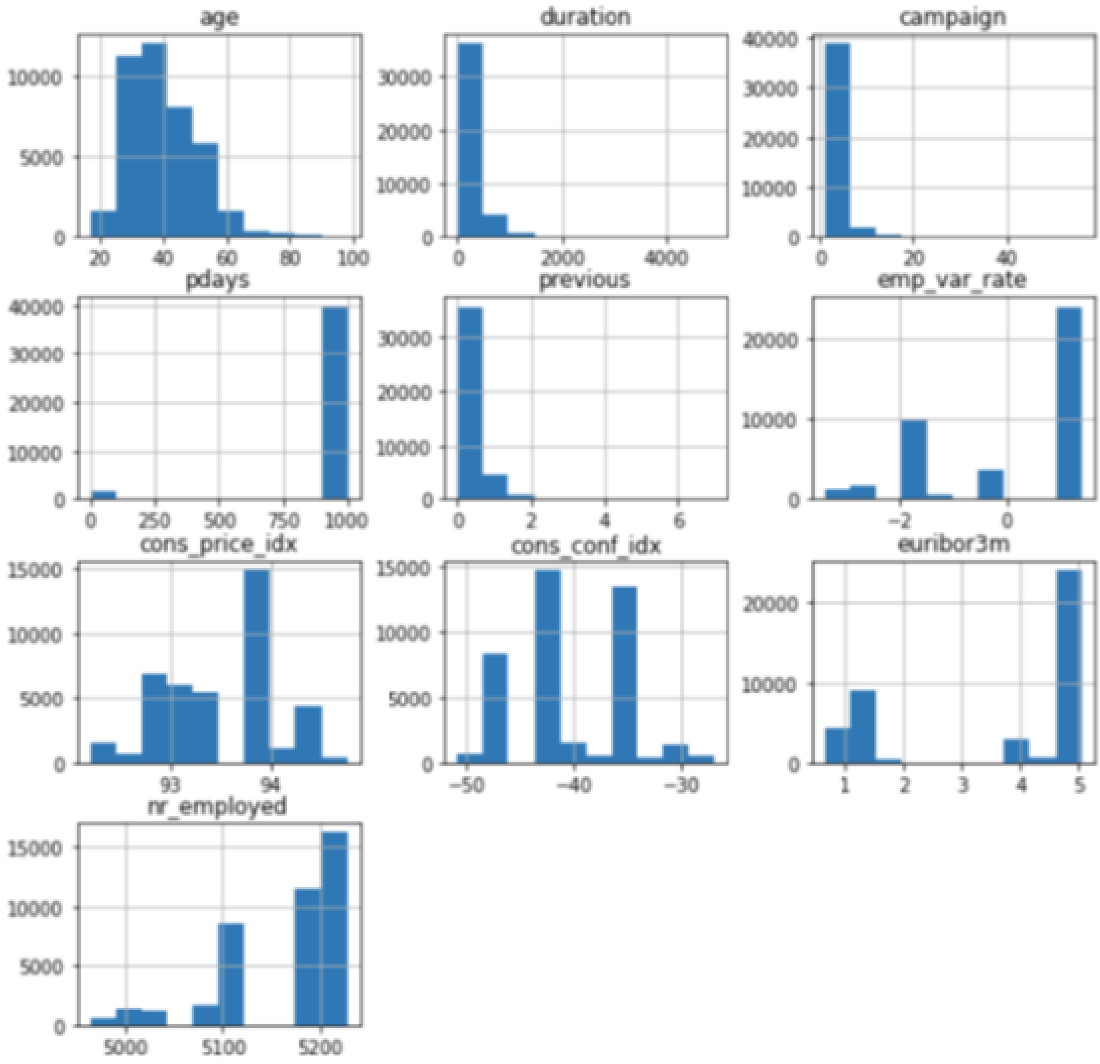






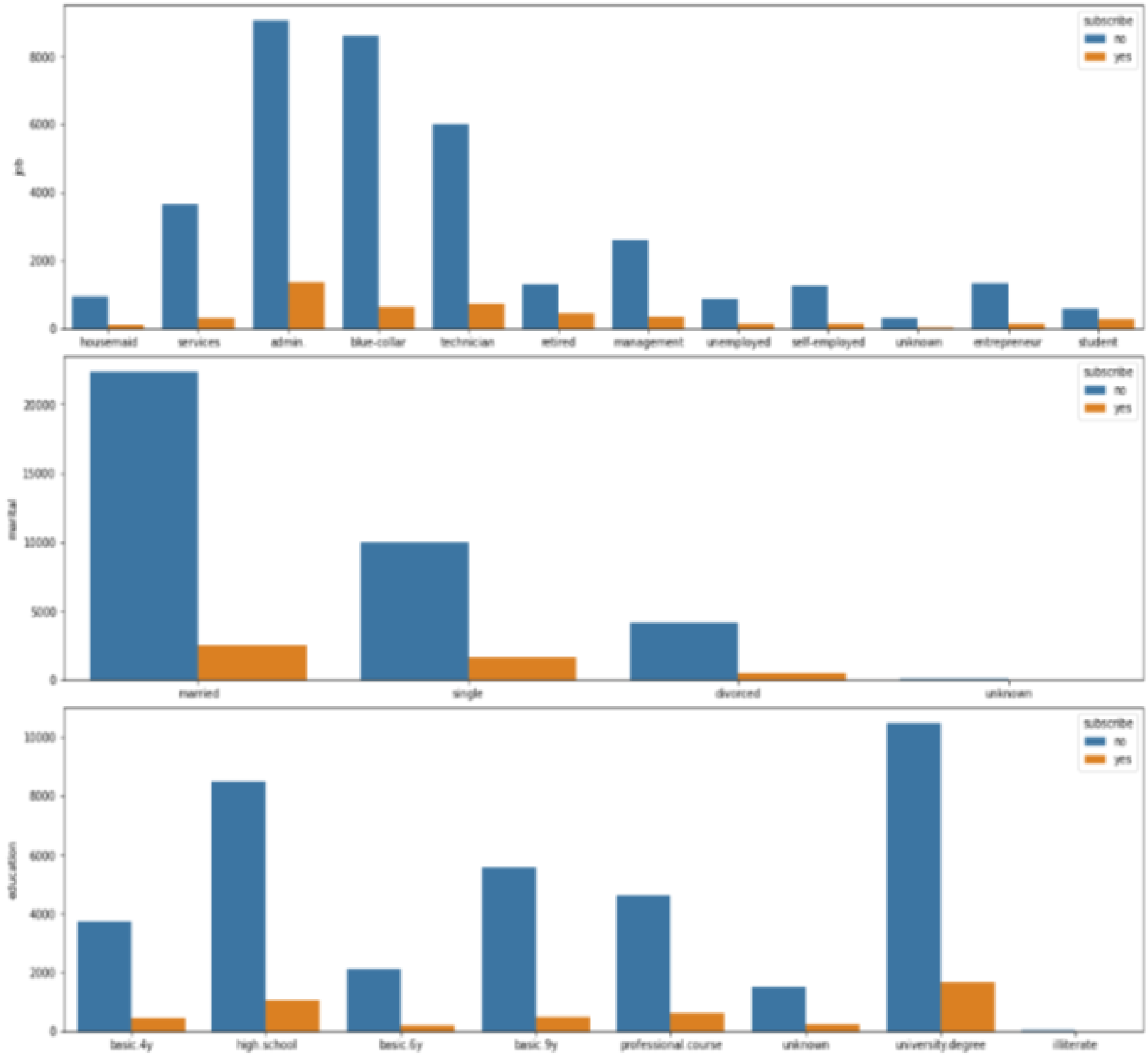
- Most of the clients are working as admin 25% and blue collar 22% job category.
- 60% of the clients are married.
- Most of the clients 29% hold University degree.
- The no of clients who defaulted on a credit, are very less. It shows 80% data for 'No'. % of yes is almost 0. so this feature doesn’t seem to be very for prediction purposes and can be dropped from the dataset.
- Housing Shows almost equal % of yes and no.
- Most of the clients do not have personal loan.
- More than 63% of all clients were contacted through cellular phone.
- Most of the clients were contacted in the month of May 33%.
- Here we see equal distribution of the data in the graph and the % amongst the days. So, there is no significant day which shows more activity than others.
- More than 86% of clients were never covered by previous marketing campaigns.
- We see there is imbalance in data. only 11.70% clients have subscribed to a term deposit.

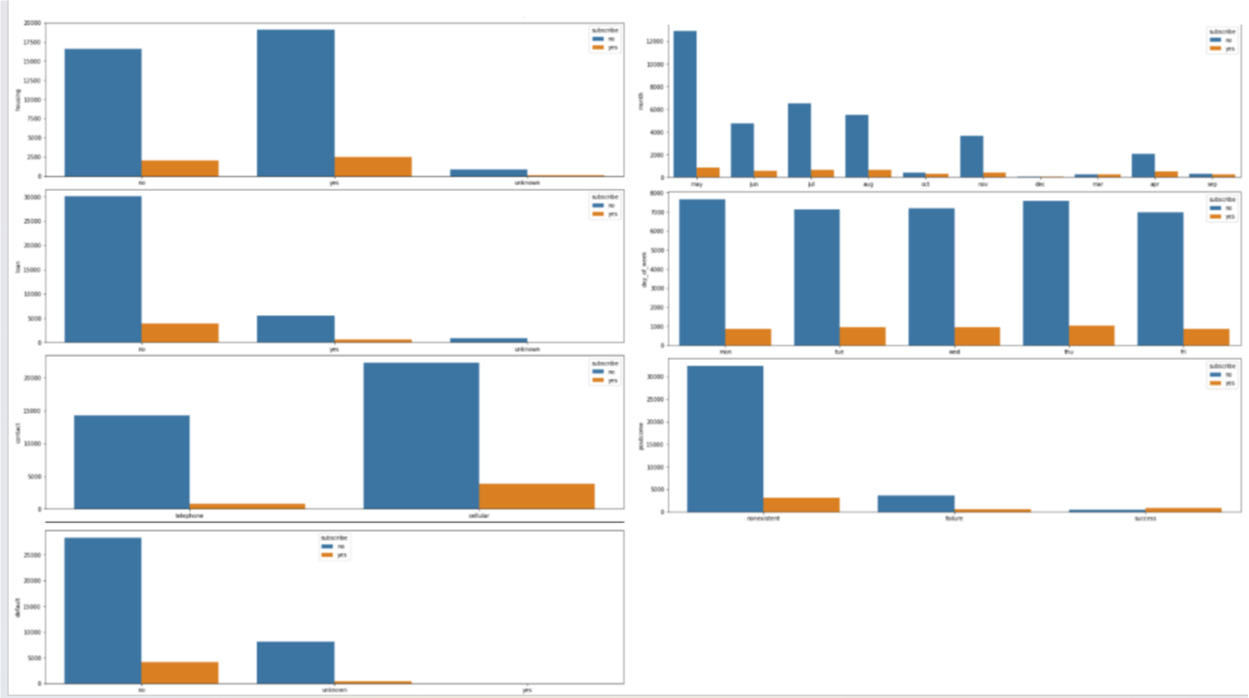
Check how Each Numerical feature is distributed using Histogram. Below are the findings:



- The graph of age shows normal distribution.
- The graph of pdays, previous, duration, campaign shows skewness.
- The variable "duration" will need to be dropped before we start building a predictive model because it highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed.
- Other graphs show several spikes in the data.
- pdays columns shows above 96% values are having 999 value. So, this column should be removed.

Bivariate Analysis Findings:

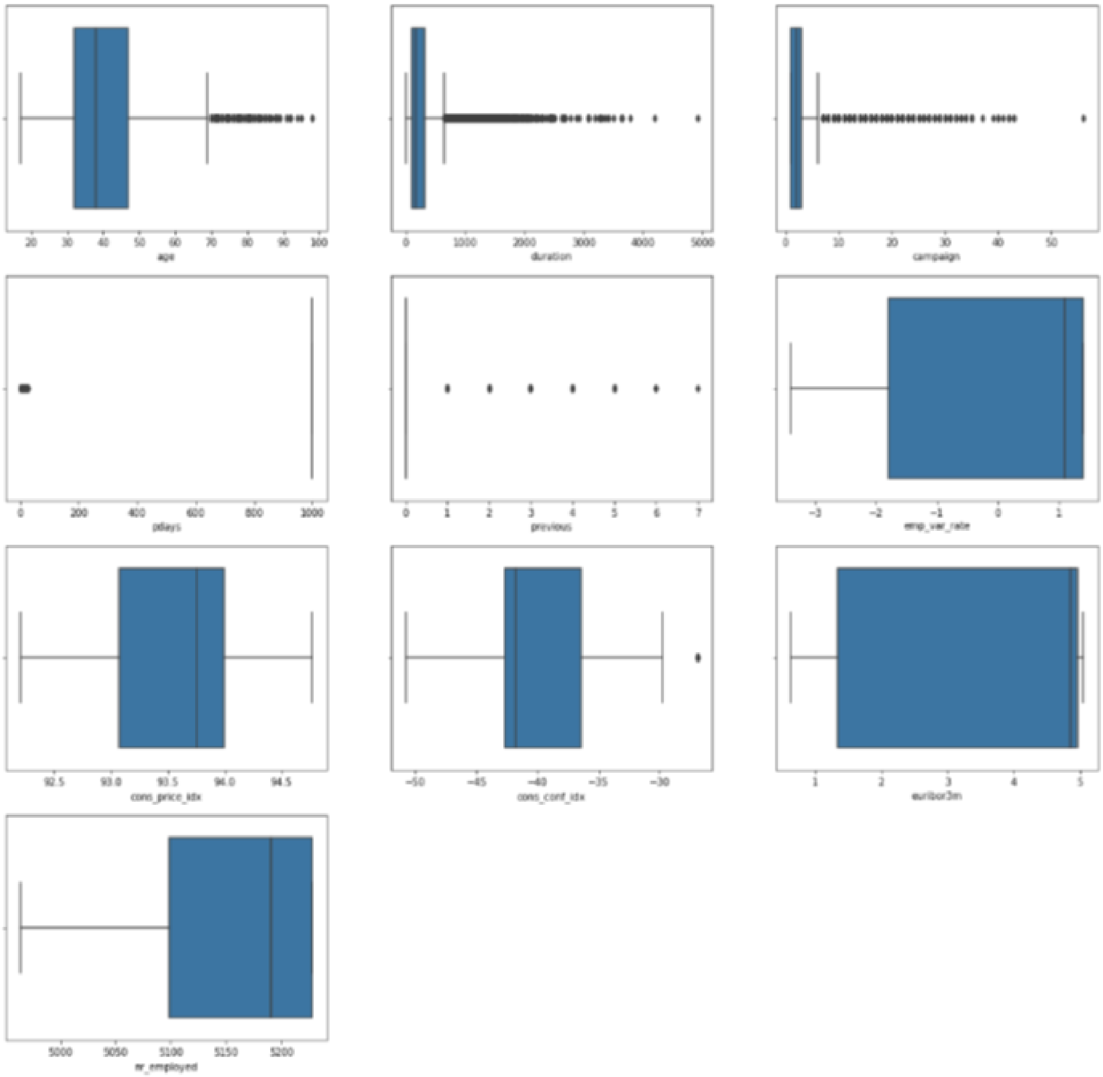




- The % of clients showing interest in 'term deposit' is more when they are married, holds university degree and no defaulted to credit.
- The retired people seem to have higher % of 'Yes' for term deposit than other job category clients.
- The clients having housing, loan have higher % or saying no to term deposit.
- The months may, jun, juy, aug shows more clients responding to term deposit
- If poutcome = success then the % of 'yes' to term deposit is high
- % of interest in 'Deposit' is more when clients are contacted via cellular mode.
- Clients having longer durations shows more interest in term deposit

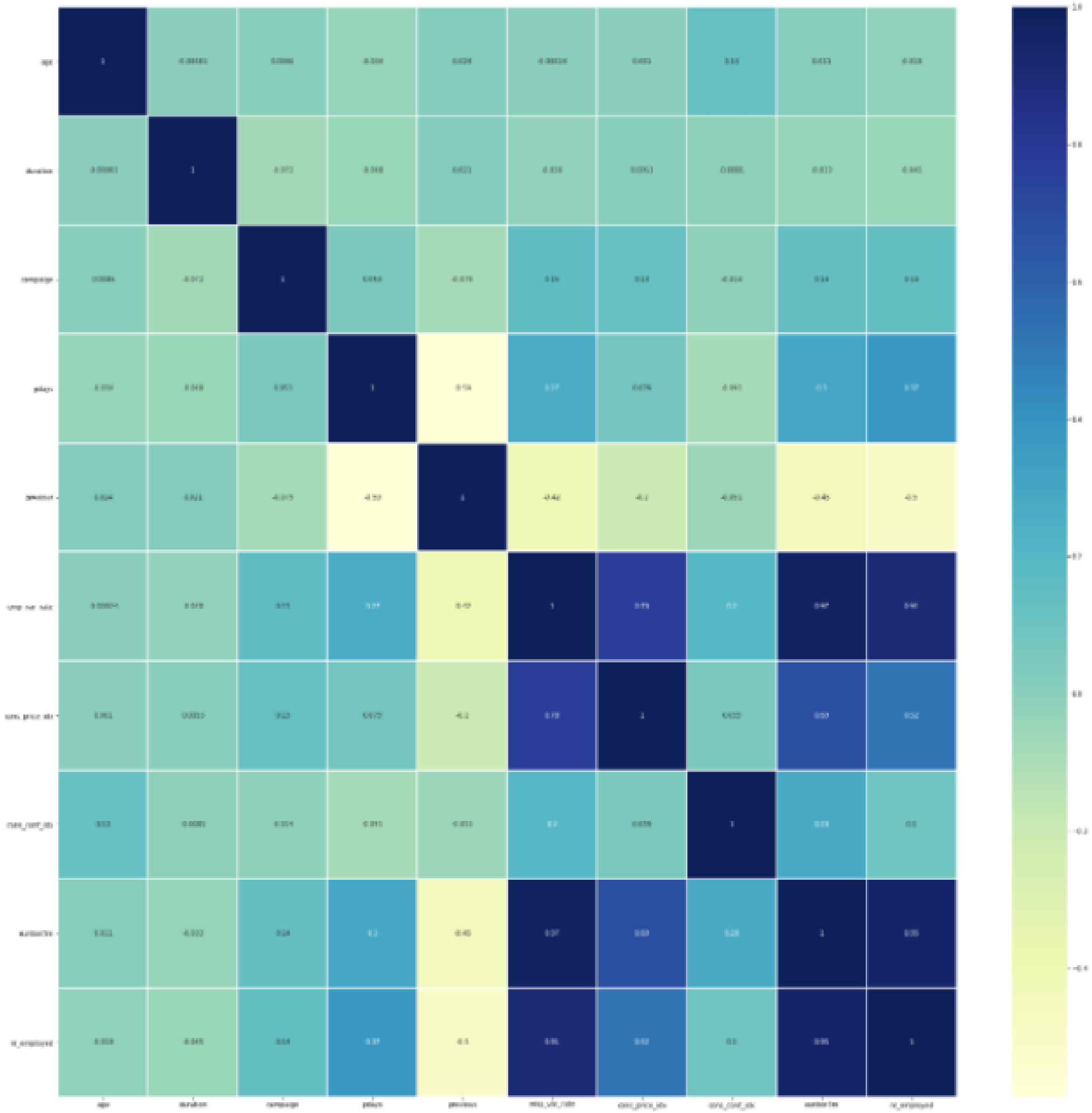
Checked Outlier using box plot.

The features like age, pdays, duration, campaign, duration has some outliers



Check correlation matrix - to check the strength of variation between two variables

Multivariate Analysis Findings:



Multivariate analysis showed that, euribor3m is highly corelated with 'emp_var_rate' and nr_employed. They will add redundancy and overfitting. This should be dropped during feature engineering.

Step 4-feature engineering
In this step, we should drop the unwanted features , handle missing data, remove outliers.

Drop below columns as per the findings in EDA :

- emp_var_rate
- nr_employed
- default
- pdays

Outlier removing:
Handle the outliers by taking 99 percentile data into consideration for the columns

- Age
- Duration
- Cons_conf_idx
- Campaign