Project: Bank Marketing (Campaign)

Week 10: Deliverables

Name: Supin Hooda
Email: hoodasupin@gmail.com
Country: Canada
Batch Code: LISUM25
Specialization: Data Science
Submission Date: 6th Nov 2023
Submitted to: Data Glacier (Individual project)

Table of Contents

Problem Description

Background:
ABC Bank is planning to launch a new term deposit product and is looking to create a predictive model to determine whether a customer is likely to subscribe to this product based on their interactions with the bank and other financial institutions. This predictive model aims to assist the bank in optimizing its marketing efforts and improving the effectiveness of its campaigns.

Objective:
The primary objective of this project is to develop a predictive model that can accurately classify customers into two groups: those who are likely to subscribe to the term deposit ("yes") and those who are not likely to subscribe ("no").

Data Source:
The dataset provided for this project contains various customer attributes and information related to the bank's marketing campaigns. These attributes will be used to build and train the predictive model.

Data Understanding

Types of Data:
The dataset comprises both numerical and categorical variables. Here is a summary of the variables:

1. Numerical Variables:
  - `age`: Customer's age (numeric)
  - `duration`: Last contact duration, in seconds (numeric)
  - `campaign`: Number of contacts performed during this campaign (numeric)
  - `pdays`: Number of days passed after the client was last contacted from a previous campaign (numeric; 999 means the client was not previously contacted)
  - `previous`: Number of contacts performed before this campaign and for this client (numeric)
  - `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)

2. Categorical Variables:
  - `job`: Type of job (categorical)
  - `marital`: Marital status (categorical)
  - `education`: Education level (categorical)
  - `default`: Has credit in default? (categorical)
  - `housing`: Has a housing loan? (categorical)
  - `loan`: Has a personal loan? (categorical)

- `contact`: Contact communication type (categorical)
- `month`: Last contact month of the year (categorical)
- `day_of_week`: Last contact day of the week (categorical)
- `poutcome`: Outcome of the previous marketing campaign (categorical)
- `y`: The target variable, whether the client subscribed to a term deposit (binary: 'yes' or 'no')

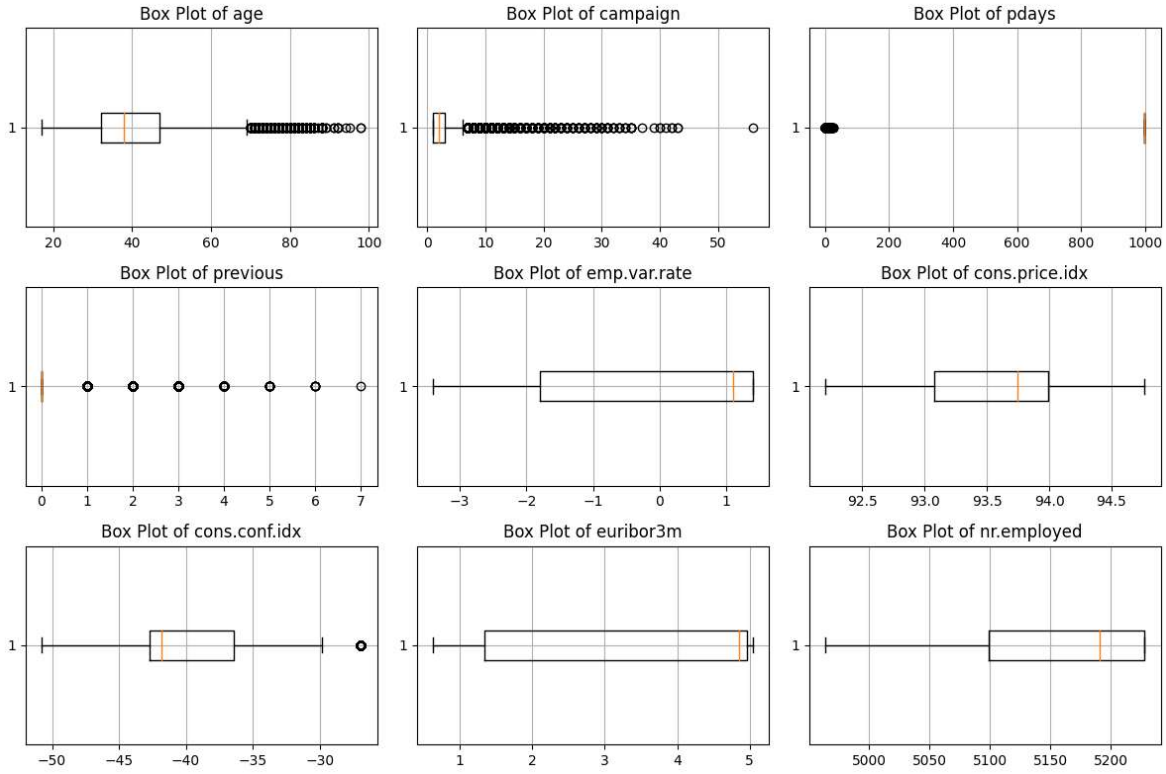
Problems in the Data

The dataset contains 41,188 rows, with 12 duplicate rows identified and removed using the `drop_duplicates` method. Fortunately, there are no missing values in the dataset.

Outlier Detection

Outliers in the data are detected using box plots. Outliers are data points that deviate significantly from the central distribution of data, impacting key statistics such as mean and mode. Managing outliers during data cleaning is essential to ensure that model performance is not compromised.

Outliers are observed in the following features: "age," "campaign," "pdays," and "previous." These outliers are identified as data points located outside the box plot whiskers.



Approaches to Overcome the Problems

- **Age**: The maximum age value of 98 appears to be realistic and is retained in the dataset.

- **Pdays**: The maximum value of 999 indicates that the client was not previously contacted, and it's present in around 96% of rows. Thus, it's considered unrealistic to drop rows with this value, and they are retained.

- **Campaign**: The "campaign" feature, which denotes the number of contacts made during the campaign, contains a maximum value of 56, which is considered noise. The portion of records with "campaign" values exceeding 20 is approximately 0.38%. It is proposed to impute these rows with the average of campaign values.

- **Previous**: The "previous" feature represents the number of contacts made before this campaign. The maximum value of 7 does not appear to be an outlier, and it is retained.

Treating Outliers

1. **Treating outliers**

Imputation using median

```
#The value which is outside the whisker
print(df['campaign'].quantile(0.95))
```

```
7.0
```

```
#replacing the values which are greater than the 95th percentile
import numpy as np
df['campaign1'] = np.where(df['campaign'] > 7, 2, df['campaign'])
df[['campaign', 'campaign1']].describe()
```

	campaign	campaign1
count	41176.000000	41176.000000
mean	2.567879	2.118127
std	2.770318	1.383215
min	1.000000	1.000000
25%	1.000000	1.000000
50%	2.000000	2.000000
75%	3.000000	3.000000
max	56.000000	7.000000

Imputation using mean

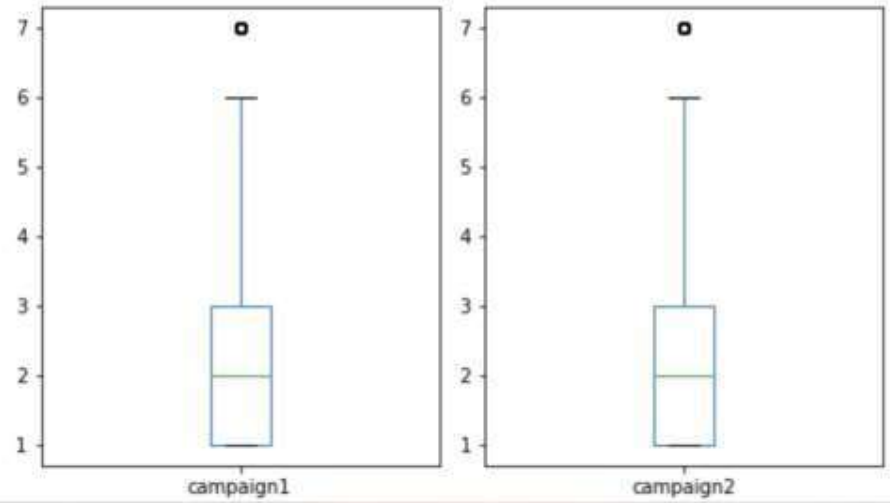
From df.describe(), the mean is 2.56

```
#replacing the values which are greater than the 95th percentile
df['campaign2'] = np.where(df['campaign'] > 7, 2.56, df['campaign'])
df[['campaign', 'campaign1','campaign2']].describe()
```

	campaign	campaign1	campaign2
count	41176.000000	41176.000000	41176.000000
mean	2.567879	2.118127	2.142295
std	2.770318	1.383215	1.385829
min	1.000000	1.000000	1.000000
25%	1.000000	1.000000	1.000000
50%	2.000000	2.000000	2.000000
75%	3.000000	3.000000	3.000000
max	56.000000	7.000000	7.000000

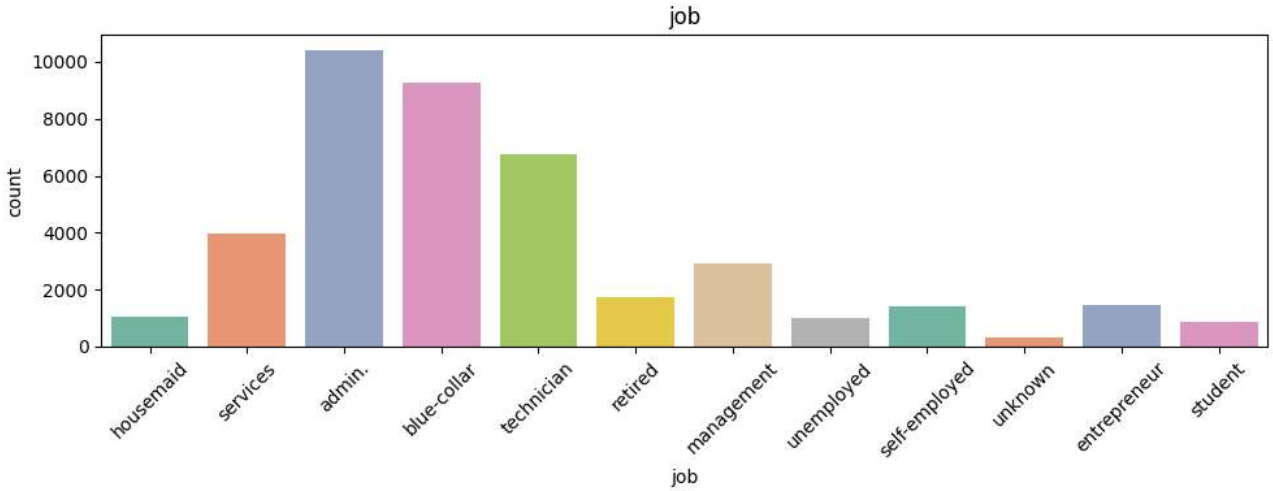
Statistics of the dataset after both median and mean imputation remainsmore or less the same.

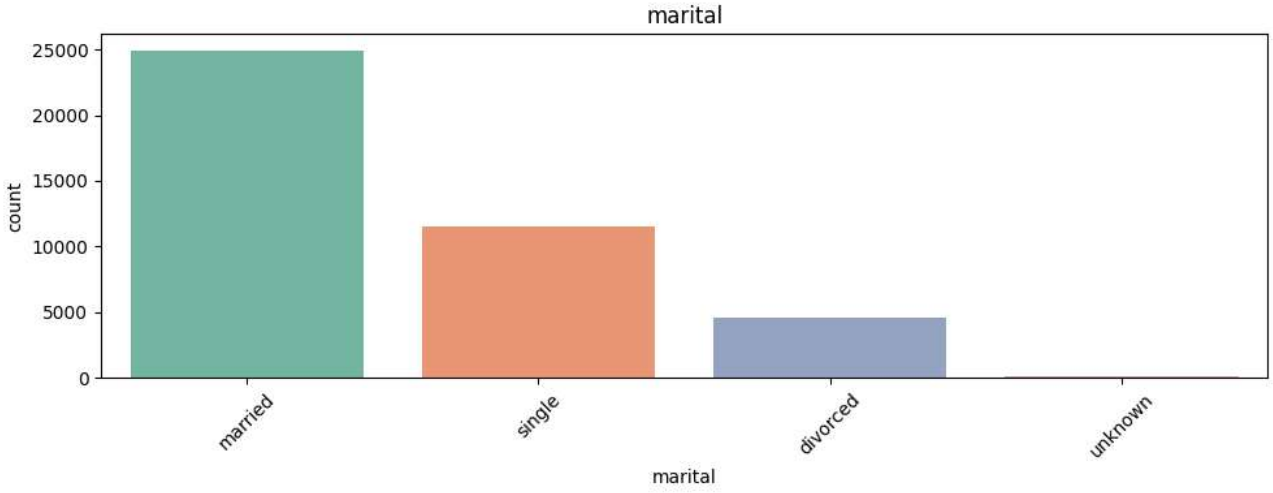
```
#outlier detection  after imputation
import matplotlib.pyplot as plt
cols = ['campaign1','campaign2']
plt.figure(figsize=(10,15))
for i, col in enumerate(cols):
    plt.subplot(4,3,i+1)
    df.boxplot(col)
    plt.grid()
    plt.tight_layout()
```

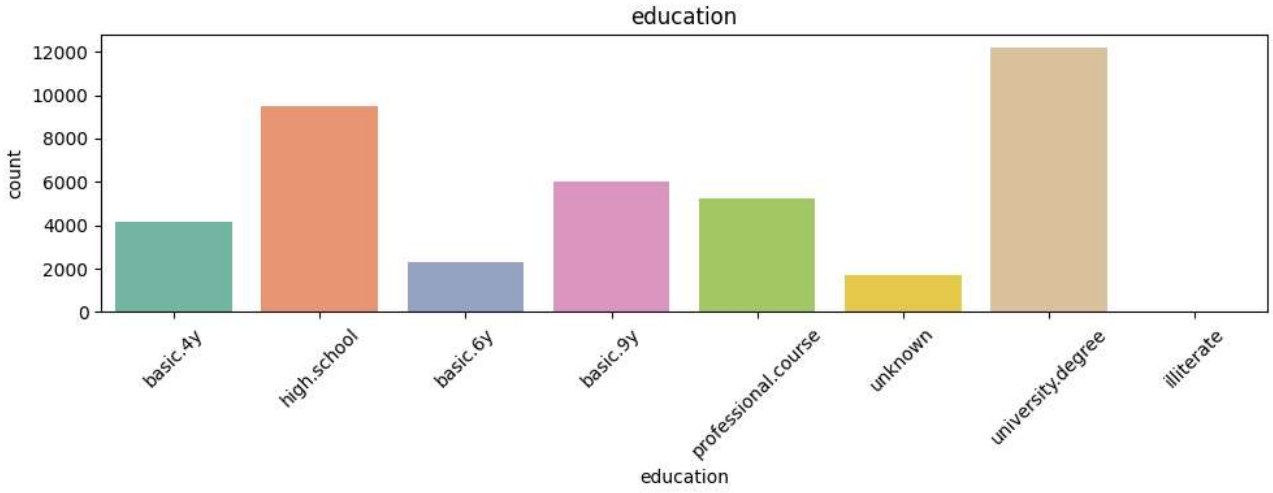


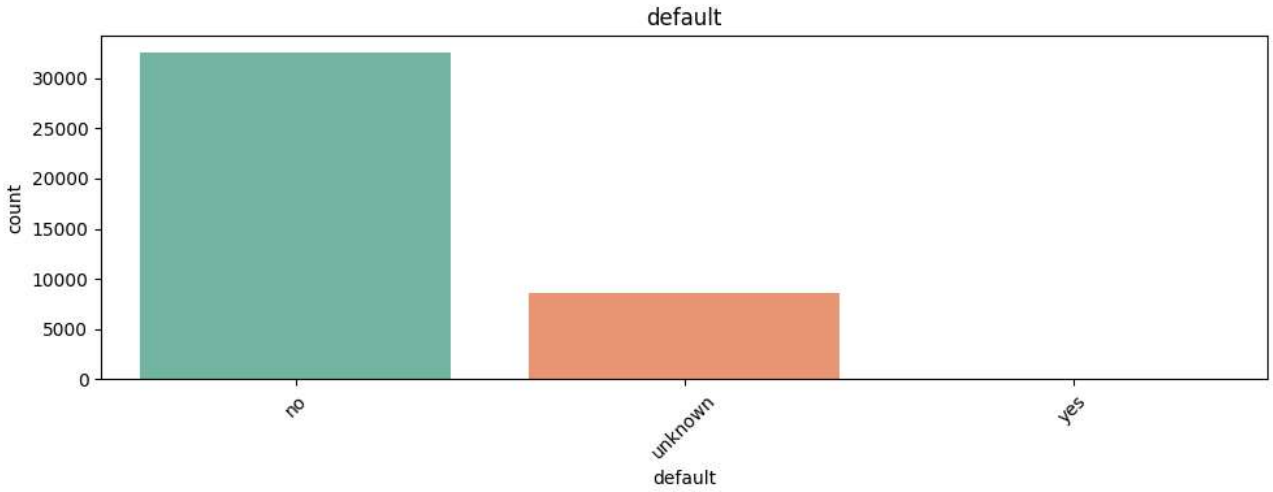
2. EDA and recommendation

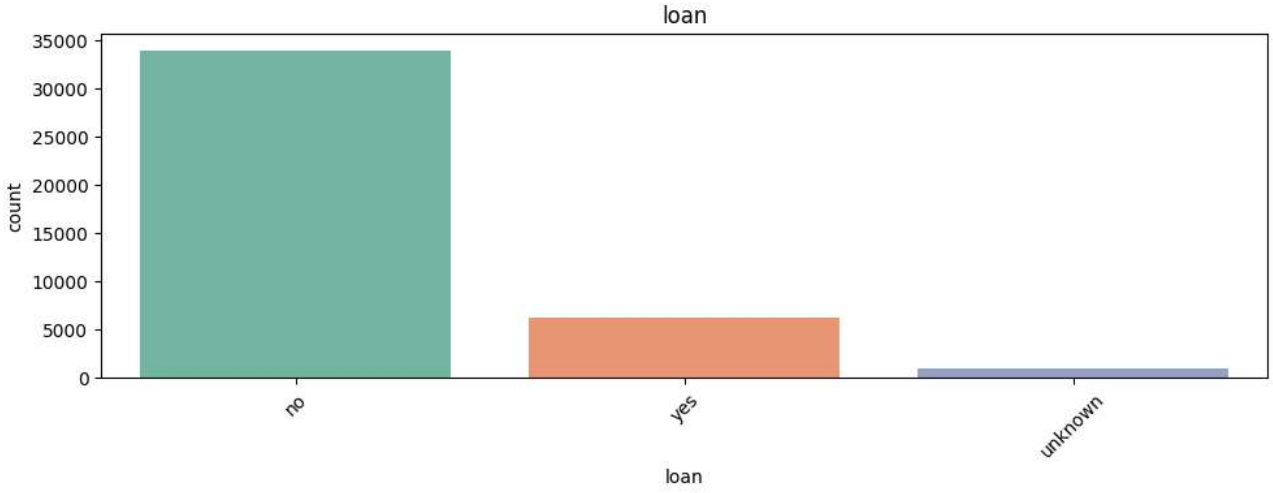
- Exploring categorical values

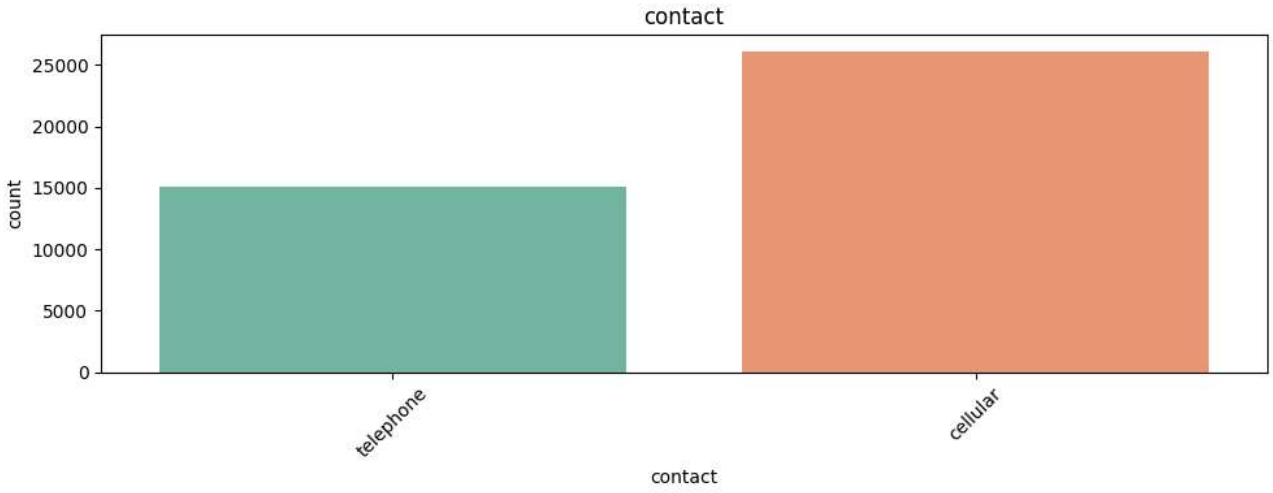


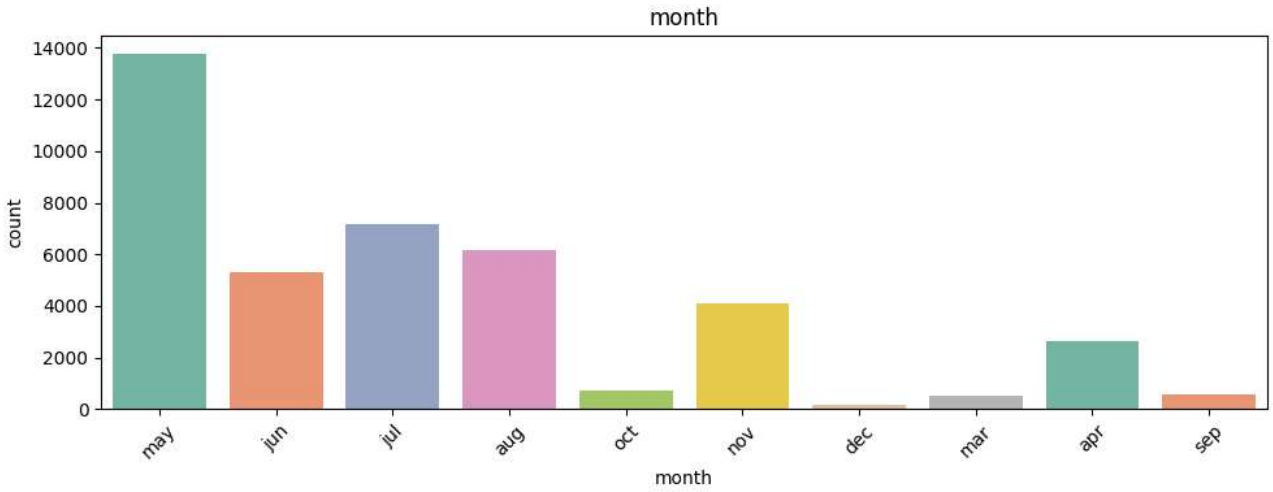

marital
count
25000
20000
15000
10000
5000
0
married
single
divorced
unknown
marital

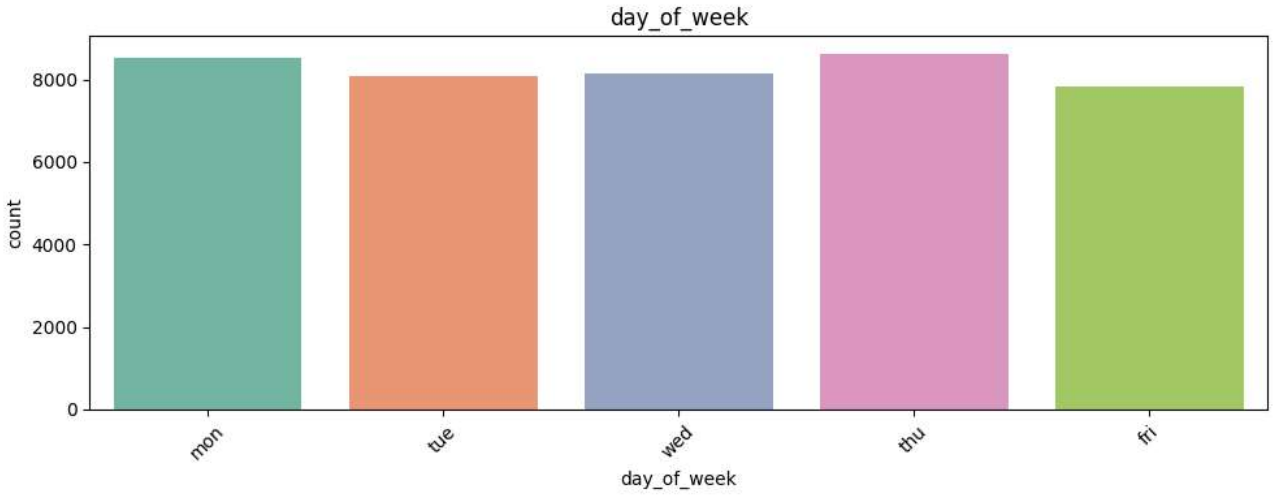

education
count
12000
10000
8000
6000
4000
2000
0
basic.4y
high.school
basic.6y
basic.9y
professional.course
unknown
university.degree
illiterate
education

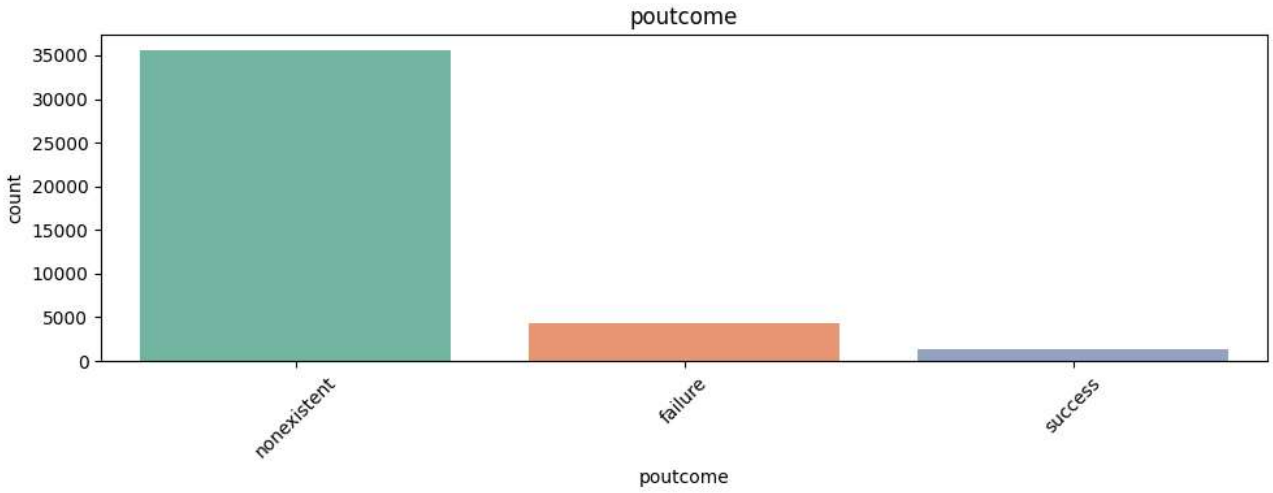

default
count
30000
25000
20000
15000
10000
5000
0
no
unknown
yes
default

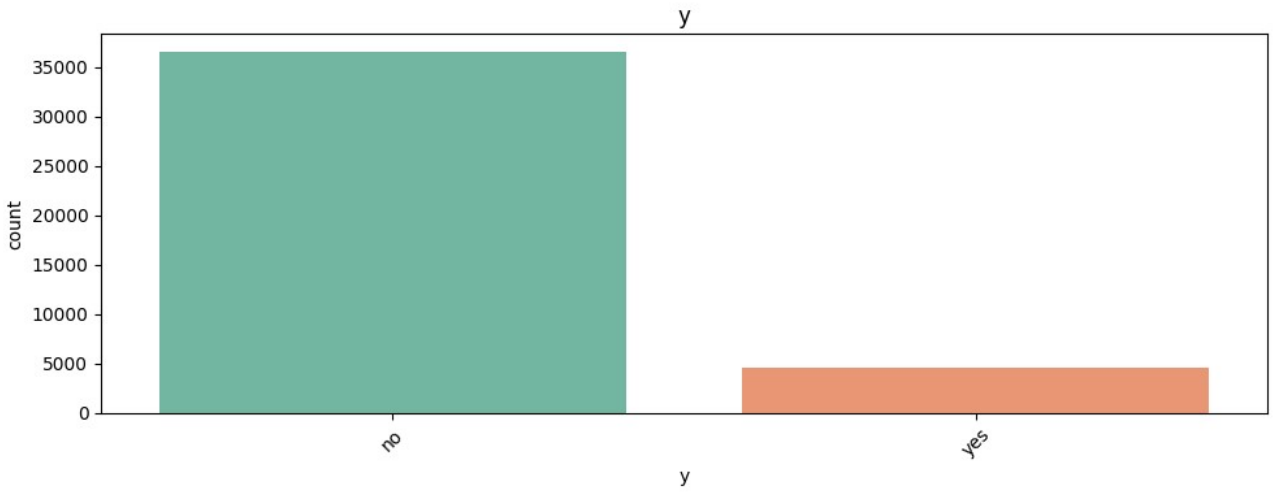

loan
35000
30000
25000
20000
15000
10000
5000
0
count
no
yes
unknown
loan


contact
25000
20000
15000
10000
5000
0
count
telephone
cellular
contact

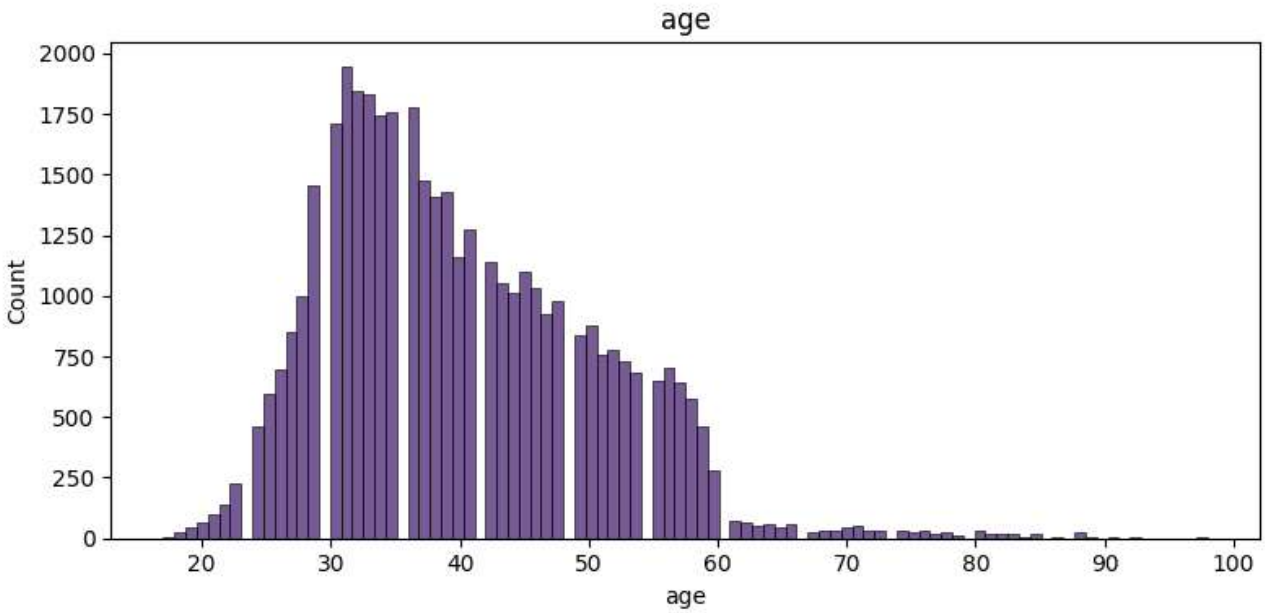

month
14000
12000
10000
8000
6000
4000
2000
0
count
may
jun
jul
aug
oct
nov
dec
mar
apr
sep
month

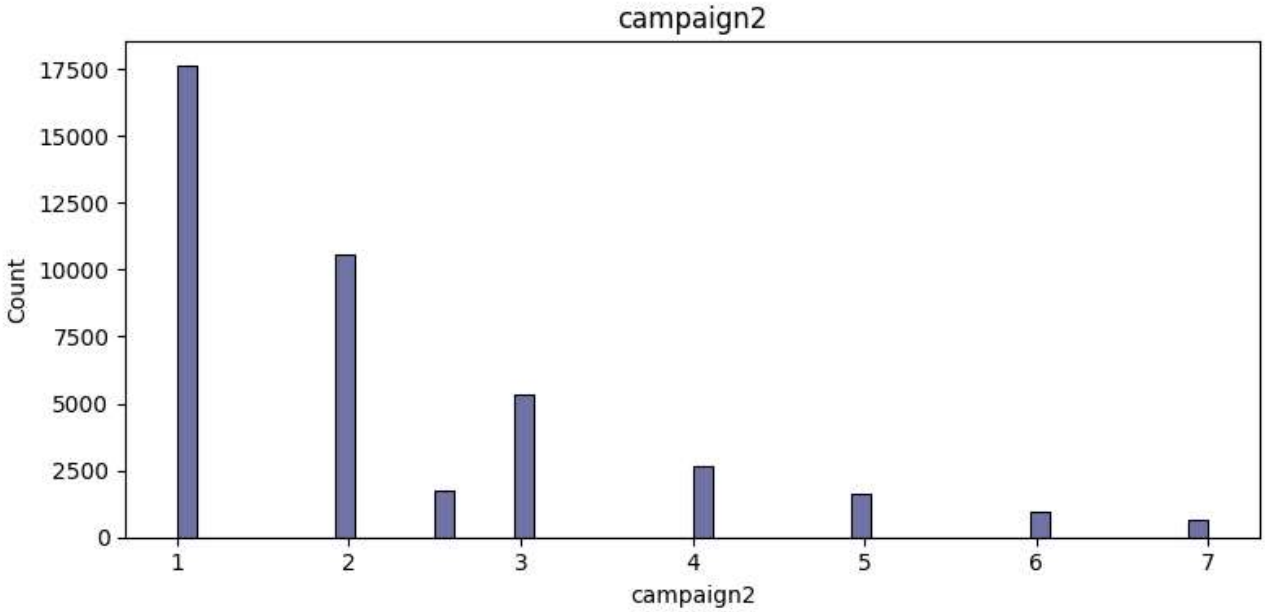


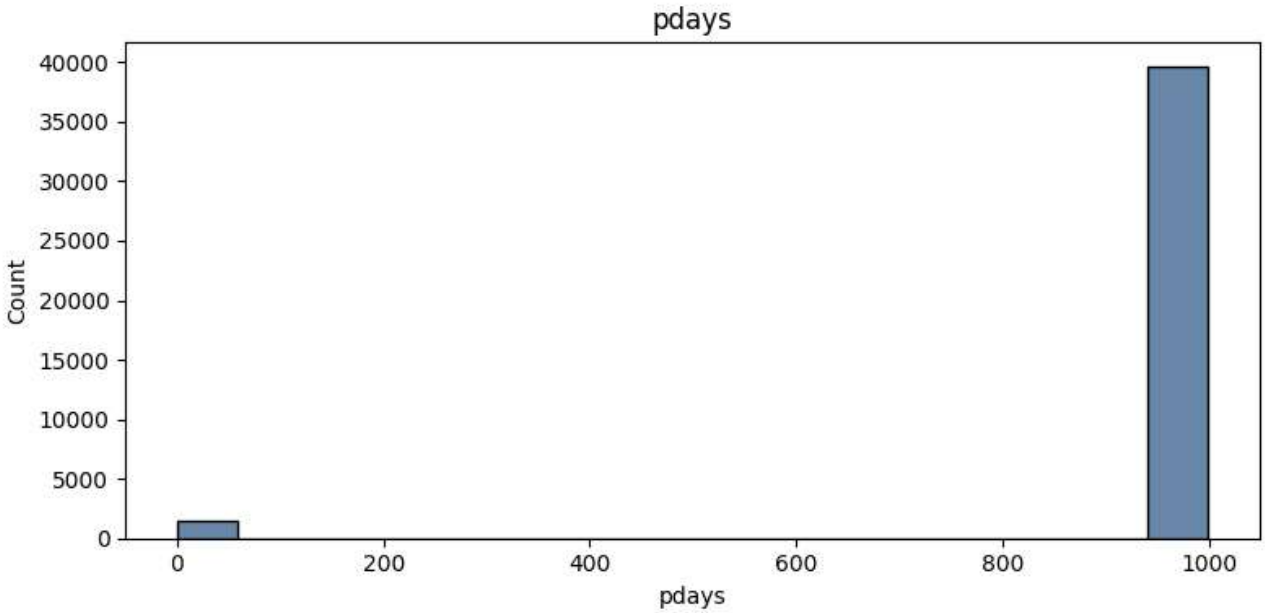


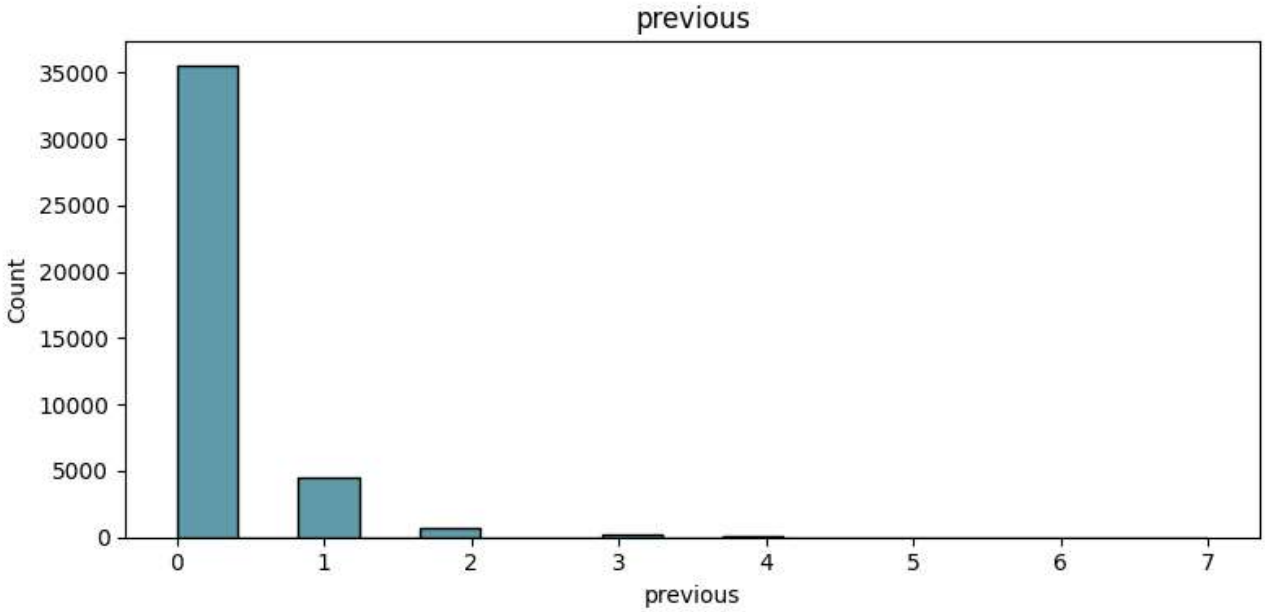


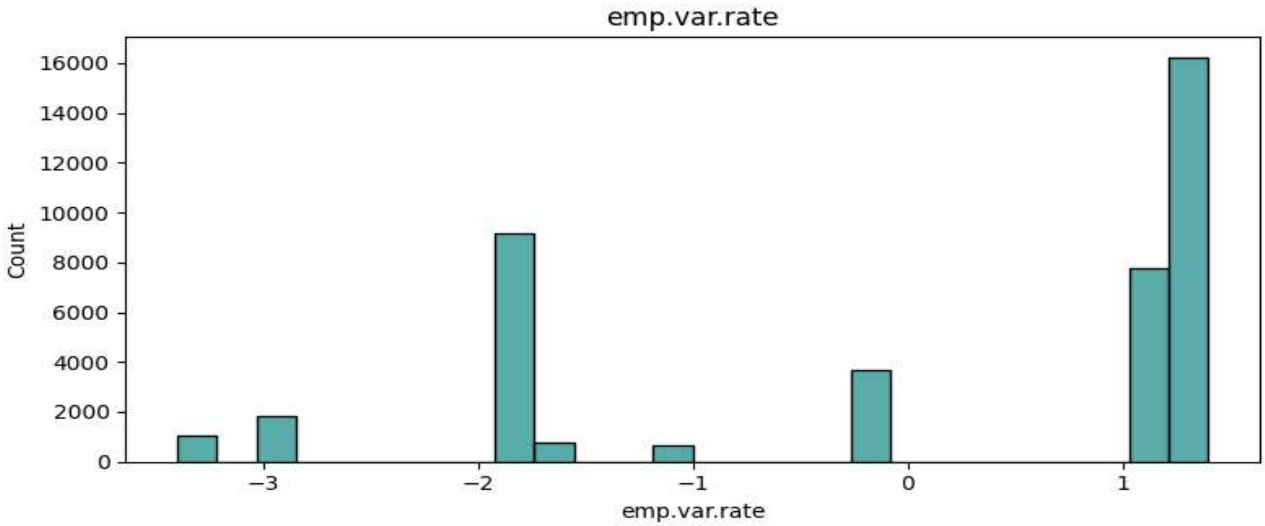
- Exploring numerical features

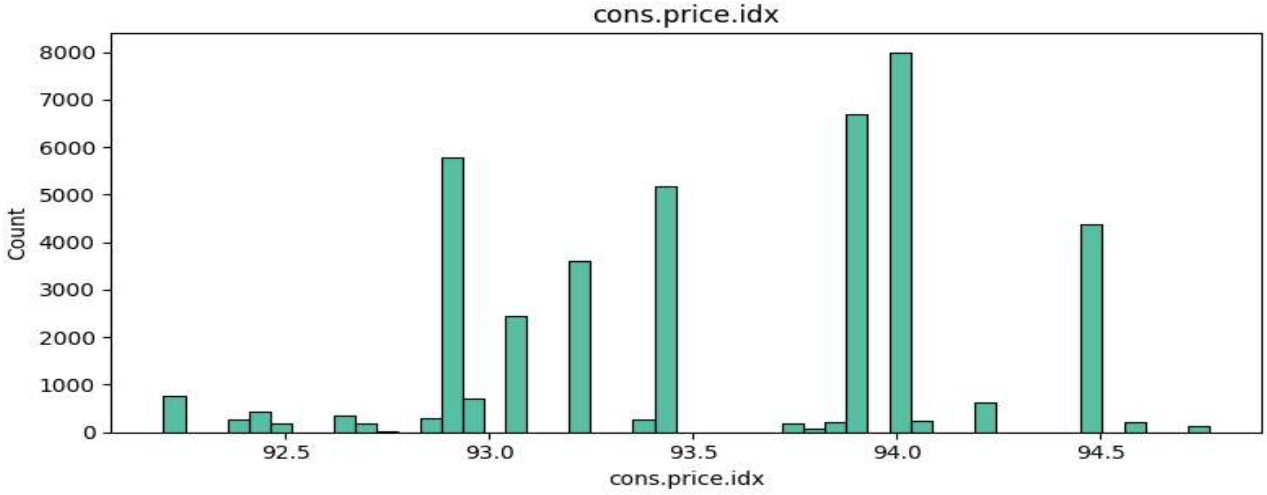

age
Count
2000
1750
1500
1250
1000
750
500
250
0
20
30
40
50
60
70
80
90
100
age

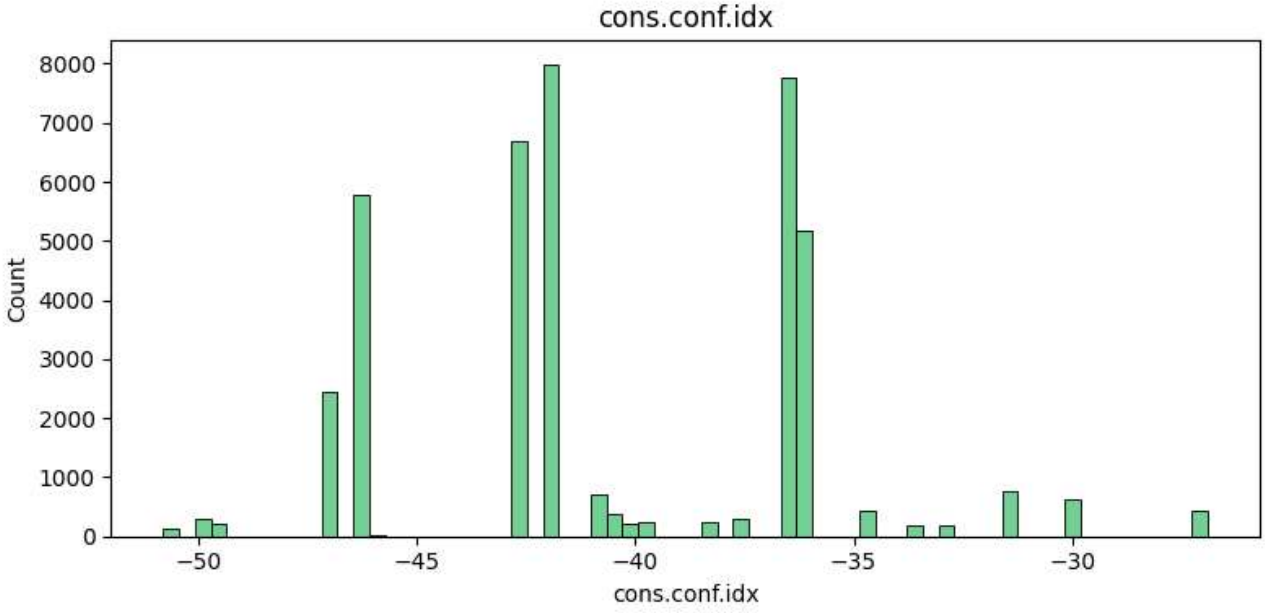

campaign2
Count
17500
15000
12500
10000
7500
5000
2500
0
1
2
3
4
5
6
7
campaign2

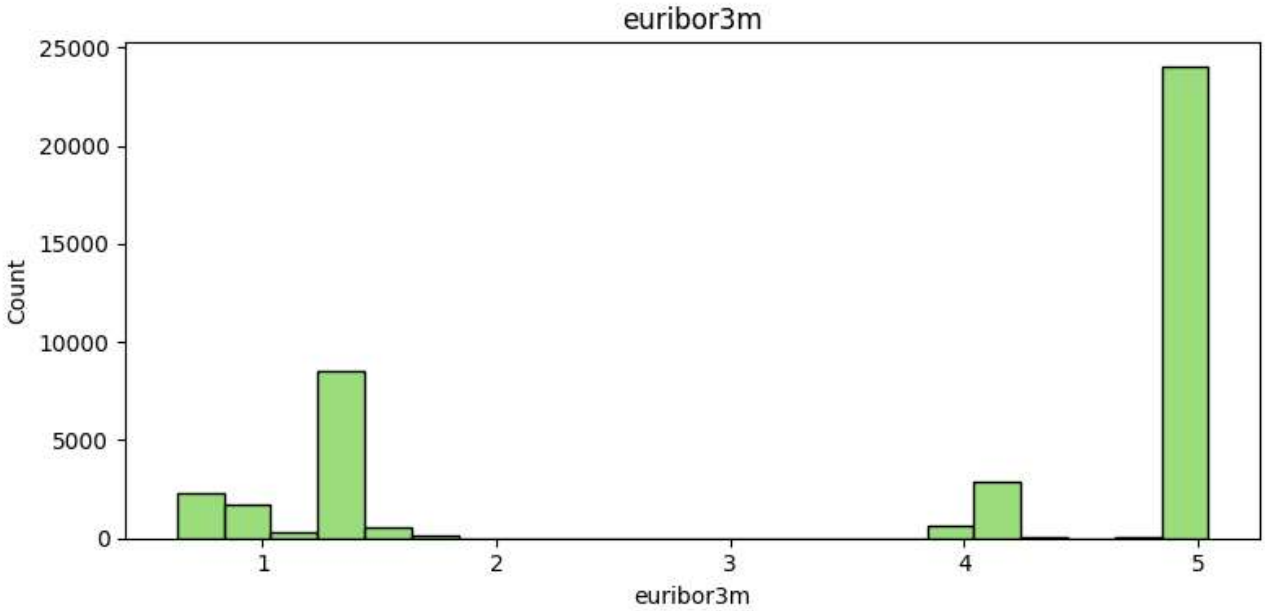

pdays
Count
pdays
0
5000
10000
15000
20000
25000
30000
35000
40000
0
200
400
600
800
1000

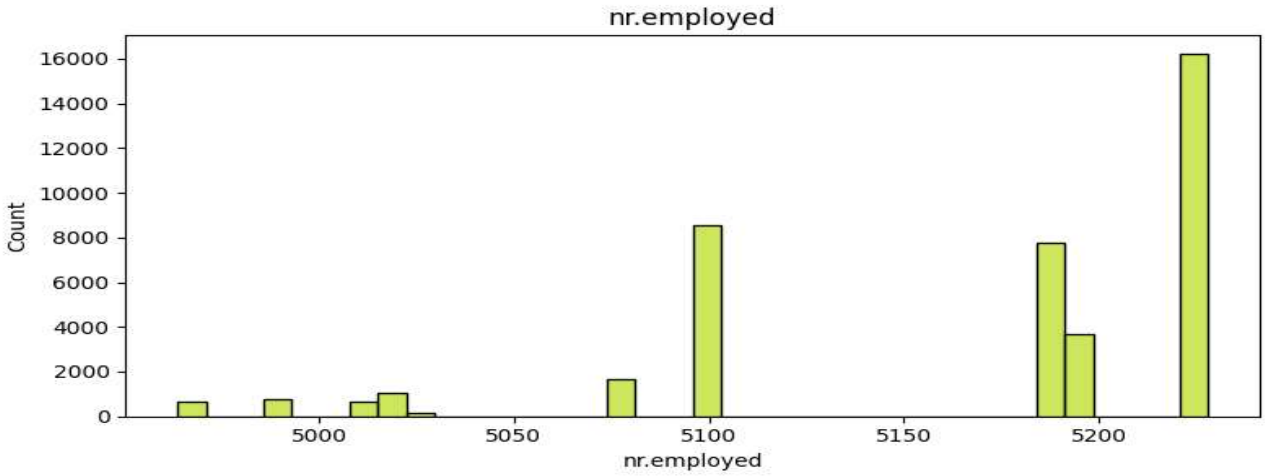

previous
Count
previous
0
5000
10000
15000
20000
25000
30000
35000
0
1
2
3
4
5
6
7


emp.var.rate
Count
emp.var.rate

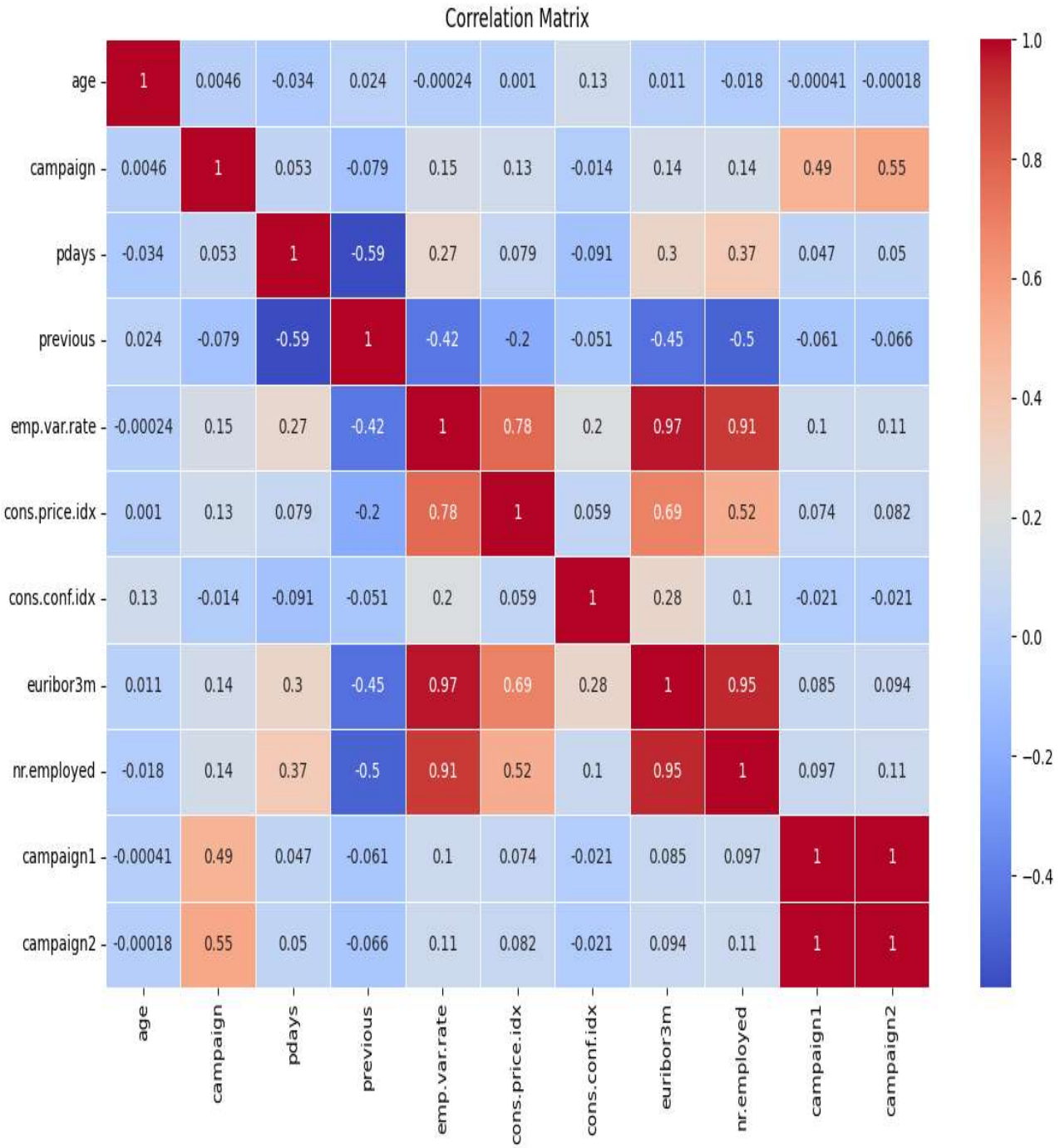

cons.price.idx
Count
cons.price.idx


cons.conf.idx
Count
cons.conf.idx

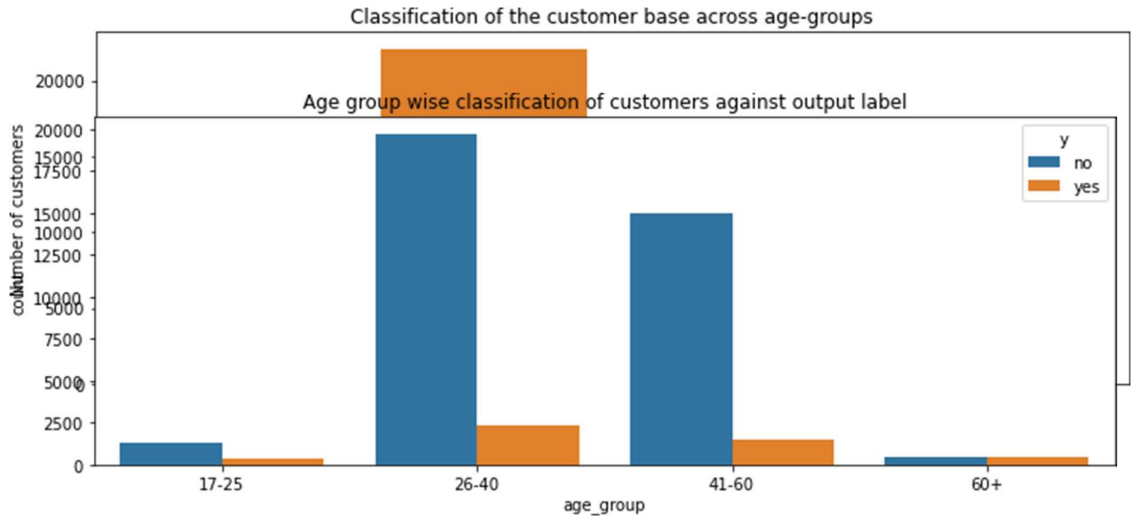




- Correlation matrix


Correlation Matrix
age 1 0.0046 -0.034 0.024 -0.00024 0.001 0.13 0.011 -0.018 -0.00041 -0.00018
campaign 0.0046 1 0.053 -0.079 0.15 0.13 -0.014 0.14 0.14 0.49 0.55
pdays -0.034 0.053 1 -0.59 0.27 0.079 -0.091 0.3 0.37 0.047 0.05
previous 0.024 -0.079 -0.59 1 -0.42 -0.2 -0.051 -0.45 -0.5 -0.061 -0.066
emp.var.rate -0.00024 0.15 0.27 -0.42 1 0.78 0.2 0.97 0.91 0.1 0.11
cons.price.idx 0.001 0.13 0.079 -0.2 0.78 1 0.059 0.69 0.52 0.074 0.082
cons.conf.idx 0.13 -0.014 -0.091 -0.051 0.2 0.059 1 0.28 0.1 -0.021 -0.021
euribor3m 0.011 0.14 0.3 -0.45 0.97 0.69 0.28 1 0.95 0.085 0.094
nr.employed -0.018 0.14 0.37 -0.5 0.91 0.52 0.1 0.95 1 0.097 0.11
campaign1 -0.00041 0.49 0.047 -0.061 0.1 0.074 -0.021 0.085 0.097 1 1
campaign2 -0.00018 0.55 0.05 -0.066 0.11 0.082 -0.021 0.094 0.11 1 1
age
campaign
pdays
previous
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
campaign1
campaign2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

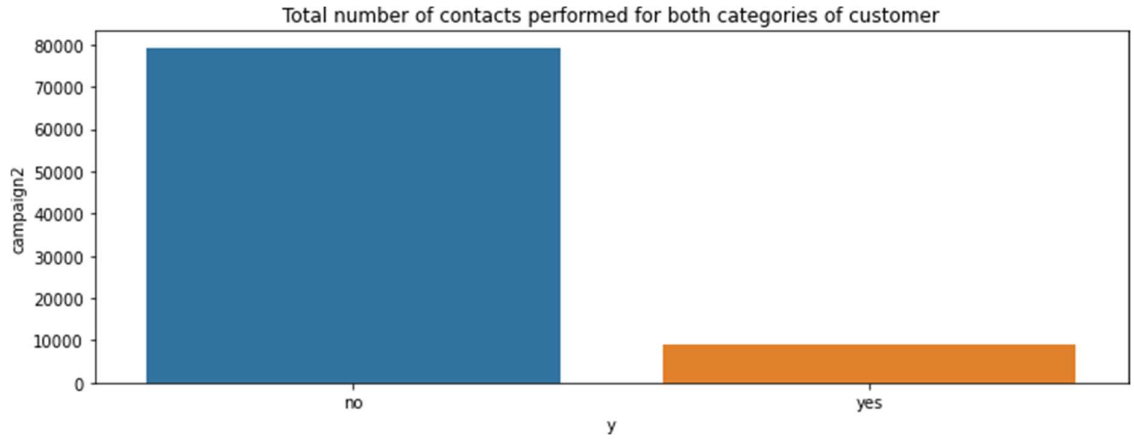
- Classification of the customer base across age-groups



- Looking at relation between different age groups and the output labely

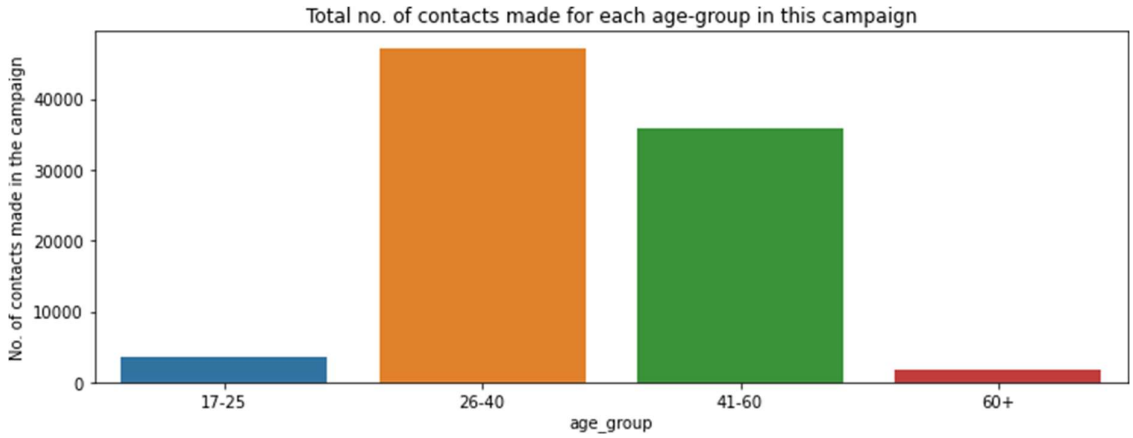
Observation: In the age-groups of 26-40 and 41-60 yrs, majority of thepeople are not subscribed to the term deposit plan

- Looking at relation between Number of contacts made to the customer (campaign) and the output label y



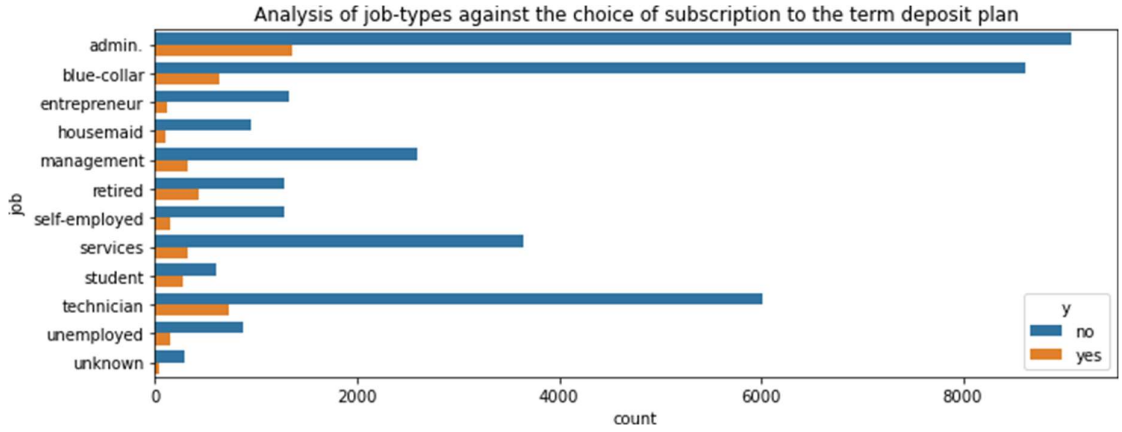
Observation: When a greater number of contacts is made to the customer,they haven't subscribed to the term deposit plan

- Looking at relation between 'age_group' and 'campaign' that is numberof contacts performed for each age group



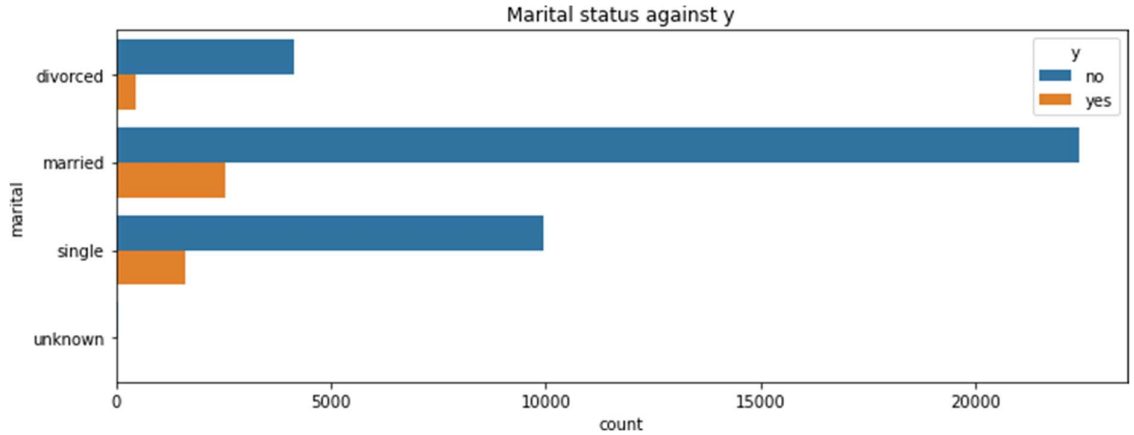
Observation: The 26-40 and 41-60 age-groups witness majority of thecontacts made in this campaign. These two age-groups seem to the target groups for the bank.

- Looking at relation between job and the output label y



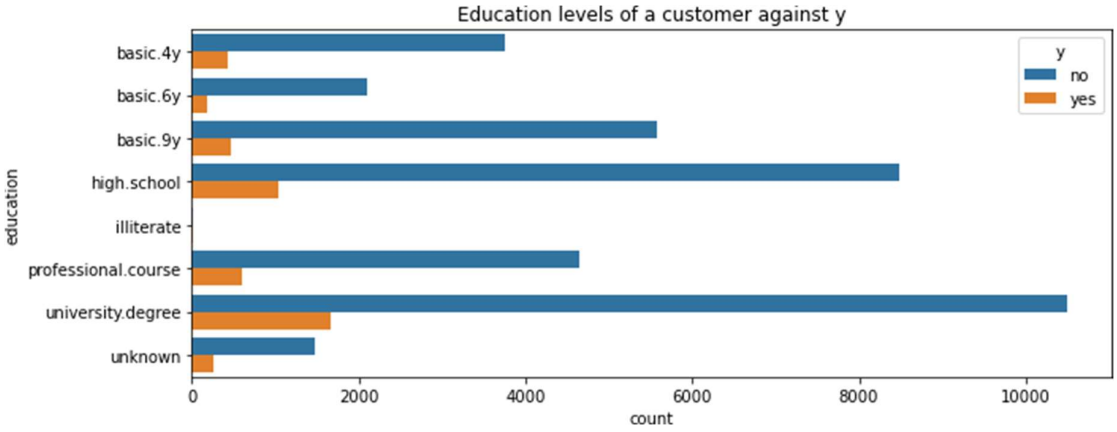
Observation: Looking at the jobs, 'admin', 'blue-collar' and 'technician'are the prominent jobs and most of the customers in these jobs have rejected the term deposit plan.

- Analysing marital status and the output label



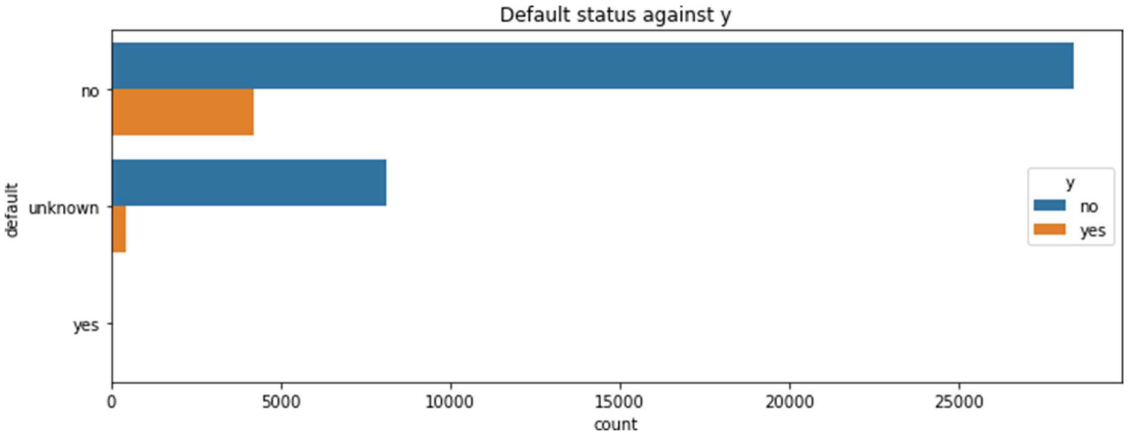
Observation: married and single customers are the majority of the customer base and comparatively married customers have taken theterm deposit

- Analysing the different education levels of a customer against thechoice of subscription

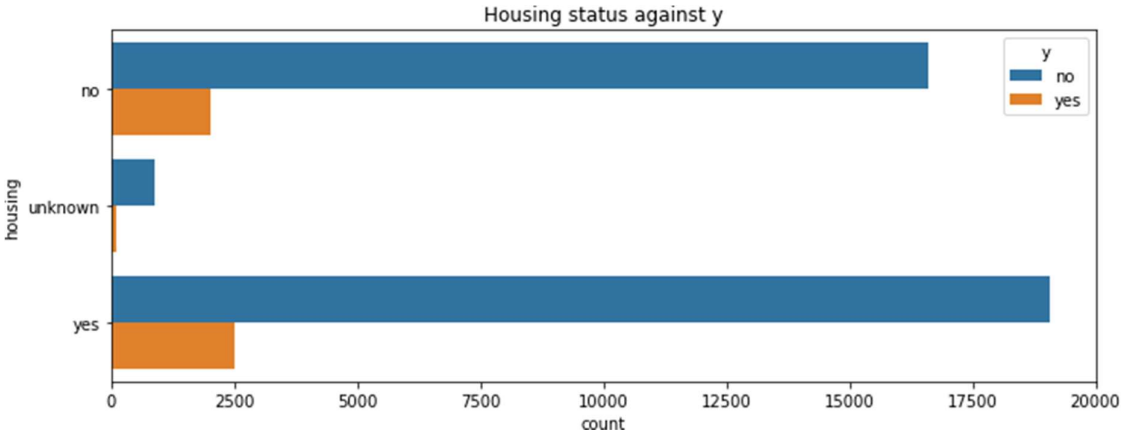


Observation: Customers with university degree have subscribed tothe term deposit more

- Analysing the default status against the choice of subscription

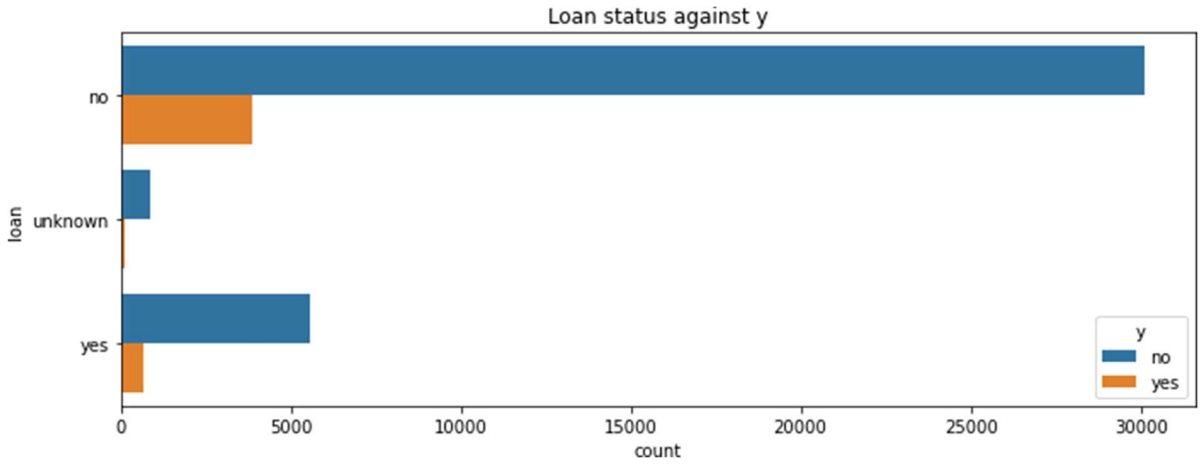


- Analysing housing status and y



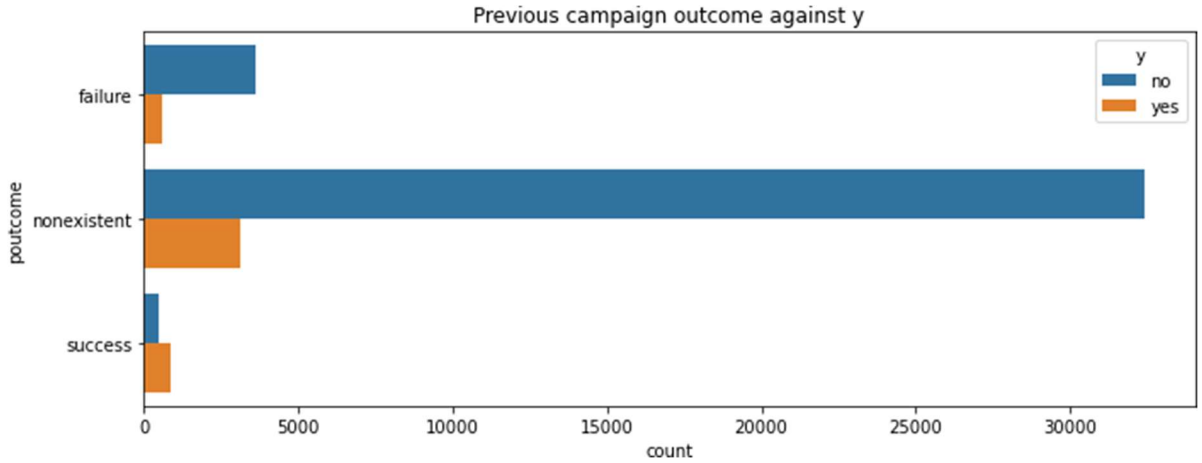
Observation: Number of customers who have subscribed to the termdeposit is comparatively more for those with housing loan

- Analysing loan status and y



Observation: Number of customers who have subscribed to the term deposit is comparatively less for those with personal loan

- Analysing poutcome and y



Observation: The success rate of previous marketing campaign has resulted in more number of people subscribing to the term deposit

Github Repo Link

https://github.com/supinhooda/assignment/tree/main/Week10