EDA and proposed modelling

Project name: Bank Marketing Campaign

Team: Data Science Master

Date: January 12th, 2022

Agenda

Executive Summary

Data Understanding

EDA

EDA Summary

Recommendations

Team Member Details

GROUP NAME: DATA SCIENCE MASTER
NAME : ABHIMANYU GANGANI
EMAIL : Agangani97@gmail.com
COUNTRY : UNITED KINGDOM
COLLEGE : ANGLIA RUSKIN UNIVERSITY
SPECIALIZATION : DATA SCIENCE

Executive Summary

- **Client :**

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which helps them understand whether a particular customer will buy their product or not.

- **Problem Statement :**

Build a Classification ML model to shortlist customers who are most likely to buy the term deposit product. This would allow the marketing team to target those customers through various channels.

- **Analysis :**

The Analysis of this data is divided into the following parts:

1. Data Understanding
2. Univariate analysis
3. Bivariate analysis
4. Model recommendations

Data Understanding :

- **Dataset Description :**

  Four Different datasets provided.
  1. bank-additional-full: 20 inputs (+1 target variable) and 41119 observations
  2. bank-additional: 20 inputs (+1 target variable) and 4119 observations b
  3. bank-full: 17 inputs (+1 target variable) and 45211 observations
  4. bank: 17 inputs (+1 target variable) and 4521 observations

- **Assumptions :**

  1. Timeline of observations - May 2008 to November 2010.
  2. ‘Duration’ feature is dropped to give realistic predictions from the classification model.
  3. A frequently occurring missing value ‘unknown’ is considered as another category for the categorical features.
  4. Duplicated rows were deleted from the dataset

Data Understanding :

Data columns (total 21 columns):

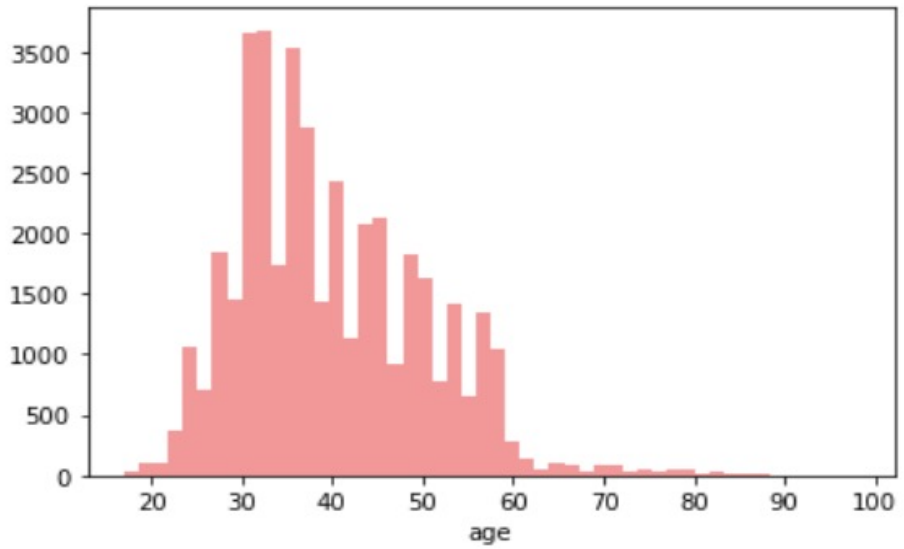
#	Column	Dtype	Description
0	age	int64	Age of Client.
1	job	object	Type of Job.
2	marital	object	Marital Status.
3	education	object	Level of Education.
4	default	object	Has credit in default?
5	housing	object	Has housing loan?
6	loan	object	Has personal loan?
7	contact	object	How client has been communicated?
8	month	object	last contacted month.
9	day_of_week	object	last contacted day.
10	duration	int64	duration of communication(seconds).
11	campaign	int64	number of contacts performed in Campaign.
12	pdays	int64	number of days passed after contact.
13	previous	int64	number of total contacts performed.
14	poutcome	object	outcome of the previous campaign.

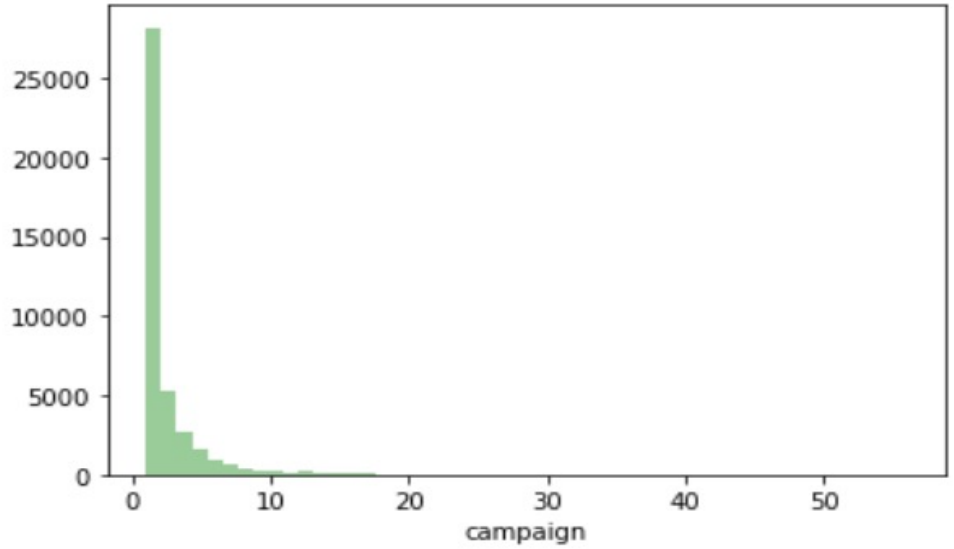
Data Understanding :

Data columns (total 21 columns):

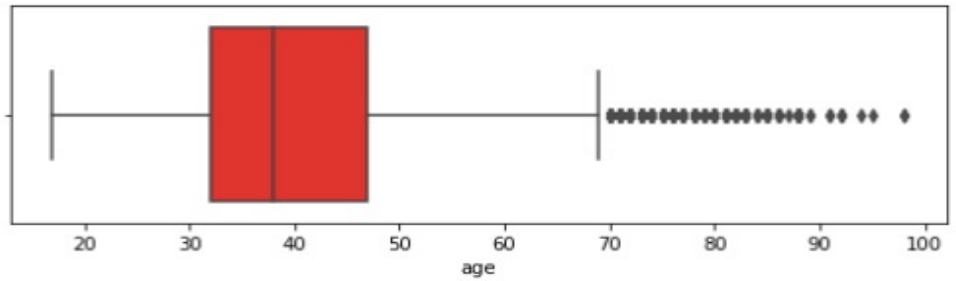
#	Column	Dtype	Description
15	emp.var.rate	float64	Employment variation rate.
16	cons.price.idx	float64	Consumer price index.
17	cons.conf.idx	float64	Consumer confidence index.
18	euribor3m	float64	Euribor 3 months rate.
19	nr.employed	float64	number of employees.
20	y	object	has the client subscribed product.

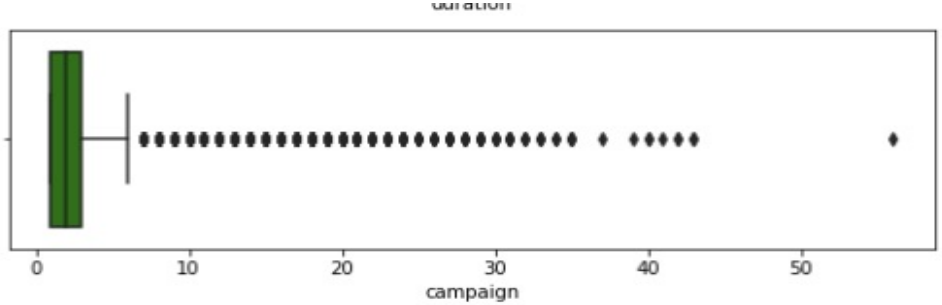
EDA (Univariate Analysis)



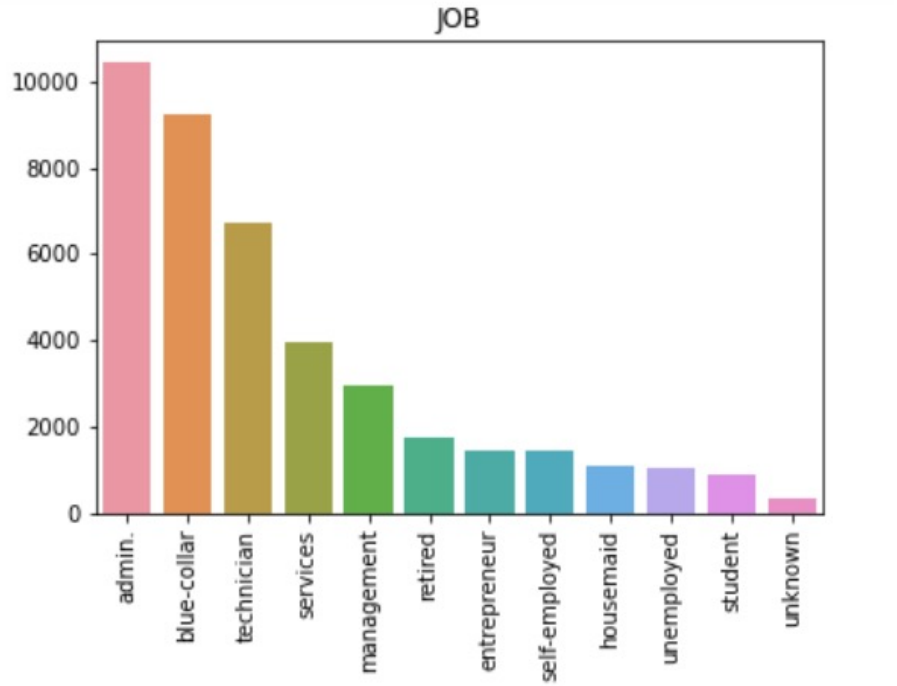


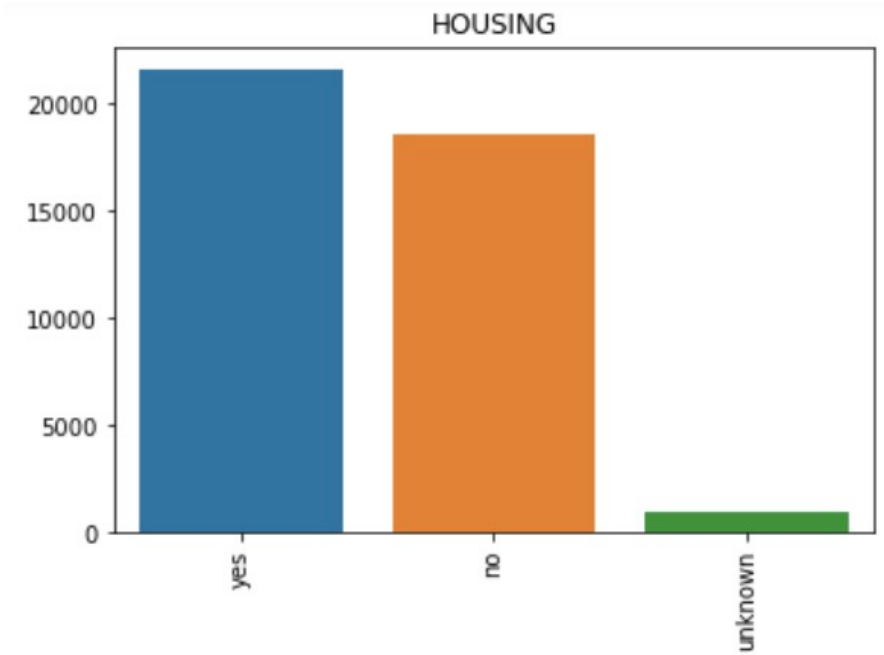
- Positive Skewness observed in Age and Campaign Features.
- 97% of clients falls in 20 to 60 years of age.
- Box plots of ‘age’ and ‘campaign’ features show that data points outside the whiskers are outliers.
- No outlier treatment is required since outliers value are realistic.





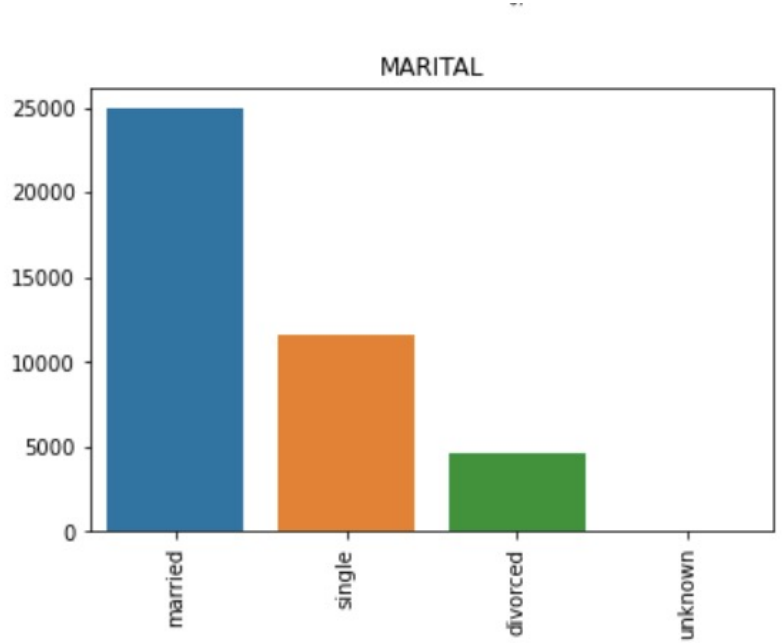
EDA (Univariate Analysis)

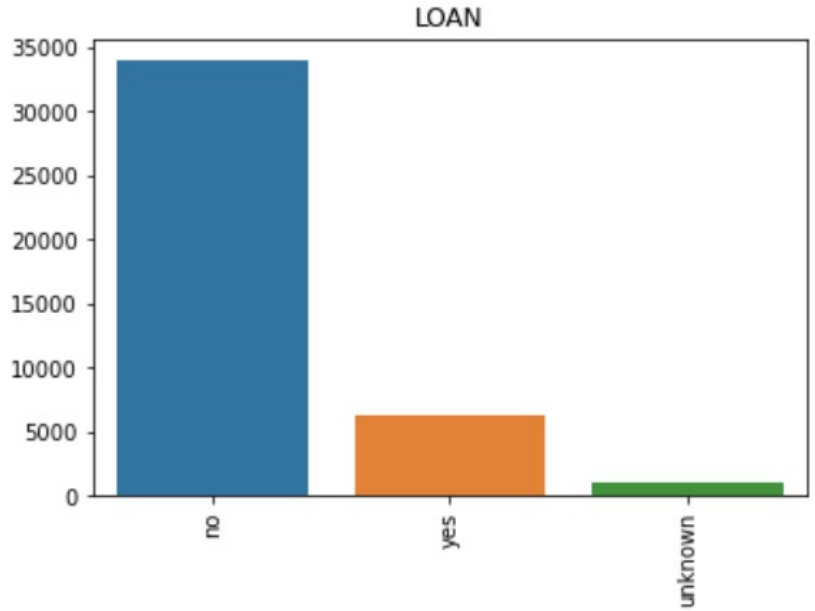




- Majority of Clients are employed as Admin, Blue-collar and technical jobs.
- There are almost equal number of people with and without the housing status.

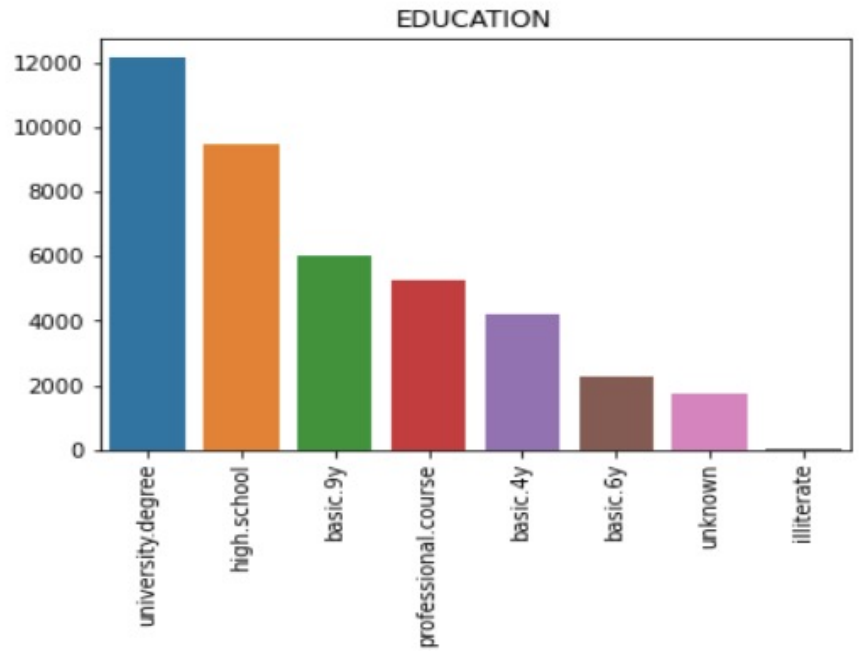
EDA (Univariate Analysis)

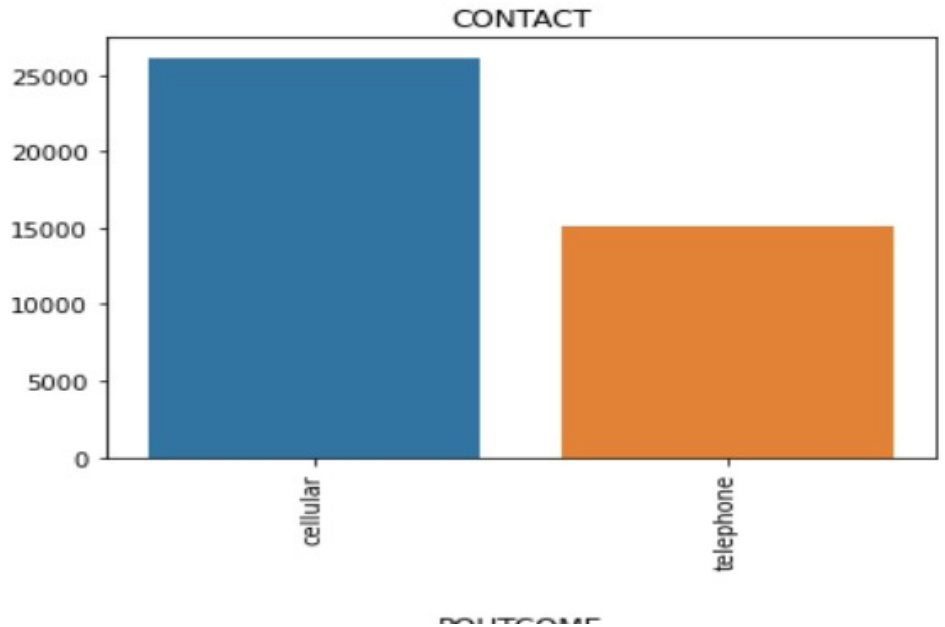




- In the given customer base only majority of peoples are married (25000) and its more than rest of categories combined.
- Most of the customer base has never opted any kind of loan.

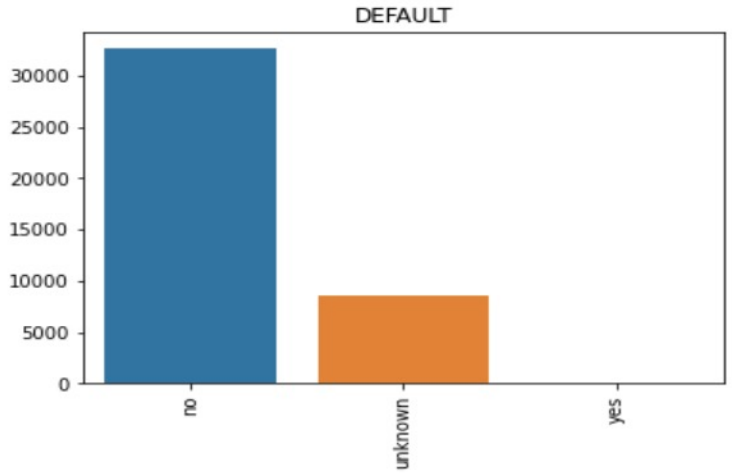
EDA (Univariate Analysis)

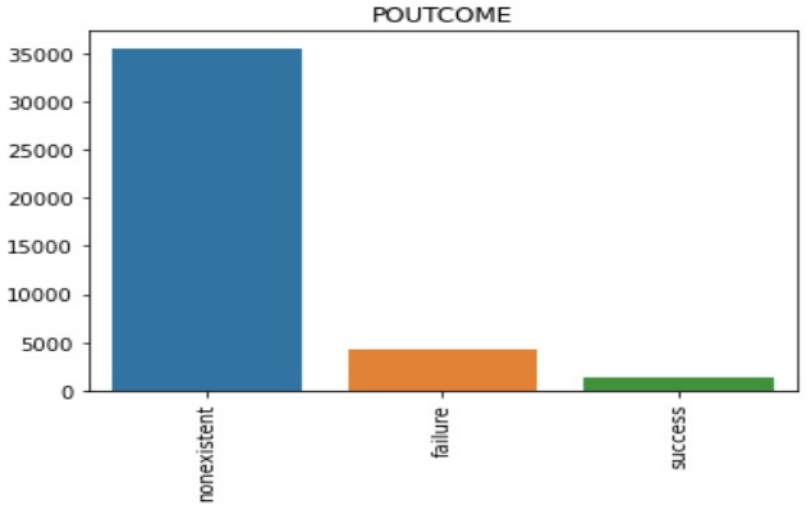


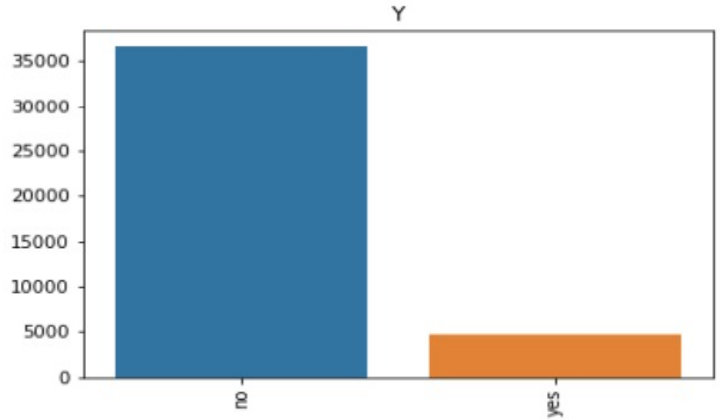


- Cellular communications were made nearly double than telephone.
- Most of the clients in the database are high school or university qualified.

EDA (Univariate Analysis)

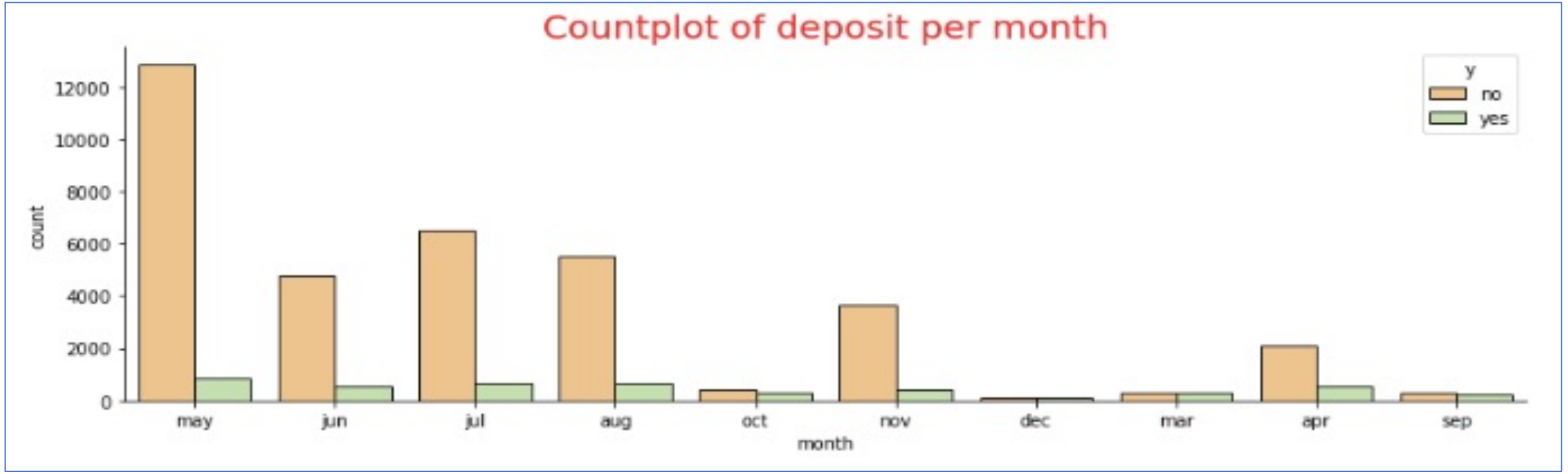






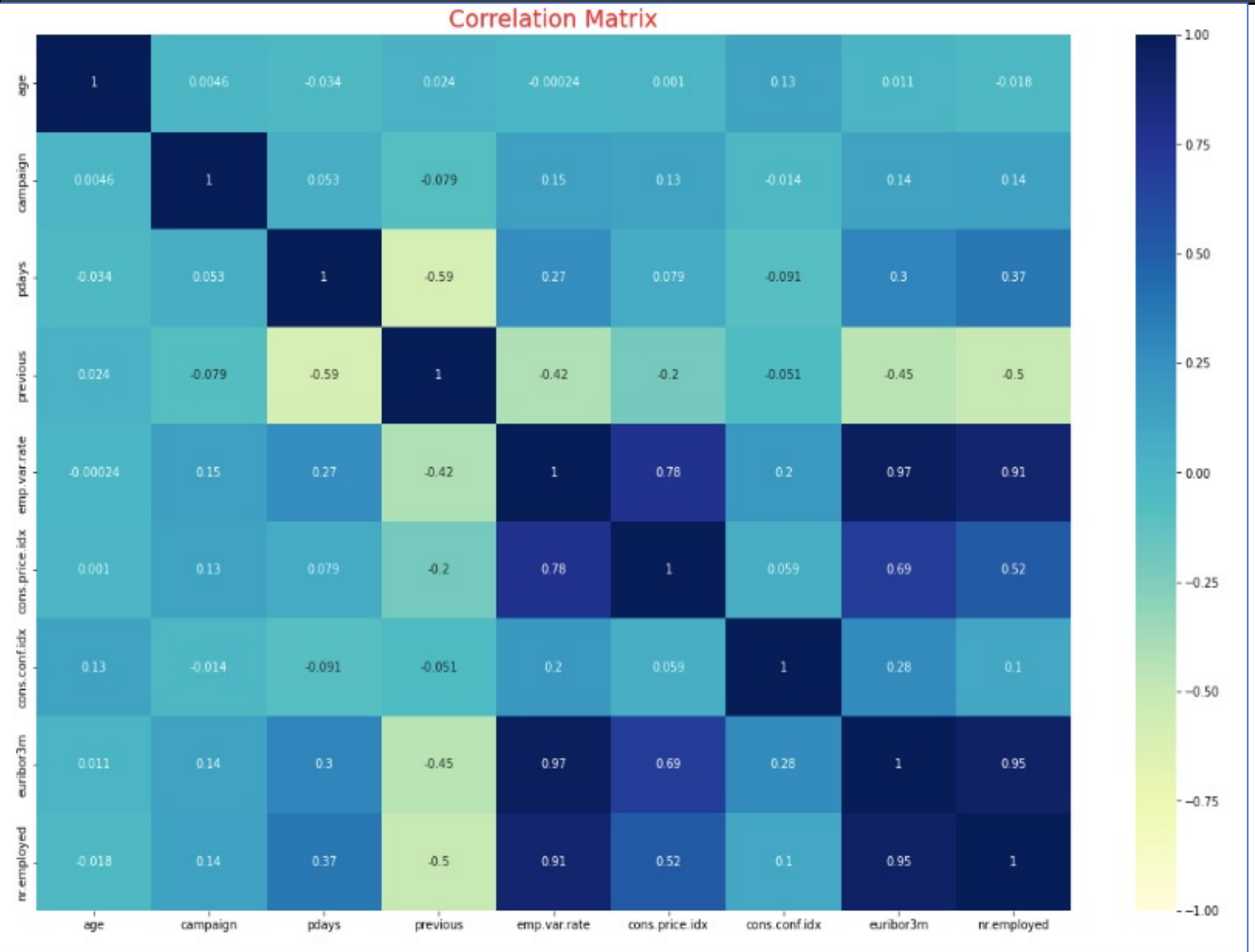
- Among the customer base, more than 35000 of them have rejected the term deposit plan
- Out of the total calls that are made, 11.3 % of them are successful.
- There are no defaulters present in data but count of unknown defaults is concerning.
- For over 35000 clients previous contact was not made.

EDA (Univariate Analysis)



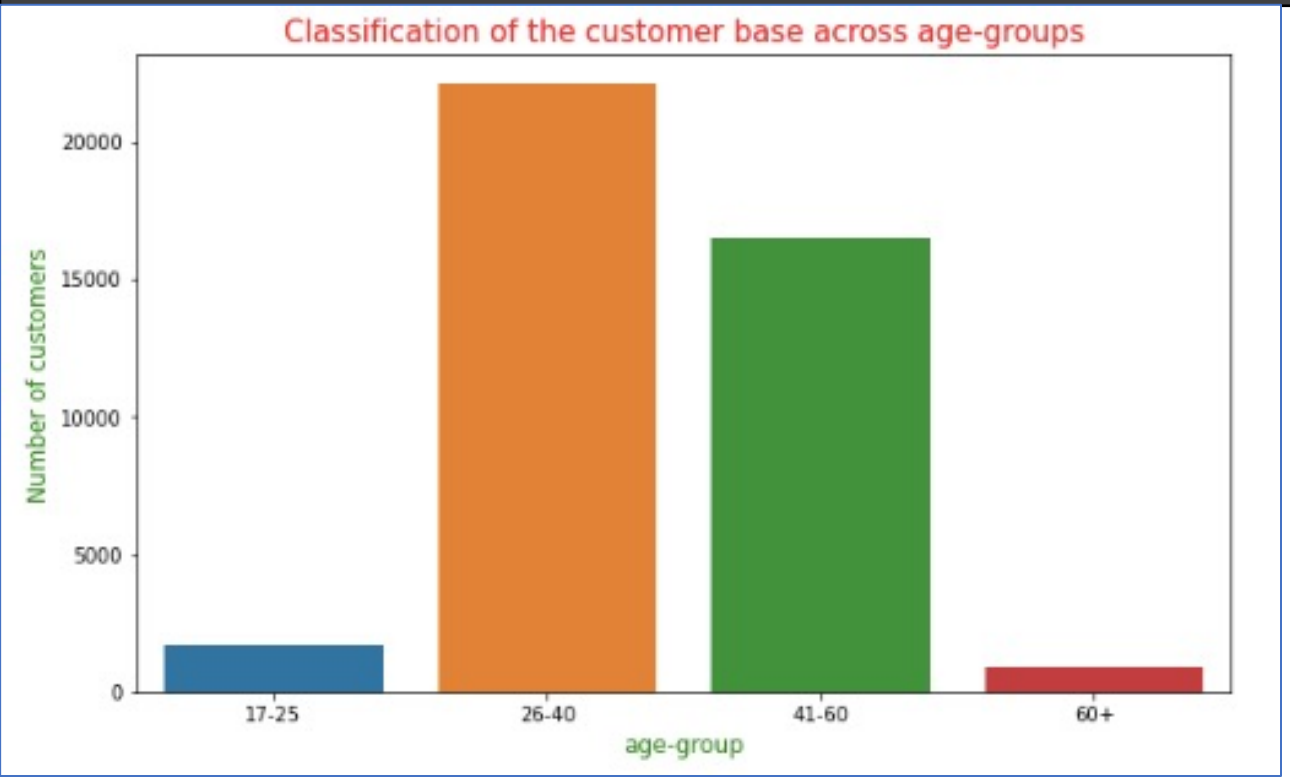
The most number of successful conversions were observed in the month of may as shown by above plot.

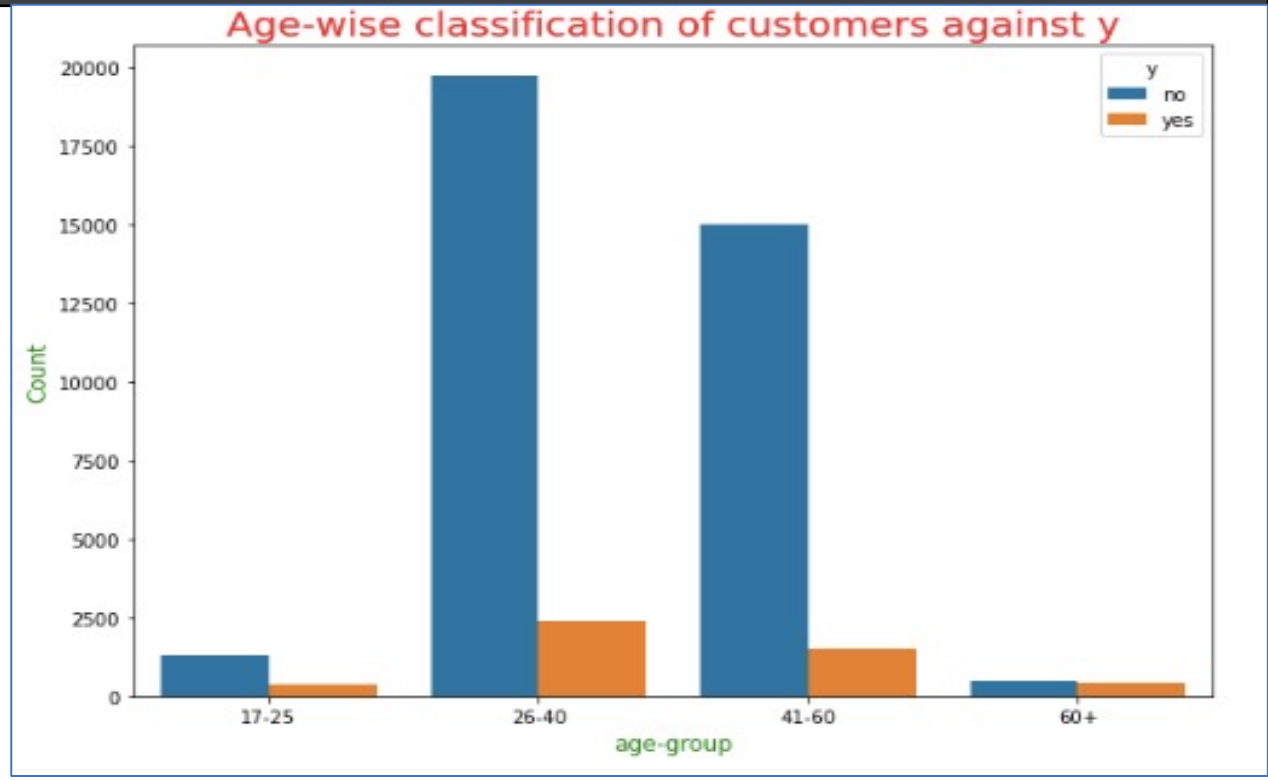
EDA Correlation Analysis(Bivariate Analysis)



Highly Correlated features (consumer price index, consumer confidence index and employment rate)may describe client state from different socio-economic angles. The variance might support model capacity for generalization.

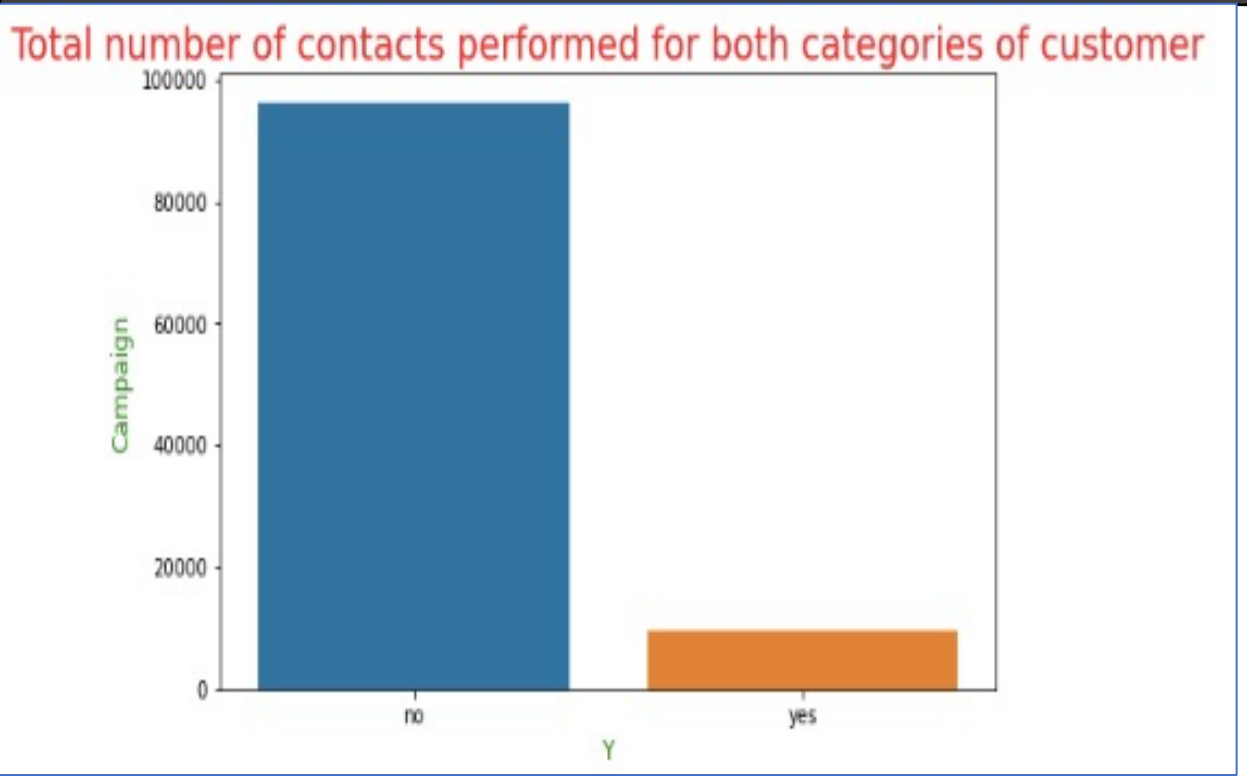
• EDA (Bivariate Analysis) # Age-Group

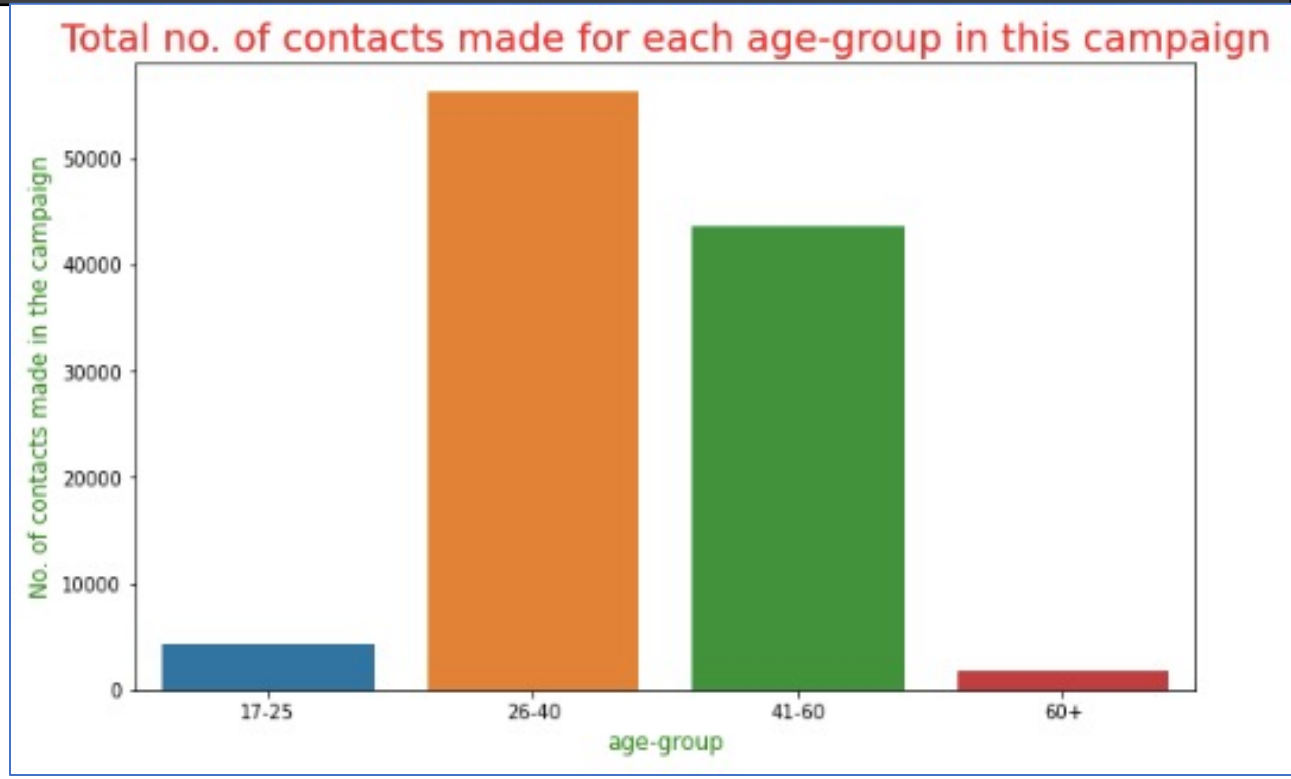




- Using binning we have created new feature age_group using attribute age to get better visualization.
- Around 97% of clients lies between age group 26-40 and 41-60.
- Approximately 82% of clients have rejected the term deposit plan in above mentioned age-groups.

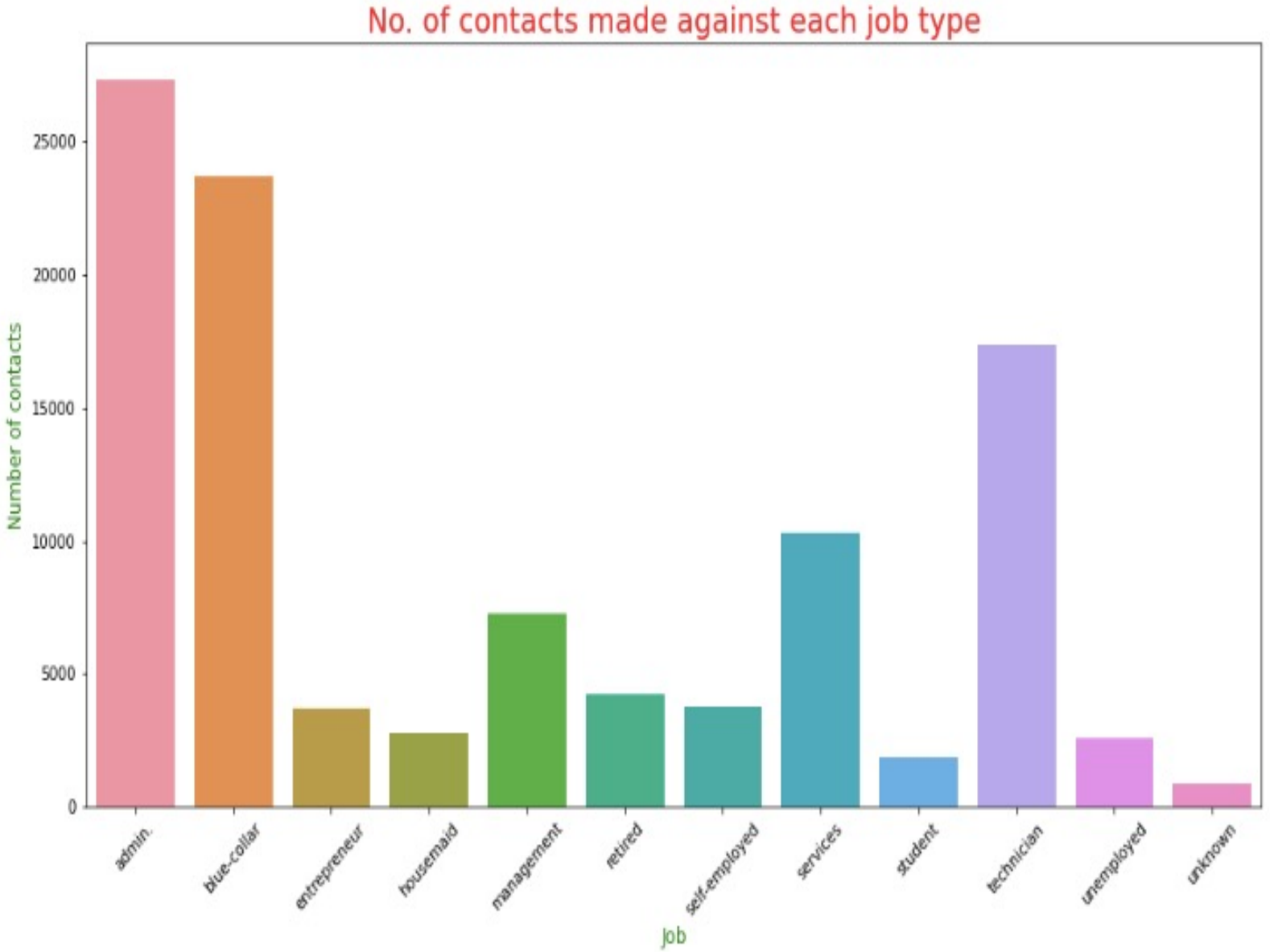
• EDA (Bivariate Analysis) # Campaign





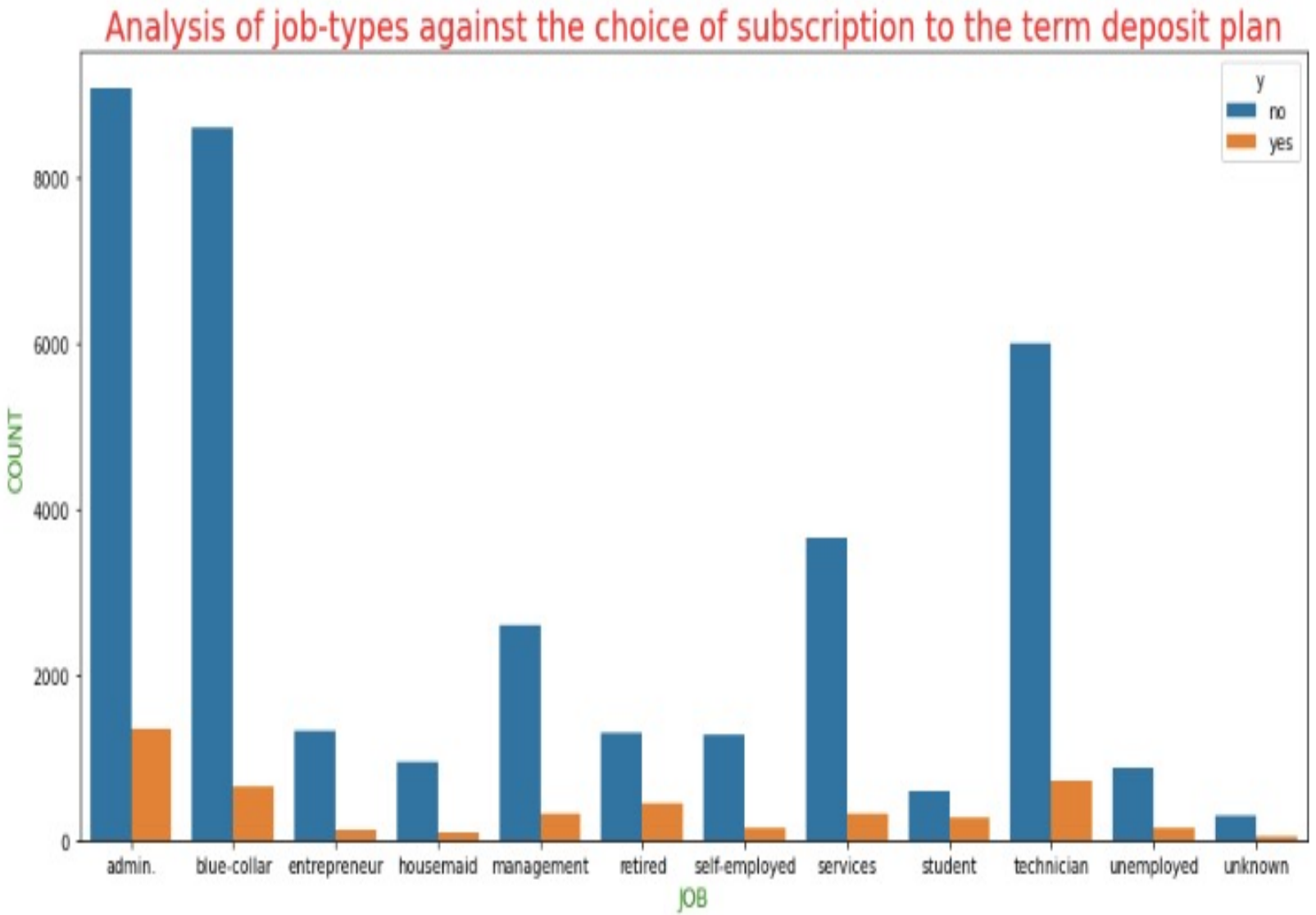
- Checking the number of contacts made in the marketing campaign.
- The age group 26-40 witnessed above 50000 contacts in the current marketing campaign.
- Over 90000 customers who have been contacted have rejected the term deposit plan.
- Contacts performed for age-group 26-40 is more then combined count of other age-group 17-25, 41-60, 60+.

• EDA (Bivariate Analysis) # JOB



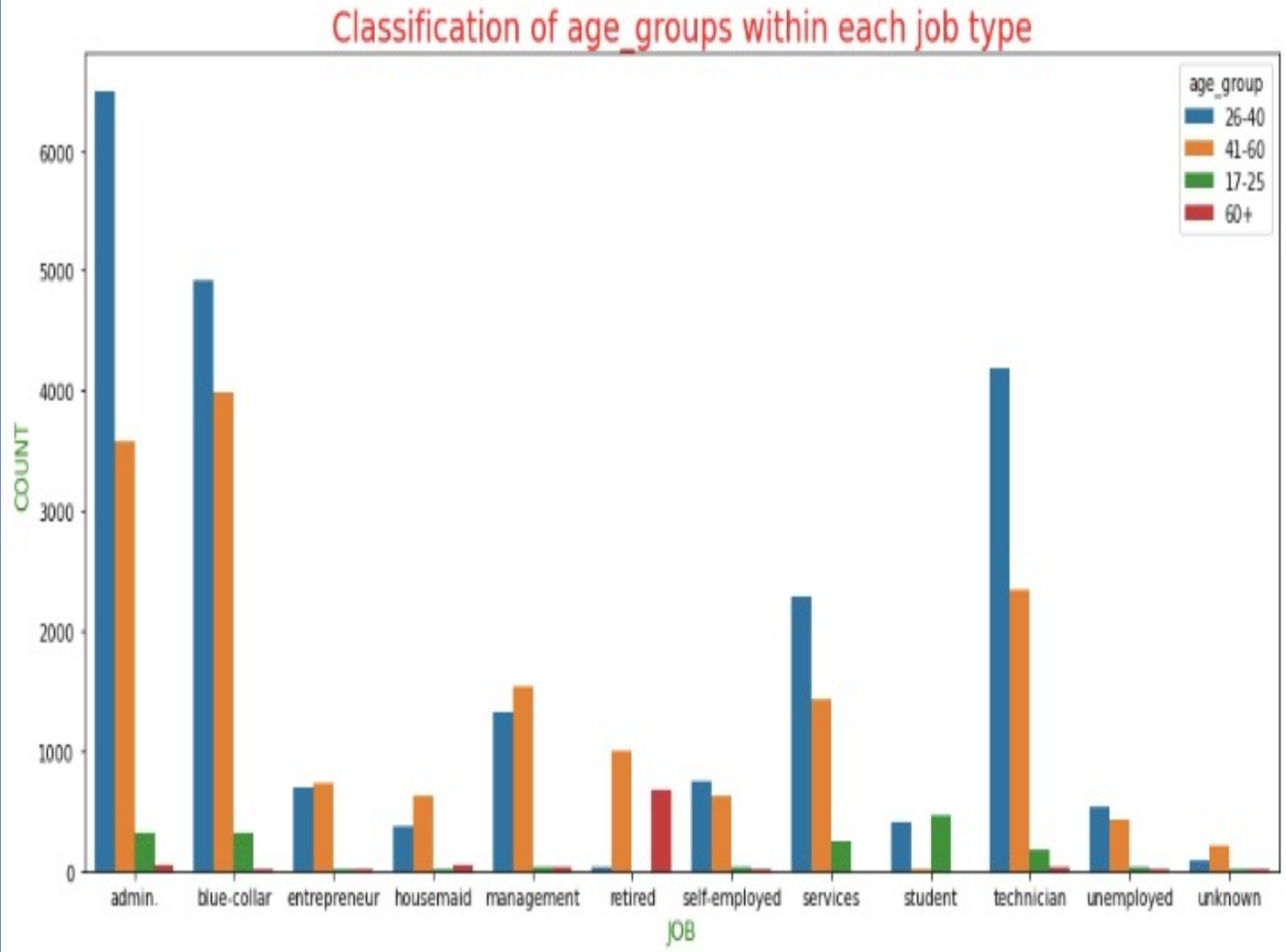
- The customers in professions of ‘admin.’, ‘blue-collar’ and ‘technician’ have majorly rejected the term deposit plan. The former three jobs comprise approximately 64% among the customer base.

• EDA (Bivariate Analysis) # JOB



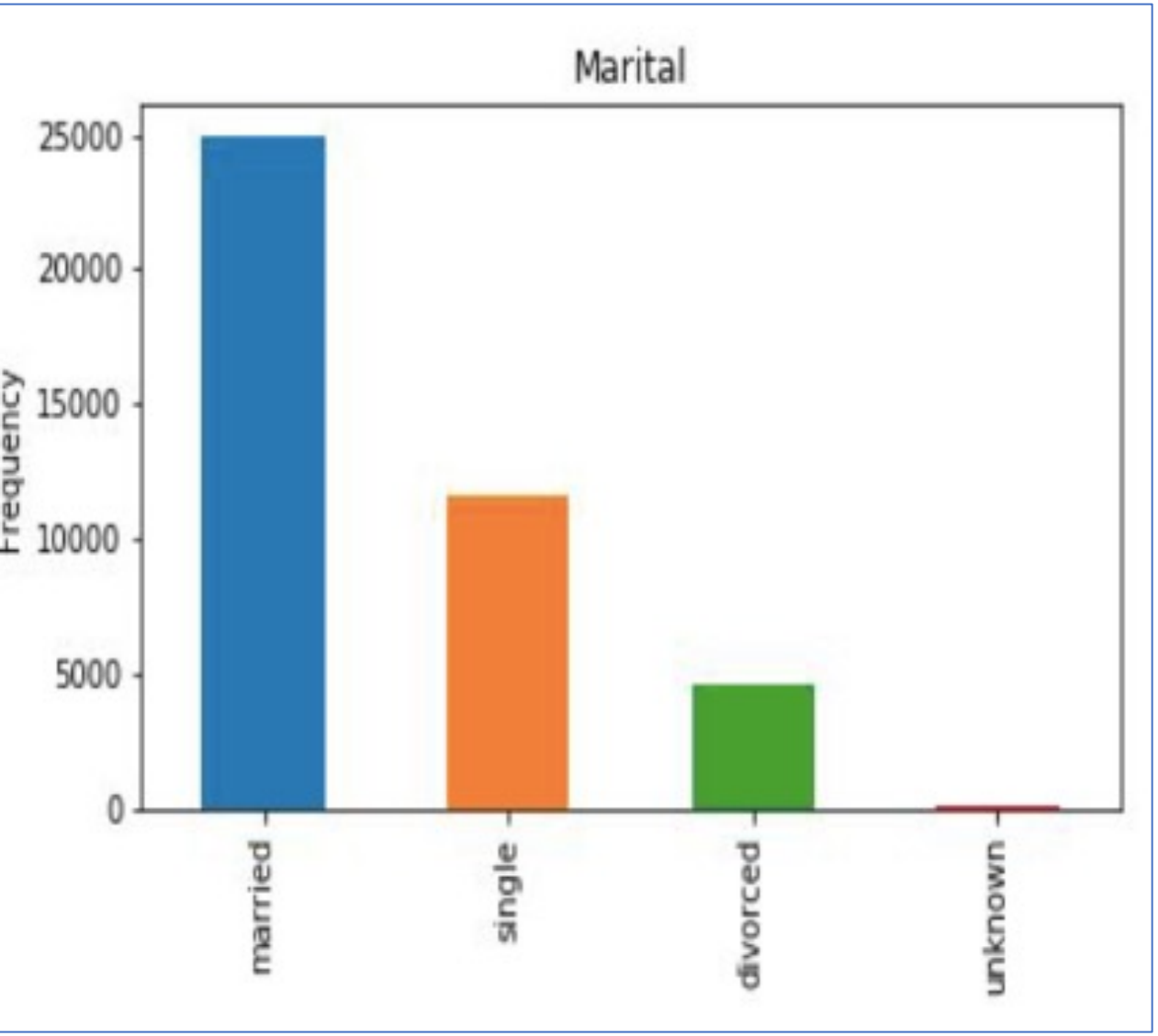
- Only 13.5% of customers in above mentioned jobs have accepted the term deposit plan.

• EDA (Bivariate Analysis) # JOB



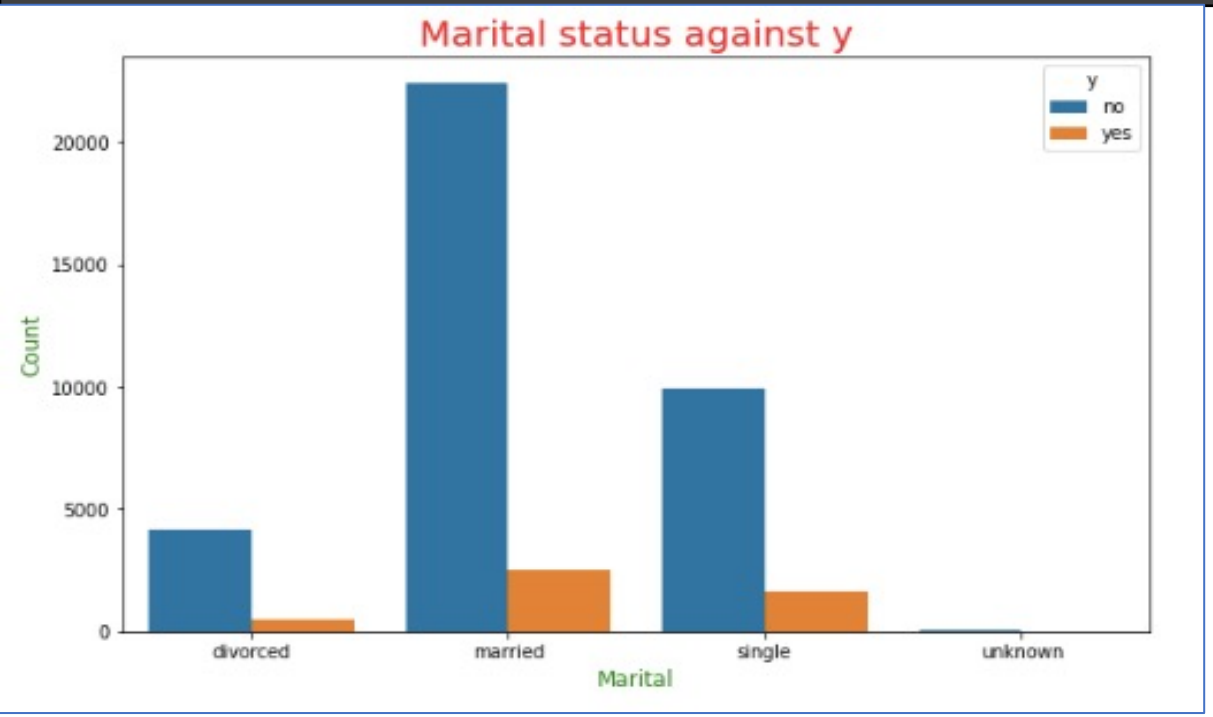
- Age-group 26-40 and 41-60 are prominent in admin, blue-collar, technician and services jobs.
- In Management, entrepreneur and housemaid age group 41-60 possess most jobs.
- There are less then 200 people who have unknown as job category.

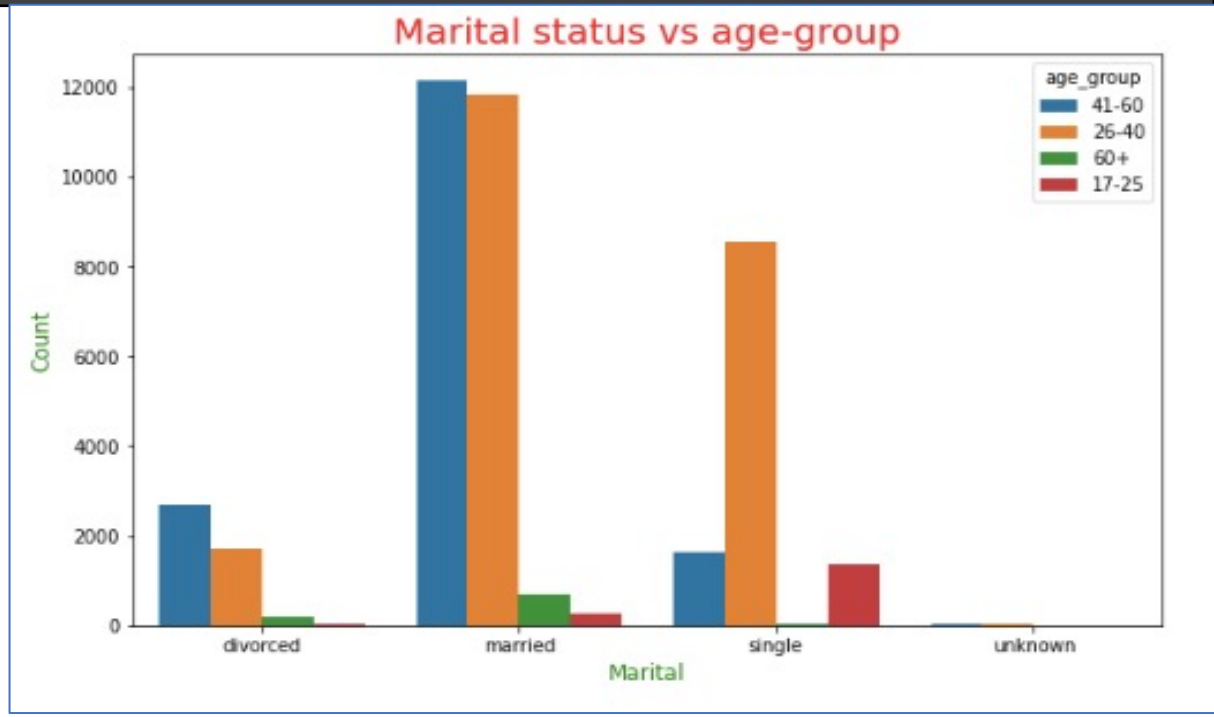
• EDA (Bivariate Analysis) # Marital



- Around 50%(25000) of customers are married and close to 10000 are singles.
- There are barely any customers with unknown marital status.

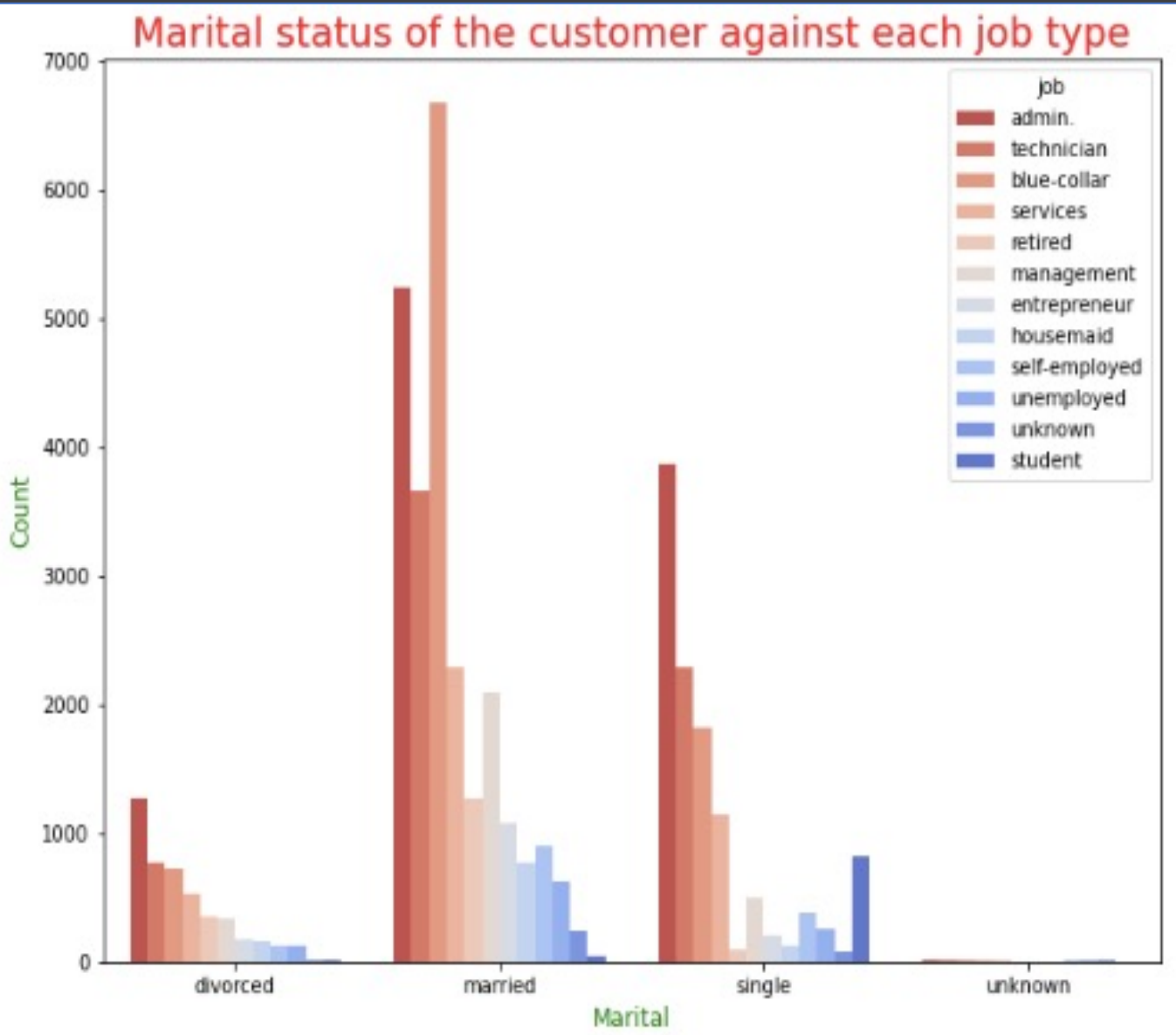
• EDA (Bivariate Analysis) # Marital





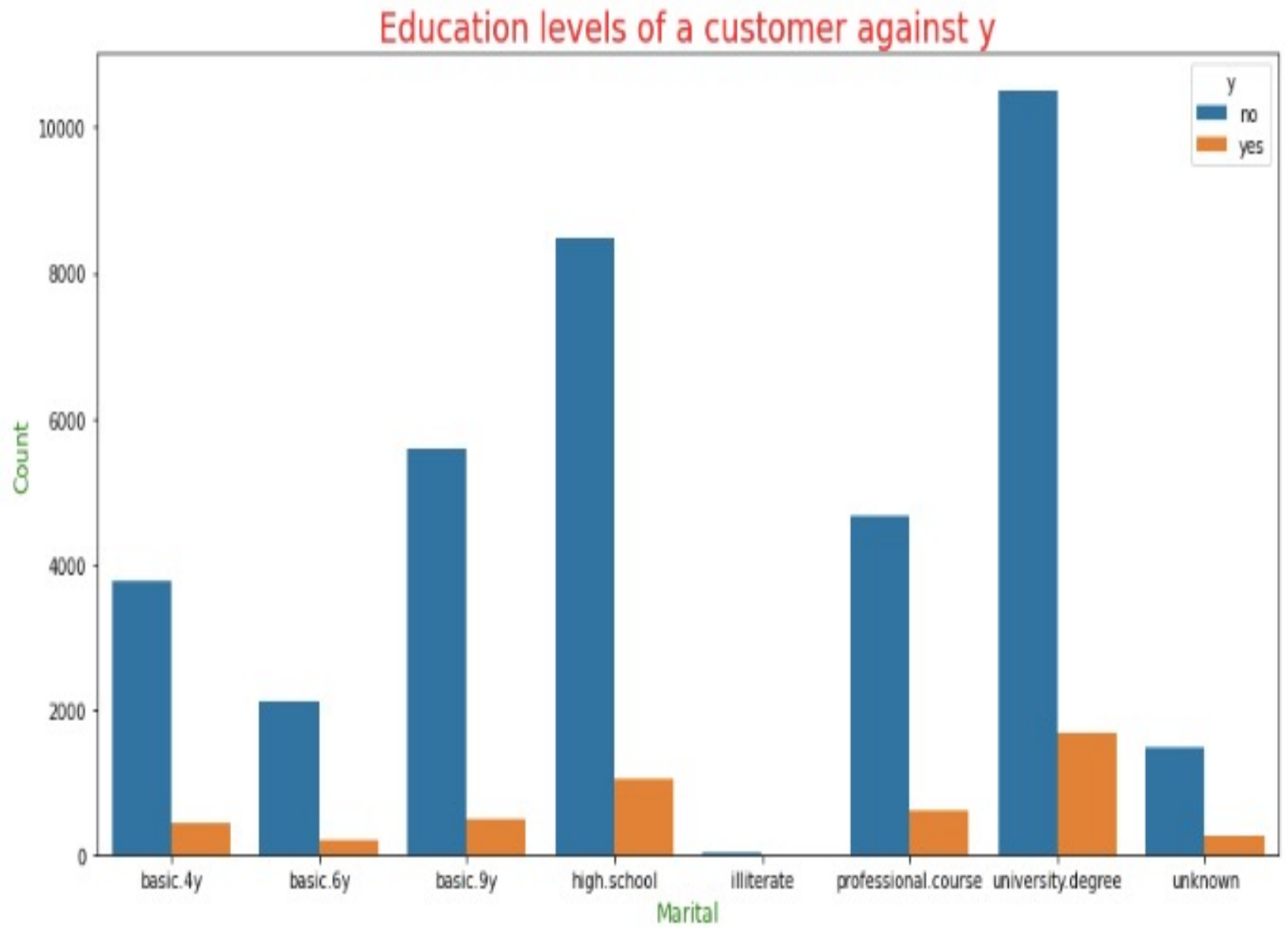
- The number of married people agrees more to the term-deposit service but relatively single people opts in more.
- Number of single people in age category 26-40 is relatively high then any other class.

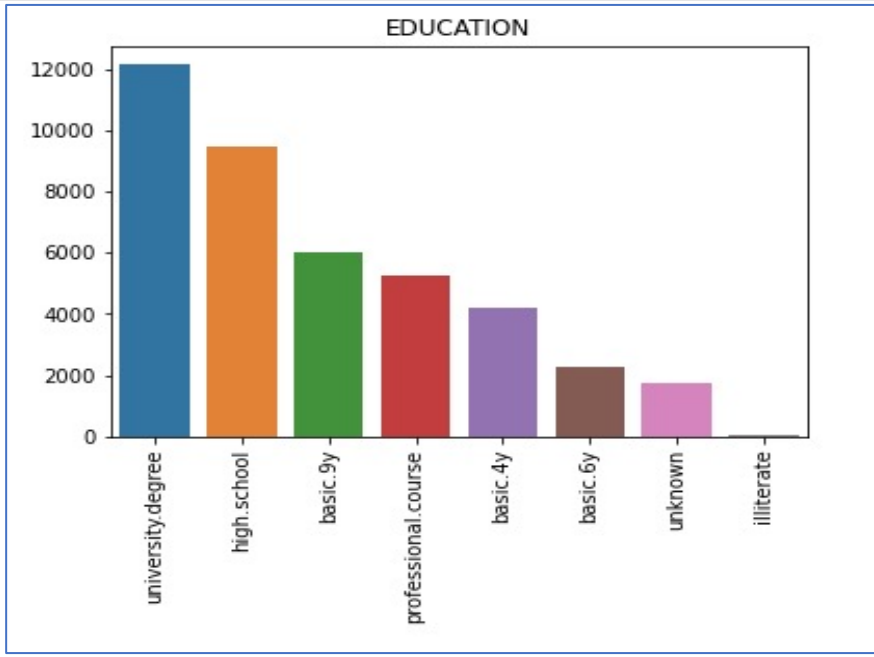
• EDA (Bivariate Analysis) # Marital



- Married people in admin, blue-collar and technician jobs are around 15000 in number.
- Among single customers, there are almost double admins then technicians.
- Married people are mostly having blue-collar jobs.

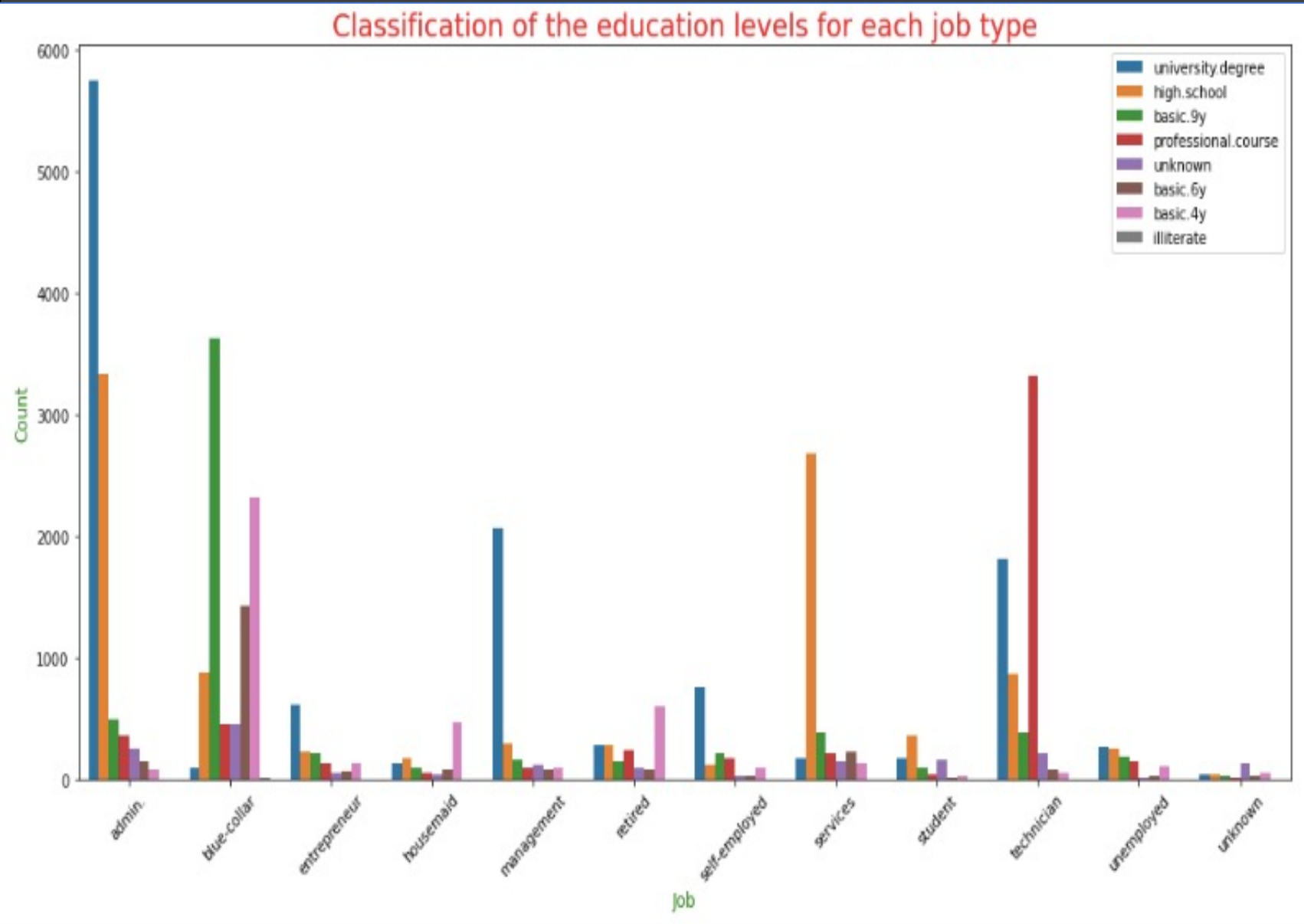
• EDA (Bivariate Analysis) # Education





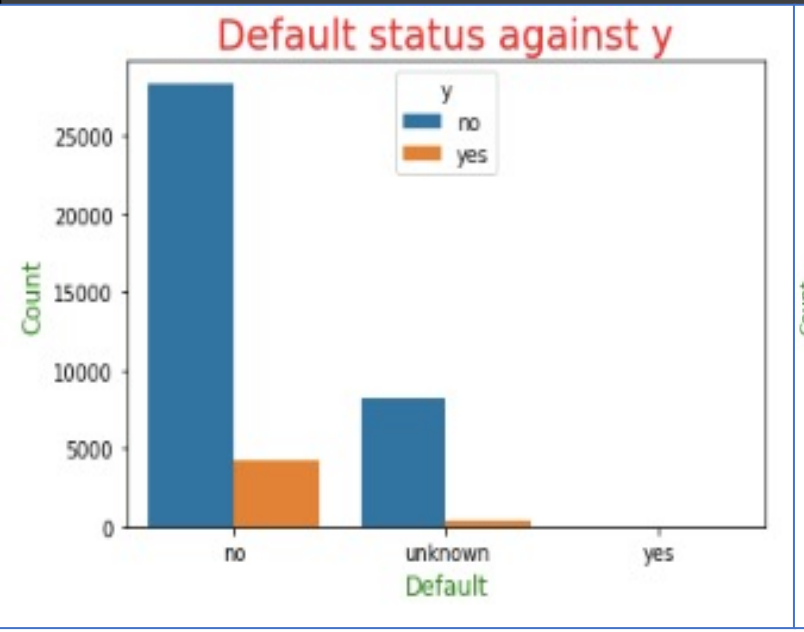
- There are nearly 1000 clients with unknown education status.
- Over 10000 customers having university have rejected the term deposit plan.

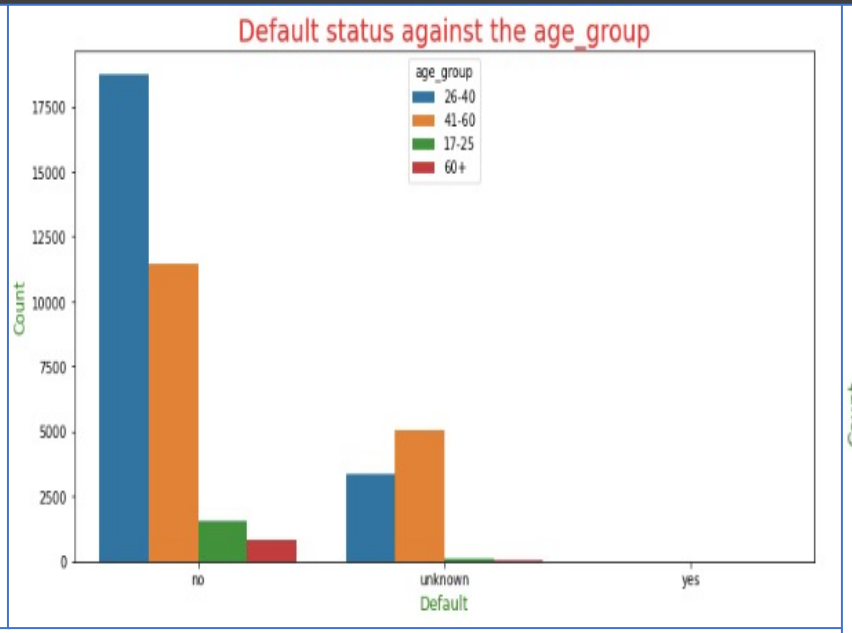
• EDA (Bivariate Analysis) # Education

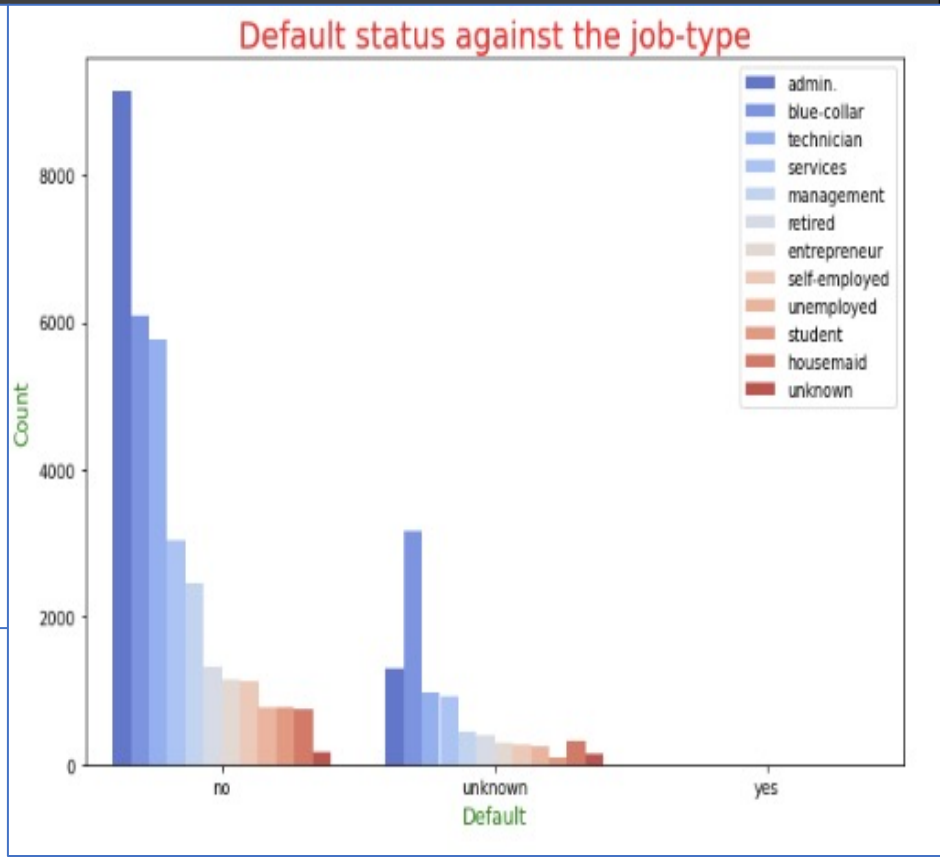


- Admins, Technician and management possessing jobs are having most of clients qualified through universities.
- Most of the clients having blue collar jobs have basic.9y qualification.
- If not university degree technicians are qualified through a professional course.

• EDA (Bivariate Analysis) # Deafult





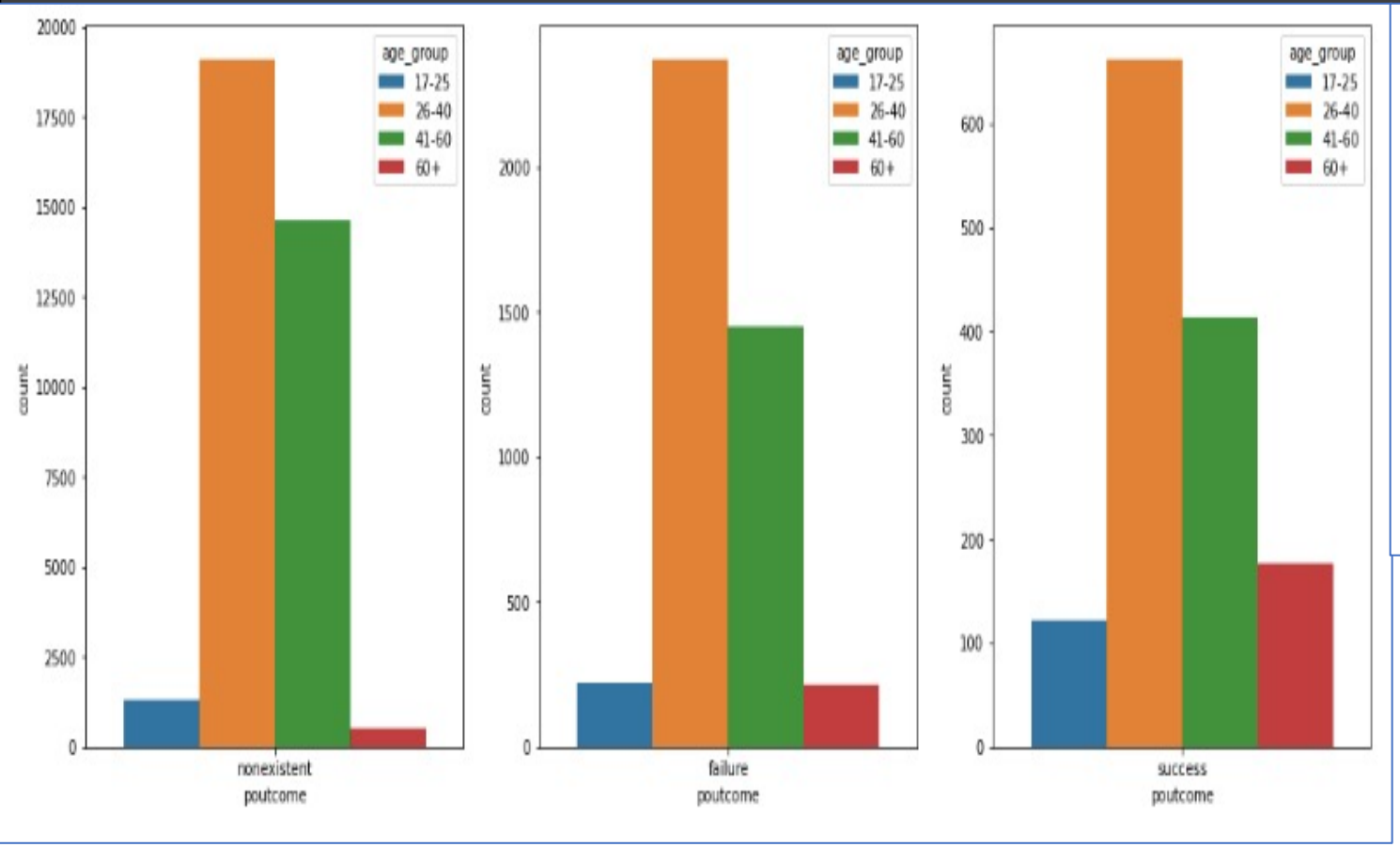


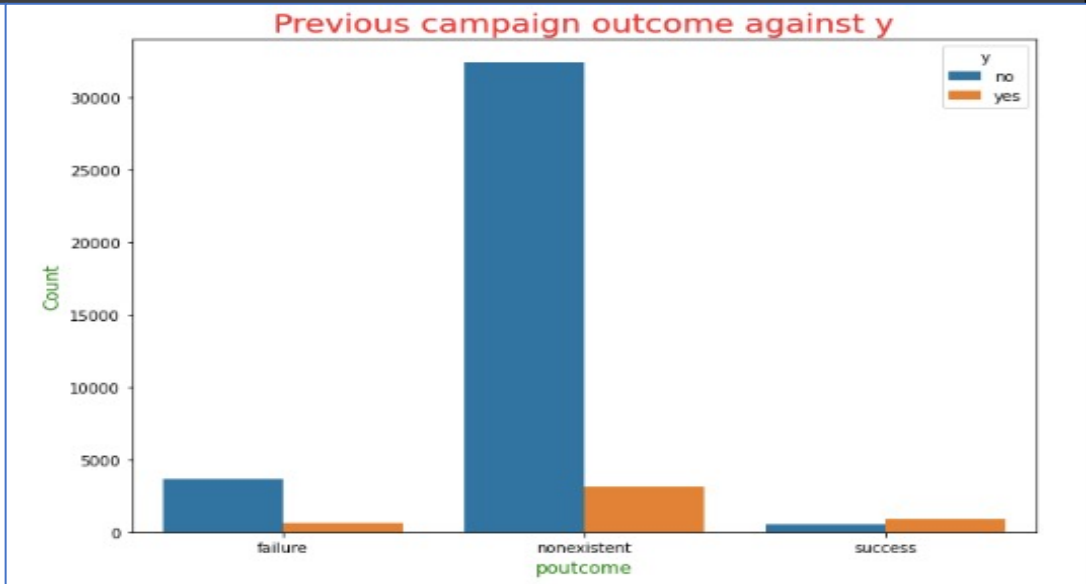
- Over 50% of customers have no credit in default. In those above 25000 customers have not accepted term deposit plan.
- Nearly 7500 customers have not declared default status.

- The 26-40 and 41-60 age group have an unknown default status.
- Similarly, these two groups are in majority among customers who have no defaults.

- Admin, blue-collar and technician professionals are greater than 20000 in number who have no defaulters.
- In addition, 50% of blue-collar professional have unknown default status.

• EDA (Bivariate Analysis) # Poutcome

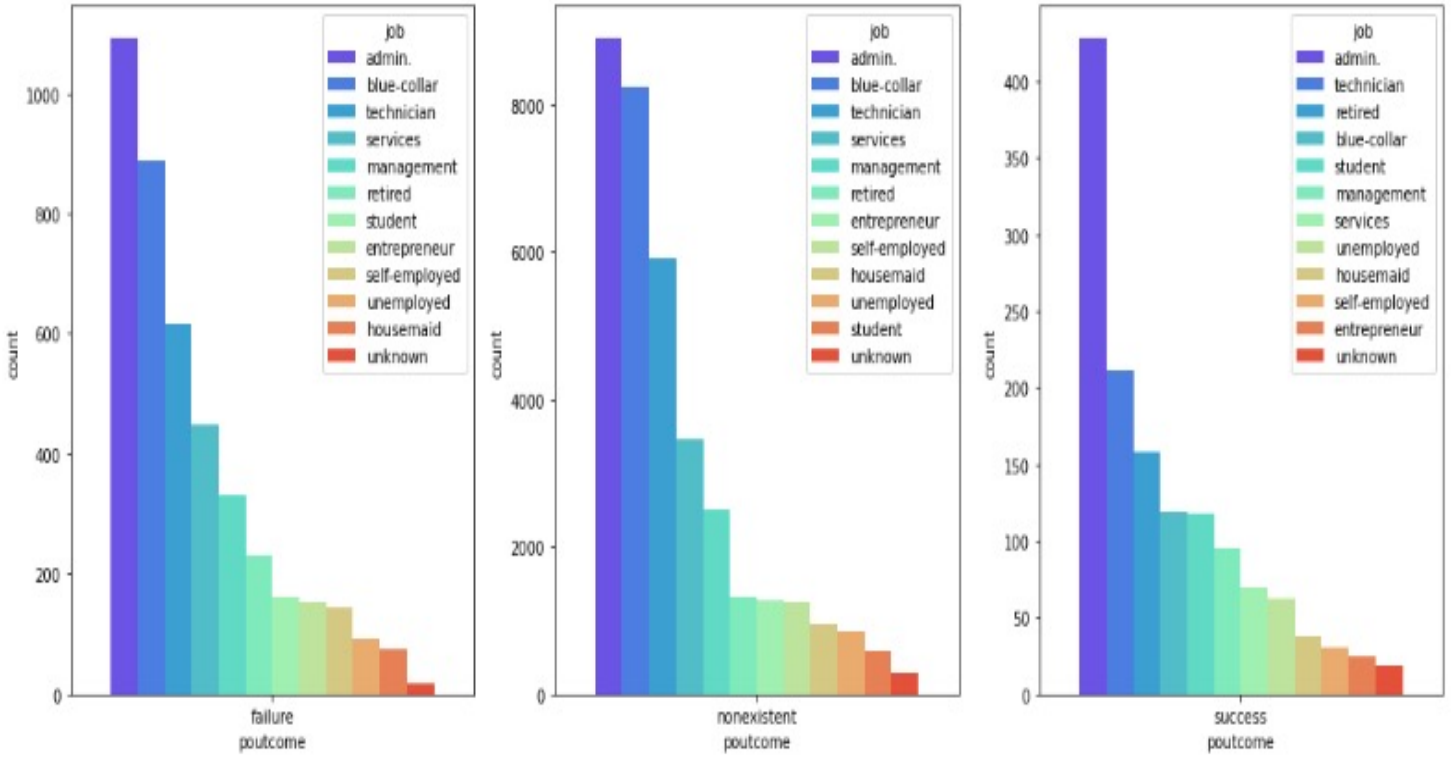


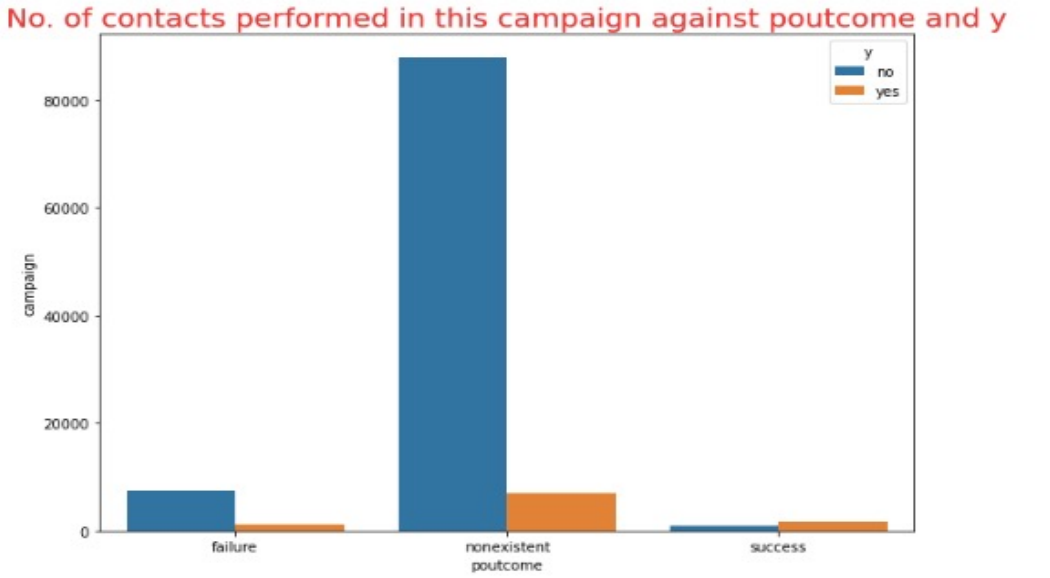


- Out of 35000 customers for whom the previous campaign was non-existent, over 30000 have rejected the term deposit plan.

- Over 90% customers for whom the outcome of previous campaign was non-existent are in 26-40 and 41-60 age groups.
- Both age-groups follows similar trend for rest of the outcome.
- On the other hand, the previous campaign was a success for nearly 1500 customers.

• EDA (Bivariate Analysis) # Poutcome

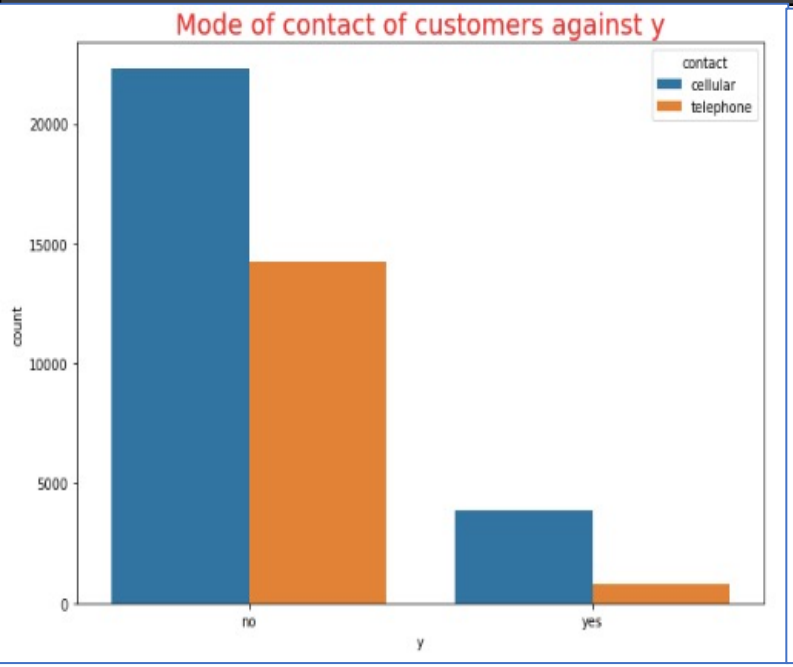


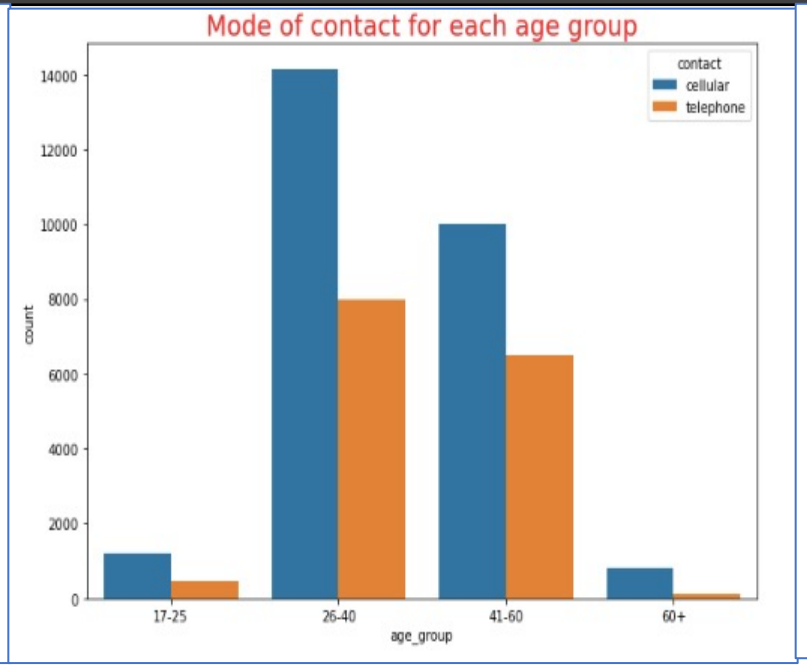


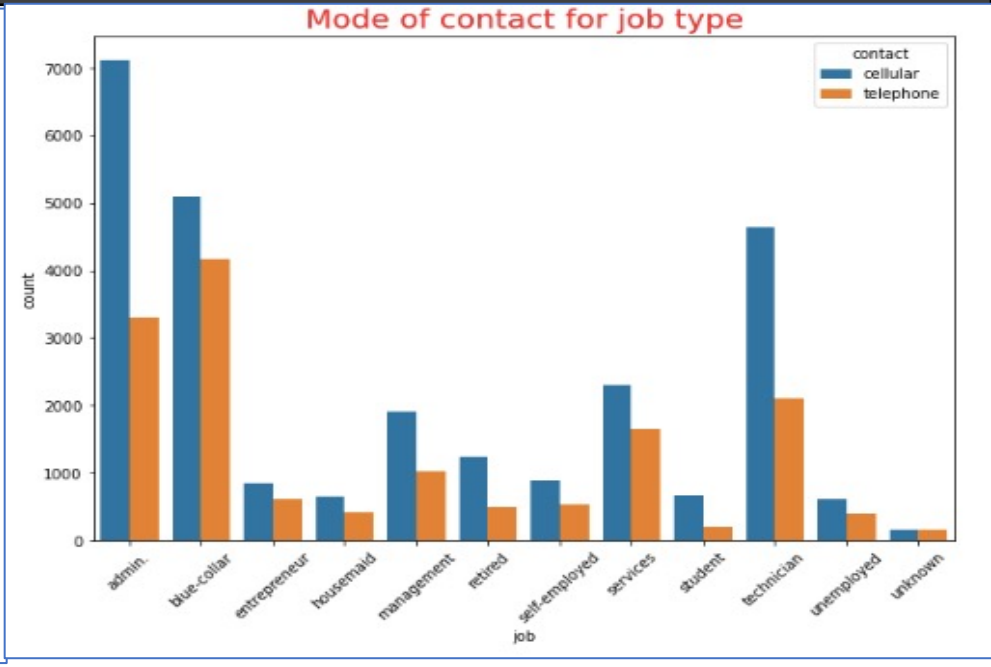
- Customers with profession admin, blue-collar and technician for whom the previous campaign outcome was non-existent are more than 20000 in number.
- However, for the outcome of success, the number of admin professionals are almost double the number of the next job category.

- For all non-existent outcomes, close to 85000 contacts have been made by the marketing team.
- This dwarfs the small number of contacts made to customers belonging to the other two outcomes of the previous campaign.

• EDA (Bivariate Analysis) # Contact







- Out of the 37000 customers not opting for the term deposit plan, 22000 of them had been contacted through cellular communication.
- For the 26-40 age-group, almost double the number of customers were contacted through cellular means than telephone. In addition, for 41-60 age-group, there is a gap of 4000 in the number of customers contacted through two modes of communication.
- For all job types, cellular mode seems to be the preferred mode of communication. For example, close to 7000 customers in admin roles are communicated through cellular means. This number is more than double the number contacted through telephone.

• EDA Recommendation.

What general recommendation can be offered to improve the campaign and make it more successful?

- Senior citizens and students responds better to their proposal and have high conversion rate.
- Looking at the customer database, the age-groups of 26-40 and 41-60 have a higher proportion. These groups present a profitable target for the marketing team.
- Month of may is most successful as we can say most term deposits are subscribed in this month.
- It is imperative to form target groups based on socioeconomic categories. Age, income level (not always high), profession and education can accurately determine the marketing profile of a potential client.
- Increasing the time with the customer and using different means of communication tends to increase the chances of customer subscribing to the term deposit plan.
- In the target groups, focus on Admin., Blue-collar, Technician, Services and Management professions

Given these factors, it is recommended to concentrate on those consumer groups that are potentially more promising.

• Model Recommendation.

What type of ML model to use?

In order to predict the client subscription for a deposit term, we will use a predictive ML model to helps us identify potential customers. As a start and after the performed EDA on the provided dataset, we choose to test out the following set of models since we don't know yet what algorithms will do well on this dataset.

The Following algorithm selected include:

- **Linear Algorithms :**
  Logistic Regression (LR)
  Linear Discriminant Analysis (LDA)
- **Ensemble Methods :**
  Boosting methods: AdaBoost (AB) and Gradient Boosting (GBM)
  Bagging methods: Random Forests (RF) and Extra Trees (ET).
- **Non Linear Algorithms :**
  Classifications and Regression Trees (CART).
  Support Vector Machines (SVM)
  Gaussian Naive Bayes (NB)
  K-nearest Neighbours (KNN)

Thank You