Exploratory Data Analysis Presentation

Bank Marketing Campaign

Efe KARASIL-Sefa SÖZER

13.08.2022

Agenda

Executive Summary

Data Understanding

EDA

EDA- Categorical Columns

EDA Recommendations

Model Recommendation

Data Glacier

Your Deep Learning Partner

Executive Summary

- **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 helps 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).

- **Problem Statement:**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).

- **Analysis:**

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

- Data Understanding
- Univariate analysis
- Bivariate analysis
- Model recommendations

Data Understanding

Data Set Information:

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

There are four datasets:

1) bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]

2) bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.

3) bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).

4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Data Understanding

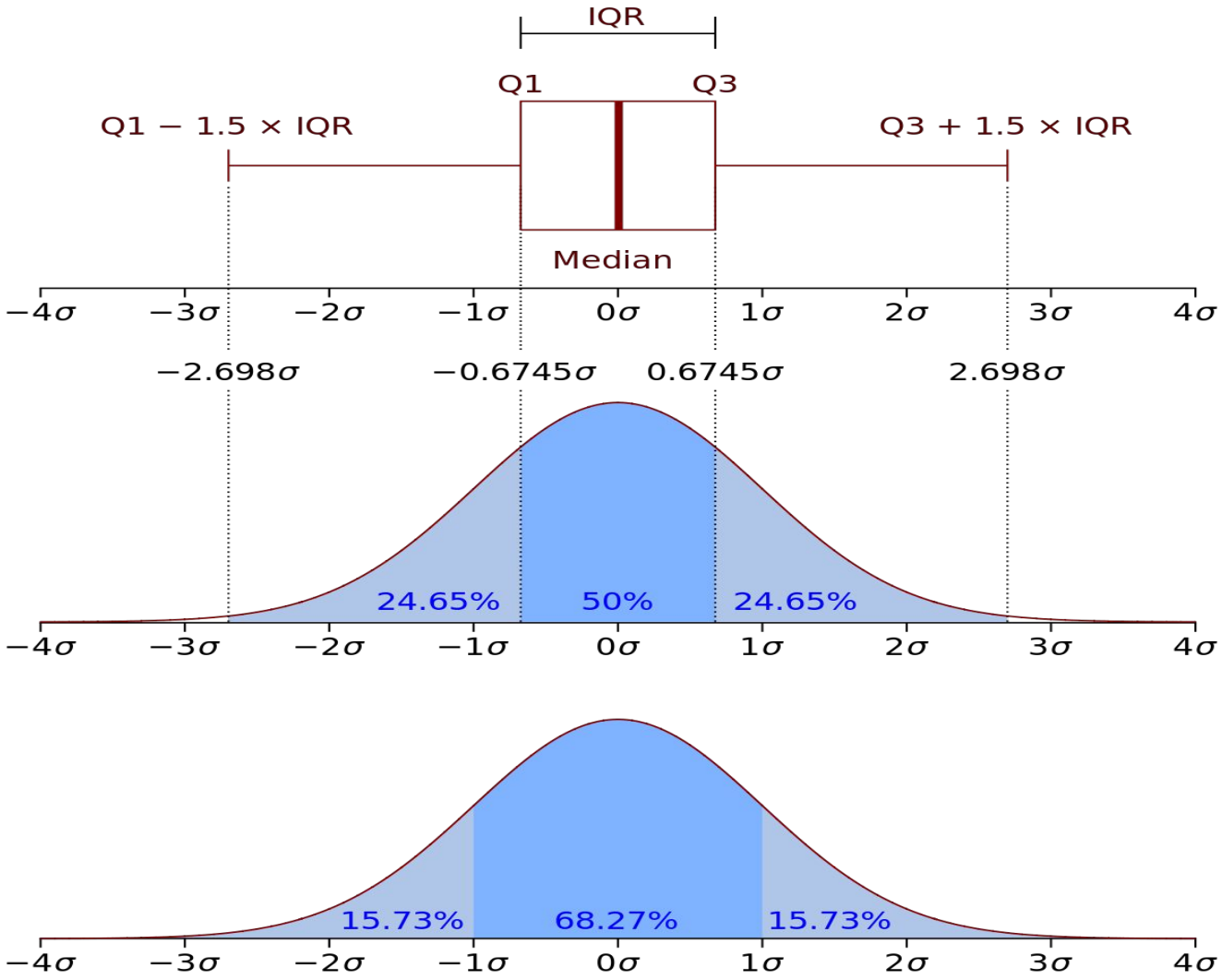
- **Attribute Information:**
- Input variables:
- bank client data:
- 1 - age (numeric)
- 2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
- 3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
- 4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
- 5 - default: has credit in default? (categorical: 'no','yes','unknown')
- 6 - housing: has housing loan? (categorical: 'no','yes','unknown')
- 7 - loan: has personal loan? (categorical: 'no','yes','unknown')
- # related with the last contact of the current campaign:
- 8 - contact: contact communication type (categorical: 'cellular','telephone')
- 9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- 10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- 11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Data Understanding

other attributes:

- 12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- 13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- 14 - previous: number of contacts performed before this campaign and for this client (numeric)
- 15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
- # social and economic context attributes
- 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
- 17 - cons.price.idx: consumer price index - monthly indicator (numeric)
- 18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
- 19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
- 20 - nr.employed: number of employees - quarterly indicator (numeric)
-
- Output variable (desired target):
- 21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

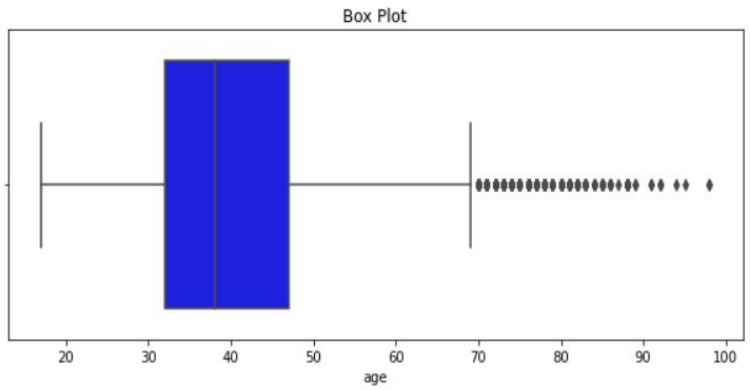
EDA- Outlier Detection and Handling

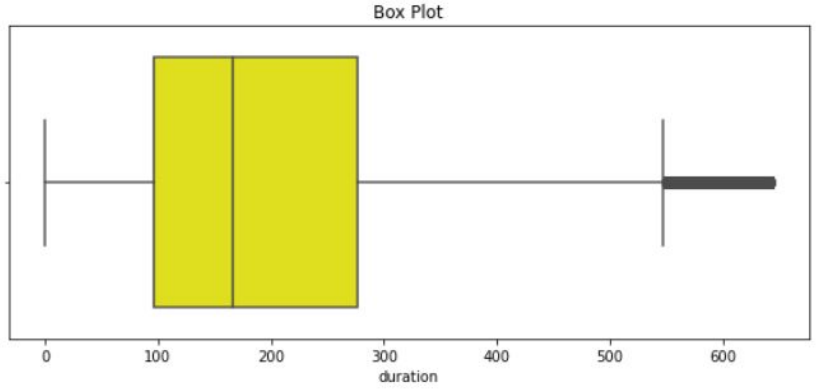


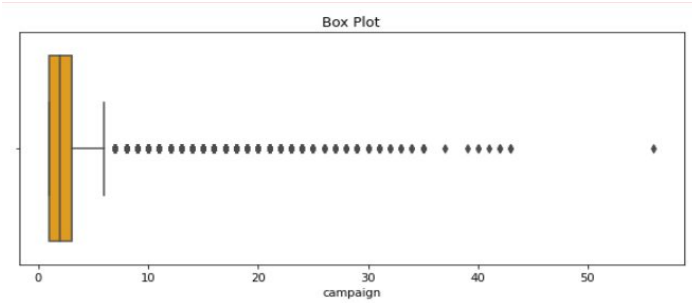
IQR(The interquartile range) method was used for outlier detection and handling.

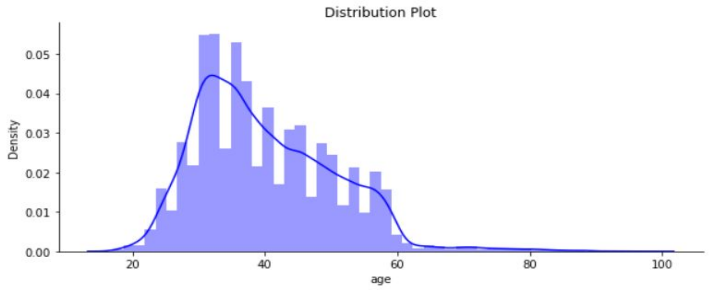
The interquartile range method defines outliers as values larger than Q3 + 1.5 * IQR or the values smaller than Q1 – 1.5 * IQR.

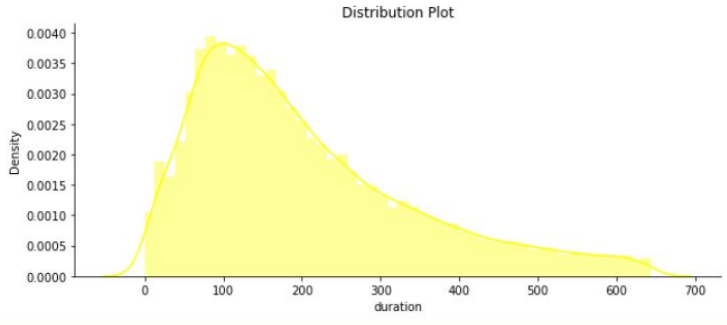
EDA- eski

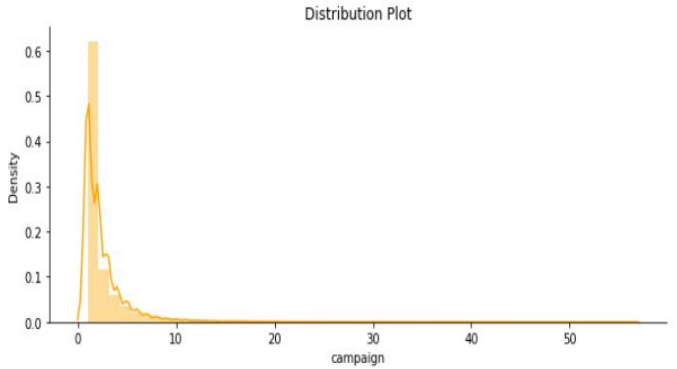








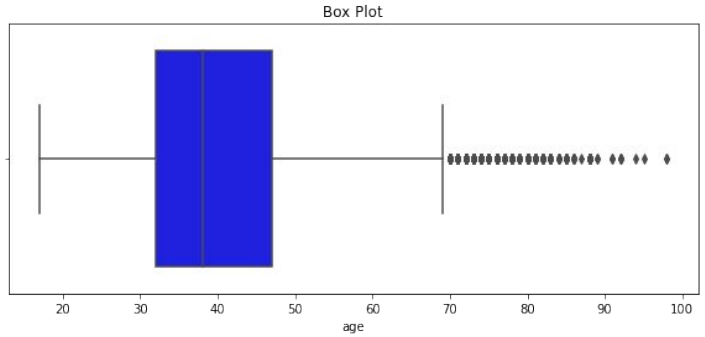


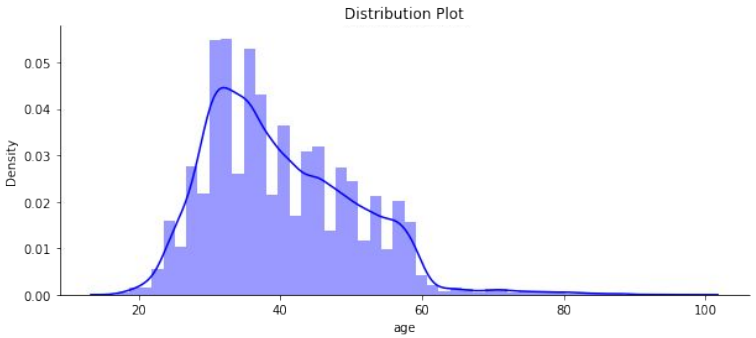


2 different methods for outliers was followed. The first method was directly removing the outliers, while the second method was to imputation of average values for outliers. In the future machine learning models, best performing methods planned to be used.

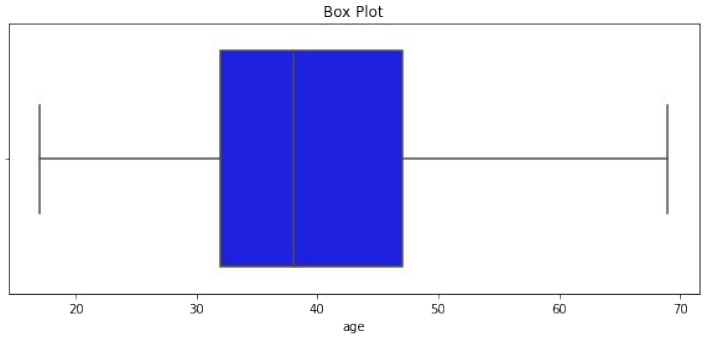
We observe positive skewness in the graphs above.

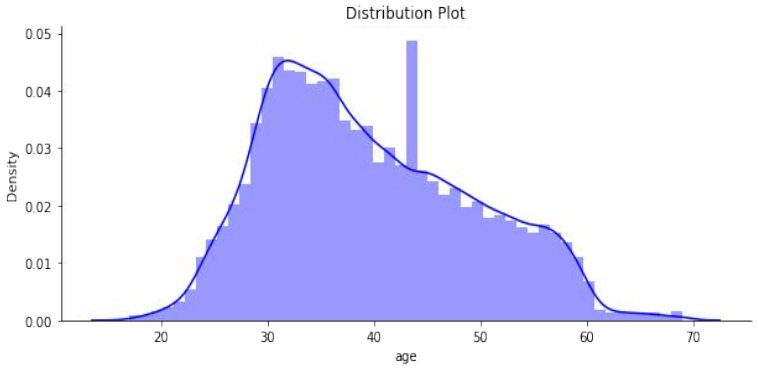
EDA- eski



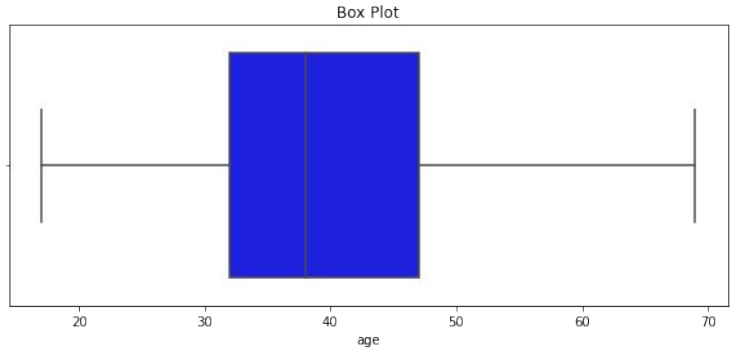


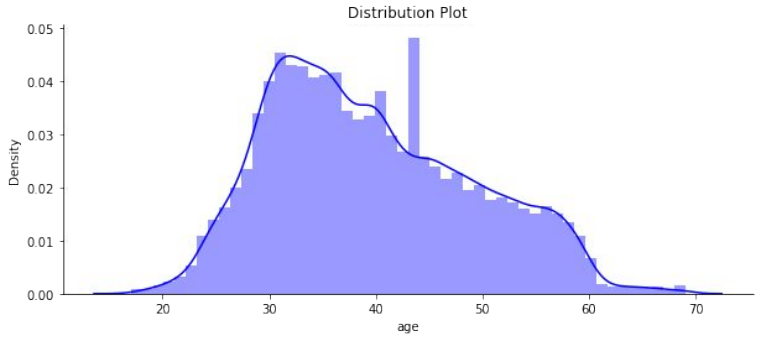
Original





Deleting outliers

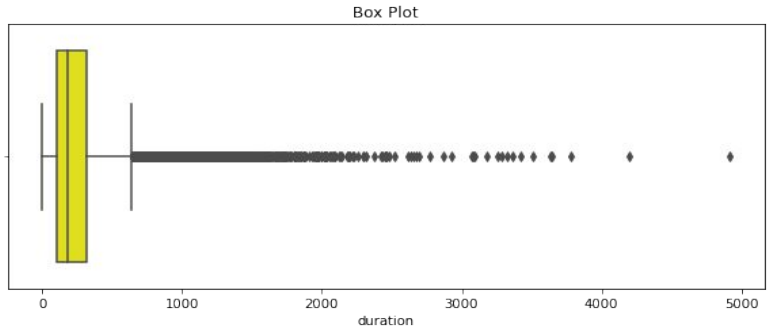


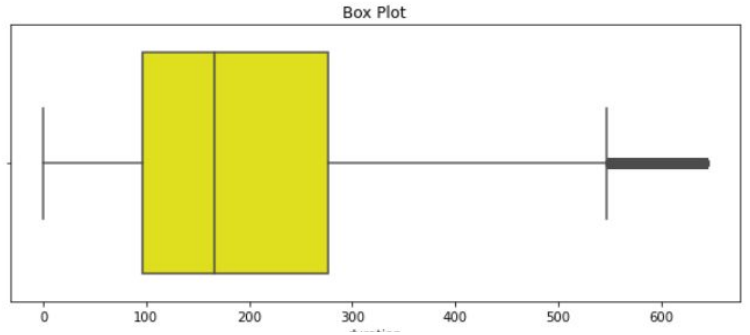


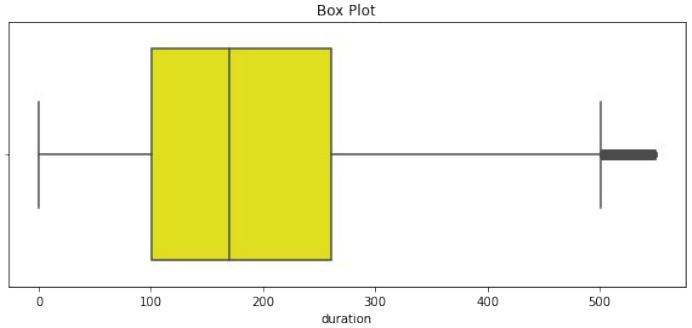
Imputing mean values instead of outliers

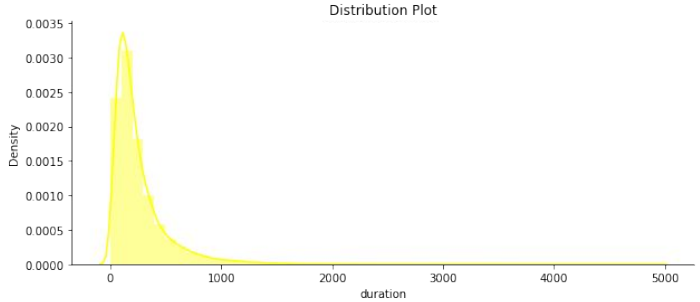
We observe positive skewness in the graphs.

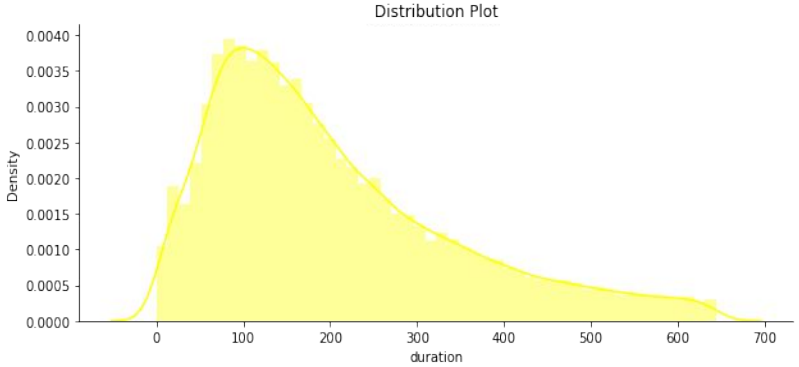
EDA- eski

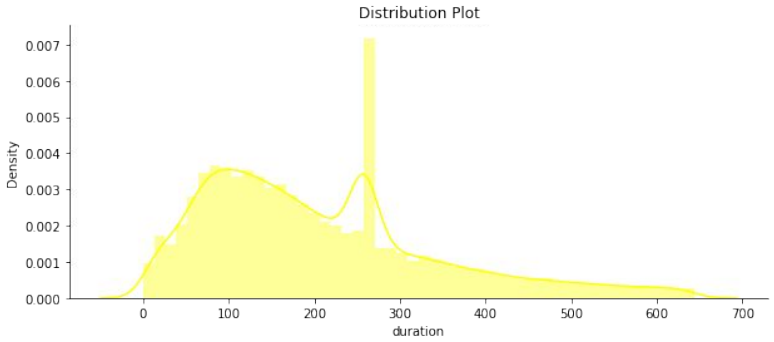












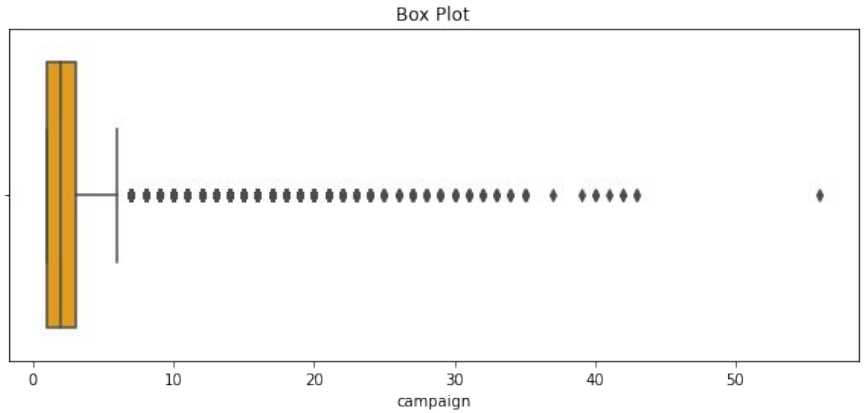
Original

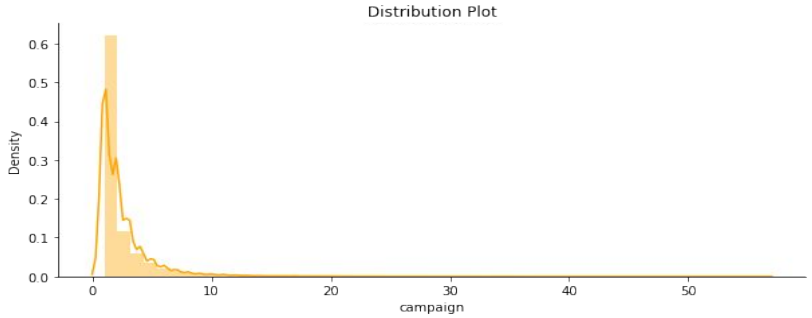
Deleting outliers

Imputing mean values instead of outliers

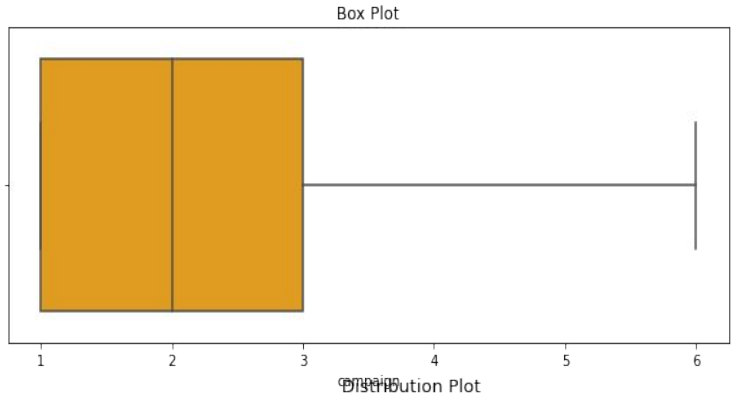
We observe positive skewness in the graphs.

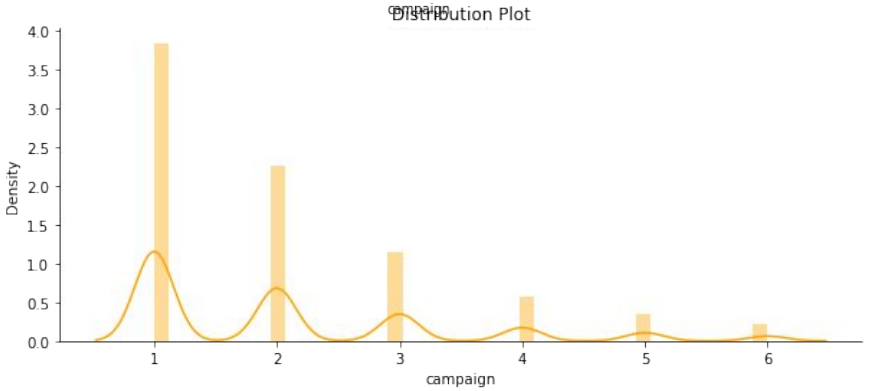
EDA- eski



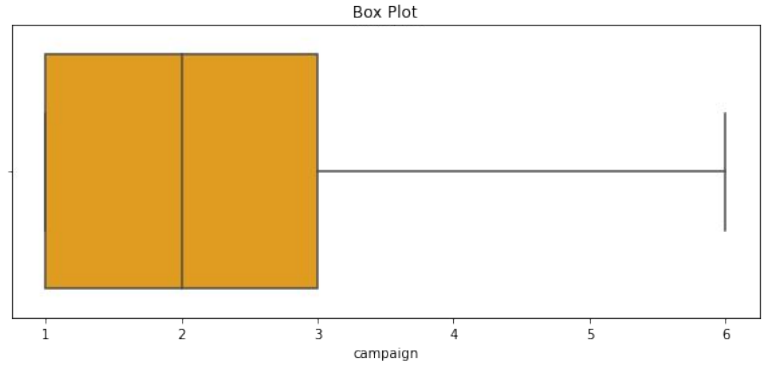


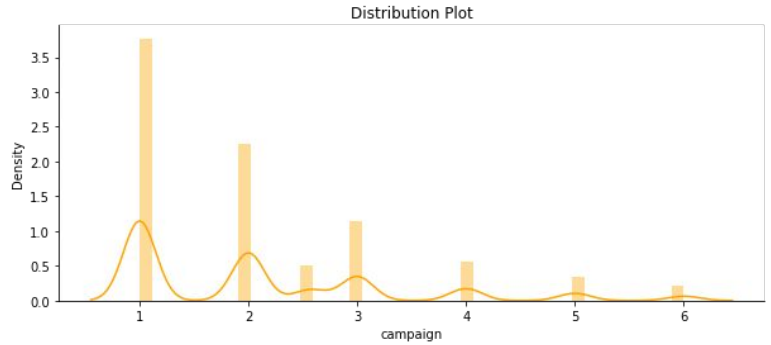
Original





Deleting outliers

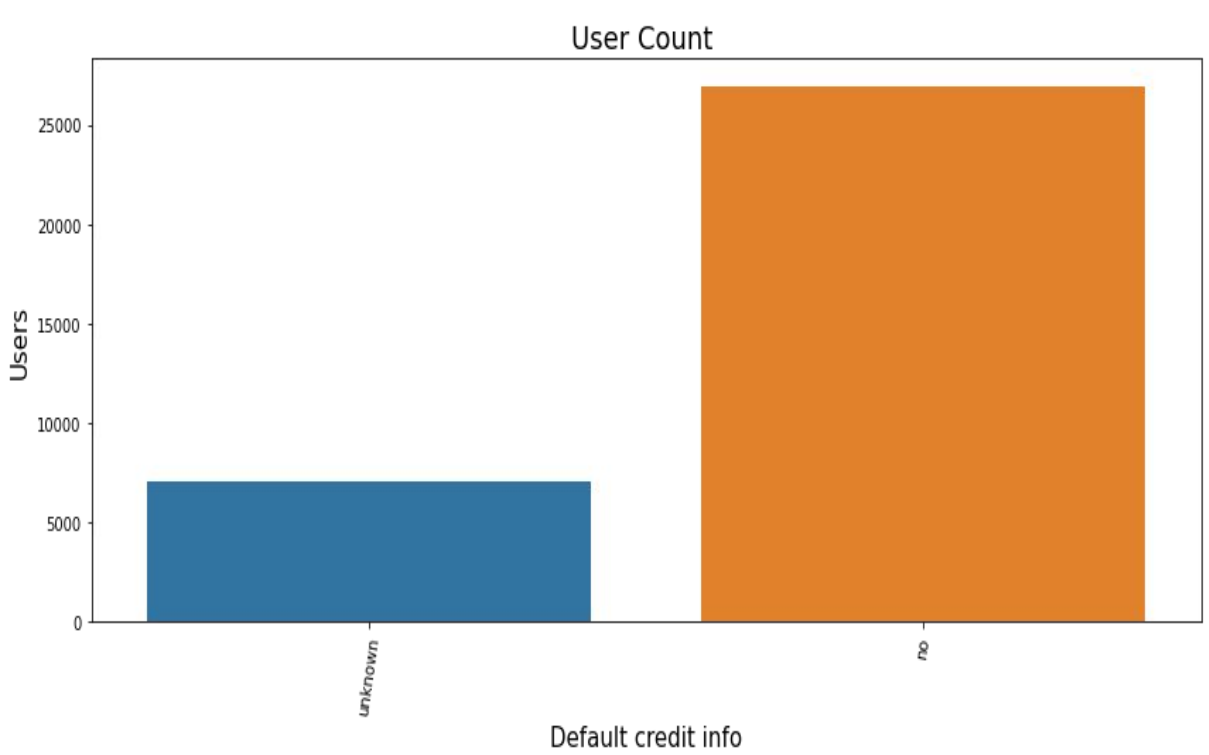




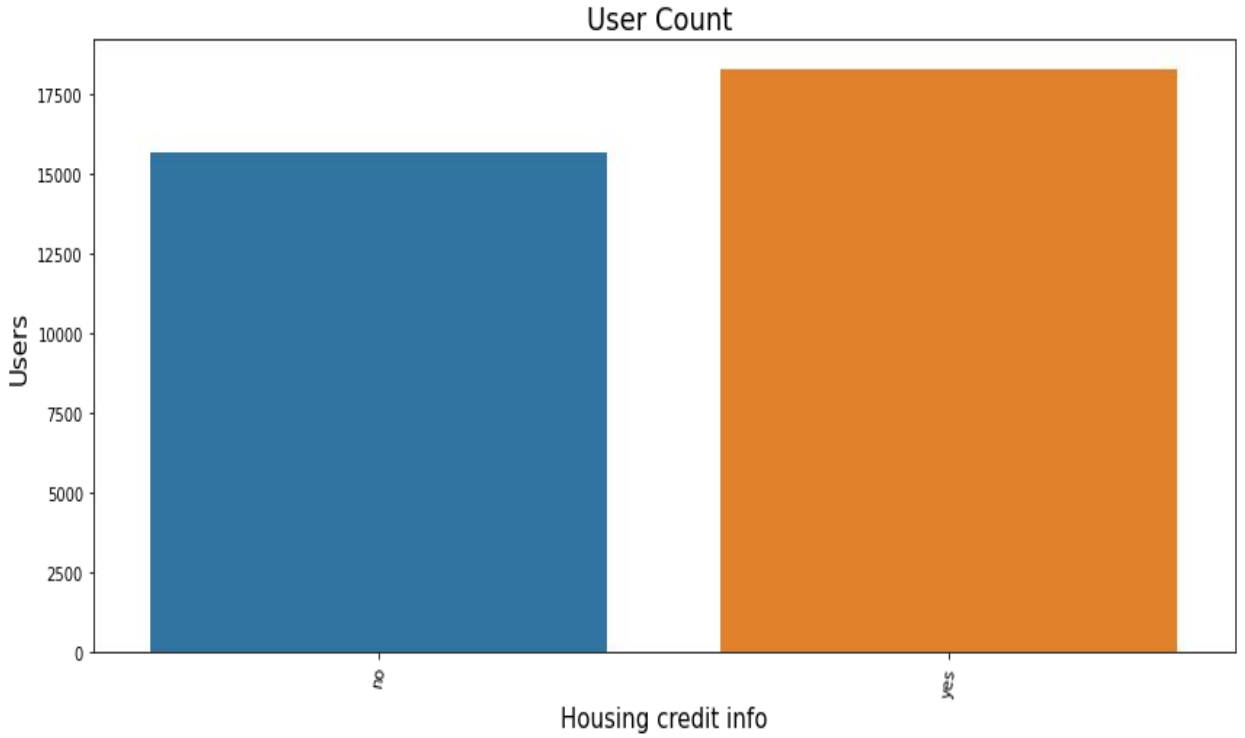
Imputing mean values instead of outliers

We observe positive skewness in the graphs.

EDA- Categorical Columns

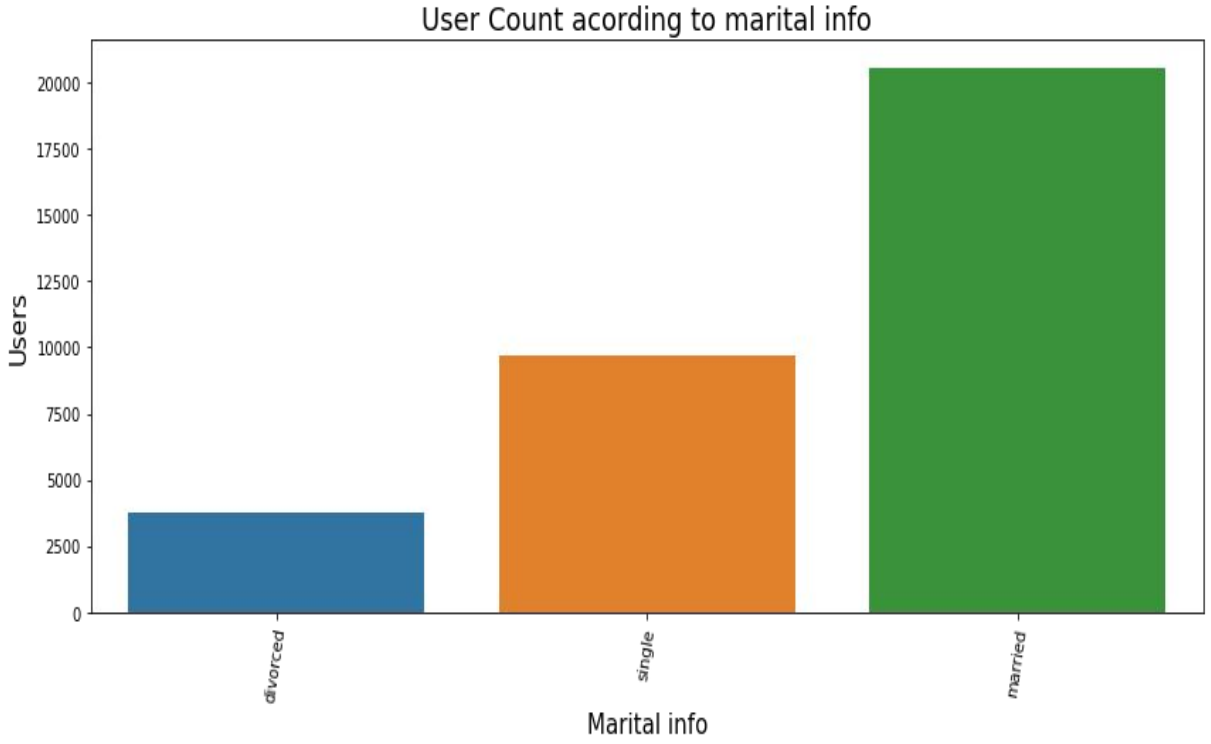


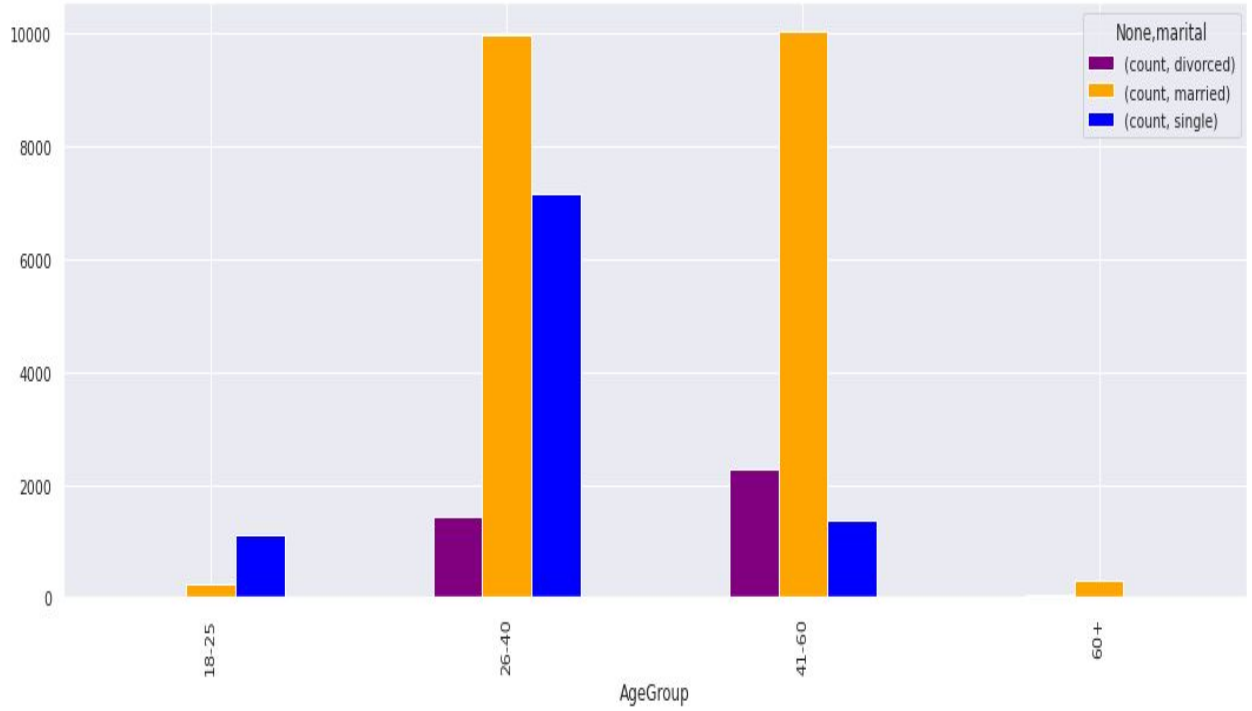
Most of the customers do not have credit in default or this information is unknown.



It can be stated that, almost half of the customers have housing credit while other half do not.

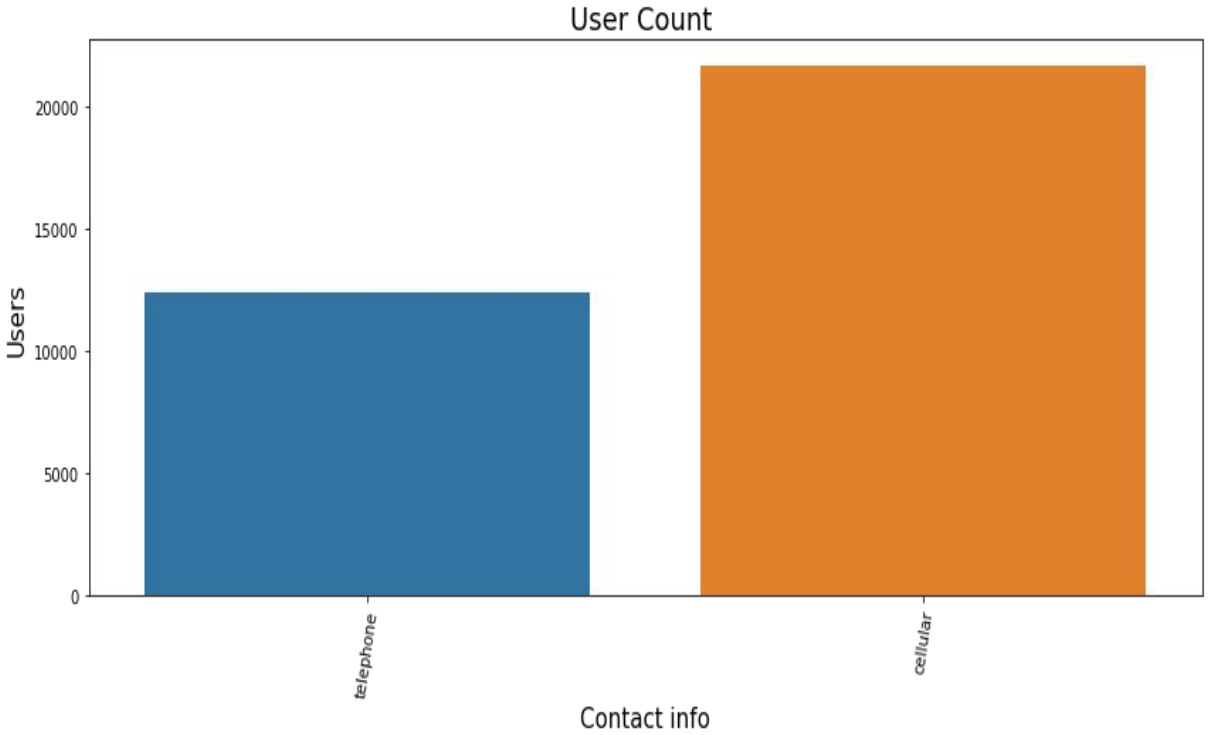
EDA- Categorical Columns

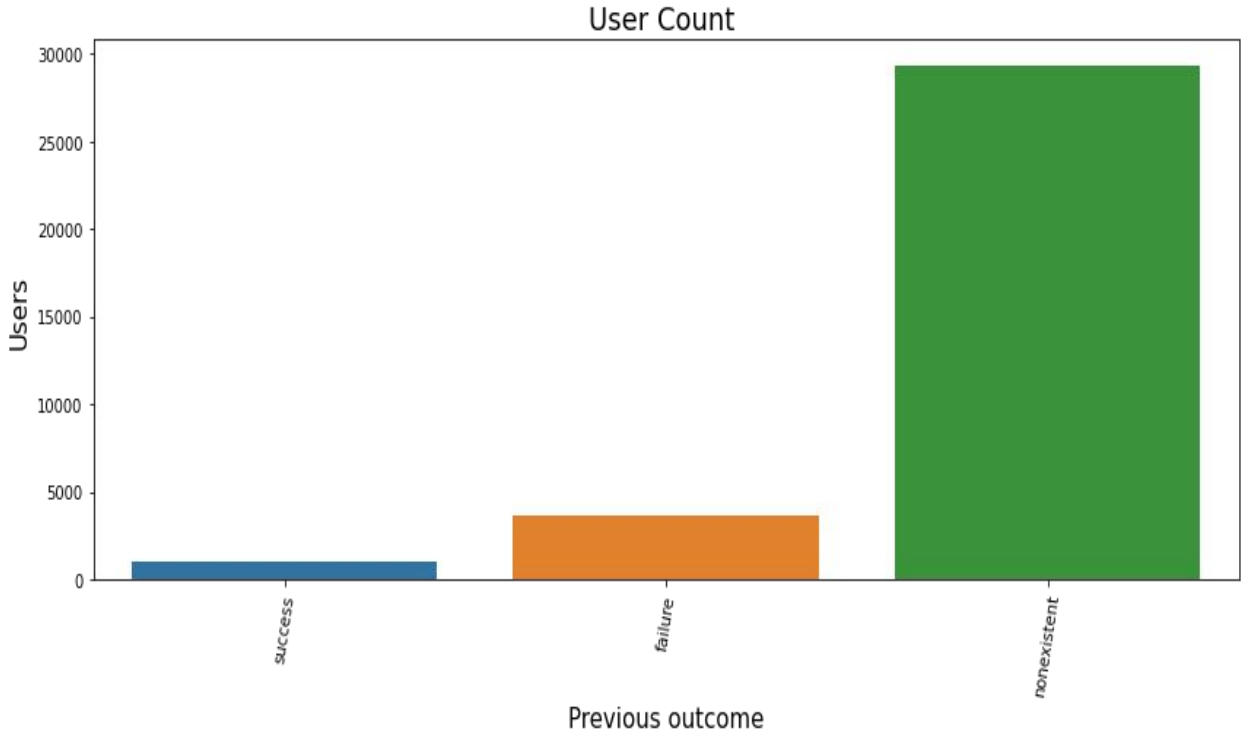




Marital information for every customers in data is given on the left. On the right marital situation of customers with their age groups are given

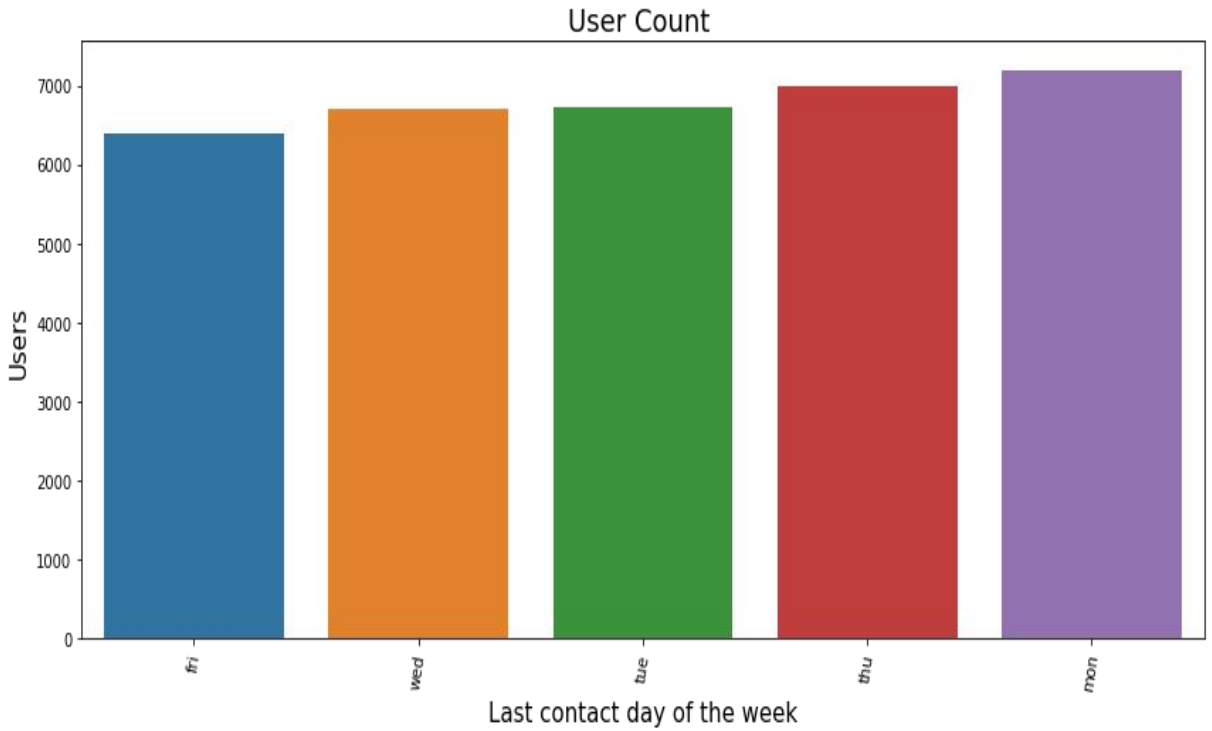
EDA- Categorical Columns

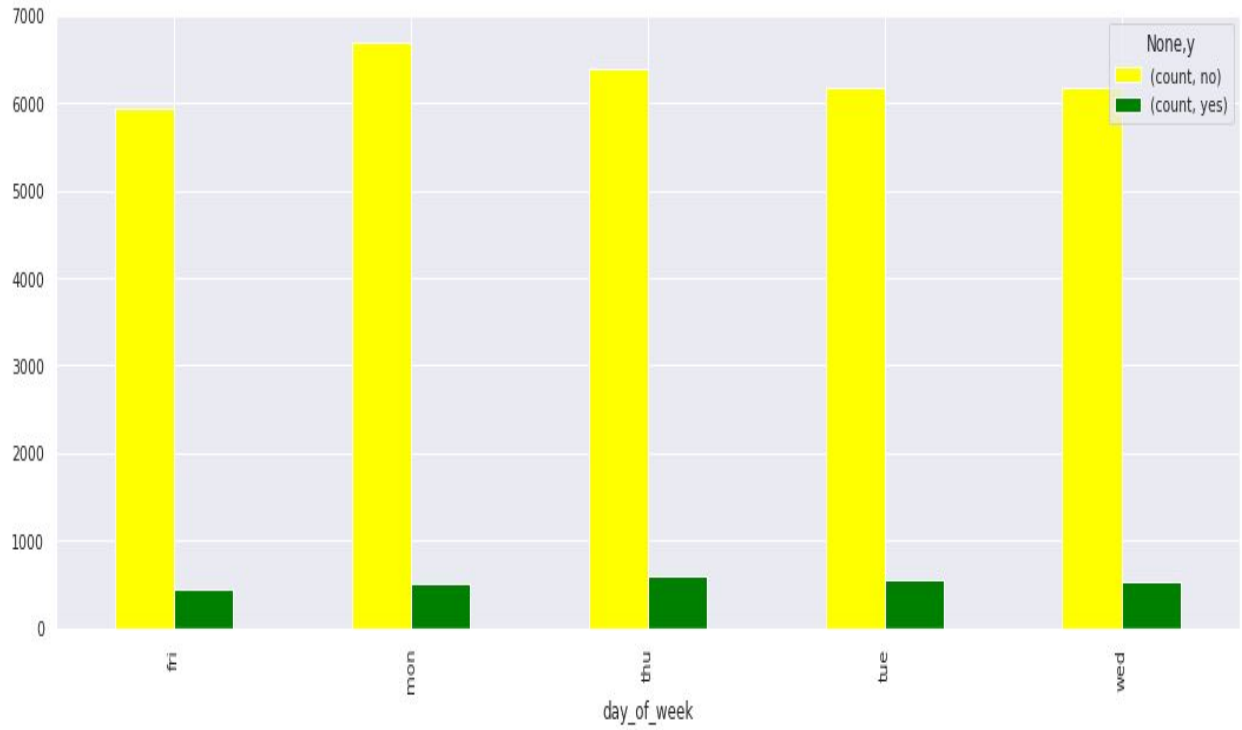




It can be said that most of the customers' previous outcome of this campaign for the bank is nonexistent.

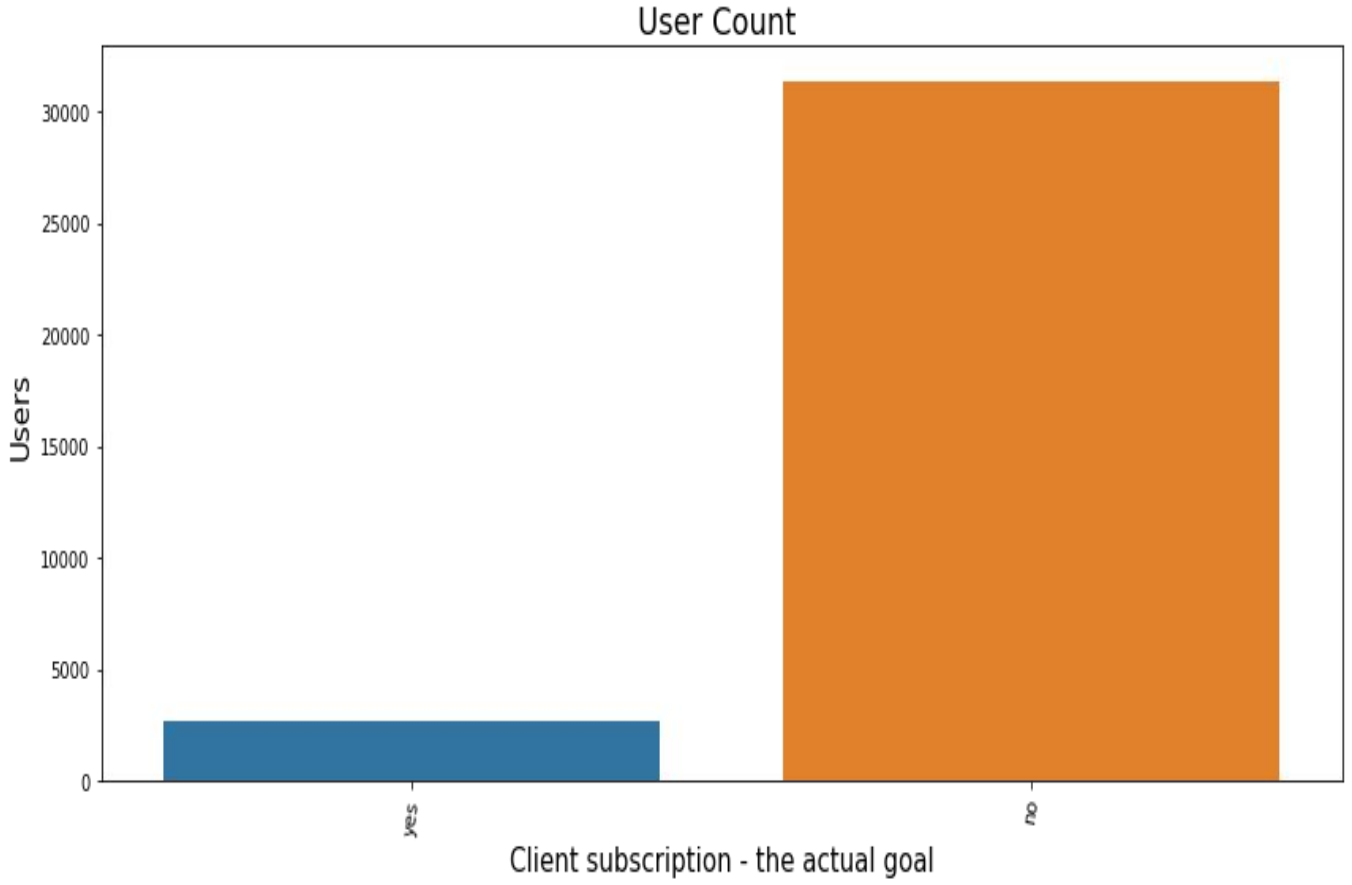
EDA- Categorical Columns





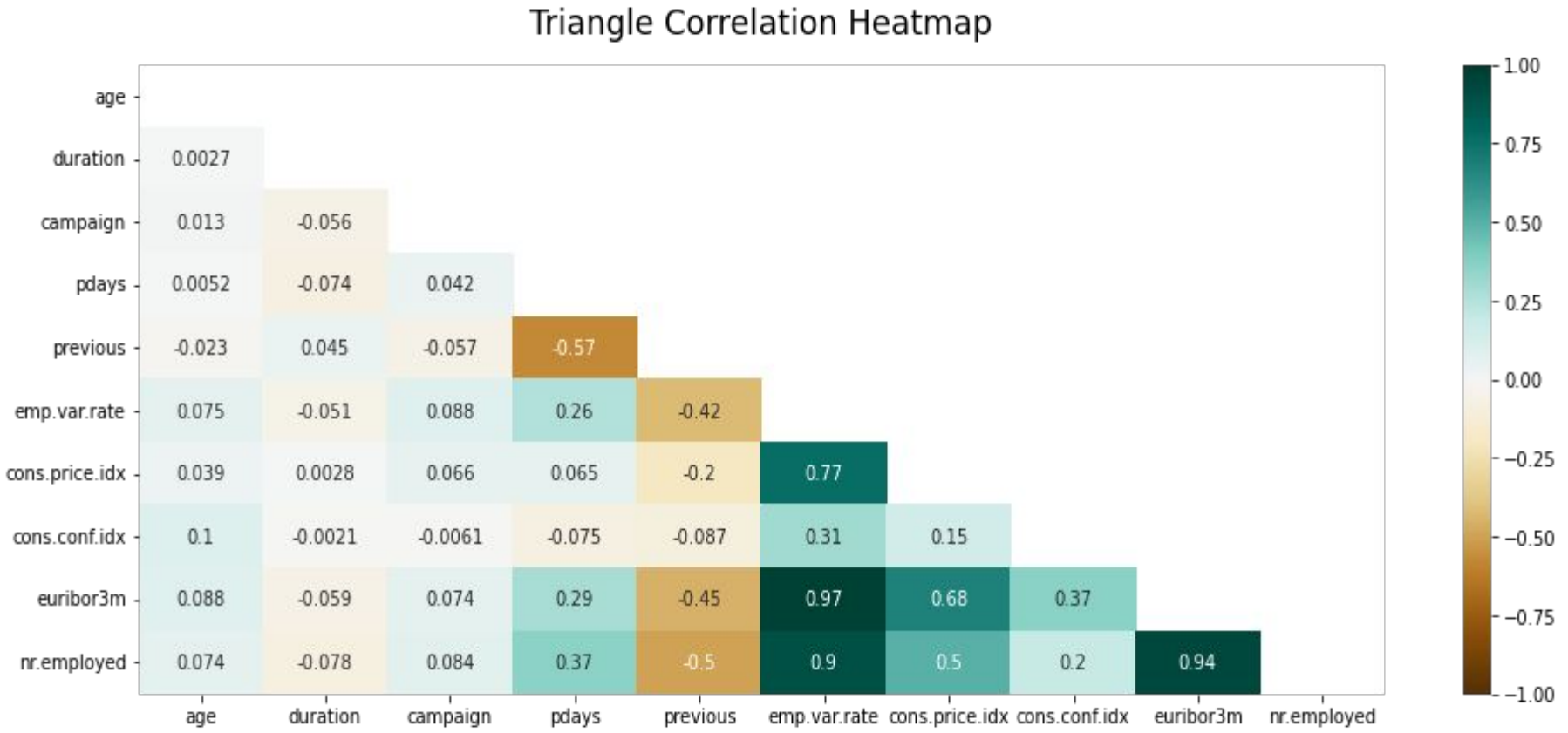
As it can be seen, last contact days of the week are almost evenly distributed for both total counts and “yes”/”no” percentages/counts. Because of this data, this feature was thought to not be in meaningful business insights/recommendations

EDA- Categorical Columns



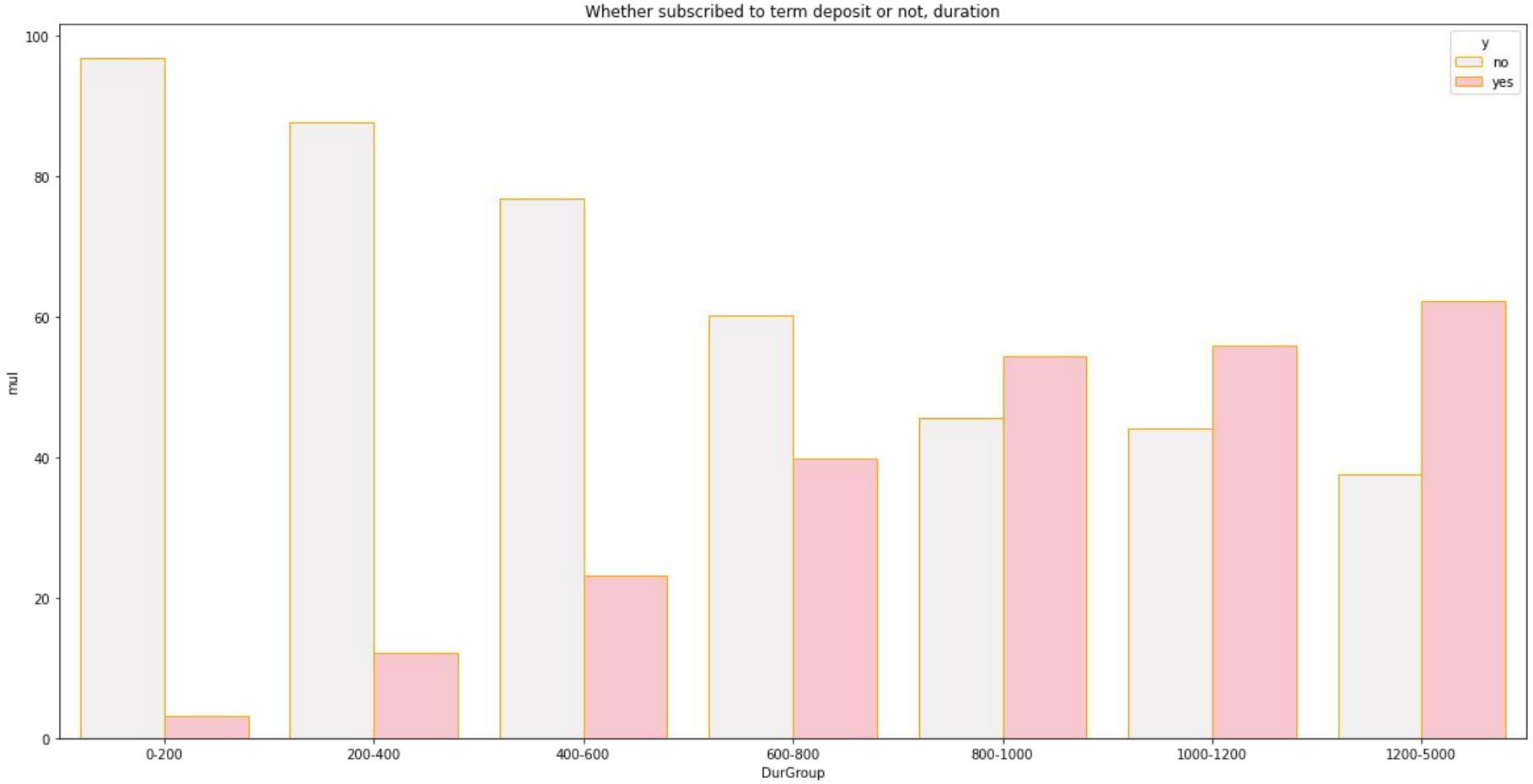
This column is the one that wanted to be predicted with the knowledge of other features, columns information.

EDA Recommendations #1



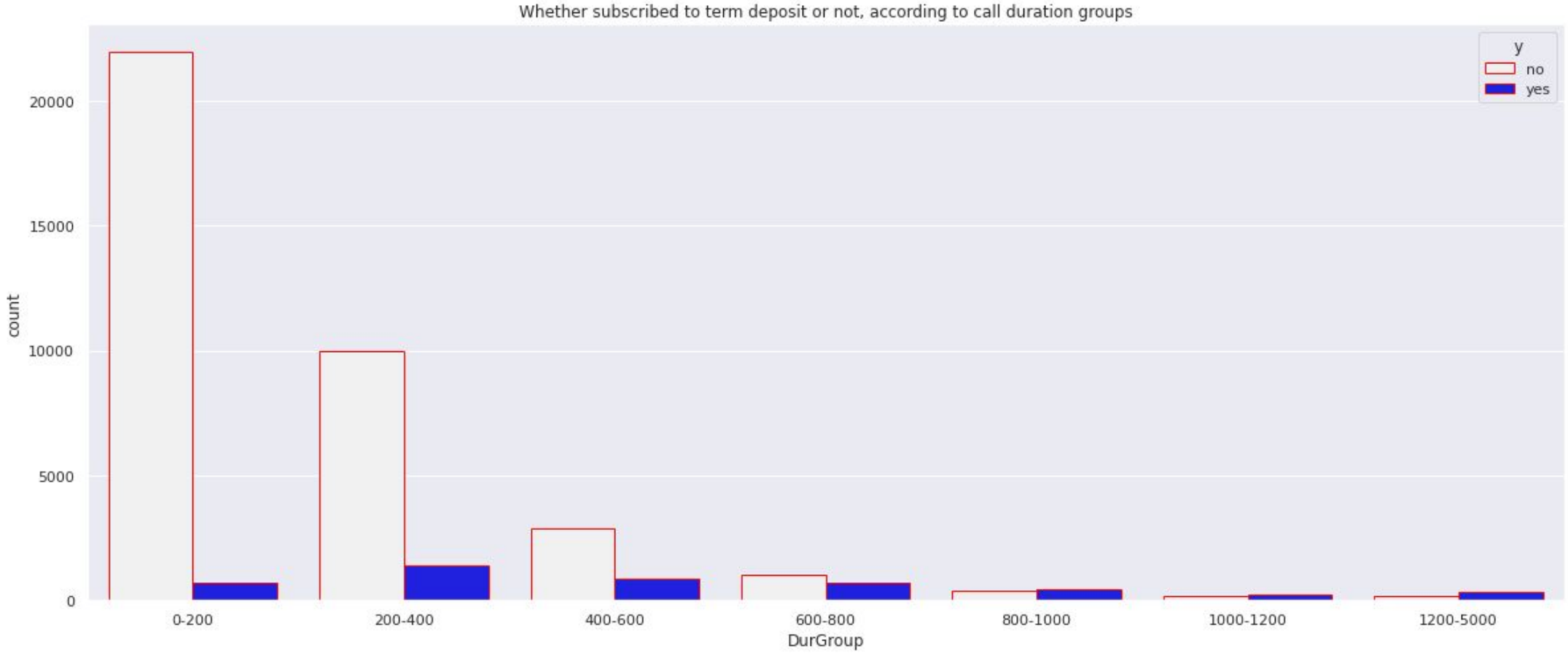
- **In correlation heat map, “euribo3rn”, “cons.priceidx”, “nr.employed” and “emp.var.rate” features have very high correlation between them. 2 or 3 of these features can be dropped from the data, since their existence will not be extra useful-providing good, new information- for machine learning model which will be deployed in next steps.**

EDA Recommendations #2



X-label represents the duration of calls(in seconds) with customer, Y-label is percentage of that groups response to this campaign, It can bee seen that when call duration is increased getting “Yes” response probability is clearly increased as well

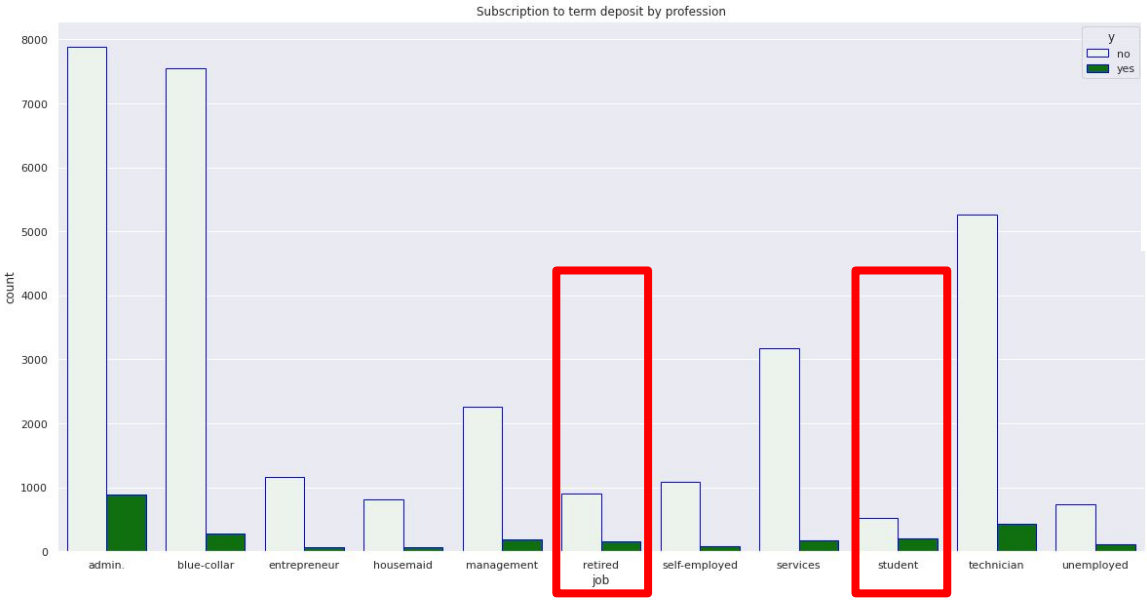
EDA Recommendations #2

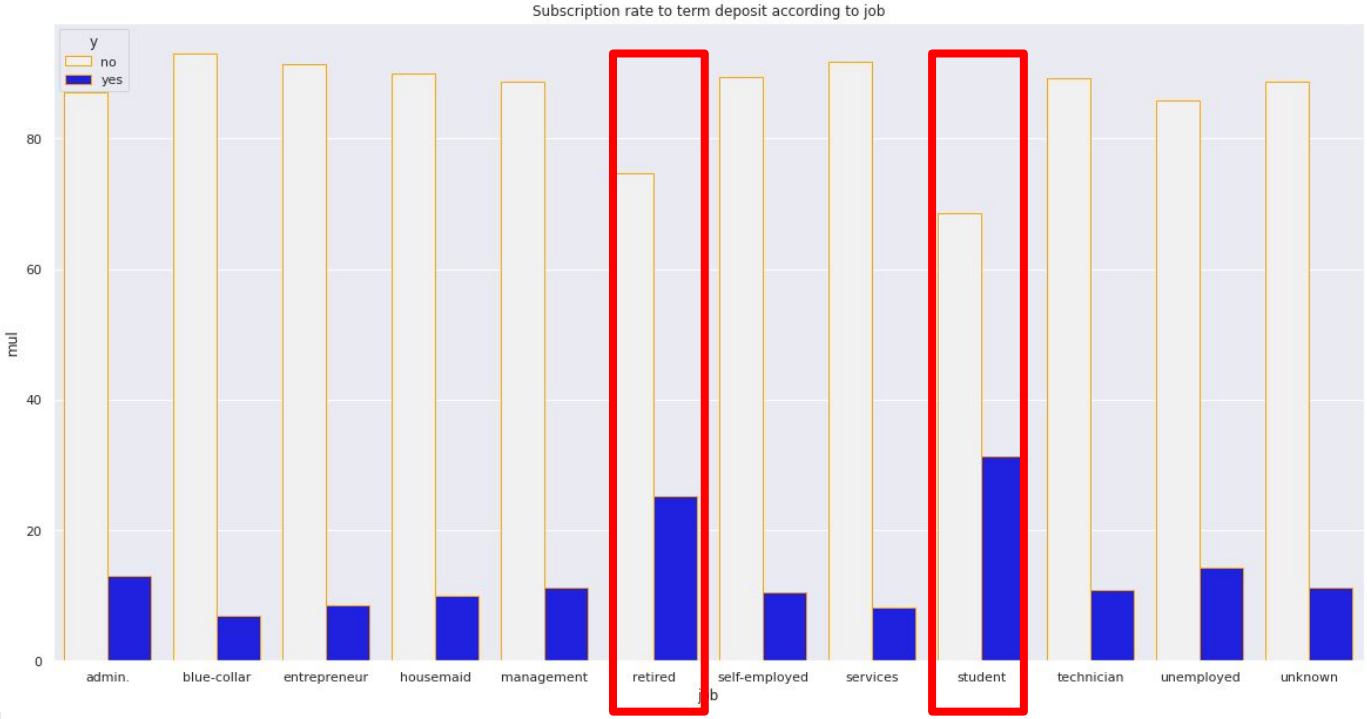


As it can be seen, most of the calls last shorter than 10 minutes(600 seconds). With the previous analysis, it can be mentioned that call duration can be increased to convince customers

EDA Recommendations #3

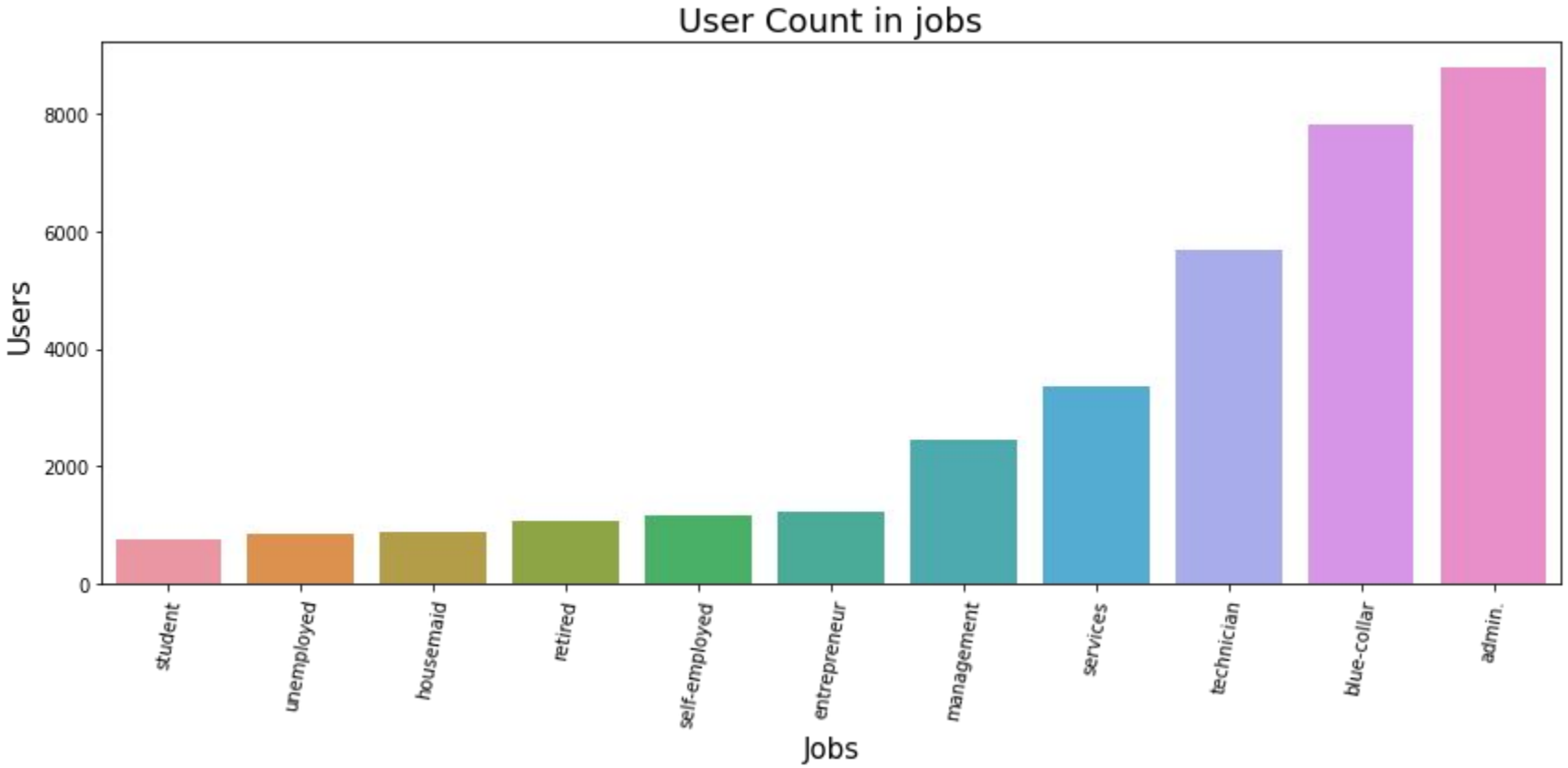
-Total counts of customers according to their job, with their responses (Yes,No) --- Below





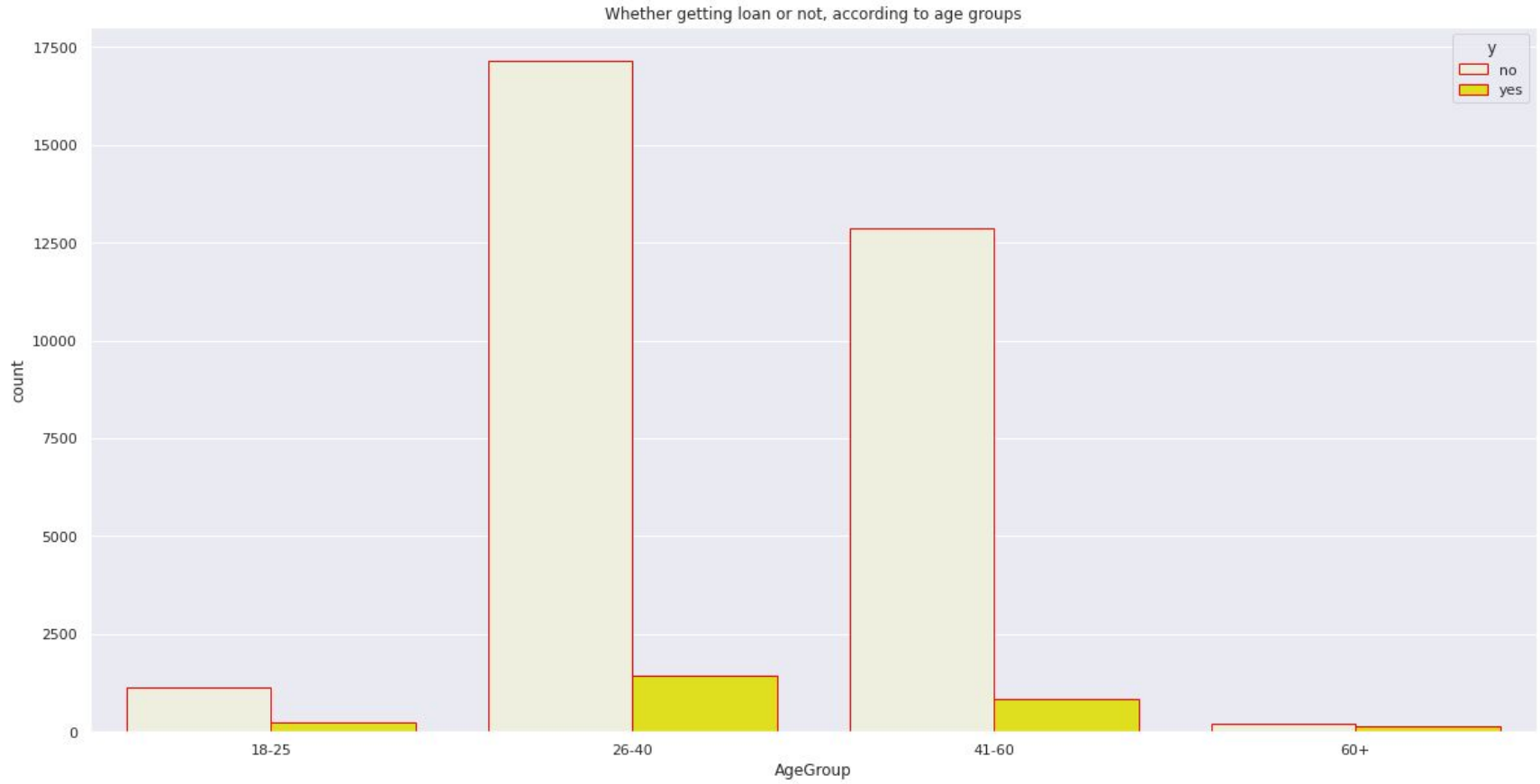
-Percentages of responses according to customer jobs, --- Above

EDA Recommendations #3



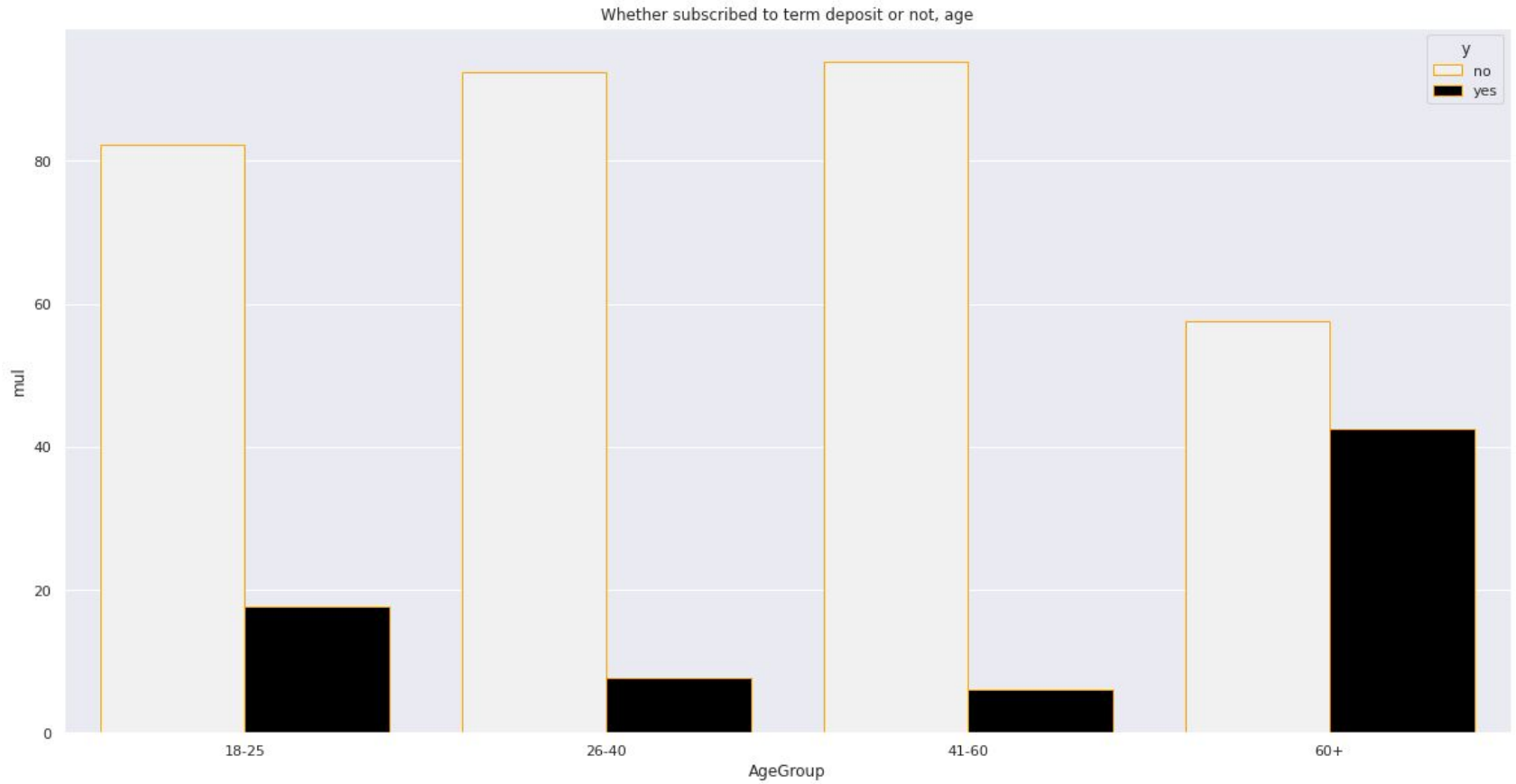
Students and retired people can be reached more as well, again their rate of acceptance are higher according to different jobs' subscriptions while they are one of the least contacted jobs

EDA Recommendations #4



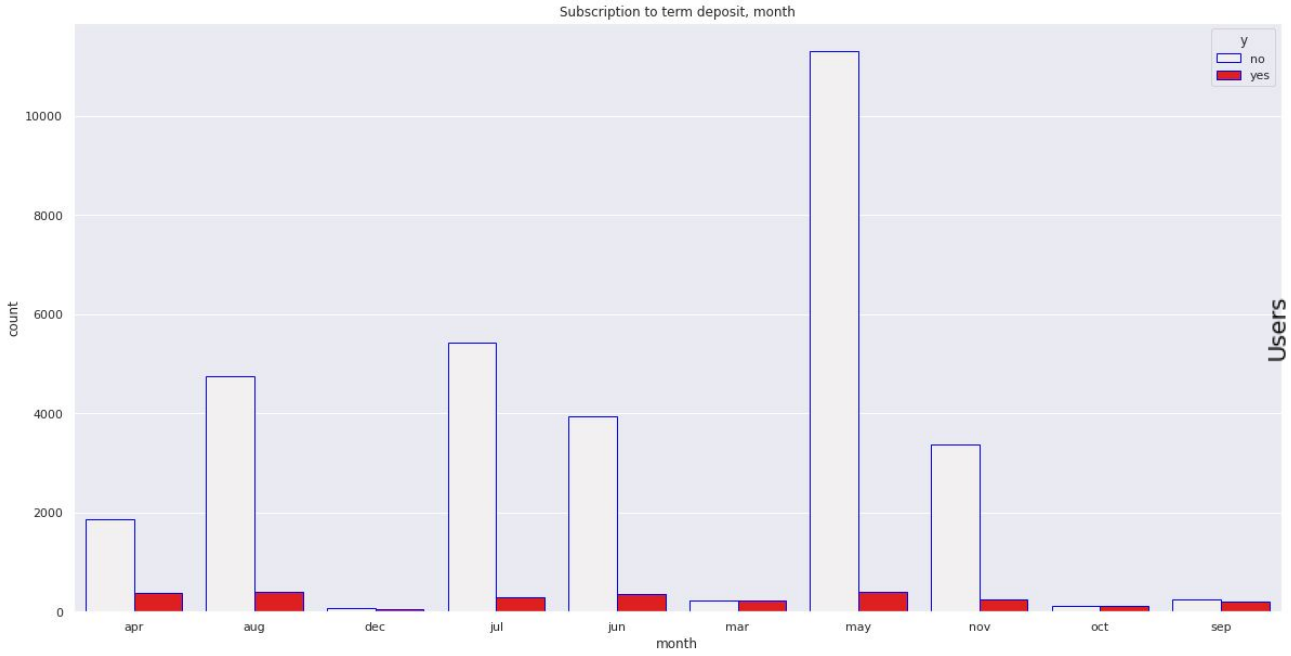
As it can be seen 26-40 and 41-60 age groups make up the vast majority of people taking part in the campaign

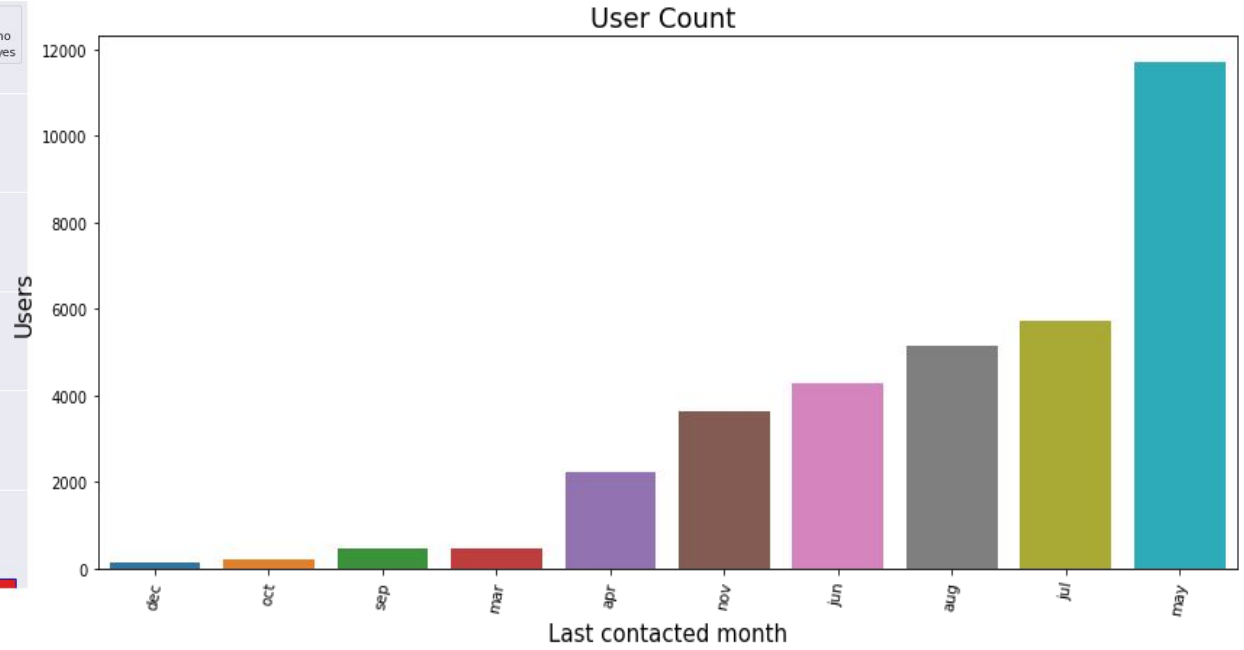
EDA Recommendations #4



Although, 26-40 and 41-60 age groups make up the vast majority of people taking part in the campaign, percentage of getting "Yes" response is higher in 18-25 and even much more higher in 60+ age groups. So, contacting these age groups more can make huge difference.

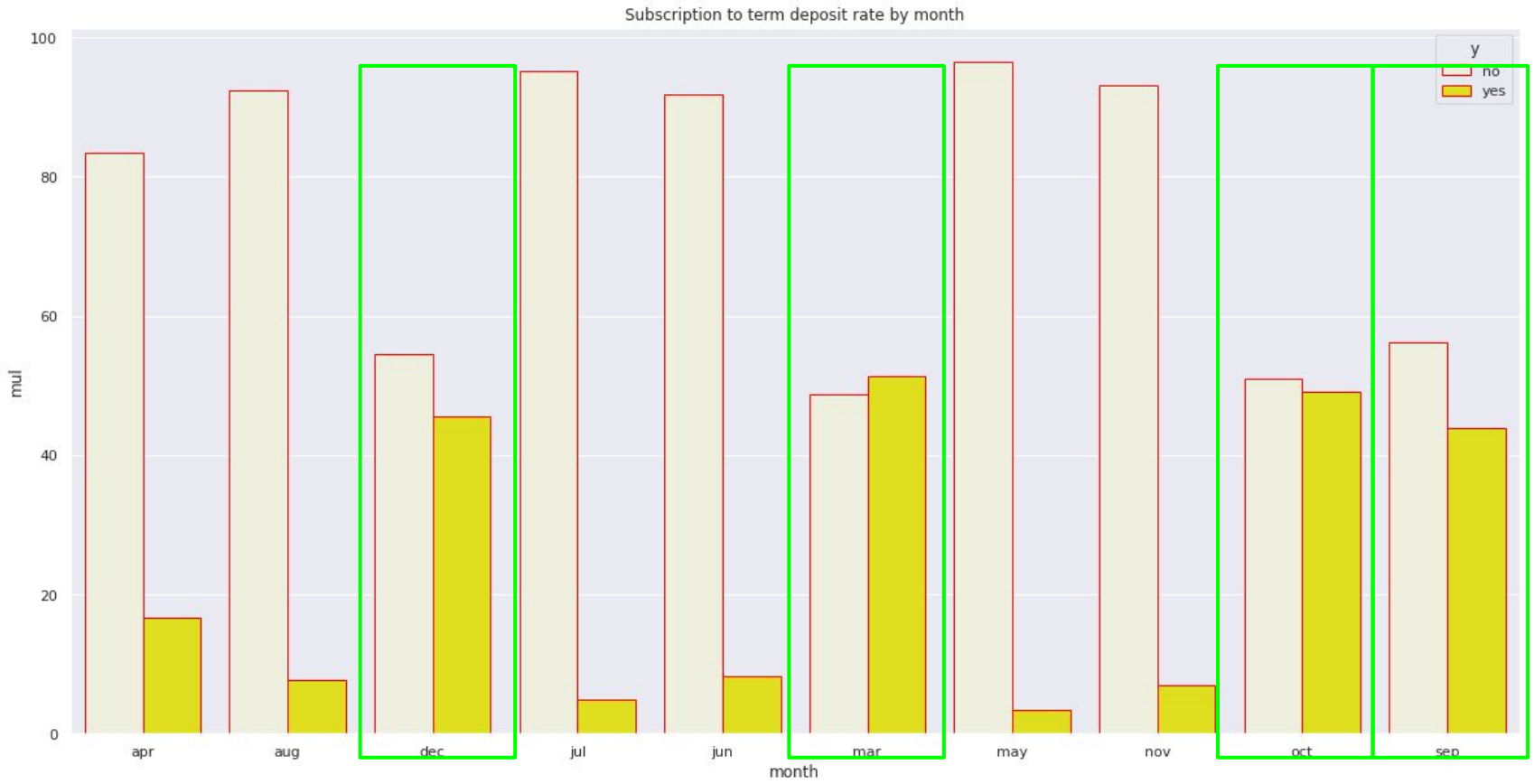
EDA Recommendations #5





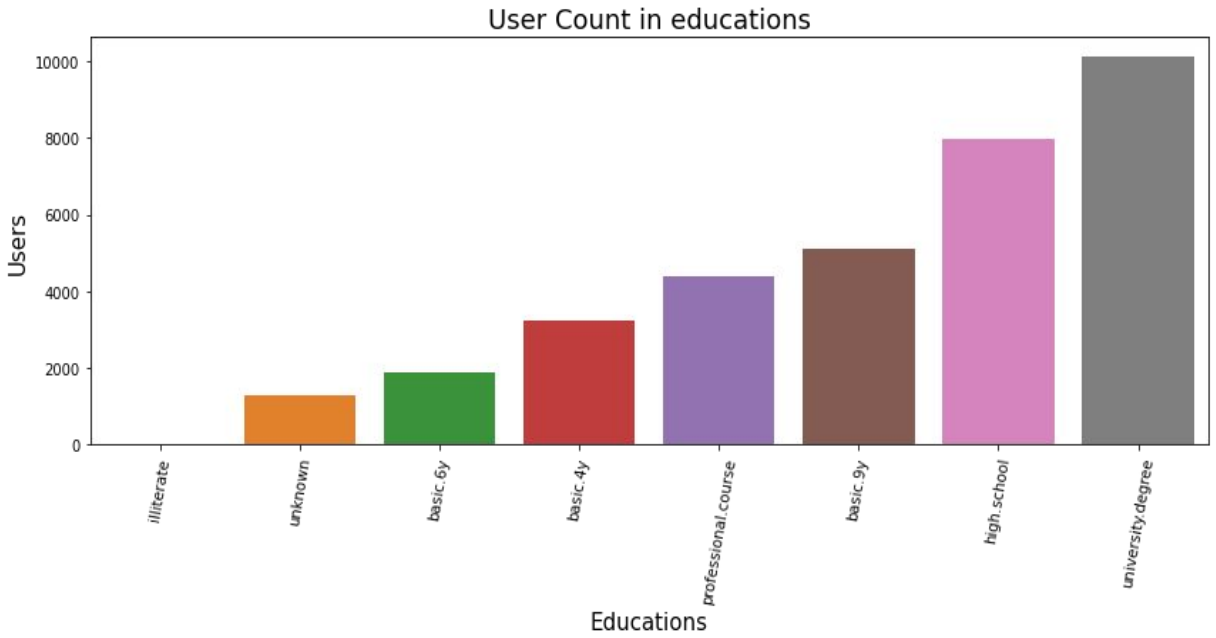
Contacts made in the December, March, October and September months can be increased, because they have the least amount of calls while they have the highest rate of acceptance

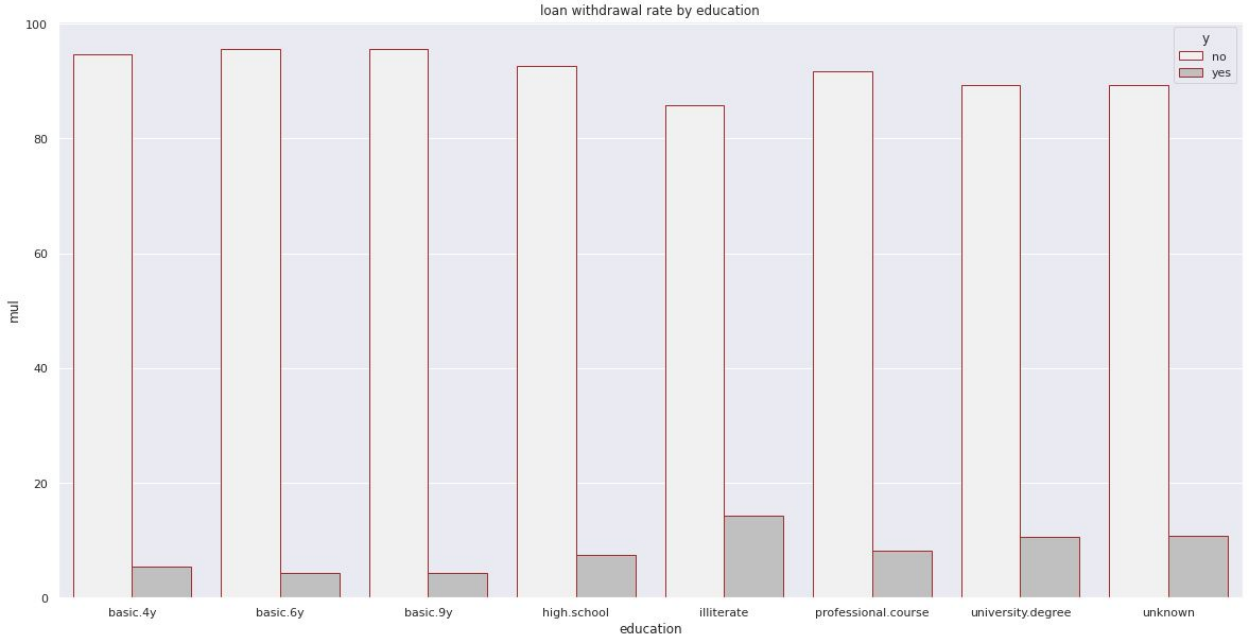
EDA Recommendations #5



Contacts made in the December, March, October and September months can be increased, because they have the least amount of calls while they have the highest rate of acceptance

EDA Recommendations #6





Even though, there is so little knowledge and data about illiterate people, increasing the amount of contacts to those people can be beneficial

Other than that other jobs seem to be similar in terms of “Yes” or “No” response ratio

Model Recommendations

- **It was thought to be that classification models provide better results, since our desired output column for the machine learning models is a categorical value.**

Thank You