Exploratory Data Analysis

and proposed modelling techniques for business users

Bank Marketing Campaign

Agenda

1. Team member details
2. Problem Statement
3. Approach
4. EDA
5. EDA recommendations with proposed models
6. Github Repository Link

Meet the team, ‘Sparagua’, specializing in Data Science:

➢ ***Daniela Alvarez*** (daniela.alvarez04@gmail.com)
Country: Peru
College/Company: Universidad de Piura (Peru), Datacamp, Kaggle Learn

➢ ***Akhil Nair*** (akhil.nair1908@gmail.com)
Country: India
College/Company: SIESGST Nerul, Mumbai University

Problem Statement

Experiencing a decrease in revenue, Portuguese bank now wants to predict which clients can subscribe to a term deposit.

Based on past activity, they want to develop a model to identify customers most likely to subscribe.

This would save their time, efforts and resources as they do not need to focus on clients that are unlikely to subscribe.

Approach

Separation

For performing EDA, we first separate our prediction target from our features.

Categorical Data

Object type data from our training set is checked along with our targets and plotted accordingly.

Numerical Data

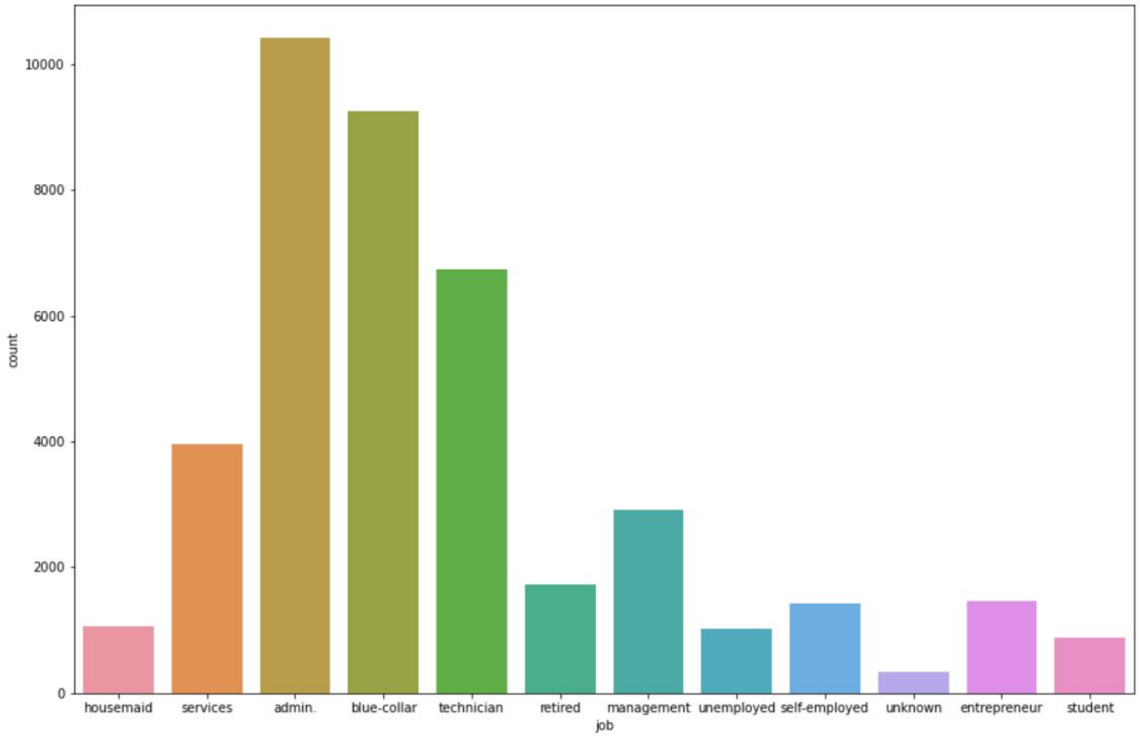
Numerical data is compared with one another in order to find hidden relations between features.

Exploratory Data Analysis (EDA)

Categorical Data

1. Jobs

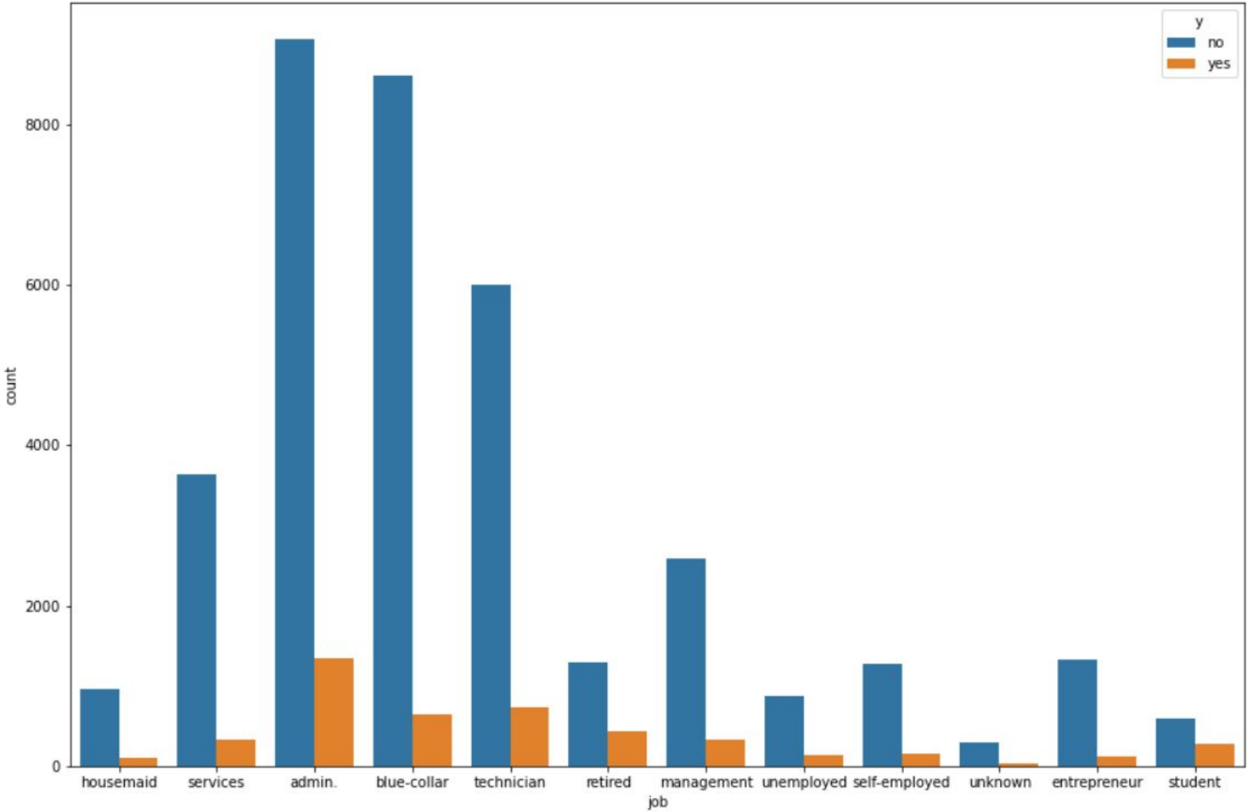
```
countplot_features('job')
countplot_targetvsfeature('job', target)
```



Categorical Data

1. Jobs

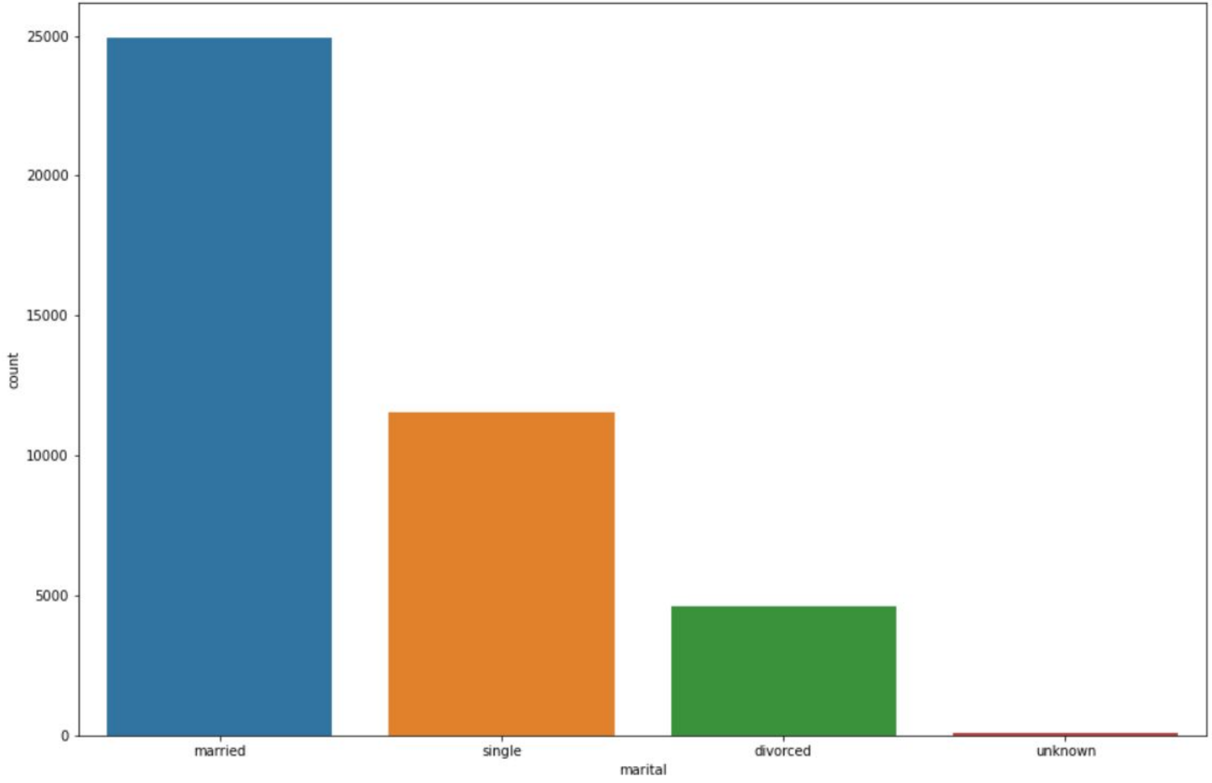
The most common jobs are administrative, blue collar and technical jobs,
whereas the least common ones are students, housemaids and unemployed individuals.



Categorical Data

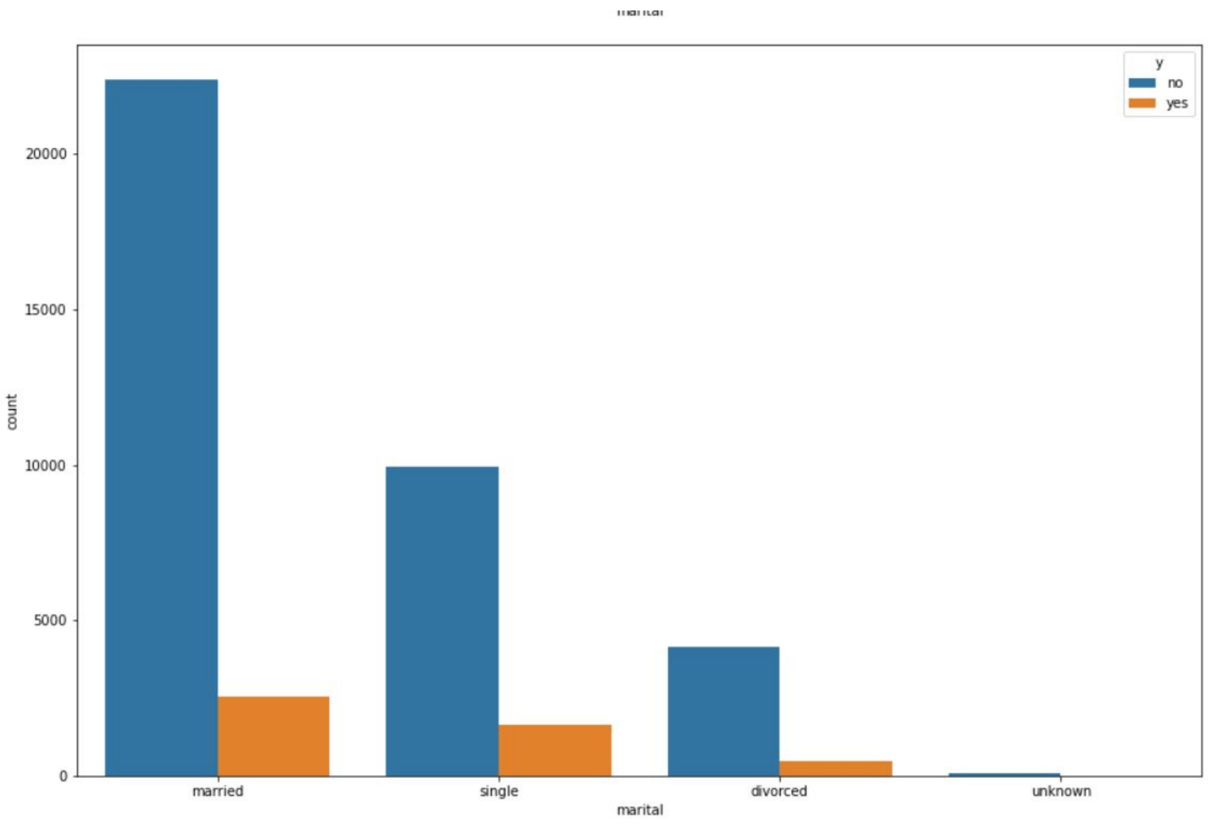
2. Marital Information

```
countplot_features('marital')
countplot_targetvsfeature('marital', target)
```



Categorical Data

2. Marital Information

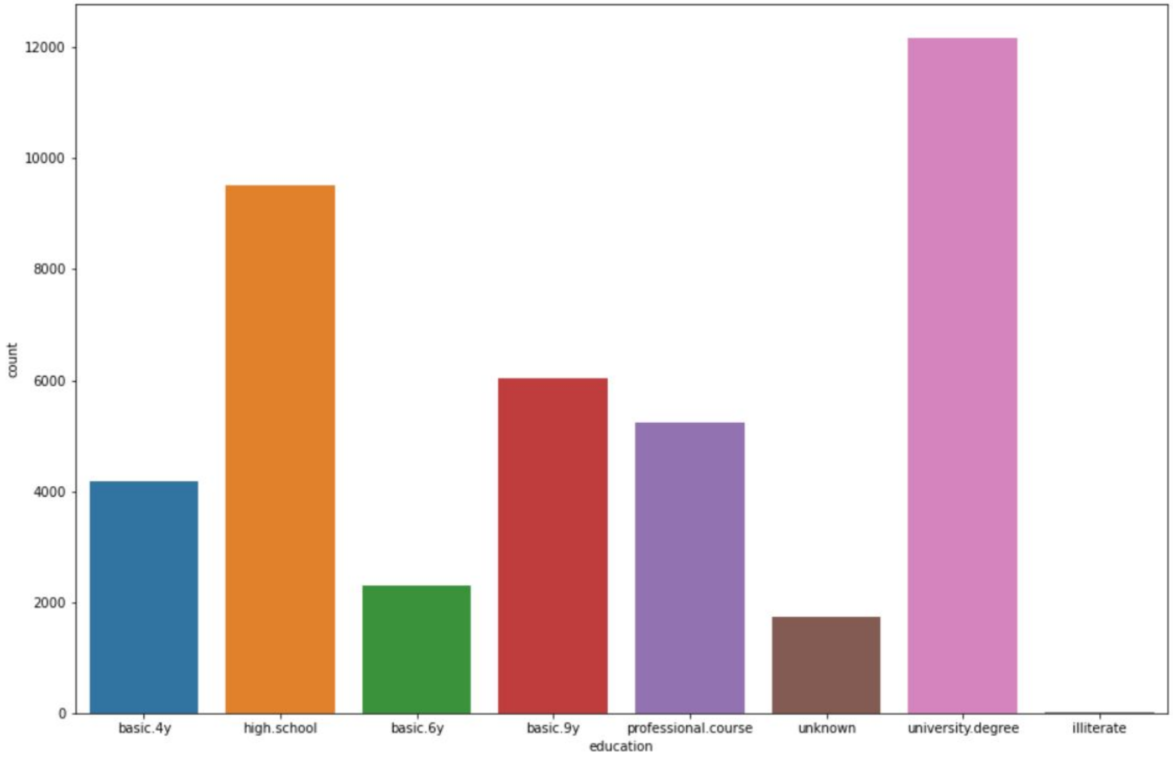


As we can see, most individuals are married.

Categorical Data

3. Education

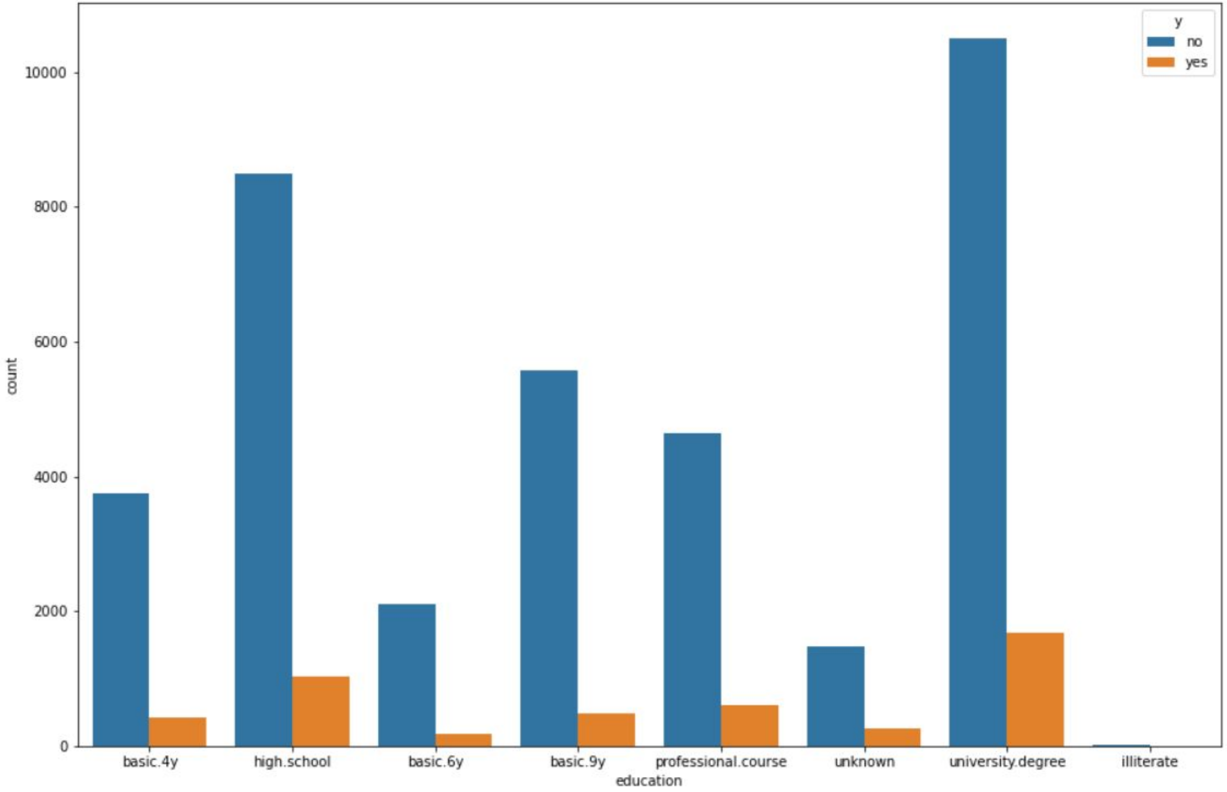
```
countplot_features('education')
countplot_targetvsfeature('education', target)
```



Categorical Data

3. Education

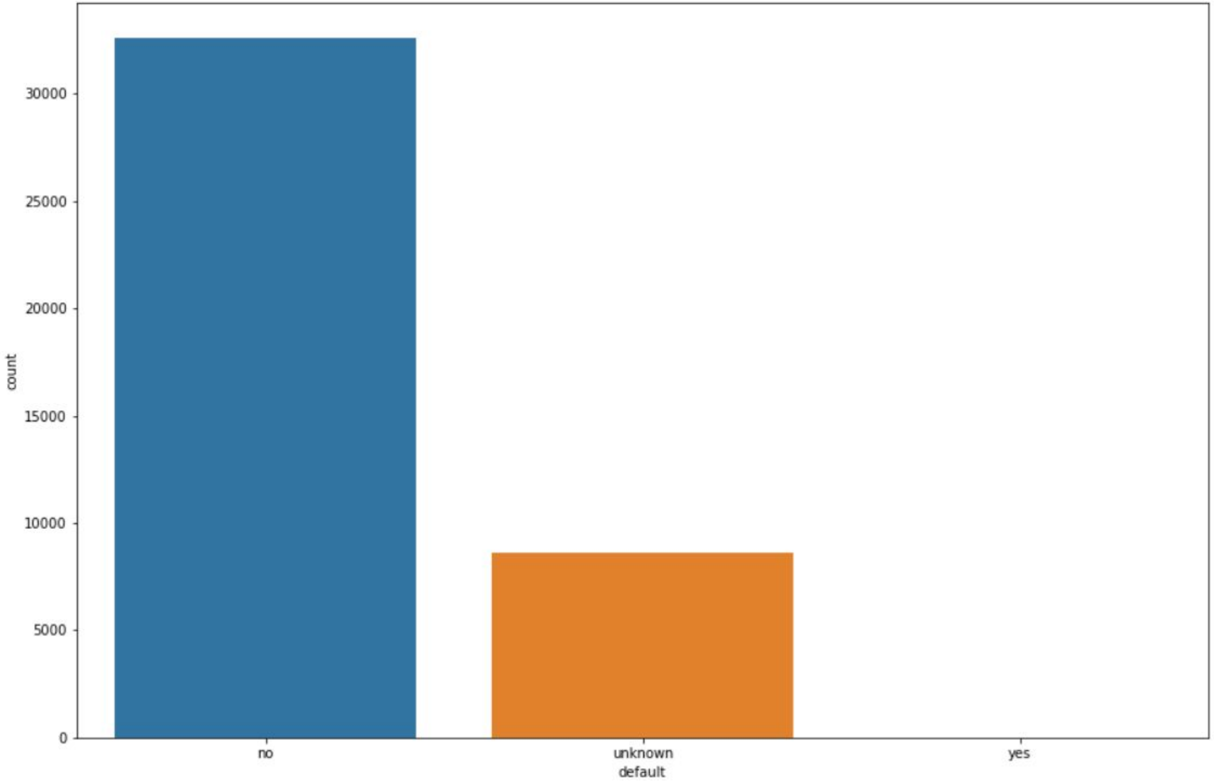
Most of the potential customers have a college degree, or a high
school degree. Very few are illiterate.



Categorical Data

4. Defaults

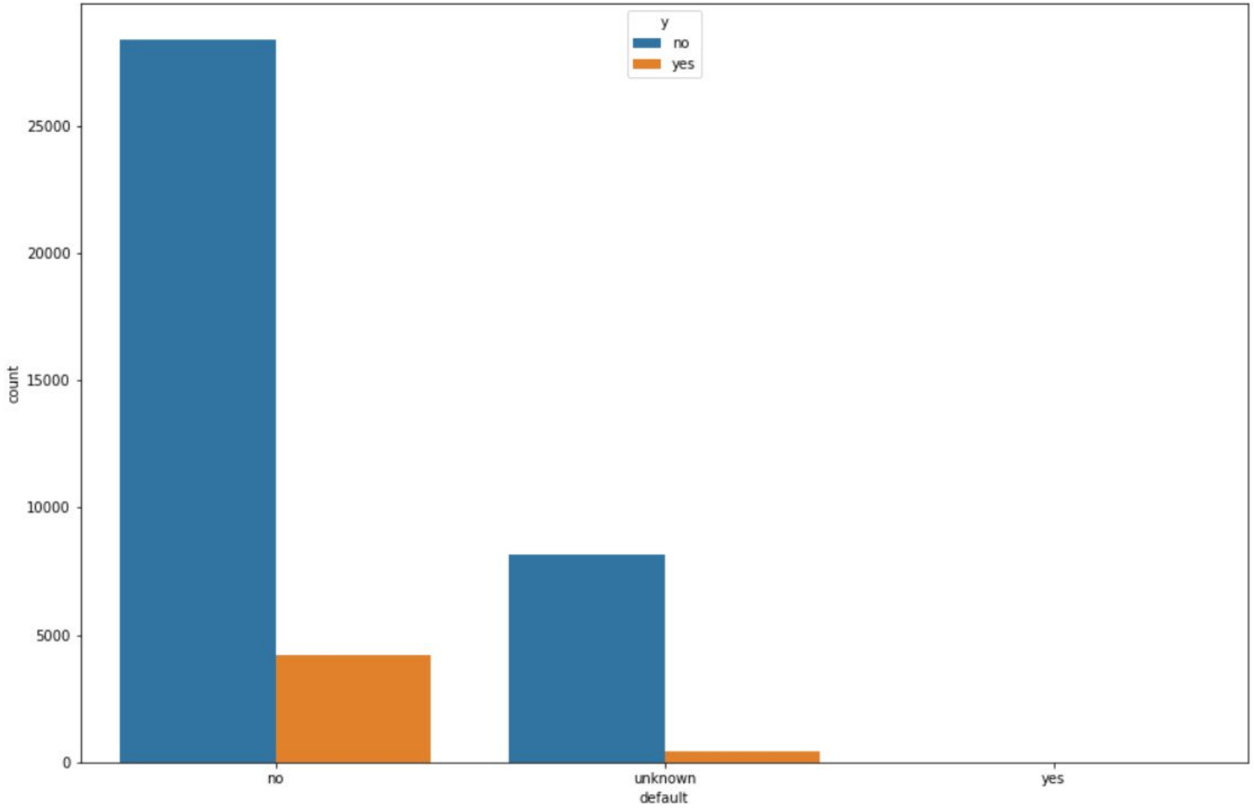
```
countplot_features('default')
countplot_targetvsfeature('default', target)
```



Categorical Data

4. Defaults

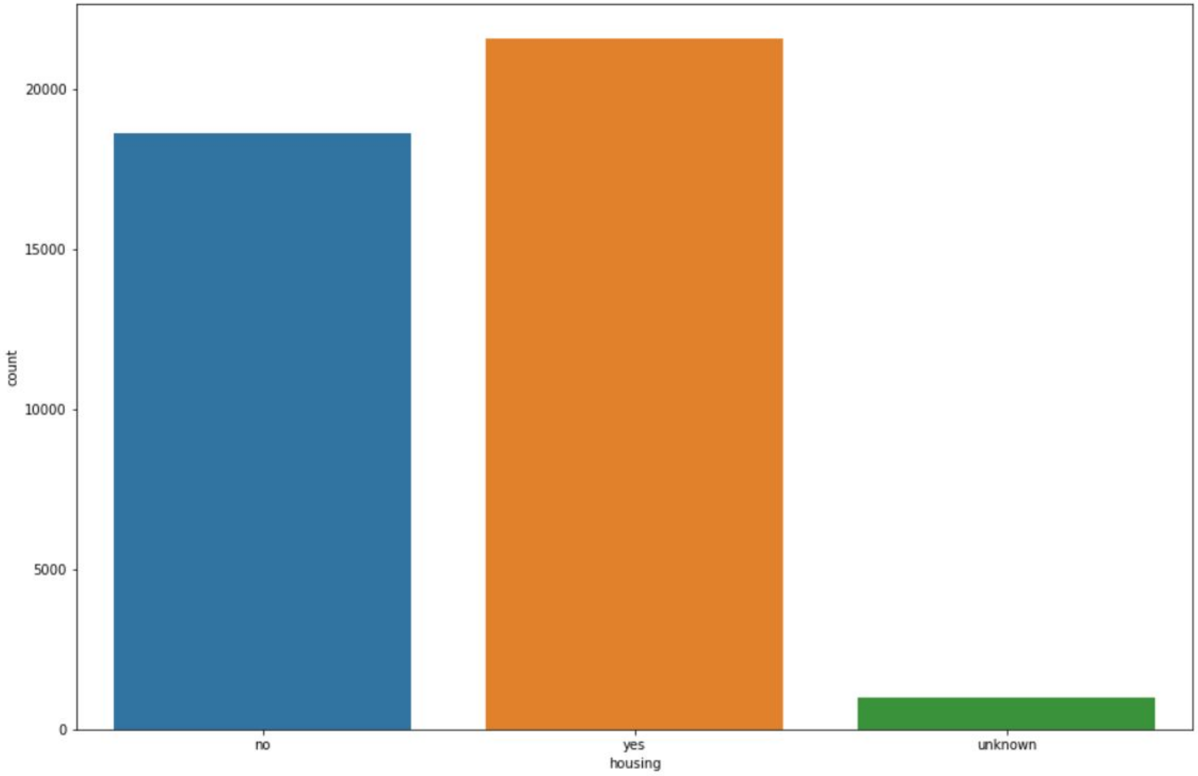
Most of the target clientele have no defaults.



Categorical Data

5. Housing Loans

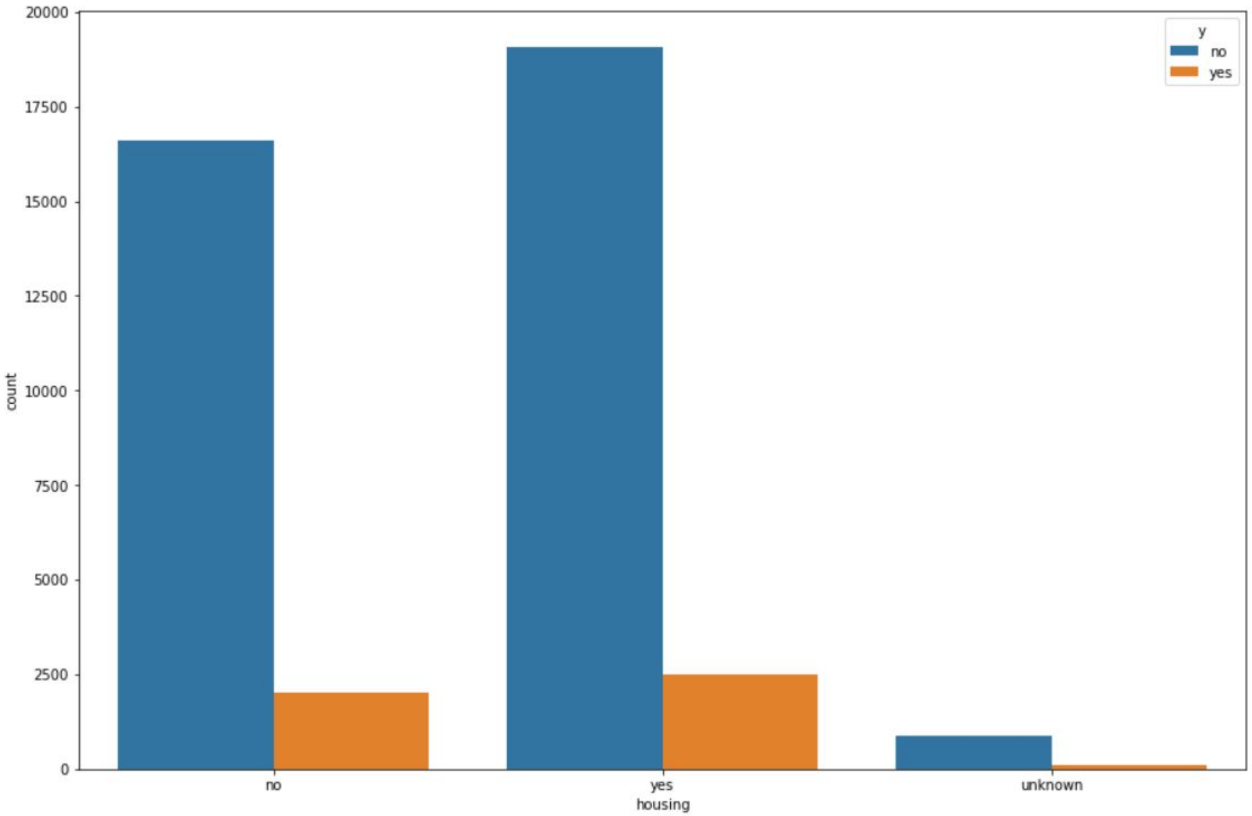
```
countplot_features('housing')
countplot_targetvsfeature('housing', target)
```



Categorical Data

5. Housing Loans

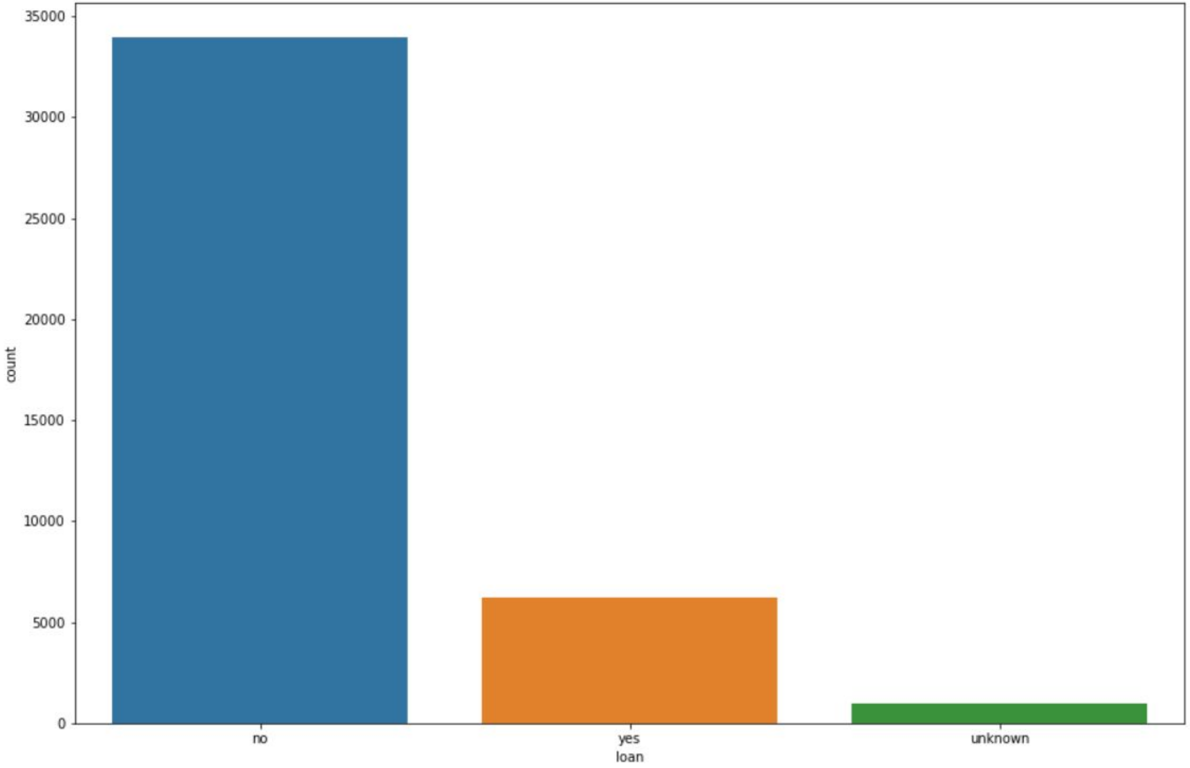
Most of the target clientele have a housing loan.



Categorical Data

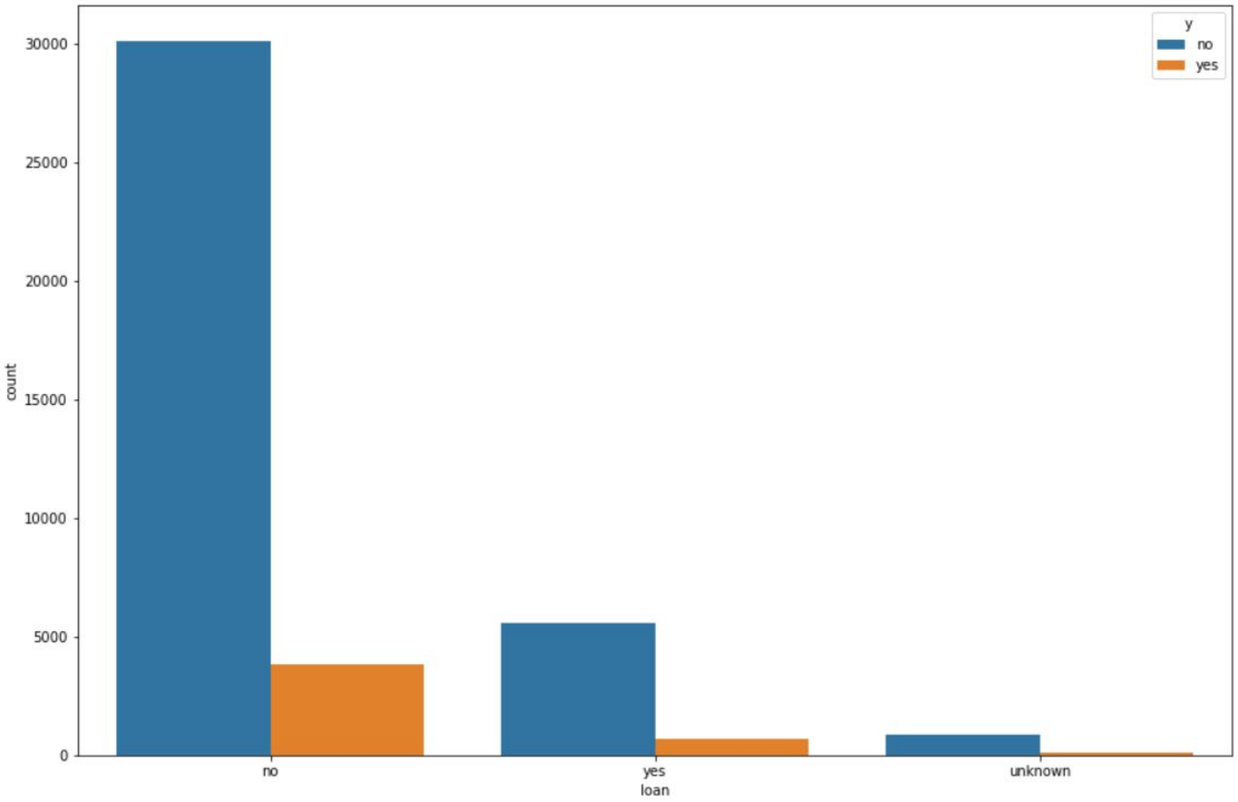
6. Personal Loans

```
countplot_features('loan')
countplot_targetvsfeature('loan', target)
```



Categorical Data

6. Personal Loans

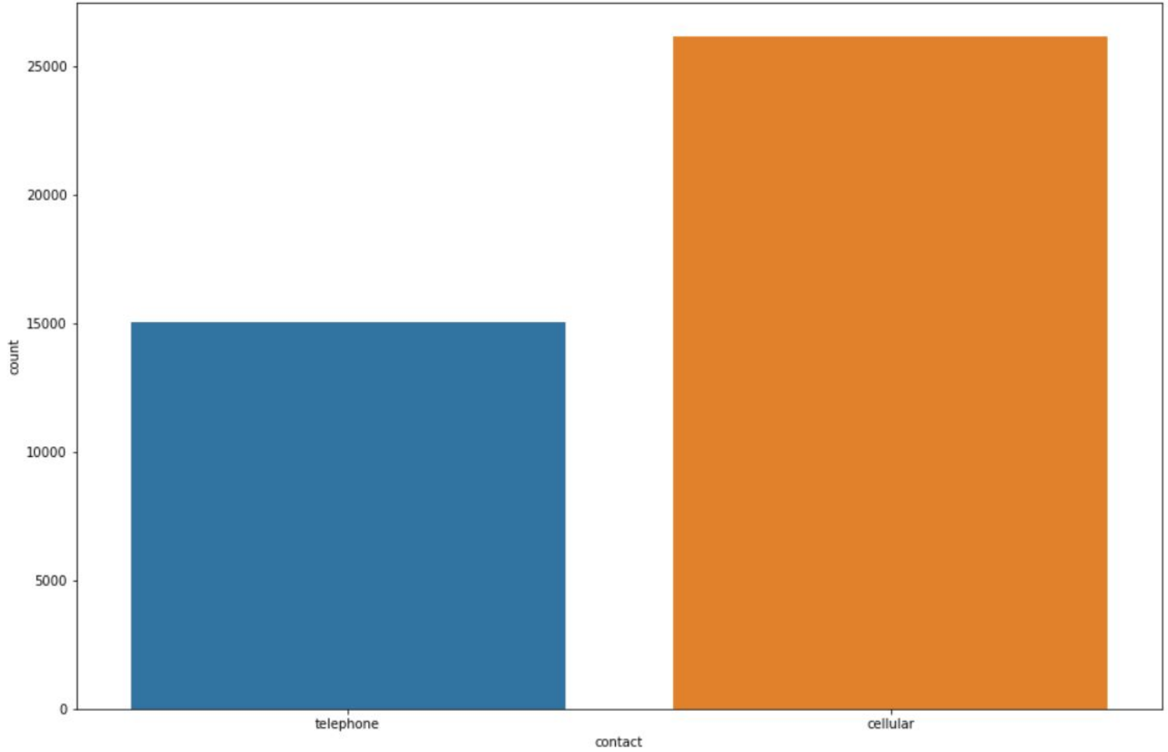


Most of the target clientele do not have any personal loans.

Categorical Data

7. Contact

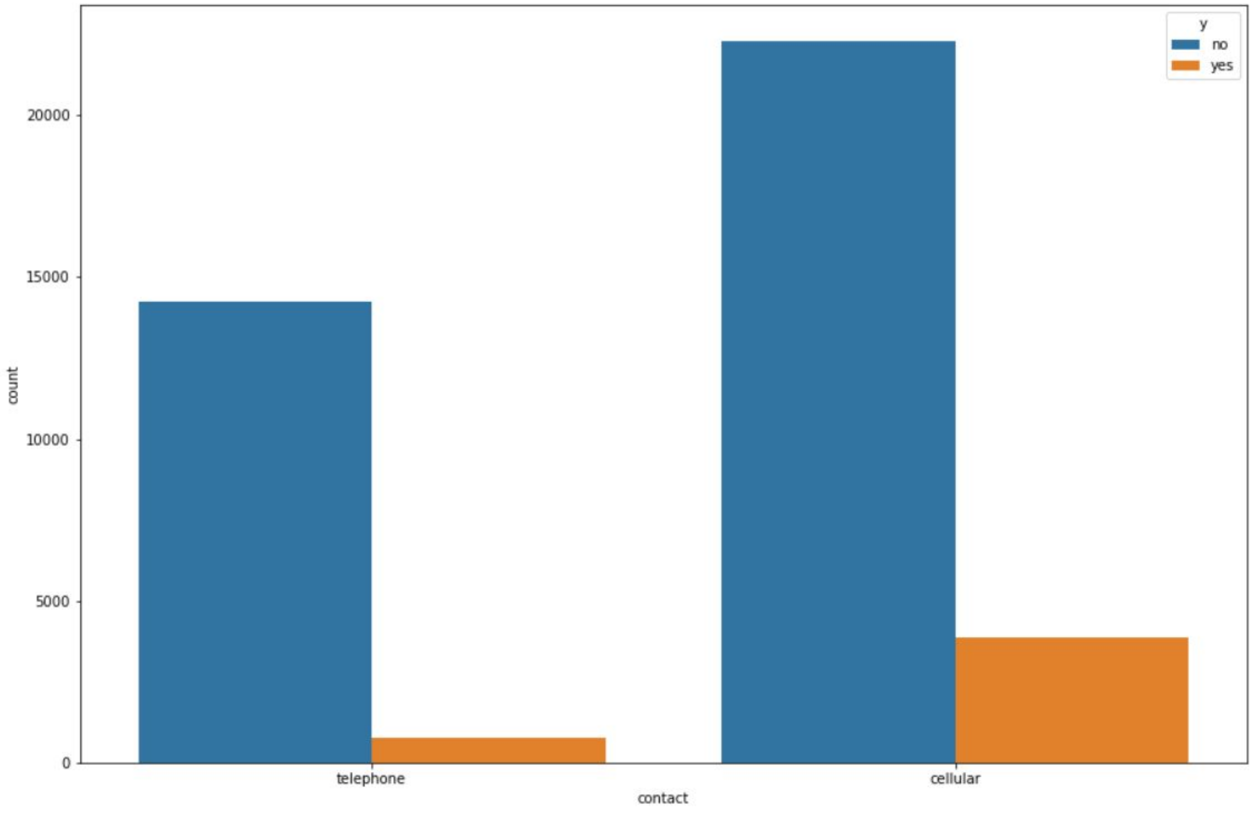
```
countplot_features('contact')
countplot_targetvsfeature('contact', target)
```



Categorical Data

7. Contact

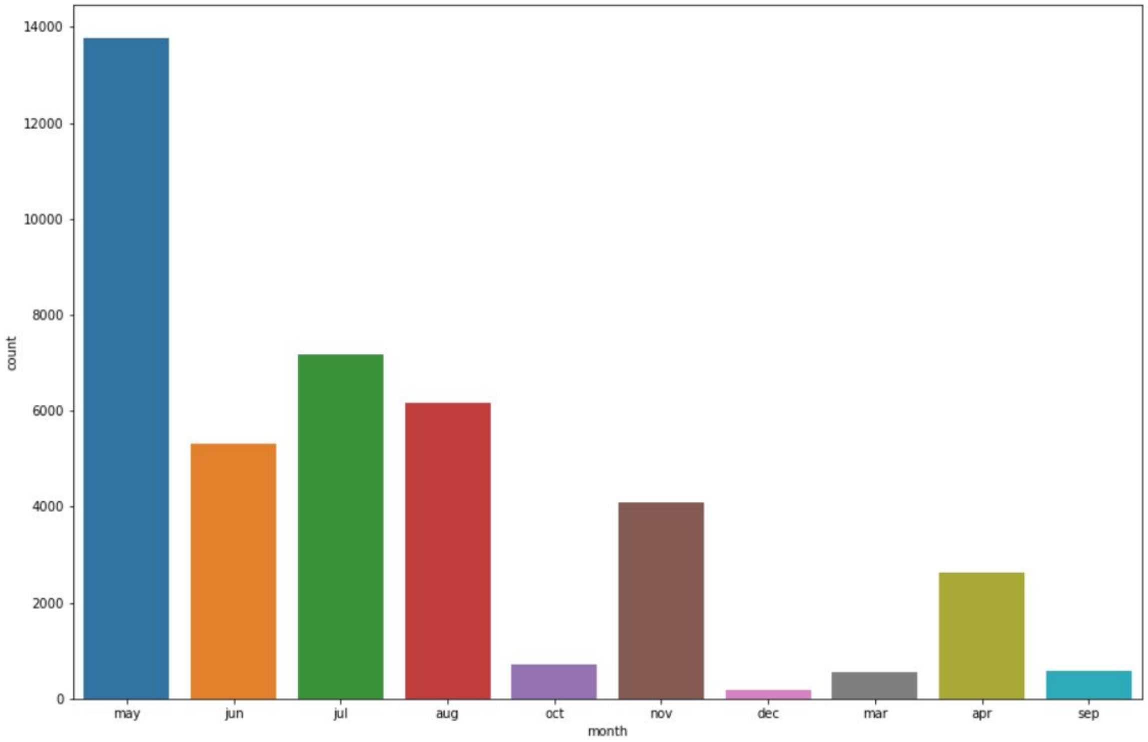
Most individuals have listed their preferred contact method as cellular phone over telephone.



Categorical Data

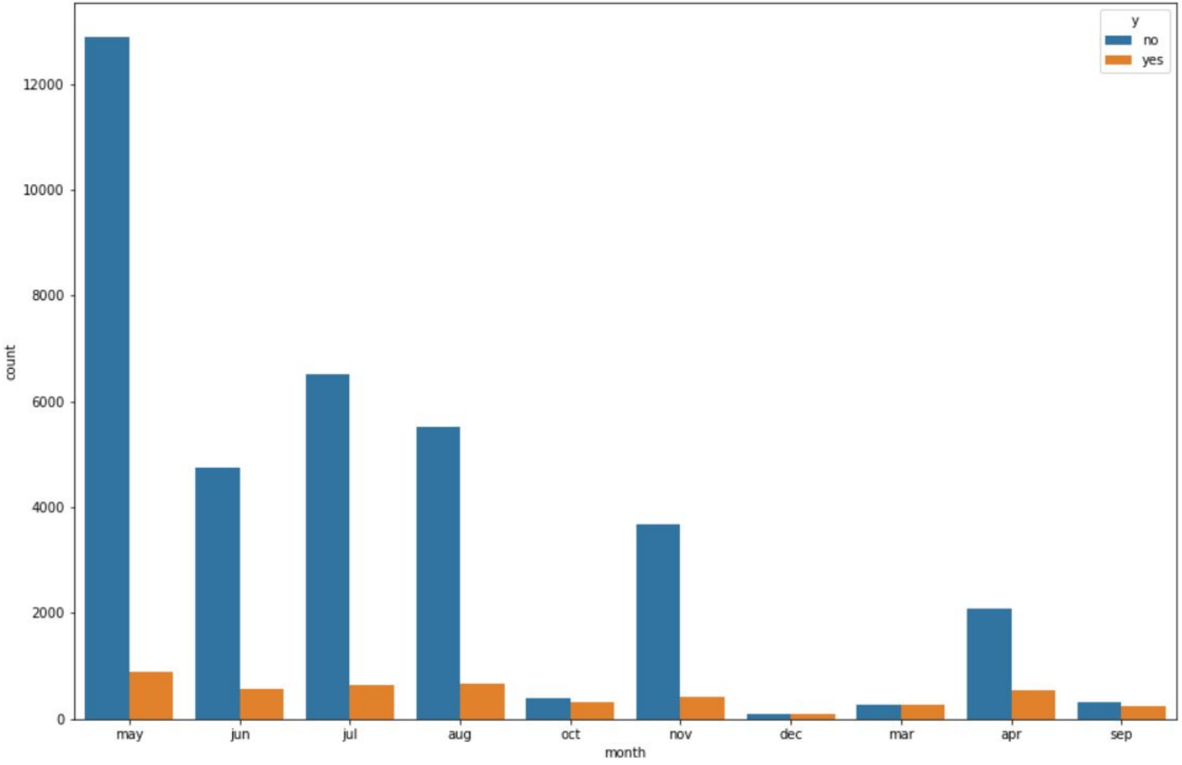
8. Month of contact

```
countplot_features('month')
countplot_targetvsfeature('month', target)
```



Categorical Data

8. Month of contact

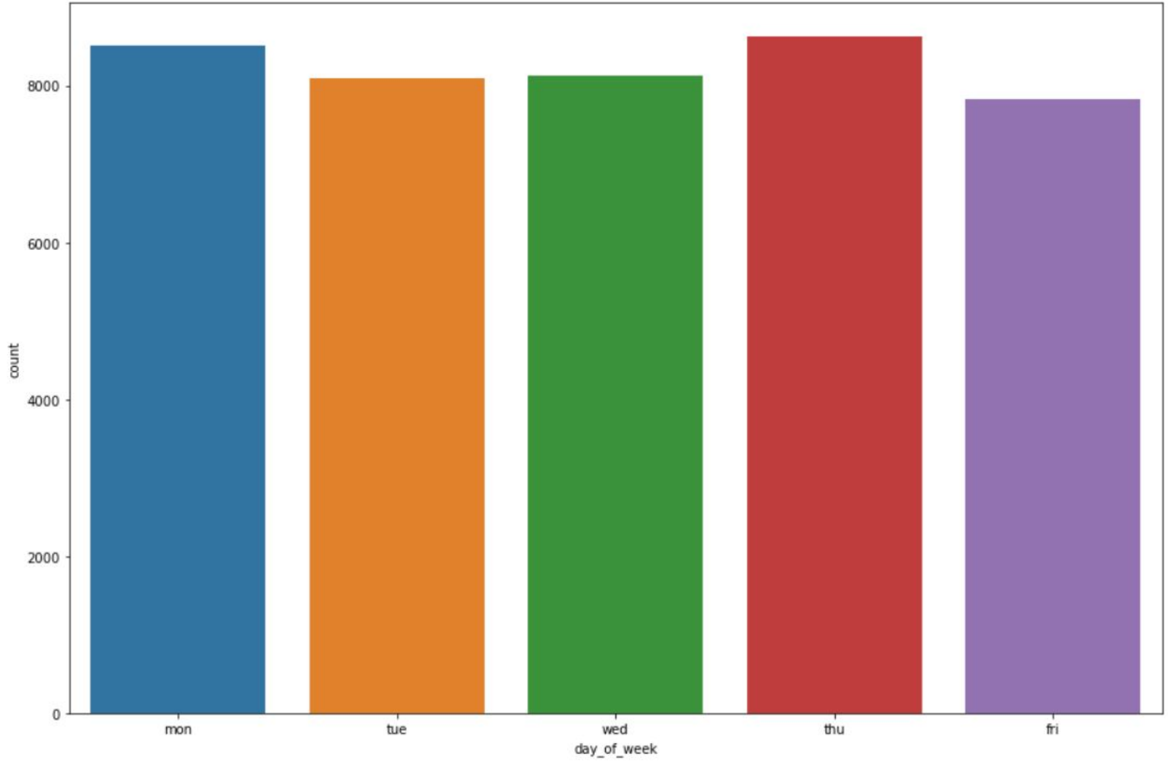


The last month of contact for most of them by far is May, followed by July, August and June.

Categorical Data

9. Day of week

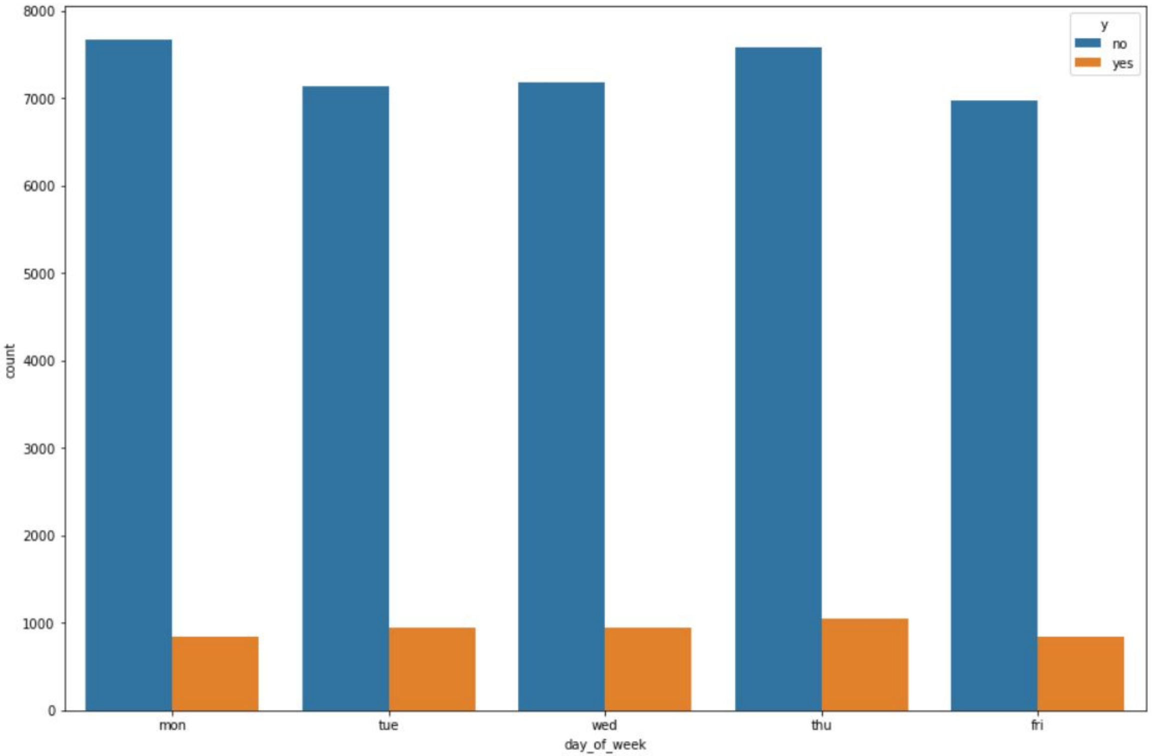
```
countplot_features('day_of_week')
countplot_targetvsfeature('day_of_week', target)
```



Categorical Data

9. Day of week

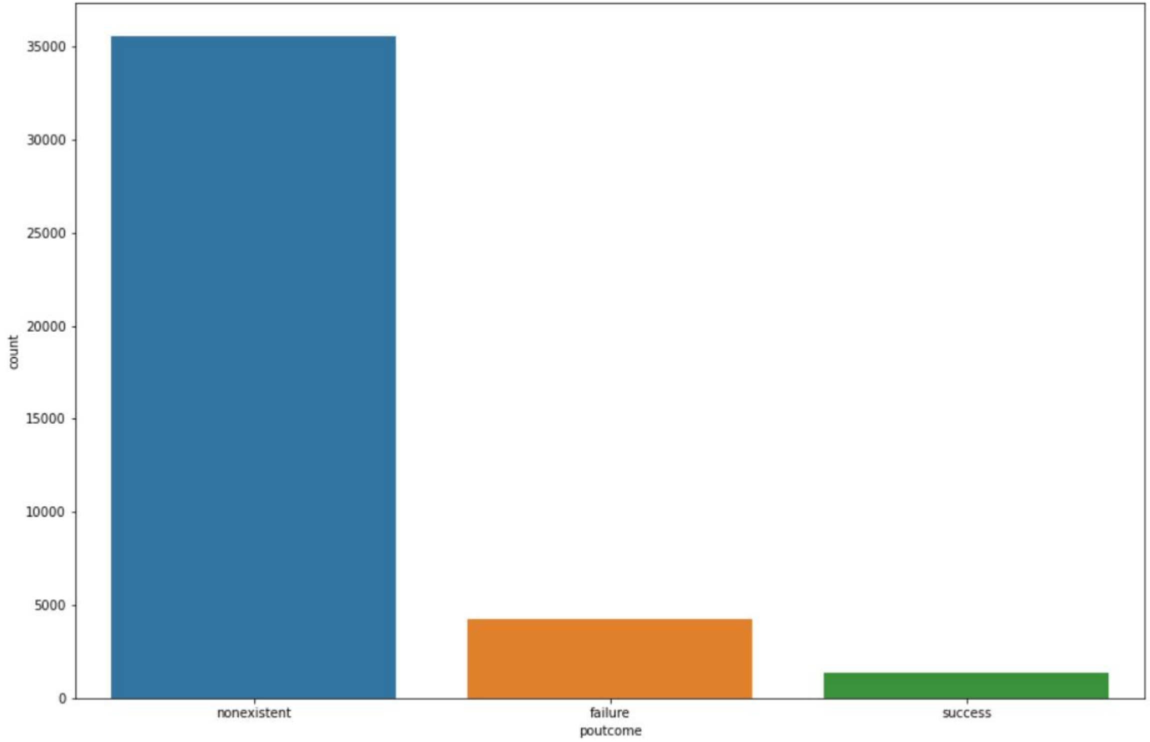
There is an even distribution in the last day of contact among the targets.



Categorical Data

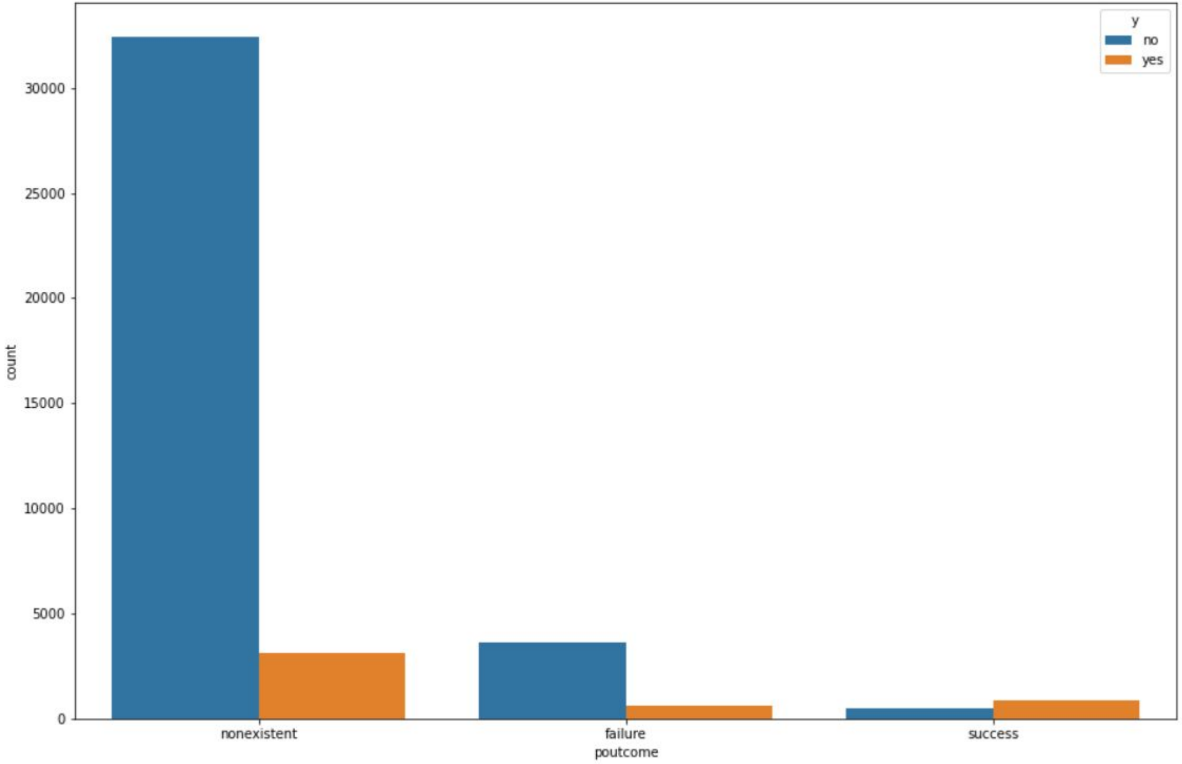
10. Previous Outcome

```
countplot_features('poutcome')
countplot_targetvsfeature('poutcome', target)
```



Categorical Data

10. Previous Outcome



Most of the past data shows us a nonexistent outcome, but this time a good portion of the individuals answered even made a term deposit.

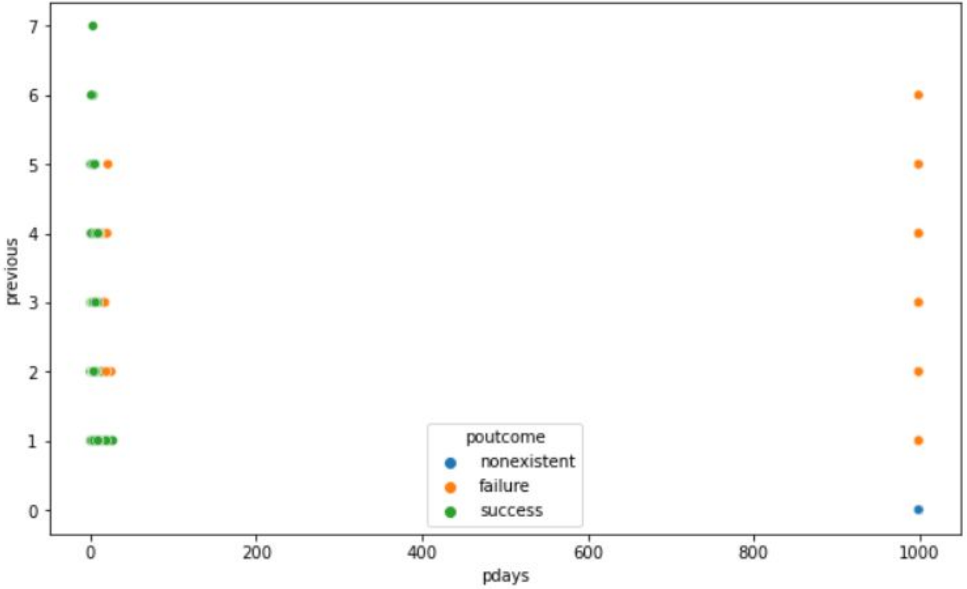
Numerical Data

1. Relation between 'pdays' and 'previous'

```
#Let's verify that there is coherence between the pdays variable
#(#of days since last contacted----> if 999 then client was never contacted before)
#and previous variable (# of times contacted in last campaign).

plt.figure(figsize=(10,6))
sns.scatterplot(x=features['pdays'], y=features['previous'], hue = features['poutcome'])
```

```
<AxesSubplot:xlabel='pdays', ylabel='previous'>
```



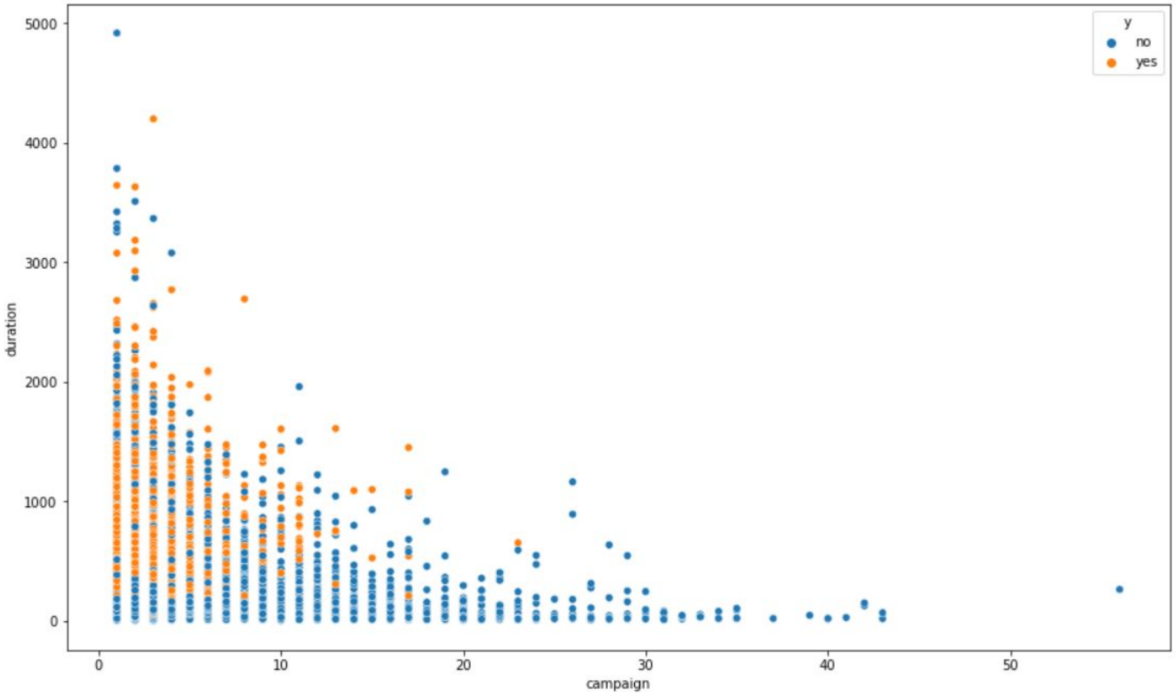
The orange dots on the left side represent negative responses from people contacted 2-5 times, and the green dots represent positive responses from people contacted 1-7 times.

Numerical Data

2. Relation between call duration and frequency

```
plt.figure(figsize=(15,9))
sns.scatterplot(x= features['campaign'], y= features['duration'], hue = target)
```

```
<AxesSubplot:xlabel='campaign', ylabel='duration'>
```



There is a negative relation between the number of calls made to a prospective client and the duration of these calls.

Clients that have been called over 12 times do not respond, or give a very brief response.

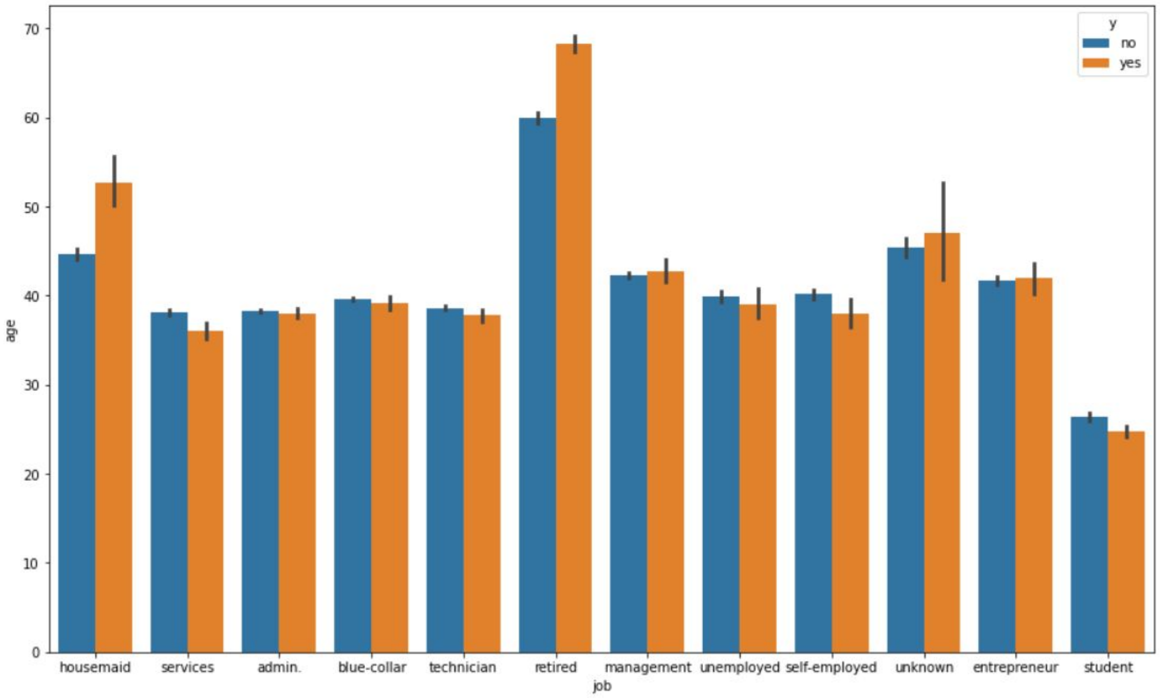
Numerical Data

3. Relation between job and age

```
#Can we visualize what job and age is a more common client for a term deposit?

plt.figure(figsize=(15,9))
sns.barplot(x= features['job'], y= features['age'], hue= target)
```

```
<AxesSubplot:xlabel='job', ylabel='age'>
```



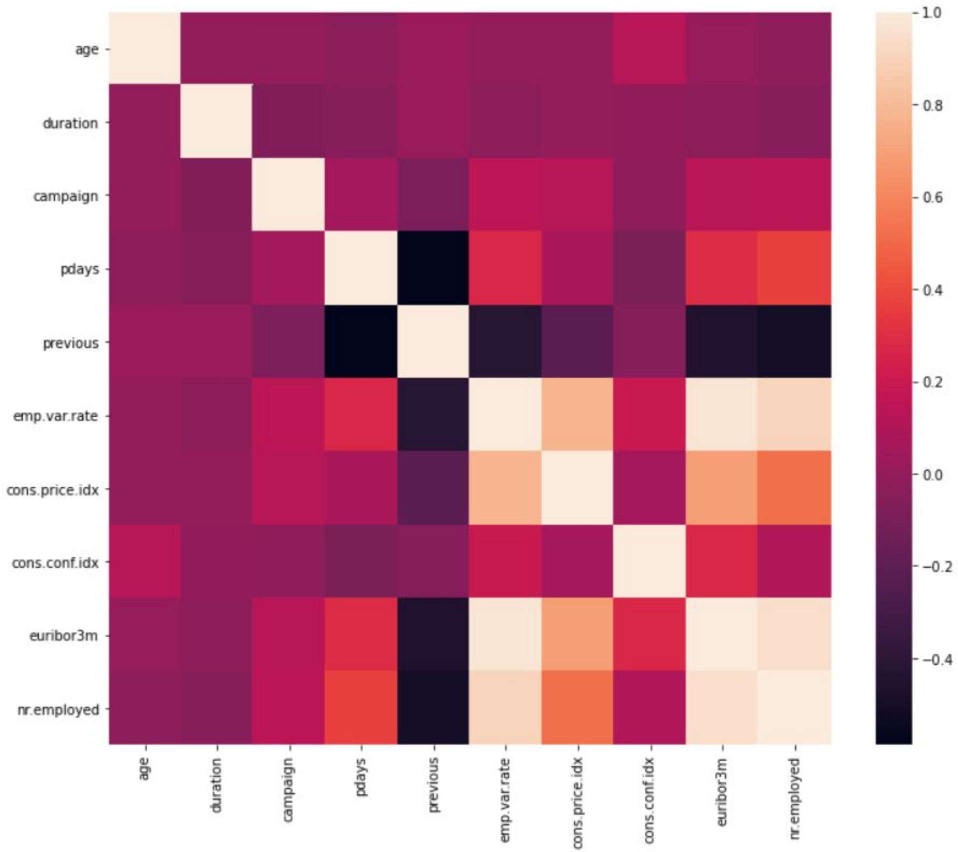
There seems to be a larger difference between the ‘yes’ and ‘no’ subscribers among the retired people aged 65-70 years old and the housemaids aged 50-55 years old than the other potential clients.

Numerical Data

4. Other insights

The consumer price index is strongly correlated with the bank's interest rates and employee variation rate, i.e., the higher the price index, the greater the interest rate.

The employee number also has a strong correlation with the employee variation rate and bank interest rates.



Final recommendation

along with proposed modelling techniques

EDA recommendation

The frequency of contacts made with the prospective clients has a very strong negative correlation with the bank's interest rates and employee variation rates, i.e, the greater the rates of interest, the lesser the number of contacts that had been performed before this campaign.

A lower interest rate could therefore increase the number of contacts made this campaign.

Proposed models for this dataset

- As our primary goal is to predict if a deposit will be made or not, the output would be binary. Classification models would therefore be our best bet.

- Performing cross validation among classification models, we found these to be the best models for this case to be:

    - Logistic Regression

    - Support Vector Classifier (SVC)

GitHub Link:

https://github.com/danielaaz04/Bank-Marketing-Campaign

Thank you.