• Exploratory Data Analysis (EDA) performed on the data

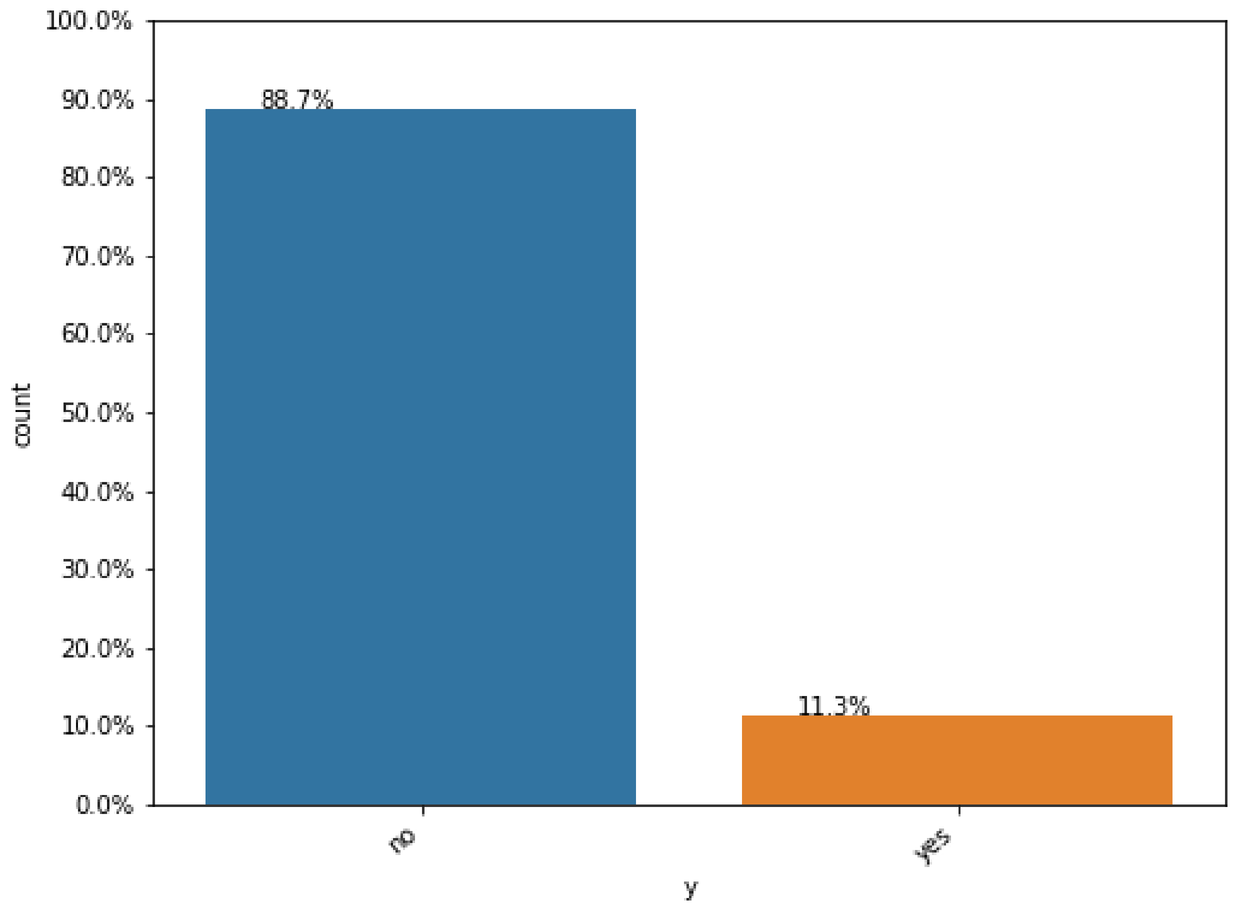
With the help of the libraries such as matplotlib and seaborn, our goal to perform some exploratory data analysis can be easily accomplished. I will reveal some insights which I've developed.

• Distribution of Class variable

➢ In which class does majority data points belongs to? Yes, or no?

I was conducting some analysis in comparing in which class does most of the data points belongs to.

As shown below, in the plot we can see that majority of datapoints belong to No class labels with 88.7% and minority of class belongs to 11.3% so the ratio of No: Yes is 8:1.



- **Univariate Analysis**

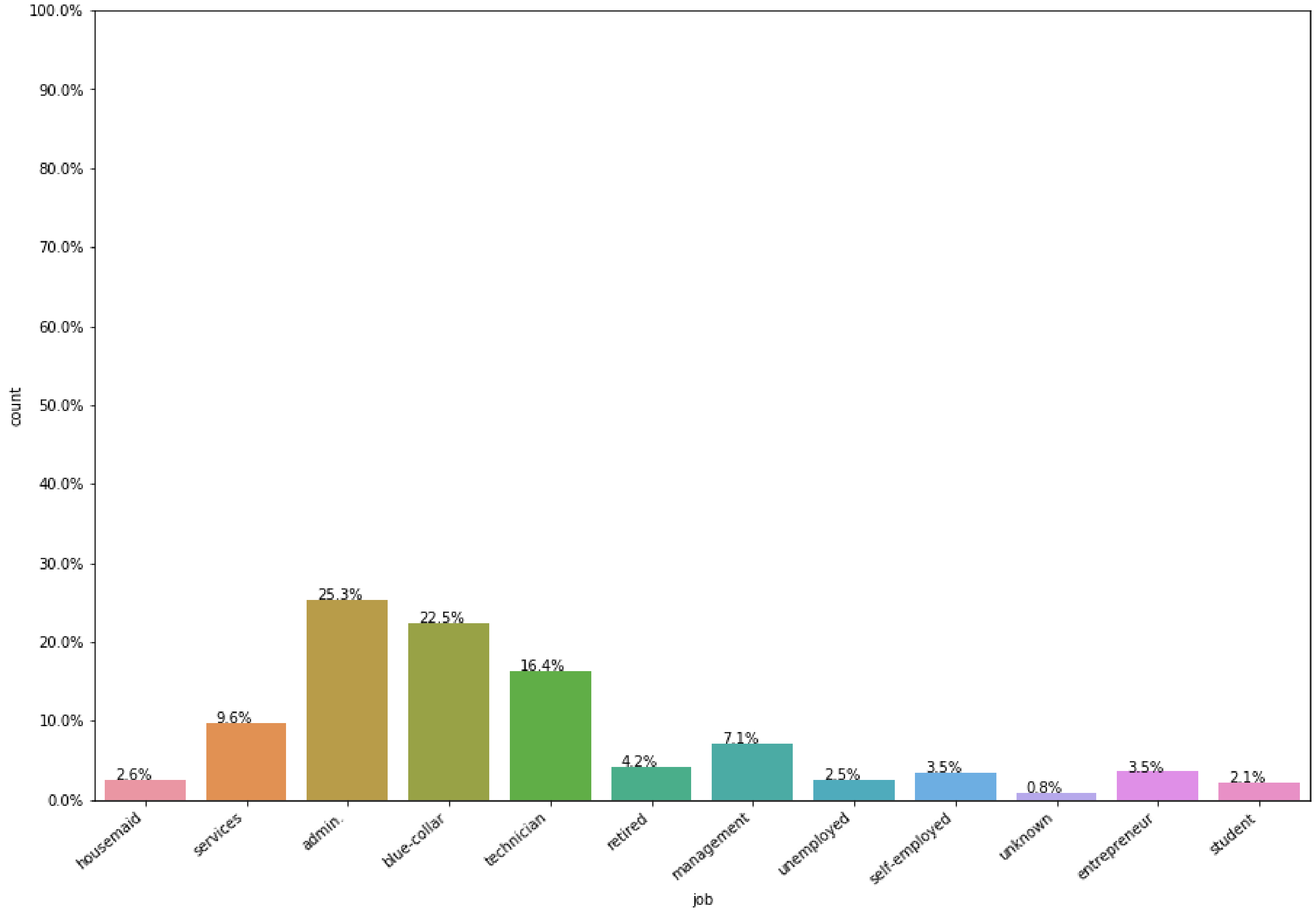
  EDA

Categorical Variables:

Let’s start doing EDA on rest of the columns of the datapoints.

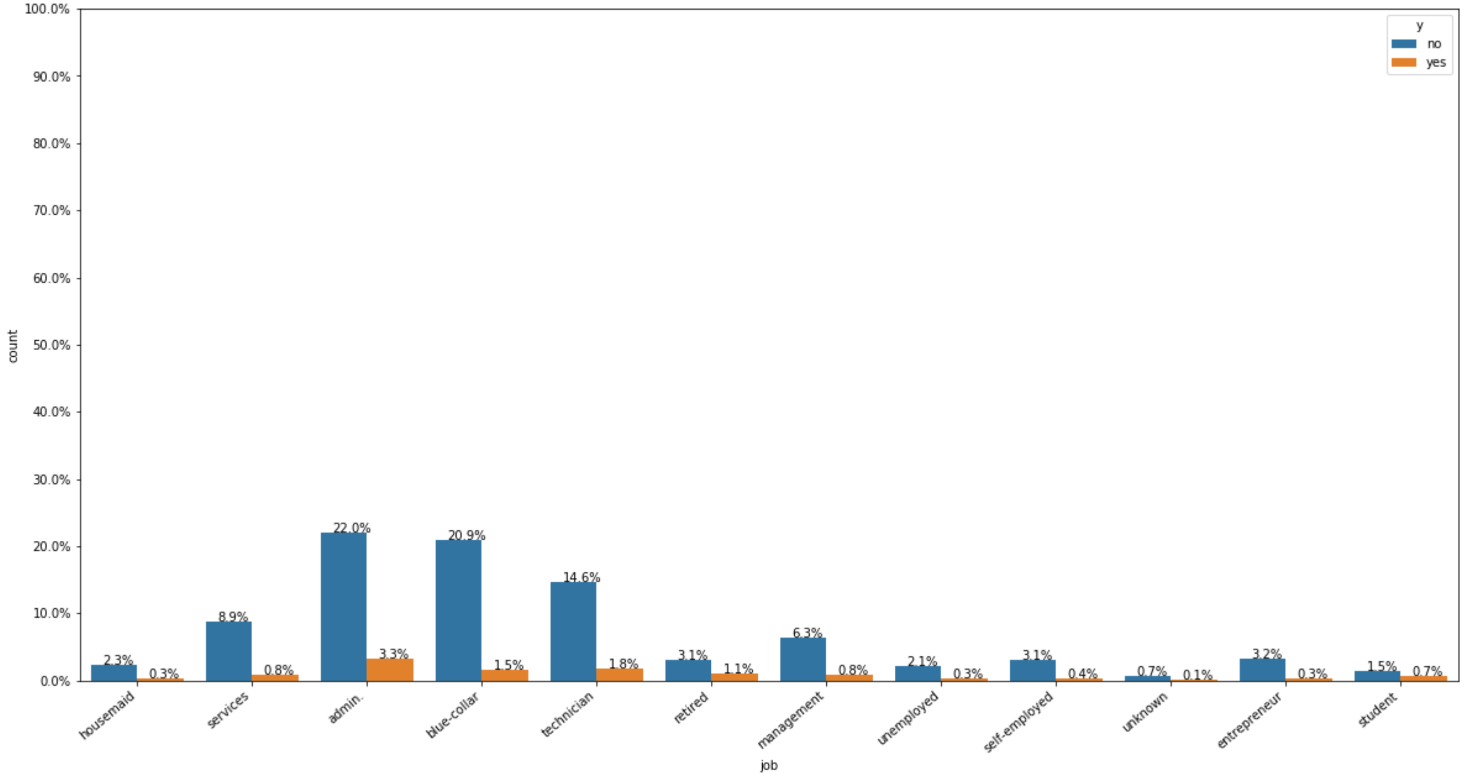
- ## Feature: Job (Categorical variable)

This a categorical feature which means the type of job, and with the values **‘admin’,’blue-collar’,’entrepreneur’,’housemaid’,’management’,’retired’,’self-employed’,’services’,’student’,’technician’,’unemployed’,’unknown’.**



- From the above distribution we can see that most of the customers that have jobs as **"admin", "blue-collar"** or **"technician"** have been contacted by the bank.
- One interesting thing to find out would be to see the distribution for each classes as well.
- For example, how many people who work as an admin have subscribed a term deposit.

- **Which job profession has the highest rate for subscribing a term deposit and which has the highest rate of not subscribing?**

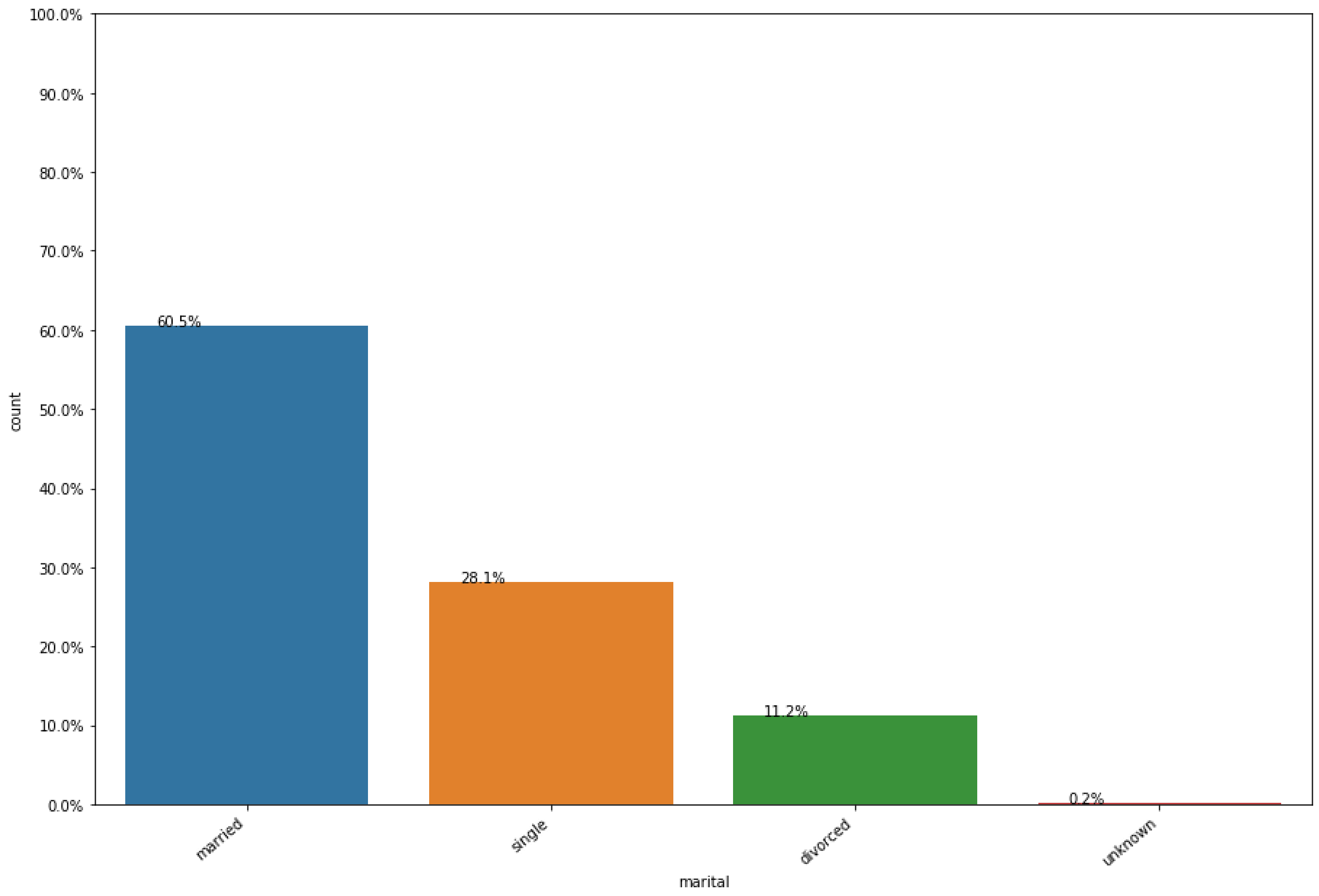


- From the plot above, we can see that the customers who have jobs of admin have the highest rate of subscribing a term deposit, but they are also the highest when it comes to not subscribing. This is simply because we have more customers working as admin than any other profession.

➢ Feature: Marital (Categorical feature)

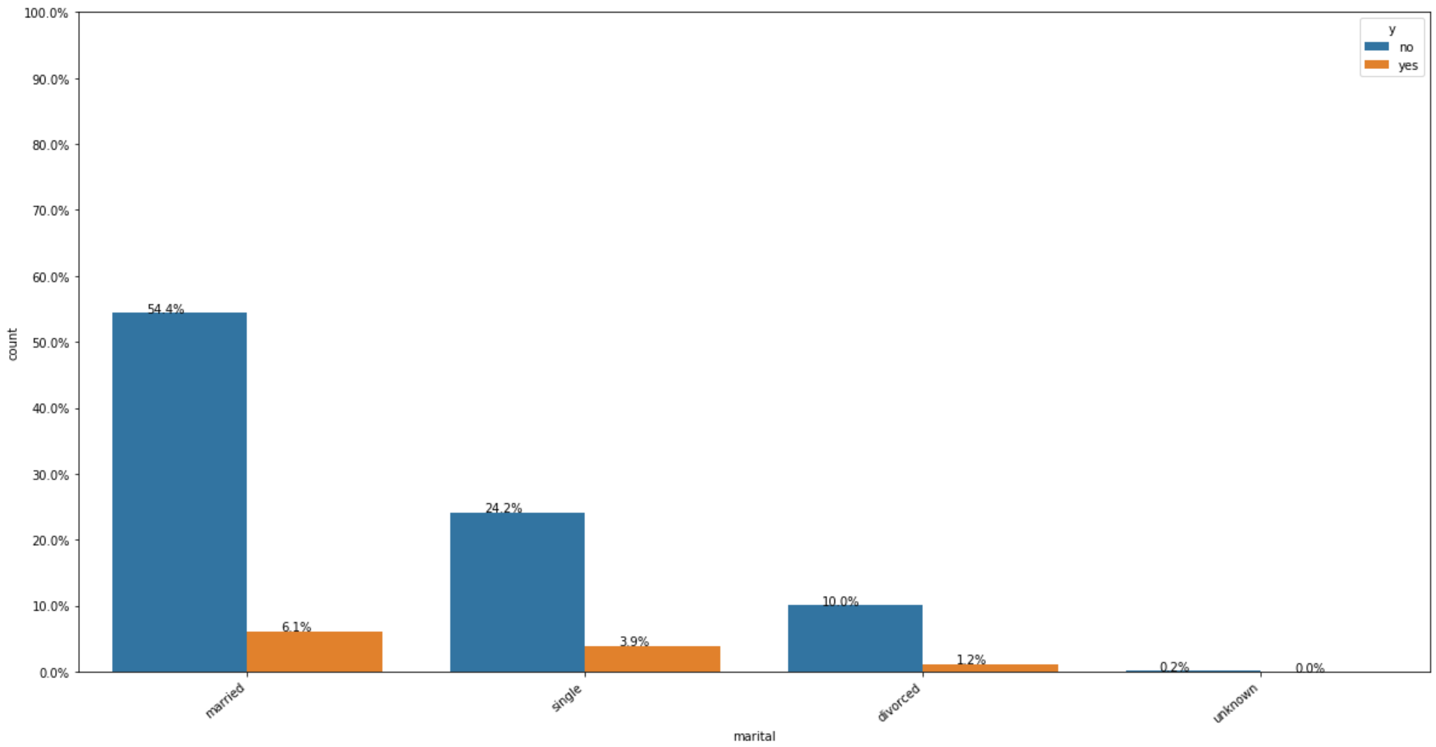
This is a categorical feature which means marital status, and with the values **‘divorced’,’married’,’single’ and ’unknown’**

note: ‘divorced’ means divorced or widowed.



- From the plot above as we can see, customers who have been contacted the most are married.
- About 0.2% of marital status of customer is unknown.

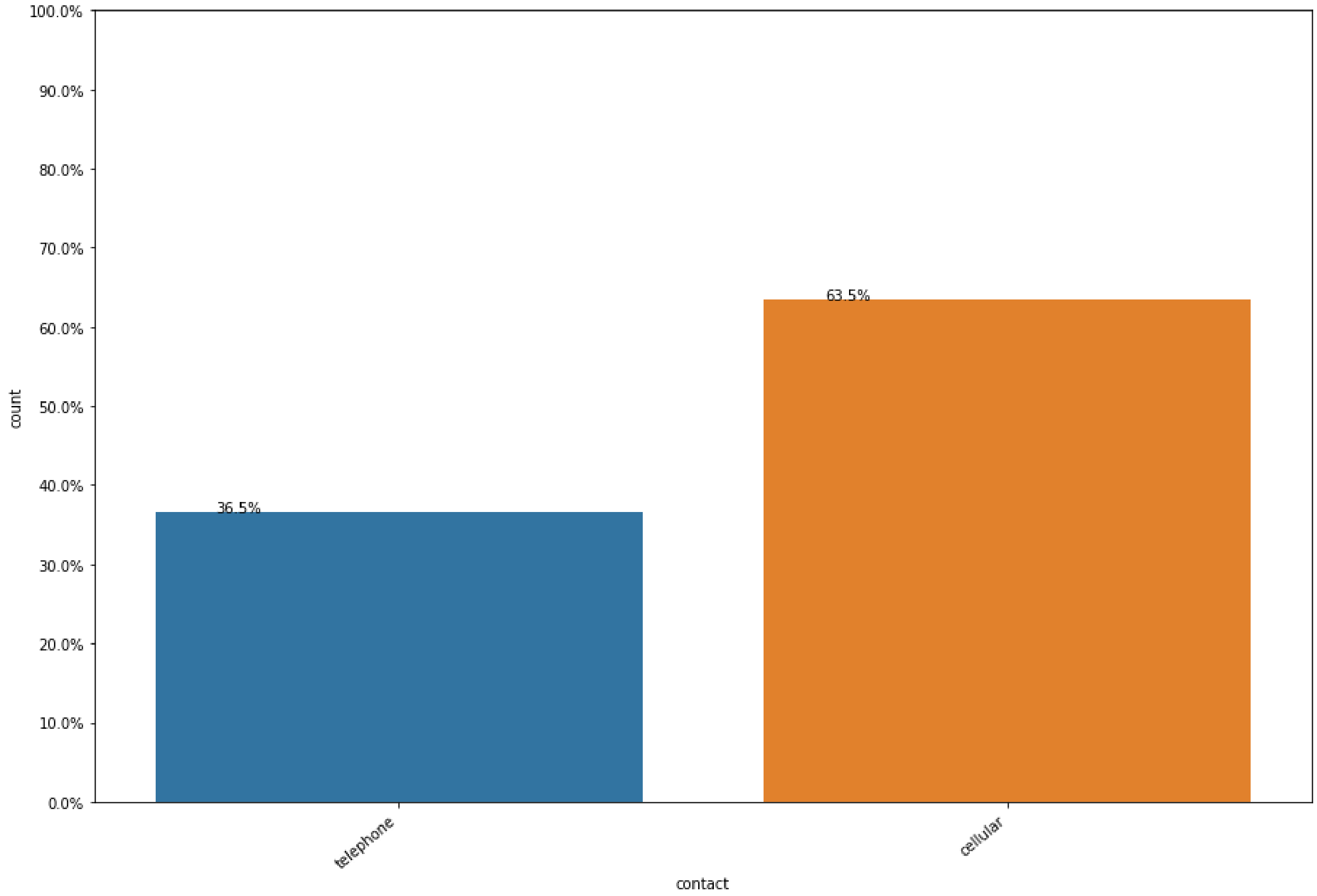
➢ Which marital status subscribes the most for long term deposits?



- From the plot above we can see that married people have subscribed for long term deposits more than any other people with any marital status.
- They are also the most one's who have turned down the deposits offered by the bank.
- People whose status is unknown have not subscribed to the long-term deposits at all.

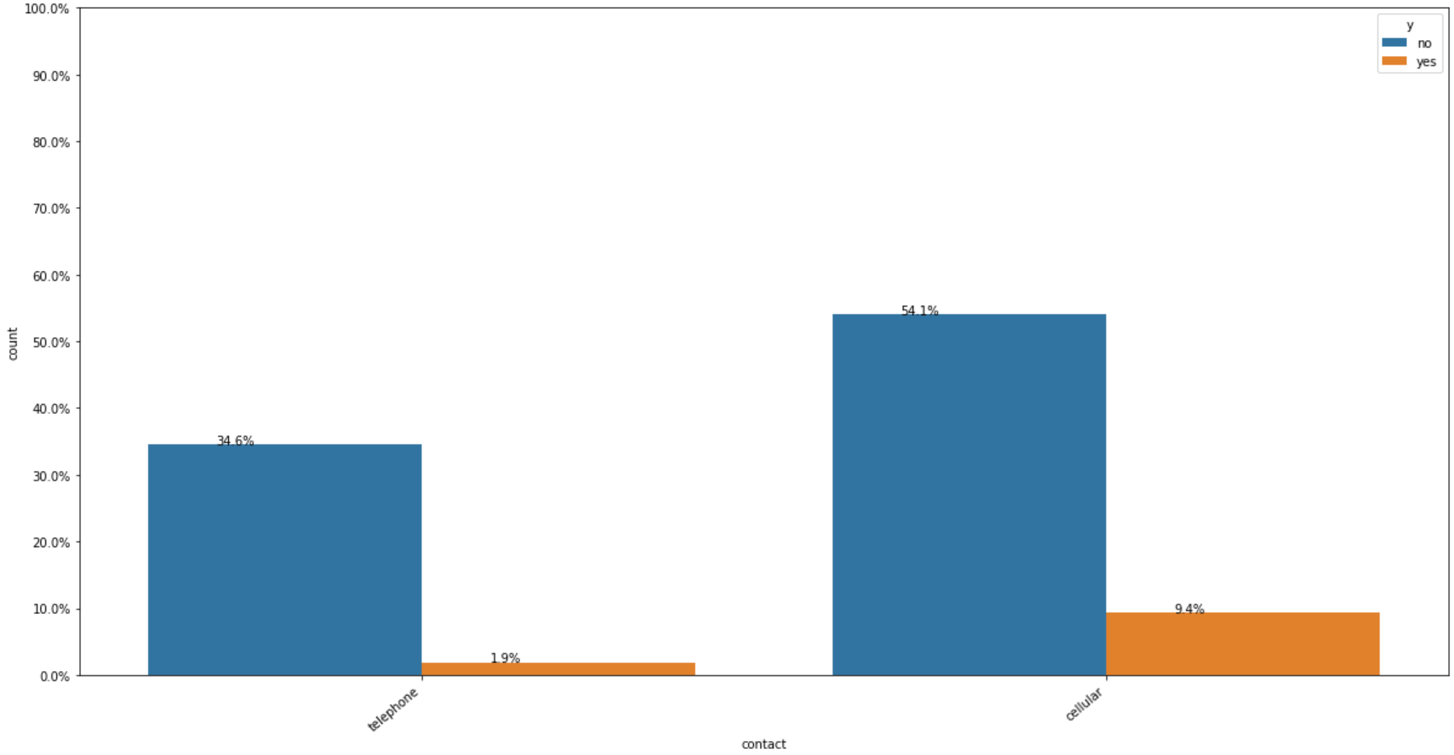
➢ Feature: contact (Categorical)

This is a categorical feature which means contact communication type, and with the values **‘cellular’ and ’telephone’.**



- As shown above, most people are contacted more in cellular than telephone.

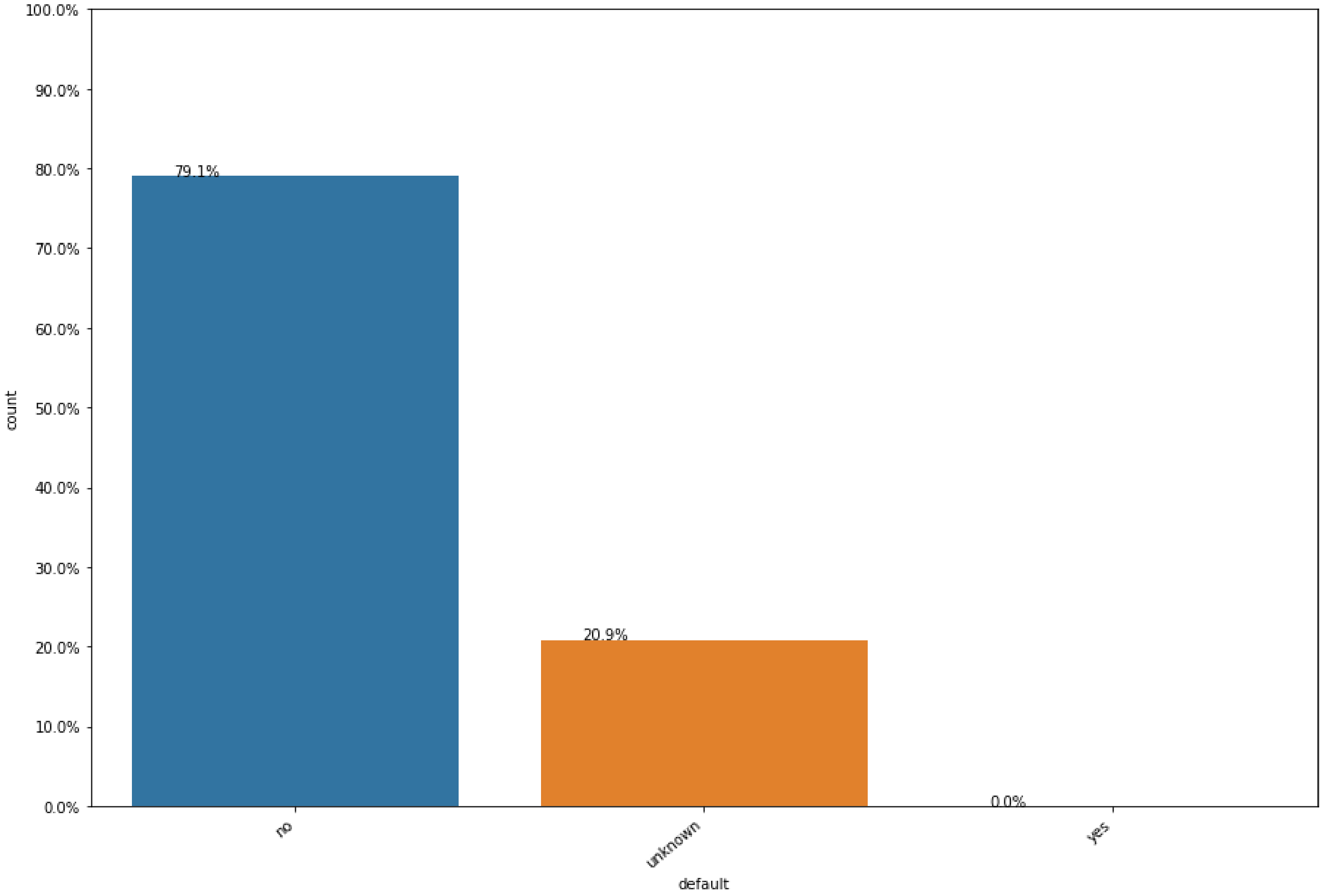
➢ Which contact type has subscribed the most for long term deposits?



- People with contact type cellular have subscribed more for long term deposits than telephone.

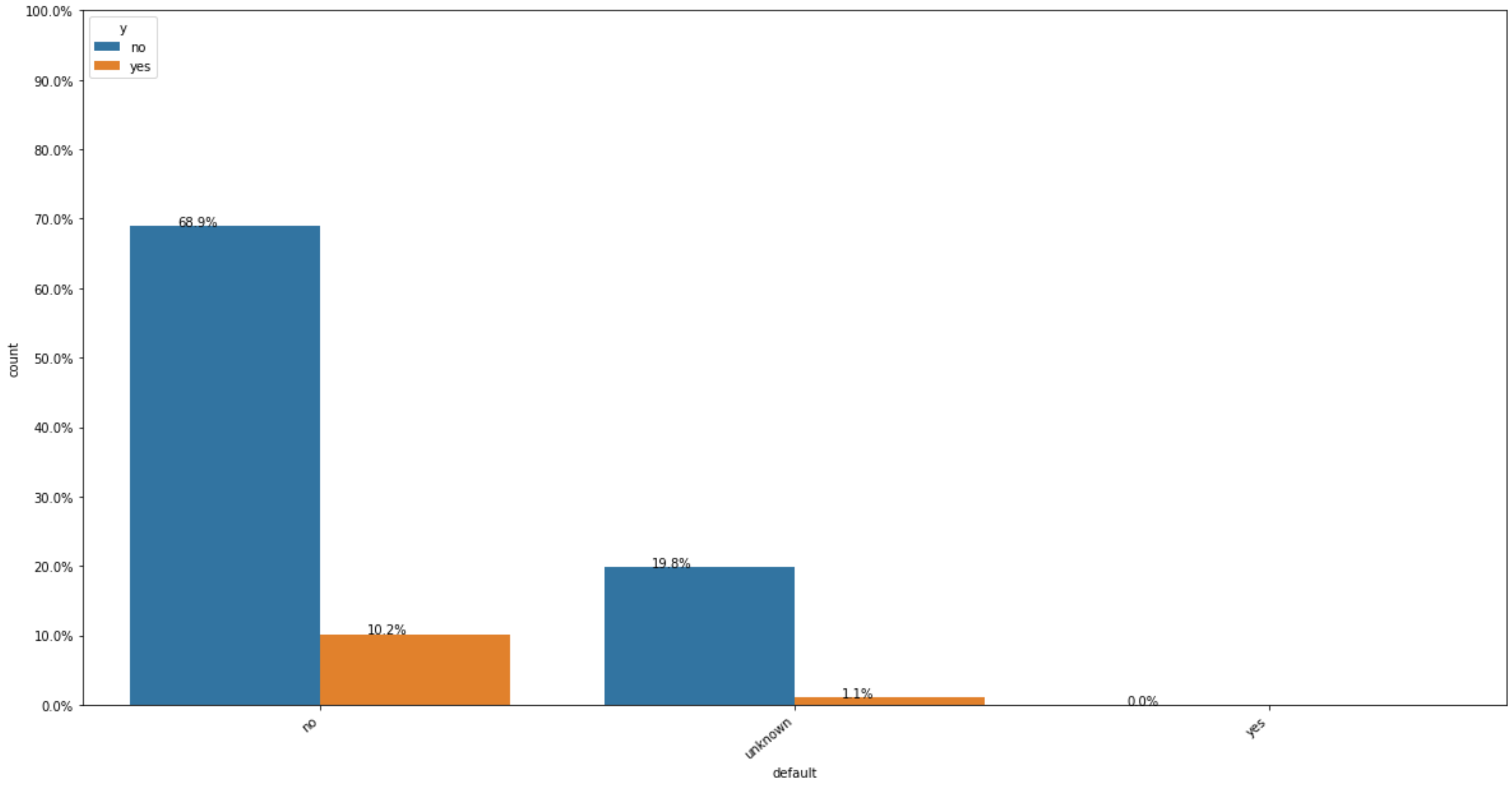
➢ Feature: default (categorical)

This is a categorical feature which means **"has credit in default",** and with the values **"yes"** and **"no"** and **"unknown".**



- As shown above, we can clearly see that the people with default status as ‘no’ are the most who have been contacted by the bank for the deposits.
- People with default status ‘yes’ have not been contacted by the bank at all.
- While very few people with unknown default status have been contacted by the bank.

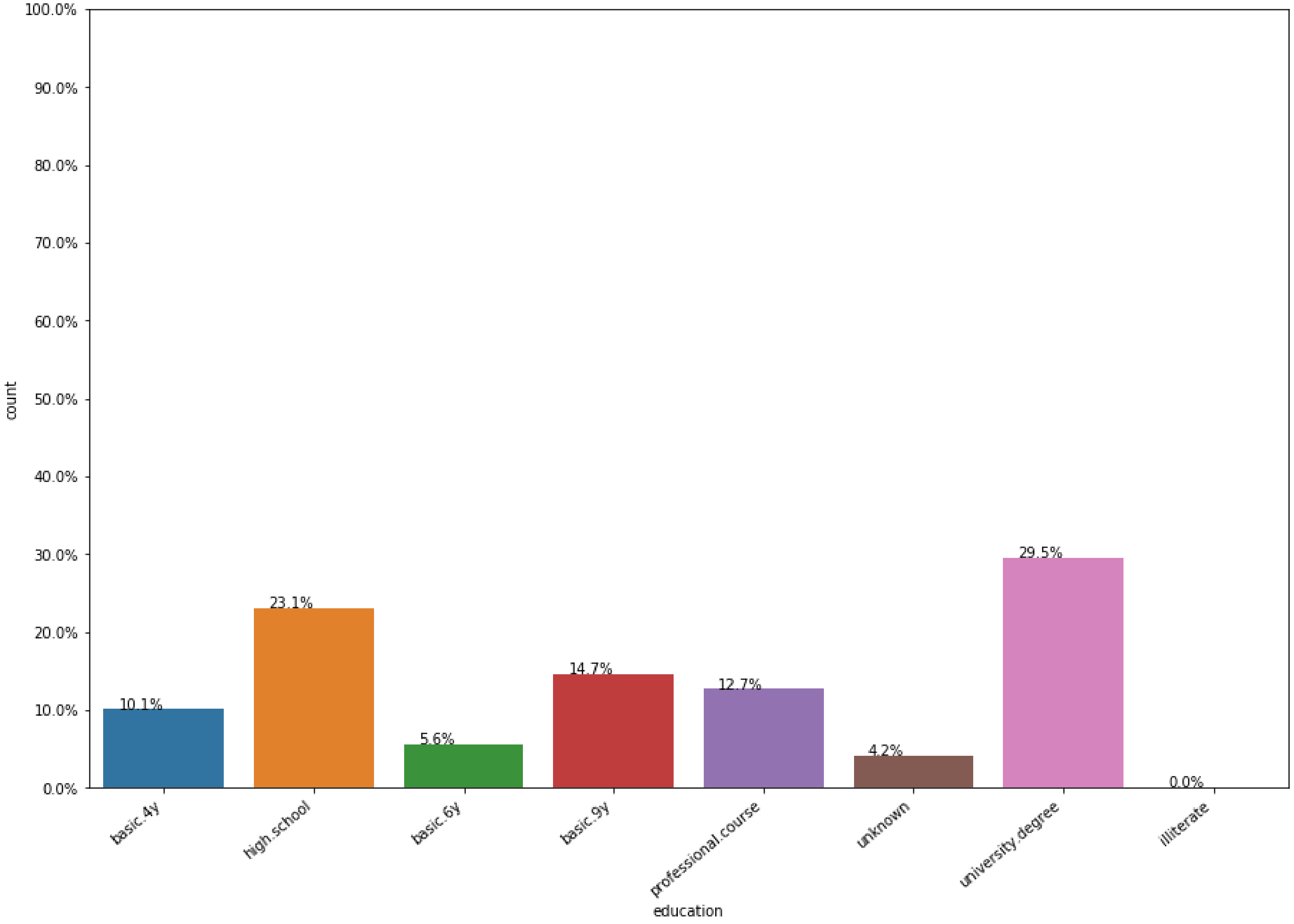
- **Do people who have or who don't have credit in default subscribe the most for long term deposits?**



- From the plot above we can observe that people with default status as no are the most one's who have and have not subscribed for bank deposits.

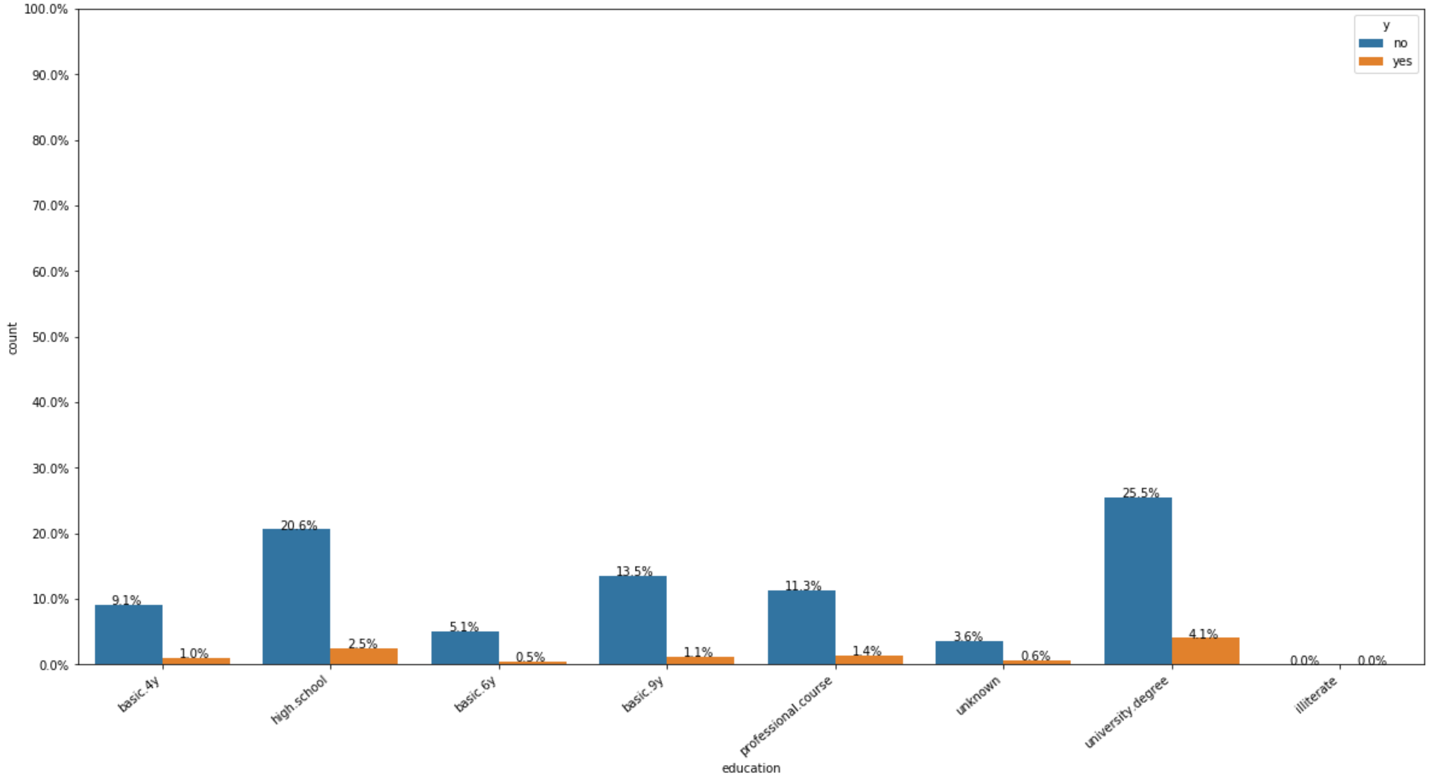
- **Feature: Education**

This is a categorical feature which is the educational qualifications, and with the values **‘basic.4y’,’basic.6y’,’basic.9y’,’high.school’,’illiterate’,’professional.course’,’university. degree’ and ’unknown’.**



- As shown above, people contacted by the bank with university degree as their educational qualification are more than the people with any other educational qualification.

➢ Do people who have completed their university degrees tend to subscribe the most for long term deposits?

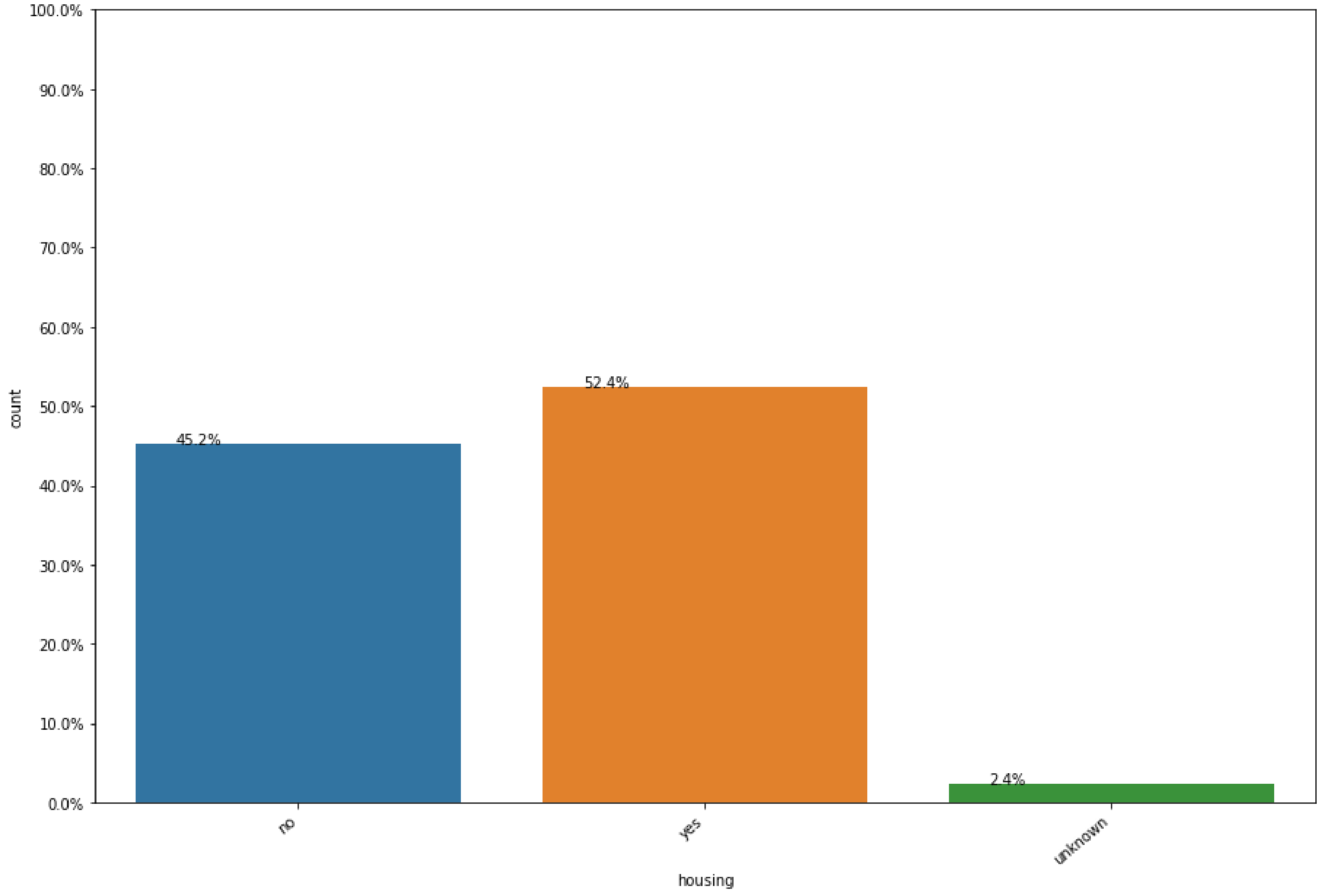


- As shown above, people with university degree as education qualification are the most who have subscribed for the deposits.
- They are also the most who have not subscribed for deposits.

➢ Feature: housing (Categorical)

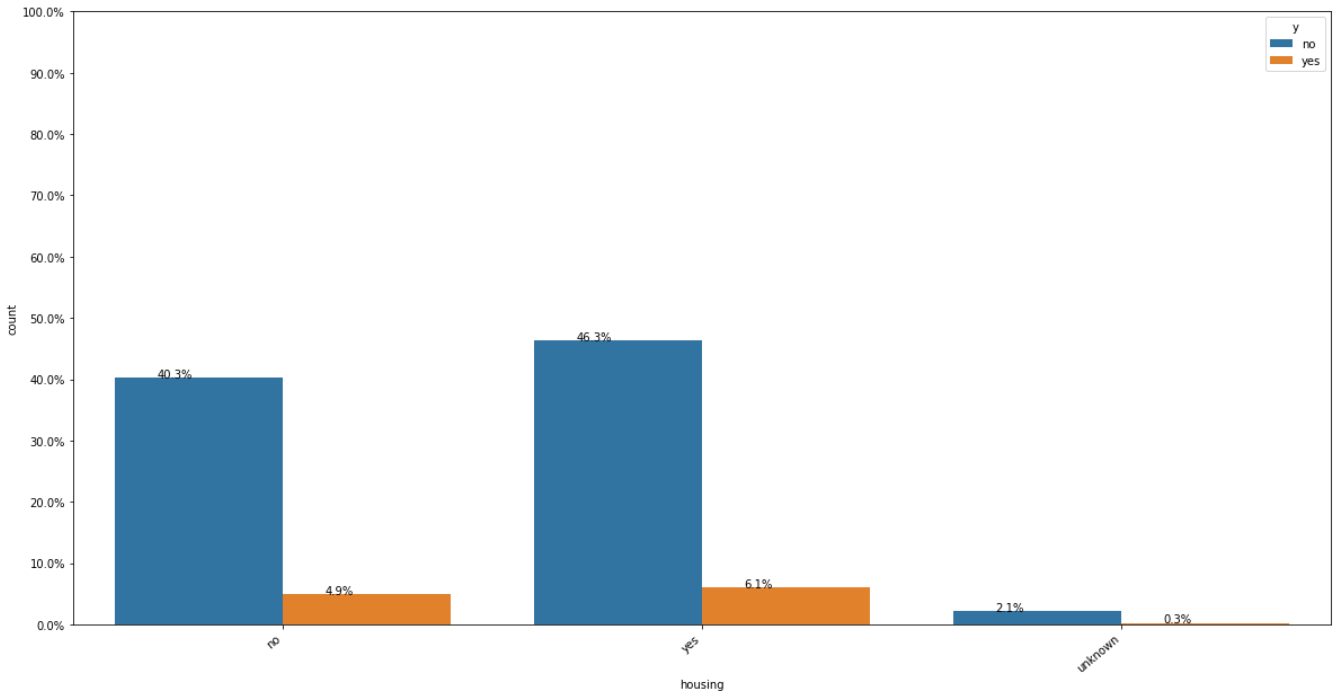
Has housing loan?

Values are **‘no’, ‘yes’** and **’unknown’.**



- People who have housing loan have been contacted the most by the bank.
- People who have no housing loan has also been contacted pretty much, and people who have status unknown have been least contacted.

➢ Do people with or without a housing loan subscribed the most for long term deposits?

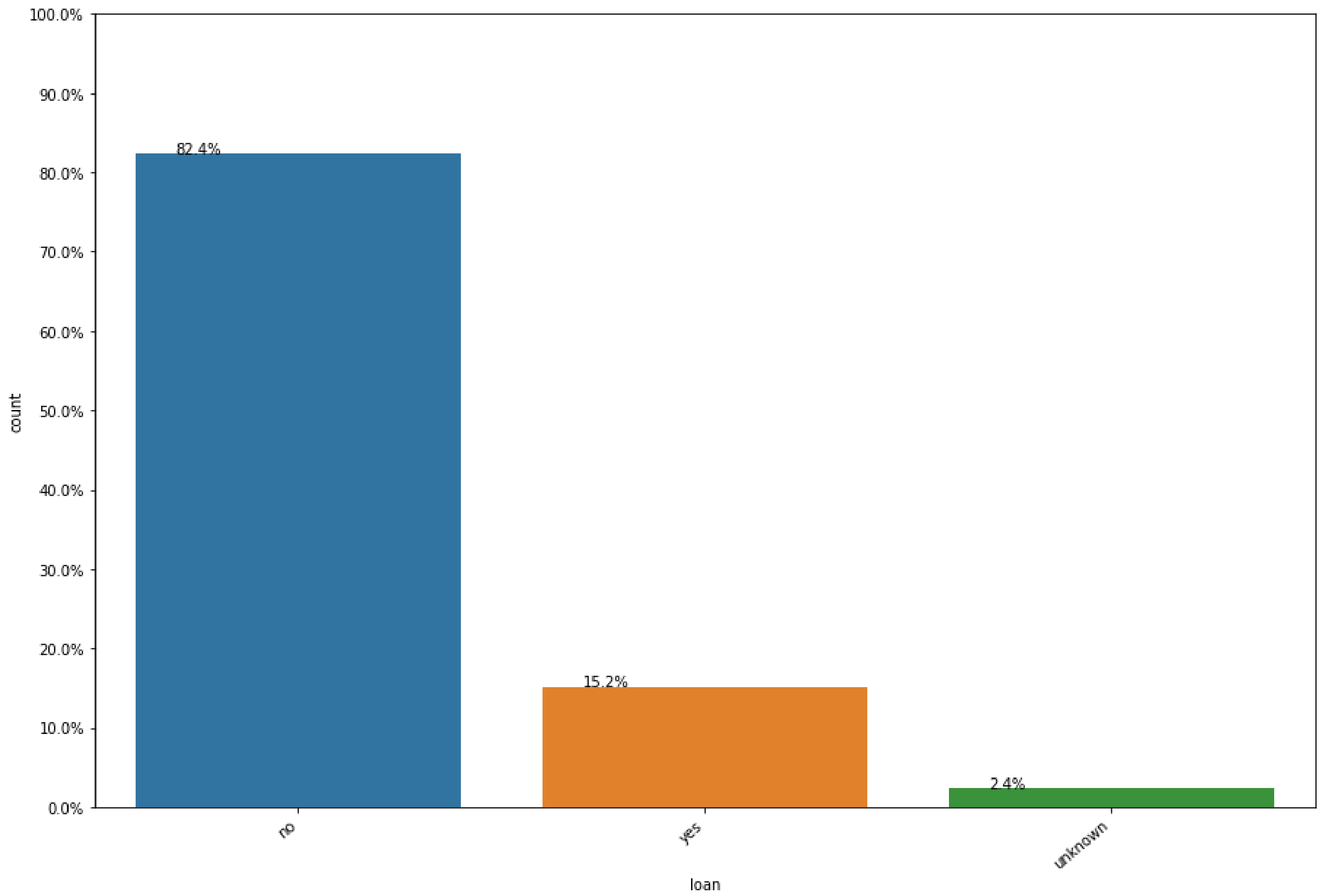


- As shown above, people who have a housing loan have subscribed the most for long term deposits, followed by the ones who does not have housing loans.
- People with a housing loan are also the most ones who have not subscribed for the deposits.

➢ Feature: loan (Categorical)

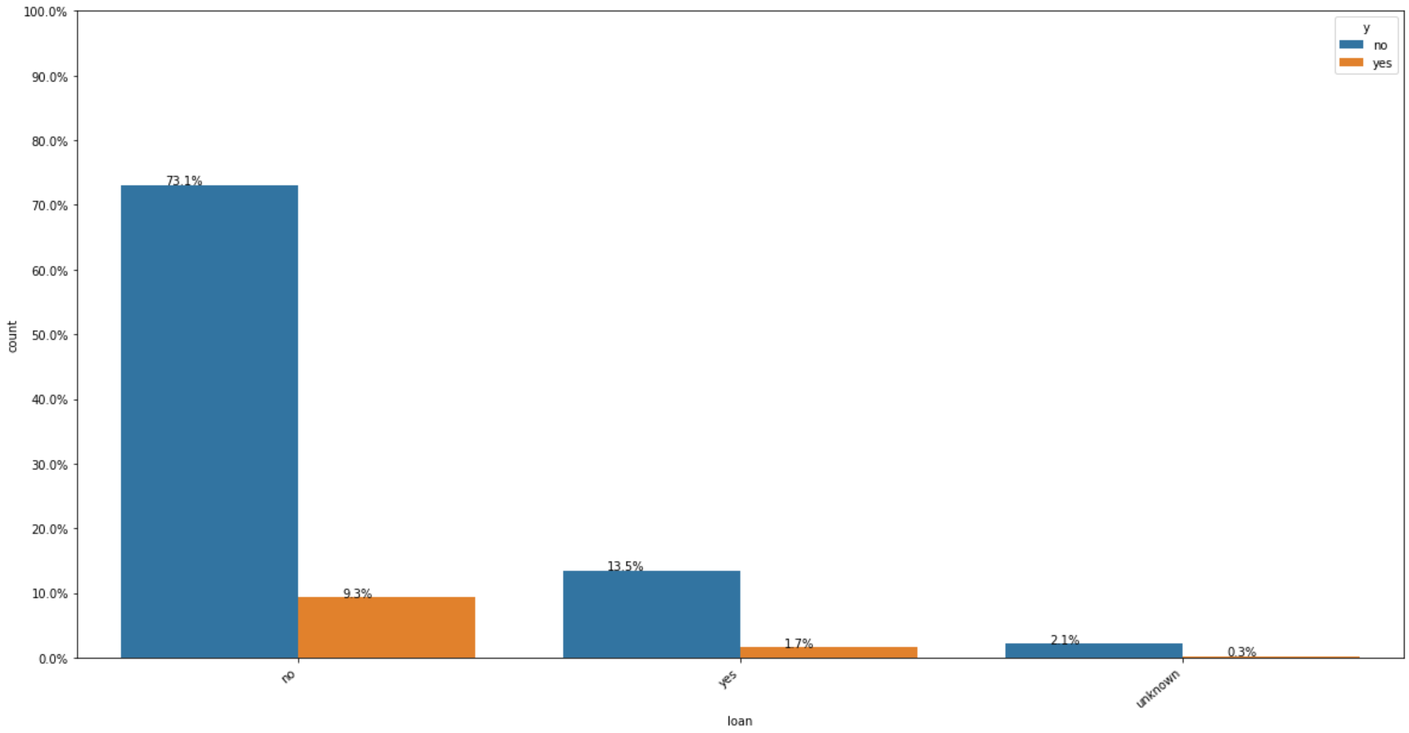
Has personal loan?

Values are **'no','yes'** and **'unknown'.**



- As shown above, people who do not have loans have been most contacted for longer term deposits than the ones who have loans.

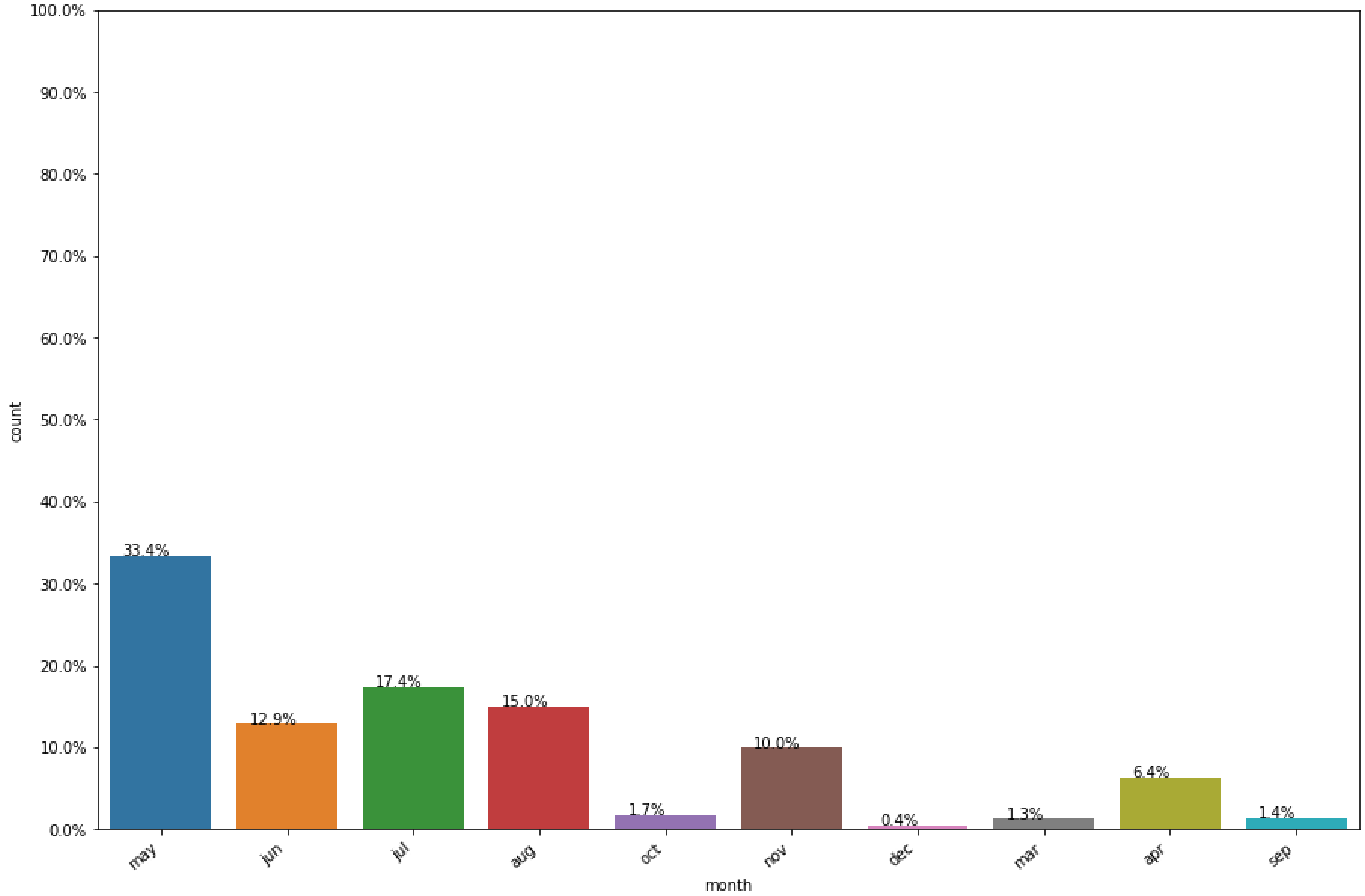
➢ Do people with or without loans subscribed the most for long term deposits?



- People with no personal loan have subscribed the most for long term deposits.
- They are also the ones who have not subscribed for long term deposits.

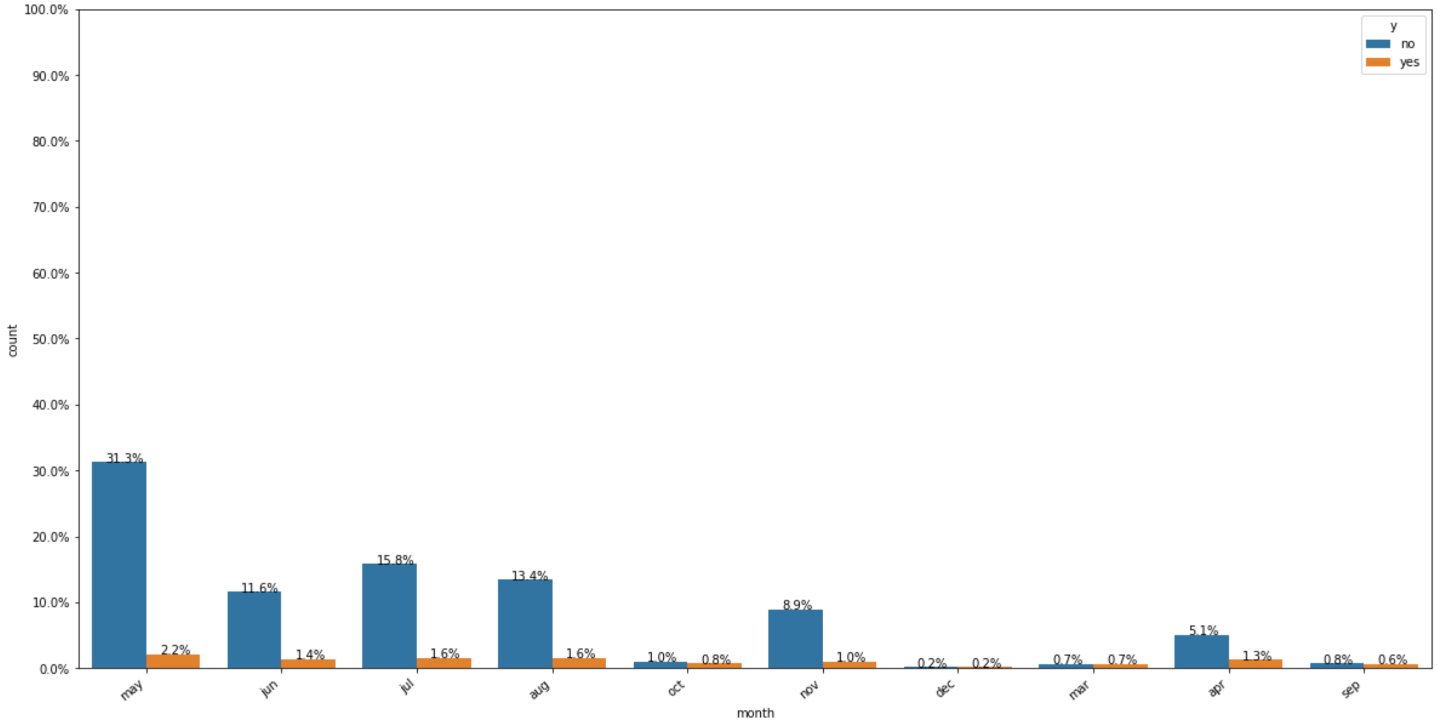
➢ Feature: month (Categorical)

Last contact month of year, values **'jan', 'feb', 'mar',** …, **'nov'** and **'dec'.**



- People are being contacted the most in the month of May than any other months.
- It is followed by July, August, June.
- Very few people have been contacted in the month of December.
- People have not been contacted in the month of January and February.

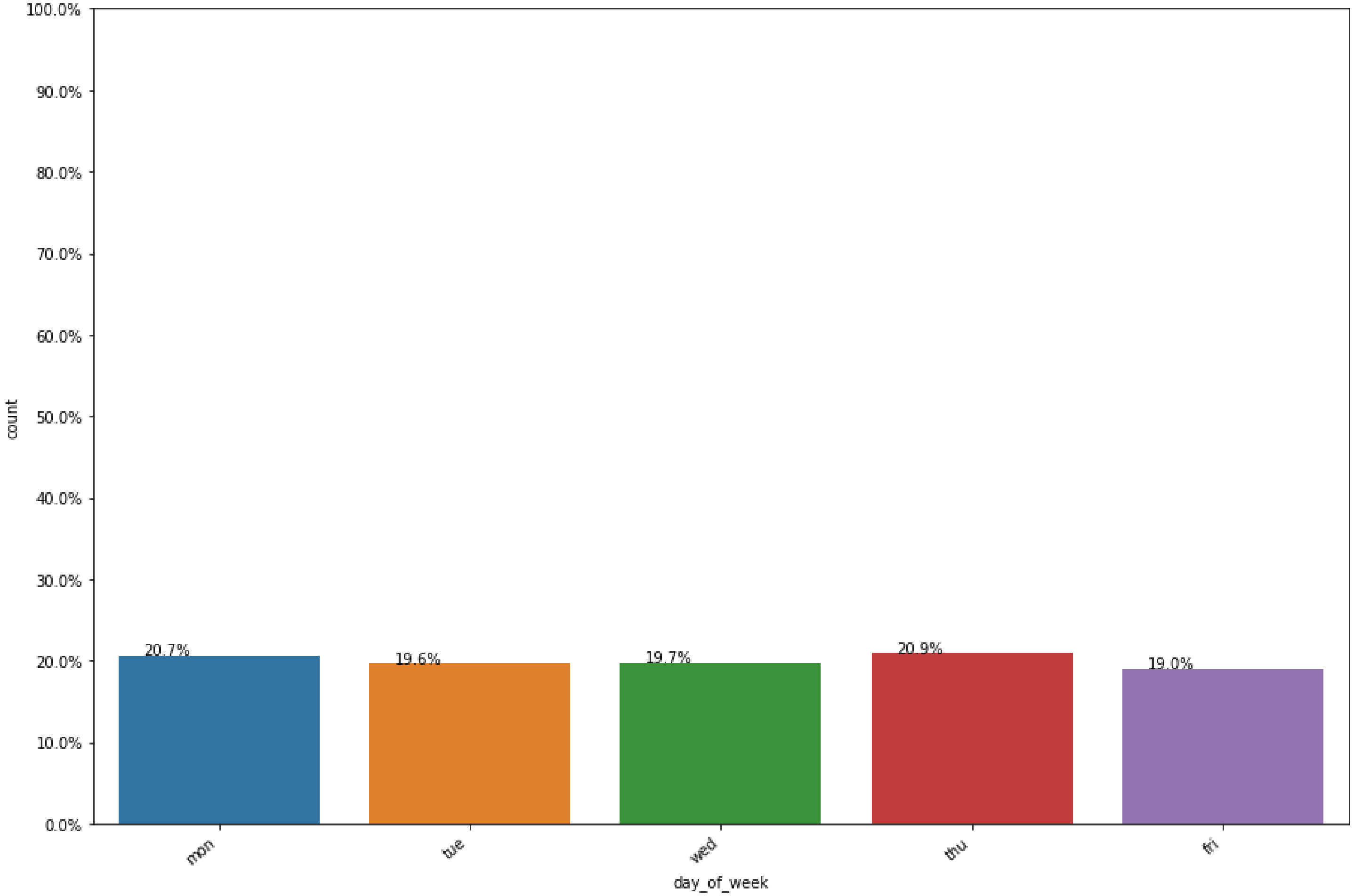
➢ In which month do people have higher chances to subscribe for longer term deposits?



- People who have been contacted in May have higher chances to subscribe for longer term deposits but have also higher chances for not subscribing the long-term deposits.
- Very few people are contacted in the month of December, March, September, and October and have almost equal chances for subscribing the deposits or not.

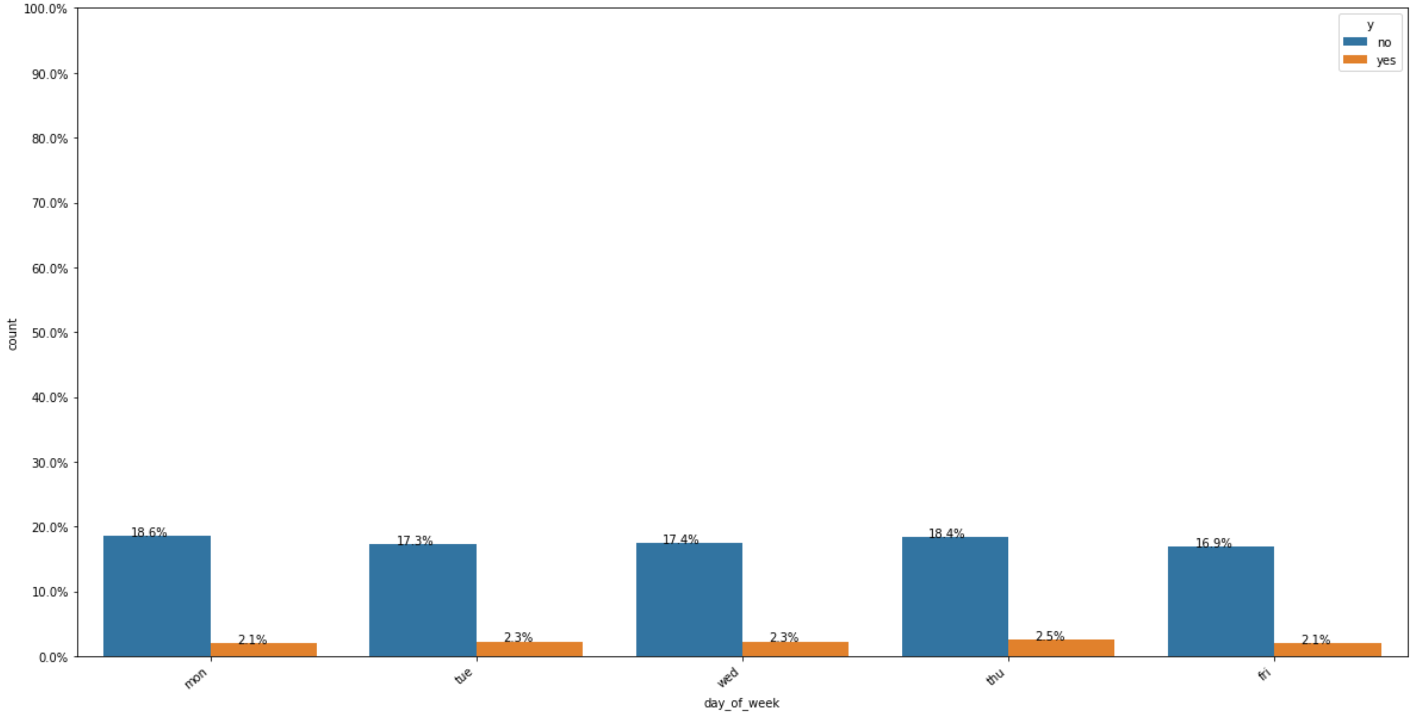
➢ Feature: day_of_week (Categorical)

Last contact day of the week, and the values are **‘mon’,’tue’,’wed’,’thu’**and **’fri’.**


100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
count
20.7%
19.6%
19.7%
20.9%
19.0%
mon
tue
wed
thu
fri
day_of_week

- From the plot above we can see that people are contacted from Monday to Friday but not on Saturday and Sunday.

Does in all days people have equal chances for subscribing or not subscribing for the long-term deposits?

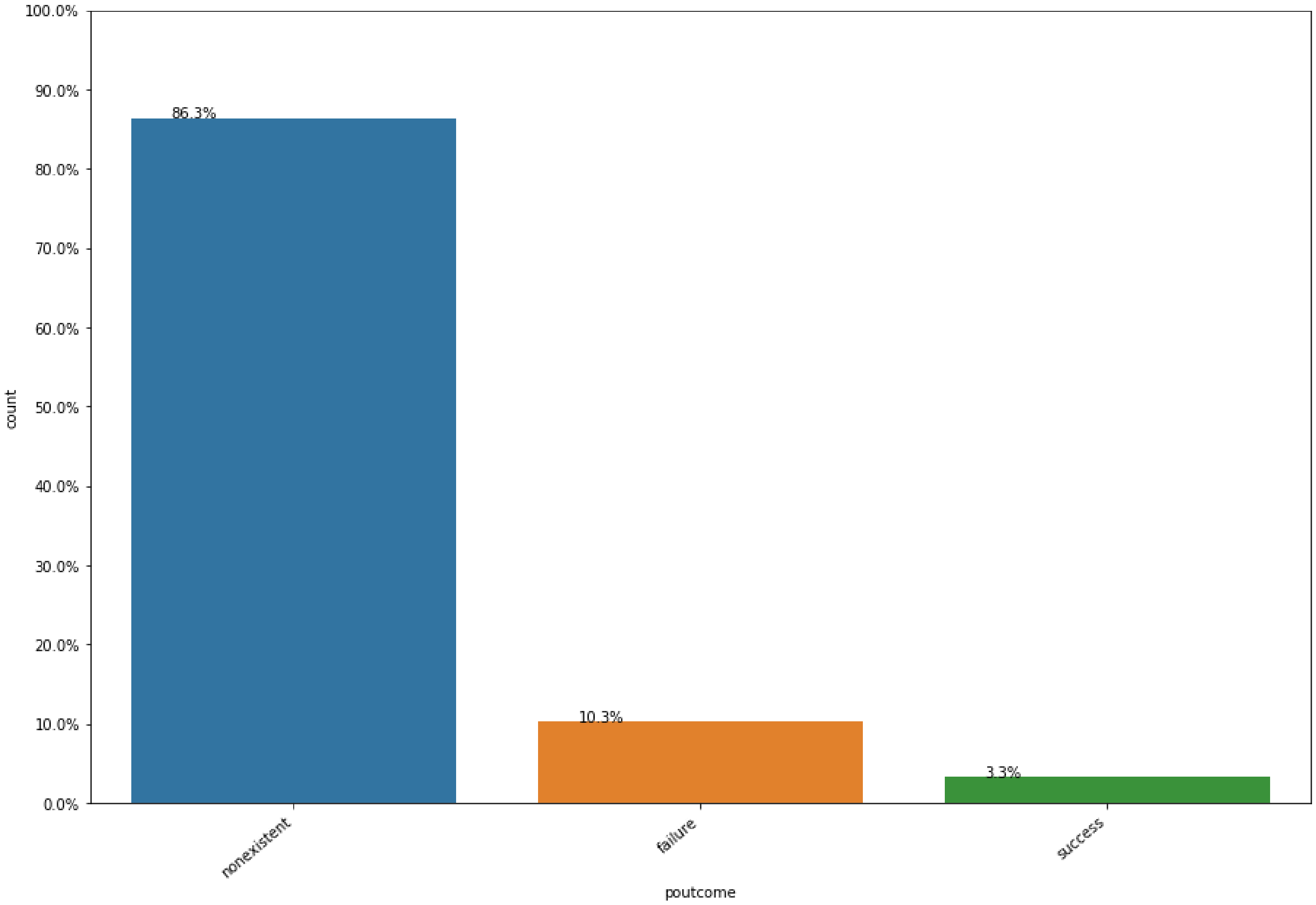


- As shown above, in all the days they have equal chances for subscribing and not subscribing the term deposits.

- Day_of_week may not be very helpful in predicting whether the customer will subscribe for long term deposits or not.

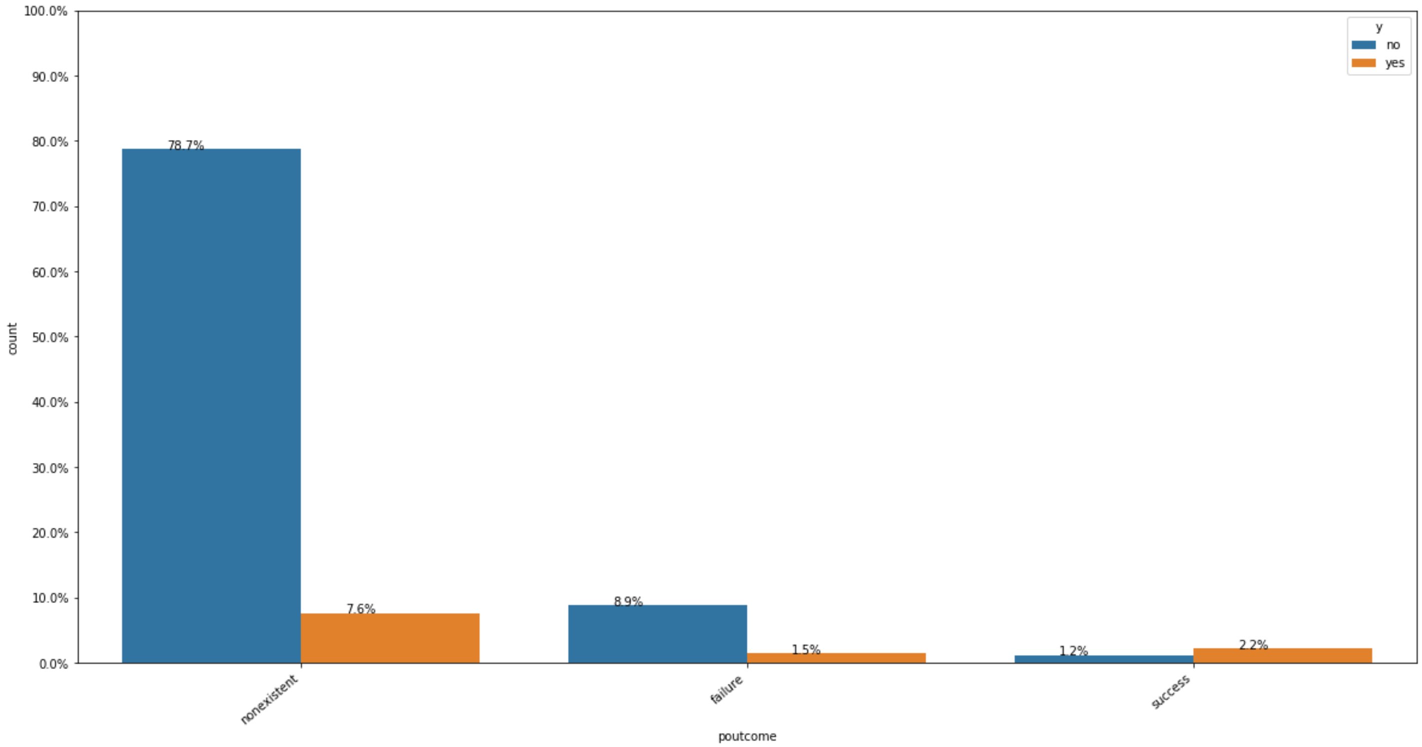
➢ Feature: poutcome (Categorical)

Outcome of the previous marketing campaign, and the values are ‘**failure’**,**’nonexistent**’ and **’success’.**



- From the above plot it is evident that majority of outcome of previous campaigns are nonexistent.
- Very few people from previous marketing campaigns have subscribed for long term deposits.

➢ Which previous outcome has subscribed the most?



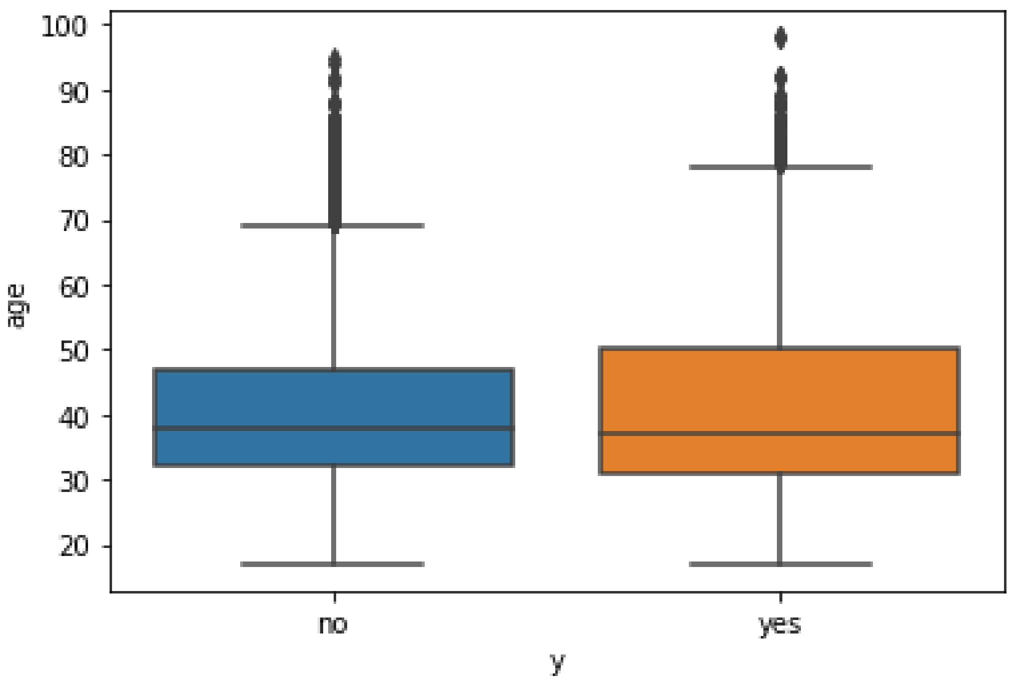
- From the above plot, people whose previous outcome is non-existent has subscribed the most than any other group of people belonging to previous outcome.
- It is also clear that people belonging to success category of previous outcome has turned down for longer deposits.

EDA

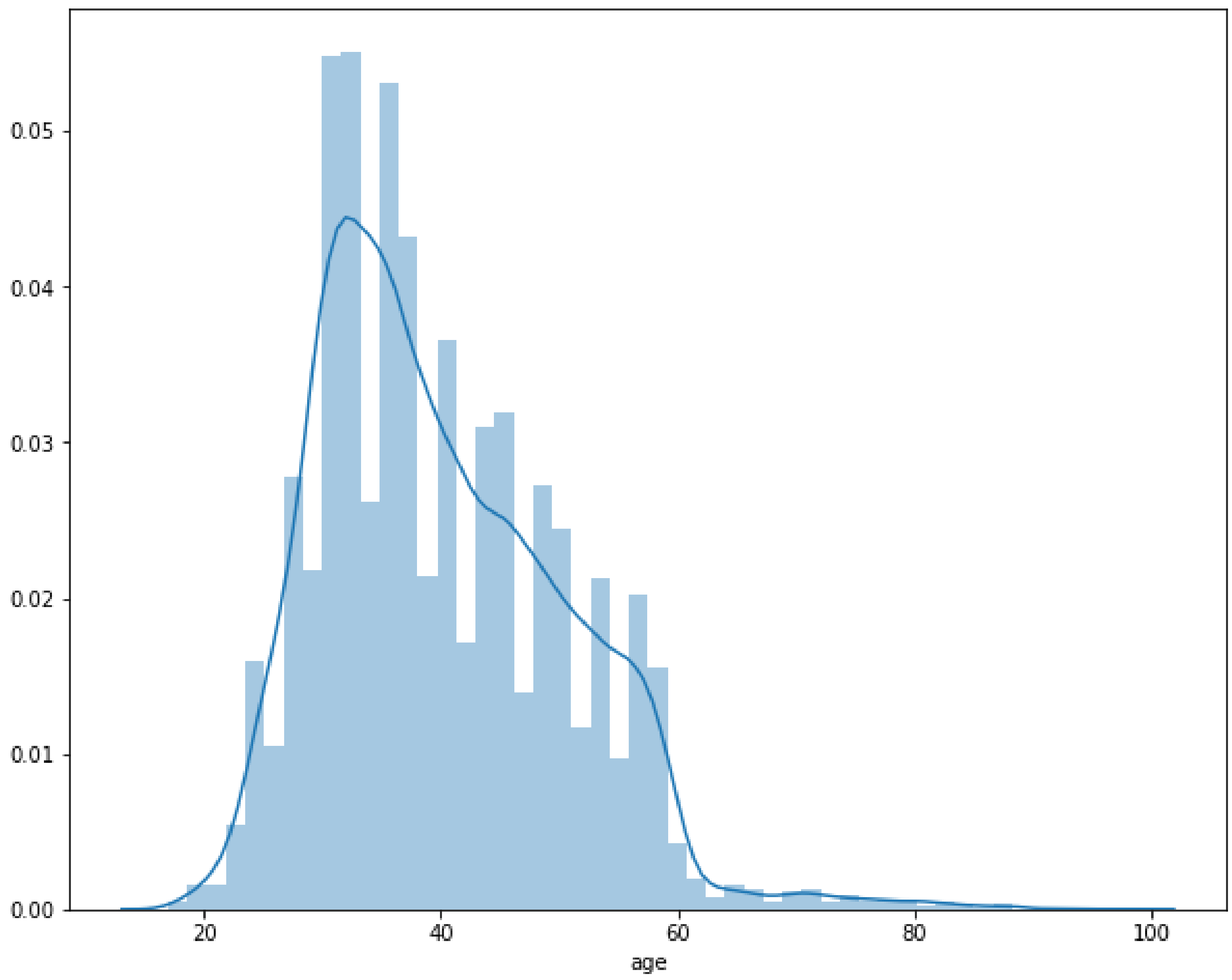
- **Numerical Variables:**

Now I perform some exploratory data analysis on the numerical variables.

- **Feature: age (Numeric)**

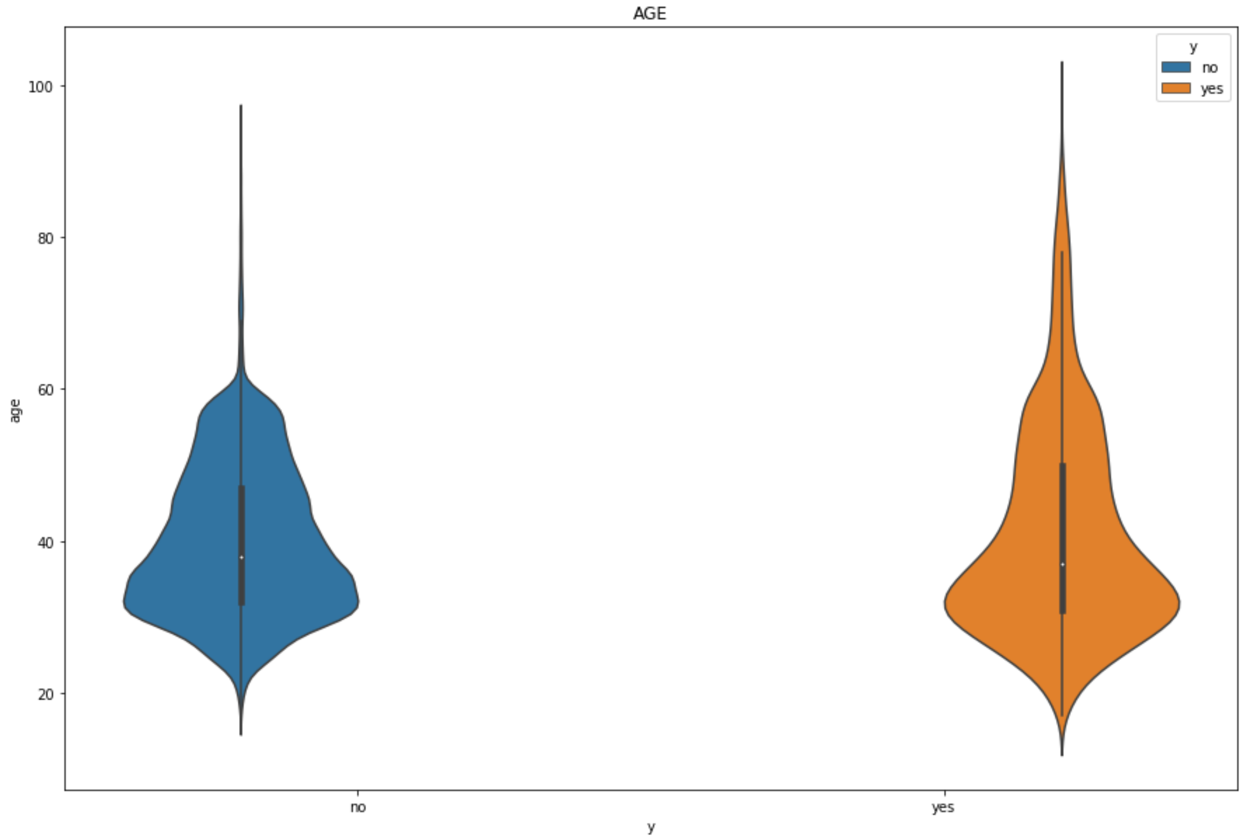


- From the boxplot above we know that for both the customers that subscribed or didn't subscribed a term deposit, has a median age of around 38-40, and the boxplot for both the classes overlaps quite a lot, which means that age isn't necessarily a good indicator for which customer will subscribe and which customer will not.

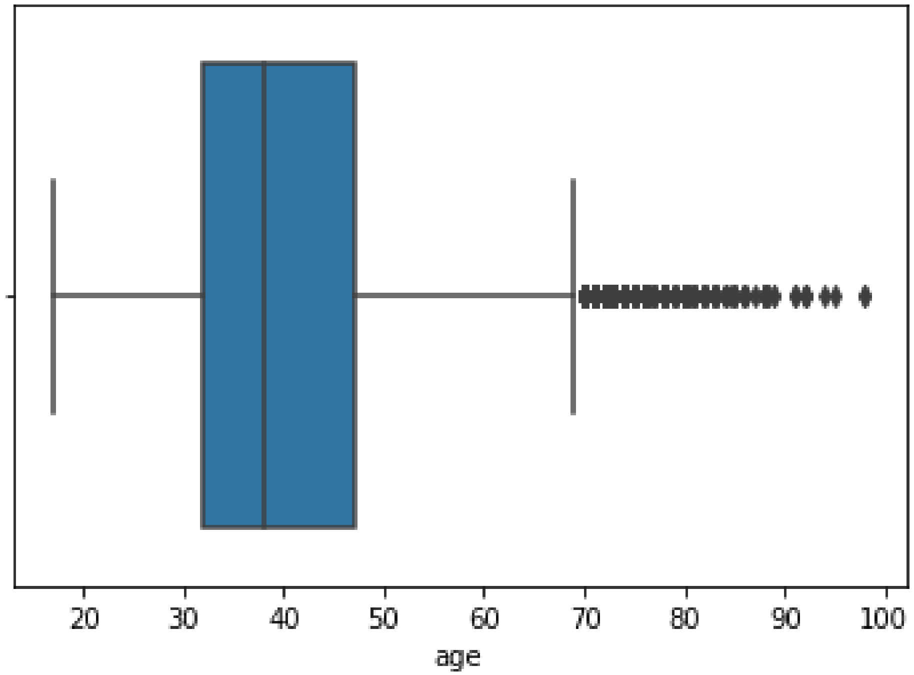


- From the plot above it is a right skewed graph and there is an evidence of outliers after the age of 60.
- We also see in the distribution that most of the customers are in the age range of 30-40.

However, we can use other plots to check for outliers, Violin plot and box plot.



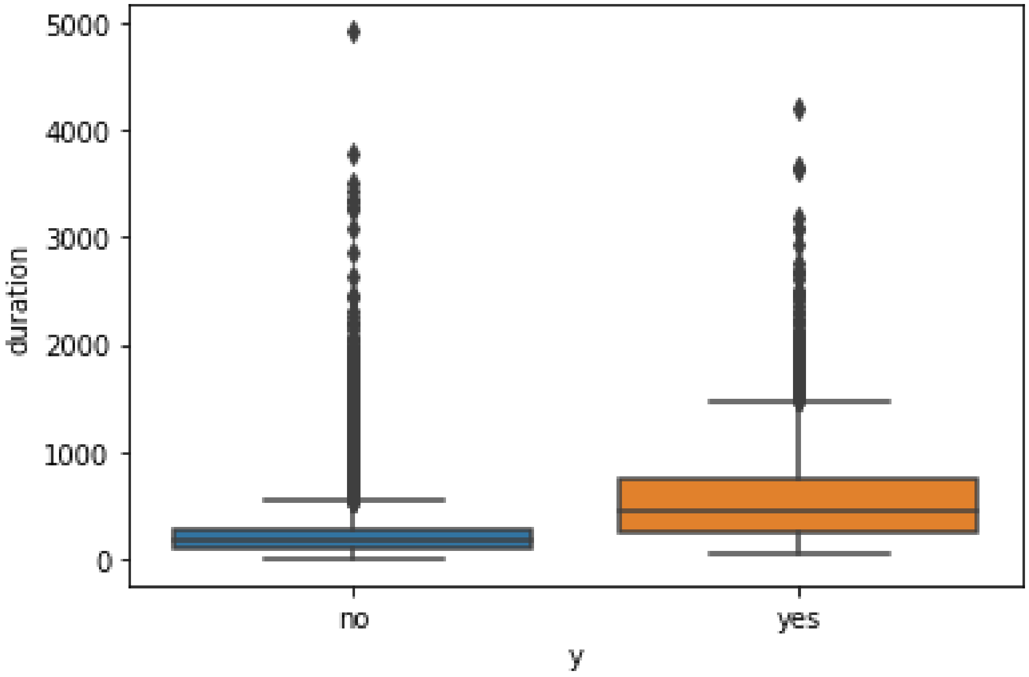
- Using the Violin plot as shown above, it is clearly visible that there are outliers present for both the class.
- In No class, outliers are present above age 70 and for Yes class, outliers are present above age 75.
- Median for No class is around 40 which is same for Yes class.
- Also, it is visible that IQR range is almost overlapping so age might not be very helpful in predicting class label.



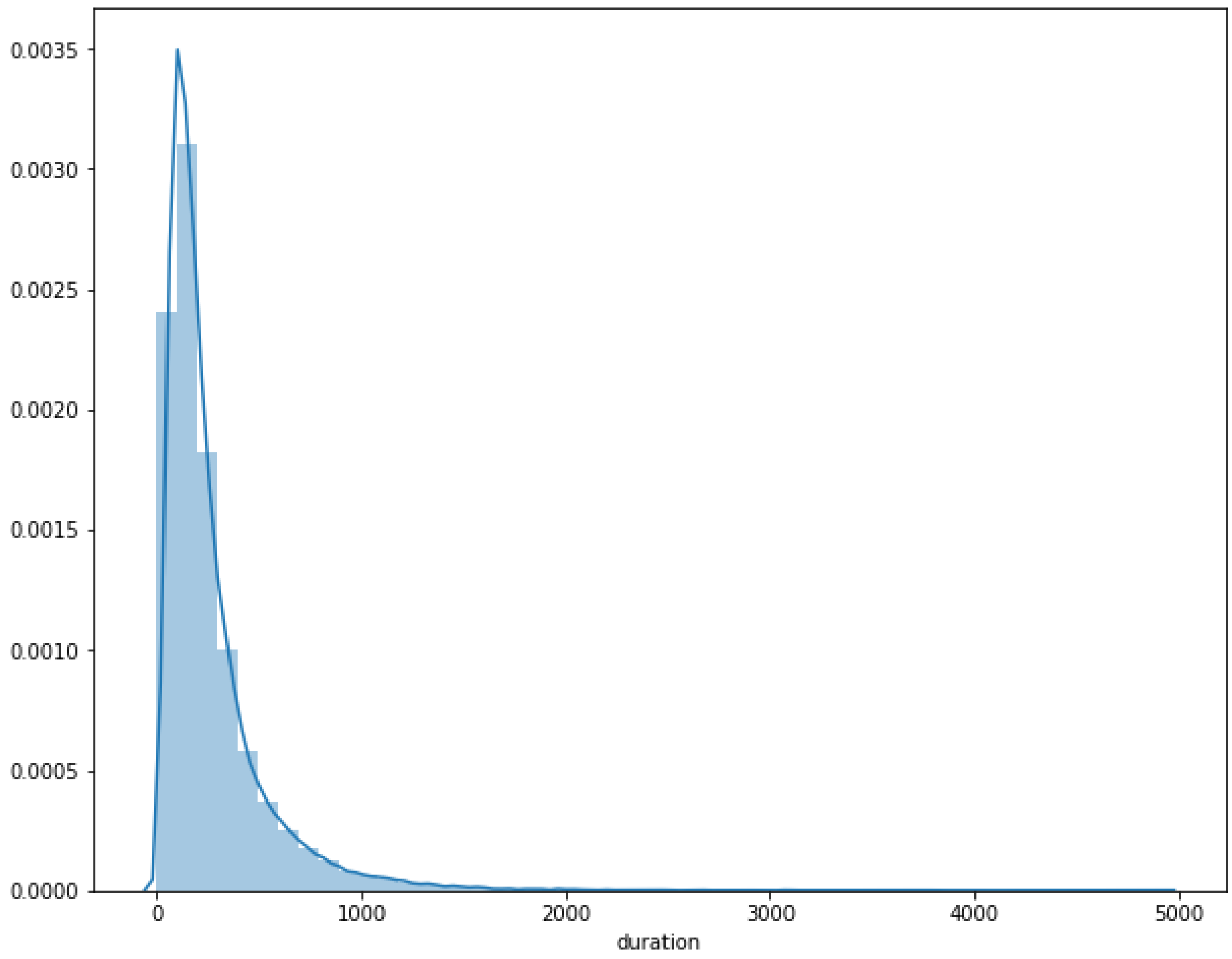
- Using the box plot as shown above, we can say that outliers are present after the age of 70 years.

➢ Feature: duration (numeric)

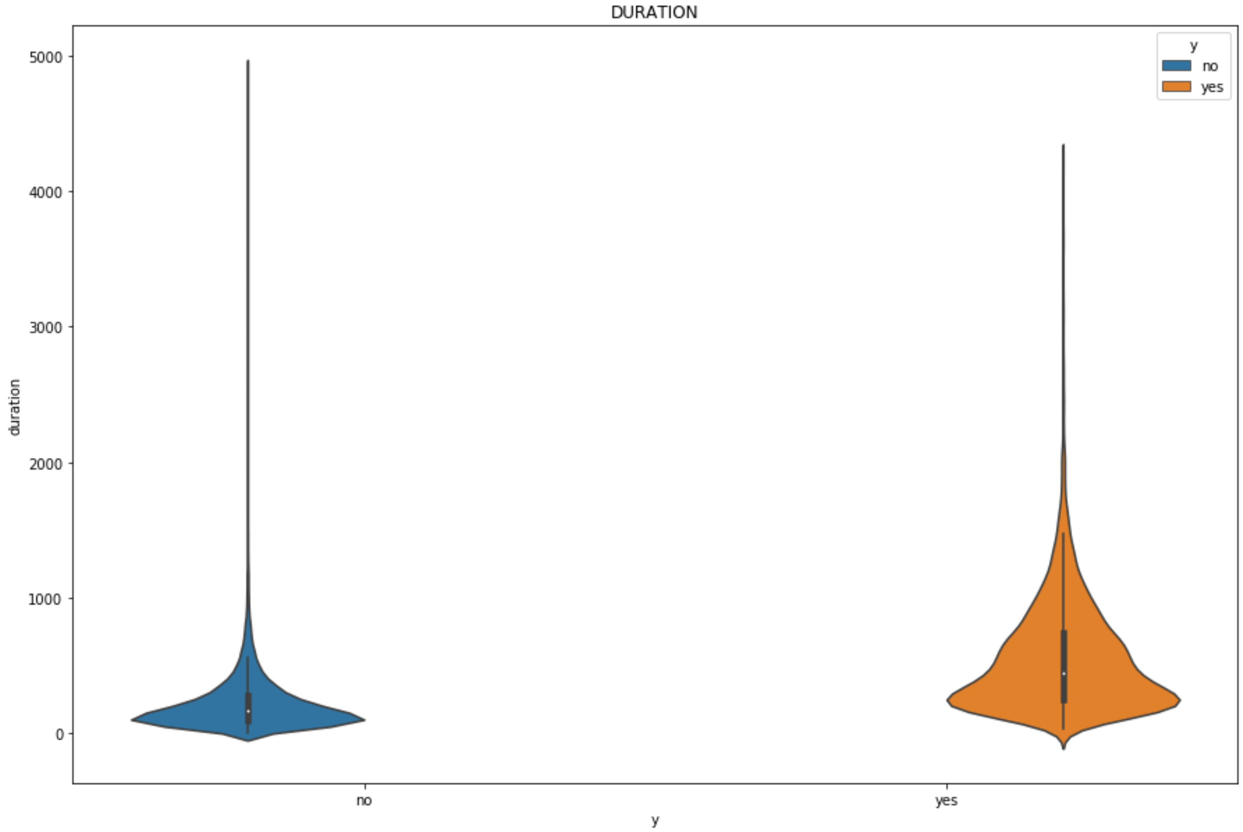
Last contact duration in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.



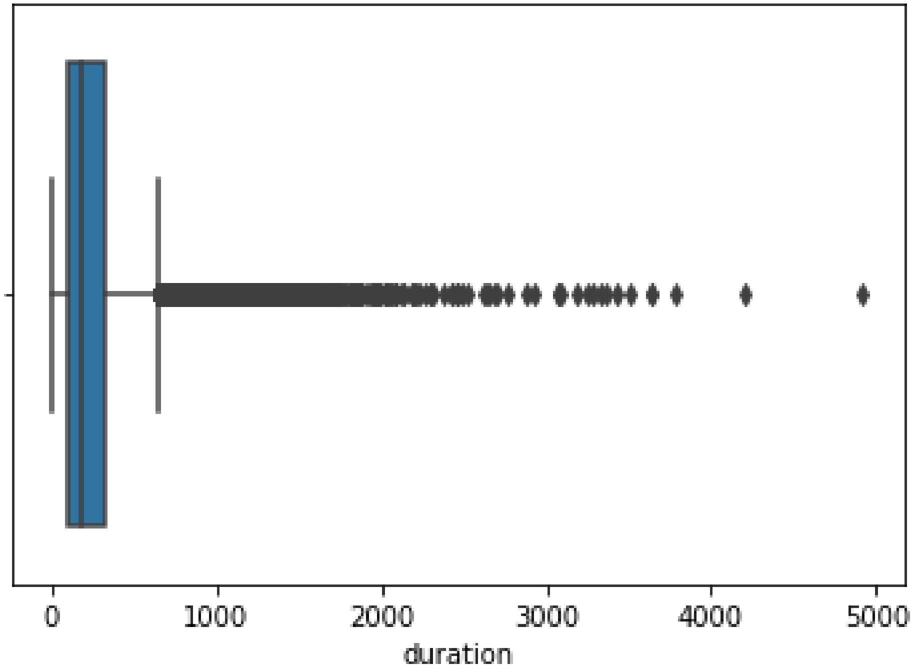
- From the plot above, the duration (last contact duration) of a customer can be useful for predicting the target variable.
- It is expected because it is already mentioned in the data overview that this field highly affects the target variable and should only be used for benchmark purposes.



- From the plot above it is clearly evident that from duration around 1500 onward outliers, are present and it is right skewed data.
- In this distribution above we find where most the values are very low and very few have high values.



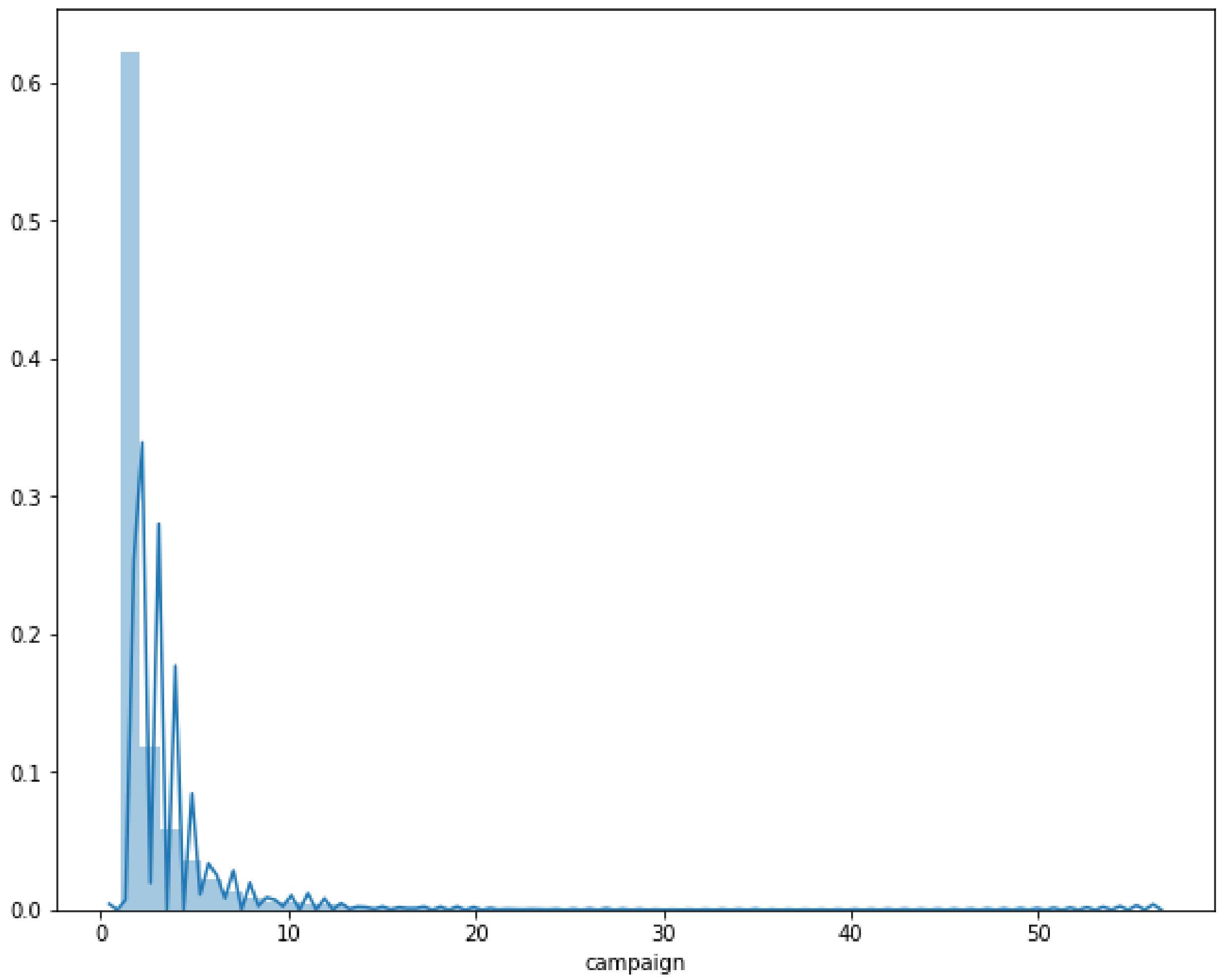
- Using the Violin plot as shown above, any duration of call with class labels as no, more than 1000 are considered as outliers while with class labels as yes, more than 1500 would be considered as outliers.



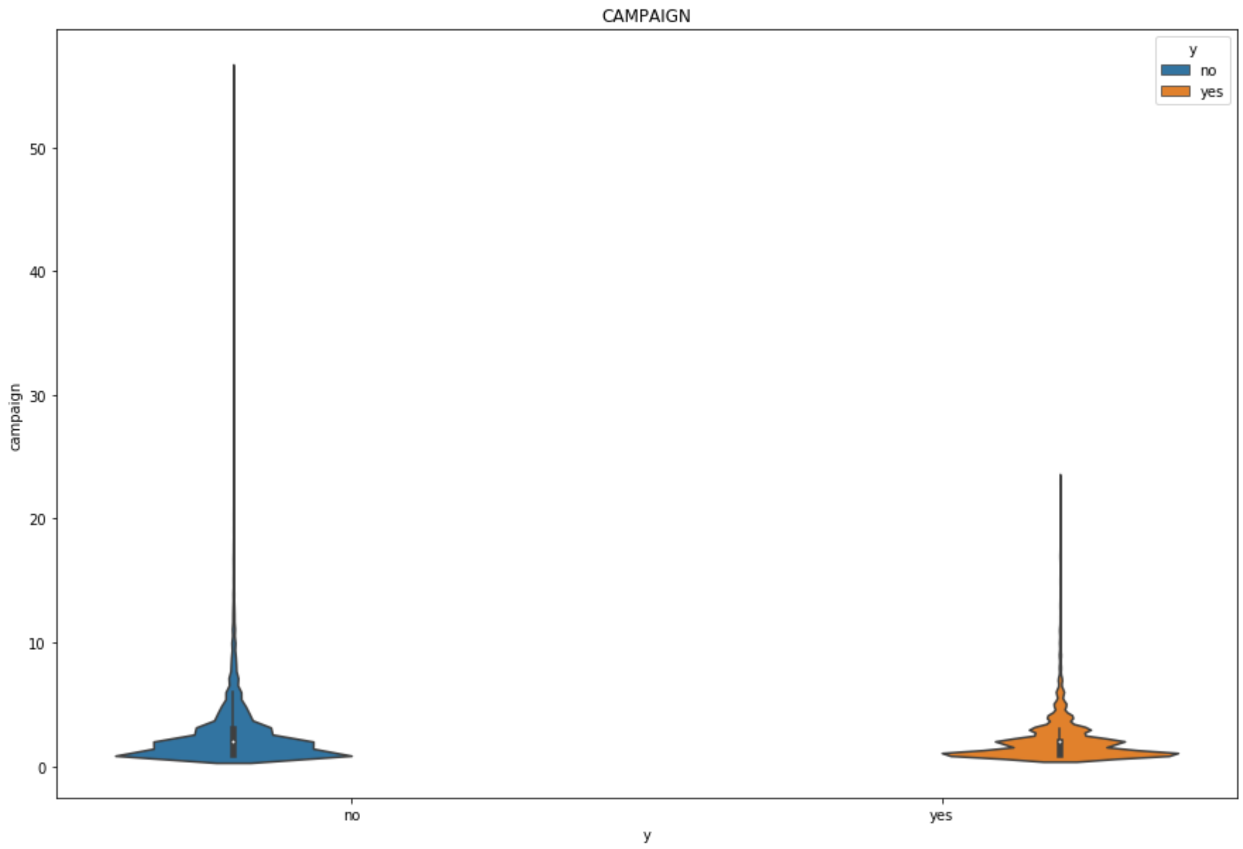
- Using the box plot as shown above, we can say that outliers are present after the duration of calls more than 1000.

➢ Feature: campaign (numeric)

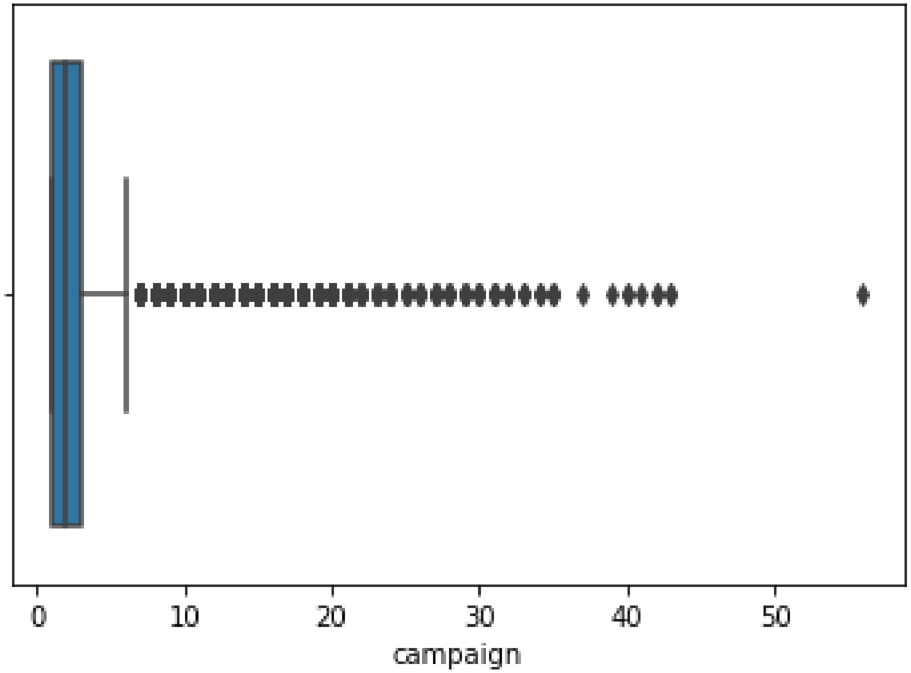
campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact).



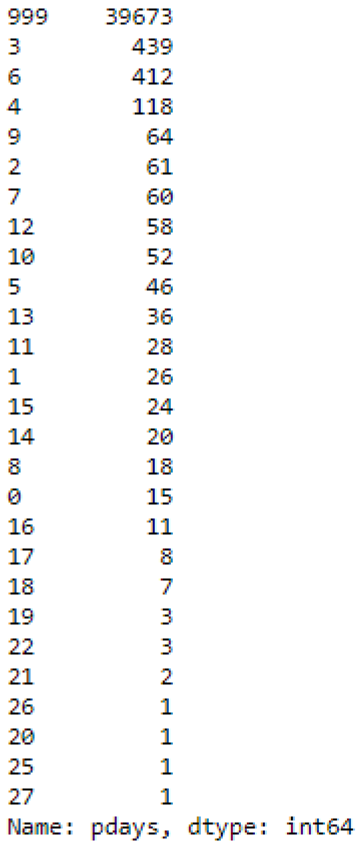
- From the plot above, outliers might be present after the number of campaigns 10.



- Outliers are present when number of campaigns are more than 10 irrespective of any class labels.

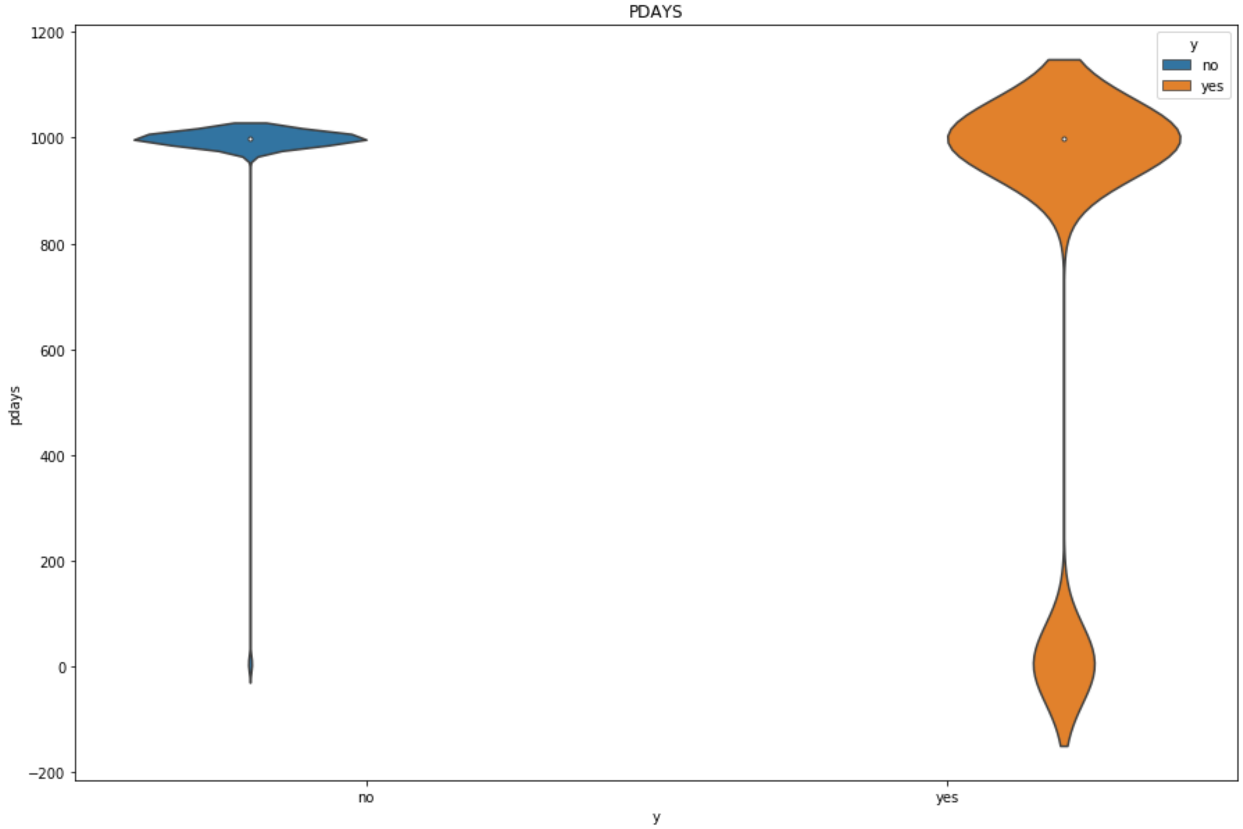


- Using the box plot as shown above, we can see those campaigns more than 10 are considered as outliers.

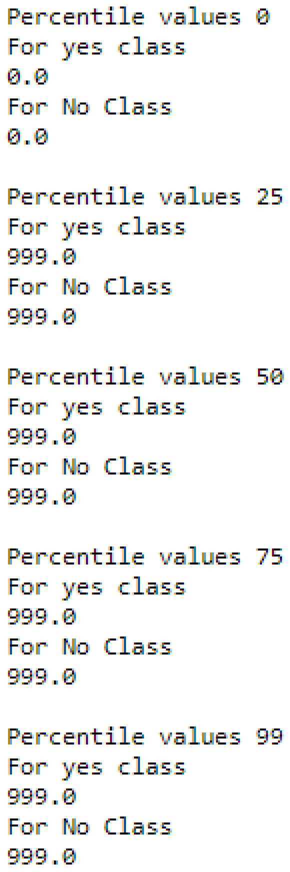


```
999    39673
3        439
6        412
4        118
9         64
2         61
7         60
12        58
10        52
5         46
13        36
11        28
1         26
15        24
14        20
8         18
0         15
16        11
17         8
18         7
19         3
22         3
21         2
26         1
20         1
25         1
27         1
Name: pdays, dtype: int64
```

- From the data 999 means the person has not been contacted before.
- It is very evident that huge number of people have not been contacted.



- From the plot above it is visible that irrespective of class labels, mostly people have not been contacted by the bank. Very people have been contacted by the bank and number of days passed for previous campaign is between 0–100.
- It means we either must compute pdays or drop the pdays depends on the percentage of values.
- Also, it is not very clear but the IQR range for the both the classes are overlapping.
- Let's try to get the 25,50,75 percentile for this feature.

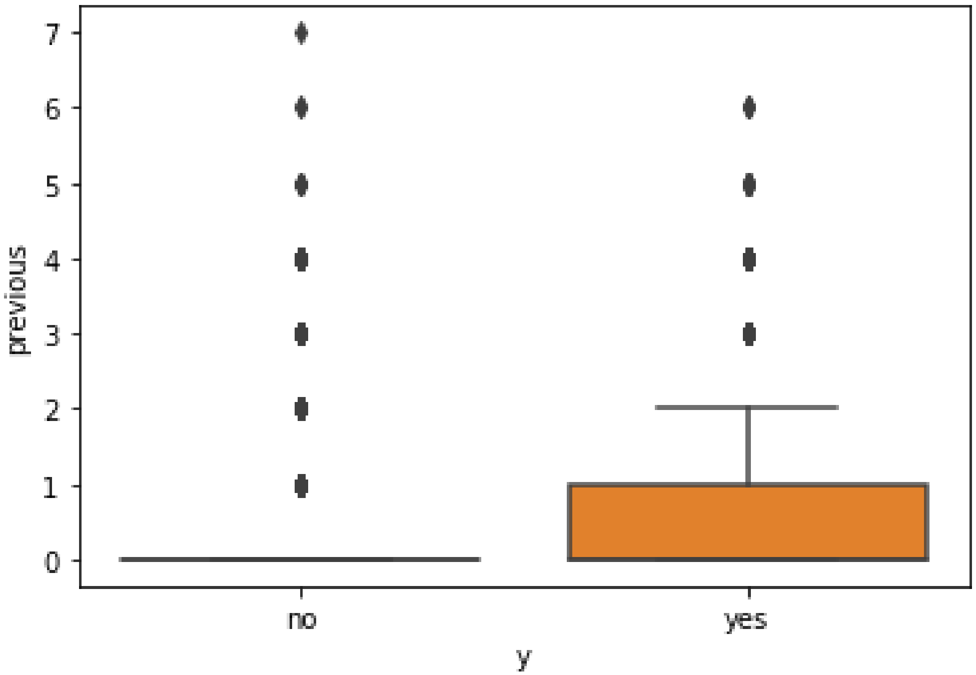


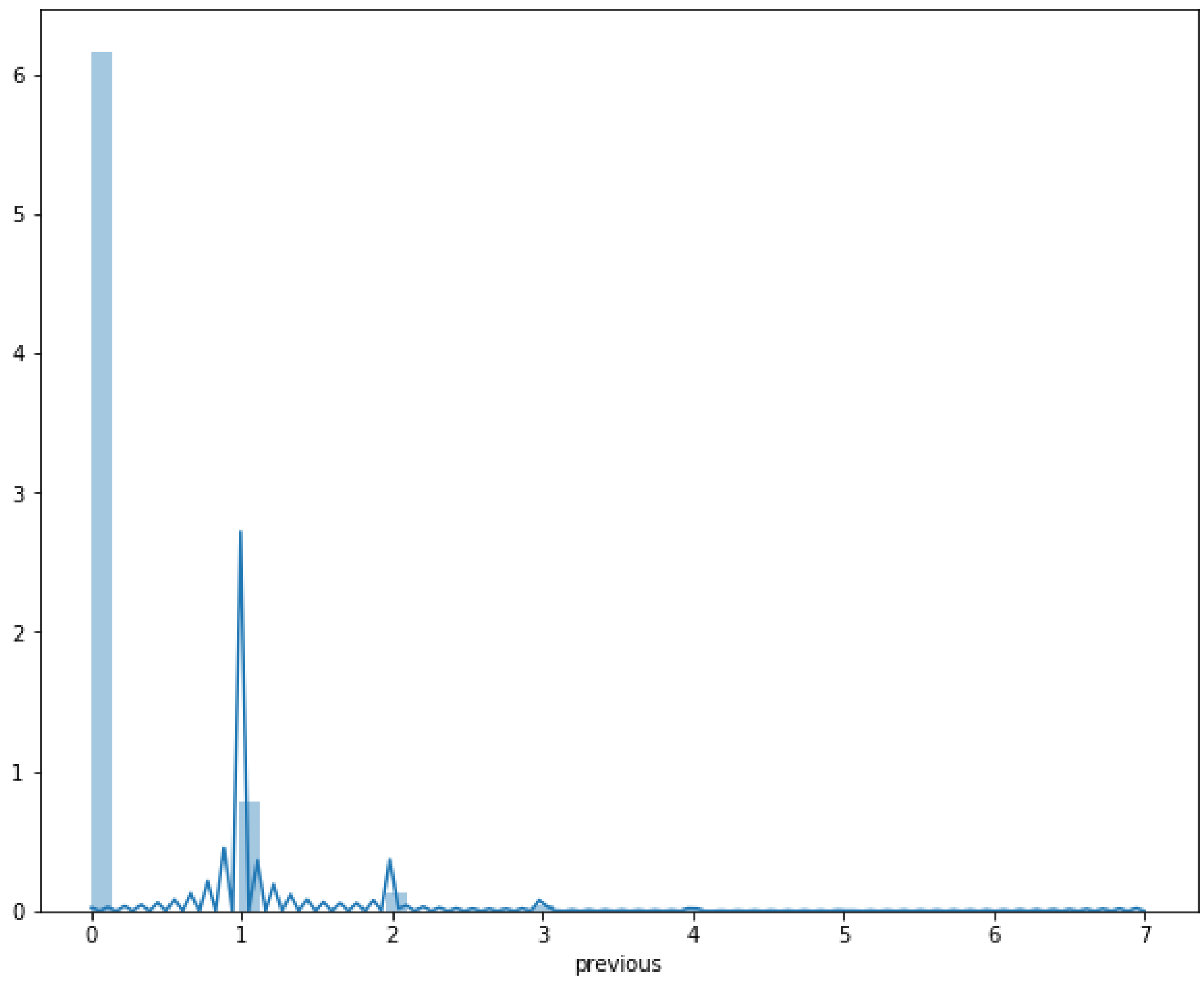
```
Percentile values 0
For yes class
0.0
For No Class
0.0

Percentile values 25
For yes class
999.0
For No Class
999.0

Percentile values 50
For yes class
999.0
For No Class
999.0

Percentile values 75
For yes class
999.0
For No Class
999.0

Percentile values 99
For yes class
999.0
For No Class
999.0
```

- This shows that pdays have almost same percentile values for both the class labels.

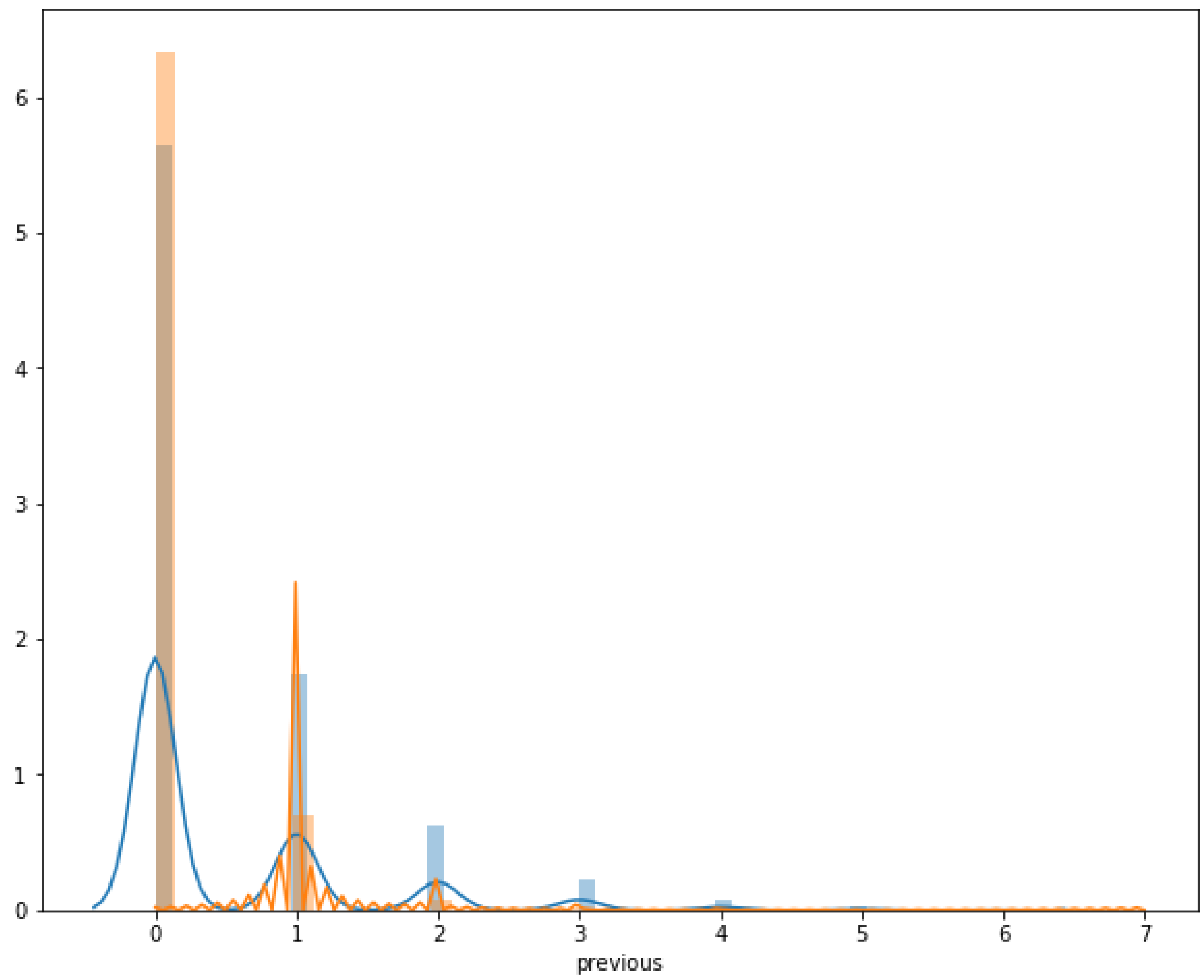
➢ Feature: previous (numeric)

Number of contacts performed before this campaign and for this client (numeric).

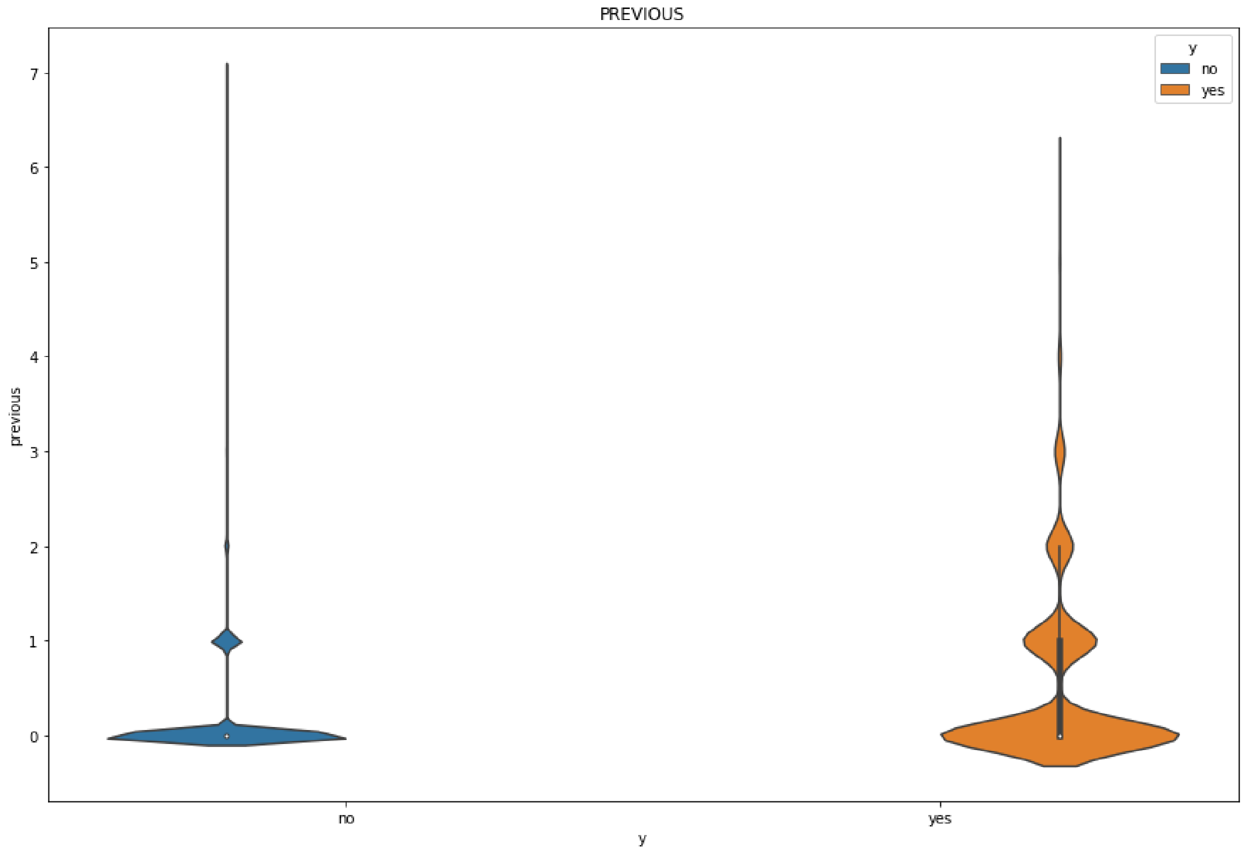




- From the plot above , there might be outliers present after the number 4.



- The previous feature is very similarly distributed for both the classes in the target variable.
- From basic EDA it is not sure how much value this individual feature has on the target variable.

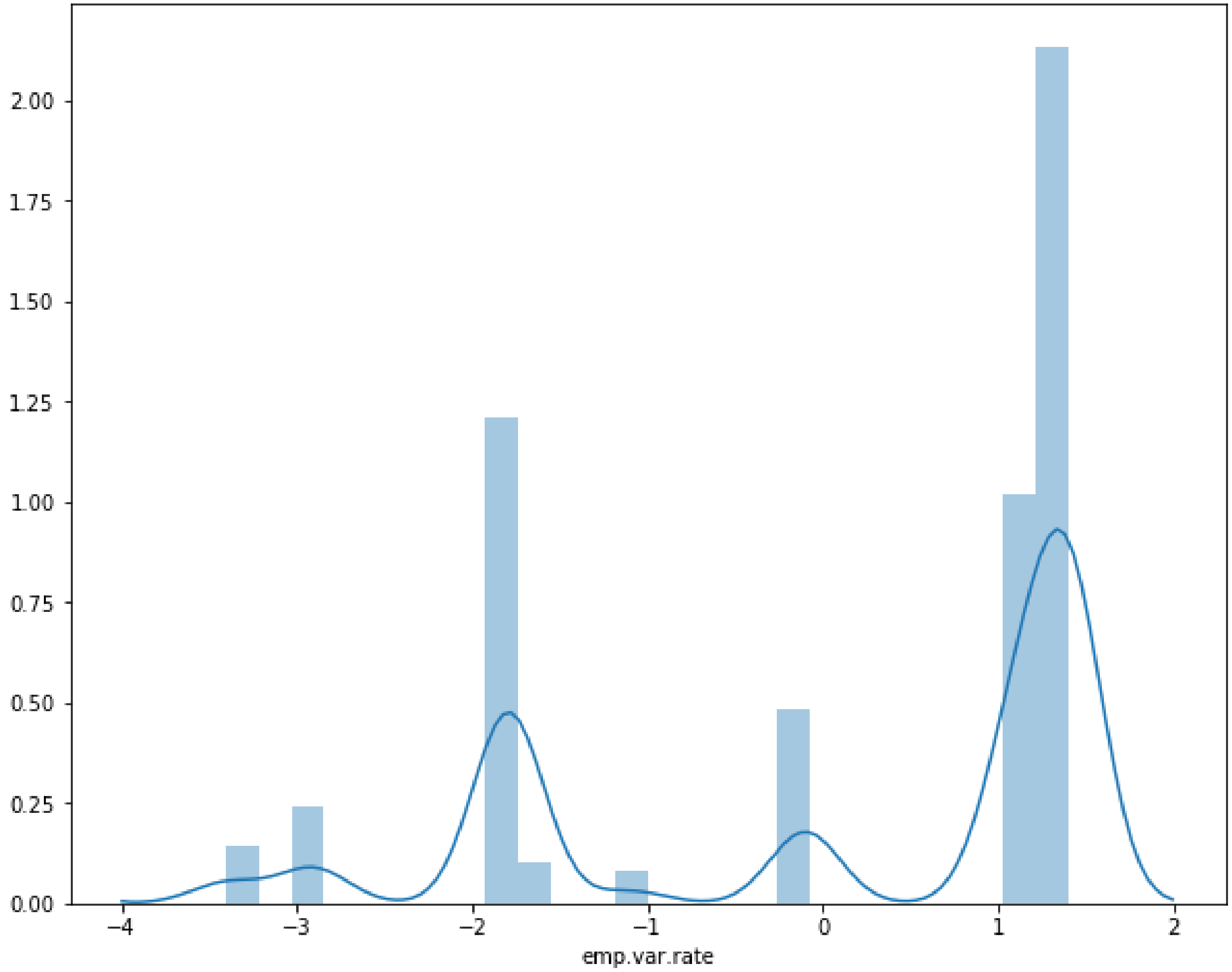


- Using the Violin plot as shown above as we can see, people who have contacted once for the previous campaign have subscribed for long term deposits.
- For class no, there are so many outliers starting with value 1 but for yes class, outliers are present from value 3.

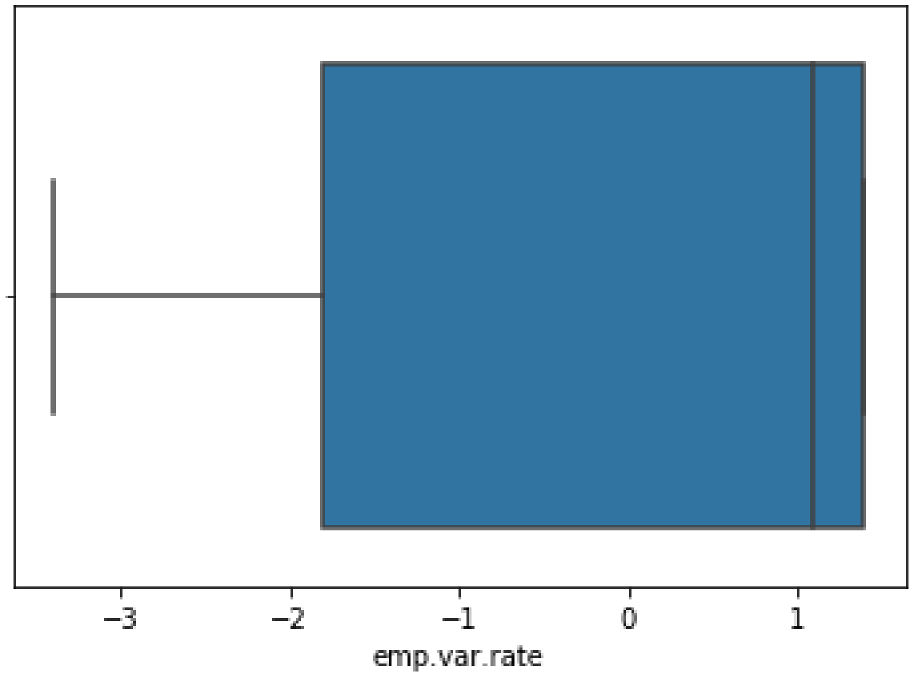
From the plot it is visible that previous feature would be helpful in predicting the class labels.

- **Feature:emp.var.rate(numeric)**

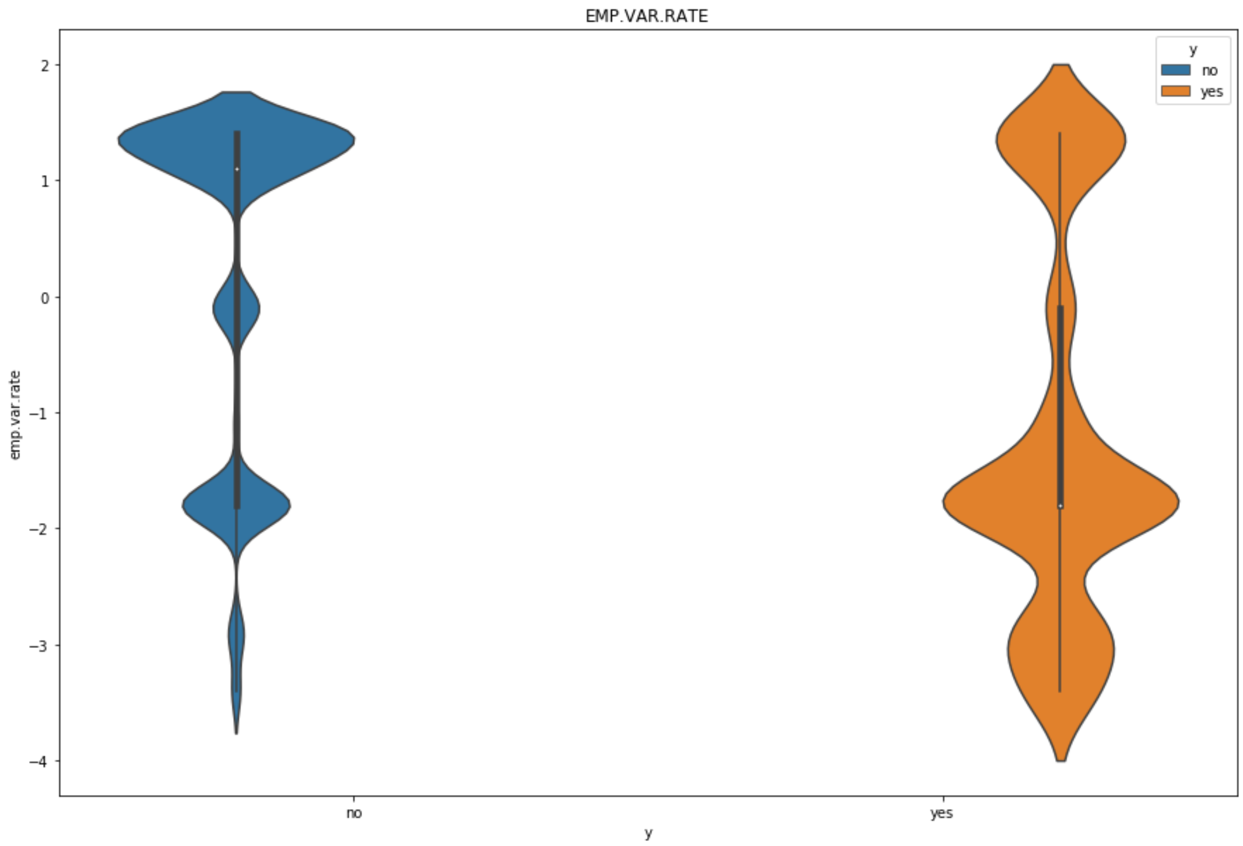
employment variation rate - quarterly indicator (numeric)



- From plot above it is visible that there are no outliers present.



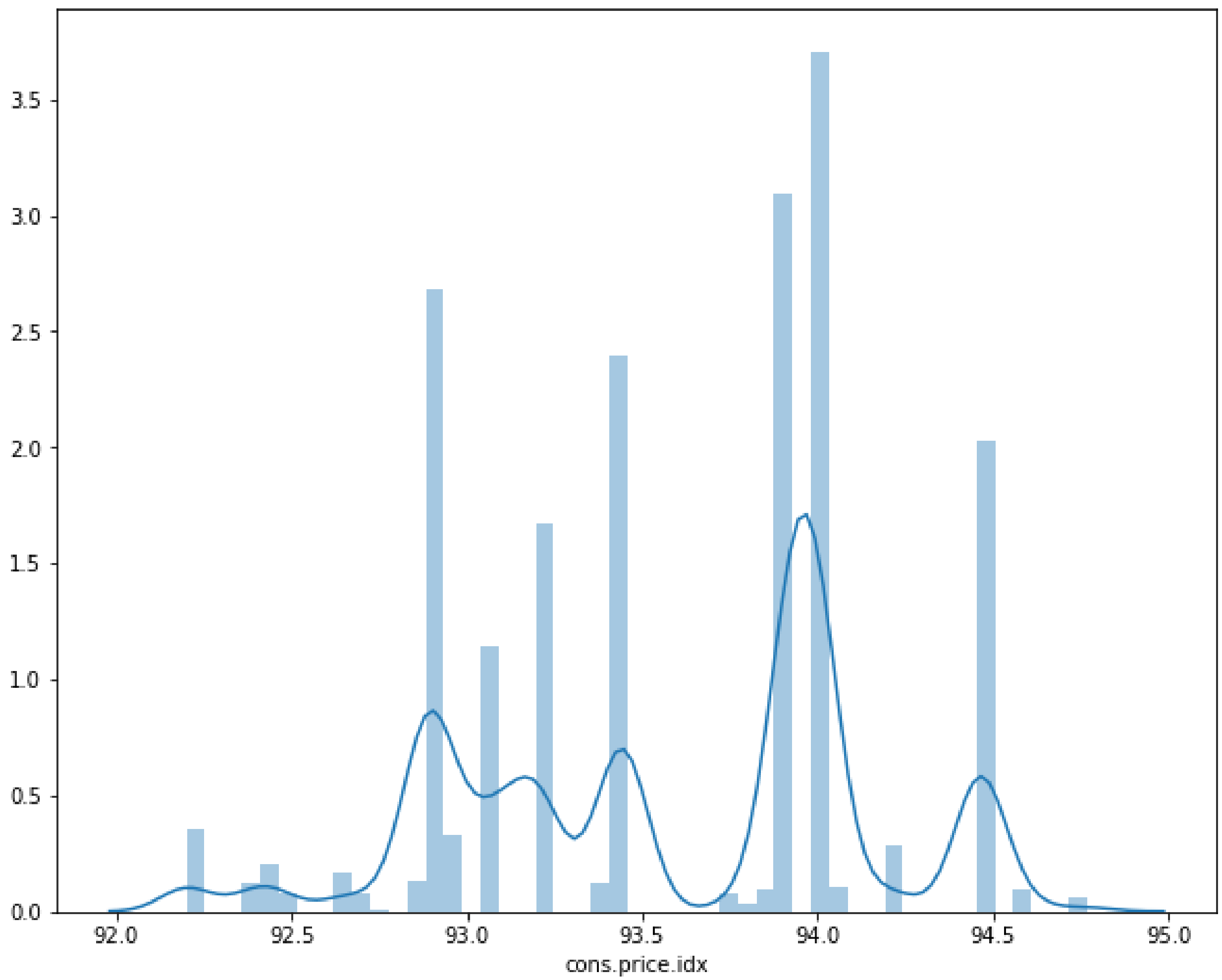
- No outliers are present as shown above.



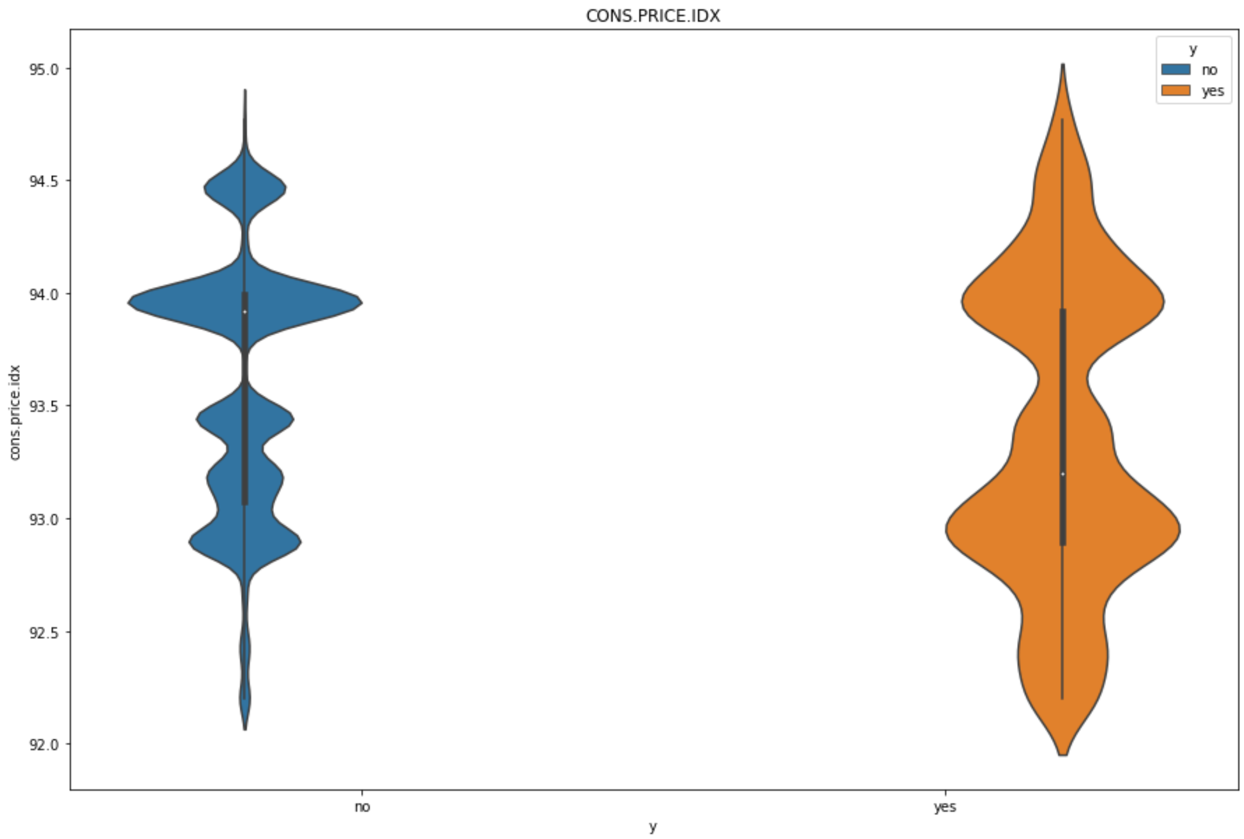
- There are no outliers present for any class for this feature and emp.var. rate fetaure would be very useful in predicting labels.

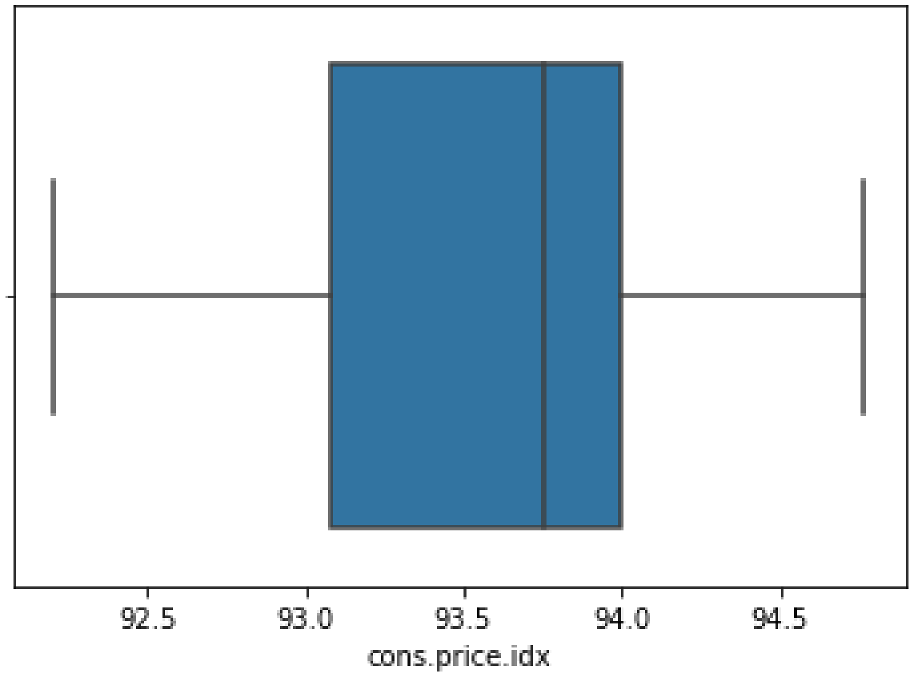
➢ Feature:cons.price.idx(numeric)

Consumer price index - monthly indicator (numeric).



- From plot above we can see that there are no outliers present.

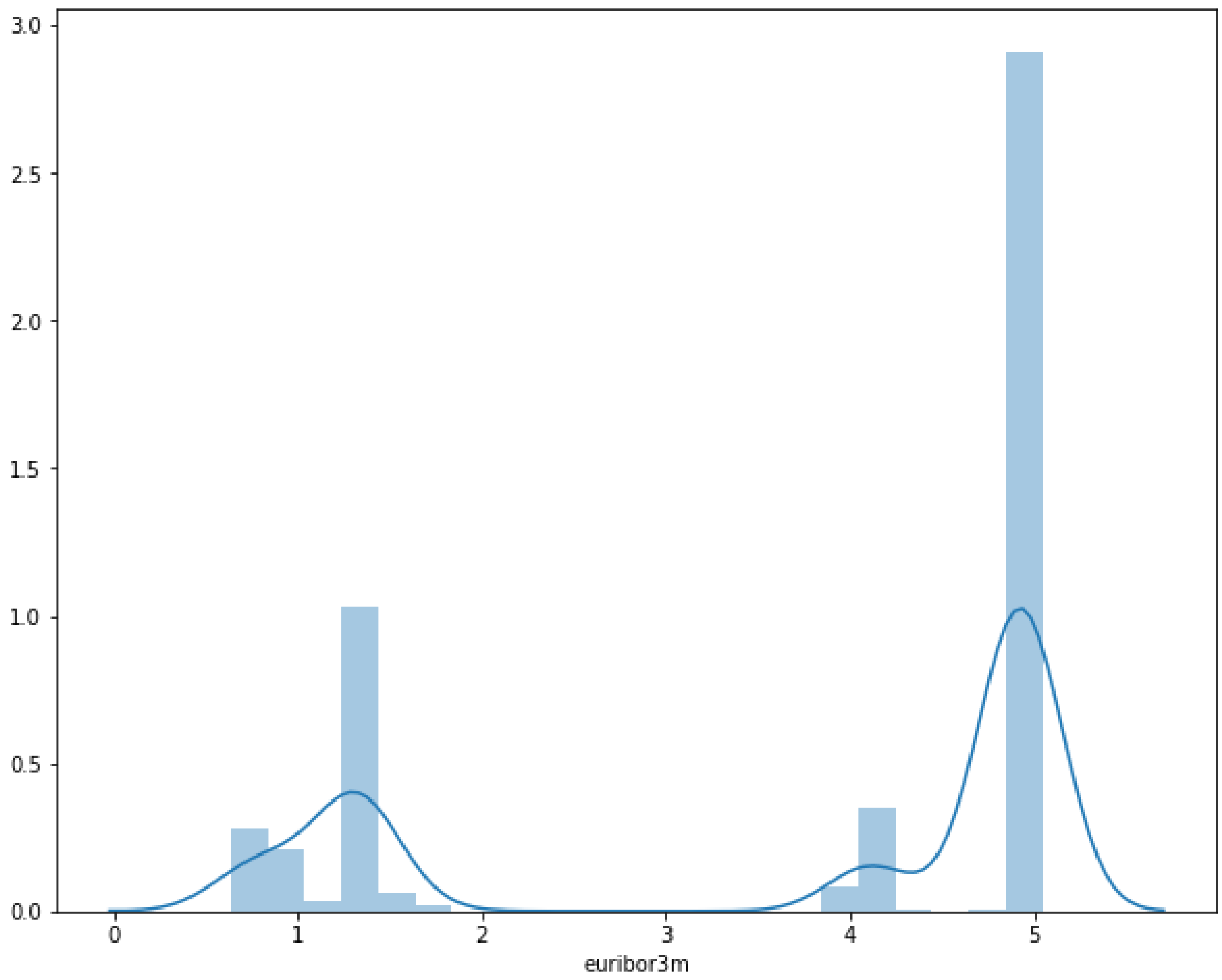

CONS.PRICE.IDX
cons.price.idx
95.0
94.5
94.0
93.5
93.0
92.5
92.0
no
yes
y
y
no
yes

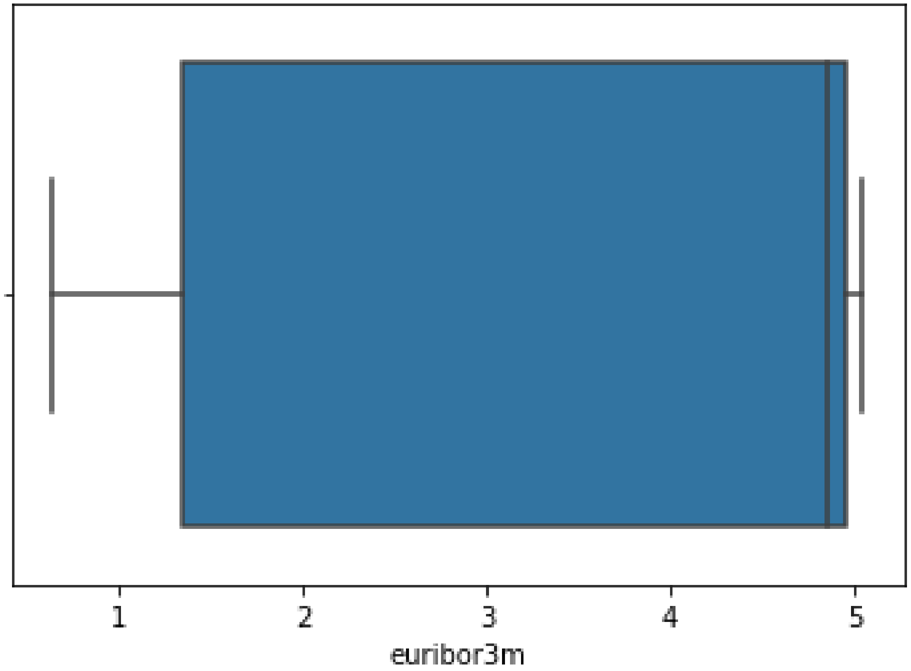


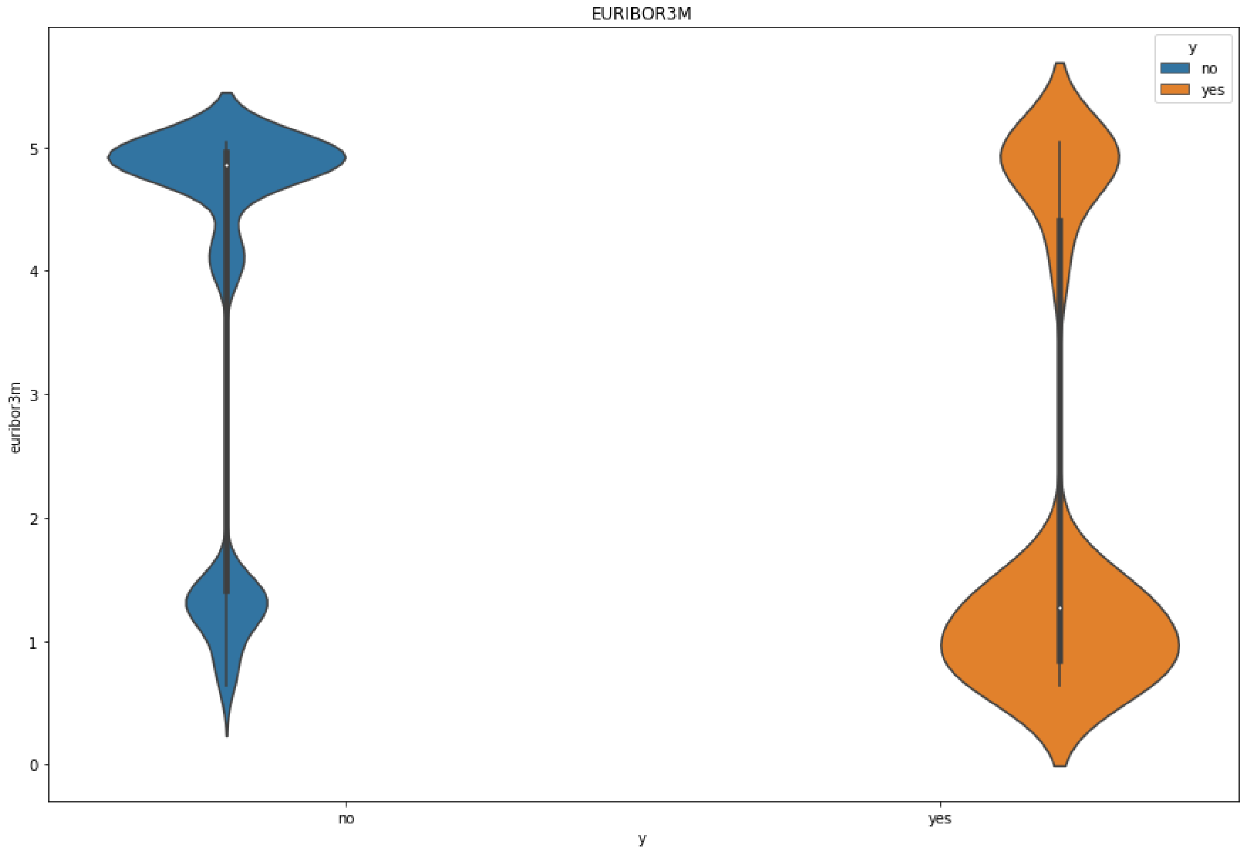
- From the plots above, cons.price.idx does not contain any outliers and they would also be very much helpful in predicting class labels.

➢ Feature:euribor3m(numeric)

Euribor 3 month rate - daily indicator (numeric).


3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
euribor3m

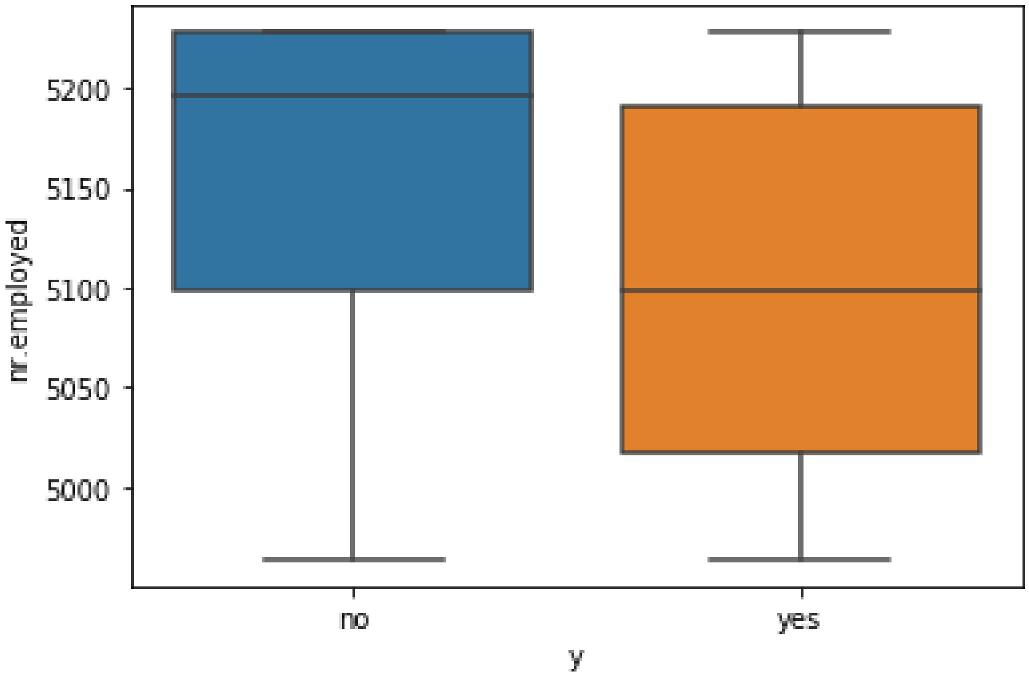

1
2
3
4
5
euribor3m

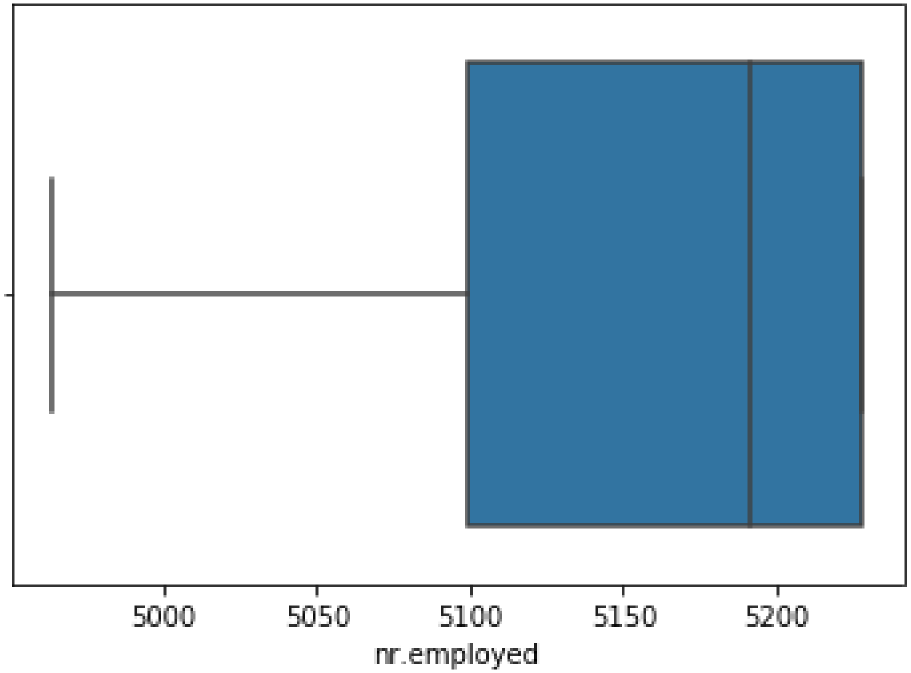


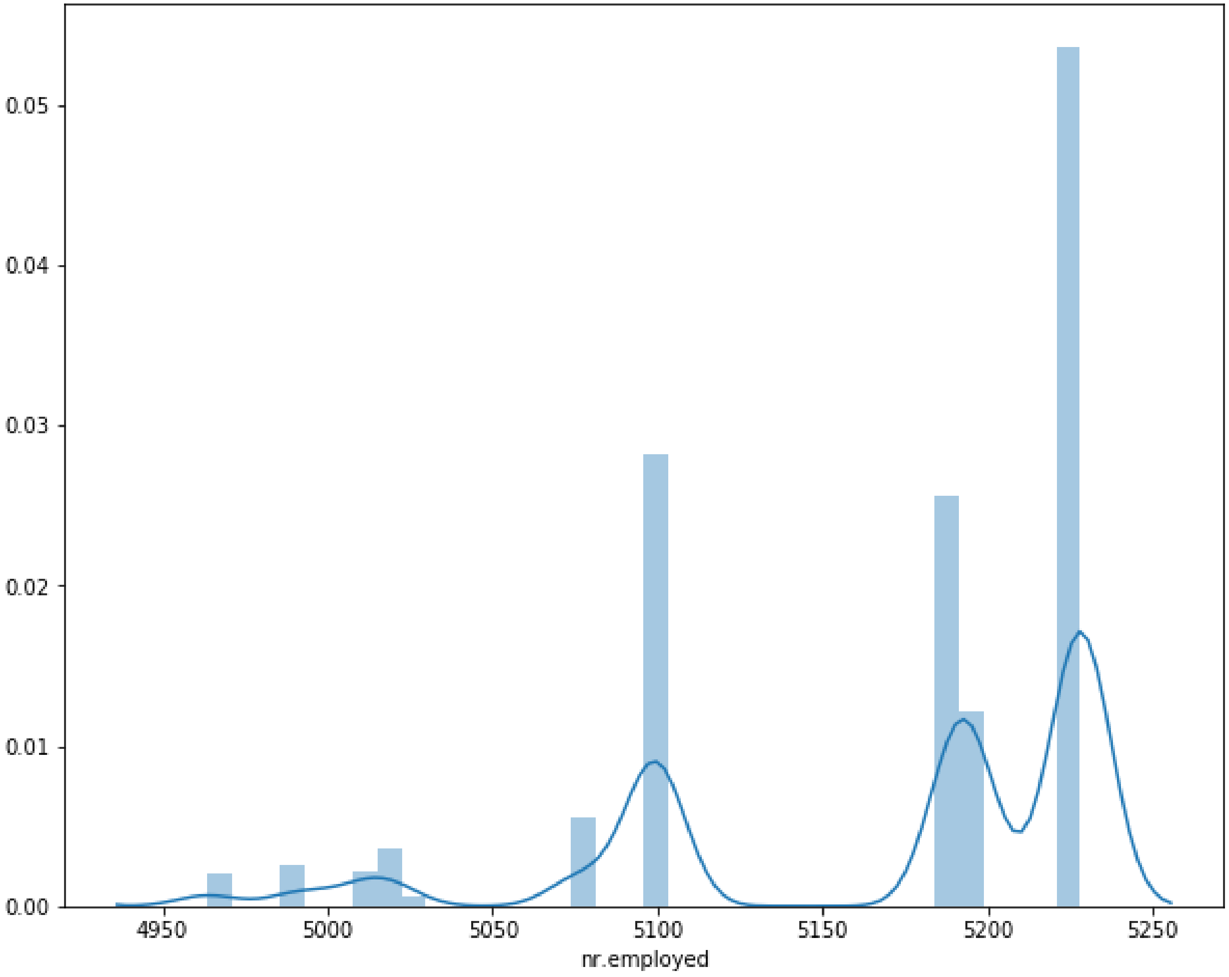
- From the plots above, Euribor3m does not contain outliers and would be very much helpful in predicting the class labels.

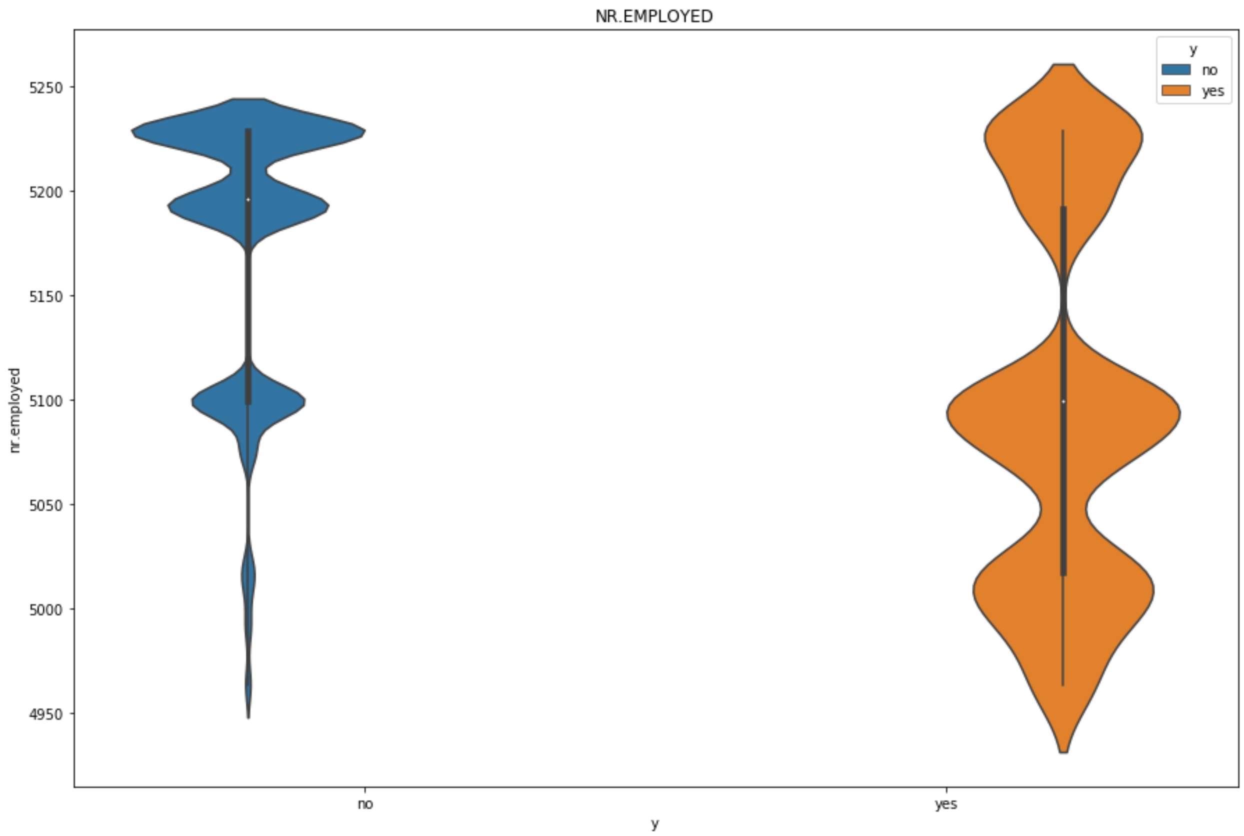
➢ Feature:nr.employed(numeric)

Number of employees - quarterly indicator (numeric).


nr.employed
5200
5150
5100
5050
5000
no
yes
y


5000
5050
5100
5150
5200
nr.employed

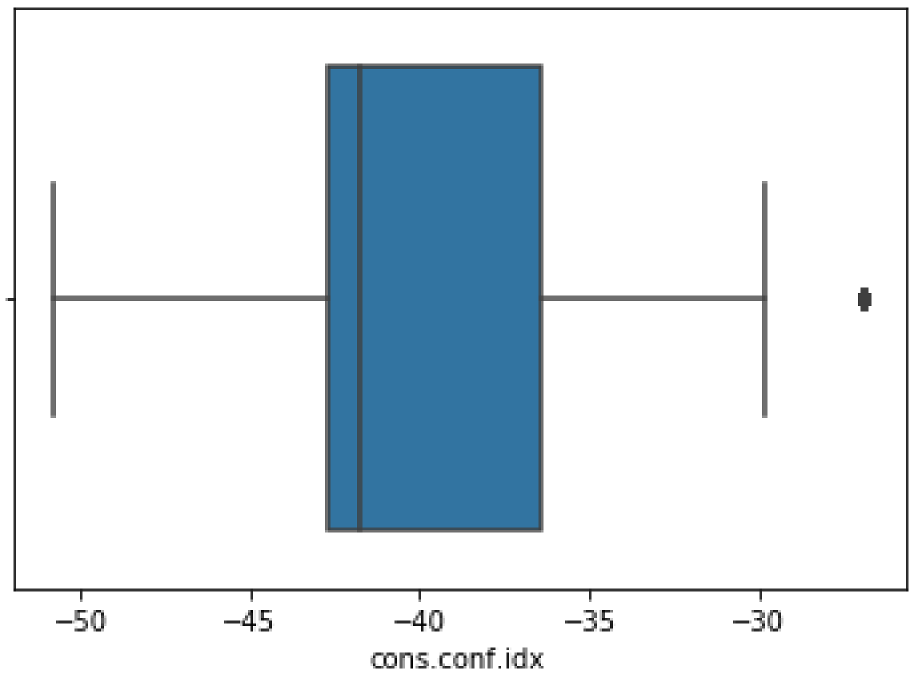

0.05
0.04
0.03
0.02
0.01
0.00
4950
5000
5050
5100
5150
5200
5250
nr.employed

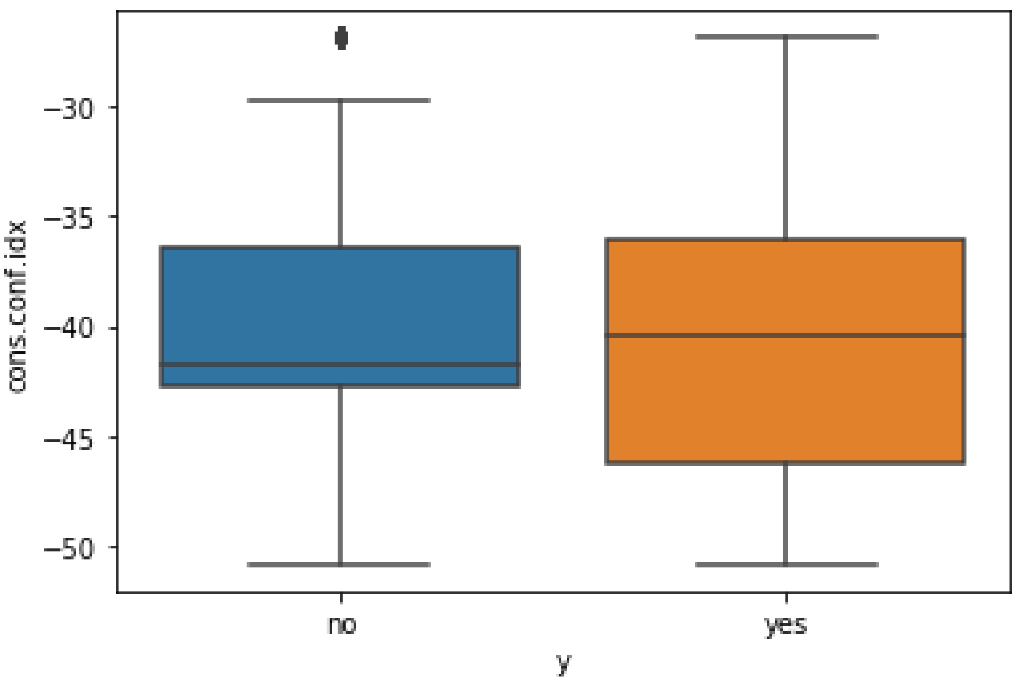


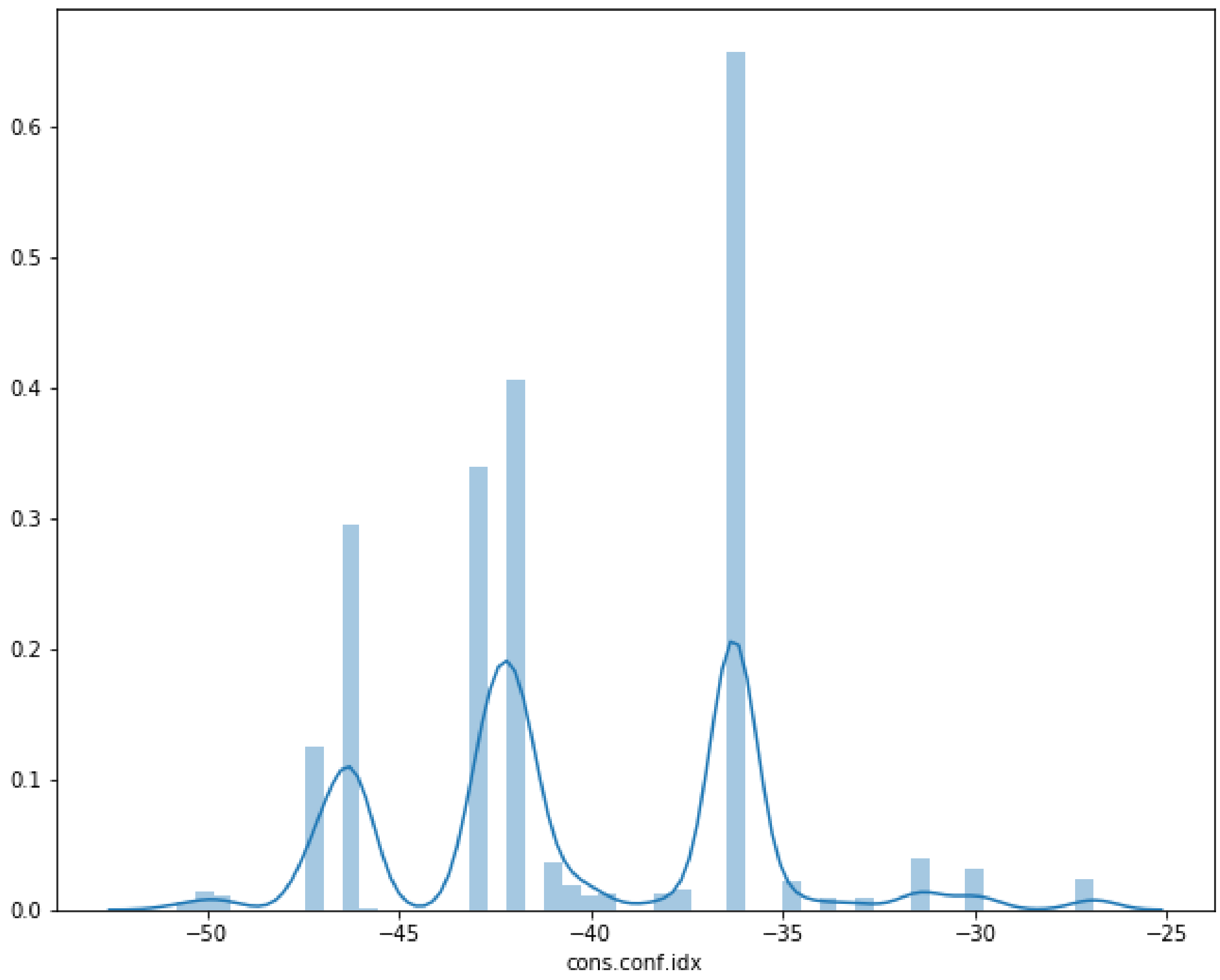
- From the above plots, nr. employed does not contain outliers and nr. employed would also be very much helpful in predicting class labels.

➢ Feature:cons.conf.idx(numeric)

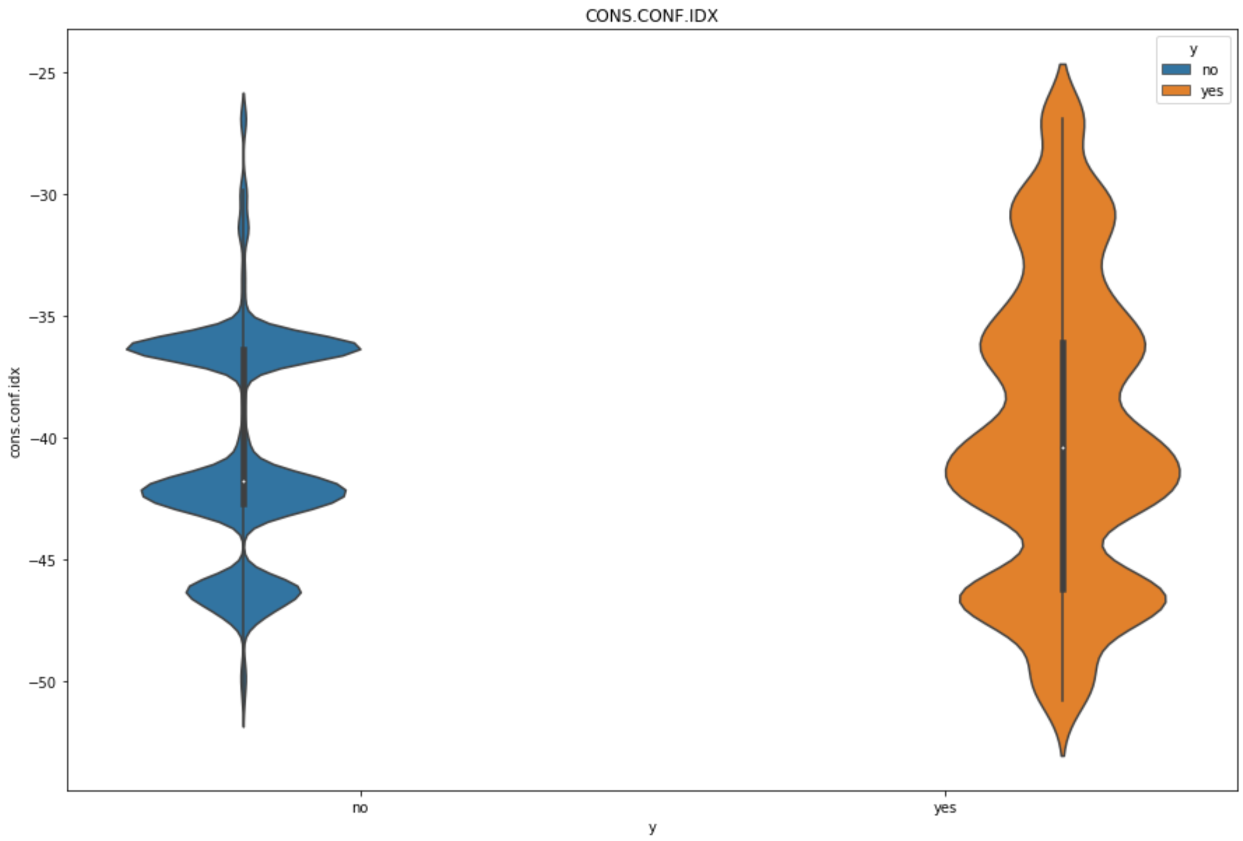
consumer confidence index - monthly indicator (numeric).


−50
−45
−40
−35
−30
cons.conf.idx


−30
−35
−40
−45
−50
cons.conf.idx
no
yes
y



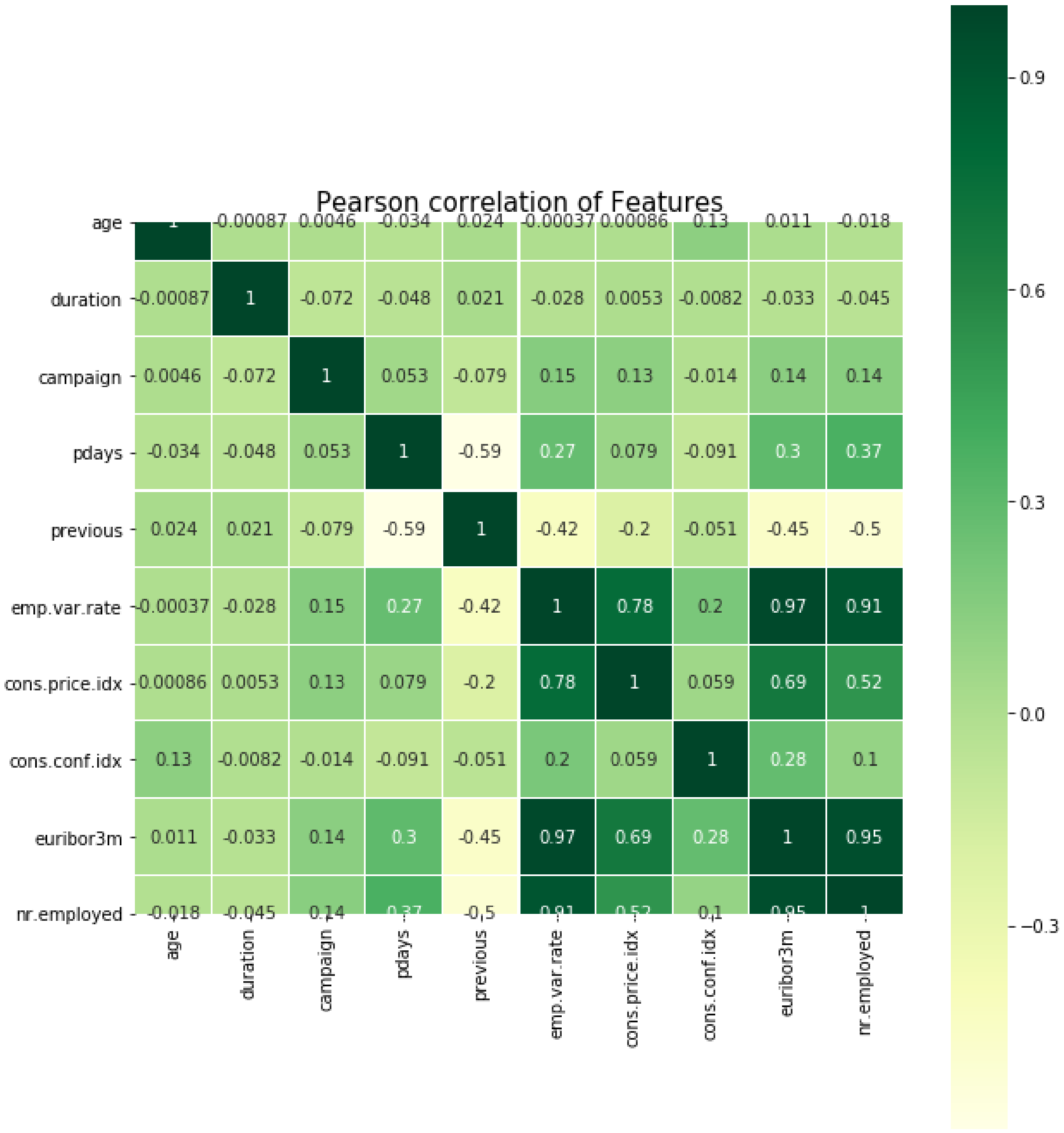
- From the above plot there might be a case of outliers. Let’s check with violinplot



- In cons.conf.idx feature for class labels no, there is an outlier present when value above -30.

- **Correlation among features:**
  I am going to use a Pearson correlation.

Let’s Check for correlation of features between the Numerical Features.


Pearson correlation of Features
age
duration
campaign
pdays
previous
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
1 -0.00087 0.0046 -0.034 0.024 -0.00037 0.00086 0.13 0.011 -0.018
-0.00087 1 -0.072 -0.048 0.021 -0.028 0.0053 -0.0082 -0.033 -0.045
0.0046 -0.072 1 0.053 -0.079 0.15 0.13 -0.014 0.14 0.14
-0.034 -0.048 0.053 1 -0.59 0.27 0.079 -0.091 0.3 0.37
0.024 0.021 -0.079 -0.59 1 -0.42 -0.2 -0.051 -0.45 -0.5
-0.00037 -0.028 0.15 0.27 -0.42 1 0.78 0.2 0.97 0.91
0.00086 0.0053 0.13 0.079 -0.2 0.78 1 0.059 0.69 0.52
0.13 -0.0082 -0.014 -0.091 -0.051 0.2 0.059 1 0.28 0.1
0.011 -0.033 0.14 0.3 -0.45 0.97 0.69 0.28 1 0.95
-0.018 -0.045 0.14 0.37 -0.5 0.91 0.52 0.1 0.95 1
0.9
0.6
0.3
0.0
−0.3

- From the above heatmap we can see that there are some numerical features which share a high correlation between them, e.g., **nr. employed** and **euribor3m** these features share a correlation value of 0.95, and **euribor3m** and **emp.var. rate** share a correlation of 0.97, which is very high compared to the other features that we see in the heatmap.

➢ Feature Engineering:

Feature engineering is the process of converting data into features that improves the prediction

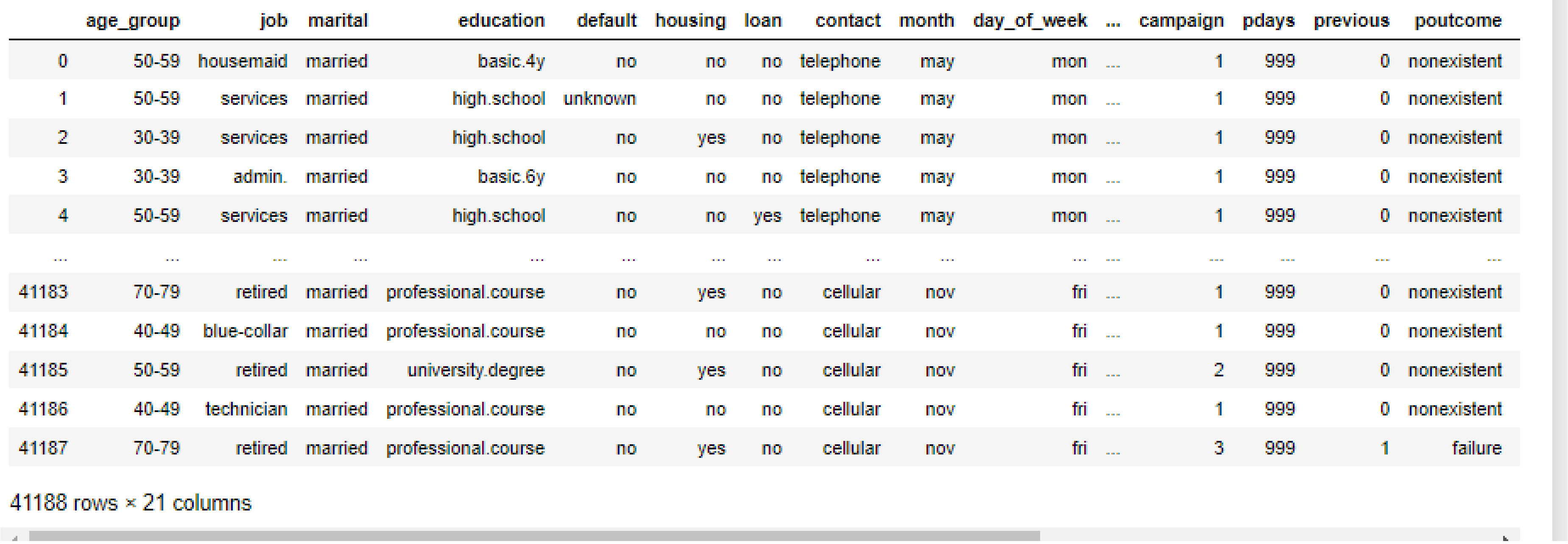
and performance of model in unseen data.

- **Converting an age(numerical) variable to a categorical variable.**

Here I have created 9 groups from minimum age 10 to maximum age 100.

After creating, inserted the age group into data frame and deleted the age column from the data frame.

The numerical **age** variable is now replaced by a new categorical **age_group** variable as shown below.



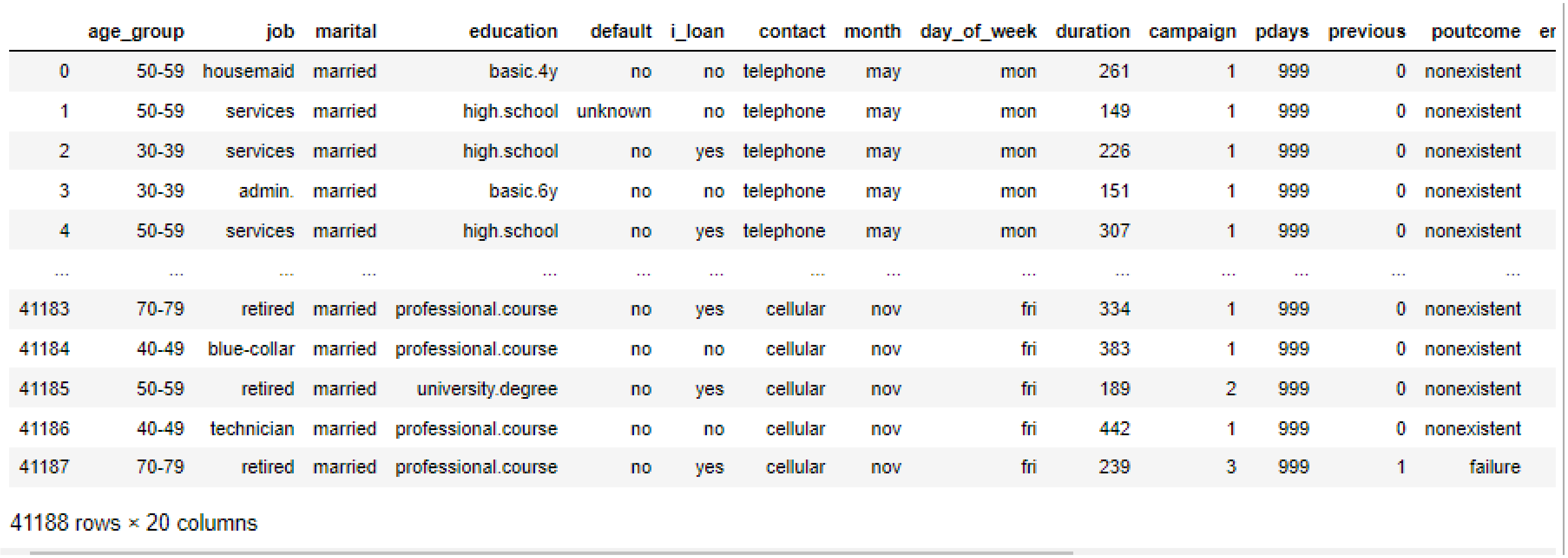
	age_group	job	marital	education	default	housing	loan	contact	month	day_of_week	...	campaign	pdays	previous	poutcome
0	50-59	housemaid	married	basic.4y	no	no	no	telephone	may	mon	...	1	999	0	nonexistent
1	50-59	services	married	high.school	unknown	no	no	telephone	may	mon	...	1	999	0	nonexistent
2	30-39	services	married	high.school	no	yes	no	telephone	may	mon	...	1	999	0	nonexistent
3	30-39	admin.	married	basic.6y	no	no	no	telephone	may	mon	...	1	999	0	nonexistent
4	50-59	services	married	high.school	no	no	yes	telephone	may	mon	...	1	999	0	nonexistent
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
41183	70-79	retired	married	professional.course	no	yes	no	cellular	nov	fri	...	1	999	0	nonexistent
41184	40-49	blue-collar	married	professional.course	no	no	no	cellular	nov	fri	...	1	999	0	nonexistent
41185	50-59	retired	married	university.degree	no	yes	no	cellular	nov	fri	...	2	999	0	nonexistent
41186	40-49	technician	married	professional.course	no	no	no	cellular	nov	fri	...	1	999	0	nonexistent
41187	70-79	retired	married	professional.course	no	yes	no	cellular	nov	fri	...	3	999	1	failure

41188 rows × 21 columns

- **Creating i_loan column; Deleting Housing and Loan column**

  - I will now create i_loan based on the columns['loan','housing'], status we have 3 statuses yes, no, unknown.
  - If any 2 columns have status yes then ,i_loan will have yes as status, if any of the columns have no as status, then i_loan will have no as status or else unknown status.

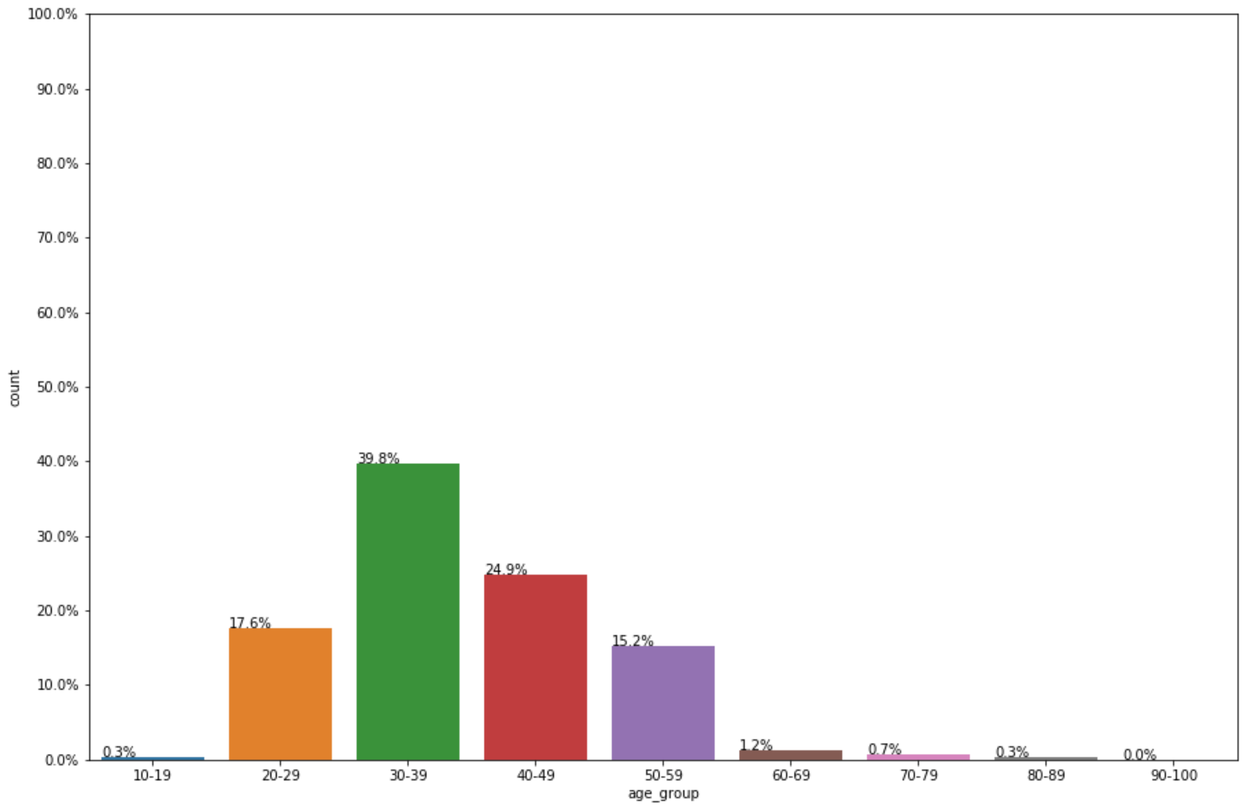
This is how a new dataframe looks, the dataframe now has a new variable **i_loan** and there are now 20 variables present as shown below.



	age_group	job	marital	education	default	i_loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	er
0	50-59	housemaid	married	basic.4y	no	no	telephone	may	mon	261	1	999	0	nonexistent	
1	50-59	services	married	high.school	unknown	no	telephone	may	mon	149	1	999	0	nonexistent	
2	30-39	services	married	high.school	no	yes	telephone	may	mon	226	1	999	0	nonexistent	
3	30-39	admin.	married	basic.6y	no	no	telephone	may	mon	151	1	999	0	nonexistent	
4	50-59	services	married	high.school	no	yes	telephone	may	mon	307	1	999	0	nonexistent	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
41183	70-79	retired	married	professional.course	no	yes	cellular	nov	fri	334	1	999	0	nonexistent	
41184	40-49	blue-collar	married	professional.course	no	no	cellular	nov	fri	383	1	999	0	nonexistent	
41185	50-59	retired	married	university.degree	no	yes	cellular	nov	fri	189	2	999	0	nonexistent	
41186	40-49	technician	married	professional.course	no	no	cellular	nov	fri	442	1	999	0	nonexistent	
41187	70-79	retired	married	professional.course	no	yes	cellular	nov	fri	239	3	999	1	failure	

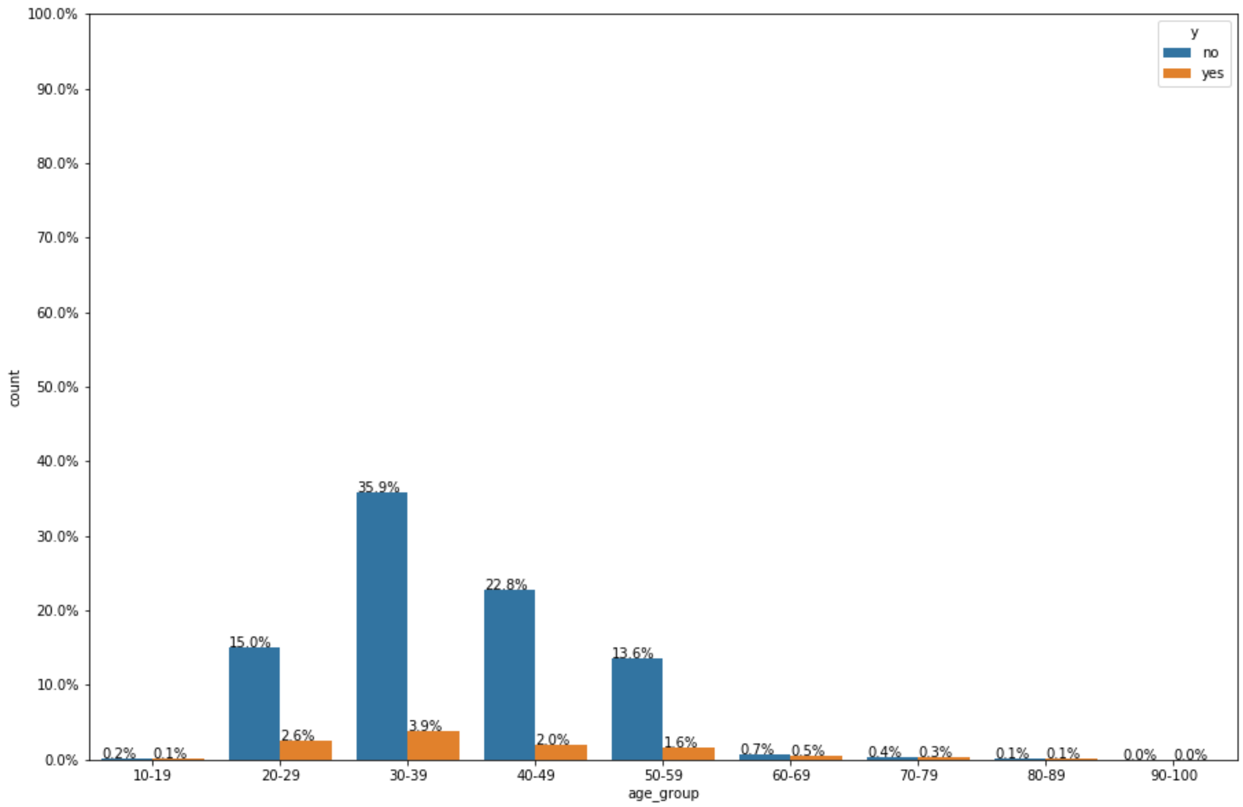
41188 rows × 20 columns

- **Which age group has the bank contacted the most?**

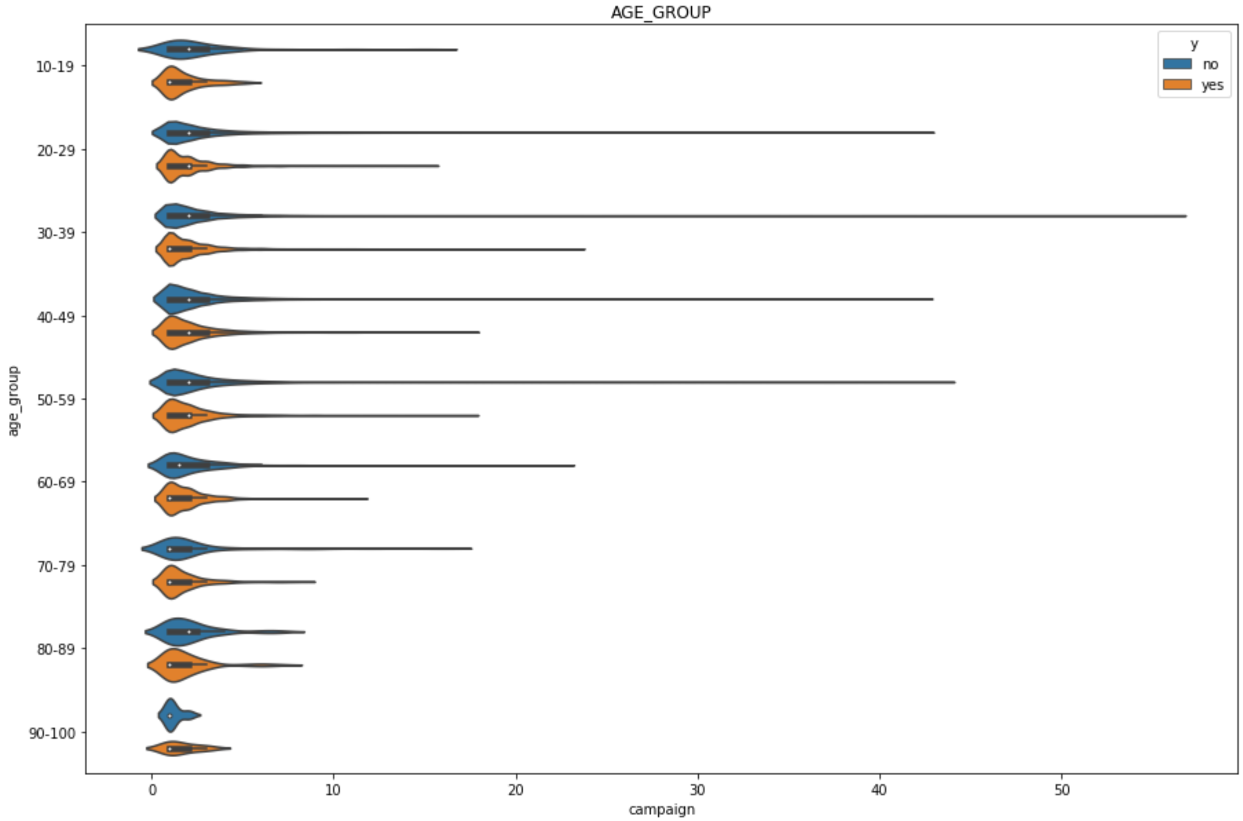


- As shown above, the bank has contacted to the most between the age group of 30-39 followed by 40-49.

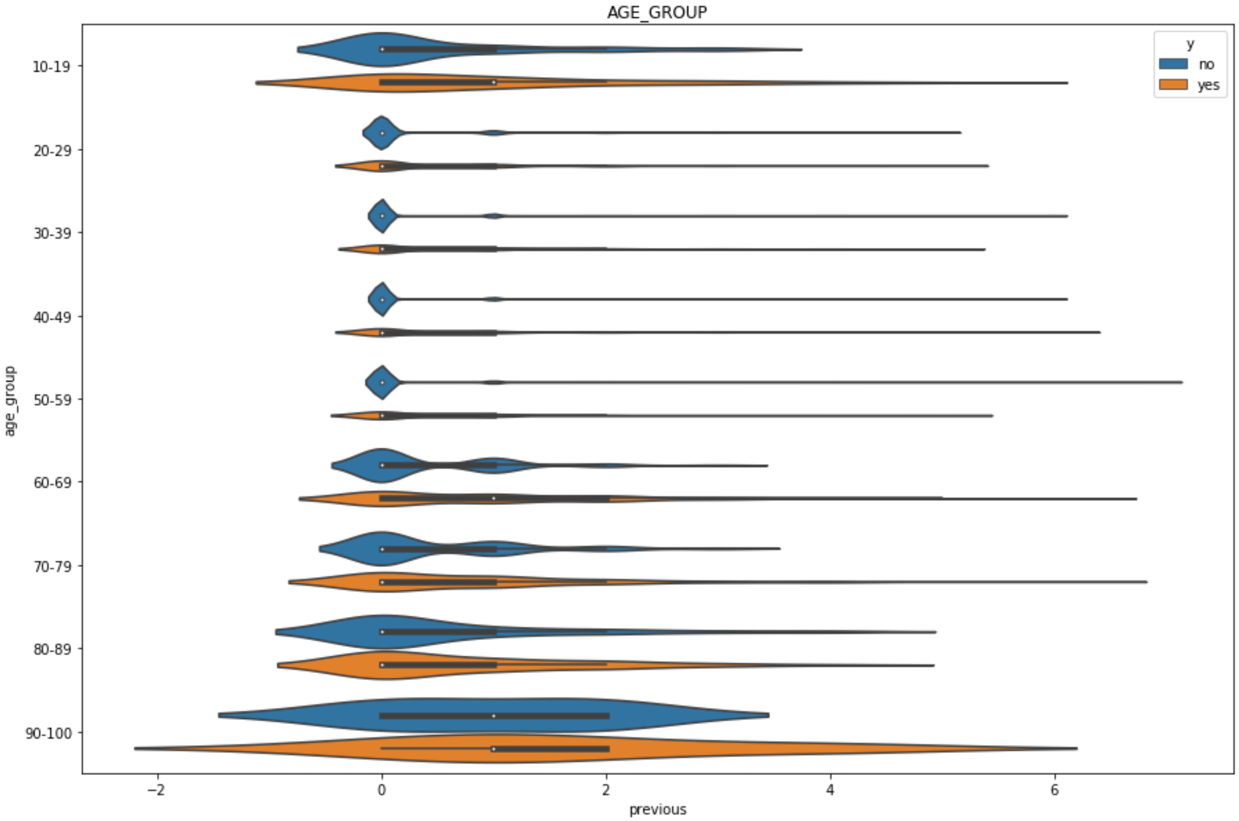
➢ Which age group is likely to subscribe for long term deposits?



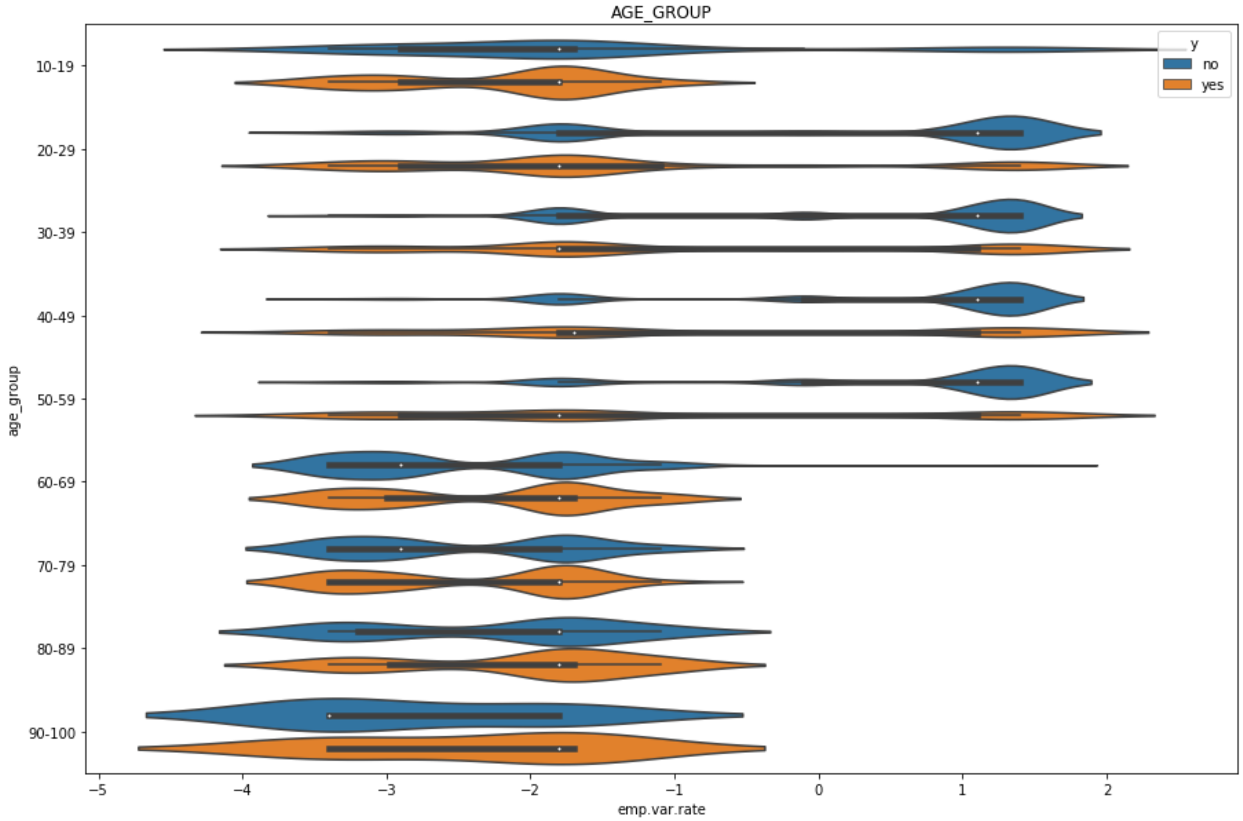
- Age group of 30-39 are the most people who have not subscribed for the deposits.
- They are also the most who have subscribed for the deposits.



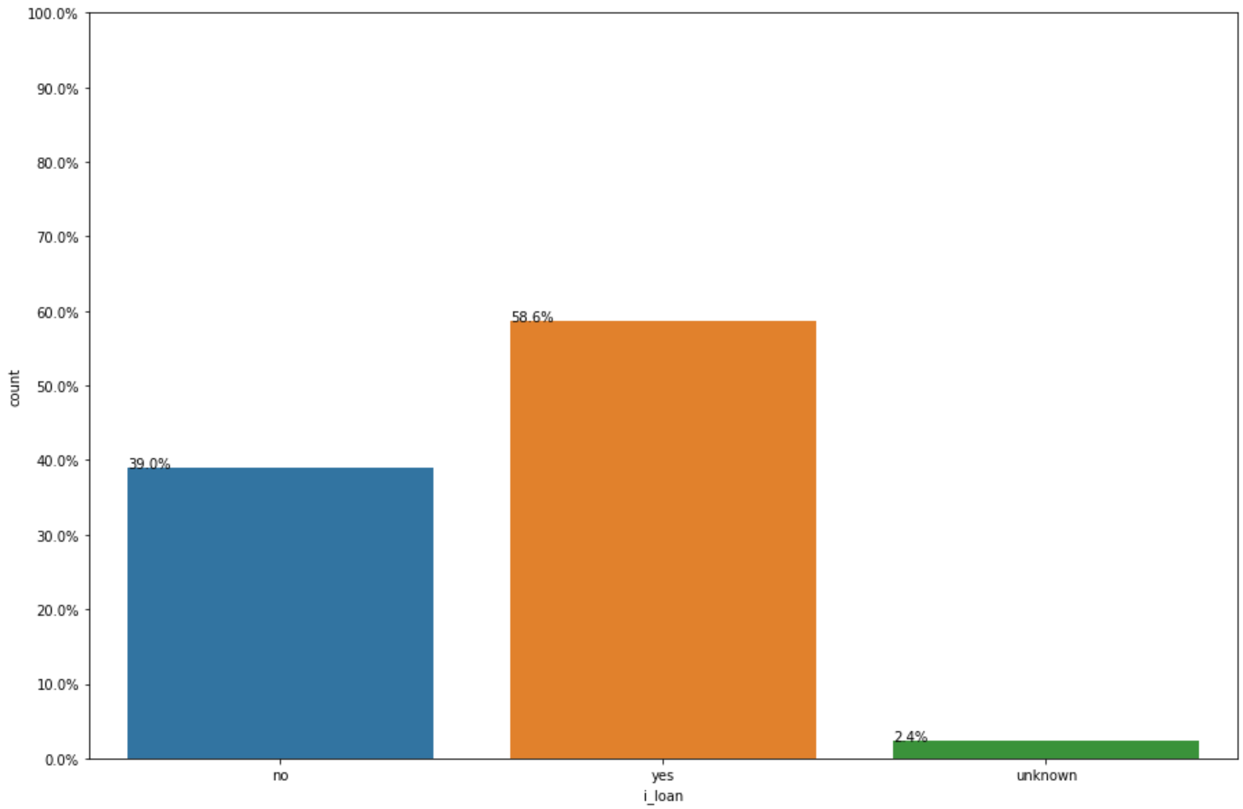
- As shown above, number of campaigns from 1-7 and age above 70 has possible outliers.



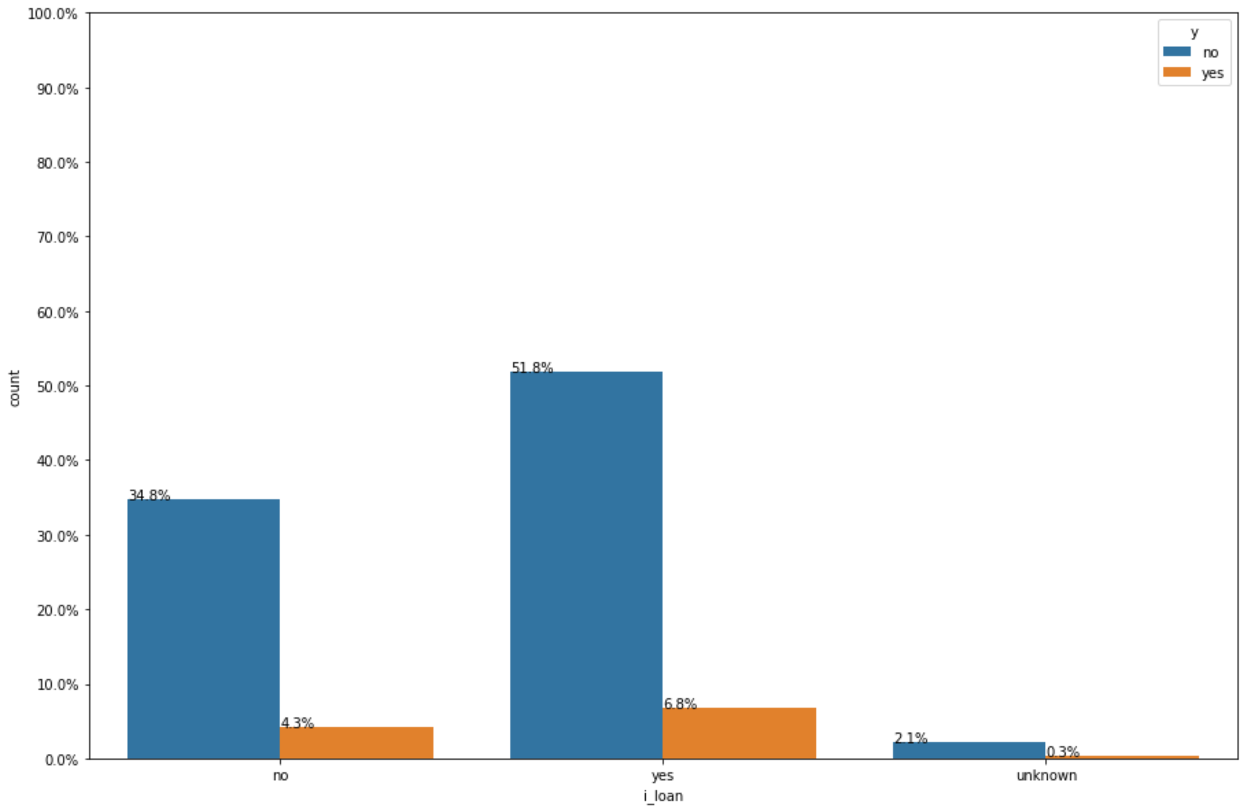
- For number of previous campaigns from 0-1 age group above 70 are possible outliers.
- For number of contacts for previous campaign as 2, age around 90 are possible outliers.
- For number of previous campaigns from 3-4 age around 80 are possible outliers.



- For emp.var. rate with -1.8 has most outliers above age around 60.

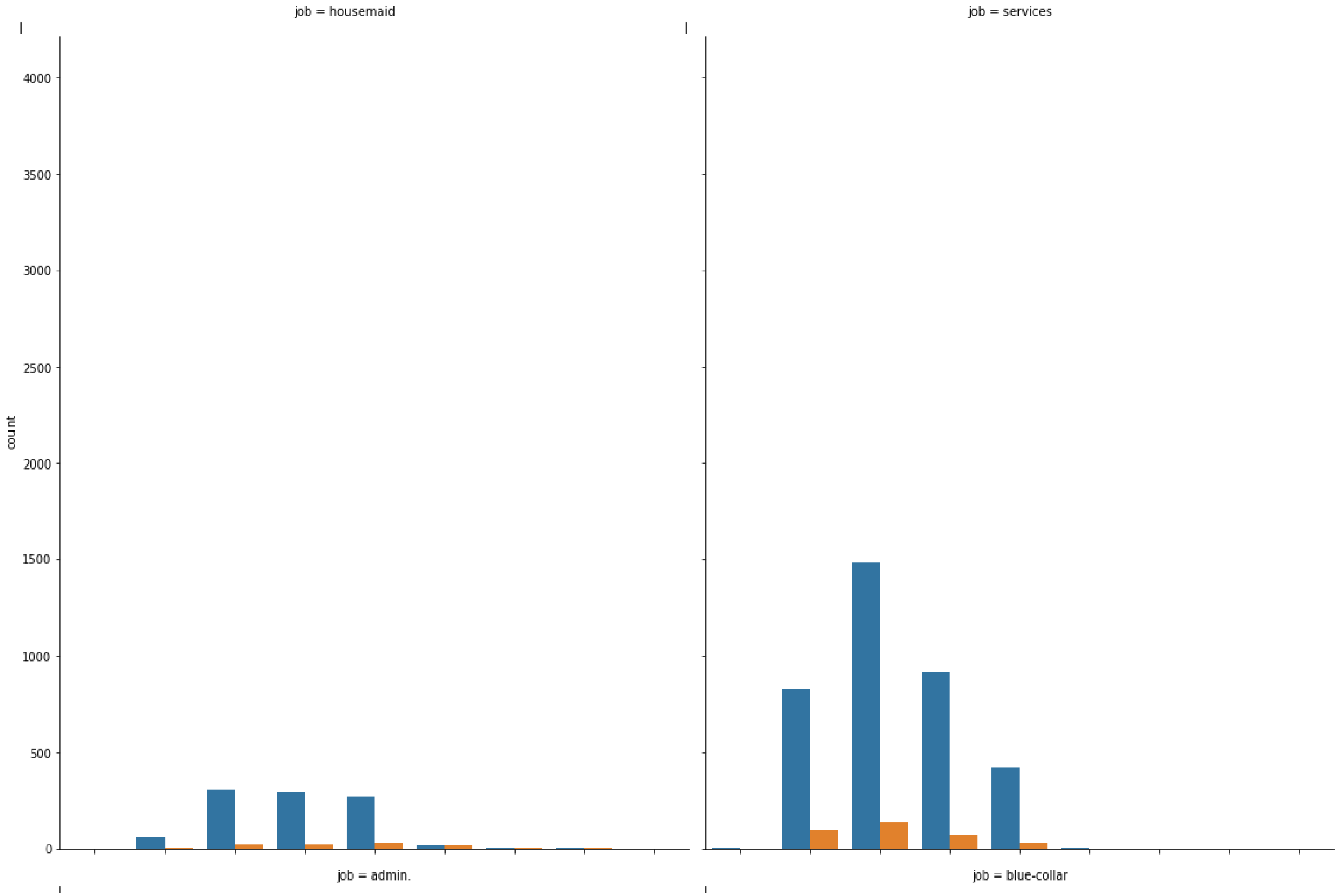


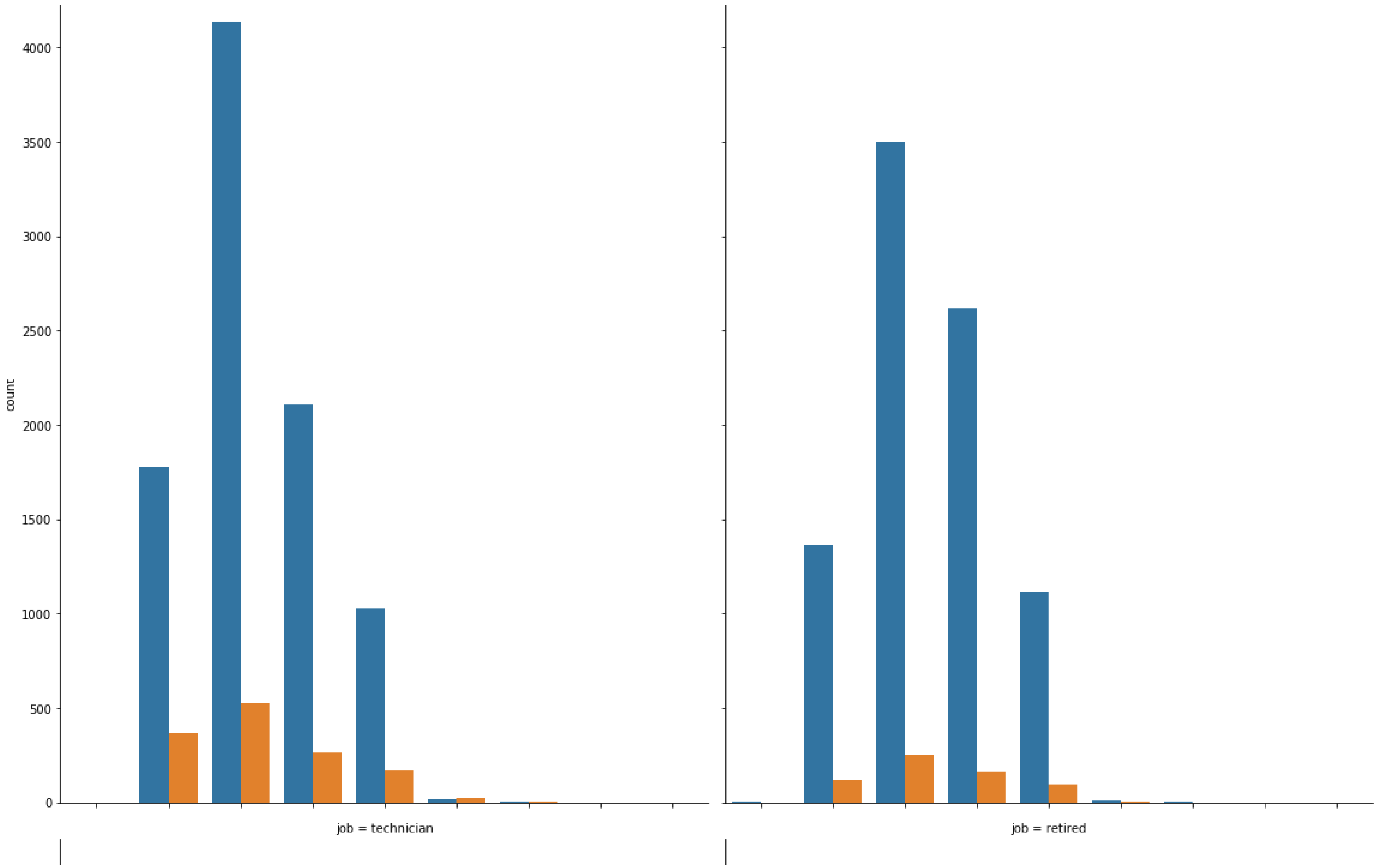
- Most of the people the bank has contacted either have personal or housing loan.
- Very few of the status of loan is unknown.

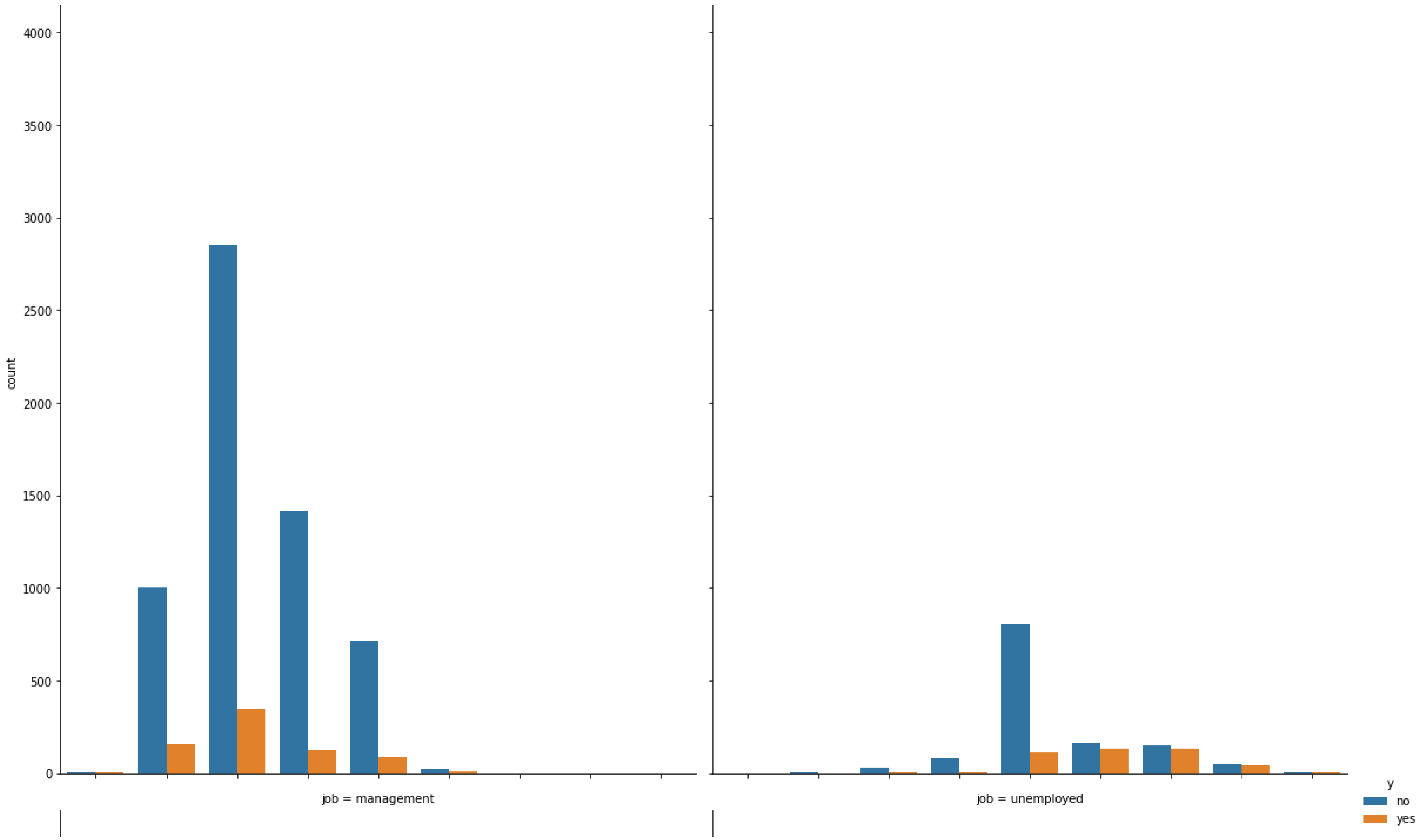


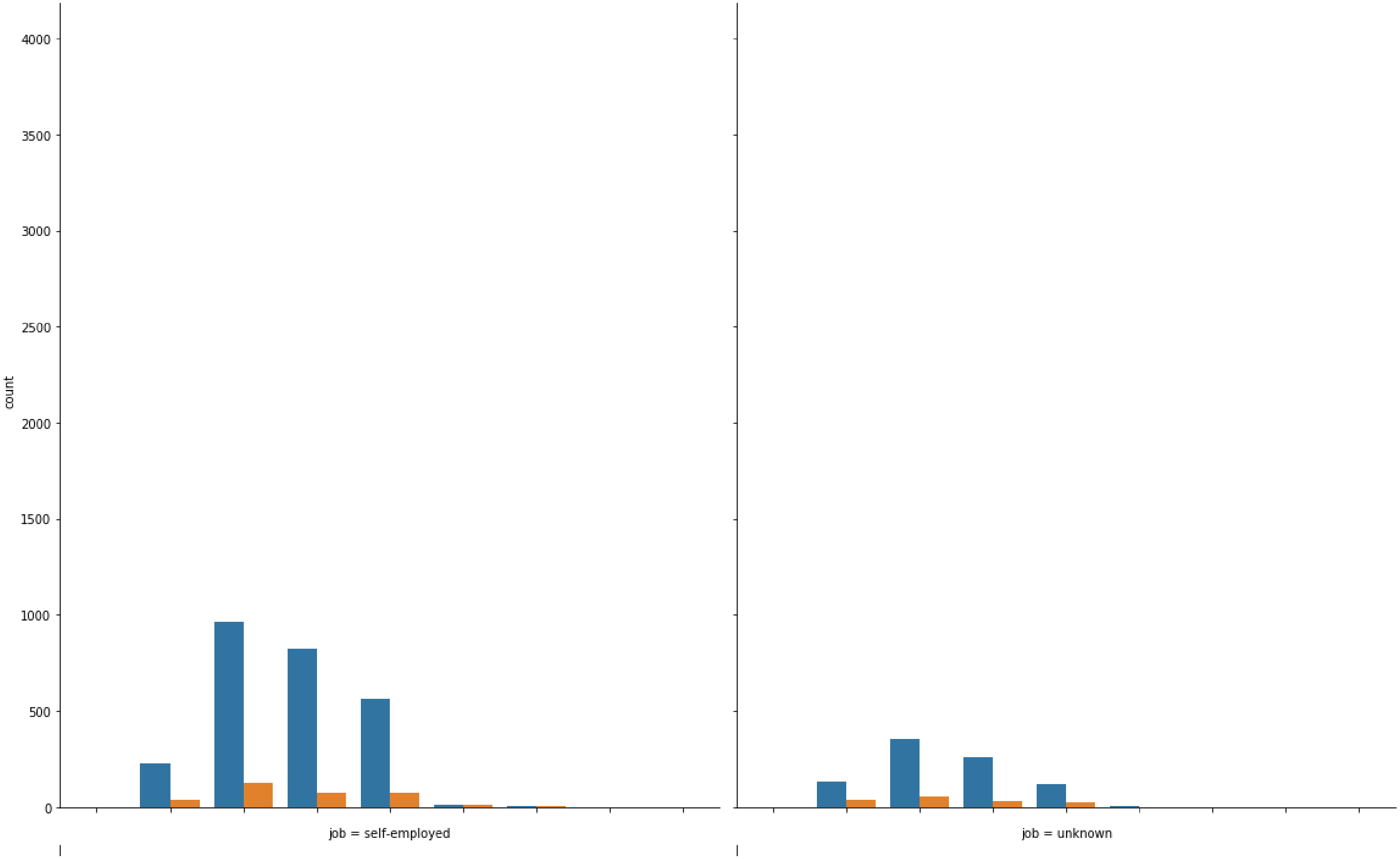
- People who have loans are in majority who have subscribed for deposits.
- They are also the ones who have not subscribed for the deposits.

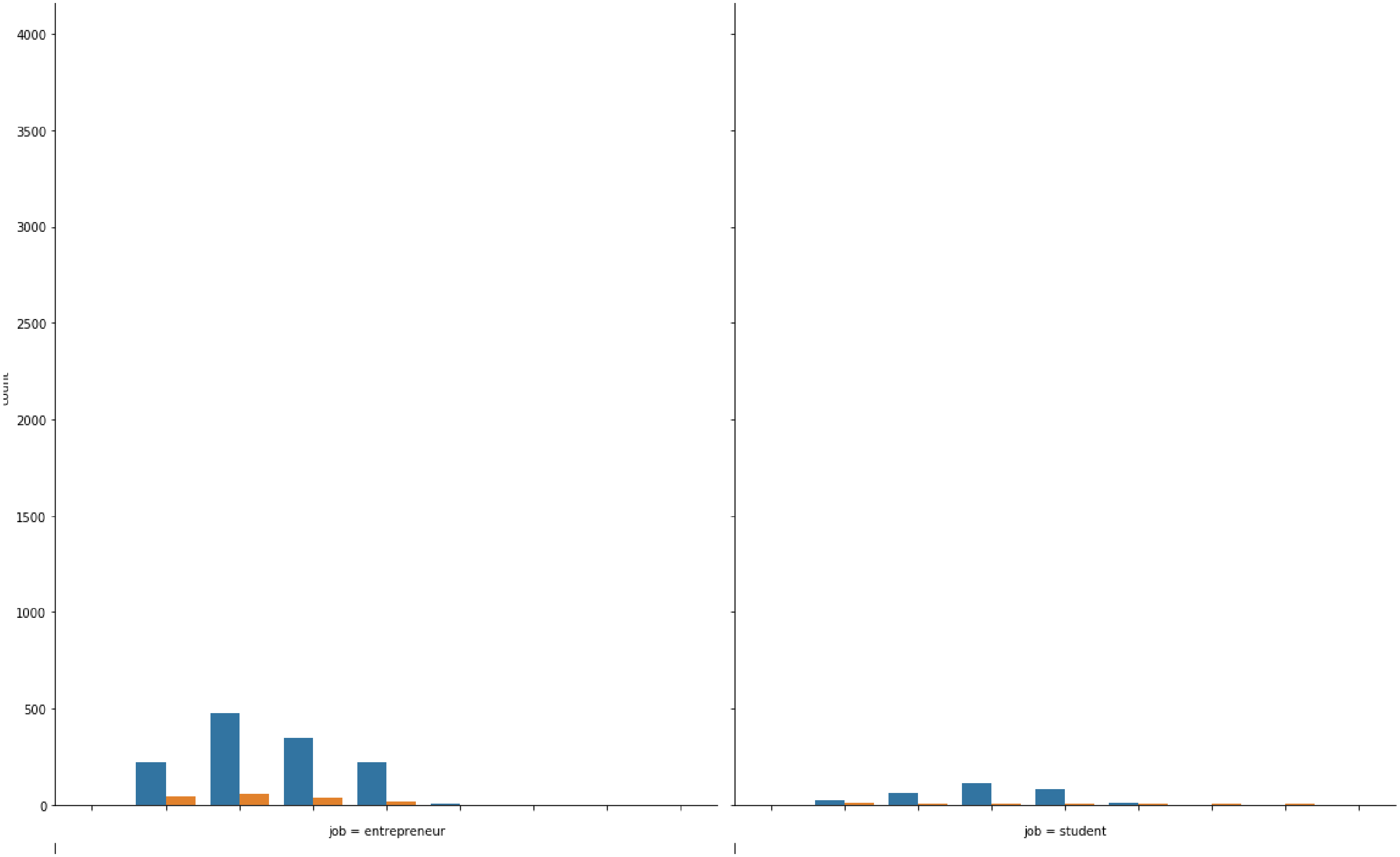
➢ Age Group and Job

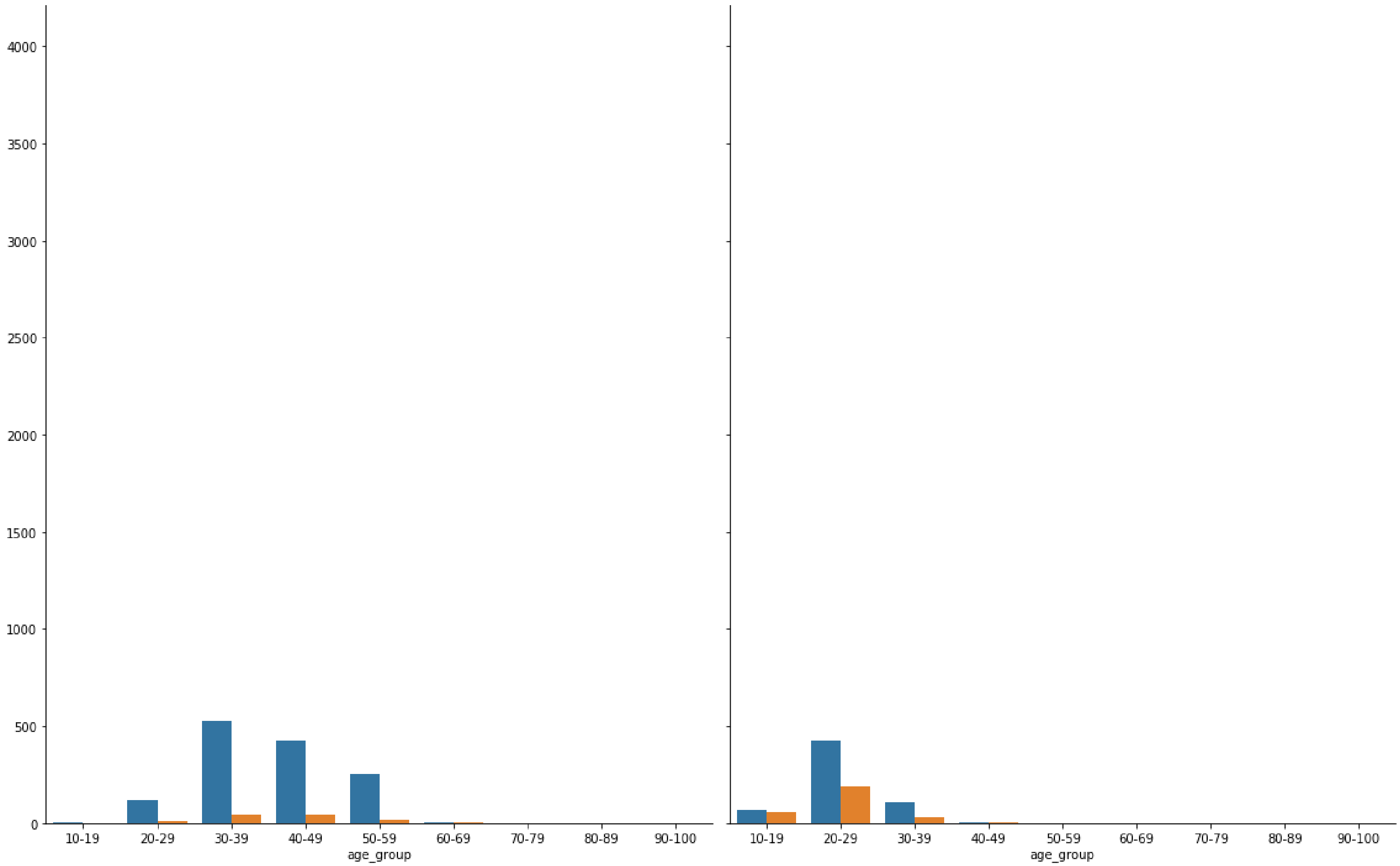



count
4000
3500
3000
2500
2000
1500
1000
500
0
job = technician
job = retired


count
4000
3500
3000
2500
2000
1500
1000
500
0
job = management
job = unemployed
y
no
yes

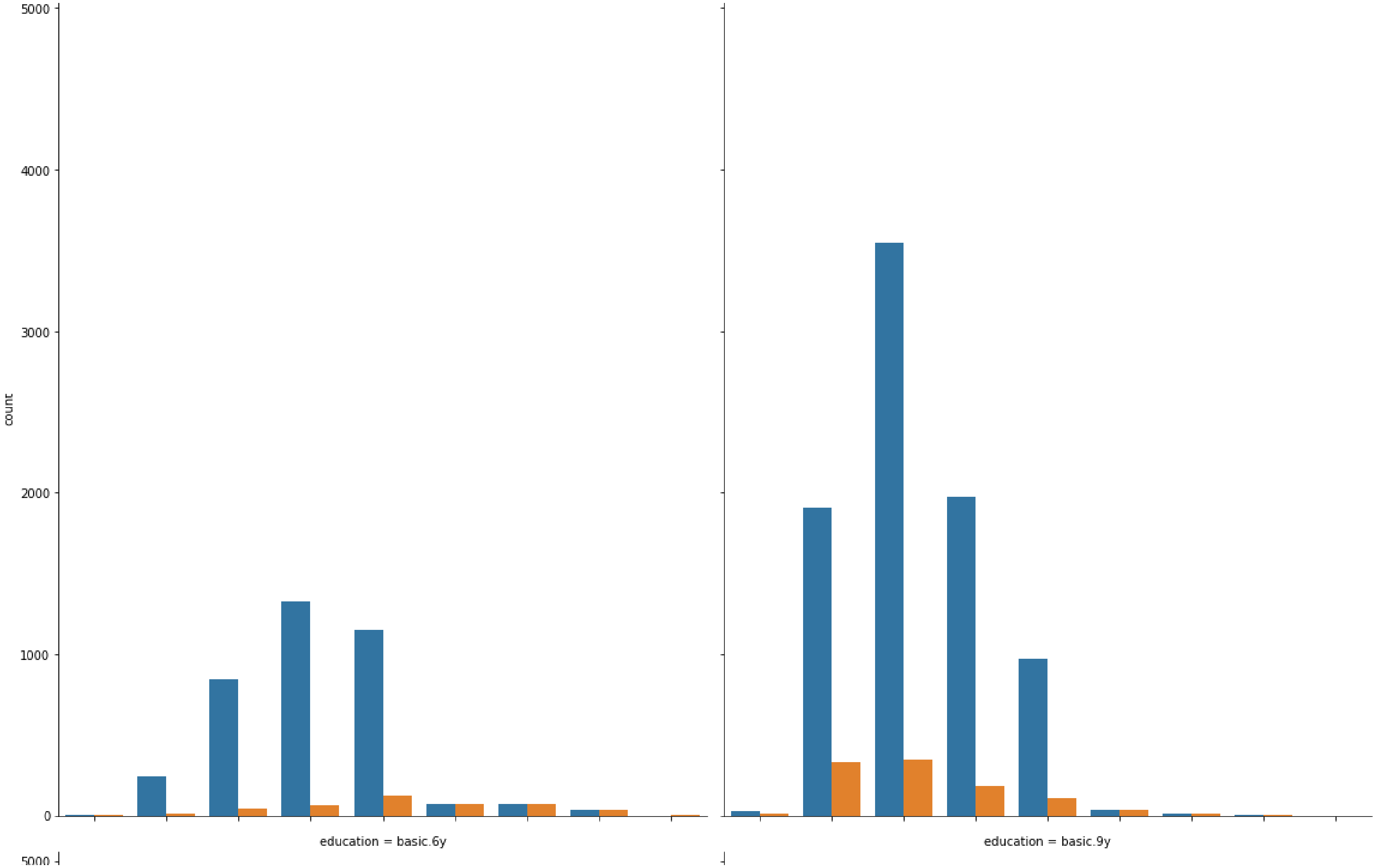

4000
3500
3000
2500
2000
1500
1000
500
0
count
job = self-employed
job = unknown

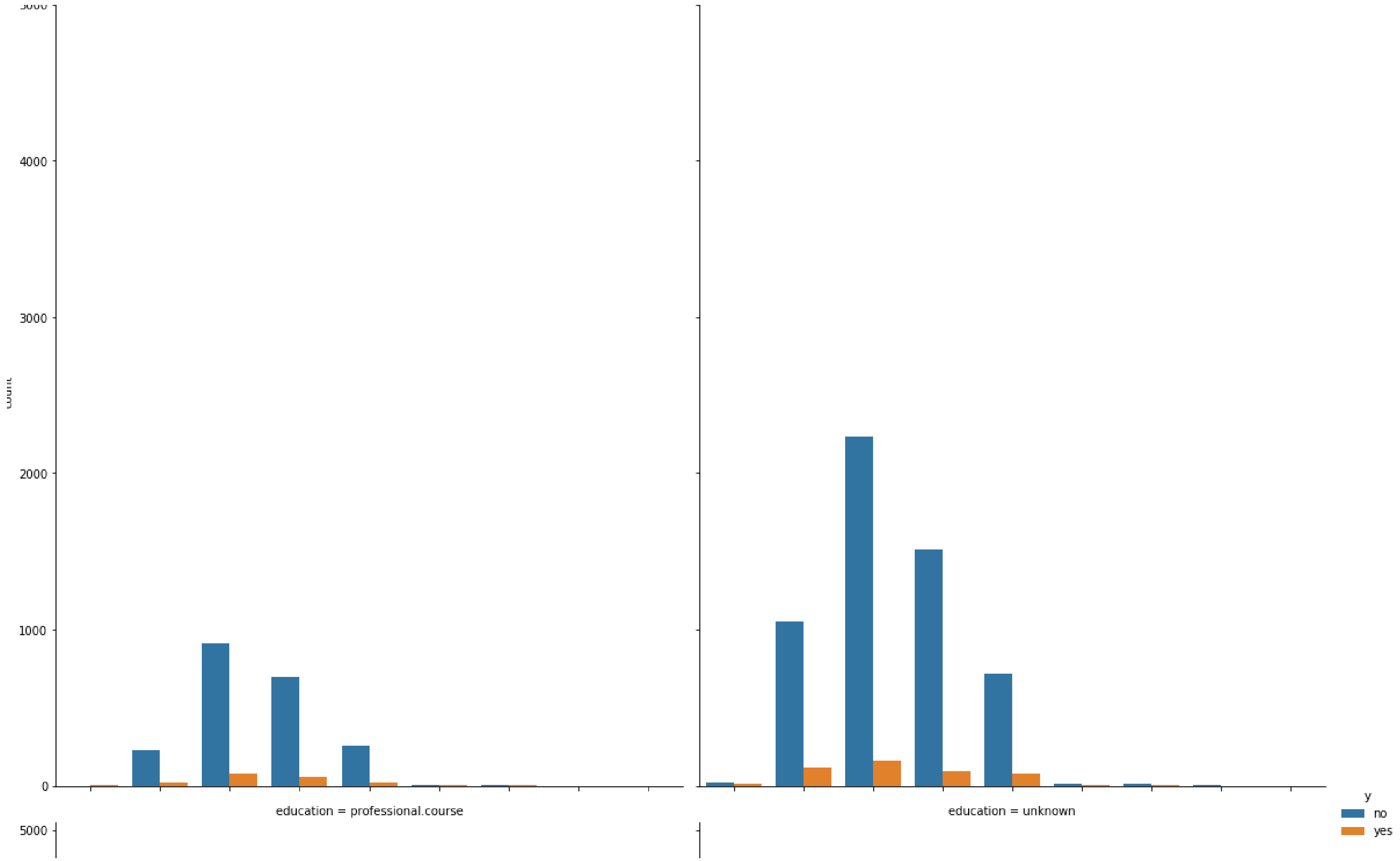

4000
3500
3000
2500
2000
1500
1000
500
0
count
job = entrepreneur
job = student

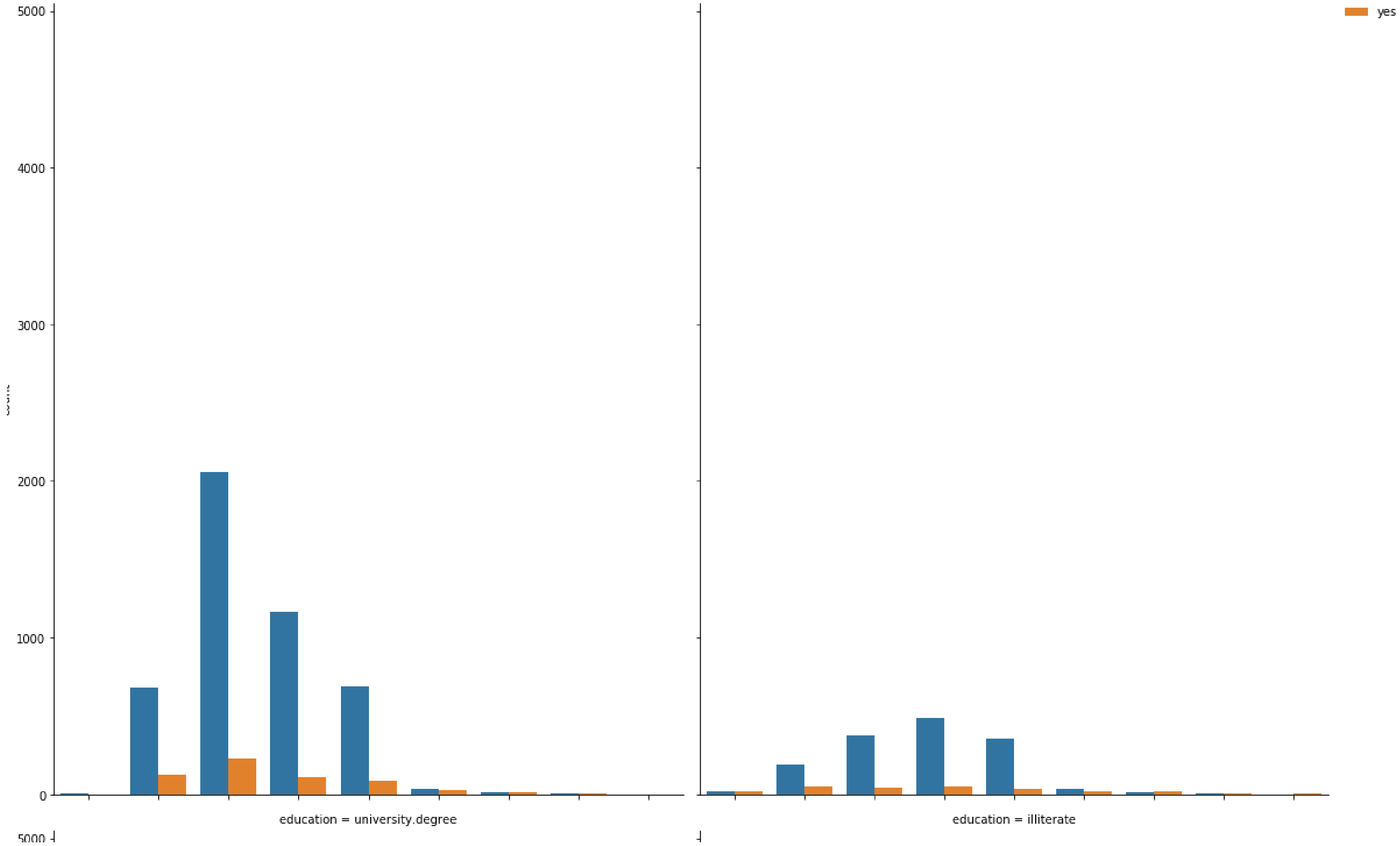

4000
3500
3000
2500
2000
1500
1000
500
0
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90-100
age_group
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90-100
age_group

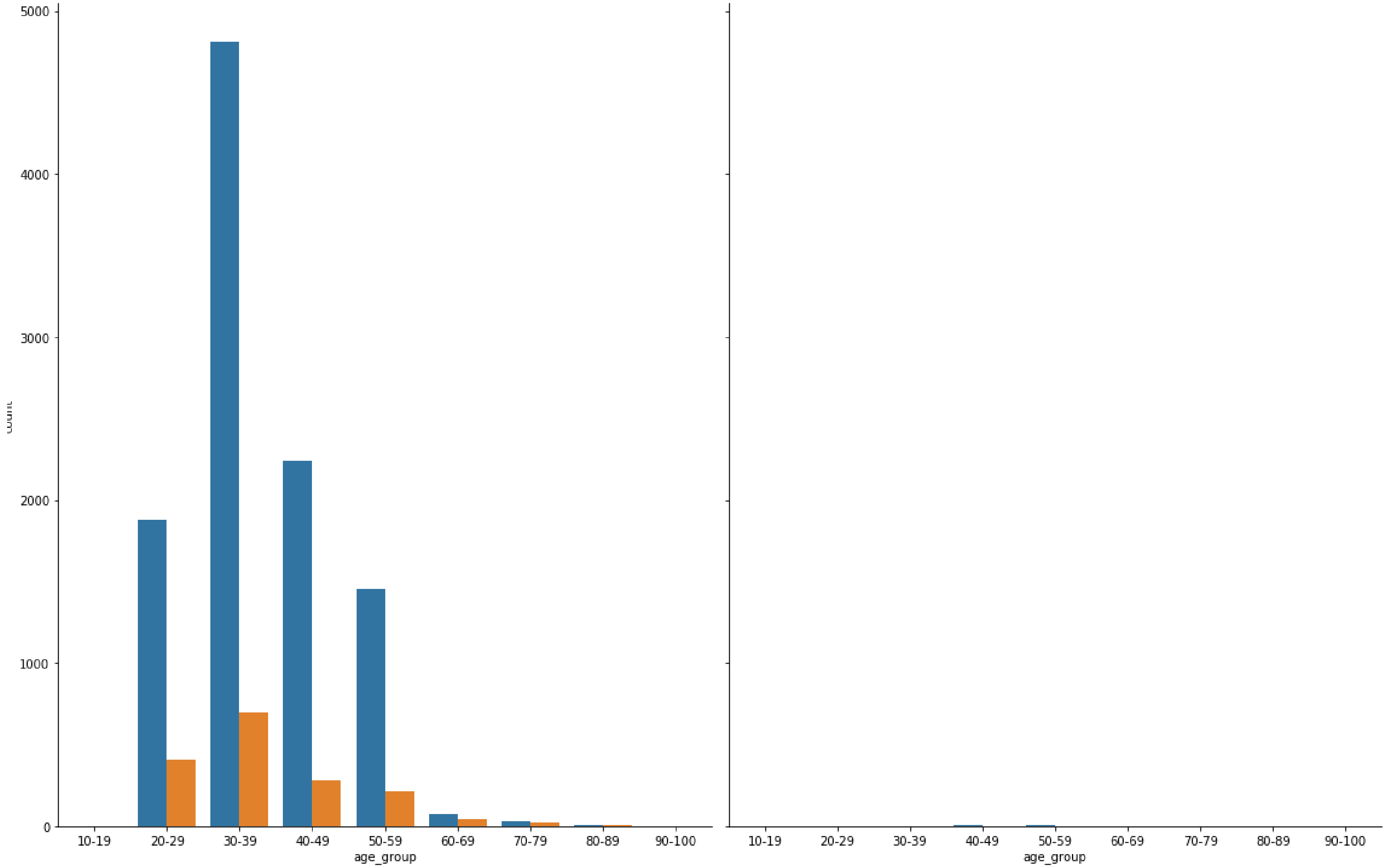
- **As shown above, people whose job professions are being admins and age group between 30-39 has the highest number for subscribing for deposits, followed by the age groups between 20-29 and 40-49 in the same profession.**

➢ Age Group and Education


5000
4000
3000
2000
1000
0
count
education = basic.6y
education = basic.9y
5000

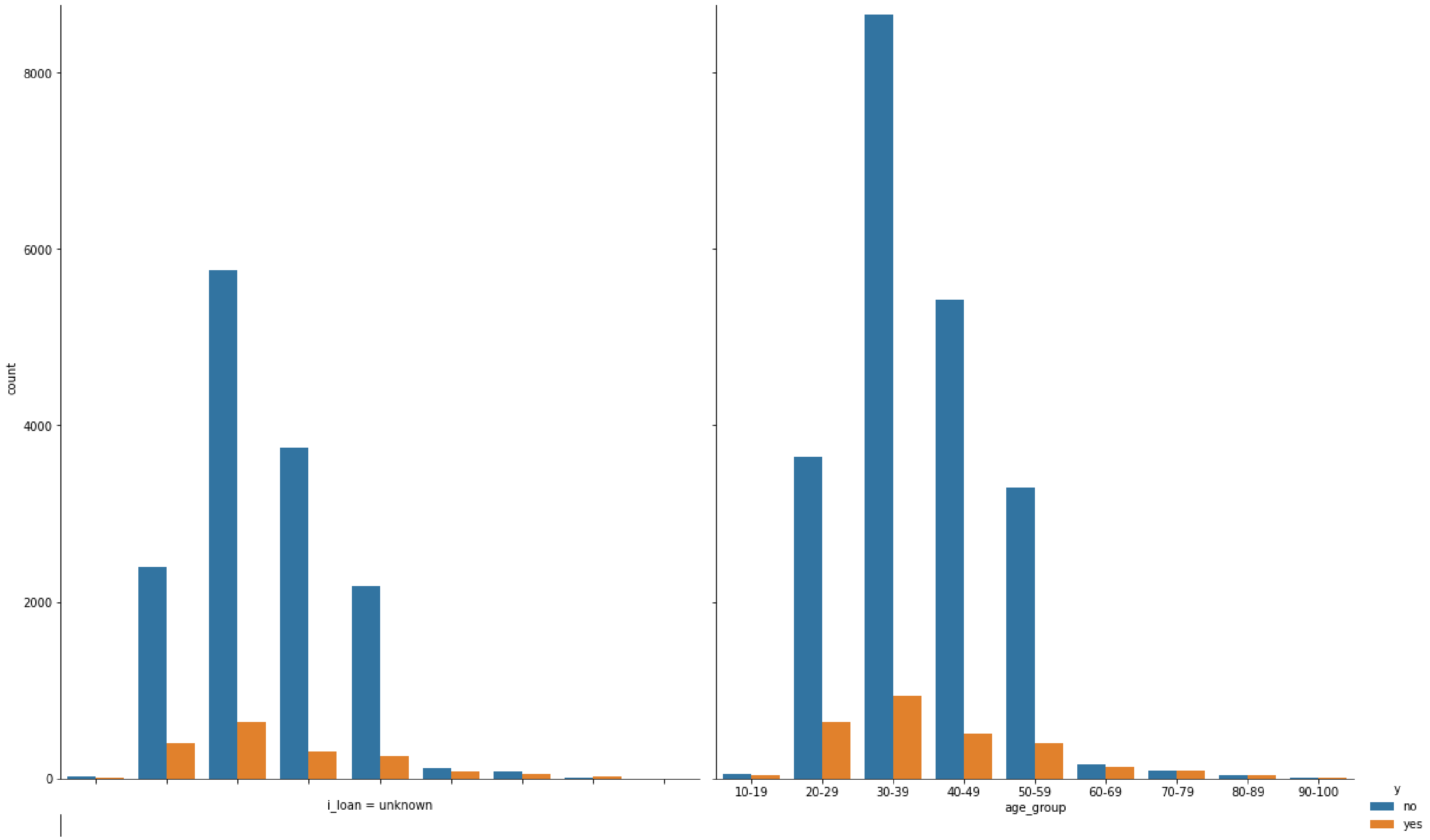

5000
4000
3000
2000
1000
0
count
education = professional.course
education = unknown
y
no
yes
5000

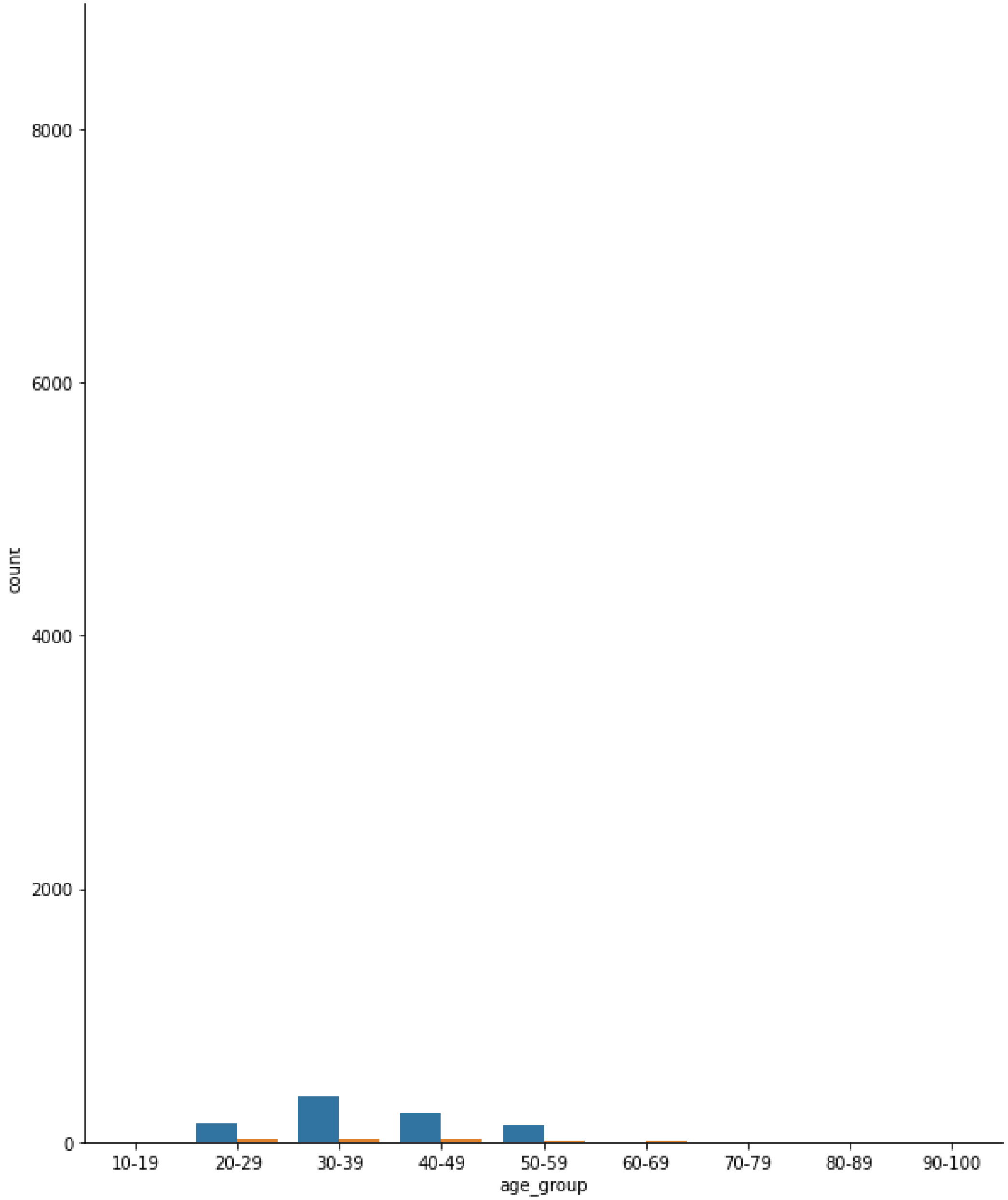

5000
4000
3000
2000
1000
0
education = university.degree
education = illiterate
yes
5000


5000
4000
3000
2000
1000
0
count
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90-100
age_group
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90-100
age_group

- As shown above, people who are in age group of 30-39 and have completed university are the most ones who have and have not subscribed for deposits.

- **Age Group and I_loan**

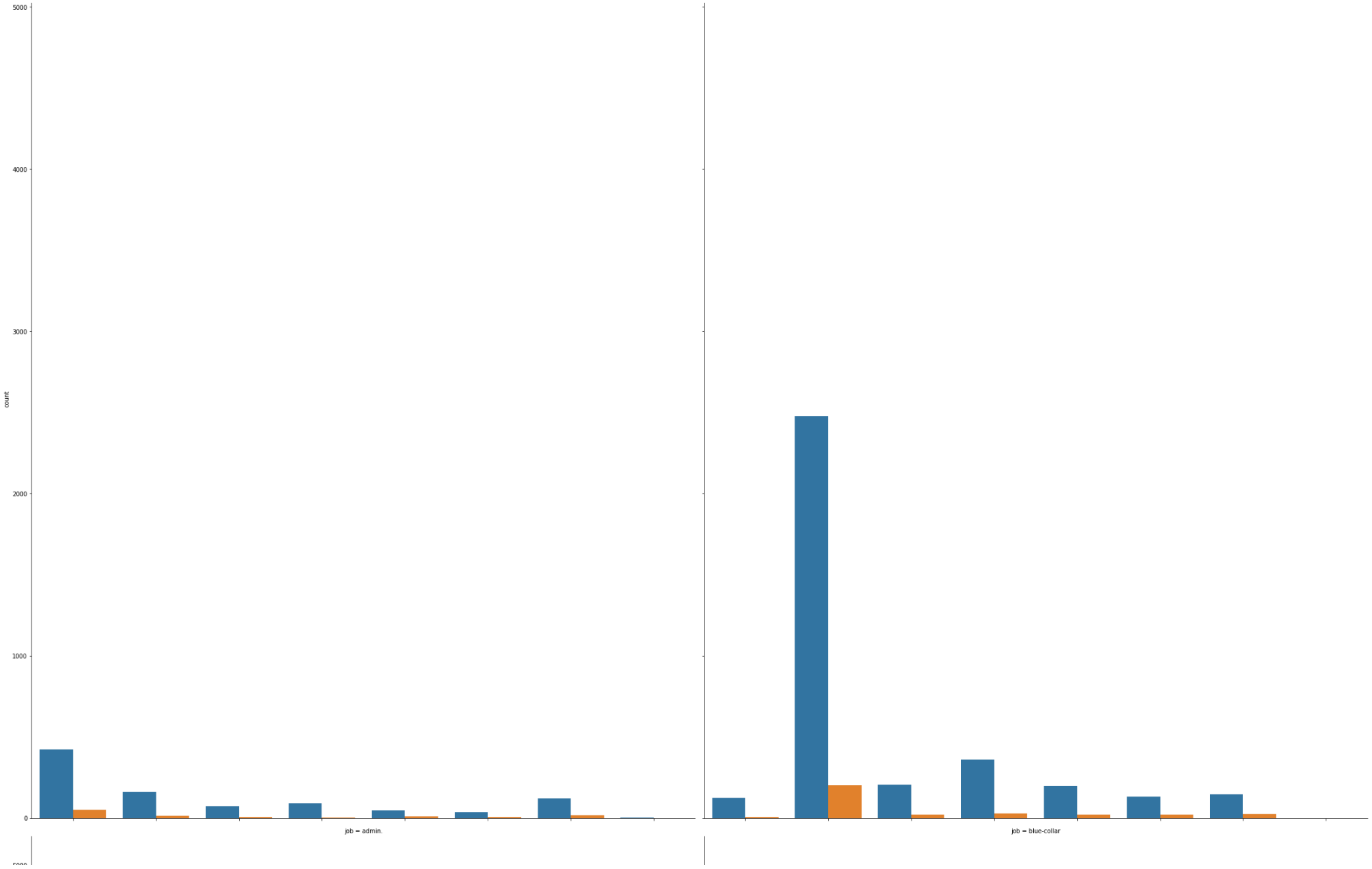

8000
6000
4000
2000
0
count
i_loan = unknown
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90-100
age_group
y
no
yes

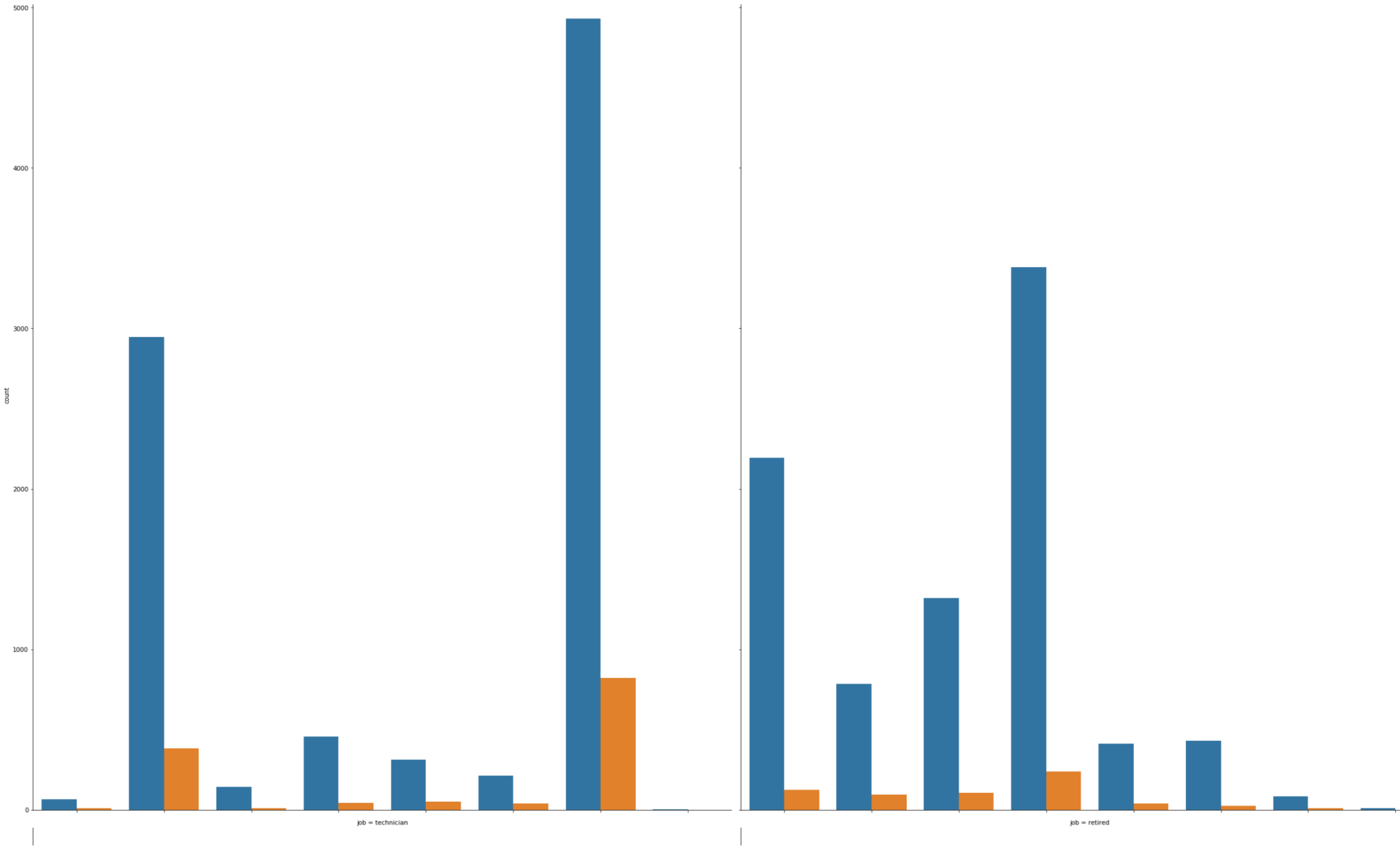


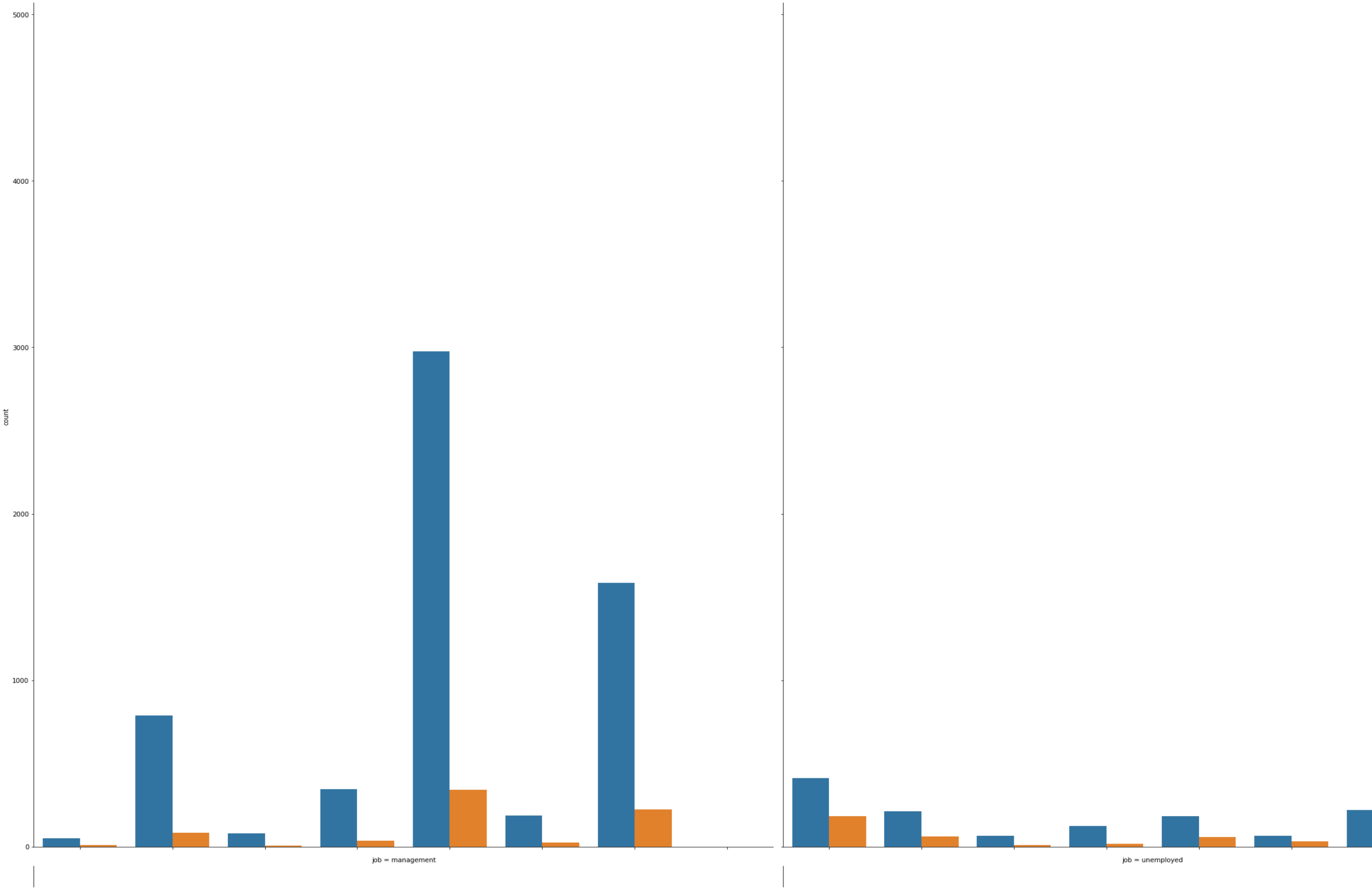
count
8000
6000
4000
2000
0
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90-100
age_group

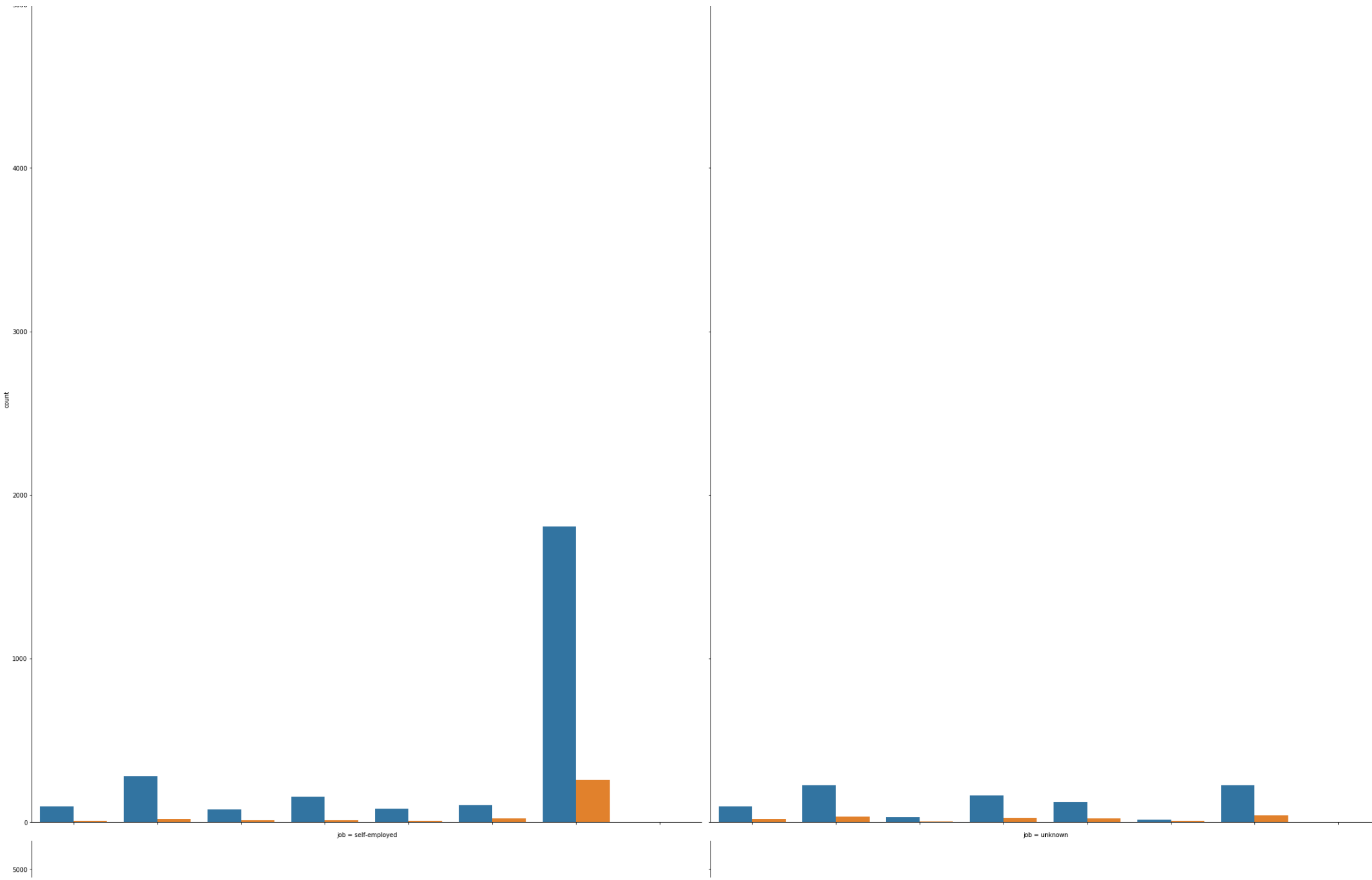
- People who are in age group of 30-39 and have loans are the most ones who have subscribed for long term deposits.
- They are also the most ones who have not subscribed for it.

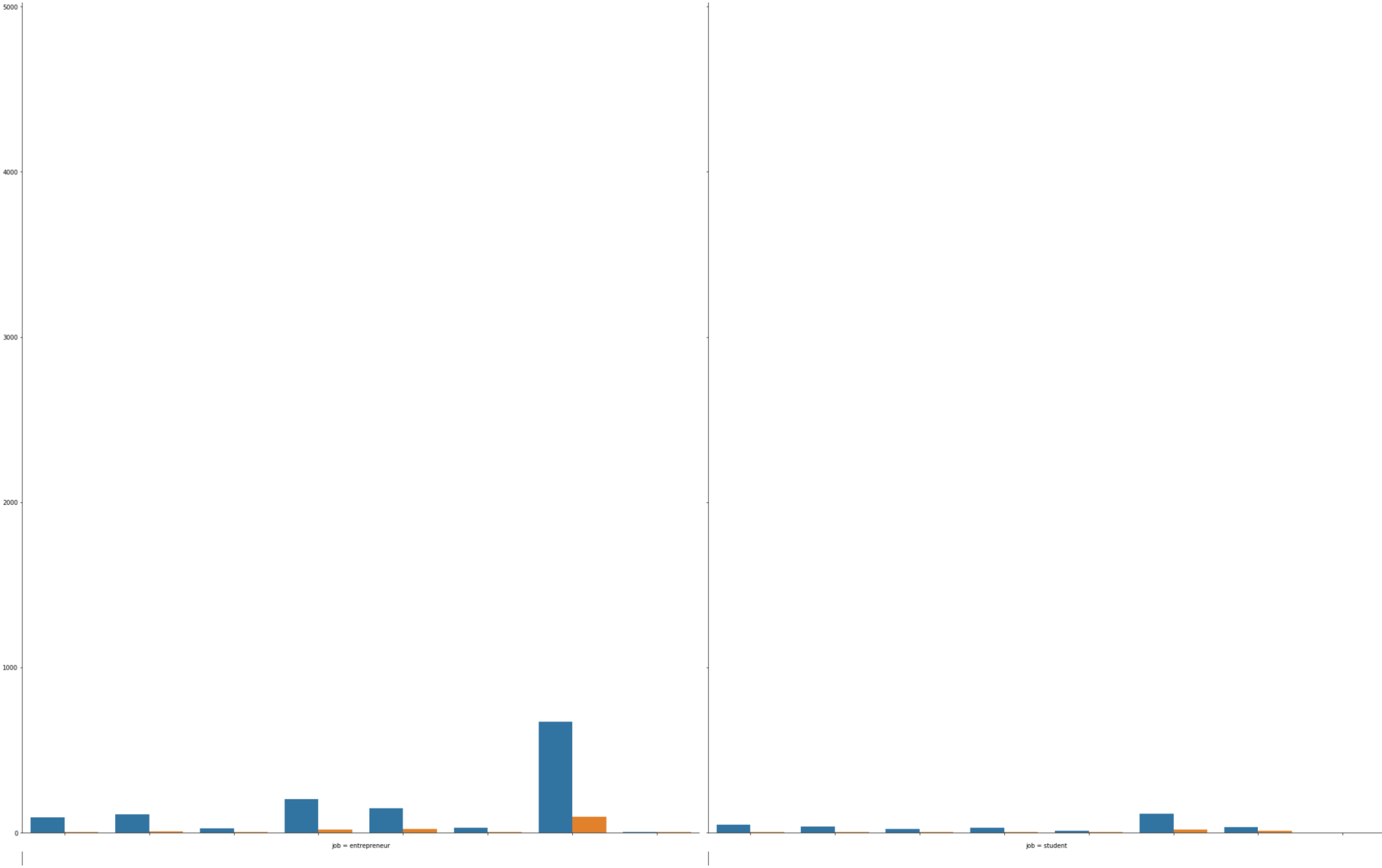
➢ Education and Job

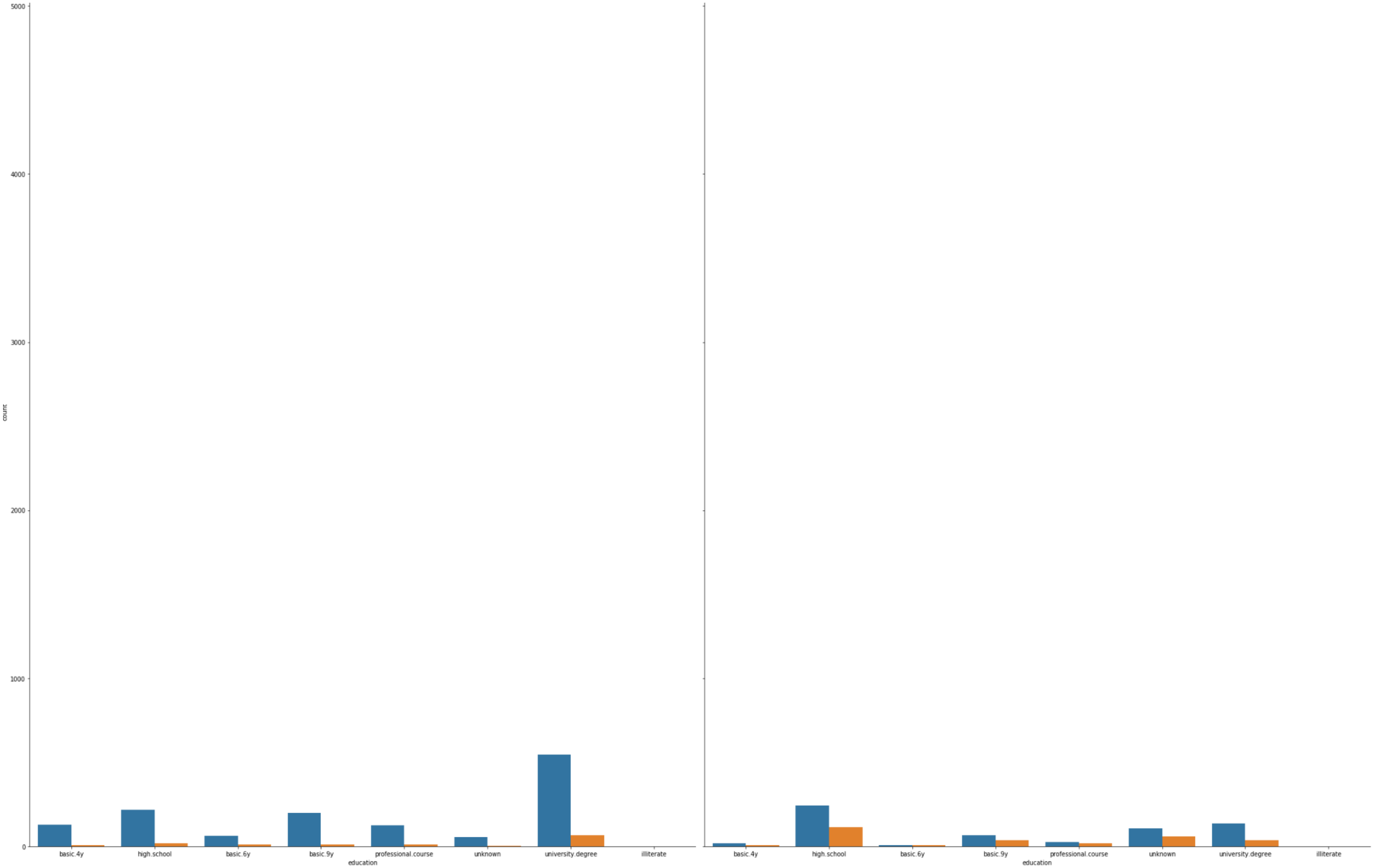

5000
4000
3000
2000
1000
0
count
job = admin.
job = blue-collar
5000


5000
4000
3000
2000
1000
0
count
job = technician
job = retired


count
5000
4000
3000
2000
1000
0
job = management
job = unemployed

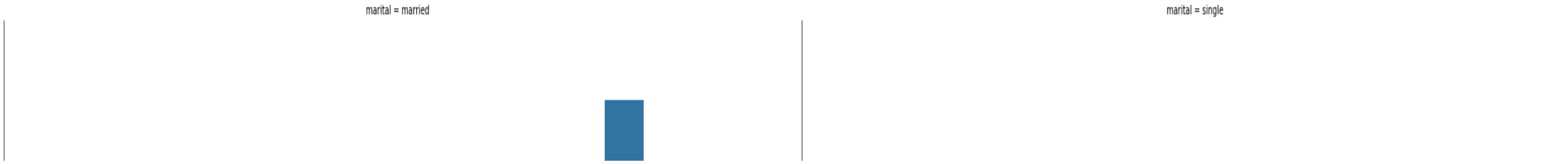

5000
4000
3000
2000
1000
0
count
job = self-employed
job = unknown
5000

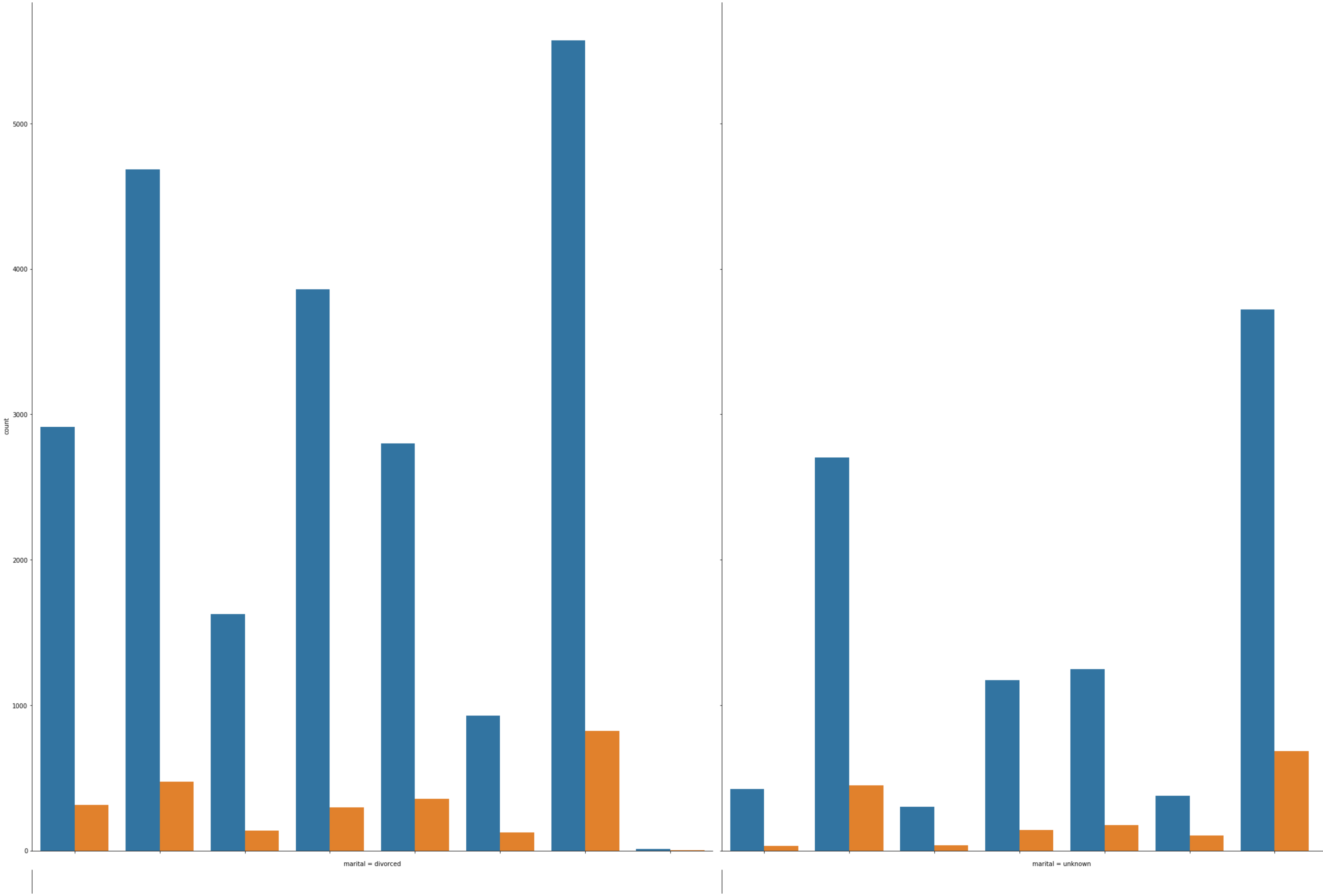

5000
4000
3000
2000
1000
0
job = entrepreneur
job = student

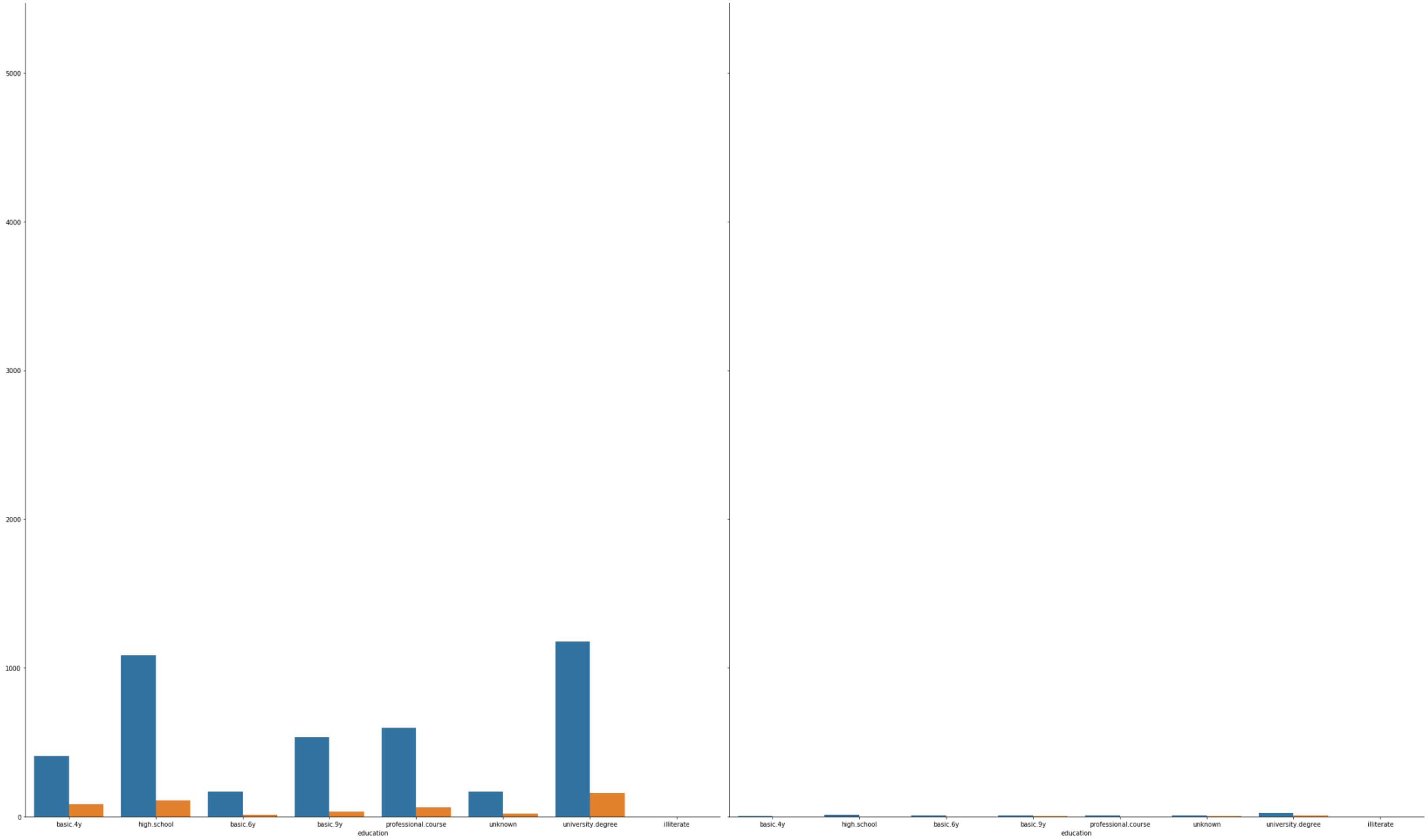

5000
4000
3000
2000
1000
0
count
basic.4y
high.school
basic.6y
basic.9y
professional.course
unknown
university.degree
illiterate
education
basic.4y
high.school
basic.6y
basic.9y
professional.course
unknown
university.degree
illiterate
education

- People who are admins and have completed university are the most ones who have subscribed for long term deposits.
- They are also the most ones who have not subscribed for long term deposits.

➢ Education and Marital



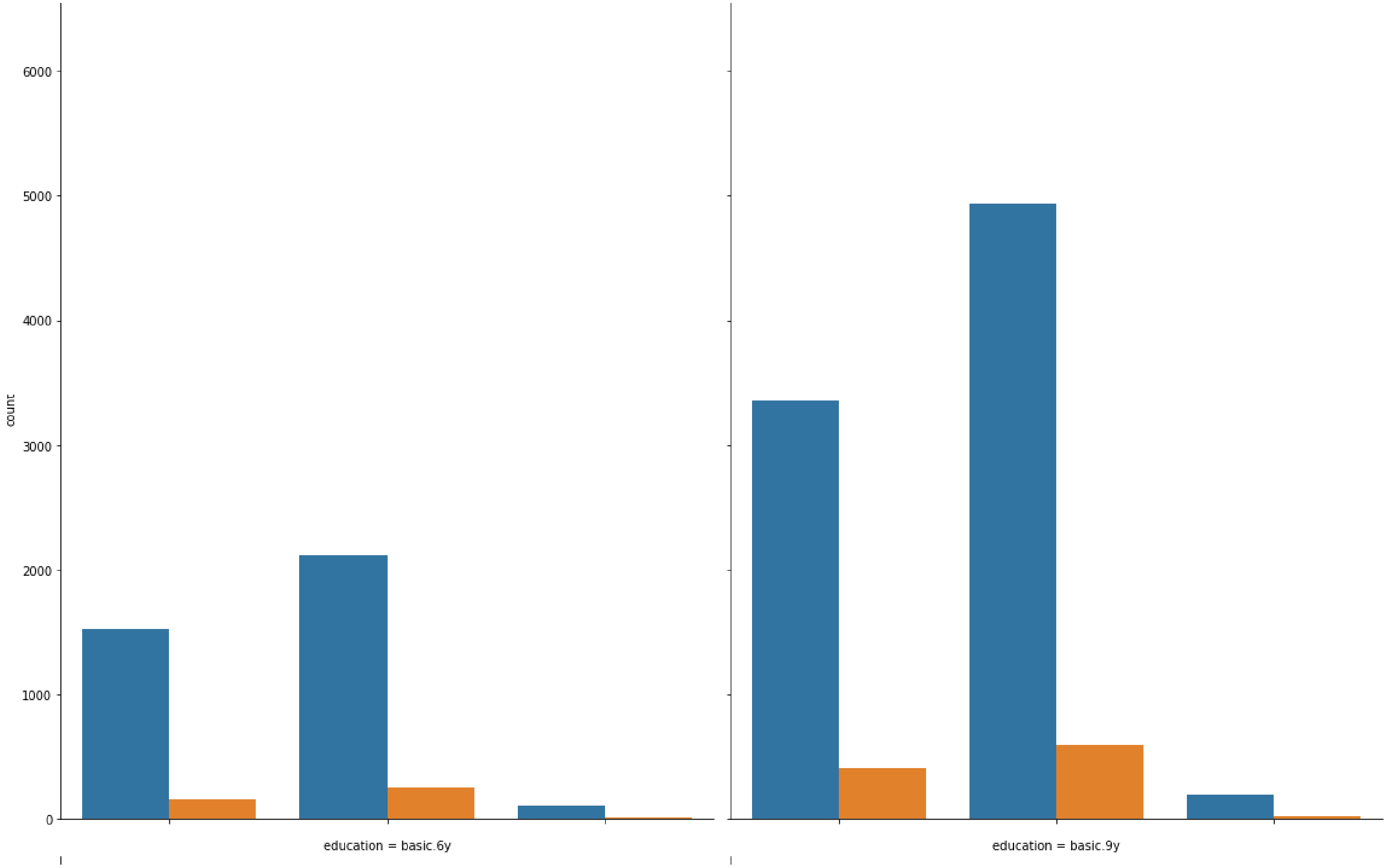

count
5000
4000
3000
2000
1000
0
marital = divorced
marital = unknown

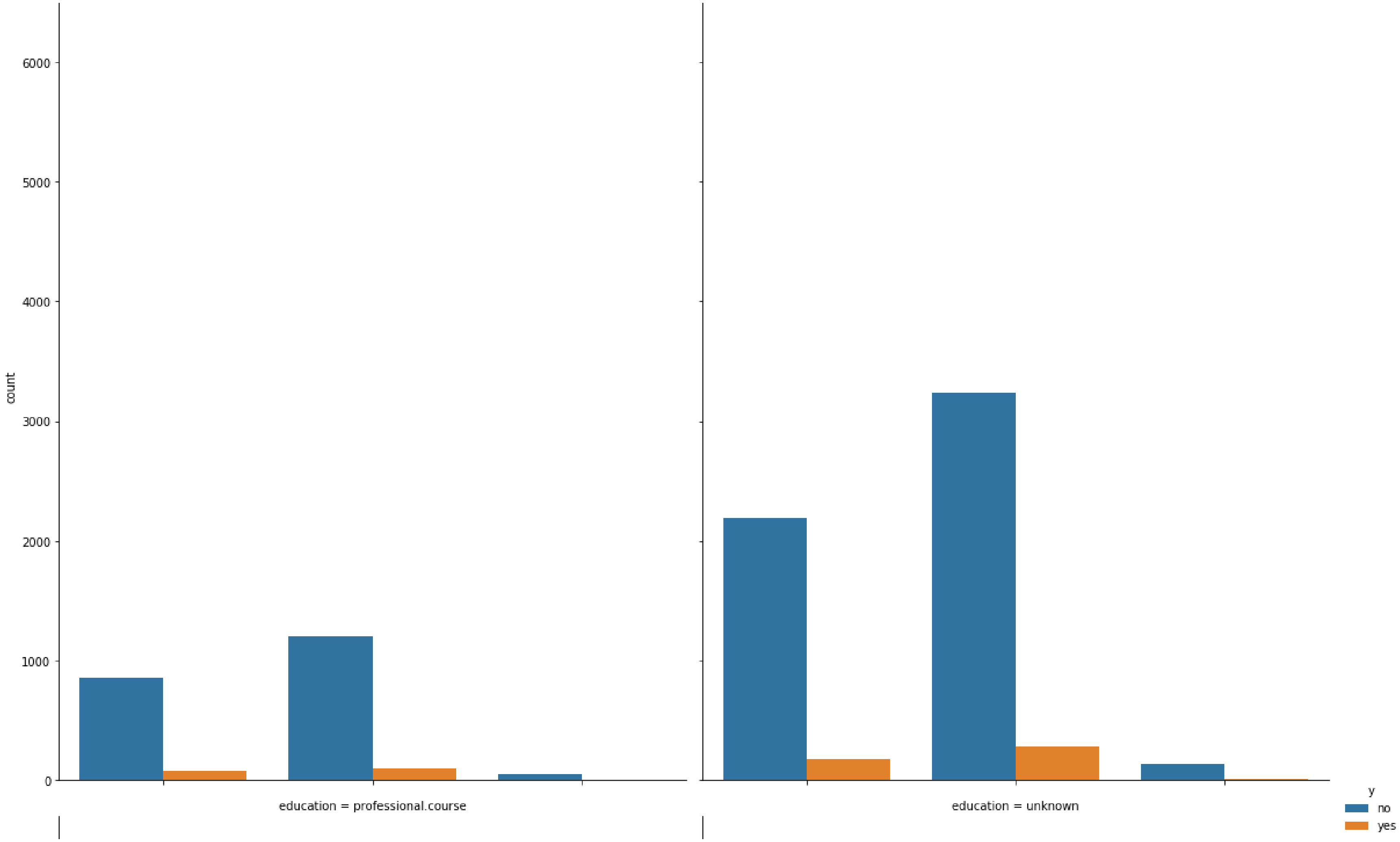


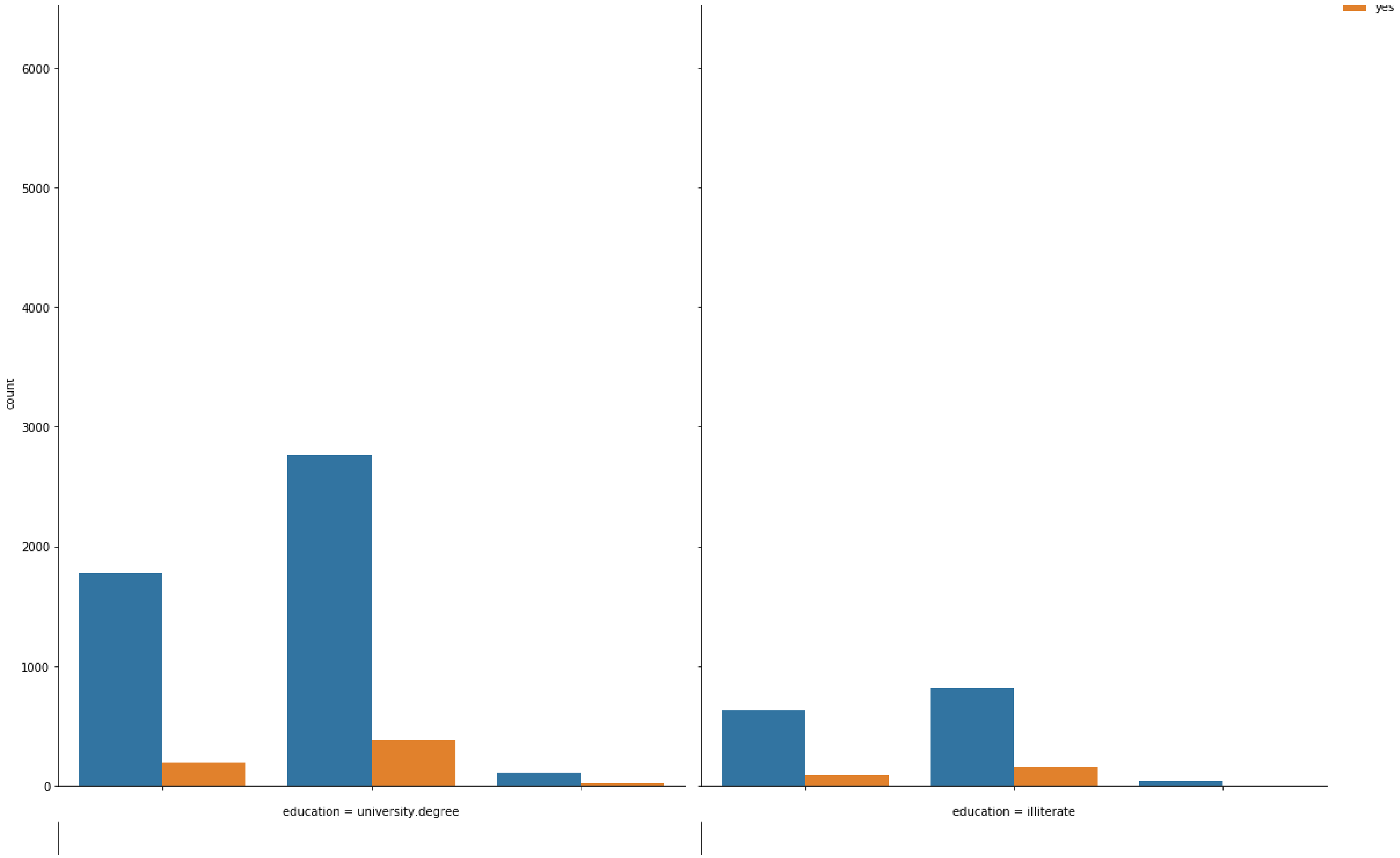
- People who have done university and are either single or married are the most ones who have subscribed for the deposits.

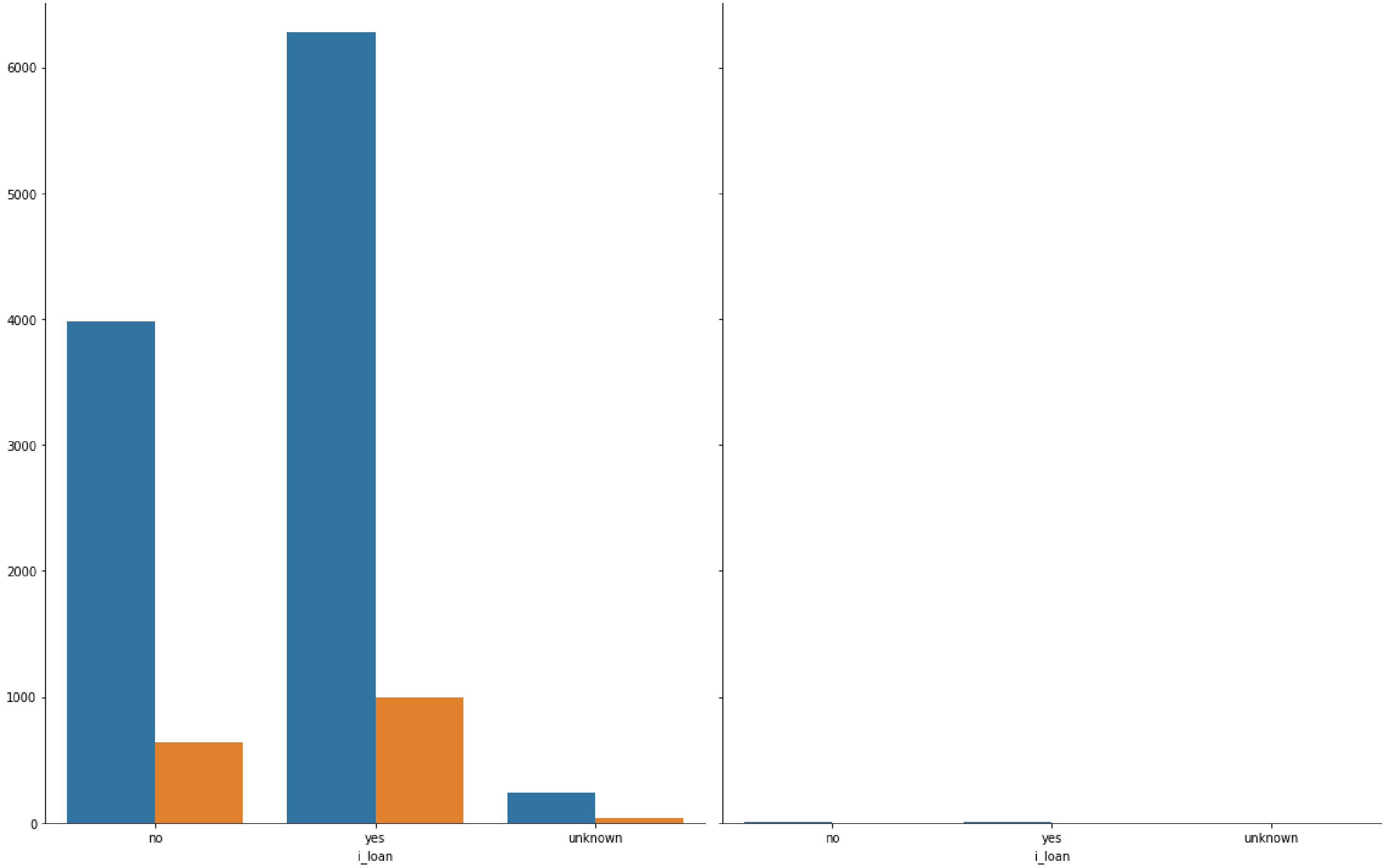
- People who are married and have done university are the most ones who have not subscribed for long term deposits.

➢ I_loan and Education


count
6000
5000
4000
3000
2000
1000
0
education = basic.6y
education = basic.9y


count
6000
5000
4000
3000
2000
1000
0
education = professional.course
education = unknown
y
no
yes


yes
count
6000
5000
4000
3000
2000
1000
0
education = university.degree
education = illiterate

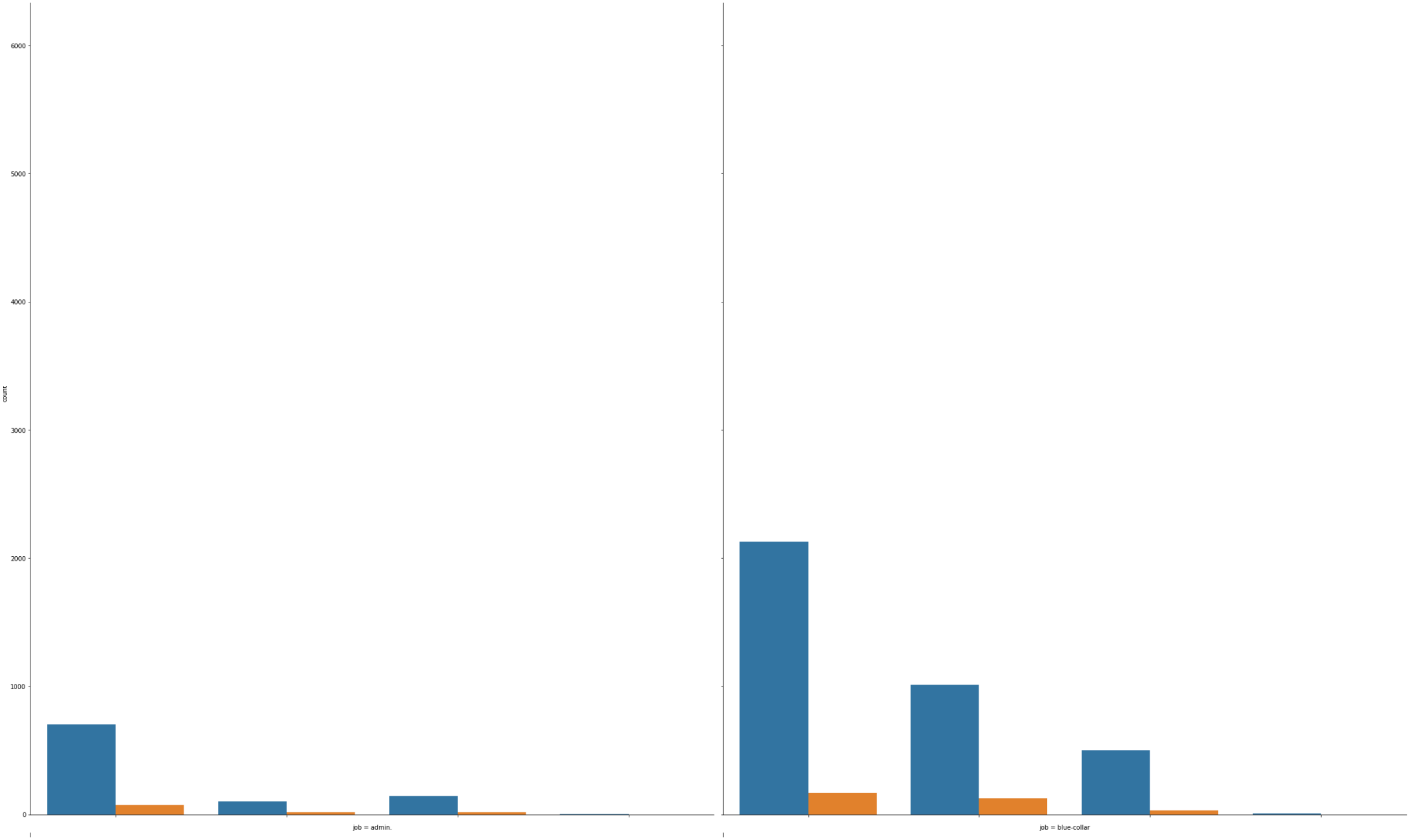


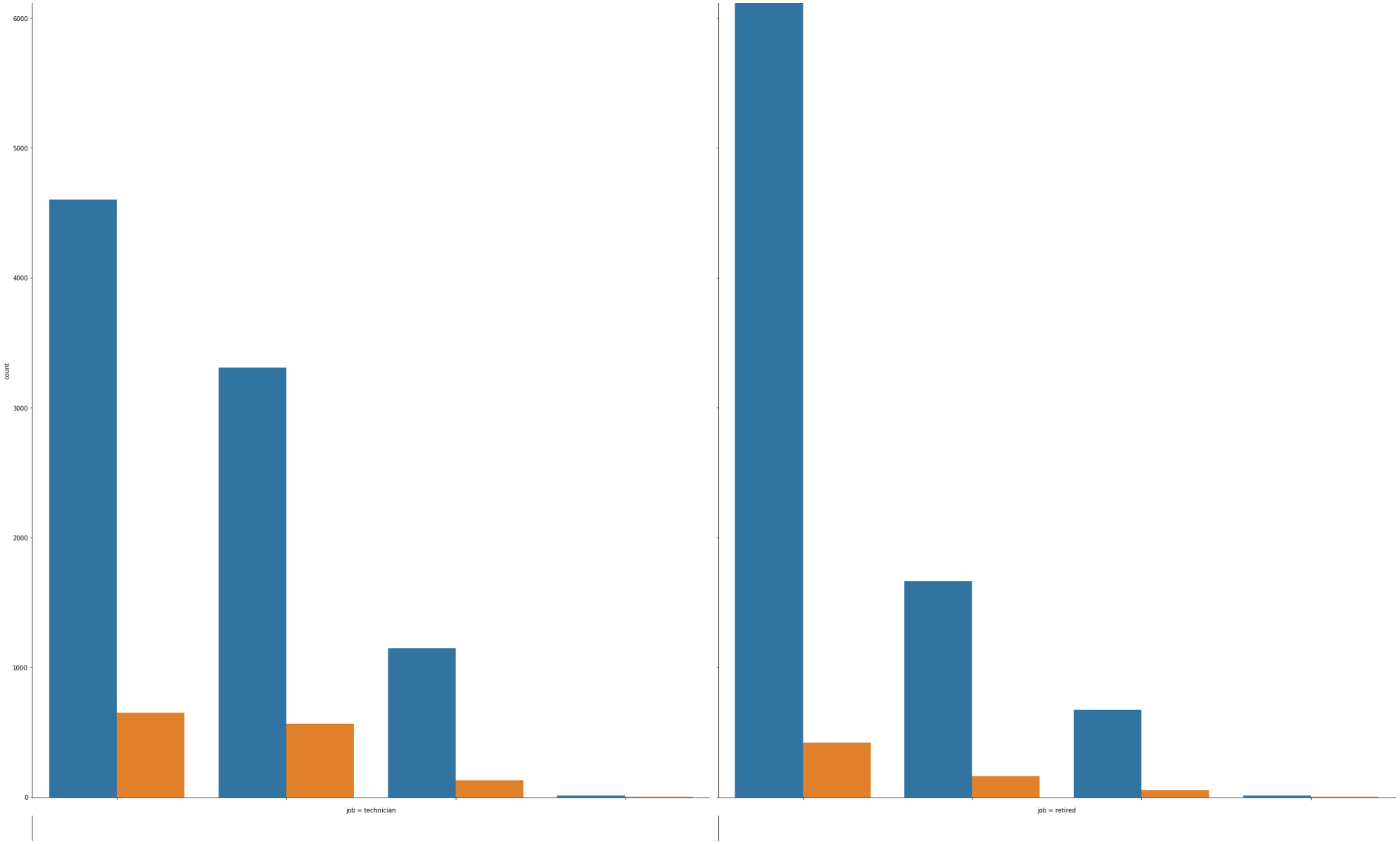
- People who are in university and have loans are the most ones who have subscribed for deposits.
- They are also the most people who have not subscribed for their deposits.

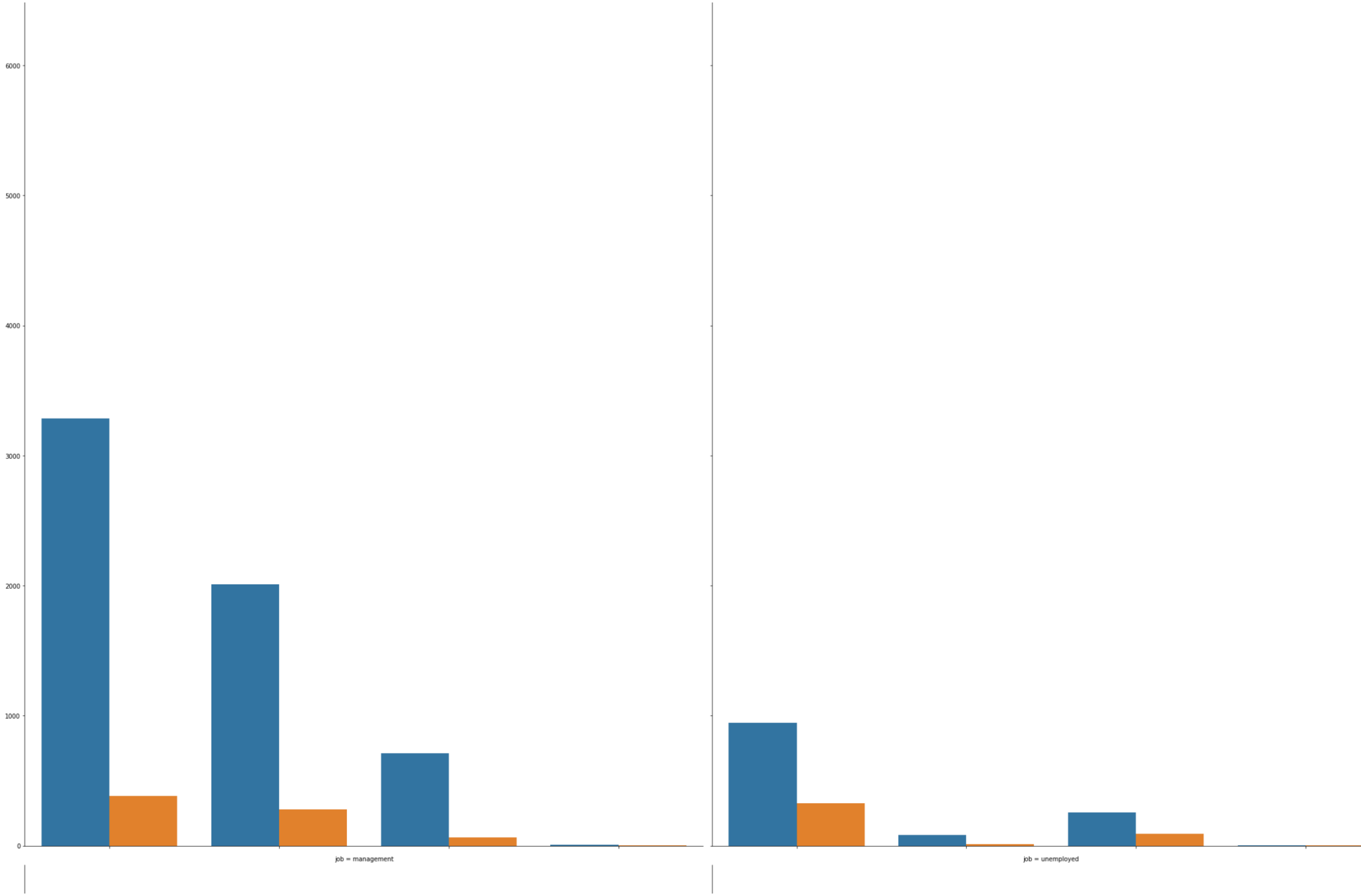
- **Marital and Job**

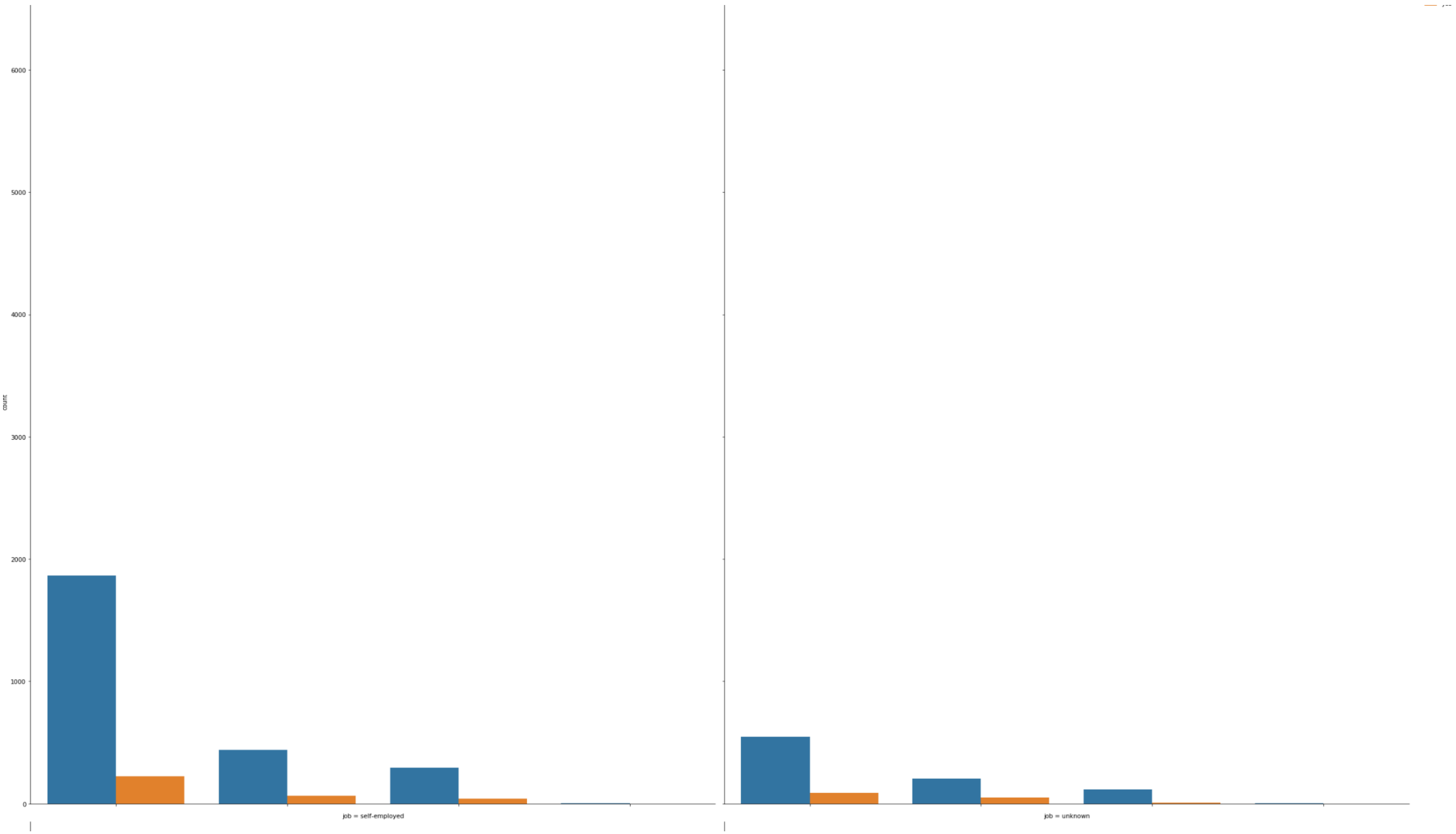
job = housemaid

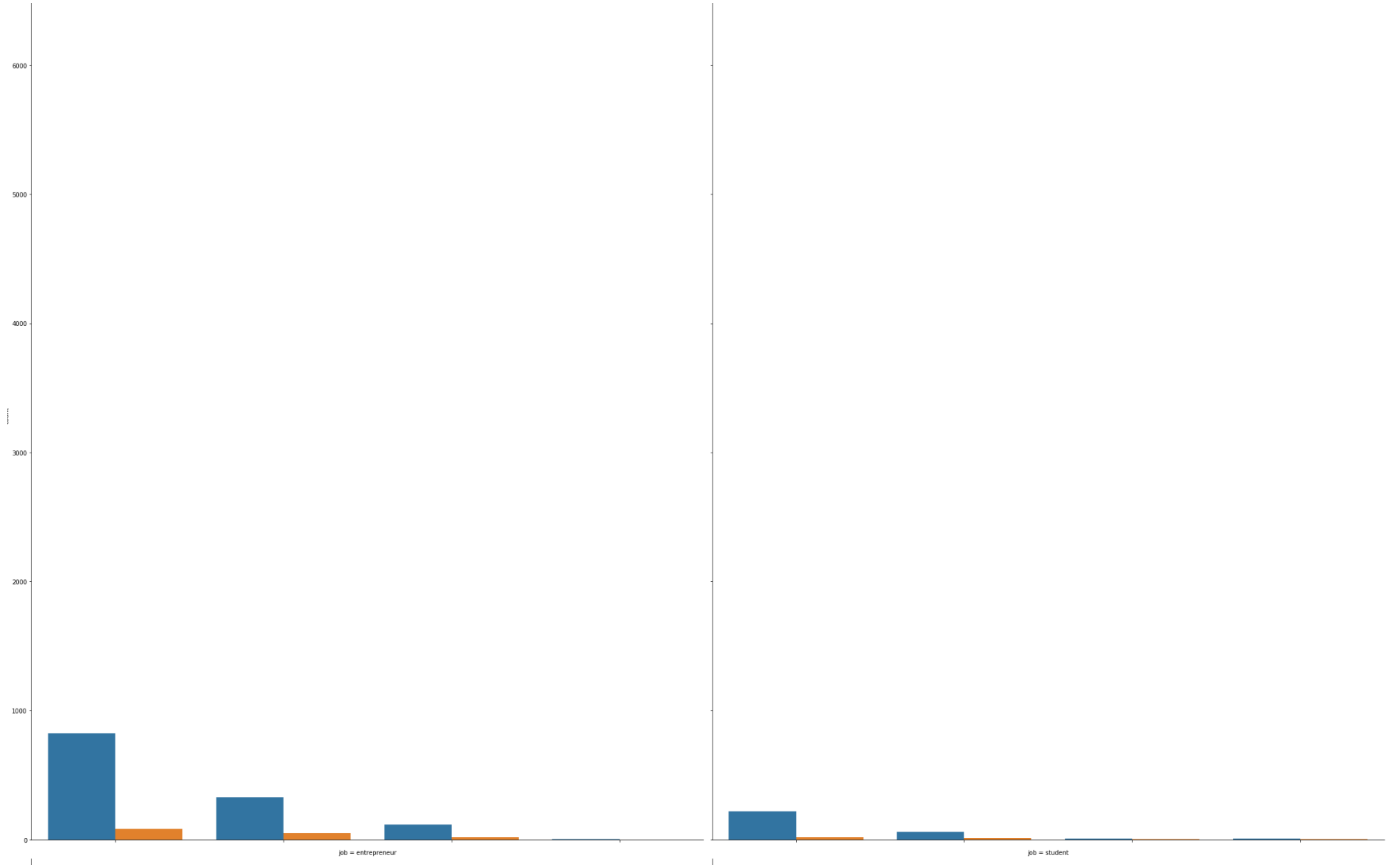
job = services

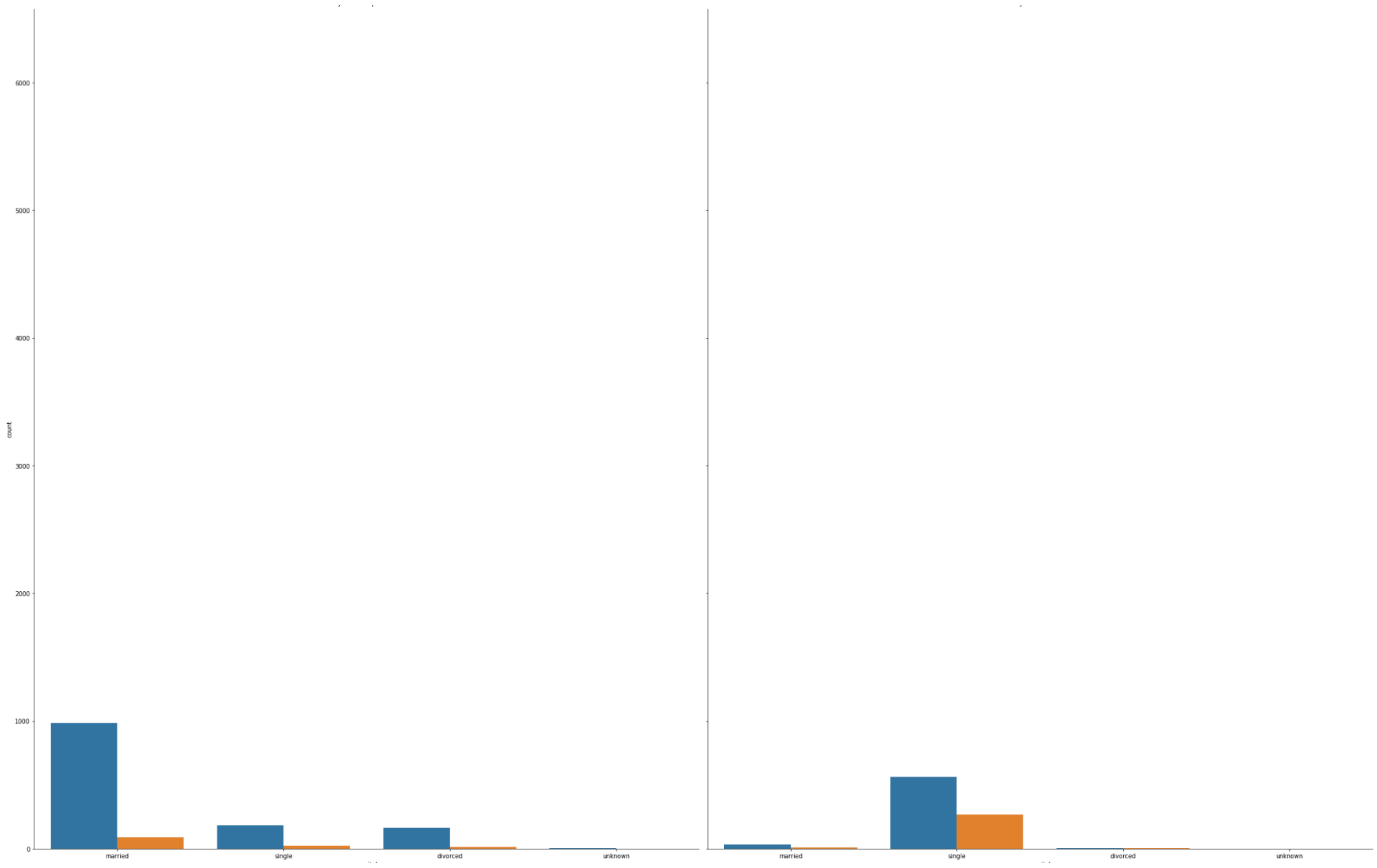

count
6000
5000
4000
3000
2000
1000
0
job = admin.
job = blue-collar


6000
5000
4000
3000
2000
1000
0
count
job = technician
job = retired


6000
5000
4000
3000
2000
1000
0
job = management
job = unemployed

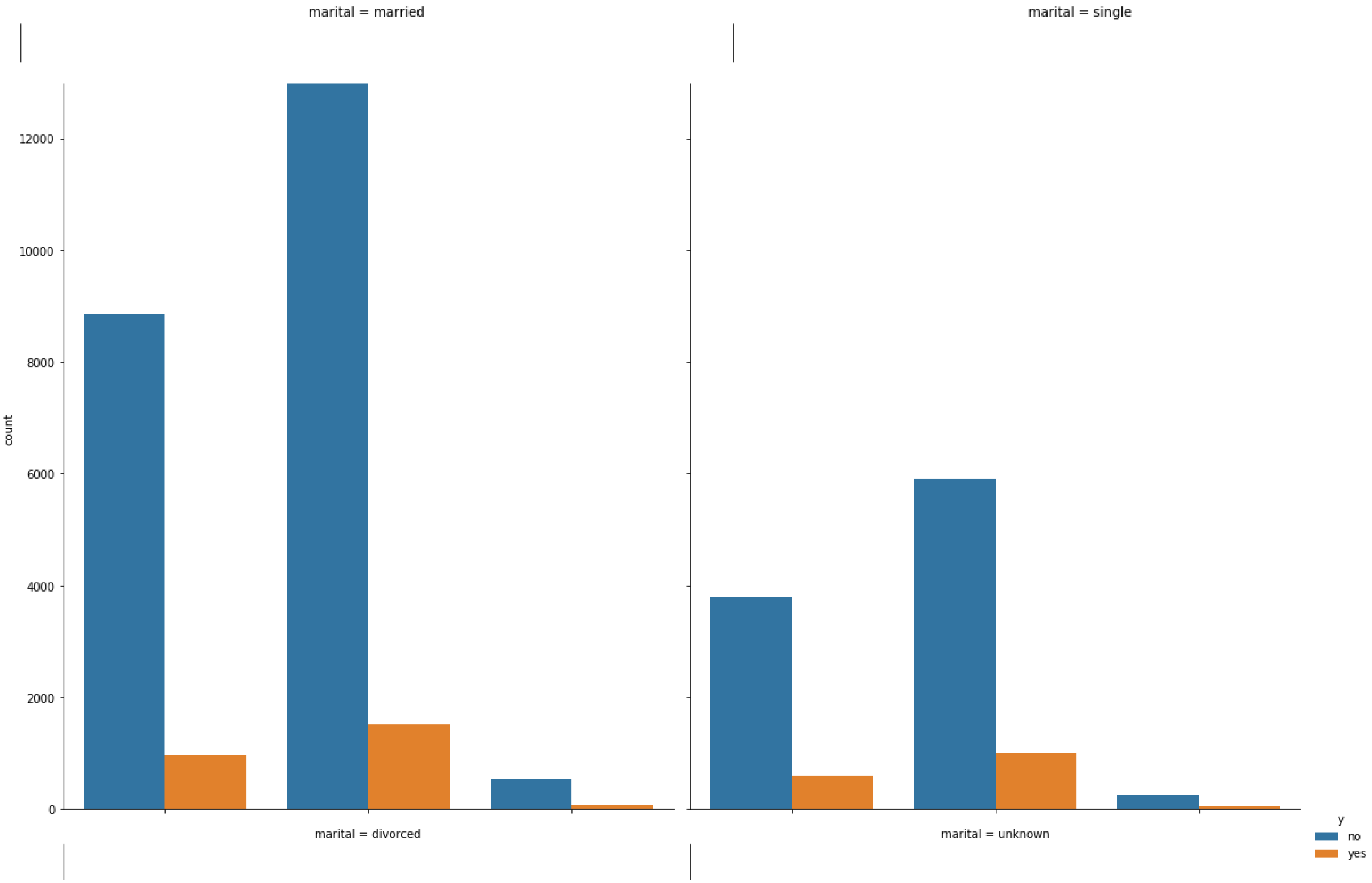

count
6000
5000
4000
3000
2000
1000
0
job = self-employed
job = unknown

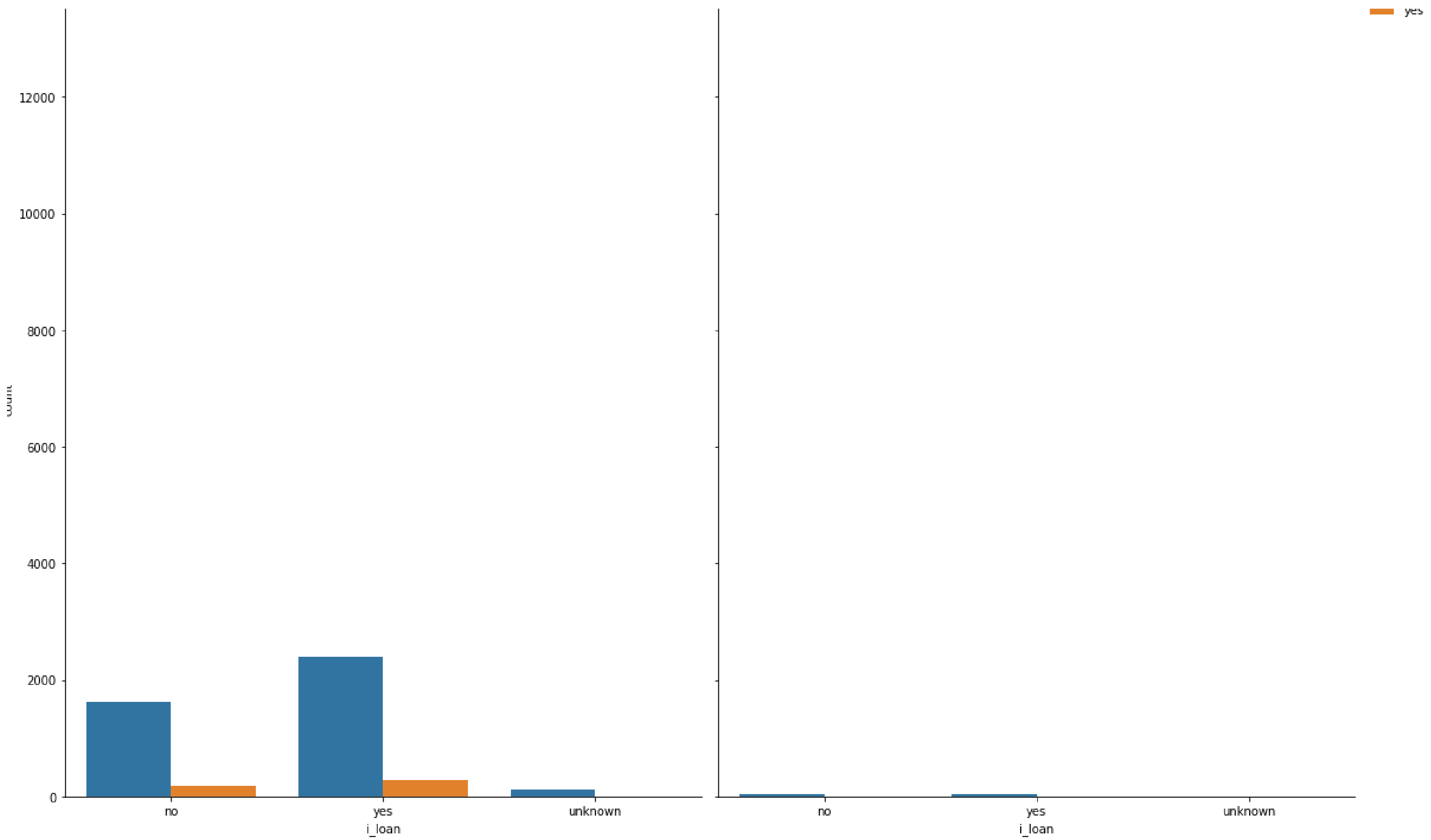

count
0
1000
2000
3000
4000
5000
6000
job = entrepreneur
job = student



- People who are in admin job with marital status married and single and most ones who have subscribed for their deposits.
- People who are in blue collar job and are married are the most ones who has not subscribed for their deposits.

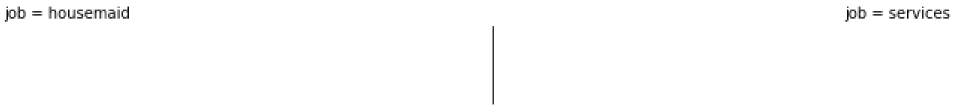
➢ I_loan and Marital

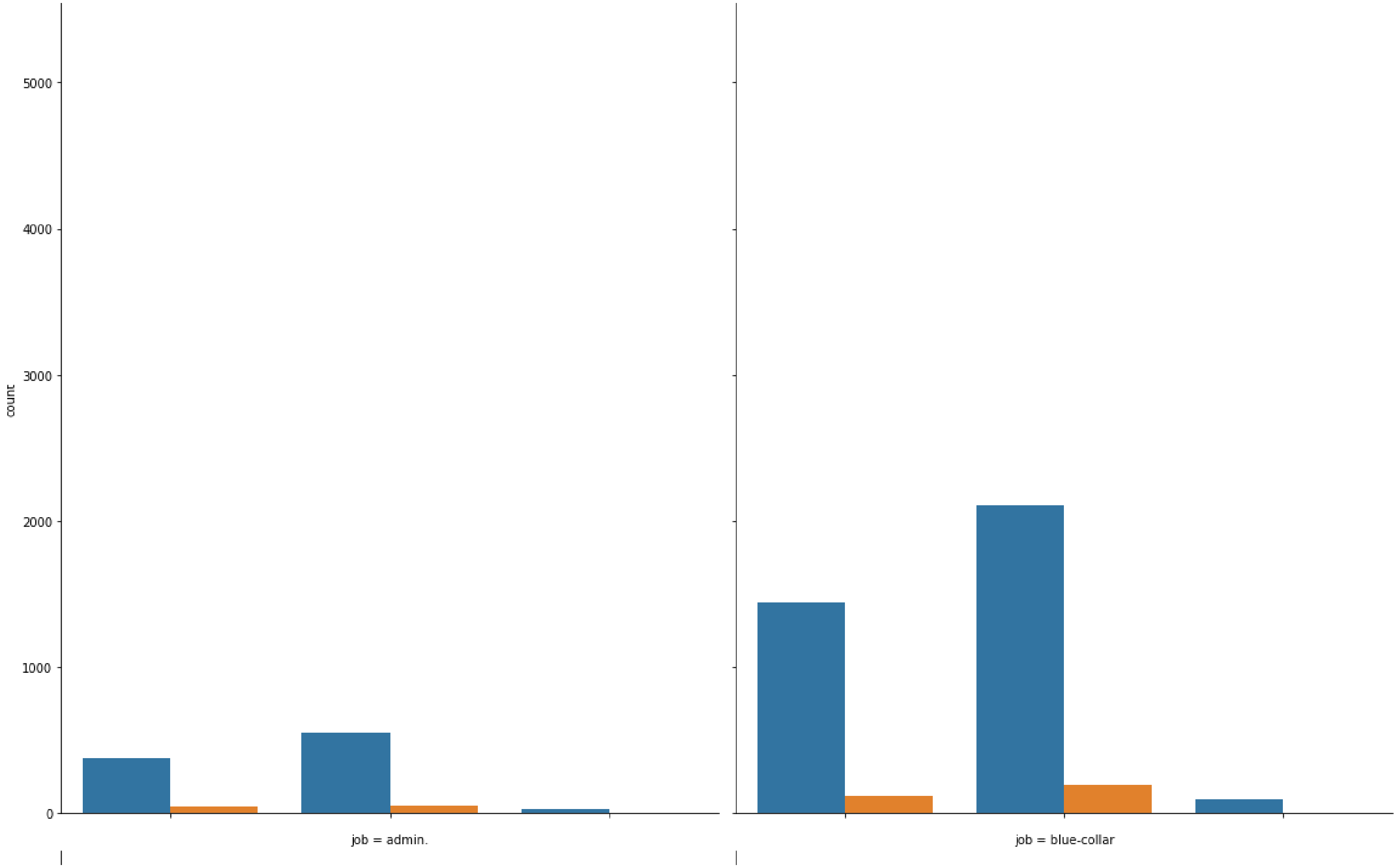


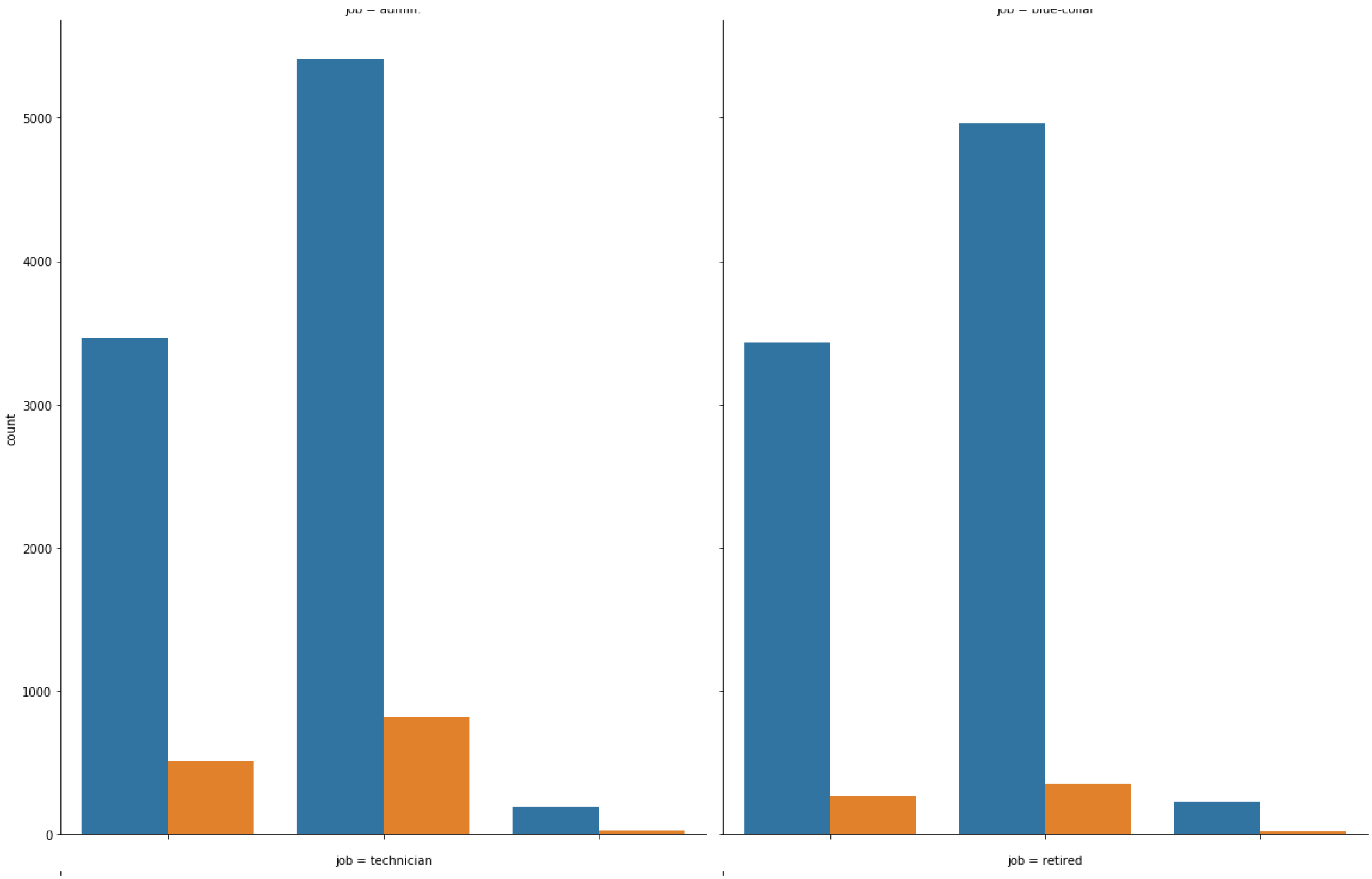


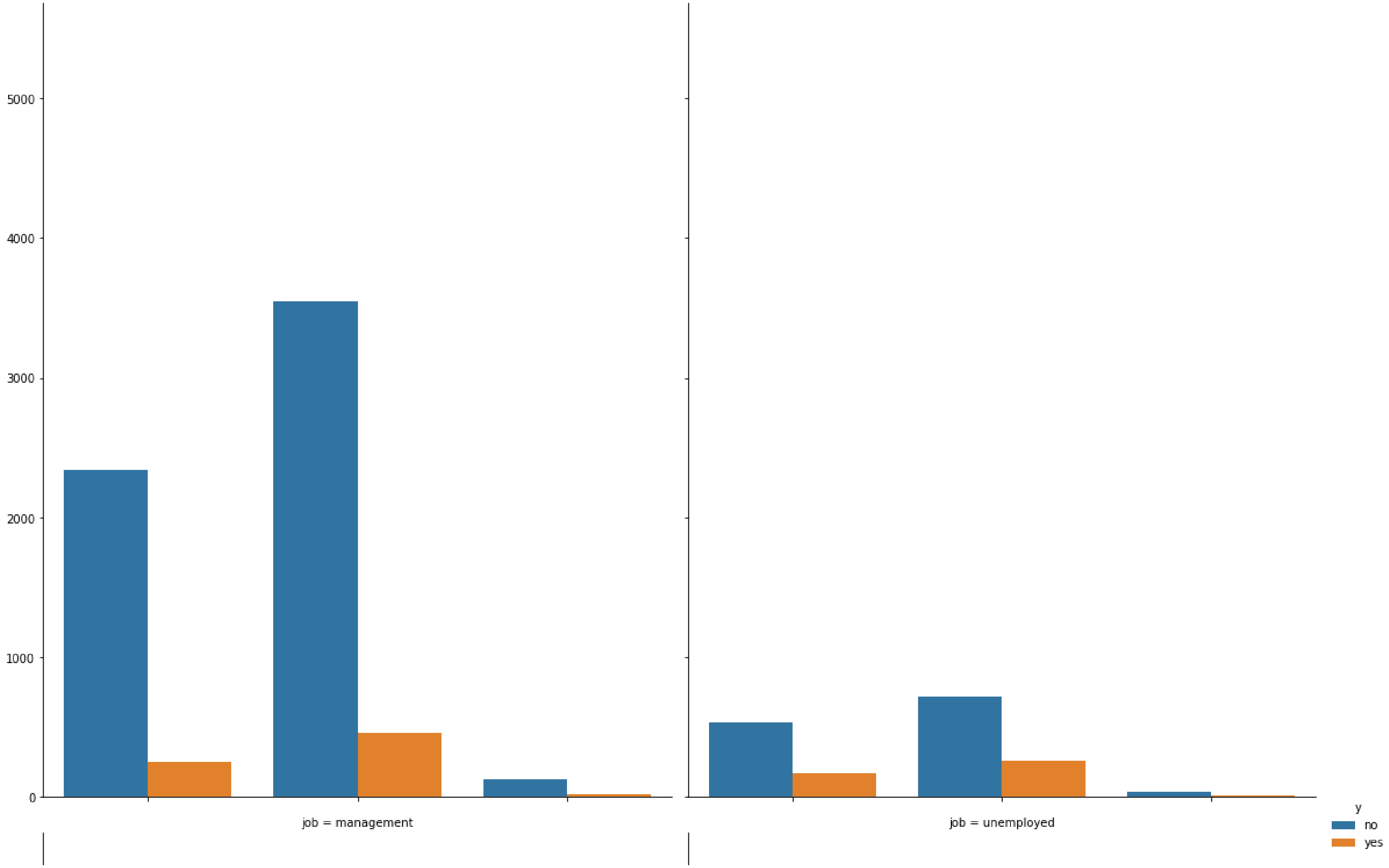
- People who are married and have loans are the most ones who have subscribed for their deposits.
- They are the most ones in any marital status for not subscribing for their deposits.

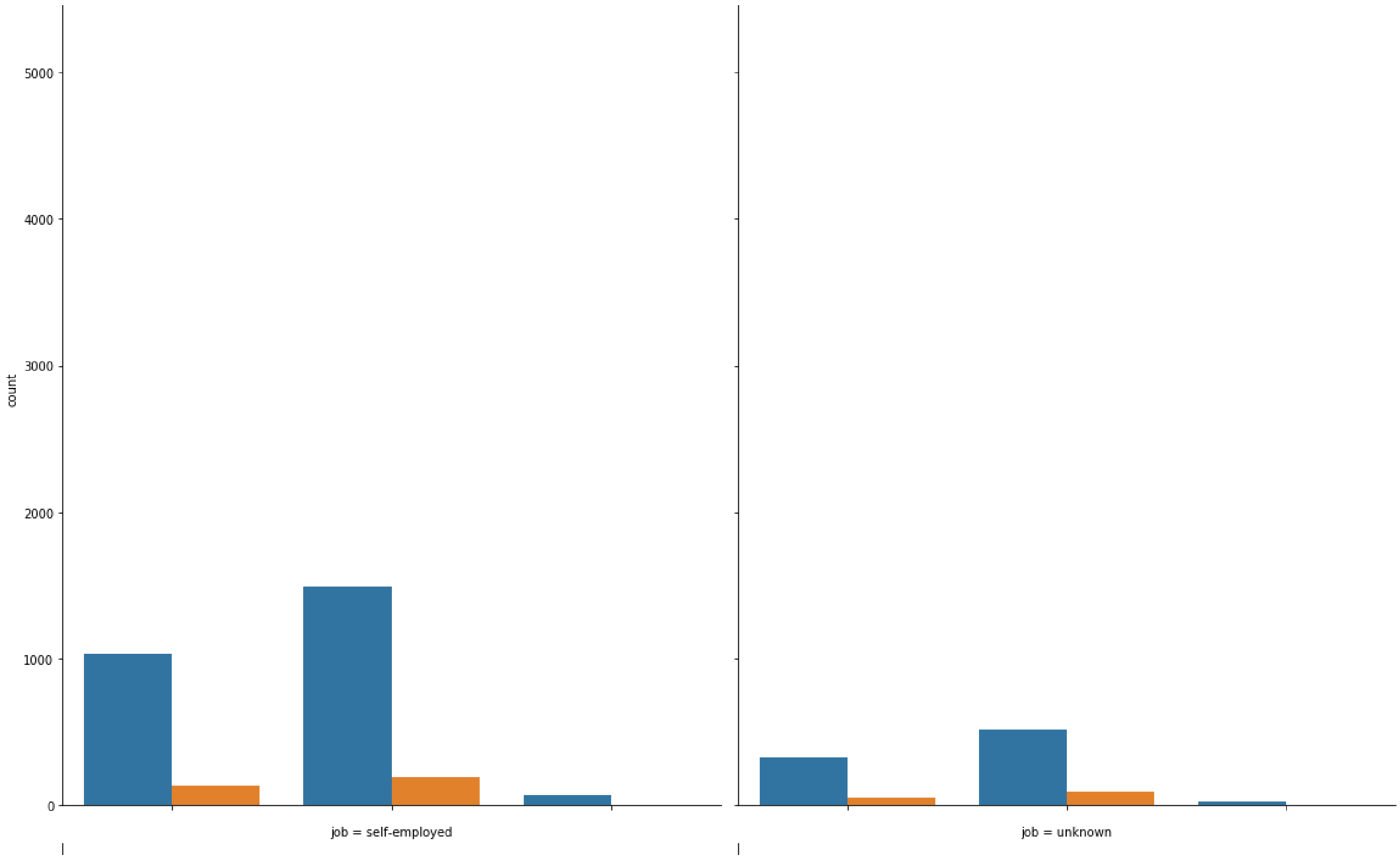
➢ l_loan and Job

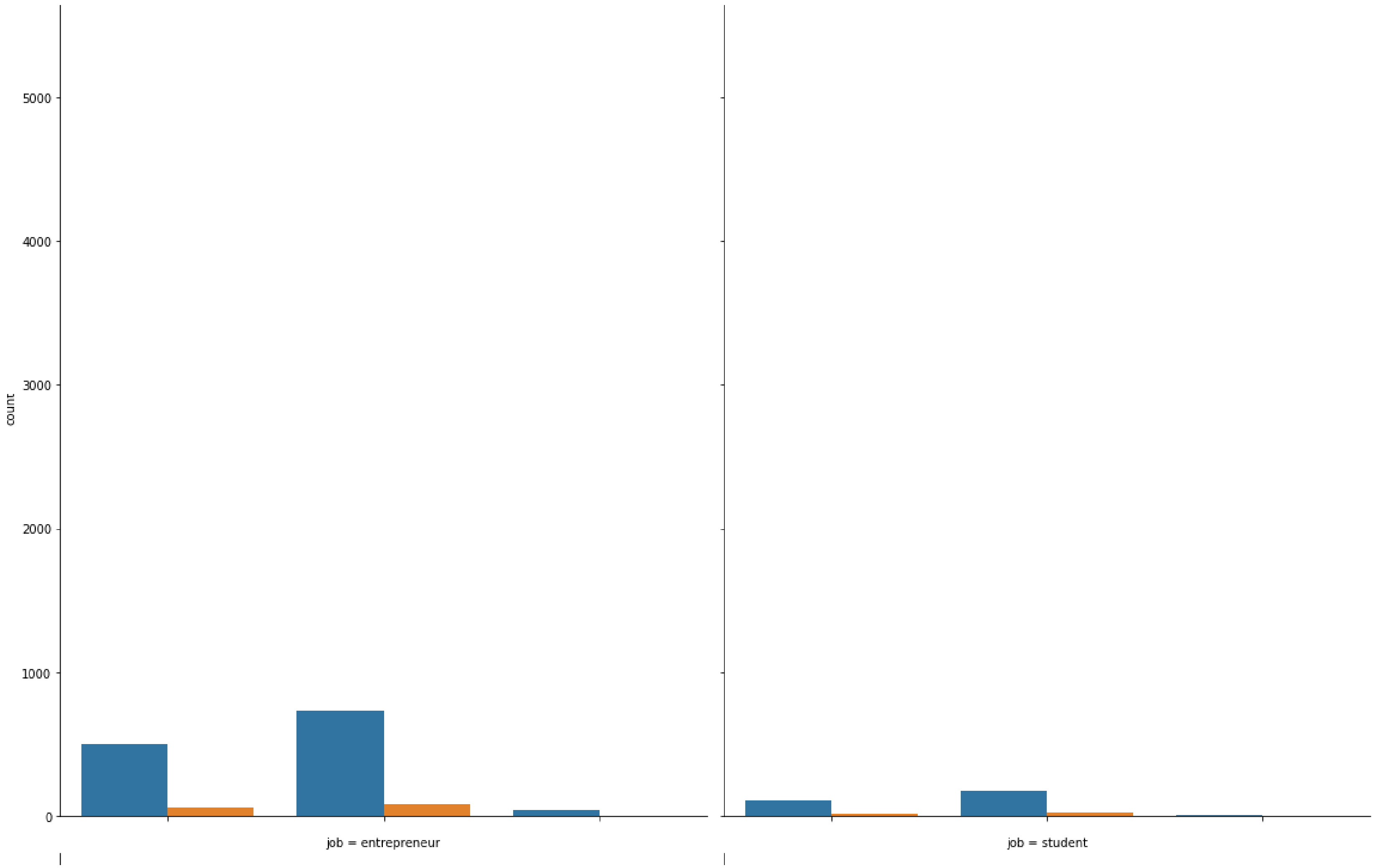


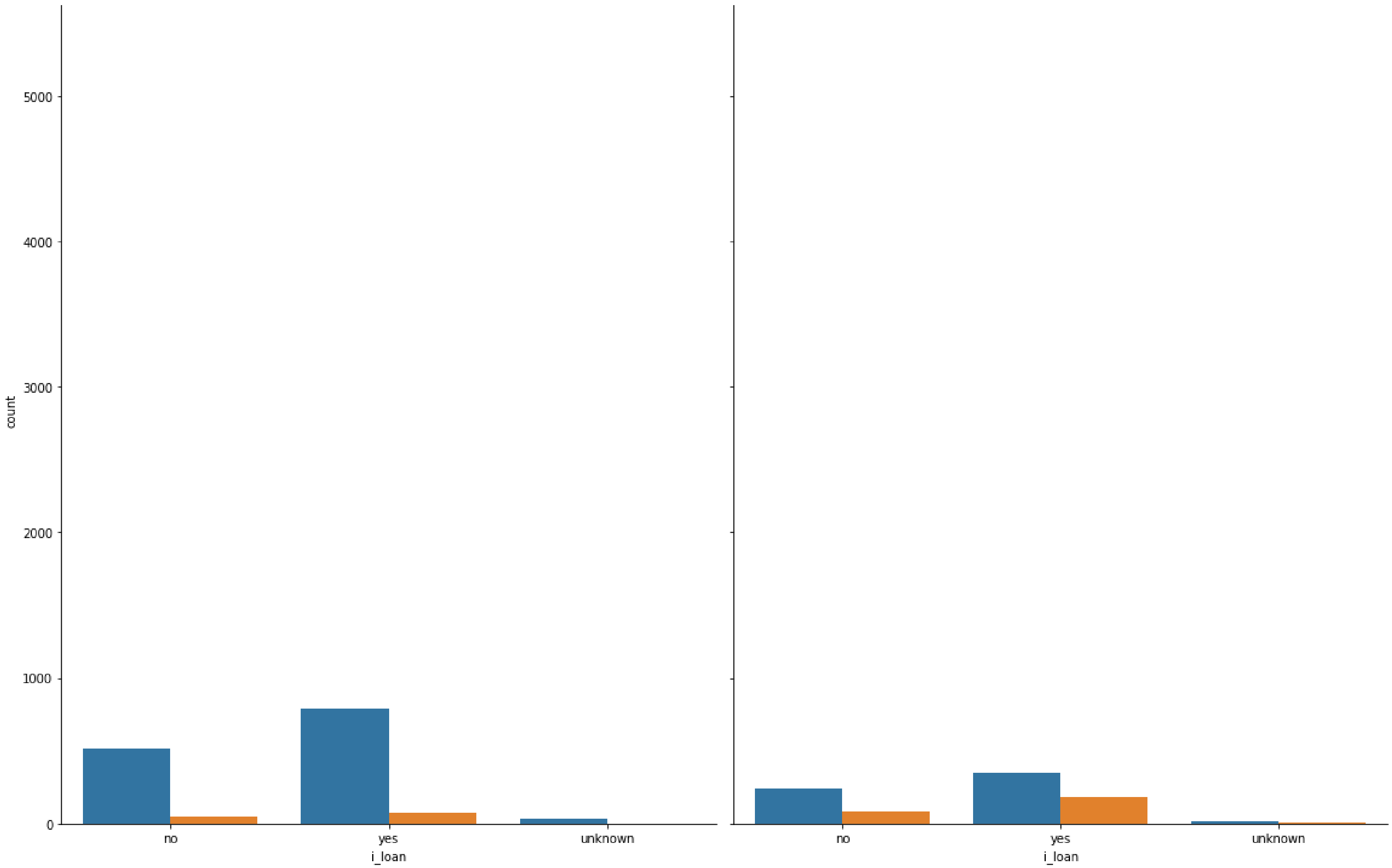

count
0
1000
2000
3000
4000
5000
job = admin.
job = blue-collar


job = admin.
job = blue-collar
5000
4000
3000
2000
1000
0
count
job = technician
job = retired


5000
4000
3000
2000
1000
0
job = management
job = unemployed
y
no
yes


count
5000
4000
3000
2000
1000
0
job = self-employed
job = unknown


count
5000
4000
3000
2000
1000
0
job = entrepreneur
job = student



- People who are in the profession of admin with loans are the ones who have subscribed for deposits,
  followed by the admins who does not have loans subscribed for the deposits.

- People who are technicians and admins and have loan status as yes or no are most ones who have not subscribed for their deposits.

Final Recommendation Summary

After I've conducted some analysis through visualizations using plots, it revealed as follows:

- People who are in admin job has been more contacted for the deposits by the bank.
- People who are married has been contacted more for the deposits by the bank.
- People who have been contacted more on the cellular than the telephone.
- People have been contacted more in the month of May than any other month.
- They have not been contacted in January and February at all.
- People have not been contacted on Saturday and Sunday.
- People with no default status has been contacted more by the bank.
- People who have housing loan has been contacted more by the bank.
- People with no personal loan have been contacted more by the bank.
- People who are in university has been contacted more by the bank.
- Age, Duration, Campaign has outliers and are rightly skewed.
- Pdays have more than 70% of data imputed so it is better either to impute or remove the column.
- Euribor3m with nr.employed and emp.var.rate with nr.employed with the highest correlation

The full source code is done on an ipynb file.