Week 10: Deliverables

Individual Project:
This project is done individually
Group Name:
Name:Nkululeko Freedom Mqadi
Email:mqadinf@gmail.com
Country:South Africa
College/Company:Deviare
Specialization:Data Science

Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Bank Marketing(Campaign)

15 August 2021

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 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).

I'm going to perform some Exploratory Data Analysis (EDA) using the **"bank-additional-full.csv"** dataset.

The main reason for doing some **Exploratory Data Analysis(EDA)**

- is used to ease the burden when doing some machine learning tasks. Machine learning tasks typically start with a comprehensive exploration of the datasets.
- Its helps ML practitioners to gain a deeper understanding of the properties of the data (schema, statistical properties, and so on), the quality of the data (missing values, inconsistent data types, and so on) and the predictive power of the data(for example, the correlation of features with the target).
- we can accomplish the task of performing some descriptive analysis which provides an understanding of the characteristics of the dataset and performing some visualizations which presents data in a pictorial or graphical format.

- The analysis has been divided into the following parts:
- Data Understanding
- Data Exploration
- EDA on categorical variables
- EDA on numerical variables

Data understanding

- The dataset consists of direct marketing campaigns data of a banking institution.
- Picked from UCI Machine Learning Repository which is an amazing source for publicly available datasets.
- There were four variants of the datasets out of which we chose “ bank-additional-full.csv”
- Consists of 41188 data points with 20 independent variables out of which 10 are numeric features and 10 are categorical features. The list of features available to us are given below:

Data set description:

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.

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')

Data Exploration

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
age               41188 non-null int64
job               41188 non-null object
marital           41188 non-null object
education         41188 non-null object
default           41188 non-null object
housing           41188 non-null object
loan              41188 non-null object
contact           41188 non-null object
month             41188 non-null object
day_of_week       41188 non-null object
duration          41188 non-null int64
campaign          41188 non-null int64
pdays             41188 non-null int64
previous          41188 non-null int64
poutcome          41188 non-null object
emp.var.rate      41188 non-null float64
cons.price.idx    41188 non-null float64
cons.conf.idx     41188 non-null float64
euribor3m         41188 non-null float64
nr.employed       41188 non-null float64
y                 41188 non-null object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```

- Above we can see that there are 10 numerical columns having 5 columns of integer datatype and 5 columns on float datatype and 11 are categorical datatype including target variable which is named as y in the dataset.

Checking for null/missing values

```
age               0
job               0
marital           0
education         0
default           0
housing           0
loan              0
contact           0
month             0
day_of_week       0
duration          0
campaign          0
pdays             0
previous          0
poutcome          0
emp.var.rate      0
cons.price.idx    0
cons.conf.idx     0
euribor3m         0
nr.employed       0
y                 0
dtype: int64
```

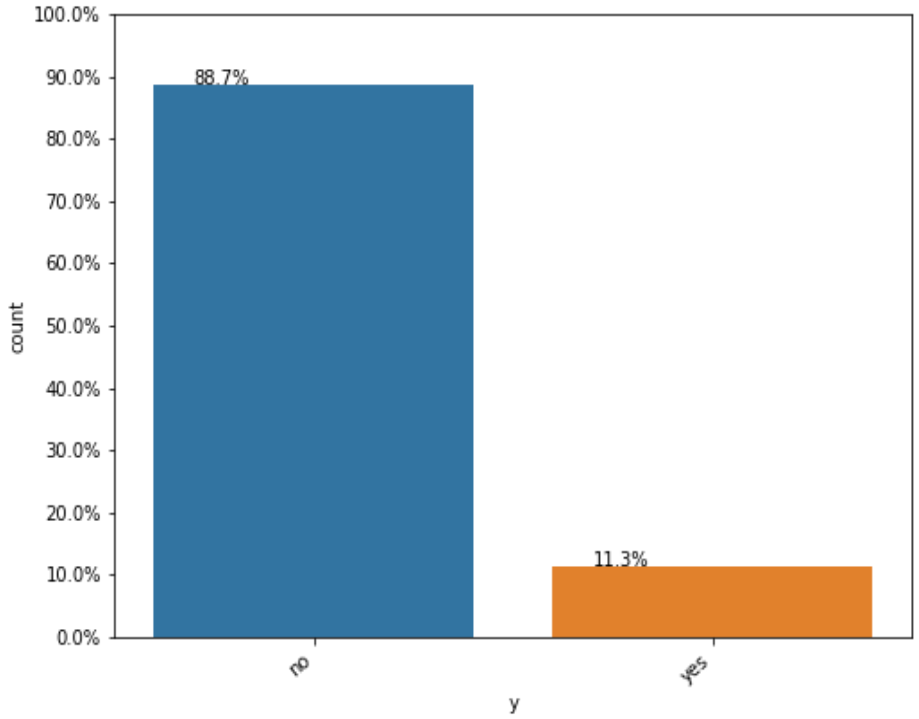
- There are no null/missing values in a dataset.

- # EDA
  ## Categorical Variables:

I will first perform some exploratory data analysis on the categorical variables.

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

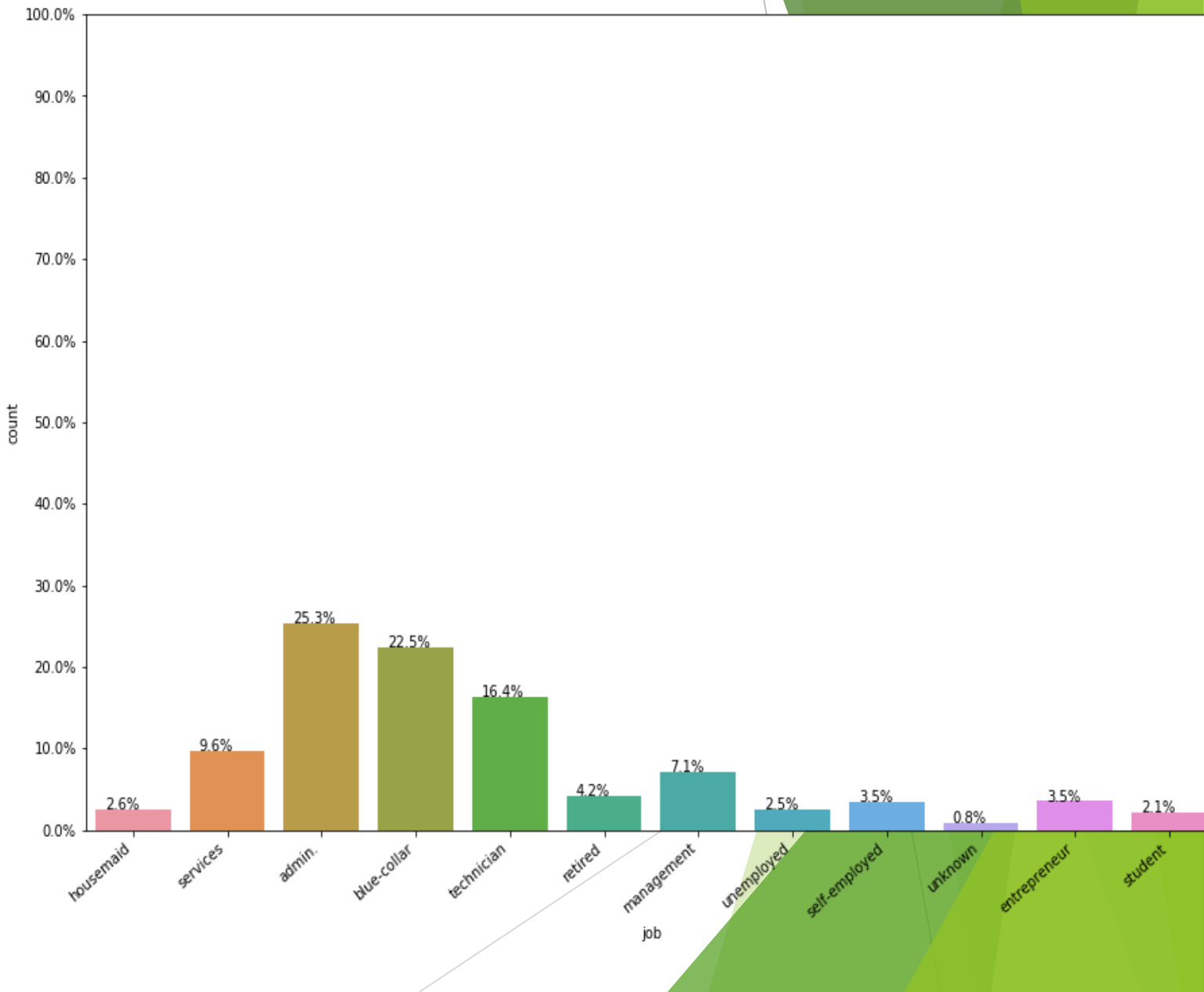
As shown from the plot, we can see that majority of datapoints belong to No class labels with 88.7% and minority of class belongs to 11.3% so the ratio of No:Yes is 8:1.



Univariate Analysis EDA

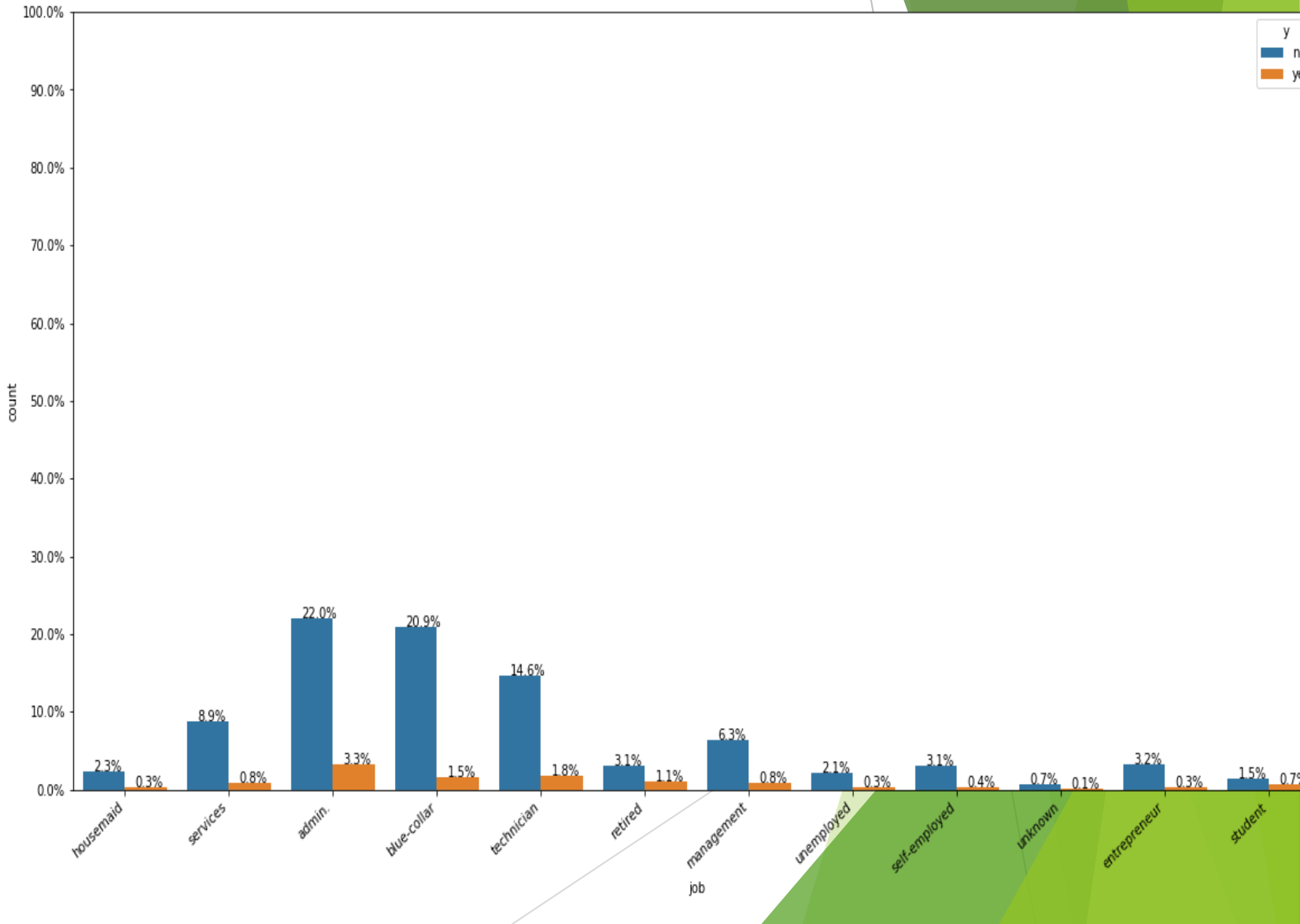
Let’s start doing EDA on rest of the columns of the datapoints.

- **Feature: Job (Categorical variable)**
- From the graph 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.



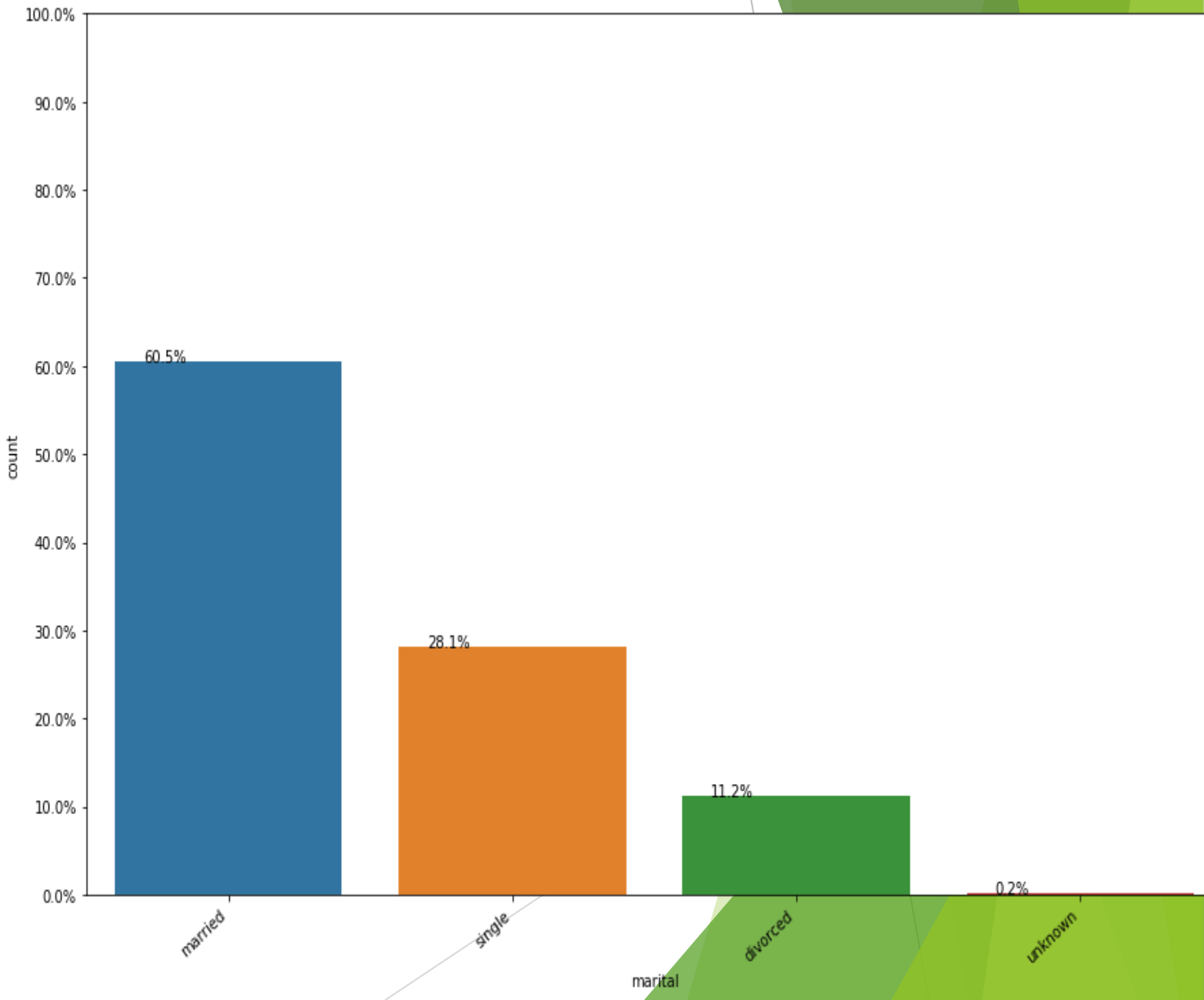
➢ Which job profession has the highest rate for subscribing a term deposit and which has the highest rate of not subscribing?

- From the plot, we can see that the customers who have jobs of admin have the highest rate of subscribing a term deposit
- 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.



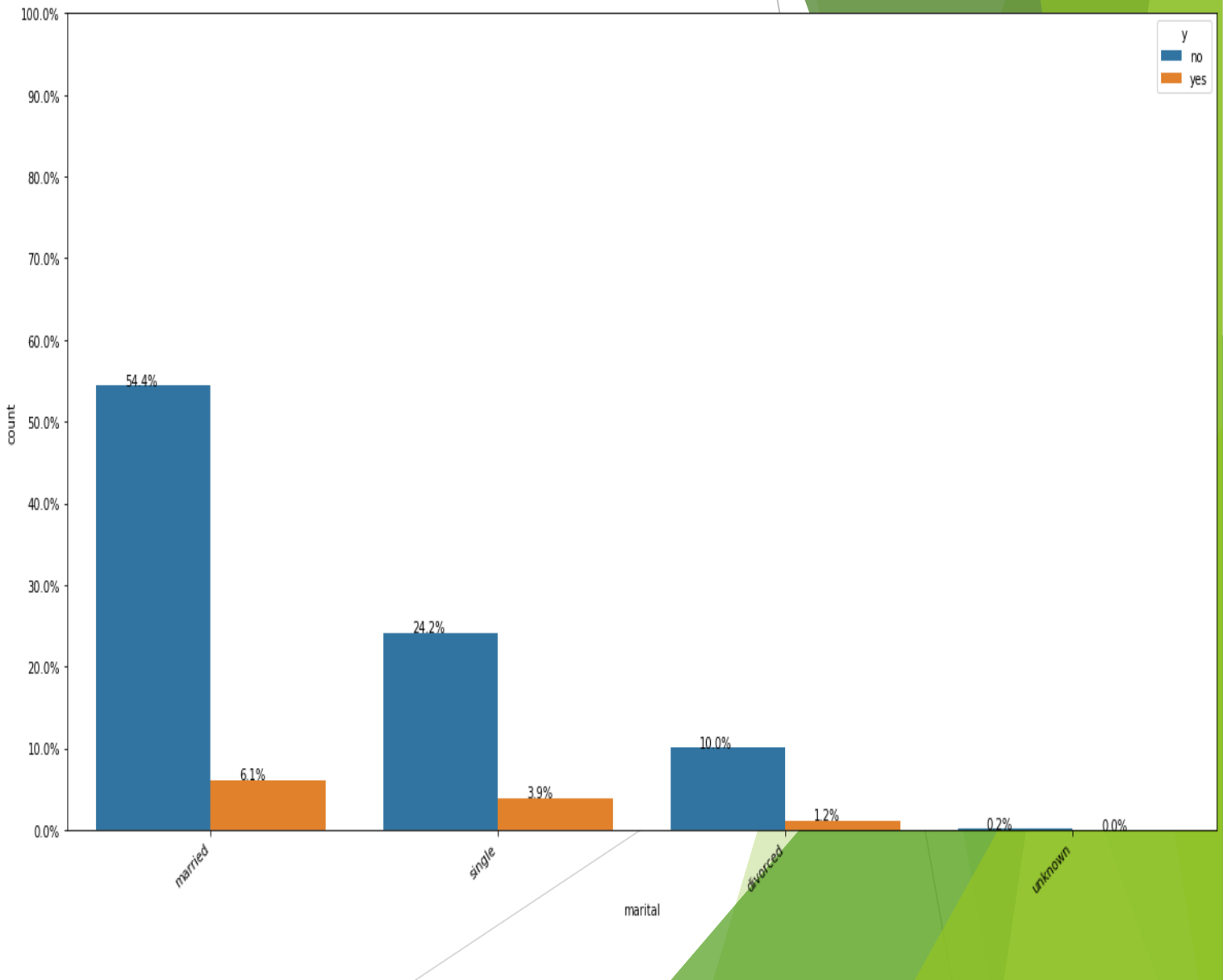
Feature: Marital (Categorical feature)

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



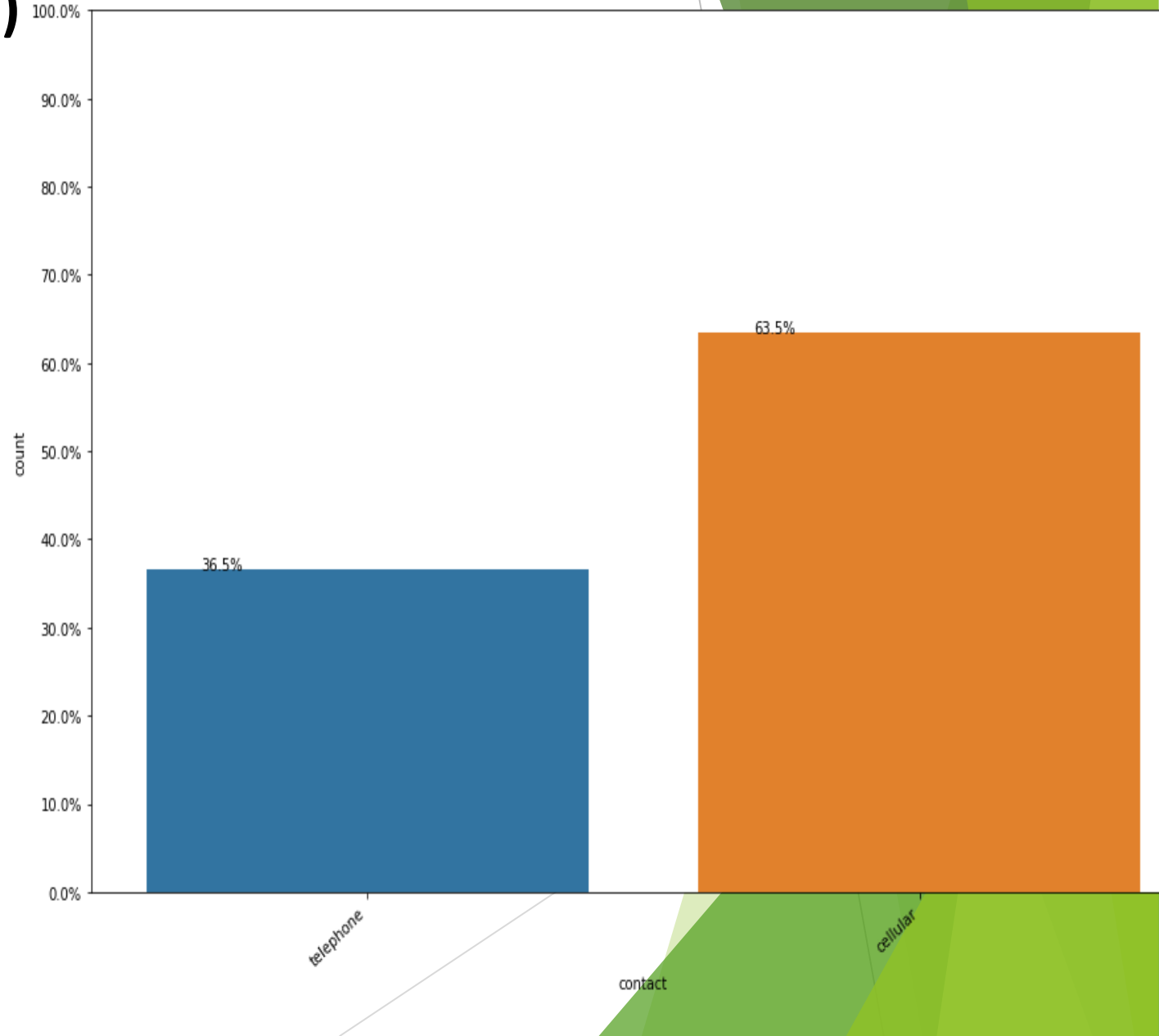
➢ Which marital status subscribes the most for long term deposits?

- From the plot 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 has not subscribed to the long-term deposits at all.



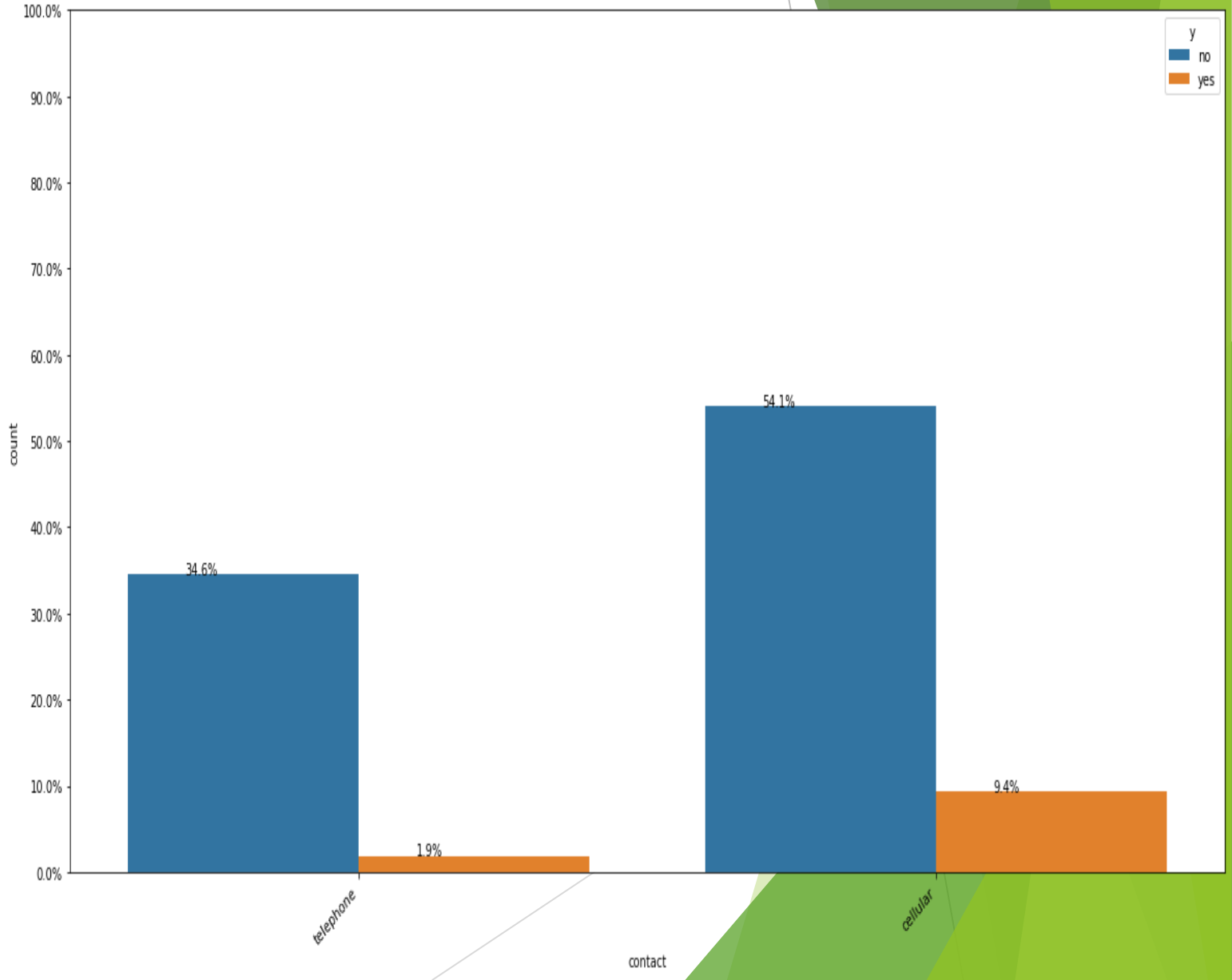
➢ Feature: contact (Categorical)

- As shown from the plot, most people are contacted more in cellular than telephone.



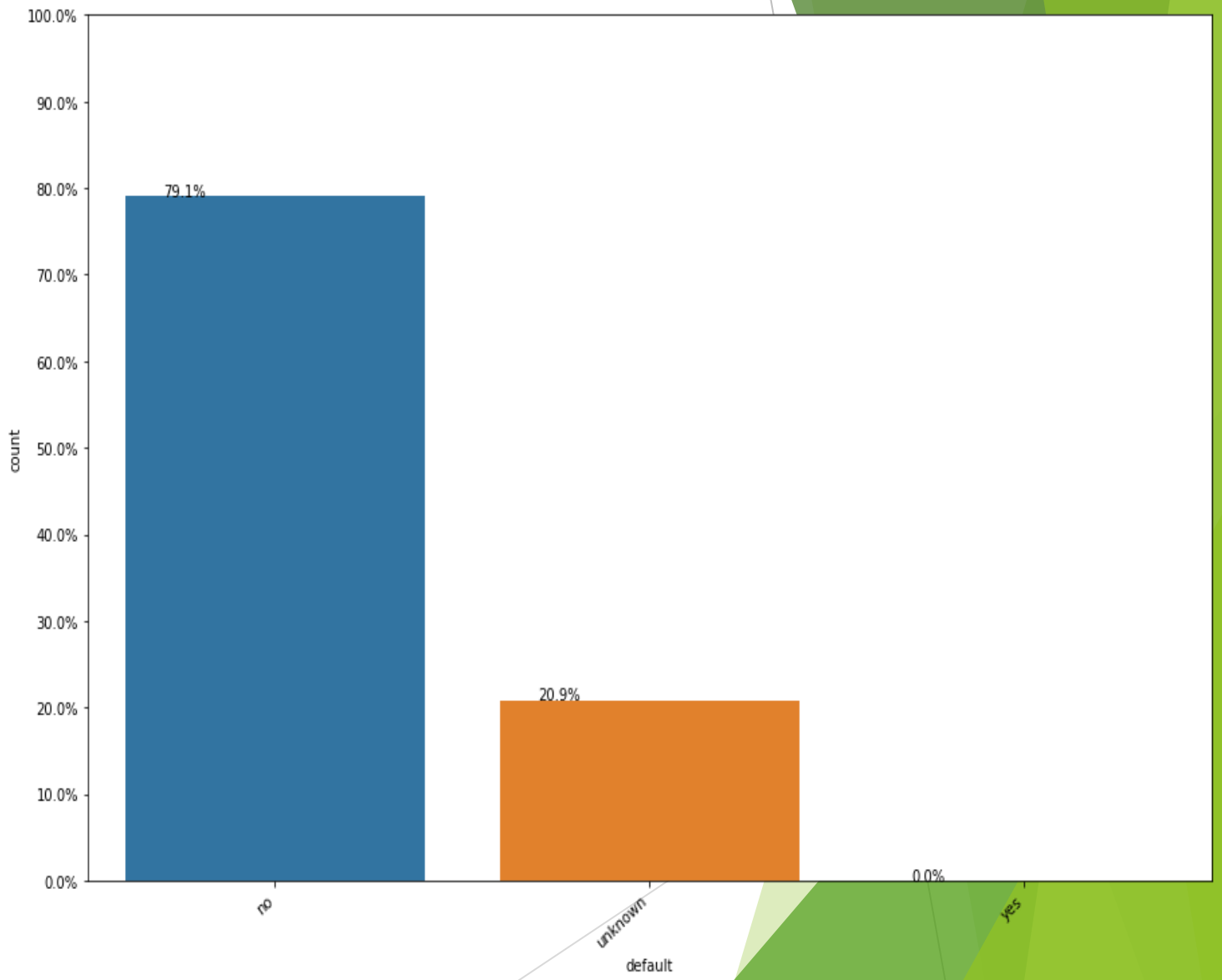
➢ Which contact type has subscribed the most for long term deposits?

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



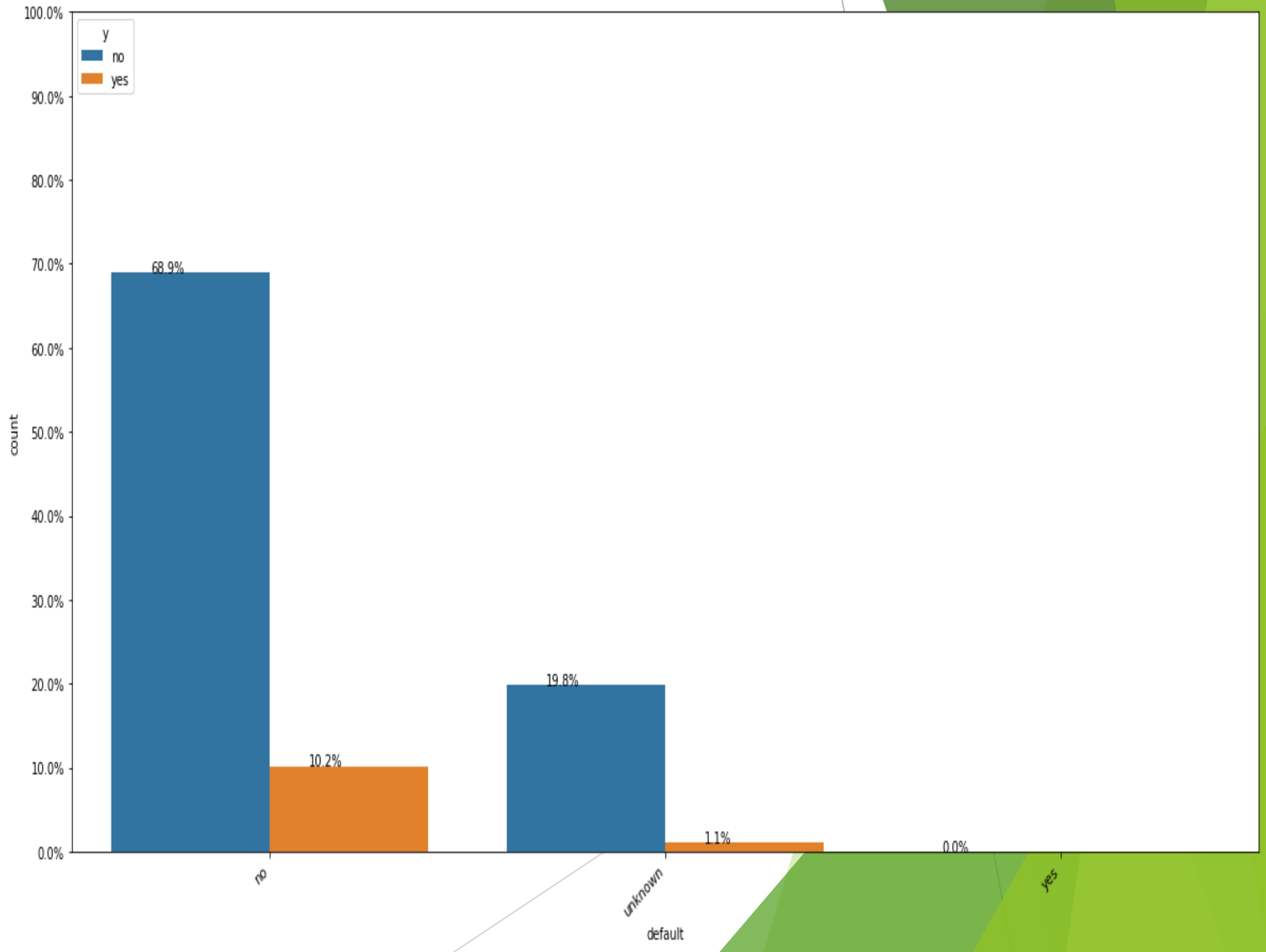
➢ Feature: default (categorical)

- As shown from the plot, 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.



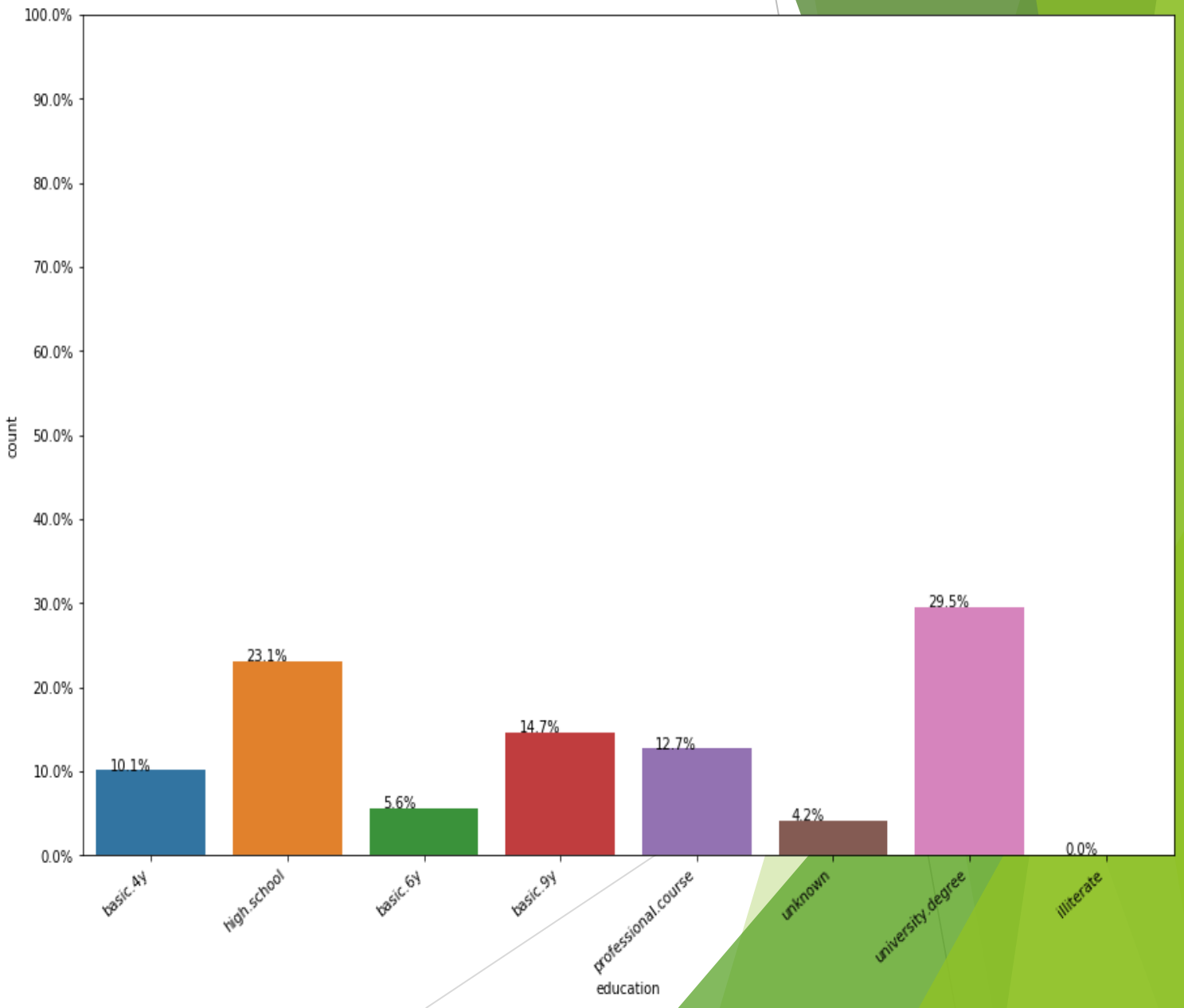
➢ Do people who have or who don't have credit in default subscribe the most for long term deposits?

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



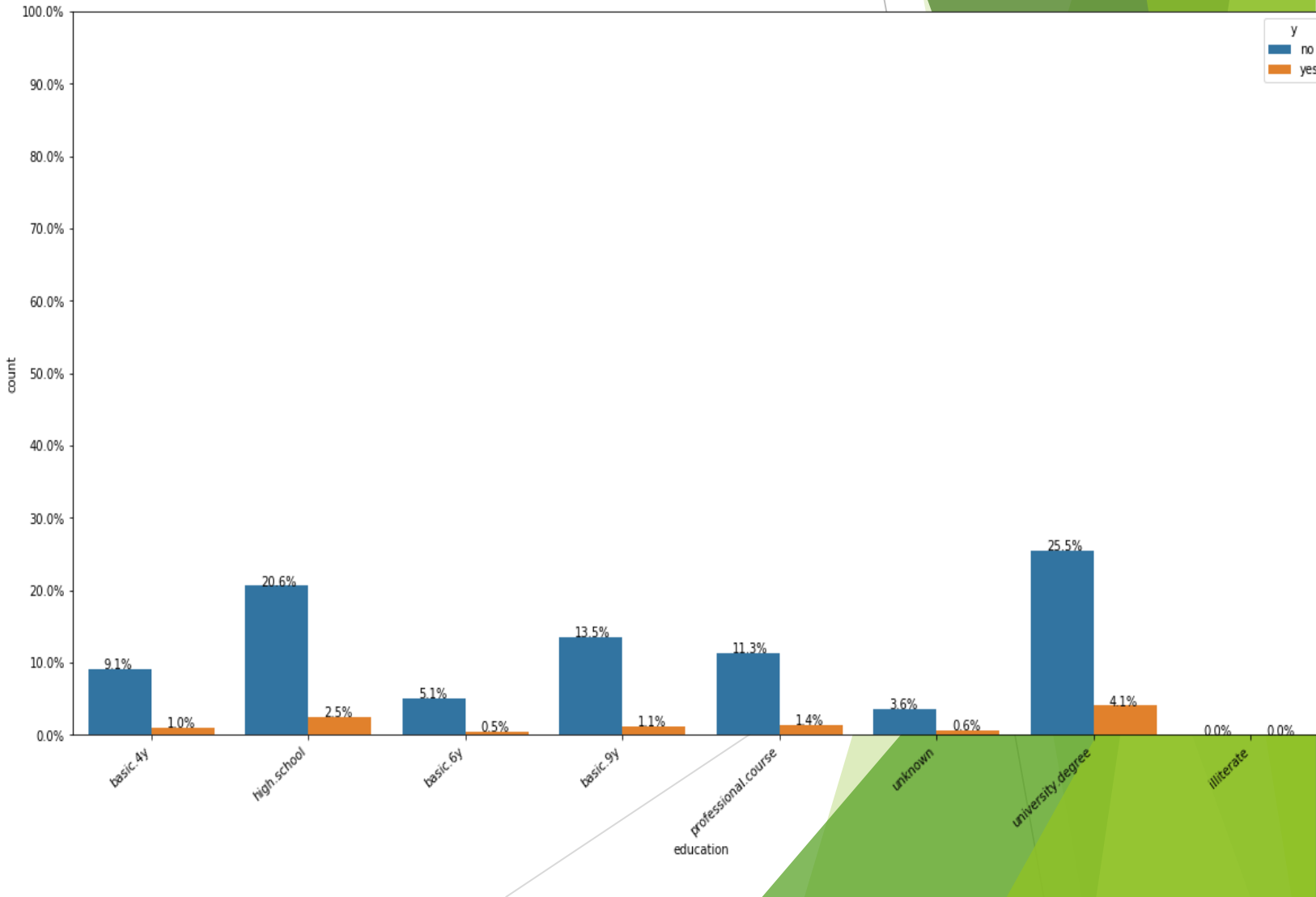
Feature: Education

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



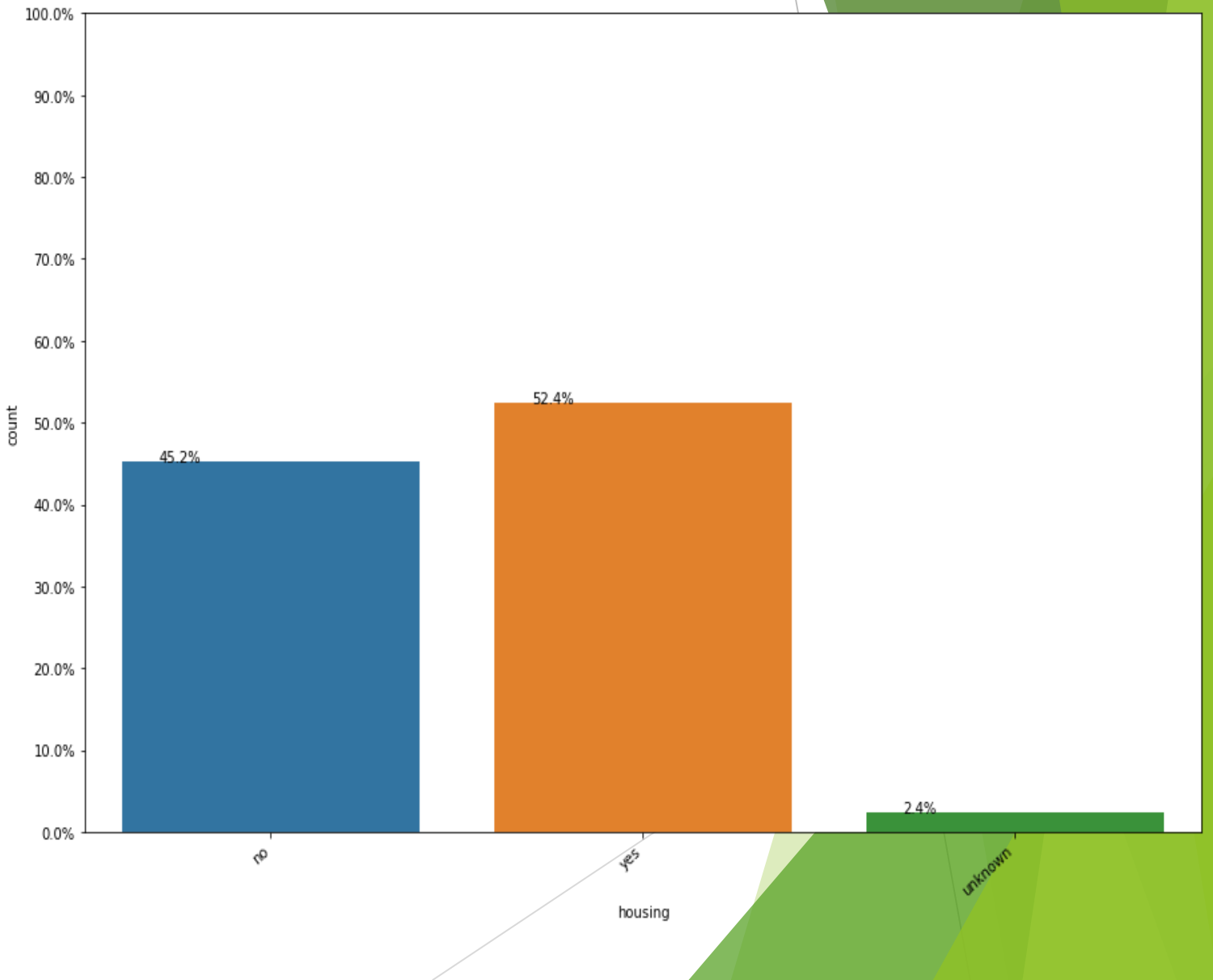
➢ Do people who have completed their university degrees tend to subscribe the most for long term deposits?

- As shown from the plot, 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.



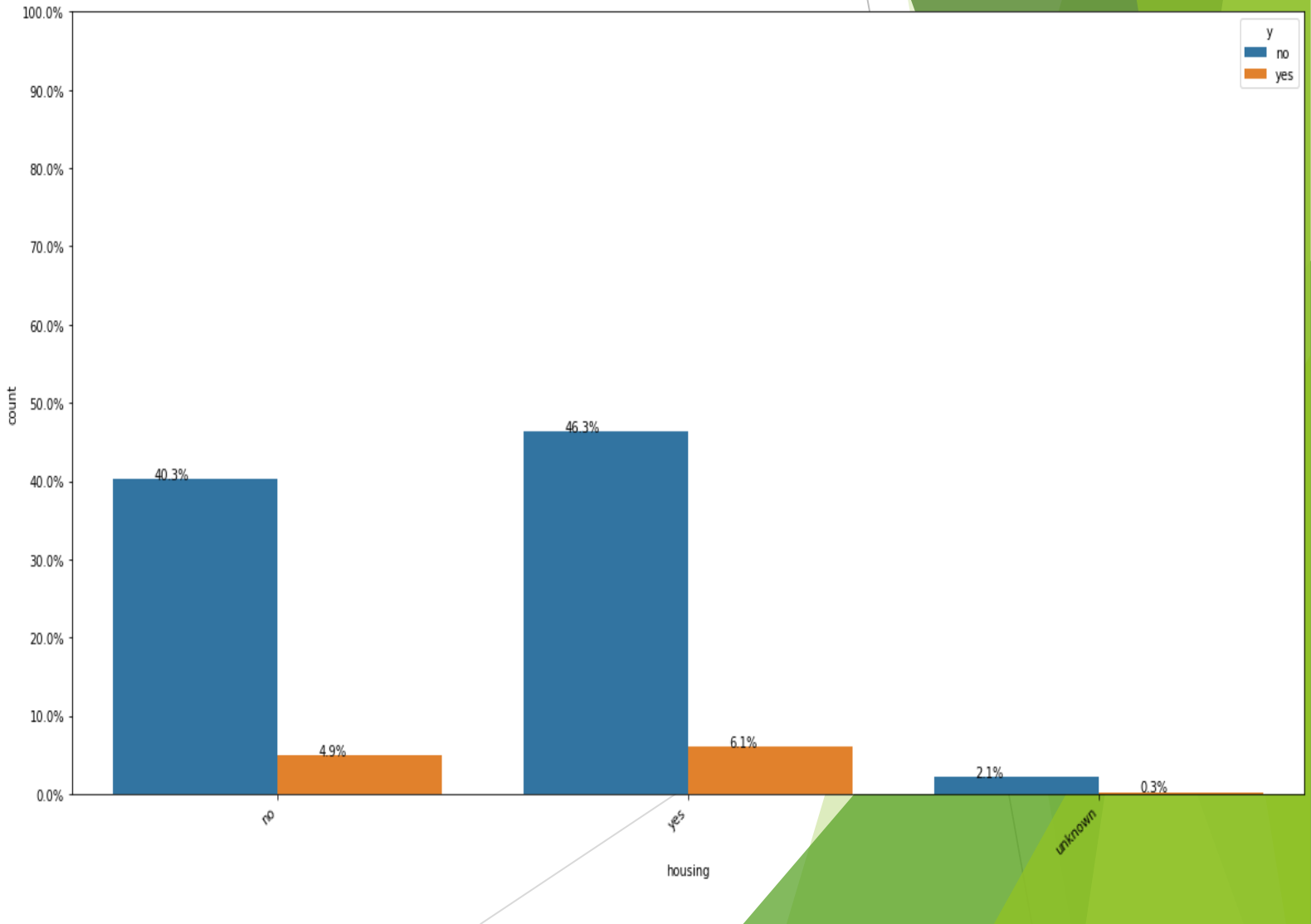
➢ Feature: housing (Categorical)

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



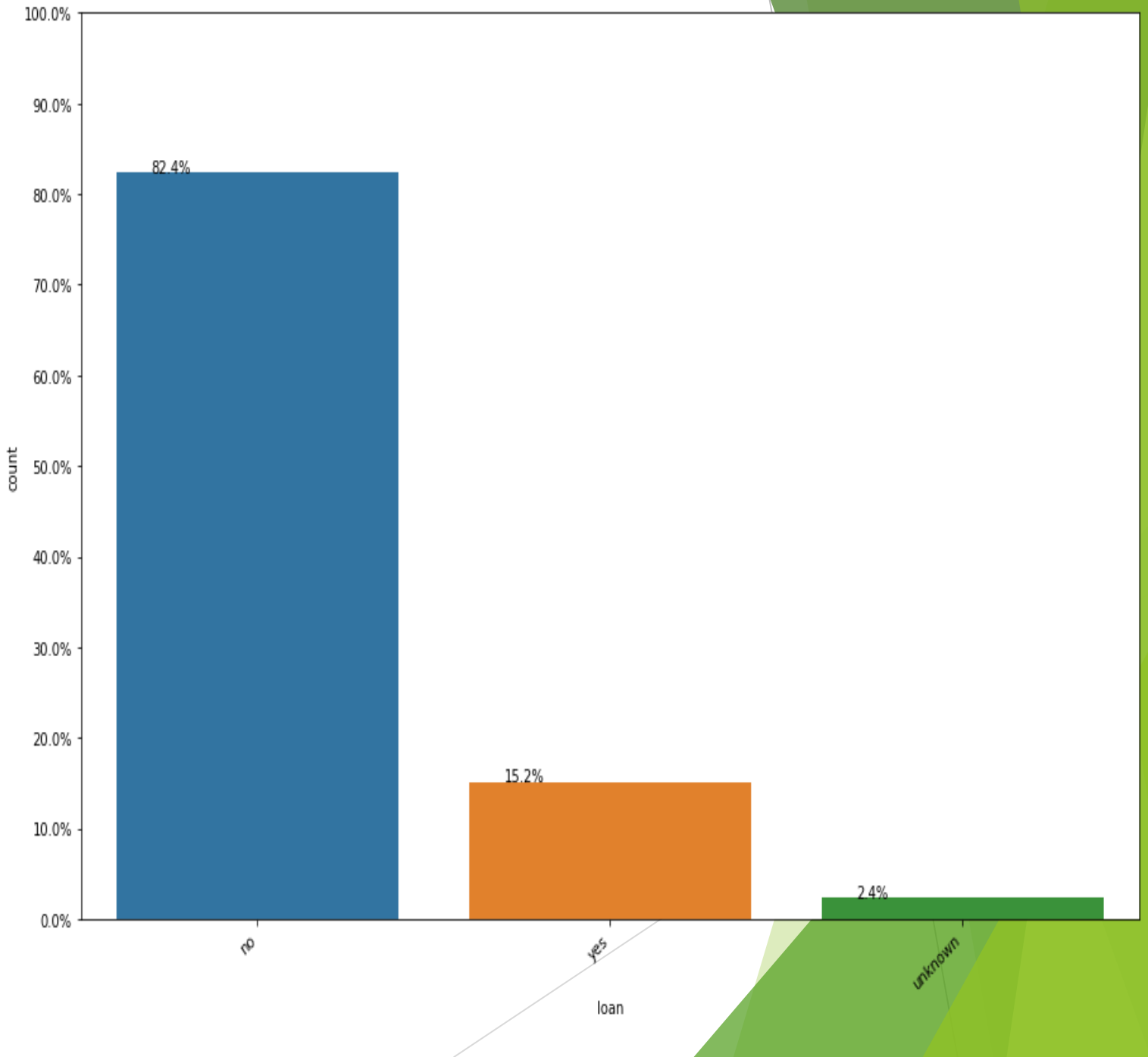
➢ Do people with or without a housing loan subscribe the most for long term deposits?

- 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 has also the most ones who have not subscribed for the deposits.



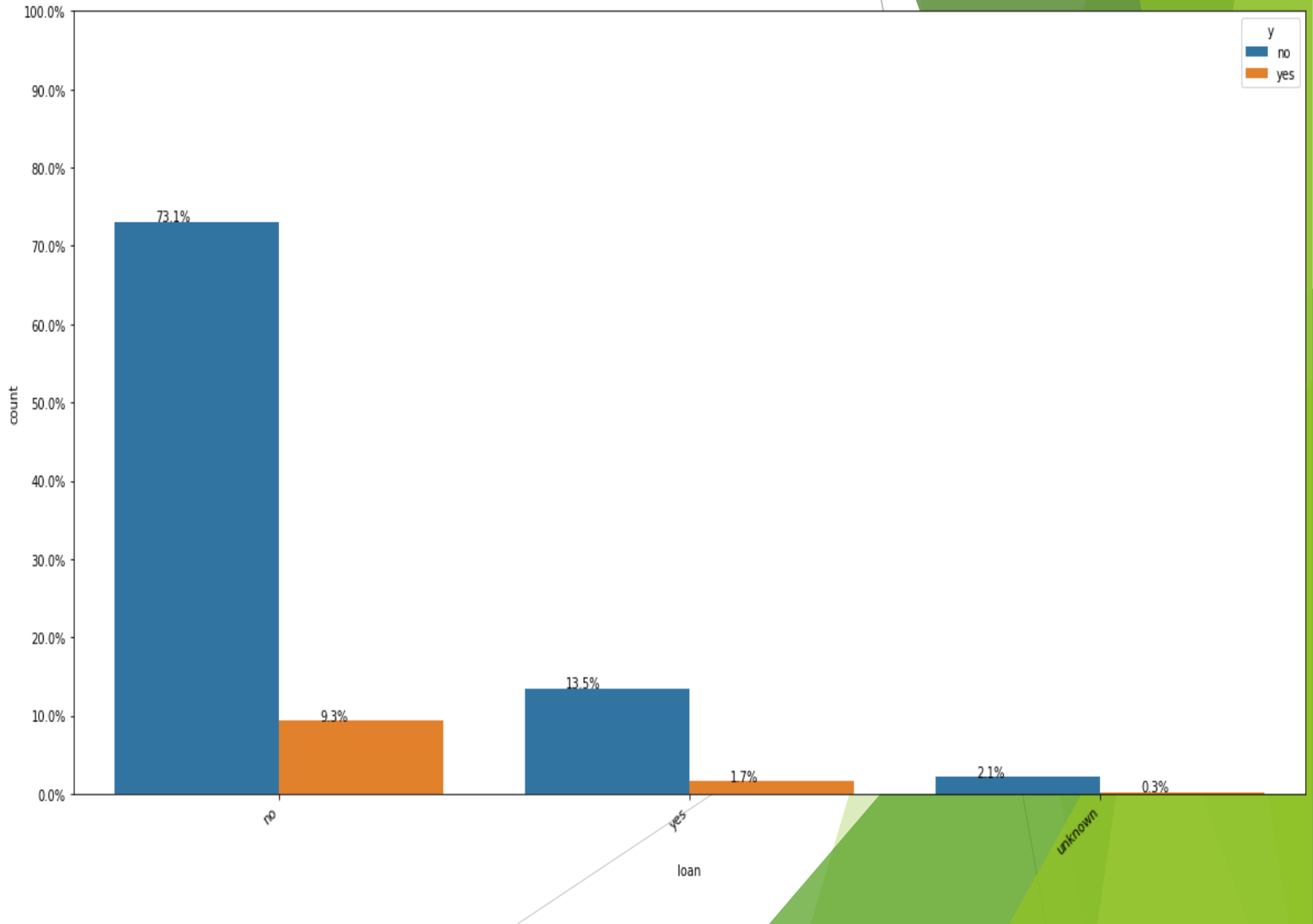
➢ Feature: loan (Categorical)

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



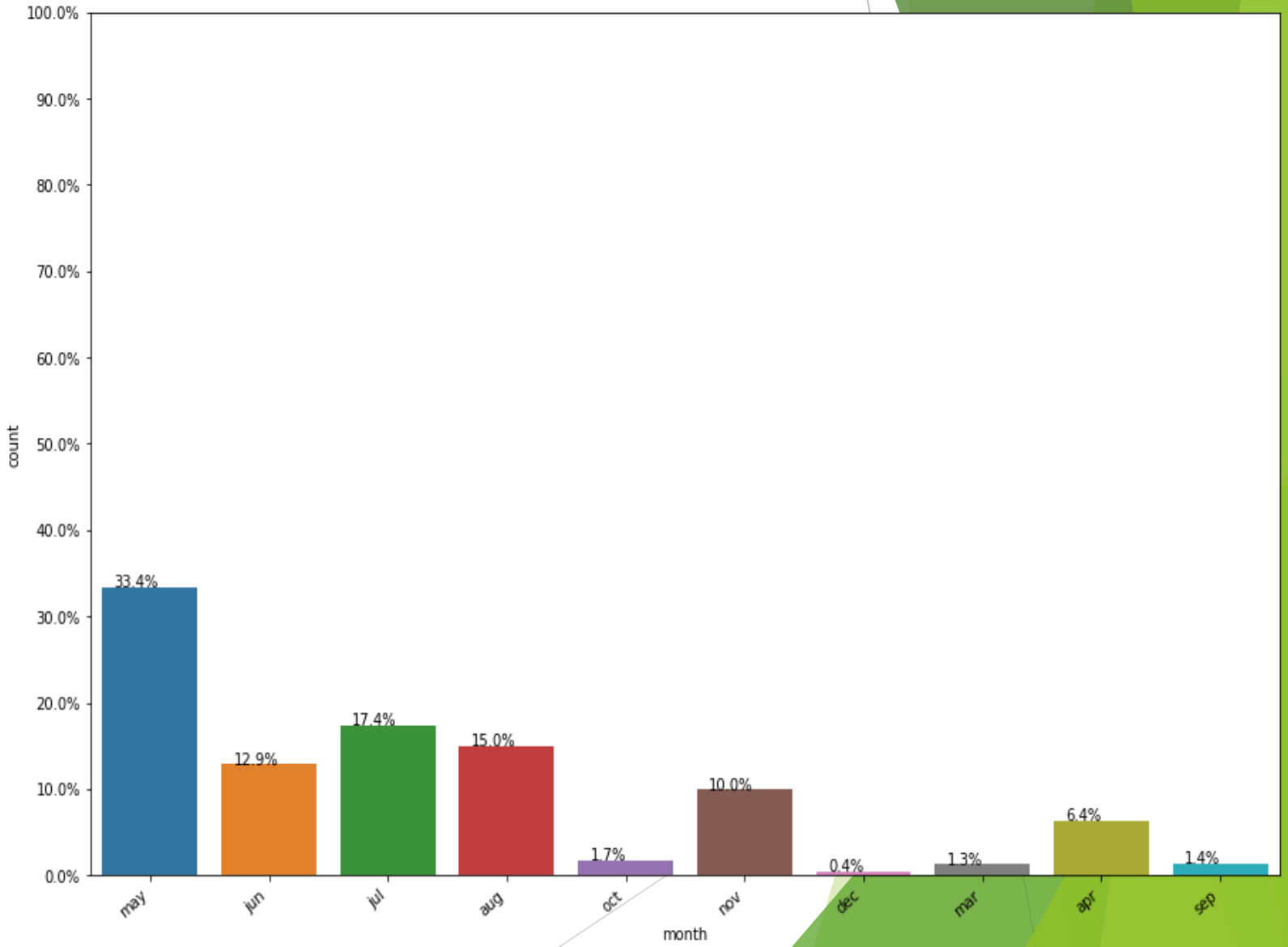
➢ Do people with or without loans subscribed the most for long term deposits?

- 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.



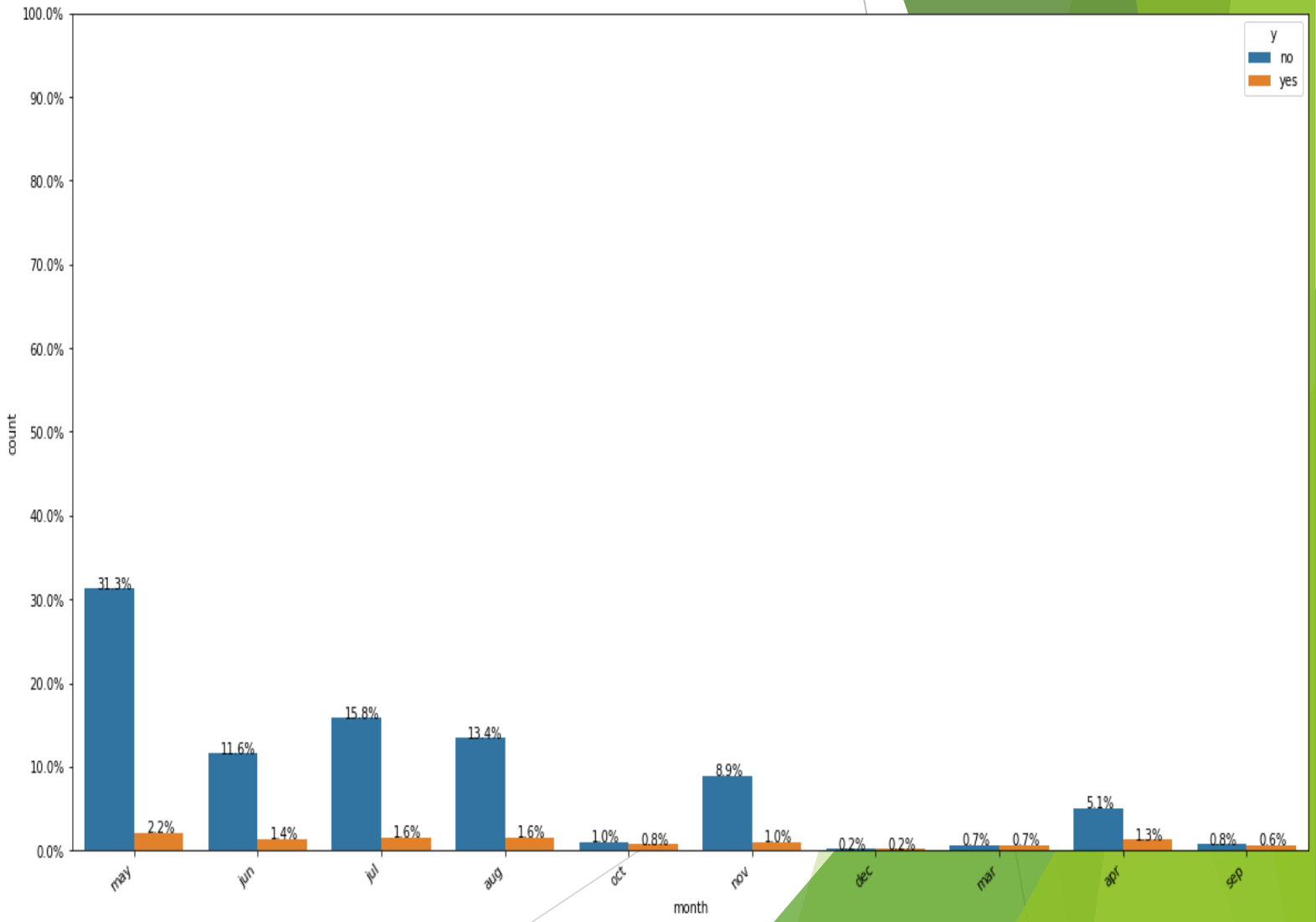
➢ Feature: month (Categorical)

- 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.



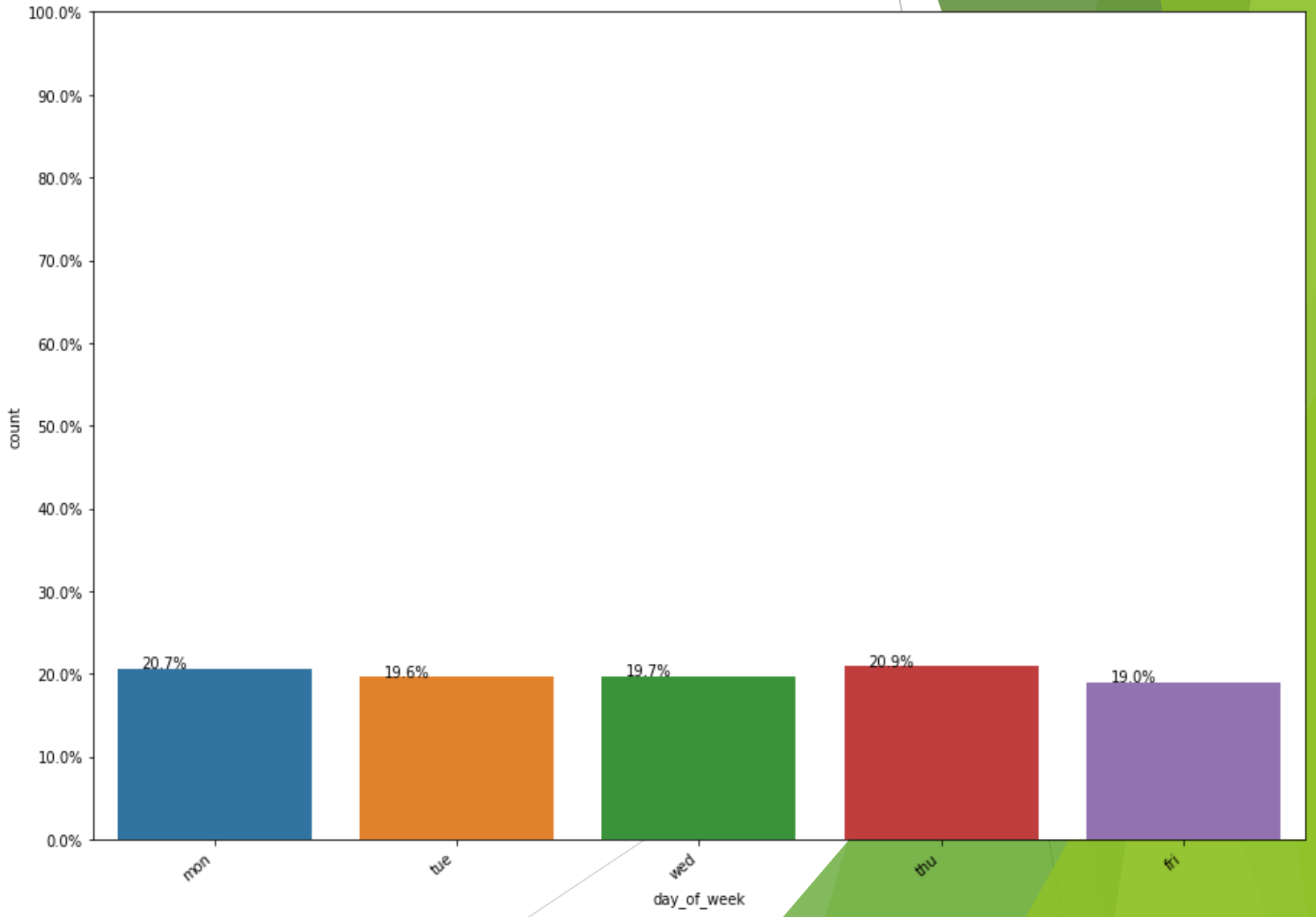
➢ In which month do people have higher chances to subscribe for longer term deposits?

- 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.



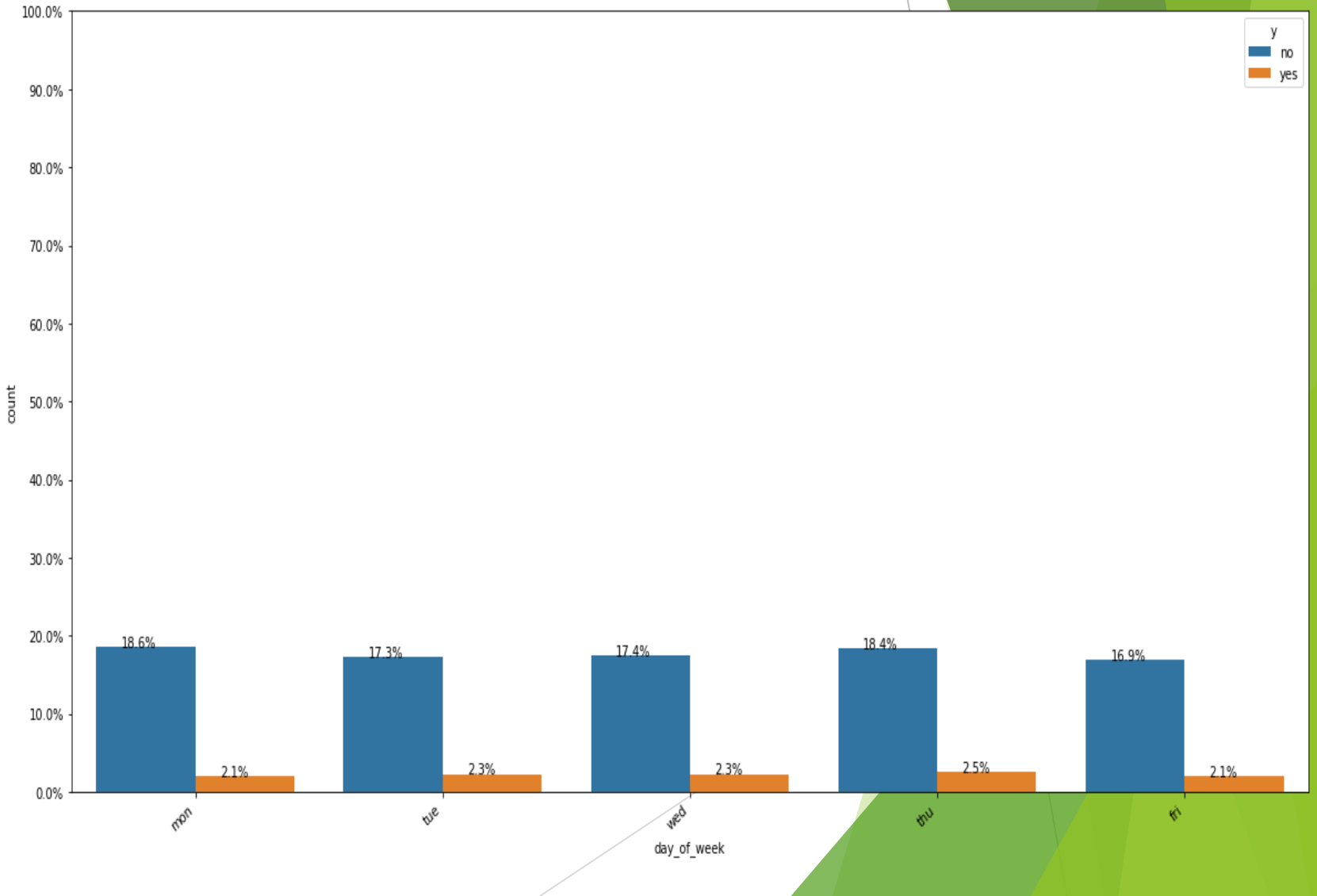
➢ Feature: day_of_week (Categorical)

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



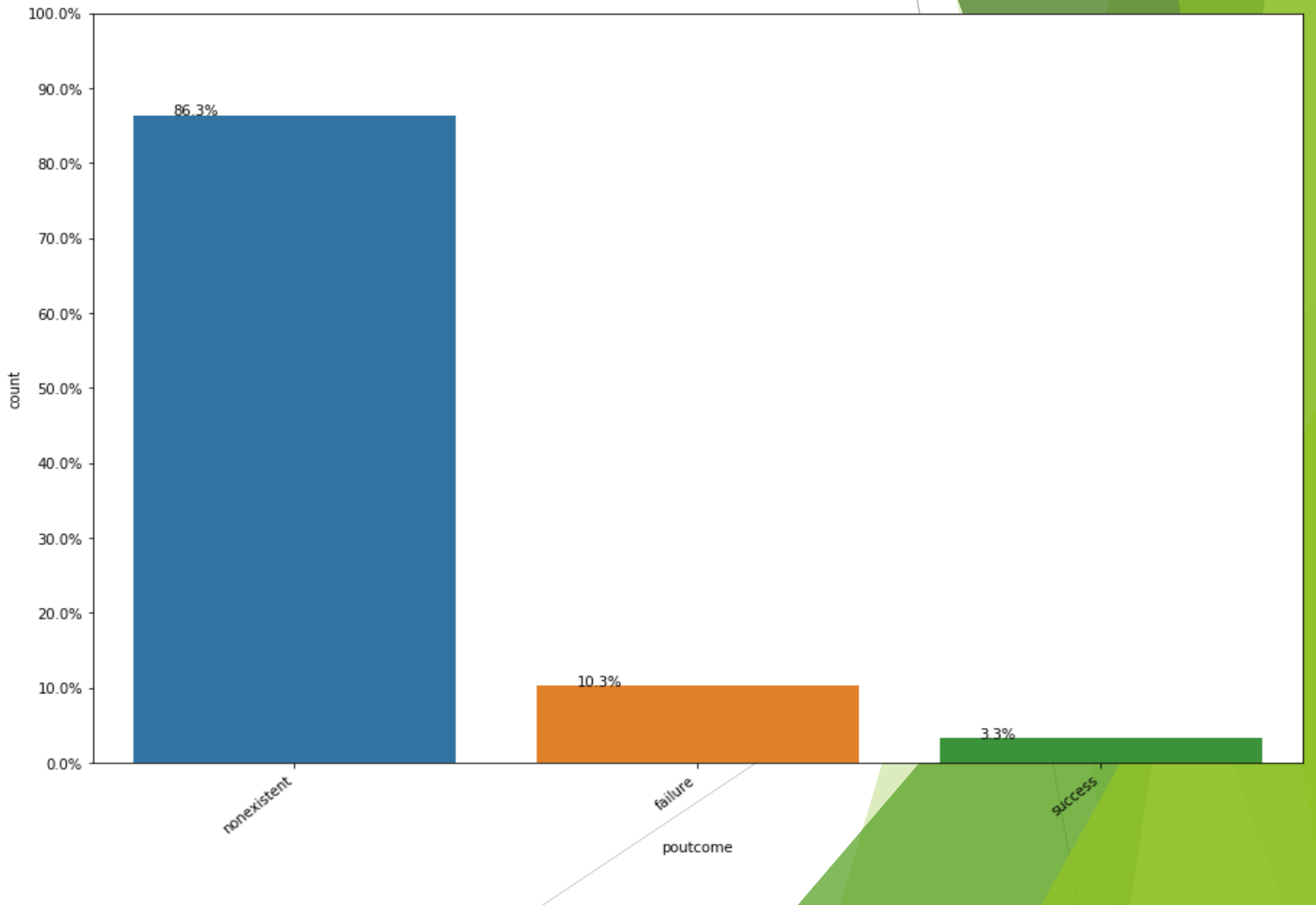
➢ Does in all days people have equal chances for subscribing or not subscribing for the long-term deposits?

- As shown from the plot, 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.



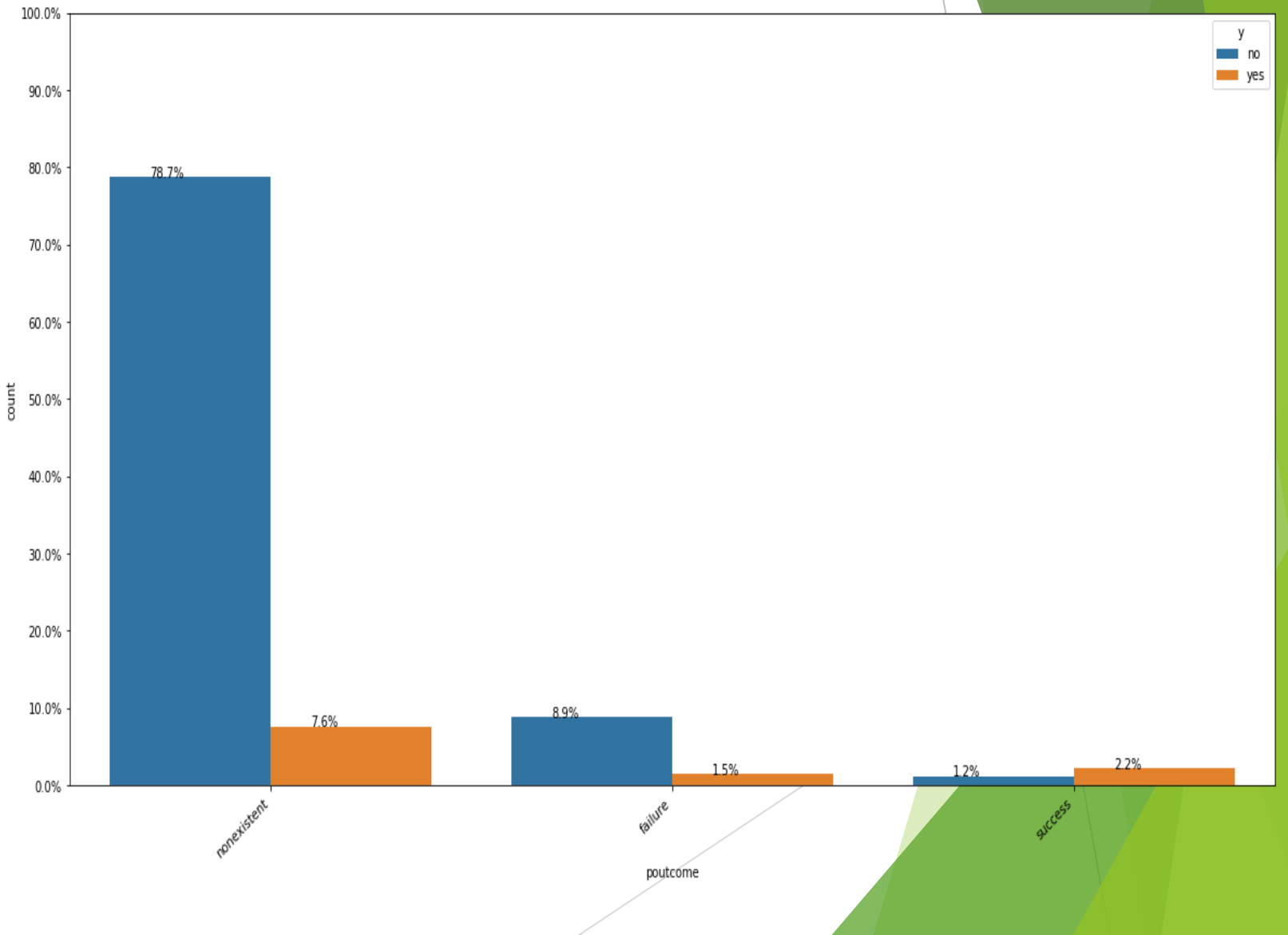
➢ Feature: poutcome (Categorical)

- As shown from the 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?

- As we can see, 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.

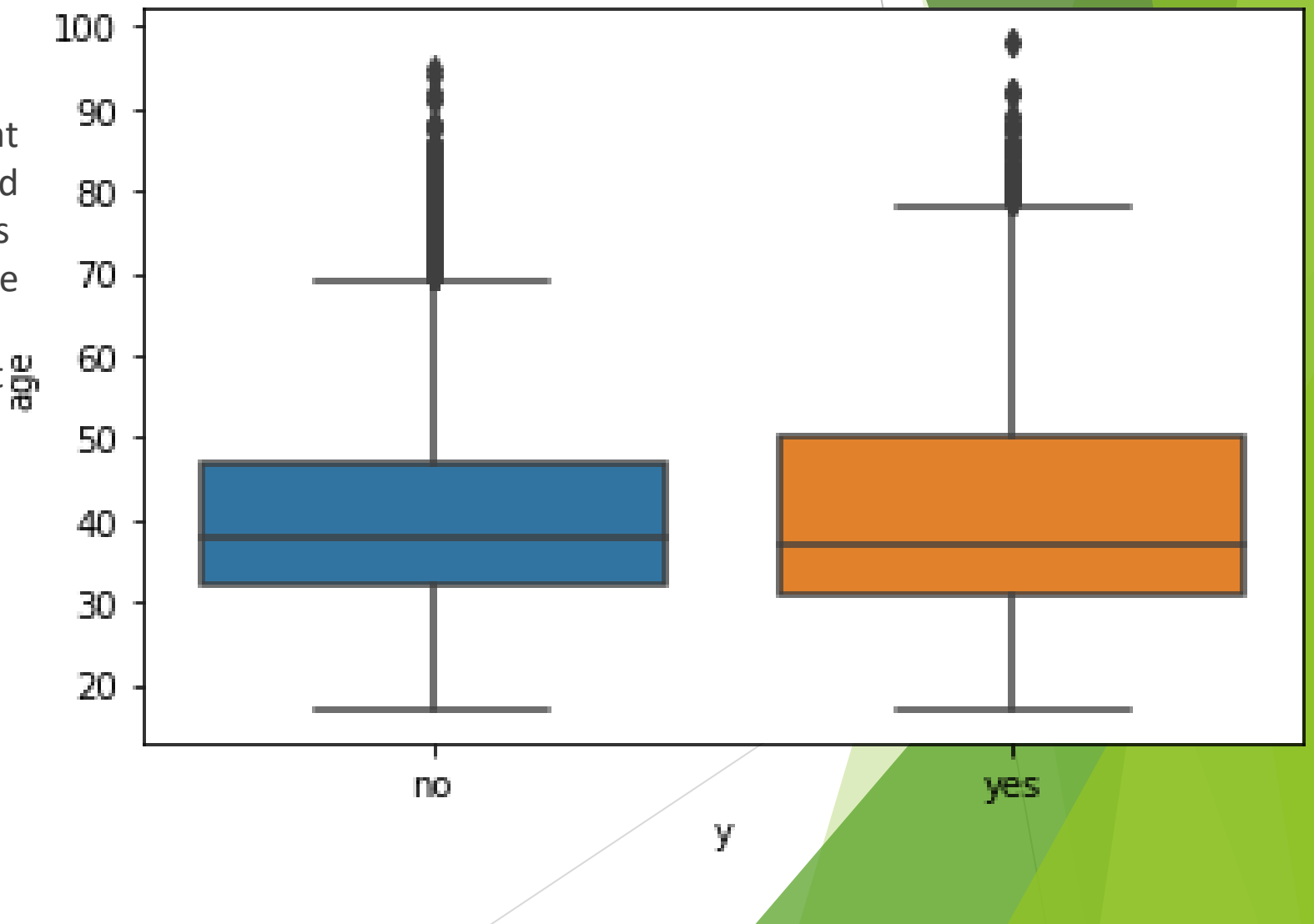


- **EDA**
  Numerical Variables:

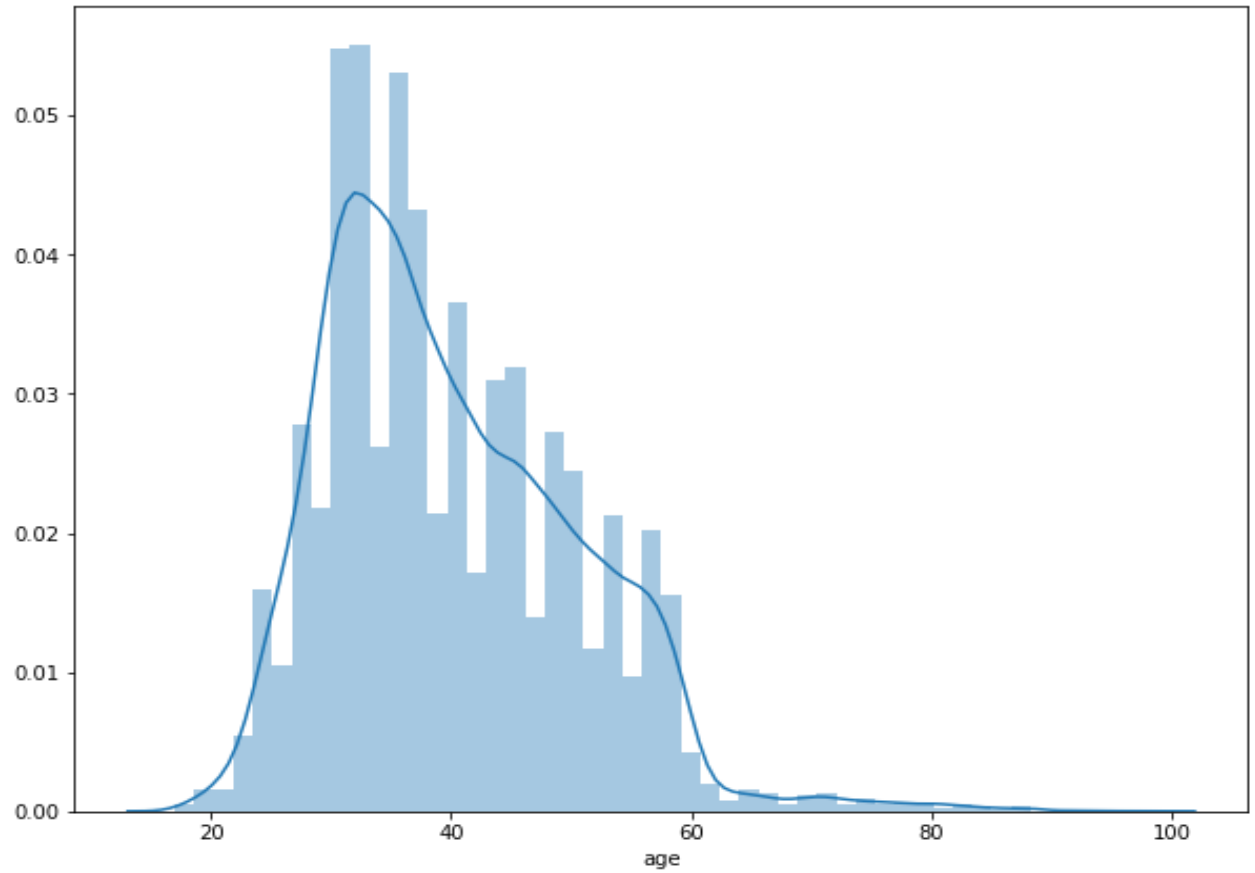
- Now will I perform some exploratory data analysis on the numerical variables.

➢ Feature: age (Numeric)

- From the plot as shown ,we know that for both the customers that subscribed or didn't subscribe 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.

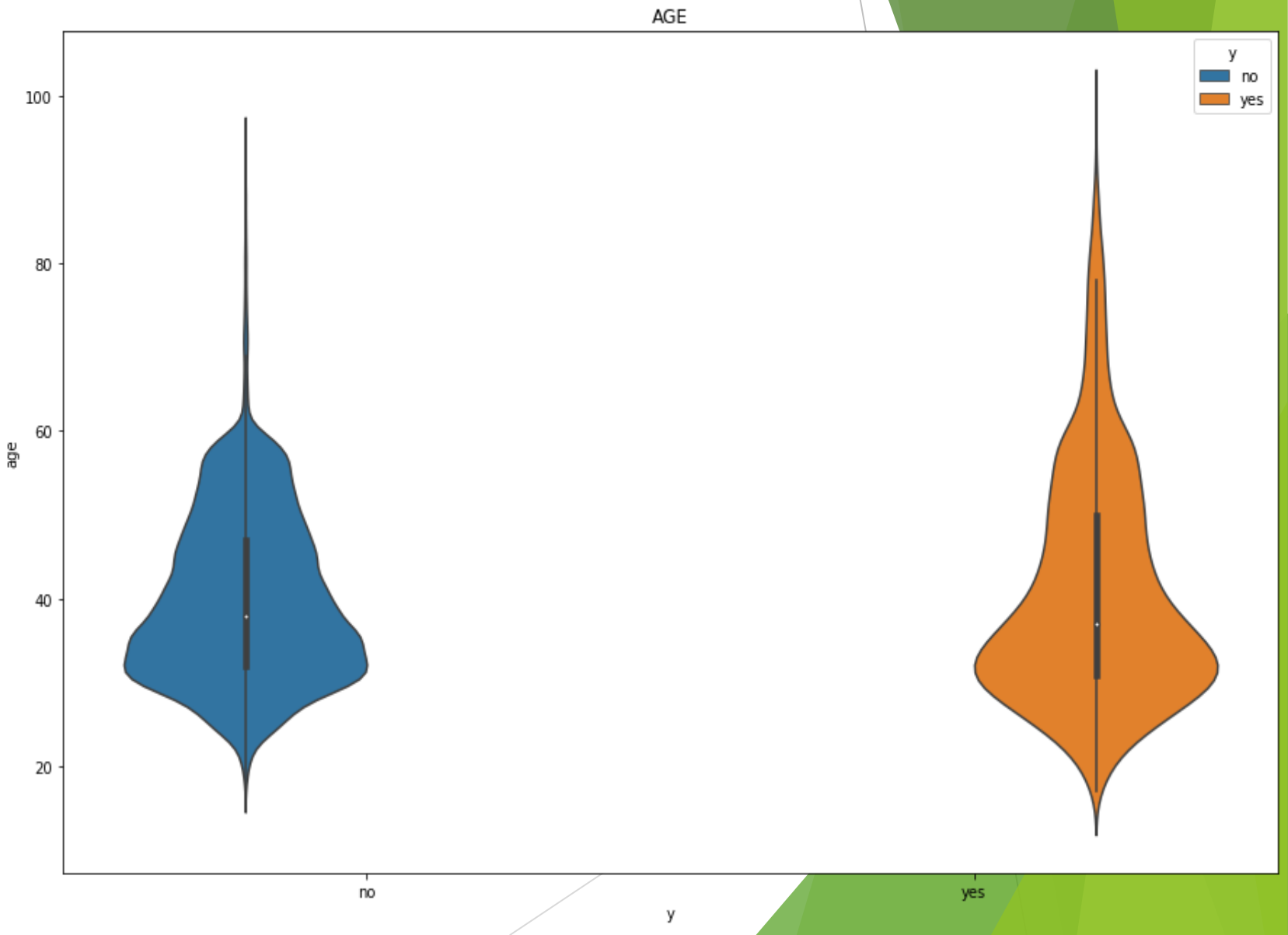


- From the plot as shown below ,it is a right skewed graph and there is an evidence of outliers after the age of 60.
  We also see in the distribution that most of the customers are in the age range of 30-40.

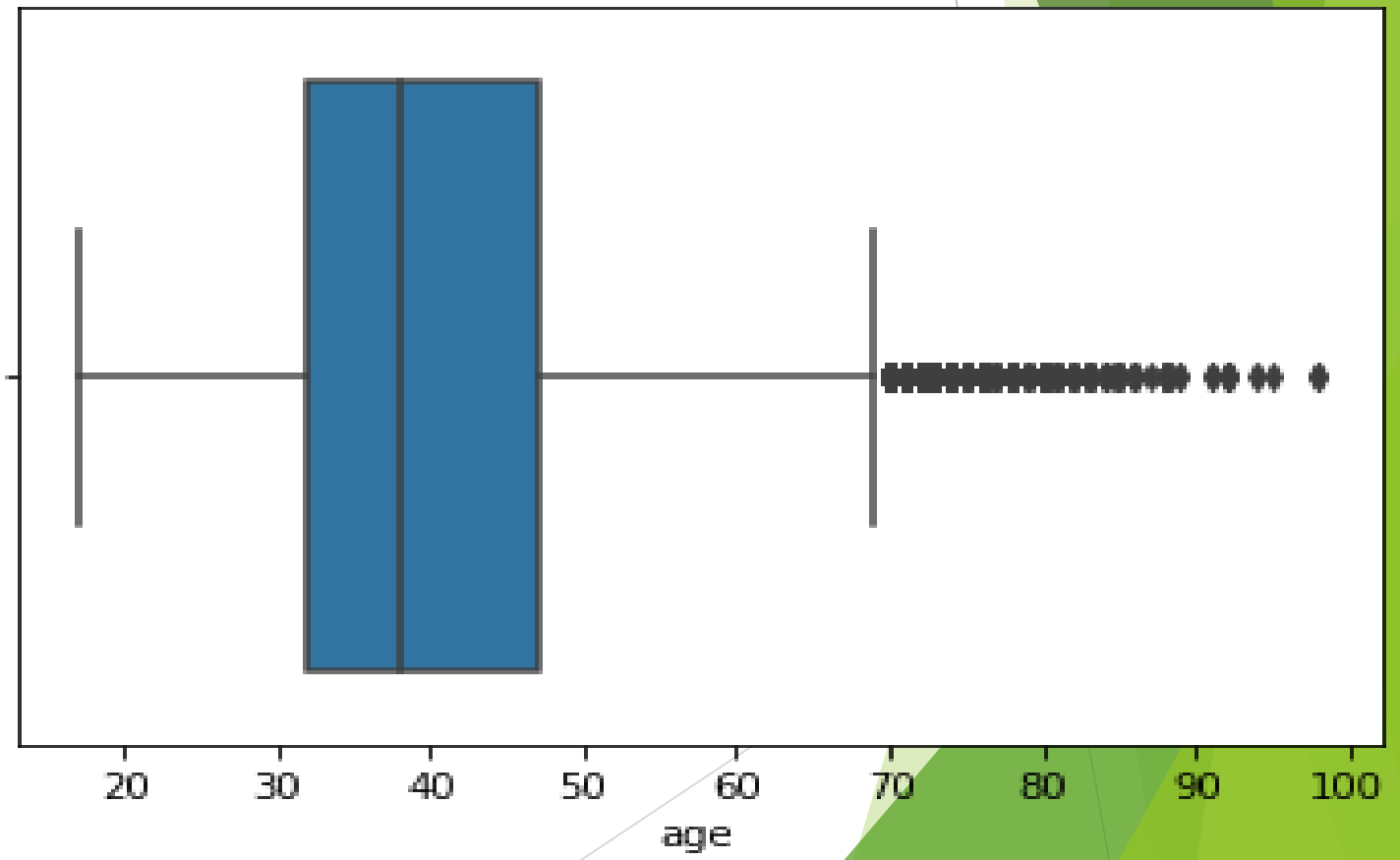


However, we can use other plots to check for outliers ,Violin plot and box plot.

- Using the Violin plot as shown , 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.

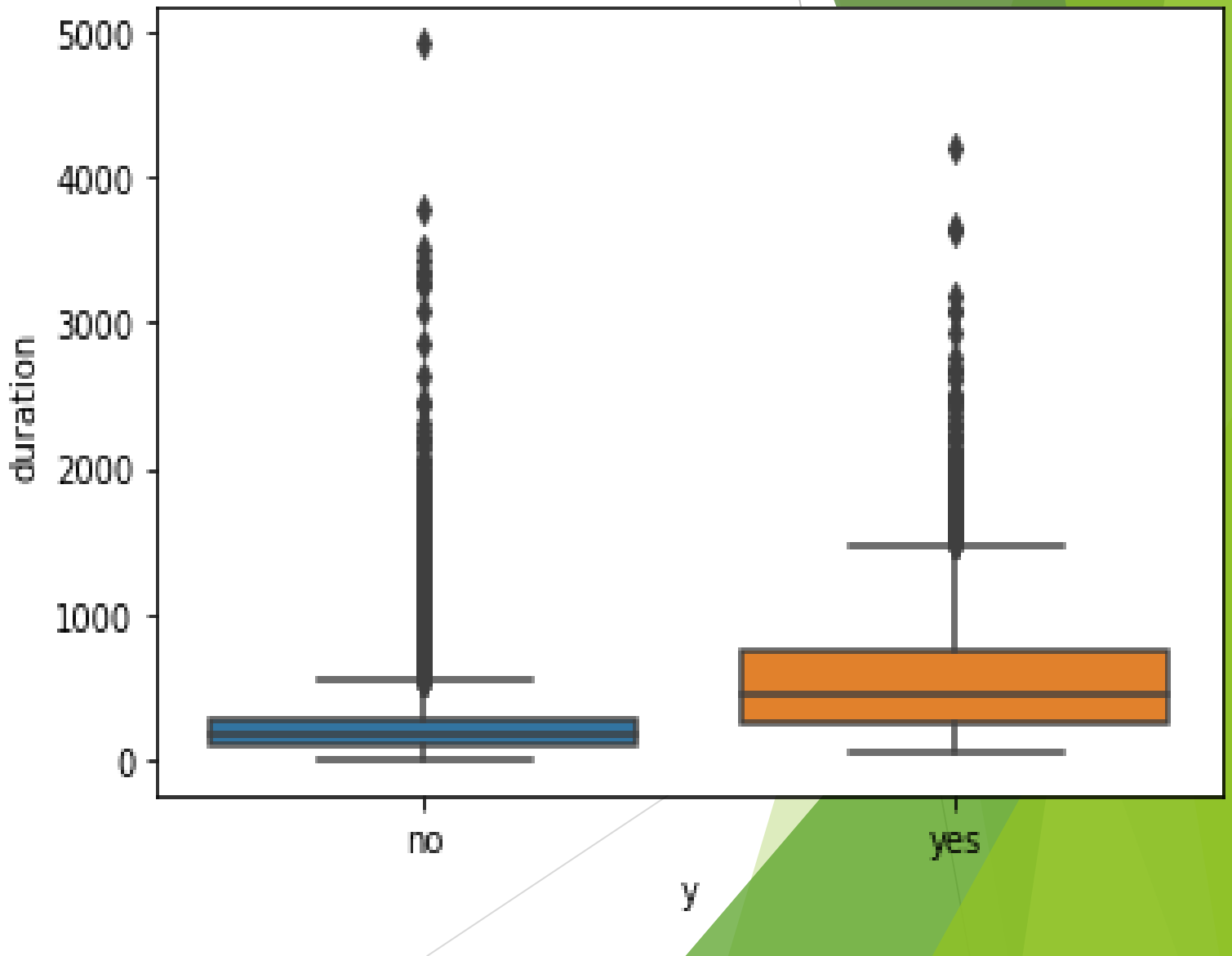


- Using the box plot as shown , we can say that outliers are present after the age of 70 years.

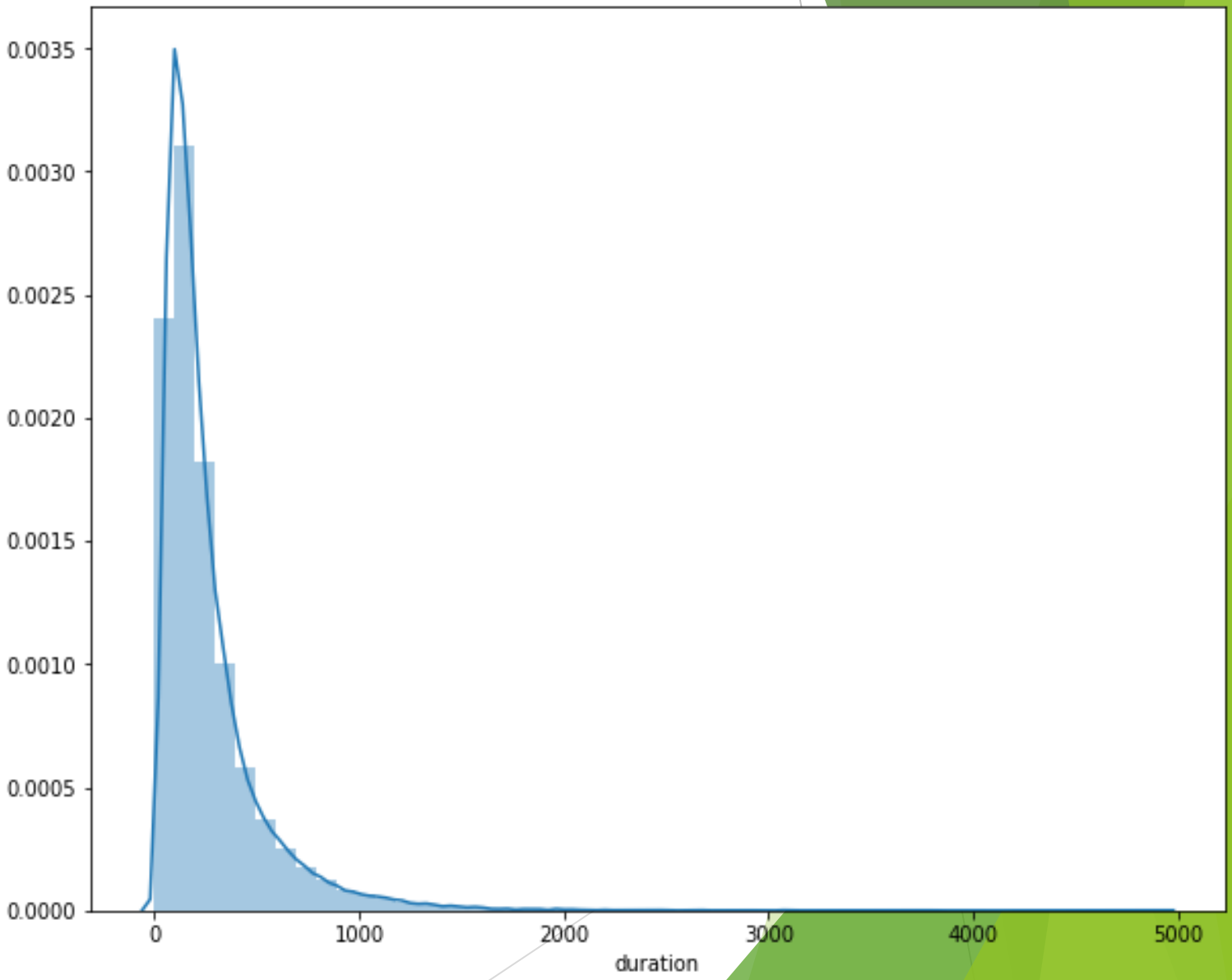


➢ Feature: duration (numeric)

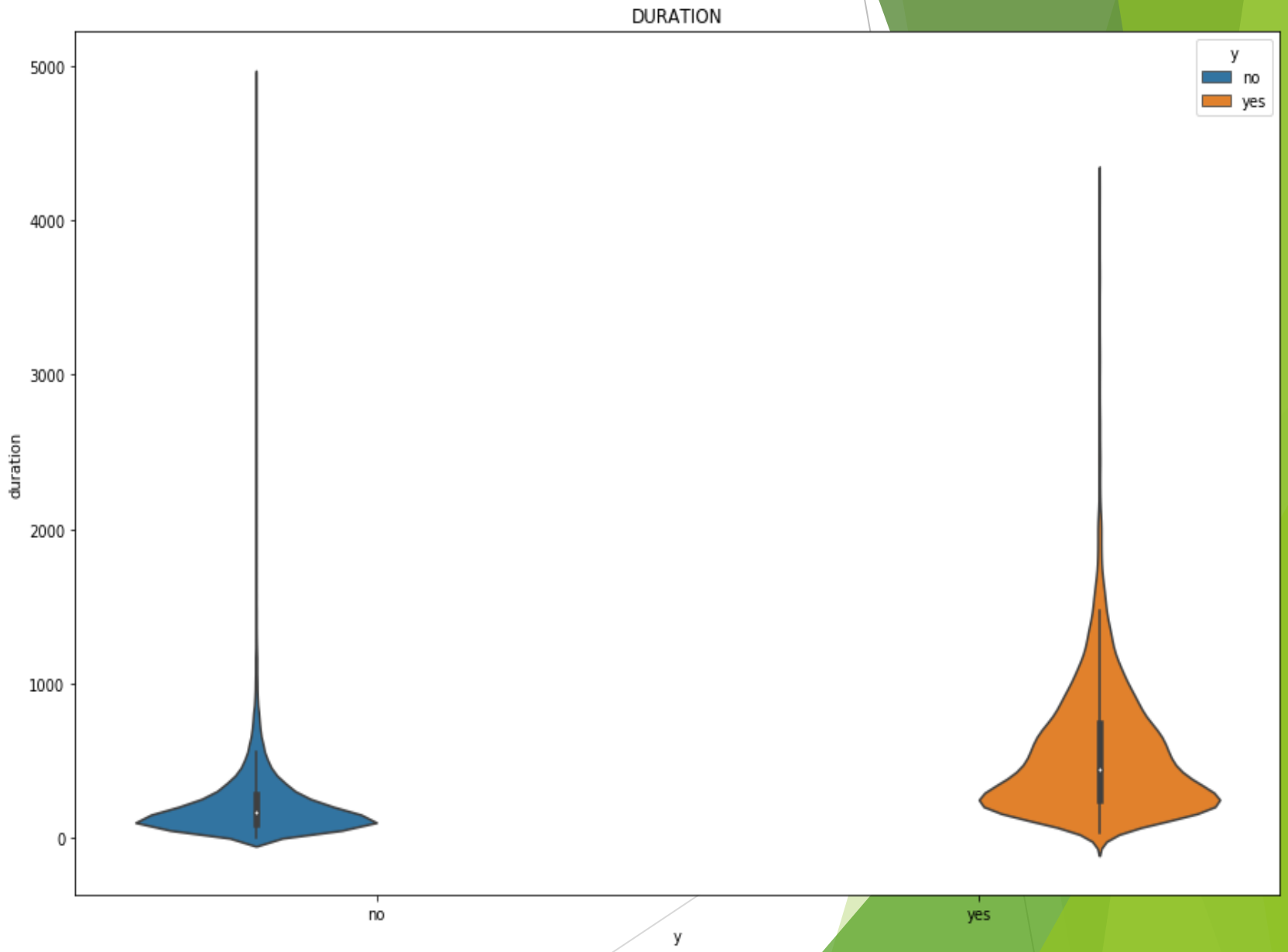
- From the plot as shown, the duration (last contact duration) of a customer can be useful for predicting the target variable.
- It is expected because it is already mentioned in the data overview that this field highly affects the target variable and should only be used for benchmark purposes.



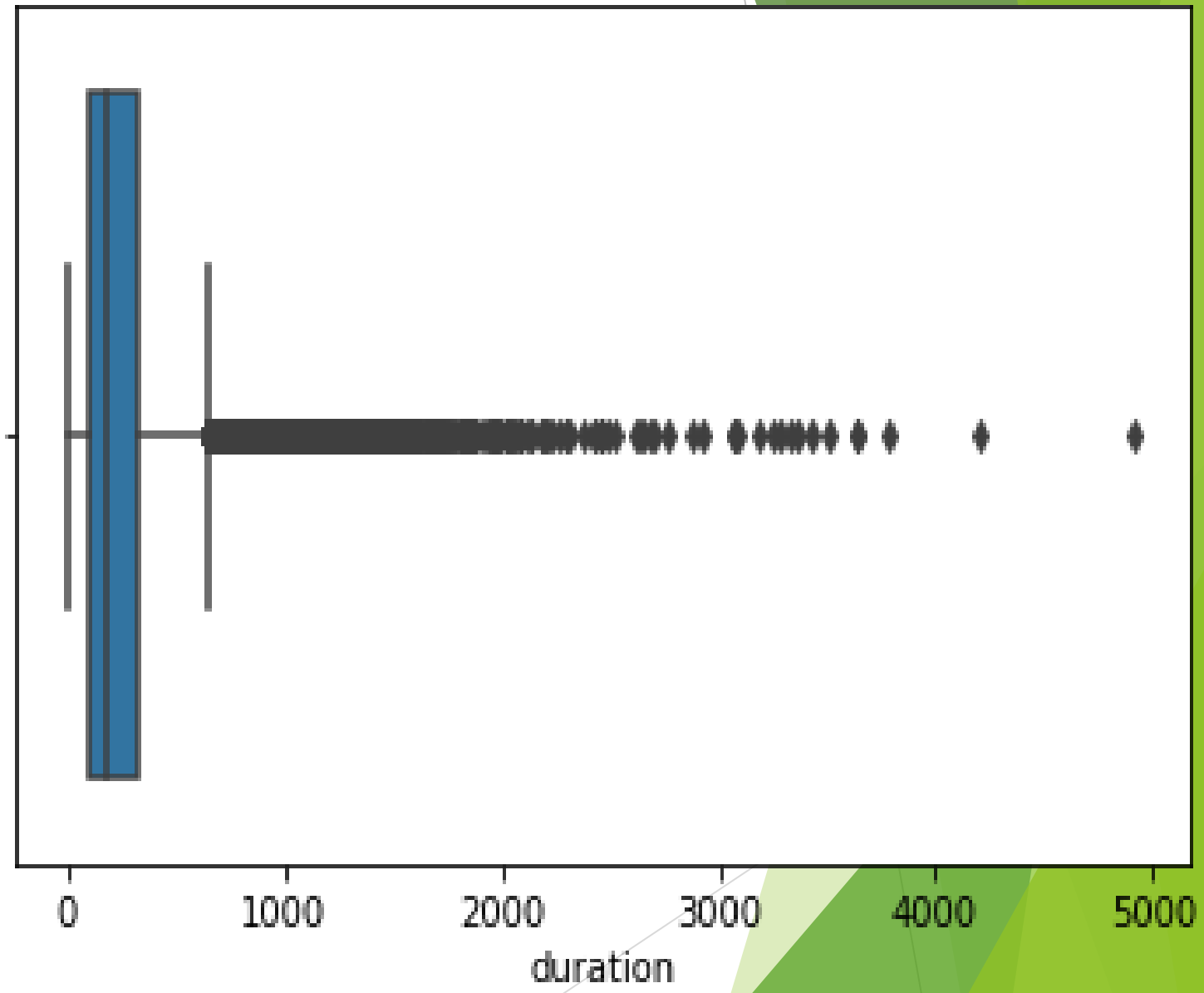
- As shown from the plot ,it is clear that from duration around 1500 onward outliers, are present and it is right skewed data.
- In this distribution above we find where most the values are very low and very few have high values.



- Using the Violin plot as shown , 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.

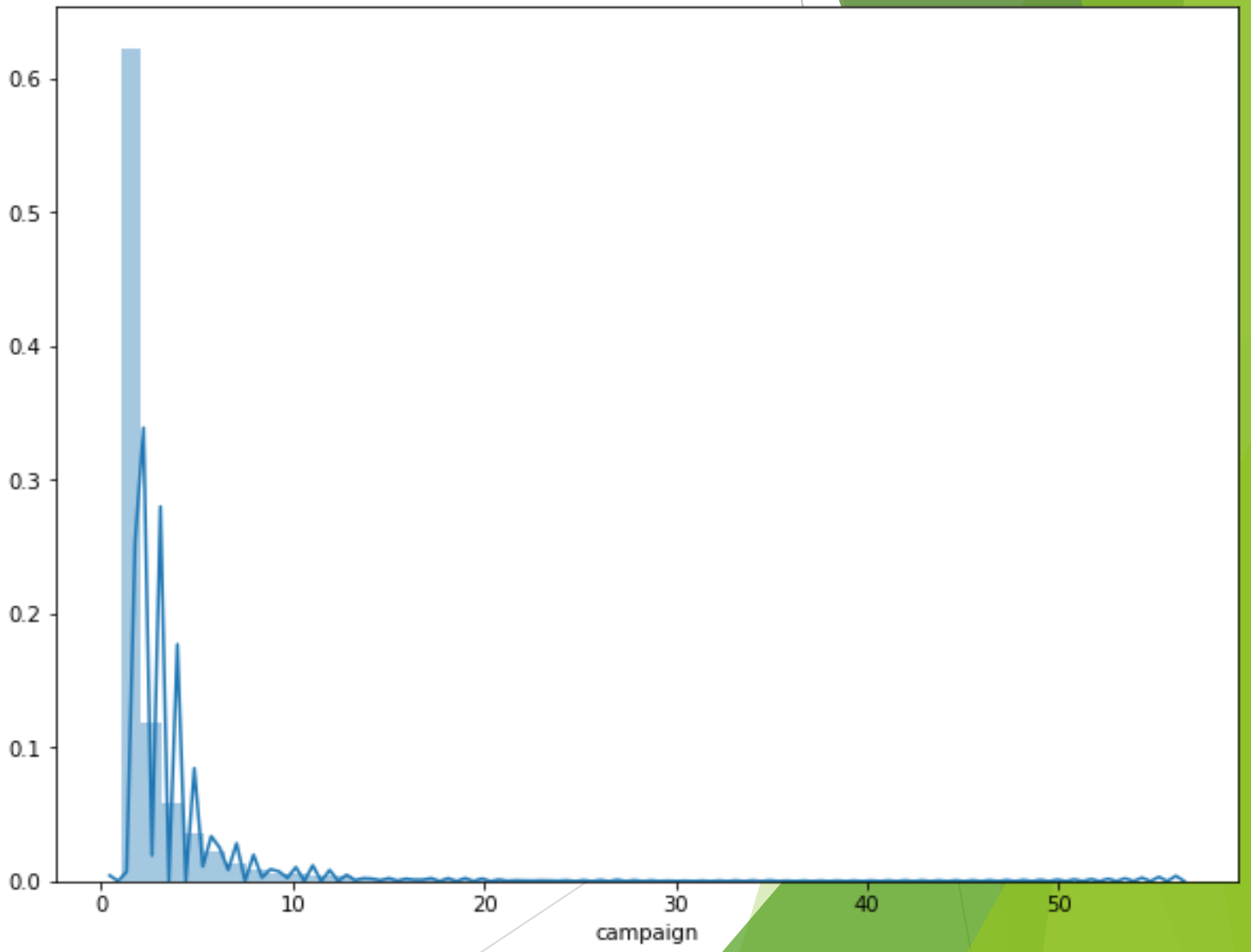


- Using the box plot as shown , we can say that outliers are present after the duration of calls more than 1000.

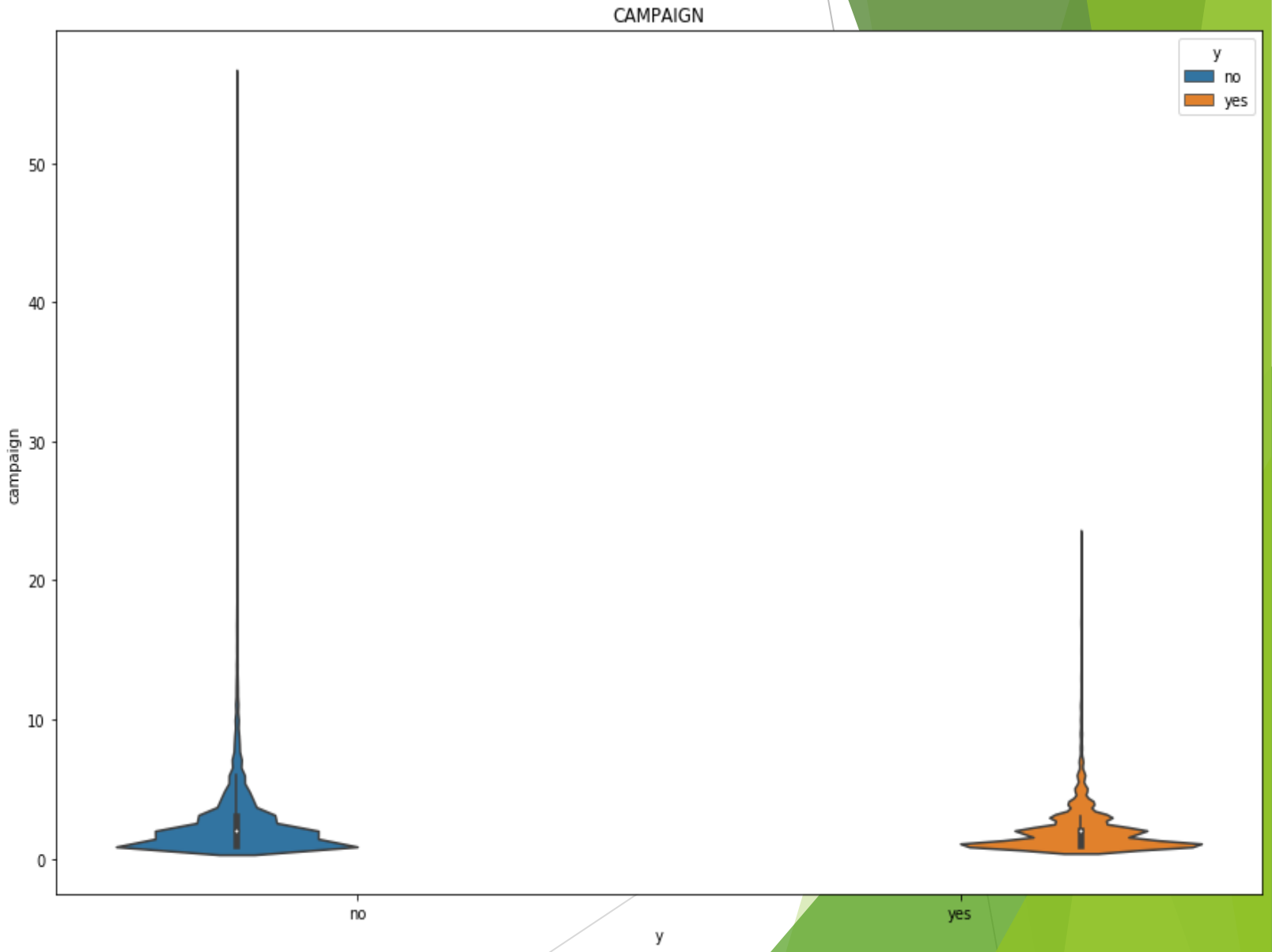


➢ Feature: campaign (numeric)

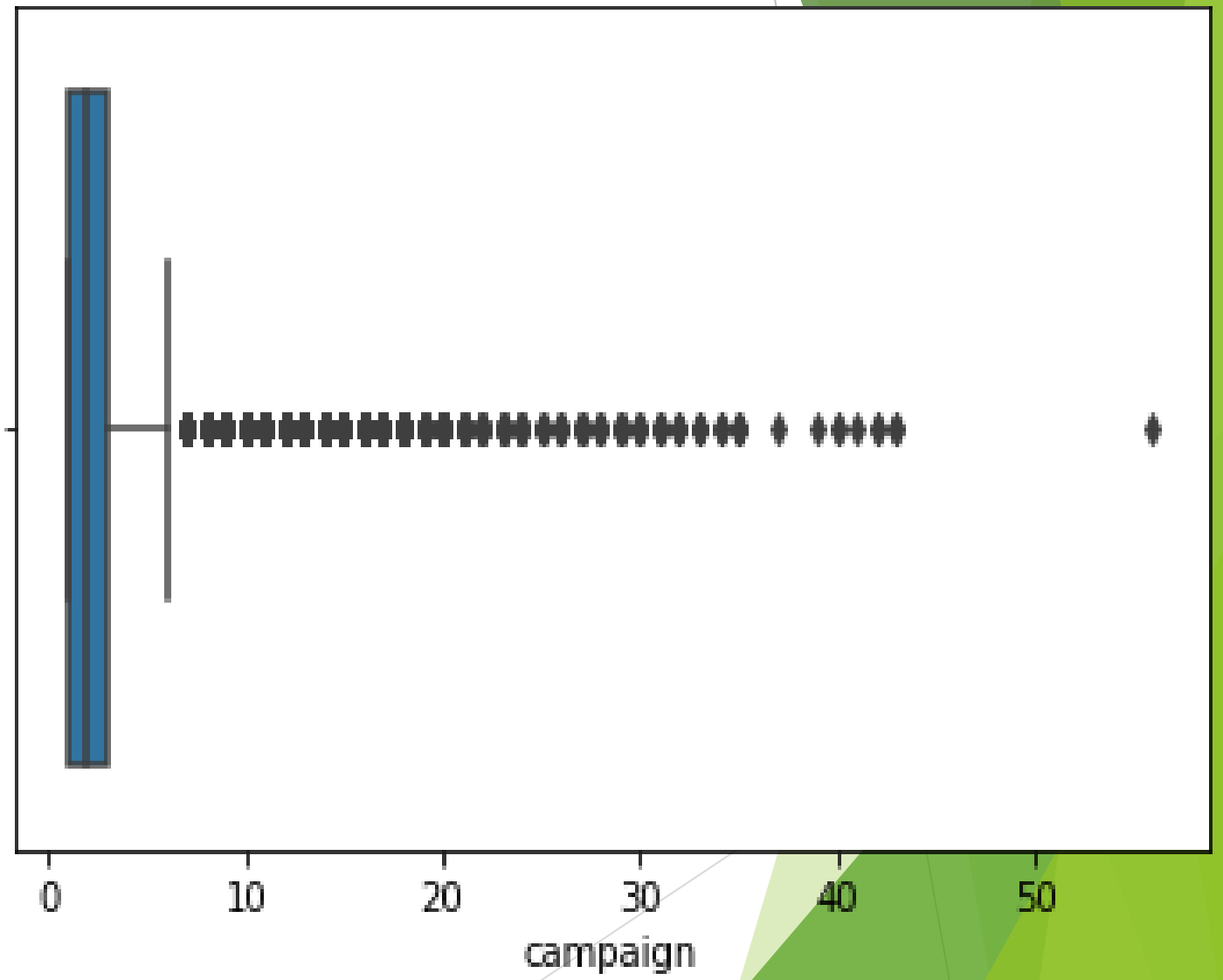
- As we can see from the plot, outliers might be present after the number of campaigns 10.



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

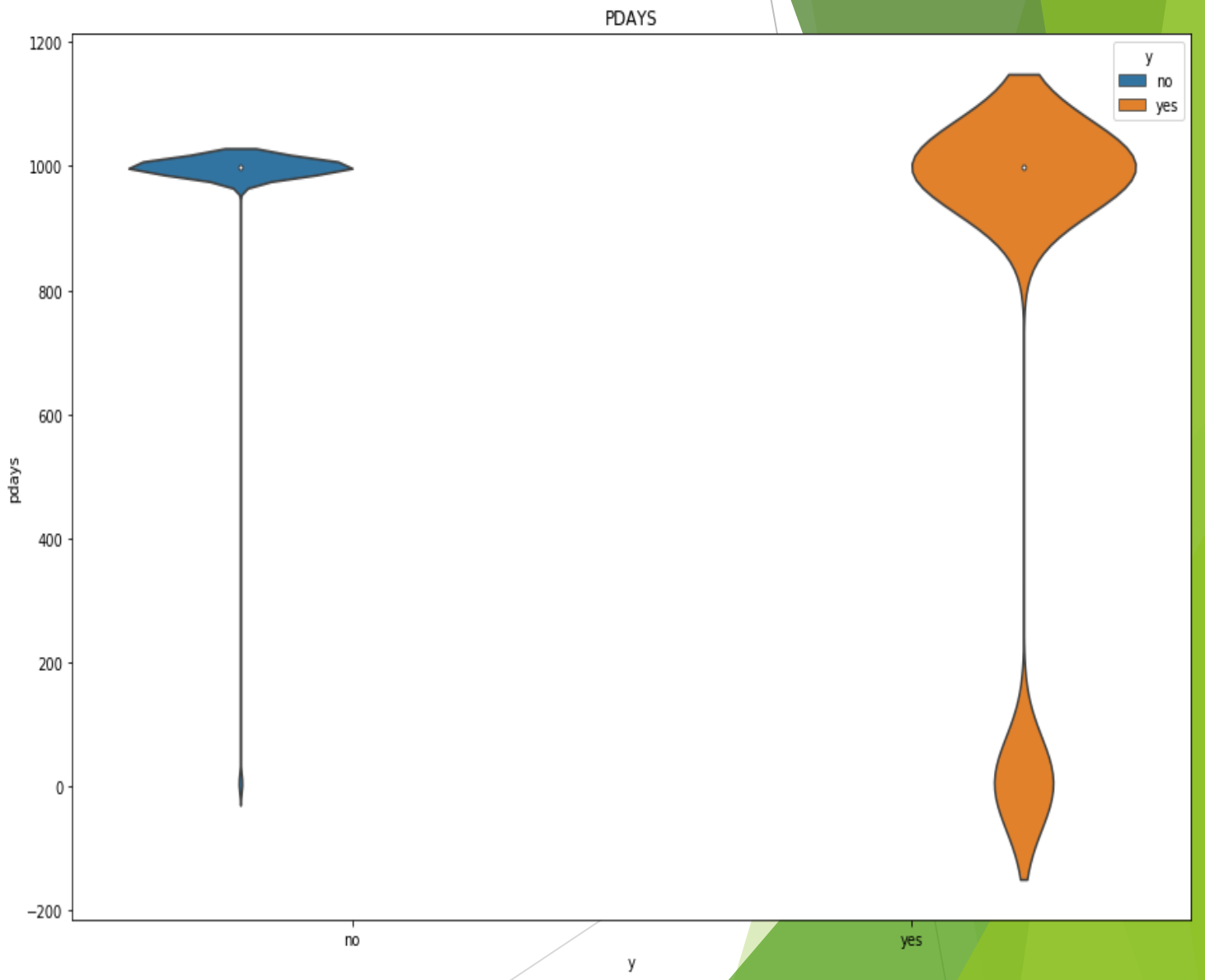


- Using the box plot as shown, we can see those campaigns more than 10 are considered as outliers.

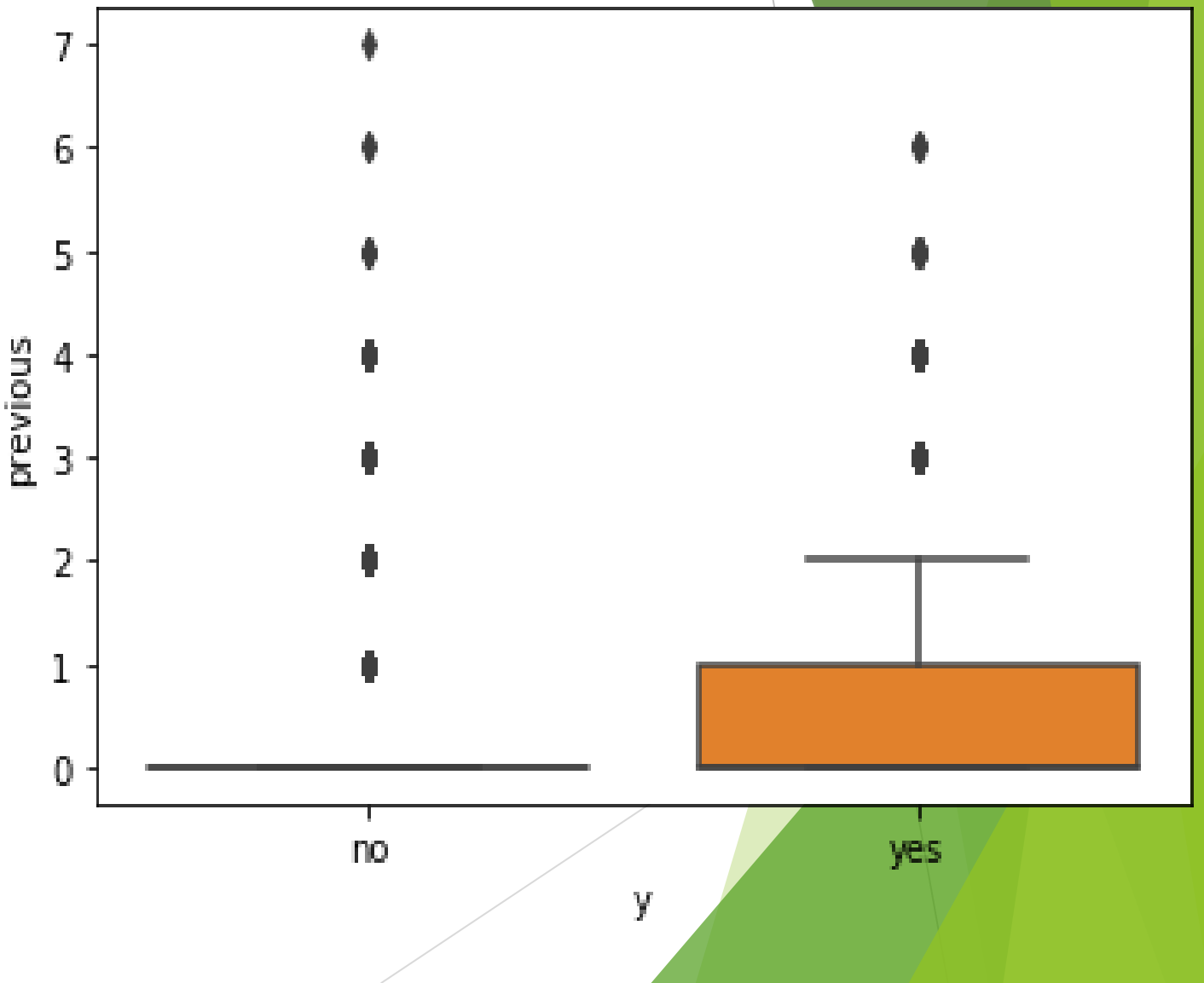


➢ Feature: pdays (numeric)

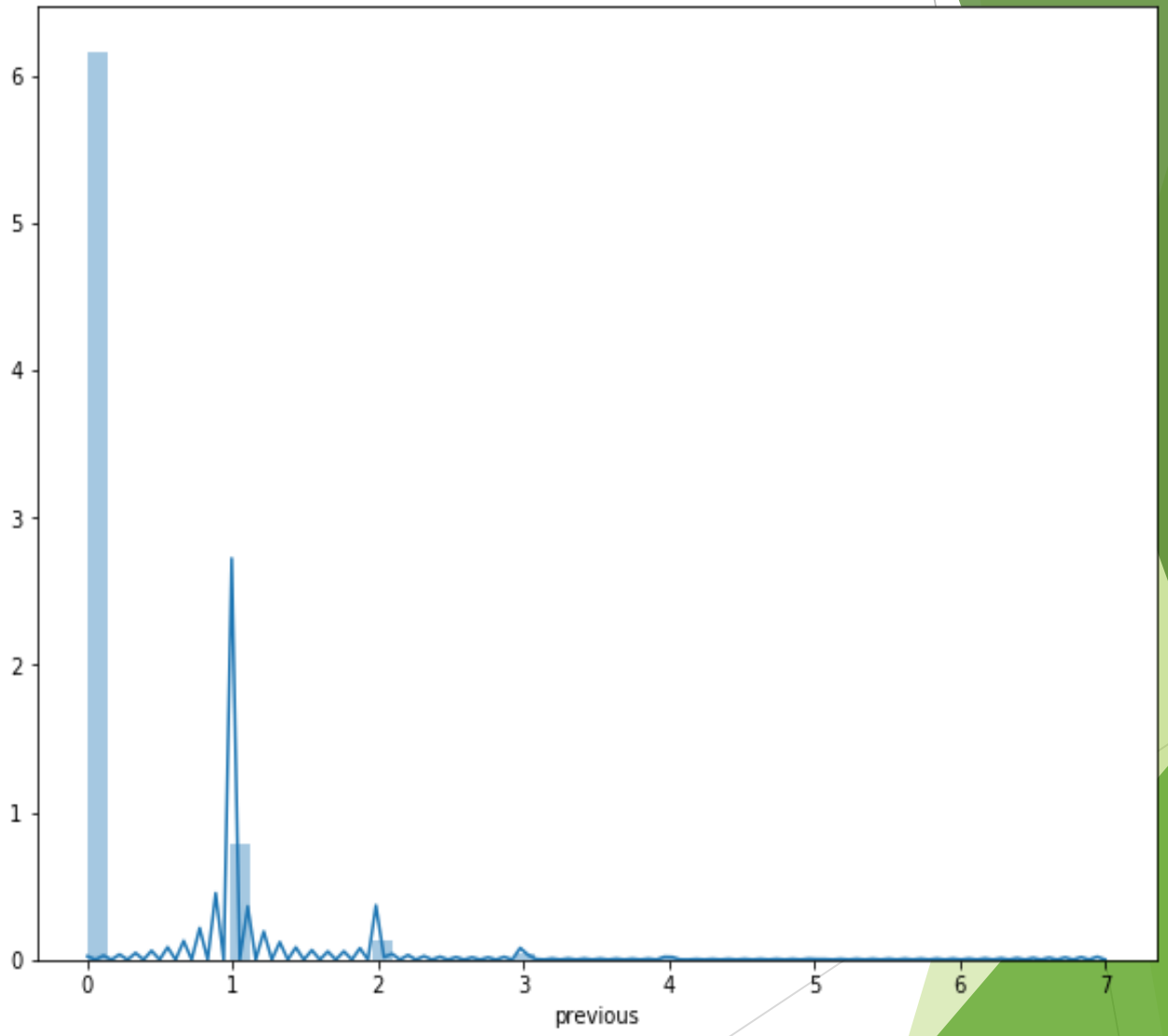
- From the Violin plot it is visible that irrespective of class labels, mostly people have not been contacted by the bank. Very people have been contacted by the bank and number of days passed for previous campaign is between 0–100.
- It means we either must compute pdays or drop the pdays depends on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



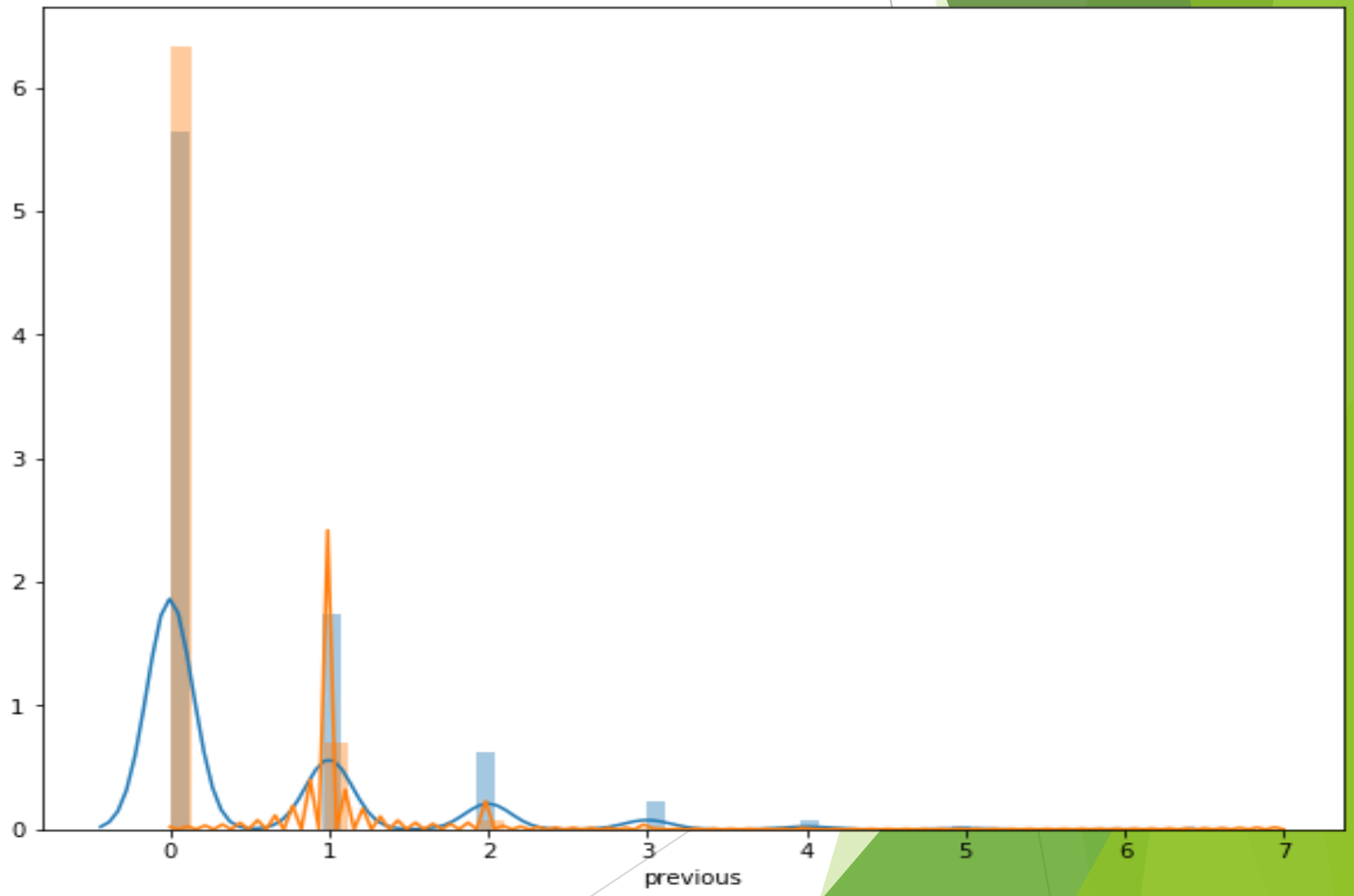
- **Feature: previous (numeric)**



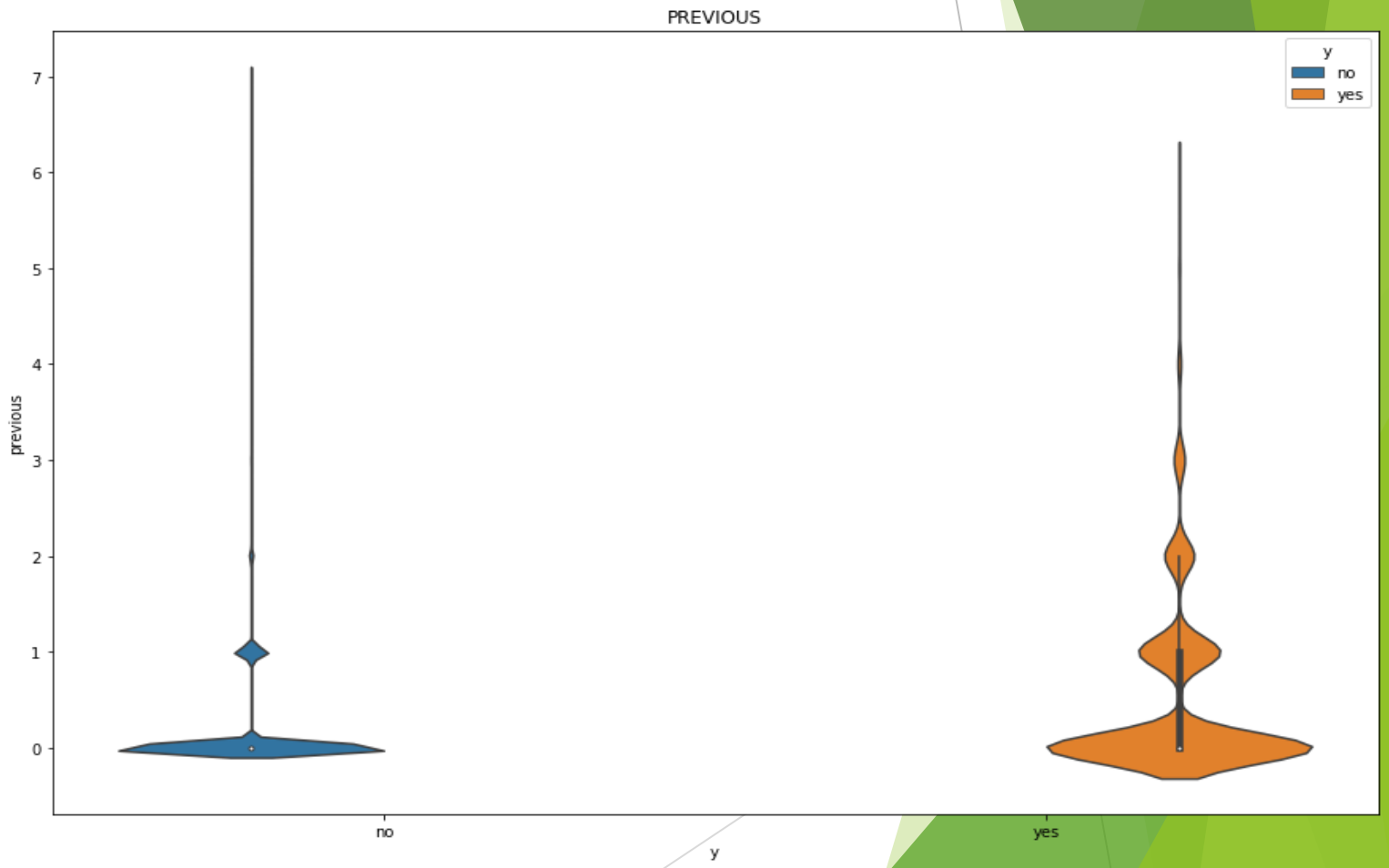
- From the plot as shown , there might be outliers present after the number 4.



- 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.

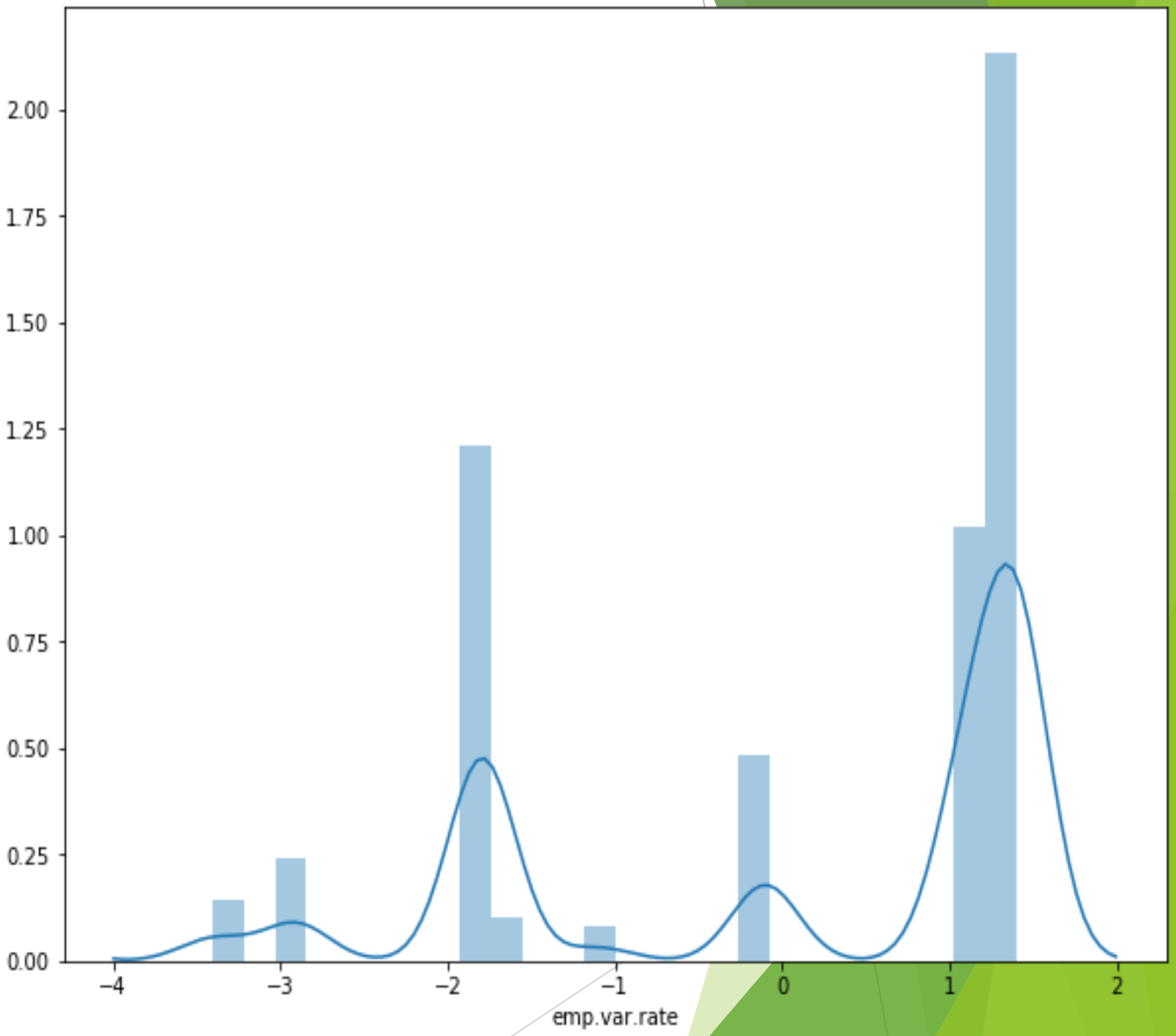


- Using the Violin plot as shown as we can see, people who have contacted once for the previous campaign has subscribed for long term deposits.
- For class no, there are so many outliers starting with value 1 but for yes class, outliers are present from value 3.
- From the plot it is visible that previous feature would be helpful in predicting the class labels.

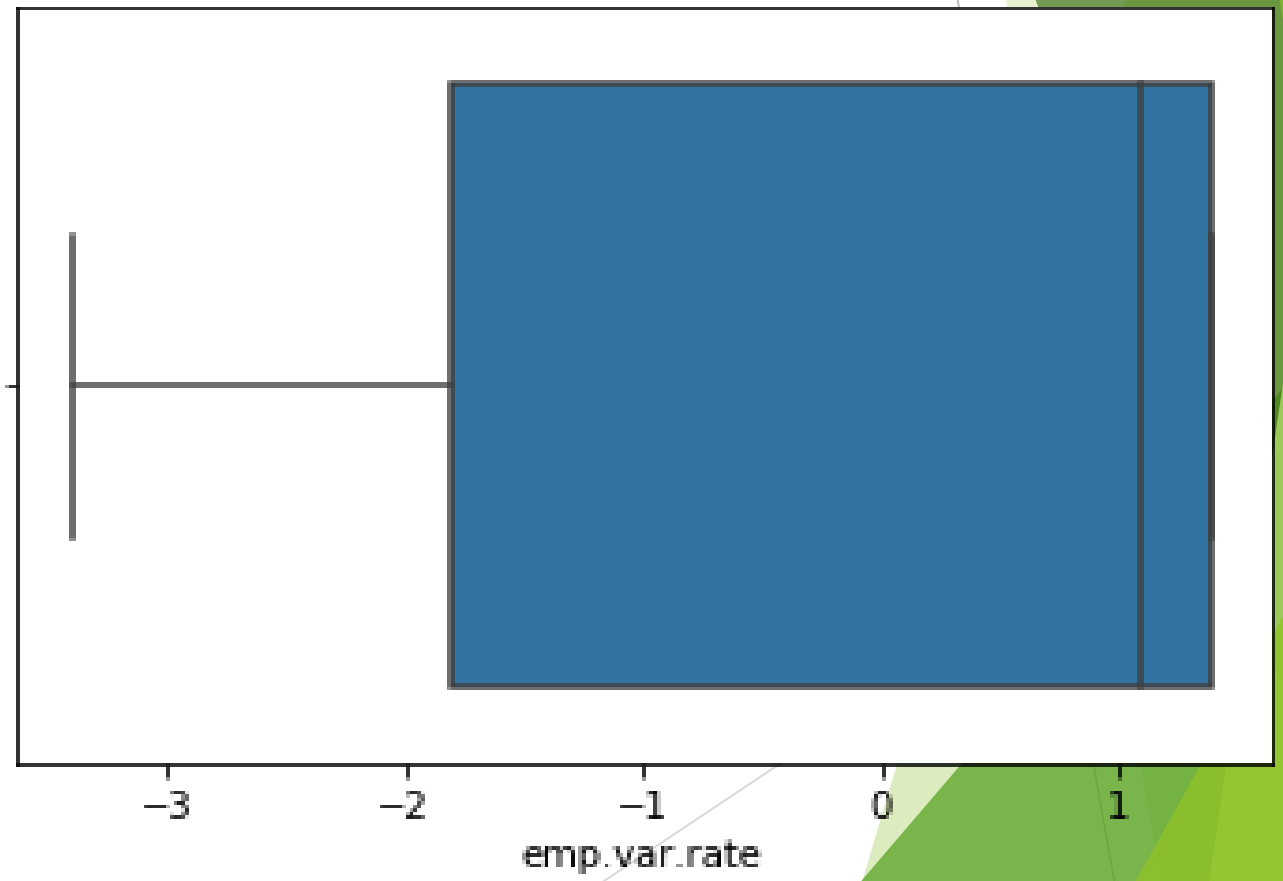


➢ Feature:emp.Var.Rate(numeric)

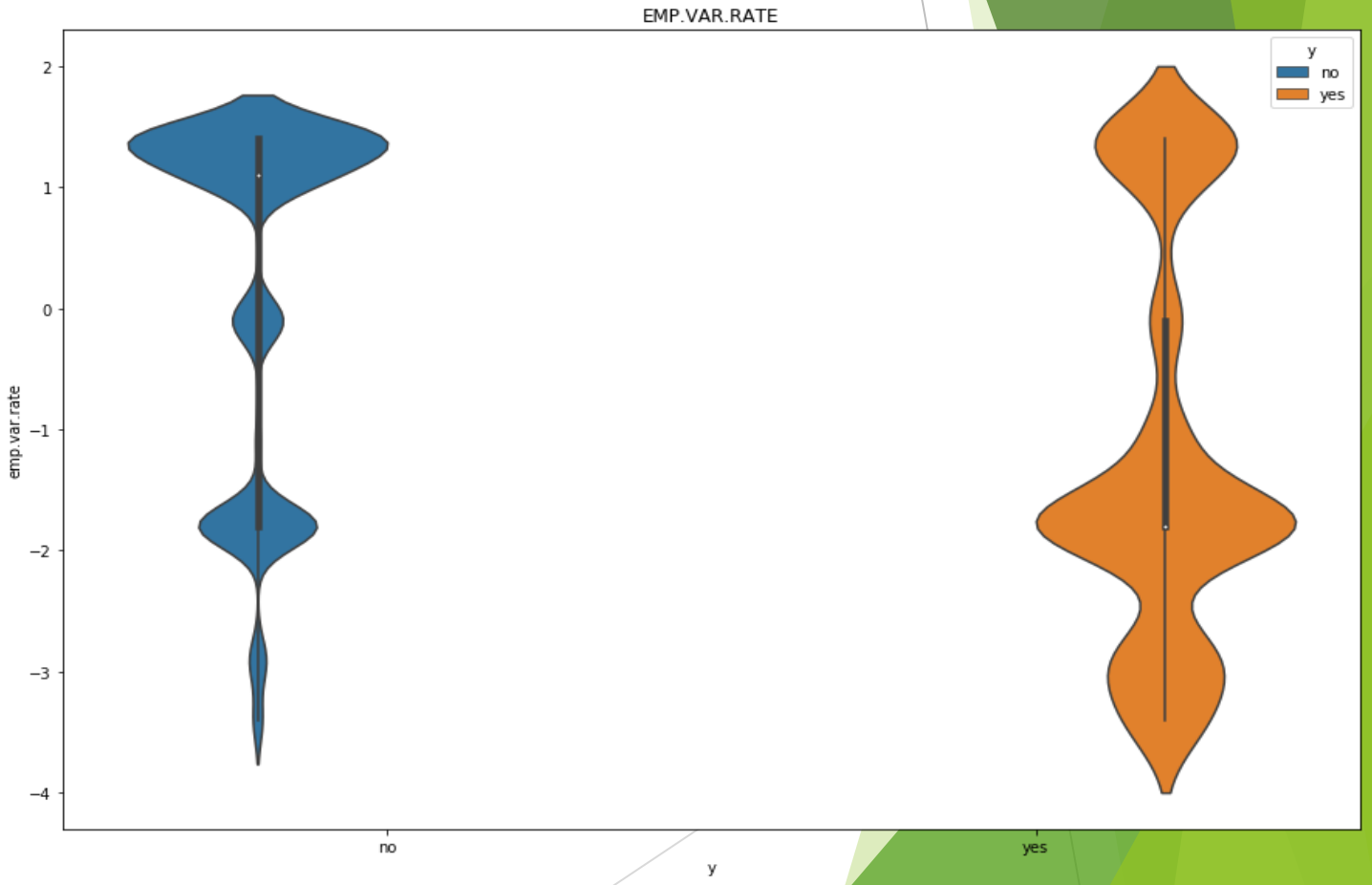
- From the plot it is visible that there are no outliers present.



- No outliers are present as shown on the boxplot.

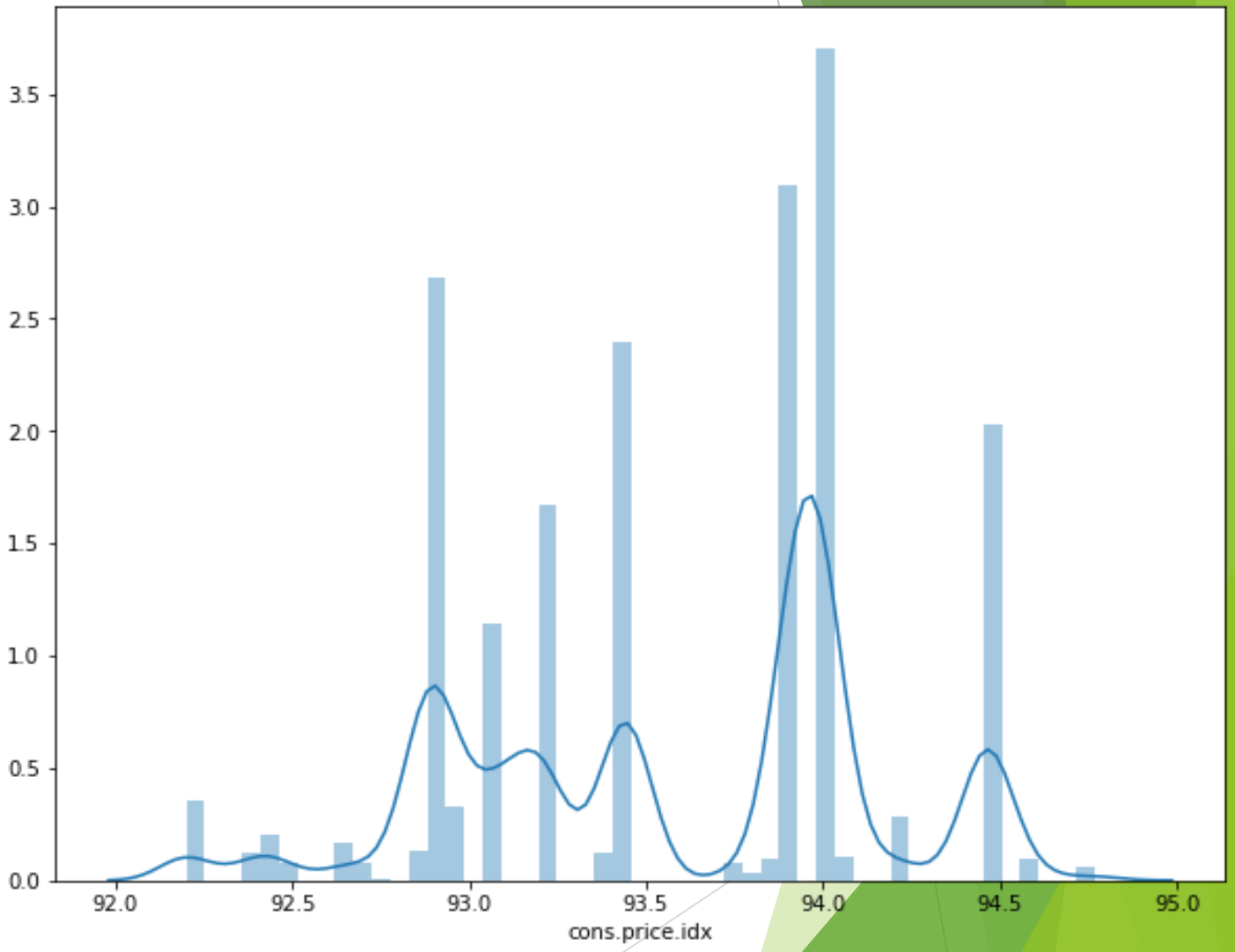


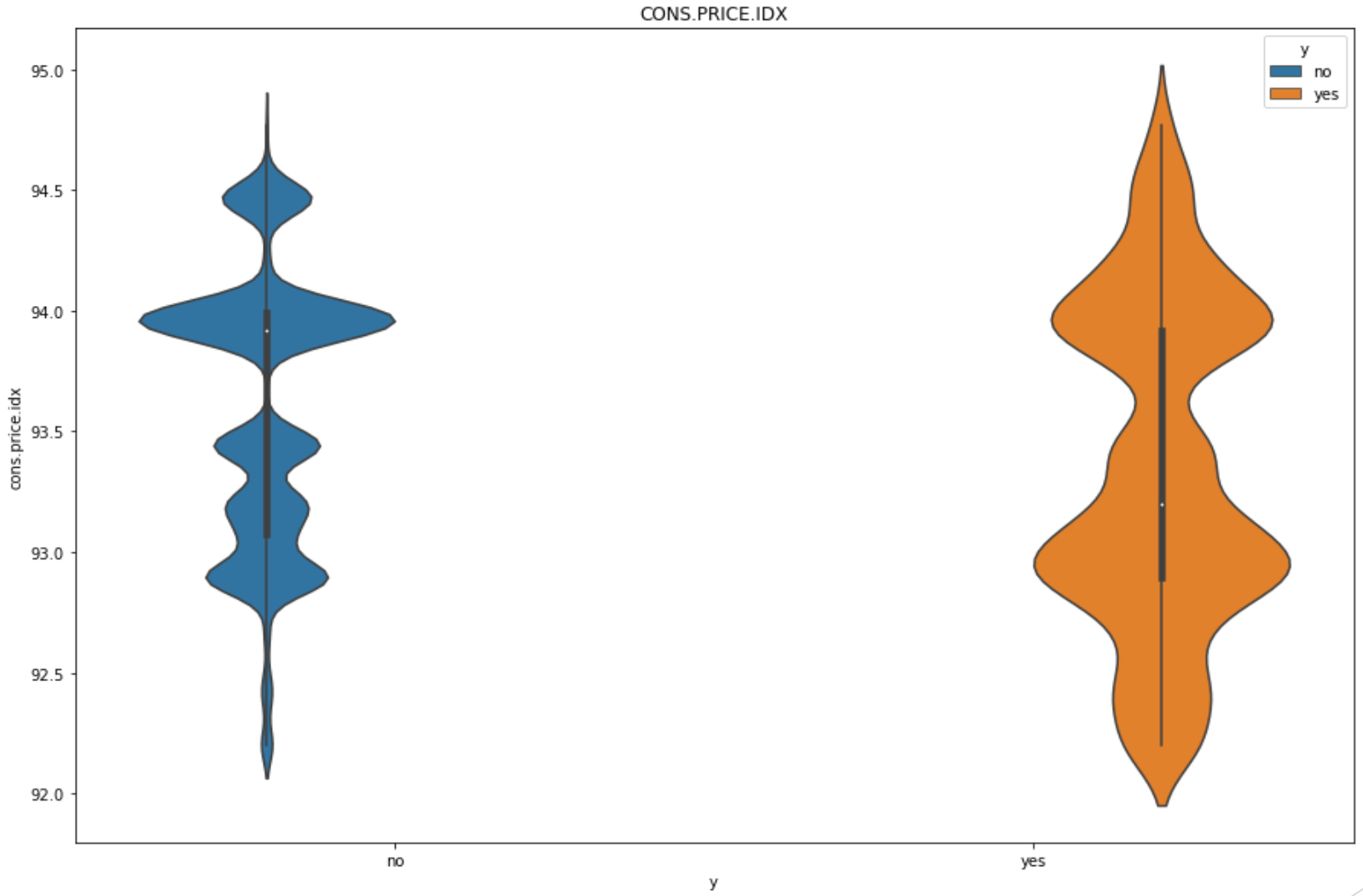
- There are no outliers present for any class for this feature and emp.var. rate feature would be very useful in predicting labels.

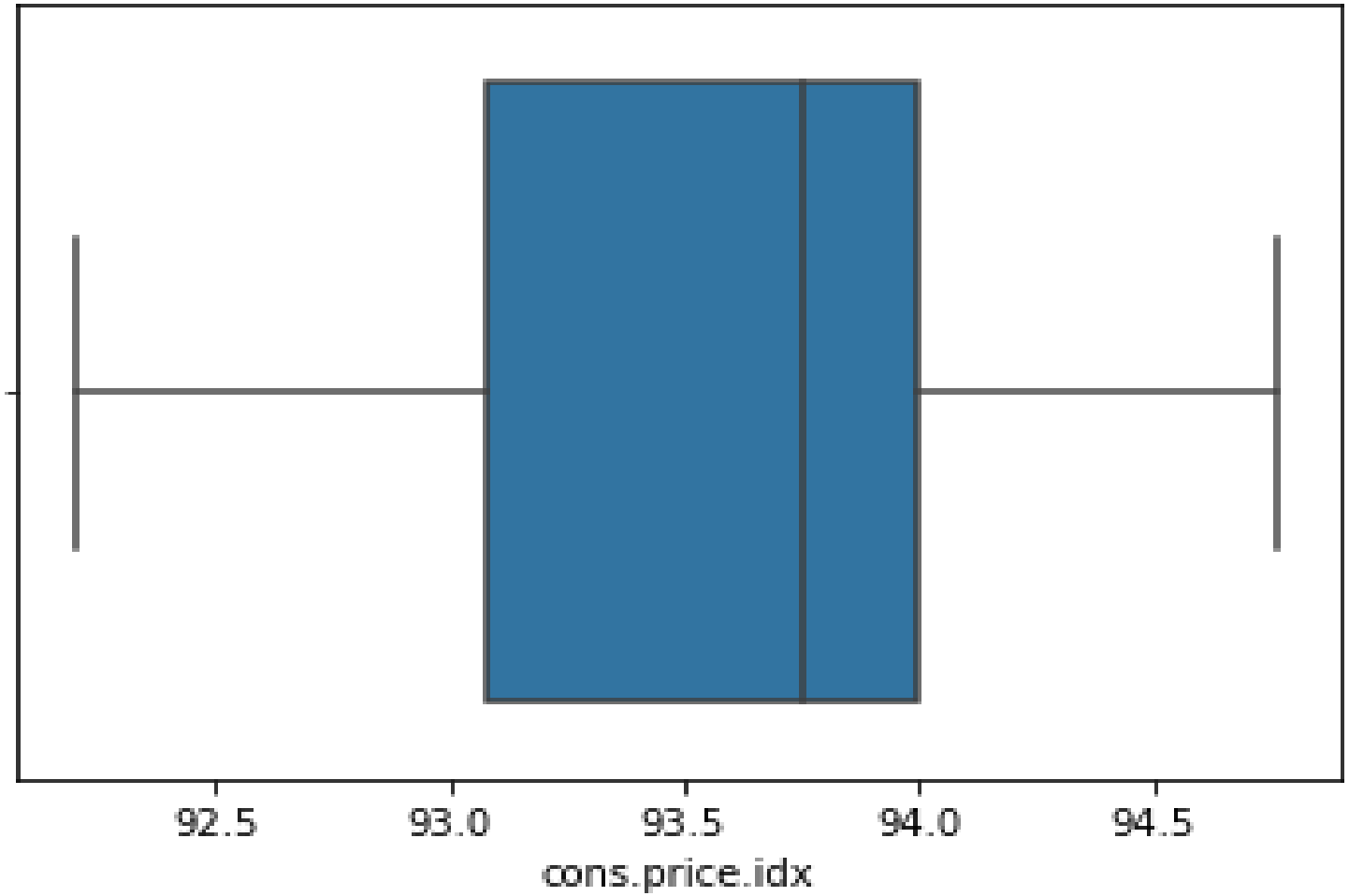


➢ Feature:cons.price.idx(numeric)

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

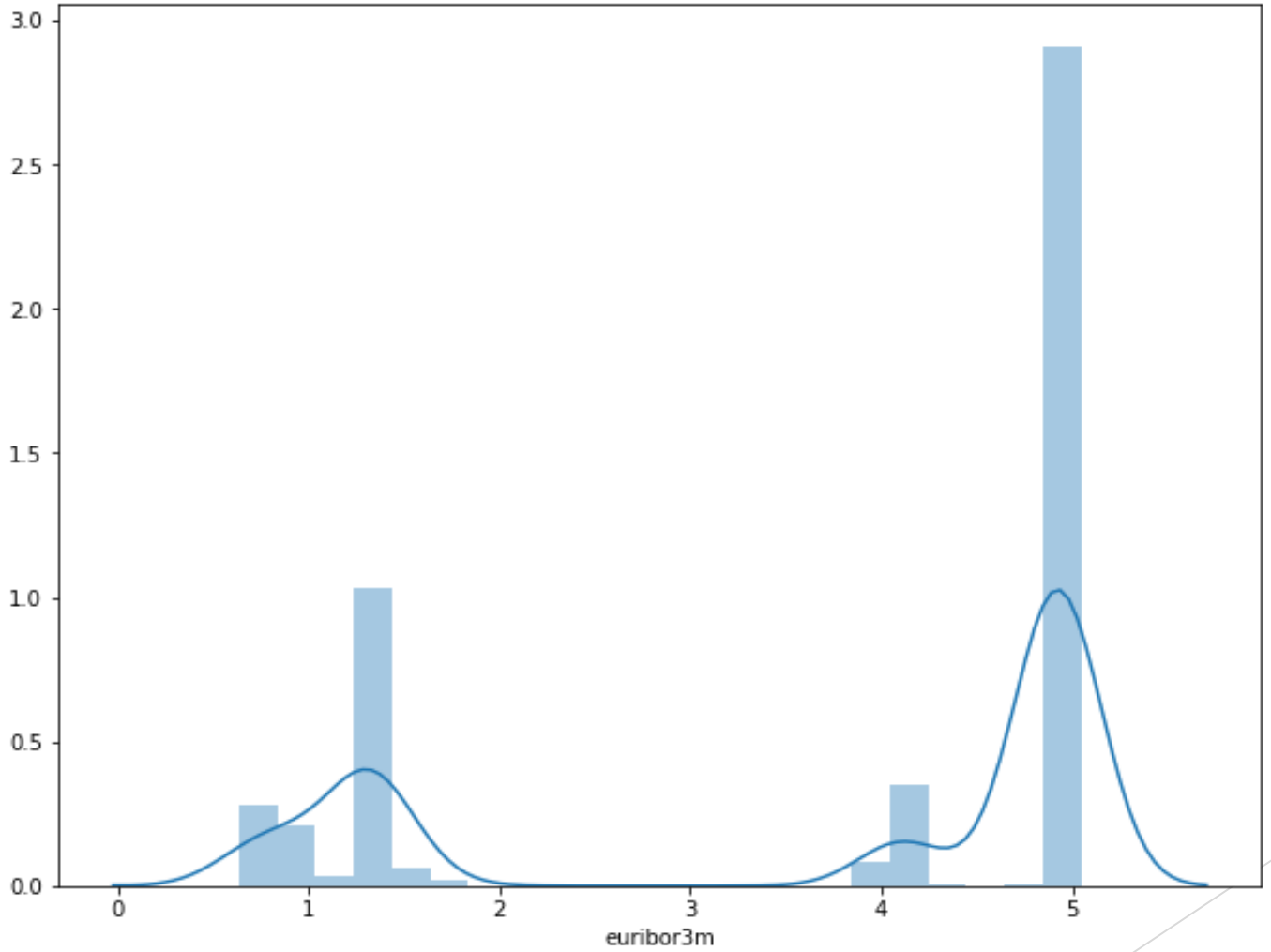


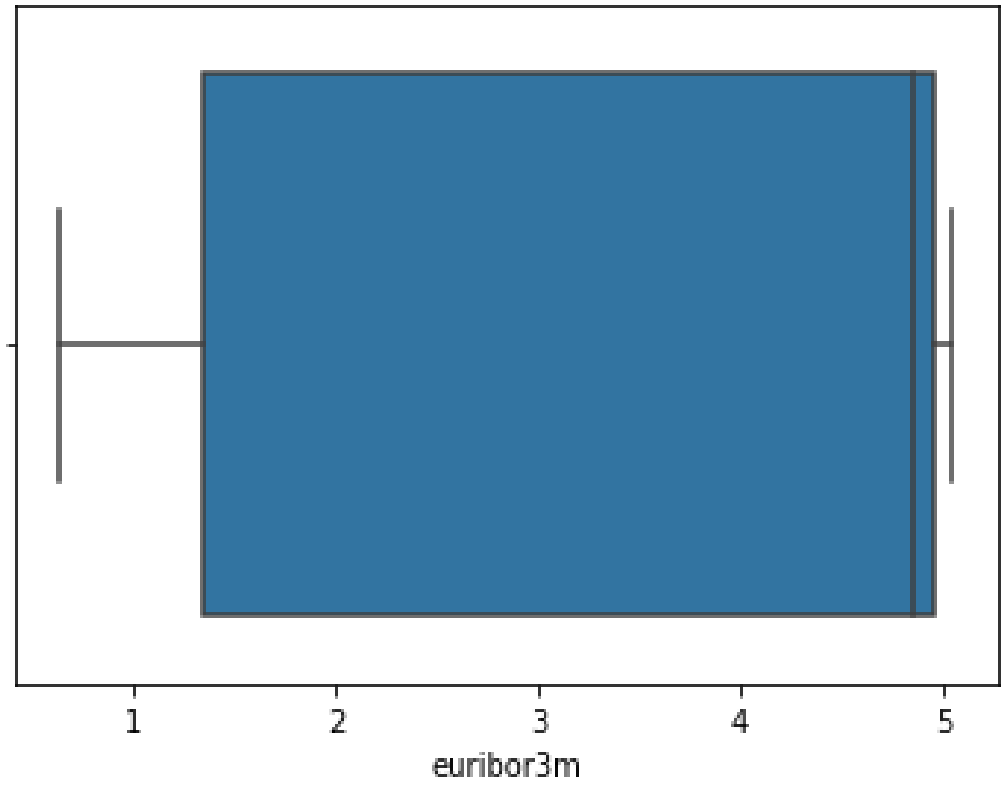

CONS.PRICE.IDX
y
no
yes
95.0
94.5
94.0
93.5
93.0
92.5
92.0
cons.price.idx
no
yes
y

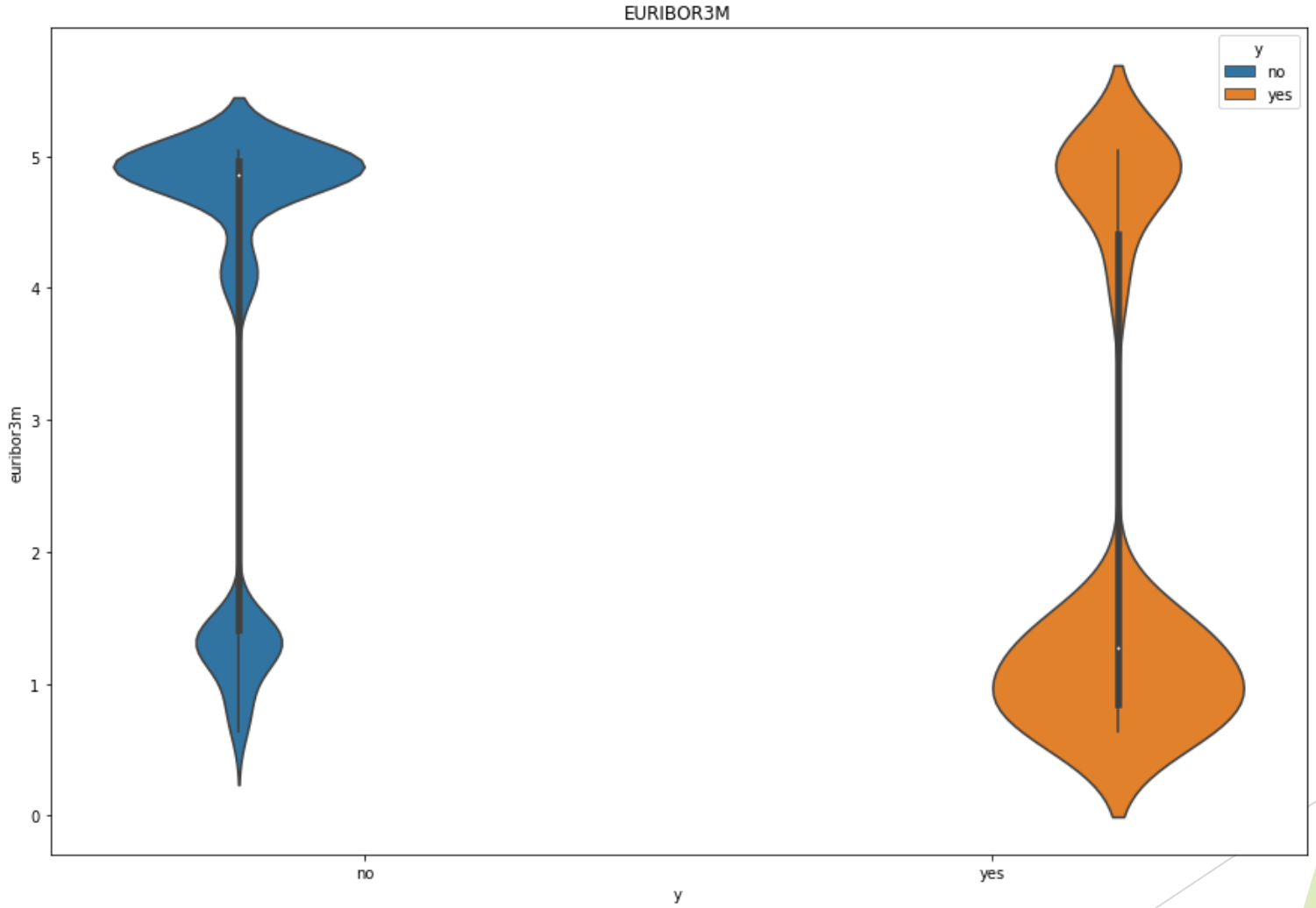

92.5
93.0
93.5
94.0
94.5
cons.price.idx

- From the 2 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)

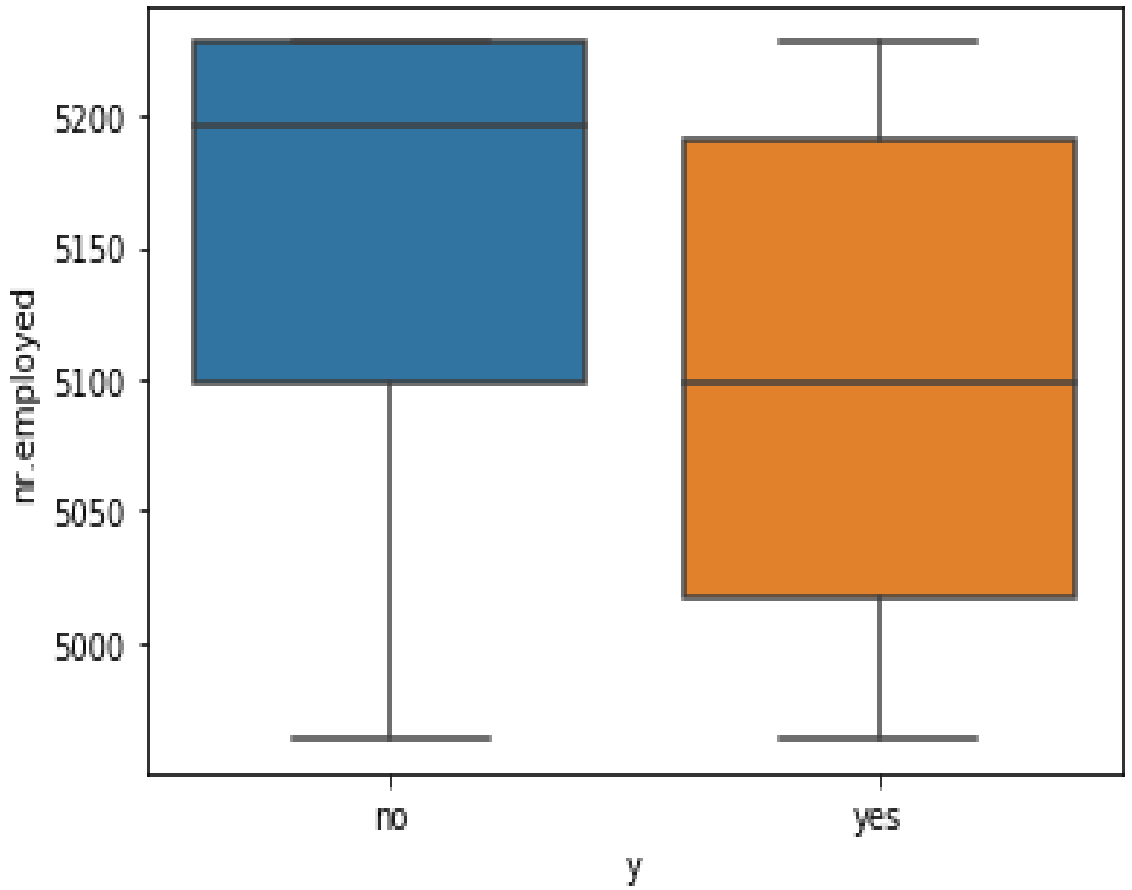


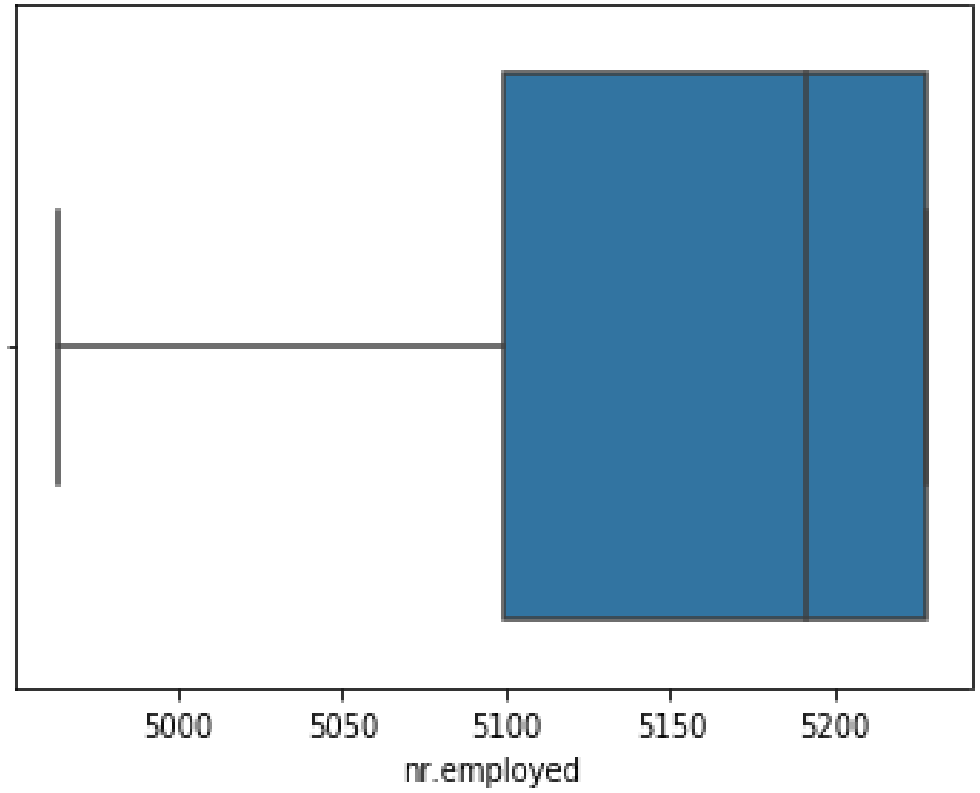

1
2
3
4
5
euribor3m

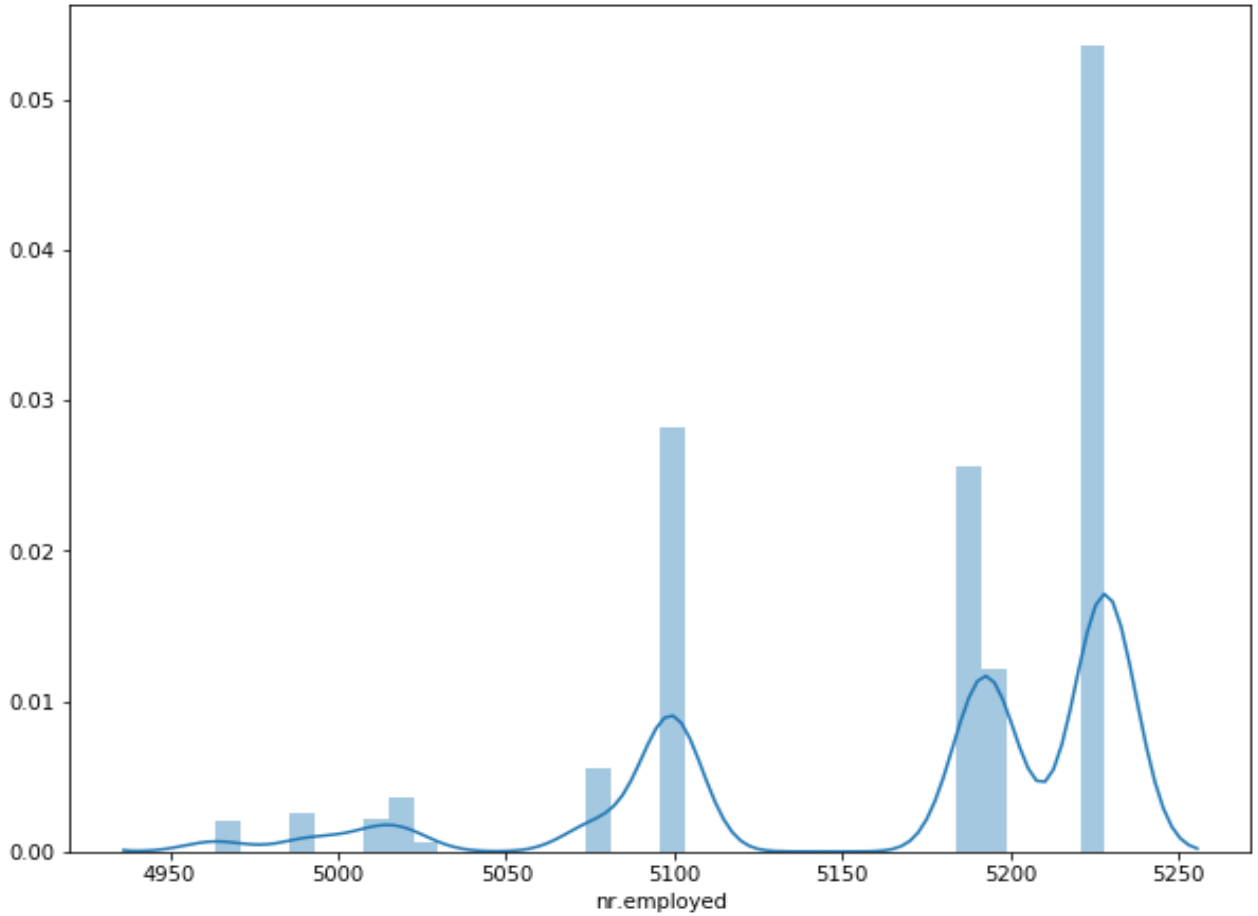

EURIBOR3M
y
no
yes
euribor3m
5
4
3
2
1
0
no
yes
y

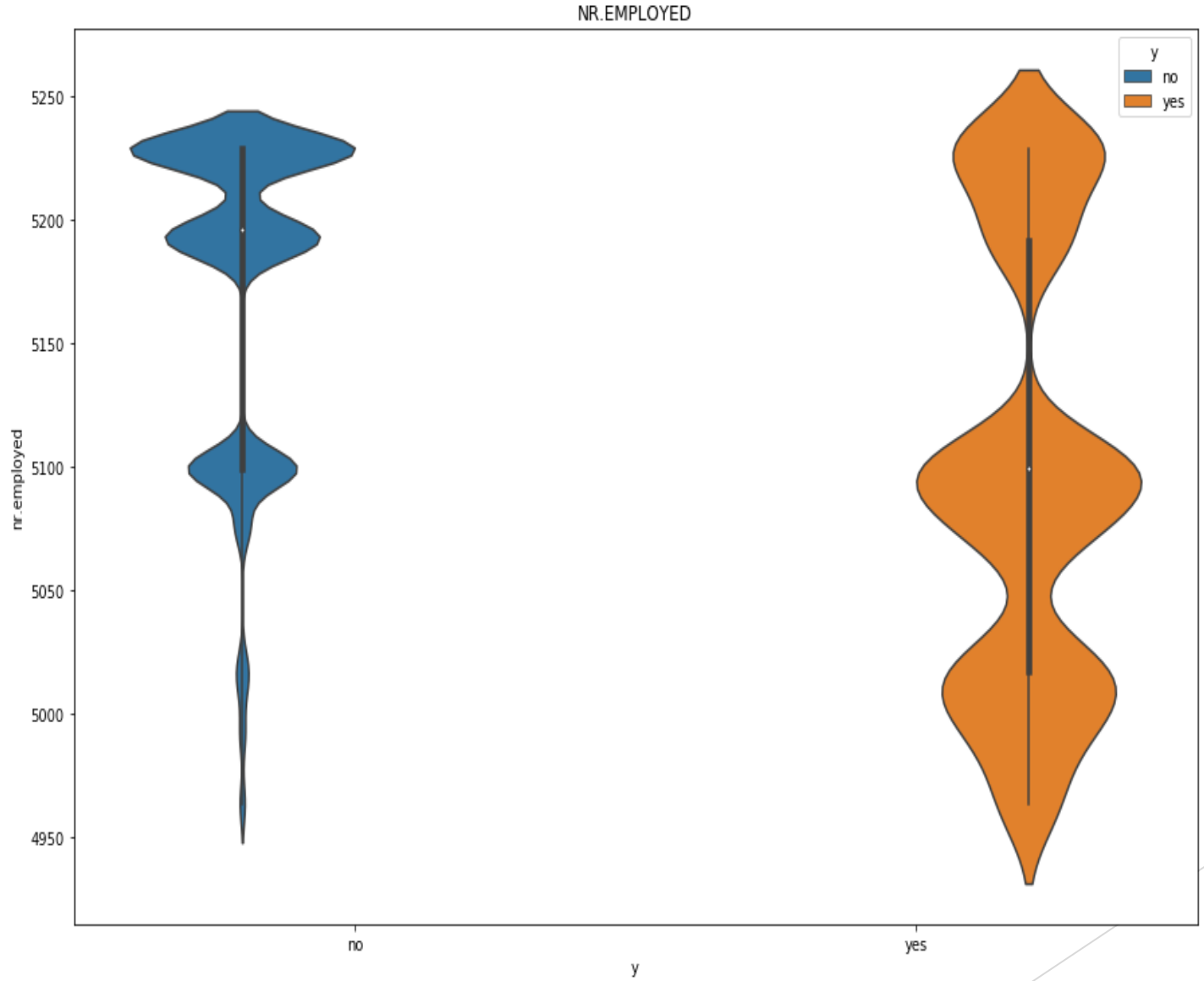
- From the above plots,Euribor3m does not contain outliers and would be very much helpful in predicting the class labels.

- **Feature:nr.employed(numeric)**



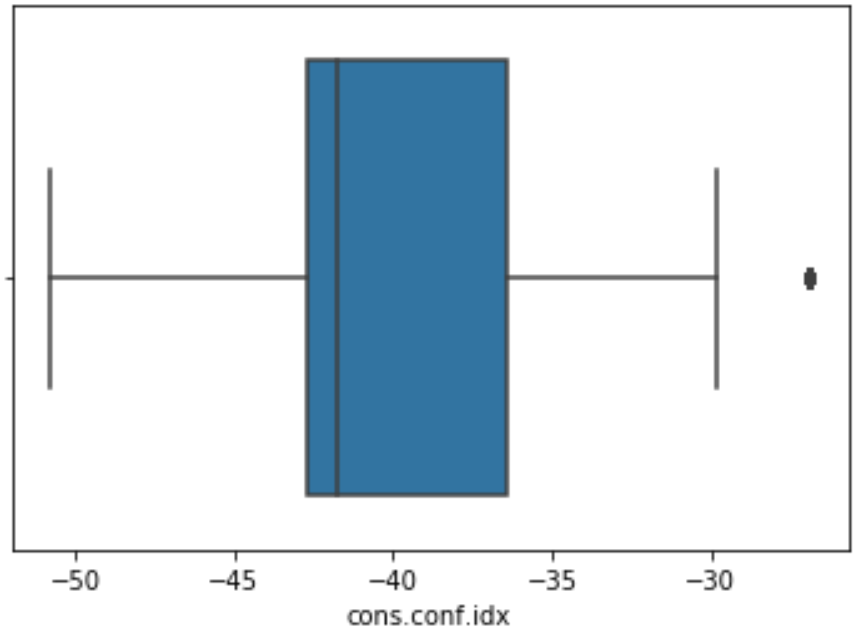

5000
5050
5100
5150
5200
nr.employed

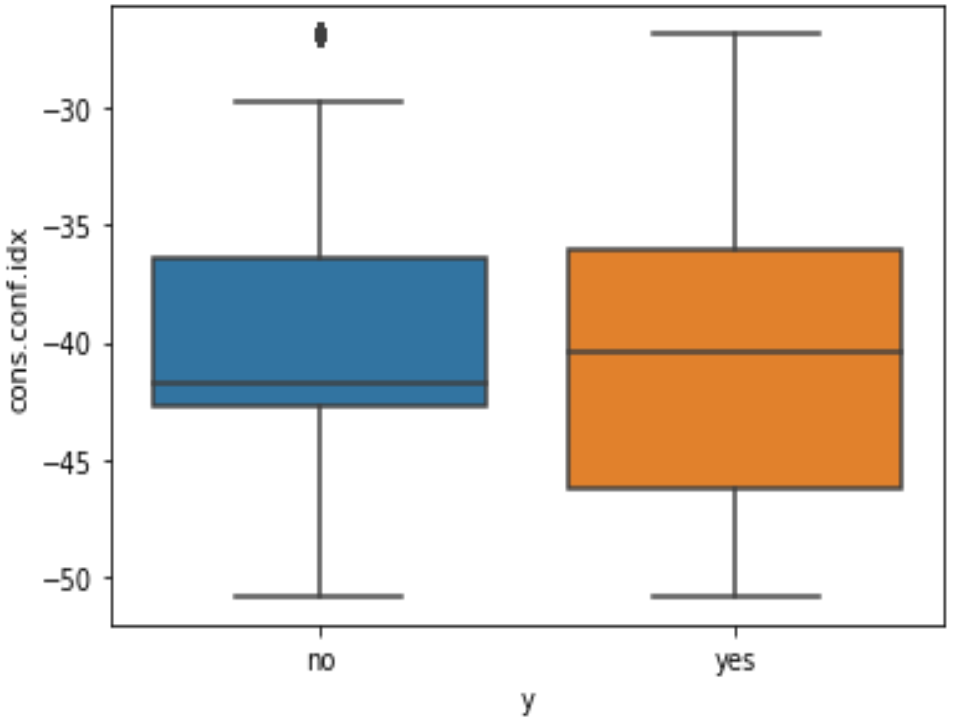

0.05
0.04
0.03
0.02
0.01
0.00
4950
5000
5050
5100
5150
5200
5250
nr.employed


NR.EMPLOYED
y
no
yes
5250
5200
5150
5100
5050
5000
4950
nr.employed
no
yes
y

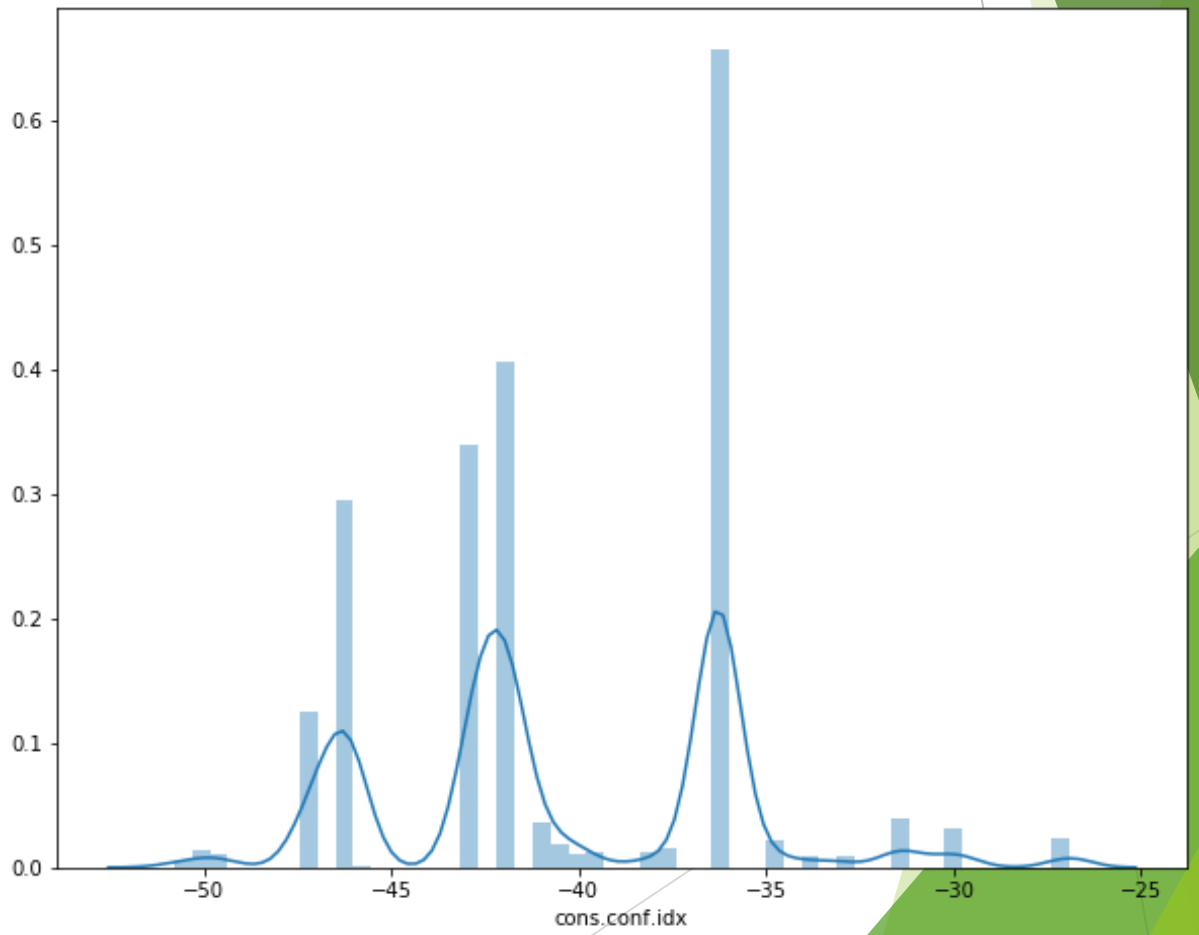
- From the plots above , nr. employed does not contain outliers and nr. employed would also be very much helpful in predicting class labels.

- **Feature:cons.conf.idx(numeric)**

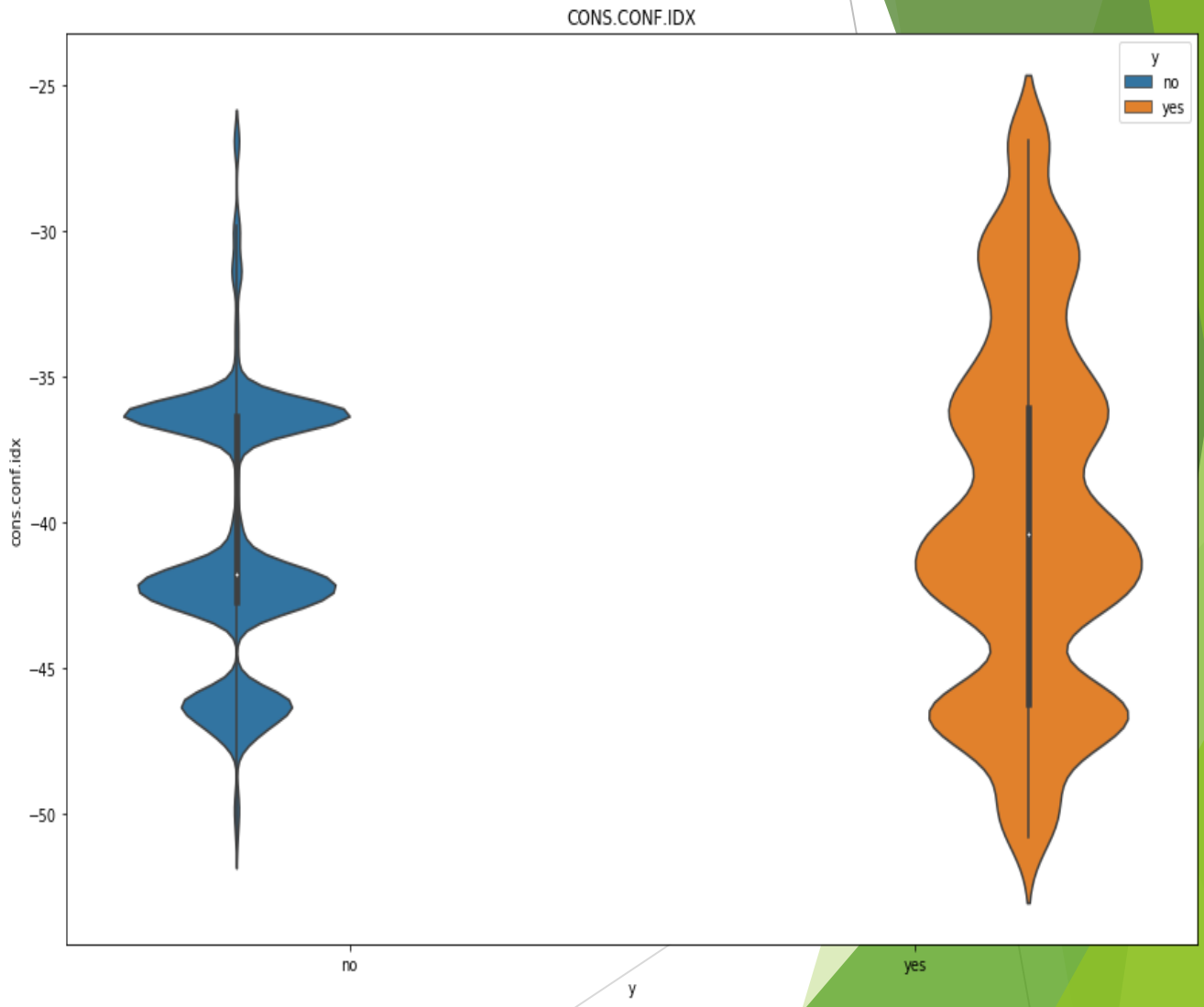



−30
−35
−40
−45
−50
cons.conf.idx
no
yes
y

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

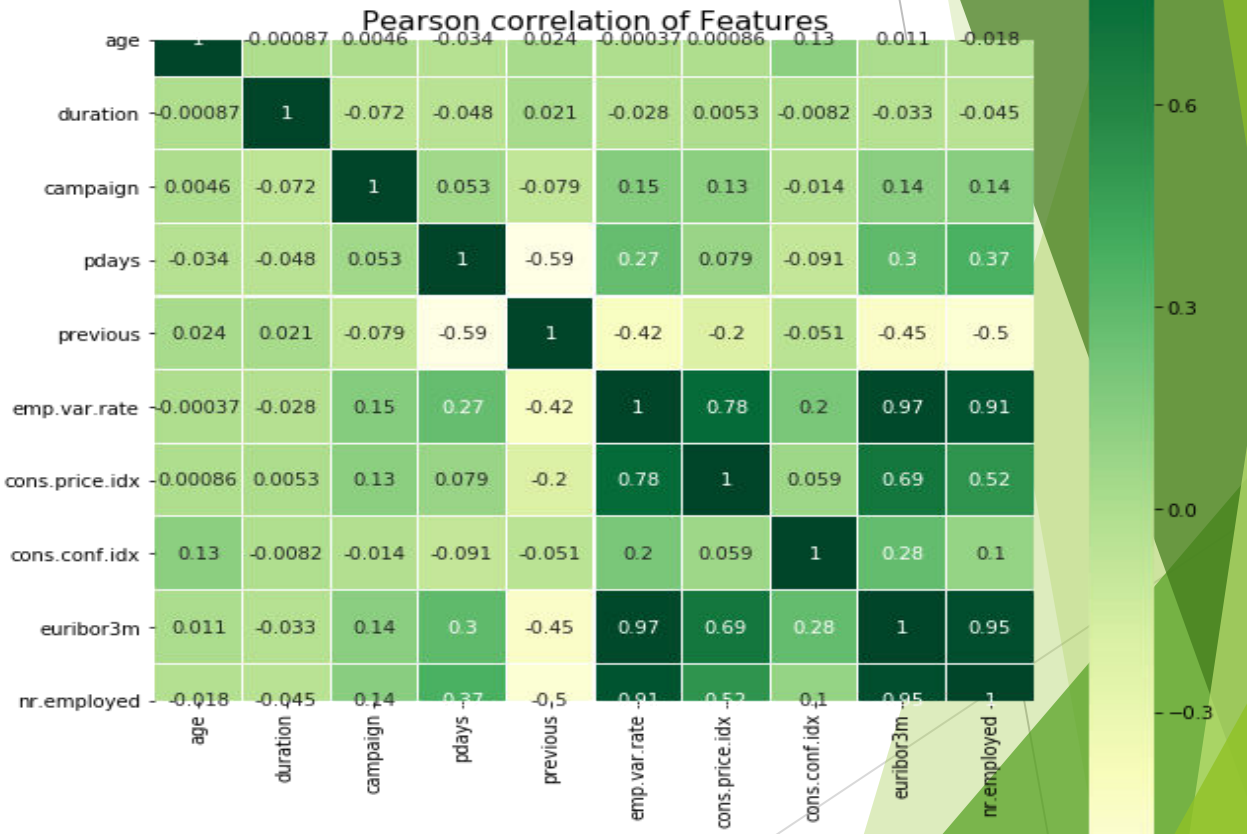


- In cons.conf.idx feature for class labels no, there is an outlier present when value above -30.



Correlation among features:

- Pearson correlation used.
- Let’s check for correlation of features between the numerical features.



- 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:

- 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.

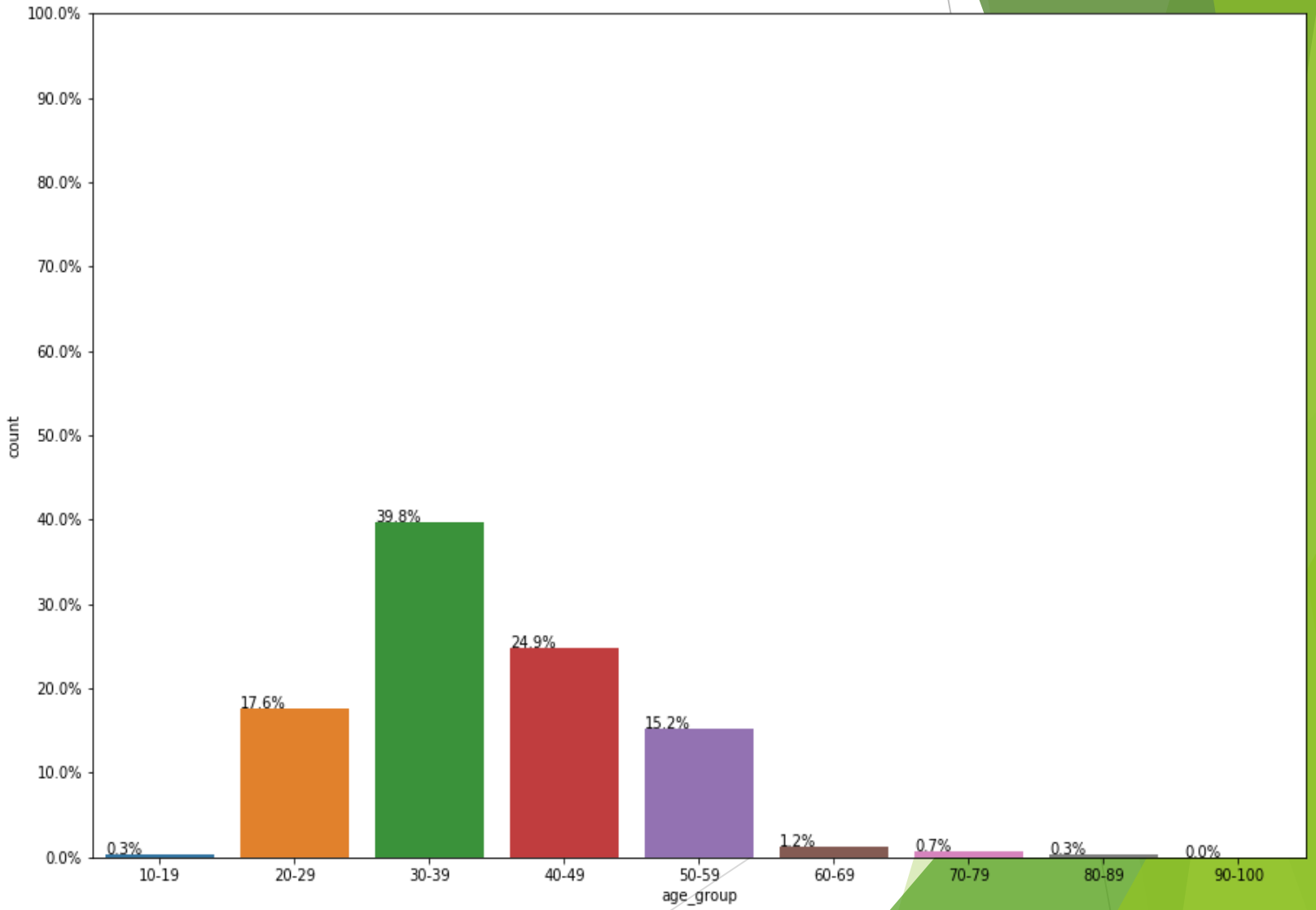
- Here I have created 9 groups from minimum age 10 to maximum age 100.
- After creating, inserted the age group into data frame and deleted the age column from the data frame.

➢ Creating i_loan column; Deleting Housing and Loan column

- I have created 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.

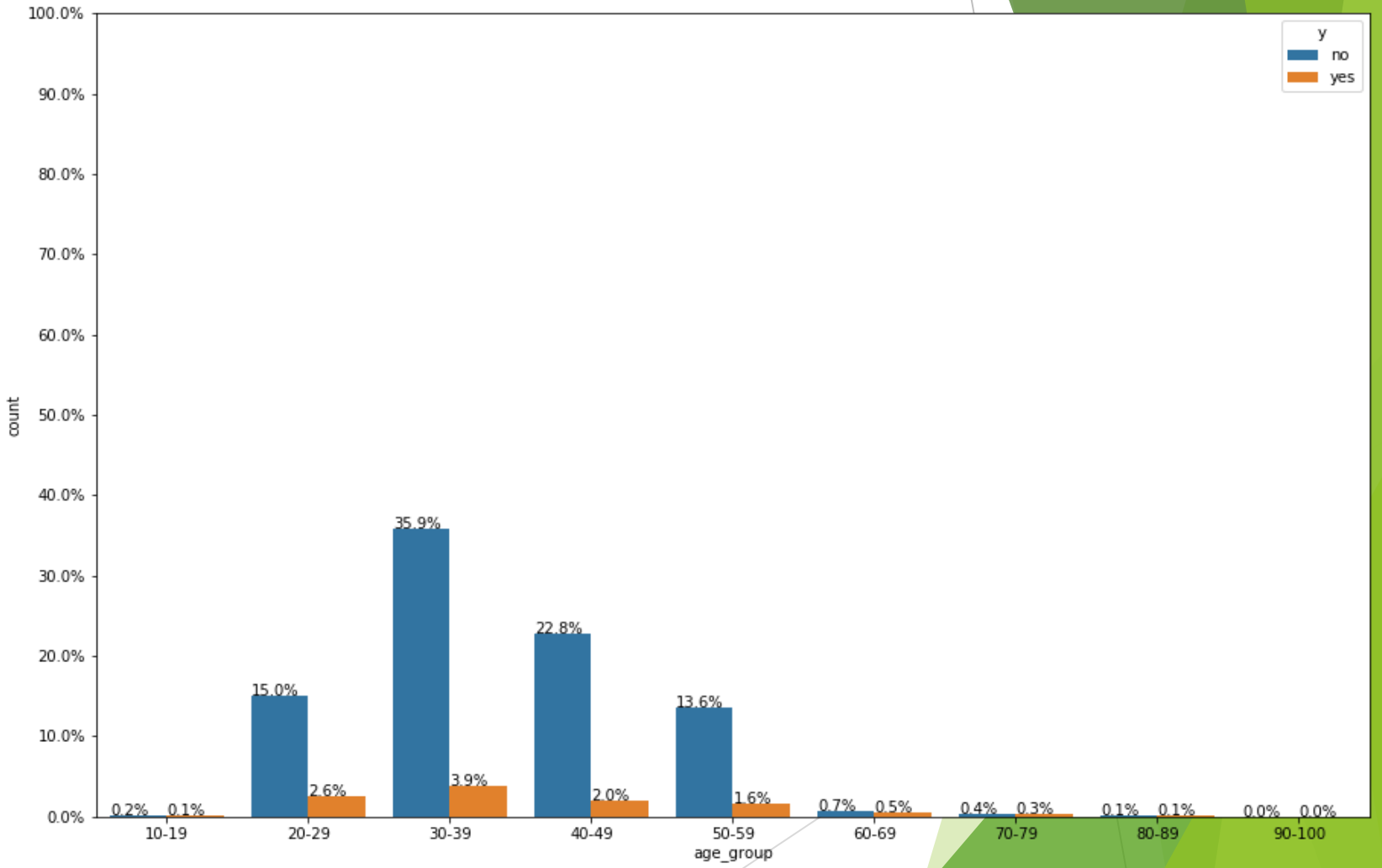
➢ Which age group has the bank contacted the most?

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



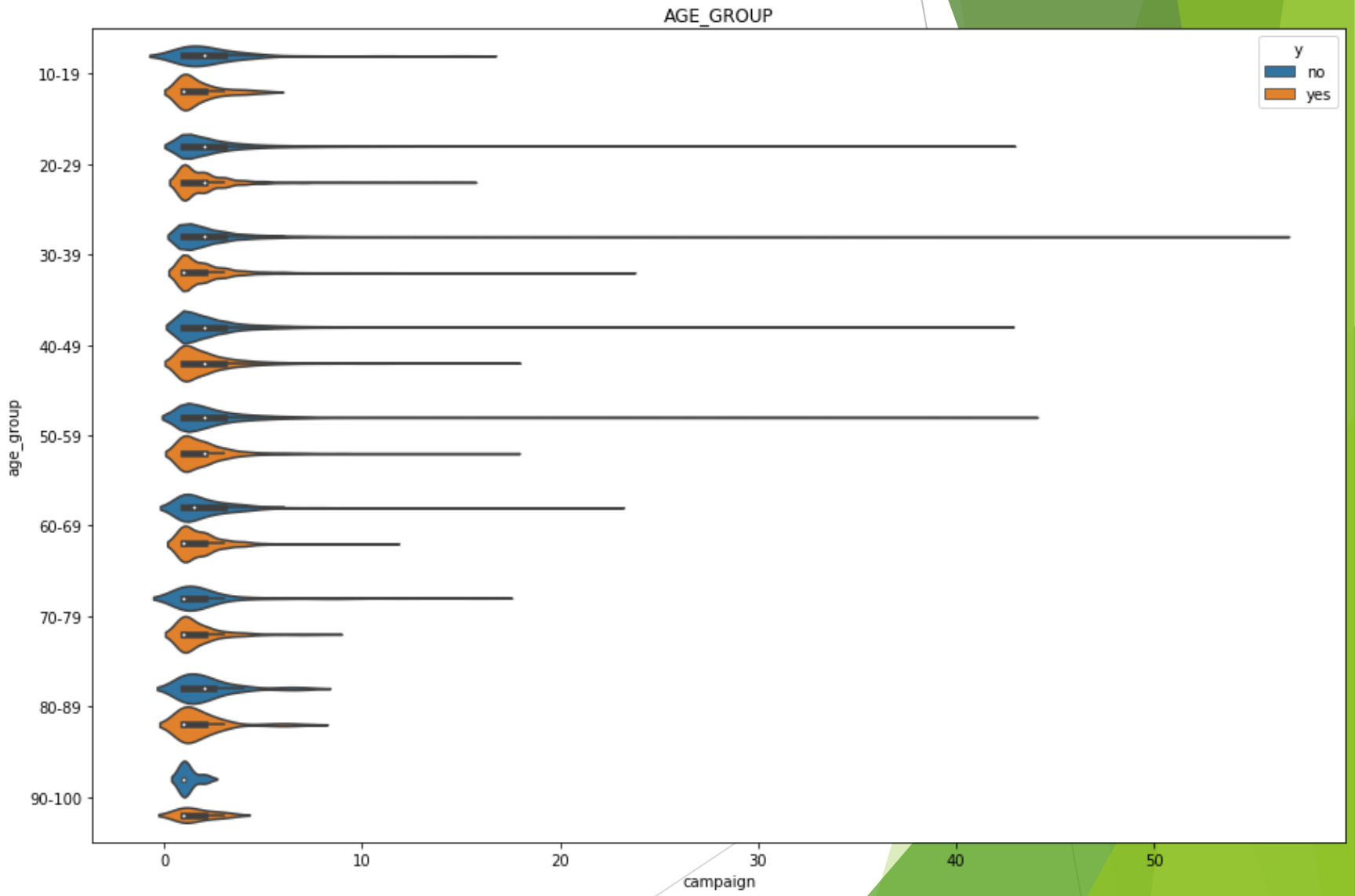
➢ Which age group is most likely to subscribe for long term deposits?

- 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.



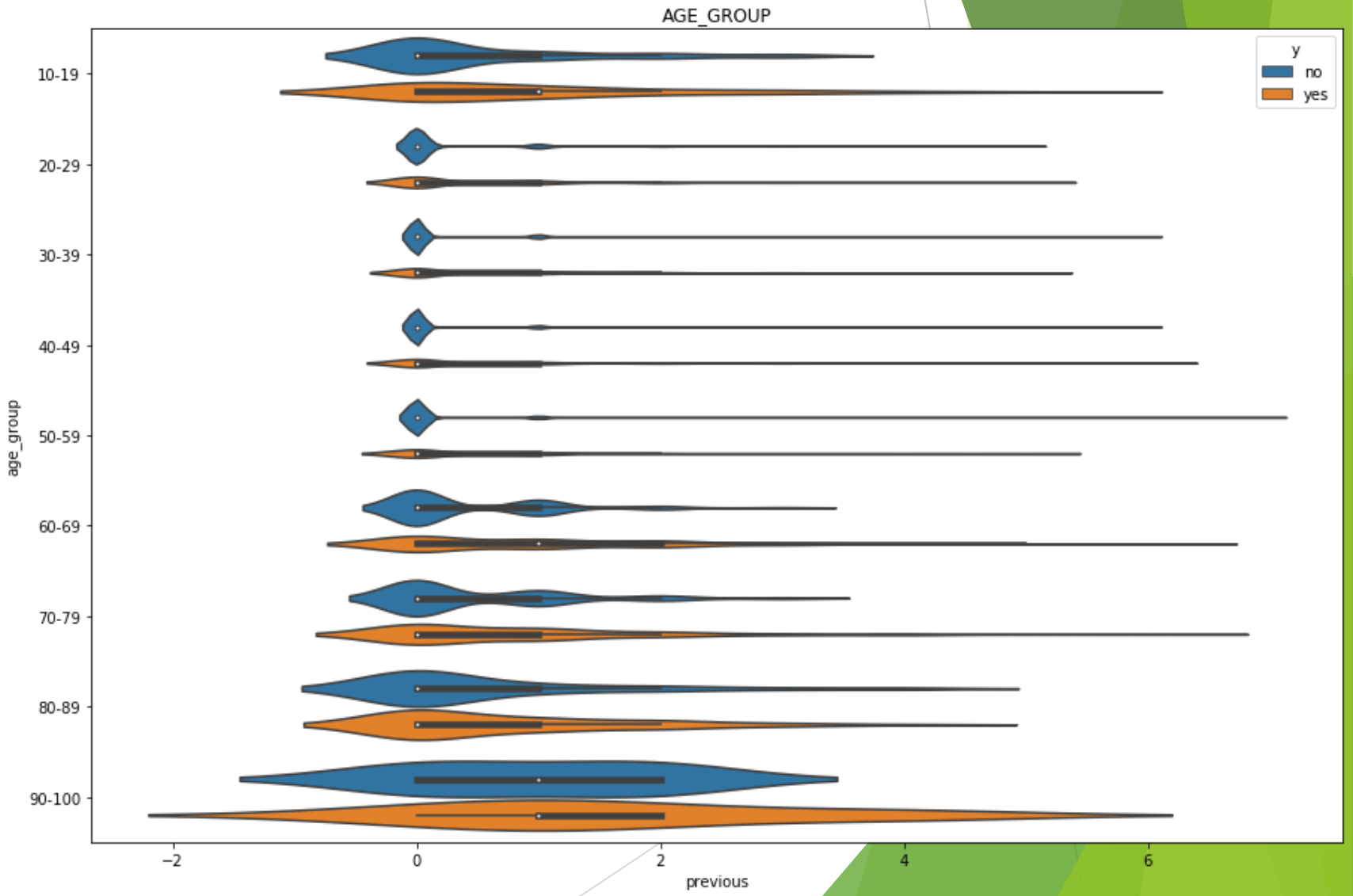
➢ Age_group and campaign

- Number of campaigns from 1-7 and age above 70 has possible outliers.



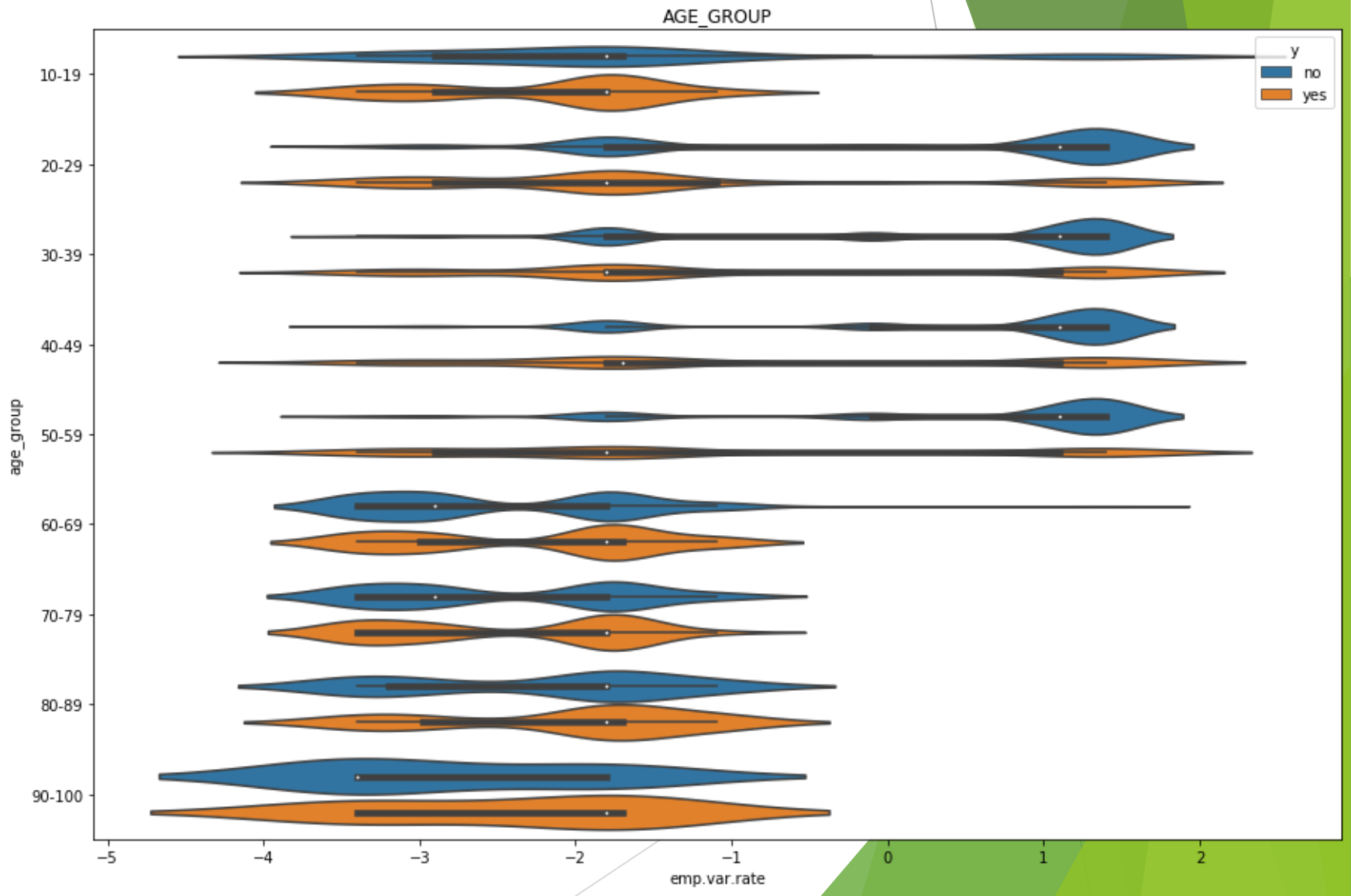
➢ Age_group and previous

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



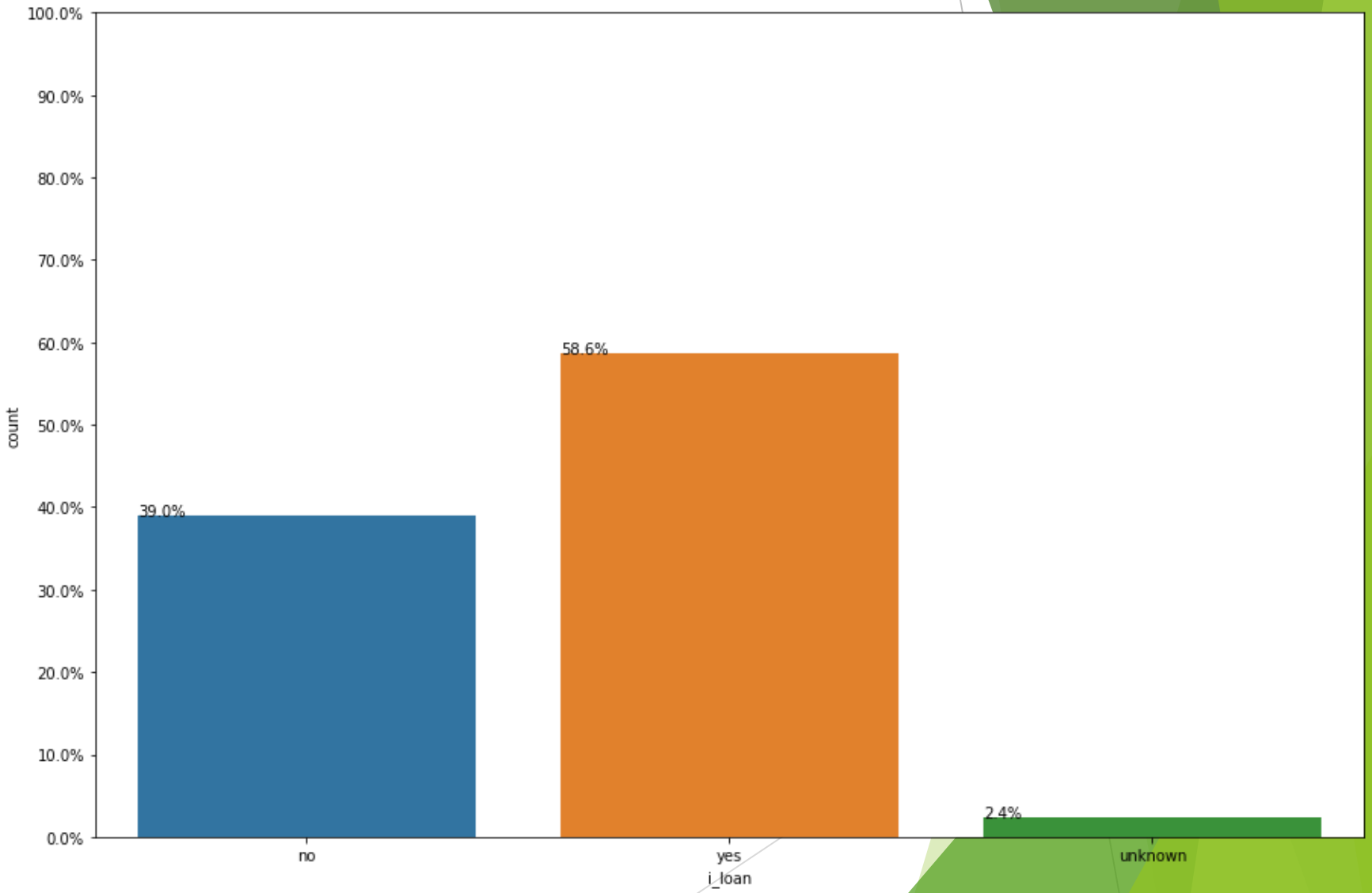
➢ Age_group and emp.var.rate

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



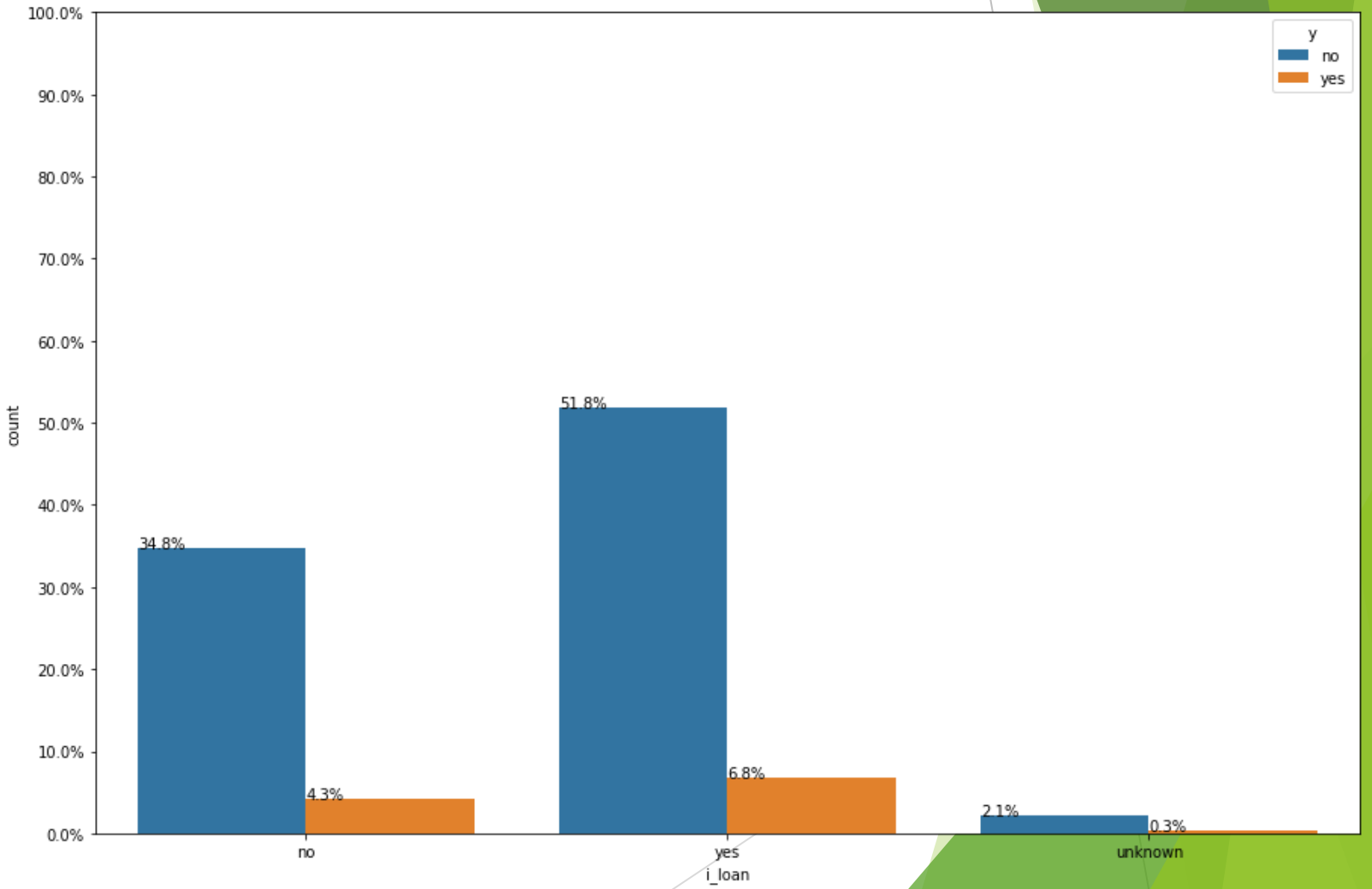
➢ i_loan

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



➢ Do people who have loans are the most that have subscribed for deposits?

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



- For the other EDA of the following:
- Age_group and job
- Age_group and education
- Age_group and i_loan
- Education and job
- Education and marital
- I_loan and education
- Marital and job
- I_loan and marital
- I_loan and job

you can check out and review the full code on the EDA ipynb file

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 has been contacted more on the cellular than the telephone.
- People has been contacted more in the month of May than any other month. They have not been contacted in January and February at all.
- People has not been contacted on Saturday and Sunday.
- People with no default status has been contacted more by the bank.
- People who has housing loan has been contacted more by the bank
- People with no personal loan has been contacted more by the bank.
- People who are in university has been contacted more by the bank.
- Age,Duration,Campaign have 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.

Thank You

Data Glacier
Your Deep Learning Partner