DATA SCIENCE A

DEPOSIT SUBSCRIPTION

EDA MINI PROJECT

TEAM 1

Our Team

1. Annisa Dwi Ari
2. Djeremiah Christofel
3. Farhan Imam Naufal
4. Mikhael Aditha S M
5. Salwa Ahla Amania

AGENDA

01 THE OBJECTIVES

02 DATA PREPARATION

03 ANALYSIS & FINDING

04 SUMMARY & CONCLUSION

THE OBJECTIVES

Bank-Full Exploratory Data Analysis

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

Features Description

Input variables :

bank client data :

1. age (numeric)
2. job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid", "entrepreneur", "student", "blue-collar","self-employed", "retired", "technician", "services")
3. marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4. education (categorical: "unknown","secondary","primary","tertiary")
5. default: has credit in default? (binary: "yes","no")
6. balance: average yearly balance, in euros (numeric)
7. housing: has housing loan? (binary: "yes","no")
8. loan: has personal loan? (binary: "yes","no")

Features Description

Irelated with the last contact of the current campaign:

1. contact: contact communication type (categorical: "unknown", "telephone", "cellular")
2. day: last contact day of the month (numeric)
3. month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
4. duration: last contact duration, in seconds (numeric)

Features Description

Other attributes :

1. campaign : number of contacts performed during this campaign and for this client (numeric, includes last contact)
2. pdays : number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
3. previous : number of contacts performed before this campaign and for this client (numeric)
4. poutcome : outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Output variable (desired target):

1. y - has the client subscribed a term deposit? (binary: "yes","no")

DATA PREPARATION

1.IMPORT LIBRARY

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn
from sklearn.preprocessing import LabelEncoder
import scipy.stats as ss
import warnings
warnings.filterwarnings("ignore")
```

2. Reading the dataset

In [2]:
```
df = pd.read_csv("bank-full.csv", sep=";")
```

3. Check the top 5 of data

In [3]:
```
df.head()
```

Out[3]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

4. Check the bottom 5 of data

In [4]:
```
df.tail()
```

Out[4]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
45206	51	technician	married	tertiary	no	825	no	no	cellular	17	nov	977	3	-1	0	unknown	yes
45207	71	retired	divorced	primary	no	1729	no	no	cellular	17	nov	456	2	-1	0	unknown	yes
45208	72	retired	married	secondary	no	5715	no	no	cellular	17	nov	1127	5	184	3	success	yes
45209	57	blue-collar	married	secondary	no	668	no	no	telephone	17	nov	508	4	-1	0	unknown	no
45210	37	entrepreneur	married	secondary	no	2971	no	no	cellular	17	nov	361	2	188	11	other	no

5. Check the sample of data

In [5]:
```
df.sample(3)
```

Out[5]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
40594	28	management	single	tertiary	no	5474	no	no	cellular	28	jul	486	5	-1	0	unknown	yes
7446	48	admin.	married	tertiary	no	960	yes	no	unknown	29	may	286	3	-1	0	unknown	no
14666	38	housemaid	married	secondary	no	0	no	no	cellular	15	jul	105	7	-1	0	unknown	no

6. Check the information of the data

```
In [6]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   age        45211 non-null  int64 
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64 
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64 
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64 
 12  campaign   45211 non-null  int64 
 13  pdays      45211 non-null  int64 
 14  previous   45211 non-null  int64 
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

7. Summary of each feature in the dataset

```
data = pd.DataFrame({'Columns' : df.columns, 'dataType' : df.dtypes.values,
                     'null' : [df[i].isna().sum() for i in df.columns],
                     'null percentage' : [((df[i].isna().sum()/len(df[i]))*100).round(1)
                                          for i in df.columns],
                     'Nunique' : [df[i].nunique() for i in df.columns],
                     'uniqueSample' : [list(pd.Series(df[i].unique()).sample(2))
                                        for i in df.columns]}).reset_index(drop = True)
data.sort_values(by = "null", ascending=False)
```

	Columns	dataType	null	null percentage	Nunique	uniqueSample
0	age	int64	0	0.0	77	[79, 19]
9	day	int64	0	0.0	31	[25, 27]
15	poutcome	object	0	0.0	4	[unknown, other]
14	previous	int64	0	0.0	41	[30, 40]
13	pdays	int64	0	0.0	559	[558, 315]
12	campaign	int64	0	0.0	48	[30, 31]
11	duration	int64	0	0.0	1573	[565, 917]
10	month	object	0	0.0	12	[feb, jan]
8	contact	object	0	0.0	3	[cellular, unknown]
1	job	object	0	0.0	12	[housemaid, self-employed]
7	loan	object	0	0.0	2	[no, yes]
6	housing	object	0	0.0	2	[yes, no]
5	balance	int64	0	0.0	7168	[3202, -216]
4	default	object	0	0.0	2	[no, yes]
3	education	object	0	0.0	4	[primary, secondary]
2	marital	object	0	0.0	3	[single, married]
16	y	object	0	0.0	2	[yes, no]

8. Summary of integer features

In [8]:

```
df.describe().T
```

Out[8]:

	count	mean	std	min	25%	50%	75%	max
age	45211.0	40.936210	10.618762	18.0	33.0	39.0	48.0	95.0
balance	45211.0	1362.272058	3044.765829	-8019.0	72.0	448.0	1428.0	102127.0
day	45211.0	15.806419	8.322476	1.0	8.0	16.0	21.0	31.0
duration	45211.0	258.163080	257.527812	0.0	103.0	180.0	319.0	4918.0
campaign	45211.0	2.763841	3.098021	1.0	1.0	2.0	3.0	63.0
pdays	45211.0	40.197828	100.128746	-1.0	-1.0	-1.0	-1.0	871.0
previous	45211.0	0.580323	2.303441	0.0	0.0	0.0	0.0	275.0

9. Summary of object features

```
In [9]: df.describe(include = "object").T
```

Out[9]:

	count	unique	top	freq
job	45211	12	blue-collar	9732
marital	45211	3	married	27214
education	45211	4	secondary	23202
default	45211	2	no	44396
housing	45211	2	yes	25130
loan	45211	2	no	37967
contact	45211	3	cellular	29285
month	45211	12	may	13766
poutcome	45211	4	unknown	36959
y	45211	2	no	39922

10. Checking and adding up the missing values

```
In [10]: df.isnull().sum().sort_values(ascending=False)

Out[10]: age          0
         day          0
         poutcome     0
         previous     0
         pdays        0
         campaign     0
         duration     0
         month        0
         contact      0
         job          0
         loan         0
         housing      0
         balance      0
         default      0
         education    0
         marital      0
         y            0
         dtype: int64
```

11. Label encoding

```
label_encoder = LabelEncoder()
df["job_encode"] = label_encoder.fit_transform(df["job"])
df["marital_encode"] = label_encoder.fit_transform(df["marital"])
df["housing_encode"] = label_encoder.fit_transform(df["housing"])
df["default_encode"] = label_encoder.fit_transform(df["default"])
df["loan_encode"] = label_encoder.fit_transform(df["loan"])
df["poutcome_encode"] = label_encoder.fit_transform(df["poutcome"])
df["education_encode"] = label_encoder.fit_transform(df["education"])
df["y_encode"] = label_encoder.fit_transform(df["y"])

categorical_column = ["job_encode", "marital_encode",
                      "housing_encode", "default_encode",
                      "loan_encode", "poutcome_encode",
                      "education_encode", "y_encode"]
```

12. Checking the duplicate data

In [13]:

```
#Mencari data yang duplicate
df.duplicated().sum()
```

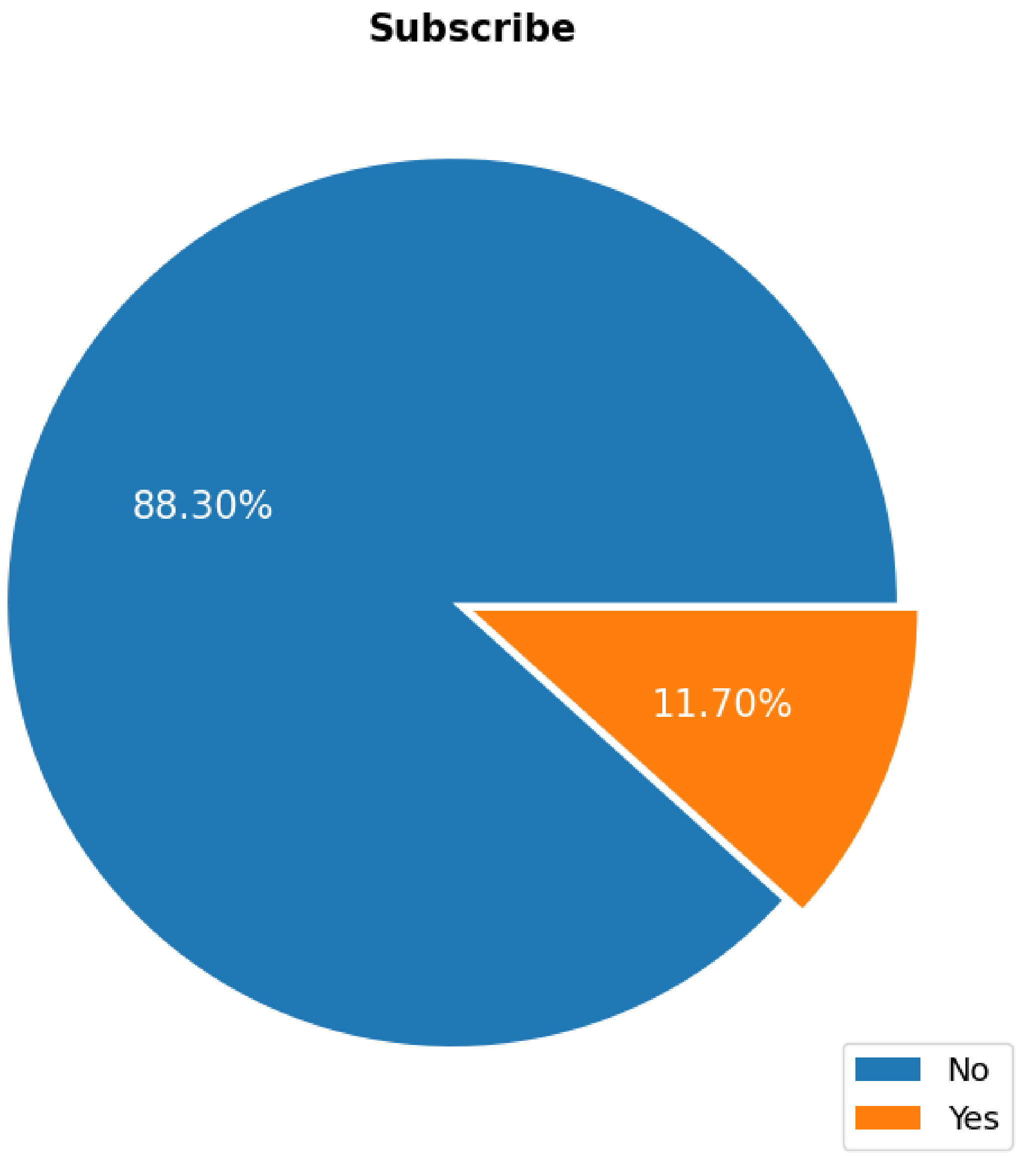
Out[13]: 0

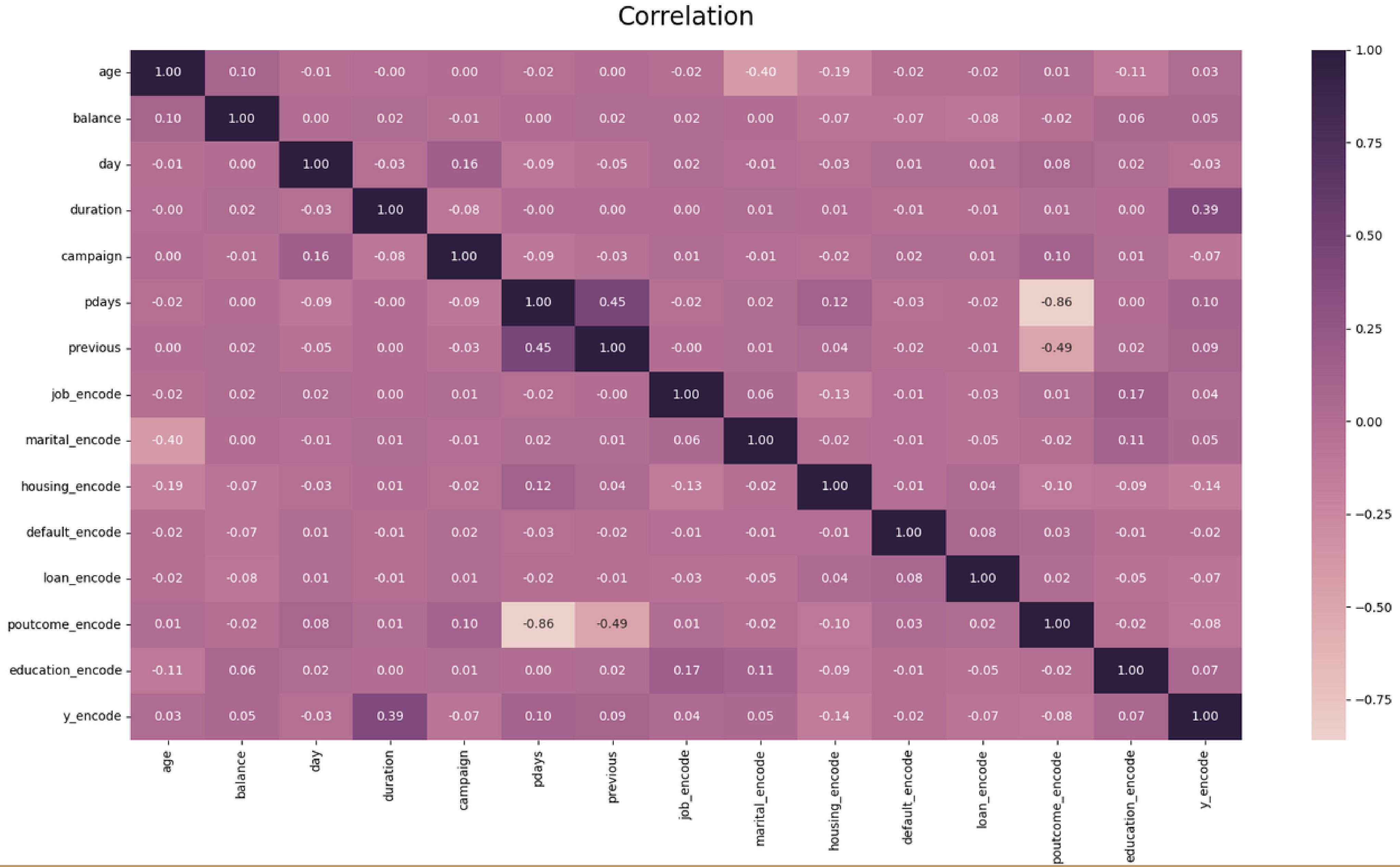
13. Check the information of the data after update

In [14]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 25 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   age               45211 non-null  int64 
 1   job               45211 non-null  object
 2   marital           45211 non-null  object
 3   education         45211 non-null  object
 4   default           45211 non-null  object
 5   balance           45211 non-null  int64 
 6   housing           45211 non-null  object
 7   loan              45211 non-null  object
 8   contact           45211 non-null  object
 9   day               45211 non-null  int64 
 10  month             45211 non-null  object
 11  duration          45211 non-null  int64 
 12  campaign          45211 non-null  int64 
 13  pdays             45211 non-null  int64 
 14  previous          45211 non-null  int64 
 15  poutcome          45211 non-null  object
 16  y                 45211 non-null  object
 17  job_encode        45211 non-null  int32 
 18  marital_encode    45211 non-null  int32 
 19  housing_encode    45211 non-null  int32 
 20  default_encode    45211 non-null  int32 
 21  loan_encode       45211 non-null  int32 
 22  poutcome_encode   45211 non-null  int32 
 23  education_encode  45211 non-null  int32 
 24  y_encode          45211 non-null  int32 
dtypes: int32(8), int64(7), object(10)
memory usage: 7.2+ MB
```

ANALYSIS AND FINDING

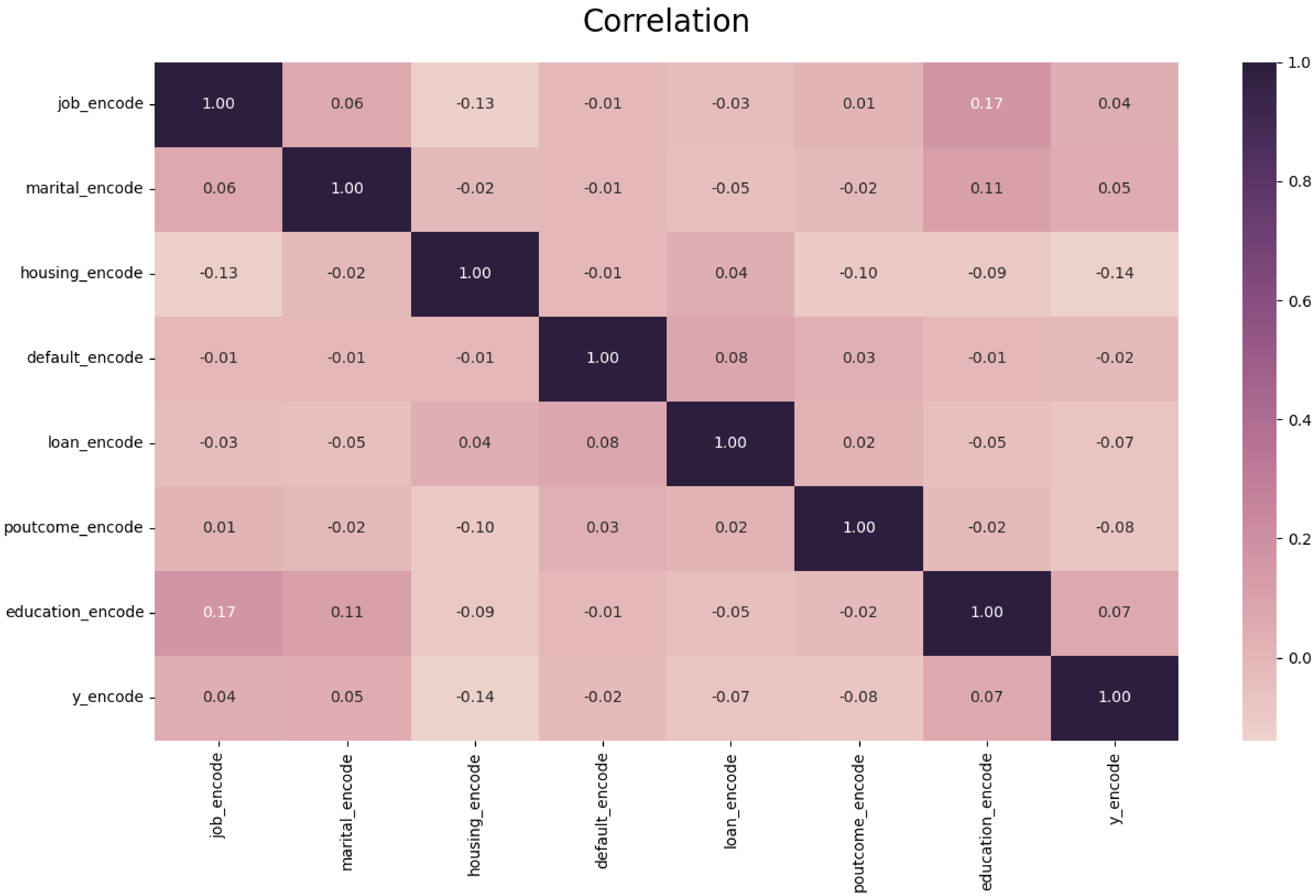




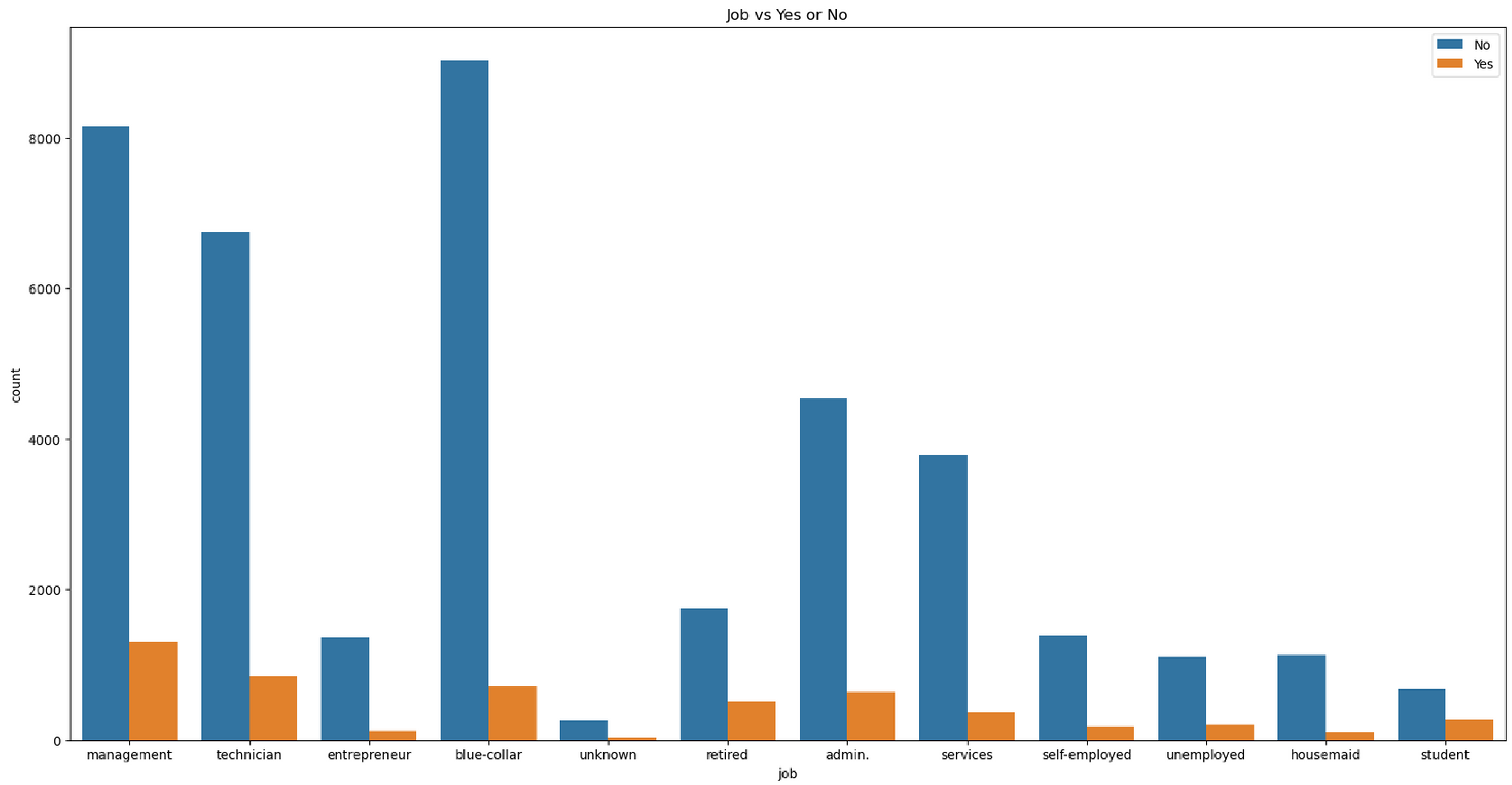
From the output obtained, there are 5289 (11.70%) who are subscribed, while those who are not subscribed are 39922 (88.3%)

There are not many large correlation values in the target column with the features column, the biggest correlation with the target column is the duration column. However, there are other columns that have a high correlation, such as poutcome_encode with pdays and previous with potcome_encode.

a. Categorical Data Analysis



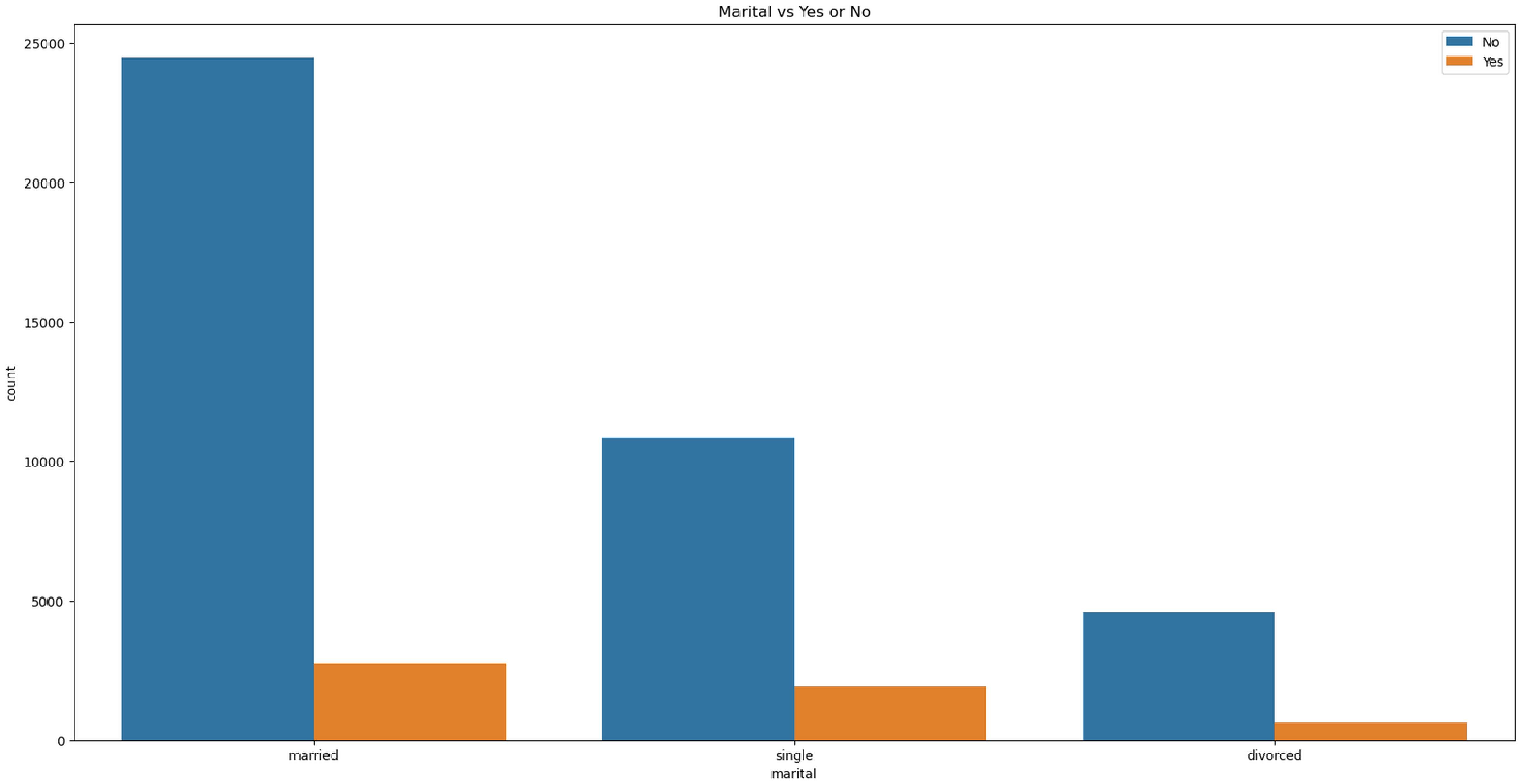
1. JOB



y / job	no	yes	yes percentage	no percentage	percentage of job
management	8157	1301	0.137556	0.862444	20.92
technician	6757	840	0.110570	0.889430	16.80
blue-collar	9024	708	0.072750	0.927250	21.53
admin.	4540	631	0.122027	0.877973	11.44
retired	1748	516	0.227915	0.772085	5.01
services	3785	369	0.088830	0.911170	9.19
student	669	269	0.286780	0.713220	2.07
unemployed	1101	202	0.155027	0.844973	2.88
self-employed	1392	187	0.118429	0.881571	3.49
entrepreneur	1364	123	0.082717	0.917283	3.29
housemaid	1131	109	0.087903	0.912097	2.74
unknown	254	34	0.118056	0.881944	0.64

Based on the graphs and tables, the job with the most subscriptions to deposits is management with a total of 1301. This is reasonable because management work is the most common job in the data, which is 20.92%.
Meanwhile, the largest percentage based on work is retired and student, respectively 28.7% and 22.7%.

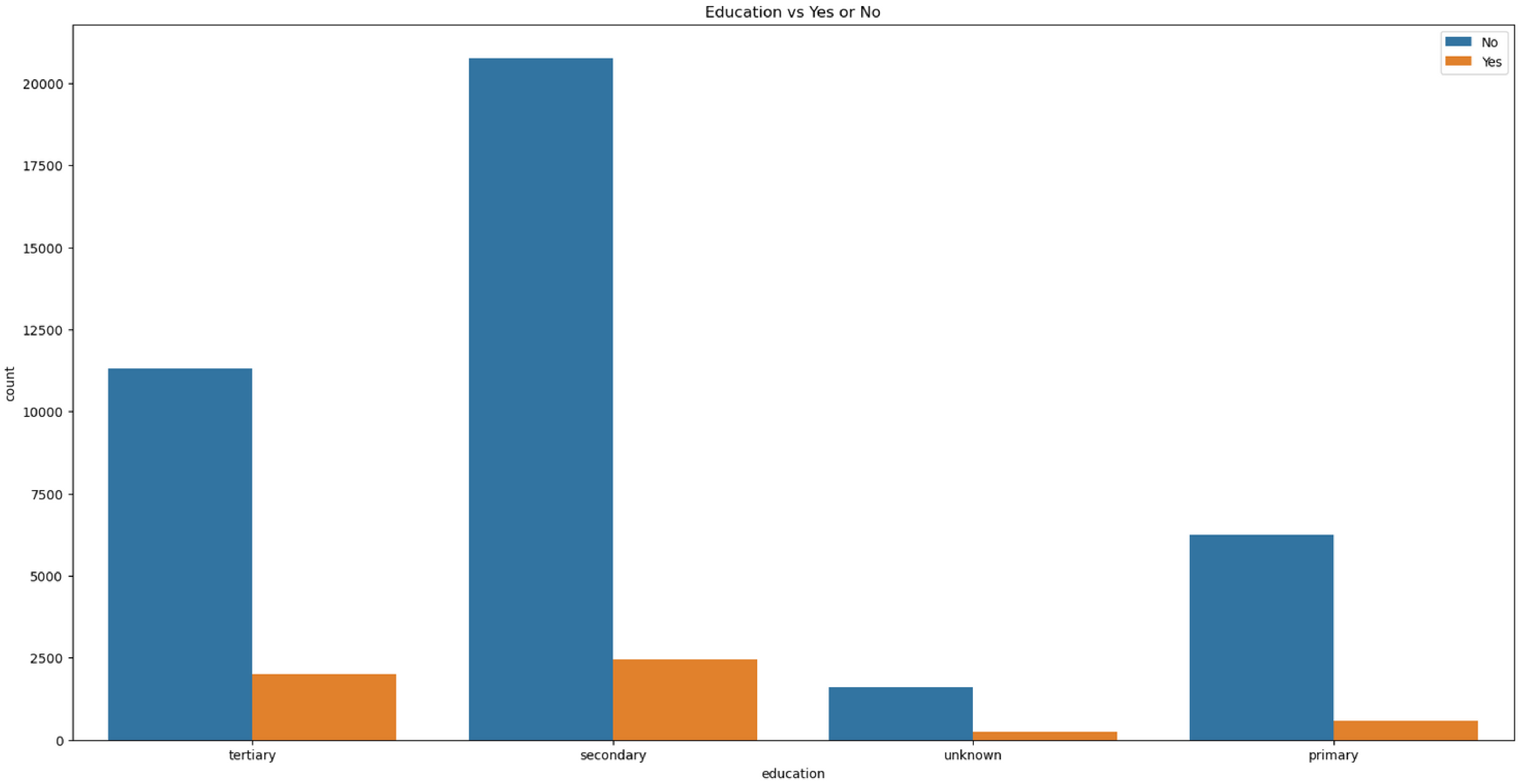
2. MARITAL



y	no	yes	yes percentage	no percentage	percentage of marital
marital					
married	24459	2755	0.101235	0.898765	60.19
single	10878	1912	0.149492	0.850508	28.29
divorced	4585	622	0.119455	0.880545	11.52

Based on the graphs and tables, the job with the most subscriptions to deposits is management with a total of 1301. This is reasonable because management work is the most common job in the data, which is 20.92%.
Meanwhile, the largest percentage based on work is retired and student, respectively 28.7% and 22.7%.

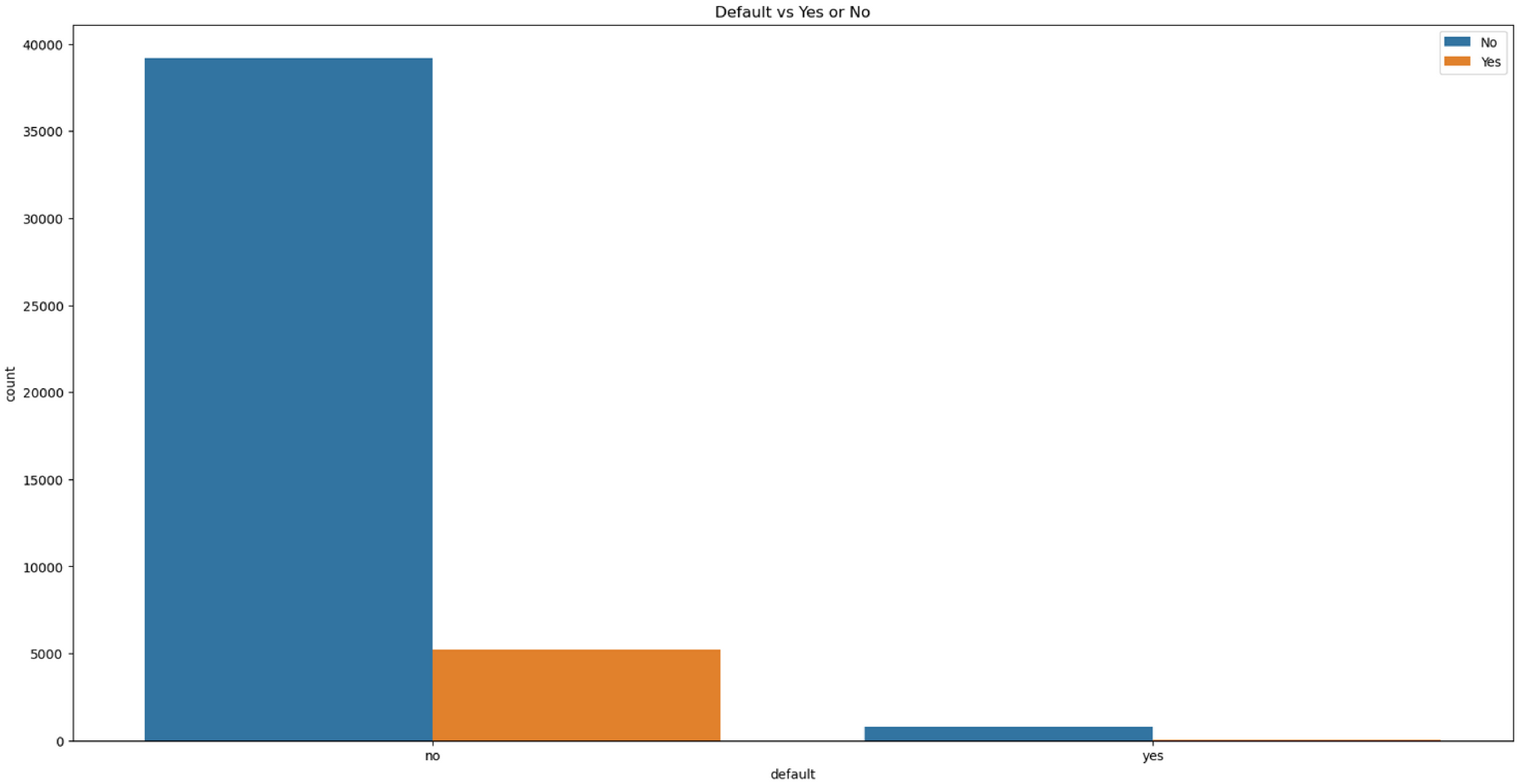
3. EDUCATION



y education	no	yes	yes percentage	no percentage	percentage of education
secondary	20752	2450	0.105594	0.894406	51.32
tertiary	11305	1996	0.150064	0.849936	29.42
primary	6260	591	0.086265	0.913735	15.15
unknown	1605	252	0.135703	0.864297	4.11

The education with the highest yes percentage is tertiary, which is 15%, but the difference in yes percentage between other levels of education is not too far so that education does not significantly affect subscriptions.

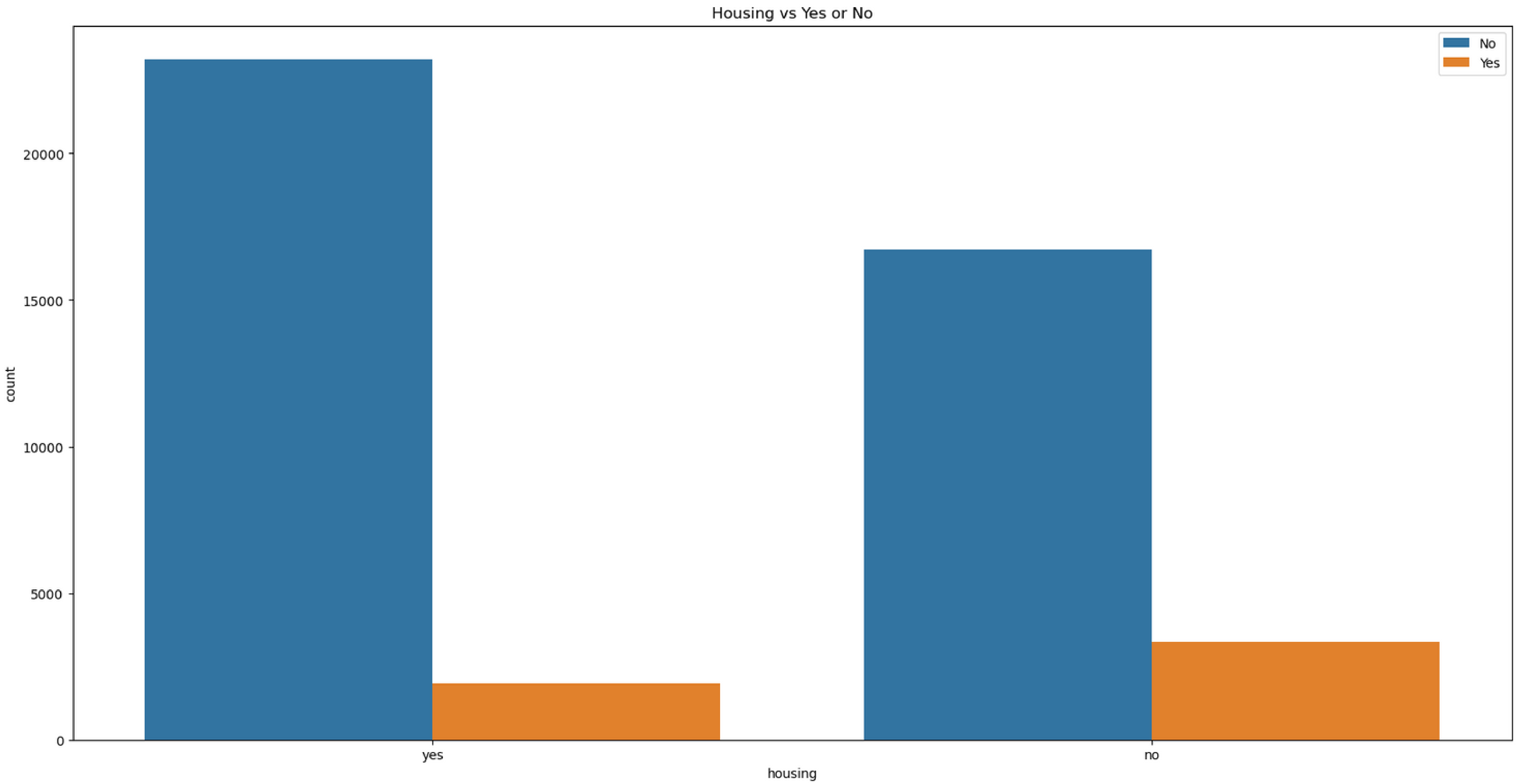
4. DEFAULT



y	no	yes	yes percentage	no percentage	percentage of default
default					
no	39159	5237	0.117961	0.882039	98.2
yes	763	52	0.063804	0.936196	1.8

Most of the marketing done goes to people whose credit defaults are not, and after the campaign there are 11% who subscribe to deposits.

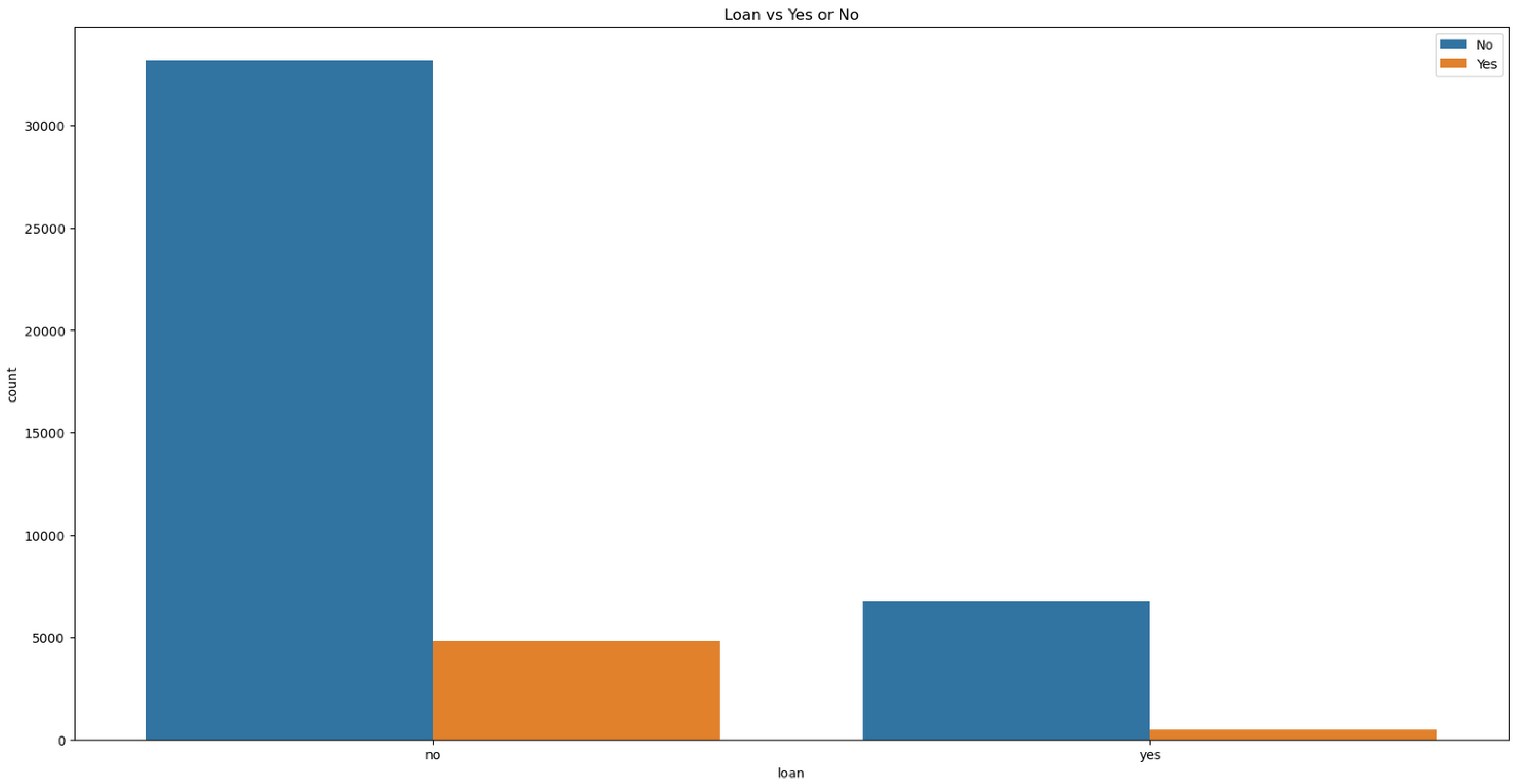
5. HOUSING



y	no	yes	yes percentage	no percentage	percentage of housing
housing					
no	16727	3354	0.167024	0.832976	44.42
yes	23195	1935	0.077000	0.923000	55.58

Based on the graphs and tables, it is quite obvious that there is a significant effect of the presence or absence of a home loan. Of people who do not have a home loan, there are 16% who subscribe to deposits, while those who have a home loan are only 7.7%. Although the percentage who subscribe to deposits is not too large, it is quite significant when compared to people who have loans for houses. Therefore, the presence or absence of a home loan is one of the factors that can cause someone to decide to subscribe to a deposit or not.

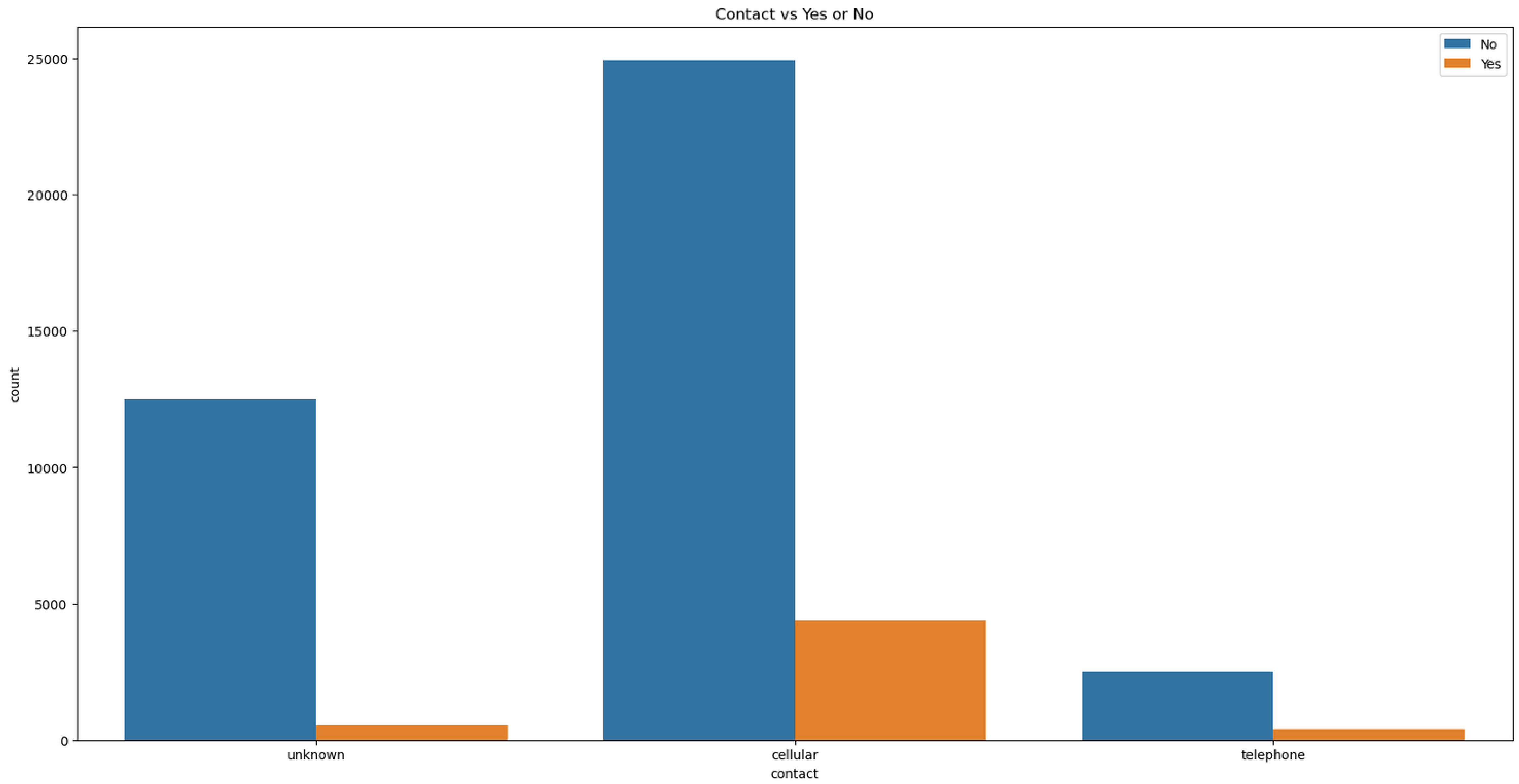
6. LOAN



y	no	yes	yes percentage	no percentage	percentage of loan
loan					
no	33162	4805	0.126557	0.873443	83.98
yes	6760	484	0.066814	0.933186	16.02

Based on the graphs and tables, the proportion obtained is higher for subscribing when someone does not have a loan, which is 12.6%. While someone who has a loan is only 6.7%. This can be achieved because in general someone who does not have a loan will have more money to save, people who have a loan do not because they are lending to other people.

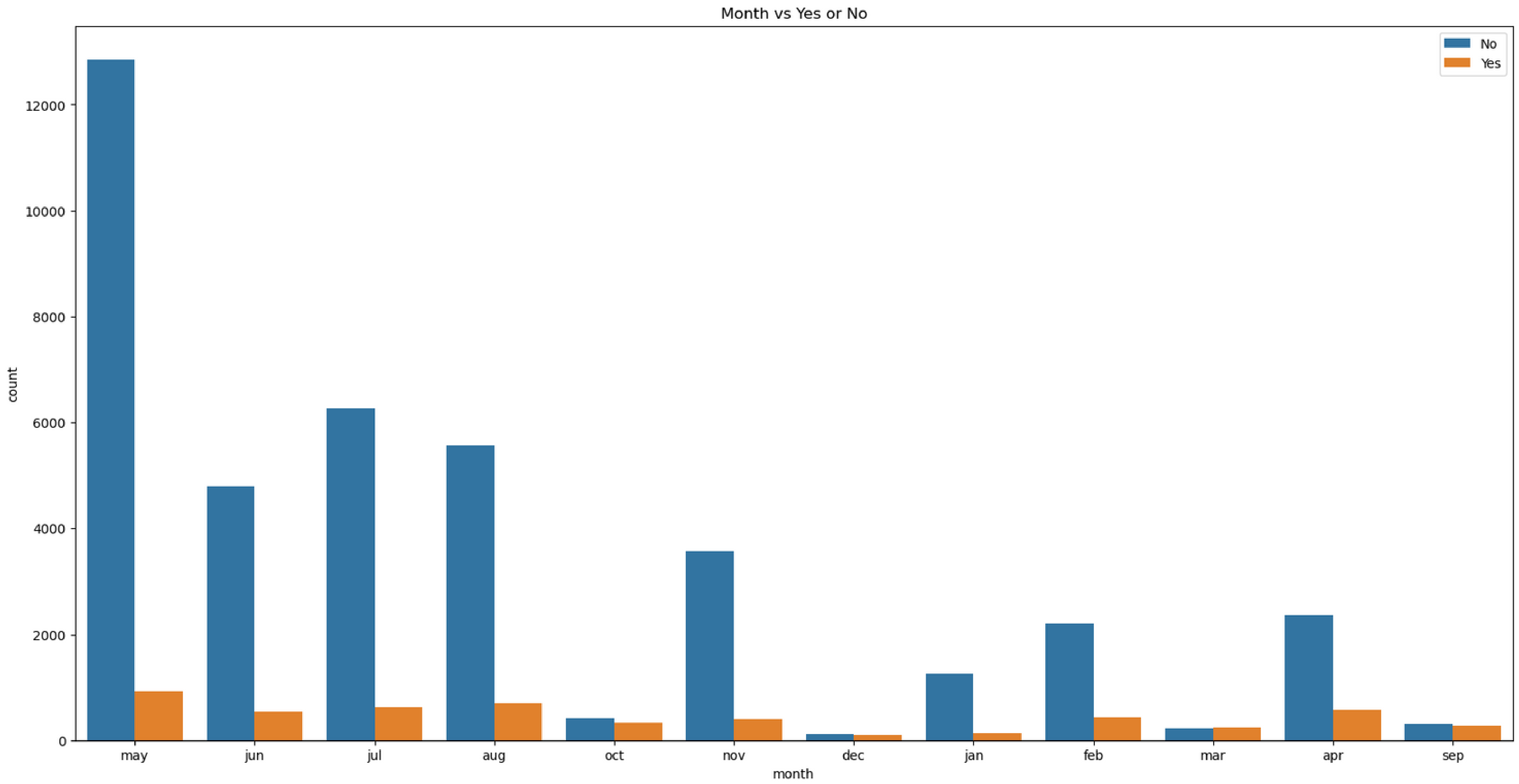
7. CONTACTS



y	no	yes	yes percentage	no percentage	percentage of contact
contact					
cellular	24916	4369	0.149189	0.850811	64.77
unknown	12490	530	0.040707	0.959293	28.80
telephone	2516	390	0.134205	0.865795	6.43

The way of campaigning through cellular and telephone has almost the same percentage yes. While the unknown has a very small yes percentage.

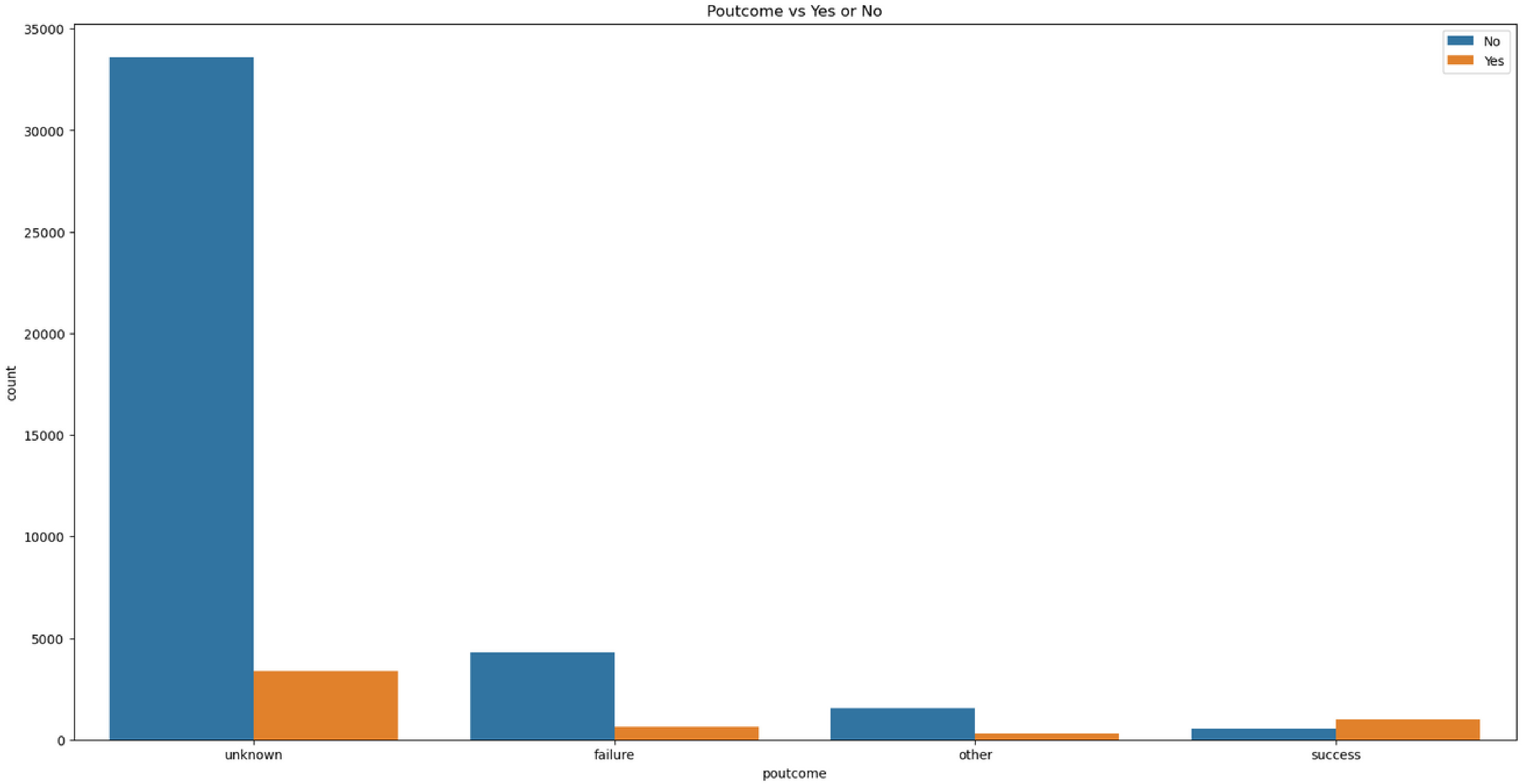
8. MONTH



month	no	yes	yes percentage	no percentage	percentage of month
may	12841	925	0.067195	0.932805	30.45
aug	5559	688	0.110133	0.889867	13.82
jul	6268	627	0.090935	0.909065	15.25
apr	2355	577	0.196794	0.803206	6.49
jun	4795	546	0.102228	0.897772	11.81
feb	2208	441	0.166478	0.833522	5.86
nov	3567	403	0.101511	0.898489	8.78
oct	415	323	0.437669	0.562331	1.63
sep	310	269	0.464594	0.535406	1.28
mar	229	248	0.519916	0.480084	1.06
jan	1261	142	0.101212	0.898788	3.10
dec	114	100	0.467290	0.532710	0.47

Most recent contacts in a year were in May. This is directly proportional to the number of people who subscribe to deposits, which is 925 (6%). However, the month with the highest percentage yes was in April, which was 19% with a total of 577.

9. POUTCOME

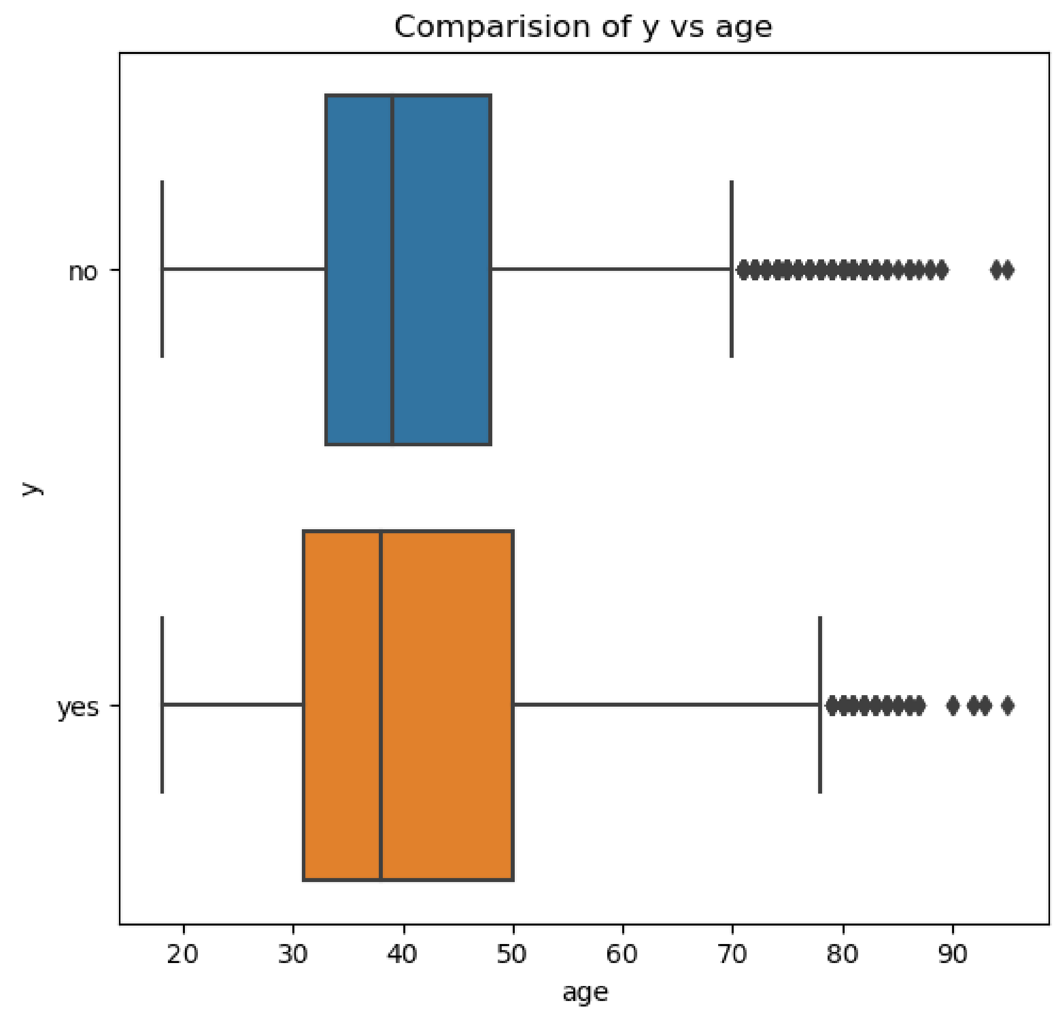


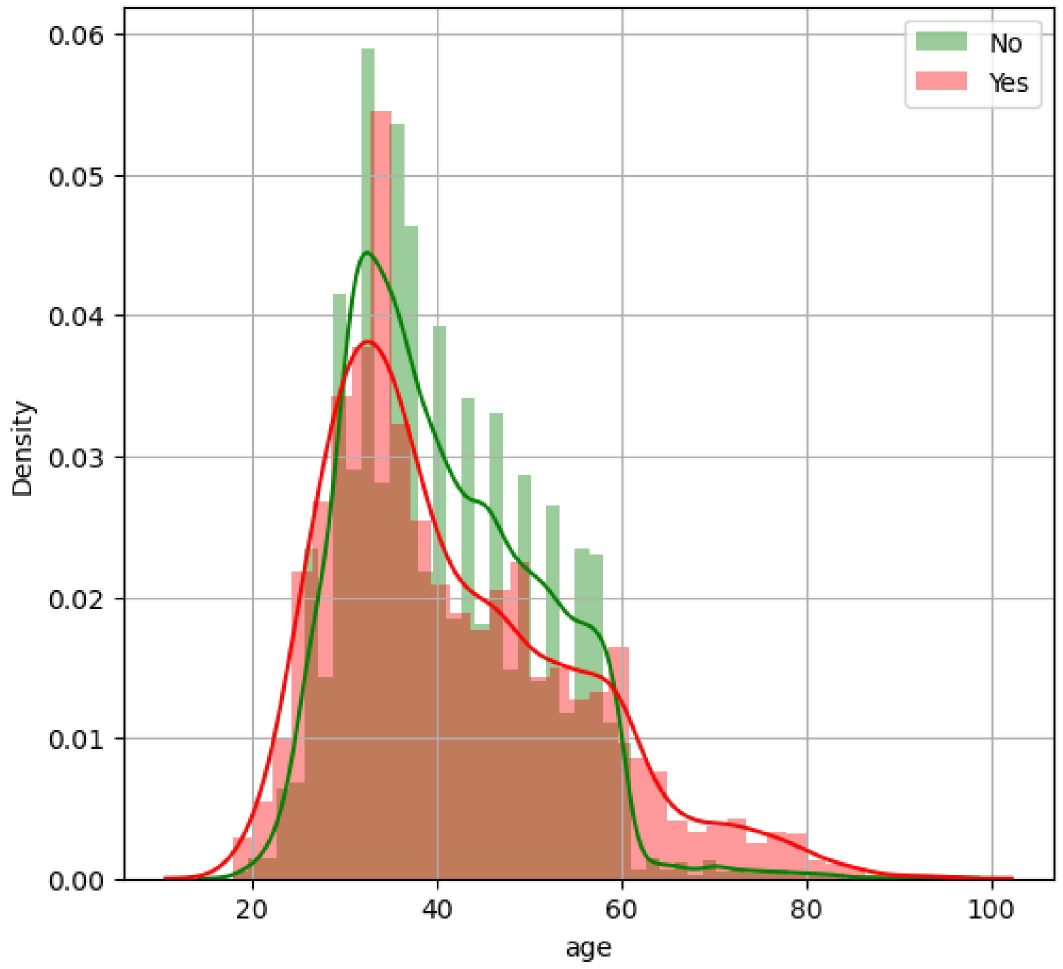
y	no	yes	yes percentage	no percentage	percentage of poutcome
poutcome					
unknown	33573	3386	0.091615	0.908385	81.75
success	533	978	0.647253	0.352747	3.34
failure	4283	618	0.126097	0.873903	10.84
other	1533	307	0.166848	0.833152	4.07

Poutcomes are the result of previous marketing campaigns. In the tables and graphs it can be seen that the results of a successful marketing campaign have a large enough impact for someone to subscribe to a deposit. This can be seen from the success column yes percentage, where a value of 64.7% is obtained with a total of 978 who subscribe to deposits. Therefore, the results of the previous successful campaign had a great impact and became one of the significant factors on the number of deposit customers.

b. Numerical Data Analysis

1. AGE



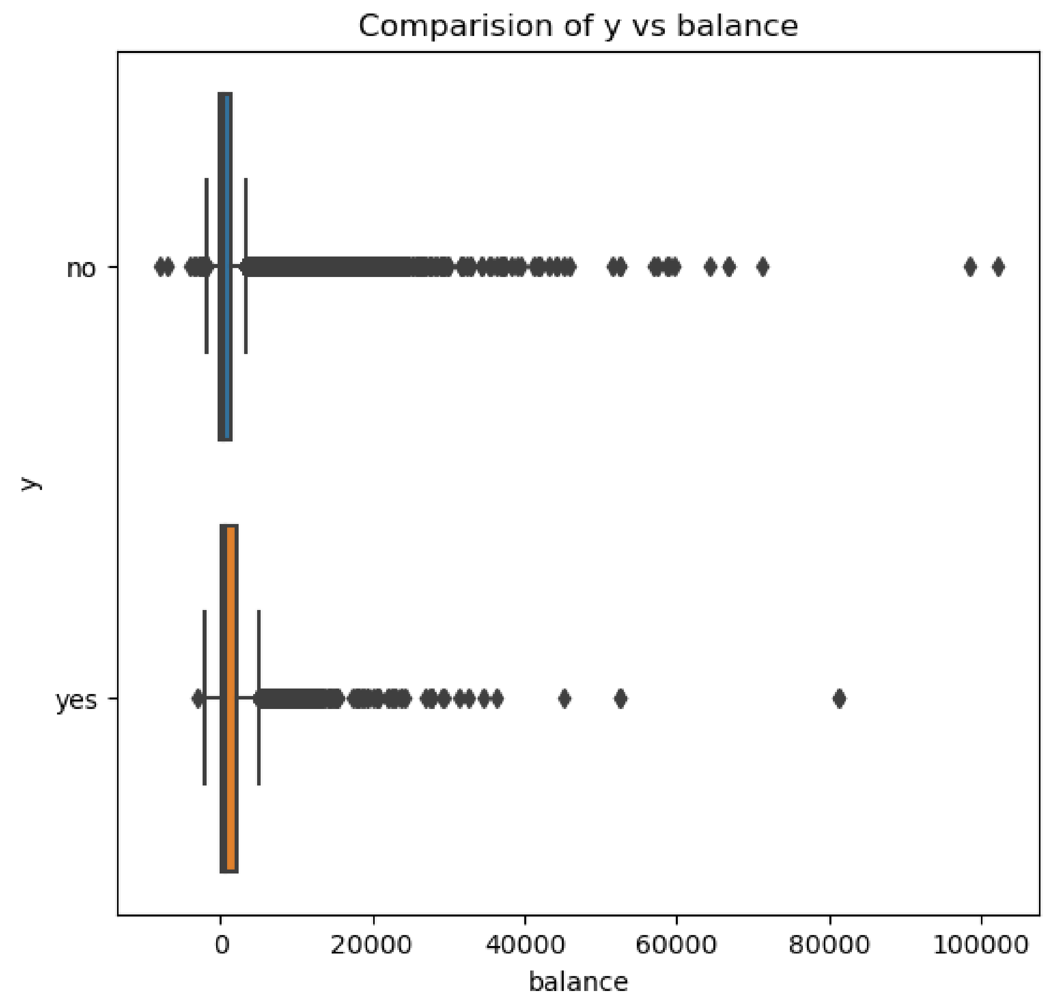


y age	no	yes	yes percentage	no percentage
32	1864	221	0.105995	0.894005
30	1540	217	0.123506	0.876494
33	1762	210	0.106491	0.893509
35	1685	209	0.110348	0.889652
31	1790	206	0.103206	0.896794
...	...	...	...	...
93	0	2	1.000000	0.000000
95	1	1	0.500000	0.500000
88	2	0	0.000000	1.000000
89	3	0	0.000000	1.000000
94	1	0	0.000000	1.000000

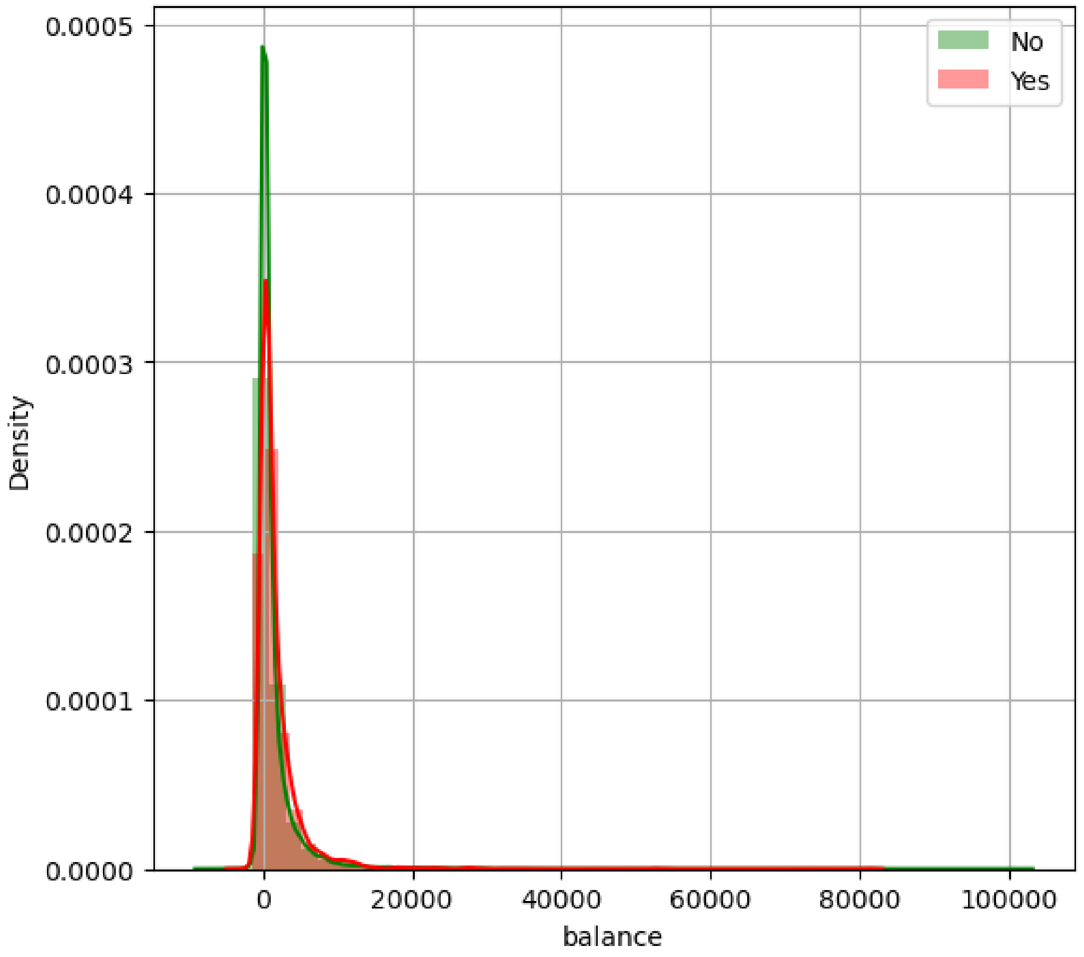
Nilai minimum pada kolom age adalah 18
Nilai maksimum pada kolom age adalah 95

Based on the age distribution table, it can be seen that yes the percentage will be large in the age range > 60. This is in accordance with the previous analysis on Categorical Jobs, where the retired type is the largest percentage.

2. BALANCE



<Figure size 600x400 with 0 Axes>
Nilai minimum pada kolom balance adalah -8019
Nilai maksimum pada kolom balance adalah 102127

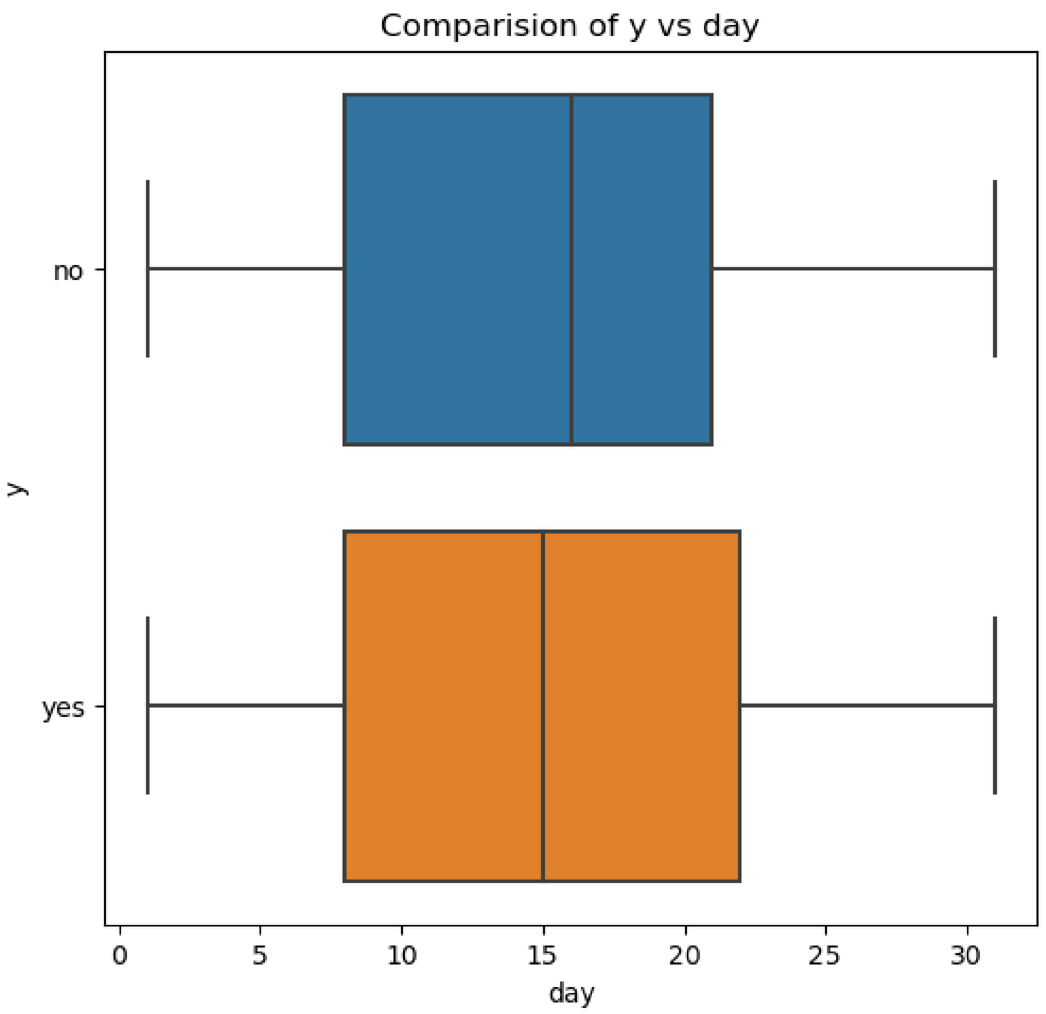


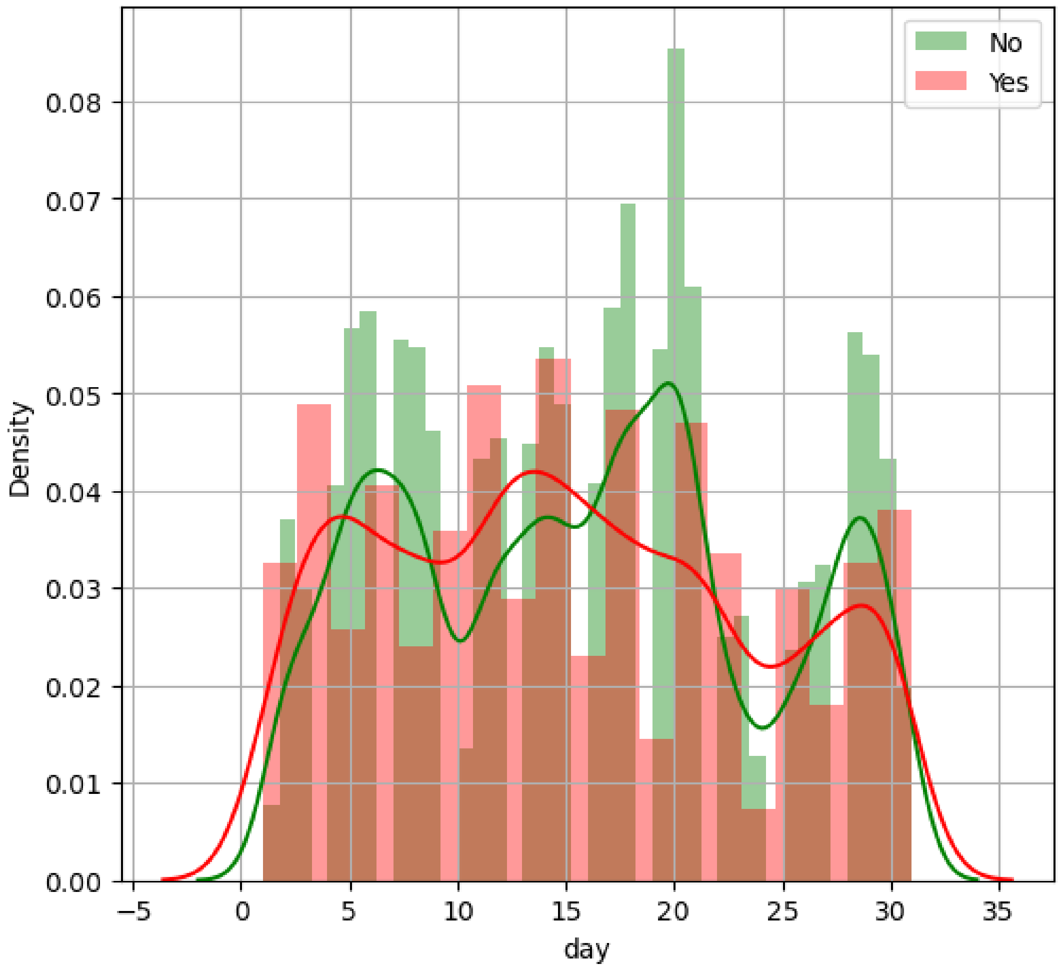
y	no	yes	yes percentage	no percentage
balance				
0	3222	292	0.083096	0.916904
1	178	17	0.087179	0.912821
2	140	16	0.102564	0.897436
5	98	15	0.132743	0.867257
294	22	12	0.352941	0.647059
...	...	...	...	...
66653	1	0	0.000000	1.000000
66721	1	0	0.000000	1.000000
71188	1	0	0.000000	1.000000
98417	1	0	0.000000	1.000000
102127	1	0	0.000000	1.000000

7168 rows × 4 columns

The highest contribution is at the average annual balance of 0. There are 292 customers who are deposit customers at balance = 0.

3. DAY



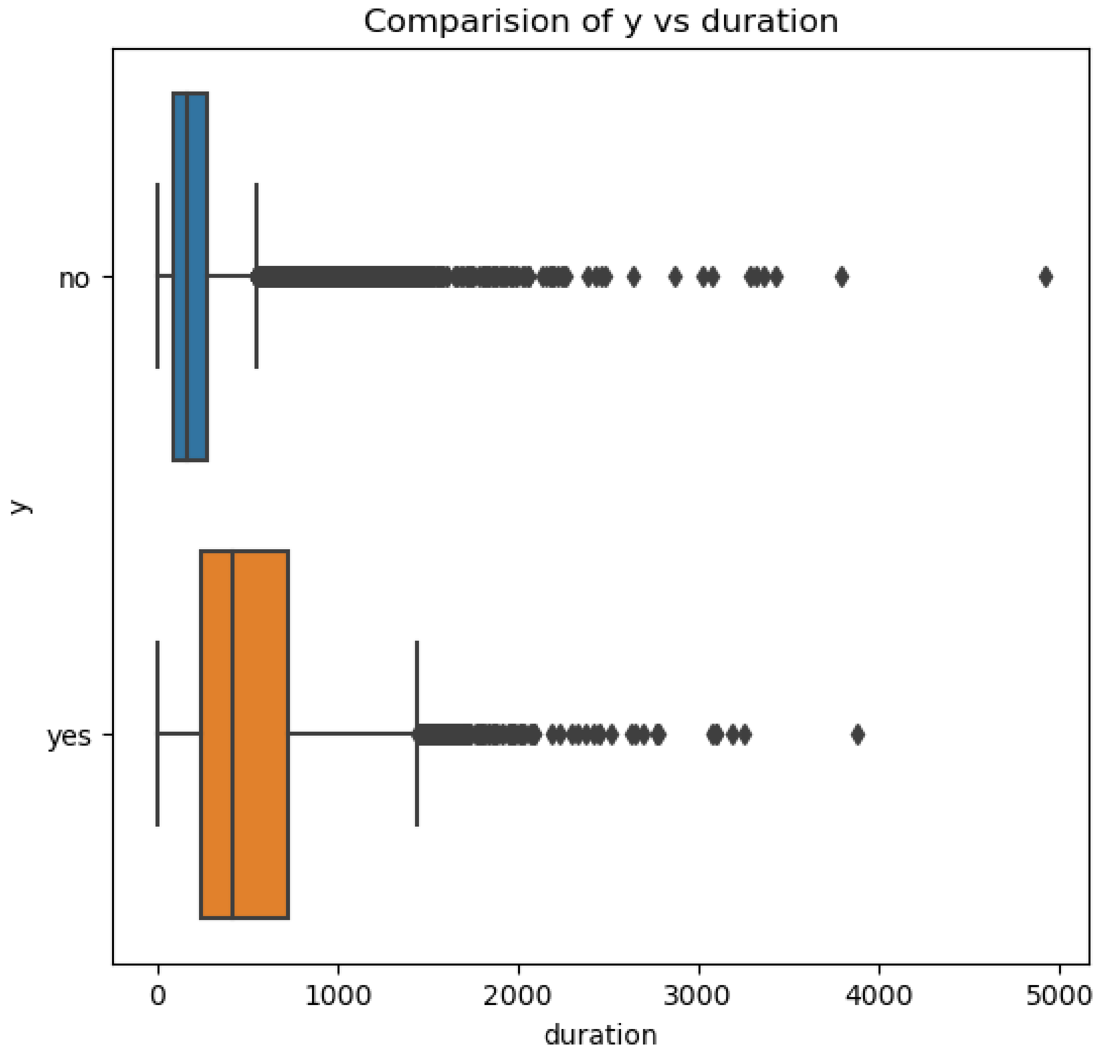


y day	no	yes	yes percentage	no percentage
30	1295	271	0.173052	0.826948
12	1359	244	0.152215	0.847785
13	1344	241	0.152050	0.847950
15	1465	238	0.139753	0.860247
4	1215	230	0.159170	0.840830
18	2080	228	0.098787	0.901213
5	1695	215	0.112565	0.887435
14	1638	210	0.113636	0.886364
8	1641	201	0.109121	0.890879
21	1825	201	0.099210	0.900790
16	1223	192	0.135689	0.864311
20	2560	192	0.069767	0.930233
2	1111	182	0.140758	0.859242
11	1298	181	0.122380	0.877620
6	1751	181	0.093685	0.906315
9	1382	179	0.114670	0.885330
3	901	178	0.164968	0.835032
17	1763	176	0.090768	0.909232
7	1660	157	0.086406	0.913594
22	751	154	0.170166	0.829834
27	971	150	0.133809	0.866191
28	1687	143	0.078142	0.921858
25	707	133	0.158333	0.841667
29	1616	129	0.073926	0.926074
23	813	126	0.134185	0.865815
19	1635	122	0.069437	0.930563
10	403	121	0.230916	0.769084
26	919	116	0.112077	0.887923
1	232	90	0.279503	0.720497
24	385	62	0.138702	0.861298
31	597	46	0.071540	0.928460

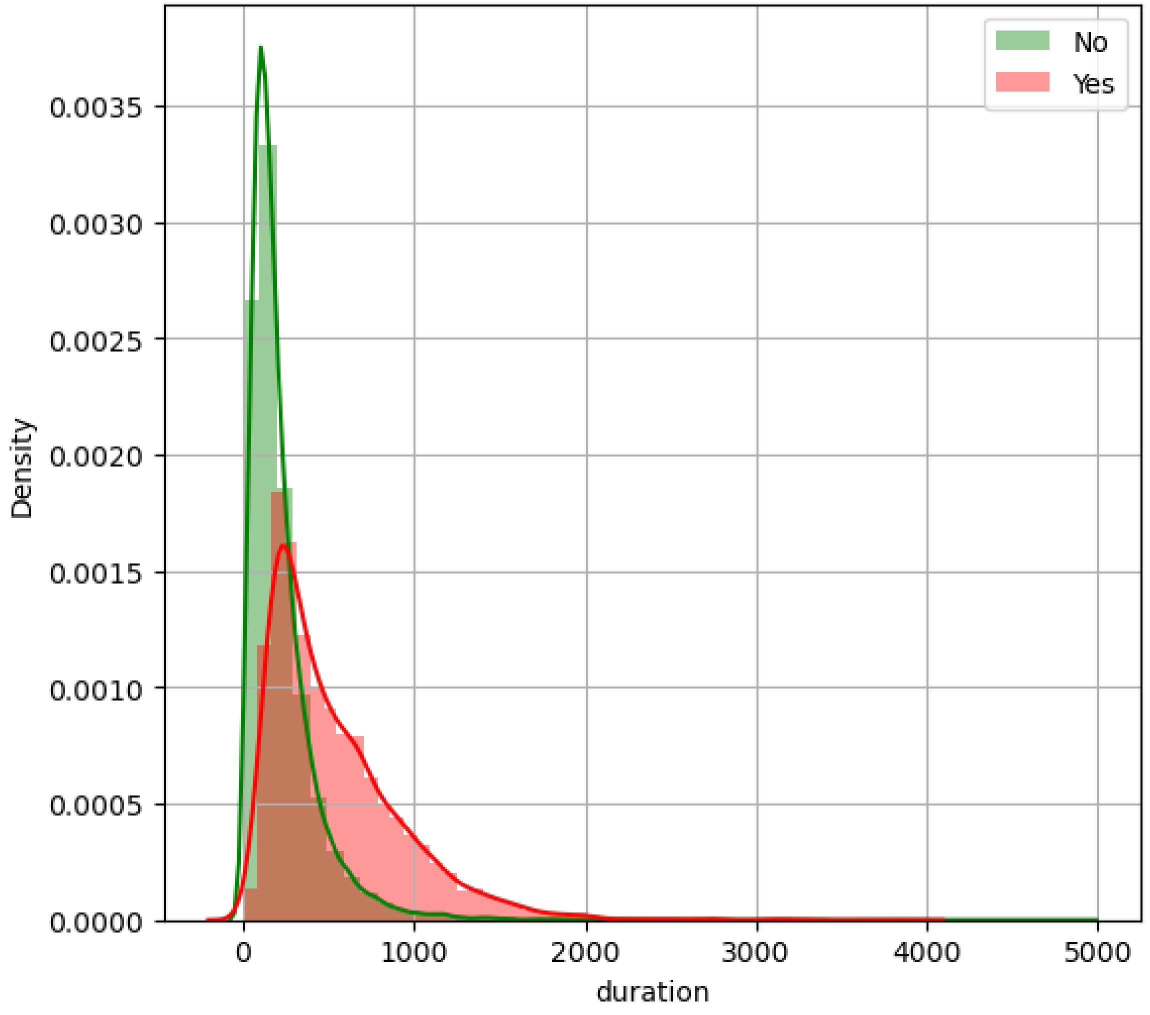
<Figure size 600x400 with 0 Axes>
Nilai minimum pada kolom day adalah 1
Nilai maksimum pada kolom day adalah 31

Seen at the end of the month (30) is the highest number for someone to subscribe to deposits.

4. DURATION



<Figure size 600x400 with 0 Axes>
Nilai minimum pada kolom duration adalah 0
Nilai maksimum pada kolom duration adalah 4918

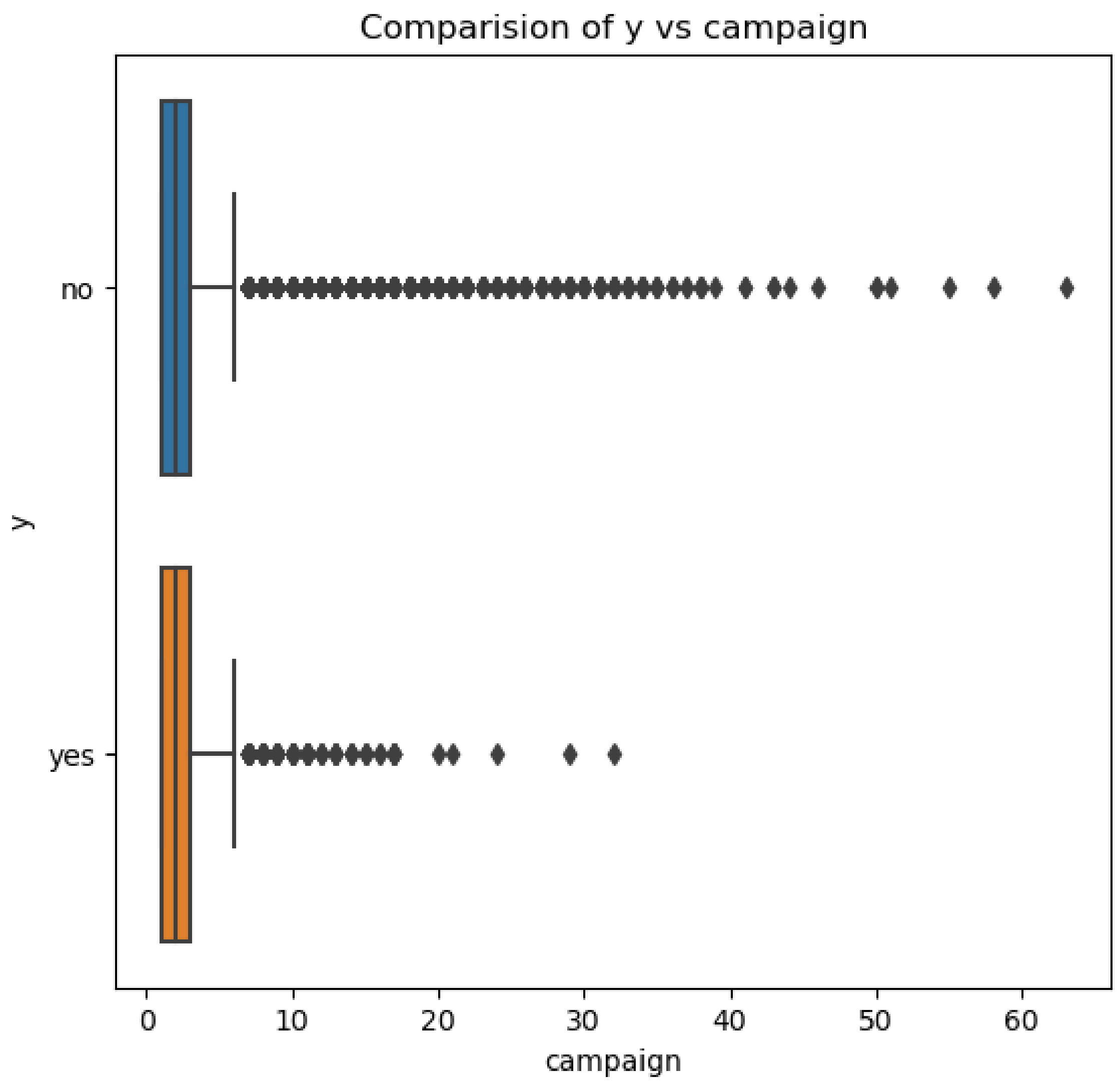


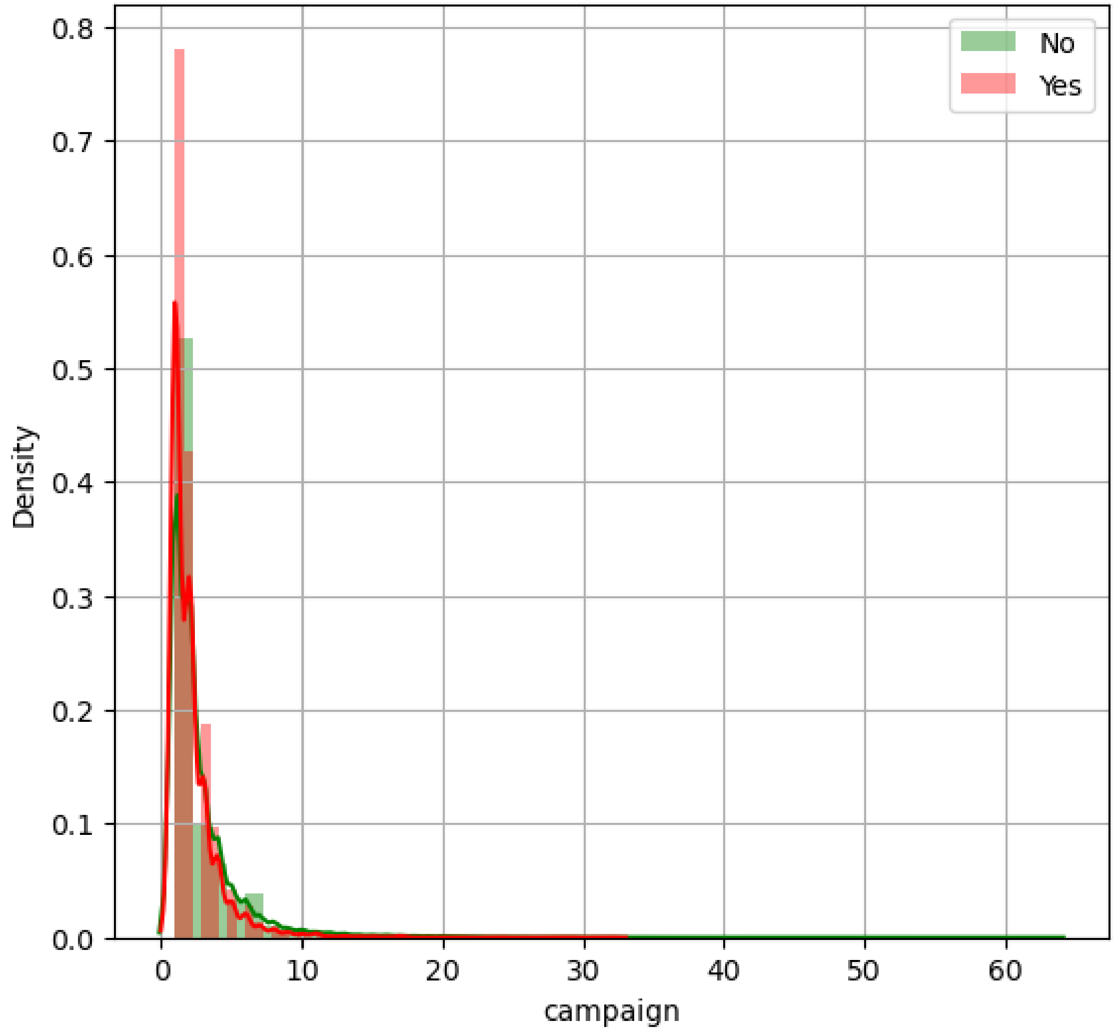
y	no	yes	yes percentage	no percentage
duration				
261	54	19	0.260274	0.739726
226	101	17	0.144068	0.855932
229	71	16	0.183908	0.816092
232	74	16	0.177778	0.822222
187	104	16	0.133333	0.866667
...	...	...	...	...
3322	1	0	0.000000	1.000000
3366	1	0	0.000000	1.000000
3422	1	0	0.000000	1.000000
3785	1	0	0.000000	1.000000
4918	1	0	0.000000	1.000000

1573 rows × 4 columns

The last contact duration in seconds seems successful enough to attract deposit customers at more than 537 seconds. This makes sense because when the communication is long, it proves that the intended person is someone who is interested in the deposit campaign.

5. CAMPAIGN



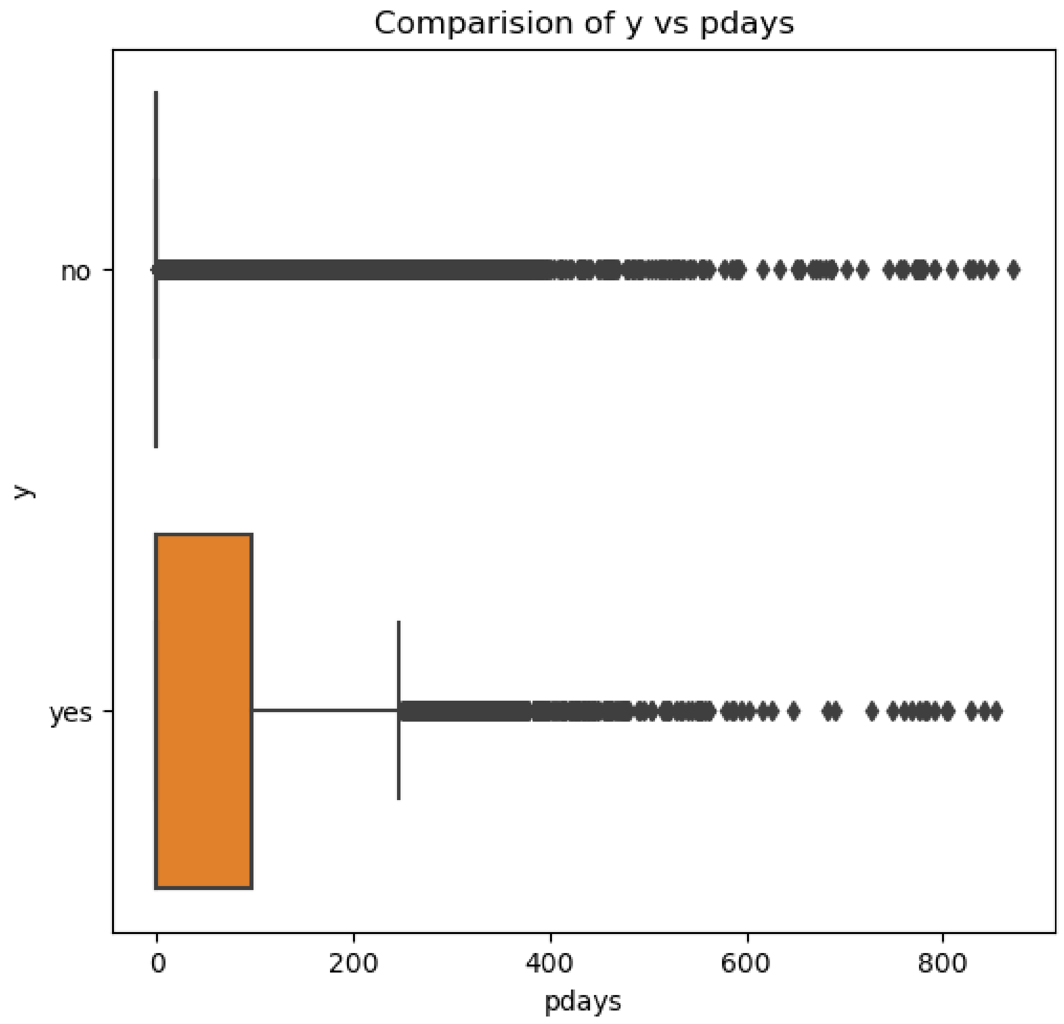


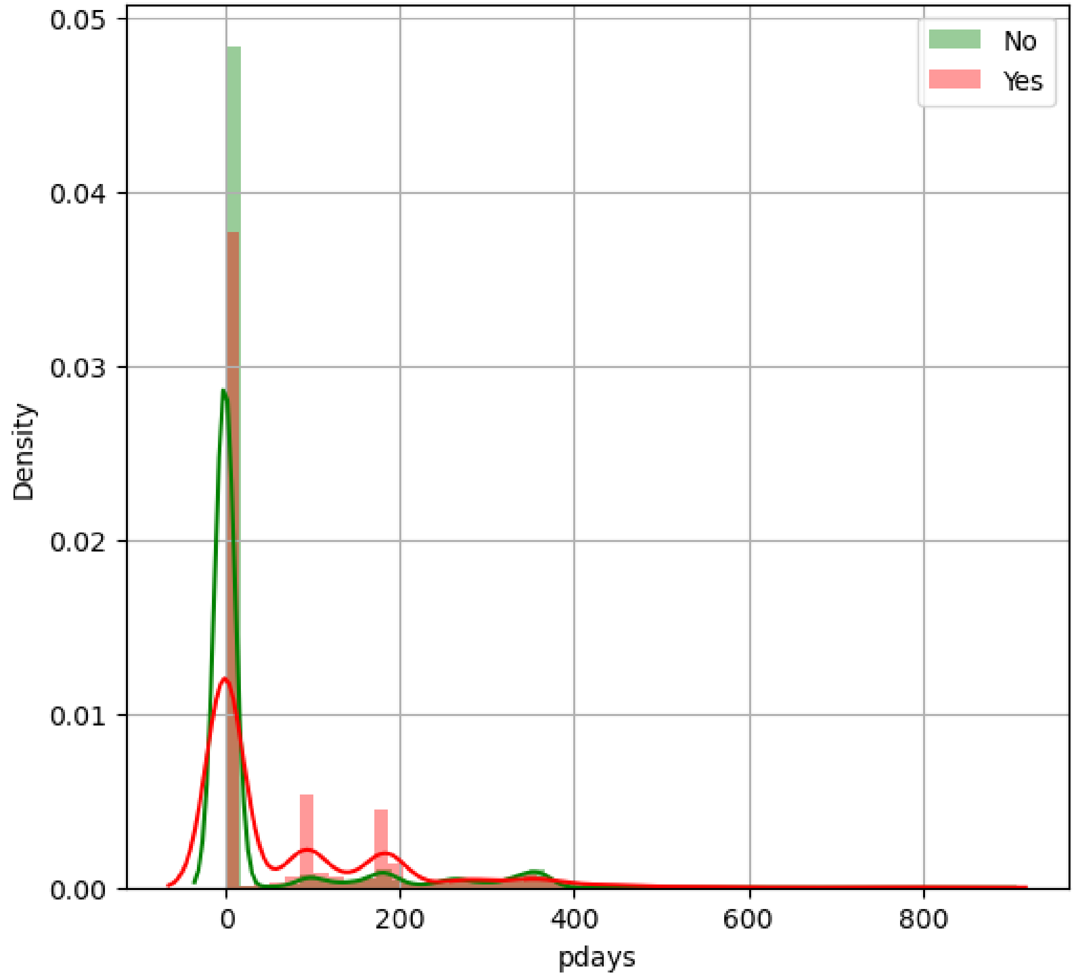
<Figure size 600x400 with 0 Axes>
Nilai minimum pada kolom campaign adalah 1
Nilai maksimum pada kolom campaign adalah 63

y campaign	no	yes	yes percentage	no percentage
1	14983	2561	0.145976	0.854024
2	11104	1401	0.112035	0.887965
3	4903	618	0.111936	0.888064
4	3205	317	0.090006	0.909994
5	1625	139	0.078798	0.921202
6	1199	92	0.071263	0.928737
7	688	47	0.063946	0.936054
8	508	32	0.059259	0.940741
9	306	21	0.064220	0.935780
11	185	16	0.079602	0.920398
10	252	14	0.052632	0.947368
17	63	6	0.086957	0.913043
13	127	6	0.045113	0.954887
15	80	4	0.047619	0.952381
14	89	4	0.043011	0.956989
12	151	4	0.025806	0.974194
16	77	2	0.025316	0.974684
32	8	1	0.111111	0.888889
29	15	1	0.062500	0.937500
24	19	1	0.050000	0.950000
21	34	1	0.028571	0.971429
20	42	1	0.023256	0.976744
18	51	0	0.000000	1.000000
19	44	0	0.000000	1.000000
22	23	0	0.000000	1.000000
23	22	0	0.000000	1.000000
25	22	0	0.000000	1.000000
26	13	0	0.000000	1.000000
27	10	0	0.000000	1.000000
28	16	0	0.000000	1.000000
30	8	0	0.000000	1.000000
31	12	0	0.000000	1.000000
33	6	0	0.000000	1.000000
34	5	0	0.000000	1.000000
35	4	0	0.000000	1.000000
36	4	0	0.000000	1.000000
37	2	0	0.000000	1.000000
38	3	0	0.000000	1.000000
39	1	0	0.000000	1.000000
41	2	0	0.000000	1.000000
43	3	0	0.000000	1.000000
44	1	0	0.000000	1.000000
46	1	0	0.000000	1.000000
50	2	0	0.000000	1.000000
51	1	0	0.000000	1.000000
55	1	0	0.000000	1.000000
58	1	0	0.000000	1.000000
63	1	0	0.000000	1.000000

The number of contacts made during the campaign was mostly done at the beginning and very few at the end of the campaign.

6. PDAYS



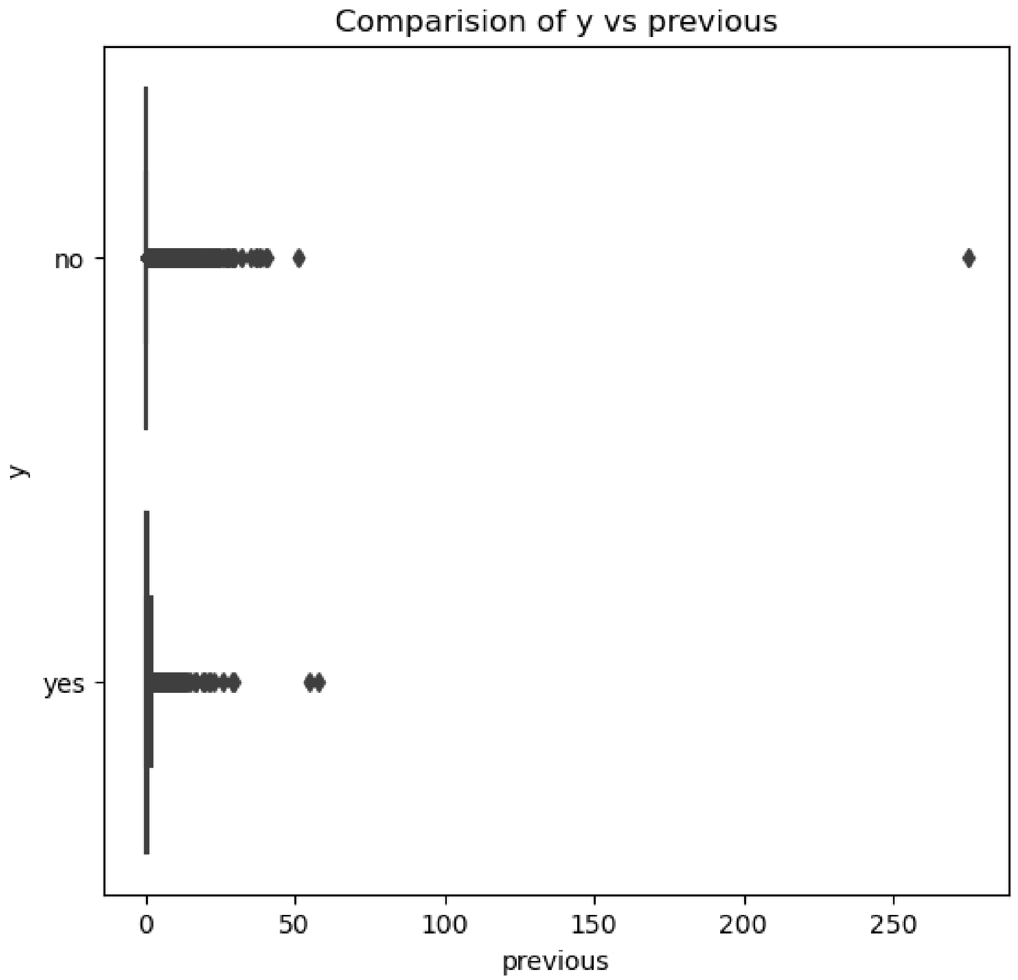


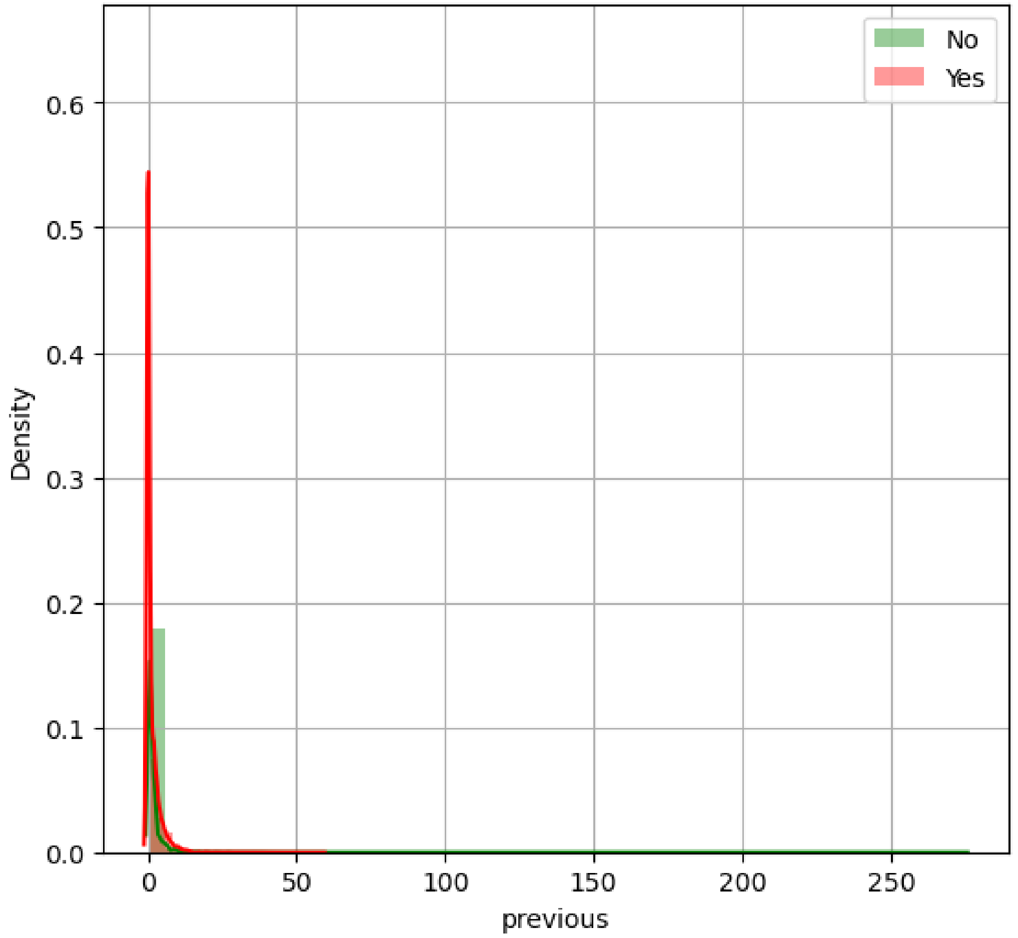
y pdays	no	yes	yes percentage	no percentage
-1	33570	3384	0.091573	0.908427
92	51	96	0.653061	0.346939
182	87	80	0.479042	0.520958
91	51	75	0.595238	0.404762
181	43	74	0.632479	0.367521
...	...	...	...	...
826	1	0	0.000000	1.000000
831	1	0	0.000000	1.000000
838	1	0	0.000000	1.000000
850	1	0	0.000000	1.000000
871	1	0	0.000000	1.000000

```
<Figure size 600x400 with 0 Axes>
Nilai minimum pada kolom pdays adalah -1
Nilai maksimum pada kolom pdays adalah 871
```

-1 means that the client was not contacted in the previous days. This is very because -1 has the most total but very few new subscribers are subscribed. In contrast to 92, 182, 91, and 181 which have the most total new clients and a fairly good proportion. In the table it can also be said that a good day to contact the previous client is 3 - 6 months in advance.

7. PREVIOUS

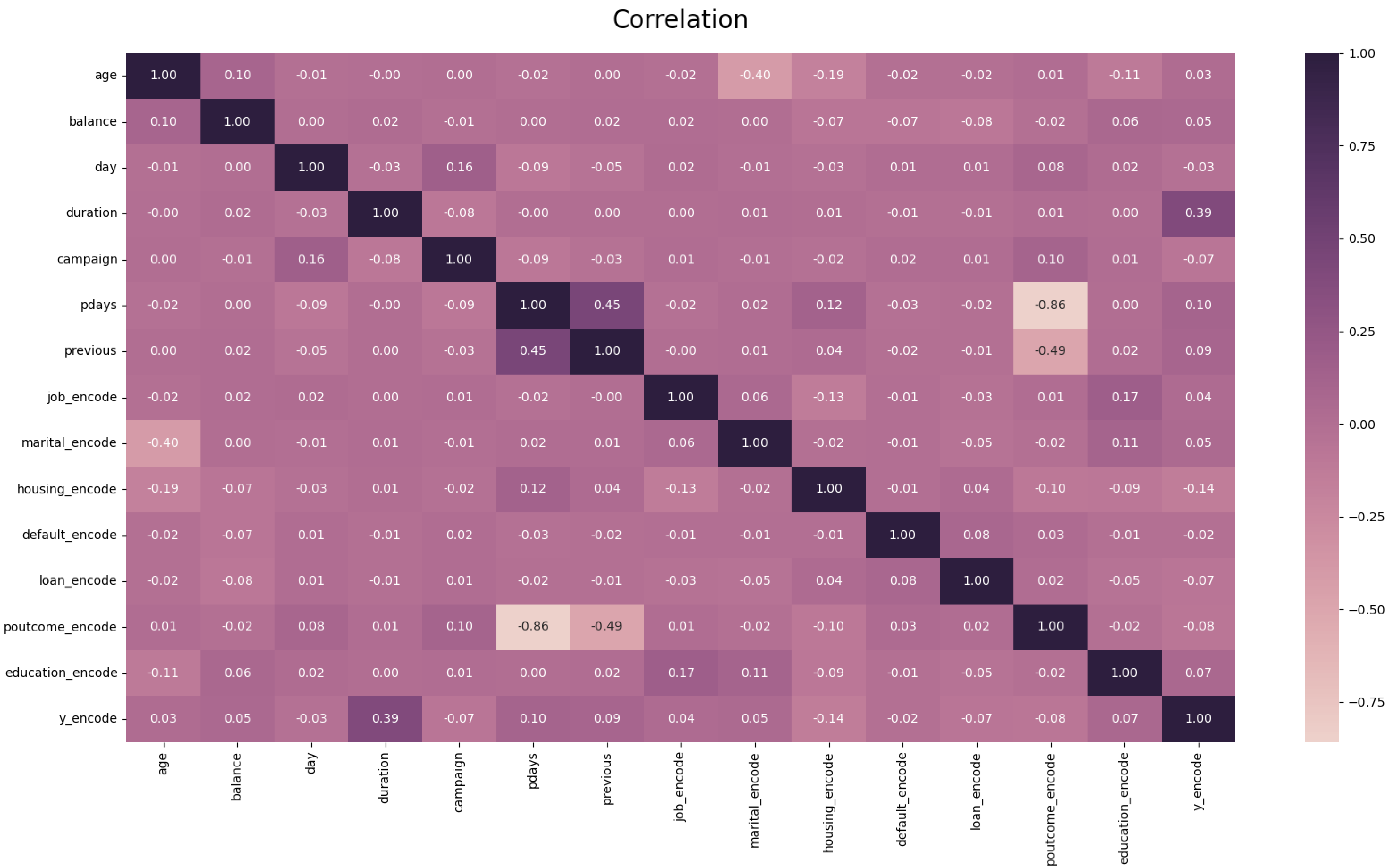




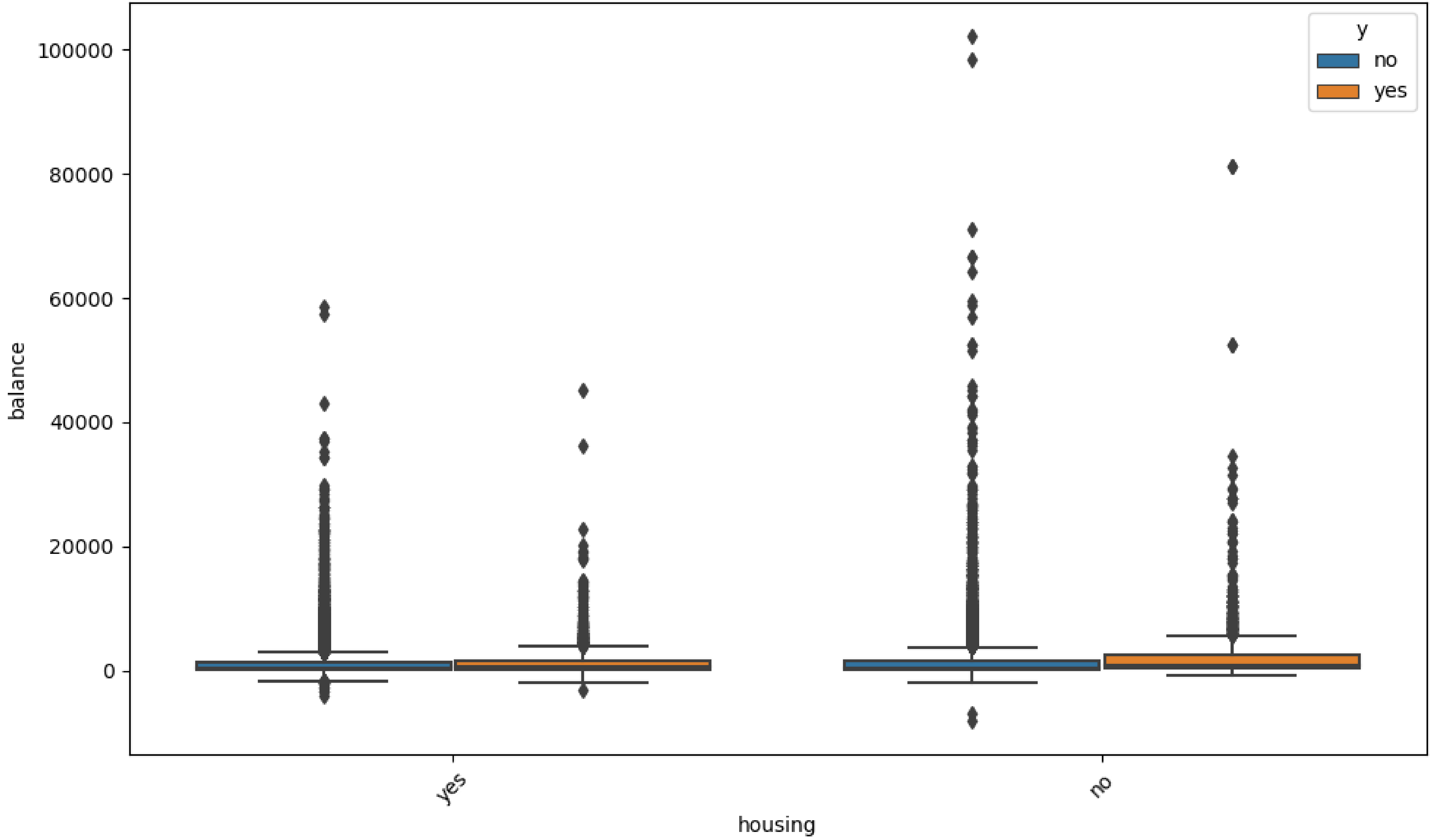
<Figure size 600x400 with 0 Axes>
Nilai minimum pada kolom previous adalah 0
Nilai maksimum pada kolom previous adalah 275

y previous	no	yes	yes percentage	no percentage
0	33570	3384	0.091573	0.908427
1	2189	583	0.210317	0.789683
2	1650	456	0.216524	0.783476
3	848	294	0.257443	0.742557
4	543	171	0.239496	0.760504
5	338	121	0.263617	0.736383
6	194	83	0.299639	0.700361
7	151	54	0.263415	0.736585
8	90	39	0.302326	0.697674
10	41	26	0.388060	0.611940
9	68	24	0.260870	0.739130
11	50	15	0.230769	0.769231
12	34	10	0.227273	0.772727
13	29	9	0.236842	0.763158
14	14	5	0.263158	0.736842
17	12	3	0.200000	0.800000
19	9	2	0.181818	0.818182
55	0	1	1.000000	0.000000
58	0	1	1.000000	0.000000
26	1	1	0.500000	0.500000
30	2	1	0.333333	0.666667
21	3	1	0.250000	0.750000
29	3	1	0.250000	0.750000
22	5	1	0.166667	0.833333
20	7	1	0.125000	0.875000
23	7	1	0.125000	0.875000
15	19	1	0.050000	0.950000
16	13	0	0.000000	1.000000
18	6	0	0.000000	1.000000
24	5	0	0.000000	1.000000
25	4	0	0.000000	1.000000
27	5	0	0.000000	1.000000
28	2	0	0.000000	1.000000
32	1	0	0.000000	1.000000
35	1	0	0.000000	1.000000
37	2	0	0.000000	1.000000
38	2	0	0.000000	1.000000
40	1	0	0.000000	1.000000
41	1	0	0.000000	1.000000
51	1	0	0.000000	1.000000
275	1	0	0.000000	1.000000

c. Multivariate Analysis

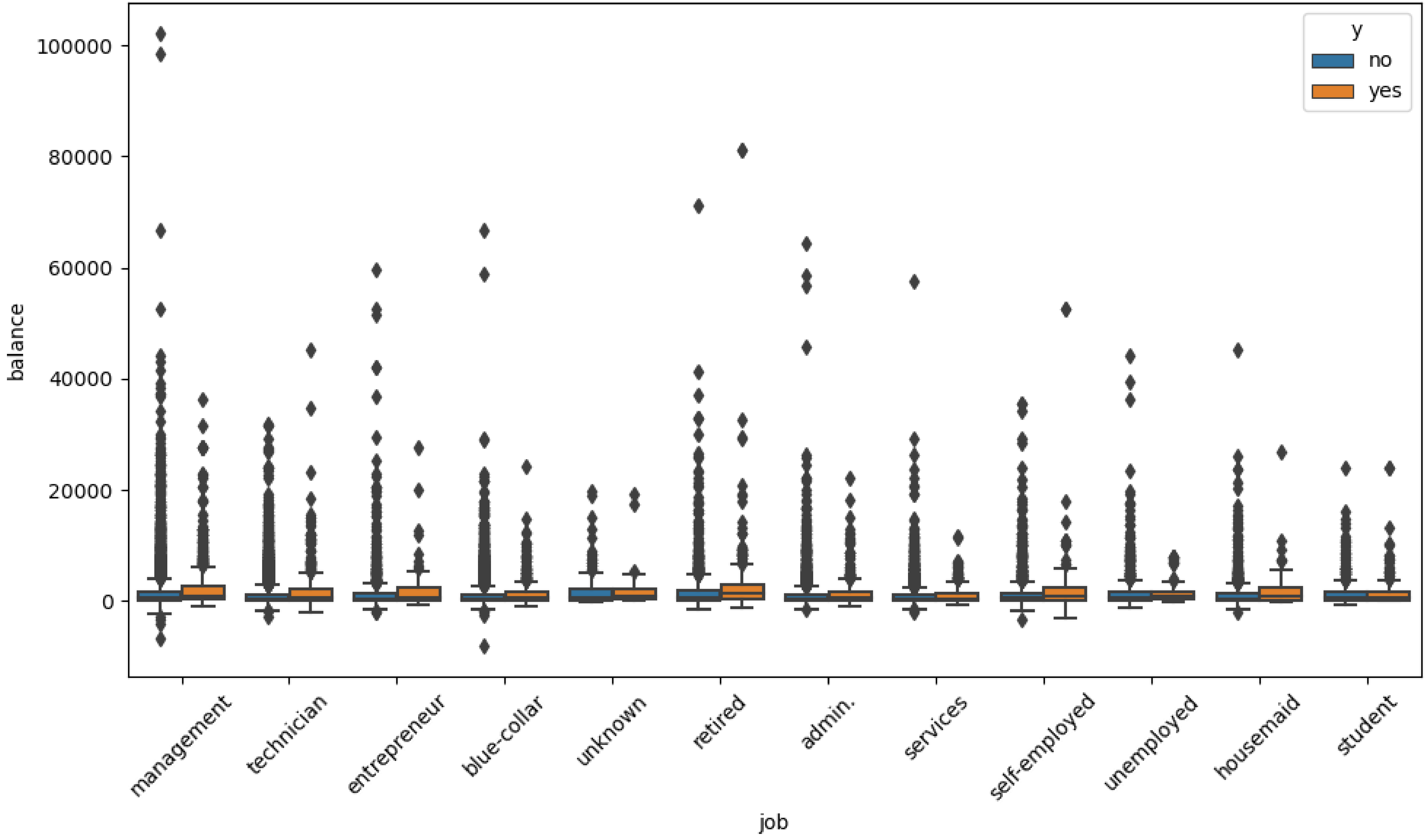


HOUSING AND BALANCE WITH Y



Here it can be seen that people who do not have a home loan have a larger mean balance than those who have a home loan. Therefore, the boxplot reinforces that someone who does not have a home loan will make a deposit because it has a higher mean balance.

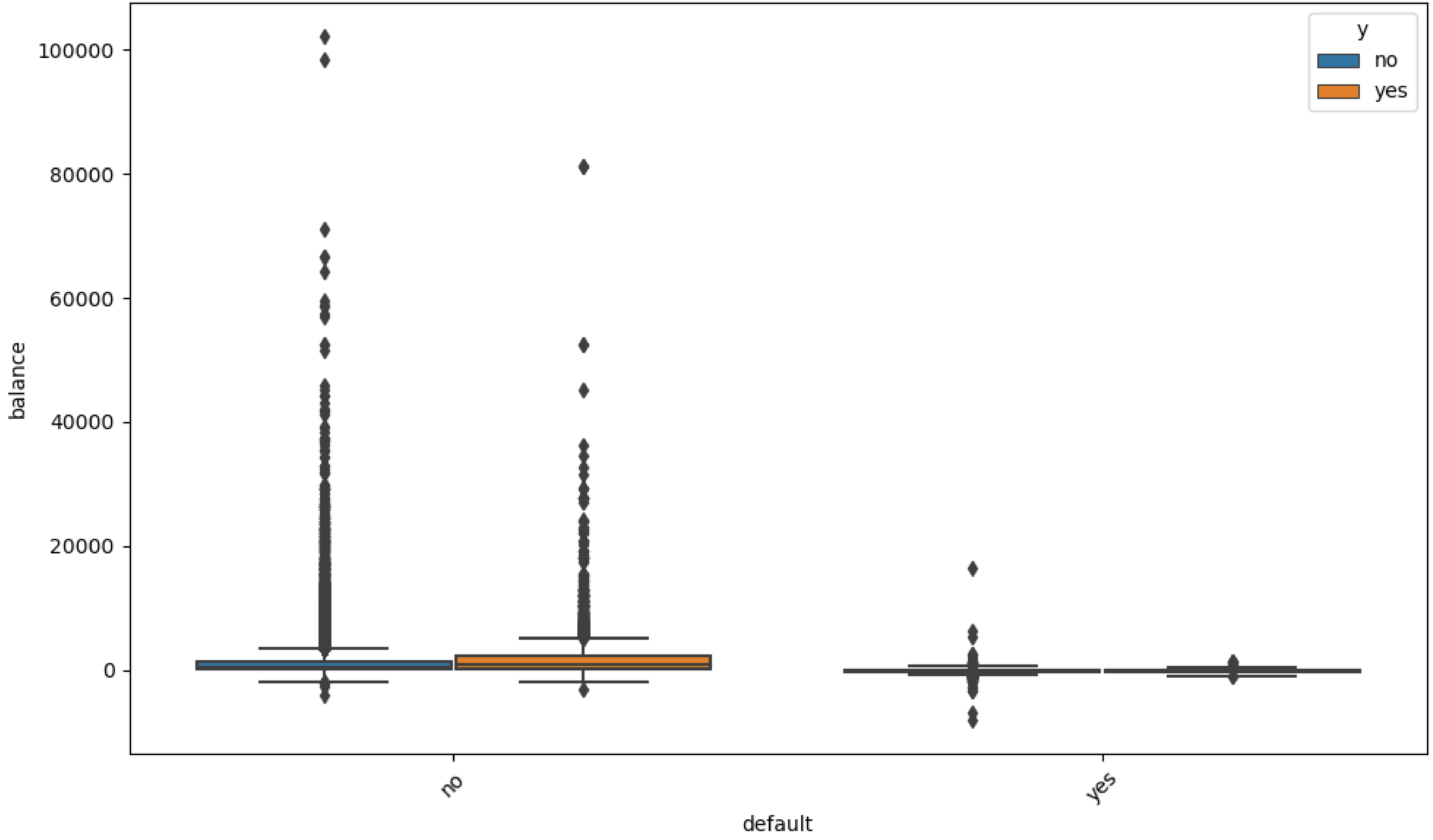
JOB AND BALANCE WITH Y



job	min	max	sum	mean
retired	-1598	81204	4492263	1984.215106
unknown	-295	19706	510439	1772.357639
management	-6847	102127	16680288	1763.616832
self-employed	-3313	52587	2602146	1647.970868
unemployed	-1270	44134	1982835	1521.745971
entrepreneur	-2082	59649	2262426	1521.470074
housemaid	-1941	45141	1726570	1392.395161
student	-679	24025	1302001	1388.060768
technician	-2827	45248	9516246	1252.632092
admin.	-1601	64343	5873423	1135.838909
blue-collar	-8019	66653	10499141	1078.826654
services	-2122	57435	4141904	997.088108

In terms of job categories, the average balance is not much different. But there is something that interests me, namely the retirees who subscribe to deposits. Retirees who subscribe to deposits have a higher average than retirees who do not subscribe.

DEFAULT AND BALANCE WITH Y



The box plot shows that the composition of no by default is the target market for this campaign. Because it looks very different data between no and yes. In addition, the mean balance of those who become deposit customers is higher than that of those who do not follow deposits.

Summary

In the campaign, it succeeded in attracting 5289 subscribers, which is 11.70% of the total. After doing the analysis, the largest percentage of jobs based on occupation are retirees and students, respectively 22.7% and 28.7%. This makes sense because people who have retired usually make deposits on their money. Furthermore, other factors that influence a person to subscribe to a deposit or not have a home loan. It also makes sense that when a person does not have a mortgage on a home, then the burden of not having a mortgage on a home is that they can save and subscribe to a deposit. The next influencing factor is the success of the previous marketing campaign. When the previous marketing campaign was successful, it could have a considerable impact, where 64.7% of the previous campaign succeeded in becoming a deposit subscription. And the last factor is the duration of the last contact, i.e. the longer the contact, the greater the subscription percentage. This is because it is interested in developing that a person is interested in what he is interested in.

TEAM 1

THANK YOU