DATA-A-THON:

TEAM newSENSE (#3)

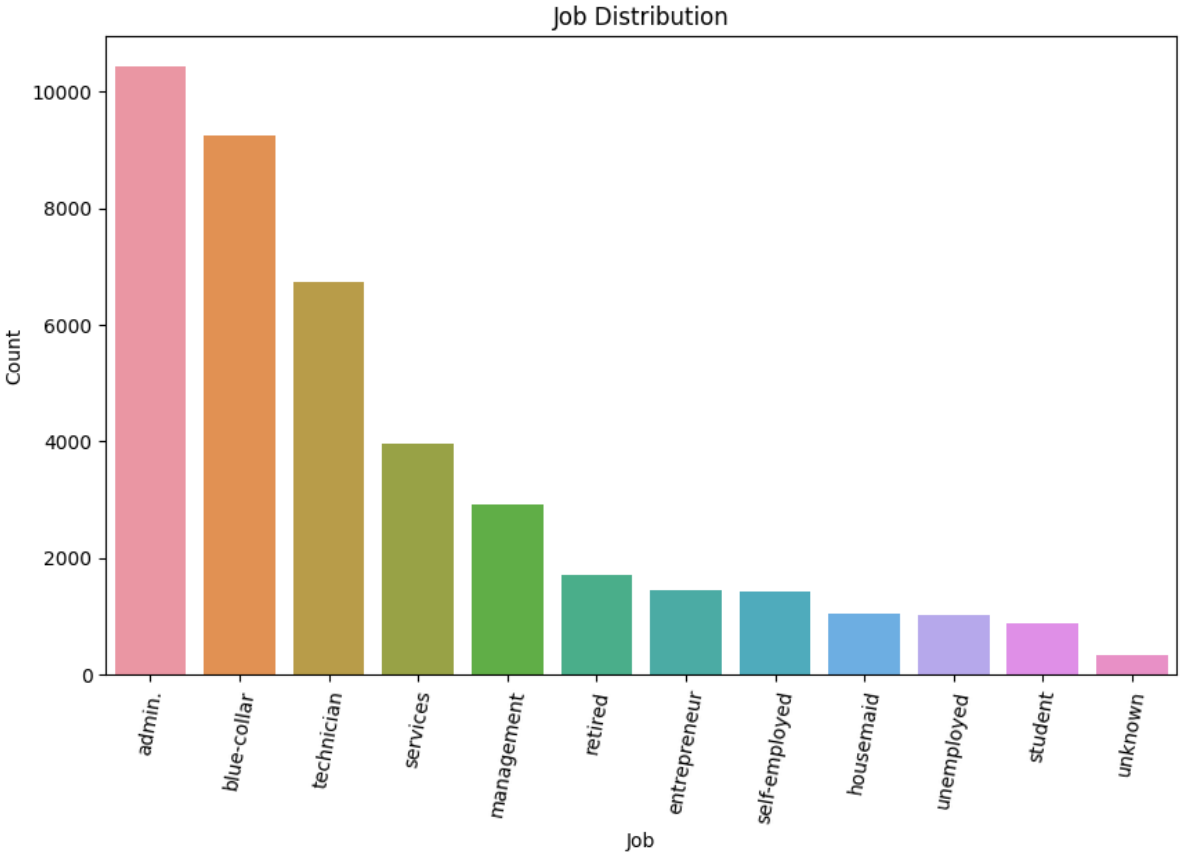
S.Mithilesh Gopalakrishnan

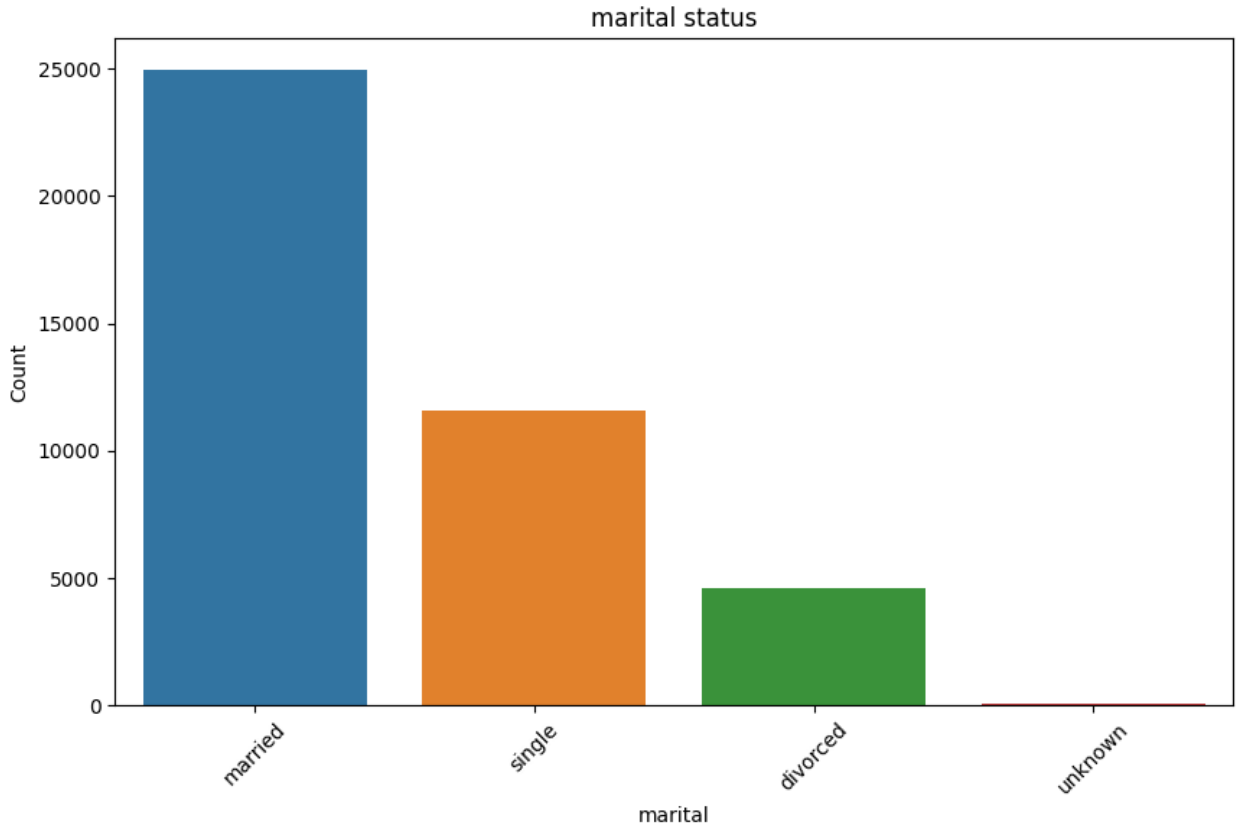
Gayathrisri . G

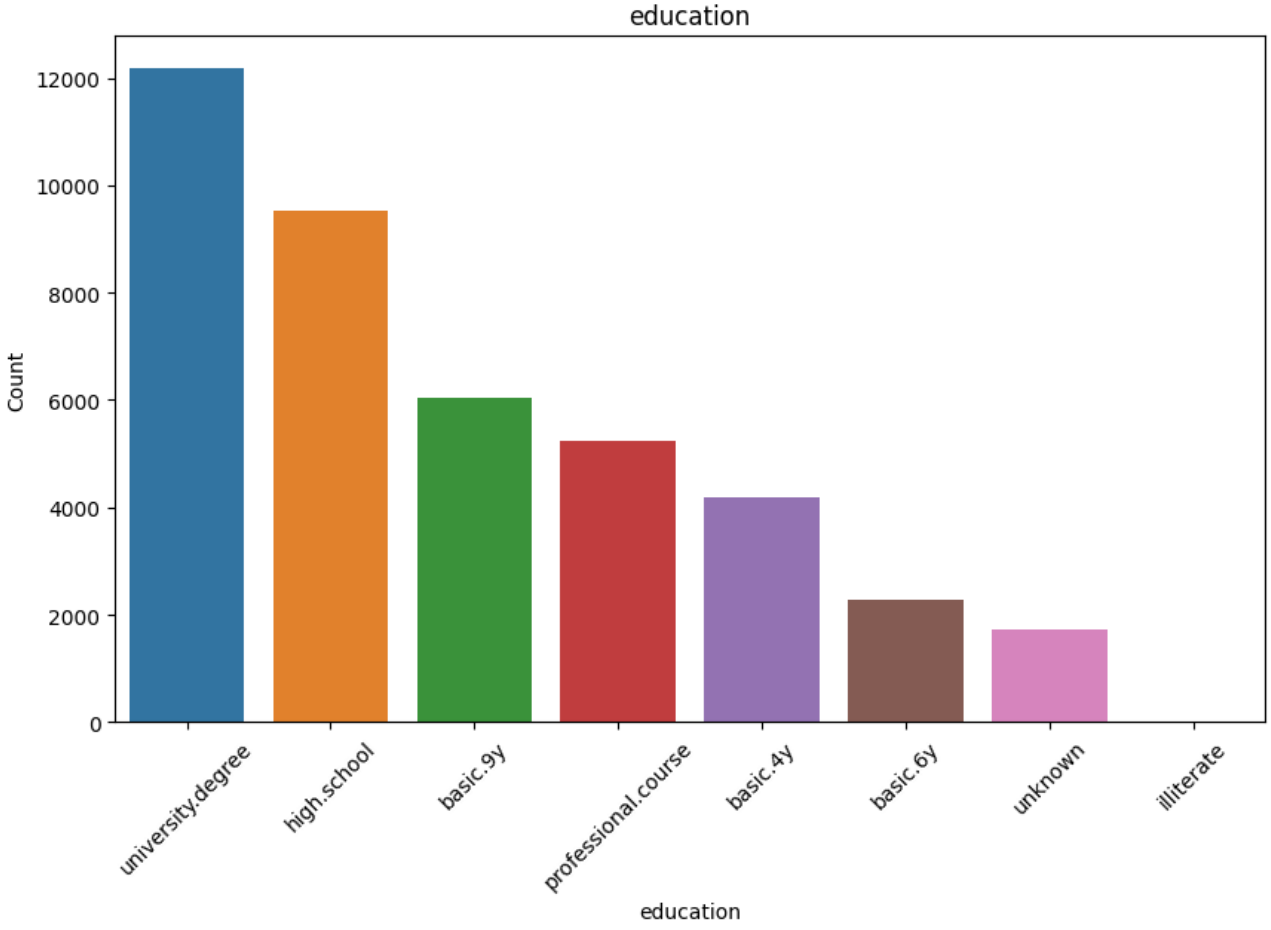
SLOT – 1:

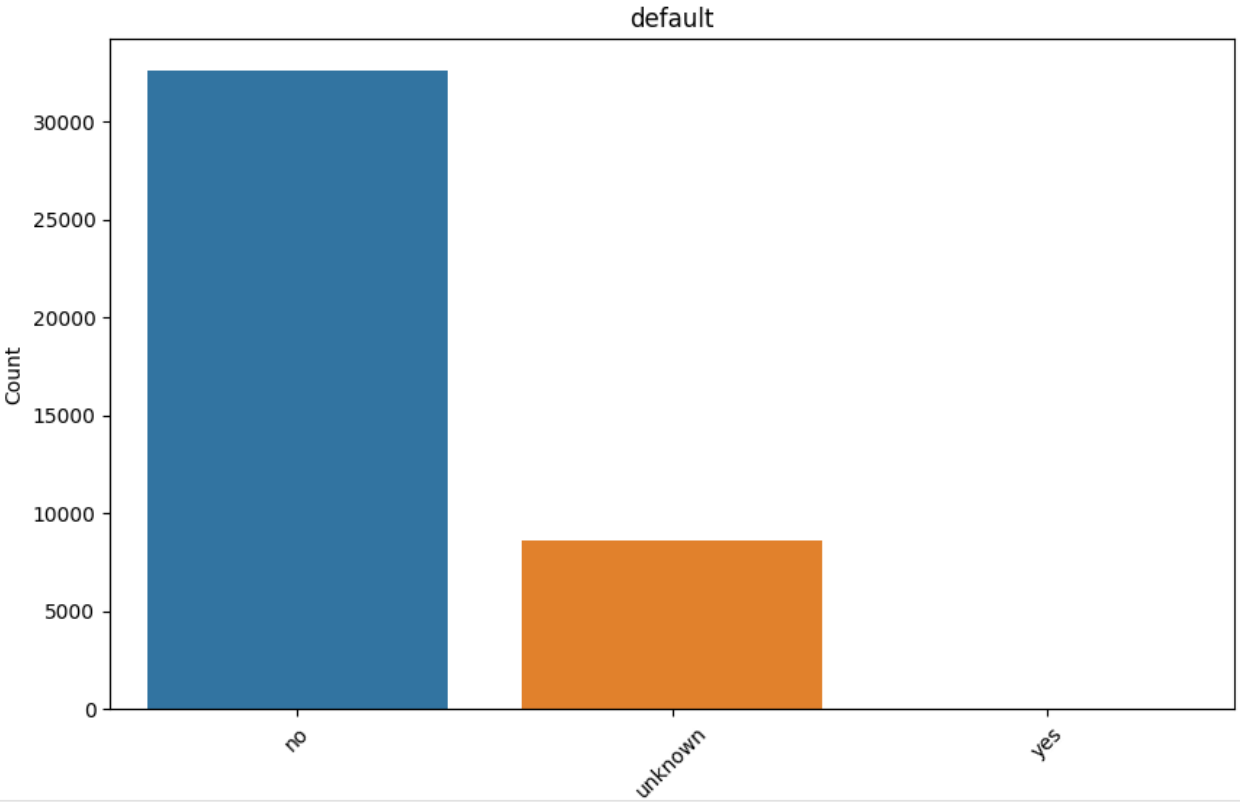
1. Exploratory Data Analysis (What does all the analysis of numerical and categorical features reveal):

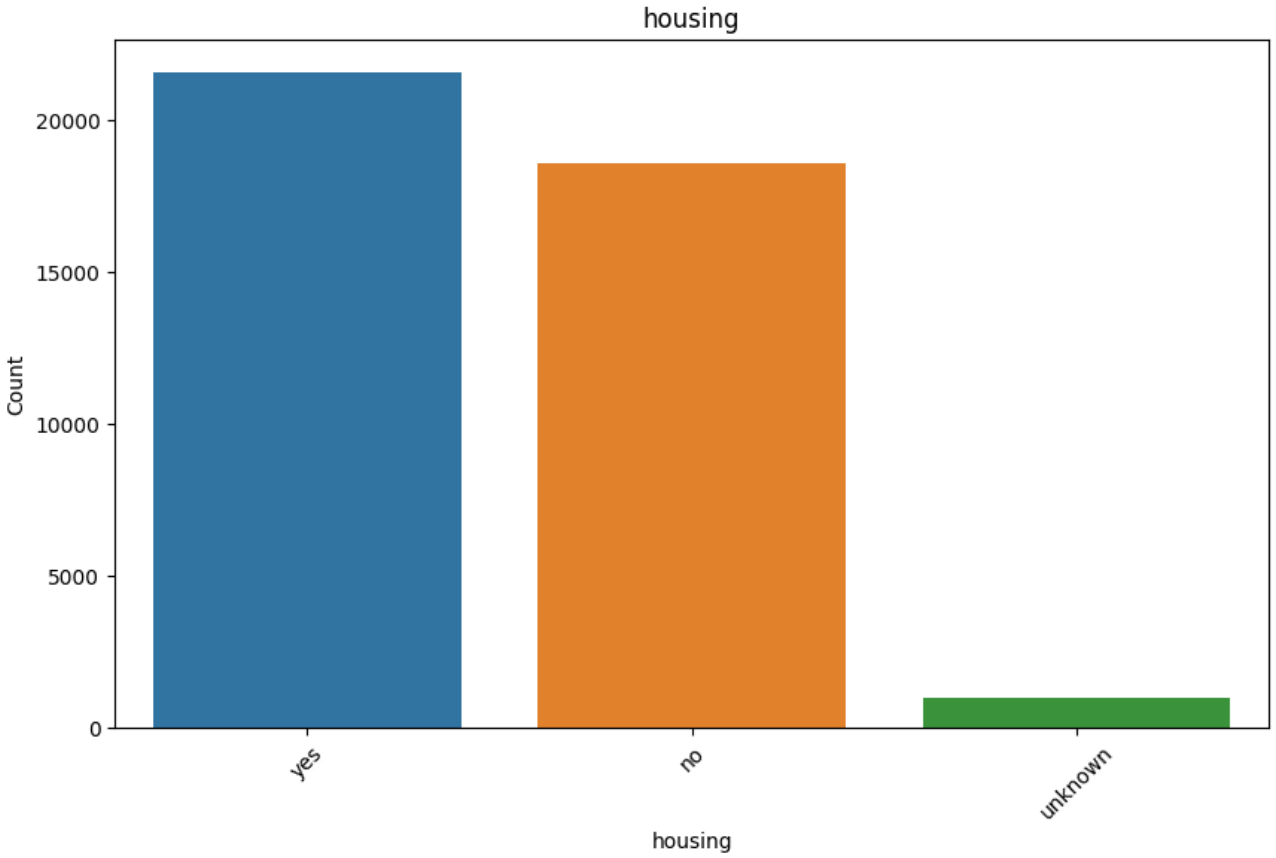
a) The Categorical Features are 'job', 'marital','education','default','housing','loan','contact','month','day_of_week','poutcome', 'y' (target variable)

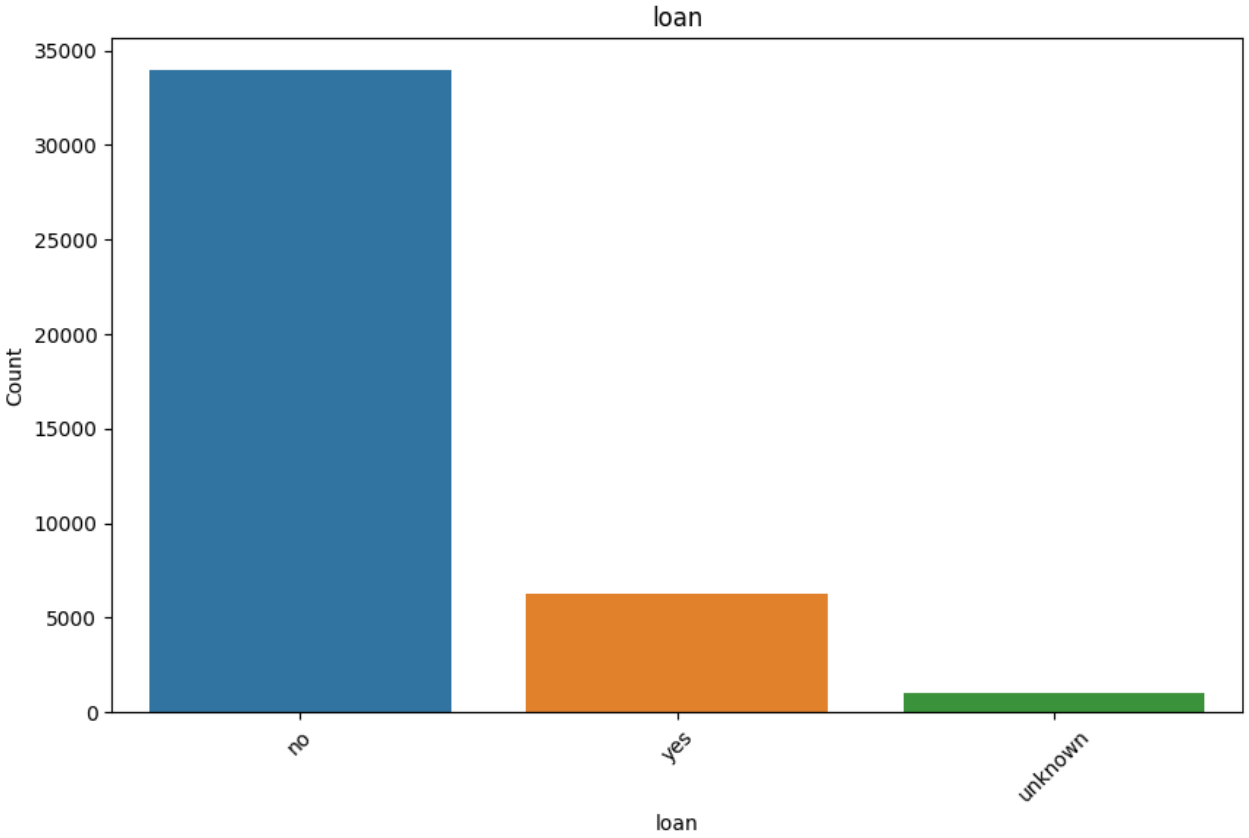


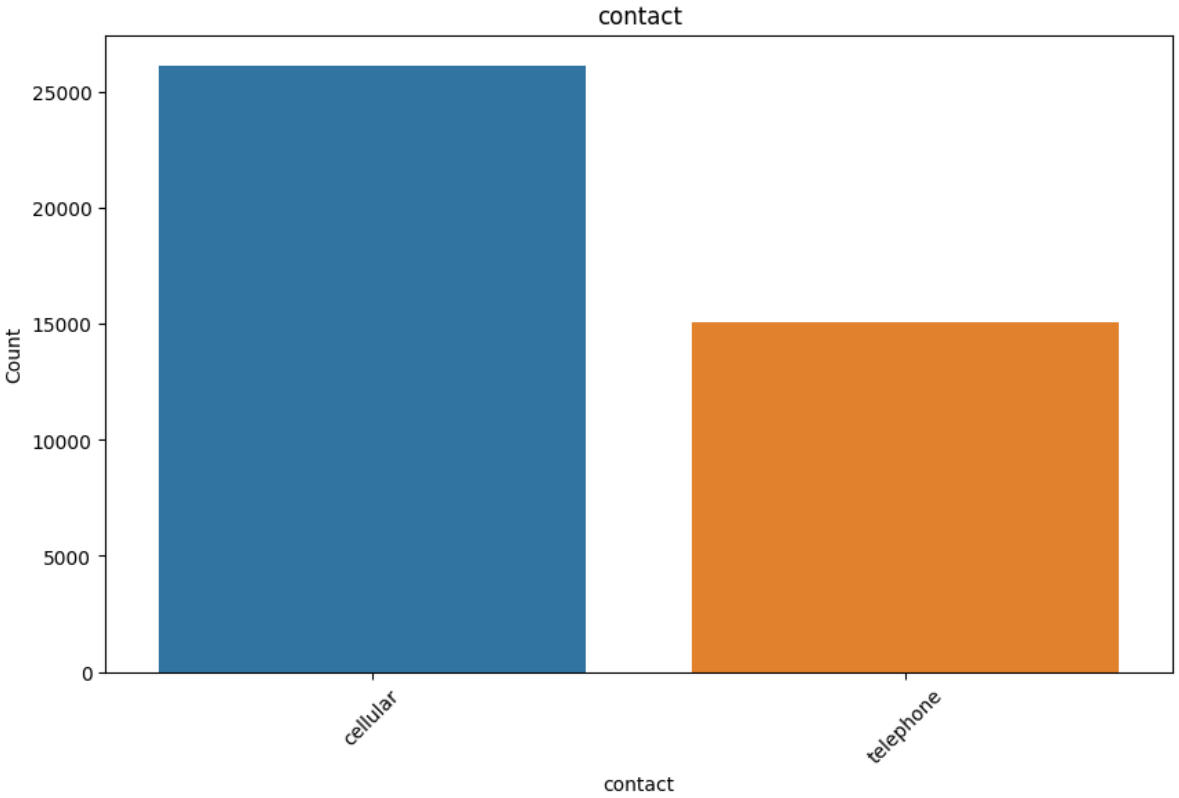

marital status
25000
20000
15000
10000
5000
0
Count
married
single
divorced
unknown
marital

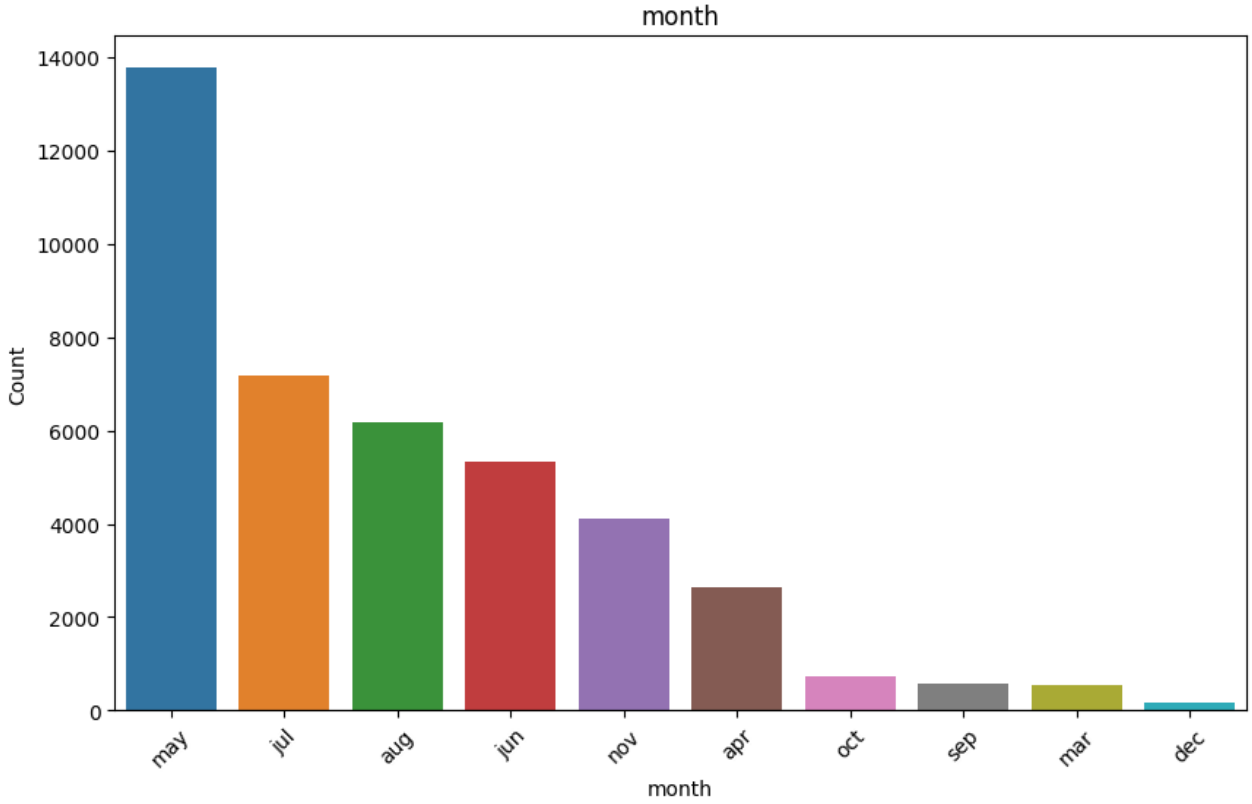

education
12000
10000
8000
6000
4000
2000
0
Count
university.degree
high.school
basic.9y
professional.course
basic.4y
basic.6y
unknown
illiterate
education

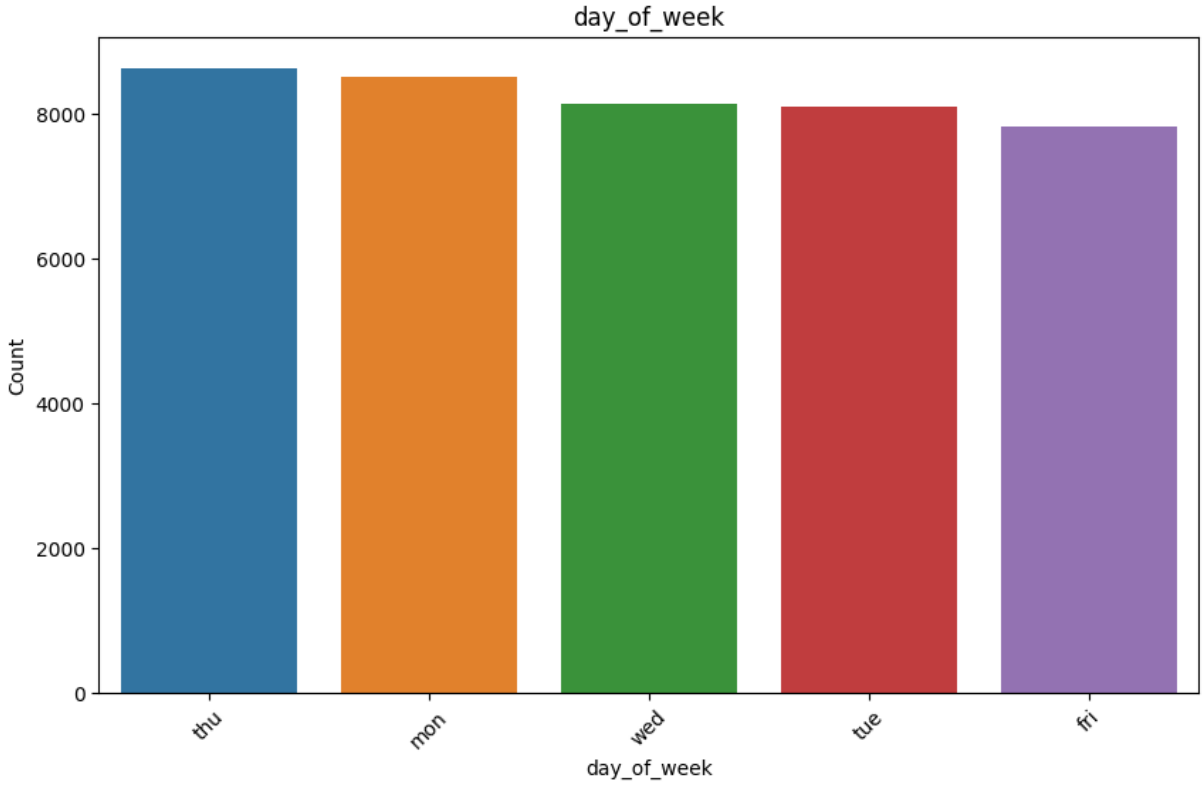

default
Count
30000
25000
20000
15000
10000
5000
0
no
unknown
yes

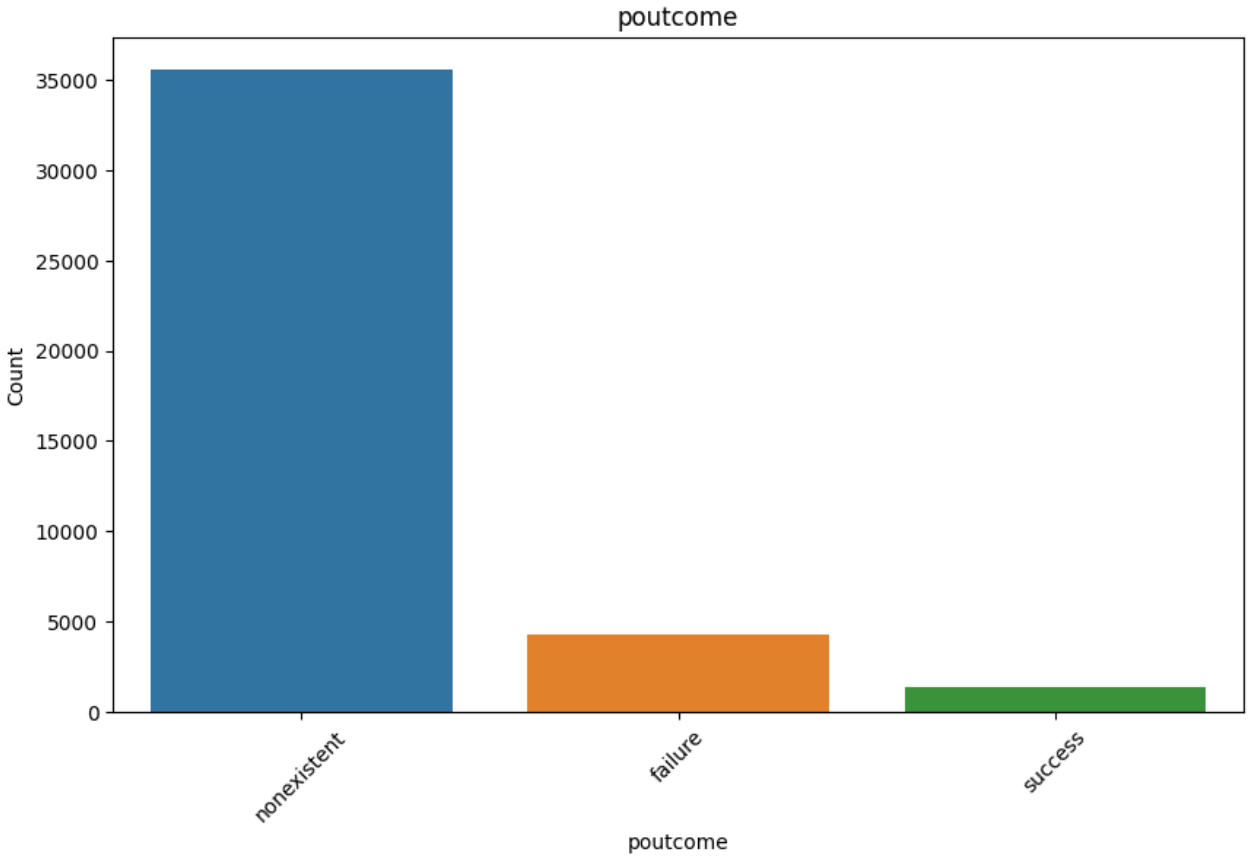

housing
Count
20000
15000
10000
5000
0
yes
no
unknown
housing

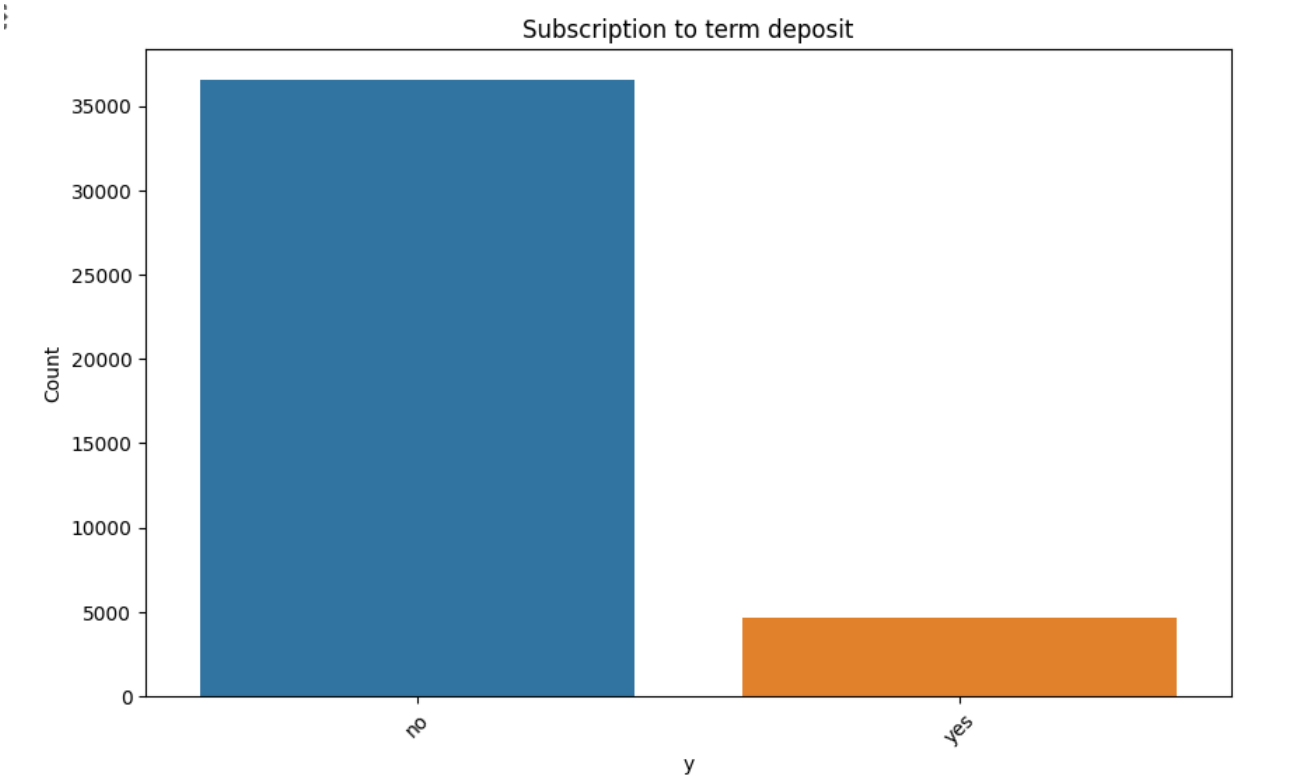

loan
35000
30000
25000
20000
15000
10000
5000
0
Count
no
yes
unknown
loan


contact
25000
20000
15000
10000
5000
0
Count
cellular
telephone
contact

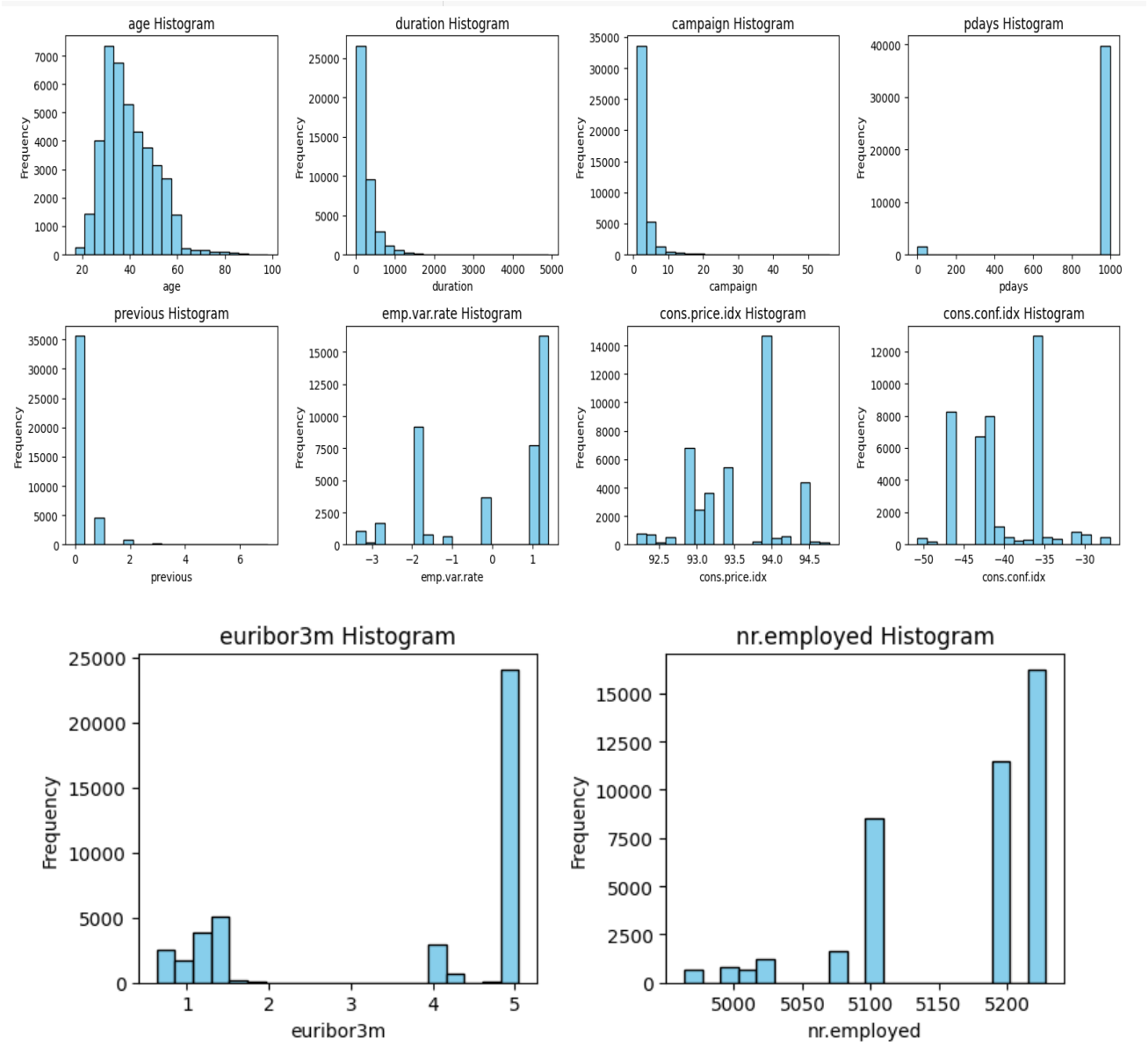

month
14000
12000
10000
8000
6000
4000
2000
0
Count
may
jul
aug
jun
nov
apr
oct
sep
mar
dec
month


day_of_week
8000
6000
4000
2000
0
Count
thu
mon
wed
tue
fri
day_of_week





b) Numerical features:



Correlation heatmap:

2. Perform all the necessary Feature Engineering Techniques.

a) Missing Value Treatment

b) Label Encoding

c) Imbalanced Data Handling

Code for the given problem:

```
#2. Feature Engineering
# a) Missing Value Treatment
data.dropna(inplace=True)
```

```
[68]  #b) Label Encoding
label_encoder = LabelEncoder()
categorical_columns = ['job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'poutcome']
for column in categorical_columns:
    data[column] = label_encoder.fit_transform(data[column])
```

```
[69] # c) Imbalanced Data Handling (using SMOTE)
X = data.drop(columns=['y'])
y = data['y']
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X, y)
```

```
[70] #Missing values were handled by dropna method
#Categorical variables were encoded using label encoding.
#Imbalanced data was addressed using SMOTE(Synthetic Minority Over-sampling Technique), resulting in a balanced dataset.
```