Lead Score Case Study

Group Members:

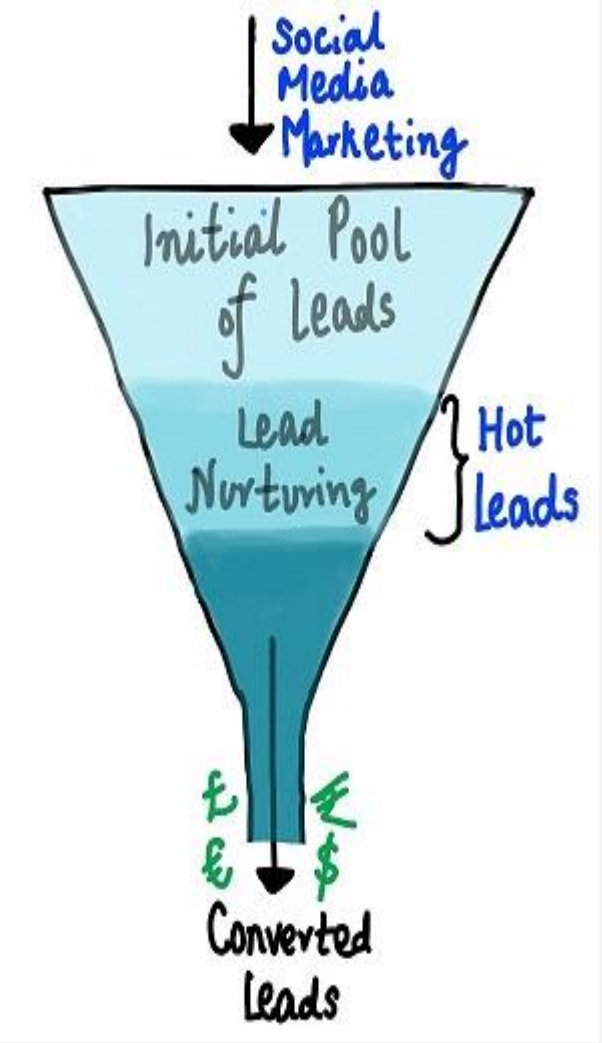
- Vijay Agrawal
- Shawaz Jahangiri

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.



Goals And Objectives

There are quite a few goals for this case study.

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Solution Approach

- **Read Data & Inspect data**
- **Cleaning Data**
    1) **Replacing 'Select' with NaN (Since it means no option is selected)**
    2) **Dropping columns with one unique value as it won't affect our analysis**
    3) **Dropping columns having high number of null values**
    4) **Categorization of column 'Country'**
    5) **Drop rows having null values in any column.**
- **EDA(Exploratory data analysis)**

1) **Univariate Analysis**
    a) **Univariate Analysis for Categorical Variables**
    b) **Univariate Analysis for Numerical Variables**
    c) **Relation between categorical variables to Converted**
    d) **Correlation among variables**
    e) **Outliers handling**

Solution Approach(Cnotd.)

- **Dummy Variables**
- **Test-Train Split**
- **Model Building**
- **Prediction**
- **Model Evaluation**
- **ROC Curve**

Solution Approach(Cnotd.)

- **Prediction on Test set**
- **Precision-Recall**
- **Precision and recall tradeoff**
- **Conclusion**

Data Cleaning

1. While submitting detail by the customer/visitors, he may not select any value in particular field. In that condition it contains 'Select'. Which is not the right value. So we changed it to Null value.

2. Dropped columns which contains only one unique value. As it won't affect our analysis.

    1) Magazine
    2) Receive More Updates About Our Courses
    3) Update me on Supply Chain Content
    4) Get updates on DM Content
    5) I agree to pay the amount through cheque

Data Cleaning

3 Dropped columns having high number of null values

1) How did you hear about X Education
2) Tags
3) Lead Quality
4) Lead Profile
5) City
6) Asymmetrique Activity Index
7) Asymmetrique Profile Index
8) Asymmetrique Activity Score
9) Asymmetrique Profile Score

Data Cleaning

4 Categorization of column 'Country':

Out of all the visitors, 6492 visitors are from India, 2461 visitors didn't mention the country and others are from various countries.

Hence we created three categories for this column:

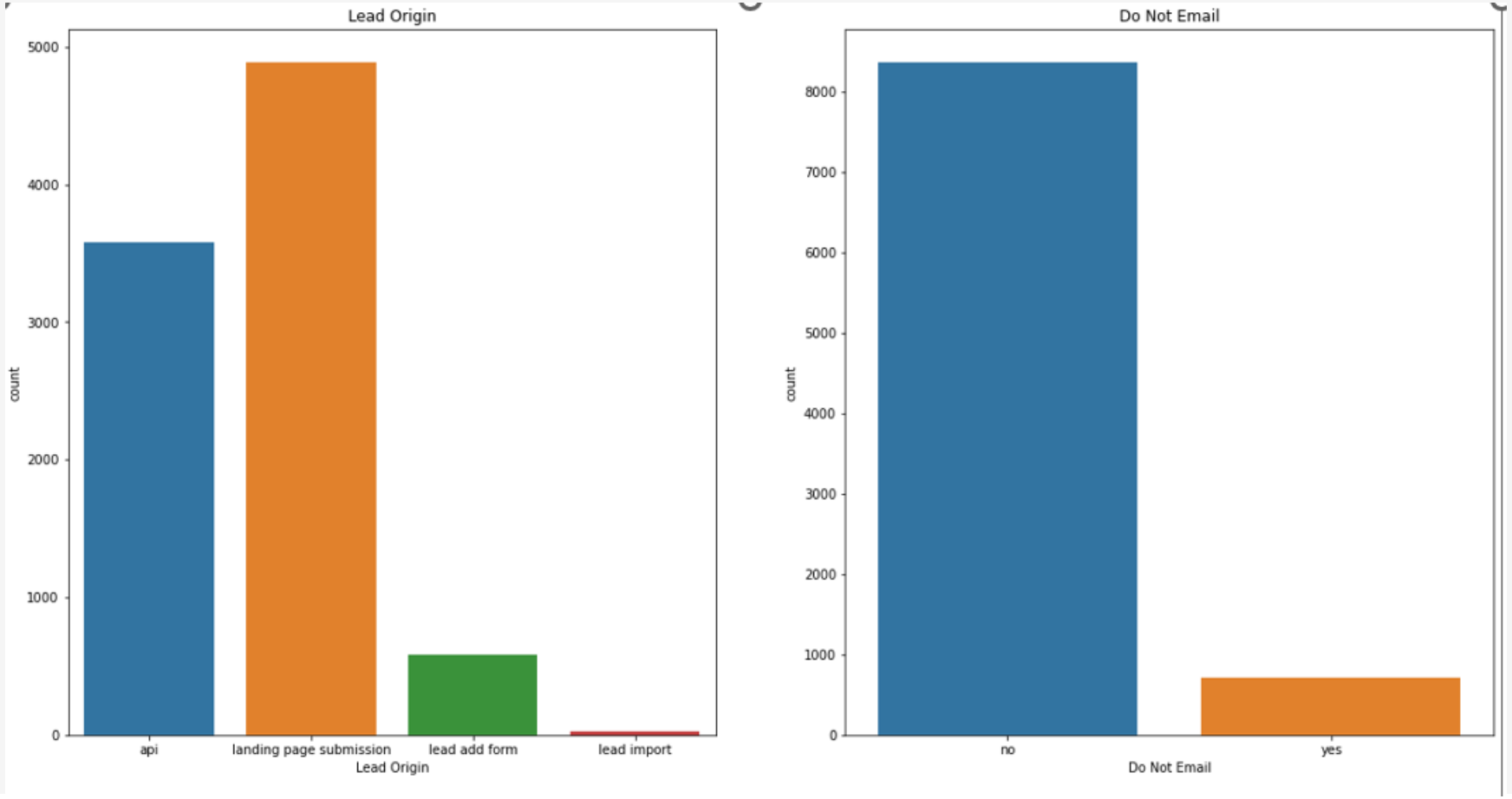
1) India
2) not provided
3) outside india

5 Drop rows having null values in any column.

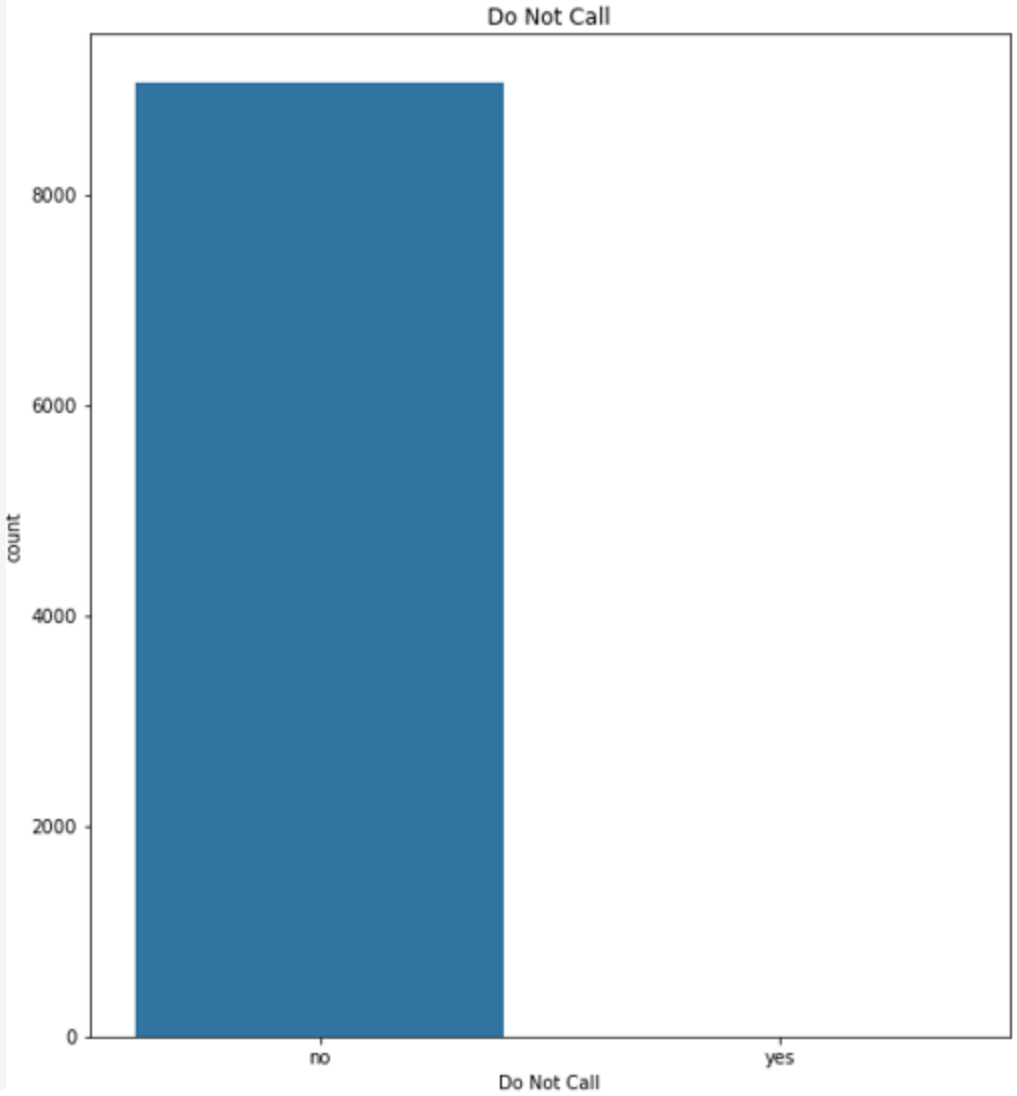
EDA(Exploratory data analysis)

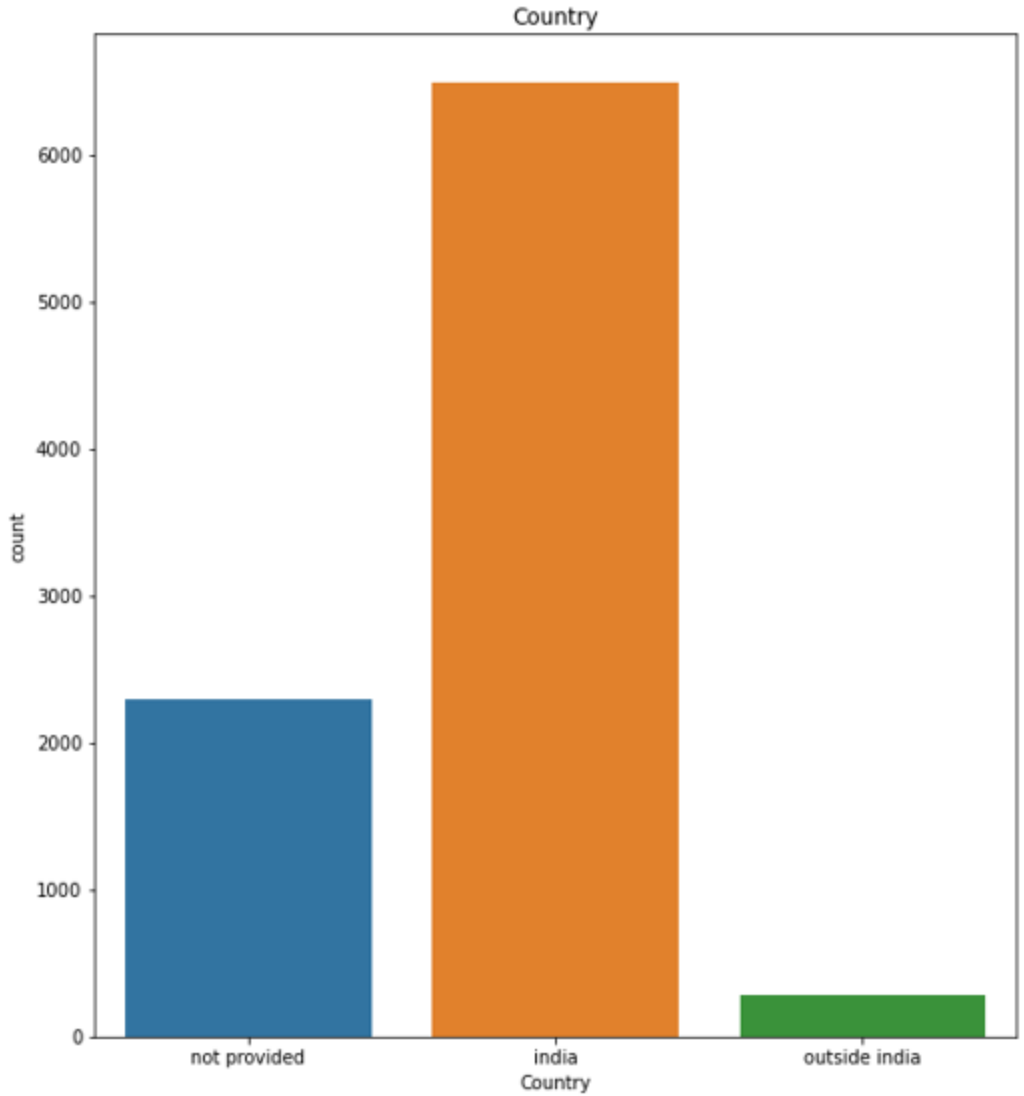
Univariate Analysis

a) **Univariate Analysis for Categorical Variables**

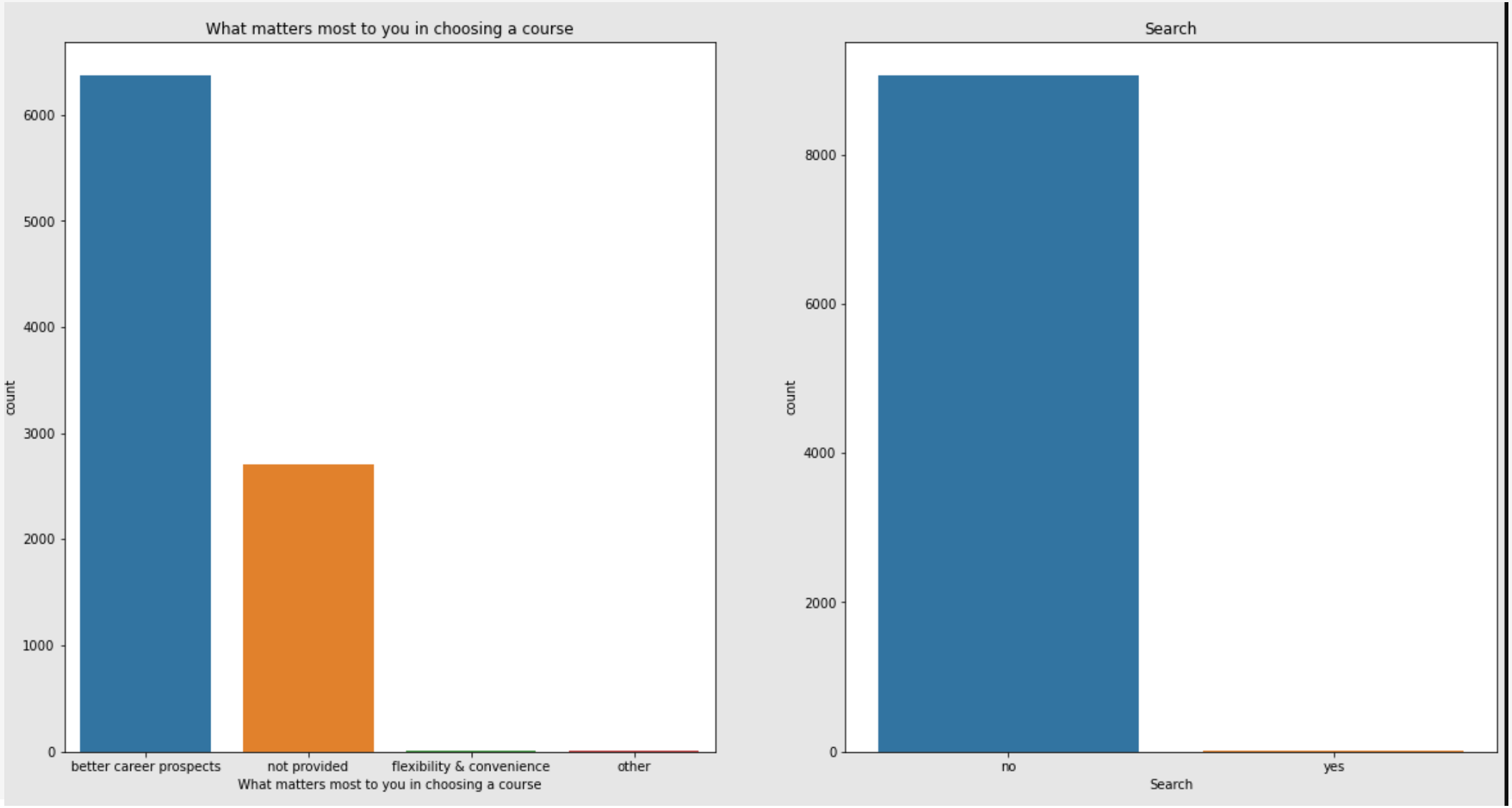


EDA(Exploratory data analysis)

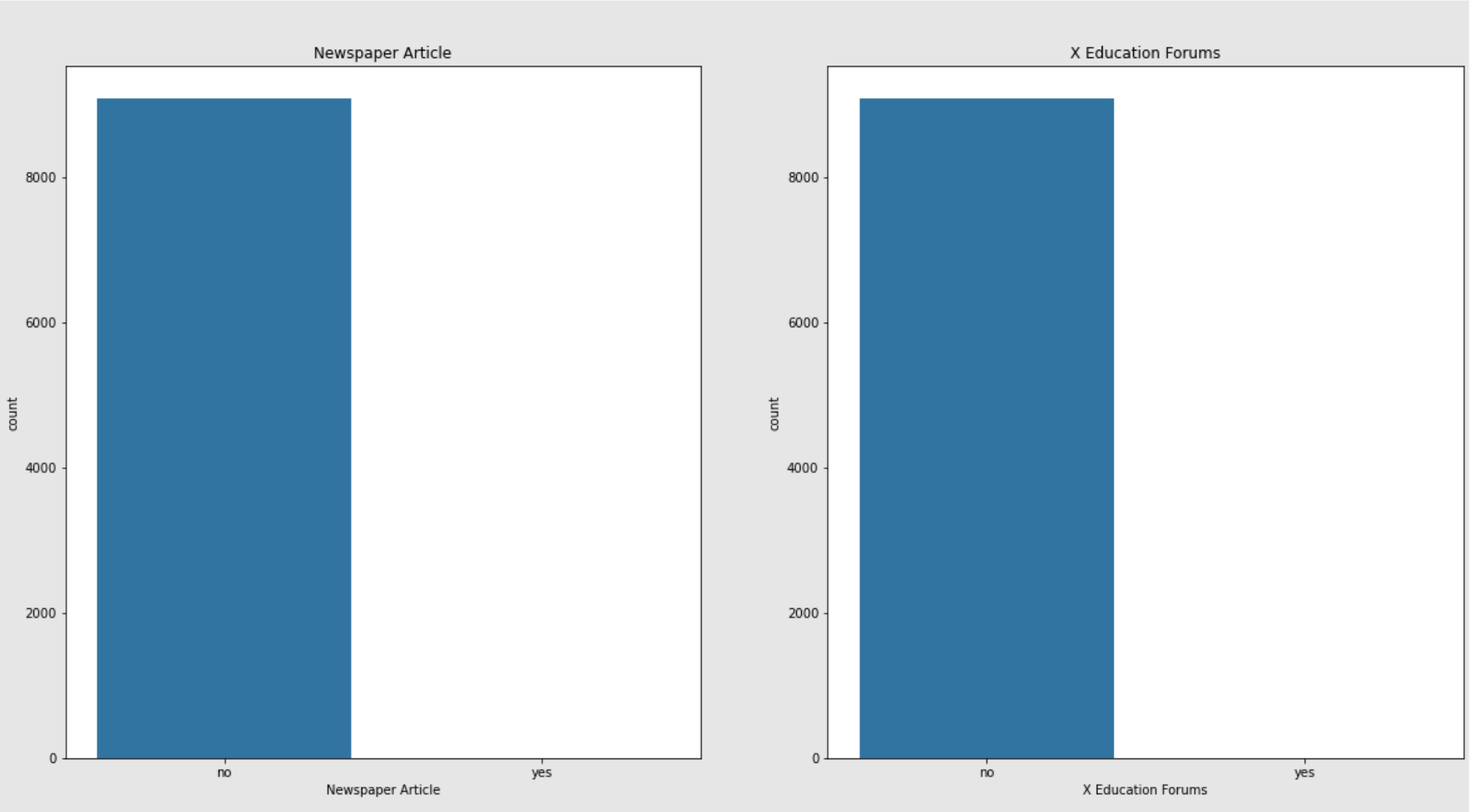




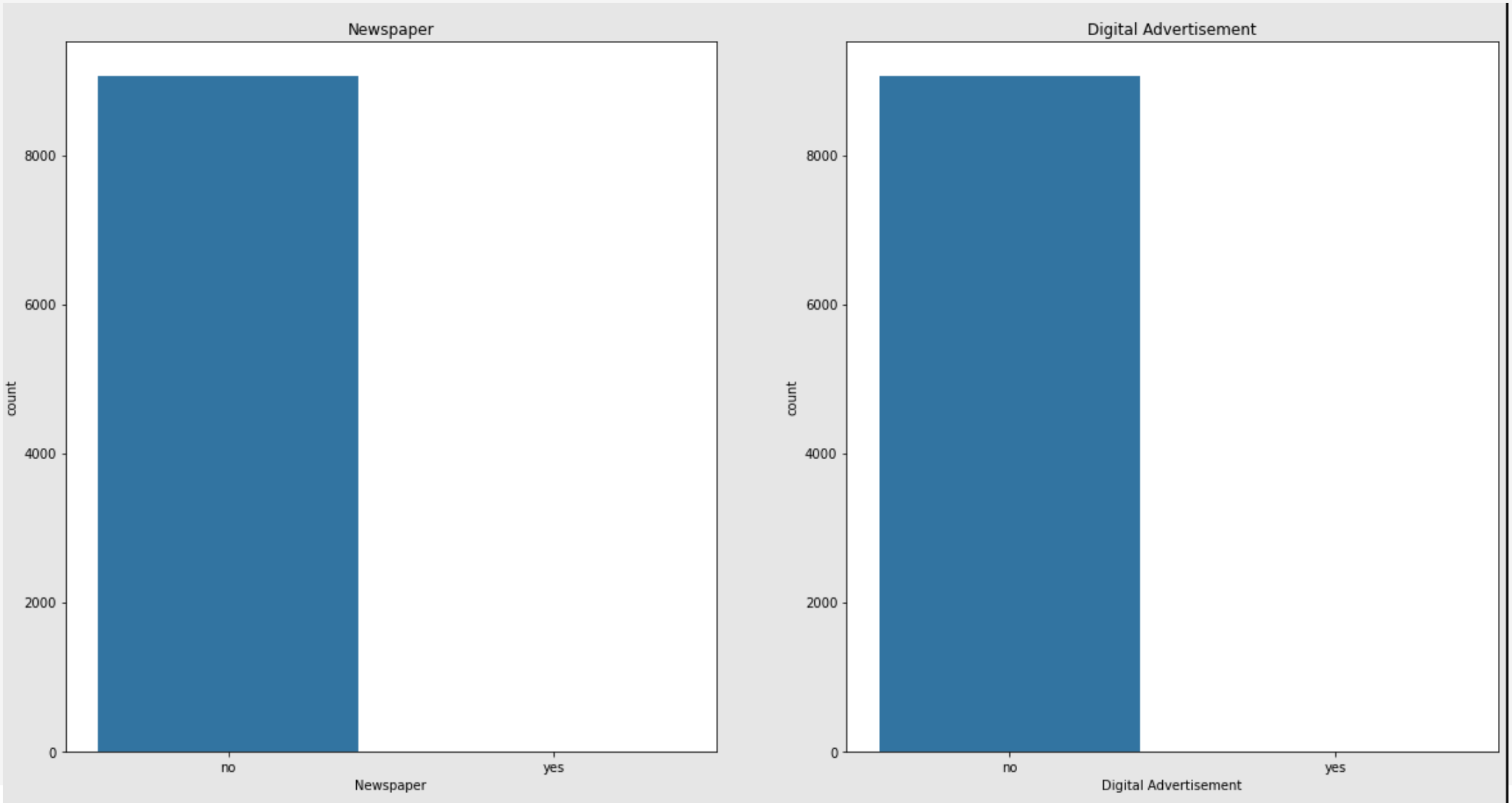
EDA(Exploratory data analysis)



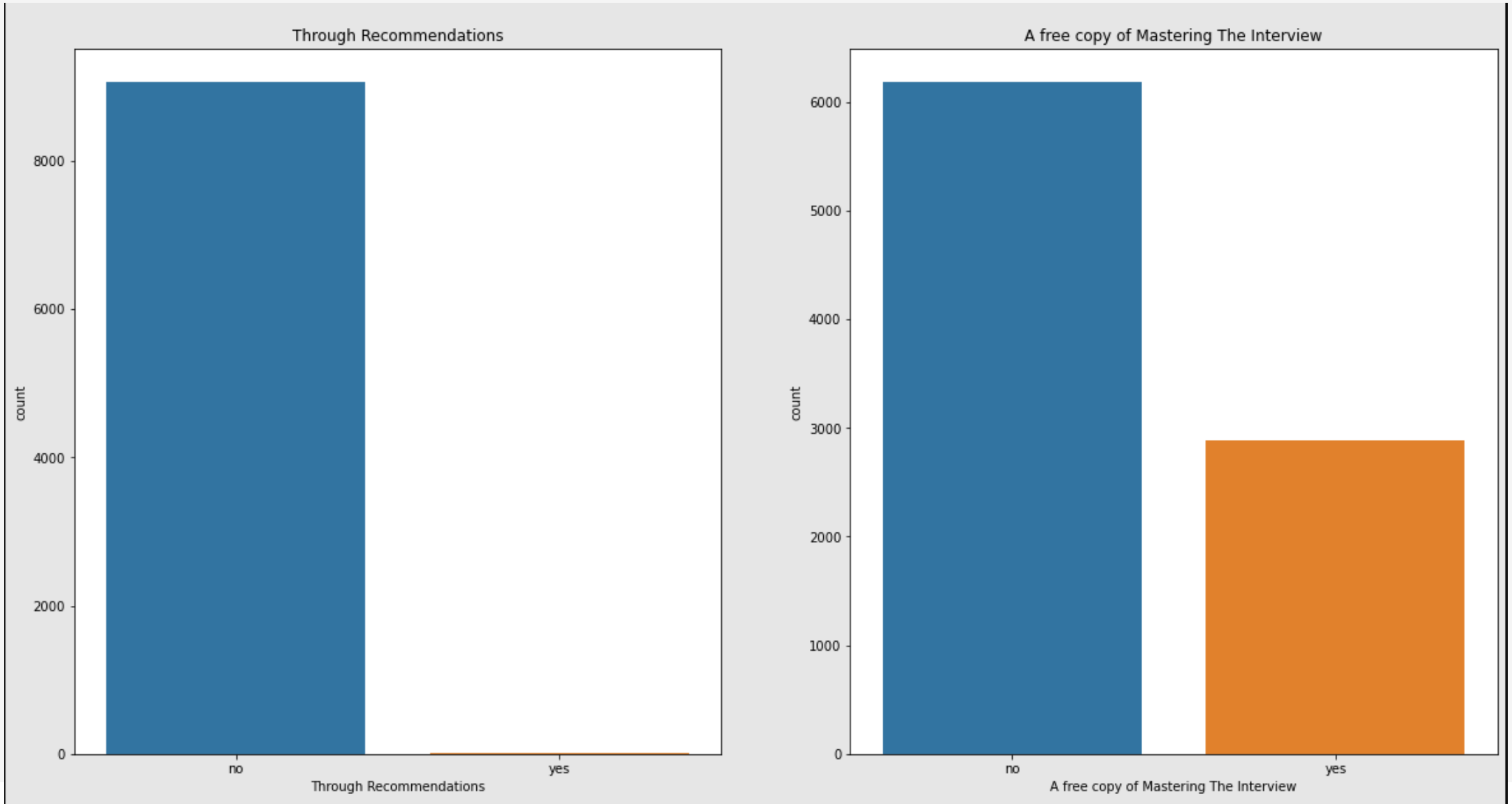
EDA(Exploratory data analysis)



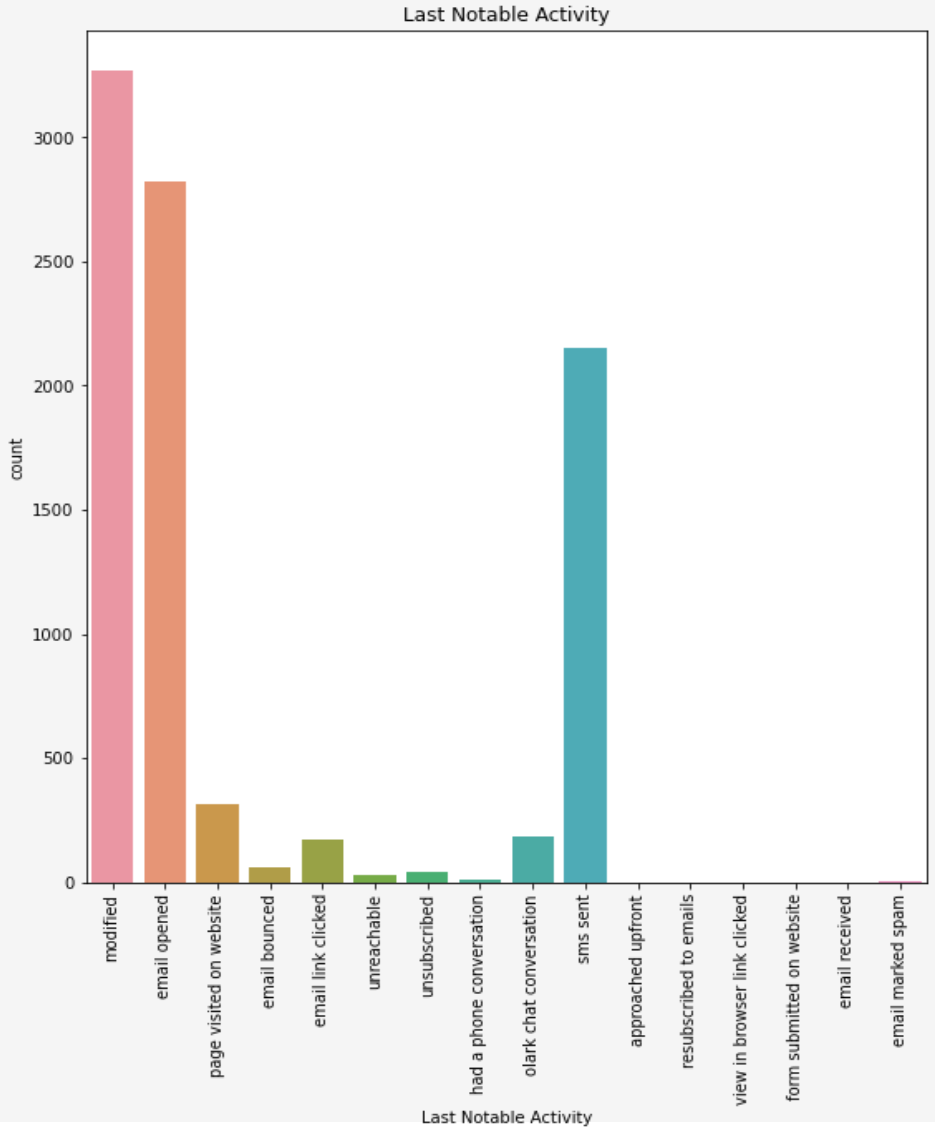
EDA(Exploratory data analysis)

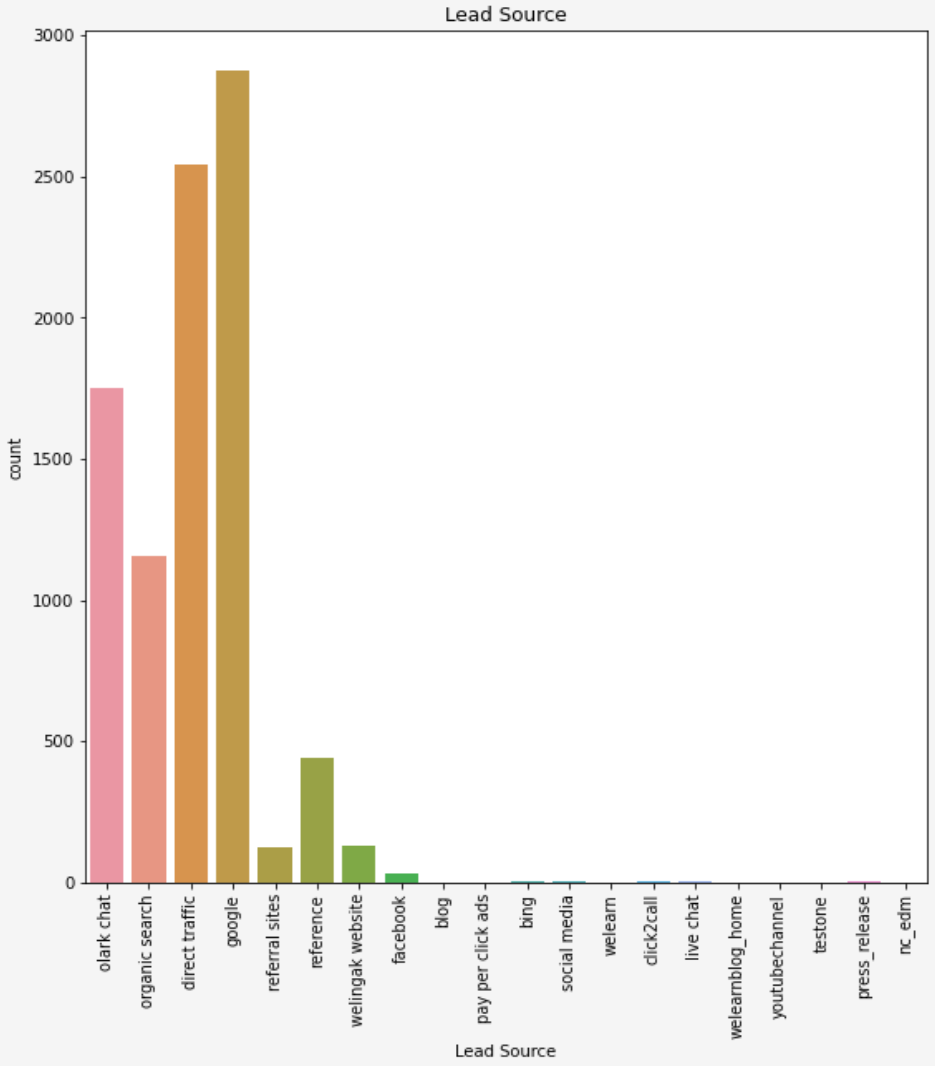


EDA(Exploratory data analysis)

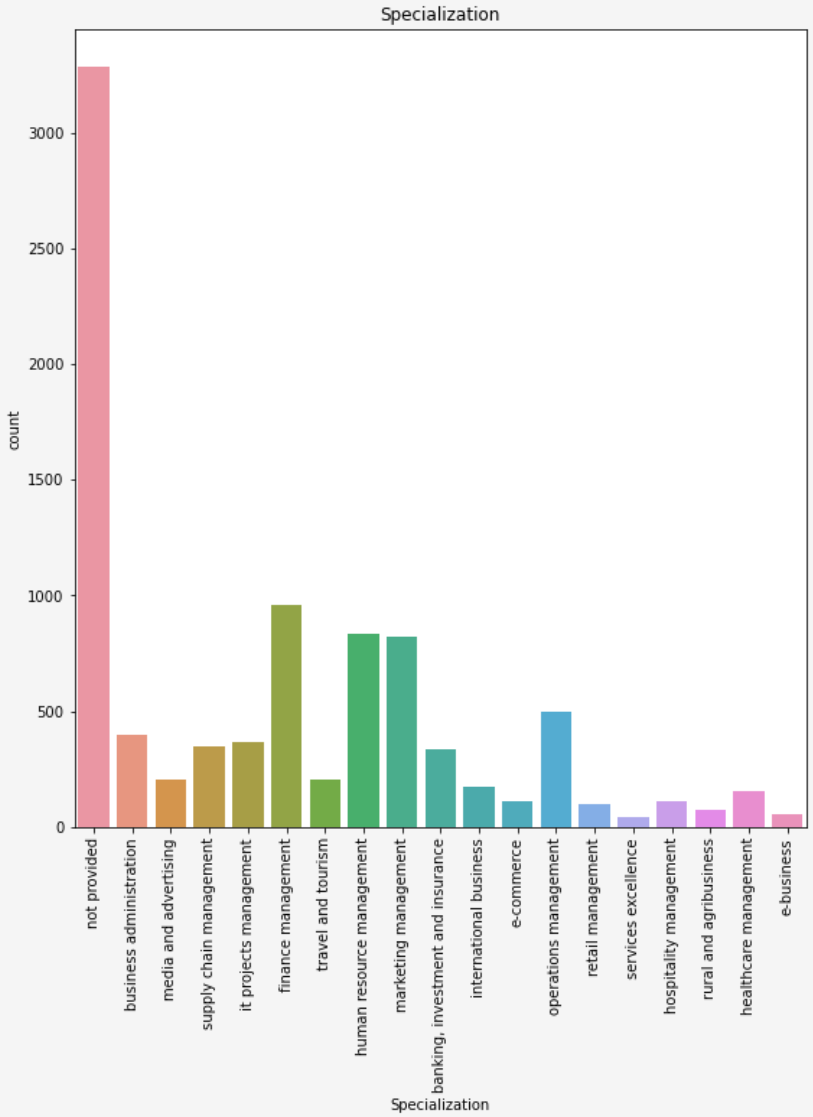


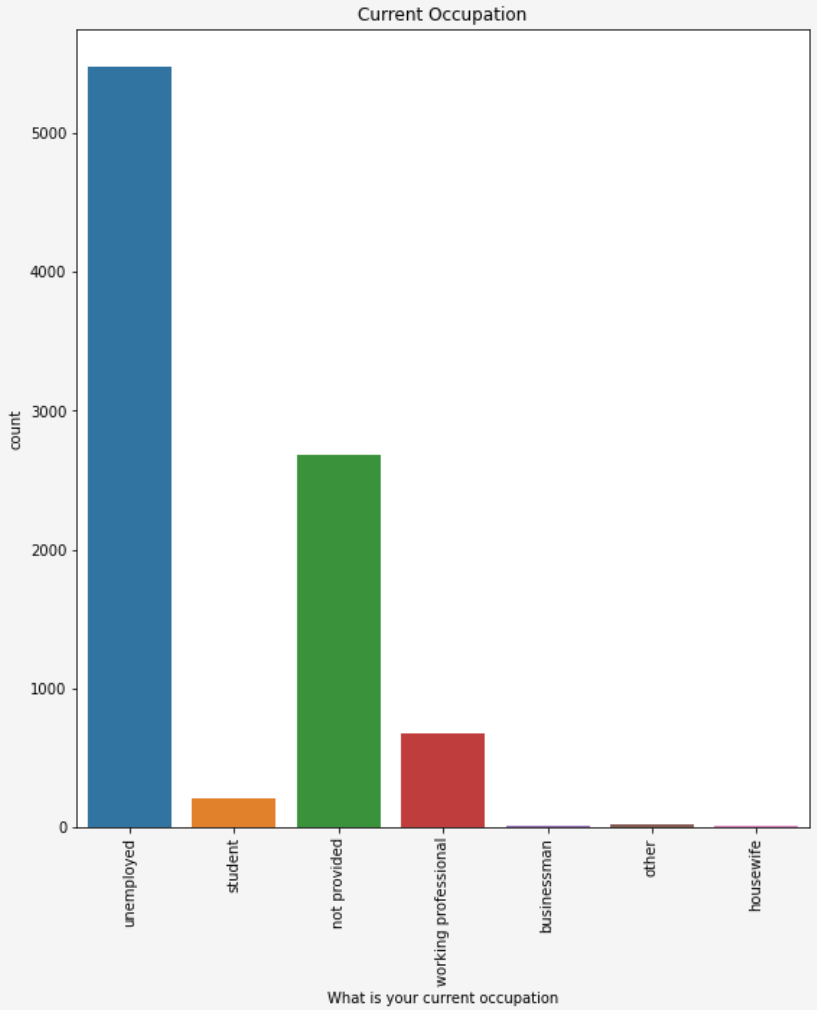
EDA(Exploratory data analysis)



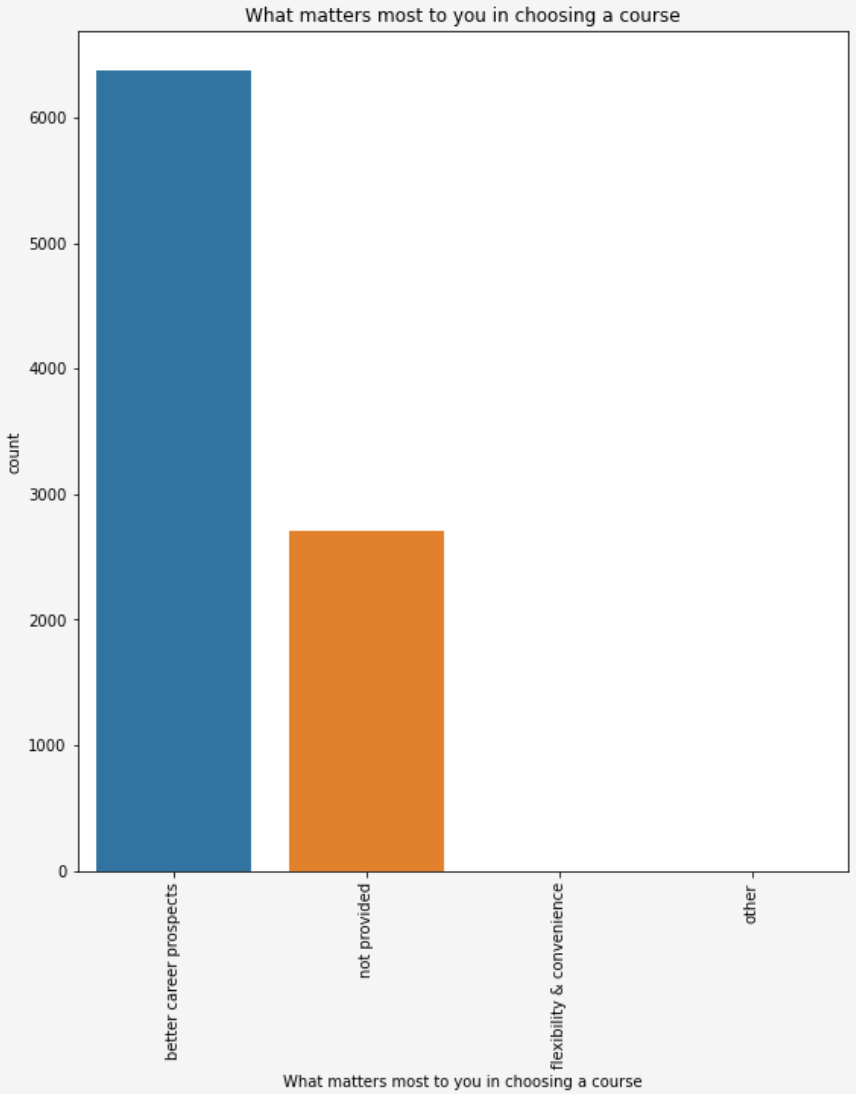


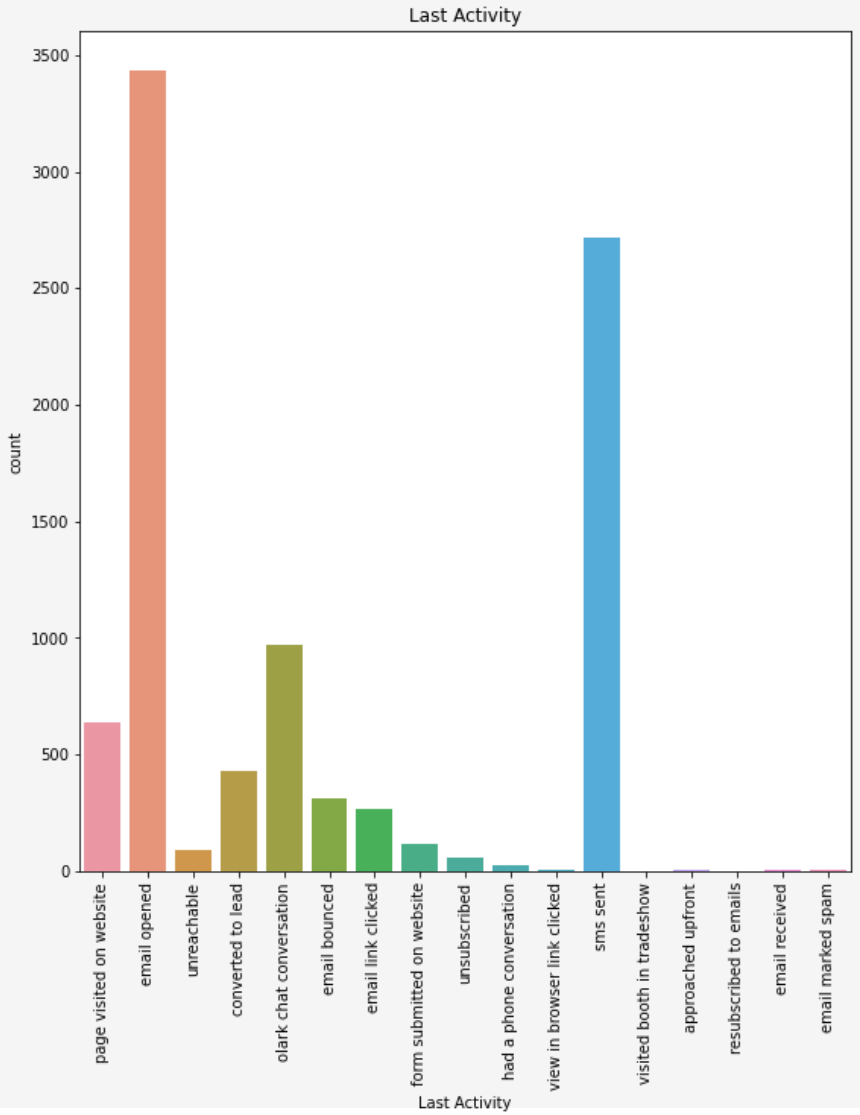
EDA(Exploratory data analysis)



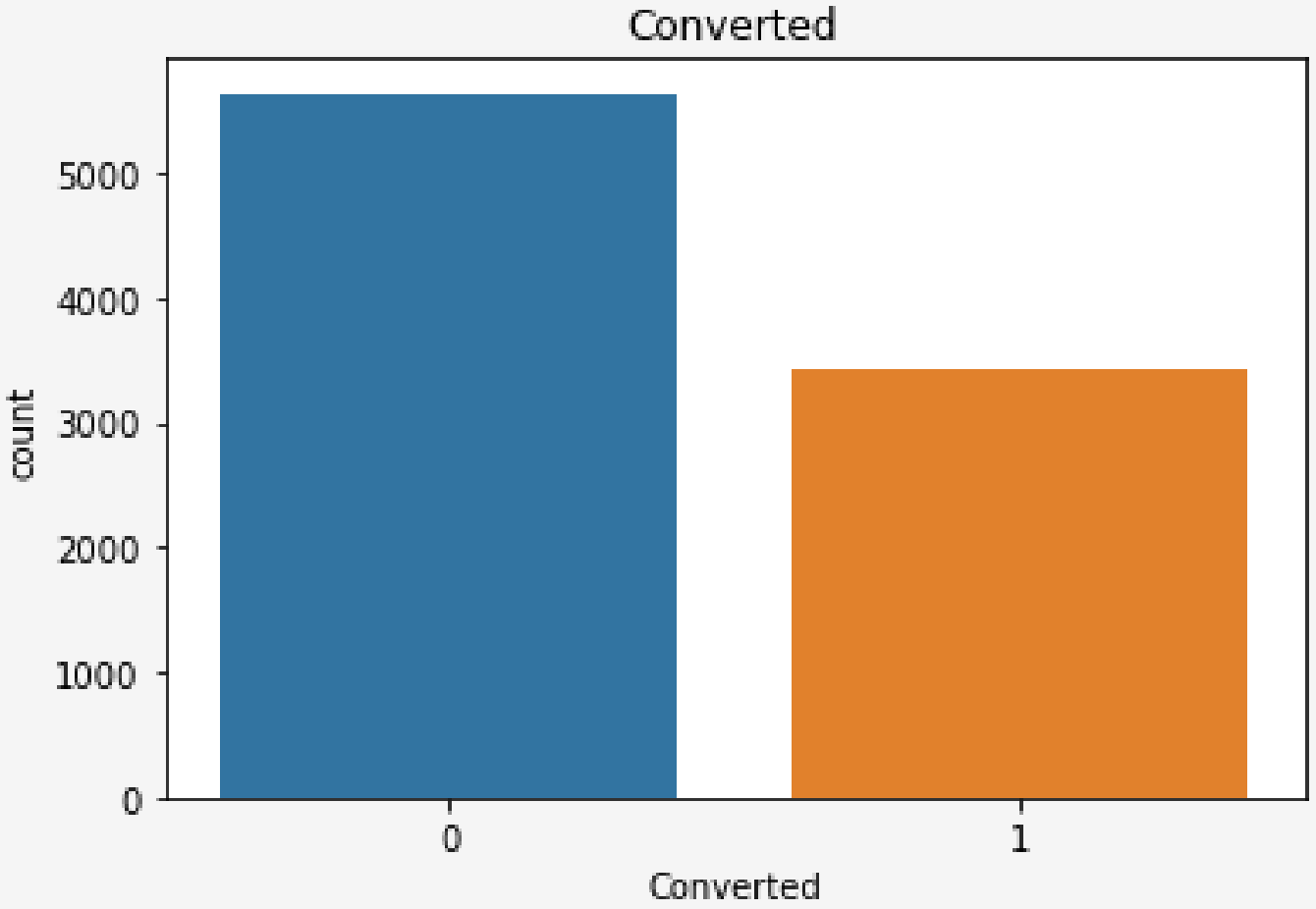


EDA(Exploratory data analysis)





EDA(Exploratory data analysis)

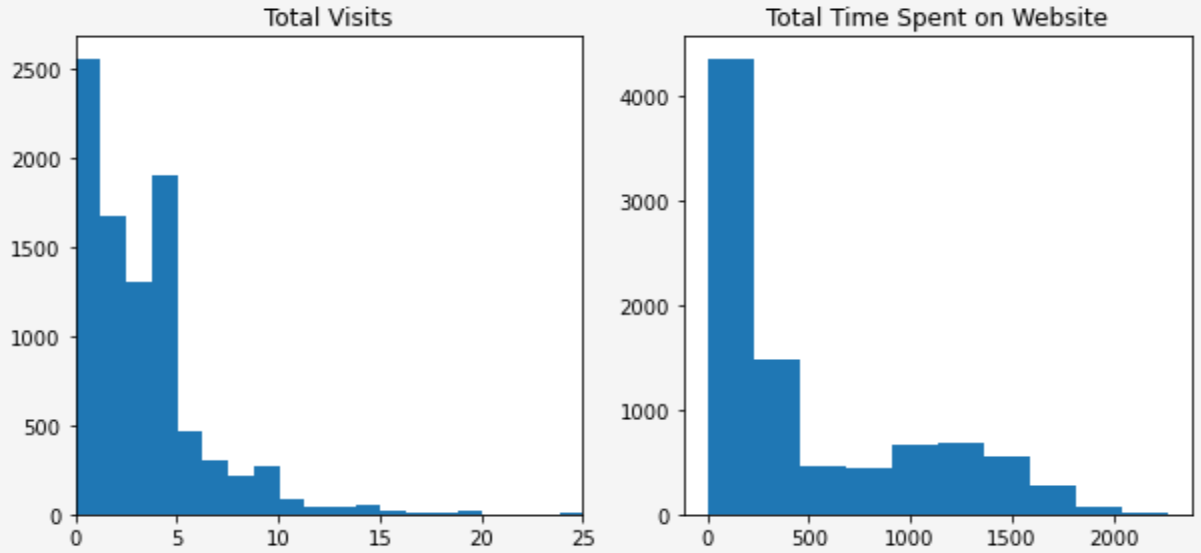


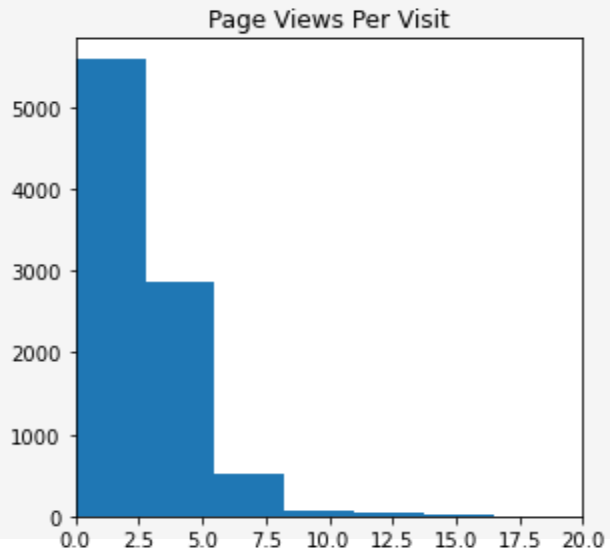
EDA(Exploratory data analysis)

Univariate Analysis

a) **Univariate Analysis for Categorical Variables**

b) **Univariate Analysis for Numerical Variables**

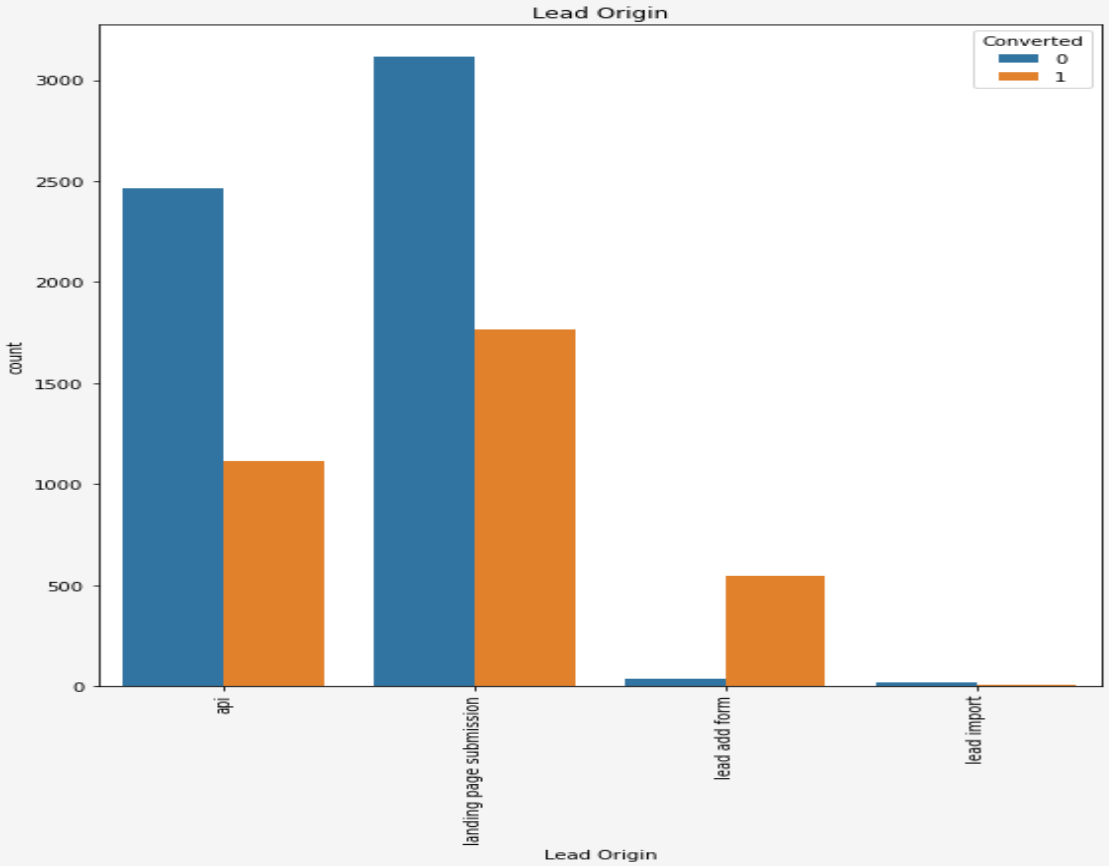


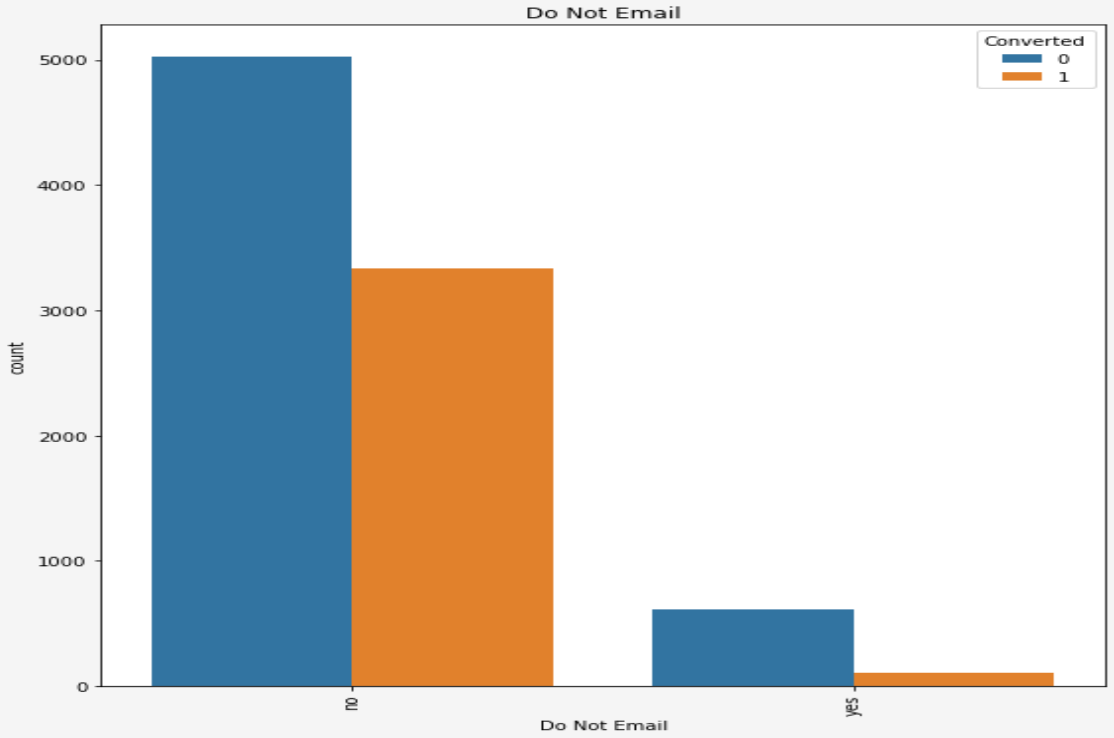


EDA(Exploratory data analysis)

Univariate Analysis

Relation between categorical variables to Converted

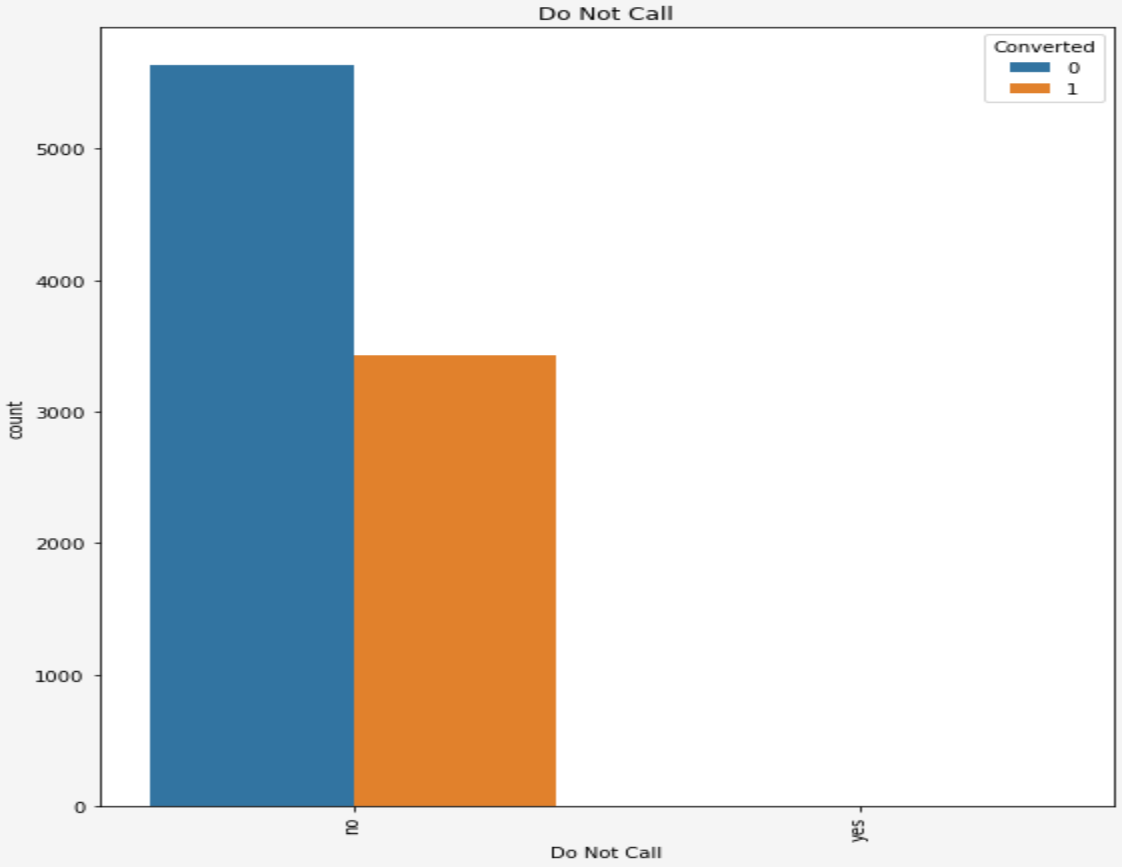


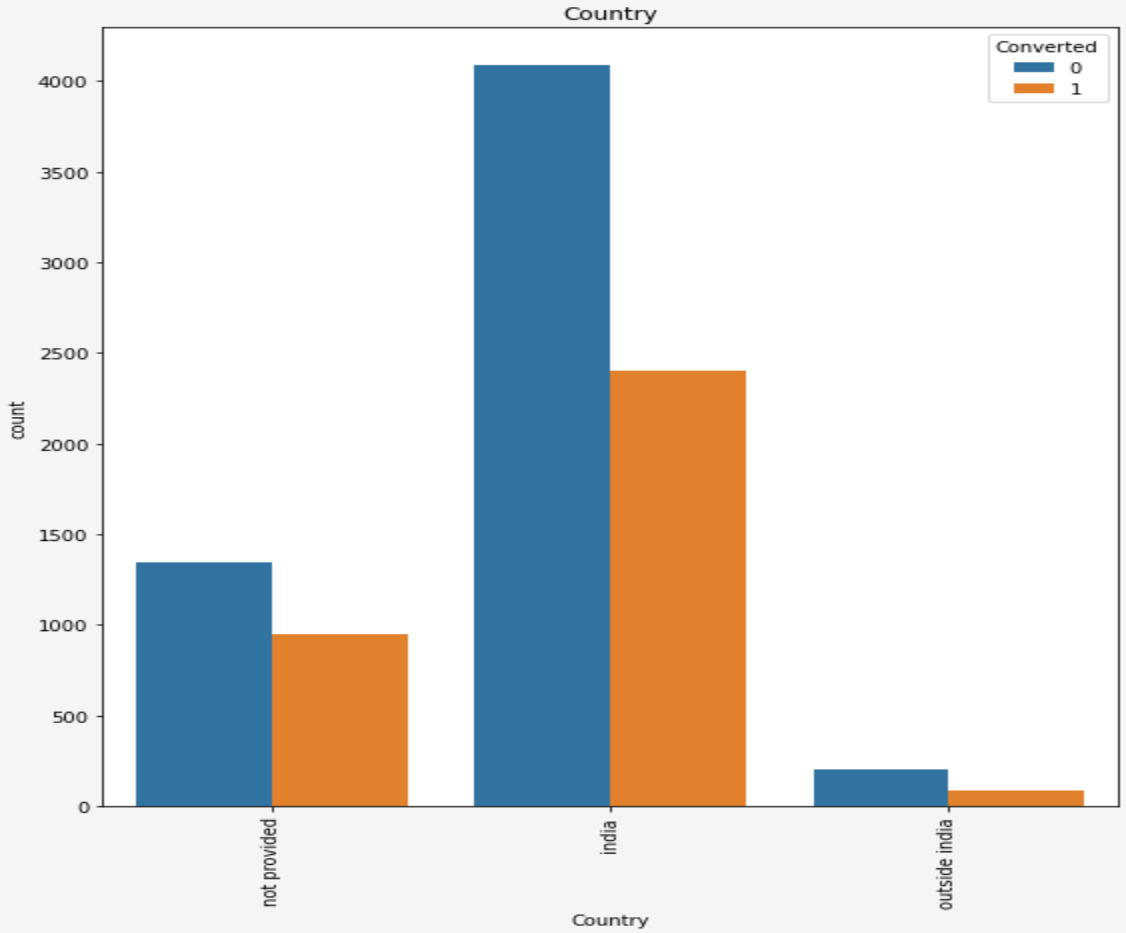


EDA(Exploratory data analysis)

Univariate Analysis

Relation between categorical variables to Converted

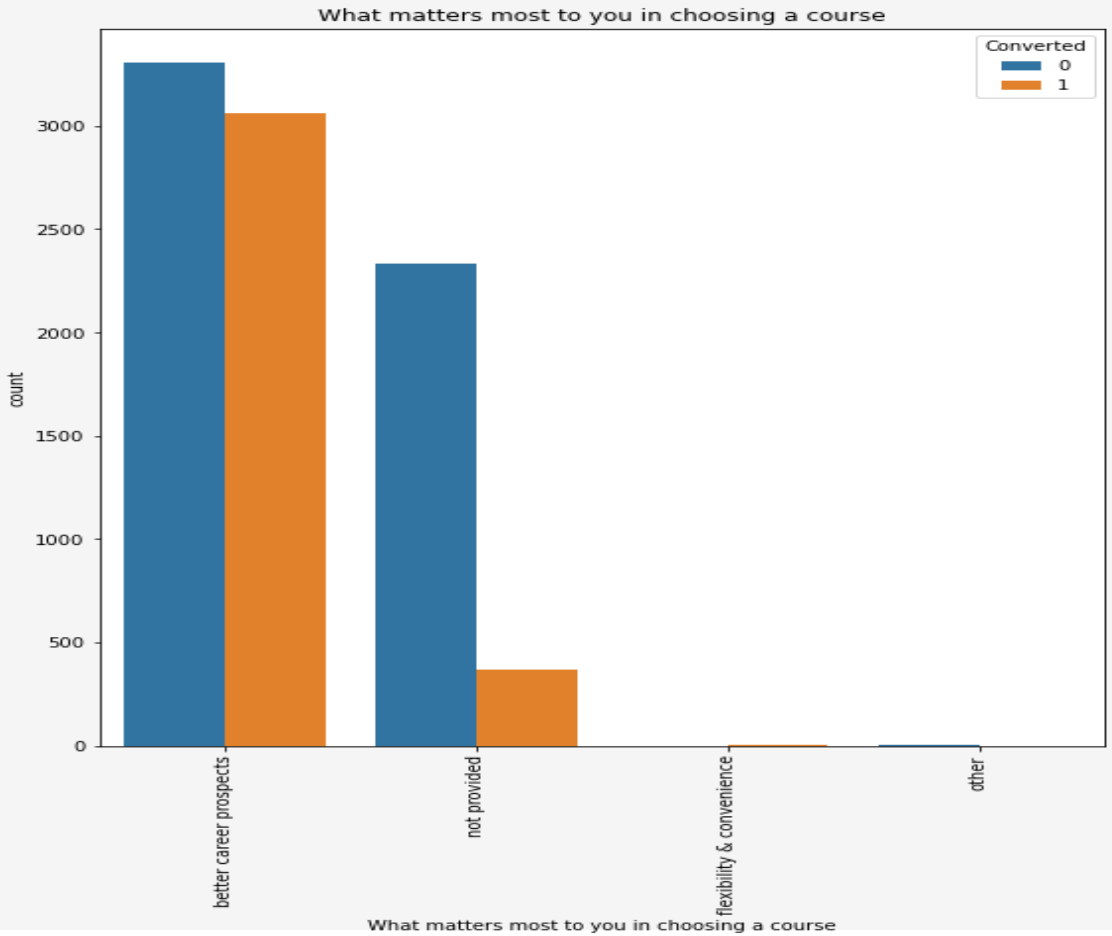


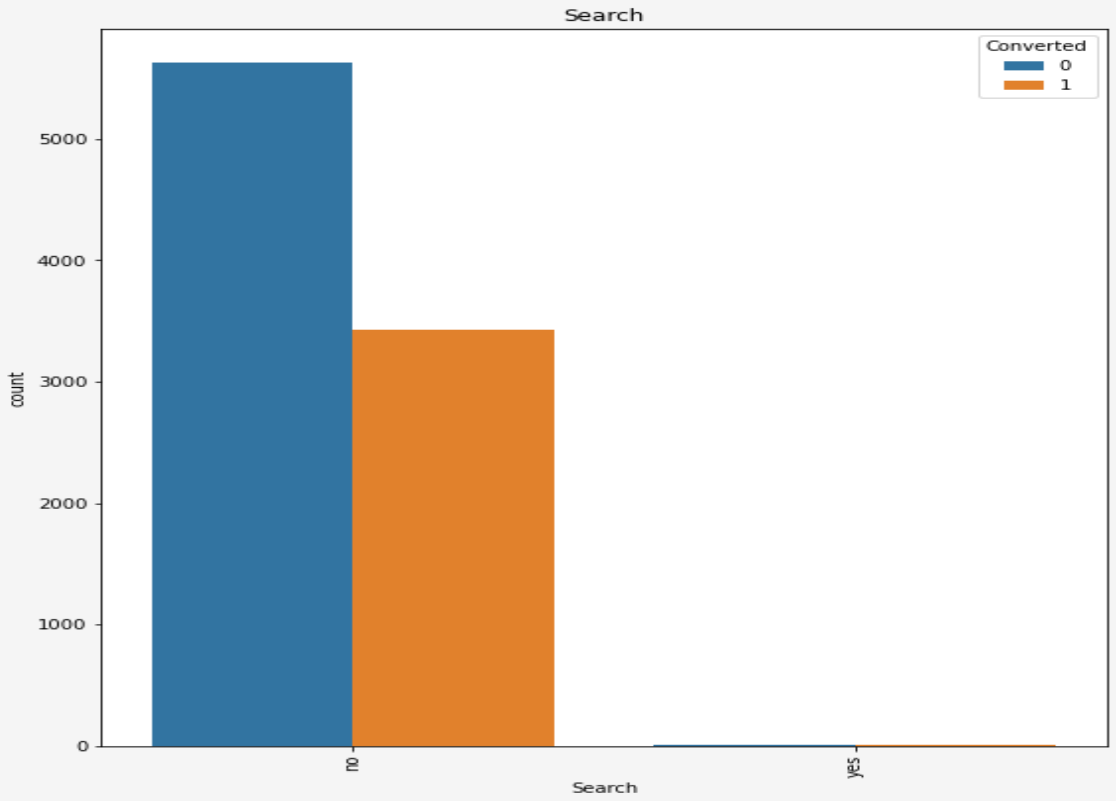


EDA(Exploratory data analysis)

Univariate Analysis

Relation between categorical variables to Converted

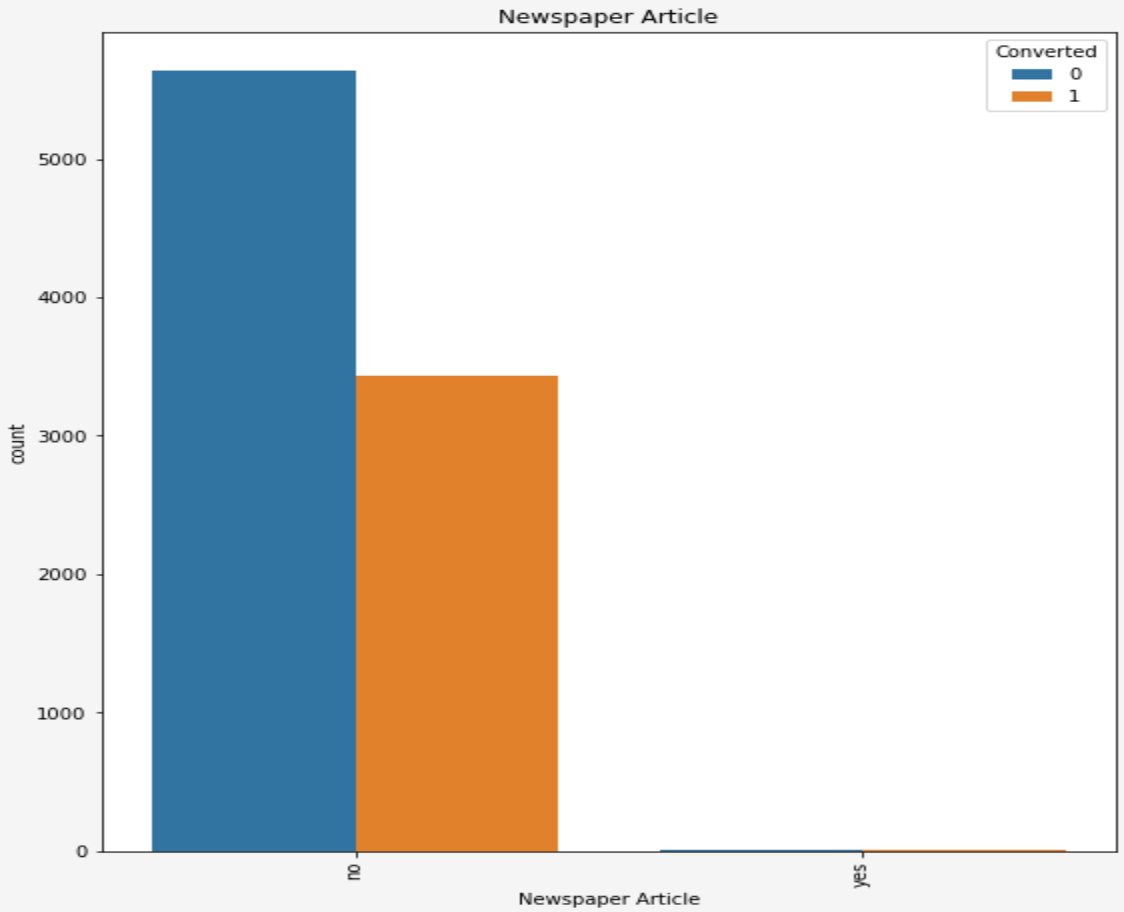


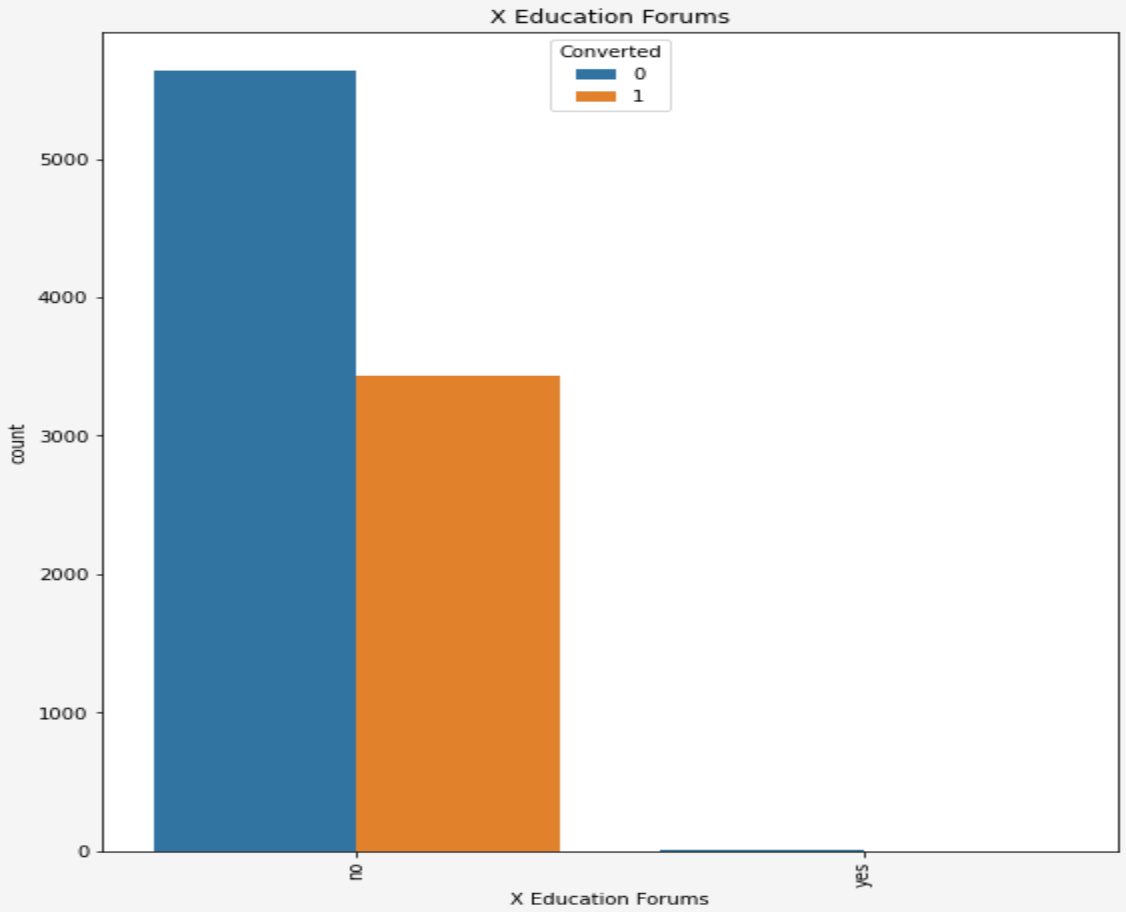


EDA(Exploratory data analysis)

Univariate Analysis

Relation between categorical variables to Converted

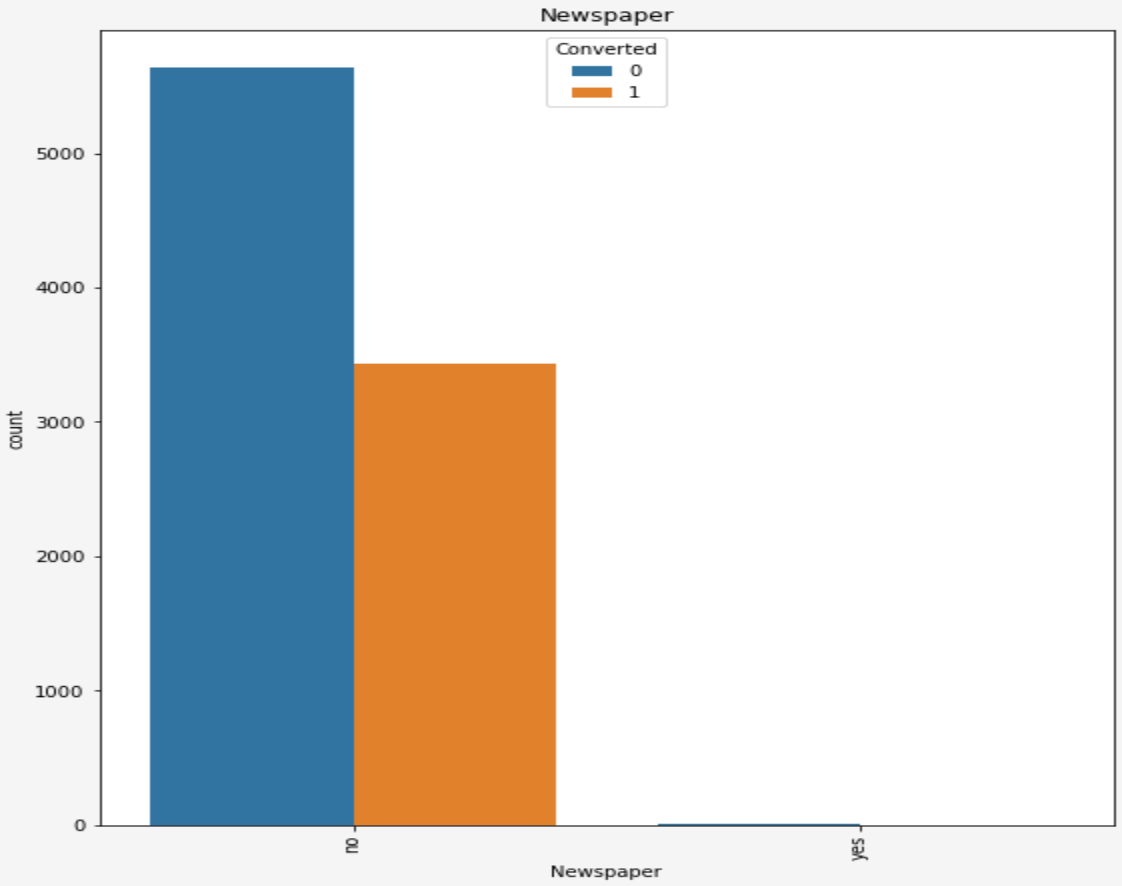


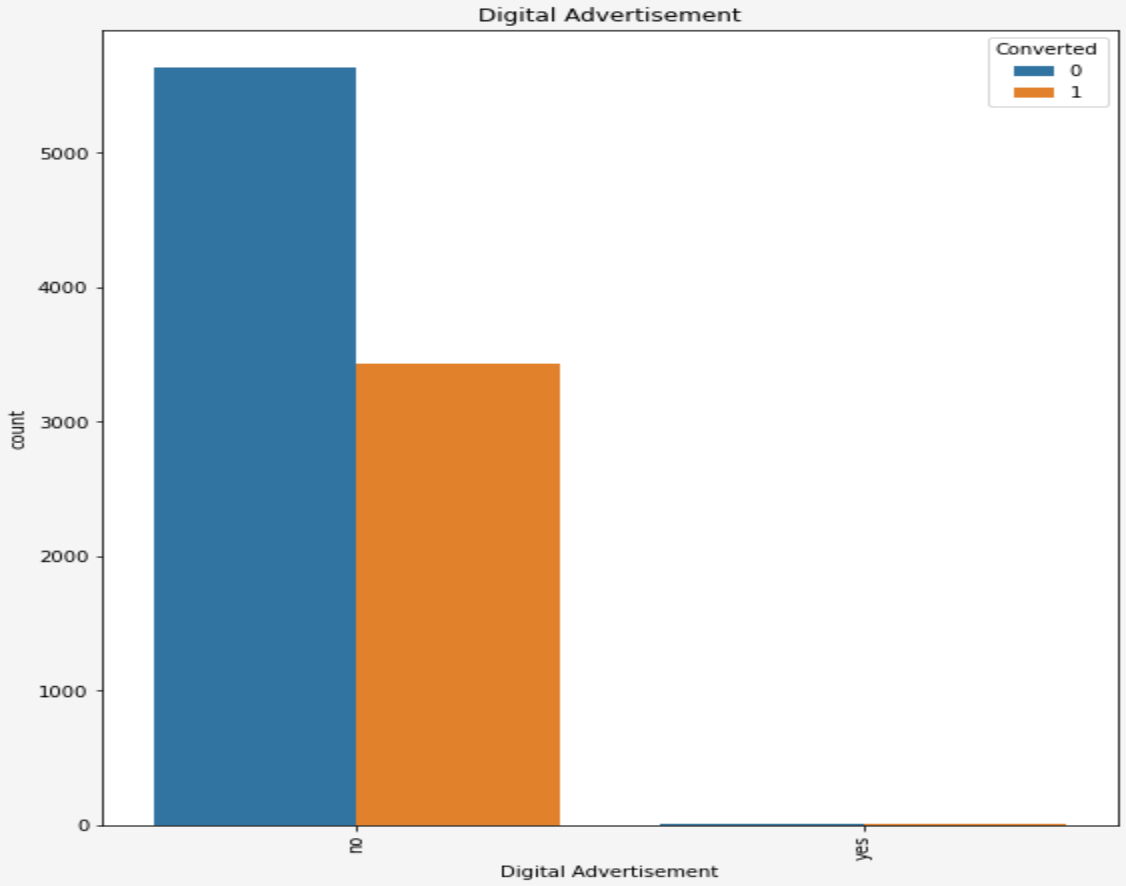


EDA(Exploratory data analysis)

Univariate Analysis

Relation between categorical variables to Converted

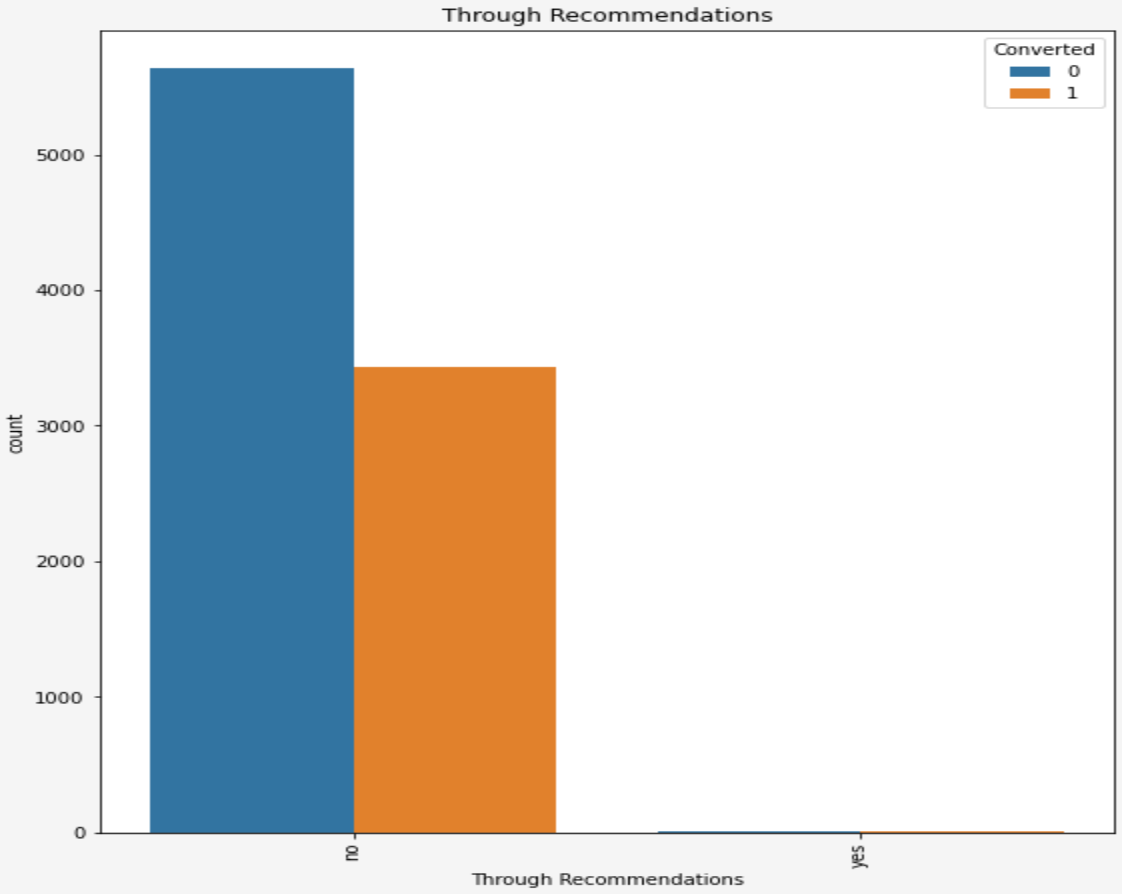


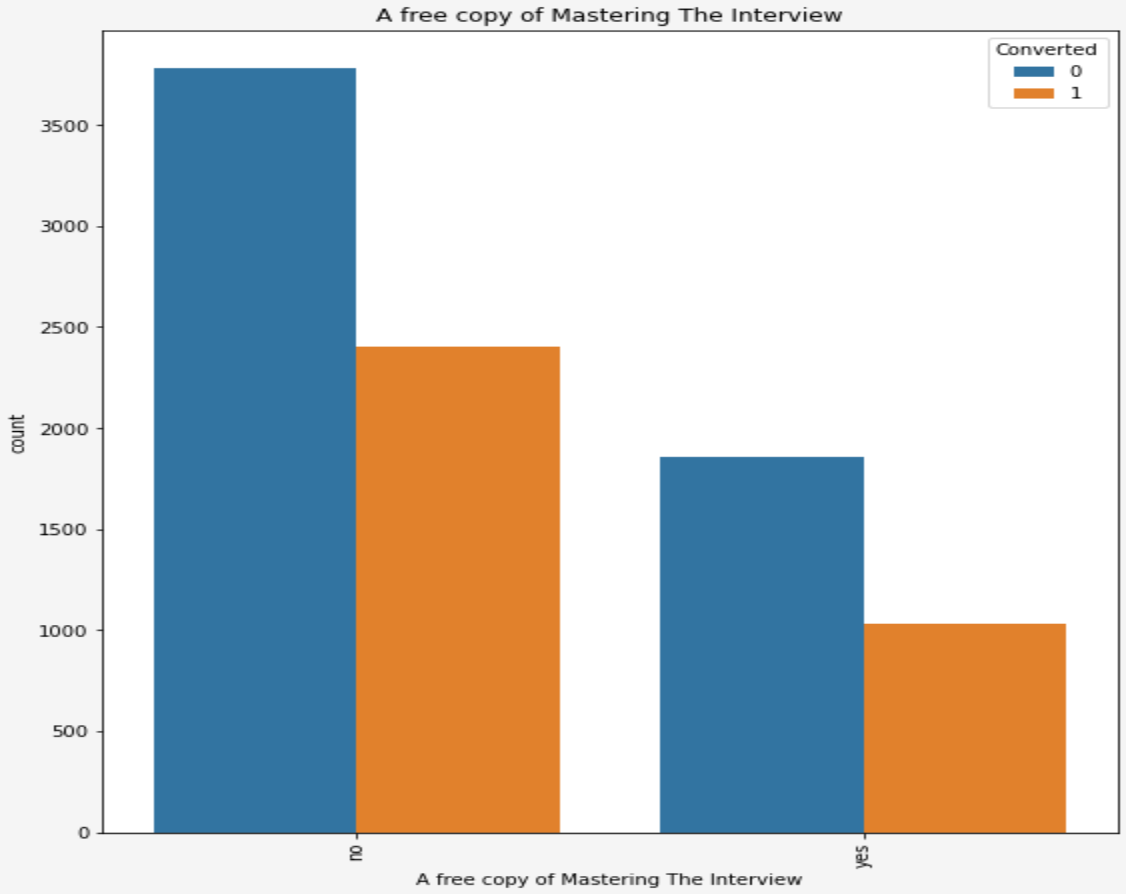


EDA(Exploratory data analysis)

Univariate Analysis

Relation between categorical variables to Converted

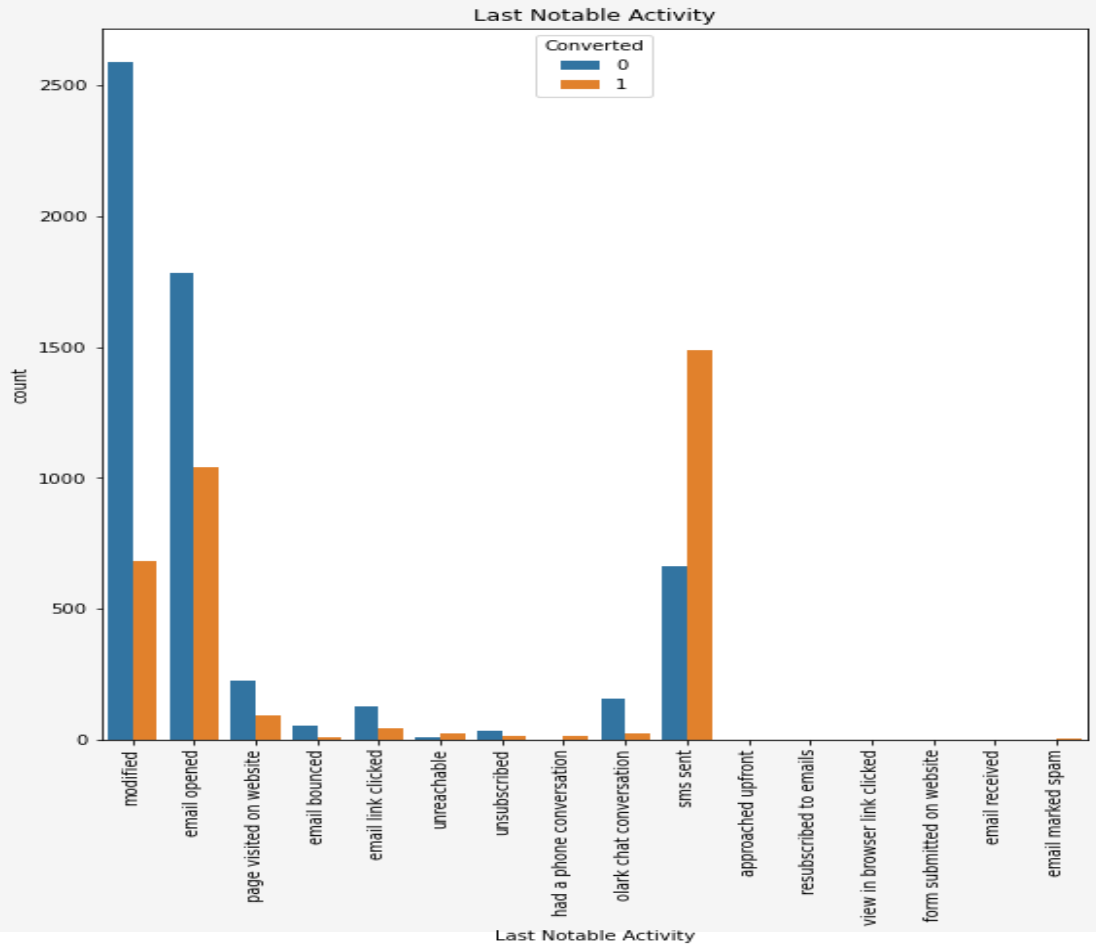


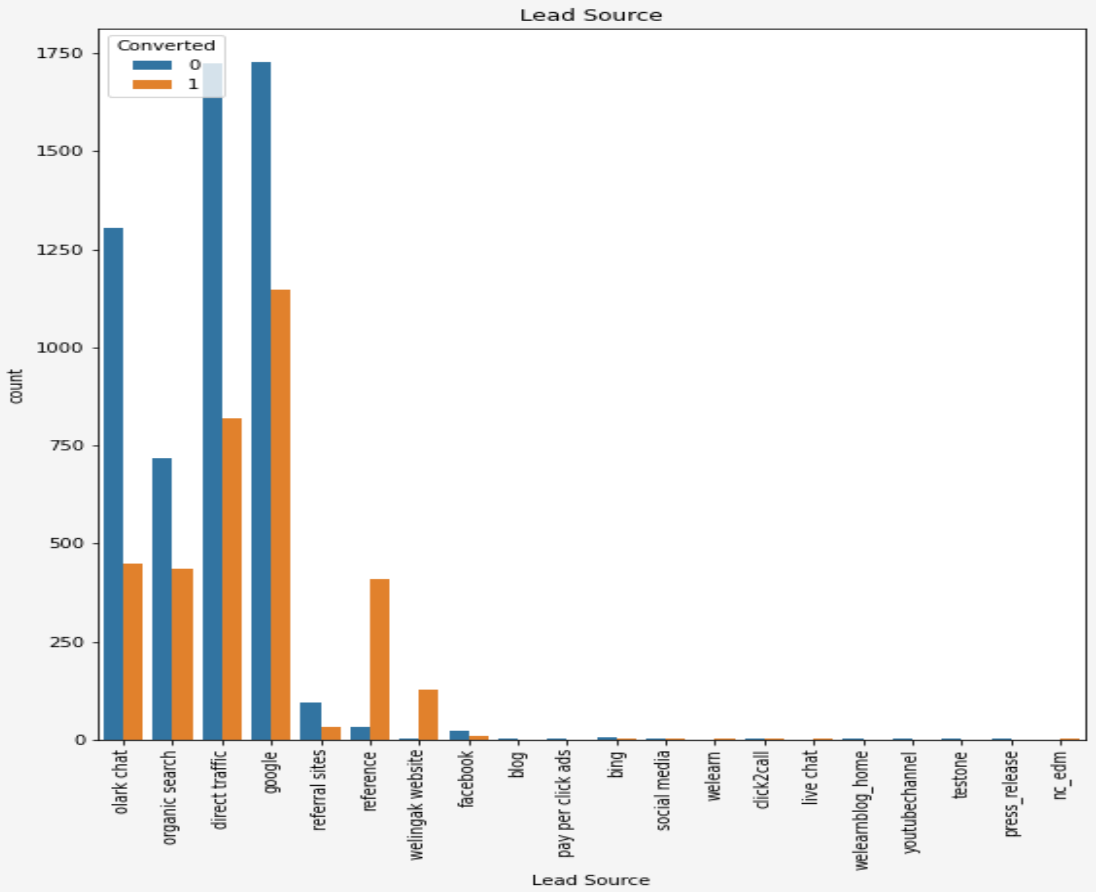


EDA(Exploratory data analysis)

Univariate Analysis

Relation between categorical variables to Converted

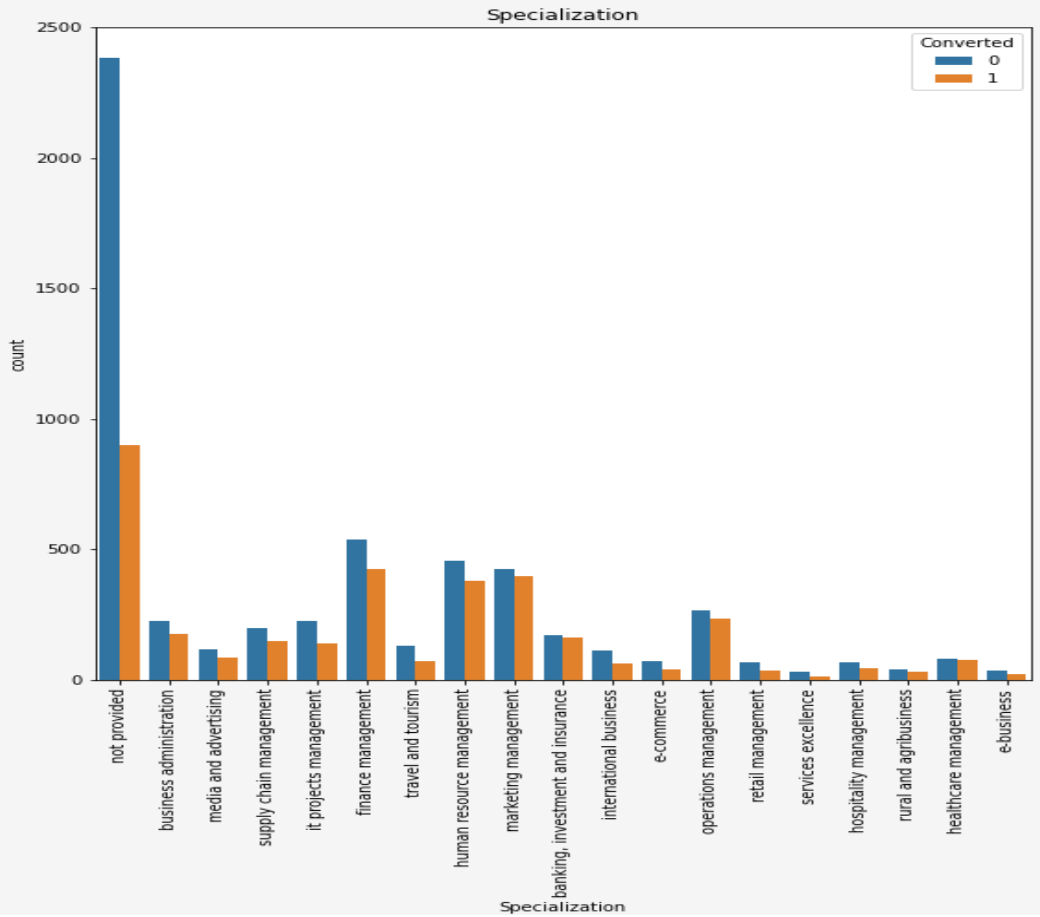


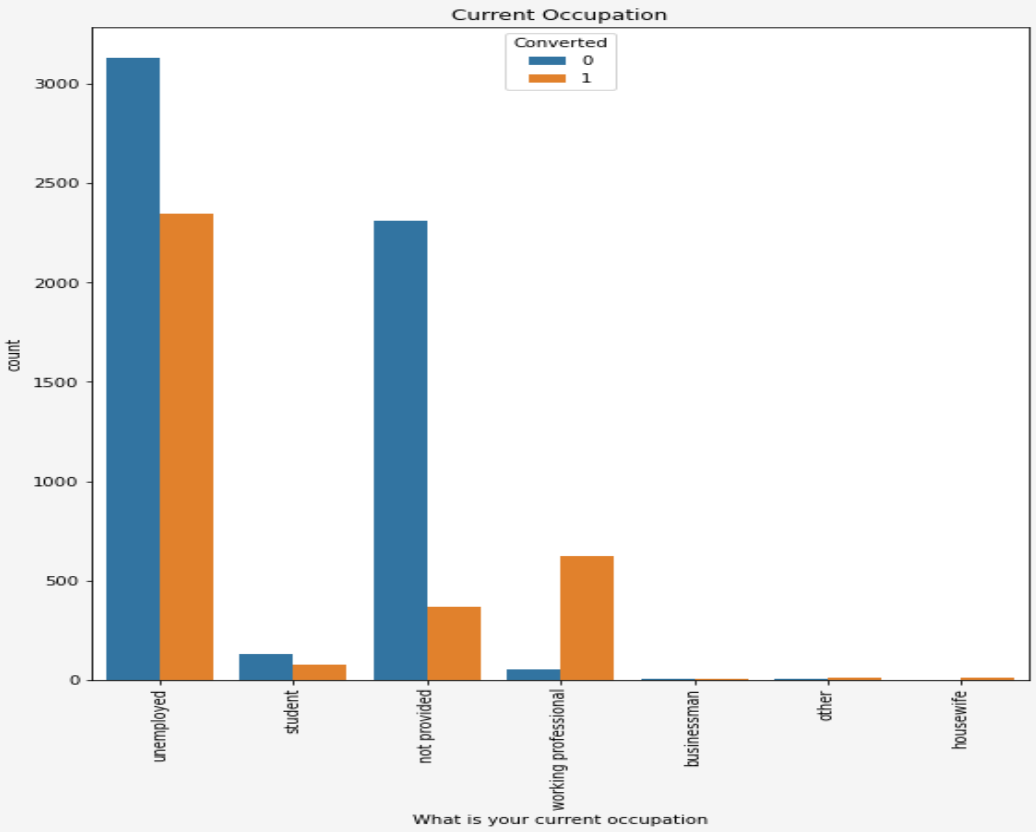


EDA(Exploratory data analysis)

Univariate Analysis

Relation between categorical variables to Converted

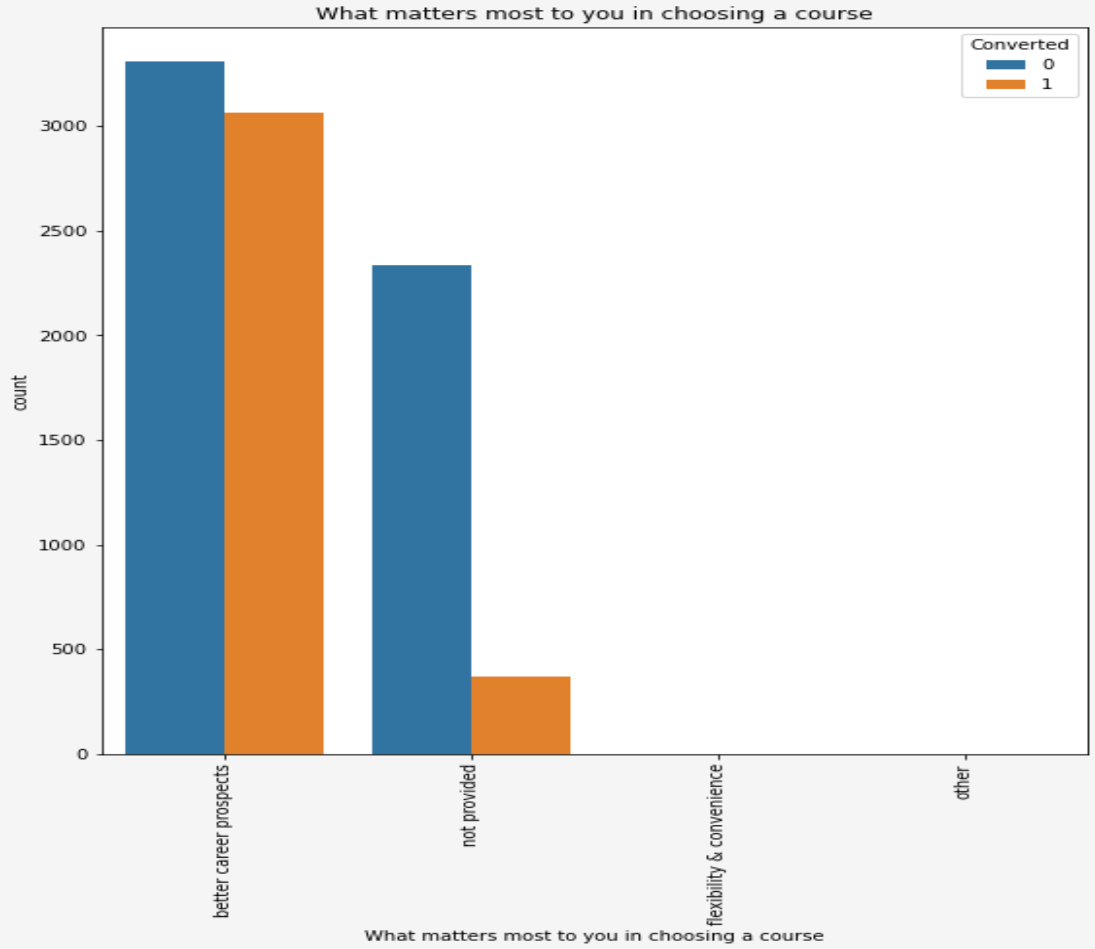


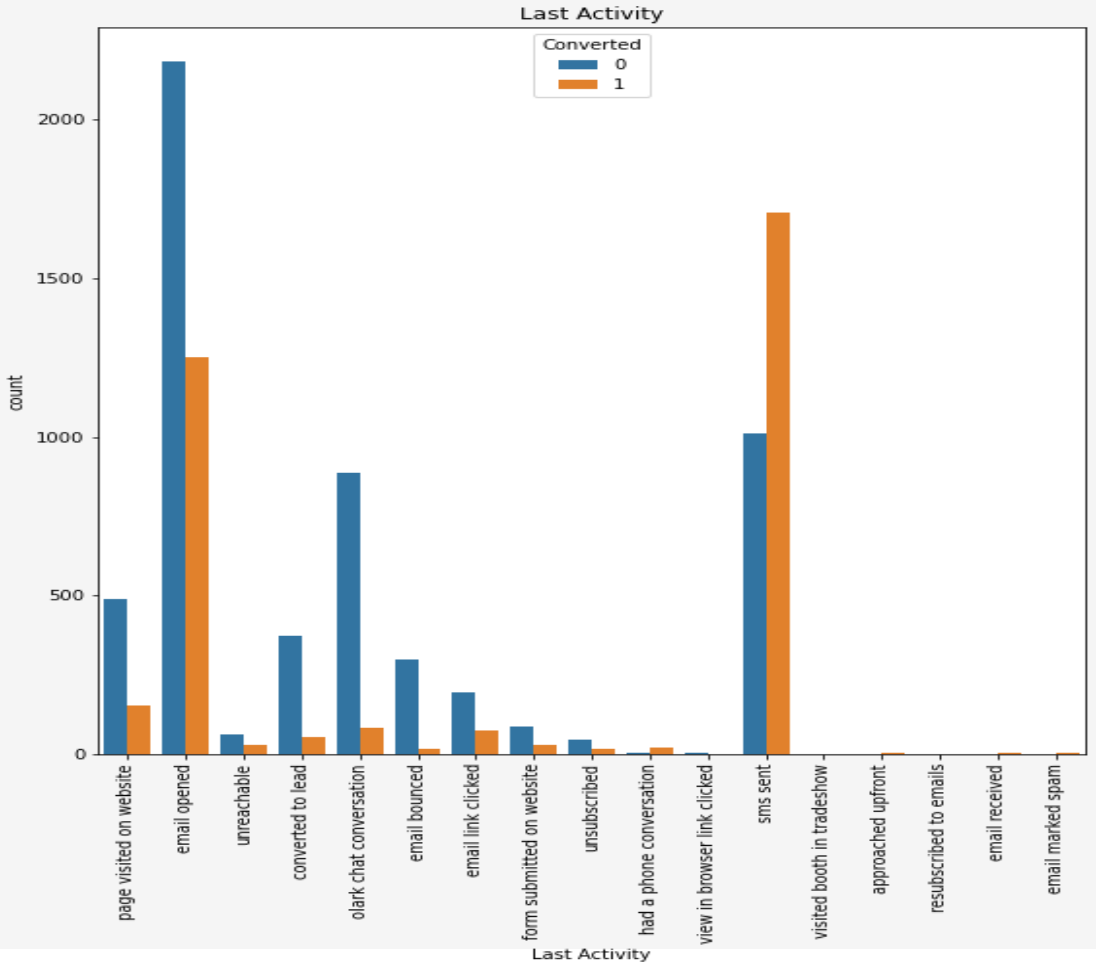


EDA(Exploratory data analysis)

Univariate Analysis

Relation between categorical variables to Converted





EDA(Exploratory data analysis)

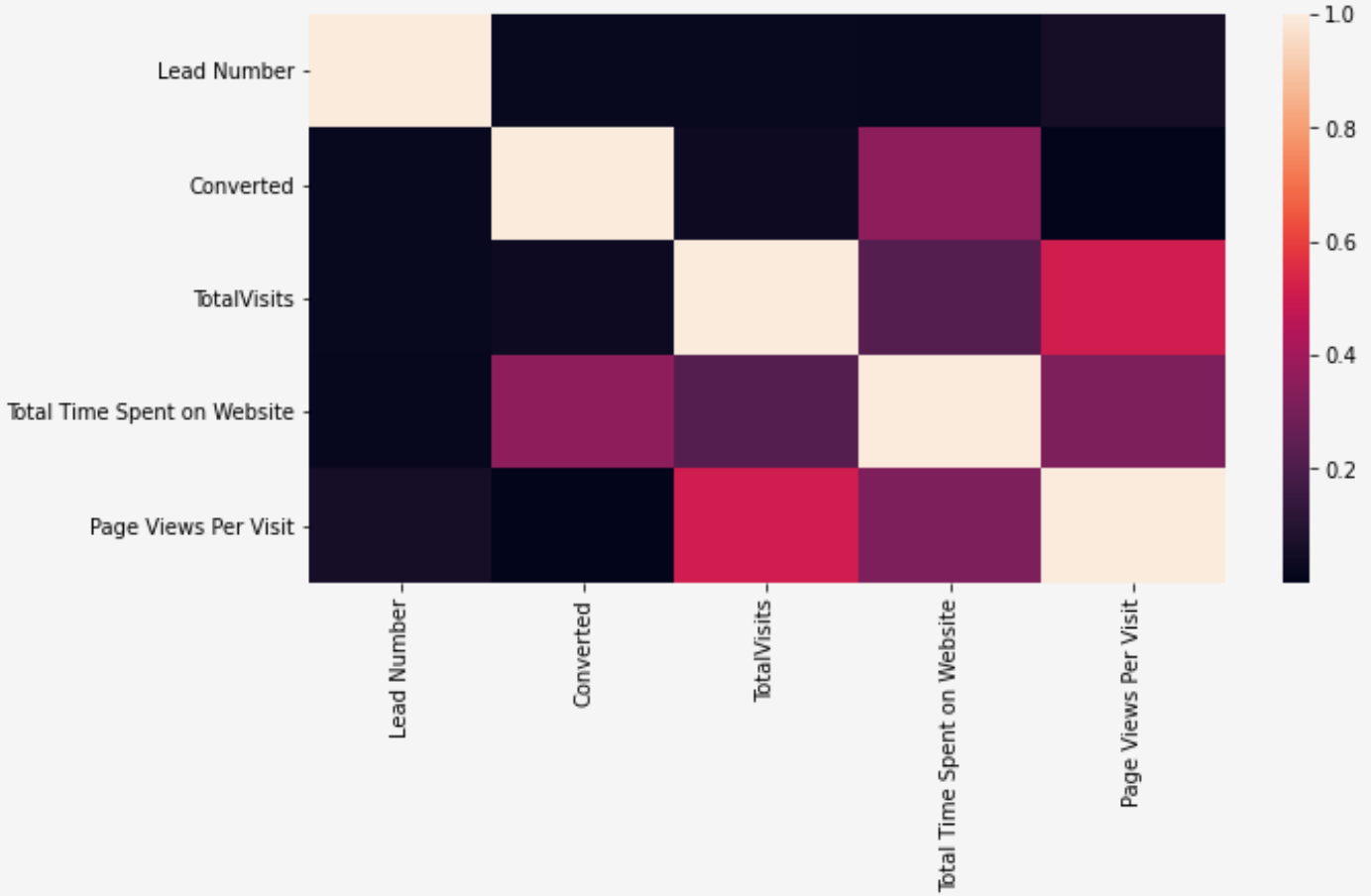
Univariate Analysis

a) **Univariate Analysis for Categorical Variables**

b) **Univariate Analysis for Numerical Variables**

c) **Relation between categorical variables to Converted**

d) **Correlation among variables**



EDA(Exploratory data analysis)

Univariate Analysis

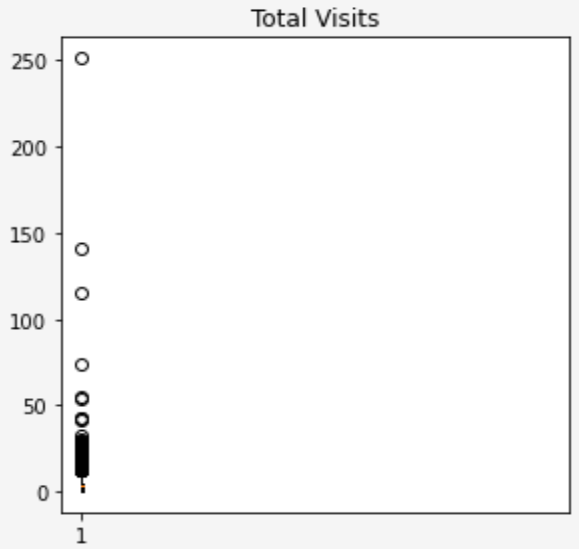
a) **Univariate Analysis for Categorical Variables**

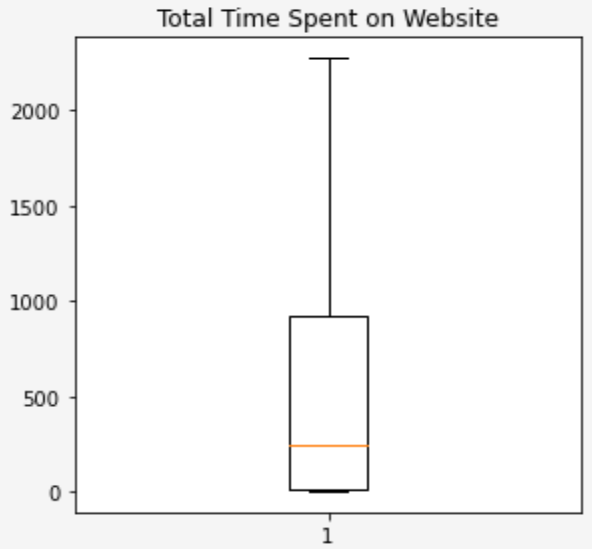
b) **Univariate Analysis for Numerical Variables**

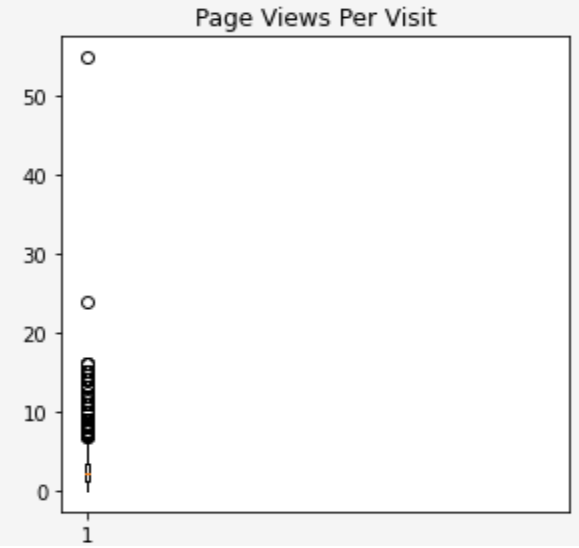
c) **Relation between categorical variables to Converted**

d) **Correlation among variables**

e) **Outliers handling**







Dummy Variables

Total 9074 rows and 22 columns to be Analyzed

Dummy Variables are created for Object type variables

Model Building

Splitting the data into testing and training set. Split the data into 75:25 ratio.

Use RFE for feature selection

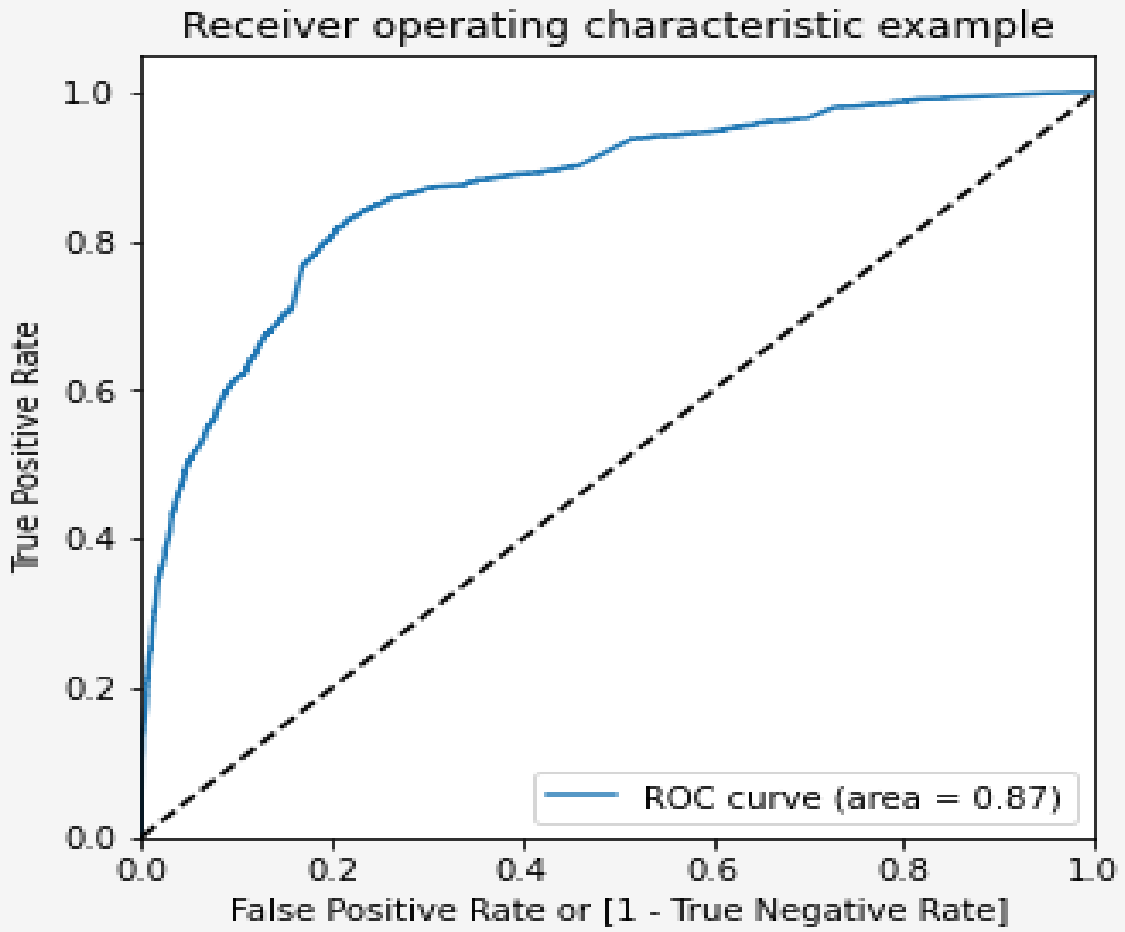
Selected top 15 variables from RFE

Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5

Predictions on test data set

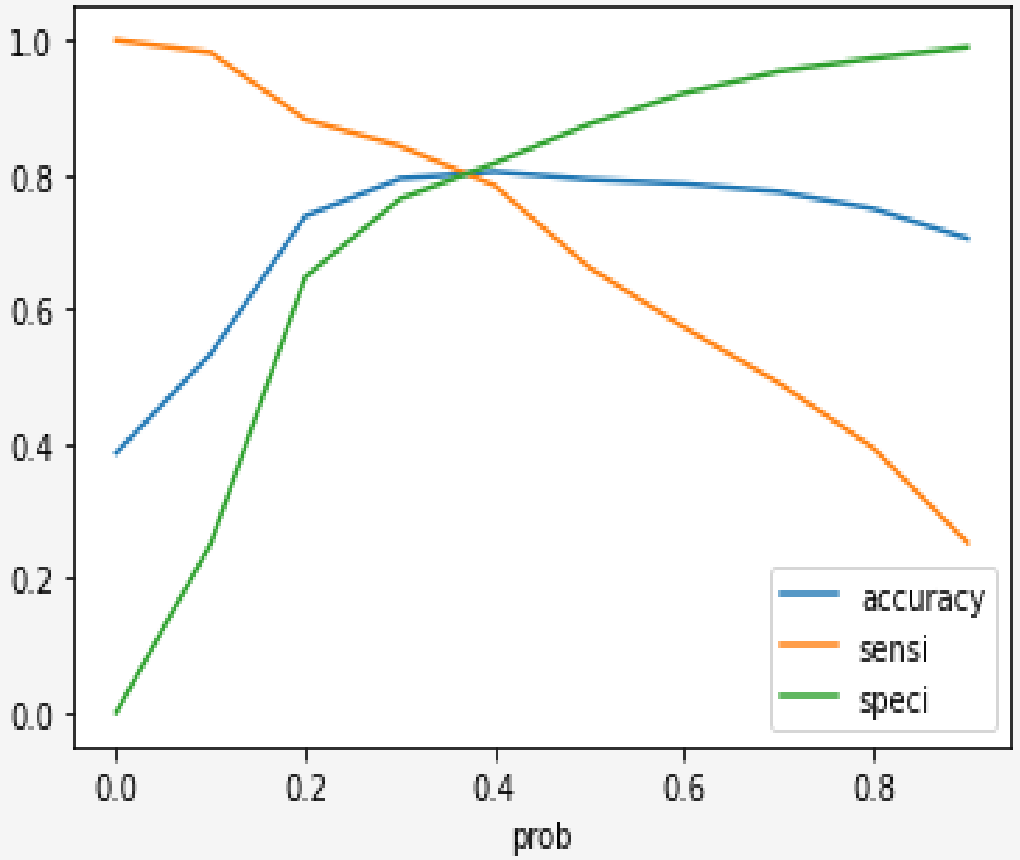
Accuracy is around 79%

ROC Curve



The area under ROC curve is 0.87 which is a very good value.

ROC Curve



From the graph it is visible that the optimal cut off is at 0.35.

Conclusion

It was found that the variables that mattered the most in the potential buyers are :

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
    a. direct traffic
    b. welingak website

4. When the last activity was Olark chat conversation.
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.