LEAD SCORE CASE STUDY

SUBMITTED BY –

SAURABH TANEJA
SARGAM AGARWAL
SHAIL SHIKHAR

PROBLEM STATEMENT :-

X education sells online courses on different websites. When people fill up the form and share the email or phone number, they are considered the leads .

These leads are followed by the sales team and when the customer buys the courses these leads are converted. This conversion rate is 30% which is very low.

We need to build a logistic regression model to predict the most potential leads (hot leads) on which sales team can work and do follow ups. This will save the effort and time of the company .

DATA PROVIDED :

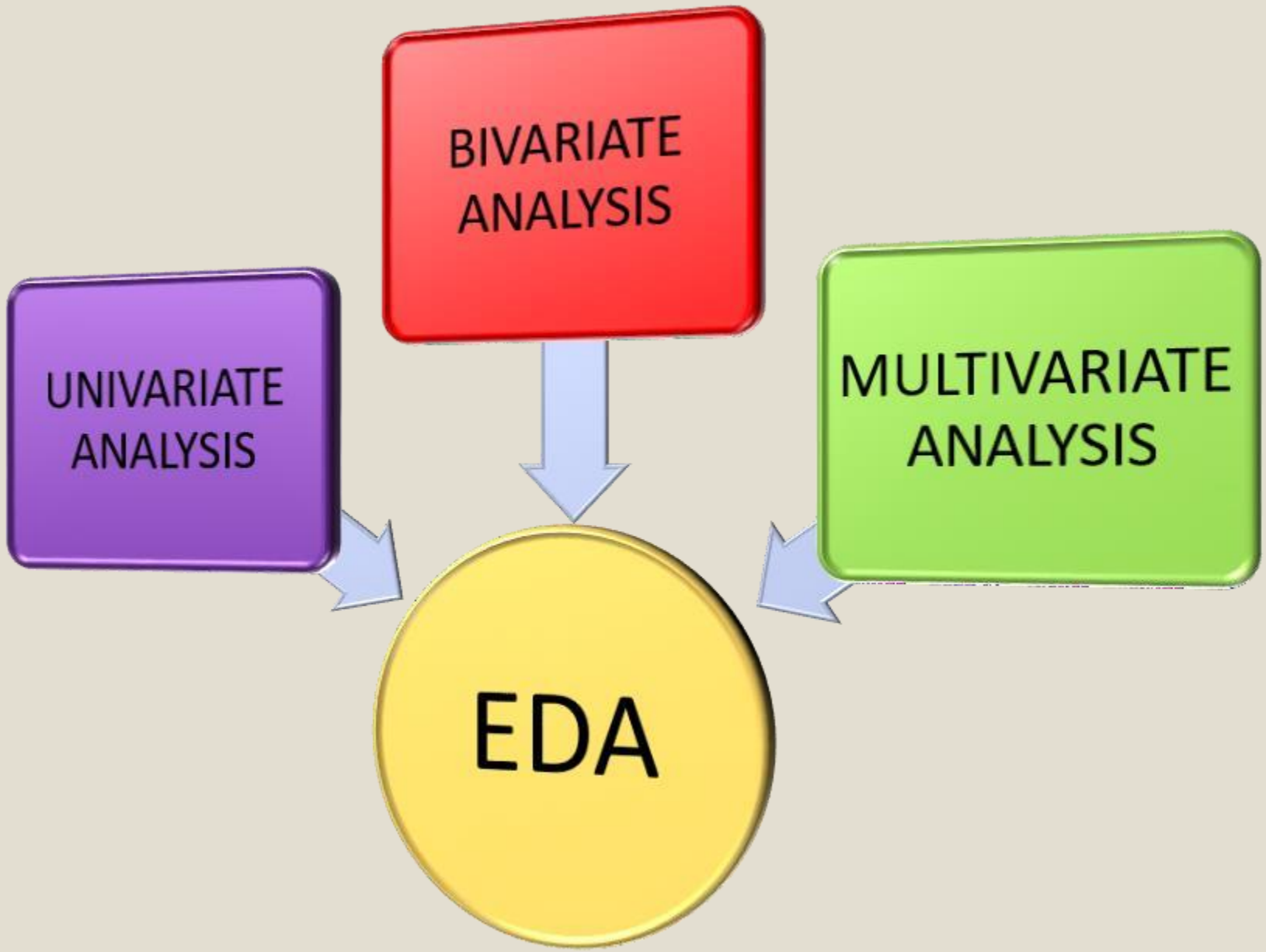
- **Leads.csv :- This dataset consists of various attributes such as Lead Score, Total Time spent on the website , Total Visits , Last Activity etc. which may or may not be useful in deciding whether a lead will converted or not .**
- **The Target variable which is Converted which tells whether a past lead was converted or not whereby 1 indicates that it was converted and 0 means it was not converted .**
- **Leads Data Dictionary.csv :- Ensembles the description of all the columns .**

Approach and Methodology

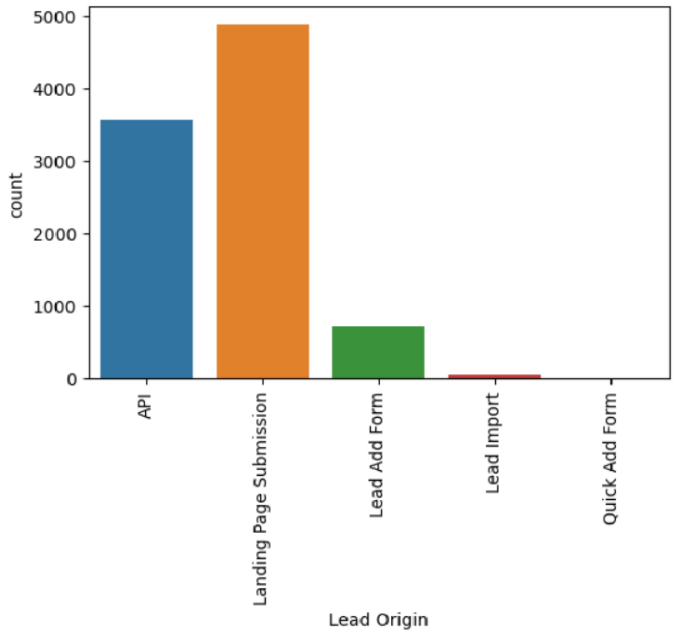
1. **Data loading and Understanding:** Loading the Leads.csv file and understanding the shape ,columns ,info and description of the dataset .

2. **Data Cleaning:** If the missing value count is more than or equal to 40 % the drop those columns from the dataset .

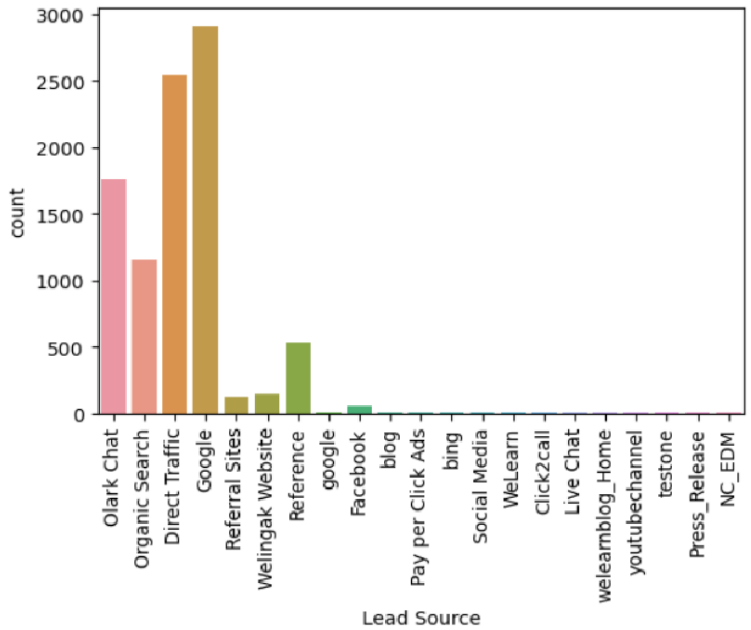
3. **Data Imputation :** For a Categorical variable impute the missing values with Mode and for a continuous variable impute the missing values with Median.


BIVARIATE ANALYSIS
UNIVARIATE ANALYSIS
MULTIVARIATE ANALYSIS
EDA

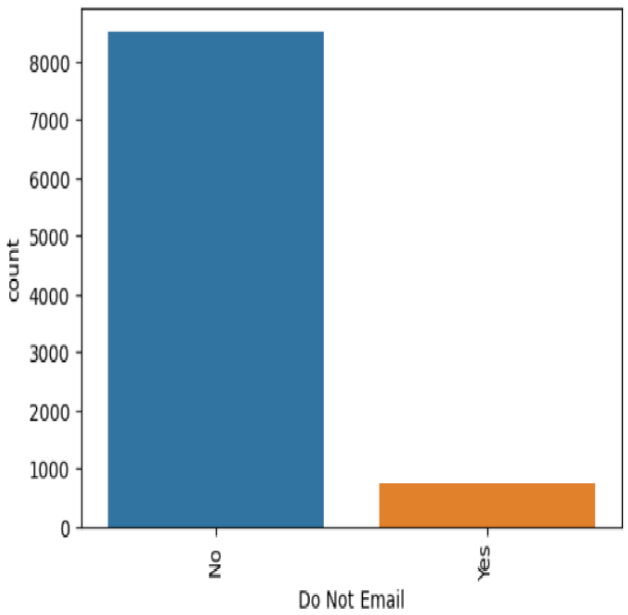
Countplot of Lead Origin



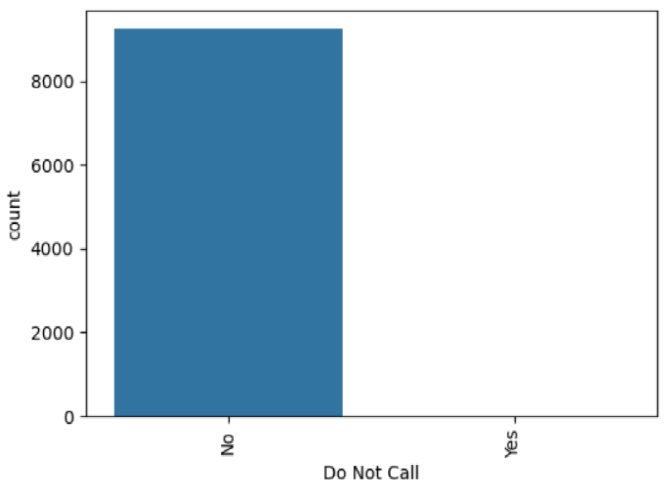
Countplot of Lead Source



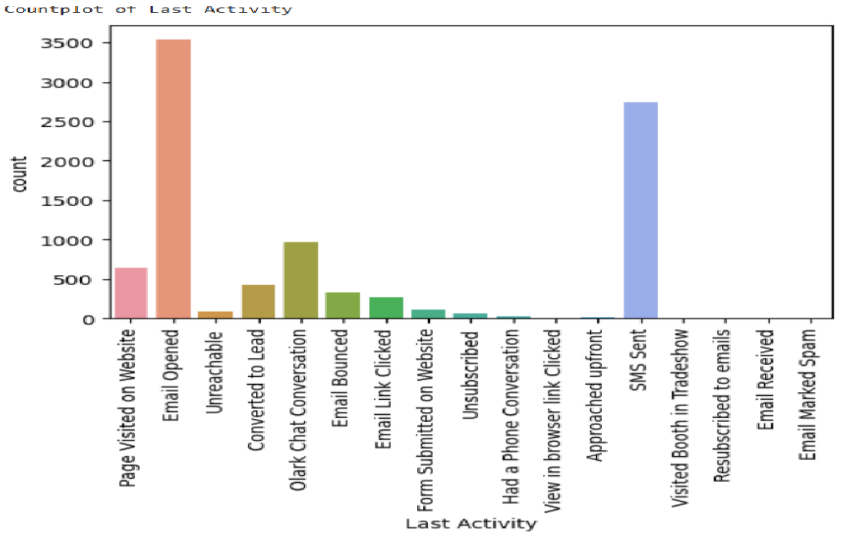
Countplot of Do Not Email



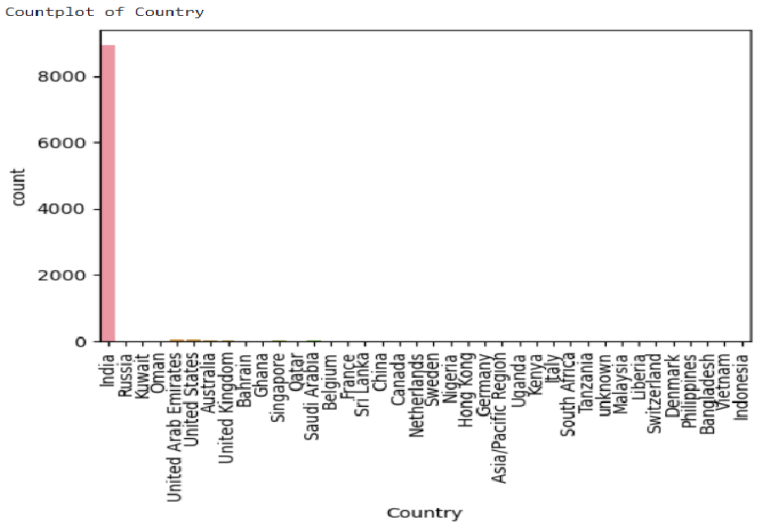
Countplot of Do Not Call

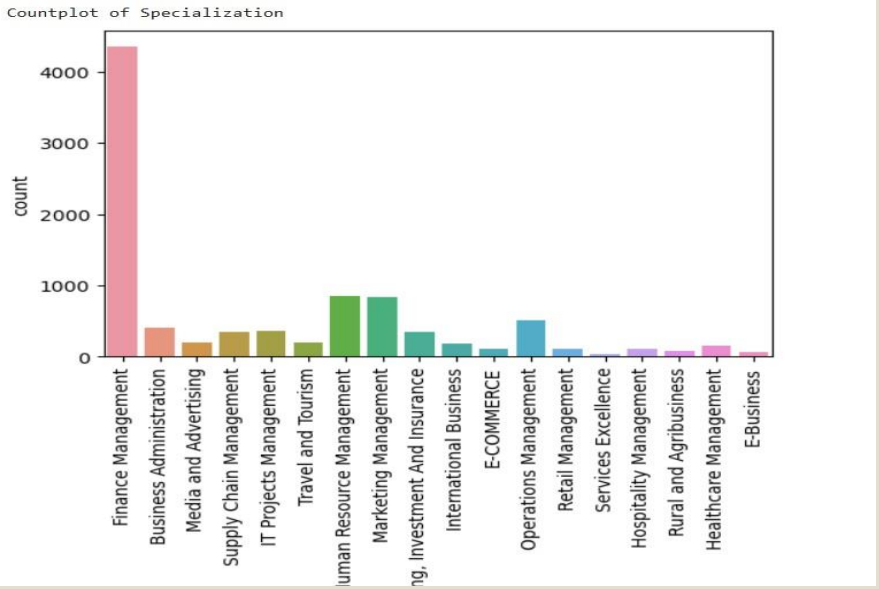


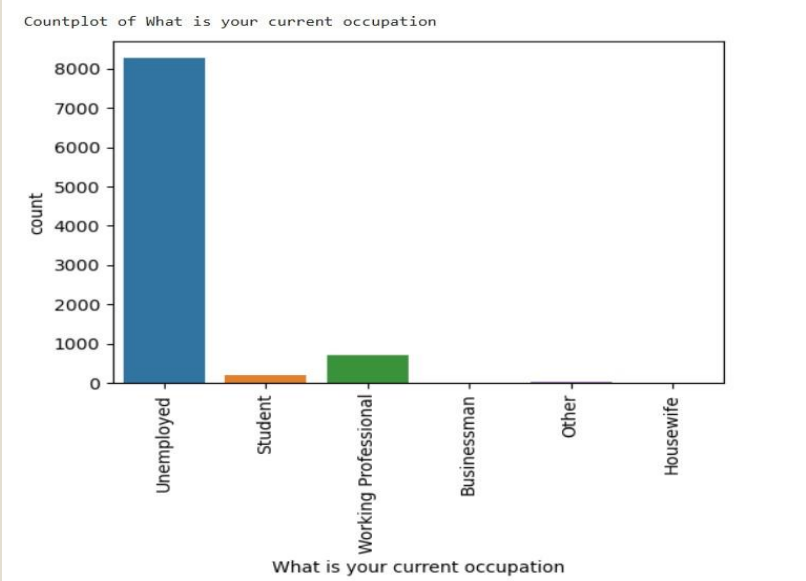
Countplot of Last Activity

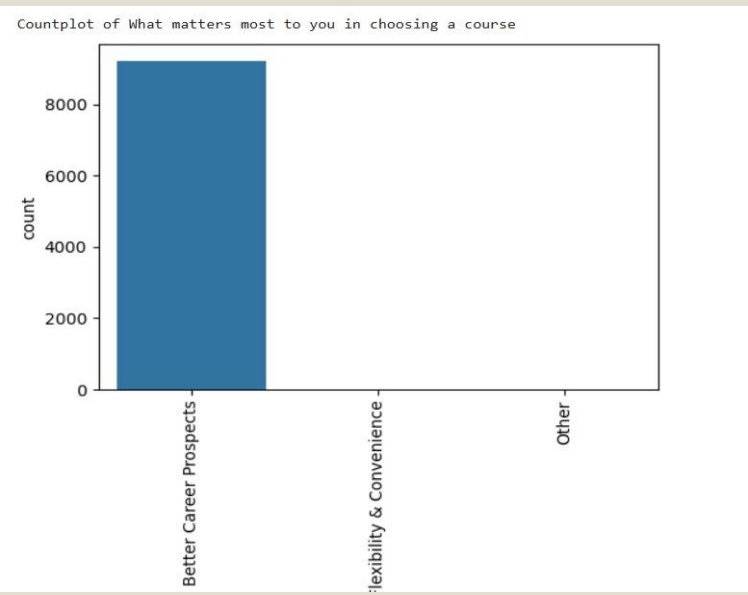


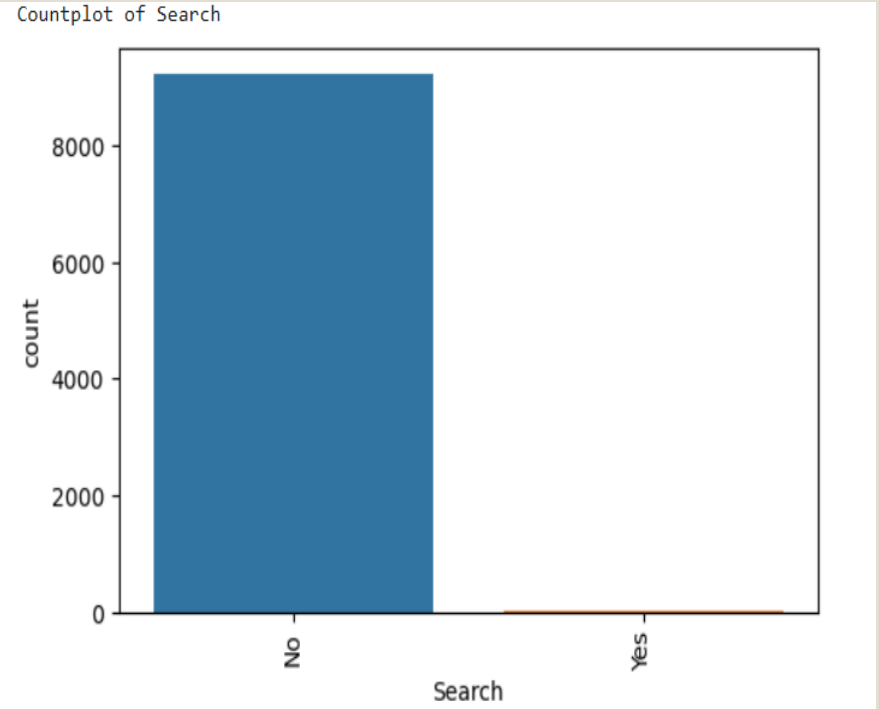
Countplot of Country

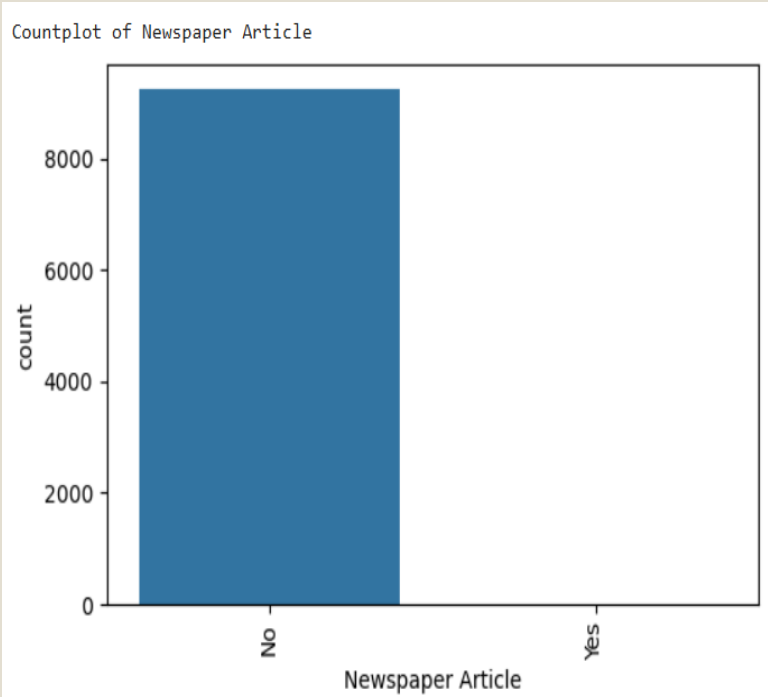


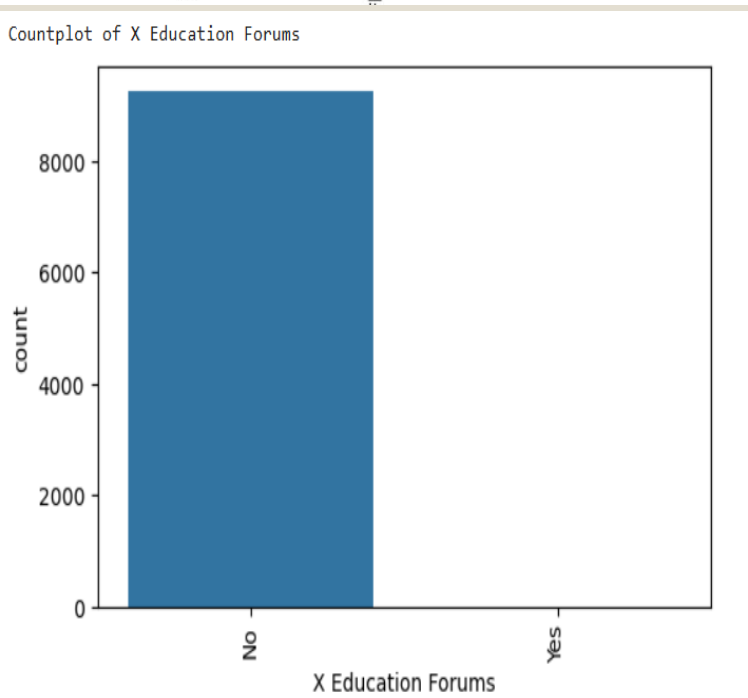

Countplot of Specialization
count
Finance Management
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Travel and Tourism
Marketing Management
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business


Countplot of What is your current occupation
count
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

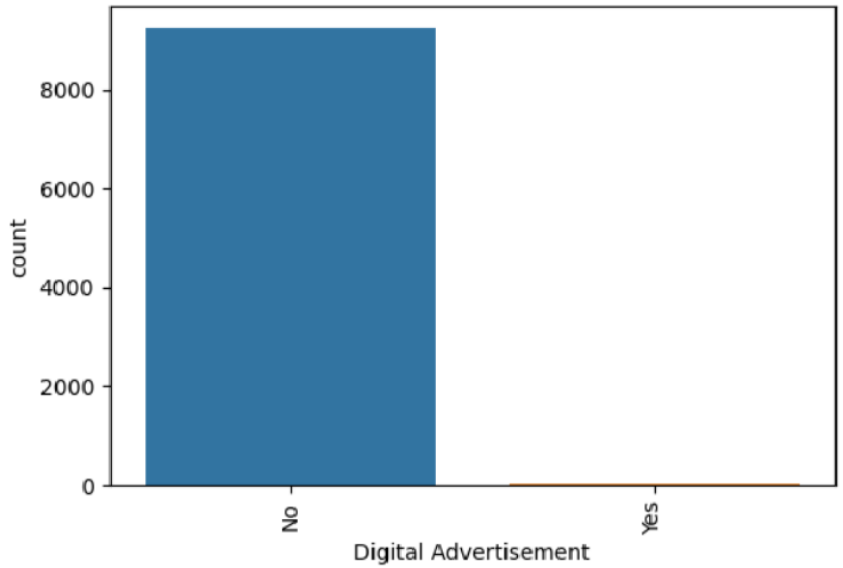

Countplot of What matters most to you in choosing a course
count
Better Career Prospects
Other


Countplot of Search
count
No
Yes
Search

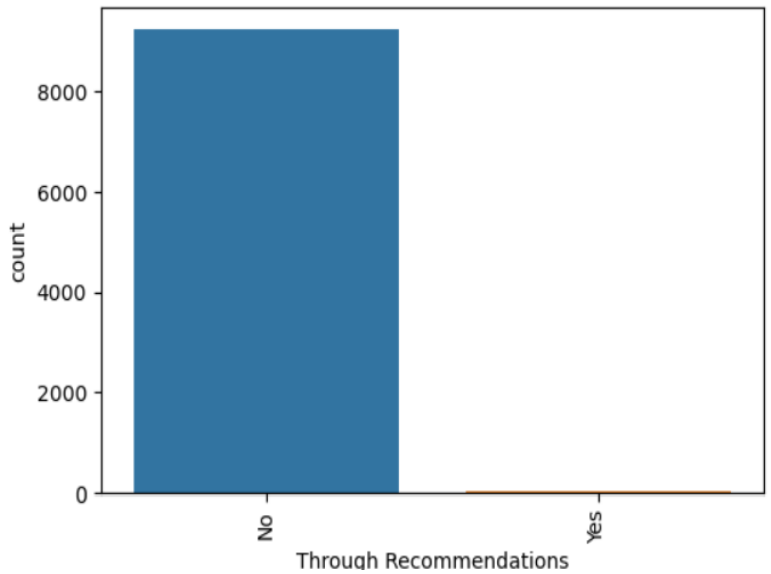

Countplot of Newspaper Article
count
No
Yes
Newspaper Article


Countplot of X Education Forums
count
No
Yes
X Education Forums

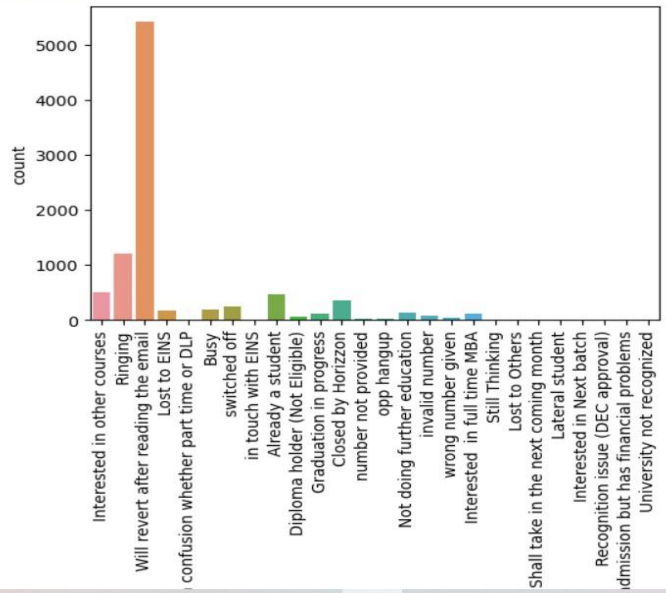
Countplot of Digital Advertisement



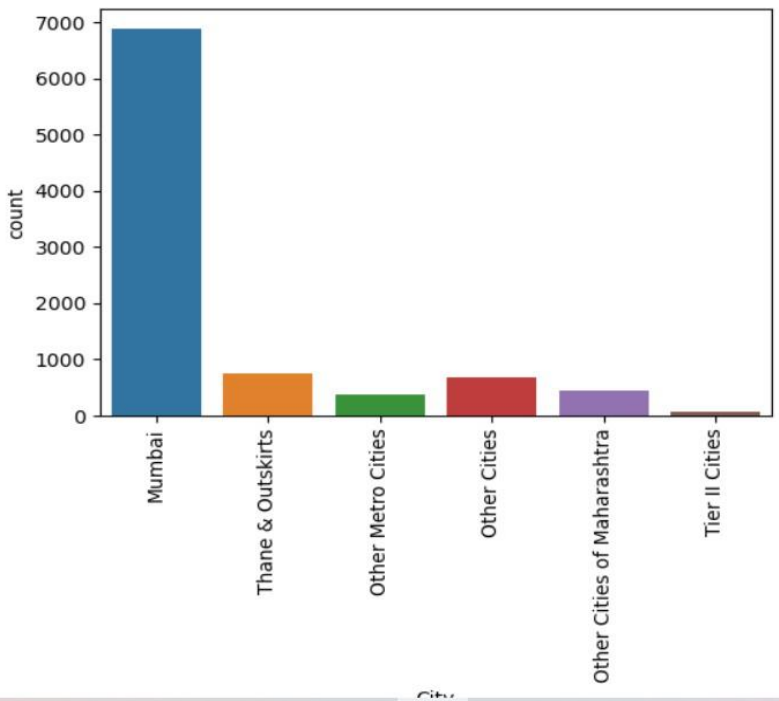
Countplot of Through Recommendations



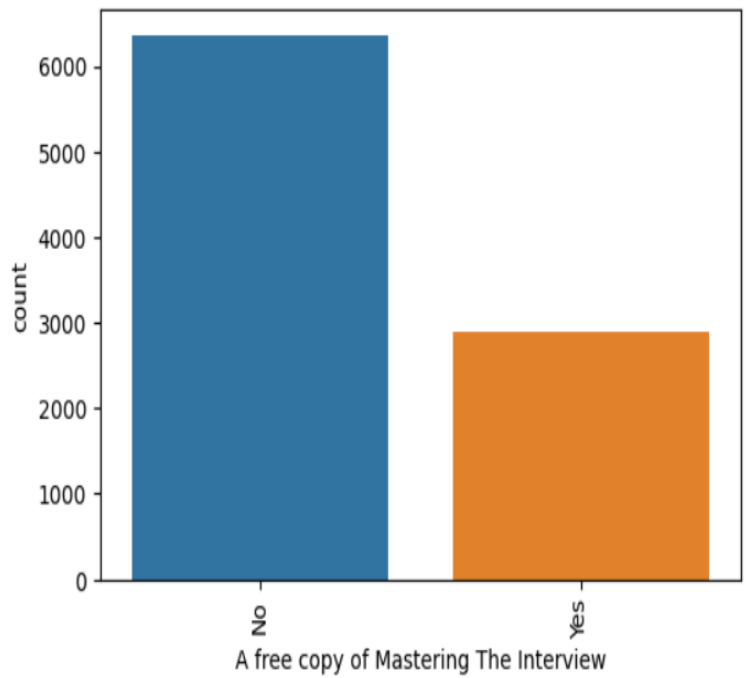
Countplot of Tags

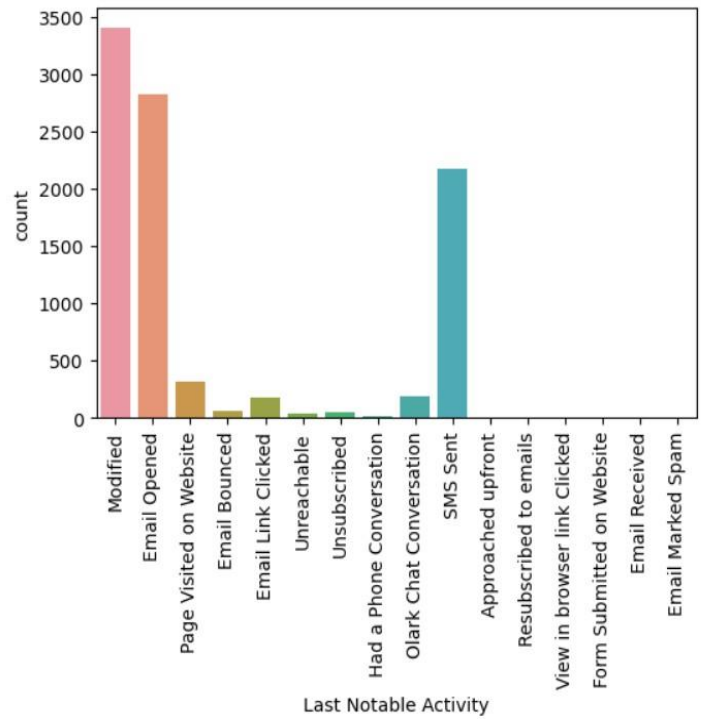


Countplot of City



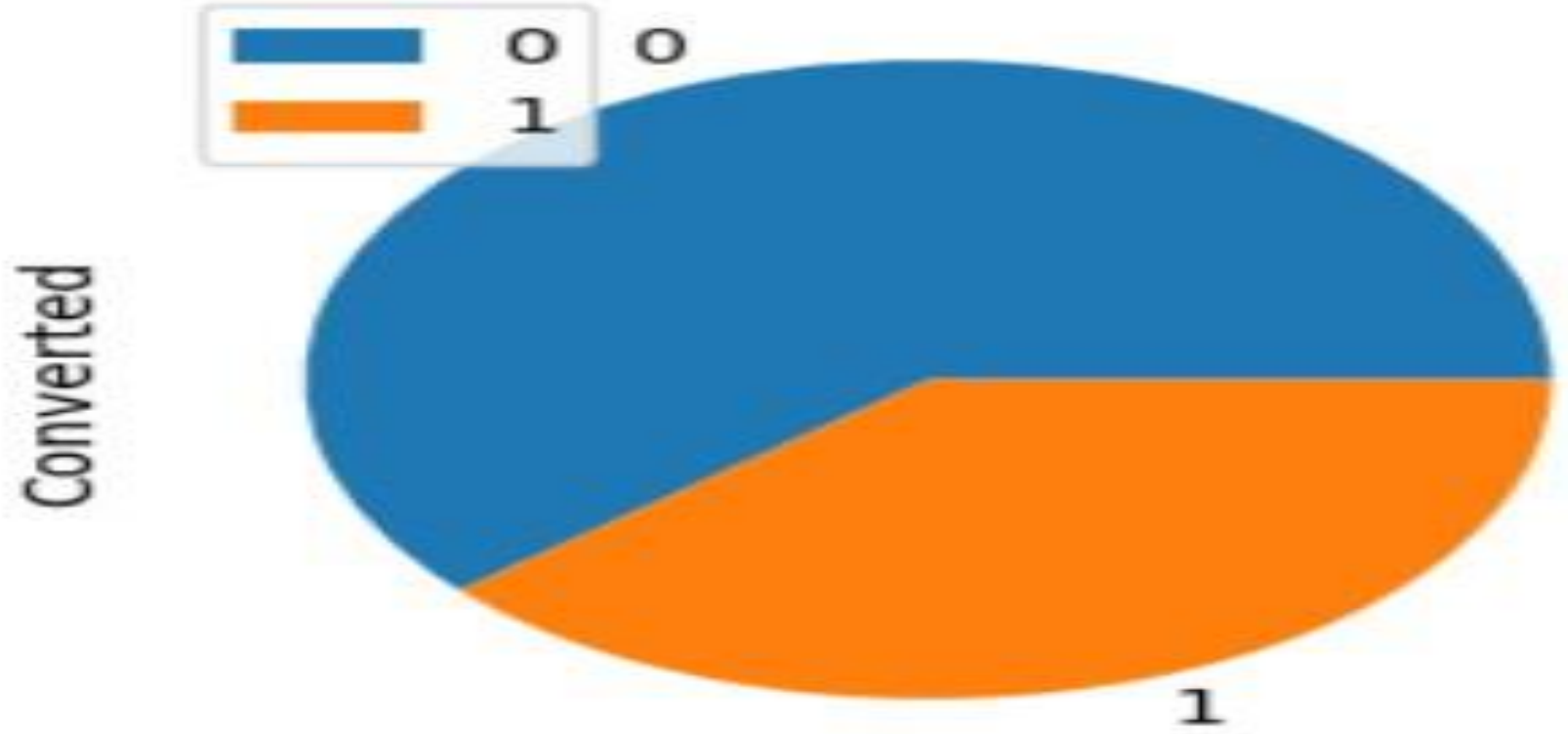
Countplot of A free copy of Mastering The Interview

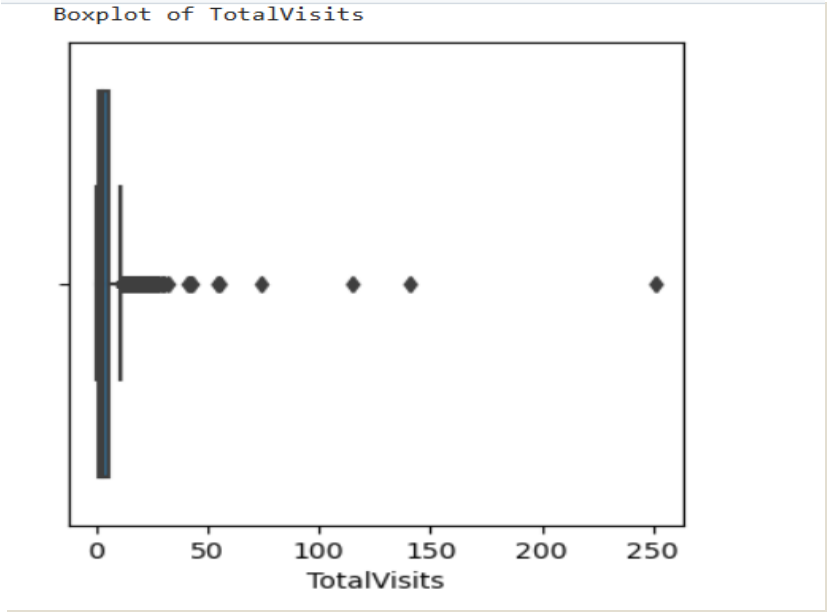


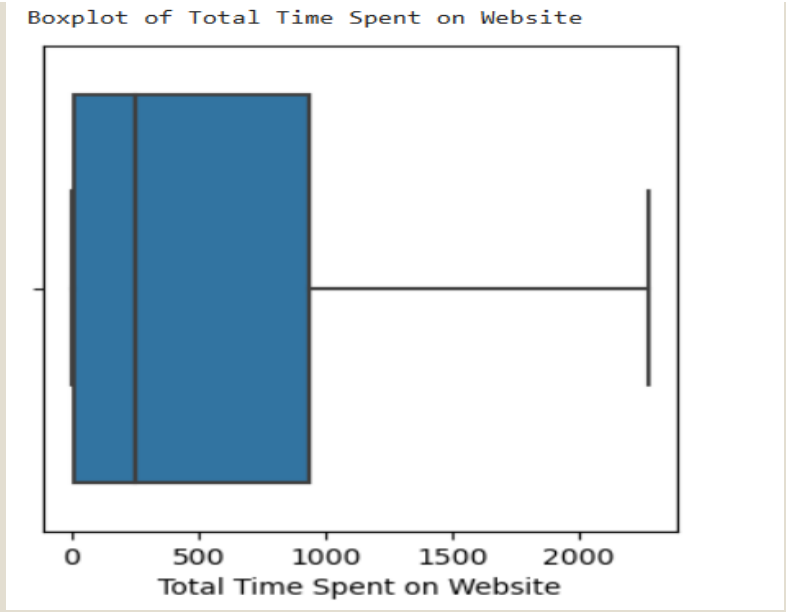


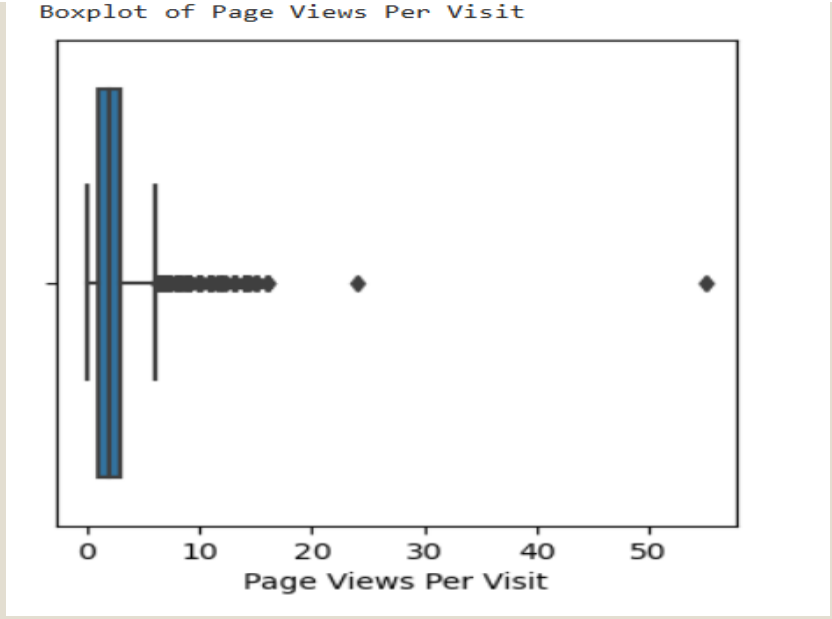
Analysing the Target Variable

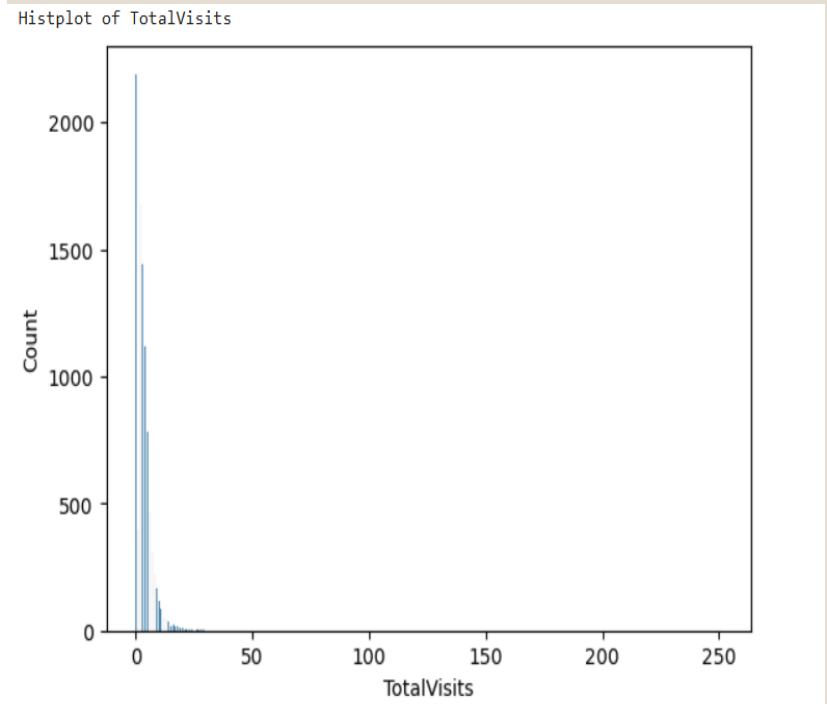
```
0     61.461039
1     38.538961
Name: Converted, dtype: float64
```

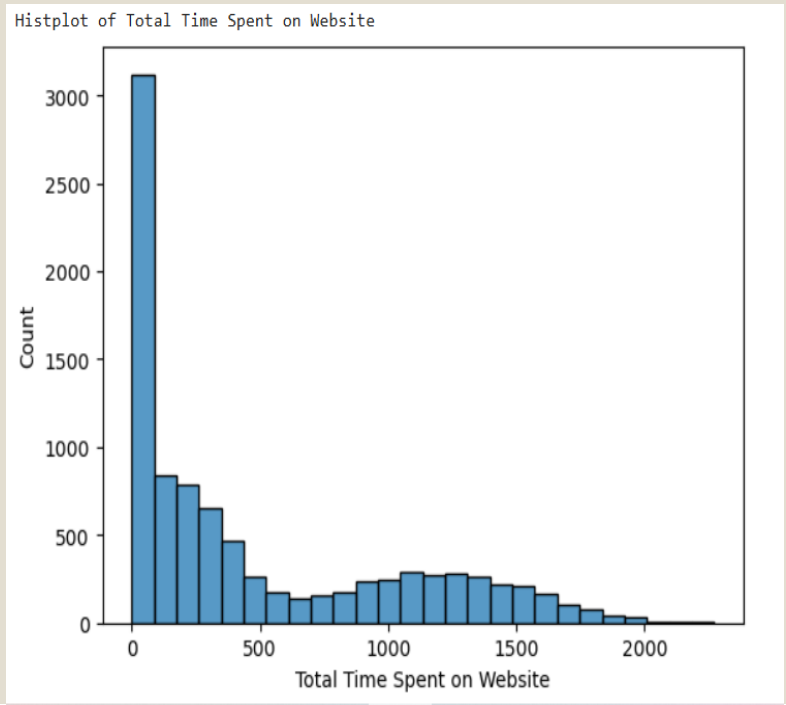


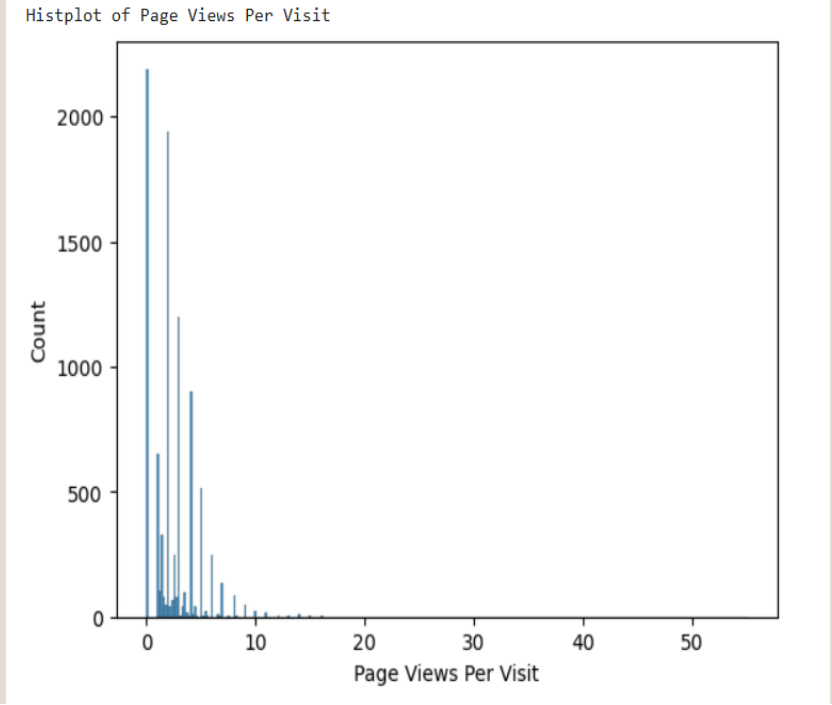

Boxplot of TotalVisits
0
50
100
150
200
250
TotalVisits

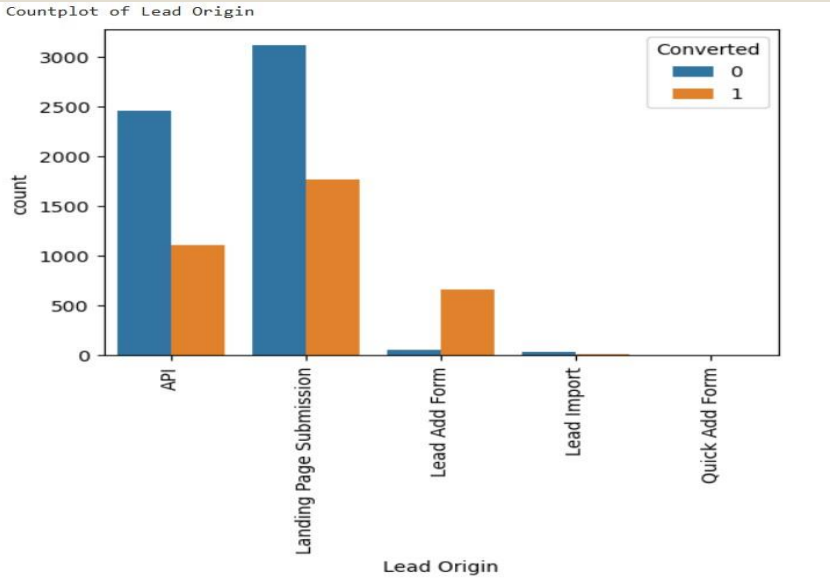

Boxplot of Total Time Spent on Website
0
500
1000
1500
2000
Total Time Spent on Website

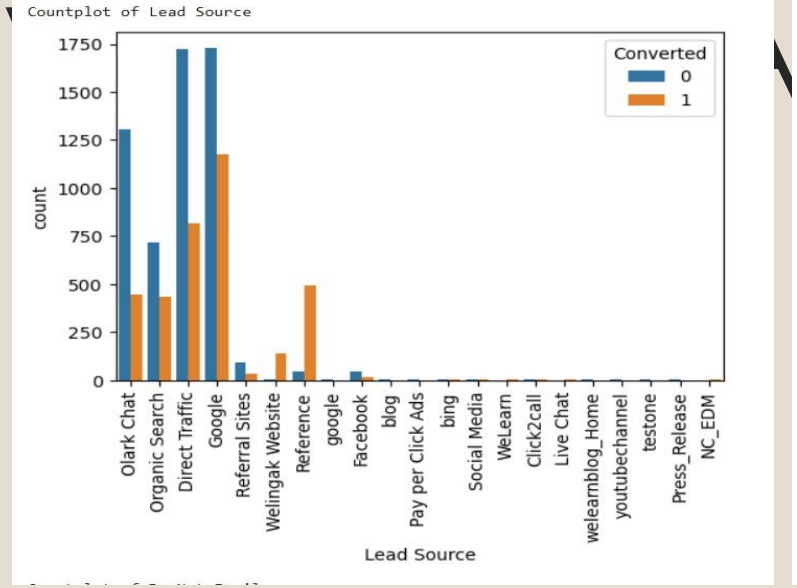

Boxplot of Page Views Per Visit
0
10
20
30
40
50
Page Views Per Visit

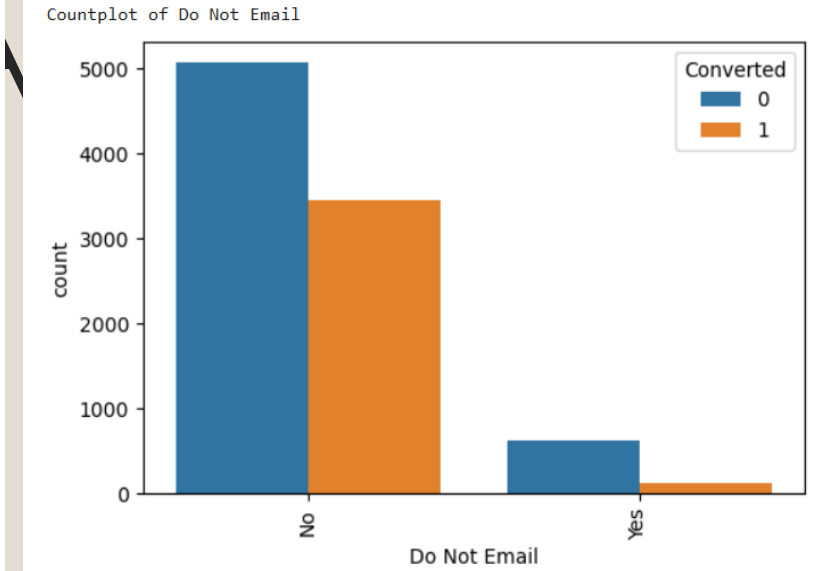

Histplot of TotalVisits
Count
0
500
1000
1500
2000
0
50
100
150
200
250
TotalVisits

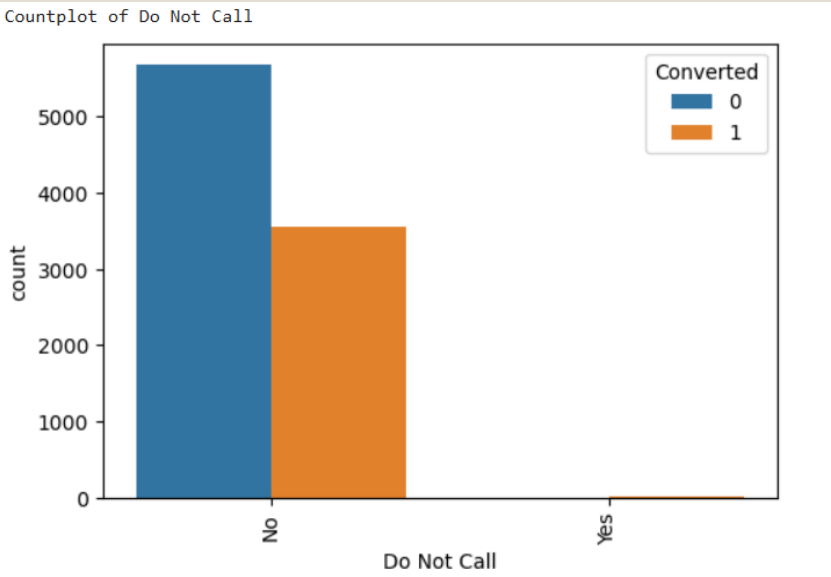

Histplot of Total Time Spent on Website
Count
0
500
1000
1500
2000
2500
3000
0
500
1000
1500
2000
Total Time Spent on Website

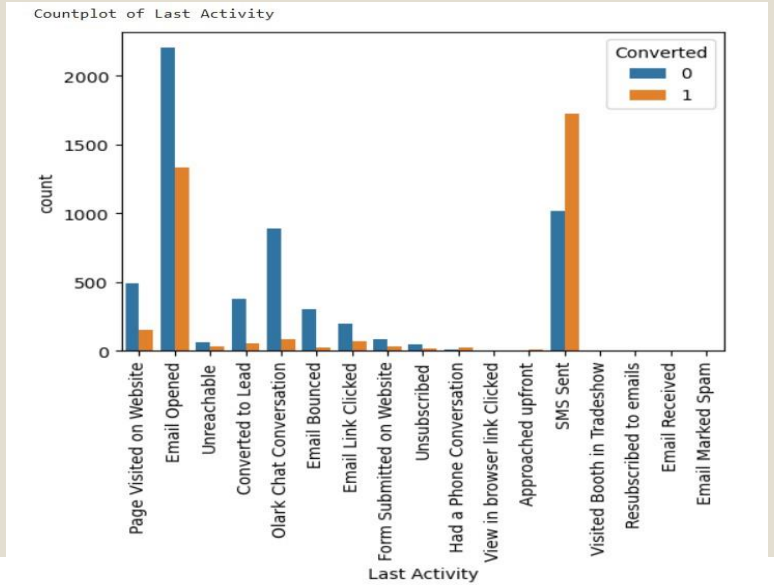

Histplot of Page Views Per Visit
Count
0
500
1000
1500
2000
0
10
20
30
40
50
Page Views Per Visit

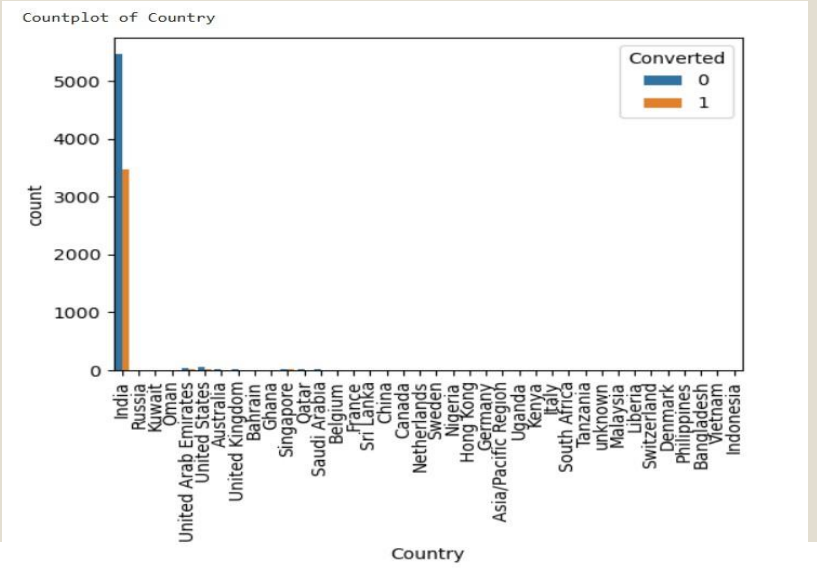

Countplot of Lead Origin
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin


Countplot of Lead Source
Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
weleamblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source

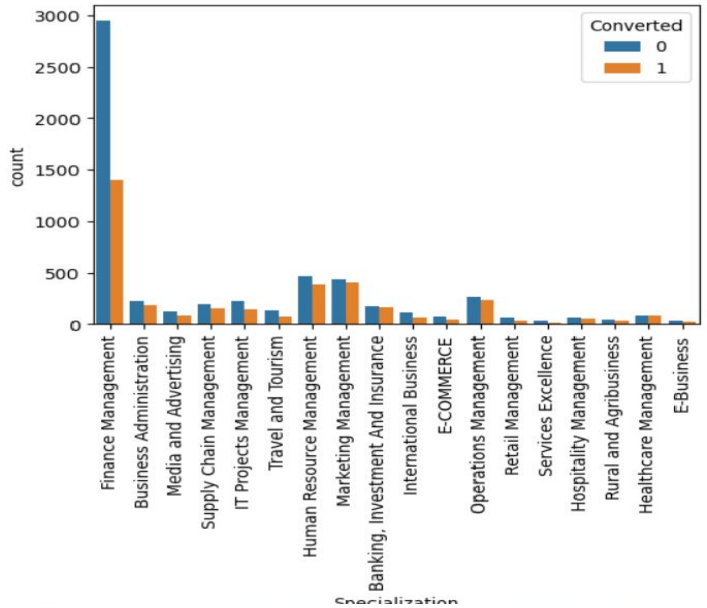

Countplot of Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email


Countplot of Do Not Call
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Call

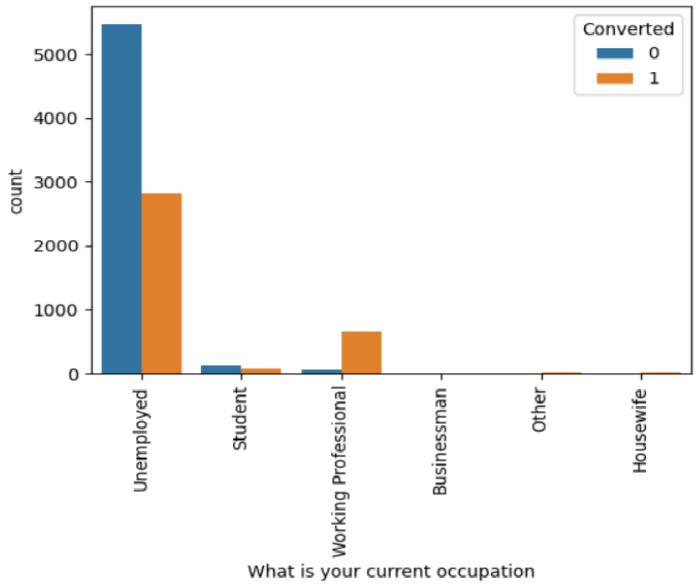

Countplot of Last Activity
Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity


Countplot of Country
Converted
0
1
count
5000
4000
3000
2000
1000
0
India
Russia
Kuwait
Oman
United Arab Emirates
United States
Australia
United Kingdom
Bahrain
Ghana
Singapore
Qatar
Saudi Arabia
Belgium
France
Sri Lanka
China
Canada
Netherlands
Sweden
Nigeria
Hong Kong
Germany
Asia/Pacific Region
Uganda
Kenya
Italy
South Africa
Tanzania
unknown
Malaysia
Liberia
Switzerland
Denmark
Philippines
Bangladesh
Vietnam
Indonesia
Country

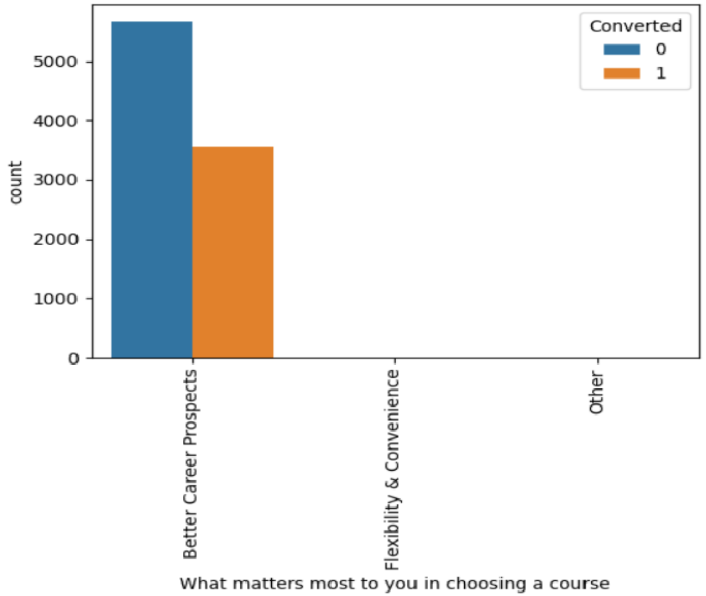
Countplot of Specialization



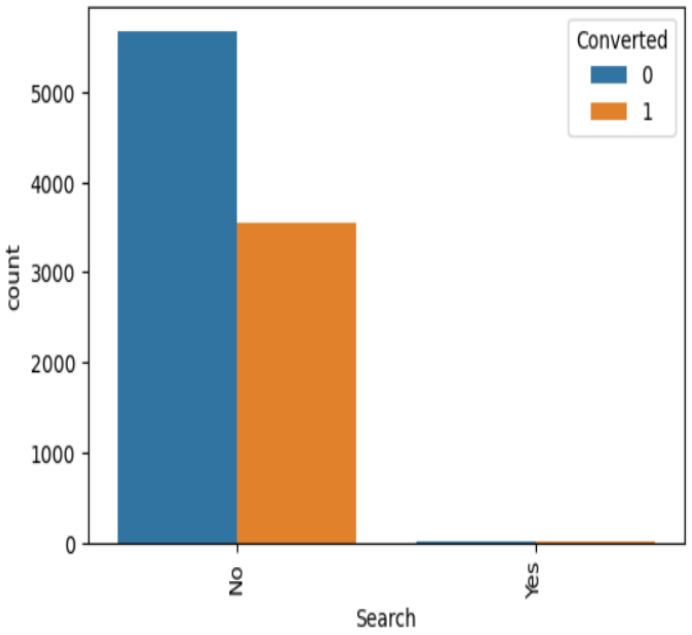
Countplot of What is your current occupation



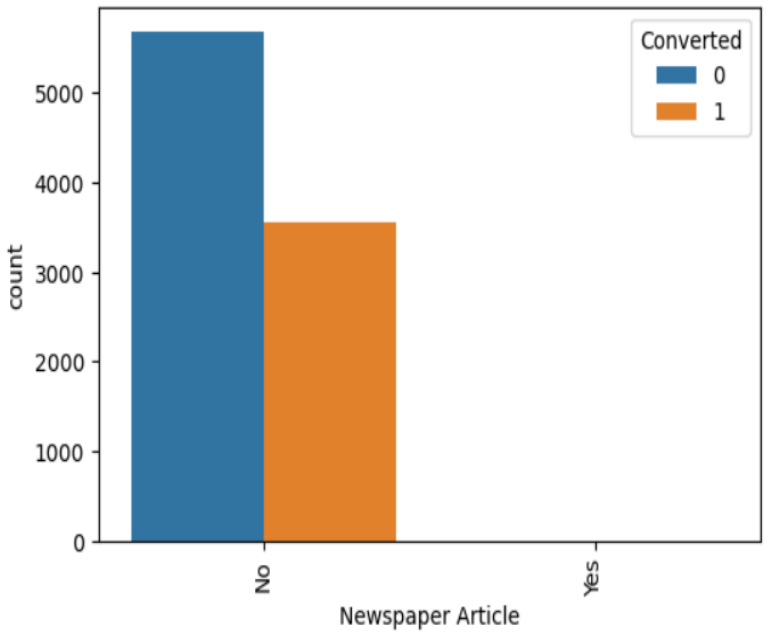
Countplot of What matters most to you in choosing a course



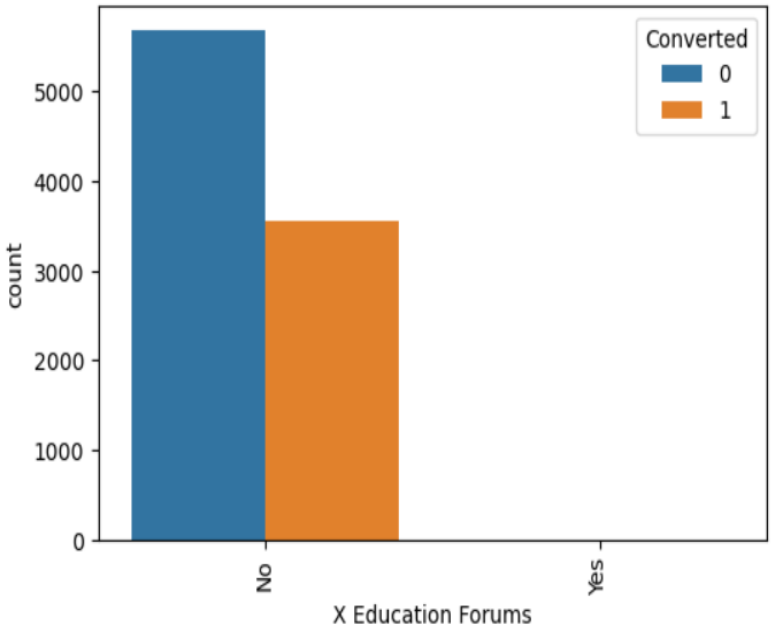
Countplot of Search



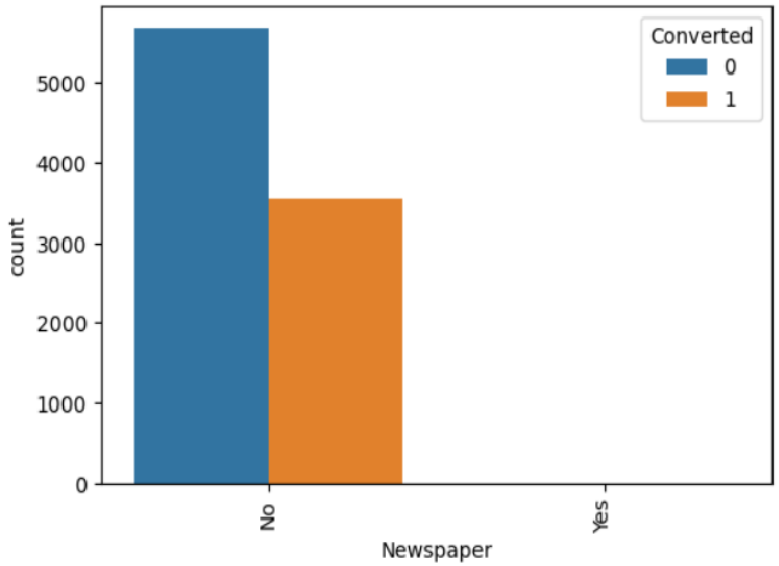
Countplot of Newspaper Article



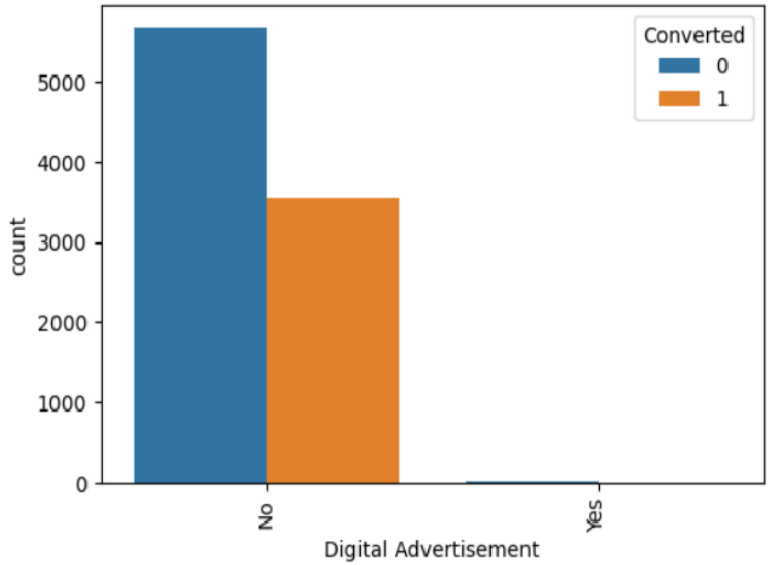
Countplot of X Education Forums



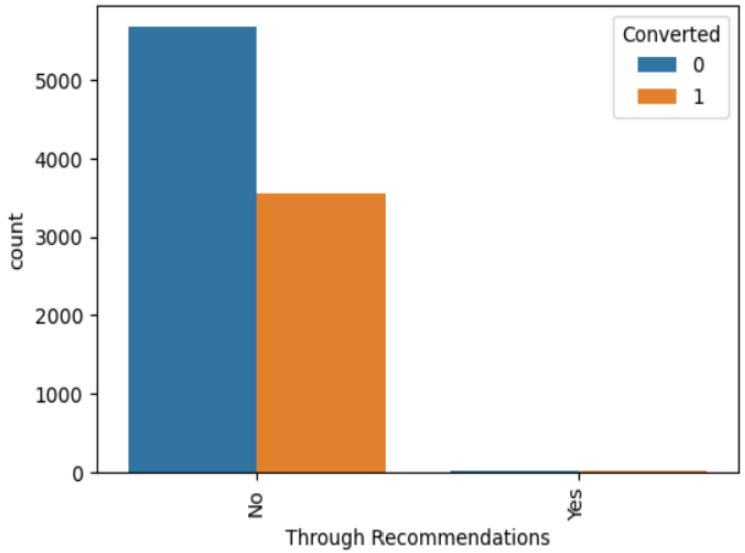
Countplot of Newspaper



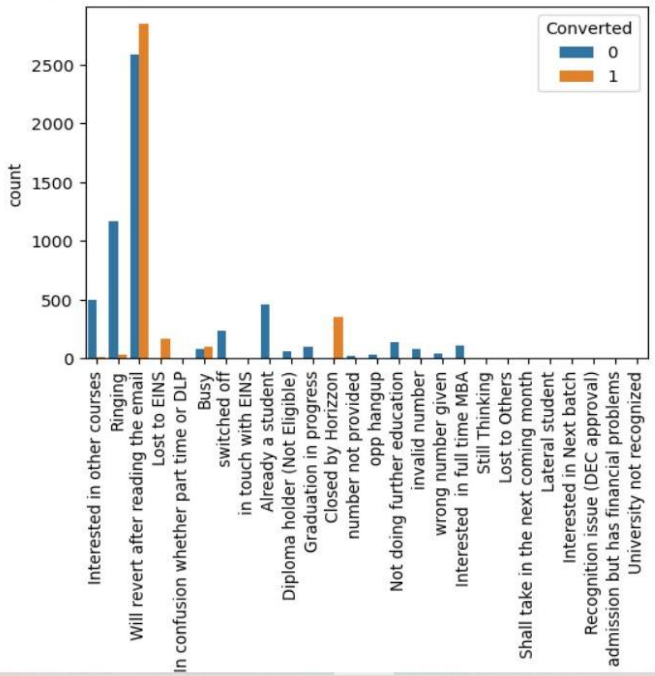
Countplot of Digital Advertisement



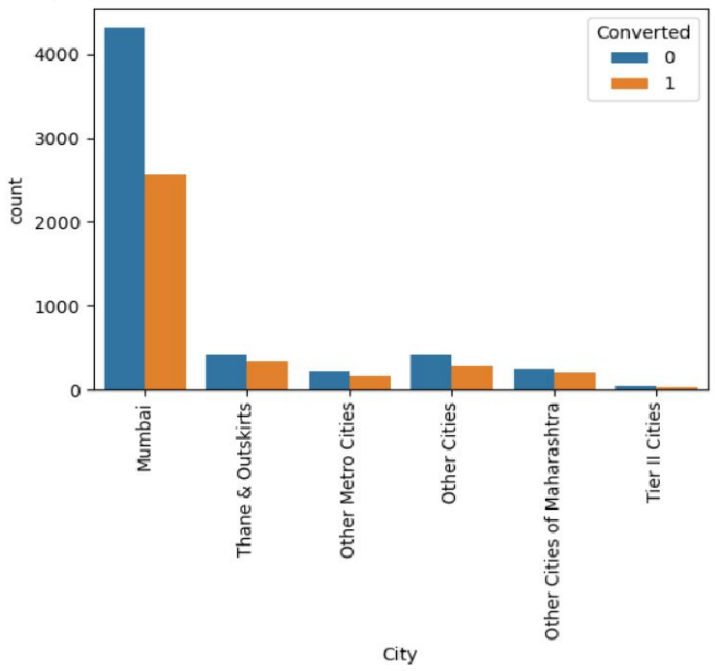
Countplot of Through Recommendations



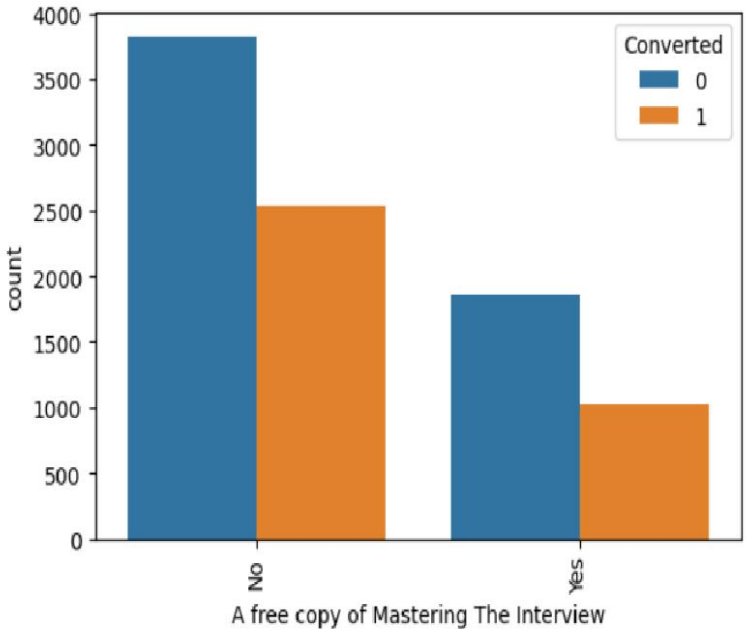
Countplot of Tags

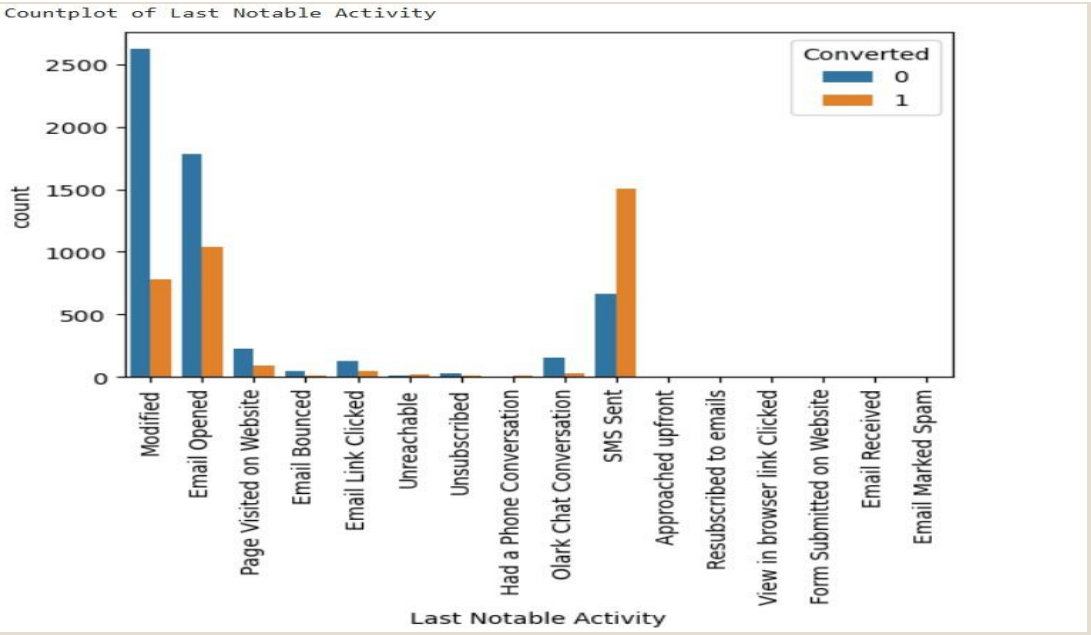


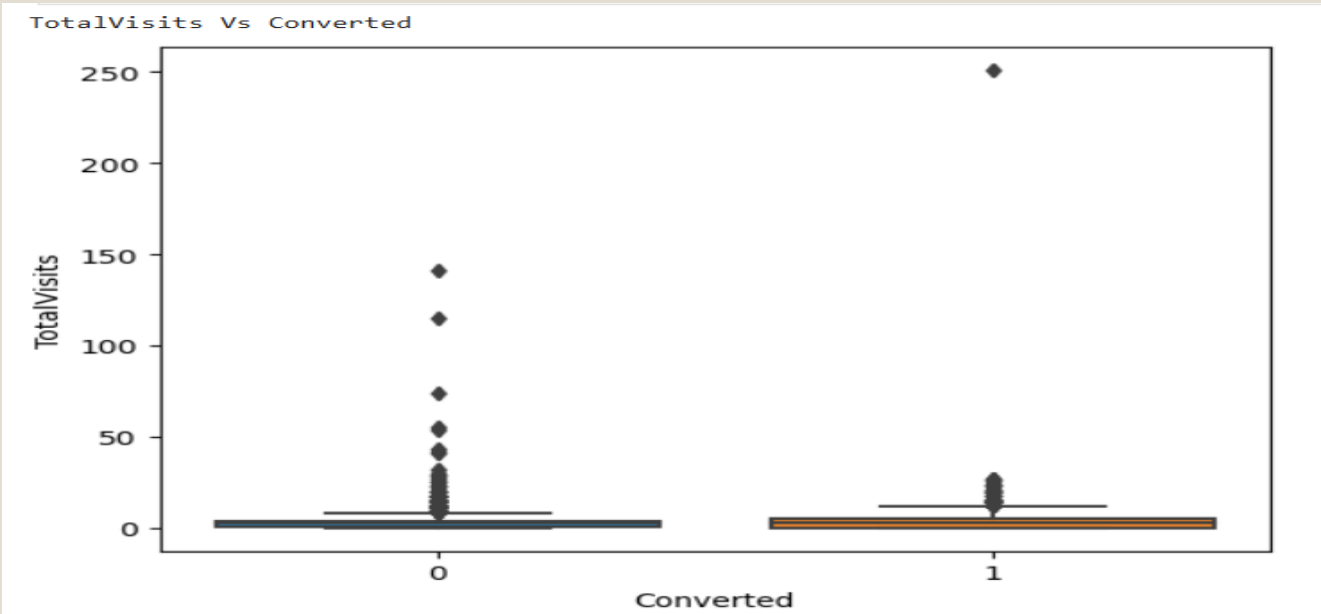
Countplot of City

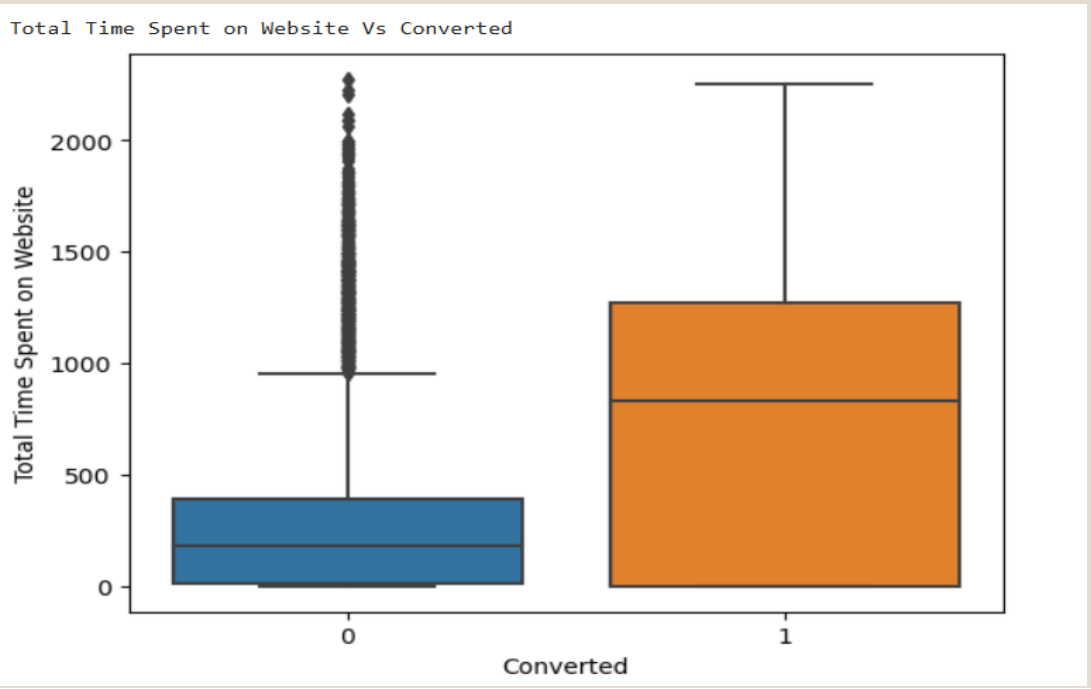


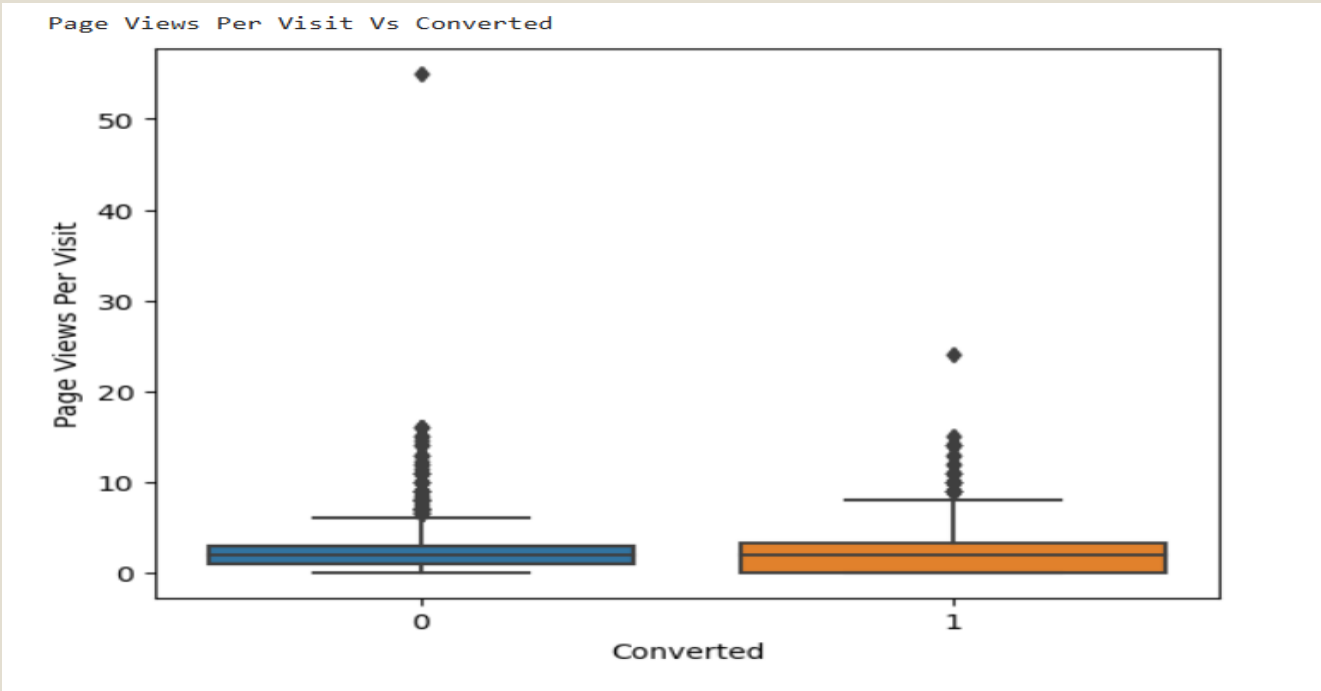
Countplot of A free copy of Mastering The Interview



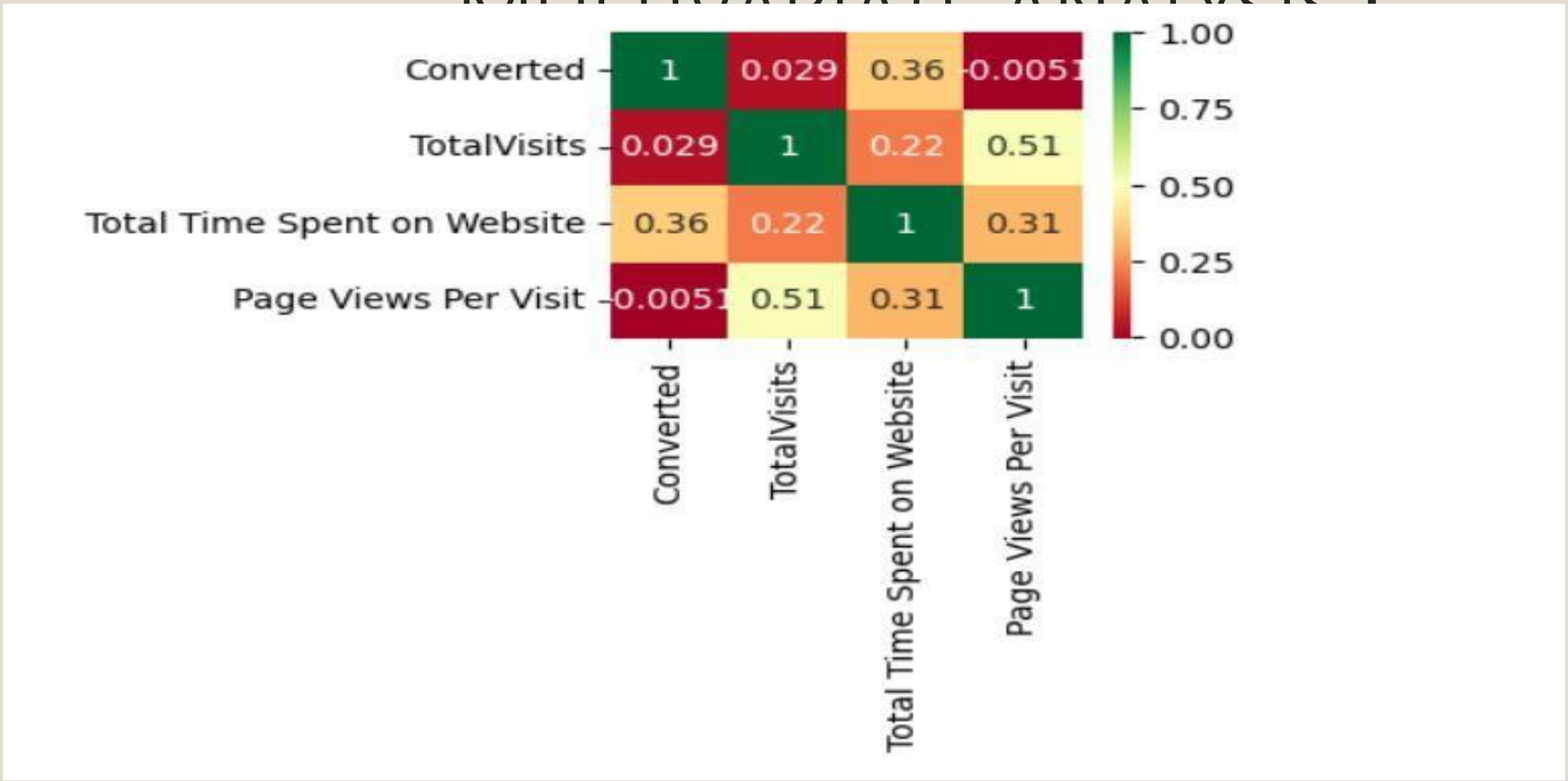

Countplot of Last Notable Activity
Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity


TotalVisits Vs Converted
TotalVisits
250
200
150
100
50
0
0
1
Converted


Total Time Spent on Website Vs Converted
Total Time Spent on Website
2000
1500
1000
500
0
0
1
Converted


Page Views Per Visit Vs Converted
Page Views Per Visit
50
40
30
20
10
0
0
1
Converted

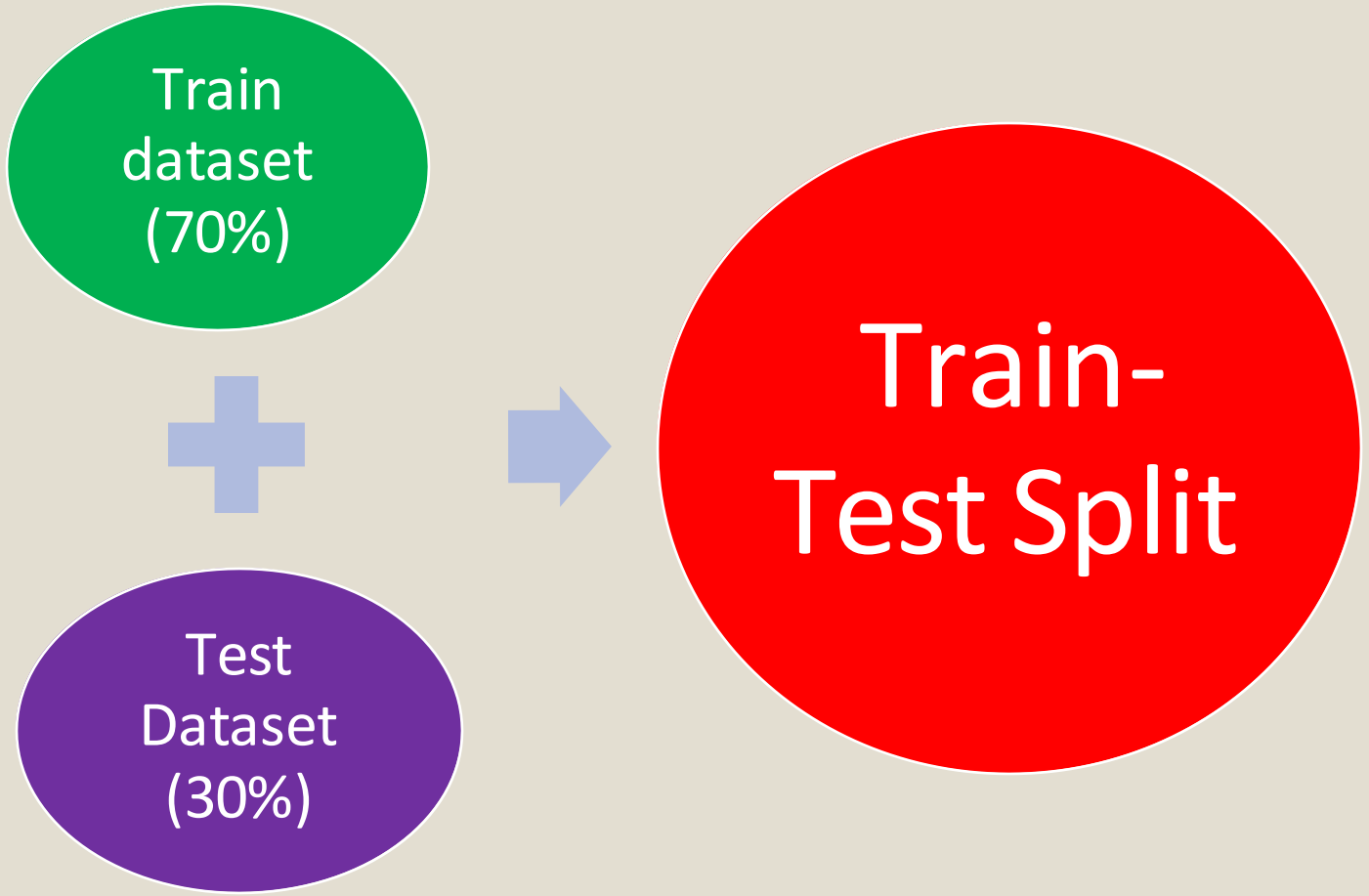
MULTIVARIATE ANALYSIS :



Creation Of the Dummy Variables :-

For logistic regression model all the Categorical Variables are Divided into Dummy variables to change the values to numeric form (0,1).This is done to avoid any Multicollinearity which further leads to unstable estimates of the regression coefficients . That's why by dropping the first category we can avoid creating the redundant feature that is perfectly correlated with other dummy variables .

Creation of Train-Test Split :-



Feature Selection and Model Building :

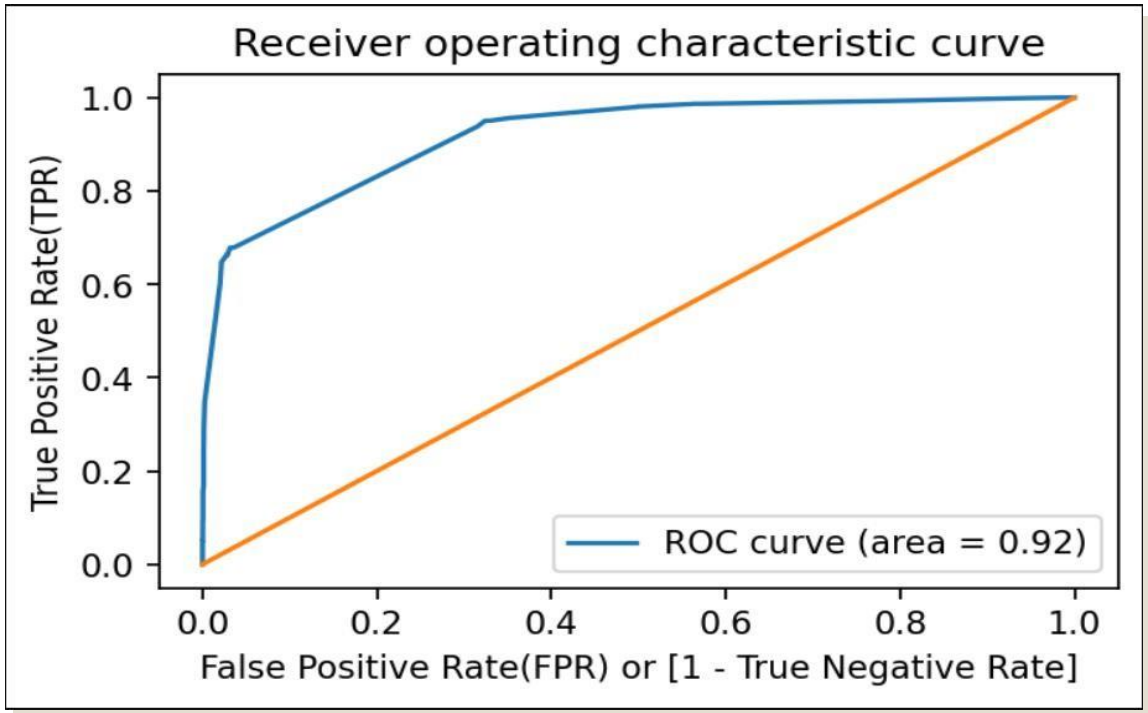
Using Hybrid approach to select Features

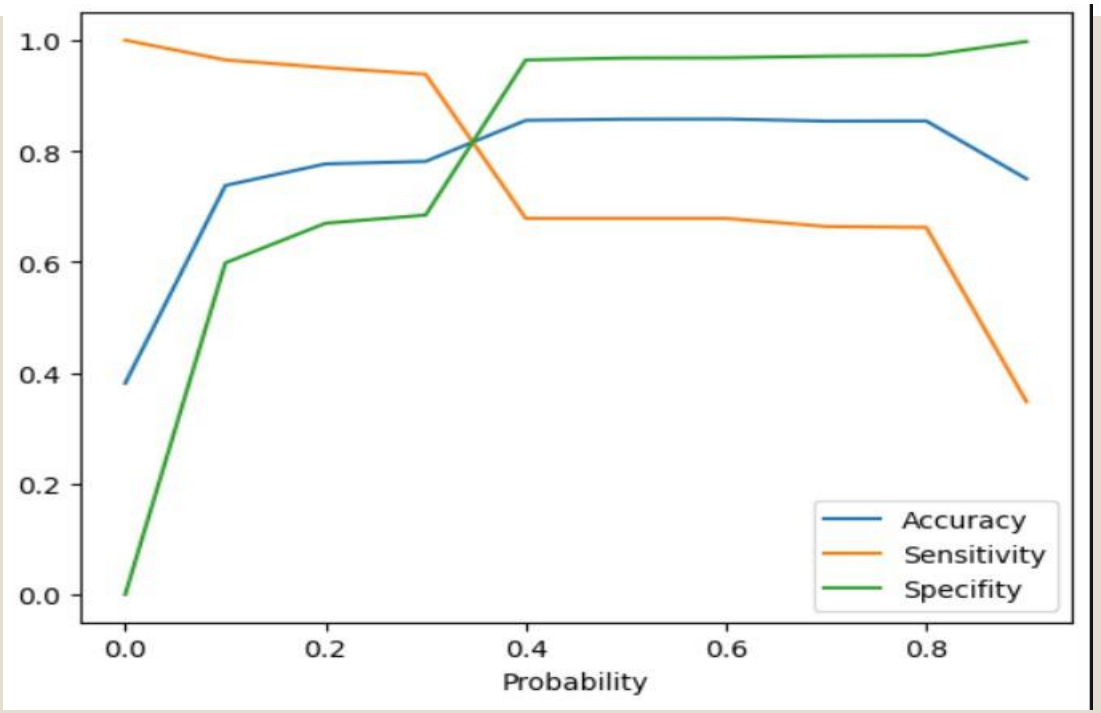
RFE is used to select 15 features

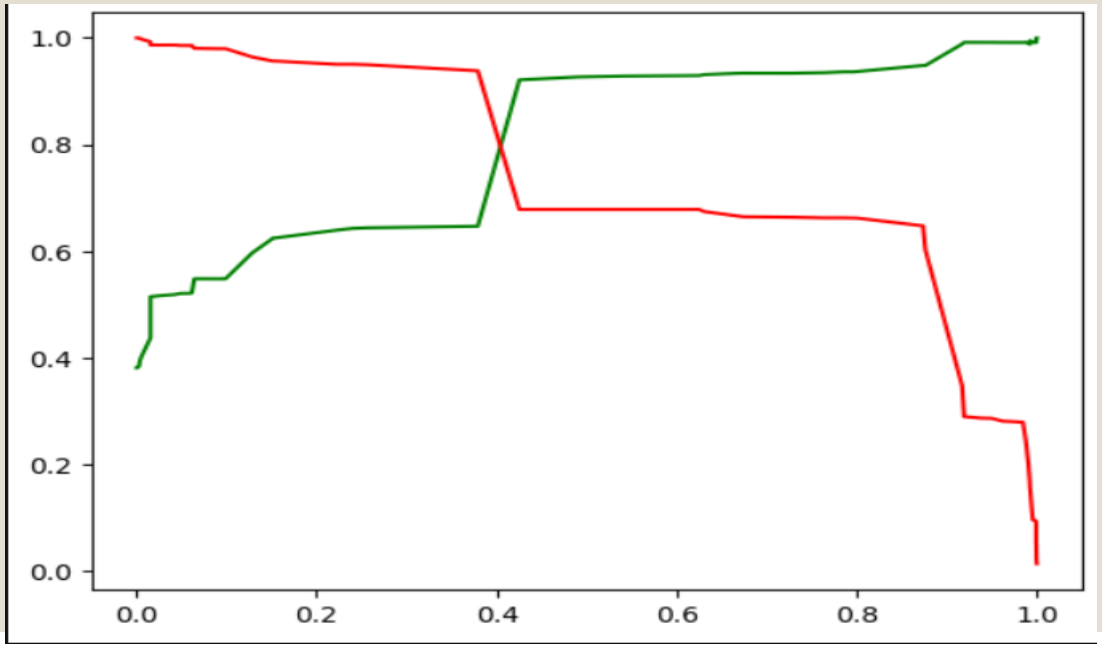
Analysing the p-values and VIFs, if these are more than 0.05 dropped them. At the end ,11 features are found from the Model5.

Model Evaluation and Prediction on the Test Dataset

- **Model Evaluation :-** Predicting from the train dataset and evaluating the Accuracy, Recall , Precision and F1_Score.
- Plotting the Roc curve and then check the optimum cutoff . It came out to be 0.35
- **Prediction the Test Dataset :-** Normalizing the numeric data types in the test dataset. Predicting the Test Dataset and evaluating the Accuracy, Recall, Precision and F1 score
- By comparing the train and test dataset we have found same level of model efficiency .


Receiver operating characteristic curve
True Positive Rate(TPR)
False Positive Rate(FPR) or [1 - True Negative Rate]
ROC curve (area = 0.92)
0.0
0.2
0.4
0.6
0.8
1.0


Accuracy
Sensitivity
Specifity
Probability
0.0
0.2
0.4
0.6
0.8
1.0


0.0
0.2
0.4
0.6
0.8
1.0

CONCLUSION :-

Keys variables to identify the hot leads are :-

- **When the customer is a working professional, it has a high chance of conversion**
- **When the Lead origin is Lead add form .**
- **Last Activity is identified as Olark Chat Conversion and SMS sent .**
- **When customer has permitted for email.**
- **When customer is tagged as ‘lost to EINS’ :, closed by horizon ,’ Will revert after reading mail, ’In touch with EINS ‘.**

Through EDA we can see the ‘time spent on websites ‘,’ total visits ‘,’lead source as Google ’ seems to give fruitful results .

Recommendations

Some of the quick suggestions for reducing effort for sales team are as follow :-

- Phone calls must be done to people who are spending more of a time on the websites, filling forms, coming back with queries .
- The one who have permitted to call as well as emails .
- Reverting back as ping you after reading it .
- Target the working professionals .
- Use the automated SMS service which can reduce the calls counts .