

# **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 .



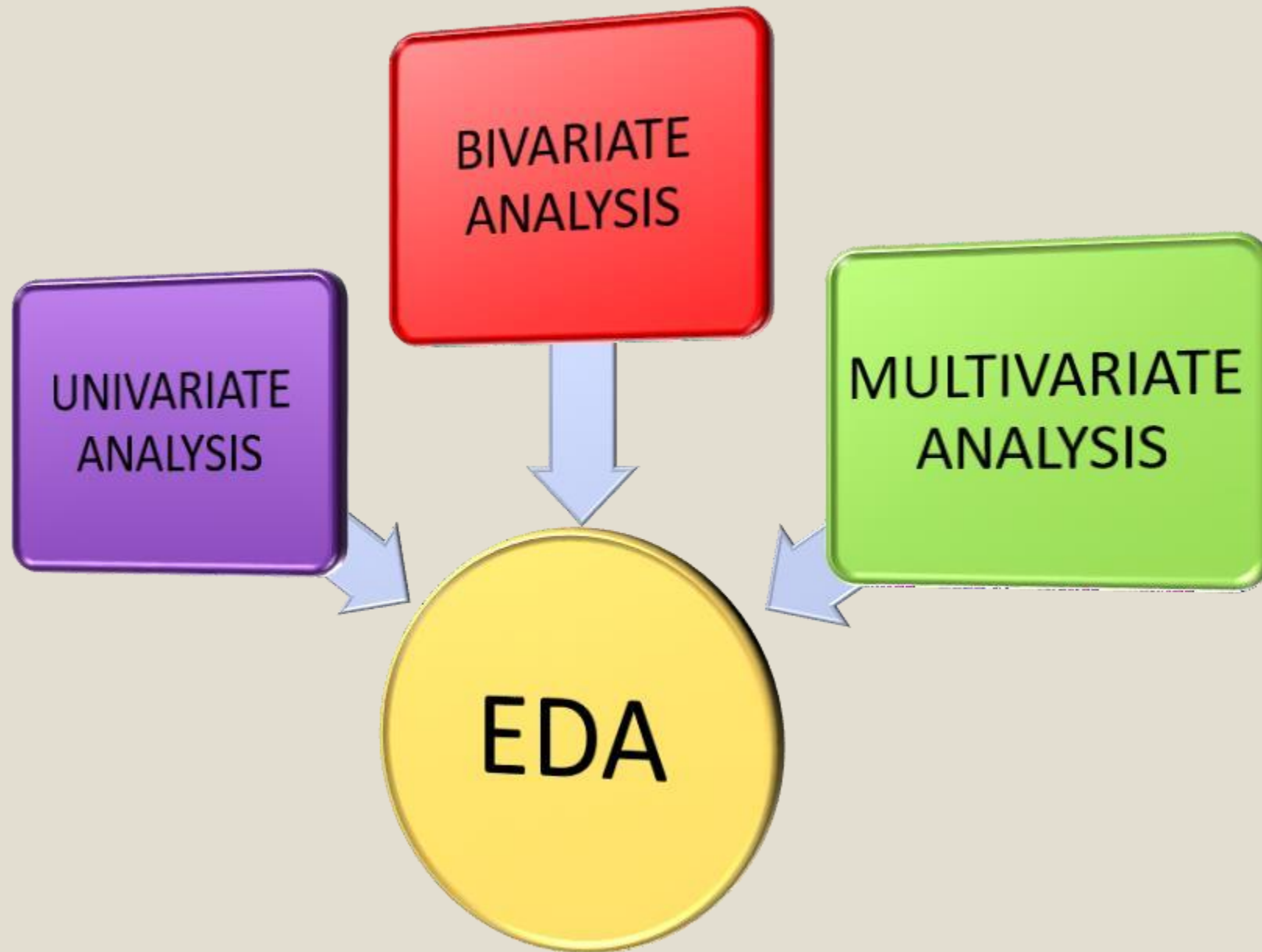
2.

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



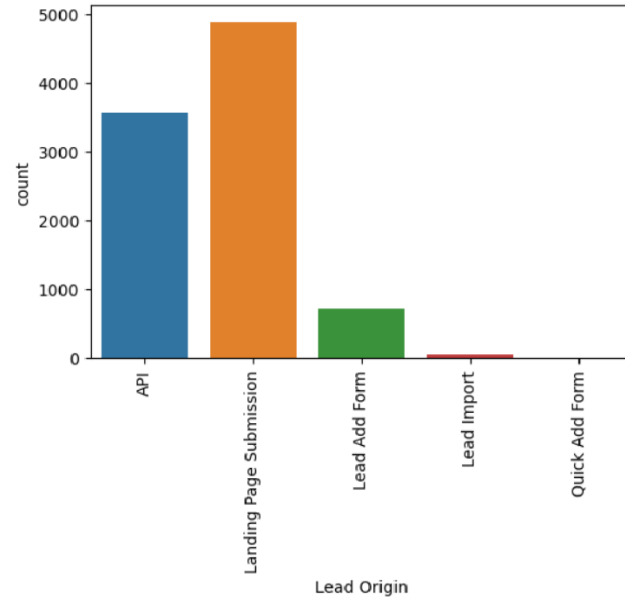
3.

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

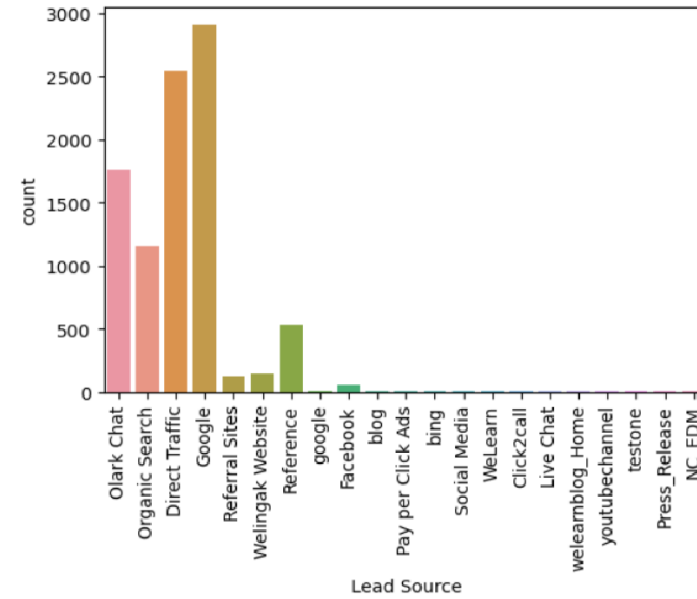


# UNIVARIATE ANALYSIS:-

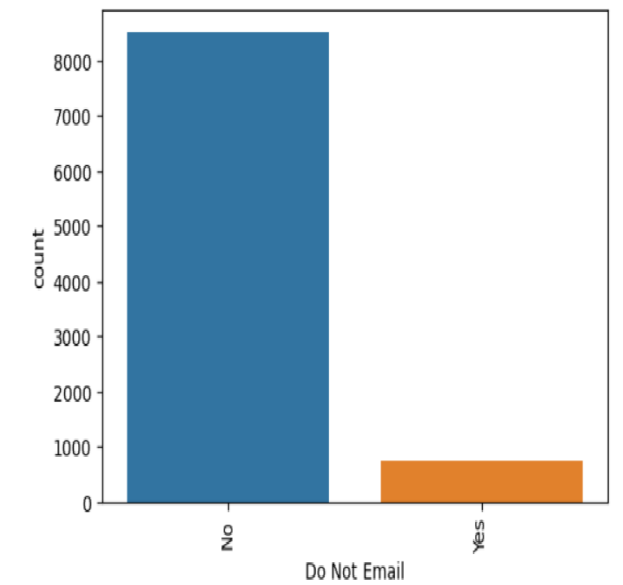
Countplot of Lead Origin



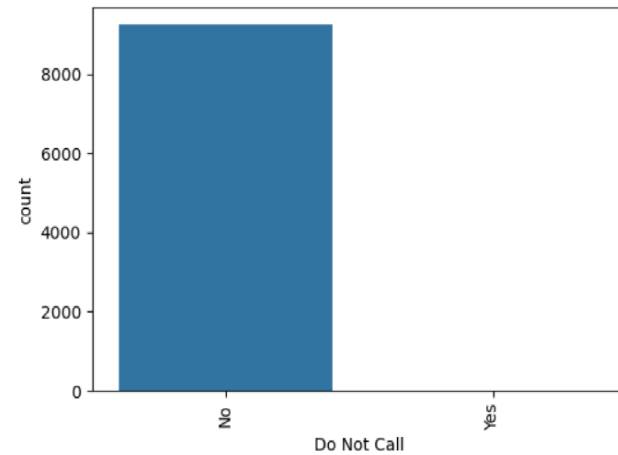
Countplot of Lead Source



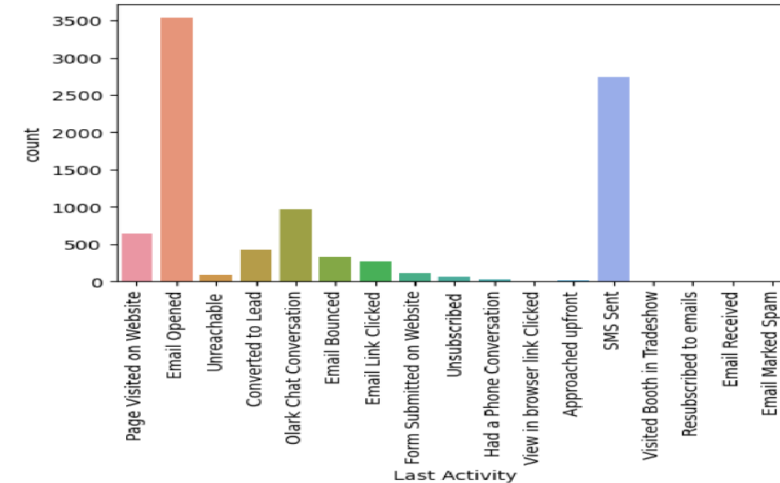
Countplot of Do Not Email



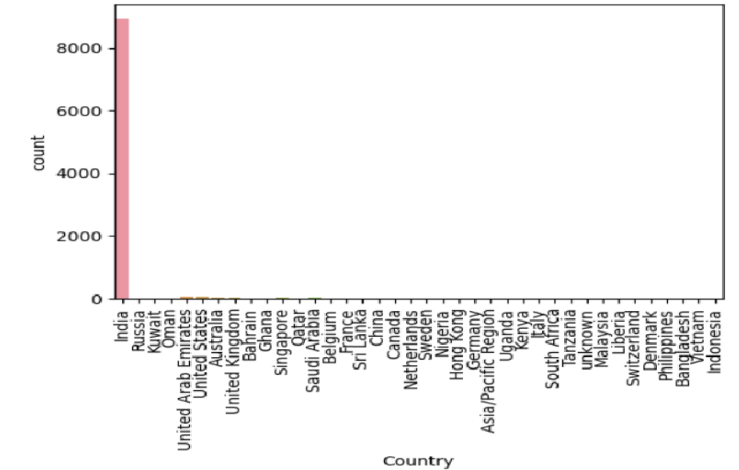
Countplot of Do Not Call



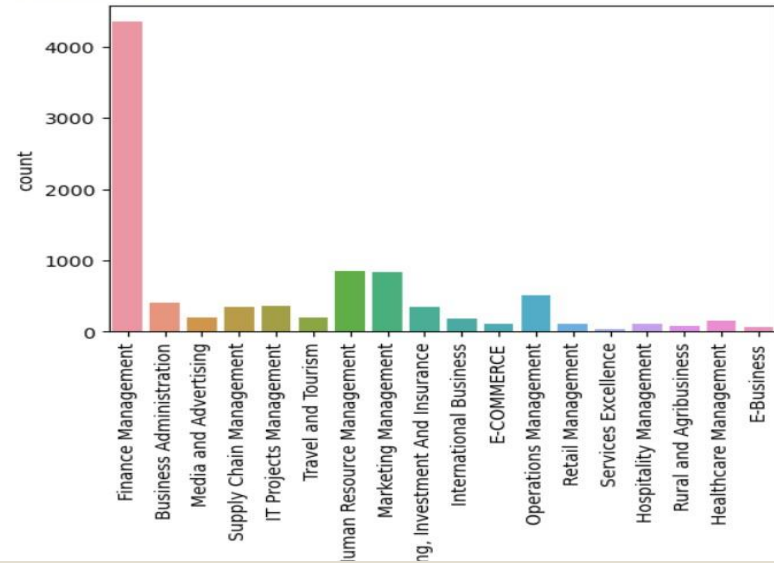
Countplot of Last Activity



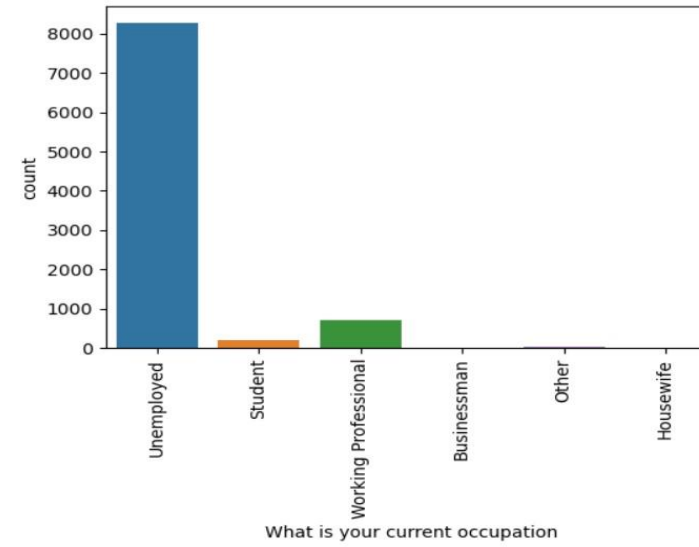
Countplot of Country



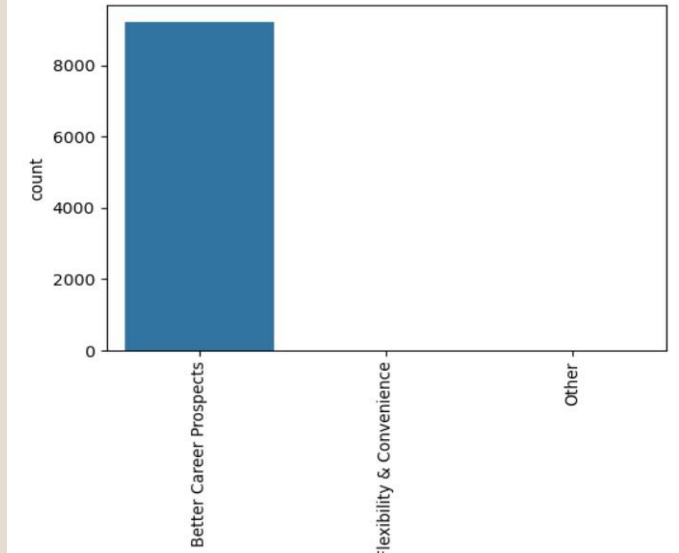
Countplot of Specialization



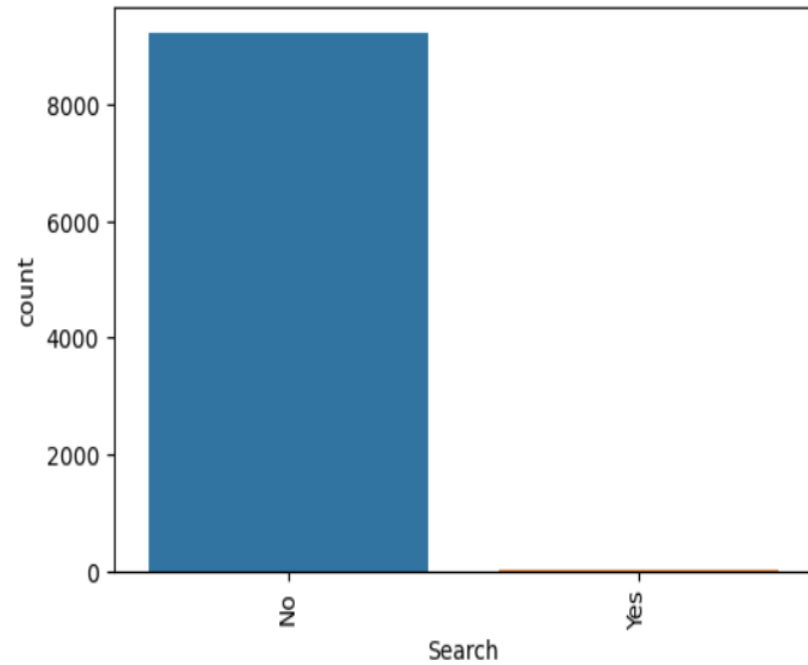
Countplot of What is your current occupation



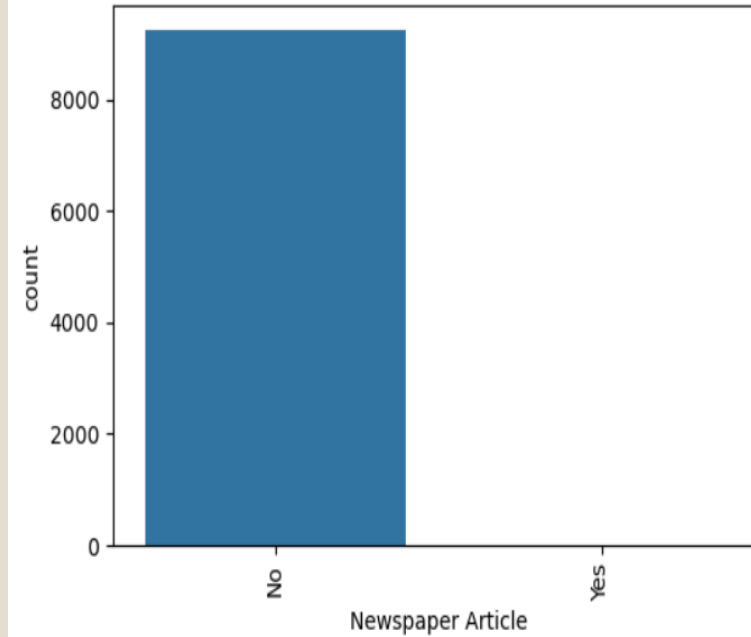
Countplot of What matters most to you in choosing a course



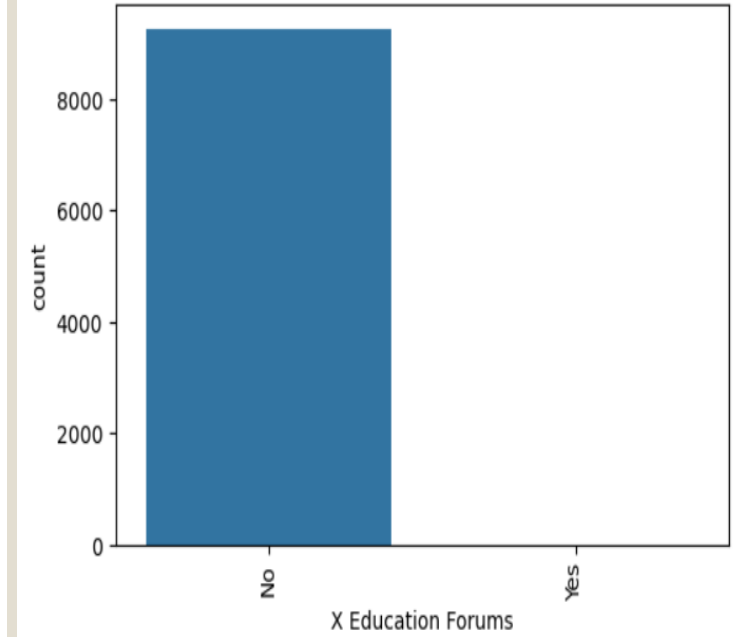
Countplot of Search



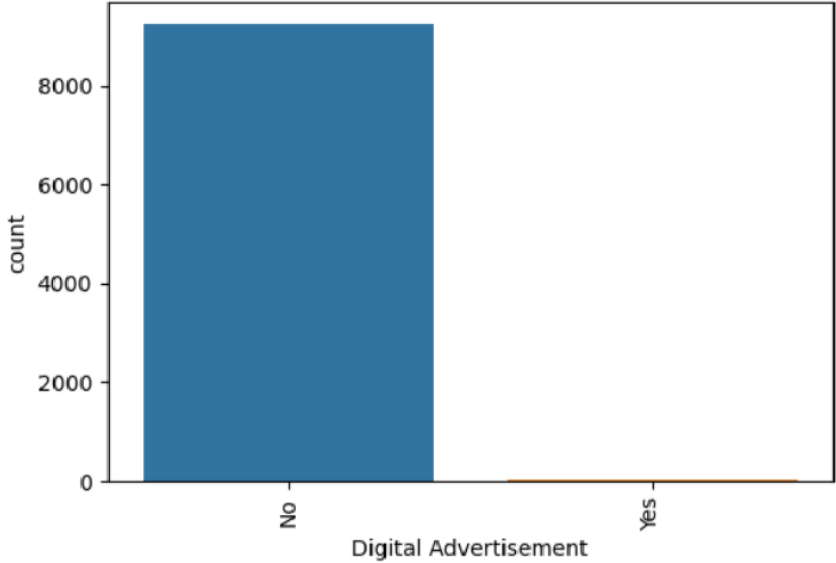
Countplot of Newspaper Article



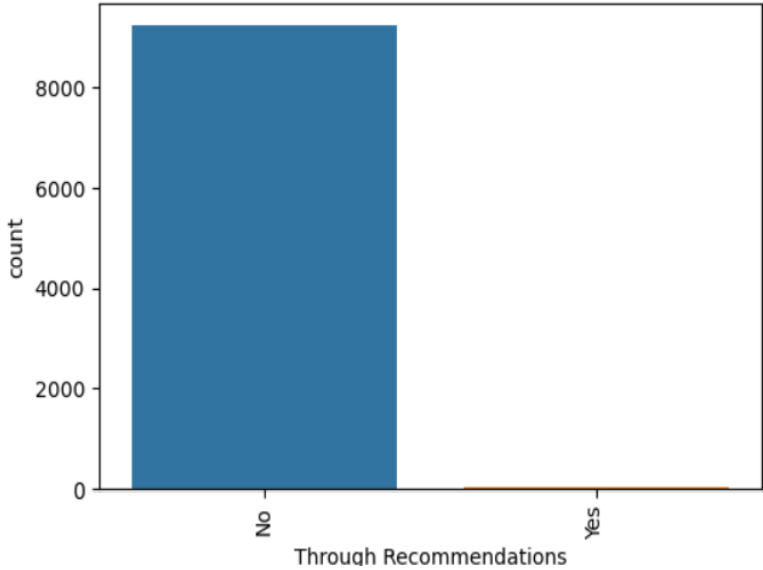
Countplot of X Education Forums



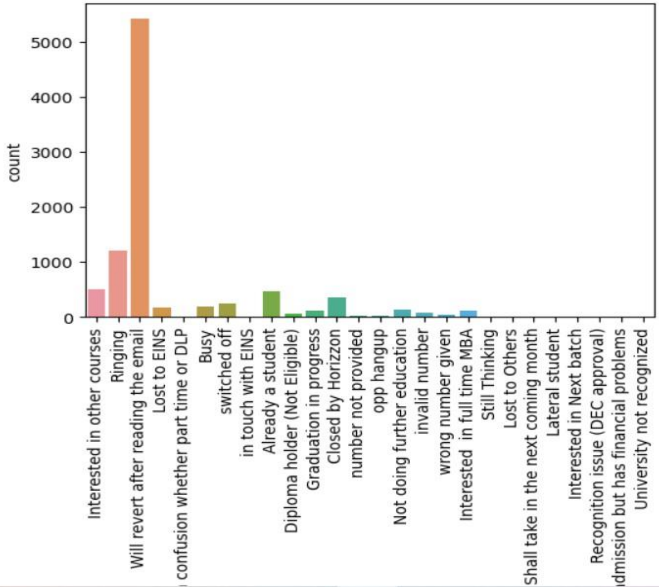
Countplot of Digital Advertisement



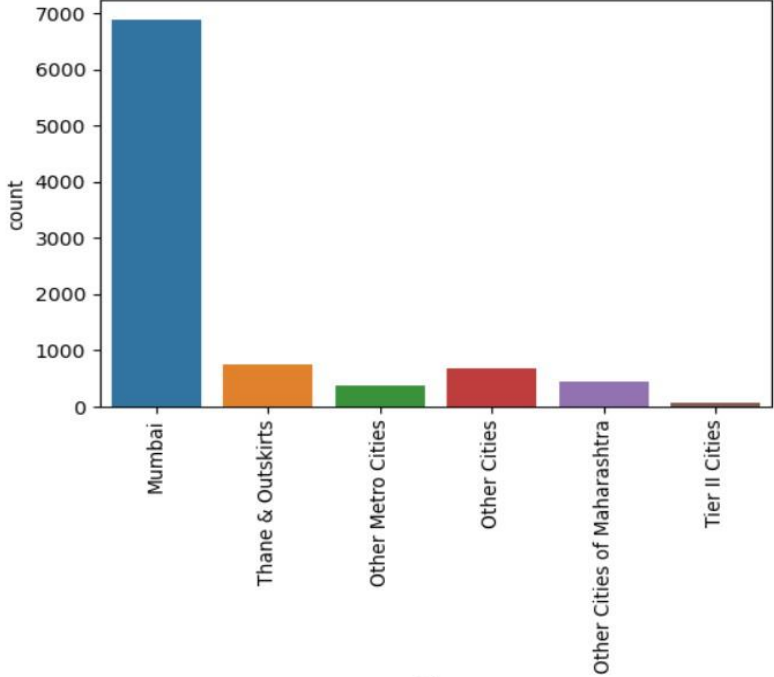
Countplot of Through Recommendations



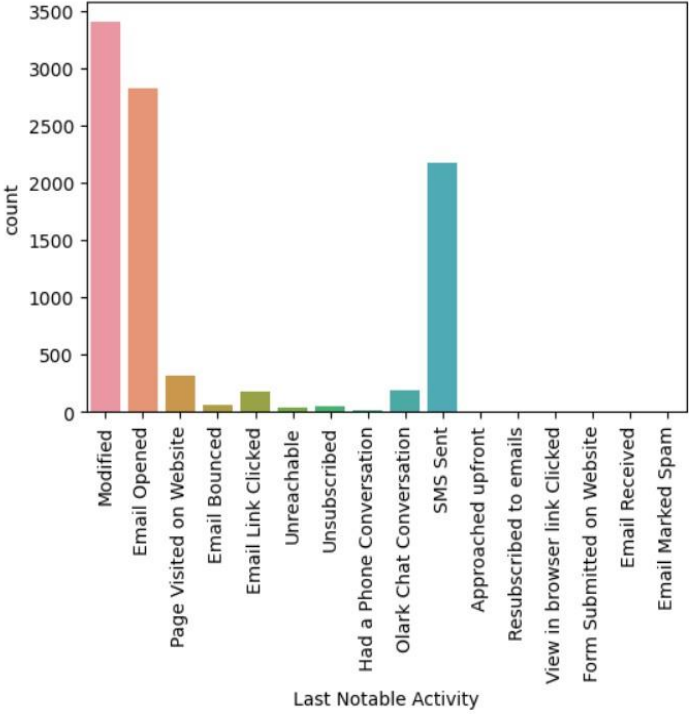
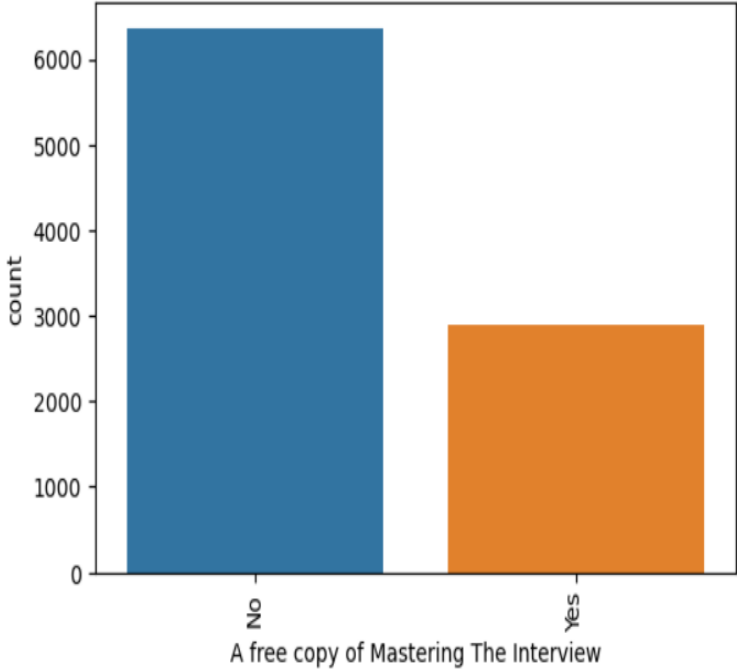
Countplot of Tags



Countplot of City



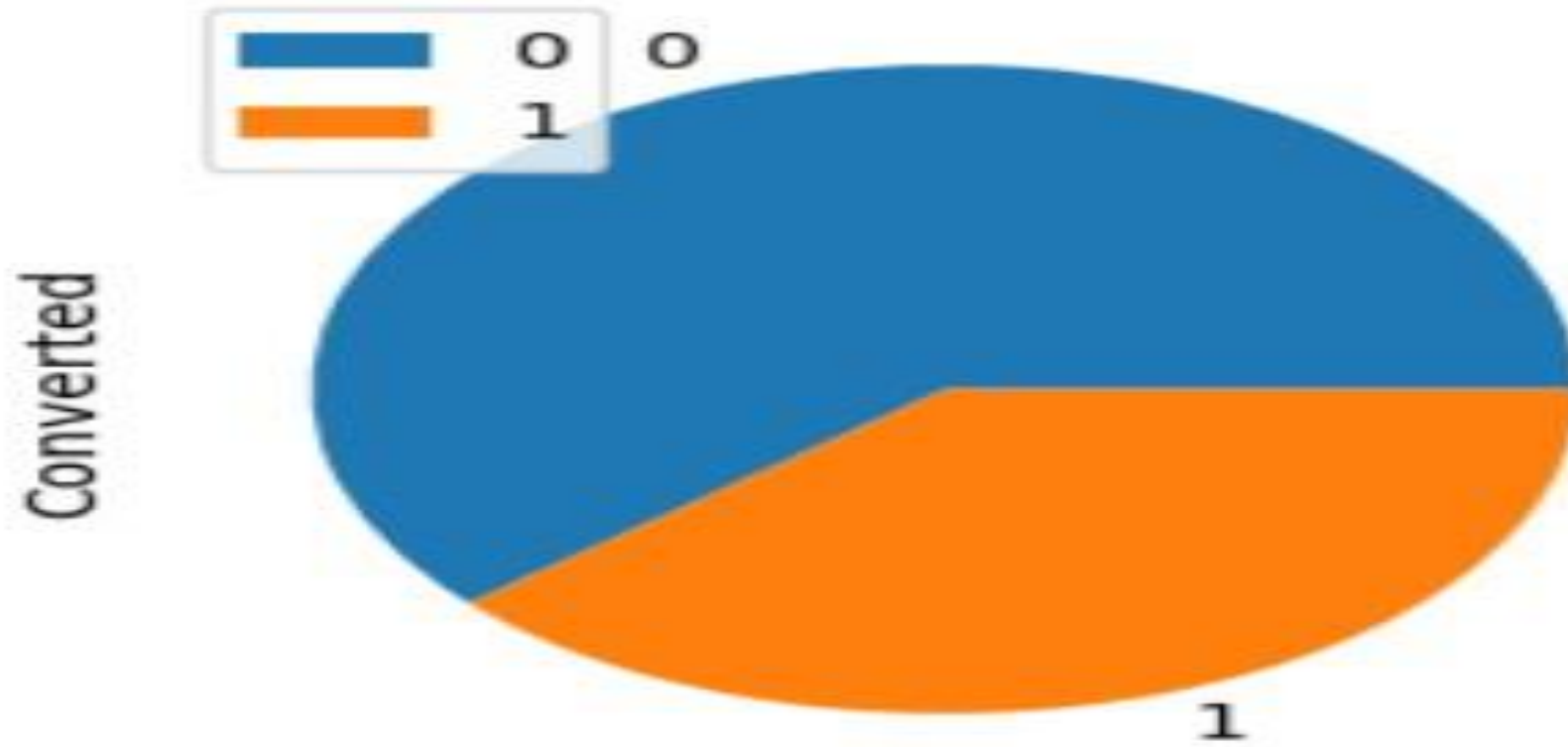
Countplot of A free copy of Mastering The Interview



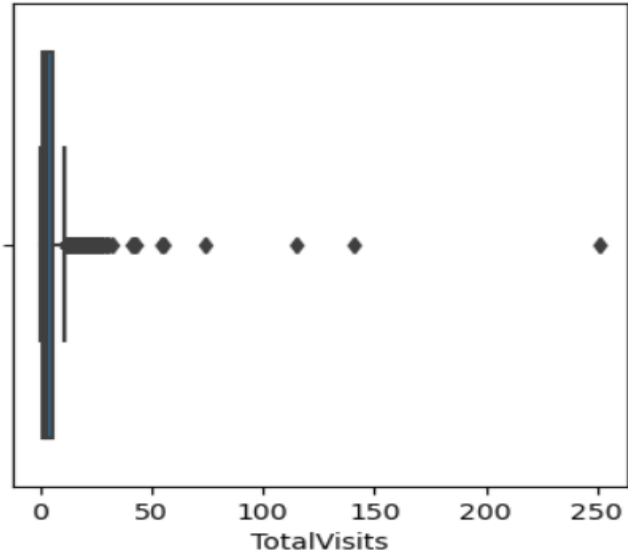


# Analysing the Target Variable :-

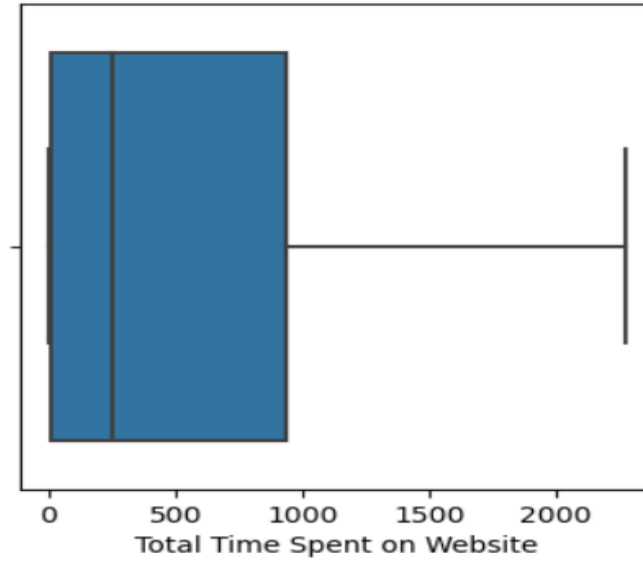
```
0    61.461039  
1    38.538961  
Name: Converted, dtype: float64
```



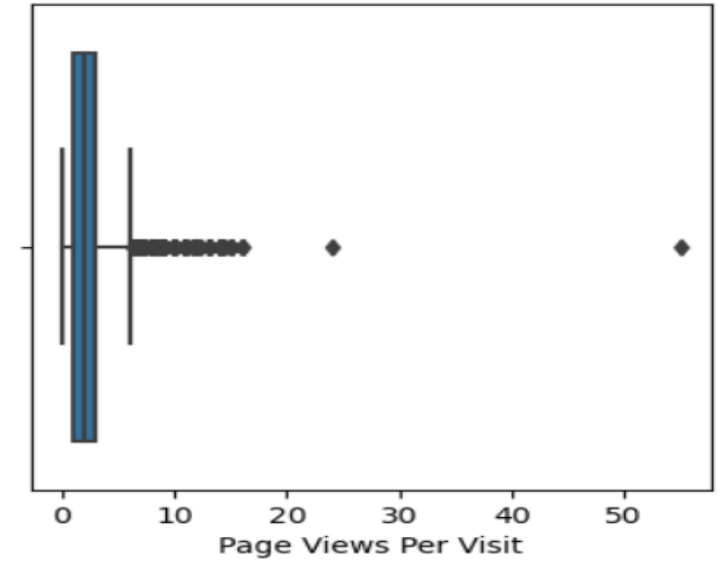
Boxplot of TotalVisits



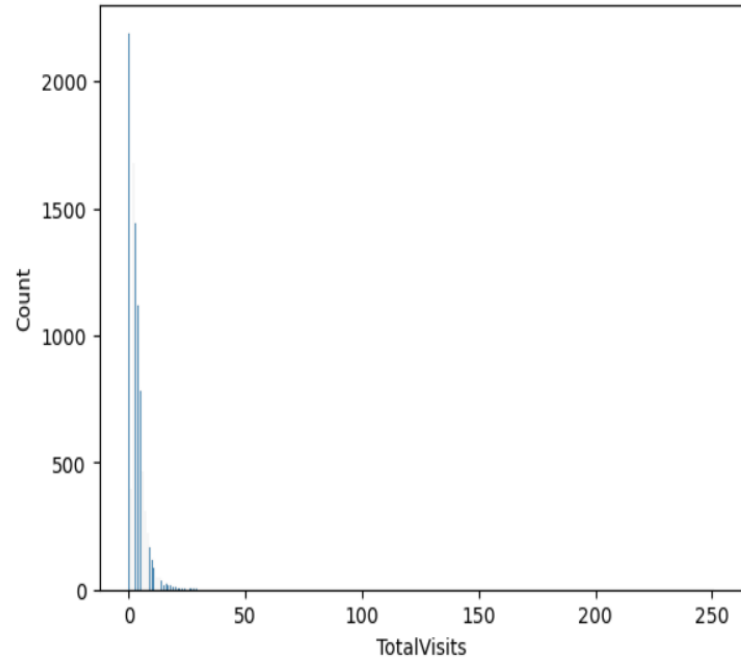
Boxplot of Total Time Spent on Website



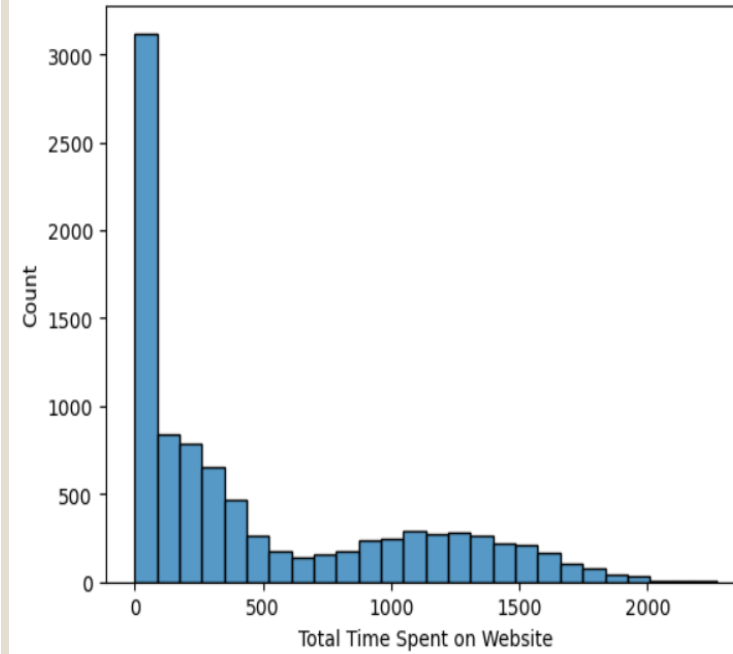
Boxplot of Page Views Per Visit



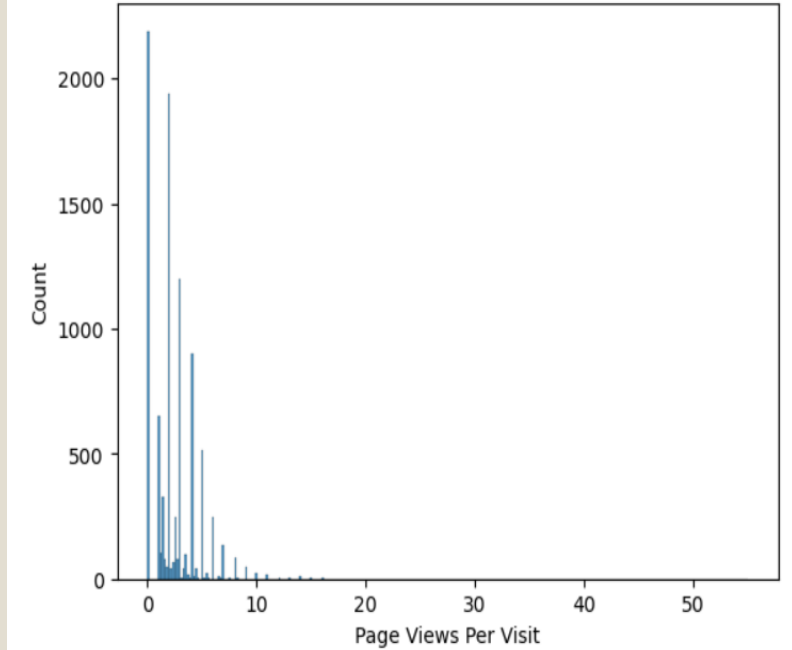
Histplot of TotalVisits



Histplot of Total Time Spent on Website

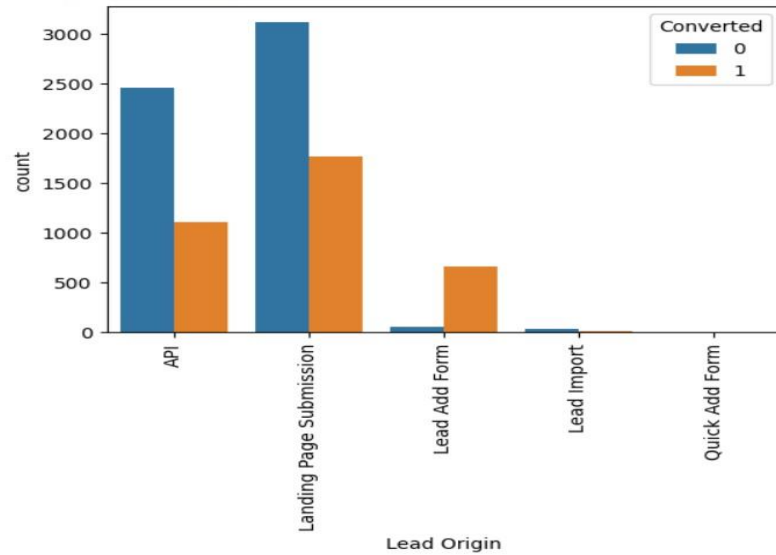


Histplot of Page Views Per Visit

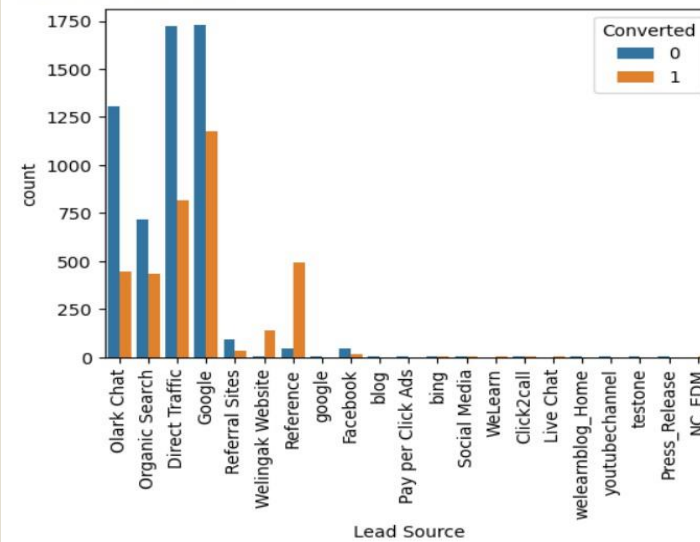


# BIVARIATE ANALYSIS:-

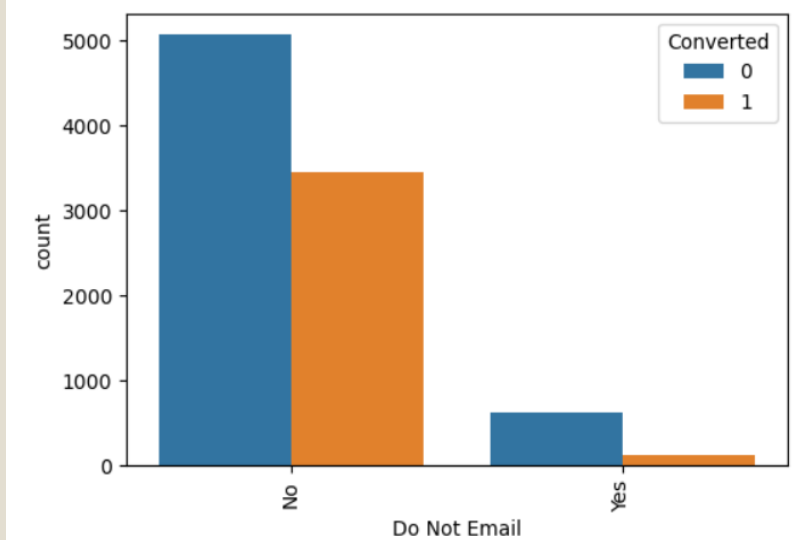
Countplot of Lead Origin



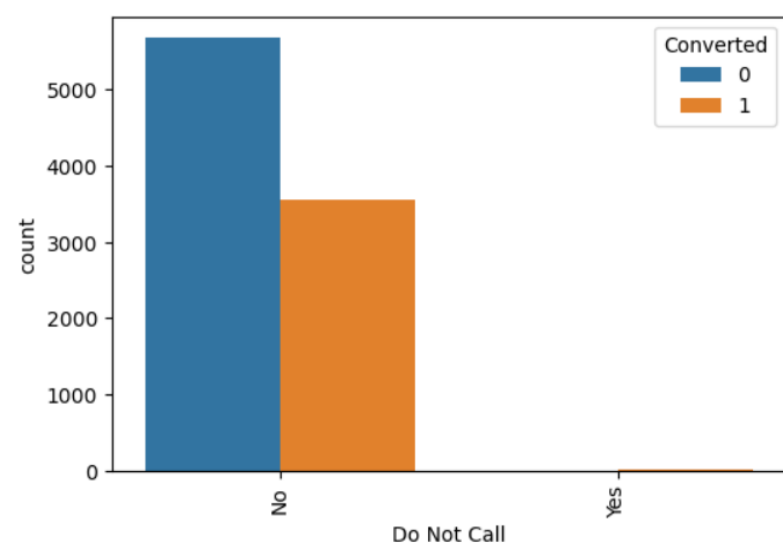
Countplot of Lead Source



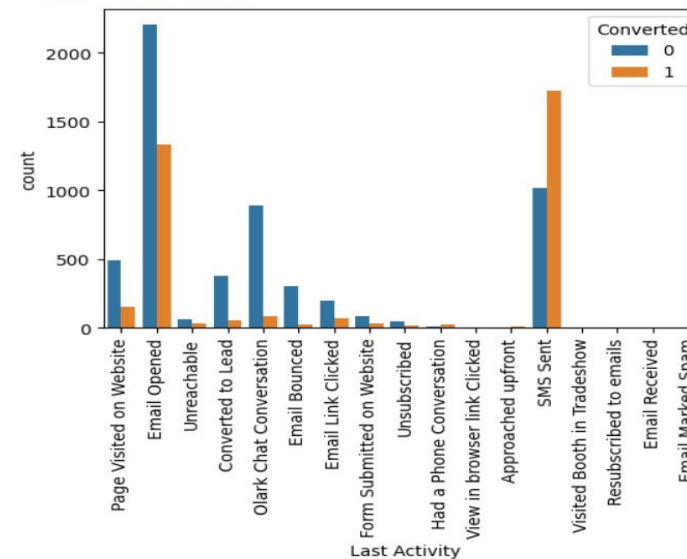
Countplot of Do Not Email



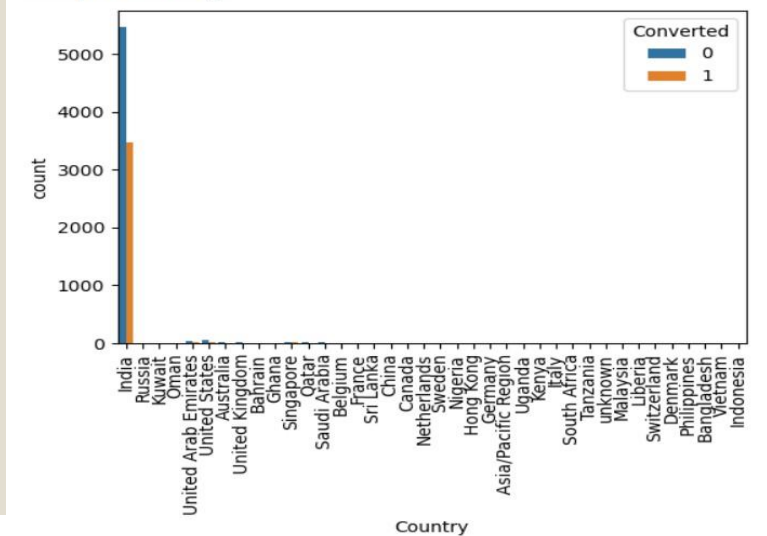
Countplot of Do Not Call



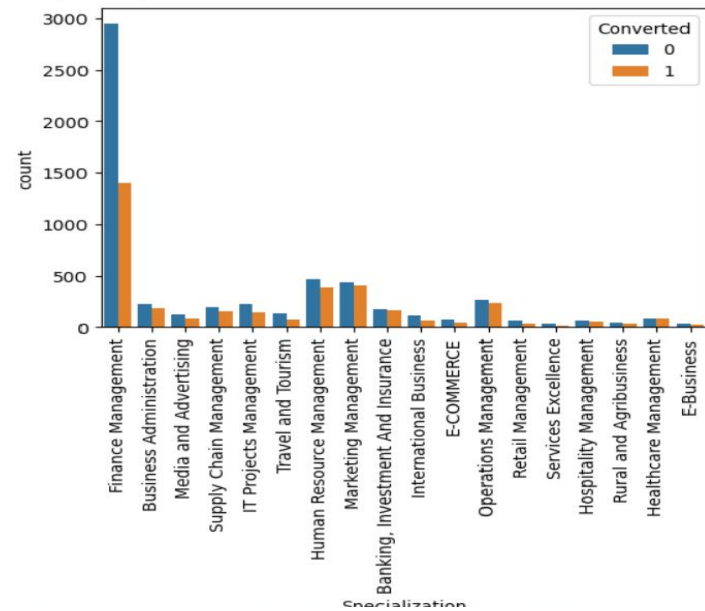
Countplot of Last Activity



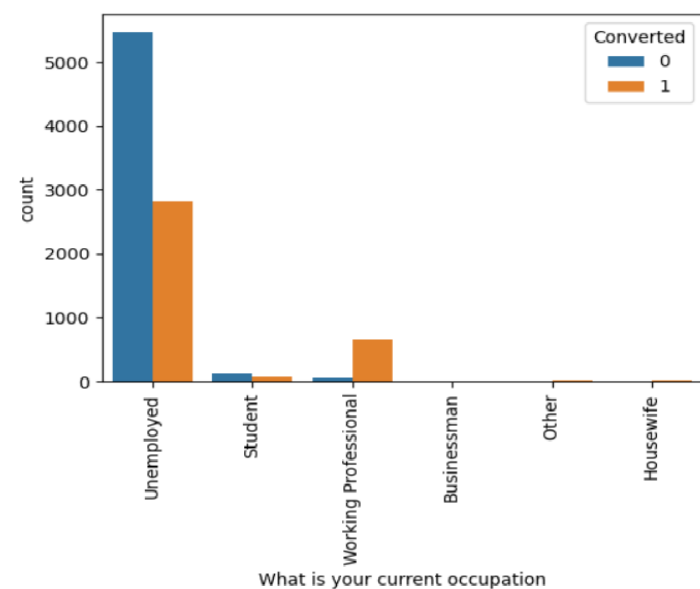
Countplot of Country



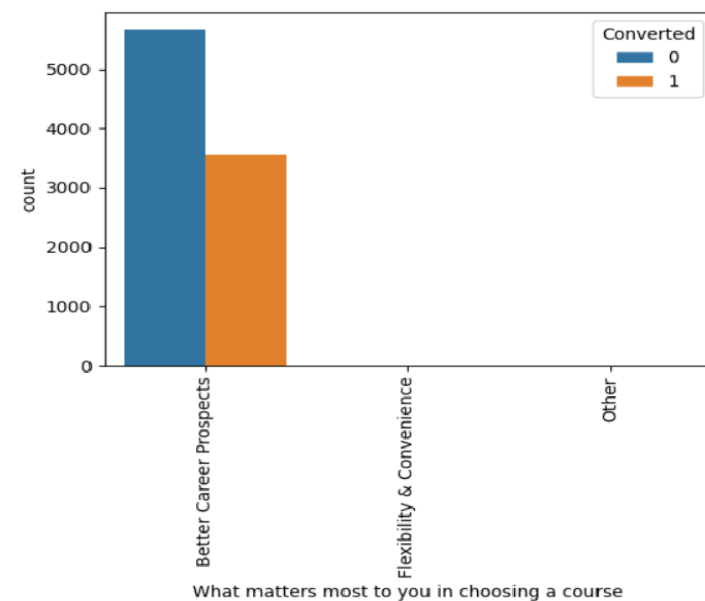
Countplot of Specialization



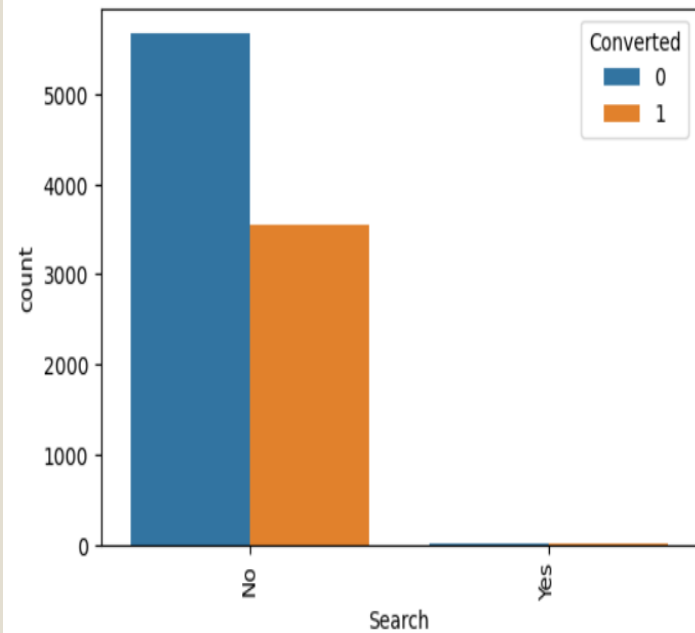
Countplot of What is your current occupation



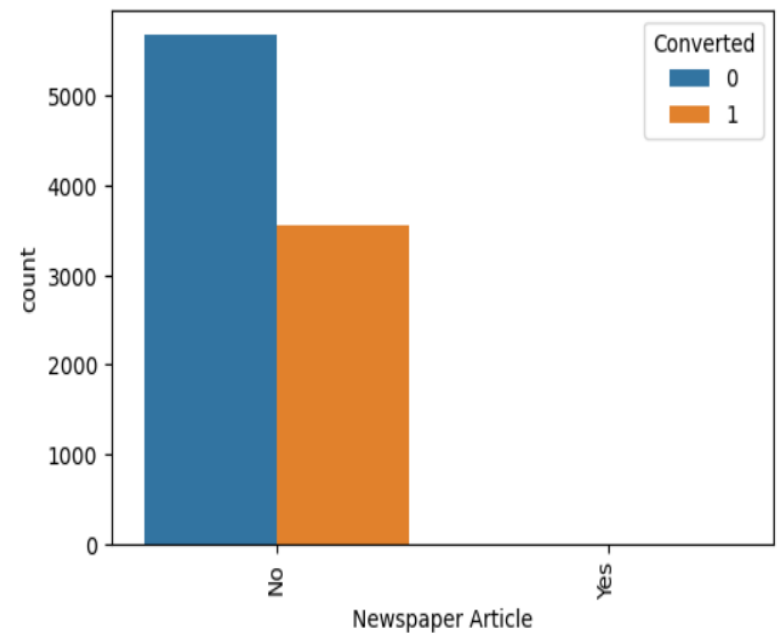
Countplot of What matters most to you in choosing a course



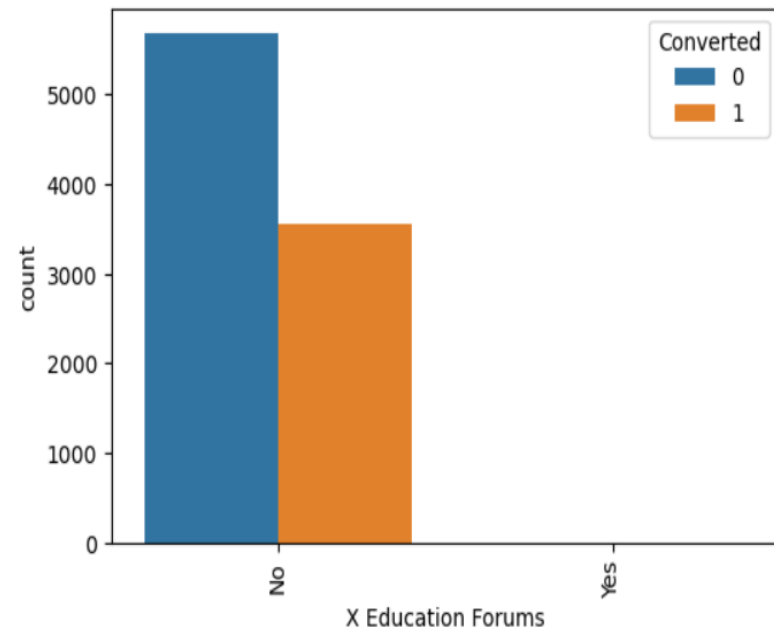
Countplot of Search



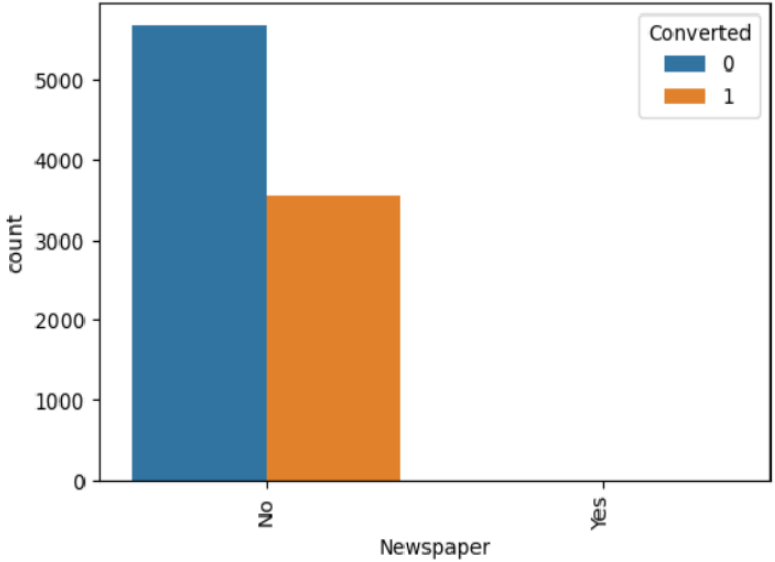
Countplot of Newspaper Article



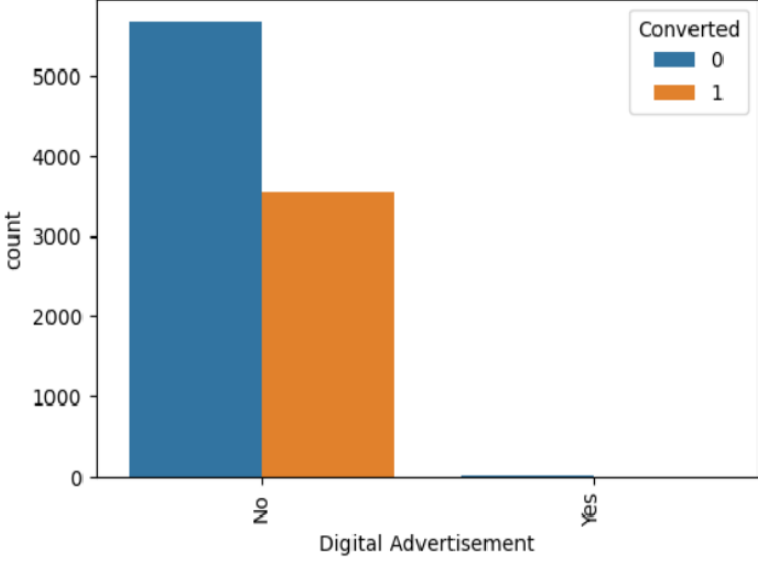
Countplot of X Education Forums



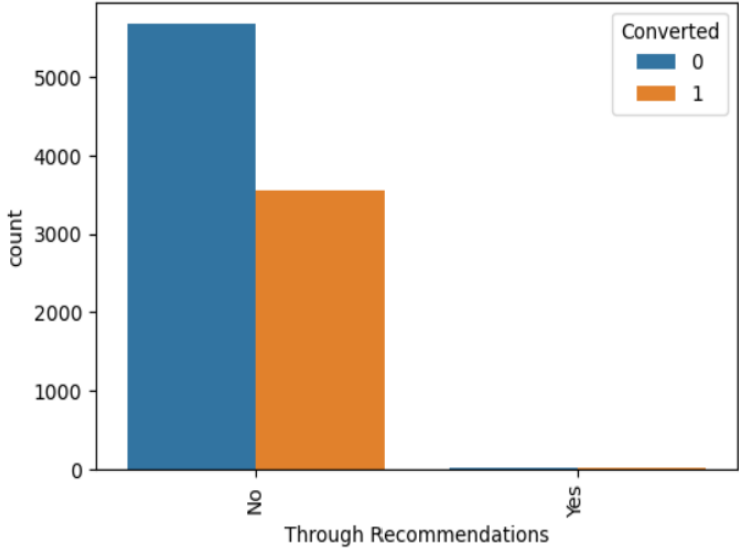
Countplot of Newspaper



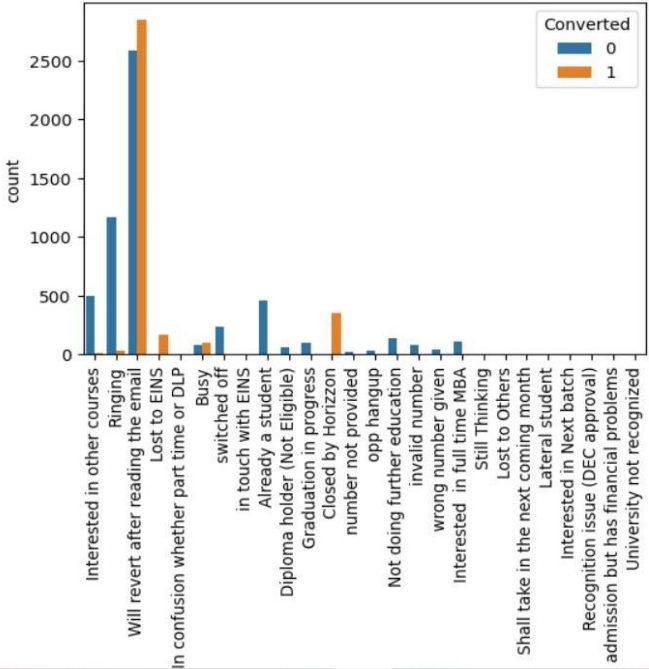
Countplot of Digital Advertisement



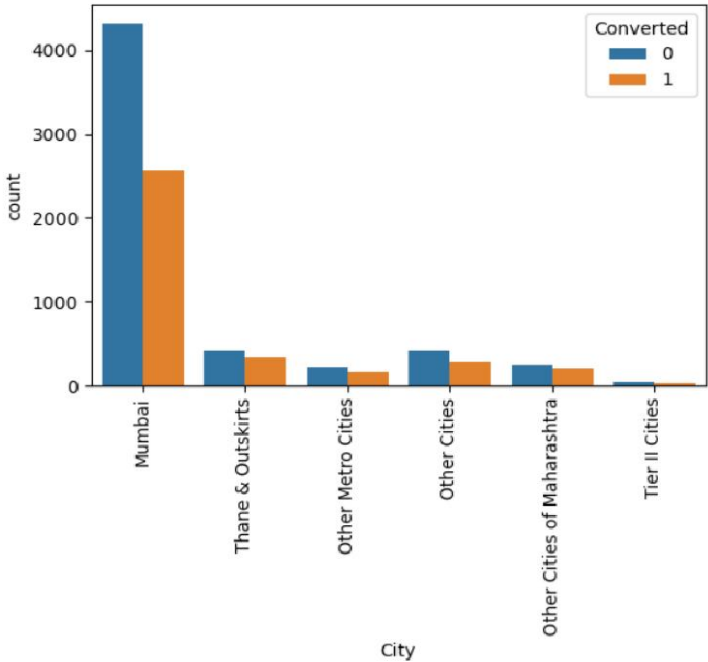
Countplot of Through Recommendations



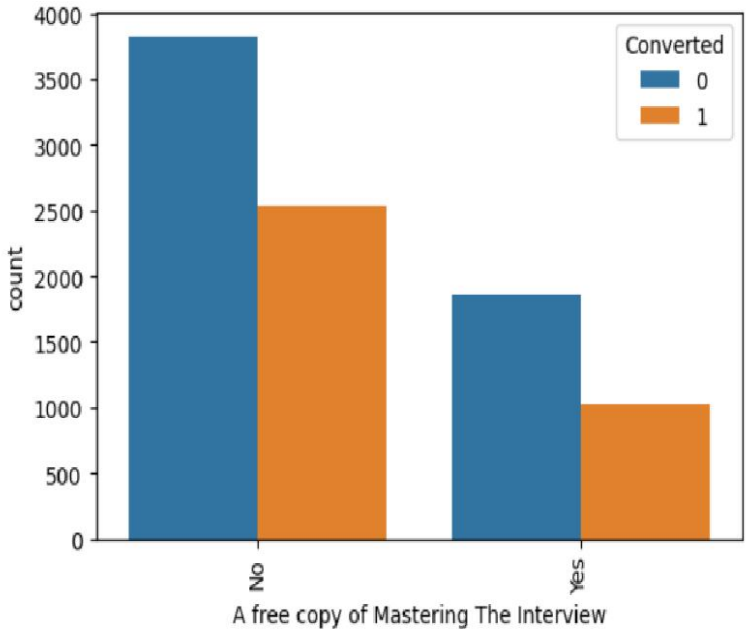
Countplot of Tags



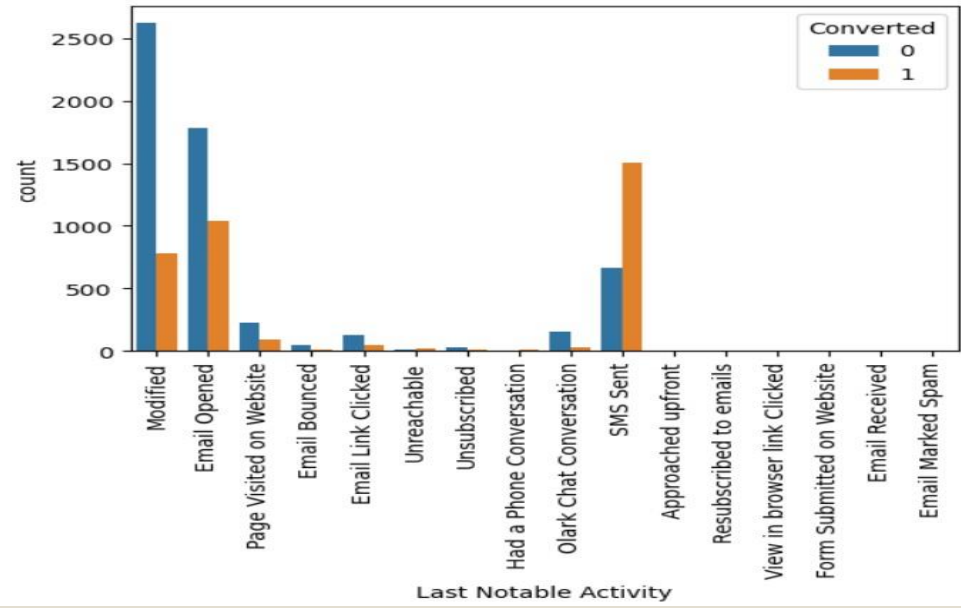
Countplot of City



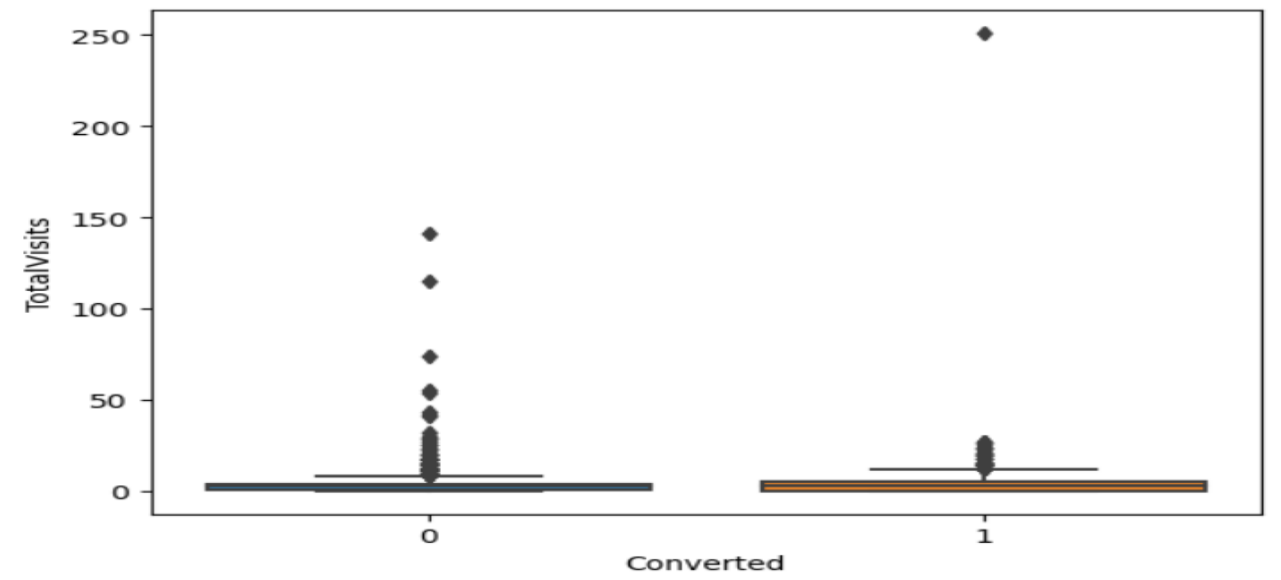
Countplot of A free copy of Mastering The Interview



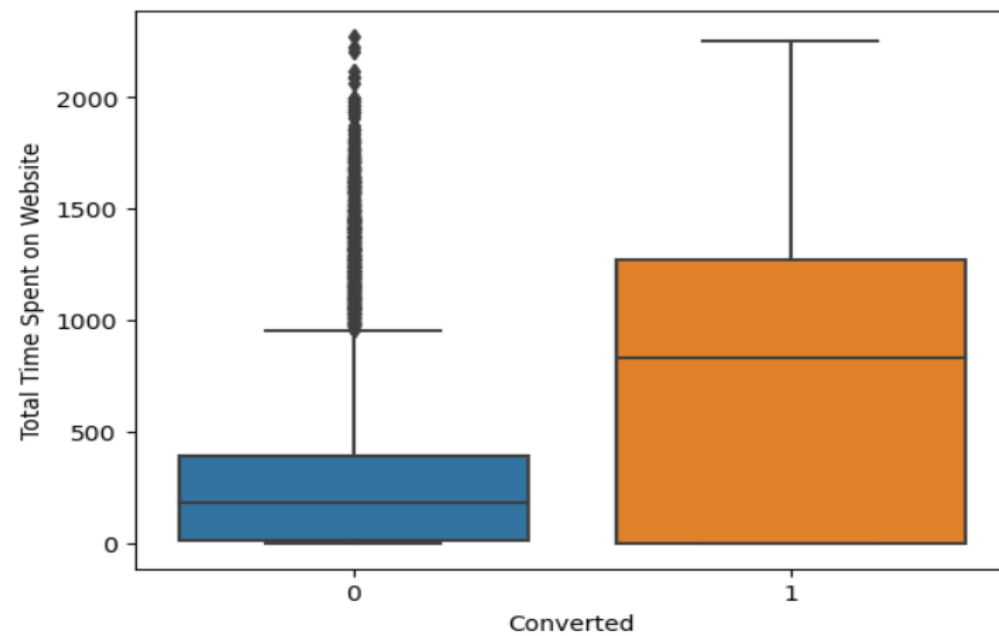
Countplot of Last Notable Activity



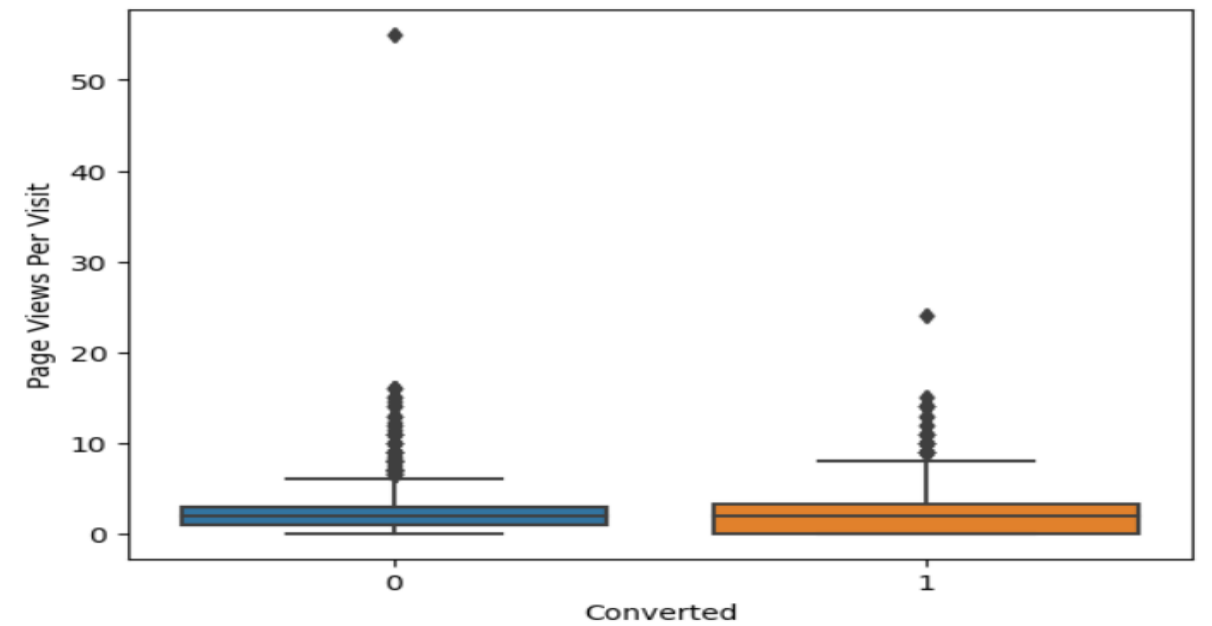
TotalVisits Vs Converted



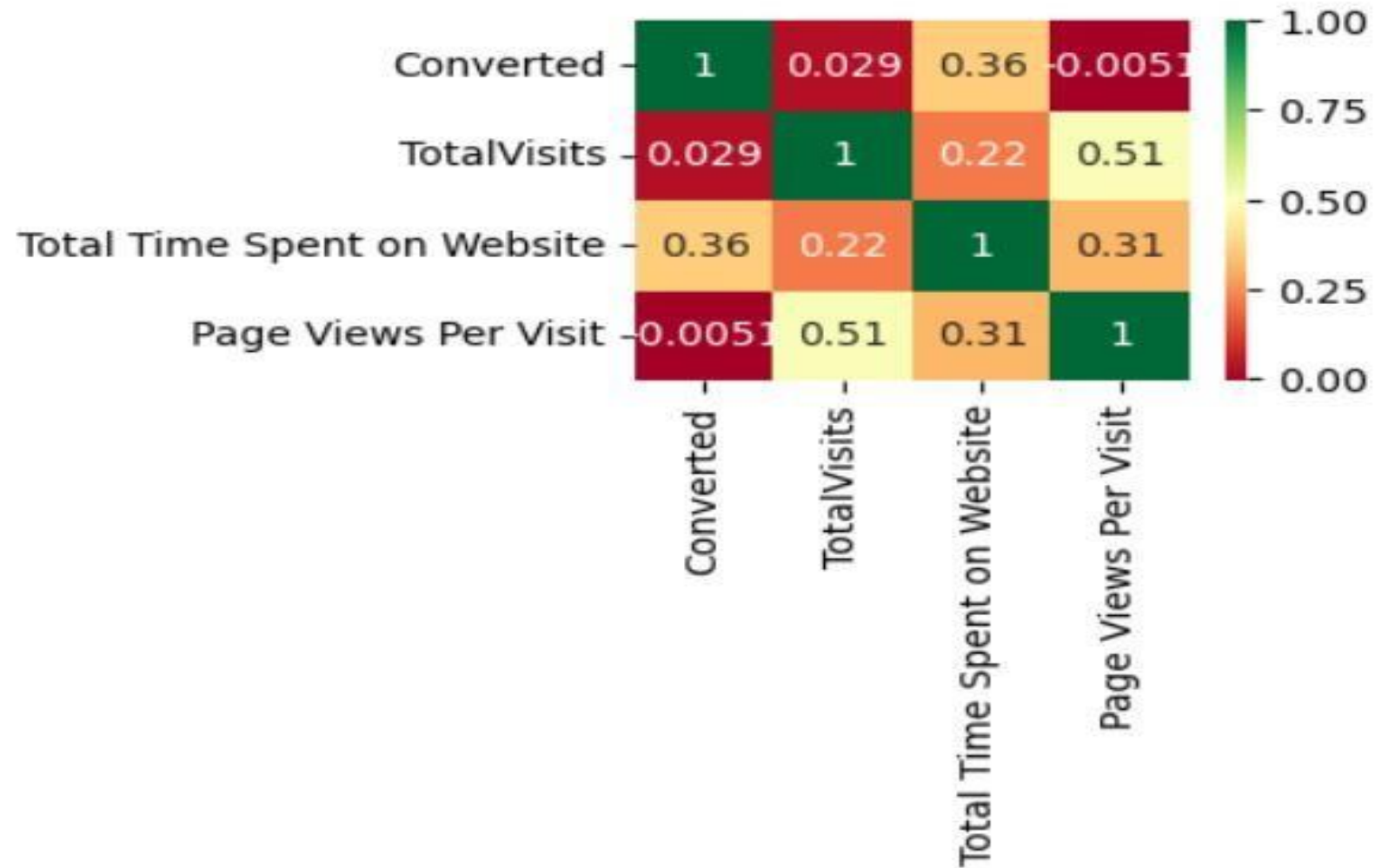
Total Time Spent on Website Vs Converted



Page Views Per Visit Vs Converted



# MULTIVARIATE ANALYSIS :-

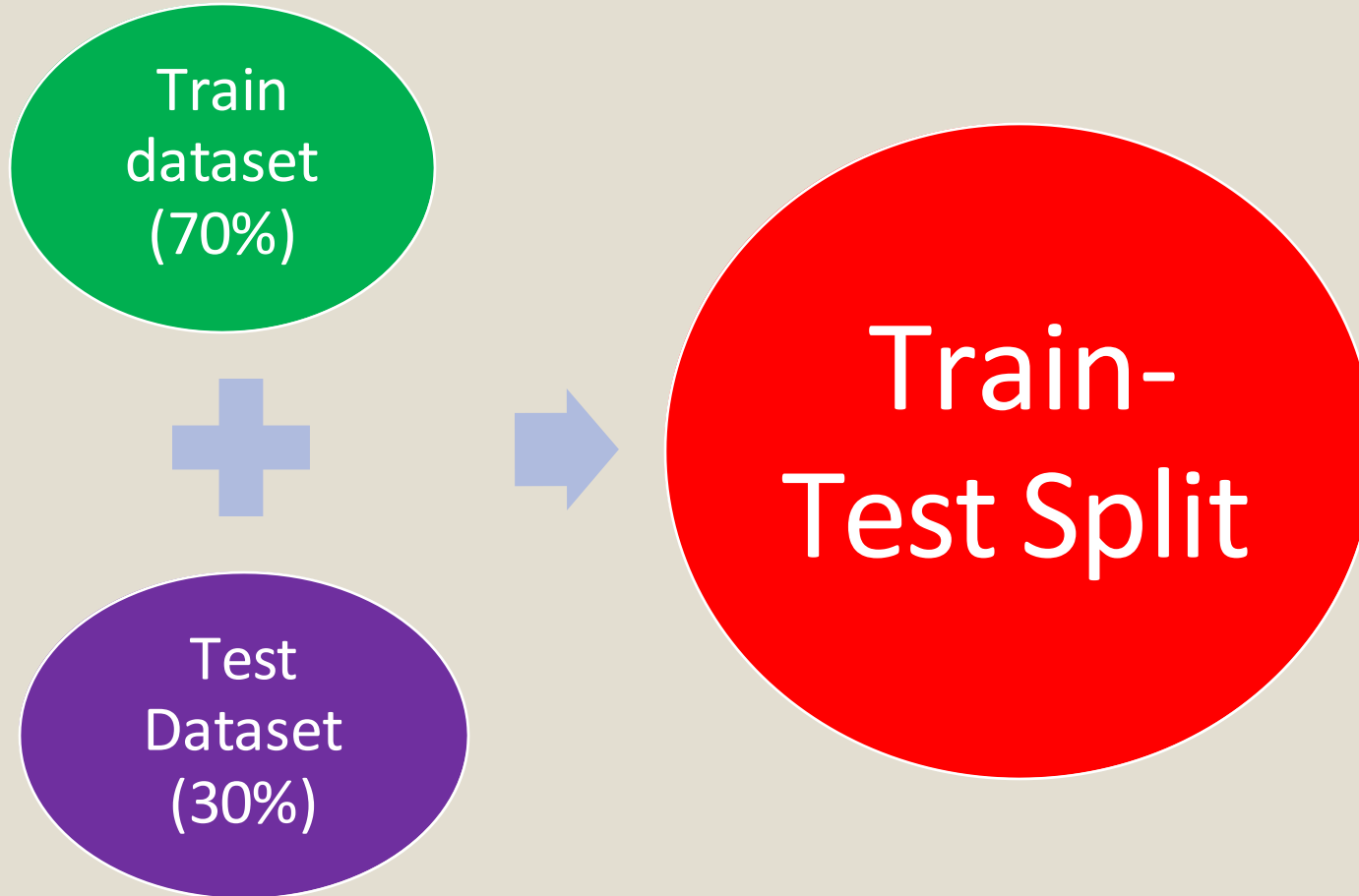


# Creation Of the Dummy Variables :-

**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

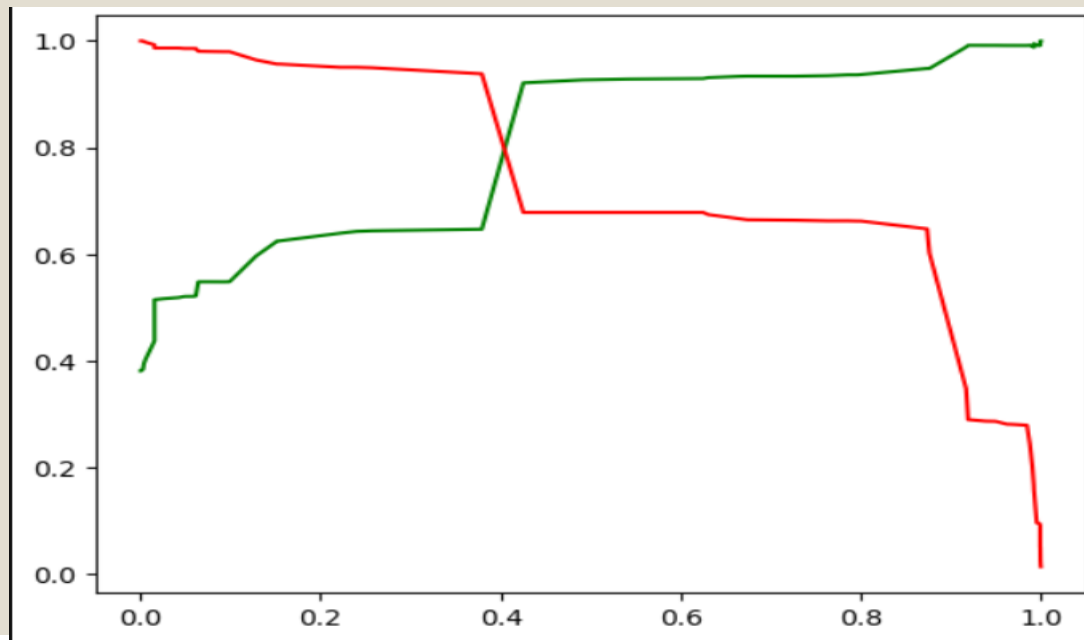
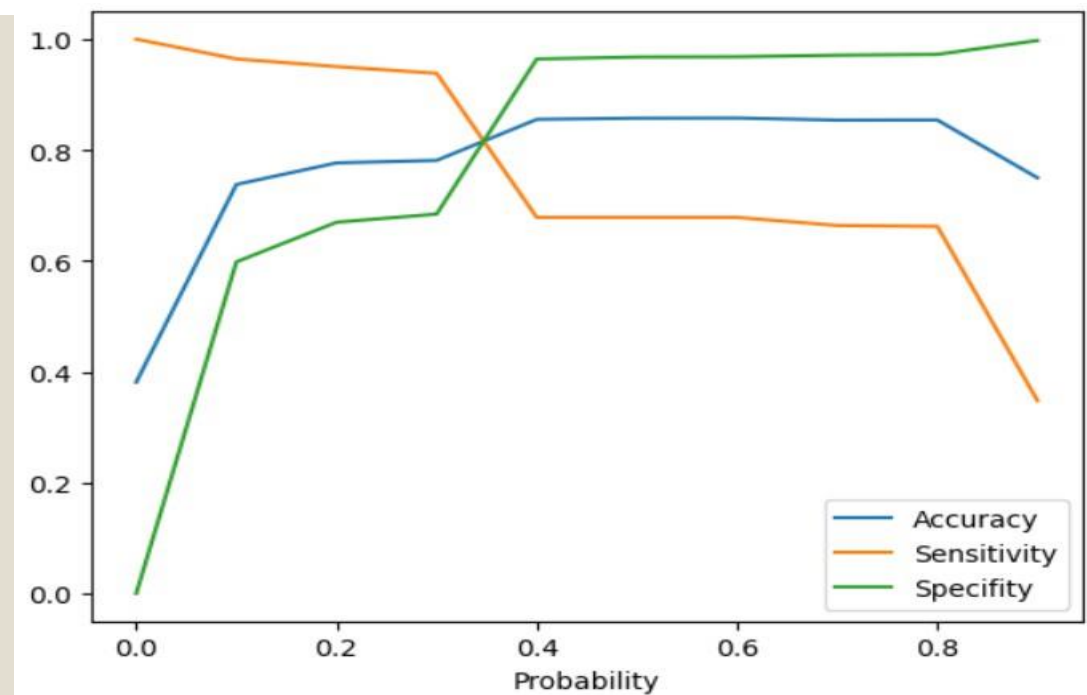
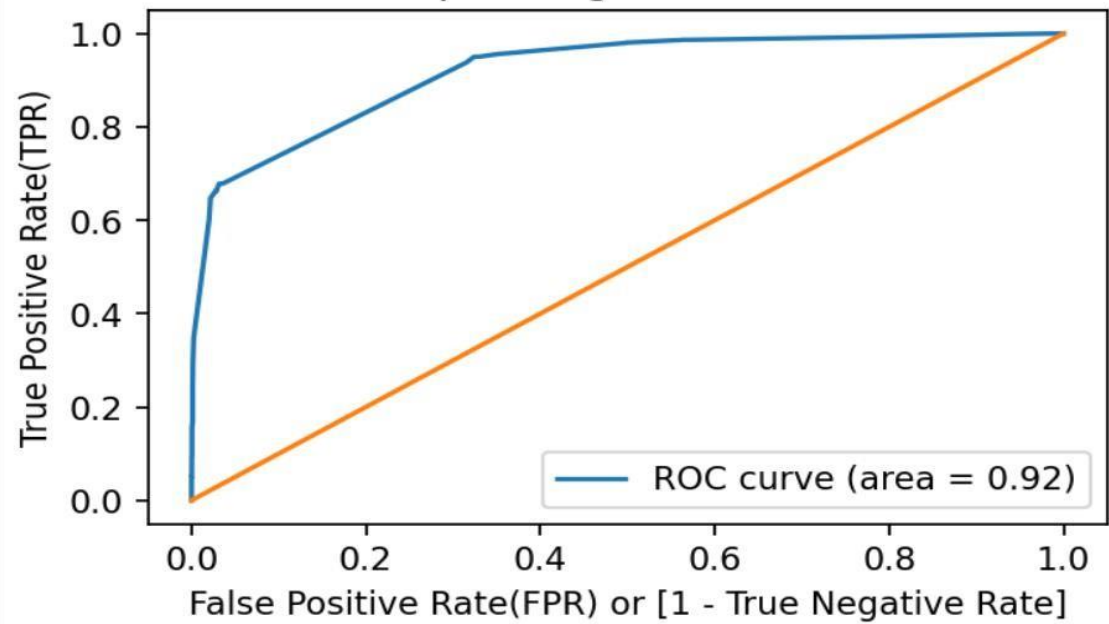
RFE is used to select 15 features

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



# 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 .