# LEAD SCORING CASE STUDY

*Done by*
*Kiruthika P*

# 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.
- 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.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor.
- The company now focusing on sales, so the lead conversion rate should be high that should be achieved through the identification of potential clients

# OBJECTIVE

- The company wants us as the data analyst to assist them to select the potential leads
- They need a model to identify more promising leads who can be converted to the customers by assigning lead score to each leads
- Lead score should be used to predict the chance of coversion rate of lead to customer.
- The target set by the CEO is 80%. Based on the data given we should calculate the lead score with some analysis

# PROBLEM APPROACH

- Data Setup(Import the required libraries and the data)
- Data Inspection
- Data Cleanup
- Dummy variable creation
- EDA and Data Preparation
- Scaling and Correlation
- Model building and evaluation
- Making the predictions

# DATA SET-UP

In [1]:
```
# Suppressing Warnings
import warnings
warnings.filterwarnings('ignore')
# Importing Pandas and NumPy
import pandas as pd, numpy as np

import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
# Importing lead dataset
lead_data = pd.read_csv("Leads.csv")
lead_data.head()
```

Out[2]:

| | Prospect ID | Lead Number | Lead Origin | Lead Source | Do Not Email | Do Not Call | Converted | TotalVisits | Total Time Spent on Website | Page Views Per Visit | ... | Get updates on DM Content | Lead Profile | City | Asymmetrique Activity Index | Asymmetriqu Profile Inde |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 7927b2df-8bba-4d29-b9a2-b6e0beafe620 | 660737 | API | Olark Chat | No | No | 0 | 0.0 | 0 | 0.0 | ... | No | Select | Select | 02.Medium | 02.Mediur |
| 1 | 2a272436-5132-4136-86fa-dcc88c88f482 | 660728 | API | Organic Search | No | No | 0 | 5.0 | 674 | 2.5 | ... | No | Select | Select | 02.Medium | 02.Mediur |
| 2 | 8cc8c611-a219-4f35-ad23-fdfd2656bd8a | 660727 | Landing Page Submission | Direct Traffic | No | No | 1 | 2.0 | 1532 | 2.0 | ... | No | Potential Lead | Mumbai | 02.Medium | 01.Hig |
| 3 | 0cc2df48-7cf4-4e39-9de9-19797f9b38cc | 660719 | Landing Page Submission | Direct Traffic | No | No | 0 | 1.0 | 305 | 1.0 | ... | No | Select | Mumbai | 02.Medium | 01.Hig |
| 4 | 3256f628-e534-4826-9d63-4a8b88782852 | 660681 | Landing Page Submission | Google | No | No | 1 | 2.0 | 1428 | 1.0 | ... | No | Select | Mumbai | 02.Medium | 01.Hig |

# DATA INSPECTION

*We have 9240 rows and 37 columns in our leads dataset*
*All the dataypes of the variables are in correct format.*
*Missing values found given below*

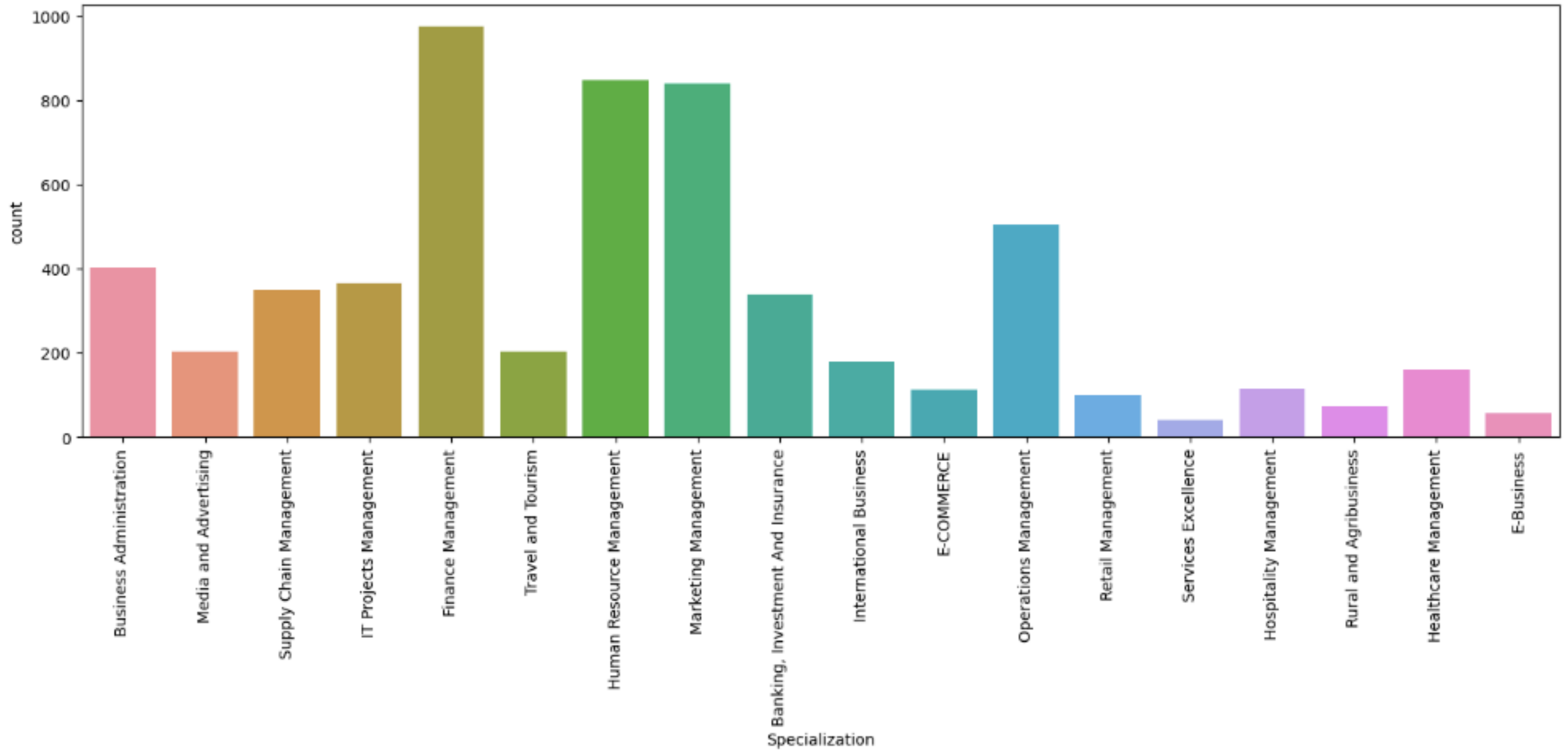
In [5]:
```
lead_data.describe()
```

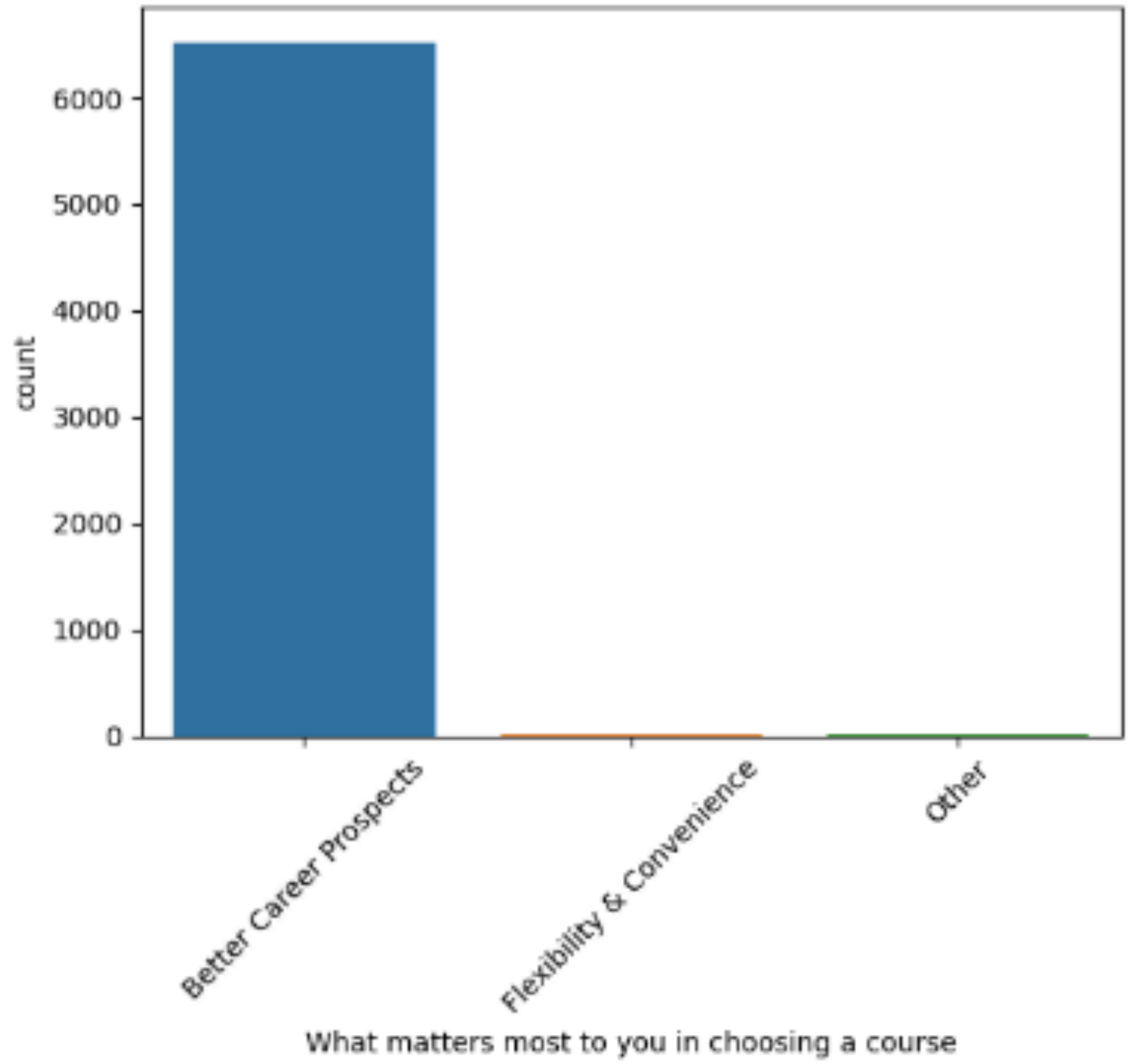
Out[5]:

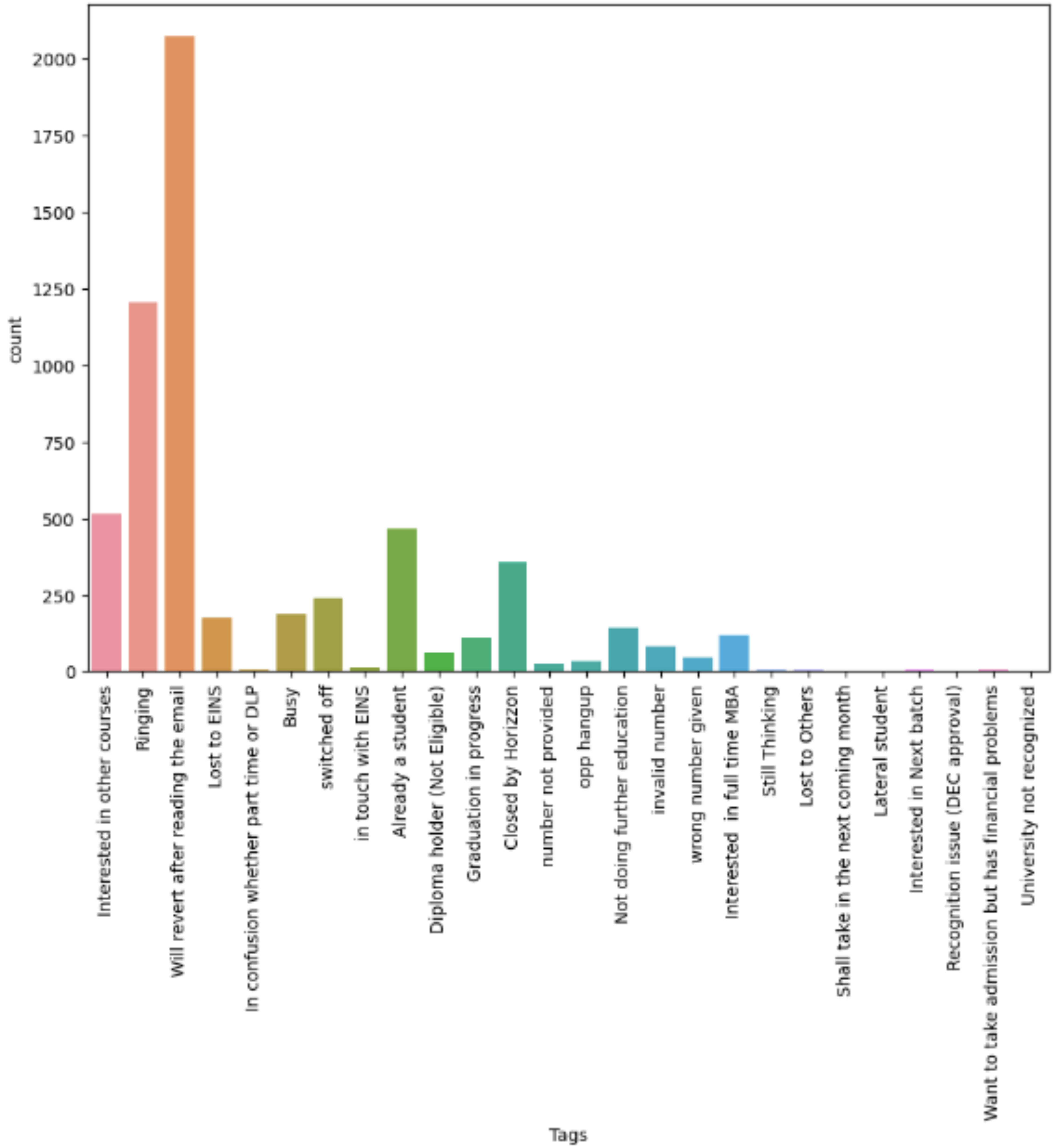
| | Lead Number | Converted | TotalVisits | Total Time Spent on Website | Page Views Per Visit | Asymmetrique Activity Score | Asymmetrique Profile Score |
|---|---|---|---|---|---|---|---|
| **count** | 9240.000000 | 9240.000000 | 9103.000000 | 9240.000000 | 9103.000000 | 5022.000000 | 5022.000000 |
| **mean** | 617188.435606 | 0.385390 | 3.445238 | 487.698268 | 2.362820 | 14.306252 | 16.344883 |
| **std** | 23405.995698 | 0.486714 | 4.854853 | 548.021466 | 2.161418 | 1.386694 | 1.811395 |
| **min** | 579533.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 7.000000 | 11.000000 |
| **25%** | 596484.500000 | 0.000000 | 1.000000 | 12.000000 | 1.000000 | 14.000000 | 15.000000 |
| **50%** | 615479.000000 | 0.000000 | 3.000000 | 248.000000 | 2.000000 | 14.000000 | 16.000000 |
| **75%** | 637387.250000 | 1.000000 | 5.000000 | 936.000000 | 3.000000 | 15.000000 | 18.000000 |
| **max** | 660737.000000 | 1.000000 | 251.000000 | 2272.000000 | 55.000000 | 18.000000 | 20.000000 |

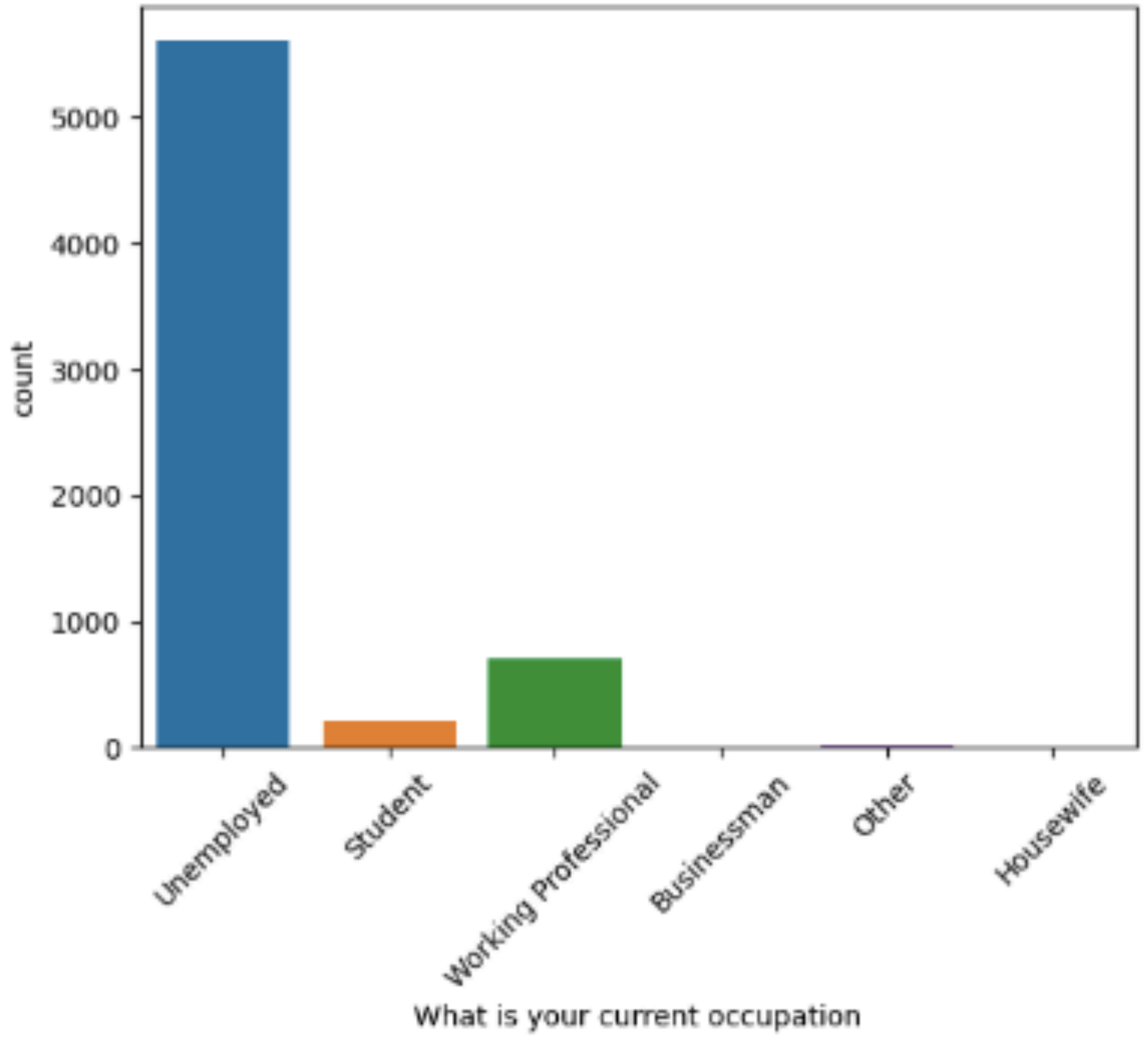
# DATA CLEANUP

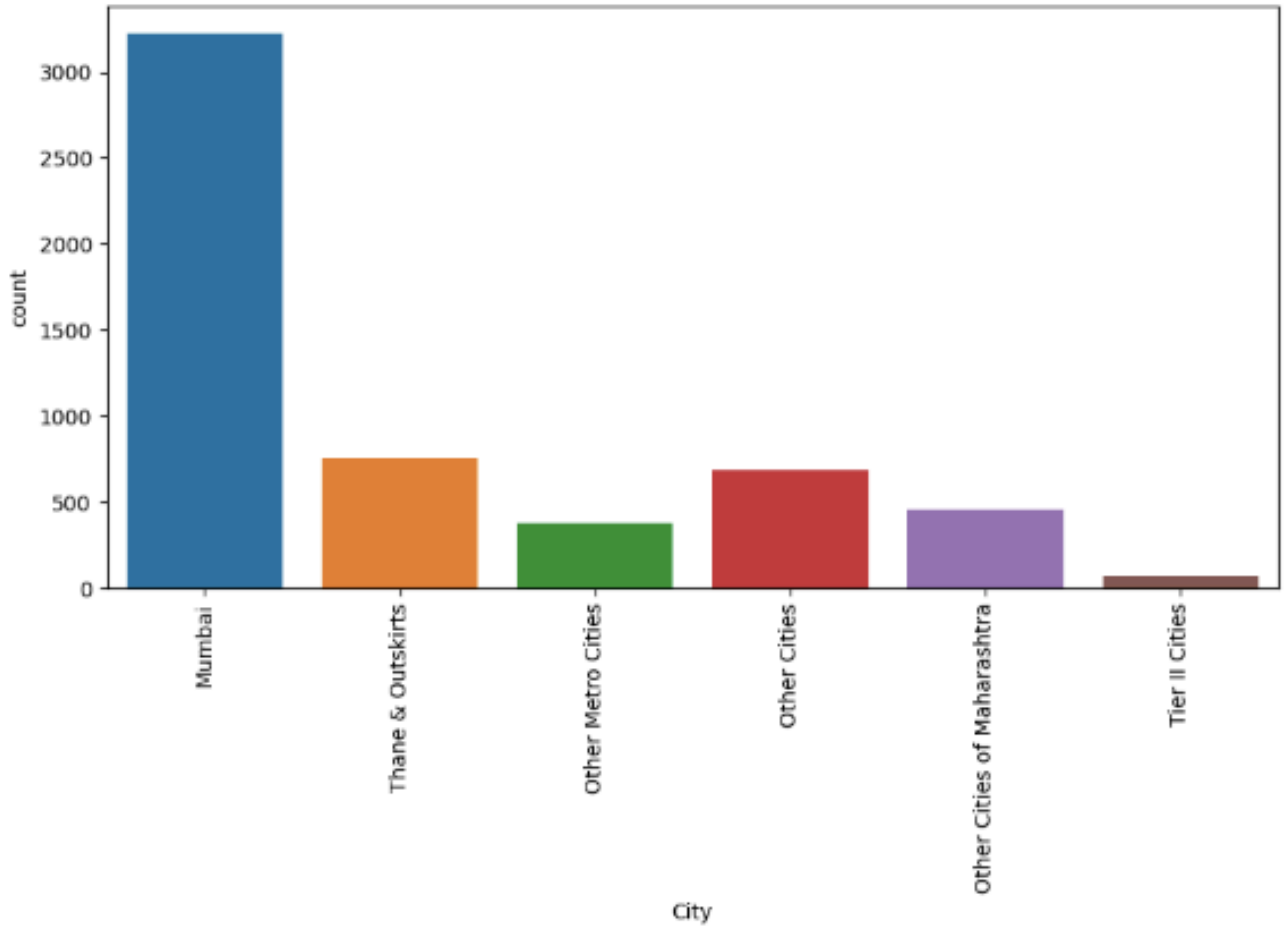
- We observe that there are 'Select' values in many columns'Select' values are as good as NULL. So we can convert these values to null values.
- We see that for some columns we have high percentage of missing values. We can drop the columns with missing values greater than 40%.
- Some of the columns and the % of missing values
    - Specialisation - 37%
    - Tags - 36%
    - What matters most to you in choosing a course - 29%
    - What is your current occupation - 29%
    - Country - 37%
    - City - 40%
- We have retained 98% of the rows after cleanup.

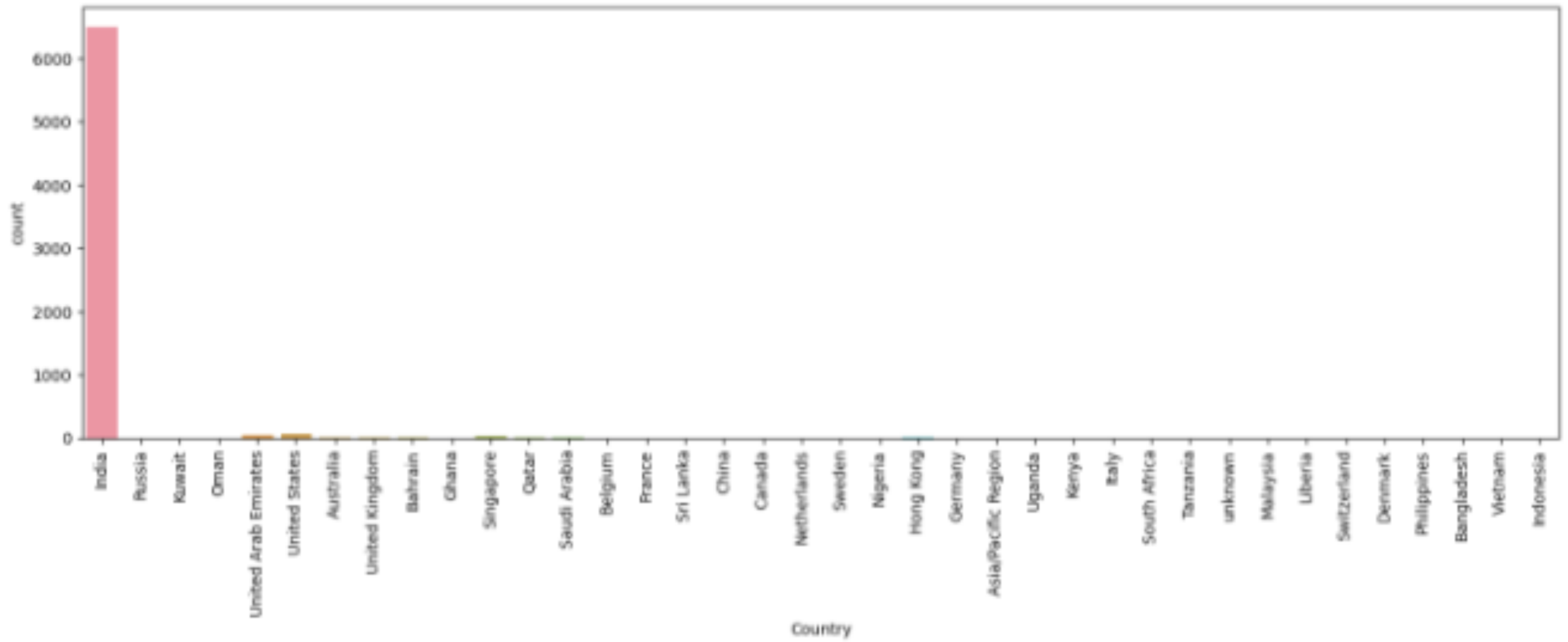

count
0
200
400
600
800
1000
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization


count
0
1000
2000
3000
4000
5000
6000
Better Career Prospects
Flexibility & Convenience
Other
What matters most to you in choosing a course


count
0
250
500
750
1000
1250
1500
1750
2000
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
Closed by Horizzon
number not provided
opp hangup
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
Tags



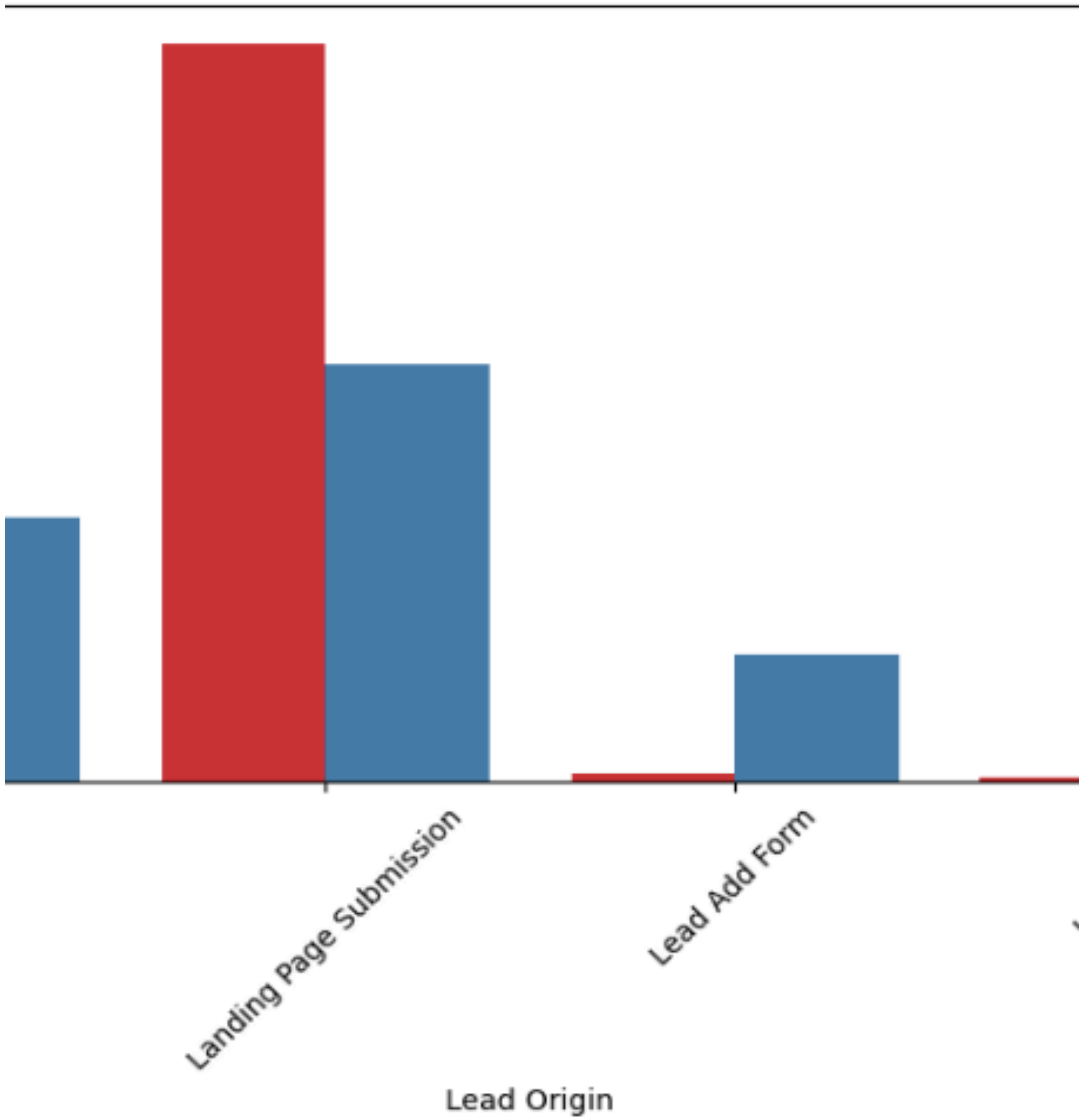




```
Last Notable Activity                          0.0
dtype: float64
```

In [29]:
```
# Percentage of rows retained
(len(lead_data.index)/9240)*100
```

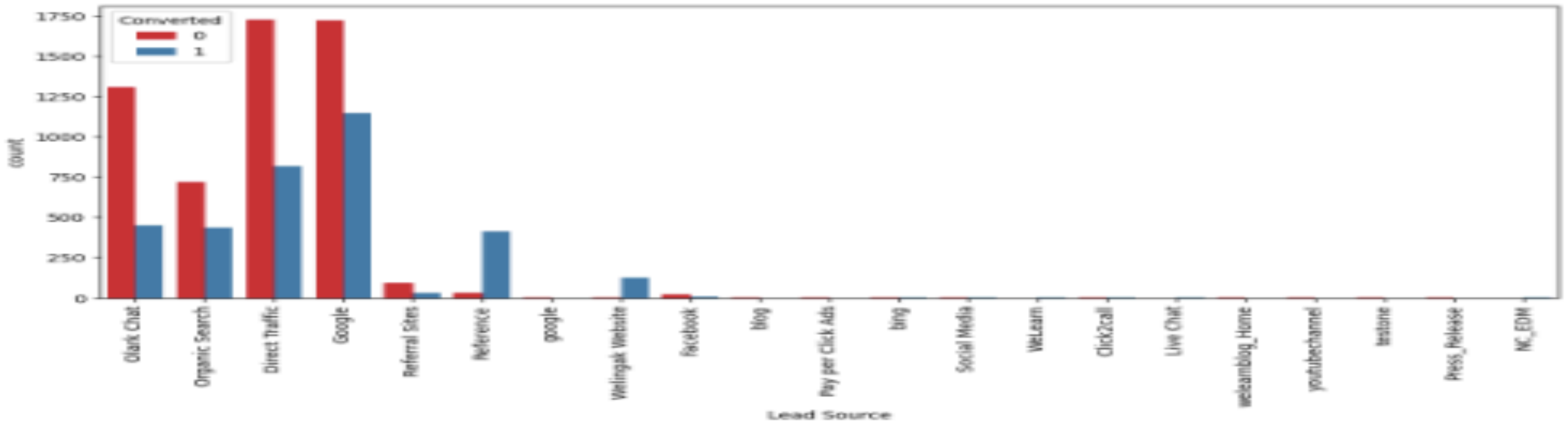
Out[29]: `98.2034632034632`



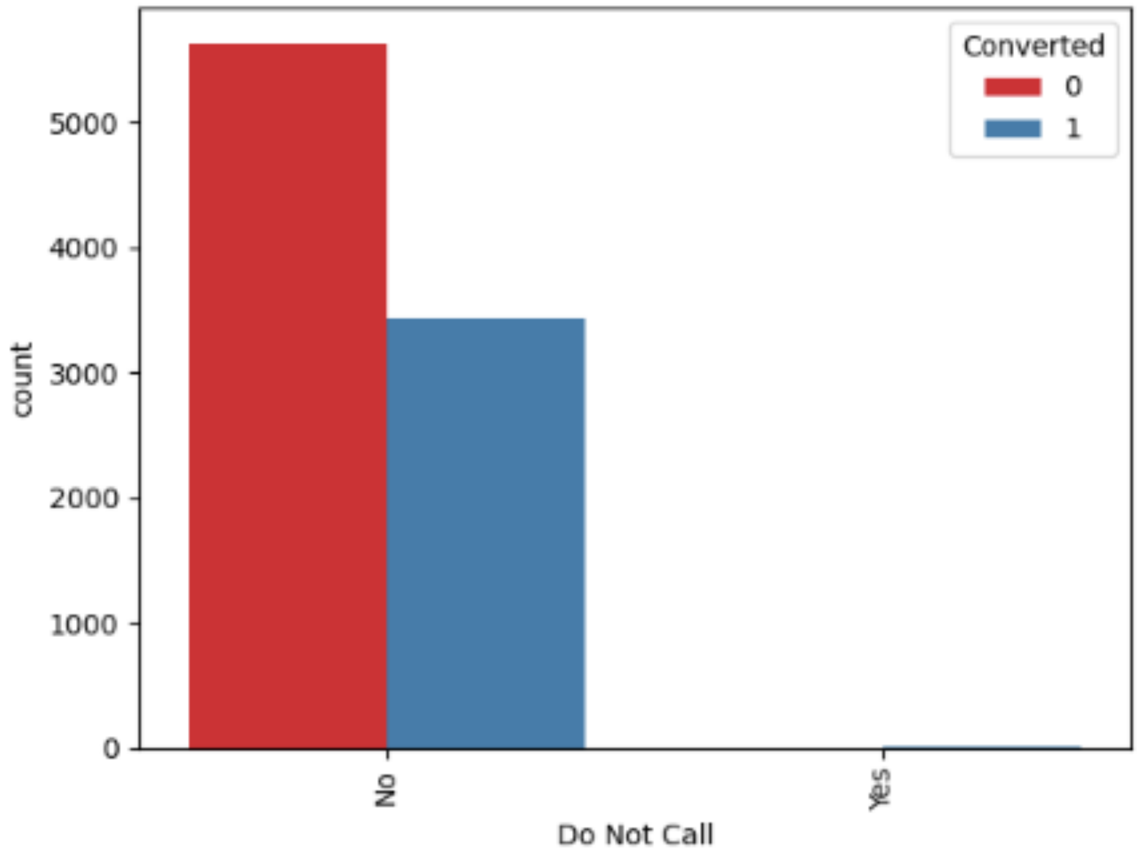

# EDA

- Converted is the target variable, Indicates whether a lead has been successfully converted (1) or not (0). The lead conversion rate is 38%
- Lead Origin
  - API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
  - Add Form has more than 90% conversion rate and Lead import are very less in count

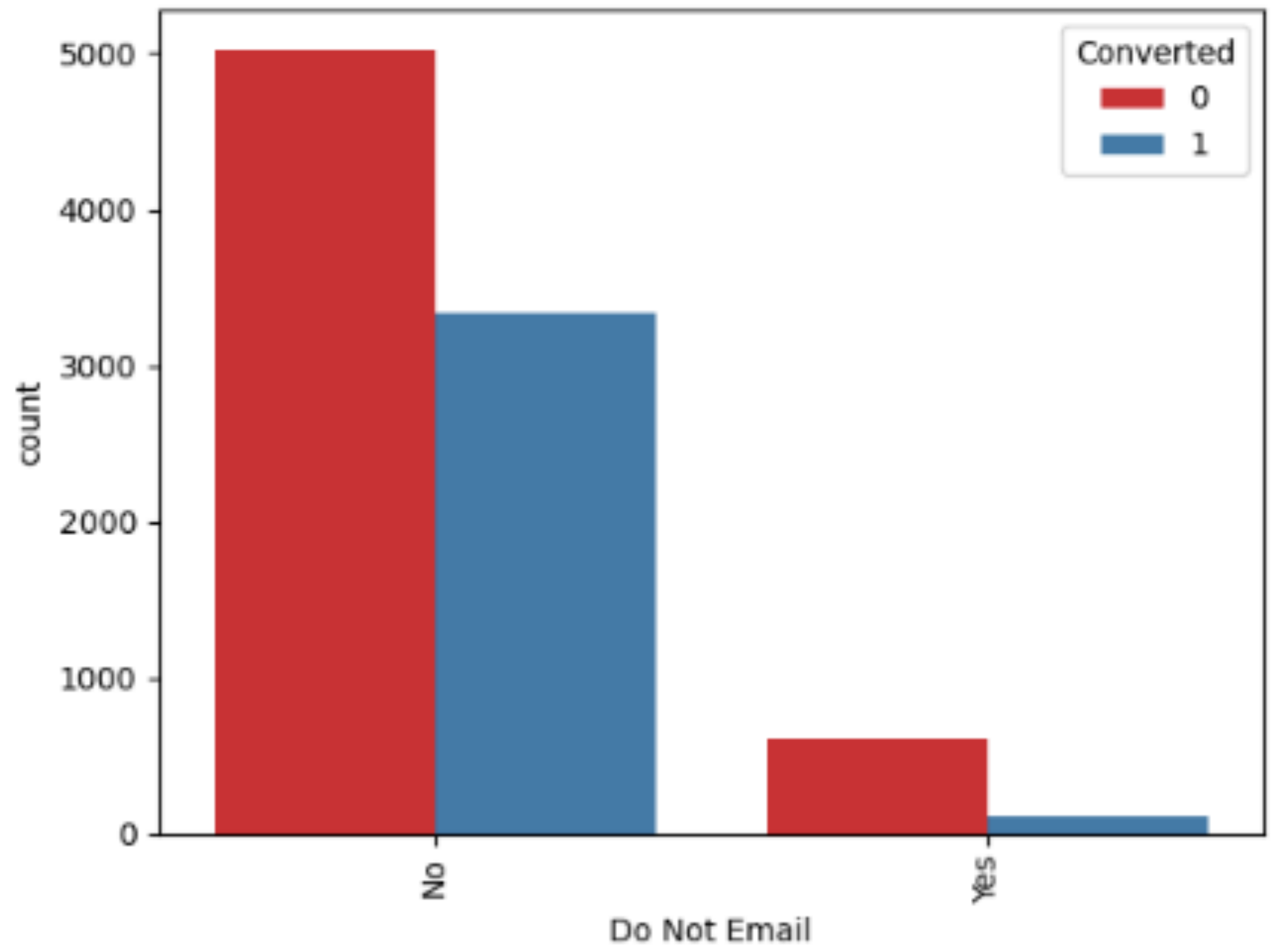
```
In [35]: plt.figure(figsize=(13,5))
         sns.countplot(x = "Lead Source", hue = "Converted", data = lead_data, palette='Set1')
         plt.xticks(rotation = 90)
Out[35]: (array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
                 17, 18, 19, 20]),
          [Text(0, 0, 'Olark Chat'),
           Text(1, 0, 'Organic Search'),
           Text(2, 0, 'Direct Traffic'),
           Text(3, 0, 'Google'),
           Text(4, 0, 'Referral Sites'),
           Text(5, 0, 'Reference'),
           Text(6, 0, 'google'),
           Text(7, 0, 'Welingak Website'),
           Text(8, 0, 'Facebook'),
           Text(9, 0, 'blog'),
           Text(10, 0, 'Pay per Click Ads'),
           Text(11, 0, 'bing'),
           Text(12, 0, 'Social Media'),
           Text(13, 0, 'WeLearn'),
           Text(14, 0, 'Click2call'),
           Text(15, 0, 'Live Chat'),
           Text(16, 0, 'welearnblog_Home'),
           Text(17, 0, 'youtubechannel'),
           Text(18, 0, 'testone'),
           Text(19, 0, 'Press_Release'),
           Text(20, 0, 'NC_EDM')])
```

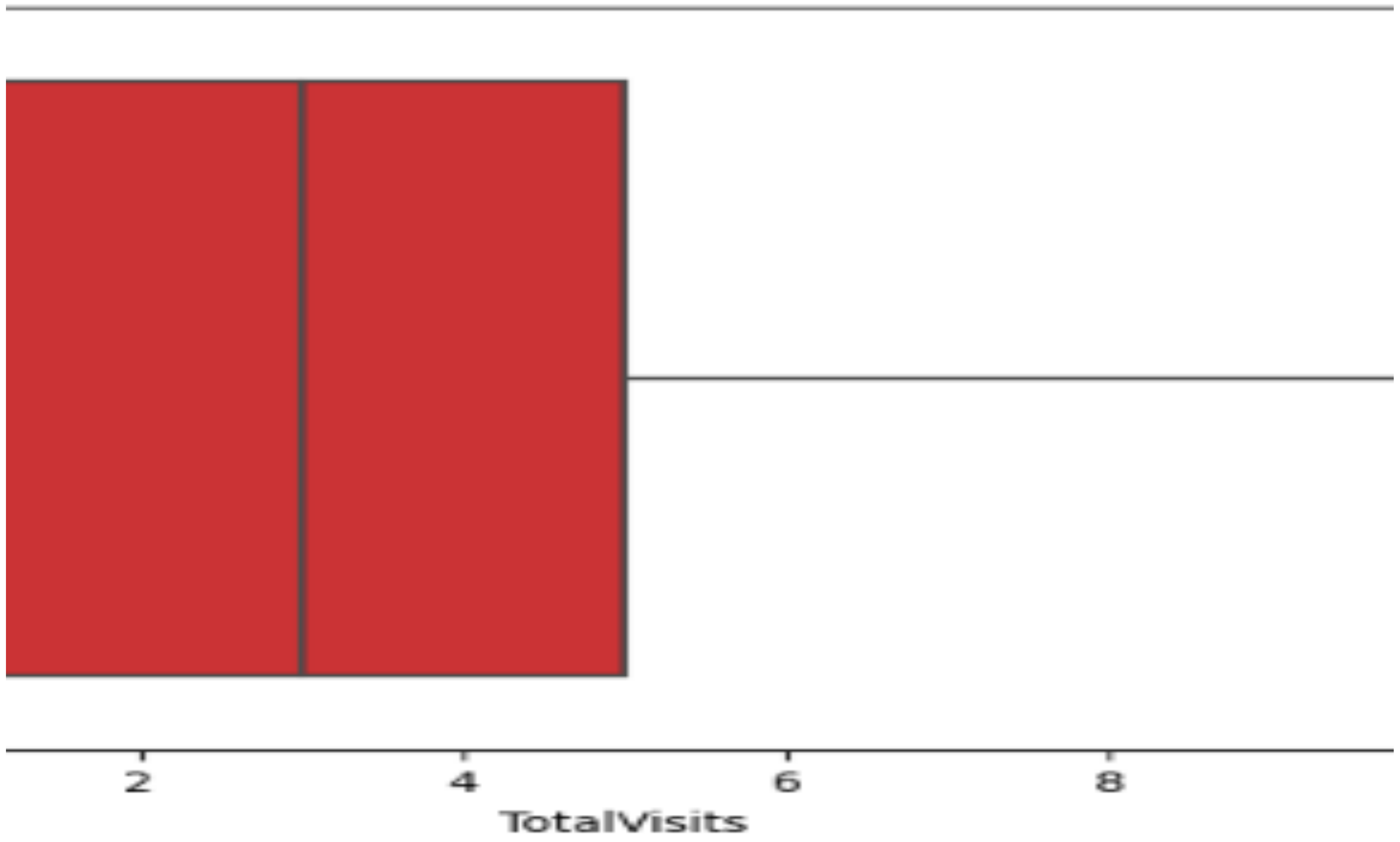



- ➤ Lead Source
  - ➤ Google and Direct traffic generates maximum number of leads.
  - ➤ Conversion Rate of reference leads and leads through welingak website is high
- ➤ Do not Email & Do not Email
  - ➤ Most entries are 'No'. No Inference can be drawn with this parameter.

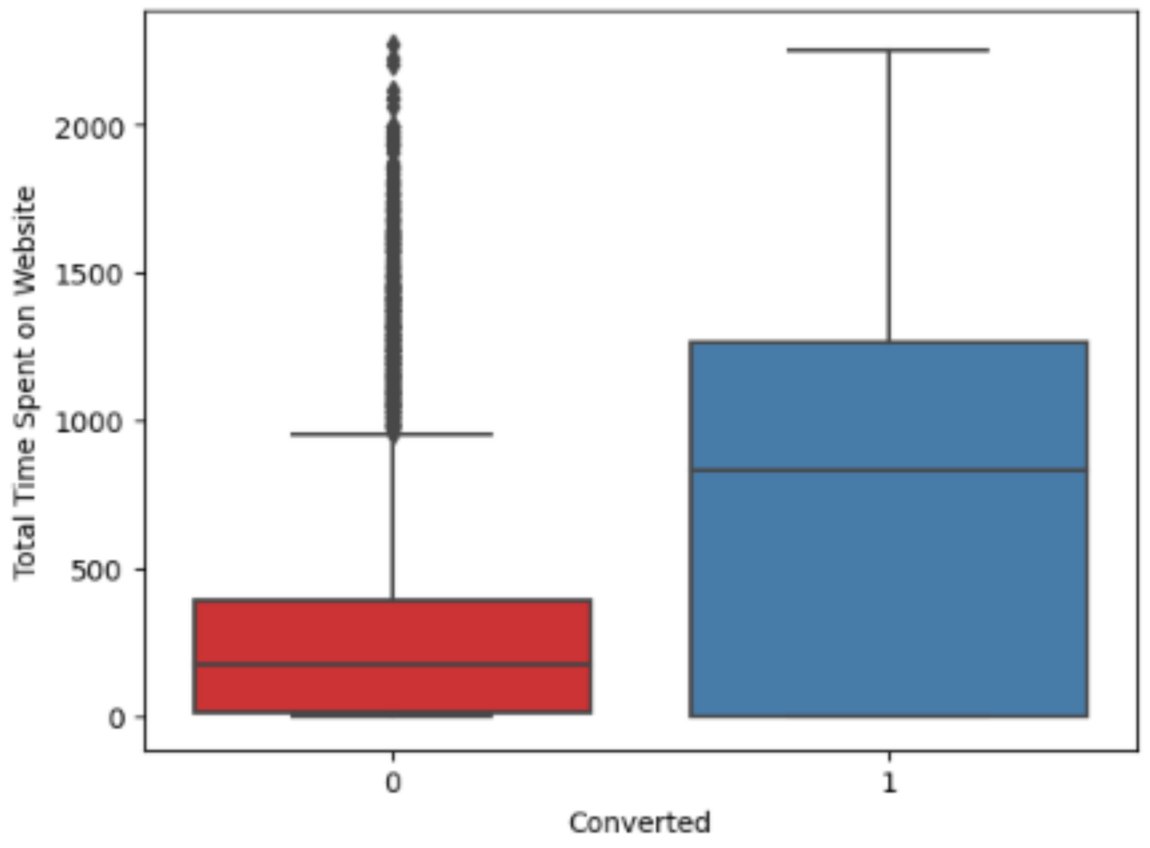


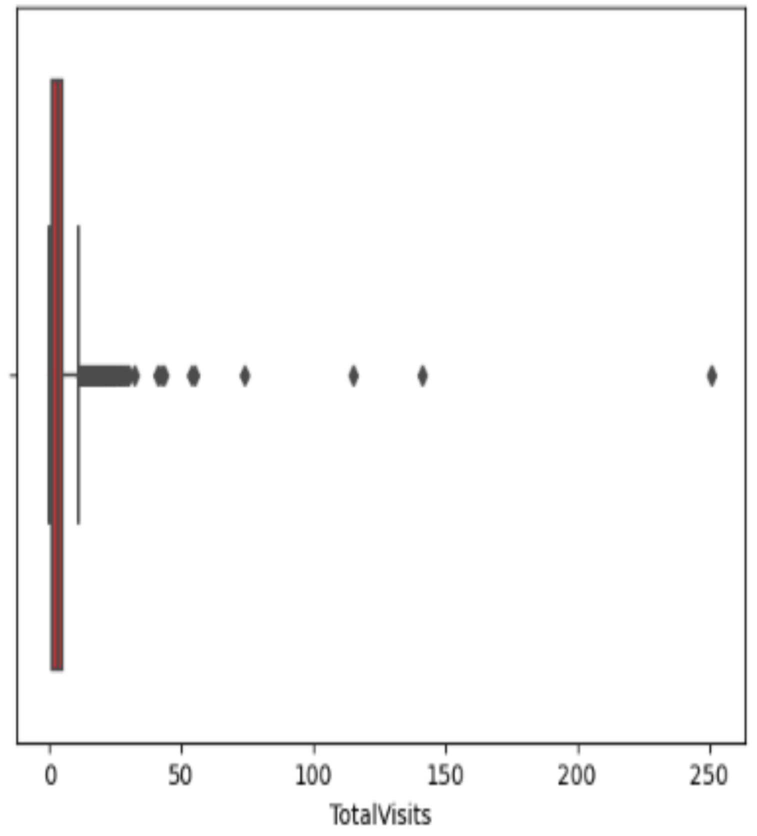

```
In [39]: sns.countplot(x = "Do Not Email", hue = "Converted", data = lead_data,palette='Set1')
         plt.xticks(rotation = 90)
Out[39]: (array([0, 1]), [Text(0, 0, 'No'), Text(1, 0, 'Yes')])
```

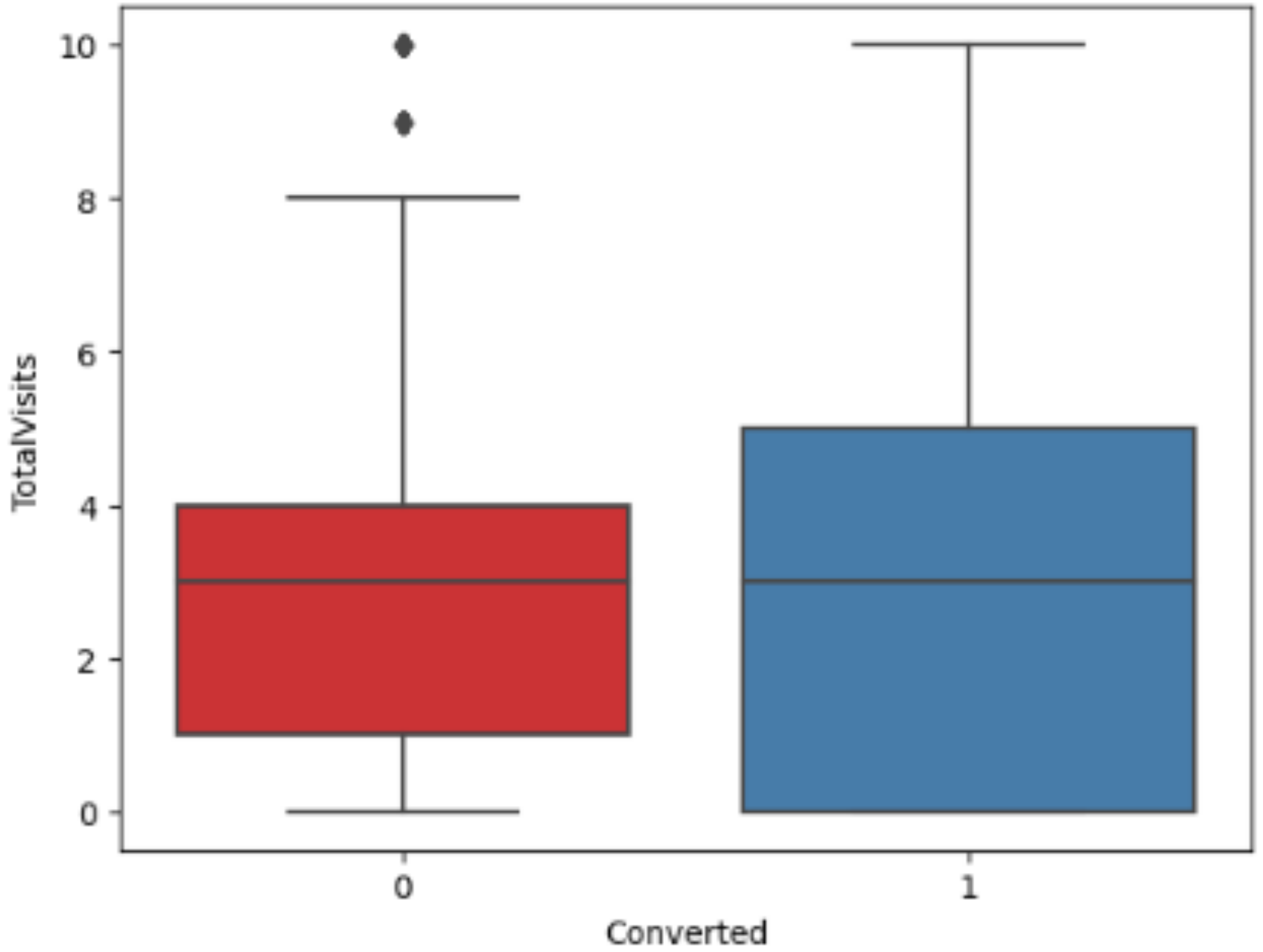





- Total Visits
  - As we can see there are a number of outliers in the data. We will cap the outliers to 95% value for analysis.
  - Median for converted and not converted leads are the same.
- Total time spent on website
  - Leads spending more time on the weblise are more likely to be converted.



# DATA PREPARATION

- Converting some binary variables to 1 or 0
- Creating dummy variables for categorical features like Lead origin, lead source, last activity, specialization and etc.,
- Drop the columns for which dummies are created
- Splitting the data into train and test set
- Scale the features
- Feature selection using RFE

# MODEL BUILDING

- Created different models and dropped the models which has the high values
- Since the P values of all variables is 0 and VIF values are low for all the variables, model-7 is our final model. We have 12 variables in our final model.

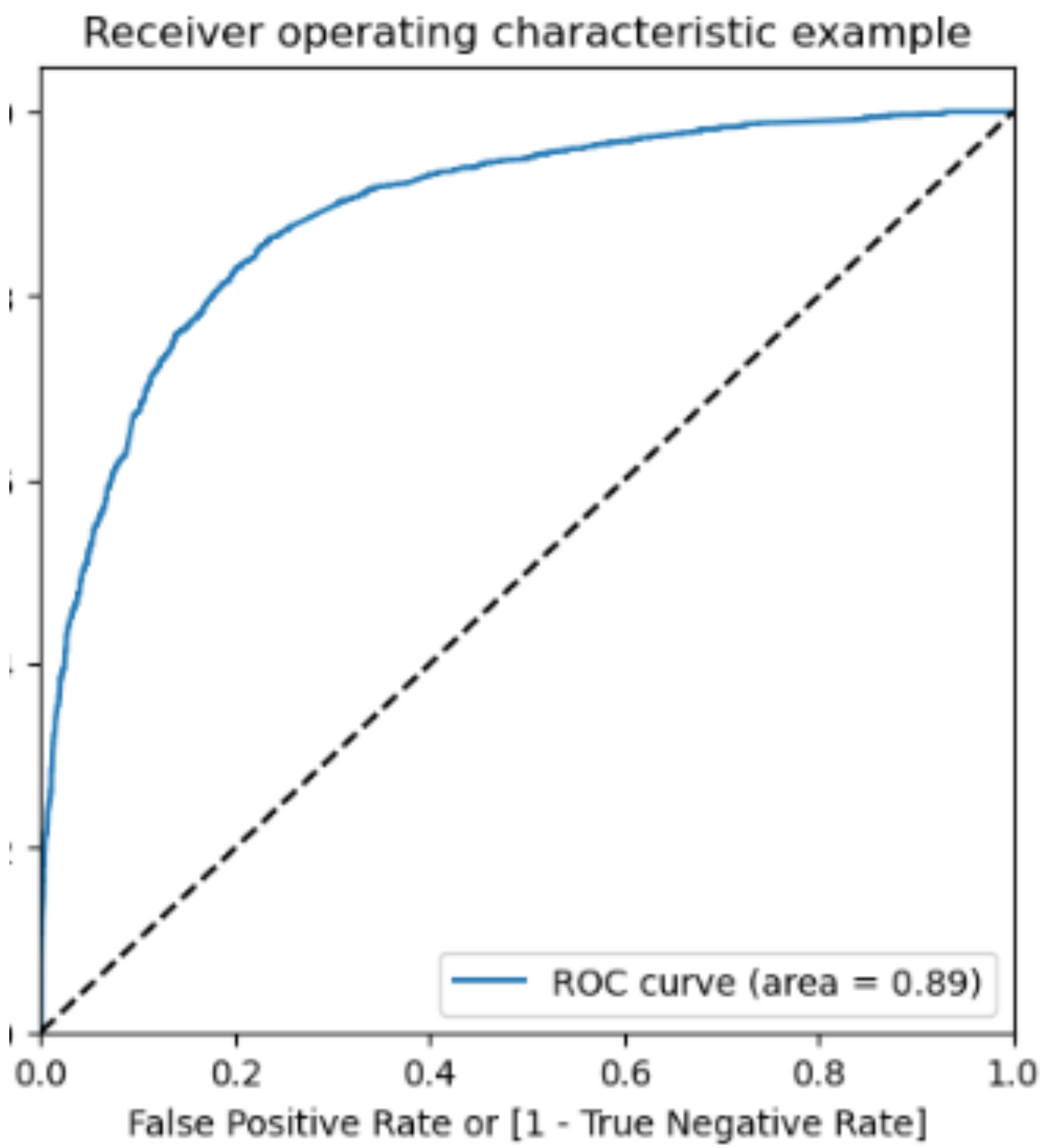
Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Converted | **No. Observations:** | 6351 |
| **Model:** | GLM | **Df Residuals:** | 6320 |
| **Model Family:** | Binomial | **Df Model:** | 30 |
| **Link Function:** | Logit | **Scale:** | 1.0000 |
| **Method:** | IRLS | **Log-Likelihood:** | -2577.3 |
| **Date:** | Wed, 16 Aug 2023 | **Deviance:** | 5154.5 |
| **Time:** | 00:14:51 | **Pearson chi2:** | 6.45e+03 |
| **No. Iterations:** | 20 | **Pseudo R-squ. (CS):** | 0.4063 |
| **Covariance Type:** | nonrobust | | |

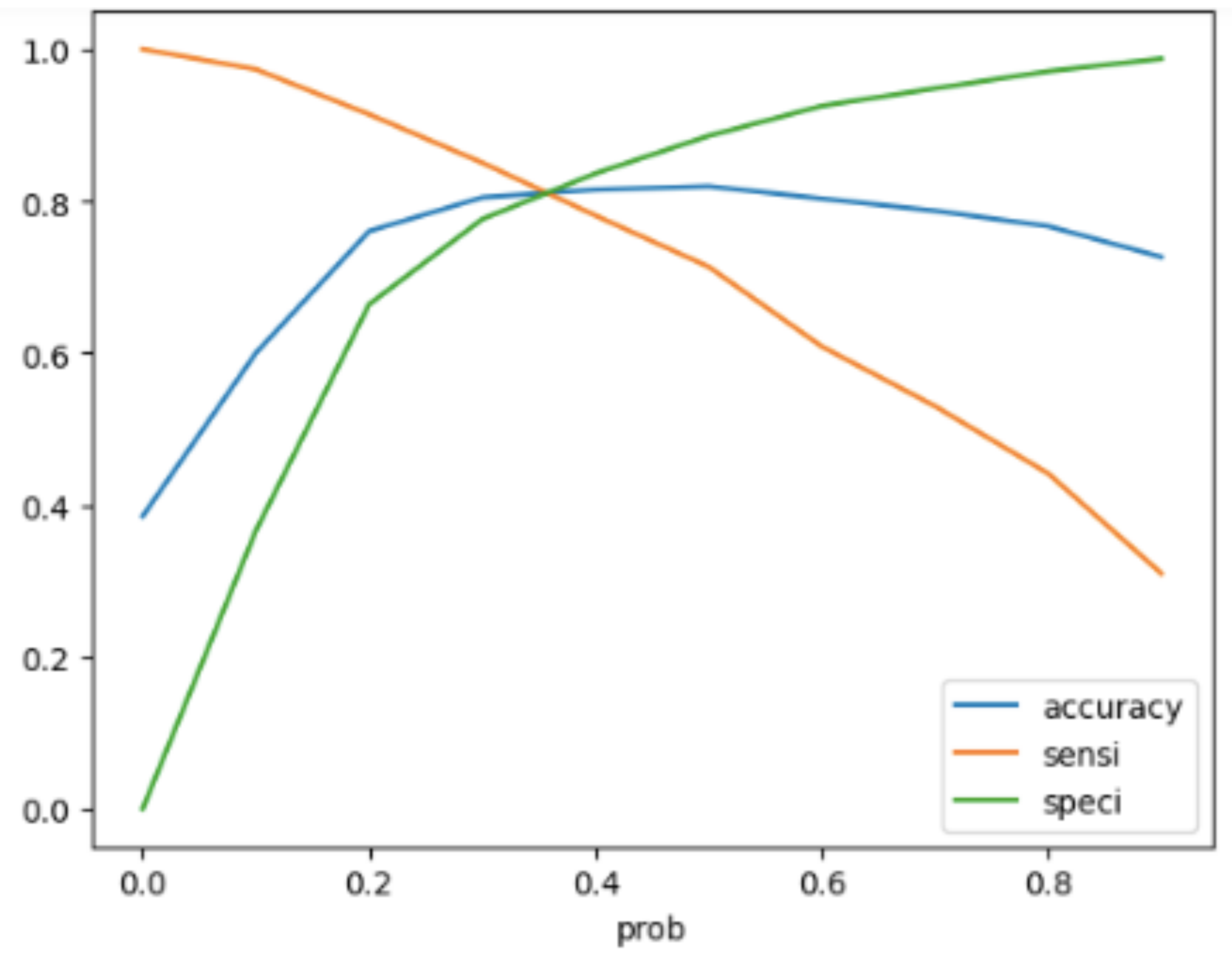
# MODEL PREDICTIONS

- Created a new column 'predicted' with 1 if Converted_Prob > 0.5 else 0
- Confusion matrix
- Calculated positive and negative Predictive values
- We found out that our specificity was good 87% but our sensitivity was only 70%. Hence, this needed to be taken care of.
- We have got sensitivity of 70% and this was mainly because of the cut-off point of 0.5 that we had arbitrarily chosen. Now, this cut off point had to be optimised in order to get a decent value of sensitivity and for this we will use the ROC curve.

```
In [91]: from sklearn import metrics

         # Confusion matrix
         confusion = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.predicted )
         print(confusion)

[[3459  446]
 [ 701 1745]]

In [92]: # Let's check the overall accuracy.
         print('Accuracy :',metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.predicted))

Accuracy : 0.8193985199181232

In [93]: TP = confusion[1,1] # true positive
         TN = confusion[0,0] # true negatives
         FP = confusion[0,1] # false positives
         FN = confusion[1,0] # false negatives

In [94]: # Sensitivity of our logistic regression model
         print("Sensitivity : ",TP / float(TP+FN))

Sensitivity :  0.7134096484055601

In [95]: # Let us calculate specificity
         print("Specificity : ",TN / float(TN+FP))

Specificity :  0.885787451984635

In [96]: # Calculate false postive rate - predicting converted lead when the lead actually was not converted
         print("False Positive Rate :",FP/ float(TN+FP))

False Positive Rate : 0.11421254801536491

In [97]: # positive predictive value
         print("Positive Predictive Value :",TP / float(TP+FP))

Positive Predictive Value : 0.7964399817434962

In [98]: # Negative predictive value
         print ("Negative predictive value :",TN / float(TN+ FN))

Negative predictive value : 0.8314903846153846
```




- ➤ Since we have higher area under the ROC curve , therefore our model is a good one.
- ➤ From the curve in the below diagram 0.34 is the optimum point to take it as a cutoff probability



- Observations:
  - Test Data
    - Accuracy : 80.4%
    - Sensitivity : 80.4%
    - Specificity : 80.5%
- Results
- Train Data:
  - Accuracy : 81.0 %
  - Sensitivity : 81.7 %
  - Specificity : 80.6 %
- Test Data:
  - Accuracy : 80.4%
  - Sensitivity : 80.4%
  - Specificity : 80.5%

- Thus we have achieved our goal of getting a ballpark of the target lead conversion rate to be around 80% . The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate of 80%.

# CONCLUSION

- The company can communicate to the leads that gets from the following fields. Those are lead sources, reference and websites which has the high chance of conversion to customers.
- Concentrate on Working professionals and the persons who spent more time on the websites. Last activity was sms sent are more likely to be converted.
- Specialization was others and who asked not to email those are likely not interested. Lead origin is Landing page submission are also not likely to be converted.