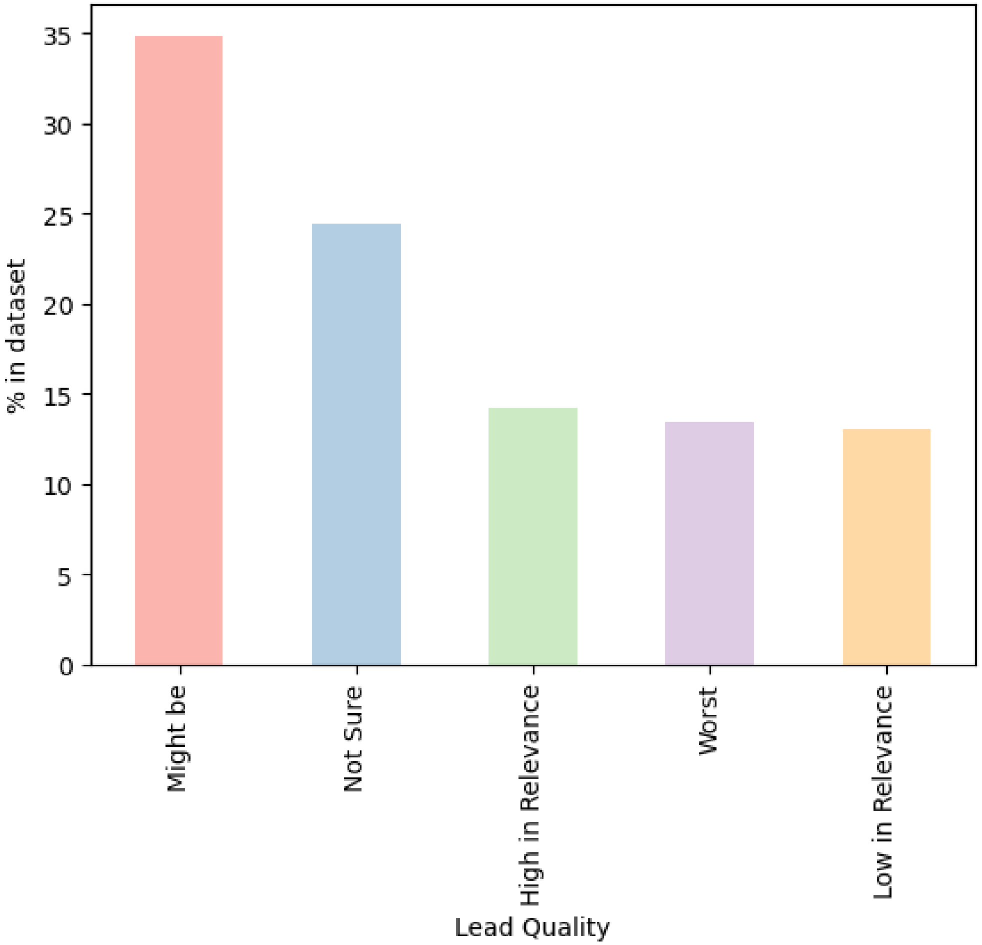
LEAD SCORING CASE STUDY

METHODOLGY: -

- DATA CLEANING
- DATA PREPARATION
- EDA
- MODEL BUILDING
- MODEL EVALUATION

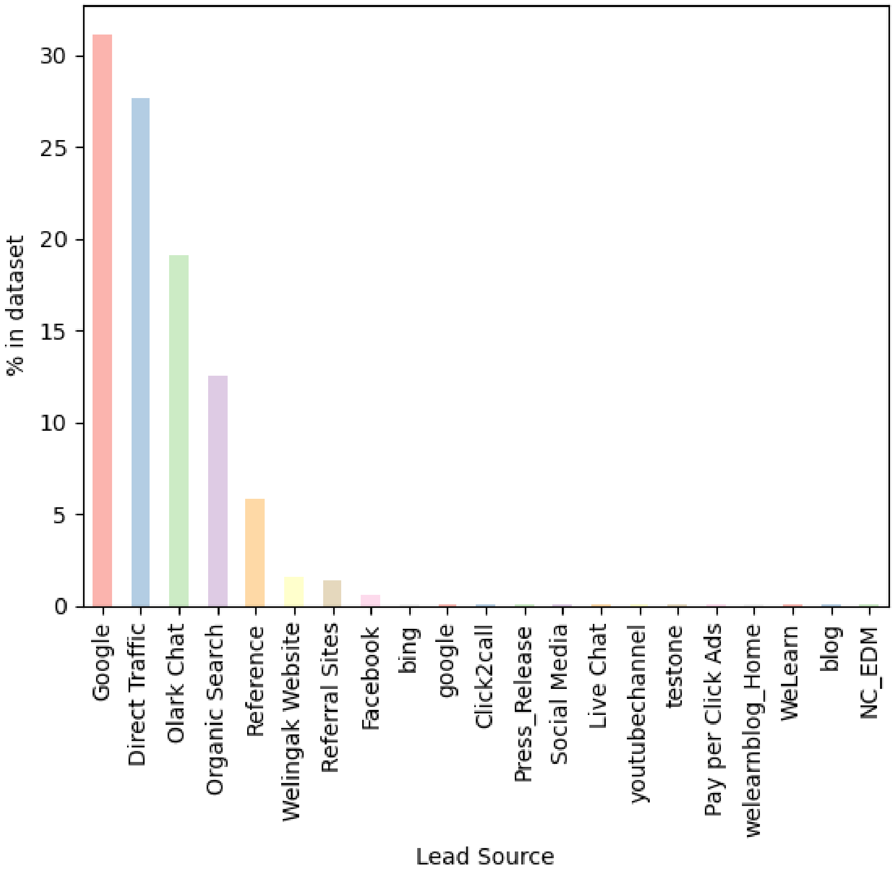
EXPLORATORY DATA ANALYSIS

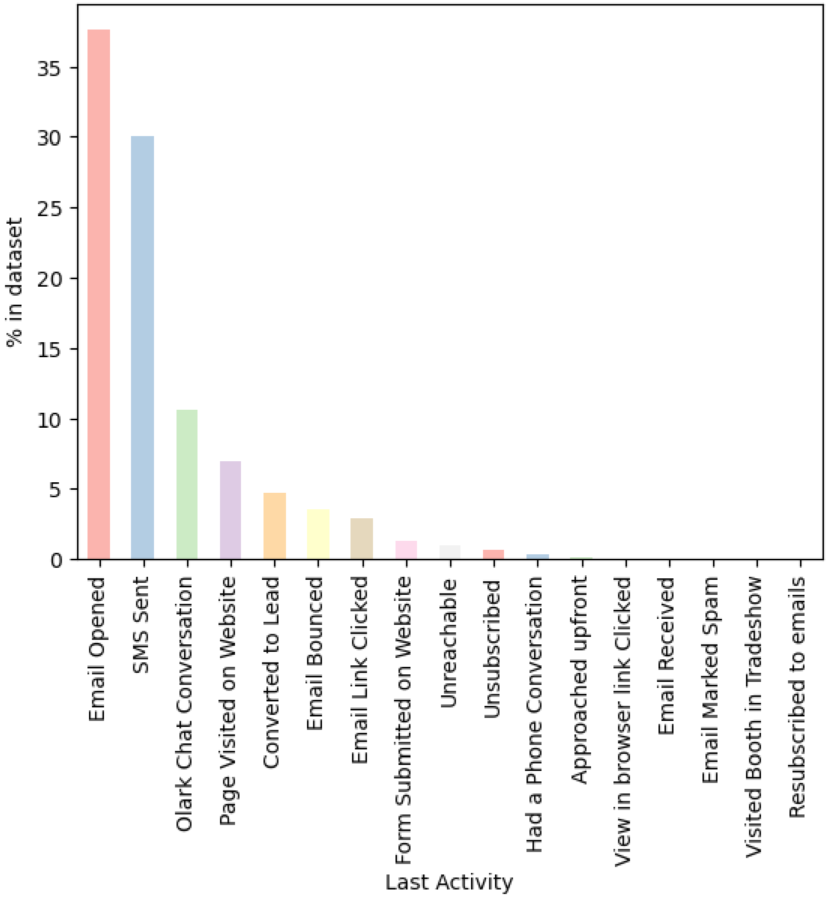
CHECKING FOR LEAD QUALITY:-



Null values in 'Lead Quality' column can be imputed with the value 'Not Sure' as we can assume that not filling in a column means the employee does not know or is not sure about the option.

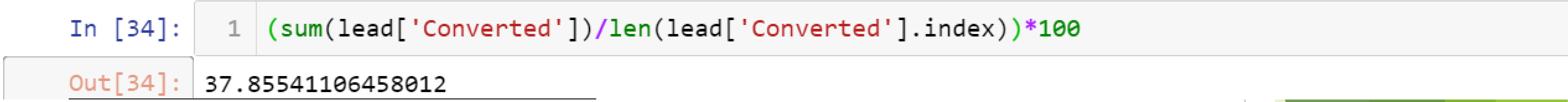
INSIGHTS FOR LEAD SOURCE LAST ACTIVITY:-



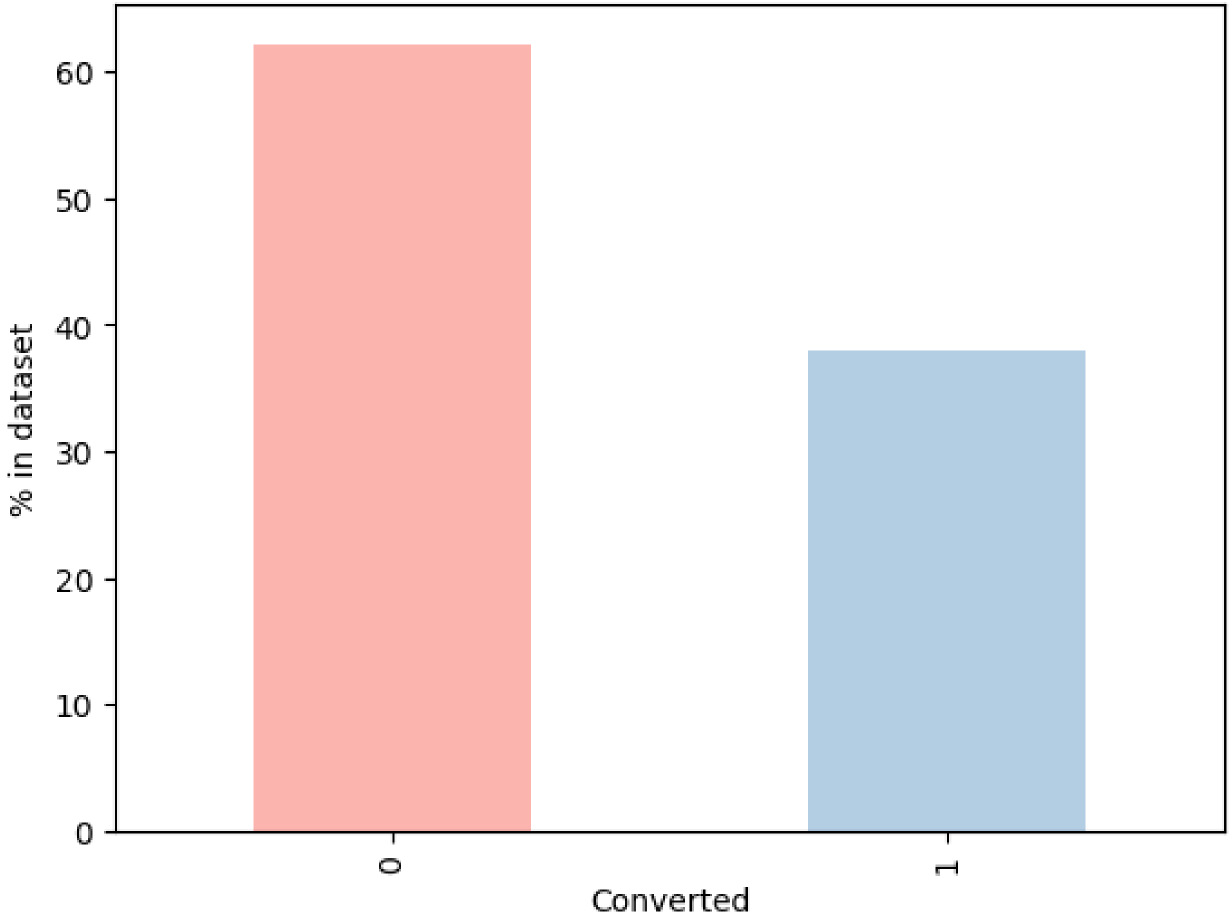


In these categorical variables, imputing with the most frequent value is not accurate as the next most frequent value has similar frequency. Also, as these variables have very little missing values, it is better to drop the rows containing these missing values.

INSIGHTS FOR COLUMN CONVERTED:-



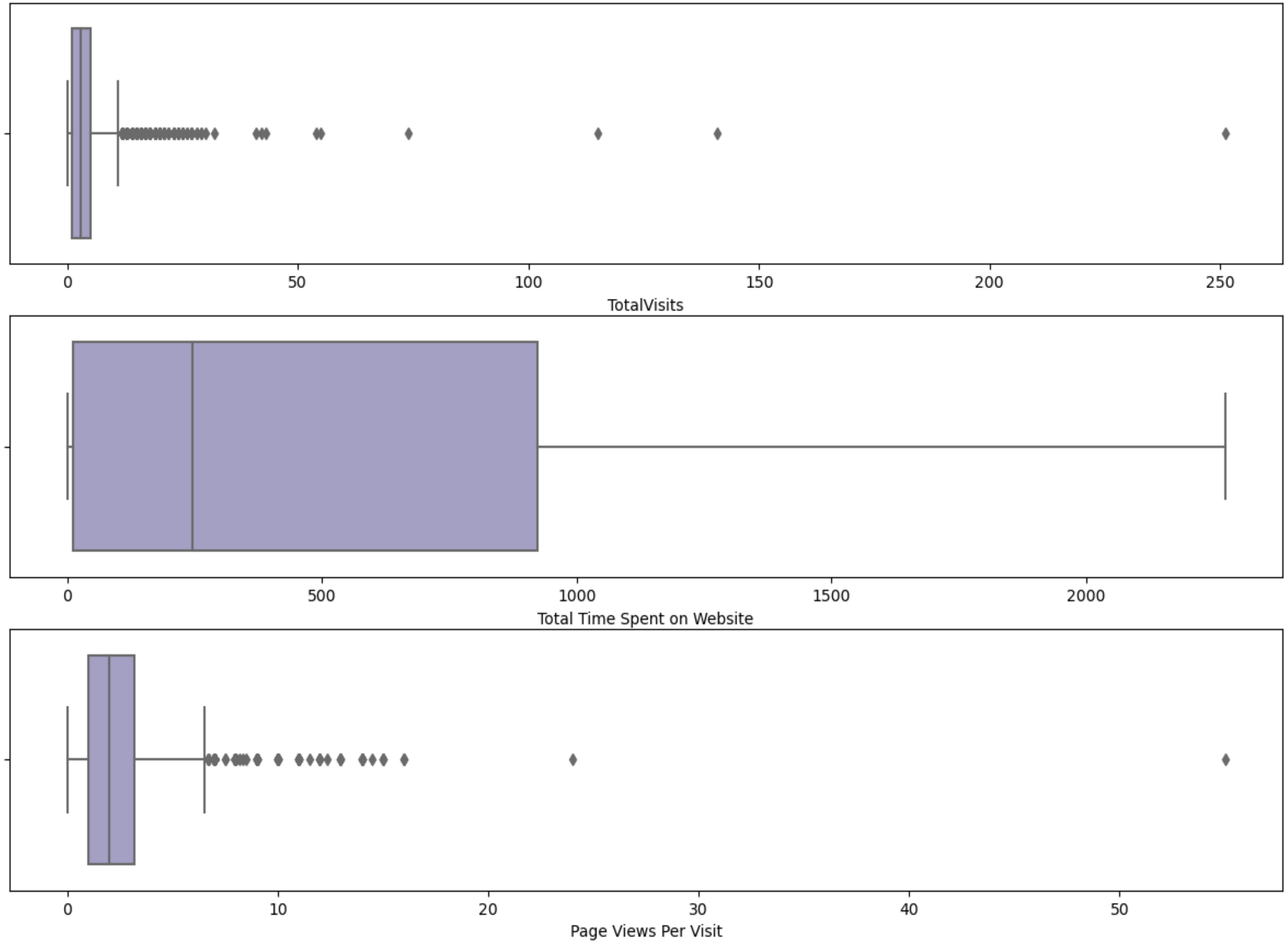
```
In [34]: (sum(lead['Converted'])/len(lead['Converted'].index))*100
Out[34]: 37.85541106458012
```



37.8% of the 'Converted' data is 1 i.e. 37.8% of the leads are converted.

This means we have enough data of converted leads for modelling.

EDA ON NUMERICAL VARIABLES:-

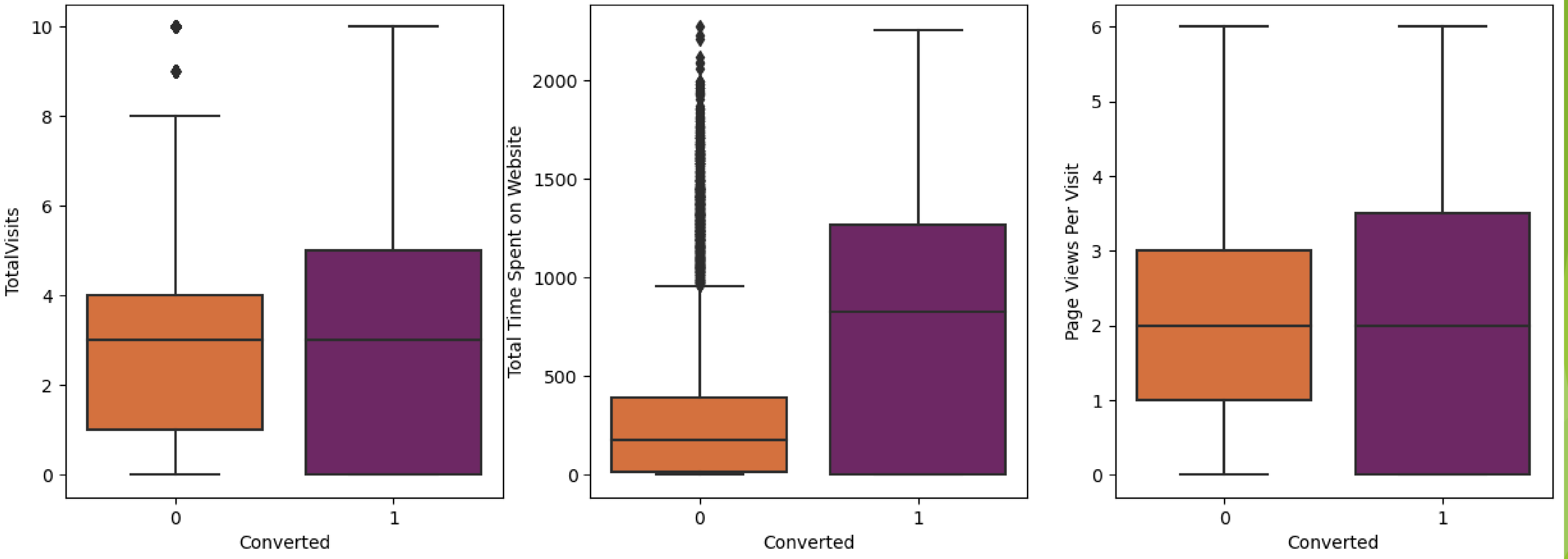


- From the boxplots, we can see that there are outliers present in the variables.

- For 'TotalVisits', the 95% quantile is 10 whereas the maximum value is 251. Hence, we can cap these outliers at 95% value.

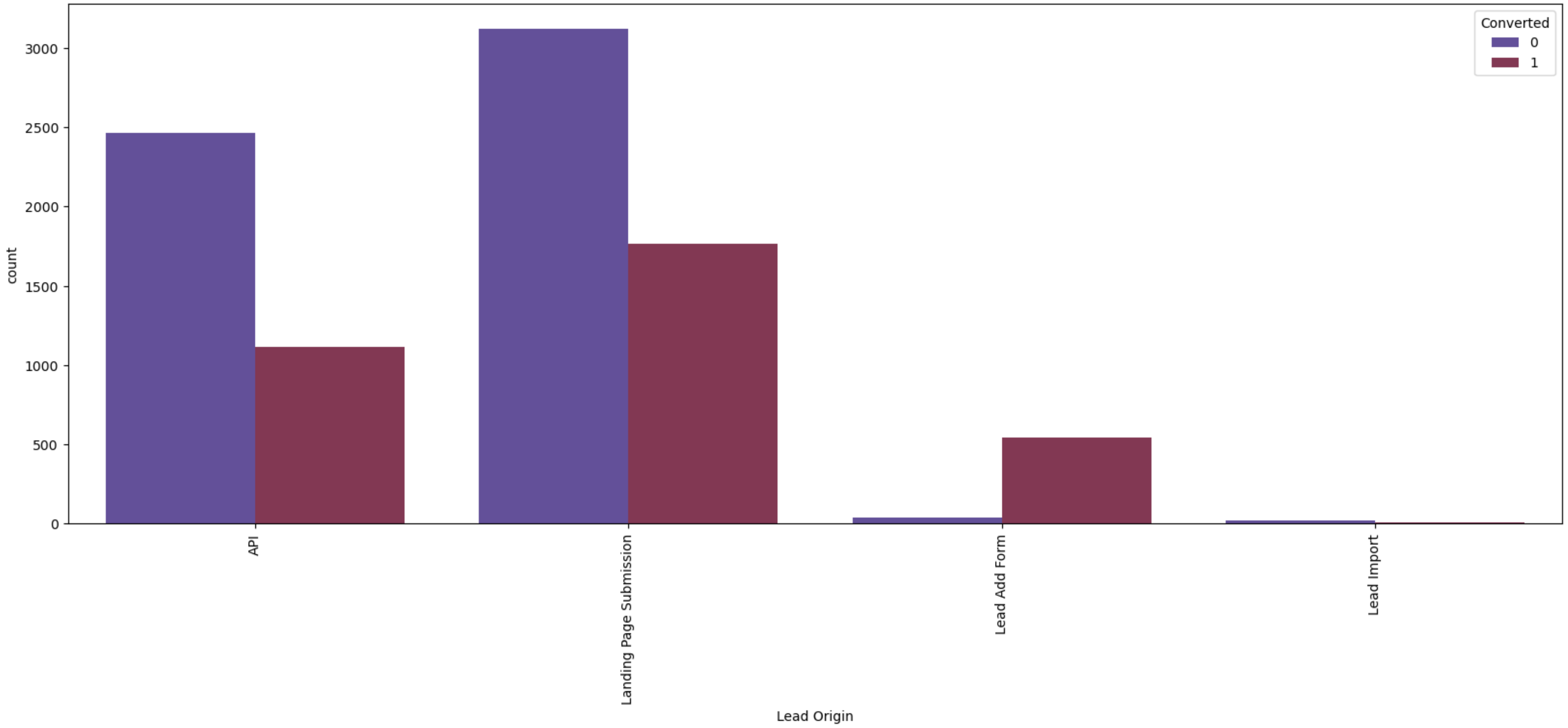
- There are no significant outliers in 'Total Time Spent on Website' for 'Page Views Per Visit', similar to 'TotalVisits', we should cap outliers at 95% value.

PLOTTED NUMERICAL VARIABLES AGAINST TARGET VARIABLE:-



- 1) 'TotalVisits' has same median values for both outputs of leads. No conclusion can be drawn from this.
- 2) People spending more time on the website are more likely to be converted, this is also aligned with our general knowledge.
- 3) 'Page Views Per Visit' also has same median values for both outputs of leads and hence are inconclusive.

COUNTPLOTS FOR CATEGORICAL VARIABLES AGAINST TARGET VARIABLE

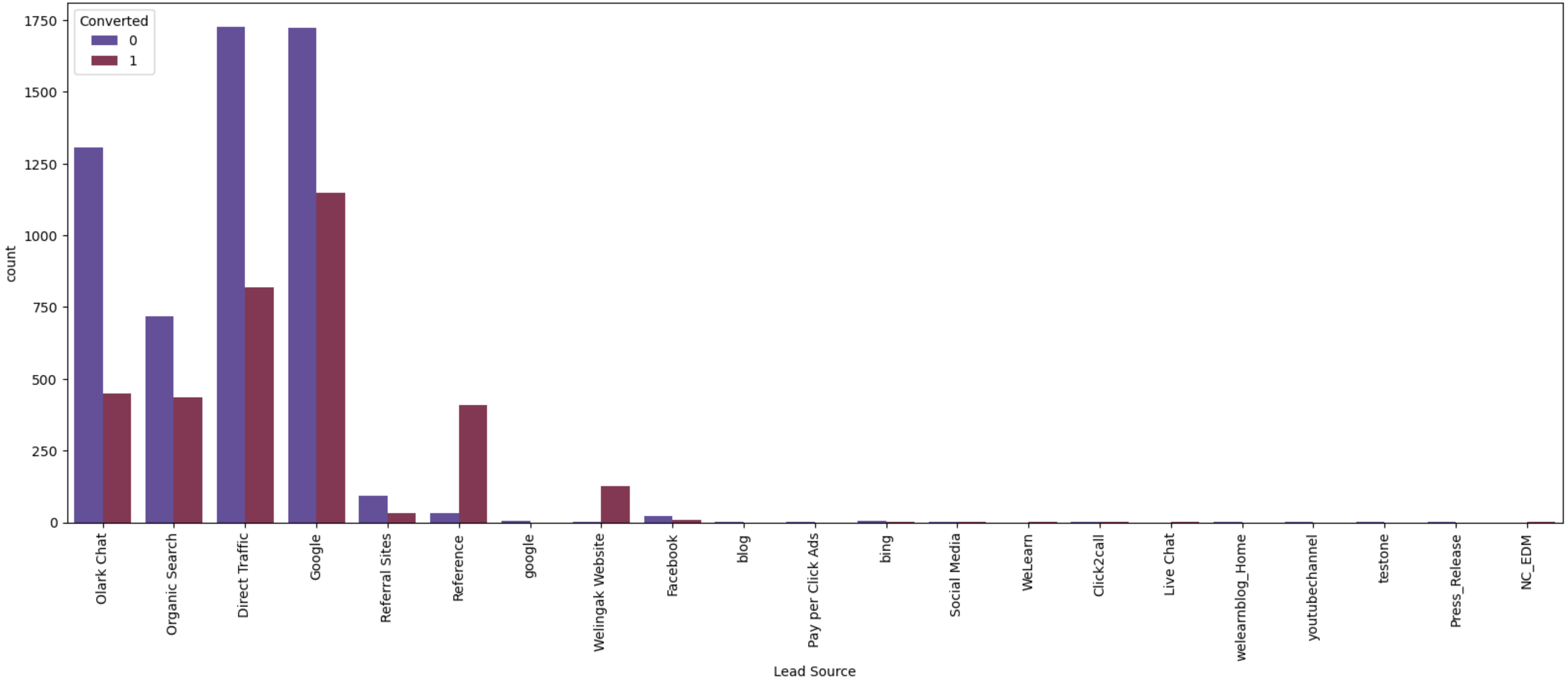


- *Observations for Lead Origin :'API' and 'Landing Page Submission' generate the most leads but have less conversion rates of around 30%. Whereas, 'Lead Add Form' generates less leads but conversion rate is great.*

  We should try to increase conversion rate for 'API' and 'Landing Page Submission', and increase leads generation using 'Lead Add Form'.

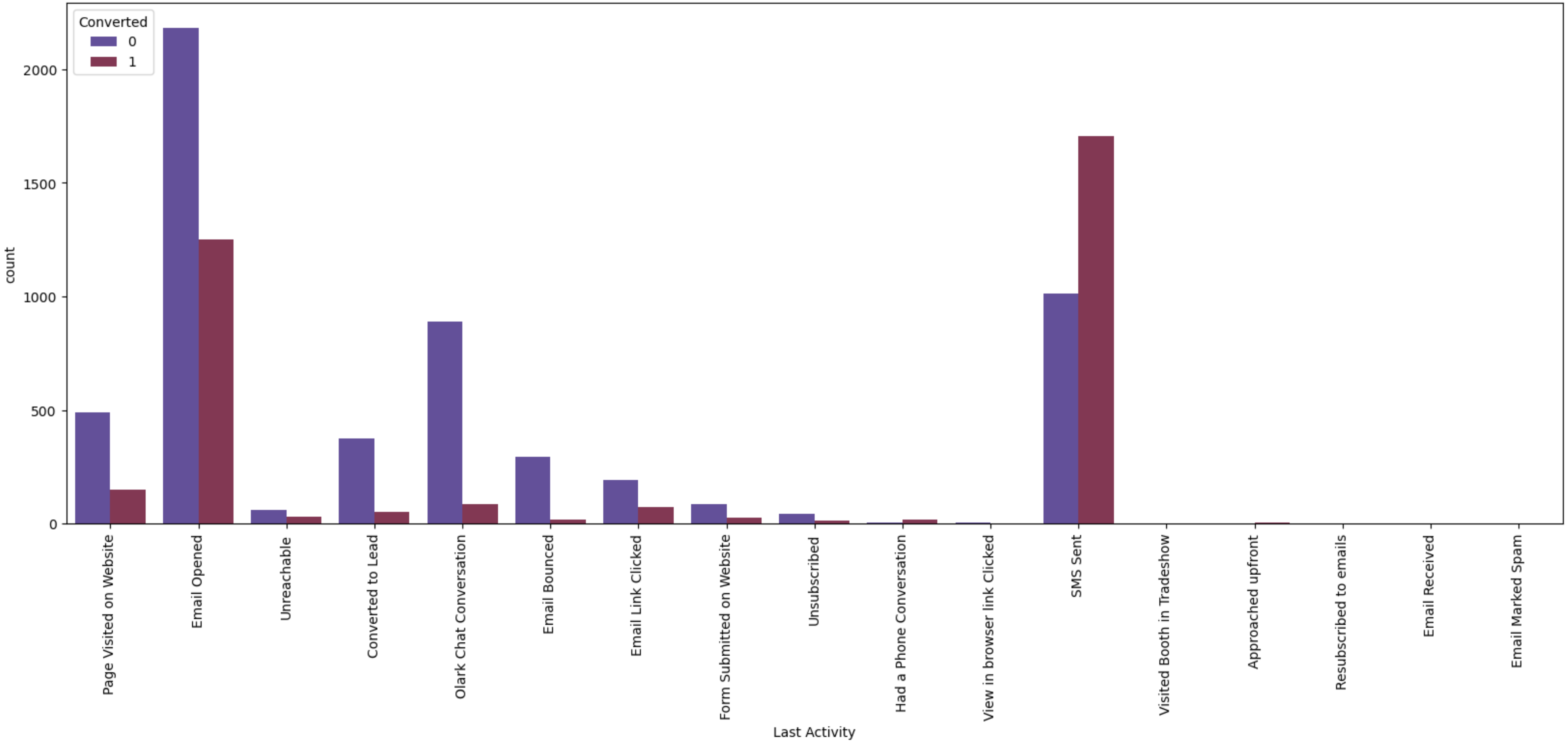
As it can be seen from the graph, number of leads generated by many of the sources are negligible. There are sufficient numbers till Facebook. We can convert all others in one single category of 'Others'.

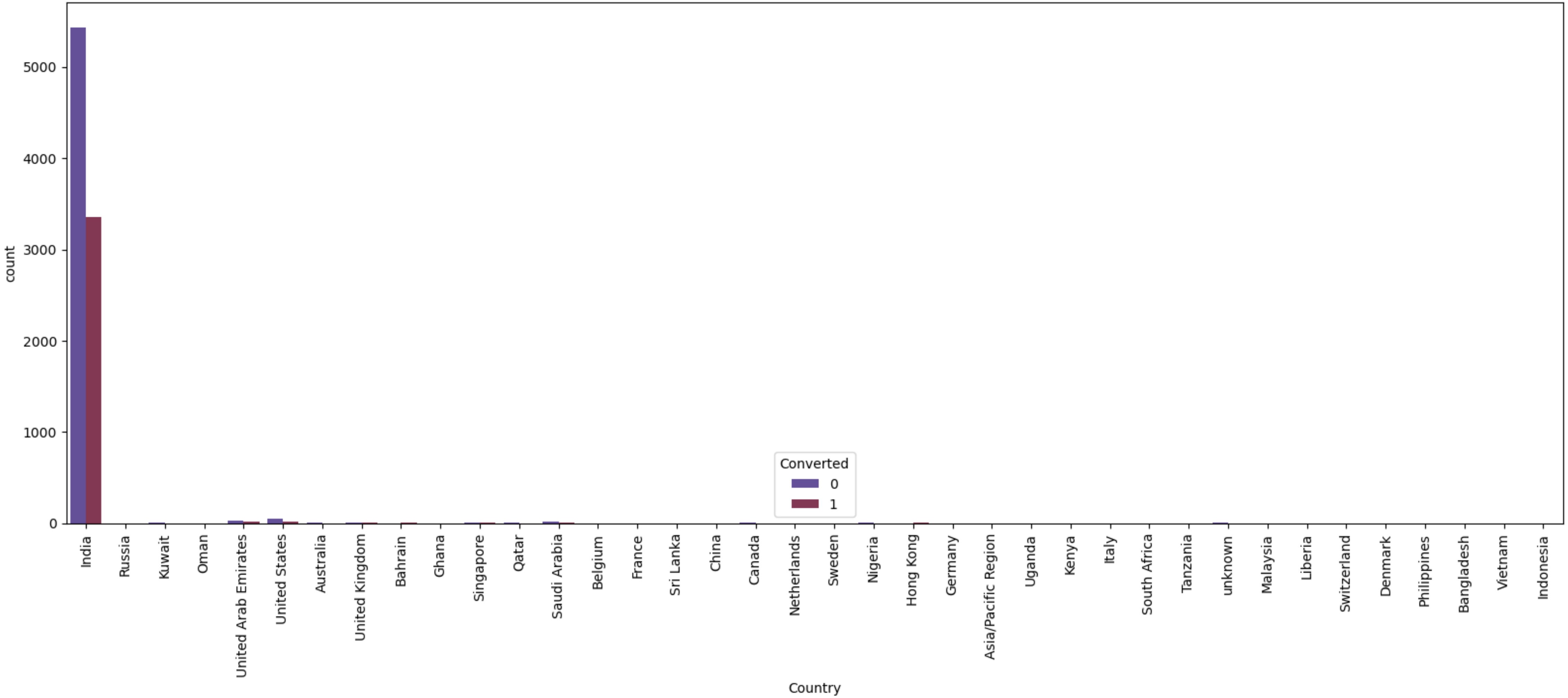
'Direct Traffic' and 'Google' generate maximum number of leads while maximum conversion rate is achieved through 'Reference' and 'Welingak Website'.



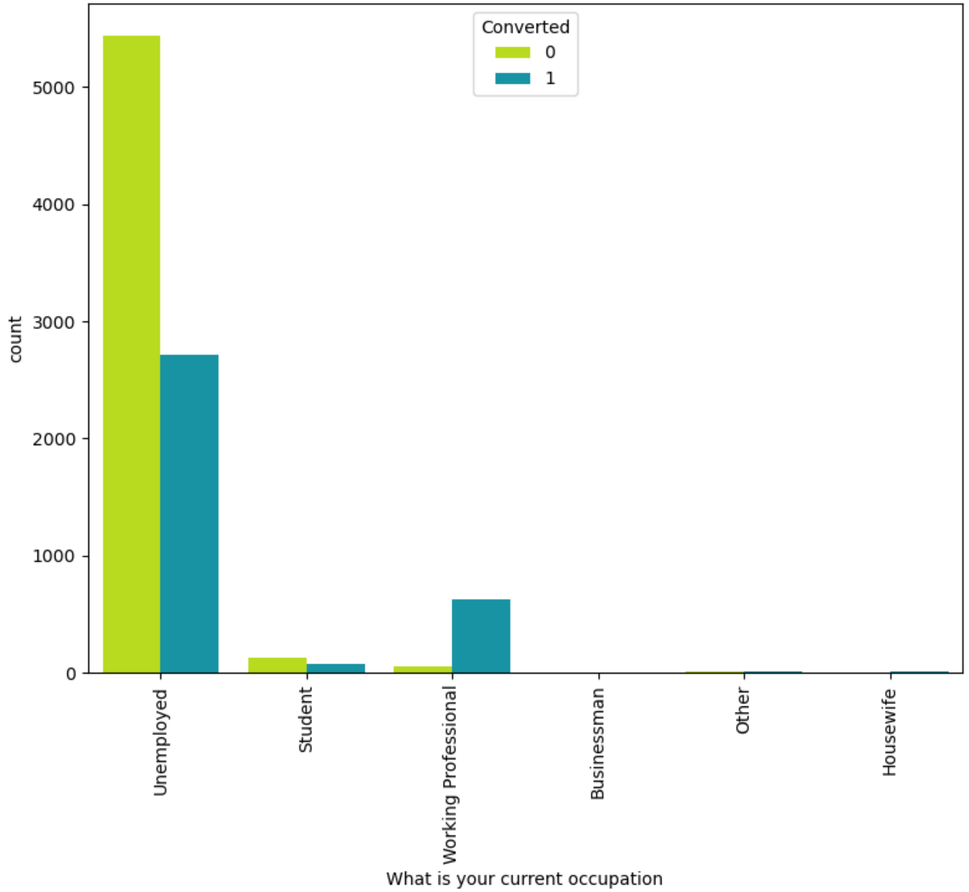
- *Highest number of lead are generated where the last activity is 'Email Opened' while maximum conversion rate is for the activity of 'SMS Sent'. Its conversion rate is significantly high.*

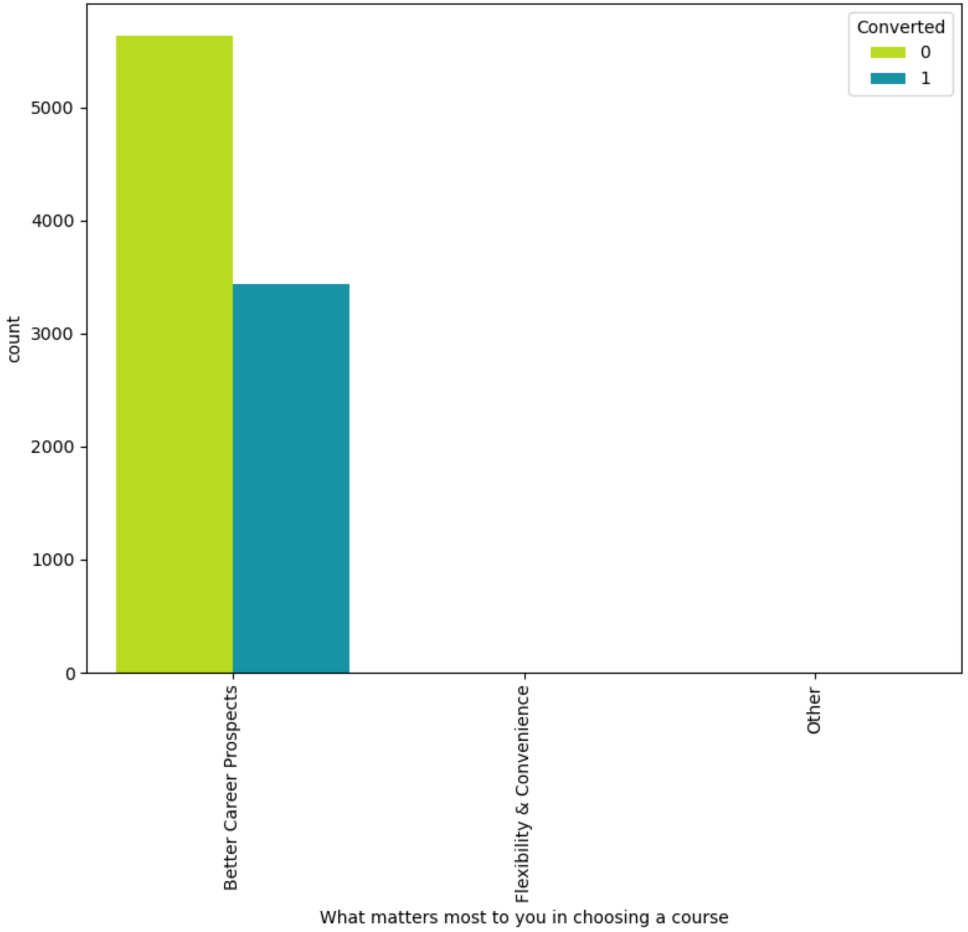
  Categories after the 'SMS Sent' have almost negligible effect. We can aggregate them all in one single category.





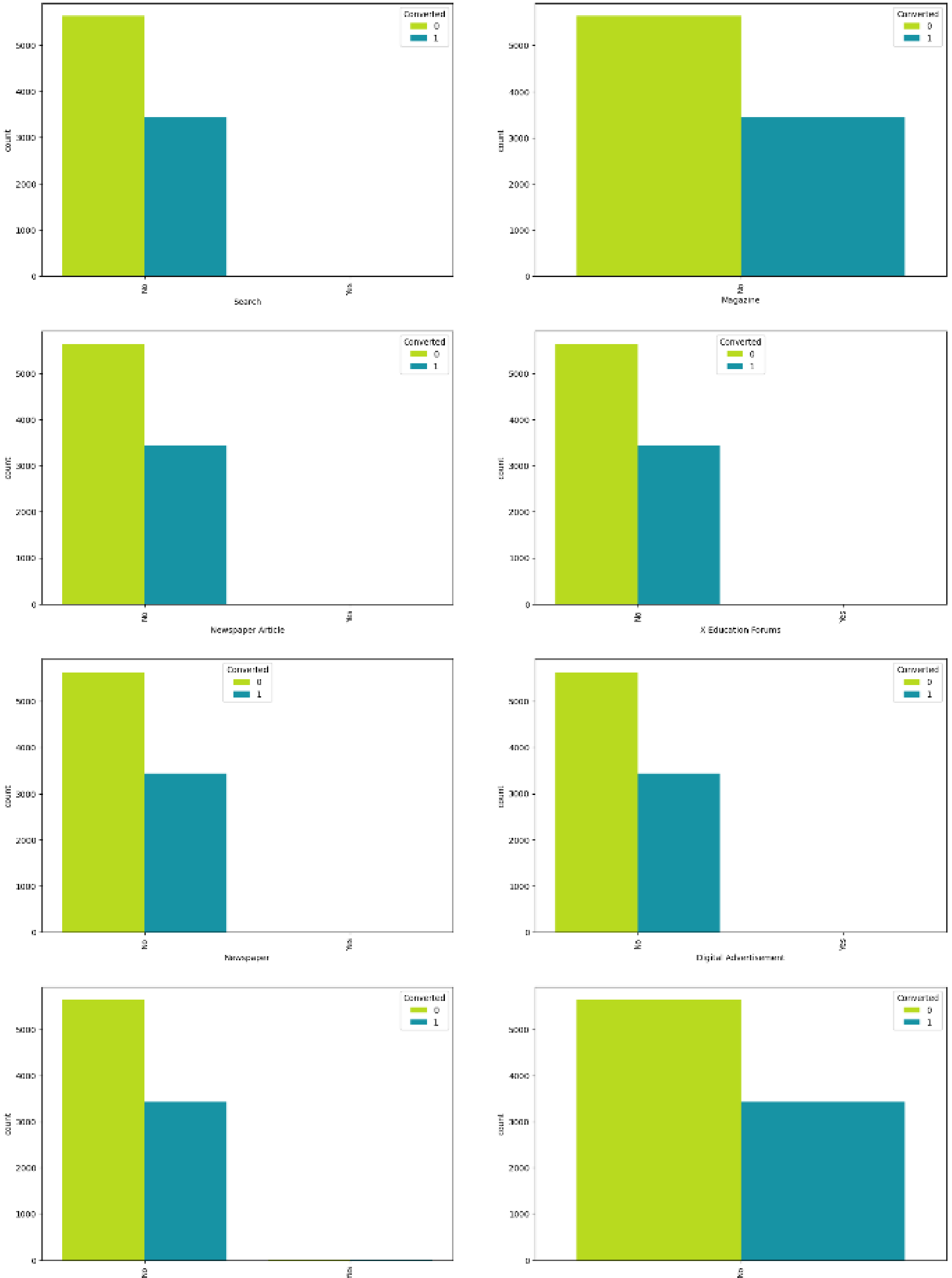
Observations for Country : Most of the responses are for India and others are not significant.





Observations for What is your current occupation and What matters most to you in choosing a course :
1) The highest conversion rate is for 'Working Professional'. High number of leads are generated for 'Unemployed' but conversion rate is low.

2) Variable 'What matters most to you in choosing a course' has only one category with significant count.



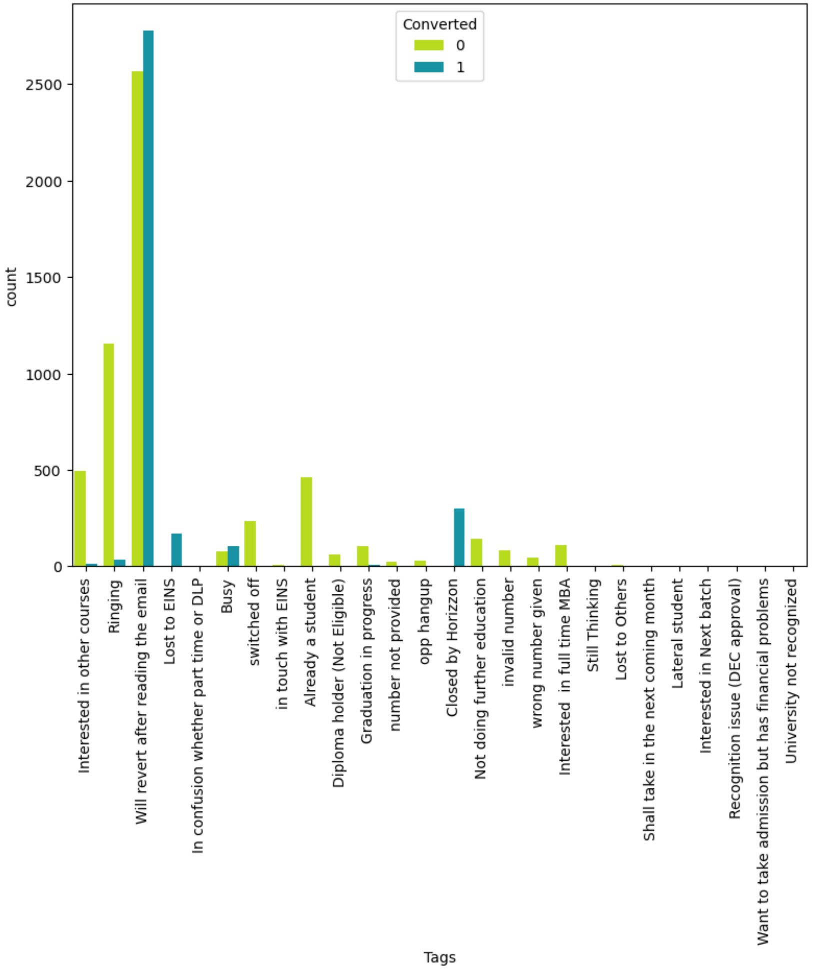
Observations for Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, and Receive More Updates About Our Courses:

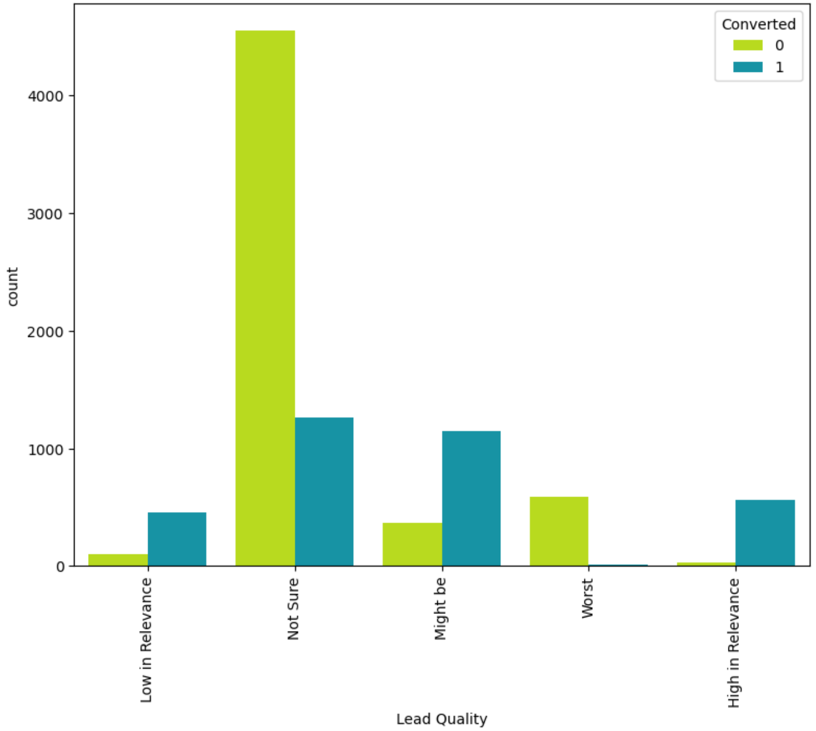
As all the above variables have most of the values as no, nothing significant can be inferred from these plots.

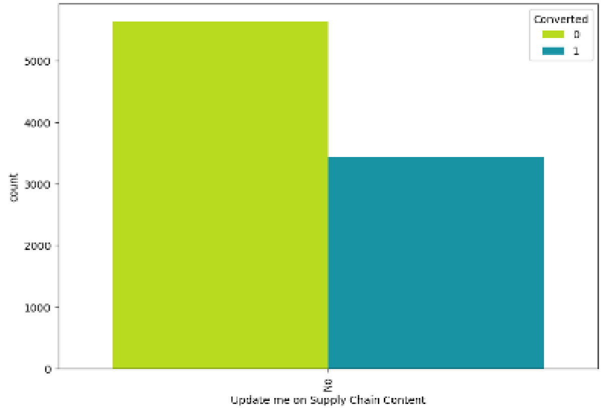
In Tags, categories after 'Interested in full time MBA' have very few leads generated, so we can combine them into one single category.

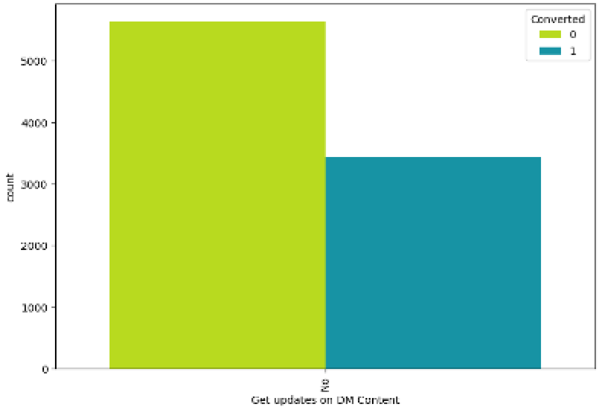
Most leads generated and the highest conversion rate are both attributed to the tag 'Will revert after reading the email'.

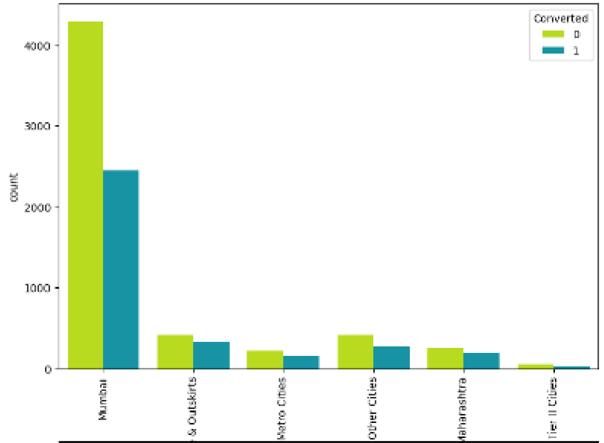
In Lead quality, as expected, 'Might be' as the highest conversion rate while 'Worst' has the lowest.

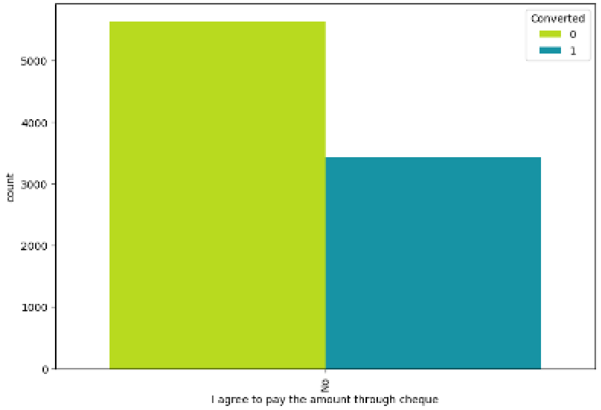


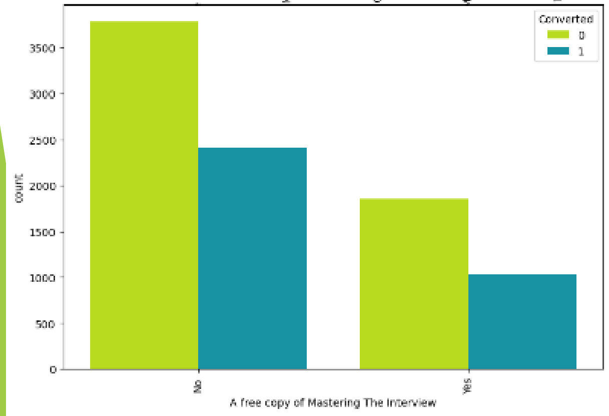


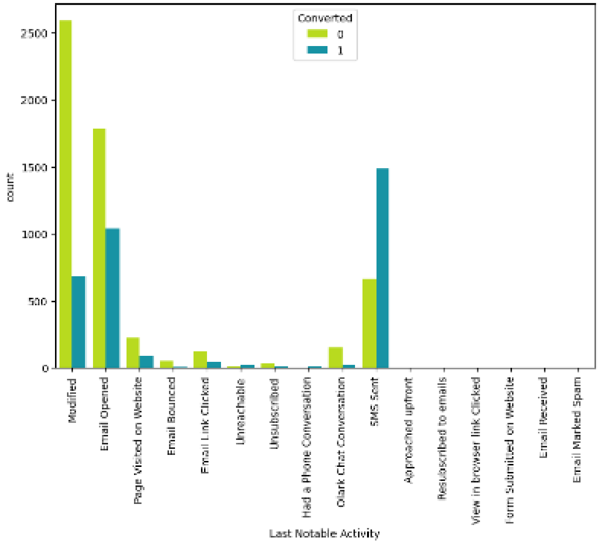










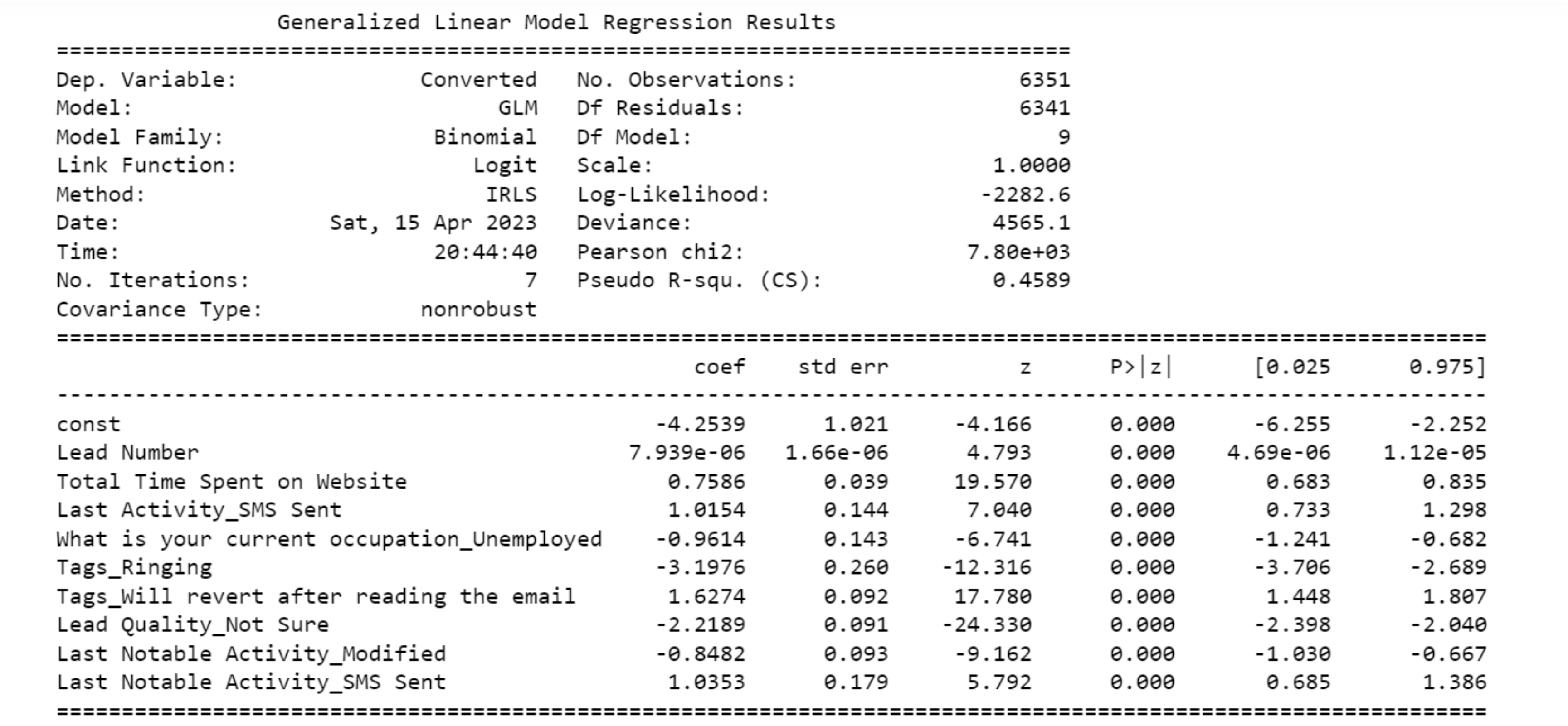


Observations for Update me on Supply Chain Content, Get updates on DM Content, City, I agree to pay the amount through cheque, A free copy of Mastering The Interview, and Last Notable Activity :

1) Most of these variables are insignificant in analysis as many of them only have one significant category 'NO'.

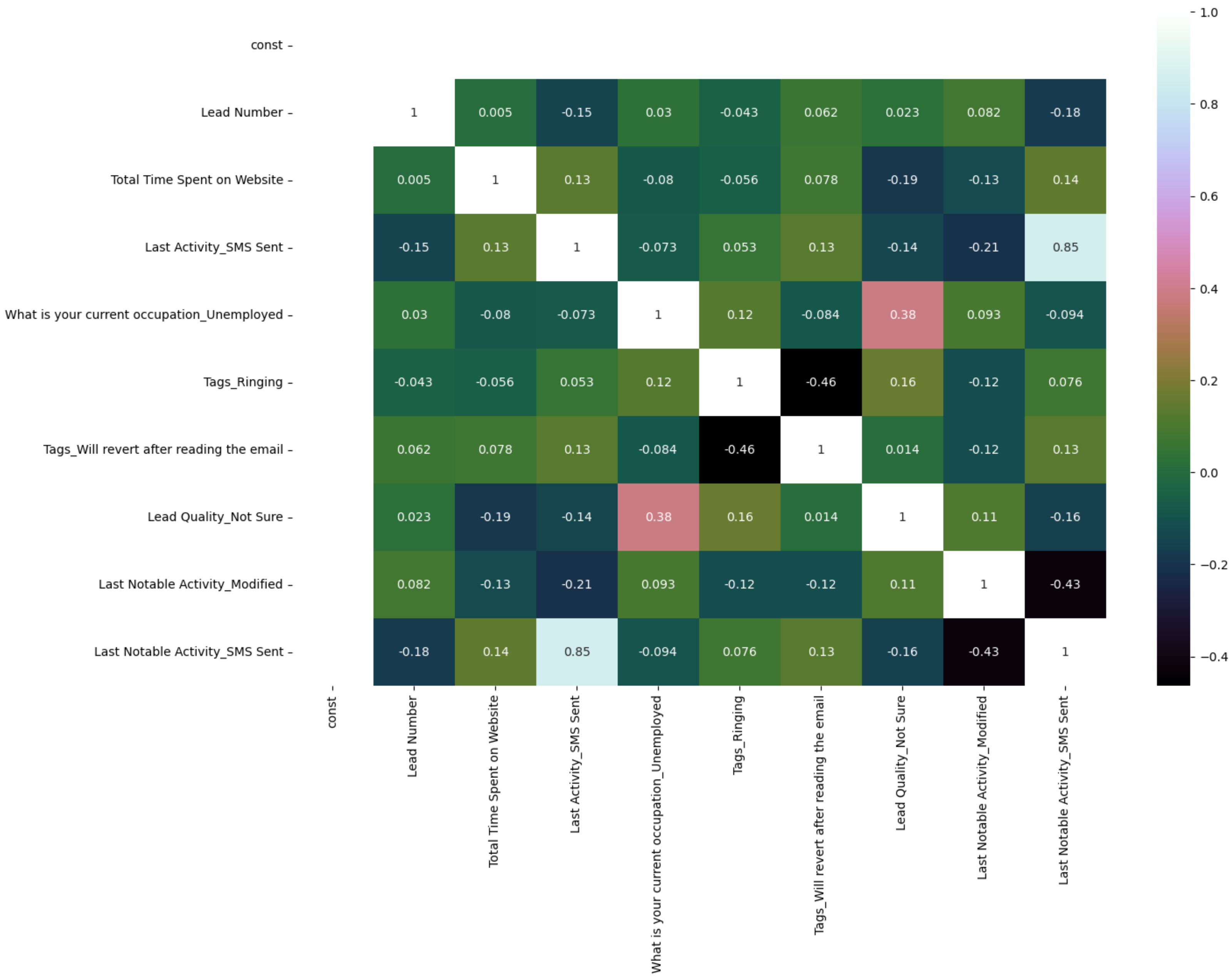
2) In City, most of the leads are generated for 'Mumbai'.

3) In 'A free copy of Mastering The Interview', both categories have similar conversion rates.

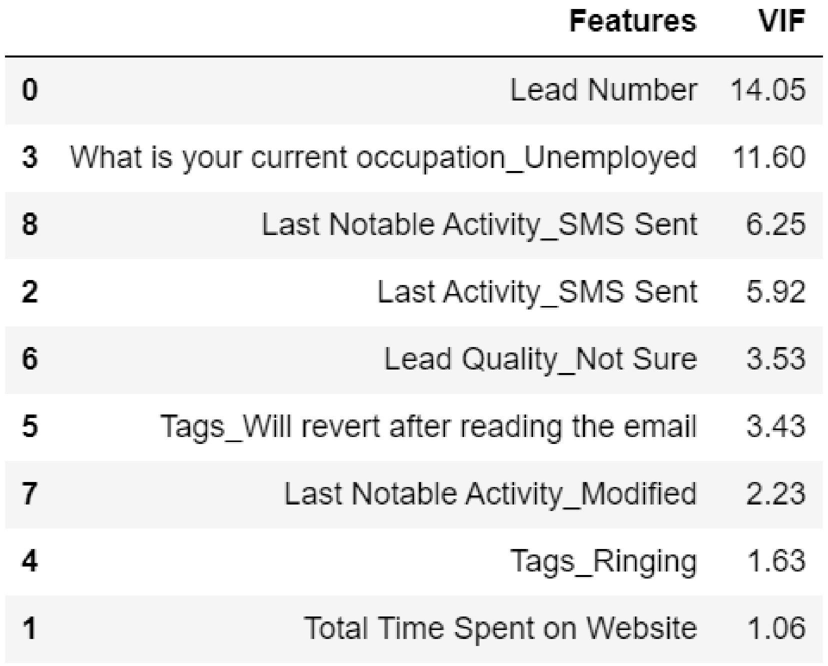


```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6351
Model:                            GLM   Df Residuals:                     6341
Model Family:                Binomial   Df Model:                            9
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2282.6
Date:                Sat, 15 Apr 2023   Deviance:                       4565.1
Time:                        20:44:40   Pearson chi2:                 7.80e+03
No. Iterations:                     7   Pseudo R-squ. (CS):             0.4589
Covariance Type:            nonrobust
==============================================================================================================
                                                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------
const                                         -4.2539      1.021     -4.166      0.000      -6.255      -2.252
Lead Number                                 7.939e-06   1.66e-06      4.793      0.000    4.69e-06    1.12e-05
Total Time Spent on Website                    0.7586      0.039     19.570      0.000       0.683       0.835
Last Activity_SMS Sent                         1.0154      0.144      7.040      0.000       0.733       1.298
What is your current occupation_Unemployed    -0.9614      0.143     -6.741      0.000      -1.241      -0.682
Tags_Ringing                                  -3.1976      0.260    -12.316      0.000      -3.706      -2.689
Tags_Will revert after reading the email       1.6274      0.092     17.780      0.000       1.448       1.807
Lead Quality_Not Sure                         -2.2189      0.091    -24.330      0.000      -2.398      -2.040
Last Notable Activity_Modified                -0.8482      0.093     -9.162      0.000      -1.030      -0.667
Last Notable Activity_SMS Sent                 1.0353      0.179      5.792      0.000       0.685       1.386
==============================================================================================================
```

All of the features have p-value are zero i.e. they all seem significant.

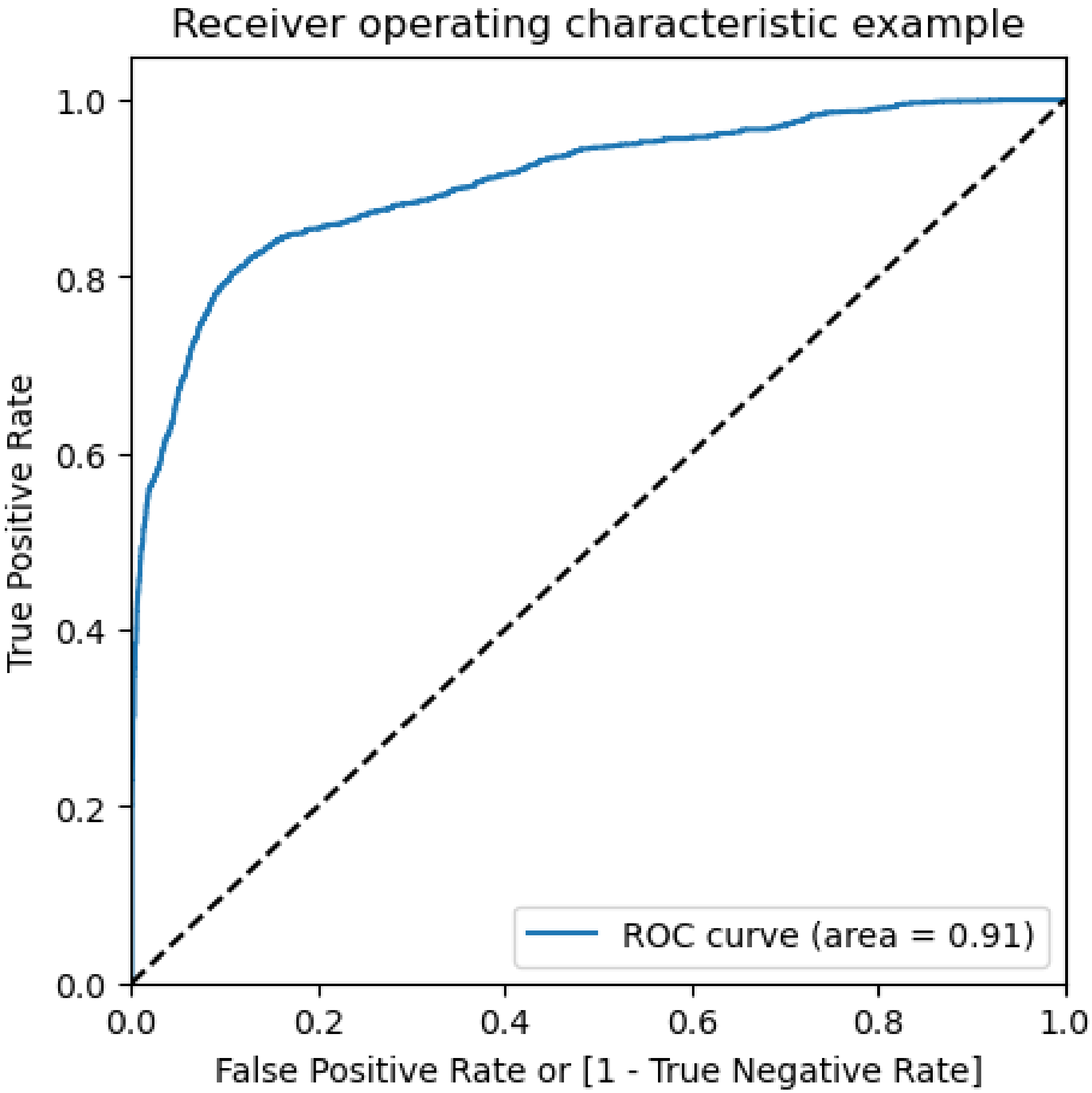


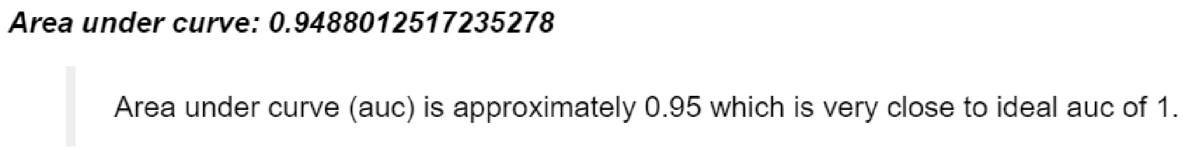
From VIF values and heat maps, we can see that there is not much multicollinearity present.



	Features	VIF
0	Lead Number	14.05
3	What is your current occupation_Unemployed	11.60
8	Last Notable Activity_SMS Sent	6.25
2	Last Activity_SMS Sent	5.92
6	Lead Quality_Not Sure	3.53
5	Tags_Will revert after reading the email	3.43
7	Last Notable Activity_Modified	2.23
4	Tags_Ringing	1.63
1	Total Time Spent on Website	1.06

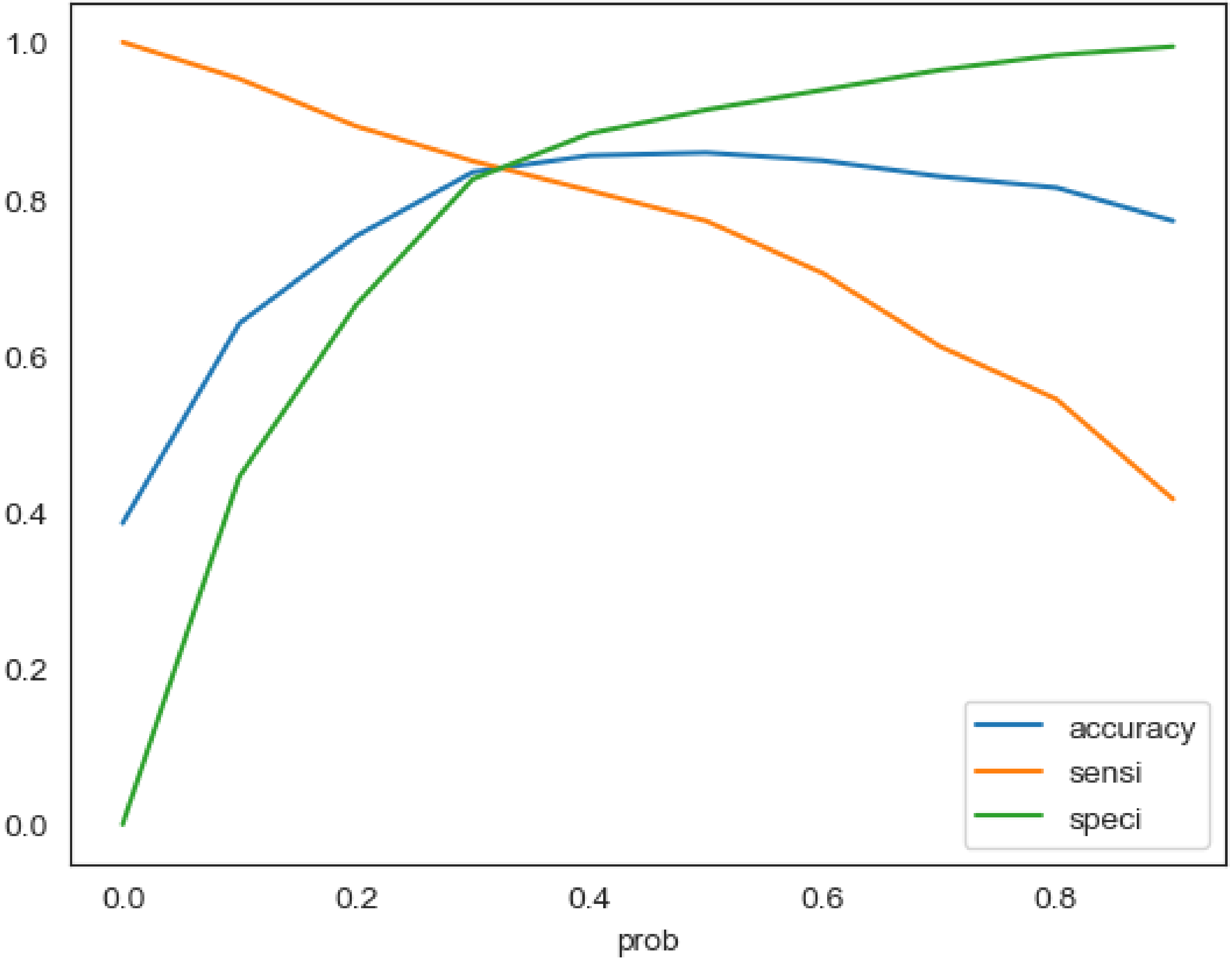
ROC CURVE




Area under curve: 0.9488012517235278

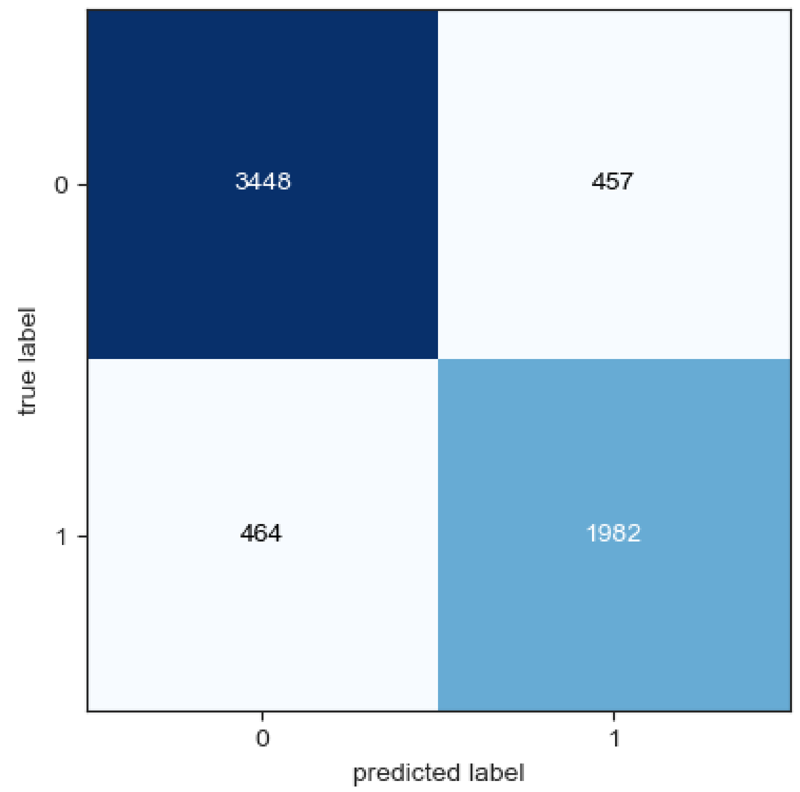
> Area under curve (auc) is approximately 0.95 which is very close to ideal auc of 1.

OPTIMAL CUTOFF POINT:-

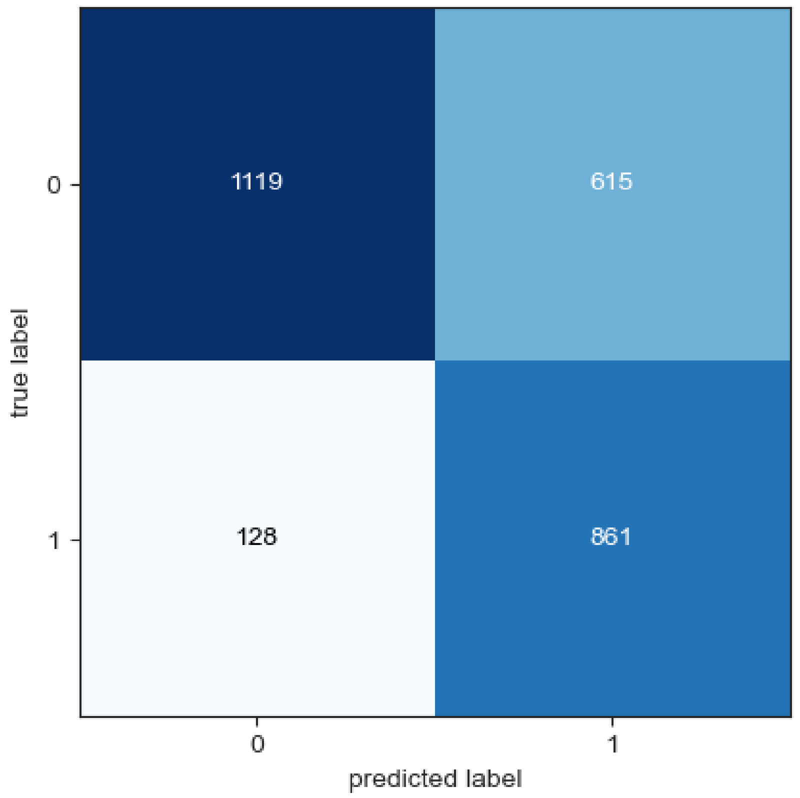


From the curve above, 0.37 is the optimum point to take as a cutoff probability.

CONFUSION MATRIX FOR TRAIN AND TEST SET:-



TRAIN DATA



TEST DATA