Logistic Regression

Lead Score Case-Study

Presented by:

Nikita Pise & Mohit Patil

(nikitapise11@gmail.com) (mohitz4418@gmail.com)

Problem Statement

An education company named X-Education sells online courses to the industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on these websites, 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. Now, although X education gets a lot of leads, its lead conversation rate is very poor. For example, if say, they acquire 100 leads in a day, only about 30 of them are converted.

- X Education has appointed you to help and select the most promising leads.
- The company wants you to build a model wherein you need to assign a lead score to each of the leads such that the customer. with higher lead score have higher conversion chance and the customer with the lower lead score have lower conversion chance.

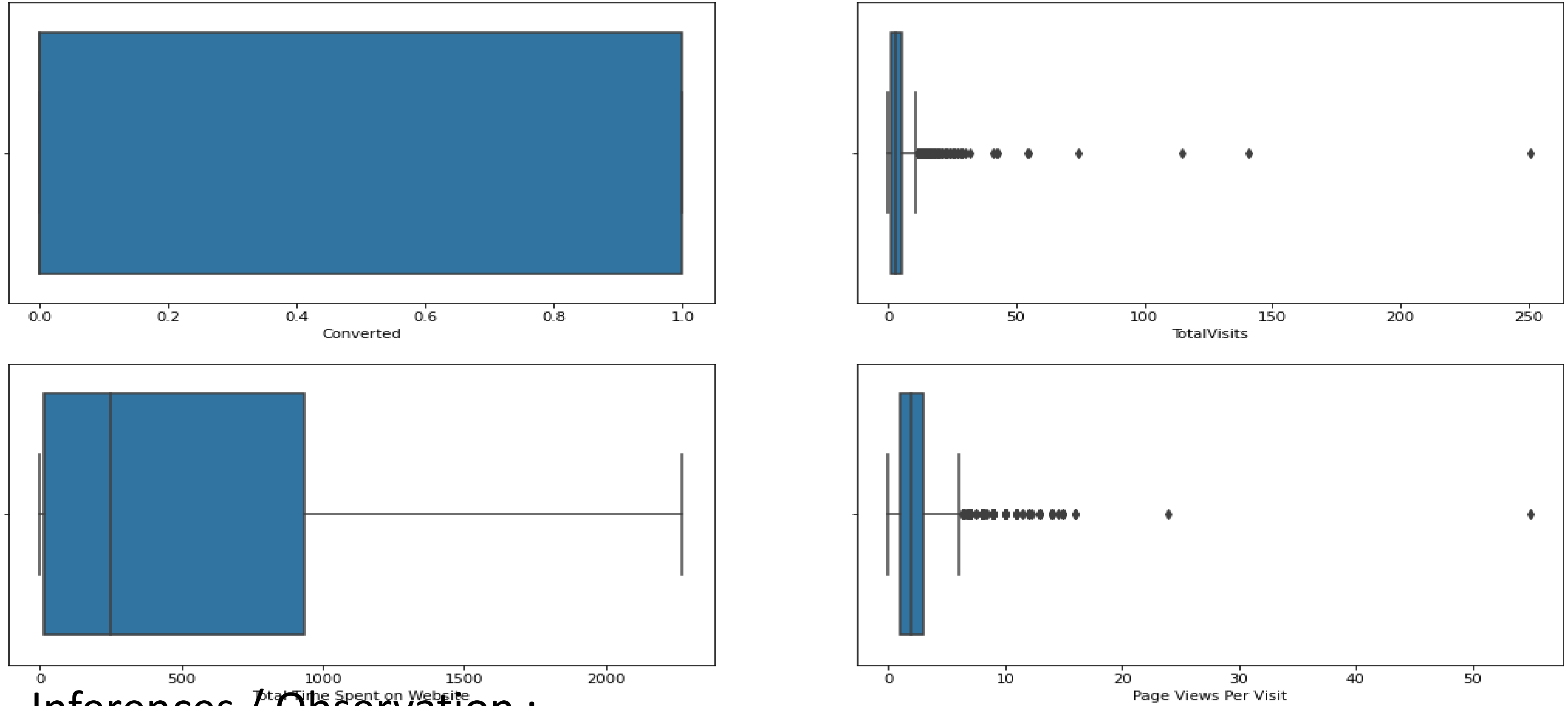
Data Cleaning and Imputation –

Data quality check and missing values

- There are Categories named ‘Select’ in all the categorical columns; This is nothing but the empty field left by the customers ;Therefore we proceeded replacing the ‘Select’ field by np.nan.
- There are many columns with large amount of missing values and categorical columns with skewed data .
- Variables with more than 50% of missing values are dropped because it may affect the analysis.
- Also, we just checked the best matrix such as mean, median or mode to impute the missing values of variables having relatively less missing values
- There were few levels in categorical columns with very less amount of percentage so those categories are clubbed in a separate category called other and such categories are removed from all the categorical columns.

Data Cleaning and outlier treatment –

Numerical variables – Checked the presence of outliers

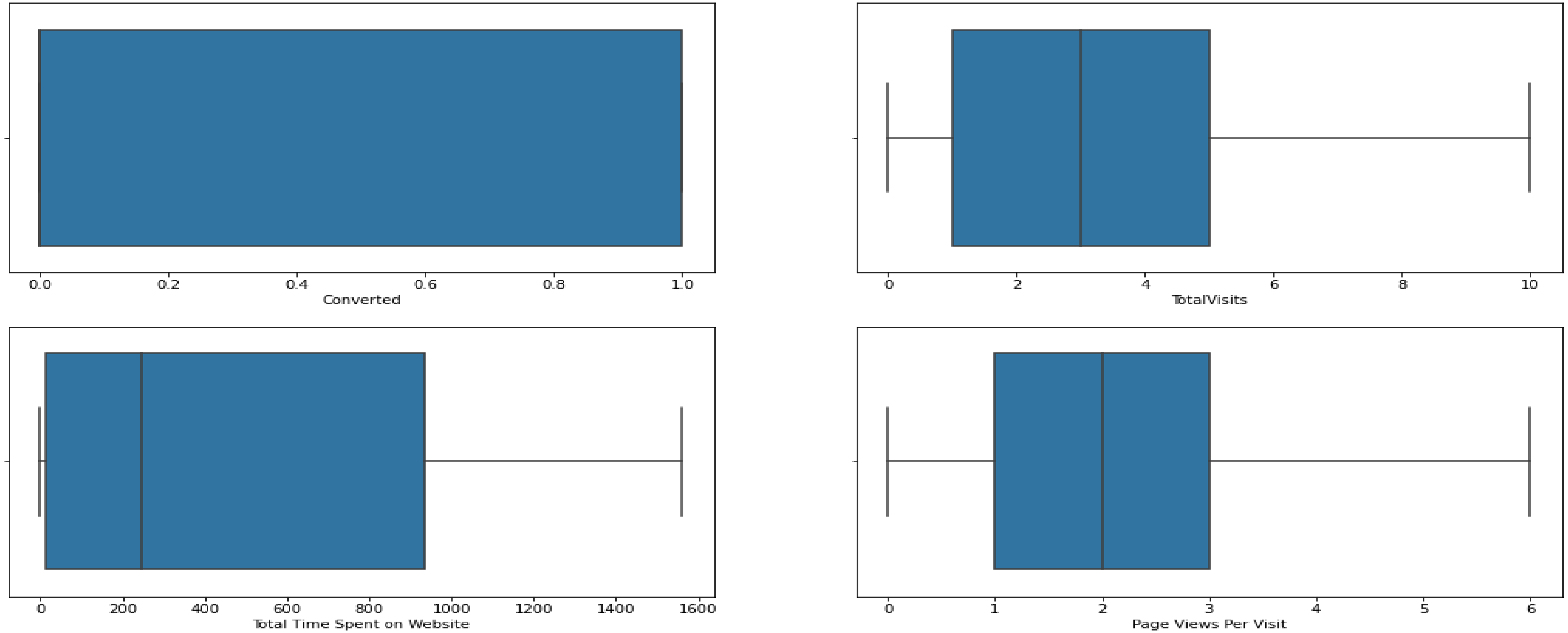


Inferences / Observation :-

There are so many outliers present in only two variables. Hence , the we have capped the outliers within the soft range.

Data Cleaning and outlier treatment –

Numerical variables –Outliers Capping



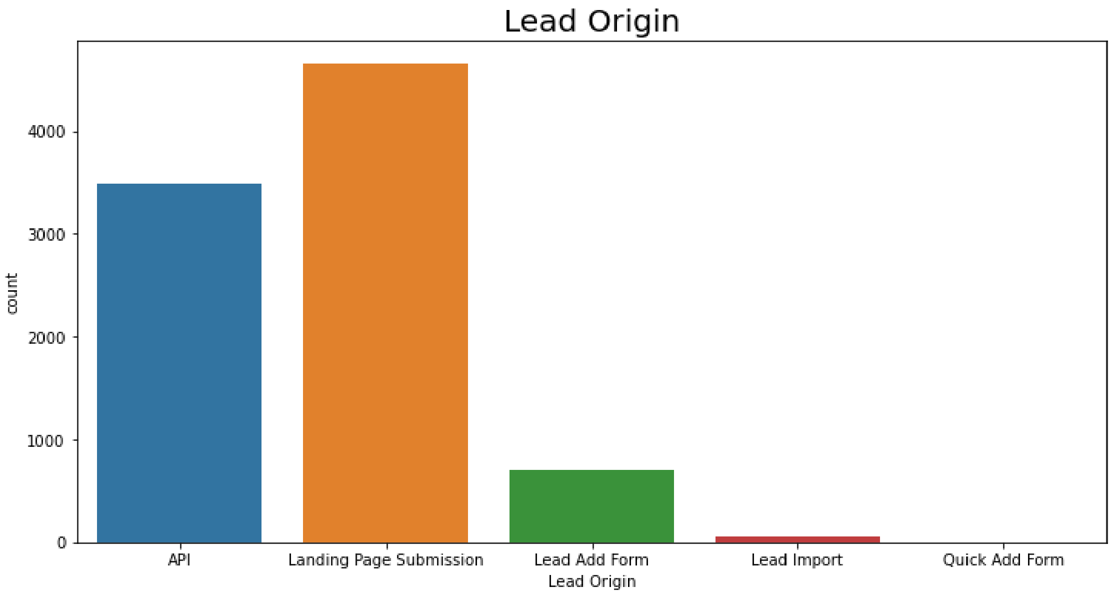
Inferences / Observation :-

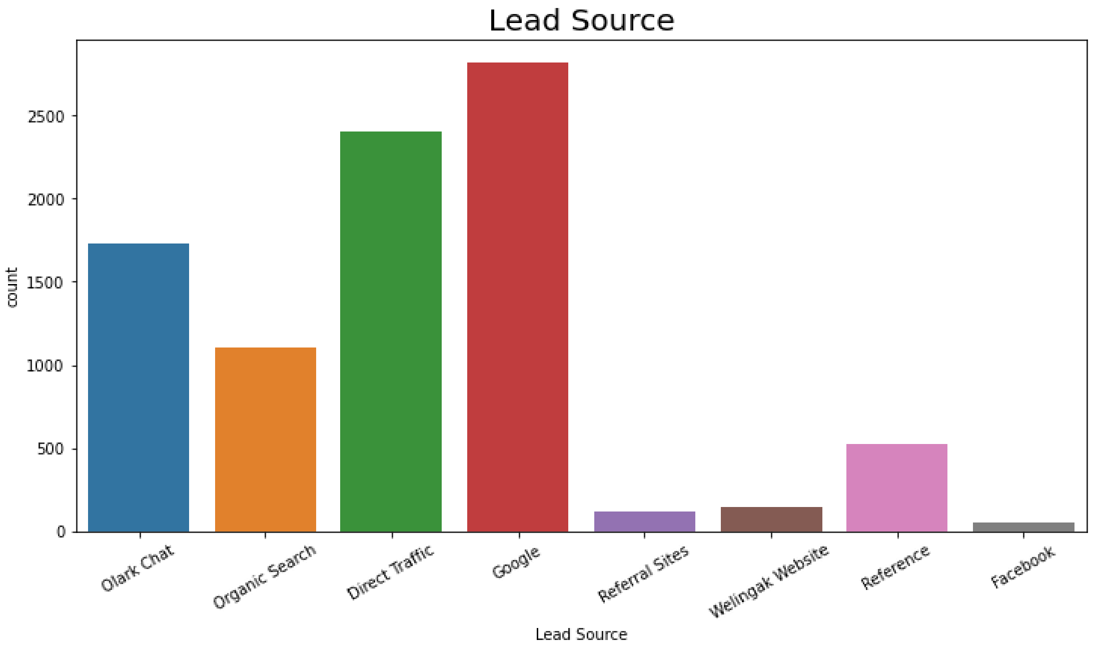
- Now there are no outliers as we have performed capping now we can proceed with the further cleaning

Analysis–

Univariate analysis for categorical variables.

- Univariate analysis are carried out on 6 categorical variables i.e. Lead Origin, Lead Source, Last Activity, Specialization, What is your current occupation and City.



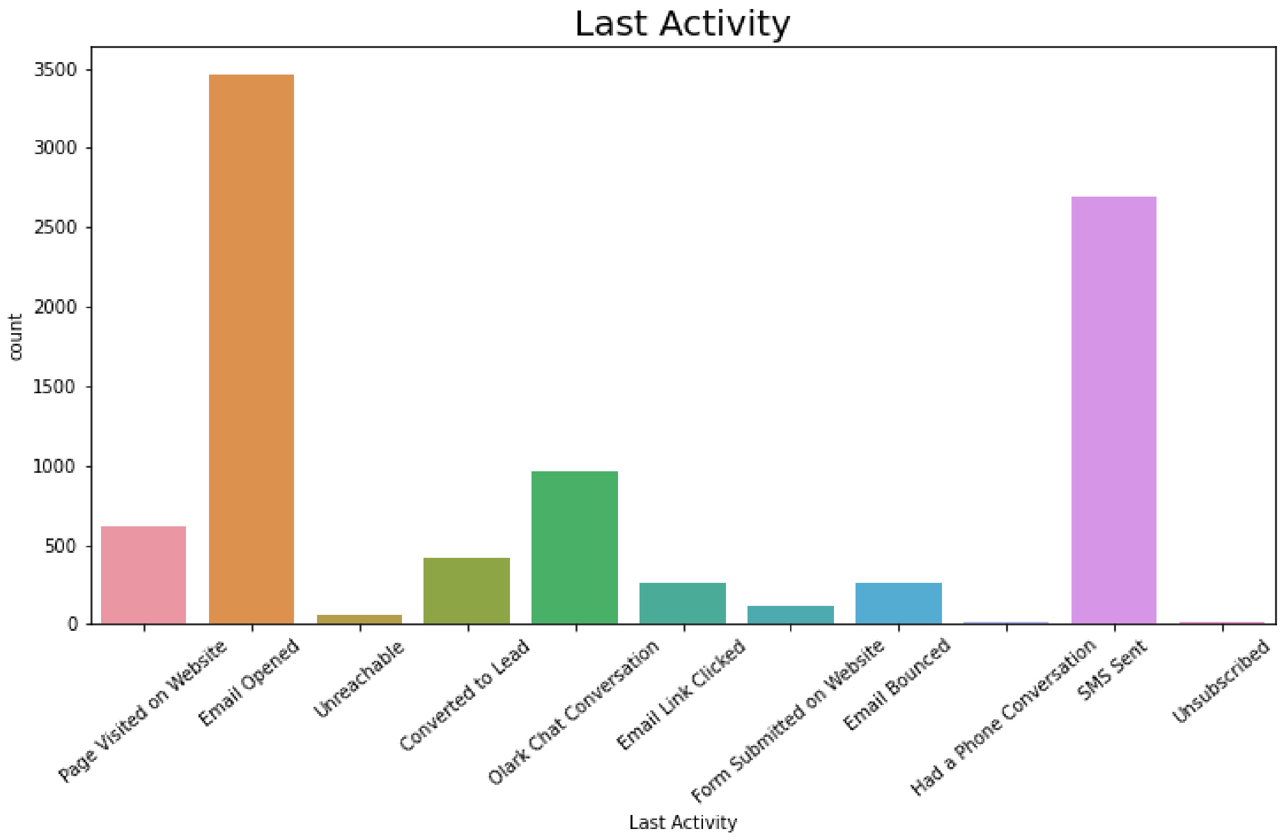


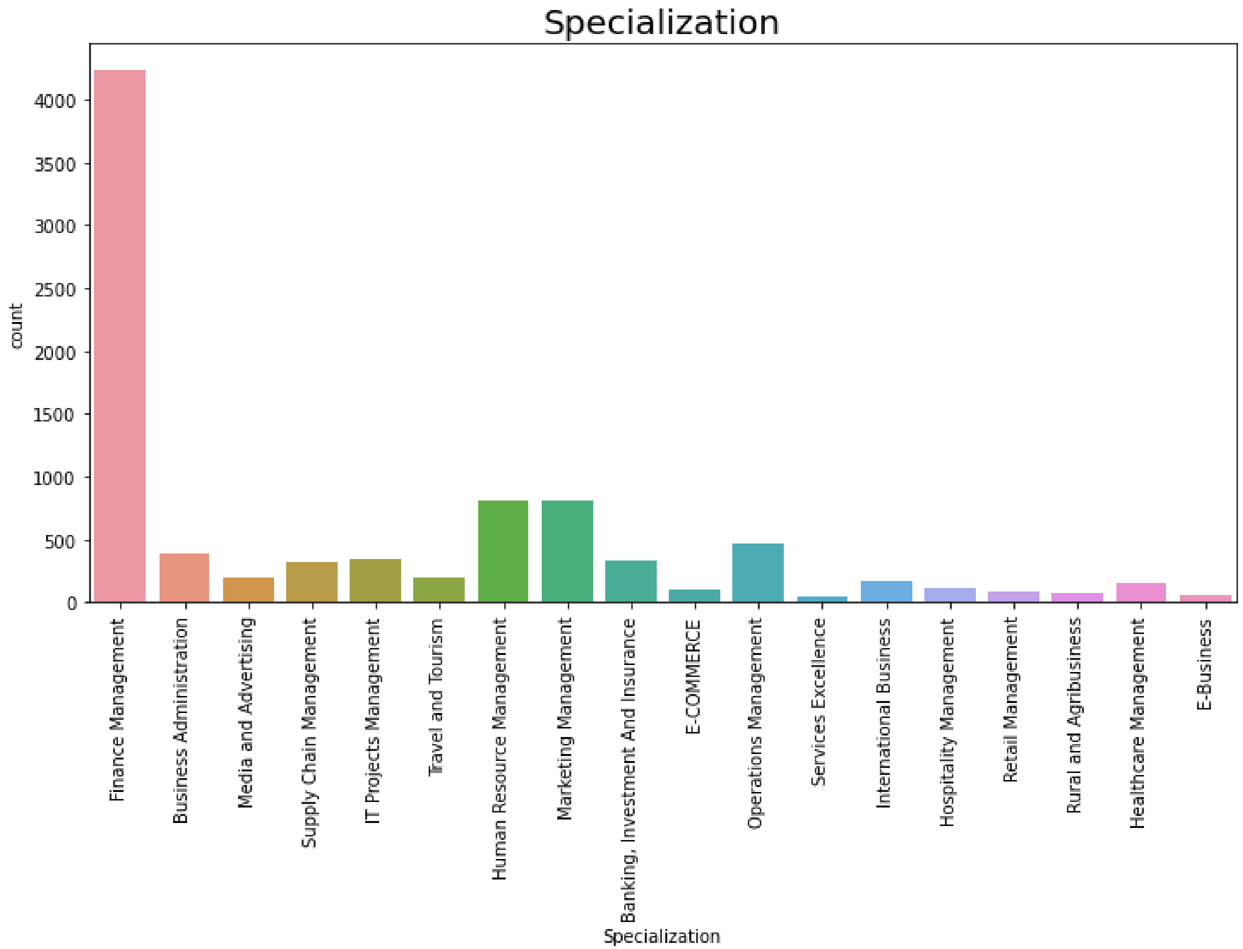
- Most of the customer is identified by Landing Page Submission.

•Most of the customer are from Google and Direct traffic to X education website.

Analysis–

Univariate analysis for categorical variables

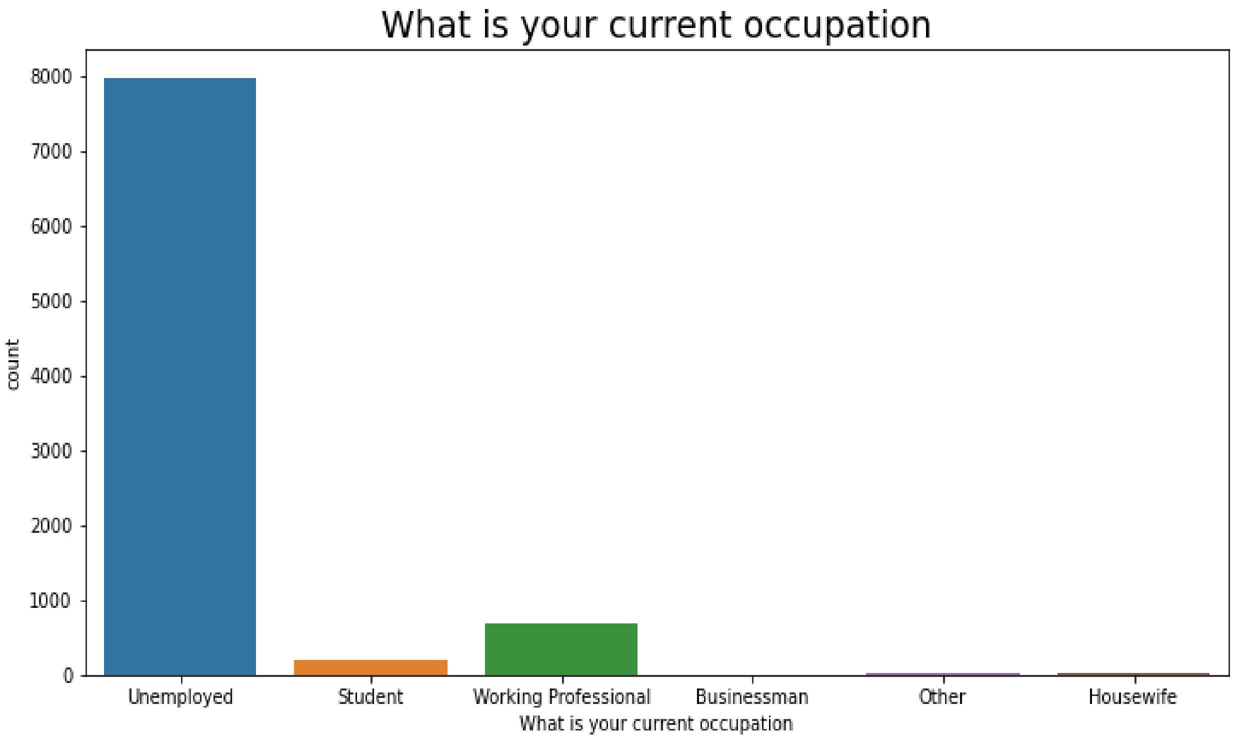


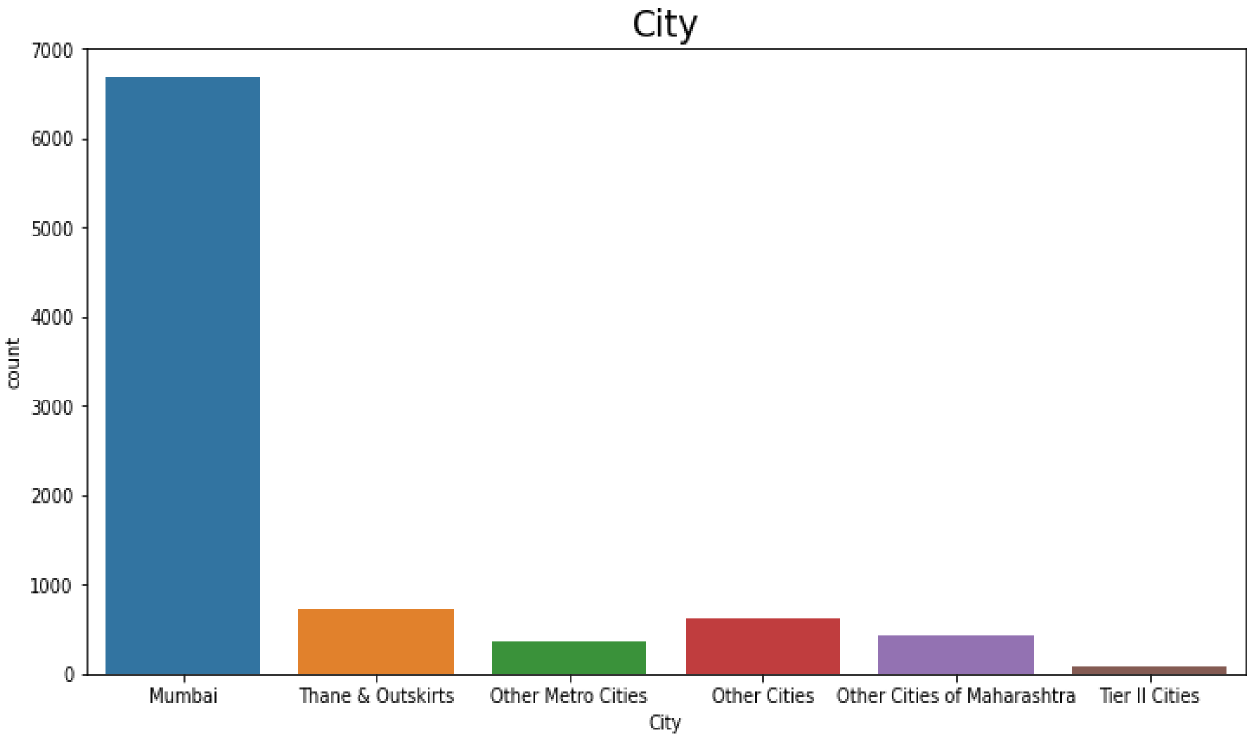


- Most of the customers have performed last activity as Email Opened and SMS Sent.
- Most of the customers are from Finance Management specialization and worked there before.

Analysis–

Univariate analysis for categorical variables

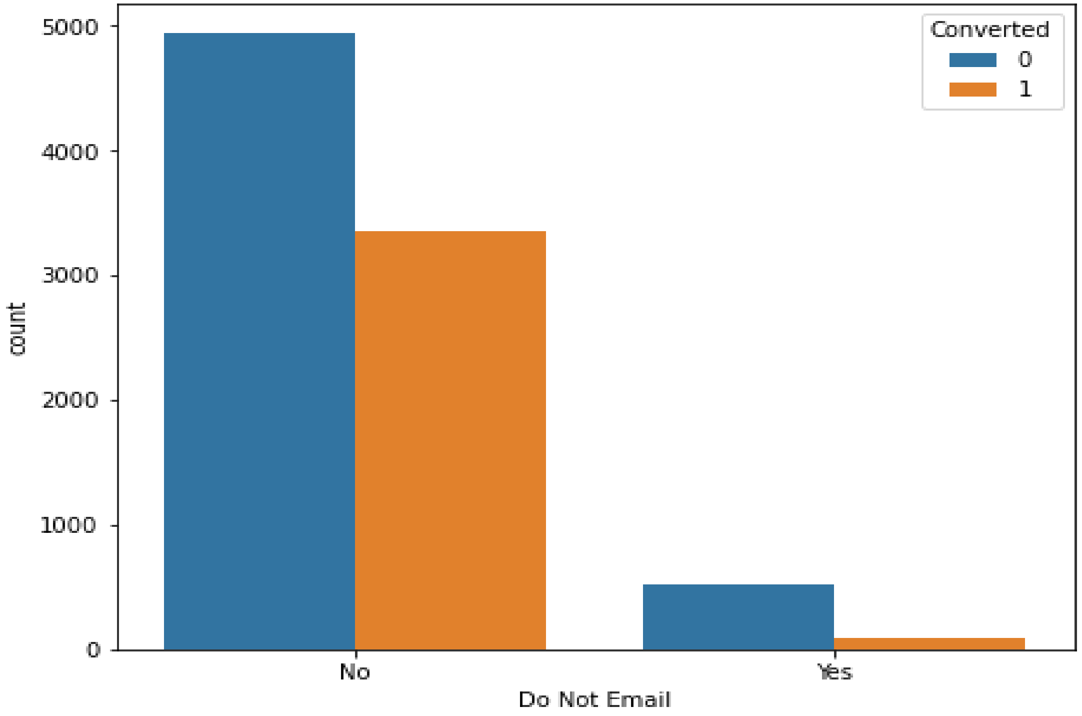


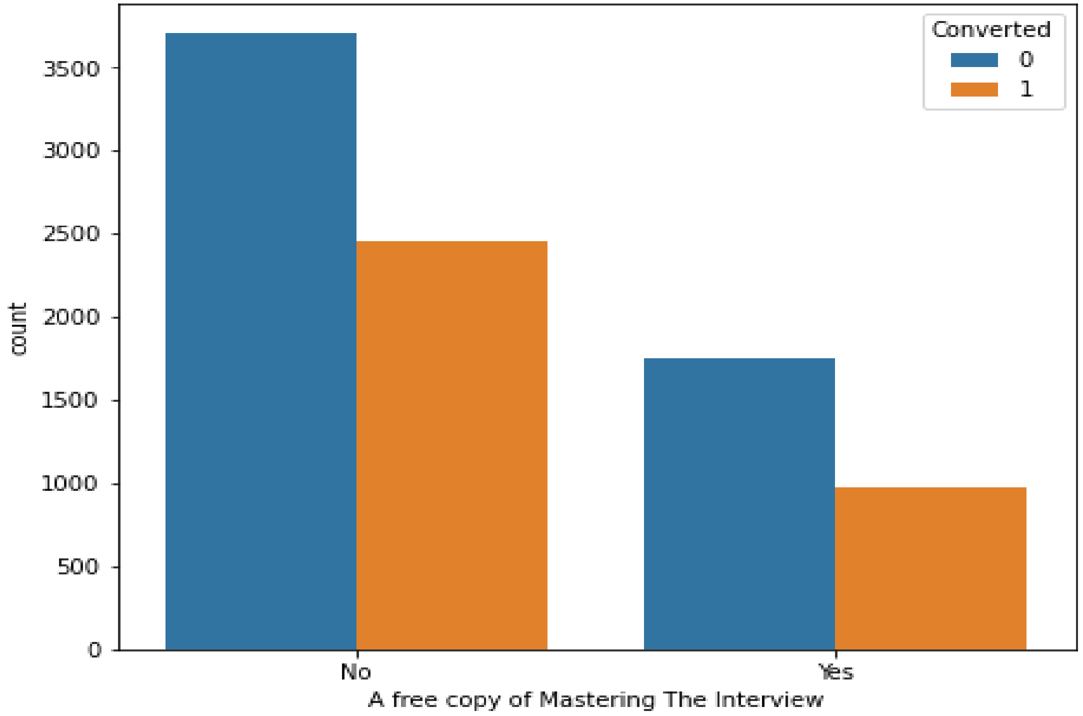


- Most of the customers are Unemployed.
- Most of the customers are from Mumbai City.

Analysis–

Univariate analysis for categorical variables with target variable

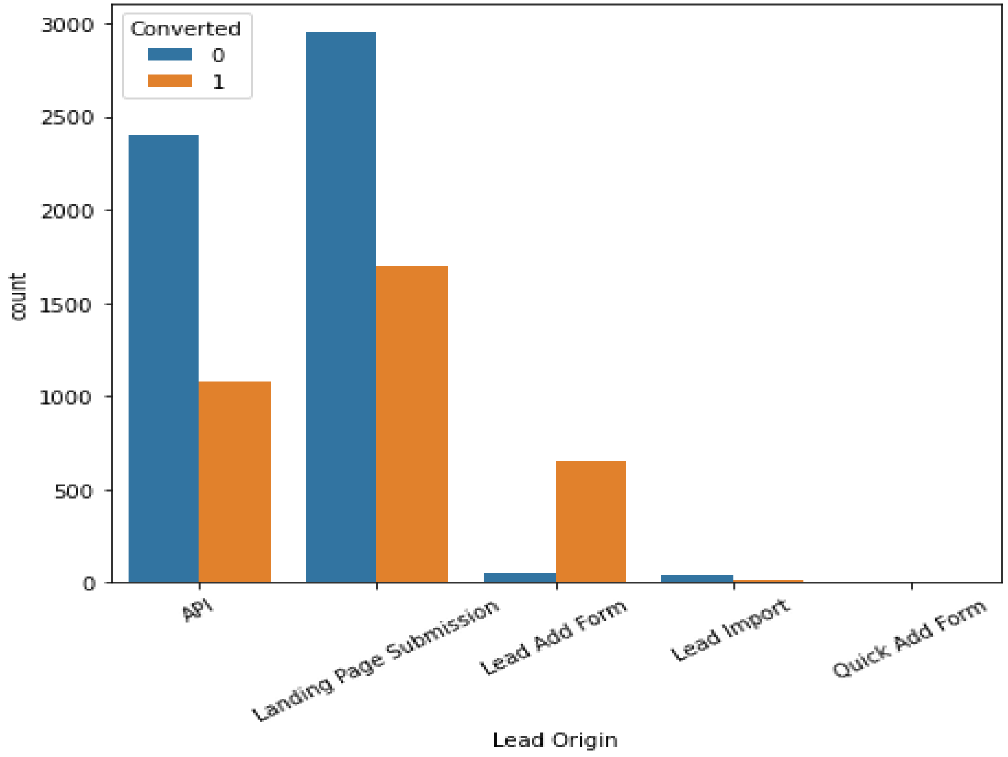


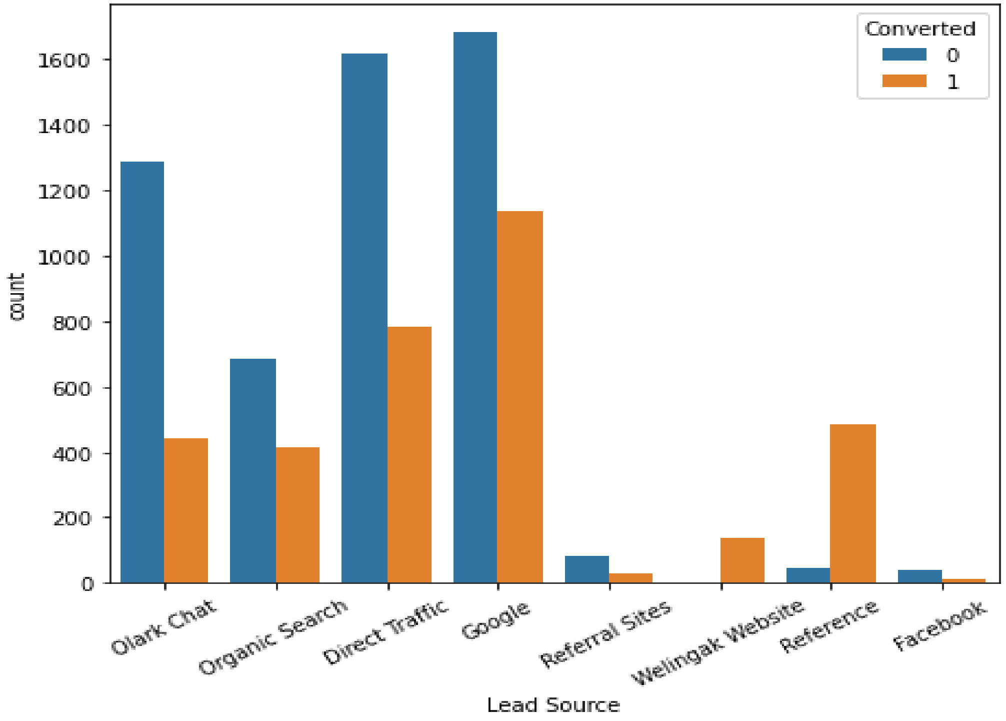


- Most of the leads do not want to be emailed about the course and also do not want the free copy of mastering the interview.
- Those leads who do not want to be emailed have high chances of getting converted.

Analysis–

Univariate analysis for categorical variables with target variable





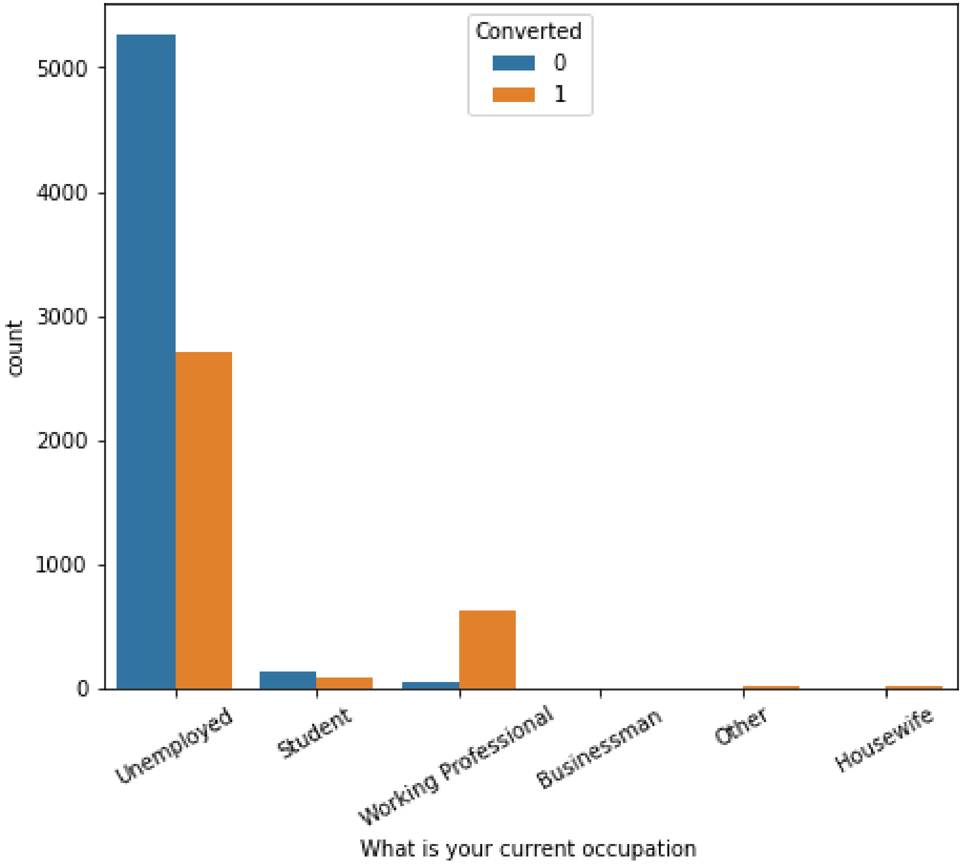
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- API and Landing Page Submission bring higher number of leads as well as conversion.
- In order to improve overall lead conversion rate, we need to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

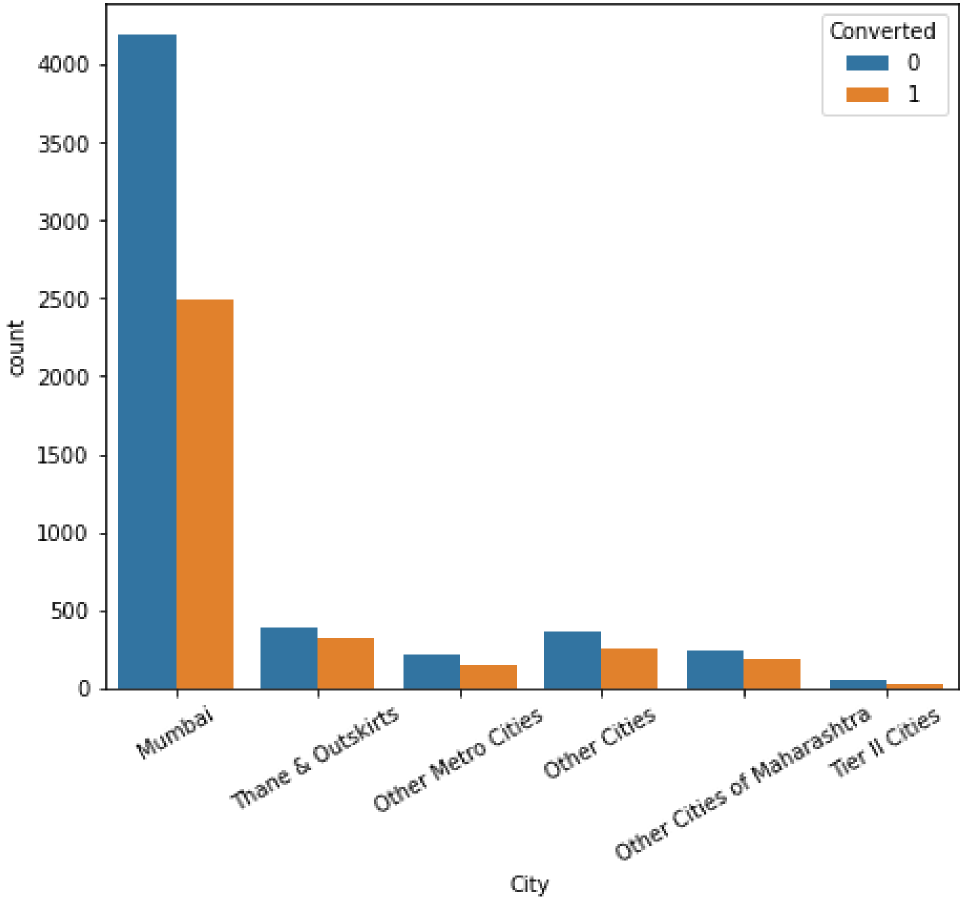
•Reference and Welingak Website have very high conversion rate but count of leads are not very high.

•Google, Direct Traffic and Olark Chat bring higher number of leads as well as conversion.

Analysis–

Univariate analysis for categorical variables with target variable

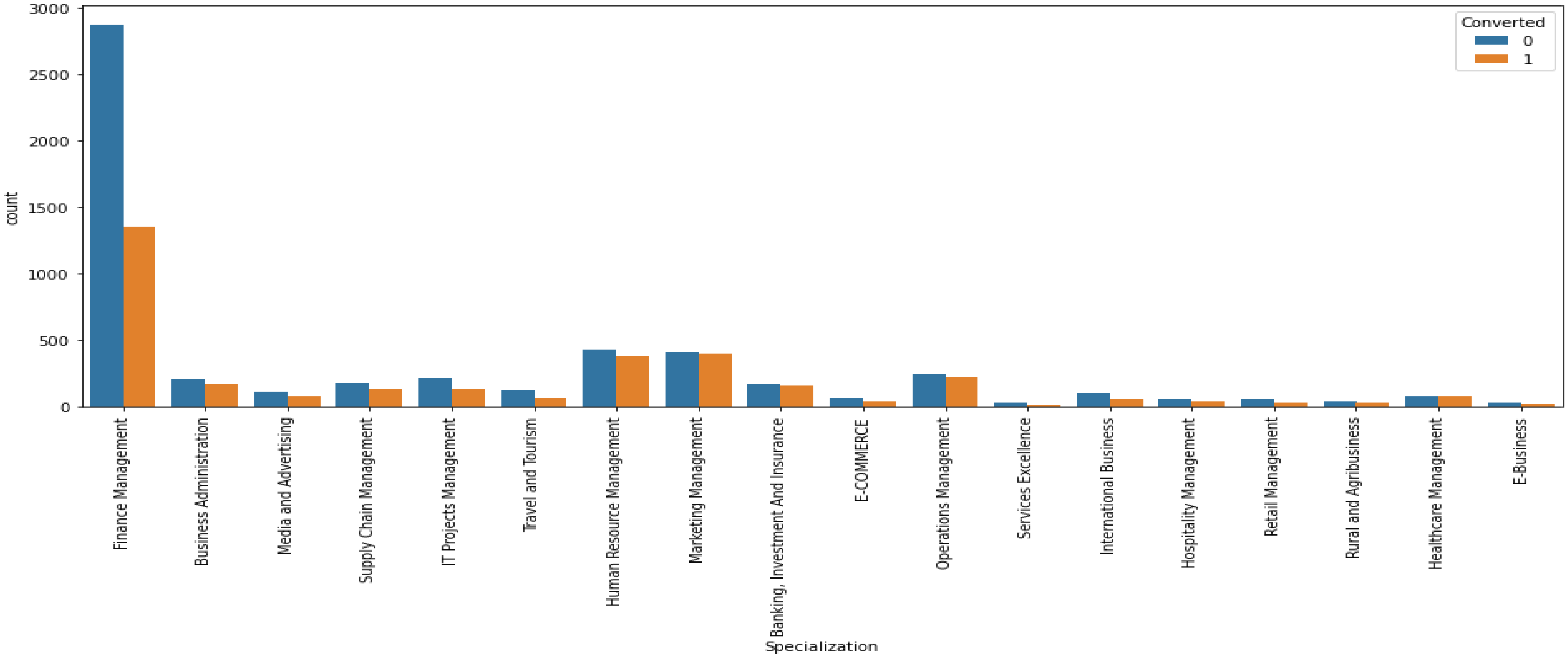




- Working professional checking about the course have high chances of joining the course.
- Higher number of leads as well as conversion from Unemployed category.
- Mumbai has the highest number of leads as well as conversion.

Analysis–

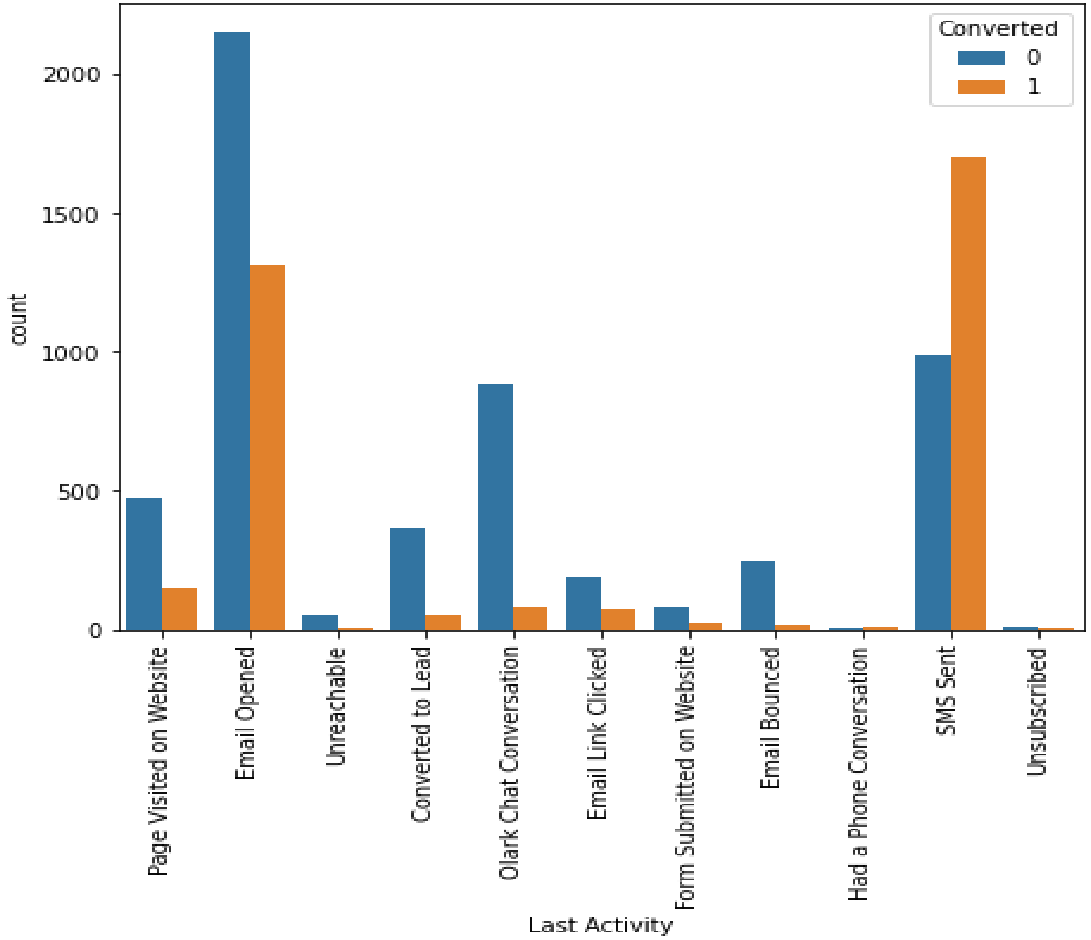
Univariate analysis for categorical variables with target variable

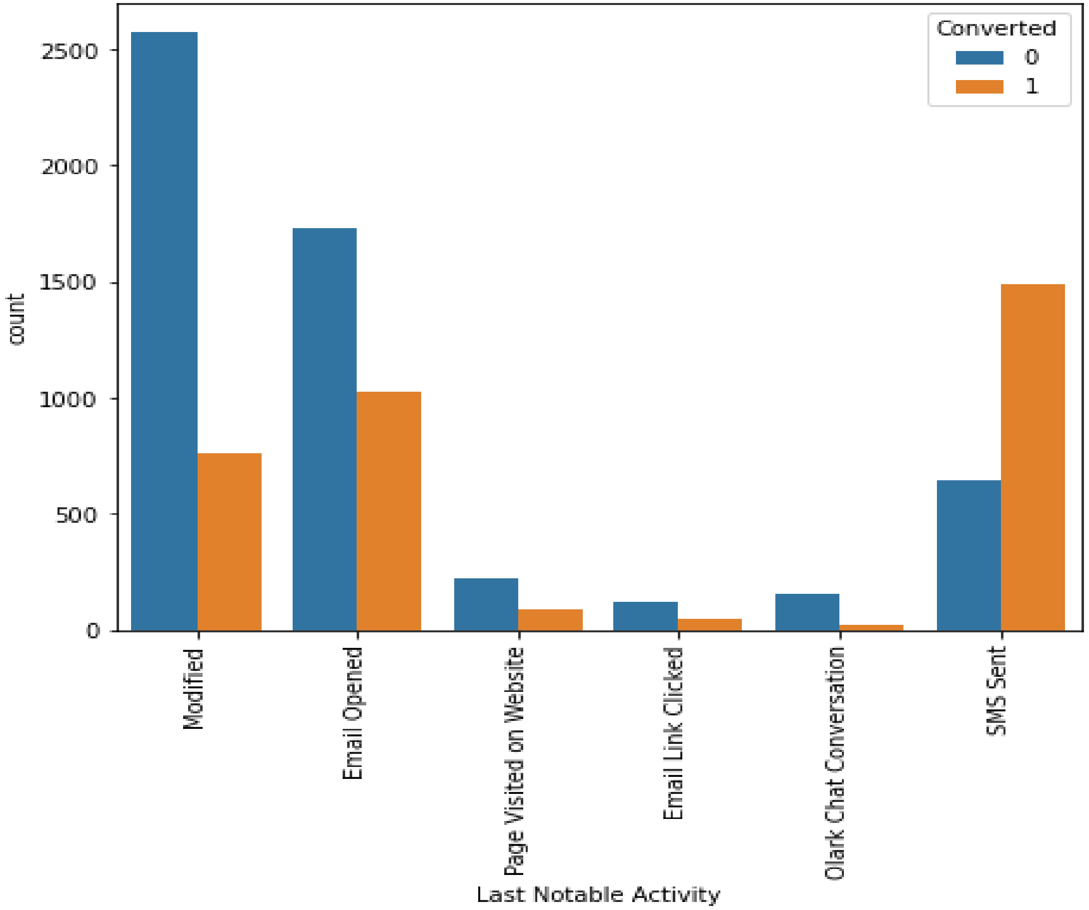


- Highest number of leads as well as conversion fall in 'Others' category. However, this category is the bucket of 'Not Specified' Values.
- Finance Management, Human Resource Management, Marketing Management, Operations Management are showing reasonably good results in terms of count of leads as well as conversion.

Analysis–

Univariate analysis for categorical variables with target variable

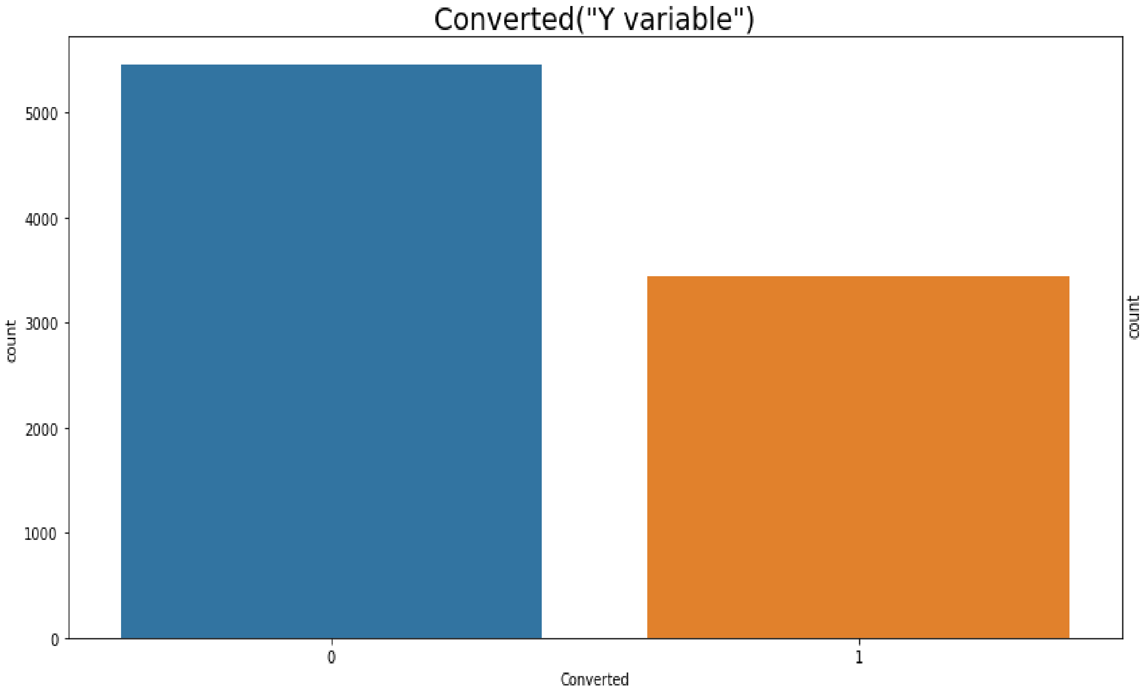


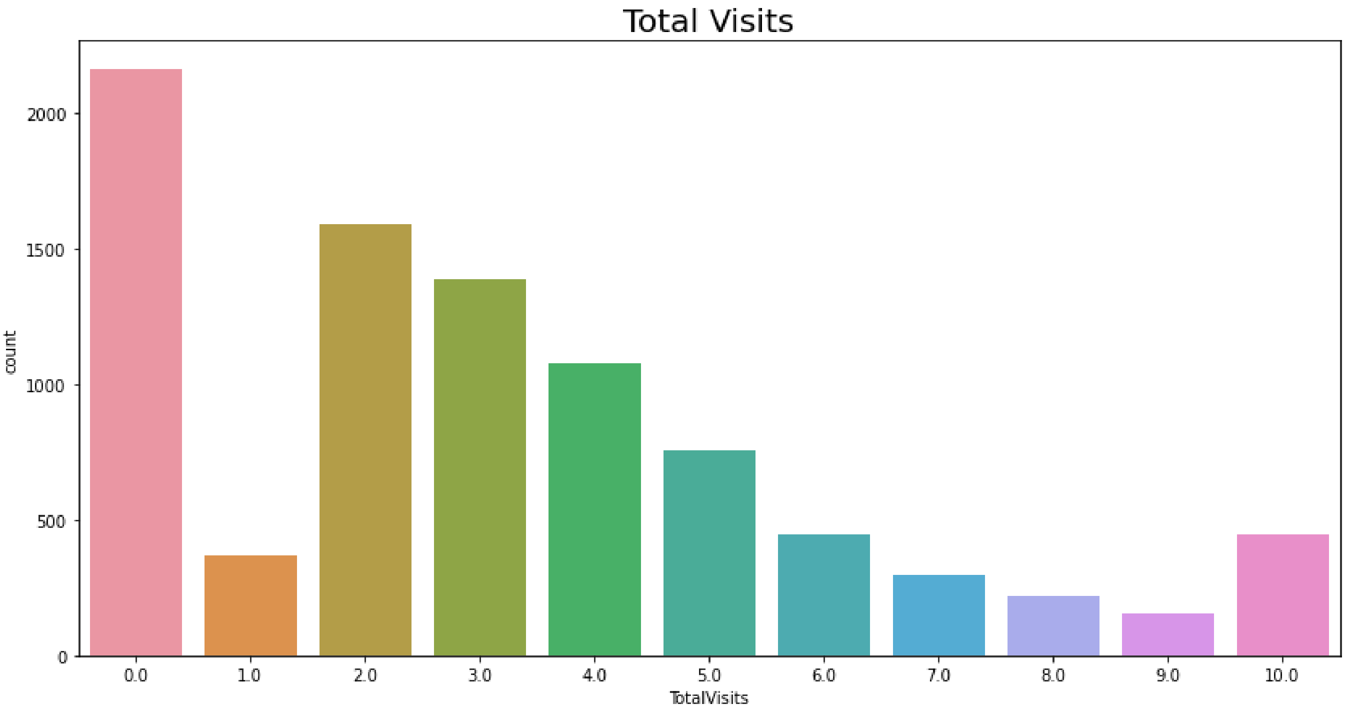


- Although the count is high for 'Email Opened', but the highest conversion rate from 'SMS Sent' Category.
- High Conversion rate is for 'Email Opened' and 'SMS Sent' Category.
- Lead count is highest for 'Modified' and 'Email Opened' category.

Analysis–

Univariate analysis for Numerical variables

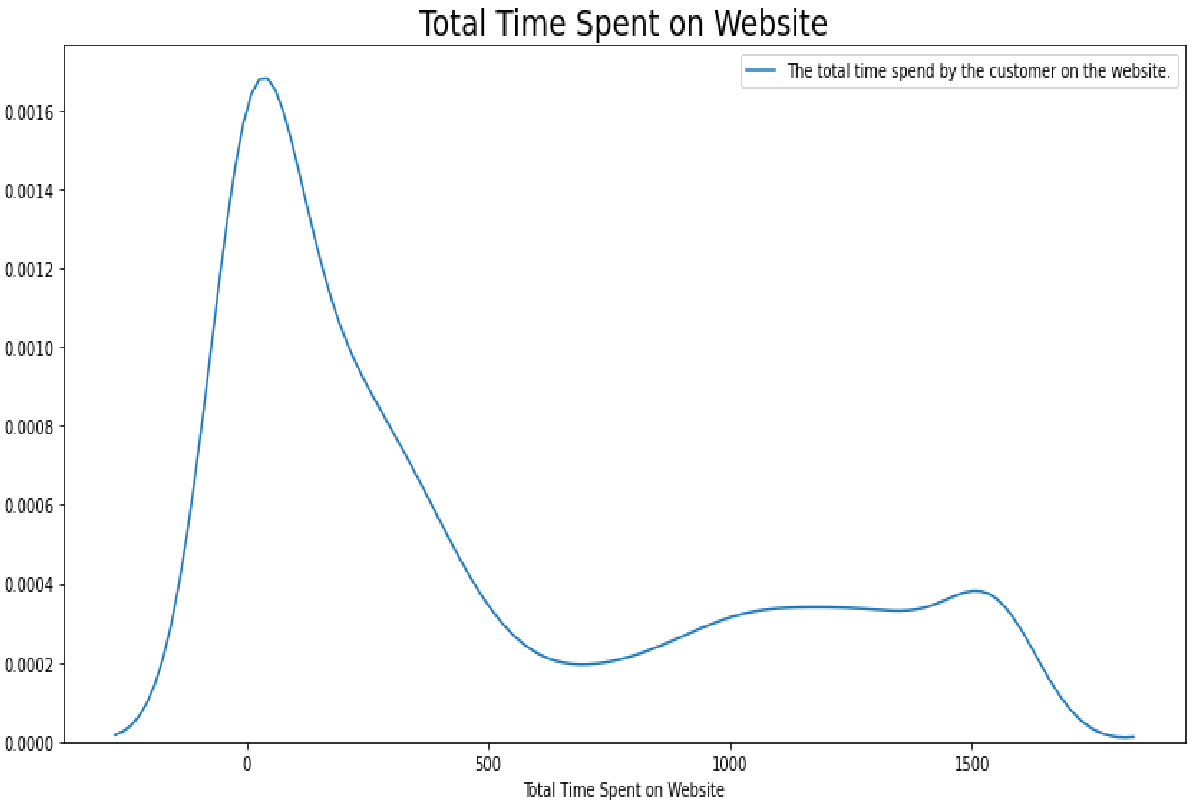




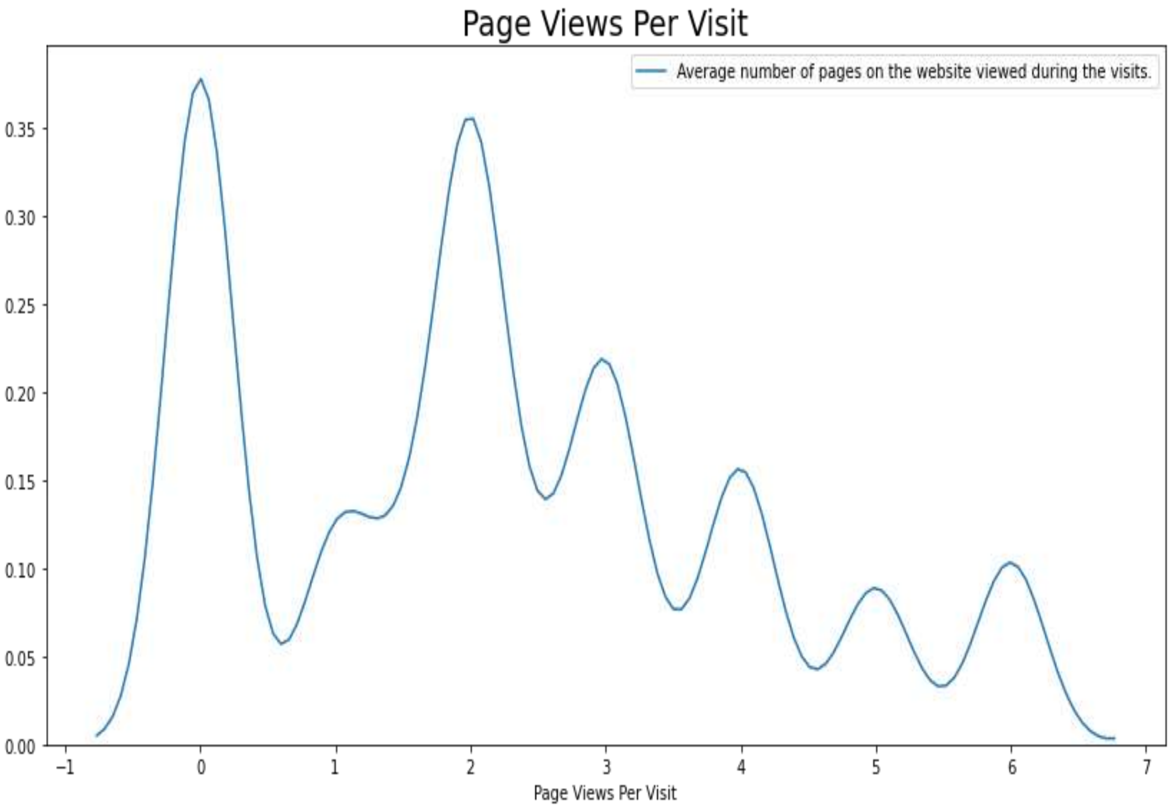
- Most of the customers are not converted.
- Total Visits are on an average ranging from 2 to 4. More customers are having 2 visits

Analysis–

Univariate analysis for numerical variables


Total Time Spent on Website

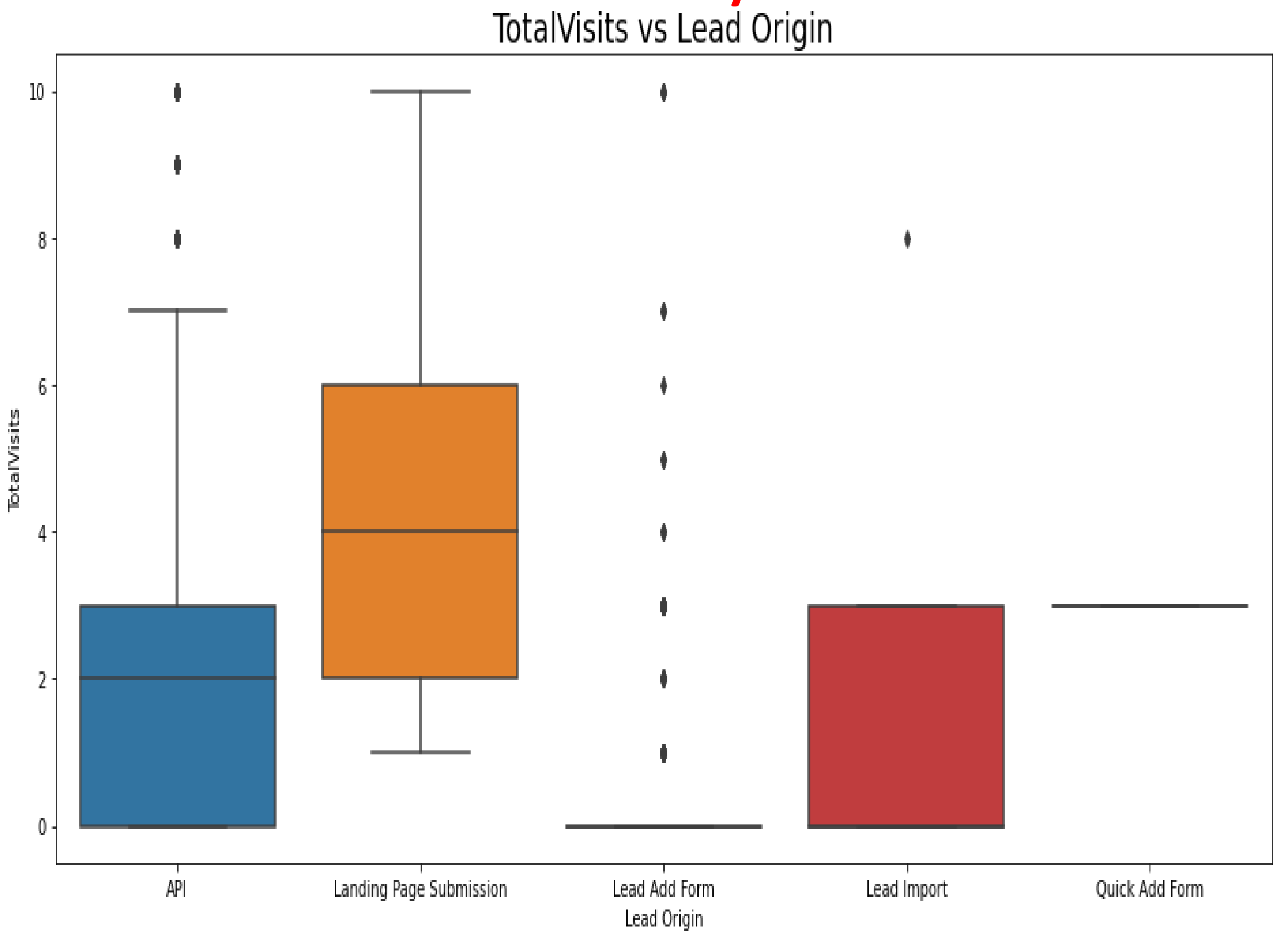
- Maximum number of customers are having a time spent between 0 to 500 seconds.
- There are very low customers who spends more time on website.

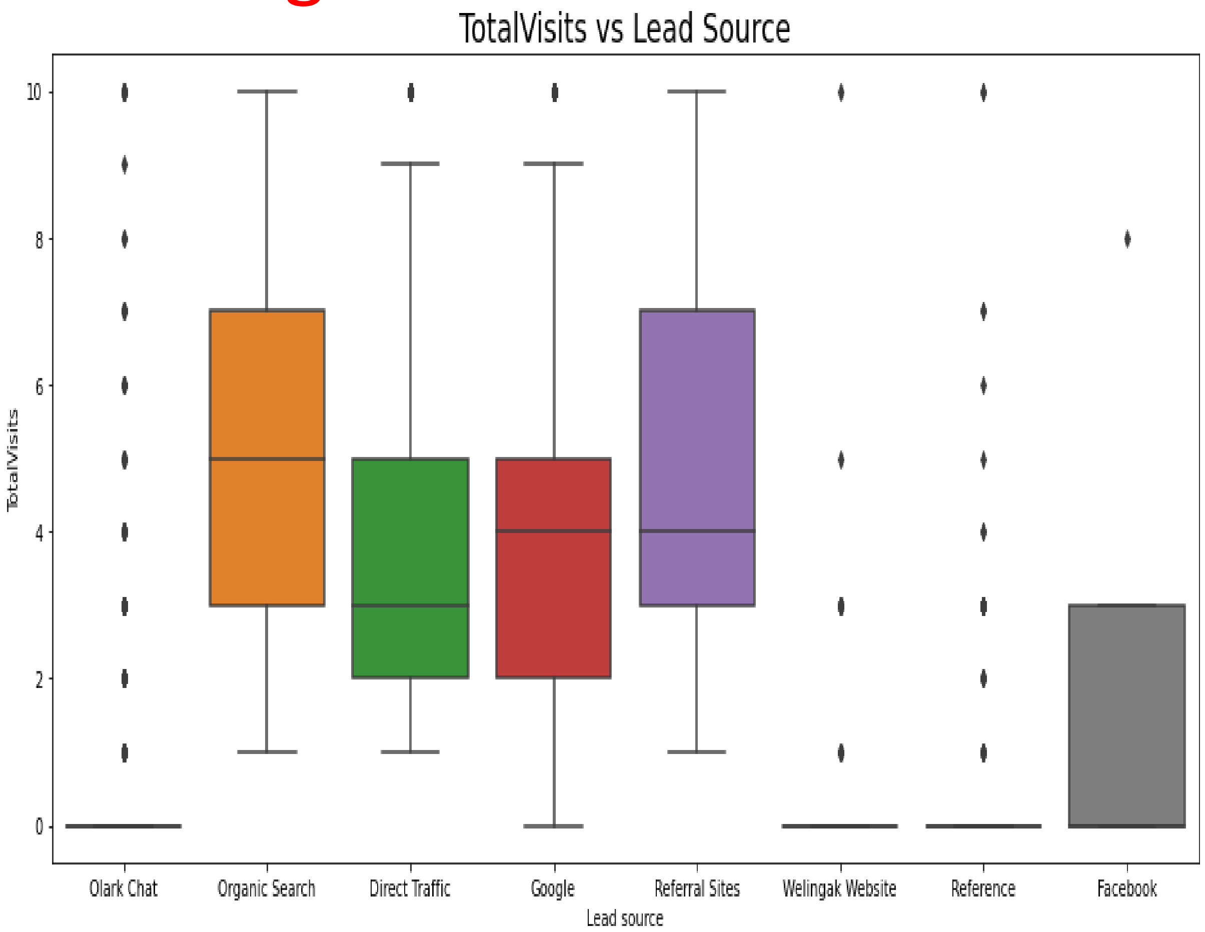

Page Views Per Visit

- Most customers are visited 4 pages and there are very less customers who visited more than 4 pages.
- Maximum number of page views is 2 to 3.

Analysis–

a. Bivariante Analysais for Continuos - Categorical variables.

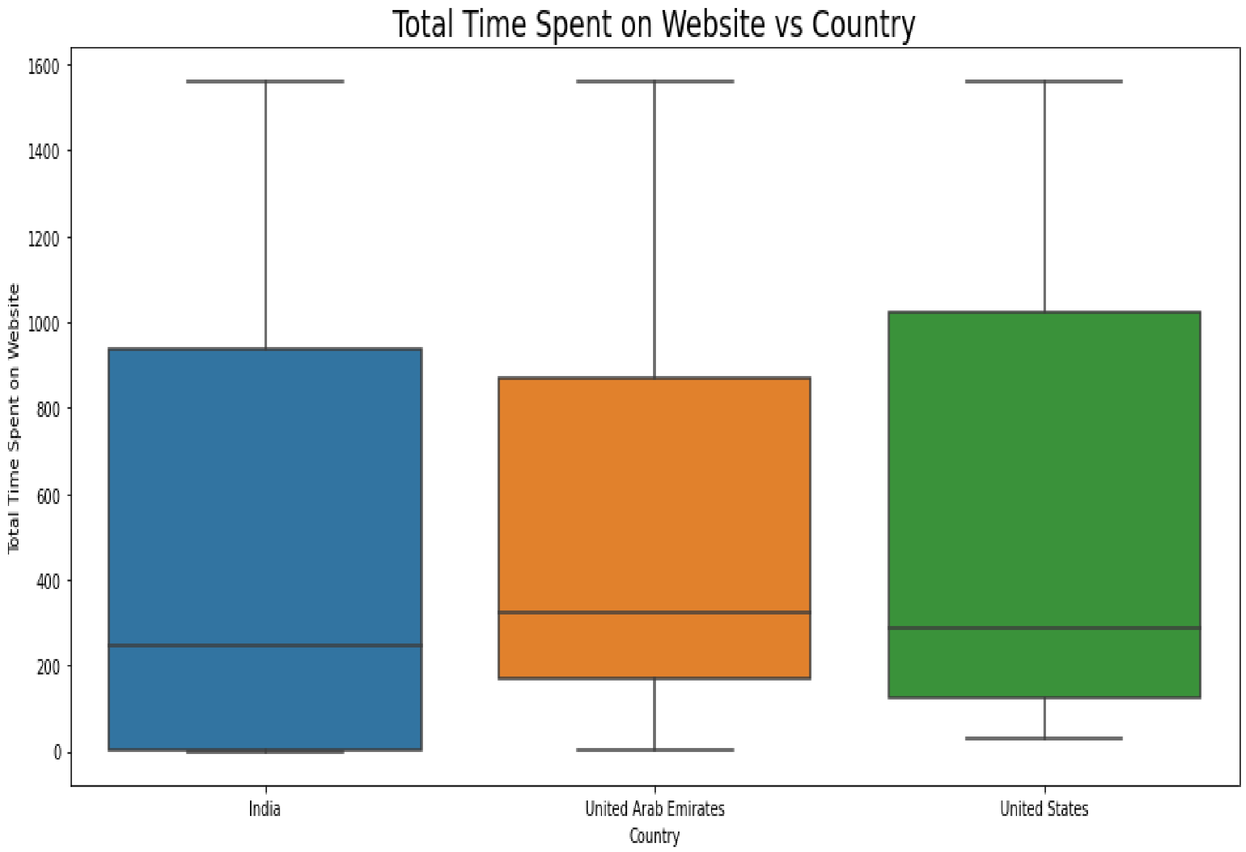


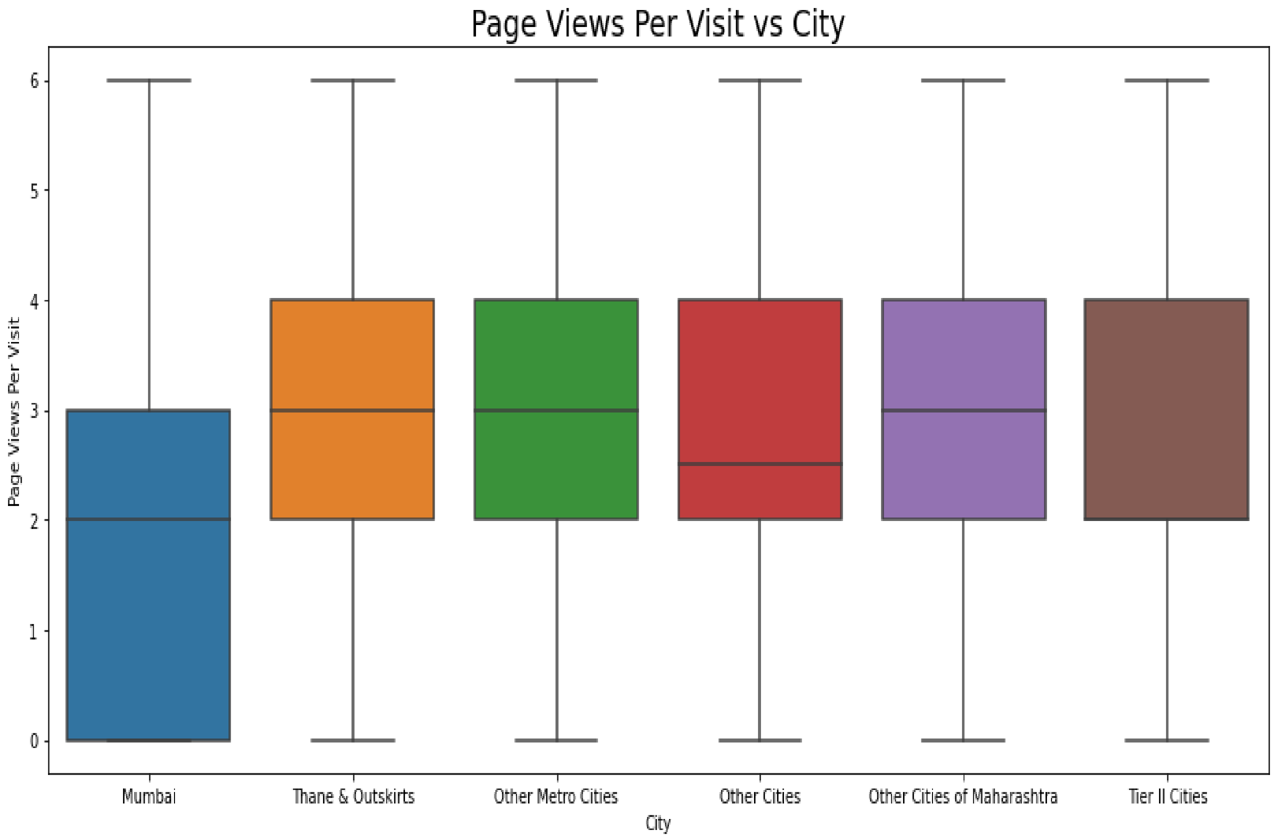


- The customers who have origin as Landing Page Submission have higher amount of total visits followed by API and Lead import.
- The customers who have source as Organic Search have higher amount of total visits followed by Referral Sites Lead Source, Direct Traffic and Google.

Analysis–

b. Bivariate Analysis for Continuous - Categorical variables..

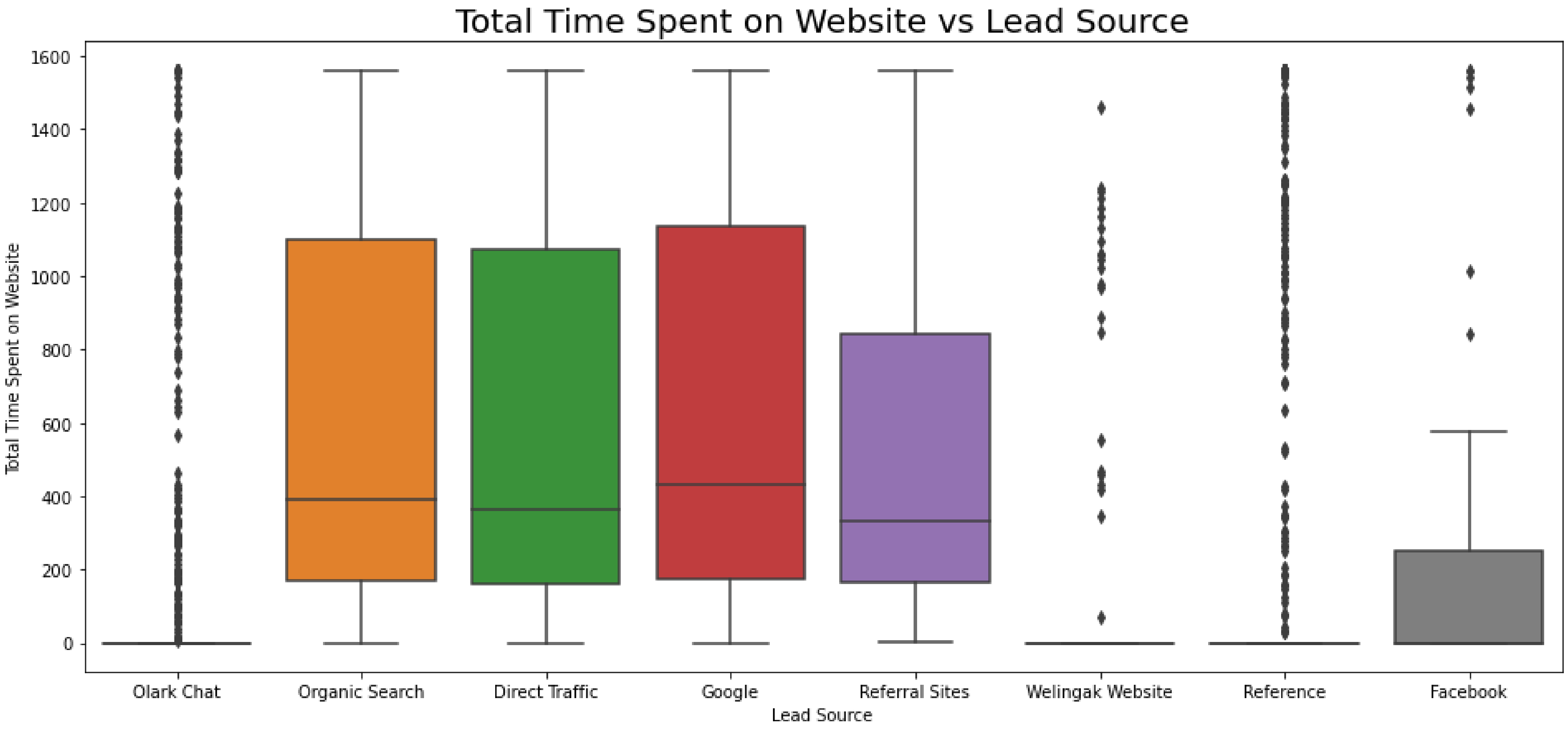




- Almost every country has same amount of time spent on website.
- Every city has almost same number of page views per visit except Mumbai has less number of Page Views Per Visit.

Analysis–

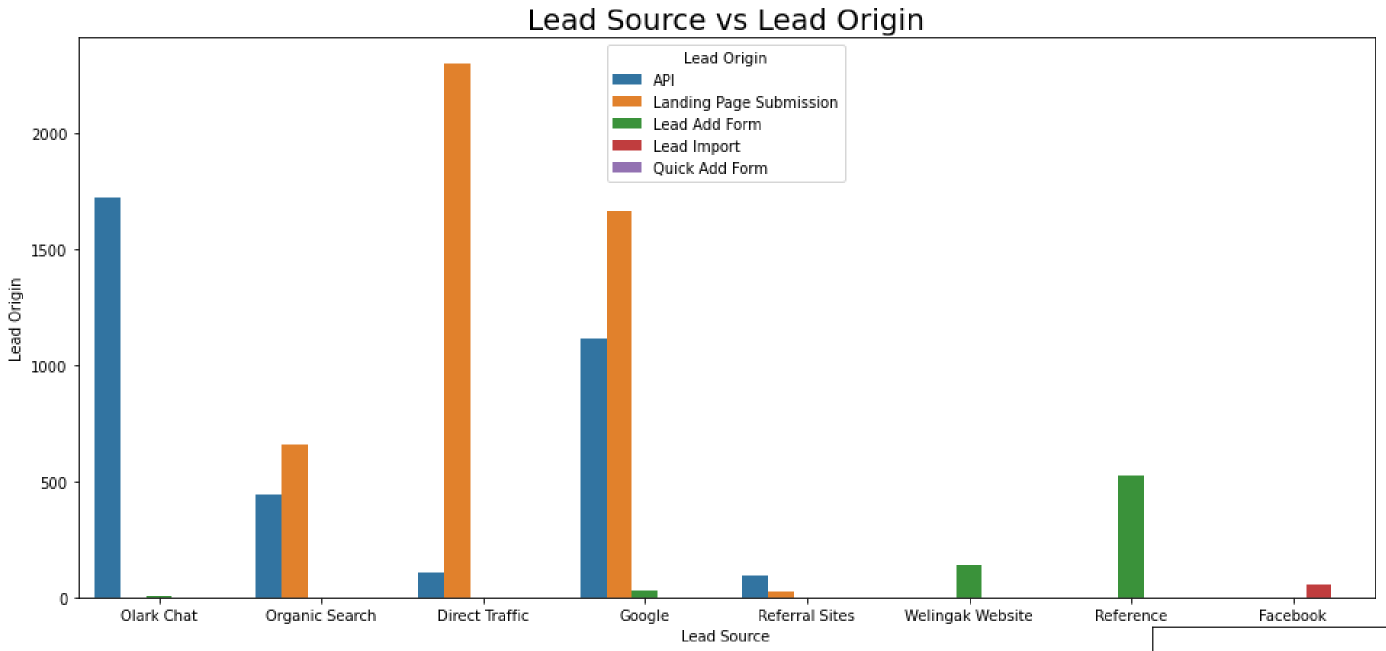
b. Bivariate Analysis for Continuous - Categorical variables..



- The customer who has lead source as "Google", "Direct Traffic", "Organic search" has higher time spent of website.

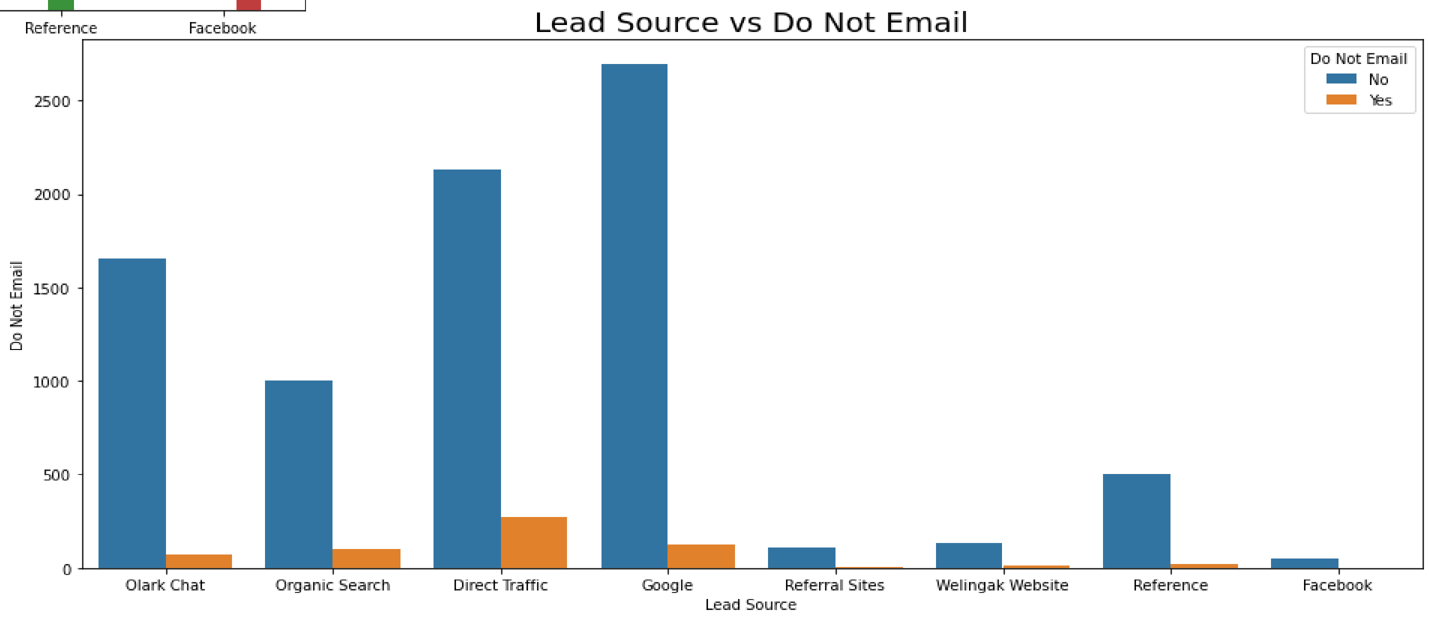
Analysis–

b. Bivariate Analysis for Continuous - Categorical variables..



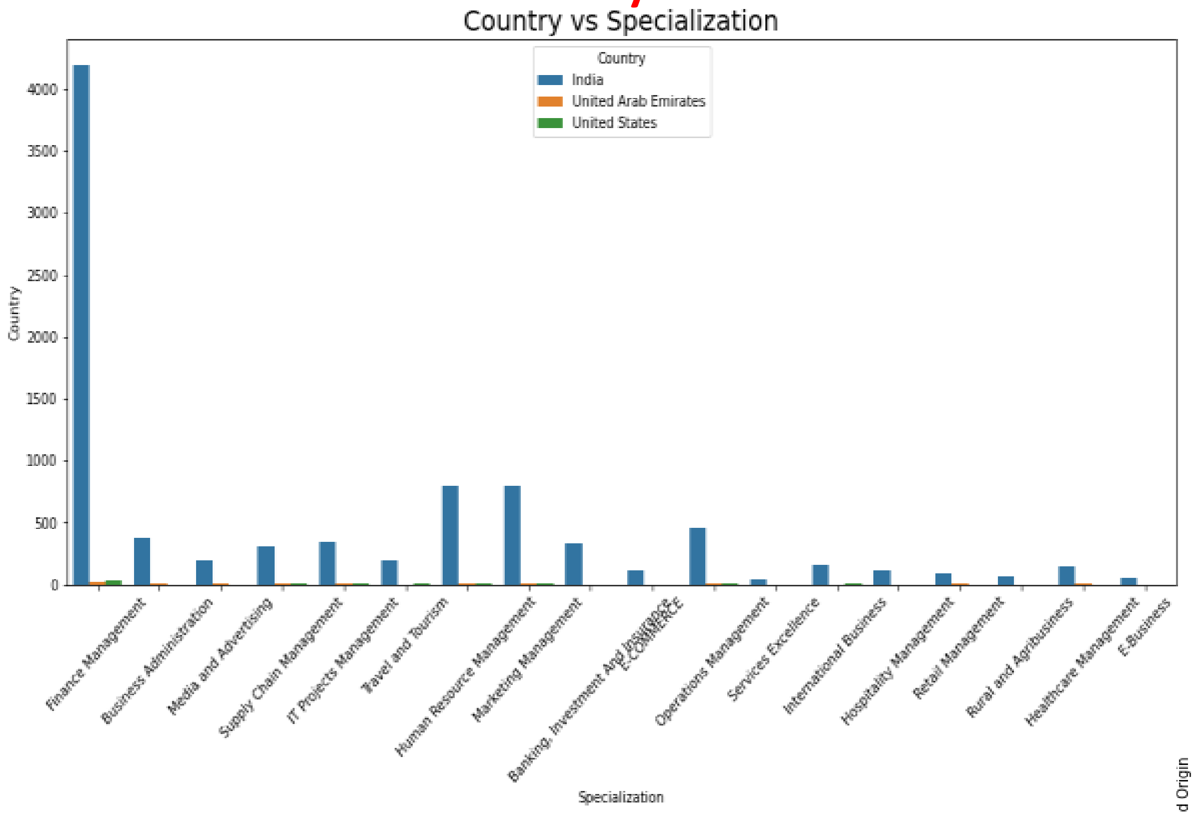
- The Olark Chat source has API as its origin most of times.
- The most customer which are from Direct Traffic source has origin as Landing Page Submission.
- The most customer which are from Google source has origin as Landing Page Submission.

- The customer who has specialization as Finance Management has Origin as API and Landing Page Submission.
- From every specialization most of the customers has origin as Landing Page submission.

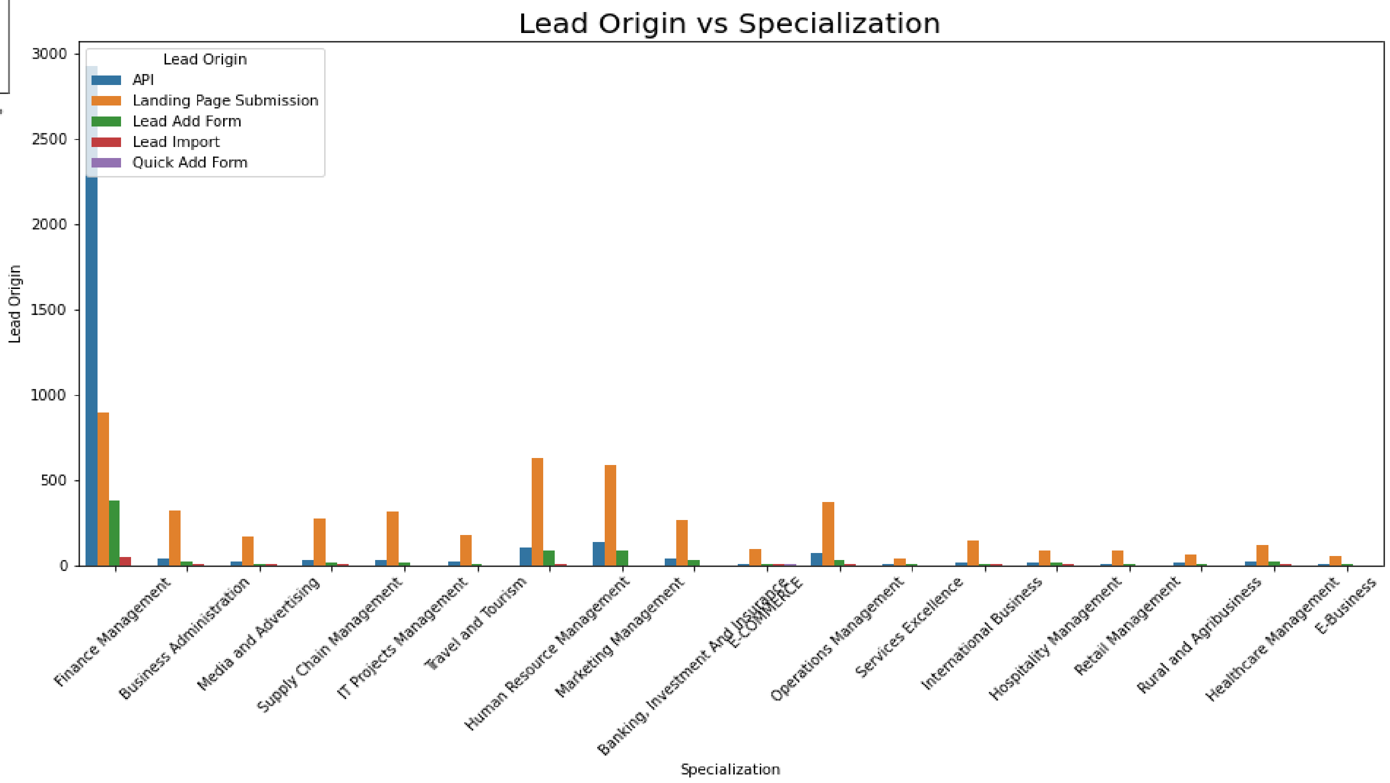


Analysis–

b. Bivariate Analysis for Continuous - Categorical variables..



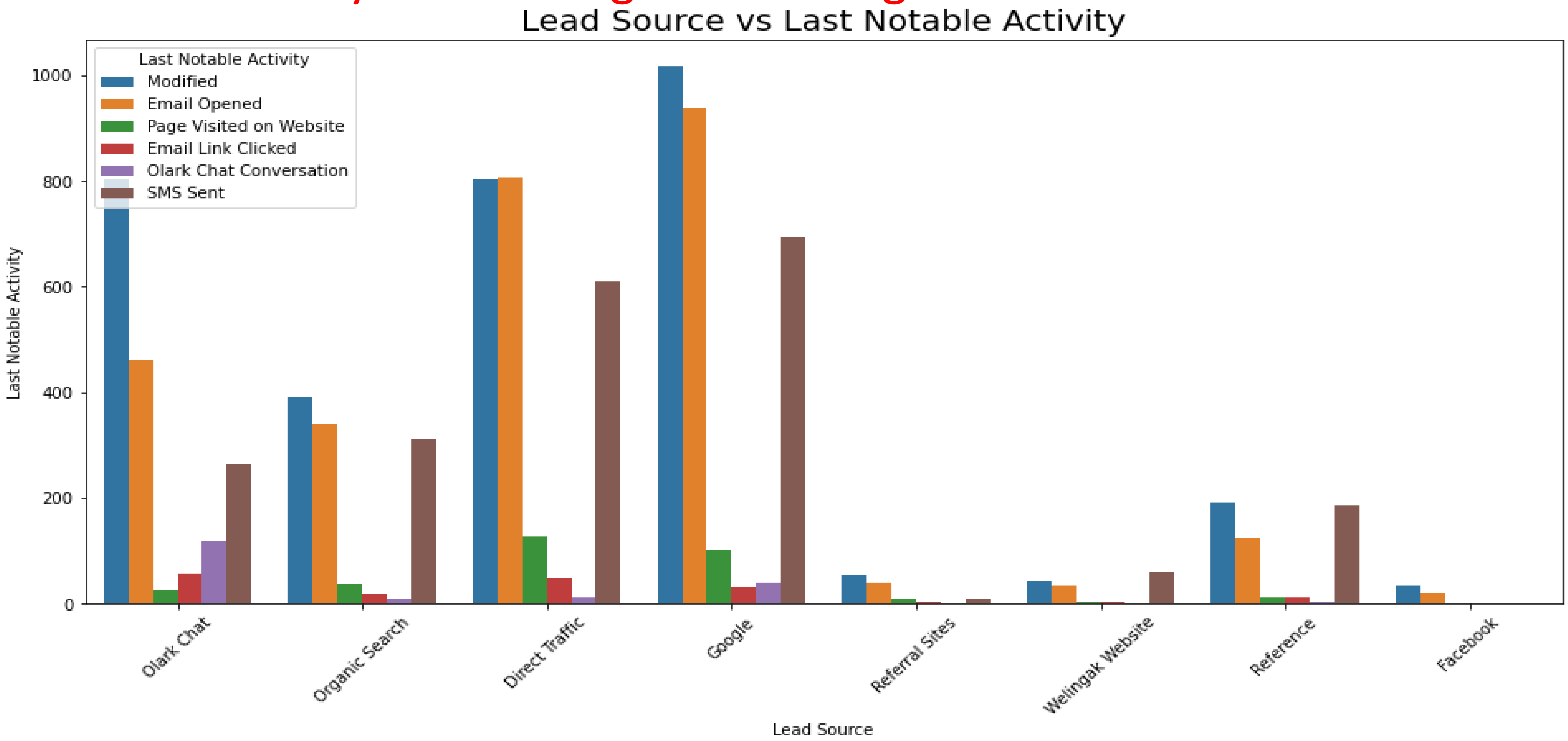
- •Customers which are from India has highest Specialization as Finance Management.



- The customer who has specialization as Finance Management has Origin as API and Landing Page Submission.
- •From every specialization most of the customers has origin as Landing Page submission.

Analysis–

c. Bivariate Analysis for Categorical- Categorical variables.



- Most of the customers who have Source as Google, Direct traffic, Olark Chat and Organic Search have Last notable activity as Modified and Email Opened.

Data Preparation

- In this step we converted the binary column into zero's and one's. All the categorical columns are replaced by their respective dummies.

- After the dummification we divided the data in the ratio of 0.7 : 0.3.

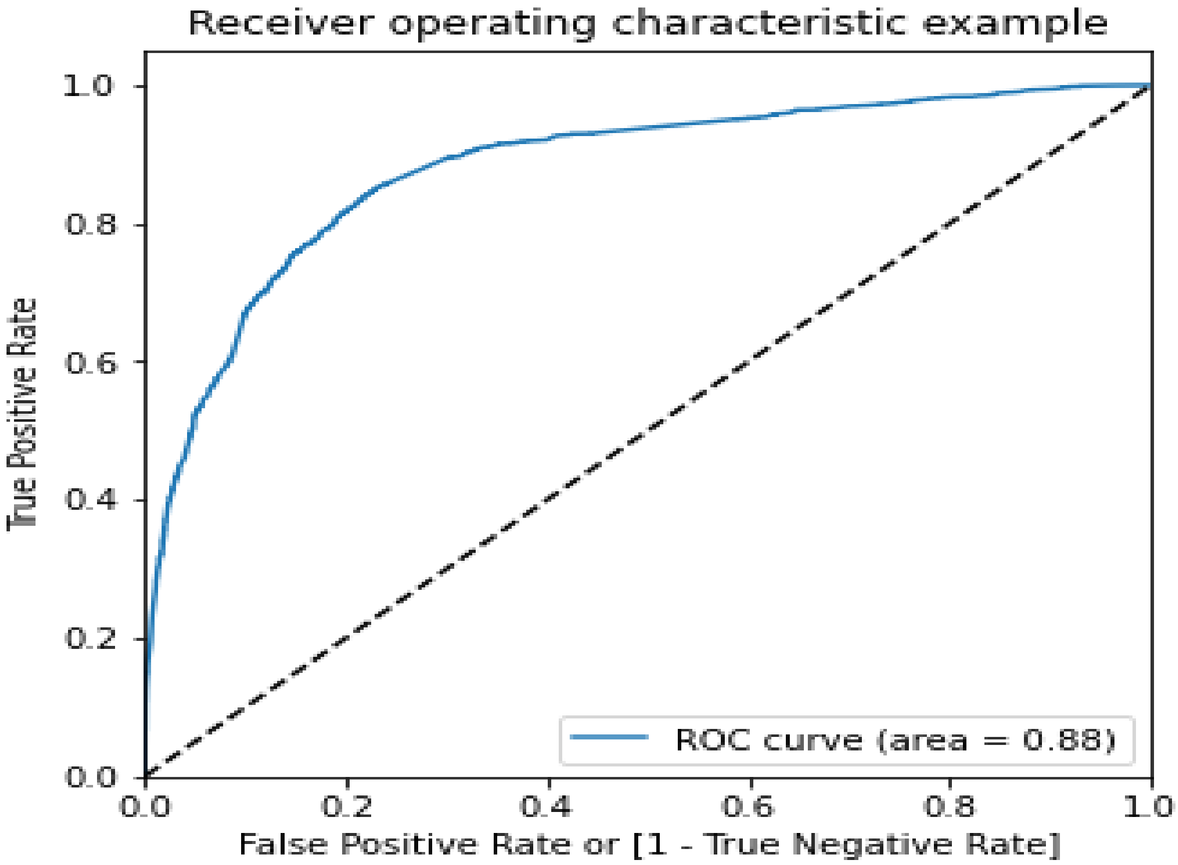
- Feature Scaling : In this step all the numeric columns are scaled using standard scaler function. The correlation metrics is also checked.

Modeling

- RFE is used to perform variable selection effectively and to eliminate the insignificant columns
- .
- The statsmodel library is used to build the logistic regression model.

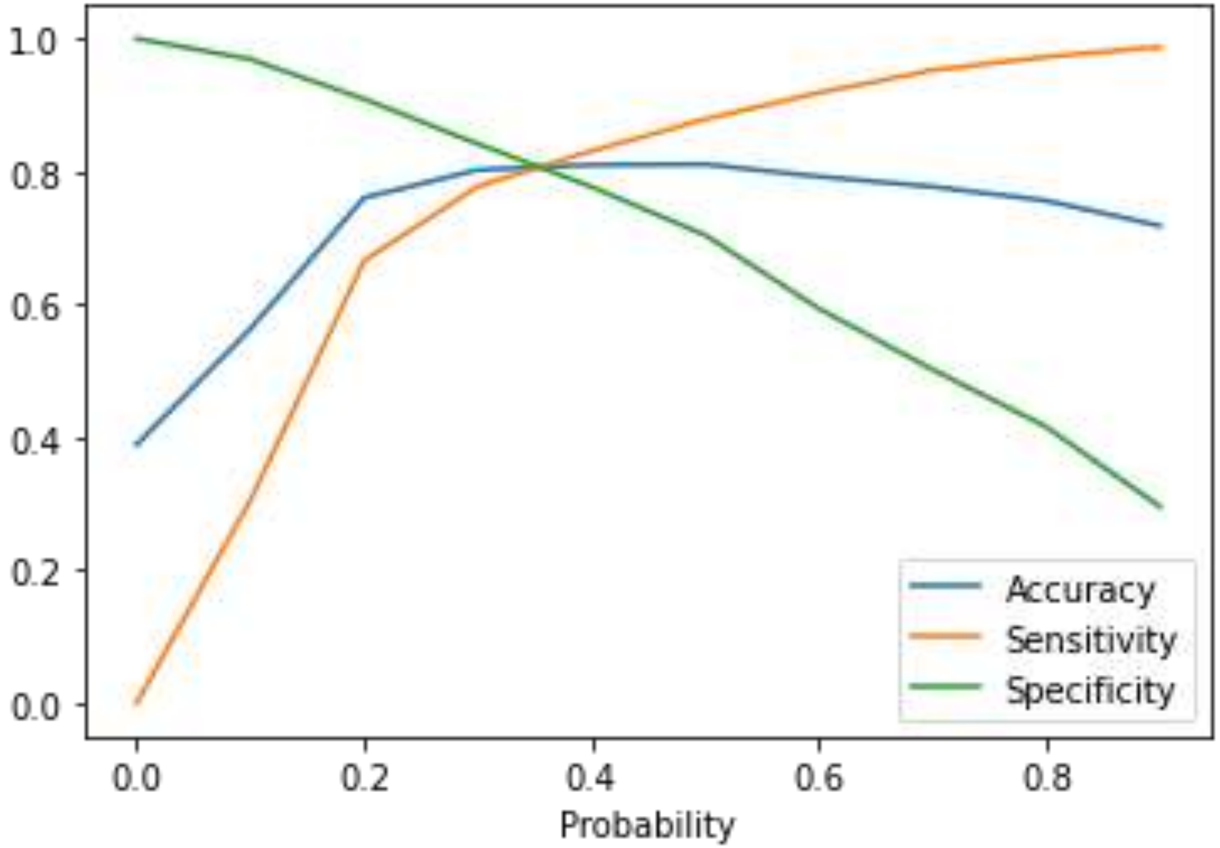
- The P-value and VIF is checked and the insignificant columns are dropped.

ROC curve



- We can observe that the area under the curve is 88% and the ROC curve is more towards the upper-left corner of the graph, it means that the model is very good.

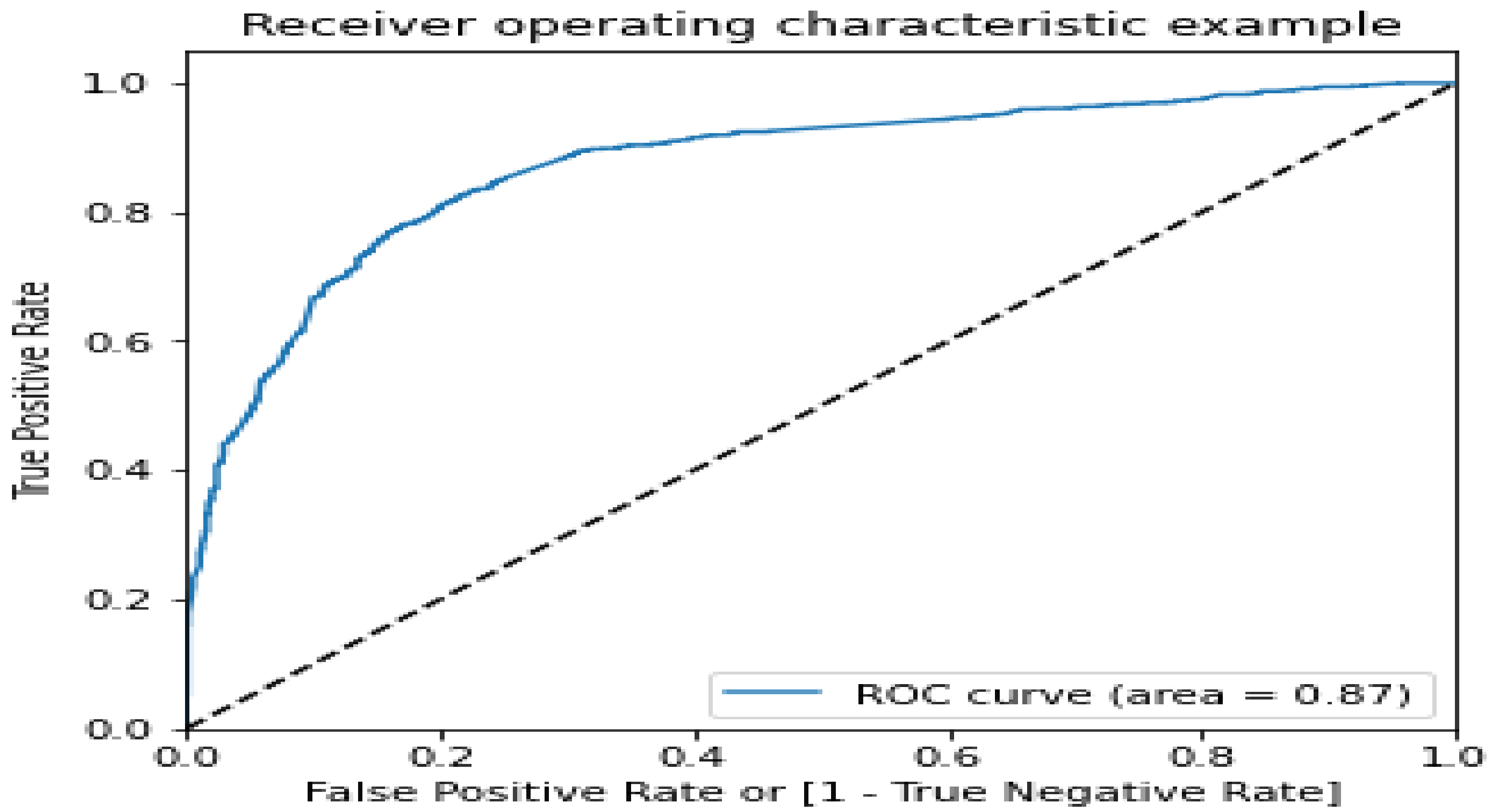
Optimal probability cutoff



- From the above curve we can observe that the 0.36 is the optimal point to take it as a cutoff probability.

Making predictions on the test data

ROC curve of the test data



The area under the ROC curve is 0.88, which is a very good value.

Summary

Final words

Based on analysis, defining the results and conclusion.

Training Set :

Accuracy	80.76 %
Sensitivity	80.43 %
Specificity	80.97 %
Precision	78.65 %
Recall	70.41 %
F1-Score	76.47 %
False Positive Rate	19.02 %
Positive Predictive Value	72.89 %
Negative Predictive Value	86.67 %

Test Set :

Accuracy	80.34%
Sensitivity	81.04%
Specificity	79.91%
Precision	71.36%
Recall	81.04%
F1-Score	75.89%
False Positive Rate	20.08%
Positive Predictive Value	71.36%
Negative Predictive Value	87.21%

Summary

Final words

Based on analysis, defining the results and conclusion.

The features which are most mattered in lead conversion are : (Arranging from most important to less important by comparing the coefficient.)

1. **Lead Origin_Lead Add Form**
2. **Lead Source_Welingak Website**
3. **What is your current occupation_Working Professional**
4. **Last Notable Activity_SMS Sent**
5. **Do Not Email**
6. **Total Time Spent on Website**
7. **Last Notable Activity_Olark Chat Conversation**
8. **Lead Source_Olark Chat**
9. **Last Notable Activity_Email Opened**
10. **Lead Origin_Landing Page Submission**