LEAD SCORING CASE STUDY

By:
DSC-40

Kunal Kashyap
Vedant Tele
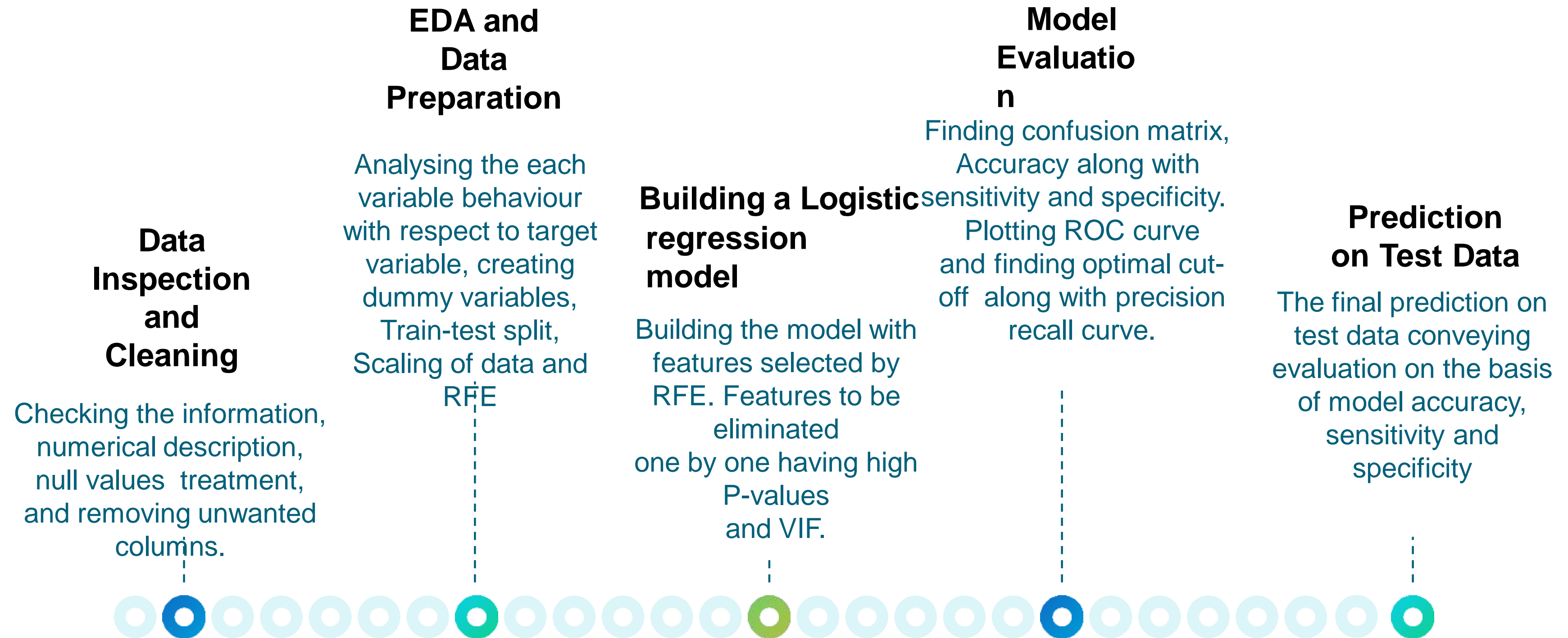
Problem Statement

- X education sells online courses to Industry Professionals.
- X education gets a lots of leads, its lead conversion rate it very poor. For example, if they acquire 100 leads in a day , only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as "Hot Leads".
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.
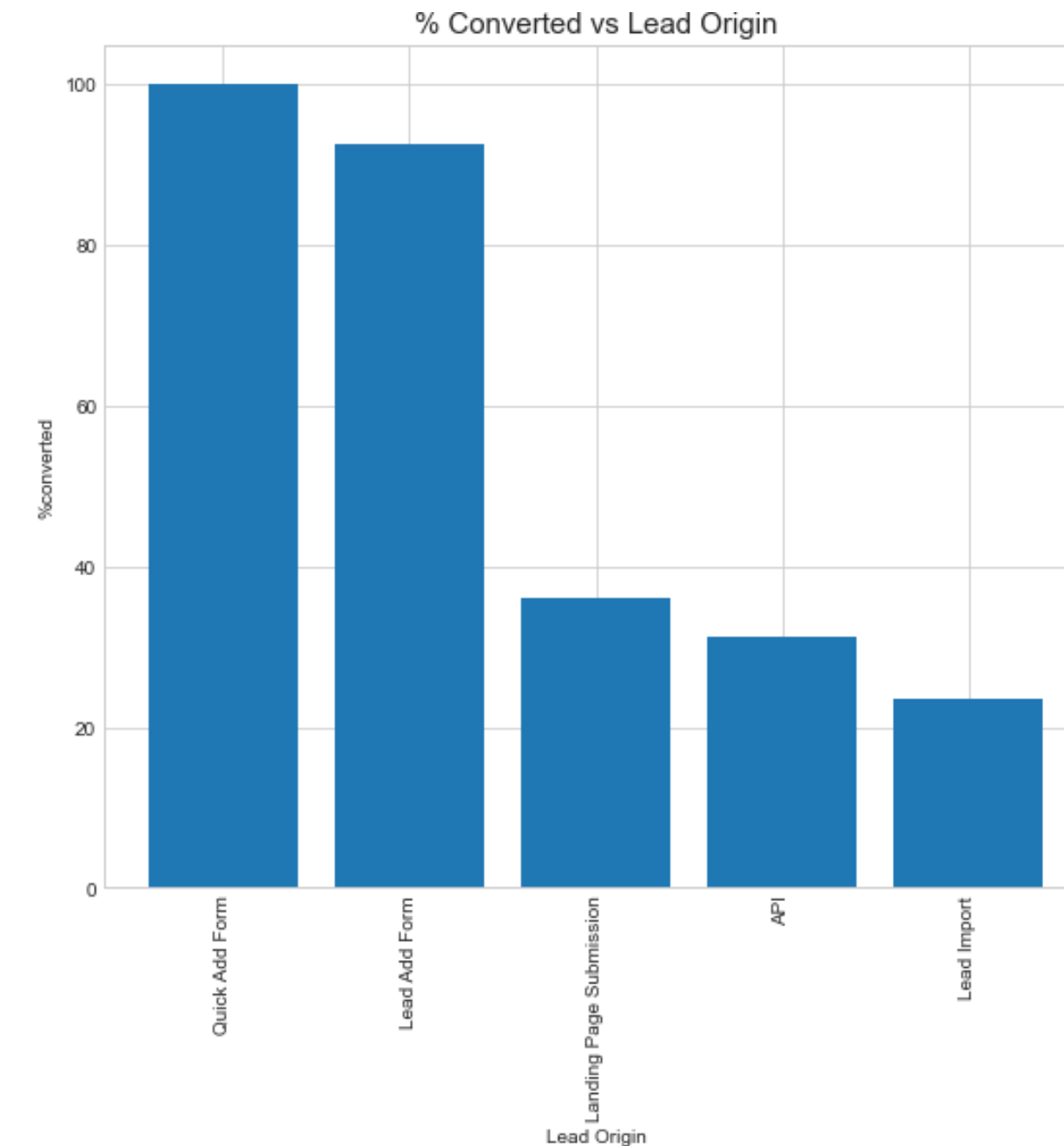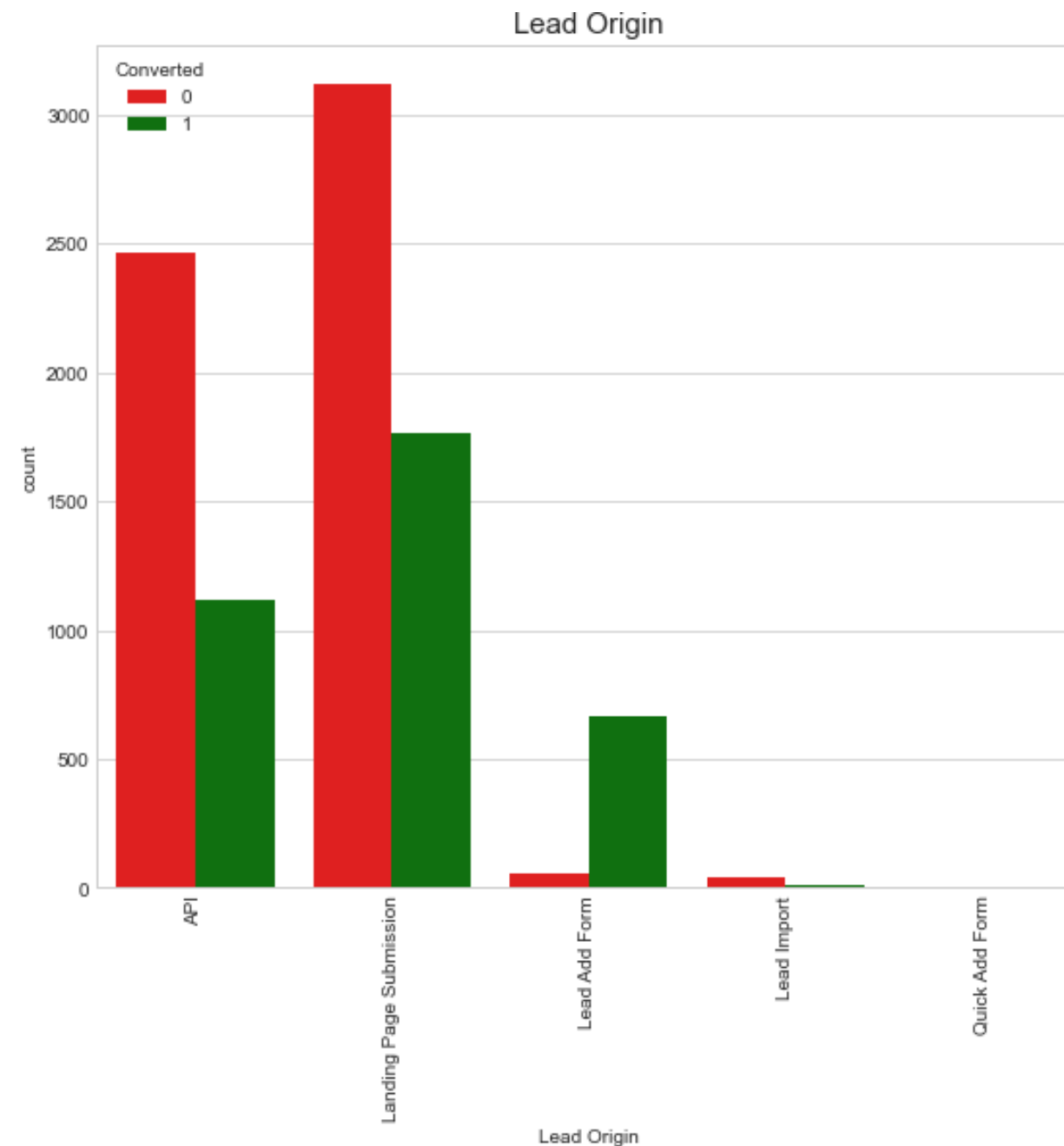
Business Objective

- Help X education to select the Most Promising Leads (Hot Leads)
- Build a Logistic regression model to assign a lead score value between 0 to 100 to each of the leads which can be used by the company to target Potential Leads

Analysis Approach

EDA and Data Preparation

Analysing the each variable behaviour with respect to target variable, creating dummy variables, Train-test split, Scaling of data and RFE

Model Evaluation

Finding confusion matrix, Accuracy along with sensitivity and specificity. Plotting ROC curve and finding optimal cut-off along with precision recall curve.

Data Inspection and Cleaning

Checking the information, numerical description, null values treatment, and removing unwanted columns.

Building a Logistic regression model

Building the model with features selected by RFE. Features to be eliminated one by one having high P-values and VIF.

Prediction on Test Data

The final prediction on test data conveying evaluation on the basis of model accuracy, sensitivity and specificity

Analysis - Lead Origin

- **Observation:**

The majority of the leads came from submissions on the landing page, followed by API, where approximately 30% are converted.Leads from the Quick Add Form are 100% Converted; there was only one lead from that category.Leads from the Lead Add Form have the highest conversion rate in this category, accounting for approximately 90% of the 718 leads.Lead imports are few in number, and the conversion rate is also low.To increase overall lead conversion rates, we must concentrate on improving lead conversion from API and Landing Page Submission origins and generating more leads from Lead Add Forms.

Analysis - Lead Source
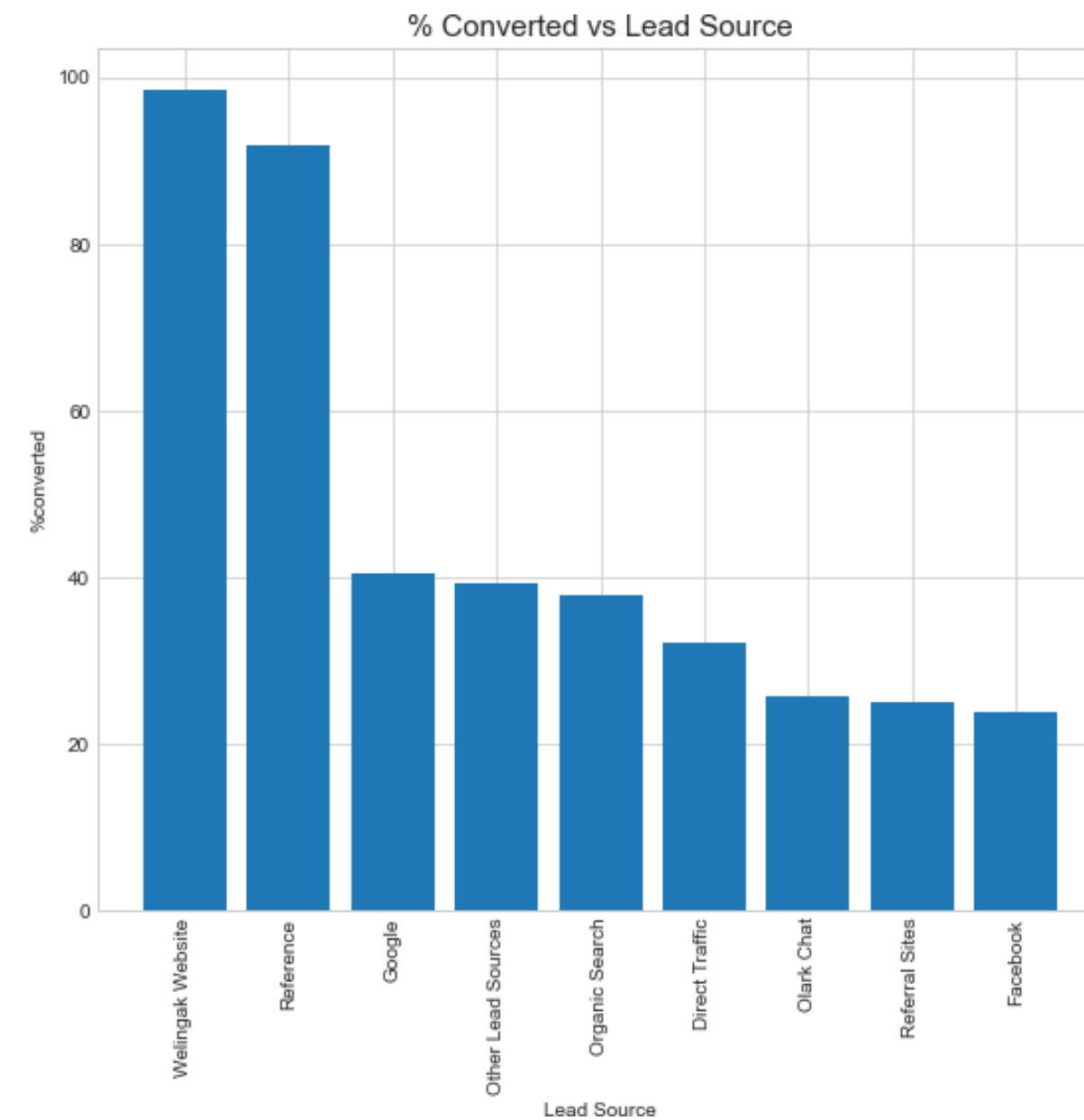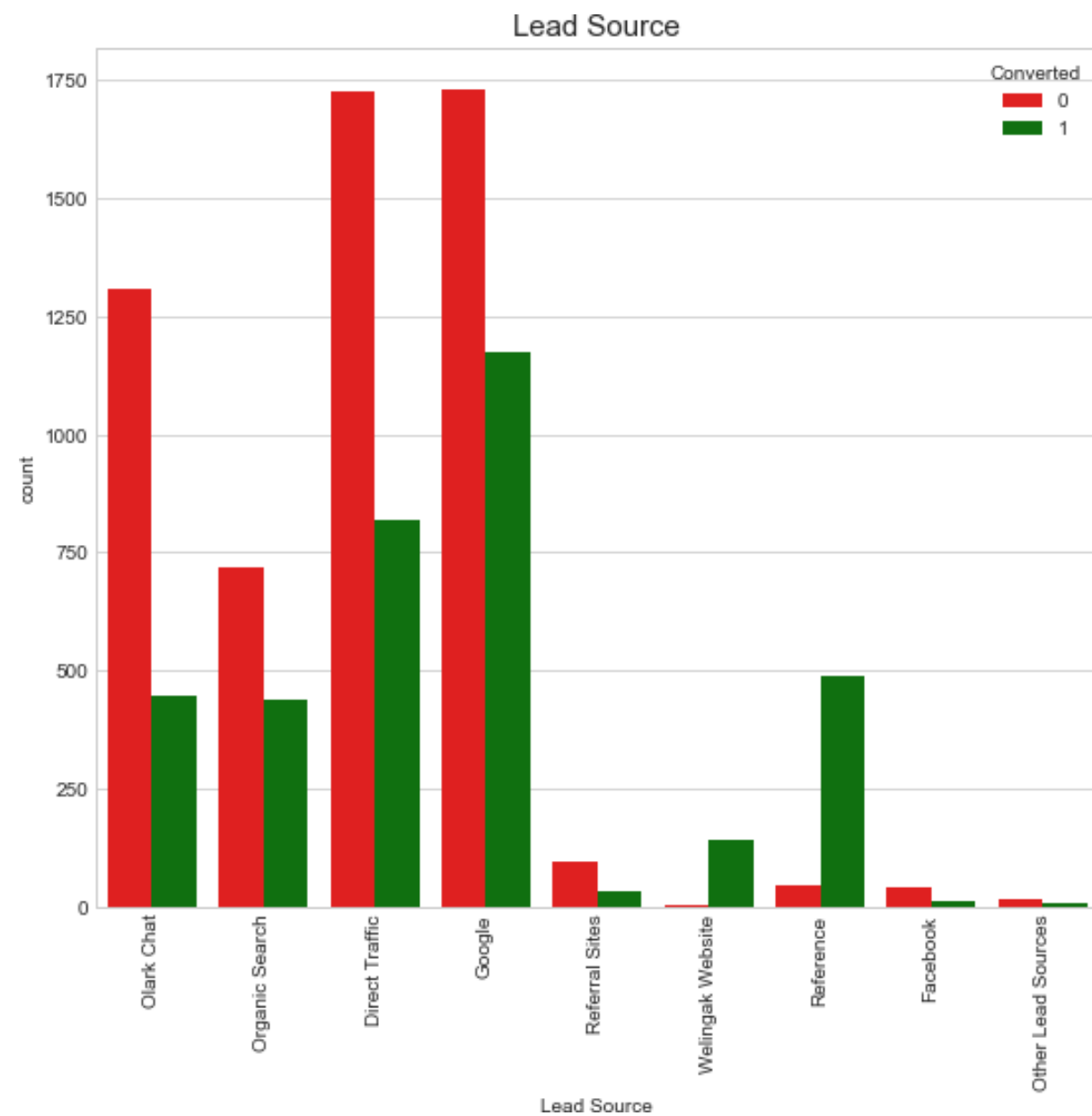
Observation:

Google generated the most leads, accounting for 40% of all conversions, followed by Direct Traffic, Organic Search, and Olark Chat. From a total of 534 leads, a lead generated by a referral has a conversion rate of more than 90%.Welingak's website has a nearly perfect lead conversion rate. This option should be investigated further in order to improve lead conversion. To increase lead count, initiatives should be implemented so that existing members increase their referrals.
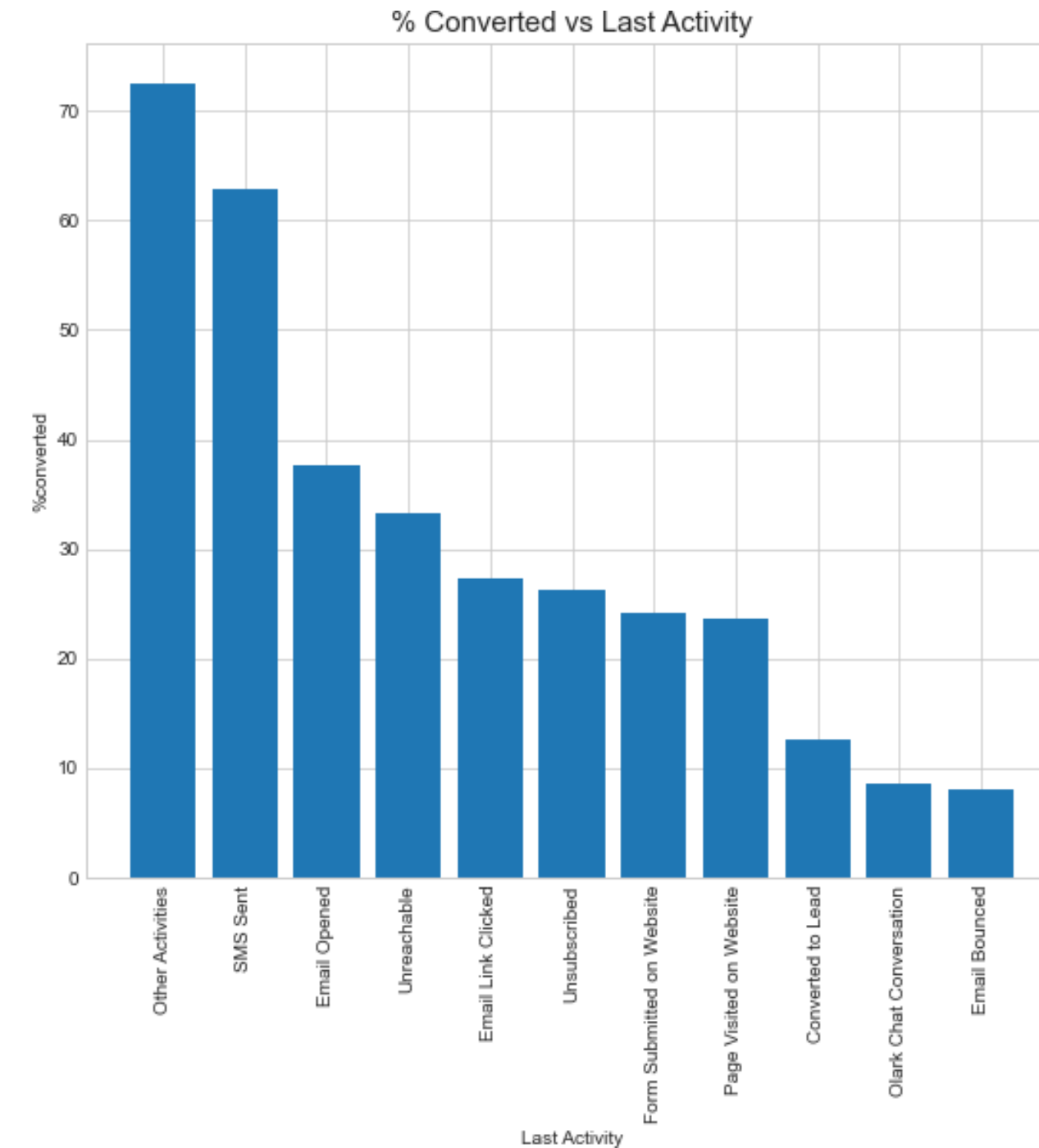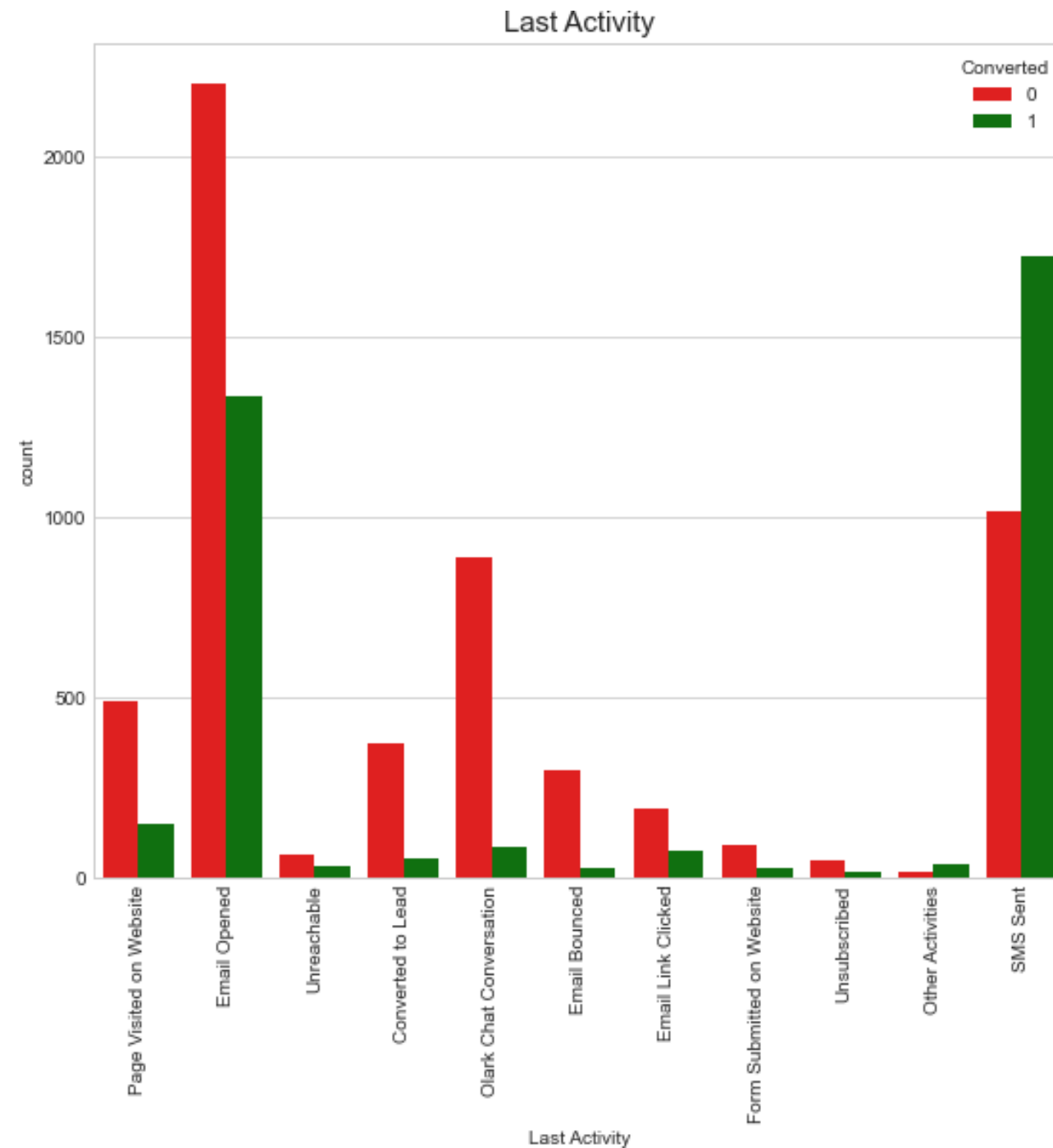
Analysis - Do Not Email
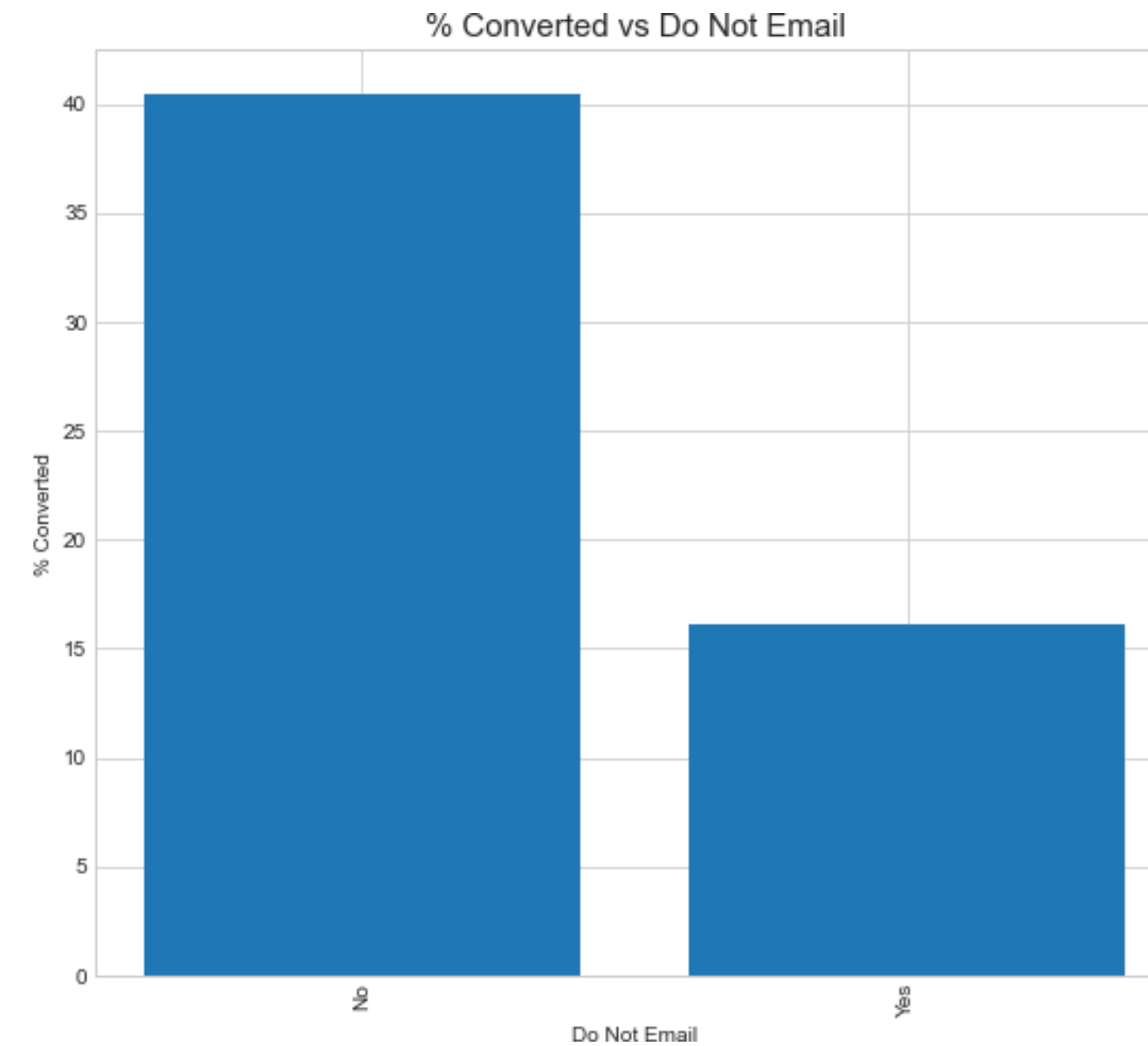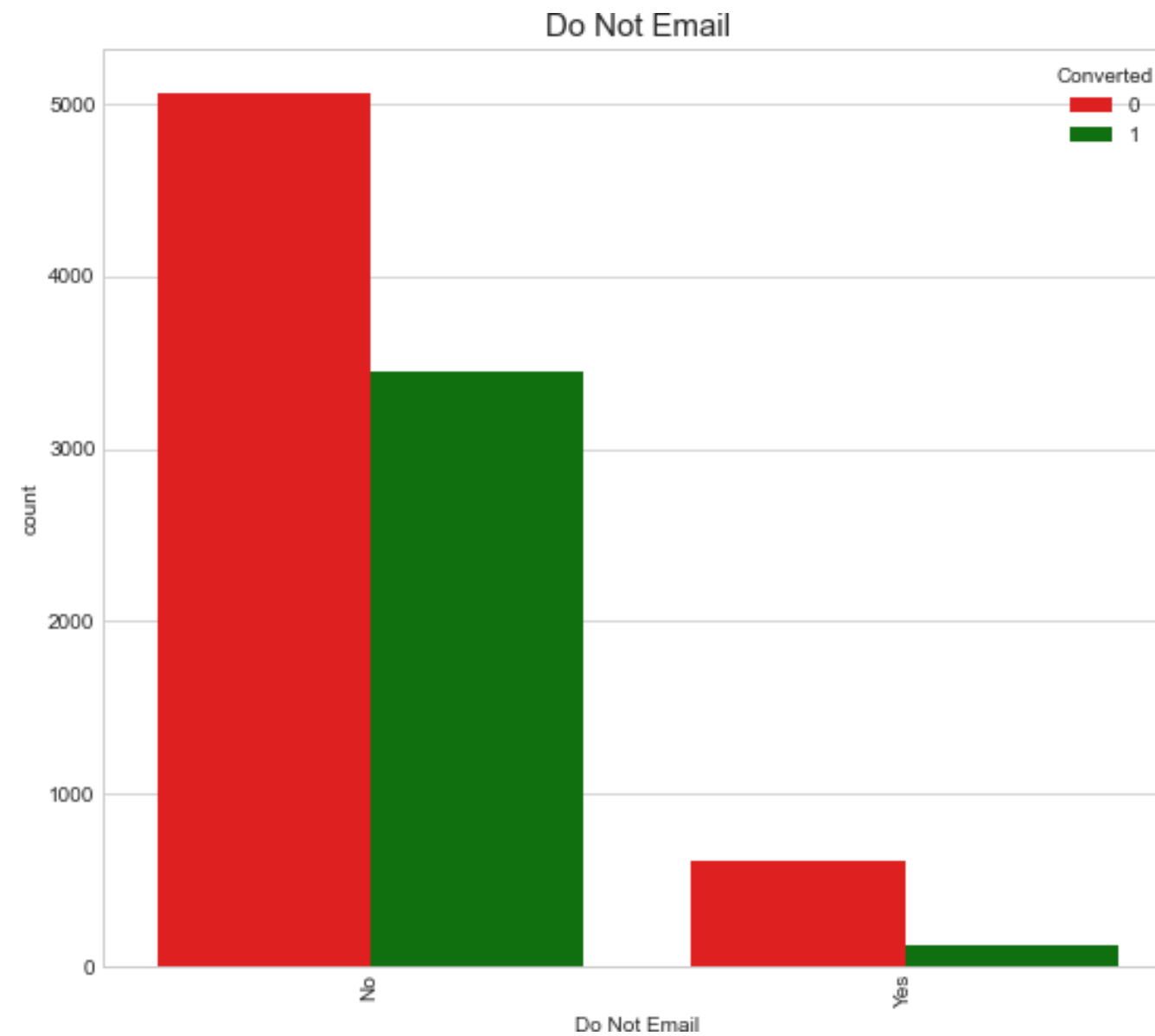
Observation:

- **Majority of the people want Email (~92%)**
- **People who have opted to receive Email has higher rate of conversion (40%)**
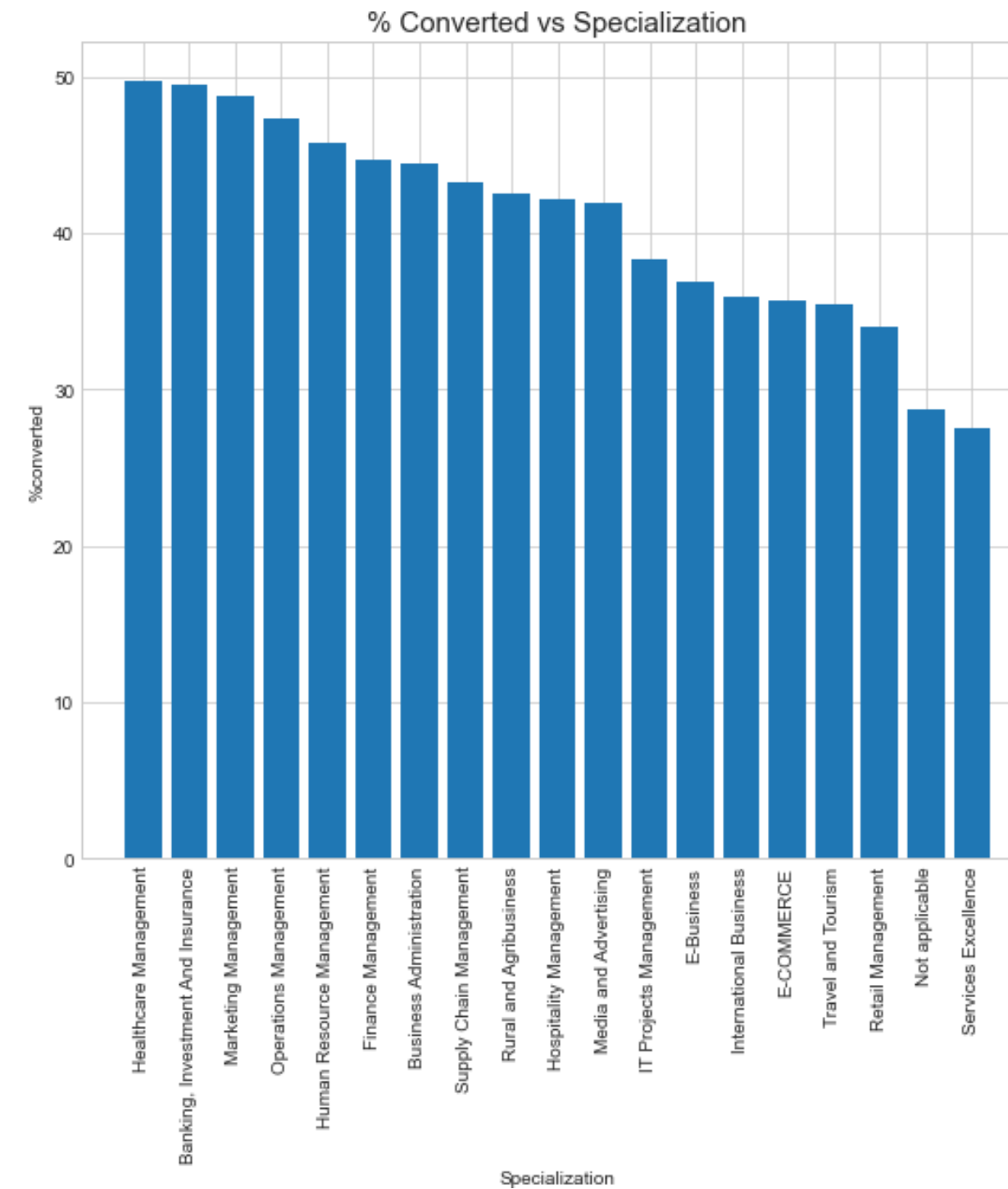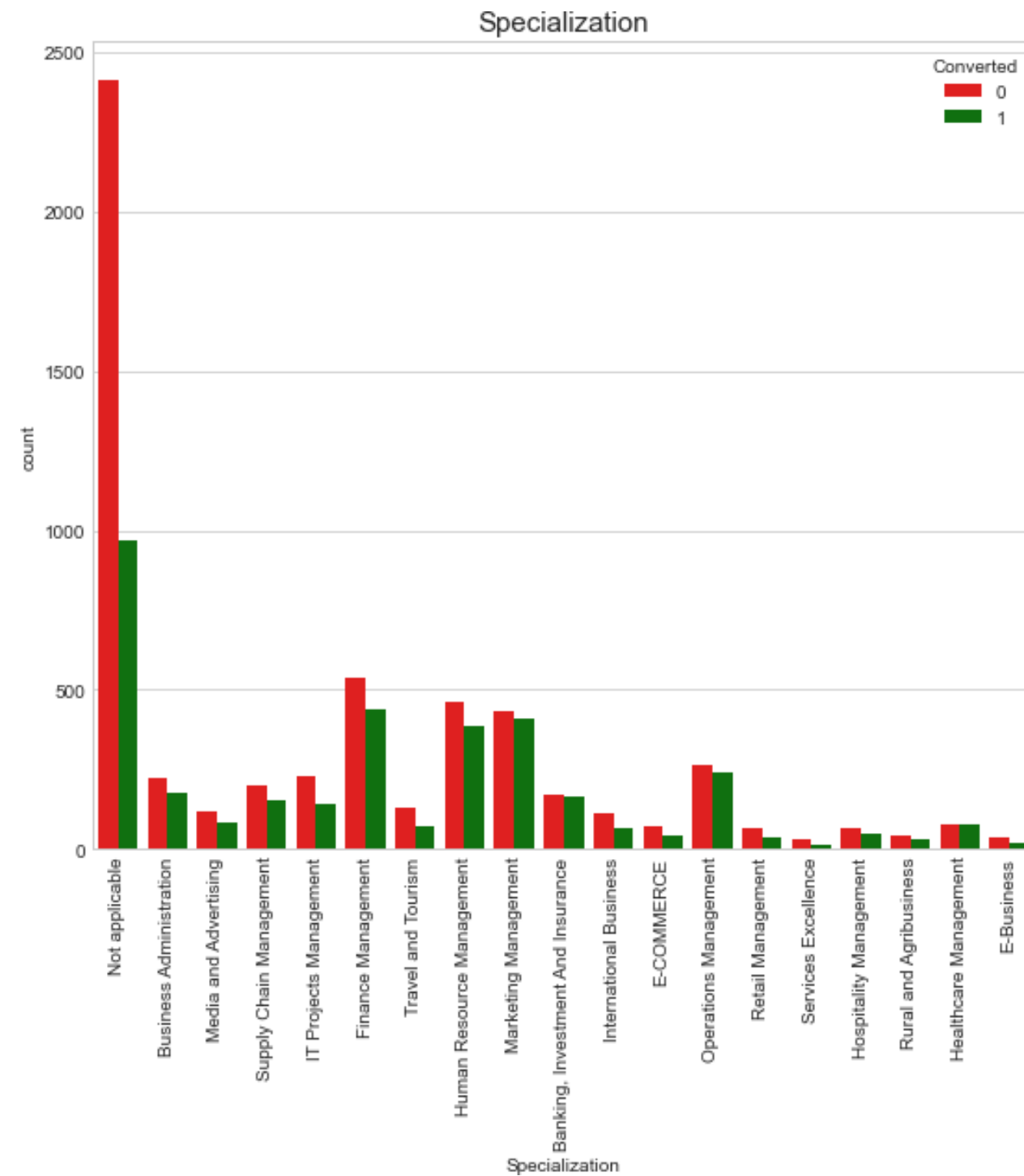
Analysis - Last Activity

Observation:

• As their most recent activity, the majority of the leads have opened their email.Lead conversion is very high (70 percent) after combining smaller Last Activity types as Other Activities.The conversion rate for leads with the most recent activity as SMS Sent is nearly 60%.
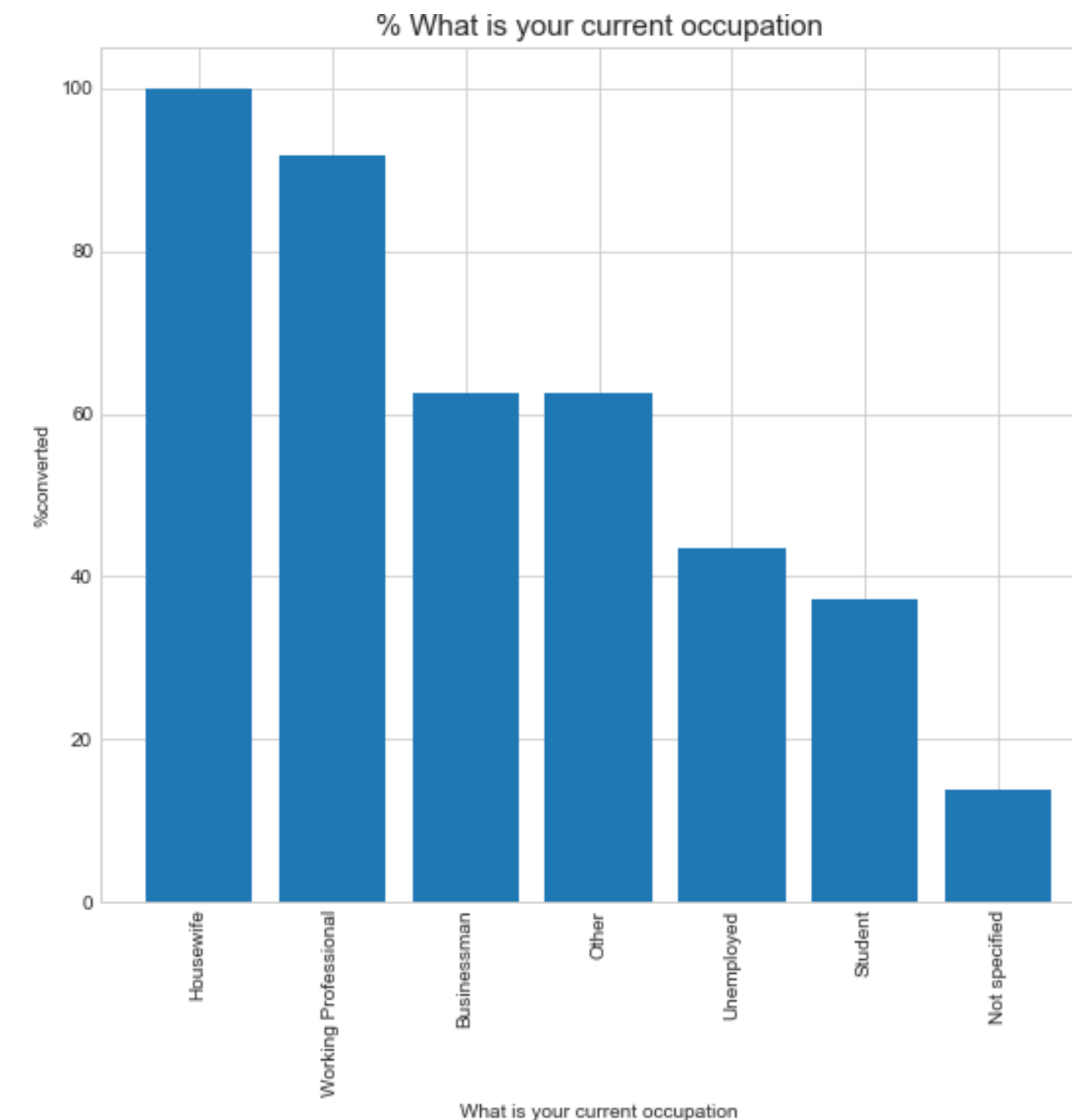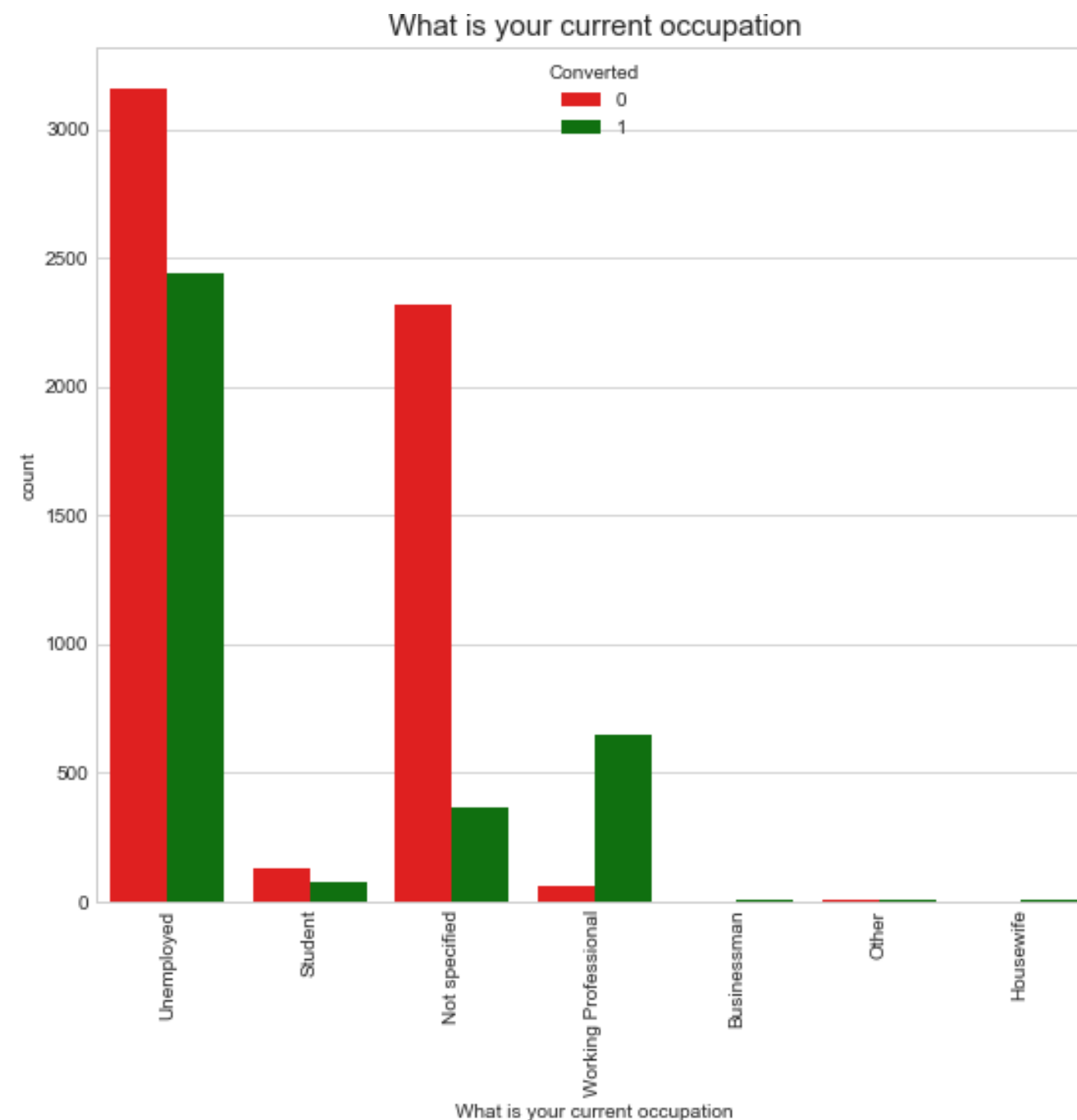
Analysis - Specialization

Observation:

- **Most of the leads have not mentioned a specialization and around 28% of those converted**
- **Leads with Banking Investment and insurance and Marketing Management - Over 45% Converted**
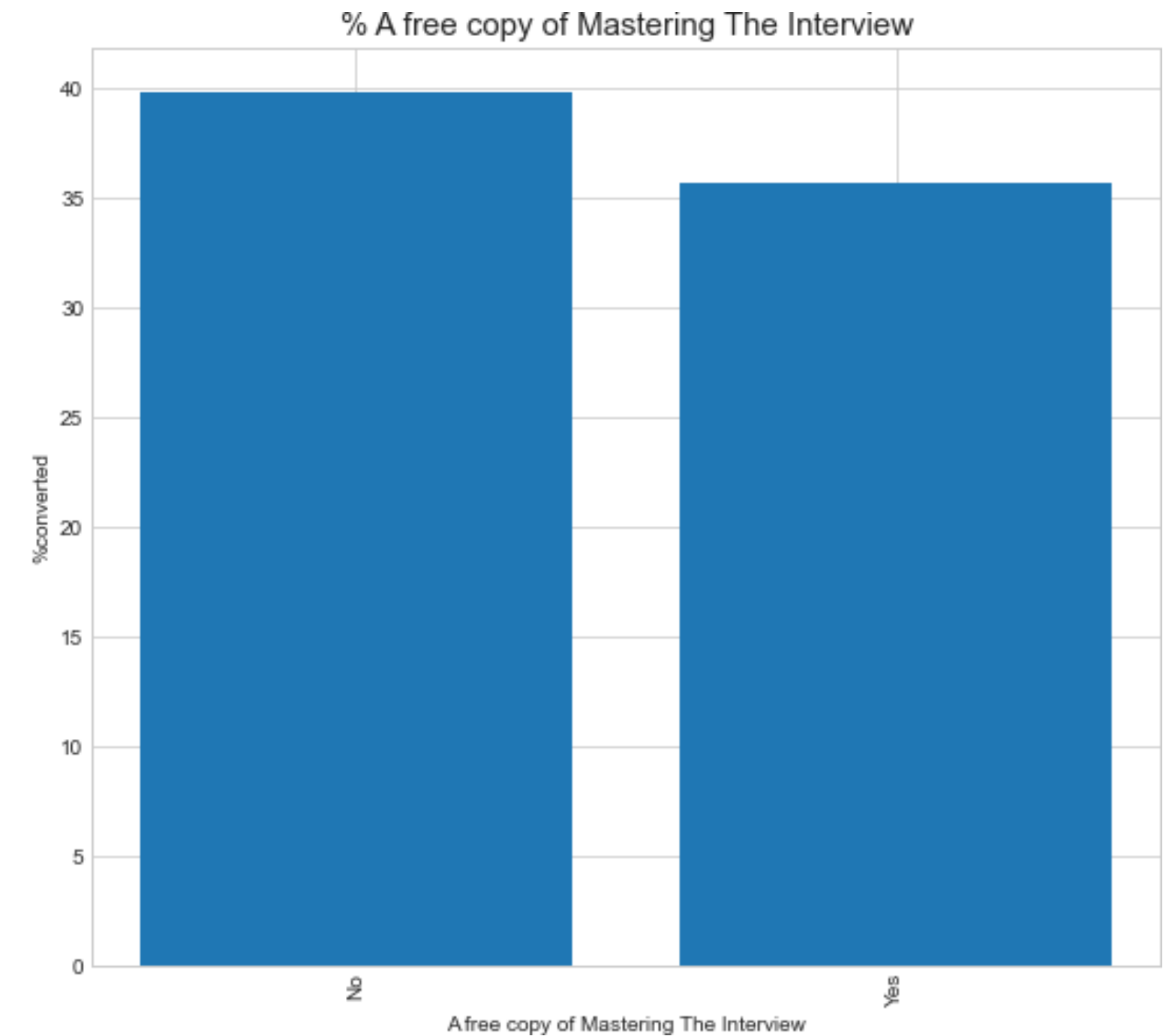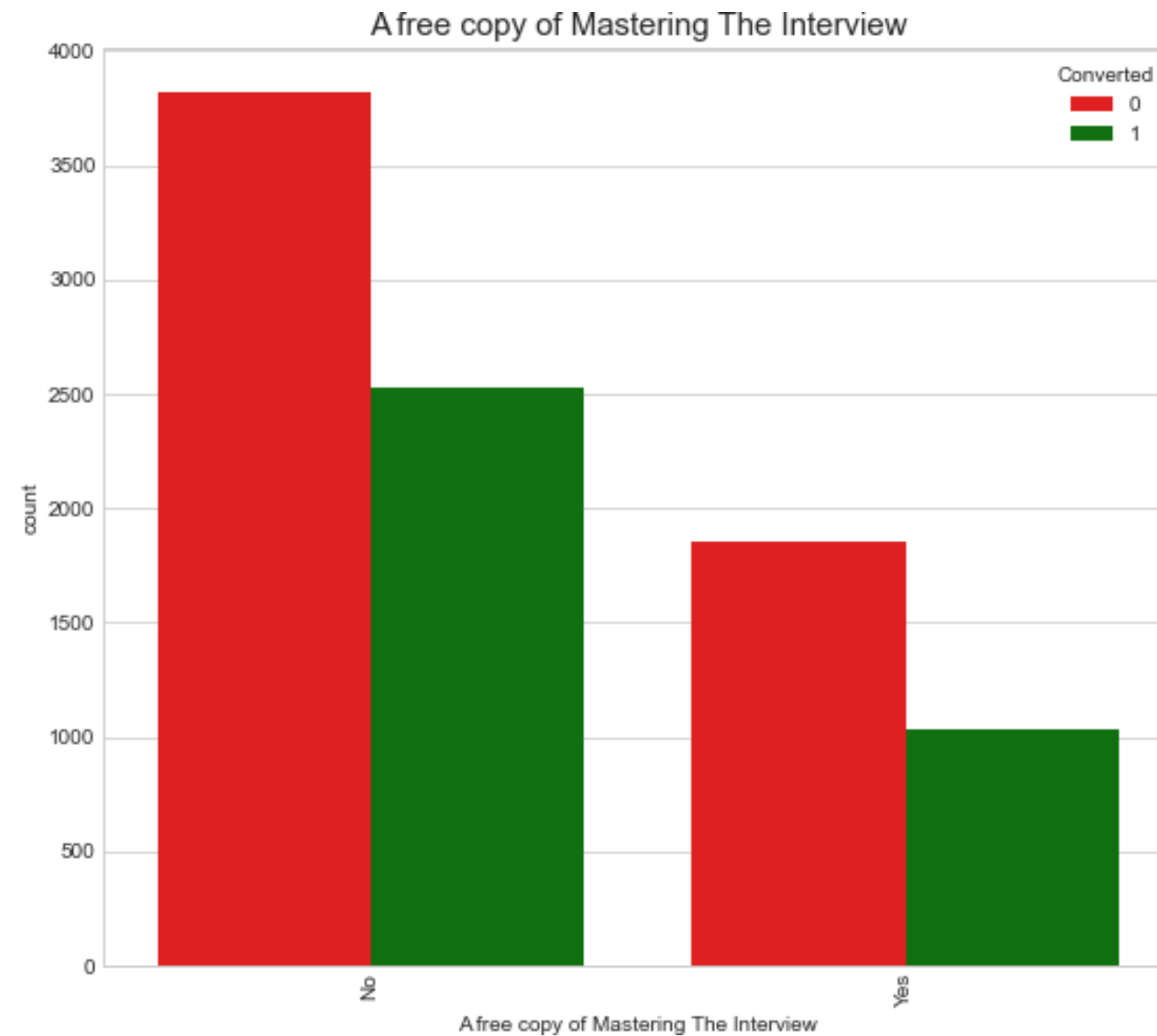
Analysis - Current Occupation

Observation:

- **Housewives are less in numbers, but have 100% conversion rate**
- **Working professionals, Businessmen and Other have high conversion rate**
- **Leads with Unemployed occupation is highest in number, but the conversion rate is low (~40%)**
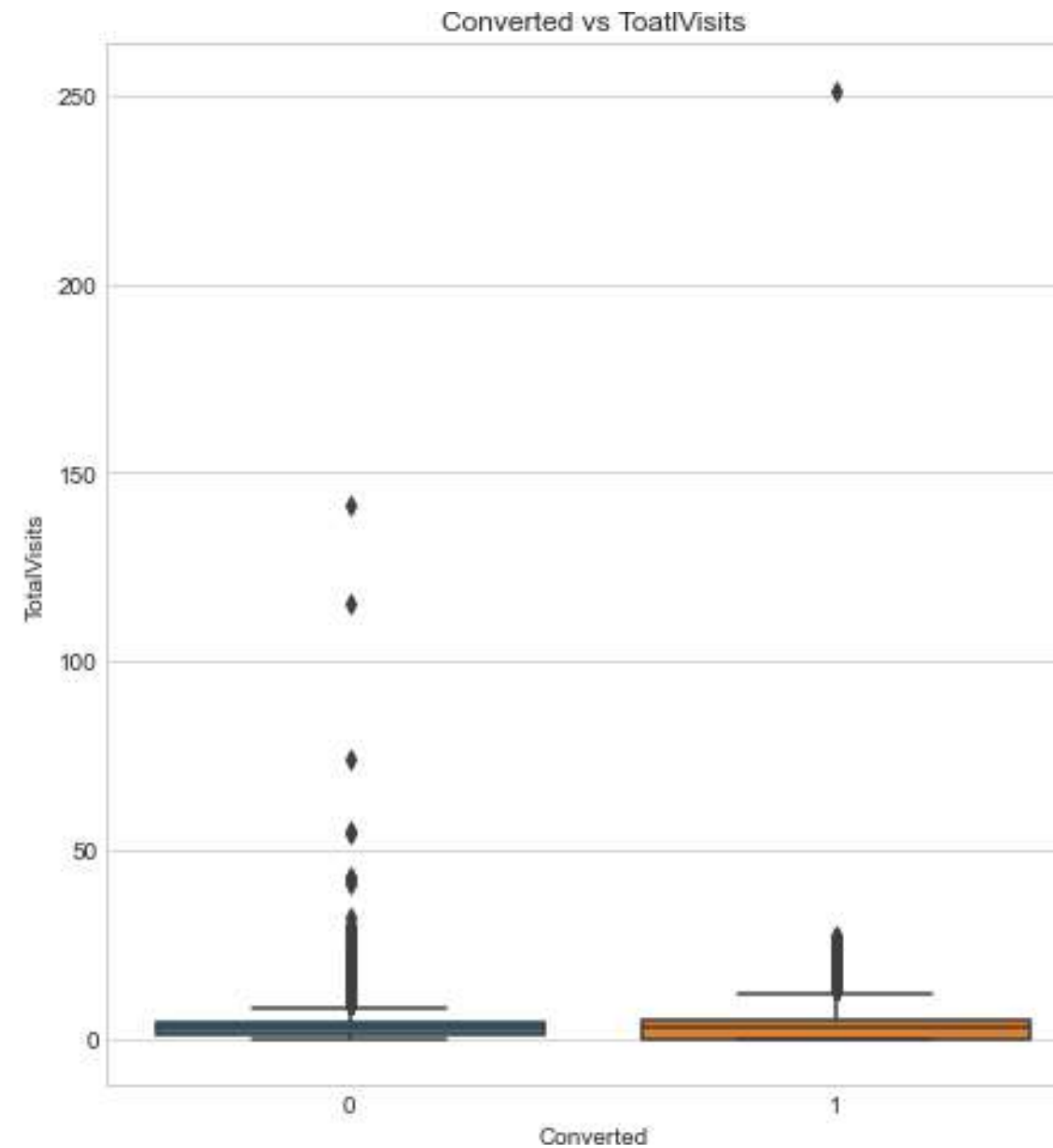
Analysis - A free copy of Mastering The Interview

Observation:

- **Distributing Free-Copy of Mastering Interview doesn't seem affect the conversion as the conversion rate is almost same.**

Analysis – Total Visits
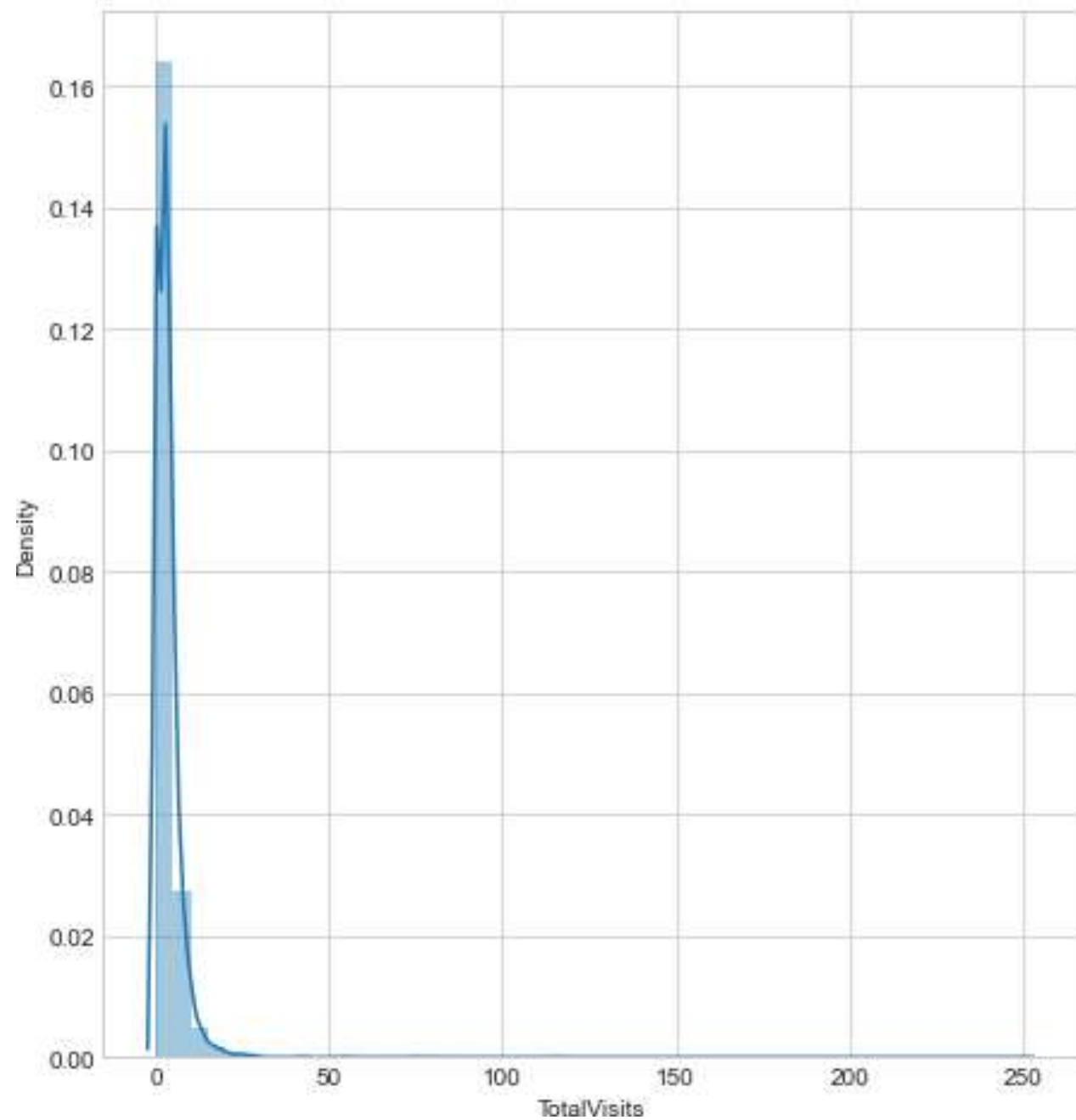
Observation:
- **Presence of outliers in Total Visit column**

Analysis – Total Time Spent on Website
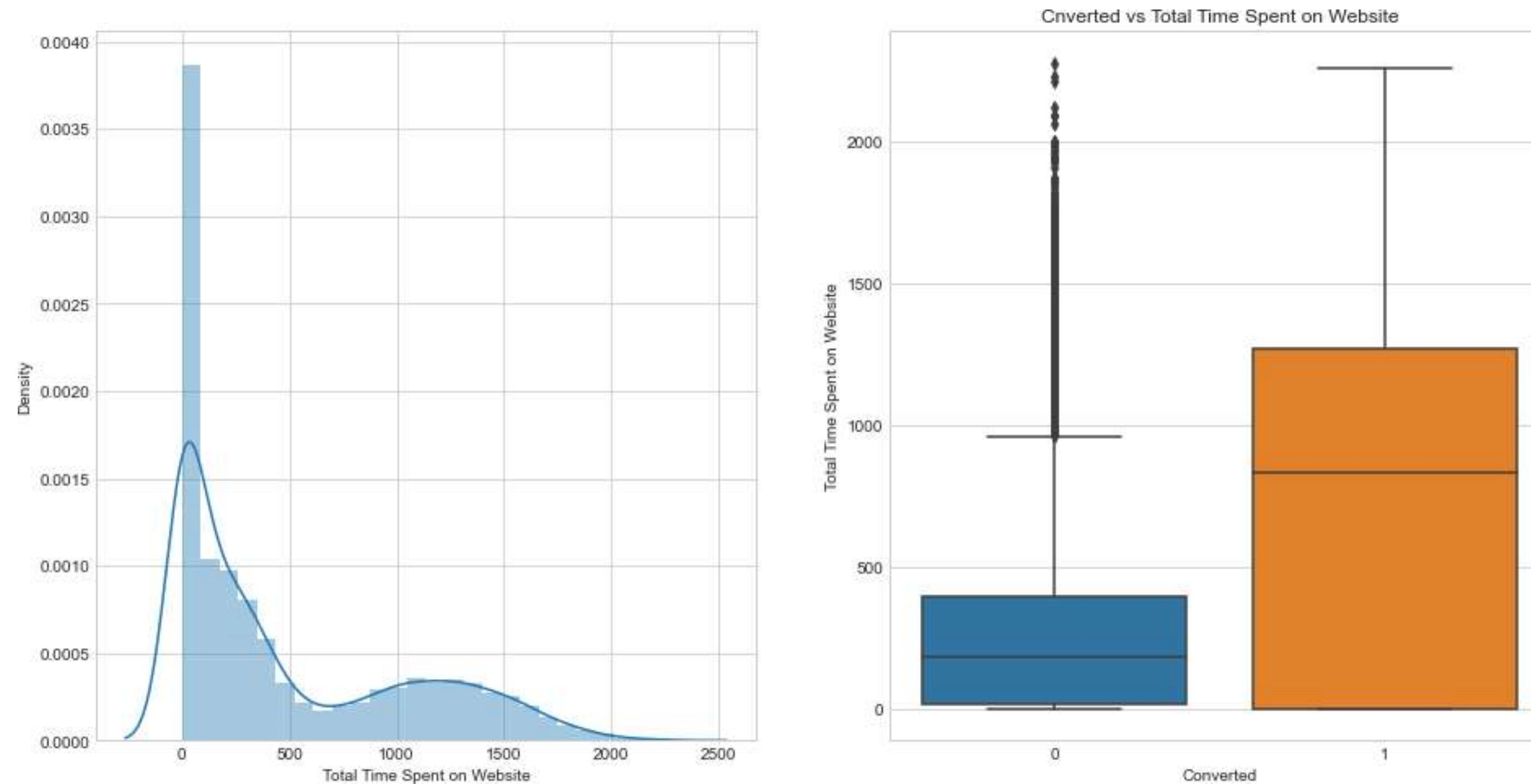
Observation:
- **Presence of outliers in Total Time Spent on Website column**
- **Leads spending more time on the website are more likely to be converted**

Analysis – Page Views Per Visit
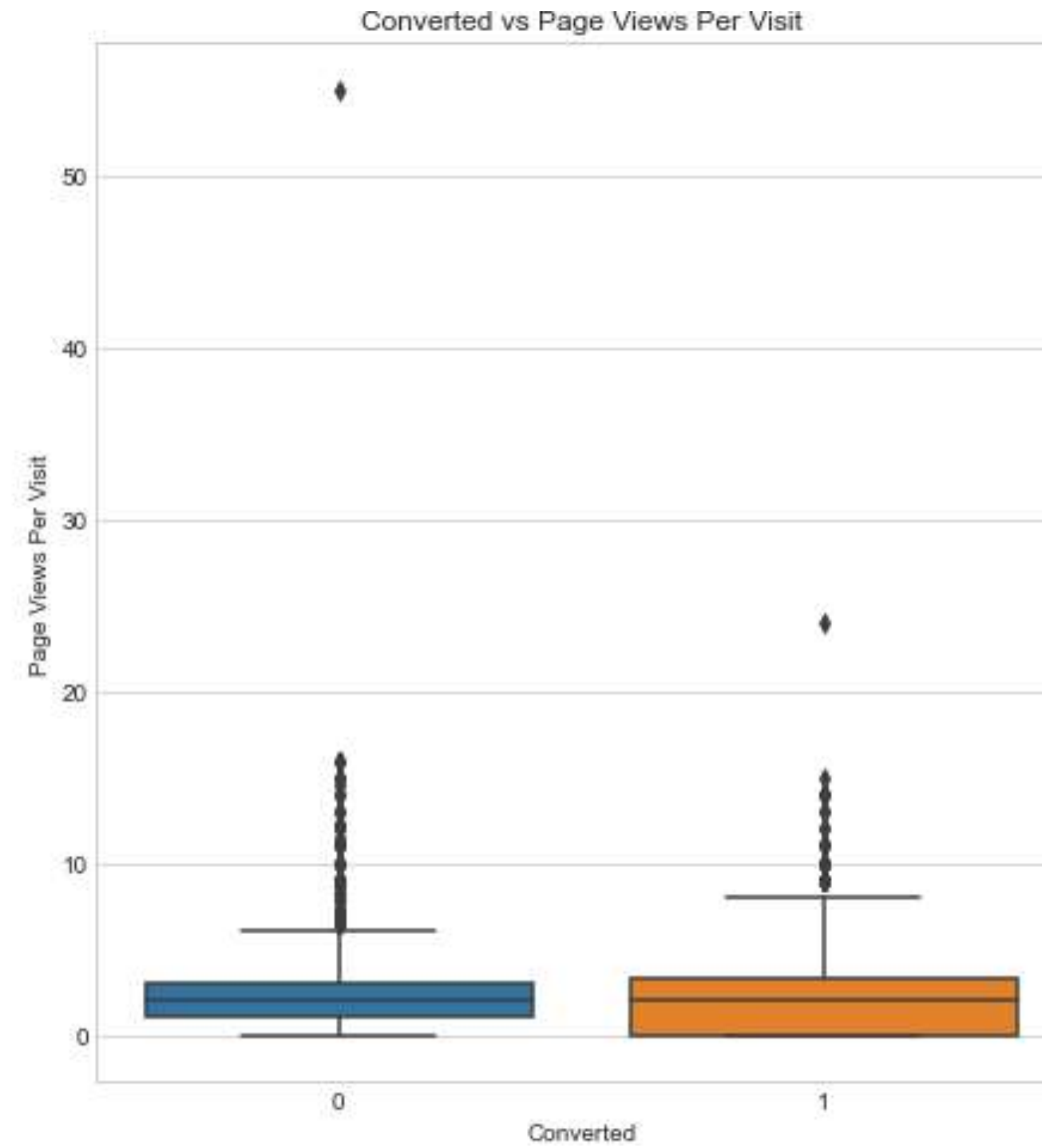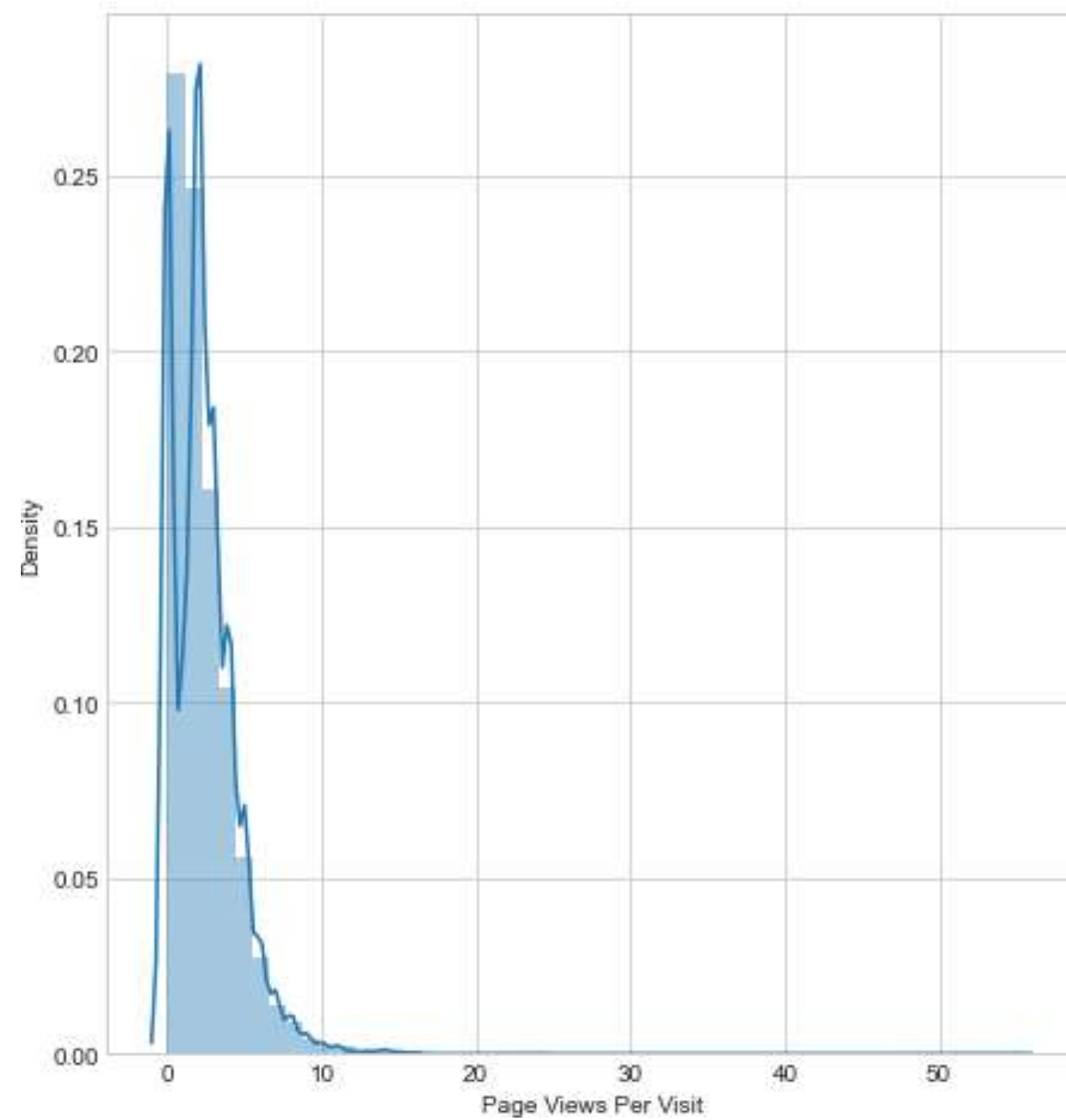
Observation:

- **Presence of outliers in Page Views Per Visit column**

Analysis- Multivariate
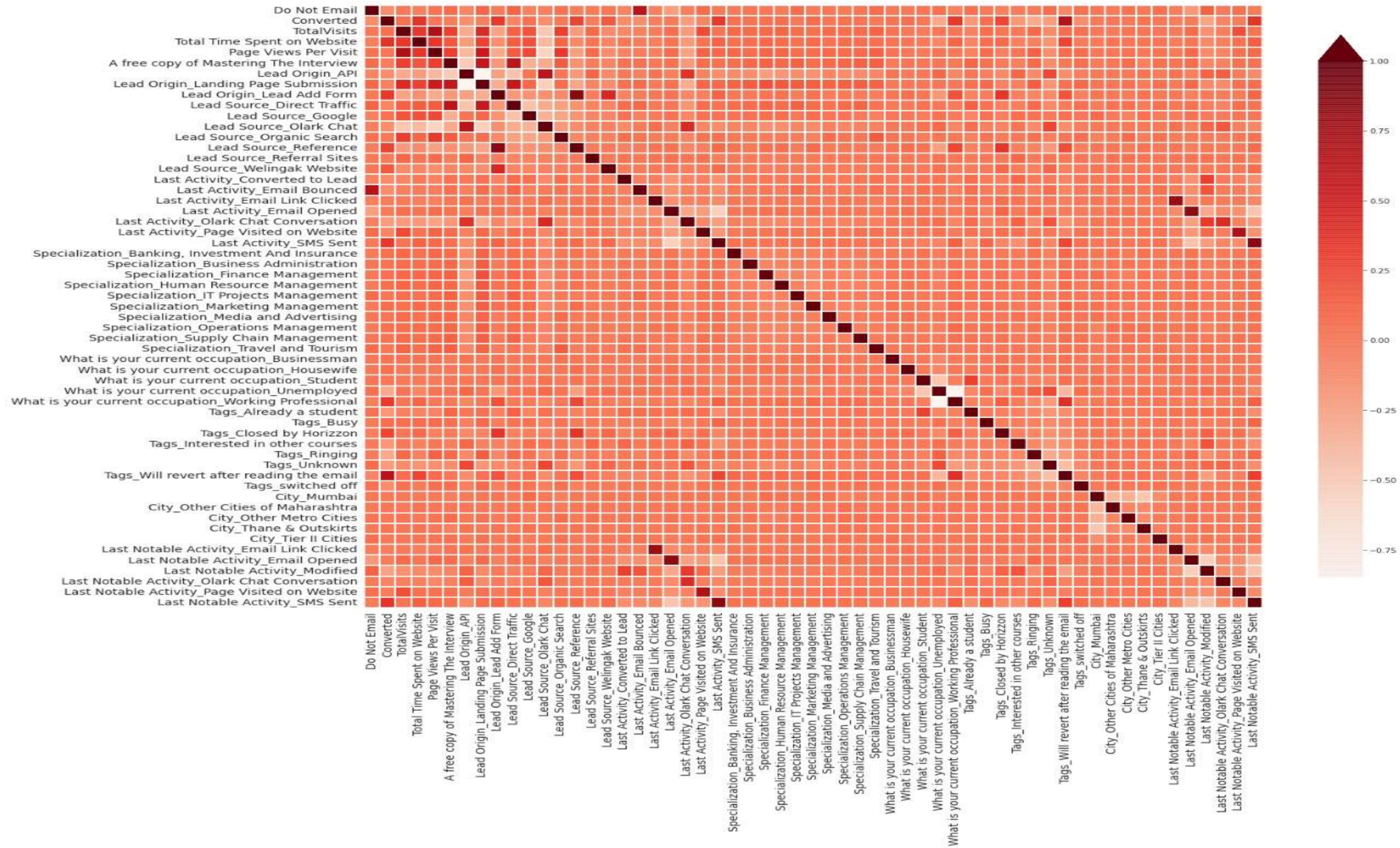
Model Evaluation – Train Data

0.335 **Probability Threshold**

Accuracy	Sensitivity	Specificity	AUC	F1 Score
80.55	80.29	80.70	0.88	75.89

Recall	Precision	Positive predictive value	Negative predictive value
80.29	71.94	71.94	86.92

Model Evaluation – Test Data

0.335 **Probability Threshold**

Accuracy	Sensitivity	Specificity	AUC	F1 Score
80.48	80.09	80.73	0.88	76.42

Recall	Precision	Positive predictive value	Negative predictive value
80.09	73.08	73.08	86.13

Final Prediction

	Converted	Cust_ID	Converted_probability	final_predicted	lead_score
0	1	4269	0.741716	1	74
1	1	2376	0.921619	1	92
2	1	7766	0.742460	1	74
3	0	9199	0.067694	0	6
4	1	4359	0.794282	1	79

Lead Score = 100 * Conversion Probability

- Higher the lead score, higher is the probability of a lead getting converted and vice versa,
- Since, we had used 0.335 as our final Probability threshold for deciding if a lead will convert or not, any lead with a lead score of

33.5 or above will have a value of '1' in the Final predicted column.

Feature Importance

1. **'Tags_Closed by Horizzon': If this variable is True or 1, then the log-odds go up by 6.84.**
2. **'Lead Source_Welingak Website': If the Lead source is Welingak website then the log odds increase by 4.26.**
3. **'Tags_Will revert after reading the email': If the current status / tag is 'Will revert after reading the email' then the log odds increase by 4.07.**
4. **'Tags_switched off': If the current status / tag is 'switched off', then the log odds decrease by 3.48.**
5. **'Tags_Already a student': If the current status / tag is 'Already a student', then the log odds decrease by 3.44.**
6. **'Tags_Ringing': If the current status / tag is 'Ringing', then the log odds decrease by 3.29.**

Top three variables in our model that contribute most towards the probability of a lead getting converted

- **Lead Origin_Lead Add Form**
- **What is your current occupation_Working Professional**
- **Lead Source_Welingak Website**

Recommendation

- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and Google leads and generate more leads from reference and welingak website.

- Lead conversion rate, can be improved by focusing more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form

- Though Google is the highest source to get leads, the lead conversion through Google is low comparatively.

- Focus on Working Professional which has high conversion

- Website should be made more engaging to make leads spend more time

- Improve the Olark Chat service since this is affecting the conversion negatively