LEAD SCORE CASE-STUDY

BY USING LOGISTIC REGRESSION

Submitted By,
Biraj Mukherjee
Chaitanya Pande
Chaithanya. U

CONTENTS

- Problem Statement
- Methodology
- EDA
- Outliers
- Correlations
- Model Evaluation
- Observation
- Conclusion

PROBLEM STATEMENT

- X Education, an online course provider for industry professionals, aims to improve its lead conversion process. Currently, they acquire leads through form submissions on their website and have a conversion rate of about 30%.
- This means out of 100 daily leads, only 30 are typically converted. To enhance efficiency, X Education wants to identify "Hot Leads"—the leads with the highest conversion potential.
- By focusing on these high-potential leads, the sales team can improve the conversion rate, making their efforts more effective and targeted.

BUSINESS OBJECTIVE

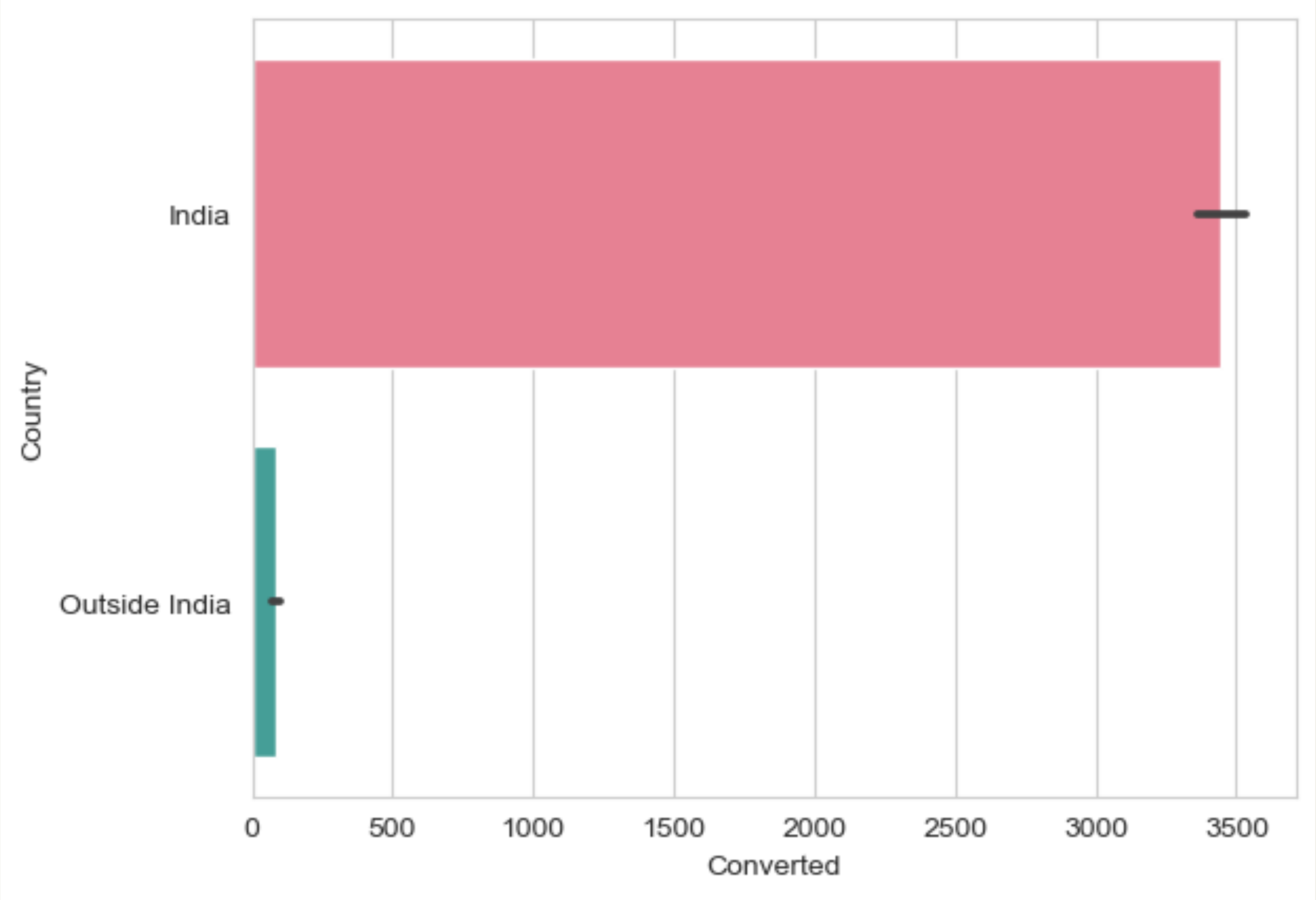
X Education is an education company that sells online courses to industry professionals. When interested professionals visit their website and browse courses, they fill out a form, which makes them leads. However, the lead conversion rate is around 30%, meaning that out of 100 leads, only about 30 are converted. To make the sales process more efficient, X Education wants to identify the most potential leads, or "Hot Leads," so that the sales team can focus on communicating with them instead of making calls to everyone. This should increase the lead conversion rate.

METHODOLOGY

- Importing and inspecting the data
- Data preparation
- Exploratory Data Analysis (EDA)
- Creating dummy variables
- Splitting data into training and testing sets
- Scaling features
- Analyzing correlations
- Building models (RFE, R-squared, p-values)
- Evaluating models
- Predicting on the test set

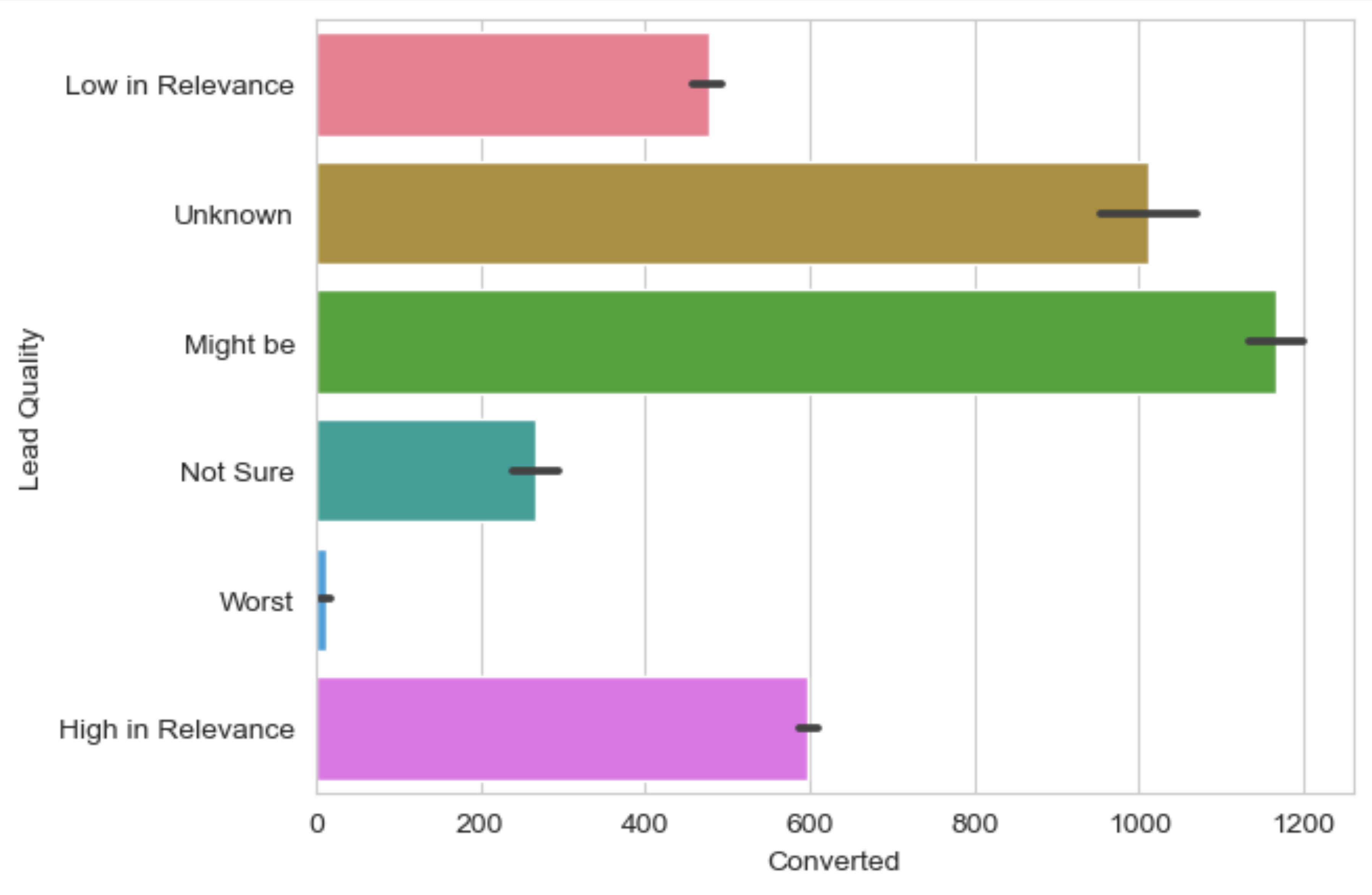
E D A

- Some columns contain a category labeled 'Select', which has been addressed.
- Assigning 'Select' values a unique category enriches data, altering variance. Common due to non-mandatory fields on the dataset



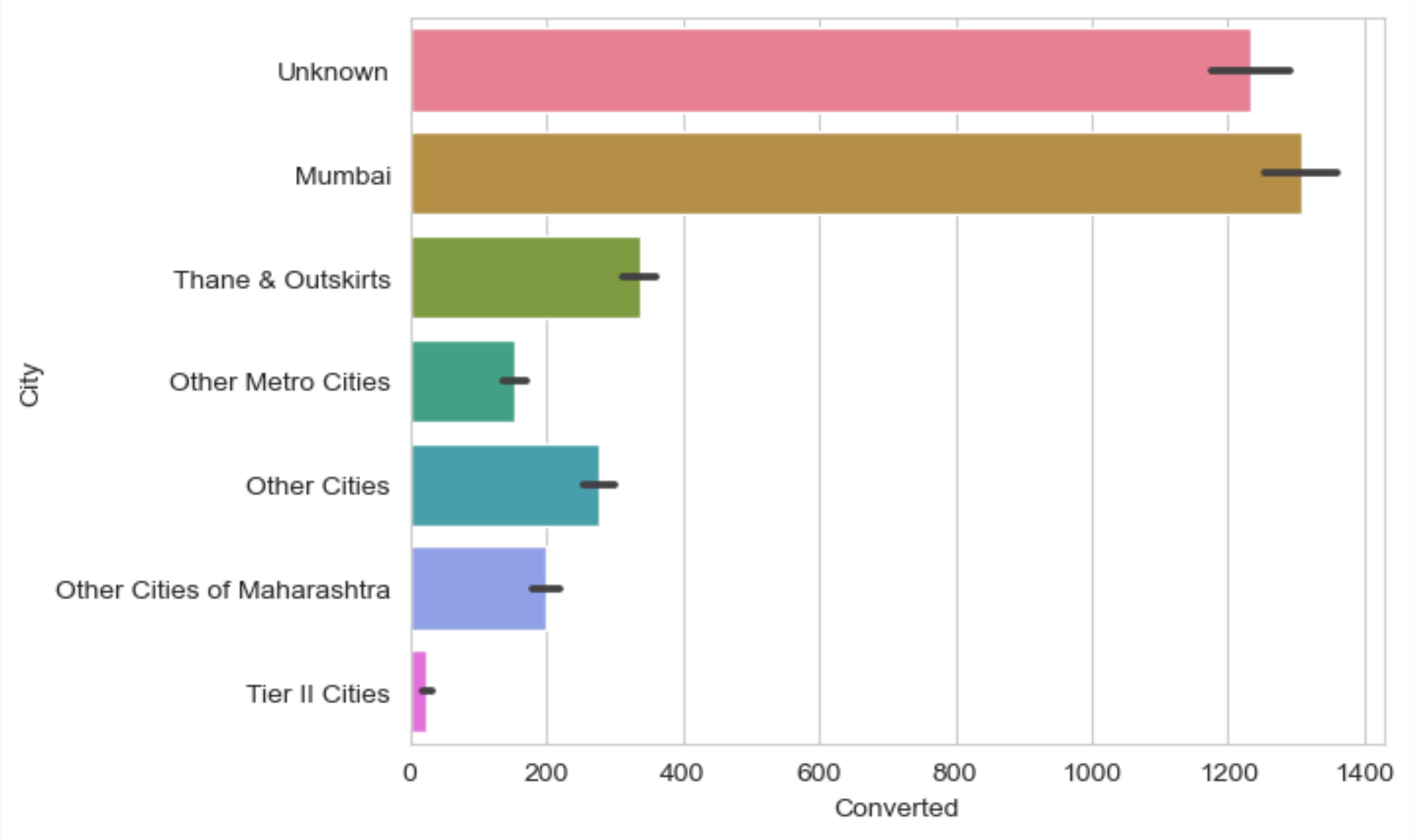
LEAD QUALITY VS CONVERTED

- This particular graph depicts the variation of lead quality across the target variable converted.
- x = Converted(Target variable)
- y = Lead quality (Quality of leads for conversion)



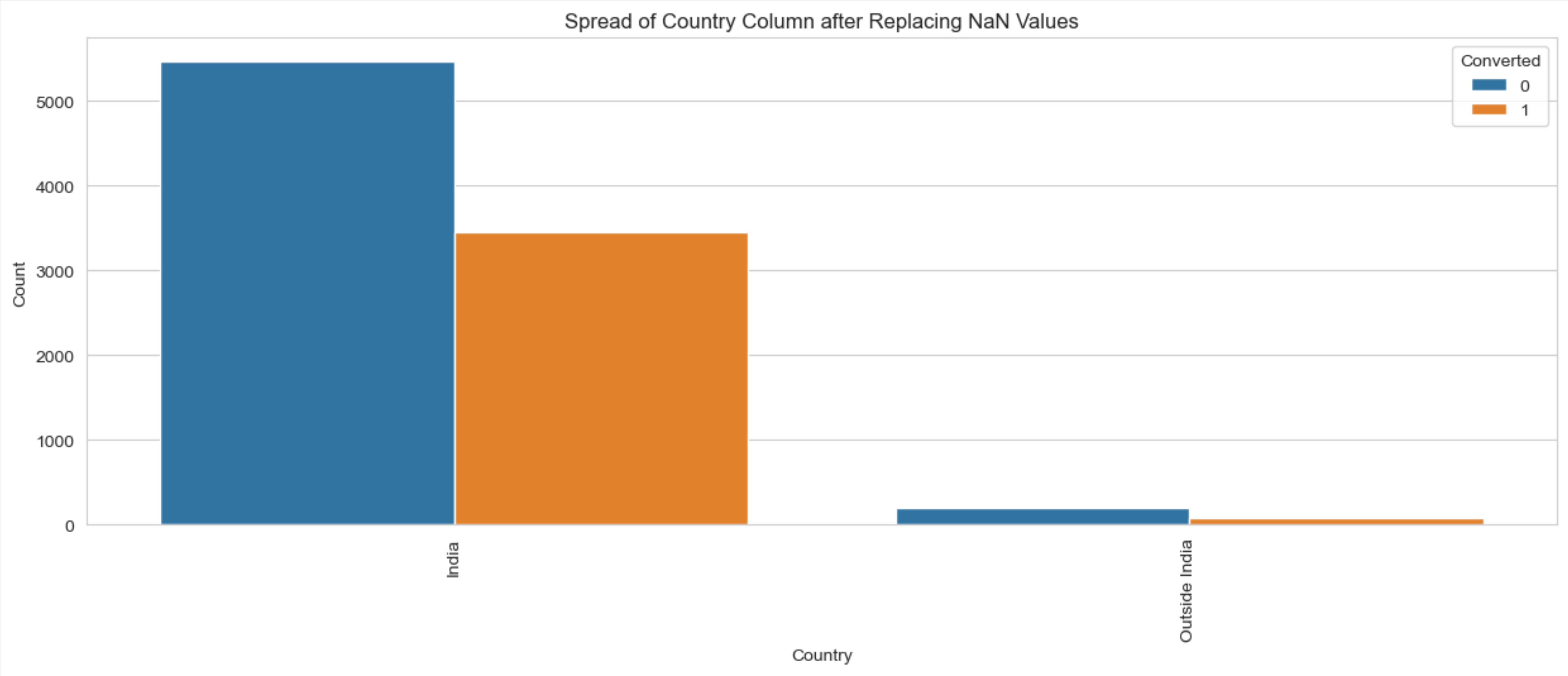
CITY VS CONVERTED

- This particular graph depicts the variation of converted target variable across various cities.
- x = Converted(Target variable)
- y = Cities



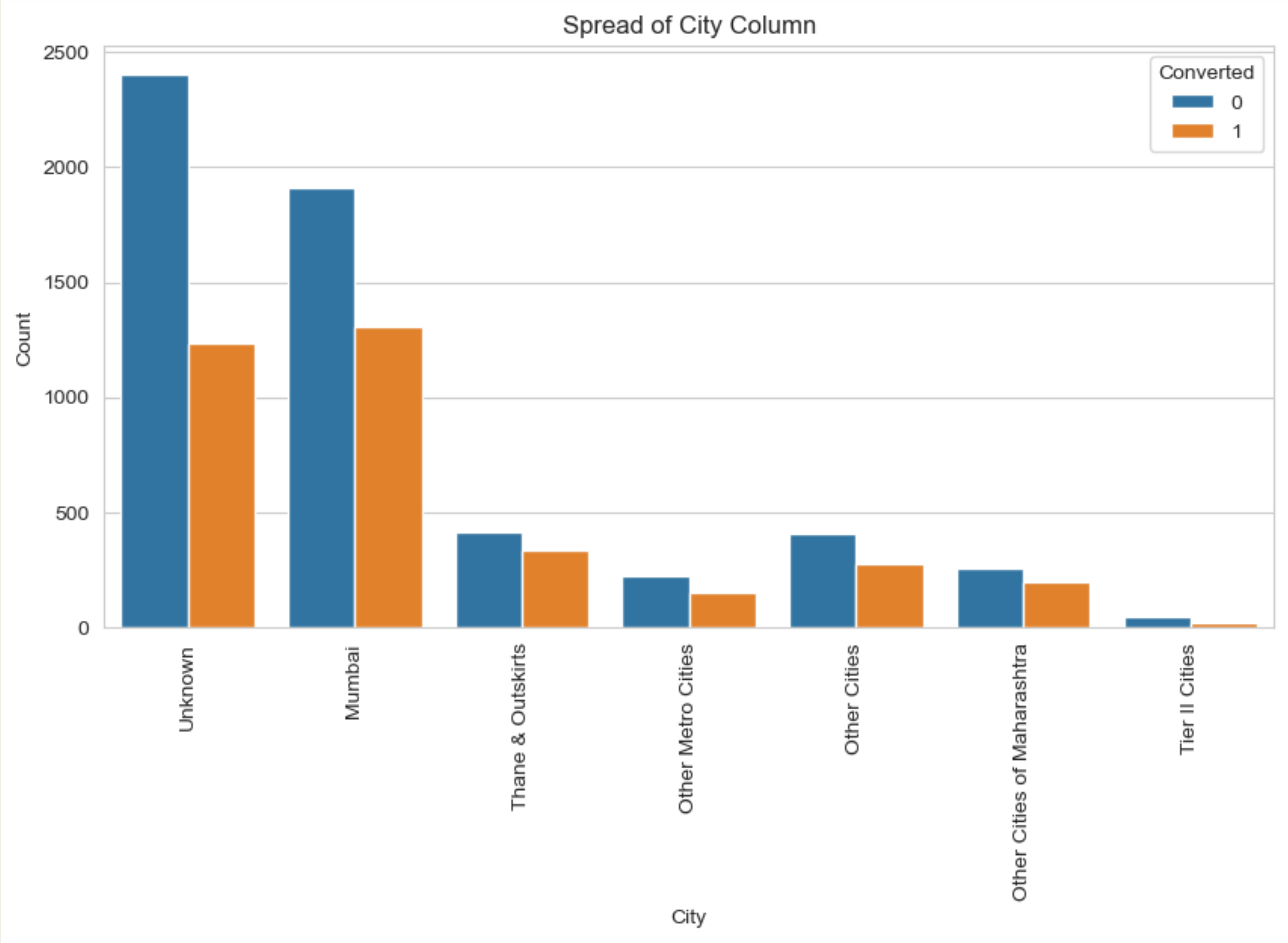
CATEGORICAL DATA VISUALIZATION

Upon the categorical segmentation of country, we got to visualize that :- Given that the vast majority of entries (~97%) in the 'Country' column are for India, it can be safely removed.



COMPETITIVE ANALYSIS

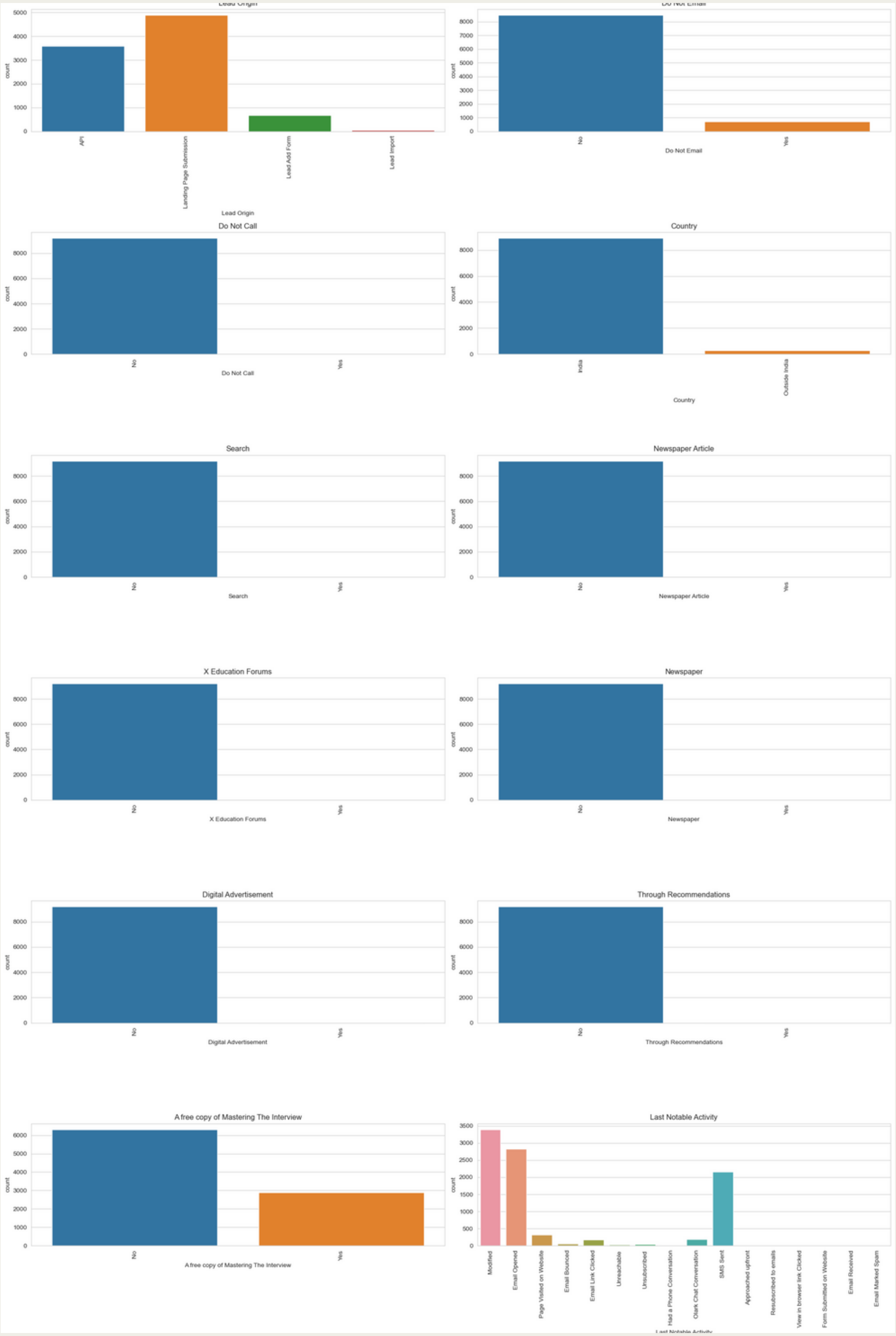
Distribution of leads or conversions by city in Maharashtra, India. The graph shows the count of leads or conversions for each city, with the 'Unknown' category representing leads or conversions that could not be attributed to a specific city. The 'Converted' category represents leads that have been successfully converted into customers.



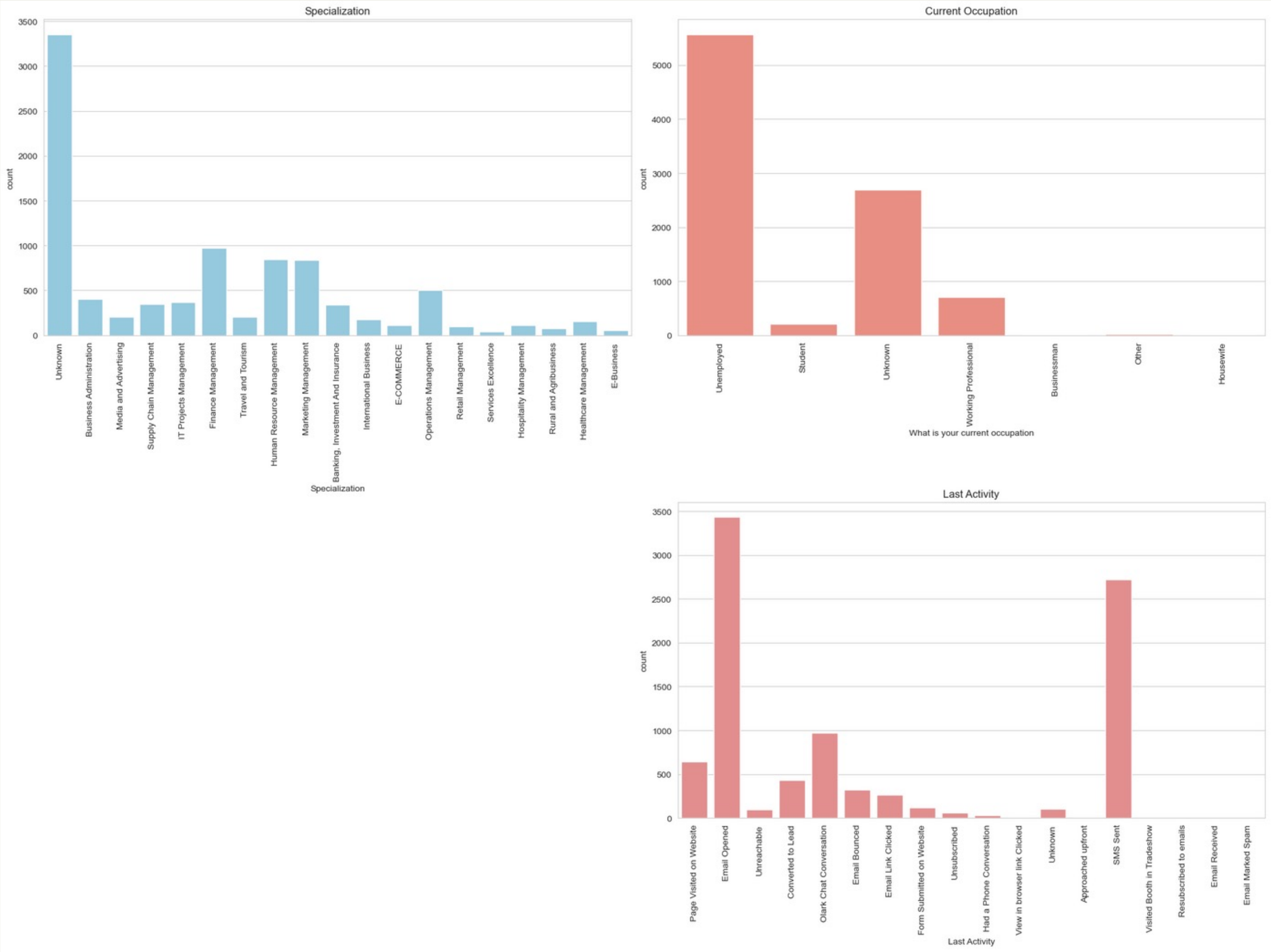
Distribution of leads or conversions by various lead origins, contact preferences, and marketing channels.

The graph shows the count of leads or conversions for each category within each column, with the x-axis labels rotated for readability.

The columns represented include lead origin, contact preferences (Do Not Email, Do Not Call, and Last Notable Activity), and marketing channels (Search, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, and A free copy of Mastering The Interview)



Visualization of the distribution of leads or conversions by specialization, current occupation, and last activity. The resulting plot would show the count of leads or conversions for each unique value in each of these columns, with the x-axis labels rotated for readability.



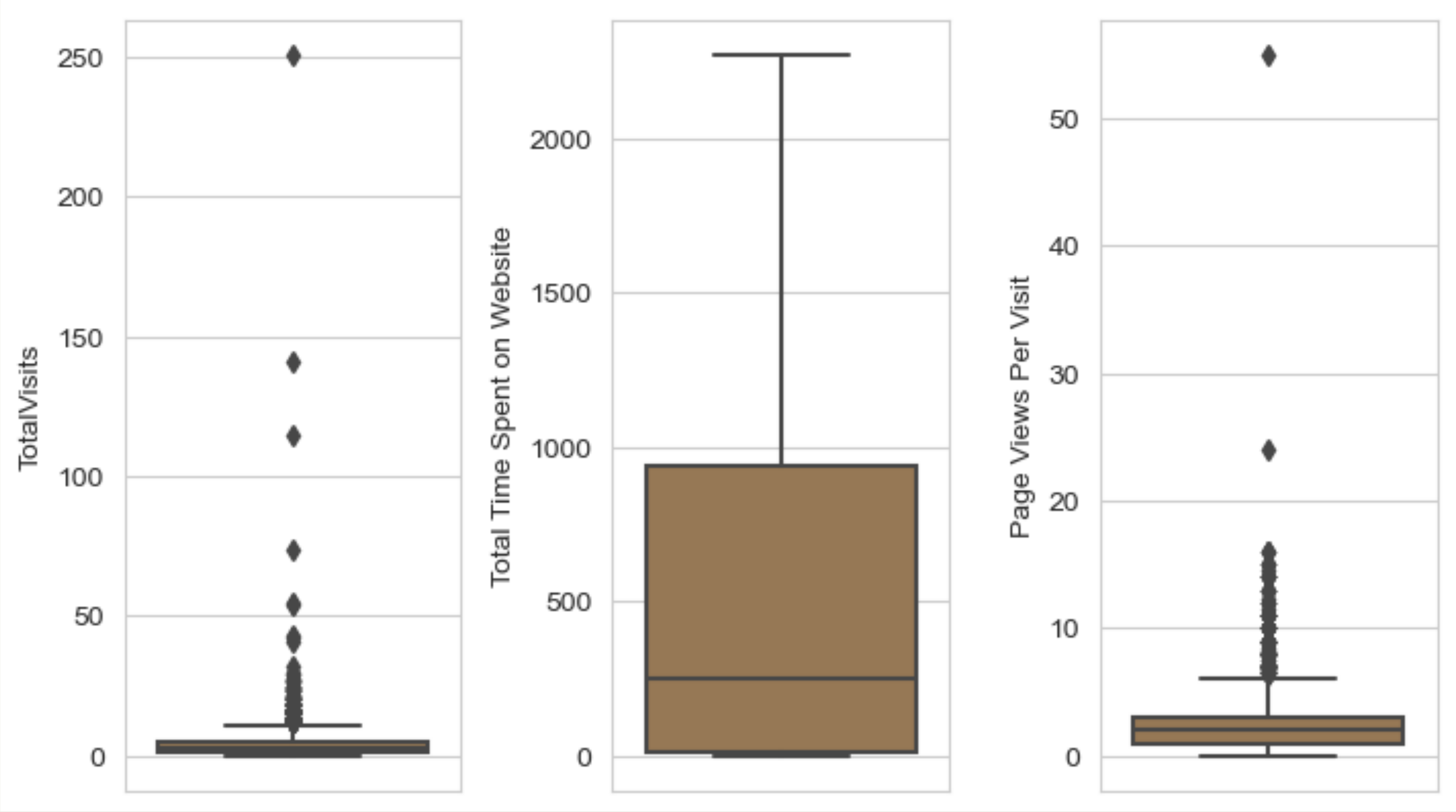
Classification: Internal

OUTLIERS

This outlier boxplots represents three numerical variables , namely,

1. Total Visits
2. Total time spent on the website
3. Page views per visit

We can clearly visualize from the plot that total visits and page view per visit attributes has an outlier.

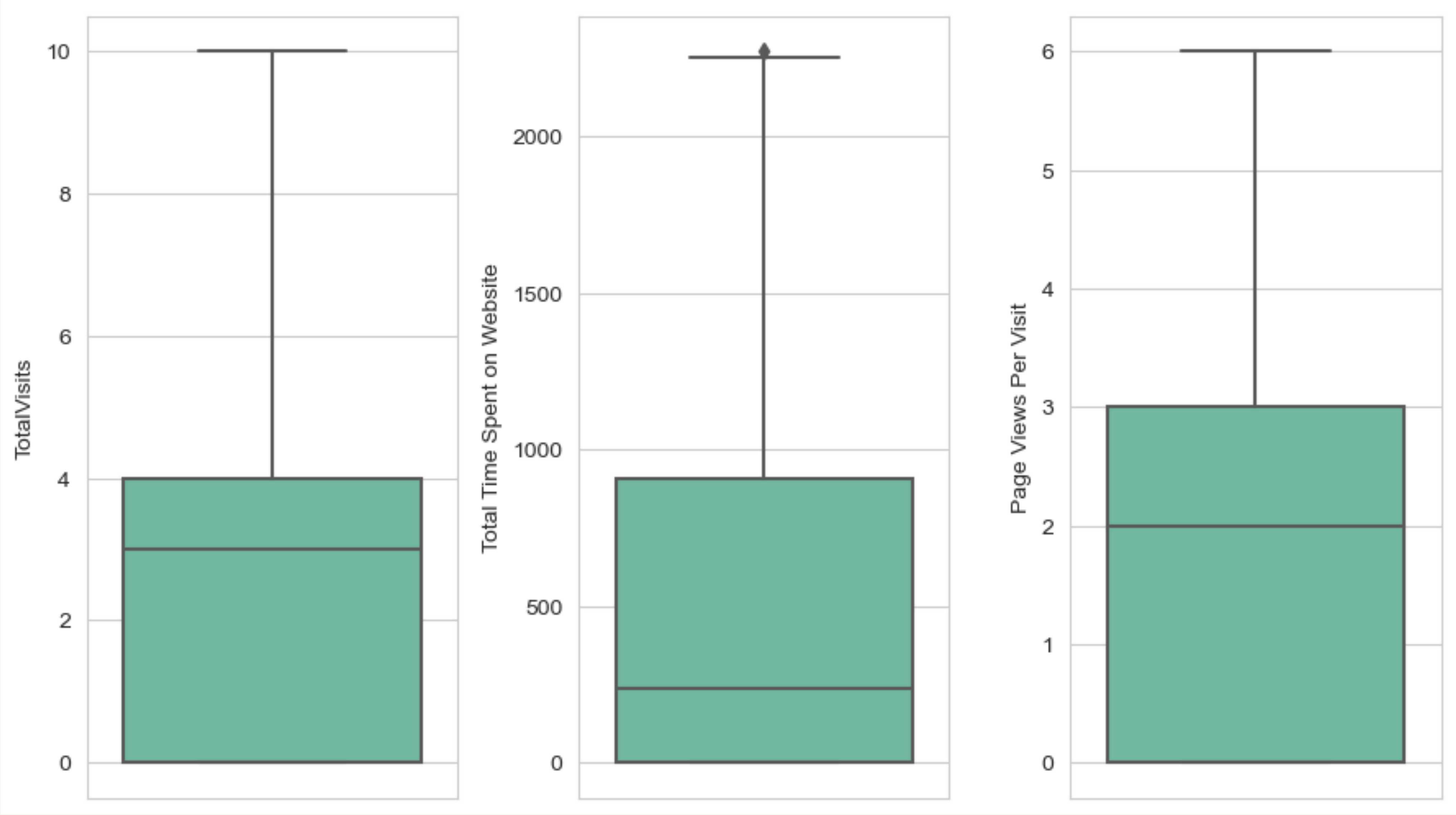


Classification: Internal

OUTLIERS(CONTINUATION)

Removing outlier values based on the Interquartile distance for some of the continuous variable.

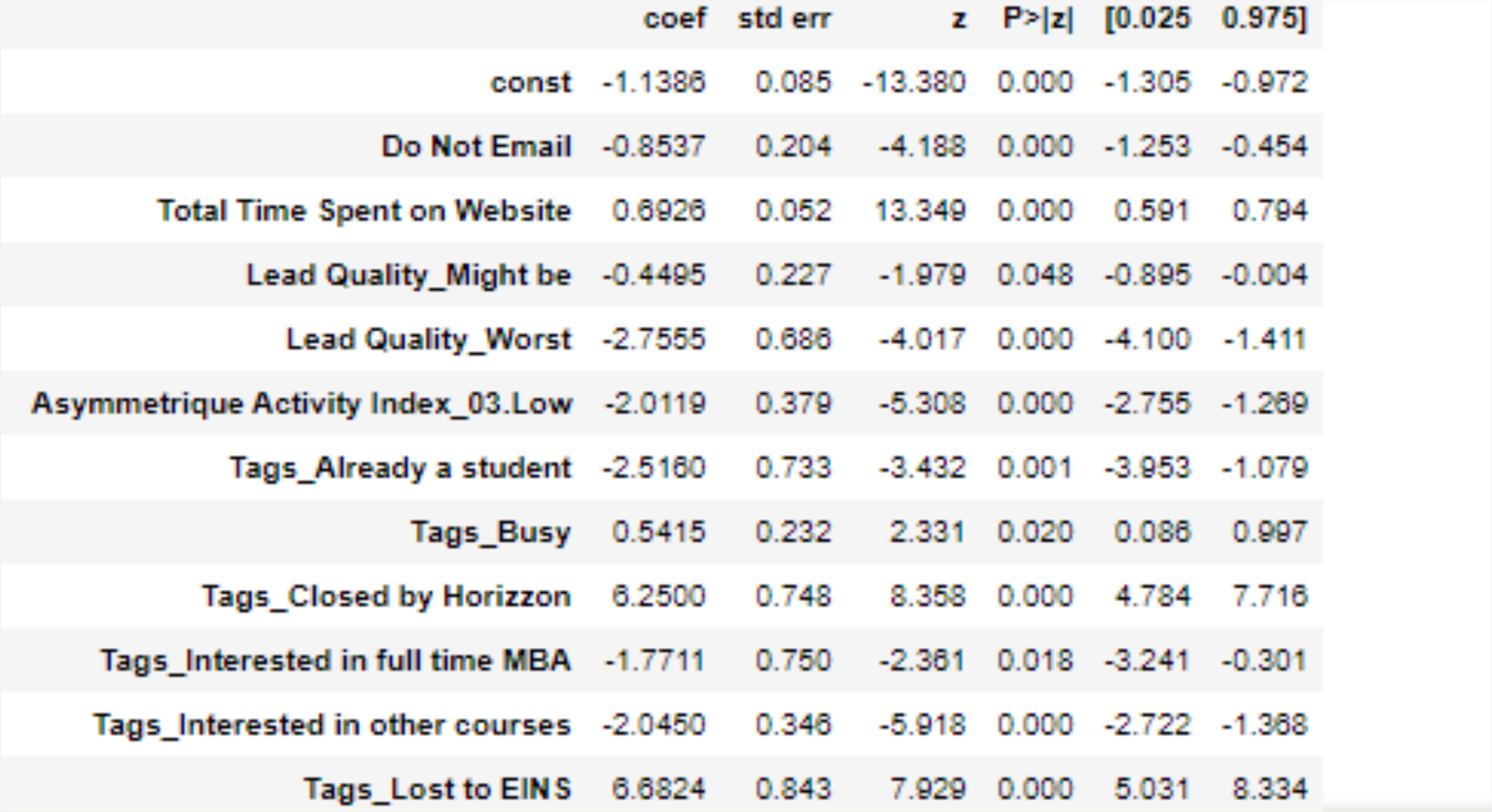
Below is the boxplot is shown:



Classification: Internal

FINAL MODEL

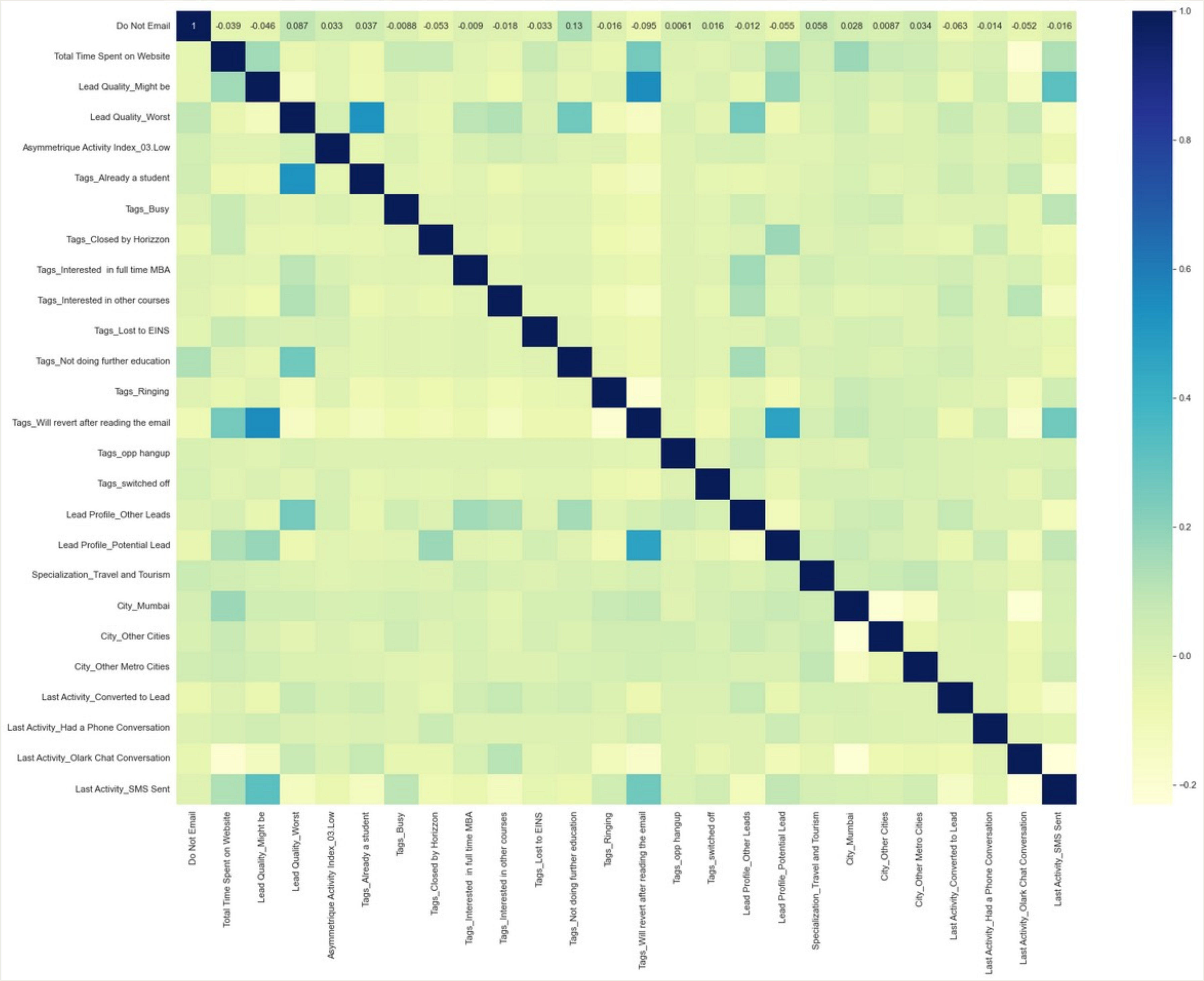
This is the final model we build, and the below stats shows the p value for all the feature variables are >0.05.So, we conclude that the model which we have created is meeting the criteria.



	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1386	0.085	-13.380	0.000	-1.305	-0.972
Do Not Email	-0.8537	0.204	-4.188	0.000	-1.253	-0.454
Total Time Spent on Website	0.6926	0.052	13.349	0.000	0.591	0.794
Lead Quality_Might be	-0.4495	0.227	-1.979	0.048	-0.895	-0.004
Lead Quality_Worst	-2.7555	0.686	-4.017	0.000	-4.100	-1.411
Asymmetrique Activity Index_03.Low	-2.0119	0.379	-5.308	0.000	-2.755	-1.269
Tags_Already a student	-2.5160	0.733	-3.432	0.001	-3.953	-1.079
Tags_Busy	0.5415	0.232	2.331	0.020	0.086	0.997
Tags_Closed by Horizzon	6.2500	0.748	8.358	0.000	4.784	7.716
Tags_Interested in full time MBA	-1.7711	0.750	-2.361	0.018	-3.241	-0.301
Tags_Interested in other courses	-2.0450	0.346	-5.918	0.000	-2.722	-1.368
Tags_Lost to EINS	6.6824	0.843	7.929	0.000	5.031	8.334

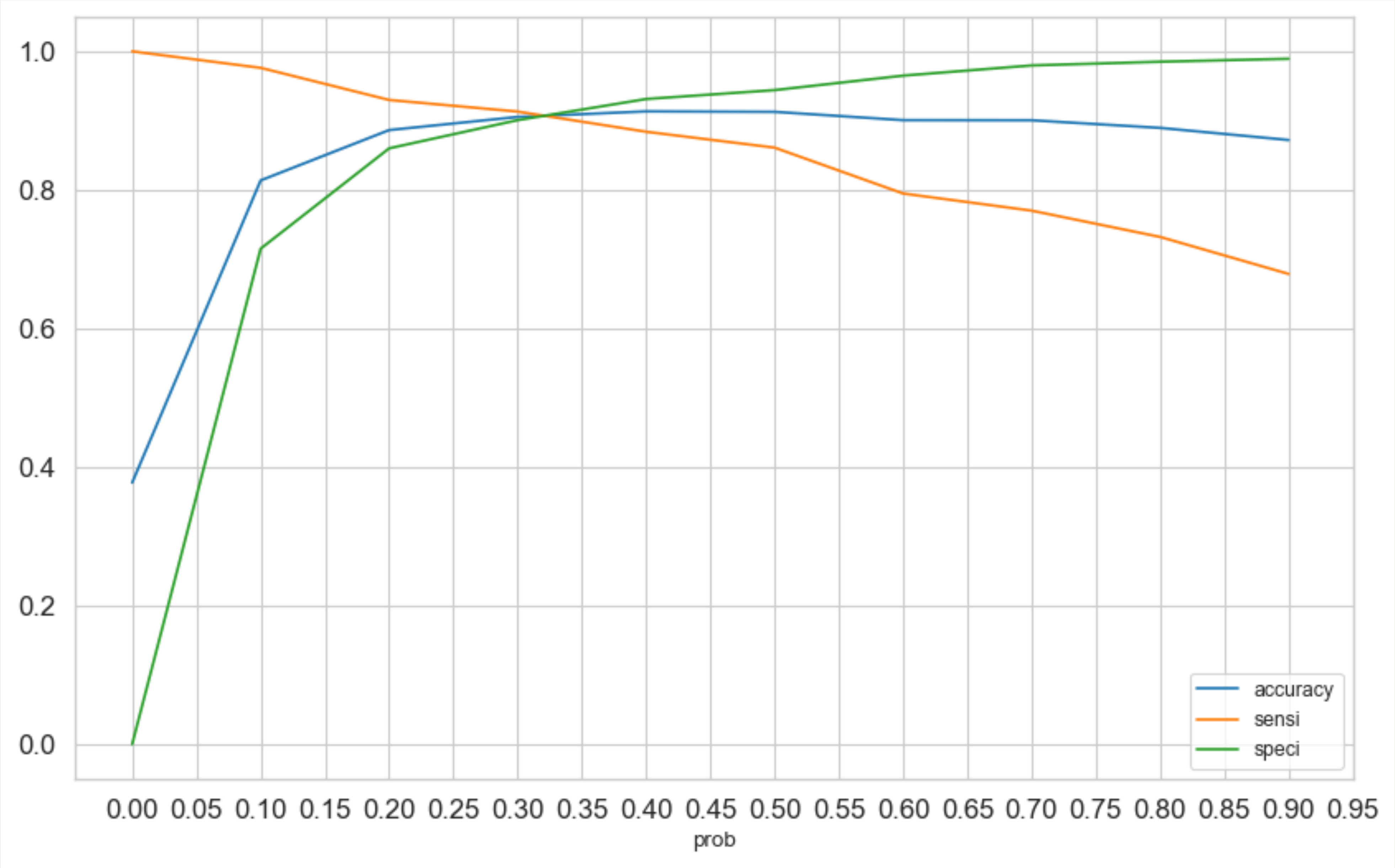
CORRELATION

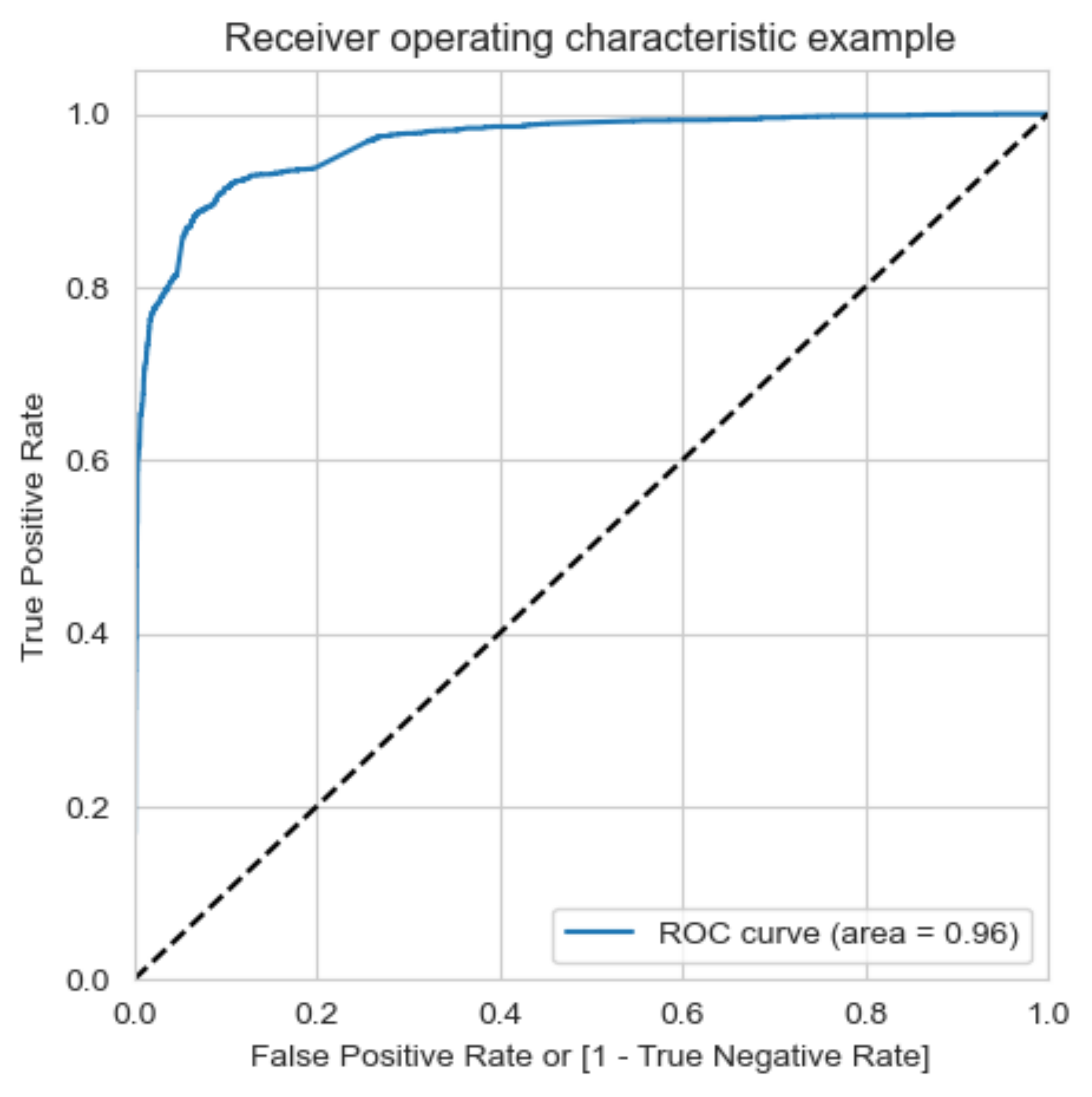
It is a correlation matrix which depicts the correlation or a heatmap between the variables.



MODEL EVALUATION-ROC CURVE

From the second curve , the cut off point is between 0.3 ~ 0.35 which comes around 0.33 as curtoff probability.





CONCLUSION

After trying several models, we finally chose a model with the following characteristics:

- All variables have p-value < 0.05.
- All the features have very low VIF values, meaning, there is hardly any muliticollinearity among the features. - This is also evident from the heat map.
- The overall accuracy of 91.06% at a probability threshold of 0.33 on the test dataset is also very acceptable.

SUMMARY

The important features responsible for a good conversion rate or the ones that contribute more towards the probability of a lead getting converted are:

1. Tags_Lost to EINS 100.00

2. Tags_Closed by Horizzon 93.53

3. Tags_Will revert after reading the email 69.14