X Education - Lead Scoring Case Study

HAMZA

MANISHA

HARSH H

Background

X Education Company

- X Education , An education company named sells online courses to industry professionals
- Many interested professionals land on their website
- The company markets its courses on several websites 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
- When these people fill up a form providing their email address or phone number, they are classified to be a lead
- 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
- The typical lead conversion rate at X education is around 30%

Problem Statement

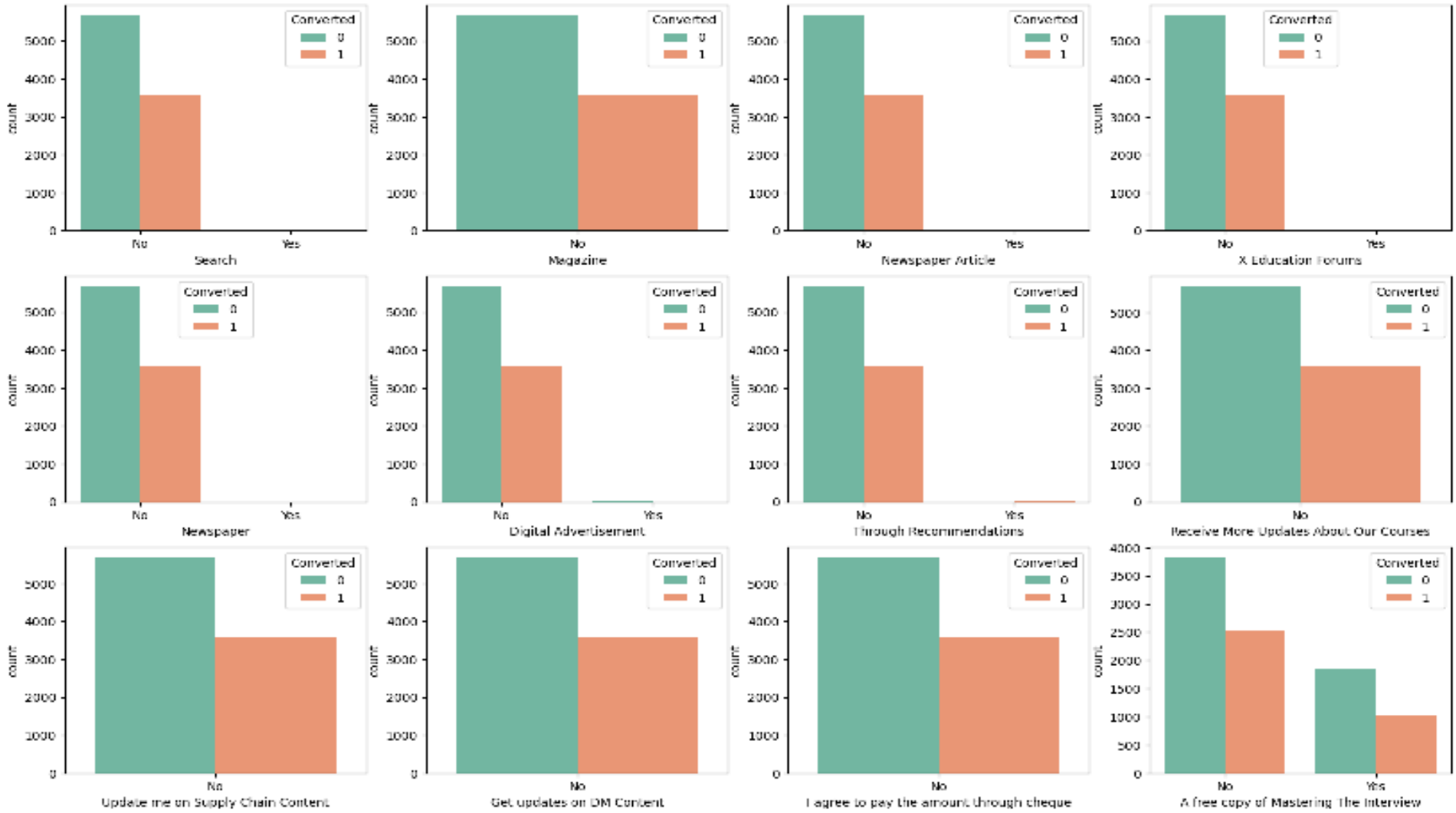
X Education Company’s Problem

- X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

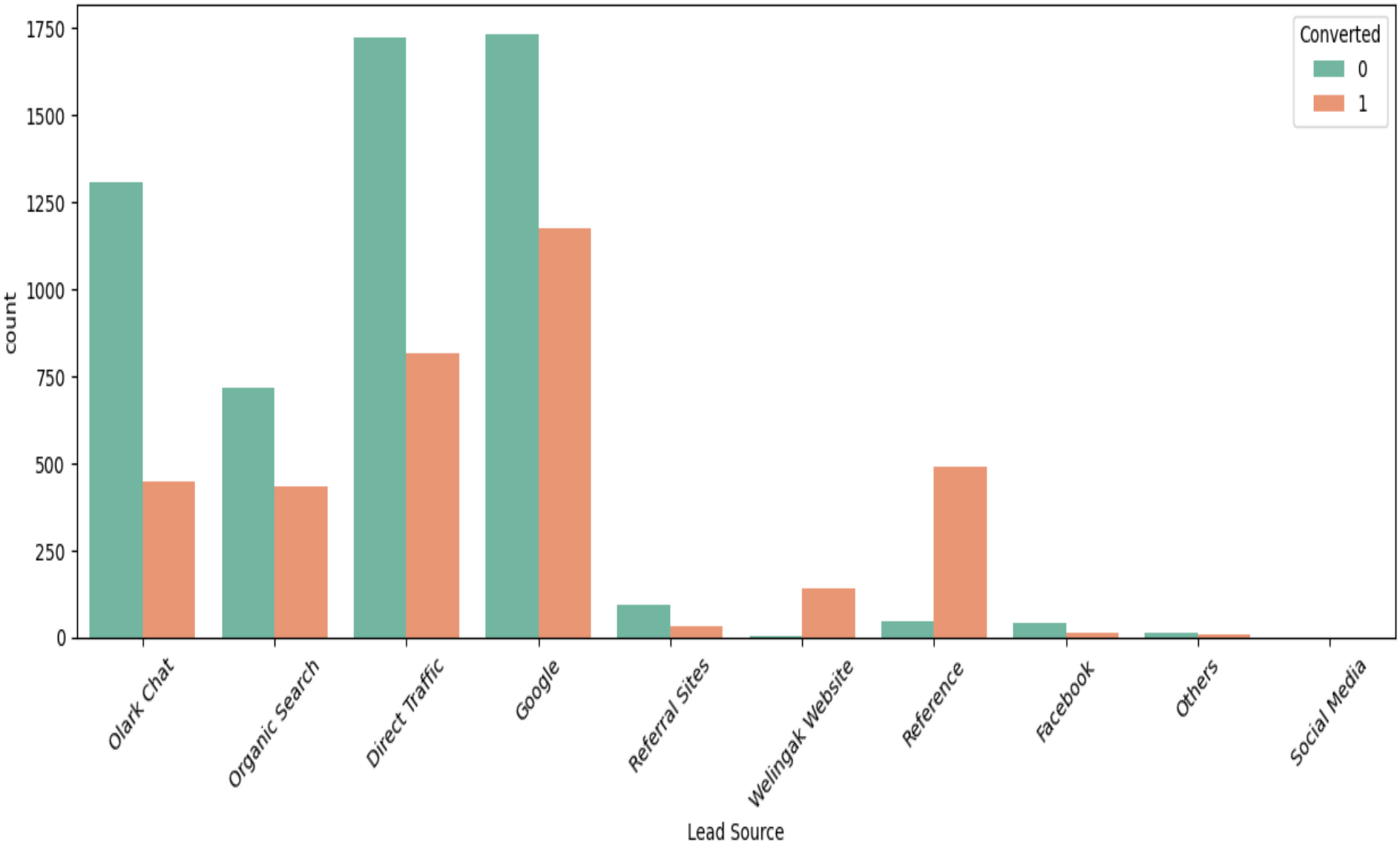
Implementation

- Data Gathering
- Data Cleaning
- Performing EDA
- Data Preparation
- Model Building
- Feature Selection
- Model Improvement
- Final Model
- Verifying with PCA

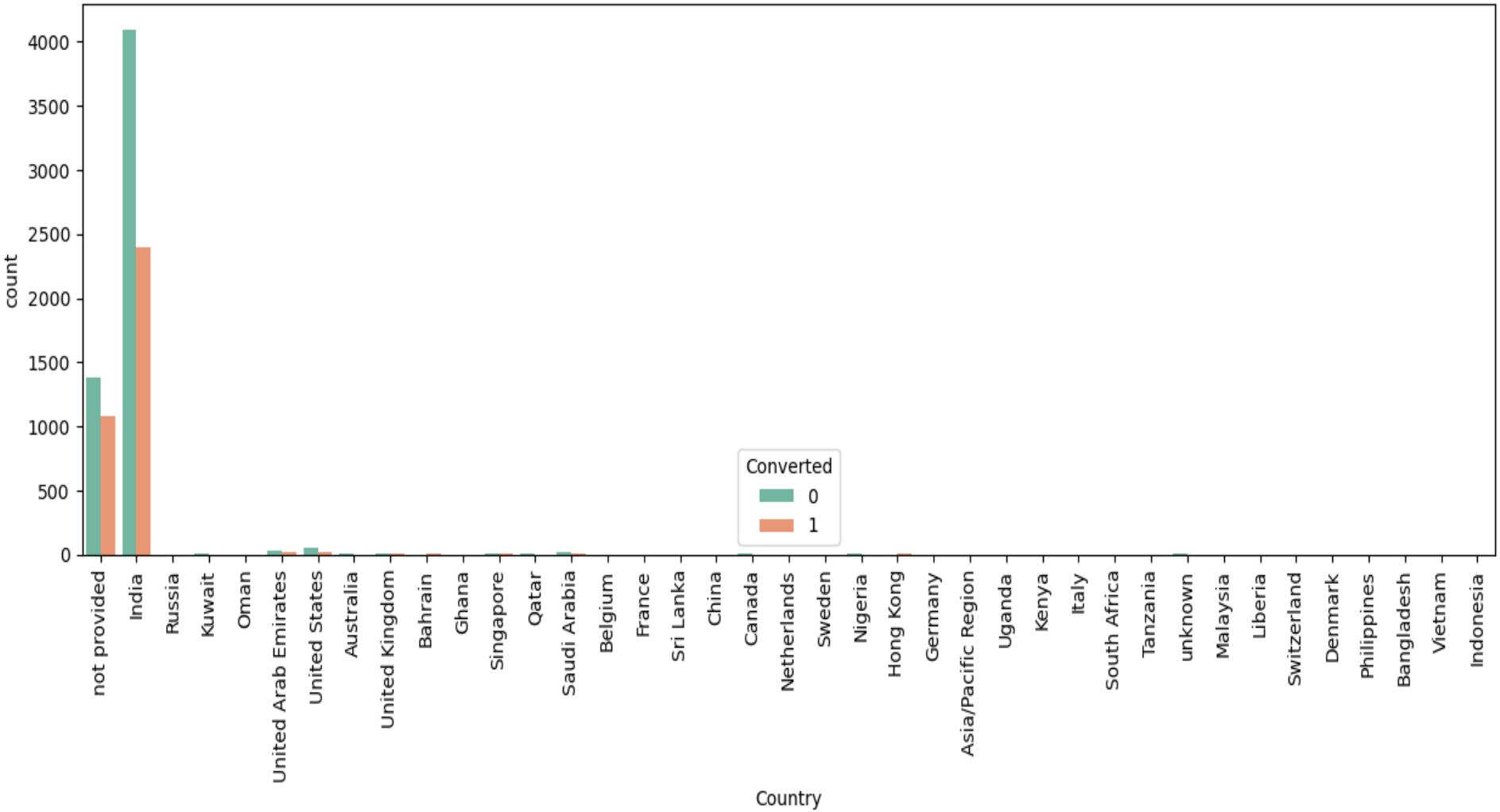
Plots (Visualization)



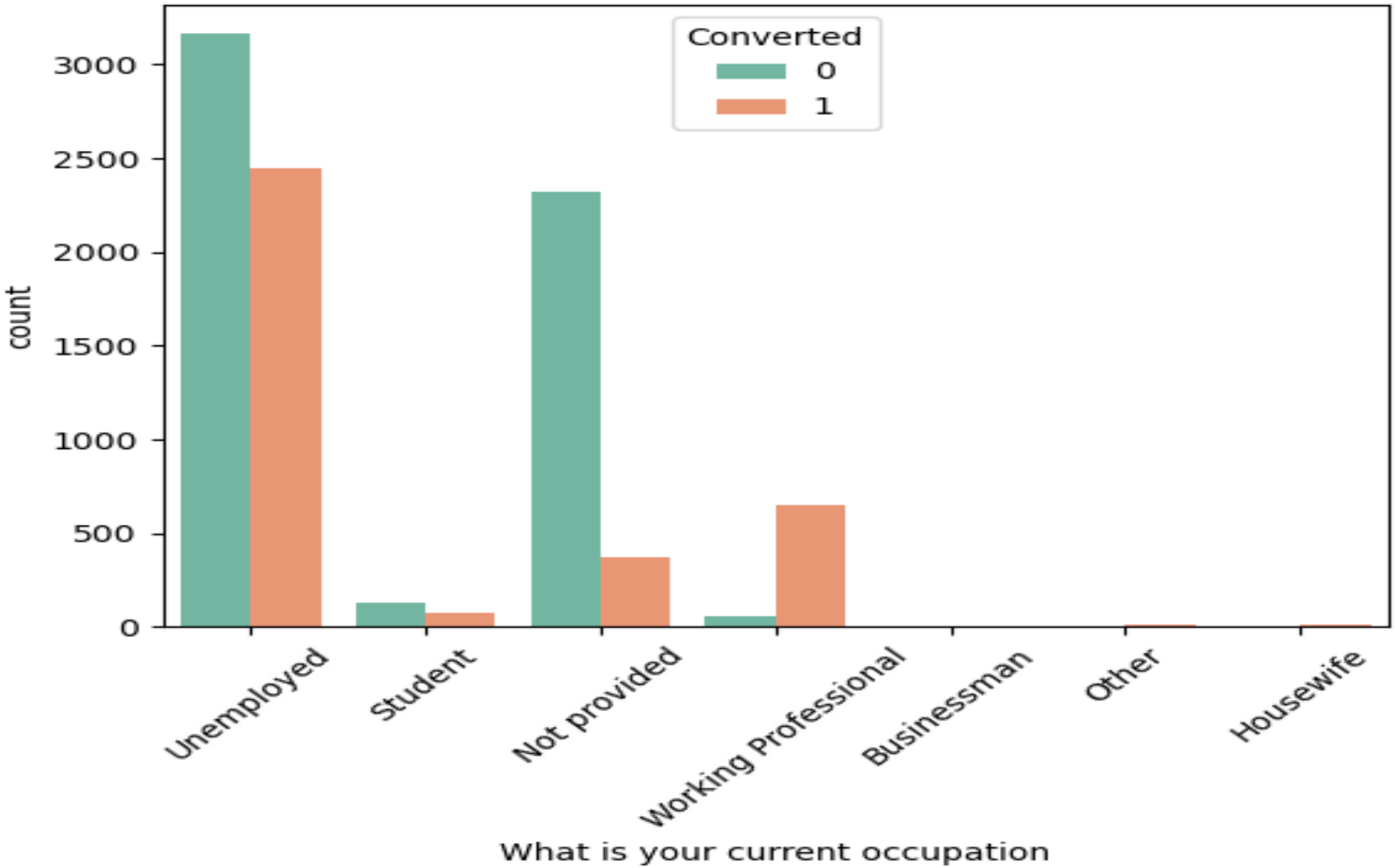
Visualizing variables for imbalancing



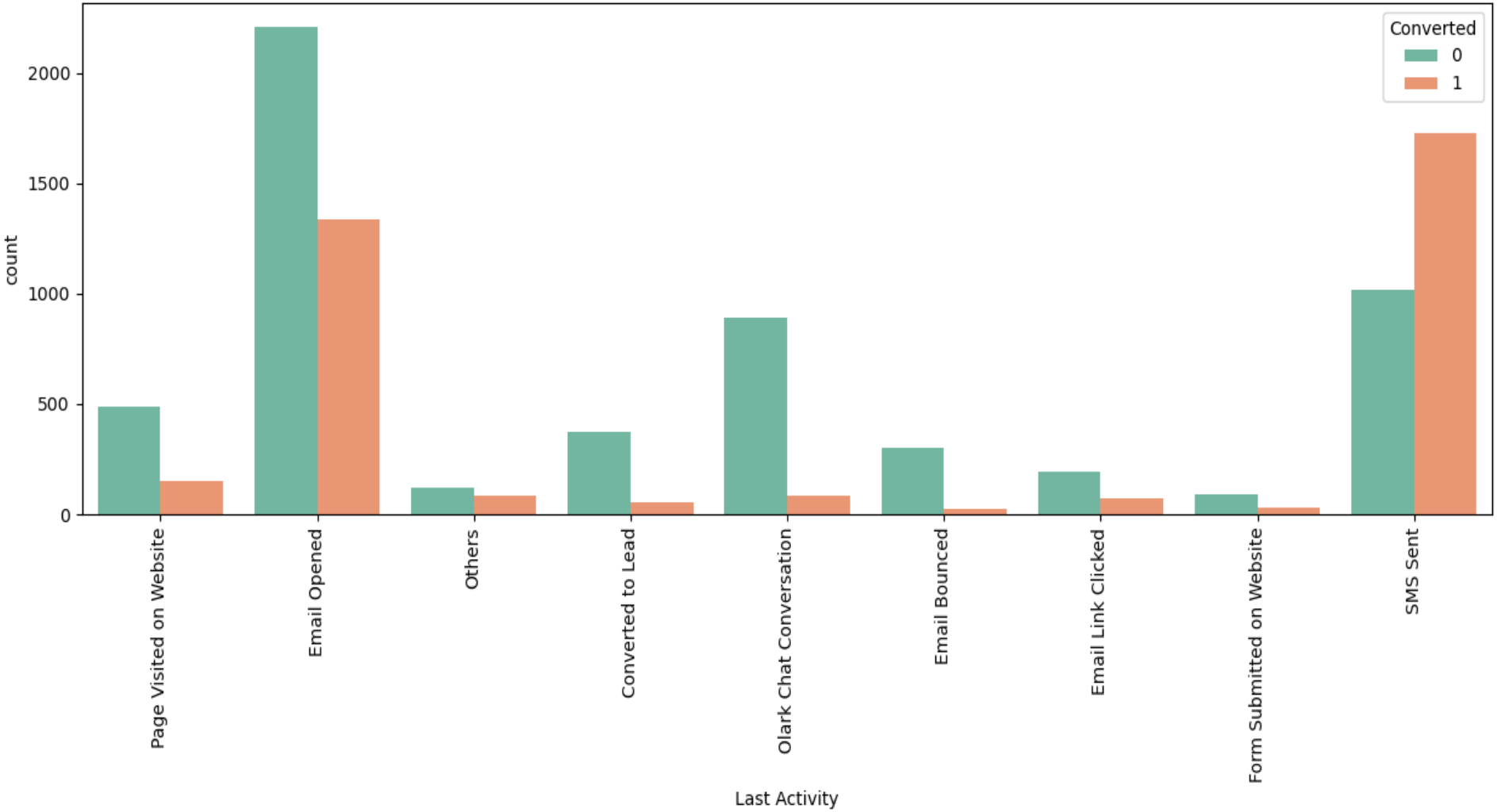
Visualizing count of Lead source variable
based on converted value



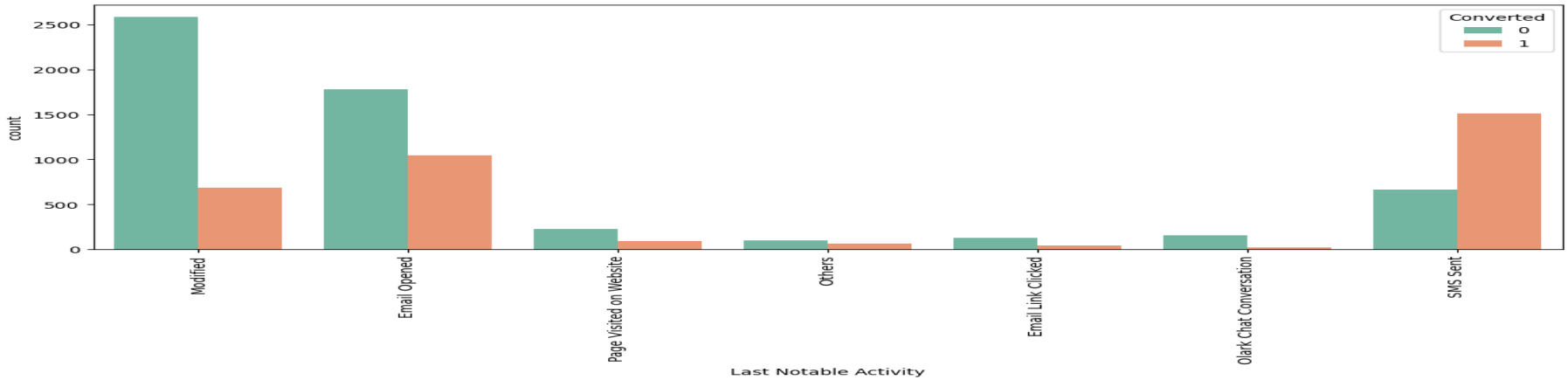
Visualizing country variable after imputation



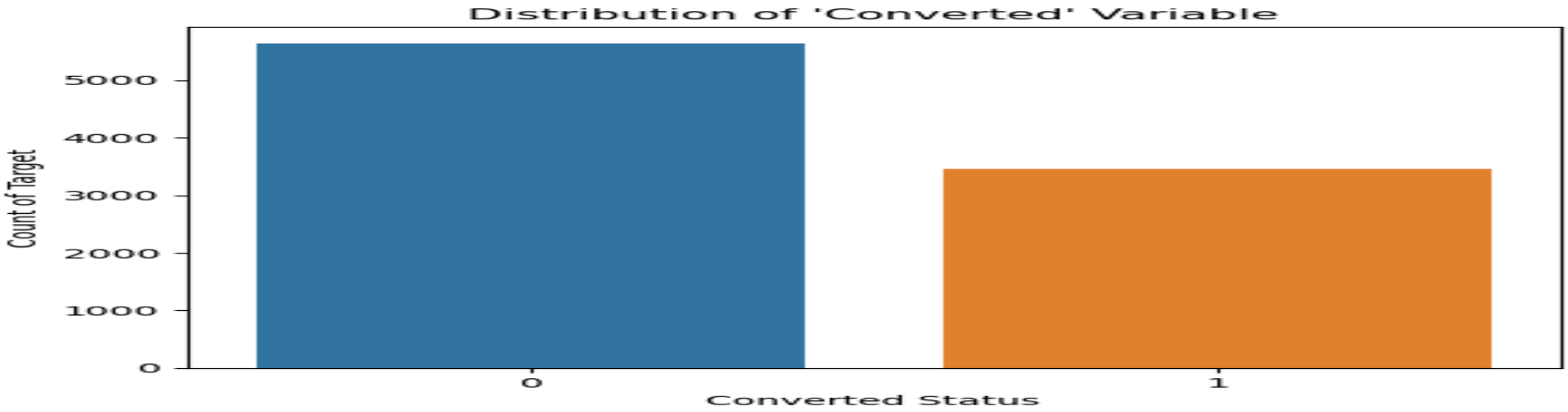
Visualizing count of variable based on converted value



Visualizing count of last activity variable



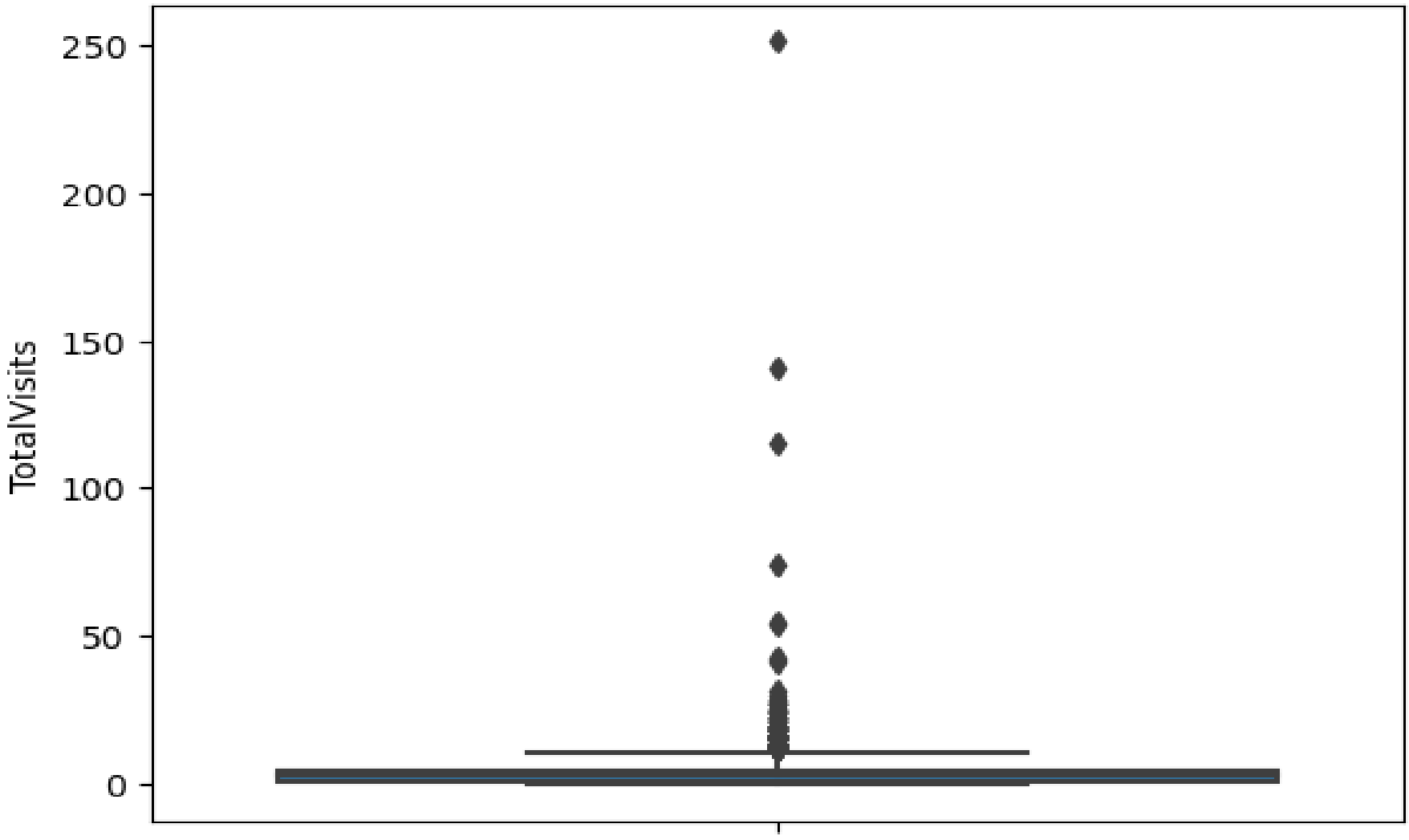
Last notable activity



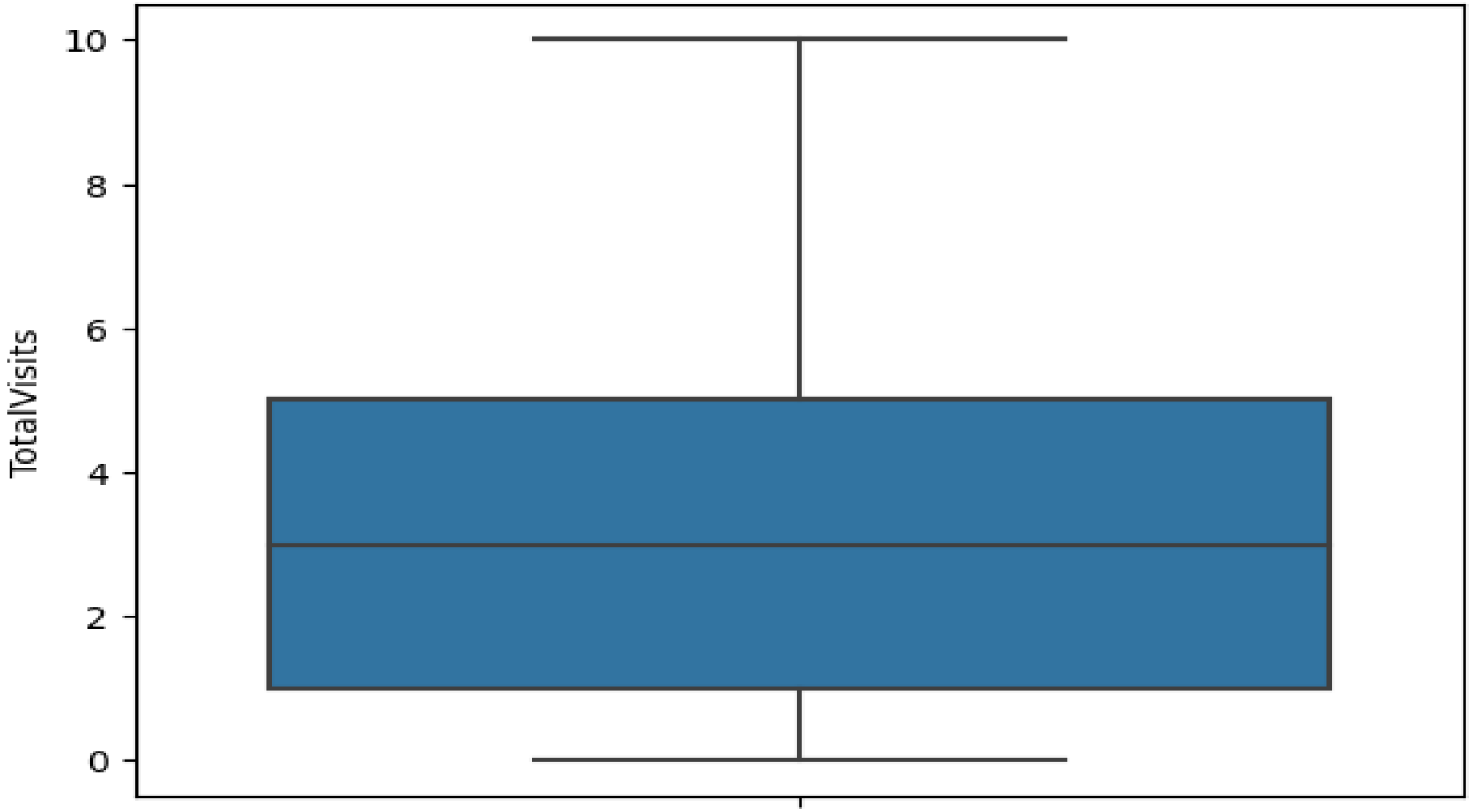
Distribution of converted value


Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.03
0.35
-0.0033
0.03
1
0.22
0.51
0.35
0.22
1
0.32
-0.0033
0.51
0.32
1
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2
0.0

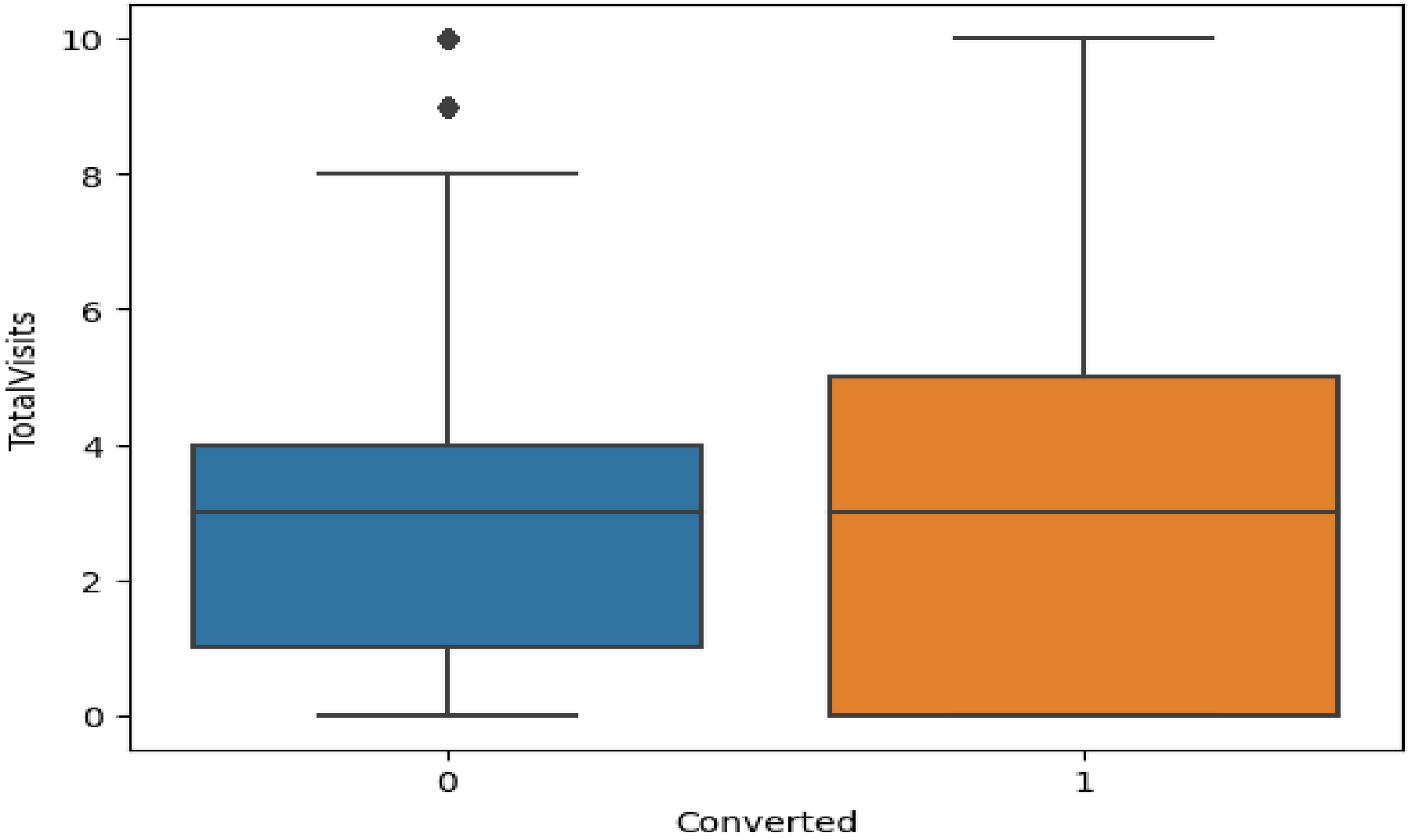
Heat map



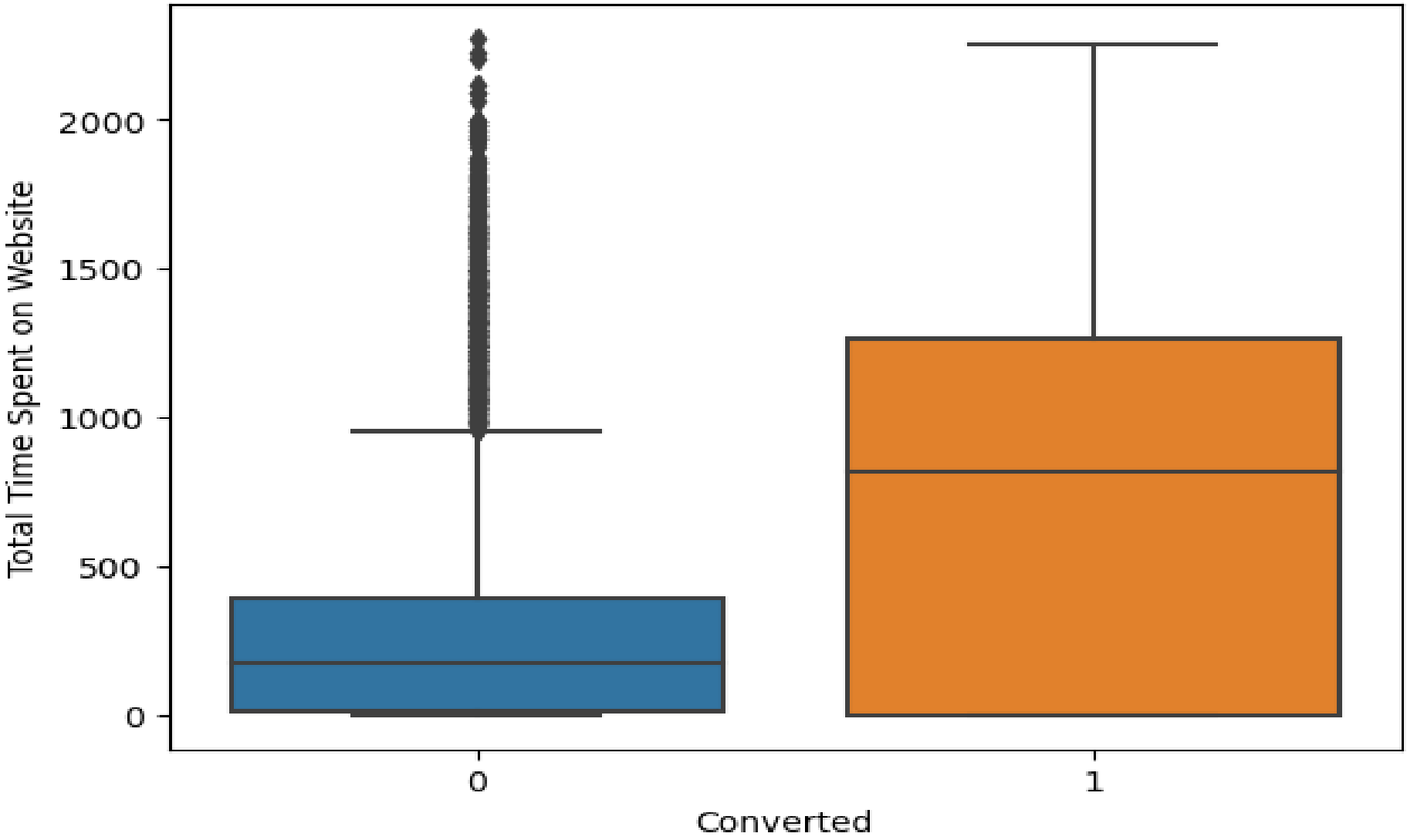
Visualizing spread of variable Total Visits



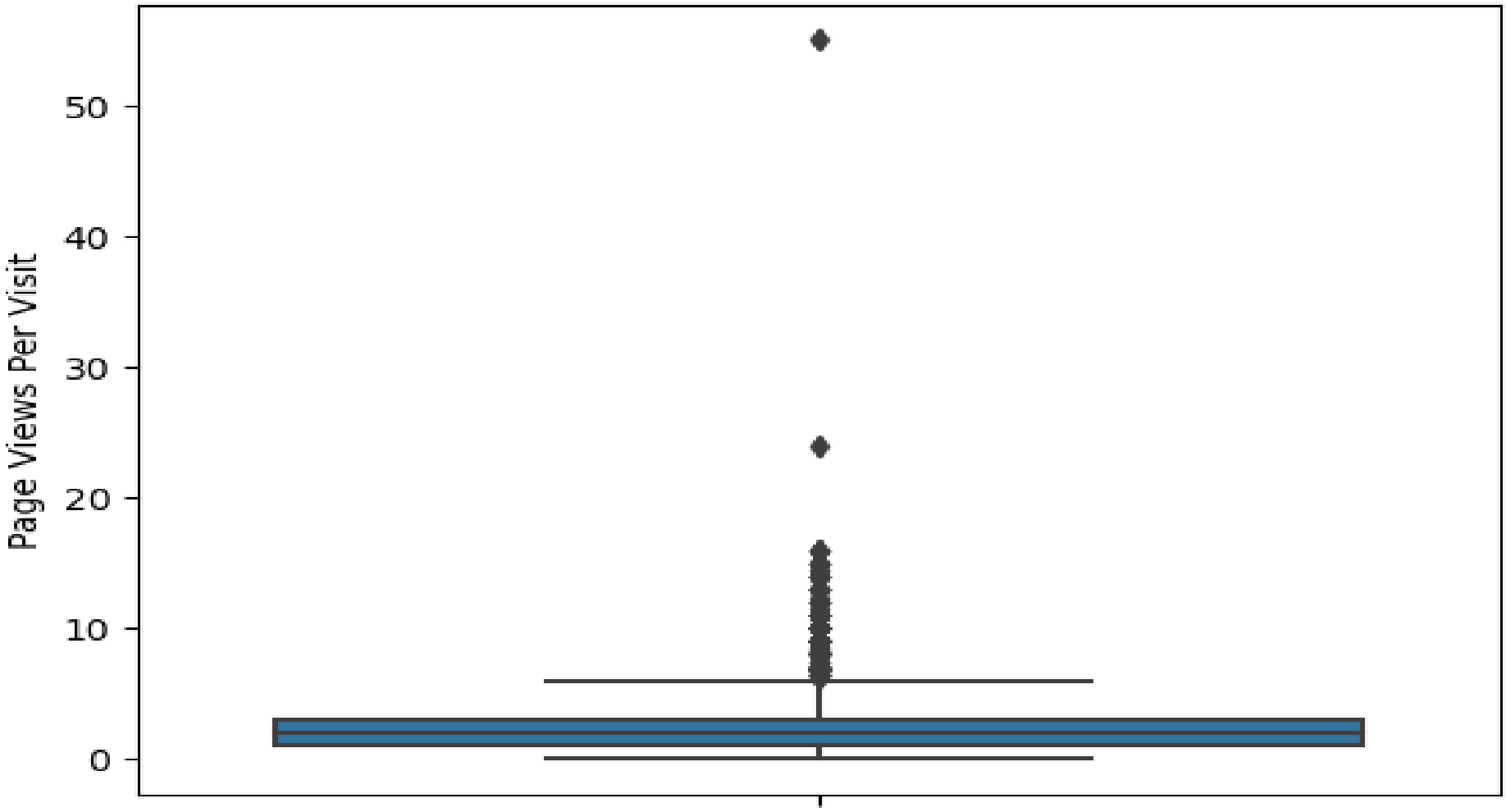
Visualizing Variable After Outlier Treatment



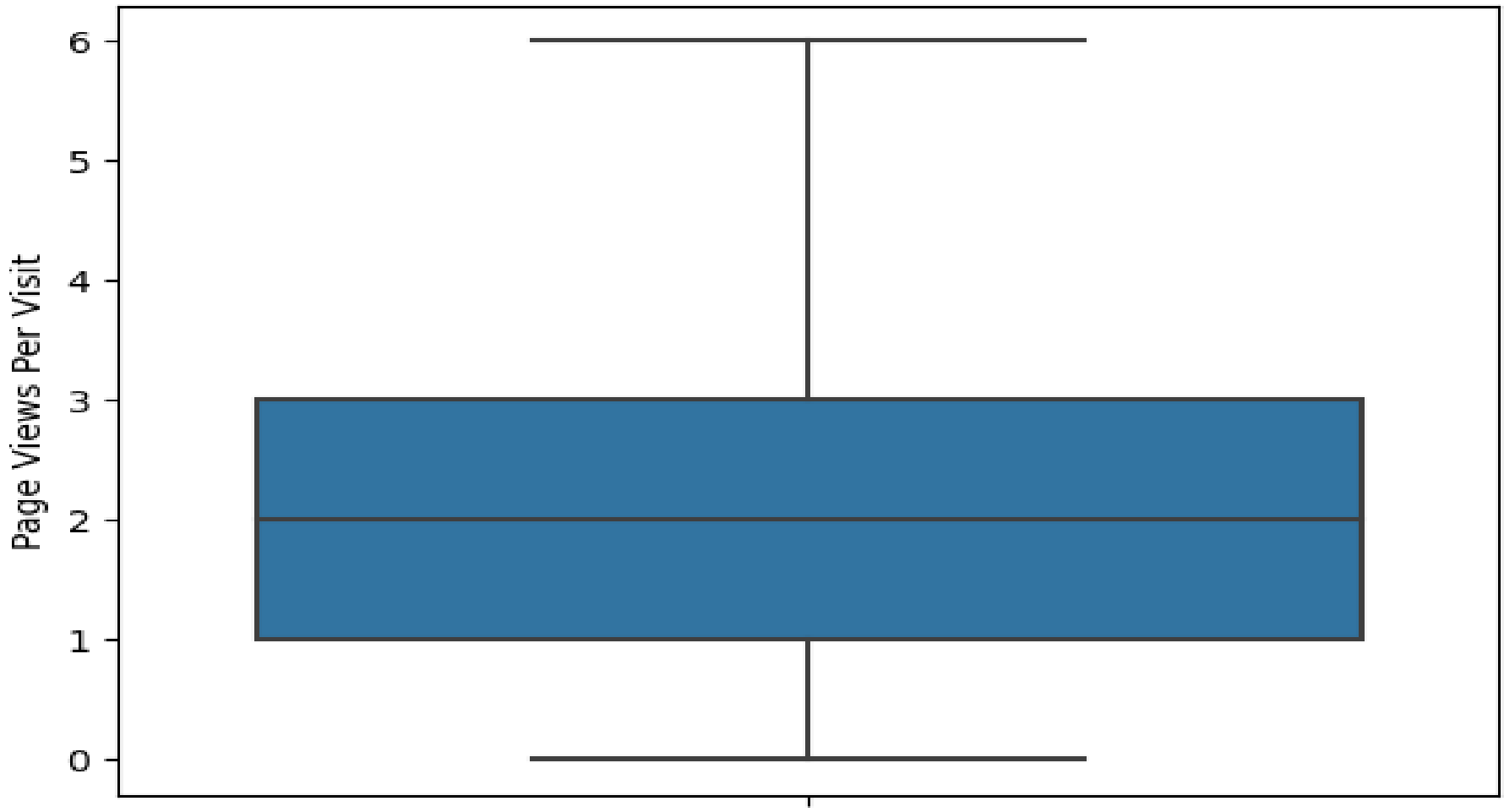
Visualizing Total Visits w..r.t Target Variable 'Converted'



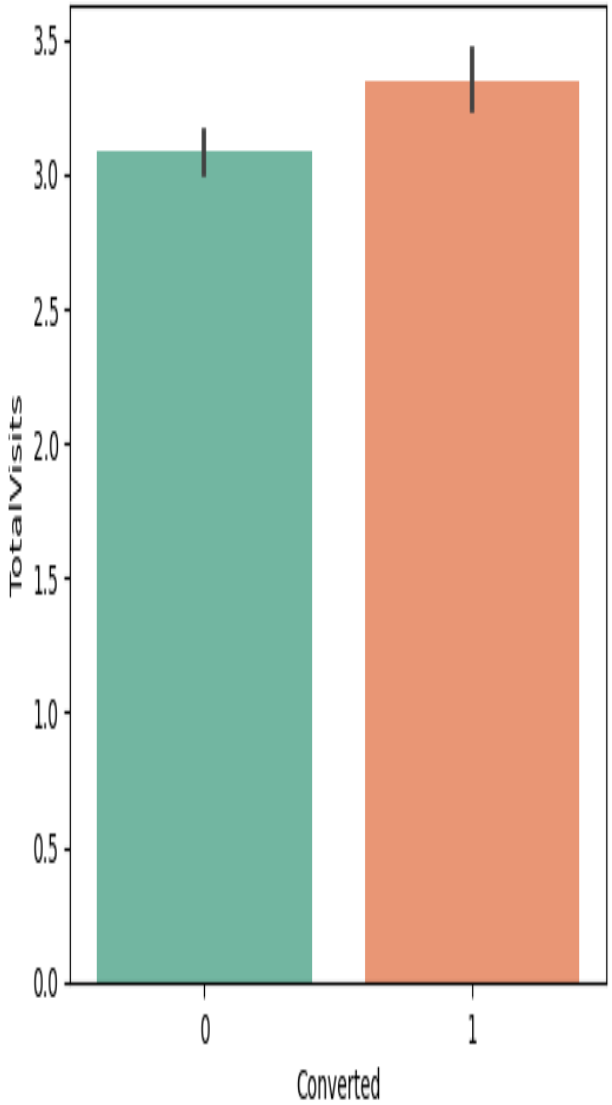
Visualizing spread of variable 'Total Time Spent on Website'

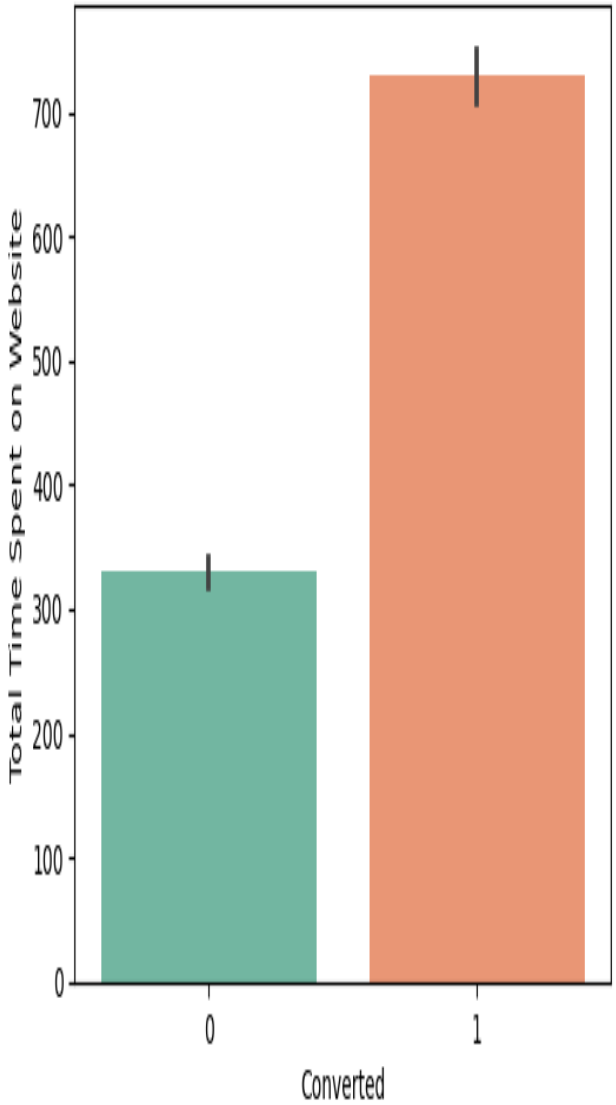


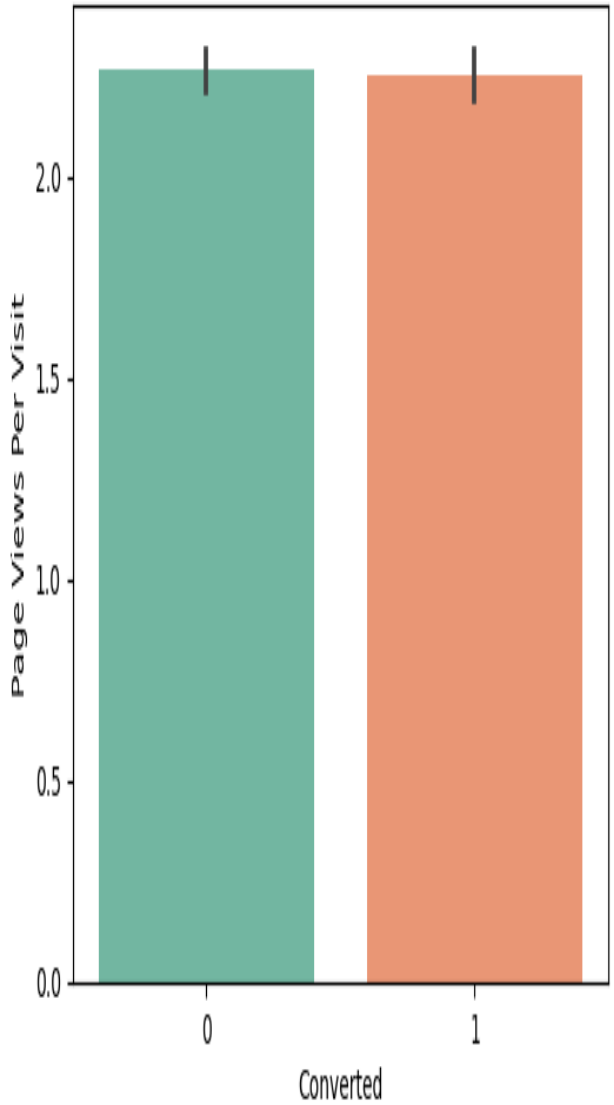
Visualizing variable after outlier treatment



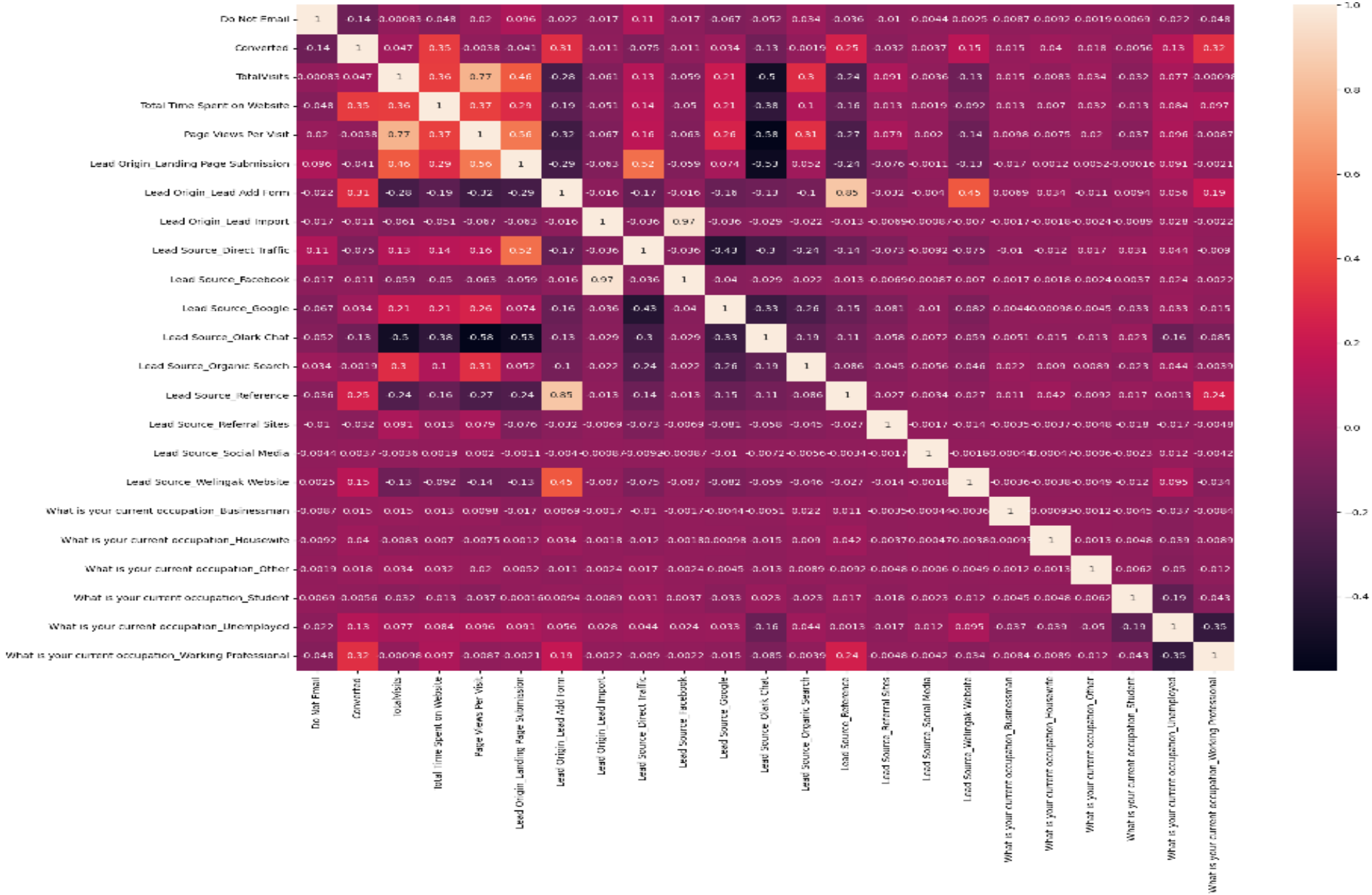
Visualizing 'Page Views Per Visit'
w.r.t Target variable 'Converted'



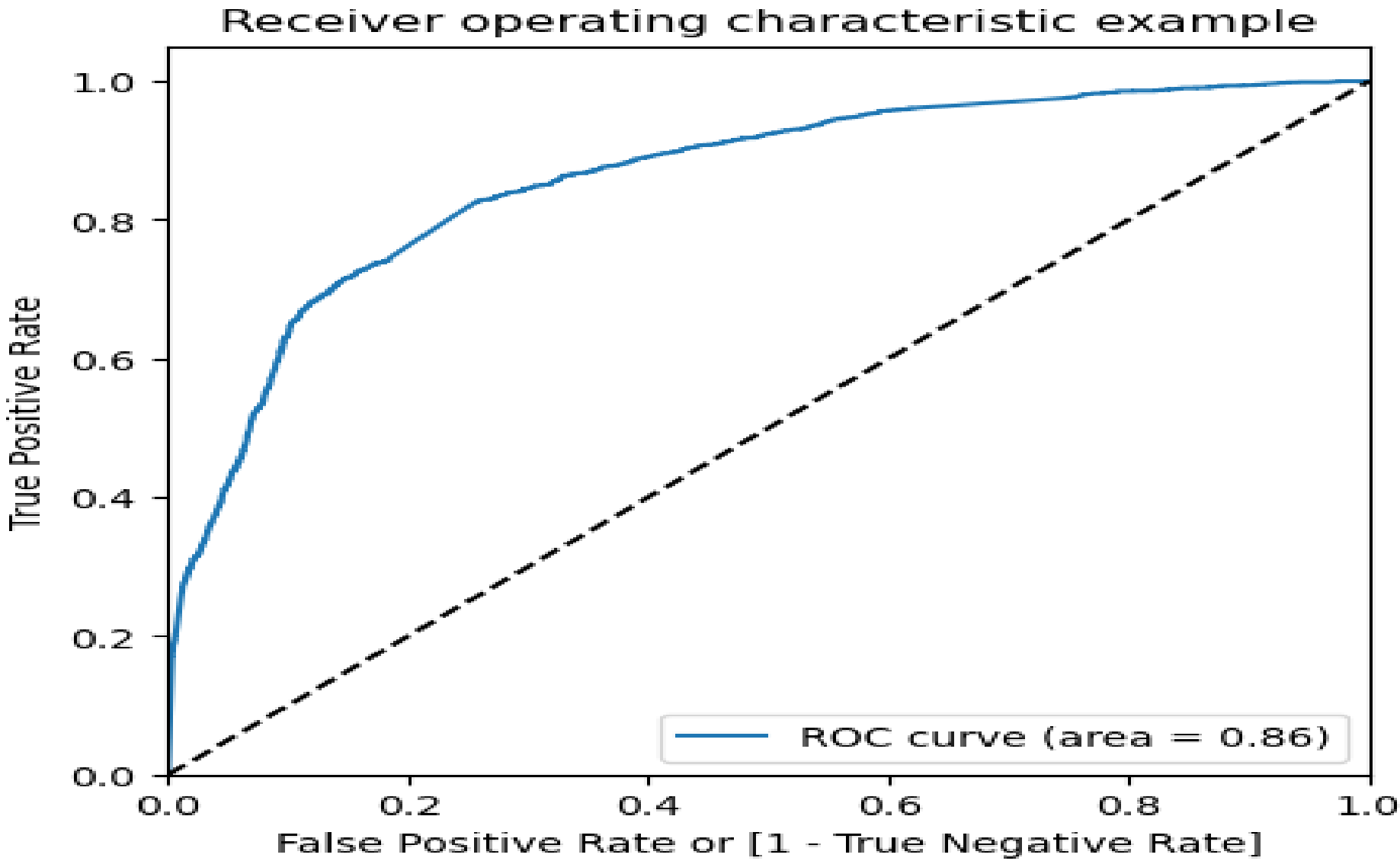




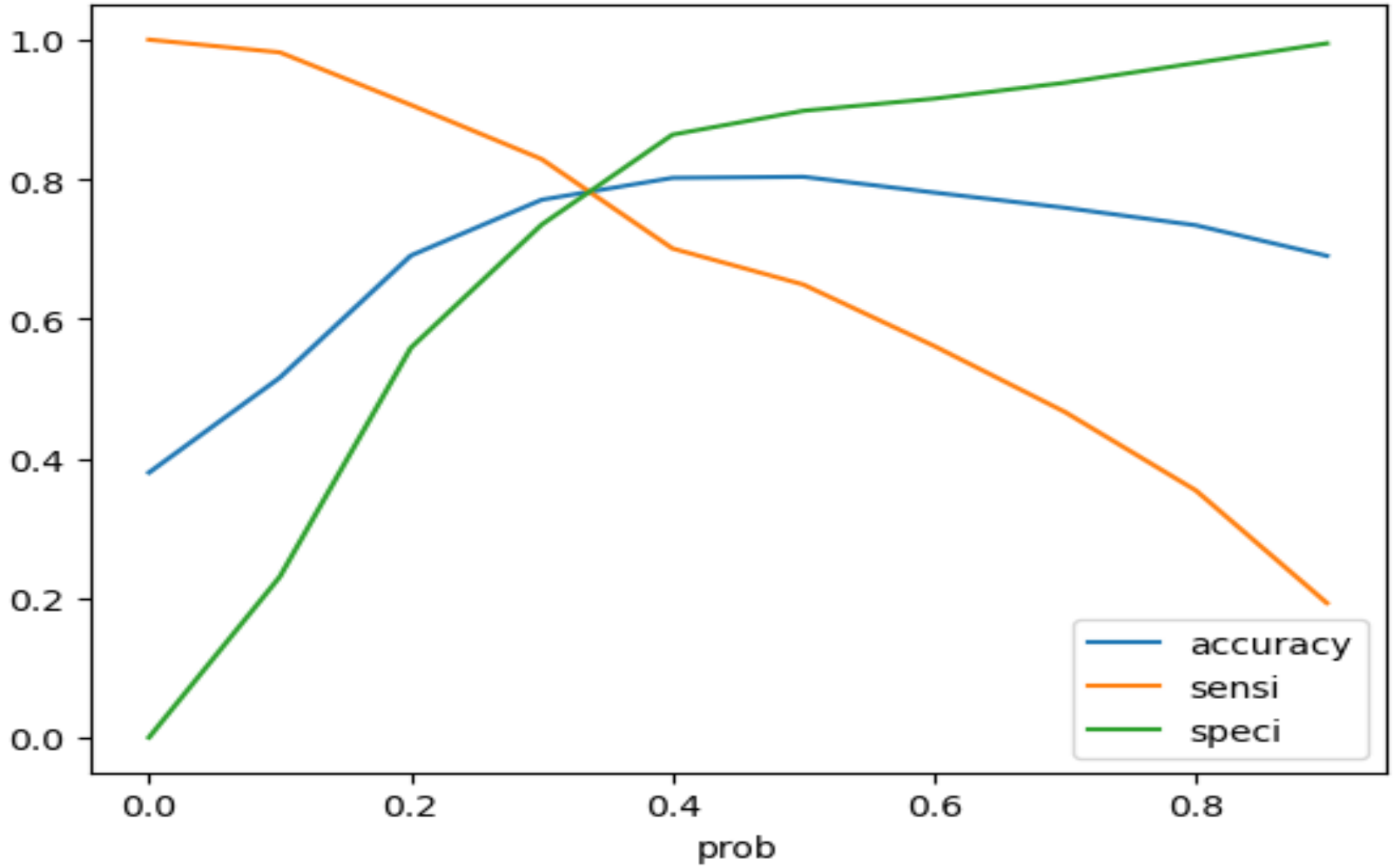
Conversions for all numeric values



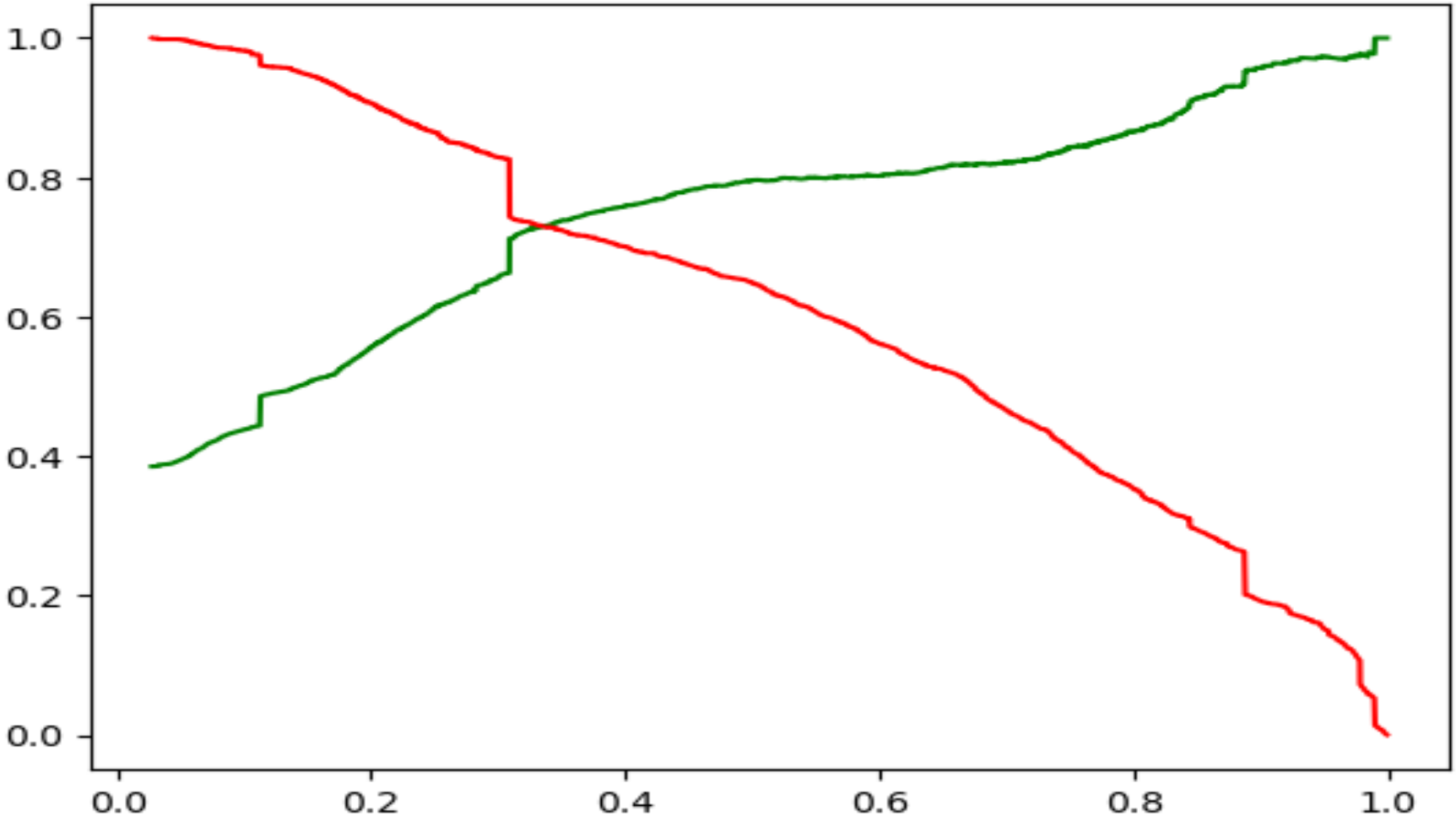
Correlation Matrix


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.86)
0.0
0.2
0.4
0.6
0.8
1.0

ROC Curve



Accuracy sensitivity and specificity for various probabilities.



Precision Recall Curve

Model Analysis

Overall accuracy on Test set: 77.52%

Performance of our Final Model

Sensitivity of our logistic regression model: 83.01%

Specificity of our logistic regression model: 74.13%

Inferences from Model

Top 3 variables in model, that contribute towards lead conversion are:

- 1.Lead Origin_Lead Add Form .
- 2.What is your current occupation _ Working Professional .
- 3. Total Time Spent on Website.

Inferences from Model

Top 3 variables in model, that should be focused are:

- Lead Add Form (from Lead Origin)
- Had a Phone Conversation (from Last Notable Activity)
- Working Professional (from What is your current occupation)

Conclusion

❖While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.

❖Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which are approximately closer to the respective values calculated using trained set.

❖Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%

❖Hence overall this model seems to be good.