Leads Scoring Case Study Presentation

By Satvik Praveen, Manika Gupta and Animesh Gupta
DS C46 July Batch

Problem Statement

An education company named X Education needs to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

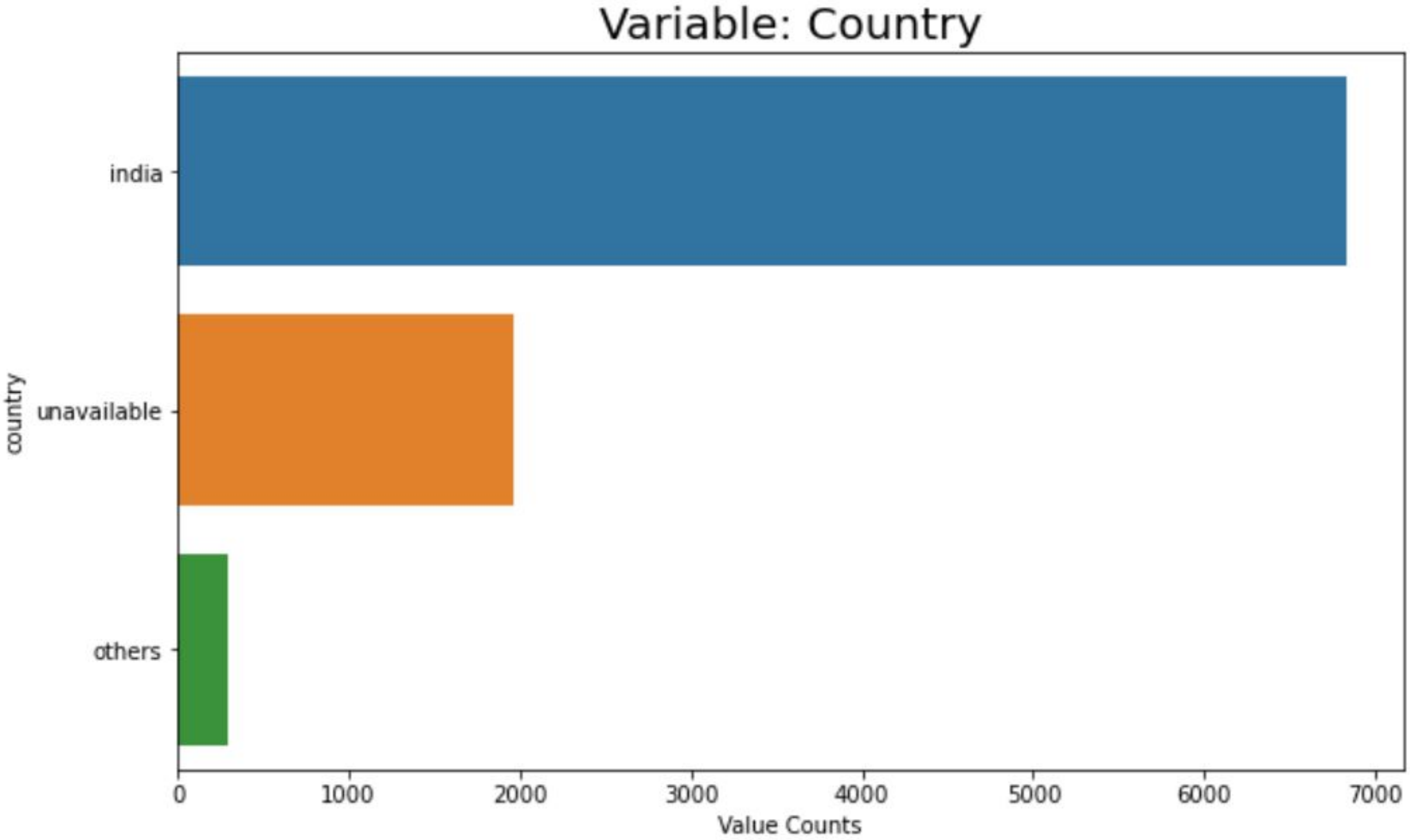
EDA

There were 3 different analyses done:

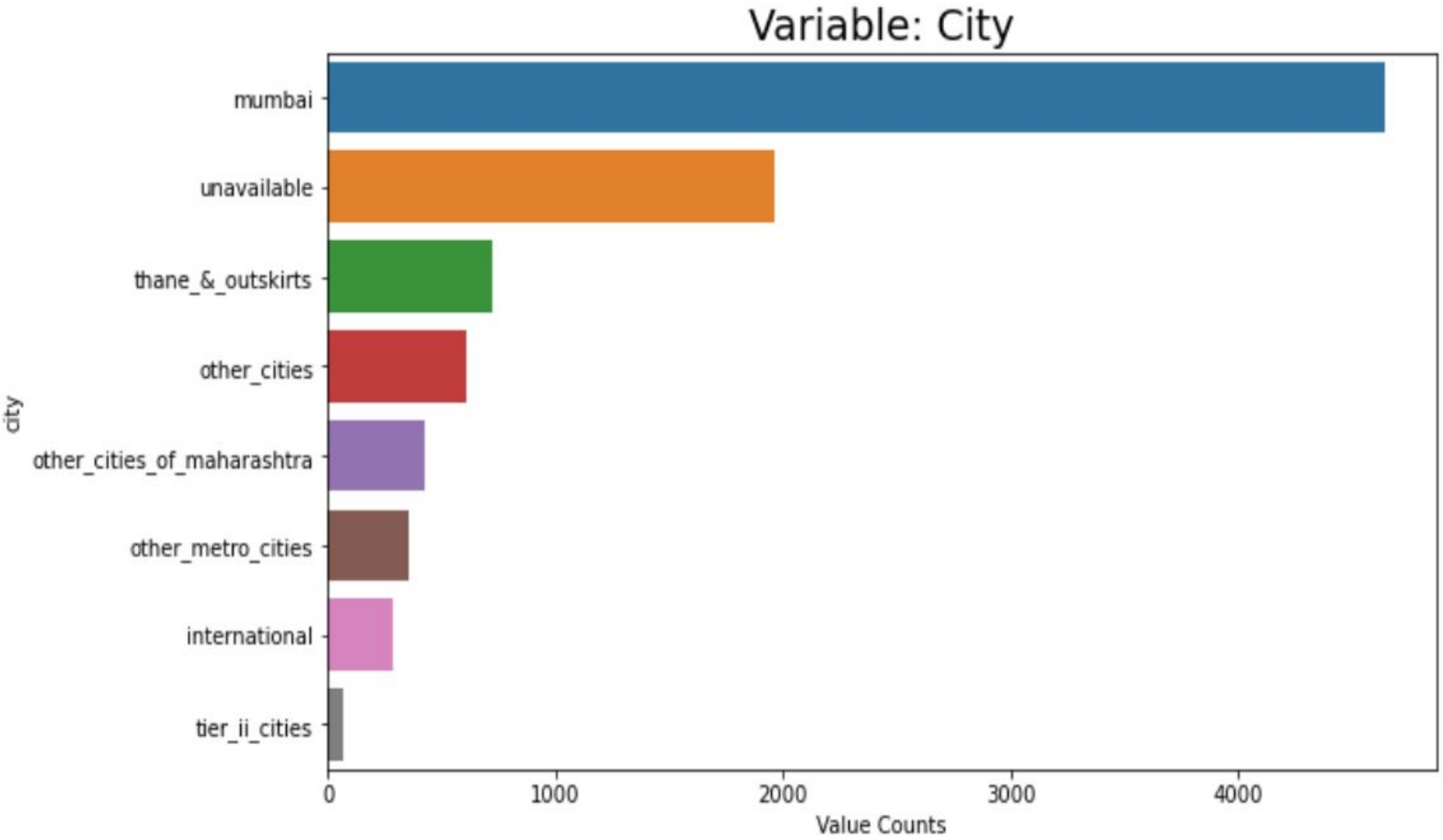
Univariate Analysis: the analysis of a single variable.

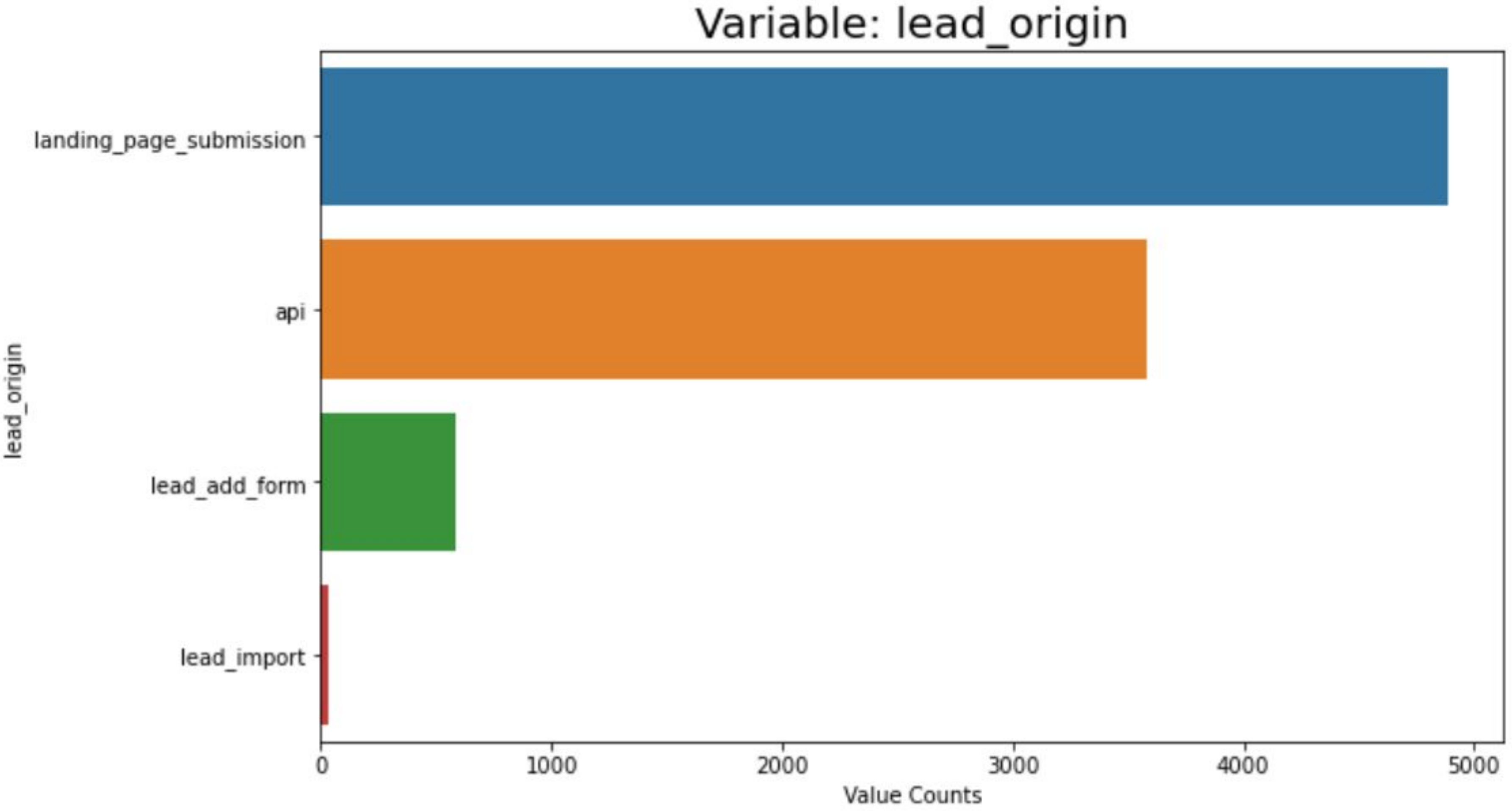
Bivariate Analysis: the analysis of two variables taken together.

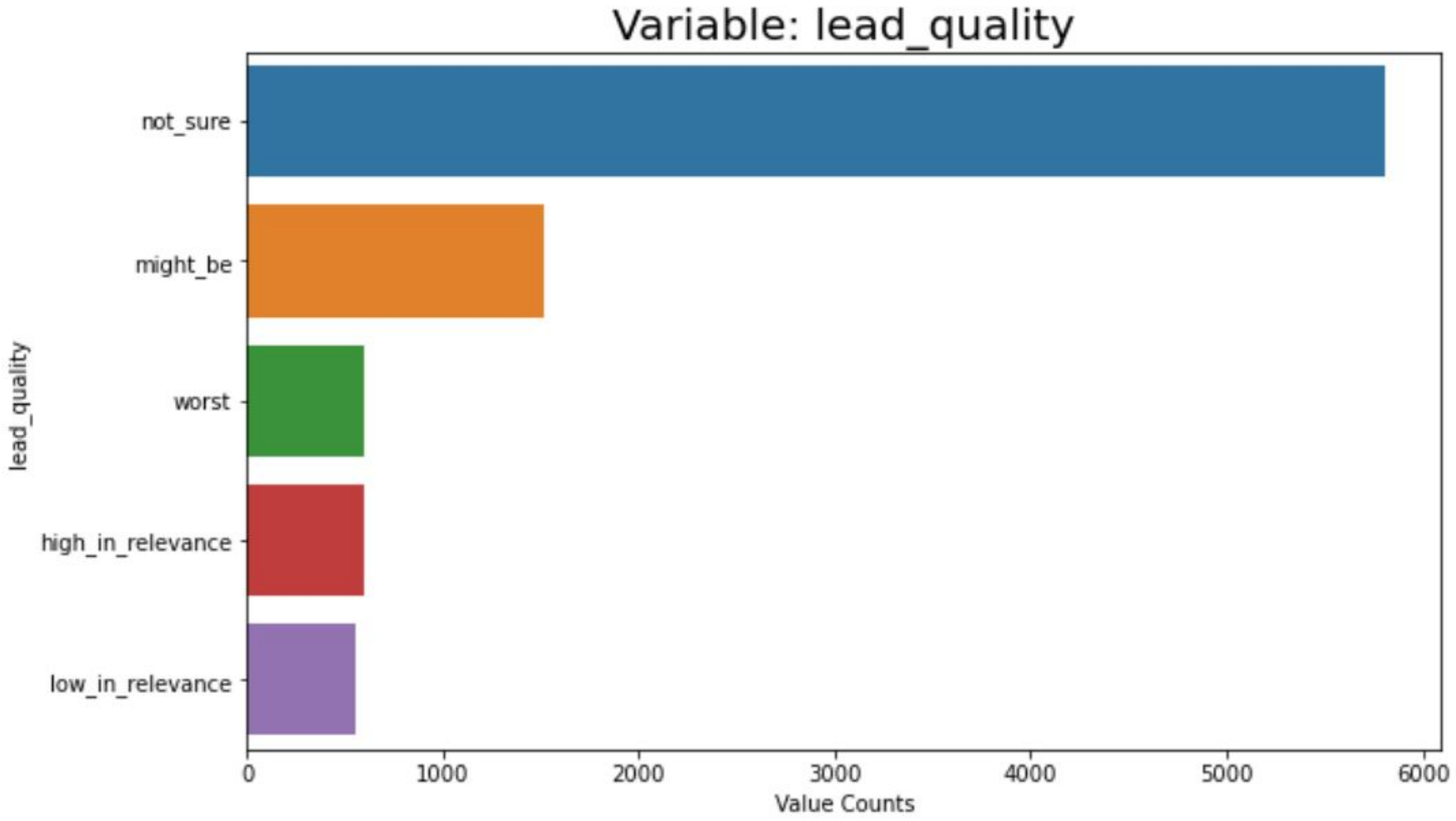
Multivariate Analysis: the analysis of multiple variables taken together.



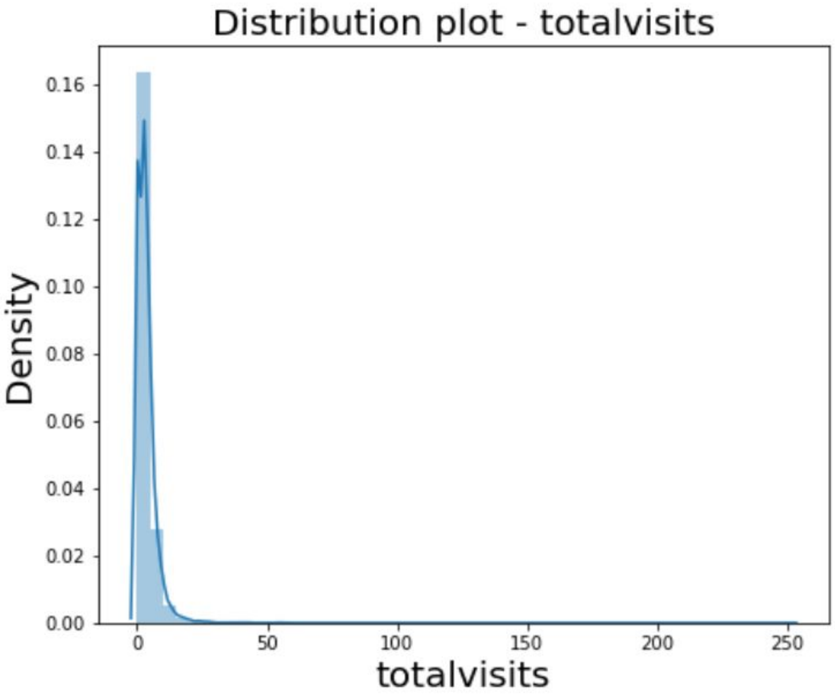
Variable: Country

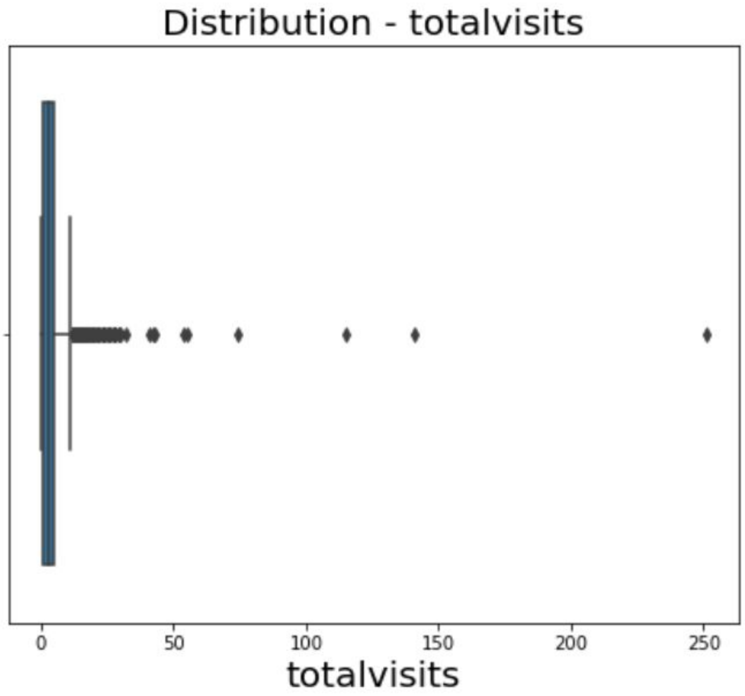

Variable: City
mumbai
unavailable
thane_&_outskirts
other_cities
other_cities_of_maharashtra
other_metro_cities
international
tier_ii_cities
city
0
1000
2000
3000
4000
Value Counts


Variable: lead_origin
lead_origin
landing_page_submission
api
lead_add_form
lead_import
0
1000
2000
3000
4000
5000
Value Counts

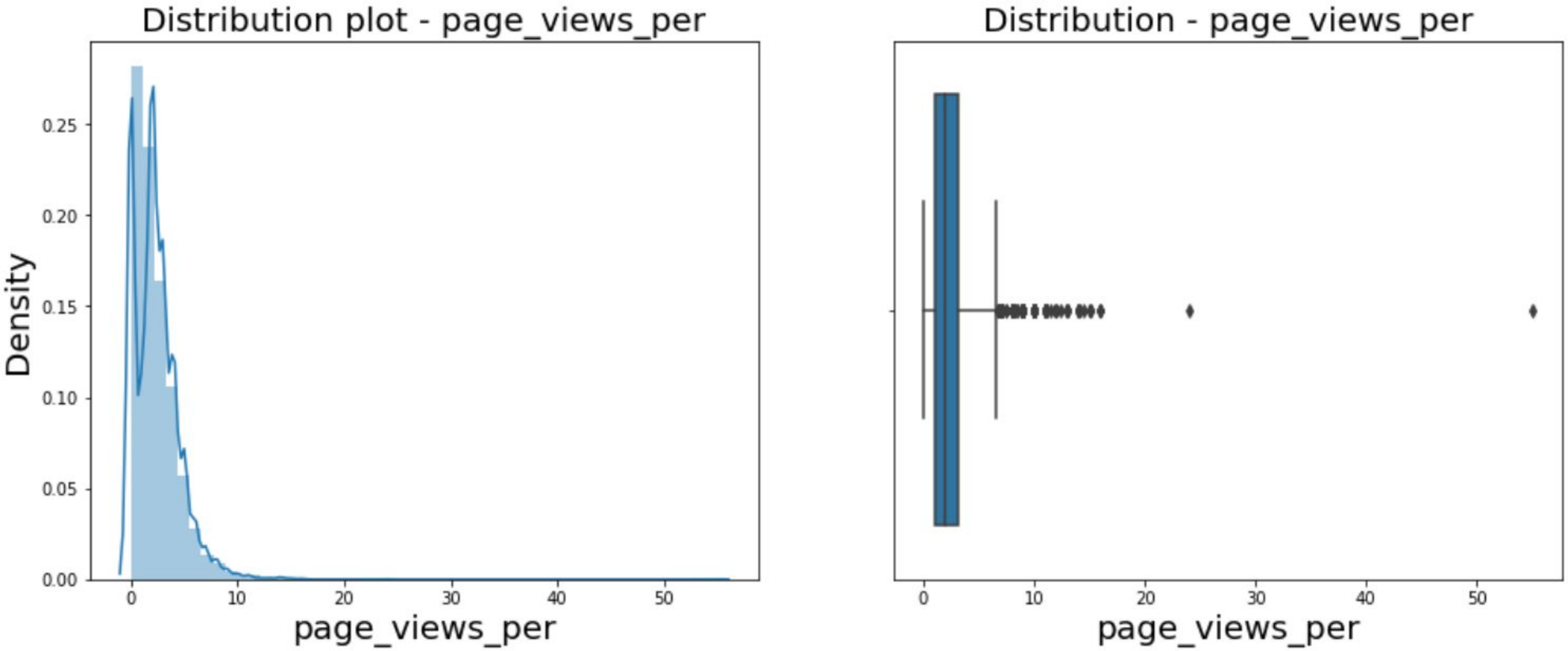

Variable: lead_quality
lead_quality
not_sure
might_be
worst
high_in_relevance
low_in_relevance
0
1000
2000
3000
4000
5000
6000
Value Counts

Numerical Variable: totalvisits

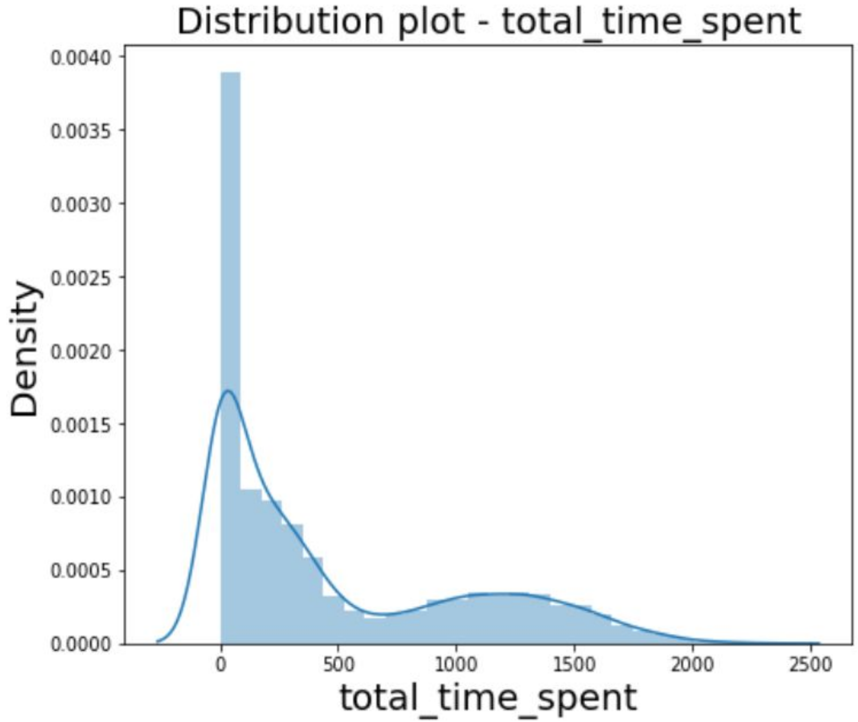




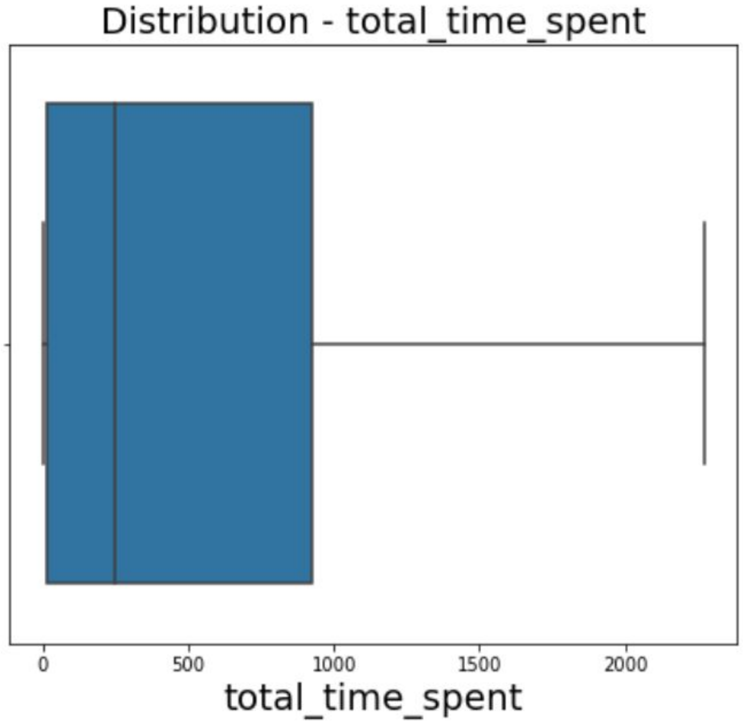
Numerical Variable: page_views_per



Numerical Variable: total_time_spent

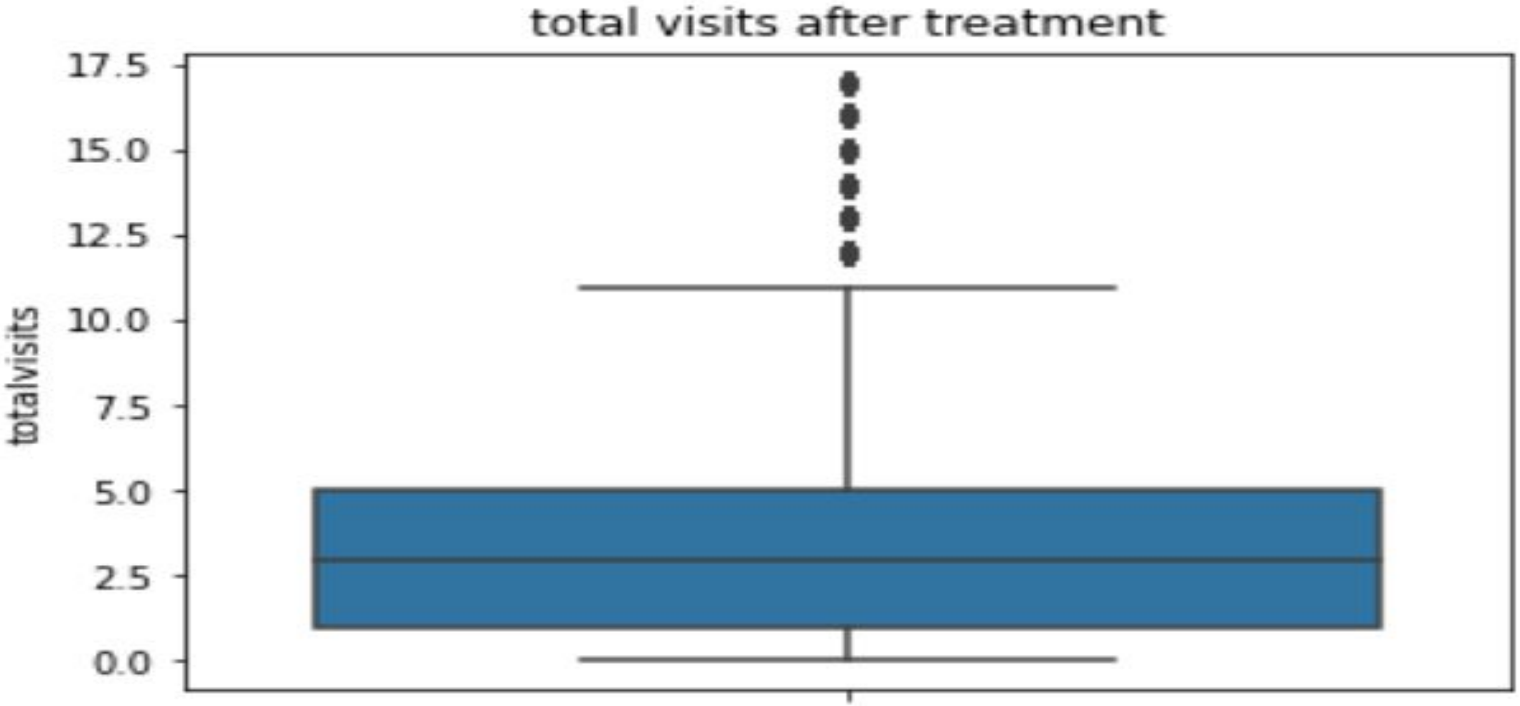


Distribution plot - total_time_spent

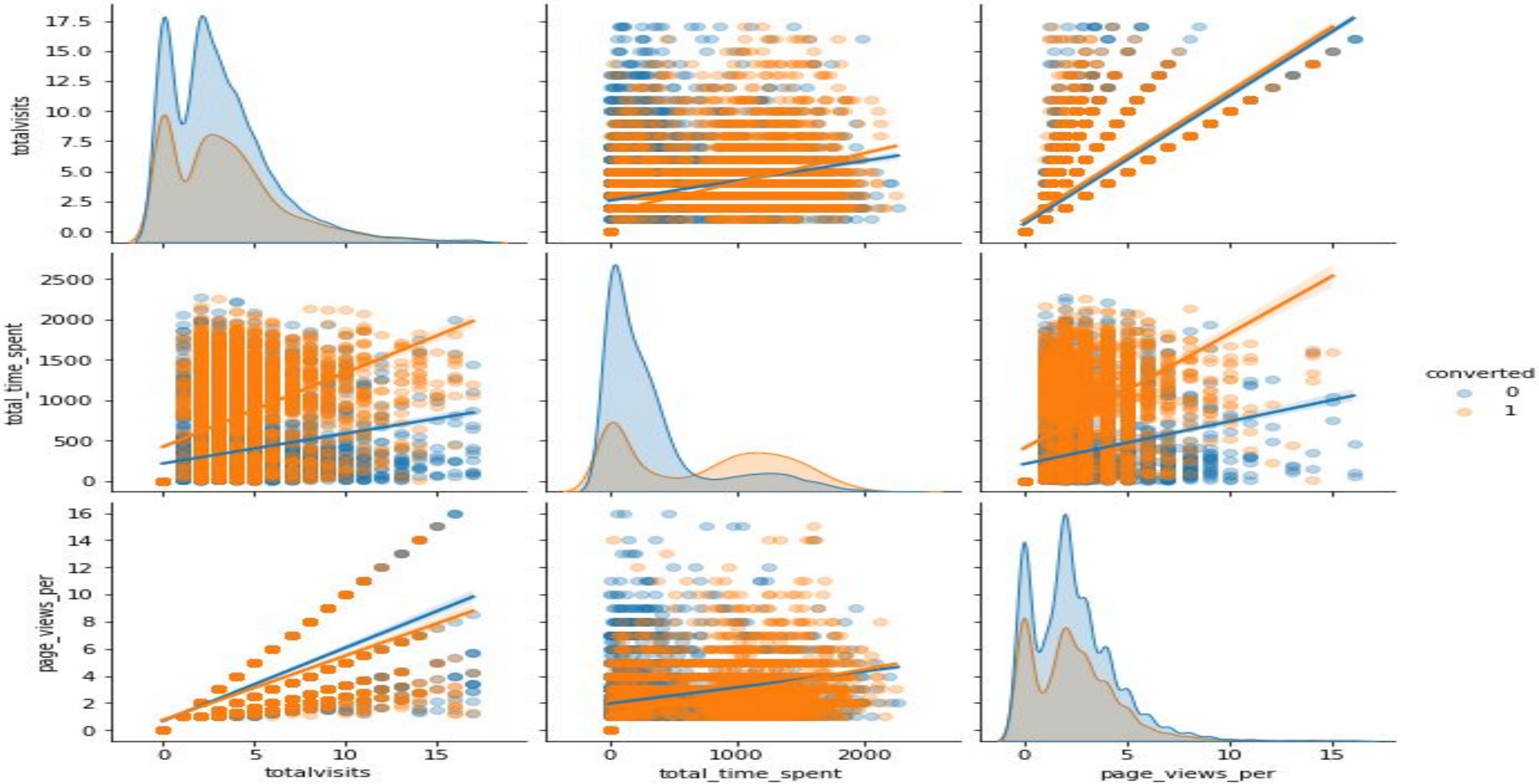


Distribution - total_time_spent

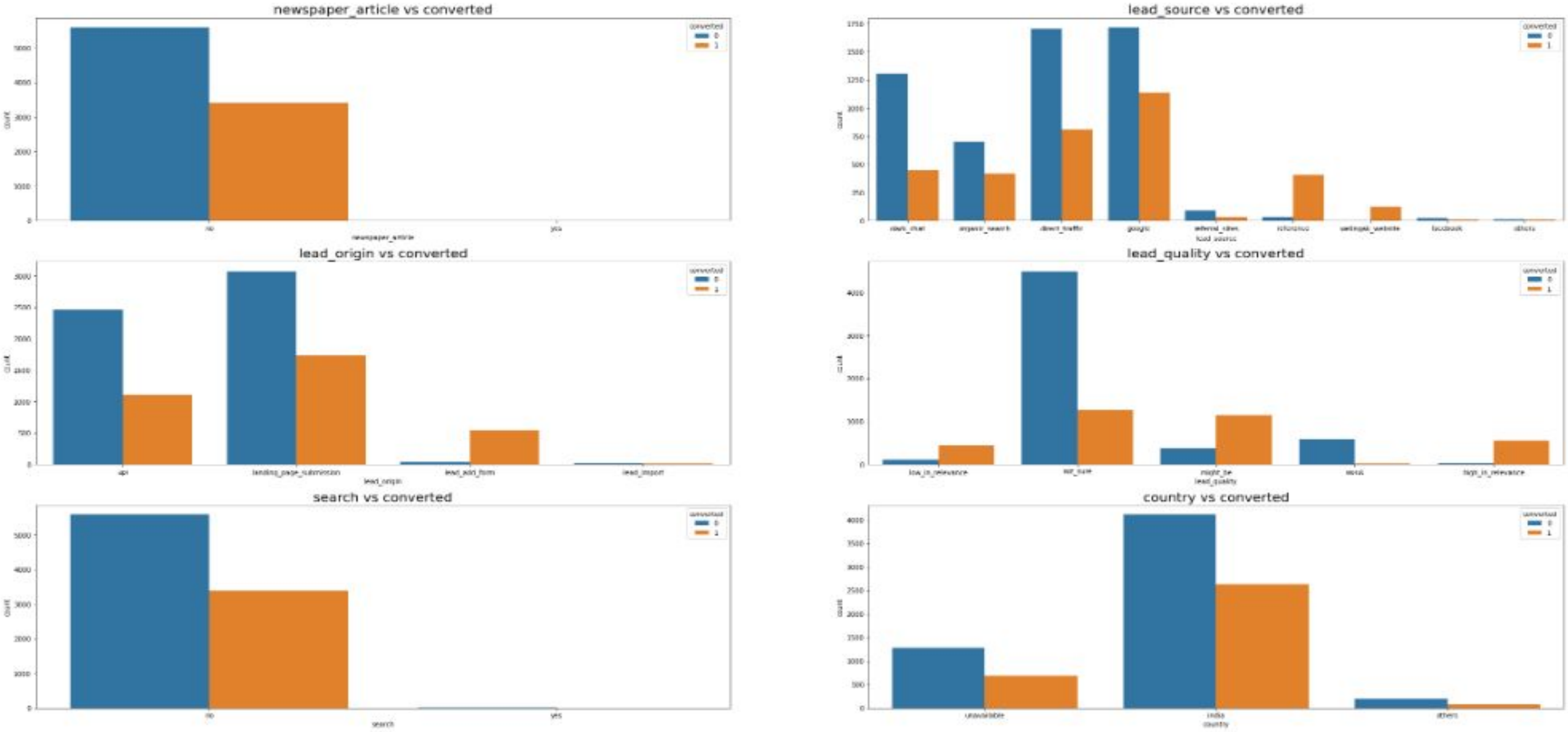
Total visits post treatment



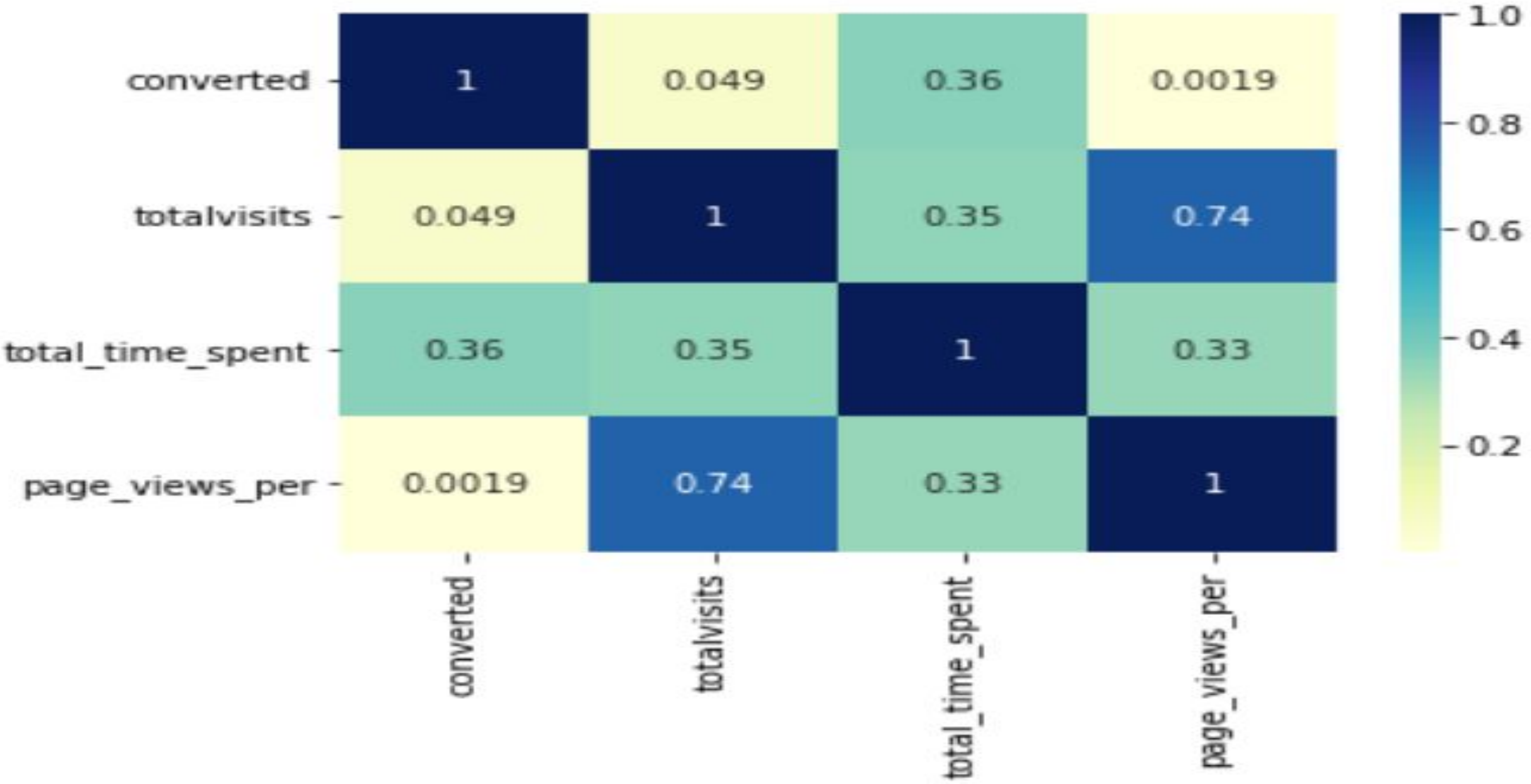
Numerical Variables and Target variable



Categorical Variable and target variable



Multivariate Analysis



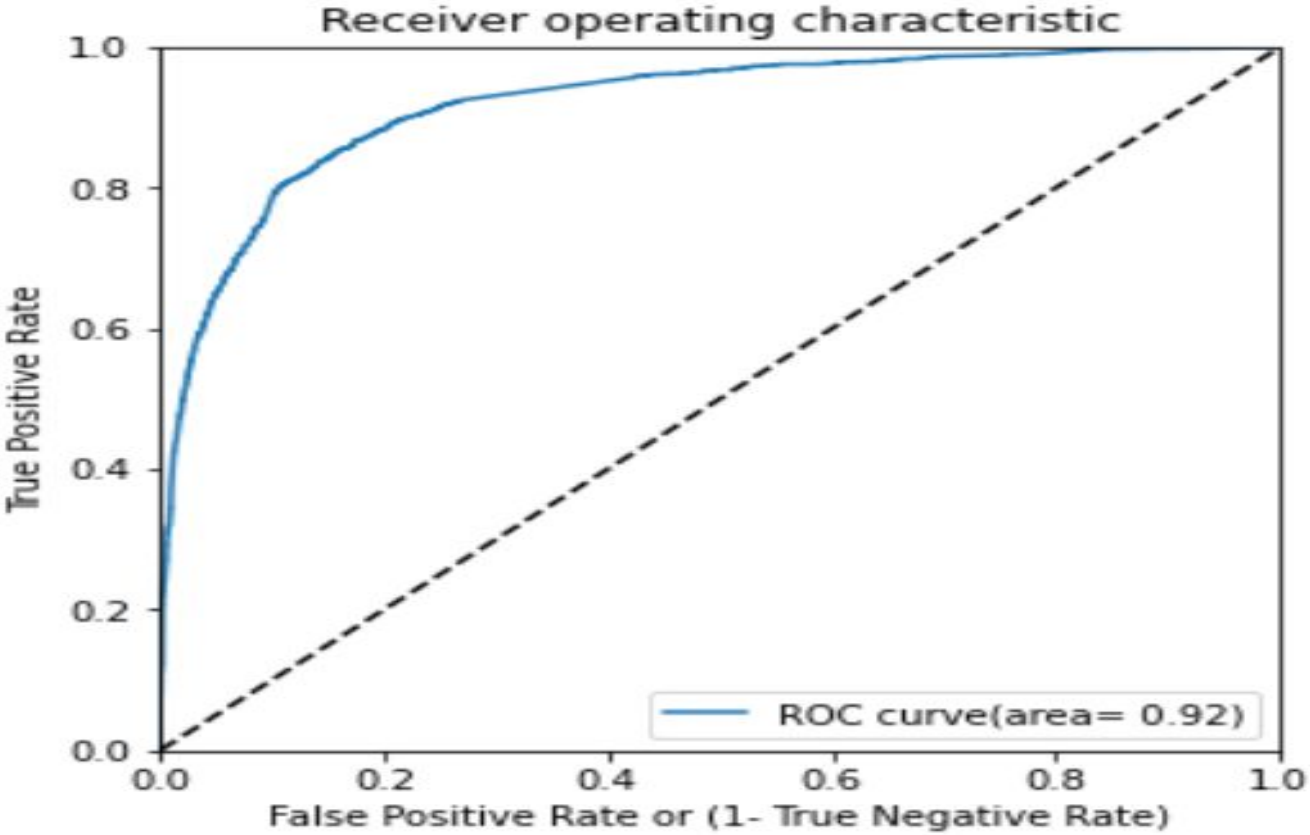
Model Building

Logistic Regression Model building:

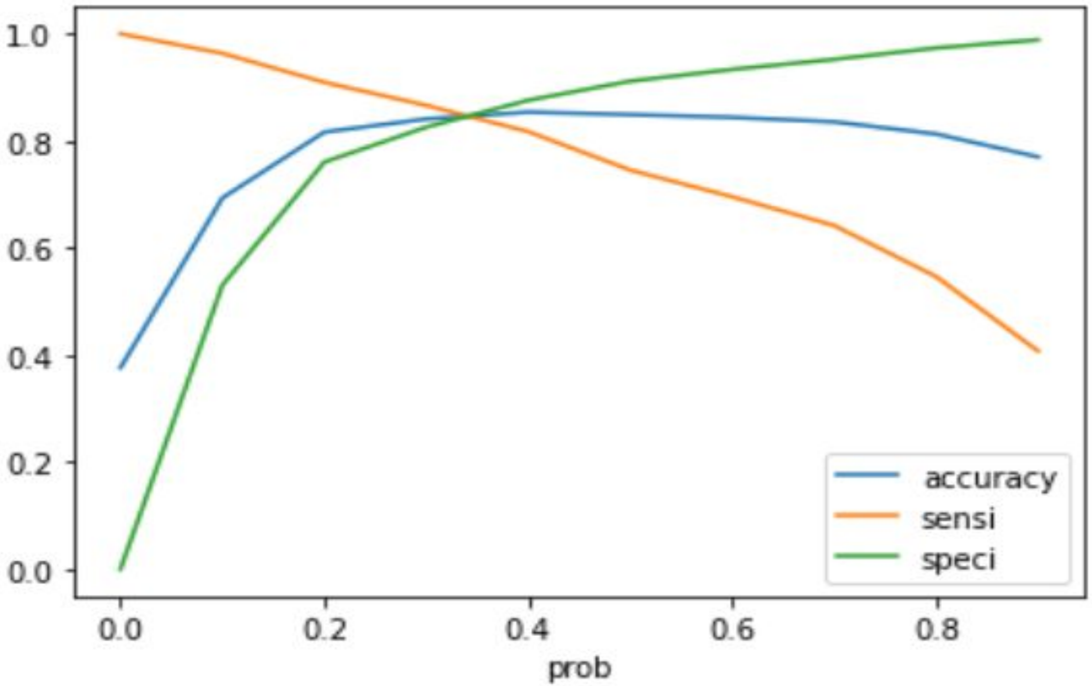
The Feature selection work is done based on Recursive Feature Elimination procedure and then manual selection of variables using the following criteria:

- P-value less than 0.05 for a variable to be kept.
- VIF value less than 5 for a variable to be kept.

ROC Curve


Receiver operating characteristic
True Positive Rate
False Positive Rate or (1- True Negative Rate)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve(area= 0.92)

Plot- accuracy, specificity, sensitivity



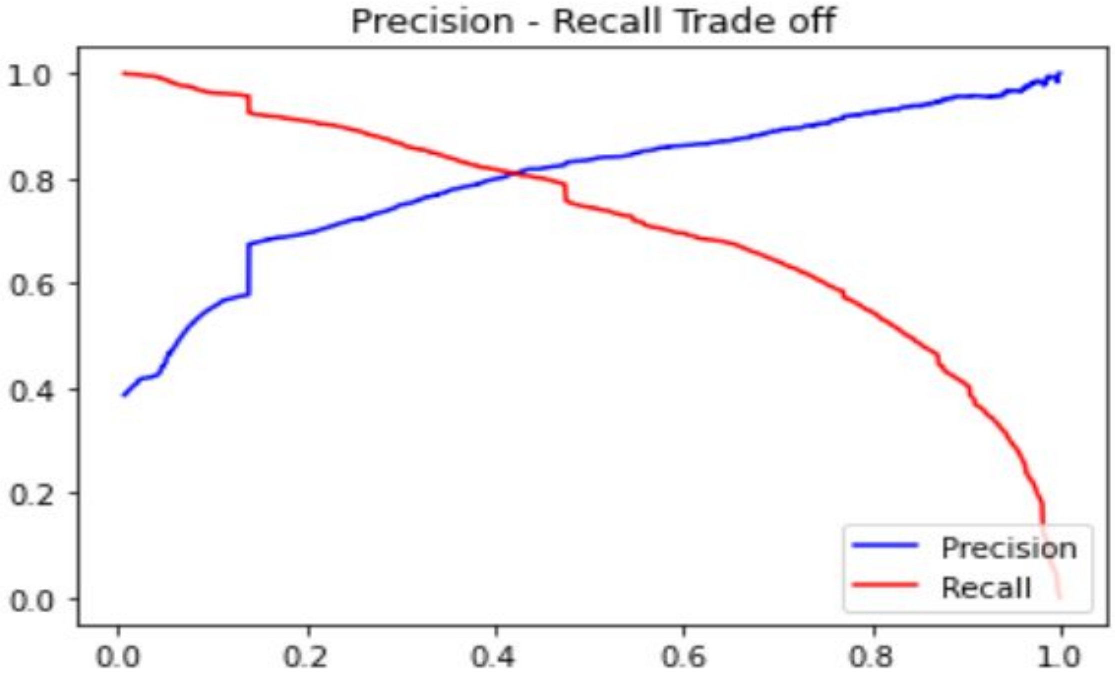
Explanation

The observation made from the previous two graphs are as follows:

- The area under the ROC curve is 92% which is great.
- The cut-off obtained from the plot is about 0.37 which we shall use to perform the prediction.

Metrics\Dataset	**Train**	**Test**
Accuracy	0.8495	0.8443
Sensitivity (Recall)	0.8288	0.8362
Specificity	0.6329	0.6211
Precision	0.7835	0.7746
F-score	0.8055	0.8042

Precision-Recall Tradeoff curve



Explanation

From the precision-recall curve, it is very clear that the cut-off is around 41.

We shall employ this cut-off in our prediction.

Metrics\Dataset	Train	Test
Accuracy	0.8544	0.8495
Sensitivity (Recall)	0.8136	0.8197
Specificity	0.6418	0.6308
Precision	0.8021	0.7936
F-score	0.8078	0.8064

Conclusion

The variables contributing to the conversion are as follows:

1. totalvisits
2. total_time_spent
3. lead_source: olark_chat, reference, welingak_website.
4. lead_quality: not_sure, might_be, worst, low_in_relevance
5. last_activity: sms_sent
6. last_notable_activity: olark_chat_conversation, unreachable
7. do_not_email_yes
8. asymmetric_activity_index_low

The model has achieved an accuracy of about 85% and hence, has sufficiently met the business goal.