LEAD SCORING - Case Study

Premjeet Singh & Shubham Tatad

Business Problem Statement

- Business Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines 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. Moreover, the company also gets leads through past referrals. 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%

Goal:

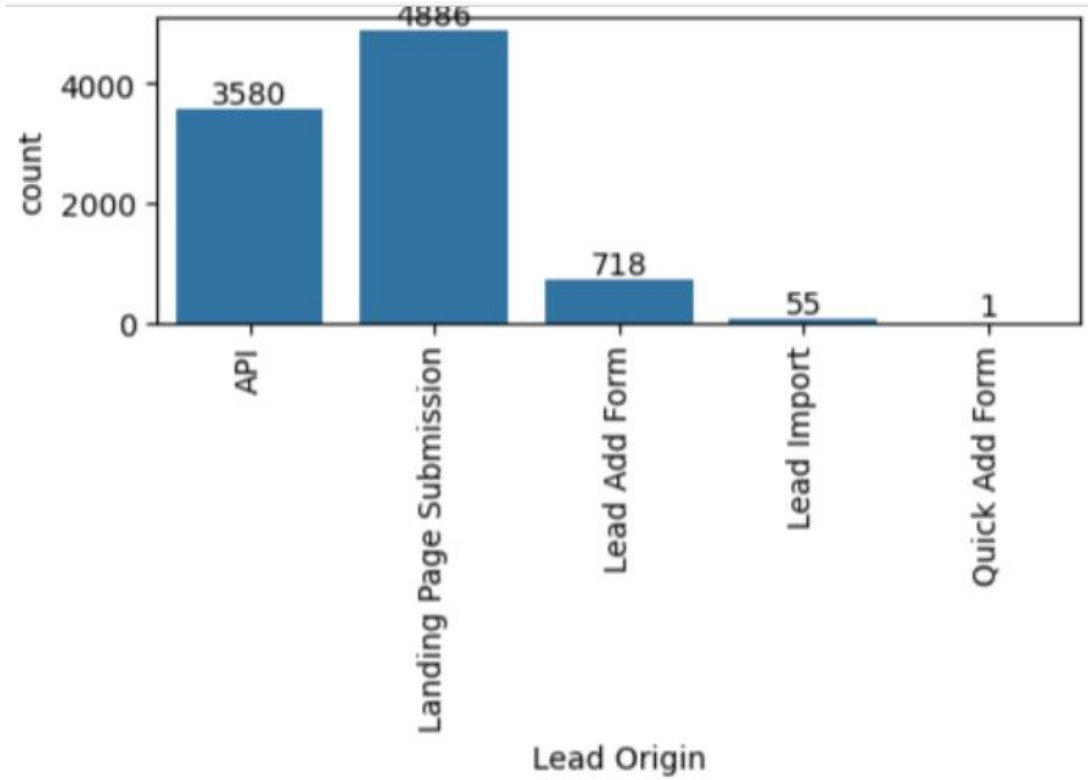
- To identify the features that contributes to predict Lead Conversion.
- Identifying Hot Leads by generating Lead Score for all leads, so that leads having higher Lead Scores can be contacted with priority for achieving Higher Lead Conversion Rate.

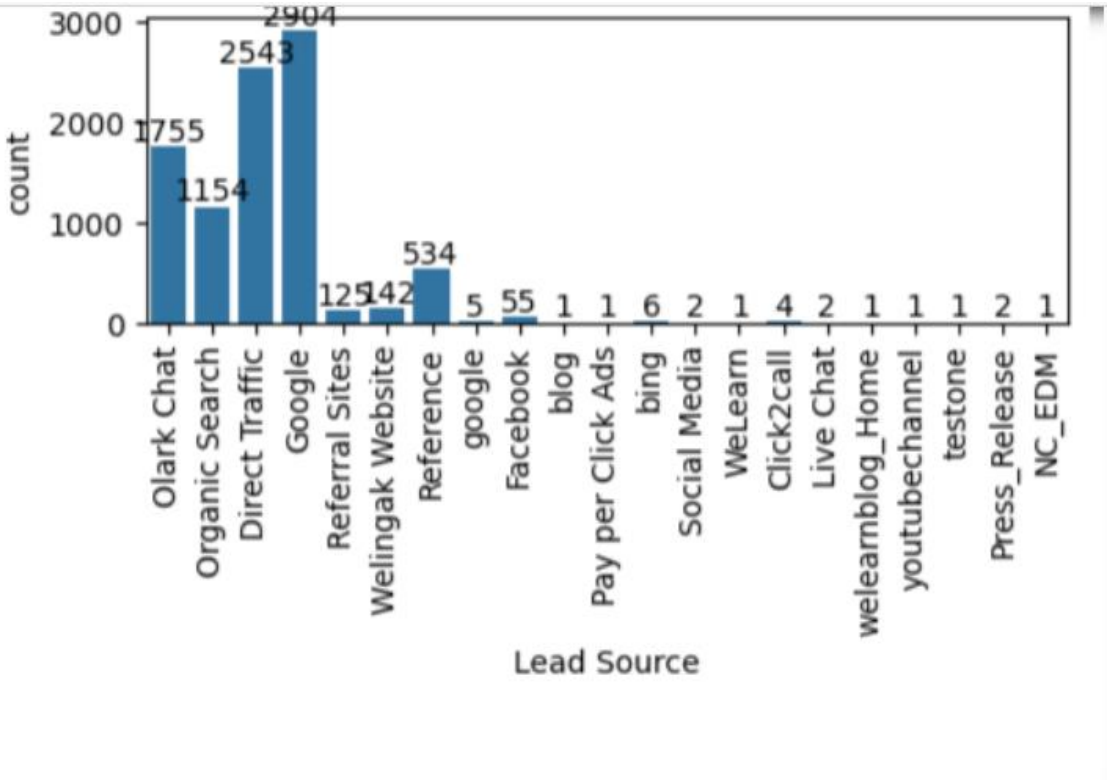
Reading and Understanding the Data:

- When checked and found 9240 records in leads.csv file and it has 37 columns which include 30 categorical and 7 numerical columns.

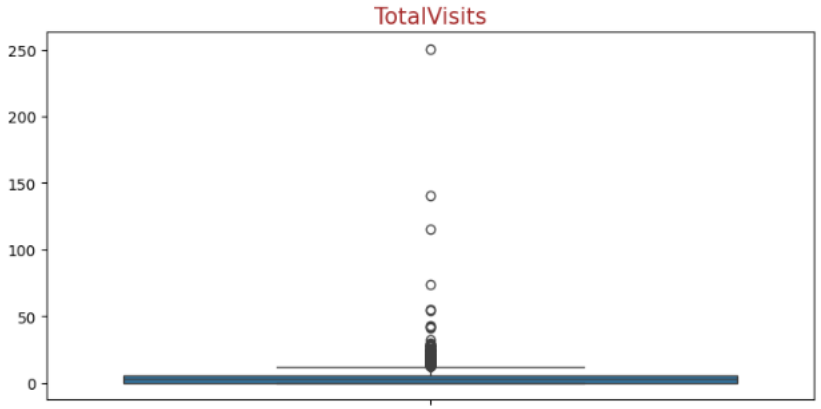
- Select is present as a class in different columns like: ➢ Specialization ➢ How did you hear about X Education ➢ Lead Profile ➢ City
- ‘Select’ is not a valid class, we found that the Select might be the default value set in the form dropdowns. So, we replaced ‘Select’ with NaN.

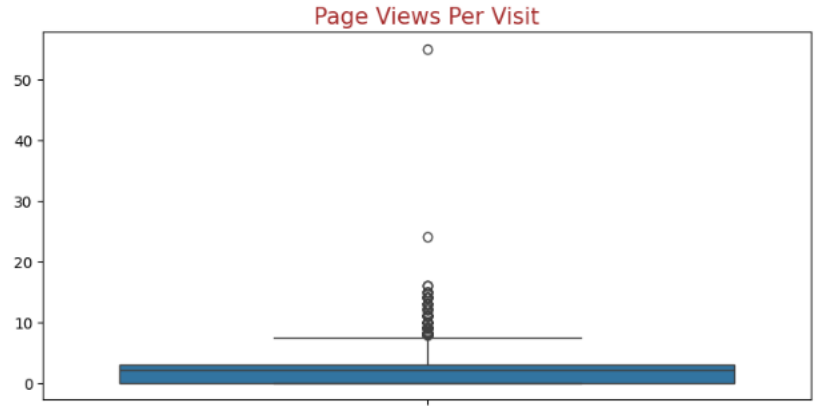
Checking the Uniqueness of the data:

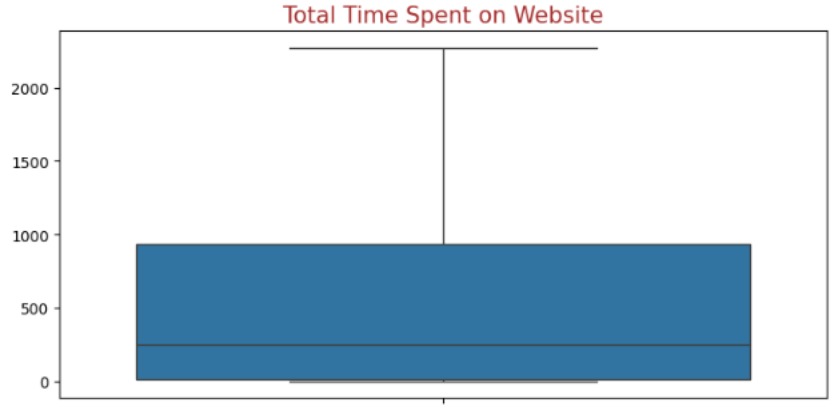




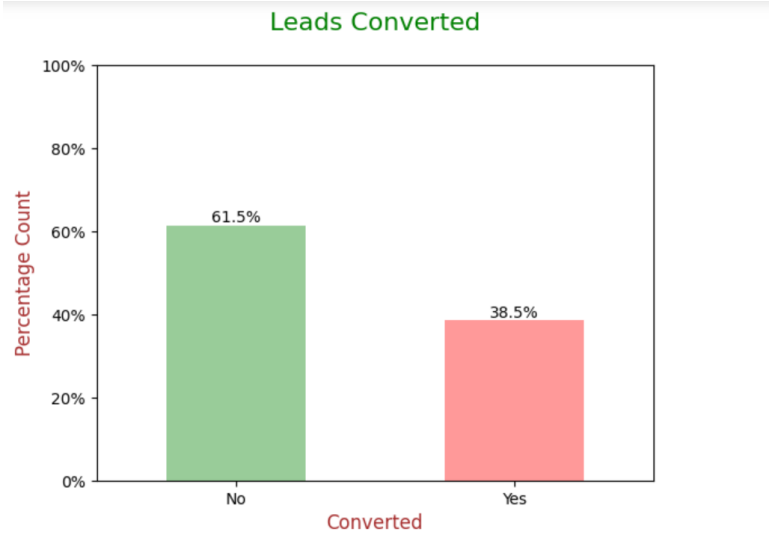
Checking Outliers Using Boxplot:

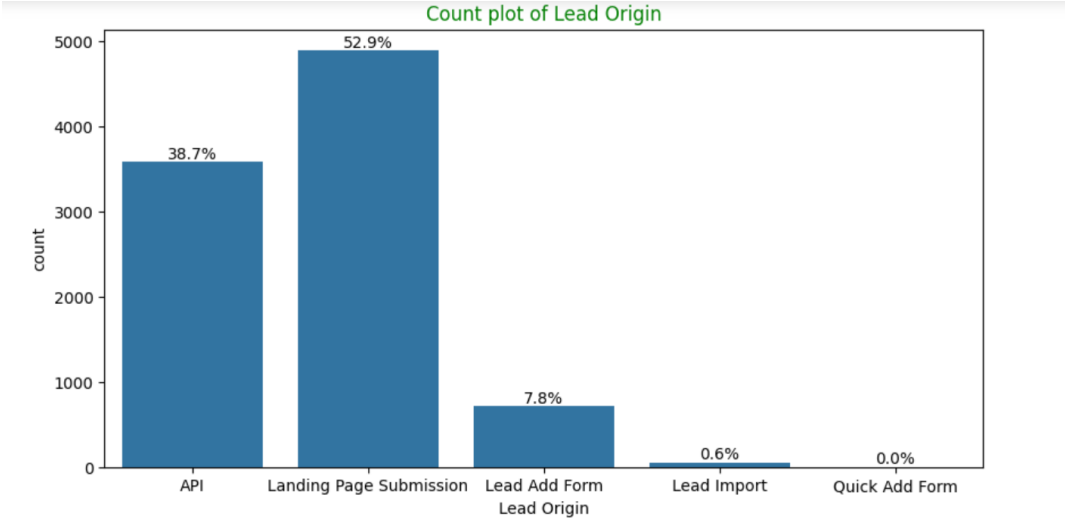


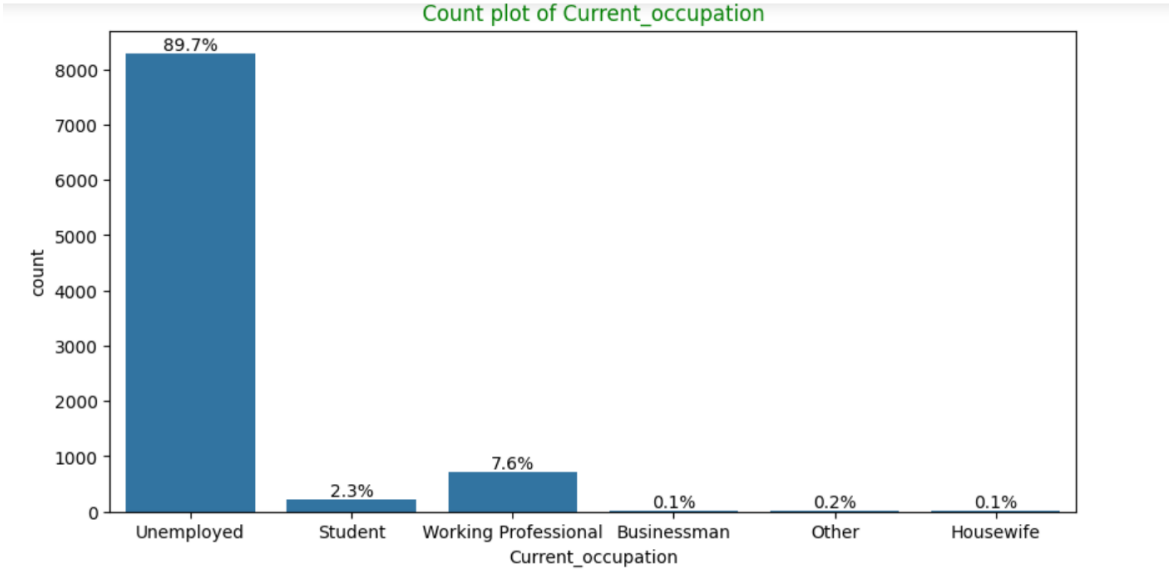




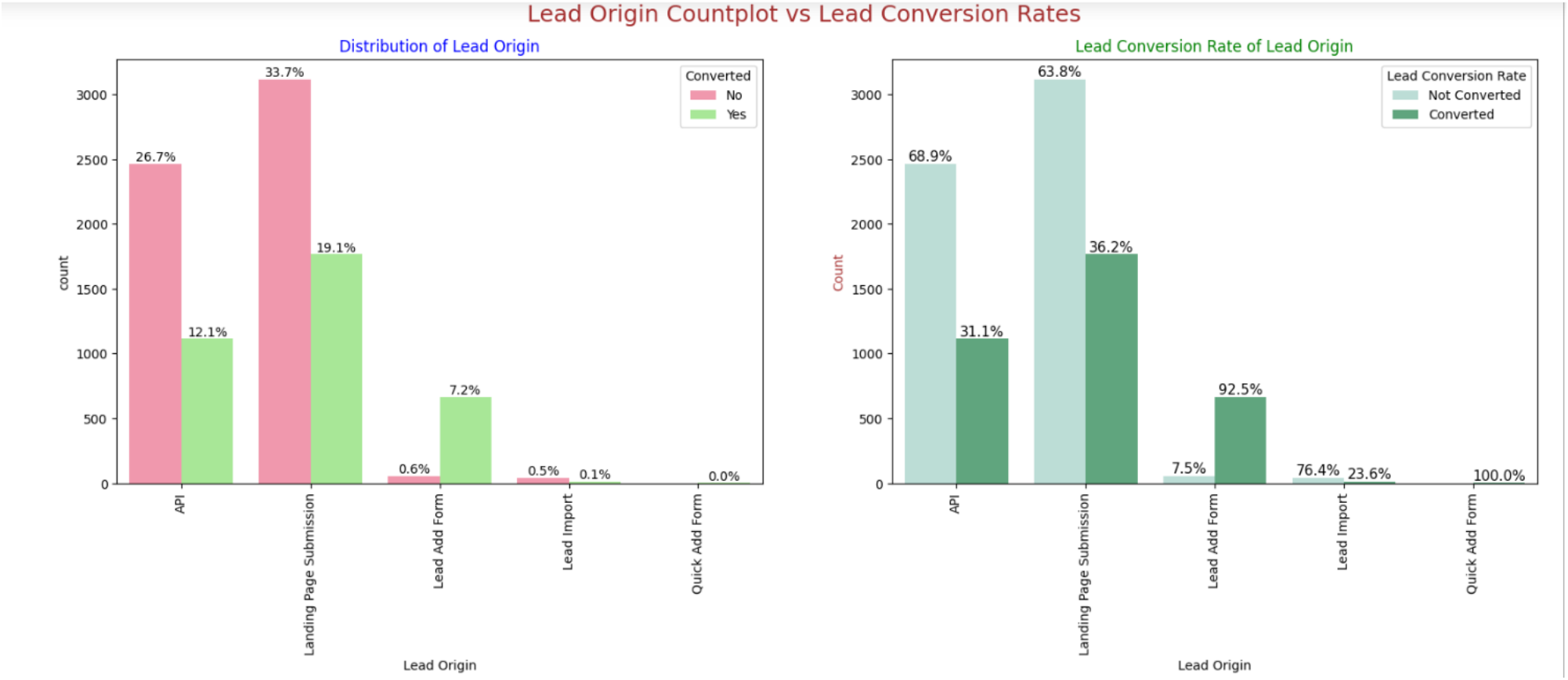
Exploratory Data Analysis:



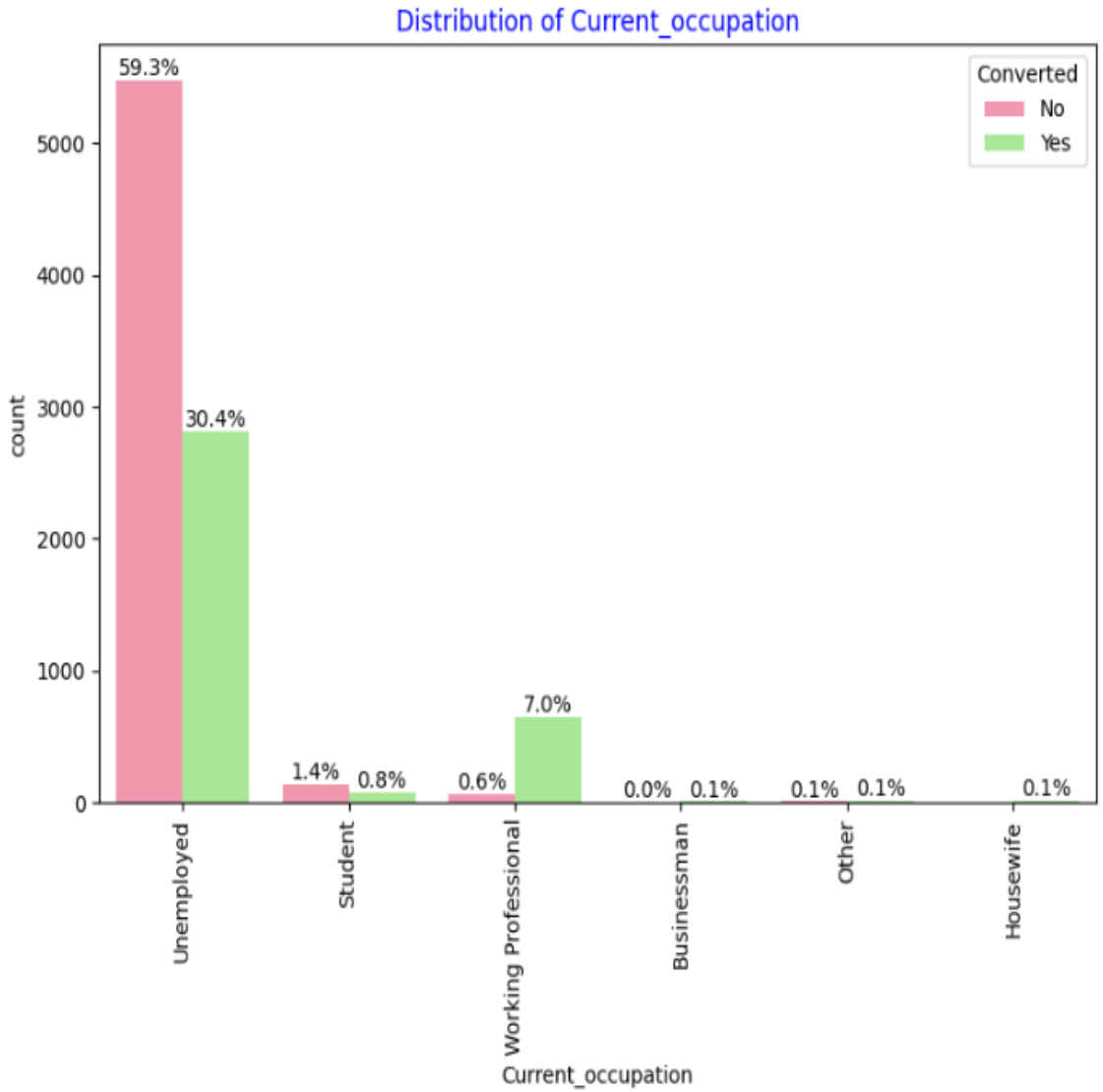


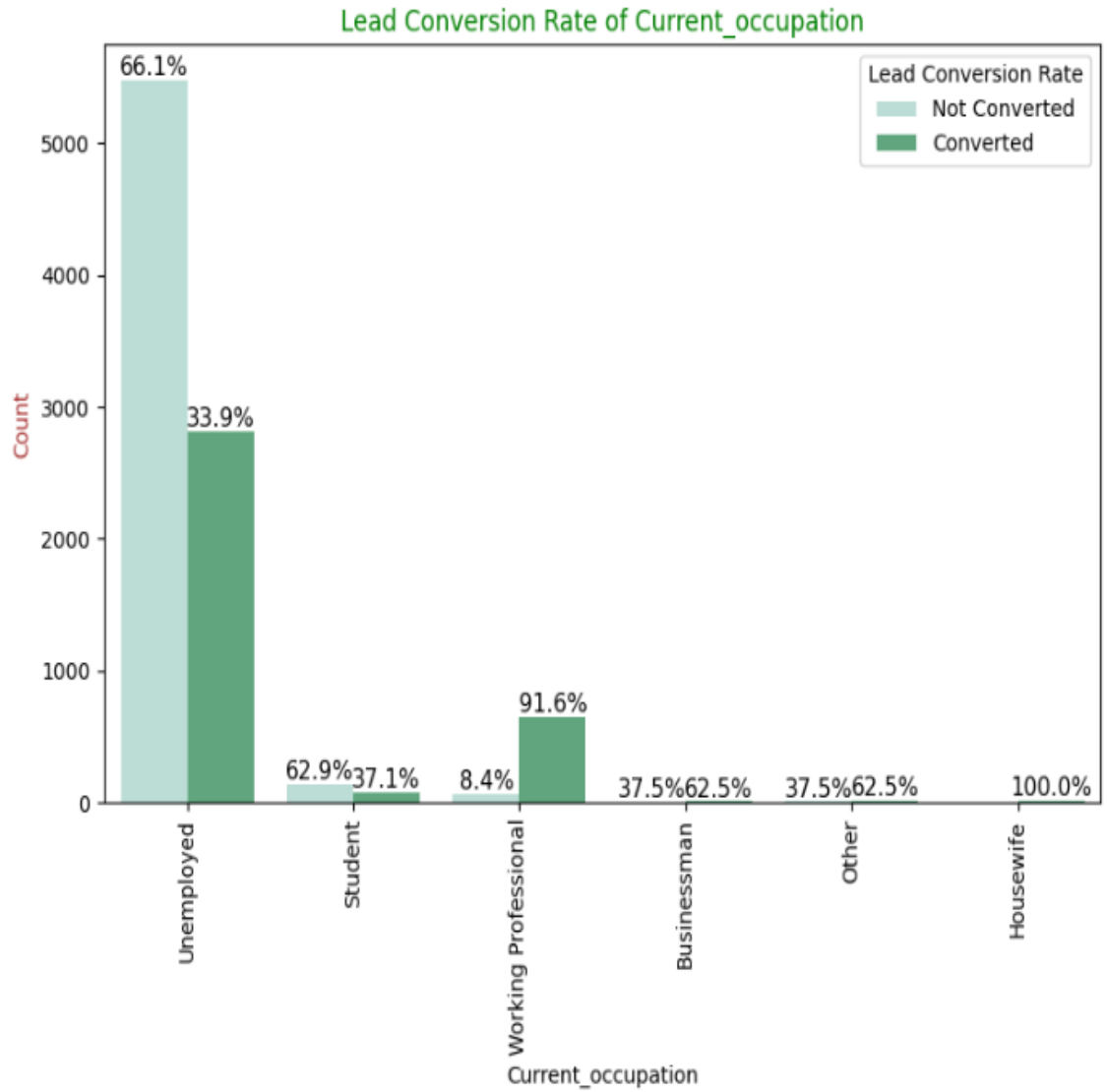


Bivariate Analysis:

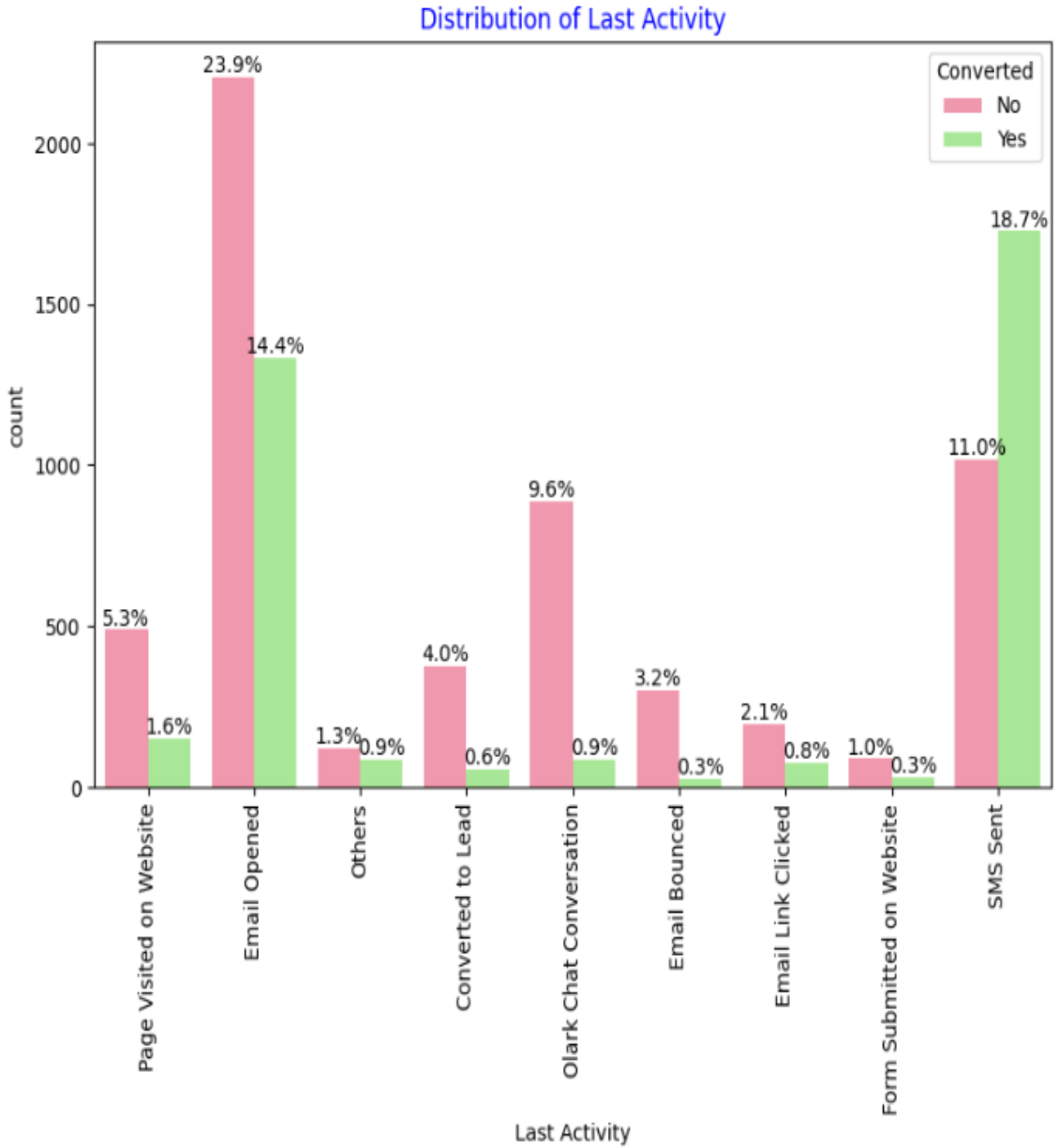


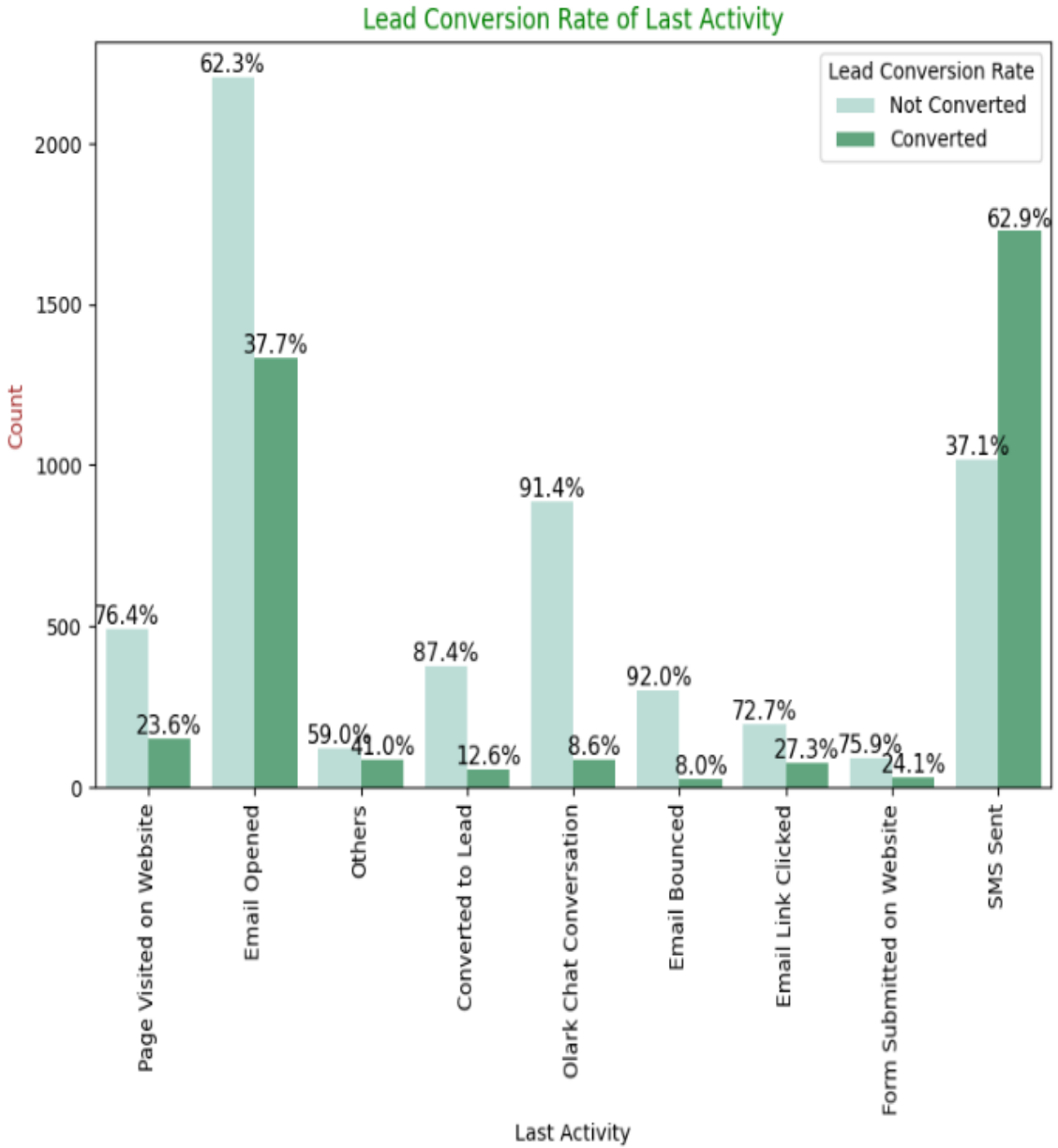
Current_occupation Countplot vs Lead Conversion Rates



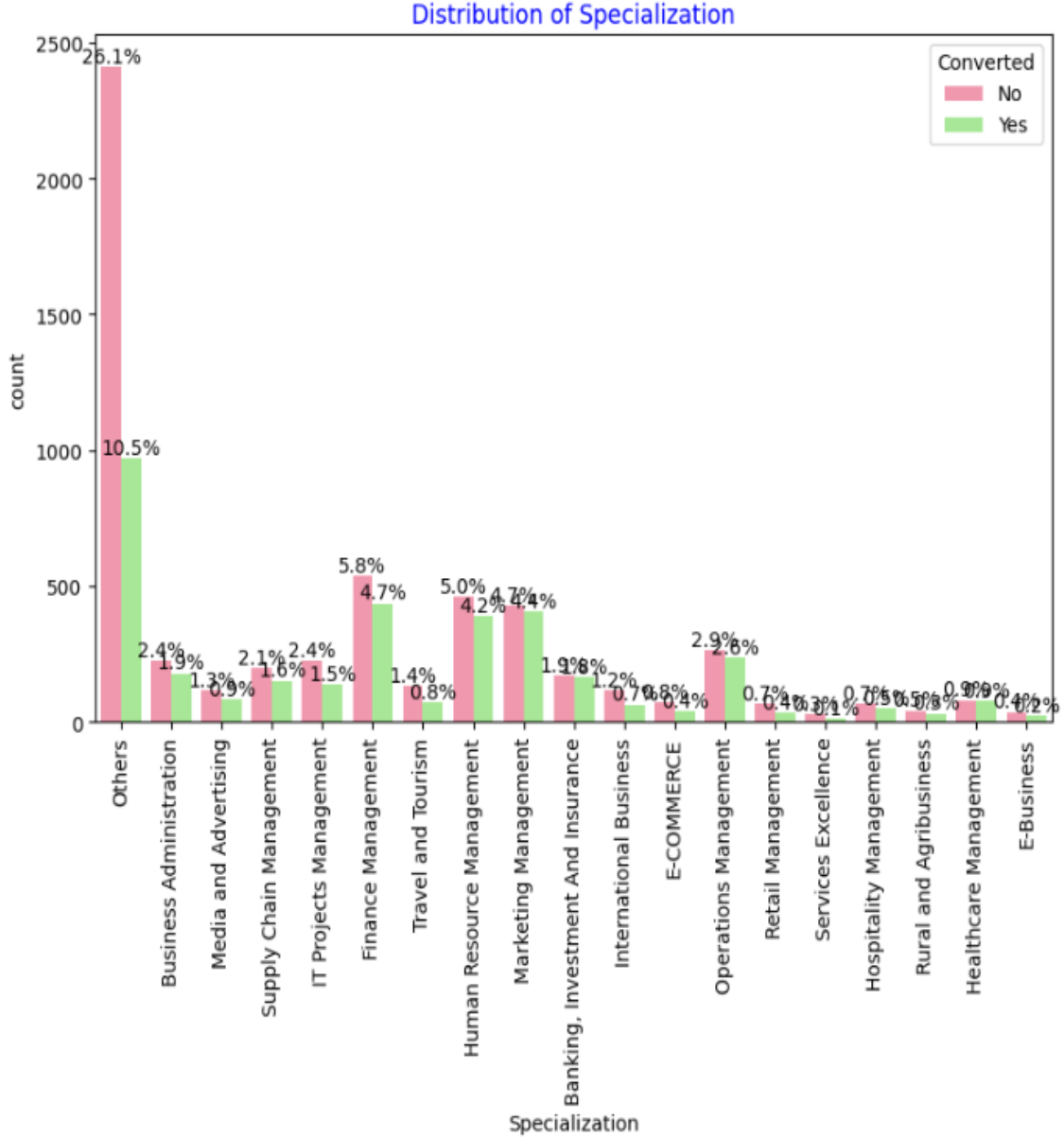


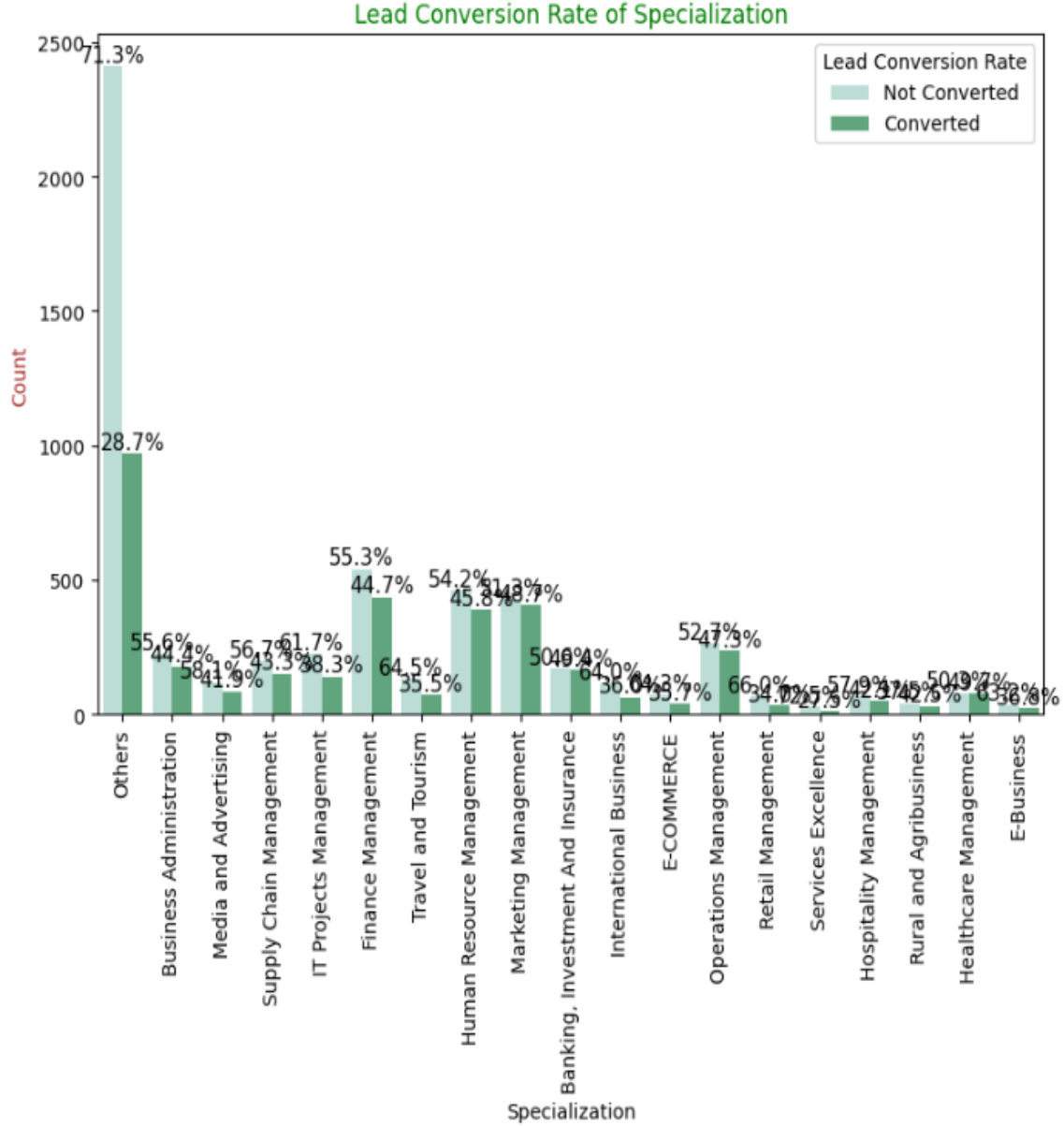
Last Activity Countplot vs Lead Conversion Rates



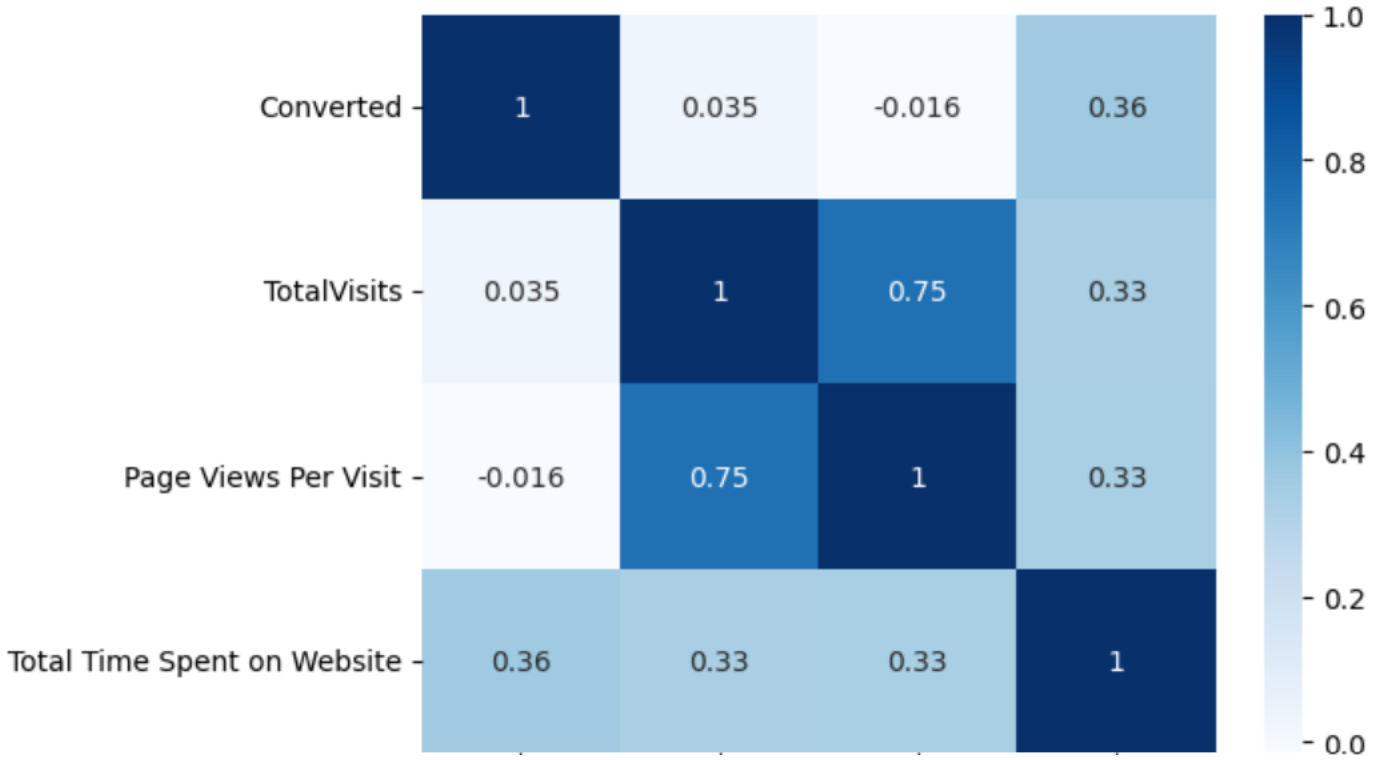


Specialization Countplot vs Lead Conversion Rates





Heatmap to understand Correlation:



Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6453
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2698.2
Date:	Tue, 23 Jul 2024	Deviance:	5396.5
Time:	21:55:07	Pearson chi2:	8.14e+03

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5937	0.039	-15.324	0.000	-0.670	-0.518
Do Not Email	-0.2902	0.046	-6.325	0.000	-0.380	-0.200
TotalVisits	0.3426	0.050	6.902	0.000	0.245	0.440
Total Time Spent on Website	1.0369	0.039	26.685	0.000	0.961	1.113
Page Views Per Visit	-0.3165	0.056	-5.628	0.000	-0.427	-0.206
Lead Origin_Landing Page Submission	-0.5836	0.063	-9.241	0.000	-0.707	-0.460
Lead Source_Olark Chat	0.3579	0.052	6.945	0.000	0.257	0.459
Lead Source_Reference	0.6971	0.053	13.103	0.000	0.593	0.801
Lead Source_Welingak Website	0.6457	0.087	7.458	0.000	0.476	0.815
Last Activity_Email Opened	0.4081	0.053	7.707	0.000	0.304	0.512
Last Activity_Olark Chat Conversation	-0.2015	0.057	-3.510	0.000	-0.314	-0.089
Last Activity_Others	0.2090	0.035	5.975	0.000	0.140	0.278
Last Activity_SMS Sent	0.9407	0.051	18.344	0.000	0.840	1.041
Specialization_Others	-0.5576	0.059	-9.467	0.000	-0.673	-0.442
Current_occupation_Working Professional	0.7121	0.051	13.918	0.000	0.612	0.812

Conclusion:

The variables which decide the probability of the lead getting converted are as follows:

- Total time spent on website
- Lead source
- Current occupation (if working professional or not)