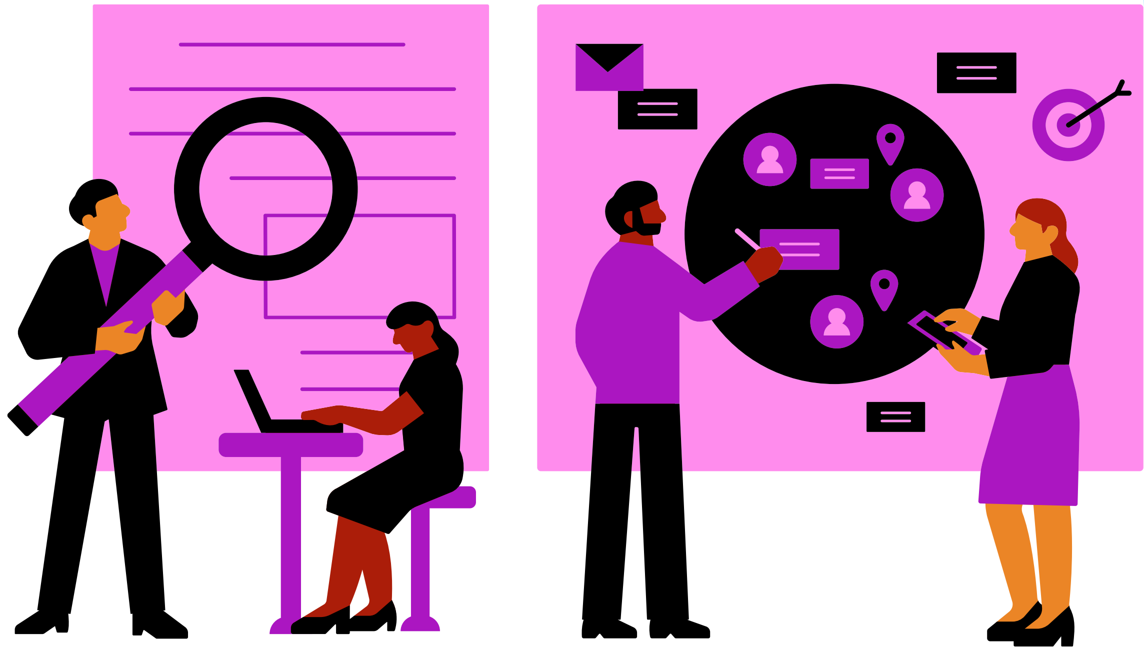
LEAD SCORING MODEL: IDENTIFYING POTENTIAL LEADS FOR HIGHER CONVERSION RATES

By,
Nisarga Priya

CONTENTS



1 Problem statement

2 Data cleaning and Preparation

3 Exploratory data analysis and Logistic regression model

4 Summary and Recommendations

LET'S BEGIN!

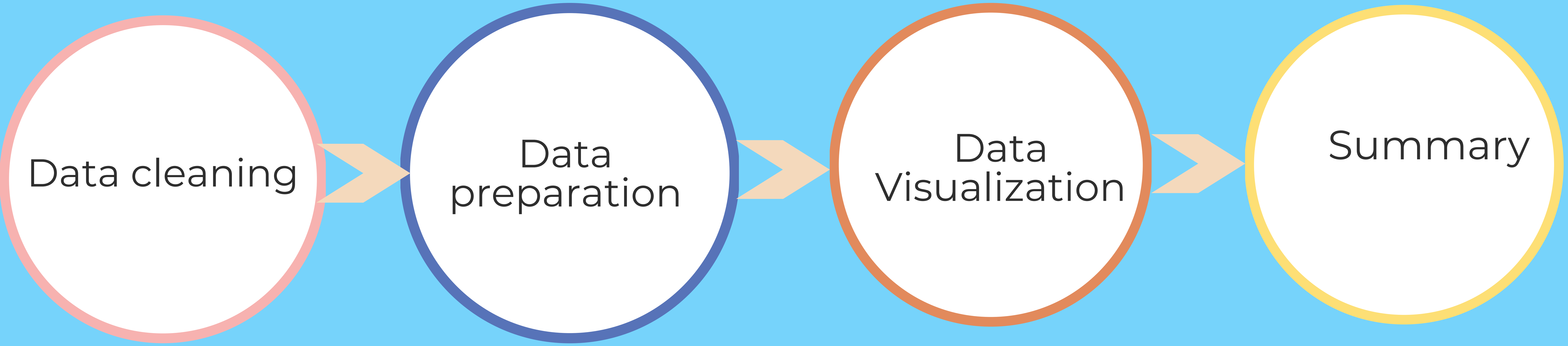
PROBLEM STATEMENT:

Below are some of the reasons for why we are conducting this analysis:

- X Education receives a significant number of leads daily but has a low conversion rate.
- Only around 30 out of 100 acquired leads are converted into paying customers.
- The company wants to identify 'Hot Leads' with higher chances of conversion.
- The goal is to develop a lead scoring model assigning scores from 0 to 100 to each lead.
- The model will help prioritize potential leads for more effective communication.
- The objective is to optimize the lead conversion process and enhance the sales team's efficiency.

STEPS TAKEN TO ANALYSE THE DATA

We clean and prepare the data set to perform EDA which includes following process:



Data cleaning

Data preparation

Data Visualization

Summary

DATA CLEANING

Process followed for data cleaning:

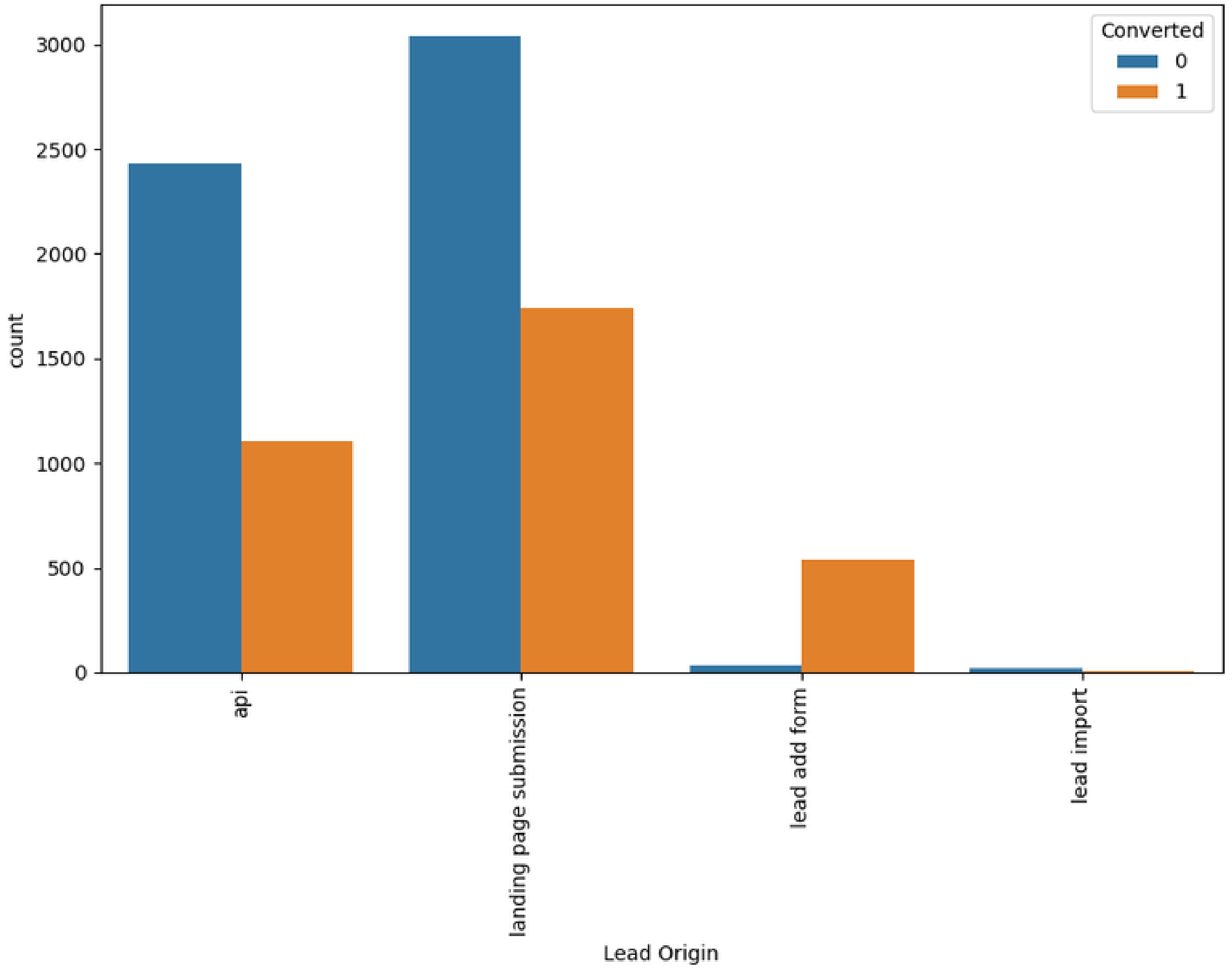
1. Duplicate data was checked for and handled.
2. NA values and missing values were checked and handled.
3. Columns with a large amount of missing values and deemed not useful for the analysis were dropped.
4. Imputation of values was performed with median and mode values
5. Outliers in the data were checked and outlier treatment was performed

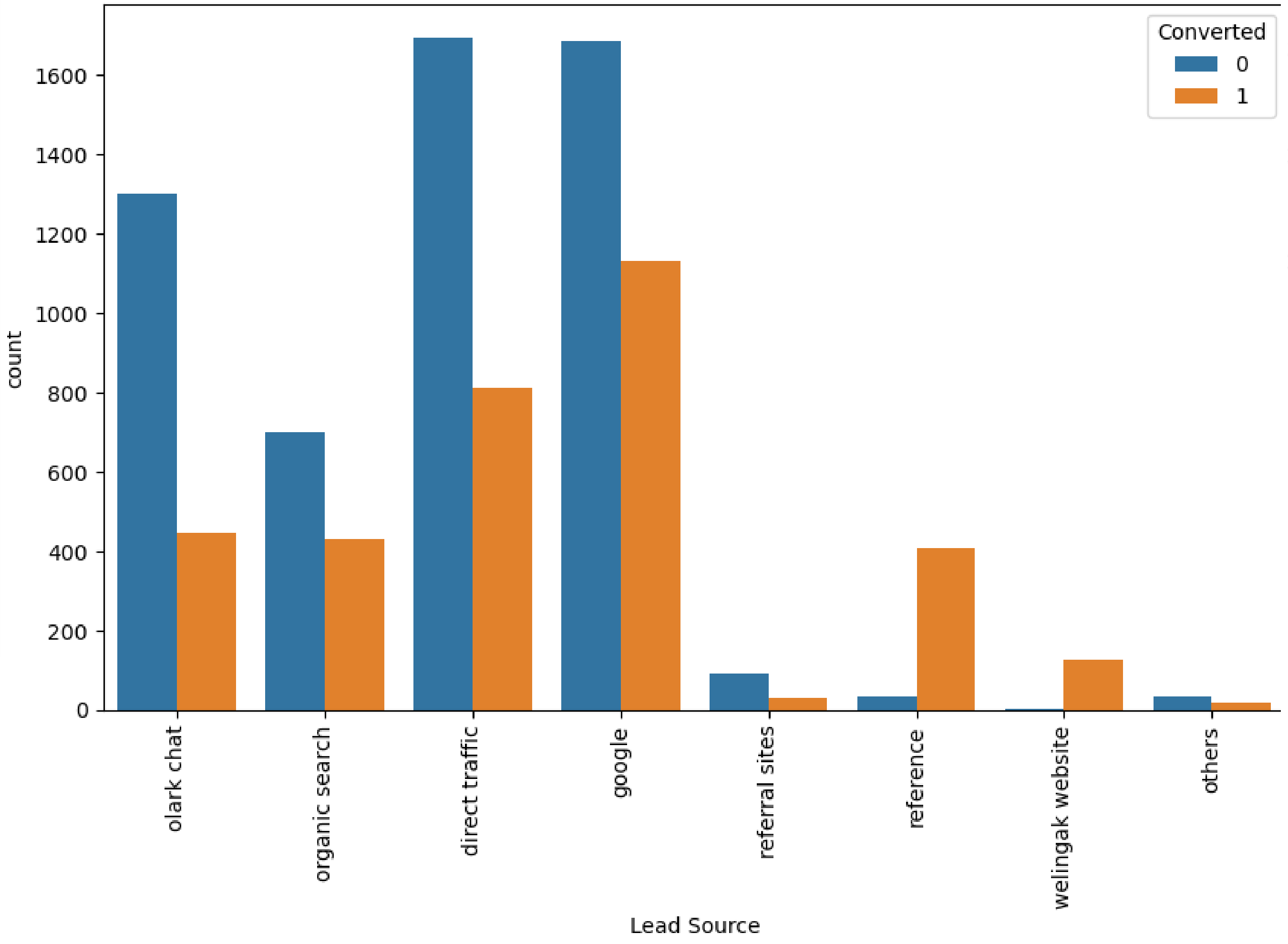
DATA PREPARATION

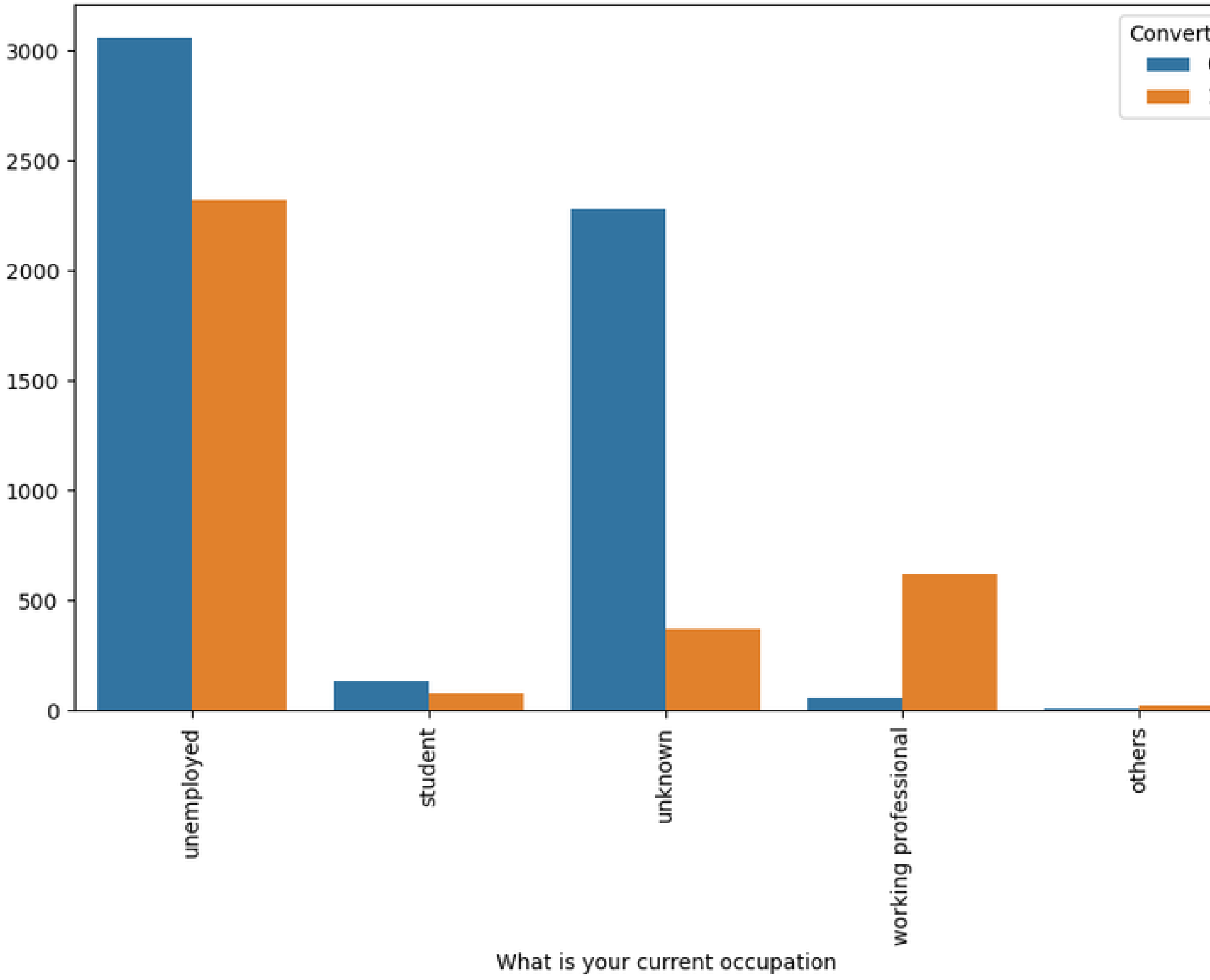
Steps taken to prepare the data:

1. Single value features like "Magazine," "Receive More Updates About Our Courses," "Update me on Supply Chain Content," "Get updates on DM Content," "I agree to pay the amount through cheque," etc., were dropped.
2. The "Prospect ID" and "Lead Number" columns, which were not necessary for the analysis, were removed.
3. Features with low variance, such as "Do Not Call," "What matters most to you in choosing a course," "Search," "Newspaper Article," "X Education Forums," "Newspaper," "Digital Advertisement," etc., were dropped. •
4. Columns with more than 45% missing values were dropped.

EXPLORATORY DATA ANALYSIS: UNIVARIATE ANALYSIS



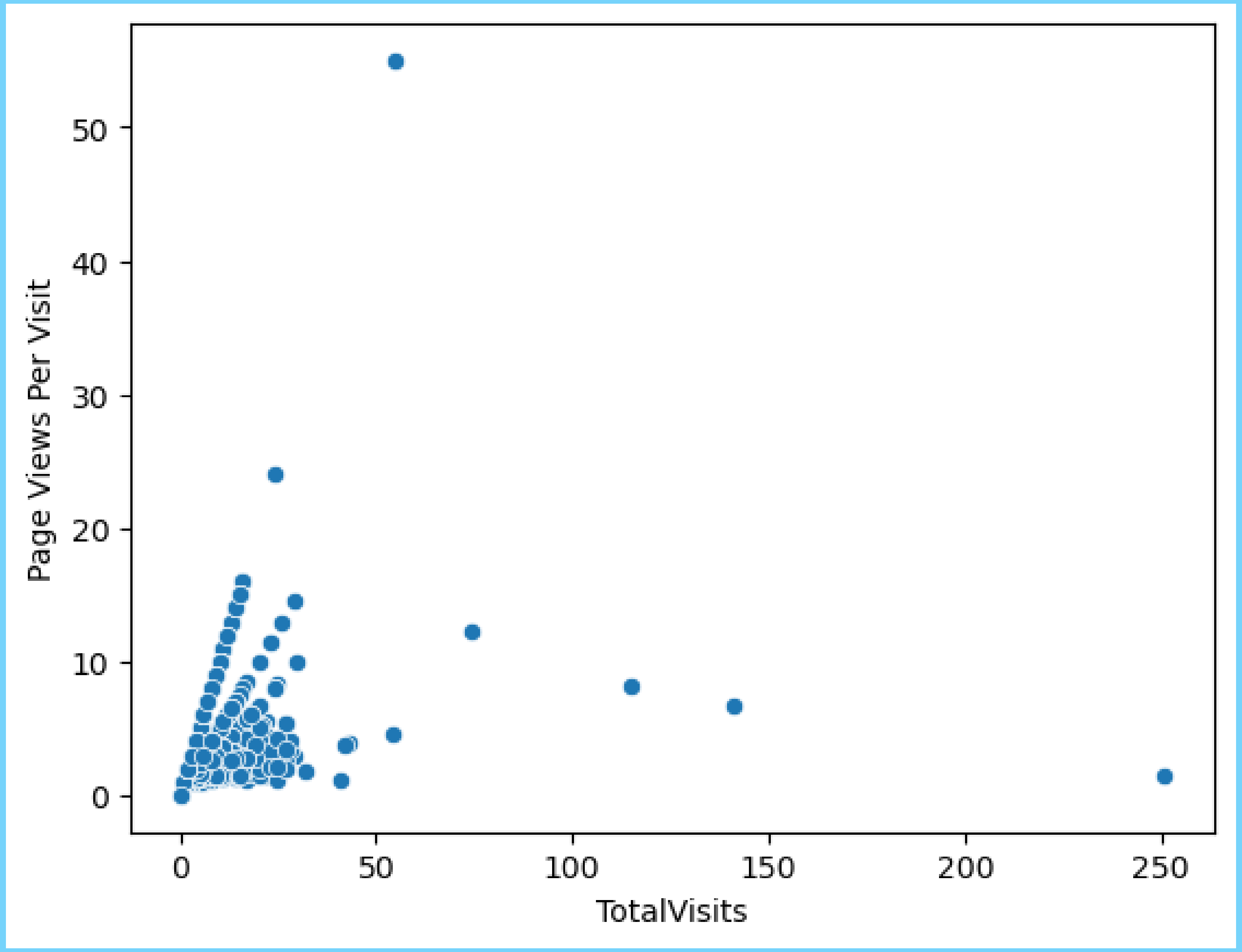


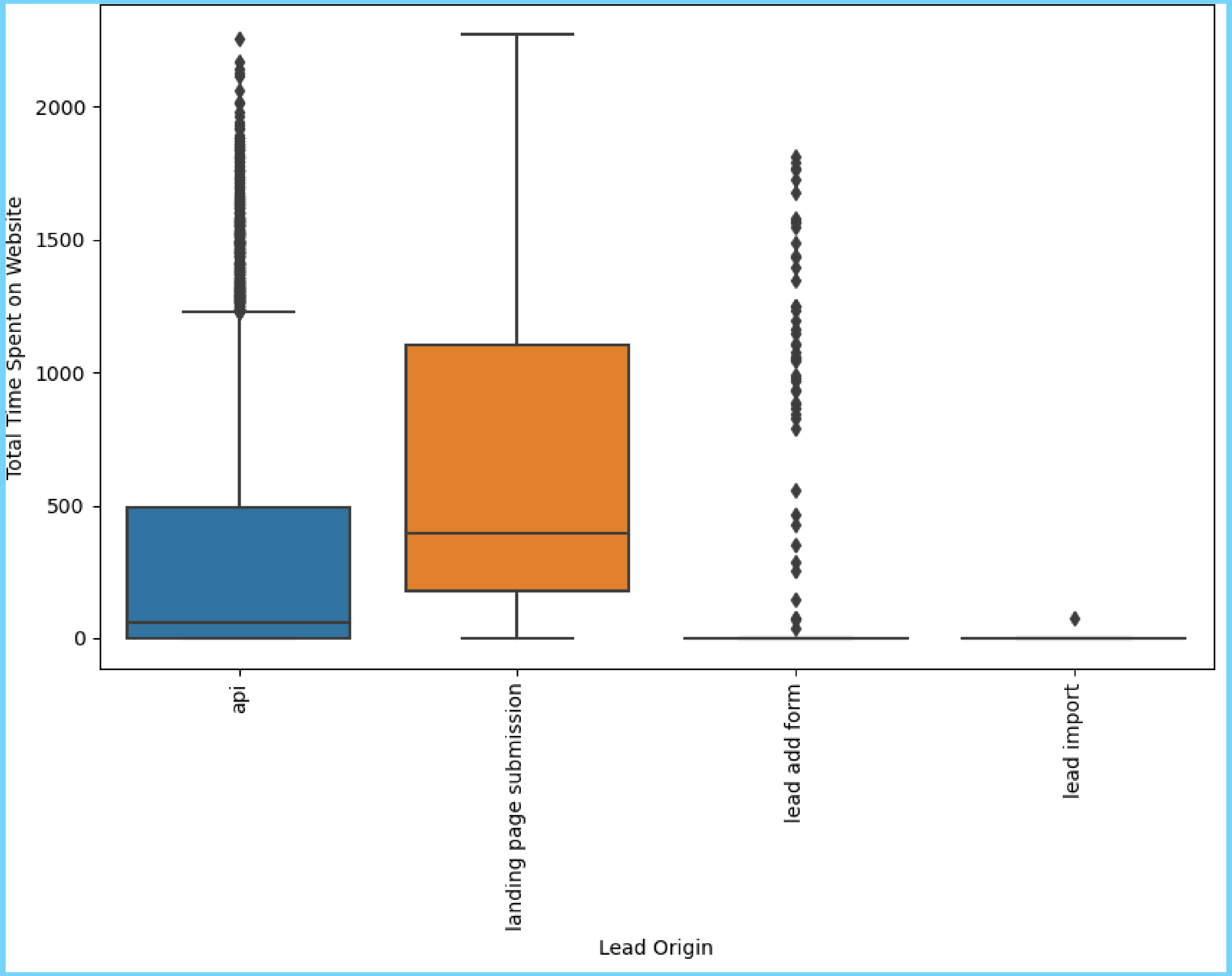


Insights:

- Working professionals are the most frequently converted leads.
- Unemployed individuals constitute a significant portion of the leads.

EXPLORATORY DATA ANALYSIS: BIVARIATE ANALYSIS

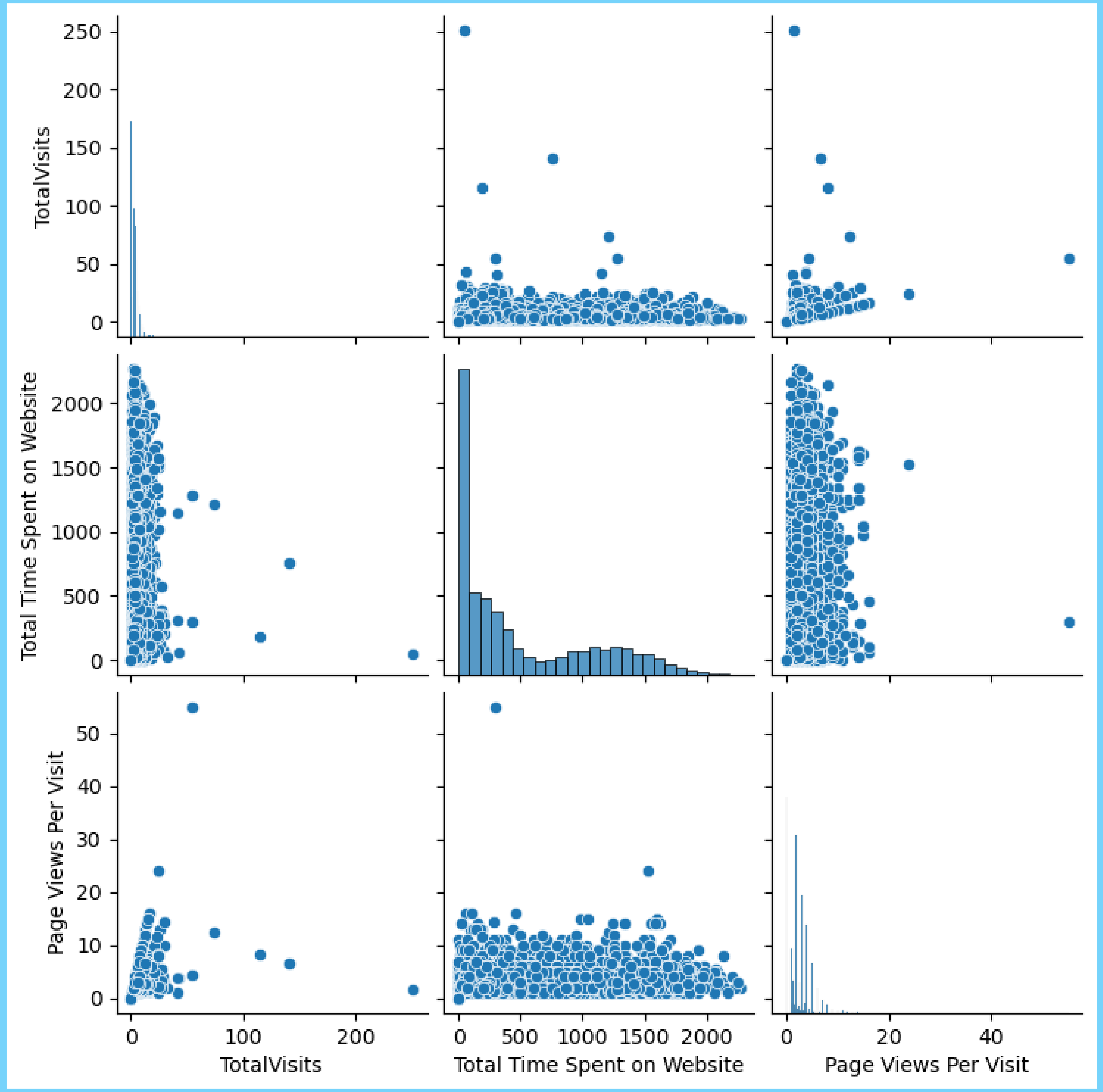


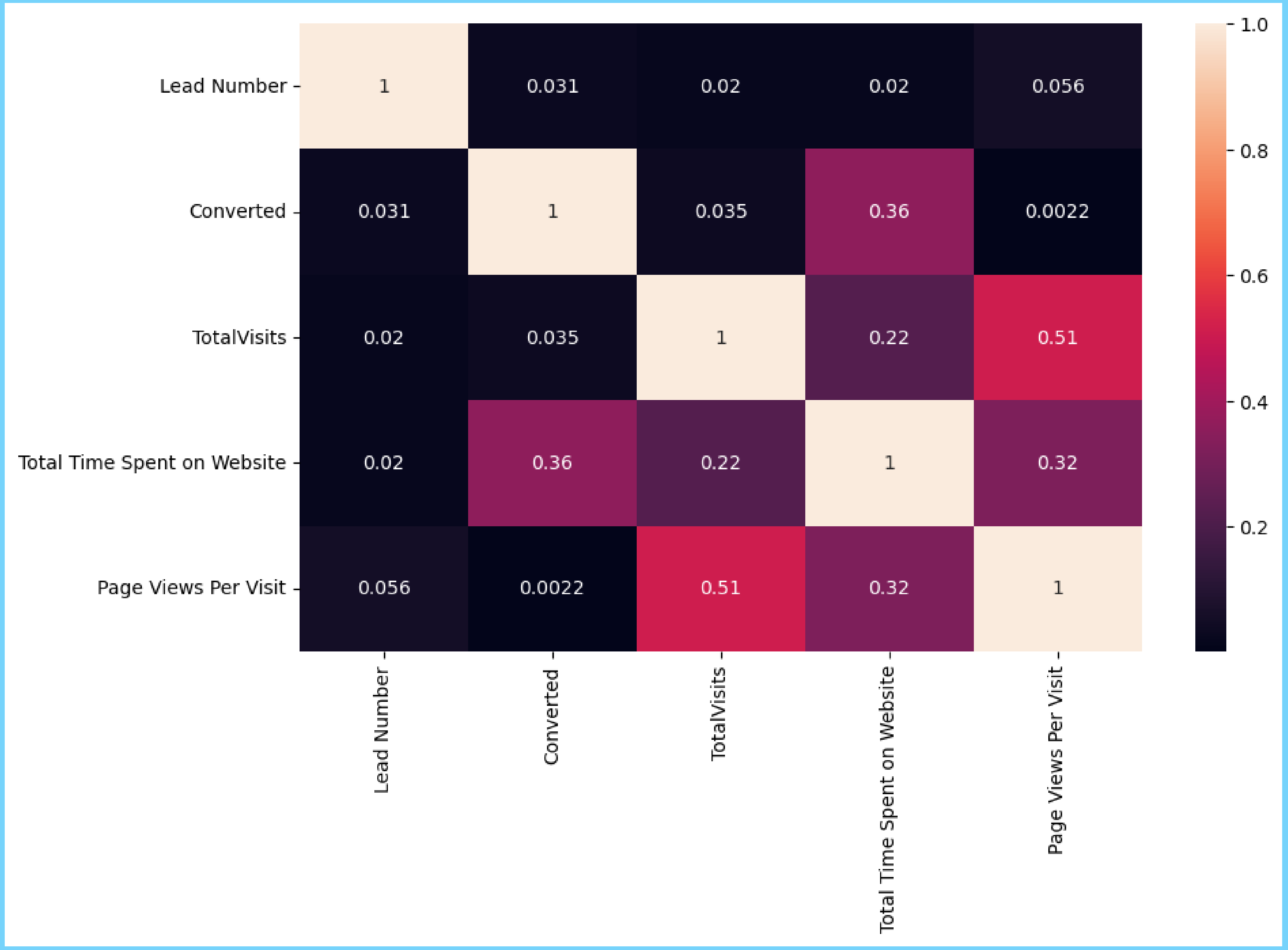


Insights:

- There is a consistent relationship between the total number of visits and the total time spent on the website.
- Time spent online through api second highest , first being the landing page

EXPLORATORY DATA ANALYSIS: MULTIVARIATE ANALYSIS

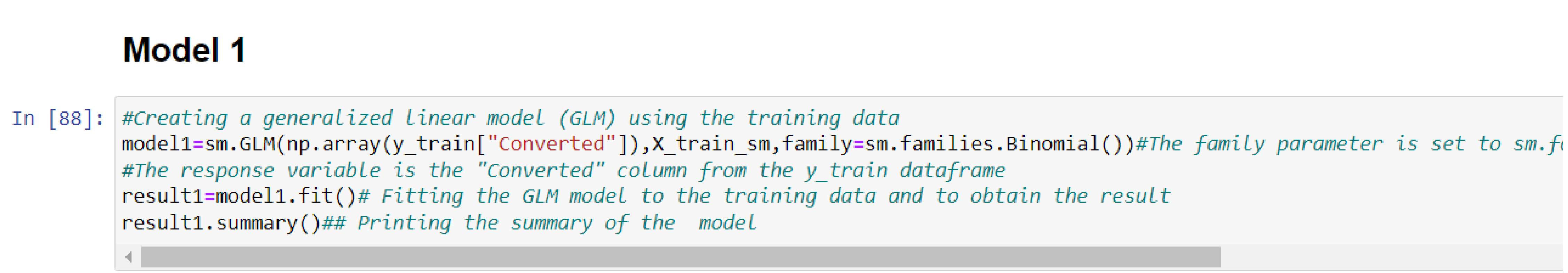




Insights:

- There is positive correlation between continuous columns 'Total Visits' and 'Page views per visit
- There is a positive correlation between 'converted column' and 'total time spent on website'
- There is positive correlation between 'converted' and 'total visits'

LOGISTIC REGRESSION MODEL:

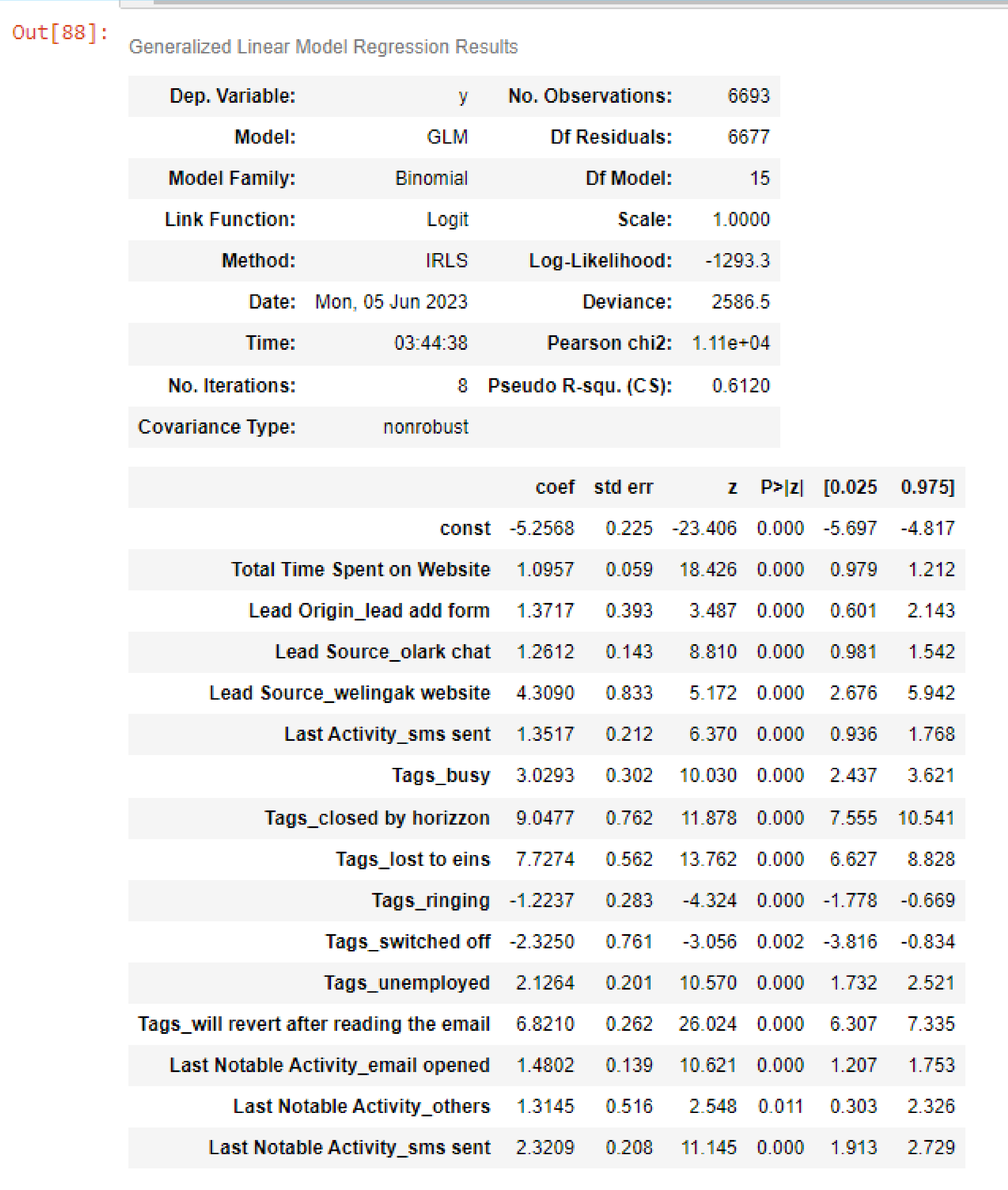


Model 1

In [88]:

```
#Creating a generalized linear model (GLM) using the training data
model1=sm.GLM(np.array(y_train["Converted"]),X_train_sm,family=sm.families.Binomial())#The family parameter is set to sm.f
#The response variable is the "Converted" column from the y_train dataframe
result1=model1.fit()# Fitting the GLM model to the training data and to obtain the result
result1.summary()## Printing the summary of the  model
```

- The data was split into training and testing sets using a 75:25 ratio.
- Recursive Feature Elimination (RFE) was performed to select the top 15 variables.
- The model was built by removing variables with a p-value greater than 0.05 and a VIF value greater than 5.
- Predictions were made on the test dataset, and the overall accuracy of the model was determined to be 92%.



Out[88]:

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	6693
Model:	GLM	Df Residuals:	6677
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1293.3
Date:	Mon, 05 Jun 2023	Deviance:	2586.5
Time:	03:44:38	Pearson chi2:	1.11e+04
No. Iterations:	8	Pseudo R-squ. (CS):	0.6120
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.2568	0.225	-23.406	0.000	-5.697	-4.817
Total Time Spent on Website	1.0957	0.059	18.426	0.000	0.979	1.212
Lead Origin_lead add form	1.3717	0.393	3.487	0.000	0.601	2.143
Lead Source_olark chat	1.2612	0.143	8.810	0.000	0.981	1.542
Lead Source_welingak website	4.3090	0.833	5.172	0.000	2.676	5.942
Last Activity_sms sent	1.3517	0.212	6.370	0.000	0.936	1.768
Tags_busy	3.0293	0.302	10.030	0.000	2.437	3.621
Tags_closed by horizzon	9.0477	0.762	11.878	0.000	7.555	10.541
Tags_lost to eins	7.7274	0.562	13.762	0.000	6.627	8.828
Tags_ringing	-1.2237	0.283	-4.324	0.000	-1.778	-0.669
Tags_switched off	-2.3250	0.761	-3.056	0.002	-3.816	-0.834
Tags_unemployed	2.1264	0.201	10.570	0.000	1.732	2.521
Tags_will revert after reading the email	6.8210	0.262	26.024	0.000	6.307	7.335
Last Notable Activity_email opened	1.4802	0.139	10.621	0.000	1.207	1.753
Last Notable Activity_others	1.3145	0.516	2.548	0.011	0.303	2.326
Last Notable Activity_sms sent	2.3209	0.208	11.145	0.000	1.913	2.729

SUMMARY AND RECOMMENDATIONS

After analysing and visualizing the data set we can conclude that:

- Key features that significantly influence lead conversion were , total time spent on the website, lead origin through lead add forms, olark chats and google as a lead source, and welingak website as a lead source.
- Focusing on engaging individuals who spend a significant amount of time on the website by sending attractive offers and personalized content will convince them to go through courses.
- Target unemployed individuals with programs highlighting job guarantees or interview guarantees to capture their interest.
- Optimize search engine visibility, particularly on platforms like Google, to ensure X Education appears at the top of search results for relevant certifications or education-related queries.
- Implementing email campaigns to provide informative content, such as free courses and master classes, to nurture leads and increase conversion chances.
- starting a campaign to encouraging women who are housewives to get educated through courses and provide interview guarantee programs