LEAD SCORING PROJECT

BY-

Y. SWETHA ASHERVADAM

Problem Statement

- *X Education, an online education company, faces challenges in converting leads into paying customers despite generating a significant number of leads daily.*
- *With a conversion rate of only 30%, the company aims to improve its lead identification process to target potential customers more effectively.*
- *They seek to develop a model that assigns lead scores to prioritize leads based on their likelihood of conversion, ultimately aiming for an 80% conversion rate.*

Data Scenario Categorization

- The dataset provided contains approximately **9000** data points, including attributes like Lead Source, Total Time Spent on Website, Total Visits, and Last Activity.
- The target variable, **'Converted'**, indicates whether a lead was converted **(1)** or not **(0)**.
- The dataset also includes categorical variables with levels such as **'Select'**, which need handling as they are equivalent to null values.

Objectives

1) Build a logistic regression model to assign lead scores ranging from 0 to 100 to each lead, aiding the company in targeting potential customers effectively. Higher scores indicate hotter leads with a higher likelihood of conversion, while lower scores represent colder leads less likely to convert.

2) Ensure the model's adaptability to accommodate potential changes in the company's requirements in the future, addressing any additional problems or challenges that may arise.

STEPS FOLLOWED

Importing Data and inspecting the Dataframe

- Reading the data from the csv file using pandas library

Data Exploration

- Handling missing values
- Handling the 'Select' value present in four variables
- Checking Data Imbalance

Exploratory Data Analysis

- Univariate Analysis
- Bivariate Analysis
- Segmented Analysis on Target variable 'Converted'
- Multivariate Analysis

Data cleaning and preparation

- Dropping the imbalanced variables
- Replacing lower frequency values if required
- Treatment of Outliers

Feature Engineering

- Converting the binary variables (Yes/No) to 0/1
- Handling categorical variables
  - Mapping categorical variables to integers
  - Dummy variable creation

Test-train split and scaling

- Train data – 70% and Test data – 30%

Model Building

- Feature elimination based on correlations
- Feature selection using RFE (Coarse Tuning)
- Manual feature elimination (using p-values and VIFs)

Model Evaluation

- Accuracy
- Sensitivity and Specificity
- Optimal cut-off using ROC curve
- Precision and Recall

Predictions on the test set

- Predictions using the test data

Data Manipulation

- Handling Missing Values
- Dropping single value variables or imbalanced variables
- Capping the outliers

Handling Missing Values

This dataset has 9240 rows and 37 columns.
Identified and removed columns with null values exceeding 40% threshold for analysis.
After removal, 12 columns still had null values.
Imputed missing values in 'Tags', 'What matters most to you in choosing a course' and 'What is your current occupation' column with 'Not Specified'.
Imputed missing values in 'Country' column with mode and clubbed lower frequency values.

Addressed the variables containing 'Select' values, requiring handling as follows:

- Dropped 'Lead Profile' column due to combined presence of 'Select' and null values exceeding 70%.
- Dropped 'City' column due to combined presence of 'Select' and null values exceeding 39%.
- Dropped 'How did you hear about X Education' column due to combined presence of 'Select' and null values constituting 74%.
- Imputed 'Specialization' column's 'Select' and null values with 'Not Specified' due to constituting 36% of the values.

Eliminated rows with negligible null values (ranging from approximately 1% to 2%) in remaining four key variables.

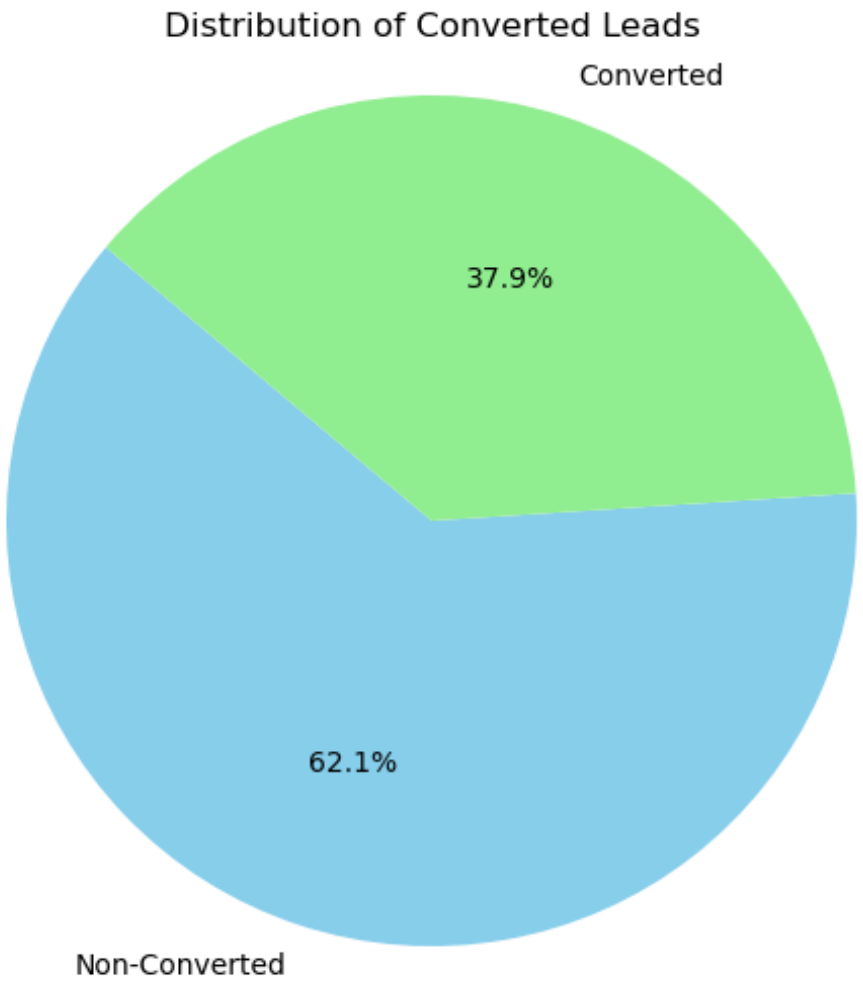
Dropped columns deemed imbalanced:

- Prospect ID, Lead Number, Do Not Call, Magazine, Search, Newspaper Article, X Education Forums, Newspaper, Through Recommendations, Digital Advertisement, I agree to pay the amount through cheque, Update me on Supply Chain Content, Receive More Updates About Our Courses, Get updates on DM Content.

Capped outliers in 'TotalVisits' and 'Page Views Per Visit' to the 95th percentile.

After data cleaning, the dataset contains 9074 rows and 15 columns.

Data Imbalance

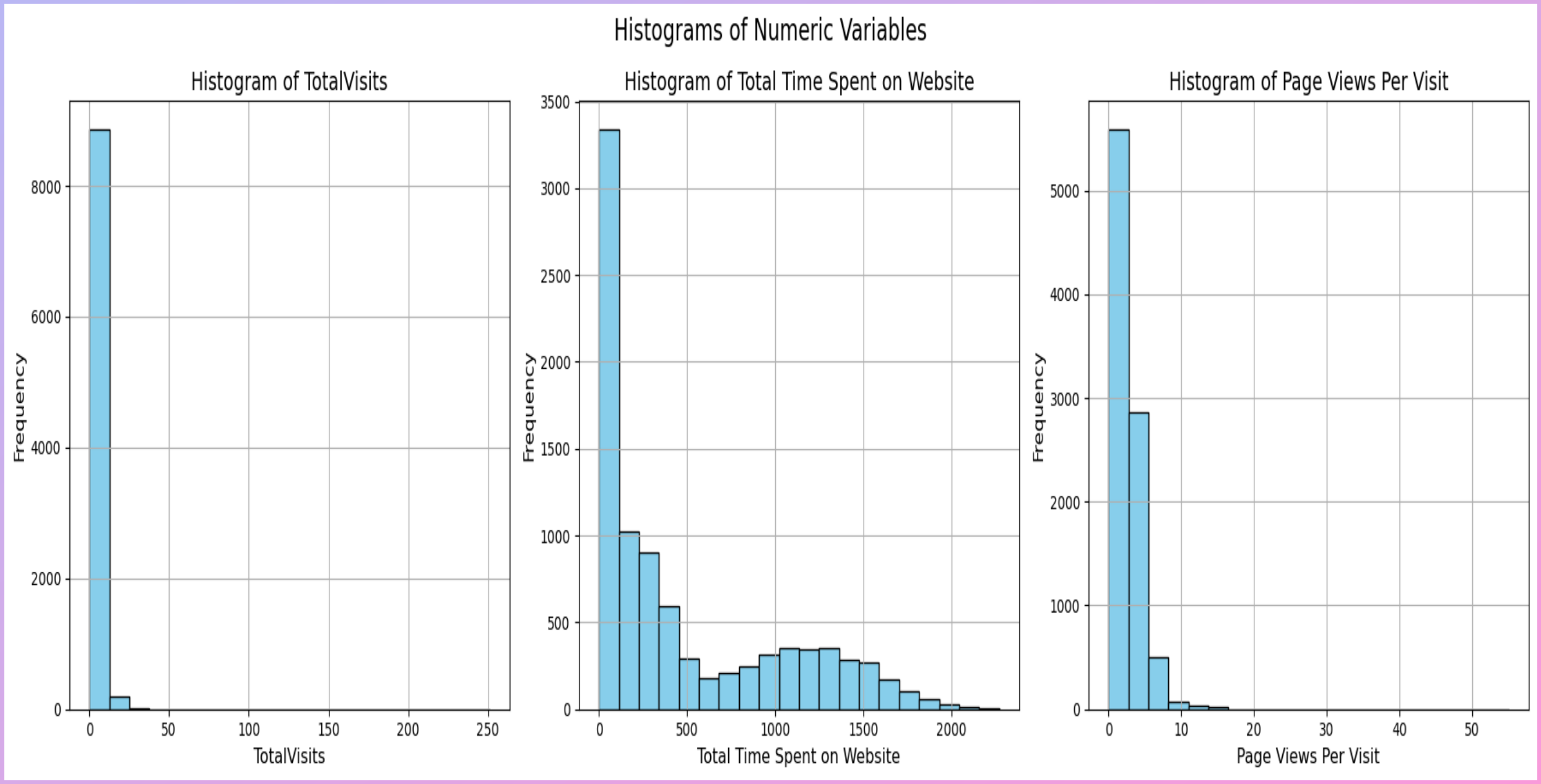


The number of converted leads are (3435, 29)

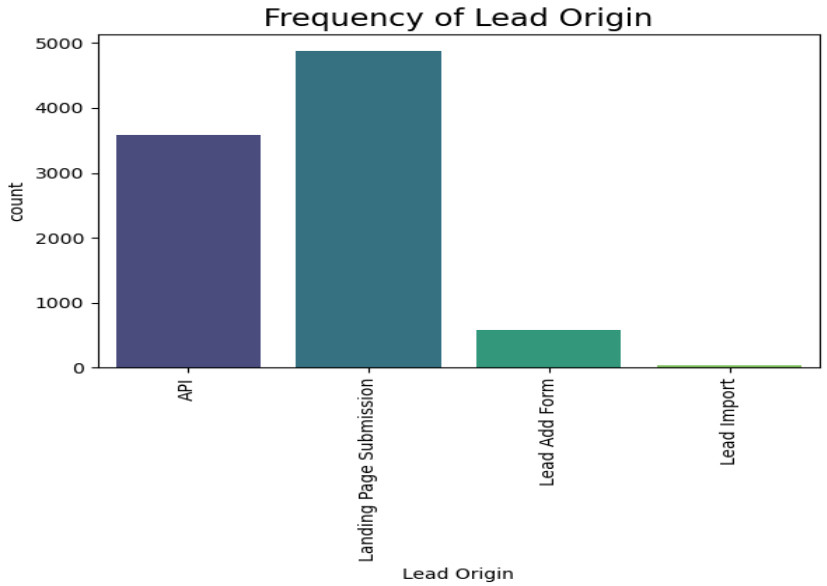
The number of non converted leads are (5639, 29)

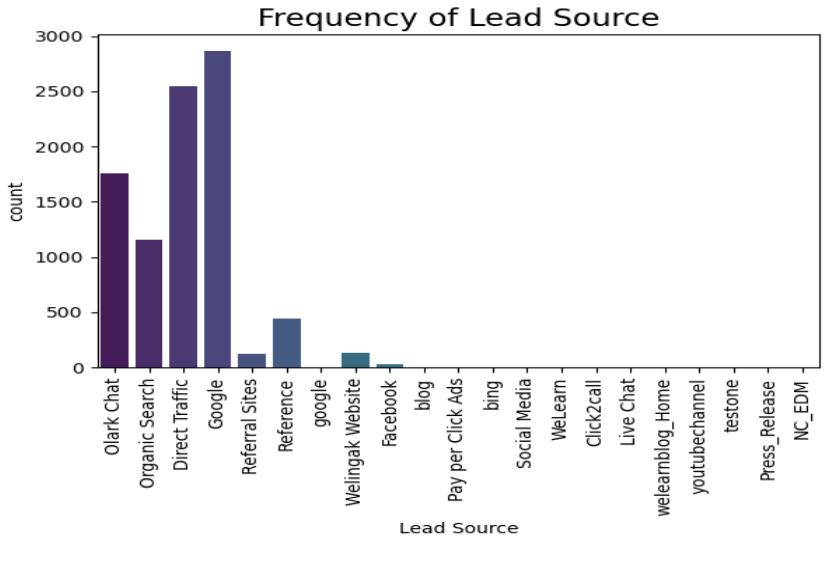
Exploratory Data Analysis

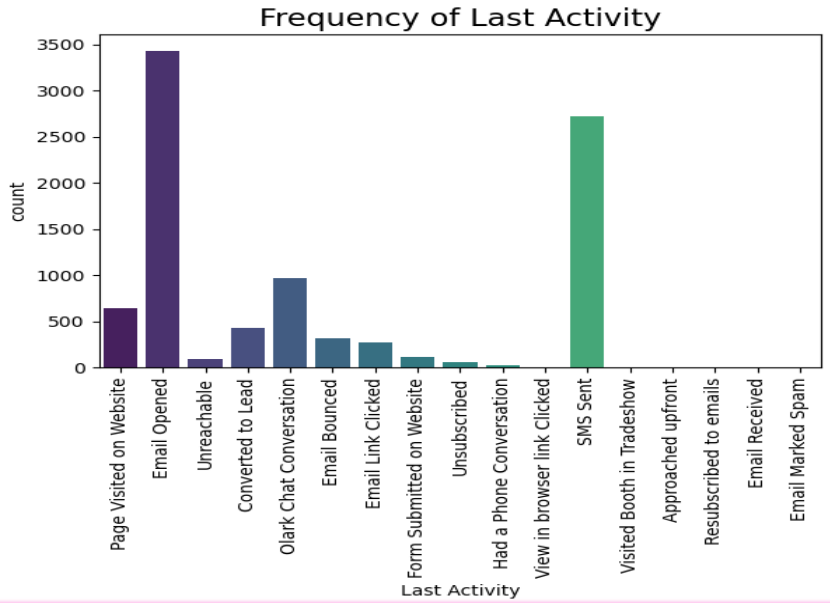
UNIVARIATE ANALYSIS

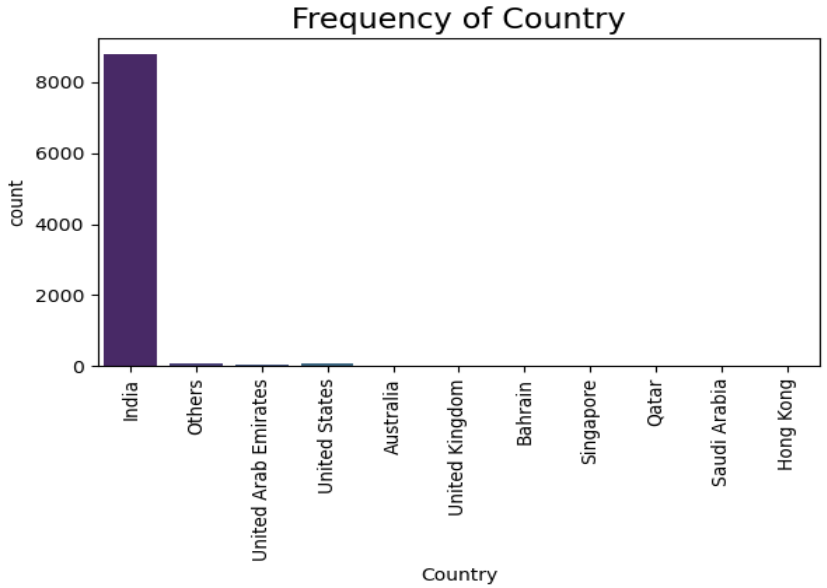

Histograms of Numeric Variables
Histogram of TotalVisits
Frequency
TotalVisits
0
2000
4000
6000
8000
0
50
100
150
200
250
Histogram of Total Time Spent on Website
Frequency
Total Time Spent on Website
0
500
1000
1500
2000
2500
3000
3500
0
500
1000
1500
2000
Histogram of Page Views Per Visit
Frequency
Page Views Per Visit
0
1000
2000
3000
4000
5000
0
10
20
30
40
50

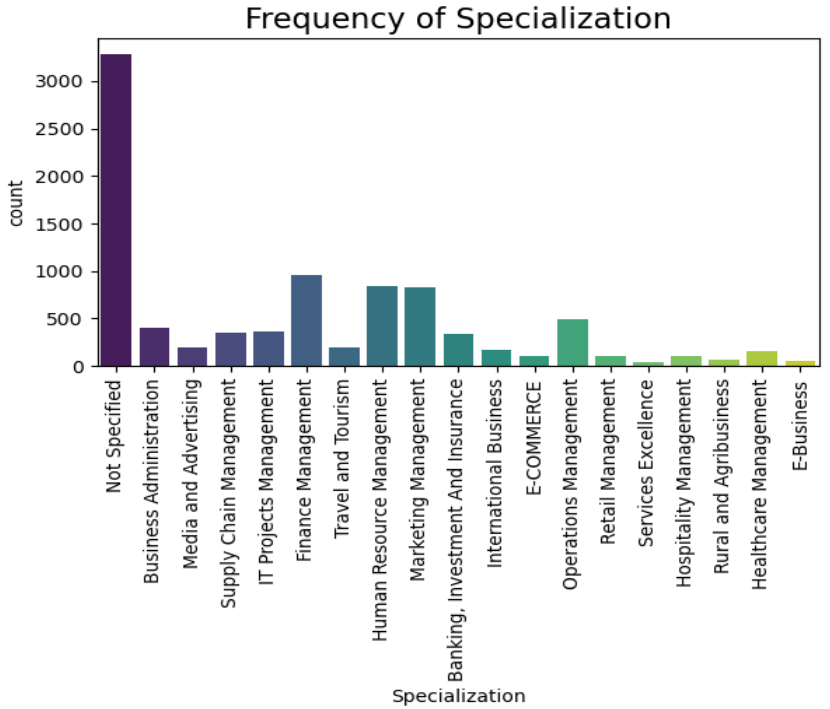
Bar charts for categorical variables

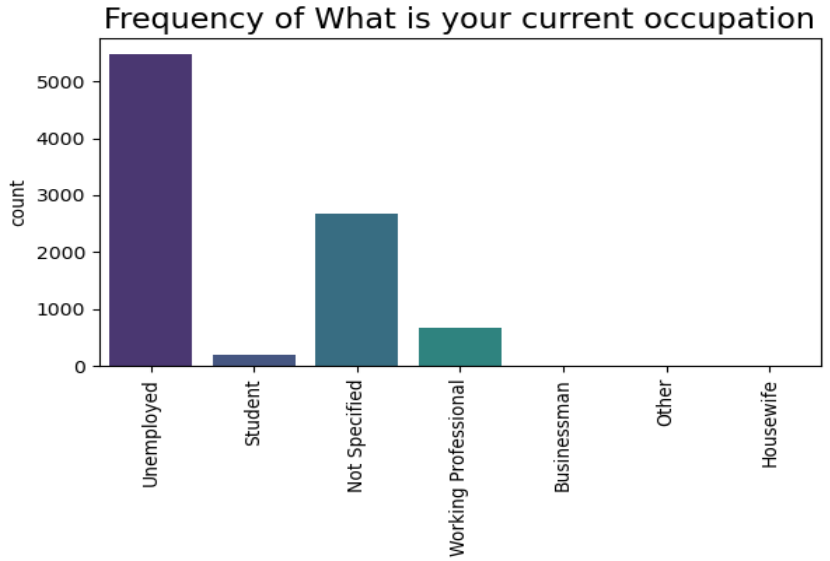




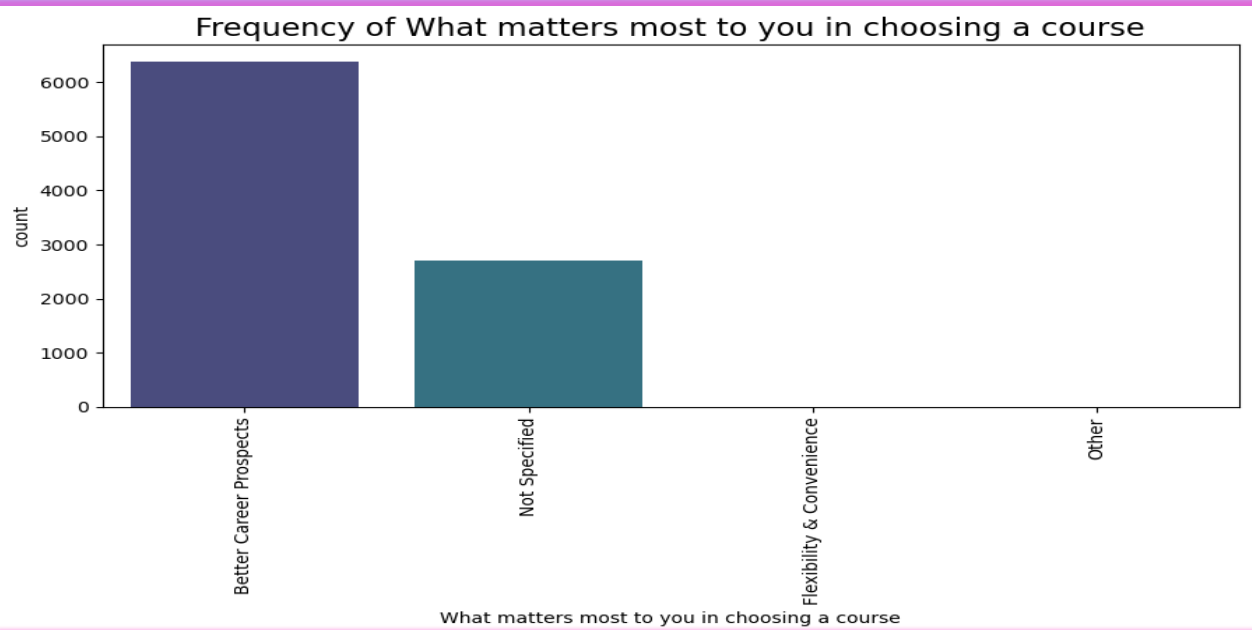


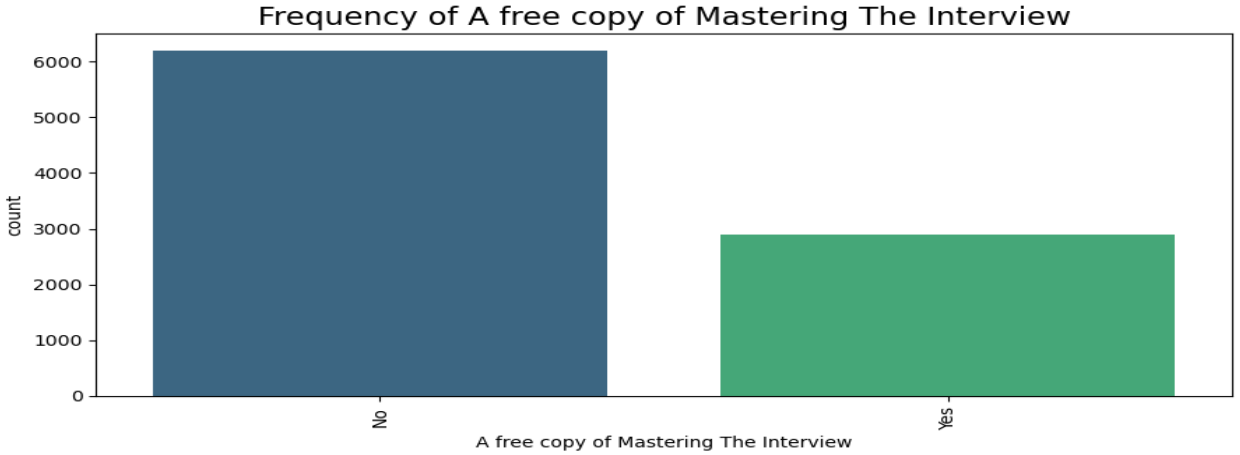


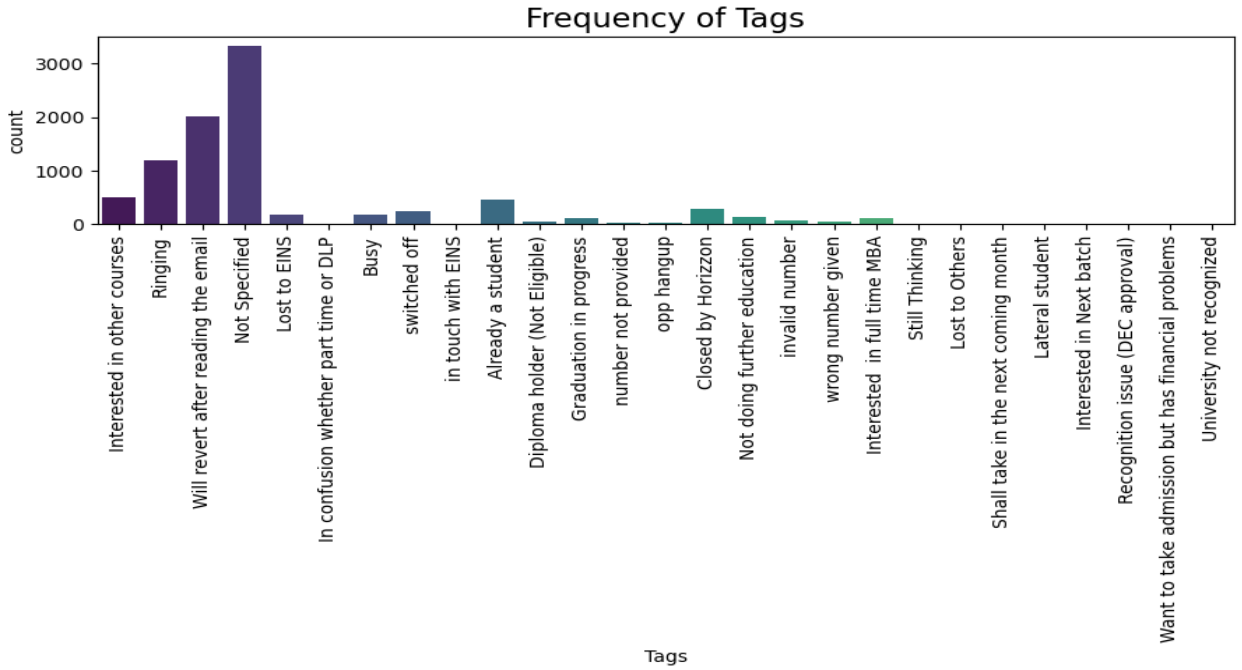


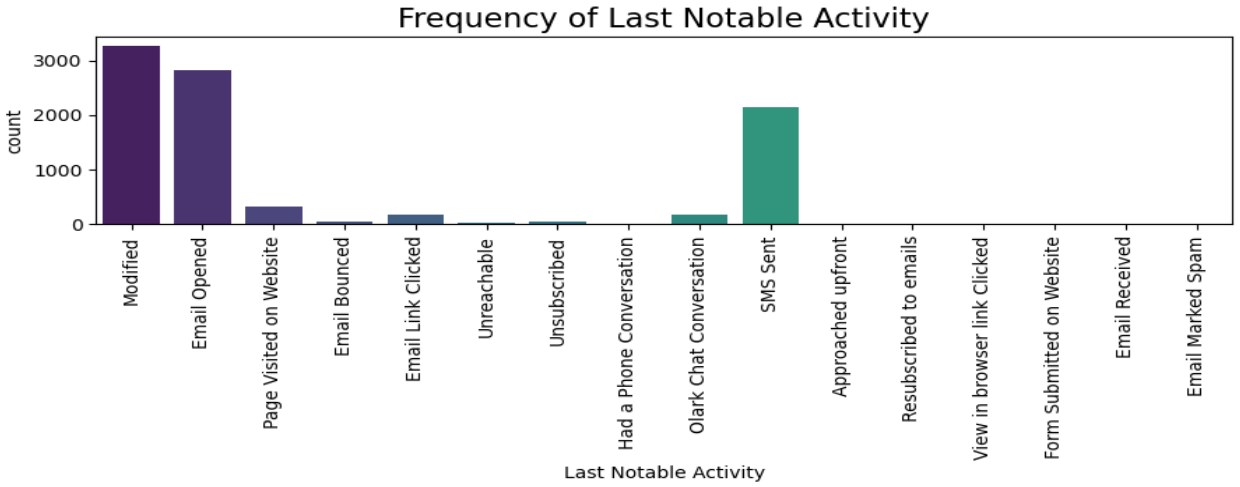


Bar charts for categorical variables



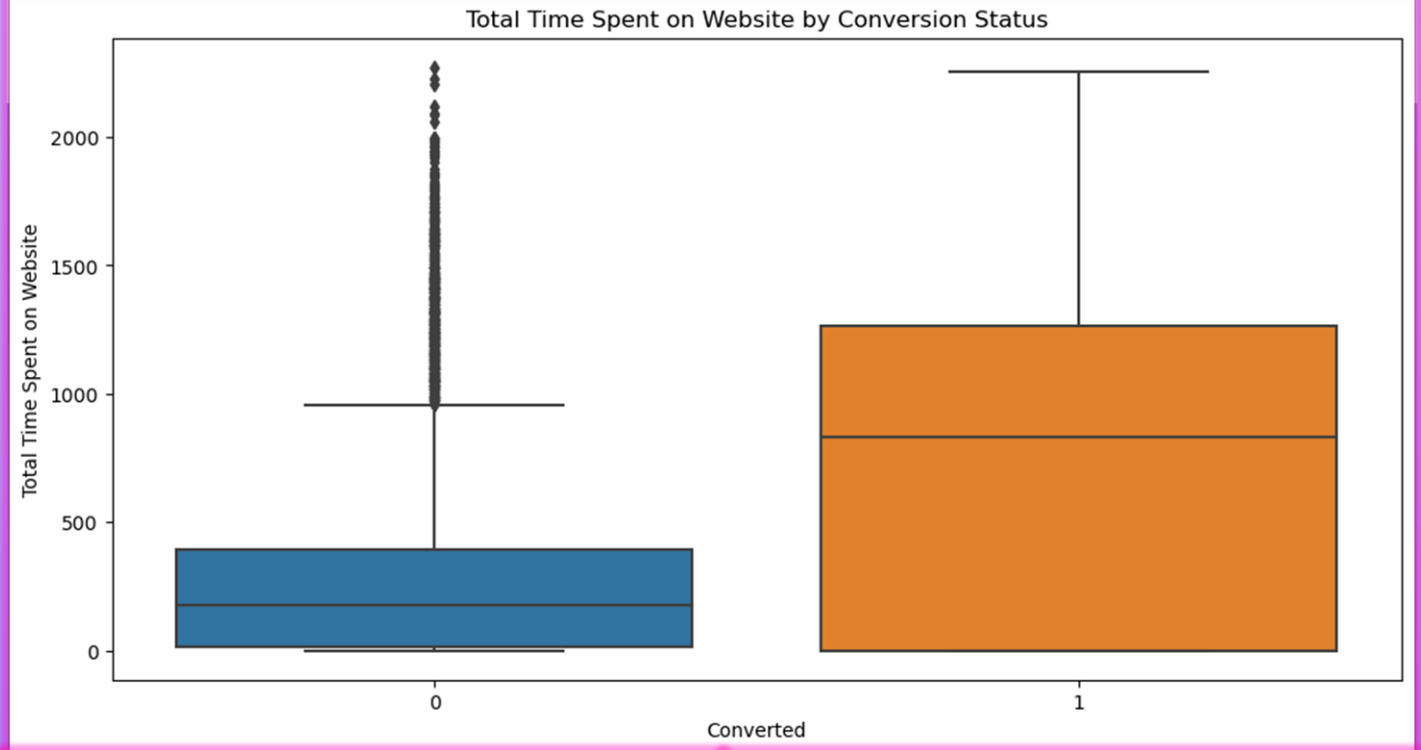




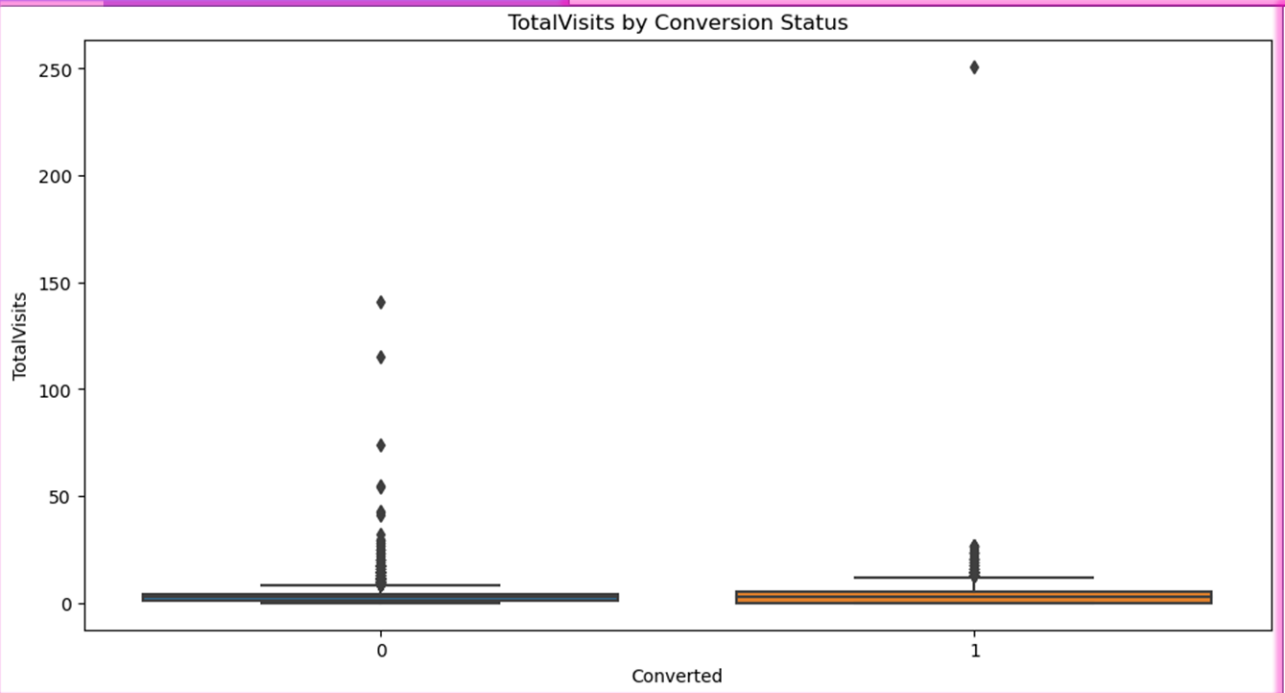


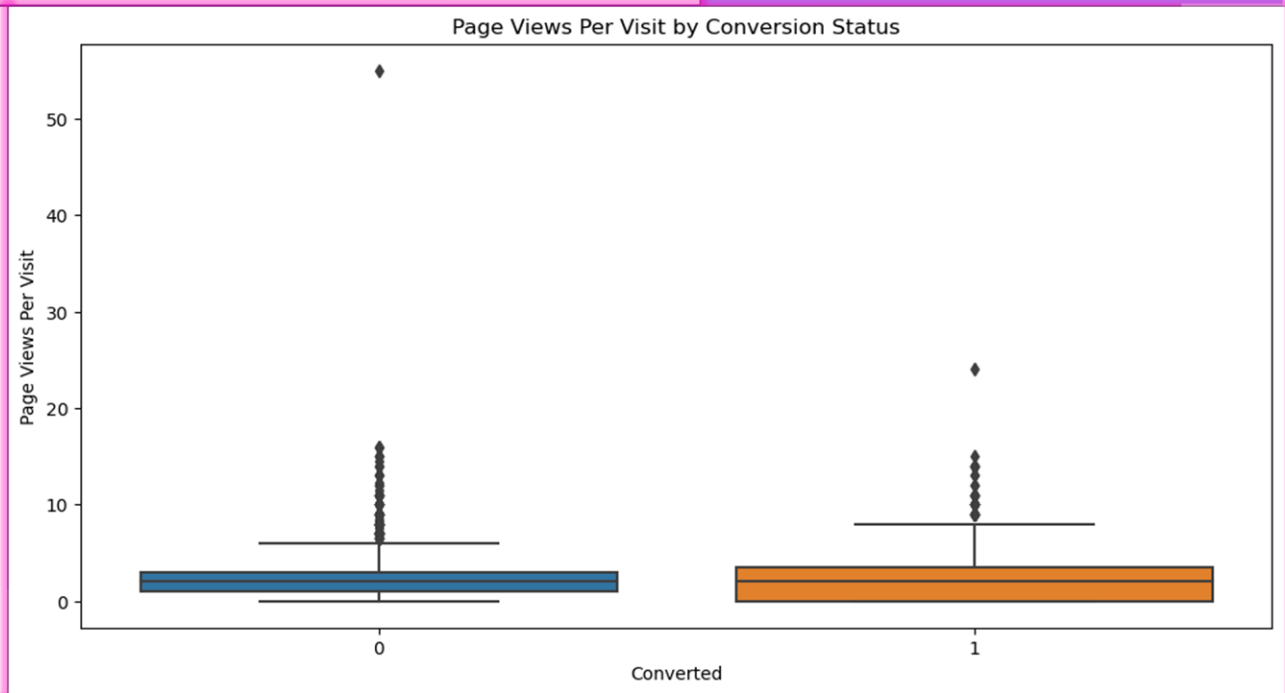
BIVARIATE ANALYSIS

Box plots for numerical variables

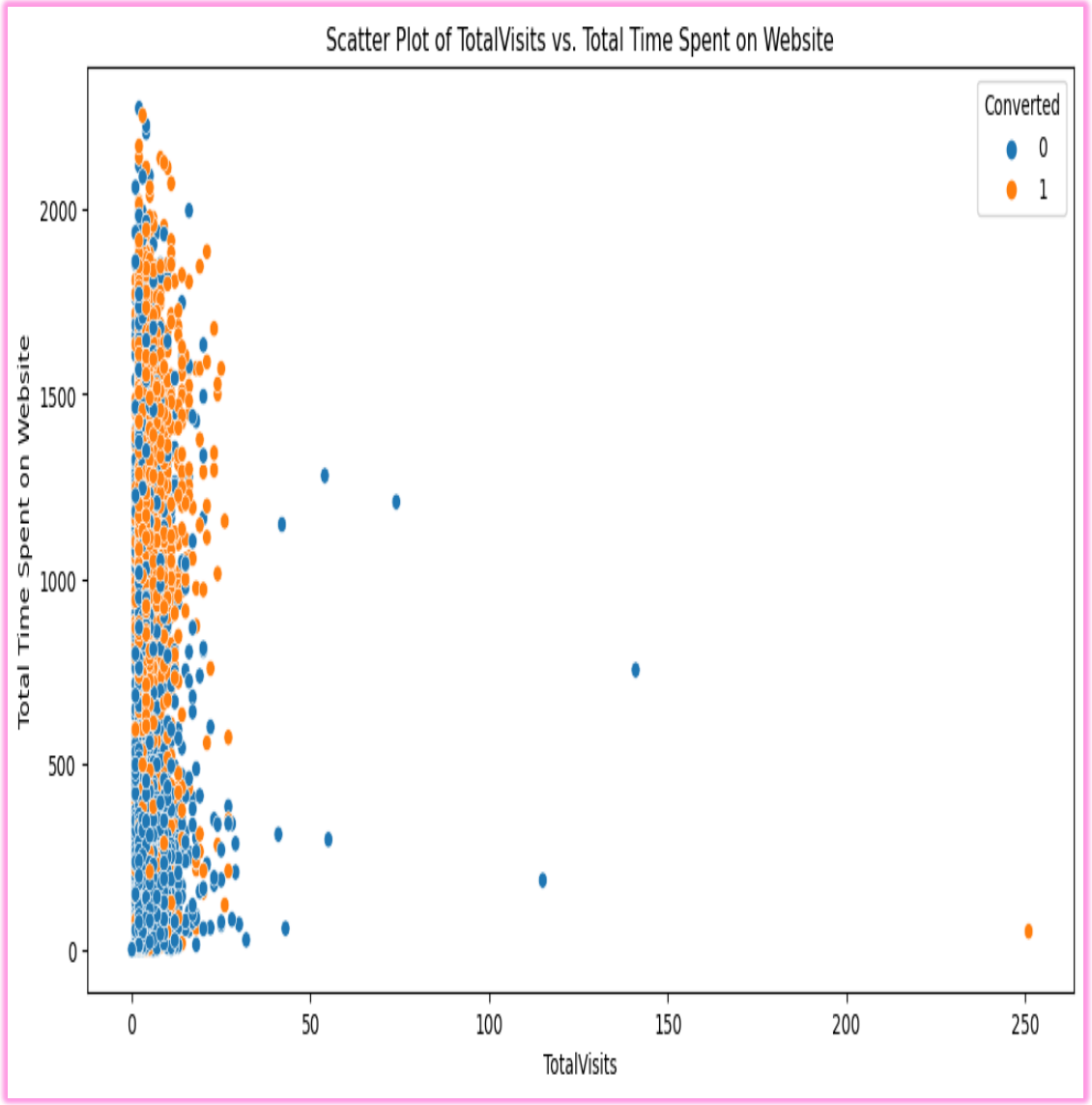


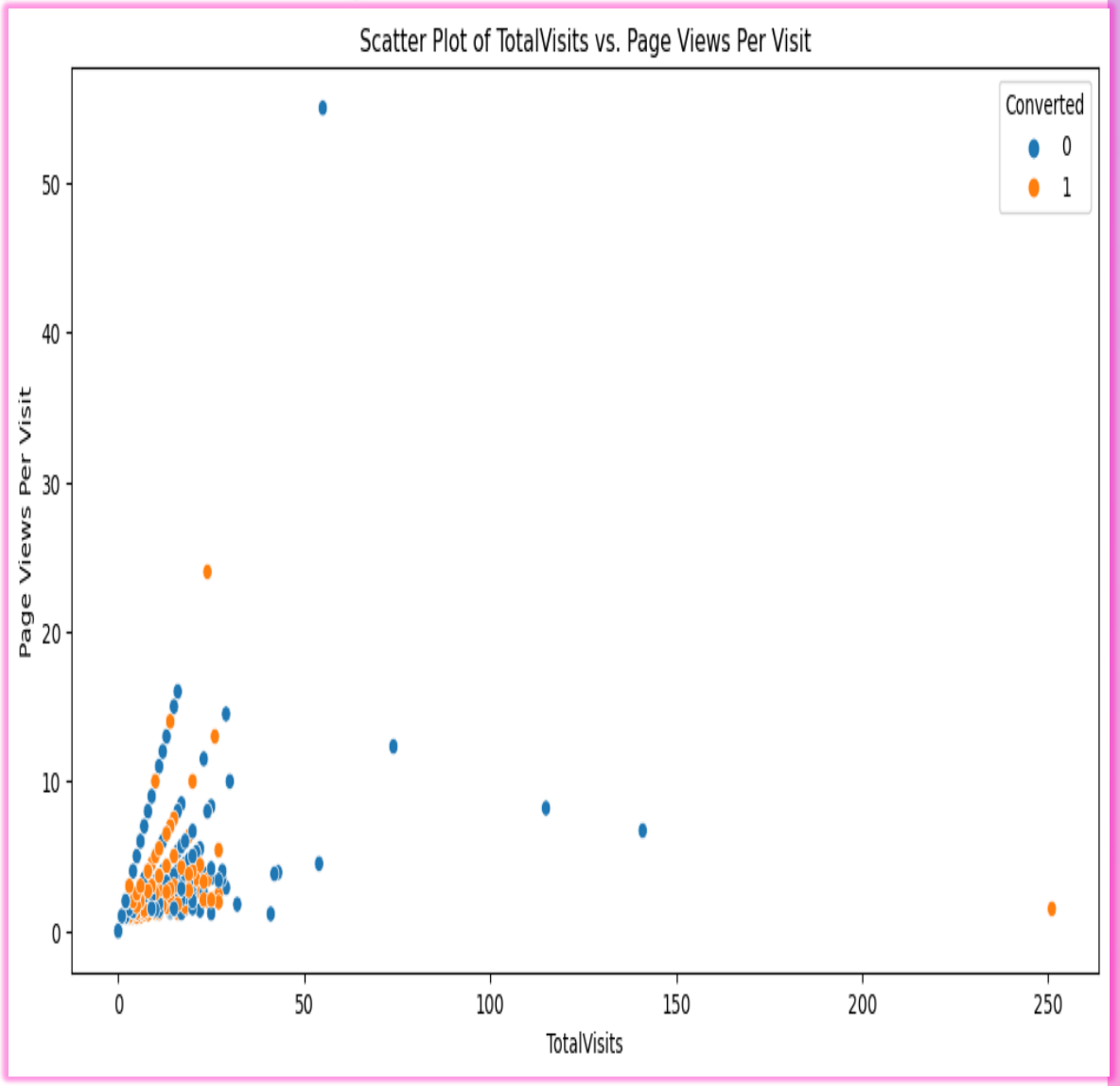
Against Conversion status





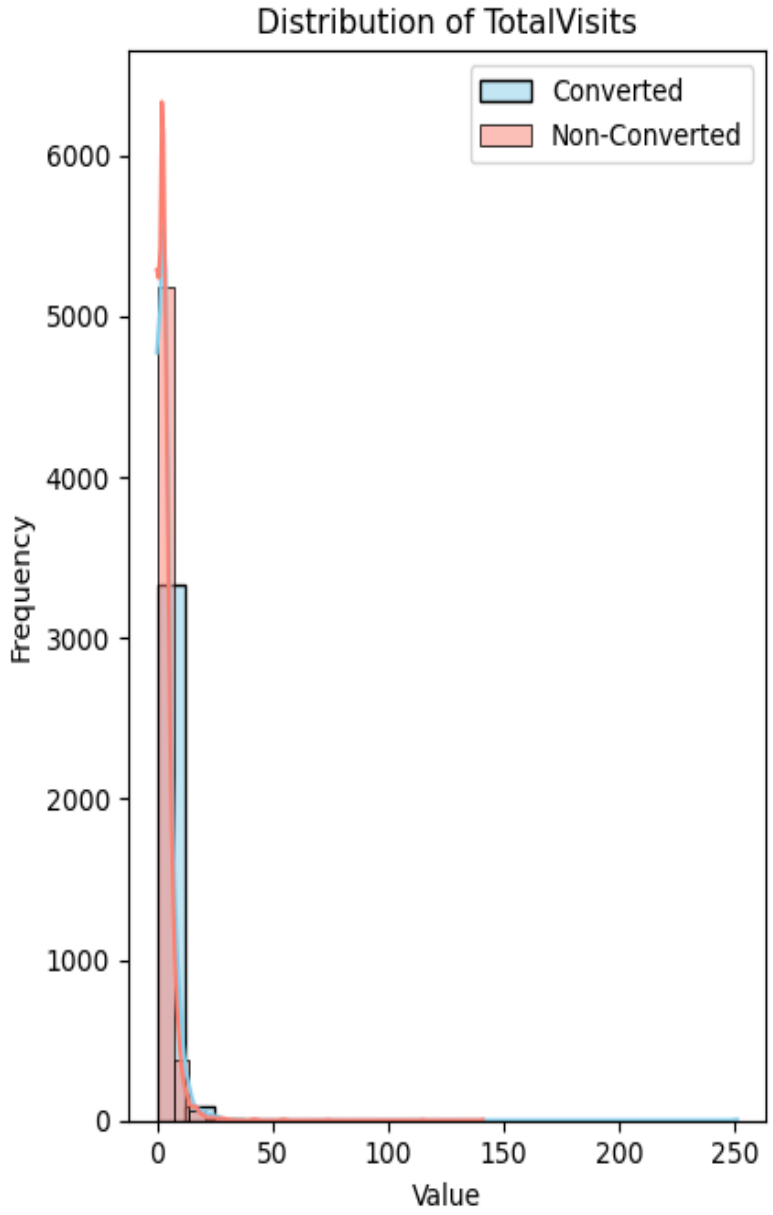
Scatter plots for numerical variables against the target variable *'Converted'*

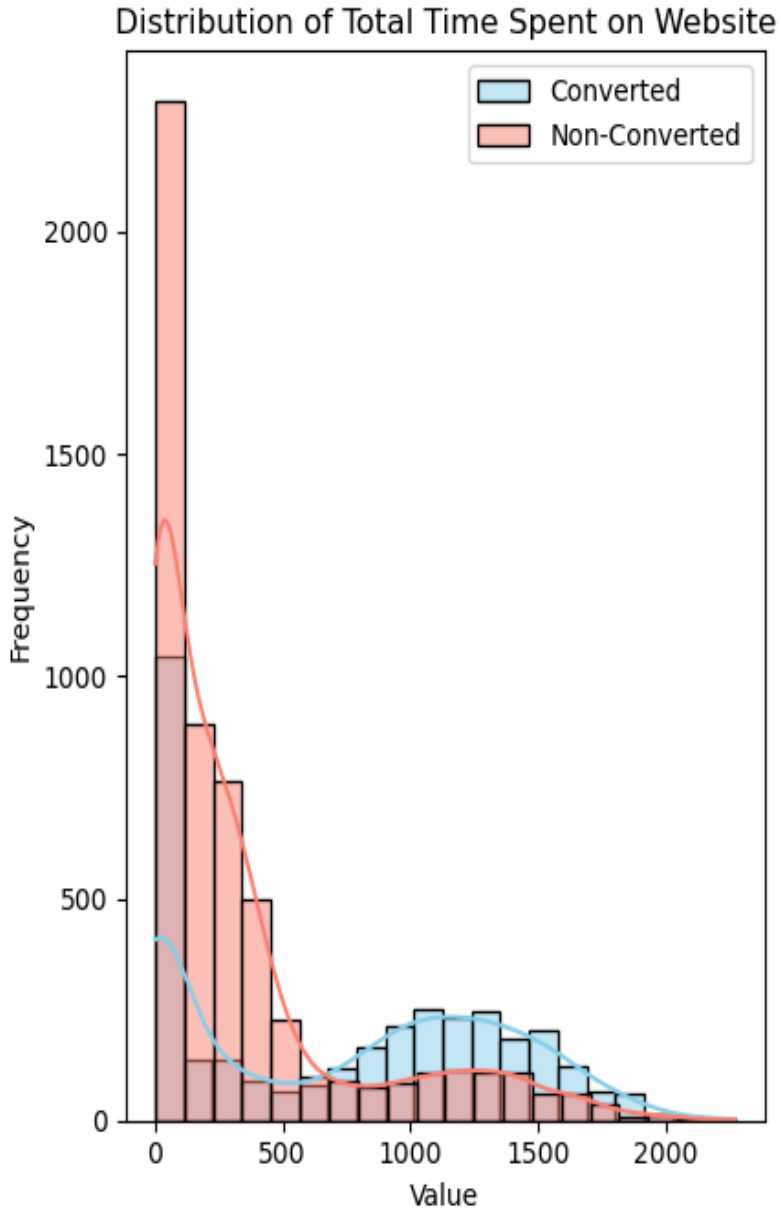


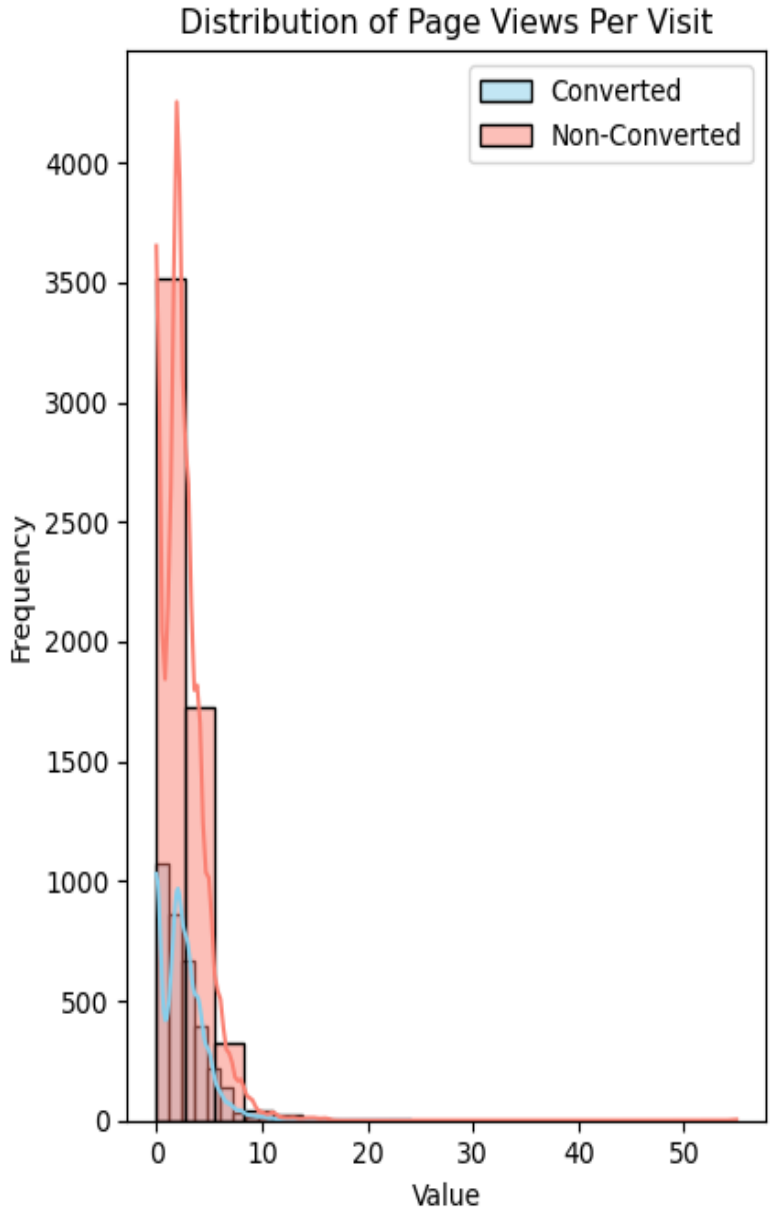


SEGMENTED ANALYSIS

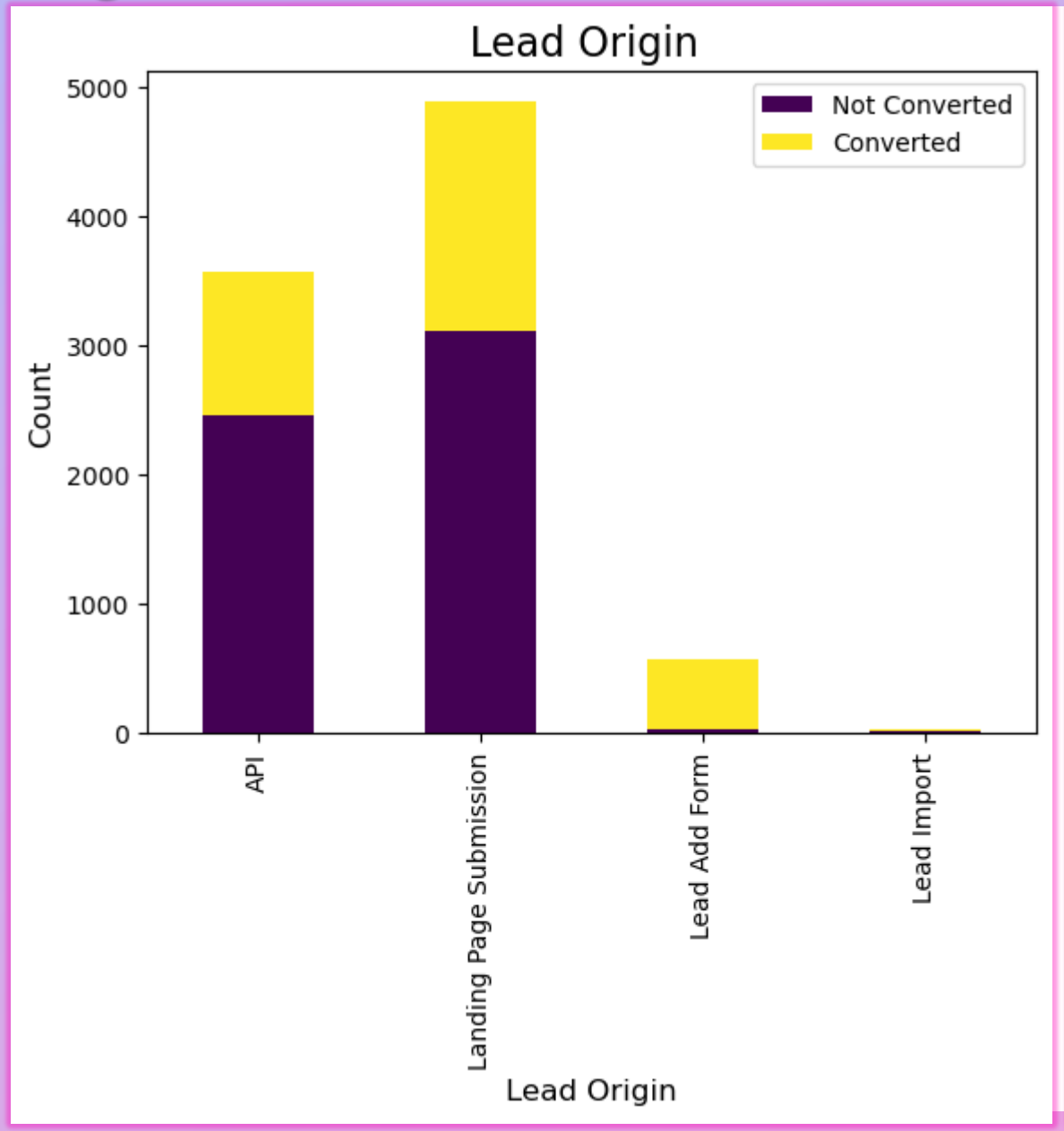
Numeric variables

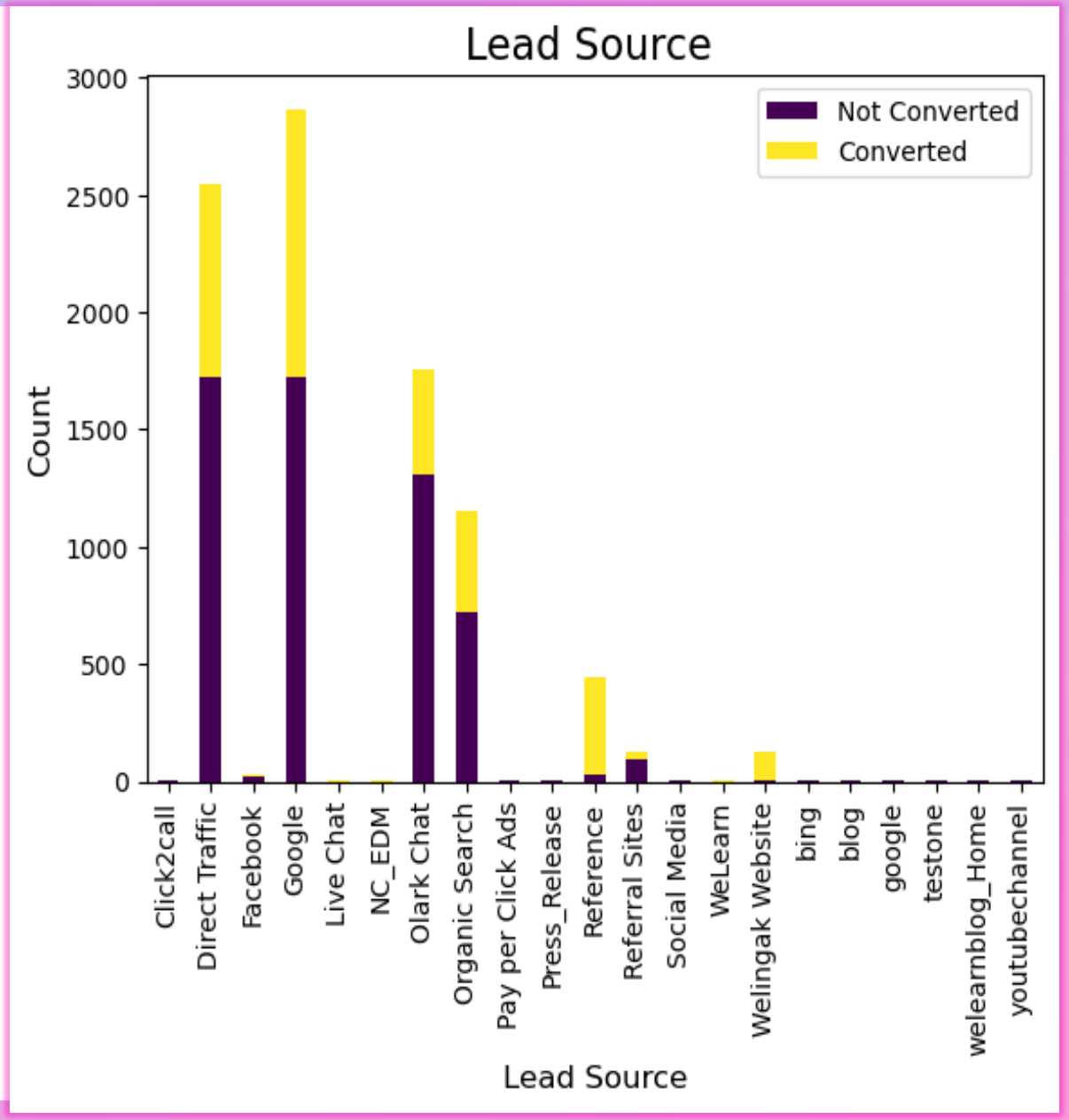




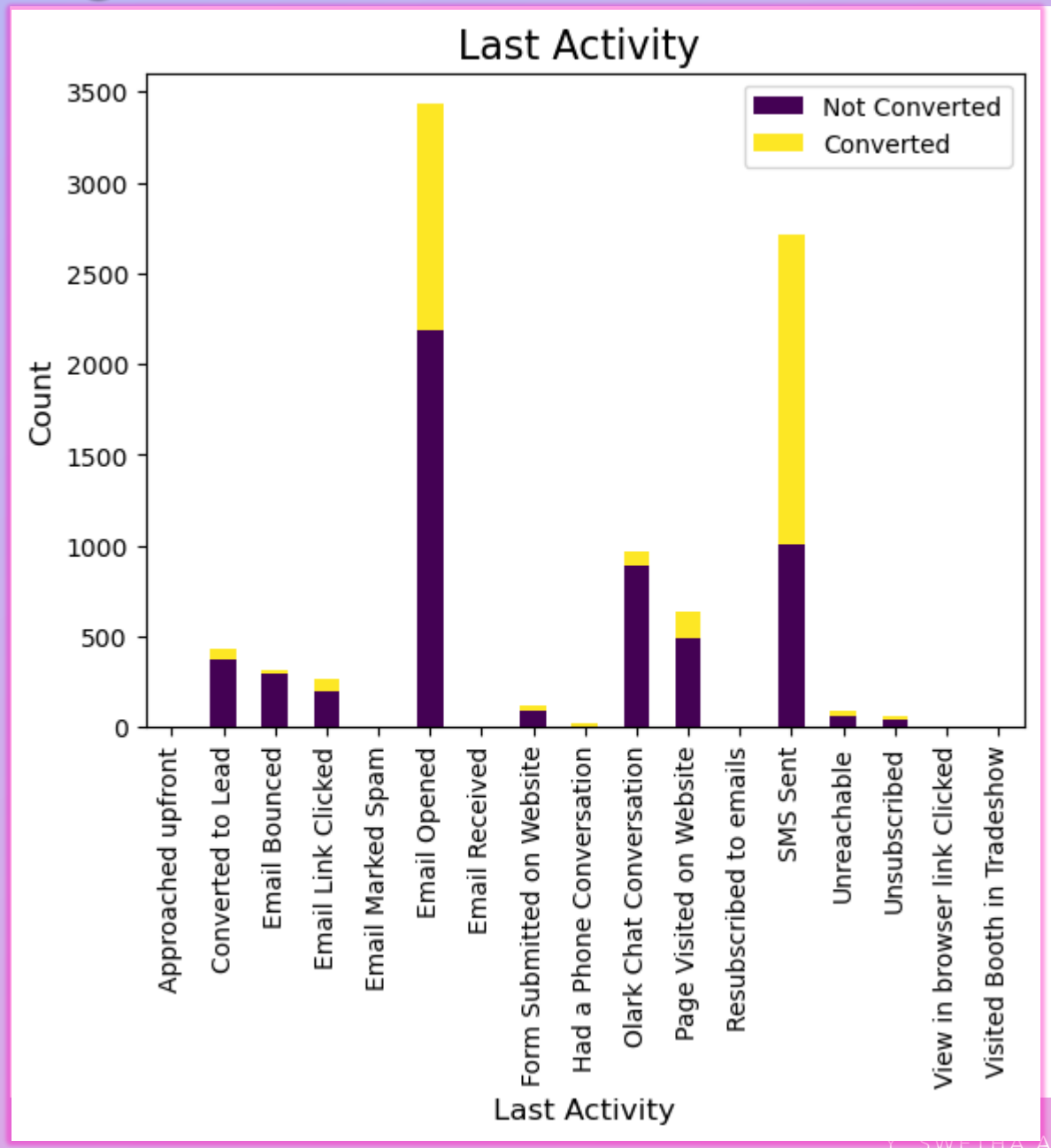


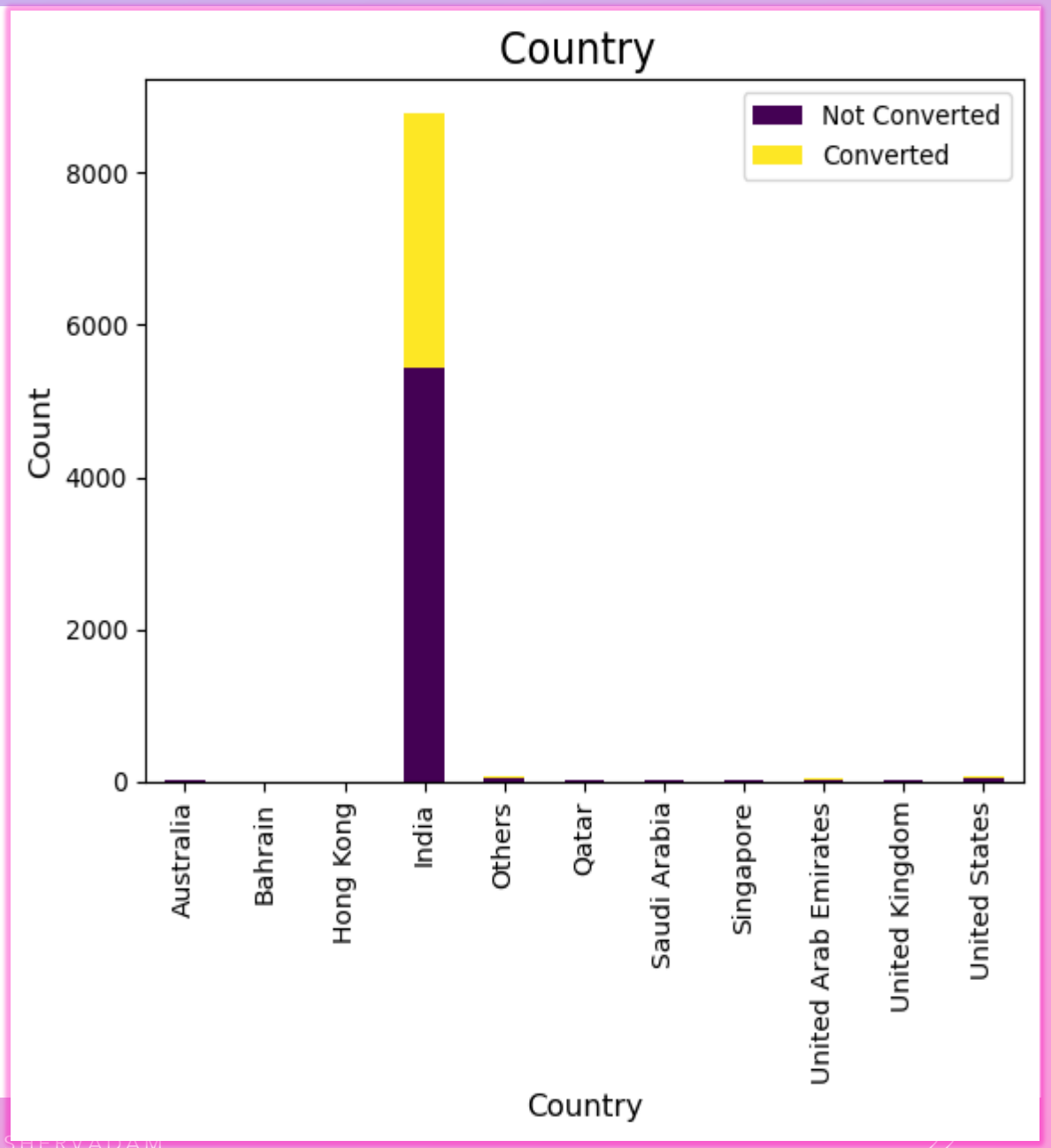
Categorical variables



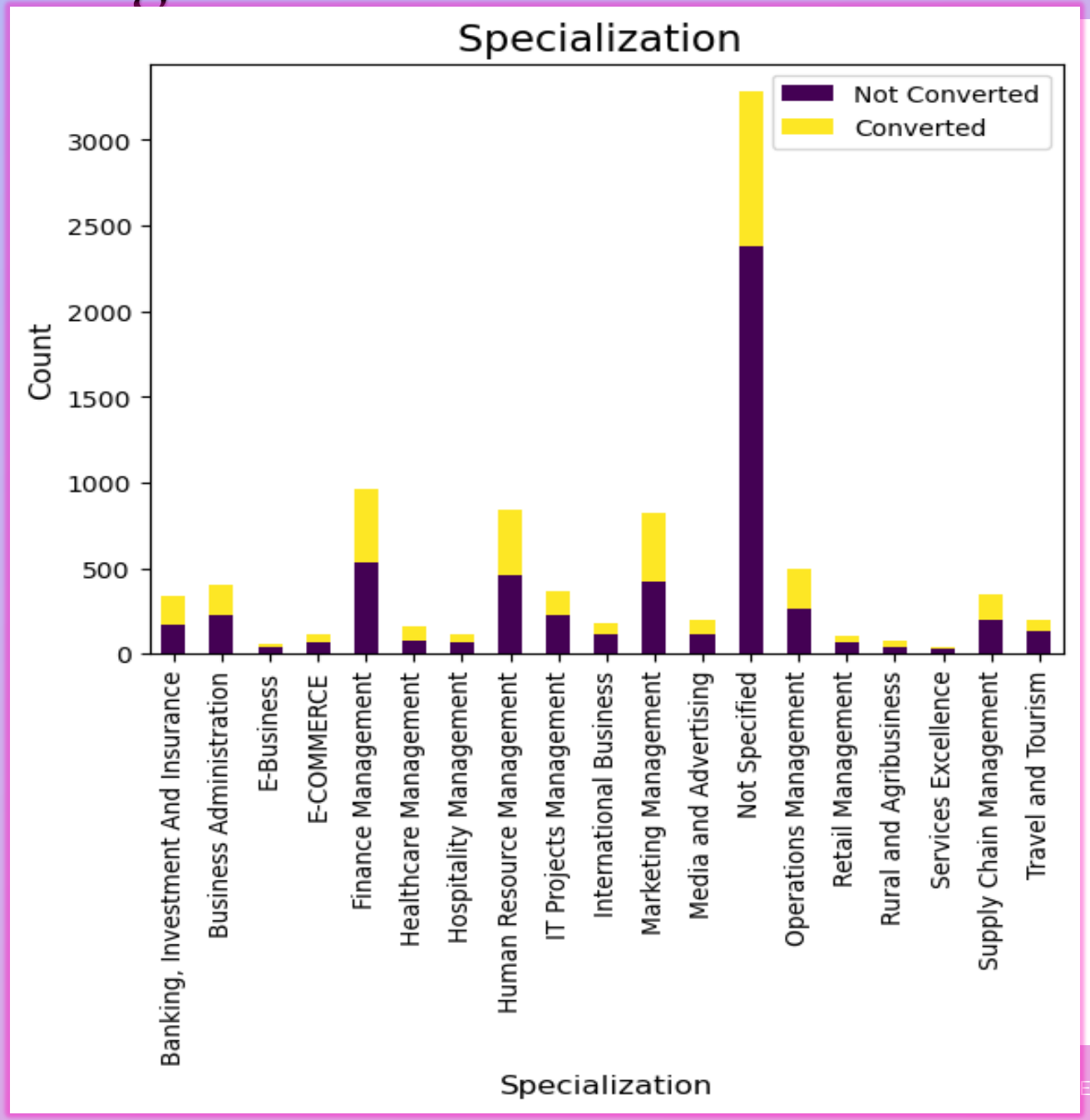


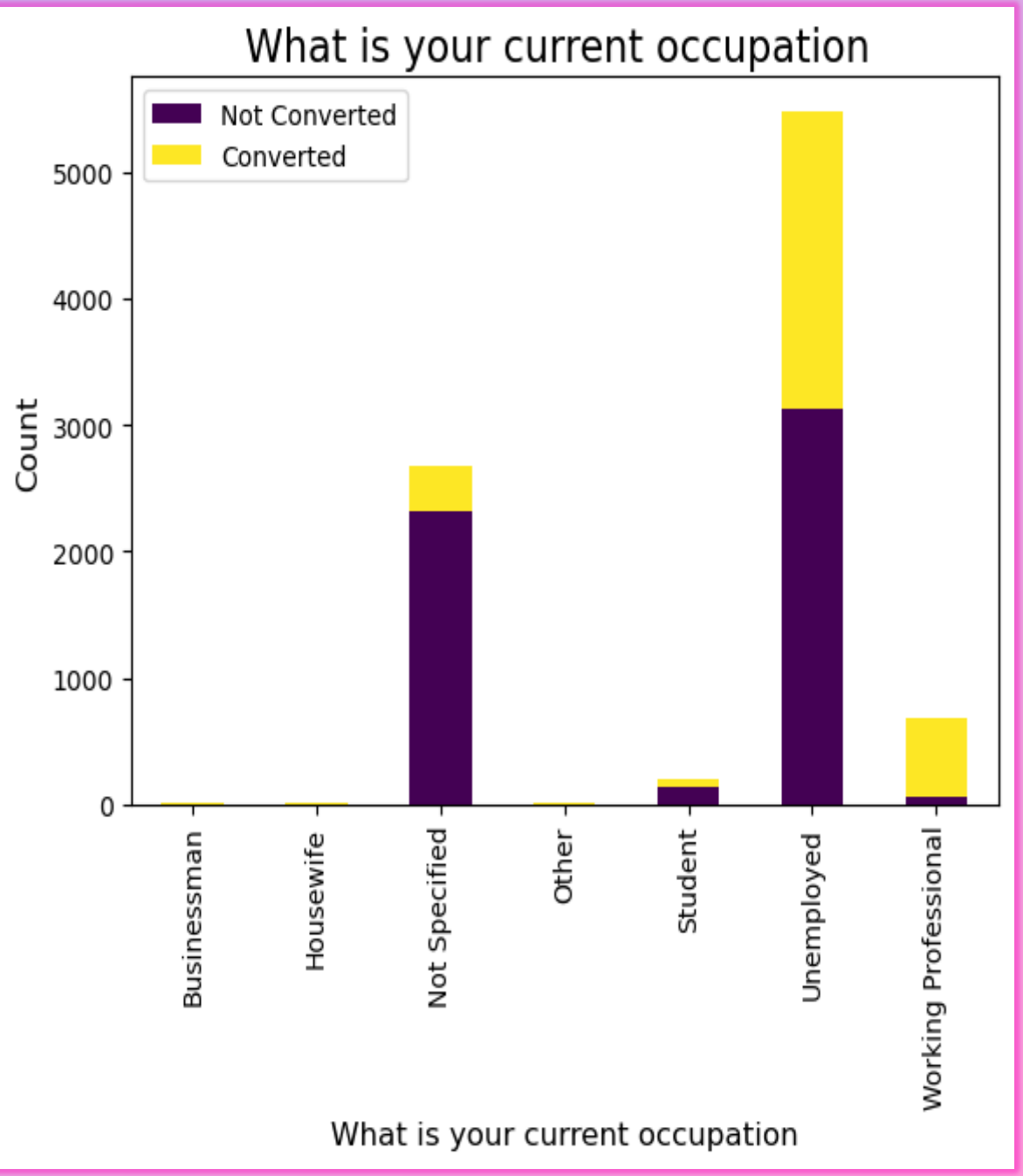
Categorical variables



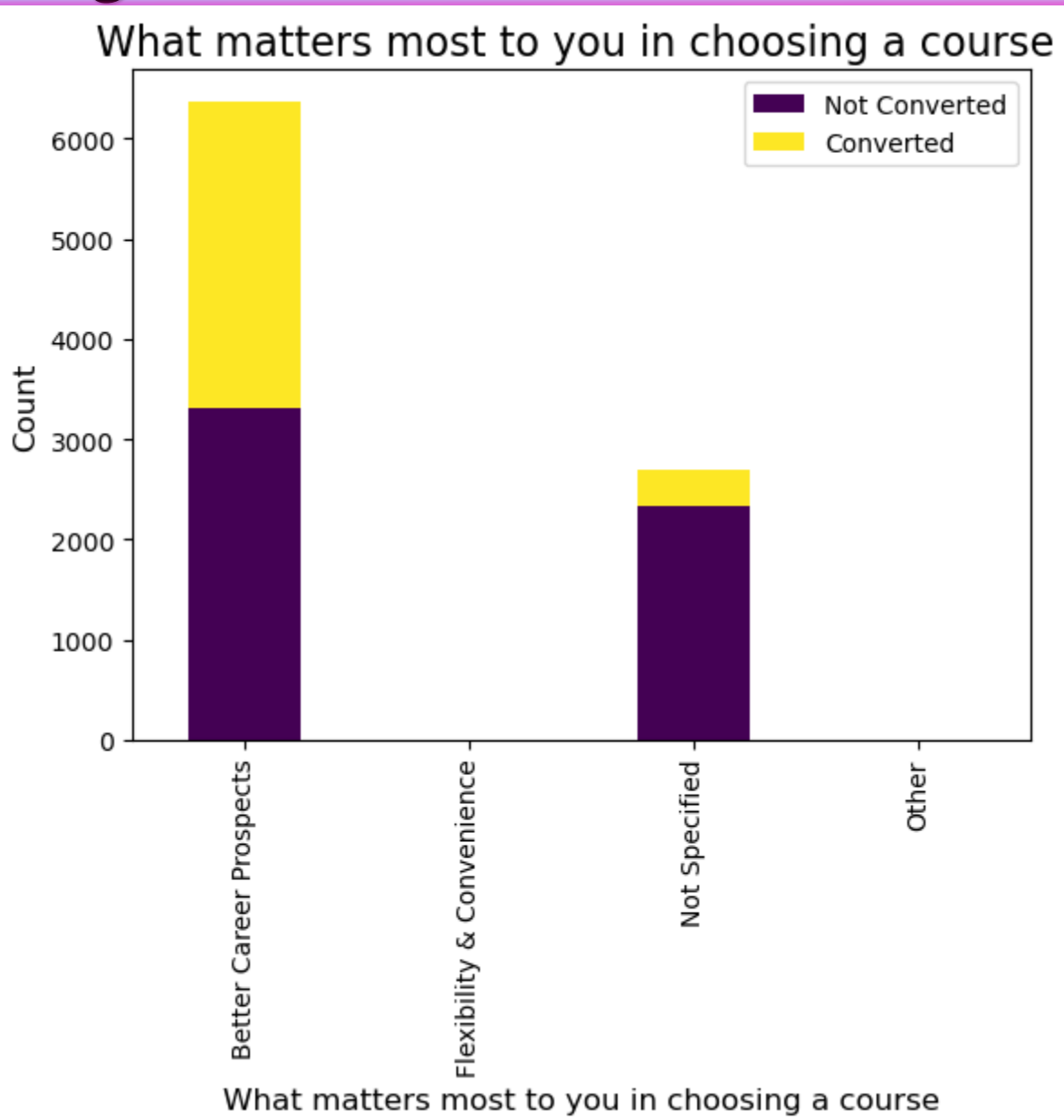


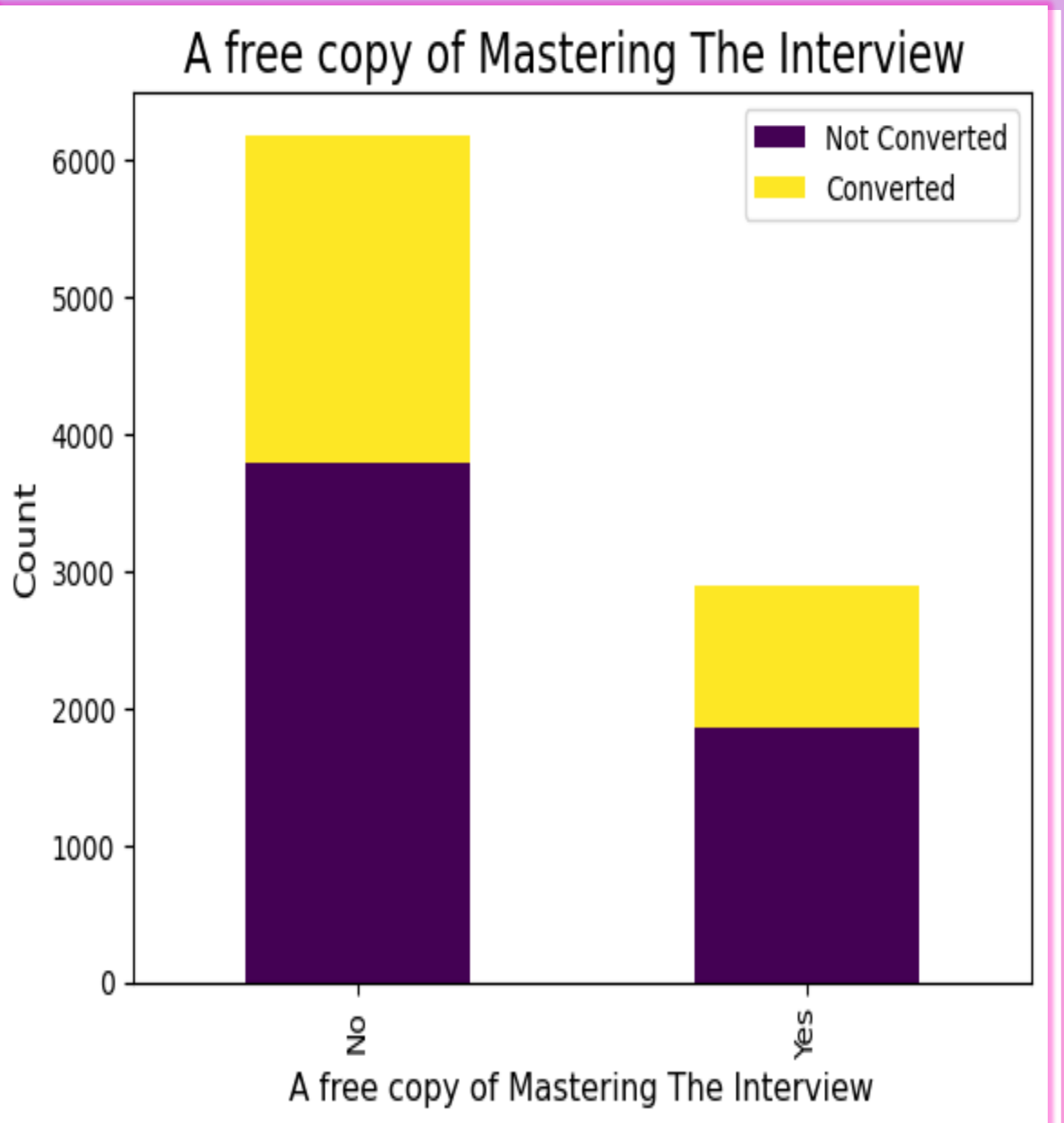
Categorical variables

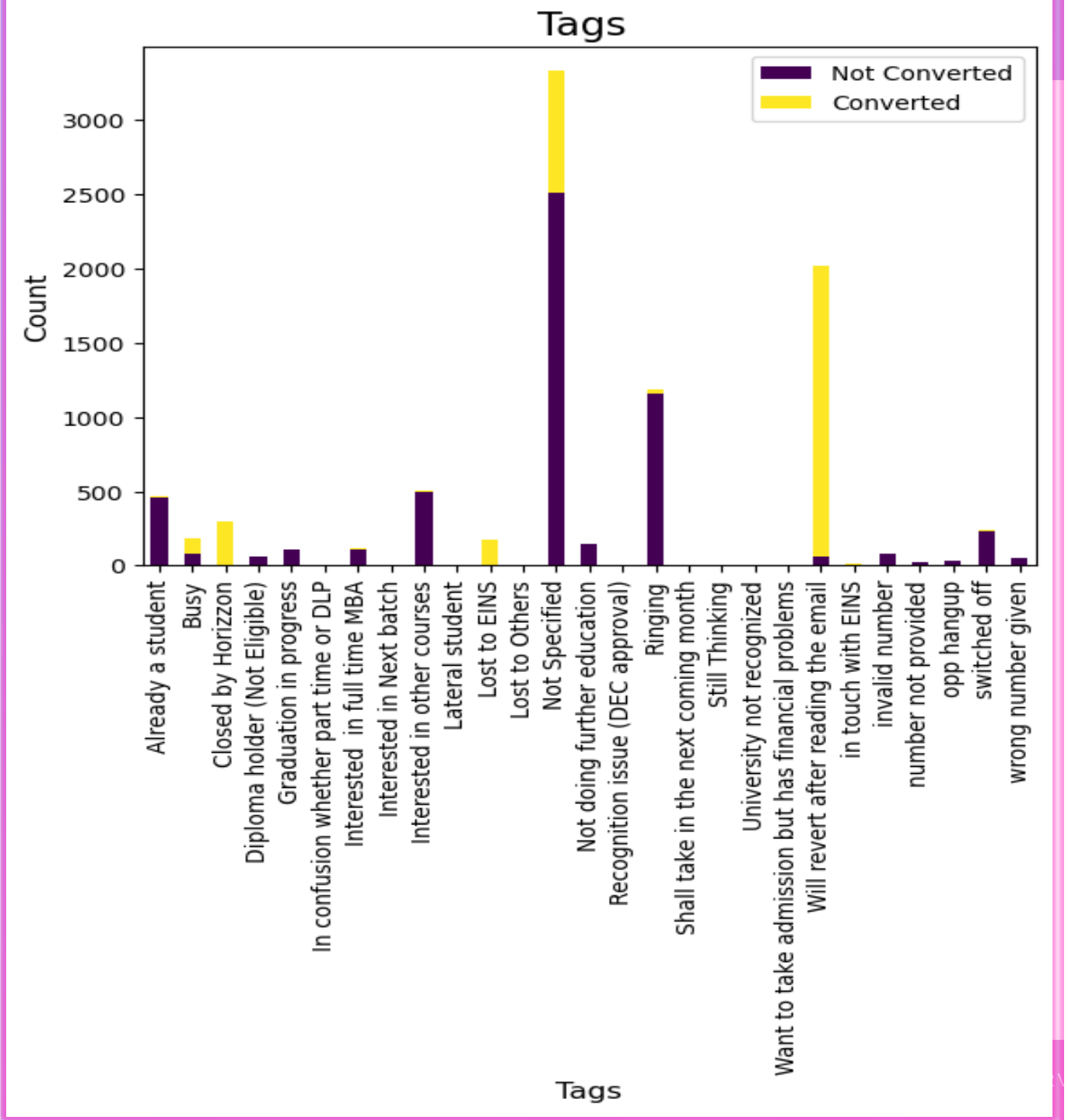


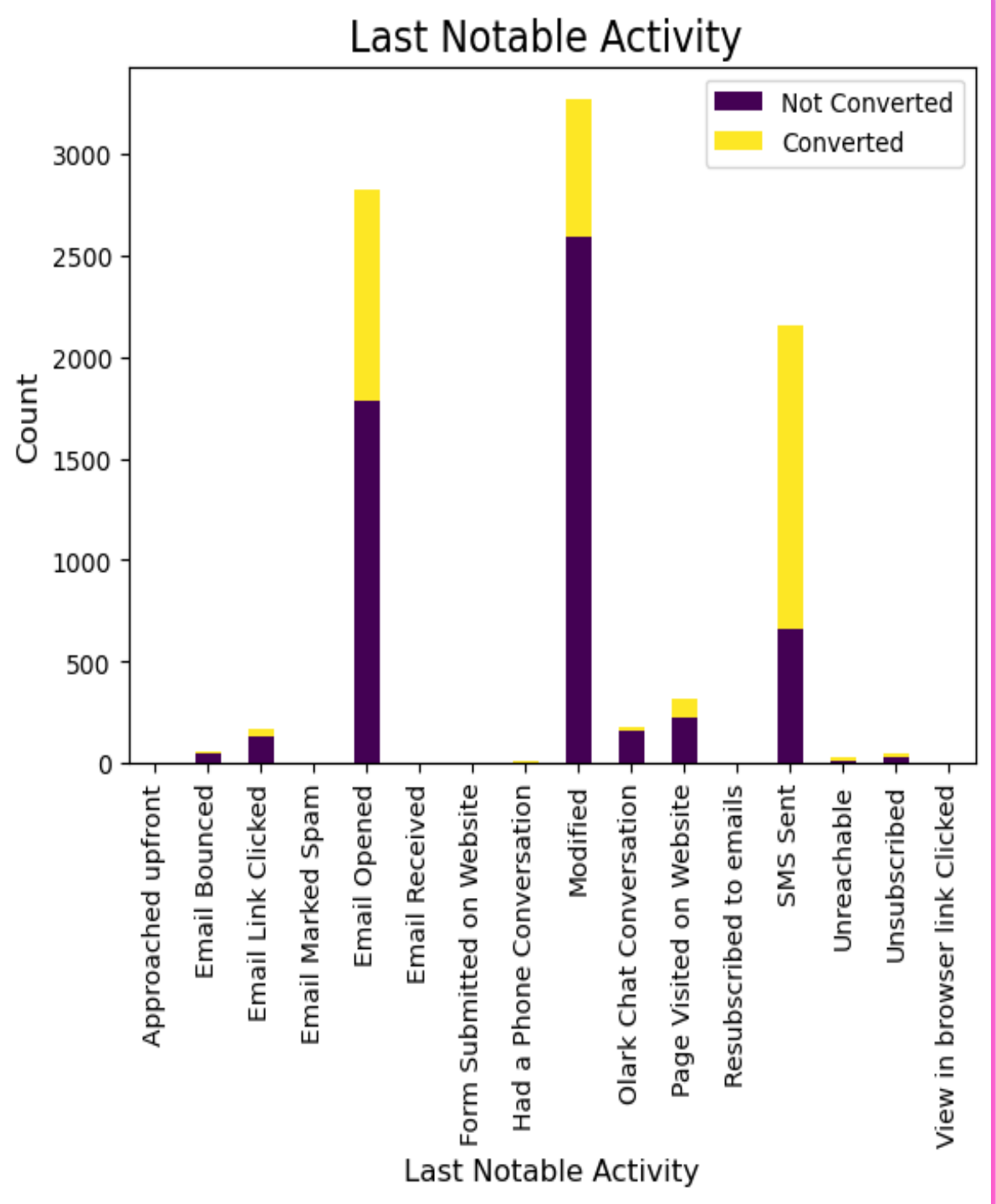


Categorical variables

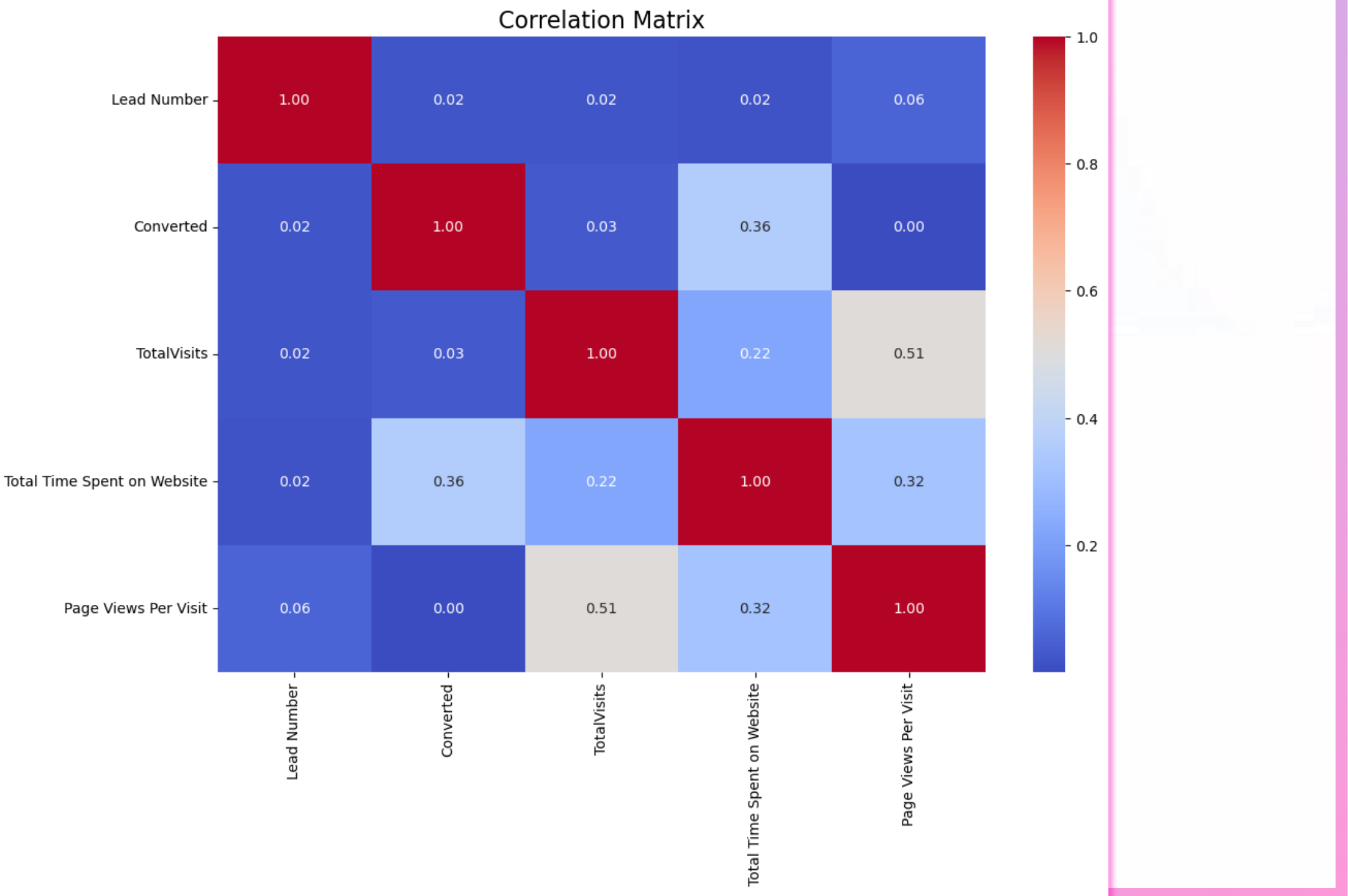


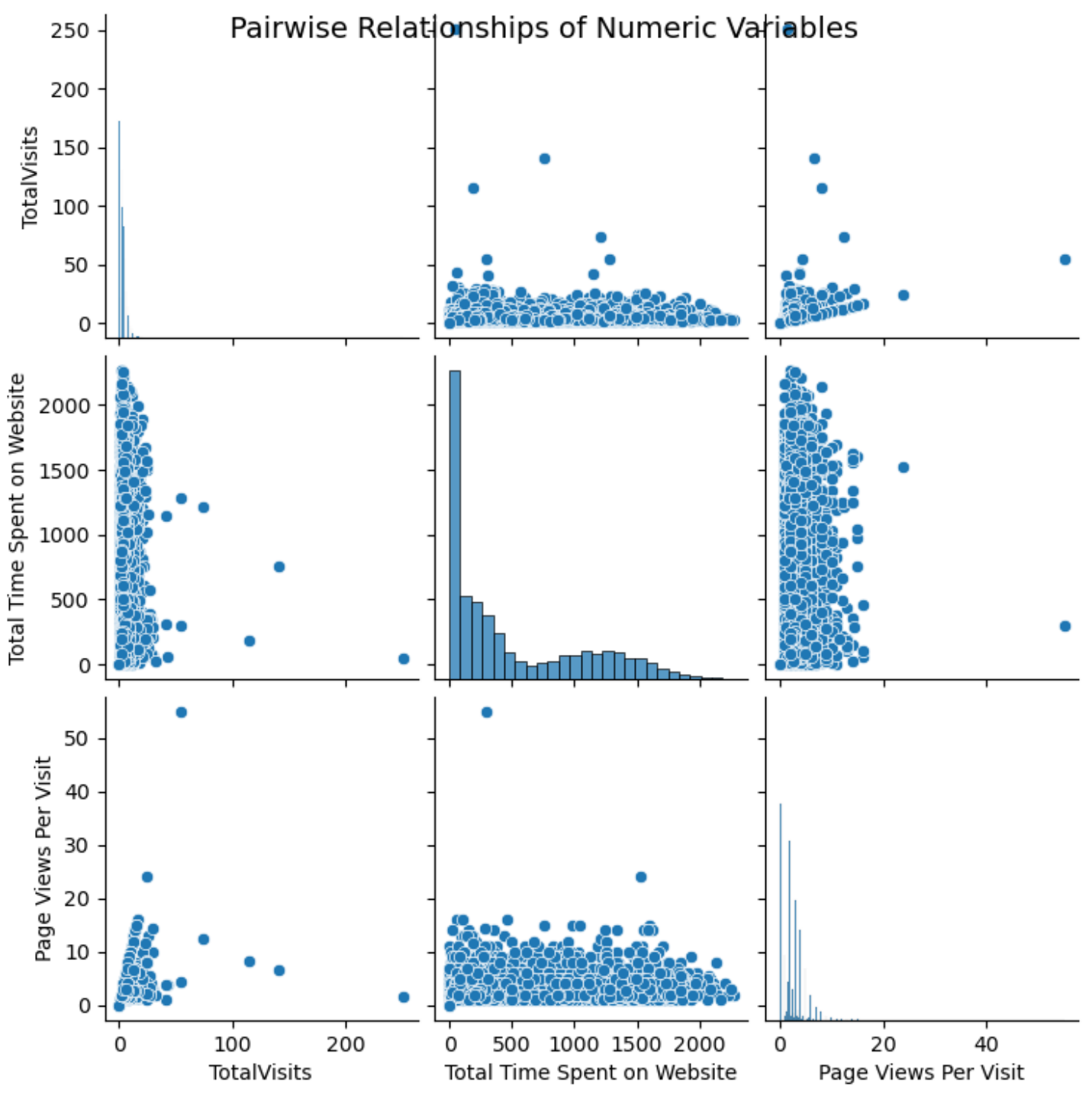



Tags
Not Converted
Converted
Count
3000
2500
2000
1500
1000
500
0
Already a student
Busy
Closed by Horizzon
Diploma holder (Not Eligible)
Graduation in progress
In confusion whether part time or DLP
Interested in full time MBA
Interested in Next batch
Interested in other courses
Lateral student
Lost to EINS
Lost to Others
Not Specified
Not doing further education
Recognition issue (DEC approval)
Ringing
Shall take in the next coming month
Still Thinking
University not recognized
Want to take admission but has financial problems
Will revert after reading the email
in touch with EINS
invalid number
number not provided
opp hangup
switched off
wrong number given
Tags


Last Notable Activity
Not Converted
Converted
Count
3000
2500
2000
1500
1000
500
0
Approached upfront
Email Bounced
Email Link Clicked
Email Marked Spam
Email Opened
Email Received
Form Submitted on Website
Had a Phone Conversation
Modified
Olark Chat Conversation
Page Visited on Website
Resubscribed to emails
SMS Sent
Unreachable
Unsubscribed
View in browser link Clicked
Last Notable Activity

MULTIVARIATE ANALYSIS


Correlation Matrix
Lead Number
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.00 0.02 0.02 0.02 0.06
0.02 1.00 0.03 0.36 0.00
0.02 0.03 1.00 0.22 0.51
0.02 0.36 0.22 1.00 0.32
0.06 0.00 0.51 0.32 1.00
Lead Number
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2


Pairwise Relationships of Numeric Variables
TotalVisits
Total Time Spent on Website
Page Views Per Visit
0
50
100
150
200
250
500
1000
1500
2000
20
30
40

DATA CONVERSION

Converted the following columns to binary variables (0,1)

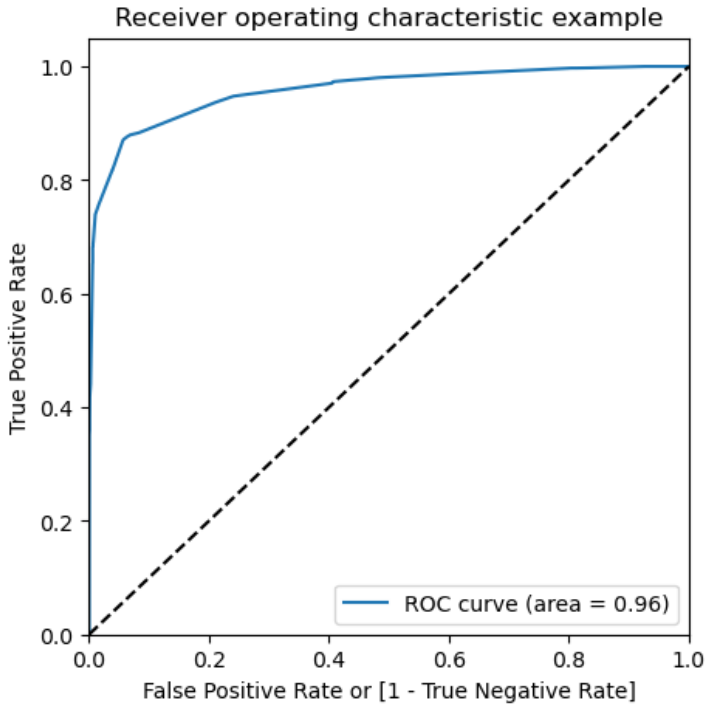
- Do Not Email
- A free copy of Mastering The Interview'

Created dummy variables- one hot encoded for the following columns-

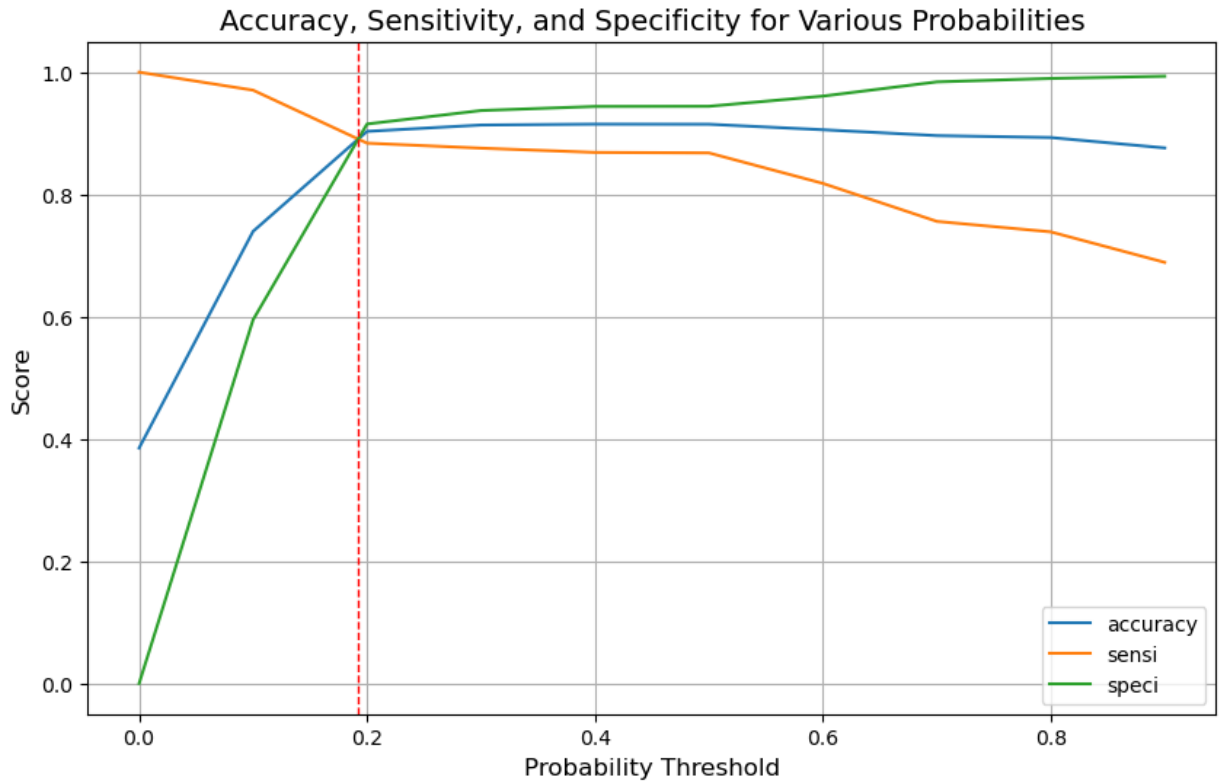
- Lead Origin
- What matters most to you in choosing a course
- Lead Source
- Last Activity
- Country
- Specialization
- Tags
- What is your current occupation
- 'Last Notable Activity'

MODEL BUILDING

ROC CURVE



Accuracy, Sensitivity, And Specificity Curves



- With the current cut off as seen in the graph as **0.192**, the **accuracy** is **90.32%**, **sensitivity** is **88.39%** and **specificity** is **91.52%**
- The **precision** is **86.72%** , **Recall** is **88.39%** and **F1** score is **88.39%**.

Predictions on the Test data

- Performed scaling on the required variables as part of preprocessing.
- Trained the test set using the final model derived from the feature selection process.
- Utilized the trained model to make predictions on the test data.
- Employed a cutoff value of 0.192 for classification.
- Achieved an **Accuracy** of **89.94%** on the test data.
- **Sensitivity**, representing the true positive rate, stood at **87.46%**.
- **Specificity**, representing the true negative rate, was measured at **91.35%**.

Conclusion

The most important variables that mattered the most in the potential buyers are –

- The Total time spent on the website and the Page views per visit
- The 'Lead Add Form' is the lead origin with the highest proportion of converted leads among all the leads.
- When the Lead source is
  - Google
  - Direct Traffic
  - Organic Search
  - Weliangk website
- When the Last Activity was –
  - Reading the ‘SMS Sent’
  - Opening the ‘Emails’ received
- When their current occupation is a working professional.
- When the tag is ‘Will revert after reading the email’ , ‘Closed by Horizon’, and ‘Lost to EINS’


THANK YOU!