# Lead Scoring Case Study

Team – Vaishnavi Herur & Ujwal Mishra

Submission – 21st May, 2024



Public

# Problem Statement

- ✓ An education company, X Education, wants to improve its lead conversion rate.
- ✓ The typical lead conversion rate at X education is around 30%.
- ✓ They aim to identify the most potential leads (hot leads) to focus their sales efforts more efficiently.
- ✓ The company needs a model to assign lead scores to prioritize leads for follow-up, aiming for an 80% conversion rate

*Note : Conversion rate is for example, if, say, they acquire 100 leads in a day, only about 30 of them are converted*

# Business Objective

- Build a logistic regression model to predict lead conversion probabilities.
  - Assign lead scores based on these probabilities.
  - Use the lead scores to prioritize leads for follow-up
- Increase lead conversion rate from the current 30% to 80%.
- Improve efficiency and effectiveness of the sales team.
  - Identify hot leads for more targeted sales efforts.

# Solution Methodology

1. Data cleaning and data manipulation.
   - Check and handle duplicate data.
   - Check and handle NA values and missing values.
   - Drop columns, if it contains large amount of missing values and not useful for the analysis.
   - Imputation of the values, if necessary.
   - Check and handle outliers in data.

2. EDA
   - Univariate data analysis
   - Bivariate data analysis

3. Model building
   - Feature Scaling & Dummy Variables and encoding of the data.
   - Classification technique: logistic regression used for the model making and prediction.
   - Validation of the model.

4. Model presentation.

5. Conclusions and recommendations.

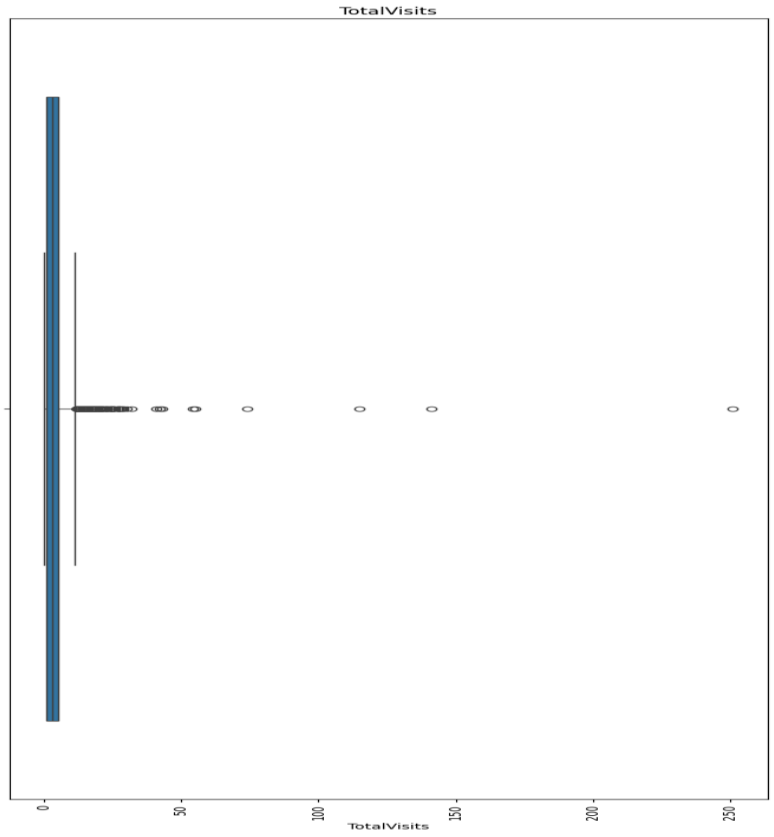
# Data Cleaning and Manipulation

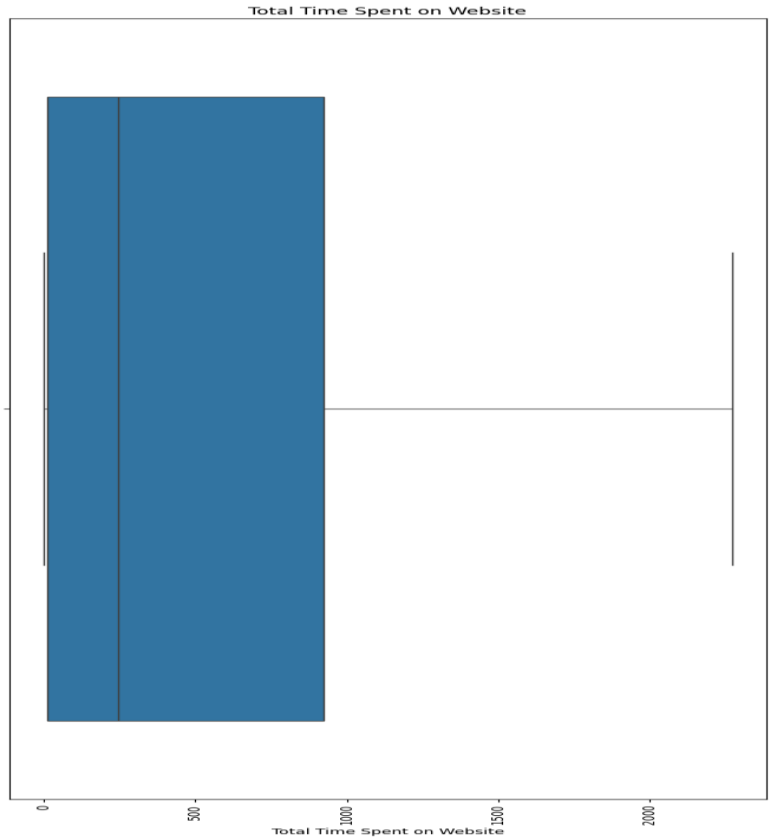
- Shape of data set : Rows – 9240 | Columns – 37
- Columns with high % of null values ~30% and above are dropped
  - 'Last Activity', 'What matters most to you in choosing a course', 'Tags', 'Lead Quality', 'Lead Profile', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score'
- Columns with no variance or having only single value are dropped
  - 'Magazine', 'Receive More Updates About Our Courses', 'Get updates on DM Content', 'Update me on Supply Chain Content', 'I agree to pay the amount through cheque'
- Columns which are of no relevance are dropped
  - 'City', 'A free copy of Mastering The Interview', 'Last Notable Activity'

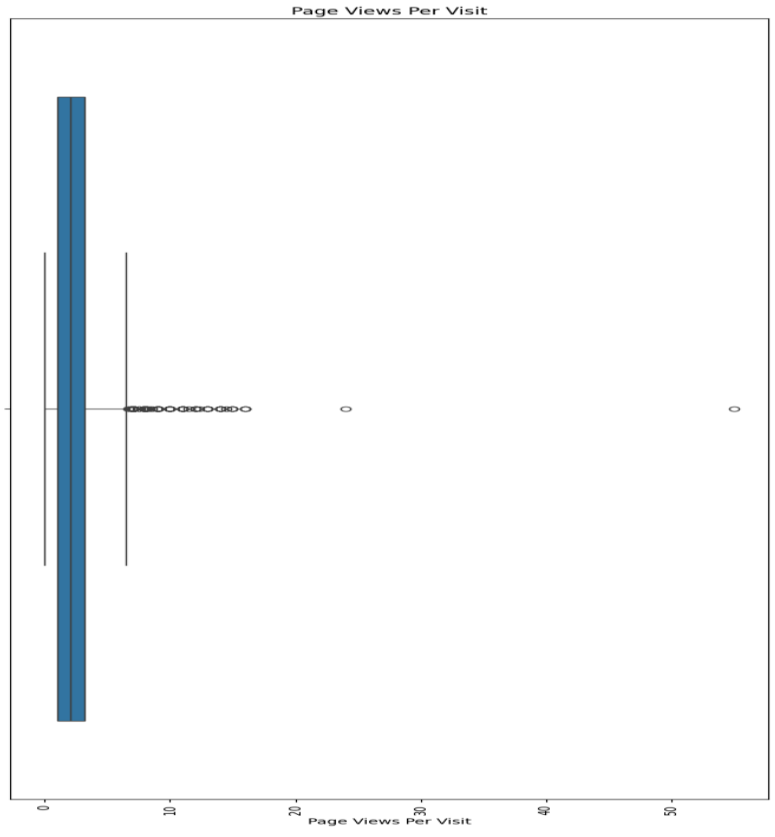
# Data Cleaning and Manipulation

- These columns are dropped due to data imbalance. Refer the univariate analysis
  - 'Do Not Email', 'Do Not Call', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations'
- Data in these columns has been bucketed based on similarity and redundancy
  - 'Lead Origin', 'Lead Source', 'Specialization'

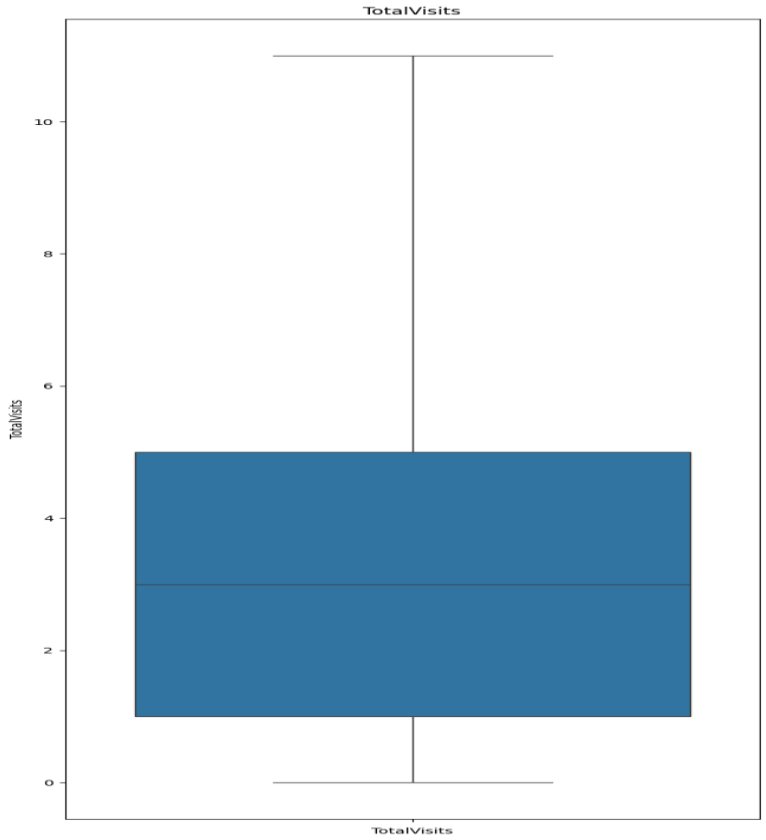
# Handling Outlier Data

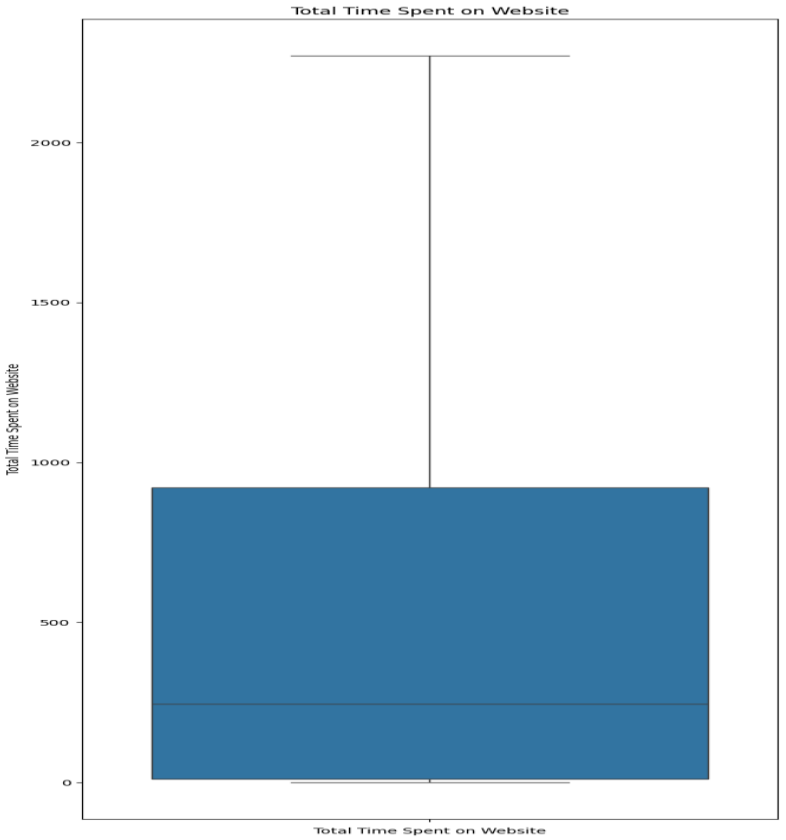


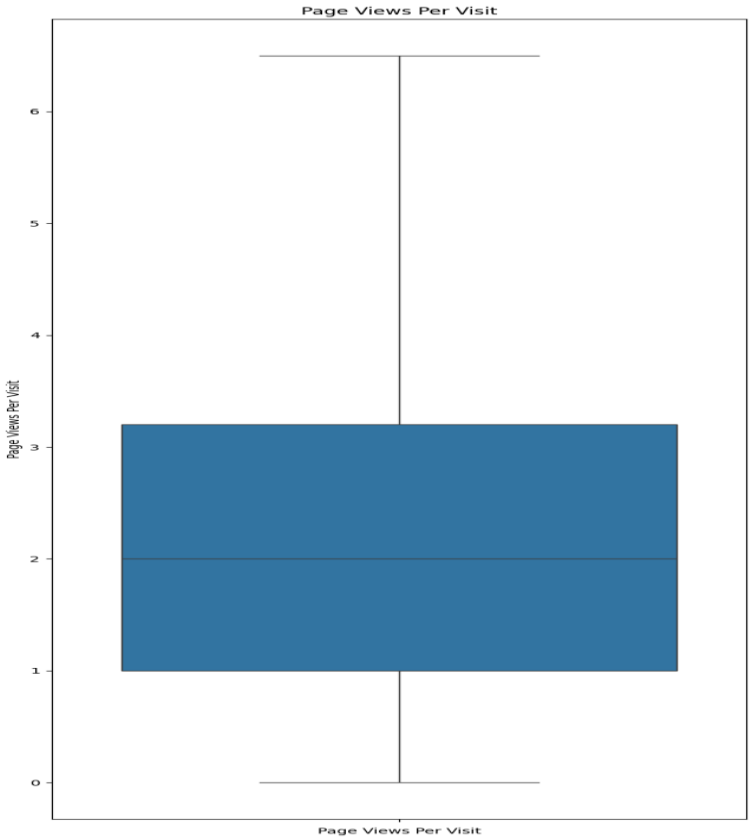



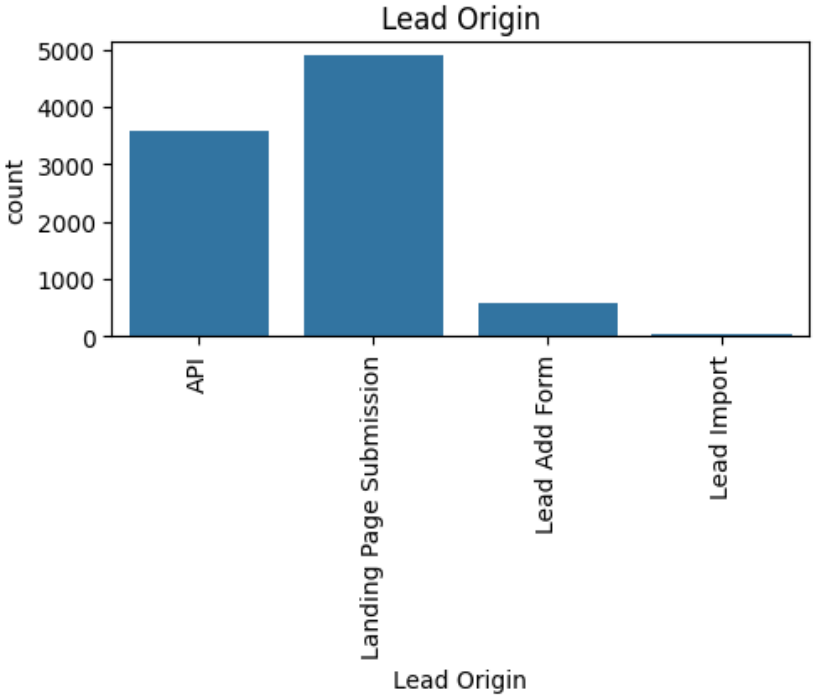
# Handling Outlier Data

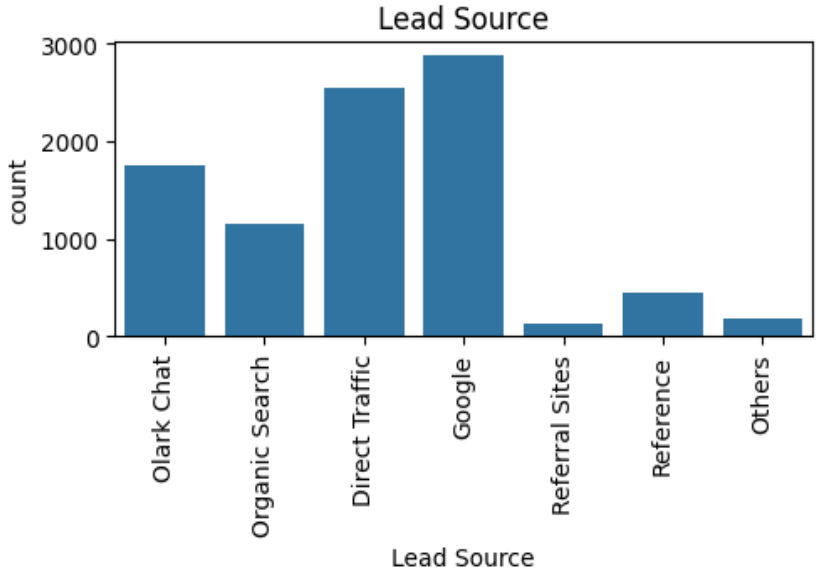


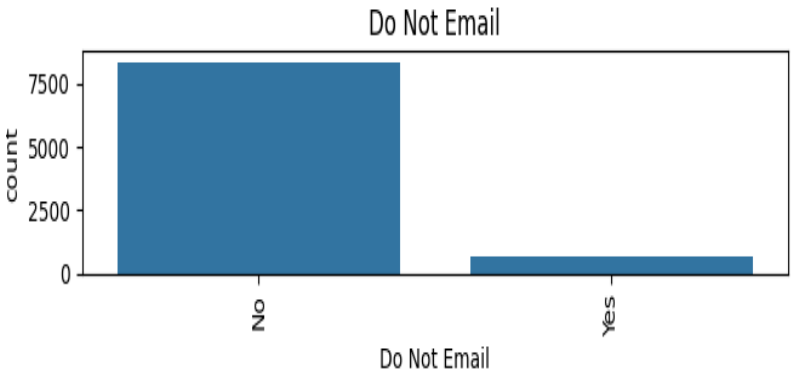



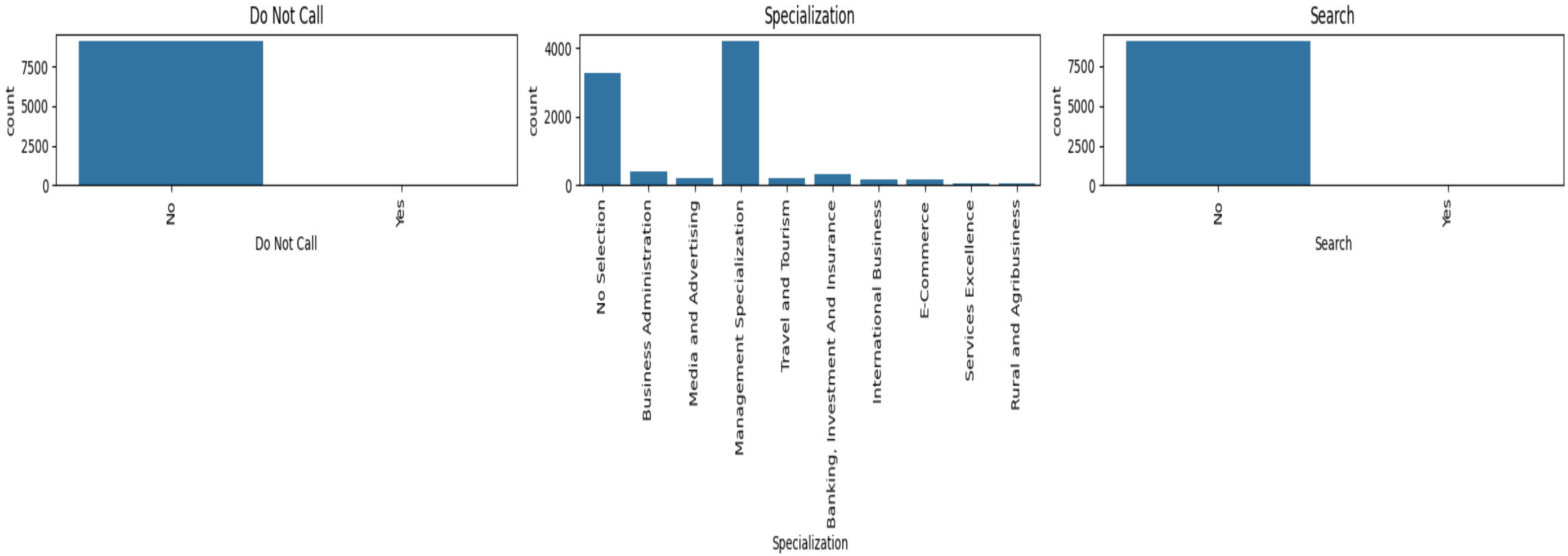
# Univariate Analysis



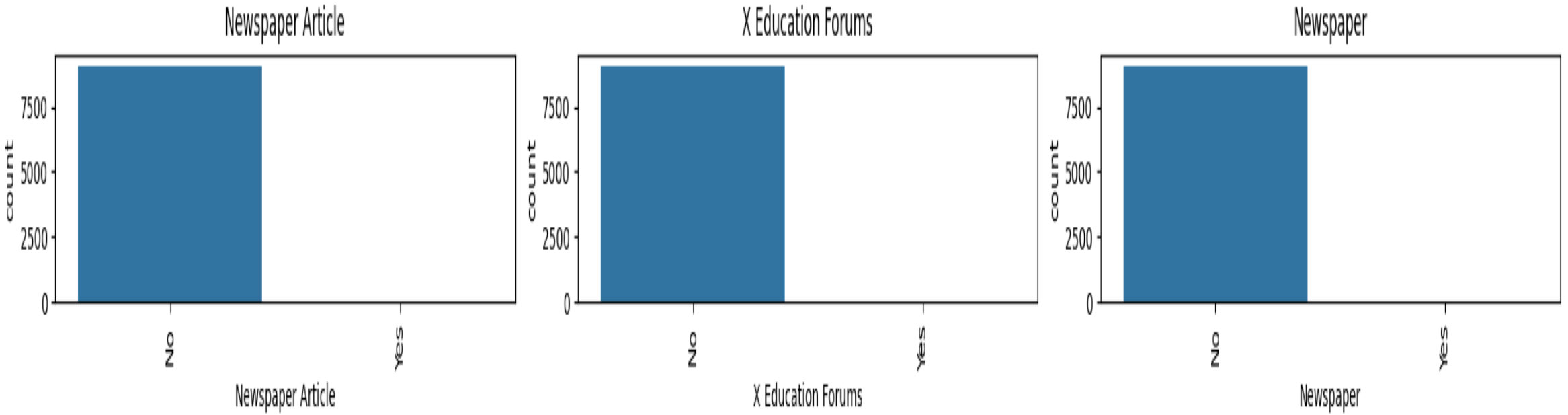



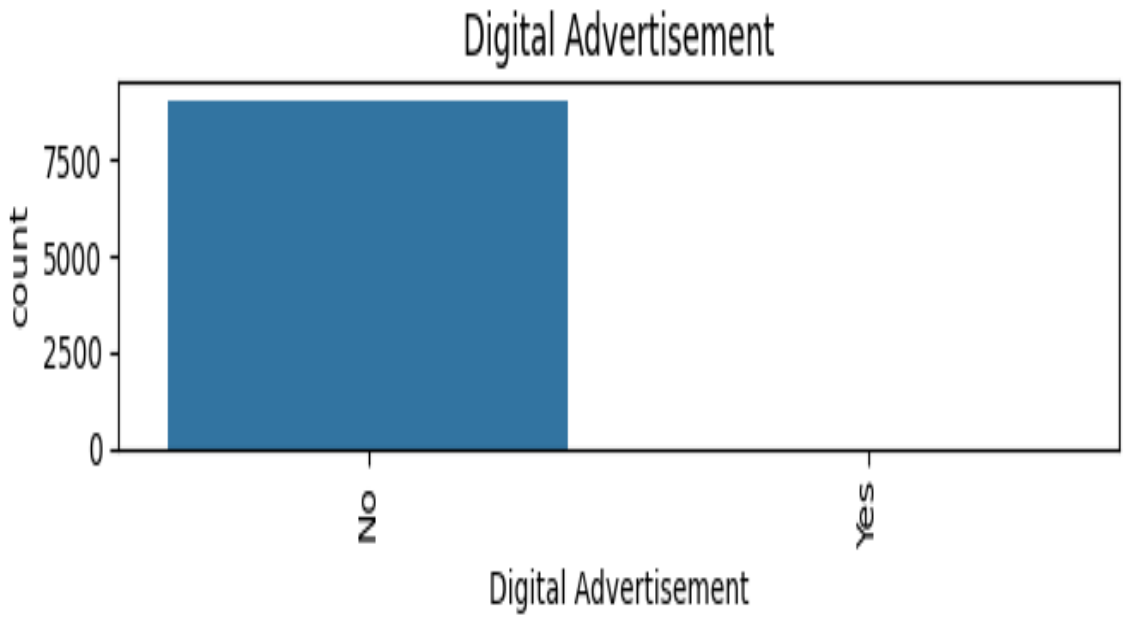
# Univariate Analysis

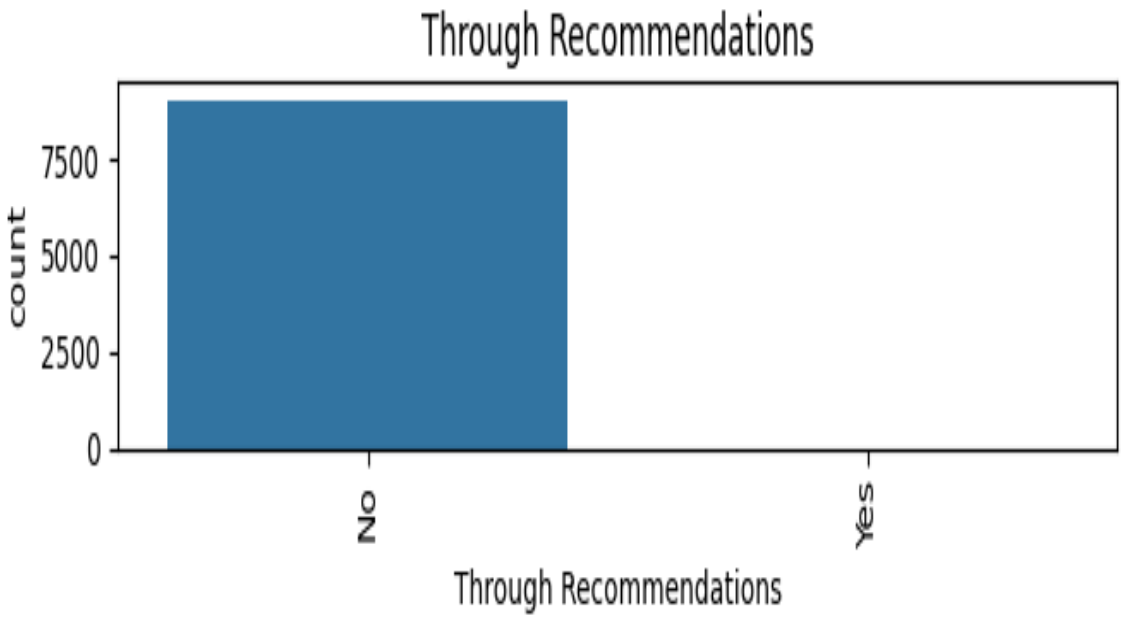


# Data Cleaning and Manipulation

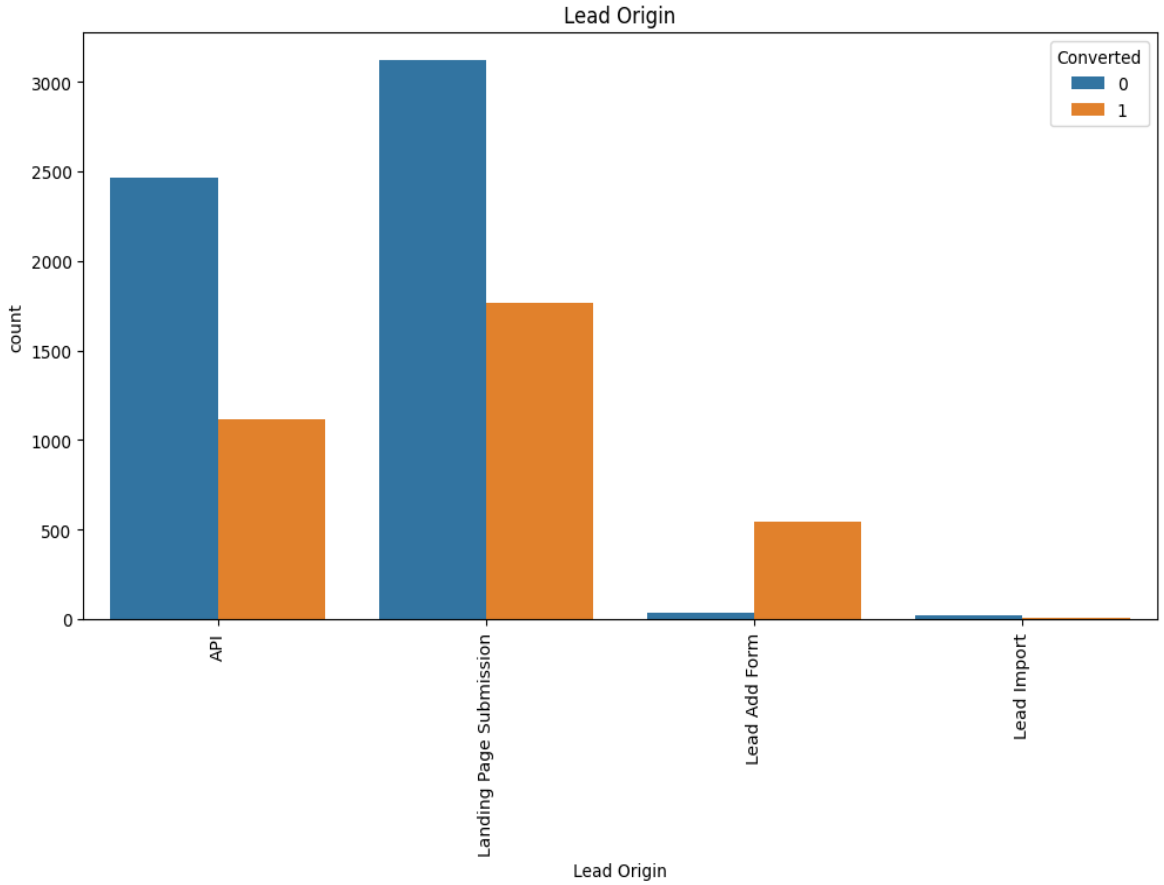


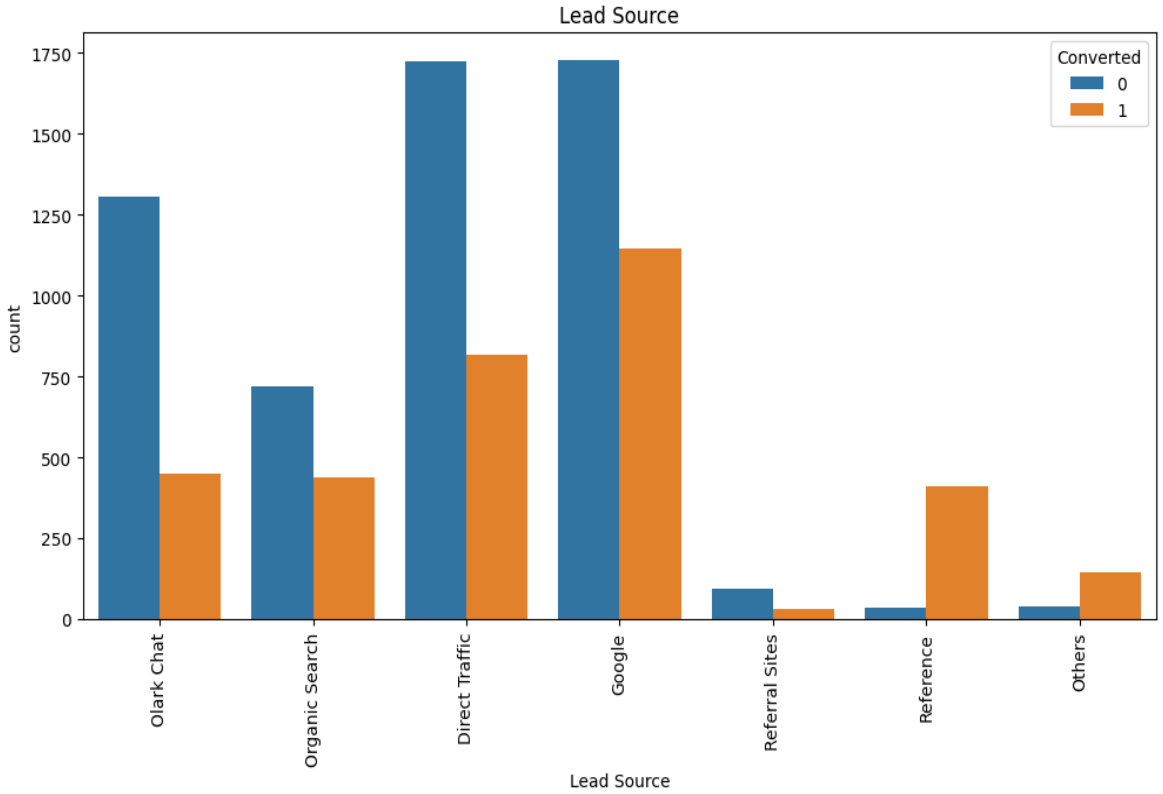
# Data Cleaning and Manipulation



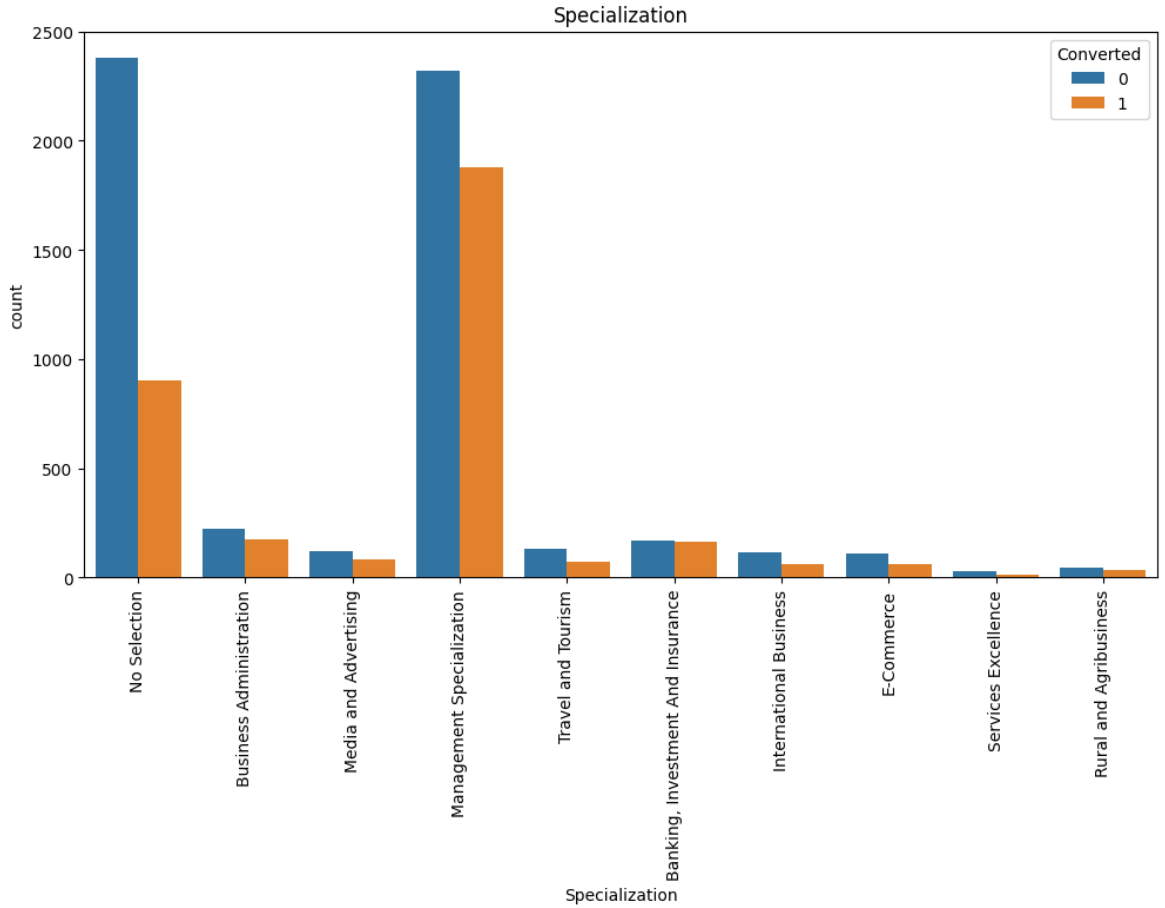
# Bivariate Analysis

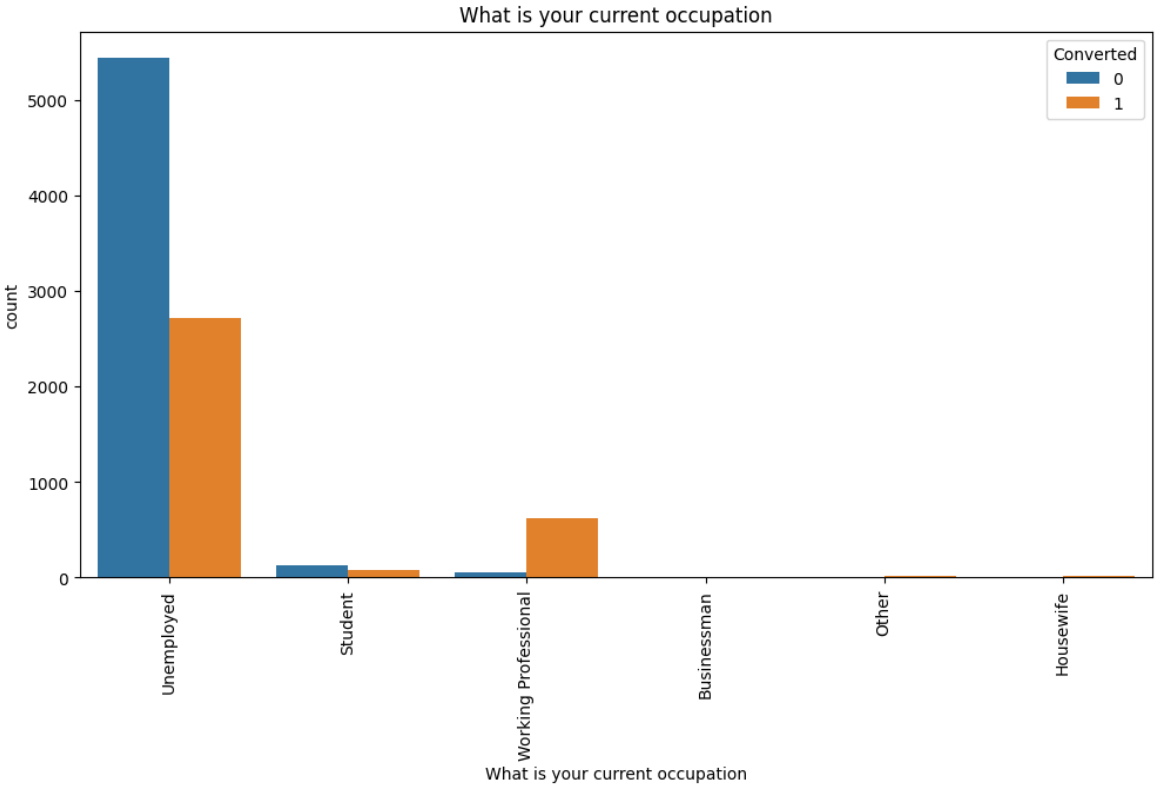



Publi

# Bivariate Analysis



# Bivariate Analysis

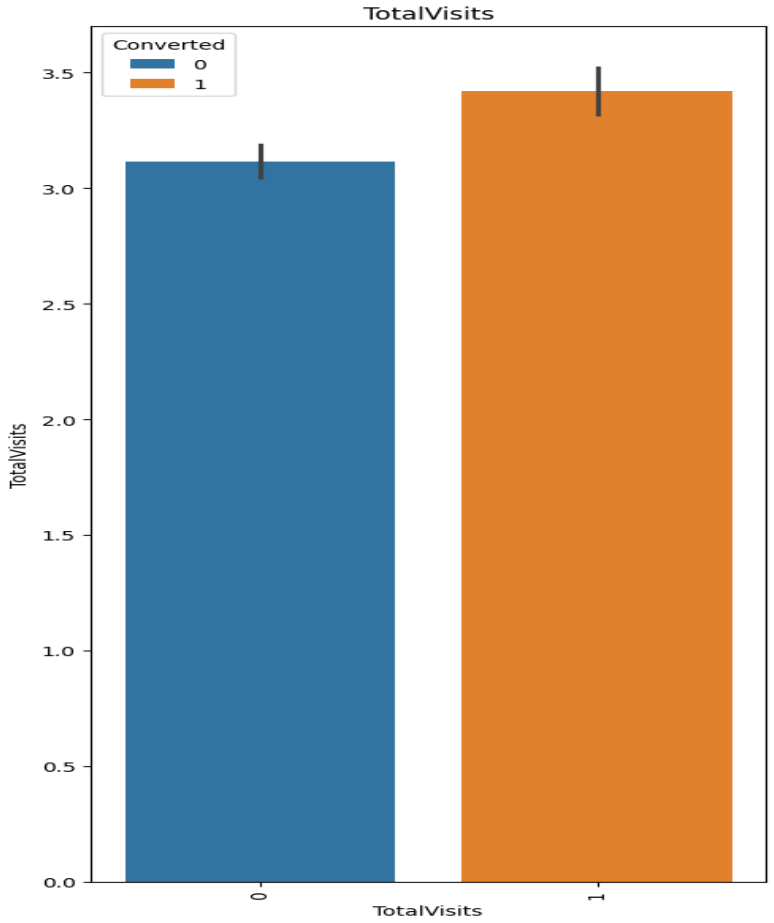
- ✓ Lead Origin: Leads coming from a "Landing Page Submission," about 36.17% of them end up converting into customers, while the remaining 63.83% do not convert. This indicates that while landing page submissions are effective in attracting leads, there's still room for improvement in converting these leads into customers.

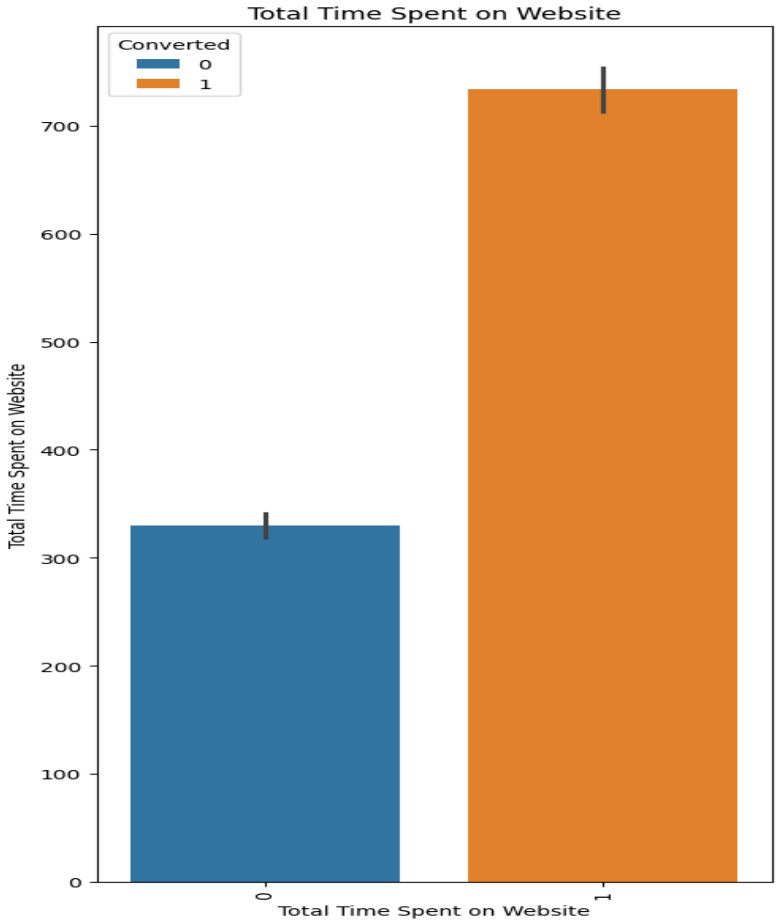
- ✓ Lead Source: The data suggests that while around 60.08% of leads from Google did not convert, approximately 39.92% did convert. This indicates that Google is a significant source of leads, but there is still potential to improve conversion rates by optimizing strategies targeting leads from this platform.

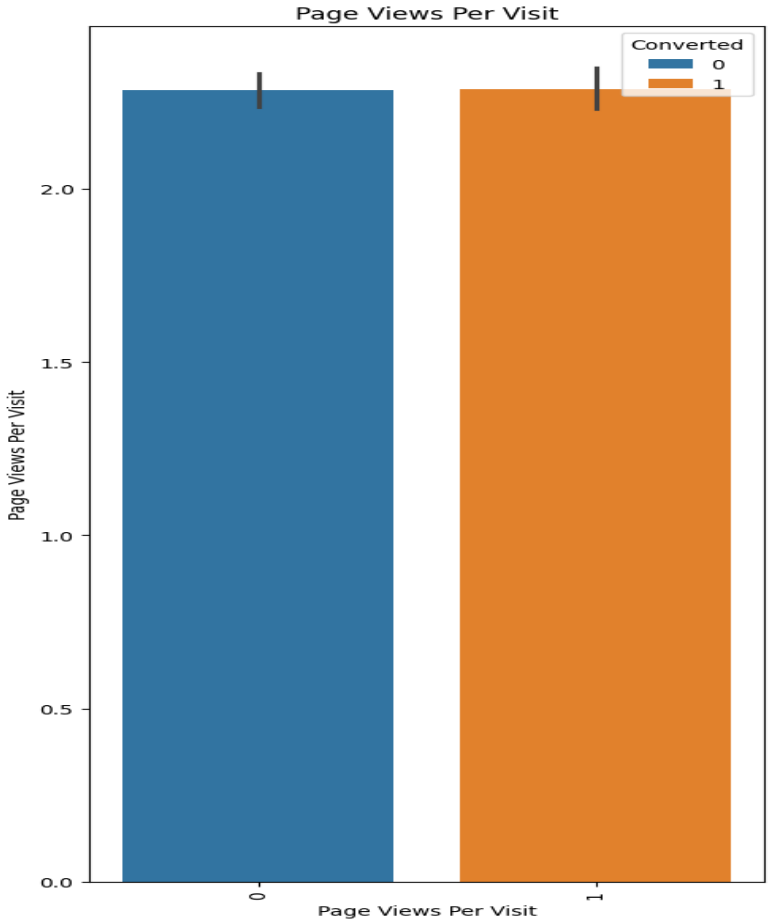
- ✓ Specialization: The provided data indicates that among leads with a management specialization, approximately 44.70% ended up converting into customers, while the remaining 55.30% did not convert. This suggests that while there is some success in converting leads with a management specialization, there may be opportunities to further enhance conversion rates within this group.

- ✓ Current occupation: The data indicates that 33.29% of unemployed leads convert into customers, while 66.71% do not. This suggests that while there is some potential in targeting unemployed individuals, the conversion rate is relatively low, highlighting a need for tailored strategies or support to better engage and convert this demographic

# Bivariate Analysis (Continuous)




Pub

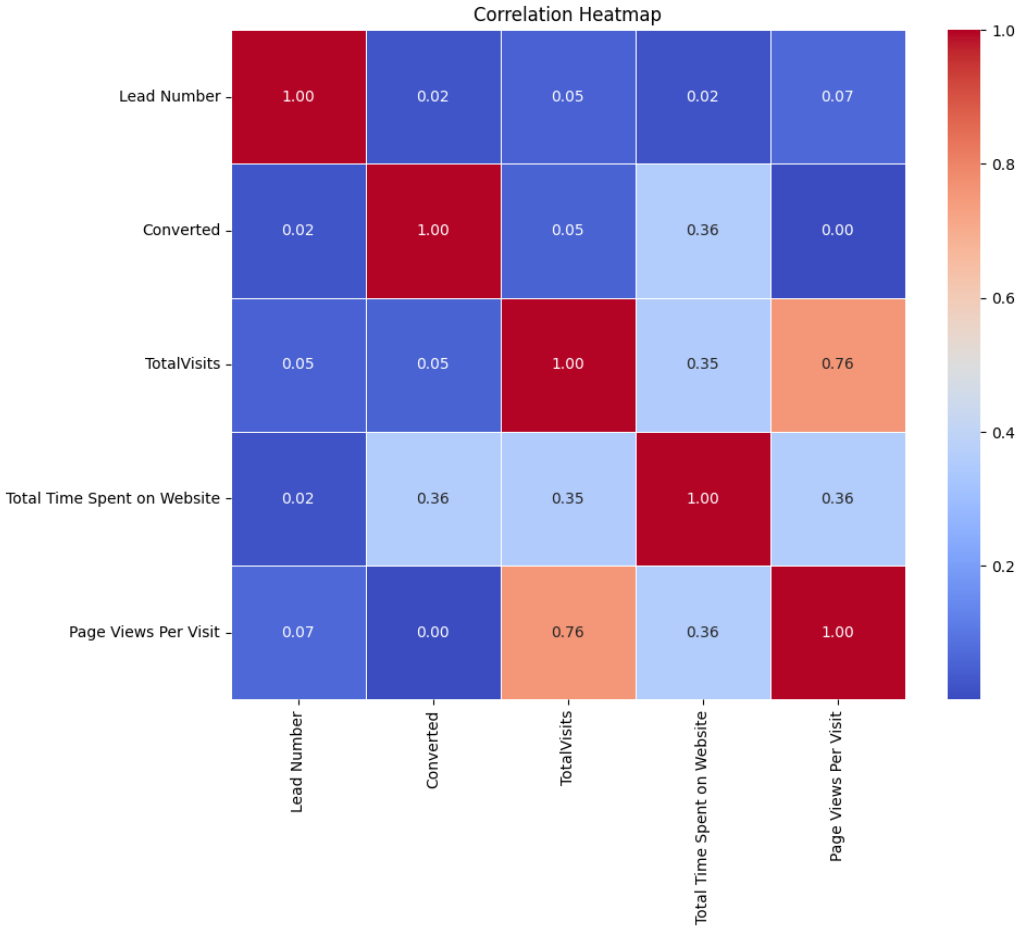
# Bivariate Analysis (Continuous)

Total Visits - This slight difference suggests that there's a small increase in the average number of visits for leads who eventually converted, indicating a potential trend towards higher engagement among converting leads.

Total time spent on website - Leads who did not convert (Converted=0) spent approximately 330 seconds (or about 5.5 minutes) on the website, whereas those who did convert (Converted=1) spent significantly more time, around 734 seconds (or about 12.2 minutes). This suggests a substantial difference in engagement levels between converting and non-converting leads, with converting leads spending notably more time on the website.

Page view per visit - On average, both converting (Converted=1) and non-converting (Converted=0) leads have a similar number of page views per visit, with converting leads having a slightly higher average of approximately 2.29 page views per visit compared to 2.29 for non-converting leads. This suggests that there's not a significant difference in the average number of pages viewed per visit between converting and non-converting leads.

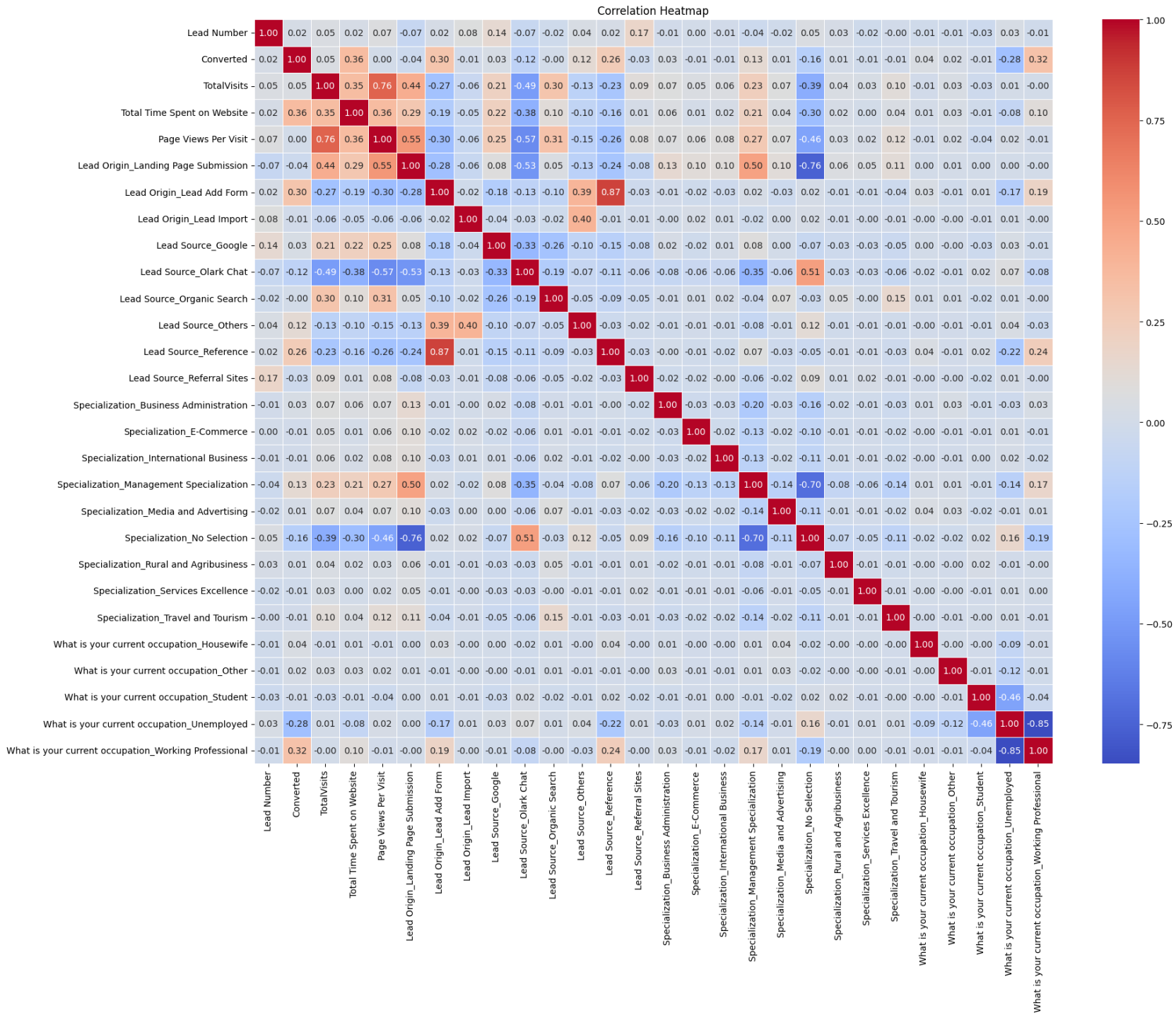
# Heatmap




- There is a moderate positive correlation (correlation coefficient = 0.3546) between the total number of visits and the total time spent on the website, indicating that leads who visit the website more often tend to spend more time on it.
- There is a strong positive correlation (correlation coefficient = 0.7552) between the total number of visits and the average page views per visit, suggesting that leads who visit the website more often tend to view more pages per visit

# Model Building

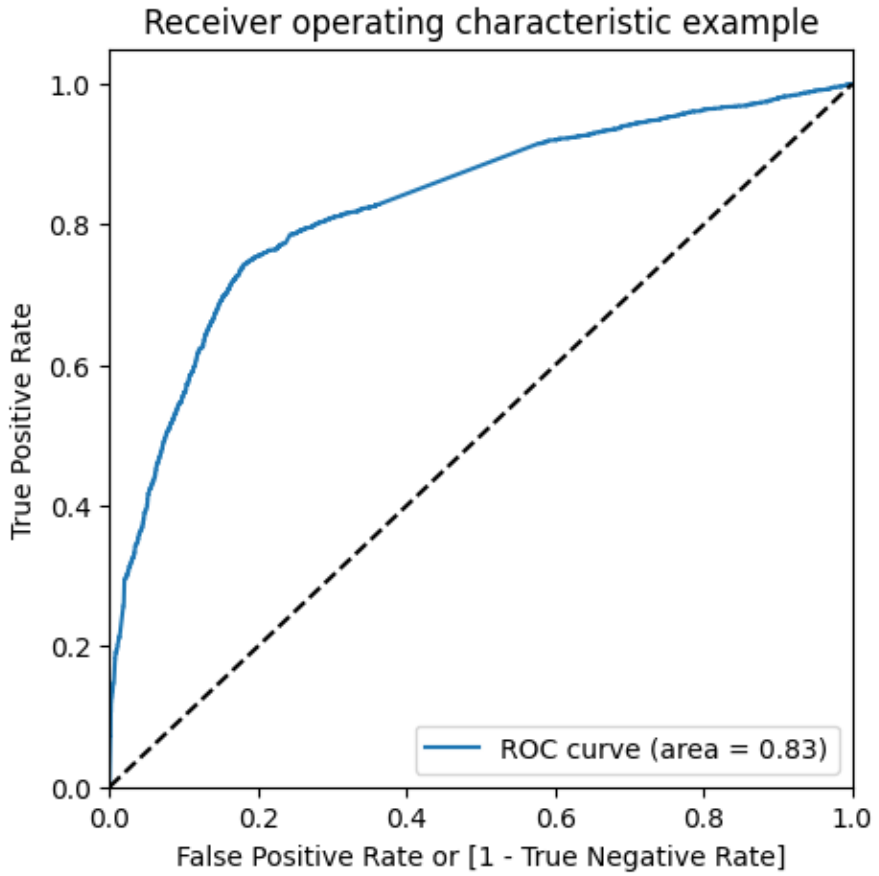
- Splitting dataset into train and test
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p-value is greater than 0.05 and vif value is greater than 5
- Overall accuracy 79.39%

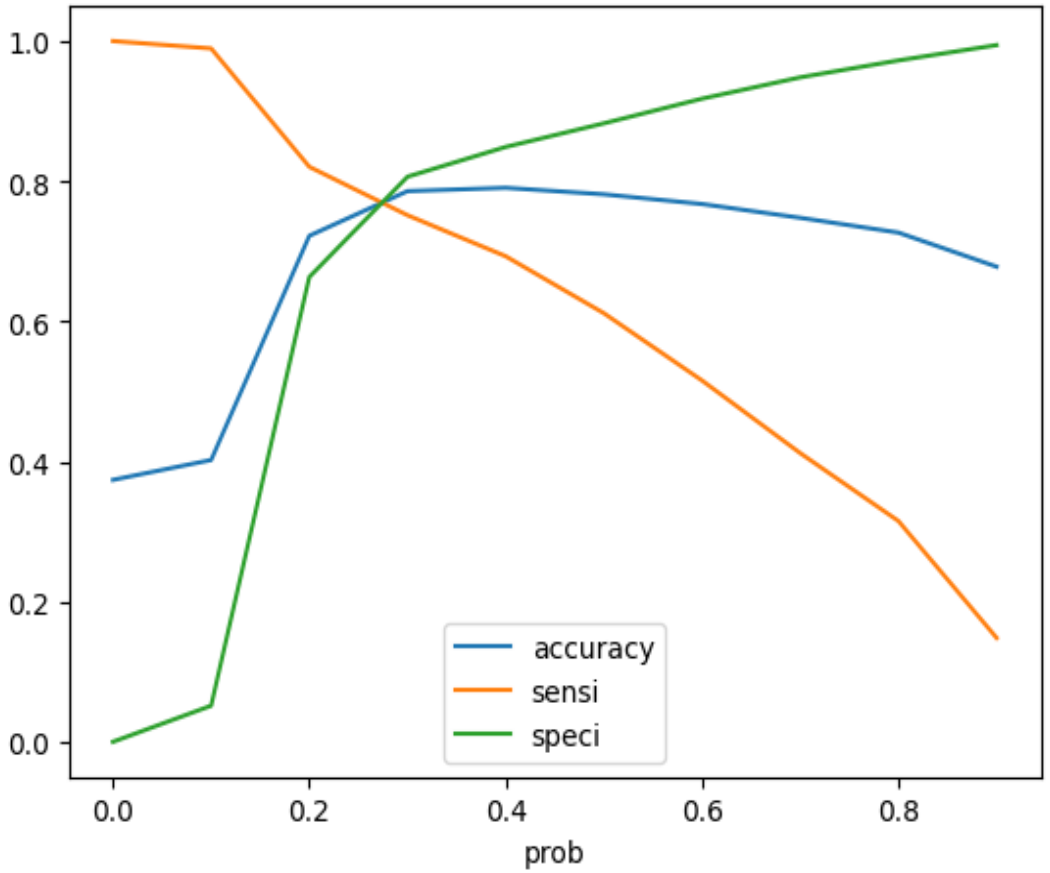



- ✓ Dummy variables created for categorical variables
  - 'Lead Origin'
  - 'Lead Source'
  - 'Specialization'
  - 'What is your current occupation'

- ✓ Dropping the following columns due to multi collinearity
  - 'Page Views Per Visit',
  - 'Specialization_No Selection'
  - 'Lead Origin_Lead Add Form'

# ROC Curve



# Conclusion

Top 3 variables that contribute most towards probability of lead getting converted

- Total time spend on the website (1.053)
- Lead Source_Reference(0.9118)
- What is your current occupation_Working Professional(0.7711)

The company should focus on building strategy around the following

- Engaging and informative website content that keeps visitors longer can significantly enhance conversion rates
- Leveraging customer testimonials, referral programs, and encouraging satisfied customers to spread the word can be beneficial strategies
- Aggressively perform marketing addressing working professional