Lead Scoring Analysis Assignment

Prepared for X Education Company

Problem Statement

- X Education needs to improve their lead conversion rate by identifying potential 'hot leads'.
- Goal: Build a model to assign lead scores that indicate conversion probability.

Data Understanding

- Dataset: 9000+ leads
- Key Attributes: Lead Source, Total Time Spent, Last Activity, etc.
- Target Variable: Converted (1 = Converted, 0 = Not Converted).

Data Cleaning and Preparation

1. Removed columns with >50% missing values.
2. Imputed missing numerical values with median.
3. Imputed missing categorical values with mode.
4. Removed duplicate rows.
5. Standardized numerical features.

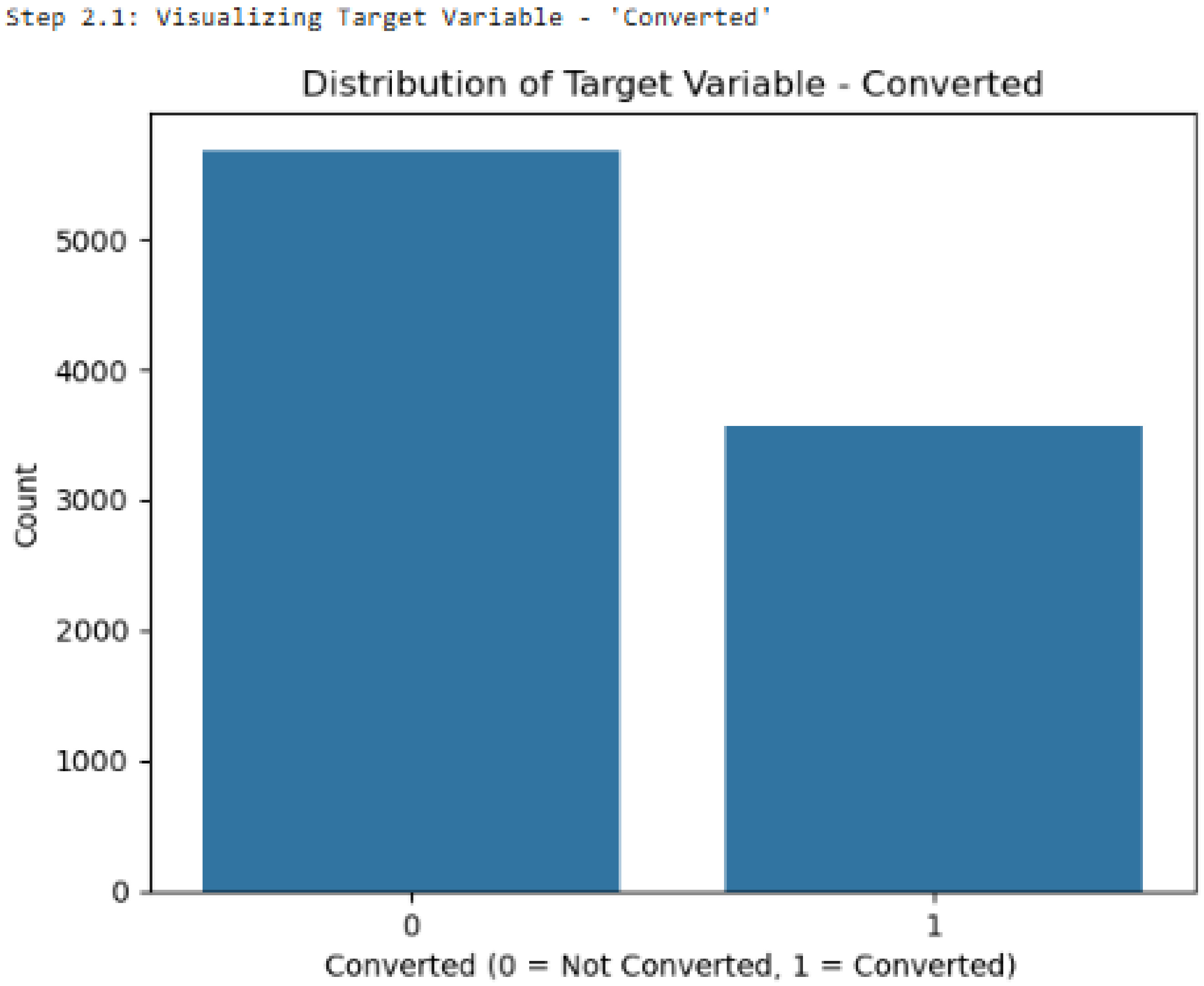
Final result of Data Cleaning:

```
Step 1.4: Checking for duplicate rows
Number of Duplicate Rows: 0
Final Data Shape After Cleaning: (9240, 34)
```

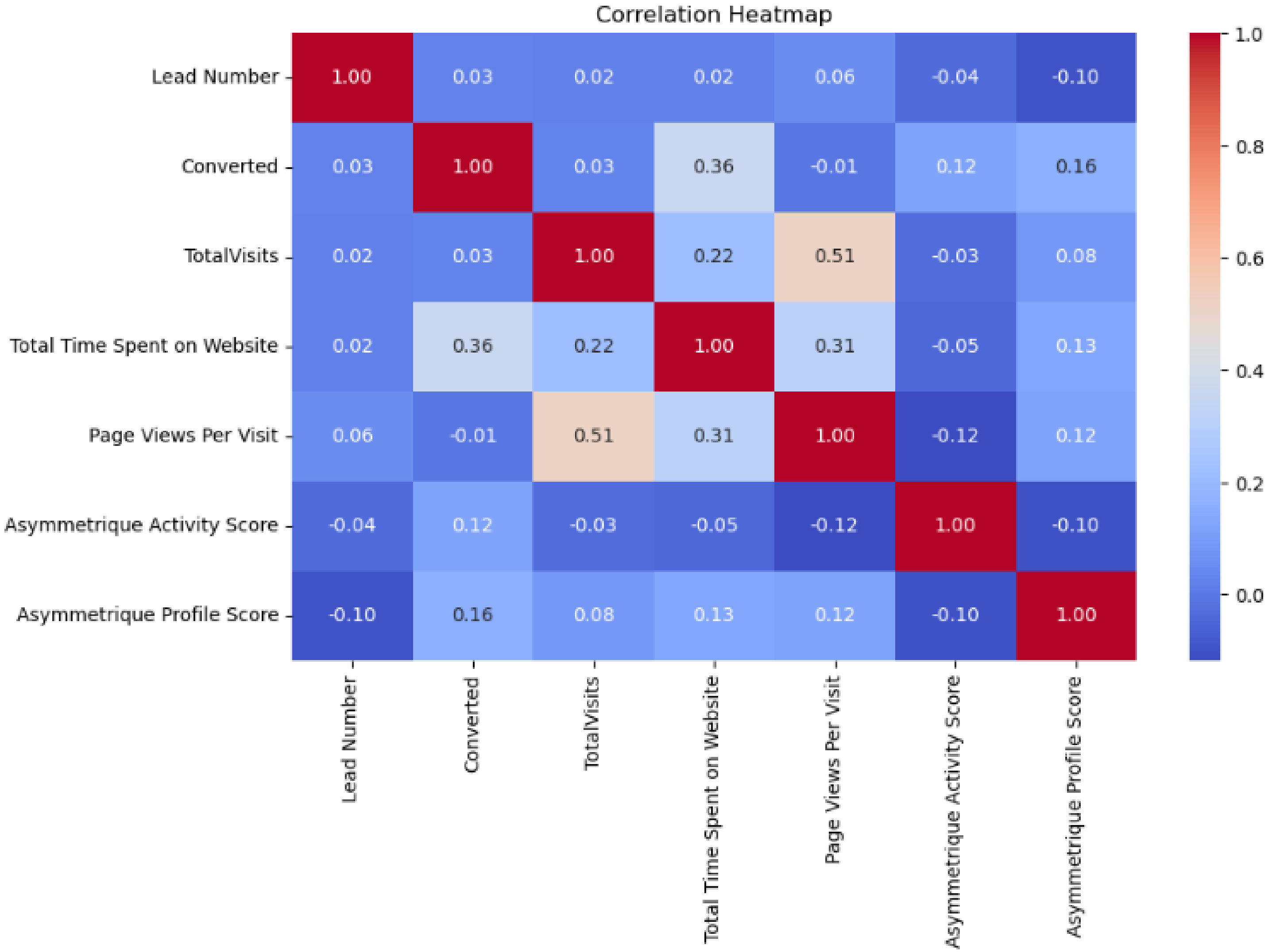
Exploratory Data Analysis

Key Insights:

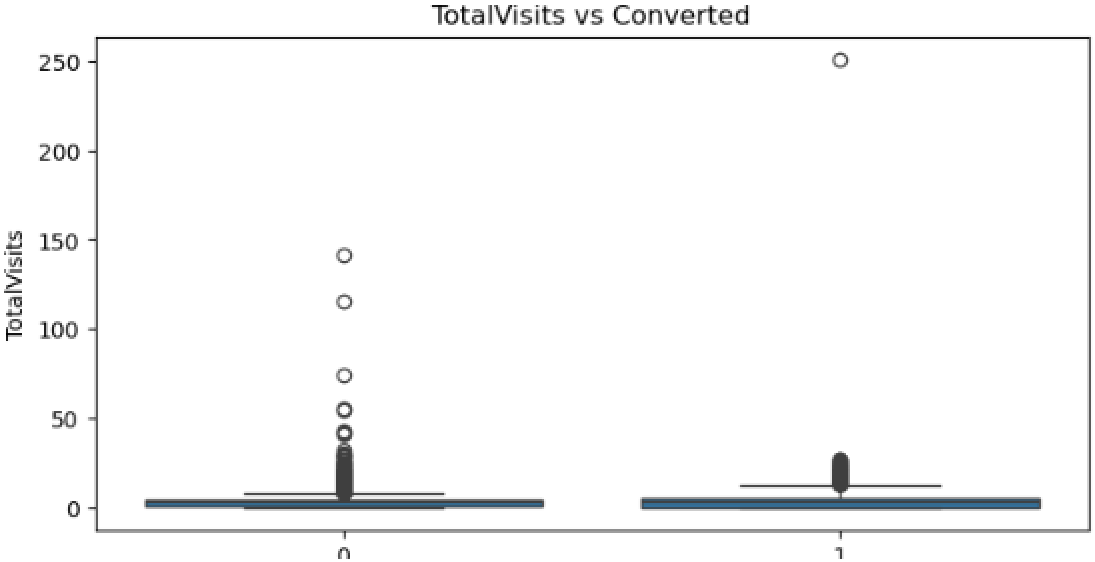
- - Positive correlation between 'Time Spent on Website' and conversion.
- - Lead Source and Last Activity show patterns of conversion rates.
- - Imbalance in target variable (Converted ~30%).



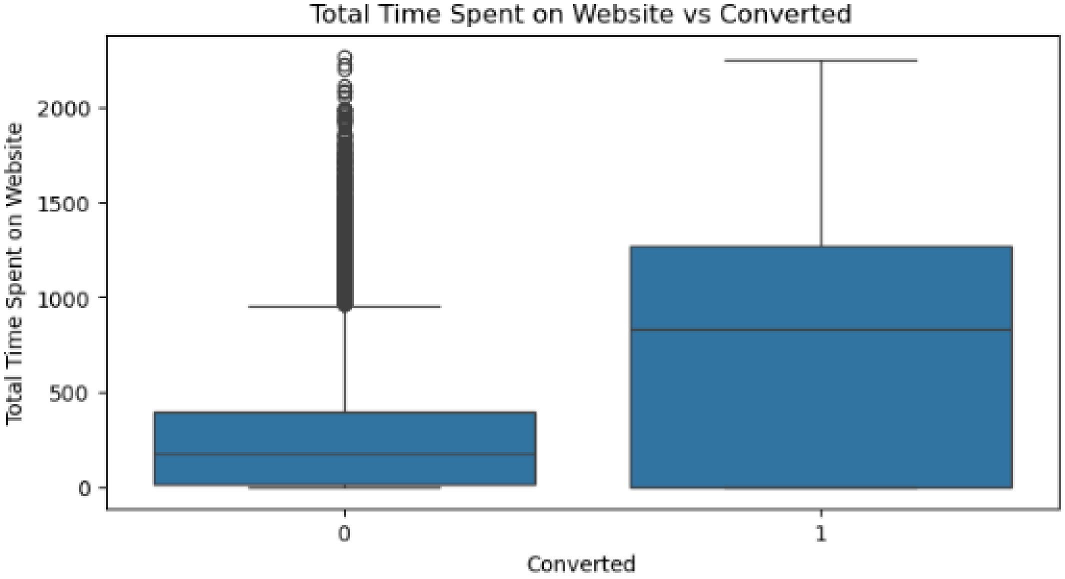
```
Step 2.1: Visualizing Target Variable - 'Converted'
Step 2.2: Correlation Heatmap for Numerical Features
```



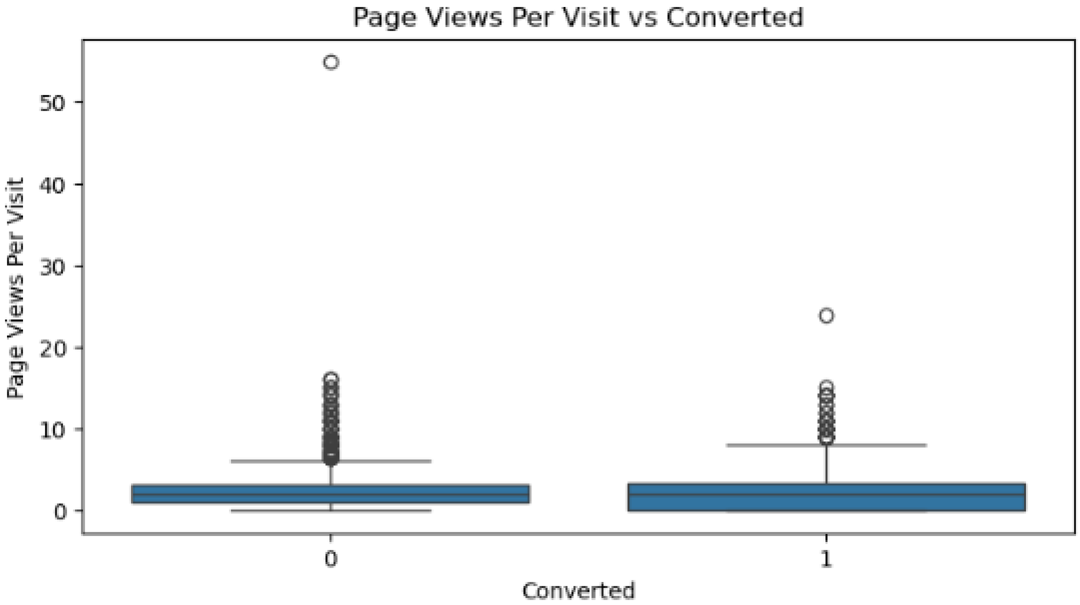
```
Step 2.3: Boxplots for Numerical Features
```



```
Analysis for TotalVisits: Boxplot displayed
```

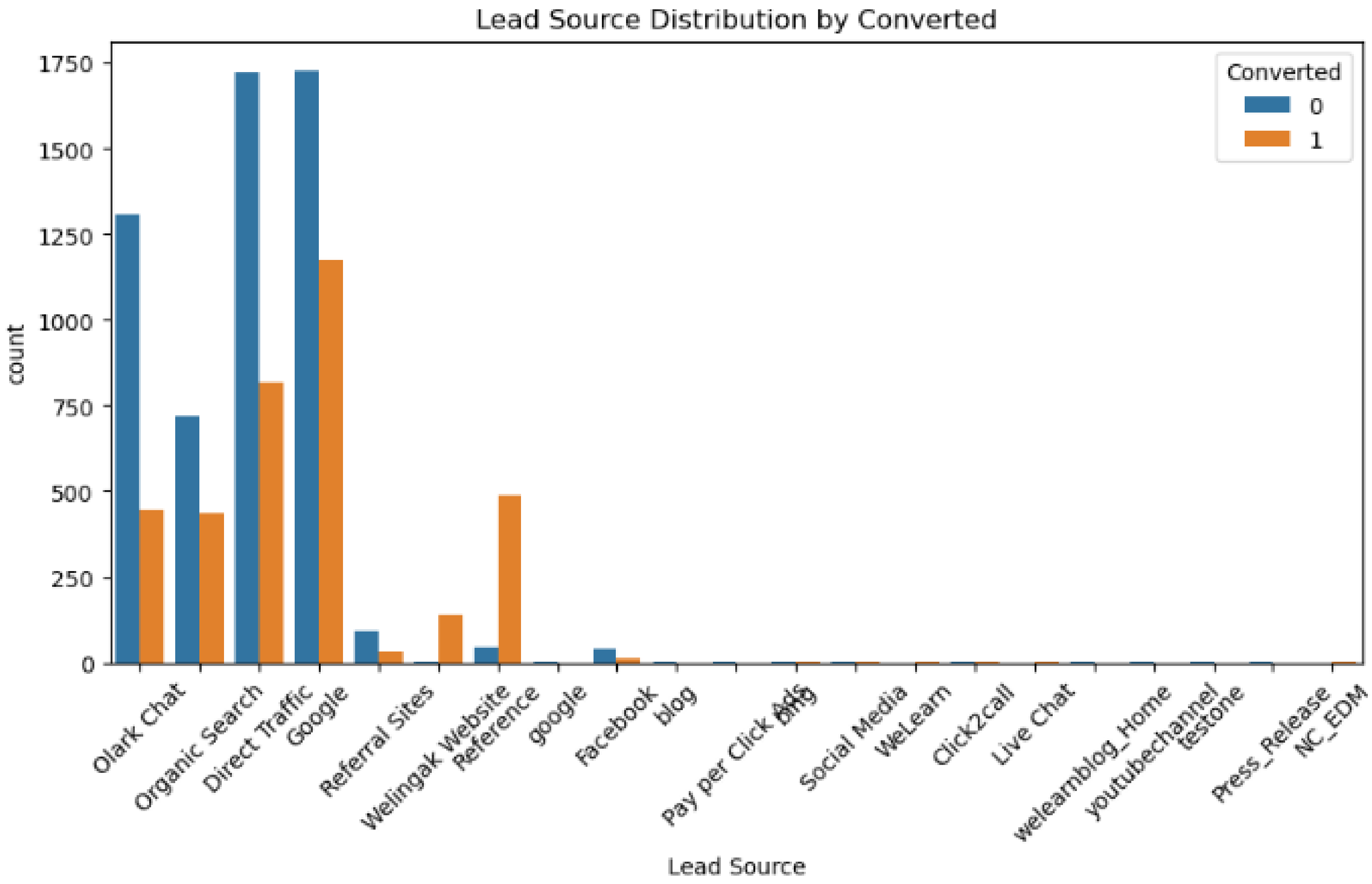


```
Analysis for Total Time Spent on Website: Boxplot displayed
```

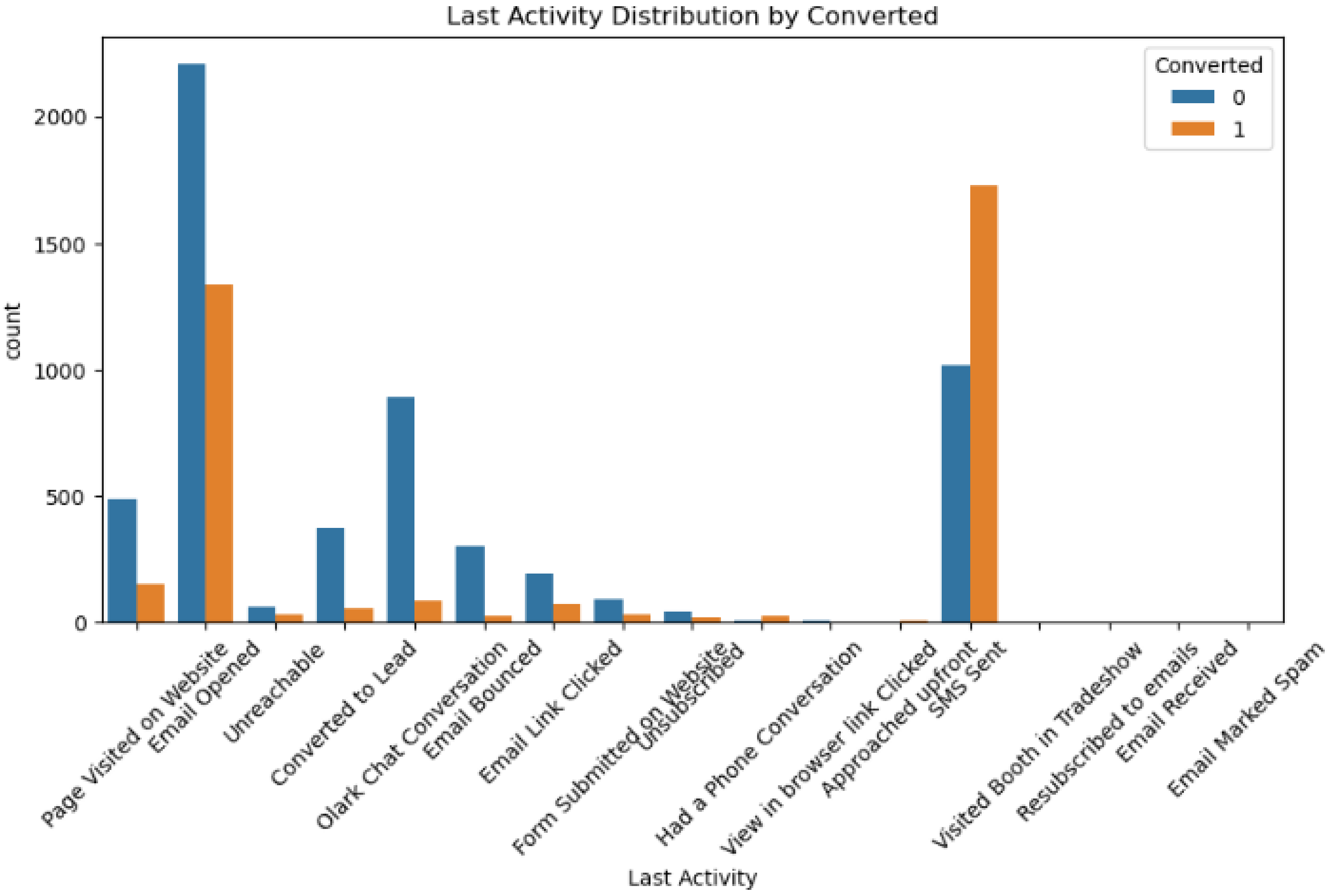


```
Analysis for Page Views Per Visit: Boxplot displayed
```

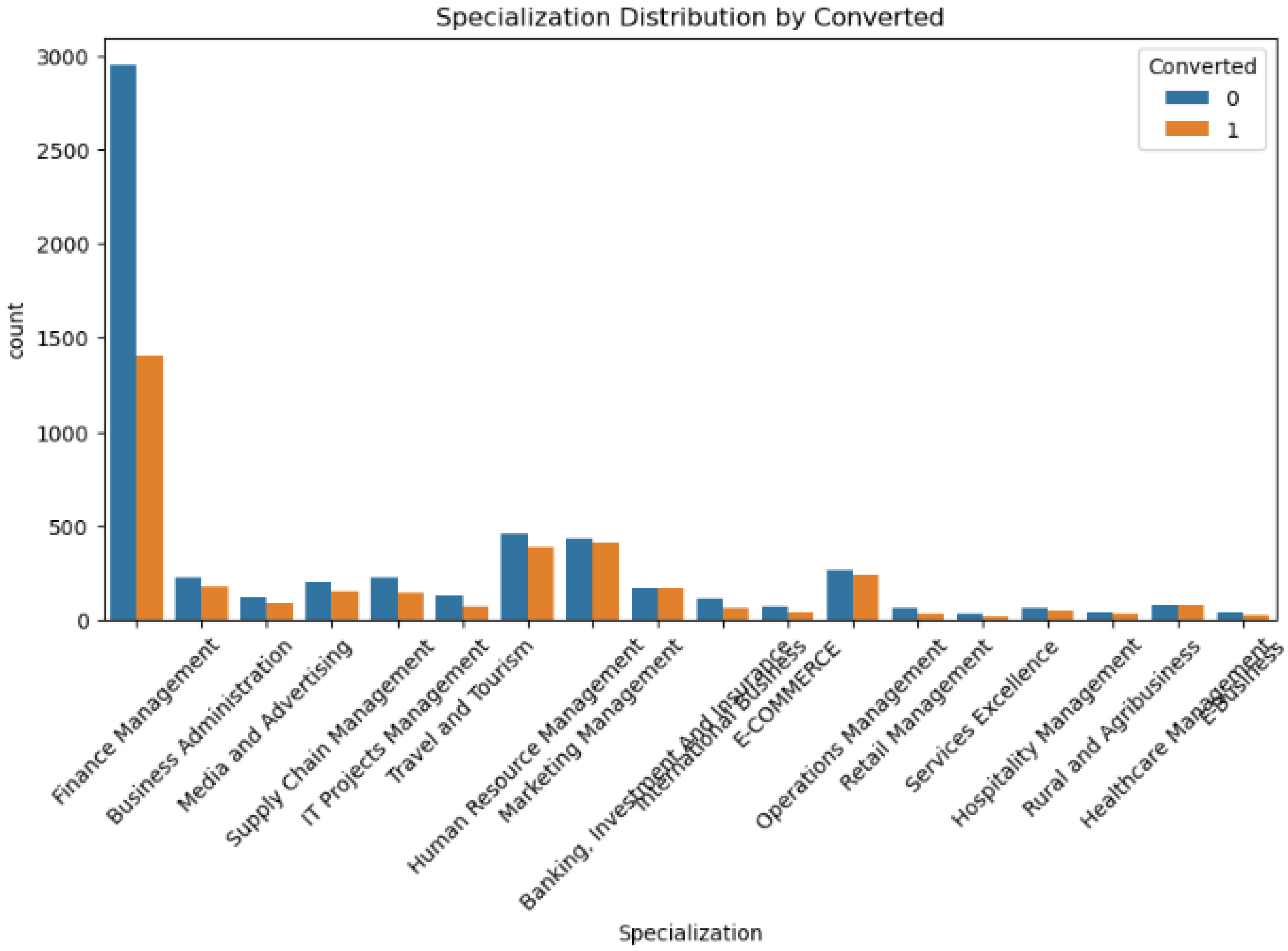
```
Step 2.4: Countplots for Categorical Features
```



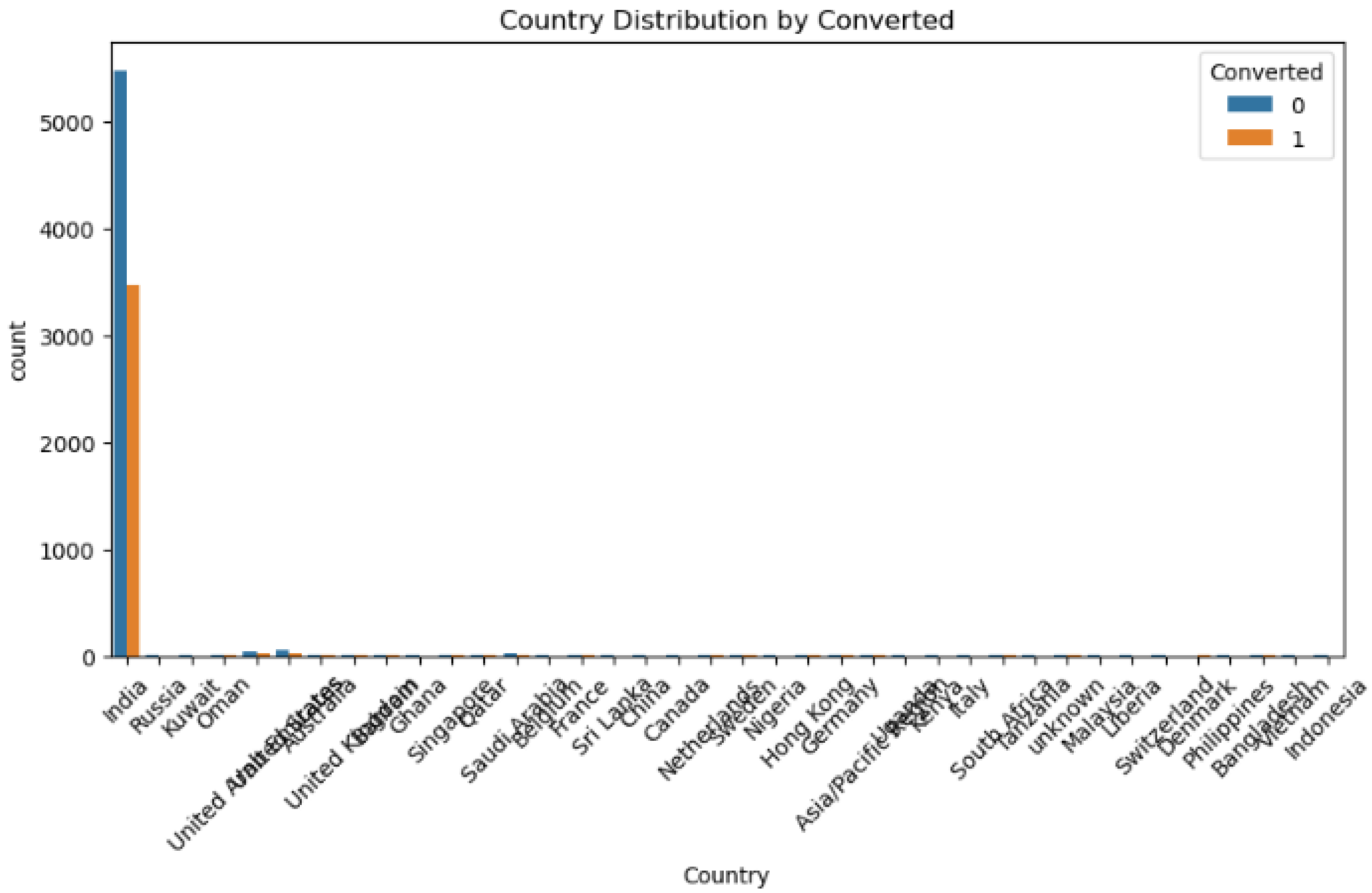
```
Analysis for Lead Source: Countplot displayed
```



```
Analysis for Last Activity: Countplot displayed
```



```
Analysis for Specialization: Countplot displayed
```



```
Analysis for Country: Countplot displayed
```

Feature Engineering

- 1. Dropped irrelevant columns (ID, etc.).
- 2. Created dummy variables for categorical features.
- 3. Standardized numerical features using StandardScaler.

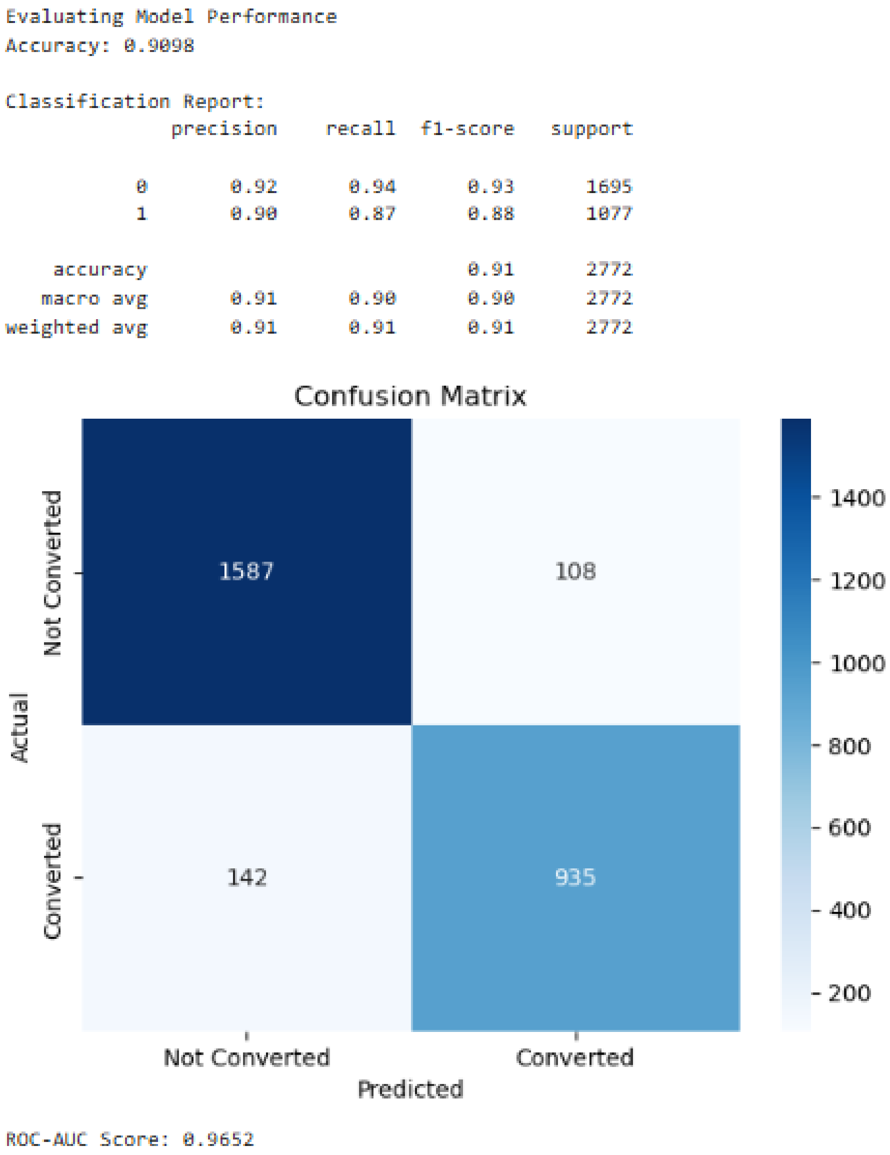
Model Building

- Algorithm: Logistic Regression
- Data Split: 70% training, 30% testing
- Max Iterations: 1000

Model Evaluation

- Accuracy: 90.98%
- Precision, Recall, F1-Score: Achieved high performance
- ROC-AUC Score: 91%

- Evaluation Metrics:
- - Confusion Matrix
- - ROC-AUC Curve



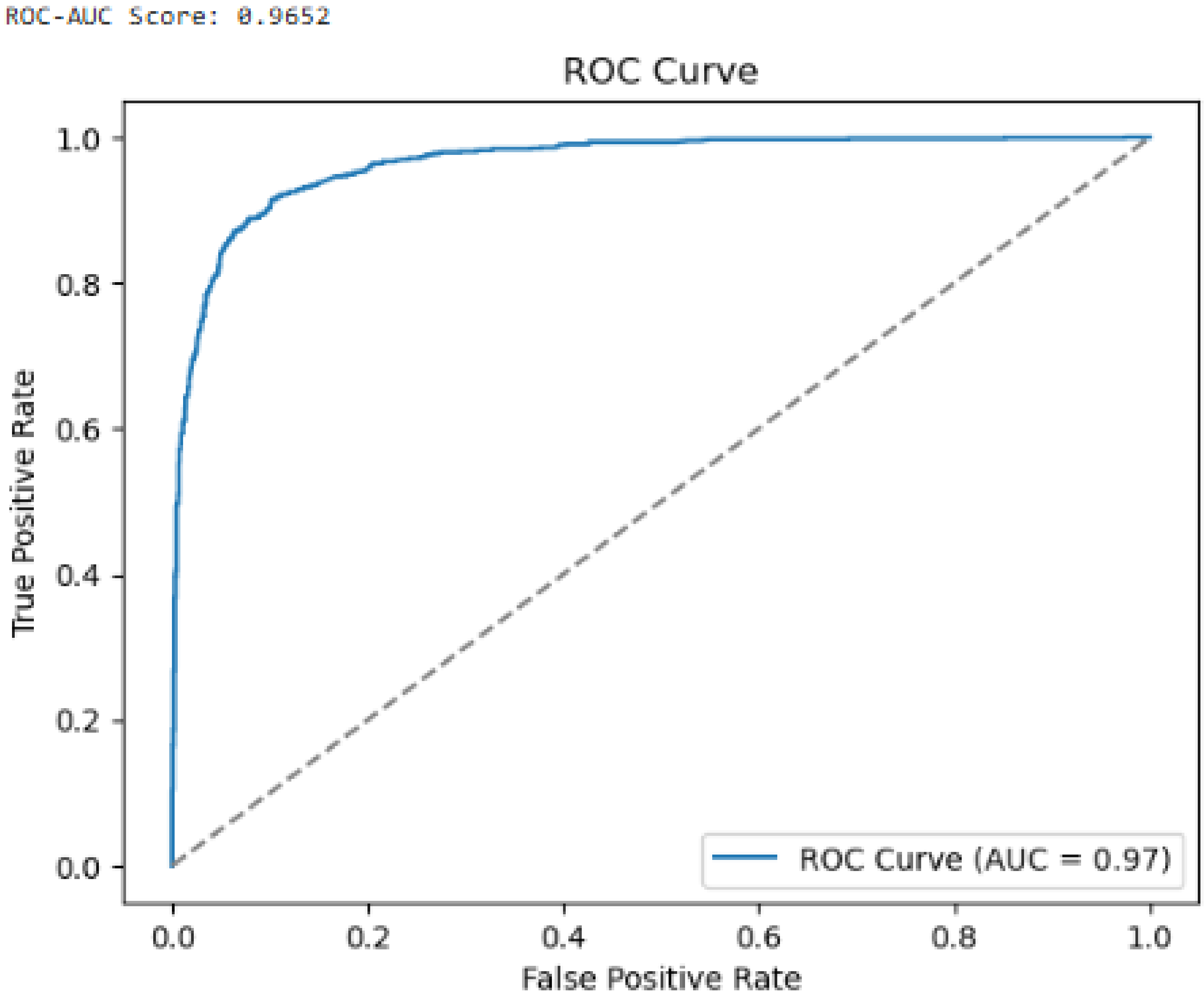
```
Evaluating Model Performance
Accuracy: 0.9098

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.94      0.93      1695
           1       0.90      0.87      0.88      1077

    accuracy                           0.91      2772
   macro avg       0.91      0.90      0.90      2772
weighted avg       0.91      0.91      0.91      2772
```

```
ROC-AUC Score: 0.9652
```



```
ROC-AUC Score: 0.9652
```

Lead Scoring

- Lead scores calculated based on model probabilities.
- Scale: 0-100
- Output File: Lead_Scores.csv

```
Step 5: Assigning Lead Scores
Lead Scoring Completed: Lead scores saved as 'Lead_Scores.csv'

Sample Lead Scores:
      TotalVisits  Total Time Spent on Website  Page Views Per Visit  \
4608     0.946584                    -0.430113              0.145670
7935    -0.298549                     0.805307             -0.166587
4043    -0.506071                    -0.886324             -0.632643
7821    -0.298549                    -0.300549             -0.166587
856     -0.091027                    -0.523180              0.299469

      Asymmetrique Activity Score  Asymmetrique Profile Score  \
4608                         14.0                        16.0
7935                         14.0                        16.0
4043                         14.0                        16.0
7821                         14.0                        18.0
856                          13.0                        18.0

      Lead Origin_Landing Page Submission  Lead Origin_Lead Add Form  \
4608                                 True                      False
7935                                 True                      False
4043                                 True                      False
7821                                 True                      False
856                                  True                      False
```

```
     Lead Origin_Lead Import  Lead Origin_Quick Add Form  \
4608                   False                       False
7935                   False                       False
4043                   False                       False
7821                   False                       False
856                    False                       False

     Lead Source_Direct Traffic  ...  \
4608                      False  ...
7935                       True  ...
4043                       True  ...
7821                       True  ...
856                       False  ...

     Last Notable Activity_Olark Chat Conversation  \
4608                                         False
7935                                         False
4043                                         False
7821                                         False
856                                          False

     Last Notable Activity_Page Visited on Website  \
4608                                         False
7935                                         False
4043                                         False
7821                                         False
856                                          False

     Last Notable Activity_Resubscribed to emails  \
4608                                        False
7935                                        False
4043                                        False
7821                                        False
856                                         False
```

```
   Last Notable Activity_SMS Sent  Last Notable Activity_Unreachable  \
4608                        False                              False
7935                        False                              False
4043                        False                              False
7821                        False                              False
856                         False                              False

   Last Notable Activity_Unsubscribed  \
4608                             False
7935                             False
4043                             False
7821                             False
856                              False

   Last Notable Activity_View in browser link Clicked  Lead_Score  \
4608                                              False       77.29
7935                                              False       10.01
4043                                              False        2.81
7821                                              False        3.39
856                                               False       15.93

   Actual_Converted  Predicted_Converted
4608               NaN                    1
7935               NaN                    0
4043               NaN                    0
7821               NaN                    0
856                0.0                    0

[5 rows x 167 columns]
```

Conclusion and Recommendations

1. Focus on leads with high lead scores (>80).
2. Improve engagement strategies for 'warm' leads (60-80).
3. Automate lead prioritization based on model output.
4. Monitor model performance and retrain periodically.