# Optimizing Lead Conversion : A Data-Driven Approach

## CASE STUDY PRESENTATION FOR X EDUCATION :


- VIBHUTI KOLADIYA
- MANISH KUMAR
- CHIRAG KHURANA

# Overview and Objectives :

- X Education faces challenges with lead conversion despite generating numerous leads daily.
- The company aims to enhance lead conversion by identifying potential leads or 'Hot Leads'.
- The lead conversion process involves nurturing potential leads to increase conversion rates.
- Objective: Build a logistic regression model to assign lead scores, prioritizing leads with higher conversion potential.
- CEO's target: Achieve an 80% lead conversion rate.
- Goals of the Case Study:
  - Develop a logistic regression model to assign lead scores ranging from 0 to 100.
  - Address additional company requirements and potential future adjustments in the model.

# Data Overview:

- The dataset comprises approximately 9000 data points.
- It includes various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.
- The dataset aims to predict whether a lead will be converted or not.

# Target Variable:

- The target variable is 'Converted', indicating whether a lead was converted (1) or not (0).

# Data Quality Considerations:

- Some categorical variables contain a level called 'Select', which needs to be addressed as it is treated as a null value.

# Data Preparation and Cleaning :

1. Handling Missing Values:
   - Columns with more than 35% missing values were dropped.
   - Imputation of missing values for other columns.
   - 'Not Specified' was used for some missing categorical values to retain data integrity.
2. Categorical Variables:
   - Conversion of categorical variables to dummy variables.
   - Removal of 'Select' levels as they represent missing values.
3. Numeric Variables:
   - Applied MinMaxScaler to scale numeric columns for consistency.
4. Data Splitting:
   - Split the dataset into training (70%) and testing (30%) sets.

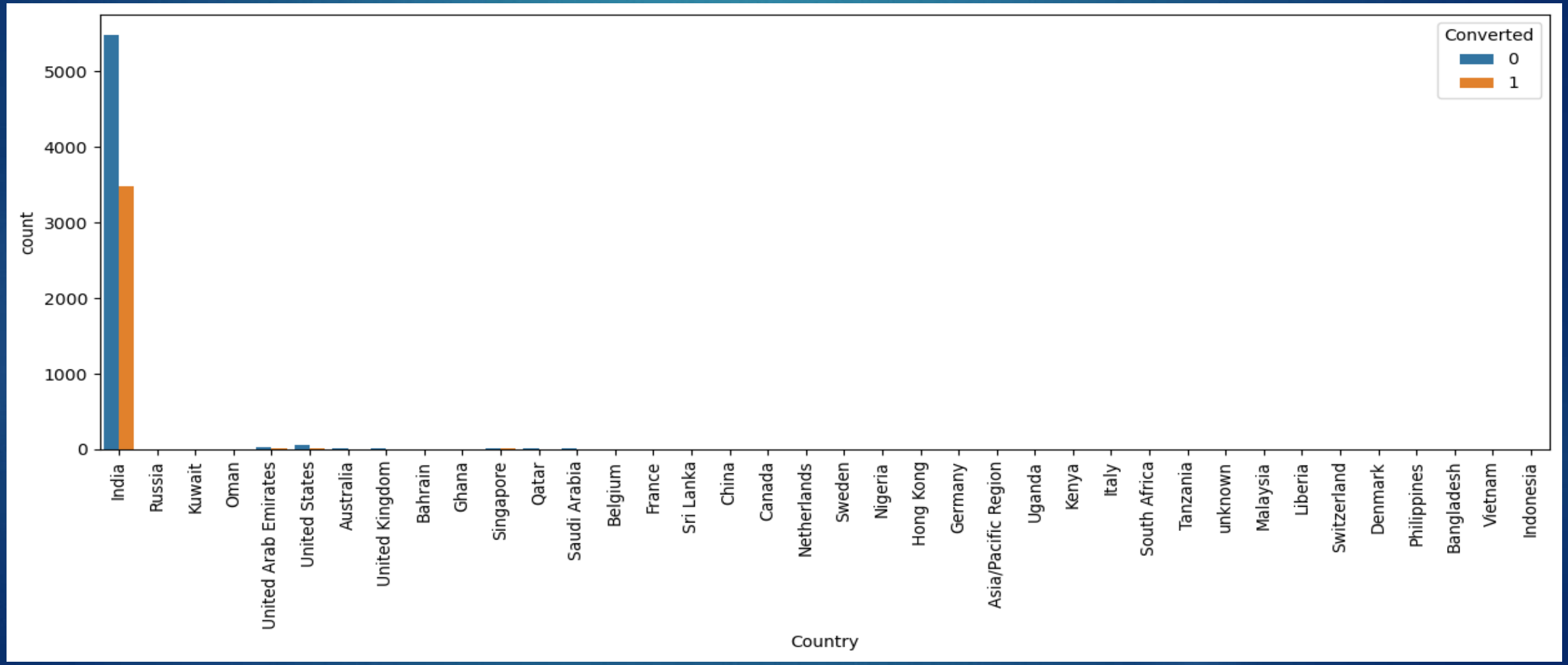
# Exploratory Data Analysis (EDA) :

- Distribution of Converted Leads:
  - Visualized the distribution of the target variable 'Converted’.
  - Approximately 30% of leads were converted.
- Lead Source Analysis:
  - Identified top lead sources with significant differences in conversion rates.
- Time Spent on Website:
  - Leads spending more time on the website had higher conversion rates, highlighting the importance of engaging website content.
- Lead Origin:
  - Certain origins, like 'Lead Add Form', had higher conversion probabilities.
- Last Activity:
  - Activities like 'SMS Sent' and 'Olark Chat Conversation' correlated with higher conversion rates.

# EDA Univariate Analysis :

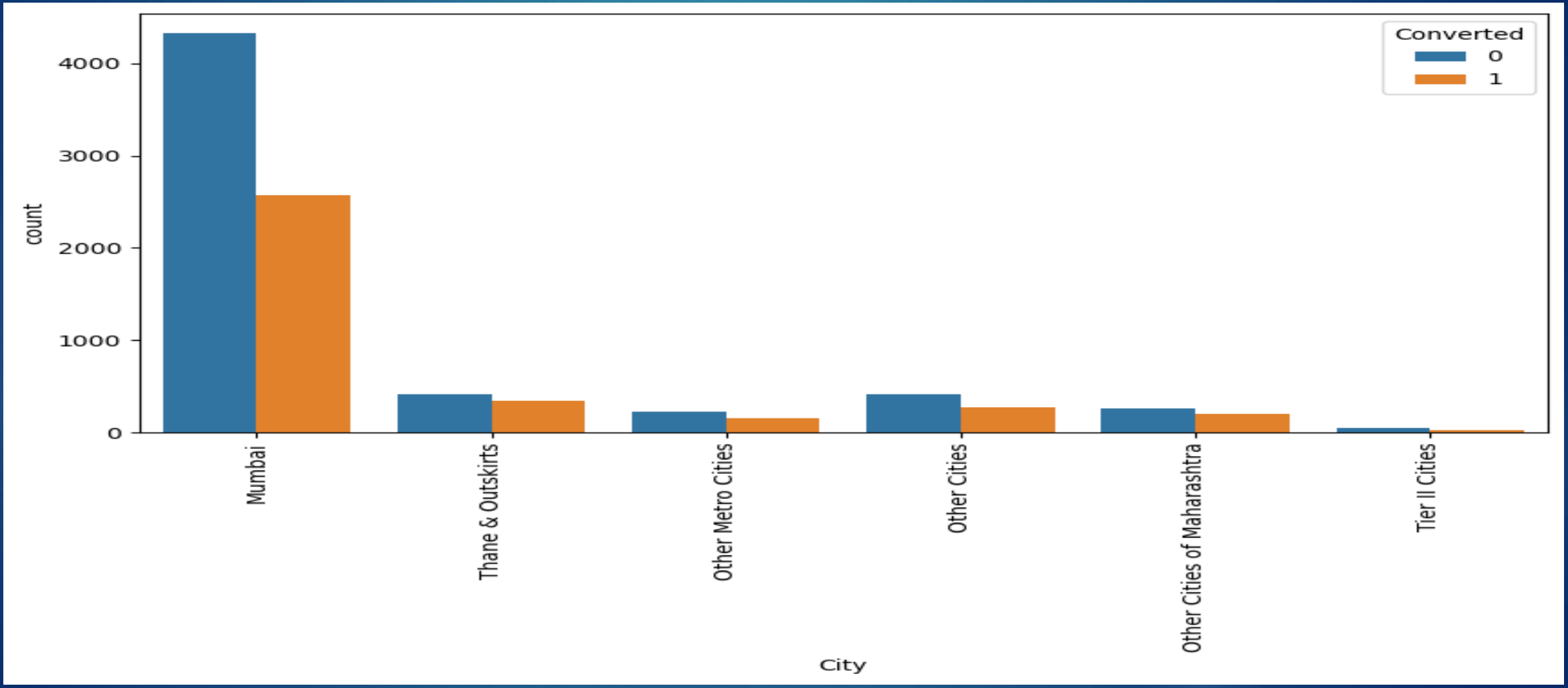
# Distribution of Leads by Country :

- As we can see, the number of leads from India is overwhelmingly high, comprising nearly 97% of the data.



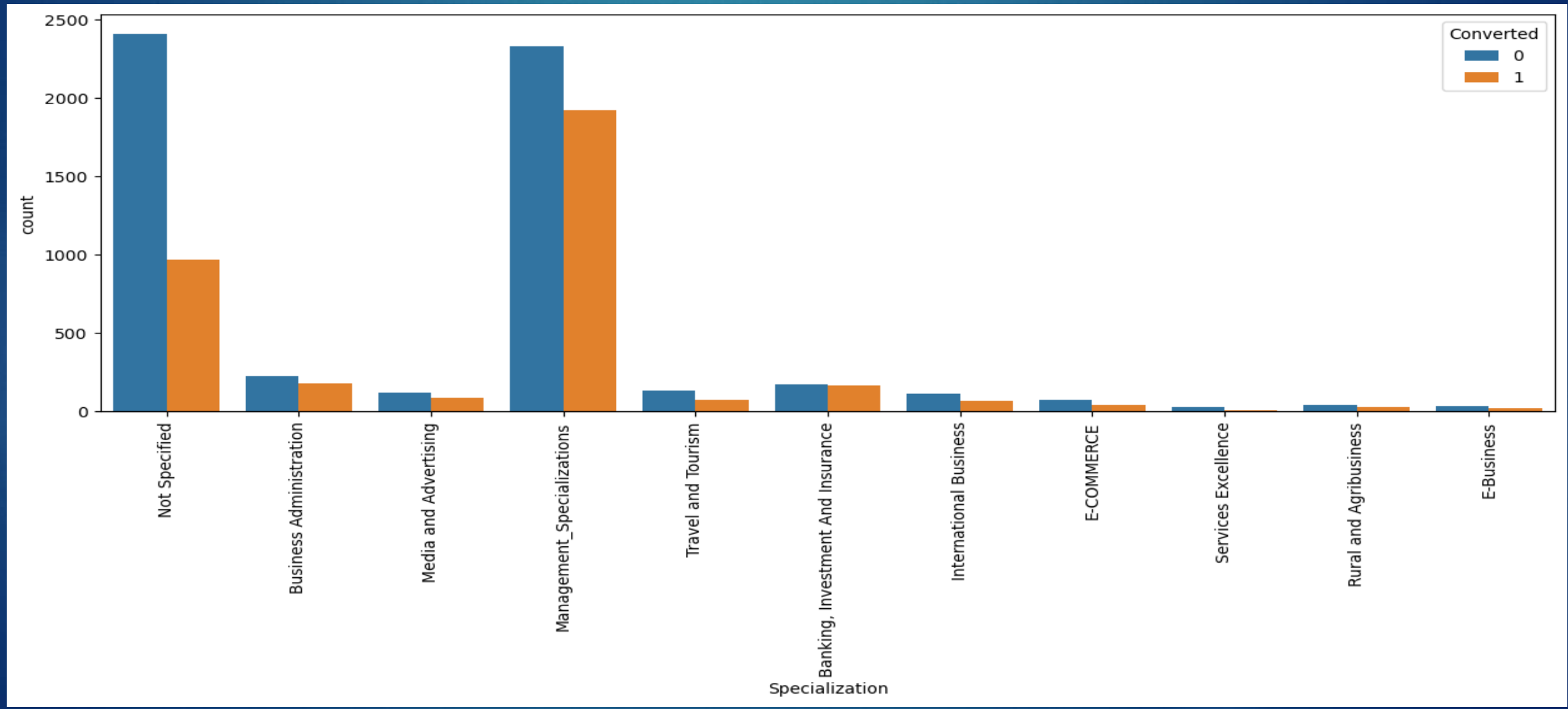
# Distribution of Leads by City :

- Most leads are from Mumbai, with few from other cities.



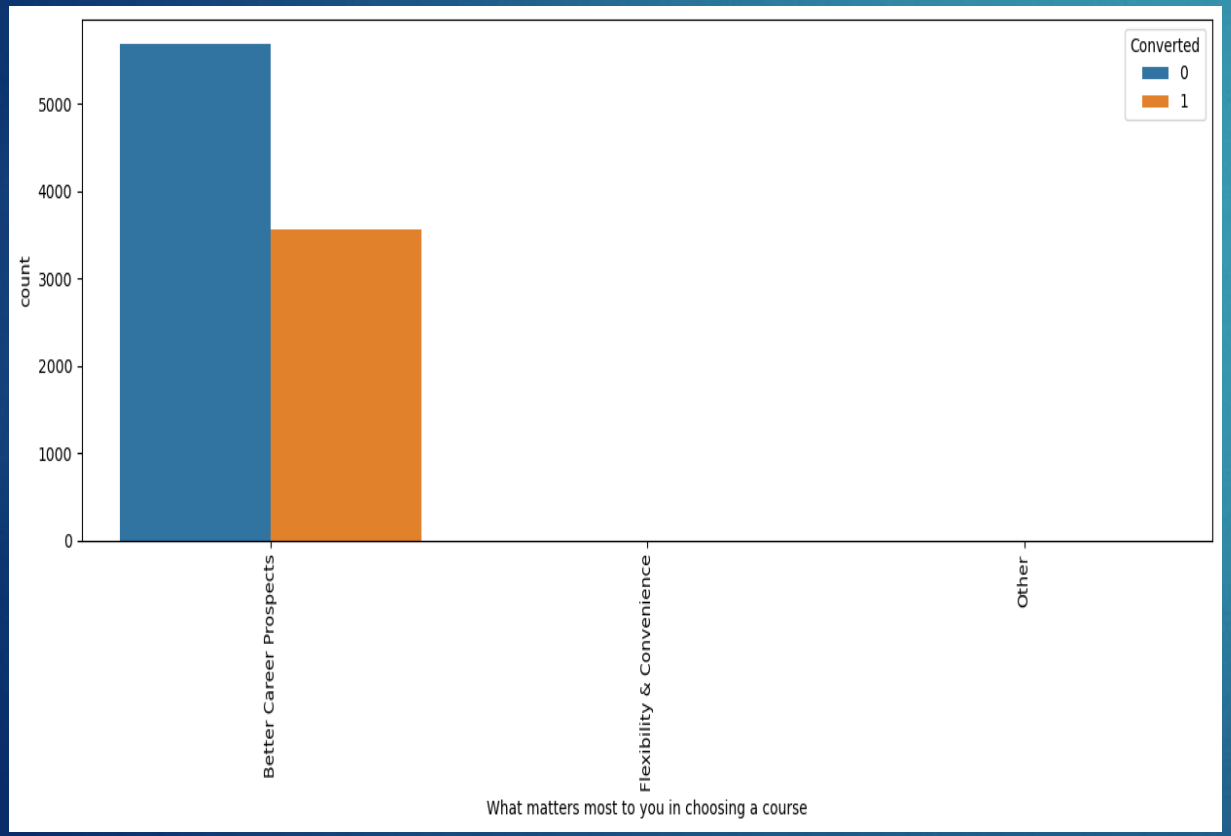
# Lead Conversion by Specialization :

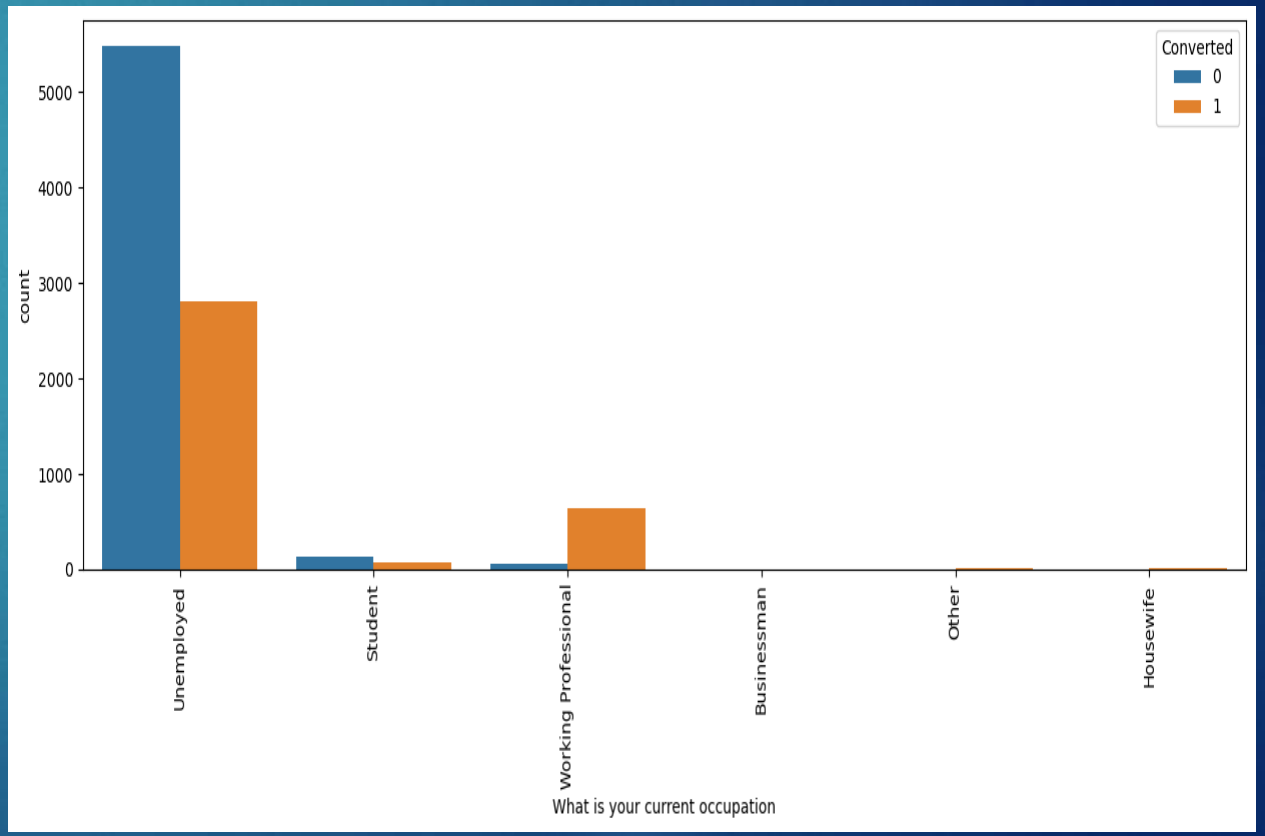
- Most leads are in 'Not Specified' or 'Management Specializations,' suggesting these categories dominate the dataset and should be closely analyzed.



# Lead Conversion by Current Occupation & by Course Selection Criteria :

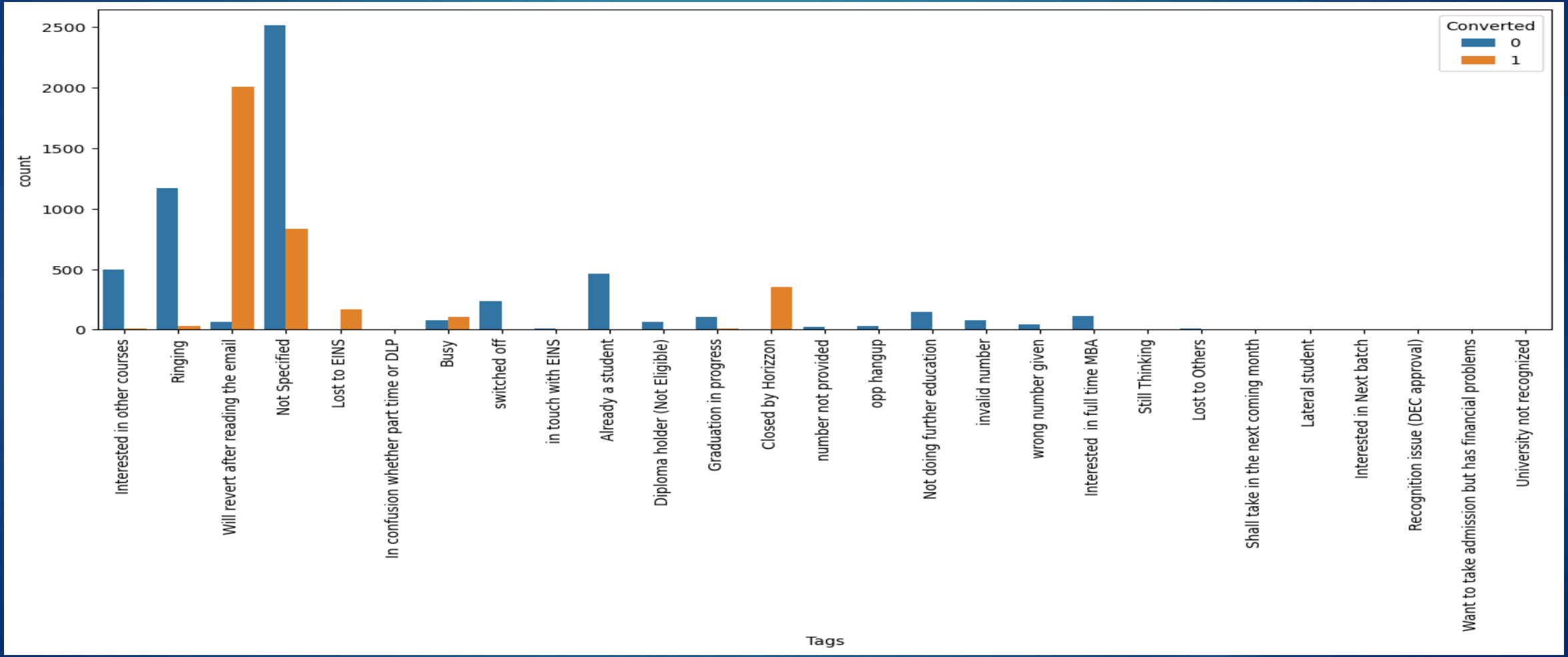
- People with unemployment are more likely to choose the courses for better career opportunities.
- Working professionals can be the next target for opting the courses for better career prospects.



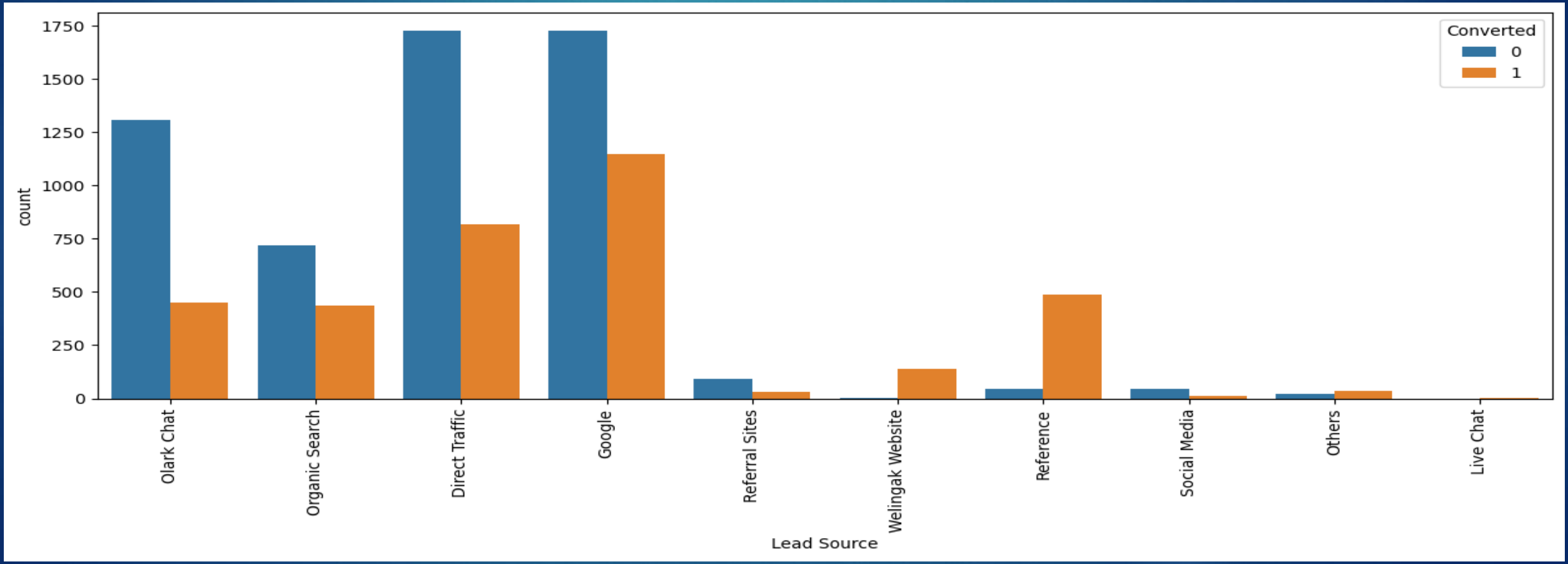
# Lead Conversion by Response to Email :

- People who respond to emails are more likely to opt for courses, while those tagged as 'Ringing' have a lower conversion rate and can be deprioritized.



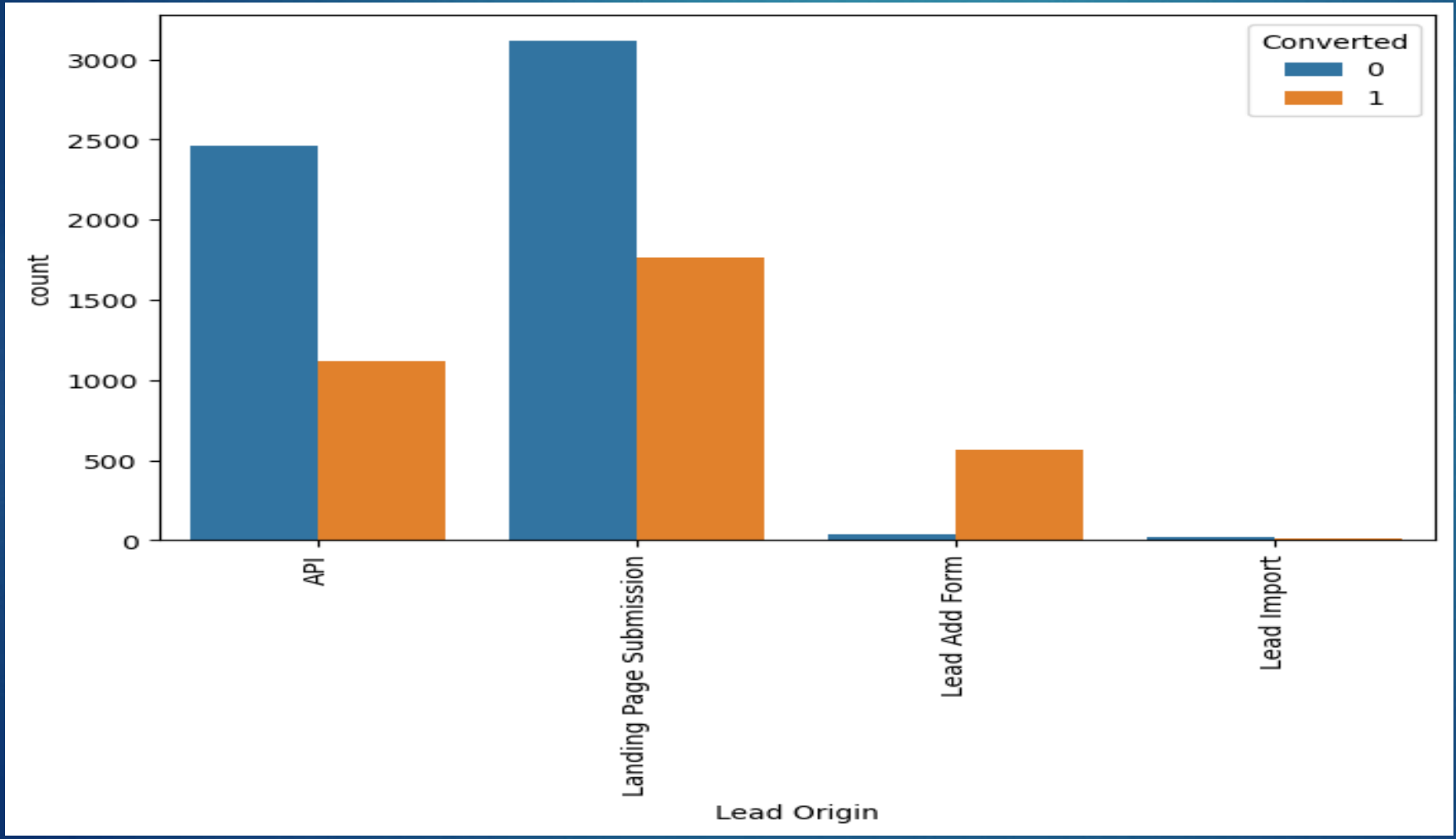
# Leveraging Lead Source Diversity for Enhanced Conversion :

- By directing focus towards optimizing conversion rates of Olark Chat, organic search, and Google leads, while also increasing lead generation from reference and Welingak website, we can amplify overall conversion rates and foster robust growth.



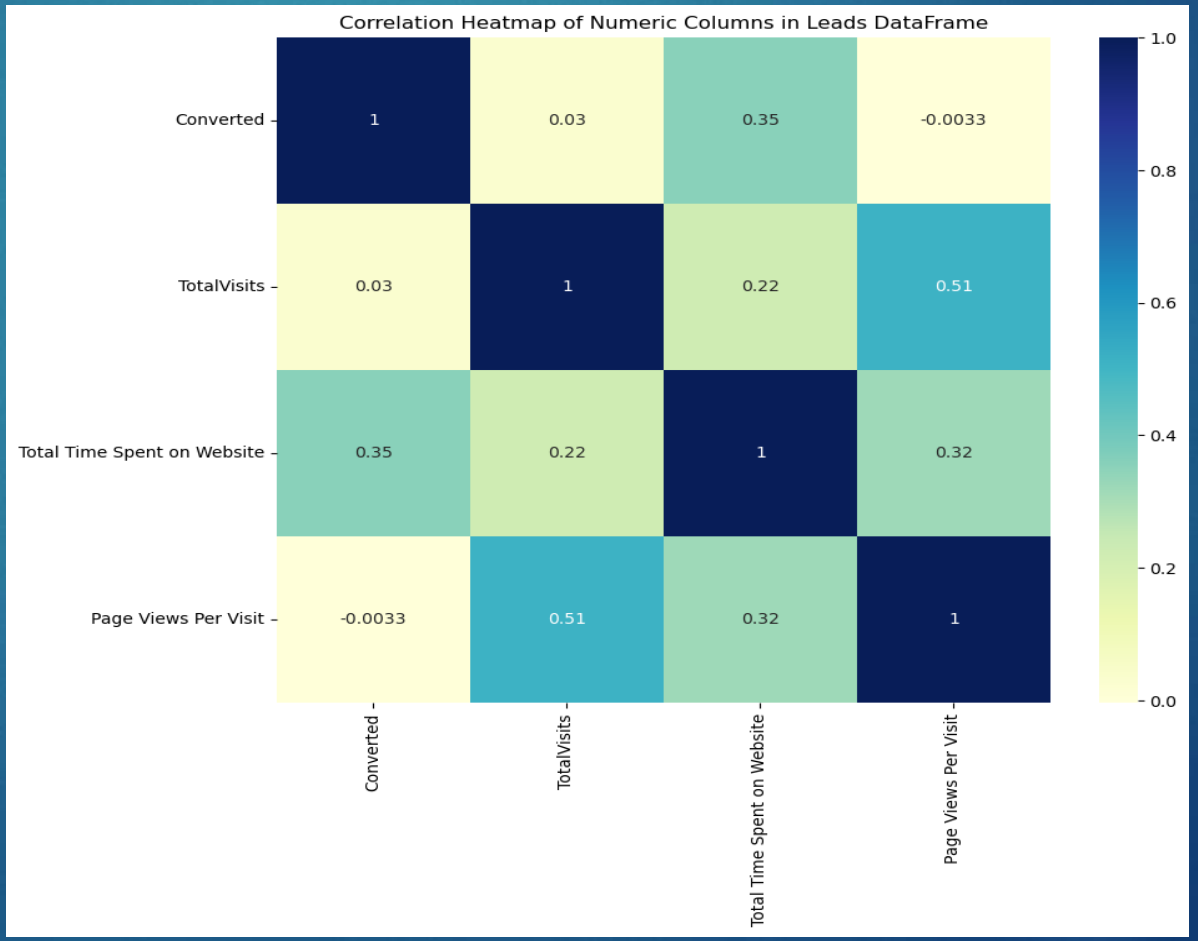
# Originating Conversion Pathways

- Applicant who landed on the submission page is likely to opt for the courses.
- Followed by API and Lead Ad form.



# Unveiling Numeric Relationships: Correlation Heatmap of Leads Data Frame

- Conversion positively links with time spent on site but negatively with page views per visit. Total visits align with page views. Total time on site positively affects page views but is less tied to total visits.



# Feature Selection and Model Building :

- Feature Selection:
  - Select relevant features to improve model accuracy.
  - Use techniques like RFE or VIF analysis.
- Model Building:
  - Develop predictive models using selected features.
  - Iterate through preprocessing, model selection, and evaluation.
- Integration:
  - Feature selection directly impacts model performance.
  - Ensure selected features align with model objectives.
- Optimization:
  - Regularization and cross-validation enhance model robustness.
  - Validate assumptions to ensure model reliability.

# Model Evaluation & Performance Metrics :

- Evaluation Process:
  - Assess model performance using various metrics.
  - Compare predicted outcomes with actual data.
- Key Metrics:
  - Accuracy, precision, recall, and F1-score.
  - ROC curve and AUC for binary classification.
- Interpretation:
  - Understand strengths and weaknesses of the model.
  - Choose metrics aligned with project goals.
- Improvement:
  - Iteratively refine models based on evaluation results.
  - Optimize hyperparameters for better performance.

# Model Evaluation and Performance Metrics :

- Overview of Model Evaluation Techniques:
  - Assess model performance with various metrics (e.g., accuracy, recall) and tools (e.g., confusion matrix, ROC curve).
- Confusion Matrix Analysis and Performance Metrics.
  - Confusion Matrix :
    - ✓ TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative
  - Key Metrics:
    - ✓ Accuracy: Correct predictions: 88%
    - ✓ Sensitivity (Recall): Identified positives: 82%
    - ✓ Specificity: Identified negatives: 91%
- ROC Curve Analysis and AUC:
  - ROC Curve: Plots sensitivity vs. 1-specificity.
  - AUC: 0.93

# Insights & Recommendations :

- Key Insights:
  - Top Predictive Features: Customer demographics, interaction history, and campaign engagement are key influencers of lead conversion.
  - Model Reliability :
    - ✓ Train Data: 92.29% accuracy, 91.70% sensitivity, 92.66% specificity.
    - ✓ Test Data: 92.78% accuracy, 91.98% sensitivity, 93.26% specificity.
- Recommendations:
  - People with unemployment are more likely to choose the courses for better career opportunities.
  - Working professionals can be the next target for opting the courses for better career prospects.
  - Google search engine can bring lot of business to the company as the Lead conversion is very high compared to other sources followed by Reference.
  - People who revert the email are more likely to opt for the courses

# Conclusion and Data-Driven Strategies :

- Summary of Findings:
  - Identified key predictors of lead conversion.
  - Achieved high model performance with accuracy rates exceeding 92%.
  - Discovered distinct behavioral patterns among customer segments.
- Implications for X Education:
  - Insights offer opportunities to enhance lead conversion efforts.
  - Implementation of data-driven strategies can improve marketing effectiveness.
- Recap of Data-Driven Strategies for Lead Conversion Optimization:
  1. Feature Prioritization
  2. Data Quality Enhancement
  3. Regular Model Updates
  4. Segment-Specific Approaches

# Thank You!