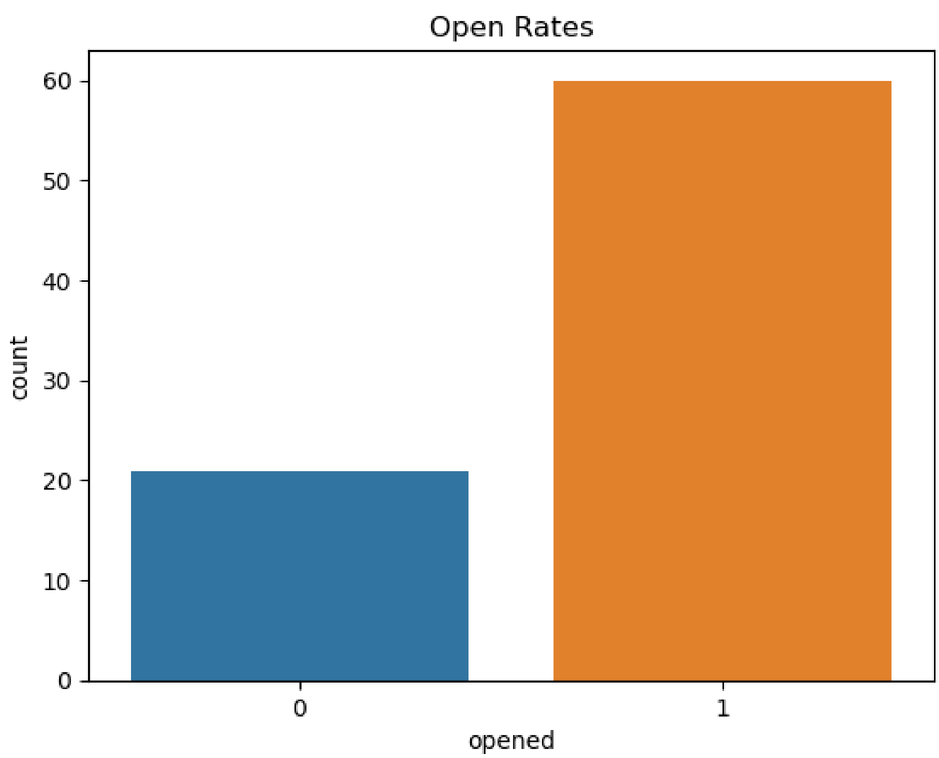
Email engagement model

➢ **Exploratory Data Analysis (EDA) and Visualization:**

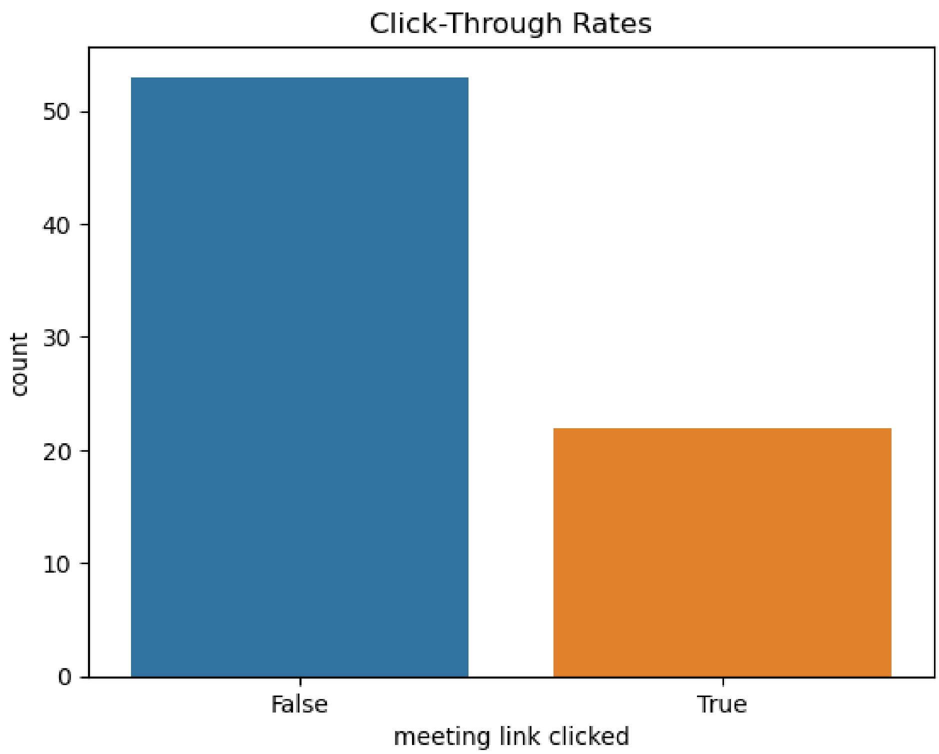
Open Rate Visualization:

- Analyzed the distribution of email open rates.
- Open rates were categorized as 0 (not opened) and 1 (opened).



Click-Through Rate Visualization:

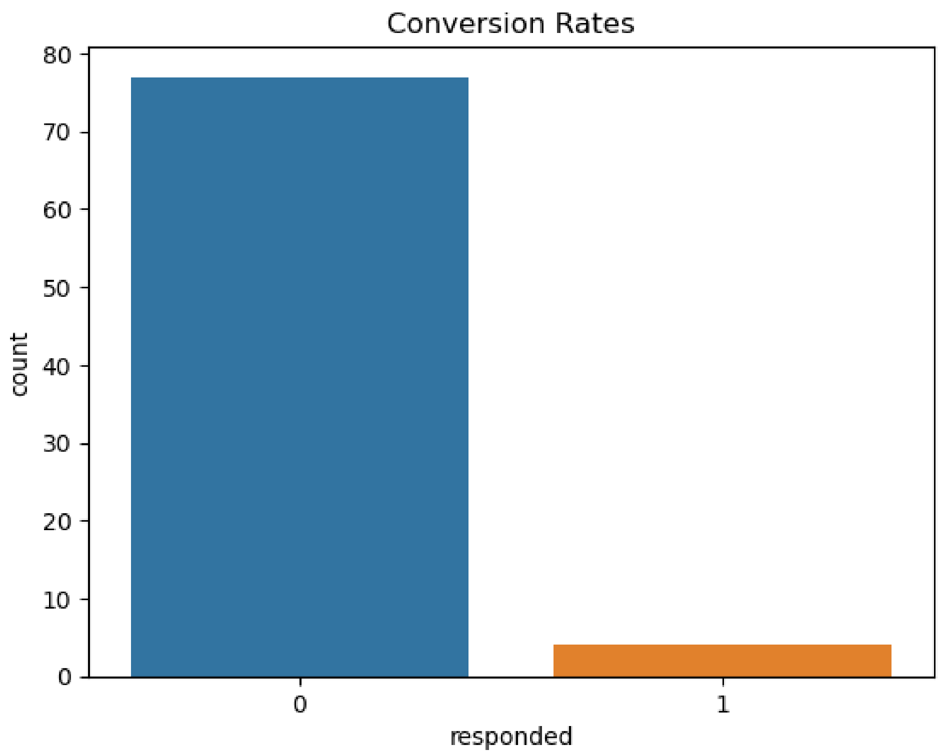
- Investigated the click-through rate based on the presence of a meeting link in the email body.
- Utilized a binary indicator (False: no click, True: click) to determine engagement.



Conversion Rate Visualization:

- Explored the conversion rate by analyzing user responses to emails.
- Responses were represented as 0 (no response) and 1 (response).

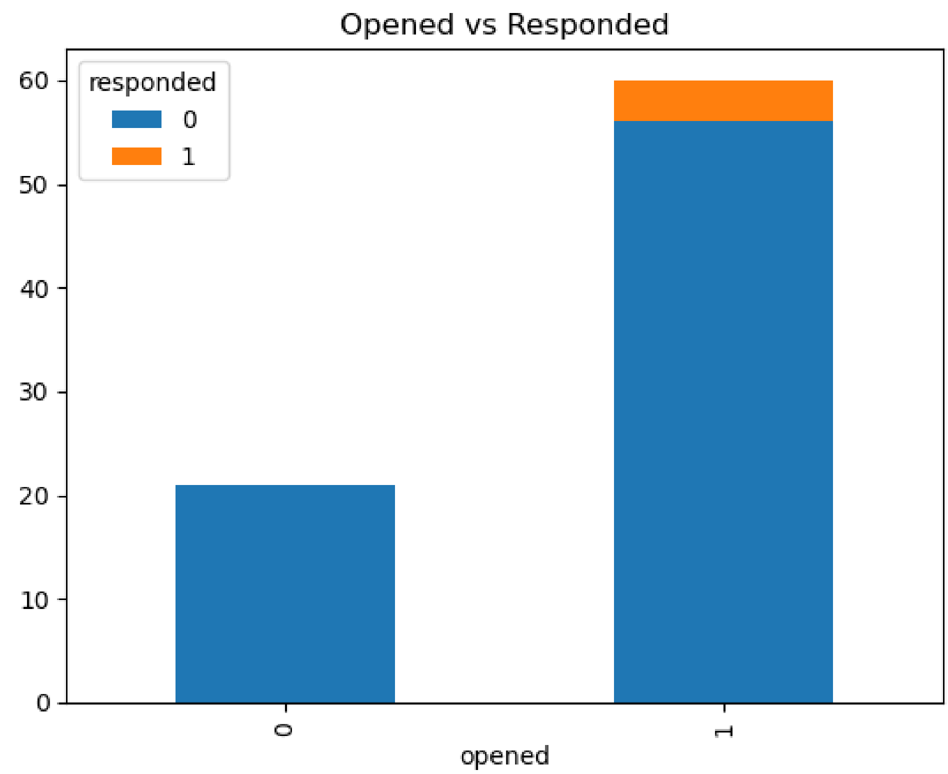
Responses serve as a key metric for email engagement, indicating user interest and interaction. A higher response rate signifies effective communication and potential for conversion. Analyzing responses enables continuous improvement in content and targeting strategies for impactful email campaigns.



Cross-tab Visualization:

- Examined the relationship between open rate and conversion using a cross-tabulation technique.

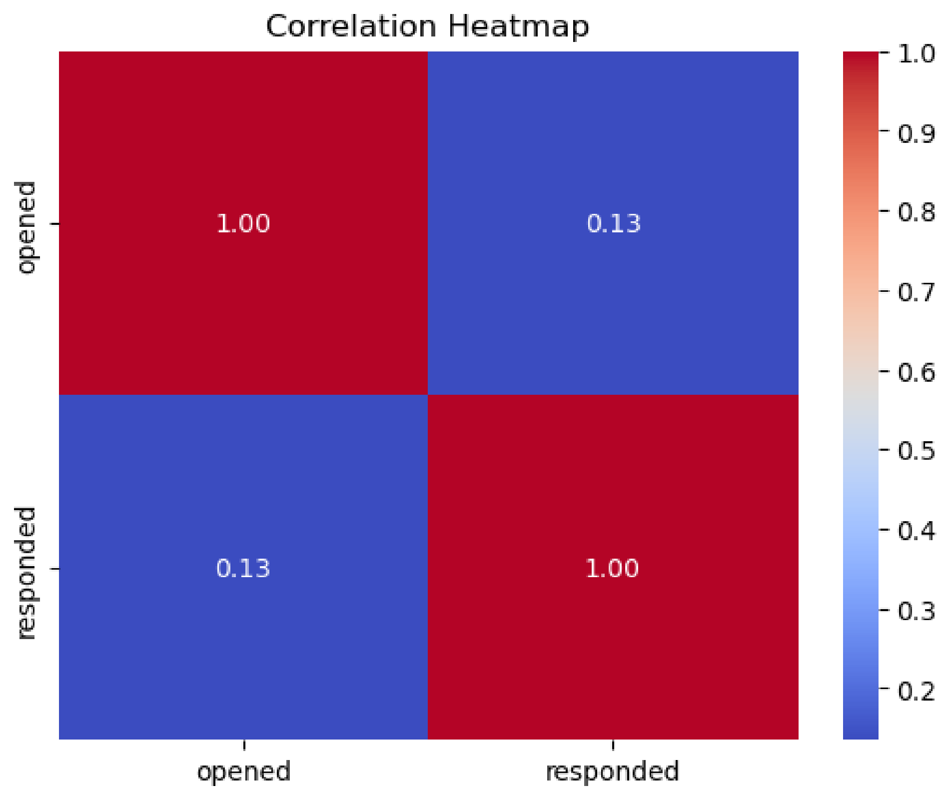
Cross-tab visualization, also known as a contingency table, illustrates the joint distribution of two categorical variables. Each cell in the table represents the count or percentage of observations that fall into a specific combination of categories. This technique helps identify patterns, dependencies, and associations between the variables, providing valuable insights for categorical data analysis. By visually exploring the cross-tabulation, researchers can gain a comprehensive understanding of the relationships within the dataset. This method is widely employed in fields such as market research, social sciences, and data exploration to unveil categorical data patterns and trends.



Correlation Heatmap:

- Constructed a heatmap to visualize correlations between attributes, particularly focusing on open rate and response rate.

A correlation heatmap visually represents the correlation coefficients between pairs of variables in a dataset. Using a color scale, it quickly highlights the strength and direction of linear relationships between variables. Positive correlations are often depicted in warmer colors, while negative correlations are in cooler colors. This visualization aids in identifying patterns, multicollinearity, and potential insights for feature selection in statistical analyses. Widely utilized in data exploration, the heatmap is a valuable tool for assessing the interdependencies of variables and guiding further modeling decisions in areas such as finance and machine learning.



- **Data Preprocessing:**

Handling Missing Values:

- Executed data preprocessing steps, specifically removing rows with missing values in the dataset.

- **Model Development:**

Model Selection:

- Chose the Random Forest Classifier as the initial model due to its flexibility and ability to handle complex datasets.

Base Prototype:

- Implemented a base prototype of the Random Forest Classifier.
- Achieved an initial accuracy of 60%.

Hyperparameter Tuning:

- Conducted hyperparameter tuning using GridSearchCV to optimize model performance.
- Achieved a significant accuracy improvement to 73% with the tuned Random Forest Classifier.

- **Ensemble Model Exploration:**

Model Selection:

- Explored an ensemble approach with three models (Random Forest Classifier, Support Vector Classifier, Gradient Boosting Classifier).

Time Constraints:

- Due to time constraints, deferred the implementation of the ensemble model.
- Reserved it for future refinement and optimization.

- **Suggestions for Further Improvement:**

Feature Engineering:

- We can consider further feature engineering to identify additional variables that may enhance model predictive power.

Model Refinement:

- We can also explore advanced algorithms or ensemble techniques to improve model accuracy when time permits.