TELECOM CHURN PREDICTION PROJECT REPORT

BUSINESS UNDERSTANDING

Overview

Customer churn is a critical challenge for telecommunications companies, including SyriaTel. Churn occurs when customers stop using the company's services, leading to revenue loss and increased acquisition costs to replace those customers. Retaining existing customers is far more cost-effective than acquiring new ones, making churn prediction vital for business sustainability.

Challenges:

- **Identifying patterns of churn behavior**: Customer decisions can be influenced by various factors such as service quality, pricing, or competitive offers.
- **Data quality and complexity**: Handling large datasets with diverse customer attributes and ensuring accuracy.
- **Actionable insights**: Converting predictions into strategies that effectively reduce churn.

Stakeholders:

- **Business Executives**: Focused on financial impact and strategic planning.
- **Customer Service Teams**: Need insights to engage and retain at-risk customers.
- **Data Science Team**: Responsible for developing, validating, and deploying predictive models.

Proposed Solution (Analysis & Modelling):

We will use historical customer data to build and evaluate machine learning models. The process involves:

1. Data preprocessing and exploratory analysis.
2. Training and comparing multiple classifiers (e.g., logistic regression, Decision Trees).
3. Selecting the best model based on performance metrics.

Projected Conclusion:

The project aims to identify key indicators of customer churn and provide a robust predictive model. This will enable SyriaTel to proactively address churn risks, potentially improving customer retention and increasing revenue.

Problem Statement

How can SyriaTel predict and reduce customer churn using historical data to improve customer retention and minimize revenue loss?

Objectives:

1. Develop a machine learning model to predict customer churn.
2. Identify key factors contributing to churn.
3. Provide actionable insights for customer retention strategies.

Metrics of Success:

- **Accuracy Score:** Measures overall correctness of predictions.
- **Precision Score:** Ensures focus on true churners among predicted churners.
- **Additional Metrics:** ROC-AUC and F1-score for balanced evaluation.

Business Metrics:

- Reduced churn rate.
- Increased Customer Lifetime Value (CLV).
- Improved ROI on retention strategies.

These metrics ensure that the model accurately identifies churn cases while minimizing false positives, allowing for efficient targeting of retention efforts.

Data Understanding

Data Source:

The dataset is sourced from Kaggle through github(**SyriaTel Customer Churn**).

The data is based on a Syrian Telecom company called SyriaTel, containing customer attributes and churn status.

Data Size:

- **Rows:** 3,333
- **Columns:** 21

Columns:

- **Churn:** Target variable (1 = churned, 0 = not churned).

Categorical Features:

- state
- area code
- international plan
- voicemail plan

Continuous Columns:

- account length
- number vmail messages
- total day minutes
- total day calls
- total day charge
- total eve minutes
- total eve calls
- total eve charge
- total night minutes
- total night calls
- total night charge
- total intl minutes
- total intl charge
- customer service calls

DATA PREPARATION AND ANALYSIS

Data Checks and Handling:

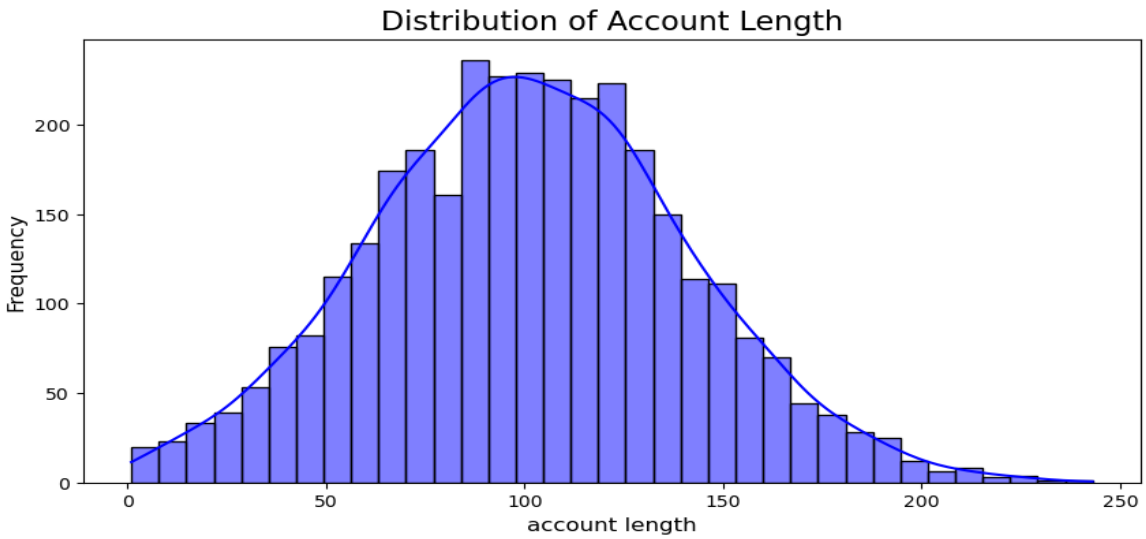
1. **Missing Values:**
    - There were no missing values.
2. **Duplicate Values:**
    - There were no duplicate values.
3. **Outliers:**
    - Detected in numeric columns using boxplots.
    - Addressed using IQR-based clipping.
4. **Null Values:**
    - There were no null values.

The data had already been cleaned from our source.

Exploratory Data Analysis :

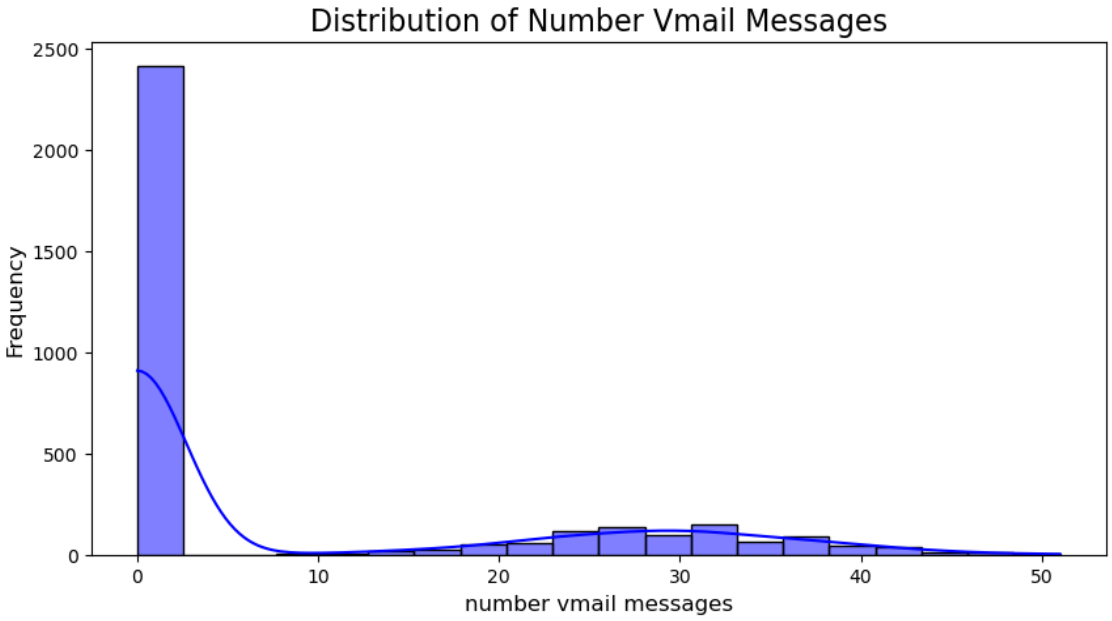
- **Univariate Analysis:**
  Distribution plots for numeric features and bar charts for categorical features.

  Most of the continuous features have a normal distribution **.**



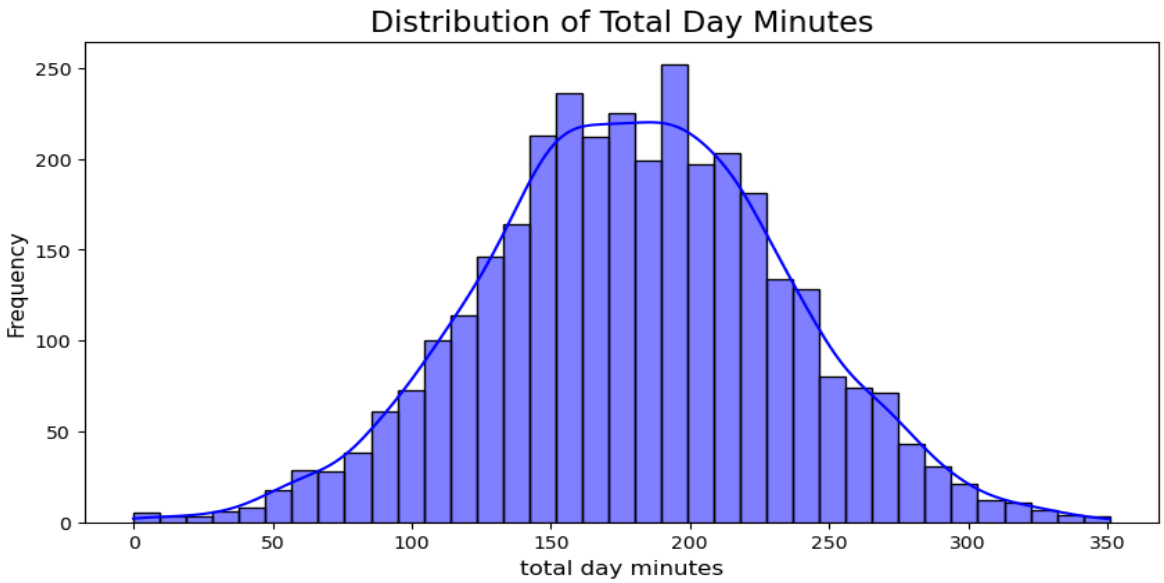
Observations

- The account length appears to have a normal distribution (Skewness = 0.10).



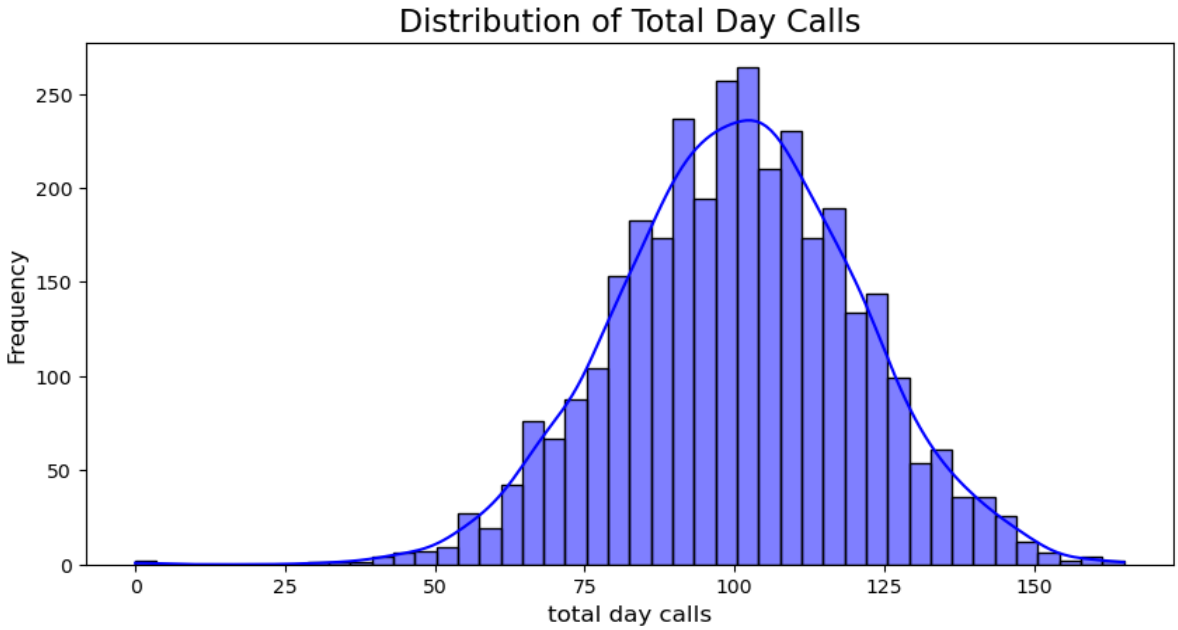
Observations

- The number vmail messages appears to have a skewed distribution (Skewness = 1.26).



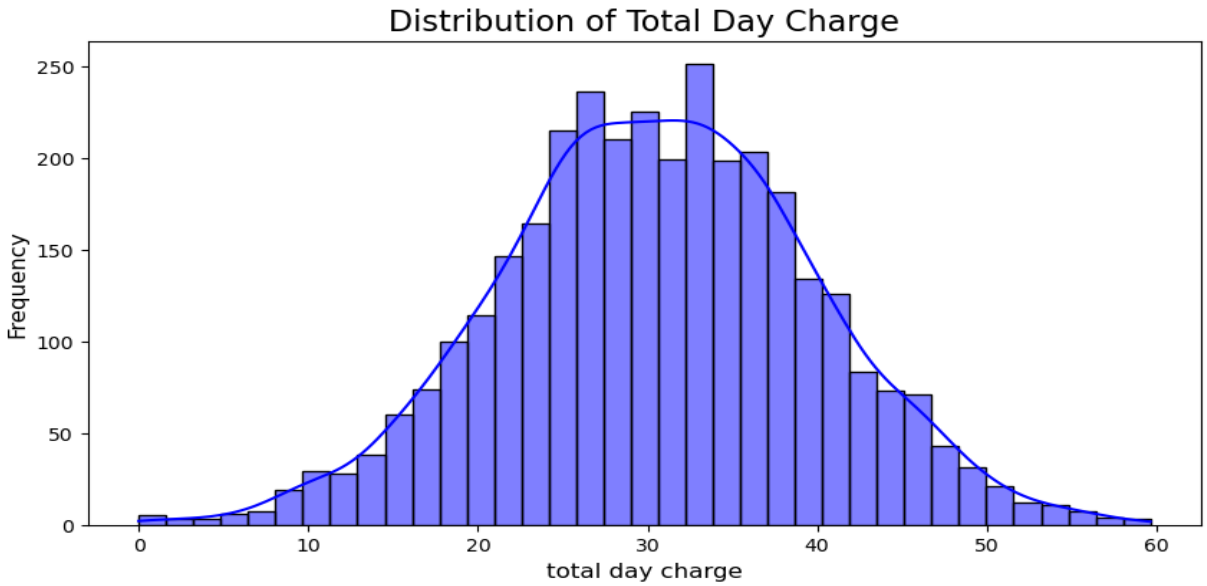
Observations

- The total day minutes appears to have a normal distribution (Skewness = -0.03).



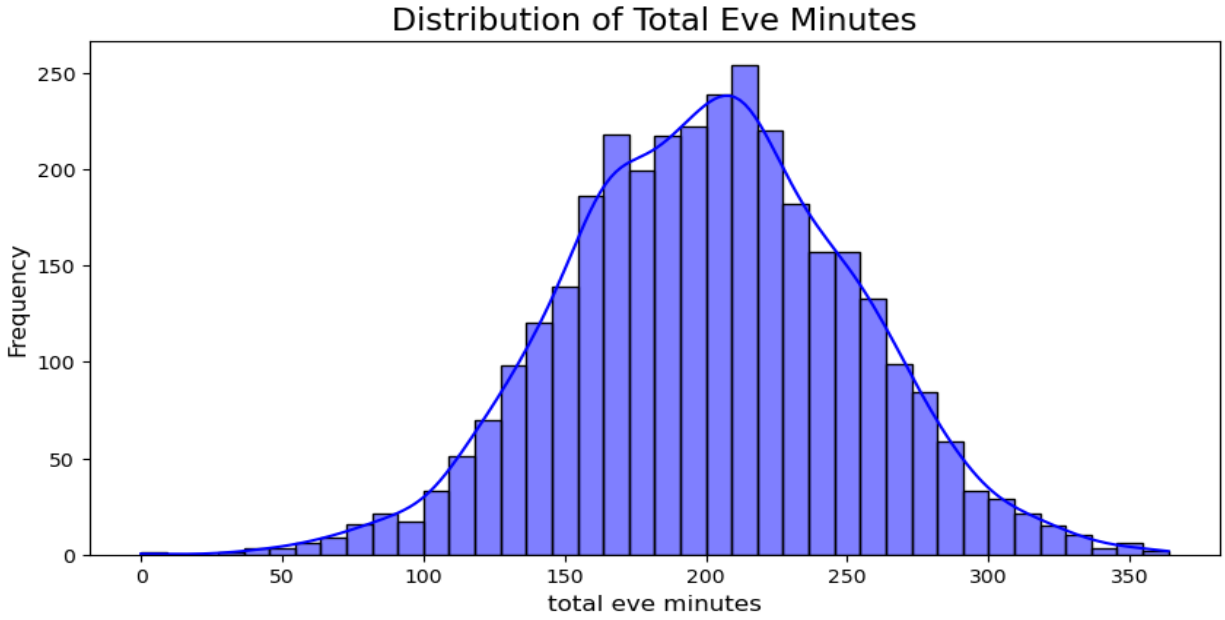
Observations

- The total day calls appears to have a normal distribution (Skewness = -0.11).



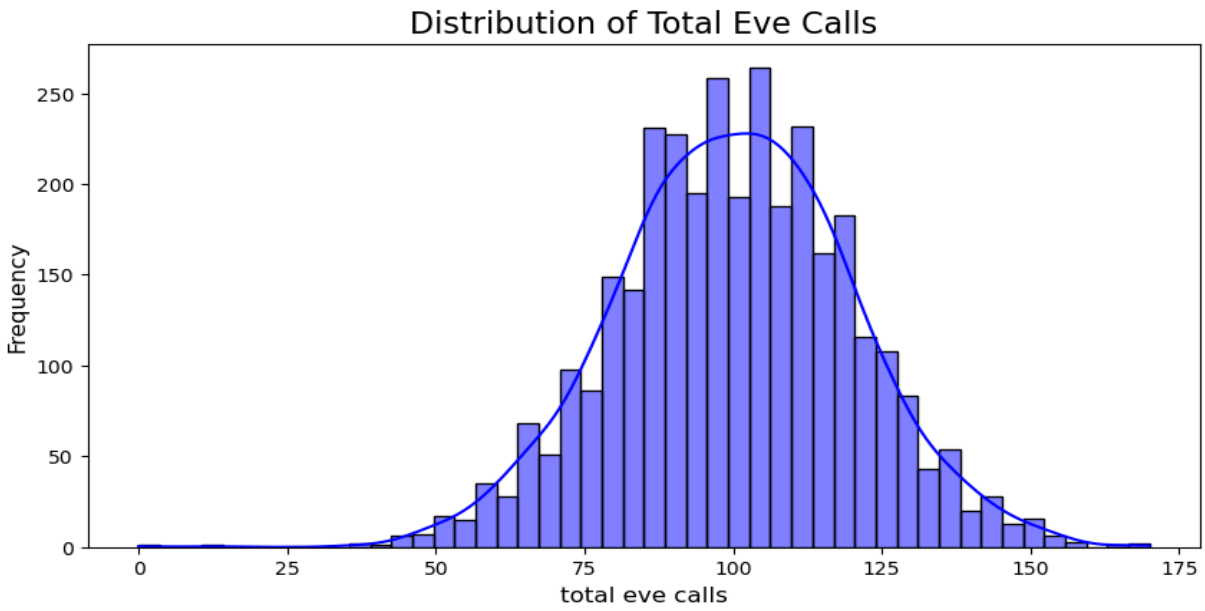
Observations

- The total day charge appears to have a normal distribution (Skewness = -0.03).



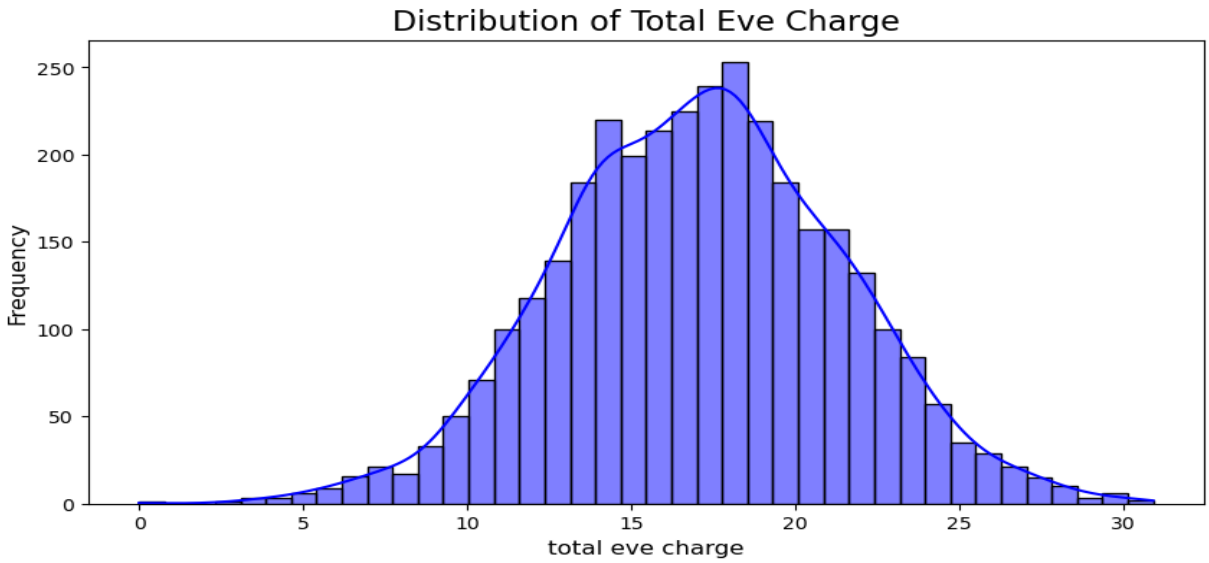
Observations

- The total eve minutes appears to have a normal distribution (Skewness = -0.02).



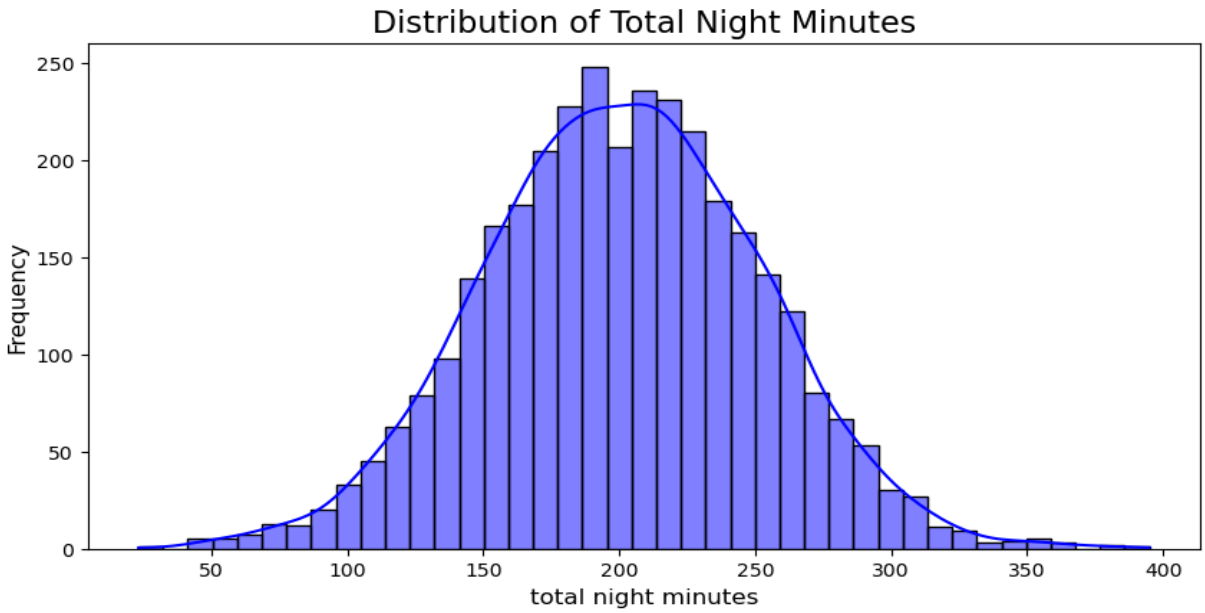
Observations

- The total eve calls appears to have a normal distribution (Skewness = -0.06).



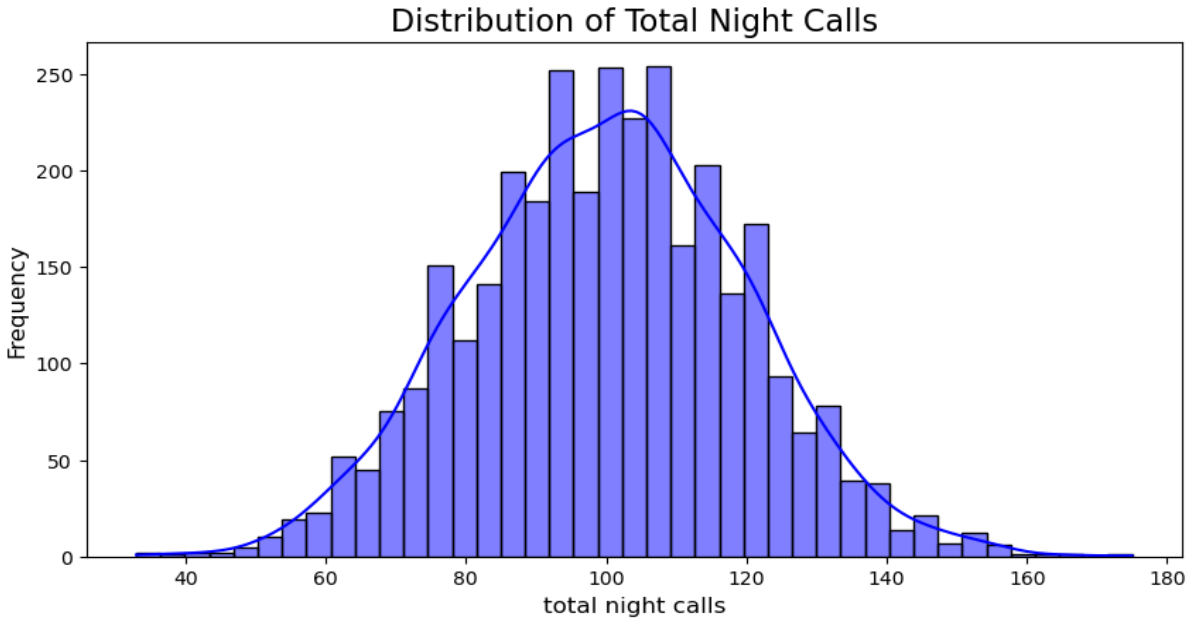
Observations

- The total eve charge appears to have a normal distribution (Skewness = -0.02).



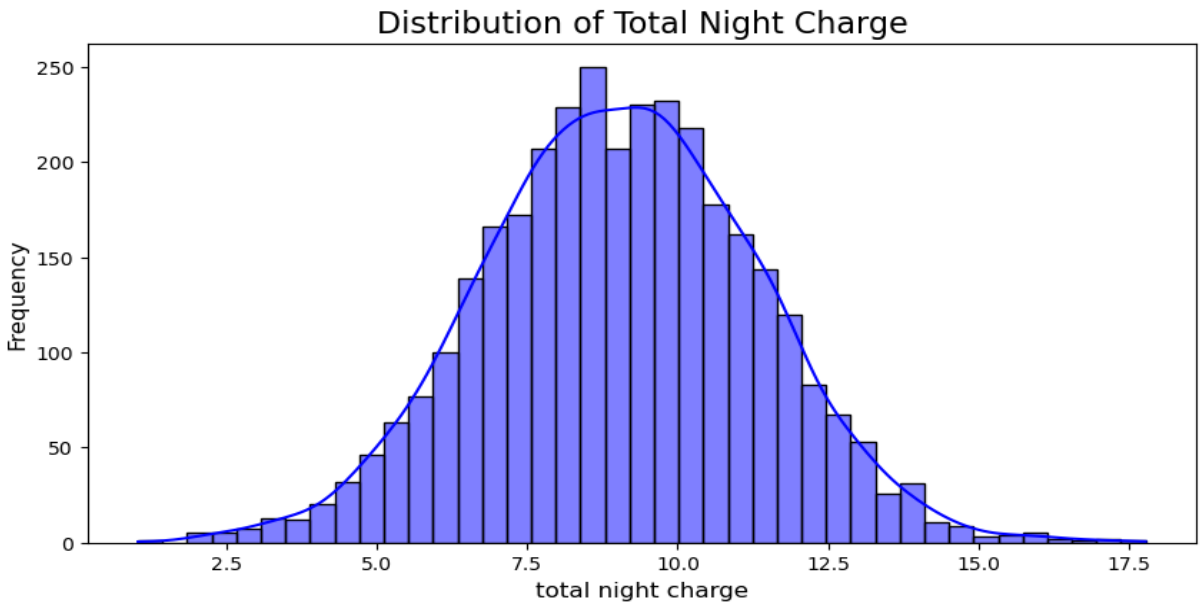
Observations

- The total night minutes appear to have a normal distribution (Skewness = 0.01).



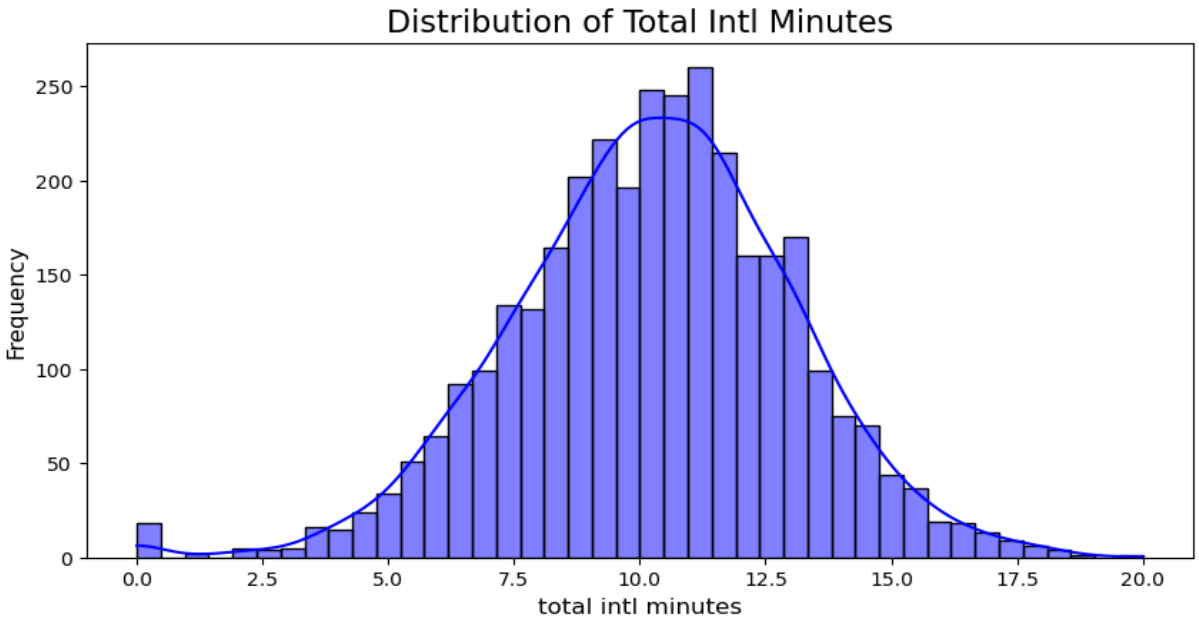
Observations

- The total night calls appear to have a normal distribution (Skewness = 0.03).



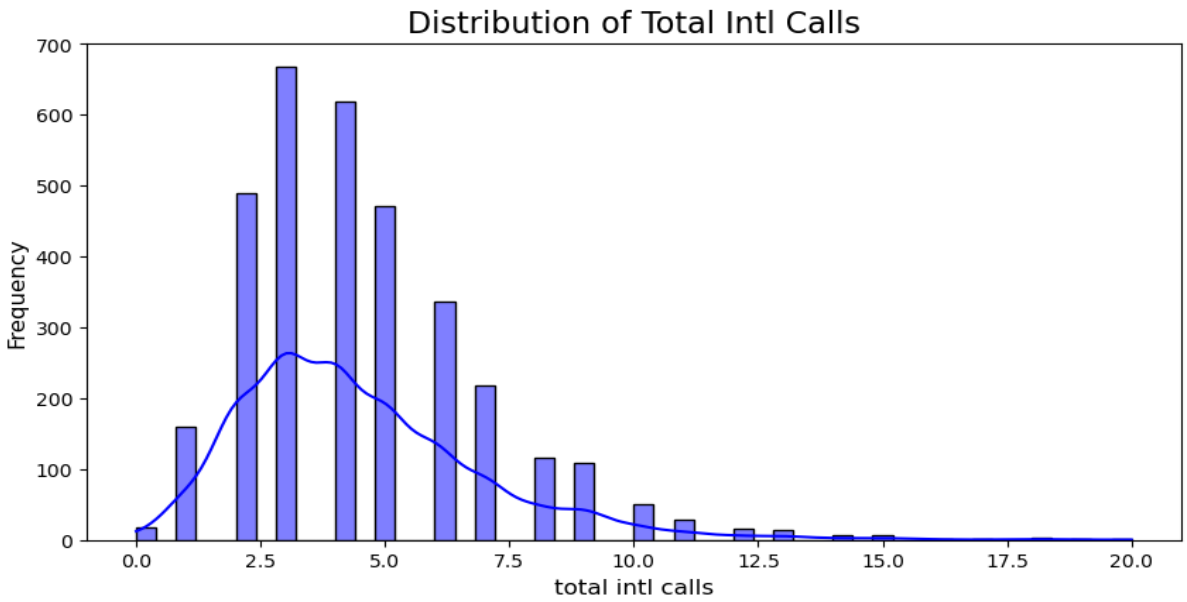
Observations

- The total night charge appears to have a normal distribution (Skewness = 0.01).



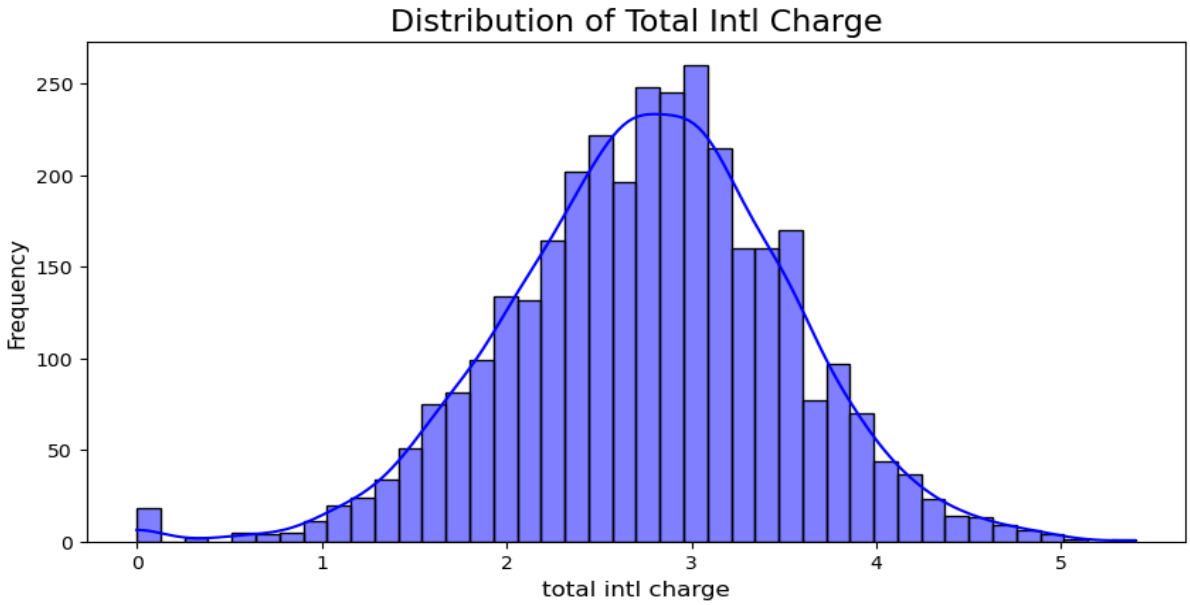
Observations

- The total intl minutes appear to have a normal distribution (Skewness = -0.25).



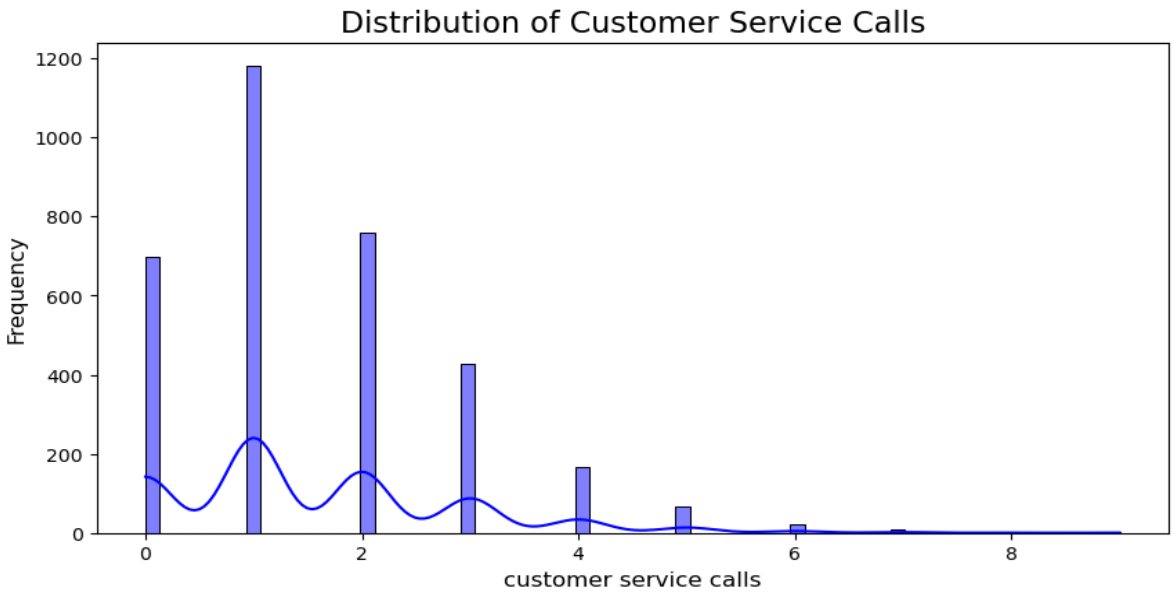
Observations

- The total intl calls appears to have a skewed distribution (Skewness = 1.32).



Observations

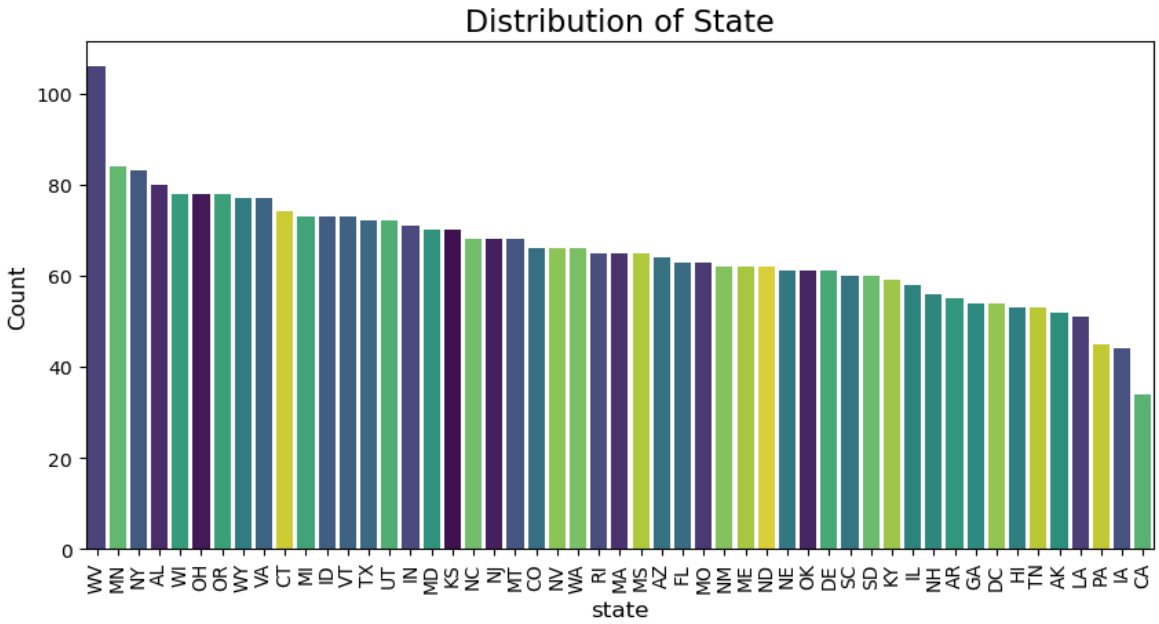
- The total intl charge appears to have a normal distribution (Skewness = -0.25).



Observations

- The customer service calls appears to have a skewed distribution (Skewness = 1.09).

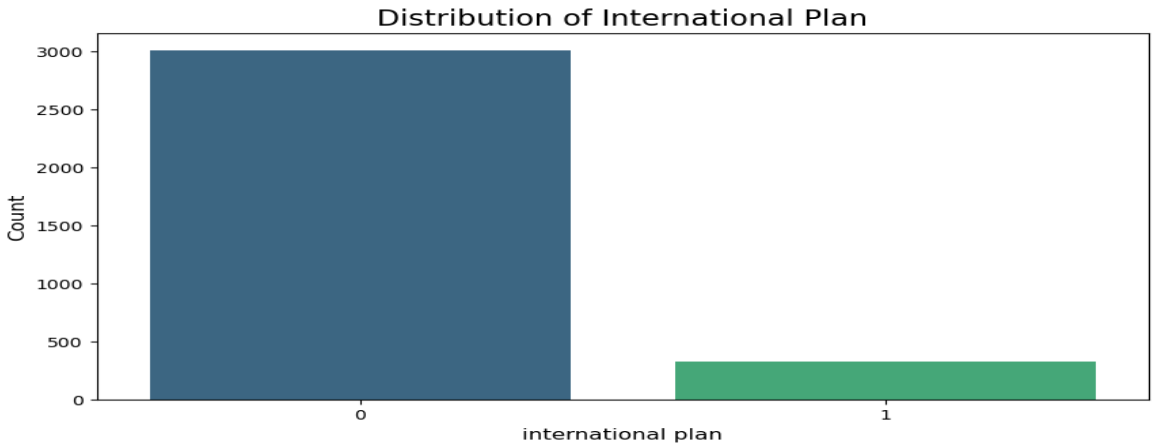
Categorical Features



Observation

- Distribution is relatively uniform across all states, with slight variations in customer counts.

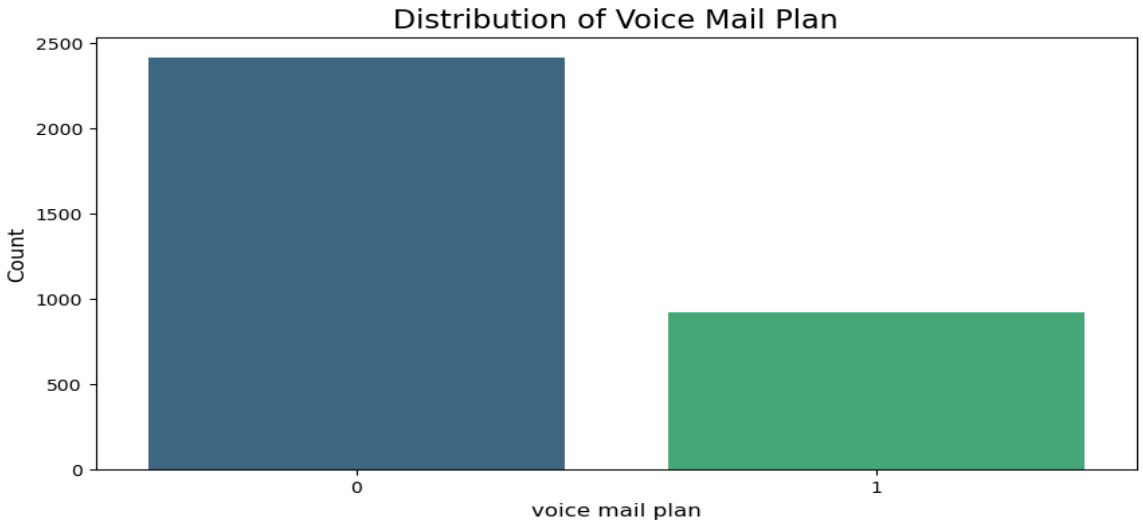
- States like WV and IA have higher customer counts compared to others.



Observation

- Most customers do not subscribe to the international plan (~90%).

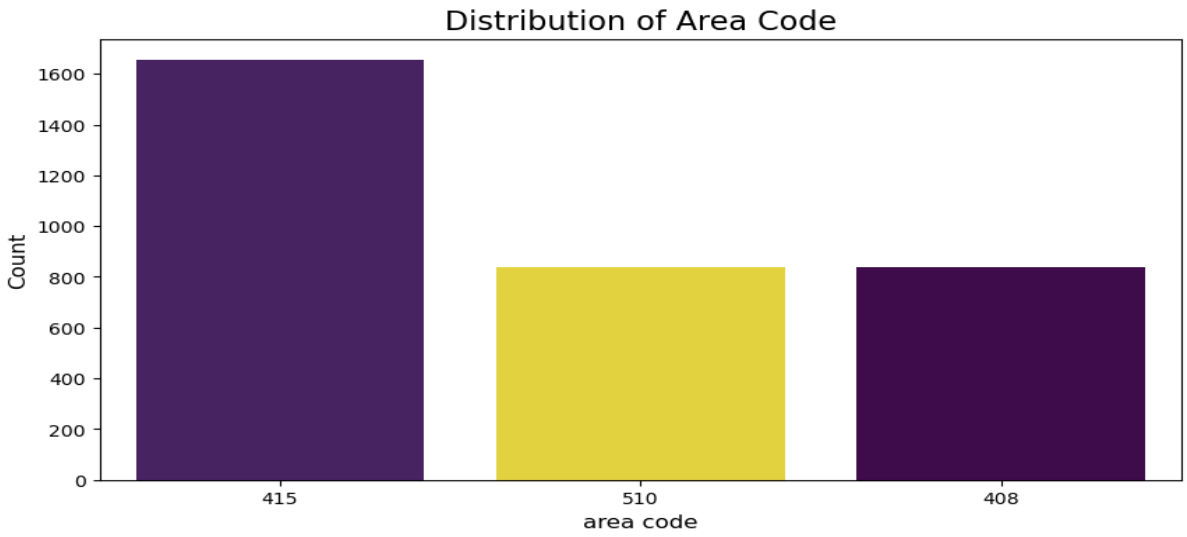
- Non-subscribers dominate, indicating potential low relevance unless linked to churn.



Observation

- A majority of customers (~65%) do not have a voicemail plan.

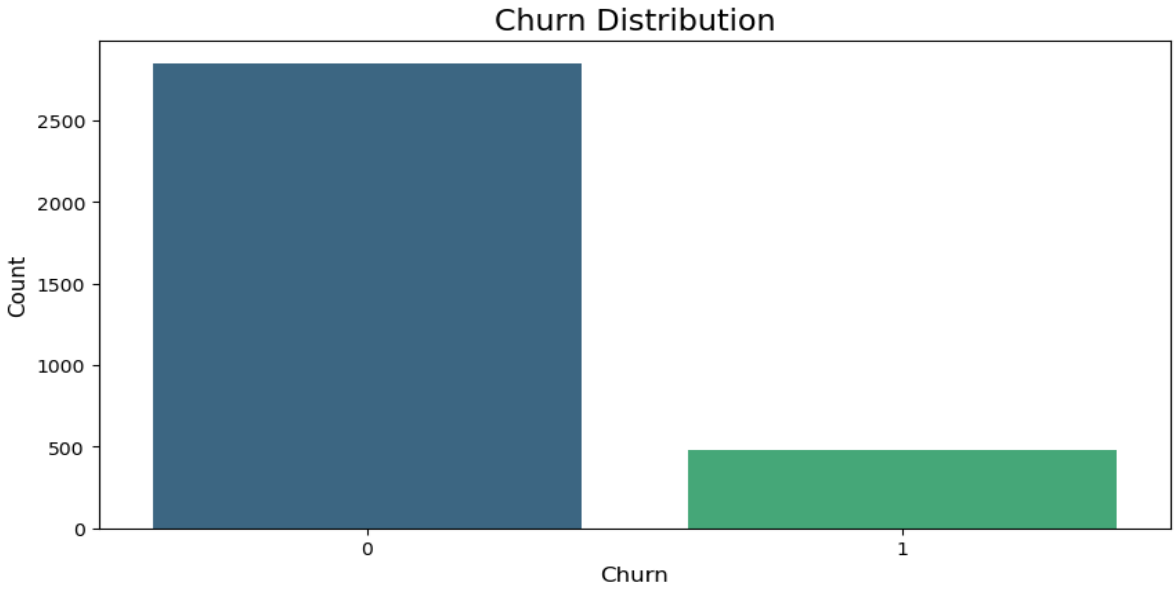
- This feature could be explored further for its impact on churn.



Observation

- Customers are grouped into three distinct area codes (415, 408, 510).

- Area code 415 has the highest customer count, but differences are minor.



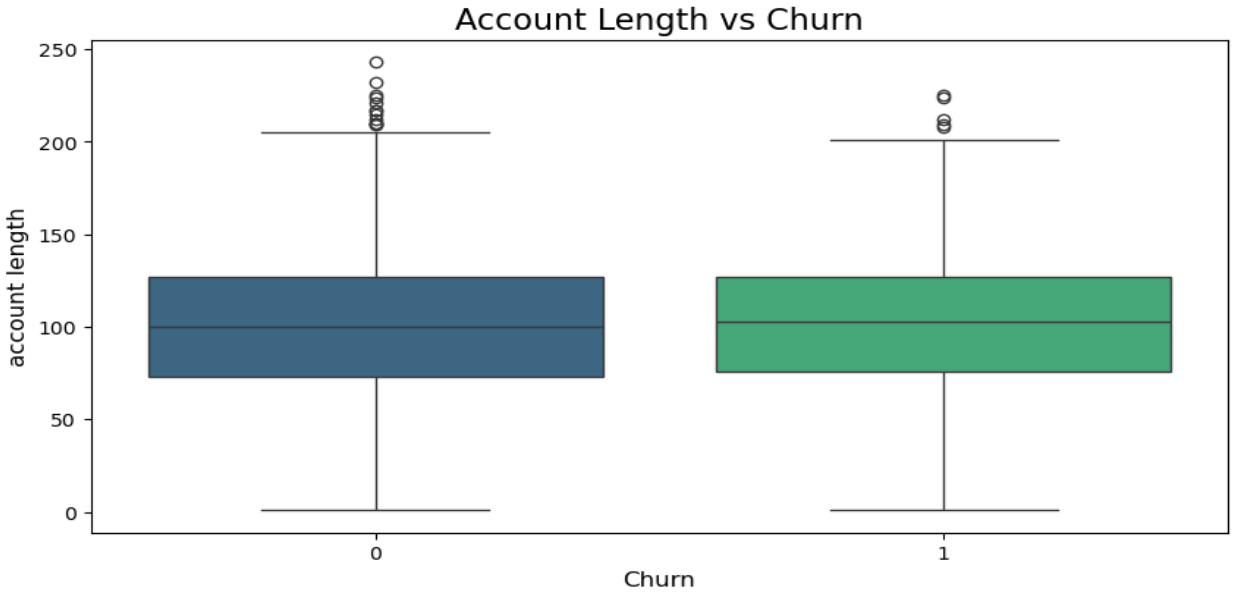
Observations

483 customers have churned, which represents 14.49% of the total 3333 customers, for which it confirms class imbalance.

- **Bivariate Analysis:**

  Relationships between Churn and other features.

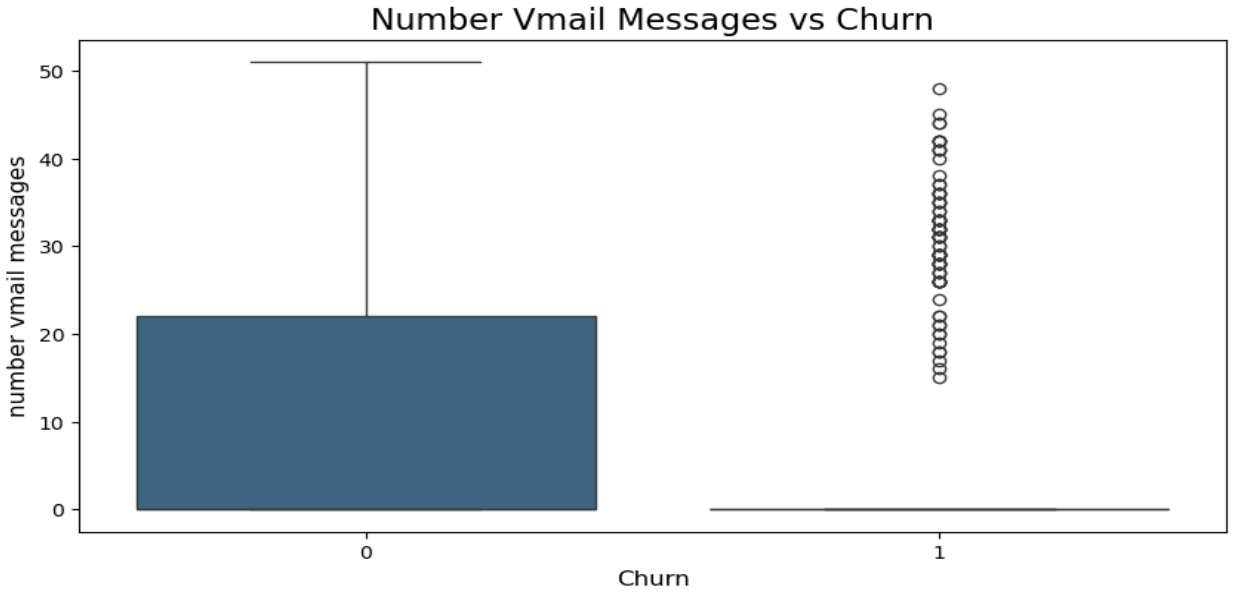
a) Comparing continuous variables with churn



Observations

The distribution of account length appears to differ between churned and non-churned customers.

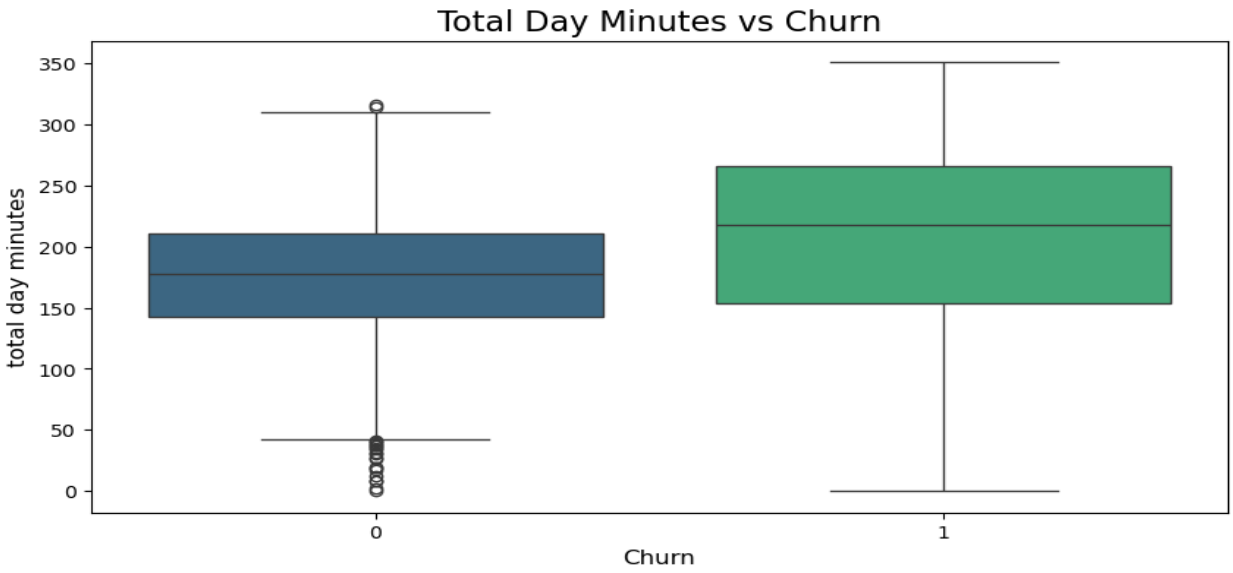
- Churned customers tend to have higher values of account length.



Observations

The distribution of number vmail messages appears to differ between churned and non-churned customers.

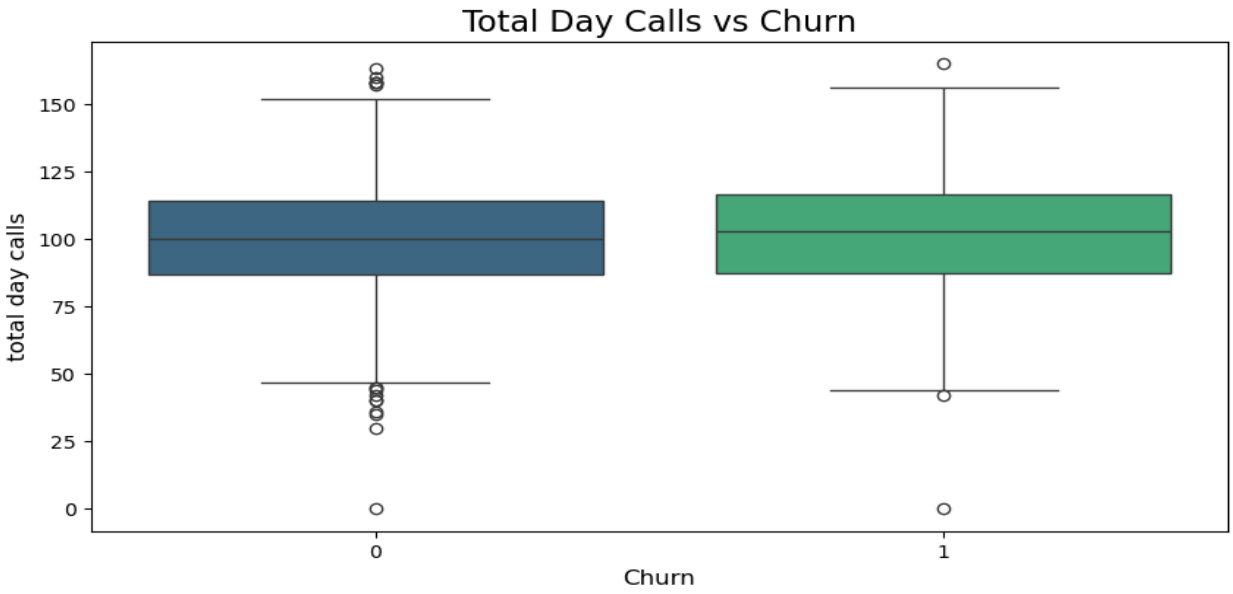
- Non-churned customers tend to have higher values of number vmail messages.



Observations

The distribution of total day minutes appears to differ between churned and non-churned customers.

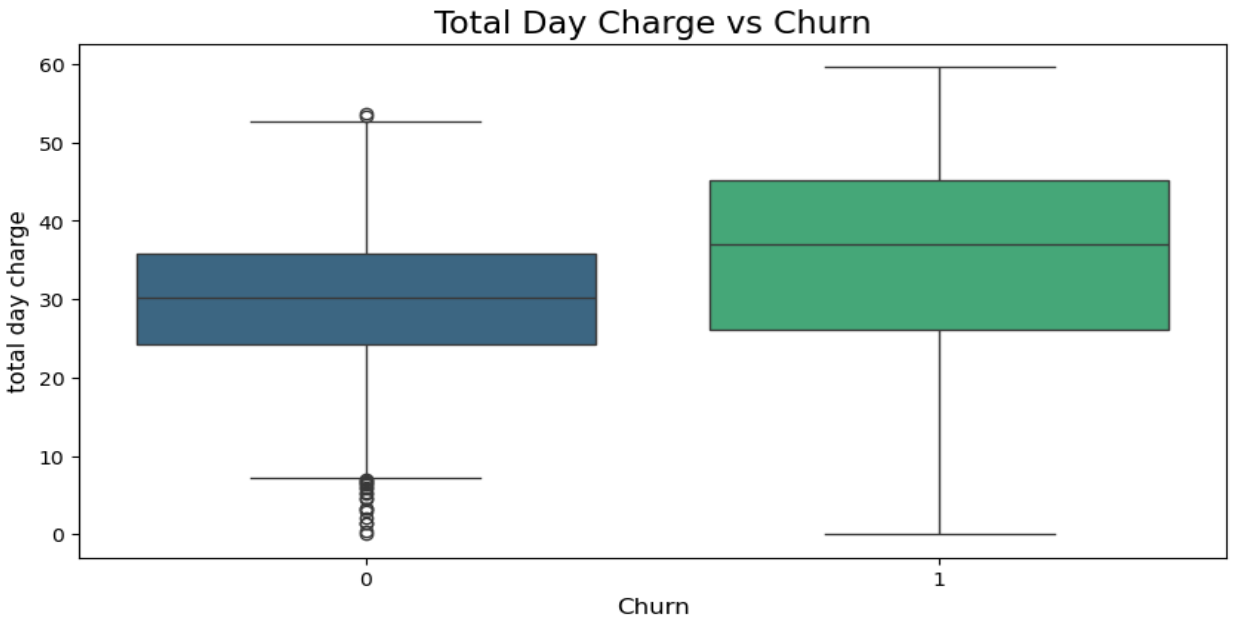
- Churned customers tend to have higher values of total day minutes.



Observations

The distribution of total day calls appears to differ between churned and non-churned customers.

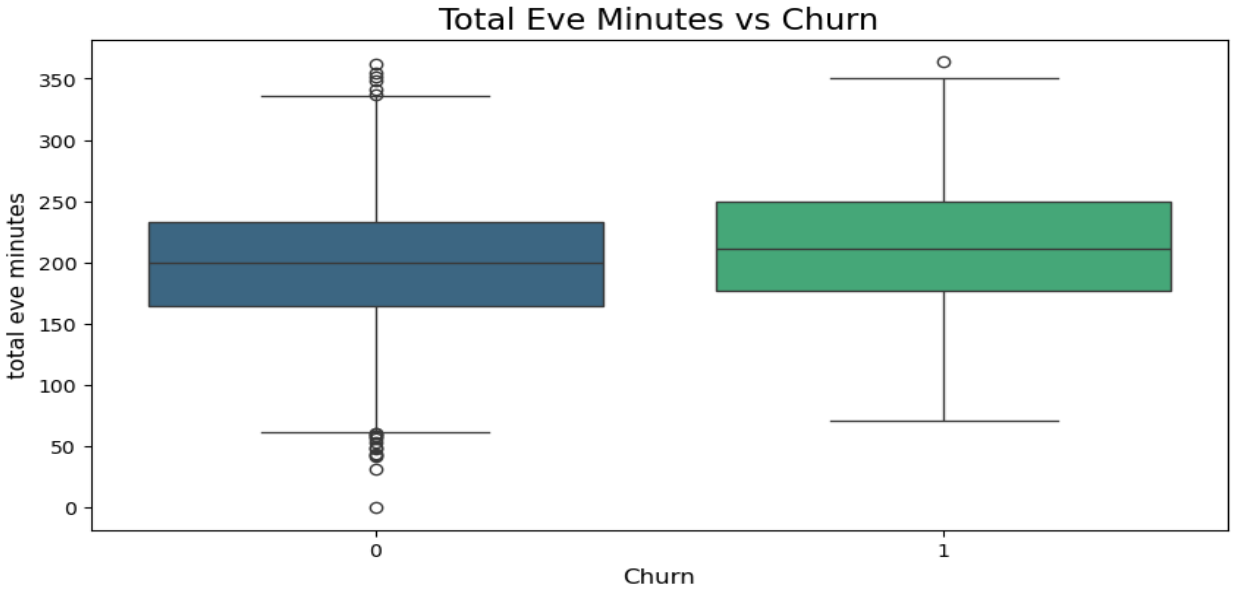
- Churned customers tend to have higher values of total day calls.



Observations

The distribution of total day charge appears to differ between churned and non-churned customers.

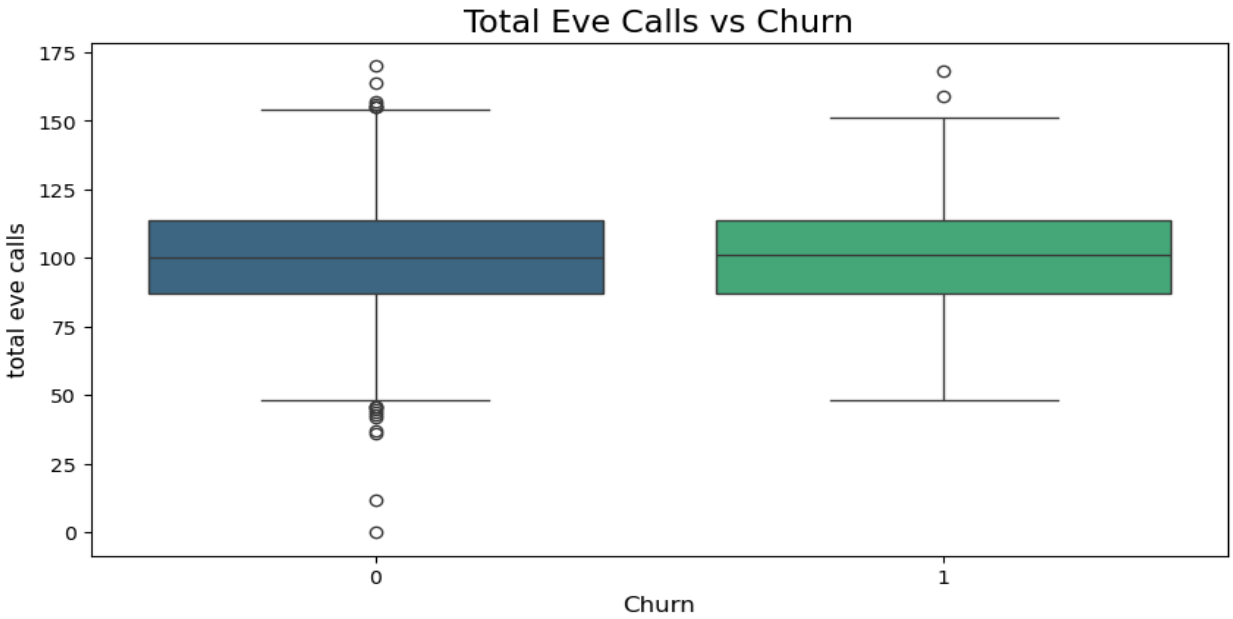
- Churned customers tend to have higher values of total day charge.



Observations

The distribution of total eve minutes appears to differ between churned and non-churned customers.

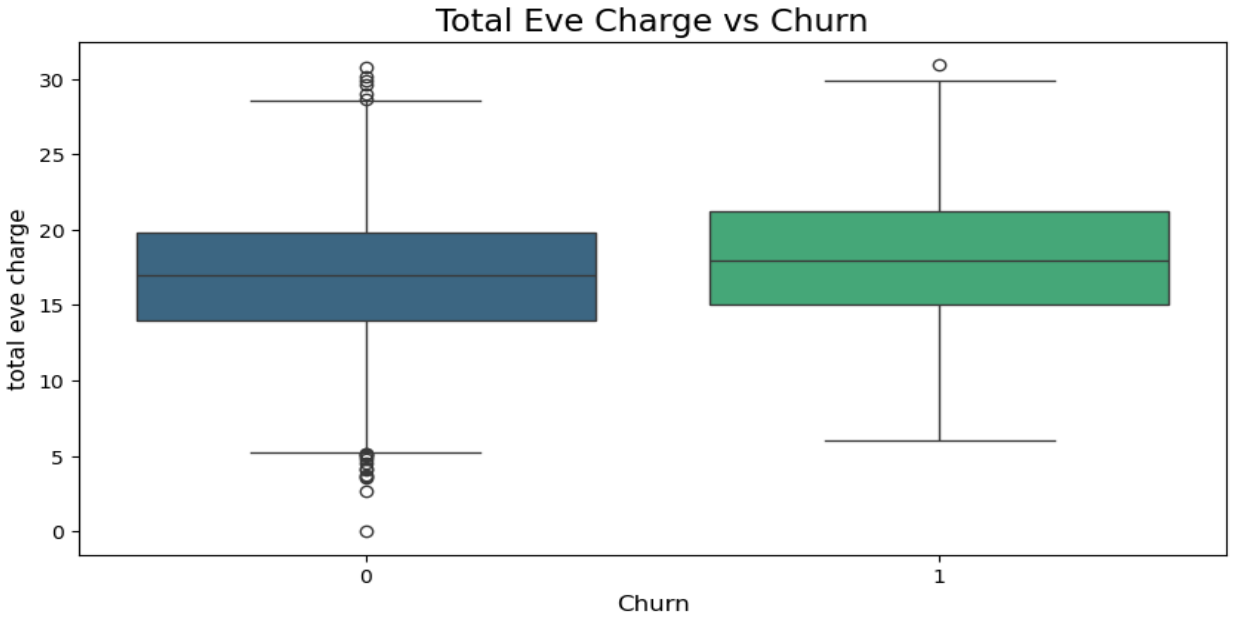
- Churned customers tend to have higher values of total eve minutes.



Observations

The distribution of total eve calls appears to differ between churned and non-churned customers.

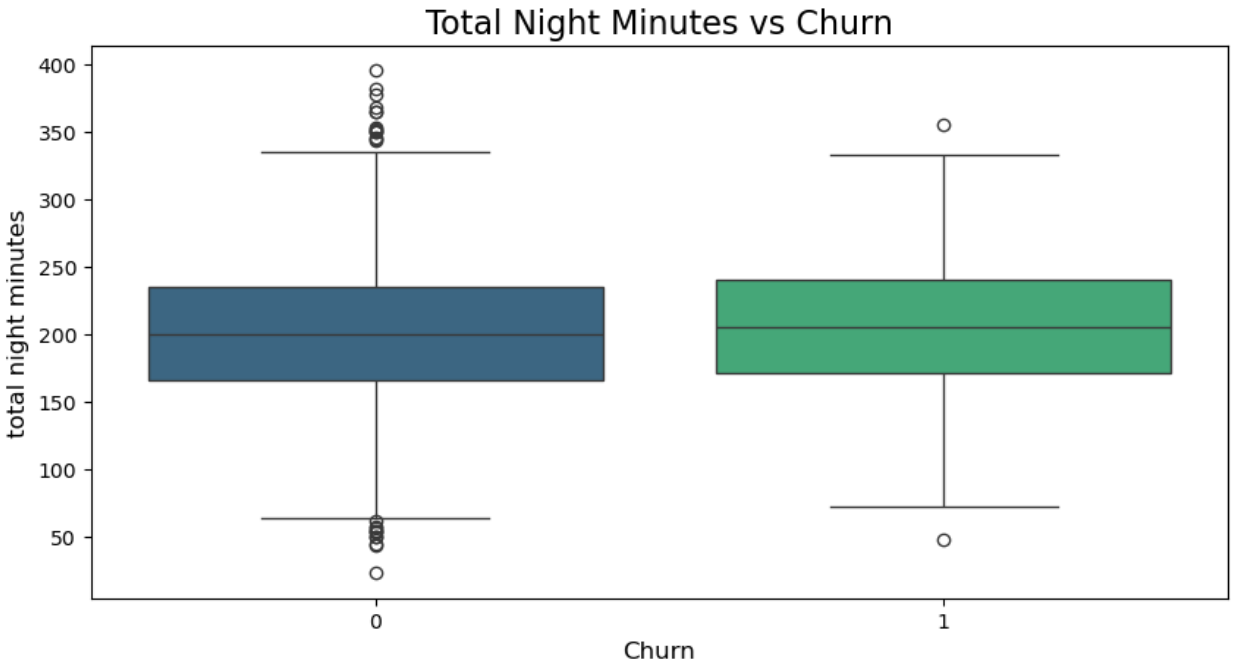
- Churned customers tend to have higher values of total eve calls.



Observations

The distribution of total eve charge appears to differ between churned and non-churned customers.

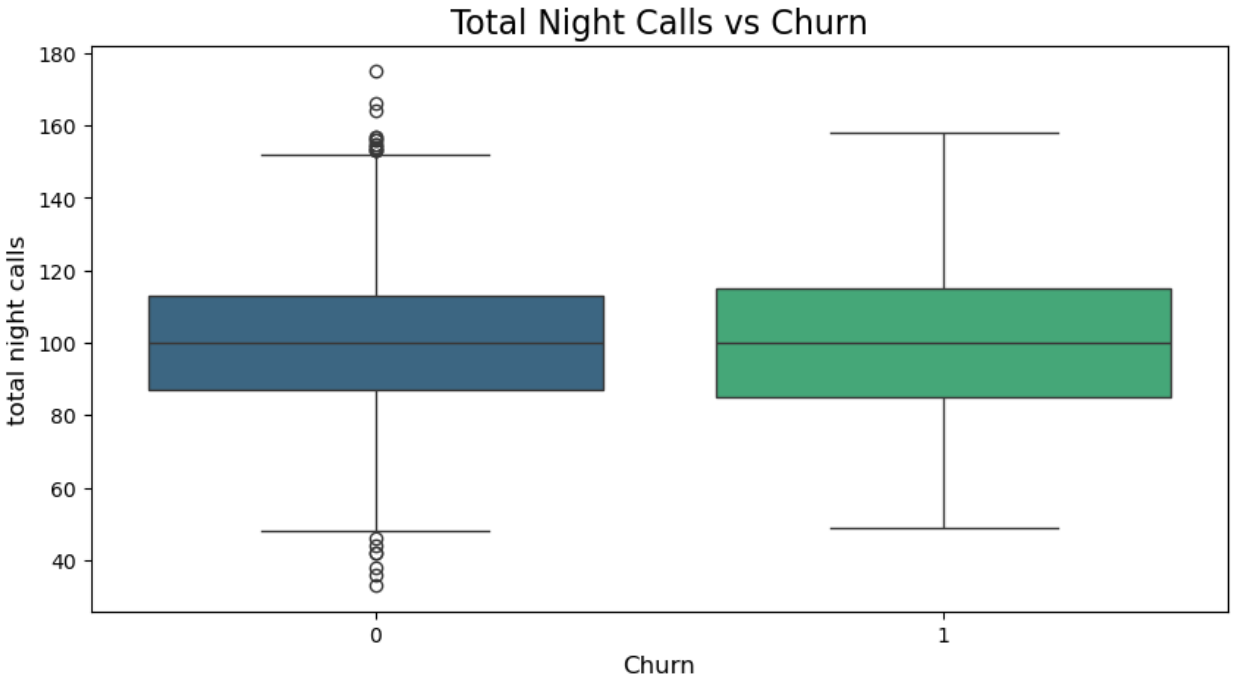
- Churned customers tend to have higher values of total eve charge.



Observations

The distribution of total night minutes appears to differ btwn churned and non-churned customers.

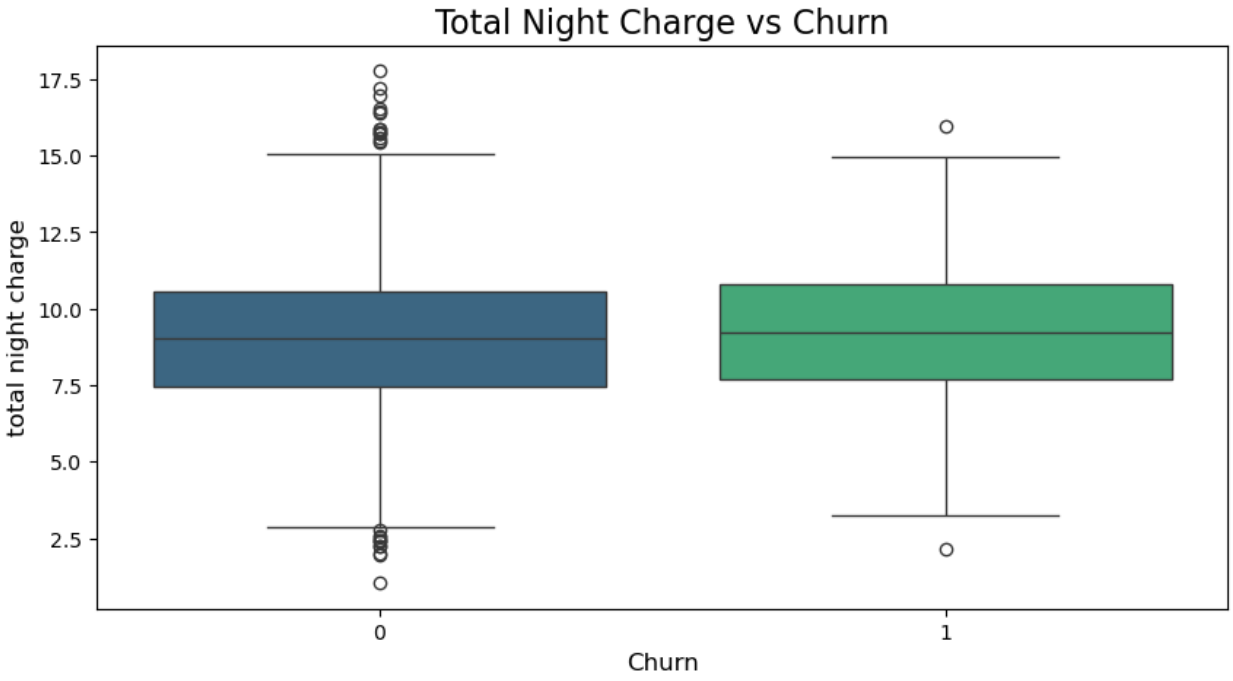
- Churned customers tend to have higher values of total night minutes.



Observations

The distribution of total night calls appears to differ between churned and non-churned customers.

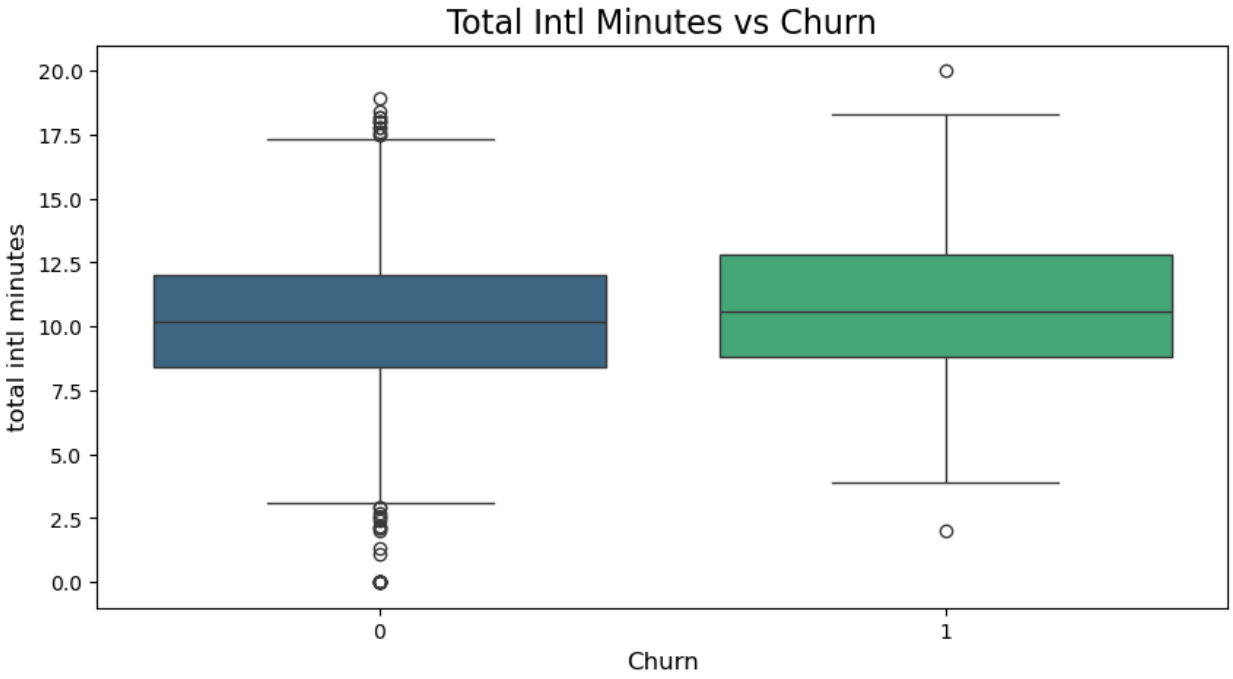
- Churned customers tend to have higher values of total night calls.



Observations

The distribution of total night charge appears to differ between churned and non-churned customers.

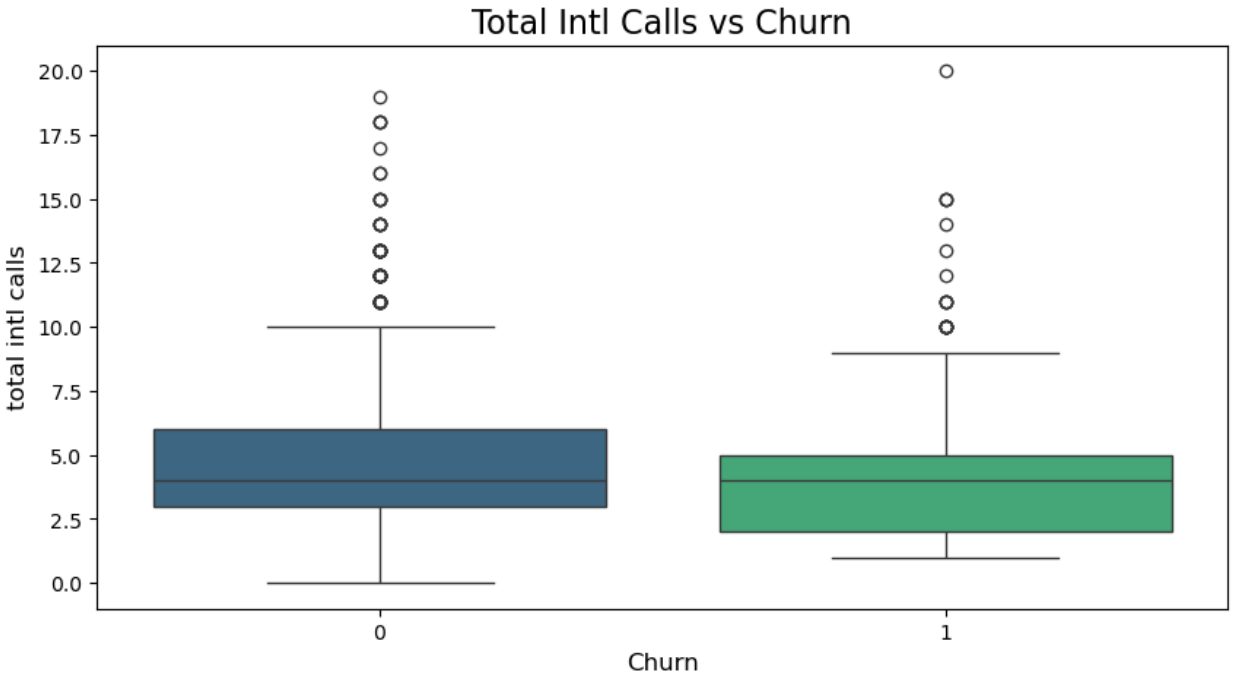
- Churned customers tend to have higher values of total night charge.



Observations

The distribution of total intl minutes appears to differ between churned and non-churned customers.

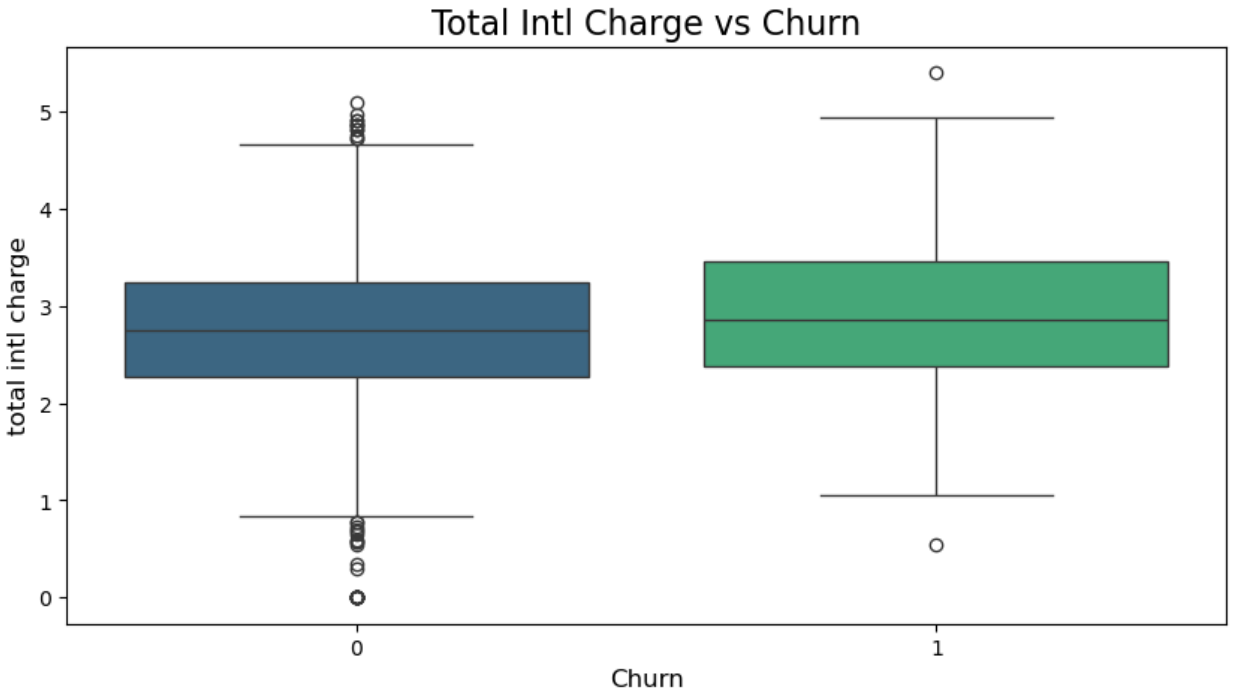
- Churned customers tend to have higher values of total intl minutes.



Observations

The distribution of total intl calls appears to differ between churned and non-churned customers.

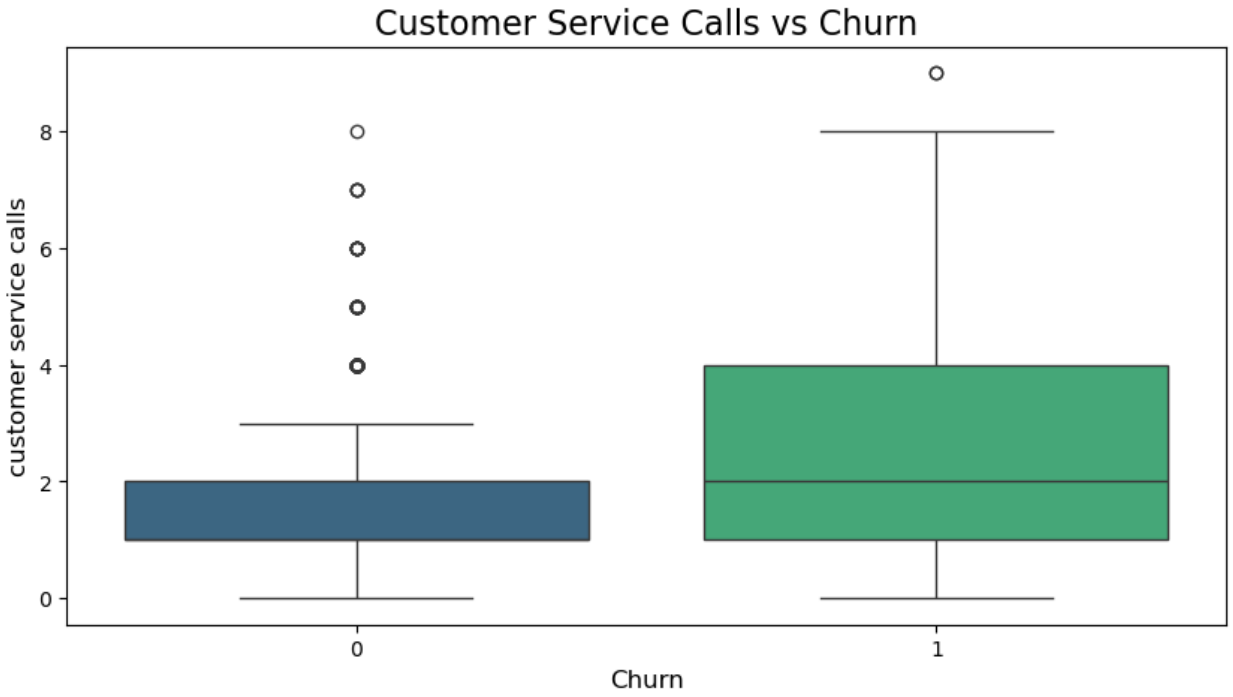
- Non-churned customers tend to have higher values of total intl calls.



Observations

The distribution of total intl charge appears to differ between churned and non-churned customers.

- Churned customers tend to have higher values of total intl charge.

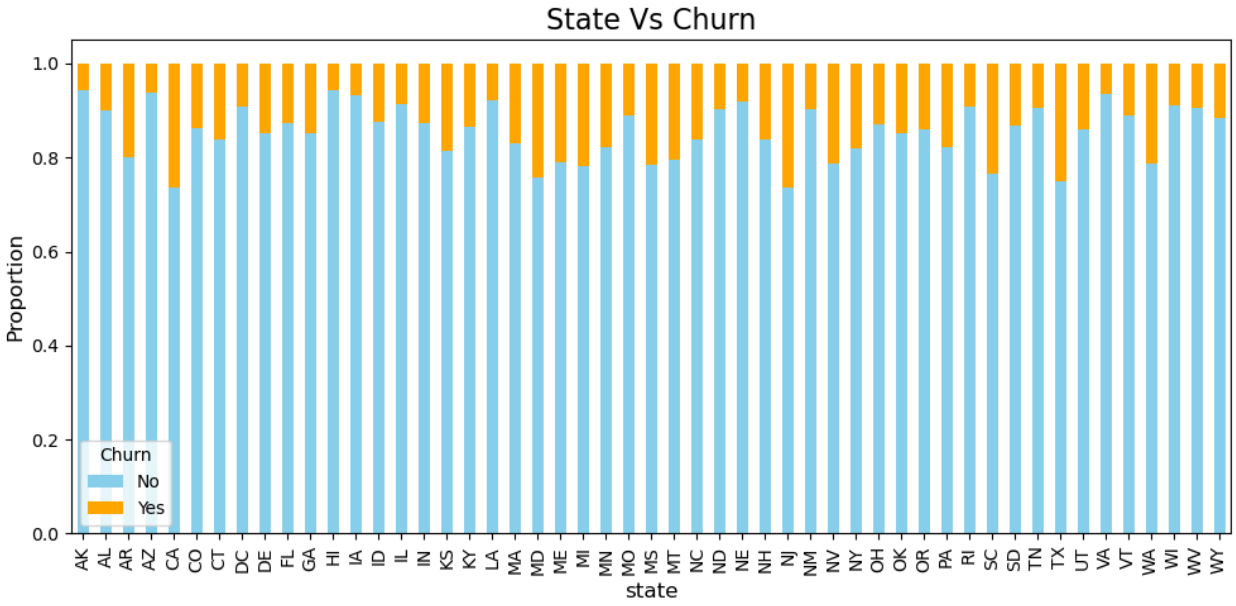


Observations

The distribution of customer service calls appears to differ between churned and non-churned custo mers.

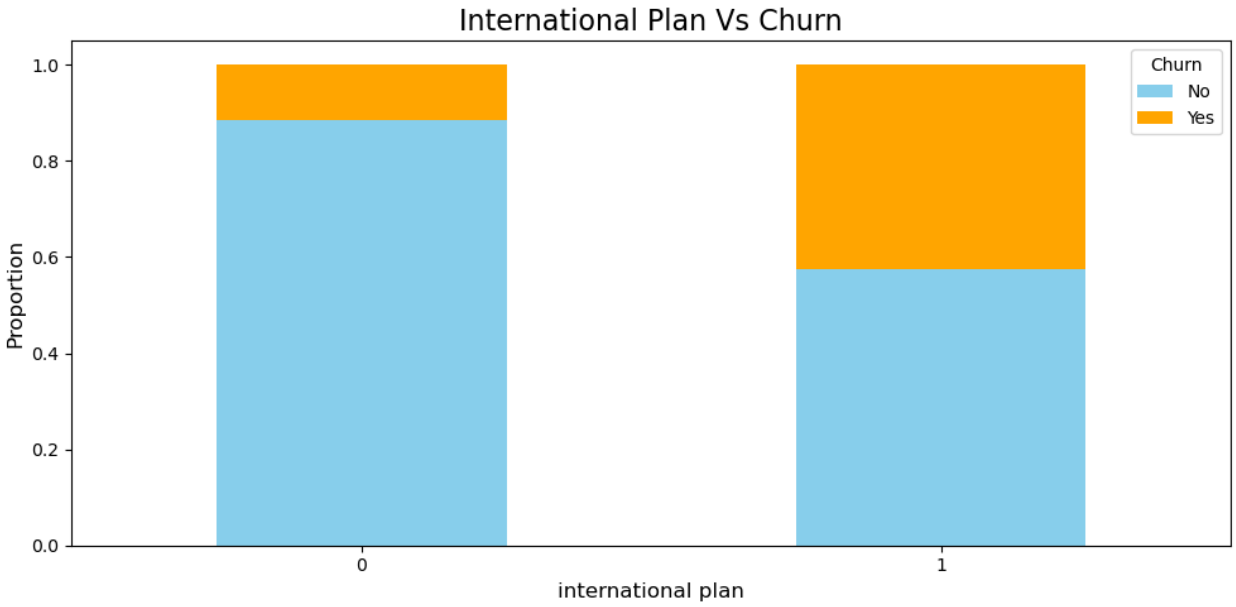
- Churned customers tend to have higher values of customer service calls.

b) Categorical Features by Churn



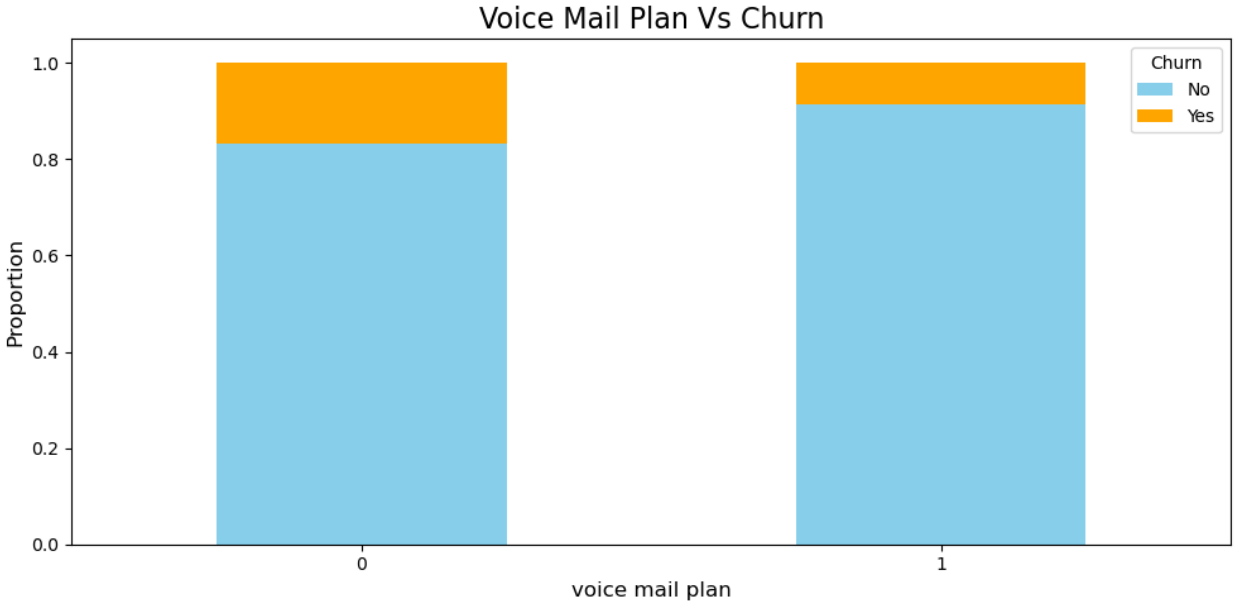
Observations

- The distribution of churn varies across state.
- Significant variation in churn rates is observed among the state categories.



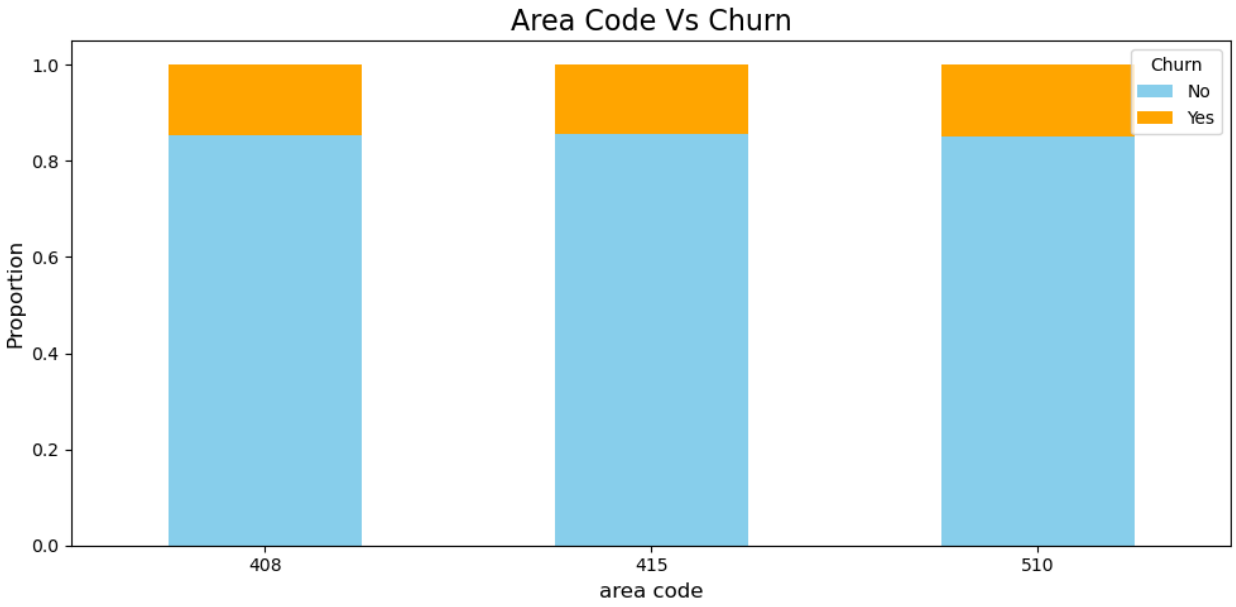
Observations

- The distribution of churn varies across international plan.
- Significant variation in churn rates is observed among the international plan categories.



Observations

- The distribution of churn varies across voice mail plan.
- Minimal variation in churn rates is observed among the voice mail plan categories.



Observations

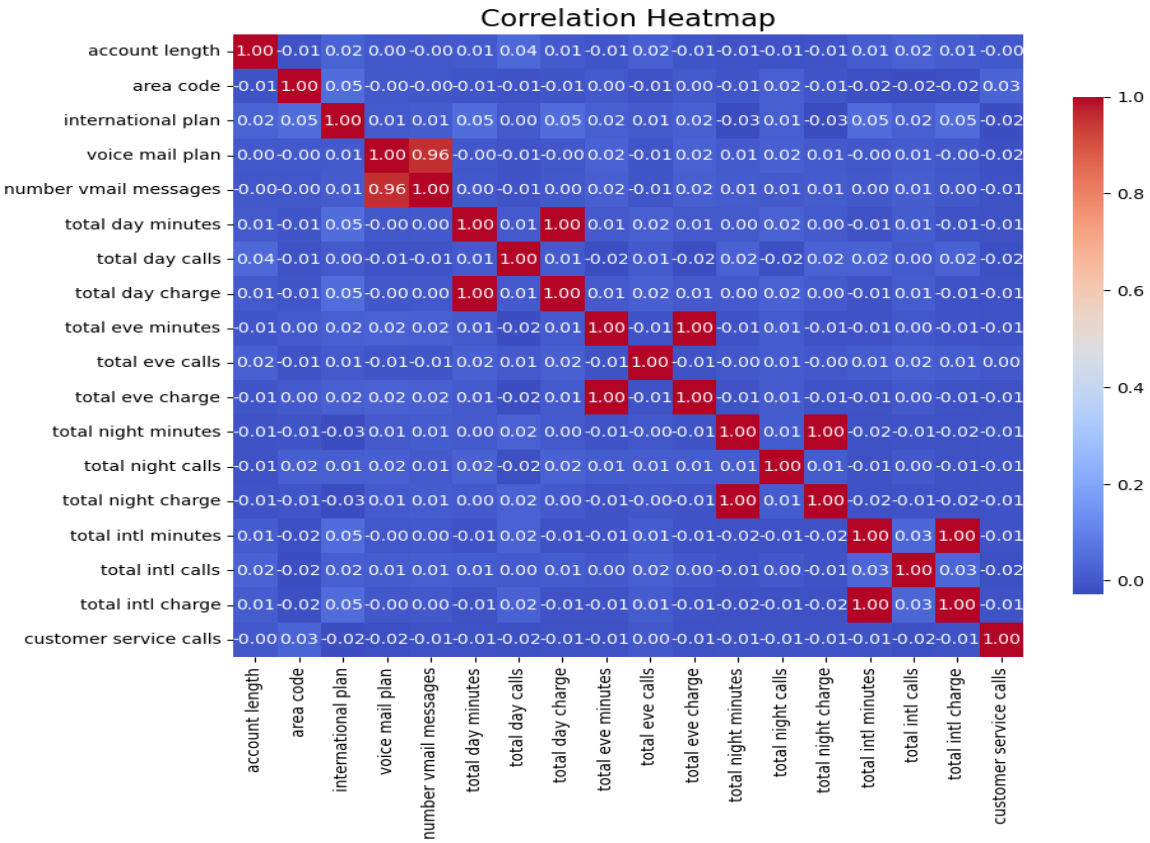
- The distribution of churn varies across area code.
- Minimal variation in churn rates is observed among the area code categories.

Multivariate Analysis:

Correlation matrix and interaction effects between key features and churn.

a) Correlation heatmap

- Visualize a correlation heatmap, which will help in detecting relationships among numerical features.



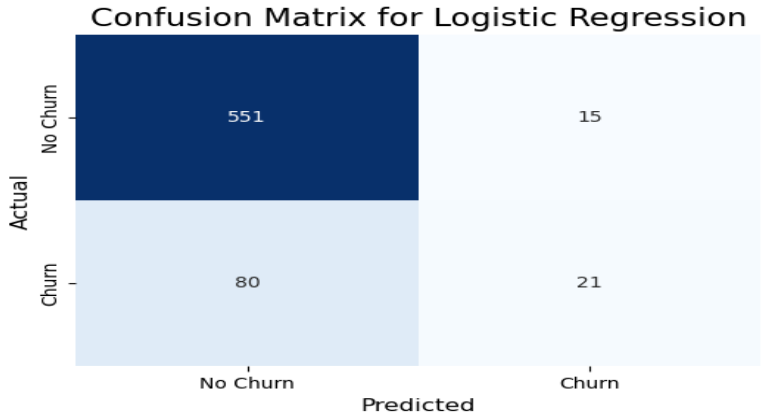
Observations

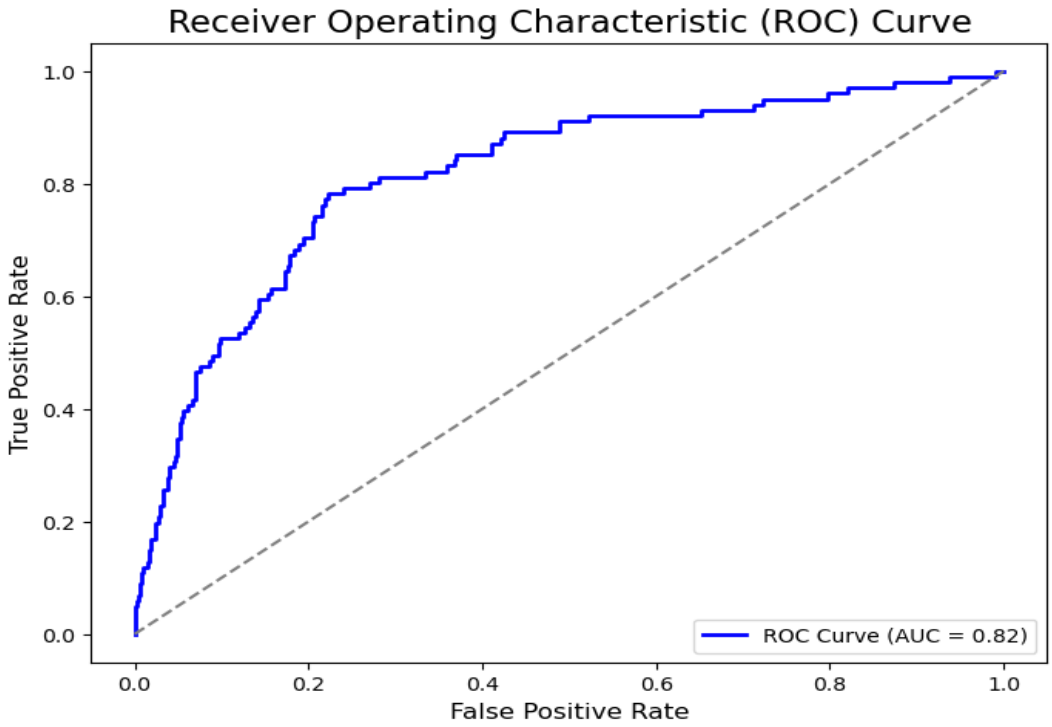
The highest correlations are observed between [('total day minutes', 'total day charge'), ('total day charge', 'total day minutes'), ('total eve minutes', 'total eve charge')].

MODELING

Models Used:

1. **Baseline Model:** Logistic Regression (interpretable and simple).



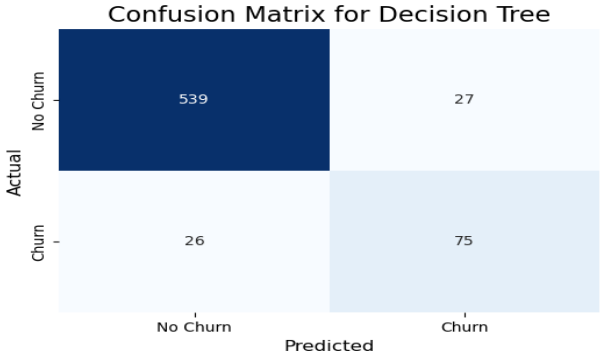


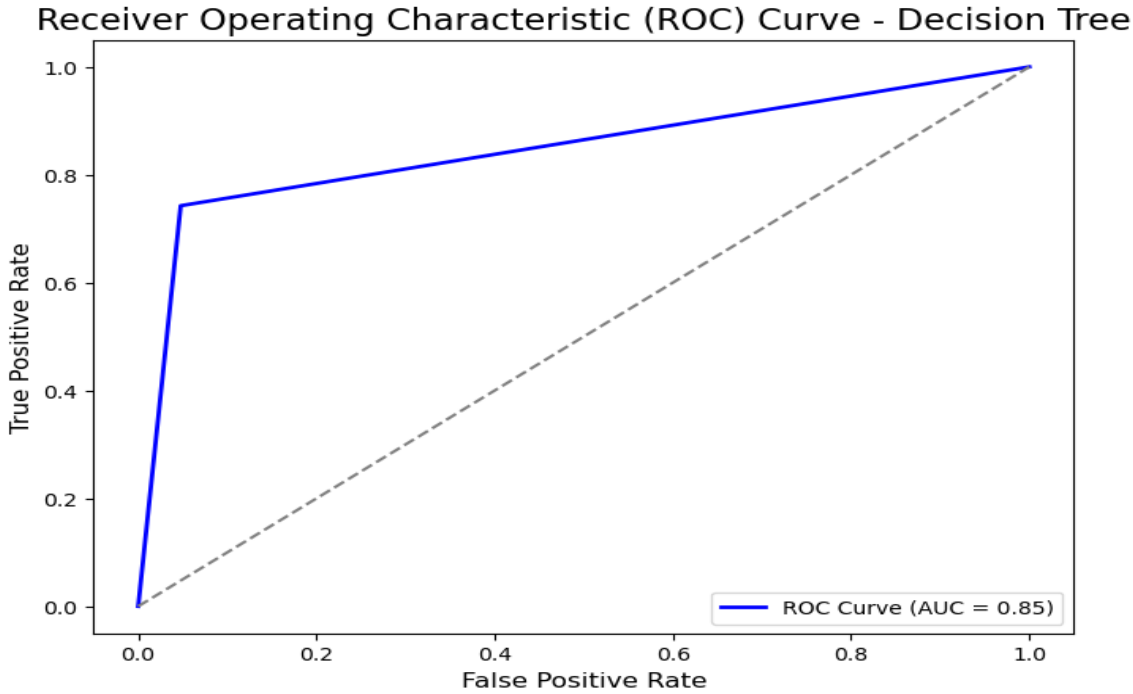
Observations and Findings

1. Logistic Regression Accuracy: 0.86
   - The accuracy of the model is 85.76%, indicating how well the model is classifying both churned and non-churned customers.
2. Logistic Regression Precision: 0.58
   - The precision of the model is 58.33%, which means that of all the customers predicted to churn, 58.3 3% actually did churn.
3. Logistic Regression Recall: 0.21
   - The recall of the model is 20.79%, indicating how well the model identifies actual churned customer.

4. Logistic Regression F1-Score: 0.31
- The F1-score is 0.31, which is the harmonic mean of precision and recall.
5. Confusion Matrix: The confusion matrix reveals the following:
- True Positives (TP): 21 - Churned customers correctly predicted.
- True Negatives (TN): 551 - Non-churned customers correctly predicted.
- False Positives (FP): 15 - Non-churned customers wrongly predicted as churned.
- False Negatives (FN): 80 - Churned customers wrongly predicted as non-churned.
6. ROC Curve: The ROC curve shows the trade-off between the true positive rate and the false positive r ate
- The AUC (Area Under the Curve) score is 0.82, indicating the model's ability to distinguish betwen the classes.
- An AUC score close to 1.0 indicates excellent model performance, while a score near 0.5 suggests r andom guessing.

2. **Second Model:** Decision Tree Classifier (non-linear relationships).





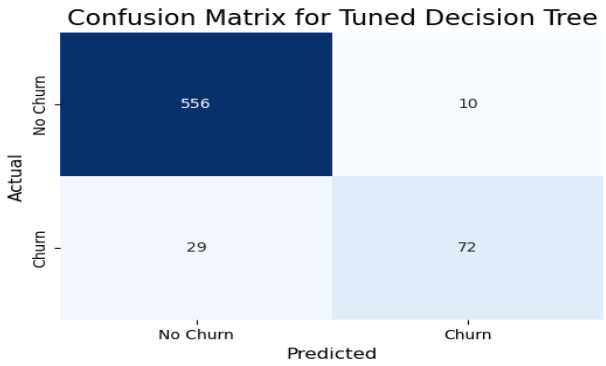
Observations and Findings

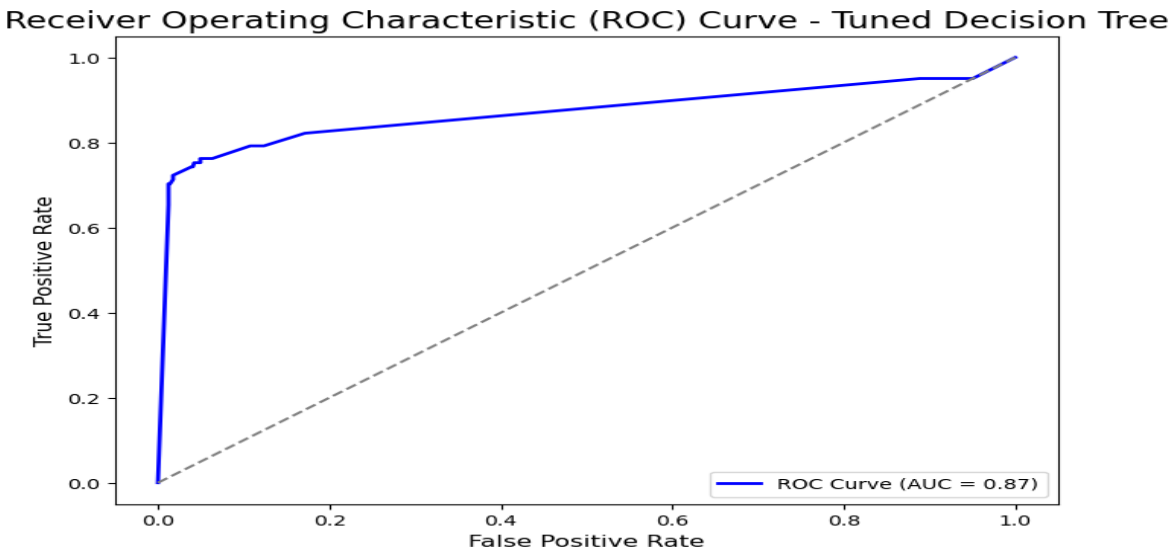
1. Decision Tree Accuracy: 0.92

- The accuracy of the model is 92.05%, indicating how well the model is classifying both churned and non-churned customers.

2. Decision Tree Precision: 0.74
   - The precision of the model is 73.53%, which means that of all the customers predicted to churn, 73.53% actually did churn.
3. Decision Tree Recall: 0.74
   - The recall of the model is 74.26%, which means that 74.26% of actual churned customers were correctly identified by the model.
4. Decision Tree F1 Score: 0.74
   - The F1 score is 73.89%, indicating the balance between precision and recall.
5. Confusion Matrix: The confusion matrix reveals the following:
   - True Positives (TP): 75 - Churned customers correctly predicted.
   - True Negatives (TN): 539 - Non-churned customers correctly predicted.
   - False Positives (FP): 27 - Non-churned customers wrongly predicted as churned.
   - False Negatives (FN): 26 - Churned customers wrongly predicted as non-churned.
6. ROC Curve: The ROC curve shows the trade-off between the true positive rate and the false positive rate.
   - The AUC (Area Under the Curve) score is 0.85, indicating the model's ability to distinguish between the classes.
   - An AUC score close to 1.0 indicates excellent model performance, while a score near 0.5 suggests random guessing.

3. **Tuned Model:** Hyperparameter-tuned Decision Tree using GridSearchCV.





Observations and Findings

1. Tuned Decision Tree Accuracy: 0.94
   - The accuracy of the tuned model is 94.15%, showing improvement from the baseline.
2. Tuned Decision Tree Precision: 0.88
   - The precision of the tuned model is 87.80%, indicating better identification of churned customers com pared to the baseline.
3. Tuned Decision Tree Recall: 0.71
   - The recall of the tuned model is 71.29%, which means that 71.29% of actual churned customers were correctly identified.
4. Tuned Decision Tree F1 Score: 0.79
   - The F1 score is 78.69%, indicating a better balance between precision and recall.
5. Confusion Matrix: The confusion matrix reveals the following:
   - True Positives (TP): 72 - Churned customers correctly predicted.
   - True Negatives (TN): 556 - Non-churned customers correctly predicted.
   - False Positives (FP): 10 - Non-churned customers wrongly predicted as churned.
   - False Negatives (FN): 29 - Churned customers wrongly predicted as non-churned.

6. ROC Curve: The ROC curve shows the trade-off between the true positive rate and the false positive rat e.
   - The AUC (Area Under the Curve) score is 0.87, indicating the model's improved ability to distinguish b etween churned and non-churned customers.
   - A higher AUC value indicates that the model is better at making distinctions.

These models are used because we are dealing with a classification problem.

Metrics Used:

Evaluation based on accuracy, precision, recall, F1-score, and AUC (Area Under the Curve).

EVALUATION

Model Performance:

1. **Logistic Regression:**
   - Accuracy: 86%
   - Precision: 58%
   - Recall: 21%
   - F1 Score:31%
   - AUC: 82%

2. **Decision Tree:**
    - Accuracy: 92%
    - Precision: 74%
    - Recall: 74%
    - F1 Score:74%
    - AUC: 85%
3. **Tuned Decision Tree:**
    - Accuracy: 94%
    - Precision: 88%
    - Recall: 71%
    - F1 Score: 79%
    - AUC: 87%

Best Model:

The Tuned Decision Tree model performs best due to its high precision, AUC, and interpretability.

CONCLUSION

1. The **Tuned Decision Tree** is the most suitable model for this project, as it met and exceeded the success criteria (accuracy ≥ 85%, precision ≥ 80%). This model is robust and can effectively predict customer churn while minimizing false positives.
2. The **Basic Decision Tree** performed well in accuracy but fell short on precision, showing that hyperparameter tuning plays a crucial role in improving model performance.
3. The **Logistic Regression** model is unsuitable for deployment due to its significantly lower precision, meaning it might misclassify too many customers as churned, leading to inefficient allocation of retention efforts.
4. Insights highlight high churn risks among customers with month-to-month contracts.

RECOMMENDATIONS

1. Develop targeted retention campaigns for high-risk segments (e.g., short-tenure customers).
2. Offer incentives for customers to shift to longer contracts.
3. Monitor churn risk factors continuously and update models with new data.

NEXT STEPS

1. **Deployment:**
   Integrate the model into production systems for real-time predictions.
2. **Data Collection:**
   - Add features capturing customer satisfaction and interactions.
   - Gather longitudinal data for improved trend analysis.
3. **Enhancements:**
   - Regularly refine models to adapt to changing patterns.
   - Incorporate feedback loops for continuous improvement.