PREDICTION BY:

UCHE SAMUEL

CONNECTTEL TELECOMMUNICATION

CUSTOMER CHURN

Predictive Analysis using Supervised Machine Learning

TABLE OF CONTENT

The project involves developing a customer churn prediction system for a leading telecommunications company, ConnectTel. The company faces the pressing need to address customer churn, which poses a significant threat to its business sustainability and growth. The goal is to leverage advanced analytics and machine learning techniques on available customer data to accurately forecast customer churn and implement targeted retention initiatives.

ABOUT THE COMPANY

ConnectTel is a top telecom company known for innovation and reliable services worldwide, offering voice, data, and Internet solutions for individuals and businesses.

OBJECTIVES

1 **Develop a Robust Prediction System:**

2 **Identify At-Risk Customers**

3 **Implement Targeted Retention Initiatives**

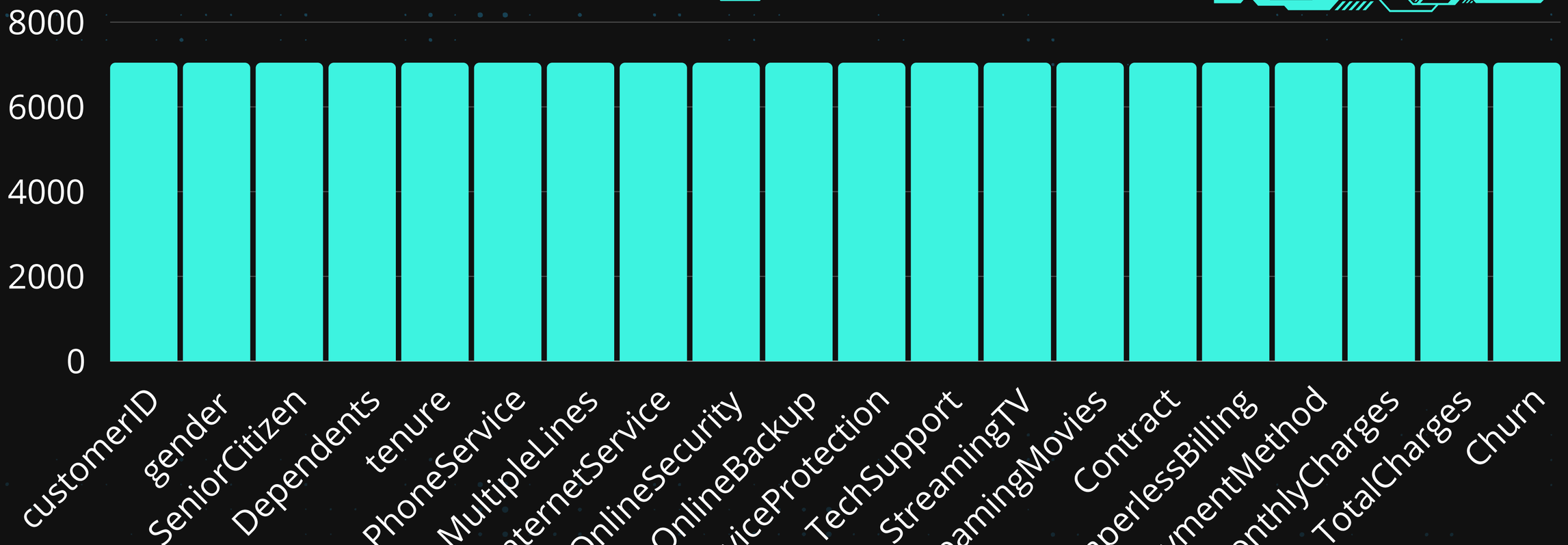
4 **Maintain a Competitive Edge**

DATASET

The dataset comprises various customer attributes such as gender, senior citizen status, partner and dependents status, tenure, types of services subscribed, contract details, billing preferences, and financial information, which are essential for predicting and understanding customer churn behavior in the telecommunications industry.

DATA INFO

11 MISSING VALUES IN TOTAL CHARGES



DATA DESCRIPTION

NUMERICAL OBSERVATION

- SeniorCitizen is denoted in Binary indicator, however, it's a categorical features of if a customer is a senior citizen or not. This will be treated in EDA.
- We can see that the median tenure of a customer is 29 months, maximum is 72 months min is 0 month and mean is 32 Months
- We can also see that the average monthly charge of a customer is $64, the min is $18 and the max is $118.
- The Average total charge is $2283, the min is $18.80, and the max is $8684.80

DATA DESCRIPTION

CATEGORICAL OBSERVATION

- The largest Customer Gender is the Male with 3555 Customers
- Most of the Customers don't have a partner
- Most of the Customers don't have Dependants
- Most of the Customers have a PhoneService
- Most of the Customers don't have multiple lines
- Fibre Optic is the most subscribed internet service
- Most of the Customers don't have an online Security
- Most of the Customers don't have an online backup service
- Most of the Customers don't have device protection
- Most of the Customers don't have technical Support
- Most of the customers don't have a streaming service
- Most of the customers don't have a streaming movie service
- Most of the customers apply to monthly subscriptions
- Most of the Customers have applied for paperless billing
- Most of the customers pay with electronic checks
- Most of the customers have not churned and they are 5174 in number. While over 1500 has churned.

DATA CLEANING/PRE-PROCESSING

- We checked for duplicates and none of the datas were duplicated.
- We did a search of missing values and we noticed that the TotalCharges had 11 values missing.
- Since Machine Learning doesn't work with missing values, we decided to drop those rows.
- Next, we created some new features to improve the performance of our machine learning. And they are, Type of Customer, Total Services/Customer and Service Enagement/Customer

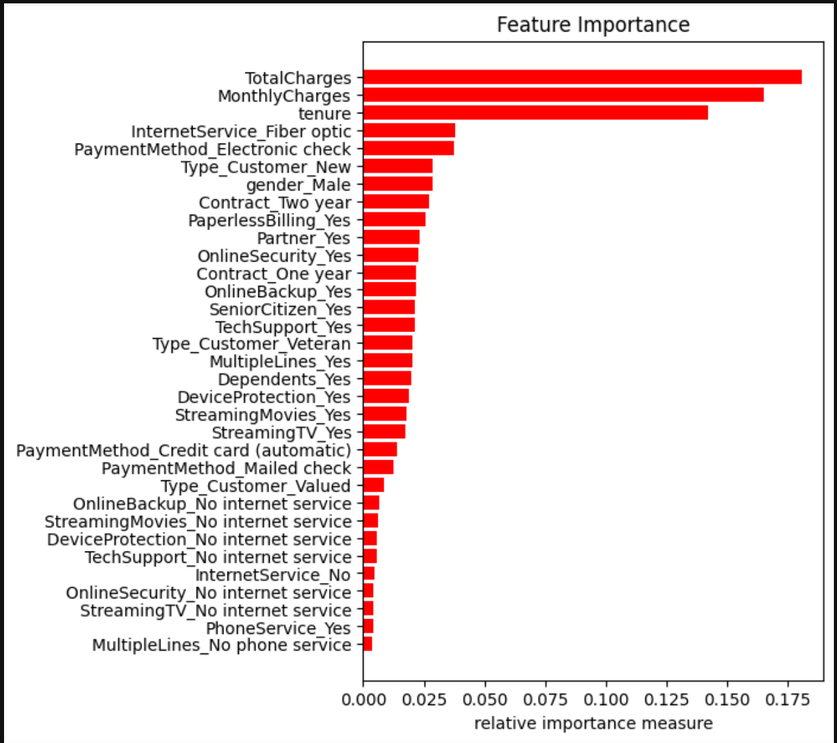
DATA CLEANING/PRE-PROCESSING

- We also identified a redundant feature which is the CustomerID, and it is redundant because it is a unique identification.
- We converted the numerical features (SeniorCitizen) into categorical feature of Yes or No.
- We also converted the categorical feature/Target (Churn) into a binary feature of 0 or 1
- Next we segmented our pre-processed data into Data and Target label.

DATA CLEANING/PRE-PROCESSING

- Next, we scaled the segmented dataset into binary denomination of 1 and 0 using the MinMax Scaler.
- Now that we have our binary data, we need to plot our feature importance Chart and see which of the features would be necessary for our predictive machine learning.
- We used the Random Classifier Model to find these importance features.

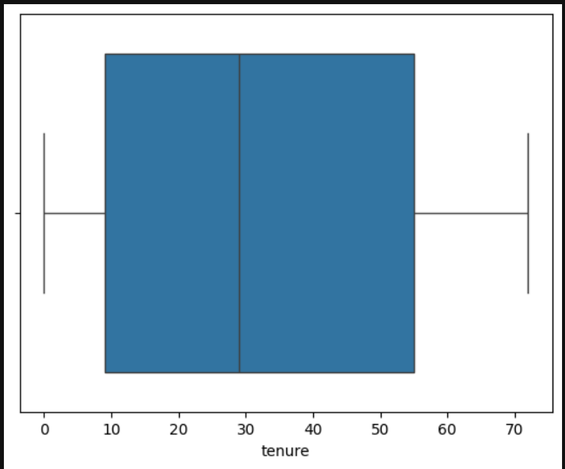
FEATURE IMPORTANCE PLOT

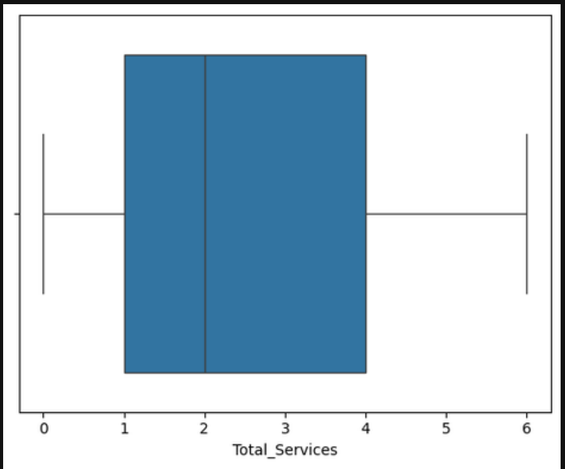


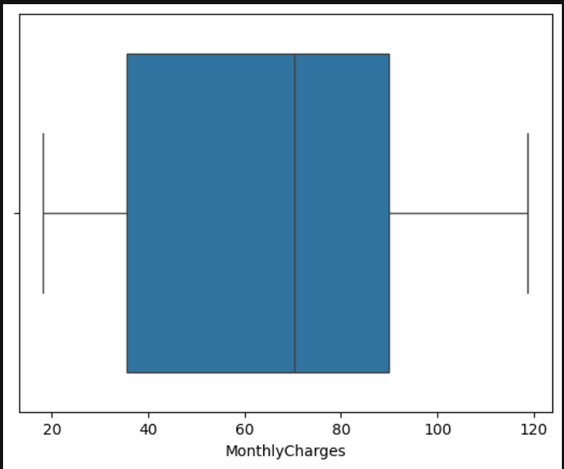
- From the feature importance chart, we identified that our most important features are:
  - Total Charges
  - Monthly Charges
  - Tenure
- These features are important for predicting the customer churn.

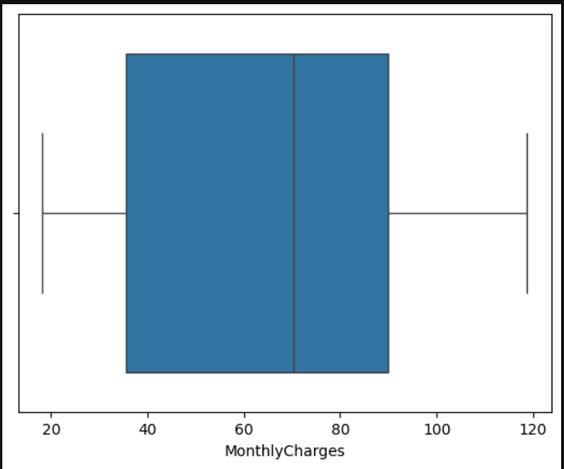
EXPLORATORY DATA ANALYSIS INSIGHTS (EDA)

UNIVARIATE ANALYSIS







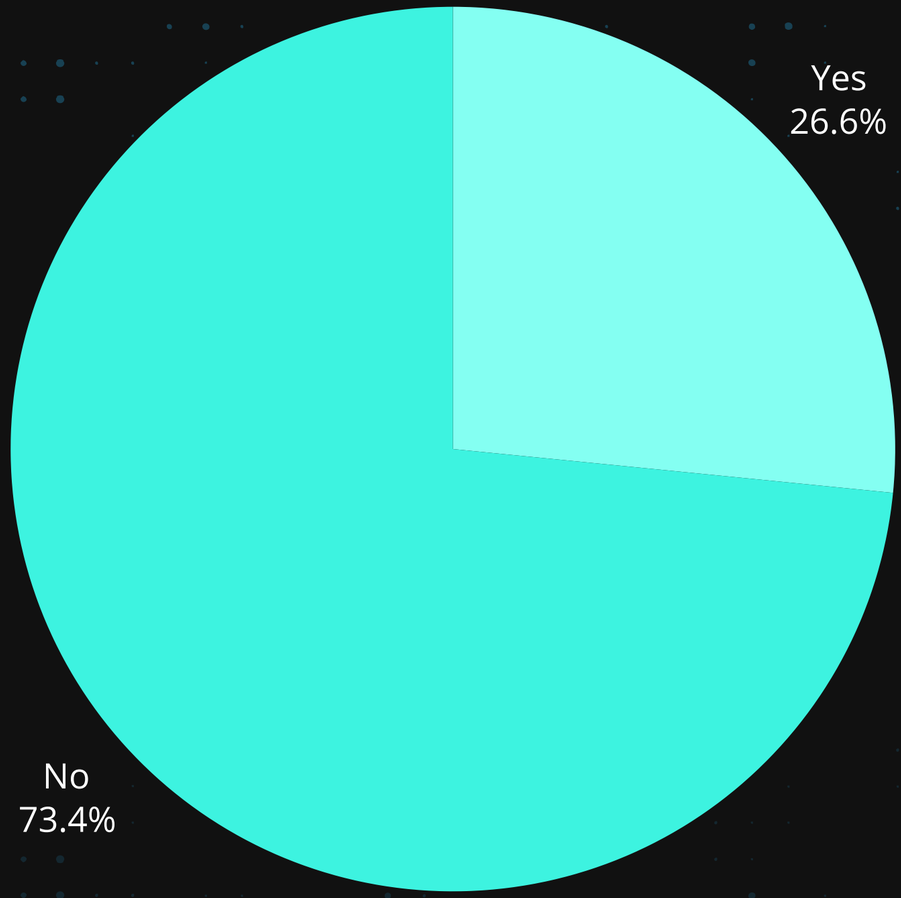


We analysed each numerical feature to identify any potential anomaly or outliers and we found out that there's none using the box plotting charts.

UNIVARIATE ANALYSIS

FEATURE

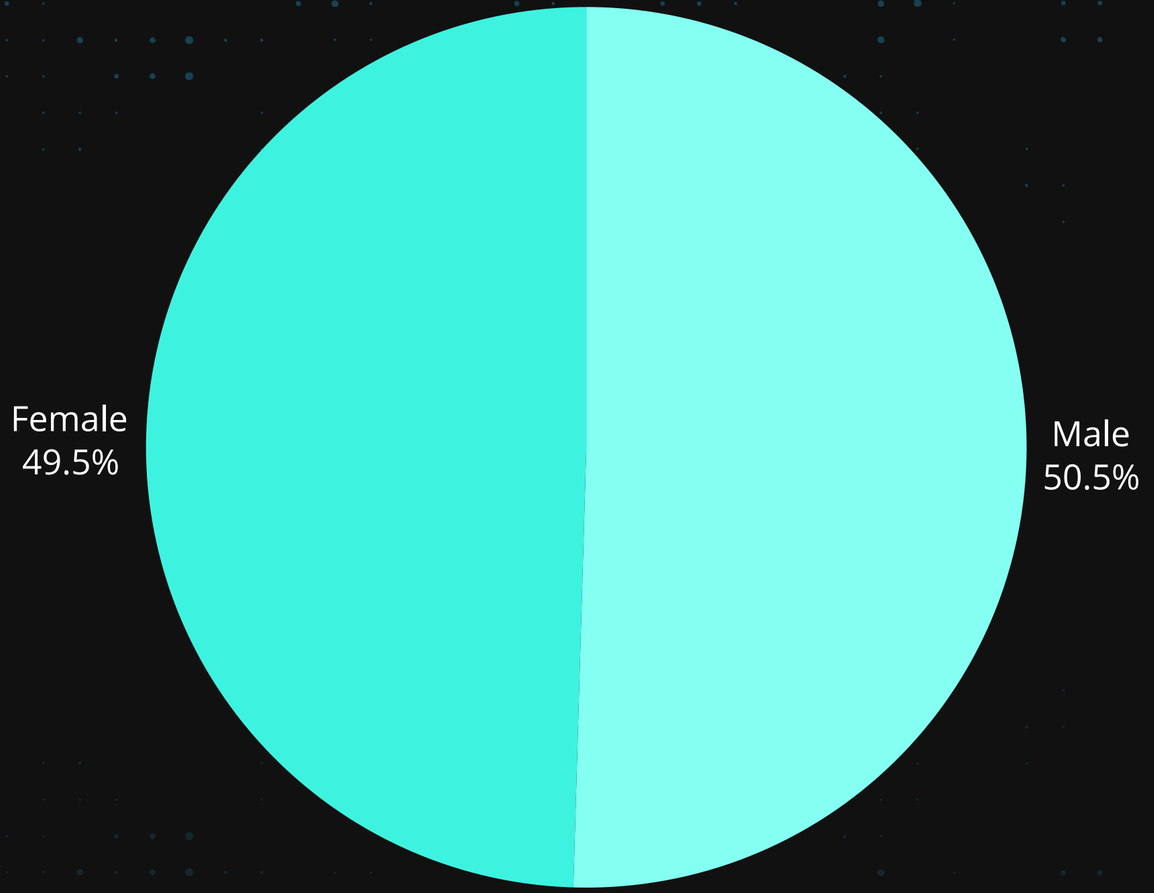
Churn



UNIVARIATE ANALYSIS

FEATURE

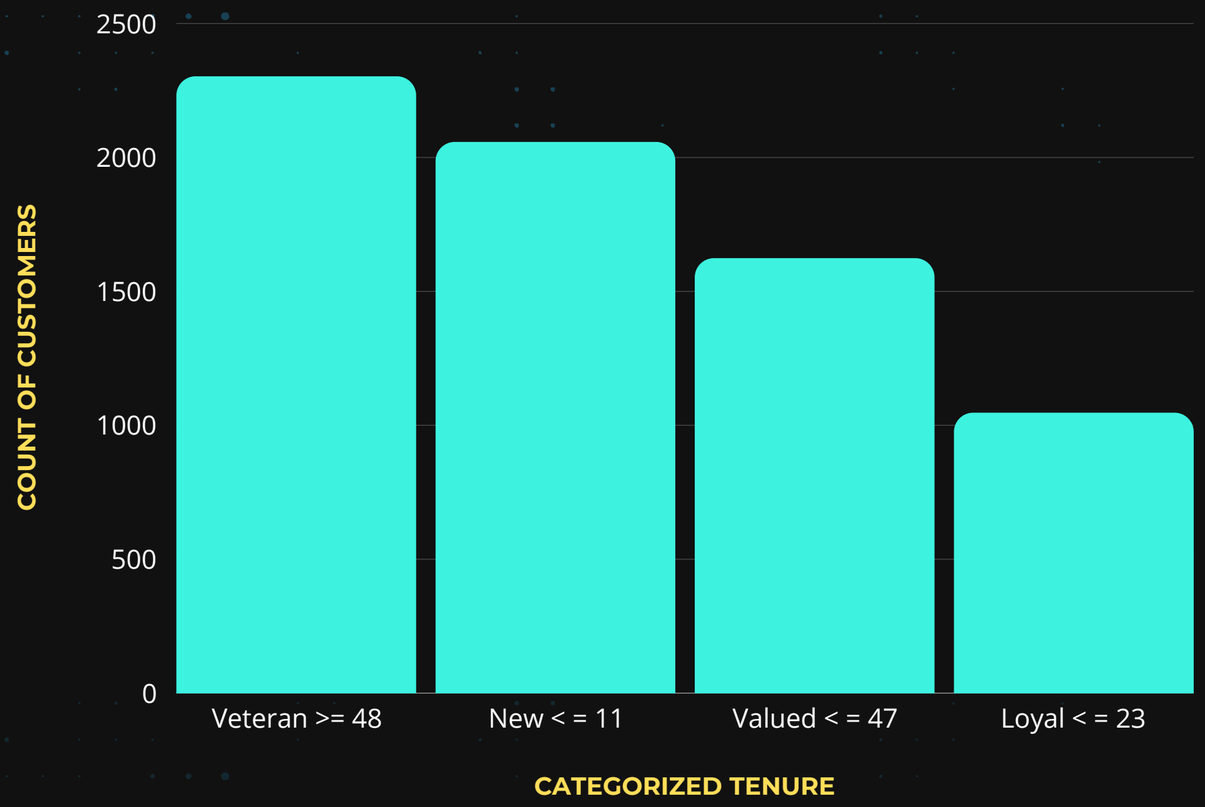
Gender



UNIVARIATE ANALYSIS

FEATURE

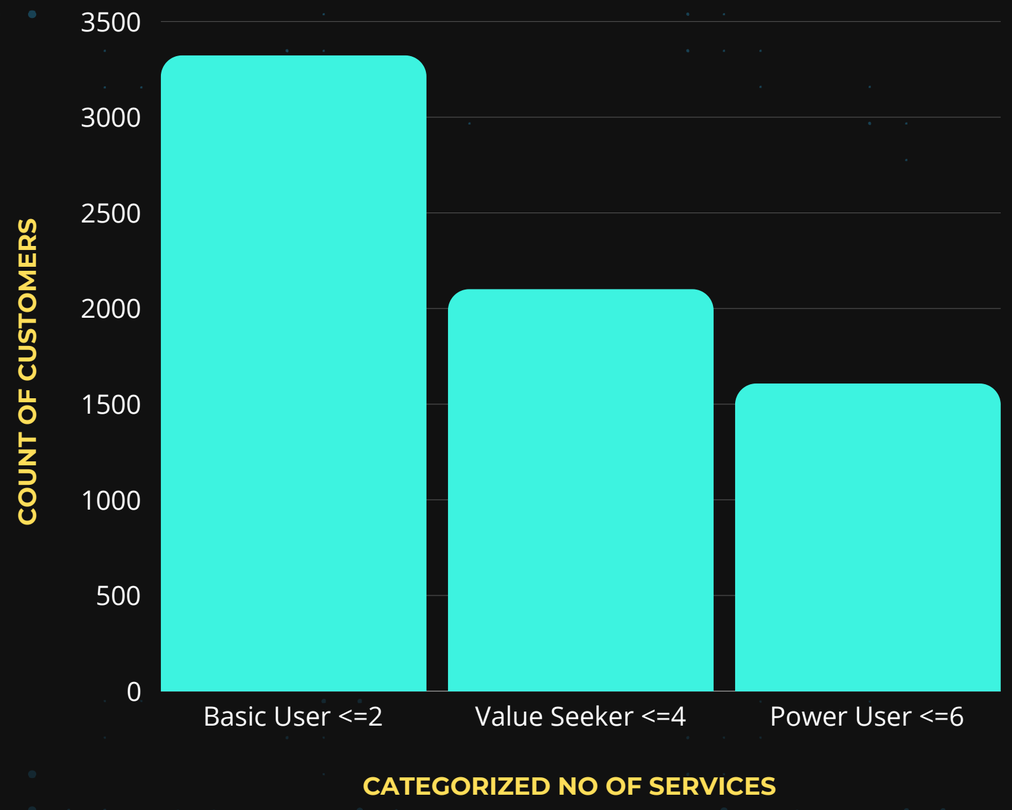
Type of Customer



UNIVARIATE ANALYSIS

FEATURE

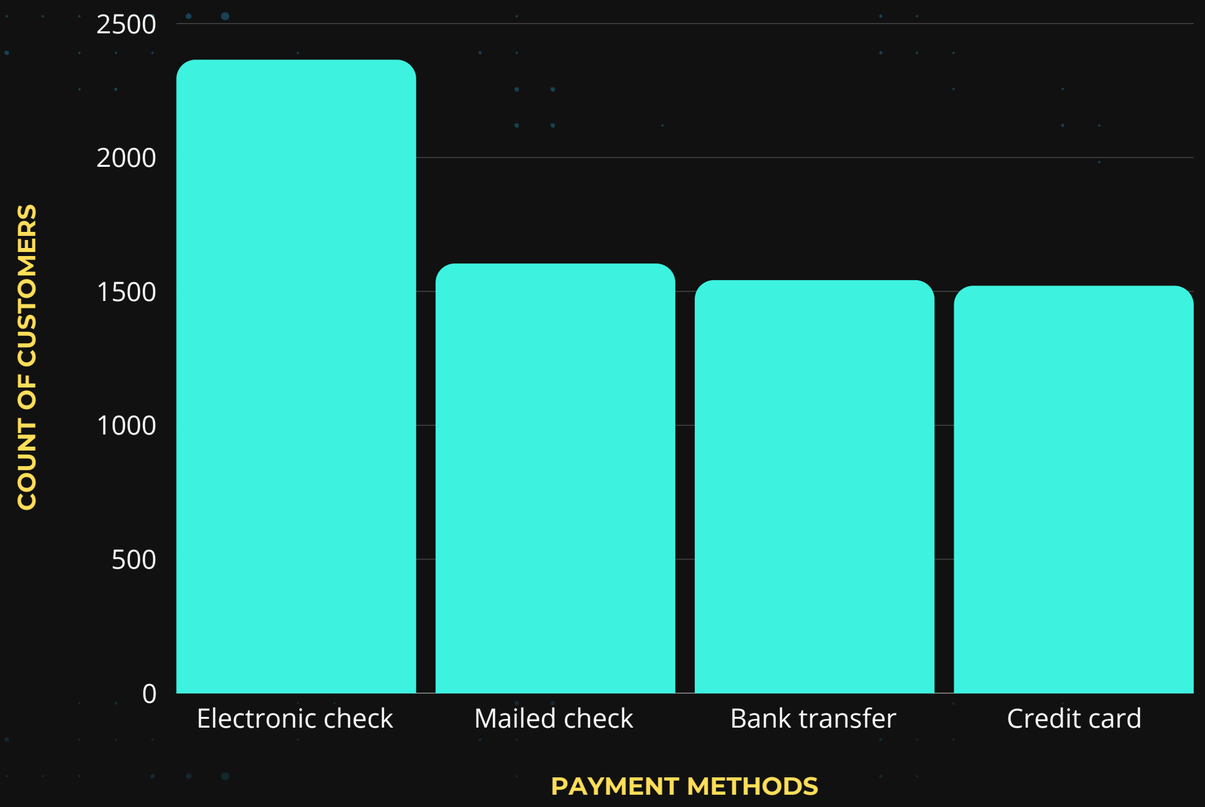
Type of Service Engagement



UNIVARIATE ANALYSIS

FEATURE

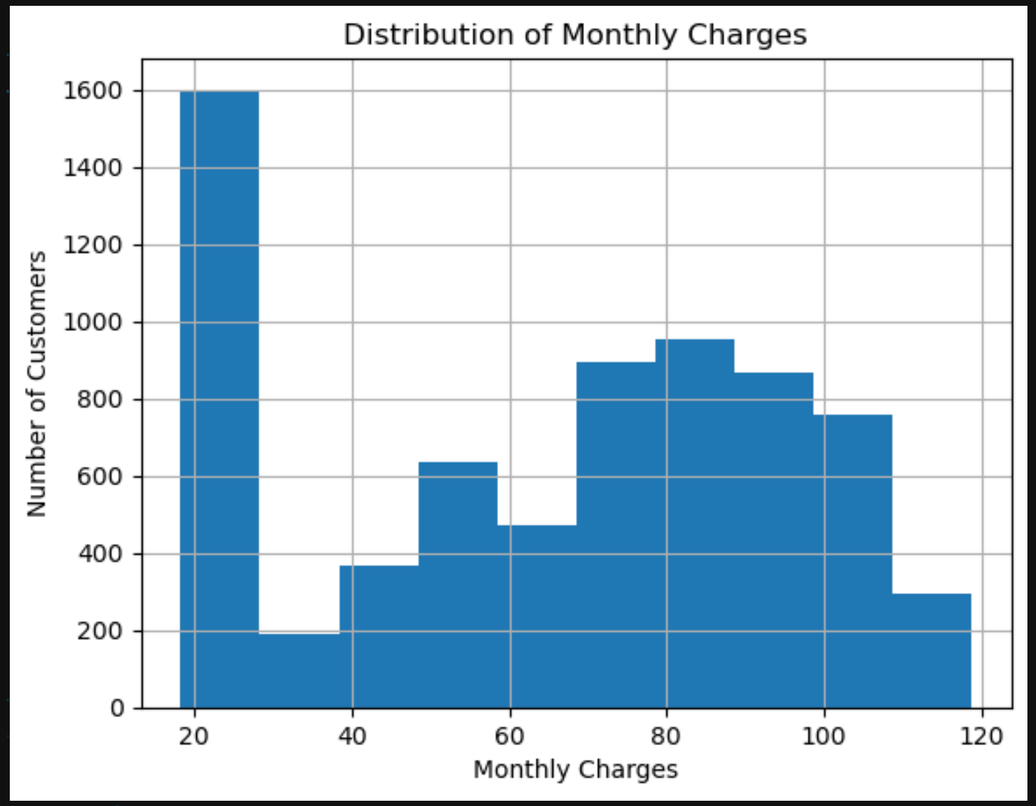
Payment Method



UNIVARIATE ANALYSIS

FEATURE

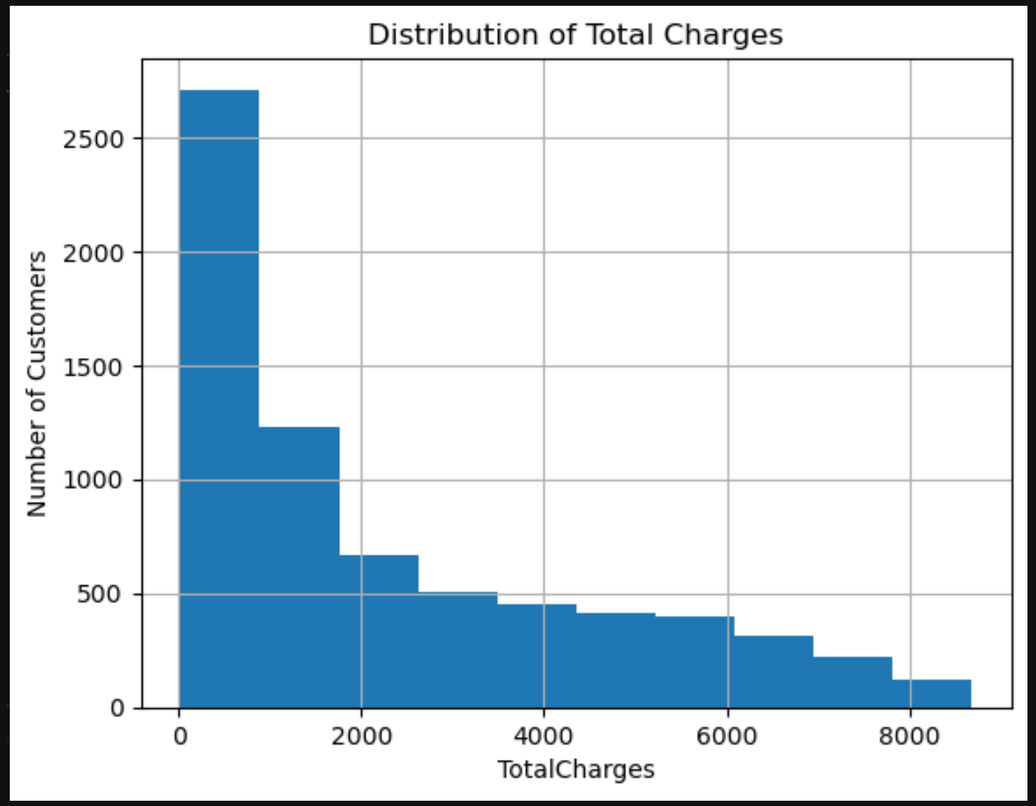
Monthly Charges



UNIVARIATE ANALYSIS

FEATURE

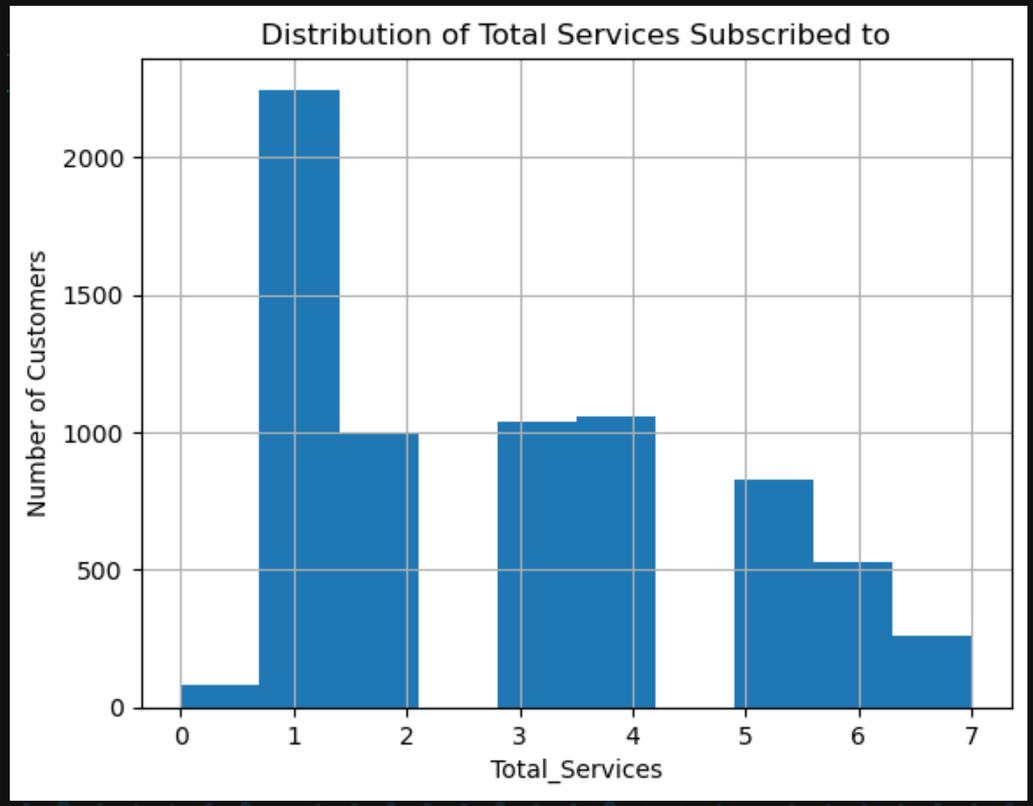
Total Charges



UNIVARIATE ANALYSIS

FEATURE

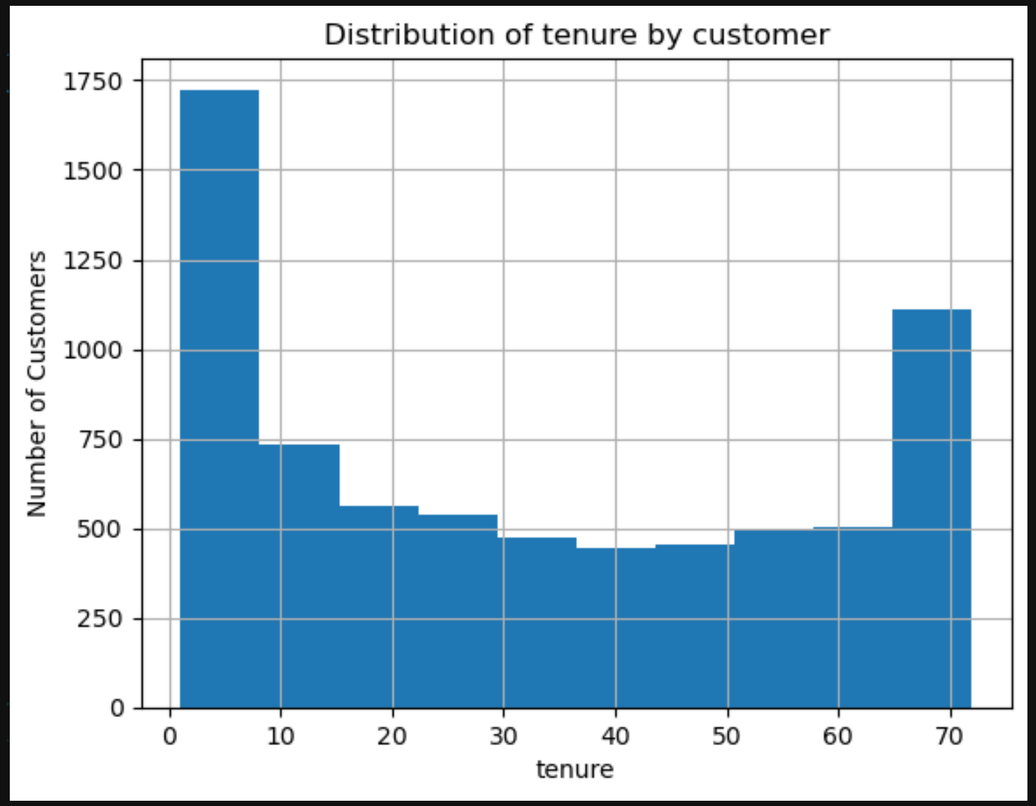
Total Services



UNIVARIATE ANALYSIS

FEATURE

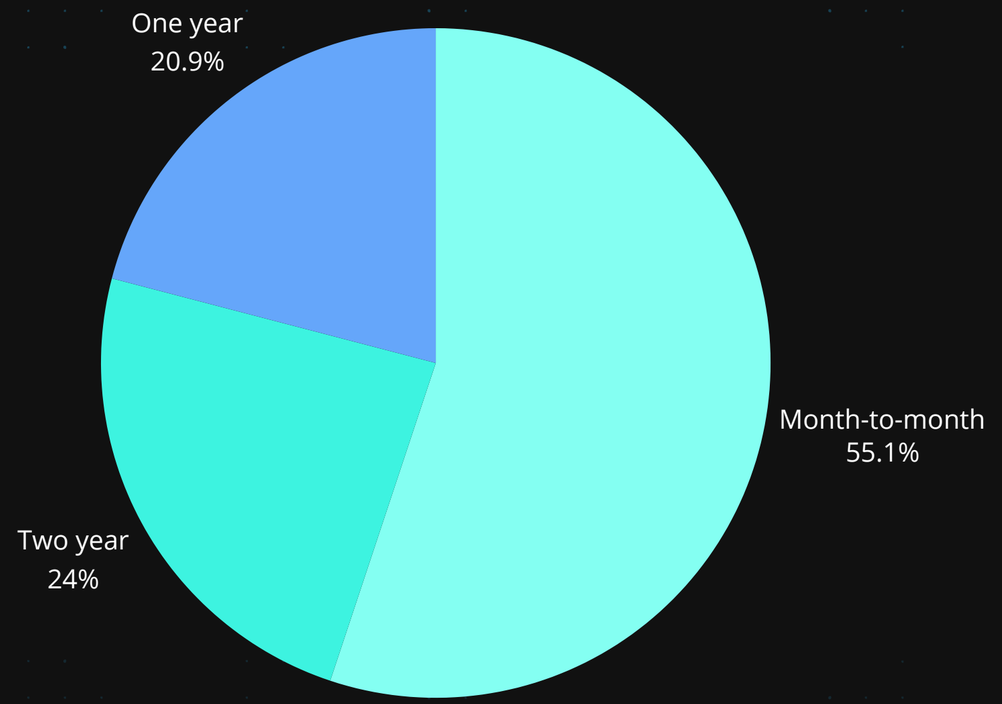
Tenure



UNIVARIATE ANALYSIS

FEATURE

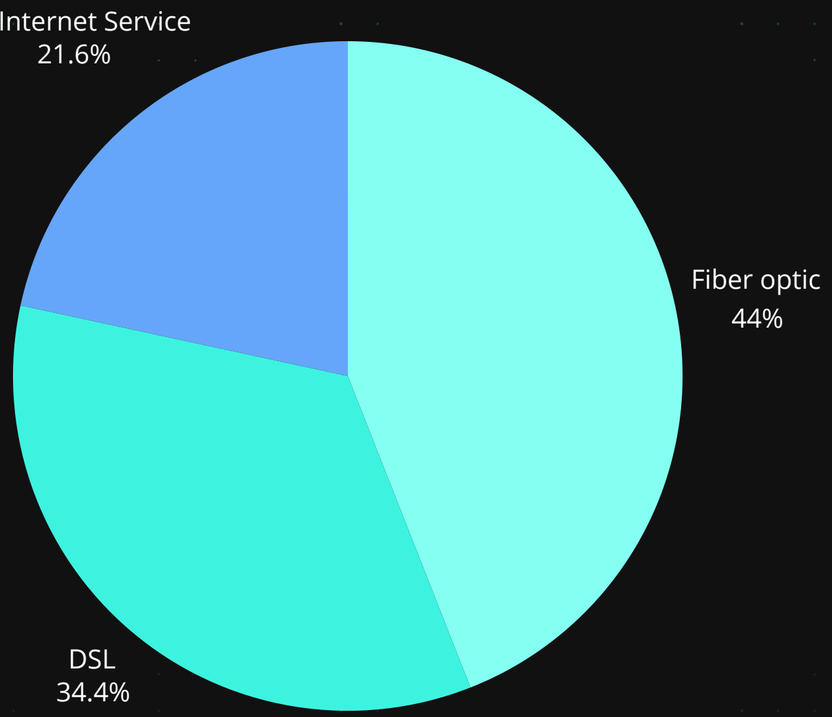
Contract



UNIVARIATE ANALYSIS

FEATURE

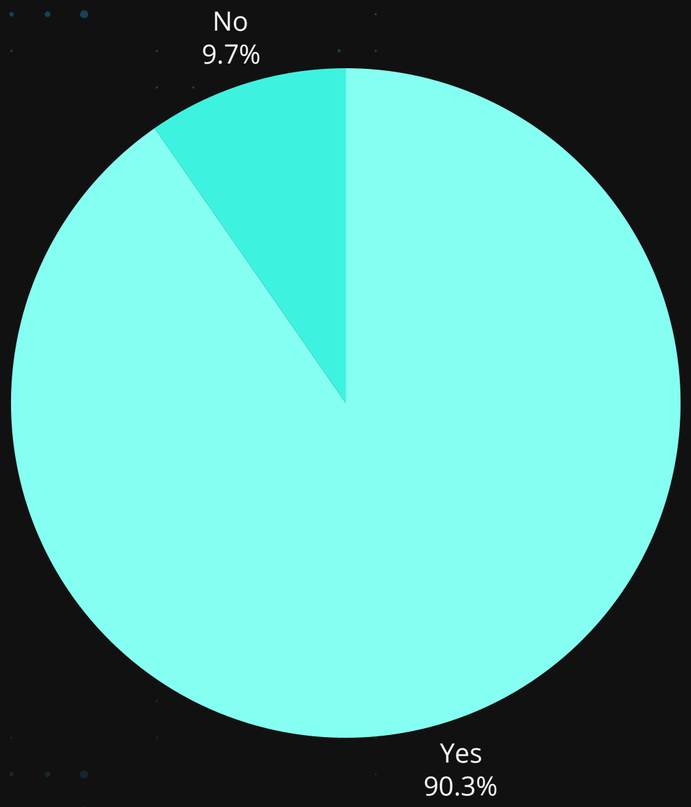
Internet Service



UNIVARIATE ANALYSIS

FEATURE

Phone Service



UNIVARIATE ANALYSIS

CUSTOMER SEGMENTATION - KEY FINDINGS

- Service Engagement:
  - Most users (52.1%) are low-engagement (Basic Users) with 1-2 services.
- Churn:
  - Active customers: 73.4%
  - Churned customers: 26.6%
- Tenure:
  - Veteran customers (48+ months) are the largest group (32.8%).
- Payment:
  - Electronic Check is the preferred method (33.6%).
- Demographics:
  - Gender split is nearly even (50.5% male, 49.5% female).

UNIVARIATE ANALYSIS

CUSTOMER SEGMENTATION - KEY FINDINGS

- Charges:
  - Most users pay 19–28 monthly.
  - Most have a total charge between 0���1000.
- Tenure Distribution:
  - Most users have been with you for 0-8 months.
- Contract:
  - Monthly contracts are most popular (55.1%).
- Senior Citizens:
  - Most users (83.8%) are not senior citizens.
- Partners & Dependents:
  - Most users have no partner (51.7%) and no dependents (70.2%).

UNIVARIATE ANALYSIS

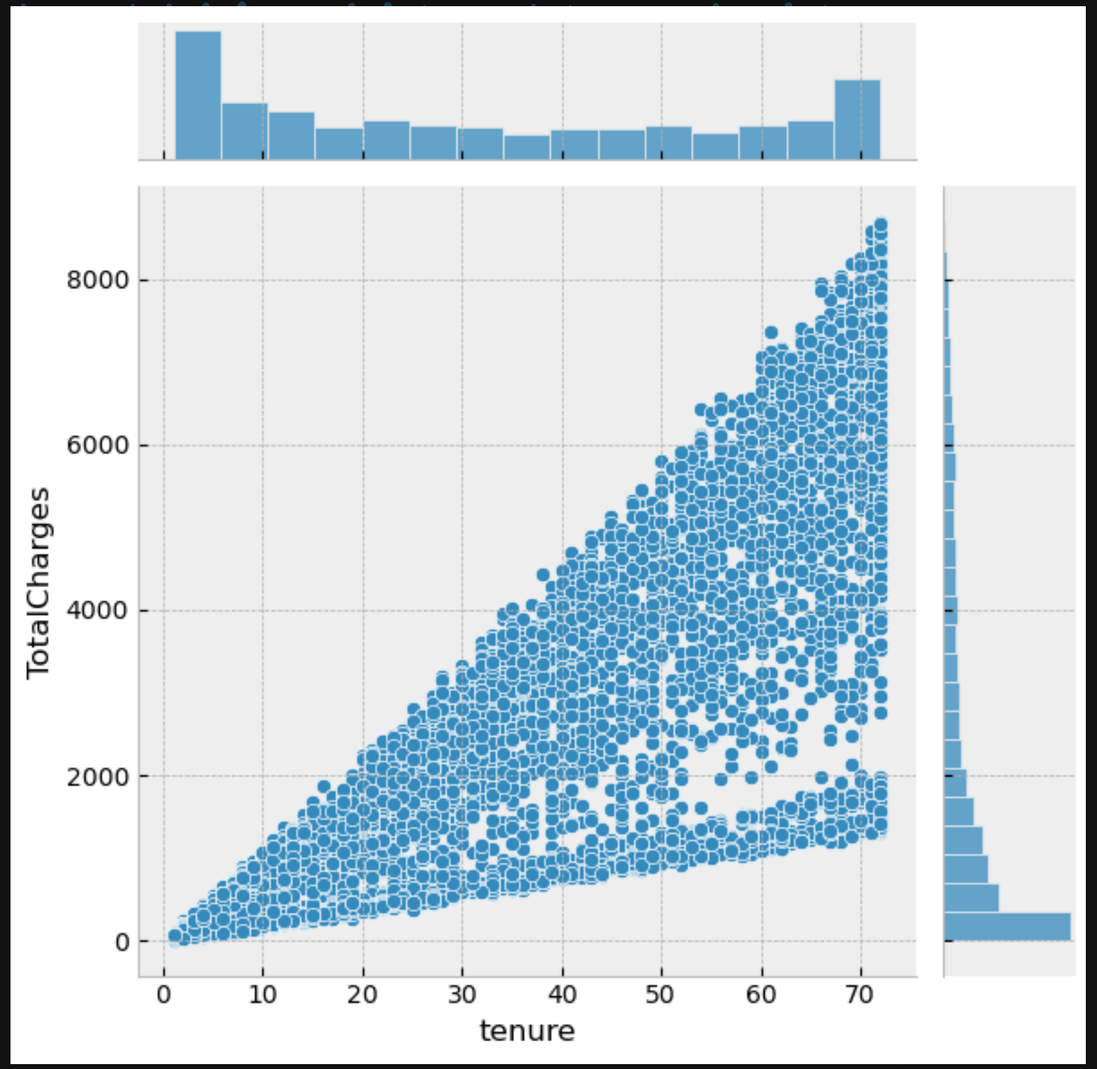
CUSTOMER SEGMENTATION - KEY FINDINGS

- Service Usage:
  - Phone service is dominant (90.3%).
  - Fiber Optic internet leads (44.0%) over DSL (34.4%).
- Additional Services:
  - Many lack online security (49.7%), backup (43.9%), and movie streaming (39.5%).
- Billing:
  - Paperless billing is preferred (59.3%).
- Number of Services:
  - Most users subscribe to only 1 service (34%).
  - No one subscribes to all services.

BIVARIATE ANALYSIS

FEATURE

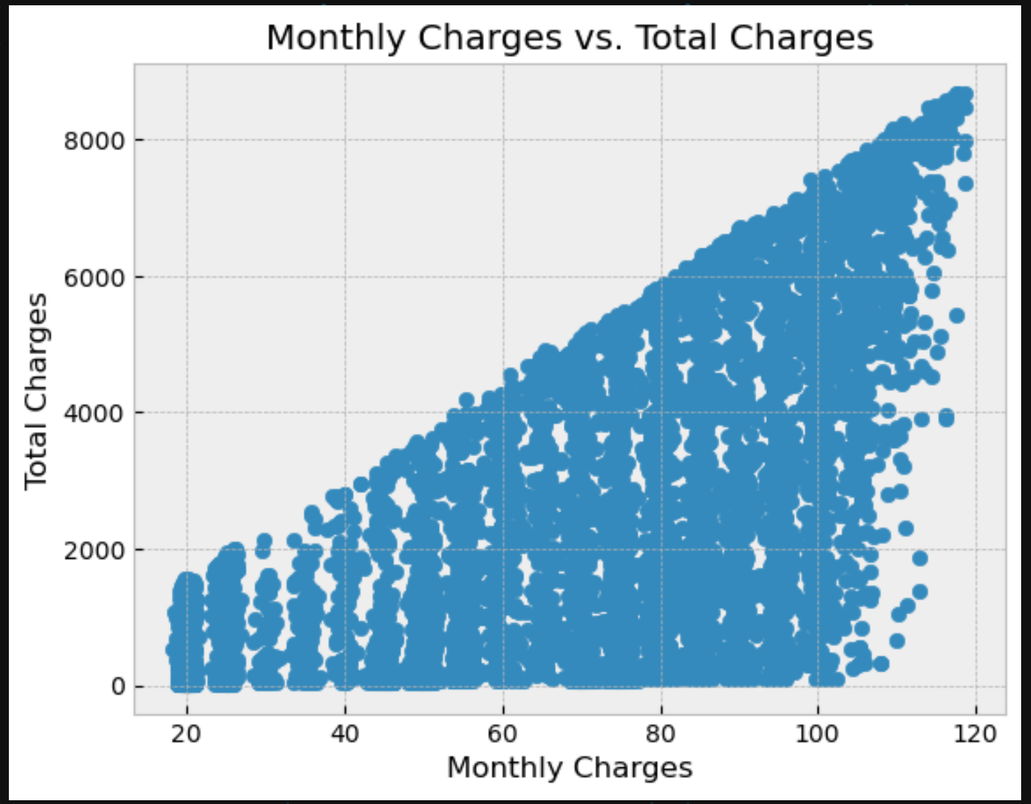
TotalCharges & Tenure



BIVARIATE ANALYSIS

FEATURE

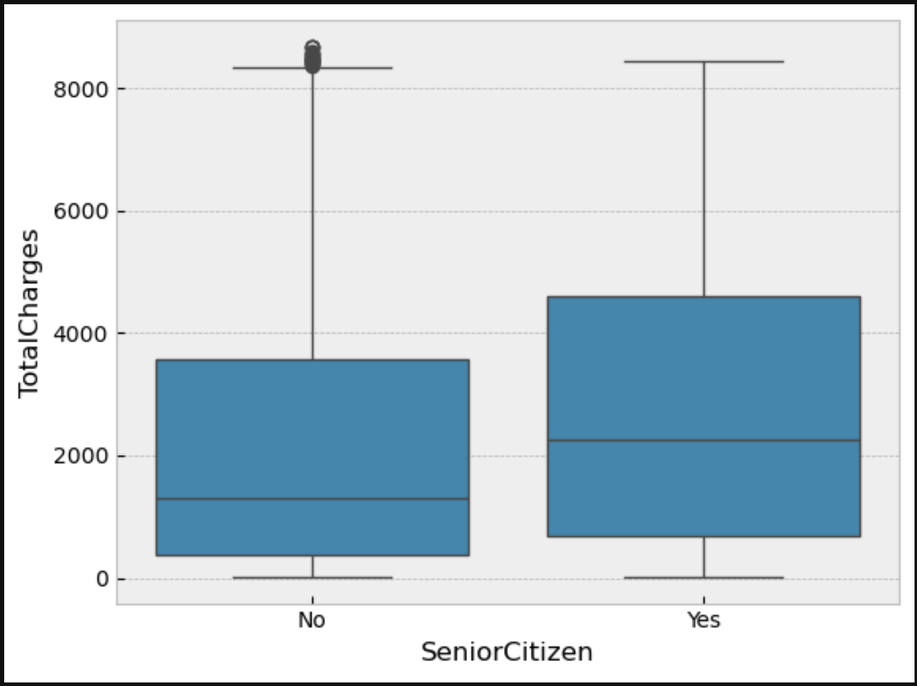
TotalCharges & MonthlyCharges

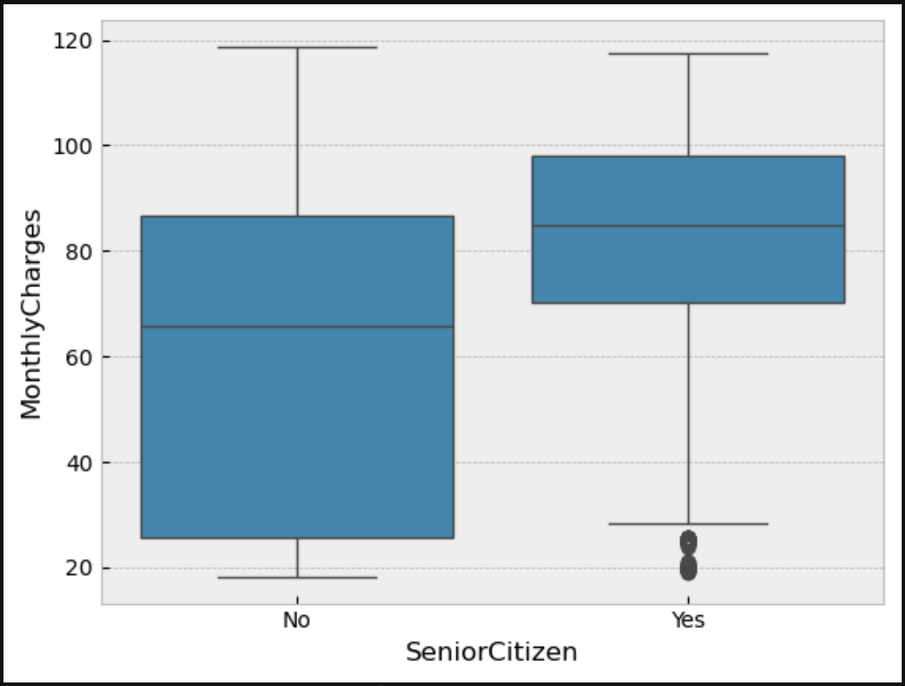


BIVARIATE ANALYSIS

FEATURE

Senior Citizen with monthly and total charges

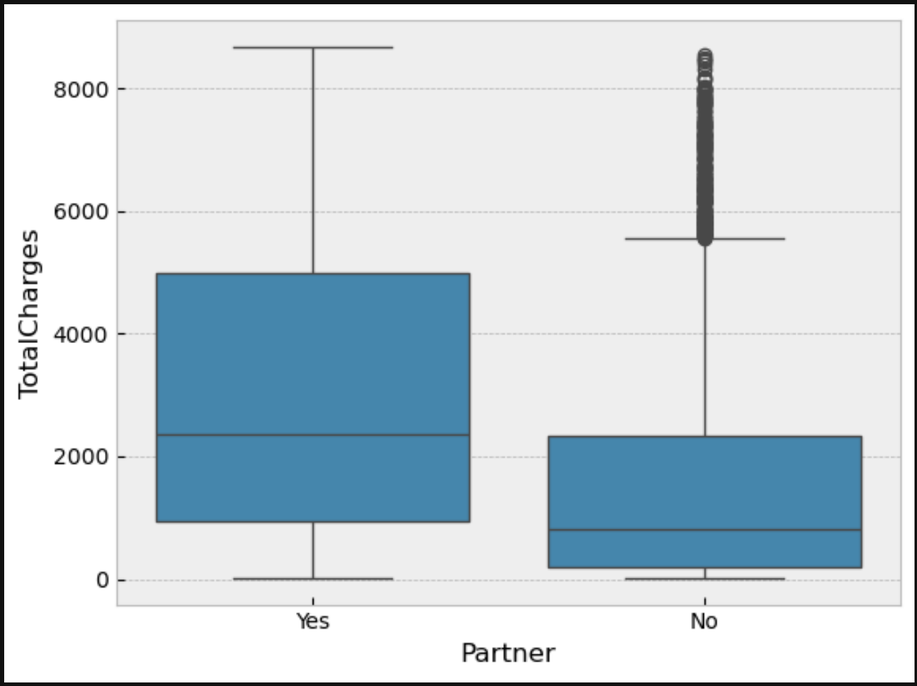


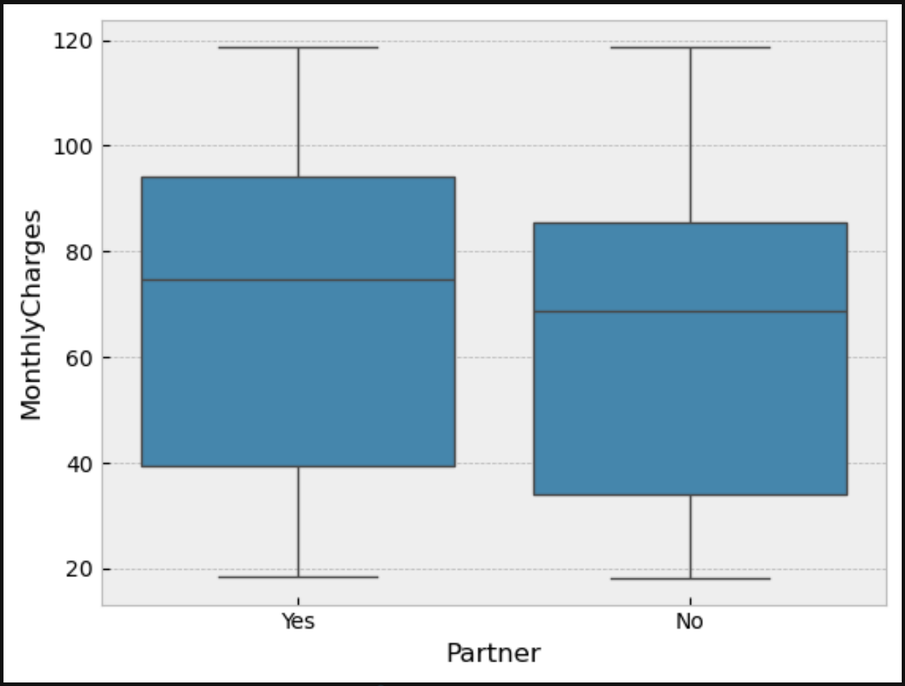


BIVARIATE ANALYSIS

FEATURE

Partner with monthly and total charges

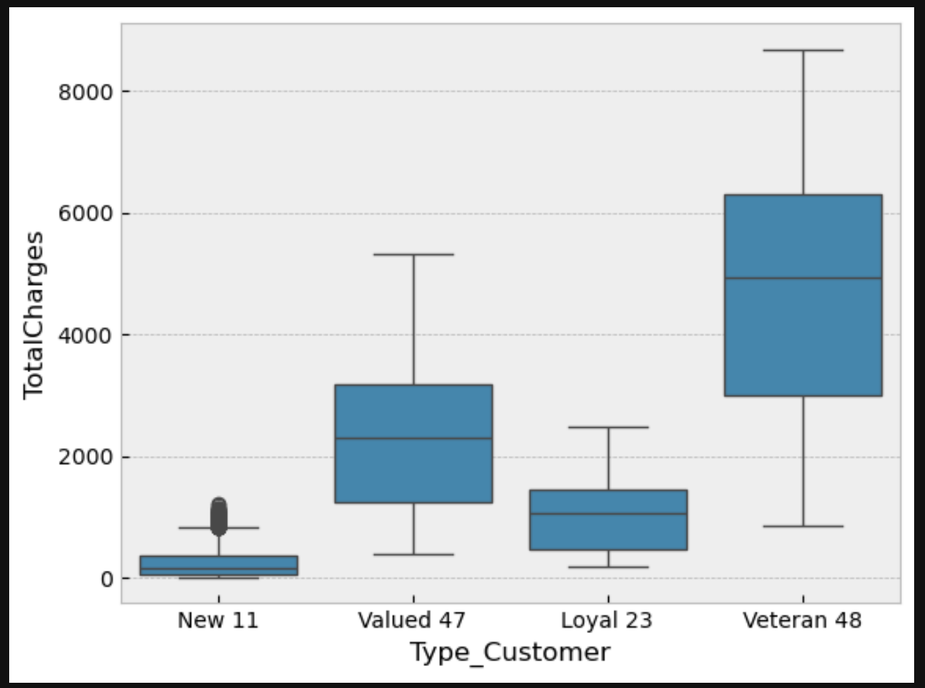


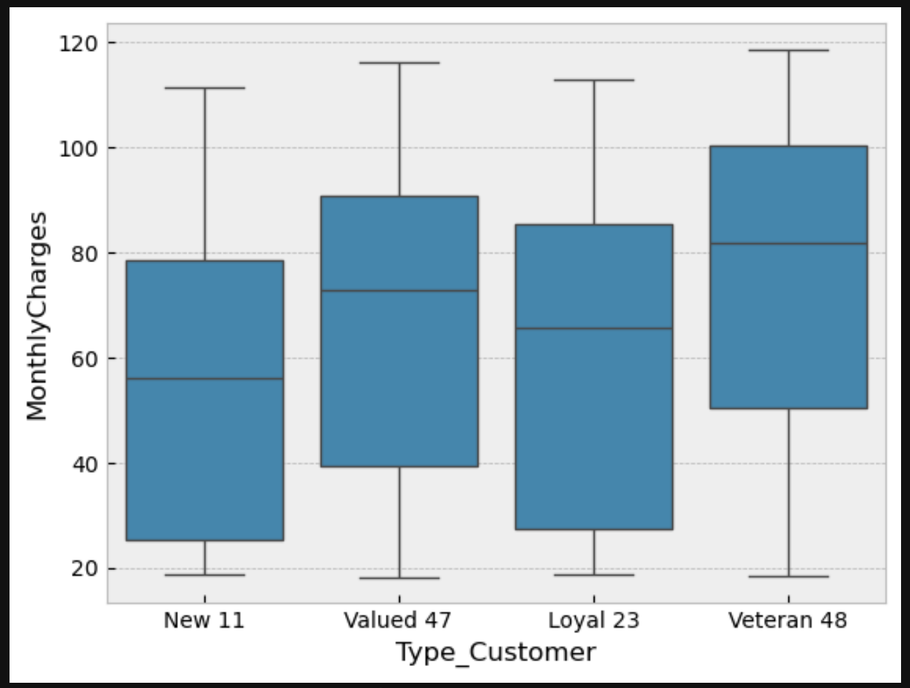


BIVARIATE ANALYSIS

FEATURE

Type of Customer with monthly and total charges

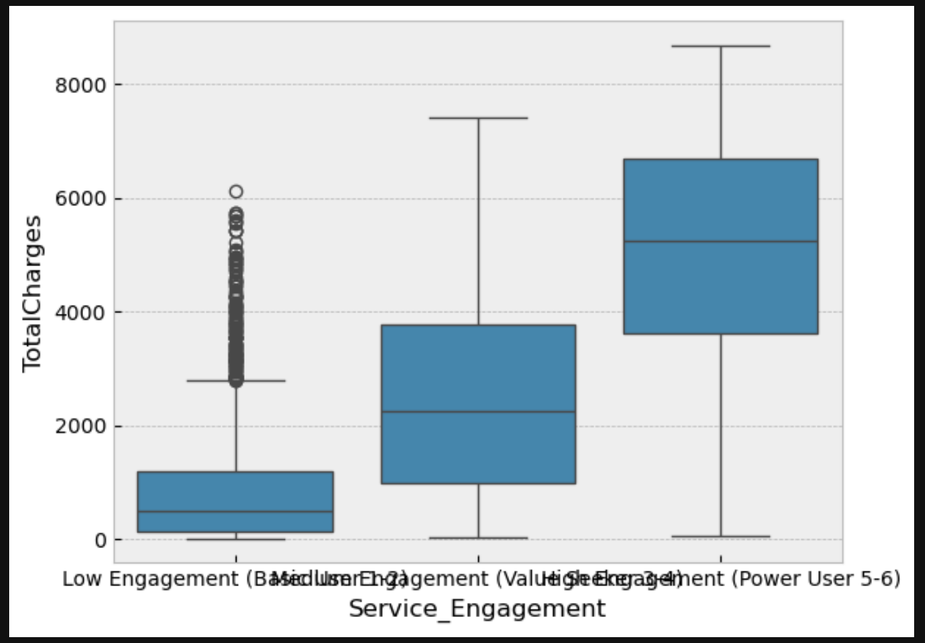


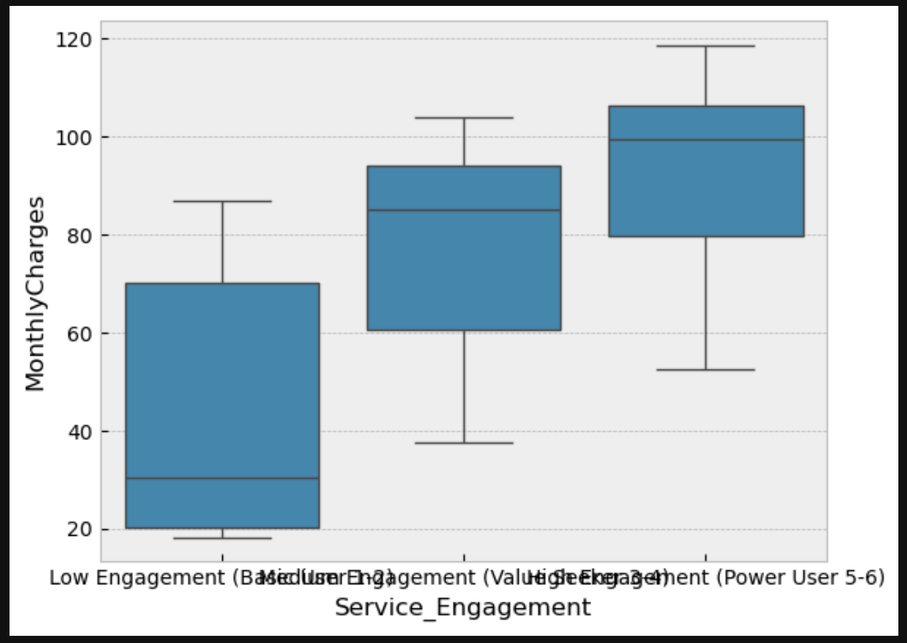


BIVARIATE ANALYSIS

FEATURE

Service Engagement with monthly and total charges

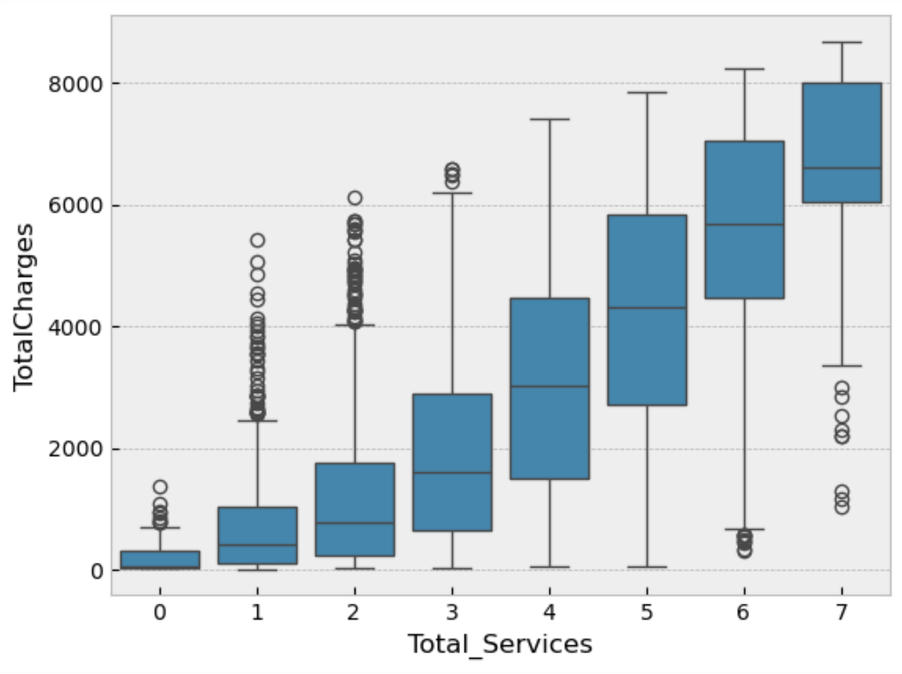


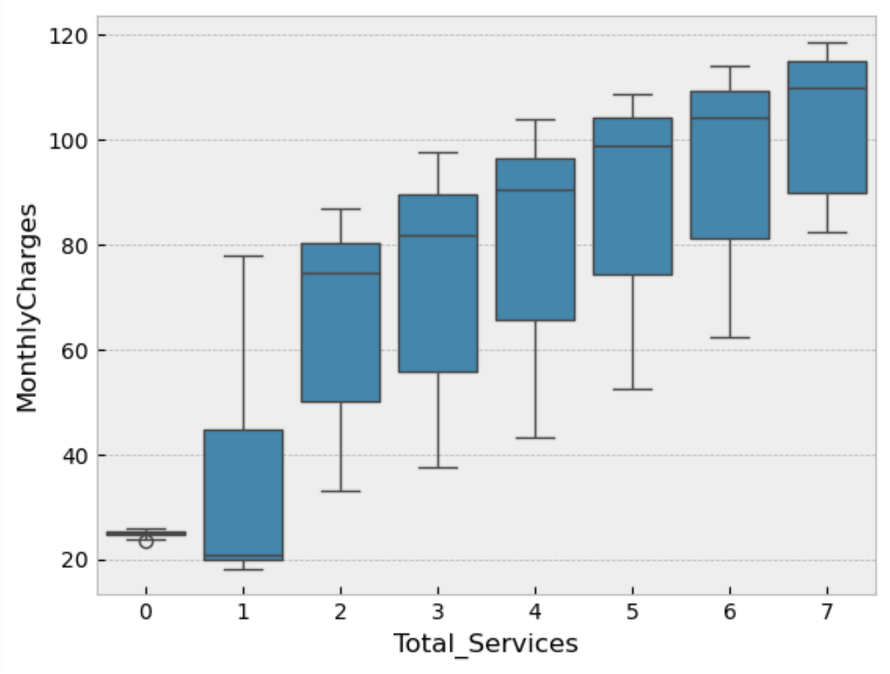


BIVARIATE ANALYSIS

FEATURE

Total Services with monthly and total charges

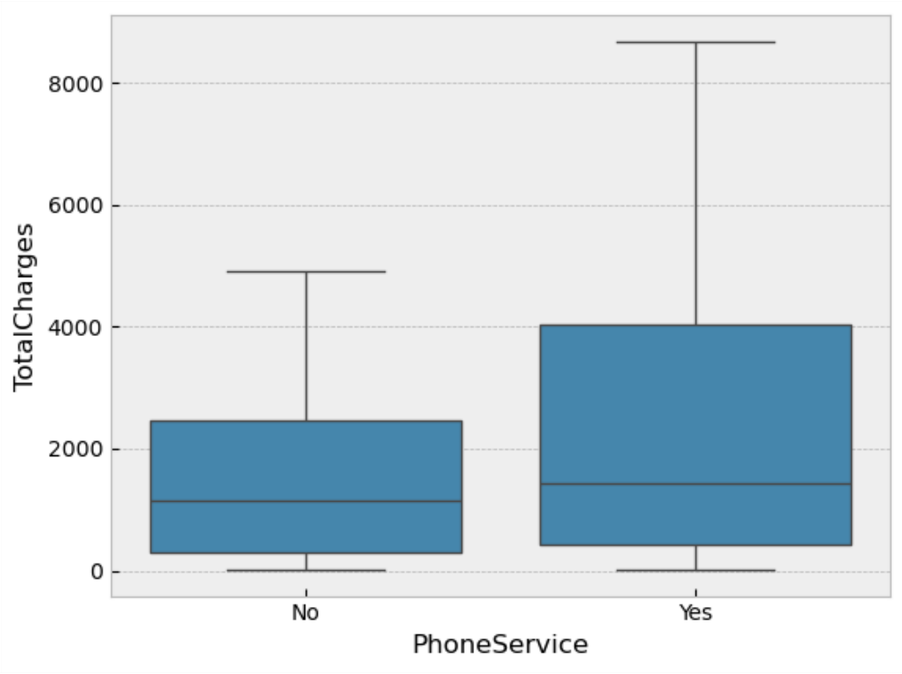


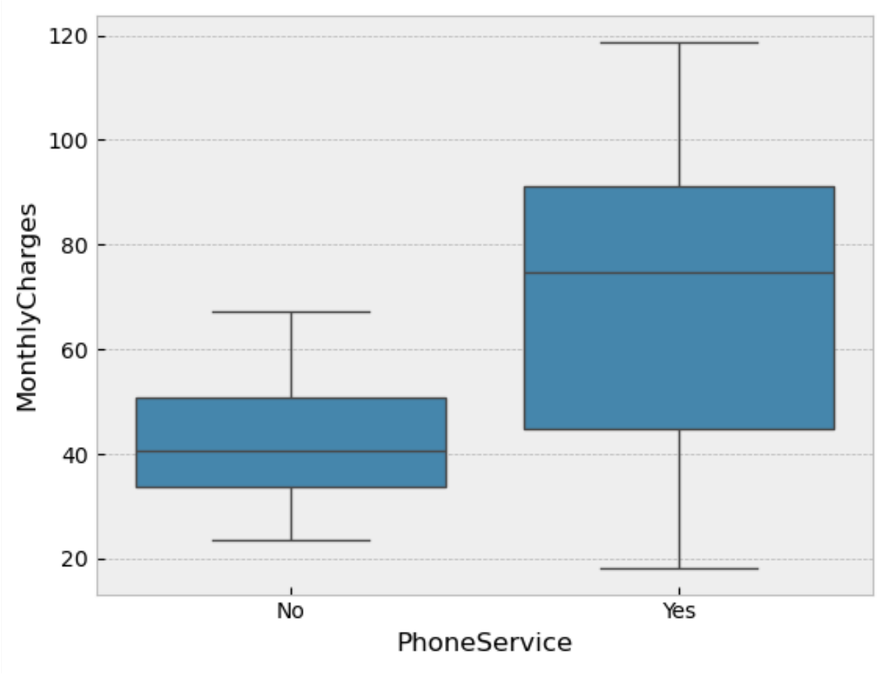


BIVARIATE ANALYSIS

FEATURE

Phone Services with monthly and total charges

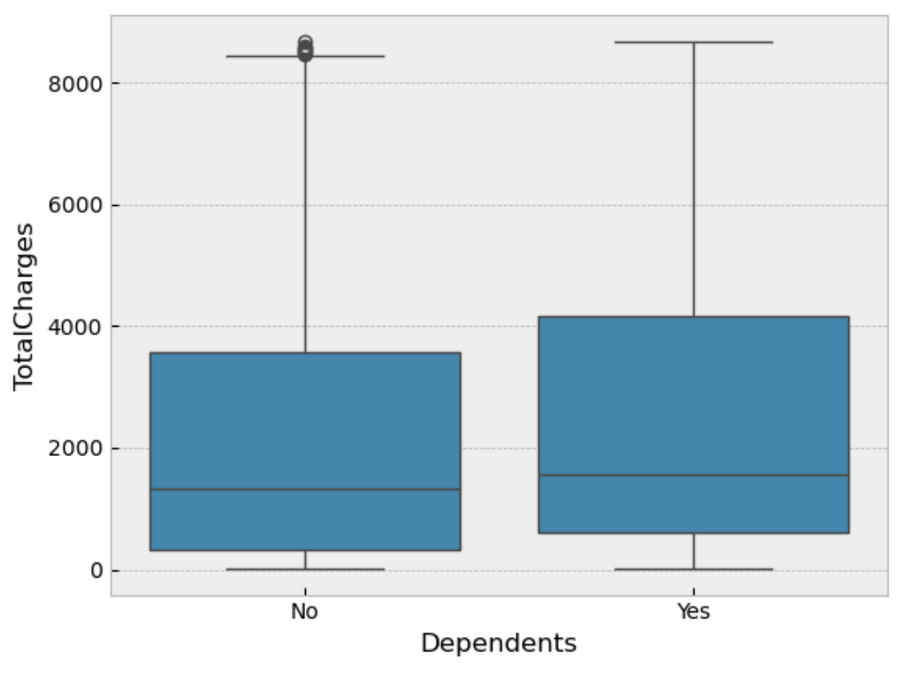


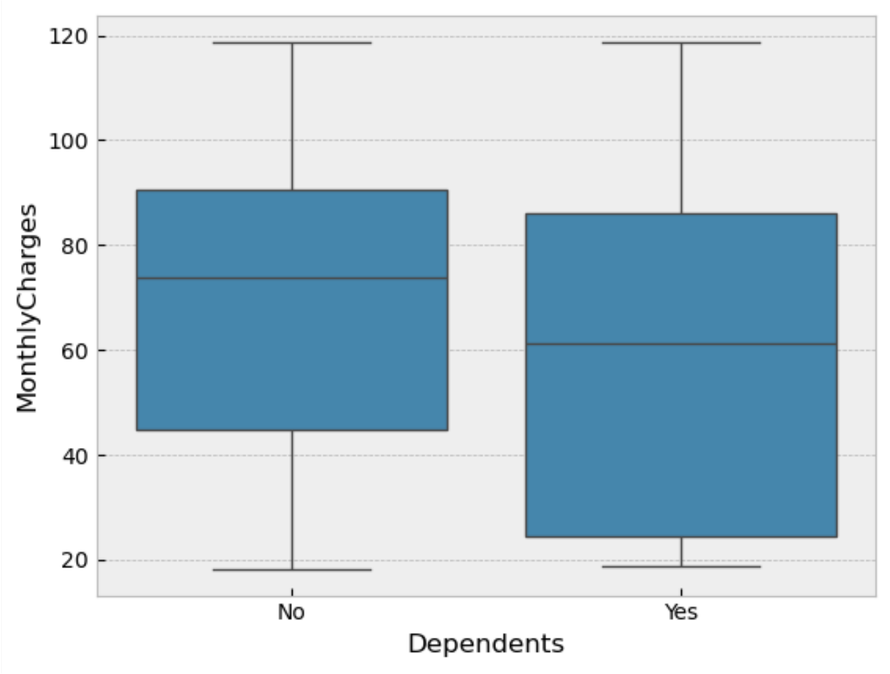


BIVARIATE ANALYSIS

FEATURE

Dependents with monthly and total charges

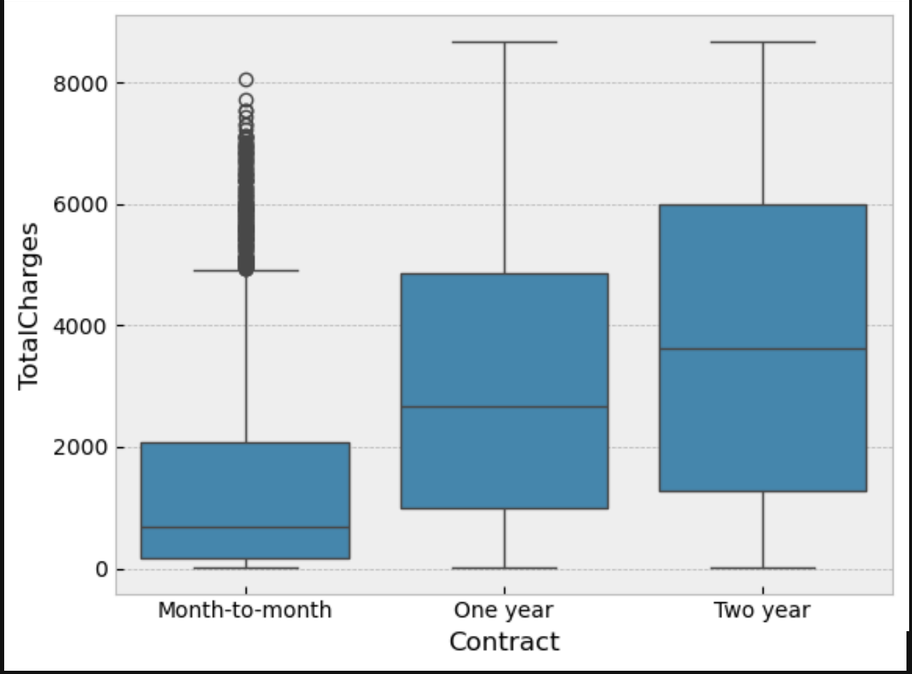


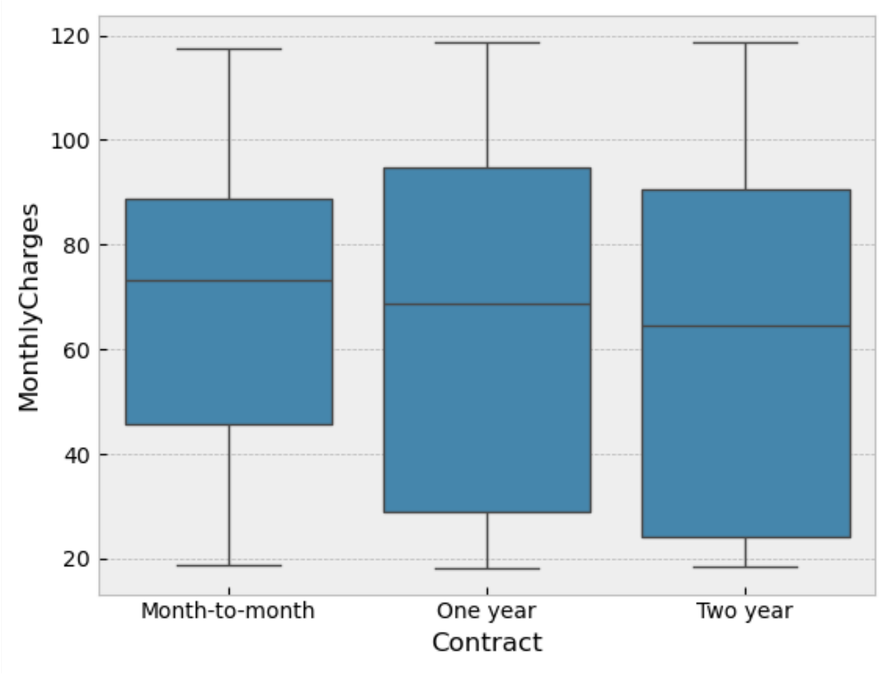


BIVARIATE ANALYSIS

FEATURE

Contract with monthly and total charges

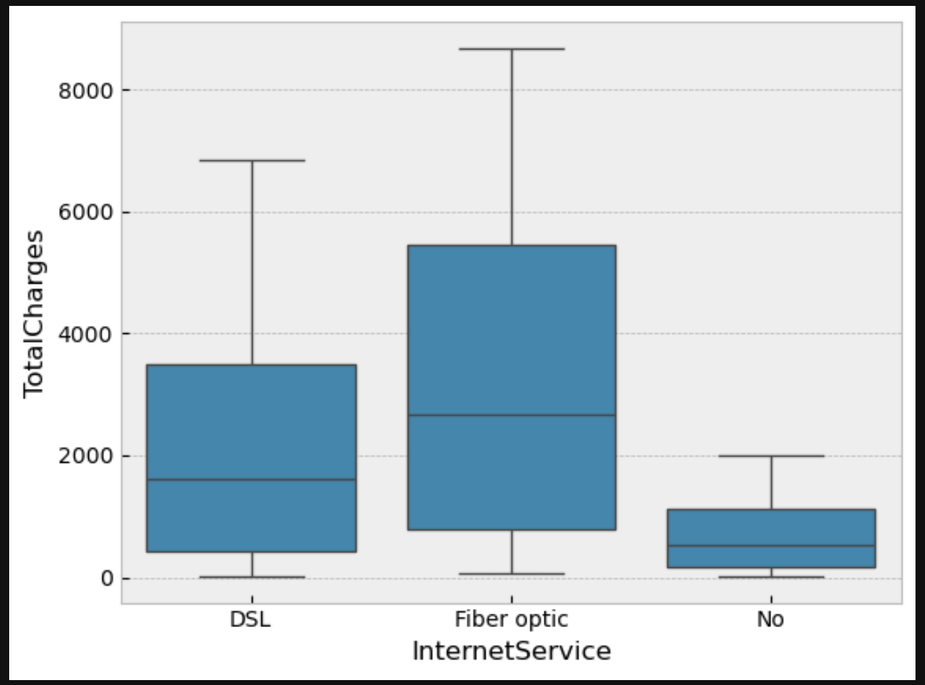


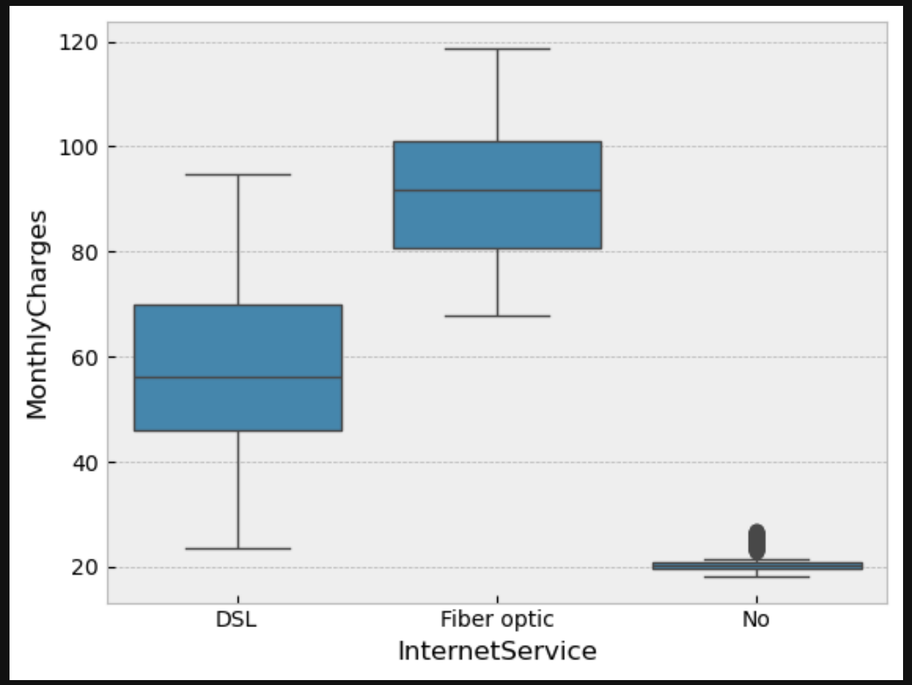


BIVARIATE ANALYSIS

FEATURE

Internet Service with monthly and total charges

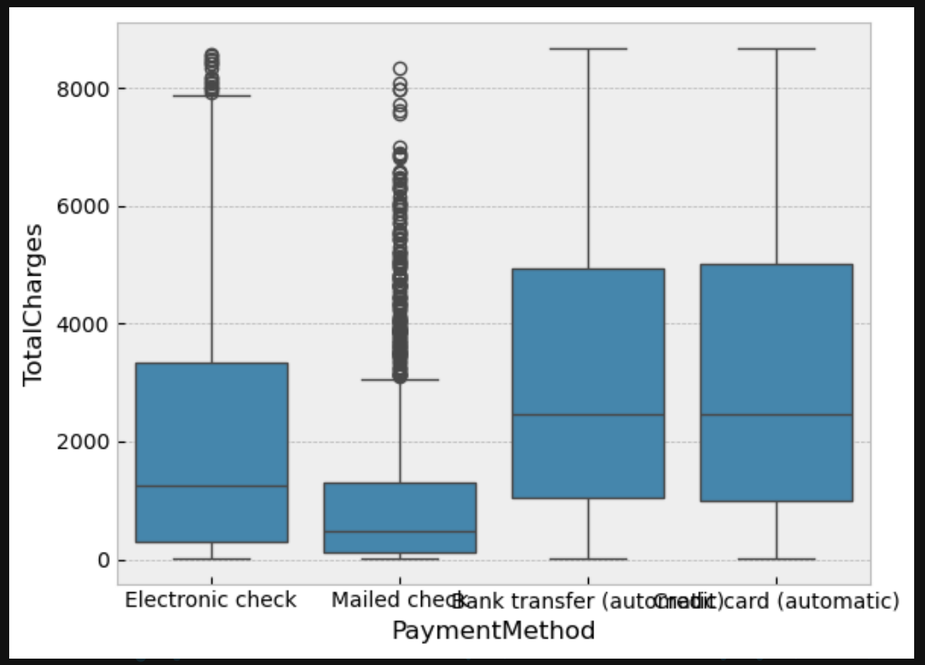


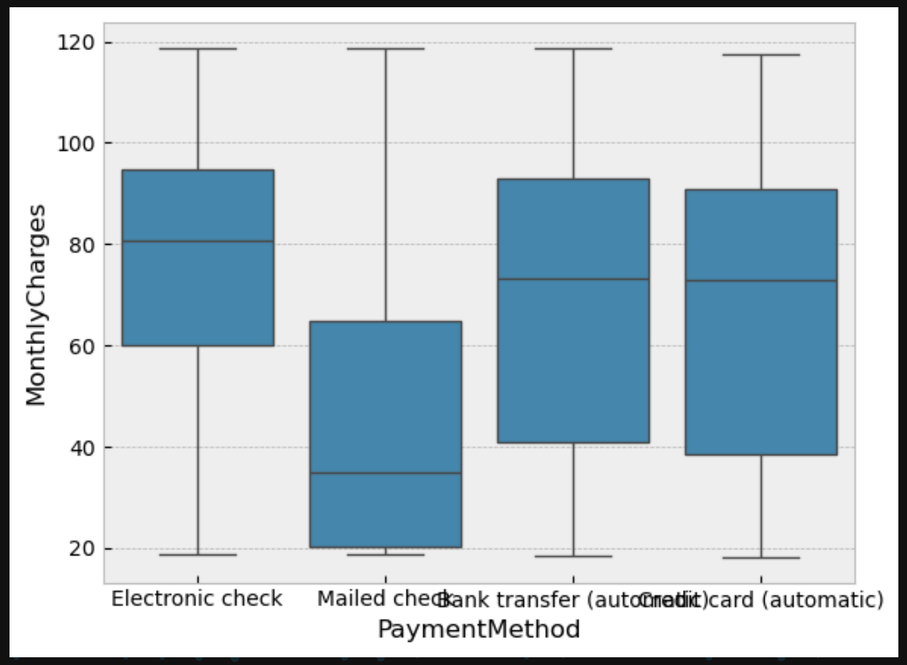


BIVARIATE ANALYSIS

FEATURE

Payment Method with monthly and total charges

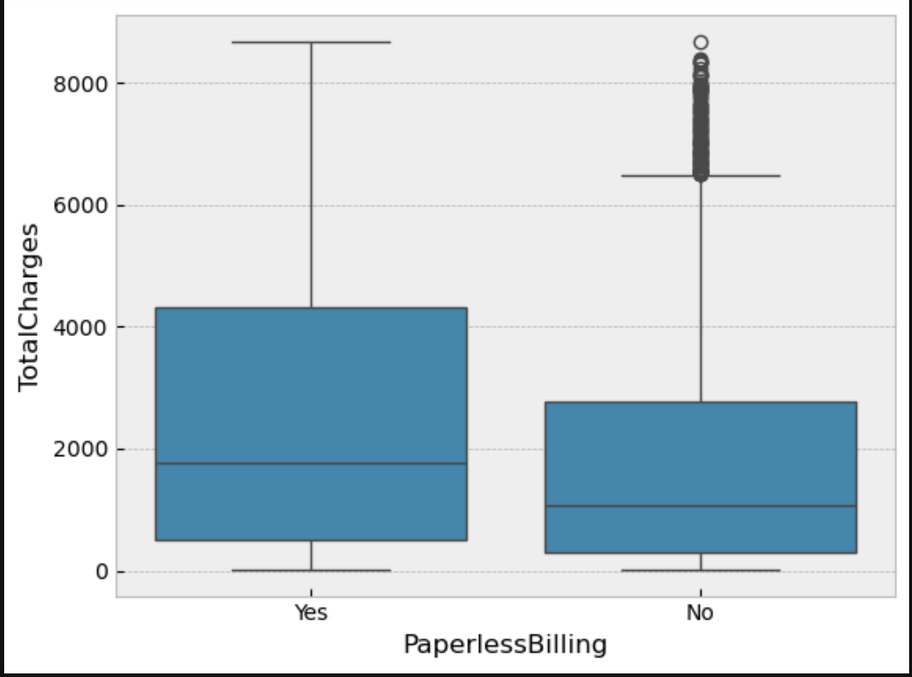


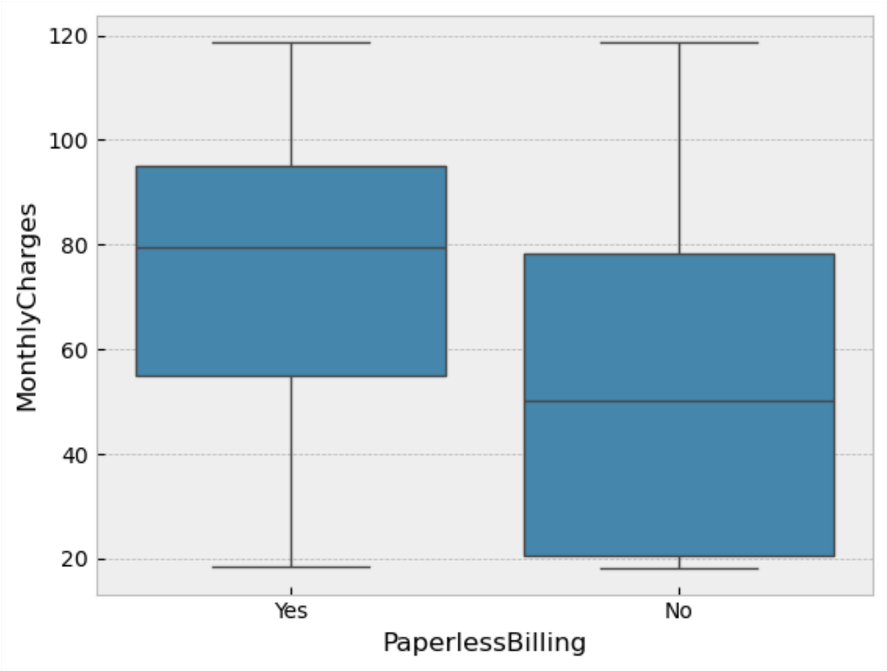


BIVARIATE ANALYSIS

FEATURE

Paperless Billing with monthly and total charges

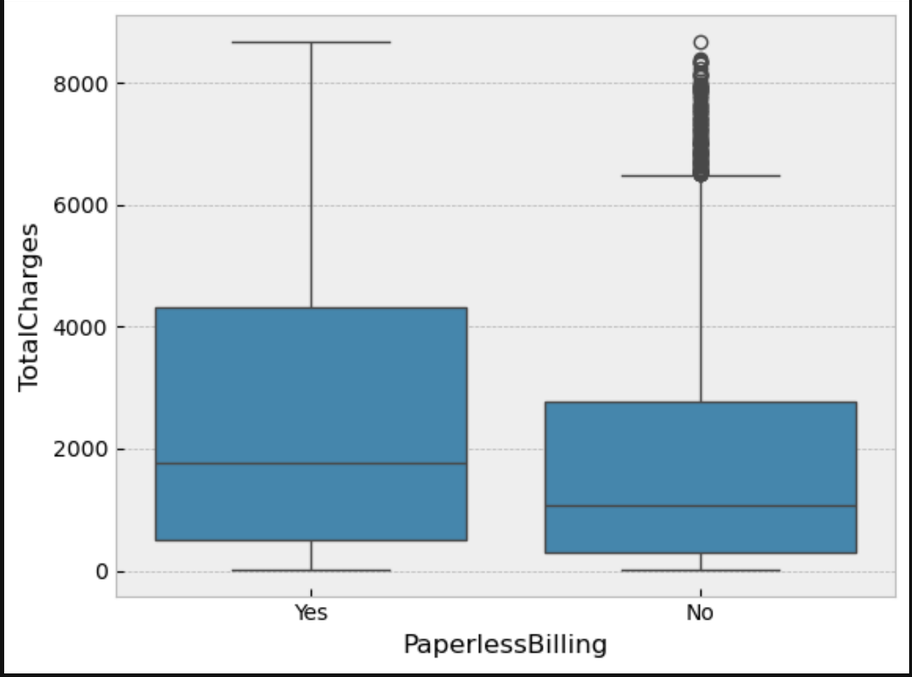


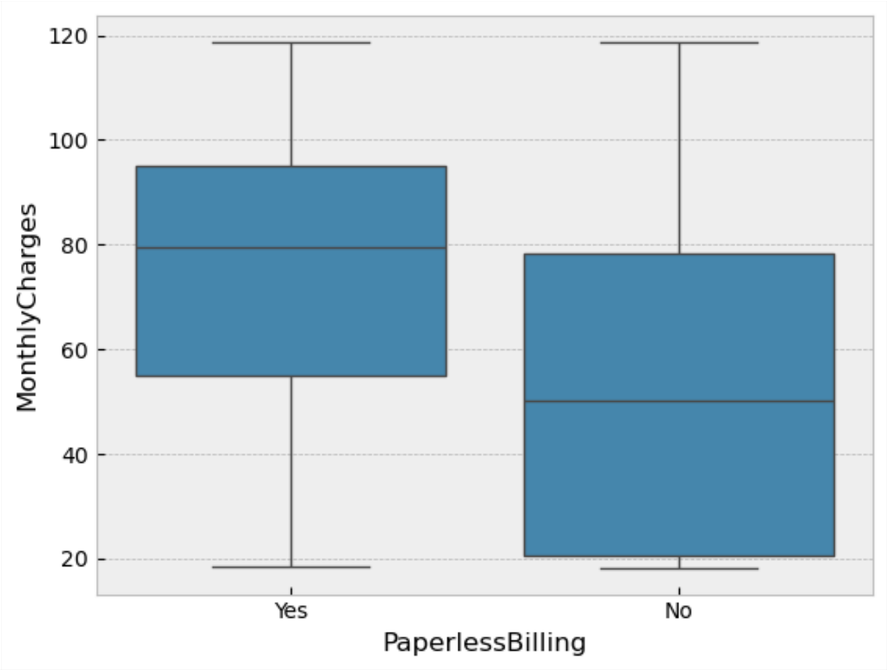


BIVARIATE ANALYSIS

FEATURE

Paperless Billing with monthly and total charges





BIVARIATE ANALYSIS

CUSTOMER SEGMENTATION - KEY FINDINGS

- - There's a positive relationship between monthly charges and total charges.
- - There's a positive relationship between tenure and total charges.
- - The probability of male customers churning and female customers is 50:50
- - Based on the numbers of senior citizen, those who are not senior citizens are more likely to Churn
- - The probability of senior customers churning and not senior customers are 16:84

BIVARIATE ANALYSIS

CUSTOMER SEGMENTATION - KEY FINDINGS

- Senior citizen customers, customers with partners, veterans, and power users tend to have higher total and monthly charges, while new customers and basic users have lower charges.
- Customers with phone services, dependents, and those using fiber optic services also tend to have higher charges.
- Payment methods such as bank transfers and credit cards result in higher total charges, while mailed checks have the lowest.
- Electronic checks have the highest monthly charges.

BIVARIATE ANALYSIS

CUSTOMER SEGMENTATION - KEY FINDINGS

- Additionally, customers supporting paperless billing have higher charges.
- Tenure is positively correlated with total charges.
- Outliers exist in total charges for various customer segments and payment methods, as well as in monthly charges for certain demographics.

BIVARIATE ANALYSIS

COUNT PLOT RELATIONSHIPS BETWEEN IMPORTANT FEATURES AND TARGET CHURN

Most churned customers are male, not senior citizens, have no partners or dependents, use phone service, Fiber Optic internet, lack online security, backup, tech support, device protection, have monthly plans, use paperless billing, electronic check, are basic and new users.

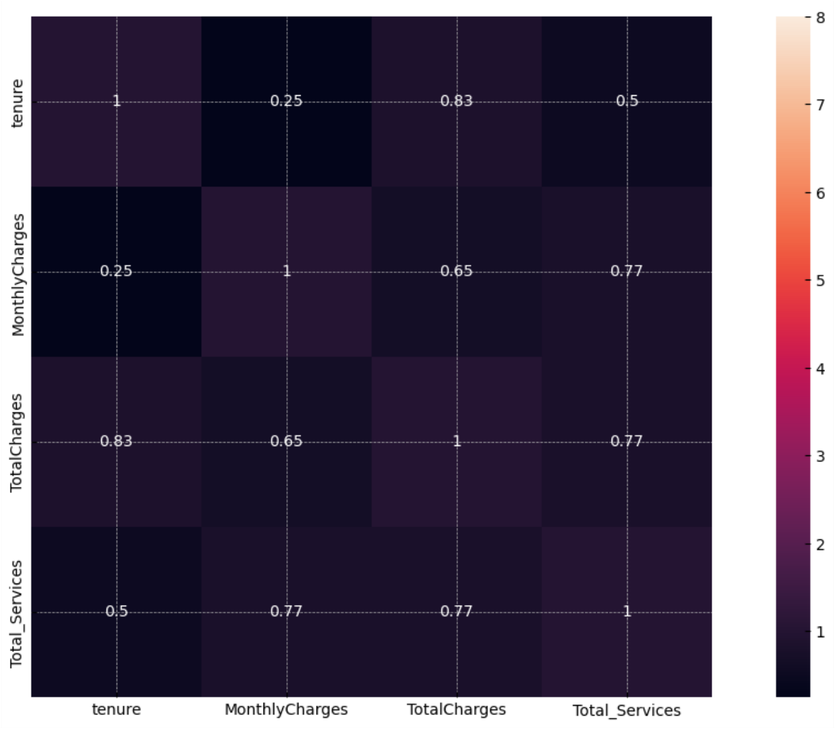
We've been able to identify at risk customers

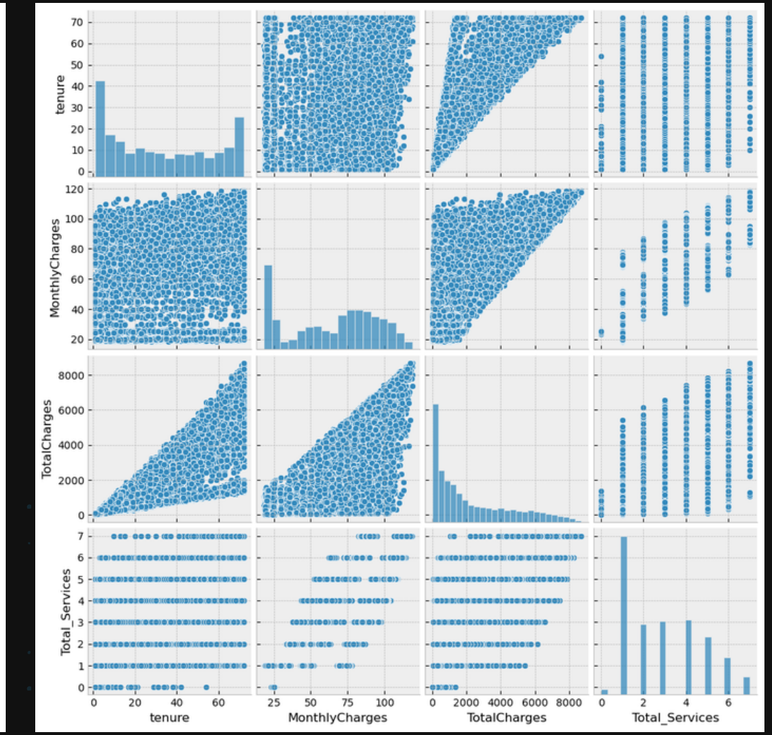
MULTIVARIATE ANALYSIS

FEATURE

Relationships between

Tenure | Monthly Charges | Total Charges | Total Services





- Tenure is correlated with total charges
- Tenure is correlated with total services
- Monthly charges are correlated with total charges
- Monthly charges are correlated with total services
- Total charges are correlated with total services

MACHINE LEARNING MODEL

LOGISTICS REGRESSION

TASKS

WORKING WITH OUTLIERS AND UNDERFITTING METHOD

1. Separated our normalized data into target and data.
2. Splitted the data into x_train, x_test, y_train, y_test where the training data is 80% of the original dataset and test is 20% of the original dataset
3. Our first training model is the Logistics Regression Model.
4. We instantiated the model
5. We fit the model on the 80% training dataset
6. We predicted with the 20% test dataset
7. We evaluated our logistics regression model and we found out that:

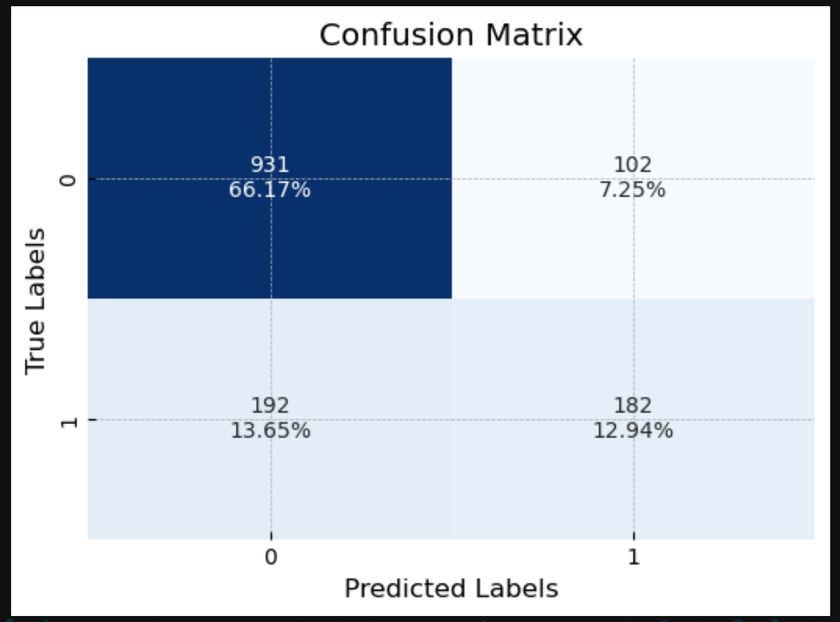
MODEL EVALUATION METRICS

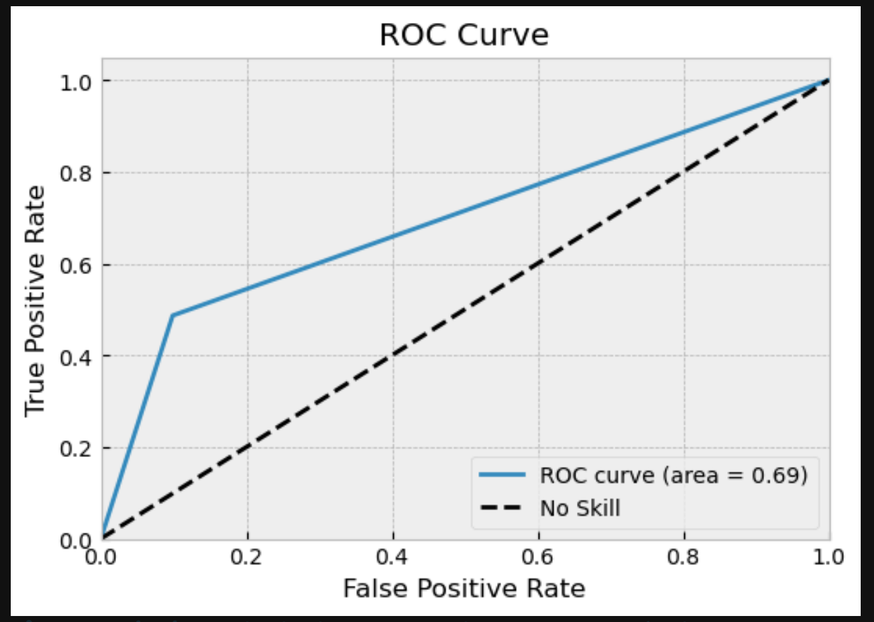
METRIC SCORES

FROM THE LOGISTICS REGRESSION MODEL,

	precision	recall	f1-score	support
0	0.83	0.90	0.86	1033
1	0.64	0.49	0.55	374
accuracy			0.79	1407
macro avg	0.73	0.69	0.71	1407
weighted avg	0.78	0.79	0.78	1407

CONFUSION MATRIX

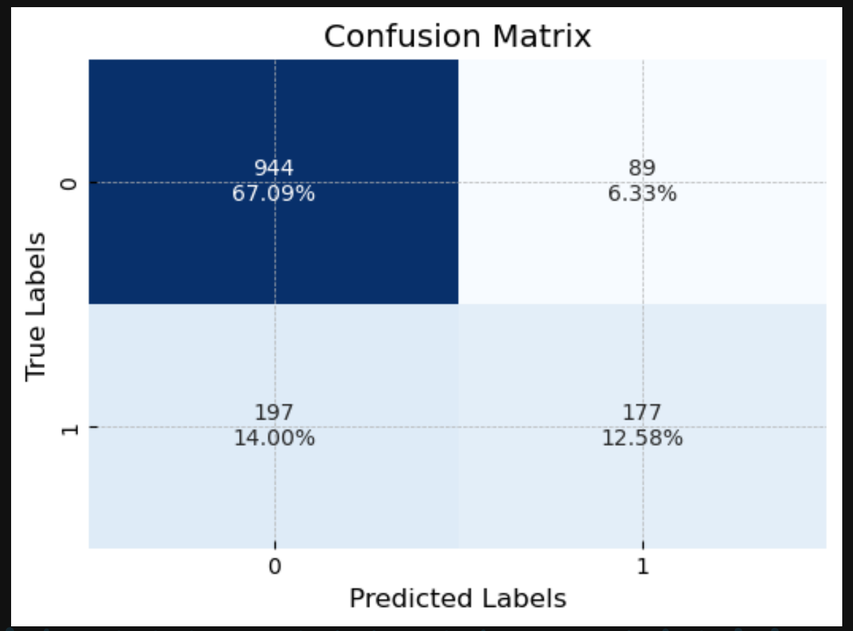


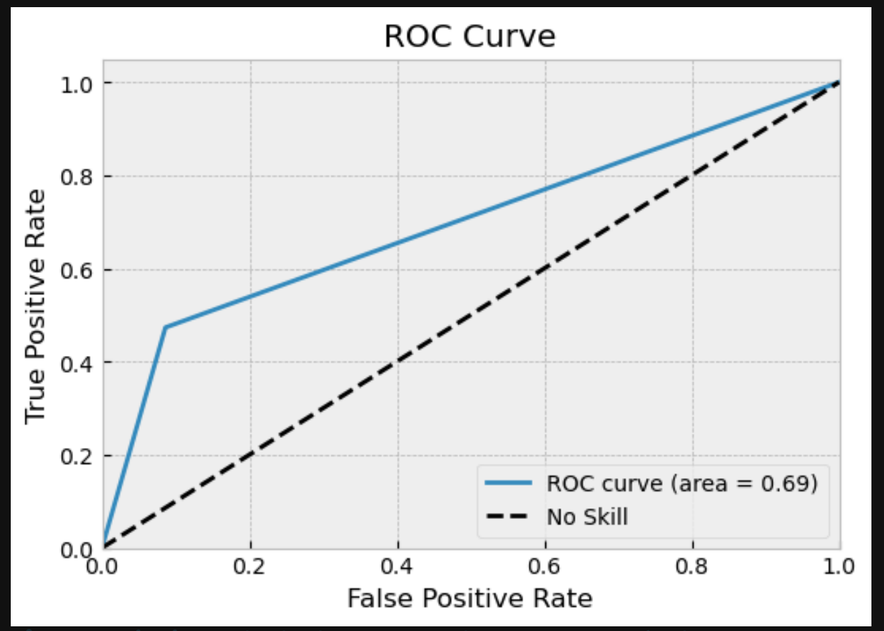


- True Postive is 931 at 66.17%
- False Positive is 102 at 7.25%
- False Negative is 192 at 13.65%
- True Negative is 182 at 12.94%
- ROC Curve fell at 0.69 which is a fair prediction model

A 0.69 ROC Curve is decent, so we opted to perform Hyper Parameter Optimization and Re-training to enhance our training performance.

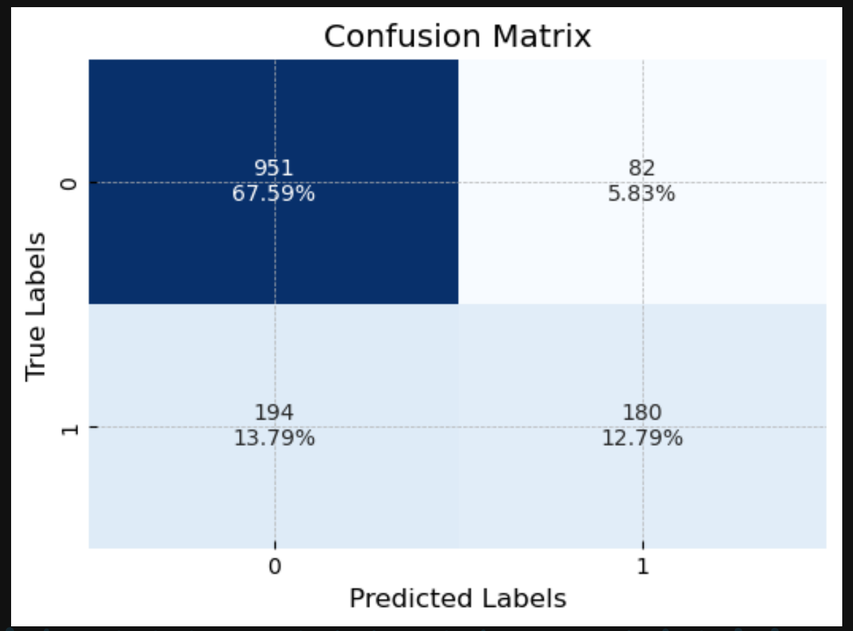
HYPER PARAMETER OPT.

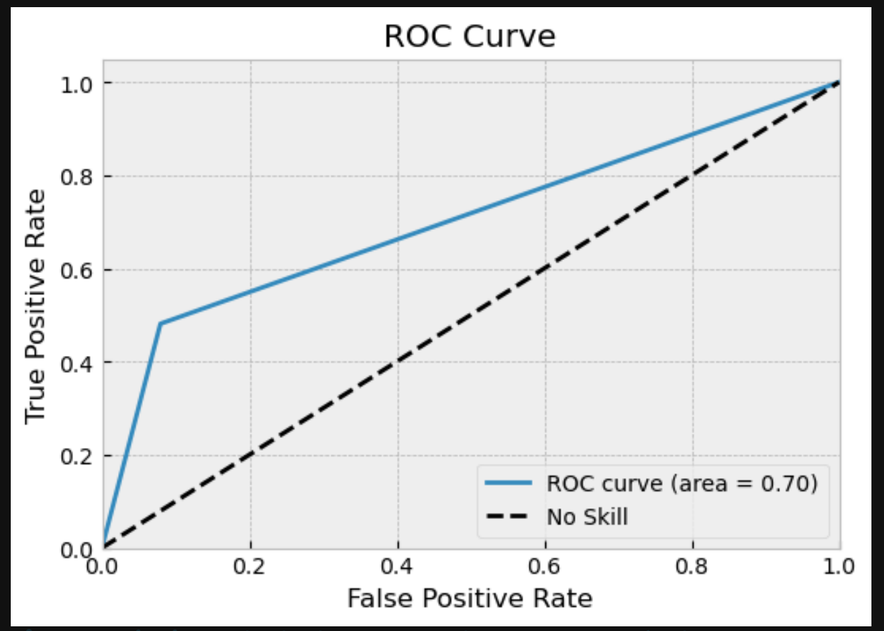




- True Postive is 944 at 67.09%
- False Positive is 89 at 6.33%
- False Negative is 197 at 14.00%
- True Negative is 177 at 12.58%
- ROC Curve is still at 0.69 which is a fair prediction model

RE-TRAINING THE MODEL





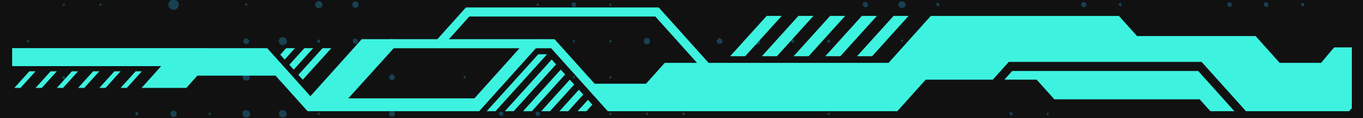
- True Postive is 951 at 67.59%
- False Positive is 82 at 5.83%
- False Negative is 194 at 13.79%
- True Negative is 180 at 12.79%
- ROC Curve increased to 0.70 which is still a fair prediction.

After retraining the logistic regression model, we observed an improvement in the ROC Curve to 0.70. Nevertheless, this is considered moderate, thus we should further enhance feature engineering and eliminate any outliers from the bivariate analysis.

MACHINE LEARNING MODEL

OTHER MODELS

TASKS



- XGBCLASSIFIER
- RANDOMFORESTCLASSIFIER
- KNEIGHBORSCLASSIFIER
- SGDCLASSIFIER
- SVC
- GAUSSIANNB
- DECISIONTREECLASSIFIER

MODEL EVALUATION METRICS

OTHER MODELS

METRIC SCORES

Accuracy Score

	XGB Classifier	Random Forest	K-Nearest Neighbours	SGD Classifier	SVC	Naive Bayes	Decision tree	Logistic Regression
0	78.25%	78.68%	74.56%	74.56%	78.96%	66.52%	71.22%	79.1%

Precision

	XGB Classifier	Random Forest	K-Nearest Neighbours	SGD Classifier	SVC	Naive Bayes	Decision tree	Logistic Regression
0	60.3%	63.03%	52.23%	51.9%	65.0%	43.44%	46.25%	64.08%

METRIC SCORES

Recall

	XGB Classifier	Random Forest	K-Nearest Neighbours	SGD Classifier	SVC	Naive Bayes	Decision tree	Logistic Regression
0	53.21%	47.86%	50.0%	58.56%	45.19%	85.83%	51.07%	48.66%

Receiver Operating Characteristics

	XGB Classifier	Random Forest	K-Nearest Neighbours	SGD Classifier	SVC	Naive Bayes	Decision tree	Logistic Regression
0	70.26%	68.85%	66.72%	69.45%	68.19%	72.68%	64.79%	69.39%

OBSERVATION

1. - The best model accuracy score is the logistics regression at 79.1%
2. - The best model precision score is Logistics Regression at 64.08%
3. - The best model recall score is Naive Bayes at 85.83%
4. - The best model ROC score is Naive Bayes at 72.68%

BUSINESS IMPACT

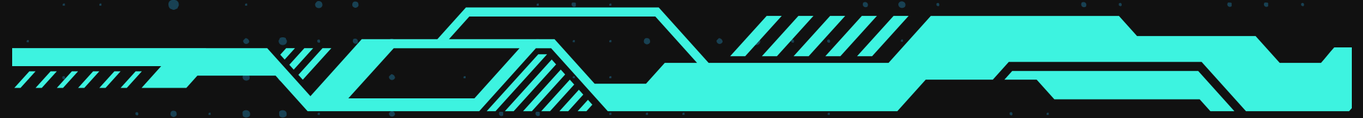
OBSERVATION

WHAT METRICS ARE MOST IMPORTANT FOR THE PROBLEM?

Connecttel needs to focus more on reducing the False Negatives.

NEXT STEPS

DEPLOYMENT



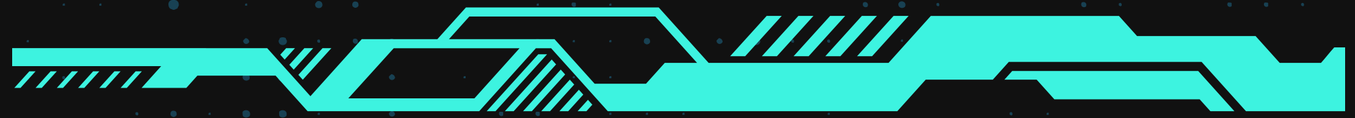
1. Prepare the model ensuring it is trained and validated
2. Choose deployment method eg API, Embedded or Batch Processing
3. Containerize the Model
4. Deploy to a Server for accessing
5. Monitor and maintain
6. Ensure deployment meets security and compliance requirements
7. Documentation and Testing

CONCLUSION

1. We've been able to identify customers that are likely to churn
2. We've also developed a robust prediction system

STAR RECAP

SITUATION



A dataset was provided containing both numerical and categorical features with missing values. The goal was to preprocess the data, perform cleaning, handle missing values, encode categorical features to numerical, encode numerical features to categorical, create new features, and analyze the data through exploratory data analysis (EDA) and then predict using at least 3 training models.

TASK

1. Preprocess and cleaned the dataset.
2. Handle missing values and encoded categorical features to numerical.
3. Encode numerical features to categorical where necessary.
4. Create new features to enhance the dataset.
5. Plot the feature importance chart to understand key predictors.
6. Perform EDA including univariate, bivariate, and multivariate analysis.
7. Separate the dataset into features and target label.
8. Split the data into training and testing sets.
9. Train a logistic regression model on the dataset.
10. Optimize hyperparameters to improve the model performance.
11. Retrain the logistic regression model with the optimized hyperparameters.
12. Train with 7 other supervised machine learning models

ACTION

1. Conducted preprocessing and cleaning of the dataset.
2. Handled missing values and encoded categorical features using techniques like one-hot encoding.
3. Encoded numerical features to categorical using binning or other methods.
4. Created new features based on domain knowledge or feature engineering.
5. Plotted the feature importance chart using techniques like feature importance scores.ed hyperparameters.

ACTION

1. Conducted EDA including univariate, bivariate, and multivariate analysis to explore relationships within the data.
2. Separated the dataset into features (X) and target label (y).
3. Split the data into training and testing sets using techniques like train-test split.
4. Trained a logistic regression model on the training data.
5. Used hyperparameter optimization techniques like GridSearchCV or RandomizedSearchCV to find the best parameters for the logistic regression model.
6. Retrained the logistic regression model using the optimized hyperparameters.

RESULT

The dataset was successfully preprocessed and cleaned, missing values were handled, categorical and numerical features were encoded appropriately, new features were created, and the dataset was analyzed through EDA. A logistic regression model was trained, optimized using hyperparameters, and retrained to improve its performance. The final model is ready for evaluation and deployment with 0.70 ROC Curve.

THE END

PREDICTION BY:

UCHE
SAMUEL

VIEW PRESENTATION