Analysis Report

-Aman Ali

Feature Engineering

- Average Monthly Data Usage:
    - Formula: Average Monthly Data Usage = Total Usage in GB / Subscription Length in Months
    - Description: This feature calculates the average data usage per month by dividing the total data usage in gigabytes (GB) by the length of the subscription in months.
- Billing Change Rate:
    - Formula: Billing Change Rate = Monthly Bill - Previous Month's Monthly Bill
    - Description: The billing change rate calculates the difference in the monthly bill from the current month to the previous month, indicating how the monthly bill has changed.
- Billing Amount as a Percentage:
    - Formula: Billing Amount as Percentage = (Monthly Bill / Mean Monthly Bill) * 100
    - Description: This feature calculates the monthly bill as a percentage of the mean (average) monthly bill. It provides insight into how a customer's bill compares to the average.
- Customer Tenure in Months:
    - Formula: Customer Tenure in Months = Subscription Length in Months
    - Description: This feature represents the tenure or duration of the customer's subscription in months. It is equal to the length of the subscription.
- Churn History:
    - Formula: Churn History = Lagged Churn Column (Previous Month's Churn Status)

    - Description: Churn history is a lagged version of the churn column, indicating whether the customer churned (canceled their subscription) in the previous month.
- Age Group Indicator:
    - Formula: Age Group Indicator = Categorization based on Age Bins
    - Description: Age group indicator categorizes customers into different age groups such as "Young," "Middle-Aged," or "Senior" based on predefined age bins. Customers' ages determine their respective groups.
- Remaining Subscription Length:
    - Formula: Remaining Subscription Length = Subscription Length in Months - Current Month
    - Description: This feature calculates the remaining duration of the customer's subscription in months. It represents how many months are left in the subscription.
- Average Bill Change:
    - Formula: Average Bill Change = Rolling Mean of Billing Change Rate (Last 3 Months)
    - Description: Average bill change calculates the mean of the billing change rate over the last three months. It indicates the average rate of change in the monthly bill.

These formulas provide a clear understanding of how each feature is computed and its significance in the dataset for further analysis or modeling.

Feature importances

- **Remaining_Subscription_Length (0.1219):** This feature has the highest importance score. It suggests that the remaining subscription length is a significant predictor of customer churn. Customers with shorter remaining subscription lengths might be more likely to churn.

- **Billing_Change_Rate (0.1208):** Billing change rate is the second most important feature. It indicates that fluctuations in the monthly billing amount play a crucial role in predicting churn. Higher billing change rates might be associated with a higher likelihood of churn.

- **Average_Bill_Change (0.1208):** This feature is very similar in importance to Billing_Change_Rate. It also captures billing fluctuations but might represent a smoothed or averaged version of those changes.

- **Average_Monthly_Data_Usage (0.1096):** The average monthly data usage is the fourth most important feature. It suggests that customers' data usage patterns can impact their likelihood of churning. Higher data usage may correlate with lower churn rates.

- **Total_Usage_GB (0.1068):** This feature represents the total data usage in gigabytes and is also important. It's related to Average_Monthly_Data_Usage but considers the overall data consumption.

- **Billing_As_Percentage (0.1046):** The percentage of the monthly bill relative to some reference value is the sixth most important feature. This might capture information about customers' budget constraints or sensitivity to price changes.

- **Monthly_Bill (0.1045):** The actual monthly billing amount is an important feature. High monthly bills could be a factor leading to churn.

- **Age (0.0863):** Age is also significant but less so than some of the billing-related features. It indicates that customer age plays a role in predicting churn, with younger or older customers potentially having different behaviors.

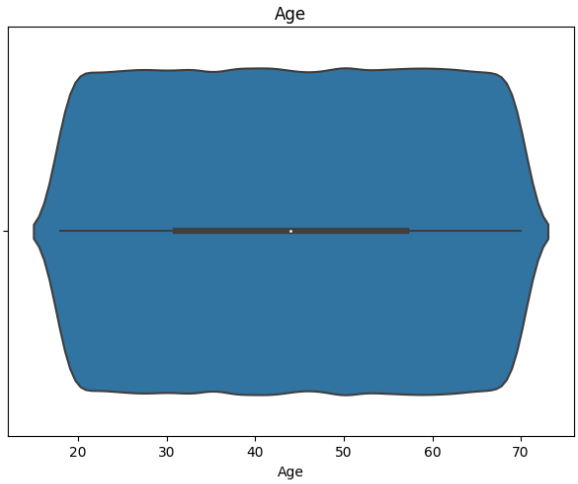
- **Subscription_Length_Months (0.0481):** Subscription length in months is less important compared to other features. However, it still contributes to the model's predictions, suggesting that longer subscription commitments might reduce churn.

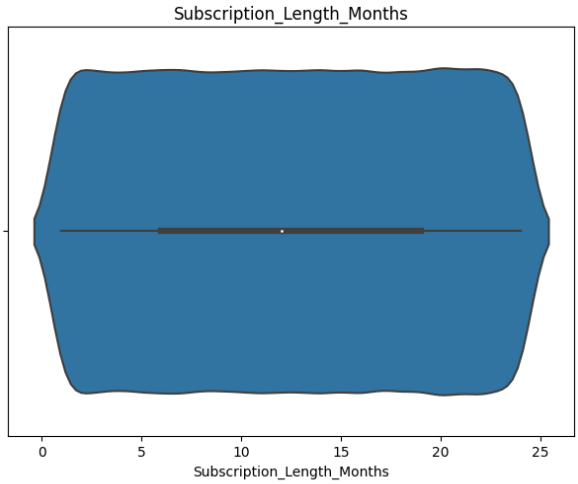
- **Customer_Tenure_Months (0.0476):** Customer tenure, or how long a customer has been with the provider, is also less important but still relevant. Longer tenure might indicate loyalty and reduce churn.

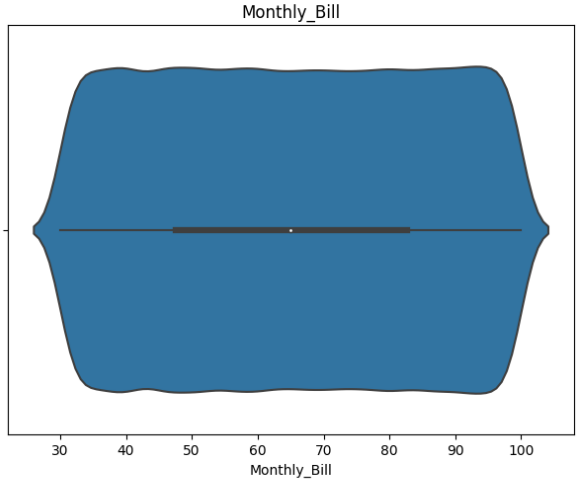
- **Churn_History (0.0159):** Churn history, which captures whether a customer has churned in the past, has some importance. It's less critical than the other features, indicating that current behavior and characteristics are more predictive.

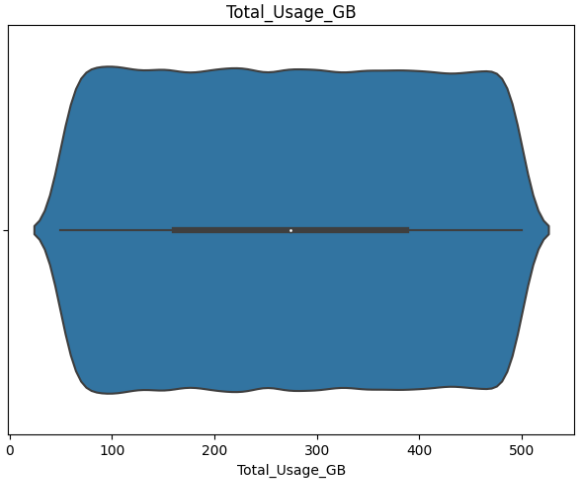
- **Age_Group_Indicator (0.0131):** Age group indicator has the lowest importance in your model. It suggests that the specific age group category might not be as crucial as the overall age feature.

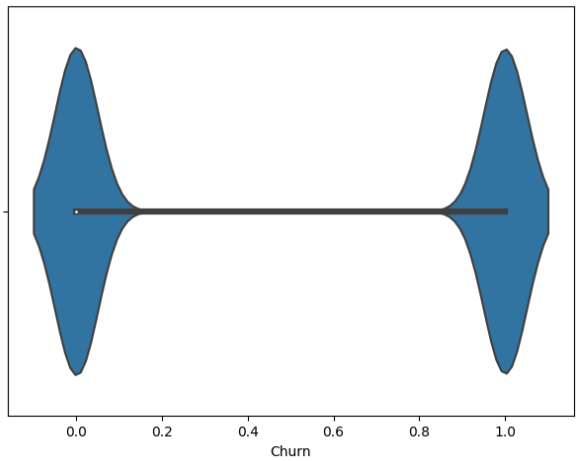
Visualisations

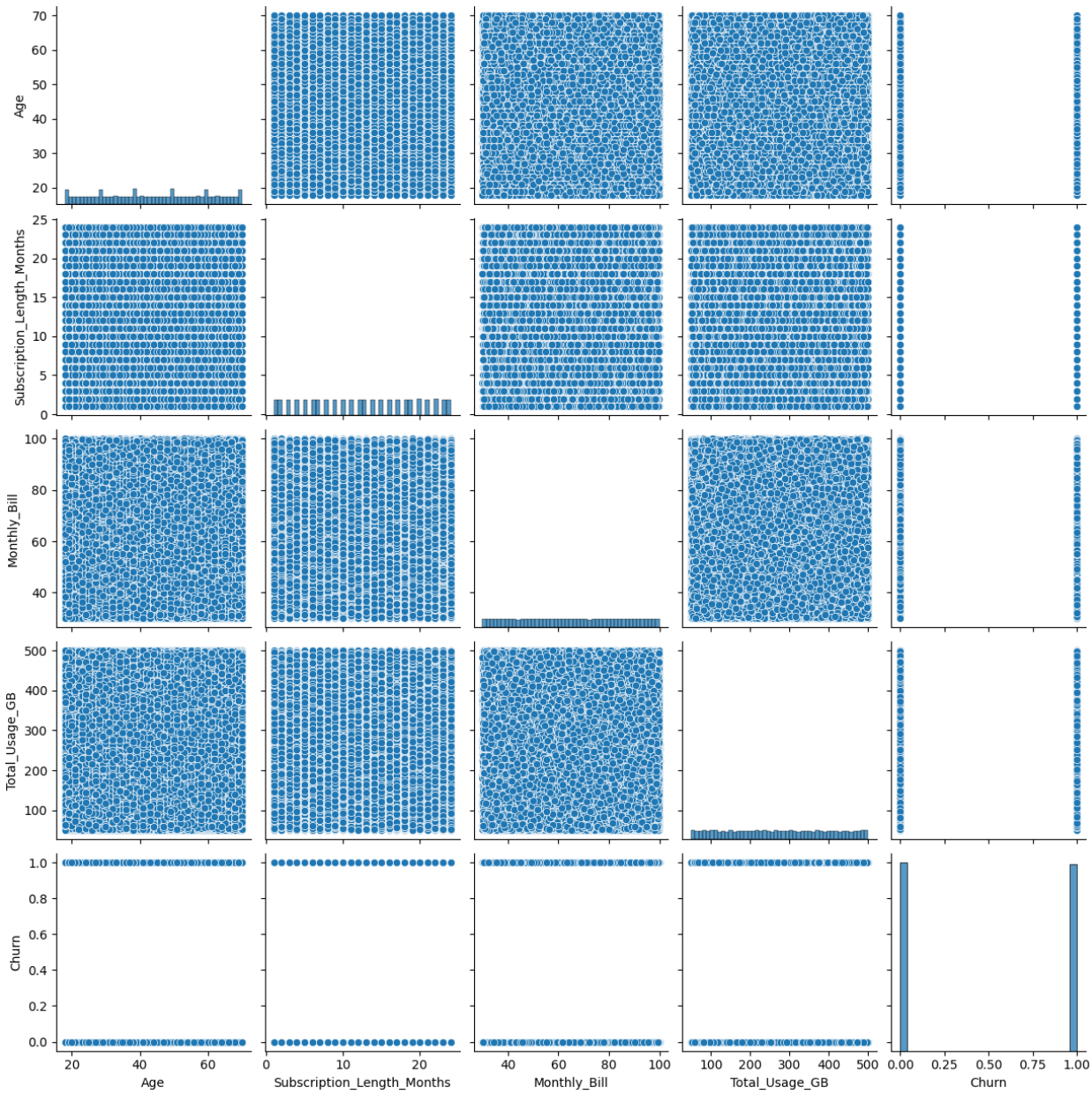


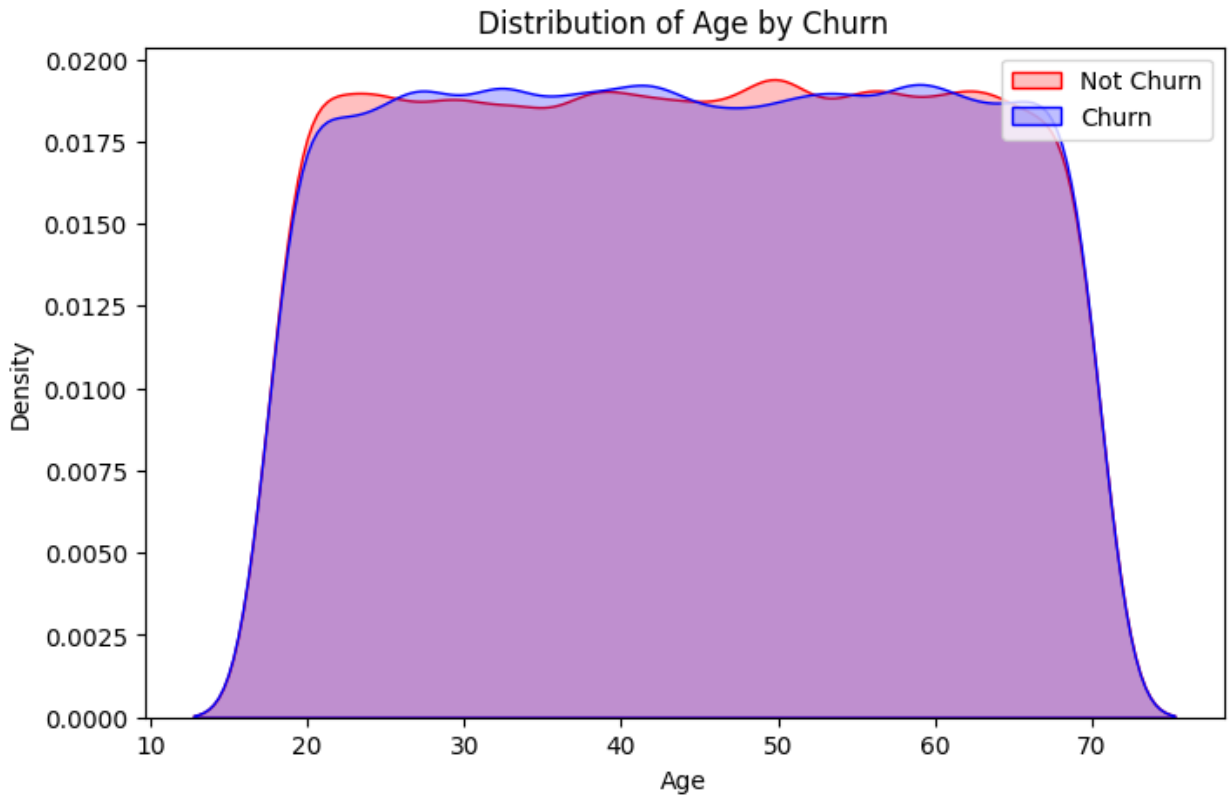


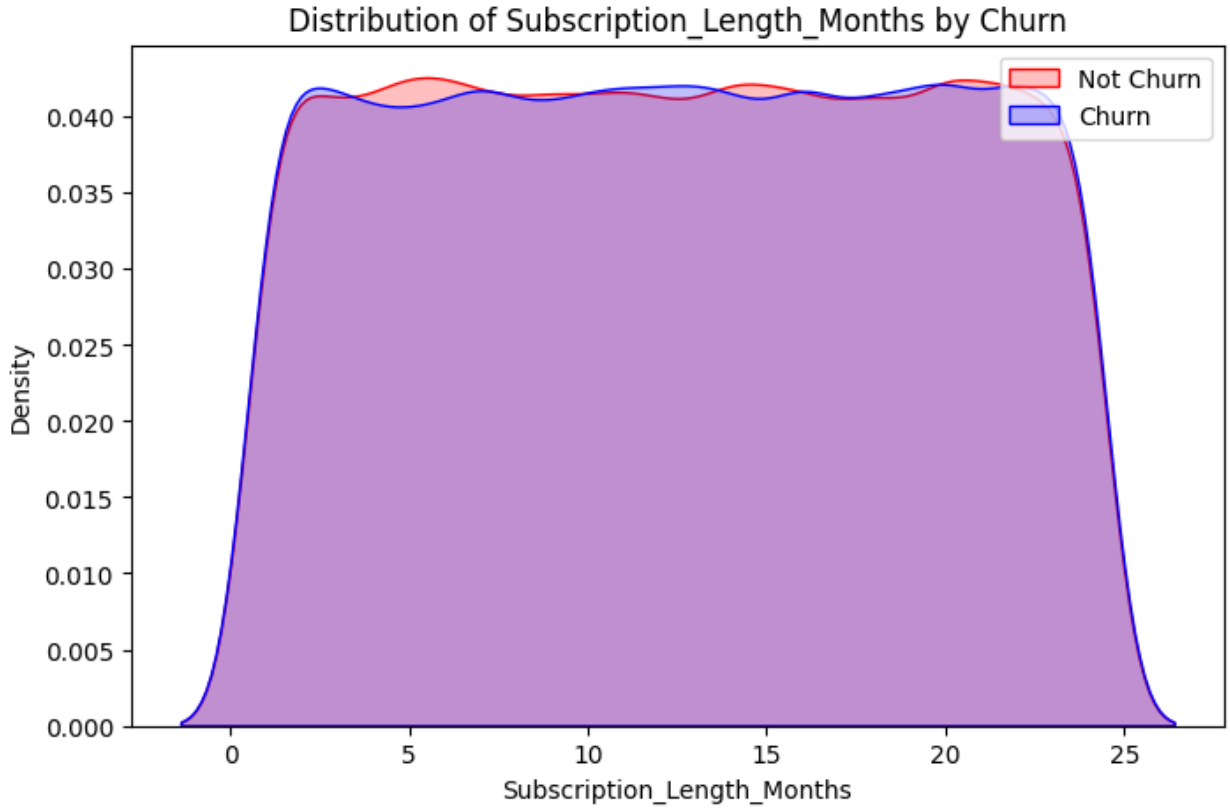


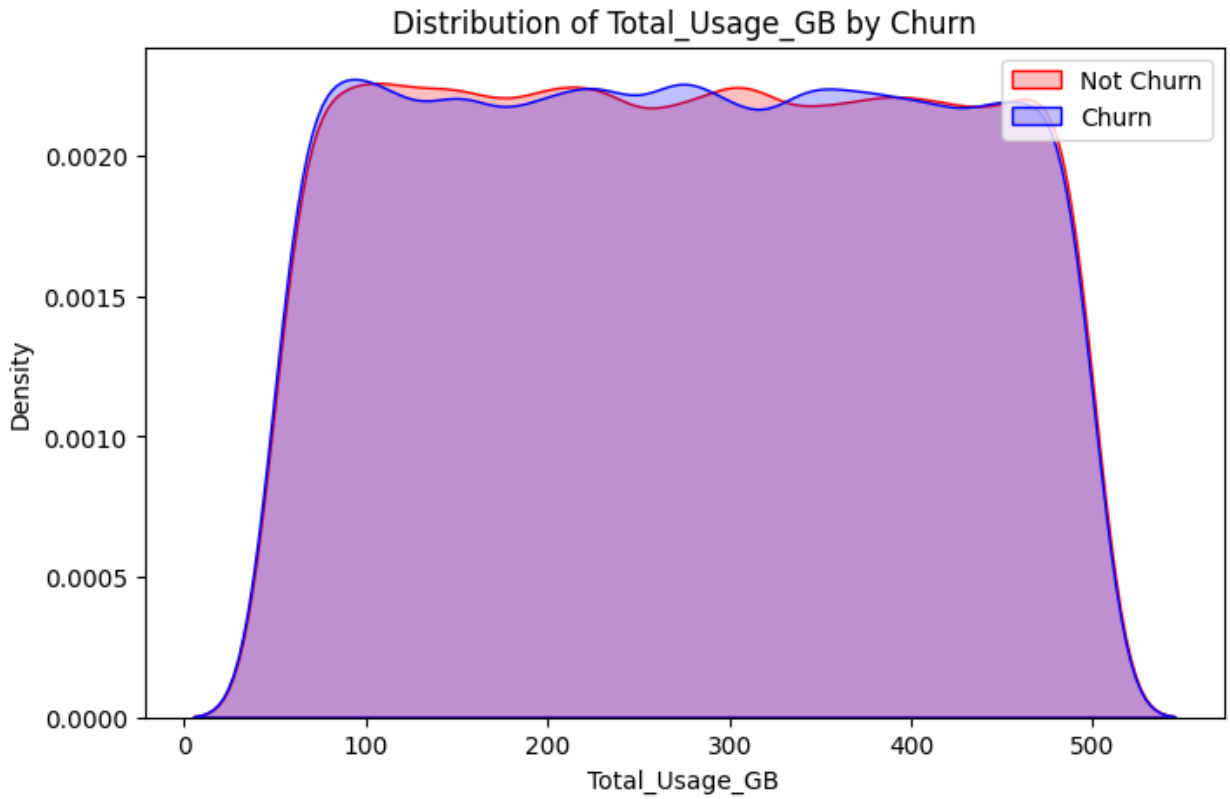


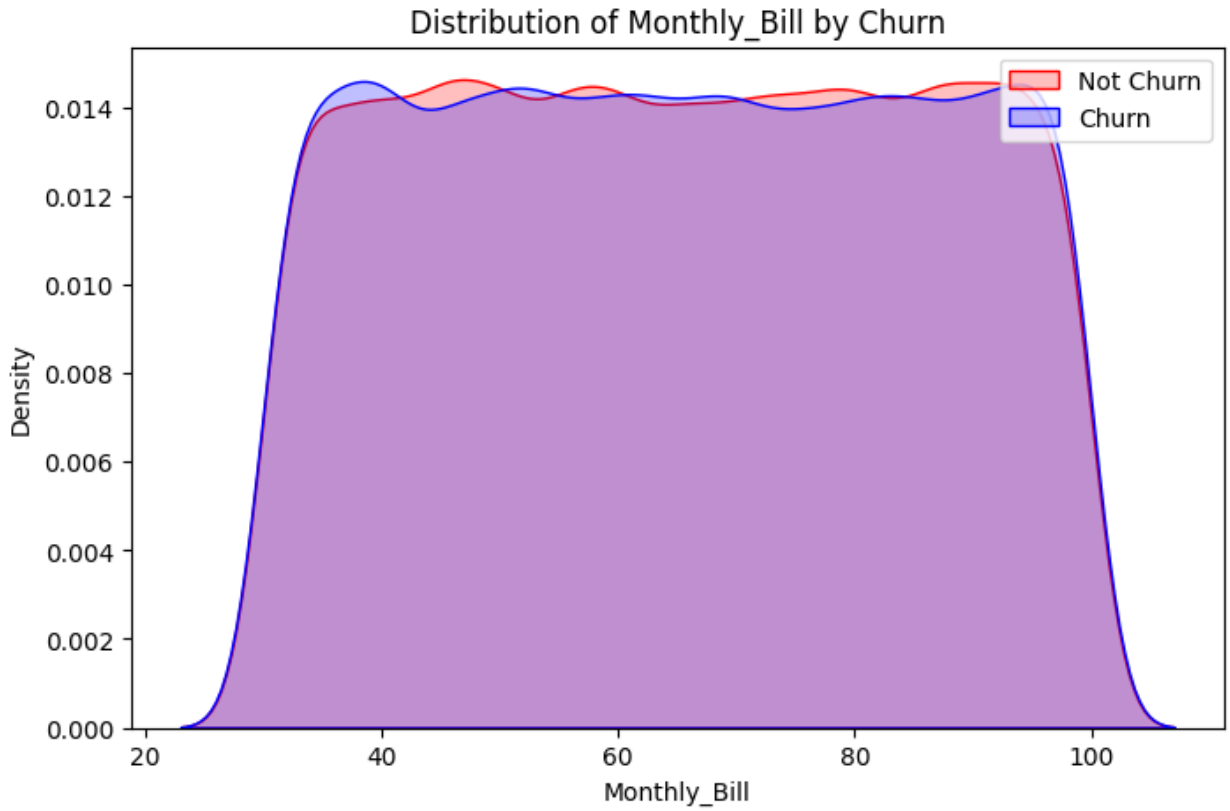


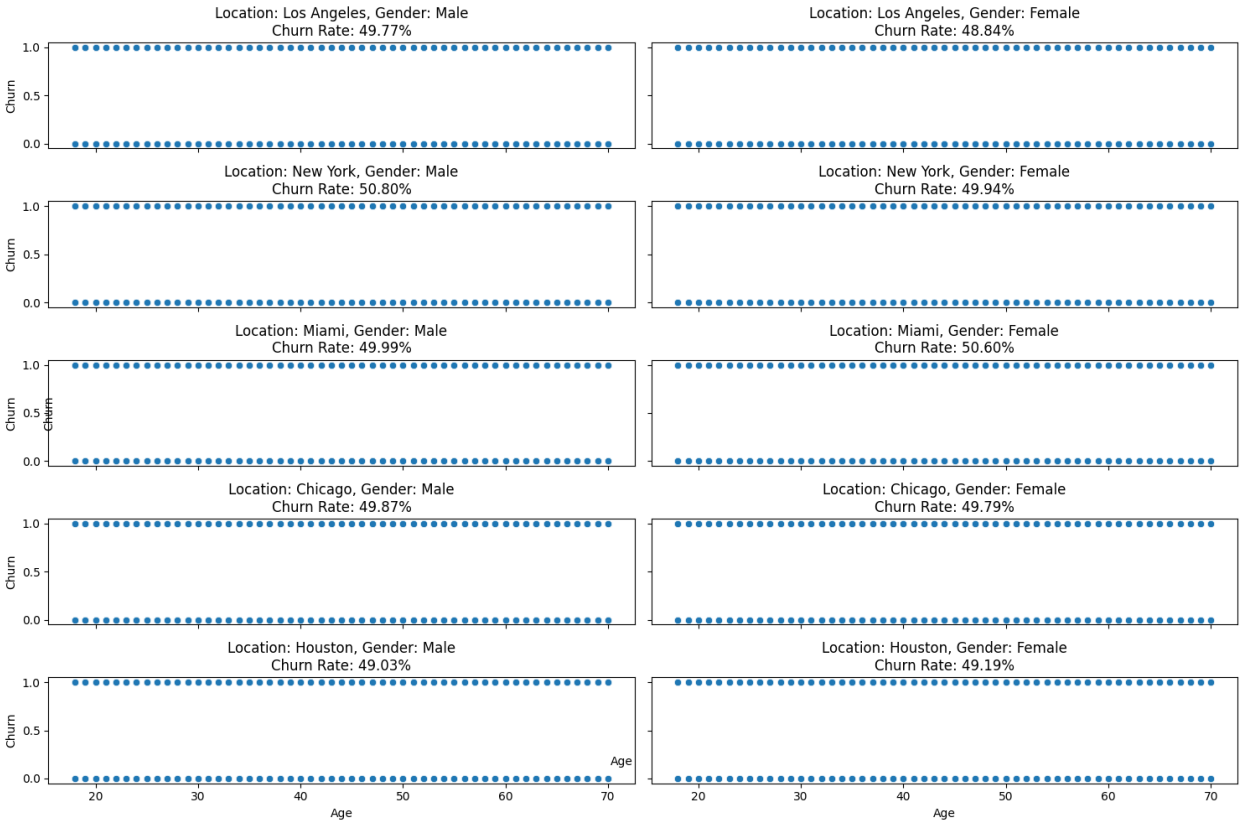

Age
Subscription_Length_Months
Monthly_Bill
Total_Usage_GB
Churn
Age
Subscription_Length_Months
Monthly_Bill
Total_Usage_GB
Churn

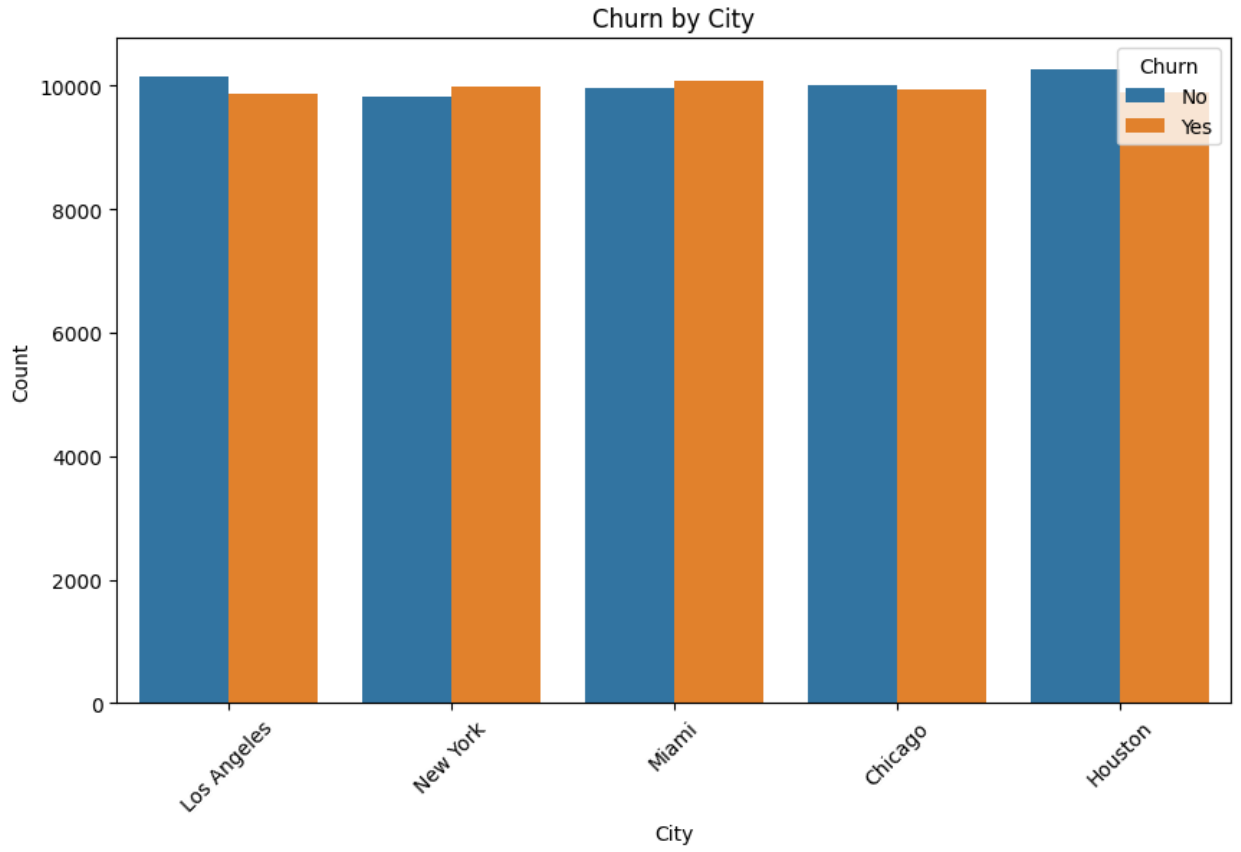

Distribution of Age by Churn
Not Churn
Churn
Density
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
10
20
30
40
50
60
70
Age

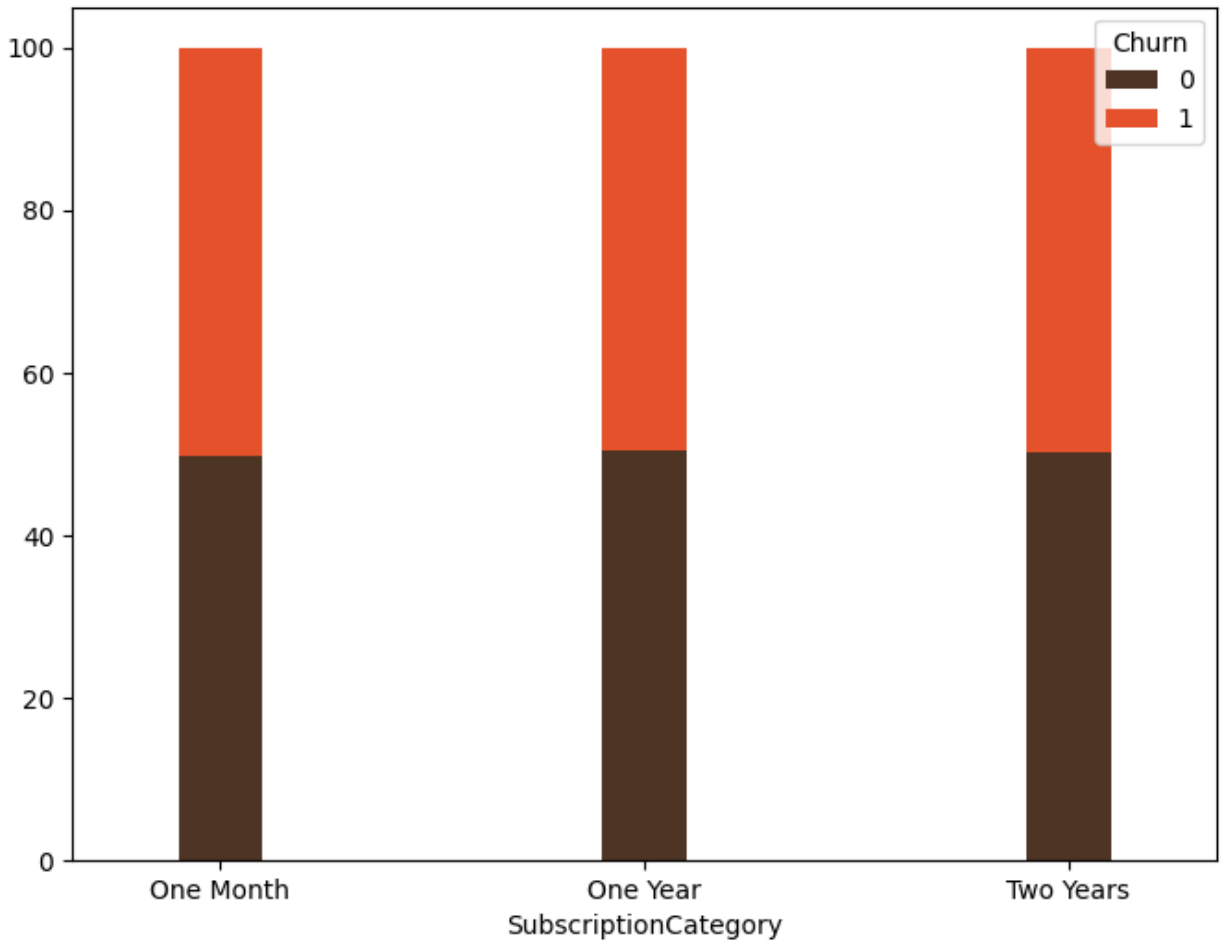

Distribution of Subscription_Length_Months by Churn
Not Churn
Churn
Density
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
5
10
15
20
25
Subscription_Length_Months

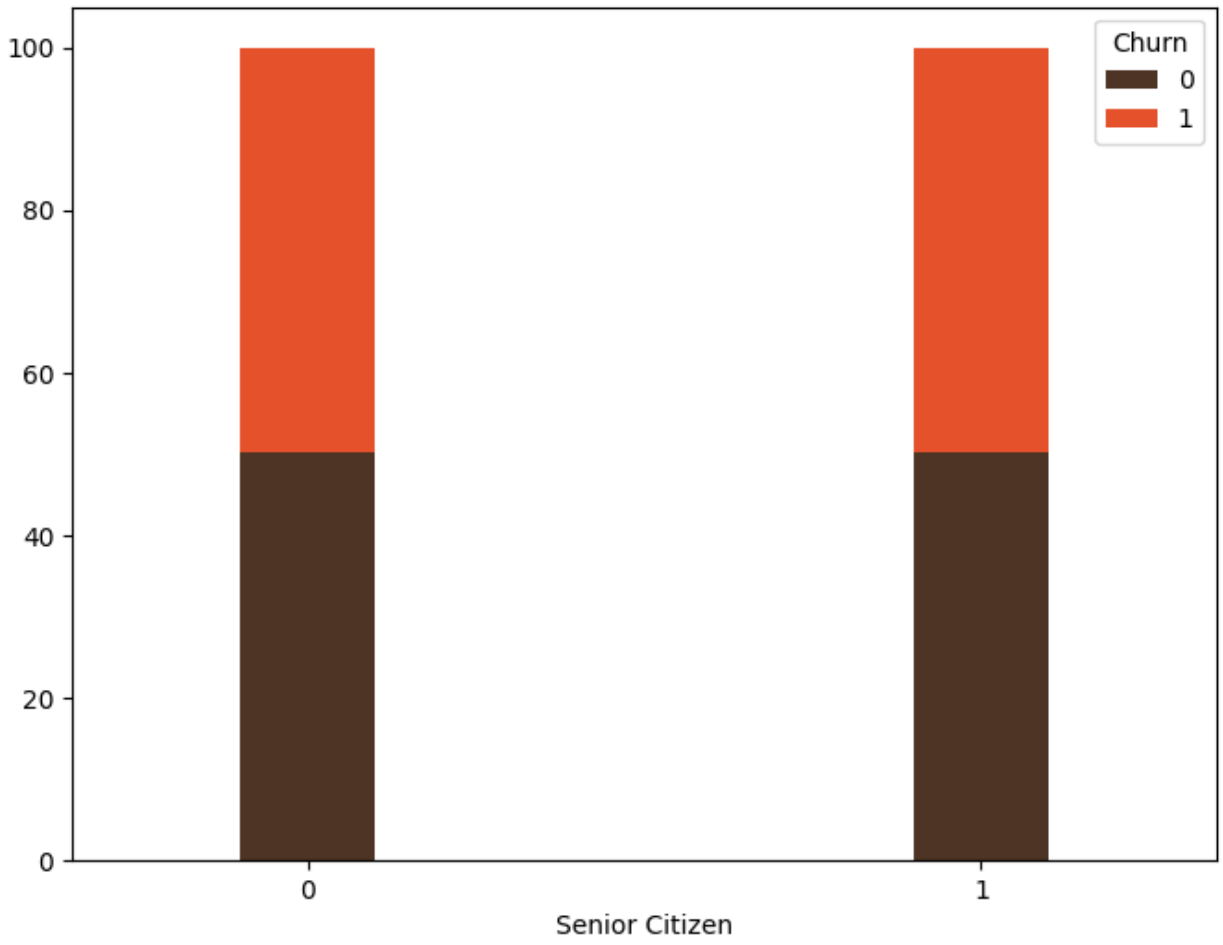

Distribution of Total_Usage_GB by Churn
Not Churn
Churn
Density
0.0020
0.0015
0.0010
0.0005
0.0000
0
100
200
300
400
500
Total_Usage_GB

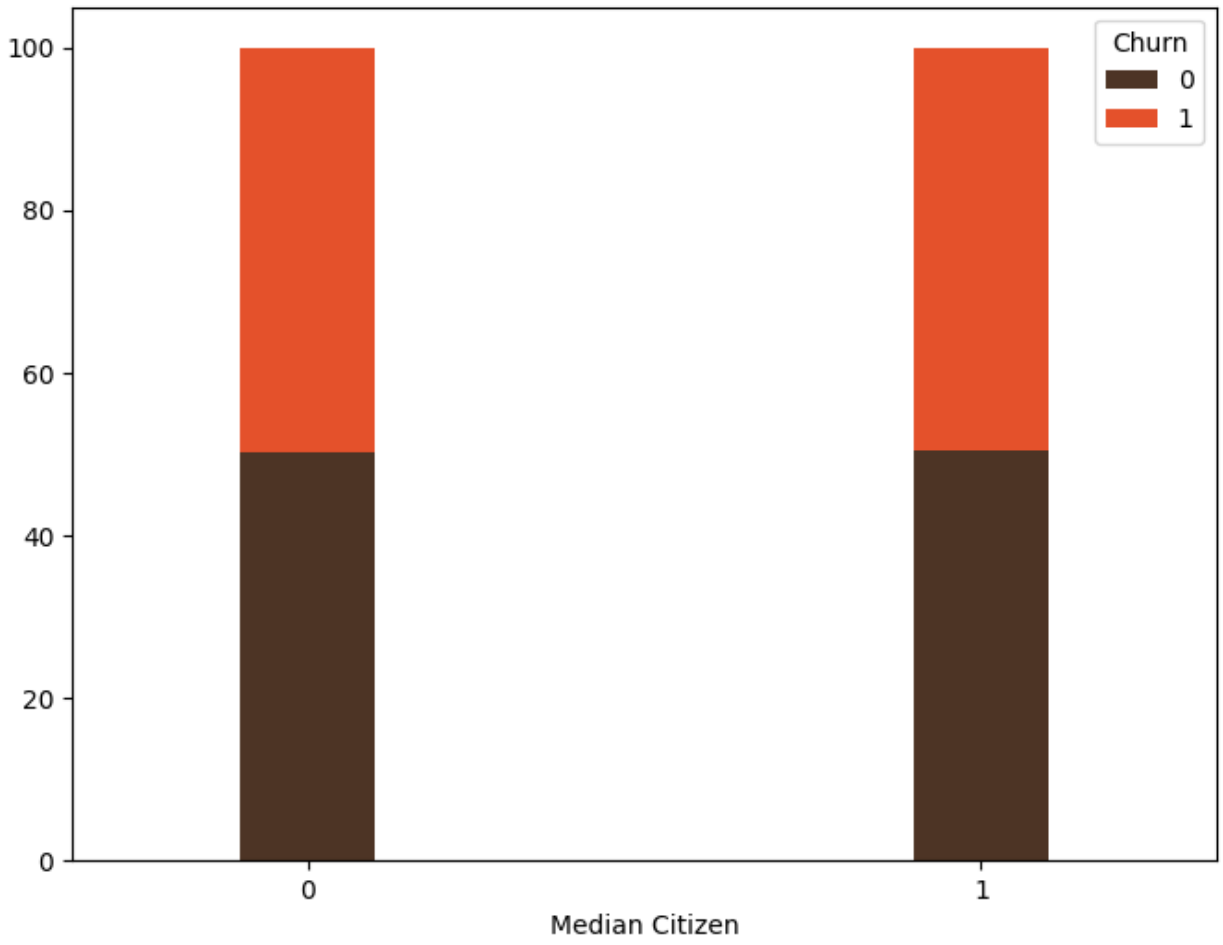

Distribution of Monthly_Bill by Churn
Not Churn
Churn
Density
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
20
40
60
80
100
Monthly_Bill

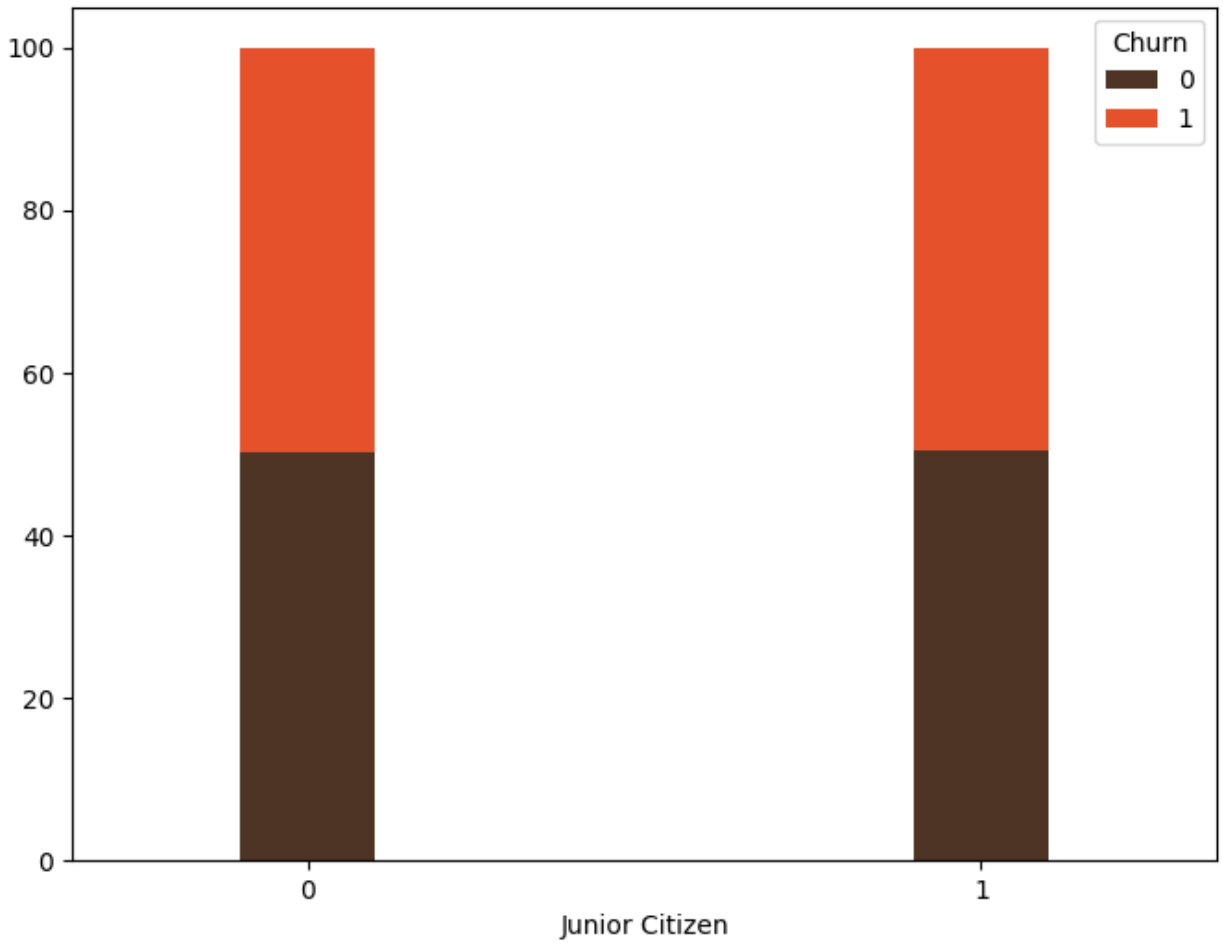

Location: Los Angeles, Gender: Male
Churn Rate: 49.77%
Location: Los Angeles, Gender: Female
Churn Rate: 48.84%
Location: New York, Gender: Male
Churn Rate: 50.80%
Location: New York, Gender: Female
Churn Rate: 49.94%
Location: Miami, Gender: Male
Churn Rate: 49.99%
Location: Miami, Gender: Female
Churn Rate: 50.60%
Location: Chicago, Gender: Male
Churn Rate: 49.87%
Location: Chicago, Gender: Female
Churn Rate: 49.79%
Location: Houston, Gender: Male
Churn Rate: 49.03%
Location: Houston, Gender: Female
Churn Rate: 49.19%
Churn
Age
0.0
0.5
1.0
20
30
40
50
60
70

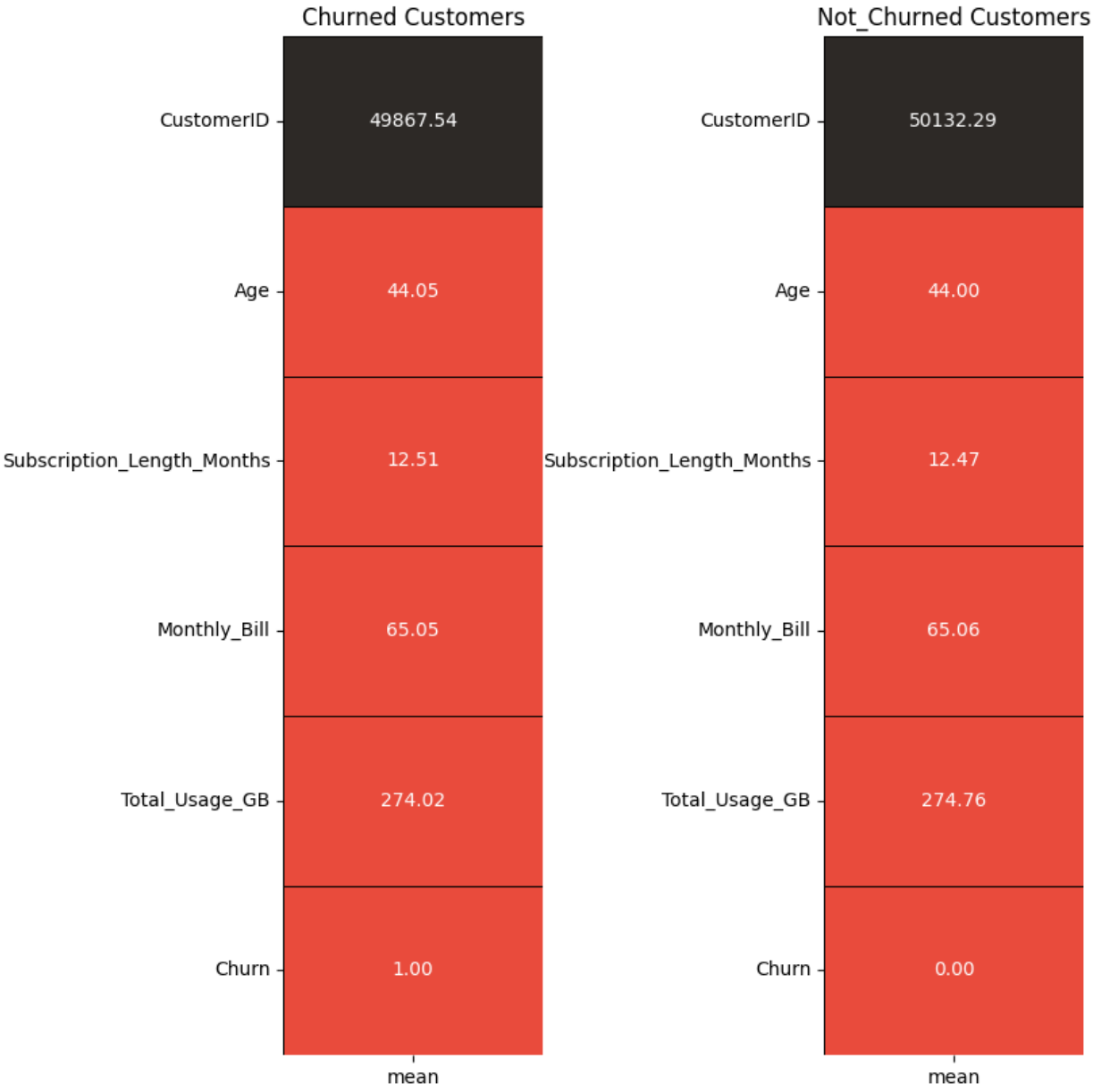

Churn by City
Churn
No
Yes
Count
City
0
2000
4000
6000
8000
10000
Los Angeles
New York
Miami
Chicago
Houston

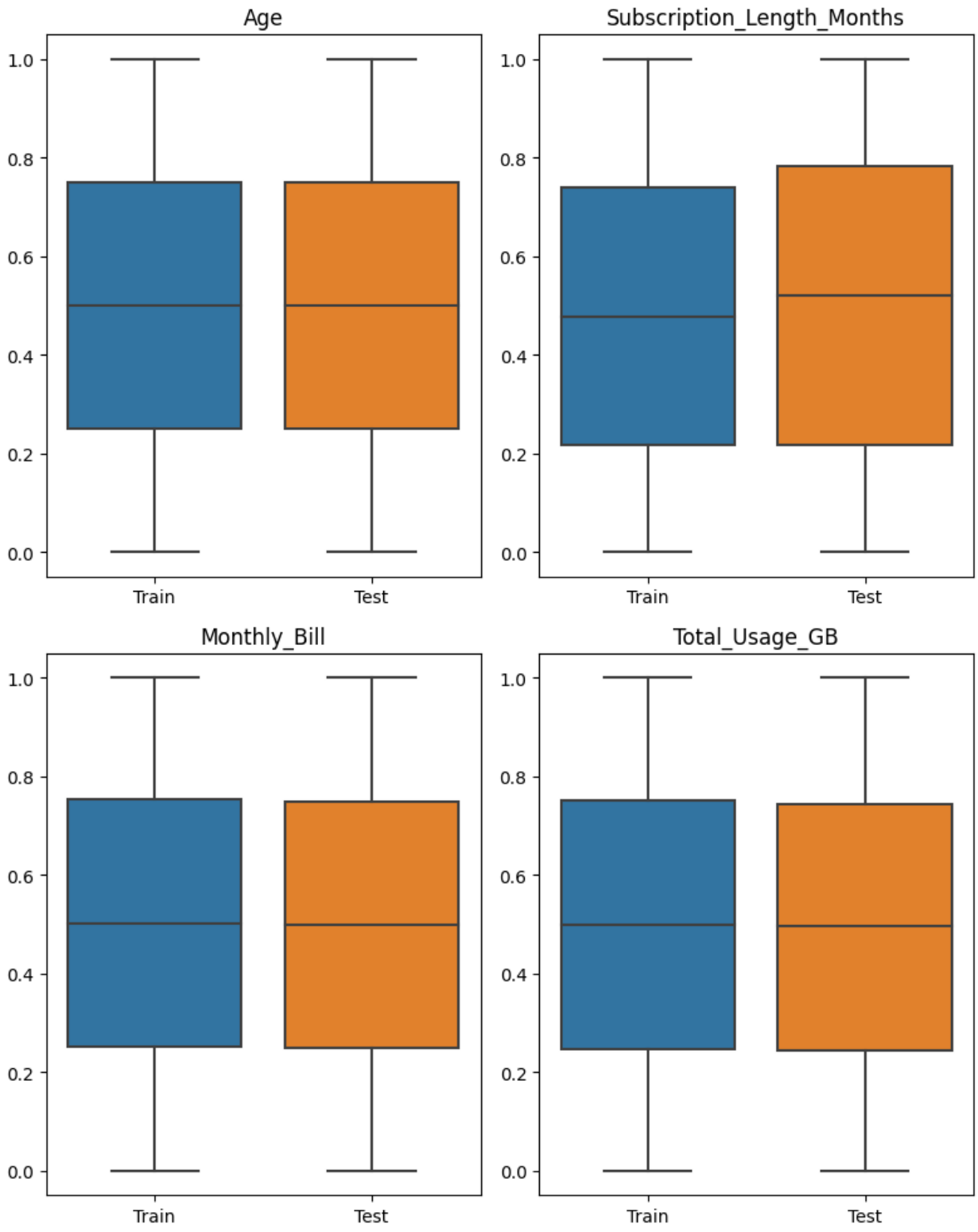

Churn
0
1
100
80
60
40
20
0
One Month
One Year
Two Years
SubscriptionCategory


Churn
0
1
100
80
60
40
20
0
0
1
Senior Citizen

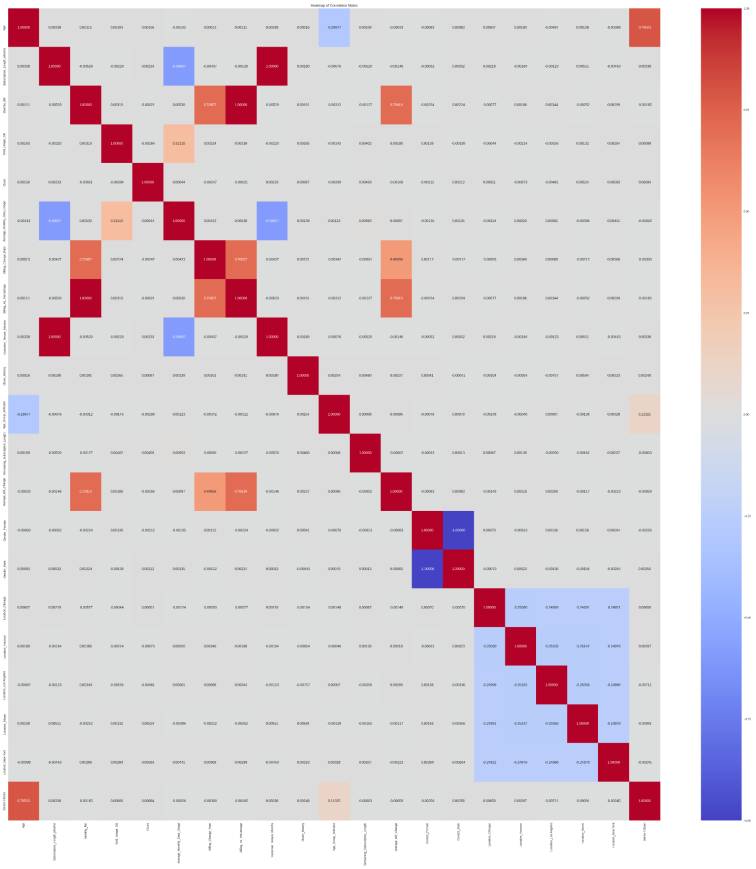

Churn
0
1
100
80
60
40
20
0
0
1
Median Citizen


Churn
0
1
100
80
60
40
20
0
0
1
Junior Citizen

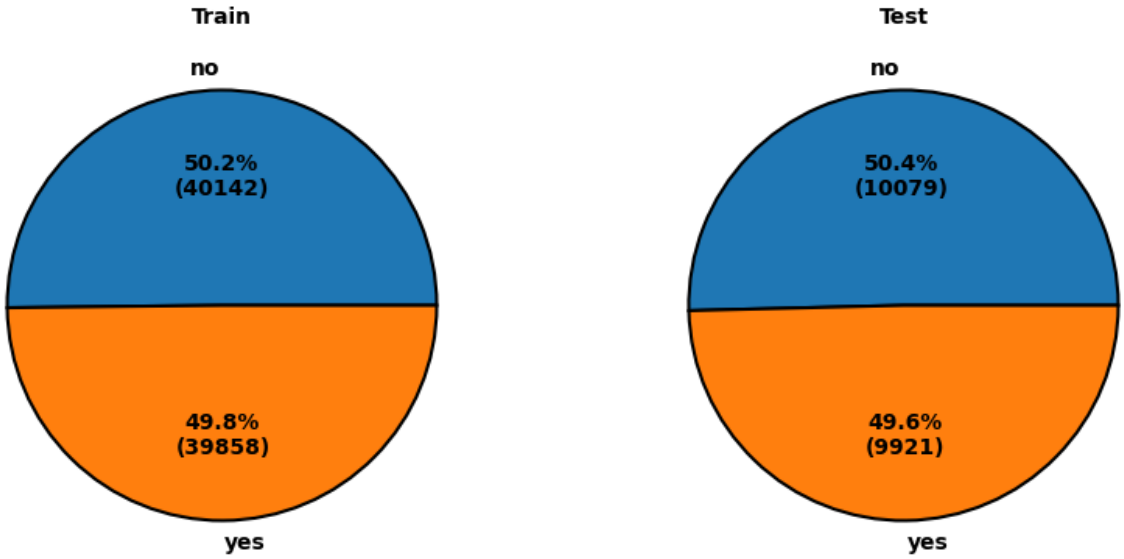

Churned Customers
CustomerID
49867.54
Age
44.05
Subscription_Length_Months
12.51
Monthly_Bill
65.05
Total_Usage_GB
274.02
Churn
1.00
mean
Not_Churned Customers
CustomerID
50132.29
Age
44.00
Subscription_Length_Months
12.47
Monthly_Bill
65.06
Total_Usage_GB
274.76
Churn
0.00
mean

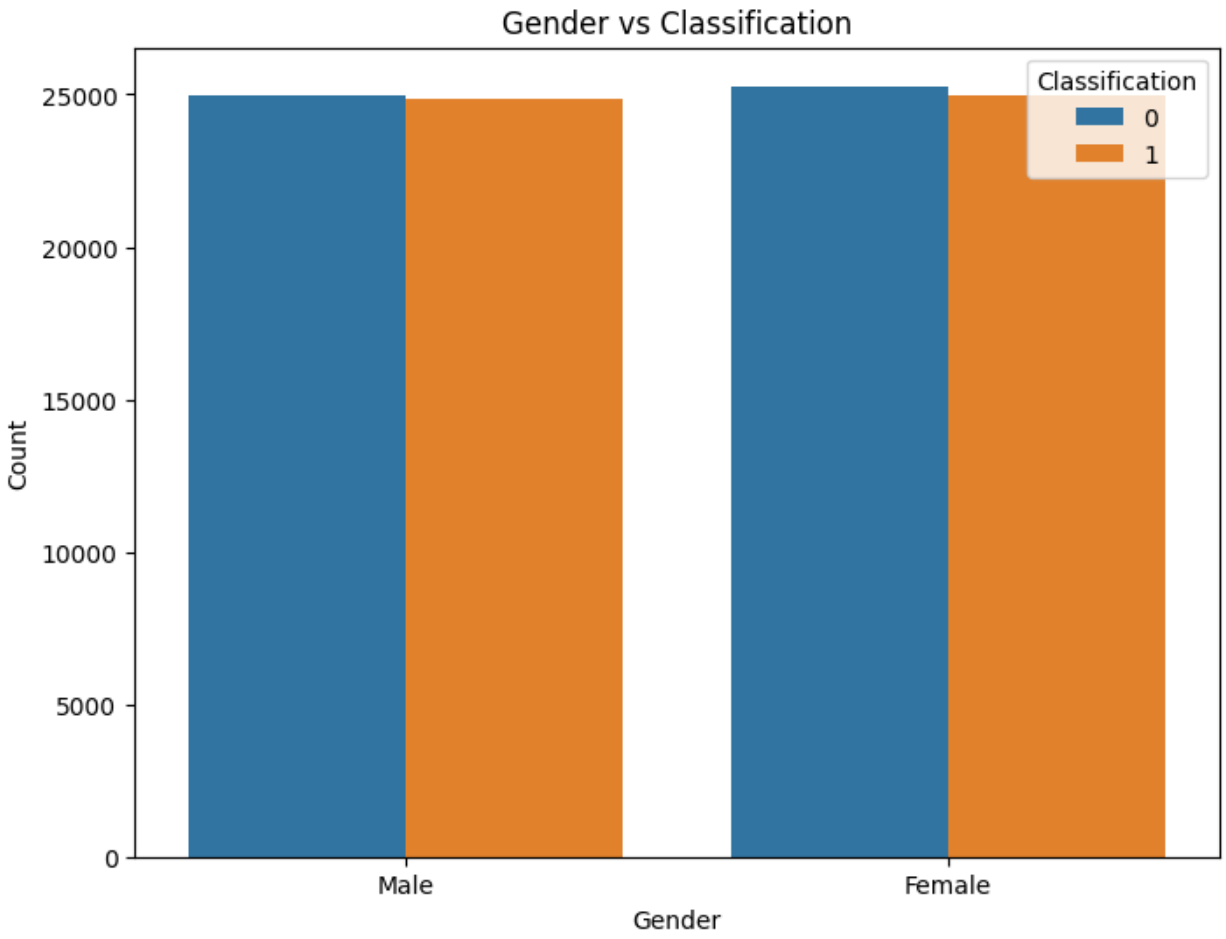

Age
Subscription_Length_Months
Monthly_Bill
Total_Usage_GB
1.0
0.8
0.6
0.4
0.2
0.0
Train
Test

Correlation plot with new features:



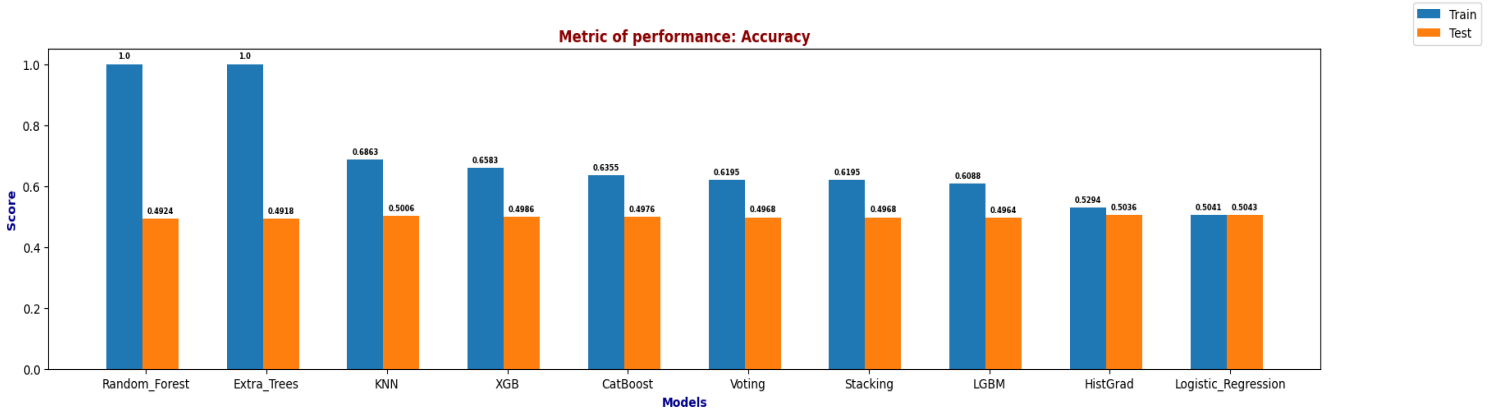
Data distribution of Churn in Train and Test sets




Gender vs Classification
Classification
0
1
25000
20000
15000
10000
5000
0
Count
Male
Female
Gender

	Description	Value
0	Session id	4474
1	Original data shape	(80000, 20)
2	Transformed data shape	(80000, 20)
3	Numeric features	20
4	Rows with missing values	0.0%
5	Preprocess	True
6	Imputation type	simple
7	Numeric imputation	mean
8	Categorical imputation	mode
9	CPU Jobs	-1
10	Use GPU	False
11	Log Experiment	False
12	Experiment Name	anomaly-default-name
13	USI	1f84

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
xgboost	Extreme Gradient Boosting	0.5157	0.5235	0.4977	0.5146	0.5060	0.0312	0.0312
catboost	CatBoost Classifier	0.5136	0.5178	0.4888	0.5126	0.5004	0.0270	0.0270
lightgbm	Light Gradient Boosting Machine	0.5095	0.5136	0.4609	0.5088	0.4836	0.0187	0.0188
rf	Random Forest Classifier	0.5091	0.5115	0.4716	0.5081	0.4892	0.0179	0.0180
gbc	Gradient Boosting Classifier	0.5085	0.5086	0.3930	0.5092	0.4435	0.0162	0.0167
ada	Ada Boost Classifier	0.5035	0.5059	0.4465	0.5023	0.4727	0.0067	0.0067
et	Extra Trees Classifier	0.5024	0.5047	0.4753	0.5009	0.4877	0.0045	0.0046
lr	Logistic Regression	0.5007	0.5011	0.4285	0.4989	0.4610	0.0009	0.0009
mlp	MLP Classifier	0.4982	0.4987	0.4331	0.4960	0.4617	-0.0041	-0.0041



Hence we can look at the visualisations above to conclude that the best model is the XGBoost model and we can say that the data had no outliers and there was no class imbalance observed.