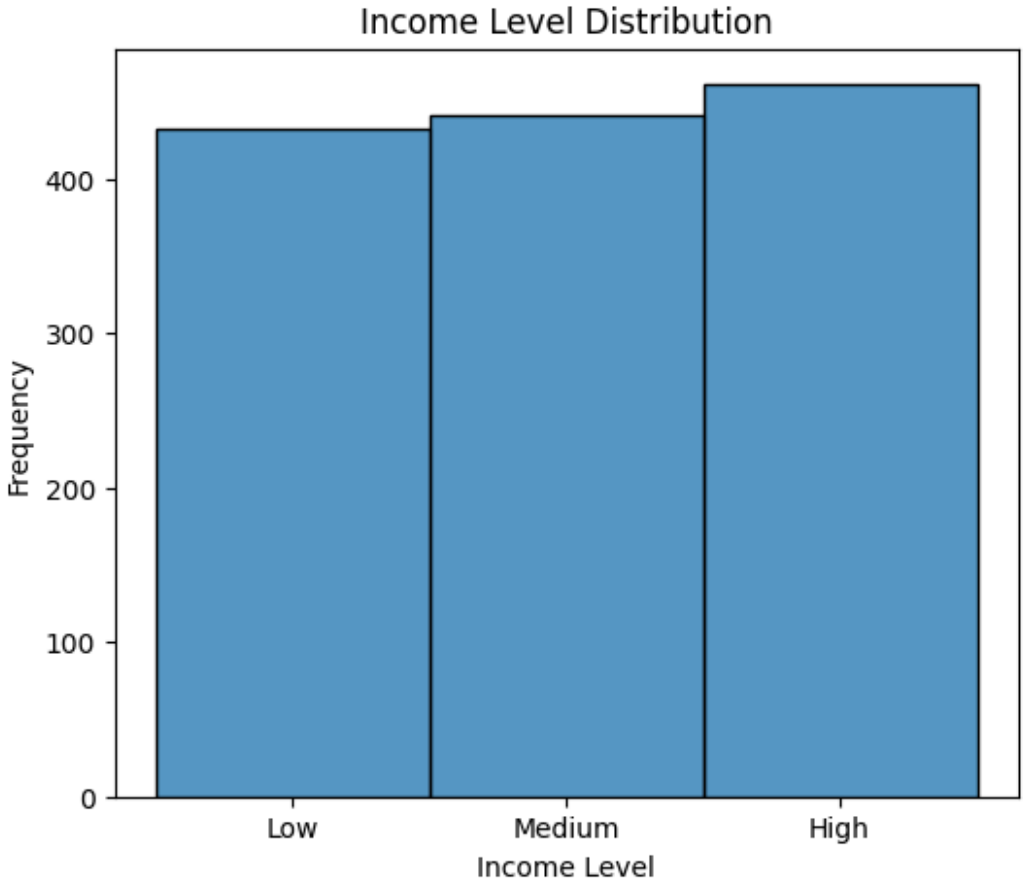
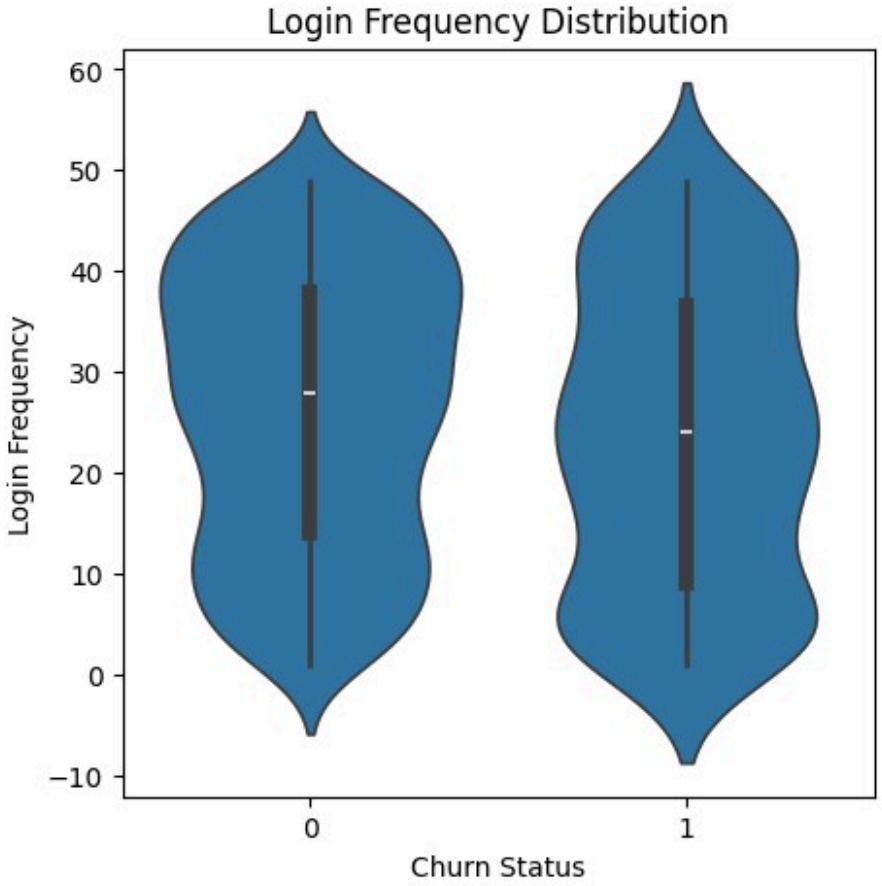
AYUSH SAIN
DATA ANALYSIS REPORT

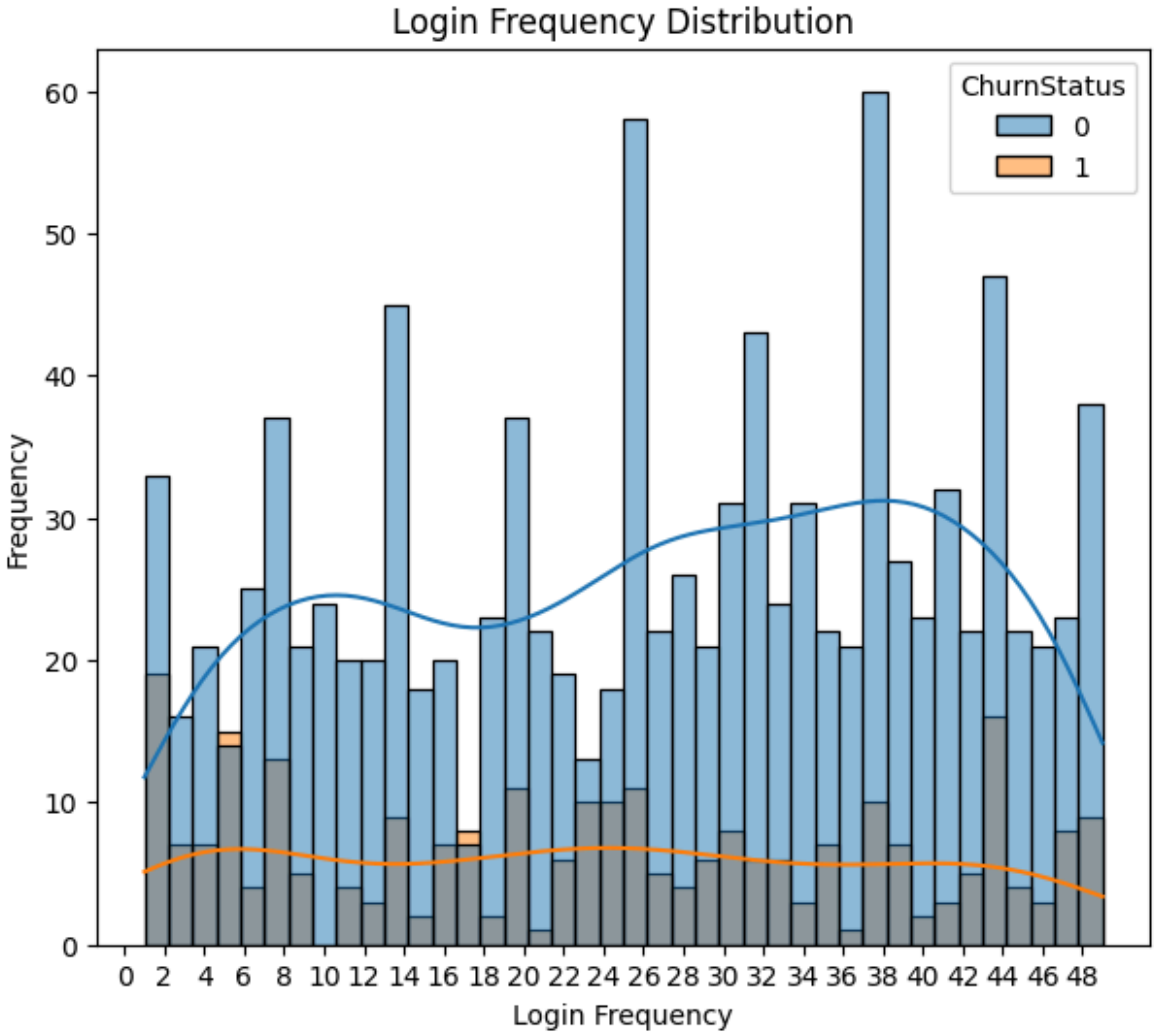
Data Preprocessing and Cleaning Report

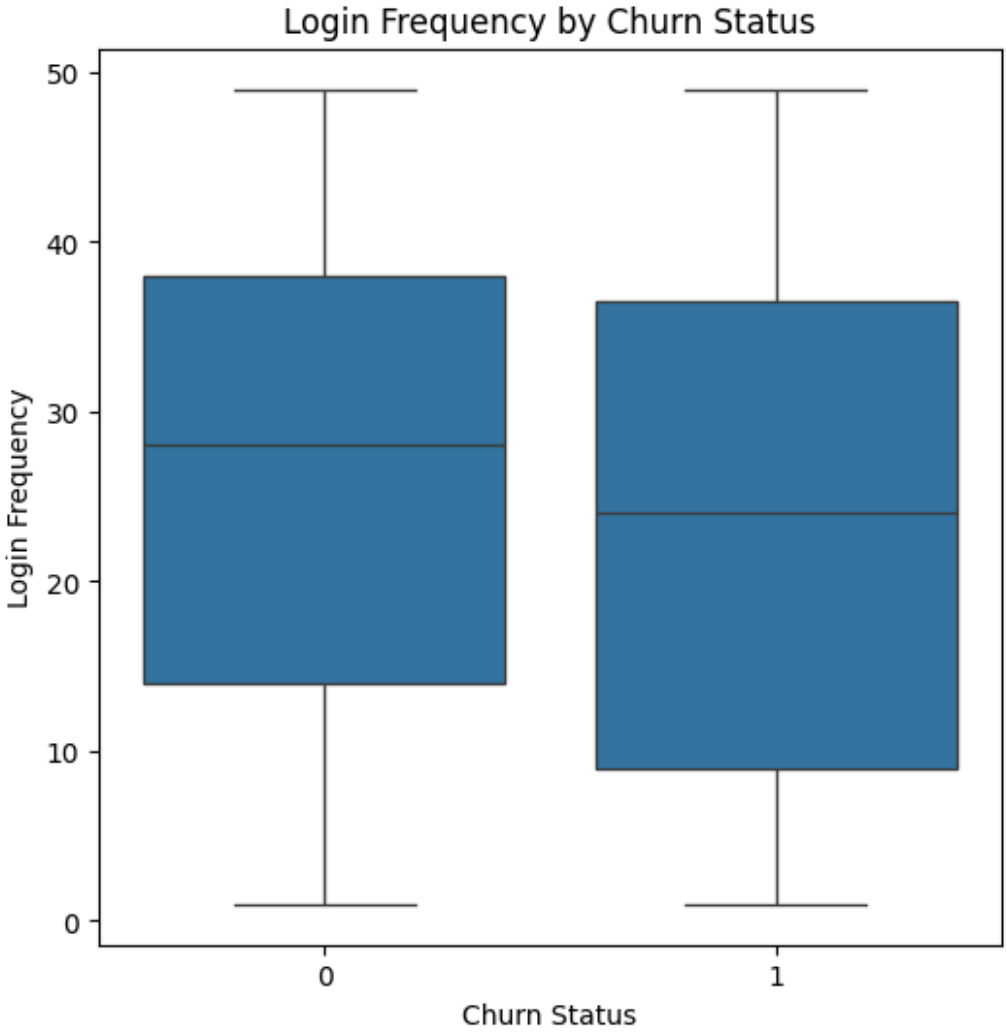
1. Introduction

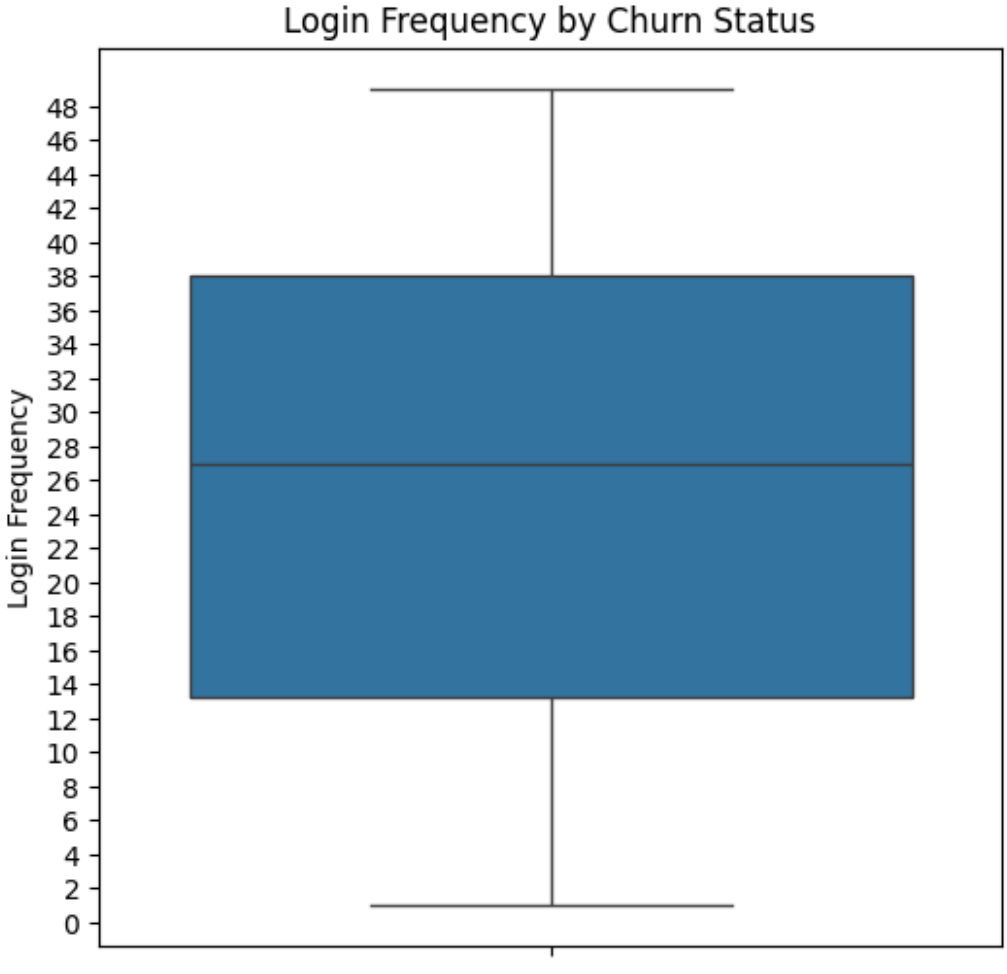
This section of the report details the systematic steps taken to prepare the customer churn dataset for analysis and model building. The dataset comprises multiple features related to customer demographics, behavior, and service usage. The primary goal of this preprocessing phase was to ensure the data is clean, consistent, and in a format suitable for machine learning tasks, particularly churn prediction.

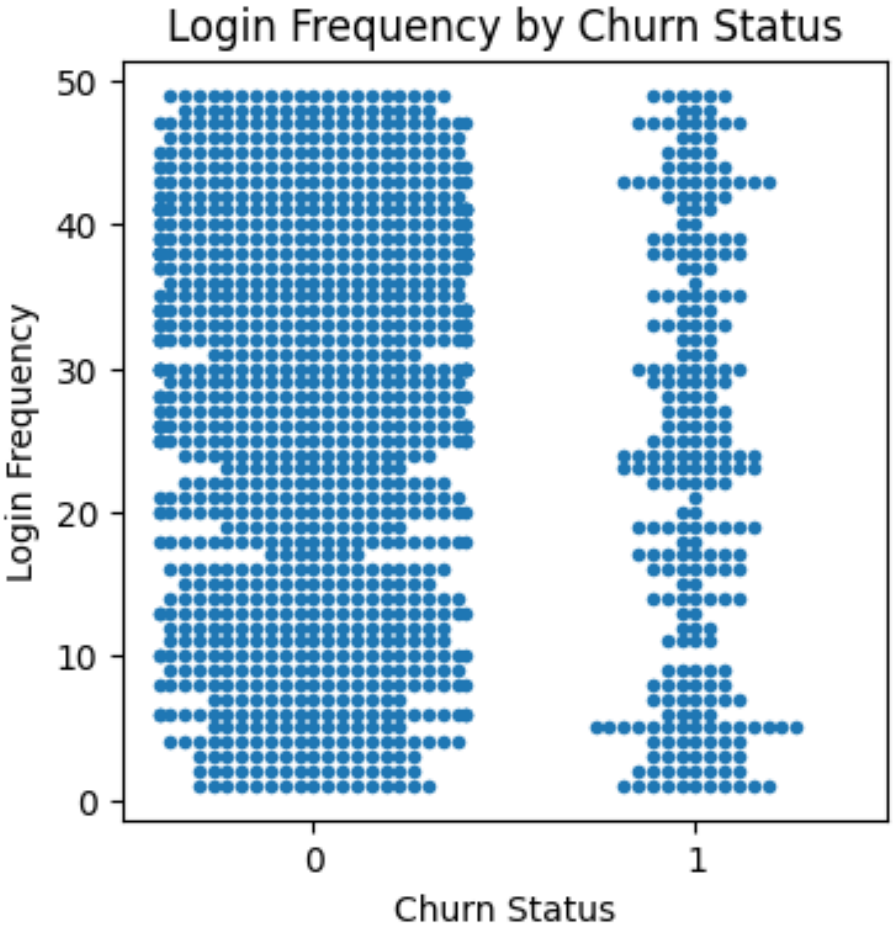


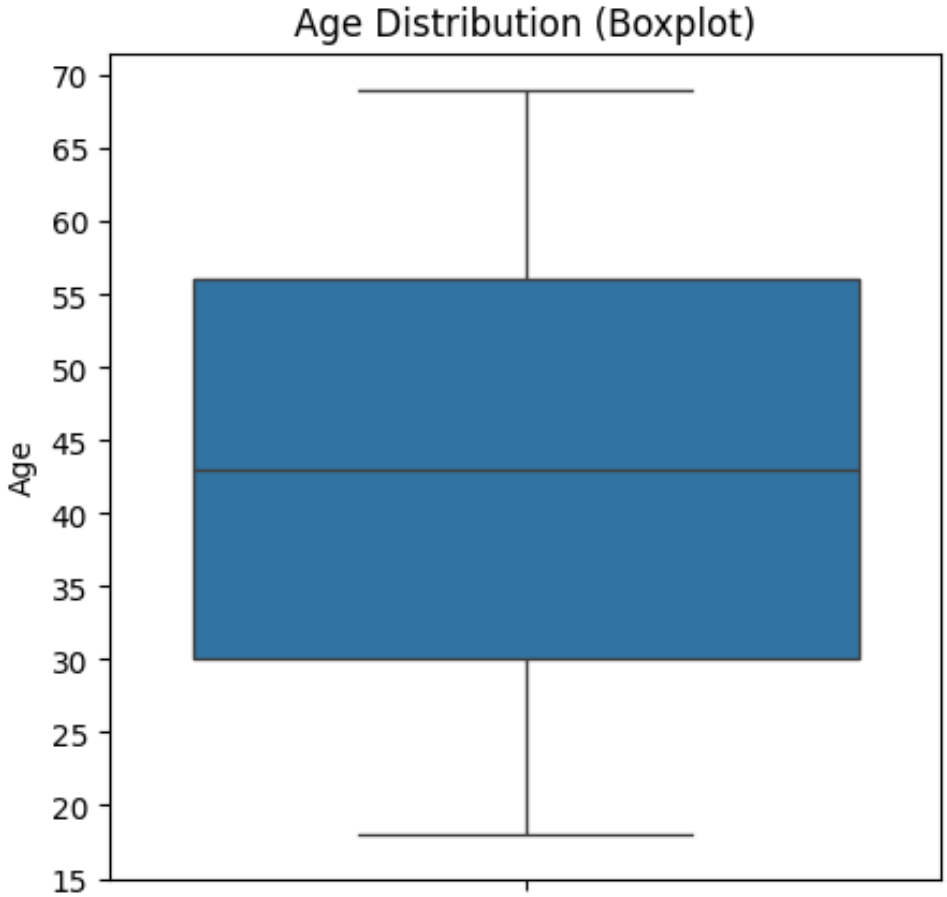

Login Frequency Distribution
Login Frequency
60
50
40
30
20
10
0
−10
0
1
Churn Status


Login Frequency Distribution
ChurnStatus
0
1
Frequency
60
50
40
30
20
10
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48
Login Frequency

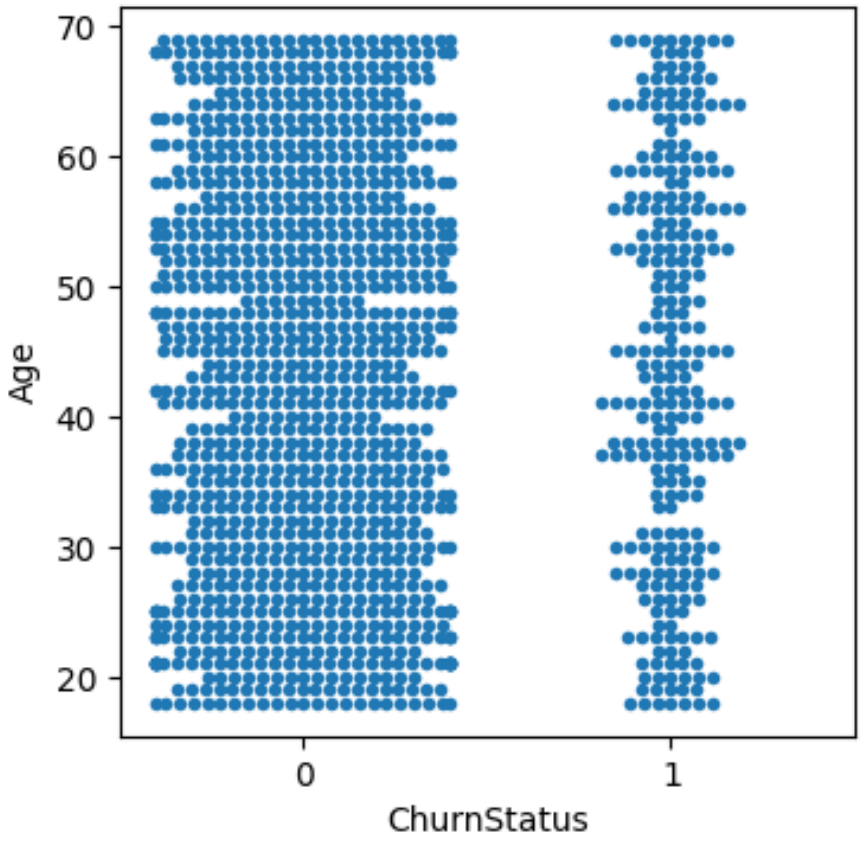

Login Frequency by Churn Status
Login Frequency
Churn Status
0
1
0
10
20
30
40
50

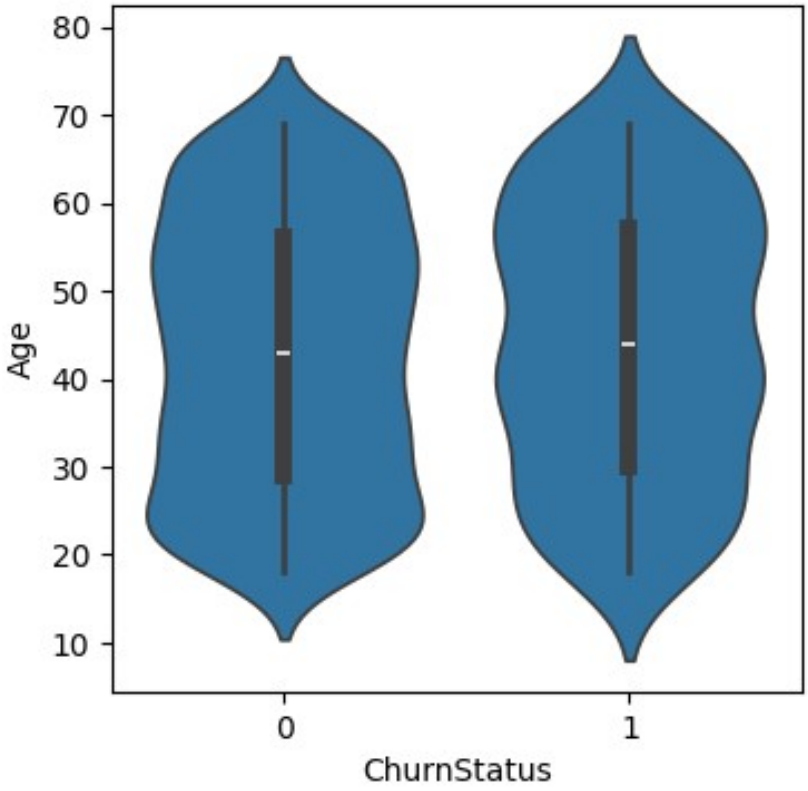

Login Frequency by Churn Status
Login Frequency
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48

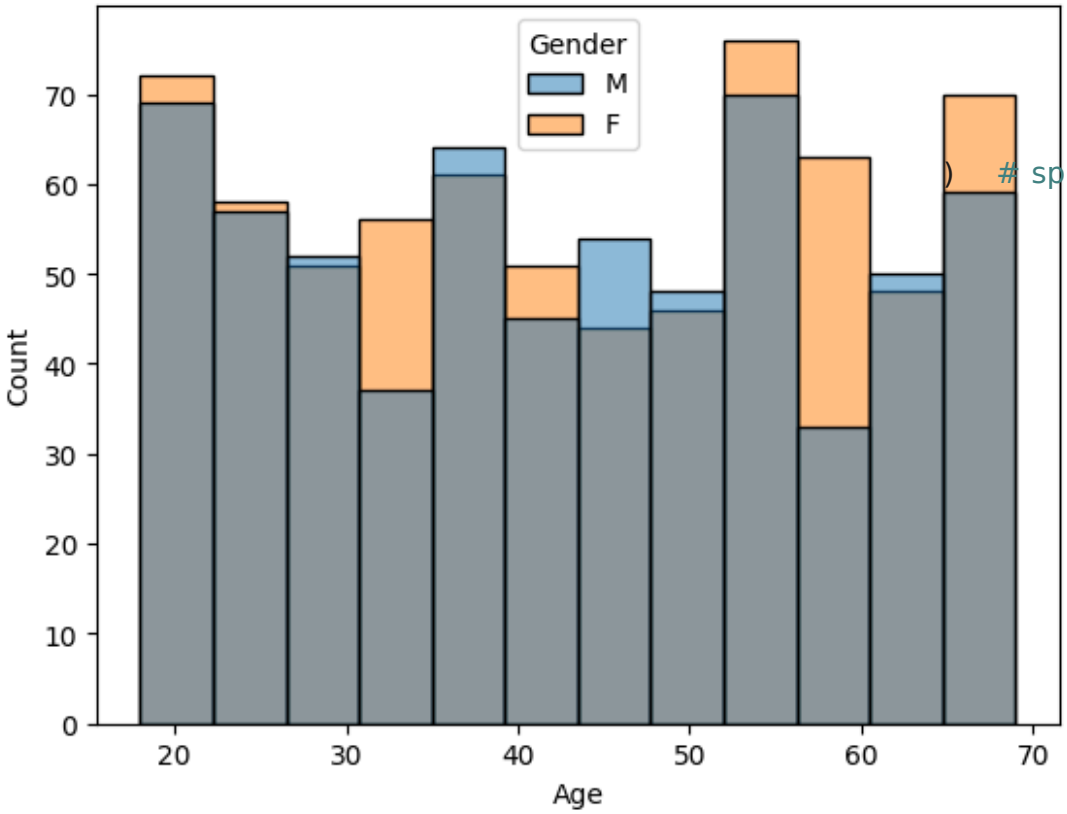

Login Frequency by Churn Status
Login Frequency
50
40
30
20
10
0
0
1
Churn Status


Age Distribution (Boxplot)
Age
70
65
60
55
50
45
40
35
30
25
20
15

bi variate analysis using the target variable

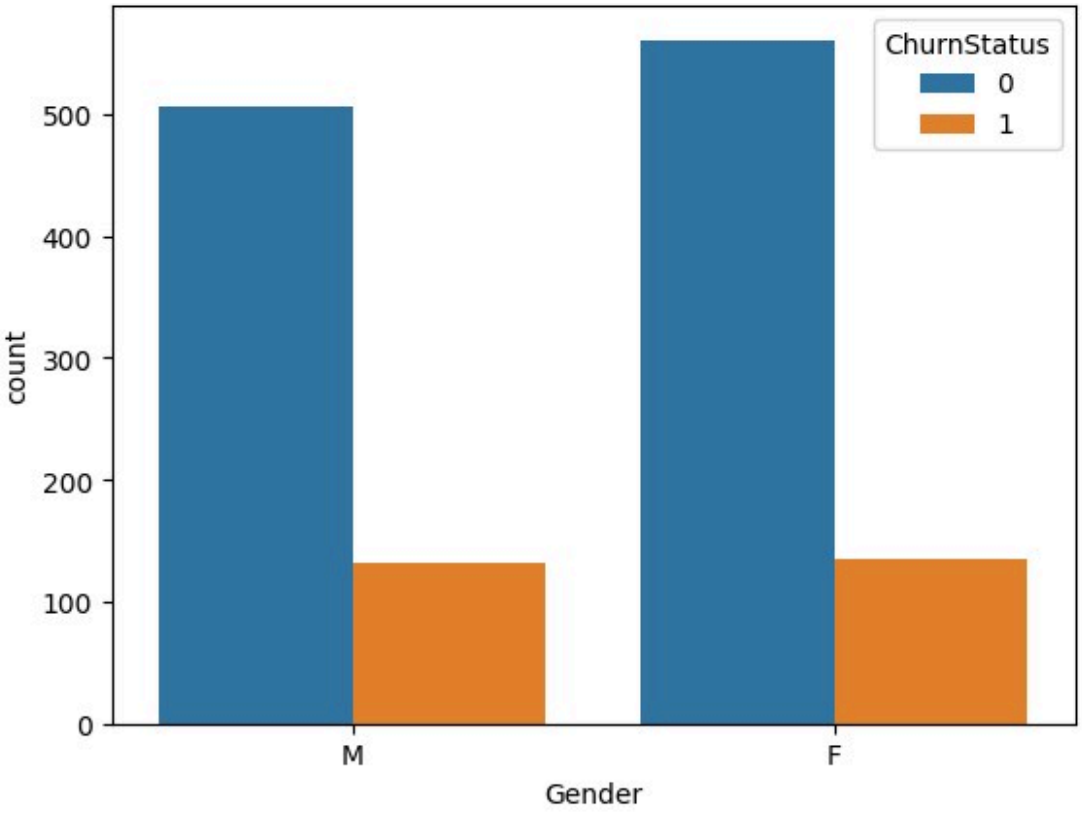


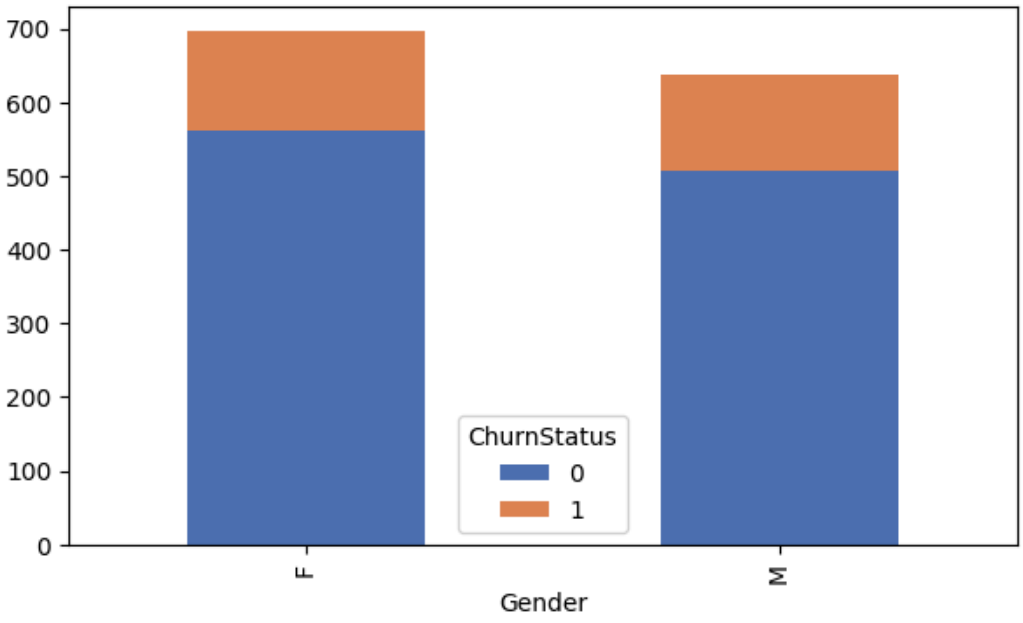


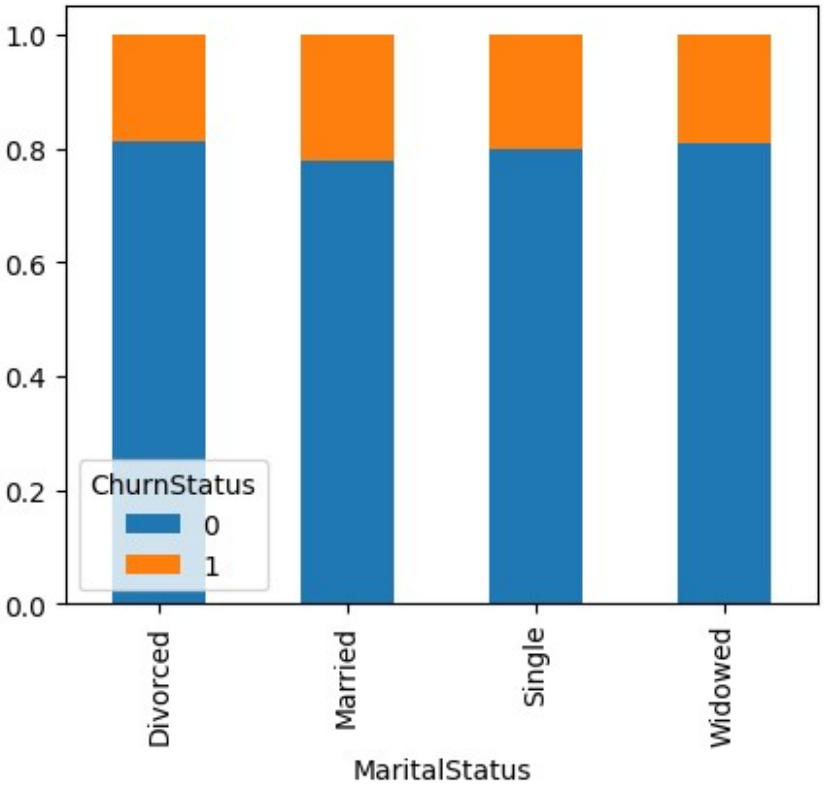


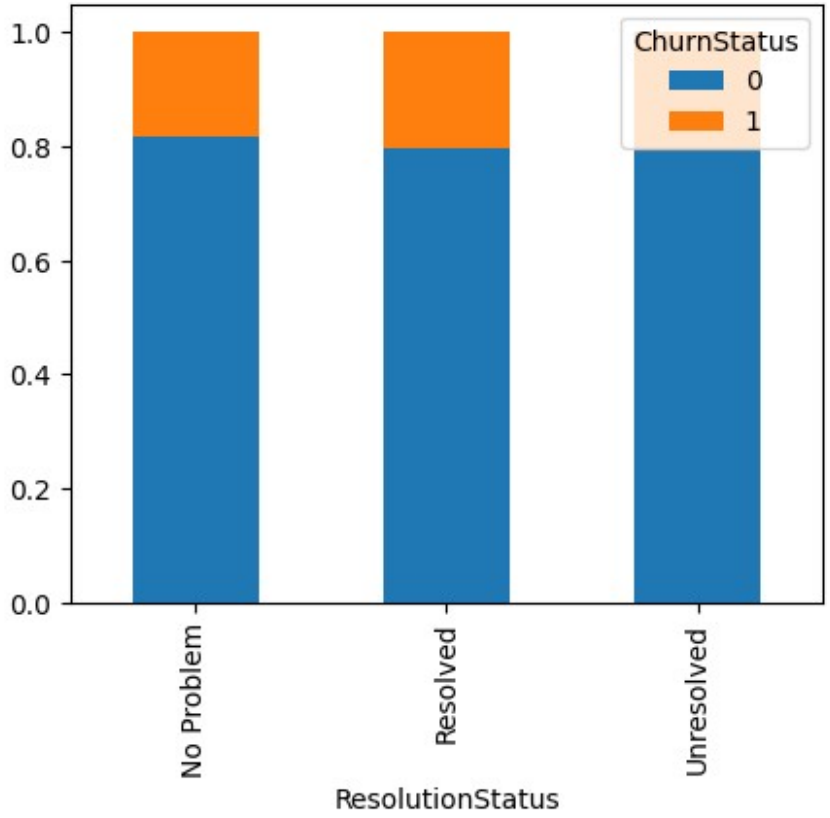

```
) # split by
```

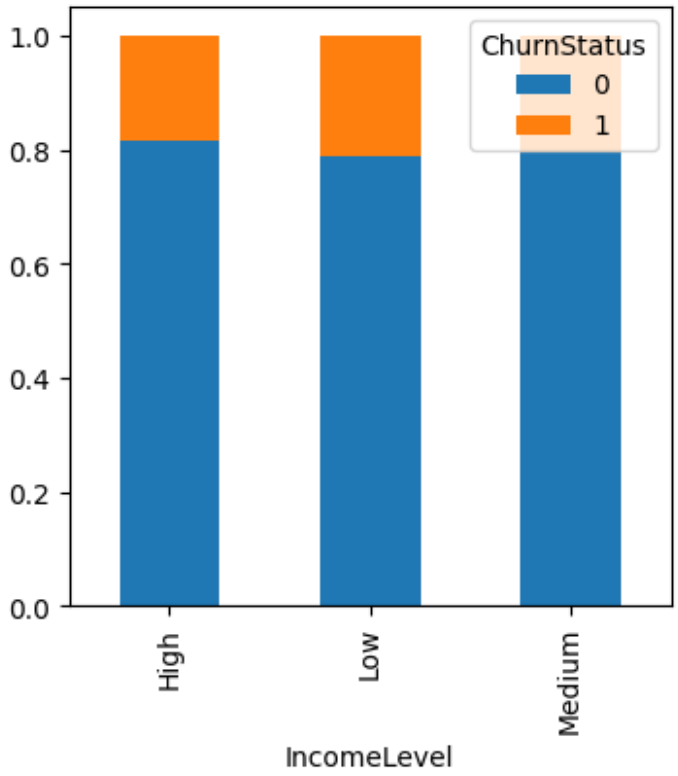
categorical vs target variable



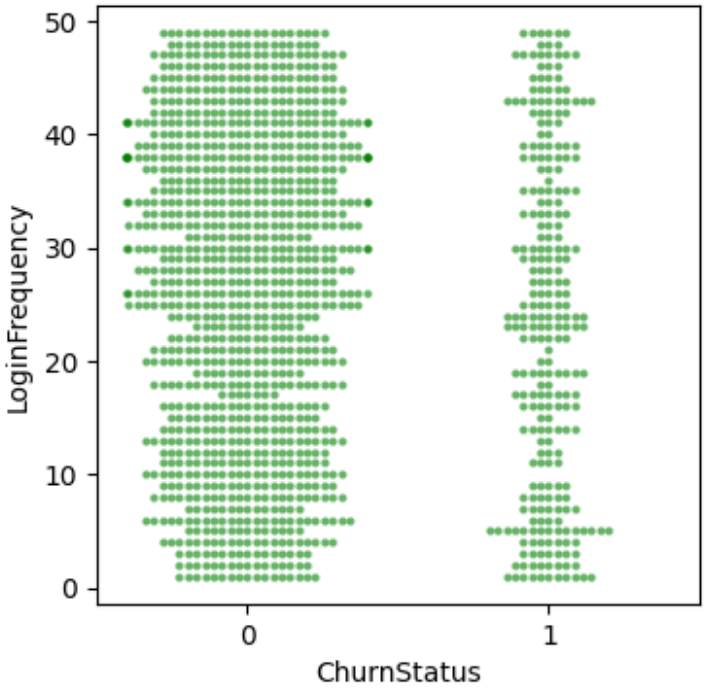


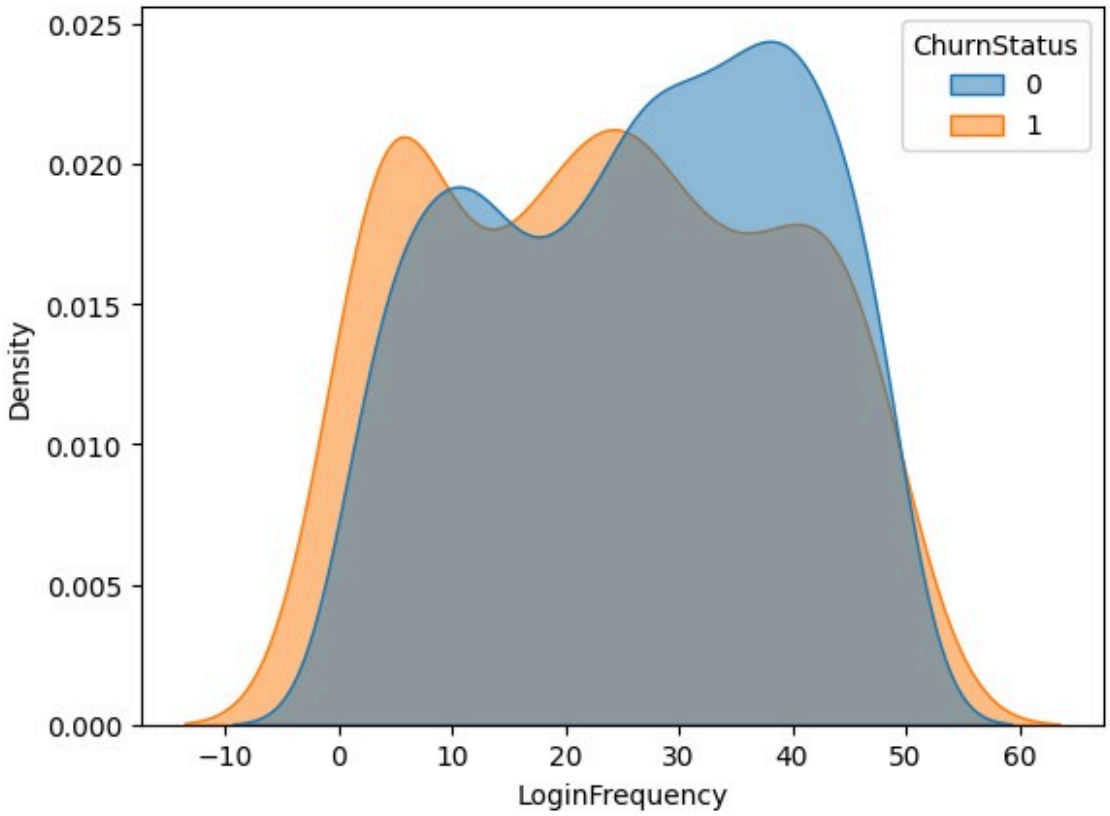

1.0
0.8
0.6
0.4
0.2
0.0
ChurnStatus
0
1
Divorced
Married
Single
Widowed
MaritalStatus

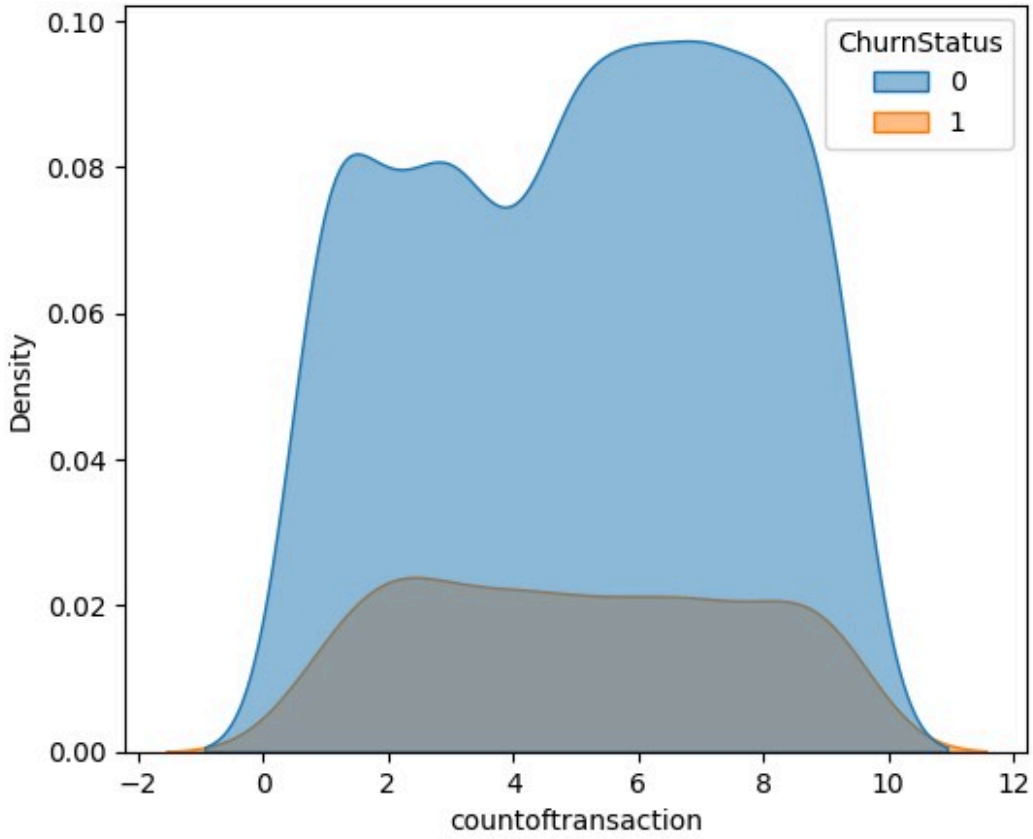

1.0
0.8
0.6
0.4
0.2
0.0
ChurnStatus
0
1
No Problem
Resolved
Unresolved
ResolutionStatus

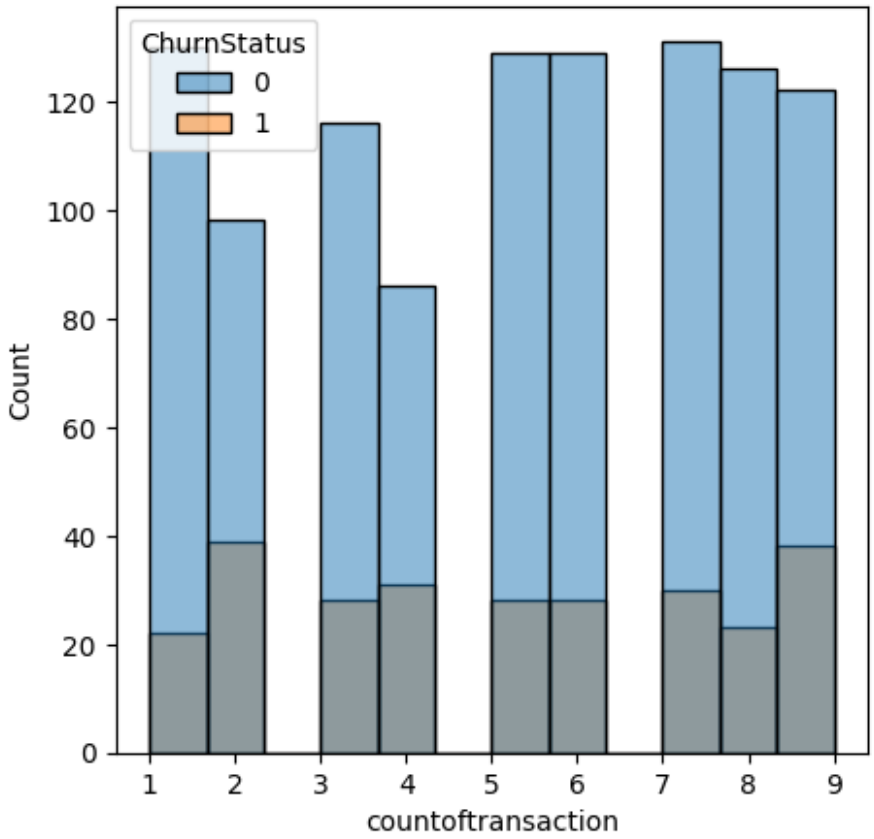


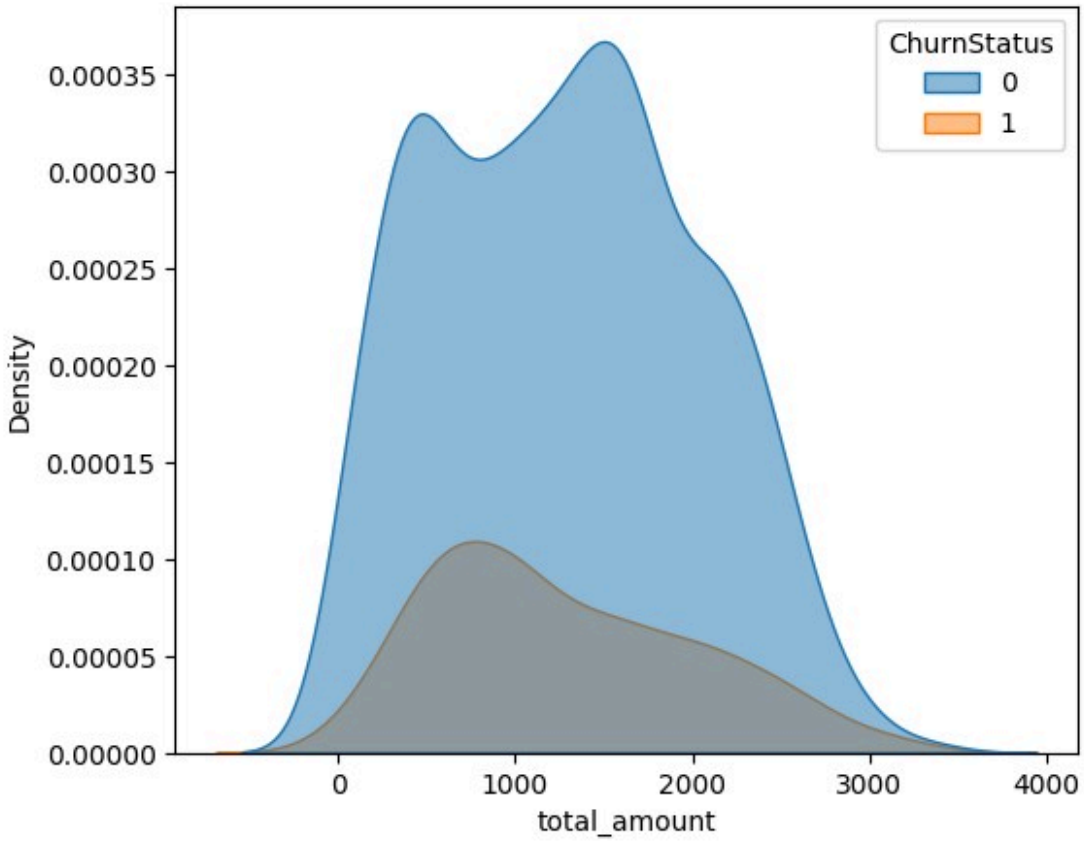
numerical vs target

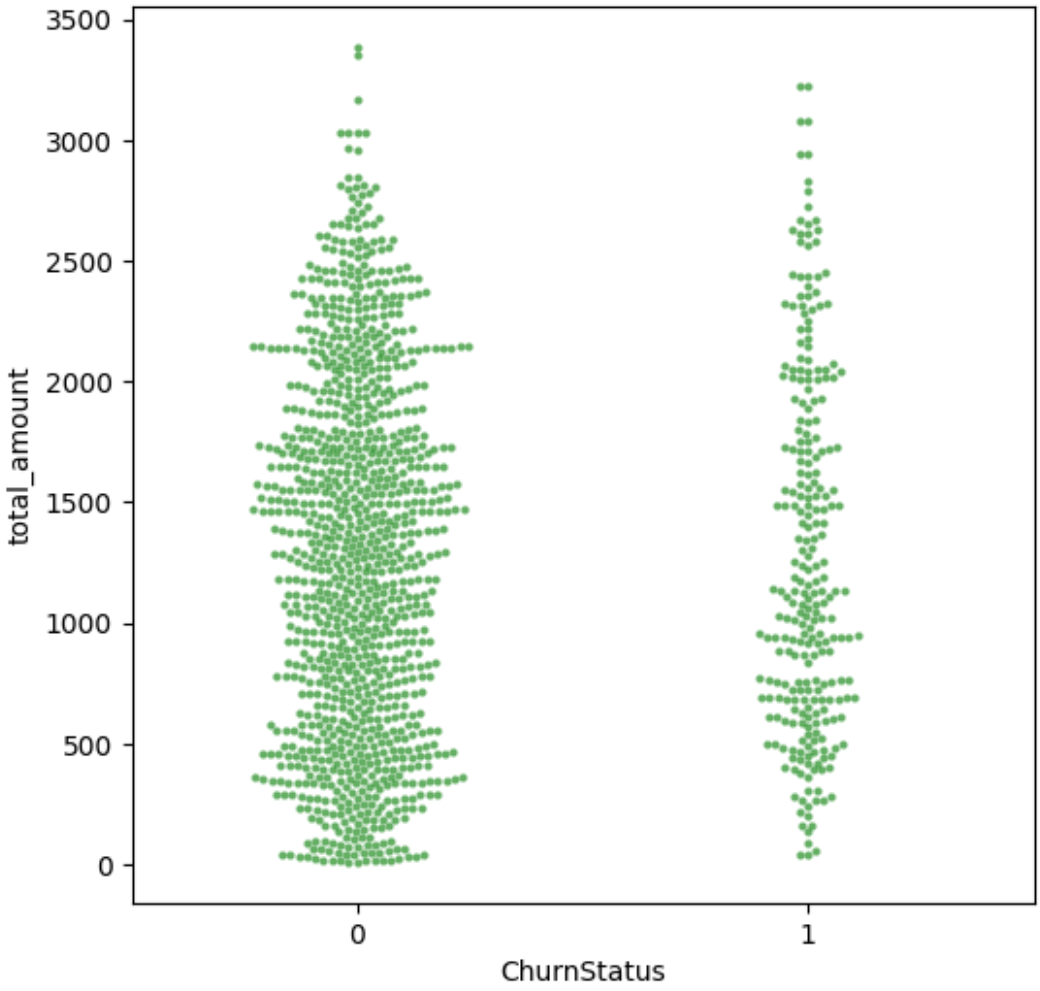


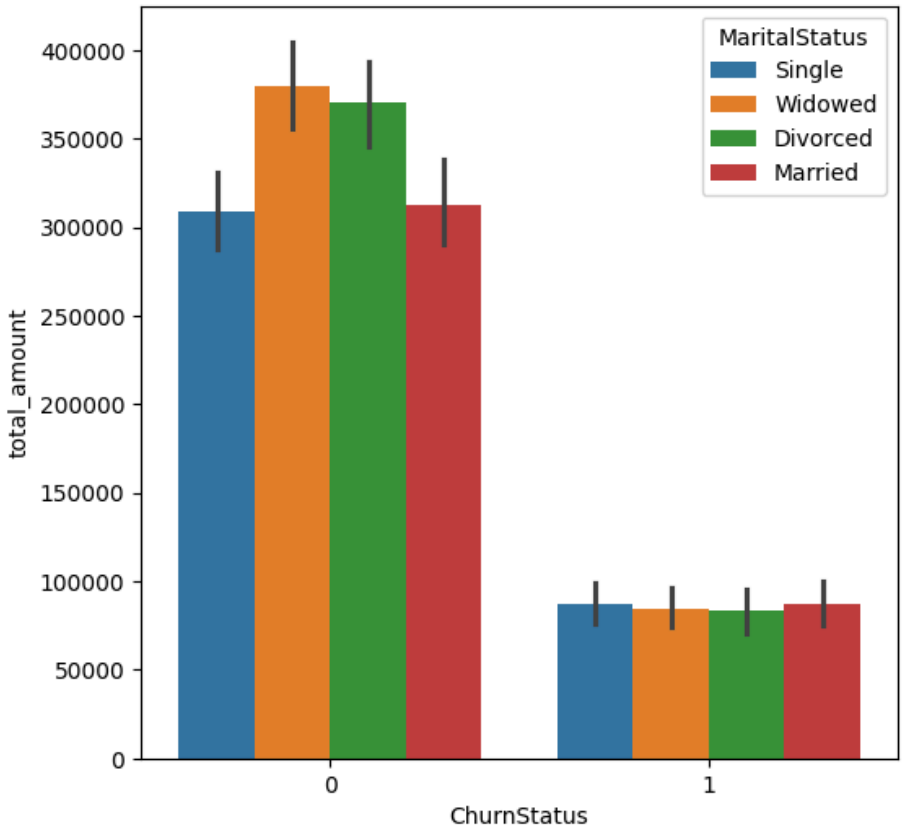

ChurnStatus
0
1
Density
0.025
0.020
0.015
0.010
0.005
0.000
−10
0
10
20
30
40
50
60
LoginFrequency

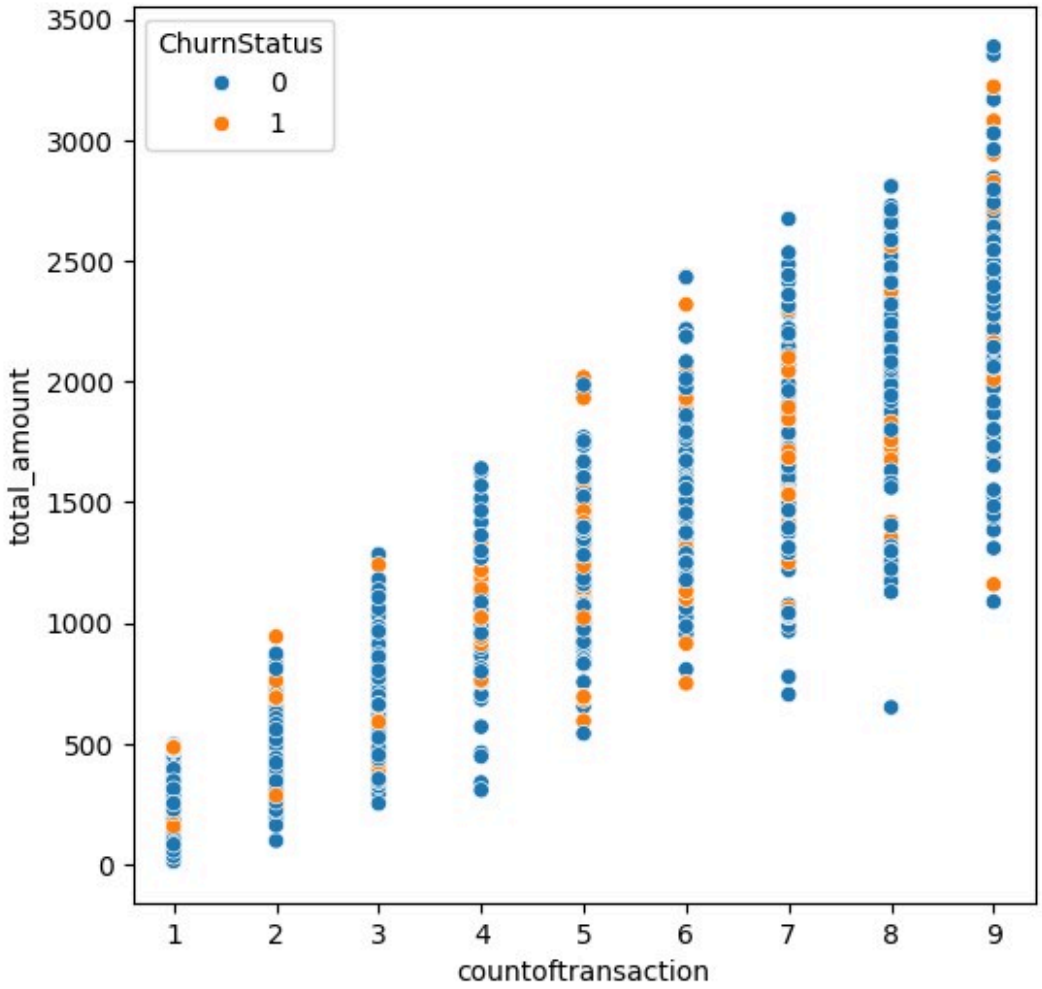

ChurnStatus
0
1
Density
0.10
0.08
0.06
0.04
0.02
0.00
−2
0
2
4
6
8
10
12
countoftransaction

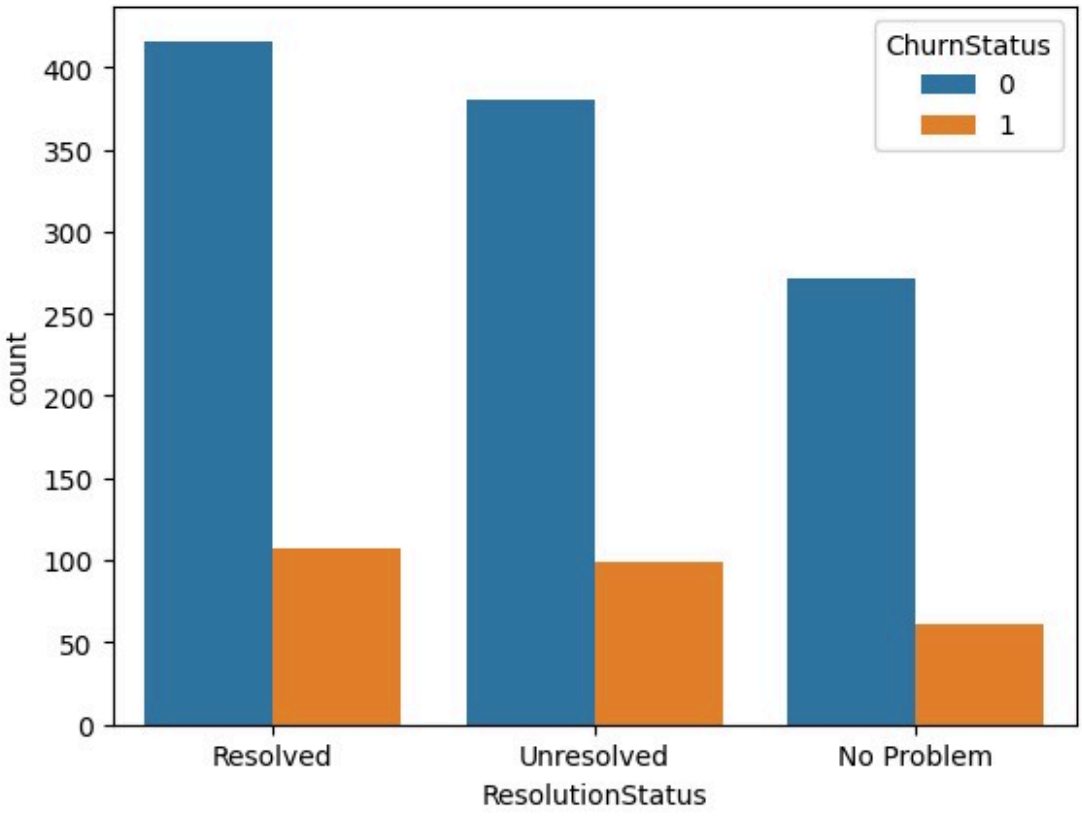

ChurnStatus
0
1
Count
countoftransaction
0
20
40
60
80
100
120
1
2
3
4
5
6
7
8
9

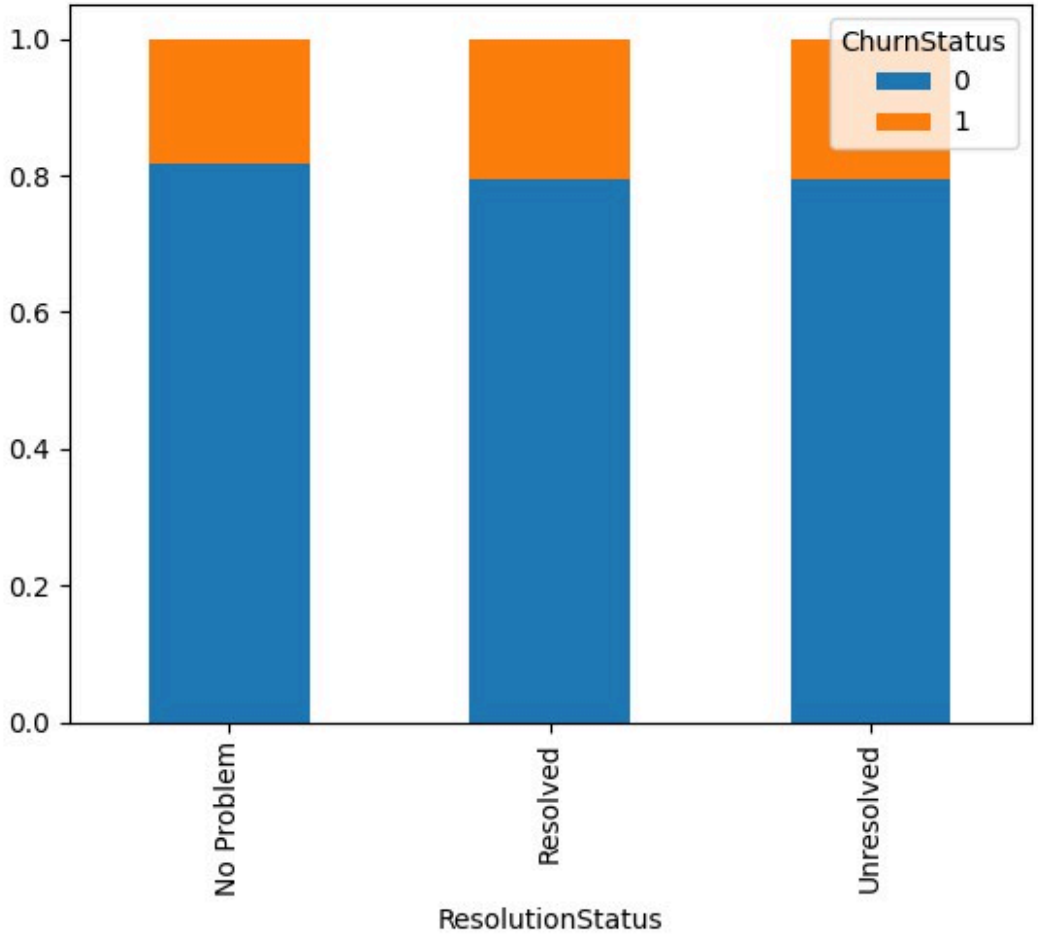

ChurnStatus
0
1
Density
total_amount
0.00000
0.00005
0.00010
0.00015
0.00020
0.00025
0.00030
0.00035
0
1000
2000
3000
4000

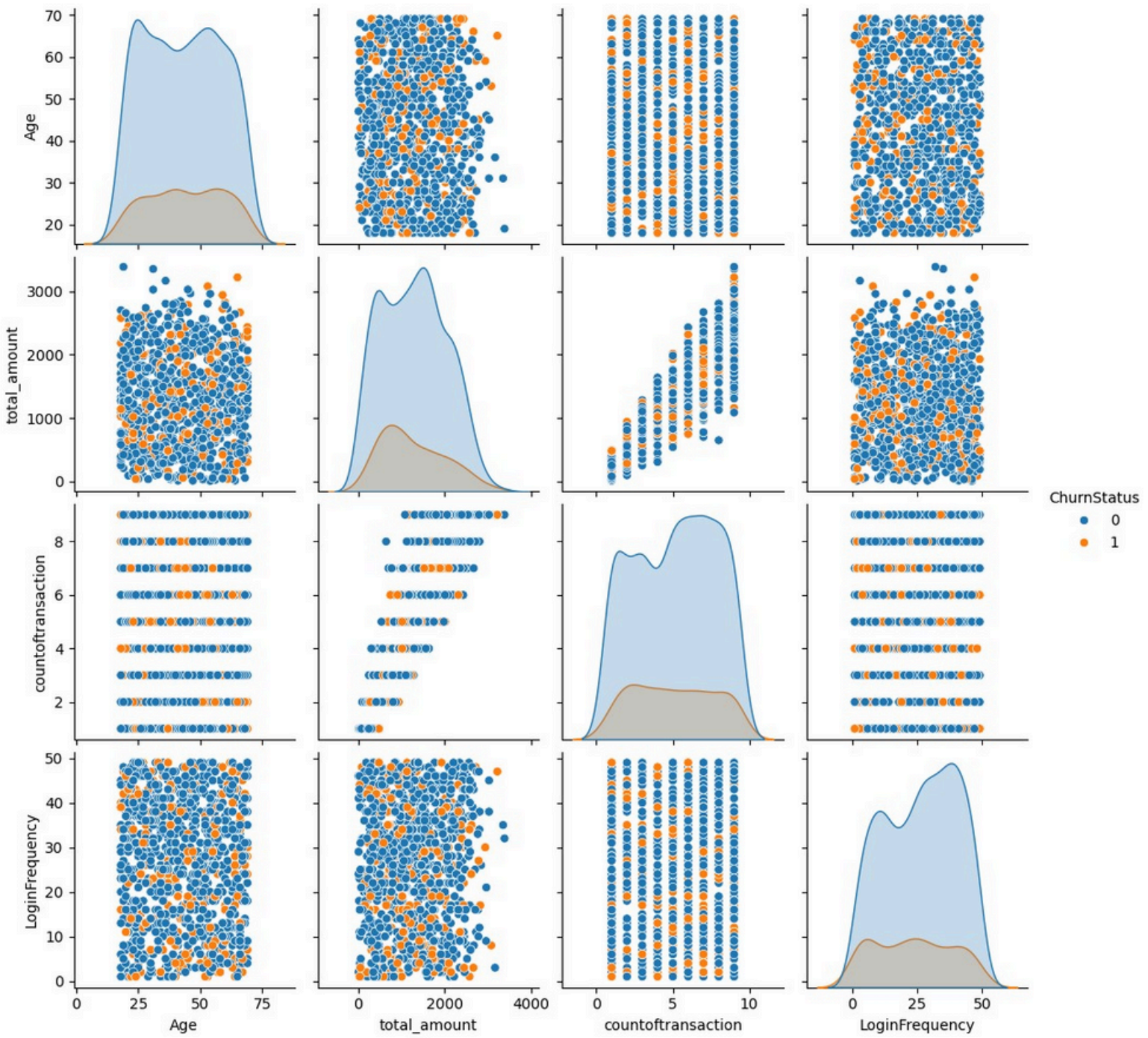

3500
3000
2500
2000
1500
1000
500
0
total_amount
0
1
ChurnStatus

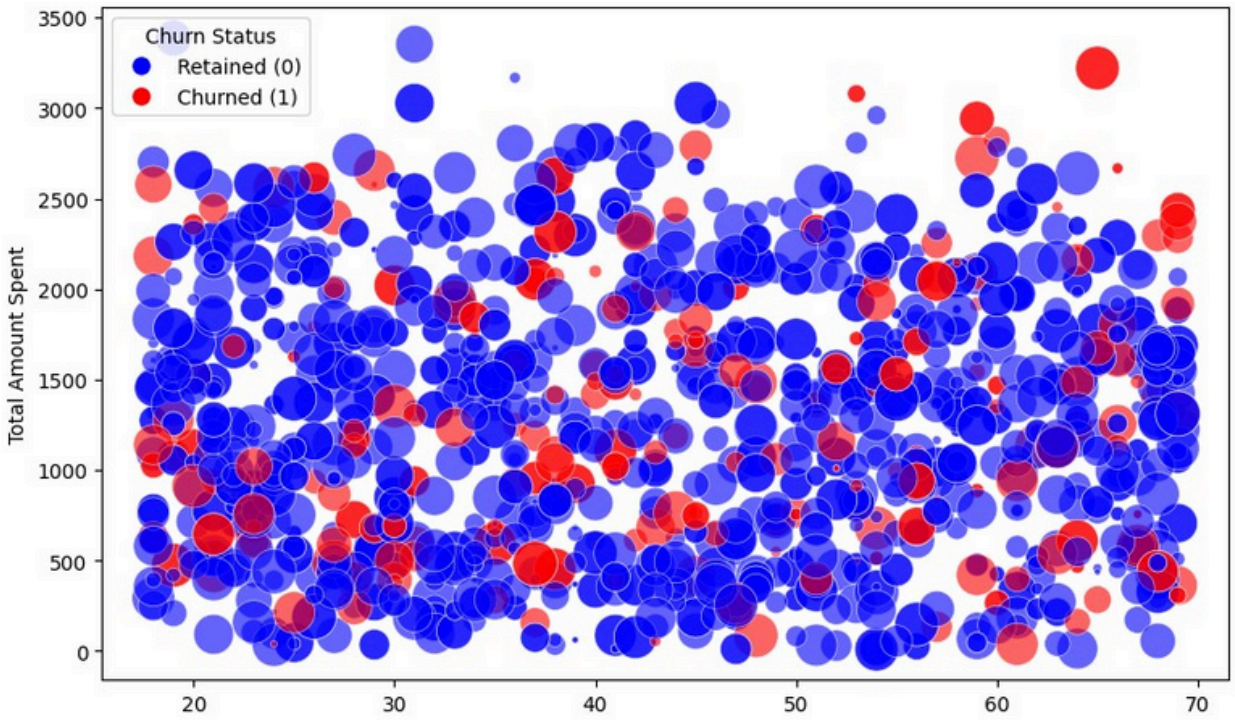

MaritalStatus
Single
Widowed
Divorced
Married
400000
350000
300000
250000
200000
150000
100000
50000
0
total_amount
0
1
ChurnStatus

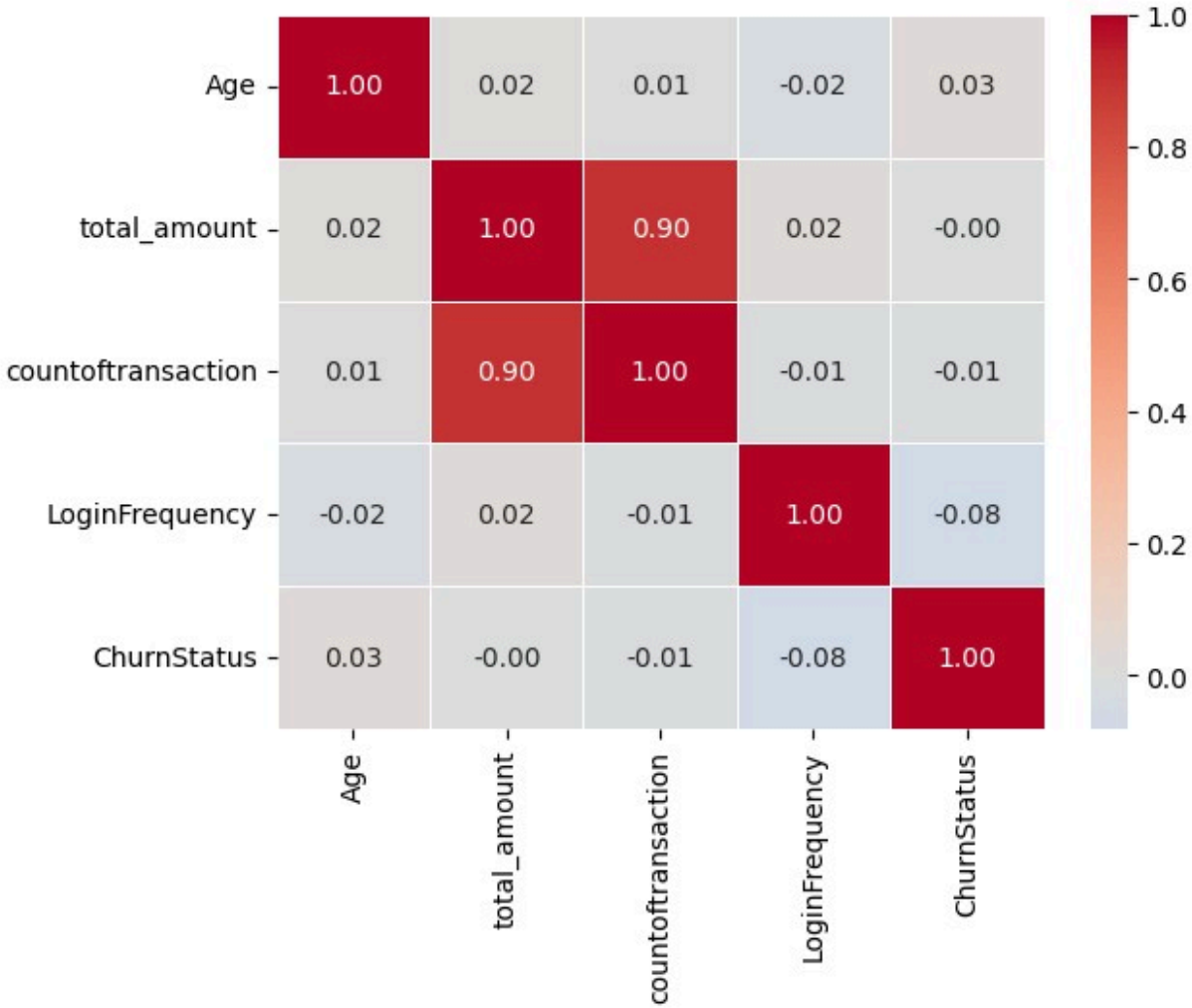

ChurnStatus
0
1
total_amount
countoftransaction
0
500
1000
1500
2000
2500
3000
3500
1
2
3
4
5
6
7
8
9


ChurnStatus
0
1
count
ResolutionStatus
0
50
100
150
200
250
300
350
400
Resolved
Unresolved
No Problem


ChurnStatus
0
1
1.0
0.8
0.6
0.4
0.2
0.0
No Problem
Resolved
Unresolved
ResolutionStatus


Age
total_amount
countoftransaction
LoginFrequency
ChurnStatus
0
1


Churn Status
Retained (0)
Churned (1)
Total Amount Spent
3500
3000
2500
2000
1500
1000
500
0
20
30
40
50
60
70

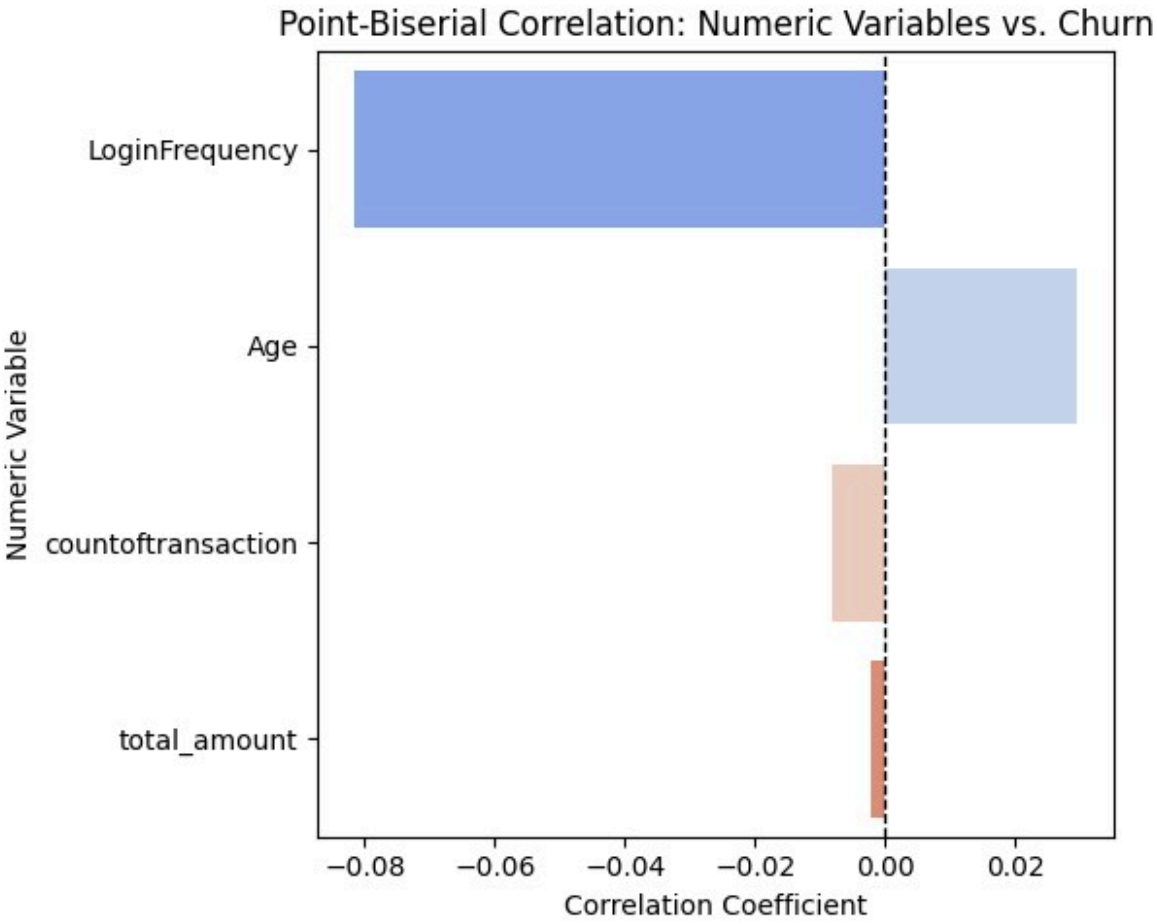

Age
total_amount
countoftransaction
LoginFrequency
ChurnStatus
1.00 0.02 0.01 -0.02 0.03
0.02 1.00 0.90 0.02 -0.00
0.01 0.90 1.00 -0.01 -0.01
-0.02 0.02 -0.01 1.00 -0.08
0.03 -0.00 -0.01 -0.08 1.00
1.0
0.8
0.6
0.4
0.2
0.0

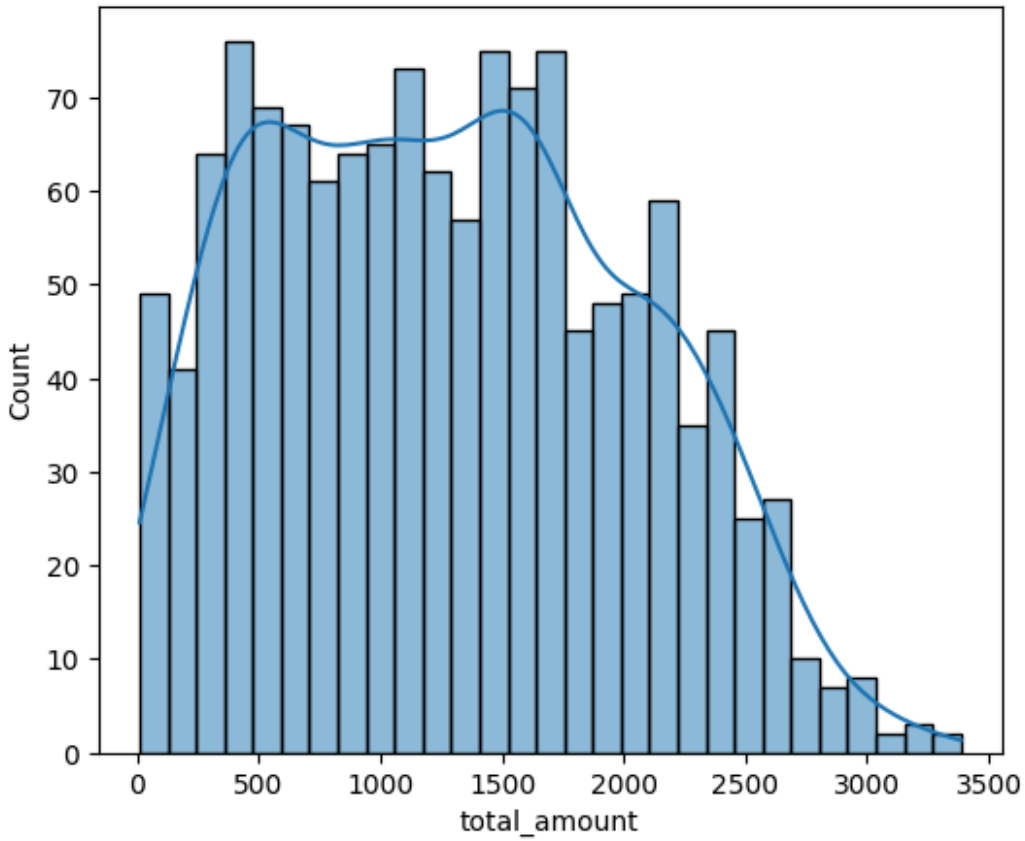
2. Handling Missing Values

Missing values were identified across both numerical and categorical features. Numerical columns such as Age, IncomeLevel, and LoginFrequency had minimal missing data, which were imputed using the median value to preserve distribution and avoid skewness. For categorical columns like Gender and ResolutionStatus, missing values were imputed using a new category labelled "Unknown" to avoid dropping informative instances. A binary flag indicating the presence of missing values was also created where necessary, to allow the model to capture any potential predictive signal associated with missingness. Records with missing target labels (ChurnStatus) were dropped, as they cannot contribute to supervised learning.

3. Outlier Detection and Treatment

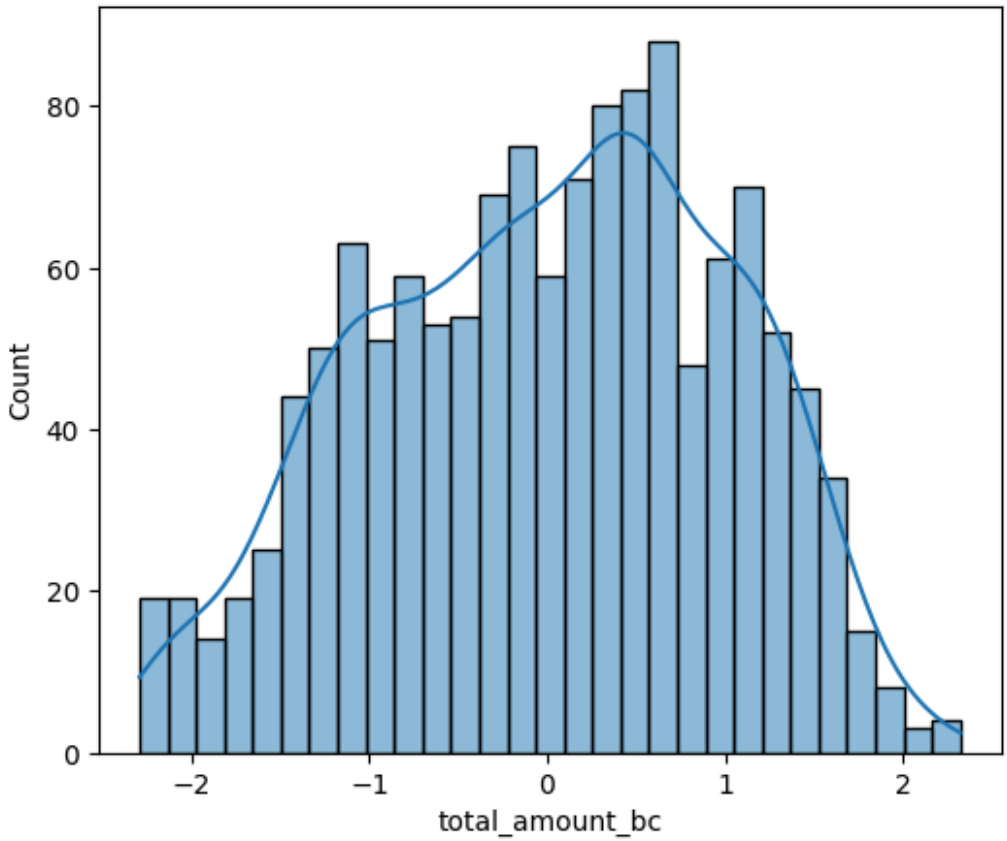
Outliers were assessed using the interquartile range (IQR) method and visual inspection via boxplots. Key numeric features such as total_amount, countoftransaction, and IncomeLevel exhibited significant right-skewed distributions, indicating the presence of high-end outliers. Rather than removing these values outright—which could eliminate high-value customers—extreme values were capped at the 99th percentile. Additionally, log transformation was applied to reduce skewness and compress scale differences for better model performance. Erroneous values in Age (e.g., below 0 or above 120) were considered data entry errors and removed.

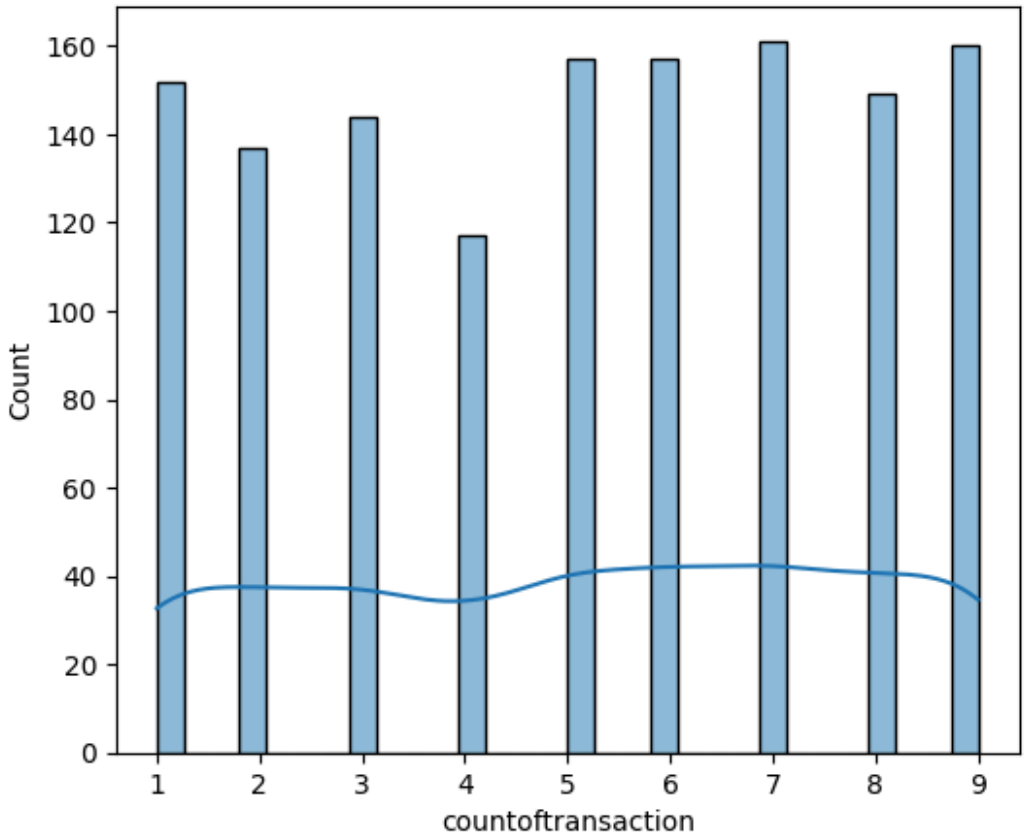

Point-Biserial Correlation: Numeric Variables vs. Churn
LoginFrequency
Age
countoftransaction
total_amount
Numeric Variable
−0.08
−0.06
−0.04
−0.02
0.00
0.02
Correlation Coefficient

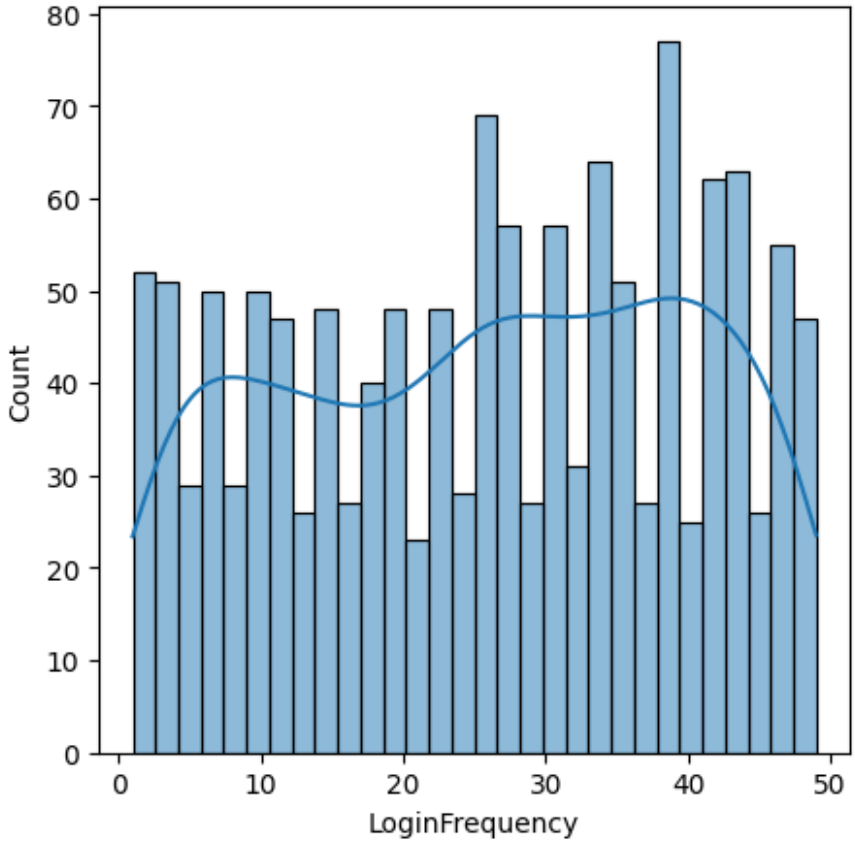


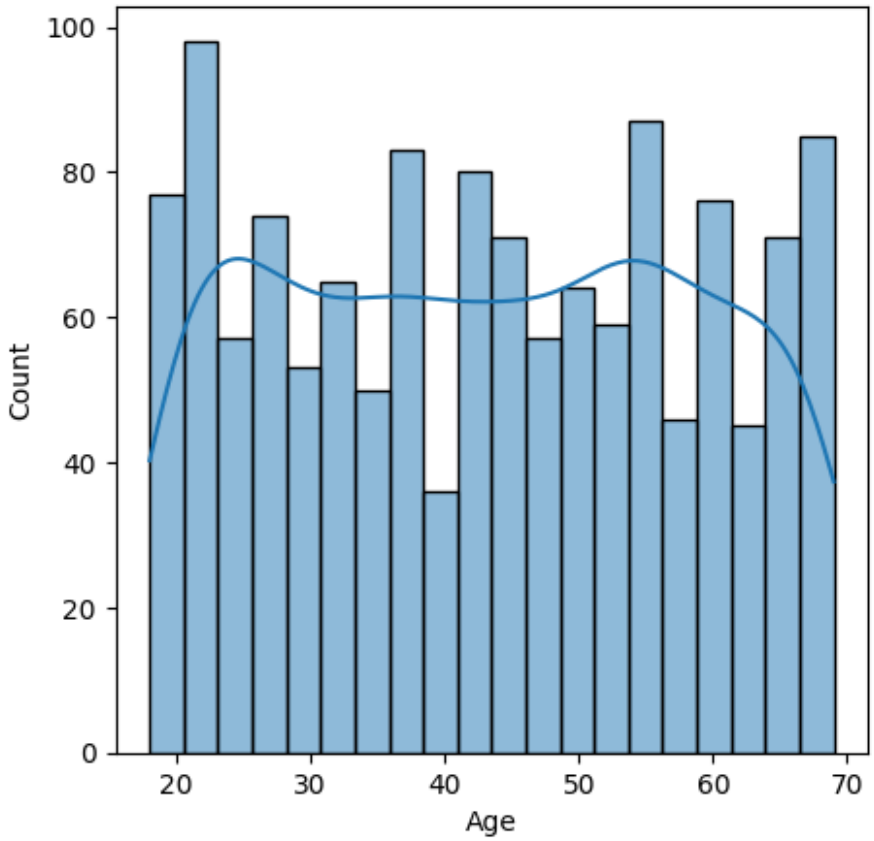

```
from sklearn.preprocessing import PowerTransformer

# For strictly positive data (Box-Cox)
pt = PowerTransformer(method='box-cox')
df_final['total_amount_bc'] =
pt.fit_transform(df_final[['total_amount']])
plt.figure(figsize=(6, 5))
sns.histplot( data = df_final                , x ='total_amount_bc' , bins = 29          ,
kde= True )
```

```
<Axes: xlabel='total_amount_bc', ylabel='Count'>
```


Count
80
60
40
20
0
−2
−1
0
1
2
total_amount_bc


Count
countoftransaction
0
20
40
60
80
100
120
140
160
1
2
3
4
5
6
7
8
9


Count
LoginFrequency
0
10
20
30
40
50
60
70
80
0
10
20
30
40
50


100
80
60
40
20
0
Count
20
30
40
50
60
70
Age

4. Feature Scaling

To ensure consistent scales across features, especially for distance-based and gradient-based models, all numeric variables were standardized using Z-score normalization. This method centers the features around a mean of zero and standard deviation of one, helping algorithms like logistic regression and SVM converge faster and perform more accurately. Log-transformed features were scaled after transformation to maintain consistency in range and distribution.

5. Encoding Categorical Variables

Categorical variables including Gender, MaritalStatus, and ResolutionStatus were transformed using one-hot encoding. This approach preserves the non-ordinal nature of these categories and prevents the model from inferring an artificial ranking. To avoid dimensionality explosion from sparse categories, rare levels were grouped under an "Other" category where necessary. The target variable ChurnStatus was already binary (0 for retained customers, 1 for churned customers) and required no further encoding.

6. CONCLUSION

Through careful handling of missing data, strategic treatment of outliers, consistent feature scaling, and appropriate encoding of categorical variables, the dataset is now clean and structured for effective modeling. These preprocessing steps lay a robust foundation for the next phase of the project: exploratory data analysis and predictive modeling. The final dataset balances statistical integrity with business relevance, ensuring high-quality insights into customer churn behavior.