CAPSTONE PROJECT CUSTOMER CHURN (NOTE-1)

BY — *Harsh Patel*

8th December 2024

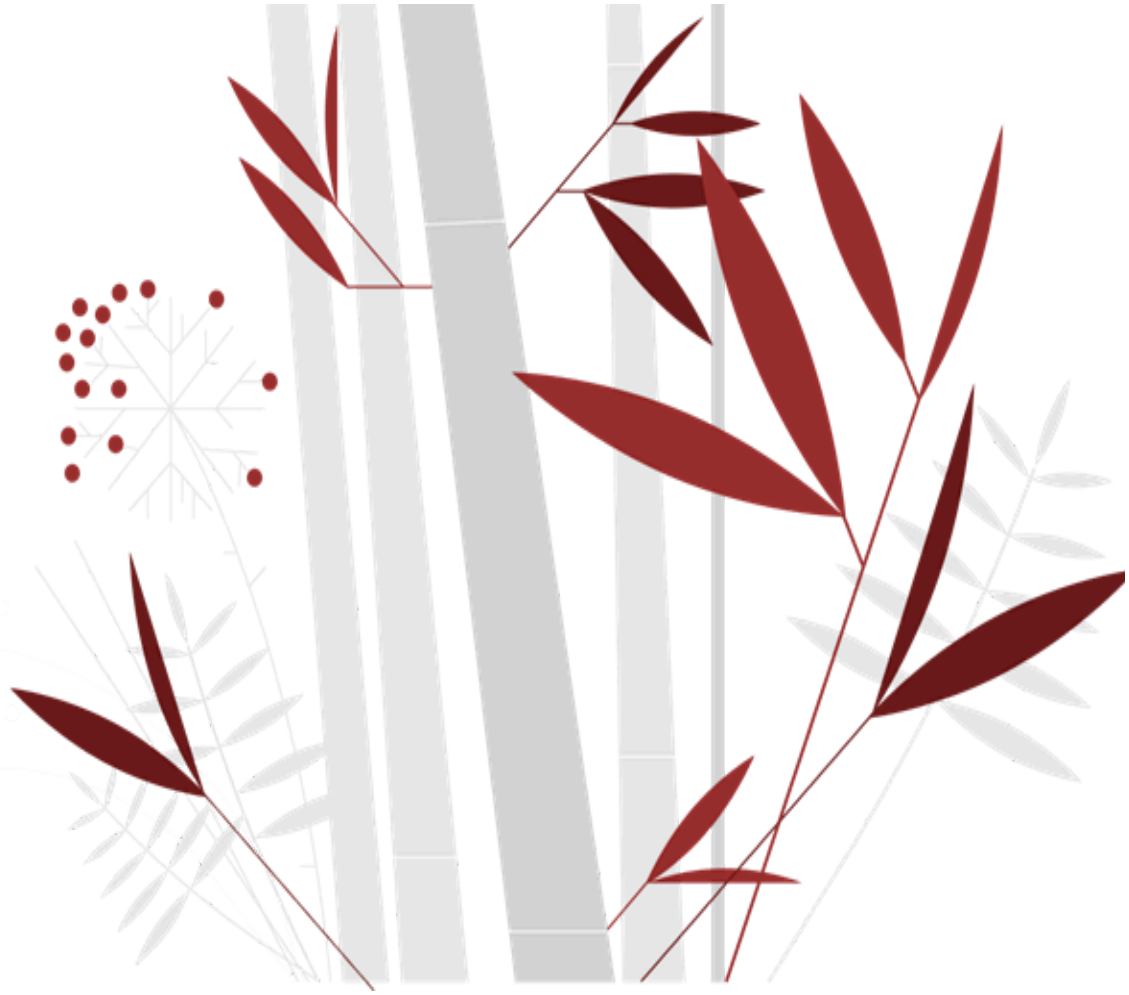


Table No.	Chart and Graphs Names	Pg no.
1.	Null values in Dataset	16
2.	Before Outlier treatment	18
3.	After Outlier treatment	35
4.	Univariate Analysis of Numerical variables	36
5.	Univariate Analysis of Categorical Variables	44
6.	Pairplot of Numerical variables	50
7.	Heatmap for Numerical variables	51
8.	Bivariate Analysis of Numerical vs Categorical variables	61
9.	Bivariate Analysis of Categorical Variables	71
10.	Dendrogram Chart	81
11.	Elbow Chart	82
12.	Silhouette Score Plot	82
13.	K Means vs Churn	84

1.Problem Statement for E-commerce/DTH company:

In the face of intense competition, DTH/E-commerce is finding it challenging to retain its existing customers. Therefore, the company wants to develop a churn prediction model to identify accounts likely to leave and provide targeted offers to these potential churners. This is critical since one account can represent multiple customers, making account churn a significant issue.

Our task is to create a churn prediction model for this company and offer business recommendations for a customer retention campaign. The model must be precise in identifying churners and the offers suggested should ensure a win-win situation for both the company and the customers, aiming to retain customers without significantly impacting revenue.

1.1 Problem Definition:

Given the competitive market, the current project focuses on developing an effective churn prediction model and providing strategic business recommendations. The proposed campaign should be cost-effective and within budget, aiming to retain customers without high expenditure.

In this competitive market, DTH/E-commerce companies face challenges in retaining customers. They aim to develop a system for predicting customer churn and offering appealing deals to retain them. The goal is to design a unique, targeted campaign

that balances customer retention with financial stability, ensuring higher customer retention and overall profitability.

1.2 Need of Study/Project of this Problem:

This project is vital for the client to strategically plan their future product designs, sales strategies, and targeted promotions for different client segments. The project's results will provide a comprehensive understanding of the firm's current position and its risk-taking capacity. Additionally, it will shed light on the organization's future potential, helping them to plan more effectively and retain customers in the long term.

For a DTH /E-commerce company, the most significant expense is acquiring a new customer. A new customer must be retained for several years to recover the initial acquisition cost and make the account profitable. Therefore, customer churn directly affects the profitability of a DTH/E-commerce operator. DTH/E-commerce must constantly expand their customer base to maintain profitability, as they often incur fixed broadcaster/content provider fees regardless of the number of customers. Consequently, it is crucial to both grow the customer base and protect the existing one.

Acquiring a new customer can cost five times more than retaining an existing one. Increasing customer retention by just 5% can boost profits by 25-95%. Given that customer churn affects both top-line and bottom-line revenue, protecting the existing

customer base is essential. Offering retention deals to all customers would harm profitability, so it is crucial to focus on those at a higher risk of churning.

This study is essential because the DTH/E-commerce company is facing fierce competition and struggling to retain current customers. Predicting churn is critical to identifying customers who might leave so the company can take action to retain them. Understanding why customers churn enables the company to offer special deals to encourage them to stay. By predicting churn, the company can focus on retaining its most valuable customers and improving loyalty.

The study will provide valuable insights into customer behavior and preferences, helping design targeted and effective marketing campaigns to increase retention and revenue. A successful churn prediction model and effective campaign recommendations can lead to increased customer satisfaction, positive brand perception, and improved financial performance.

Financial Impact: Account churn directly contributes to revenue loss and profitability decline. Retaining customers is more cost-effective than acquiring new ones.

Customer Lifetime Value: Churned customers represent lost opportunities for future revenue streams. By retaining customers, businesses can maximize the lifetime value of each customer.

Brand Reputation: High churn rates can damage a company's reputation and make it difficult to attract new customers. Retaining customers demonstrates a strong brand and customer satisfaction.

Competitive Advantage: In a competitive market, retaining existing customers is a key differentiator. Businesses that effectively manage churn can gain a competitive edge.

Customer Satisfaction and Loyalty: Retaining customers indicates they are satisfied with the products or services and are willing to continue their relationship with the company.

Addressing account churn through accurate churn prediction and targeted retention strategies is crucial for long-term financial success and customer satisfaction.

1.3 Understanding Business/Social Opportunity:

The E-commerce company is grappling with intense competition, making customer retention a major challenge in the current market. Developing a churn prediction model will enable the company to identify potential churners in advance and take proactive measures to retain them. Tailored offers and incentives can be provided to these potential churners, catering to their specific needs and preferences, thereby increasing the chances of retention. Reducing customer churn will help the company maintain a stable revenue stream and achieve long-term sustainability, as retaining existing customers is more cost-effective than acquiring new ones. Data-driven decision-making, based on the churn prediction model, will enable the company to make informed business choices, enhancing the customer experience and providing a competitive advantage over rivals.

Using churn prediction and targeted offers allows the E-commerce company to better understand its business and social opportunities. This helps build stronger customer relationships, increase revenue, and stay ahead of the competition. By making informed decisions, the company can achieve long-term success in a challenging market.

This case study focuses on a company where customers are assigned unique account IDs, and a single account ID can represent multiple customers (such as in a family plan). Customers are segmented across various plans based on their usage and the devices they use (computer or mobile), with flexibility in payment modes and cashback on bill payments. The business relies heavily on customer loyalty and stickiness, which stems from providing quality and value-added services. Running various promotional and festival offers may help in acquiring new customers and retaining existing ones.

We can conclude that retaining a customer ensures regular income for the organization, acquiring a new customer brings additional income, and losing a customer has a negative impact, especially since a single account ID can represent multiple customers. The closure of one account ID means losing multiple customers. This presents a great opportunity for the company, as having a DTH connection is a necessity for almost every family, which also leads to increased competition. The question arises: how can the company differentiate itself from competitors? What parameters are crucial for maintaining customer loyalty and retention? Addressing these social responsibilities will determine the best player in the market.

1.4 Understanding the Data Collection Methodology:

- I will assume that data has been collected from a random sample of 11,260 unique account IDs, covering various genders and marital statuses.
- Analysis of the variables such as "CC_Contacted_L12m", "rev_per_month", "Complain_l12m", "rev_growth_yoy", "coupon_used_l12m", "Day_Since_CC_connect", and "cashback_l12m" indicates that the data spans the last 12 months.
- The dataset comprises 19 variables, with 18 independent variables and 1 dependent or target variable, which indicates whether a customer has churned or not.
- The data captures the services used by customers, their payment options, and basic individual details.
- The dataset includes a mix of categorical and continuous variables.

2. Data Exploration and Data-Cleaning:

2.1 Visual Inspection of Data:

We will import the necessary python library files and upload the excel file “Customer Churn Data” using the read_excel function and open the sheet named “Data for DBSA”. Then, we will see the first five and last five rows of the dataset to see and and get the visual representation of the given data.

	AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count
0	20000	1	4	3.0	6.0	Debit Card	Female	3.0	3
1	20001	1	0	1.0	8.0	UPI	Male	3.0	4
2	20002	1	0	1.0	30.0	Debit Card	Male	2.0	4
3	20003	1	0	3.0	15.0	Debit Card	Male	2.0	4
4	20004	1	0	1.0	12.0	Credit Card	Male	2.0	3

Table-1 First five rows of dataset

	AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_coun
11255	31255	0	10	1.0	34.0	Credit Card	Male	3.0	
11256	31256	0	13	1.0	19.0	Credit Card	Male	3.0	
11257	31257	0	1	1.0	14.0	Debit Card	Male	3.0	
11258	31258	0	23	3.0	11.0	Credit Card	Male	4.0	
11259	31259	0	8	1.0	22.0	Credit Card	Male	3.0	

Table-2 Last five rows of dataset

Now, we will use shape, info and describe function on this data set and we got the following tables.We obtained the shape of the dataset has 11260 rows and 19 columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11260 entries, 0 to 11259
Data columns (total 19 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   AccountID               11260 non-null  int64
 1   Churn                   11260 non-null  int64
 2   Tenure                  11158 non-null  object
 3   City_Tier               11148 non-null  float64
 4   CC_Contacted_LY         11158 non-null  float64
 5   Payment                 11151 non-null  object
 6   Gender                  11152 non-null  object
 7   Service_Score           11162 non-null  float64
 8   Account_user_count      11148 non-null  object
 9   account_segment         11163 non-null  object
 10  CC_Agent_Score          11144 non-null  float64
 11  Marital_Status          11048 non-null  object
 12  rev_per_month           11158 non-null  object
 13  Complain_ly             10903 non-null  float64
 14  rev_growth_yoy          11260 non-null  object
 15  coupon_used_for_payment 11260 non-null  object
 16  Day_Since_CC_connect    10903 non-null  object
 17  cashback                10789 non-null  object
 18  Login_device            11039 non-null  object
dtypes: float64(5), int64(2), object(12)
```

Table-3 Information of dataset.

Table-4 shows below the summary for the continuous numerical variables five important summaries like min,max,count,quantile values.In addition it has mean, standard deviation values too.

	count	mean	std	min	25%	50%	75%	max
AccountID	11260.0	25629.500000	3250.626350	20000.0	22814.75	25629.5	28444.25	31259.0
Churn	11260.0	0.168384	0.374223	0.0	0.00	0.0	0.00	1.0
City_Tier	11148.0	1.653929	0.915015	1.0	1.00	1.0	3.00	3.0
CC_Contacted_LY	11158.0	17.867091	8.853269	4.0	11.00	16.0	23.00	132.0
Service_Score	11162.0	2.902526	0.725584	0.0	2.00	3.0	3.00	5.0
CC_Agent_Score	11144.0	3.066493	1.379772	1.0	2.00	3.0	4.00	5.0
Complain_ly	10903.0	0.285334	0.451594	0.0	0.00	0.0	1.00	1.0

Table-4 Description of dataset.

Several columns that should be read as numeric are currently read as object types. For example, the "Tenure" column, which is numeric, has been read as an object. These columns need to be checked for special characters and cleaned before converting them to numeric for further processing.

There are no duplicate rows in the dataset; each account ID is unique.

Several columns contain null values.

The following table shows the number of rows containing nulls and special characters that require data cleaning. All special characters present in the dataset were treated as nulls for imputation.

Columns such as "Gender" and "account_segment" had multiple values representing the same category, e.g., 'M' and 'Male'. These values have been cleaned up to ensure consistency.

It can be observed that there are few categorical variables that are included in the numerical variables and also we found that the dataset has anomalies or bad values as well as null/empty cells in the dataset.

Data Description:

S.no	Column	Column Description	Data Description
1	AccountID	account unique identifier	Unique ID. Hence, it will not be used in modelling
2	Churn	account churn flag (Target)	Target variable. Contains 1 for churned and 0 for non-churned
3	Tenure	Tenure of account	Continuous field. Contains values ranging from 0 to 99
4	City_Tier	Tier of primary customer's city	Categorical ordinal - values 1,2,3
5	CC_Contacted_LY	How many times all the customers contacted customer care in last 12 months	Continuous field. Contains values ranging from 4 to 132
6	Payment	Preferred Payment mode of the customers in the account	Categorical nominal - values Credit card, debit card, E wallet, UPI, Cash on Delivery
7	Gender	Gender of the primary customer	Categorical nominal - values Male, Female, M and F (M and F need to be converted to Male and Female)
8	Service_Score	Satisfaction score given by customers	Categorical ordinal - values 0 to 5
9	Account_user_count	Number of customers tagged with this account	Limited range. Can be treated as categorical - values 1 to 6
10	account_segment	Account segmentation on the basis of spend	Categorical nominal - values HNI, Regular, Regular Plus, Super, Super plus and variations with +
11	CC_Agent_Score	Satisfaction score given on customer care service	Categorical ordinal - values 1 to 5
12	Marital_Status	Marital status of primary customer	Categorical nominal - contains values Married, Single and Divorced

13	rev_per_month	Monthly average revenue from account in last 12 months	Continuous field. Contains values ranging from 1 to 140
14	Complain_ly	Complaints raised by account in last 12 months	Categorical - 0 (for no) or 1 (for yes)
15	rev_growth_yoy	Revenue growth percentage of the account (last 12 months vs last 24 to 13 month)	Continuous field. Contains values ranging from 4 to 28
16	coupon_used_for_payment	How many times customers have used coupons for payment in last 12 months	Continuous field, but with limited range. Contains values ranging from 0 to 16
17	Day_Since_CC_connect	Number of days since no customers contacted customer care	Continuous field. Contains values ranging from 0 to 47
18	Cashback	Monthly average cashback generated by account in last 12 months	Continuous field. Contains values ranging from 0 to 1997
19	Login_device	Preferred login device of the customers in the account	Categorical nominal - contains values Mobile, Computer

Table-5 Dataset Variables

2.2 Understanding of Attributes:

The table below shows the attribute names, their descriptions, and the type of values they contain. Although some variable names are slightly long, they do not contain blanks or special characters. Therefore, it has been decided to retain the current column names as they are self-explanatory and easy to understand when interpreting plots in univariate and bivariate analysis. The variable names will be changed later to shorten them or ensure uniformity during the one-hot encoding process.

Column	Values present	% Rows with values present	Number of Nulls	% Rows with nulls	Data clean-up needed?	% Rows needing data cleaning
AccountID	11260	100.00%	0	0%	None	0.00%
Churn	11260	100.00%	0	0%	None	0.00%
Tenure	11158	99.09%	102	0.91%	Yes - #	1.03%
City_Tier	11148	99.01%	112	0.99%	None	0.00%
CC_Contacted_LY	11158	99.09%	102	0.91%	None	0.00%
Payment	11151	99.03%	109	0.97%	None	0.00%
Gender	11152	99.04%	108	0.96%	Yes - M,F	5.74%
Service_Score	11162	99.13%	98	0.87%	None	0.00%
Account_user_count	11148	99.01%	112	0.99%	Yes - @	2.95%
account_segment	11163	99.14%	97	0.86%	Yes-Regular + Super +	2.74%
CC_Agent_Score	11144	98.97%	116	1.03%	None	0.00%
Marital_Status	11048	98.12%	212	1.88%	None	0.00%
rev_per_month	11158	99.09%	102	0.91%	Yes - +	6.12%
Complain_ly	10903	96.83%	357	3.17%	None	0.00%
rev_growth_yoy	11260	100.00%	0	0.00%	Yes - $	0.03%
coupon_used_for_payment	11260	100.00%	0	0.00%	Yes - $, *, #	0.03%
Day_Since_CC_connect	10903	96.83%	357	3.17%	Yes - $	0.01%
Cashback	10789	95.82%	471	4.18%	Yes - $	0.02%
Login_device	11039	98.04%	221	1.96%	Yes - &&&&	4.79%

Table-6 Dataset Anomalies and Null values

2.3 Exploratory Data Analysis:

Before we dive in for Univariate and Bivariate Analysis. We will take care of the dataset and fix it for better understanding. First of all we found out that there are no duplicate values. We can drop the “AccountID” variable as it served it purpose for unique identification in detecting any duplicated rows. After that we will replace all

the anomalies present in the dataset in all variables to null/empty values.

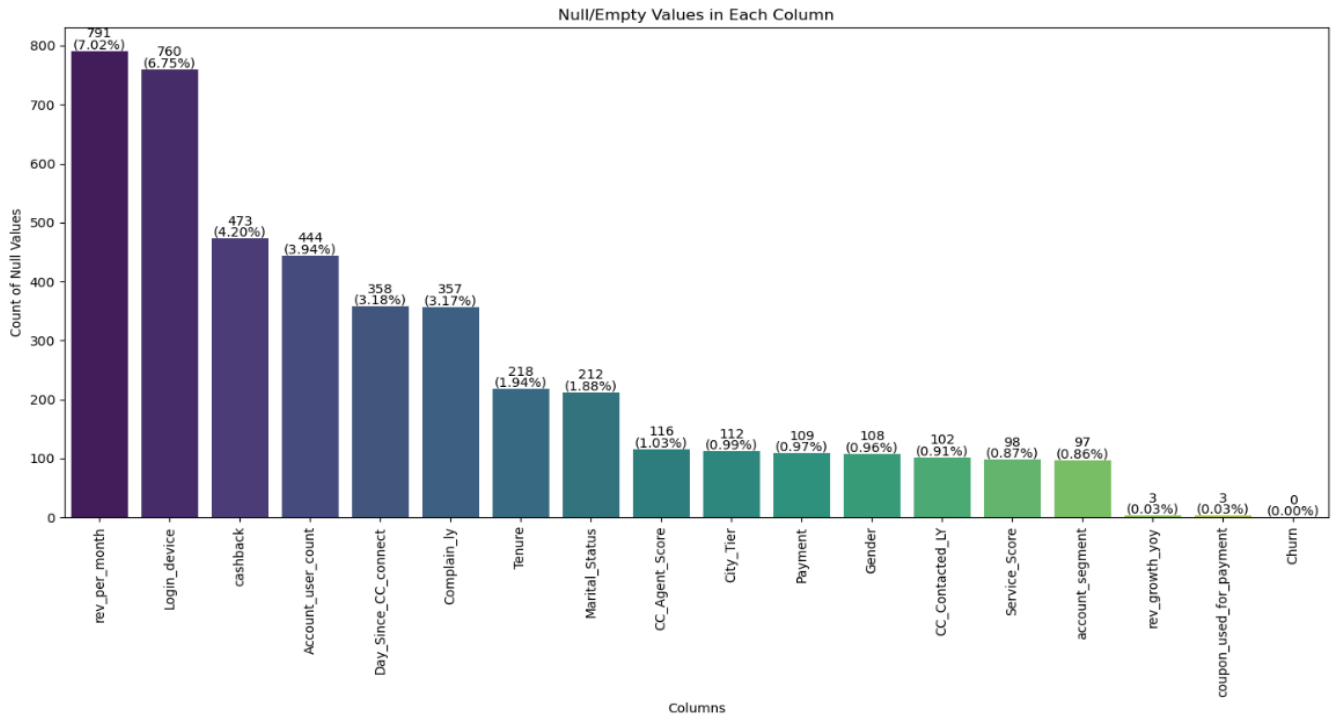


Figure-1 Null Values in dataset

As seen in the above figure all the null values are less than 30% in fact the maximum a variable has is only 7% null values. So, we can not drop any rows or columns but, instead we will impute them with mean, or median or mode values.

But, before that we will split the dataset into numerical and categorical dataset. And then we find out the skewness and kurtosis values for each numerical variables dataset and find the below tables.

Kurtosis and skewness of dataset is as below

	Kurtosis	Skewness
Tenure	23.37	3.90
CC_Contacted_LY	8.23	1.42
rev_per_month	86.96	9.09
CC_Agent_Score	-1.12	-0.14
Service_Score	-0.67	0.00
Complain_ly	-1.10	0.95
Account_user_count	0.59	-0.39
rev_growth_yoy	-0.22	0.75
coupon_used_for_payment	9.10	2.58
Day_Since_CC_connect	5.33	1.27
cashback	81.11	8.77

Table-7 Kurtosis and Skewness

As observed the dataset has skewness and therefore we will impute the null/empty values in numerical variables with the median and in categorical variables with mode and counts for each variable is shown in below table.

```
Count values for Tenure is:
Tenure
1.0     1351
0.0     1231
9.0      714
8.0      519
7.0      450
10.0     423
3.0      410
5.0      403
4.0      403
11.0     388
6.0      363
12.0     360
13.0     359
2.0      354
14.0     345
15.0     311
16.0     291
19.0     273
18.0     253
20.0     217
17.0     215
21.0     170
23.0     169
22.0     151
24.0     147
30.0     137
28.0     137
99.0     131
27.0     131
```

```
Count values for CC_Contacted_LY is:
CC_Contacted_LY
16.0     765
14.0     682
9.0      655
13.0     655
15.0     623
12.0     571
8.0      538
17.0     525
11.0     524
10.0     489
7.0      391
18.0     374
19.0     364
20.0     319
6.0      311
21.0     310
22.0     282
23.0     241
24.0     214
25.0     197
32.0     192
29.0     181
28.0     178
34.0     178
30.0     175
27.0     174
26.0     169
35.0     165
31.0     165
```

```
Count values for rev_per_month is:
rev_per_month
5.0      2128
3.0      1746
2.0      1585
4.0      1218
6.0      1085
7.0       754
8.0       643
9.0       564
10.0      413
1.0       402
11.0      278
12.0      166
13.0       93
14.0       48
15.0       24
102.0       8
124.0       5
107.0       5
123.0       5
140.0       4
118.0       4
129.0       4
133.0       4
136.0       4
117.0       3
108.0       3
101.0       3
116.0       3
110.0       3
```

```
Count values for CC_Agent_Score is:
CC_Agent_Score
3.0    3476
1.0    2302
5.0    2191
4.0    2127
2.0    1164
Name: count, dtype: int64
------------------------------------
Count values for Service_Score is:
Service_Score
3.0    5588
2.0    3251
4.0    2331
1.0      77
0.0       8
5.0       5
Name: count, dtype: int64
------------------------------------
Count values for Complain_ly is:
Complain_ly
0.0    8149
1.0    3111
Name: count, dtype: int64
```

```
Count values for Account_user_count is:
Account_user_count
4.0    5013
3.0    3261
5.0    1699
2.0     526
1.0     446
6.0     315
Name: count, dtype: int64
----------------------------------------
Count values for rev_growth_yoy is:
rev_growth_yoy
14.0    1524
13.0    1427
15.0    1286
12.0    1210
16.0     949
18.0     708
17.0     704
19.0     619
20.0     562
11.0     523
21.0     433
22.0     403
23.0     345
24.0     229
25.0     188
26.0      98
27.0      35
28.0      14
4.0        3
```

```
Count values for coupon_used_for_payment is:
coupon_used_for_payment
1.0     4376
2.0     2656
0.0     2150
3.0      698
4.0      424
5.0      284
6.0      234
7.0      184
8.0       88
9.0       34
10.0      34
11.0      30
12.0      26
13.0      22
14.0      12
15.0       4
16.0       4
```

```
Count values for Day_Since_CC_connect is:
Day_Since_CC_connect
3.0     2174
2.0     1574
1.0     1256
8.0     1169
0.0      964
7.0      911
4.0      893
9.0      622
5.0      479
10.0     339
6.0      229
11.0     183
12.0     146
13.0     117
14.0      74
15.0      37
17.0      34
16.0      26
18.0      26
30.0       2
31.0       2
47.0       2
46.0       1
Name: count, dtype: int64
```

```
Count values for cashback is:
cashback
165.25    478
155.62     10
145.08      9
149.36      9
154.73      9
          ...
131.55      1
245.64      1
130.78      1
299.72      1
191.42      1
Name: count, Length: 5692, dtype: int64
----------------------------------------
Count values for Churn is:
Churn
0    9364
1    1896
Name: count, dtype: int64
----------------------------------------
Count values for City_Tier is:
City_Tier
1.0    7375
3.0    3405
2.0     480
Name: count, dtype: int64
```

```
Count values for Payment is:
Payment
Debit Card          4696
Credit Card         3511
E wallet            1217
Cash on Delivery    1014
UPI                  822
Name: count, dtype: int64
-----------------------------------
Count values for Gender is:
Gender
Male      6812
Female    4448
Name: count, dtype: int64
-----------------------------------
Count values for account_segment is:
account_segment
Regular Plus    4221
Super           4062
HNI             1639
Super Plus       818
Regular          520
Name: count, dtype: int64
-----------------------------------
Count values for Marital_Status is:
Marital_Status
Married     6072
Single      3520
Divorced    1668
Name: count, dtype: int64
```

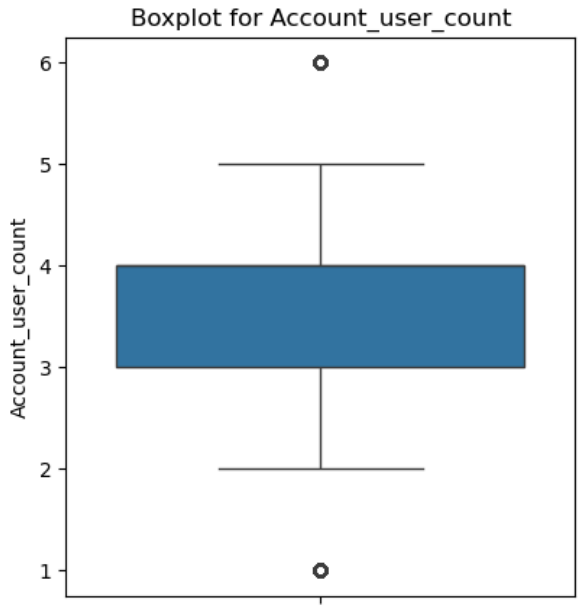
```
Count values for Login_device is:
Login_device
Mobile      8242
Computer    3018
Name: count, dtype: int64
```

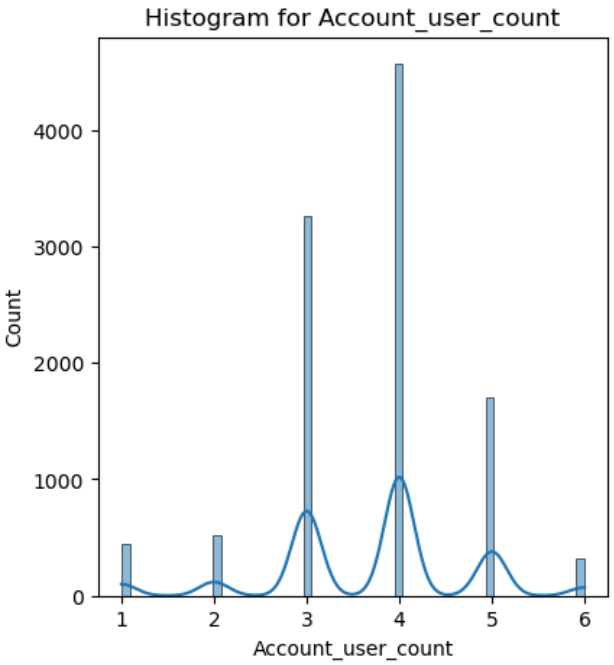
Table-8 Value Counts for Variables.

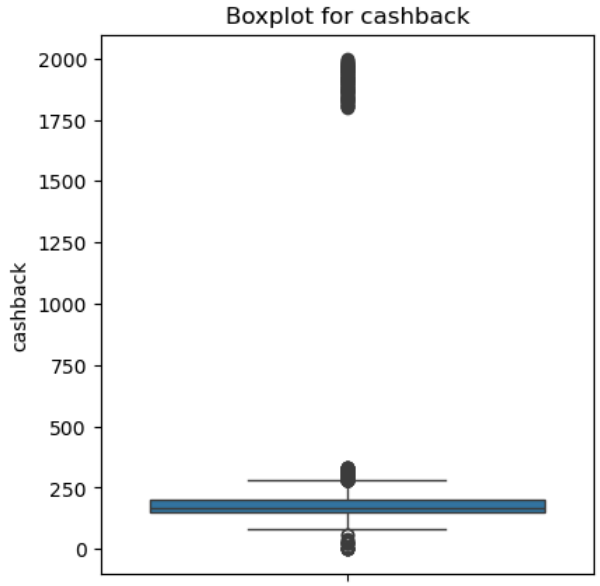
From the above count table we can say that all the data of the variables have been fixed and we also can see that some variables have biases towards one value which indicates data imbalance.

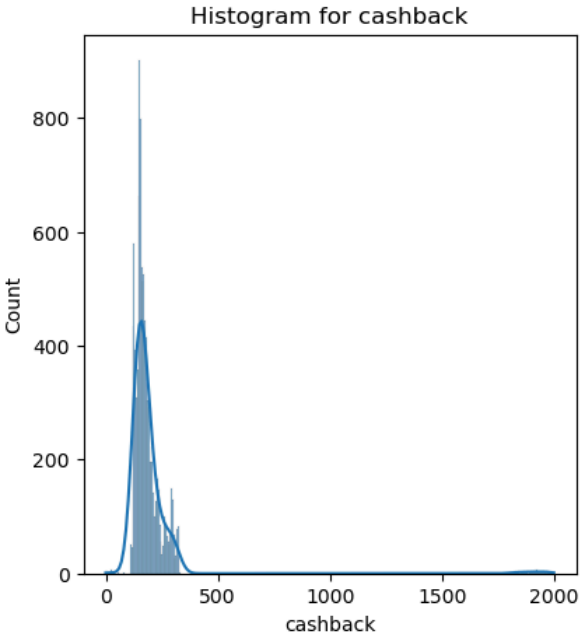
2.3.1 Outlier Treatment:

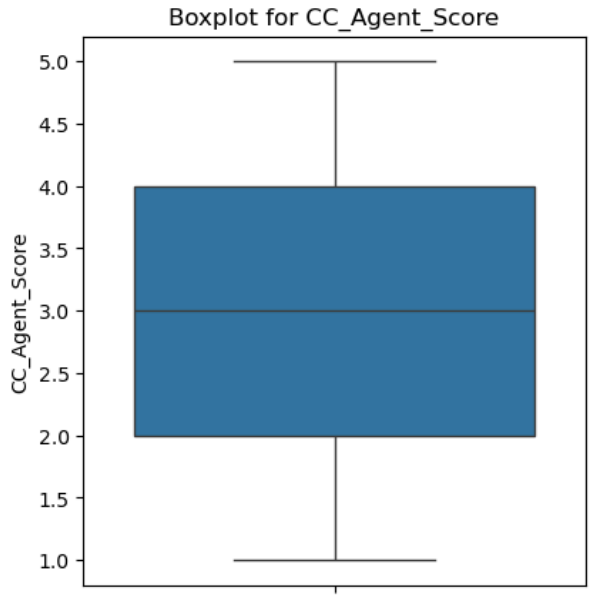
In figure-2 below we can see the boxplot and histogram for numerical variables before outlier treatment.

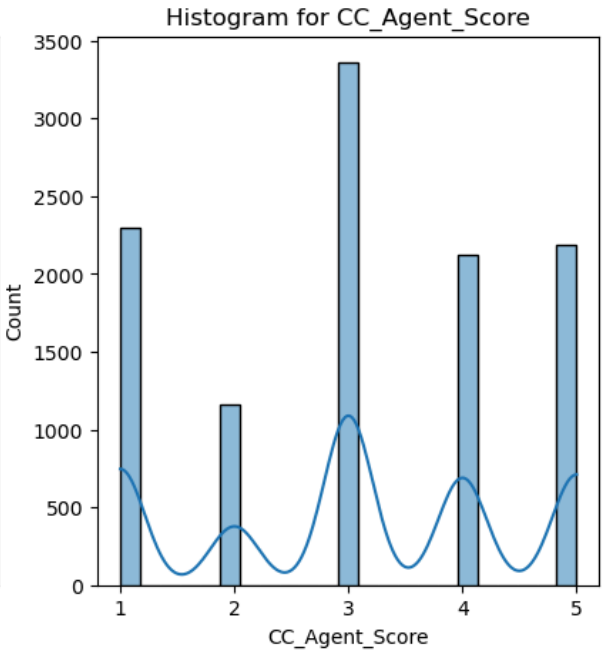

Boxplot for Account_user_count
Account_user_count
1
2
3
4
5
6

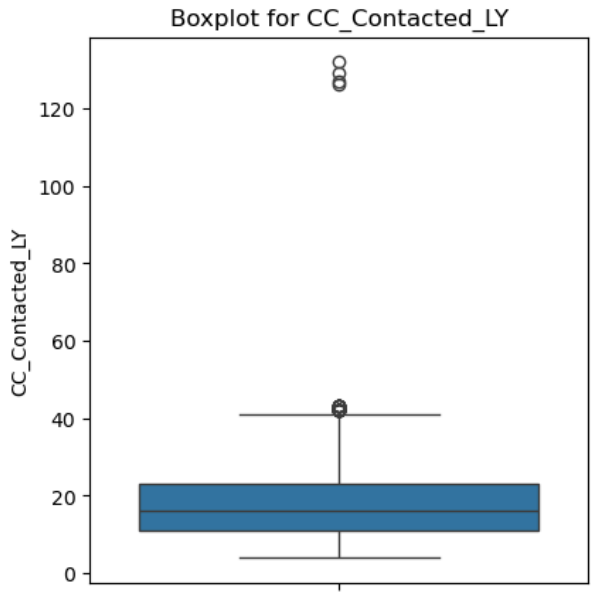

Histogram for Account_user_count
Count
0
1000
2000
3000
4000
1
2
3
4
5
6
Account_user_count

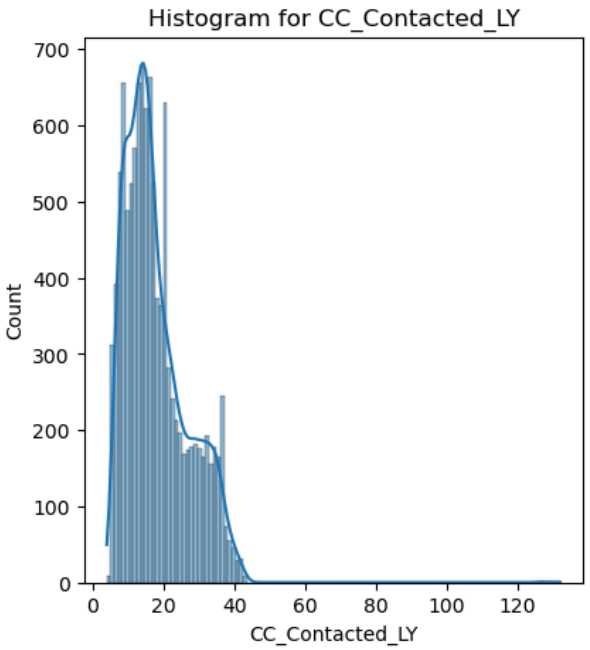

Boxplot for cashback
cashback
0
250
500
750
1000
1250
1500
1750
2000

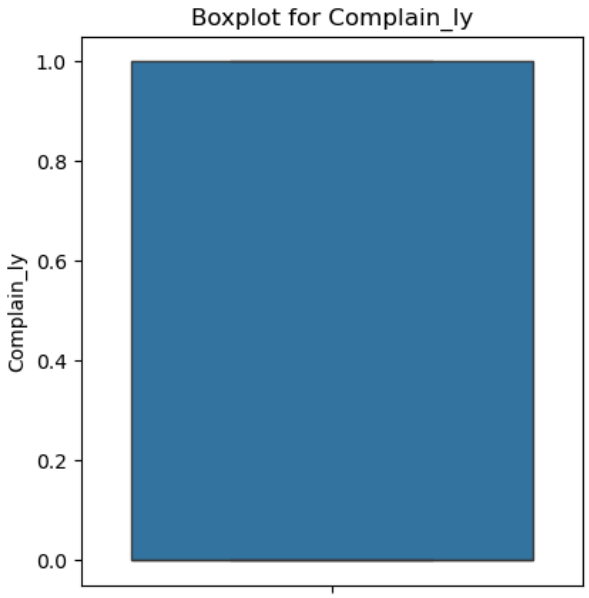

Histogram for cashback
Count
0
200
400
600
800
0
500
1000
1500
2000
cashback

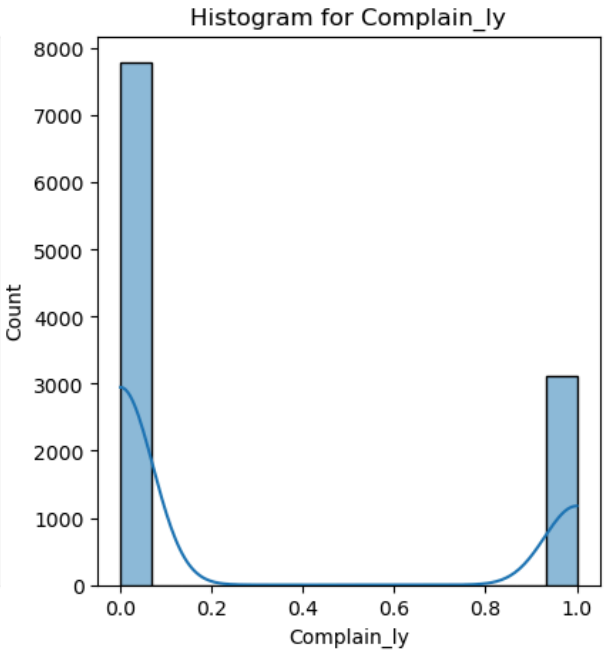

Boxplot for CC_Agent_Score
CC_Agent_Score
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0

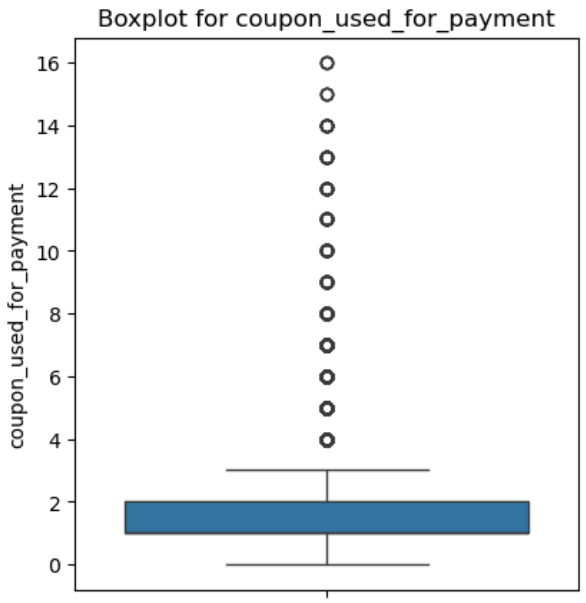

Histogram for CC_Agent_Score
Count
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
CC_Agent_Score

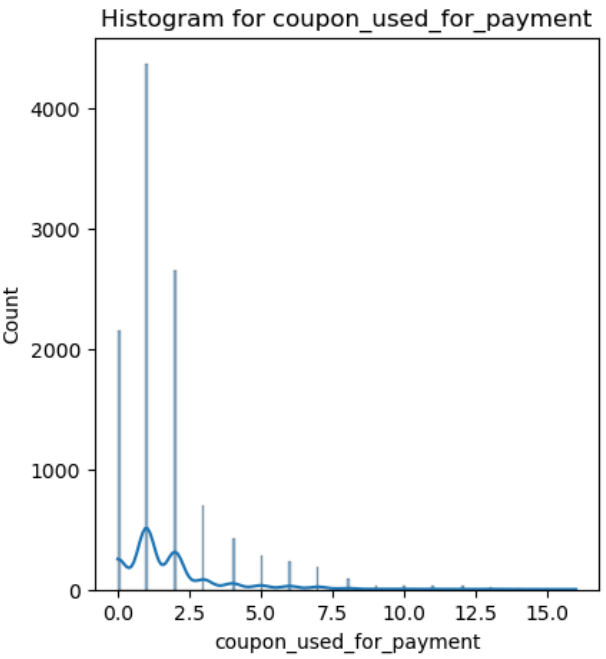

Boxplot for CC_Contacted_LY
CC_Contacted_LY
120
100
80
60
40
20
0

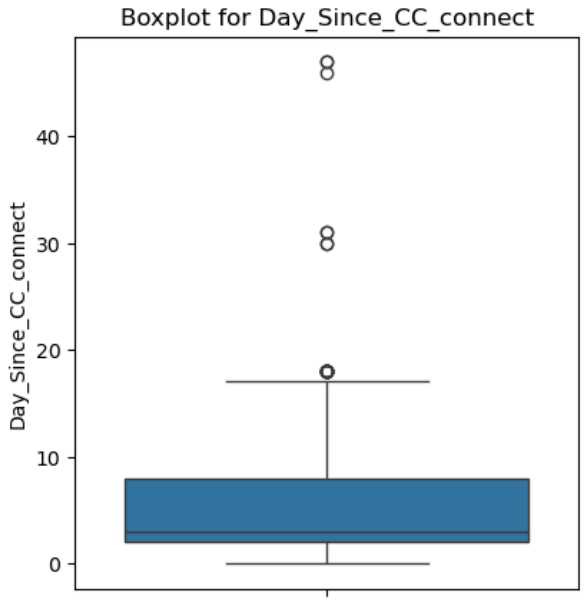

Histogram for CC_Contacted_LY
Count
700
600
500
400
300
200
100
0
0
20
40
60
80
100
120
CC_Contacted_LY

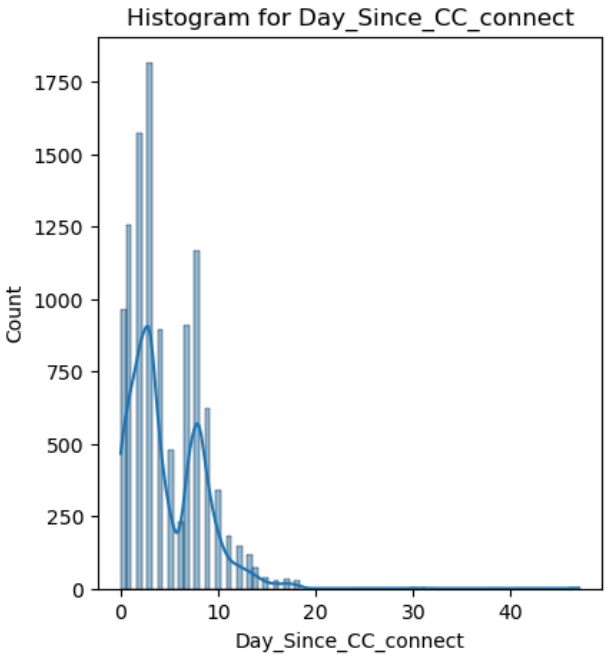

Boxplot for Complain_ly
Complain_ly
1.0
0.8
0.6
0.4
0.2
0.0

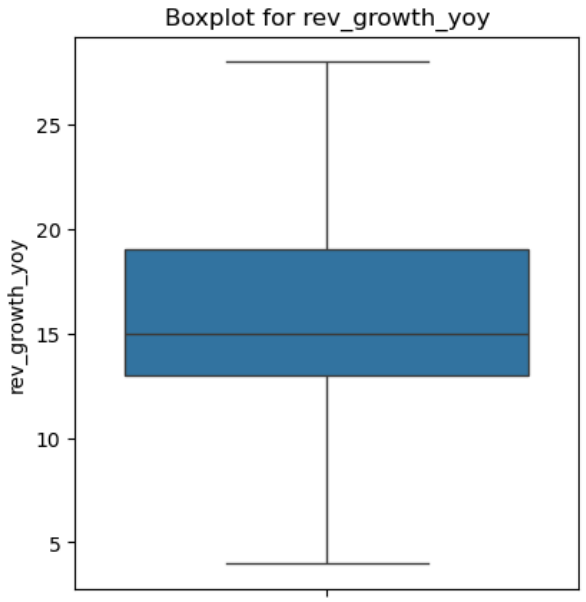

Histogram for Complain_ly
Count
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
Complain_ly

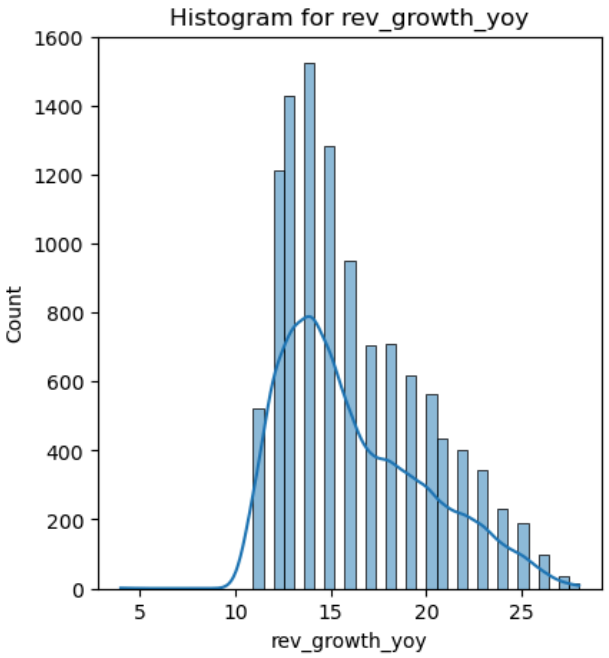

Boxplot for coupon_used_for_payment
coupon_used_for_payment
16
14
12
10
8
6
4
2
0

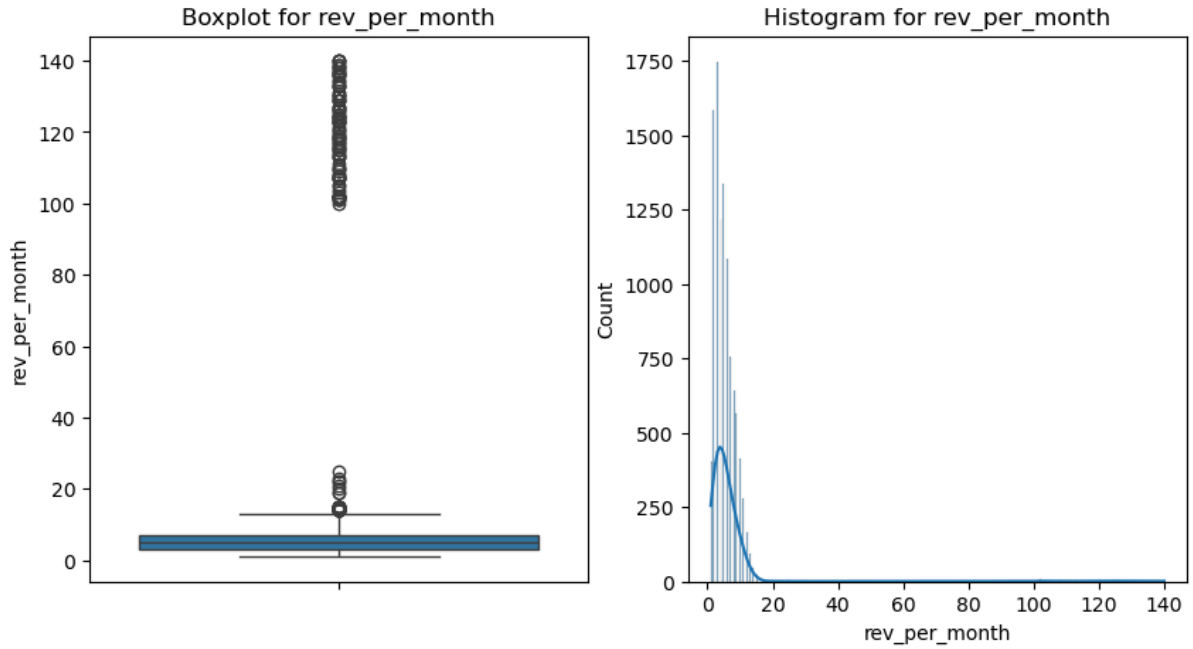

Histogram for coupon_used_for_payment
Count
4000
3000
2000
1000
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
coupon_used_for_payment


Boxplot for Day_Since_CC_connect
Day_Since_CC_connect
0
10
20
30
40


Histogram for Day_Since_CC_connect
Count
0
250
500
750
1000
1250
1500
1750
0
10
20
30
40
Day_Since_CC_connect


Boxplot for rev_growth_yoy
rev_growth_yoy
5
10
15
20
25


Histogram for rev_growth_yoy
Count
0
200
400
600
800
1000
1200
1400
1600
5
10
15
20
25
rev_growth_yoy


Boxplot for rev_per_month
rev_per_month
0
20
40
60
80
100
120
140
Histogram for rev_per_month
Count
0
250
500
750
1000
1250
1500
1750
0
20
40
60
80
100
120
140
rev_per_month

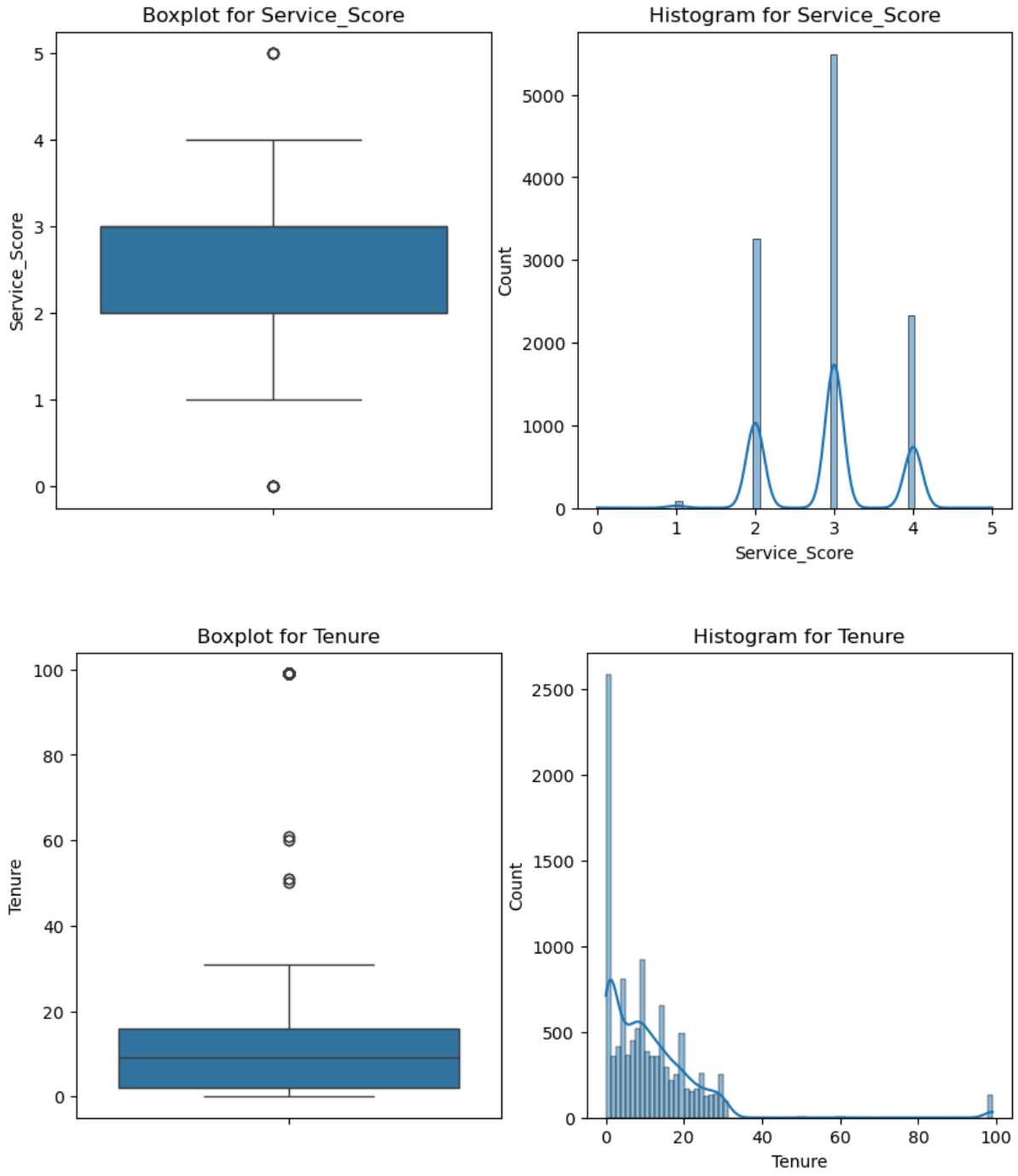


Figure-2 Before Outlier Treatment

Some variables have outliers closer to the whisker, while others have extreme outliers with no intermediate values. For example, the "rev_per_month" variable shows a significant gap between 30 and 100, indicating a lack of values in that

range. These extreme outliers do not correlate with outliers in the "cashback" field. While we cannot dismiss these outliers as incorrect, as they may belong to accounts with high activity, models like logistic regression are sensitive to outliers and may underperform if these are left untreated.

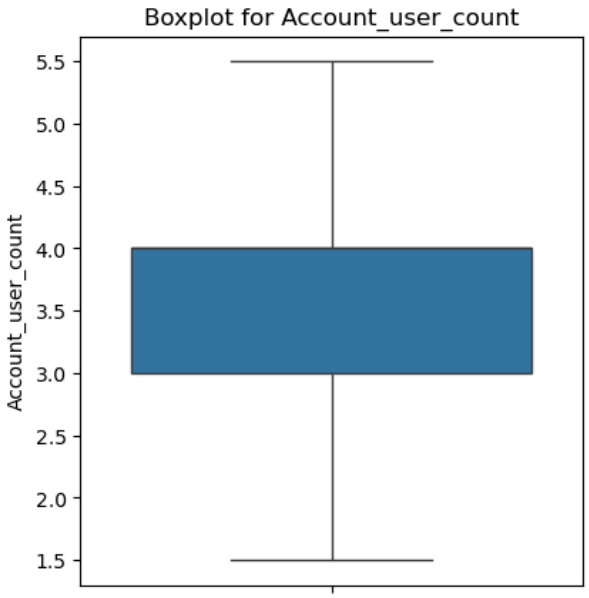
Therefore, two approaches to modeling were undertaken:

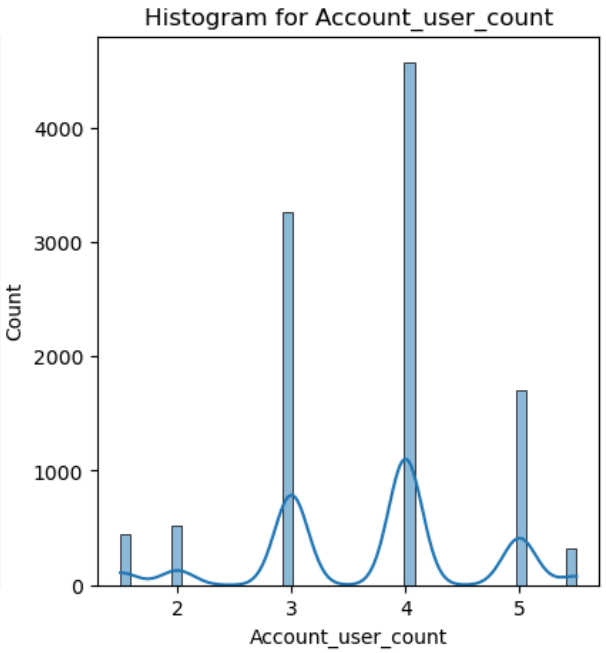
- One dataset with outliers treated for outlier-sensitive models.
- Another dataset with outliers left untreated for outlier-resistant algorithms such as Random Forest, and during the tuning/trial phase of other algorithms.

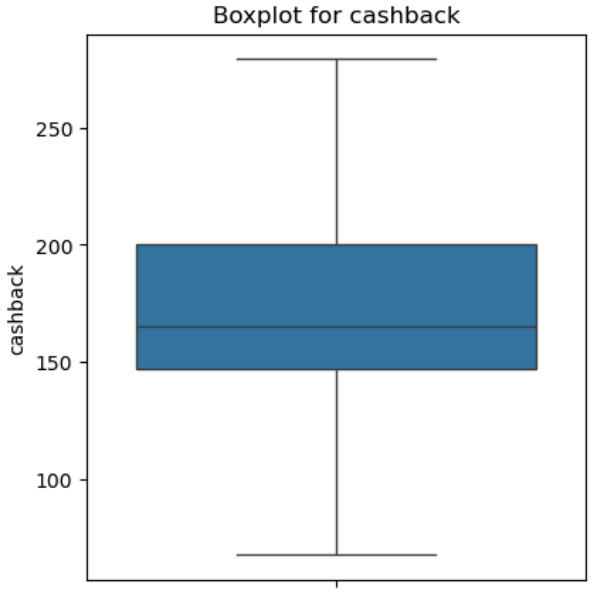
The "Coupon_used_for_payment" variable has a limited range from 0 to 16. Hence, for this analysis, outliers in this variable will not be treated, similar to categorical variables.

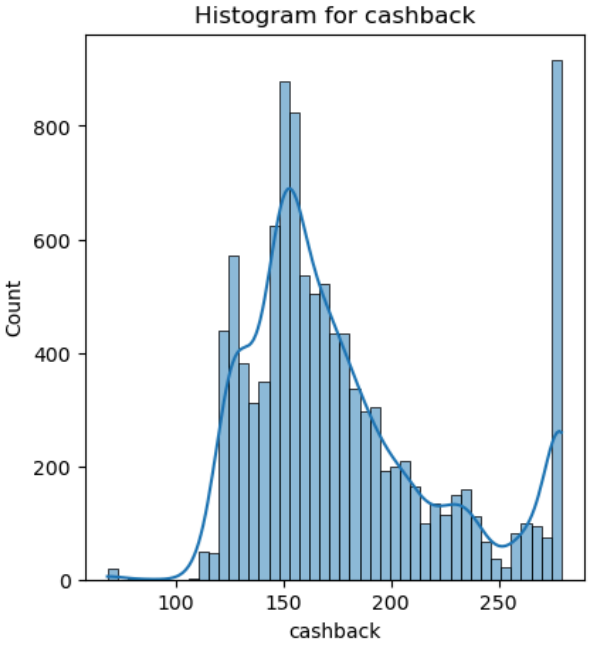
For the dataset where outliers are treated, values beyond the upper and lower whiskers were capped using the following method:

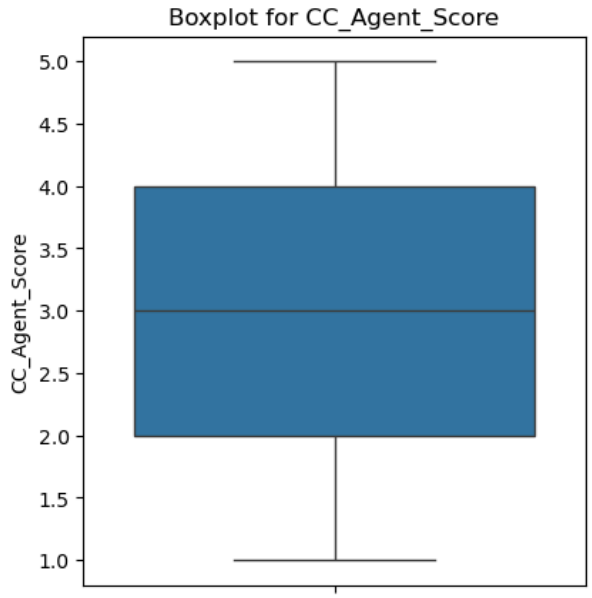
- **Lower Range:** 1st quartile - (1.5 * IQR)
- **Upper Range:** 3rd quartile + (1.5 * IQR)
- Where IQR = 3rd quartile - 1st quartile value.

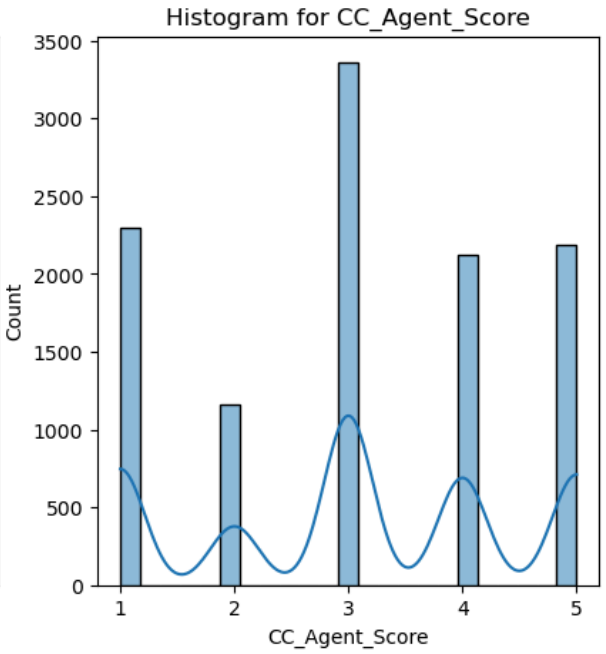

Boxplot for Account_user_count
Account_user_count
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5

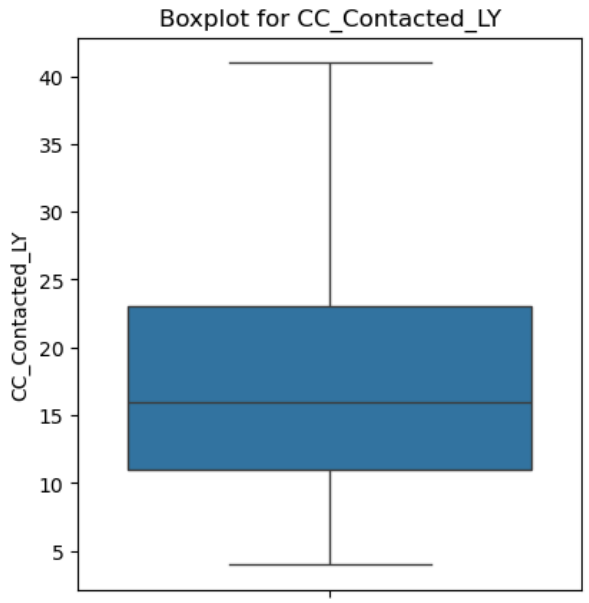

Histogram for Account_user_count
Count
4000
3000
2000
1000
0
2
3
4
5
Account_user_count

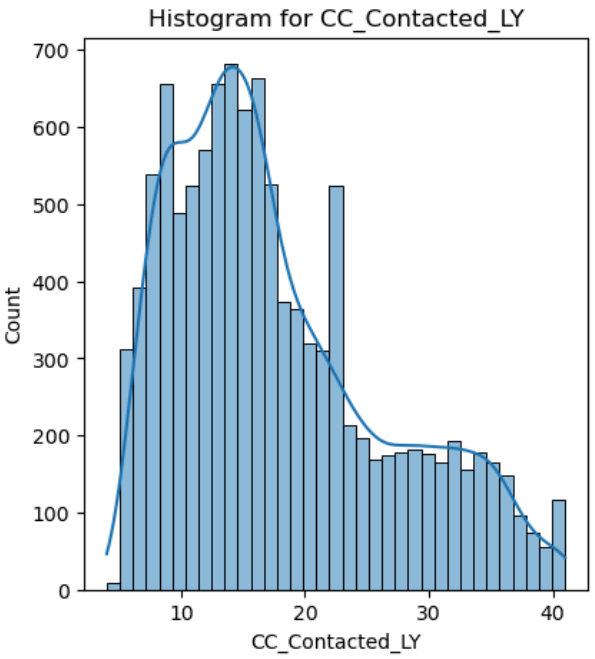

Boxplot for cashback
cashback
250
200
150
100

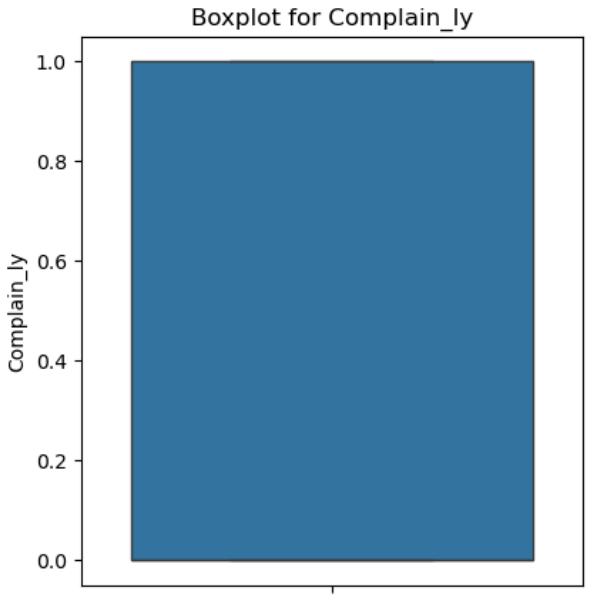

Histogram for cashback
Count
800
600
400
200
0
100
150
200
250
cashback

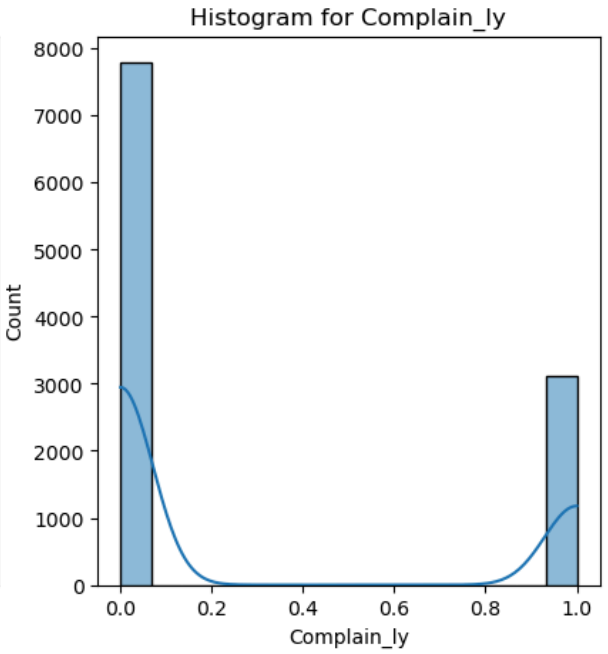

Boxplot for CC_Agent_Score
CC_Agent_Score
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0

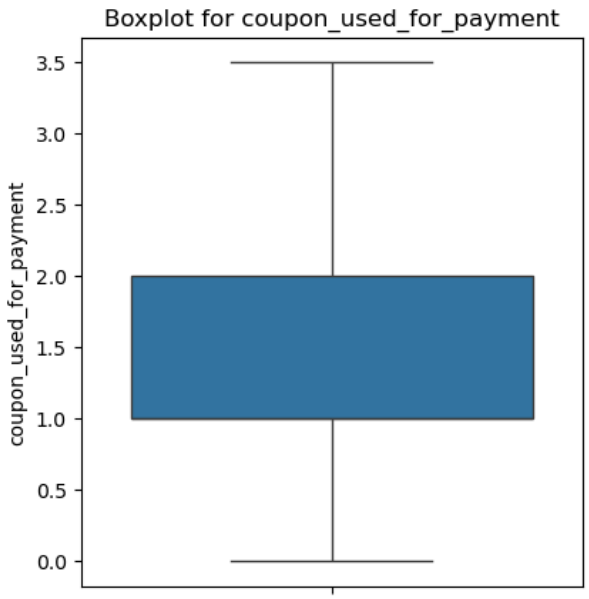

Histogram for CC_Agent_Score
Count
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
CC_Agent_Score

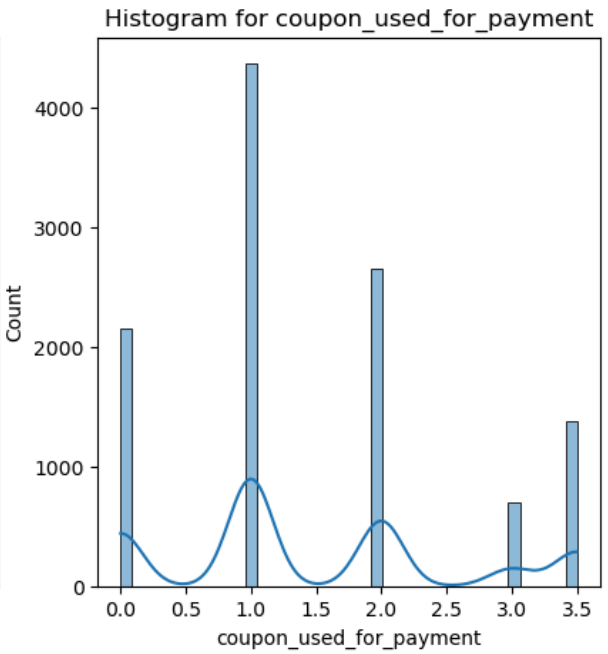

Boxplot for CC_Contacted_LY
CC_Contacted_LY
40
35
30
25
20
15
10
5

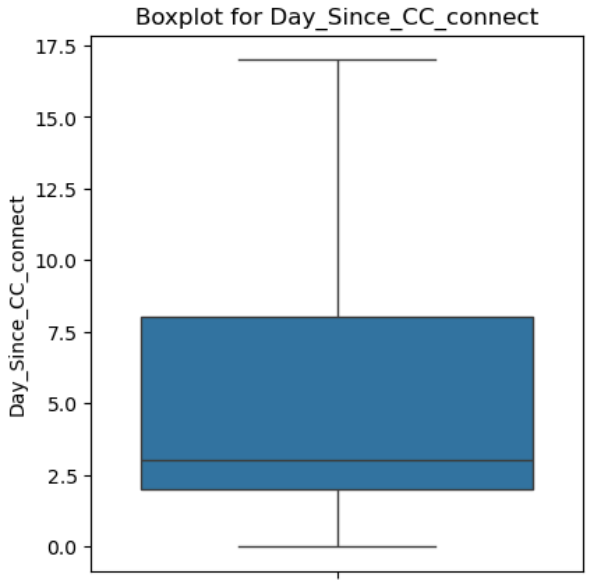

Histogram for CC_Contacted_LY
Count
700
600
500
400
300
200
100
0
10
20
30
40
CC_Contacted_LY

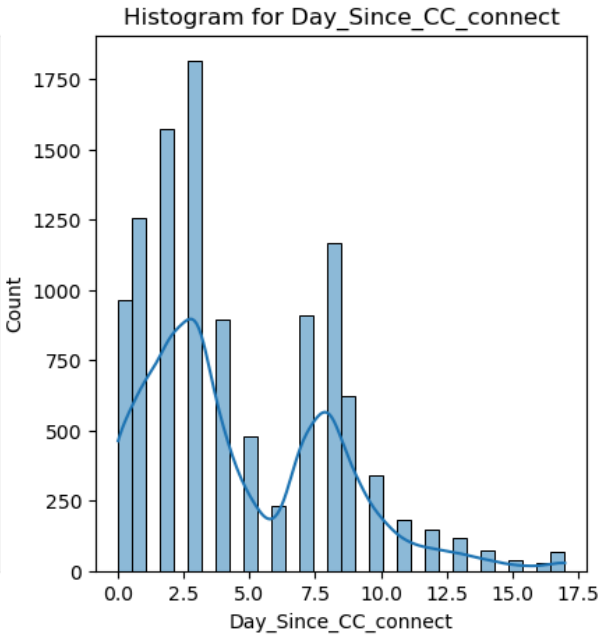

Boxplot for Complain_ly
Complain_ly
1.0
0.8
0.6
0.4
0.2
0.0

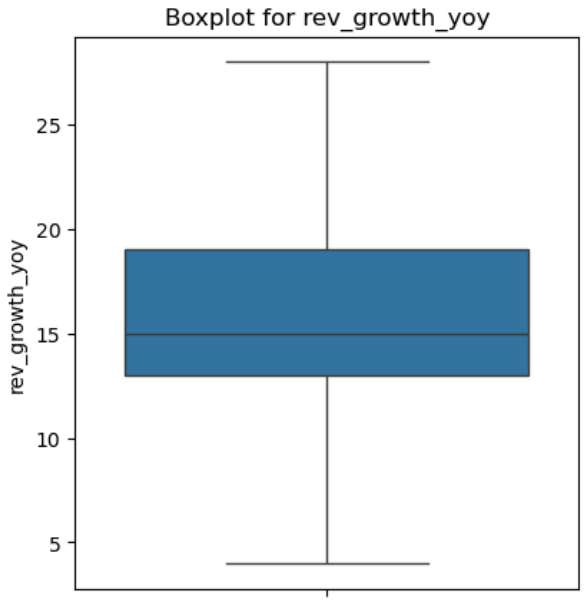

Histogram for Complain_ly
Count
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
Complain_ly

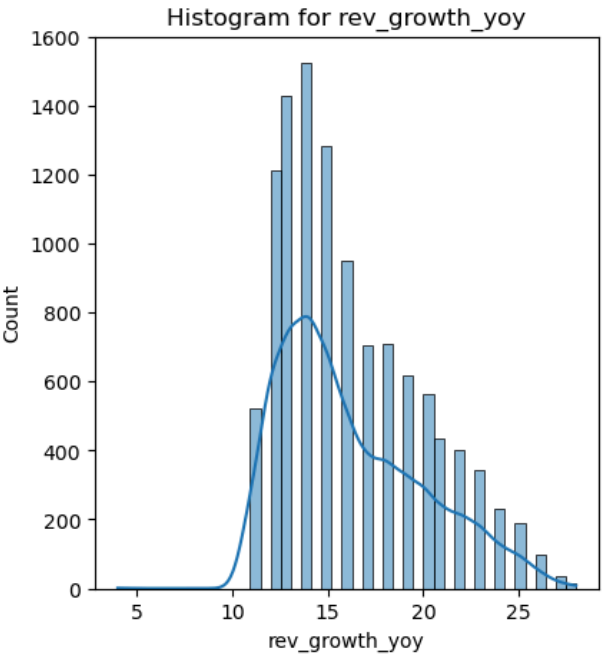

Boxplot for coupon_used_for_payment
coupon_used_for_payment
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

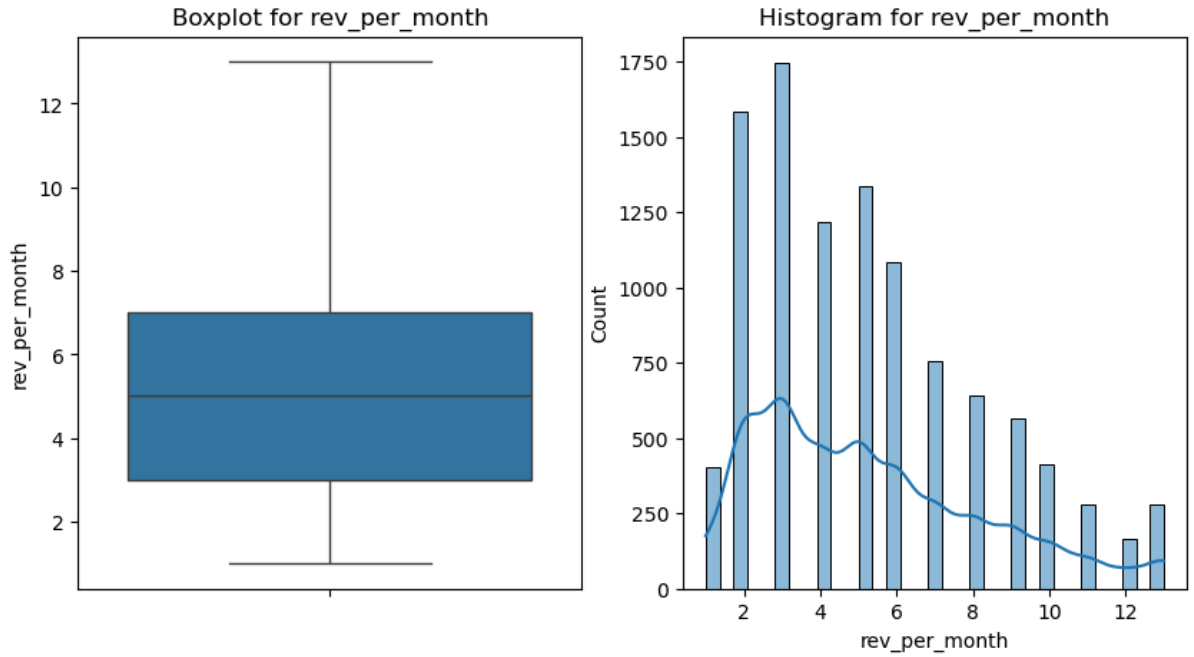

Histogram for coupon_used_for_payment
Count
4000
3000
2000
1000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
coupon_used_for_payment


Boxplot for Day_Since_CC_connect
Day_Since_CC_connect
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0


Histogram for Day_Since_CC_connect
Count
1750
1500
1250
1000
750
500
250
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Day_Since_CC_connect


Boxplot for rev_growth_yoy
rev_growth_yoy
25
20
15
10
5


Histogram for rev_growth_yoy
Count
1600
1400
1200
1000
800
600
400
200
0
5
10
15
20
25
rev_growth_yoy


Boxplot for rev_per_month
rev_per_month
12
10
8
6
4
2
Histogram for rev_per_month
Count
1750
1500
1250
1000
750
500
250
0
2
4
6
8
10
12
rev_per_month

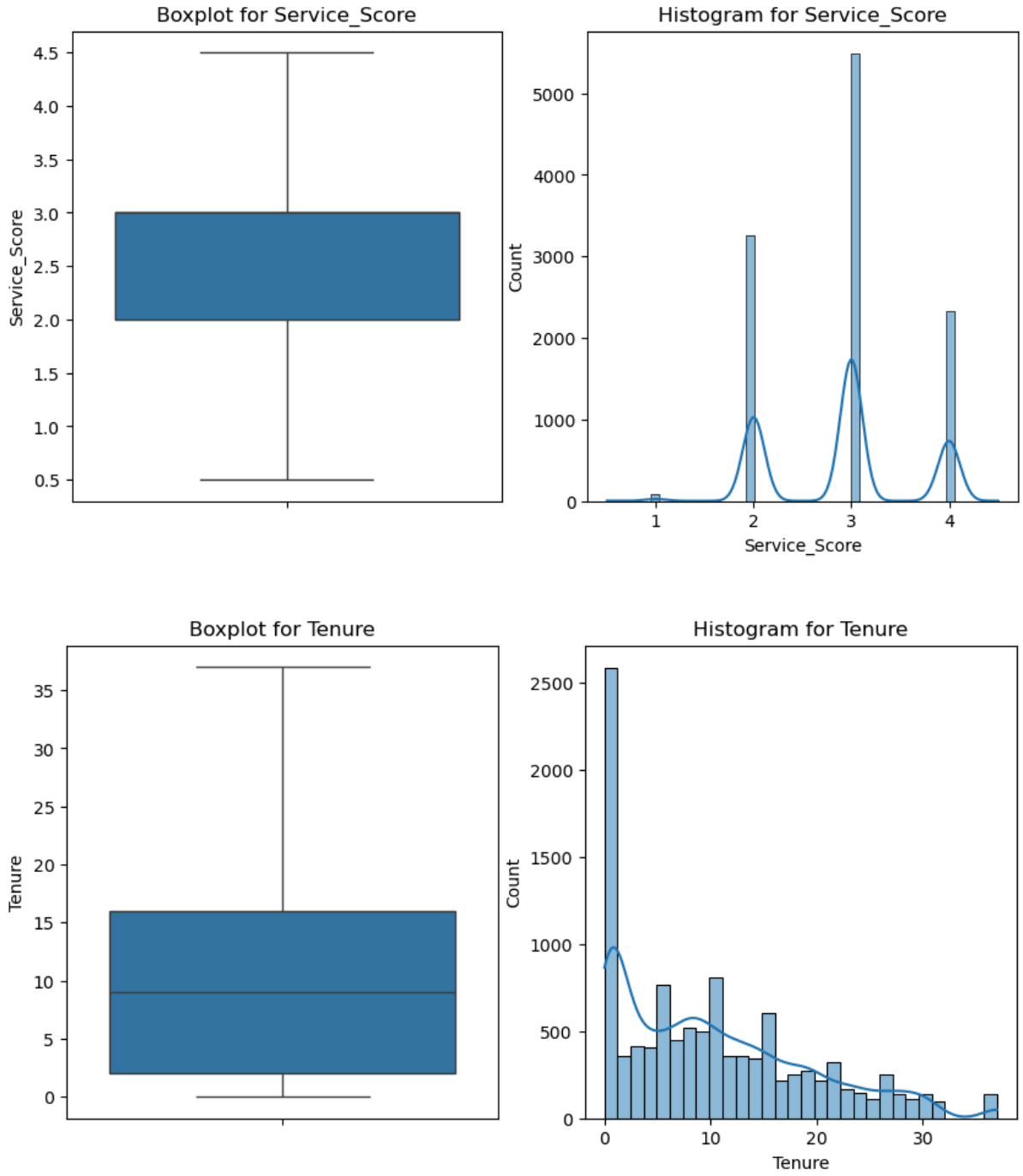


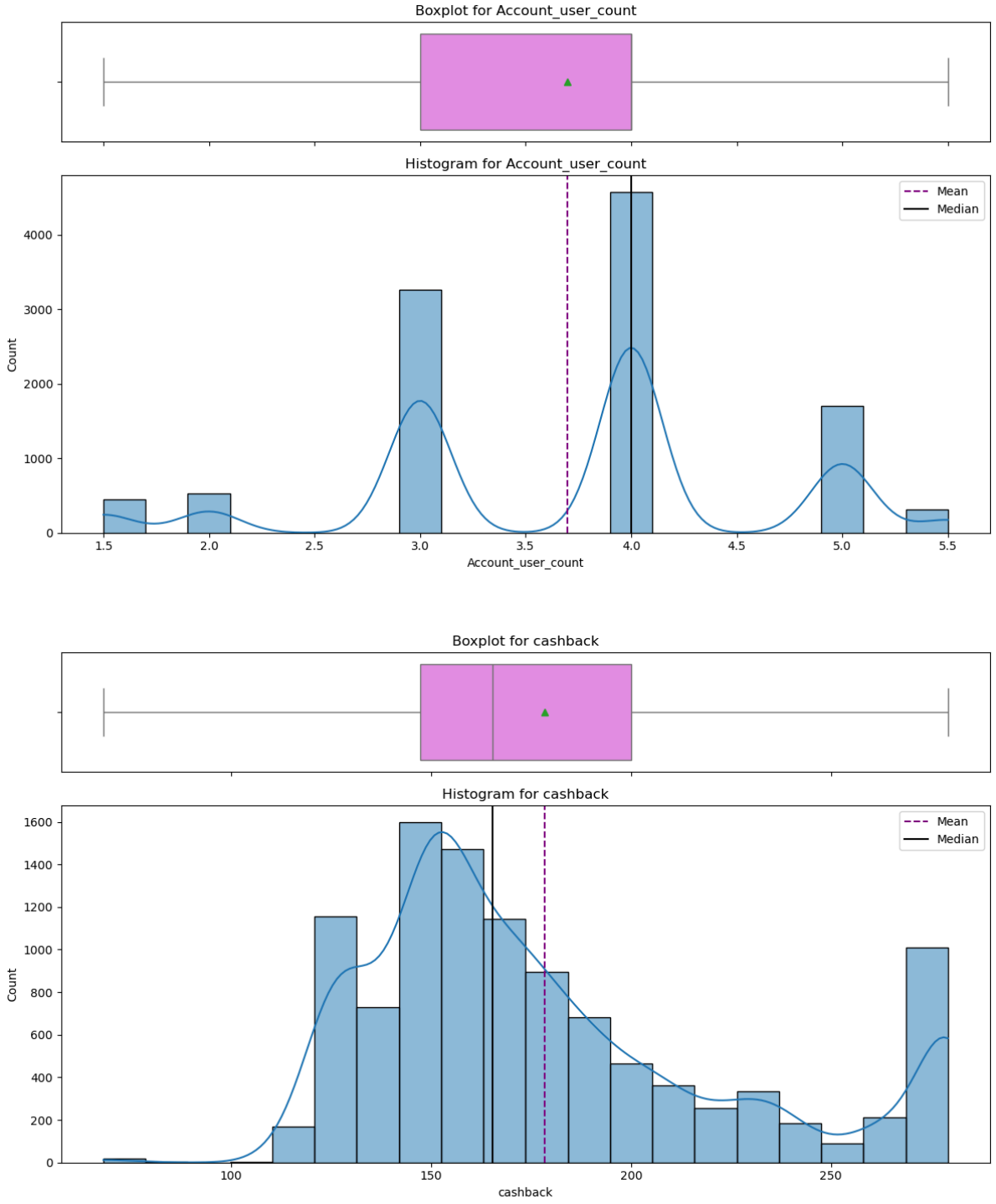
Figure-3 After Outlier Treatment

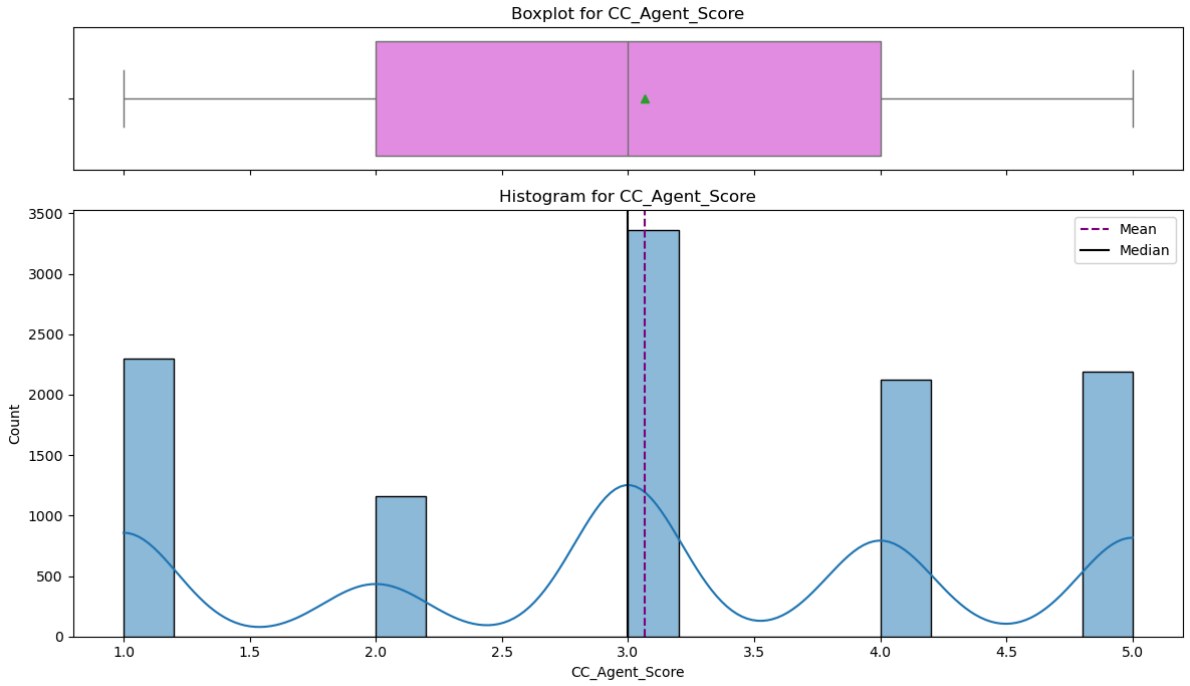
Now we will perform Univariate and Bivariate Analysis.

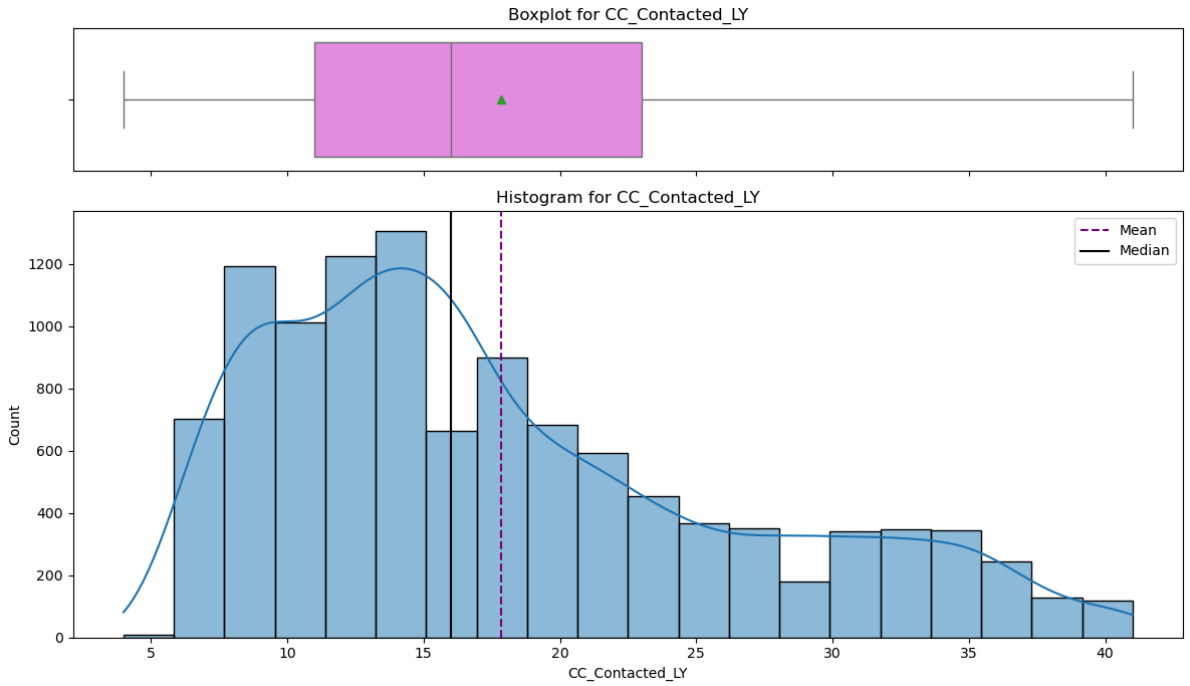
2.3.2 Univariate Analysis:

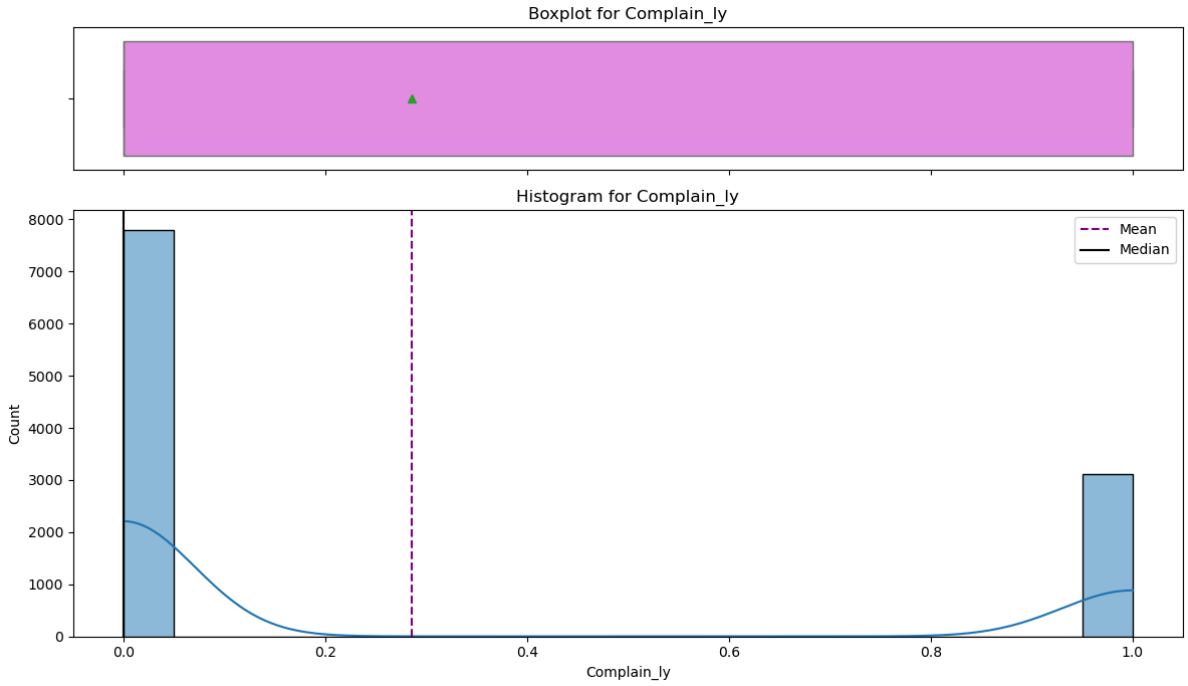
Univariate Analysis of Numerical Variables.

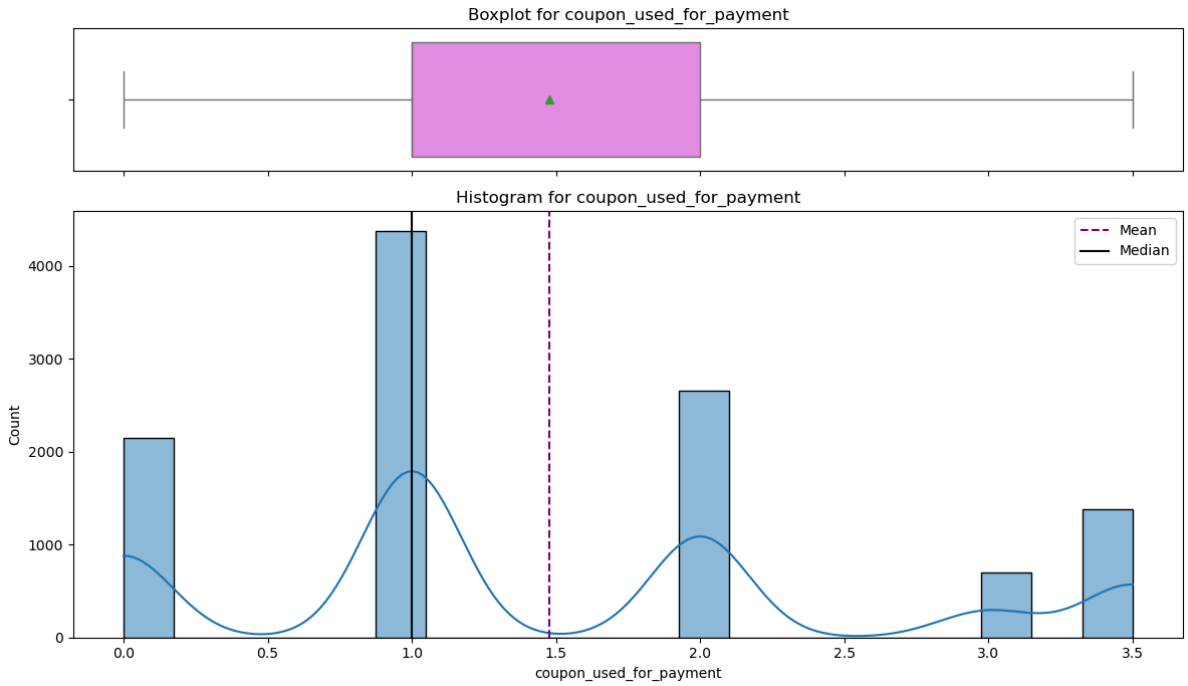
The below figure-4 is for the Univariate Analysis of Numerical Variables.

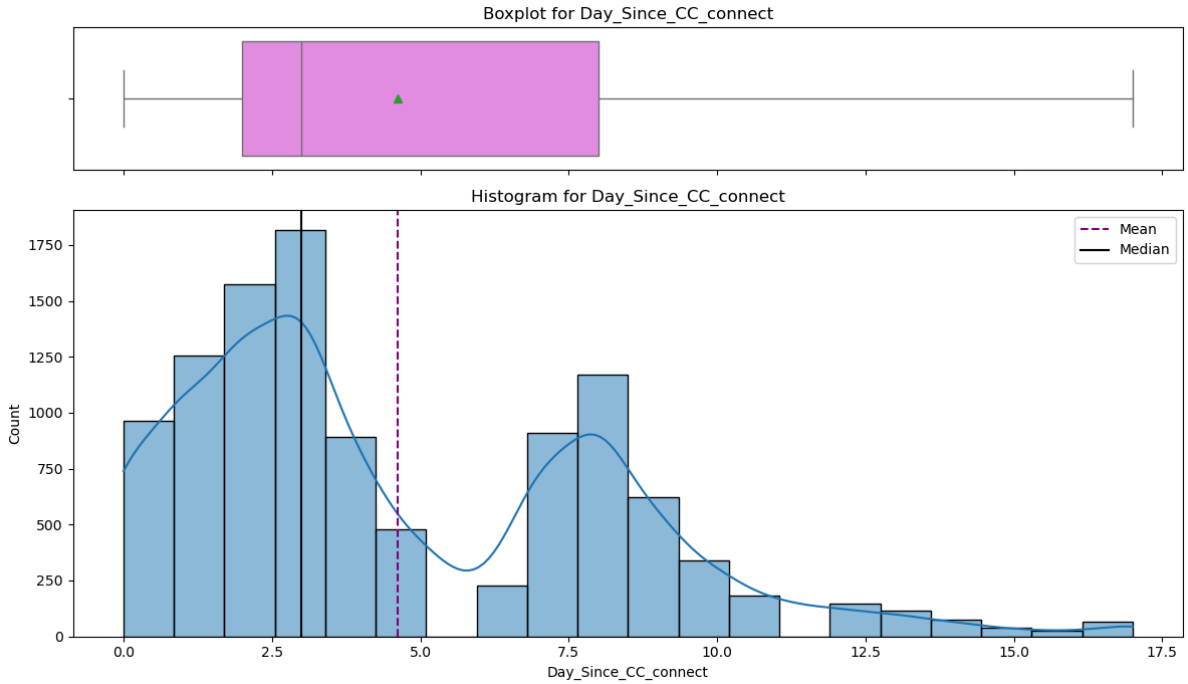


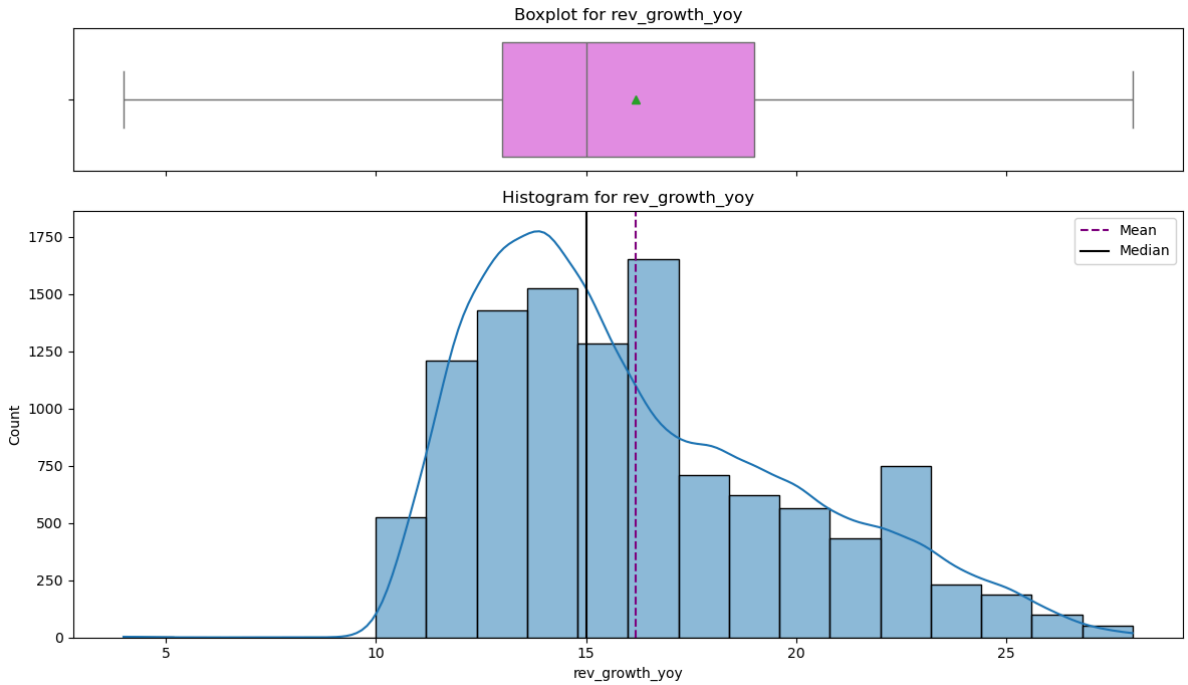

Boxplot for CC_Agent_Score
Histogram for CC_Agent_Score
Mean
Median
Count
CC_Agent_Score
0
500
1000
1500
2000
2500
3000
3500
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

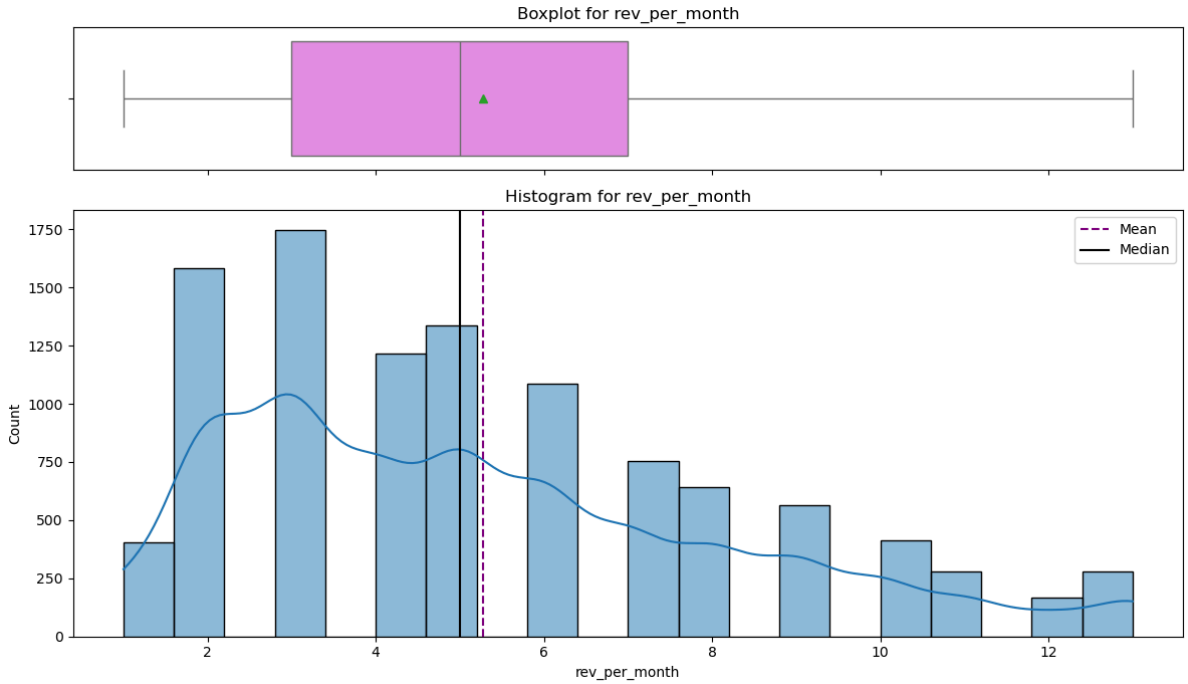

Boxplot for CC_Contacted_LY
Histogram for CC_Contacted_LY
Mean
Median
Count
CC_Contacted_LY
0
200
400
600
800
1000
1200
5
10
15
20
25
30
35
40


Boxplot for Complain_ly
Histogram for Complain_ly
Mean
Median
Count
Complain_ly
0
1000
2000
3000
4000
5000
6000
7000
8000
0.0
0.2
0.4
0.6
0.8
1.0


Boxplot for coupon_used_for_payment
Histogram for coupon_used_for_payment
Mean
Median
Count
coupon_used_for_payment
0
1000
2000
3000
4000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


Boxplot for Day_Since_CC_connect
Histogram for Day_Since_CC_connect
Mean
Median
Count
Day_Since_CC_connect
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0
250
500
750
1000
1250
1500
1750


Boxplot for rev_growth_yoy
Histogram for rev_growth_yoy
Mean
Median
Count
rev_growth_yoy
5
10
15
20
25
0
250
500
750
1000
1250
1500
1750


Boxplot for rev_per_month
Histogram for rev_per_month
Mean
Median
Count
rev_per_month
0
250
500
750
1000
1250
1500
1750
2
4
6
8
10
12

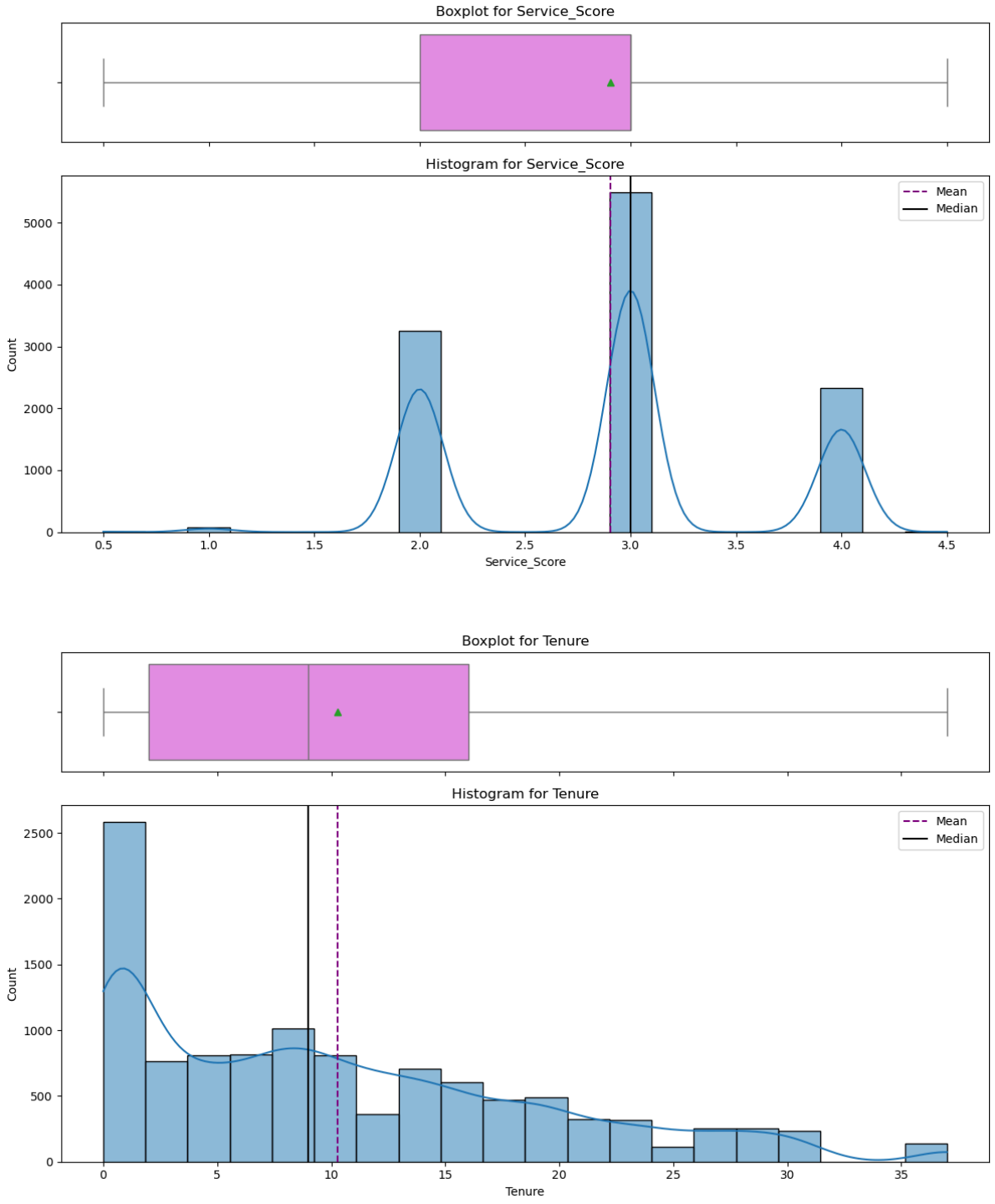


Figure-4 Univariate Analysis for Numerical Variables.

Observation:

Tenure

- **Median:** 9 months
- **Observation:** 50% of the tenure data is less than 9 months.
- **Insights:** Most customers have relatively short tenure. There are a few outliers indicating long-tenure customers who might be loyal and valuable.

CC_Contacted_LY

- **Median:** 16
- **Observation:** Represents typical customer care contacts in a year.
- **Insights:** Most interactions range from 0 to 50. Indicates robust customer support. High contact frequency may indicate issues.

Rev_per_month

- **Box Plot Observation:**
    - 75% of accounts generate less than INR 10,000 monthly.
    - Outliers show higher revenue, potentially from multi-account customers.
- **Insights:** Majority of accounts have moderate revenue. Identifying high-revenue outliers can help target high-value customers.

Day_Since_CC_connect

- **Median:** 3 days
- **Observation:** Indicates time since last customer care contact.
- **Insights:** 50% of cases have swift resolution within 3 days. Some instances show delays up to 47 days.

Cashback

- **Median:** INR 165.25
- **Observation:** Monthly average given in the past year.
- **Insights:** Typical cashback, but outliers indicate higher amounts for specific accounts, possibly high-spending customers.

Rev_growth_yoy

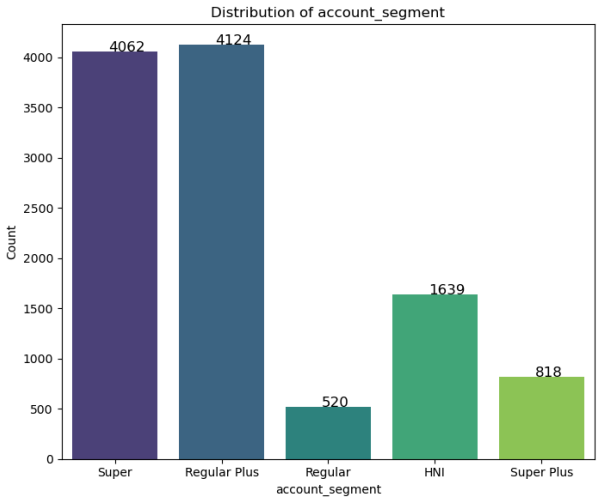
- **Median:** Under 15%
- **Observation:** Revenue growth percentage.
- **Insights:** Moderate expansion for many accounts. Exceptional cases show up to 28% growth.

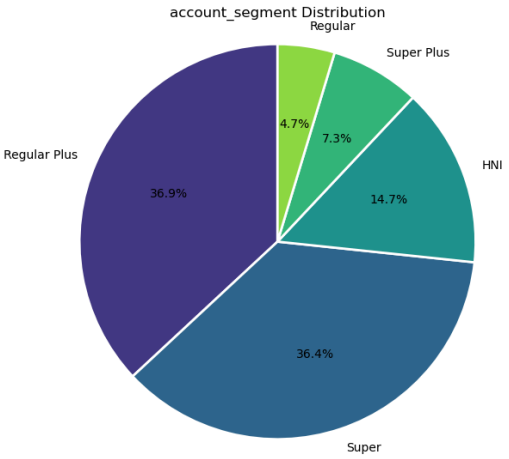
Coupon_used_for_payment

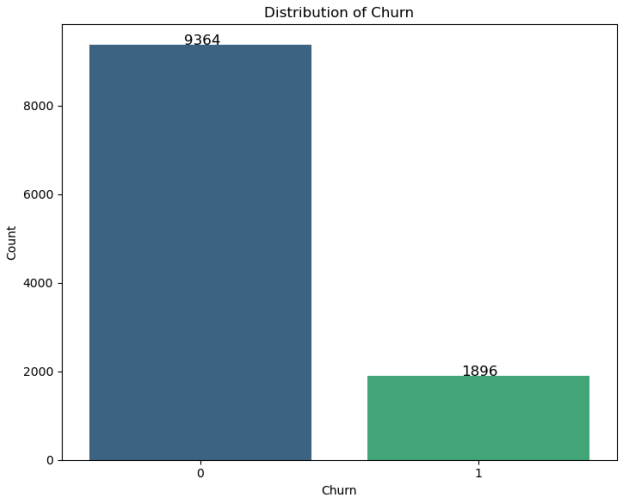
- **Median:** 1
- **Observation:** Coupon usage behavior.
- **Insights:** Most customers use coupons sparingly. Outliers maxing at 14 uses highlight heavy coupon users.

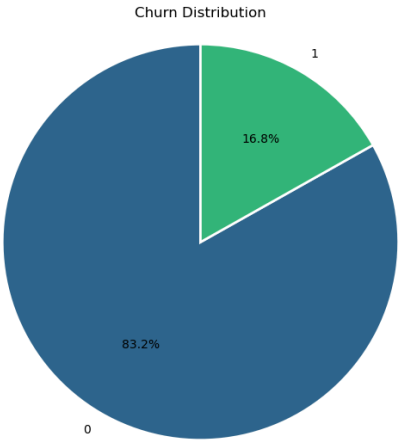
Univariate Analysis of Categorical variables:

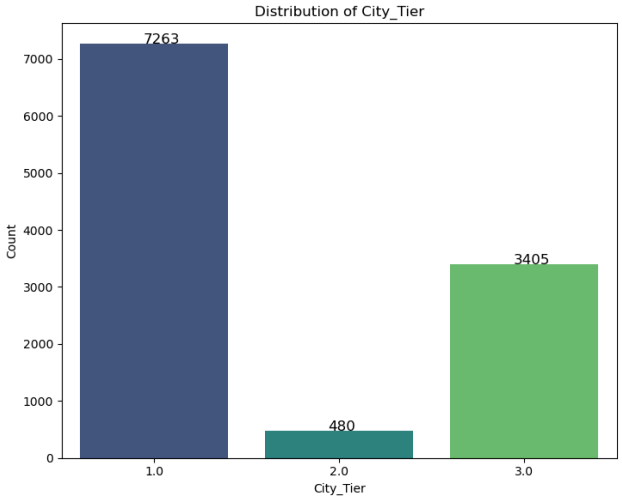
Now, the figure-5 gives the Univariate Analysis of Categorical variables.

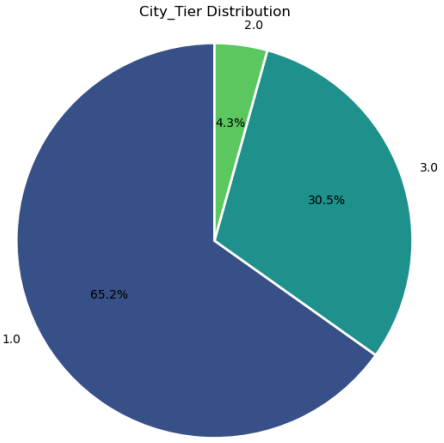

Distribution of account_segment
4062
4124
520
1639
818
Count
account_segment
Super
Regular Plus
Regular
HNI
Super Plus

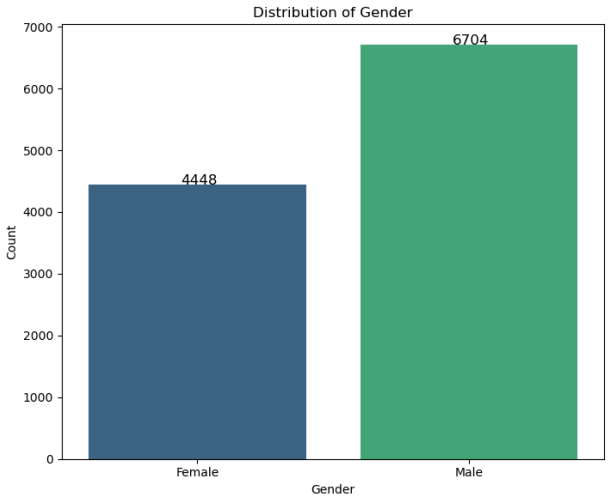

account_segment Distribution
Regular
Super Plus
HNI
Regular Plus
Super
4.7%
7.3%
14.7%
36.9%
36.4%

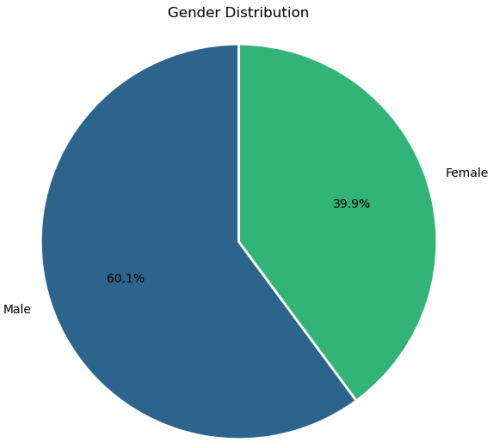

Distribution of Churn
9364
1896
Count
Churn

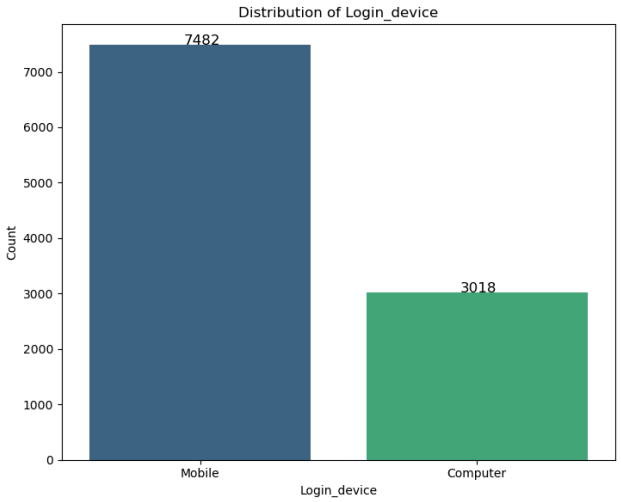

Churn Distribution
1
0
16.8%
83.2%

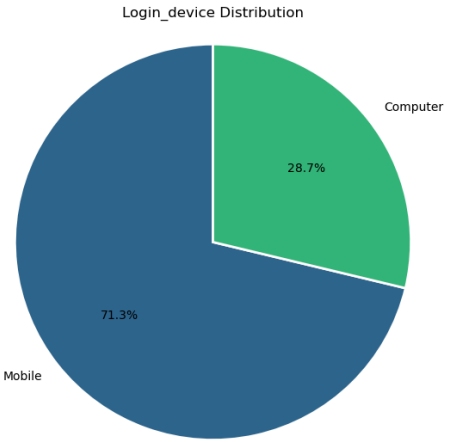

Distribution of City_Tier
7263
480
3405
Count
City_Tier


City_Tier Distribution
2.0
3.0
1.0
4.3%
30.5%
65.2%


Distribution of Gender
7000
6000
5000
4000
3000
2000
1000
0
Count
4448
6704
Female
Male
Gender


Gender Distribution
Female
39.9%
60.1%
Male


Distribution of Login_device
7000
6000
5000
4000
3000
2000
1000
0
Count
7482
3018
Mobile
Computer
Login_device


Login_device Distribution
Computer
28.7%
71.3%
Mobile

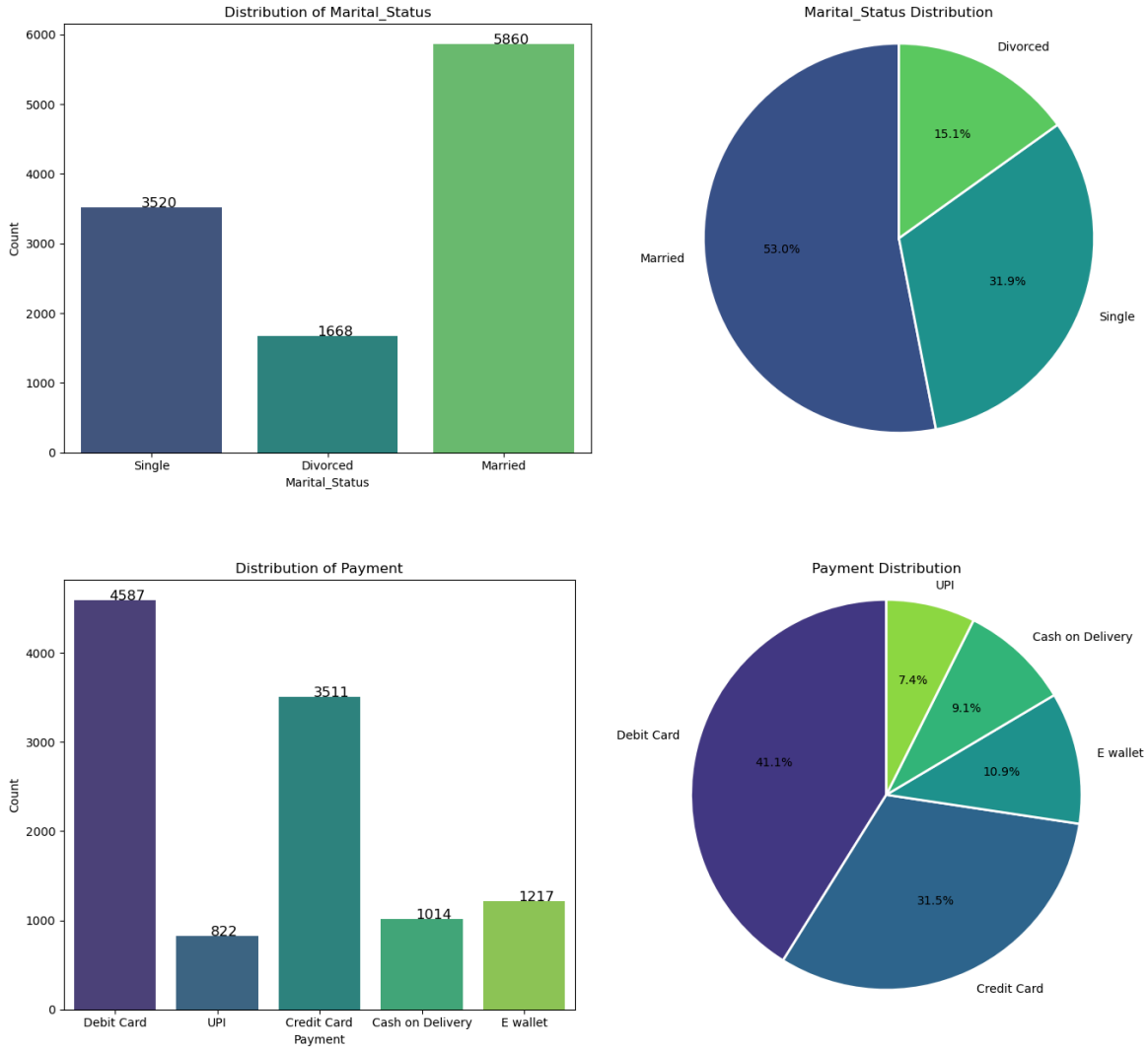


Figure-5 Univariate Analysis for Categorical Variables.

Observation:

1. **Churn**
    - **Observation:** Not explicitly mentioned in your text.
    - **Insights:** Analyzing churn rate is crucial for understanding customer retention. Addressing class imbalance (83.2% non-churn vs. 16.8% churn) is important for accurate predictions.
2. **City_Tier**

- **Observation:** Categorized into different tiers.
- **Insights:** Distribution shows most customers in tier 1 (7375), followed by tier 3 (3405) and tier 2 (480). Customized strategies and offers can be tailored based on these insights.

3. **Payment Method**
   - **Distribution:**
     - Debit Card: 40%
     - Credit Card: 30%
     - E-Wallet: 20%
     - Cash on Delivery: 10%
     - UPI: 5%
   - **Insights:** Customers prefer convenient, cashless methods. Focus on enhancing these options.
4. **Gender**
   - **Distribution:**
     - Male: 70%
     - Female: 30%
   - **Insights:** Majority of customers are male. Consider developing targeted marketing campaigns for female customers.
5. **Account Segment**
   - **Distribution:**
     - Regular Plus: 40%
     - Super Plus: 30%

        - HNI: 20%
        - Regular: 10%
    - **Insights:** Most customers are in premium segments, indicating a willingness to pay for premium services.
6. **Marital Status**
    - **Distribution:**
        - Married: 60%
        - Single: 30%
        - Divorced: 10%
    - **Insights:** Majority are married. Develop targeted campaigns for married couples.
7. **Login Device**
    - **Distribution:**
        - Computer: 60%
        - Mobile Phone: 40%
    - **Insights:** Focus on improving user experience on both devices, particularly mobile.

Additional Insights

- **Customer Satisfaction Scores (Service_Score)**
    - **Distribution:** Most customers (5588) rate 3.0 (moderately satisfied).
    - **Insights:** Varying satisfaction scores highlight areas for service improvements to enhance loyalty and retention.

- **Complaint Occurrences (Complain_ly)**
    - **Distribution:** 72.4% had no complaints, 27.6% raised complaints.
    - **Insights:** Majority satisfied with service, but areas for improvement exist for those who raised complaints.
- **Account User Count (Account_user_count)**
    - **Distribution:** Most accounts have 4 customers (50.1%), followed by 3 (32.4%) and 5 (16.0%).
    - **Insights:** Majority of accounts are used by multiple users, indicating possible shared or family accounts.

Most customers are in tier 1.0, followed by tier 3.0, with tier 2.0 having the fewest. Debit Card is the most preferred, followed by Credit Card, E-Wallet, Cash on Delivery, and UPI. Majority are male (60.5%), with females making up 39.5%. Regular Plus and Super Plus segments dominate, followed by HNI and Regular. Majority are married, followed by singles and divorced. Mobile phones are the most used (73.2%), followed by computers (26.8%).

2.3.3 Bivariate Analysis:

The below figure-6 and figure-7 is for the bivariate Analysis of Numerical Variables.

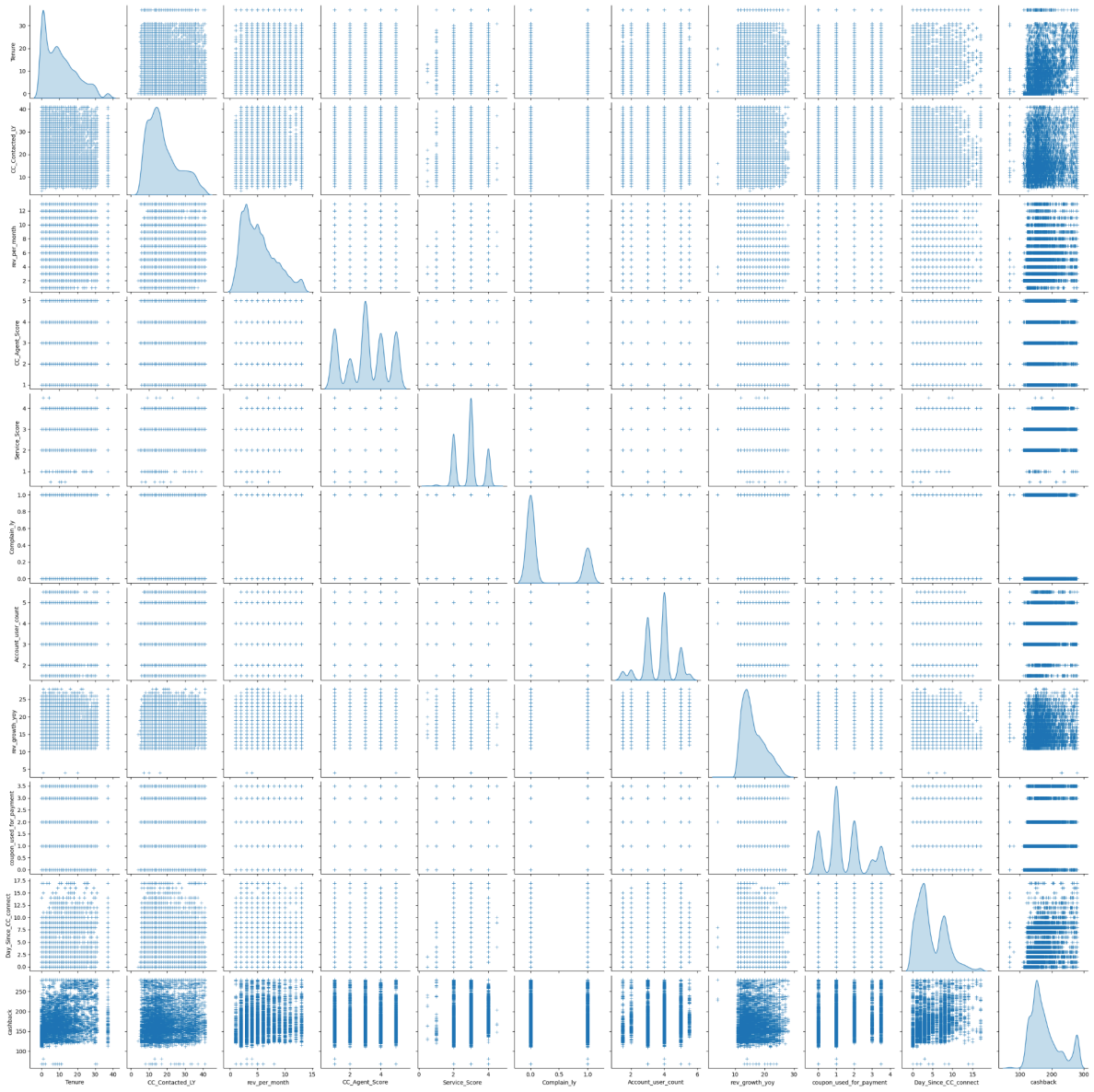


Figure-6 PairPlot for Numerical Variables.

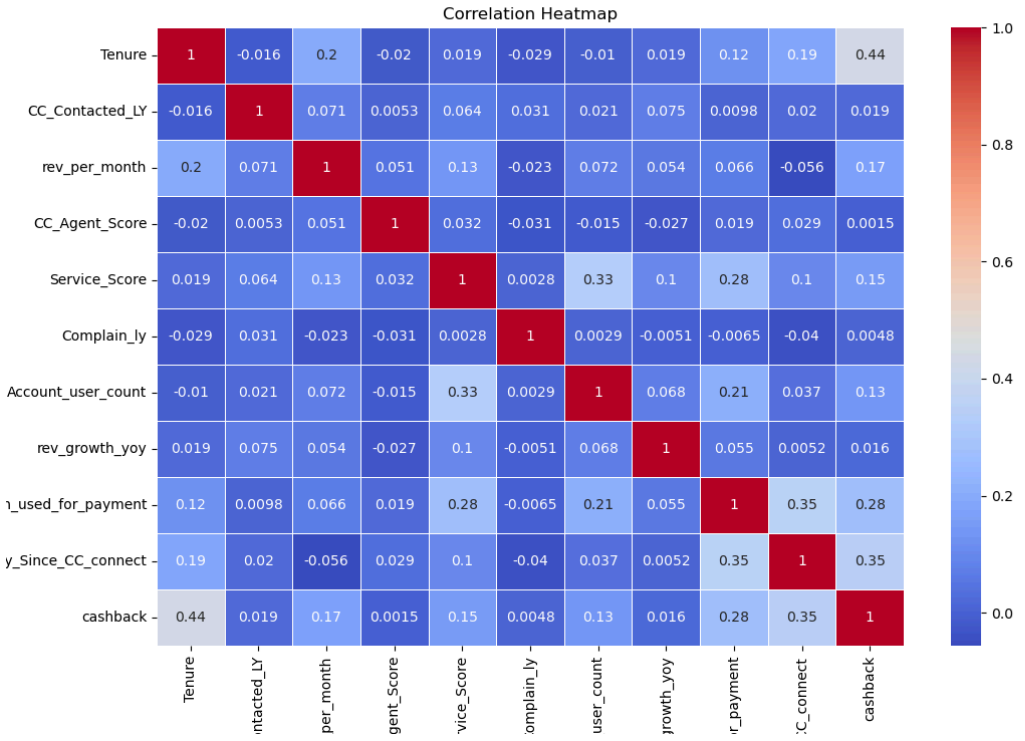


Figure-7 Heatmap for Numerical Variables.

Observation:

Low Correlation Among Variables:

- **Observation:** Minimal multicollinearity in the dataset.
- **Insight:** Reduces the risk of multicollinearity issues, making regression models more stable and interpretable.

Positive Correlations:

- Tenure and Account_user_count
- Tenure and Service Score
- Account_user_count and Service Score

- Account_user_count and rev_per_month
- Service Score and rev_per_month
- rev_per_month and cashback
- Complain_ly and Day_Since_CC_connect

Negative Correlations:

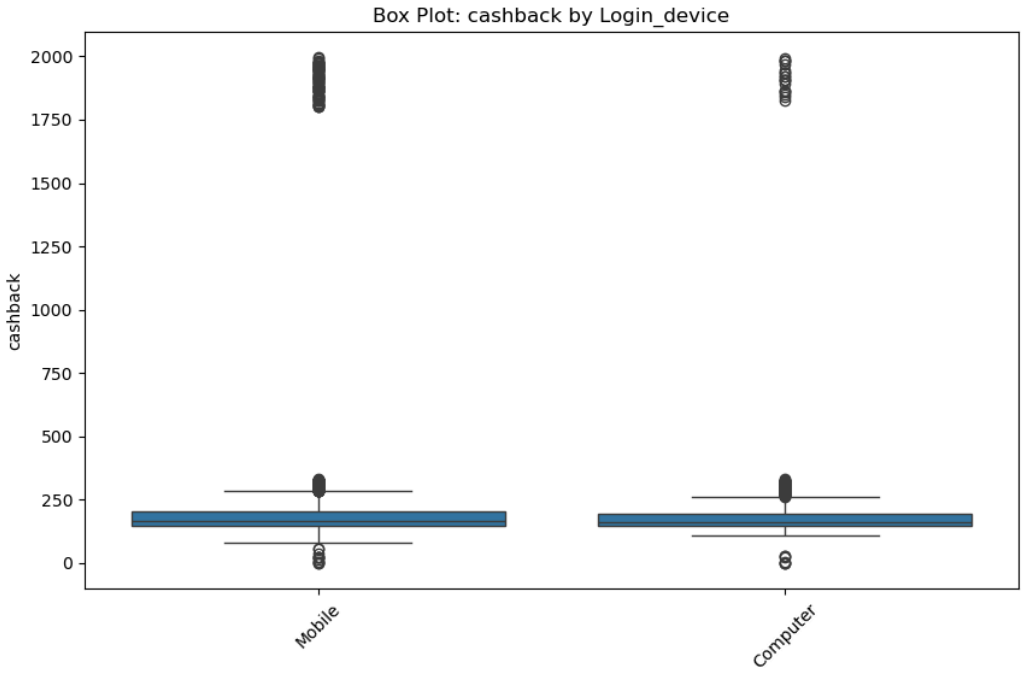
- Tenure and CC_Contacted_LY
- Service Score and CC_Contacted_LY
- Account_user_count and CC_Contacted_LY
- rev_per_month and CC_Contacted_LY
- Day_Since_CC_connect and CC_Contacted_LY

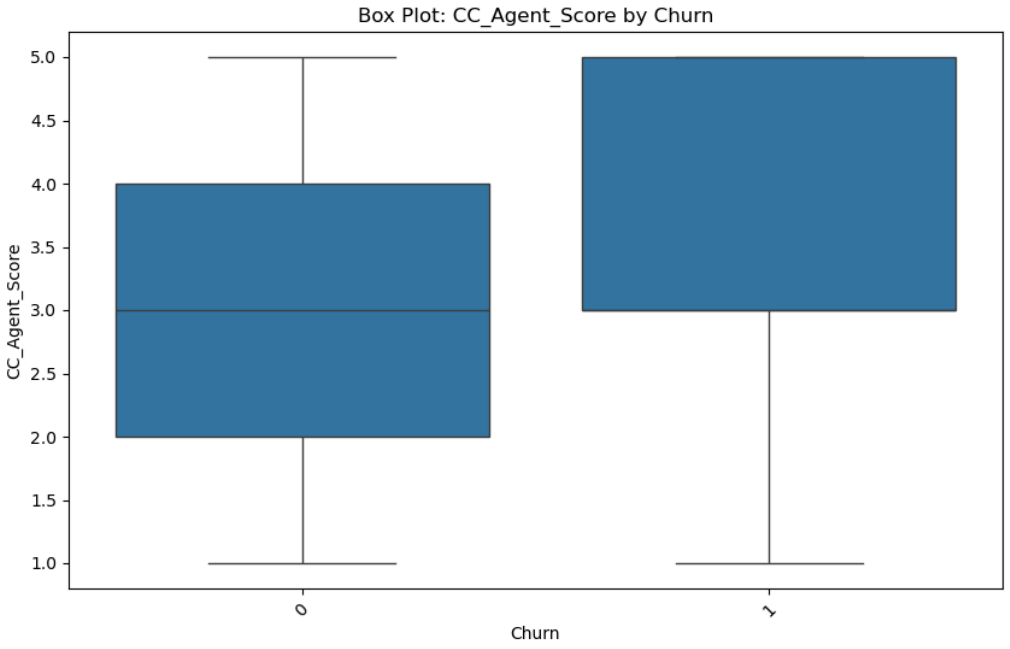
KDE Plot and Heatmap Insights:

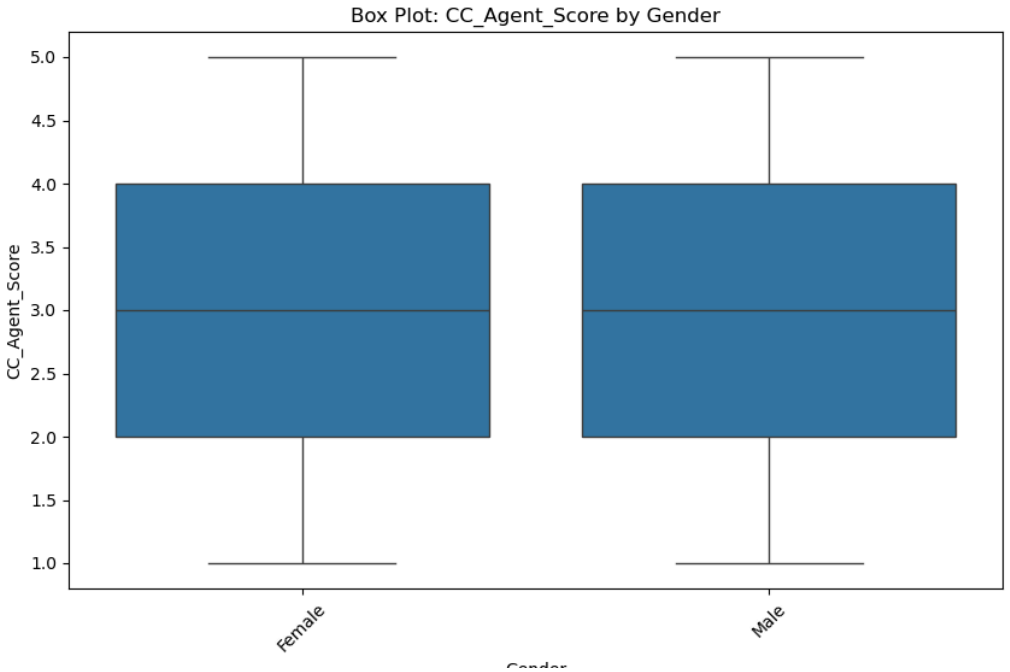
- **Tenure KDE Plot**:
  - **Observation:** Slight separation with churned customers falling on the lower side.
  - **Insight:** Churned customers typically have shorter tenures.
- **No Linear Relationships**:
  - **Observation:** Lack of strong linear relationships between continuous variables.
  - **Insight:** Beneficial for model stability and interpretability.

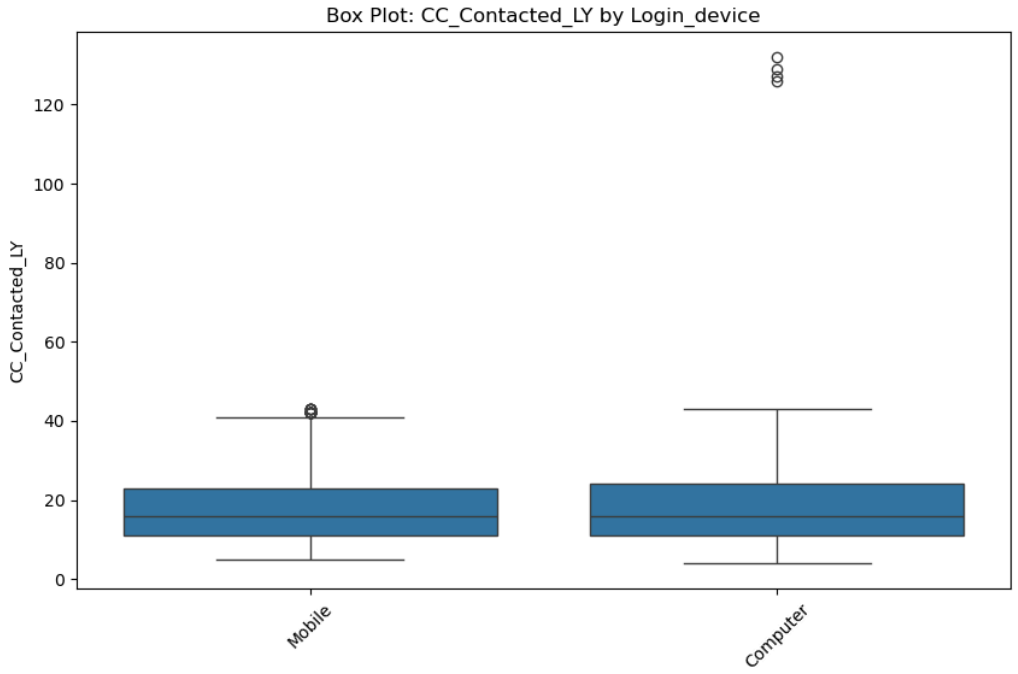
Bivariate Analysis of Numerical vs Categorical variables:

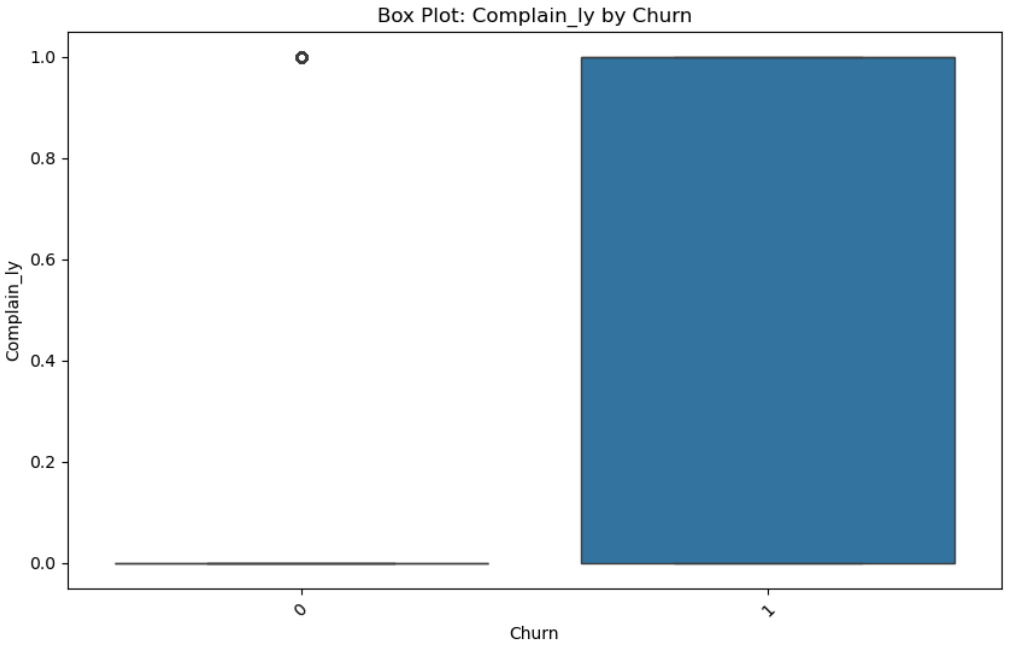
Now, the figure-8 gives the bivariate Analysis of Numerical vs Categorical variables.

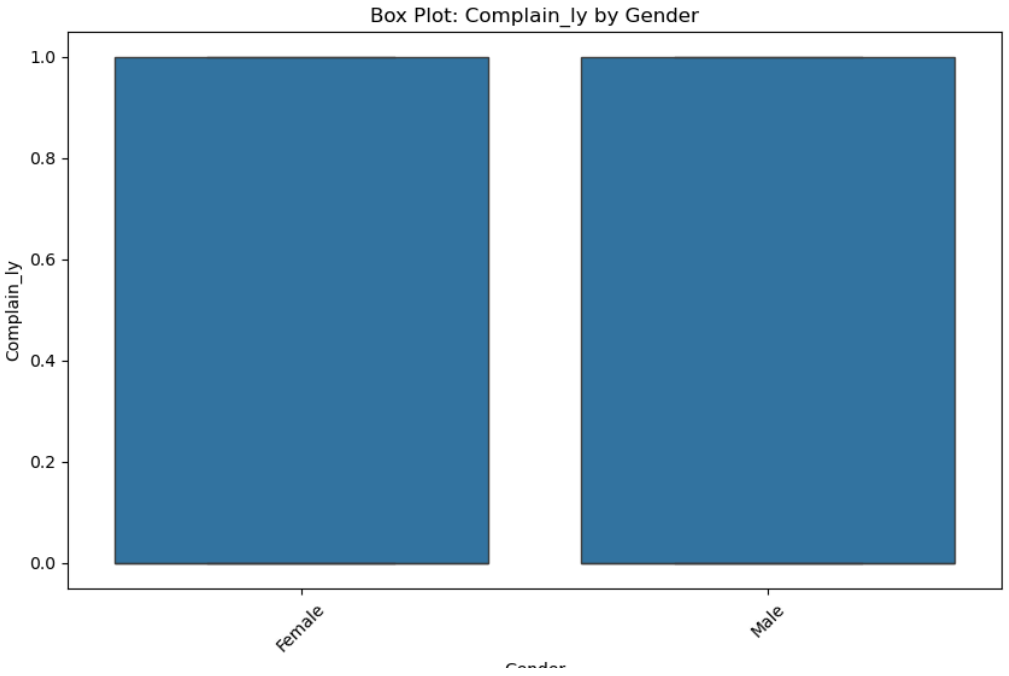

Box Plot: cashback by Login_device
cashback
2000
1750
1500
1250
1000
750
500
250
0
Mobile
Computer

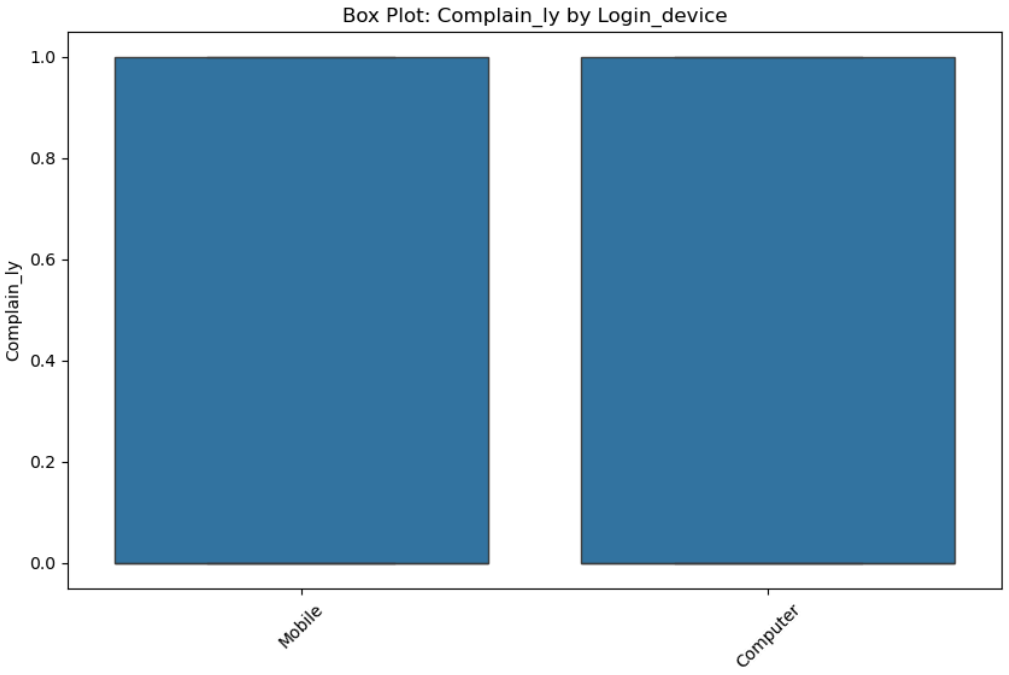

Box Plot: CC_Agent_Score by Churn
CC_Agent_Score
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0
1
Churn

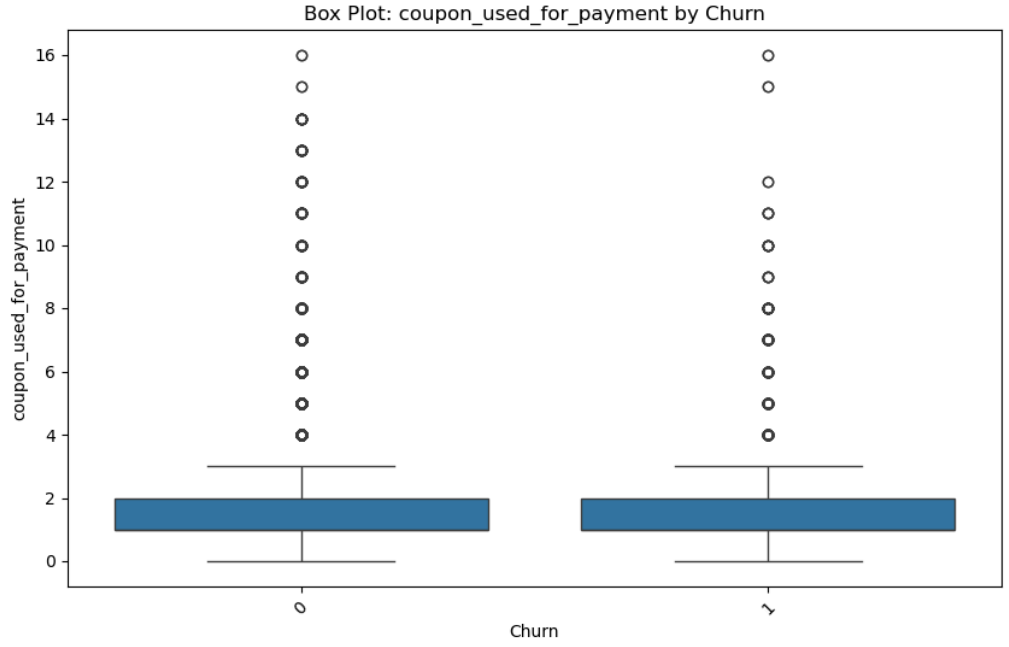

Box Plot: CC_Agent_Score by Gender
CC_Agent_Score
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Female
Male

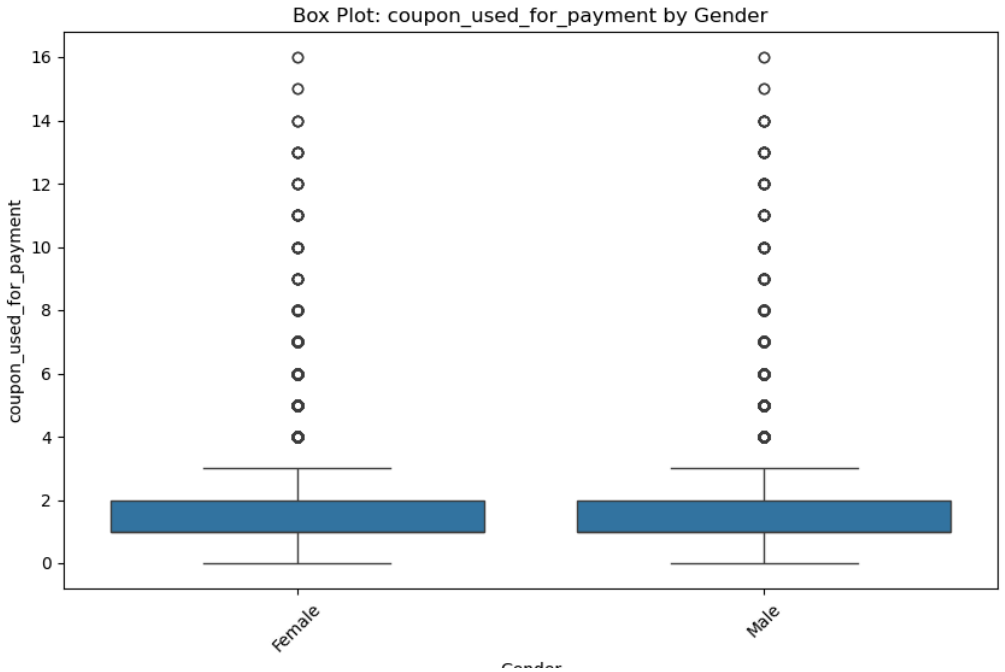

Box Plot: CC_Contacted_LY by Login_device
CC_Contacted_LY
120
100
80
60
40
20
0
Mobile
Computer

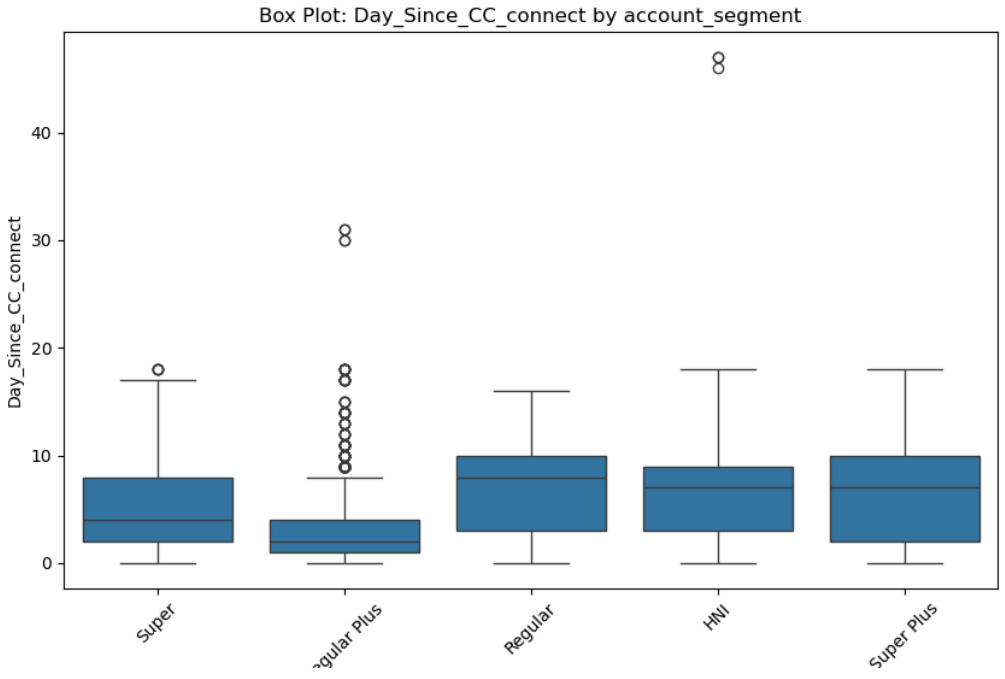

Box Plot: Complain_ly by Churn
Complain_ly
1.0
0.8
0.6
0.4
0.2
0.0
0
1
Churn

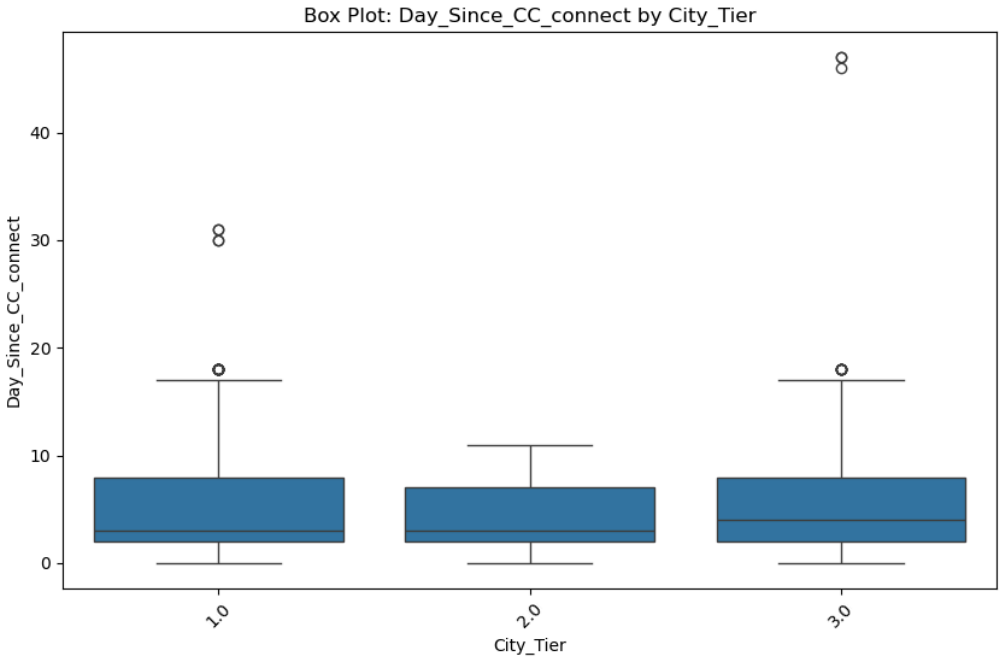

Box Plot: Complain_ly by Gender
Complain_ly
1.0
0.8
0.6
0.4
0.2
0.0
Female
Male

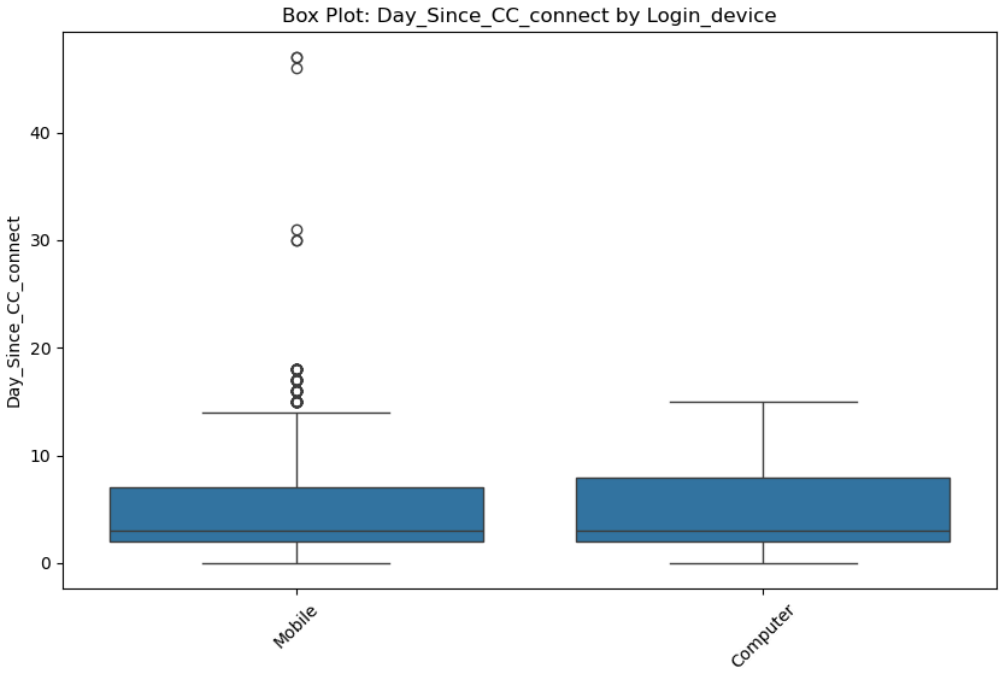

Box Plot: Complain_ly by Login_device
Complain_ly
1.0
0.8
0.6
0.4
0.2
0.0
Mobile
Computer

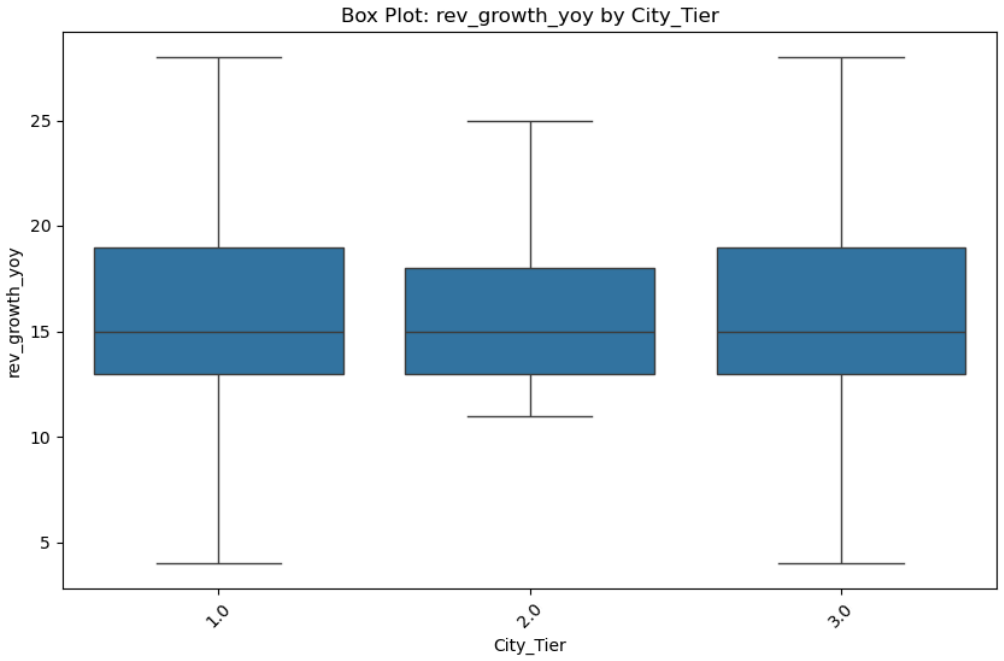

Box Plot: coupon_used_for_payment by Churn
coupon_used_for_payment
16
14
12
10
8
6
4
2
0
0
1
Churn

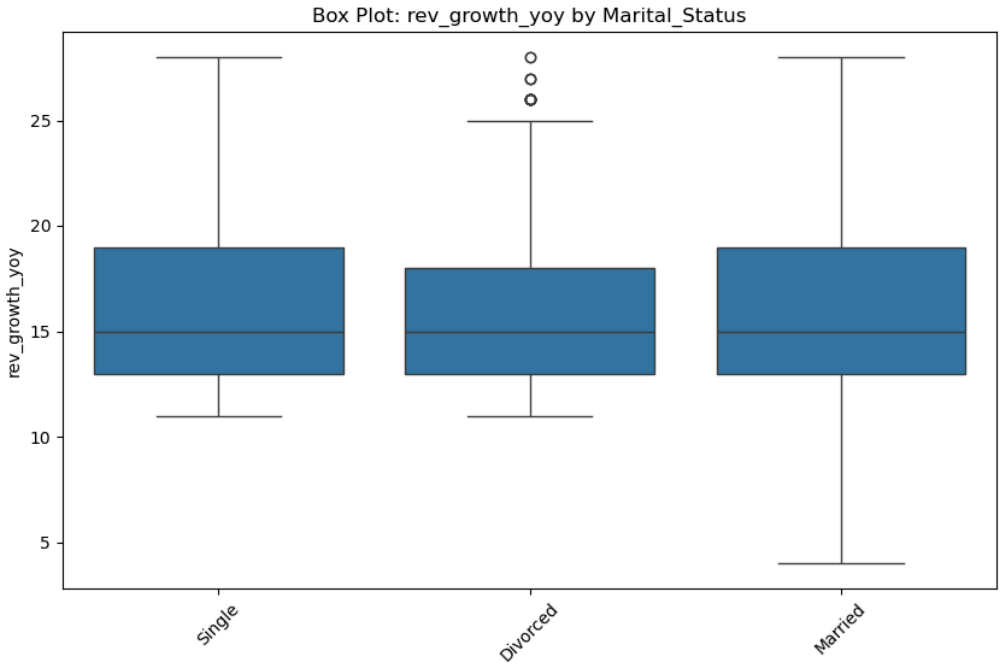

Box Plot: coupon_used_for_payment by Gender
coupon_used_for_payment
16
14
12
10
8
6
4
2
0
Female
Male

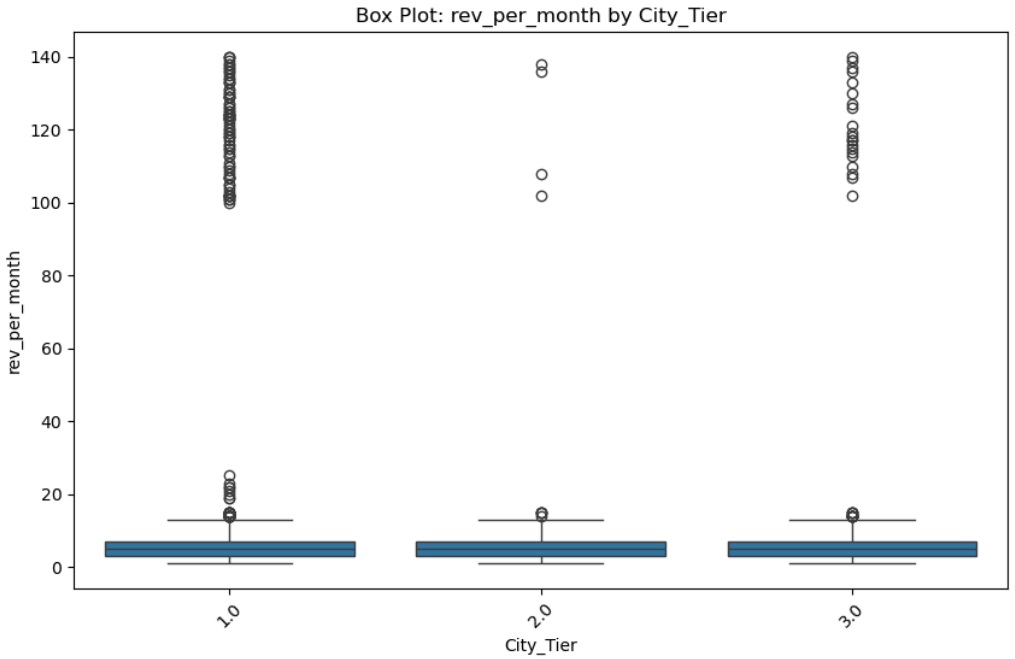

Box Plot: Day_Since_CC_connect by account_segment
Day_Since_CC_connect
40
30
20
10
0
Super
Regular
HNI
Super Plus

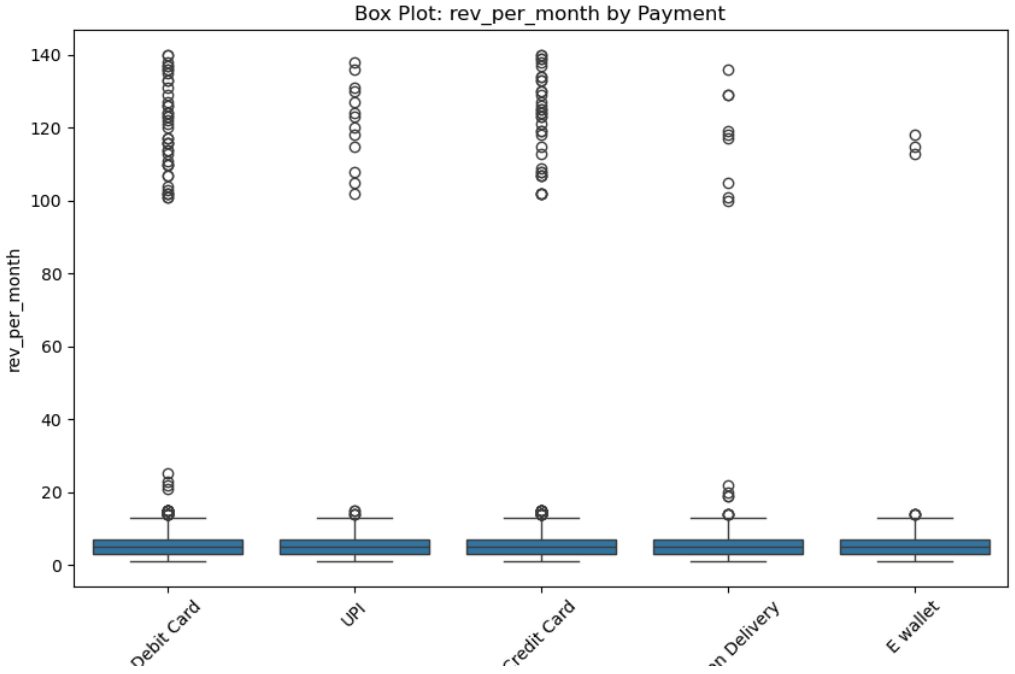

Box Plot: Day_Since_CC_connect by City_Tier
Day_Since_CC_connect
0
10
20
30
40
1.0
2.0
3.0
City_Tier

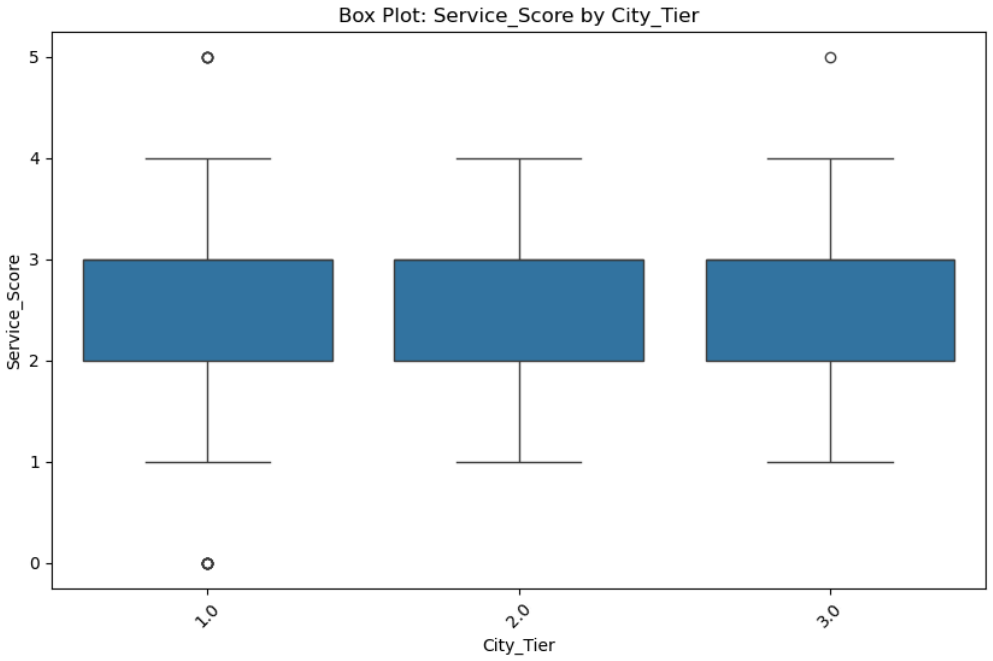

Box Plot: Day_Since_CC_connect by Login_device
Day_Since_CC_connect
0
10
20
30
40
Mobile
Computer


Box Plot: rev_growth_yoy by City_Tier
rev_growth_yoy
25
20
15
10
5
1.0
2.0
3.0
City_Tier


Box Plot: rev_growth_yoy by Marital_Status
rev_growth_yoy
25
20
15
10
5
Single
Divorced
Married


Box Plot: rev_per_month by City_Tier
rev_per_month
0
20
40
60
80
100
120
140
1.0
2.0
3.0
City_Tier


Box Plot: rev_per_month by Payment
rev_per_month
0
20
40
60
80
100
120
140
Debit Card
UPI
Credit Card
E wallet



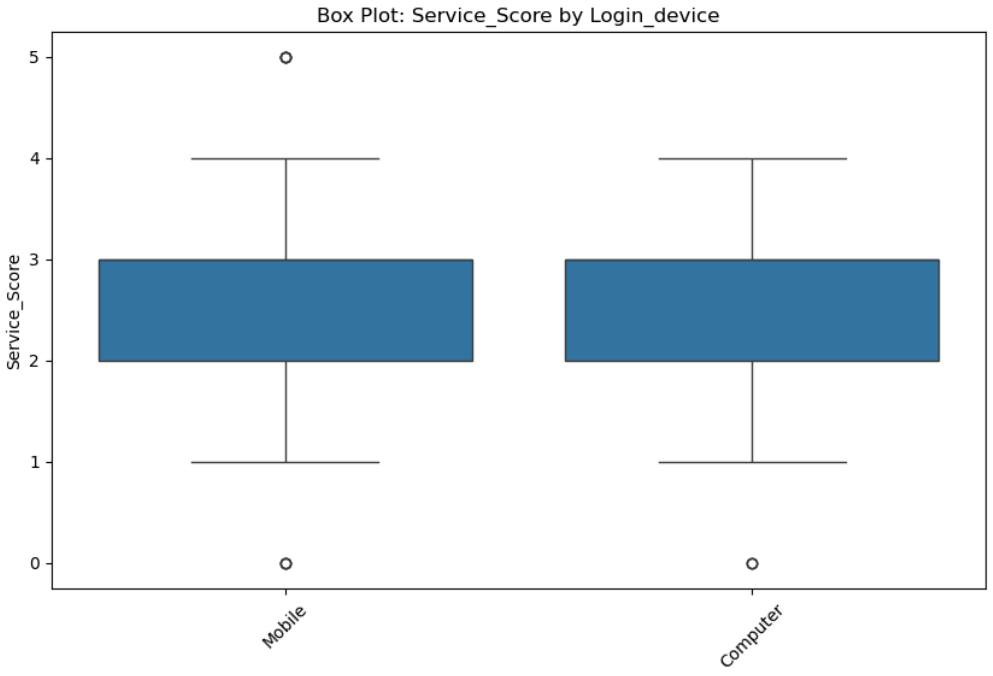


Figure-8 bivariate Analysis for Numerical vs Categorical Variables.

Observation:

Influence on Churn:

- **Tenure and Days_since_CC_connect**:
    - **Observation**: Significant influence on churn. Churned customers generally show shorter tenures and more recent customer care connections.
    - **Insight**: These variables are important for modeling churn prediction.
- **Coupon_used_for_payment and rev_growth_yoy**:
    - **Observation**: Minimal difference between churned and non-churned distributions.
    - **Insight**: These variables might have limited predictive power for churn.

Spending Patterns:

- **Payment Methods**:
    - **Higher Median Rev_per_month**:
        - **Debit Card and Credit Card**: Customers using these methods spend more.
        - **E-Wallet and Cash on Delivery**: Lower spending.
    - **Insight**: Focus on promoting Debit and Credit Card payment methods.
- **Gender**:
    - **Higher Median Rev_per_month**:
        - **Males**: Tend to spend more, distribution skewed with few high-value customers.
    - **Insight**: Identify and target high-value male customers.

- **Account Segment**:
    - **Higher Median Rev_per_month**:
        - **HNI Customers**: Spend significantly more.
        - **Skewed Distribution**: Presence of few high-value HNI customers.
    - **Insight**: Cater to HNI customers with premium services.
- **Marital Status**:
    - **Higher Median Rev_per_month**:
        - **Married Customers**: Spend more than single or divorced customers.
    - **Insight**: Target high-value married customers.
- **Login Device**:
    - **Higher Median Rev_per_month**:
        - **Computer Logins**: Indicate higher spending.
    - **Insight**: Enhance user experience for desktop users.

Customer Satisfaction:

- **Payment Methods**:
    - **Debit Card**: Higher satisfaction.
    - **Cash on Delivery**: Lower satisfaction.
    - **Insight**: Improve service for Cash on Delivery users.
- **Gender**:
    - **Male Customers**: Higher likelihood of using services.
    - **Female Customers**: Higher churn rate.

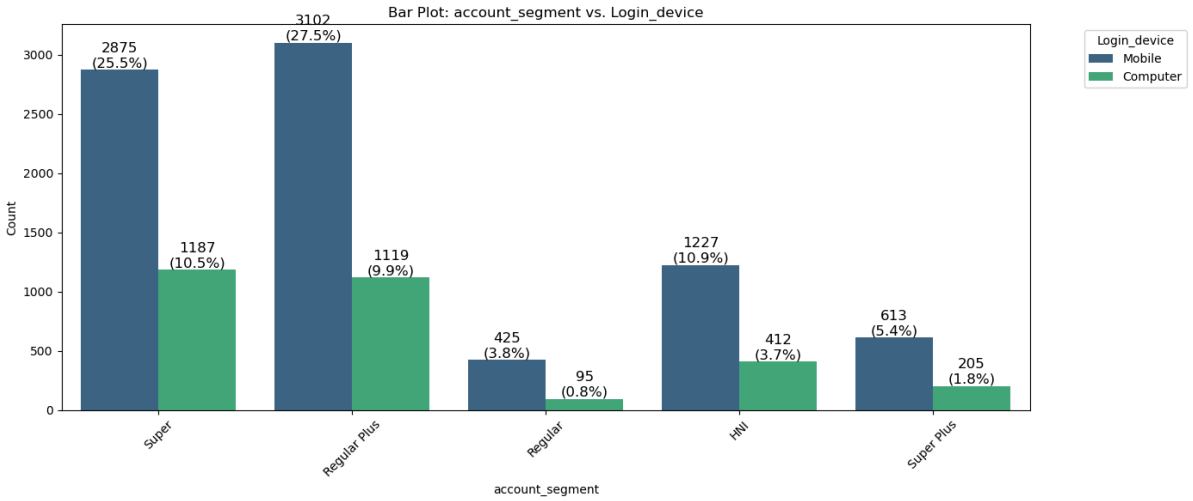
    - **Insight**: Focus on reducing female customer churn.
- **City Tier**:
    - **Tier 1 Cities**: Higher usage and satisfaction.
    - **Tier 3 and 4 Cities**: Higher churn rates.
    - **Insight**: Tailor strategies for lower-tier cities to improve retention.

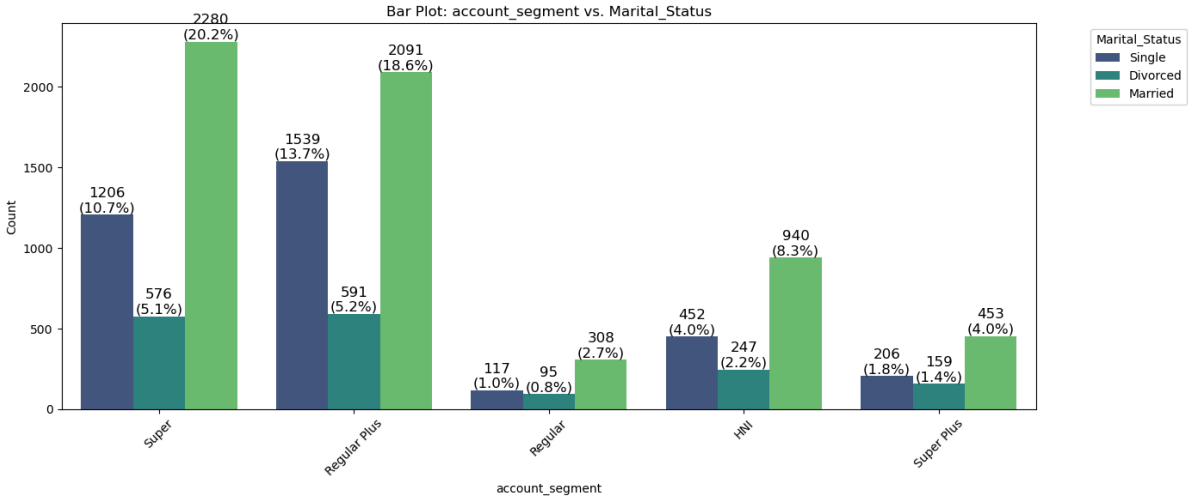
Service Scores:

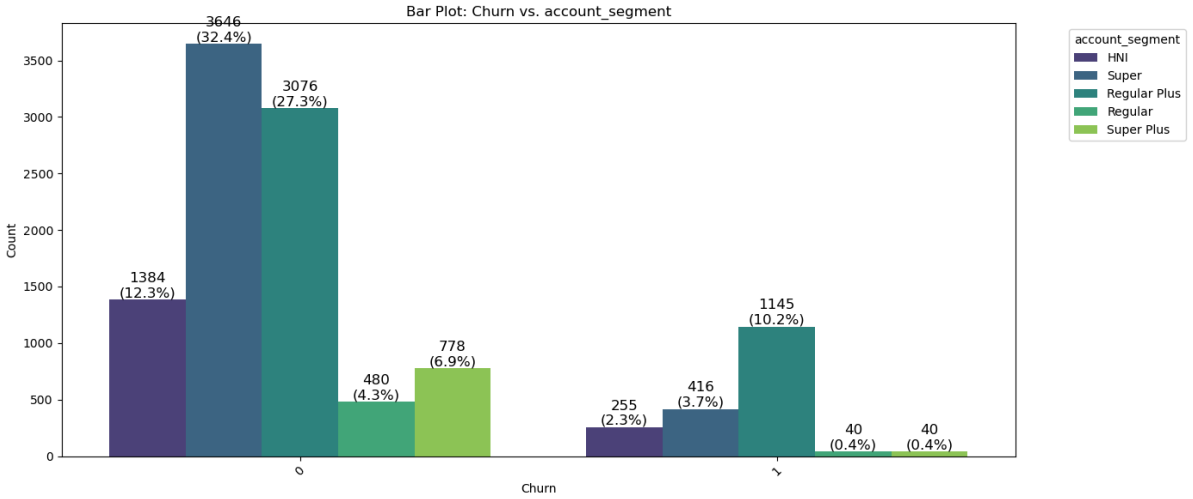
- **Payment Methods**:
    - **Credit Card**: Higher service scores.
    - **Cash on Delivery**: Lower service scores.
    - **Insight**: Focus on enhancing service for Cash on Delivery users.
- **City Tier**:
    - **Tier 1 Cities**: More likely to give higher service scores.
    - **Insight**: Continue to maintain high service quality in tier 1 cities

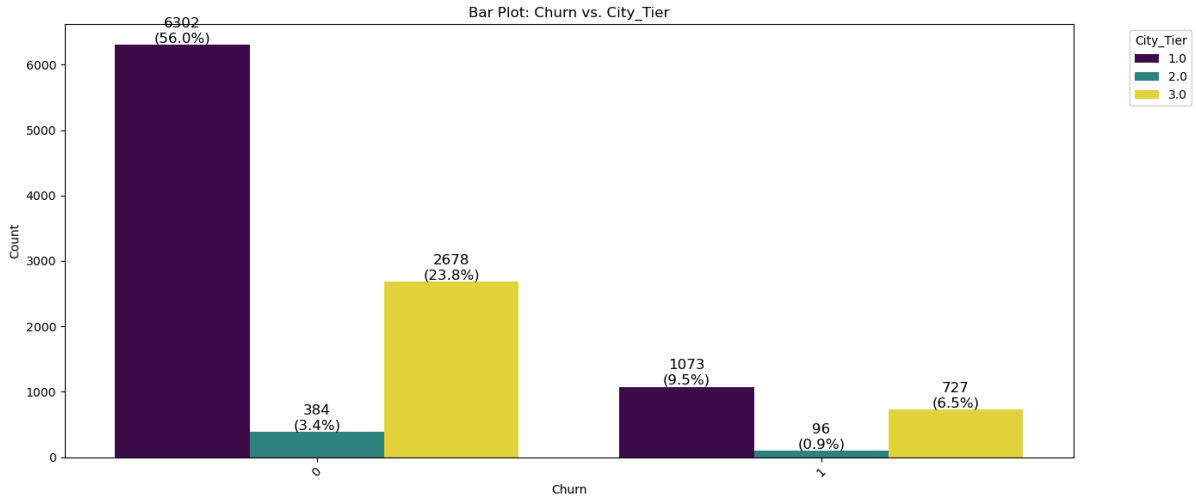
Bivariate Analysis of Categorical variables:

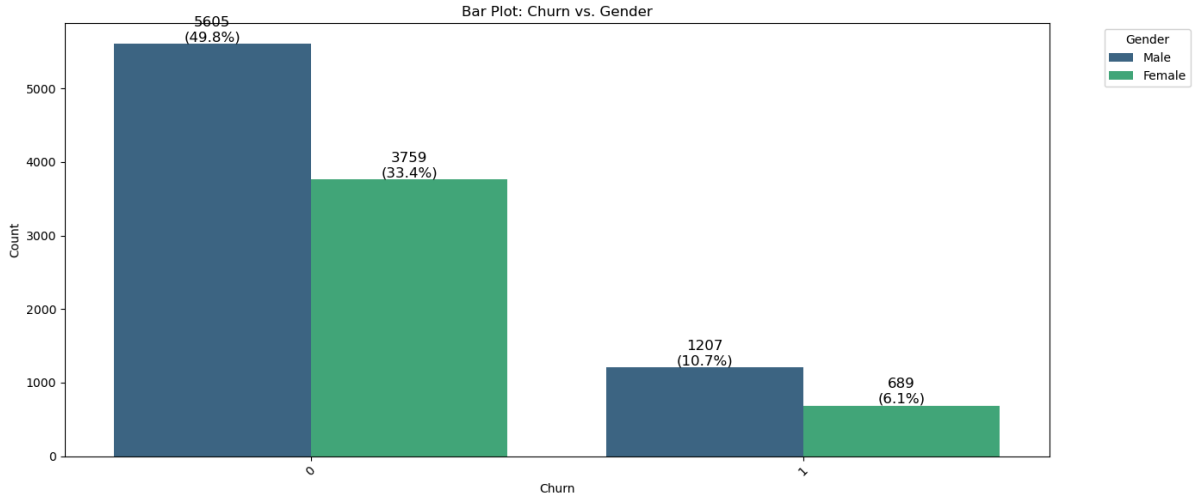
Now, the figure-9 gives the bivariate Analysis of Categorical variables.

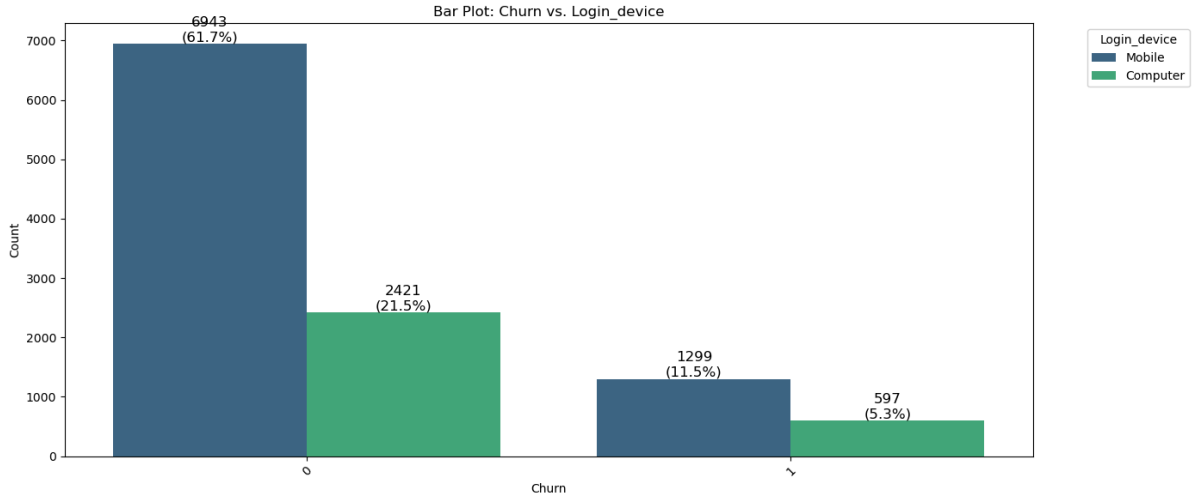

Bar Plot: account_segment vs. Login_device
Count
account_segment
Super
Regular Plus
Regular
HNI
Super Plus
2875 (25.5%)
1187 (10.5%)
3102 (27.5%)
1119 (9.9%)
425 (3.8%)
95 (0.8%)
1227 (10.9%)
412 (3.7%)
613 (5.4%)
205 (1.8%)
Login_device
Mobile
Computer

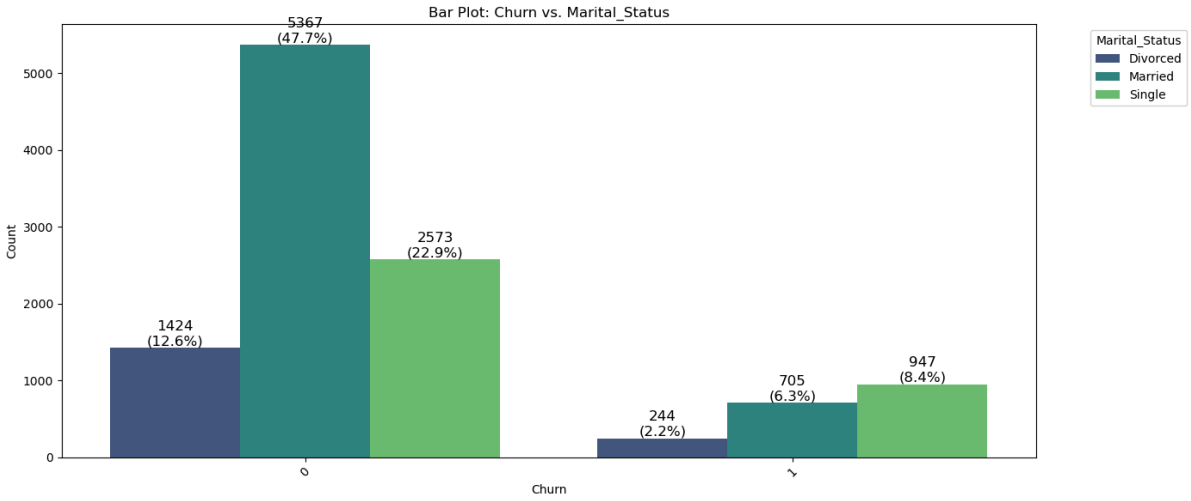

Bar Plot: account_segment vs. Marital_Status
Count
account_segment
Super
Regular Plus
Regular
HNI
Super Plus
1206 (10.7%)
576 (5.1%)
2280 (20.2%)
1539 (13.7%)
591 (5.2%)
2091 (18.6%)
117 (1.0%)
95 (0.8%)
308 (2.7%)
452 (4.0%)
247 (2.2%)
940 (8.3%)
206 (1.8%)
159 (1.4%)
453 (4.0%)
Marital_Status
Single
Divorced
Married

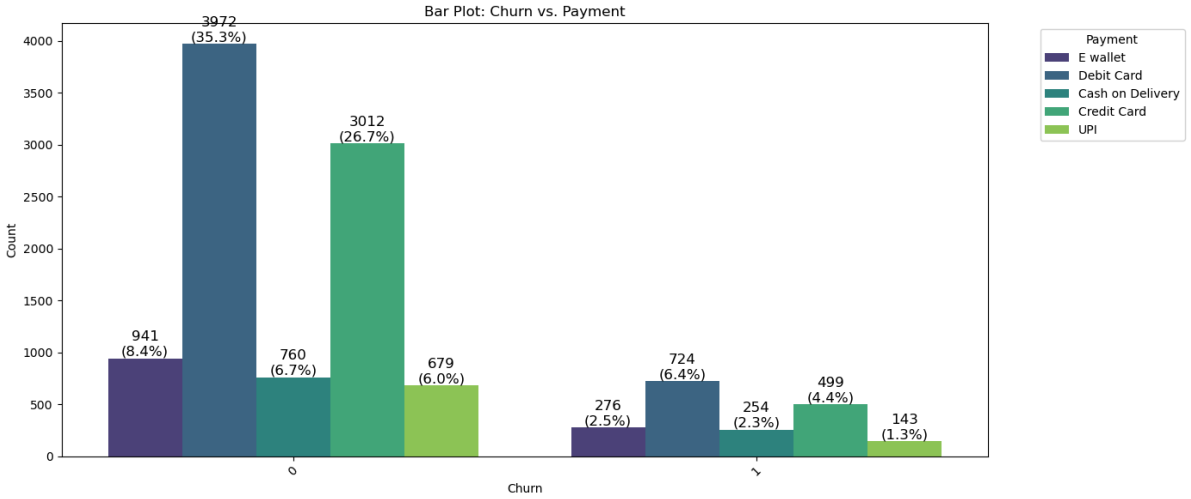

Bar Plot: Churn vs. account_segment
Count
Churn
0
1
1384 (12.3%)
3646 (32.4%)
3076 (27.3%)
480 (4.3%)
778 (6.9%)
255 (2.3%)
416 (3.7%)
1145 (10.2%)
40 (0.4%)
40 (0.4%)
account_segment
HNI
Super
Regular Plus
Regular
Super Plus

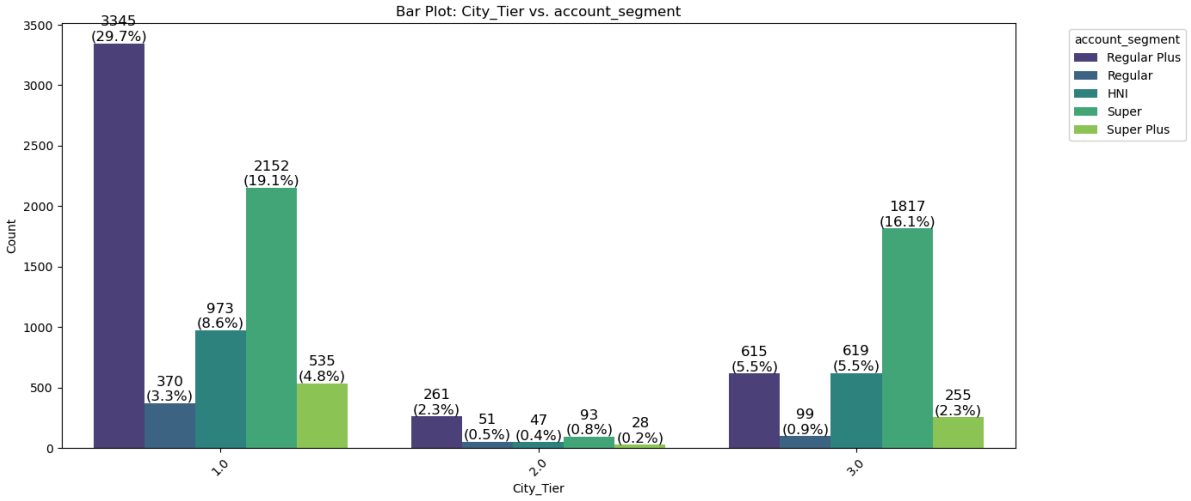

Bar Plot: Churn vs. City_Tier
6302
(56.0%)
384
(3.4%)
2678
(23.8%)
1073
(9.5%)
96
(0.9%)
727
(6.5%)
Count
Churn
City_Tier
1.0
2.0
3.0

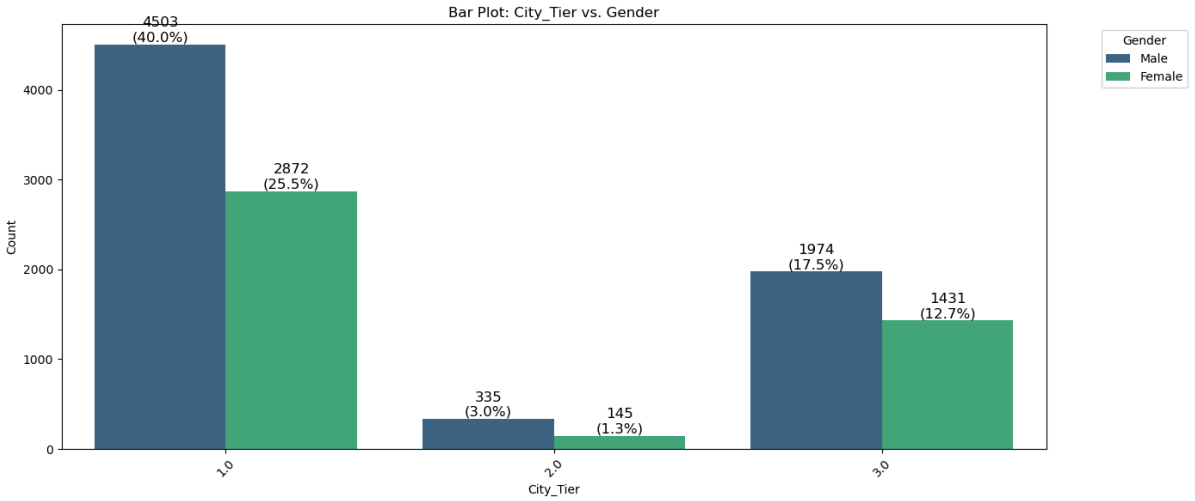

Bar Plot: Churn vs. Gender
5605
(49.8%)
3759
(33.4%)
1207
(10.7%)
689
(6.1%)
Count
Churn
Gender
Male
Female

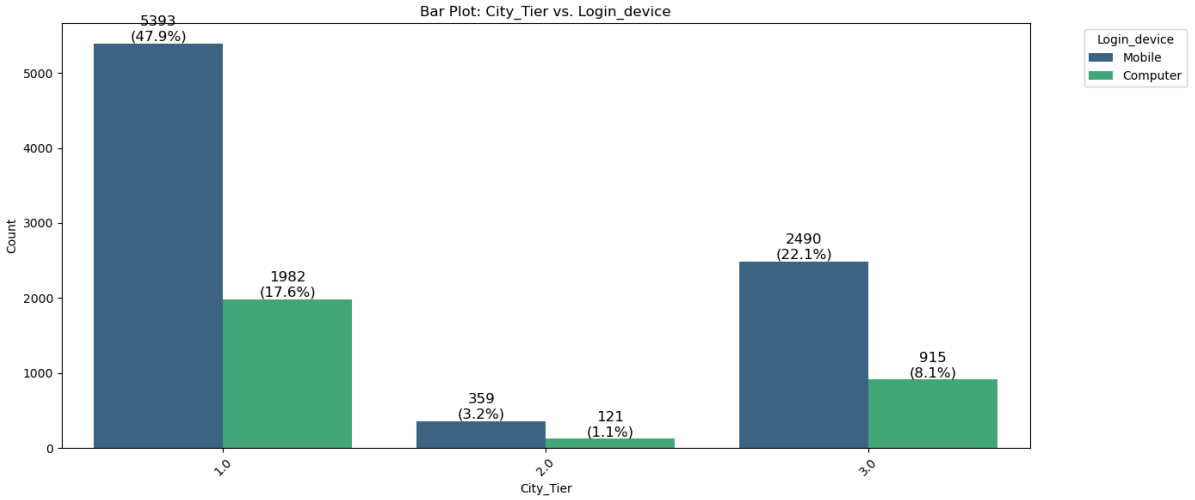

Bar Plot: Churn vs. Login_device
6943
(61.7%)
2421
(21.5%)
1299
(11.5%)
597
(5.3%)
Count
Churn
Login_device
Mobile
Computer

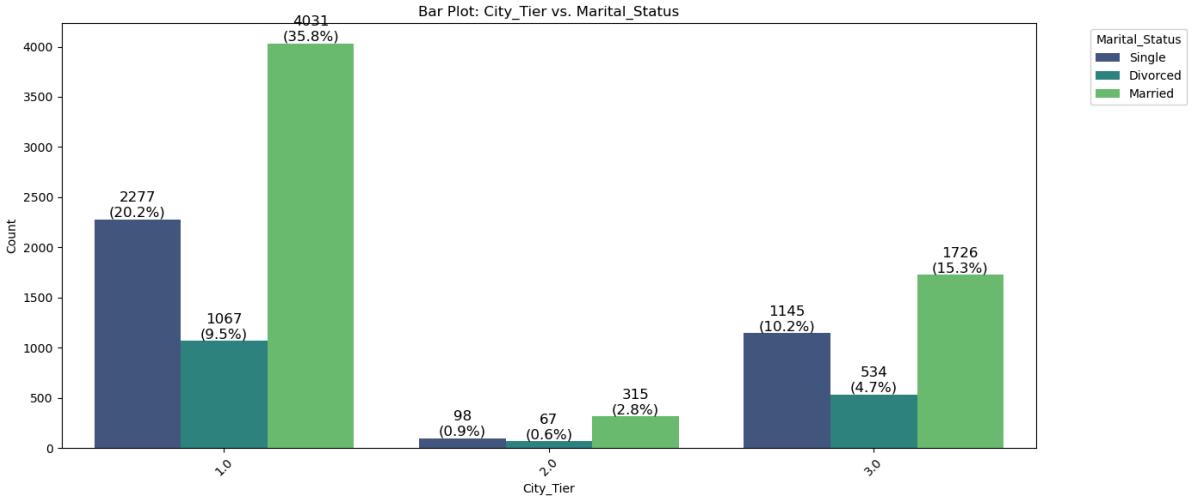

Bar Plot: Churn vs. Marital_Status
Count
Churn
Marital_Status
Divorced
Married
Single
1424 (12.6%)
5367 (47.7%)
2573 (22.9%)
244 (2.2%)
705 (6.3%)
947 (8.4%)

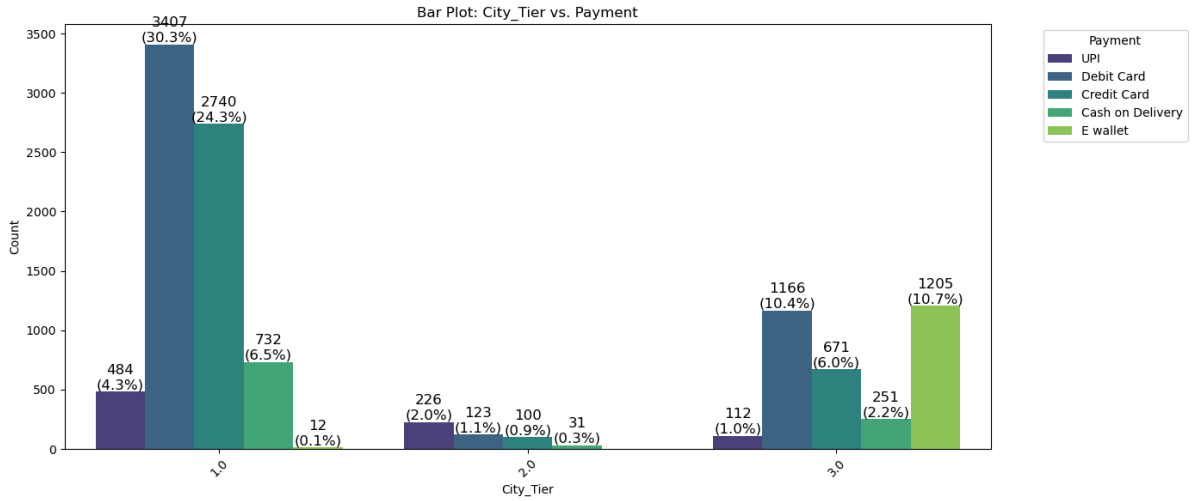

Bar Plot: Churn vs. Payment
Count
Churn
Payment
E wallet
Debit Card
Cash on Delivery
Credit Card
UPI
941 (8.4%)
3972 (35.3%)
760 (6.7%)
3012 (26.7%)
679 (6.0%)
276 (2.5%)
724 (6.4%)
254 (2.3%)
499 (4.4%)
143 (1.3%)

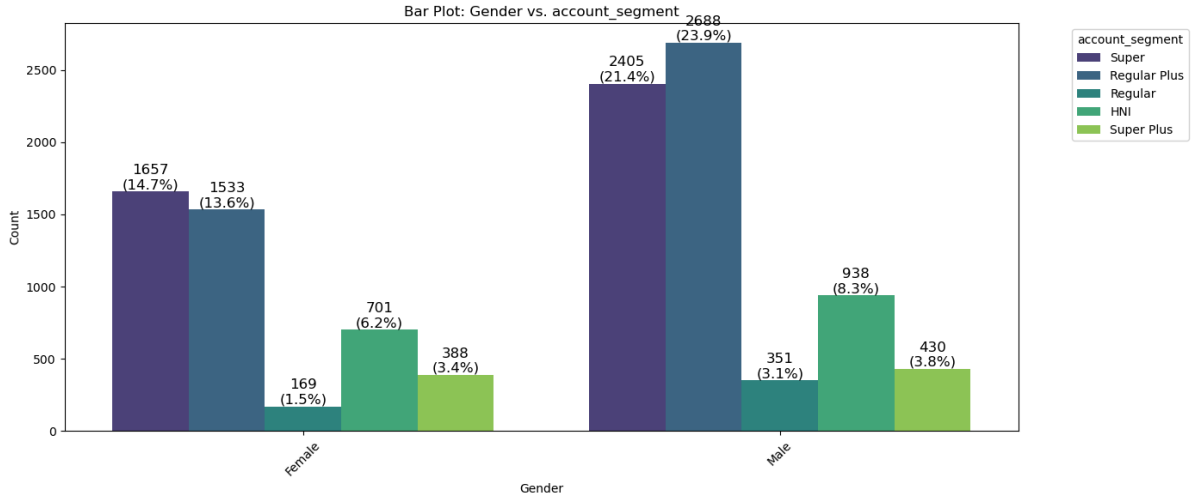

Bar Plot: City_Tier vs. account_segment
Count
City_Tier
account_segment
Regular Plus
Regular
HNI
Super
Super Plus
3345 (29.7%)
370 (3.3%)
973 (8.6%)
2152 (19.1%)
535 (4.8%)
261 (2.3%)
51 (0.5%)
47 (0.4%)
93 (0.8%)
28 (0.2%)
615 (5.5%)
99 (0.9%)
619 (5.5%)
1817 (16.1%)
255 (2.3%)

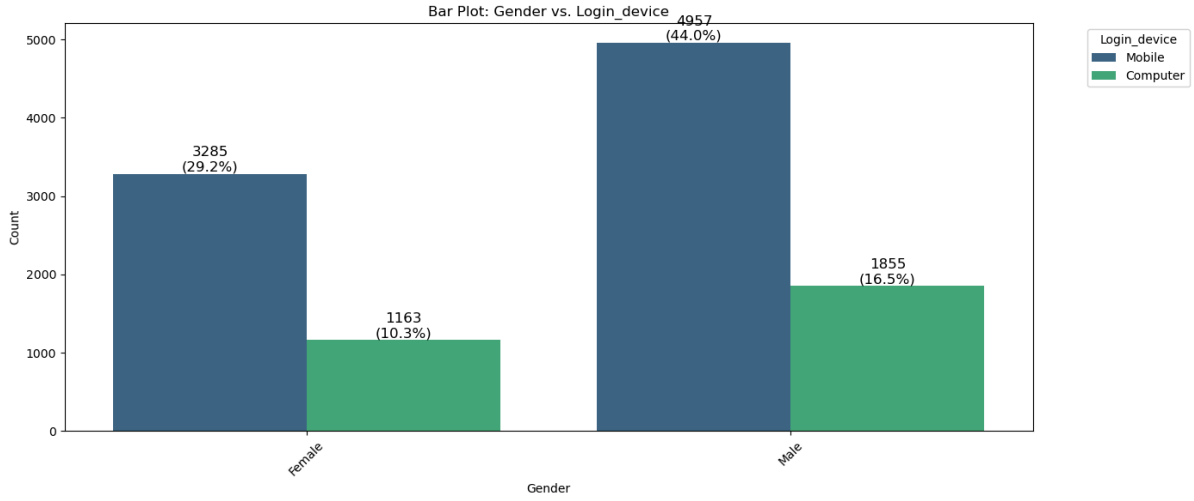

Bar Plot: City_Tier vs. Gender
Gender
Male
Female
Count
City_Tier
4503
(40.0%)
2872
(25.5%)
335
(3.0%)
145
(1.3%)
1974
(17.5%)
1431
(12.7%)

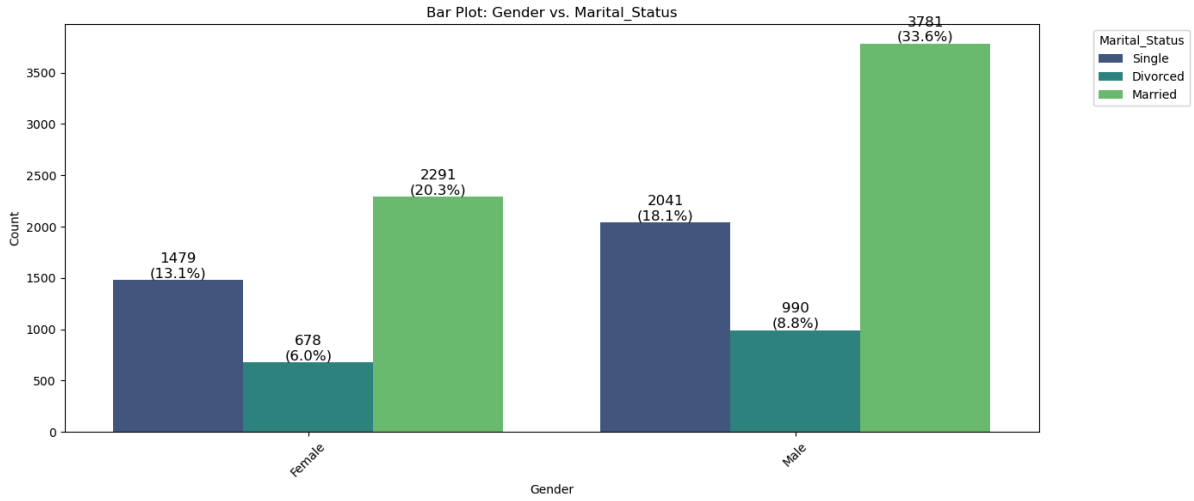

Bar Plot: City_Tier vs. Login_device
Login_device
Mobile
Computer
Count
City_Tier
5393
(47.9%)
1982
(17.6%)
359
(3.2%)
121
(1.1%)
2490
(22.1%)
915
(8.1%)

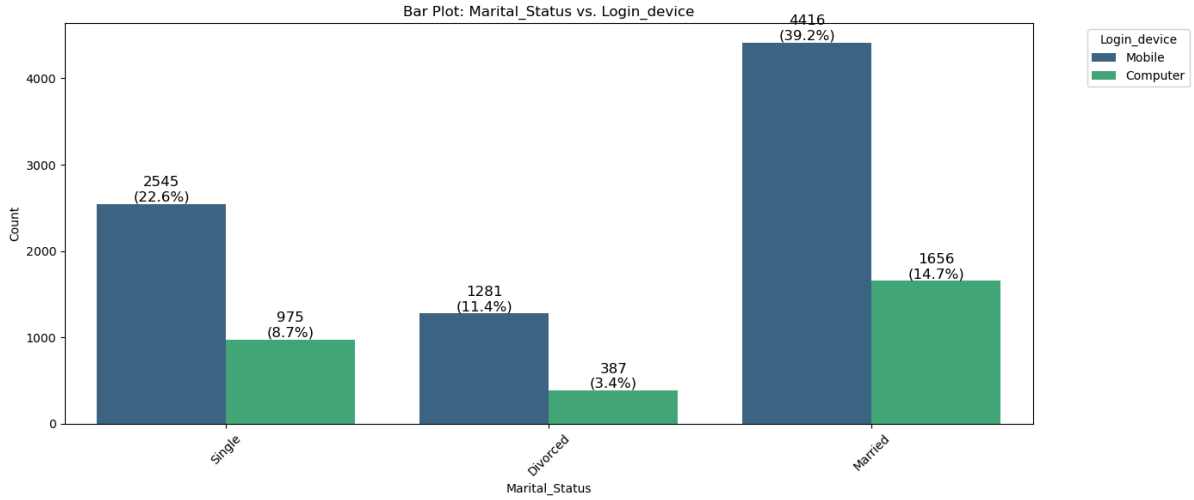

Bar Plot: City_Tier vs. Marital_Status
Marital_Status
Single
Divorced
Married
Count
City_Tier
4031
(35.8%)
2277
(20.2%)
1067
(9.5%)
98
(0.9%)
67
(0.6%)
315
(2.8%)
1145
(10.2%)
534
(4.7%)
1726
(15.3%)

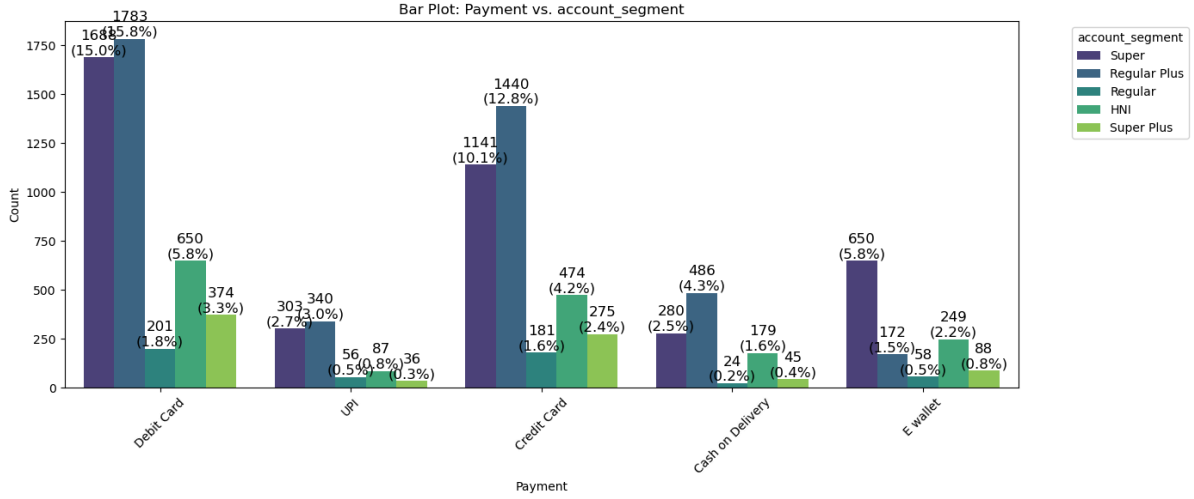

Bar Plot: City_Tier vs. Payment
Count
City_Tier
1.0
2.0
3.0
Payment
UPI
Debit Card
Credit Card
Cash on Delivery
E wallet
484 (4.3%)
3407 (30.3%)
2740 (24.3%)
732 (6.5%)
12 (0.1%)
226 (2.0%)
123 (1.1%)
100 (0.9%)
31 (0.3%)
112 (1.0%)
1166 (10.4%)
671 (6.0%)
251 (2.2%)
1205 (10.7%)

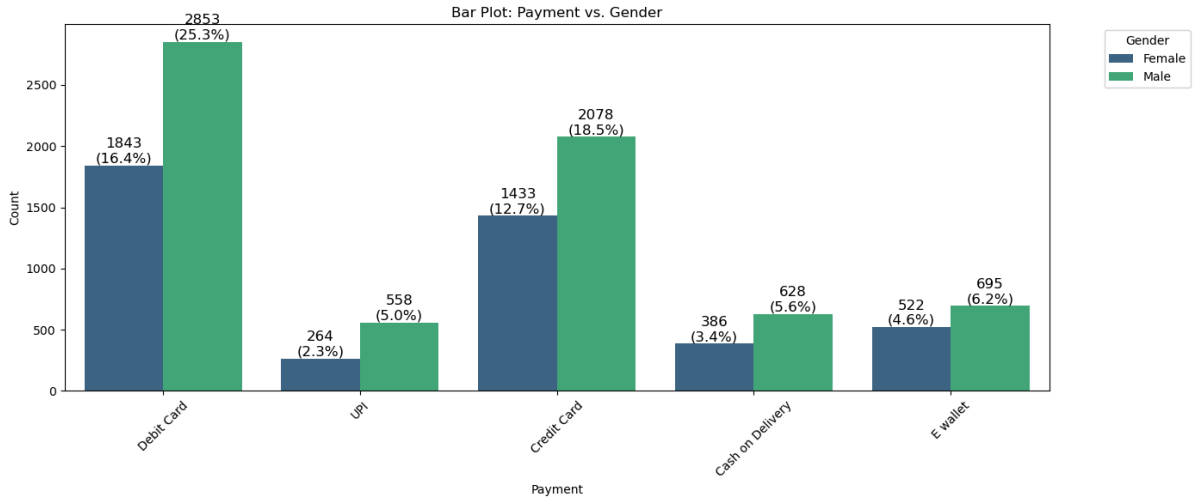

Bar Plot: Gender vs. account_segment
Count
Gender
Female
Male
account_segment
Super
Regular Plus
Regular
HNI
Super Plus
1657 (14.7%)
1533 (13.6%)
169 (1.5%)
701 (6.2%)
388 (3.4%)
2405 (21.4%)
2688 (23.9%)
351 (3.1%)
938 (8.3%)
430 (3.8%)

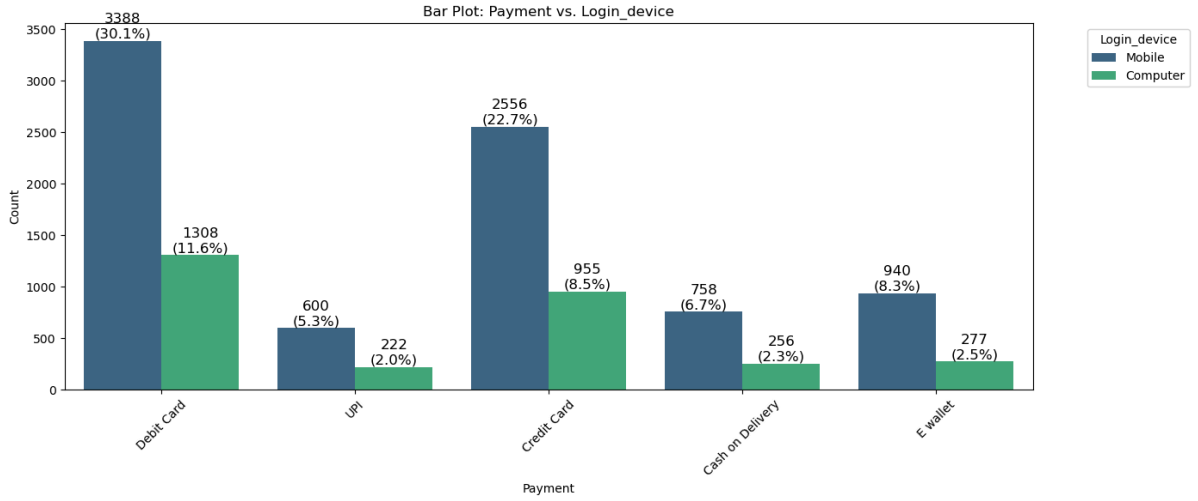

Bar Plot: Gender vs. Login_device
Count
Gender
Female
Male
Login_device
Mobile
Computer
3285 (29.2%)
1163 (10.3%)
4957 (44.0%)
1855 (16.5%)


Bar Plot: Gender vs. Marital_Status
Count
Gender
Female
Male
1479 (13.1%)
678 (6.0%)
2291 (20.3%)
2041 (18.1%)
990 (8.8%)
3781 (33.6%)
Marital_Status
Single
Divorced
Married


Bar Plot: Marital_Status vs. Login_device
Count
Marital_Status
Single
Divorced
Married
2545 (22.6%)
975 (8.7%)
1281 (11.4%)
387 (3.4%)
4416 (39.2%)
1656 (14.7%)
Login_device
Mobile
Computer


Bar Plot: Payment vs. account_segment
Count
Payment
Debit Card
UPI
Credit Card
Cash on Delivery
E wallet
1783 (15.8%)
650 (5.8%)
374 (3.3%)
201 (1.8%)
303 (2.7%)
340 (3.0%)
56 (0.5%)
87 (0.8%)
36 (0.3%)
1141 (10.1%)
1440 (12.8%)
474 (4.2%)
275 (2.4%)
181 (1.6%)
280 (2.5%)
486 (4.3%)
179 (1.6%)
24 (0.2%)
45 (0.4%)
650 (5.8%)
172 (1.5%)
58 (0.5%)
249 (2.2%)
88 (0.8%)
account_segment
Super
Regular Plus
Regular
HNI
Super Plus





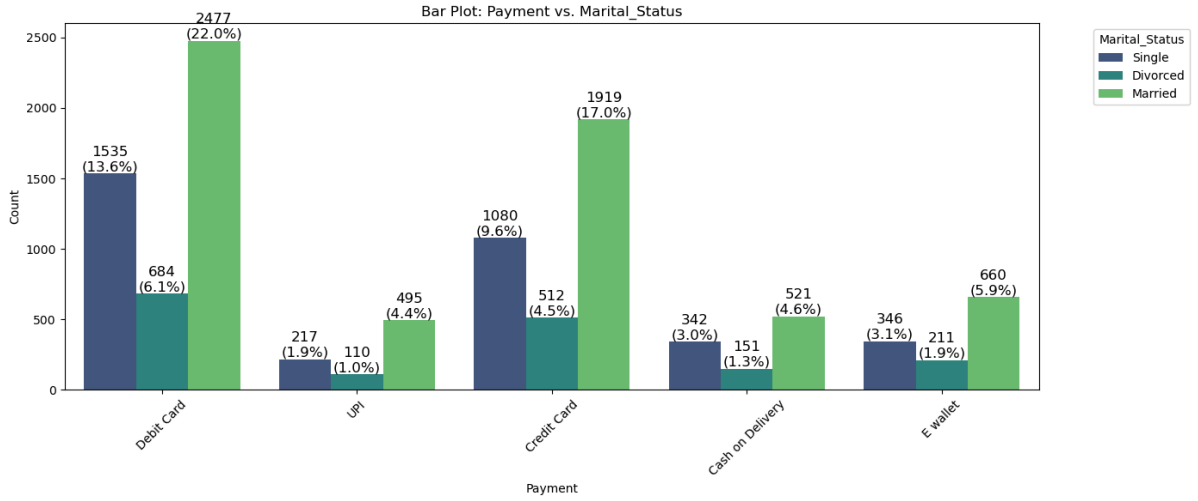


Figure-9 bivariate Analysis for Categorical Variables.

Observation:

1. **City Tier**
    - **Observation:** Shows a noticeable difference in distribution between churned and current customers.
    - **Insight:** Indicates that the city tier might influence customer churn and could be a useful predictor in the model.
2. **Account Segment**
    - **Observation:** Different distributions for churned vs. current customers.
    - **Insight:** Suggests that different account segments have varying churn rates and can impact the model's predictions.
3. **Marital Status**
    - **Observation:** Shows differences in the distribution.
    - **Insight:** Marital status might have an effect on churn behavior and could be relevant for the model.

Less Significant Variables:

1. **Gender**
    - **Observation:** Similar distributions within churned and current customers.
    - **Insight:** Gender may not significantly impact the churn prediction model.

Interpretation:

- The percentage distribution within each categorical variable provides insights into how these variables might contribute to churn prediction.
- Variables showing significant distribution differences between churned and current customers are likely to be more influential in the model.
- Those with similar distributions might have less impact on the model's performance.

2.3.4 Data Imbalance:

The dataset exhibits a clear imbalance regarding the target variable, which indicates whether a customer has churned or not. As shown in the plot below, for every 100 customers acquired, 17 have churned while 83 remain active. This distribution skews towards active customers. The goal of this exercise is to predict which customers will churn, focusing on the minority class '1'.

In a balanced dataset, there is an equal number of observations for each categorical value of the target variable (Churn). This balance allows the model to predict both classes with equal efficiency. However, customer churn typically involves fewer churned customers compared to active ones. In this dataset, approximately 17% of customers have churned, reflecting real-life churn data but posing modeling challenges.

Challenges of an Imbalanced Dataset:

Pattern Learning:

- With insufficient observations in the minority class, the model may struggle to learn its patterns effectively and may favor the majority class.
- **Solution**: Resampling techniques, such as oversampling the minority class or under-sampling the majority class, can help. Compared to under-sampling, which may result in data loss, oversampling using techniques like SMOTE (Synthetic Minority Oversampling) can be beneficial. SMOTE generates synthetic data for churned customers based on existing observations. However, it may not always improve model performance, and depending on the algorithm and type of SMOTE used, there is a risk of overfitting the training dataset.

Evaluation Metrics:

- Using accuracy as an evaluation metric in imbalanced classification problems is inappropriate. Even if the algorithm predicts all customers as belonging to the majority class, it would still yield an accuracy of 83.2%.
- **Solution**: Metrics such as Precision, Recall, or F1-score for the minority class offer a better evaluation approach.

Data Treatment:

SMOTE will be used as one of the data treatments. Models built on the imbalanced dataset will be compared against those built on the SMOTE-balanced dataset, and the model with the best performance metrics will be selected. Since SMOTE may not

always guarantee better performance, a comparison is essential to make an informed decision.

By addressing these challenges with appropriate techniques, the aim is to improve the model's ability to predict customer churn effectively.

2.3.5 One-way Anova:

Analysis of Variance is a statistical method, used to check if the means of two or more groups that are significantly different from each other. Hypothesis for the test is as follows:

H0: Means of all groups are equal

Ha: At least means of one pair of the groups is different

The Statsmodel library was used to perform the Anova test.

	Variable	F-statistic	Probability of > F	Inference at significance level of 5%
0	Tenure	632.814262	6.633734e-136	Reject null hypothesis. The means are differen...
1	CC_Contacted_LY	58.135818	2.643050e-14	Reject null hypothesis. The means are differen...
2	rev_per_month	5.513580	1.888659e-02	Reject null hypothesis. The means are differen...
3	CC_Agent_Score	125.902638	4.605513e-29	Reject null hypothesis. The means are differen...
4	Service_Score	0.899025	3.430638e-01	Cannot reject null hypothesis. The means are e...
5	Complain_ly	727.372270	2.691200e-155	Reject null hypothesis. The means are differen...
6	Account_user_count	124.378455	9.844343e-29	Reject null hypothesis. The means are differen...
7	rev_growth_yoy	2.156197	1.420237e-01	Cannot reject null hypothesis. The means are e...
8	coupon_used_for_payment	2.458905	1.168883e-01	Cannot reject null hypothesis. The means are e...
9	Day_Since_CC_connect	243.193686	2.919380e-54	Reject null hypothesis. The means are differen...
10	cashback	11.440972	7.208686e-04	Reject null hypothesis. The means are differen...

Table- 9 One-way Anova

At a significance level of 0.05 (5%), the tests for the variables `rev_growth_yoy`,Service_Score and `coupon_used_for_payment` returned p-values greater than 0.05. Consequently, we cannot reject the null hypothesis (Ho) for these variables. This indicates that there is no statistically significant difference in the means of the two groups (churn=0 and churn=1) for these variables. As such, they are not significant predictors of the target variable. This conclusion aligns with the visual observations made using the bivariate boxplots for these variables.

2.3.6 Chi-square:

Categorical variables were evaluated using the Chi-squared test of independence at a significance level of 0.05. This test helps determine if there's a statistically significant relationship between two categorical variables, aiding in the decision of whether to include them in the model. The Chi-square test of independence compares the frequency of each category for one variable across the categories of another variable. The results are typically displayed in a contingency table, where rows represent categories for one variable and columns represent categories for the other.

	Variable	chi2	p-value	chi2_output
0	City_Tier	80.288817	3.677095e-18	Reject Ho; Dependent.
1	Payment	103.799617	1.526348e-21	Reject Ho; Dependent.
2	Gender	8.983146	2.724812e-03	Reject Ho; Dependent.
3	account_segment	567.068402	2.073937e-121	Reject Ho; Dependent.
4	Marital_Status	379.808123	3.355165e-83	Reject Ho; Dependent.
5	Login_device	25.726928	3.933008e-07	Reject Ho; Dependent.

Table- 10 Chi-square test

A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct. Since the p-value returned for all the categorical variables is less than 0.05, the null hypothesis can be rejected. Hence at 5% level of significance, it may be concluded that churn is not independent of these categorical variables. Hence, we will proceed to retain all these categorical predictor variables at this point.

A p-value less than 0.05 (typically ≤ 0.05) is considered statistically significant, indicating strong evidence against the null hypothesis. It suggests that there is less than a 5% probability that the null hypothesis is true. Since the p-values for all the categorical variables are less than 0.05, we can reject the null hypothesis. Therefore, at the 5% significance level, we conclude that churn is not independent of these categorical variables. Consequently, we will retain all these categorical predictor variables for further analysis.

2.3.7 Encoding and Scaling of Dataset:

We will encode the dataset before the scaling process. We will do one hot encoding and label encoding on the categorical variables.

	Tenure	CC_Contacted_LY	rev_per_month	CC_Agent_Score	Service_Score	Complain_ly	Account_user_count	rev_growth_yoy	coupon_used_for_payment	Day_Since_C
0	-0.699363	-1.375600	1.249484	-0.772983	0.134508	1.582612	-0.742646	-1.382120	-0.431307	
1	-1.145091	-1.143070	0.579565	-0.048194	0.134508	1.582612	0.319919	-0.317598	-1.338485	
2	-1.145091	1.414758	0.244605	-0.048194	-1.247341	1.582612	0.319919	-0.583728	-1.338485	
3	-1.145091	-0.329216	0.914524	1.401386	-1.247341	-0.631867	0.319919	1.811447	-1.338485	
4	-1.145091	-0.678011	-0.760274	1.401386	-1.247341	-0.631867	-0.742646	-1.382120	-0.431307	

Table-11 Encoded and Scaled Dataset.

Before we perform Cluster on the dataset for Business Insight we will do Scaling on the dataset to make all the variables on the same scales. And for this we use Standard scalar method and obtain the given below encoded and scaled dataset in table-11.

Encoding the Categorical Variables

During the data pre-processing stage, different encoding techniques were applied to handle categorical variables in the dataset.

Ordinal Encoding for 'Account Segment':

- The 'Account Segment' variable, which reflects spending-based segments, was encoded using ordinal encoding.
- Categories like 'Regular', 'Regular Plus', 'Super', 'Super Plus', and 'HNI' were assigned numerical values from 1 to 5.

- This hierarchical encoding helps the model capture the varying significance of different segments based on spending behaviors.

One-Hot Encoding for Other Categorical Variables:

- Categorical variables such as 'Payment', 'Gender', 'Marital Status', and 'Login Device' were processed using one-hot encoding.
- One-hot encoding converts these categorical variables into binary columns for each category within the variable.

These encoding techniques ensure that the model can effectively utilize the categorical data during analysis and predictions.

Data Scaling and Preparation for Clustering

Scaling the data is crucial for clustering to ensure that all features receive equal treatment and to prevent any bias caused by the different scales of variables. Here's how the data has been prepared for clustering:

1. **Standardization of Numerical Variables**:
    - The numerical variables were scaled using standardization. This transformation adjusts the data to have a mean of 0 and a standard deviation of 1.
    - **Purpose**: Ensures that all numerical features are on the same scale, allowing for fair and unbiased clustering analysis.
2. **Exclusion of the Target Variable 'Churn'**:

- The target variable 'Churn' was excluded from scaling to maintain its original interpretation.
- **Purpose**: To preserve the binary nature of the target variable for accurate analysis and predictions.

2.3.8 Clustering of dataset:

Clustering is an unsupervised machine learning task that groups a set of objects such that objects in the same group (or cluster) are more similar to each other than to those in other groups.

It helps in identifying natural groupings within the dataset based on similarities.

The cluster profile was formed by grouping observations by clusters and finding the mean for all features.

This helped in understanding the characteristics of each cluster.

Clustering helps in identifying similar customer groups, enabling targeted strategies for different segments.

Including churn in the cluster profile allows for better understanding of customer retention patterns across different groups.

Hierarchical Clustering method:

For, Hierarchical Clustering we will plot a Dendrogram using ward's linkage method and Euclidean distance and truncating dendrogram up to p value to 10 and obtain the following plot as shown below.

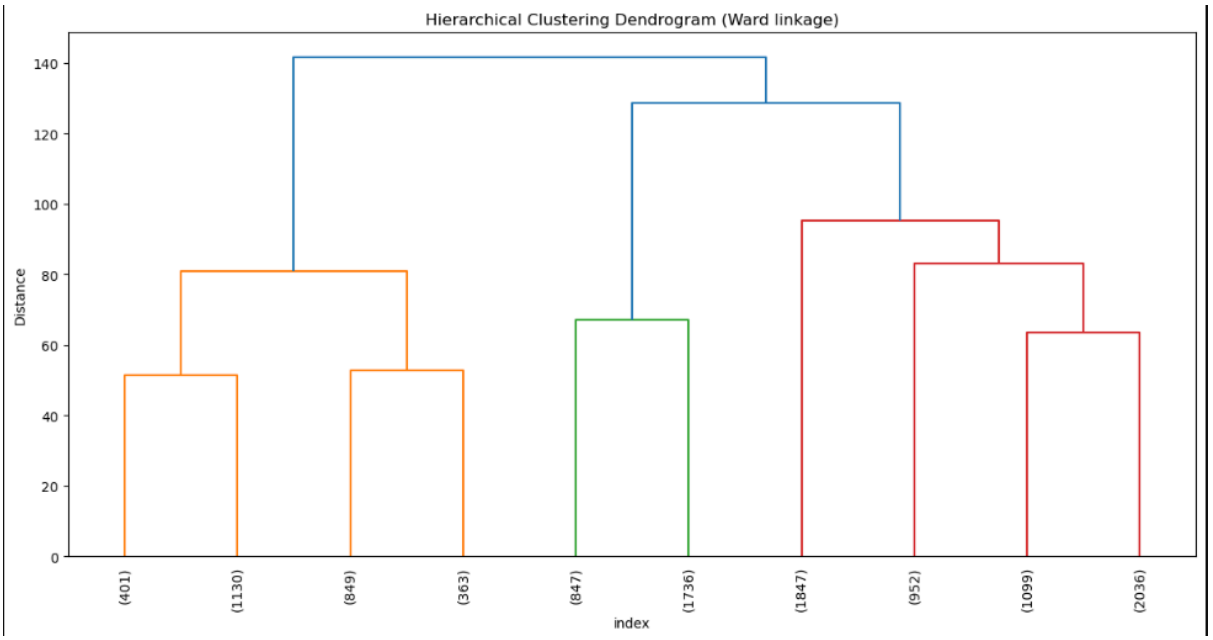


Figure10 - Dendrogram

Kmean Method for clustering:

will find out the WSS (within sum of squares) values for 10 clusters and plot them in elbow plot to find out the optimum number of K for K-Means algorithm. We can see the plot below.

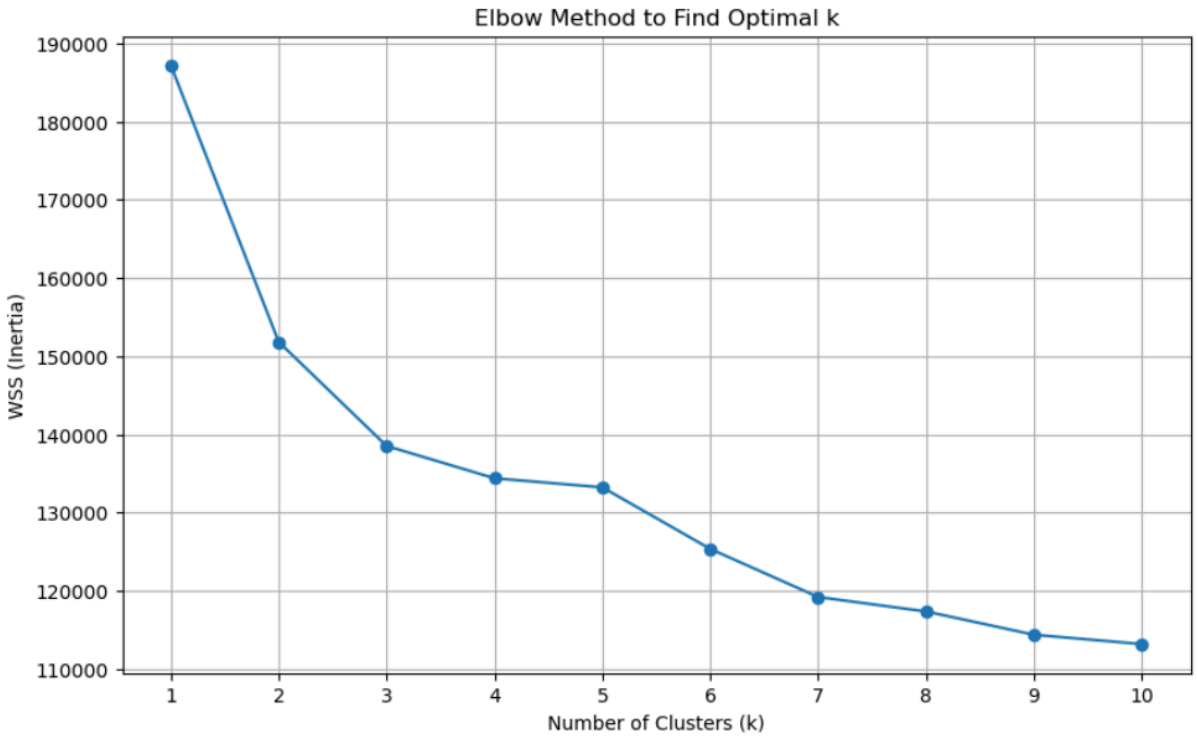


Figure-11 Elbow Plot

From the above plot and Average Silhouette score for all 10 clusters we can say that no. of clusters or K value should be 2 for carrying out the K-Means clustering.

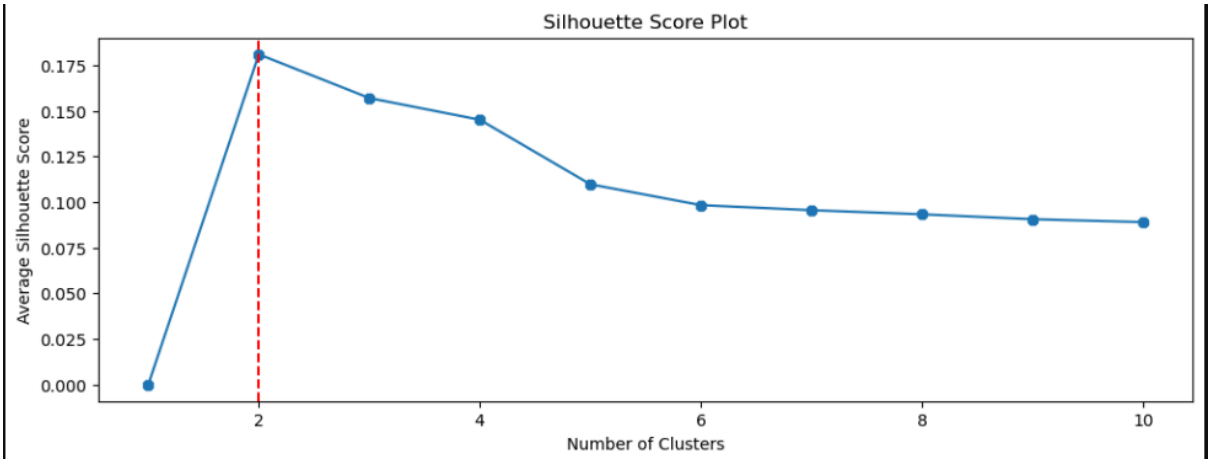


Figure-12 Silhouette score chart

Thus, We will create a cluster based on the K optimized and then add it to the dataset.

	Churn	Kmeans	clusters_maxclust	clusters_distance
0	1	0	2	2
1	1	0	2	2
2	1	0	2	2
3	1	1	3	3
4	1	1	4	4

11255	0	1	4	4
11256	0	0	1	1
11257	0	0	2	2
11258	0	0	1	1
11259	0	1	4	4

Table-12 Churn vs Kmeans and Dendrogram Clusters

Here, Cluster_maxclust and cluster_distance is obtained from Dendrogram and ward's linkage method and the Kmean column is obtained from the K-means clustering method itself.

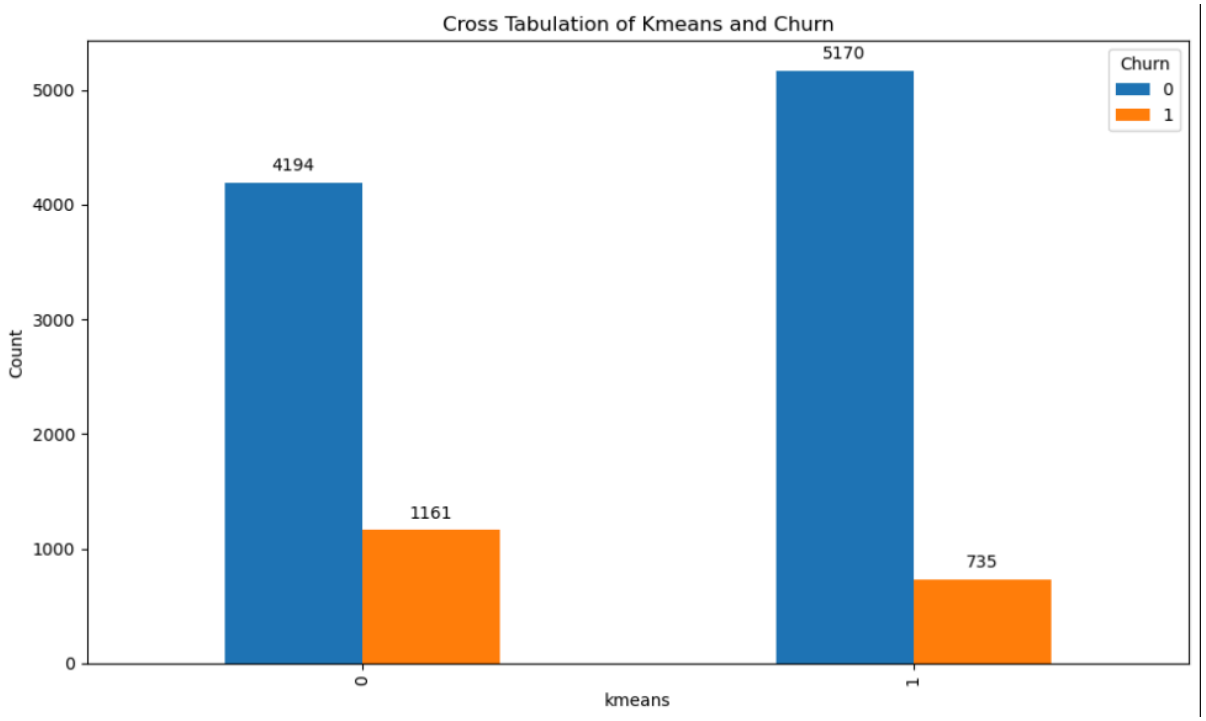


Figure-13 Kmeans vs Churn

Here we can see in figure-13 showing K Means vs churn for each cluster formed. And due to data imbalance we can see that while modeling there are going to be biassed towards some values in very variable.

3. Business Insights:

3.1 Business Insights from Cluster Profiling

1. **Clear Group Separation**:
    - Despite using an unsupervised algorithm without the target variable for clustering, clear separation of groups was achieved for features showing high F-statistic and Chi-square values in bivariate analysis.

2. **Tenure and Churn**:
    - **Observation:** Higher tenure generally indicates lower churn.
    - **Insight:** However, this trend does not hold for low tenures (clusters 1 and 0). Further analysis by binning tenure and checking the relationship with churn through a stacked bar plot would be beneficial.
3. **Account Segment and Churn**:
    - **Observation:** Regular Plus customers exhibit higher churn compared to other segments.
    - **Insight:** This indicates that the Regular Plus segment might need targeted retention strategies.
4. **Customer Care Contact**:
    - **Observation:** High churn clusters contacted customer care almost twice as often as low churn clusters.
    - **Insight:** Frequent customer care interactions could be an indicator of dissatisfaction or issues needing resolution.
5. **Cashback**:
    - **Observation:** Lower cashback in high churn clusters compared to low churn clusters.
    - **Insight:** Offering better cash back incentives might help retain customers.
6. **Complaints and Churn**:
    - **Observation:** Clusters with maximum complaints last year also have the highest churn.

- **Insight:** Resolving complaints effectively is crucial. Strengthening the follow-up process for customer complaints until resolution is achieved could improve retention.

3.2 Business Insights from Exploratory Data Analysis (EDA)

1. **Customer Feedback:**
   - **Service Ratings**: 78% of customers rated the service as 3 or lower (out of 5).
   - **Customer Care Agent Ratings**: 61% of customers rated customer care agents as 3 or lower.
   - **Insight**: Indicates dissatisfaction or minimal satisfaction in both service and customer care engagement. Focus areas for improvement.
2. **Relationship between Tenure and Churn:**
   - **Observation**: Churn is very high for low tenures. Approximately 51.85% of customers with tenure between 0 and 1 have churned.
   - **Insight**: The proportion of churn decreases with increasing tenure. Understanding and addressing reasons for early churn is crucial.
3. **Relationship between Account Segment and Churn:**
   - **Observation**: Higher churn among customers in the Regular_Plus plan.
   - **Insight**: Compare this plan with competitors' plans to determine if changes in features or pricing are needed. Collect and analyze customer feedback for this segment to understand churn reasons.

4. **Relationship between Monthly Revenue and Churn:**
    - **Observation**: Higher churn percentage among high revenue customers compared to lower revenue customers.
    - **Insight**: Concerning for the DTH provider as high-value customers are also churning. Focus on retention strategies for high-revenue customers.
5. **Relationship between Days Since Customer Care Connect and Churn:**
    - **Observation**: Churned customers have fewer days since their last customer care contact compared to active customers.
    - **Insight**: Churn tends to occur shortly after customer care contact, indicating potential dissatisfaction with the resolution process.
6. **Relationship between Customer Care Contact Last Year and Churn:**
    - **Observation**: Churned customers contacted customer care more frequently last year compared to active customers.
    - **Insight**: Frequent customer care interactions could be an indicator of underlying issues that need resolution.
7. **Relationship between Complaints Made Last Year and Churn:**
    - **Observation**: A higher proportion of churned customers made complaints compared to active customers.
    - **Insight**: Effective complaint resolution is critical for reducing churn.
8. **Relationship between User Count and Churn:**
    - **Observation**: Accounts with 5 or 6 users have a higher proportion of churned customers compared to active customers.

- **Insight**: Large user count accounts may need targeted retention strategies.

9. **Relationship between Payment Type and Churn:**
    - **Observation**: Higher churn among customers using E-wallet and Cash on Delivery compared to active customers.
    - **Insight**: Payment method might influence churn. Enhancing the experience for these payment methods could help.
10. **Relationship between City Tier and Churn:**
    - **Observation**: Higher churn among customers residing in Tier 3 cities compared to active customers.
    - **Insight**: Investigate competition and other factors in Tier 3 cities to address churn.
11. **Relationship between Marital Status and Churn:**
    - **Observation**: Single customers have a higher churn rate compared to married or divorced customers.
    - **Insight**: Develop strategies targeted at single customers to improve retention.

These insights provide a detailed understanding of the factors influencing customer churn and can guide strategic interventions to improve customer retention and satisfaction.

3.3 Business Recommendations:

The following are the Business Recommendations based solely on EDA and Clustering performed in this report.

Customer Engagement and Retention

1. **Loyalty Programs**:
    - **Insight**: Develop loyalty programs to reward long-tenure customers. This can help retain valuable customers and reduce churn rates among newer customers.
2. **Personalized Offers**:
    - **Insight**: Tailor promotional offers based on customer segments. For instance, offering special discounts or benefits to Regular Plus customers could reduce their higher churn rates.
3. **Proactive Customer Support**:
    - **Insight**: Identify customers with frequent complaints or those who contacted customer care often and provide proactive support to address their issues before they consider churning.

Payment Method Preferences

1. **Payment Method Optimization**:
    - **Insight**: Enhance and promote the most preferred payment methods like Debit and Credit Cards to encourage higher spending and improve satisfaction.
2. **Special Incentives for E-wallet and Cash on Delivery**:

- **Insight**: Introduce special incentives for customers using E-wallets and Cash on Delivery to increase their satisfaction and reduce churn.

Customer Satisfaction

1. **Improve Service Quality**:
   - **Insight**: Focus on improving overall service quality based on customer feedback, aiming to address areas causing dissatisfaction.
2. **Customer Care Training**:
   - **Insight**: Enhance training for customer care agents to improve their interaction quality and satisfaction scores.

Segment-Specific Strategies

1. **Targeting High-Value Customers**:
   - **Insight**: Pay special attention to high-revenue customers. Implement strategies like personalized services or exclusive offers to retain them.
2. **City-Specific Campaigns**:
   - **Insight**: Develop specific marketing campaigns for customers in Tier 3 cities, where churn rates are higher, to combat competition and improve retention.

Data-Driven Decision Making

1. **Regular Monitoring**:

- **Insight**: Continuously monitor and analyze key metrics such as churn rates, customer feedback, and complaint frequency to identify and address issues promptly.

2. **Utilize Predictive Analytics**:
    - **Insight**: Implement predictive analytics to identify customers at risk of churning and take preemptive actions to retain them.

These insights, combined with the ones you've already identified, can help drive strategic decisions and improve overall business performance.