Capstone Project - PGP-DSBA

Notes 1

Isha Shukla
10 August 2024

Contents

SL No.	Title	Page No.
1	Problem Statement	3
2	Need of the study/project	3
3	Understanding business/ social opportunity	3
4	Understanding how data was collected in terms of time, frequency and methodology	5
5	Visual inspection of data (rows, columns, descriptive details)	6
6	Understanding of attributes	10
7	Variable transformation	12
8	Univariate analysis	13
9	Bivariate analysis	25
10	Removal of unwanted variables	41
11	Missing Value treatment	44
12	Outlier treatment	46
13	Variable transformation(Contd)	47
14	Addition of new variables	47
15	Is the data unbalanced? If so, what can be done?	48
16	Any business insights using clustering	49
17	Any other business insights	49
18	Other Insights	51

1. Introduction of the business problem

a. Problem Statement

The DTH provider is facing increasing competition, leading to challenges in retaining existing customers. The company needs to develop a churn prediction model to identify accounts at risk of churning. Given that one account may represent multiple customers, losing an account can significantly impact the business. The goal is to accurately predict potential churners and design targeted campaigns to retain them. These campaigns must be cost-effective, as recommendations will be reviewed by the revenue assurance team, which will reject any strategies that result in a financial loss to the company. The challenge is to create a model that maximizes customer retention while minimizing costs.

b. Need of the study/project

The study is essential due to the intense competition in the DTH market, where retaining customers is increasingly challenging. Given that losing one account can mean losing multiple customers, predicting churn is critical for targeted retention efforts. The project aims to develop a data-driven churn prediction model to identify at-risk accounts and propose cost-effective, personalized retention strategies. This understanding will empower the company to make informed decisions that will help sustain and grow its customer base while minimizing losses and ensuring profitability.

c. Understanding business/social opportunity

This case study focuses on a DTH company where customers are assigned unique account IDs, with a single account ID potentially serving multiple customers (such as in a family plan). Customers are diverse in terms of gender, marital status, and payment methods, which include a variety of

options such as debit card, UPI, and cash on delivery. The company's customer base is also segmented according to the types of plans they opt for based on their usage and the devices they use (computer or mobile). Additionally, customers benefit from cashback offers on bill payments.

The success of the business largely depends on customer loyalty and the perceived value of the services provided. Offering quality service, along with value-added features, plays a crucial role in retaining customers. Running promotional and festival offers is another strategy that not only attracts new customers but also helps in retaining existing ones. The company faces a significant challenge because losing one account ID can result in losing multiple customers, which directly impacts the business's revenue and market standing.

In a market where having a DTH connection is almost a necessity for every household, competition is fierce. The company's ability to differentiate itself from competitors hinges on understanding and addressing the factors that drive customer loyalty and satisfaction.

The business opportunity in this project lies in the ability to proactively address customer churn, which is a significant challenge in the competitive DTH industry. By developing a churn prediction model, the company can identify accounts at risk of leaving and implement targeted retention strategies. This not only helps in preserving revenue by retaining existing customers but also reduces the cost associated with acquiring new customers. Furthermore, understanding the factors that contribute to churn enables the company to refine its product offerings, improve customer satisfaction, and strengthen customer loyalty. This approach positions the company to maintain a competitive edge in the market, ensuring sustained growth and profitability.

2. Data Report

Variable	Description
AccountID	account unique identifier
Churn	account churn flag (Target)
Tenure	Tenure of account
City_Tier	Tier of primary customer's city
CC_Contacted_L12m	How many times all the customers of the account has contacted customer care in last 12months
Payment	Preferred Payment mode of the customers in the account
Gender	Gender of the primary customer of the account
Service_Score	Satisfaction score given by customers of the account on service provided by company
Account_user_count	Number of customers tagged with this account
account_segment	Account segmentation on the basis of spend
CC_Agent_Score	Satisfaction score given by customers of the account on customer care service provided by company
Marital_Status	Marital status of the primary customer of the account
rev_per_month	Monthly average revenue generated by account in last 12 months
Complain_l12m	Any complaints has been raised by account in last 12 months
rev_growth_yoy	revenue growth percentage of the account (last 12 months vs last 24 to 13 month)
coupon_used_l12m	How many times customers have used coupons to do the payment in last 12 months
Day_Since_CC_connect	Number of days since no customers in the account has contacted the customer care
cashback_l12m	Monthly average cashback generated by account in last 12 months
Login_device	Preferred login device of the customers in the account

a. Understanding how data was collected in terms of time, frequency and methodology

- Dataset has 11260 rows and 19 columns.
- Target variable is “Churn” (1 dependent) and 18 variables are independent.
- Dataset has 11260 unique AccountID.
- “Churn” shows 1 if customer is churned and if it is not churned it shows 0.
- Dataset is a mix of categorical and continous variables.

1. In terms of time:

- **CC_Contacted_L12m, Complain_l12m, coupon_used_l12m, cashback_l12m:** These features suggest a 12-month rolling period, where data is collected on customer activities or metrics over the last 12 months.
- **rev_per_month, rev_growth_yoy:** These are likely calculated on a monthly basis or as year-over-year comparisons.

2. Frequency:

- **CC_Contacted_L12m:** Data on how many times the customer contacted customer care could be collected continuously, with a tally kept over the 12-month period.

- **coupon_used_l12m:** The frequency of coupon usage over the past 12 months is likely recorded each time a coupon is used.
- **cashback_l12m:** Monthly average cashback is likely calculated from individual transactions that offer cashback rewards.
- **Day_Since_CC_connect:** The count of days since the last customer care contact could be updated daily.

3. Methodology:

- **Payment, Gender, Marital_Status, account_segment**: These categorical variables might have been collected from customer account records, surveys, or registration forms.
- **Service_Score, CC_Agent_Score:** These scores could be gathered through customer satisfaction surveys or feedback forms.
- **Account_user_count:** This is likely tracked through the account management system, recording how many users are linked to each account.

b. Visual inspection of data (rows, columns, descriptive details)

	AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score
0	20000	1	4	3.0	6.0	Debit Card	Female	3.0	3	Super	2.0
1	20001	1	0	1.0	8.0	UPI	Male	3.0	4	Regular Plus	3.0
2	20002	1	0	1.0	30.0	Debit Card	Male	2.0	4	Regular Plus	3.0
3	20003	1	0	3.0	15.0	Debit Card	Male	2.0	4	Super	5.0
4	20004	1	0	1.0	12.0	Credit Card	Male	2.0	3	Regular Plus	5.0

First five rows of the dataset

- Dataset has 11260 rows and 19 columns.
- There are no duplicates present in the dataset.
- There are null values in the dataset.

```
(11260, 19)
```

Shape of the data

```
<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)
memory usage: 1.6+ MB
```

Info about the dataset

- There are 5 float datatype columns, 2 integer datatype and 12 object datatypes at the starting.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
AccountID	11260.0	NaN	NaN	NaN	25629.5	3250.62635	20000.0	22814.75	25629.5	28444.25	31259.0
Churn	11260.0	NaN	NaN	NaN	0.168384	0.374223	0.0	0.0	0.0	0.0	1.0
Tenure	11158.0	38.0	1.0	1351.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
City_Tier	11148.0	NaN	NaN	NaN	1.653929	0.915015	1.0	1.0	1.0	3.0	3.0
CC_Contacted_LY	11158.0	NaN	NaN	NaN	17.867091	8.853269	4.0	11.0	16.0	23.0	132.0
Payment	11151	5	Debit Card	4587	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	11152	4	Male	6328	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Service_Score	11162.0	NaN	NaN	NaN	2.902526	0.725584	0.0	2.0	3.0	3.0	5.0
Account_user_count	11148.0	7.0	4.0	4569.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
account_segment	11163	7	Super	4062	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CC_Agent_Score	11144.0	NaN	NaN	NaN	3.066493	1.379772	1.0	2.0	3.0	4.0	5.0
Marital_Status	11048	3	Married	5860	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rev_per_month	11158.0	59.0	3.0	1746.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Complain_ly	10903.0	NaN	NaN	NaN	0.285334	0.451594	0.0	0.0	0.0	1.0	1.0
rev_growth_yoy	11260.0	20.0	14.0	1524.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
coupon_used_for_payment	11260.0	20.0	1.0	4373.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Day_Since_CC_connect	10903.0	24.0	3.0	1816.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cashback	10789.0	5693.0	155.62	10.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Login_device	11039	3	Mobile	7482	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Describing the dataset

- The `df.describe()` function in pandas provides a summary of statistical metrics for numerical columns in a DataFrame. It returns key statistics such as the count, mean, standard deviation, minimum, maximum, and quartile values (25%, 50%, and 75%). For categorical data, using `df.describe(include='all')` provides additional information like the number of unique values, the most frequent value, and its frequency. This function is useful for quickly understanding the distribution and spread of the data in your DataFrame.
- No duplicates found in the dataset.
- Unique entries in the dataset.

```
AccountID                11260
Churn                        2
Tenure                      38
City_Tier                    3
CC_Contacted_LY             44
Payment                      5
Gender                       4
Service_Score                6
Account_user_count           7
account_segment              7
CC_Agent_Score               5
Marital_Status               3
rev_per_month               59
Complain_ly                  2
rev_growth_yoy              20
coupon_used_for_payment     20
Day_Since_CC_connect        24
cashback                  5693
Login_device                 3
dtype: int64
```

No. of unique entries in each column

- **AccountID:** 11,260 unique values
- **Churn:** 2 unique values (likely indicating binary classification: 0 = No, 1 = Yes)
- **Tenure:** 38 unique values (duration of account in months or years)
- **City_Tier:** 3 unique values (categorical, possibly representing different tiers of cities)

- **CC_Contacted_LY:** 44 unique values (number of times contacted in the last year)
- **Payment:** 5 unique values (likely representing different payment types or categories)
- **Gender:** 4 unique values (likely includes multiple categories or may need correction if intended to be binary
- **Service_Score:** 6 unique values (ratings or scores for service quality)
- **Account_user_count:** 7 unique values (number of users per account)
- **account_segment:** 7 unique values (different segments or categories of accounts)
- **CC_Agent_Score:** 5 unique values (ratings or scores for contact center agents)
- **Marital_Status:** 3 unique values (marital status categories)
- **rev_per_month:** 59 unique values (monthly revenue)
- **Complain_ly:** 2 unique values (complaints in the last year, binary)
- **rev_growth_yoy:** 20 unique values (year-over-year revenue growth)
- **coupon_used_for_payment:** 20 unique values (number of times a coupon was used)
- **Day_Since_CC_connect:** 24 unique values (days since the last contact with the customer care)
- **cashback:** 5,693 unique values (amount of cashback received)
- **Login_device:** 3 unique values (types of devices used to log in)

c. Understanding of attributes

1. **AccountID:** A unique identifier for each account. No null values present.
2. **Churn:** A binary variable indicating whether a customer has churned (1) or not (0). This is the target variable in a churn prediction model.
3. **Tenure:** The length of time an account has been active, measured in months. It can be an indicator of customer loyalty.

4. **City_Tier:** A categorical variable representing different tiers or levels of cities, which might correlate with customer behavior or service needs. It segregates cutomer in 3 parts ‘Tier 1’,’Tier 2’ and ‘Tier 3’.
5. **CC_Contacted_LY:** The number of times the customer has been contacted by customer care in the last year(or in last 12 months). This may reflect engagement or service issues.
6. **Payment:** Represents 5 different payment methods or types, which can influence customer behaviour for payment method. Categorical variable.
7. **Gender:** Categorical variable indicating the gender of the customer.
8. **Service_Score:** A score reflecting the quality of service provided to the customer, potentially impacting their satisfaction and likelihood to churn. This variable is categorical in nature.
9. **Account_user_count:** The number of users associated with an account, which can affect how the account is managed or used.
10. **account_segment:** Classification of accounts into different segments based on certain criteria, which can be useful for targeted marketing or service strategies.This variable is categorical in nature.
11. **CC_Agent_Score:** A score for the performance of customer care agents, which may influence customer experience and retention.This variable is categorical in nature.
12. **Marital_Status:** Categorical variable indicating the marital status of the customer in 3 different types of status.
13. **rev_per_month:** Monthly revenue generated from the customer. It is a key financial metric for assessing customer value and profitability.This variable is categorical in nature.
14. **Complain_ly:** A binary variable indicating whether the customer has raised a complaint in the last year. This can be a predictor of churn or satisfaction.
15. **rev_growth_yoy:** Year-over-year revenue growth, which indicates how the customer’s revenue contribution is changing over time in percentage of the account in last 12 months vs last 24 to 13 month.

16. **coupon_used_for_payment:** The number of times the customer has used a coupon for payment in last 12 months. This can influence revenue and customer loyalty.
17. **Day_Since_CC_connect:** The number of days since the customer last interacted with customer care, which can be related to engagement or support needs.
18. **cashback:** The monthly average cashback generated by account in last 12 months, which may impact customer satisfaction and loyalty.
19. **Login_device:** The type of device used by the customer to log in, which can provide insights into user preferences and behavior.

3. Exploratory Data Analysis

a. Variable transformation

1. **Payment**:
   - **Observation**: The most common payment method is Debit Card, followed by Credit Card. UPI is the least preferred.
   - **Action**: No immediate issues detected.
2. **Gender**:
   - **Observation**: There are unexpected values ('M' and 'F') in addition to 'Male' and 'Female'.
   - **Action**: Correct data entry errors by mapping 'M' to 'Male' and 'F' to 'Female'.
3. **account_segment**:
   - **Observation**: Categories like 'Regular+', 'Super+', and 'Super' are present.
   - **Action**: Standardize segments by correcting data entry errors (e.g., 'Regular +' to 'Regular Plus', 'Super +' to 'Super Plus').
4. **Marital_Status**:
   - **Observation**: Most customers are married,

```
Unique values in Payment are :
Payment
Debit Card          4587
Credit Card         3511
E wallet            1217
Cash on Delivery    1014
UPI                  822
Name: count, dtype: int64


----------------------------------------

Unique values in Gender are :
Gender
Male      6328
Female    4178
M          376
F          270
Name: count, dtype: int64


----------------------------------------

Unique values in account_segment are :
account_segment
Super           4062
Regular Plus    3862
HNI             1639
Super Plus       771
Regular          520
Regular +        262
Super +           47
Name: count, dtype: int64


----------------------------------------

Unique values in Marital_Status are :
Marital_Status
Married     5860
Single      3520
Divorced    1668
Name: count, dtype: int64


----------------------------------------

Unique values in Login_device are :
Login_device
Mobile      7482
Computer    3018
&&&&         539
Name: count, dtype: int64
```

Unique values in the categorical columns

followed by single and divorced.

- **Action**: No issue found.

5. **Login_device**:

- **Observation**: Includes ‘Mobile’, ‘Computer’, and a questionable value ‘&&&&’.
- **Action**: Investigate if ‘&&&&’ represents missing data or an unknown category. Consider replacing it with Other.

Cleaning and correcting these categories will help ensure more accurate analysis and model performance.

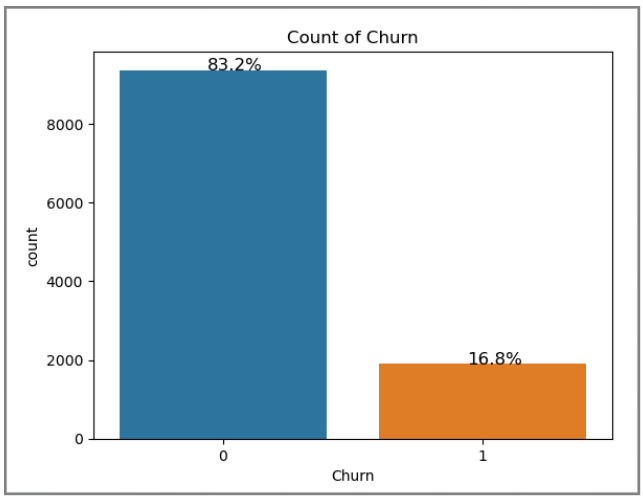
```
Unique values in Gender are :
Gender
Male      6704
Female    4448
Name: count, dtype: int64

----------------------------------------

Unique values in account_segment are :
account_segment
Regular Plus    4124
Super           4062
HNI             1639
Super Plus       818
Regular          520
Name: count, dtype: int64
----------------------------------------

Unique values in account_segment are :
Login_device
Mobile      7482
Computer    3018
Other        539
Name: count, dtype: int64
```

After correcting the values

- There are one or more values such as **'#','@','$','+'** and **'*'** present in columns ‘**Tenure’, ‘Account_user_count’, ‘rev_per_month’, 'rev_growth_yoy', 'coupon_used_for_payment', 'Day_Since_CC_connect' and ‘cashback**’.
- Replacing these values with np.nan in order to do the EDA.

b. Univariate analysis

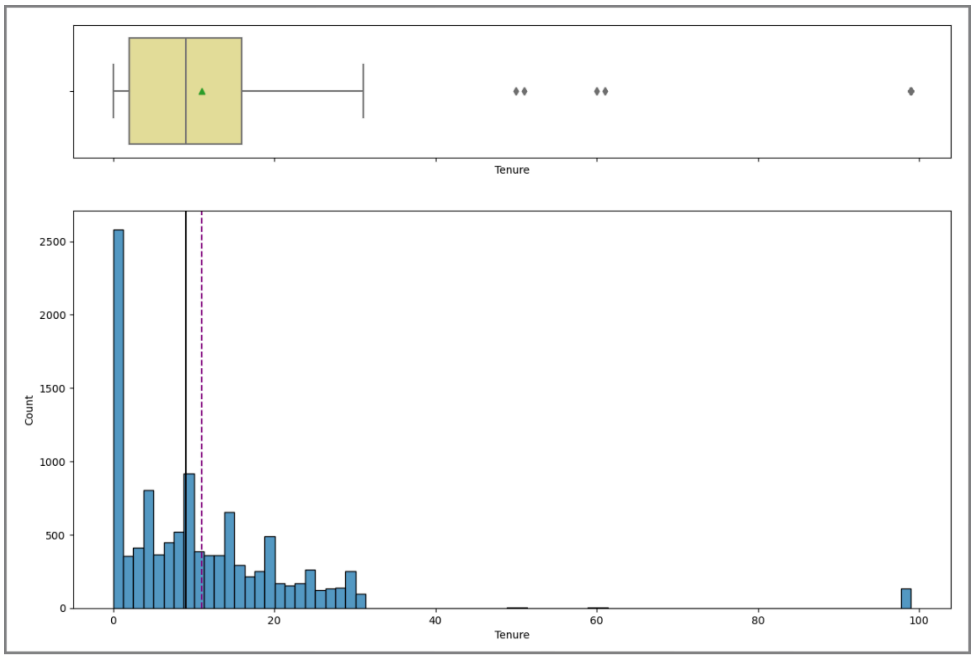
1. Churn



Count of churn

- 16.8% of customers have churned whereas 83.2% of them have not i.e., this is an imbalanced classification problem.

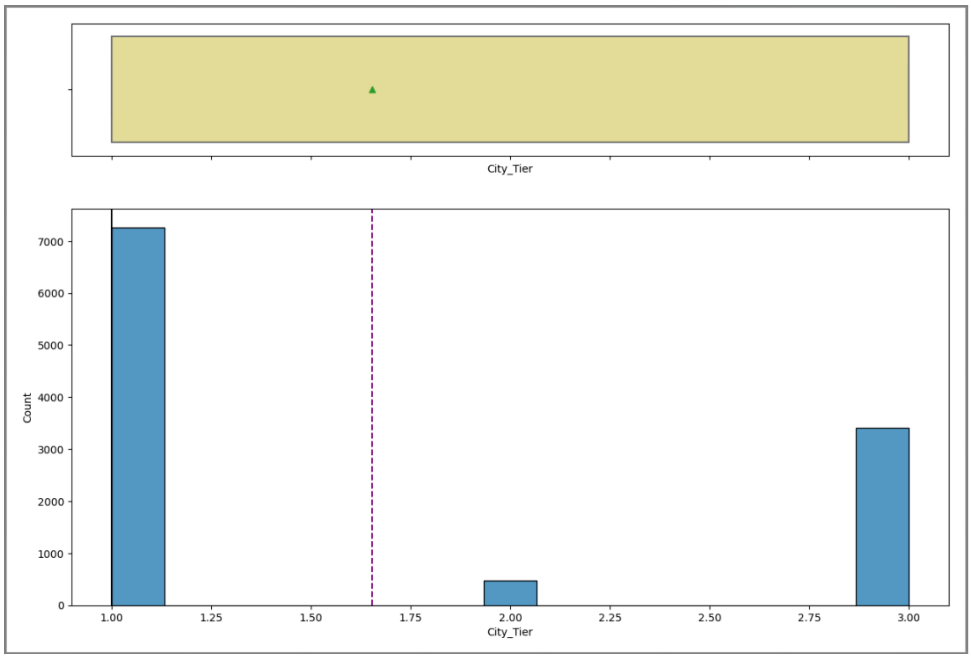
2. Tenure



Boxplot and histogram of Tenure

- Highest number of customers with a tenure of less than a month.
- Outliers around 50, 65 and about 100 months.

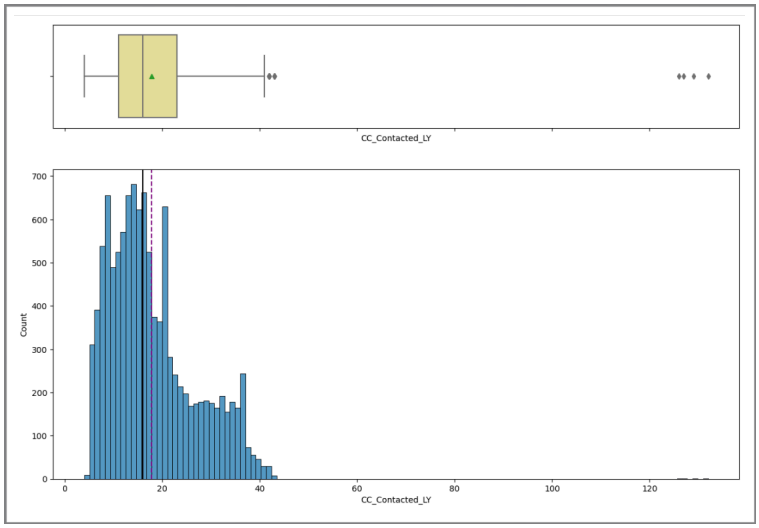
3. City_tier



Boxplot and histogram of City_tier

- Most customers are from Tier1(7263) cities followed by Tier3(3405) cities.
- Tier 2 has the least number of customers 480.
- City_Tier can be converted to categorical variable. So, we will convert it into object datatype.

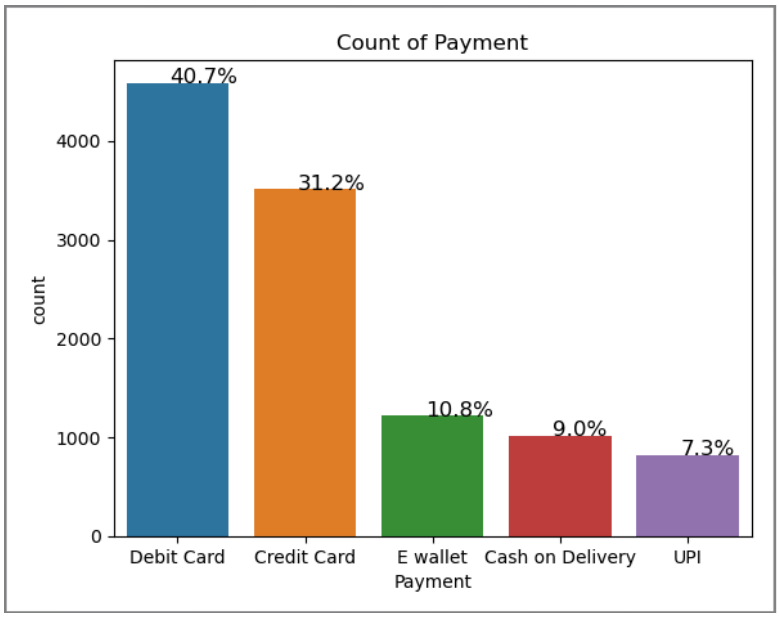
4. CC_Contacted_LY



Boxplot and histogram of CC_Contacted_LY

- The distribution suggests that the higher number of times the customer care was contacted is between 11 to 23 times.
- Outliers around 45 times and way beyond 125 times.

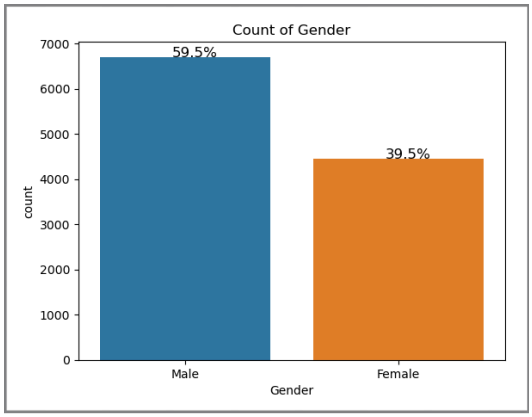
5. Payment



Payment countplot

- 40.7% of the customers prefer payment with Debit Card(highest) and 31.2% with Credit Card.
- E-wallet Payment are considered by 10.8% of the customer.
- 9% of people prefer Cash on Delievery.
- UPI is the least preferred payment mode.

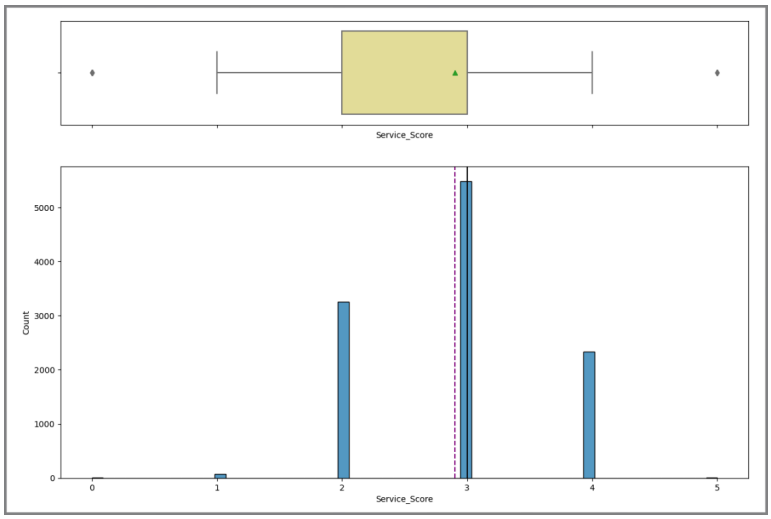
6. Gender



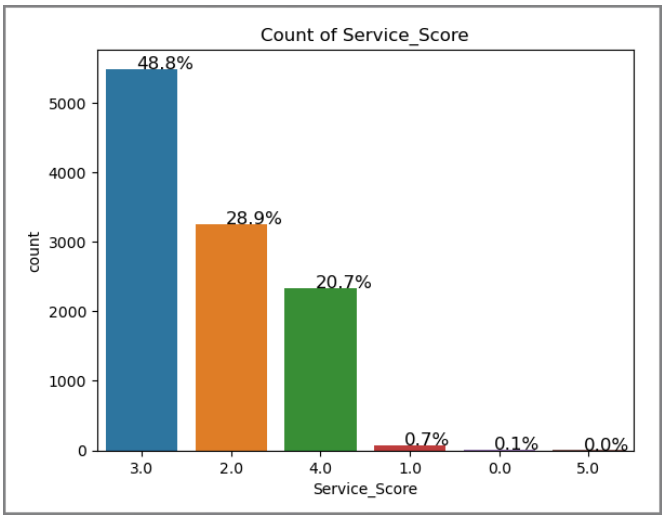
Countplot of Gender

- Most of the customers are Male (59.5%).
- 39.5% of customers are females.

7. Service_score



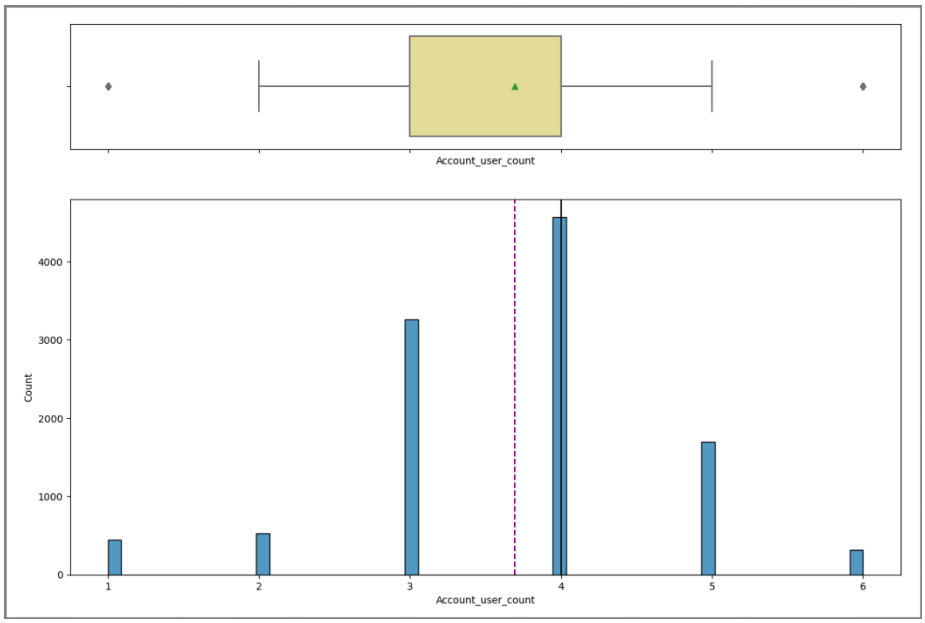
Service_score boxplot and histogram



Service_score countplot

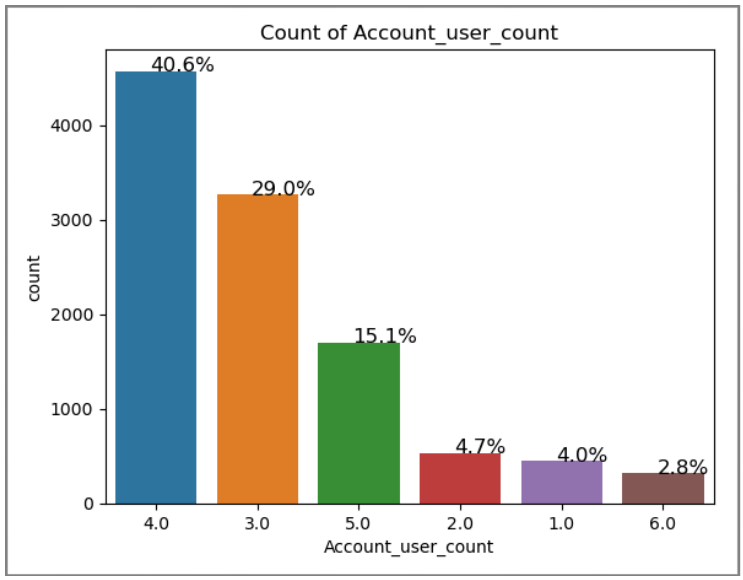
- Outliers at 0 and 5.
- Highest service score is 3 which means customers are not fully statisfied.
- We will convert this column to categorical.

8. Account_user_count



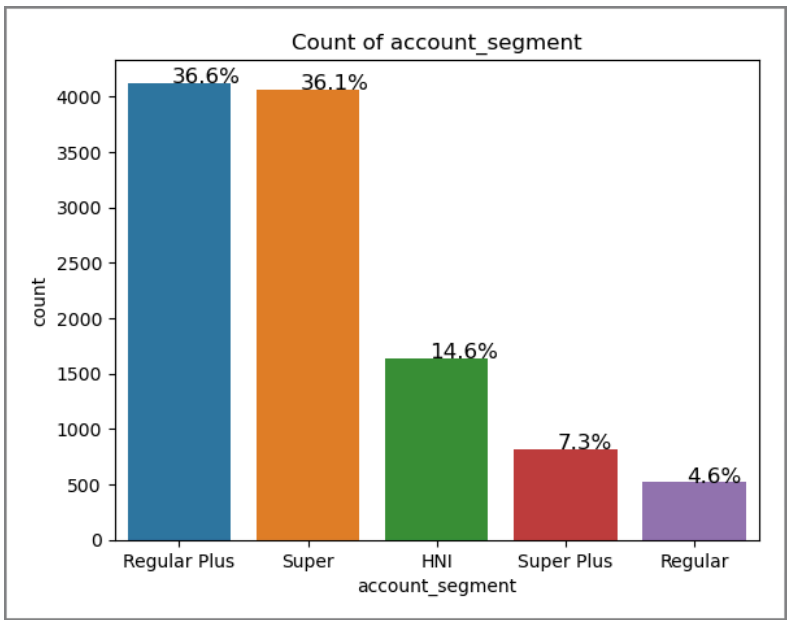
Account_user_count boxplot and histogram

- Large number of accounts are tagged with 4, followed by 3 and 5 customers per account.
- Outliers at 1 and 6 customers per account.
- Account_user_count will be converted to categorical variable.
- We will convert it into categorical datatype.



Count of Account_user_count

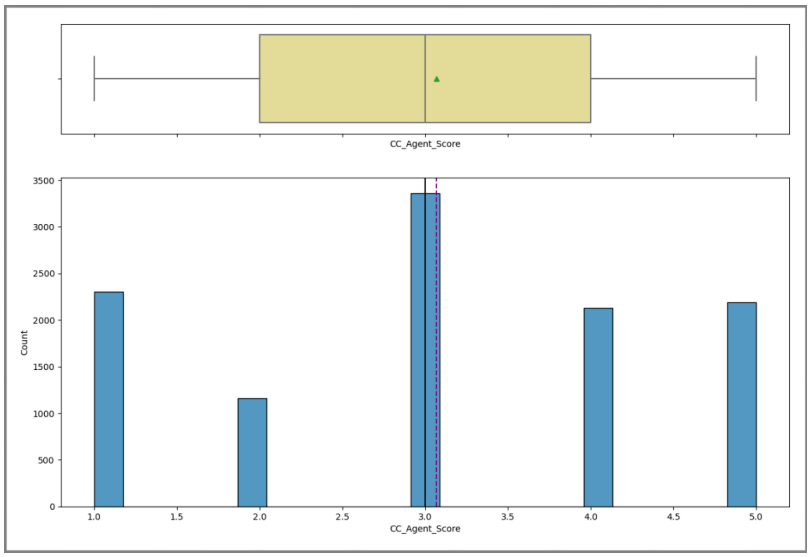
9. account_segment

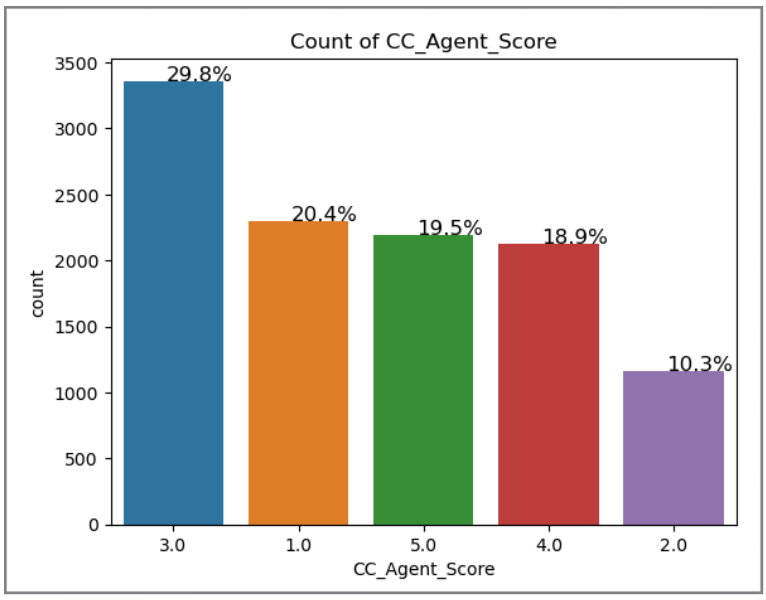


Count of account_segment

- Segments 'Regular Plus' and 'Super' have the maximum of customers above 36%.
- Segment 'Regular' has the least number of customers.
- High Net Income (HNI) segment comprise of 14.6% of customers followed by Super Plus which comprises 7.3% of customers.

10. CC_Agent_Score

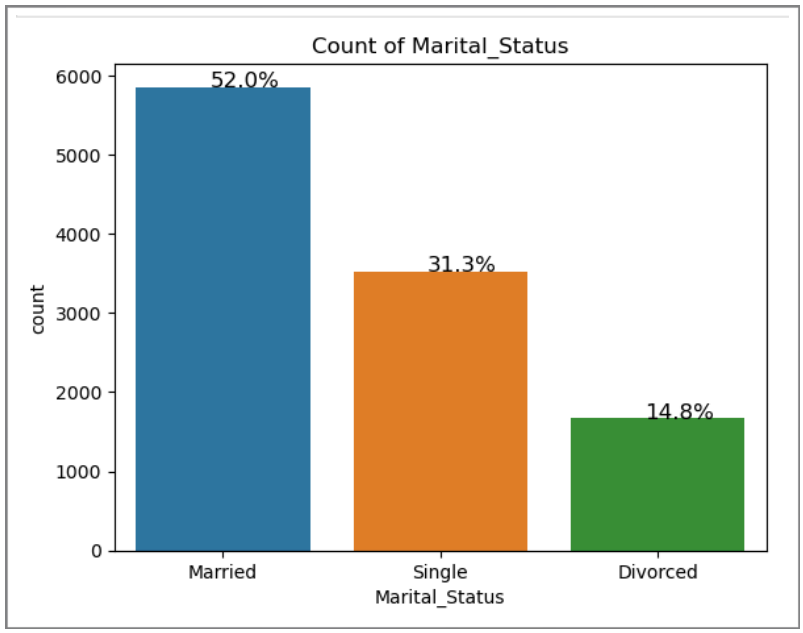




- Highest Customer Agent Score is 3.
- It appears that nearly the same number of customers have given excellent scores of 4 and 5, as well as low scores of 1.

- The distribution suggests that the customer service has almost equal chances to be either a satisfied or unsatisfied.

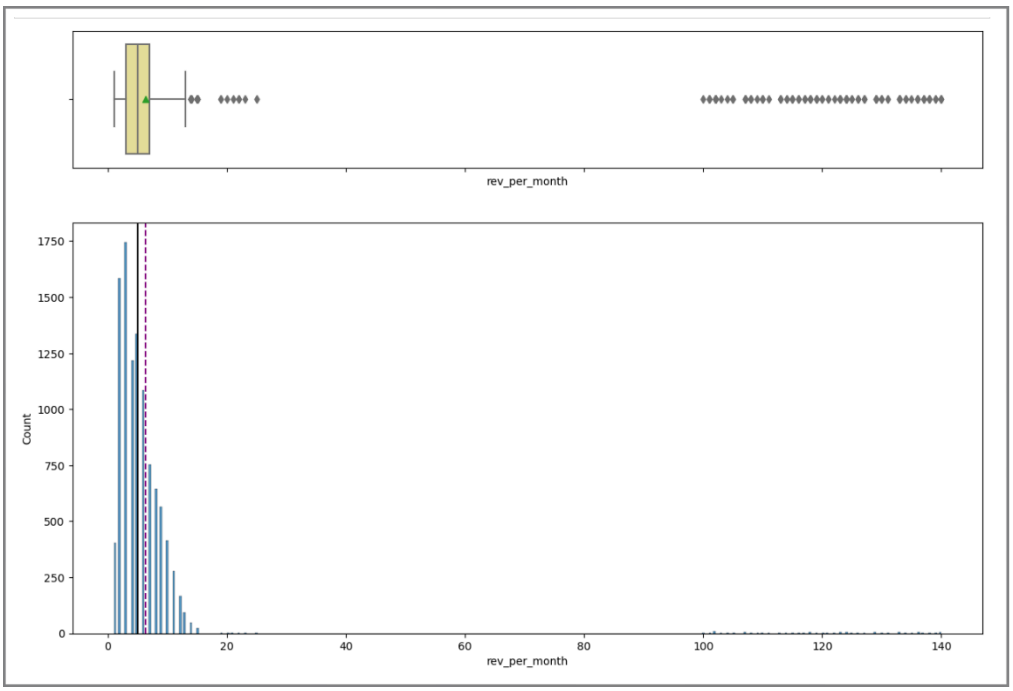
11. Marital_Status



Count of Marital_status

- 52% of the primary account holders are married.
- 31.3% of them are Single.
- 14.8% are divorced customers.

12. rev_per_month

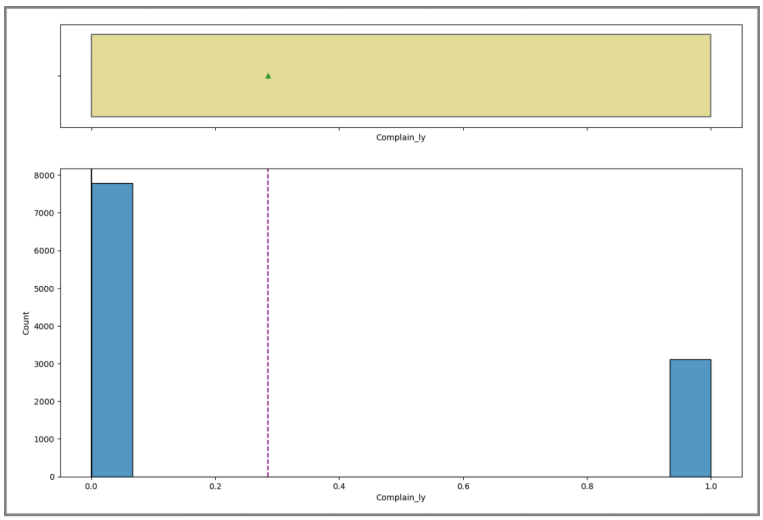


Boxplot and histogram for rev_per_month

- The currency is in thousands of INR.
- The average revenue generated by account per month is around 6362.
- Highly skewed due to the presence of outliers above 100K.

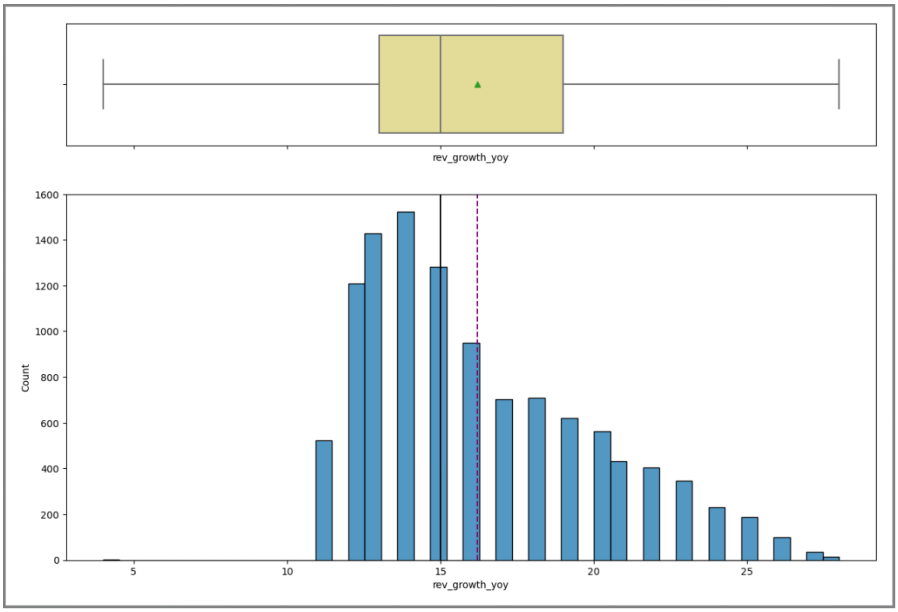
13. Complain_ly

- It is a binary data where 0 means a NO and 1 means a YES. So we will convert it into categorical.



Boxplot and histogram of Complain_ly

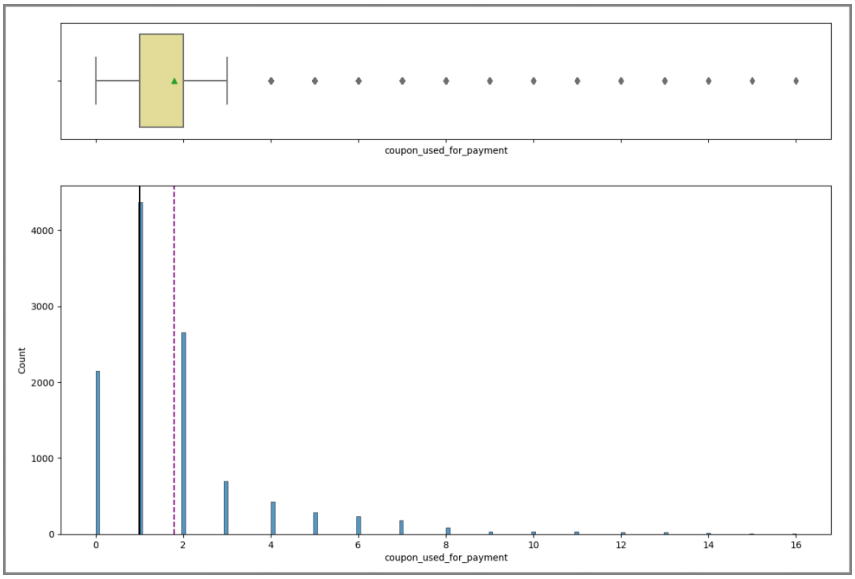
14. rev_growth_yoy



rev_growth_yoy

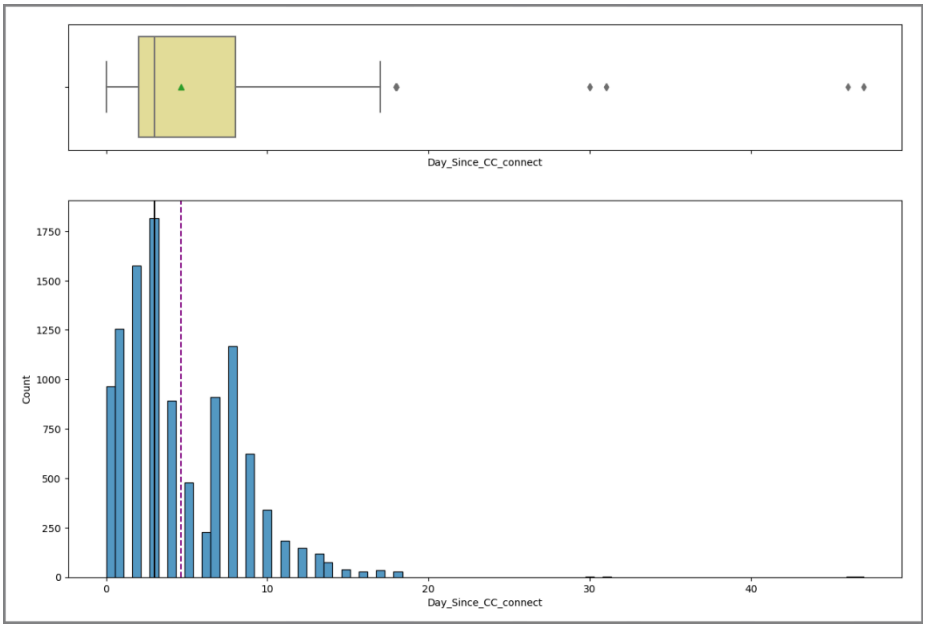
- On an average there is a 16% growth in revenue generated by the account in the past year compared to its previous year.
- The growth percentage ranges anywhere between 4% to 28%.

15. coupon_used_for_payment



- Average number of times coupon was used is 1.8.
- Outliers ranging from 4 to 16 times.

16. Day_Since_CC_connect

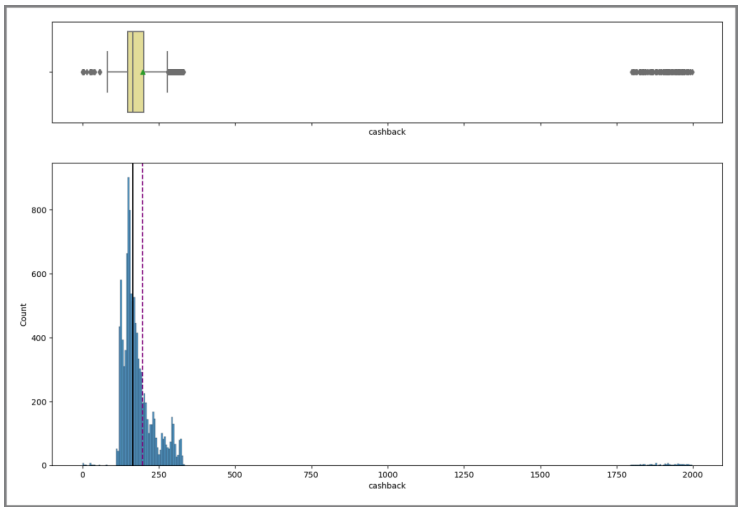


Boxplot and Histogram of Day_Since_CC_connect

- Highest number of customers reconnect within the first three days of reaching out. Customer service can be improved by tracing the recurring reasons and work towards solving it
- Average number of days taken to reconnect is around 5 days.
- The call pattern repeats again around 5 to 6 days
- Three different groups of outliers can be seen around 20, 30 and 50 days.

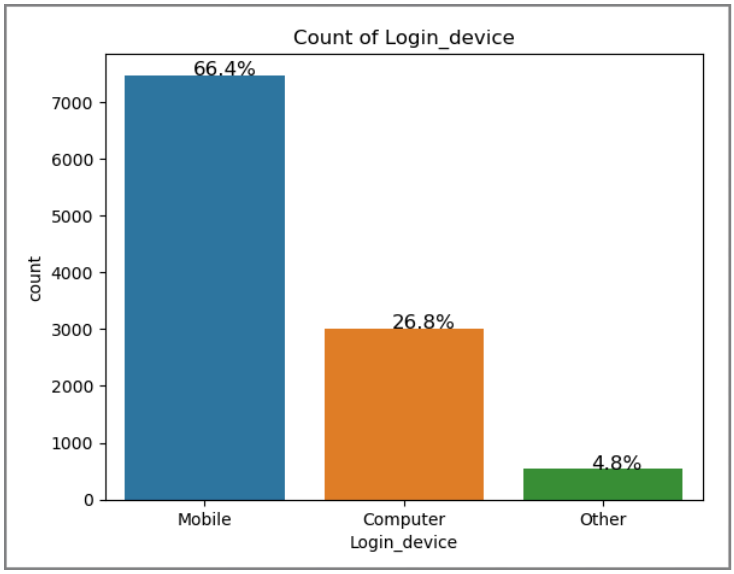
17. cashback

- On an average, the cashback generated by the account in the past year in near about 196 INR
- Outliers present between 1800 and 2000.
- Outliers are also present betwwen 0 and 100 and also between 300 to 400.



Boxplot and histogram for cashback

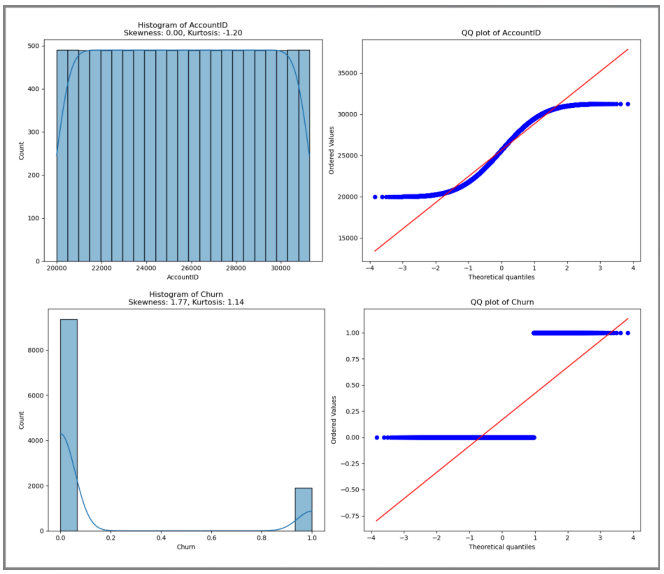
18. Login_device

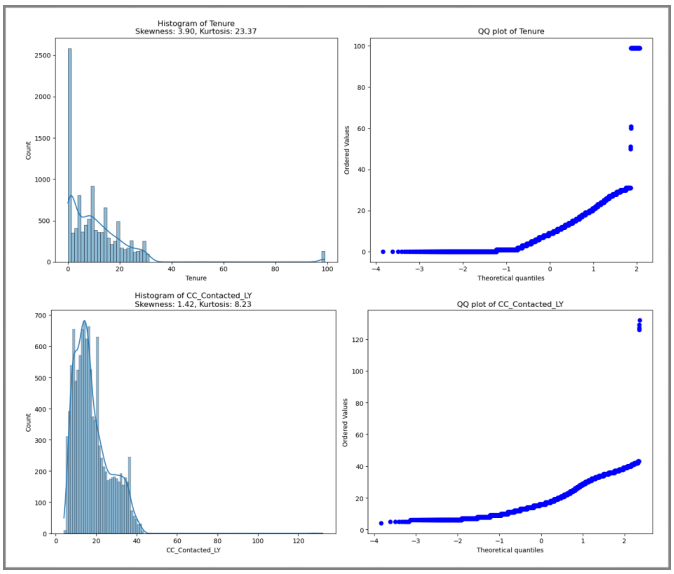


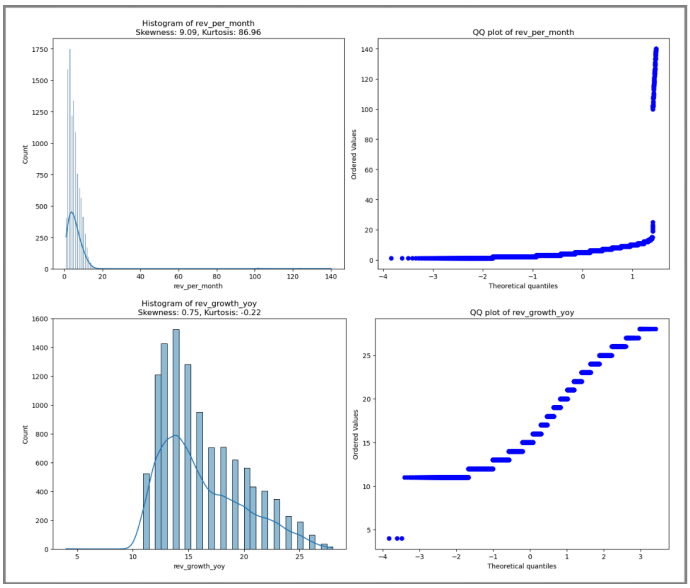
Count of Login_device

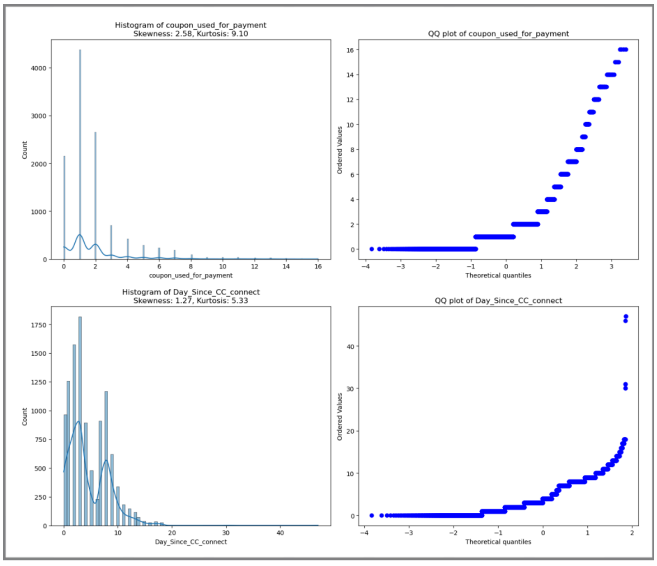
- Customers mostly use Mobile as there login device then followed by computer and Others

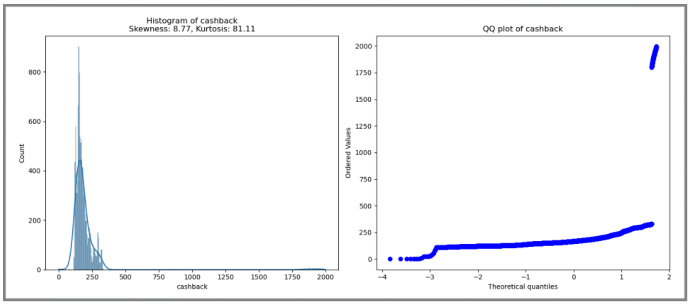
Skewness and kurtosis











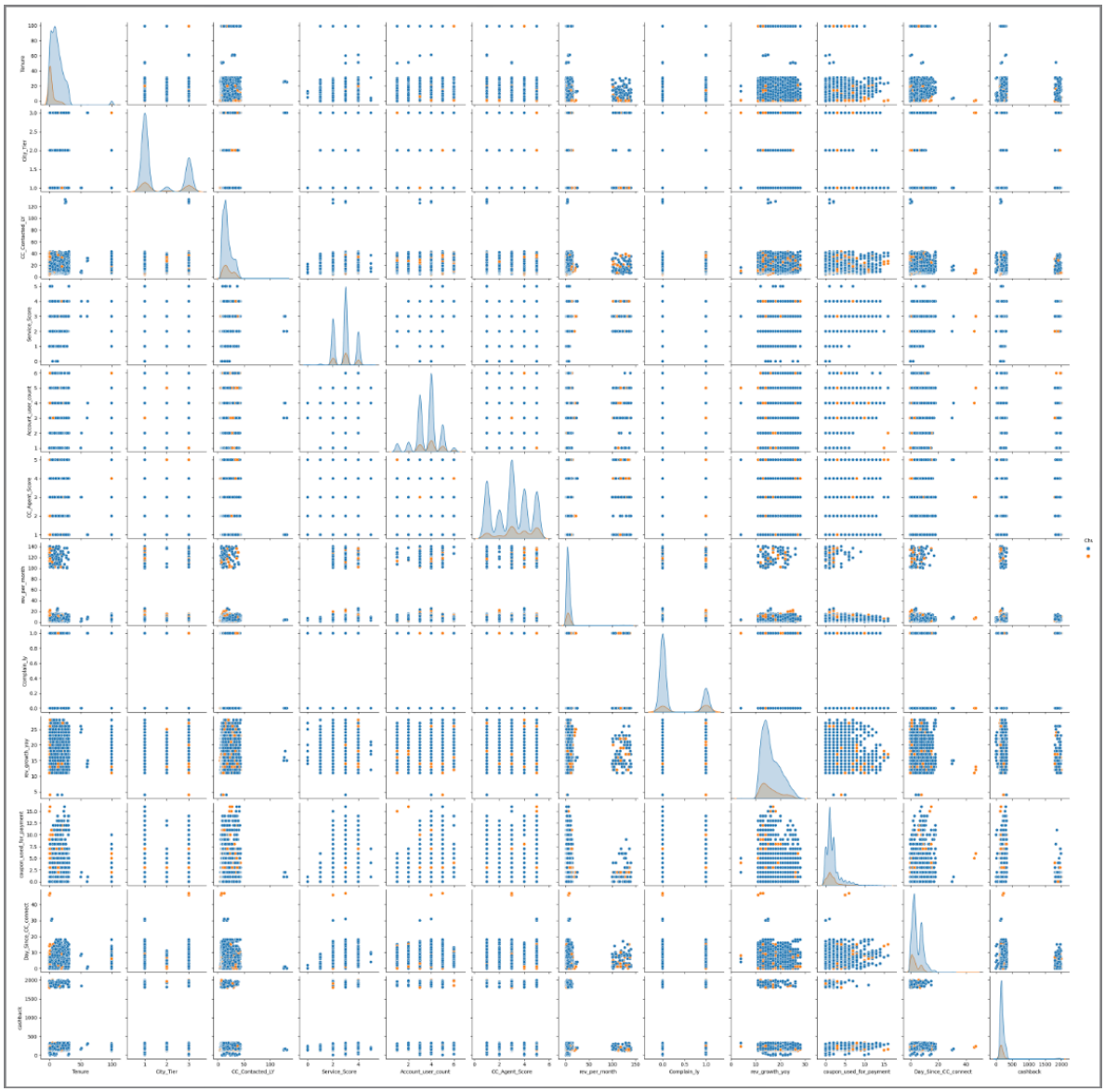
Skewness and Kurtosis of each numerical column:

	Column	Skewness	Kurtosis
0	AccountID	0.00	-1.20
1	Churn	1.77	1.14
2	Tenure	3.90	23.37
3	CC_Contacted_LY	1.42	8.23
4	rev_per_month	9.09	86.96
5	rev_growth_yoy	0.75	-0.22
6	coupon_used_for_payment	2.58	9.10
7	Day_Since_CC_connect	1.27	5.33
8	cashback	8.77	81.11

Observation of Skewness and kurtosis

- **AccountID**: Skewness and kurtosis values are close to zero, indicating a symmetrical distribution with lighter tails than a normal distribution.
- **Churn**: Positive skewness (1.77) indicates a right-skewed distribution, while a kurtosis value of 1.14 suggests a slightly peaked distribution.
- **Tenure**: High skewness (3.90) and kurtosis (23.37) indicate a highly right-skewed distribution with heavy tails, suggesting the presence of outliers.
- **CC_Contacted_LY**: Moderate right skewness (1.42) and a higher kurtosis (8.23) suggest a distribution with a long tail and a peaked shape.
- **rev_per_month**: Very high skewness (9.09) and kurtosis (86.96) indicate an extremely right-skewed distribution with very heavy tails, pointing to significant outliers.
- **rev_growth_yoy**: Slight right skewness (0.75) and near-zero kurtosis (-0.22) indicate a moderately skewed distribution with normal-like tails.
- **coupon_used_for_payment**: High skewness (2.58) and kurtosis (9.10) suggest a right-skewed distribution with heavy tails.
- **Day_Since_CC_connect**: Moderate skewness (1.27) and kurtosis (5.33) indicate a distribution that is right-skewed with some peakedness.
- **cashback**: High skewness (8.77) and kurtosis (81.11) reveal an extremely right-skewed distribution with very heavy tails, implying the presence of significant outliers.

c. Bivariate analysis



Pairplot

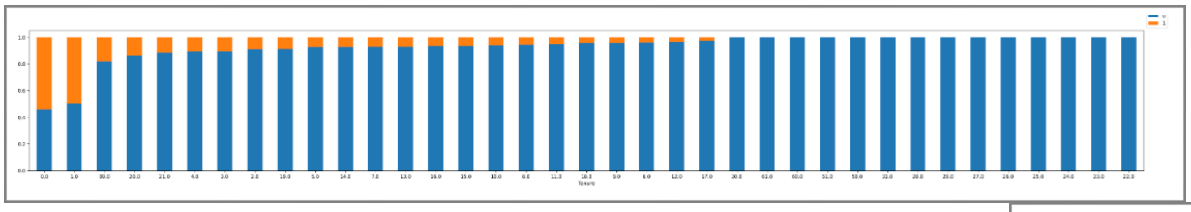
Observation from Pairplot

- Churn vs. other features shows some separation, with lower tenure and higher revenue per month correlating with churn.
- Features like rev_per_month and cashback have highly skewed distributions, with most data points clustered at lower values and some extreme outliers.

- Positive correlation observed between rev_per_month, cashback, and coupon_used_for_payment, indicating that higher spenders tend to receive more cashback and use more coupons.
- Tenure and Day_Since_CC_connect show a somewhat positive relationship, suggesting that customers with longer tenure may have recently connected with customer care.
- Several outliers are present in features like rev_per_month, cashback, and Tenure, contributing to data skewness and potentially affecting model accuracy.
- Categorical features appear as distinct clusters in the scatter plots, with some overlap indicating possible challenges in classification.

Stacked Bar Plot Inferences

i. Tenure vs Churn(Target variable)



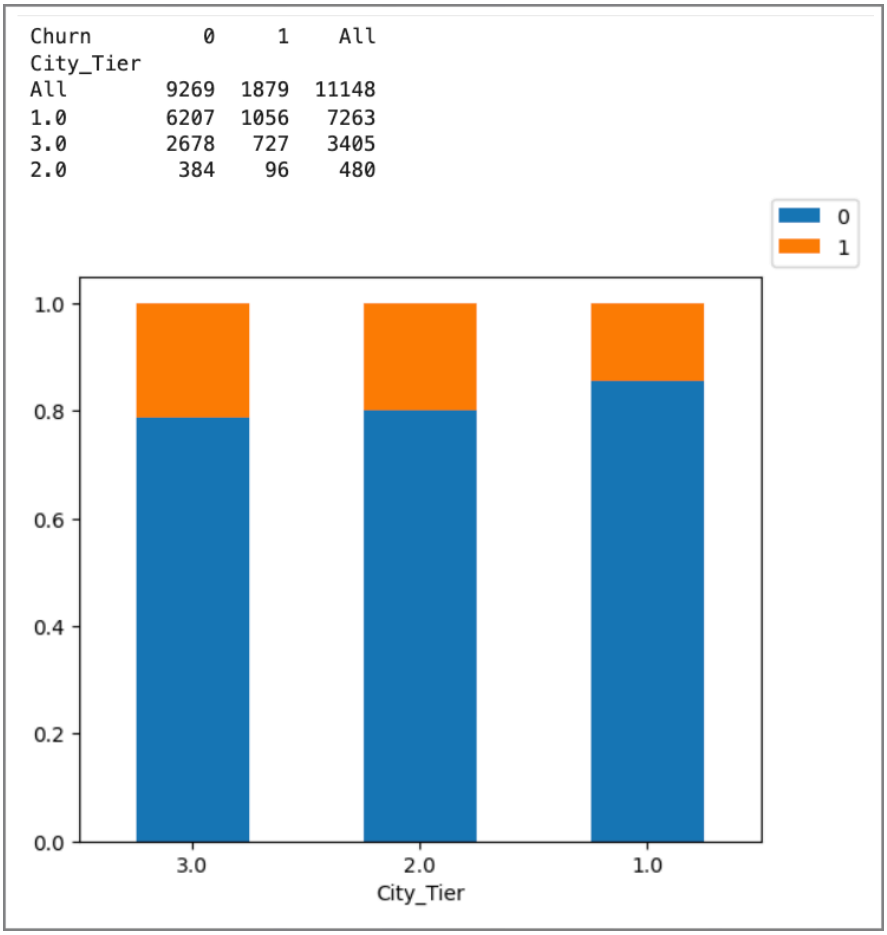
Stacked Barplot

Churn	0	1	All
Tenure			
All	9194	1848	11042
1.0	681	670	1351
0.0	562	669	1231
3.0	367	43	410
4.0	360	43	403
7.0	418	32	450
2.0	322	32	354
20.0	187	30	217
5.0	373	30	403
8.0	490	29	519
10.0	396	27	423
14.0	320	25	345
13.0	334	25	359
99.0	107	24	131
19.0	249	24	273
9.0	475	21	496
15.0	291	20	311
21.0	150	20	170
11.0	368	20	388
16.0	272	19	291
6.0	348	15	363
12.0	347	13	360
18.0	242	11	253
17.0	209	6	215
30.0	137	0	137
61.0	2	0	2
60.0	2	0	2
51.0	2	0	2
50.0	2	0	2
31.0	96	0	96
26.0	122	0	122
29.0	114	0	114
28.0	137	0	137
27.0	131	0	131
25.0	114	0	114
24.0	147	0	147
22.0	151	0	151
23.0	169	0	169

- Customers with very short tenures (e.g., 1.0 to 3.0 months) have notably high churn rates, indicating that new customers are more likely to leave. As tenure increases, the number of churned customers generally decreases, especially for tenures above 7 months, where the churn rate stabilizes and drops significantly. Interestingly, customers with tenures above 20 months show almost no churn, indicating that long-term customers are highly loyal.
- This trend suggests that early customer retention efforts could be crucial for reducing overall churn.

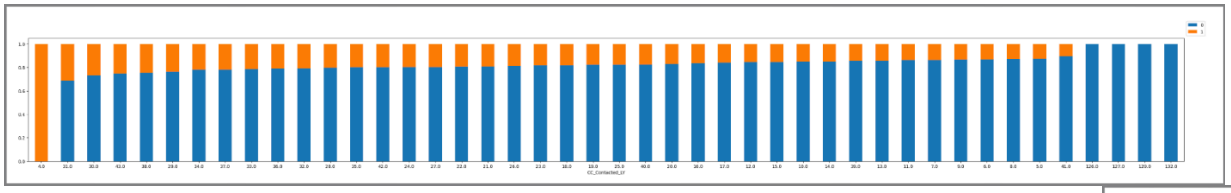
ii. City_tier vs Churn(Target variable)

Churn	0	1	All
City_Tier			
All	9269	1879	11148
1.0	6207	1056	7263
3.0	2678	727	3405
2.0	384	96	480



- Customers from **City Tier 1 show the highest retention**, with a majority (6207 out of 7263) not churning, suggesting a more stable customer base in this tier. In contrast, City Tier 3 has a higher churn rate, with 727 out of 3405 customers leaving, indicating potential challenges in customer retention. City Tier 2 has the smallest customer base, but it also has a noticeable churn rate (96 out of 480), which suggests that customers in this tier may also require targeted retention strategies.
- Overall, City Tier 1 has the most loyal customers, while Tier 3 presents the most churn risk.

iii. CC_Contacted_LY vs Churn(Target variable)

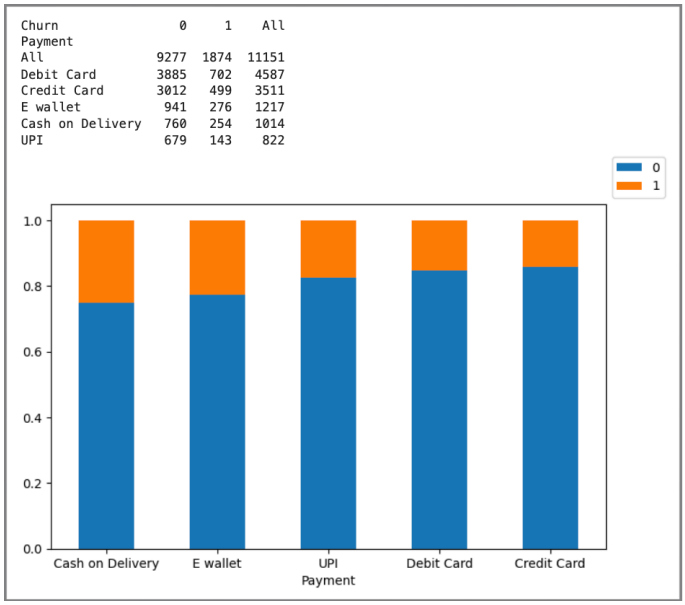


Churn	0	1	All
CC_Contacted_LY			
All	9280	1878	11158
16.0	554	109	663
14.0	580	102	682
15.0	528	95	623
13.0	561	94	655
9.0	567	88	655
12.0	483	88	571
17.0	442	83	525
10.0	415	74	489
11.0	452	72	524
8.0	469	69	538
18.0	306	68	374
19.0	299	65	364
21.0	251	59	310
22.0	227	55	282
20.0	265	54	319
7.0	338	53	391
31.0	113	52	165
30.0	128	47	175
23.0	197	44	241
29.0	138	43	181
24.0	172	42	214
6.0	270	41	311
32.0	152	40	192
34.0	139	39	178
28.0	142	36	178
25.0	162	35	197
27.0	140	34	174
35.0	132	33	165
33.0	122	33	155
26.0	137	32	169
36.0	117	31	148
37.0	75	21	96
38.0	55	18	73
39.0	47	8	55
40.0	38	8	46
42.0	24	6	30
41.0	26	3	29
43.0	6	2	8
5.0	7	1	8
4.0	0	1	1
126.0	1	0	1
127.0	1	0	1
129.0	1	0	1
132.0	1	0	1

- A pattern in churn rates relative to the number of times a customer contacted customer care in the last year (CC_Contacted_LY). Generally, a higher number of customer care contacts is associated with slightly higher churn rates. Customers with 4 or 5 contacts show minimal churn. The data suggests that a certain level of engagement with customer care may indicate dissatisfaction, potentially leading to churn, but extremely frequent or rare contacts might represent either severe issues or very satisfied customers, respectively.

iv. Payment vs Churn(Target variable)

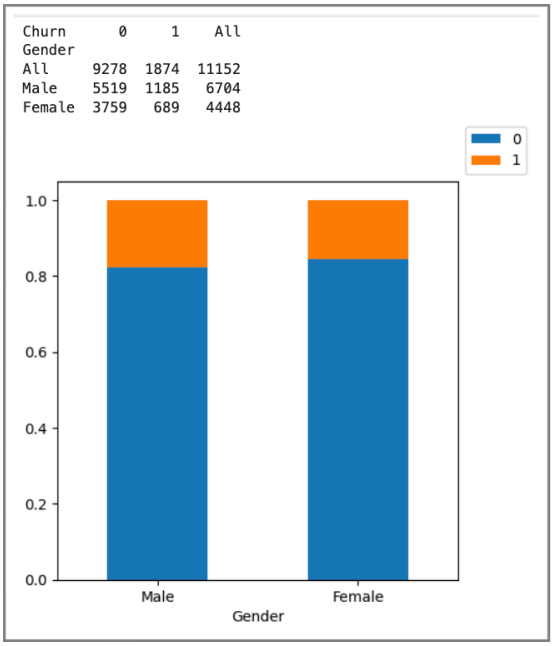
Churn	0	1	All
Payment			
All	9277	1874	11151
Debit Card	3885	702	4587
Credit Card	3012	499	3511
E wallet	941	276	1217
Cash on Delivery	760	254	1014
UPI	679	143	822



Among the 11,151 customers, those using **Debit Cards** represent the largest group, with a churn rate of approximately 15% (702 out of 4,587). **Credit Card**

users follow, with a churn rate of about 14% (499 out of 3,511). **E-wallet** users have a slightly higher churn rate of around 23% (276 out of 1,217), indicating that customers using digital wallets may be more prone to churn. **Cash on Delivery** has an even higher churn rate of 25% (254 out of 1,014). Lastly, **UPI** users exhibit a churn rate of roughly 17% (143 out of 822). These trends suggest that customers using alternative payment methods like E-wallets and Cash on Delivery may be more likely to churn compared to those using more traditional payment methods like Debit or Credit Cards.

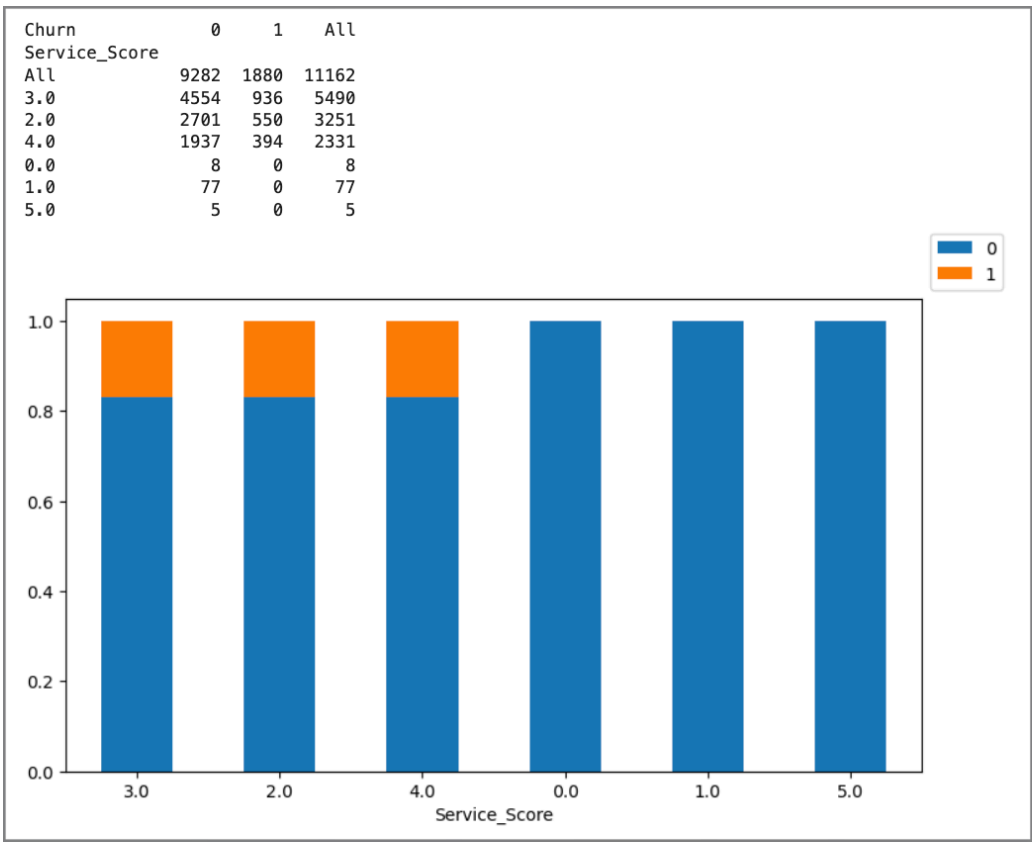
v. Gender vs Churn(Target variable)



- The churn rate for Male customers is approximately 18% (1,185 out of 6,704), whereas the churn rate for Female customers is slightly lower, at about 15% (689 out of 4,448). This indicates that, overall, male customers tend to churn at a higher rate compared to female customers. Despite this difference, the churn rates are relatively close, suggesting that both genders experience a similar propensity to churn, with males being marginally more likely to do so.

vi. Service_score vs Churn(Target variable)

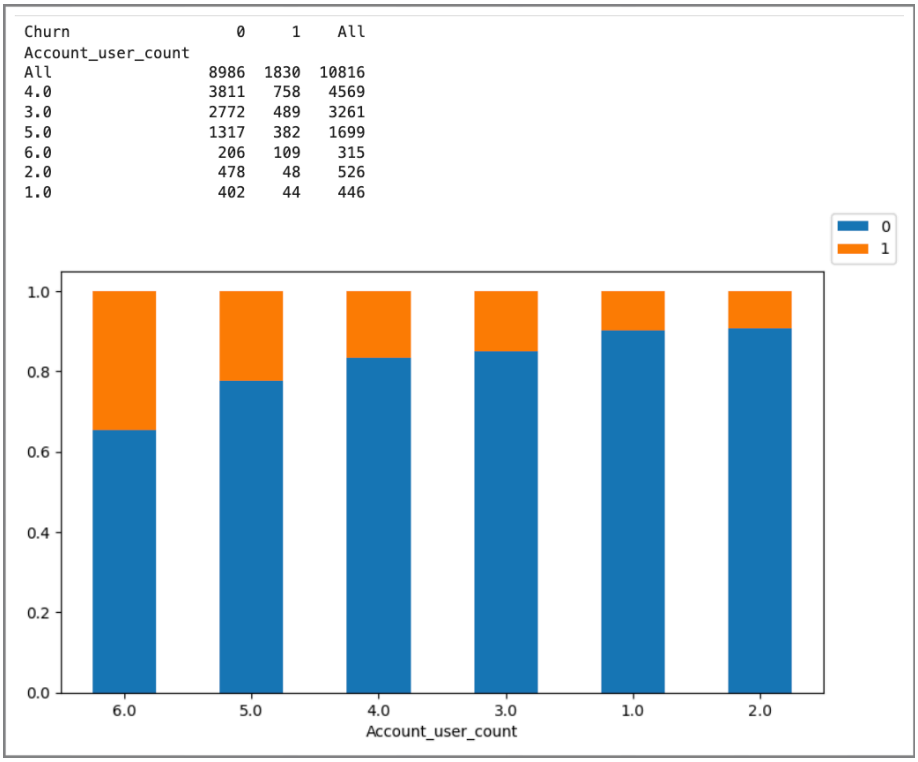
Service_Score	Churn 0	1	All
All	9282	1880	11162
3.0	4554	936	5490
2.0	2701	550	3251
4.0	1937	394	2331
0.0	8	0	8
1.0	77	0	77
5.0	5	0	5



The largest group is those with a **Service Score of 3.0**, comprising 5,490 customers, with a churn rate of 936 out of 5,490). The next significant group is the **Service Score of 2.0** with 3,251 customers, also exhibiting a churn rate around 17% (550 out of 3,251). Customers with a Service Score of 4.0 total 2,331, with a similar churn rate of about 17% (394 out of 2,331). For Service Scores of 0.0, 1.0, and 5.0, there are very few customers and no recorded churn, which might indicate limited data or potential anomalies for these scores. Overall, churn rates appear consistent across the main Service Score groups, suggesting that churn is relatively uniform within these scores.

vii.Account_user_count vs Churn(Target variable)

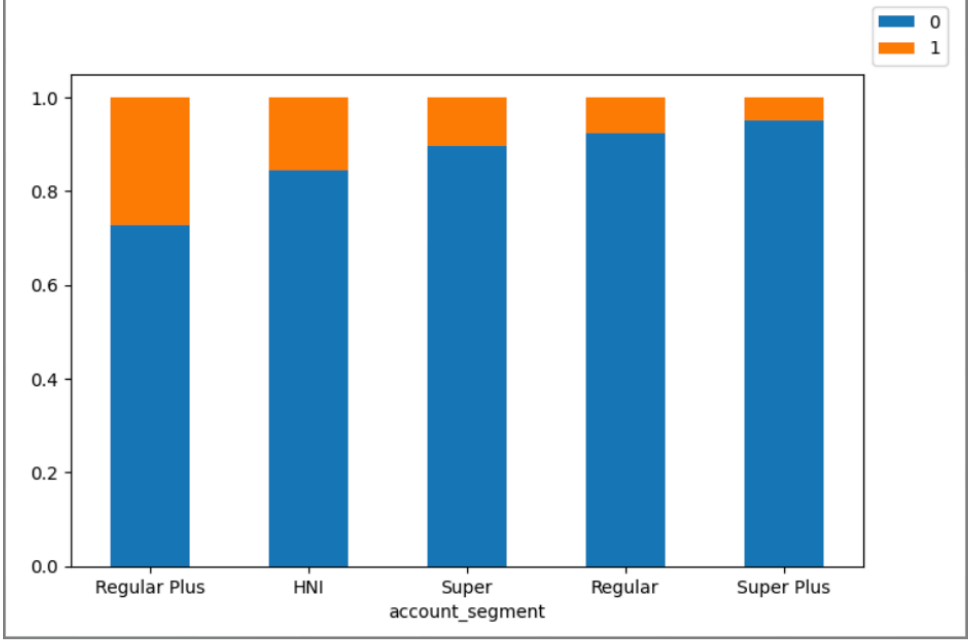
Churn	0	1	All
Account_user_count			
All	8986	1830	10816
4.0	3811	758	4569
3.0	2772	489	3261
5.0	1317	382	1699
6.0	206	109	315
2.0	478	48	526
1.0	402	44	446



The largest group, with an Account User Count of 4.0, includes 4,569 customers, of whom 758 (approximately 17%) churned. The next significant group is those with an Account User Count of 3.0, totalling 3,261 customers, with a churn rate of around 15% (489 out of 3,261). Customers with an Account User Count of 5.0 comprise 1,699 individuals, with a churn rate of about 22% (382 out of 1,699). For Account User Counts of 6.0, 2.0, and 1.0, the numbers are smaller, with churn rates varying significantly. Specifically, the churn rate for those with 6.0 accounts is about 35% (109 out of 315), which is notably higher compared to other groups. Customers with 2.0 and 1.0 accounts have relatively low churn rates, 9% (48 out of 526) and 10% (44 out of 446) respectively. Overall, the churn rate tends to increase with higher account user counts, with the highest churn rate observed among those with 6.0 accounts.

viii.account_segment vs Churn(Target variable)

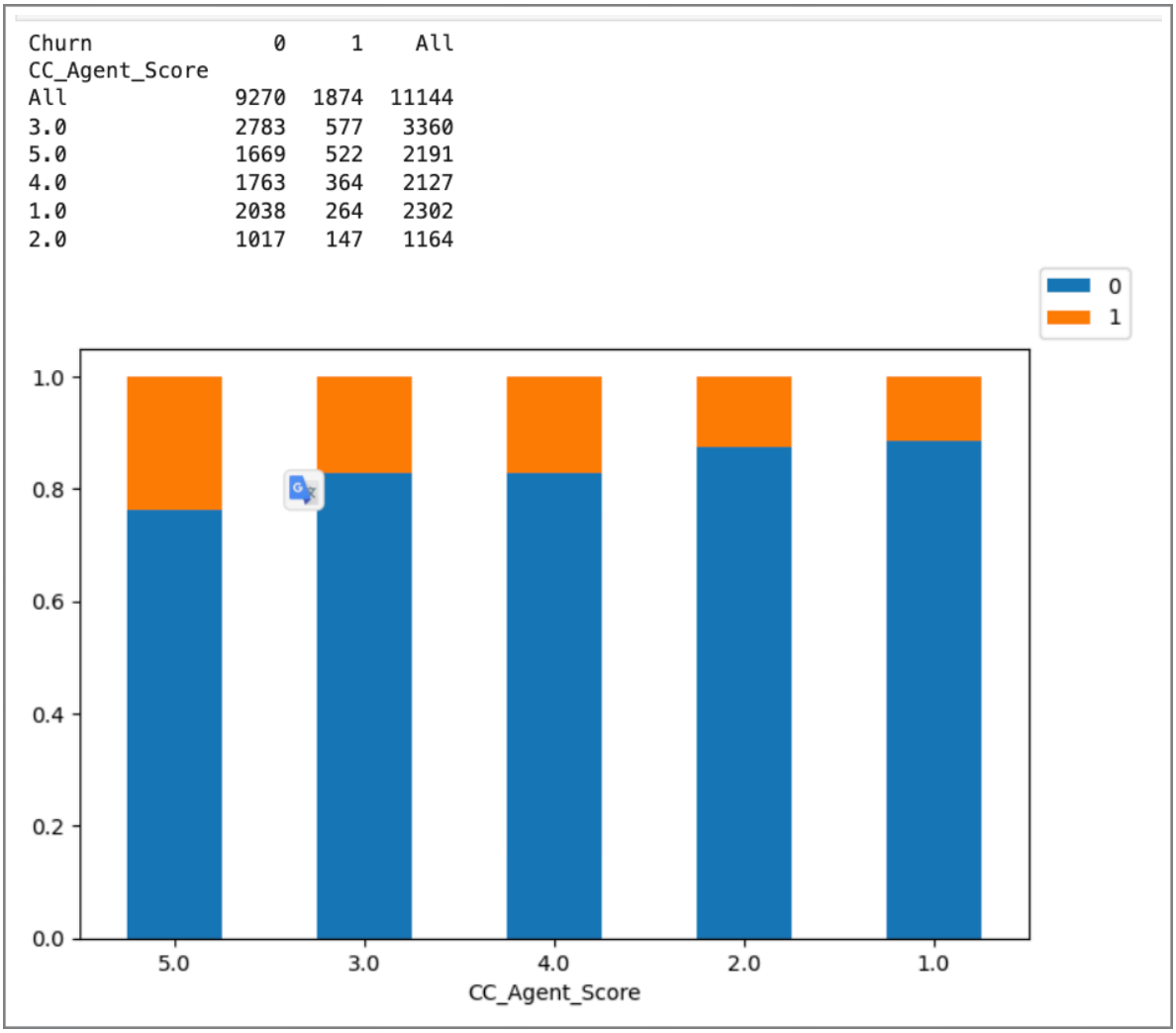
Churn	0	1	All
account_segment			
All	9285	1878	11163
Regular Plus	2997	1127	4124
Super	3646	416	4062
HNI	1384	255	1639
Regular	480	40	520
Super Plus	778	40	818



The data on churn across different account segments shows that the Regular Plus segment, with 4,124 customers, has the highest churn count at 1,127, translating to roughly 27% of its customer base. The Super segment, with 4,062 customers, experiences a lower churn rate of about 10%, with 416 customers having churned. HNI accounts, totalling 1,639 customers, have a churn rate of approximately 16% (255 churns), while the Regular segment, with only 520 customers, shows a churn rate of about 8% (40 churns). The Super Plus segment has the lowest churn rate, with 40 out of 818 customers (around 5%) having churned. Overall, the Regular Plus segment is the most affected by churn, while the Super Plus segment has the lowest churn rate.

ix. CC_Agent_Score vs Churn(Target variable)

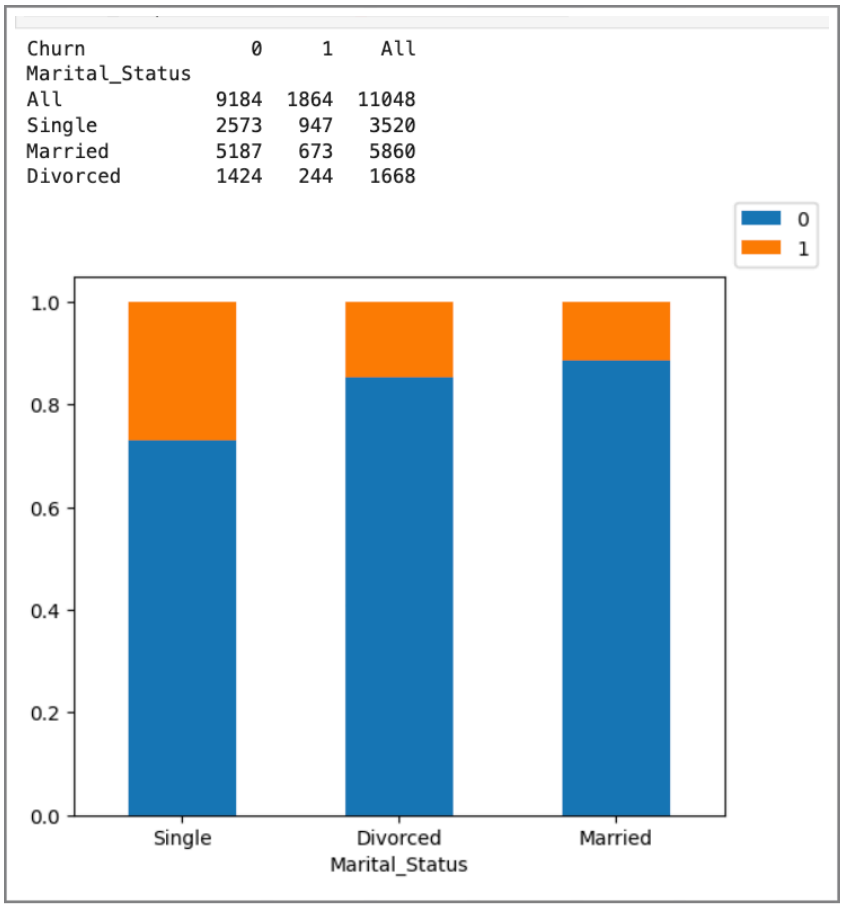
Churn	0	1	All
CC_Agent_Score			
All	9270	1874	11144
3.0	2783	577	3360
5.0	1669	522	2191
4.0	1763	364	2127
1.0	2038	264	2302
2.0	1017	147	1164



The churn data across different CC_Agent_Score categories reveals varying levels of customer retention. The 3.0 score category has the largest number of customers, totalling 3,360, with a churn count of 577, indicating a churn rate of approximately 17%. The 5.0 score category, with 2,191 customers, shows a churn of 522, resulting in a churn rate of about 24%. Customers in the 4.0 score category, totalling 2,127, have a churn rate of around 17% with 364 churns. The 1.0 score category has a high churn rate of 11%, with 264 out of 2,302 customers churning. Lastly, the 2.0 score category, with 1,164 customers, has the lowest churn rate of around 13%, with 147 churns. Overall, higher CC_Agent_Score categories tend to correlate with higher churn rates.

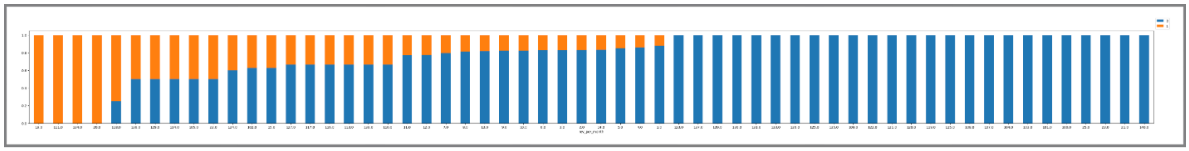
x. Marital_Status vs Churn(Target variable)

Churn	0	1	All
Marital_Status			
All	9184	1864	11048
Single	2573	947	3520
Married	5187	673	5860
Divorced	1424	244	1668



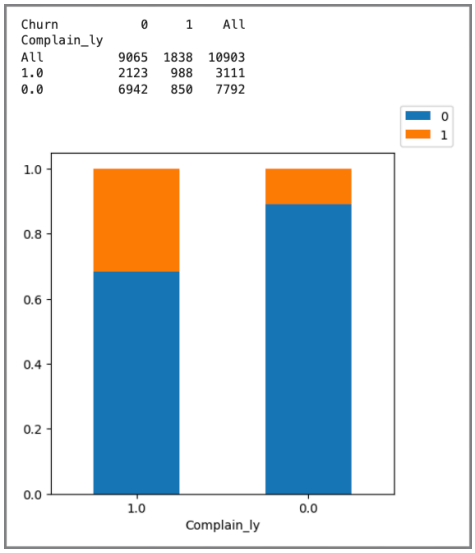
Among single customers, there is a relatively high churn rate, with 947 out of 3,520 customers churning, resulting in approximately 27% churn. Married customers have a significantly lower churn rate, with 673 out of 5,860 customers churning, equating to around 11.5%. Divorced customers fall in between, with 244 out of 1,668 customers churning, leading to a churn rate of about 14.6%. Overall, single customers exhibit the highest churn rate, while married customers are the most stable group.

xi. rev_per_month vs Churn(Target variable)



The churn data segmented by revenue per month in INR shows that the majority of customers are concentrated in the lower revenue brackets, specifically between ₹1,000 and ₹10,000 per month. In these segments, the churn rate is moderate, with the highest churn observed in the ₹3,000, ₹2,000, and ₹5,000 per month categories. For example, in the ₹3,000 per month segment, 299 out of 1,746 customers have churned, representing approximately 17.1% churn. Similarly, in the ₹2,000 per month group, 270 out of 1,585 customers have churned, accounting for about 17.0% churn. As the monthly revenue increases beyond ₹10,000, the number of customers drops significantly, and churn becomes less frequent. Outlier revenue values, such as ₹118,000, ₹102,000, ₹140,000, and others, have very few customers, with minimal churn observed. Overall, lower revenue segments exhibit higher churn rates, while higher revenue segments have lower churn and fewer customers.

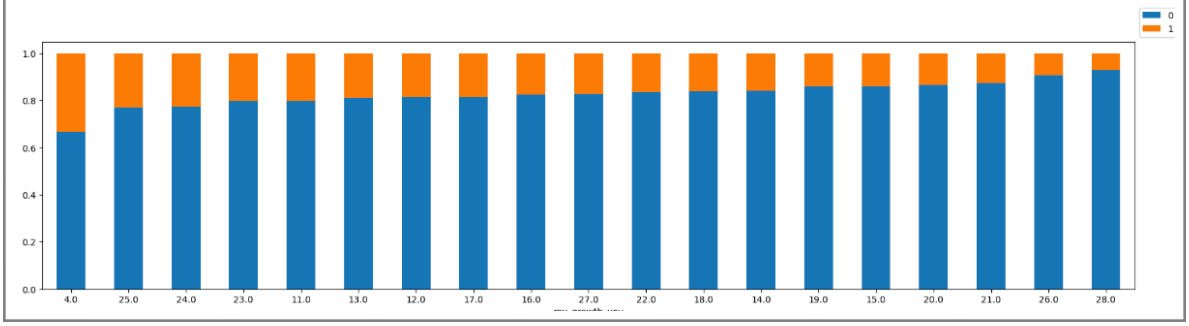
xii. Complain_ly vs Churn(Target variable)



The churn data segmented by the number of complaints lodged last year indicates that customers who lodged at least one complaint have a significantly higher churn rate compared to those who did not. Specifically, among the 3,111 customers who lodged complaints, 988 (approximately 31.8%) have churned. In contrast, only 850 out of 7,792 customers who did not lodge any complaints have churned, representing a much lower churn rate of about 10.9%. This suggests a strong correlation between customer complaints and churn, with dissatisfied customers being more likely to leave.

xiii. rev_growth_yoy vs Churn(Target variable)

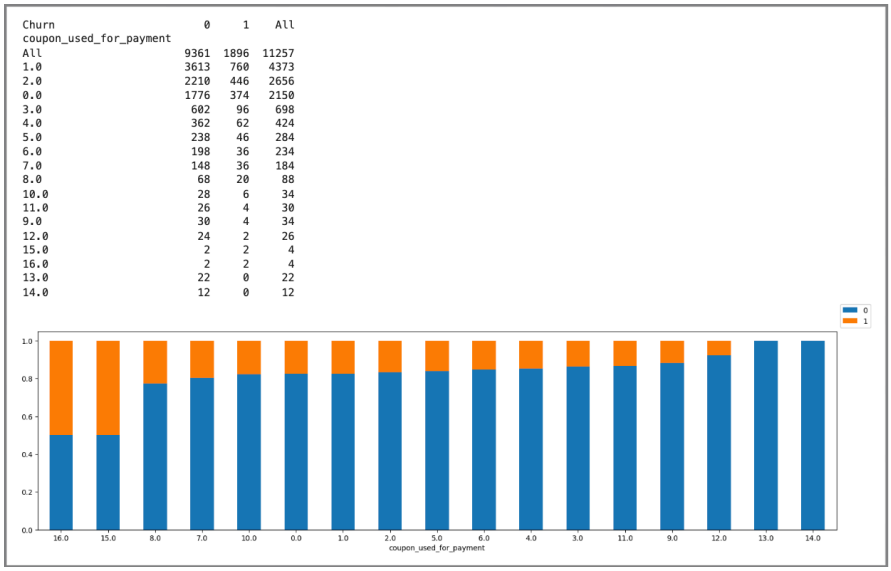
Churn	0	1	All
rev_growth_yoy			
All	9361	1896	11257
13.0	1157	270	1427
14.0	1283	241	1524
12.0	985	225	1210
15.0	1105	178	1283
16.0	782	167	949
17.0	574	130	704
18.0	593	115	708
11.0	418	105	523
19.0	533	86	619
20.0	486	76	562
23.0	275	70	345
22.0	337	66	403
21.0	378	55	433
24.0	177	52	229
25.0	145	43	188
26.0	89	9	98
27.0	29	6	35
28.0	13	1	14
4.0	2	1	3



The churn data segmented by year-over-year revenue growth indicates that customers with higher revenue growth percentages tend to have lower churn rates. Specifically, customers in the revenue growth bands of 13% to 15% have higher churn numbers (270, 241, and 225 respectively) but also represent a larger customer base. As the growth percentage increases to 16% and beyond, both the churn and customer base gradually decrease. Interestingly, customers with very low or very high revenue growth percentages, such as 4% or 28%, show negligible churn, possibly due to a smaller customer base in these bands. This suggests that moderate revenue growth is more common and may correlate with higher customer turnover.

xiv. coupon_used_for_payment vs Churn(Target variable)

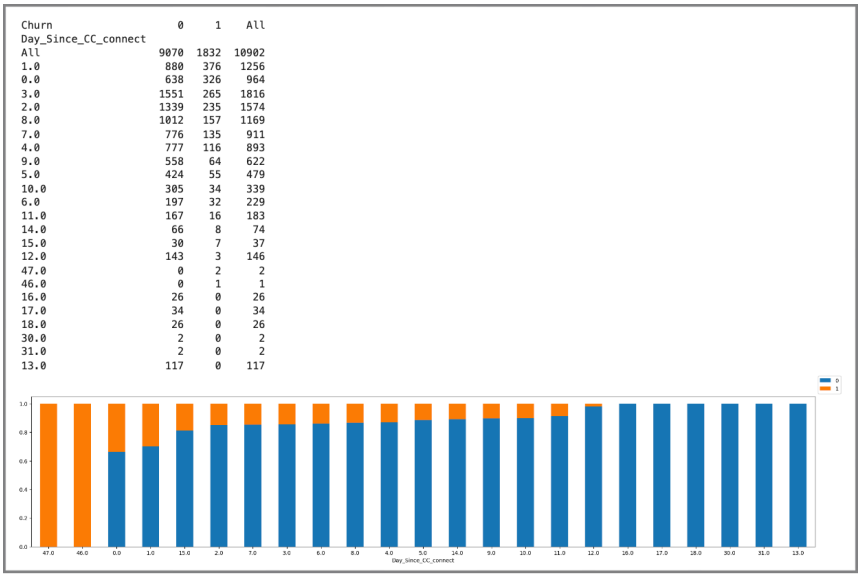
Churn	0	1	All
coupon_used_for_payment			
All	9361	1896	11257
1.0	3613	760	4373
2.0	2210	446	2656
0.0	1776	374	2150
3.0	602	96	698
4.0	362	62	424
5.0	238	46	284
6.0	198	36	234
7.0	148	36	184
8.0	68	20	88
10.0	28	6	34
11.0	26	4	30
9.0	30	4	34
12.0	24	2	26
15.0	2	2	4
16.0	2	2	4
13.0	22	0	22
14.0	12	0	12



Customers who used 1 to 2 coupons for payment have the highest representation in both the churned (760 and 446 respectively) and non-churned (3613 and 2210 respectively) categories. As the number of coupons used increases beyond 2, both the customer base and churn rates significantly decrease, with very few customers using more than 8 coupons, and these customers show minimal churn. Only two customers have used coupons the highest number of times (15 or 16) for payment, and both have churned.

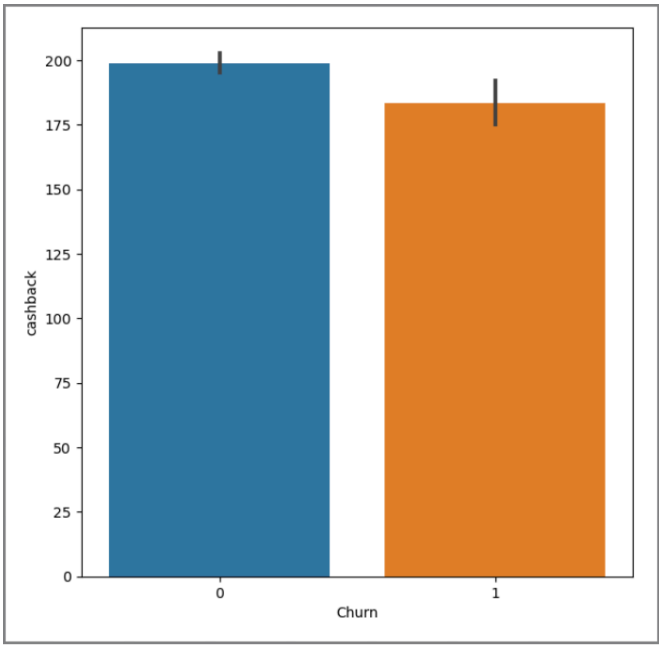
xv. Day_Since_CC_connect vs Churn(Target variable)

Churn	0	1	All
Day_Since_CC_connect			
All	9070	1832	10902
1.0	880	376	1256
0.0	638	326	964
3.0	1551	265	1816
2.0	1339	235	1574
8.0	1012	157	1169
7.0	776	135	911
4.0	777	116	893
9.0	558	64	622
5.0	424	55	479
10.0	305	34	339
6.0	197	32	229
11.0	167	16	183
14.0	66	8	74
15.0	30	7	37
12.0	143	3	146
47.0	0	2	2
46.0	0	1	1
16.0	26	0	26
17.0	34	0	34
18.0	26	0	26
30.0	2	0	2
31.0	2	0	2
13.0	117	0	117



As the days since the last call center contact increase, churn rates generally decrease. Interestingly, there are rare instances of churn even among customers who have not contacted the call center for extended periods (e.g., 47 or 46 days). This suggests that while recent interaction with the call center may be associated with higher churn, prolonged periods without contact do not necessarily guarantee retention.

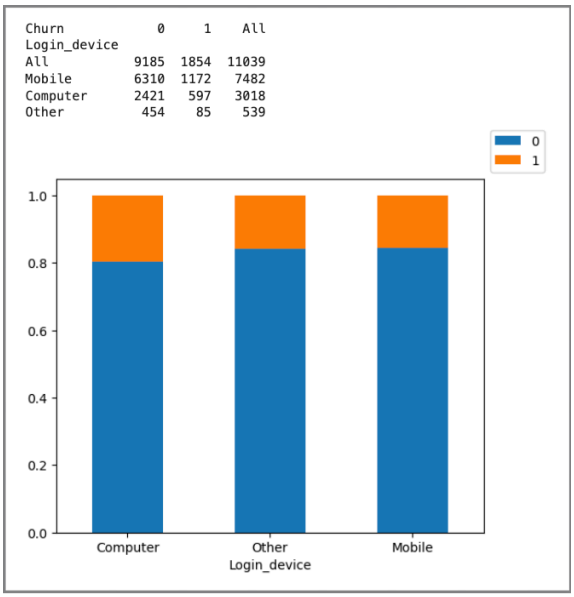
xvi. cashback vs Churn(Target variable)



- There is higher chance of receiving cashback.

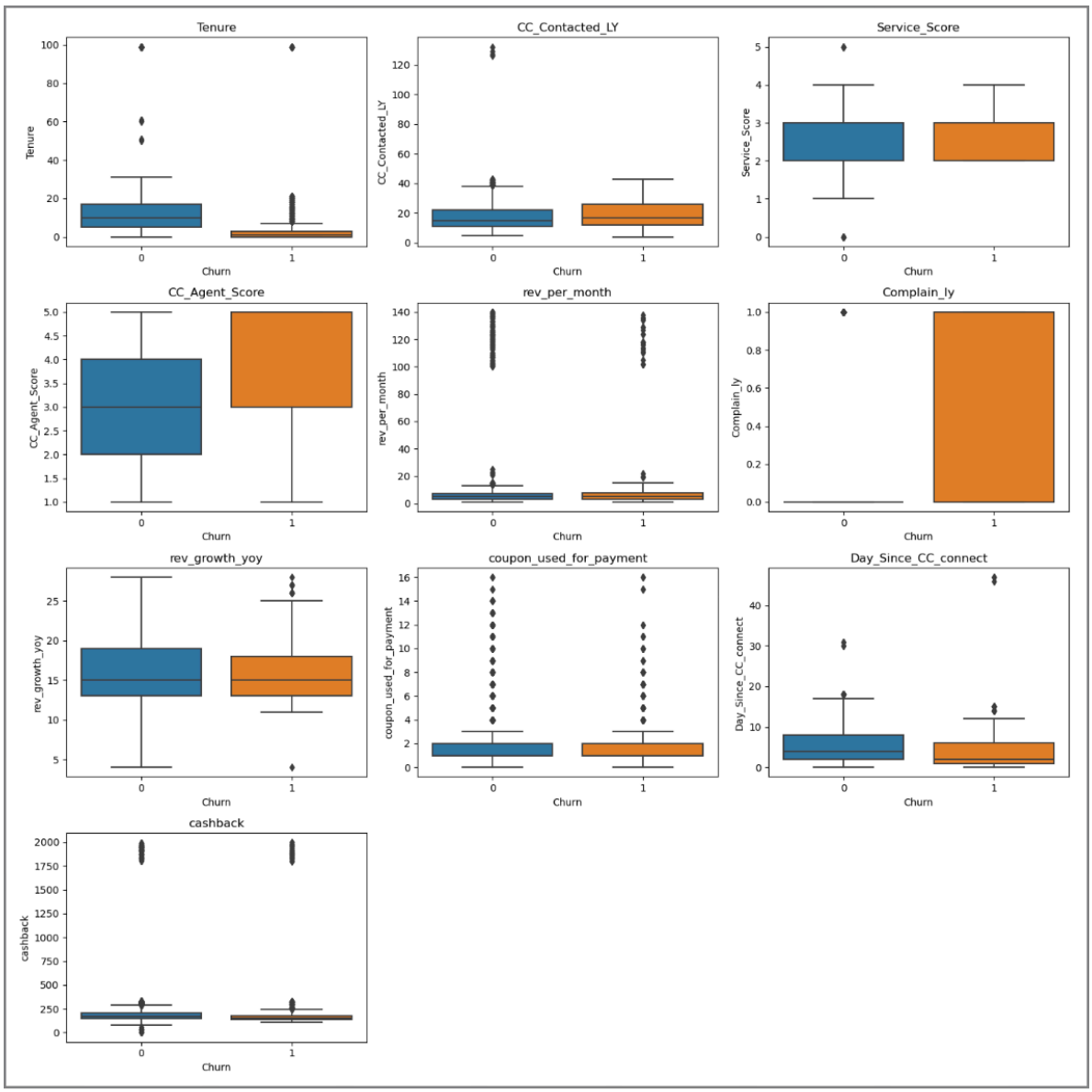
xvii. Login_device vs Churn(Target variable)

The data indicates that a majority of customers who churned were primarily using mobile devices for login, with 1,172 out of 1,854 churned users accessing services via mobile. In contrast, users logging in via computers also show a highest significant churn rate, with 597 out of 3,018 users. The "Other" category, which



includes less common devices, has the lowest churn numbers. This suggests that while mobile users make up the largest segment of the customer base, they also represent a substantial portion of those who leave, pointing to a potential area for targeted retention efforts.

Boxplot with Churn



Observations

1. **Tenure**: Customers with lower tenure tend to churn more frequently, indicating that newer customers are more prone to leaving.
2. **City_Tier**: Churn rates are higher among customers in City Tiers 1 and 3, while Tier 2 cities show significantly lower churn, suggesting variations in service experience or customer expectations across different tiers.
3. **CC_Contacted_LY**: Customers who have had more frequent interactions with customer care in the last year are more likely to churn, hinting that frequent issues or complaints correlate with higher churn rates.
4. **Payment**: Higher churn rates are observed among customers using E-wallets and Cash on Delivery, compared to those using Debit or Credit Cards, possibly reflecting differences in payment preferences or related issues.
5. **Gender**: Males exhibit a higher churn rate than females, pointing to gender-based differences in customer retention.
6. **Service_Score**: Lower service scores are strongly associated with customers who churn, underscoring the importance of service quality in retaining customers.
7. **Account_user_count**: Customers with fewer accounts, such as one or two, show higher churn rates, suggesting that those with more accounts are more engaged and less likely to leave.
8. **Account_segment**: Churn is more prevalent in the 'Regular' and 'Super' account segments, whereas premium segments like 'HNI' and 'Super Plus' have lower churn rates, indicating better retention of high-value customers.
9. **CC_Agent_Score**: Lower scores from customer care agents are linked to higher churn rates, highlighting the influence of agent performance on customer retention.
10. **Marital_Status**: Single customers have a higher churn rate compared to married or divorced customers, suggesting that marital status may affect engagement or satisfaction levels.

11. **rev_per_month**: Customers with fluctuating monthly revenue show higher churn rates, implying that consistent spending patterns may be linked to better retention.
12. **Complain_ly**: A higher number of complaints correlates with increased churn, reinforcing the idea that frequent issues are a predictor of customer attrition.
13. **rev_growth_yoy**: Lower year-over-year revenue growth is associated with a higher likelihood of churn, indicating that customers experiencing stagnant or declining revenue are more likely to leave.
14. **coupon_used_for_payment**: Frequent use of coupons for payment is linked to higher churn rates, suggesting a potential connection between reliance on discounts and lower loyalty.
15. **Day_Since_CC_connect**: Churned customers tend to have more days since their last contact with customer care, possibly indicating disengagement or unresolved issues.
16. **Cashback**: Churned customers show greater variability in cashback amounts, which may reflect differing levels of satisfaction or perceived value.
17. **Login_device**: Customers who use mobile devices to log in have higher churn rates compared to those using computers, potentially indicating differences in user experience or engagement between devices.

d. Removal of unwanted variables

Removing the `**AccountID**` column from the dataset is appropriate since it contains unique values for each record and does not contribute to meaningful analysis or prediction. Unique identifiers like `**AccountID**` are primarily used to distinguish individual records but do not provide insights into patterns or relationships within the data. By excluding this column, we can streamline the dataset, reduce computational complexity, and focus on variables that have actual predictive value or informative content.

Now, we have 18 columns in which 1 is target variable is 'Churn' and 17 independent.

```
Index(['Churn', 'Tenure', 'City_Tier', 'CC_Contacted_LY', 'Payment', 'Gender',
       'Service_Score', 'Account_user_count', 'account_segment',
       'CC_Agent_Score', 'Marital_Status', 'rev_per_month', 'Complain_ly',
       'rev_growth_yoy', 'coupon_used_for_payment', 'Day_Since_CC_connect',
       'cashback', 'Login_device'],
      dtype='object')
```

Columns of the dataset after removal of AccountID

```
(11260, 18)
```

Shape of dataset

e. Missing Value treatment

```
AccountID                0.00
Churn                    0.00
Tenure                   1.94
City_Tier                0.99
CC_Contacted_LY          0.91
Payment                  0.97
Gender                   0.96
Service_Score            0.87
Account_user_count       3.94
account_segment          0.86
CC_Agent_Score           1.03
Marital_Status           1.88
rev_per_month            7.02
Complain_ly              3.17
rev_growth_yoy           0.03
coupon_used_for_payment  0.03
Day_Since_CC_connect     3.18
cashback                 4.20
Login_device             1.96
dtype: float64
```

Missing value percentage

To handle missing values in the dataset, we will apply the KNN imputer with `n_neighbors=5`. Before performing KNN imputation, we will encode the categorical variables into numerical values to ensure the imputer can effectively calculate distances between data points. This approach allows us to estimate and fill in the missing values based on the similarity of data points within the dataset.

This code maps the categorical variables in your dataset to numerical values, which is a necessary step before applying KNN imputation. By converting categories to numerical representations, you enable the KNN algorithm to compute distances more effectively between data points, which helps in accurately imputing missing values.

Encoded Dataset:

	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_St
0	1	4.0	3.0	6.0	0.0	0.0	3.0	3.0	0.0	2.0	
1	1	0.0	1.0	8.0	1.0	1.0	3.0	4.0	1.0	3.0	
2	1	0.0	1.0	30.0	0.0	1.0	2.0	4.0	1.0	3.0	
3	1	0.0	3.0	15.0	0.0	1.0	2.0	4.0	0.0	5.0	
4	1	0.0	1.0	12.0	2.0	1.0	2.0	3.0	1.0	5.0	

Encoded Dataset

- **KNNImputer**: This is a technique to fill in missing values in your data by using the K-Nearest Neighbors (KNN) approach. The idea is that the missing value can be predicted based on the values of the closest (most similar) data points.
- **n_neighbors=5:** This parameter tells the model to look at the 5 closest data points (or “neighbors”) to predict the missing value.
- **fit_transform(X_train):** This command does two things:
    1. **Fit:** The imputer looks at the data in X_train to understand the pattern of missing values and how it can fill them in using the 5 nearest neighbors.
    2. **Transform:** After understanding the data, it replaces the missing values with the estimated ones.

Finally, the data is converted back into a DataFrame format so that it can be easily used for further analysis or modeling.

```
Tenure                      0
City_Tier                   0
CC_Contacted_LY             0
Payment                     0
Gender                      0
Service_Score               0
Account_user_count          0
account_segment             0
CC_Agent_Score              0
Marital_Status              0
rev_per_month               0
Complain_ly                 0
rev_growth_yoy              0
coupon_used_for_payment     0
Day_Since_CC_connect        0
cashback                    0
Login_device                0
dtype: int64
------------------------------
Tenure                      0
City_Tier                   0
CC_Contacted_LY             0
Payment                     0
Gender                      0
Service_Score               0
Account_user_count          0
account_segment             0
CC_Agent_Score              0
Marital_Status              0
rev_per_month               0
Complain_ly                 0
rev_growth_yoy              0
coupon_used_for_payment     0
Day_Since_CC_connect        0
cashback                    0
Login_device                0
dtype: int64
```

After imputing, there is no null values

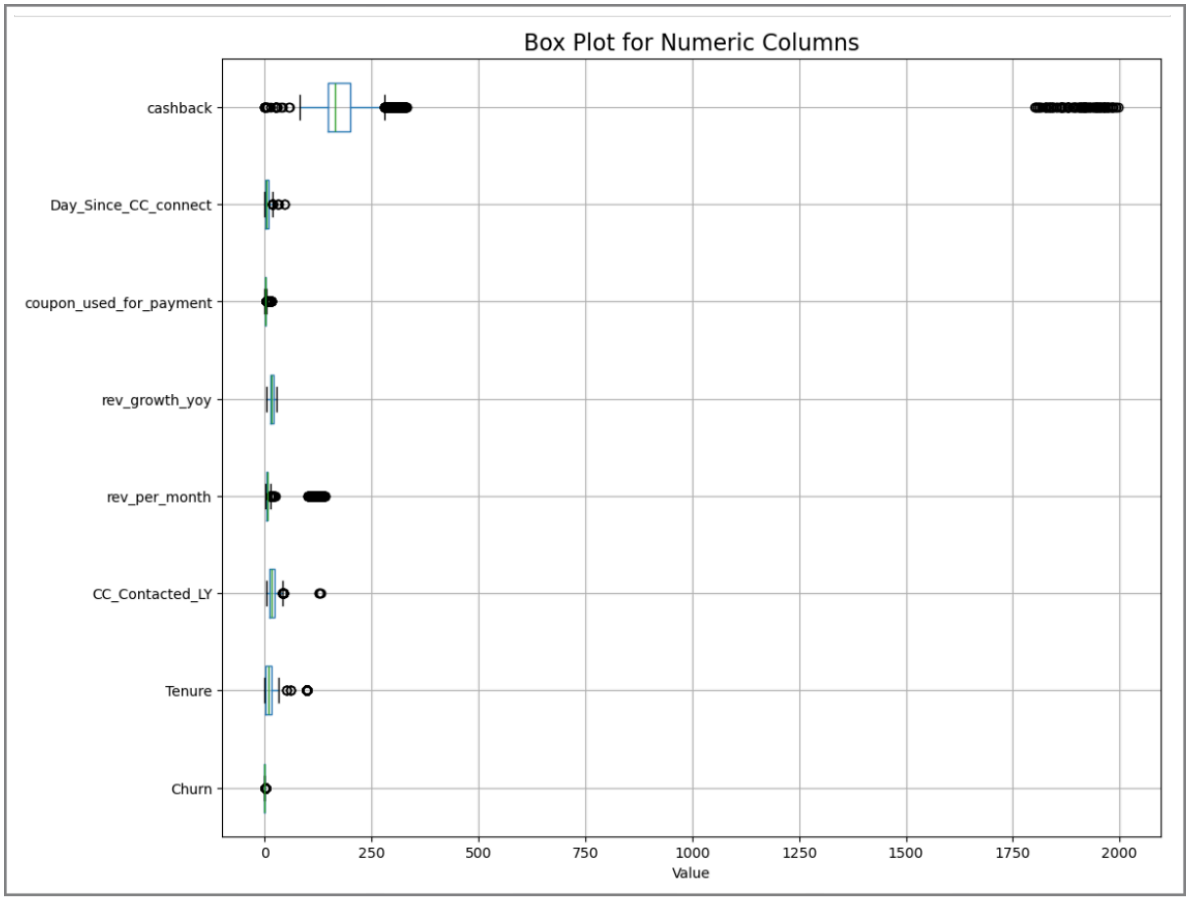
f. Outlier treatment

This dataset includes both continuous and categorical variables. Since each category represents a type of customer, it doesn't make sense to perform outlier treatment on categorical variables. Therefore, outlier treatment is applied only to continuous variables. The presence of outliers in a variable was determined using a box plot, where dots outside the upper limit of a quantile indicate outliers. There are eight continuous variables in the dataset: "Tenure," "CC_Contacted_LY," "Account_user_count," "cashback," "rev_per_month," "Day_Since_CC_connect," "coupon_used_for_payment," and "rev_growth_yoy." To address outliers, we used upper and lower limits for removal. Below is a pictorial representation of the variables before and after outlier treatment.

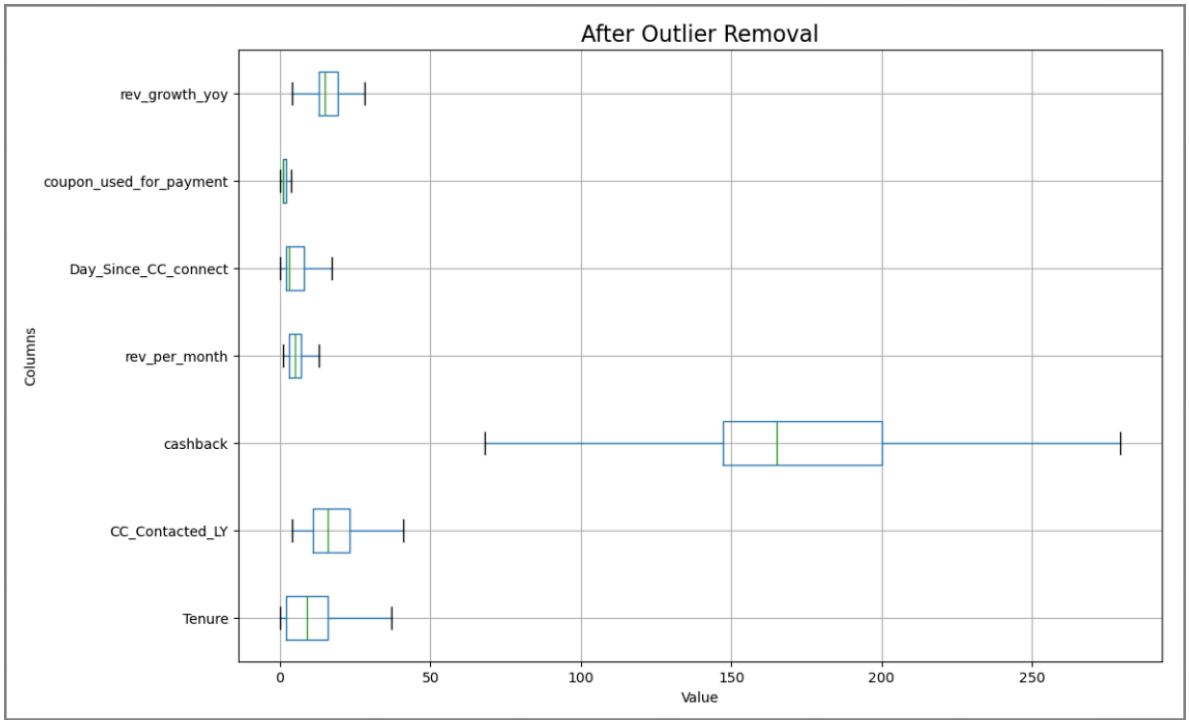
	Column	No. of outliers	Percentage of outliers
0	AccountID	0	0.000000
1	Churn	1896	16.838366
2	Tenure	139	1.234458
3	City_Tier	480	4.262877
4	CC_Contacted_LY	42	0.373002
5	Payment	0	0.000000
6	Gender	0	0.000000
7	Service_Score	90	0.799290
8	Account_user_count	1287	11.429840
9	account_segment	520	4.618117
10	CC_Agent_Score	0	0.000000
11	Marital_Status	0	0.000000
12	rev_per_month	185	1.642984
13	Complain_ly	0	0.000000
14	rev_growth_yoy	0	0.000000
15	coupon_used_for_payment	1380	12.255773
16	Day_Since_CC_connect	33	0.293073
17	cashback	879	7.806394
18	Login_device	539	4.786856

Outliers in each column

Before and after treatment of outliers



Boxplot with outliers



Boxplot after removing outliers

Variable transformation(Contd)

Variable transformation is a crucial step in data preprocessing that involves modifying variables to improve the performance of a model or to meet the assumptions of statistical techniques.

To scale features, the `StandardScaler` should be fitted only on the training data to learn the scaling parameters (mean and standard deviation close to 1), and then these parameters are used to transform both the training and test datasets. The correct approach is to use `sc.fit_transform(X_train)` to fit and scale the training data, and then apply `sc.transform(X_test)` to scale the test data based on the same parameters. This ensures that the test data is scaled consistently with the training data, avoiding data leakage and ensuring accurate model evaluation.

	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Stat
0	-1.148496	1.489238	-0.215234	-0.300658	-1.233817	-1.251189	-0.752048	-0.939742	-0.038437	0.8825
1	0.059570	-0.711240	-0.681723	-1.030607	0.815588	1.529487	1.406007	-0.152131	1.413271	0.8825
2	0.529374	1.489238	-0.565101	-1.030607	-1.233817	1.529487	0.326979	1.423089	-0.764291	-1.3616
3	0.976806	1.489238	1.767344	1.889190	0.815588	0.139149	1.406007	-0.939742	1.413271	0.8825
4	0.529374	-0.711240	-0.448478	-1.030607	-1.233817	0.139149	1.945521	0.635479	-0.038437	0.8825

Scaled train dataset

gent_Score	Marital_Status	rev_per_month	Complain_ly	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback	Login_device
-1.544121	-1.353581	1.277646	1.565011	-1.116150	-0.439236	-1.256704	-0.231599	-0.644715
-0.811490	-0.227929	-0.769584	1.565011	-0.299957	1.371880	2.020236	0.172123	-0.644715
1.386403	0.897722	-0.087174	1.565011	-0.572021	-0.439236	-0.437469	0.031917	1.126238
-1.544121	0.897722	0.254031	-0.653270	-0.844086	1.824659	0.654845	1.227724	-0.644715
1.386403	-1.353581	0.936441	-0.653270	-0.844086	-0.439236	-0.437469	-0.567083	1.126238

Scaled test dataset

Also converted all the columns to float datatype.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8445 entries, 0 to 8444
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Tenure                    8445 non-null   float64
 1   City_Tier                 8445 non-null   float64
 2   CC_Contacted_LY           8445 non-null   float64
 3   Payment                   8445 non-null   float64
 4   Gender                    8445 non-null   float64
 5   Service_Score             8445 non-null   float64
 6   Account_user_count        8445 non-null   float64
 7   account_segment           8445 non-null   float64
 8   CC_Agent_Score            8445 non-null   float64
 9   Marital_Status            8445 non-null   float64
 10  rev_per_month             8445 non-null   float64
 11  Complain_ly               8445 non-null   float64
 12  rev_growth_yoy            8445 non-null   float64
 13  coupon_used_for_payment   8445 non-null   float64
 14  Day_Since_CC_connect      8445 non-null   float64
 15  cashback                  8445 non-null   float64
 16  Login_device              8445 non-null   float64
dtypes: float64(17)
memory usage: 1.1 MB
```

g. Addition of new variables

Adding new variables is not required for the customer churn dataset, as the existing features likely provide adequate information for predicting churn. Introducing extra variables might complicate the analysis without offering significant benefits, potentially leading to overfitting or increased complexity.

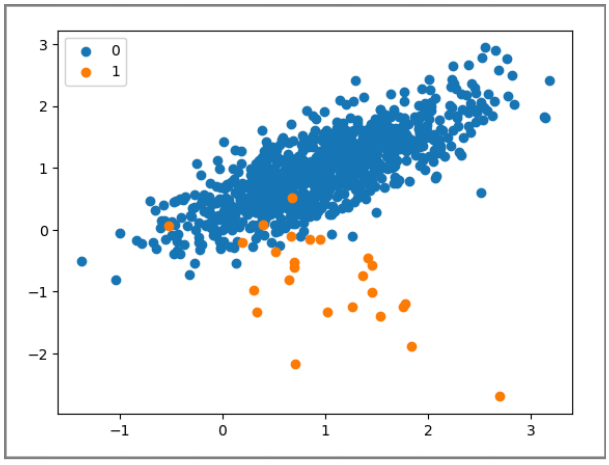
4. Business insights from EDA

a. Is the data unbalanced? If so, what can be done? Please explain in the context of the business

The provided dataset is imbalanced, with significant variation in the counts of the target variable “Churn”: 9364 instances of “0” and 1896 instances of “1”. To address this imbalance, the SMOTE technique is used to generate additional data points and balance the dataset. It is important to apply SMOTE only to the training data and not to the test data. The data has been split into training and test sets in a 75:25 ratio.

```
X_train (8445, 17)
X_test (2815, 17)
y_train (8445,)
y_test (2815,)
```

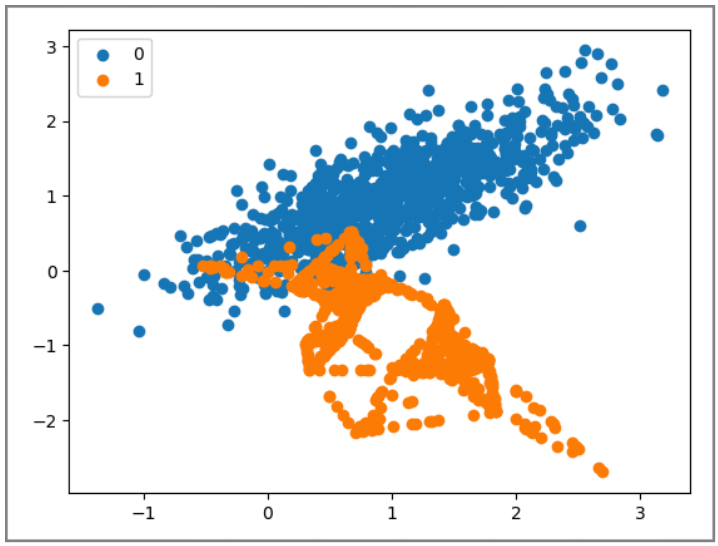
Before SMOTE is applied, shape of the data



Before SMOTE is applied

```
X_train_res (14046, 17)
y_train_res (14046,)
```

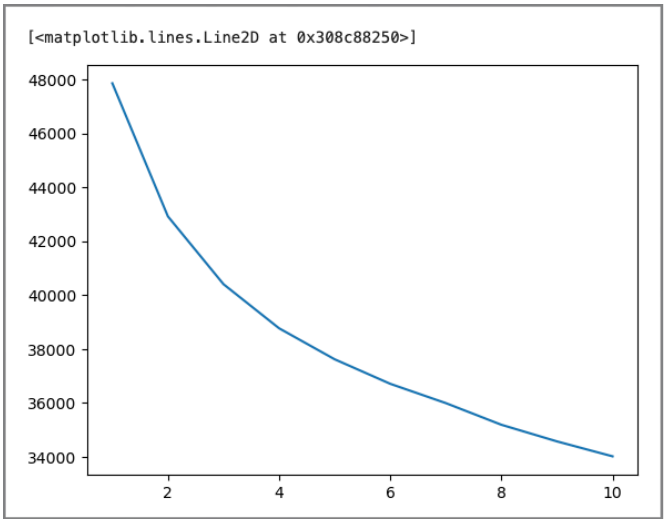
After SMOTE is applied, shape of the dataset



After SMOTE is applied

- The higher density of orange dots signifies an increase in the number of data points.

b. Any business insights using clustering



Elbow plot

- Created K- Means clustering using 3 clusters.
- The Elbow plot indicates that the optimal number of clusters for K-means clustering is 3. This is evident from the sharp bend or "elbow" in the plot around the 3-cluster mark, where the rate of decrease in inertia (within-

cluster sum of squares) significantly slows down. This suggests that adding more clusters beyond 3 does not substantially improve the model's performance, making 3 clusters an appropriate choice for segmenting the data.

c. Any business insights using clustering

- **City_Tier**: City_Tier reflects Tier of primary customer's city, influencing their spending behavior and service preferences. Expanding visibility in tier 2 cities could help acquire new customers, while ensuring services meet the expectations of higher-tier city customers to prevent churn.
- **CC_Contacted_LY:** Frequent contact with customer care, indicated by CC_Contacted_LY, may suggest dissatisfaction or recurring issues, increasing churn risk. Training customer care executives to resolve these issues effectively can enhance customer experience and reduce churn.
- **Payment:** The Payment variable could correlate with customer loyalty, as certain methods may be more convenient. Promoting hassle-free payment options like standing instructions in bank accounts or UPI can enhance convenience and encourage customer retention.
- **Gender:** The Gender variable might reveal differences in customer behavior, guiding marketing strategies. Tailoring campaigns to resonate with the preferences of different genders(Male or female) can improve engagement and reduce churn.
- **Service_Score and account_segment:** Service_Score, account_segment, and Clus_kmeans variables provide insights into how different segments perceive service quality. Improving service scores and targeting retention efforts towards segments at higher risk of churn can enhance customer satisfaction.

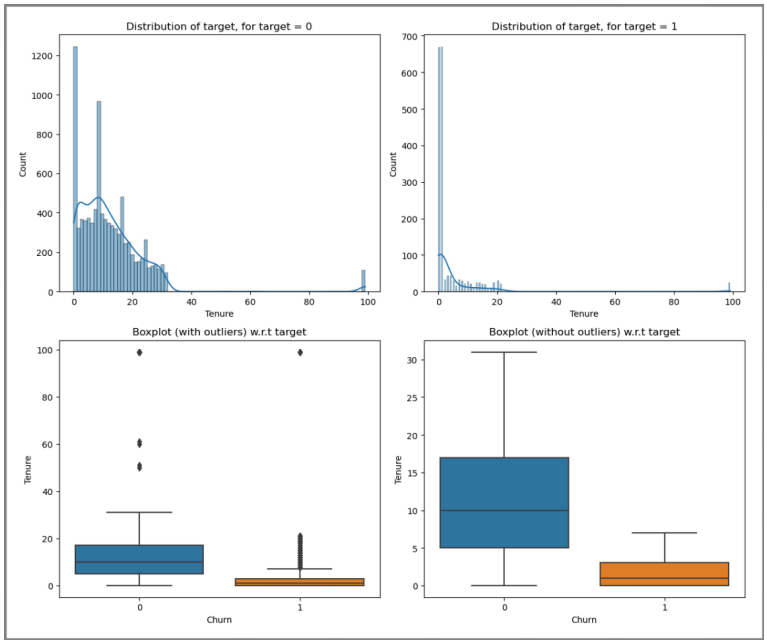
- **rev_per_month, rev_growth_yoy, and cashback:** Variables like rev_per_month, rev_growth_yoy, and cashback indicate profitability and growth potential. Customers with high revenue but low growth might need incentives to increase usage or adopt additional services, driving growth.
- **coupon_used_for_payment:** High coupon usage, as indicated by coupon_used_for_payment, suggests price sensitivity. Leveraging promotional offers in retention strategies can appeal to price-sensitive customers and encourage loyalty.
- **Complain_ly:** The Complain_ly variable highlights the importance of reducing complaints through improved service quality and proactive customer support, which can significantly lower churn rates.
- **Login_device**: The Login_device variable offers insights into customer interaction with the service. Optimizing user experience and targeting device-specific campaigns can enhance satisfaction and retention.

By combining these insights with strategic business actions, the company can effectively improve customer satisfaction, reduce churn, and drive growth across different customer segments.

More about Business Insight and Visualisation

Distribution of Target w.r.t Independent variables

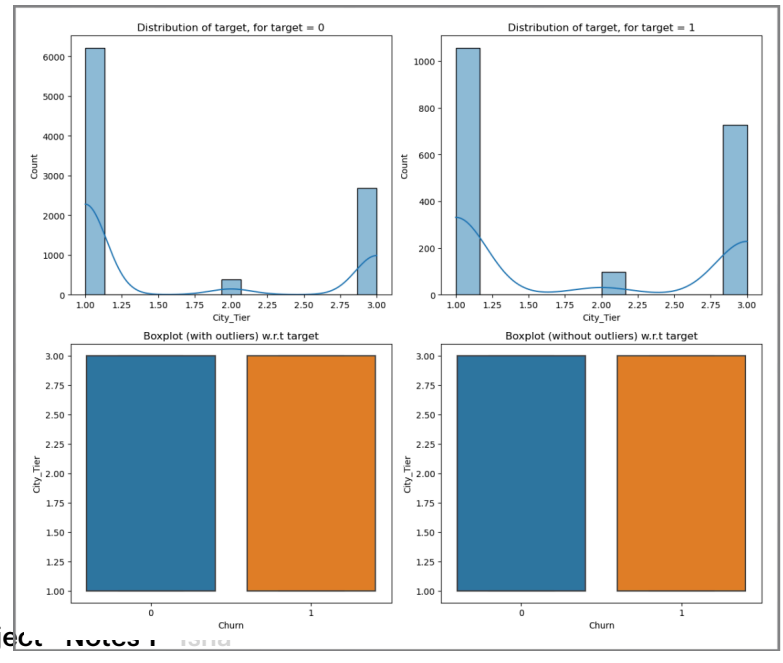
1. Churn vs Tenure



Churn vs Tenure

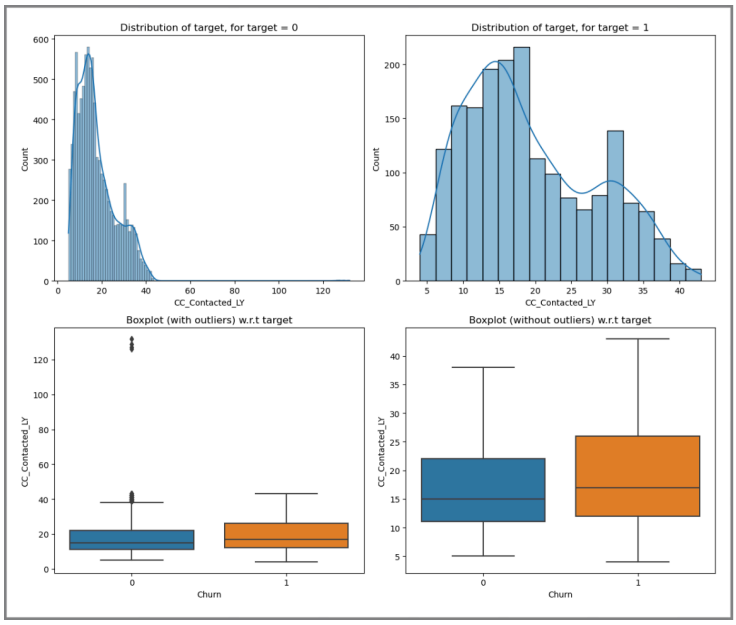
- Customers with shorter tenure, typically new customers, tend to churn more frequently.

2. **Churn vs City_Tier**



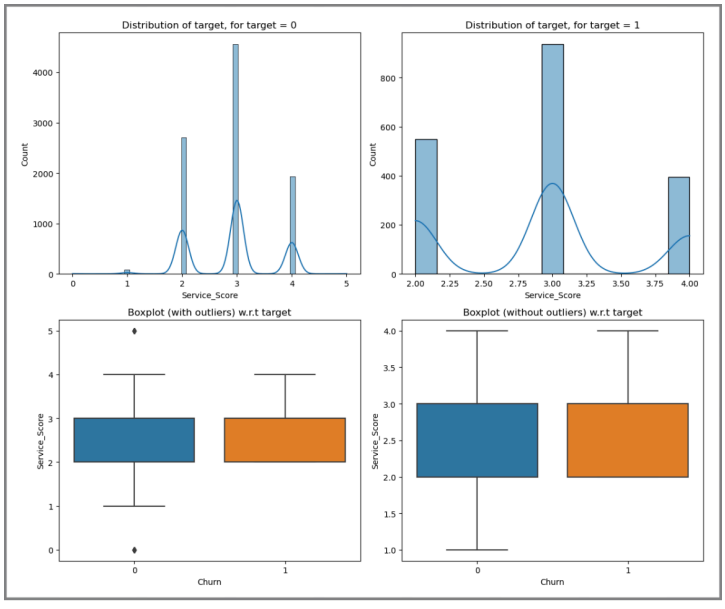
- There’s minimal difference in churn rates between churned and non-churned customers based on City Tier.

3. **Churn vs CC_Contacted_LY**



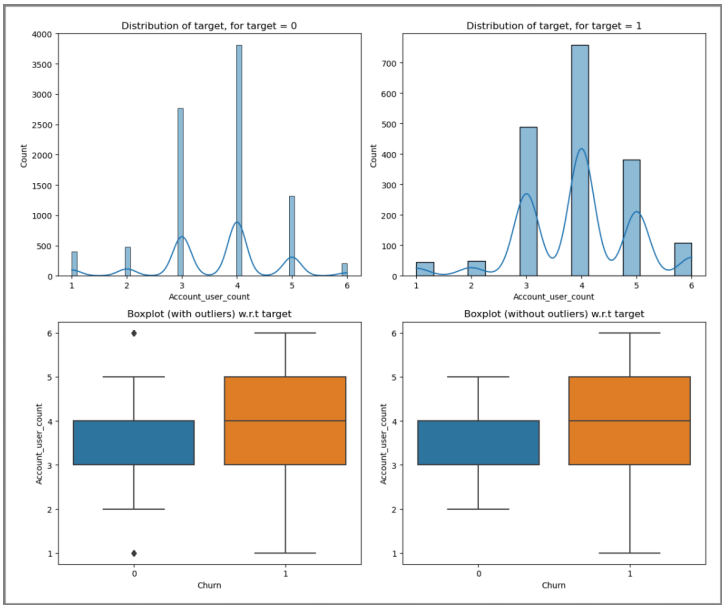
- Customers who contact customer care 12 to 27 times a year are more likely to churn.

4. **Churn vs Service_Score**



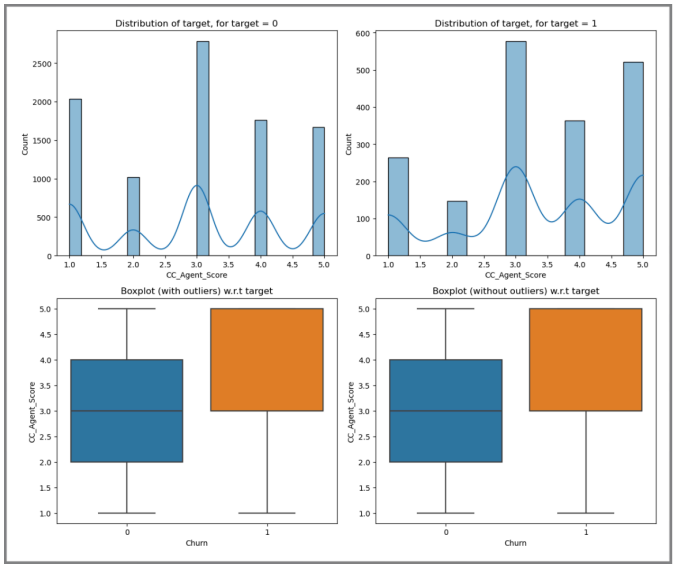
- Customers who gave a service score of 2, 3, or 4 have higher attrition rates.

5. **Churn vs Account_user_count**



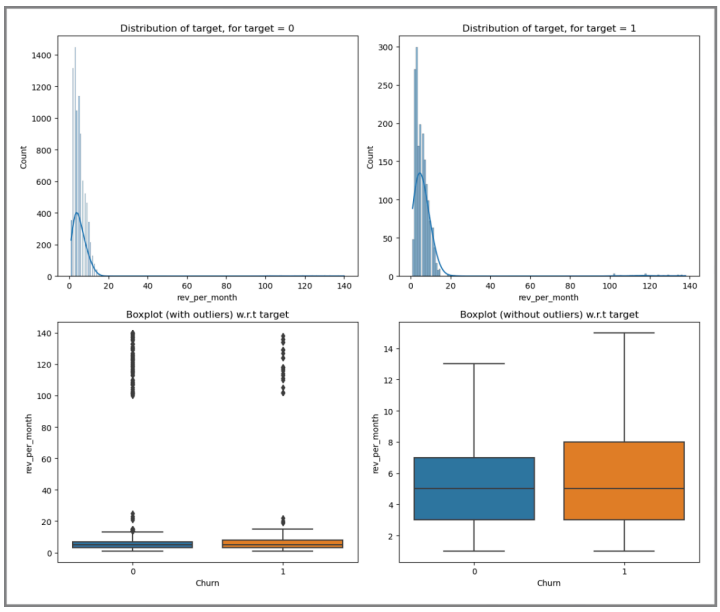
- Churn is more prevalent among customers with three or more users linked to their account.

6. **Churn vs CC_Agent_Score**



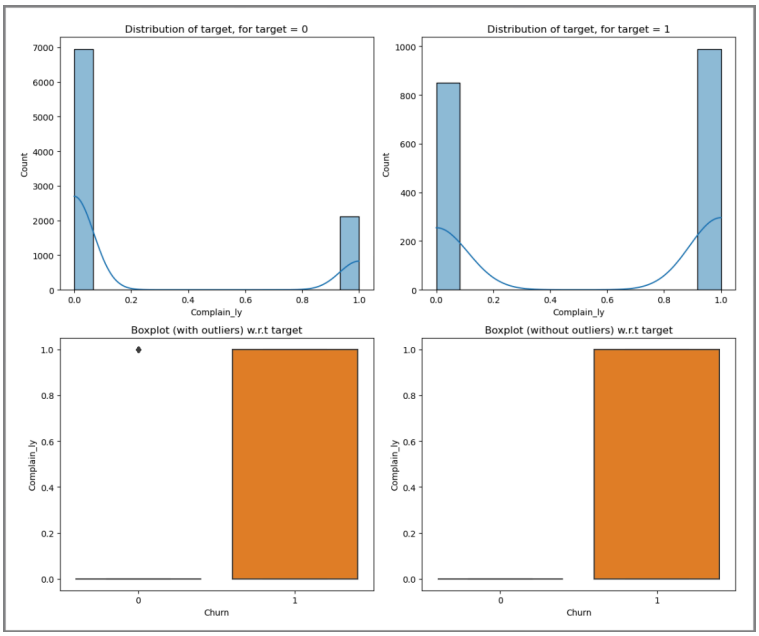
- Customers who rated the agent with a score of 3 or higher are more likely to churn.

7. **Churn vs rev_per_month**



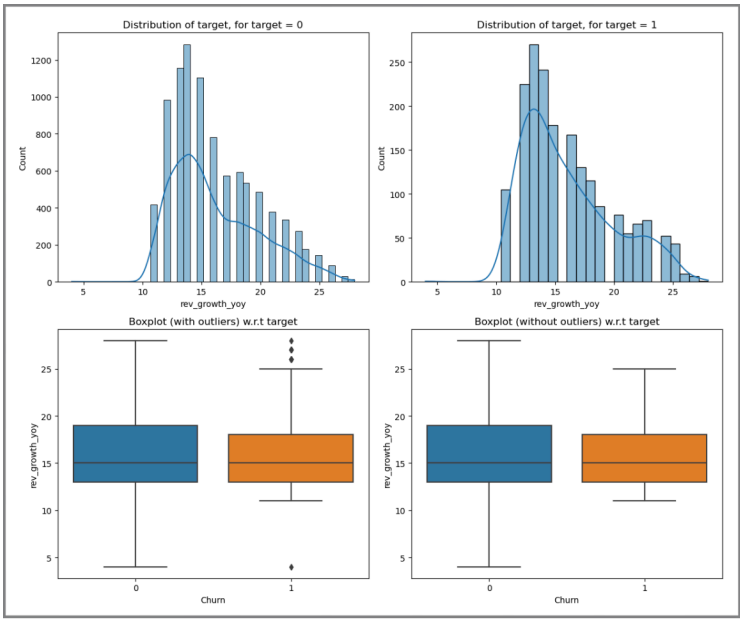
- Customers generating lower average revenue per month tend to churn at a higher rate.

8. **Churn vs Complain_ly**



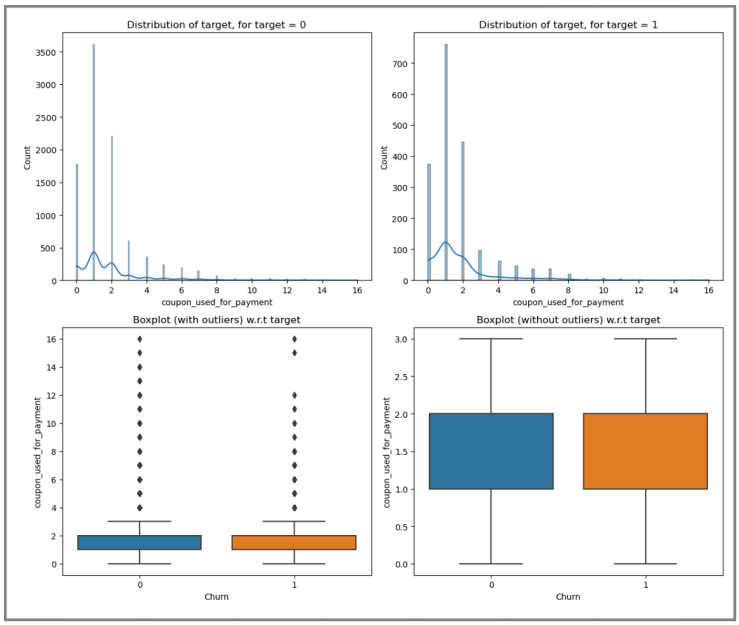
- Customers who have raised complaints are more likely to churn compared to those who haven't.

9. **Churn vs rev_growth_yoy**



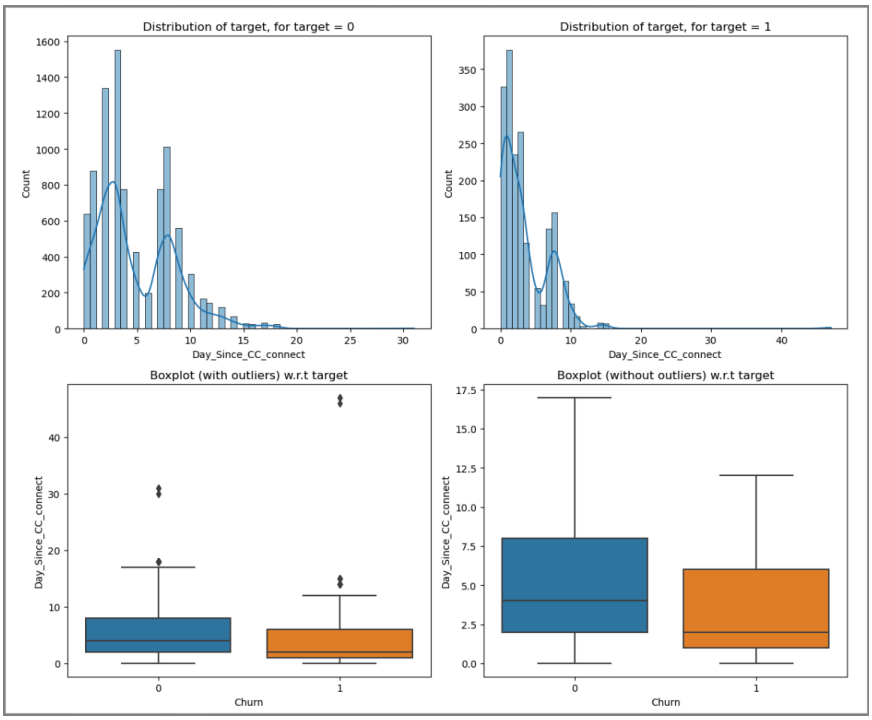
- A lower revenue growth percentage over the past year correlates with a higher likelihood of churn.

10. **Churn vs coupon_used_for_payment**



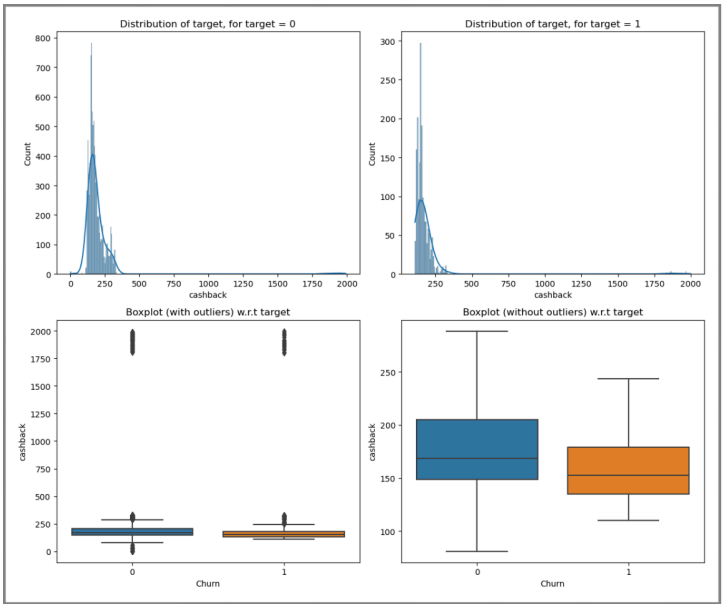
- There’s little difference in churn rates based on the number of coupons used for payment.

11. **Churn vs Day_Since_CC_connect**



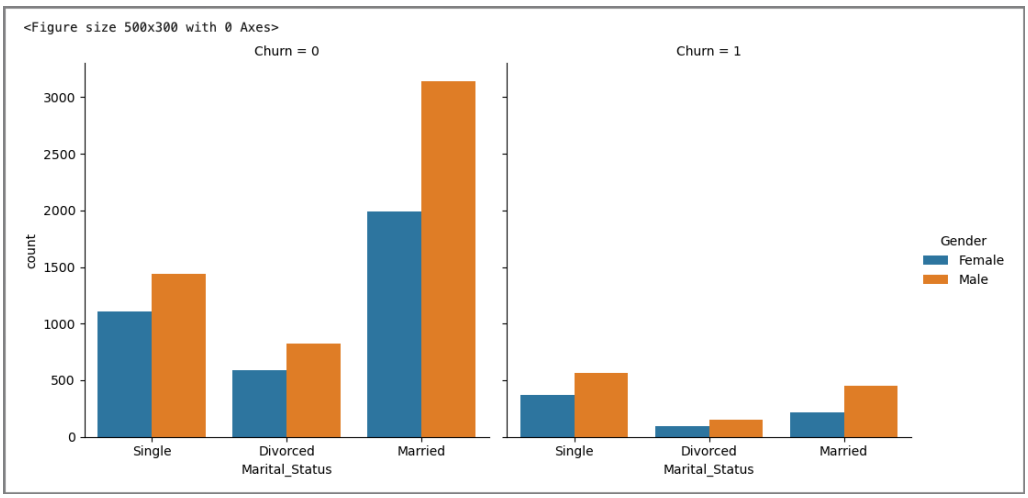
- Customers who contacted customer care within a week are at a higher risk of churning.

12. **Churn vs cashback**



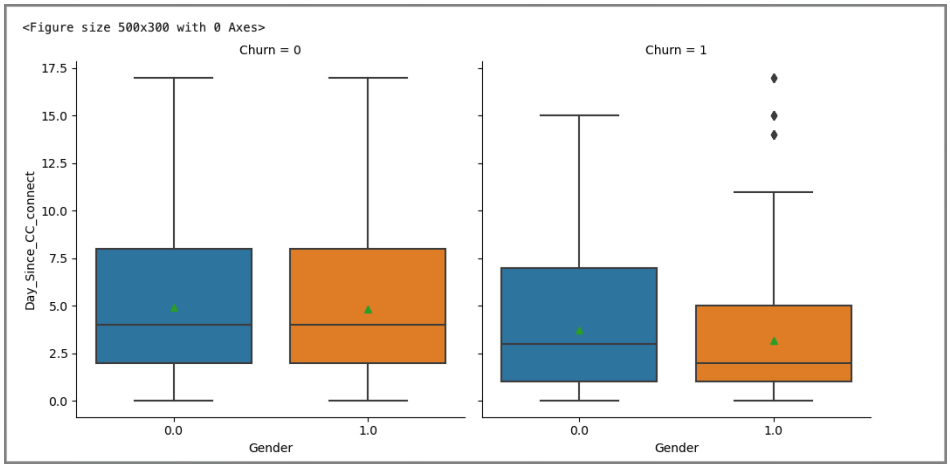
- Churn is higher among customers with an average monthly cashback under 200 INR.

13. **Churn vs Gender vs Maritial_Status**



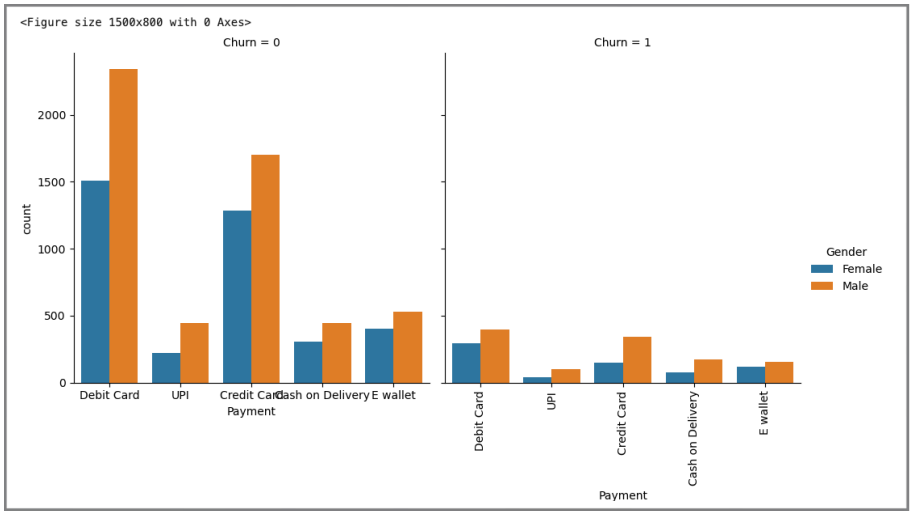
- Across different marital statuses, male customers exhibit higher churn rates than female customers.

14. **Churn vs Gender**



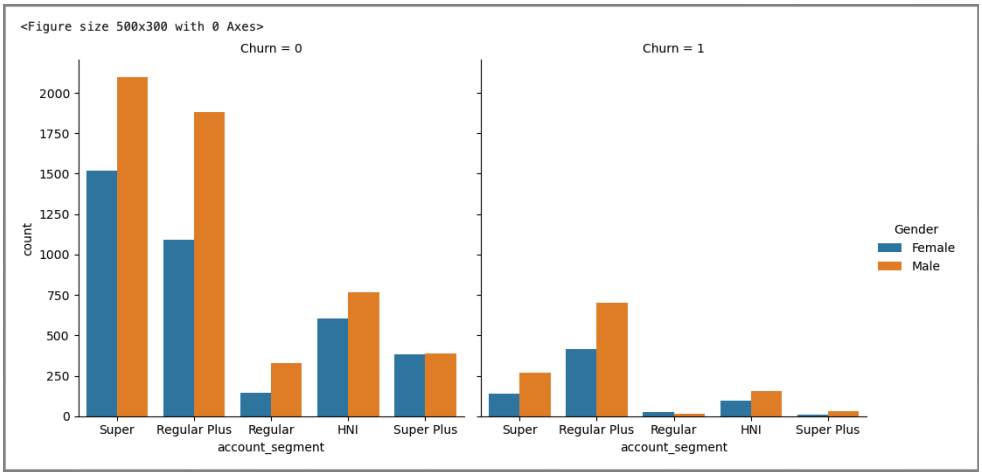
- Male customers who contact customer care more frequently than female customers are more likely to churn.

15. **Churn vs Gender vs Payment**



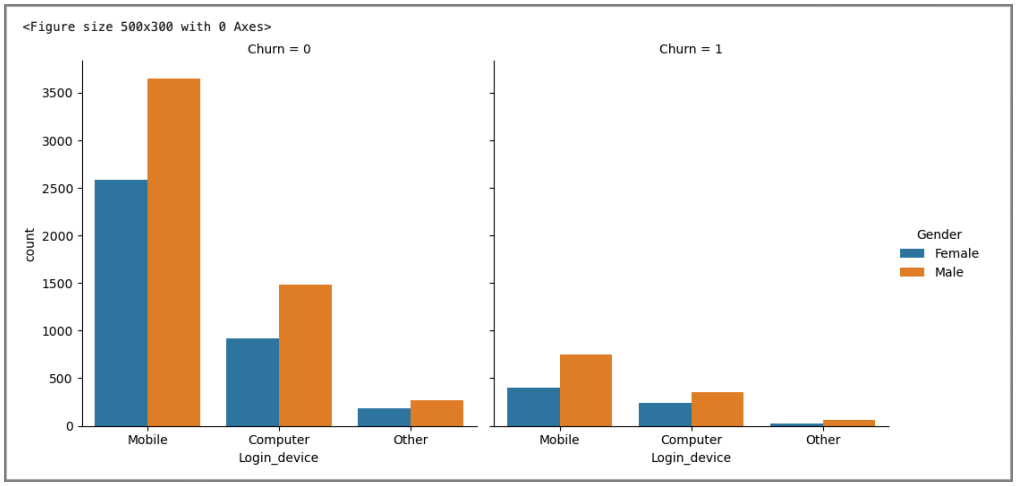
- Male customers have higher churn rates than female customers across various payment methods.

16. **Churn vs Gender vs account_segment**



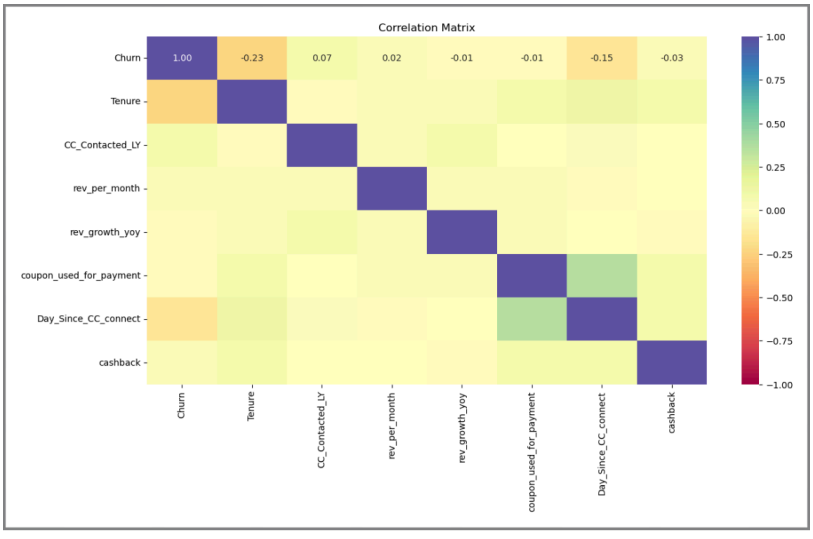
- Among ‘Regular’ account segment customers, females exhibit higher churn rates than males.Otherwise, all account segments, male customers have a higher churn rate.

17. **Churn vs Gender vs Login_device**



- Male customers churn more frequently than female customers, regardless of their preferred login device.

18. **Heatmap**



1. **Churn**:

   • Negative Correlation with Tenure (-0.233): Indicates that customers with longer tenure are less likely to churn.

   • Positive Correlation with Complain_ly (0.251): Suggests that customers who have lodged complaints in the previous year are more likely to churn.

   • Negative Correlation with Day_Since_CC_connect (-0.148): Indicates that the more recent the customer care connection, the less likely the customer is to churn.

2. **Tenure**:

   • Positive Correlation with Day_Since_CC_connect (0.123): Longer-tenured customers tend to have more days since their last customer care connection.

   • Positive Correlation with coupon_used_for_payment (0.089): Slight positive correlation indicating that longer-tenured customers might use more coupons.

3. **City_Tier:**

   • Low Correlation with all other variables: This suggests that the city tier has minimal impact on the other factors, indicating that customers across different city tiers behave similarly in these aspects.

4. **CC_Contacted_LY (Customer Care Contacted Last Year):**

   • Positive Correlation with Service_Score (0.060): Suggests that customers who contacted customer care last year had a slightly better service score.

   • Slightly Positive Correlation with Rev_growth_yoy (0.073): Indicates a marginal relationship between contacting customer care and revenue growth year-over-year.

5. **Service_Score:**

   • Strong Positive Correlation with Account_user_count (0.323): Indicates that higher service scores are associated with accounts that have more users.

   • Positive Correlation with coupon_used_for_payment (0.182): Suggests that customers with higher service scores are more likely to use coupons for payment.

6. **Account_user_count:**

   • Positive Correlation with coupon_used_for_payment (0.146): Indicates that accounts with more users tend to use more coupons for payment.

   • Positive Correlation with Service_Score (0.323): Aligns with the above point that larger accounts generally receive better service scores.

7. **CC_Agent_Score:**

   • Low correlation with other variables: This suggests that the score given to the customer care agent by the customer has minimal direct influence on other factors in the dataset.

8. **Rev_per_month (Revenue per Month):**

   • Low correlation with other variables: This suggests that monthly revenue is relatively independent of the other variables in this dataset.

9. **Complain_ly:**

   • Positive Correlation with Churn (0.251): Customers who have complained in the last year are more likely to churn, a significant observation for customer retention strategies.

10. **Rev_growth_yoy:**

    • Positive Correlation with Service_Score (0.103): Indicates a slight relationship where better service scores might be associated with revenue growth.

11. **coupon_used_for_payment:**

    • Strong Positive Correlation with Day_Since_CC_connect

	Churn	Tenure	CC_Contacted_LY	rev_per_month	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connec
Churn	1.000000	-0.233150	0.072071	0.022538	-0.013877	-0.014826	-0.14795
Tenure	-0.233150	1.000000	-0.004261	0.028431	0.018824	0.089171	0.12261
CC_Contacted_LY	0.072071	-0.004261	1.000000	0.015675	0.072913	0.004969	0.01293
rev_per_month	0.022538	0.028431	0.015675	1.000000	0.024114	0.016548	-0.00092
rev_growth_yoy	-0.013877	0.018824	0.072913	0.024114	1.000000	0.018341	0.00220
coupon_used_for_payment	-0.014826	0.089171	0.004969	0.016548	0.018341	1.000000	0.36173
Day_Since_CC_connect	-0.147956	0.122612	0.012938	-0.000923	0.002206	0.361735	1.00000
cashback	-0.032382	0.078416	0.002679	0.002974	-0.001157	0.072861	0.08446

Correlation table

(0.362): Suggests that customers who have used coupons for payment tend to have a more recent connection with customer care.

12. **Day_Since_CC_connect:**

    • Positive Correlation with coupon_used_for_payment (0.362): As mentioned above, there's a strong relationship between the last connection with customer care and coupon usage.

13. **cashback:**

    • Low correlation with most variables: Indicates that cashback offers are relatively independent in terms of their influence on other factors in the dataset.