CAPSTONE PROJECT

ON

Customer Churn Prediction

By,

Shobha Kumari Choudhary

Date: - 10th March'24

Contents

List of Figures

Introduction of the business problem

1.a) Defining problem statement:

An E Commerce company or DTH (you can choose either of these two domains) provider is facing a lot of competition in the current market and it has become a challenge to retain the existing customers in the current situation. Hence, the company wants to develop a model through which they can do churn prediction of the accounts and provide segmented offers to the potential churners. In this company, account churn is a major thing because 1 account can have multiple customers. hence by losing one account the company might be losing more than one customer.

You have been assigned to develop a churn prediction model for this company and provide business recommendations on the campaign.

Your campaign suggestion should be unique and be very clear on the campaign offer because your recommendation will go through the revenue assurance team. If they find that you are giving a lot of free (or subsidized) stuff thereby making a loss to the company; they are not going to approve your recommendation. Hence be very careful while providing campaign recommendation

- **Here the objective is to Build a Model to identify the Customers who will potentially Churn and provide useful business recommendations. Retaining a loyal customer is far more important than acquiring a new one. Hence predictive analysis of customer Retention is absolutely necessary in all business.**

1.b) Need of the study/project:

- **Customer churn is when the customers either switch from their incumbent service provider to its competitor or stop using the service altogether.**
- **There are so many competitions available in the market for any product /services we use, so it is very necessary to study about customers likes & dislikes, their preferences about the product, their needs etc.**
- **There is a need to Study about this project in detail because Churning of customers impacts the Revenue of the company.**
- **A High Churn Rate will result in decrease in Companies Revenue.**
- **Customer churn prediction using machine learning will help us to identify risky customers and understand why customers are willing to leave.**

1.c) Understanding business/social opportunity:

- **The Goal of this project is to predict the type of customers that has the potential to churn by identifying various features available that can help in minimizing the customers churn rates and get the right business insights.**
- **We need to find a way to Retain the customers as long as possible.**
- **By knowing the customer needs, the kind of services they prefer, we need to keep the customers satisfied.**

- By knowing the reason that are affecting customer churn will help us predict the churn and then avoid it.
- There can be many reasons for customers churn - like poor customer Services, another one could be prices.
- attracting new customers is much more expensive than retaining existing ones.

Data Report

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

- Data would be collected from the existing business and their past records of various customers. Data is collected w.r.t different features like gender, Marital status, Service score, Revenue generated etc.

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

❖ Sample of the Dataset:

	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

Marital_Status	rev_per_month	Complain_ly	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback	Login_device
Single	9	1.0	11	1	5	159.93	Mobile
Single	7	1.0	15	0	0	120.9	Mobile
Single	6	1.0	14	0	3	NaN	Mobile
Single	8	0.0	23	0	3	134.07	Mobile
Single	3	0.0	11	1	3	129.6	Mobile

❖ Shape of the Dataset:

- The Number of Rows present in the Dataset are: 11260
- The Number of Columns present in the Dataset are: 19

❖ Statistical Summary:

	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

- **Observations:**

- Columns like Tenure, Account_user_count, rev_per_month, rev_growth_yoy,
- coupon_used_for_payment, Day_Since_CC_connect and cashback are continuous columns but missing from the Describe output being a numeric column. This indicates the presence of some Anomalies or some Special Character in these columns.
- Columns like City_Tier, Service_Score, CC_Agent_Score and Complain_ly seems Contain Categorical data.
- Average times Customers contacted CC is around 17 times and Max 132 times.
- Average complains raised by the customers is around 28%.
- 75% of customers having service score less than or equal to 3.
- 75% of customers have given a satisfaction score of less than or equal to 4.

2.c) Understanding of attributes (variable info, renaming if required)

- **Dataset Info:**

```
<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
```

- **Observations**

- Null values are present in the dataset.
- There are 5 float64, 2 int64 and 12 object Datatypes.
- Columns like Tenure, Account_user_count, rev_per_month, rev_growth_yoy, coupon_used_for_payment, Day_Since_CC_connect and cashback contains integer values but their Datatype is object here.
- This may be due to some Anomalies or any special character present in the data.
- We need to inspect such columns for any Anomalies/Special character present in it.

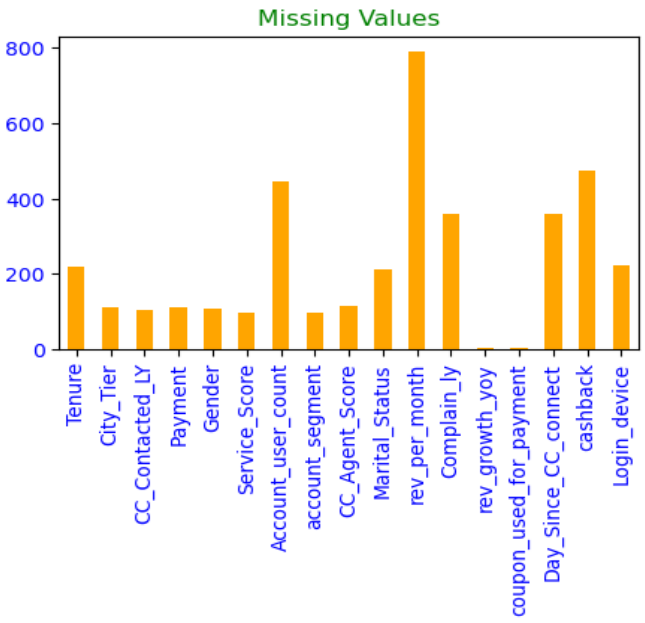
❖ Duplicates:

- There are No Duplicate Rows present in the Dataset.

❖ Missing Values:

```
Missing Values:

AccountID                    0
Churn                        0
Tenure                     102
City_Tier                  112
CC_Contacted_LY            102
Payment                    109
Gender                     108
Service_Score               98
Account_user_count         112
account_segment             97
CC_Agent_Score             116
Marital_Status             212
rev_per_month              102
Complain_ly                357
rev_growth_yoy               0
coupon_used_for_payment      0
Day_Since_CC_connect       357
cashback                   471
Login_device               221
dtype: int64
```



❖ Checking for Anomalies/Discrepancies:

- Column Tenure contains a Special character '#' as an entry making it an object datatype.
- Column Gender contains four unique values as 'Male', 'Female',' M','F', which needs to be replaced with 'Male' and 'female'.
- Account_user_count also contains special character '@' as one of the entries.
- Column account_segment contains some Anomalies like 'Regualr plus' and 'Super Plus' is written in two different ways which needs to be corrected.
- Column rev_per_month Contains '+' plus which needs to be replaced.
- Column rev_growth_yoy and Day_Since_CC_connect contains '$' character.
- Column coupon_used_for_payment contains #, $ and * characters.
- Column Login_device Contains &&&& as one of its categories.

- We need to inspect the Cashback column for any non-digit character present in it.
- We saw that Cashback column contains special character '$' due to which it is considered as 'object' datatype column.

We will Replace the special characters present in the dataset with np. Nan and later will impute the Null values.

❖ Post Removing Anomalies:

	count	mean	std	min	25%	50%	75%	max
AccountID	11260.0	25629.500000	3250.626350	20000.0	22814.75	25629.50	28444.25	31259.0
Churn	11260.0	0.168384	0.374223	0.0	0.00	0.00	0.00	1.0
Tenure	11042.0	11.025086	12.879782	0.0	2.00	9.00	16.00	99.0
City_Tier	11148.0	1.653929	0.915015	1.0	1.00	1.00	3.00	3.0
CC_Contacted_LY	11158.0	17.867091	8.853269	4.0	11.00	16.00	23.00	132.0
Service_Score	11162.0	2.902526	0.725584	0.0	2.00	3.00	3.00	5.0
Account_user_count	10816.0	3.692862	1.022976	1.0	3.00	4.00	4.00	6.0
CC_Agent_Score	11144.0	3.066493	1.379772	1.0	2.00	3.00	4.00	5.0
rev_per_month	10469.0	6.362594	11.909686	1.0	3.00	5.00	7.00	140.0
Complain_ly	10903.0	0.285334	0.451594	0.0	0.00	0.00	1.00	1.0
rev_growth_yoy	11257.0	16.193391	3.757721	4.0	13.00	15.00	19.00	28.0
coupon_used_for_payment	11257.0	1.790619	1.969551	0.0	1.00	1.00	2.00	16.0
Day_Since_CC_connect	10902.0	4.633187	3.697637	0.0	2.00	3.00	8.00	47.0
cashback	10787.0	196.236370	178.660514	0.0	147.21	165.25	200.01	1997.0

❖ Checking for Outliers:

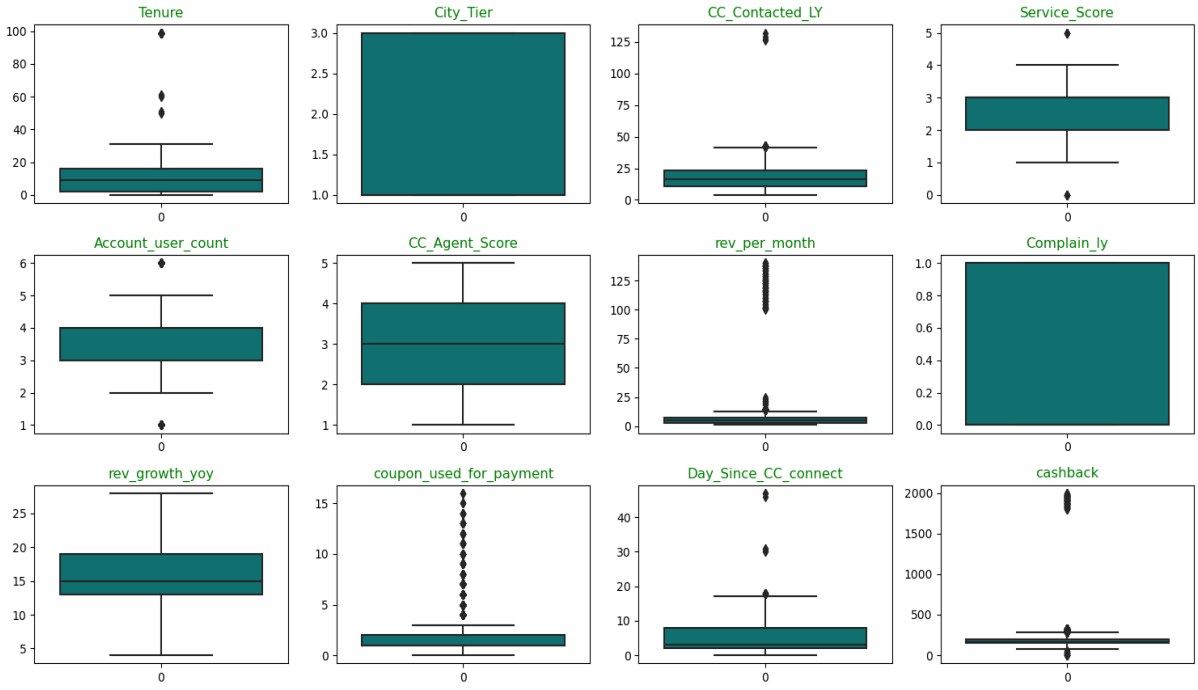


Fig no-1: Outlier Check

❖ Observations:

➢ There are many features showing the presence of Outliers.

Exploratory data analysis

Univariate Analysis

❖ Count plot of all Categorical Columns:

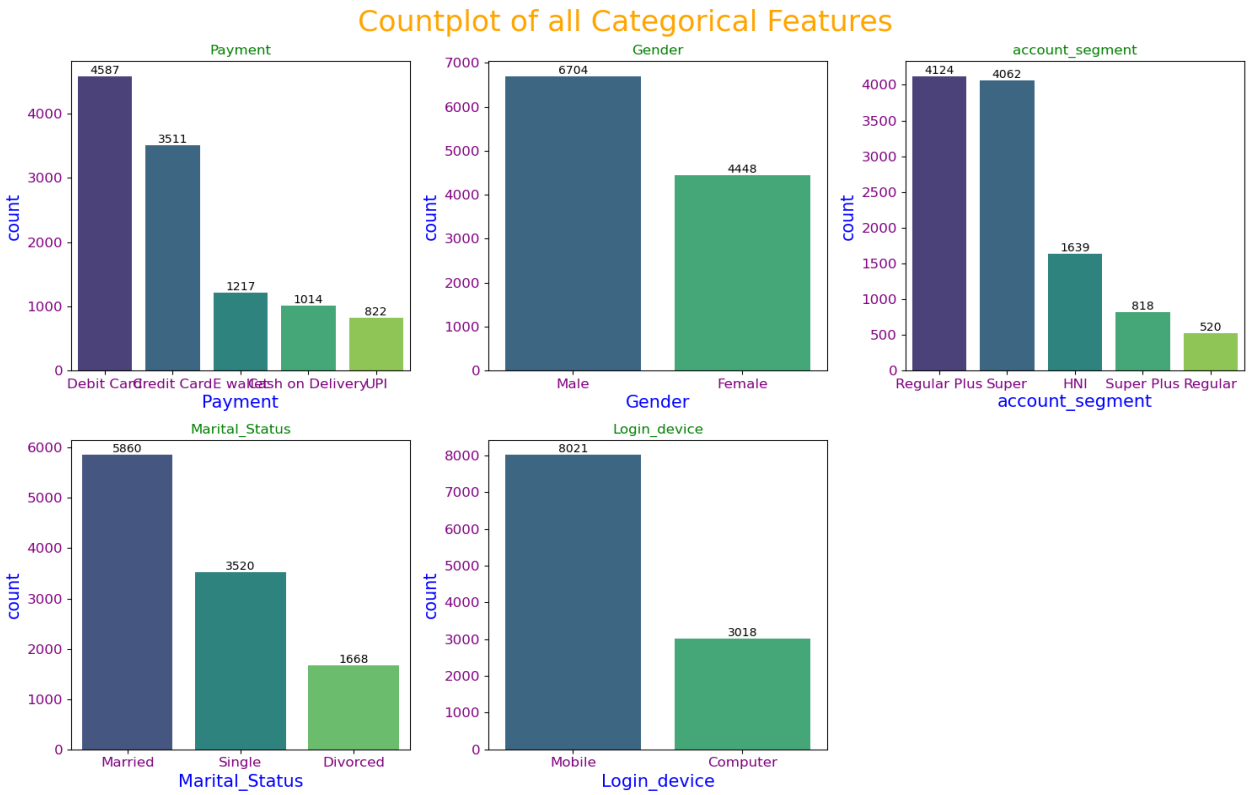


Fig no-2: Count plot

Insights:

- Most Customers Preferred Payment mode are Debit card and Credit Card.
- Very few customers make payment through UPI.
- Male customers are more as compared to female.
- There are more customers that belong to account_segment Regular plus and Super.
- Very less customers belongs to Regular segment on the basis of their spend.
- Married customers are more in number as compared to Singles and Divorced.
- Maximum of the customers prefers to Login Via Mobile as it is convenient to carry and they can Login from any Location.
- Very less customers Login via Computer.

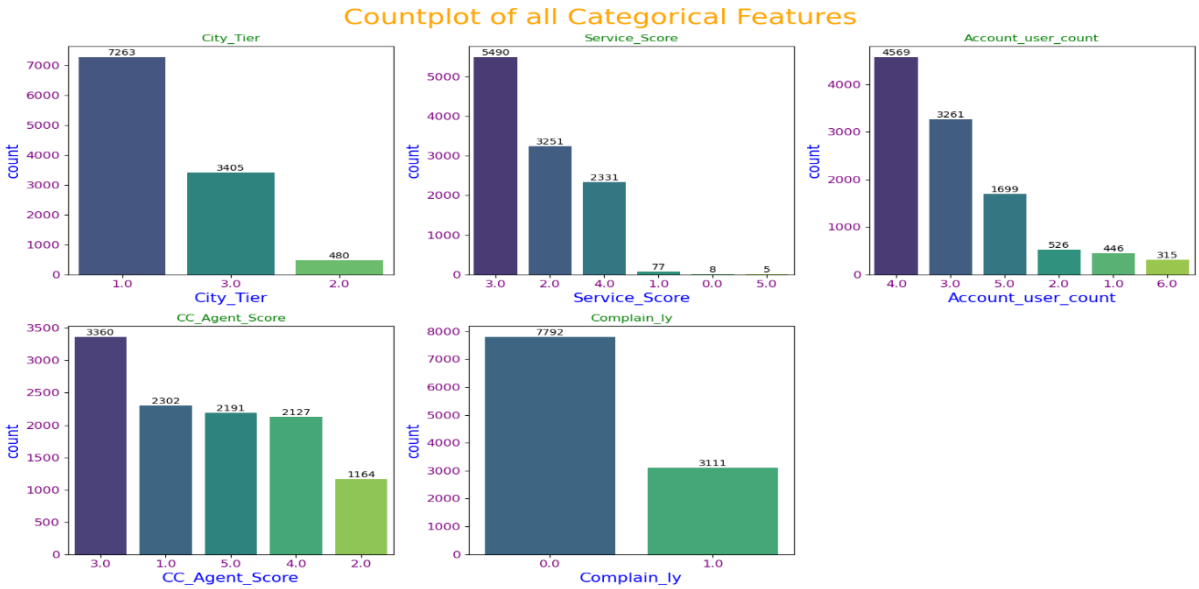


Fig no-3: Count plot

Insights:

- Maximum customers belong to Tier-1 cities followed by Tier-3 cities and very few belongs to Tier-2 cities.
- Service score given by Maximum customers is 3, which states that customers are not fully satisfied by the service provided by the company.
- More numbers of Accounts are linked with 4 members, followed by 3 and 5 members per Account.
- Very few accounts are linked with 1,2 and 6 Members.
- Satisfaction score given by customers on an Average is 3 for Customer Care Services.
- Also, Excellent Score of 4 and 5 for Customer care services are almost equal in number. So, they are Satisfied customers too.
- It seems very less Complaint has been Raised by the customers in last 12 months.

❖ Distribution of Target Variable:

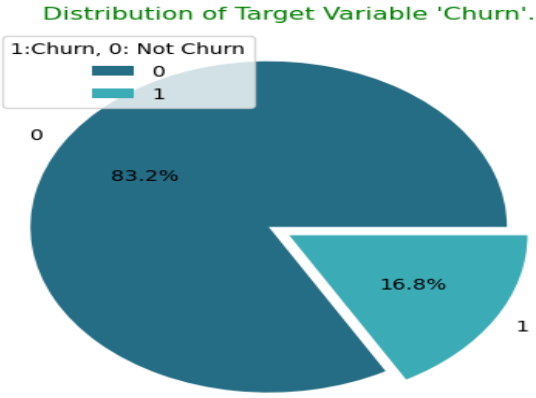


Fig no-4: Pie Chart

Insights:

- 16.8% of the customers have churned whereas 83.2% of the customers have not churned.
- seems Imbalance in Target class.

❖ Histogram and Boxplot of all continuous variables

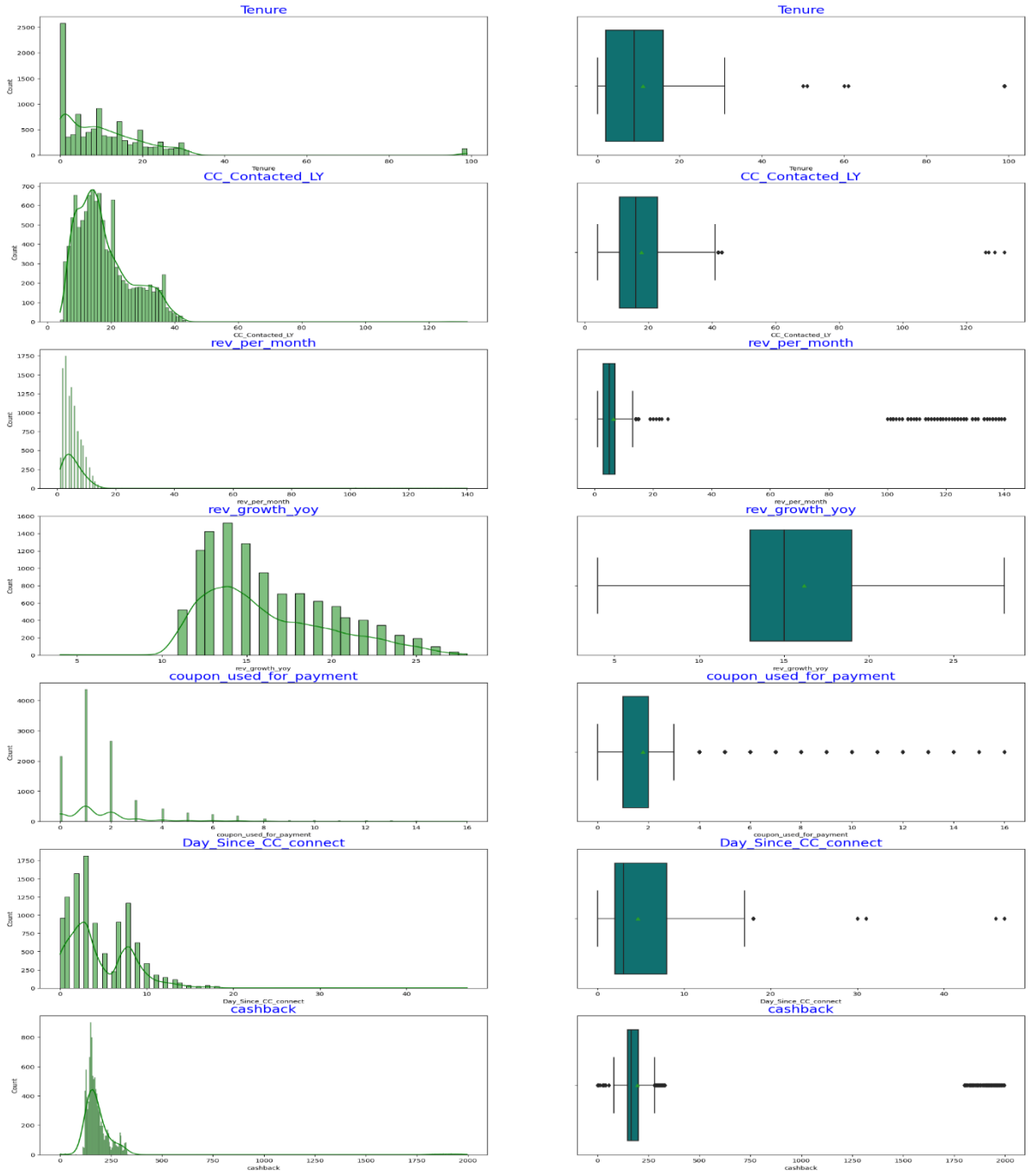


Fig no-5: Continuous variable plot

Insights:

- All the Variables are Right Skewed showing the presence of Outliers.
- Maximum customers have a Tenure of less than a month.
- There are also some customers having a Tenure of more than 50 months, Max up to 100 months.
- Maximum number of customers have contacted Customer care 11 to 23 times in last 12 months.
- The Median of the Monthly average revenue generated by the company is around 10k (Assuming the Currency is in Thousands).
- Also, this feature is highly Right Skewed showing monthly avg revenue generated by the company more than 100k.
- There is an Approx 16% growth in revenue on an average, generated by the account in the last year compared to the previous years.
- On an Average, 2 times coupons were used to do the payment.
- Also, it seems some customers used the coupons more than 4 times to max 16 times.
- Avg no of days since customers have not contacted CC is around 5 days.
- On an Average, Cashback generated by the customers is around Rs 200/- in the last 12 months.
- Also, some customers generated cashback of more than Rs 1750/- monthly.

❖ **Visualizing Tenure:**

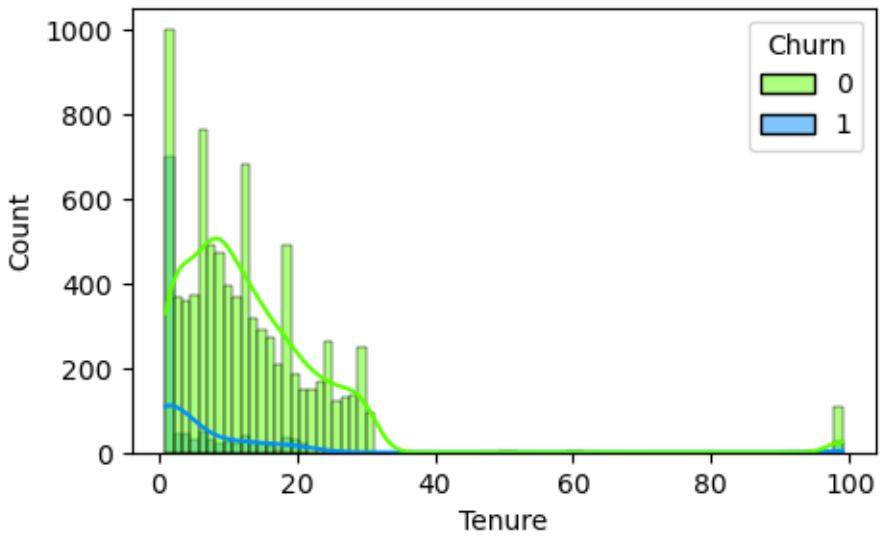


Fig no-6: Histogram of Tenure

Insights:

- Customers having a Tenure of less than 15 months, Churns more.
- There are some Loyal Customers having Tenure of more than 80 months.
- Distribution seems Overlapping.

Bivariate Analysis

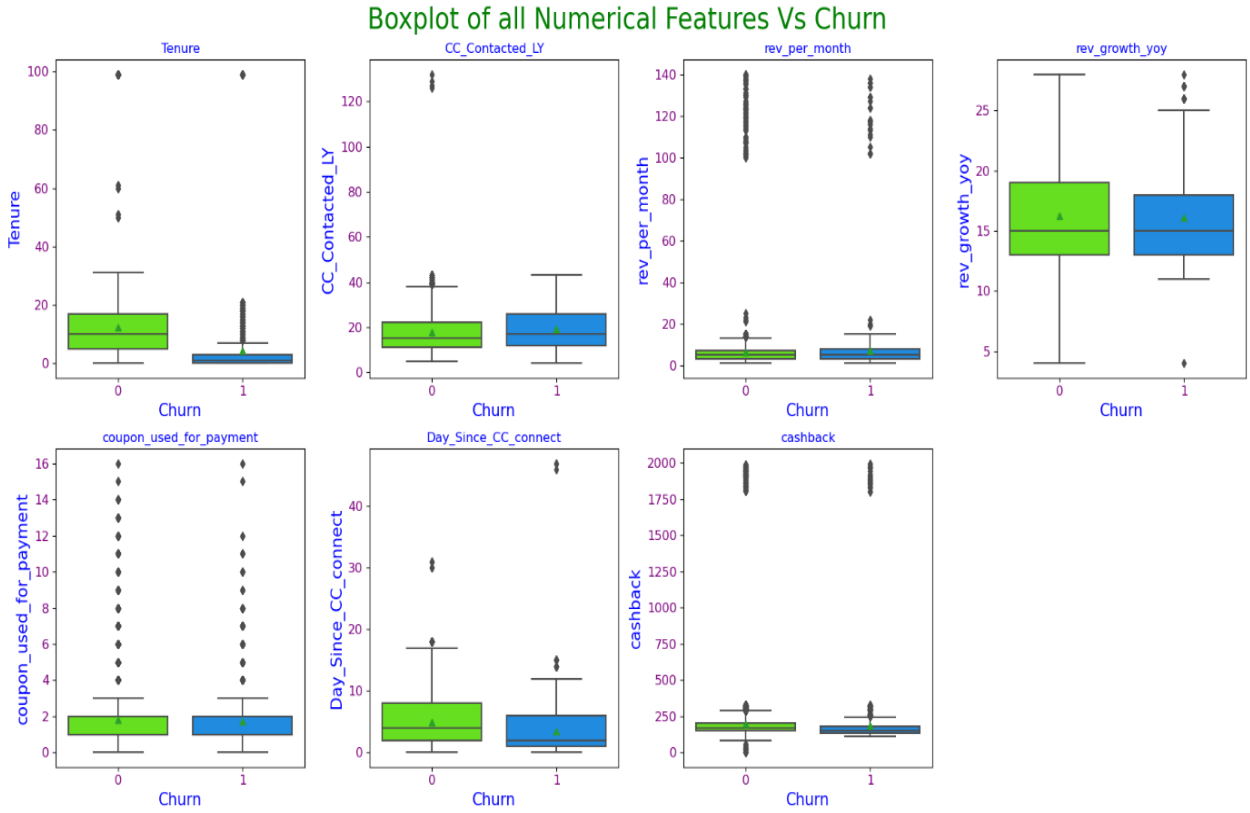


Fig no-7: Boxplots

Insights:

- It seems that when the customers contact the Customers care more their queries gets resolved due to which they don't churn and keep using the services more hence reducing the Churn rate.
- So, we can say Customer Care is also playing an Important role in Retaining Customers.
- We see that median value of the Tenure and Days_since_CC_connect for Churners is less compared to that of non-Churners.
- The Distribution of Coupon_used_for_payment, Cashback and rev_per_month is same for both Churners and Non-churners.
- There is no difference in median rev_growth_yoy between churners and non-churners.

❖ **Count plot of all Categorical Variables Vs Target Variables**

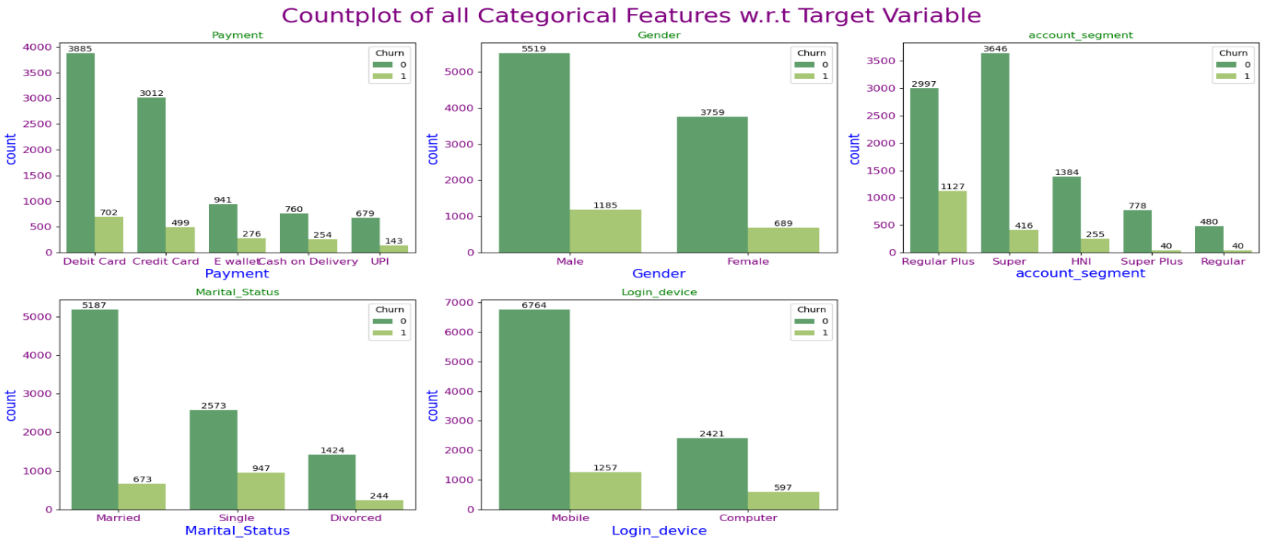


Fig no-8: Count plot w.r.t Target

Insights:

- The proportions of Churners are more of Male customers as compared to Female.
- Most of the churners make payment via Debit Card.
- Non-churners preferrable mode of payment are Debit cards and Credit Cards.
- Most of the churners belongs to Regular plus and Super account segment.
- Maximum non-Churners belongs to Super Segment followed by Regular plus.
- Most of the churners are single.
- Non-Churners are Maximum Married Couples.
- Most Preferred Login Device for both Churners and Non-Churners is Mobile since it's handy.

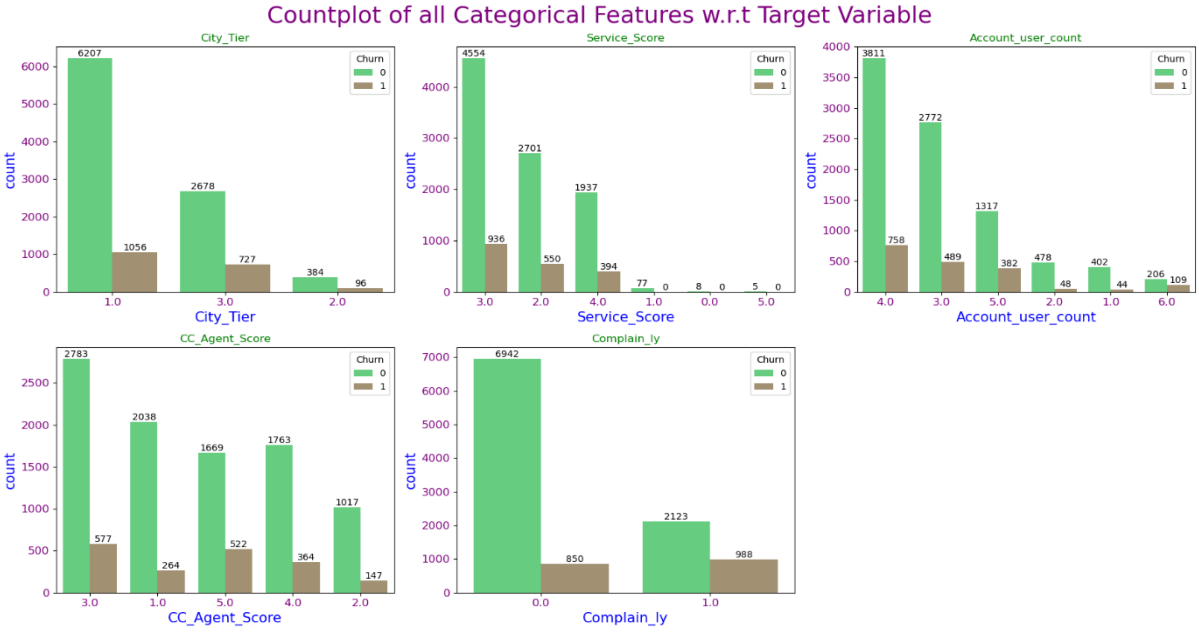


Fig no-9: Count plot w.r.t Target

Insights:

- Maximum non-churners belong to City Tier-1 followed by Tier-3.
- Churning rate of Customers from Tier-2 is very less, means customers tends continue the services seems they are more satisfied.
- Most of the Churners are from Tier-1.
- Maximum Service score given by customers is 3 by both churners and non-churners.
- Account tagged with 3,4 and 5 customers have more churning rate.
- Customers tagged with 2,1 and 6 accounts are mostly non-churners.
- Customers who have given an agent score of 3 and above show the Maximum churn rate.
- Maximum non-churners have given a Agent score of 3 and 1.
- Very few Churners have raised the complaints in the last 12 months. If they would have raised the complaint then may be their queries would have been resolved and they would not churn.
- It is evident from the fact that Customers who have raised the complaint maximum no of times are mostly non-churners. This shows that raising complaint have solved their issue and hence made them retain the use of service and decreasing churn Rate.

❖ Checking separation across the numerical variables:

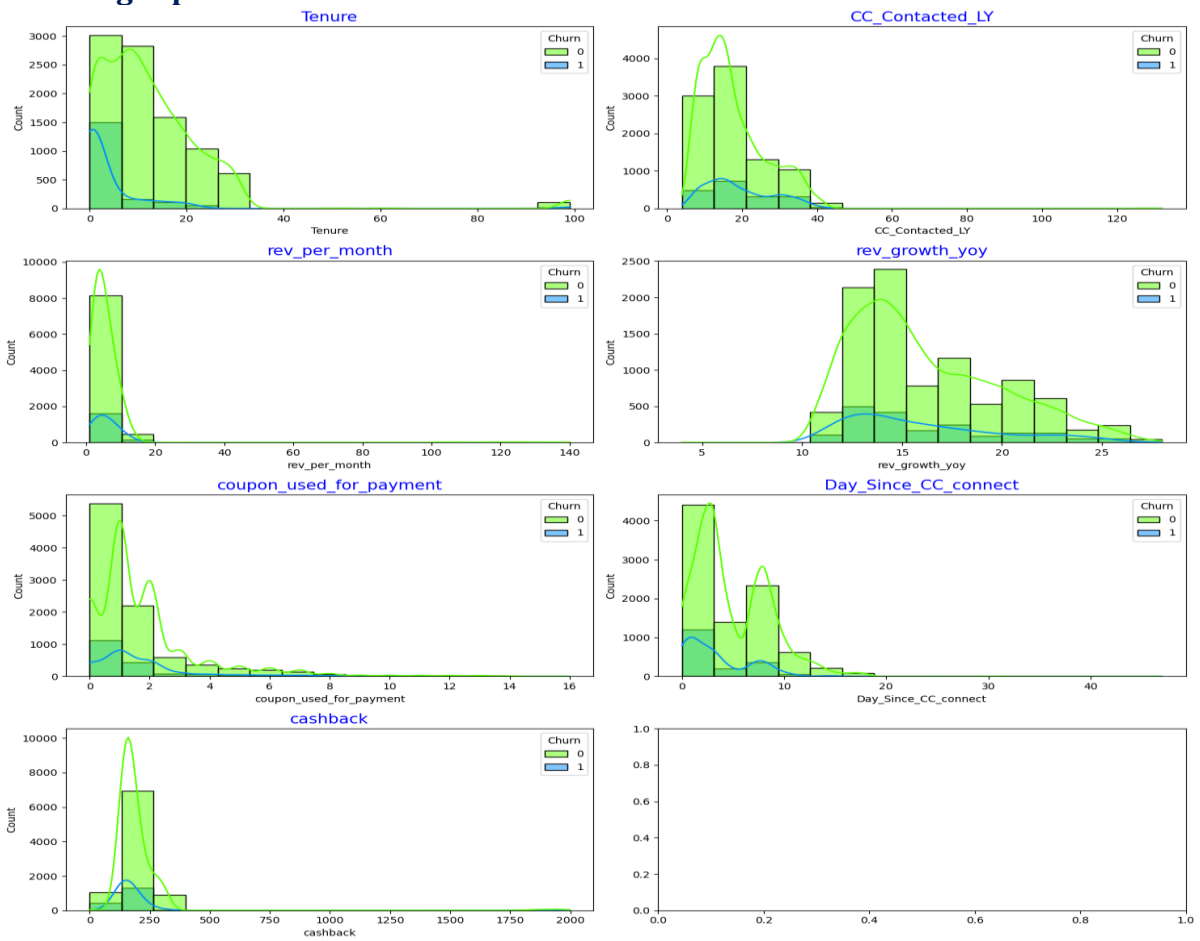


Fig no-10: Histograms

Insights:

- Most of the Distributions are Right Skewed.
- Customers having Short-term Tenure churns more.
- There are some Loyal customers also having Tenure more than 80 months.
- Maximum number of times Customer care was contacted is between 4 to 23 times.
- maximum customers use 1-2 coupons for payment.
- Most of the cashback are generated is around 0 to 300.
- Maximum cashbacks are generated by non-churners.

❖ Scatterplots:

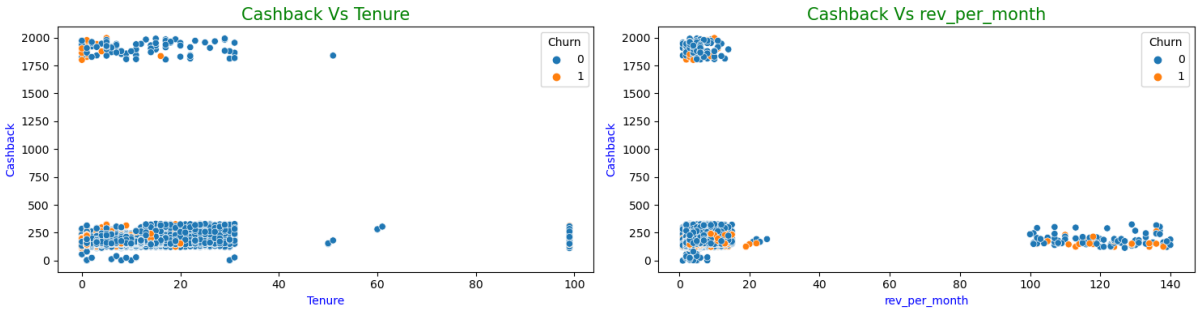


Fig no-11: Scatterplot

Insights:

- Most of the Cashback are for Tenure less than 35 months.
- Cashback Ranges between 0-280 and 1750-2000.
- Cashback of more than 1750 is mostly enjoyed by non-churners.

❖ Correlation Heatmap:

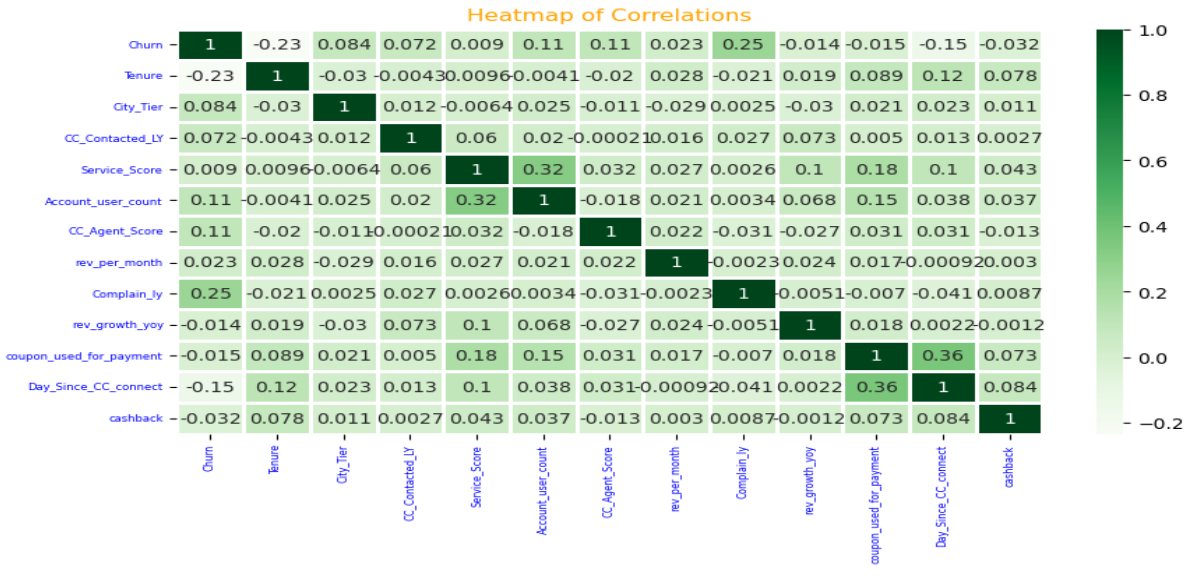


Fig no-12: Heatmap

Insights:

- Churn is negatively correlated with Tenure, having a correlation of -23%.
- The smaller the tenure value, the greater the churn rate.
- Correlation between Churn and Complain_ly is around 25%. The greater the complaint rate, the greater the churn rate.
- Service score has 32% correlation with Account user count and 18% correlation with Coupon used for payment.
- Coupon used for payment has 36% correlation with Days since CC connect, 18% with Service score and 15% with Account user count.
- There are no such features that are Highly Correlated with each other.
- There is no strong positive and Negative Correlations between the Predictors.
- There seems no Multicollinearity present in the data.

❖ Pair plot:

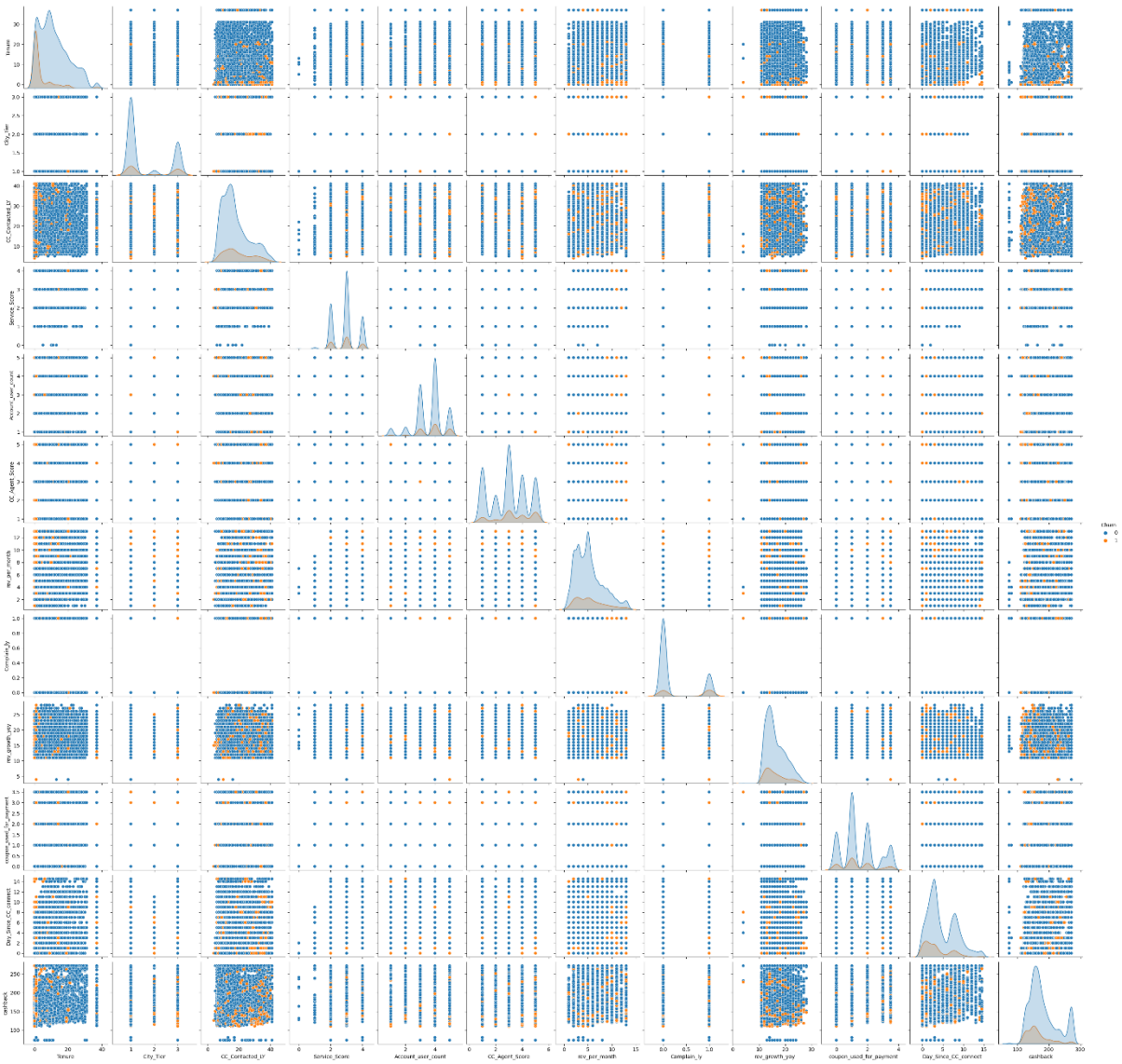


Fig no-13: Pair plot

Insights:

- Churners and non-churners seem overlapping each other in almost all the features.
- There is no Linear Pattern observed.
- Customers Churn more with Lowest Tenure.
- Tier-1 and Tier-3 Customers Churn Rate is more compared to Tier-2 cities customers.

Removal of unwanted variables (if applicable)

➢ Dropped column AccountID as it is not required for Analysis

Missing Value Treatment

```
Proportion of Missing Values in the folowing columns :

Churn                          0.000000
Tenure                         5.703820
City_Tier                      2.930403
CC_Contacted_LY                2.668760
Payment                        2.851910
Gender                         2.825746
Service_Score                  2.564103
Account_user_count            11.616954
account_segment                2.537938
CC_Agent_Score                 3.035060
Marital_Status                 5.546834
rev_per_month                 20.695971
Complain_ly                    9.340659
rev_growth_yoy                 0.078493
coupon_used_for_payment        0.078493
Day_Since_CC_connect           9.366824
cashback                      12.375720
Login_device                   5.782313
dtype: float64
```

- Since the proportion of Missing Values is less than 30%, so we will impute the missing values rather than dropping them.

❖ **Let's visually inspect the missing values in our data:**

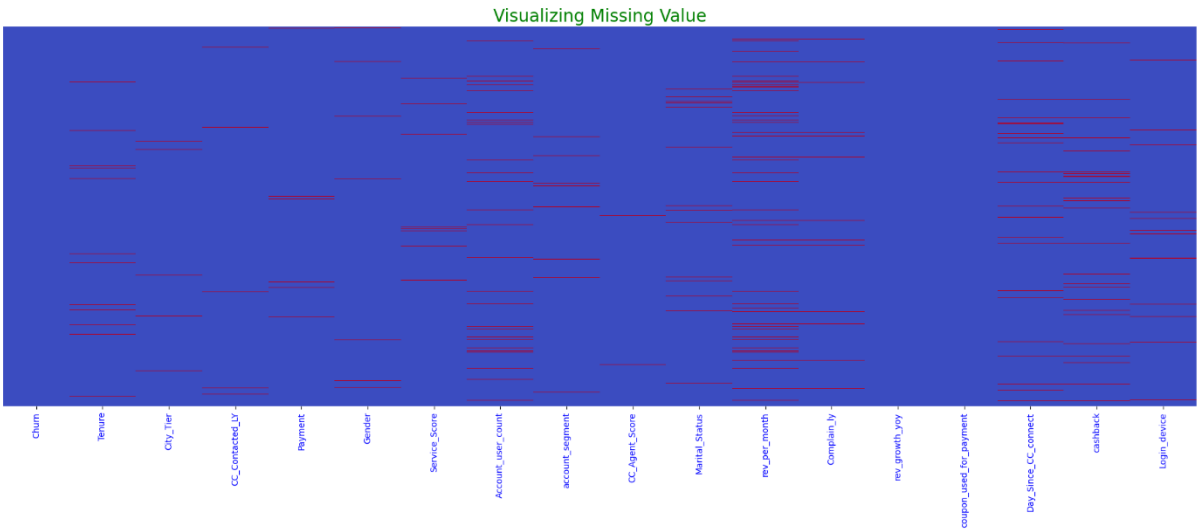


Fig no-14: Missing Value Visualization

❖ Let's Impute Categorical and Numerical Columns using Simple Imputer:

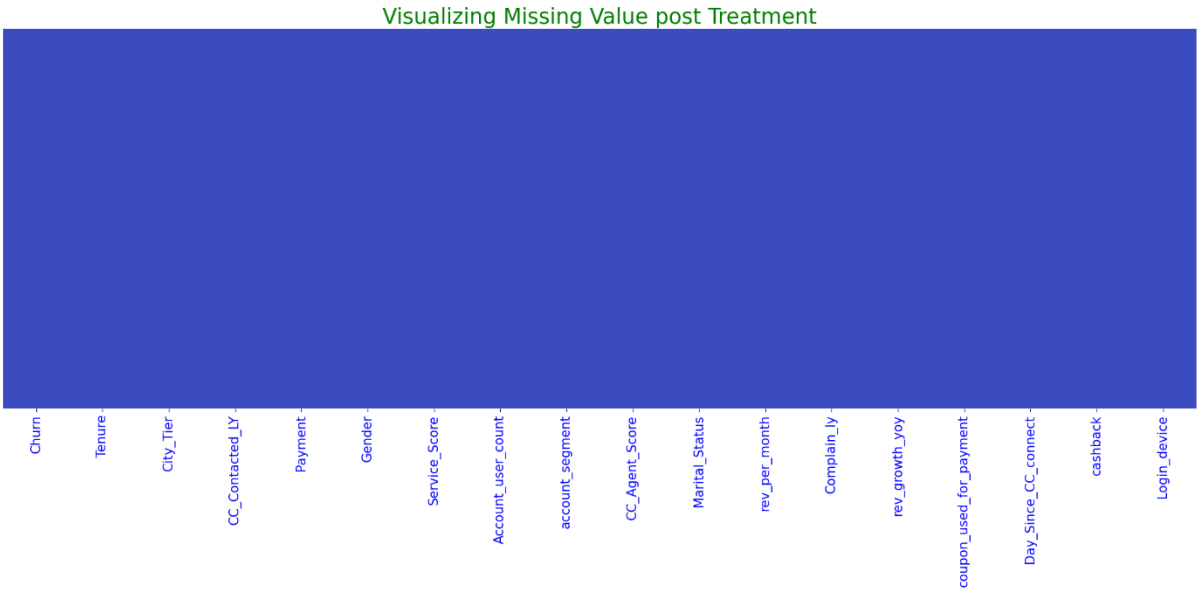


Fig no-15: Missing Value Visualization

➢ There are no Missing values Now.

Outlier Treatment

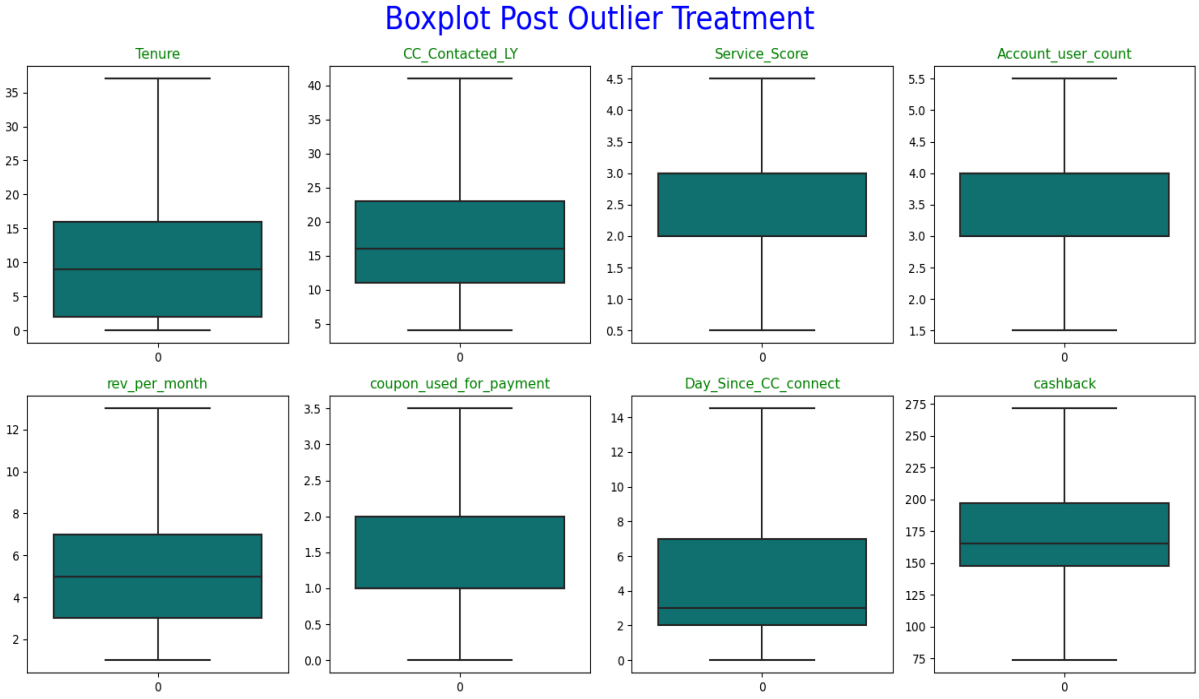


Fig no-16: Boxplot post Outlier Treatment

❖ Statistical Summary of the Customers who have Churned

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Churn	1896.0	NaN	NaN	NaN	1.0	0.0	1.0	1.0	1.0	1.0	1.0
Tenure	1896.0	NaN	NaN	NaN	3.660865	6.492737	0.0	0.0	1.0	4.0	37.0
City_Tier	1896.0	NaN	NaN	NaN	1.817511	0.957365	1.0	1.0	1.0	3.0	3.0
CC_Contacted_LY	1896.0	NaN	NaN	NaN	19.248945	8.856001	4.0	12.0	17.0	26.0	41.0
Payment	1896	5	Debit Card	724	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	1896	2	Male	1207	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Service_Score	1896.0	NaN	NaN	NaN	2.917722	0.700985	2.0	2.0	3.0	3.0	4.0
Account_user_count	1896.0	NaN	NaN	NaN	3.880802	0.90179	1.0	3.0	4.0	5.0	5.0
account_segment	1896	5	Regular Plus	1145	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CC_Agent_Score	1896.0	NaN	NaN	NaN	3.386603	1.333874	1.0	3.0	3.0	5.0	5.0
Marital_Status	1896	3	Single	947	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rev_per_month	1896.0	NaN	NaN	NaN	5.506857	3.023209	1.0	3.0	5.0	7.0	13.0
Complain_ly	1896.0	NaN	NaN	NaN	0.521097	0.499687	0.0	0.0	1.0	1.0	1.0
rev_growth_yoy	1896.0	NaN	NaN	NaN	16.077532	3.862519	4.0	13.0	15.0	18.0	28.0
coupon_used_for_payment	1896.0	NaN	NaN	NaN	1.429325	1.084681	0.0	1.0	1.0	2.0	3.5
Day_Since_CC_connect	1896.0	NaN	NaN	NaN	3.349156	3.139531	0.0	1.0	3.0	5.0	14.5
cashback	1896.0	NaN	NaN	NaN	162.960654	36.801456	110.09	136.405	153.46	176.645	271.44
Login_device	1896	2	Mobile	1299	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Useful Insights:

- **Total Churners = 1896.**
- **Average Tenure almost 4 months, Maximum Tenure is of 37 months.**
- **Tier 1 & 3 customers churns more.**
- **Most of the churners are Male and Singles.**
- **Preferred payment mode is Debit Cards holding regular Plus account and most likely Login via Mobile.**
- **Avg Cashback generated is around Rs 163 and Max cashback generated is around Rs 271.**

Data Visualization using Segmentation

Let's Segment the data based on City Tiers, Payment Mode and Genders Respectively and will Draw useful insights if any.

Segment on the basis of City Tiers

There are 3 city tiers mentioned in the dataset. Generally, tier 1 cities are considered as the major metro cities where the people tend to use more DTH Services. So, accordingly, can we say that tier 1 city customers tend to generate more avg revenue as compared to tier 2 and tier 3 city customers. Let's visualize this and find it it's True or not?

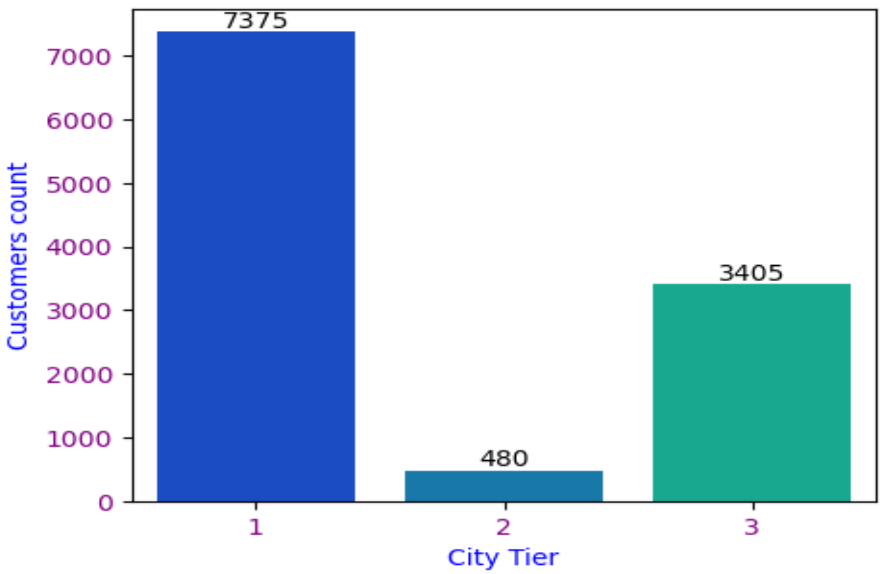


Fig no-17: Count plot of City Tier

Insights:

- Count of customers are more in Tier-1 followed by Tier-3.
- Ver less customers belong to Tier-2.

❖ City Tier Vs Monthly Average Revenue:

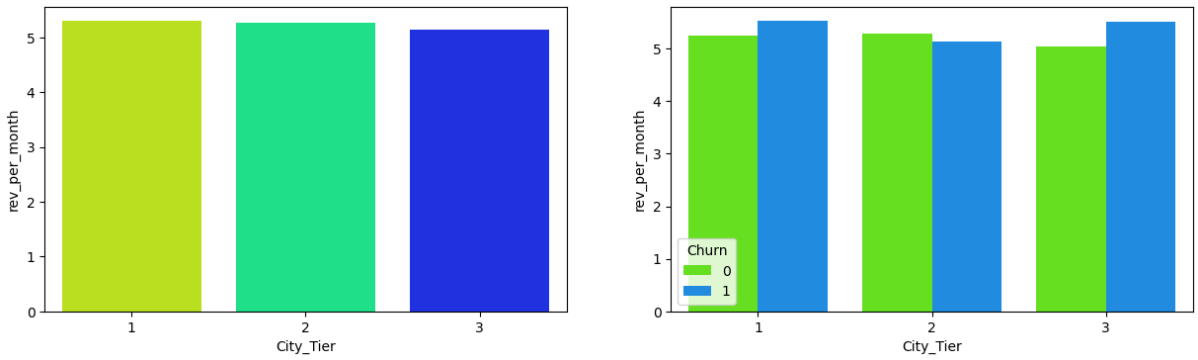


Fig no-18: City Tier Vs Revenue

Insights:

- We can see that Avg Revenue Generation is same across all the City Tiers.
- Though the Customers using the Services in Tier-2 is very less as compared to Tier-1 & 3, But the Avg Revenue generated per month by the account is same across all the Tiers.
- It seems Customers are more satisfied in Tier-2 hence they Retain to use the Services, generating greater revenue.
- As we can see from the plot, City_Tier 1 & 2 has a slightly higher mean of Avg Revenue per month as compared to City_Tier 3 which are more or less the same.

So, our assumption here that tier 1 city customers tend to generate more avg revenue cannot be validated looking at the plot.

Payment Mode

There are different payment modes (CC, DC, COD, E-wallet & UPI). Depending on the city tiers, Let's Visualize and find the preferred payment modes used by the customers

Payment City_Tier	Cash on Delivery	Credit Card	Debit Card	E wallet	UPI
1	732	2740	3407	12	484
2	31	100	123	0	226
3	251	671	1166	1205	112

- **46.2 % of customers prefer Debit card as preferred payment mode in Tier 1 cities.**
- **47.08 % of customers prefer UPI as preferred payment mode in Tier 2 cities.**
- **35.39 % of customers prefer E wallet as preferred payment mode in Tier 3 cities.**

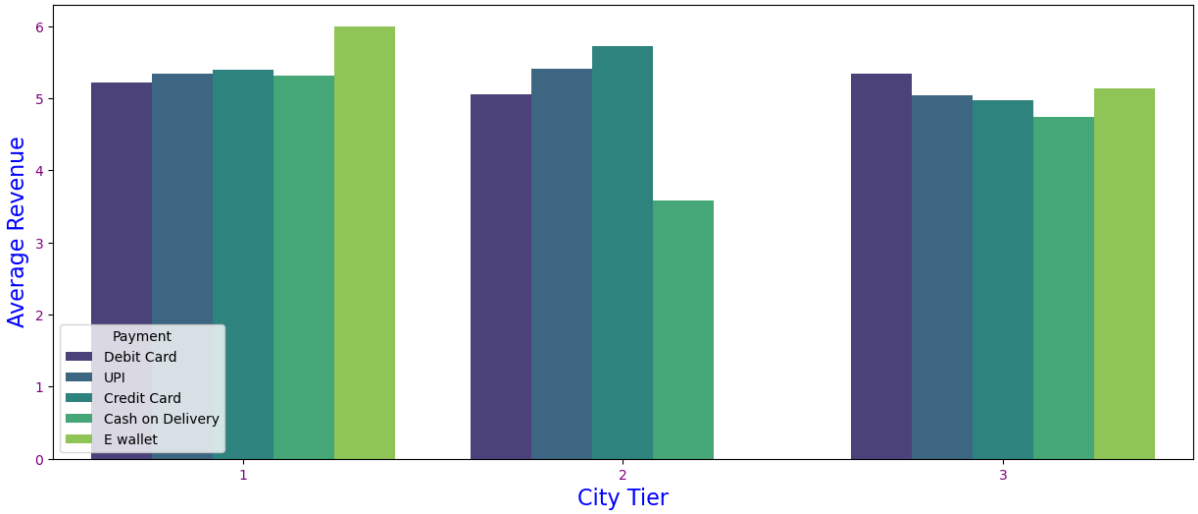


Fig no-19: City Tier Vs Payment mode

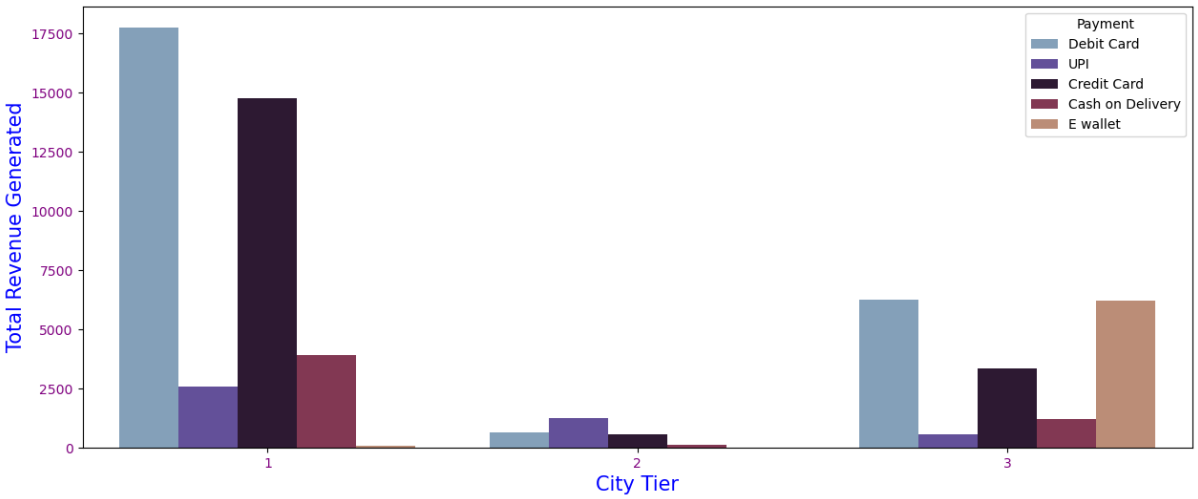


Fig no-20: City Tier Vs Payment mode Vs Total Revenue

Insights:

- As we can see, E wallet is used only by the Tier 1 & Tier 3 cities.
- Tier-2 Customers don't prefer using E-wallet as payment mode.
- Most of the customers prefer using E wallet and Debit Card in tier 3 cities.
- Almost 35% of the tier 3 city customers prefers E wallet as their payment mode, most revenue will be generated from such customers from tier 3 cities.
- Almost 46% of the tier 1 city customers prefers Debit Card as the payment mode, so most revenue will be generated from such customers from tier 1 cities.
- Almost 47% of the tier 2 city customers prefers UPI as the payment mode, so most revenue will be generated from such customers from tier 2 cities.

Segment on the Basis of Gender

❖ **Bar plot for Male Vs Female segmentation:**

- Male Count: 6812
- Female Count: 4448

❖ **Visualization for Male Vs Female segmentation:**

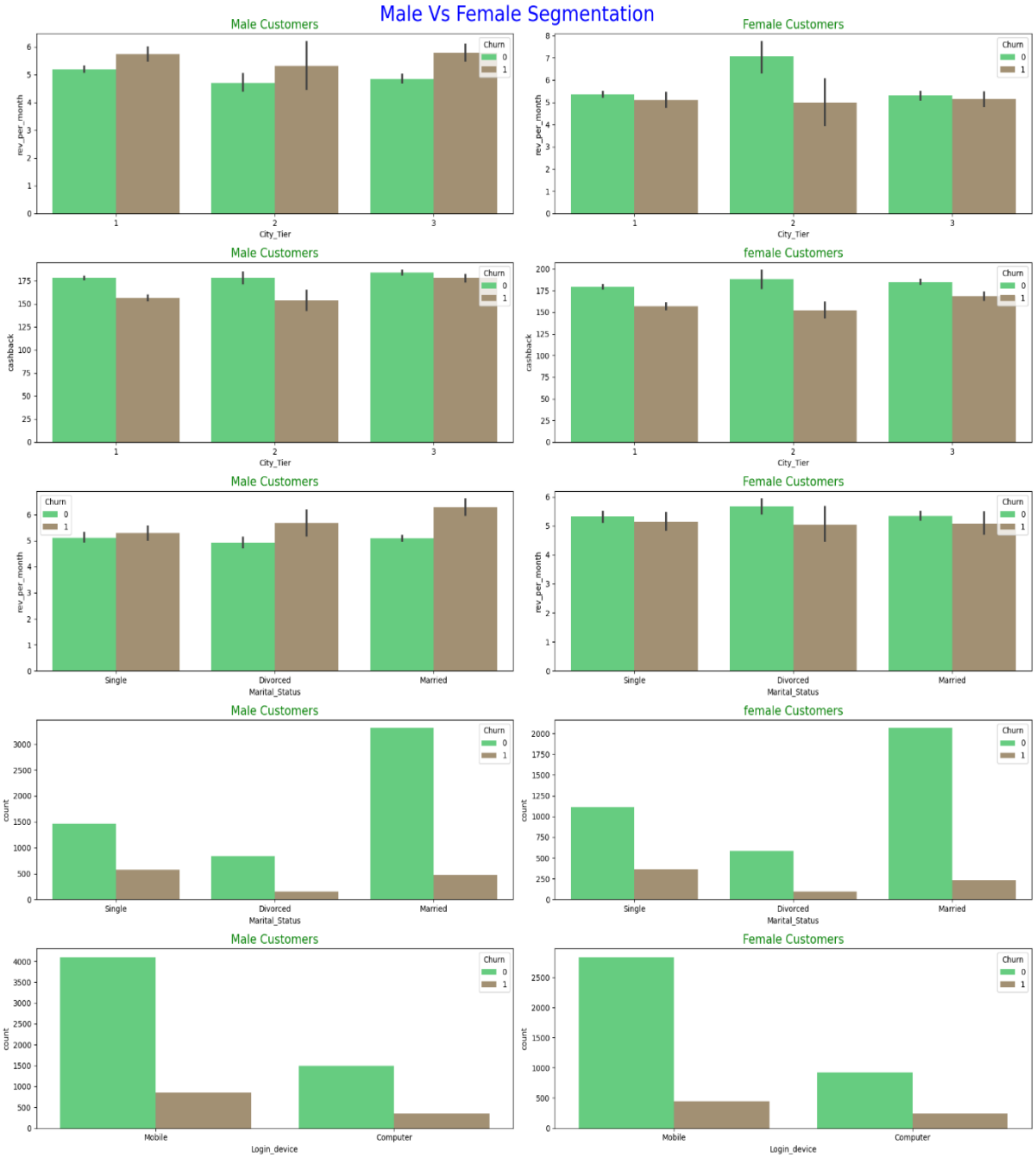


Fig no-21: Male Vs Female Segmentation

Insights:

- In Tier-1 & Tier-2, avg monthly revenue is more generated by Male customer than Females and in Tier-3 by Females.
- Cashback ratio is same for male & female in all Tier and across Marital status.
- Avg revenue generated by Singles Male and Females are almost same in all Tiers.
- Revenue generated by Divorced Females are slightly more than male.
- Count of Marital status is same across Genders.
- Mobile is the most preferred Login device among Male and Females.

❖ **Revenue Generated on the basis of Spend (Male Customers):**

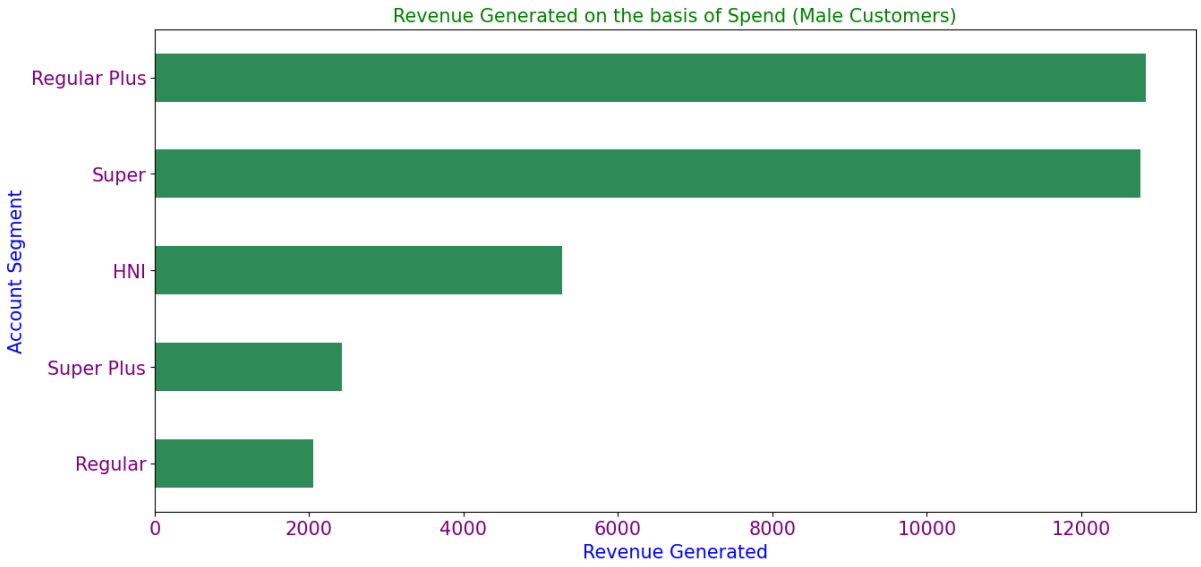


Fig no-22: Bar plot male customer

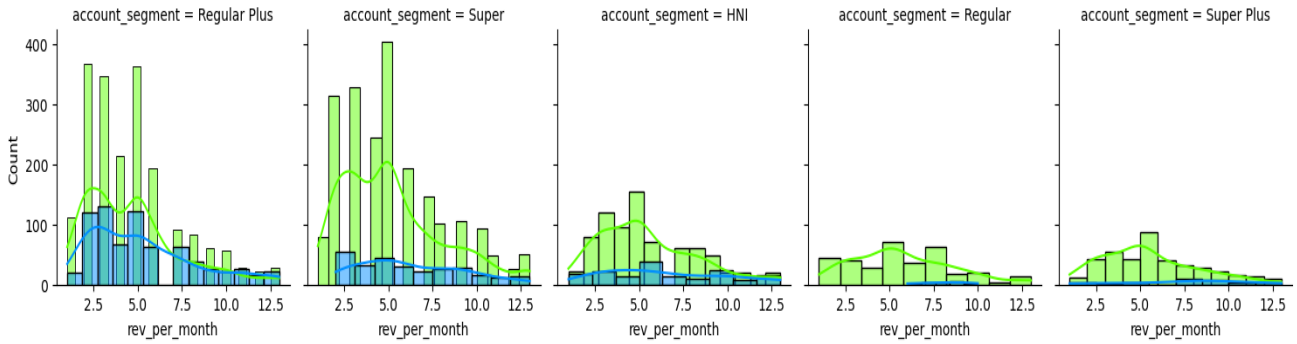


Fig no-23: Facet grid

Insights:

- **Maximum Revenue is generated by Regular plus and Super account holder by male customers.**

❖ **Revenue Generated on the basis of Spend (Male Customers):**

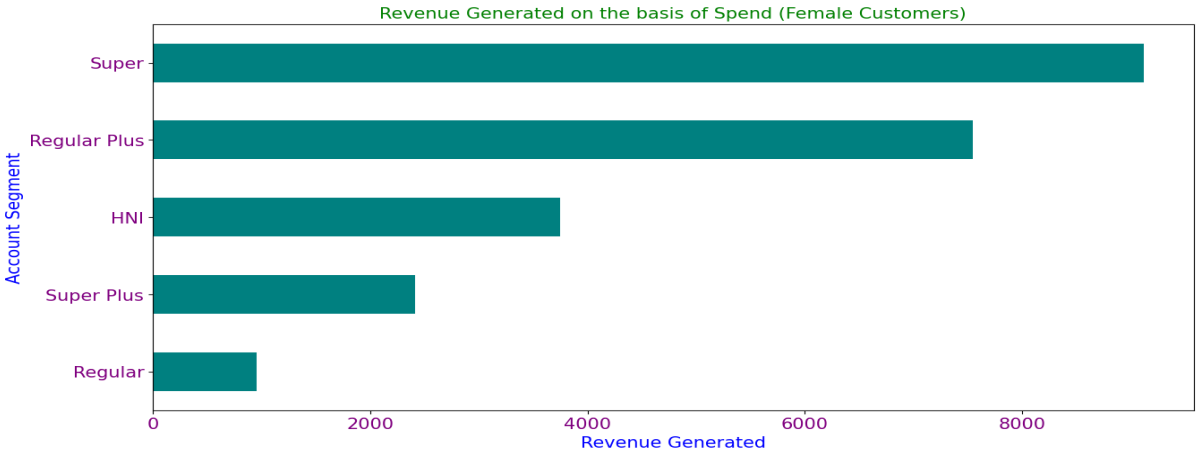


Fig no-24: Bar plot Female customer

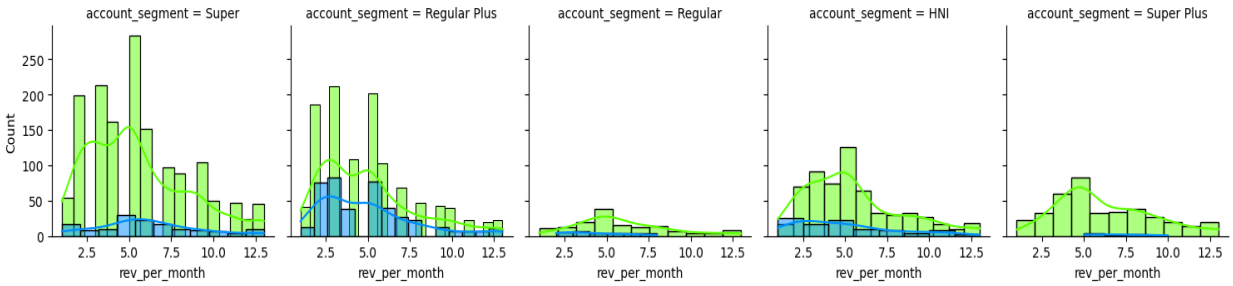


Fig no-25: Facet grid

Insights:

- Maximum revenue is generated by Super account holder followed by Regular Plus Female customers.

❖ Revenue Generated on the basis of Account_user_count (Male Customers):

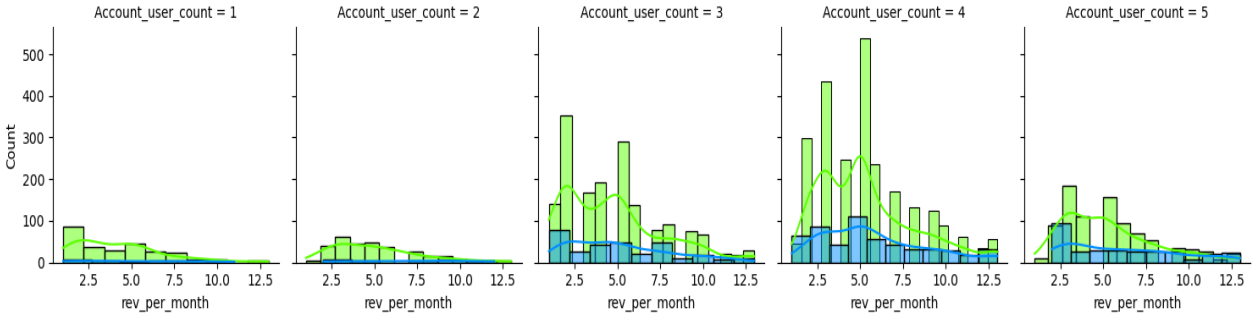


Fig no-26: Facet grid

Insights:

- Account tagged with more users generate more Revenue by Male customers.

❖ **Revenue Generated on the basis of Account_user_count (Female Customers):**

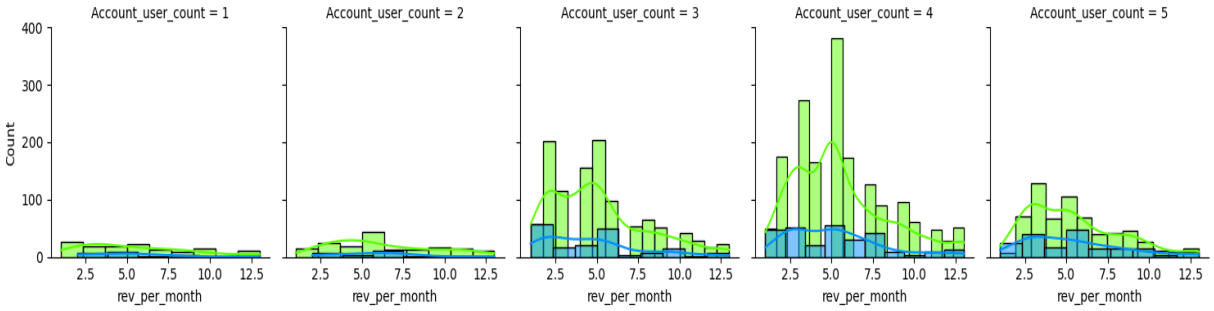


Fig no-27: Facet grid

Insights:

- Account tagged with more users generate more Revenue by Female customers.

Data Encoding

➢ **Data Encoding is done using Label Encoder for Categorical Variable.**

Feature Transformation

➢ **Predictors are Scaled using Min Max Scalar**

	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Status
0	0.108108	1.0	0.054054	0.50	0.0	0.75	0.50	0.75	0.25	1.0
1	0.000000	0.0	0.108108	1.00	1.0	0.75	0.75	0.50	0.50	1.0
2	0.000000	0.0	0.702703	0.50	1.0	0.50	0.75	0.50	0.50	1.0
3	0.000000	1.0	0.297297	0.50	1.0	0.50	0.75	0.75	1.00	1.0
4	0.000000	0.0	0.216216	0.25	1.0	0.50	0.50	0.50	1.00	1.0

rev_per_month	Complain_ly	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback	Login_device
0.666667	1.0	0.291667	0.285714	0.344828	0.435907	1.0
0.500000	1.0	0.458333	0.000000	0.000000	0.238466	1.0
0.416667	1.0	0.416667	0.000000	0.206897	0.462819	1.0
0.583333	0.0	0.791667	0.000000	0.206897	0.305089	1.0
0.166667	0.0	0.291667	0.285714	0.206897	0.282477	1.0

Addition of new variables (if required).

➢ **Addition of New variables is required in this problem.**

Clustering

- we will be performing clustering via KMeans clustering.
- Plotting Elbow plot (up to n=10) and identifying optimum number of clusters for k-means algorithm.

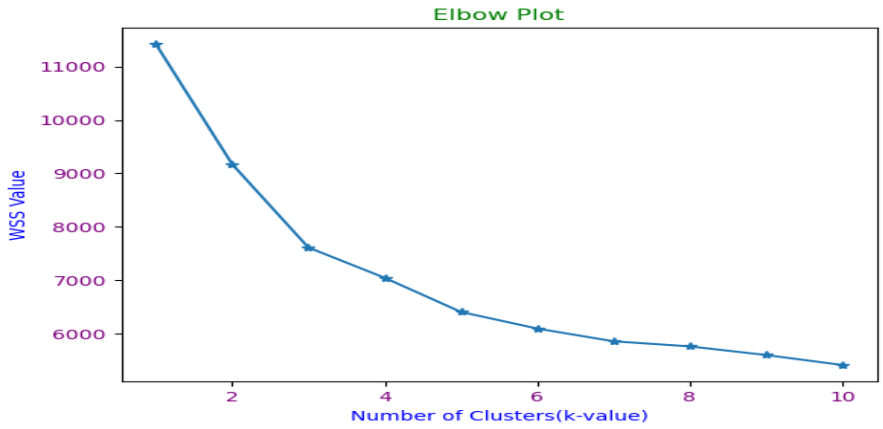


Fig no-28: Elbow plot

- From the above Elbow plot and WSS (within sum of squares) value we can choose optimum number of clusters as 3.
- Up to K=3 the drop in WSS value is significant.
- Beyond K=3, the drop is not significant.

❖ Silhouette scores for up to 10 clusters to identify optimum number of clusters.

```
The Silhouette Score for 2 clusters is 0.21014
The Silhouette Score for 3 clusters is 0.23507
The Silhouette Score for 4 clusters is 0.15923
The Silhouette Score for 5 clusters is 0.18869
The Silhouette Score for 6 clusters is 0.17312
The Silhouette Score for 7 clusters is 0.15749
The Silhouette Score for 8 clusters is 0.14826
The Silhouette Score for 9 clusters is 0.14614
The Silhouette Score for 10 clusters is 0.14788
```

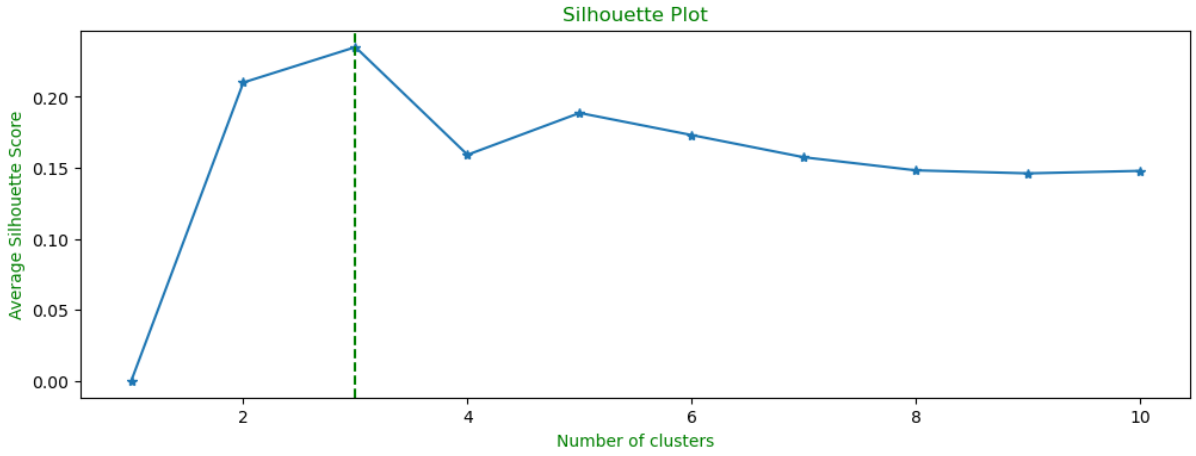


Fig no-29: Silhouette plot

- Optimal number of clusters =3.

❖ Adding Clusters to the Churn Data frame for further Visualization:

	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Status
0	1	4.0	3	6.0	Debit Card	Female	3	3	Super	2	Single
1	1	0.0	1	8.0	UPI	Male	3	4	Regular Plus	3	Single
2	1	0.0	1	30.0	Debit Card	Male	2	4	Regular Plus	3	Single
3	1	0.0	3	15.0	Debit Card	Male	2	4	Super	5	Single
4	1	0.0	1	12.0	Credit Card	Male	2	3	Regular Plus	5	Single

rev_per_month	Complain_ly	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback	Login_device	Clusters
9.0	1	11.0	1.0	5.0	159.93	Mobile	2
7.0	1	15.0	0.0	0.0	120.90	Mobile	2
6.0	1	14.0	0.0	3.0	165.25	Mobile	2
8.0	0	23.0	0.0	3.0	134.07	Mobile	1
3.0	0	11.0	1.0	3.0	129.60	Mobile	0

❖ Grouping data on the basis of Clusters and Login Device:

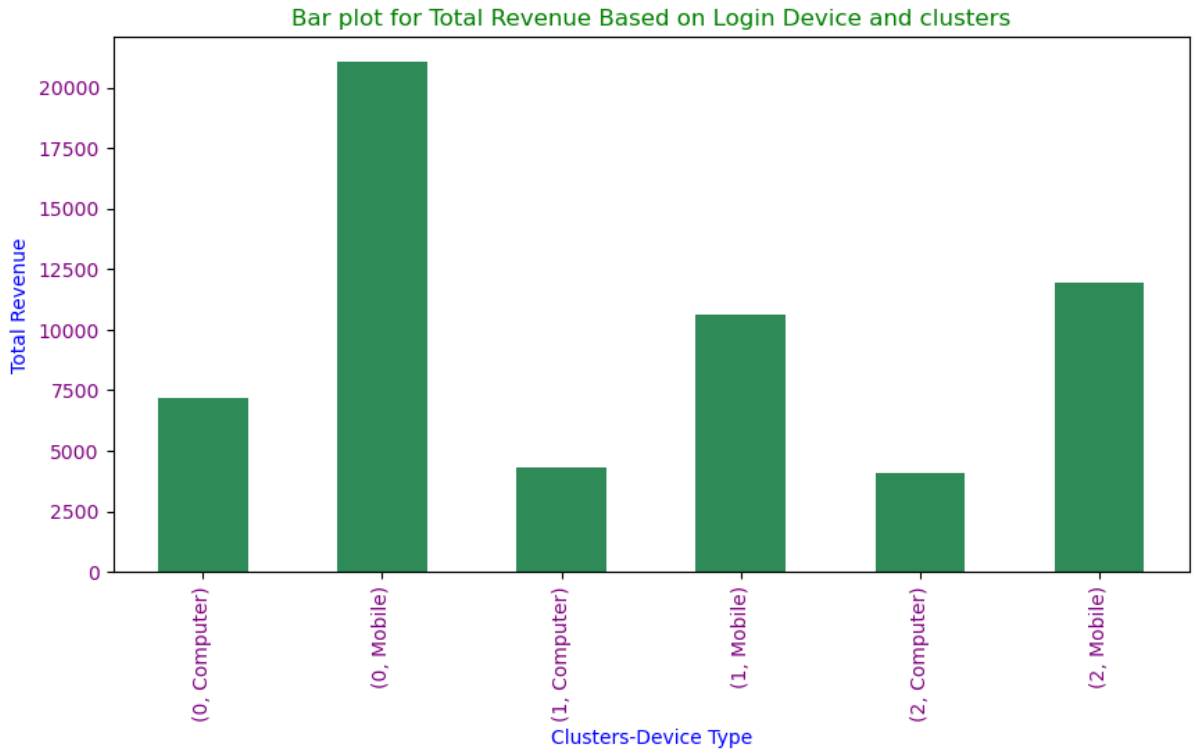


Fig no-30: Cluster-Login Device Vs Revenue

- Total Revenue generated is Maximum for Cluster 0 followed by cluster 2, through Login Device Mobile.
- Also, For Login Device Computer, Maximum Revenue is generated by Cluster 0.

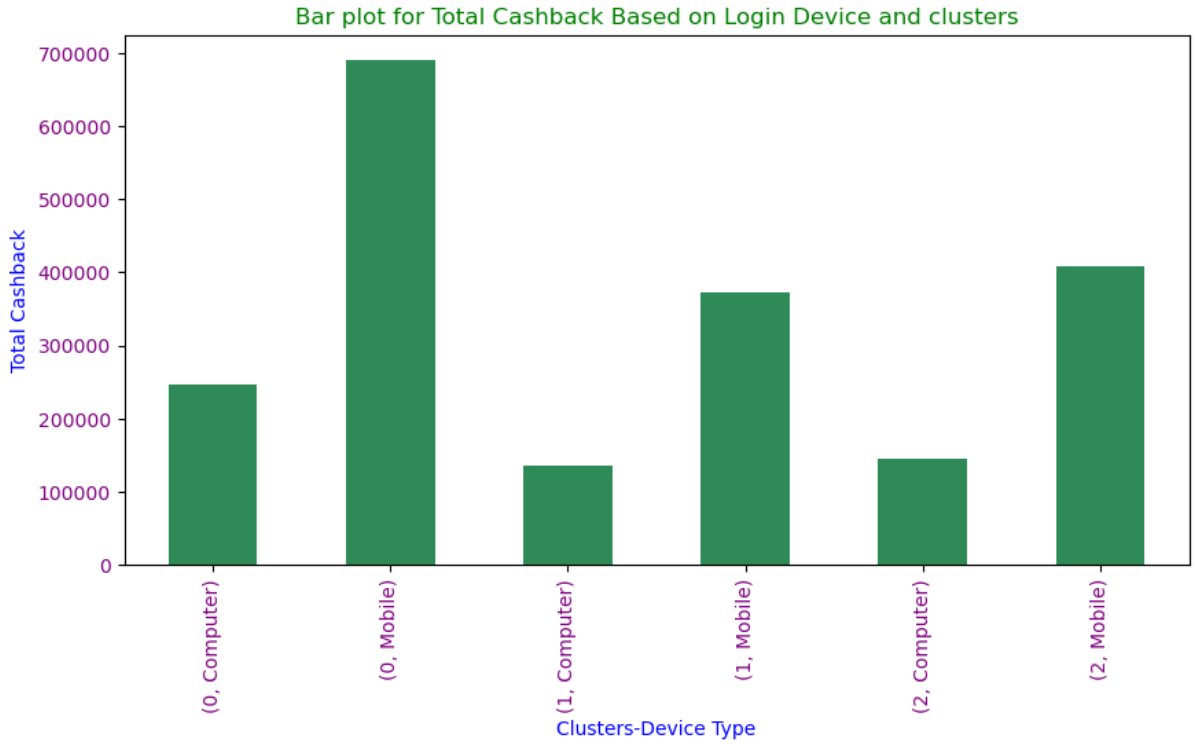


Fig no-31: Cluster-Login Device Vs Cashback

- Total Cashback is Maximum for Cluster 0 followed by cluster 2, through Login Device Mobile.
- Also, For Login Device Computer, Maximum Cashback is generated by Cluster 0.

Segmenting the Dataset based on Clusters

❖ Let's Segment the Data based on Clusters Formed

- Creating Three datasets based on Clusters formed for further Visualization.

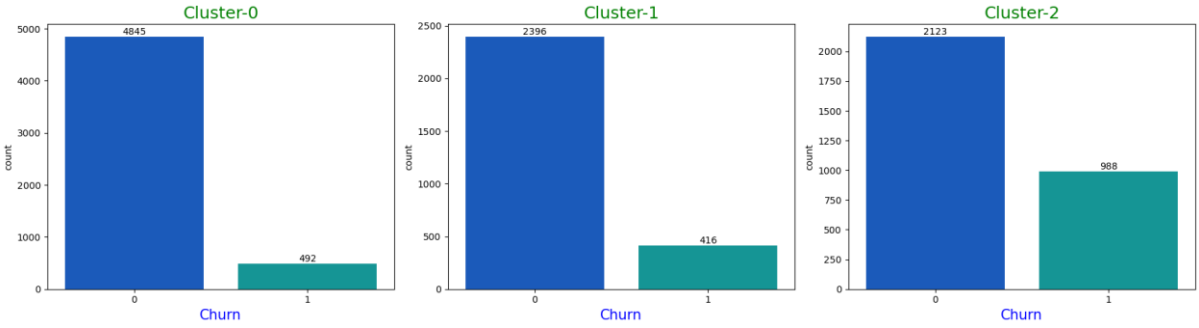


Fig no-32: Count plot Churn Vs Clusters

Insights:

- Churners are more in Cluster-2 as compared to other Clusters.

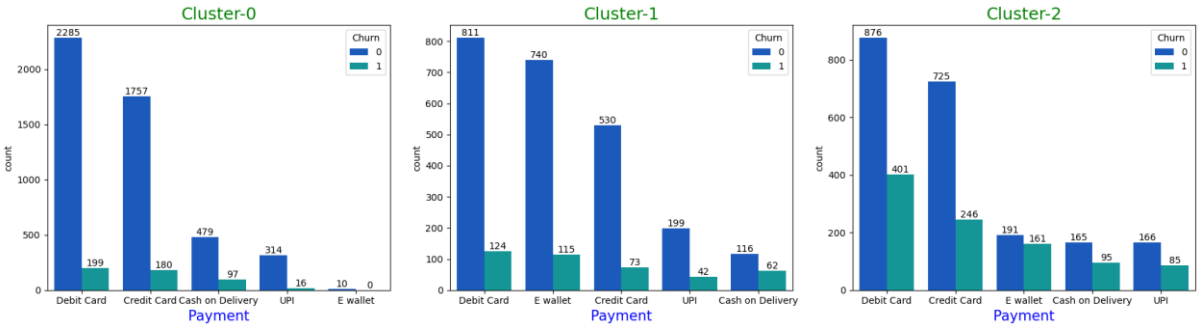


Fig no-33: Count plot Payment Vs Clusters

Insights:

- E-wallet payment mode is mostly used by Cluster-1 customers as compared to other clusters.
- Debit Card payment mode is used by all three clusters customer.
- E-wallet is Rarely used in cluster-0.
- Maximum churners use Debit card as payment mode across clusters.

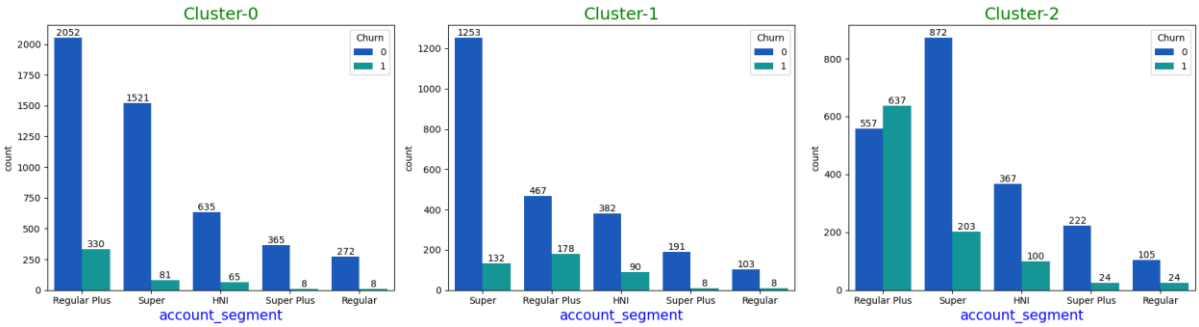


Fig no-34: Count plot account segment Vs Clusters

Insights:

- Super account holder is more in Cluster-1 followed by Cluster-2.
- Maximum Regular Plus account holders belongs to Cluster-0 and Cluster-2.
- In cluster-2 more churners belong to Regular plus account.
- More Churners belong to Regular plus account.

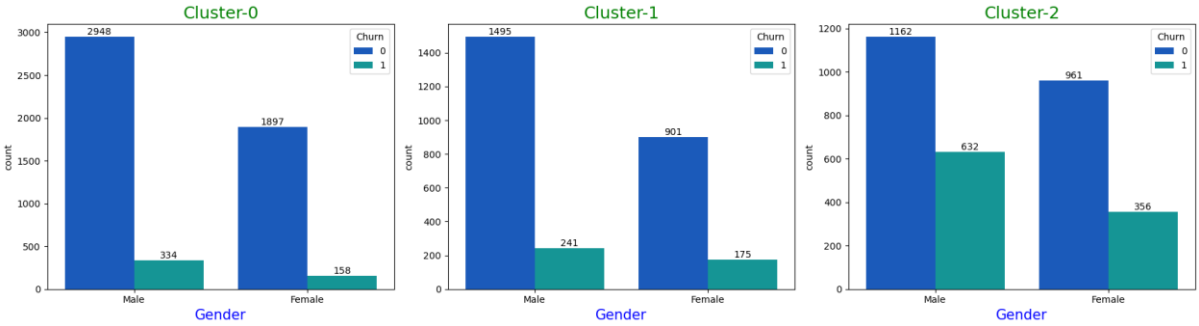


Fig no-35: Count plot Gender Vs Clusters

Insights:

- Females are slightly more in Cluster-2 as compared to other clusters.
- Gender Ratio is almost same in all the Clusters.
- Male Churners are more as compared to Females across clusters.

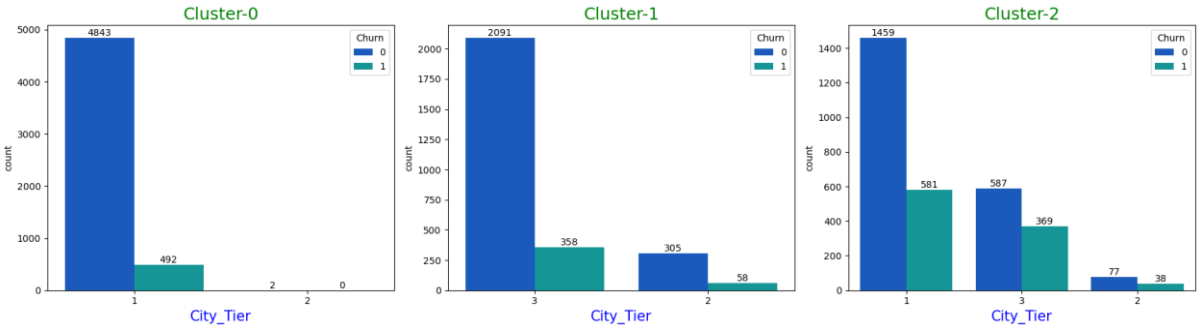


Fig no-36: Count plot City Tier Vs Clusters

Insights:

- Cluster-0 have customers belonging to mostly Tier-1 & very less from Tier-2. There are no Tier-3 customers in cluster-0.
- Cluster-1 have customers belonging to mostly Tier-3 & very less from Tier-2. There are no Tier-1 customers in cluster-1.
- In Cluster-2 maximum customers belong to Tier-1 followed by Tier-1 and then Tier-2.
- More churners belong to cluster-2.

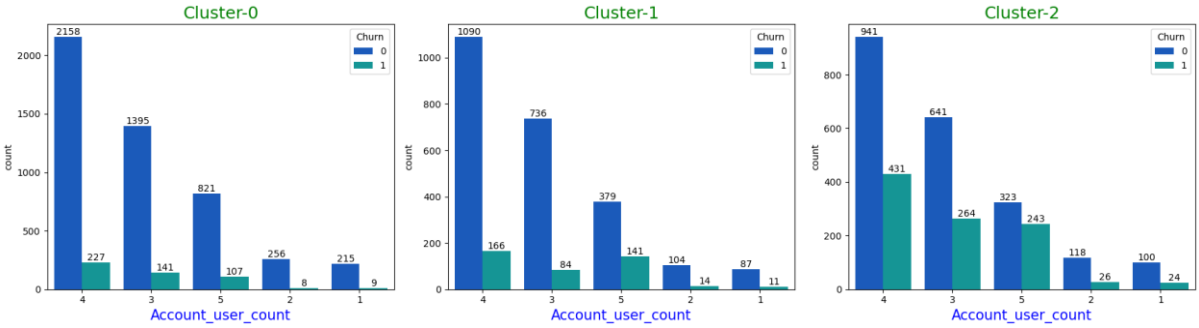


Fig no-37: Count plot Account user count Vs Clusters

Insights:

- Ratio of user tagged with single account is same across all the Clusters.
- More churners belong to account tagged with 3 or 4 users.

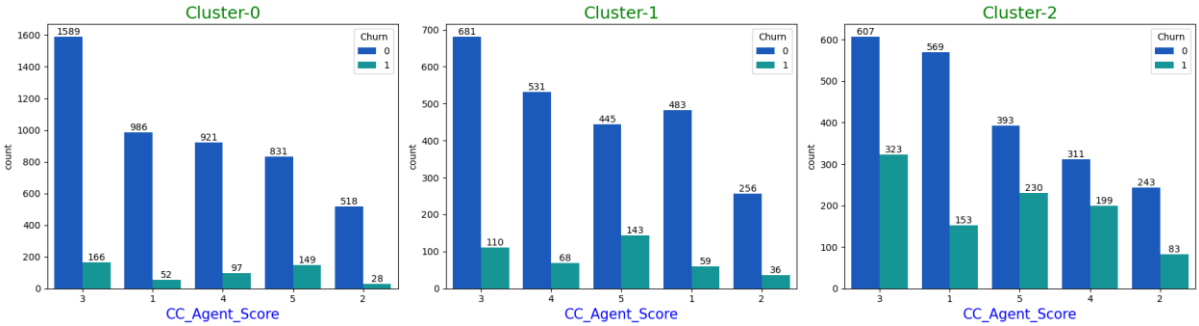


Fig no-38: Count plot CC Agent score Vs Clusters

Insights:

- In cluster-0,1 and 2, satisfaction score given by maximum customers is 3.

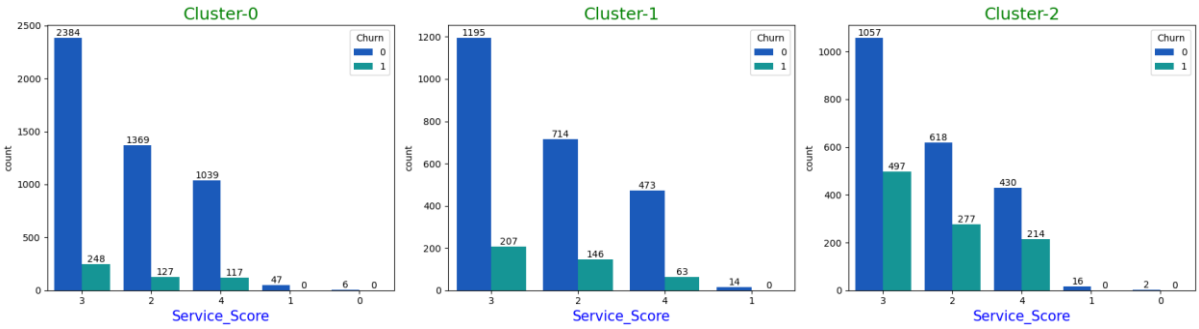


Fig no-39: Count plot Service score Vs Clusters

Insights:

- In cluster-1, no customers have given a score of 0.
- In all the Clusters maximum customers have given a score of 3,2 and 4 respectively.

❖ Tenure Vs Clusters:

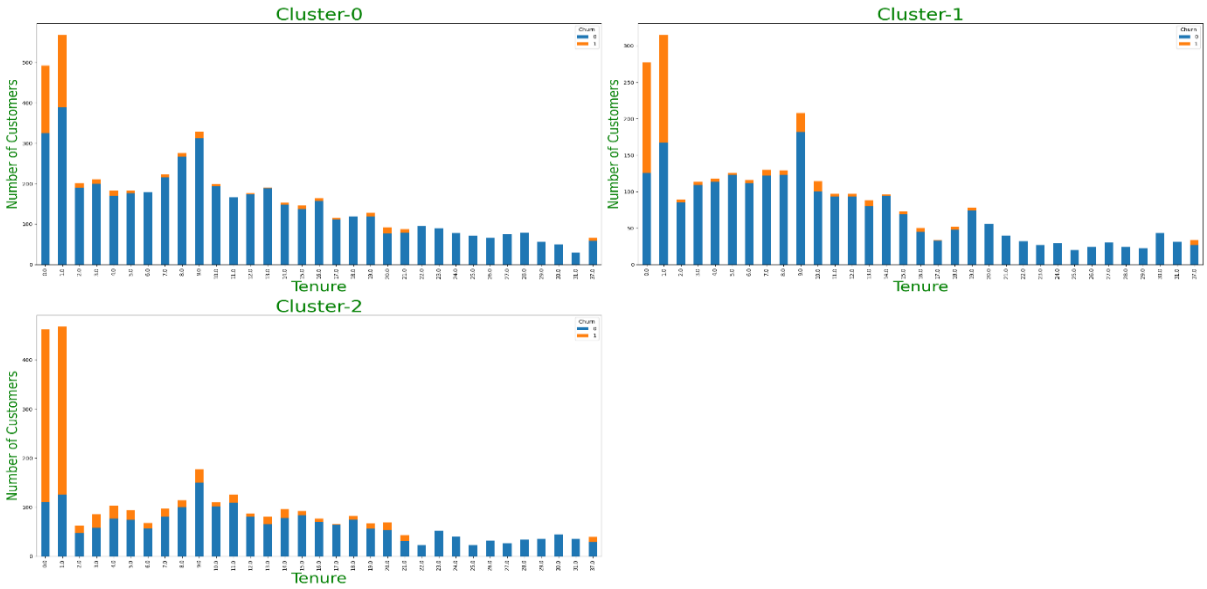


Fig no-40: Tenure Vs Clusters

Insights:

- It is observed that the newly joined customers having low Tenure Churns most, so more concern should be given to the newly joined customers and their cause of churning should be figured out.
- Most of the customers who Churns, customers with less than 2 tenures are classified as new customers.

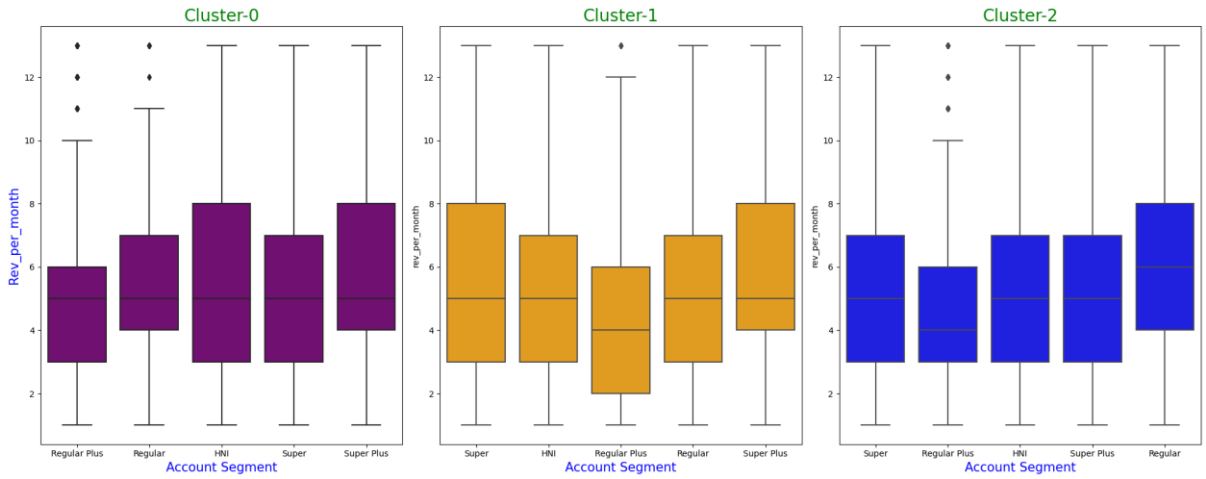


Fig no-41: Boxplot Account segment Vs Clusters

Insights:

- Maximum revenue is generated by HNI accounts in cluster 0.
- In cluster-1, Super, HNI and Super plus generate maximum revenue.
- In cluster-2, Regular accounts generate max Revenue.

Feature Selection

We will use Variance Inflation Factor to Select the Important features using the VIF value.

- ➢ **We will consider features having VIF < 5**
- ➢ **Features having VIF > 5 are removed from the datasets.**
- ➢ **We removed 15 Features Considering VIF > 5.**

	Features	VIF
5	Service_Score	17.592656
12	rev_growth_yoy	10.103968
6	Account_user_count	9.759654
15	cashback	9.384768
7	account_segment	4.447725
3	Payment	4.134365
9	Marital_Status	3.825733
2	CC_Contacted_LY	3.581888
16	Login_device	3.554259
13	coupon_used_for_payment	3.523403
10	rev_per_month	3.465482
14	Day_Since_CC_connect	3.383455
8	CC_Agent_Score	3.135966
0	Tenure	2.991562
4	Gender	2.435963
1	City_Tier	1.630030
11	Complain_ly	1.379092

- ➢ **We are left with 13 Features now and will Build the Models with these Features.**

	Features	VIF
5	account_segment	4.132428
3	Payment	3.925740
7	Marital_Status	3.477064
2	CC_Contacted_LY	3.339777
12	Login_device	3.338953
8	rev_per_month	3.232364
10	coupon_used_for_payment	3.118444
11	Day_Since_CC_connect	3.080066
6	CC_Agent_Score	2.996977
0	Tenure	2.495054
4	Gender	2.335222
1	City_Tier	1.610286
9	Complain_ly	1.362515

Business insights from EDA

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

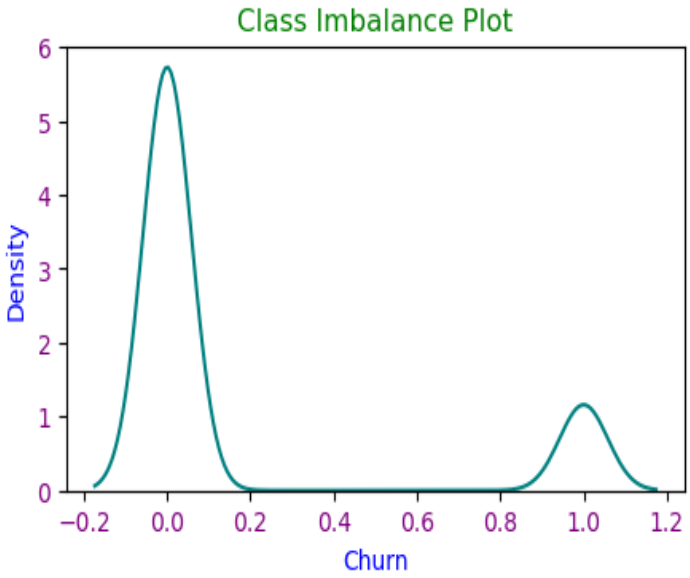


Fig no-42: Class imbalanced plot

- Here the distribution of Target variable is Skewed.
- In the Target feature 'Churn' there is a significant data imbalance, with a Majority class of 83.16% and a Minority class of 16.83%.
- To handle this class imbalance, an Up-sampling process is carried out where synthetically generated data points corresponding to minority class are injected into the datasets.
- We can use SMOTE technique for balancing the Data.

4.b) Any business insights using clustering (if applicable).

- Maximum Customers Prefers Mobile as their Login Device.
- Since Churners are more in Cluster-2, Providing special offers and discounts to cluster-2 customers can decrease the churn rate.
- Some offers should be given for making payment through UPI payment mode to increase its use cluster-0 & Cluster-2.
- E-Wallet payment mode should be provided with offers in cluster-0.
- Maximum churners use Debit card as payment mode across clusters.
- More Churners belong to Regular plus account across clusters.
- Some promotional offers should be given to these customers so that they don’t Churn.

- It is observed that the newly joined customers having low Tenure Churns most, so more concern should be given to the newly joined customers and their cause of churning should be figured out.

4.c) Any other business insights.

- Introducing Personalized offers or Promotions can decrease Churn rate.
- Tier-1 customers have high churn rate, suggesting some incentives or discounted offer can help in Retaining customers.
- UPI and E-wallet payment mode should be encouraged to increase customers.
- Tenure period should be Maximized to Retain the customers.
- Customer care services should be improvised so that the customers complaint can be resolved quickly and efficiently.
- Customized services should be offered to different account segment customers to retain customers.
- Launch tailoring offers to multi user account holders to retain customers.

Thank You