CAPSTONE PROJECT NOTES - 1
DSBA

Contents-

1. Introduction - Business problem definition……4
1.1 Problem Statement……..4
1.2 Need for the project……4
1.3 Project Objectives……..4-5
2. Data Report……6
2.1 Understanding how data was collected in terms of time, frequency and methodology……6
2.2 Visual Inspection of data……6-8
2.3 Understanding the attributes……8-10
3. Exploratory Data Analysis……10
3.1 Univariate Analysis……10
3.2 Histogram for continuous variables……11-14
3.3 Count plot for Categorical Variables……14-19
4.1 Bivariate Analysis and Multivariate Analysis……19-22
4.2 Continuous Variables - ANOVA (Analysis of Variance)……22-23
4.3 Bivariate plots for Categorical variable vs Churn……23-25
4.4 Categorical variables: Chi-squared test of independence at significance level 0.05……25-26
4.5 Pair plot for the numeric variable with hue set as target variable……26-27
4.6 Correlation Heatmap……27
5. Data cleaning and preprocessing……27
5.1 Removal of unwanted variables……27-28
5.2 Missing Value treatment……28-30
5.3 Addition of new Variable……30
5.4 Variable of transformation……30
5.5 Outlier Treatment……30-32
6. Clustering……32-33
7. Business Insights from EDA……33
a) Business Insights from cluster profiling……33
b)Is the data Unbalanced? If so,what can be done? Please explain in the context of the business……33-34
c) Other Business Insights……34
The End……34

List of Figure-

Fig.1 Box plot for numeric Variables……
Fig.2-7 Histogram for Continuous Variables……
Fig. 3-11 Count plot for Categorical Variables……
Fig. 4.1 Tenure vs Churn……
Fig. 4.2 CC_Contacted_LY vs Churn……
Fig. 4.3 Day_Since_CC_Connect vs Churn……
Fig. 4.4 Coupon used for Payment vs Churn……
Fig.4.5 Cashback vs Churn……
Fig. 4.6 Revenue_per_month vs Churn……
Fig. 4.7 Revenue_growth_yoy vs Churn……
Fig. 5-10 Stacked bar chart for categorical variables……
Fig. 6 Pair plot for numeric predictor variables……
Fig. 7 Correlation Heatmap……
Fig. 8 Visualization of nulls……
Fig. 9.1 Before outlier treatment……
Fig. 9.2 After outlier treatment……

1. Introduction - Business problem definition

1.1 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 recommendations.

1.2 Need for the Project-

A DTH provider's biggest cost is the cost of acquisition of a new customer. A customer thus acquired, will need to be retained for quite a few years so that the initial cost of acquisition is recovered back and that particular account is profitable1. Due to this reason, customer churn directly impacts the profitability of a DTH operator. DTH providers also are in a constant pressure to increase their customer base to maintain their profitability as most of them have a fixed broadcaster/content provider fee irrespective of the number of customers in their customer base.

As customer churn directly impacts both the top-line and bottom-line revenue of the business, existing customer base needs to be protected. Providing all customers with offers to retain them would make a dent in the profitability and hence it is very important to focus only on a select set of customers who are at a higher risk of churning.

1.3 Understanding business/social opportunity-

This is a case study of a DTH company where they have customers assigned with unique account ID and a single account ID can hold many customers (like family plan) across gender and marital status, customers get flexibility in terms of mode of payment they want to opt for. Customers are again segmented across various

types of plans they opt for as per their usage which is also based on the device they use (computer or mobile) moreover they earn cashbacks on bill payment. What are the parameters that play a vital role in having customers' loyalty and making them stay? All these social responsibilities will decide the best player in the market.

Data Dictionary-

- AccountID - account unique identifier
- Churn - account churn flag (Target Variable)
- 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 12 months
- 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

2. Data Report-

2.1 Understanding how data was collected in terms of time, frequency and methodology-

- Data has been collected for a random 11,260 unique account ID, across gender and marital status.
- Looking at variables "CC_Contacted_L12m", "rev_per_month", "Complain_l12m", "rev_growth_yoy", "coupon_used_l12m", "Day_Since_CC_connect" and "cashback_l12m"we can conclude that the data has been collected for last 12 month.
- Data has 19 variables, 18 independent and 1 dependent or the target variable, which shows if the customer churned or not.
- The data is the combination of services customers are using along with their payment option and also basic individual details as well.
- Data is a mixture of categorical as well as continuous variables.

2.2 Visual Inspection of data(row, columns,descriptive details)-

- Dataset has 11260 rows and 19 variables.

```
<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
```

- There are **5 columns of float type, 2 columns of integer type and 12 columns of object type.**

- There are **several columns that are supposed to be read as numeric,** instead they have been read as object type for e.g., Tenure is a numeric field but has been read as object. Those columns need to be checked for special characters and need to be treated before the column can be changed to numeric for further processing.
- There are no duplicate rows in the data set. Each account id has one unique row.
- Several columns **have null values.**
- The following table shows the number of rows containing nulls and special characters that require data cleaning. **All special characters present in the data set were treated with nulls** so that they can be imputed. Some columns such as Gender and account_segment contained multiple values to represent the same category for e.g., 'M', 'Male'. Cleaning up of those values was also performed.

Column	Values Count	% Rows With values present	Number Of Nulls	% Rows With Nulls	Data cleanup Needed?	% Rows needing data cleaning
Account ID	11260	100.00%	0	0%	None	0.00%
Churn	11260	100.00%	0	0%	None	0.00%
Tenure	11158	99.09%	102	0.91%	Yes-#	1.03%
City_Tier	11148	99.01%	112	0.99%	None	0.00%
CC_Contacted_ly	11158	99.09%	102	0.91%	None	0.00%
Payment	11151	99.03%	109	0.97%	None	0.00%
Gender	11152	99.04%	108	0.96%	Yes-M,F	5.74%
Service_Score	11162	99.13%	98	0.87%	None	0.00%
Account_user_count	11148	99.01%	112	0.99%	Yes-@	2.95%
account_segment	11163	99.14%	97	0.86%	Yes-Regular + Super +	2.74%

CC_Agent_Score	11144	98.97%	116	1.03%	None	0.00%
Marital_Status	11048	98.12%%	212	1.88%	None	0.00%
rev_per_month	11158	99.09%	102	0.91%	Yes- +	6.12%
Complain_ly	10903	96.83%	357	3.17%	None	0.00%
rev_growth_yoy	11260	100.00%	0	0.00%	Yes - $	0.03%
Coupon_used_for_ payment	11260	100.00%	0	0.00%	Yes-$,*,#	0.03%
Day_Since_CC_conne ct	10903	96.83%	357	3.17%	Yes - $	0.01%
Cashback	10789	95.82%	471	4.18%	Yes - $	0.02%
Login_device	11039	98.04%	221	1.96%	Yes-&&&	4.79%

2.3 Understanding the attributes -

This project has 18 attributes contributing towards the target variable. Let's discuss these variables one after another.

1. AccountID – This variable represents a unique ID which represents a unique customer. This is of Integer data type and there are no null values present in this.

2. Churn – This is our target variable, which represents if the customer has churned or not. This is categorical in nature and will no null values. "0" represents "NO" and "1"represents "YES".

3. Tenure – This represents the total tenure of the account since opened. This is a continuous variable with 102 null values.

4. City_Tier – These variable segregates customer into 3 parts based on city the primary customer resides. This variable is categorical in nature and has 112 null values.

5. CC_Contacted_L12m – This variable represents the number of times all the customers of the account have contacted customer care in the last 12 months. This variable is continuous in nature and has 102 null values.

6. Payment – This variable represents the preferable mve 109 null values.

7. Gender – This variable represents the gender of the pride of bill payment opted by the customer. This is categorical in nature and haimary account holder. This is categorical in nature and 108 null values.

8. Service_Score – Scores provided by the customer basis the service provided by the company. This variable is categorical in nature and has 98 null values.
9. Account_user_count – This variable gives the number of customers attached with an accountID. This is continuous in nature and has 112 null values.
10. account_segment – These variables segregate customers into different segments based on their spend and revenue generation. This is categorical in nature and has 97 null values.
11. CC_Agent_Score — Scores provided by the customer basis the service provided by the customer care representative of the company. This variable is categorical in nature and has 116 null values.
12. Marital_Status – This represents the marital status of the primary account holder. This is categorical in nature and has 212 null values.
13. rev_per_month – This represents average revenue generated per account ID in the last 12 months. This variable is continuous in nature and has 102 null values.
14. Complain_l12m – This denotes if customers have raised any complaints in the last 12 months. This is categorical in nature and has 357 null values.
15. rev_growth_yoy – This variable shows revenue growth in percentage of account for 12 months Vs 24 to 13 months. This is continuous in nature and doesn't have any null values.
16. coupon_used_l12m – This represents the number of times customer's have used discount coupons for bill payment. This is continuous in nature and doesn't have any null values.
17. Day_Since_CC_connect – This represents the number of days since customers have contacted the customer care. Higher the number of days denotes better the service. This is continuous in nature and has 357 null values.
18. cashback_l12m – This variable represents the amount of cash back earned by the customer during bill payment. This is continuous in nature and has 471 null values.
19. Login_device – This variable represents in which device the customer is availing the services if it's on phone or on computer. This is categorical in nature and has 221 null values.

The following table shows a basic statistical description of the numeric columns after data clean-up. It contains a 5-point summary of the numeric fields –

minimum, maximum, 25th percentile, 50th percentile and 75th percentile. In addition, it contains the count of values present in each column, mean and standard deviation.

	count	mean	std	min	25%	50%	75%	max
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

- Tenure seems to have a very huge range up to 99. It could be in months in which case those would be valid values.
- The maximum limit for many of the variables seems to be very far apart from the 75th percentile for many variables such as cashback, revenue per month and customer contacted last year. There seems to be significant positive skew in these variables. A look at the boxplot and histogram will confirm the presence of outliers.

3. Exploratory Data Analysis-

3.1 Univariate Analysis-

It has been done by observing:

- Box plots and histograms for continuous variables.
- Count plots for categorical variables.

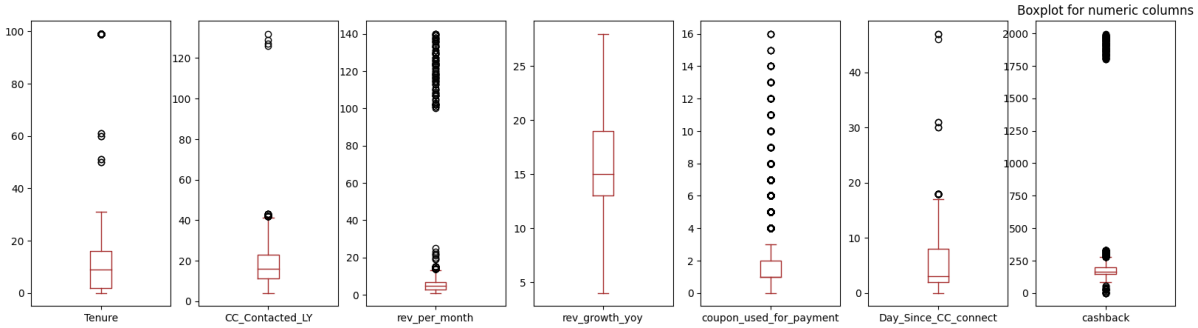
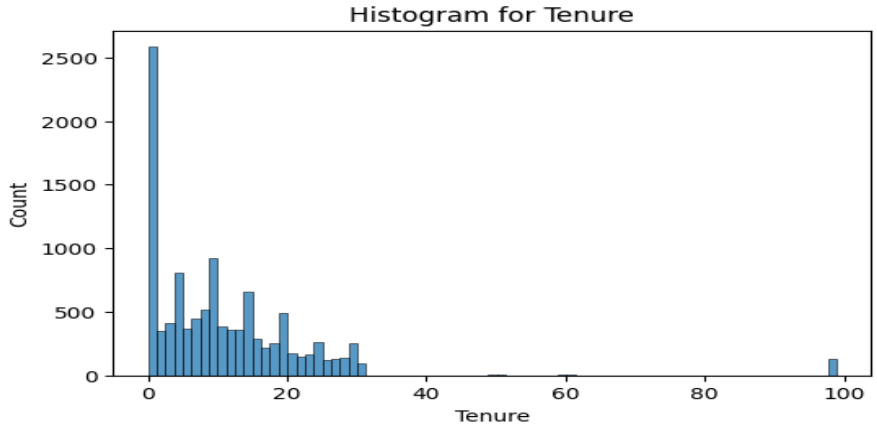
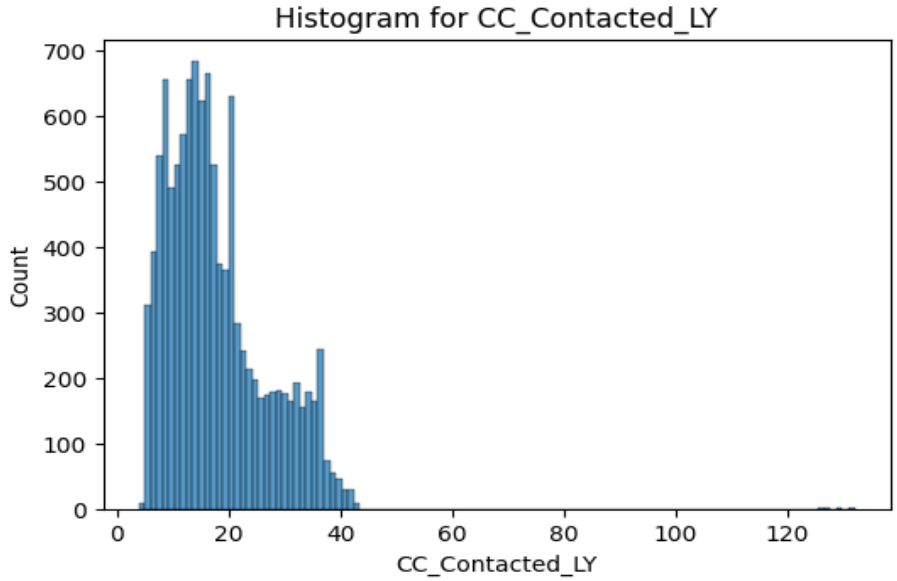
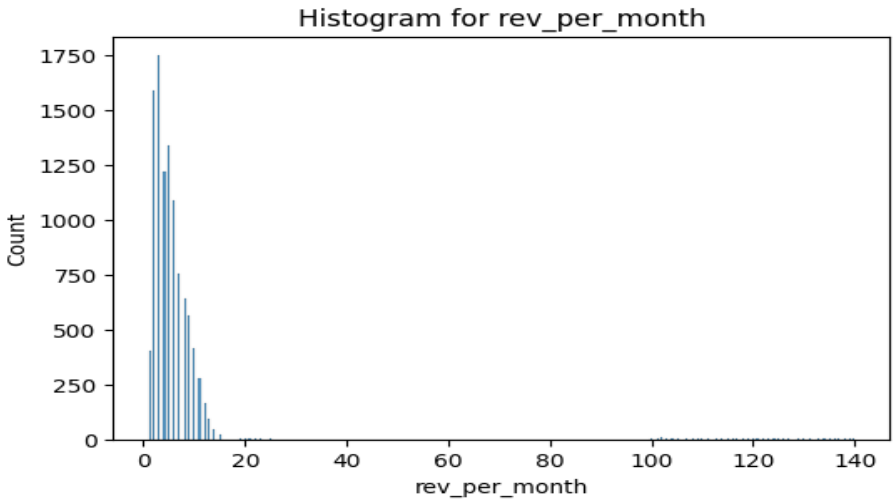


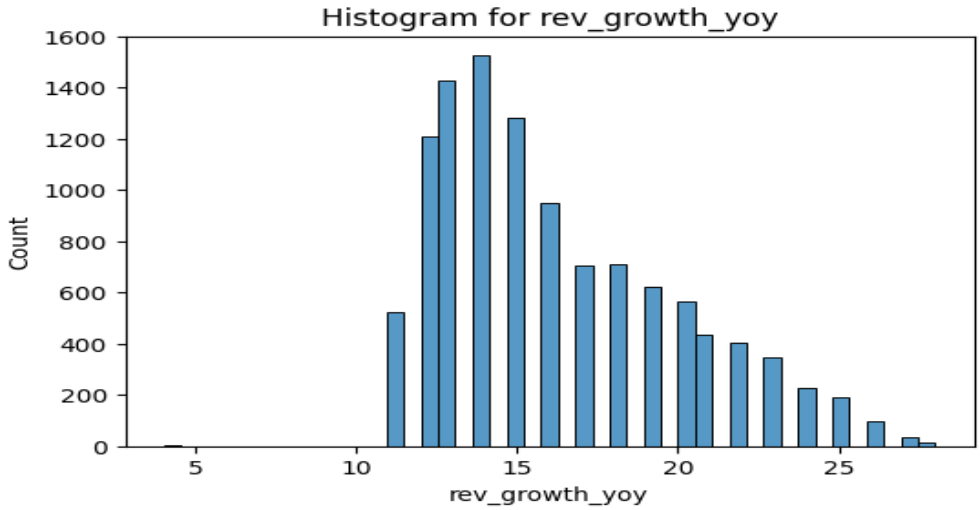
Fig 1.Boxplot for numeric variables

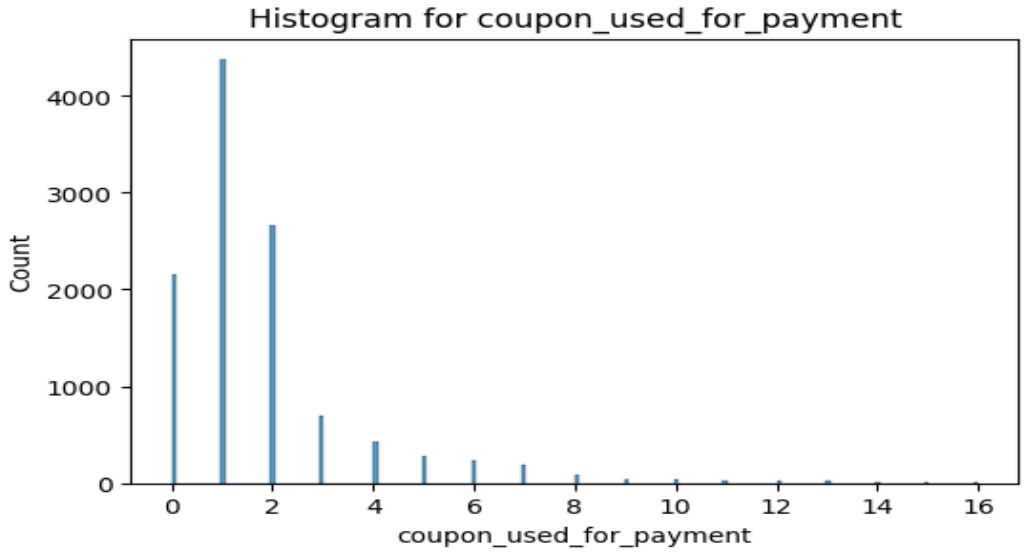
3.2 Histogram for Continuous Variables-

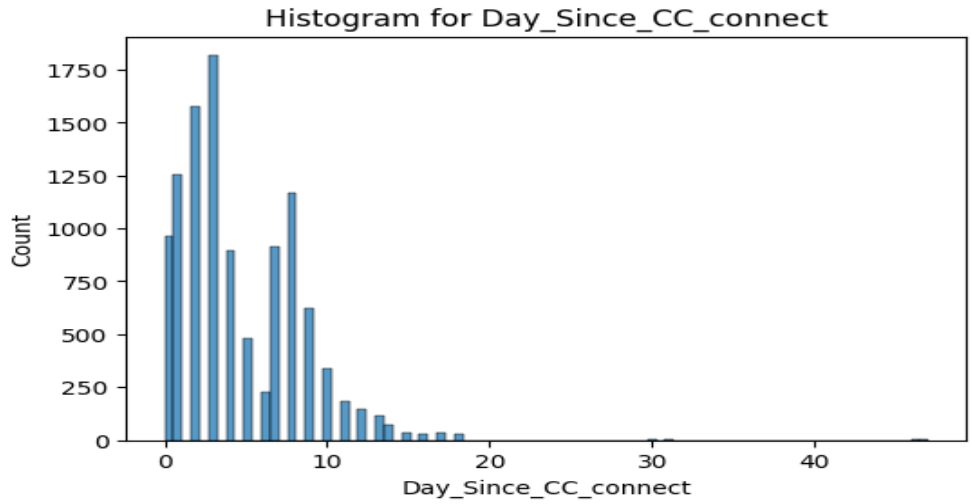





Histogram for rev_per_month
Count
rev_per_month
0
250
500
750
1000
1250
1500
1750
0
20
40
60
80
100
120
140


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


Histogram for coupon_used_for_payment
Count
coupon_used_for_payment
0
1000
2000
3000
4000
0
2
4
6
8
10
12
14
16



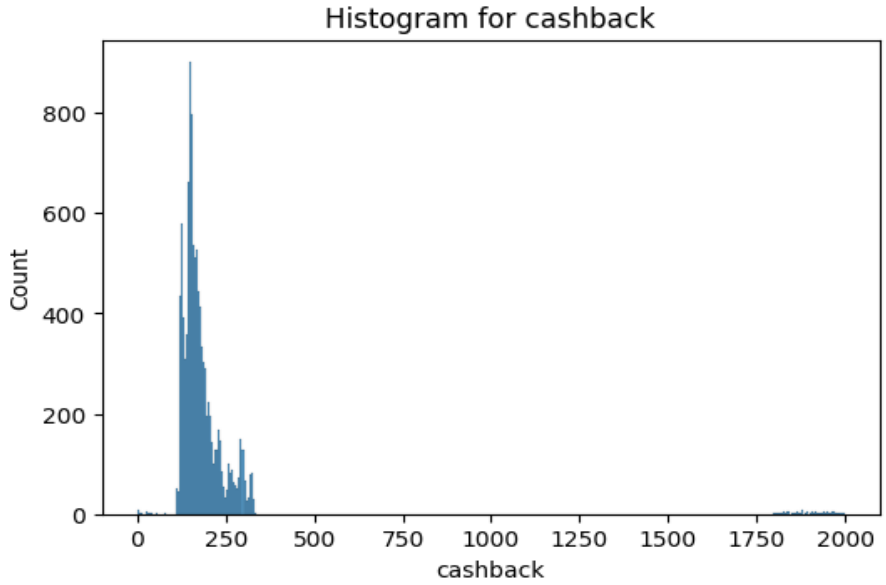


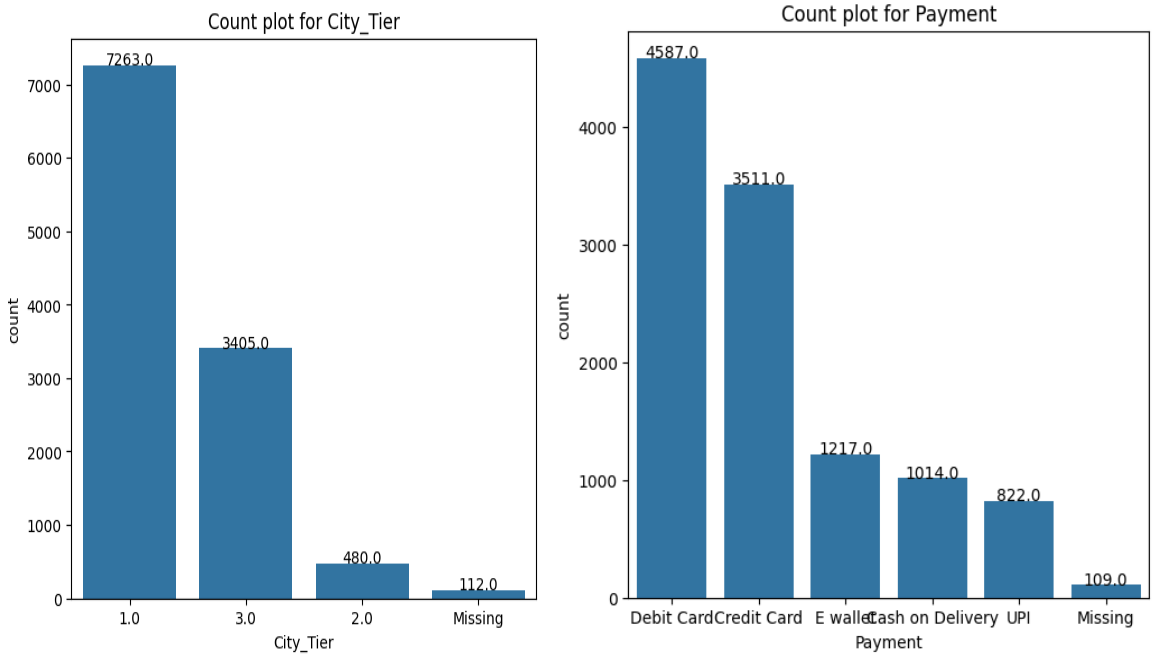
Fig.2-7 Histogram for Continuous Variables

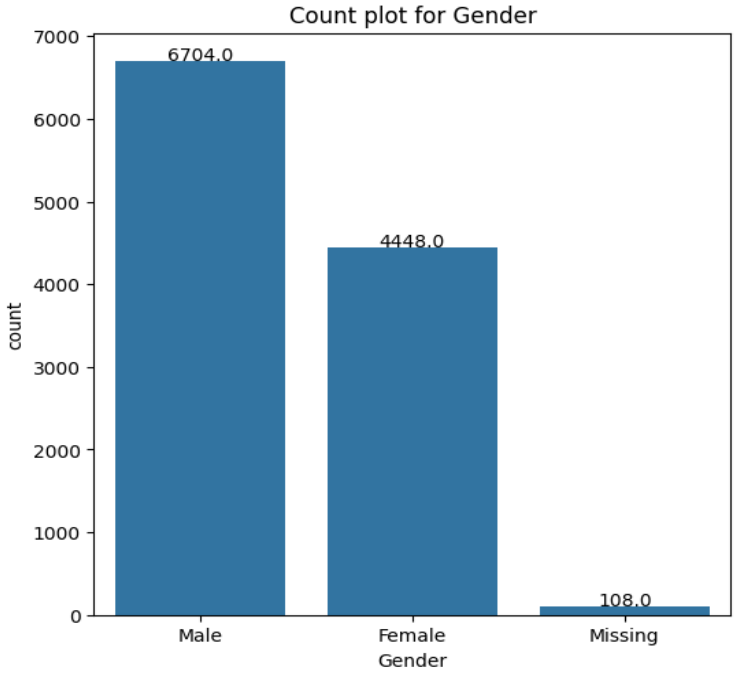
	Skewness
Tenure	3.90
CC_Contacted_LY	1.42
rev_per_month	9.09
rev_growth_yoy	0.75
coupon_used_for_payment	2.58
Day_Since_CC_connect	1.27
cashback	8.77

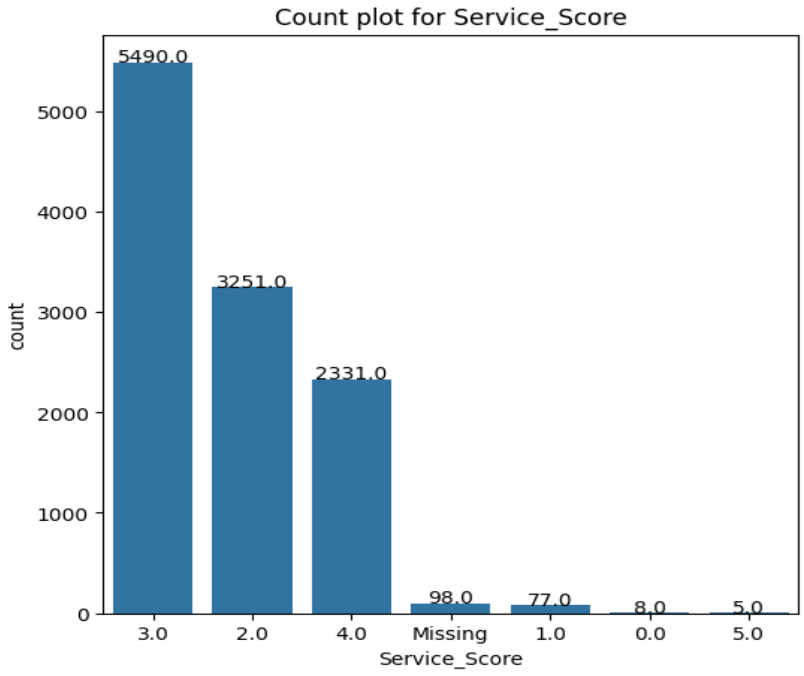
Observations-

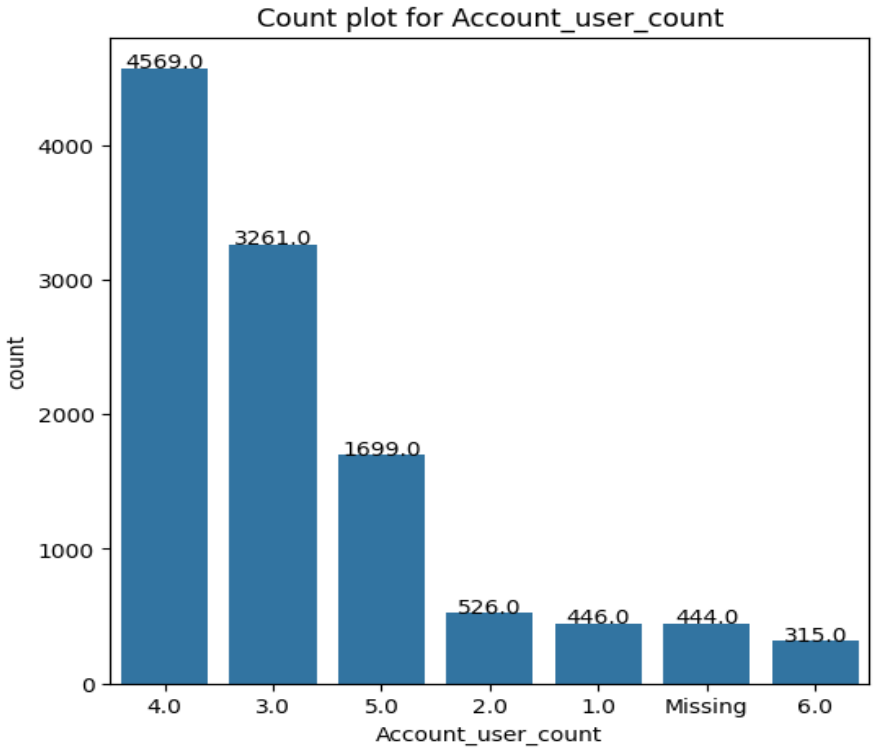
- All numeric variables with the exception of rev_growth_yoy have outliers. Some outliers for certain variables are closer to the whisker, whereas there are a group of outliers that are far beyond the whisker with no in between values.
- Coupon_used_for_payment has a very limited range 0 to 16. Hence, for the purpose of this analysis, the outliers will not be treated.
- All numeric variables with the exception of rev_growth_yoy have a high positive skew.

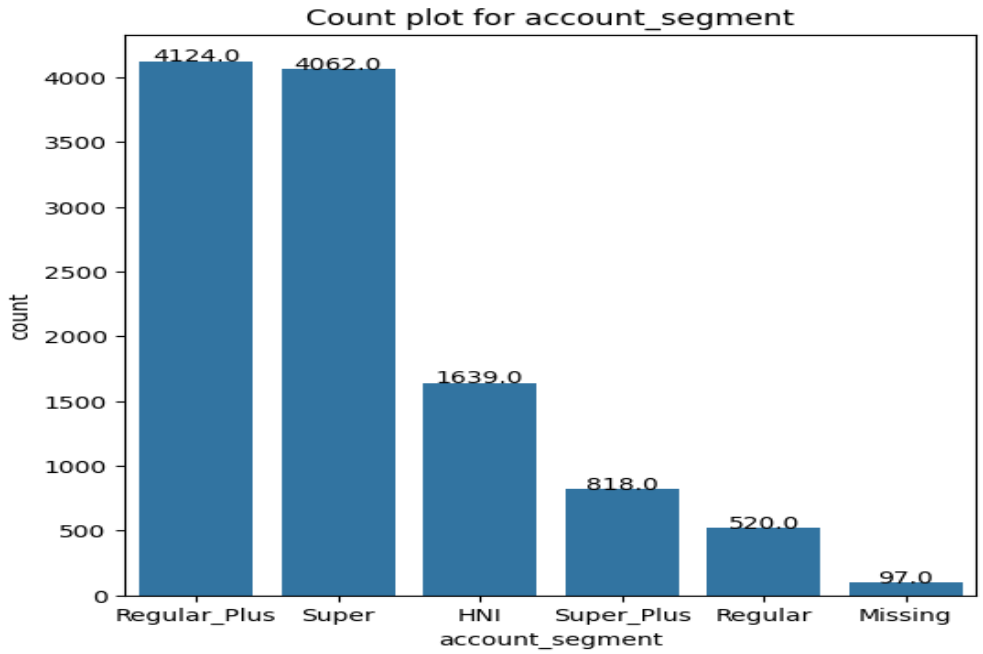
3.3 Count Plot for Categorical Variables-

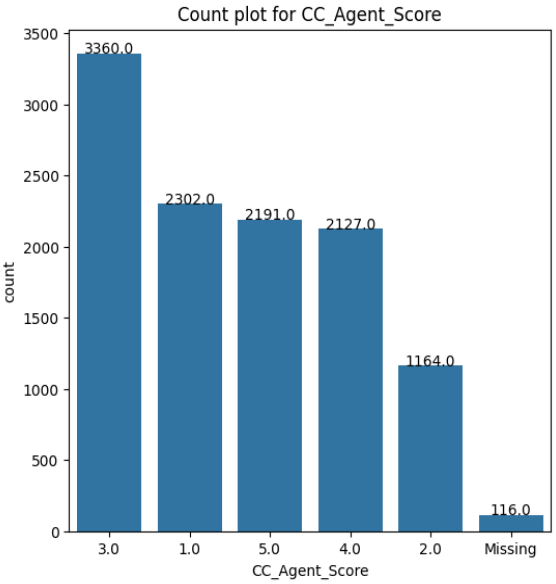


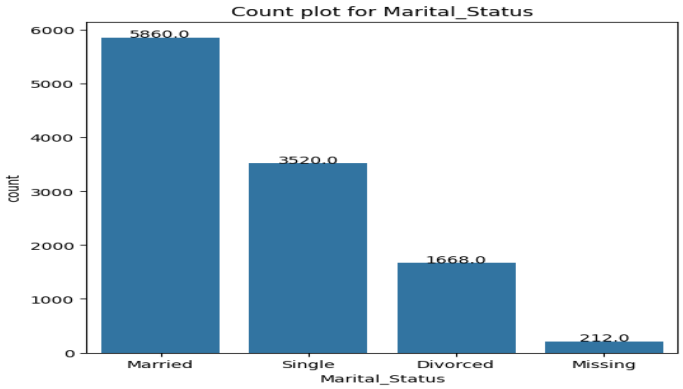

Count plot for Gender
7000
6000
5000
4000
3000
2000
1000
0
count
6704.0
4448.0
108.0
Male
Female
Missing
Gender

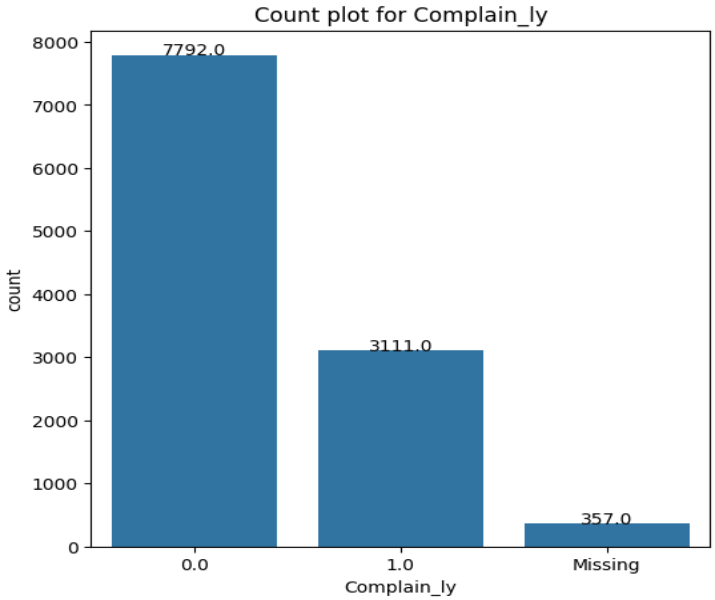

Count plot for Service_Score
5000
4000
3000
2000
1000
0
count
5490.0
3251.0
2331.0
98.0
77.0
8.0
5.0
3.0
2.0
4.0
Missing
1.0
0.0
5.0
Service_Score

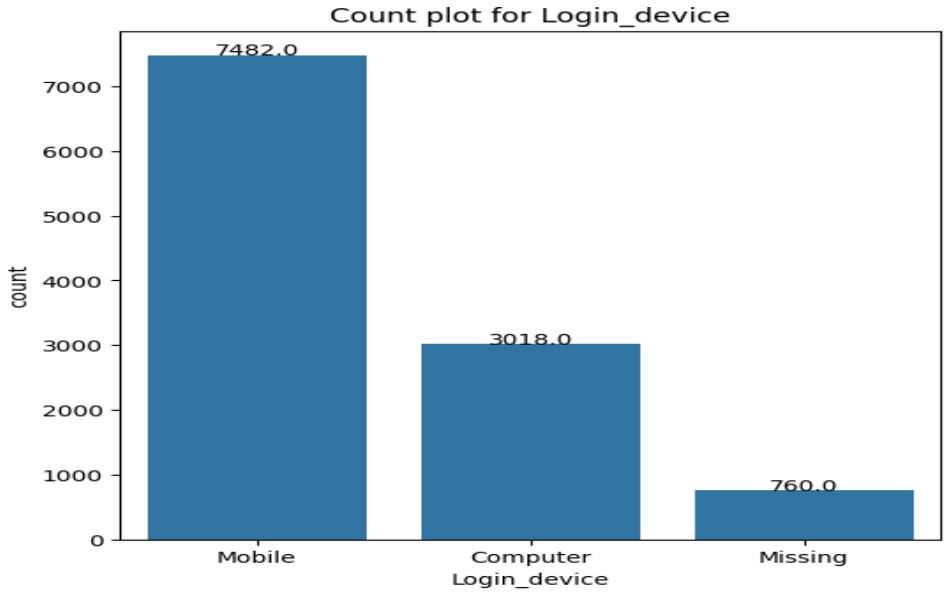

Count plot for Account_user_count
4569.0
3261.0
1699.0
526.0
446.0
444.0
315.0
4000
3000
2000
1000
0
count
4.0
3.0
5.0
2.0
1.0
Missing
6.0
Account_user_count


Count plot for account_segment
4124.0
4062.0
1639.0
818.0
520.0
97.0
4000
3500
3000
2500
2000
1500
1000
500
0
count
Regular_Plus
Super
HNI
Super_Plus
Regular
Missing
account_segment


Count plot for CC_Agent_Score
count
CC_Agent_Score
3360.0
2302.0
2191.0
2127.0
1164.0
116.0
3.0
1.0
5.0
4.0
2.0
Missing


Count plot for Marital_Status
count
Marital_Status
5860.0
3520.0
1668.0
212.0
Married
Single
Divorced
Missing


Count plot for Complain_ly
count
Complain_ly
7792.0
3111.0
357.0
0.0
1.0
Missing



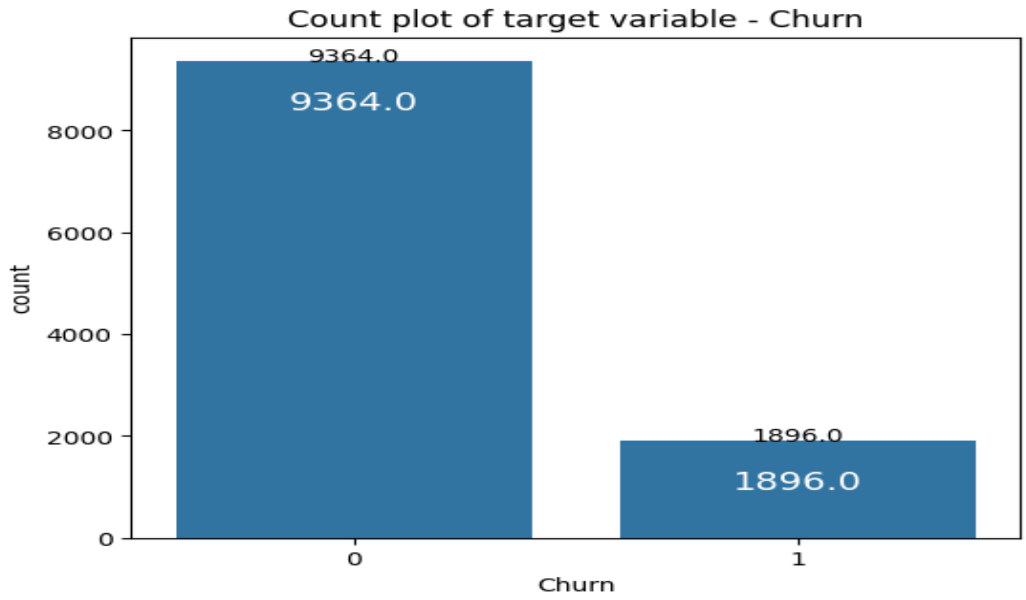


Fig 3-11 Count plot for Categorical Variable

Observations-

- This is an imbalanced dataset with target variable containing approx 16.8% churn.
- Tier 1 cities have more accounts followed by Tier 3 cities.
- Most of the accounts pay through debit card followed by credit card. UPI ranks last amongst payment methods.
- Number of male account holders outnumber females.
- Regular plus and Super are top two account segment types by number.

- Top score for both Customer service agent and Service score is 3.
- Married customers have the most accounts followed by single.
- Most accounts do not have a customer complaint filed last year.
- Most account holders use Mobile for logging and using services.

4.1 Bivariate Analysis and Multivariate Analysis-

Continuous predictor variables that show some ability to separate the target variables-

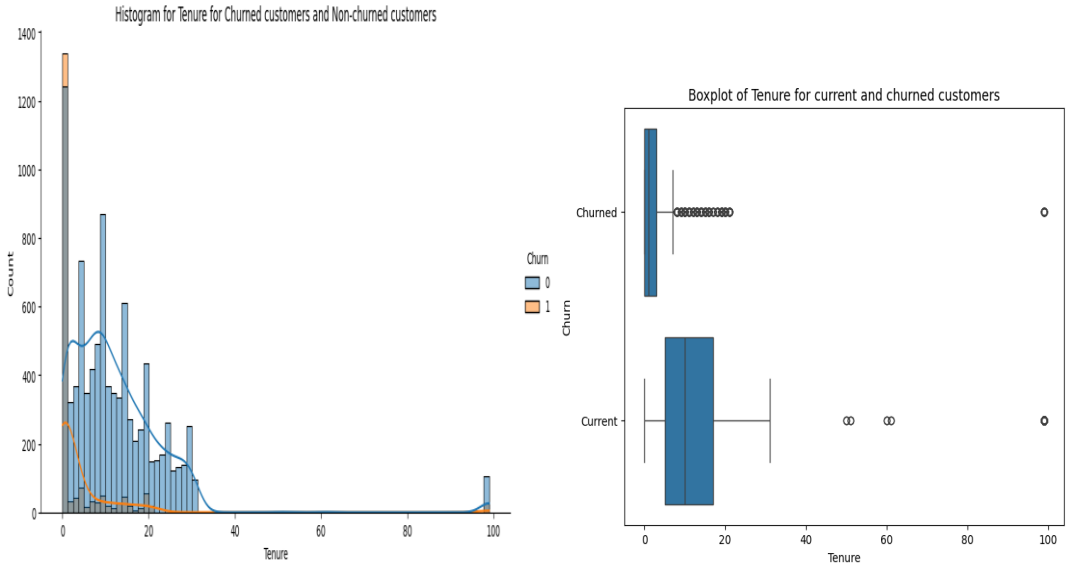


Fig 4.1 Tenure vs Churn

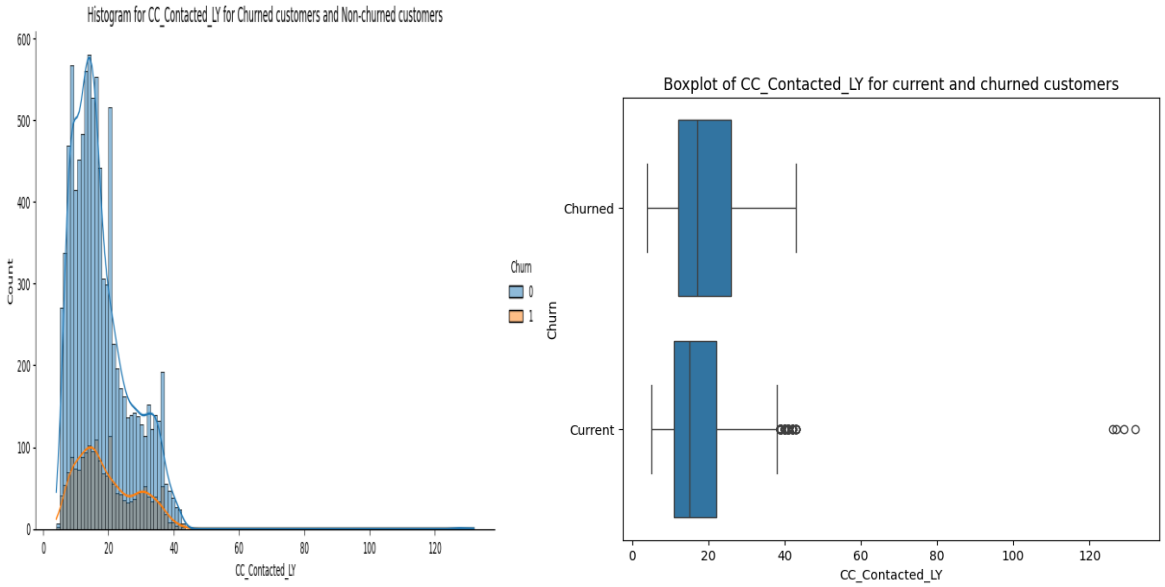


Fig 4.2 CC_Contacted_LY vs Churn

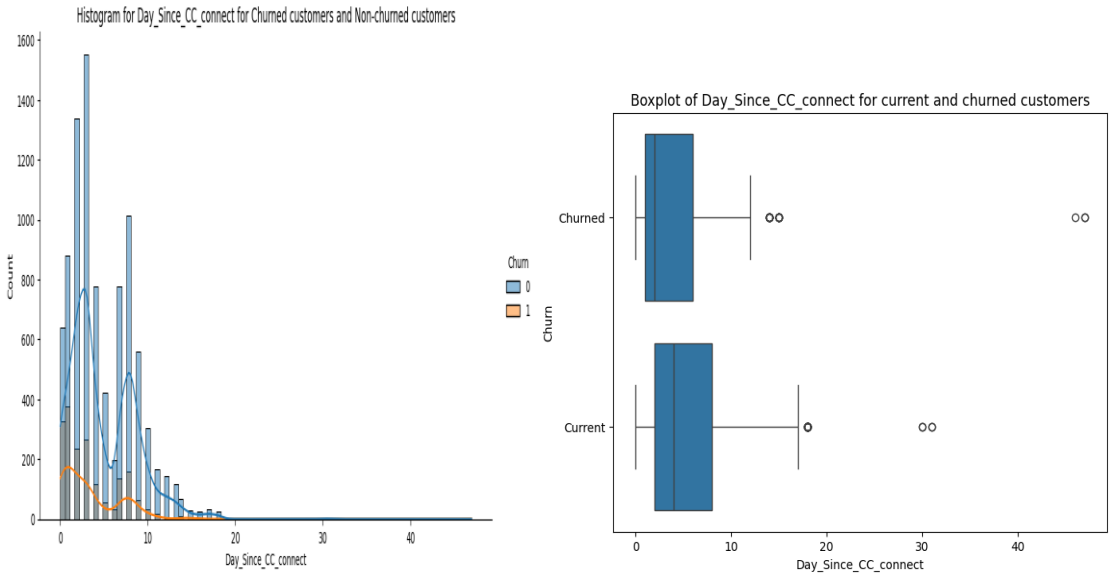


Fig 4.3 Day_Since_CC_Connect vs Churn

Continuous predictor variables that aren't able to show a clear separation between target classes-

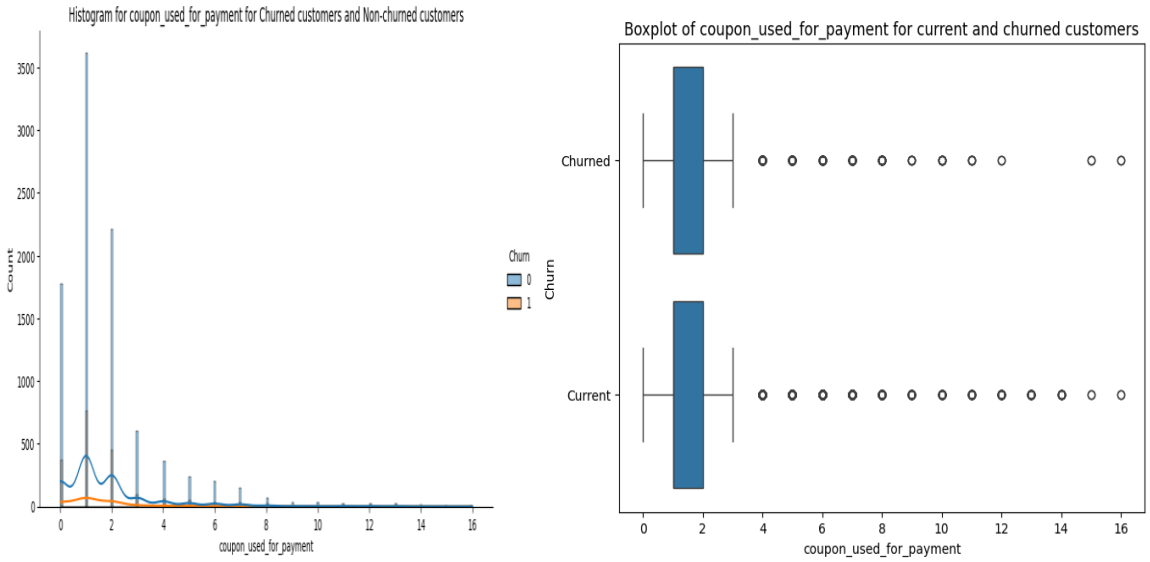


Fig 4.4 Coupon used for payment vs Churn

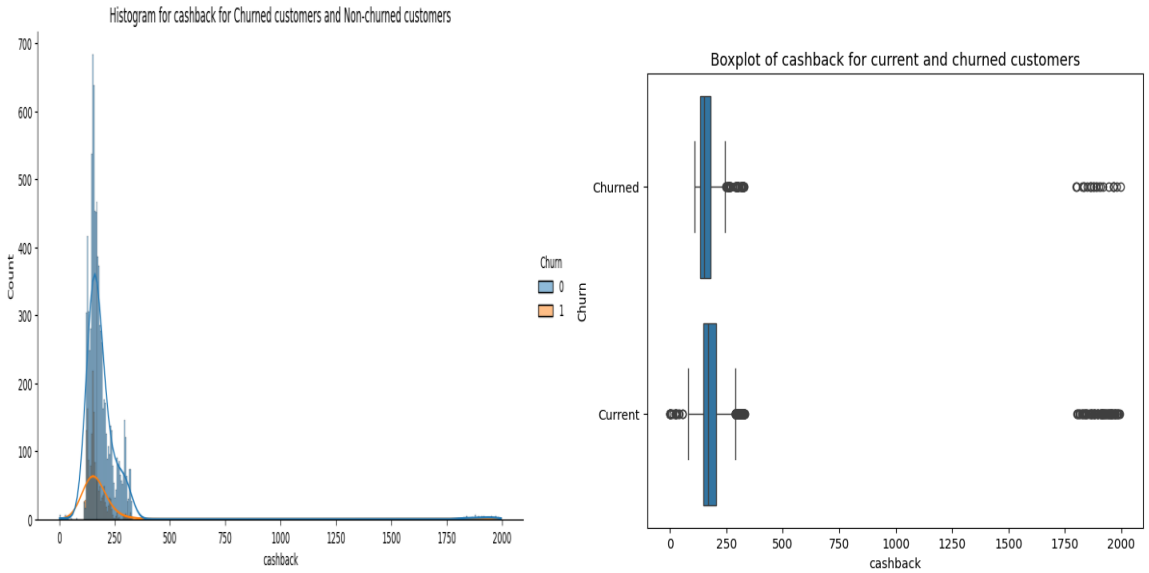


Fig 4.5 Cashback vs Churn

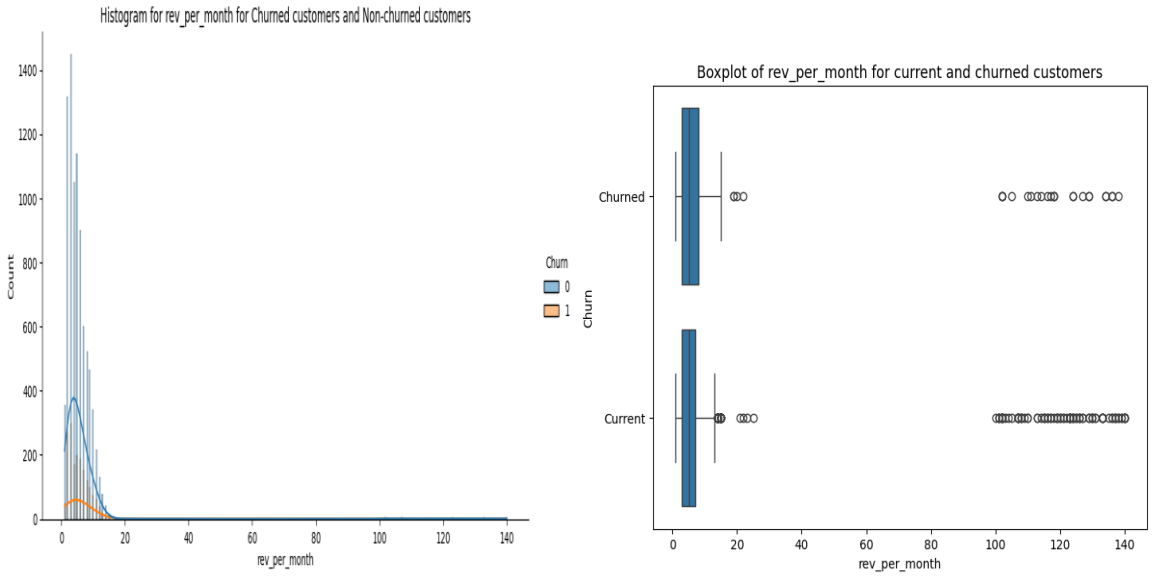


Fig 4.6 Revenue_per_month vs Churn

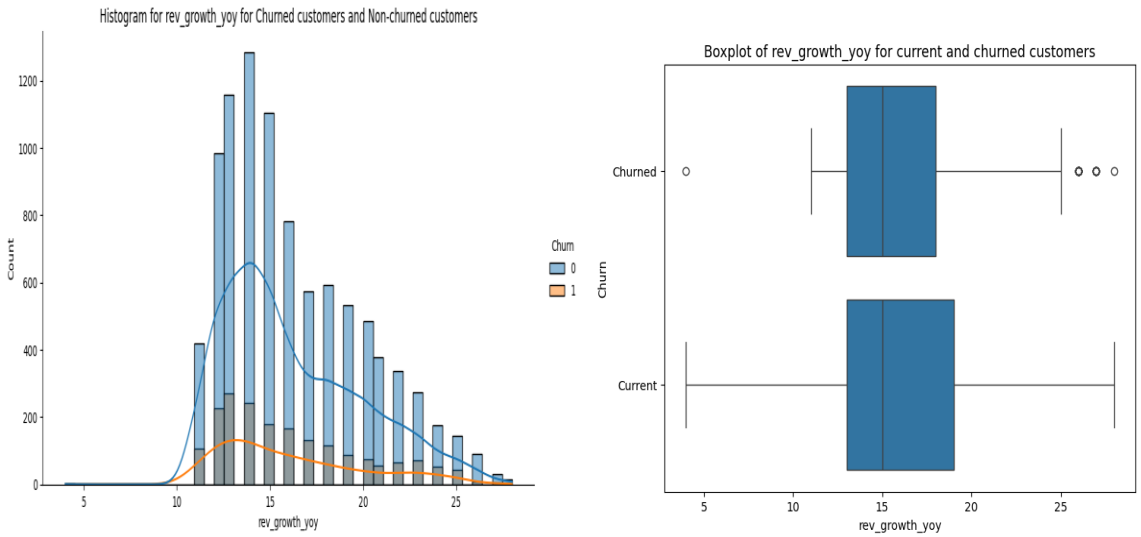


Fig 4.7 Revenue_growth_yoy vs Churn

Observations-

From the above plots we can see that variables such as Tenure, Days_since_CC_connect have some influence on the target variable. The median lines for Churned and Non-churned observations when plotted against these variables show a difference. Whereas, the medians in boxplots for variables such as coupon_used_for_payment, rev_growth_yoy do not show much difference in churned vs non-churned distributions.

4.2 Continuous Variables - ANOVA (Analysis of Variance)

H0: Means of all groups are equal

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

Results of Anova test for following variables and churn-

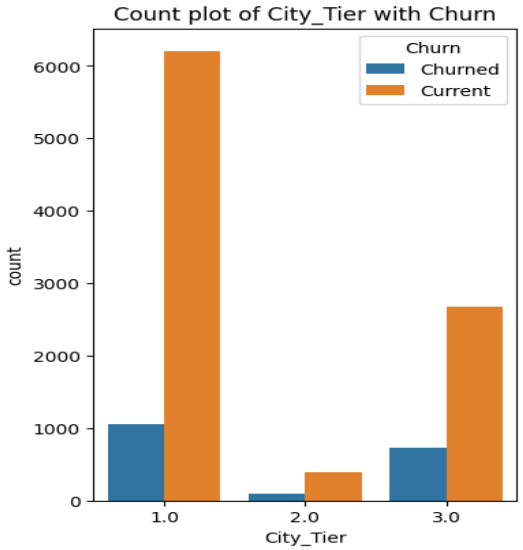
Variable	**F-Statistics**	**Probability of>F**	**Inference at significance level of 5%**
Tenure	634.6	3.3E-36	Reject null hypothesis. The means are different. Variable significant to model building.
CC_Contacted_ly	58.25	2.94E-14	Reject null hypothesis. The means are different.Variable significant to model building.
rev_per_month	5.32	0.021	Reject null hypothesis. The means are different.Variable significant to model building.

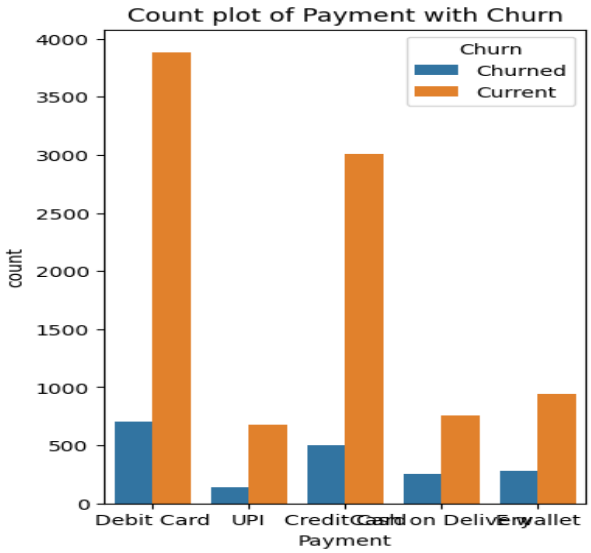
rev_growth_yoy	2.17	0.141	Cannot reject null hypothesis. The means are equal. Variable can be dropped.
Coupon_used_for_ payment	2.47	0.116	Cannot reject null hypothesis. The means are equal. Variable can be dropped.
Day_Since_CC_Connect	243.9	2.1E-54	Reject null hypothesis. The means are different.Variable significant to model building.
Cashback	11.32	0.001	Reject null hypothesis. The means are different.Variable significant to model building.

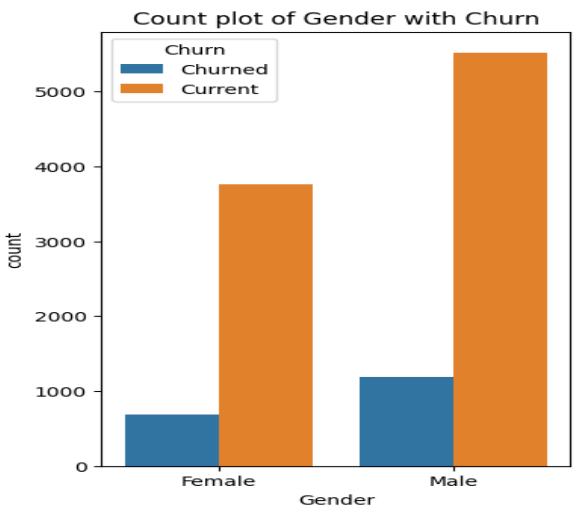
Observation-

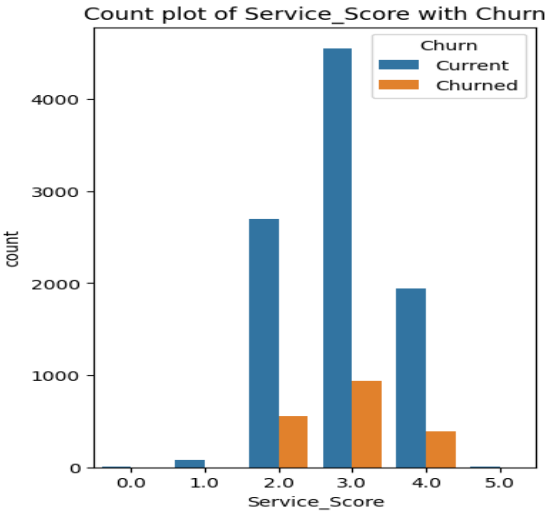
At a significance level of 0.05 (5%), the tests for the variables rev_growth_yoy and coupon_used_for_payment have given a p-value of greater than 0.05. In these two cases, Ho cannot be rejected, i.e., the means for the two groups churn=0 and churn=1 for these variables are the same. This implies that since the groups are not too different, these two variables cannot be significant predictors of the target variable.

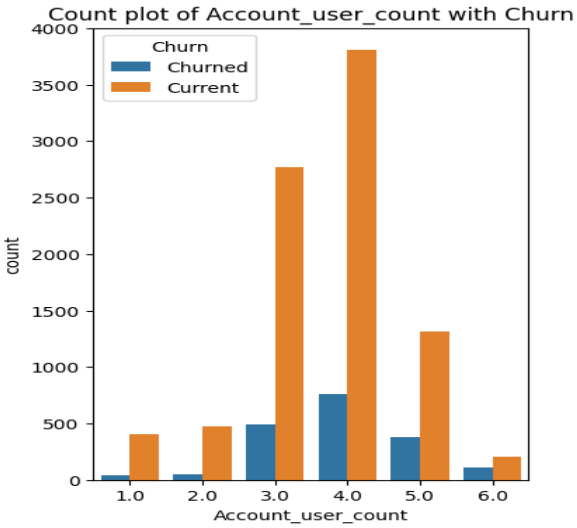
4.3 Bivariate plots for Categorical variables vs Churn-

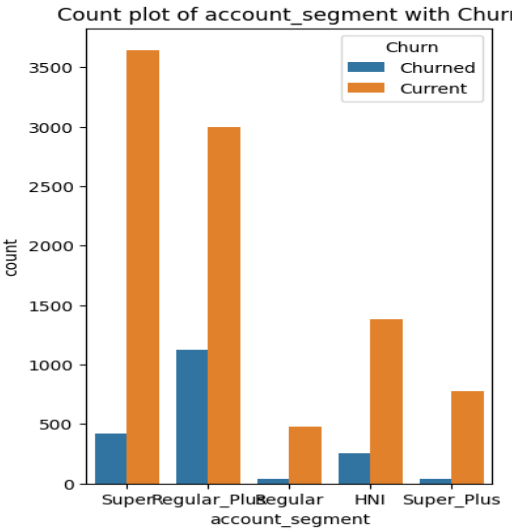


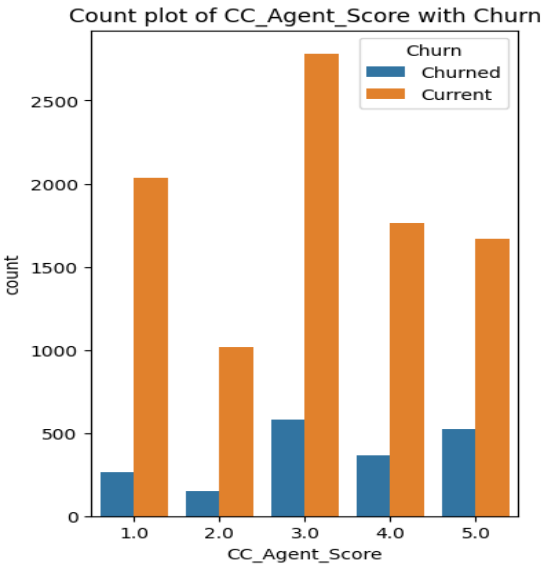


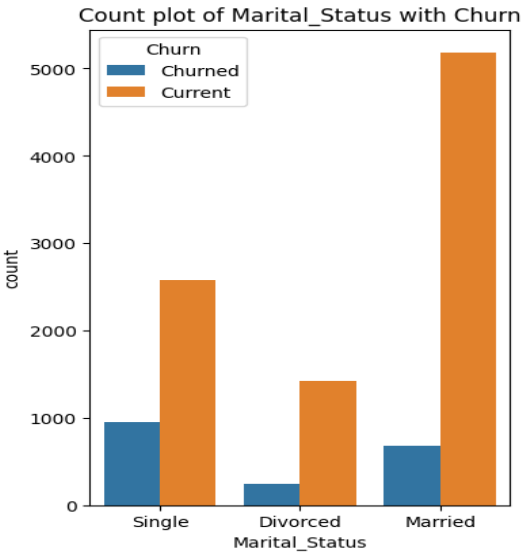

Count plot of Gender with Churn
Churn
Churned
Current
count
Female
Male
Gender


Count plot of Service_Score with Churn
Churn
Current
Churned
count
Service_Score


Count plot of Account_user_count with Churn
Churn
Churned
Current
count
Account_user_count


Count plot of account_segment with Churn
Churn
Churned
Current
count
account_segment


Count plot of CC_Agent_Score with Churn
Churn
Churned
Current
count
CC_Agent_Score


Count plot of Marital_Status with Churn
Churn
Churned
Current
count
Single
Divorced
Married
Marital_Status

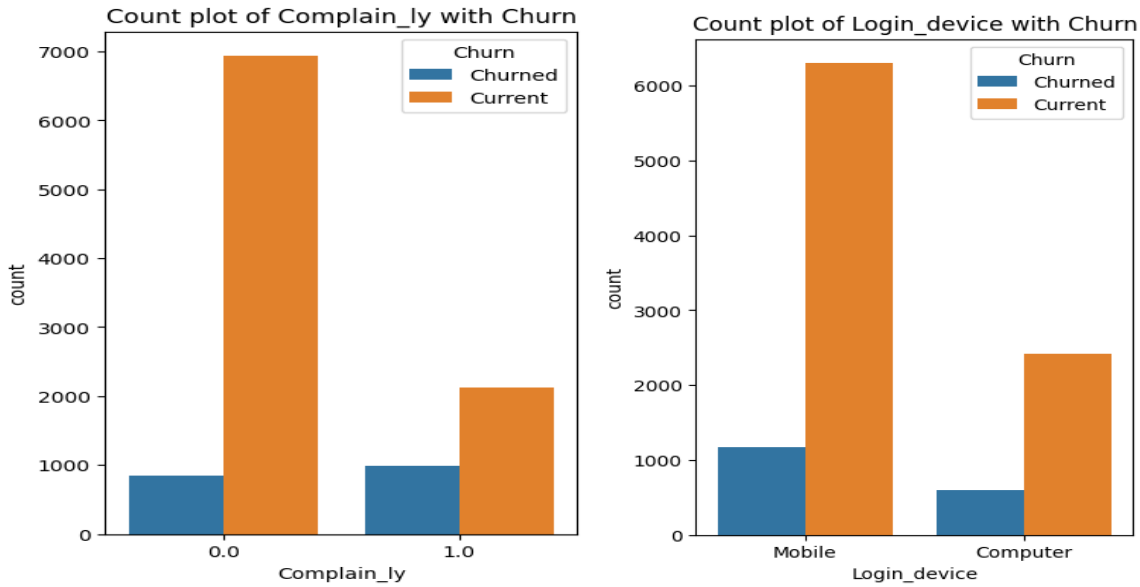


Fig 5-10 Stacked bar chart for categorical variables

Observations-

- In the above stacked bar charts, active customers are called current customers. The first bar shows distribution of the categorical predictor variable being analysed within the churned customer.
- The second bar shows distribution within active/current customers.
- From the above charts, some of the variables like city_tier, account_segment, Complain_ly, Marital_Status, CC_Agent_score seem to show a difference in distribution when churned vs current customers are considered.
- Other variables such as Gender and Service_Score have more or less similar distributions within Churned and Current/active customer bars.

4.4 Categorical variables: Chi-squared test of independence at significance level 0.05 -

Null Hypothesis: There is no relationship between two categorical variables.

Alternate Hypothesis: There is a relationship between two categorical variables.

	Variable	chi2	p-value	chi2_output
0	Gender	8.983146	2.724812e-03	Reject Ho; Dependent.
1	Service_Score	18.414690	2.469166e-03	Reject Ho; Dependent.
2	City_Tier	80.288817	3.677095e-18	Reject Ho; Dependent.
3	Payment	103.799617	1.526348e-21	Reject Ho; Dependent.
4	Account_user_count	154.959445	1.173574e-31	Reject Ho; Dependent.
5	account_segment	567.068402	2.073937e-121	Reject Ho; Dependent.
6	CC_Agent_Score	139.031565	4.549521e-29	Reject Ho; Dependent.
7	Marital_Status	379.808123	3.355165e-83	Reject Ho; Dependent.
8	Complain_ly	688.084739	1.166239e-151	Reject Ho; Dependent.

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

4.5 Pair plot for the numeric variables with hue set as target variable-

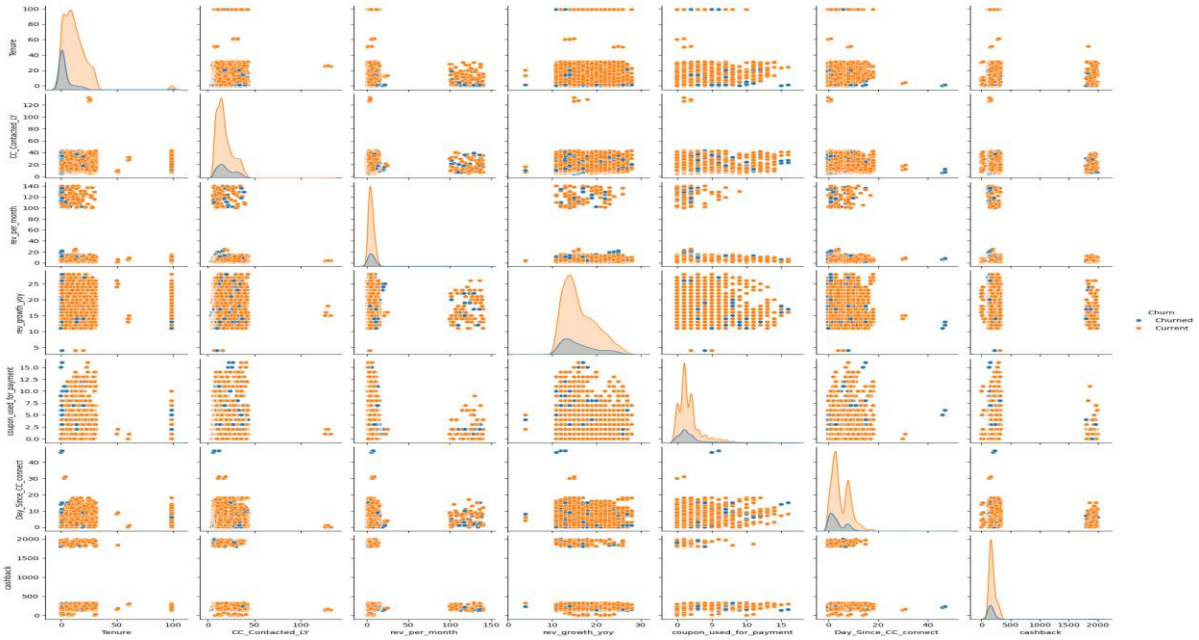


Fig 6. Pair plot for numeric predictor variables

Observations-

- Some of the variables clearly show presence of some clusters for e.g., rev_per_month and Day_Since_CC_connect.
- The diagonal kde plot for Tenure shows a slight separation with customers who churned falling on the lower side of Tenure.
- There is no linear relationship between any two continuous variables.

4.6 Correlation Heatmap-

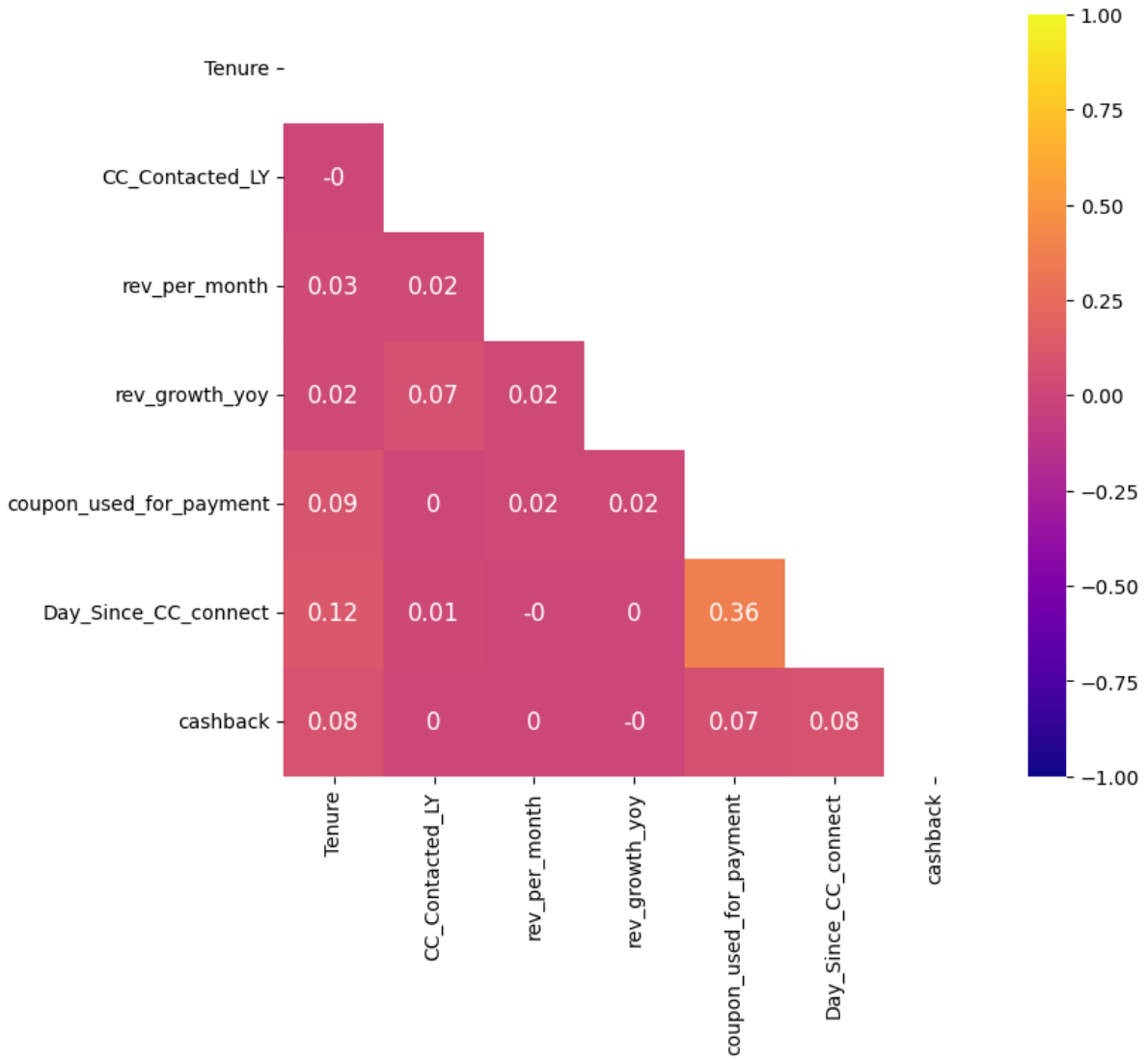


Fig 7. Correlation Heatmap

Observation- None of the numeric variables show a strong correlation. Hence there is no need to drop any variable.

5. Data cleaning and preprocessing-

5.1 Removal of unwanted variables-

- After in-depth understanding of data we conclude that removal of variables is not required at this stage of the project. We can remove the

variable “AccountID” which denotes a unique ID assigned to unique customers.

- Any variable that remains constant for all or most of the observations as this does not add any strength to prediction. As observed from the histogram and count plots/value counts (categorical), there are no variables that have constant value for all observations.
- Any predictor variable that has a very weak correlation with the target variable. As seen from the Chi square test and Anova test in the bivariate analysis.

5.2 Missing Value treatment-

Percentage of nulls-

Percentage nulls or missing values present in the predictor variables of the dataset are as follows:

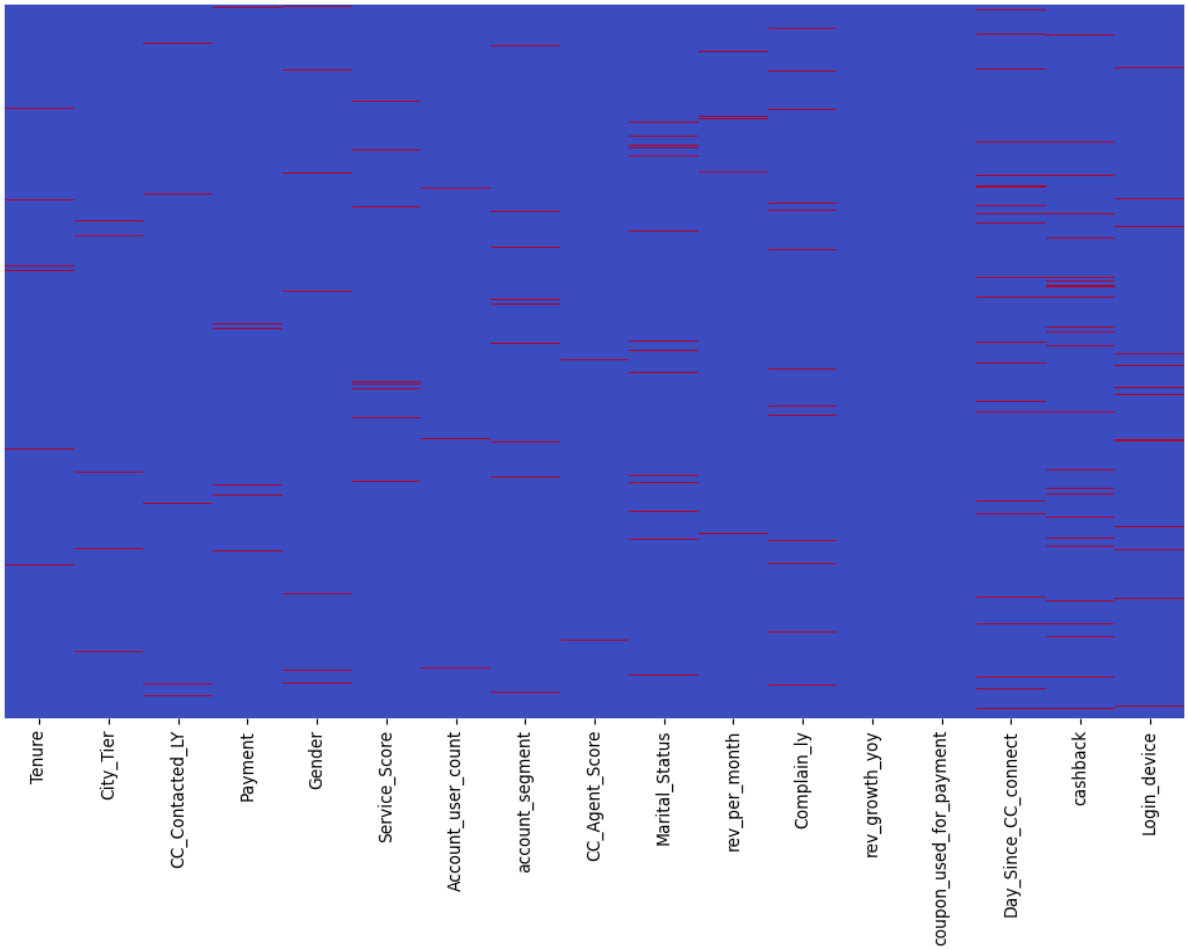


Fig 8. Visualization of nulls

```
cashback                  4.18
Day_Since_CC_connect      3.17
Complain_ly               3.17
Login_device              1.96
Marital_Status            1.88
CC_Agent_Score            1.03
Account_user_count        0.99
City_Tier                 0.99
Payment                   0.97
Gender                    0.96
rev_per_month             0.91
CC_Contacted_LY           0.91
Tenure                    0.91
Service_Score             0.87
account_segment           0.86
rev_growth_yoy            0.00
coupon_used_for_payment   0.00
dtype: float64
```

Missing Value Treatment-

Missing value treatment was done using KNN imputation, a distance-based method. The following treatments were done as they are pre-requisites for missing value treatment using KNN.

- All variables need to be numeric. Any object-type categorical variables need to be encoded suitably (label/ one-hot encoding).
- One hot encoded variables 'Payment','Gender','account_segment', 'Marital_Status','Login_device'.
- All variables need to be scaled as KNN is a distance-based algorithm. Scaling was done using Standard Scaler function from SKLearn library for the predictor variables.
- Null imputing was done using Sklearn's KNNImputer function. This algorithm imputed missing values using K-nearest neighbors.

Nulls in predictor variables after KNN Imputing-

```
Tenure                    0
City_Tier                 0
CC_Contacted_LY           0
Service_Score             0
User_Count                0
CC_Score                  0
Rev_Permonth              0
Complain_LY               0
Days_Since_CC             0
Cashback                  0
Payment_Creditcard        0
Payment_Debitcard         0
Payment_Ewallet           0
Payment_UPI               0
Gender_Male               0
ACSegment_Regular         0
ACSegment_Regularplus     0
ACSegment_Super           0
ACSegment_Superplus       0
Maritalstatus_Married     0
Maritalstatus_Single      0
Logindevice_Mobile        0
dtype: int64
```

5.3 Addition of new Variable-

Cluster code will be added to the dataset. It may be used when experimenting with model building.

5.4 Variable of transformation-

- Encoding: The variables Payment, Gender, Account_Segment, Marital_Status and Login_device are all categorical object types. They need to be converted to numeric variables. The categories in these variables do not have an order. Hence, they were one-hot encoded.
- Scaling: The dataset has been scaled as it is a pre-requisite for any distance-based algorithm like KNN imputer, K-means clustering, KNN and ANN. The scaled data can also be used by all models irrespective of whether they expect scaled input or not.
- No other transformation is expected for modeling as of now.

5.5 Outlier Treatment-

Here, outliers will be treated by capping to the lower and upper range where

- lower_range= Q1-(1.5 * IQR) and
- upper_range= Q3+(1.5 * IQR)

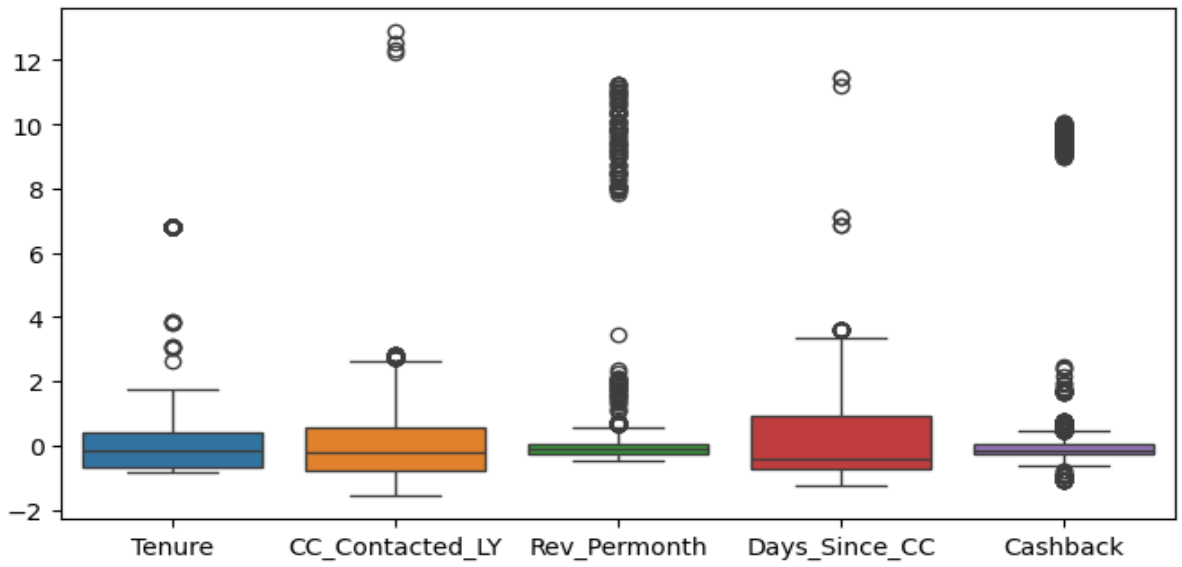


Fig 9.1 Before outlier treatment

The outliers have now been treated with lower and upper range. Plotting box plot after outlier treatment-

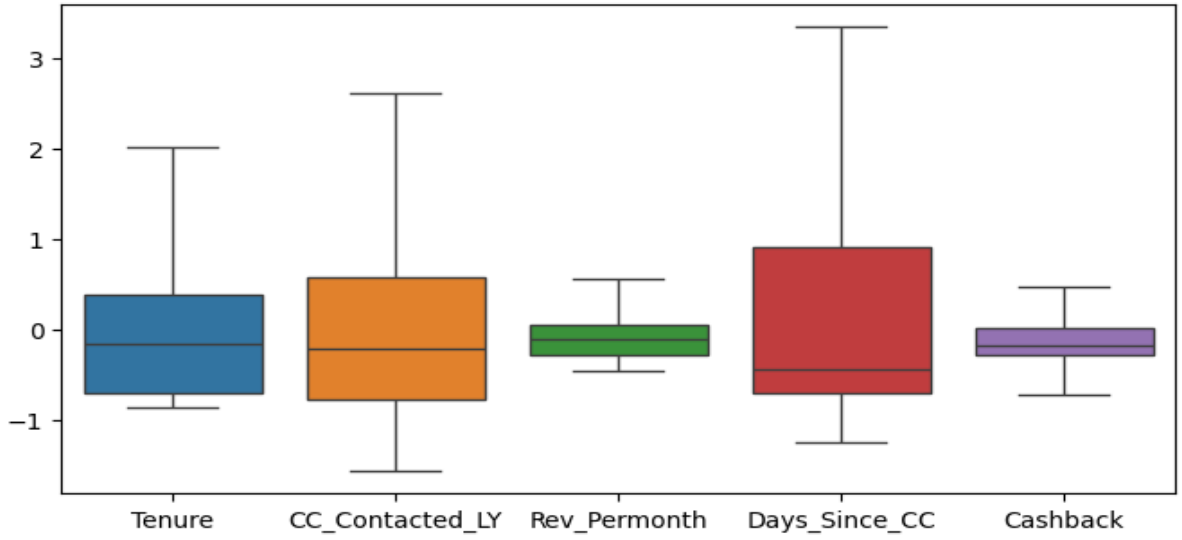


Fig 9.2 After outlier treatment

	Skewness
Tenure	0.80
CC_Contacted_LY	0.80
Rev_Permonth	0.78
Complain_LY	0.95
Days_Since_CC	0.82
Cashback	0.93

Observations-

- Some outliers for certain variables are closer to the whisker, whereas there are a group of outliers that are far beyond the whisker with no in between values.
- Two approaches to modelling were performed - one set of data with outliers treated for outlier sensitive models and another set of data with outliers not treated for outlier resistant algorithms.
- Coupon_used_for_payment has a very limited range 0 to 16. Hence, for the purpose of this analysis, the outliers will not be treated.

6. Clustering-

- Clustering is an unsupervised task and given all the features in the dataset, the clustering algorithm is allowed to group customers such that each group has similar customers and customers of different groups are dissimilar.
- For this purpose, the processed dataset (cleaned, scaled, nulls imputed, outlier treated, categorical features encoded) without the target variable was used.
- The algorithm was run for 2,3 and 4 clusters and 3 clusters seemed to have the best separation. The cluster profile was formed by grouping the observations by clusters and finding the mean for all the features.
- Although churn was not part of the features for clustering, it was added part of the cluster profile so that it is possible to appreciate how churn varies for each cluster.

kproto_3clusters	Churn	Tenure	City_Tier	CC_Contacted_LY	Complain_LY	Days_Since_CC
0	0.224396	10.669138	1.675718	30.626352	0.312352	3.846399
1	0.098399	13.556710	1.712985	14.992956	0.253772	8.880012
2	0.186470	9.535316	1.604142	13.368003	0.292514	2.212800

ACSegment_Regularplus	ACSegment_Super
0.330648	0.423092
0.135953	0.391849
0.535646	0.309122

7. Business Insights from EDA -

a) Business insights from cluster profiling-

- Even though this is an unsupervised algorithm without the use of target variable in clustering, the profiling came up with clear separation of groups only for those features that also showed a high F-statistic and high Chi square value in the bivariate analysis.
- Higher the tenure, lesser the churn. But the cluster profile also shows that for low tenures (cluster 1 and 0), this is not holding good.
- More Regular plus customers have a greater churn compared to least churn clusters.
- High churn clusters contacted customer care less times compared to low churn customers.
- The cluster with maximum complaints last year also has the maximum churn. The cluster with minimum complaints last year has the minimum churn.

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

- Dataset provided is an imbalance in nature. The categorical count of our target variable “Churn” shows high variation in counts.

- We need to apply SMOTE only to train the dataset not on the test dataset. divided data into train and test dataset in 70:30 ratio as an accepted market practice.

c) Other Business Insights-

- **Customer Feedback-** 78% of customers have rated service as 3 or less than 3 (out of a scale of 5). Likewise, 61% of customers have rated customer care agents a score of 3 or less than 3.
- **Relationship between Tenure and Churn-** In the bivariate histogram for Tenure vs Churn, it can be seen that the churn is very high for low tenures.
- **Relationship between Account segment and Churn-** More customers in Regular_Plus plan seem to churn.
- **Relationship between Monthly revenue and Churn-** The % churn in high revenue customers is slightly more than the churn in lower revenue customers.
- **Relationship between Days since customer care connect and Churn-** Days since customer care connect for churned customers is lesser than for active customers. This shows that churn has happened shortly after the customers have contacted customer care.
- **Relationship between Customer care contacted last year and Churn-** The number of times customer care was contacted previous year was more in churned customers compared to active customers.
- **Relationship between Payment type and Churn-** Proportion of customers who have paid through E-wallet and Cash on delivery is more within churned customers compared to active customers.
- **Relationship between City tier and Churn-** Proportion of customers who reside in Tier 3 cities is more within churned customers compared to active customers.
- **Relationship between Marital status and Churn-** Single customers have churned more compared to married or divorced customers.

The End…