

Capstone Project: Notes-1 Report

By

Ritusri Mohan

CONTENTS

PROBLEM UNDERSTANDING.....	3
<i>Problem Statement</i>	3
<i>Objective</i>	3
DATA REPORT.....	
<i>Data Dictionary</i>	4
<i>About the dataset</i>	4
EXPLORATORY DATA ANALYSIS & INSIGHTS.....	6
<i>Descriptive Summary</i>	6
<i>Missing Values Detection & Treatment</i>	8
<i>Outliers Detection & Treatment</i>	9
<i>Univariate Analysis</i>	11
<i>Bivariate Analysis</i>	17
<i>Multiivariate Analysis</i>	21
BUSINESS INSIGHTS	23

PROBLEM UNDERSTANDING

Problem Statement

Direct-to-Home (DTH) television is a method of receiving satellite television by means of signals transmitted from direct-broadcast satellites. A DTH 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.

Objective

By anticipating prospective churn candidates and taking proactive measures to keep them, the goal is to lower customer churn for the DTH service provider.

Customer churn is a reflection of how fast the customers terminate their services or switch to another service provider.

The general belief is that acquiring new customers is more expensive than retaining existing ones.

There are various factors that lead customers to leave their current DTH provider:

- Network problems
- Billing problems
- Bad customer support experience
- Competitive offers from another DTH service provider
- Overage charge

DATA REPORT

Data Dictionary

Variable	Description
AccountID	Account unique identifier
Churn	Account churn flag (Target)
Tenure	Tenure of account
City_Tier	Tier of primary customer's city
CC_Contacted_LY	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_ly	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_for_payment	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_112m	Monthly average cashback generated by account in last 12 months
Login_device	Preferred login device of the customers in the account

About the dataset

	AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	I
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	

- The head of the dataset is shown above.
- On checking the shape of the dataset, it is found that there are 11260 rows and 19 columns.

- From the below information it can be seen that there are 12 object type variables, 5 float type variables and 2 integer type variable.

```

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

```

- The unit for ‘Tenure’ was taken as months.
- The ‘City_Tier’ variable’s categories were taken in terms of population.

City_Tier	Population
1	1,00,000 and more
2	50,000 to 99,999
3	less than 50,000

- The unit for ‘cashback’ was taken as rupees.

EXPLORATORY DATA ANALYSIS & INSIGHTS

- The 'AccountID' variable was dropped from the dataset as it was not related to the other variables from the analysis perspective.
- After dropping the column, dataset was found to have 11260 rows and 18 columns.

Descriptive Summary

Count	Unique	Top	Freq	Mean	Std	Min	25%	50%	75%	max	
Churn	11260	NaN	NaN	NaN	0.168384	0.374223	0	0	0	0	1
Tenure	11158	38	1	1351	NaN	NaN	NaN	NaN	NaN	NaN	NaN
City_Tier	11148	NaN	NaN	NaN	1.65393	0.915015	1	1	1	3	3
CC_Contacted_LY	11158	NaN	NaN	NaN	17.8671	8.85327	4	11	16	23	132
Payment	11151	5	Debit Card	4587	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	11152	4	Male	6328	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Service_Score	11162	NaN	NaN	NaN	2.90253	0.725584	0	2	3	3	5
Account_user_count	11148	7	4	4569	NaN	NaN	NaN	NaN	NaN	NaN	NaN
account_segment	11163	7	Super	4062	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CC_Agent_Score	11144	NaN	NaN	NaN	3.06649	1.37977	1	2	3	4	5
Marital_Status	11048	3	Married	5860	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rev_per_month	11158	59	3	1746	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Complain_ly	10903	NaN	NaN	NaN	0.285334	0.451594	0	0	0	1	1
rev_growth_yoy	11260	20	14	1524	NaN	NaN	NaN	NaN	NaN	NaN	NaN
coupon_used_for_payment	11260	20	1	4373	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Day_Since_CC_connect	10903	24	3	1816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cashback	10789	321	152	208	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Login_device	11039	3	Mobile	7482	NaN	NaN	NaN	NaN	NaN	NaN	NaN

- On checking the dataset, it was seen that the following variables had wrong values for some of the records and therefore, these wrong values were replaced with suitable values.

✚ In the 'Gender' variable, 'M' and 'F' were replaced by 'Male' and 'Female' respectively.

✚ In the 'account_segment' variable, 'Regular +' and 'Super +' were replaced by 'Regular Plus' and 'Super Plus' respectively.

✚ In the 'Login_device' variable, '&&&&' was replaced by 'No_info'.

✚ In Numeric variables like 'Account_user_count', 'Tenure', 'rev_per_month', 'rev_growth_yoy', 'Day_Since_CC_connect', 'coupon_used_for_payment' and 'cashback' the wrong values were corrected using the coerce function which replaces these vales with 'Nan' value(of numeric data type).

- From the below information it can be seen that there are 12 object type variables, 5 float type variables and 1 integer type variable.

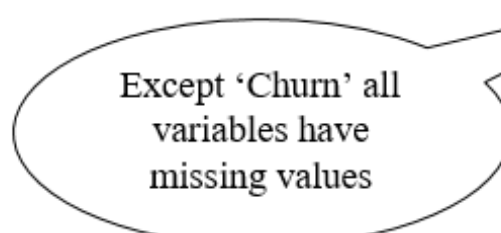
Data columns (total 18 columns):			
#	Column	Non-Null Count	Dtype
0	Churn	11260 non-null	int64
1	Tenure	11158 non-null	object
2	City_Tier	11148 non-null	float64
3	CC_Contacted_LY	11158 non-null	float64
4	Payment	11151 non-null	object
5	Gender	11152 non-null	object
6	Service_Score	11162 non-null	float64
7	Account_user_count	11148 non-null	object
8	account_segment	11163 non-null	object
9	CC_Agent_Score	11144 non-null	float64
10	Marital_Status	11048 non-null	object
11	rev_per_month	11158 non-null	object
12	Complain_ly	10903 non-null	float64
13	rev_growth_yoy	11260 non-null	object
14	coupon_used_for_payment	11260 non-null	object
15	Day_Since_CC_connect	10903 non-null	object
16	cashback	10789 non-null	object
17	Login_device	11039 non-null	object
dtypes: float64(5), int64(1), object(12)			
memory usage: 1.5+ MB			

- From the table below, it can be seen that the Numeric data is skewed.

Churn	1.772606
Tenure	3.895707
City_Tier	0.737107
CC_Contacted_LY	1.422977
Service_Score	0.003891
Account_user_count	-0.393100
CC_Agent_Score	-0.142149
rev_per_month	9.093909
Complain_ly	0.950876
rev_growth_yoy	0.752474
coupon_used_for_payment	2.575199
Day_Since_CC_connect	1.273021
cashback	8.771302
dtype: float64	

Missing Values Detection & Treatment

- The total number of missing values in the dataset was found to be 3822. The count of the missing values in each column can be seen below:



	count
Churn	0
Tenure	218
City_Tier	112
CC_Contacted_LY	102
Payment	109
Gender	108
Service_Score	98
Account_user_count	444
account_segment	97
CC_Agent_Score	116
Marital_Status	212
rev_per_month	791
Complain_ly	357
rev_growth_yoy	3
coupon_used_for_payment	3
Day_Since_CC_connect	358
cashback	473
Login_device	221

- By dividing the total number of missing values by the total number of suitable rows, the proportion of missing data was ranked. Notably, none of the columns were eliminated as the value is not more than 30% (i.e. 0.3).

rev_per_month	0.070249
cashback	0.042007
Account_user_count	0.039432
Day_Since_CC_connect	0.031794
Complain_ly	0.031705
Login_device	0.019627
Tenure	0.019361
Marital_Status	0.018828
CC_Agent_Score	0.010302
City_Tier	0.009947
Payment	0.009680
Gender	0.009591
CC_Contacted_LY	0.009059
Service_Score	0.008703
account_segment	0.008615
rev_growth_yoy	0.000266
coupon_used_for_payment	0.000266
Churn	0.000000
dtype: float64	

- The missing values were treated as follows:

✚ Variables like- 'Gender', 'Marital_Status', 'Login_device', 'account_segment', 'Payment' were imputed with string value 'No_info' as they are categorical in nature.

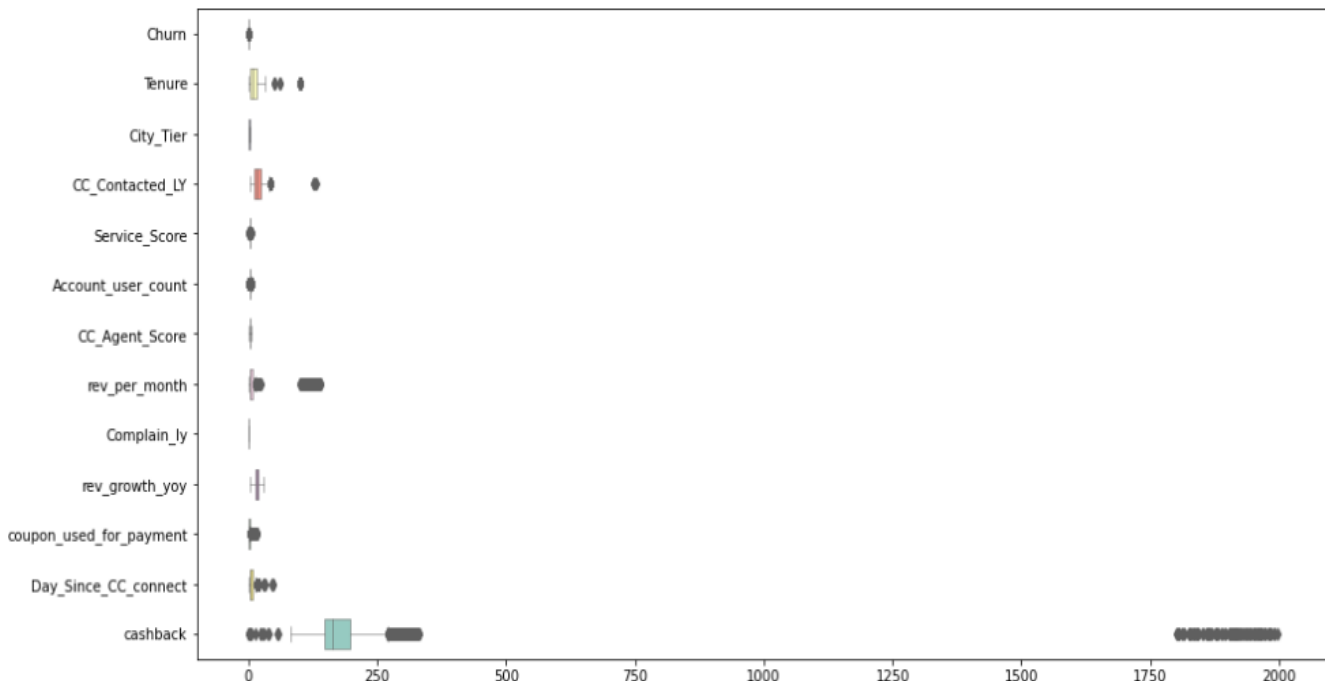
✚ Variables like- 'Account_user_count', 'Service_Score', 'Tenure', 'CC_Agent_Score', 'CC_Contacted_LY', 'Complain_ly', 'rev_per_month', 'cashback', 'rev_growth_yoy', 'coupon_used_for_payment', 'City_Tier', 'Day_Since_CC_connect', were imputed with their respective mode values.

The mode imputation method was adopted as it is considered as a good approach for skewed data.

	count
Churn	0
Tenure	0
City_Tier	0
CC_Contacted_LY	0
Payment	0
Gender	0
Service_Score	0
Account_user_count	0
account_segment	0
CC_Agent_Score	0
Marital_Status	0
rev_per_month	0
Complain_ly	0
rev_growth_yoy	0
coupon_used_for_payment	0
Day_Since_CC_connect	0
cashback	0
Login_device	0

Outliers Detection & Treatment

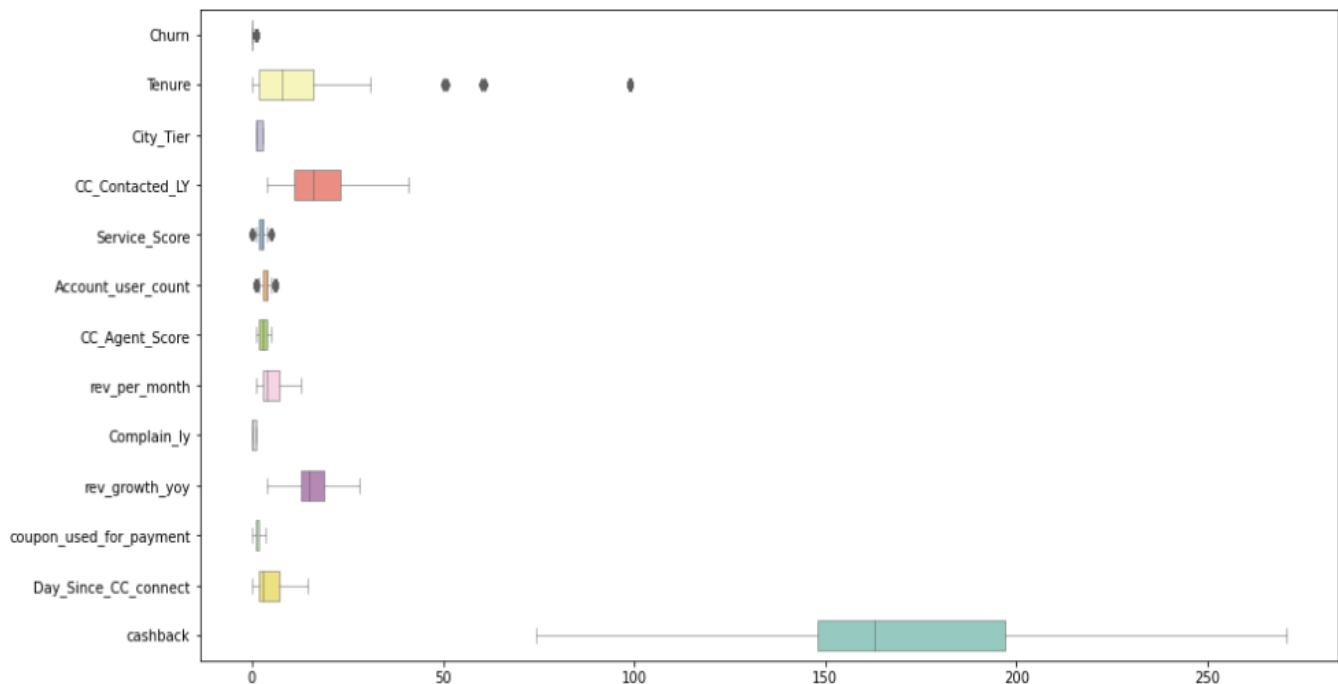
A visual representation of the outliers can be seen below.



The IQR approach was applied to identify outliers which involves creating a new range known as a decision range, and any data point that lies outside of it is regarded as an outlier. The scope is provided below.

$$\begin{aligned} \text{IQR} &= Q3 - Q1 \\ \text{Lower Bound} &= Q1 - 1.5 * \text{IQR} \\ \text{Upper Bound} &= Q3 + 1.5 * \text{IQR} \end{aligned}$$

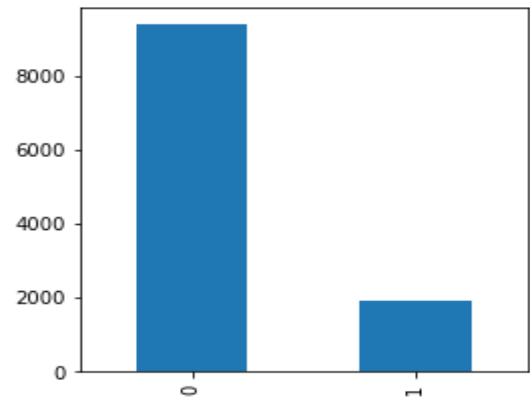
A visual representation of the outliers after treatment can be seen below. The outliers for 'Tenure', 'Service_Score' and 'CC_Agent_Score' are not treated because they are significant.



Univariate Analysis

- Churn (Target variable)

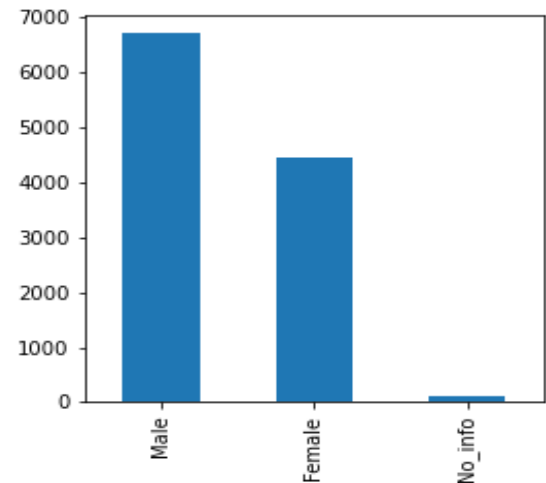
- ✚ This variable has two values: 0 & 1.
 - 0 denotes Non-Churn
 - 1 denotes Churn
- ✚ The data has higher number of Non-churn customers i.e. the customers are active.
- ✚ Non-Churn customers: 83.16%
- ✚ Churn customers: 16.84%
- ✚ Therefore, rate of churning = 16.84%



- Gender

The percentage distribution of :

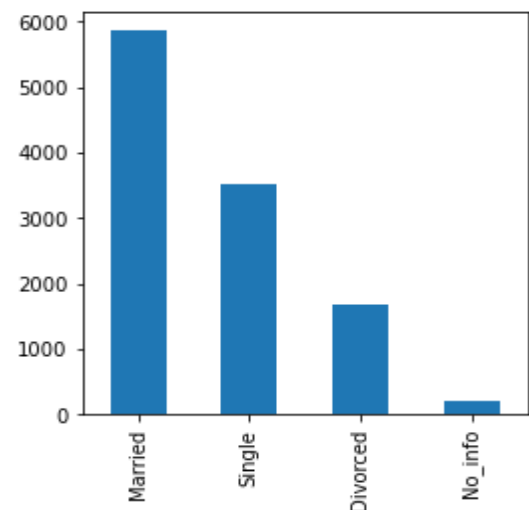
- ✚ Males: 59.5%
- ✚ Females: 39.5%
- ✚ No Information: 1%



- Marital_Status

The percentage distribution of:

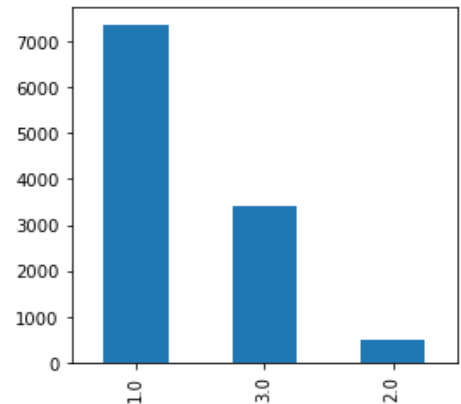
- ✚ Married customers: 52.04%
- ✚ Single customers: 31.26%
- ✚ Divorced customers: 14.81%
- ✚ No information: 1.89%



- City_Tier

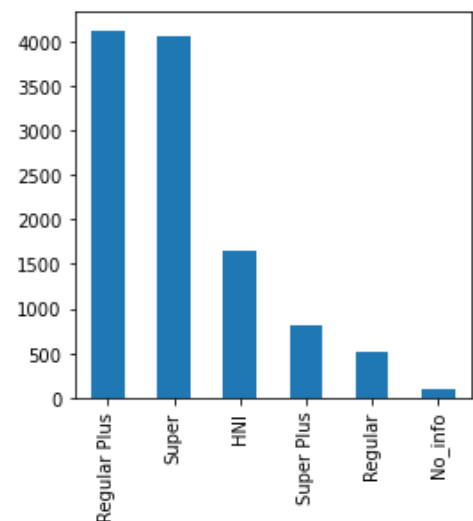
The percentage of customers from

Tier 1 > Tier 3 > Tier 2



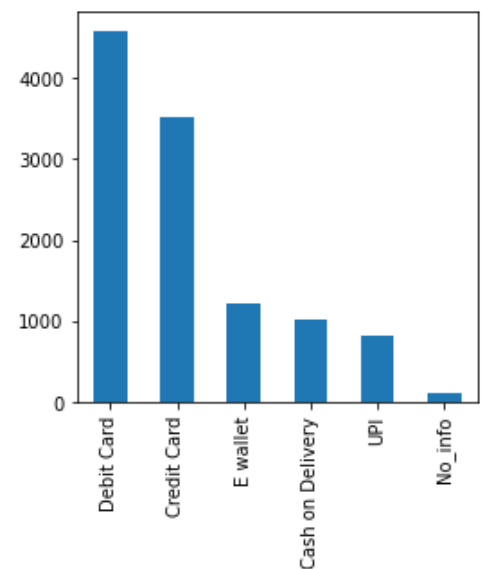
- Account_segment

Majority customers belong to Regular Plus and Super segment followed by category HNI (High net worth individuals), Super Plus and Regular.



- Payment

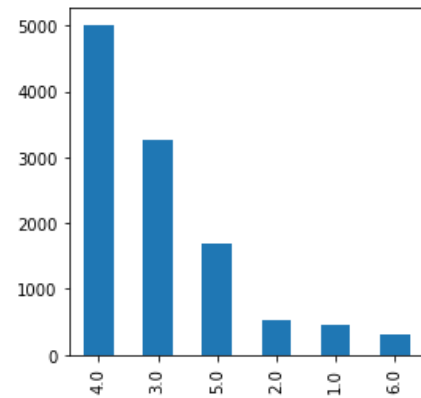
Majority customers prefer Debit card method of payment over Credit card, E wallet etc.



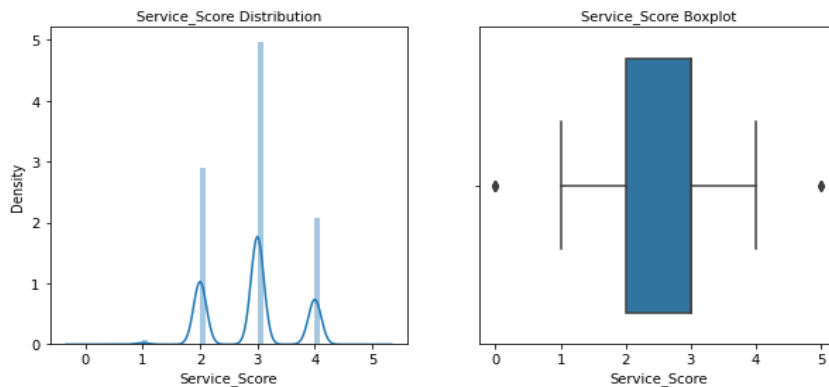
- Account_user_count

It can be seen that for maximum accounts 4 people are attached to each. It can be assumed that these are mainly nuclear families.

Minimum number of accounts have 6 users each attached to them. These are considered as joint families.

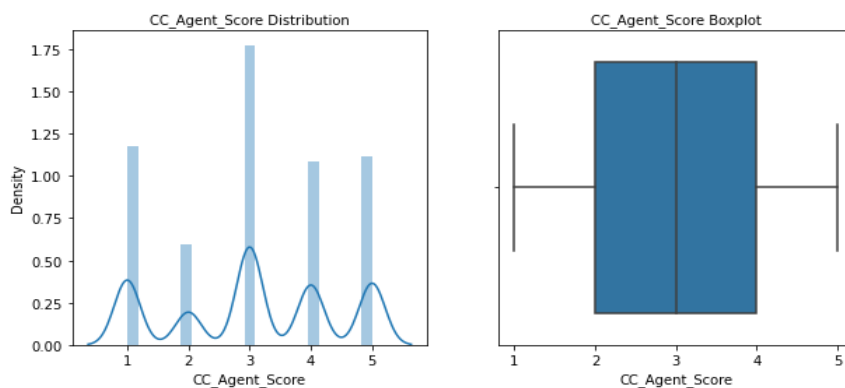


- Service_Score



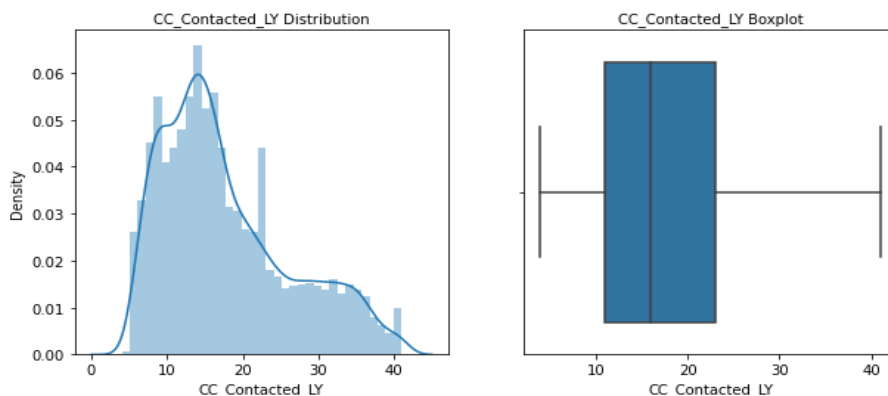
The distribution seems to be normal with barely any outliers. It can be seen that most of the customers have given a rating of 3 out of 5.

- CC_Agent_Score



The distribution seems to be normal without any outliers. It can be seen that most of the customers have given a rating of 3 out of 5.

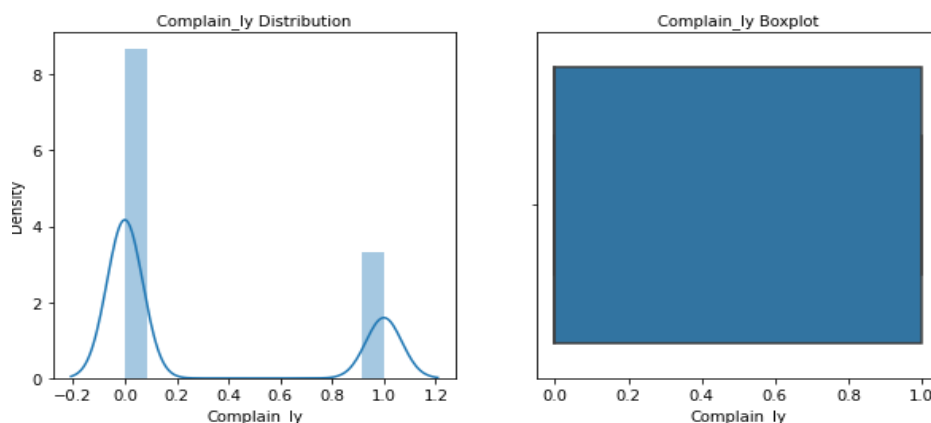
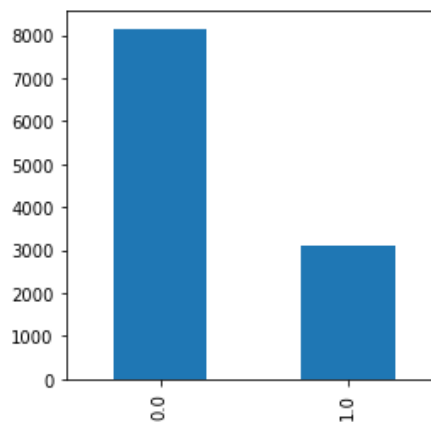
- CC_Contacted_LY



The distribution seems to be moderately skewed in a positive manner without any outliers. The frequency of customer contact in last 12 months varies from few to many as per the customers.

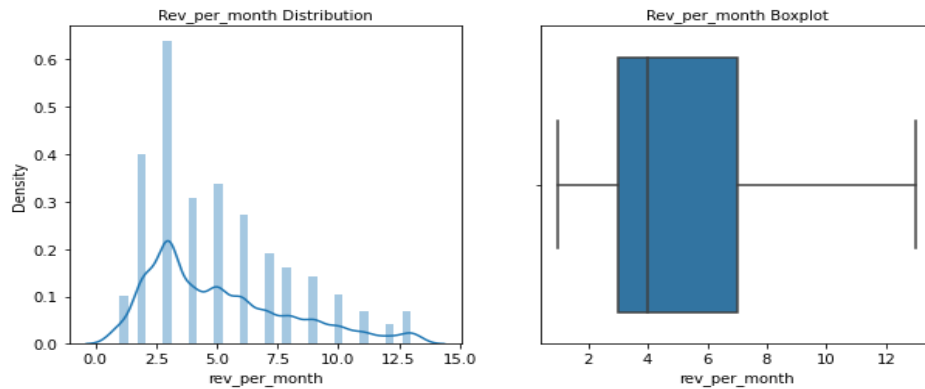
- Complain_ly

- ✚ This variable has two values: 0 & 1.
0 denotes: 'No complaints'
1 denotes: 'Yes complaints'
- ✚ The data has higher number of No complaining customers i.e. these customers are happy with the service.
- ✚ Non Complaining customers: 72.37%
Complaining customers: 27.63%



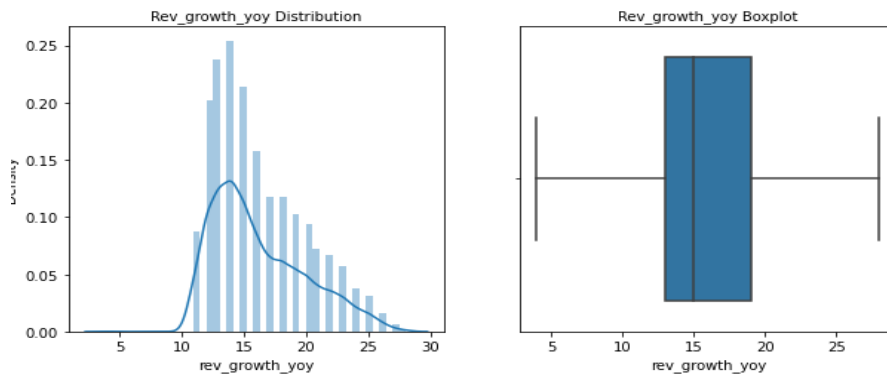
The data is positively skewed without outliers. No whiskers are visible because of the extreme skewness.

- Rev_per_month



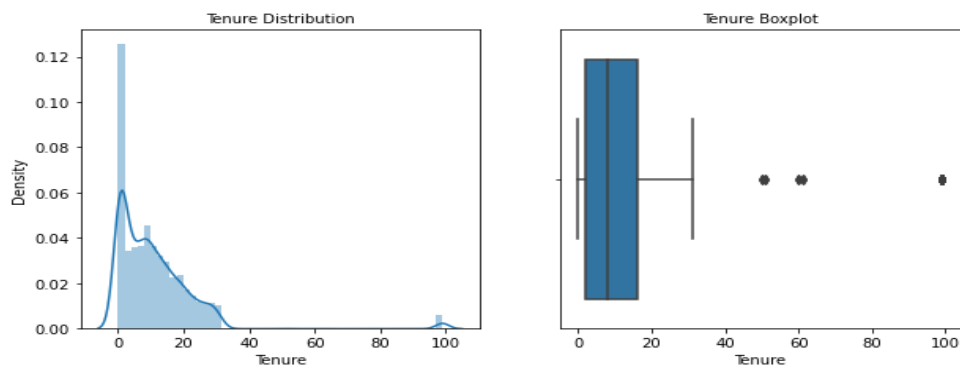
The distribution seems to be moderately skewed in a positive manner without any outliers.

- Rev_growth_yoy



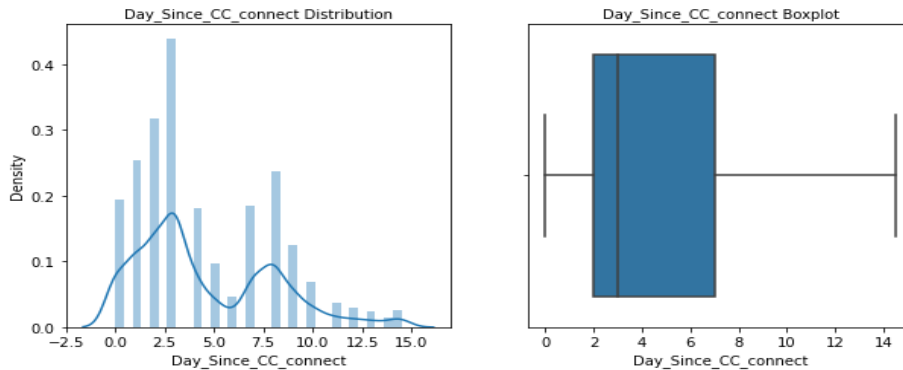
The distribution seems to be moderately skewed in a positive manner without any outliers. It seems that 50% of the times there has been high growth in revenue.

- Tenure



The data is positively skewed and it can be seen that most customer accounts have a tenure between 0 to 20 months.

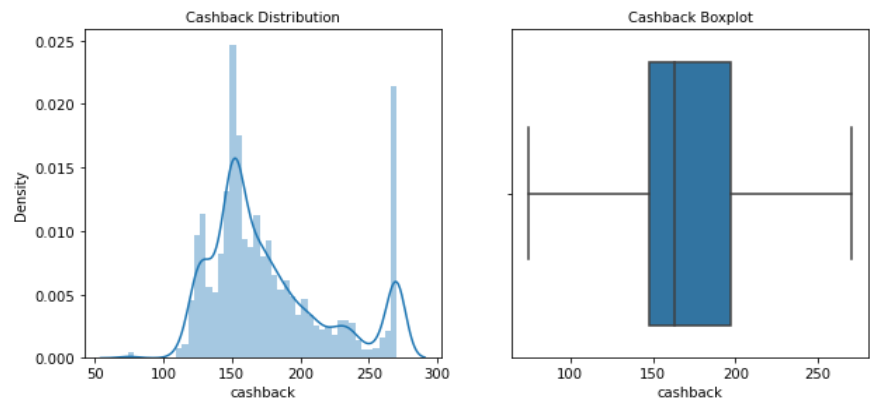
- Day_Since_CC_connect



The distribution seems to be moderately skewed in a positive manner without any outliers. It seems that maximum number of days since the customer connect lies between 3 to 5 days.

- Cashback

The distribution seems to be moderately skewed in a positive manner without any outliers. The maximum frequency of monthly cashback is for between rupees 150 and 175.



- Coupon_used_for_payment

The distribution seems to be approximately symmetric without any outliers

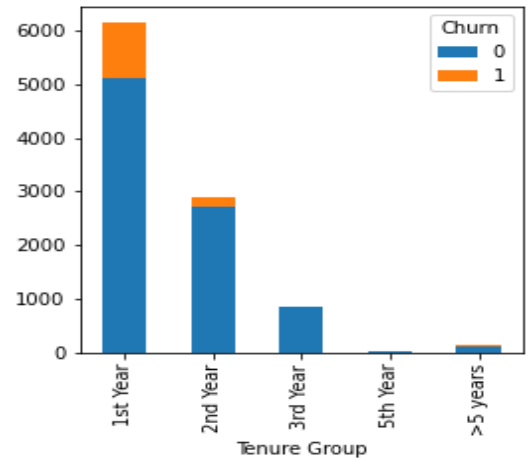


Bivariate Analysis

- Tenure Vs Churn

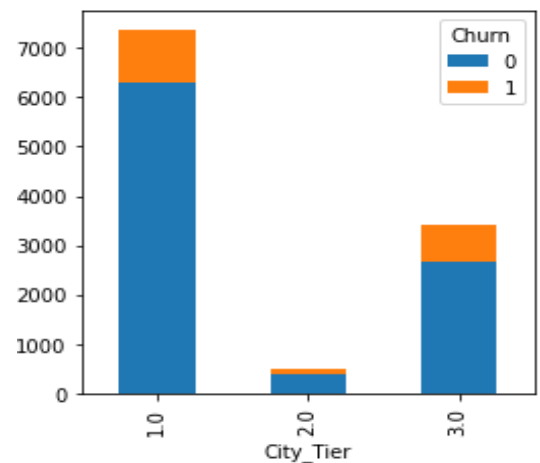
A new variable called 'Tenure Group' was created in which the months were grouped into bins (of size 12 months)

It can be seen that the highest number of customers to churn belong to the 1st year (0 to 12 months).



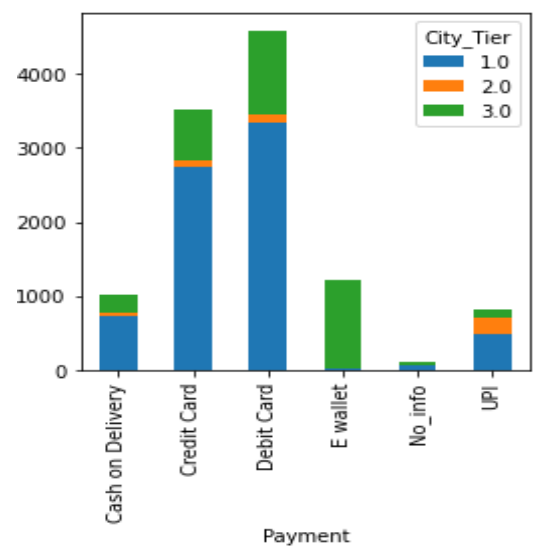
- City Tier Vs Churn

It can be seen that the number of customers to churn are from Tier 1 cities followed by Tier 2.



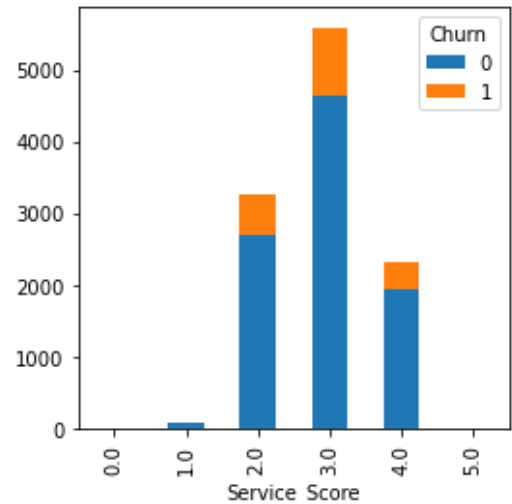
- Payment Vs City Tier

It can be seen that tier 1 customers mainly prefer debit card and credit card payment; customers from tier 2 cities prefer UPI and tier 3 customers prefer E wallet method.



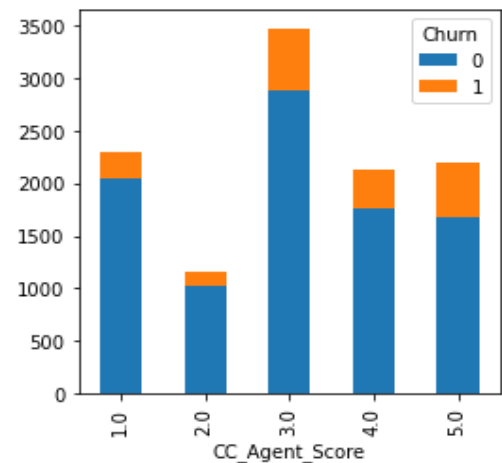
- Service Score Vs Churn

It can be seen that the maximum number of customers to churn have given a score between 2 and 3 and highest number of non- churn customers have given a rating of 3.



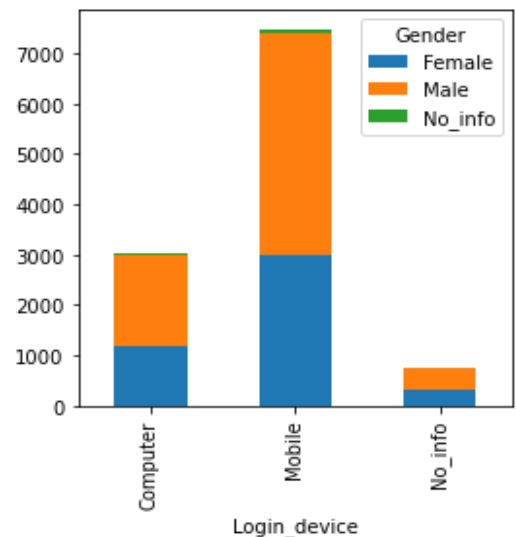
- CC Agent Score Vs Churn

It can be seen that the maximum number of customers to churn have given a score between 2 and 3 and highest number of non-churn customers have given a rating between 3 and 5.



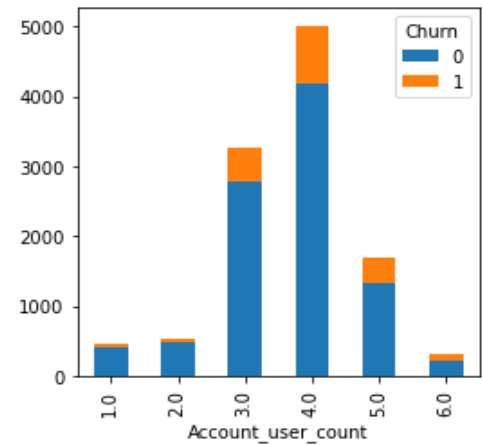
- Login device Vs Gender

Compared to female customers, male customers are more and in terms of login device mobile log in is greater.



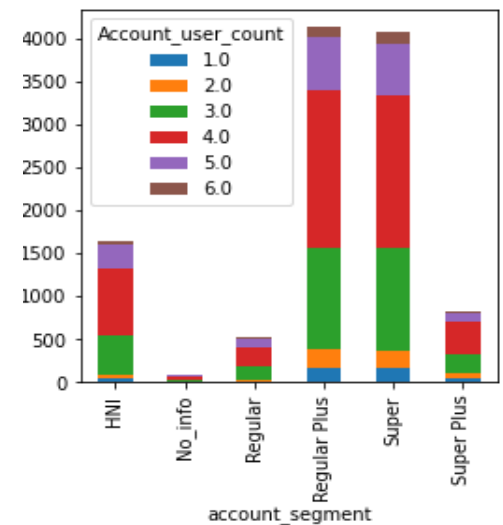
- Account user count Vs Churn

More number of accounts to churn are those who have 3-4 users attached to it including the customer. This is probably due to technical issue in multiple logins or high package price.



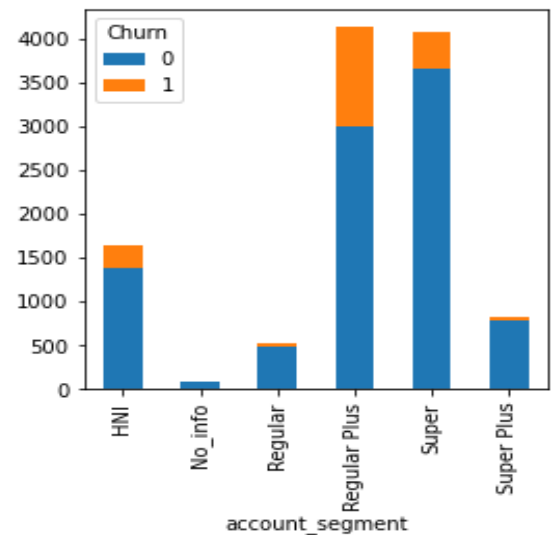
- Account segment Vs Account user count

It can be seen that across all the account segments the maximum number of accounts are those which have 3 to 4 users (nuclear families) attached to them



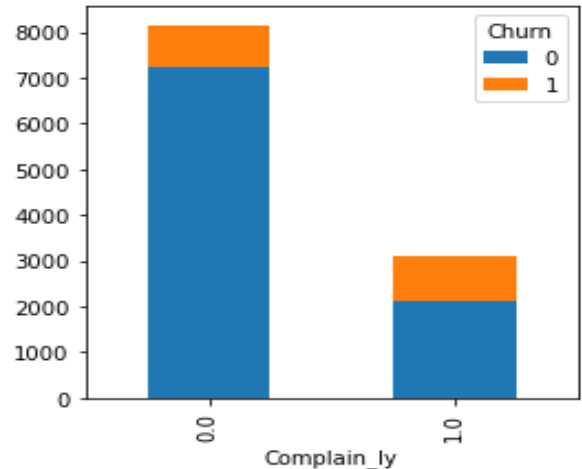
- Account segment Vs Account user count

It can be seen that most customers belong to Regular Plus and Super segment. Maximum churn customers are those who lie in the Regular Plus followed by Super.



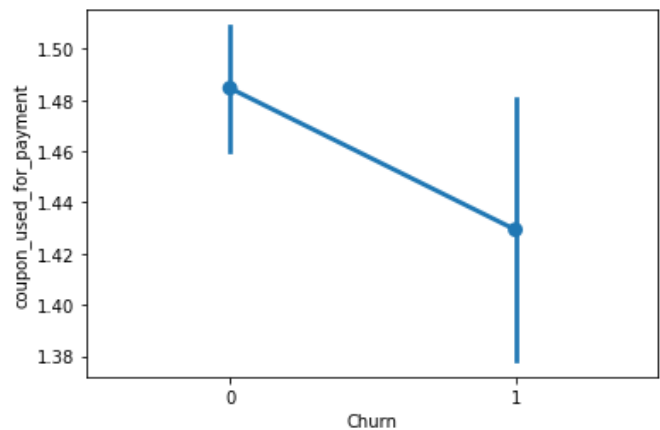
- Complain last year Vs Account user count

It can be seen that maximum count of churn customers belongs to the positive category of complaints (values 1.0 on x-axis).



- Coupon used for payment Vs Churn

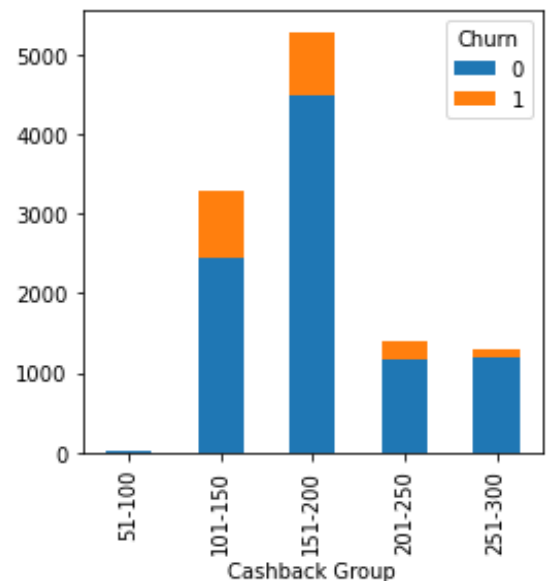
More number of churn customers are those who have low received very few coupons for payment.



- Cashback used for payment Vs Churn

A new variable called 'Cashback Group' was created in which the monthly average cashbacks were grouped into bins (of size 50 rupees).

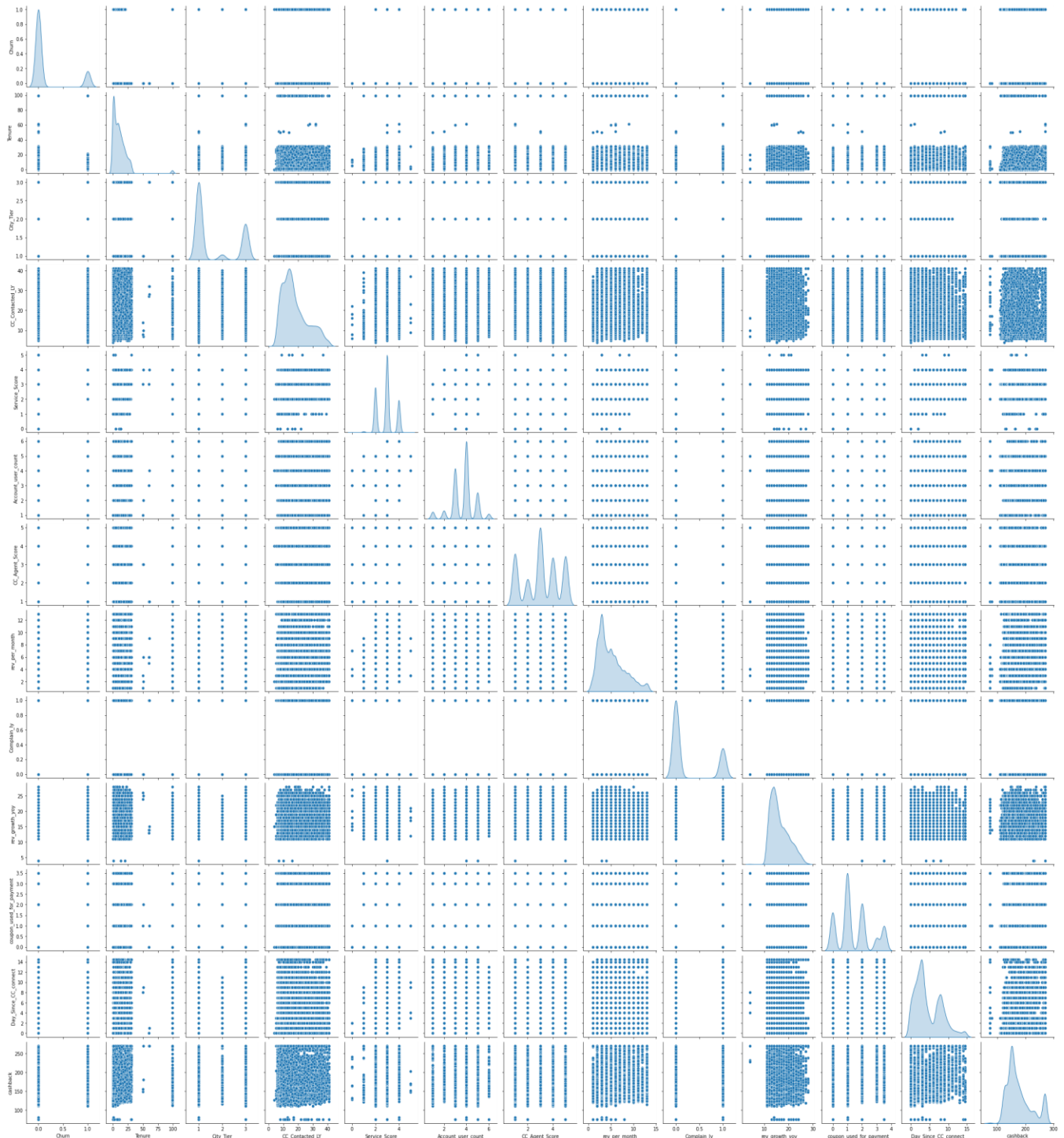
More number of churn customers are those who receive a cashback of 101-150 rupees. Maximum number of non-churn customers receive cashback of 151- 200.



Multivariate Analysis

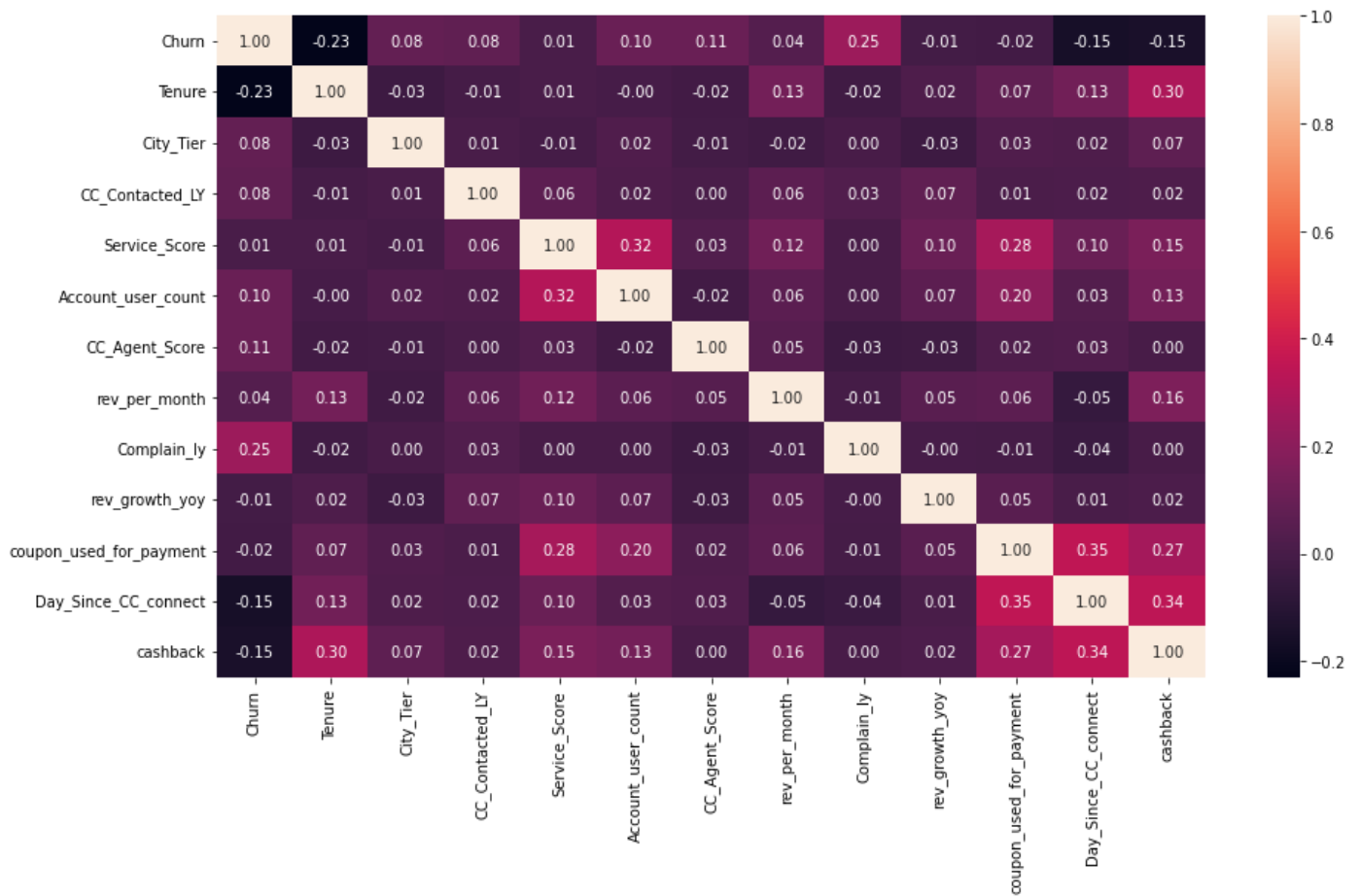
- Pairplot

The pairplot shows the relationship between various pairs of variables.



- Heatmap

From the below heatmap it was seen that there are weak correlations between various variables.



BUSINESS INSIGHTS

- ✚ Customers who rate their satisfaction higher than 3 should be solicited for higher levels of service and product satisfaction.
- ✚ Customer service should come first and complaints should be handled seriously by the business.
- ✚ Competitive marketing tactics has to be taken into account.
- ✚ New clients are churning within the first 12 months. It is very important to keep track of their subscription and extra attention should be paid to their needs.
- ✚ It is important to conduct market research to offer reasonable prices and promotions.
- ✚ Loyalty programmes should be introduced to reward and encourage devoted customers.
- ✚ So that the entire family can benefit from a single account, family plans should be made affordable.
- ✚ All members enjoy flexible plans which will give them the ability to alter their plans as they see fit and the freedom to suspend and cancel subscriptions.

THE END