1.INTRODUCTION

E-commerce, the activity of buying and selling products online, is one of the many fields revolutionized by data science. One of the essential goals for e-commerce companies is to understand user behaviour.

For this project, we have data set belongs to a leading online E-Commerce company. An online retail E-commerce company wants to know the customers who are going to churn, so accordingly they can approach customer to offer some promos.

Five major reasons for customer retention are:

- Companies save money on marketing.
- Repeat purchases from repeat customers means repeat profit.
- Free word-of-mouth advertising.
- Retained customers provide valuable feedback.
- Previous customers will pay premium prices.

Hence, it is very important for Companies to retain customers and try to find the reasons for customer attrition. If attrition rate is higher than, businesses should try to premium services and appropriate discounts so that customers continue with the company. Long term customers are an important asset to every company and hence, it is important for service providers to provide best services to their customers.

A). DEFINING PROBLEM STATEMENT

The objective of this case study is to understand the E commerce industry and the challenges they are facing in order to retain a particular customer due to high amount of competition. Hence, we need to help the company to predict customer churn and help them reduce the number of customers on the basis of data provided. The data provided comprises of details of various customers with their personal and other details.

Identifying customer churn by E commerce companies is of utmost importance since it helps them to reduce losses. If customer attrition rate is higher, it directly reflects the company's image. Therefore, the main aim of this project is to predict the customer churn by building machine learning models and hence identify churn rate. By carefully scrutinizing the reasons for churn, the company can promote their services and introduce promotional strategies. The major reasons for customer retention are:

- Lack of engagement
- Poor product-market fit
- Product Bugs
- Difficult User Experience
- Lack of Proactive Support

B). NEED OF THE STUDY/PROJECT

Indian E-commerce market is expected to grow to US$ 200 billion by 2026 from US$ 38.5 billion as of 2017. Much of the growth for the industry has been triggered by an increase in internet and smartphone penetration. As of August 2020, the number of internet connections in India significantly increased to 760 million, driven by the 'Digital India' programme. Propelled by rising smartphone penetration, launch of 4G network and increasing consumer wealth, online retail sales in India is expected to grow 31% to touch US$ 32.70 billion in 2018, led by Flipkart, Amazon India and Paytm Mall.
Even though India is a huge market for E-commerce but still many companies are not able to retain their customers.
Therefore, the need of this study is:

- To avoid customer churn
- To identify riskier customers
- To determine risk propensity faced by the company
- To reduce losses faced by companies due to customer attrition

C). UNDERSTANDING BUSINESS/SOCIAL OPPORTUNITY

Customer churn greatly affects a company as they incur high number of losses which directly impacts the company's revenue. Retaining long term customers is essential for any company to create an image. If attrition of customers is continued for any company, it signifies poor service, better prices or services offered by competitors or difficult user interface. There is a need to create a platform for analysis and coming forward with appropriate recommendations of promotional strategies and better customer service.

2.DATA REPORT

A). UNDERSTANDING HOW DATA WAS COLLECTED

The data is given to us as part of this project which is a dataset of an online E-commerce company who are losing customers. As part of the marketing department of the online retail company we need to predict the churn status effectively and help the company to retain their valuable customers.

B). VISUAL INSPECTION OF DATA (ROWS, COLUMNS, DESCRIPTIVE DETAILS)

DATA AND DATA DICTIONARY

The Dataset provided in the case study is stored as **"Customer Churn Data.xlsx"**. The variables of the dataset are:

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

DATA OVERVIEW

```
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   AccountID               11260 non-null  int64
 1   Churn                   11260 non-null  int64
 2   Tenure                  11042 non-null  float64
 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      10816 non-null  float64
 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           10469 non-null  float64
 13  Complain_ly             10903 non-null  float64
 14  rev_growth_yoy          11257 non-null  float64
 15  coupon_used_for_payment 11257 non-null  float64
 16  Day_Since_CC_connect    10902 non-null  float64
 17  cashback                10787 non-null  float64
 18  Login_device            10500 non-null  object
dtypes: float64(12), int64(2), object(5)
```

C). UNDERSTANDING OF ATTRIBUTES (VARIABLE INFO, RENAMING IF REQUIRED)

DATA INFORMATION

The variable details are as follows:

- Churn – This variable is a binary variable with output as 1 or 0. This will be our **Target Variable**.
- Binary Variables namely “Gender”, “Complain_l12m” have only two outputs.
- Numerical variables namely “Tenure”, “CC_Contacted_l12m”, “Account_user_count”,
  “rev_per_month”, “rev_growth_yoy”, “coupon_used_l12m”,

“Day_Since_CC_connect” and “cashback_l12m” are numerical or continuous variables.

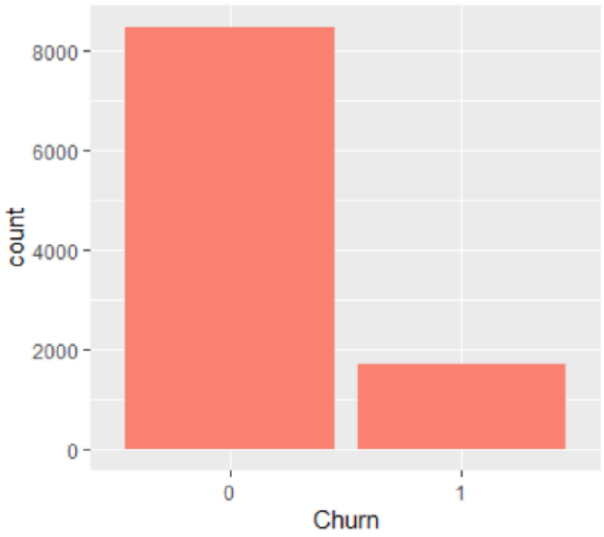
- Variables “City_Tier”, “Payment”, "Service_Score"," account_segment", "CC_Agent_Score", "Marital_Status" and "Login_device" are categorical.

- We will convert the binary and categorical variables and Account_user_count, rev_per_month and cashback to numerical variable.

EXPLORATORY DATA ANALYSIS

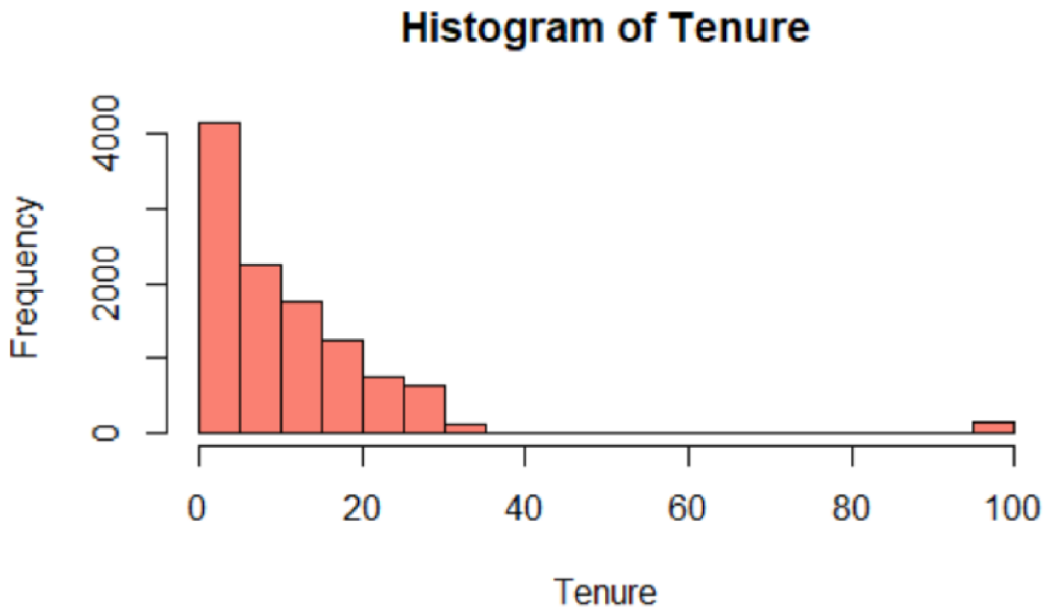
1. UNIVARIATE ANALYSIS AND BIVAIATE ANALYSIS

1. CHURN



- This is a binary variable and our target variable.
- It has values as 0: Not churned and 1: Churned customers.
- The data is unbalanced as 16.74% churned customers and remaining 83.26% non-churned customers.
- For further processing, it will be important to balance the data using SMOTE as the dataset is highly unbalanced.

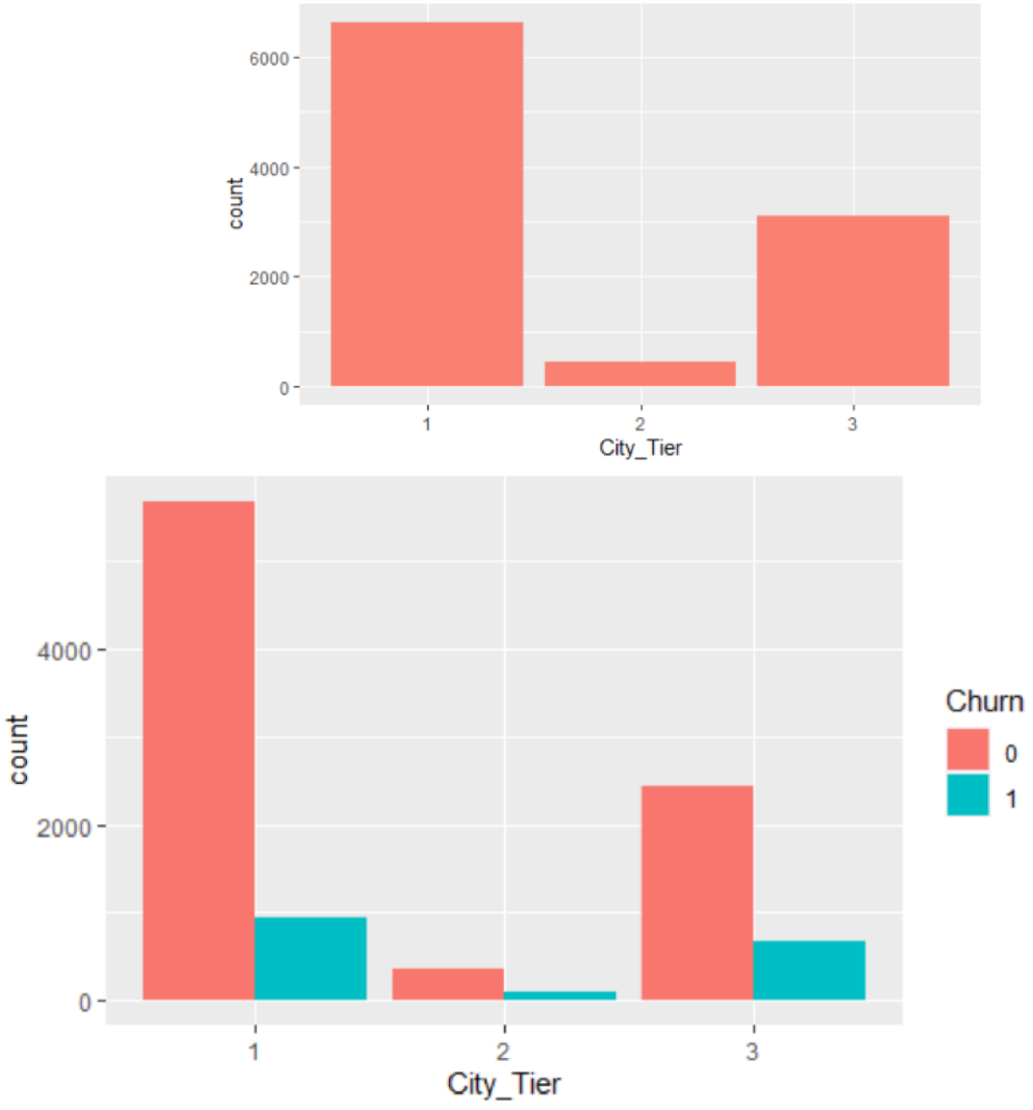
2. TENURE



Inference

- This is a type of numerical variable which is discrete in nature. It tells us for how many months a customer was retained with the company.
- The maximum tenure of customers were 0-5 months whereas average being 11 years.
- As can be seen from the above graph, that outliers are present since 2% of the data is spread over 60 months.
- As per the Bivariate analysis, we have compared the Tenure with our target variable – Churn.
- As per the graph, it can be understood that majority of "Churned" have their Tenure between 0-2 months with a few outliers present. The customers that are retained with the company range from 0-32 months with outliers present.
- However, the number of Churned customers is far less than the number of retained customers.

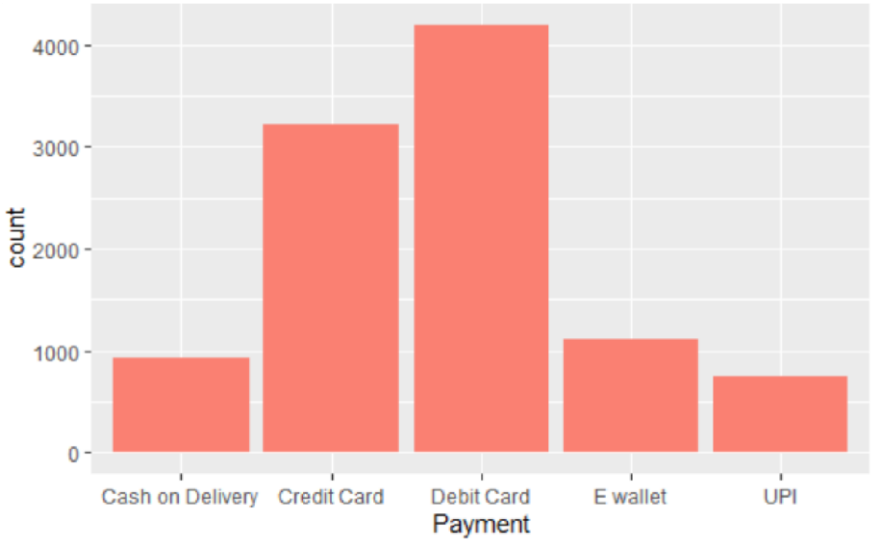
3. CITY TIER

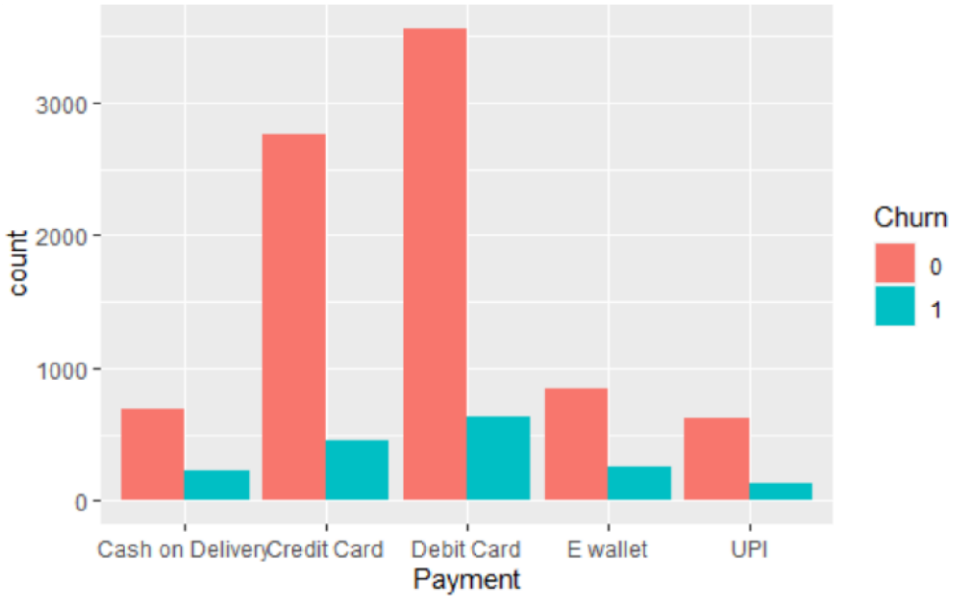


Inference

- This variable is a categorical variable with three values Tier 1, 2 and 3.
- As per the graph we can conclude that out of the total customers approximately 58% of the customers belong to Tier 1 cities, followed by 22.2% in Tier 3 and the remaining in Tier 2.
- As per the Bivariate analysis, we have compared the City Tier with our target variable – Churn.
- As per the graph, it can be understood that majority of churned customers belong to Tier 1 Cities followed by Tier 3 cities.
- However, the number of Churned customers is far less than the number of retained customers.

4. PAYMENT

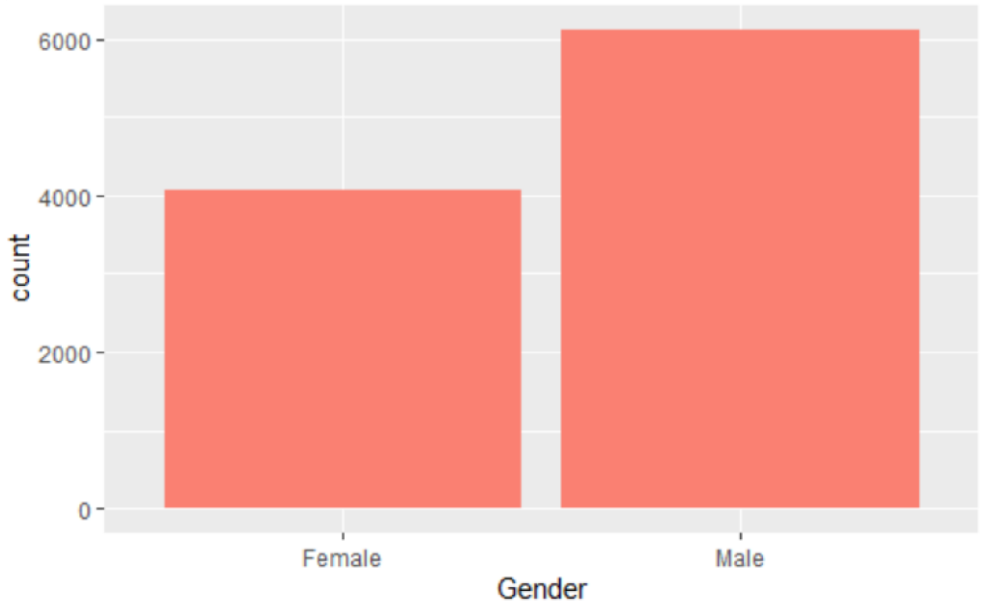


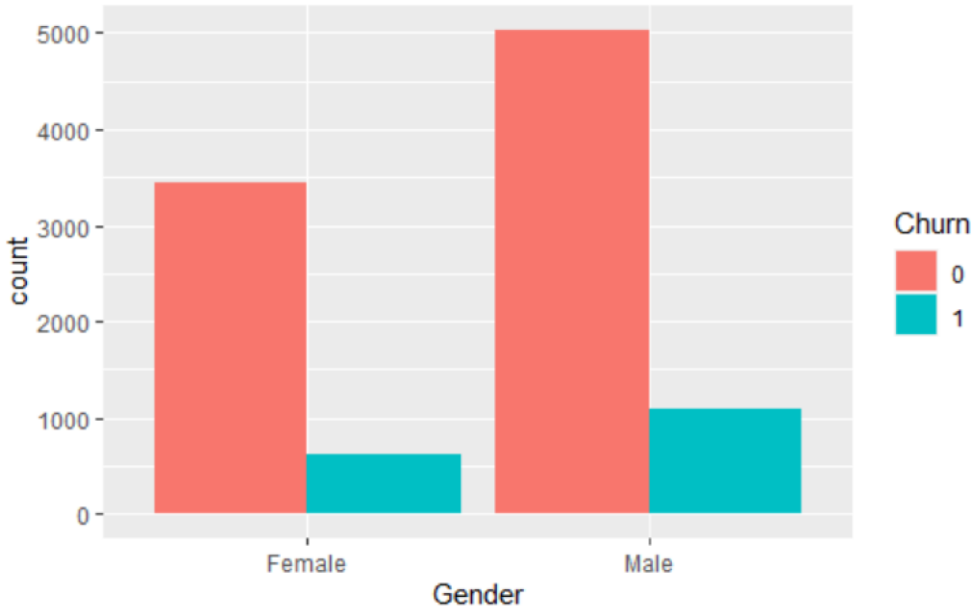


Inference

- This variable is a categorical variable with five discrete values.
- As per the graph we can conclude that out of the total customers approximately 40% of the customers have preferred payment mode as Debit card followed by 31% as Credit Card and 9.7% E wallet.
- As per the Bivariate analysis, we have compared the Preferred Payment mode with our target variable – Churn.
- As per the graph, it can be understood that majority of churned customers have preferred payment mode as Debit Card.

5. GENDER





Inference

- This variable is a binary variable with two values i.e., Male and Female. Originally in the dataset it had “F” and “M” as well which we have combined with “Female” and “Male” respectively.

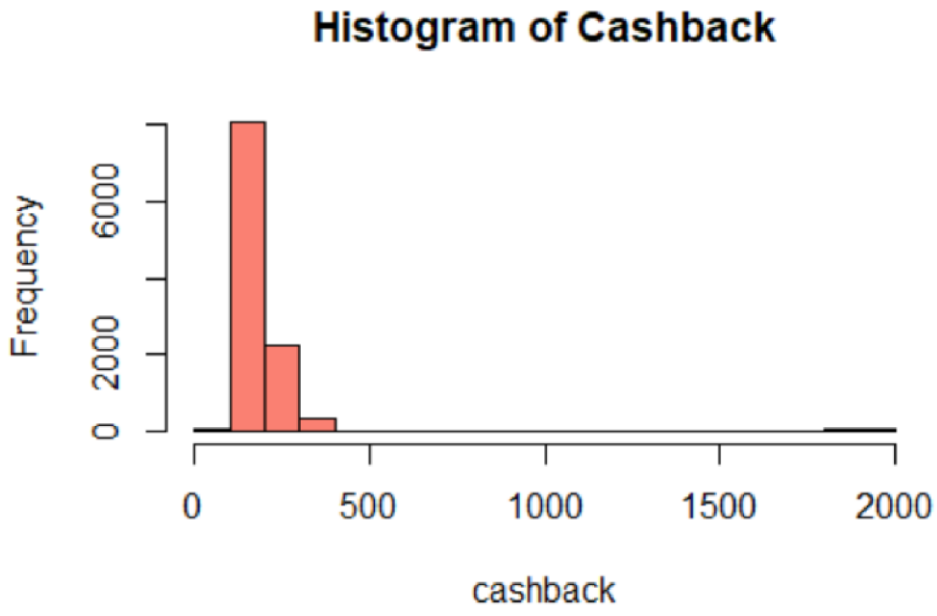
- As per the graph we can conclude that out of the total customers approximately 57.7% are males and remaining are females.

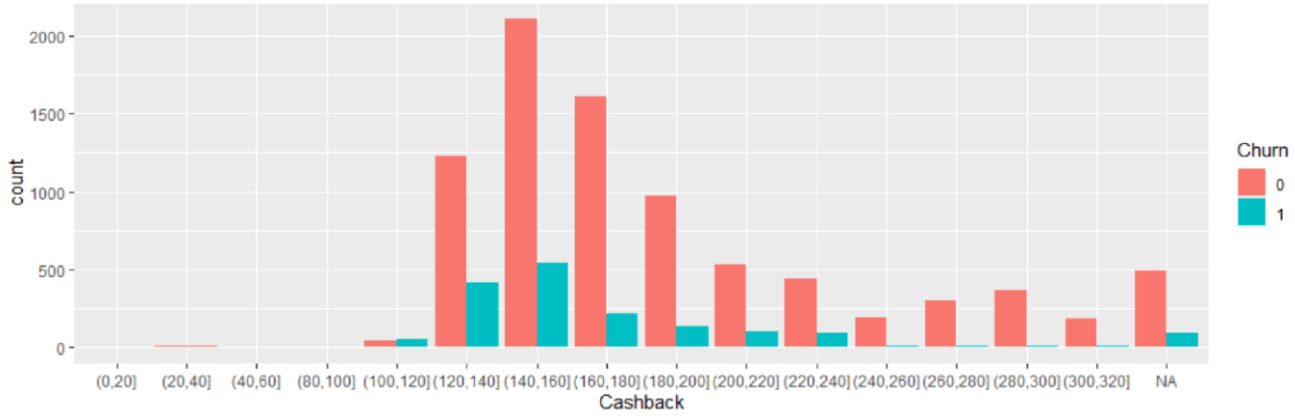
  As per the Bivariate analysis, we have compared the Gender with our target variable –Churn.

  As per the graph, it can be understood that majority of churned customers are males.

  However, the number of Churned customers is far less than the number of retained customers.

CASHBACK





Inference

- This is a type of numerical variable which is discrete in nature. It tells us the cashback amount earned by customers.

  For majority of customers the cashback amount earned range between 100-200 whereas average being 170.

  As per the Bivariate analysis, we have compared the cashback with our target variable – Churn.

- As per the graph, it can be understood that majority of “Churned” customers have earned cashback amount between 140-160 with certain outliers present. The customers that are retained with the company have earned cashback amount between 120-320 with outliers present.

- However, the number of Churned customers is far less than the number of retained customers.

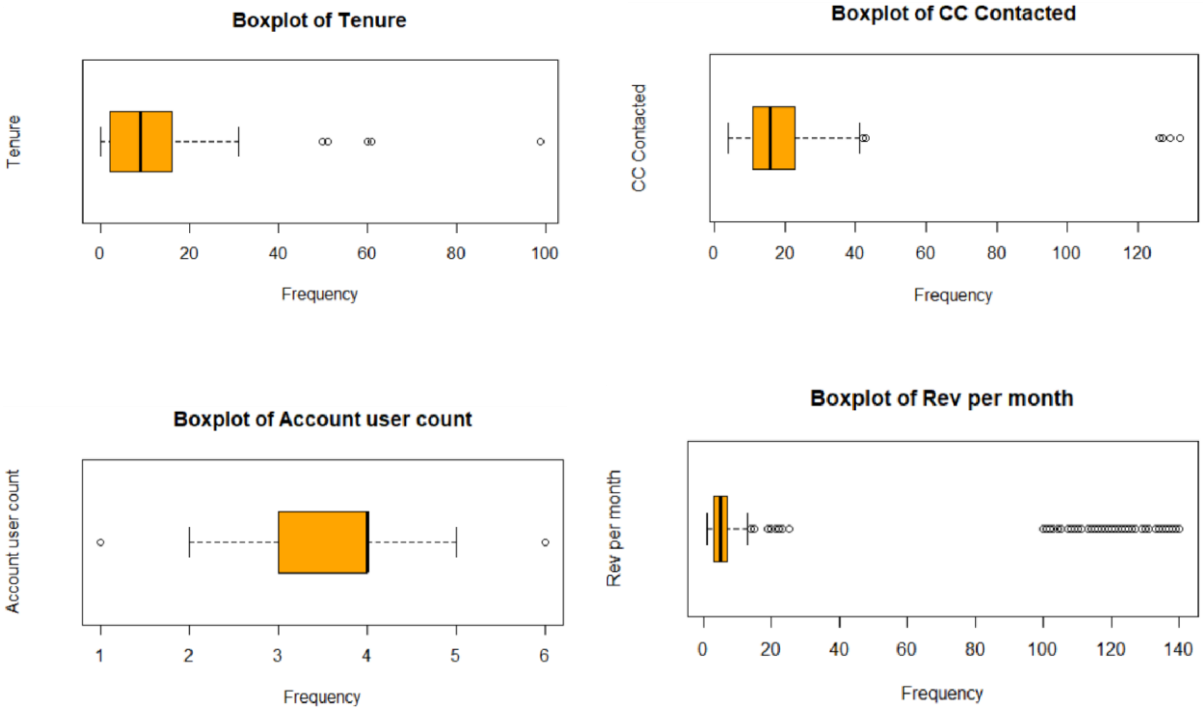
REMOVAL OF UNWANTED VARIABLES

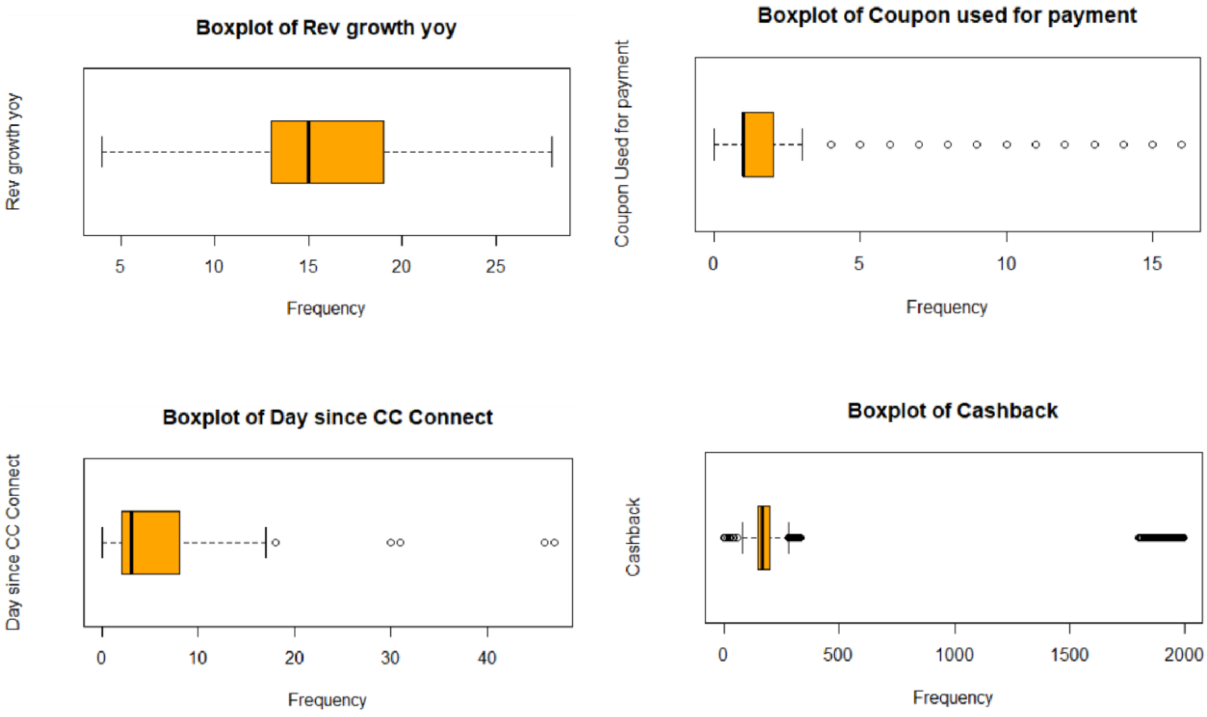
We will remove "**AccountID**" column from the dataset. This variable is unwanted for our analysis as it does not provide any valuable insights to the case study.

CHECKING FOR MISSING VALUE

By running the command to check for missing values it was found that all the variables had missing values. We will further treat them by replacing the missing values with "0" in the variables i.e. "Tenure", CC_Contacted_LY", "Account_user_count", "rev_per_month", "Complain_LY", rev_growth_yoy", "coupon_used_for_payment", "day_since_CC_Connect", "cashback". Remaining null values will be omitted from the dataset.

CHECKING FOR OUTLIERS





By running the above command, we found that all the variables have certain outliers.

VARIABLE TRANSFORMATION

While performing EDA we came across three variables namely Gender, account segment and login device which had some extra categories. By the virtue of similar names, we have combined the categories under similar headings.

Gender: “F” – “Female” and “M” – “Male”

Account segment: “Regular +” – “Regular Plus” and “Super +” – “Super Plus”

Login Device: “&&&&” – “both”

. IS THE DATA UNBALANCED?

There are 83.26% retained customers in the dataset and 16.74% churned customers only. The data is unbalanced and hence we will use SMOTE to balance the data.

MODEL BUILDING

Splitting the Data into Train and Test Dataset (70:30)

The first step of any Machine Learning Model Building is splitting the dataset into training and testing datasets. This is majorly done to evaluate the performance of an algorithm. This is done by dividing the dataset into two subsets where the training dataset is used to fit the model whereas the testing dataset is used to test the model accuracy and fit. We will split the dataset in a ratio of 70:30.

4). INSIGHTS FROM EDA

E-commerce companies these days face major issues of customer churn. It is more cost effective to retain customers than to acquire new ones which is why it's important to track customers at high risk of turnover (churn) and target them with retention strategies. As per the Exploratory Data analysis done as part of this project we came on to the following conclusions:

- New customers with tenures less than 5 years are generally churned more often.
- Churned customers have their preferred login device as Mobile phones.
- Majority of churned customers belong to Tier 1 cities and have their preferred mode of payment as Debit card.
- Churned customers have contacted the customer care almost 10-20 times in the last year.
- Majority of churned customers are males.
- The customers that are churned majorly have given a service score of 3 and customer care agent score as 3.
- Majority of the churned accounts had 2-3 users per account.
- A major number of churned customers had "regular Plus" plan.
- Majority of the customers churned are single.
- Revenue generated of majority of churned accounts is between 0-5 which is quite low and revenue growth from last year is between 12-14%.
- Churned customers have used very few coupons.
- The customers that have made several complaints are mostly churned.
- Cashback received to customers churned was between 120-160.

RECOMMENDATIONS

A few recommendations as per the EDA for the company are:

- New customers should be given more benefit in terms of discounts and promotions.
- Customers who are giving satisfaction score more than 3 should be targeted to provide more satisfaction in terms of services and products.
- Complaints should be taken seriously by the company and customer service should be the top priority.
- Competitive marketing strategies should be taken into consideration. Market research should be done to provide good pricing and discounts.
- Family plans should be made cost effective so that the entire family can avail the benefits from a single account.
- An option to customise their plans and a freedom to pause and skip their subscription plans. Flexible plans are liked by all subscribers.
- Loyalty programs to encourage and appreciate long term subscribers.
- Survey the current subscribers to understand the problem areas and asking for recommendations on improvement of services.
- The company needs to look at its performance in various cities and make sure that they maintain an equal quality of service. The company strategies must reflect that all customers are equally important to them.
- The company needs to come up with more cashback offers and in a wider variety. These can be direct monetary cashbacks or giving out coupons for discounts on future shopping.