Domain-oriented Case Study

Retail: Customer Segmentation and Churn Prediction

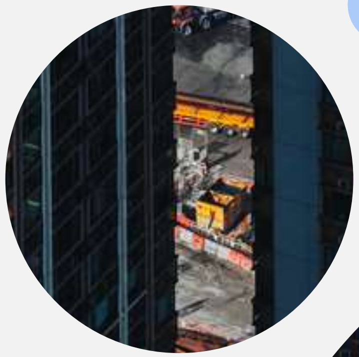
Pragya Bhargava

Karan Shet

Manikanta Sai

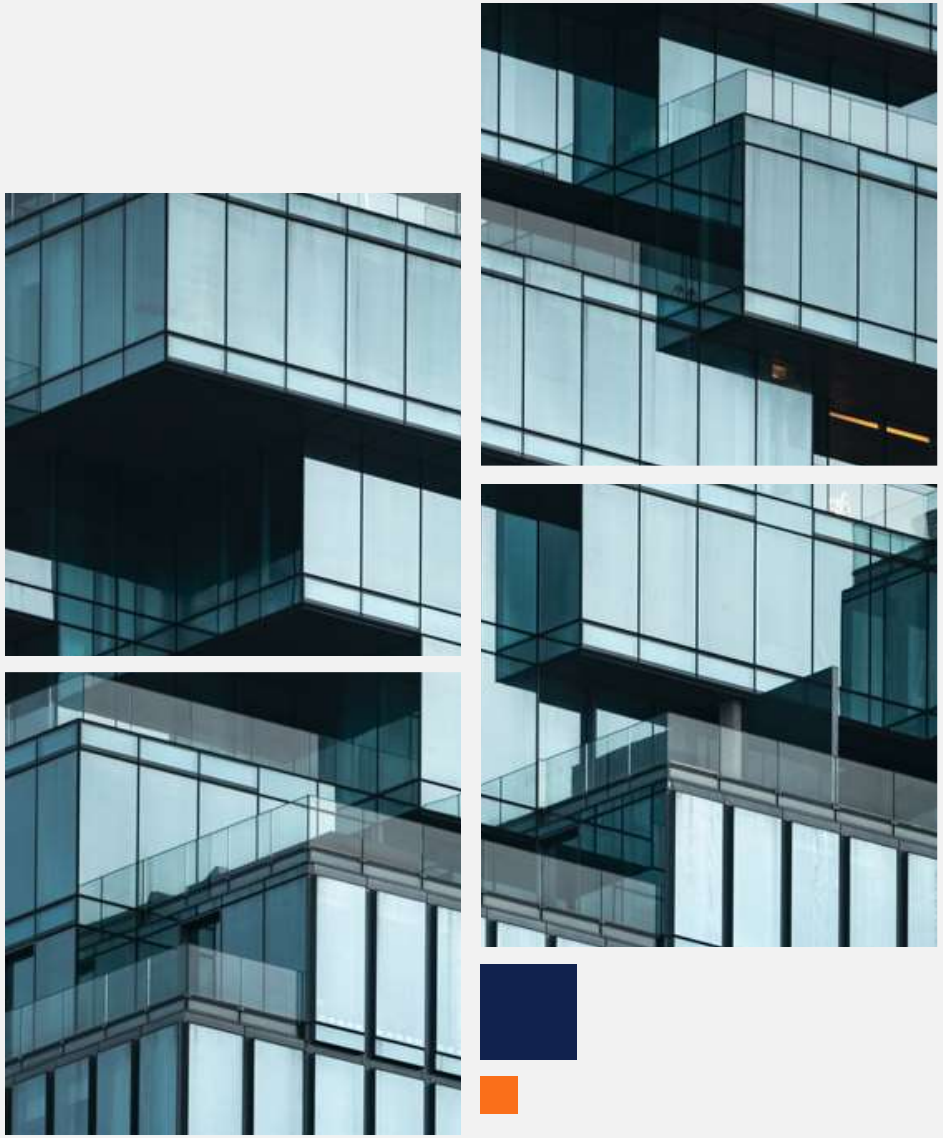
Agenda

- RetailKart.com Align on Problem Statement
- Objective
- Tools Techniques (Customer Segmentation)
  - Data Understanding and Preparation
  - Exploratory Data Analysis
  - Feature Engineering
  - Customer Segmentation
  - Clustering
- Tools Techniques (Churn Prediction)
  - Data Understanding and Preparation
  - Exploratory Data Analysis
  - Model Building
  - Model Evaluation
- Recommendation on Personalized Strategies on Customer Cohorts
- Recommendations on Personalized insights to Churn Prediction



RetailKart.com Problem Statement

- A small and medium-scale organization that majorly deals in wine, fruits and meat products, having held around 35% market share
- Organization wants to stay competitive. Challenges are to compete with quickly expanding companies
- RetailKart.com goes Online and sets the roadmap to Customer Acquisition, Increase Transactions in the platform
- One such Project is to kick started with the objective of offering Personalized Experience for its Customer Shopping Experience



Customer Segmentation

About Data

Customer Purchase Data

Shape of the Data

- 2239 Rows
- 29 Features
- 26 Numerical Features
- 3 Categorical Features

Brief Understanding of the Features

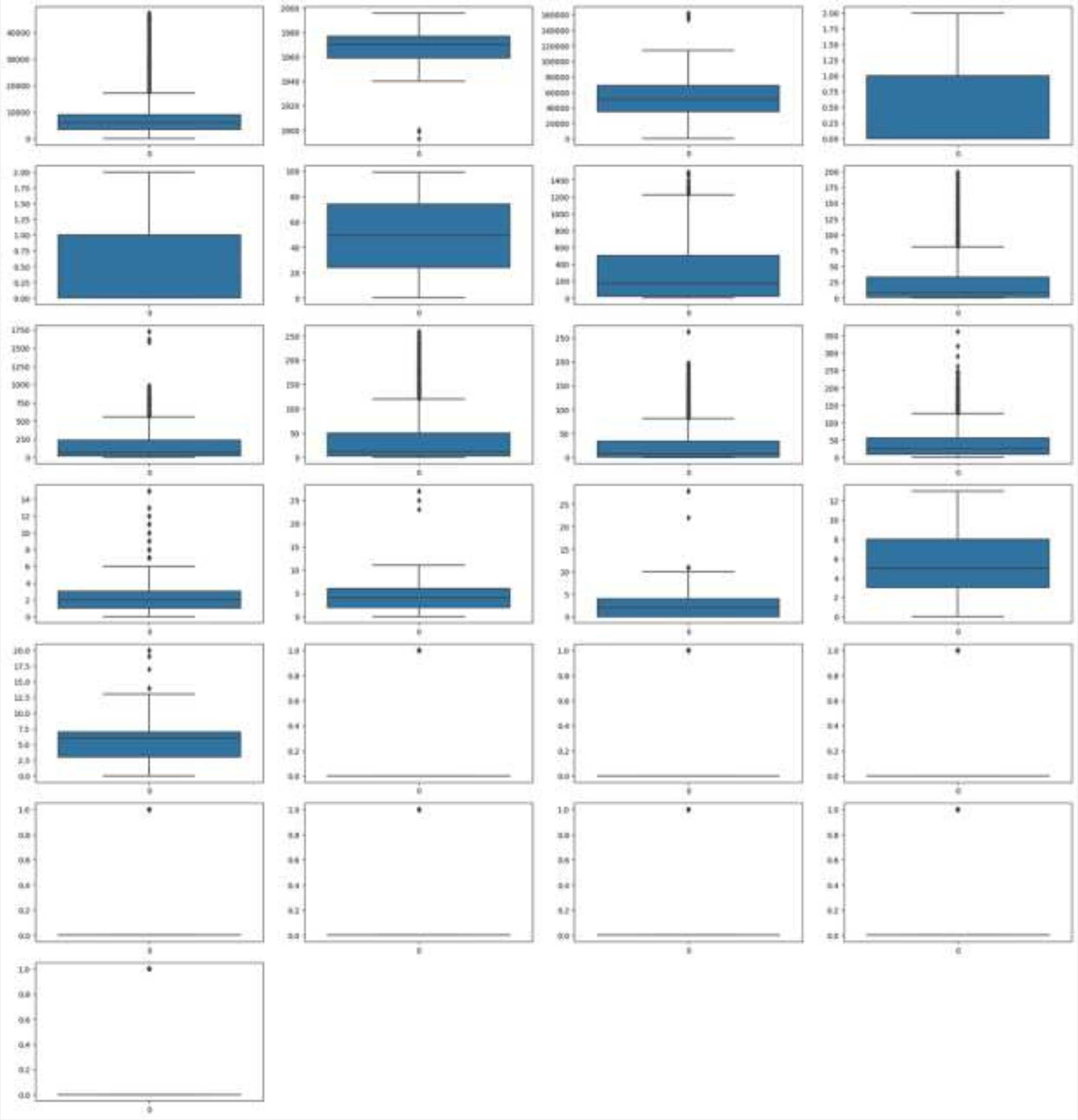
- Consists of Demographic Data (Marital Status, Education)
- Product Offering (Meat, Wine, Sweet, Fish)
- Channels (Store, Web/Online)
- Promotion response details

Missing values

- Income (1 Record)
- Response (23 Records)

Exploratory Data Analysis

- Outlier Analysis
- Outlier observed for Income and Year variable
- Treatment for such variables to be performed after Feature Engineering



Exploratory Data Analysis

Correlation > 0.5

- Income & Customer Spending, Liquor, Meat, Snacks & Fruits
- Liquor, Meat, Snacks & Fruits & NumCatalogPurchase, NumStorePurchase, NumWebPurchase

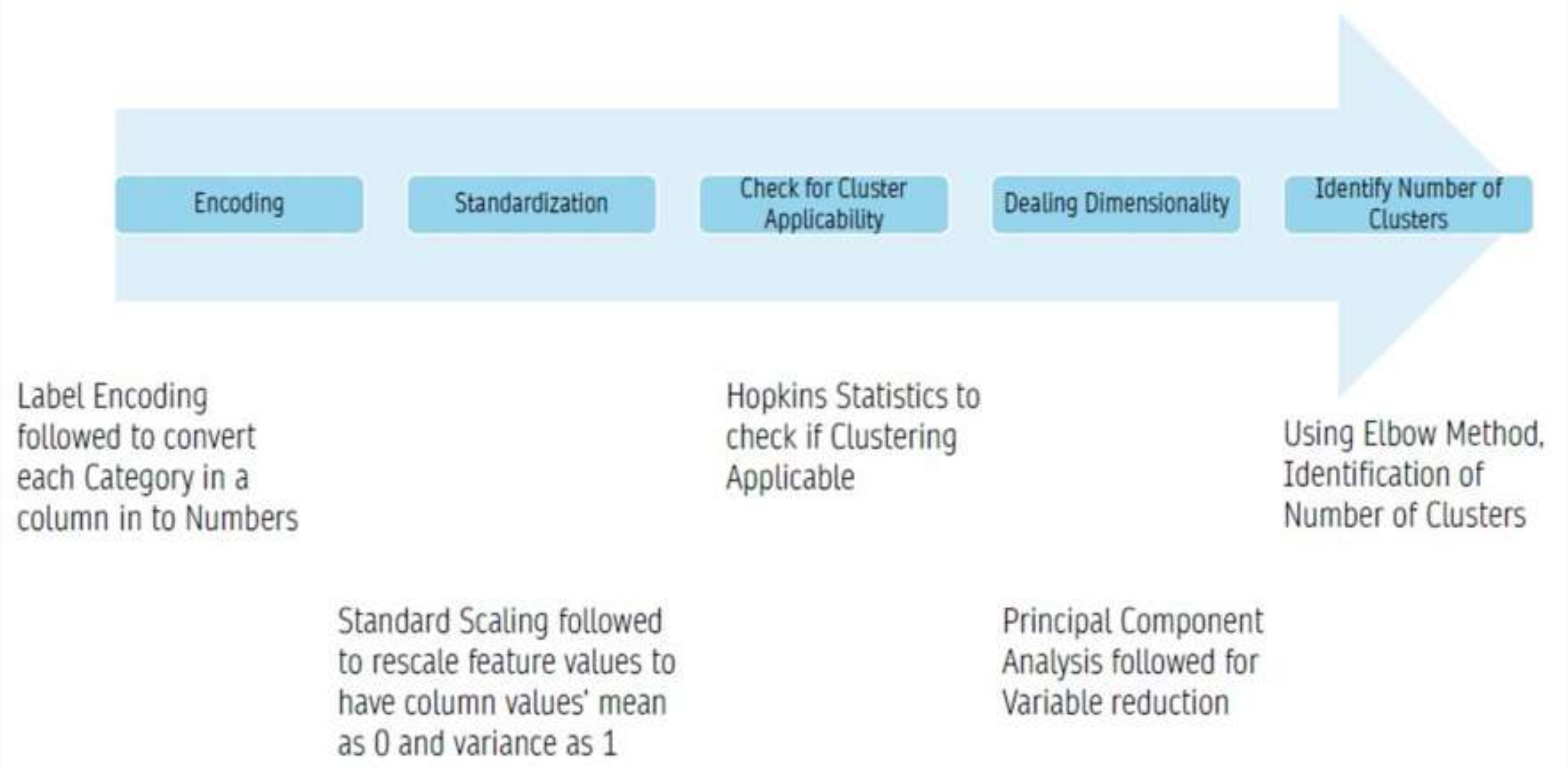
Feature Selection

Below Features need to be engineered for better clustering output

New Features Engineered	Method / Formula
Age	Based on the Year of Birth
LoyalSince	Based on the Dt_Customer
Parent	Combining count of KidsHome and TeensHome Features
Education	Simplifying Education Feature by having just two categories as UG, PG
Marital Status	Simplifying Marital Status to have Single, Not-Single
Family Size	By representing the size by combination of Marital Status and Kids
Customer Spending	By Summing up the transaction data from MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts
Categorization of Spending	Liquour <- MntWines Meat <- MntFishProducts, MntMeatProducts Snacks & Fruits <- MntSnacksProds, MntSweetProducts, MntFruits

Post feature engineering, Features ("ID", "Dt_Customer", "Year_Birth") are removed

Moving towards Clustering

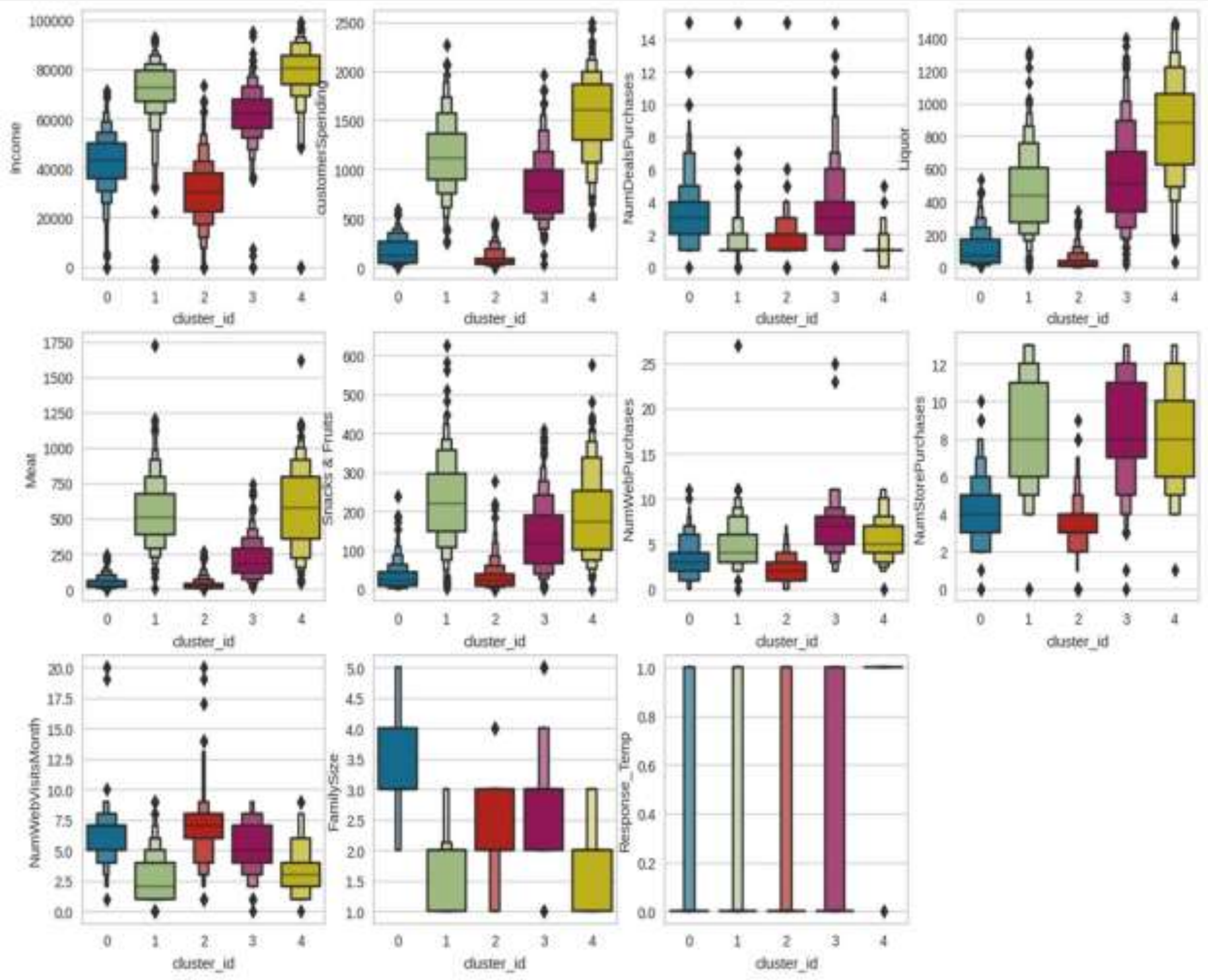


Cluster Formation and Profiling

High Income, High Spending	Very Low Income, Very Low Spending	Decent Income and Spending	Low Income & Low Spending	Customers Giving Consistent Revenue
• High Value Customers ordering Meat, Liquor • Low Exposure to Online Purchase	• Showing Interest to purchase through online platform	• Showing interest to Deals and Online Purchase	• Showing interest to purchase through online platform	• Need to be encouraged for Online Purchase

Profiling

- High income, High spending, Not choosy on deals, choosy on Liquor/Meat/Snacks consumption and potential customers to move to online; their web views are low as well. Encourage Online. Max Family Size is 3
- Very low income, very low spending, showing interest in Web Views. Family Size of 2-3
- Decent income and spending and showing interest towards deals and online purchases. Family size of 2-3
- Low Income, Low Spending; however showing interest towards online. Family size of 3-4
- Customer Segment who can give consistent revenue need to be encouraged for online purchase. Family size of 1-2



Recommendations

Higher Discounts in Online Purchases

High Value Customers and Consistent Revenue givers to be encouraged by providing some discounts.

Personalized Experience

For High Value customers, Products like Liquor & Meat can be shown once logged in.

For Low Value Customers, Sweets & Snacks, Fruits to be displayed.

Promotional Campaigns

Continue Promotional Campaigns to raise awareness and pull transactions to Online Platform.

Bringing More Deals

Deals are more opted by Medium Income Customer Segment.

Churn Prediction

About Data

Customer Purchase Data

Shape of the Data

- 5630 Rows
- 29 Features
- 20 Numerical Features
- 5 Categorical Features

Brief Understanding of the Features

- Consists of Demographic Data (Marital Status, Churn, Ordering, CityTier, Income, HourSpendonApp, Order_count, Coupons)

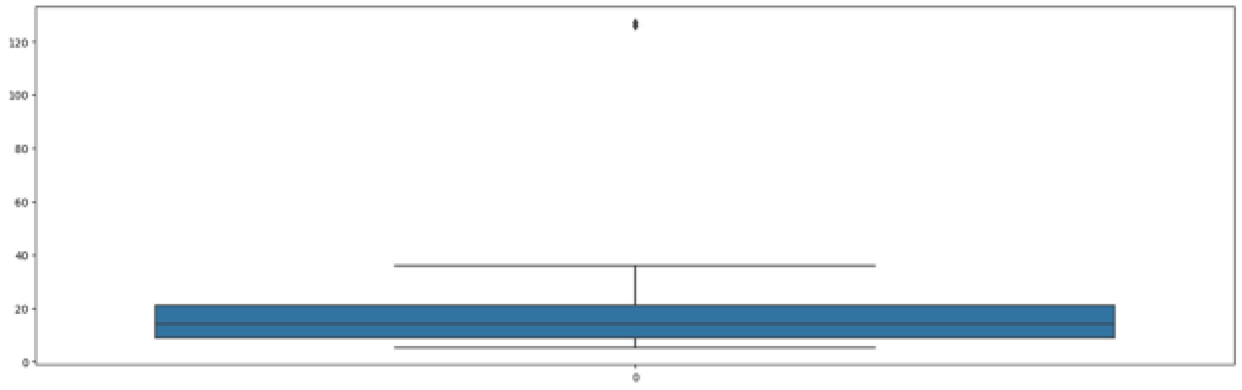
Missing values

- Tenure (264 Record)
- Warehousetohome (251 Records)
- HourSpendonApp (255 Records)
- CouponUsed (256 Records)
- OrderCount(258 Records)
- DaySinceLastOrder (307 Records)
- OrderAmountHikeFromlastYear (265 Records)

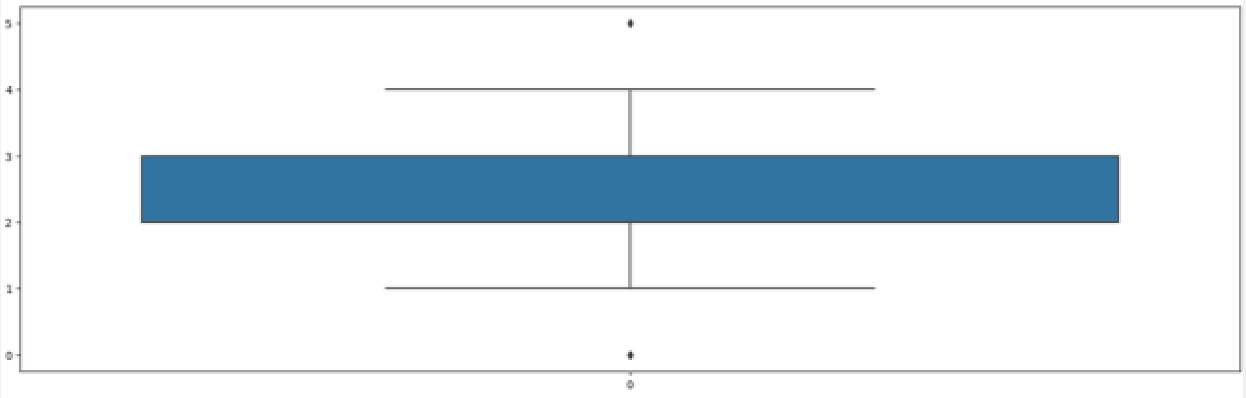
Exploratory Data Analysis

Outliers

- WarehousetoHome: 2 Outliers



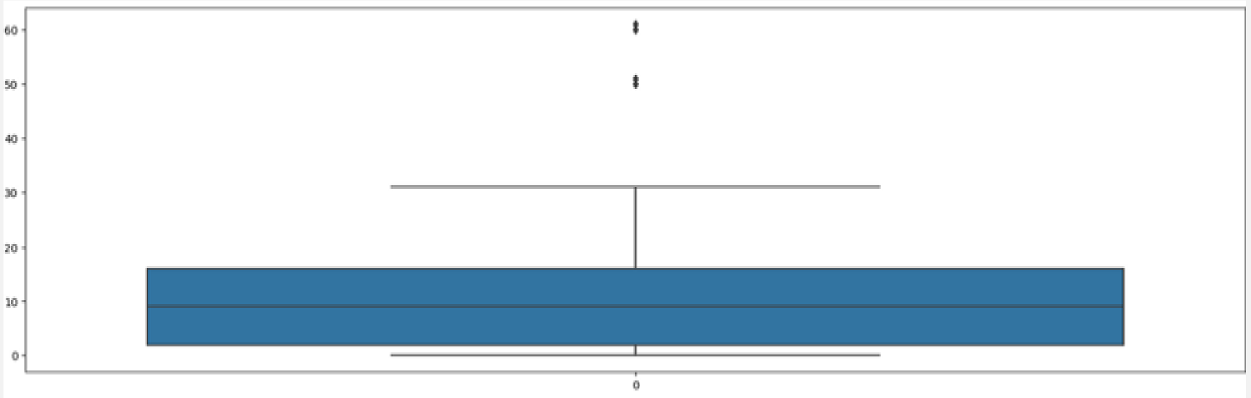
- HourSpendOnApp: 1 Outlier



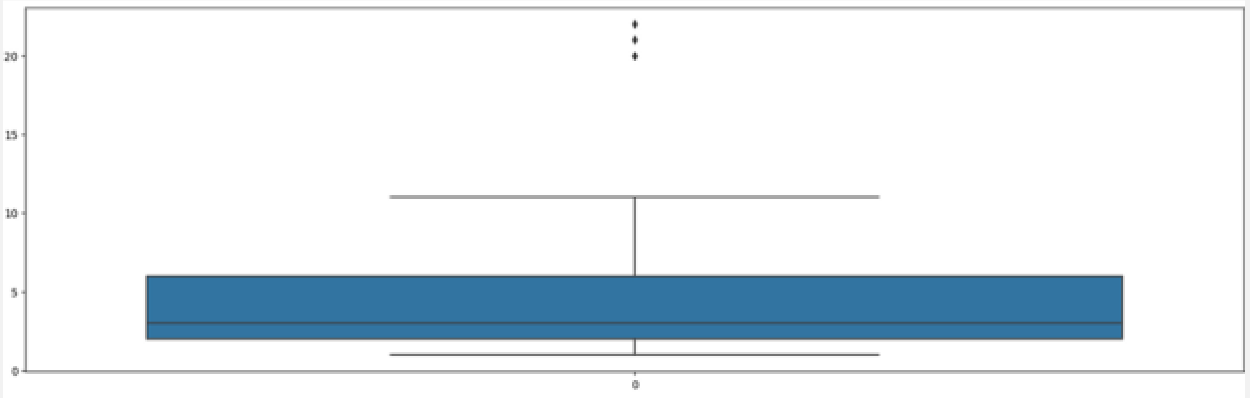
Exploratory Data Analysis

Outliers

- Tenure: 4 Outliers



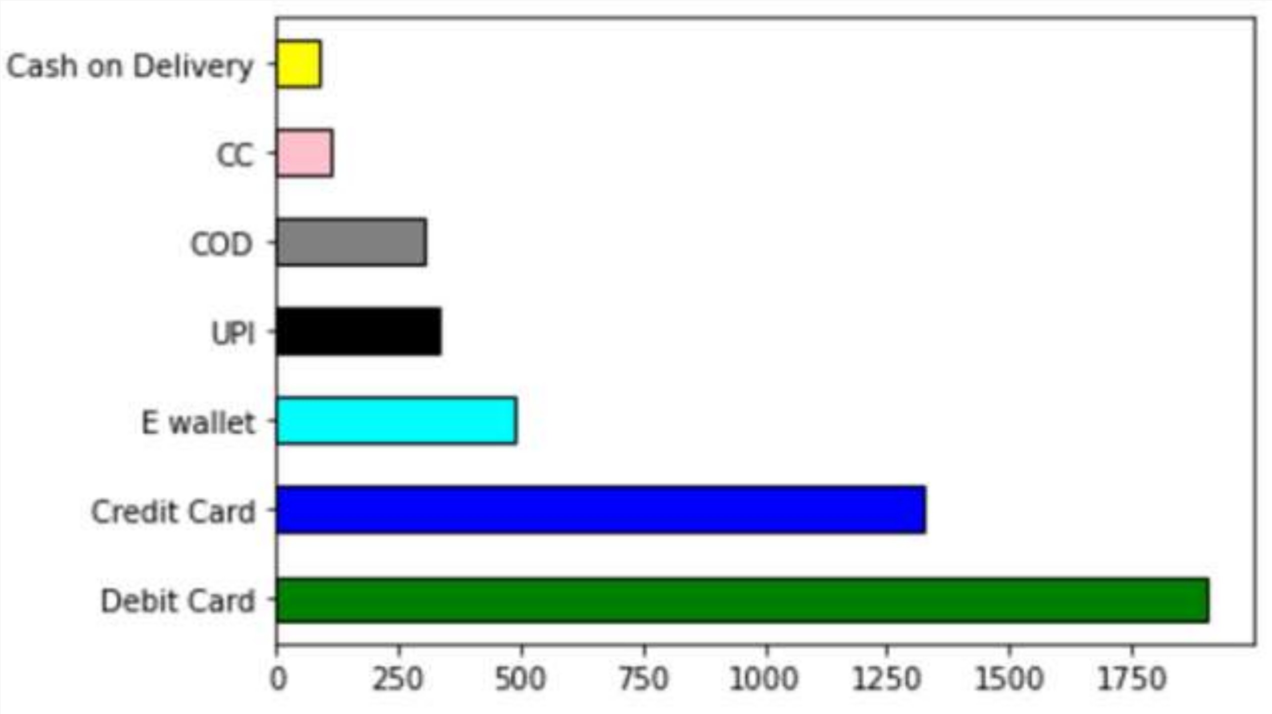
- NumberOfAddress: 3 Outlier



Exploratory Data Analysis

Correlation > 0.5

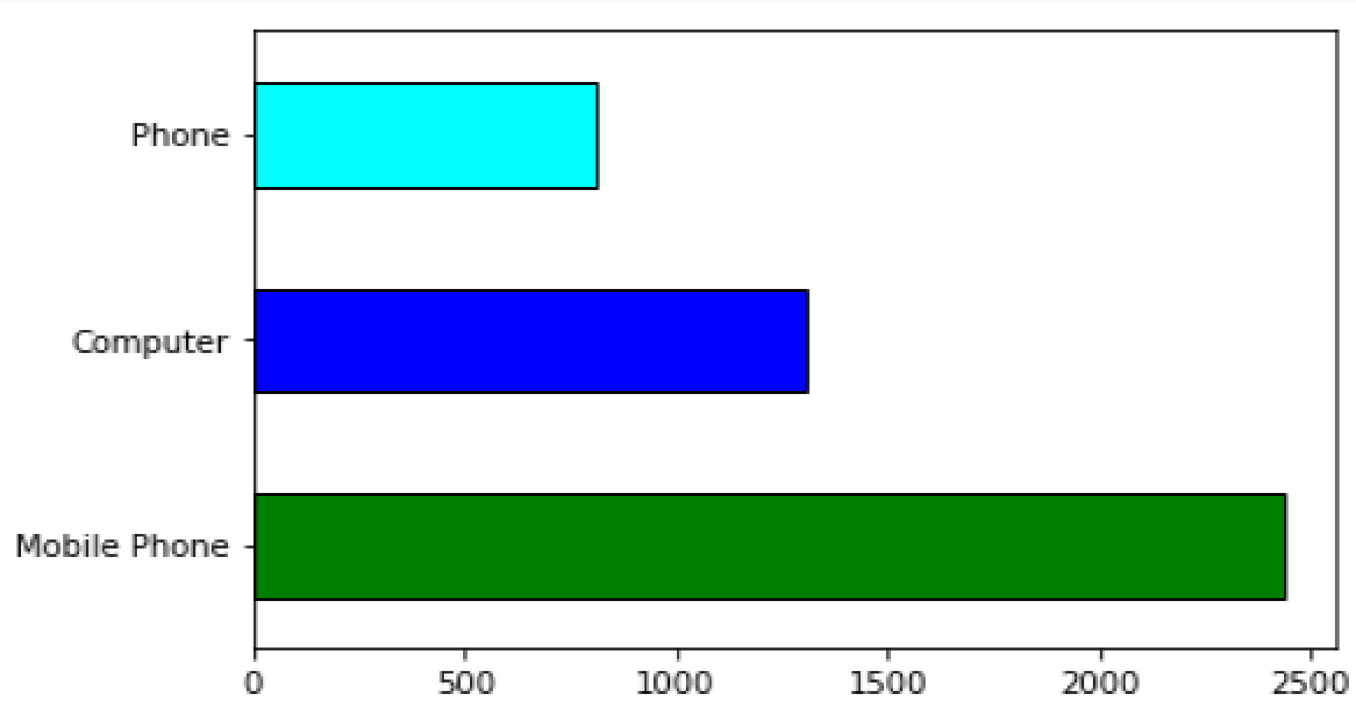
- Plot 1: Maximum users are using mobile phone as a platform for orders.



Exploratory Data Analysis

Correlation > 0.5

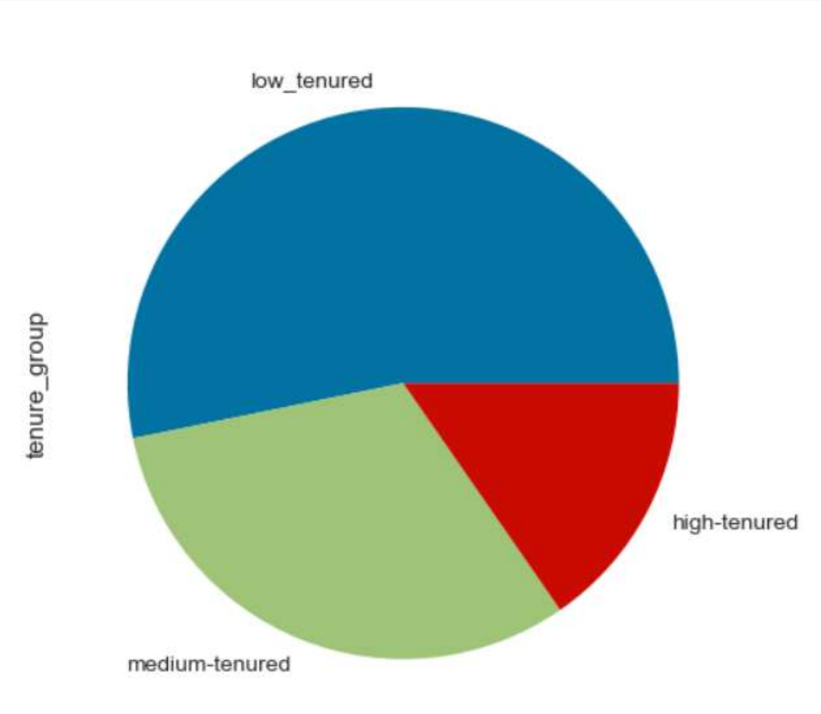
- Plot 2: The debit card transaction is the highest payment mode in the entire population.



Exploratory Data Analysis

Correlation > 0.5

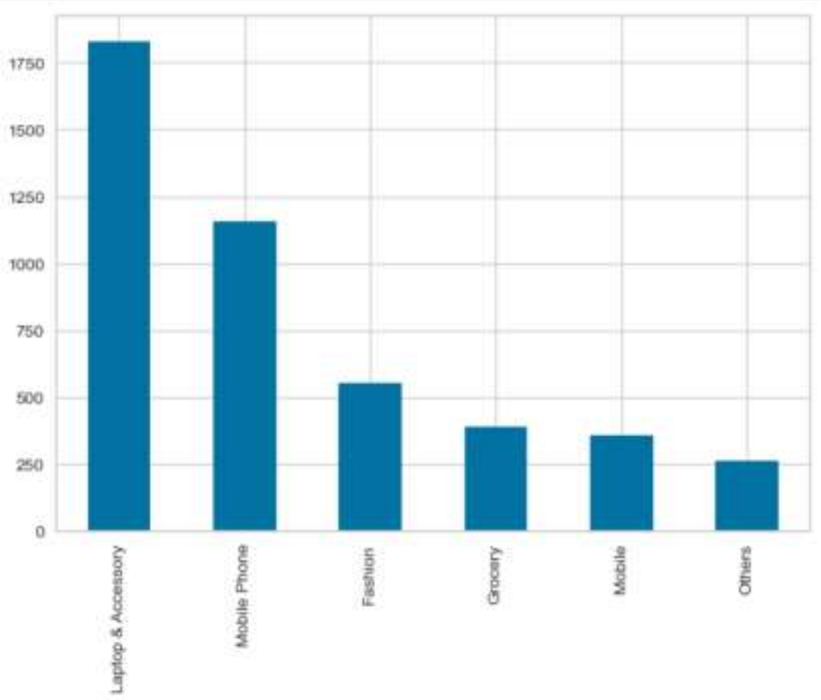
- Plot 3: The maximum population is falling under low-tenured, people who are just 8-10 months old in the platform



Exploratory Data Analysis

Correlation > 0.5

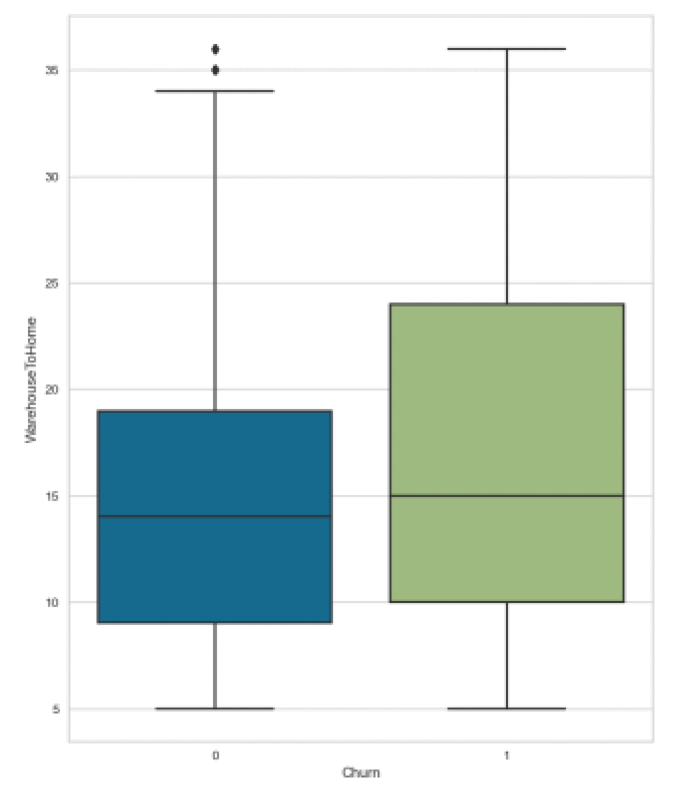
- Plot 4: Laptop accessories are the most selling items among all in the list



Exploratory Data Analysis

Correlation > 0.5

- Plot 5: In bivariate analysis, with churn variable , we can say that higher the distance from warehouse higher will be the rate of churn



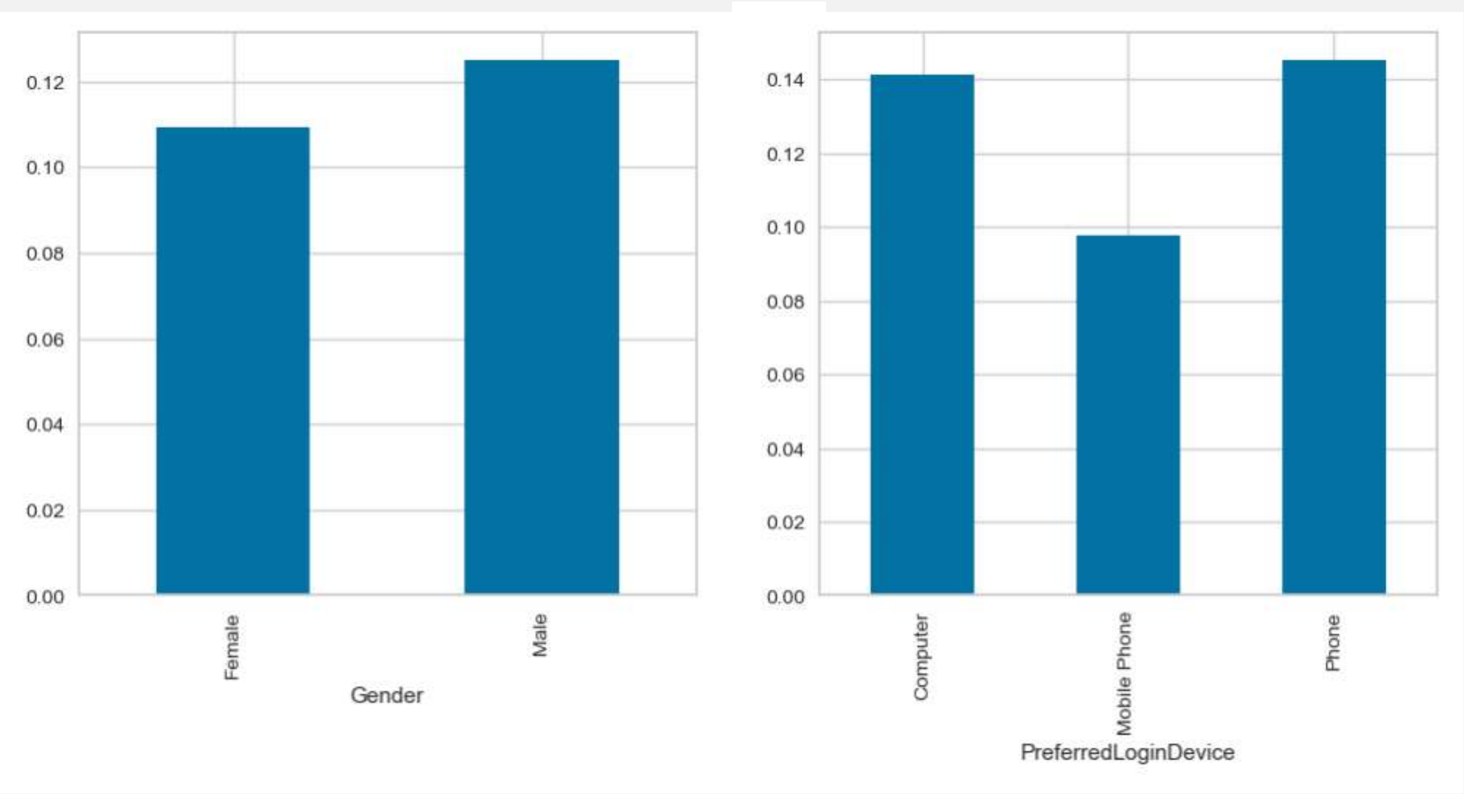
Exploratory Data Analysis

Correlation > 0.5

- Plot 6:

Males have higher chances of churn than females

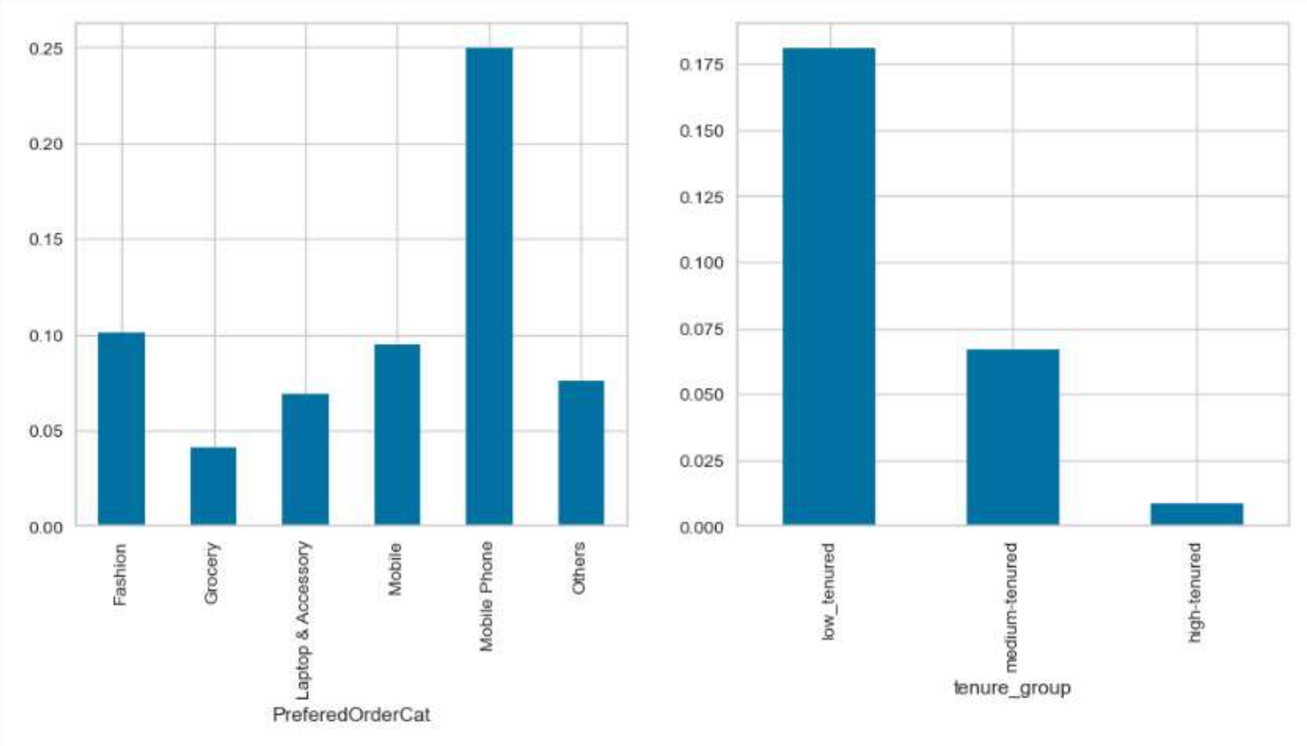
People who are calling and placing orders have a higher rate of churn



Exploratory Data Analysis

Correlation > 0.5

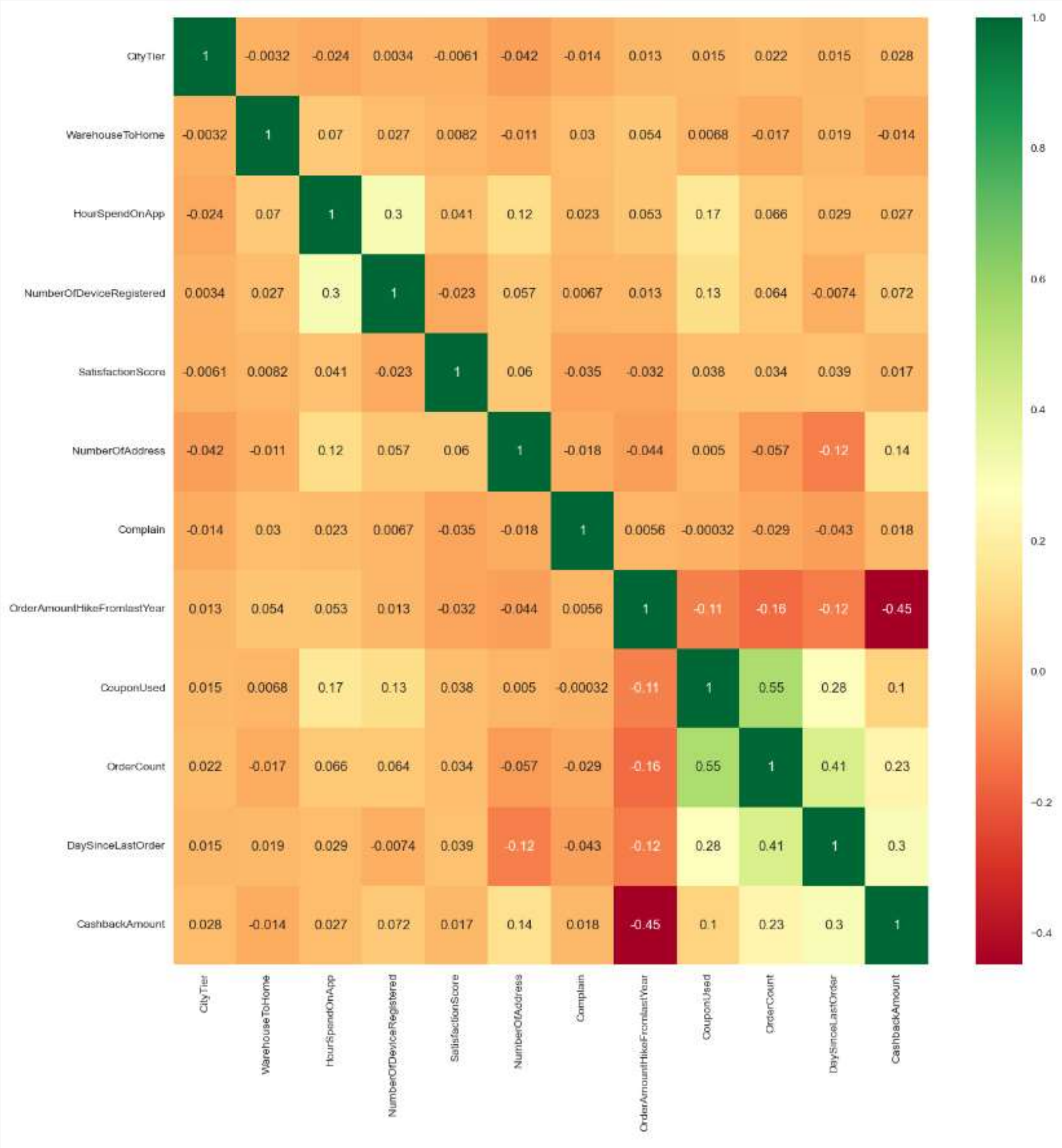
- Plot 7: Mobile Phone Category and Low-tenured customers are churning more



Exploratory Data Analysis

Correlation > 0.5

- There are certain attributes which show a high multicollinearity



Model Outcomes

Performed logistic regression taking best 25 variables in account

City tiers, payment mode like COD & credit/debit, warehouse distance , login device are some of the attributes which are very important.

Performed Scaling, RFE and class imbalanced techniques (SMOKE).

The model has achieved 87% of the sensitivity that means it's able to predict churn in training set.

Thank you