Customer Churn Prediction - Report

1. Data Gathering

Selected Datasets and Rationale

The dataset consists of multiple sheets from an Excel file. The following datasets were selected based on their relevance to Customer Churn prediction. All datasets have their respective data with the Customer ID which helps uniquely identify a customer.

1. **Customer Demographics:**

   Contains features like Age, Gender, Marital Status and Income Level. These features are critical since demographic features often influence customer behaviour.
   There are 1000 samples in the dataset, one for each customer ID.

2. **Transaction History:**

   Included features are: Transaction ID, Transaction Date, Amount Spent and Product Category. The Transaction ID was a redundant column since it is used to identify each transaction which has no effect on customer churn.
   There are 5054 samples in the dataset.
   Transaction patterns reveal spending behaviour which may affect churn.

3. **Customer Service:**

   Include features like Interaction ID, Interaction Date, Interaction Type and Resolution Status. Customer service Interactions like Unresolved complaints could be a crucial indicator for customer churn.
   Samples: 1001

4. **Online Activity:**

   This dataset contains the various online service usage of customers. Includes features like Last Login Date, Login Frequency and Service Usage. The online service usage patterns could signal a customer likely to churn. A high user may not be as likely to leave as their counterpart.
   Samples: 1000

5. **Churn Status**

   The target variable ‘ChurnStatus’ indicates whether a customer has churned (1) or not (0). 1000 unique samples in the dataset.

2. Exploratory Data Analysis - EDA

Key Insights and Visualisations

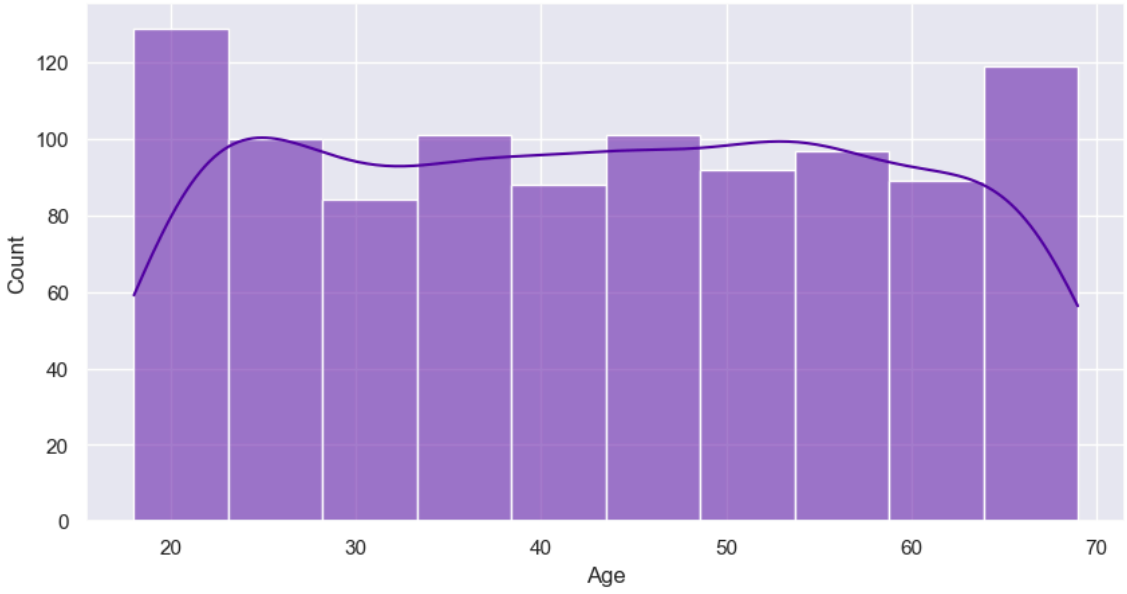
1. Customer Demographics:

Statistical summaries, Mean, Standard Deviation, Quartiles, Min and Max was calculated for the dataset.
For categorical variables the Top group and their frequency is calculated.

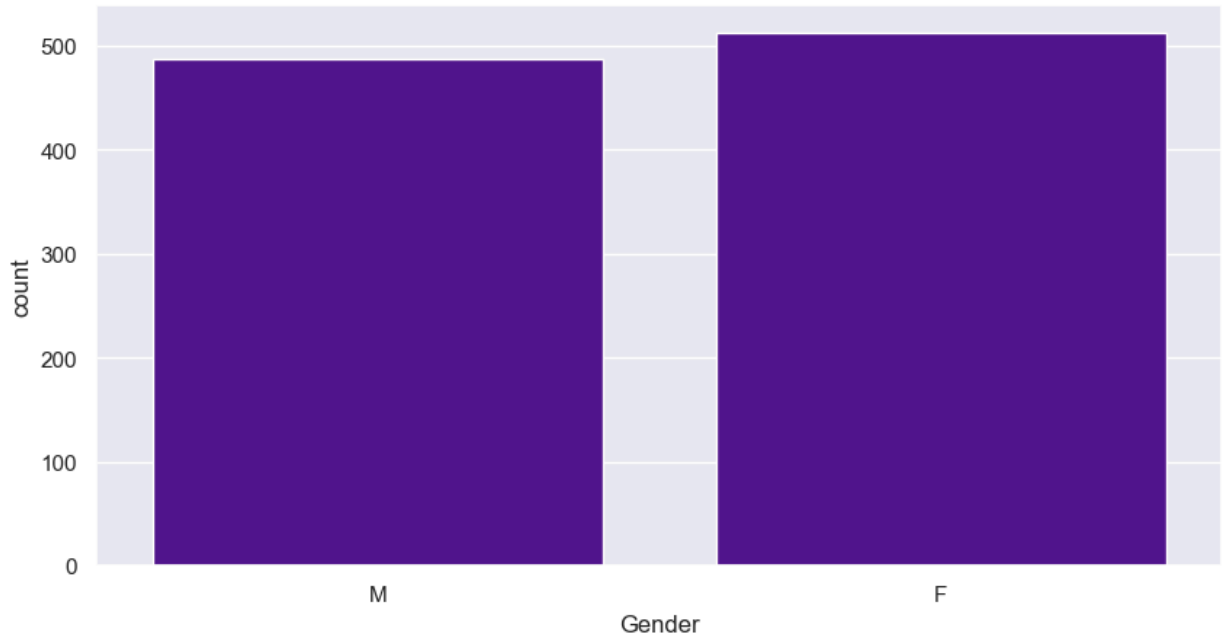
	mean	std	min	25%	50% / median	75%	max
Age	43.26	15.24	18	30	43	56	69

	Unique Classes	Top class	Frequency
Gender	2	F	513
Marital Status	4	Widowed	276
Income Level	3	High	349

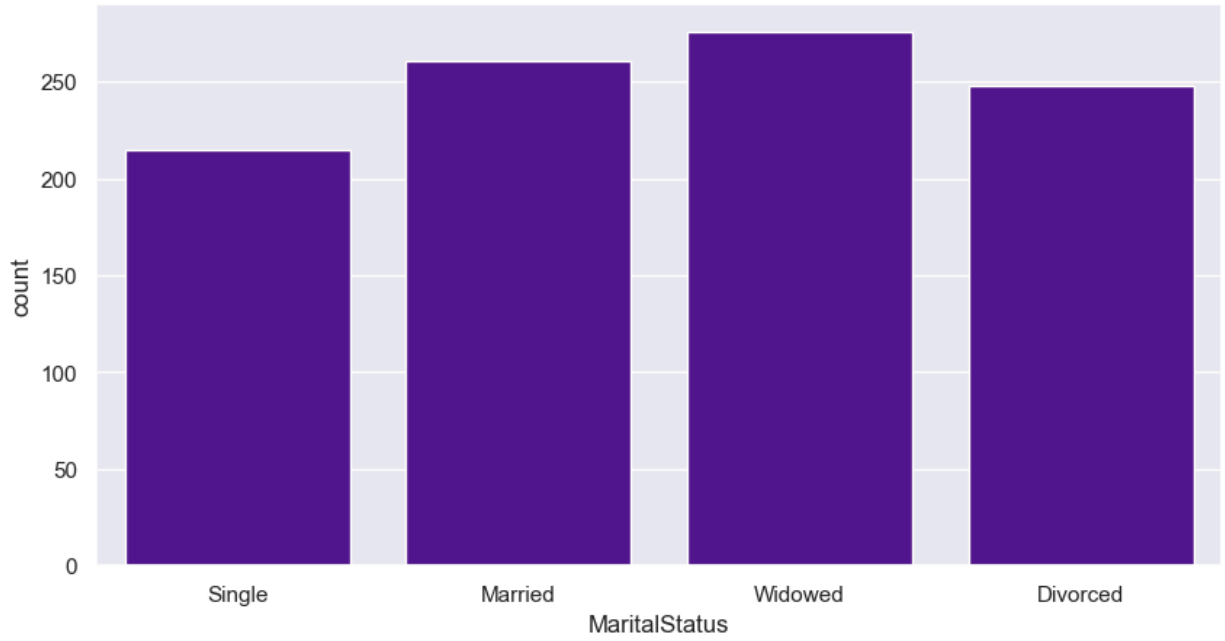
- Plotting Age column counts in bins of 5 years.



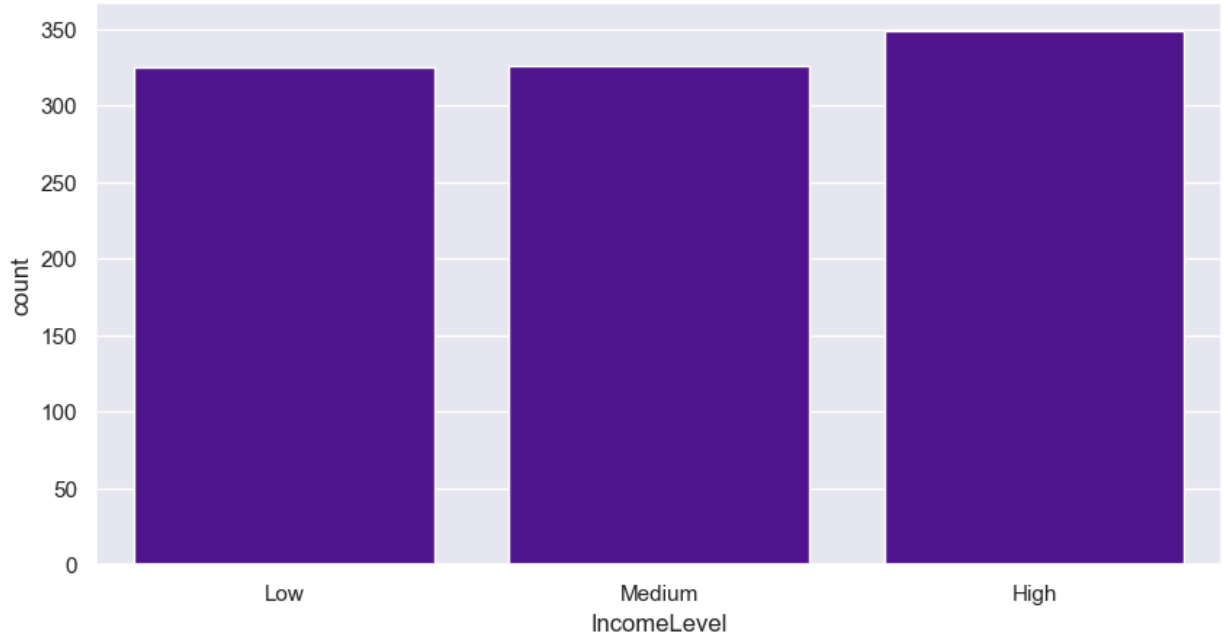
- Number of customers in by their Gender



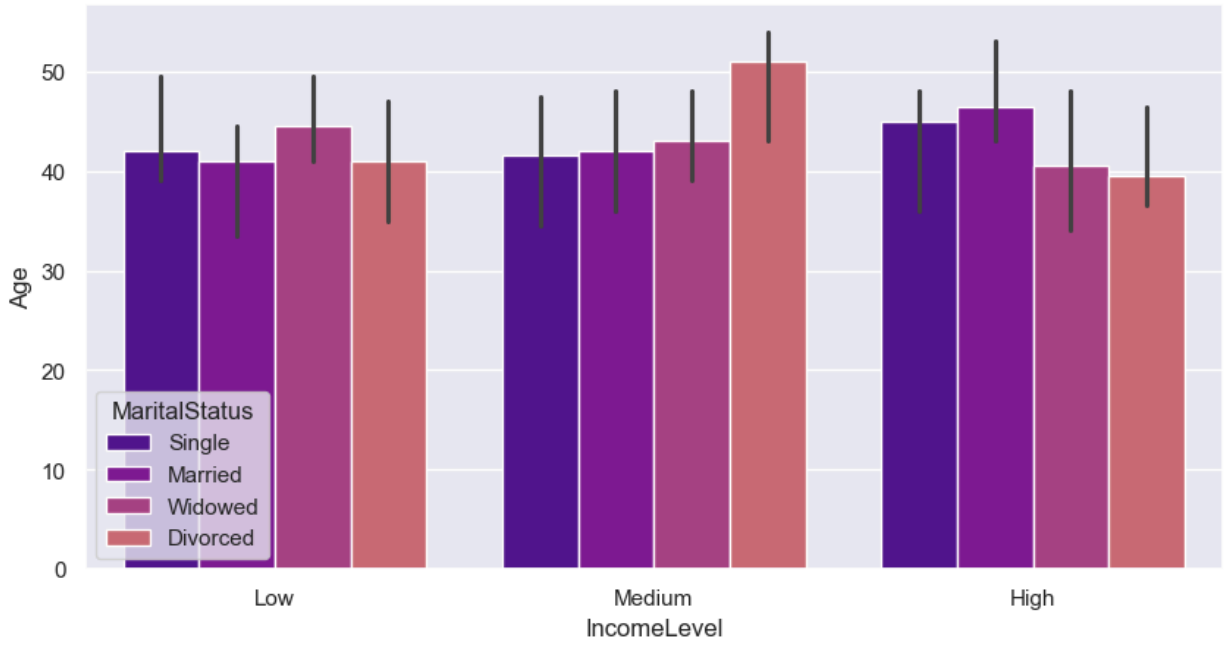
- By marital status



- By income level



- Median age of customers by their Income level and Marital status



It was evident from the EDA that the demographics data is distributed uniformly.

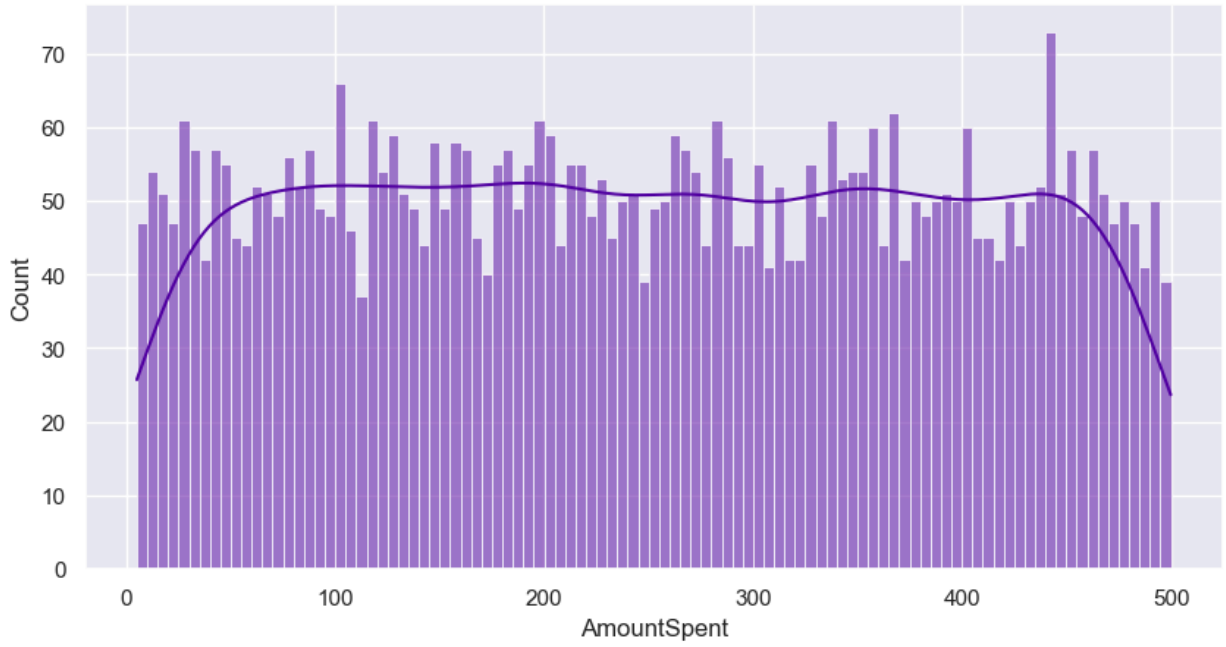
2. Transaction History:

	mean	std	min	25%	50% / median	75%	max
Amount Spent	250.70	142.25	5.18	1277.10	250.52	373.41	499.86

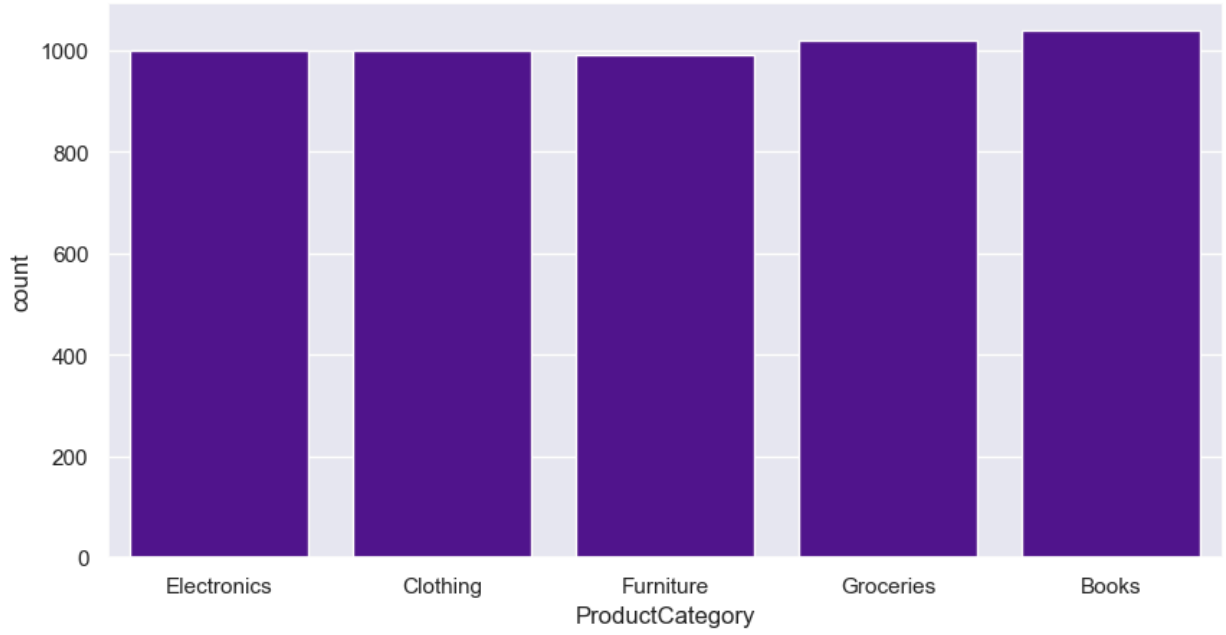
	Unique Classes	Top class	Frequency
Product Category	5	Books	1041

- Plotting the number of transactions for every 5 days

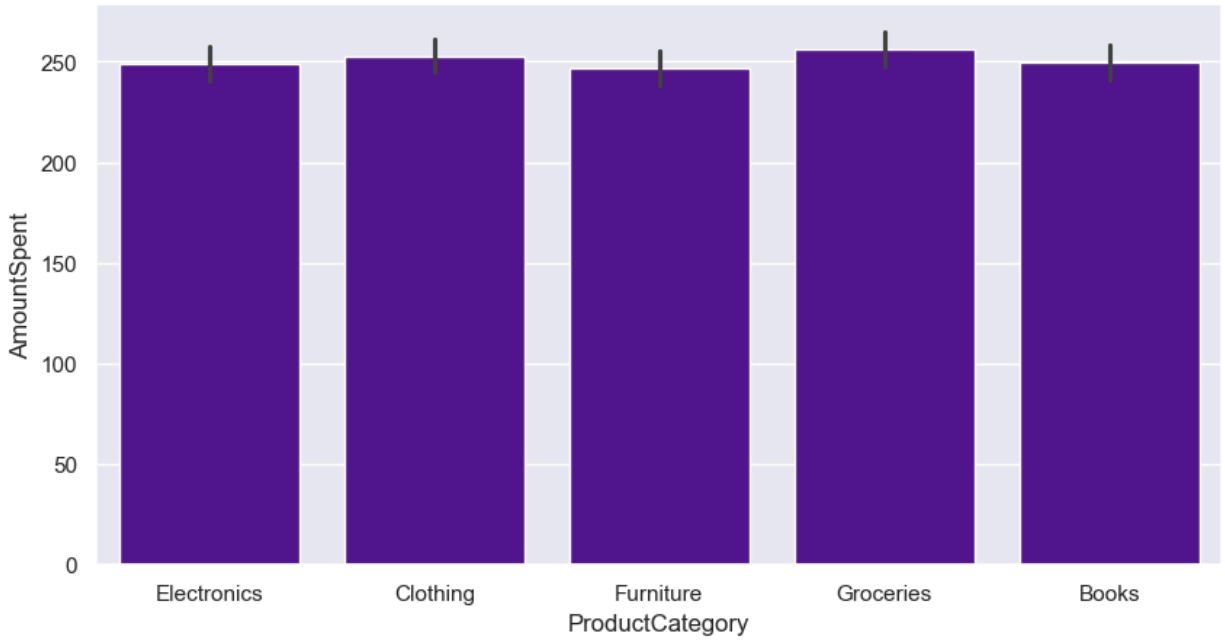
- The Amount spent on a transaction by customers, in bins of 5.



- Count by product category



- Mean amount spent by each product category

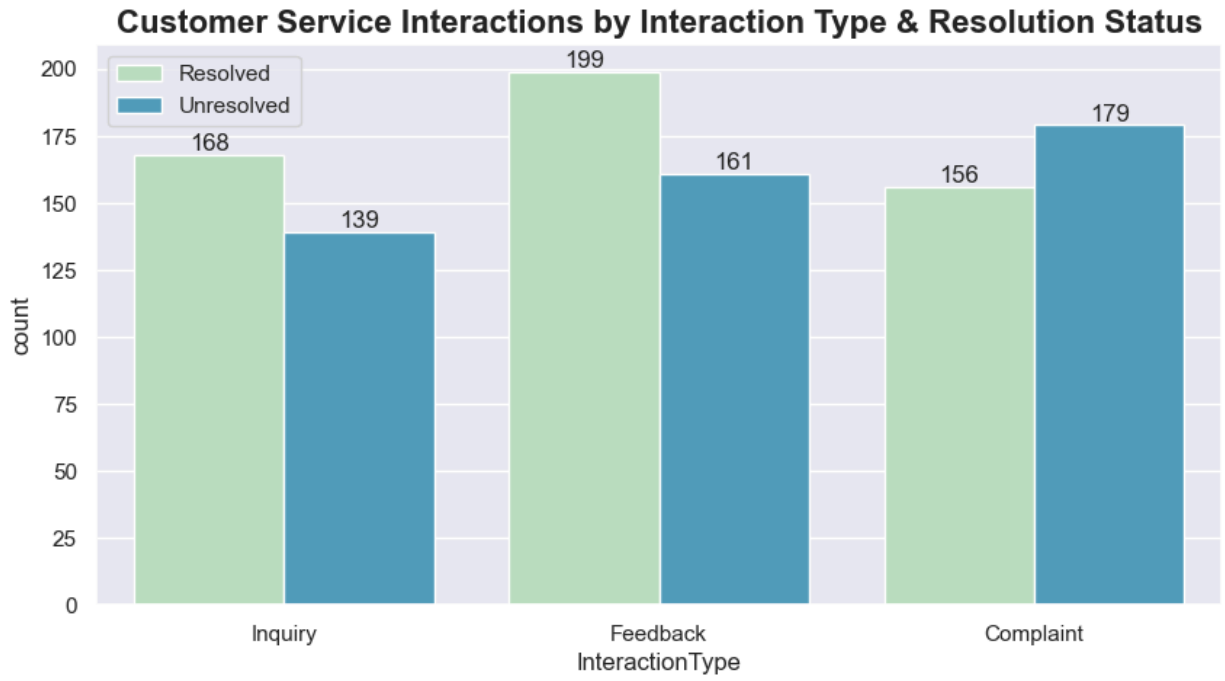


The EDA indicates uniformly distributed data.

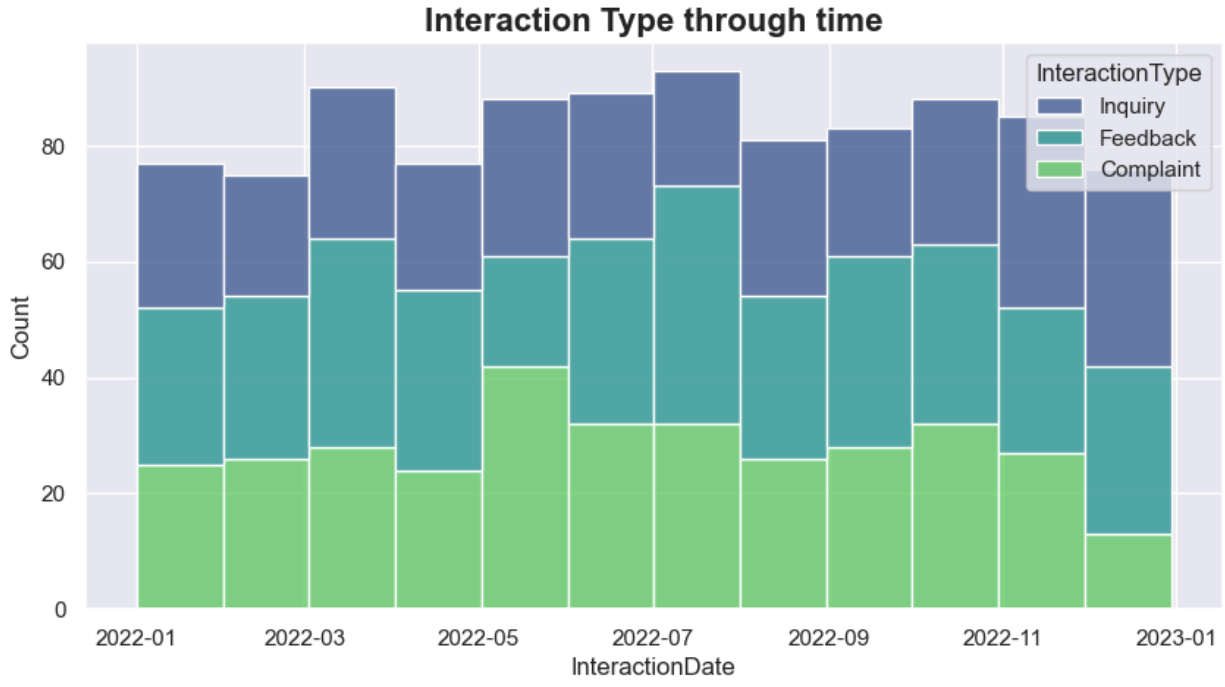
3. Customer Service:

	Unique Classes	Top class	Frequency
Interaction Type	3	Feedback	360
Resolution Status	2	Resolved	523

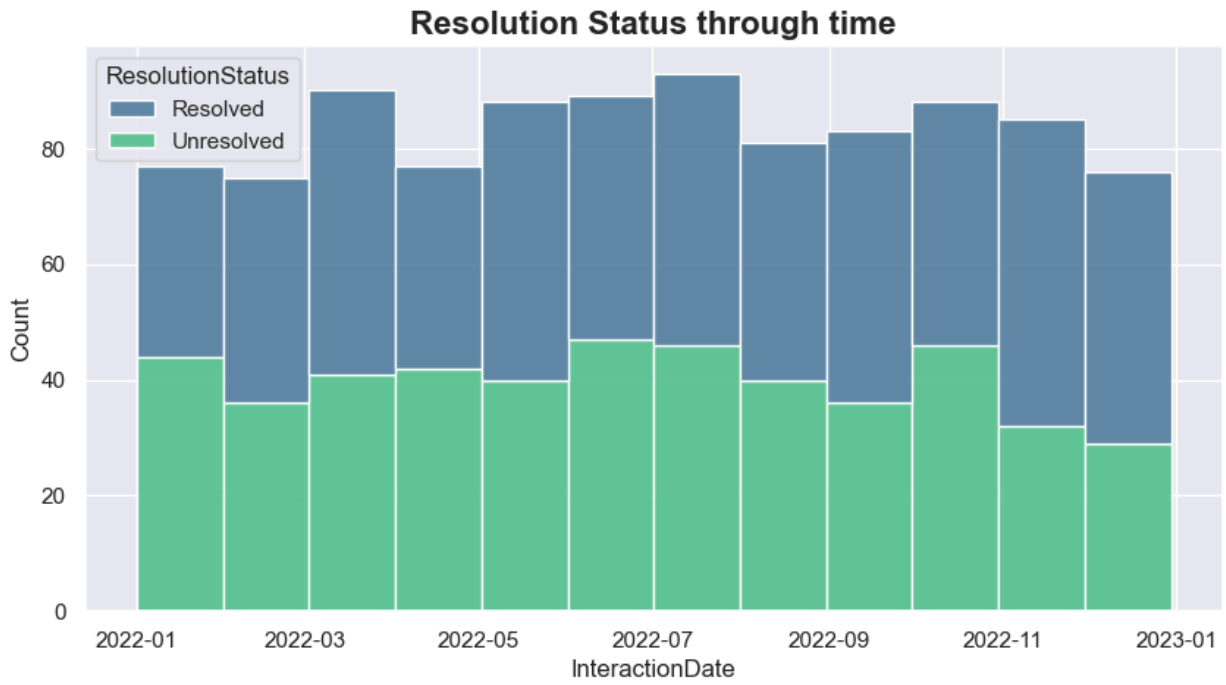
- Number of customer service interactions by Type and Status



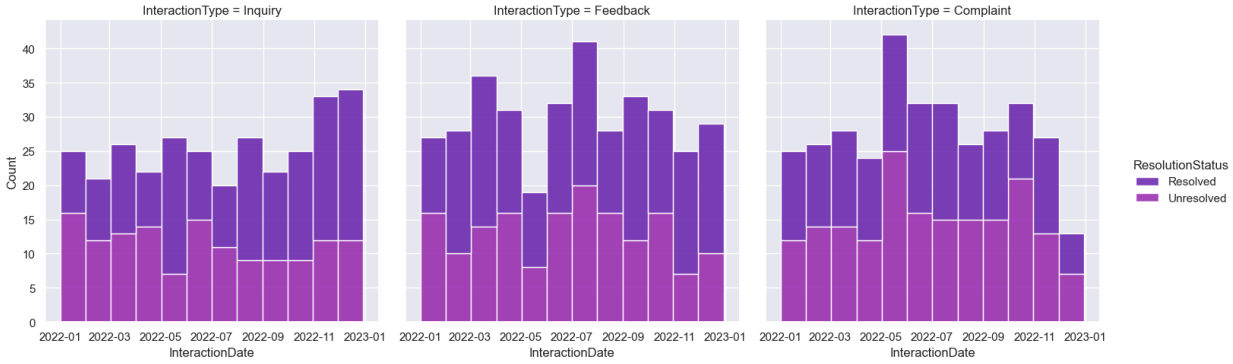
- Number of Interactions by Interaction type through time



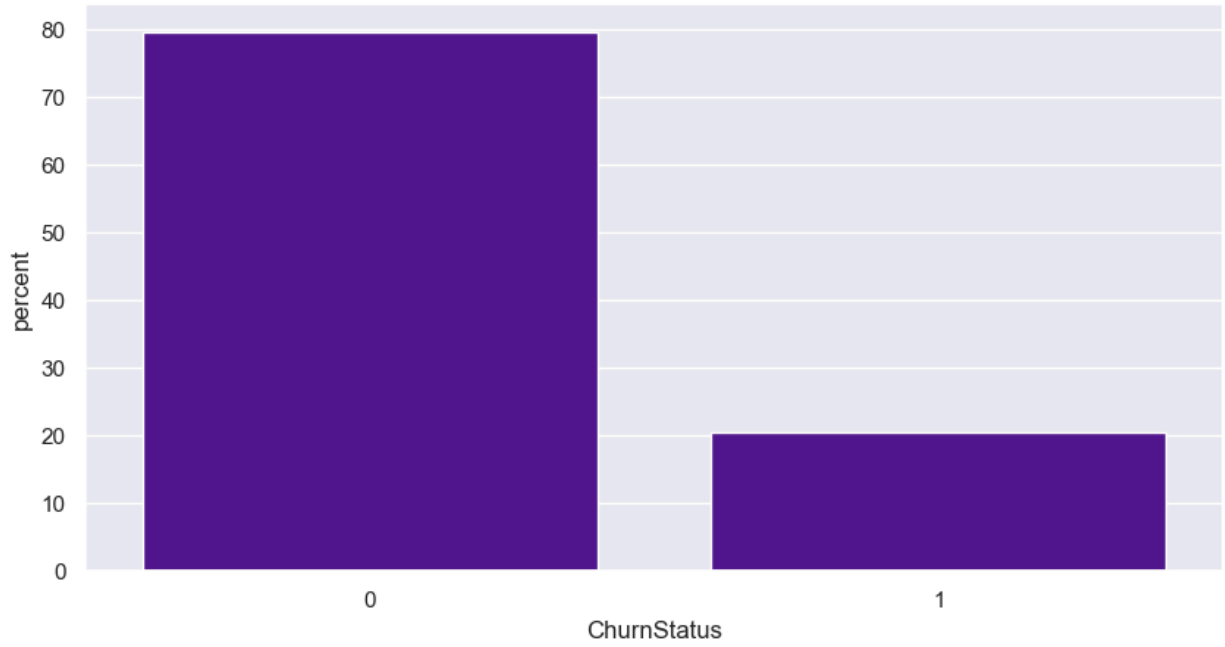
- Number of interactions through time categorized by Resolution Status



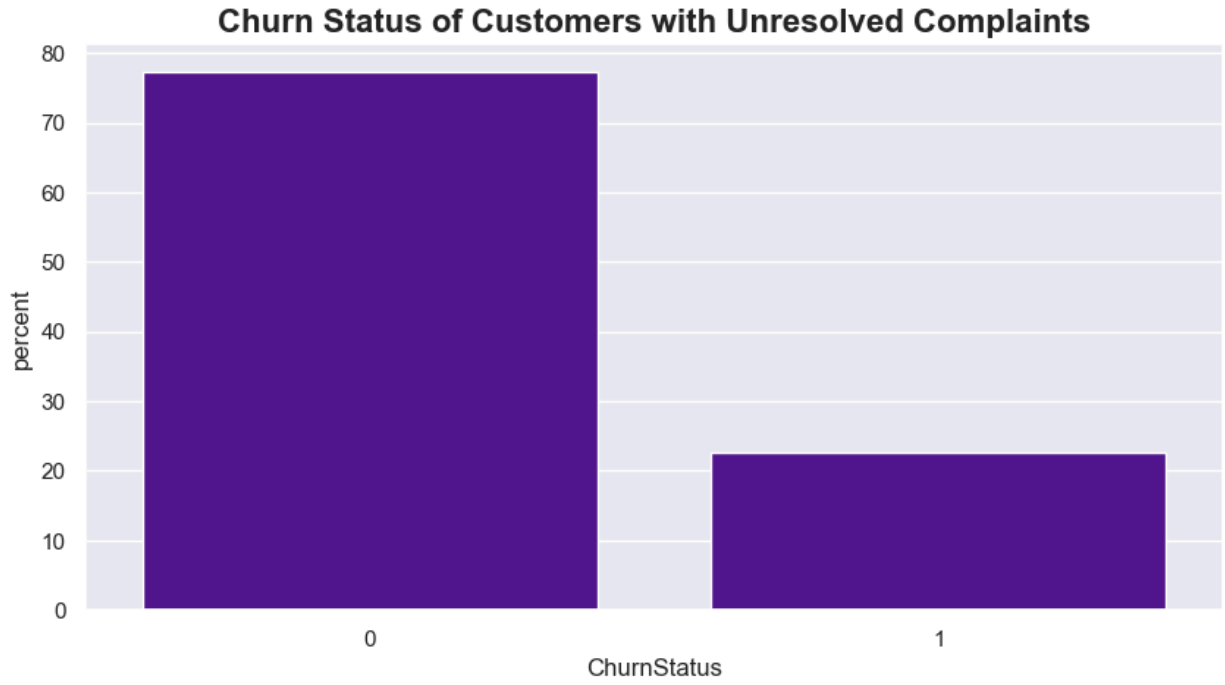
- Interaction type and Resolution status through time



- Overall churn status



- Churn status of customers with unresolved complaints



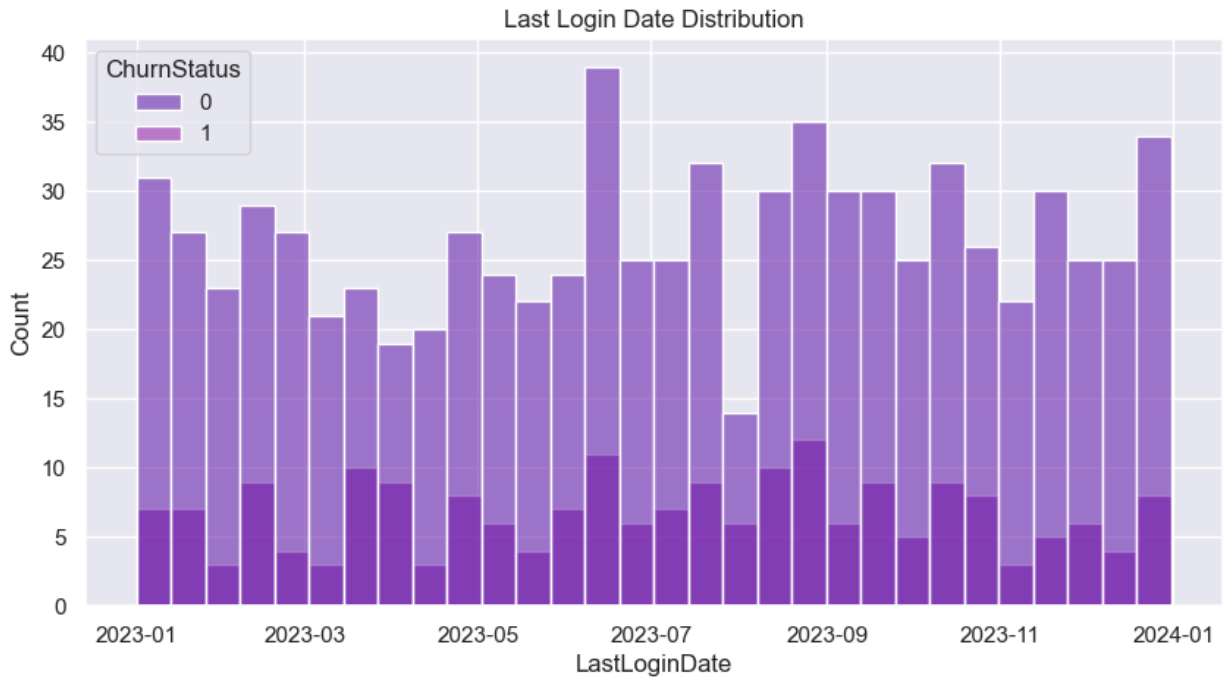
A correlation cannot be found between unresolved complaints and churning.
The data in this dataset seem to be distributed uniformly across categories.

4. Online Activity:

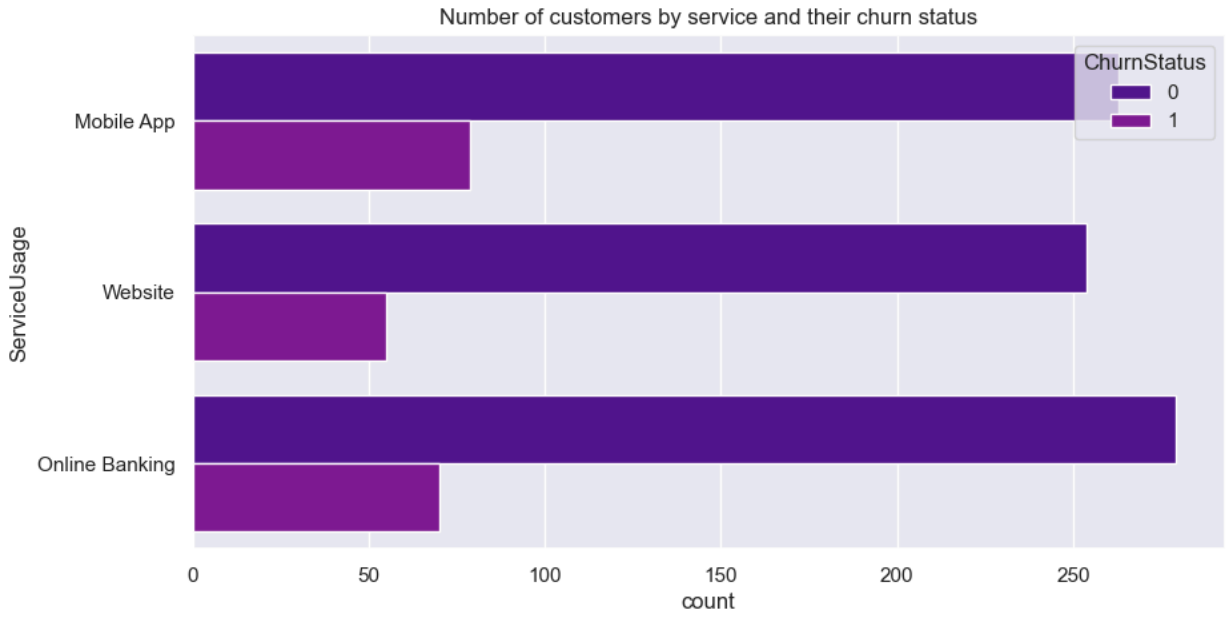
	mean	std	min	25%	50% / median	75%	max
Login Frequency	25.91	14.05	1	13..75	27	38	49

	Unique Classes	Top class	Frequency
Service Usage	3	Online banking	349

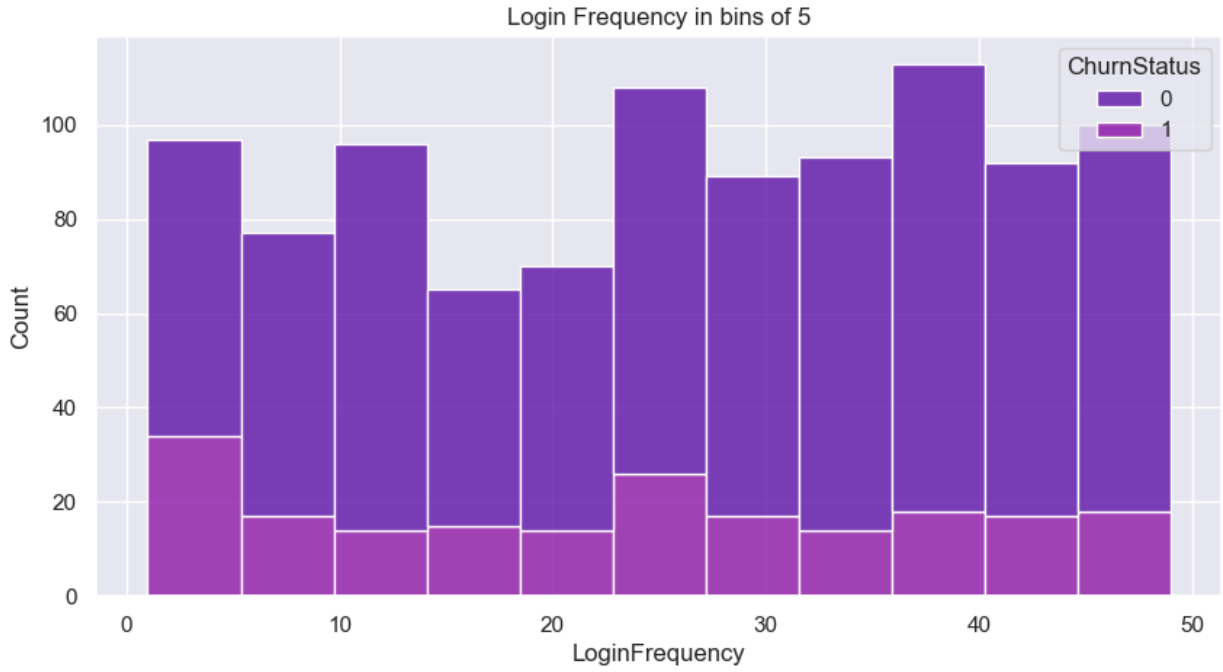
- Last login date categorised by churn status of the customer



- Number of customers by service and churn status



- Number of customer in each Login frequency (bins of 5)

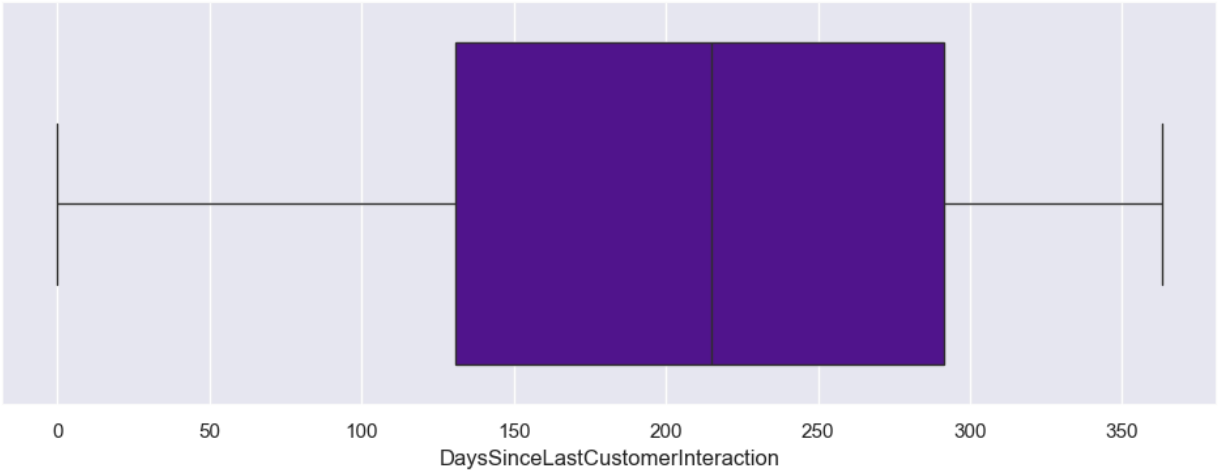


There are not any significant trends or correlations with customer churn to be found with the features.
The data in this dataset too is relatively uniform.

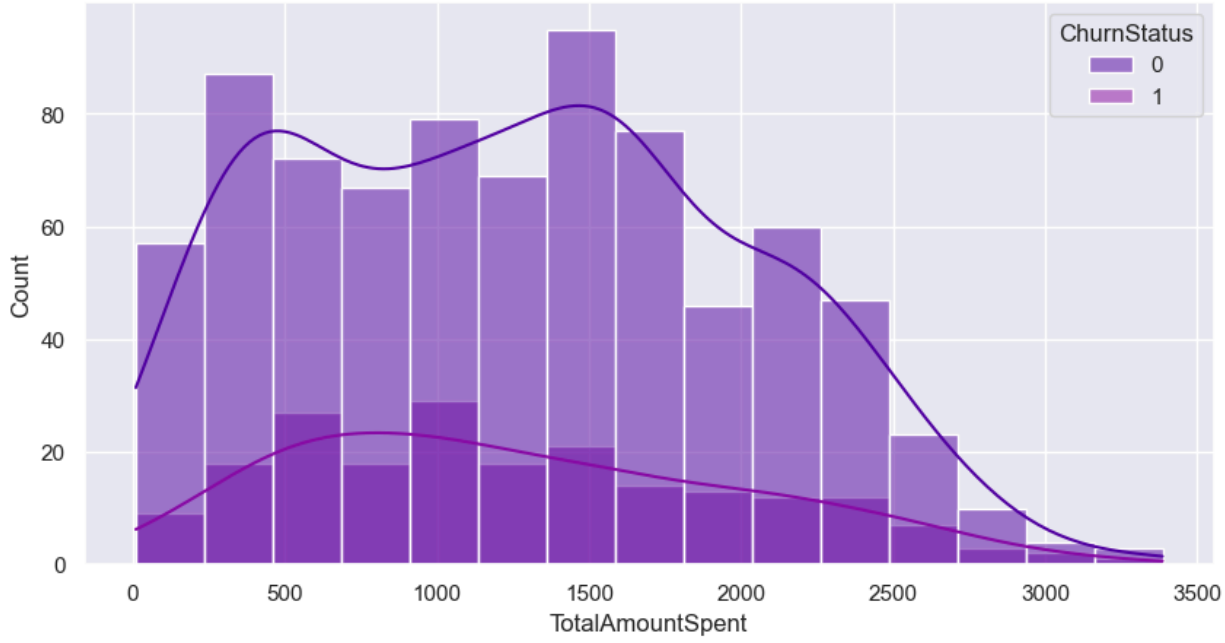
5. EDA on preprocessed data:

EDA is done on the data after preprocessing and feature engineering, several significant features have been added based on various techniques.

- **Days Since Last Customer Interaction**
  A feature derived from the customer service interaction date. Plotting Boxplot on it to view outliers.

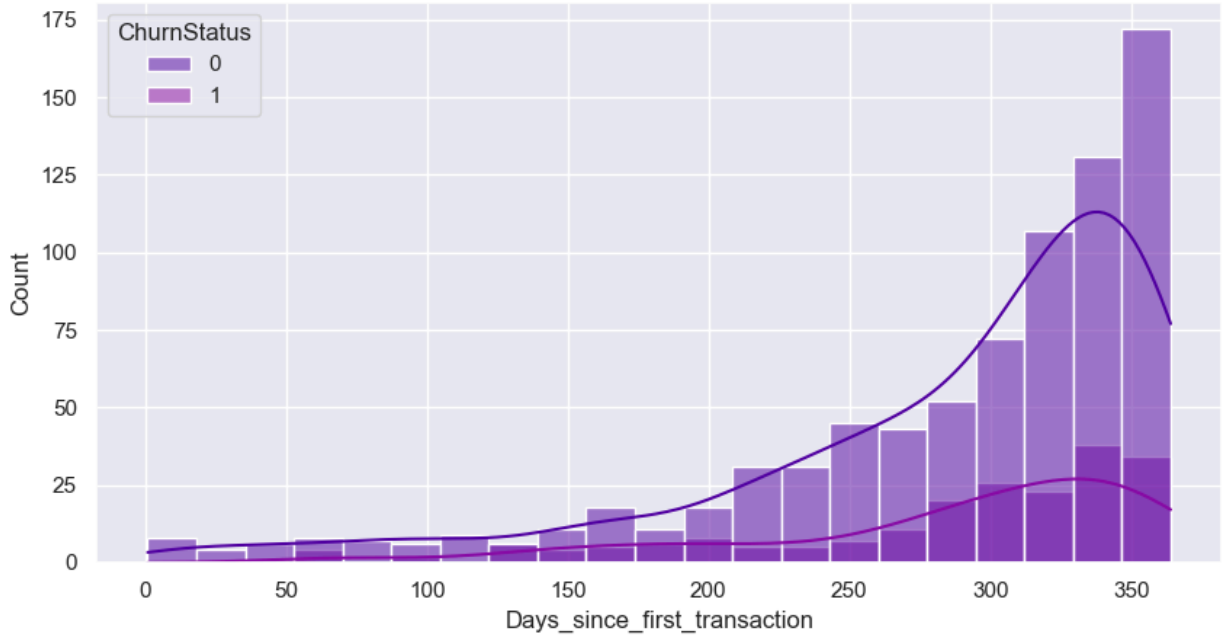


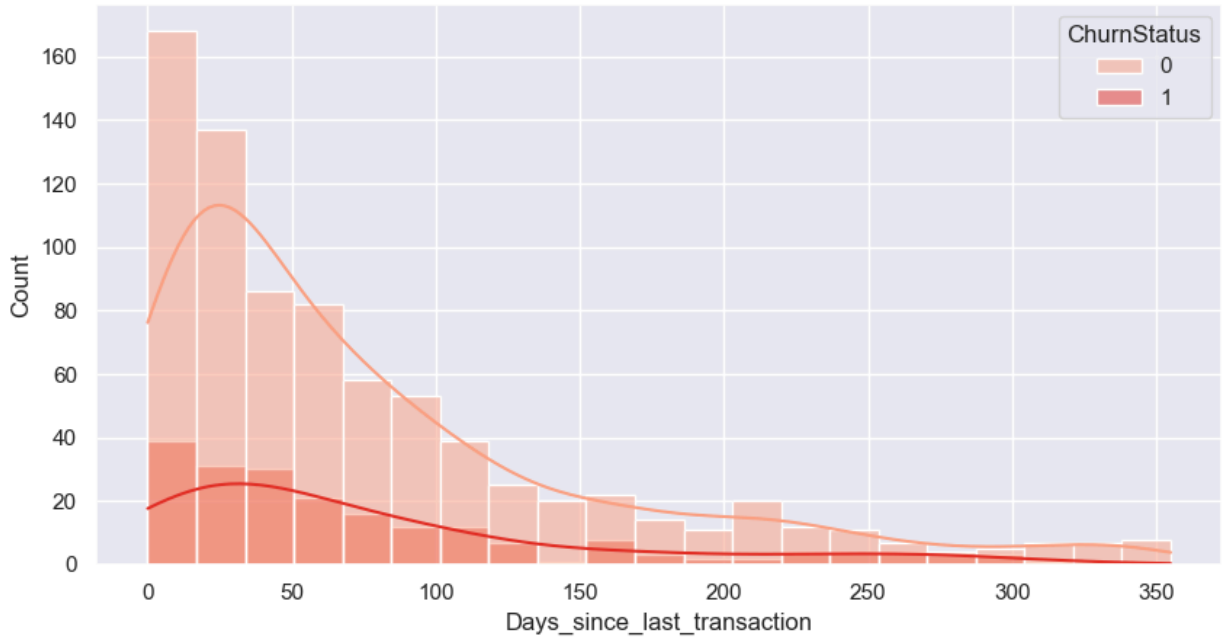
- **Total Amount Spent**
  A feature derived from customer spending.
  Plot of customer spending categorised by the churn status.



- **Days since first transaction & Days since last transaction**

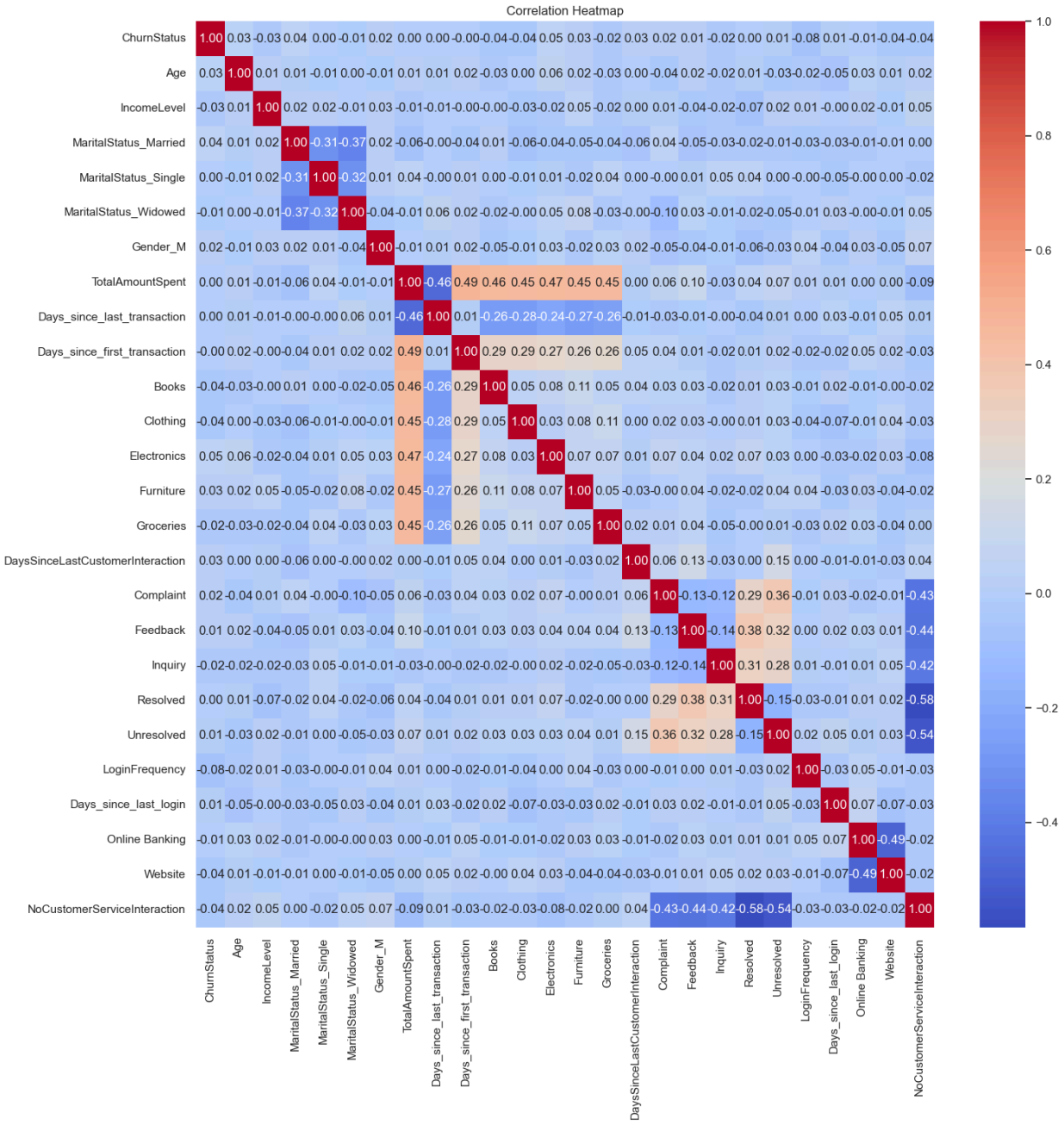
Number of customers categorized by churn status with respect to how many days have passed since.





Correlation Heatmap

- Heatmap of all the feature correlations



3. Data Cleaning & Preprocessing

1. Feature Engineering

CustomerID was a feature present in every dataset which identified a unique customer. New features were created based on the aggregation data in the dataset.

- **TotalAmountSpent**: From the transaction history dataset, the amount spent by each customer was aggregated to form this new feature.
- **Days_since_first_transaction:** The transactions of each customer were grouped and the days from the earliest transaction to the latest date in the dataset is found.
- **Days_since_last_transaction**: Same as before, but days from the latest transaction for each customer.
- **DaysSinceLastCustomeInteraction**: From the customer service dataset, the Interaction Date for each customer was grouped and the number of days from the latest date in the group to the latest date in the dataset was found.
- **Days_since_last_login**: From the online activity dataset, days since the last login of the customer was found.

2. Encoding Categorical Variables

One-hot encoding, Frequency encoding and Ordinal encoding was used to encode categorical features.

- One-hot encoded Variables

- **Gender >> Gender_M**
- **MaritalStatus >> MaritalStatus_Married, MaritalStatus_Single, MaritalStatus_Widowed**
- **ServiceUsage >> Online Banking, Website**

- Frequency Encoded Variables

- **ProductCategory >> Books, Clothing, Electronics, Furniture, Groceries**
- **InteractionType >> Complaint, Feedback, Inquiry**
- **ResolutionStatus >> Resolved, Unresolved**

- Ordinal Encoded Variables

- **IncomeLevel (**low, medium, high) >> (0,1,2)

3. Handling Missing Values

The original datasets did not have any missing values.
The features after feature engineering had missing values.
The fill value **0** for columns **Complaint, Feedback, Inquiry, Resolved and Unresolved** represent the real scenario that the customer has never had that specific interaction.

For the column **DaysSinceLastCustomerInteraction**, flagging is used to handle null values. A separate column **NoCustomerServiceInteraction** is created which flags the samples with NaN values and the existing null values are filled using the **median** of the column.

4. Normalisation

MinMaxScaler was used to normalise data because most of the data is uniformly distributed and also the order of ordinal data is preserved.

4. Cleaned and Preprocessed Dataset

The final dataset includes the features:

Demographic features:
Age
IncomeLevel
MaritalStatus_Married
MaritalStatus_Single
MaritalStatus_Widowed
Gender_M

Transaction features:
TotalAmountSpent
Days_since_last_transaction
Days_since_first_transaction
Books
Clothing
Electronics
Furniture
Groceries

Customer Service features:
DaysSinceLastCustomerInteraction
Complaint
Feedback
Inquiry
Resolved
Unresolved

Online Activity features:
LoginFrequency

Days_since_last_login
Online Banking
Website
NoCustomerServiceInteraction

Target Variable
ChurnStatus

The CustomerID column was dropped since it was redundant for machine learning.

The cleaned and preprocessed dataset was saved for model building:
Link