# THERA BANK – PROJECT 6

Sarah A. Thomas

# BUSINESS PROBLEM OVERVIEW AND SOLUTION APPROACH

- Recently, here at Thera Bank, there has been a major decline in the number of credit card users. This has made a major impact to revenue given the money earned from annual fees, balance transfer fees, cash advance fees, late payment fees, and foreign transaction fees.
- The members of the Data Science team have come up with a model to assist the bank in determining which customers are more likely to attrite. This model will help the bank re-evaluate its service offerings.
- This model is a classification model using Python. The Data Science team has used methods such as Decision Tree, AdaBoost, GradientBoost, along with oversampling and undersampling methods. In the end, we will present an AdaBoost model used with the undersampling strategy as the final, best performing model.

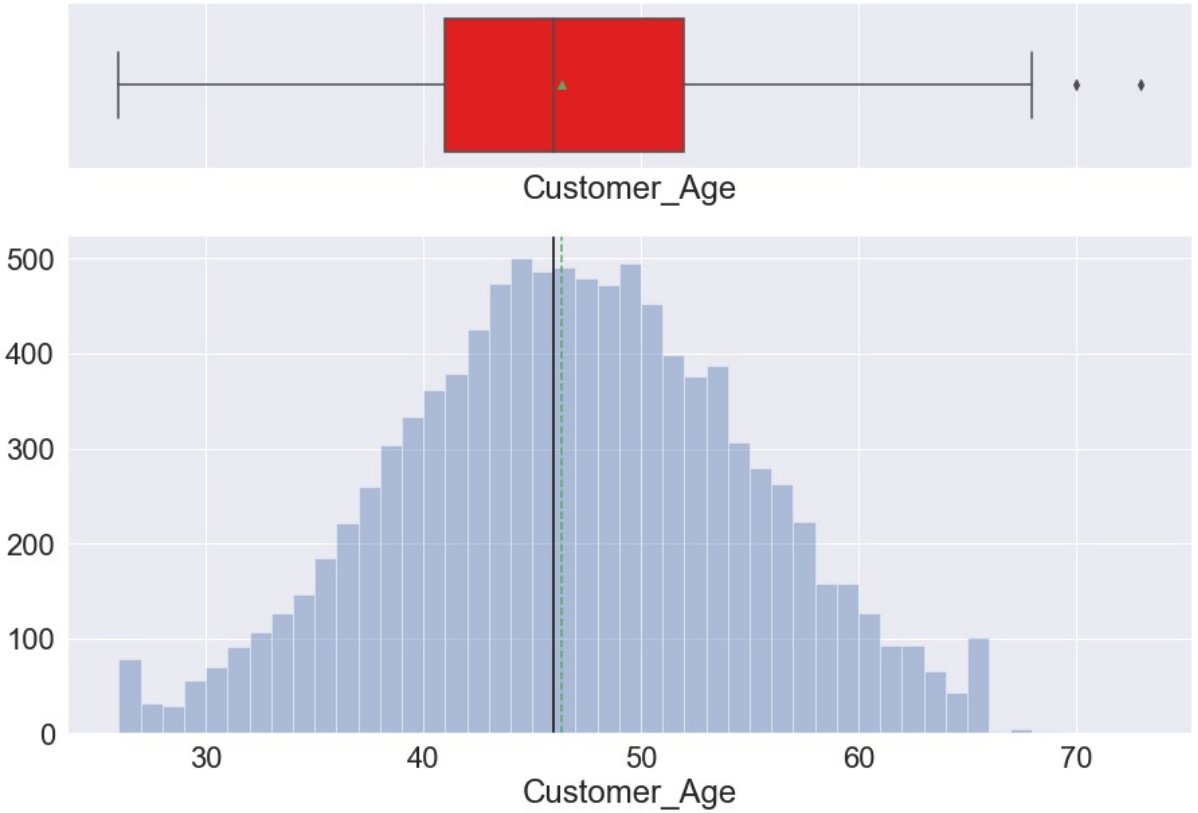
# DATA OVERVIEW (1 OF 2)

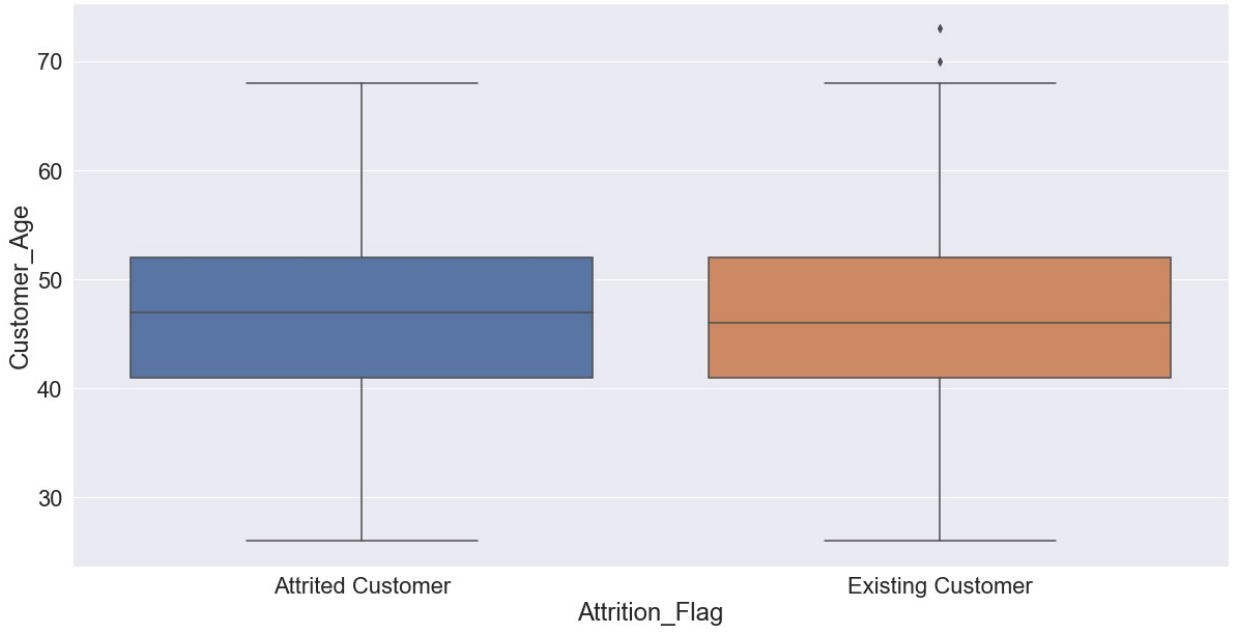
| *Variable* | *Description* |
|---|---|
| CLIENTNUM | Unique identifier – Client Number |
| Attrition_Flag (Target Variable) | "Attrited Customer" if account is closed, otherwise "Existing Customer" |
| Customer_Age | Customer age in years |
| Gender | Gender of the customer |
| Dependent_count | Number of customer's dependents |
| Education_Level | Possible values: Graduate, High School, Unknown, Uneducated, College, Post-Graduate, Doctorate |
| Marital_Status | Customer's marital status |
| Income_Category | Annual income range of the account holder |
| Card_Category | Type of card (e.g., Platinum, Silver, etc.) |
| Months_on_book | Period of relationship with the bank |

# DATA OVERVIEW (2 OF 2)

| *Variable* | *Description* |
|---|---|
| Total_Relationship_Count | Number of products held by the customer |
| Months_Inactive_12_mon | No. of months inactive in the last 12 months |
| Contacts_Count_12_mon | No. of contacts between customer and bank in past 12 months |
| Credit_Limit | Credit limit on the credit card |
| Total_Revolving_Bal | The balance that carries over from one month to the next |
| Avg_Open_To_Buy | Amount left on the credit card to use |
| Total_Trans_Amt | Total transaction amount (last 12 months) |
| Total_Trans_Ct | Total transaction count (last 12 months) |
| Total_Ct_Chng_Q4_Q1 | Ratio of the total transaction count in 4$^{th}$ quarter and in 1$^{st}$ quarter |
| Total_Amt_Chng_Q4_Q1 | Ratio of the total transaction amount in 4$^{th}$ quarter and in 1st quarter |
| Avg_Utilization_Ratio | How much of the available credit the customer spent |

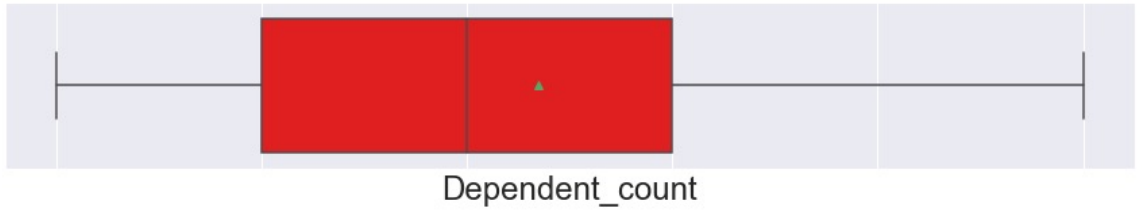
# CUSTOMER_AGE

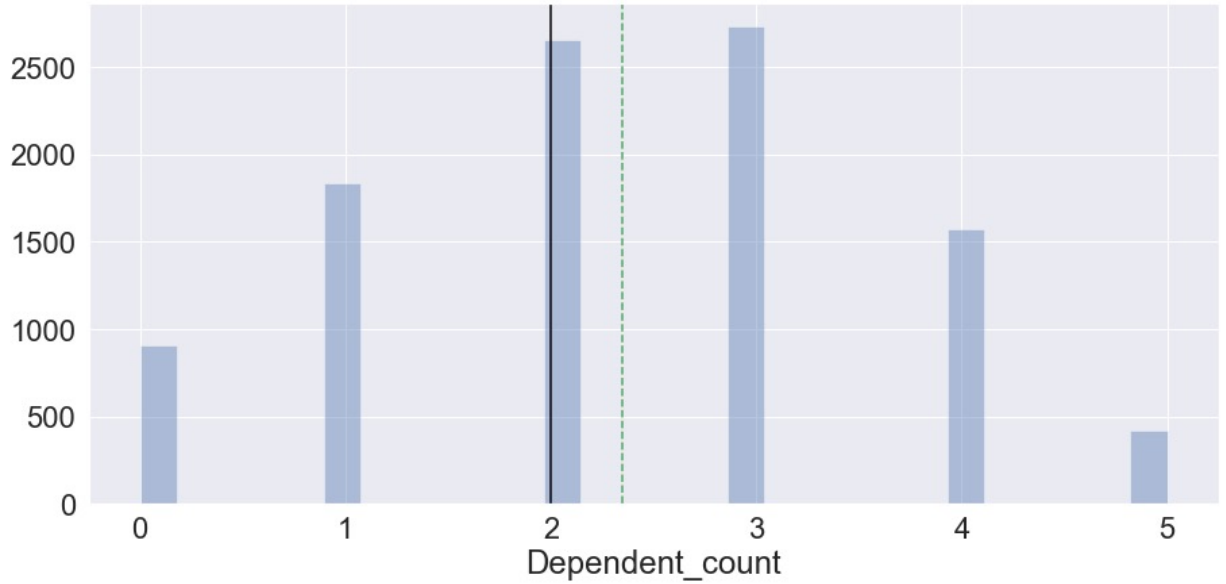


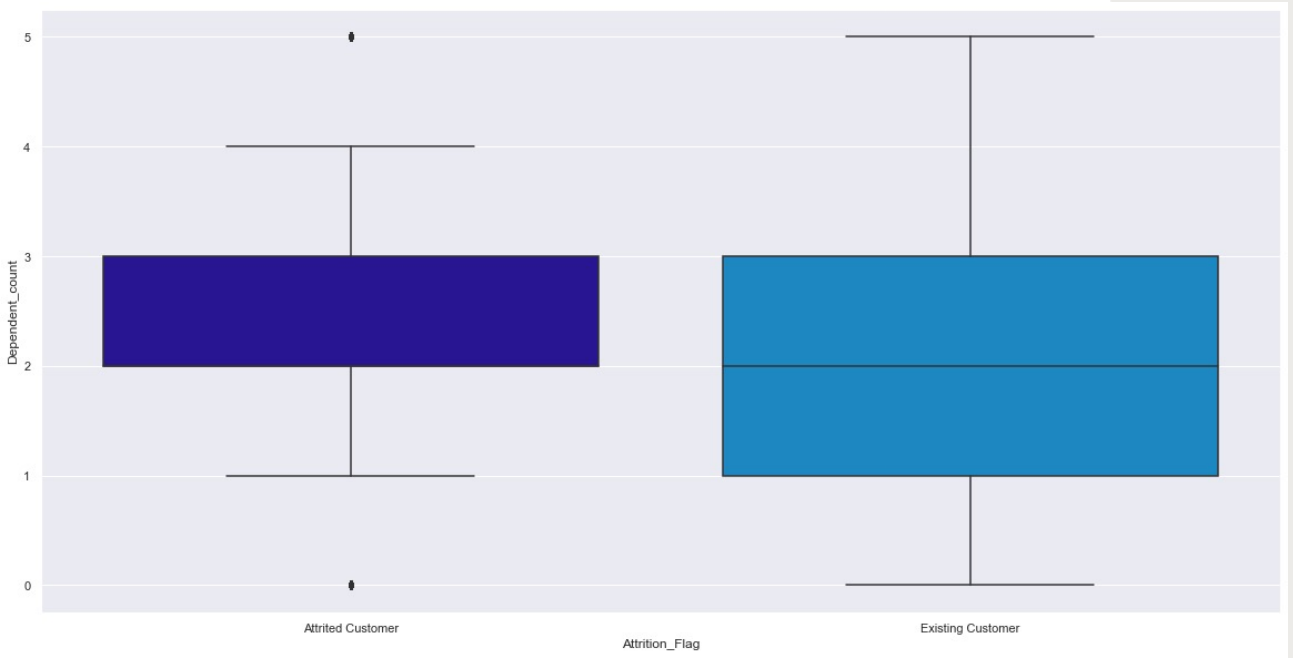




There are outliers so distribution is slightly right-skewed (not treated since so close to the rest of values). There seems to be no age difference in attrited customers vs. existing customers.

# DEPENDENT_COUNT

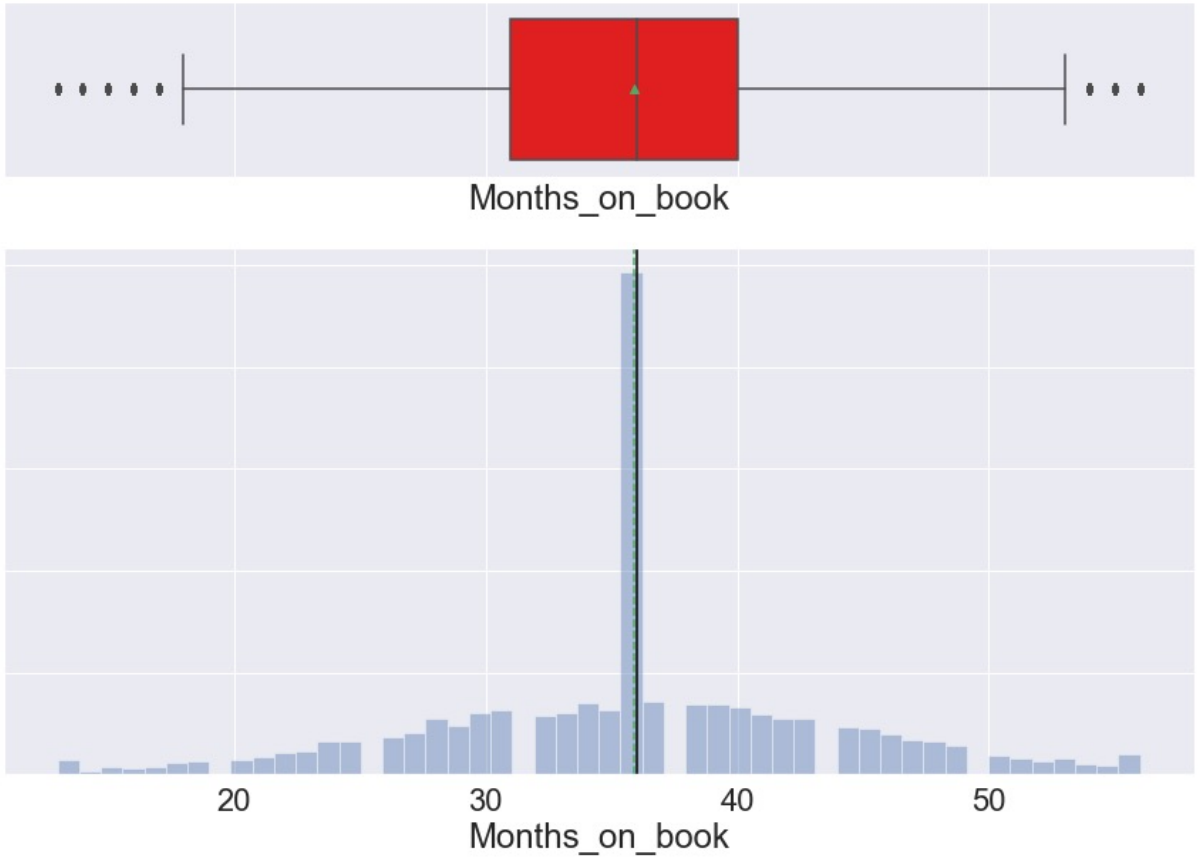


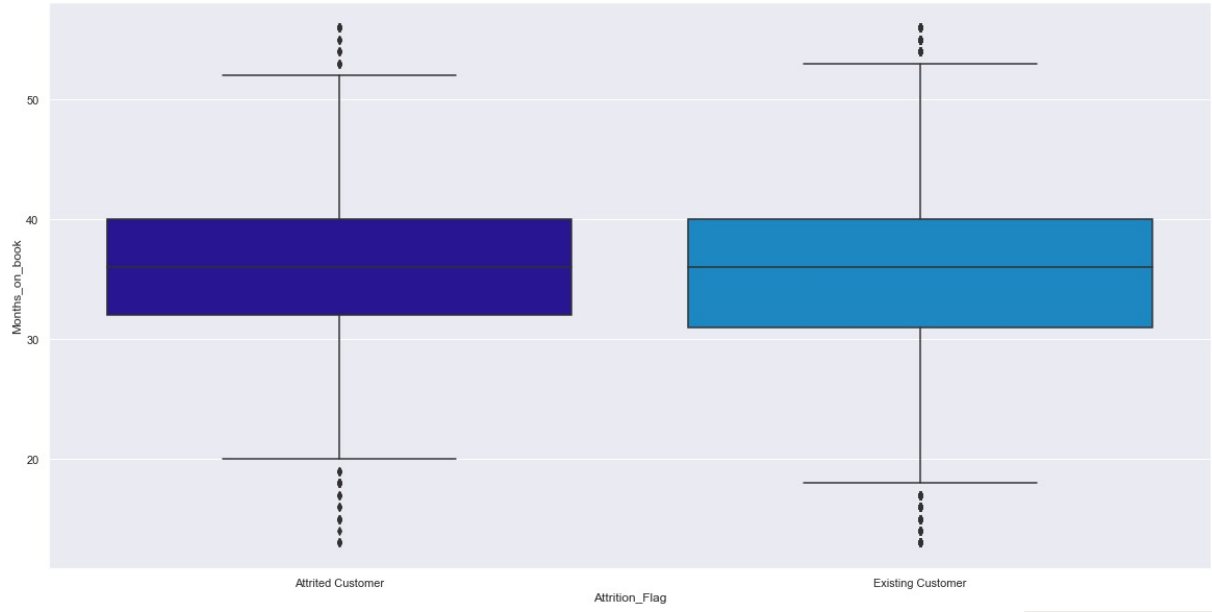




Distribution is symmetrical. There is a wider range of dependent count with existing customers.

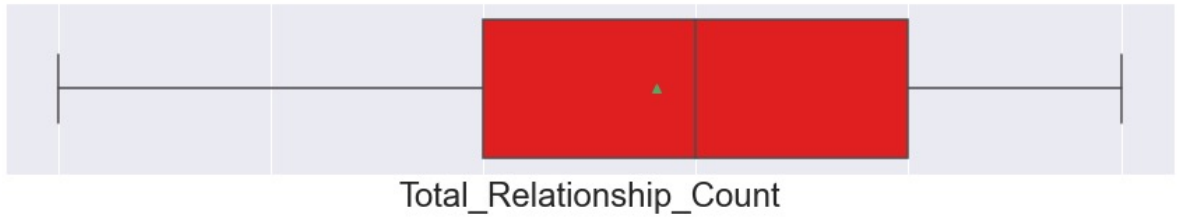
# MONTHS_ON_BOOK

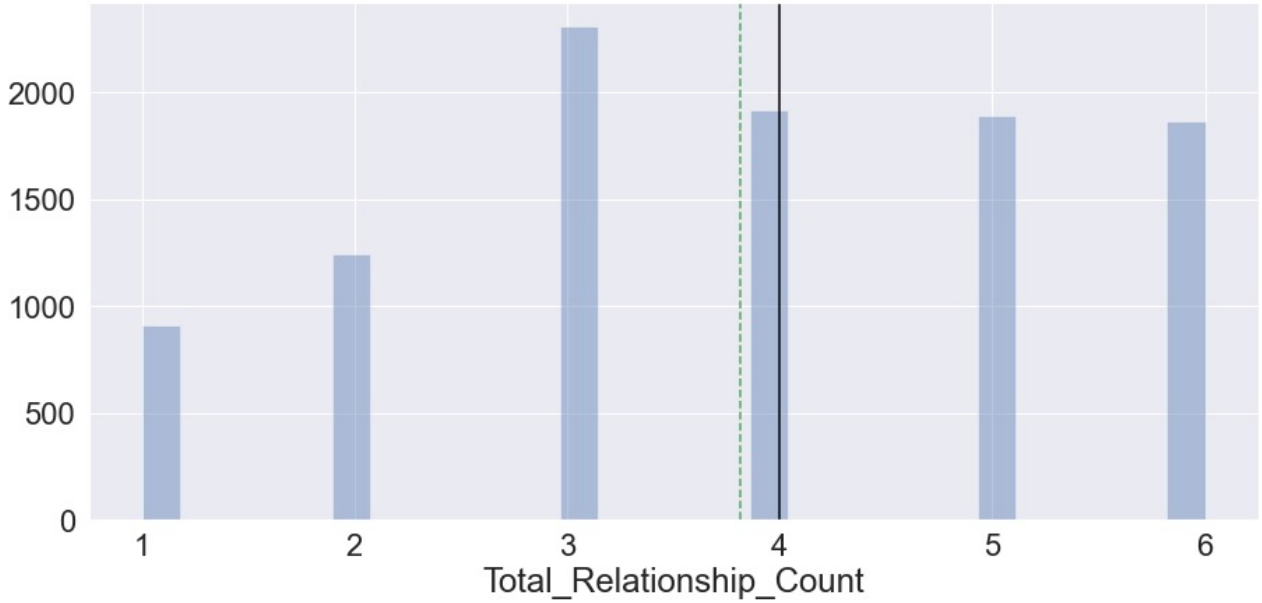


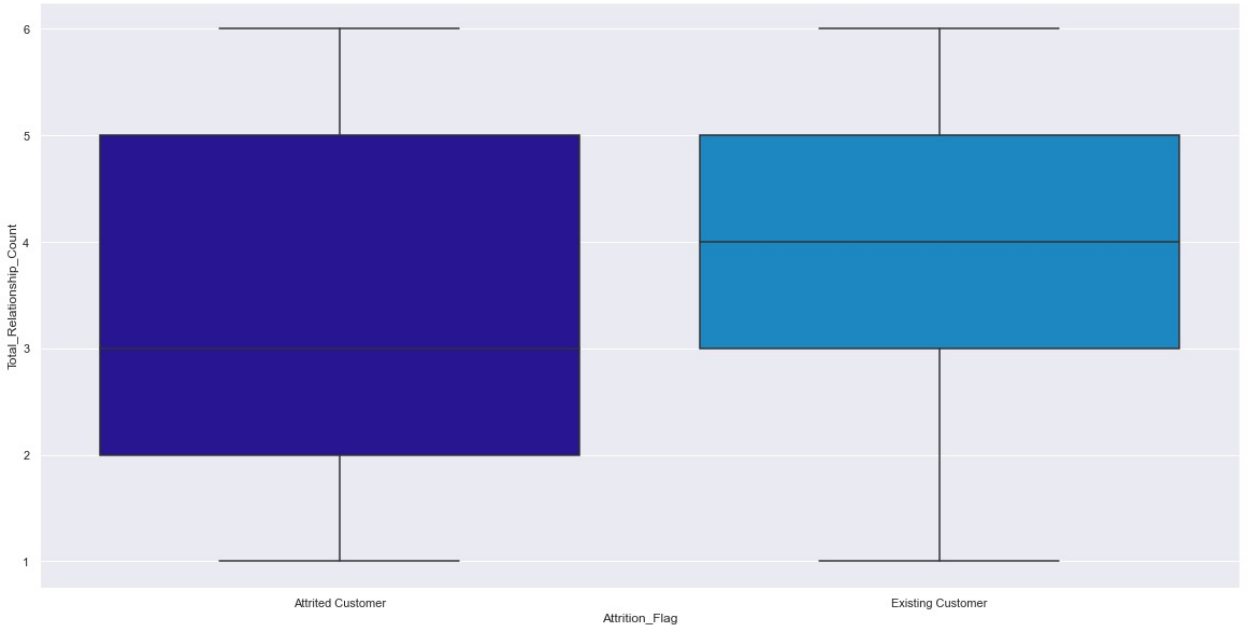




There are outliers on either side. Since these values lie close to the other values and some variation is expected, these will not be treated. Existing customers tend to have been with the bank for a slightly wider range of time.

# TOTAL_RELATIONSHIP_COUNT

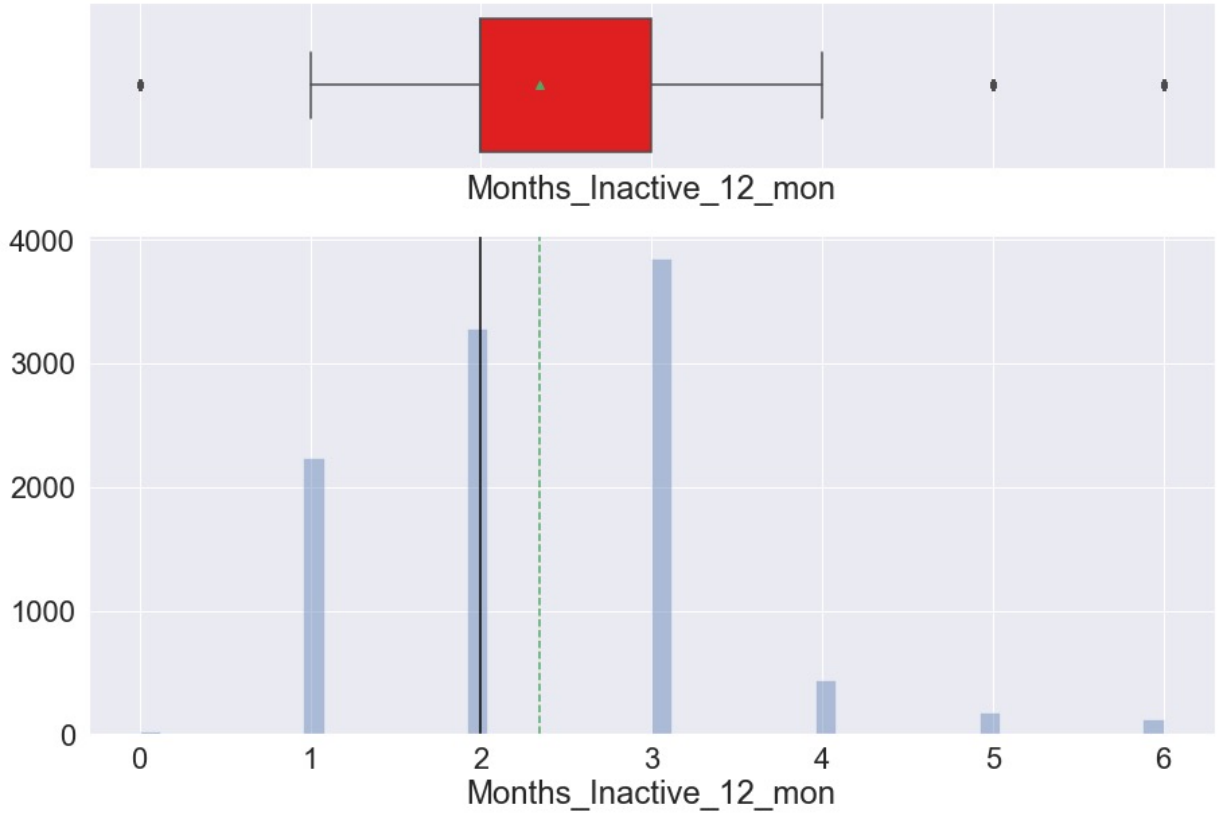


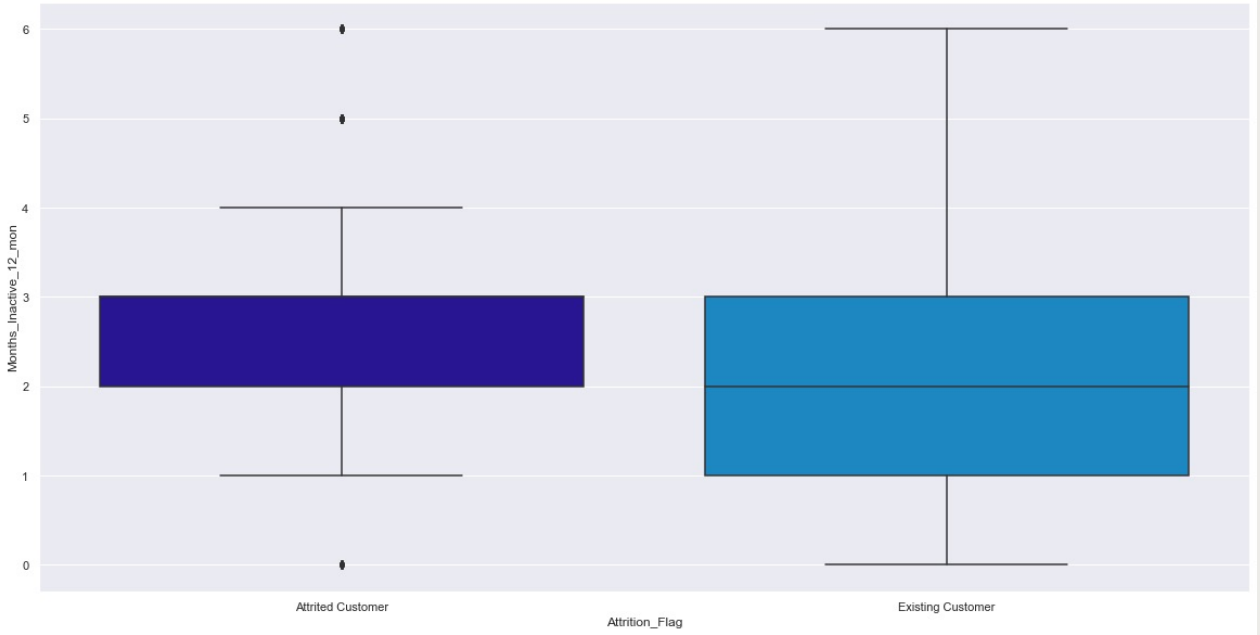




Customers on average have approx. 4 products with bank (median - 4, min - 1, max - 6). Attrited customers include those who had less number of products with the bank.

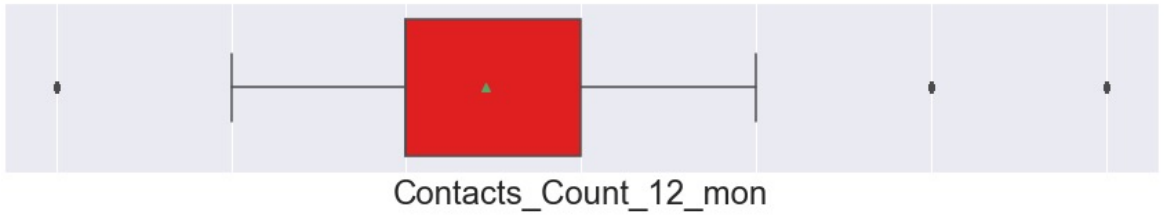
# MONTHS_INACTIVE_12_MON

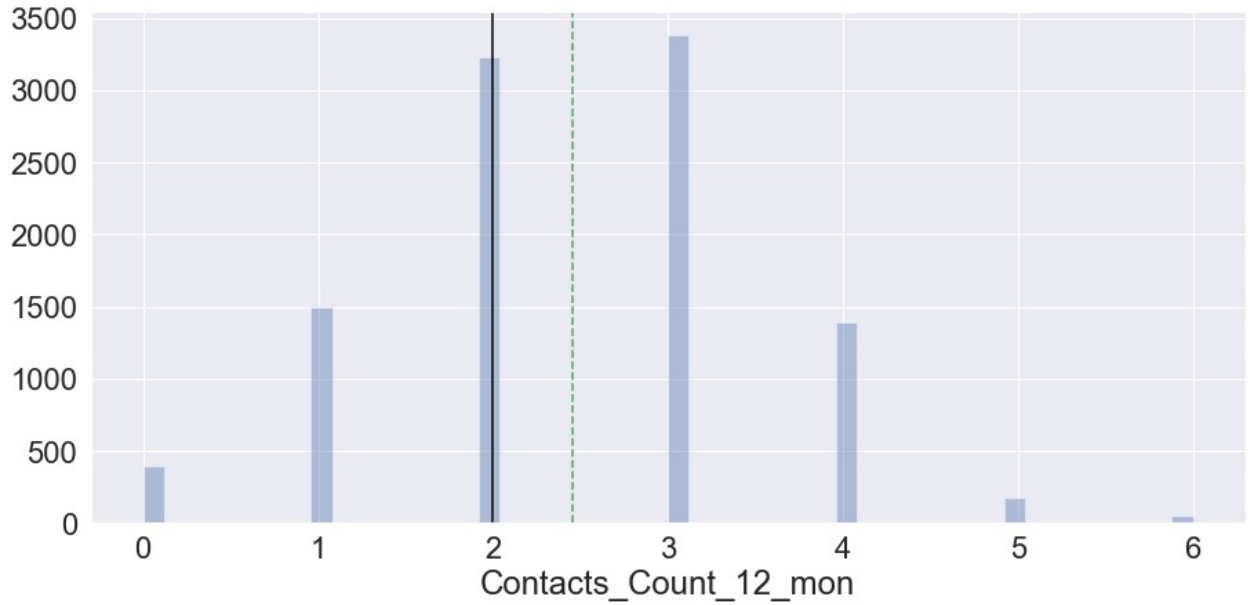


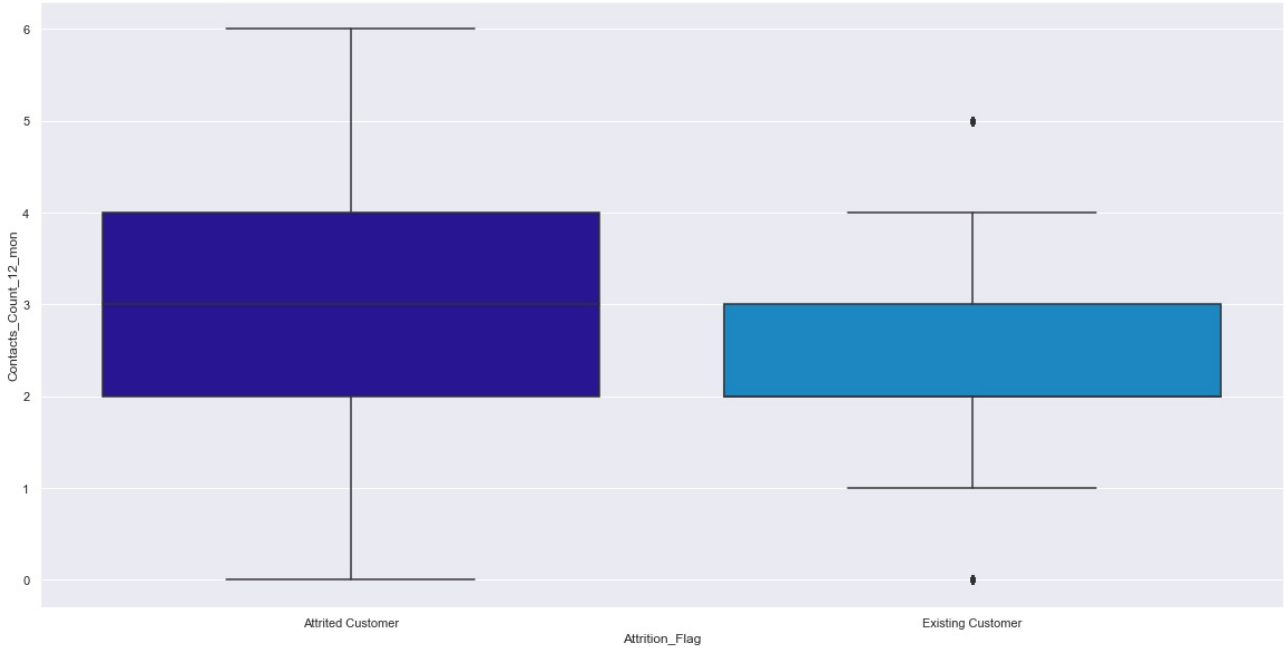




There are outliers on either side. Since these values are still close to the other values they will not be treated as outliers. Existing customers include those who have been inactive less during the last 12-month period.

# CONTACTS_COUNT_12_MON

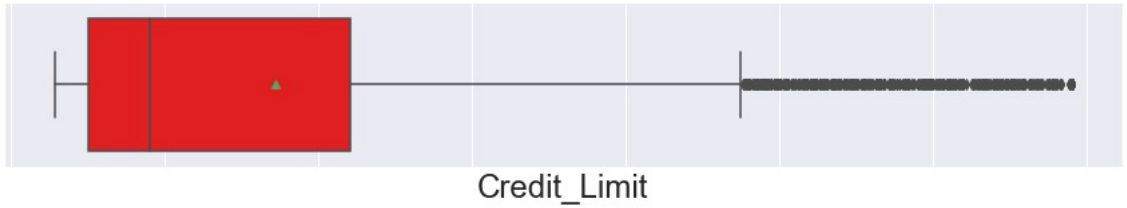


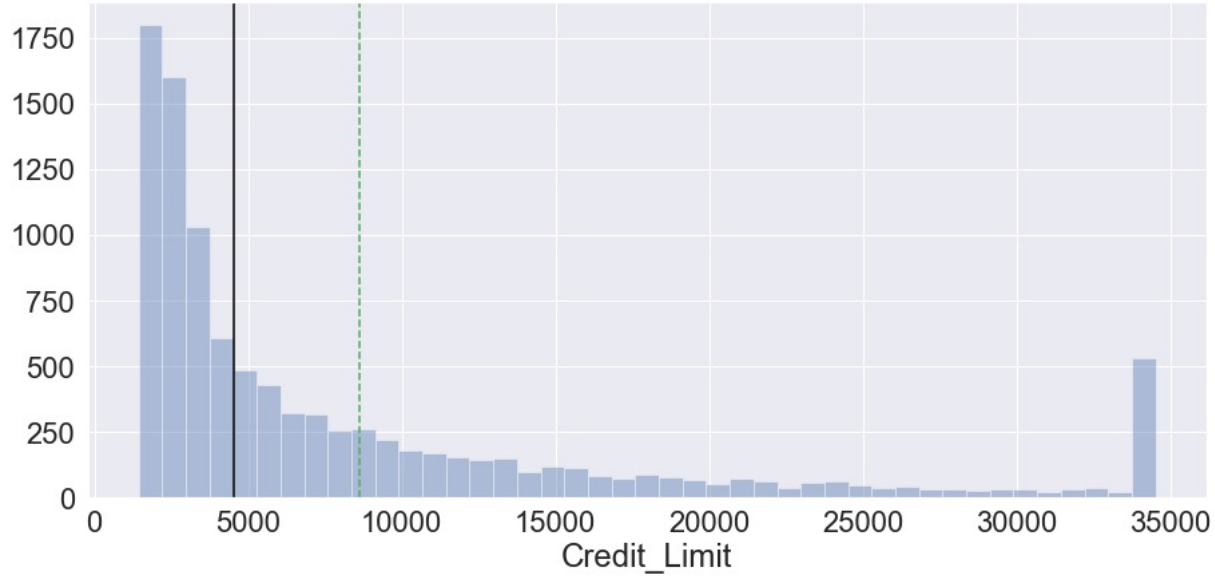



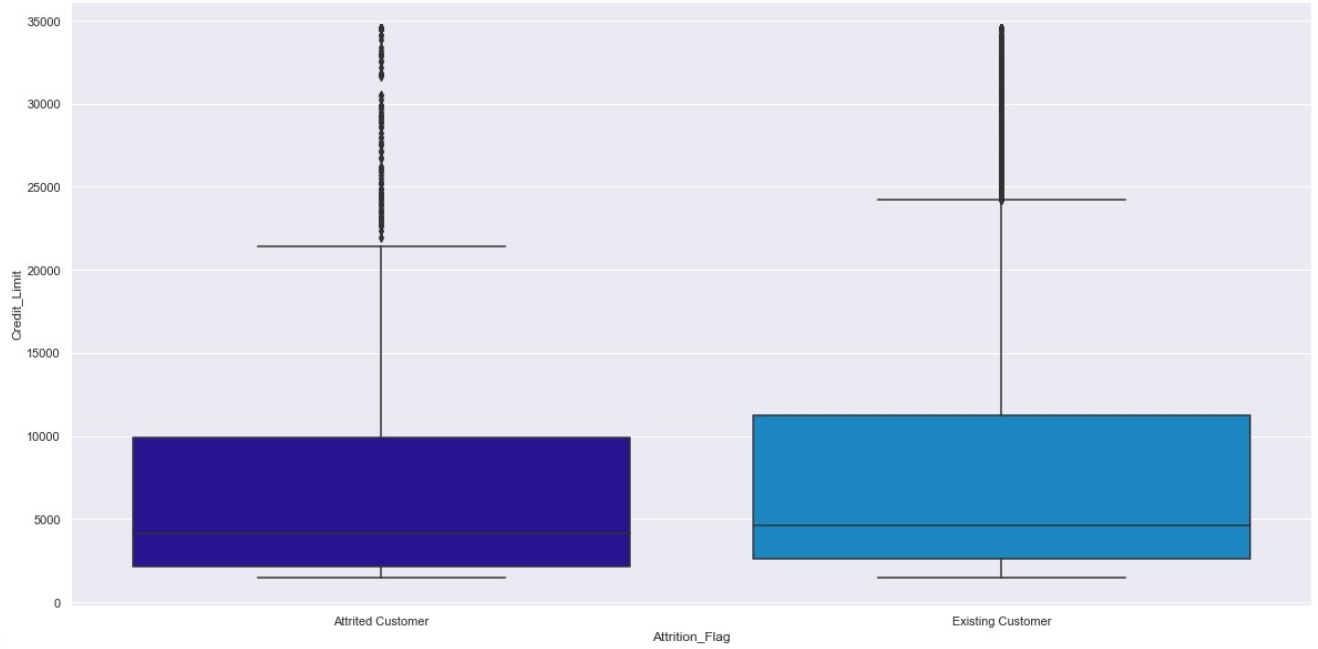




There are outliers on either side. Since these values are still close to the other values they will not be treated as outliers. Attrited customers tend to have had more contacts with the bank in the past 12 months.

# CREDIT_LIMIT

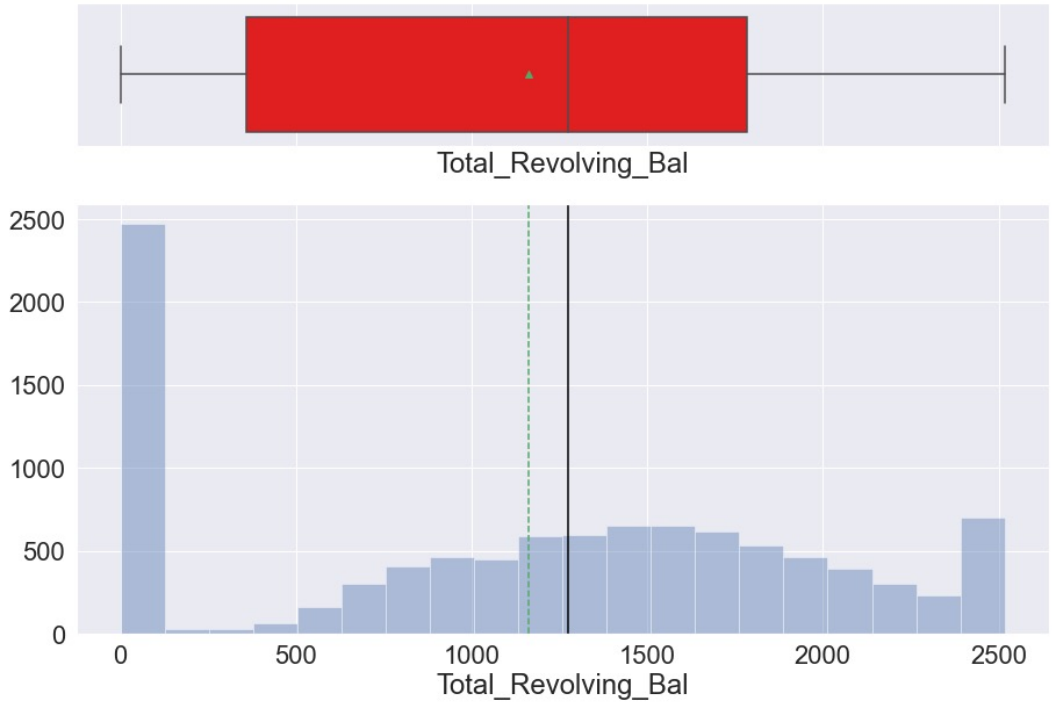


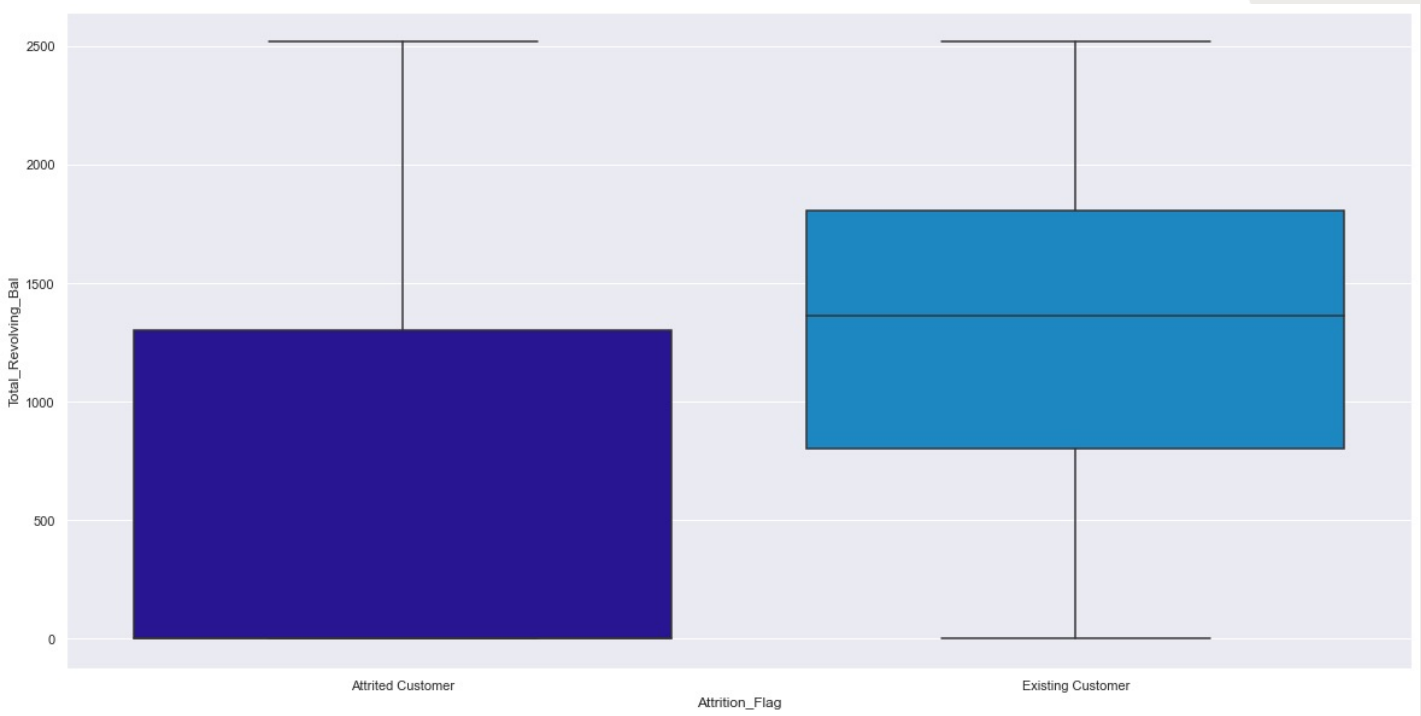




Distribution is right-skewed with outliers on the right. However, these will not be treated since some variation is expected in variables like Credit Limit. Attrited customers have a slightly smaller credit limit than existing customers.

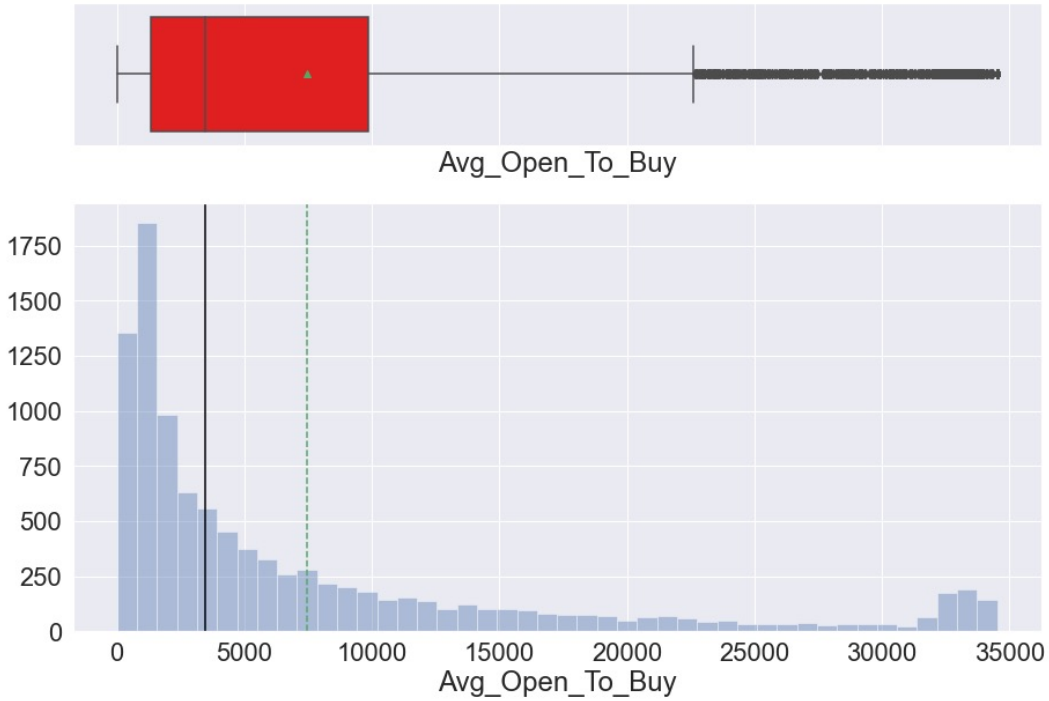
# TOTAL_REVOLVING_BAL

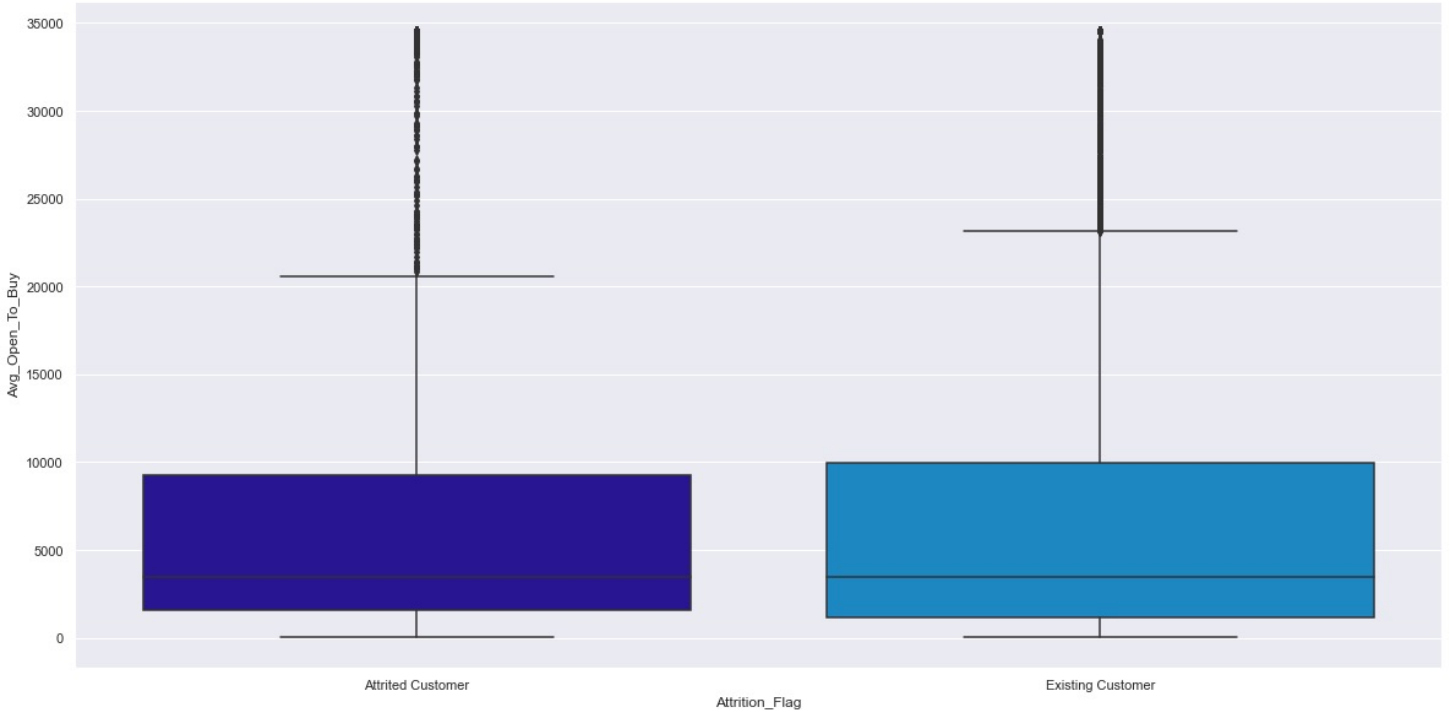



Many customers (nearly 2500) have a total revolving balance of $0. Existing customers have a higher revolving balance than attrited customers.

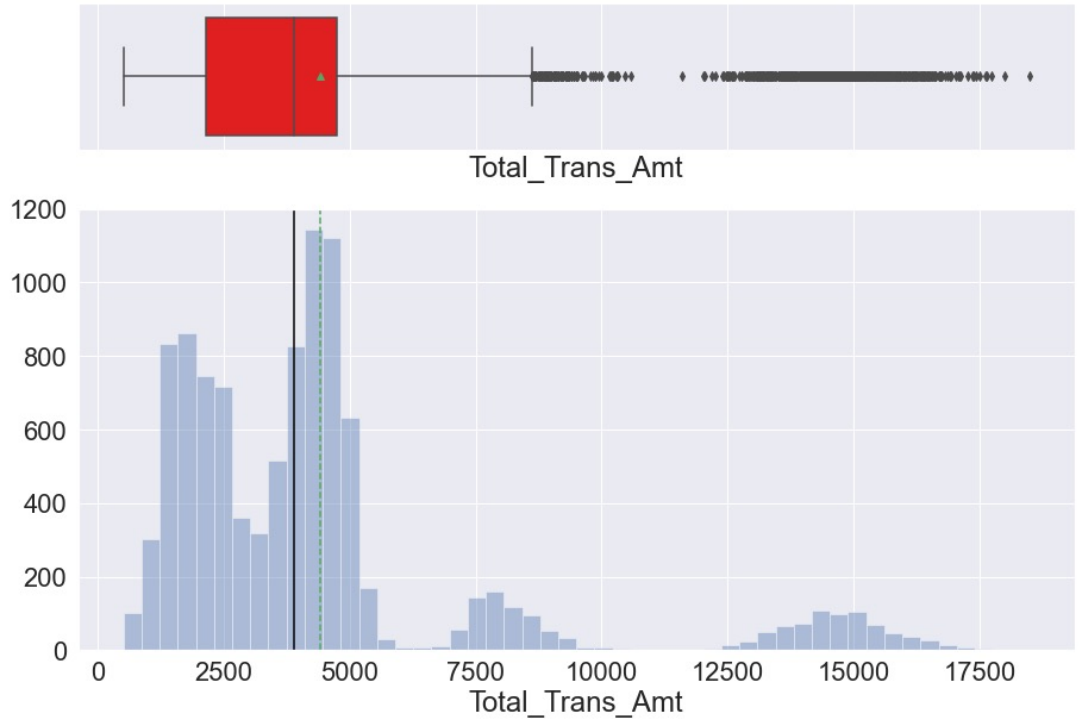
# AVG_OPEN_TO_BUY

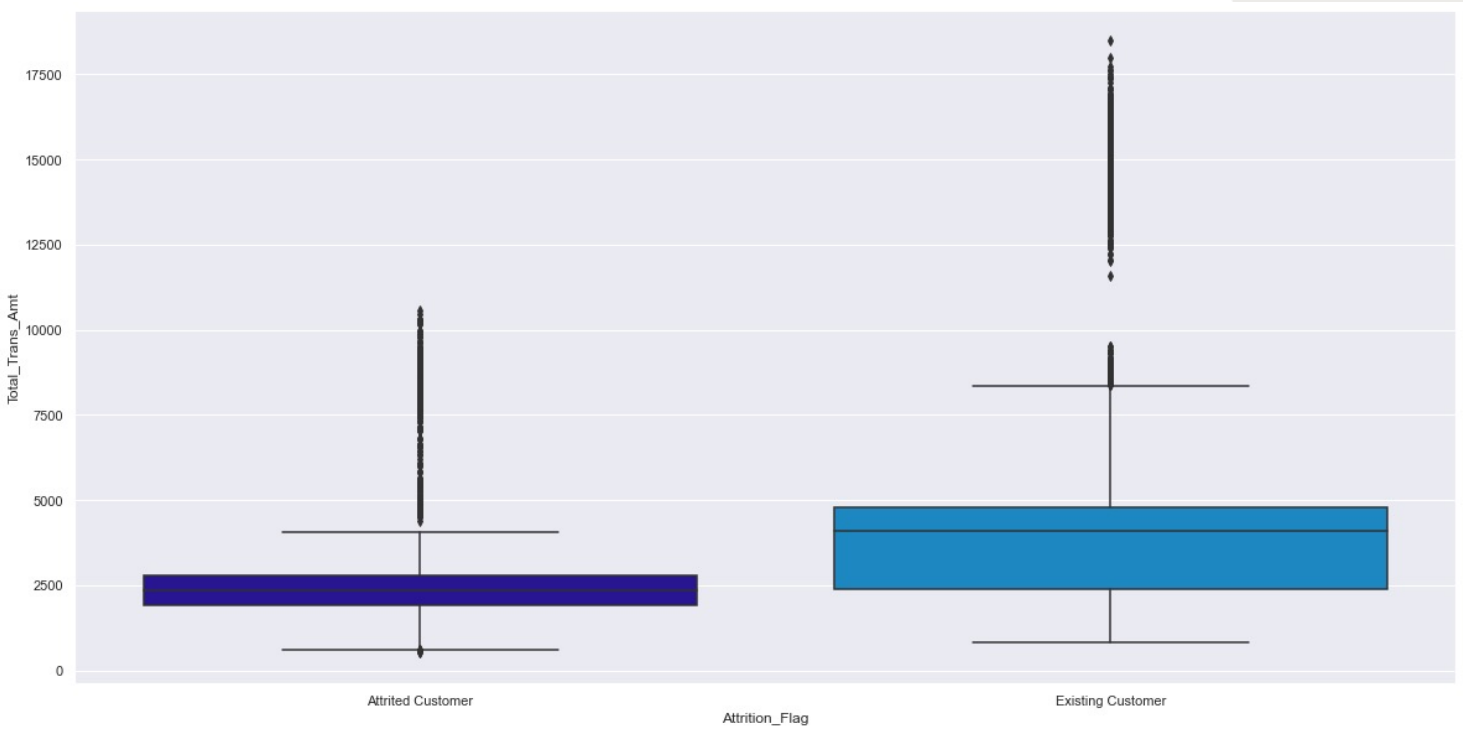



Distribution is right-skewed with outliers to the right. However, these will not be treated since some variation is expected in variables like Average Open to Buy. Average Open to Buy is comparable between attrited and existing customers.

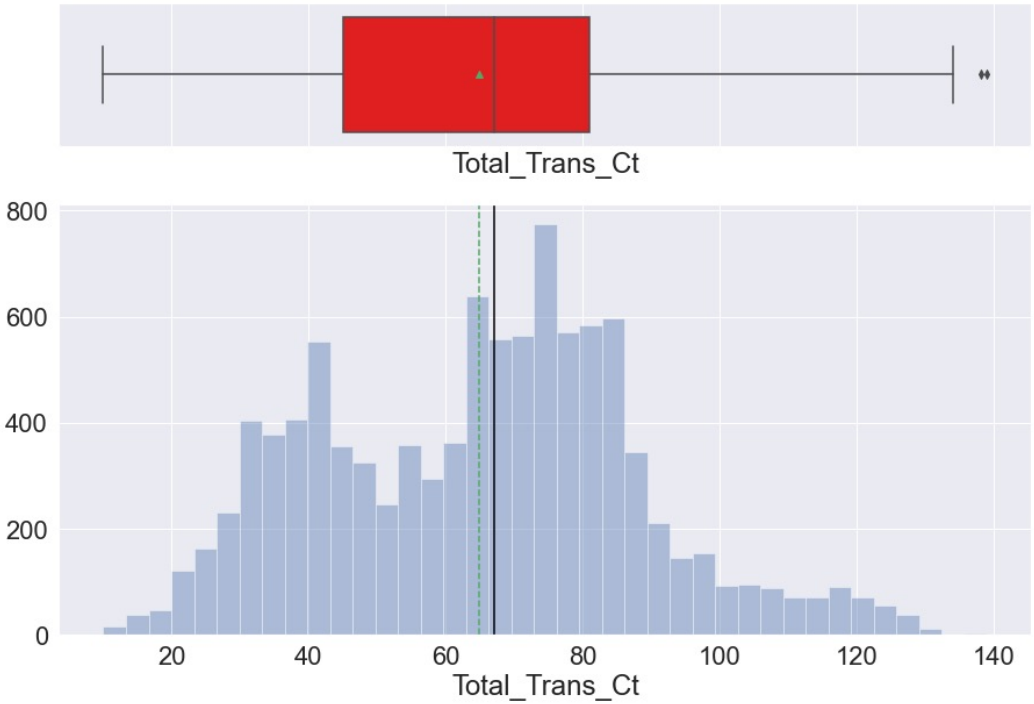
# TOTAL_TRANS_AMT

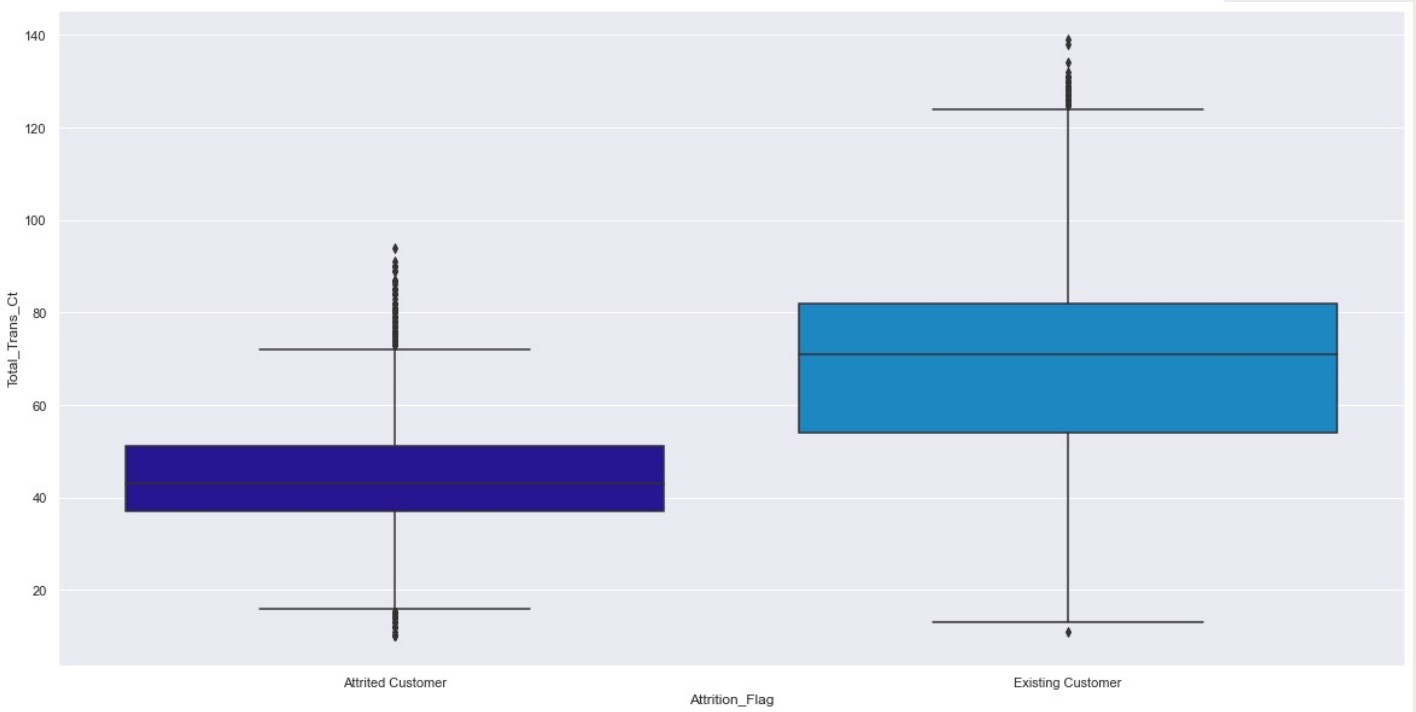



Distribution is right-skewed with outliers to the right. However, these will not be treated since some variation is expected in variables like Total Transaction Amount. Total Transaction Amount tends to be higher with existing customers.

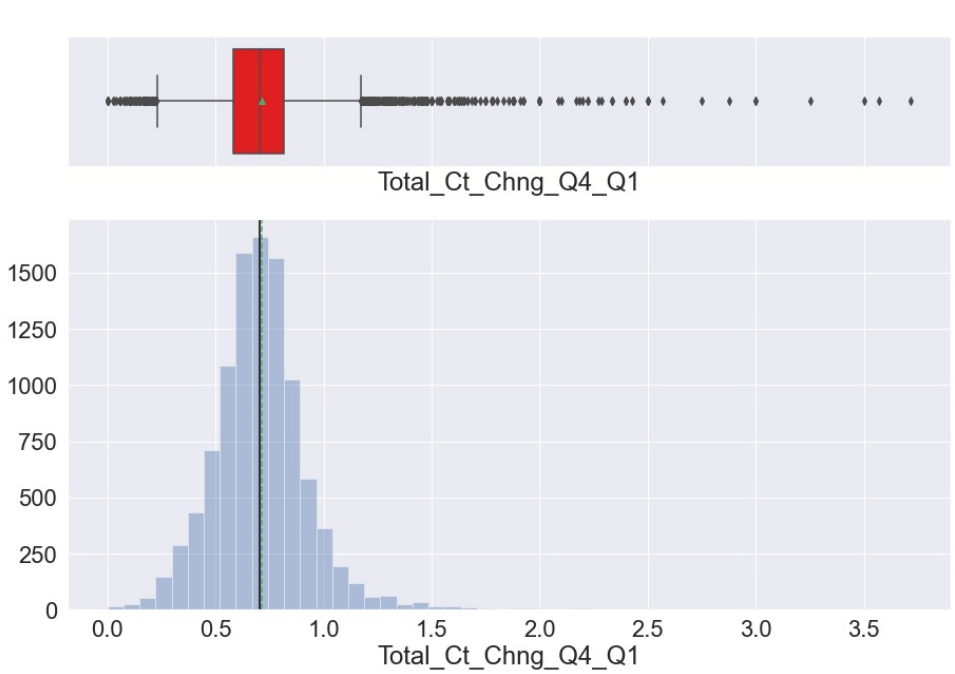
# TOTAL_TRANS_CT

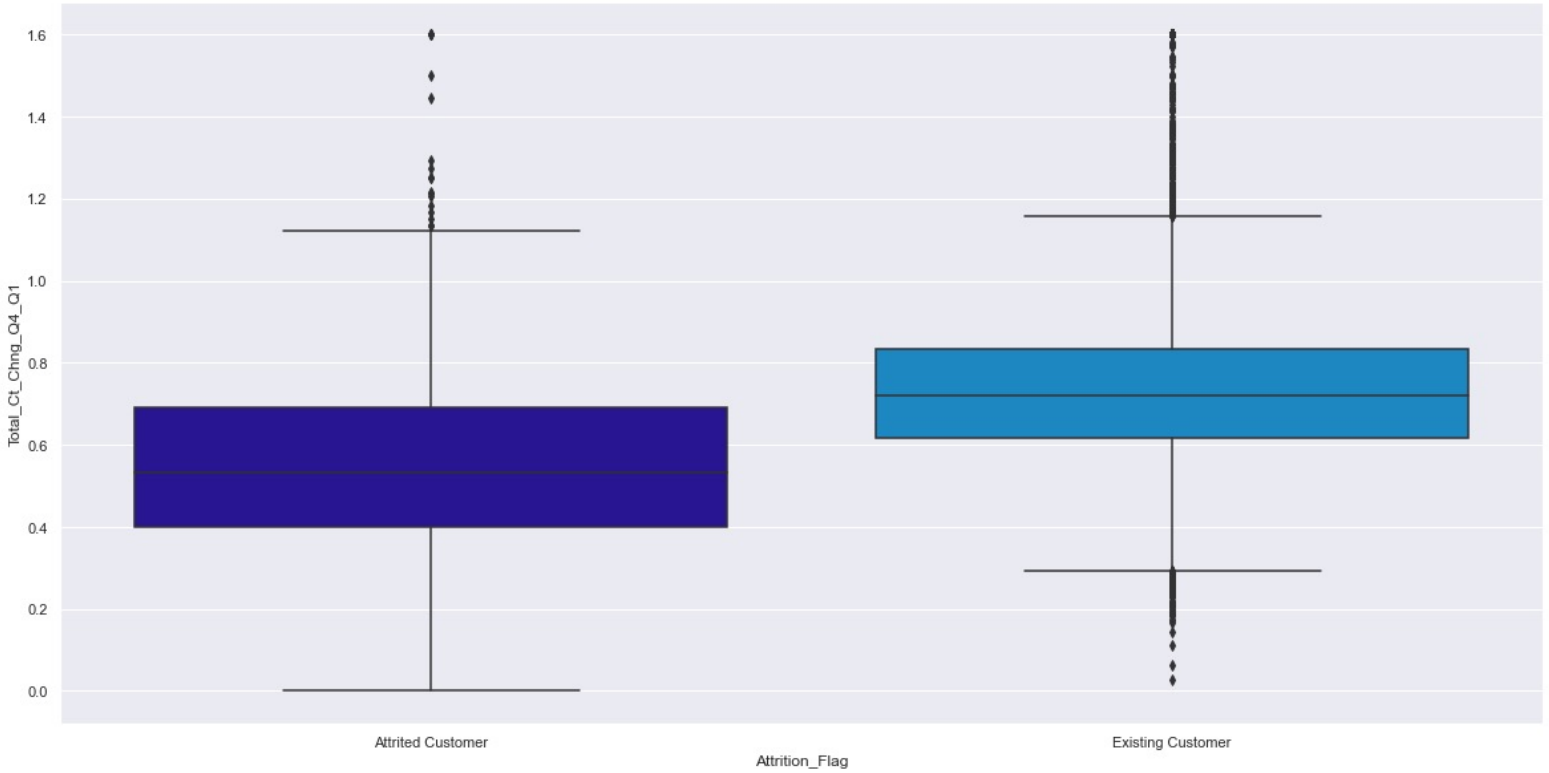



Distribution is slightly right-skewed with outliers to the right. After further investigation of data, determined that outliers were still close to other data so did not treat as outliers. Total Transaction Count tends to be higher with existing customers.

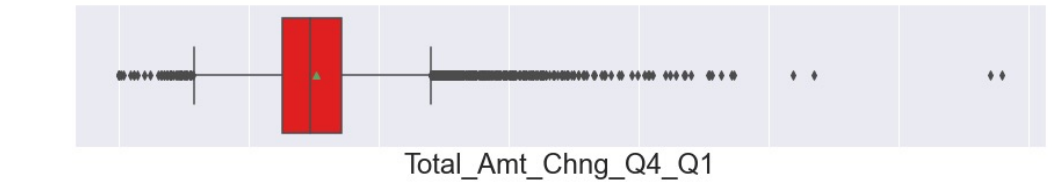
# TOTAL_CT_CHNG_Q4_Q1

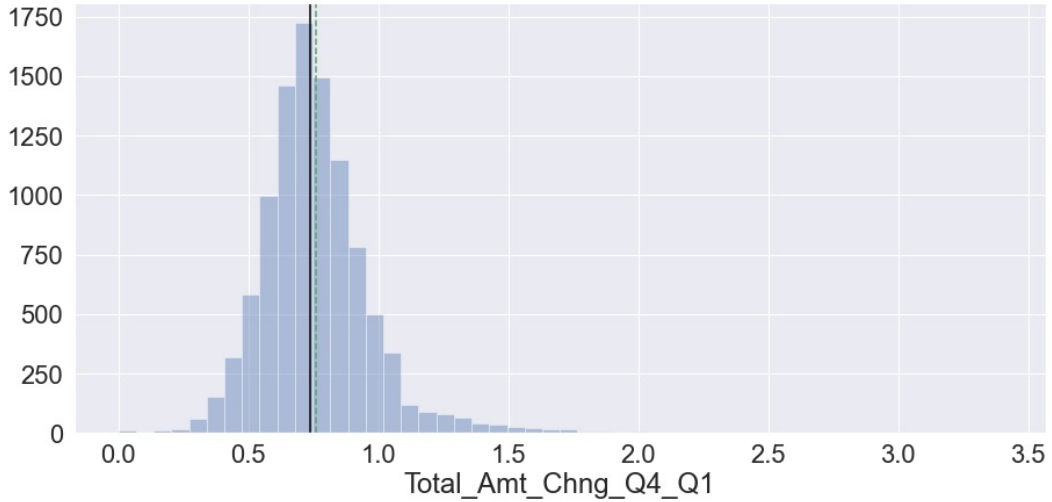


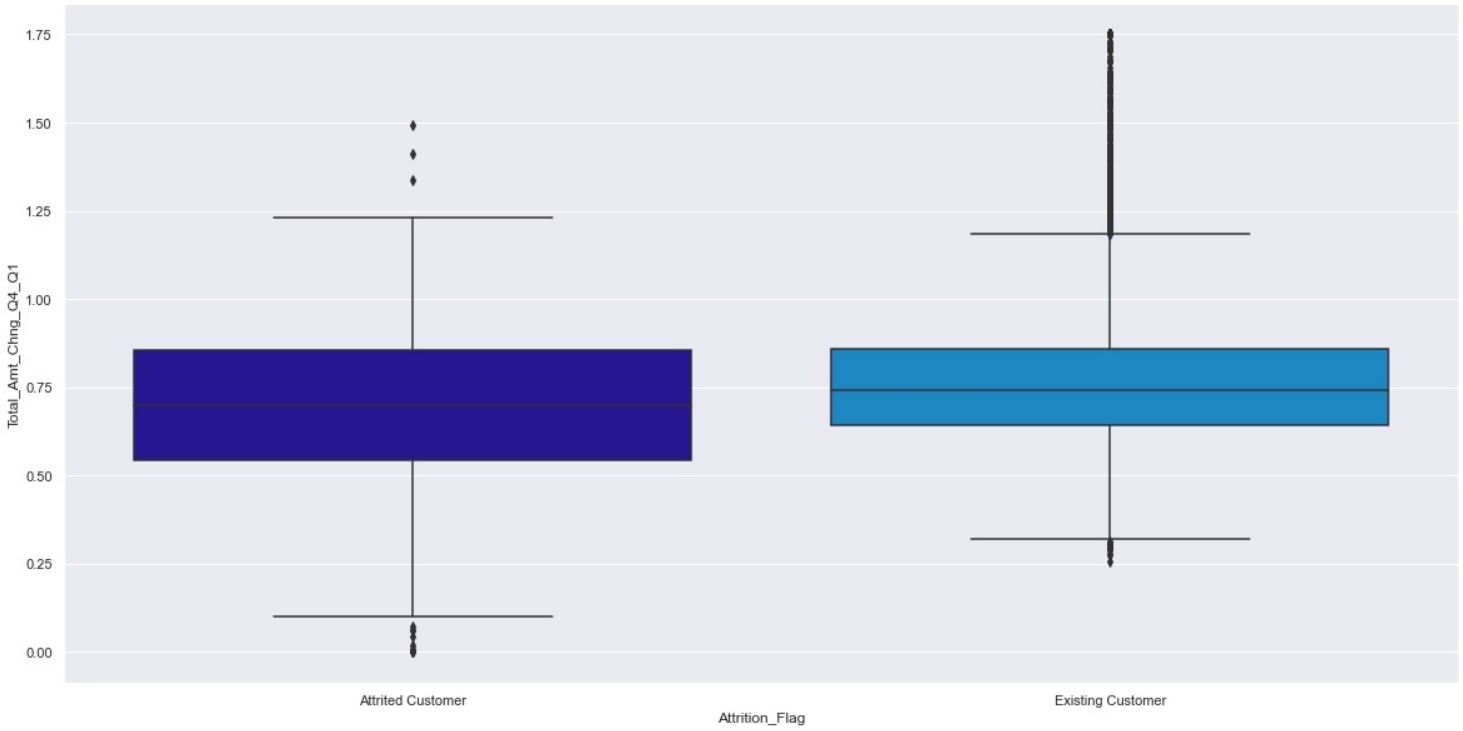




Outliers exist on both the left and right sides. Outliers on the left need not be treated since they are close to the other values. After examining the visualization together with the table of values of Total_Ct_Chng_Q4_Q1, the right outliers were treated by capping them at 1.6. Overall, the ratio tends to be lower with attrited customers.

# TOTAL_AMT_CHNG_Q4_Q1

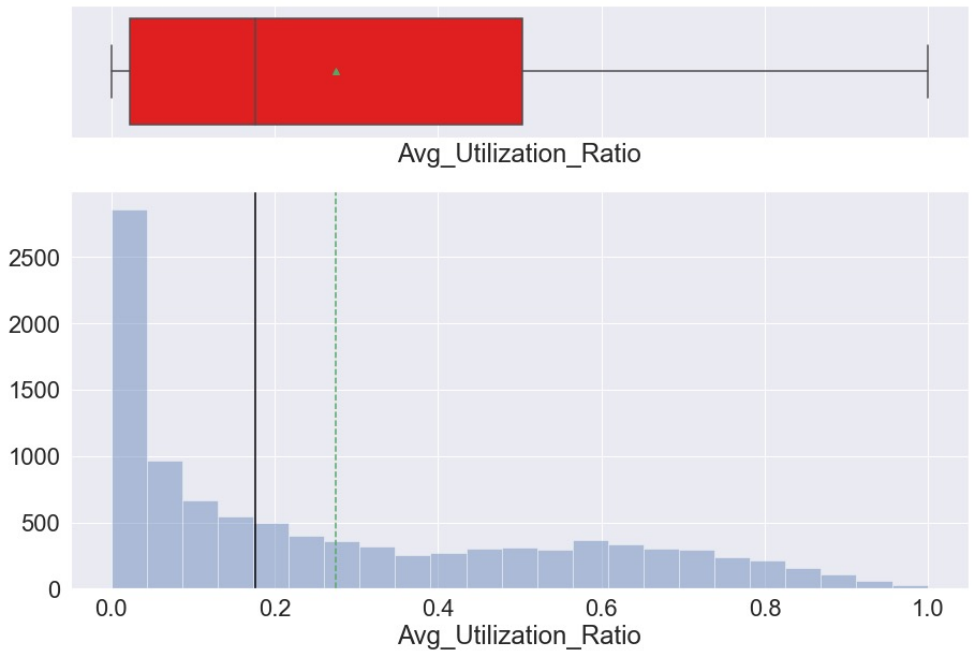


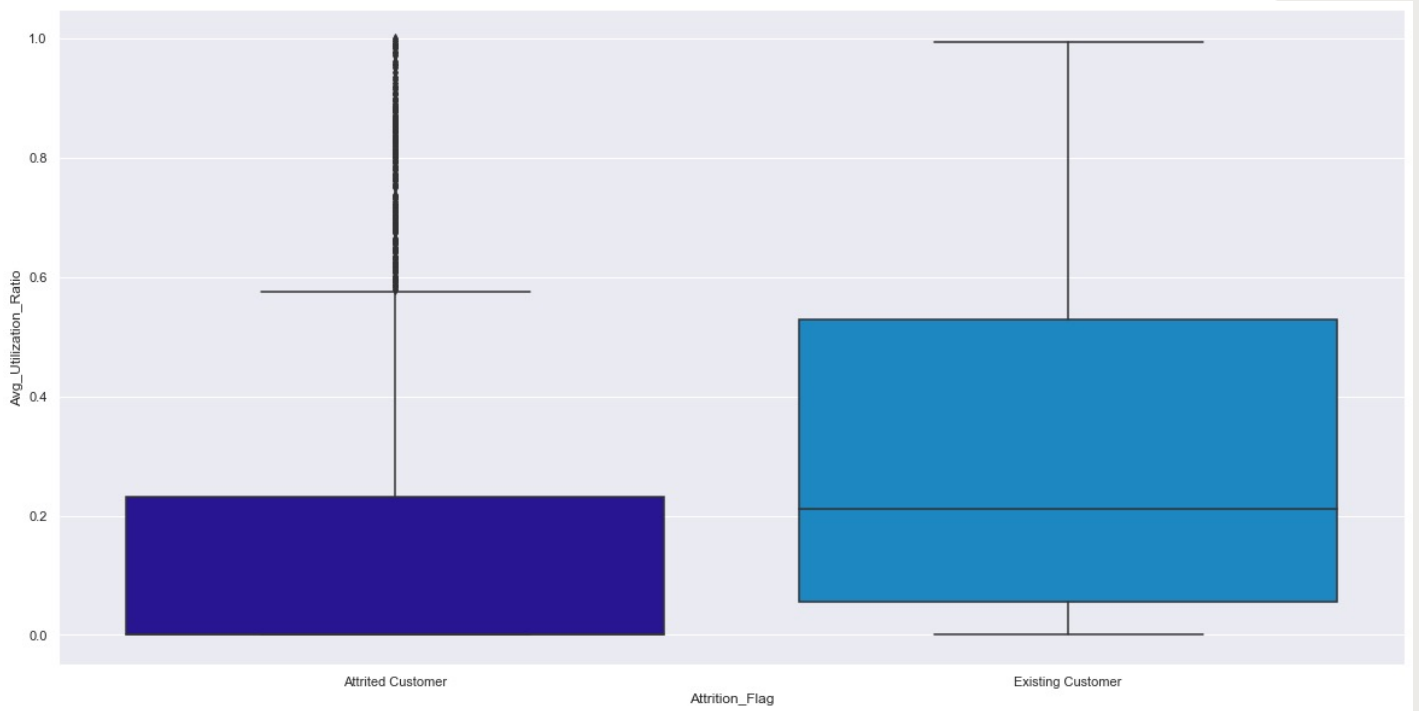




Outliers exist on both the left and right. Outliers on the left need not be treated since they are close to the other values. After examining the visualization together with the table of values of Total_Amt_Chng_Q4_Q1, the right outliers will be treated by capping them at 1.75. The ratio has a wider range (range includes lower ratios) with attrited customers.

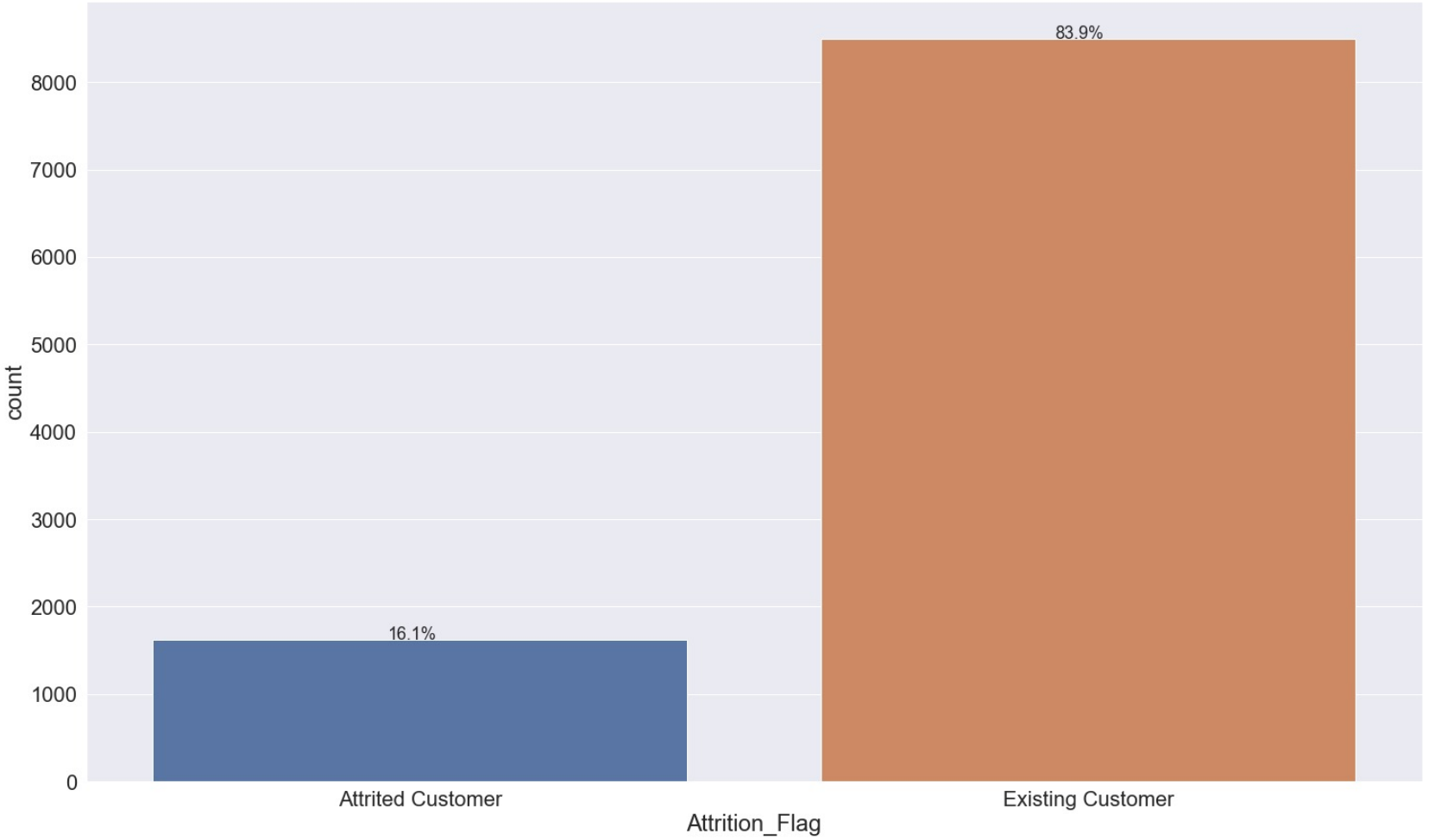
# AVG_UTILIZATION_RATIO



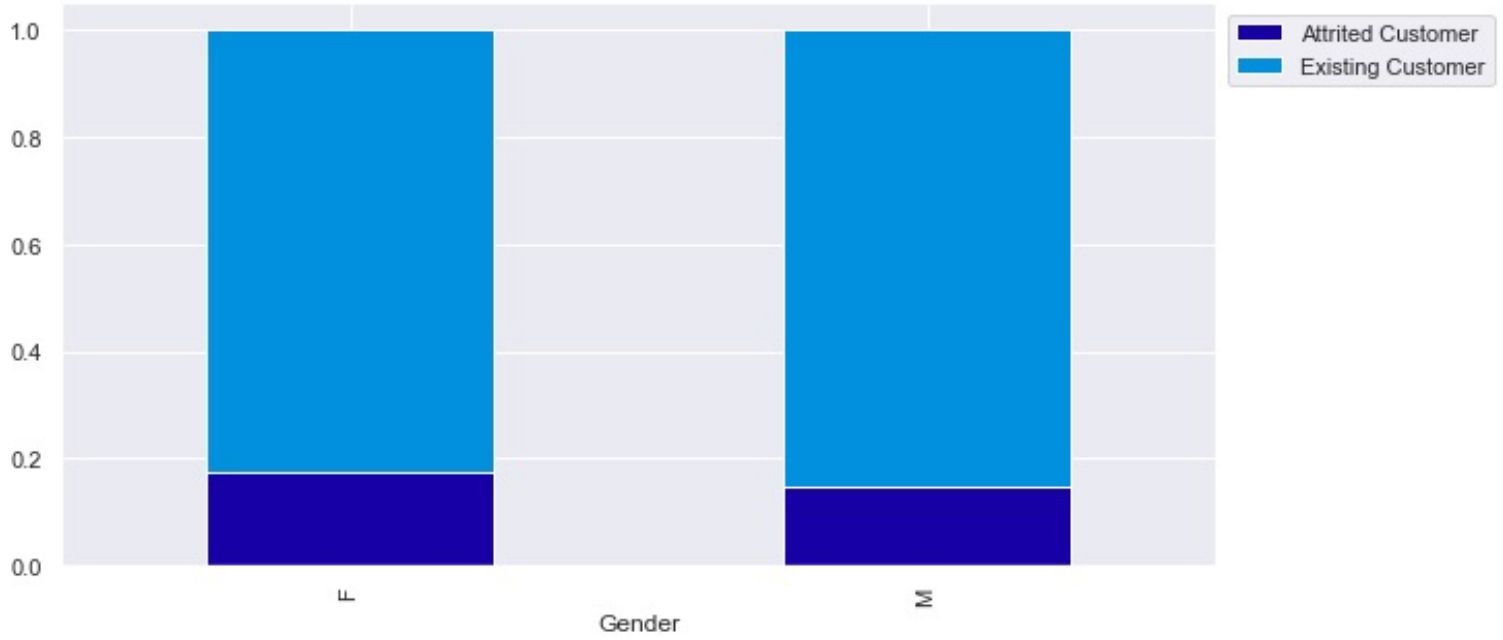

Approx. 2700 have an average utilization ratio of 0. Ratio range is wider (includes higher ratios) with existing customers.

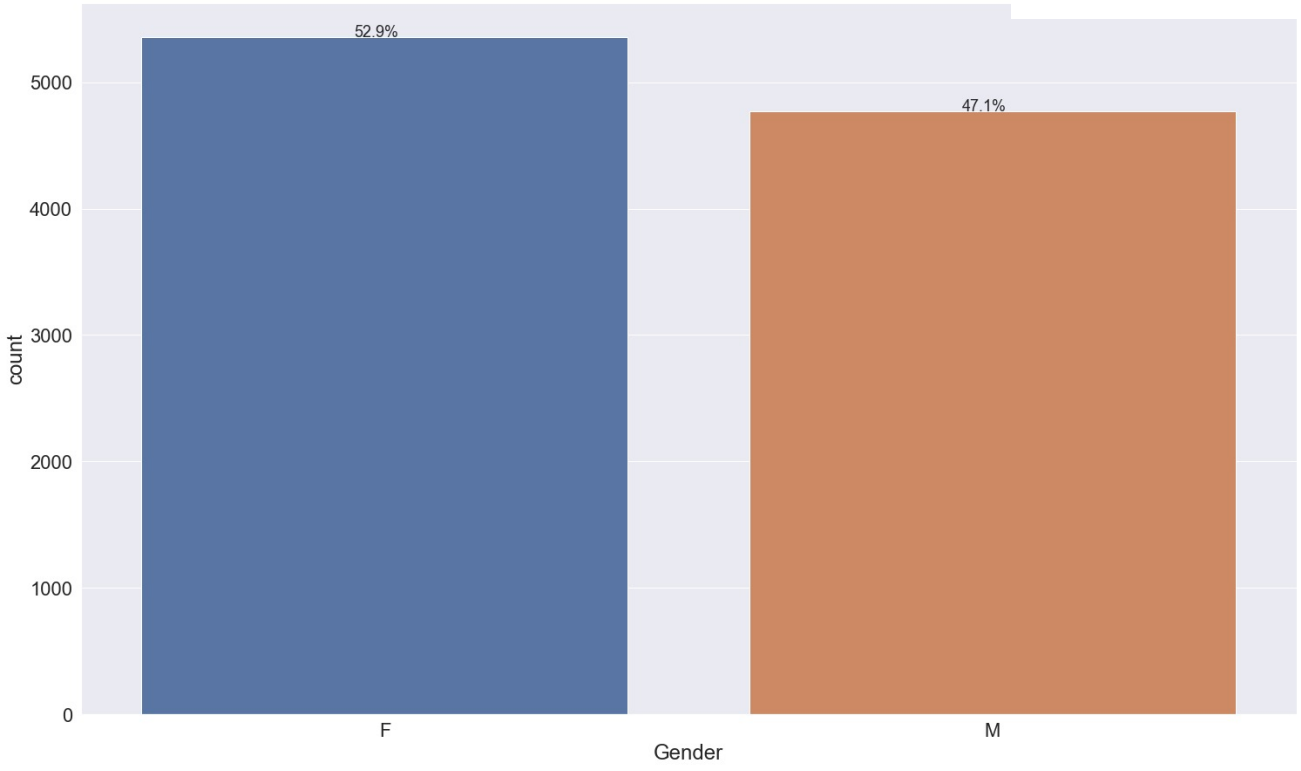
# ATTRITION_FLAG




By a large margin, most customers are current customers.

# GENDER

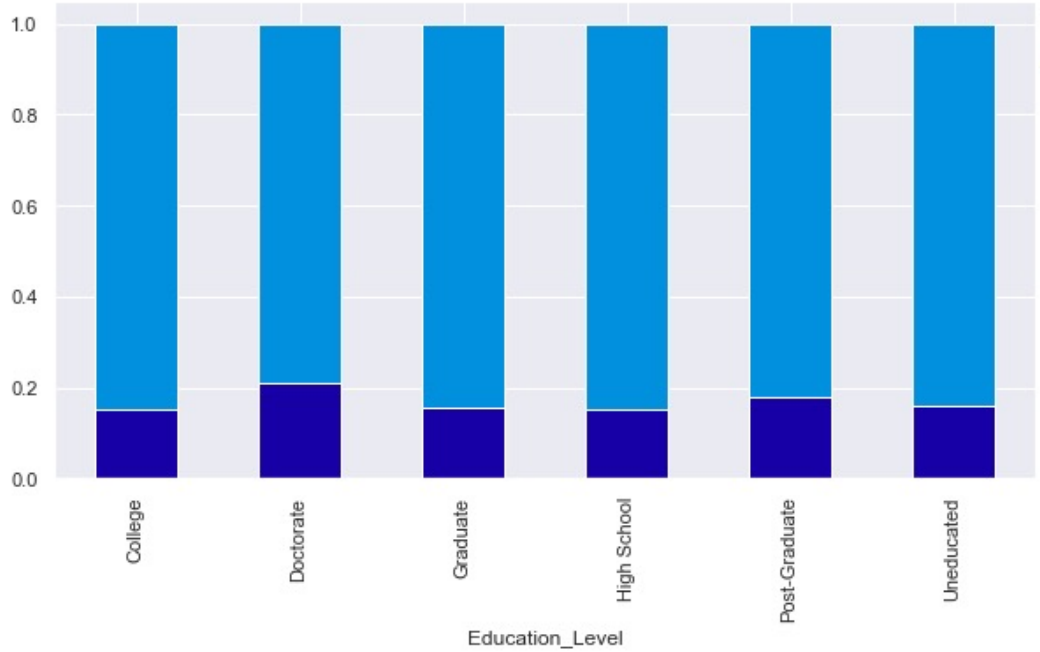


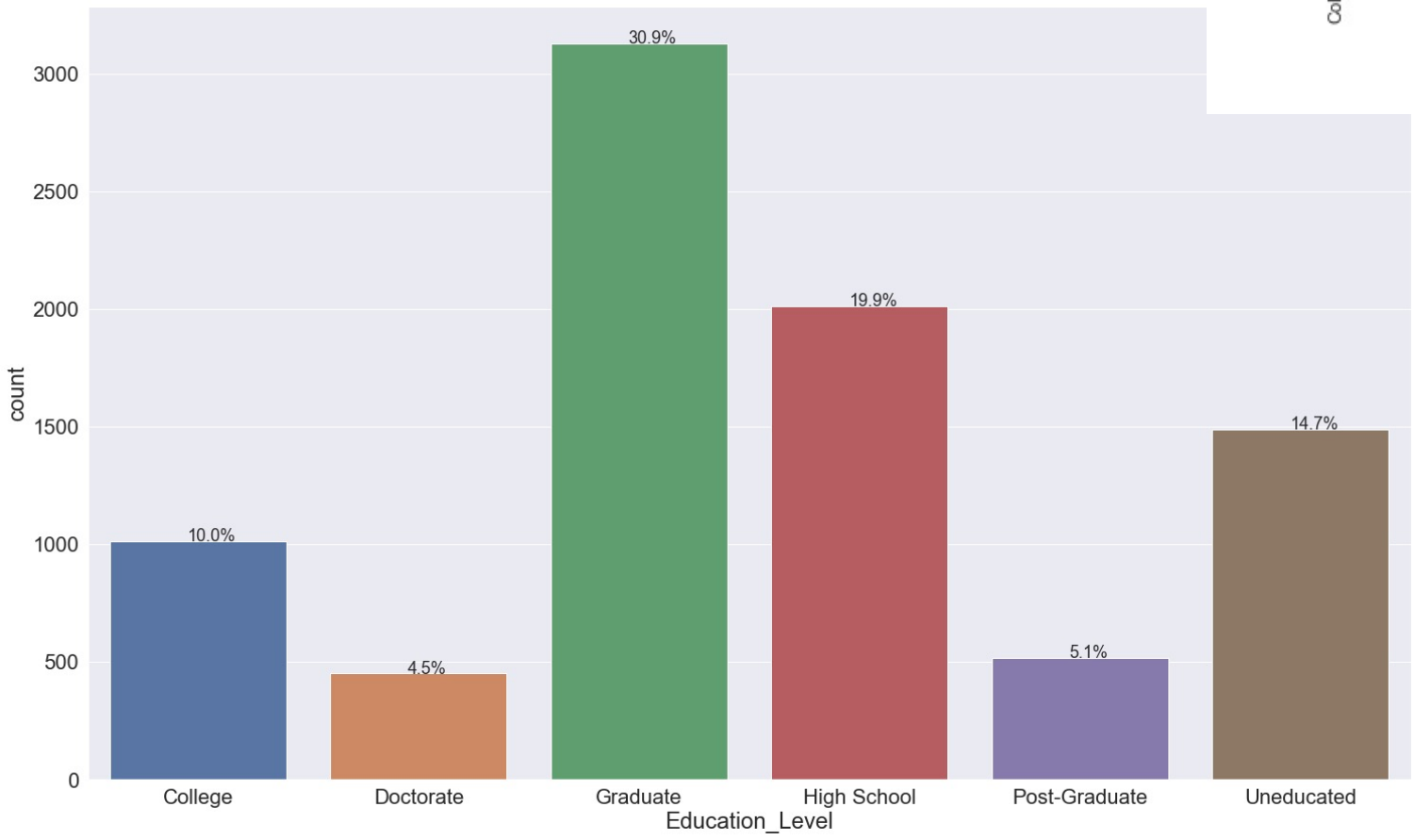




By a small margin, most customers are female.

Number of males and females does not seem to differ with attrited customers vs. existing customers.

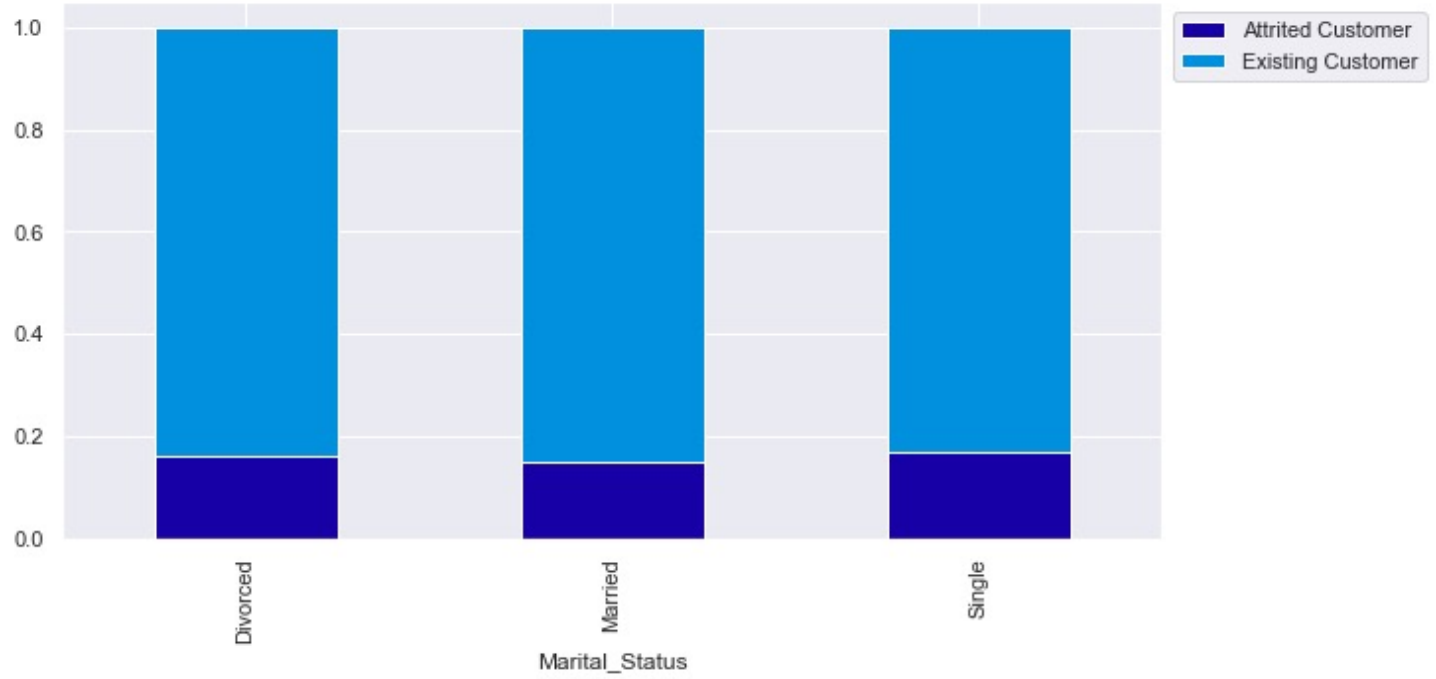
# EDUCATION_LEVEL

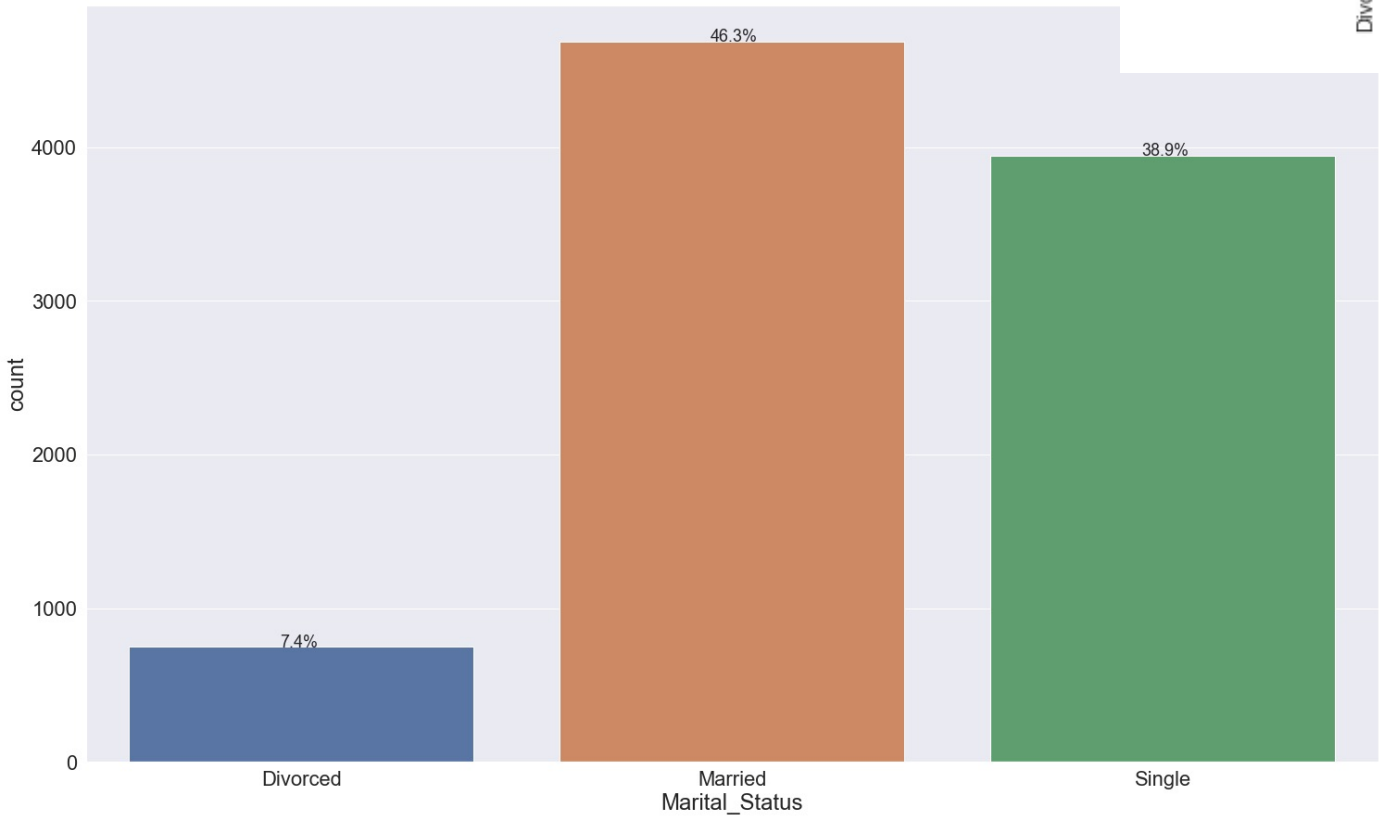



★ Most customers have a graduate degree.
★ More customers are uneducated than have a college degree.
★ Only a small percentage have doctorate degrees or post-graduate education.
★ Among the attrited customers, slightly more are doctoral and post-graduates.

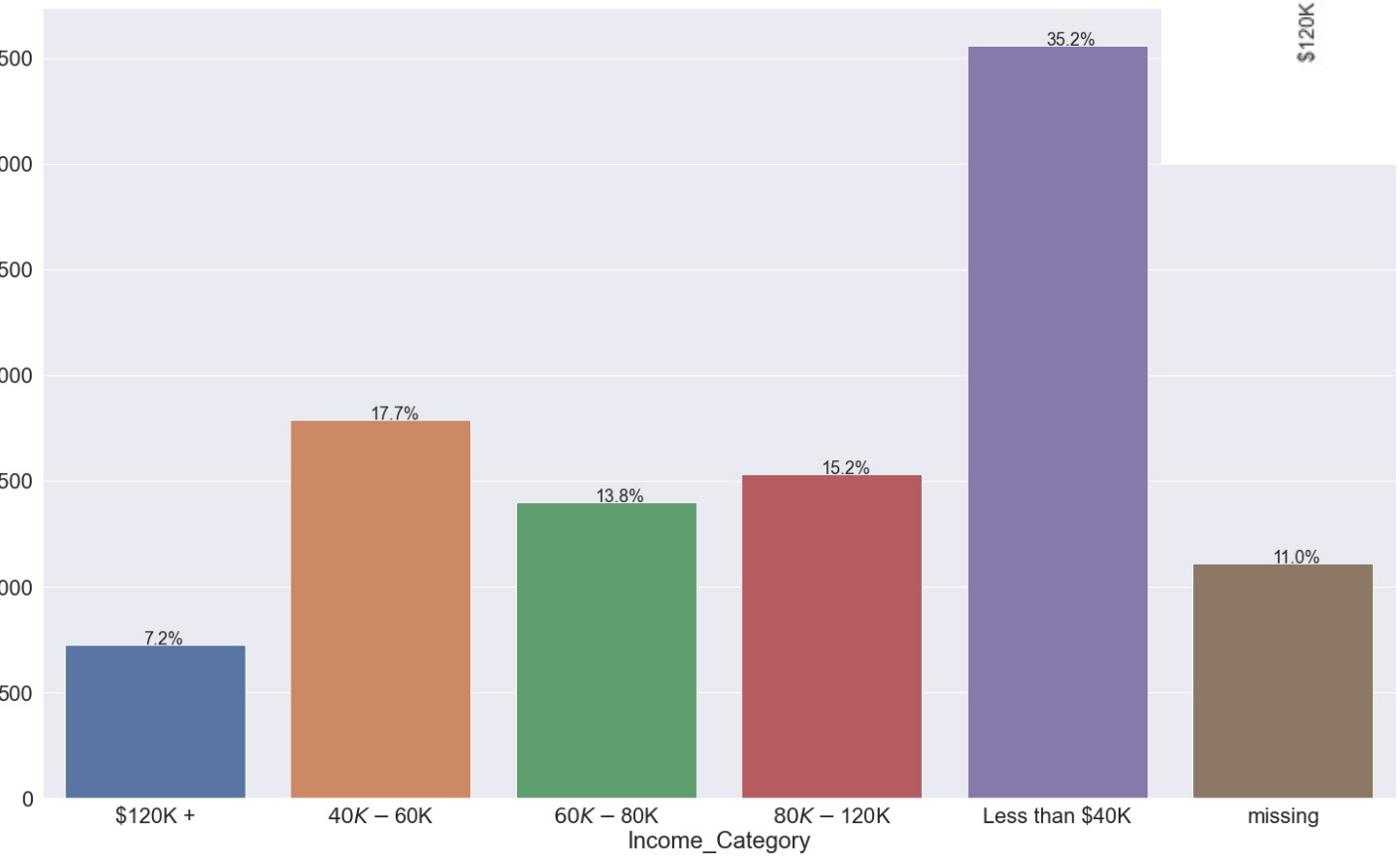
# MARITAL_STATUS

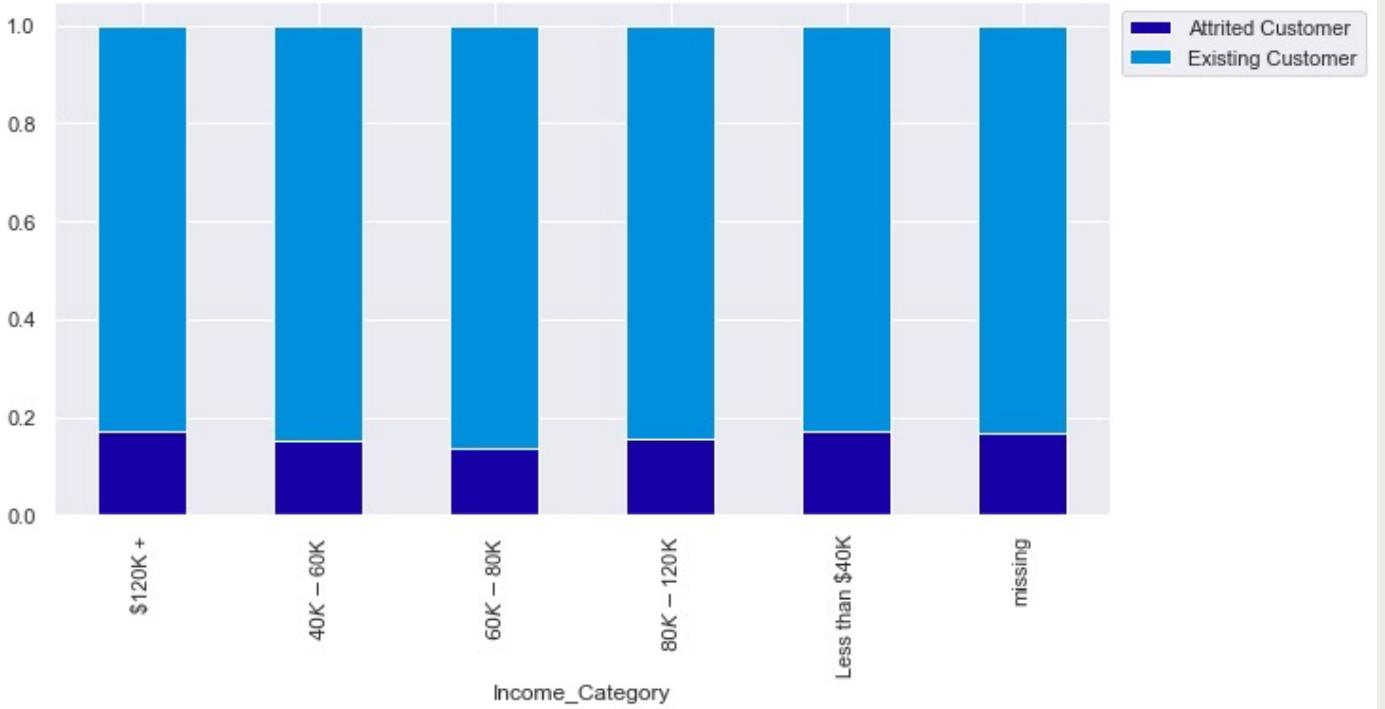



★ Most customers are married.
★ Only a small percentage of customers are divorced.
★ There seems to be no difference in the categories when it comes to being an attrited vs. an existing customer.

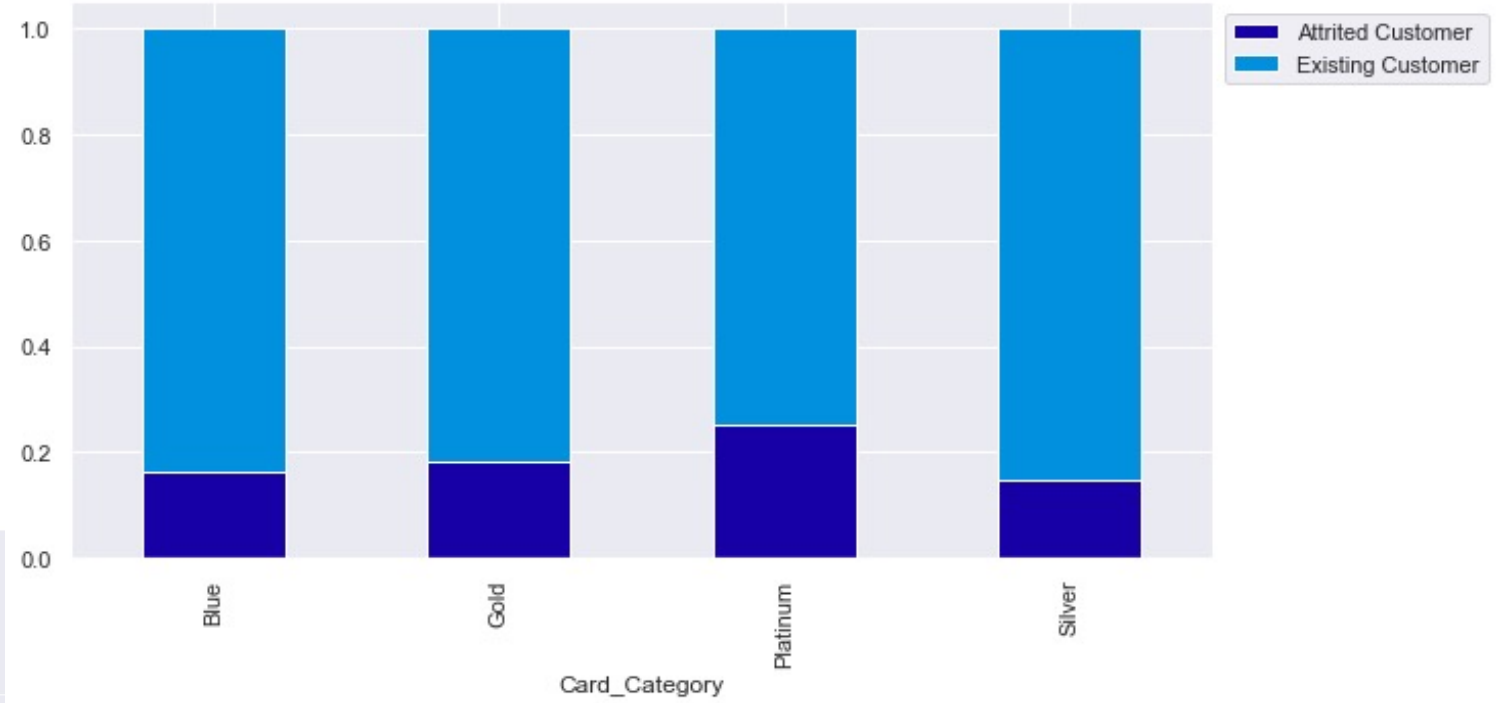
# INCOME_CATEGORY

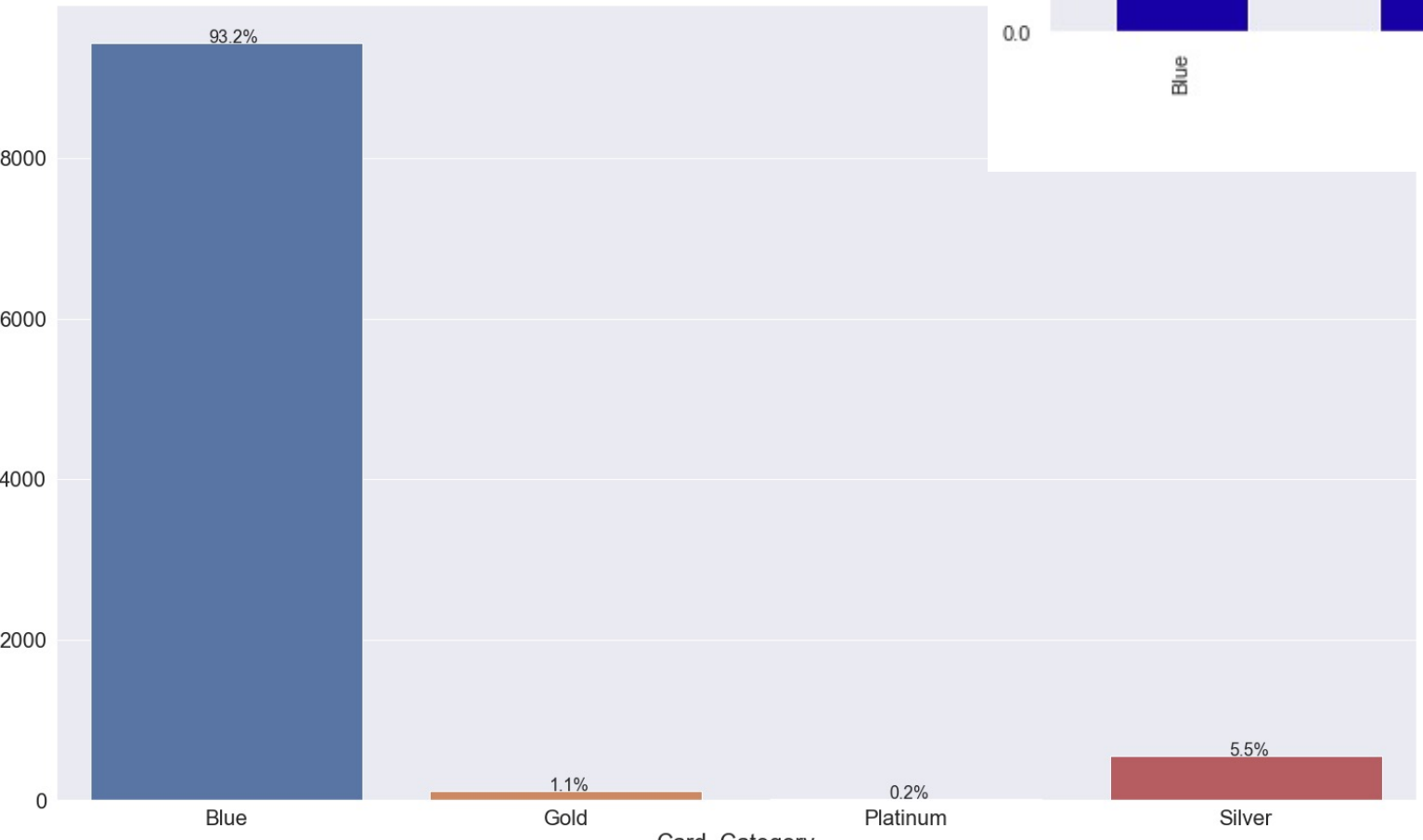



- By a large margin, most customers earn less than $40,000.
- Slightly less in the 60K-80K range tend to be attrited.

# CARD_CATEGORY




- By a significant margin, most customers are in the Blue card category.
- Slightly larger number of Platinum members tend to be attrited.

# MODEL SUMMARY

Training Performance Comparison

| | DTree Tuned | ADB Tuned | GBM Tuned | DTree Over | ADB Over | GBM Over | DTree Under | ADB Under | GBM Under |
|---|---|---|---|---|---|---|---|---|---|
| **Accuracy** | 1.000 | 0.998 | 0.986 | 1.000 | 0.962 | 0.977 | 1.000 | 0.950 | 1.000 |
| **Recall** | 1.000 | 0.991 | 0.942 | 1.000 | 0.967 | 0.981 | 1.000 | 0.953 | 1.000 |
| **Precision** | 1.000 | 0.994 | 0.969 | 1.000 | 0.958 | 0.973 | 1.000 | 0.947 | 1.000 |
| **F1** | 1.000 | 0.992 | 0.955 | 1.000 | 0.962 | 0.977 | 1.000 | 0.950 | 1.000 |

Validation Performance Comparison

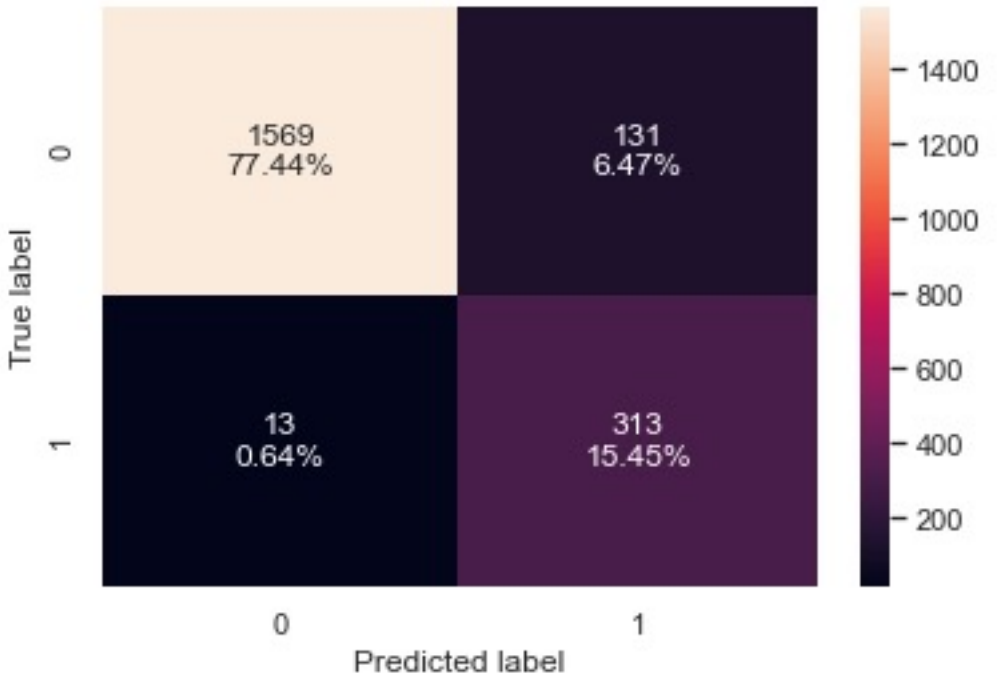
| | DTree Tuned | ADB Tuned | GBM Tuned | DTree Over | ADB Over | GBM Over | DTree Under | ADB Under | GBM Under |
|---|---|---|---|---|---|---|---|---|---|
| **Accuracy** | 0.931 | 0.966 | 0.970 | 0.932 | 0.945 | 0.957 | 0.891 | 0.929 | 0.891 |
| **Recall** | 0.776 | 0.865 | 0.877 | 0.847 | 0.883 | 0.893 | 0.896 | 0.960 | 0.896 |
| **Precision** | 0.791 | 0.919 | 0.935 | 0.758 | 0.796 | 0.846 | 0.611 | 0.705 | 0.611 |
| **F1** | 0.783 | 0.891 | 0.905 | 0.800 | 0.837 | 0.869 | 0.726 | 0.813 | 0.726 |

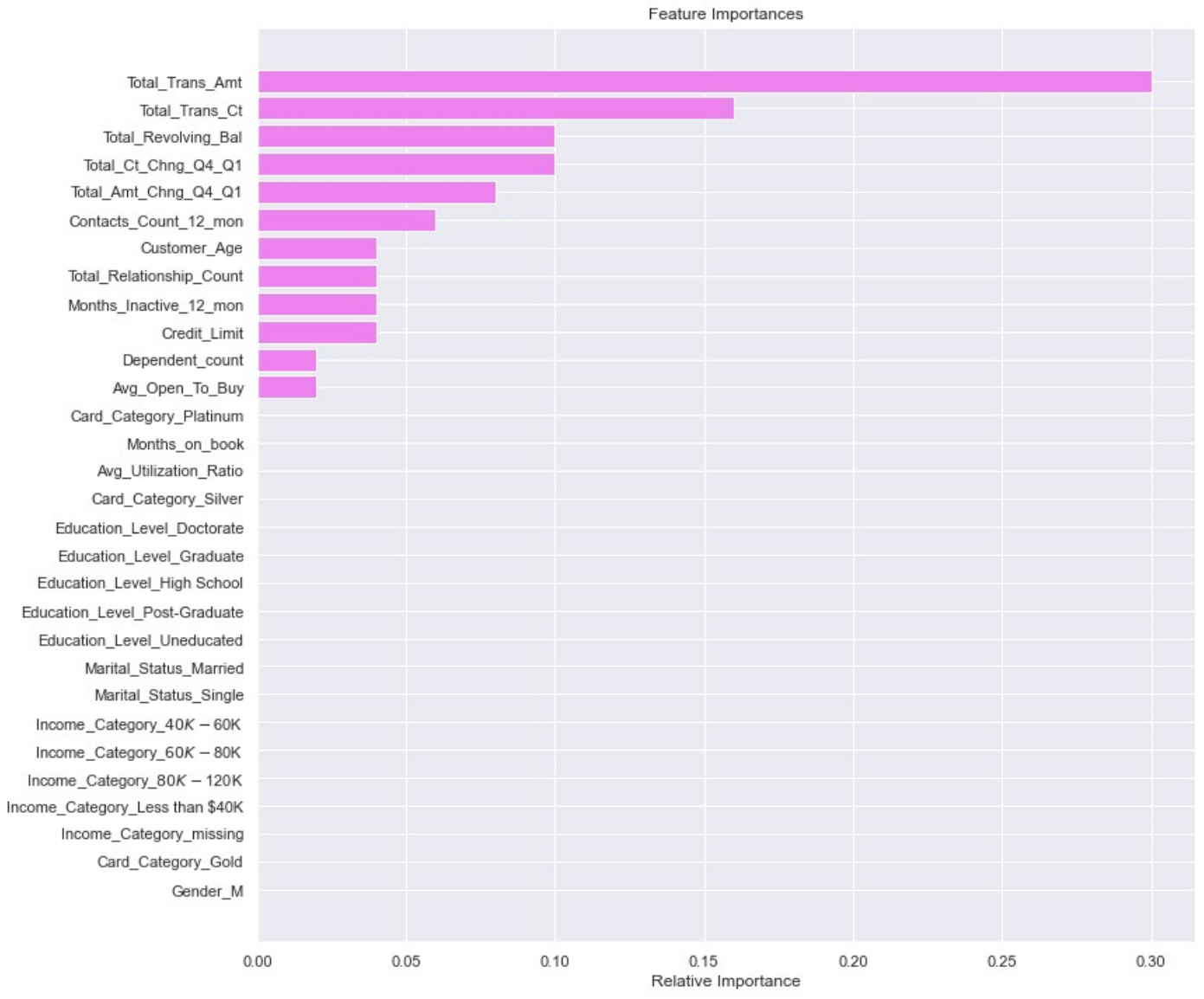
Test performance:

| | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| **0** | 0.928 | 0.960 | 0.701 | 0.810 |

- AdaBoost on the Undersampled Data gives the highest recall score on the validation set. In addition, it scores similarly on the training set which seems to point to less of an overfitting problem. Therefore, this is deemed the best model.

# ADABOOST MODEL ON UNDERSAMPLED DATA



# CONCLUSION

- Total Transaction Amount is the most important variable in terms of determining if a customer is attrited. This value tends to be higher for existing customers. Thera Bank in an effort to increase balances could offer programs for credit card holders where they earn points for dollars spent, incentivizing spending with the credit card. Similarly, they could create a balance transfer offer (0% or low percent interest rate on balance tranfers for an allotted period of time).
- Total_Trans_Ct or Total Transaction Count is also an important variable in determining if a customer is attrited. This number tends to be higher with existing customers. Thera Bank could investigate offering deals at certain merchants if their card(s) are used (e.g., discount when using your Thera Bank credit or debit card) incentivizing the increase in transactions.
- Most attrited customers make less than $40,000. Thera Bank could target this population by offering lower interest rates on cards with smaller credit limits, or cards with lower fee schedules.
- Attrited customers tend to have had less products with the bank. Thera Bank could research offering new membership plans or accounts to people with only 1 card (e.g., no fees or lower interest rates if you have both a debit and credit card with the bank).