Credit Card Customers EDA

Scott Shriver

Introduction

For this project I will be performing an EDA on a dataset of credit card customers. The goals of this EDA will be to determine more about the various credit card customers held by this organization, as well as to try and determine factors that can lead to a customer cancelling their card, referred to as churning within the dataset. For the purposes of this project, I will refer to a customer ceasing use of their credit card as churn, cancellation, and attrition interchangeably.

This report will consist of an explanation of the problem this EDA is trying to solve, a description of the dataset, a walkthrough of the data analysis and some results, and will end with a conclusion of the analysis and thoughts on potential improvements.

Problem Statement

One of the greatest problems that a company such as a bank can have is to lose their customer base, or for their customers to infrequently use the services that the company provides. For this project, I will be looking at trying to identify signs that a customer may be getting ready to cancel their credit card service. In addition, I will also try and find signs that a customer's usage of their credit card will be high or low.

Dataset

The dataset *Credit Card Customers* was found on Kaggle at the following link:

https://www.kaggle.com/sakshigoyal7/credit-card-customers

It is an adaptation of a dataset found from the following page with some slight modifications:

https://leaps.analyttica.com/sample_cases/11

This dataset contains records on over 10,000 credit card customers at an unnamed bank, along with a column explaining if this customer has closed their account or not. A description of each main column from the official documentation can be found in the following table:

Variable	Type	Description
Clientnum	Num	Client number. Unique identifier for the customer holding the account
Attrition_Flag	char	Internal event (customer activity) variable - if the account is closed then 1 else 0
Customer_Age	Num	Demographic variable - Customer's Age in Years
Gender	Char	Demographic variable - M=Male, F=Female
Dependent_count	Num	Demographic variable - Number of dependents
Education_Level	Char	Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.)
Marital_Status	Char	Demographic variable - Married, Single, Unknown
Income_Category	Char	Demographic variable - Annual Income Category of the account holder (< $40K, $40K - 60K, $60K - $80K, $80K-$120K, > $120K, Unknown)
Card_Category	Char	Product Variable - Type of Card (Blue, Silver, Gold, Platinum)

Months_on_book	Num	Months on book (Time of Relationship)
Total_Relationship_Count	Num	Total no. of products held by the customer
Months_Inactive_12_mon	Num	No. of months inactive in the last 12 months
Contacts_Count_12_mon	Num	No. of Contacts in the last 12 months
Credit_Limit	Num	Credit Limit on the Credit Card
Total_Revolving_Bal	Num	Total Revolving Balance on the Credit Card
Avg_Open_To_Buy	Num	Open to Buy Credit Line (Average of last 12 months)
Total_Amt_Chng_Q4_Q1	Num	Change in Transaction Amount (Q4 over Q1)
Total_Trans_Amt	Num	Total Transaction Amount (Last 12 months)
Total_Trans_Ct	Num	Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1	Num	Change in Transaction Count (Q4 over Q1)
Avg_Utilization_Ratio	Num	Average Card Utilization Ratio

The Kaggle version of the dataset appends two columns to the very end of the dataset based off the uploader's attempt at performing data analysis and will be excluded during the EDA.

The Kaggle page for the dataset points out that only about 16% of customers have closed their account, which will increase the difficulty of trying to identify trends in customer churn.

I chose to work on this dataset because of the large number of rows and columns available to analyze, the fact that it provides data on financial service information, and the previously mentioned challenge that this dataset was said to provide.

Experiment

In this section I will be providing an overview of the EDA process done on this dataset. The full results can be viewed through the Jupyter Notebook included in this submission.

I first start by reading in the dataset. Doing some initial data inspection such as checking the head and shape of the data help confirm that it was imported properly.

```
ccData.head()
```

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book
0	768805383	Existing Customer	45	M	3	High School	Married	$60K – 80K	Blue	39
1	818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	44
2	713982108	Existing Customer	51	M	3	Graduate	Married	$80K – 120K	Blue	36
3	769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	34
4	709106358	Existing Customer	40	M	3	Uneducated	Married	$60K – 80K	Blue	21

5 rows × 21 columns

The *Income Category, Education Level, and Card Category* of the dataset are all values that I am interested in using later in the EDA. In their initial state they are strings, so I mapped them to have a numerical value.

```
def map_data(df):
    #income category
    income_map = {'Less than $40K': 0, '$40K - $60K': 1, '$60K - $80K': 2, '$80K - $120K': 3, '$120K +': 4, 'Unknown': 5}
    df['Income_Category'] = df['Income_Category'].map(income_map)

    #education level
    education_map = {'Uneducated': 0, 'High School': 1, 'College': 2, 'Graduate': 3, 'Post-Graduate': 4, 'Doctorate': 5, 'Unknown
    df['Education_Level'] = df['Education_Level'].map(education_map)

    #card category
    card_map = {'Blue': 0, 'Silver': 1, 'Gold': 2, 'Platinum': 3}
    df['Card_Category'] = df['Card_Category'].map(card_map)

    return df

ccData = map_data(ccData)
ccData.head()
```

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book
0	768805383	Existing Customer	45	M	3	1	Married	2	0	39
1	818770008	Existing Customer	49	F	5	3	Single	0	0	44
2	713982108	Existing Customer	51	M	3	3	Married	3	0	36
3	769911858	Existing Customer	40	F	4	1	Unknown	0	0	34
4	709106358	Existing Customer	40	M	3	0	Married	2	0	21

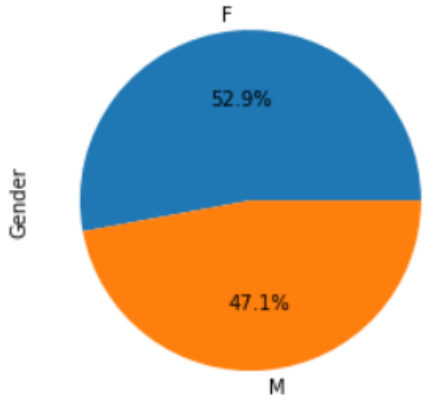
5 rows × 21 columns

Additionally, I decided to bin the *Credit Grouping and Average Utilization Grouping* values to make the easier to use for analysis.

We can see some initial visualizations below.

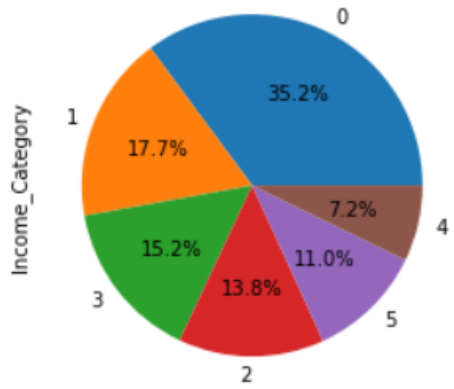
```
ccData['Gender'].value_counts().plot(kind = 'pie', autopct = '%1.1f%%')
```

```
<AxesSubplot:ylabel='Gender'>
```




```
ccData['Income_Category'].value_counts().plot(kind = 'pie', autopct = '%1.1f%%')
```

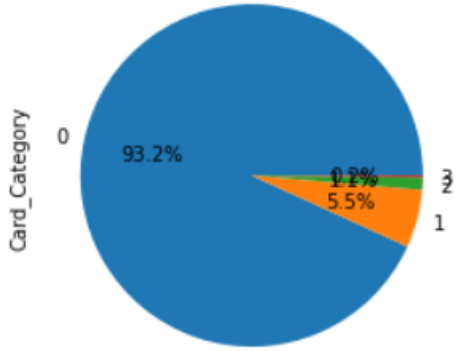
```
<AxesSubplot:ylabel='Income_Category'>
```



Slightly over half of all customers are female, and most customers are making less than $40,000 in yearly income, with a somewhat even distribution across all other income categories.

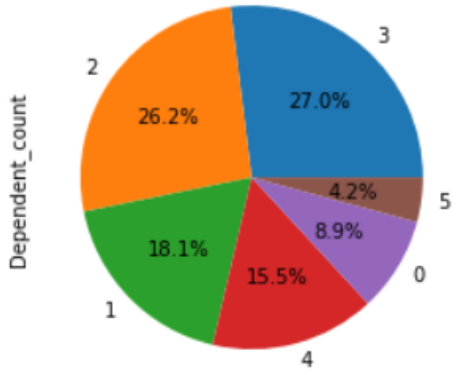
```
ccData['Card_Category'].value_counts().plot(kind = 'pie', autopct = '%1.1f%%')
```

```
<AxesSubplot:ylabel='Card_Category'>
```




```
ccData['Dependent_count'].value_counts().plot(kind = 'pie', autopct = '%1.1f%%')
```

```
<AxesSubplot:ylabel='Dependent_count'>
```

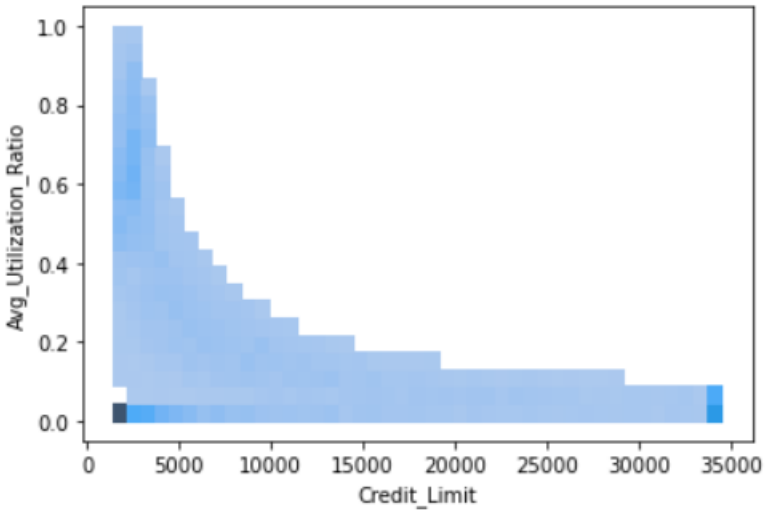


The overwhelming majority of card holders have the Blue, or default, card type, with extremely small amounts of customers having any of the premium variations. Additionally, most customers have between one to three dependents.

A key observation I found, which I had expected the opposite of, was that as a customer's credit limit increases their card utilization decreases.

```
sns.histplot(data = ccData, x = 'Credit_Limit', y = 'Avg_Utilization_Ratio')
```

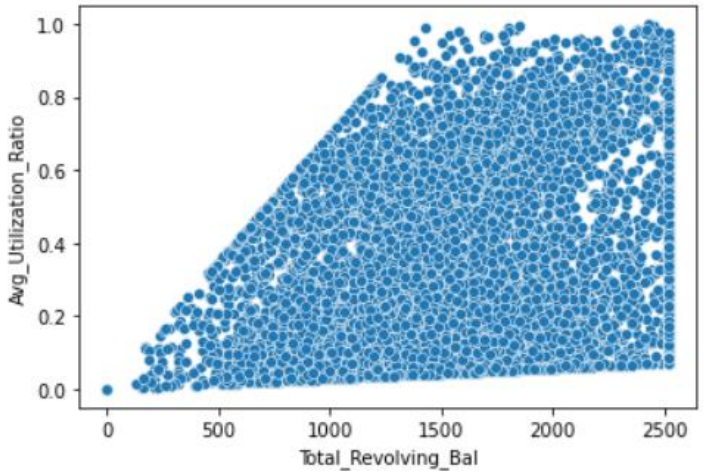
```
<AxesSubplot:xlabel='Credit_Limit', ylabel='Avg_Utilization_Ratio'>
```



In addition to this, card utilization increases with higher revolving balances for customers, but the degree to which it increases varies.

```
sns.scatterplot(data = ccData, x = 'Total_Revolving_Bal', y = 'Avg_Utilization_Ratio')
```

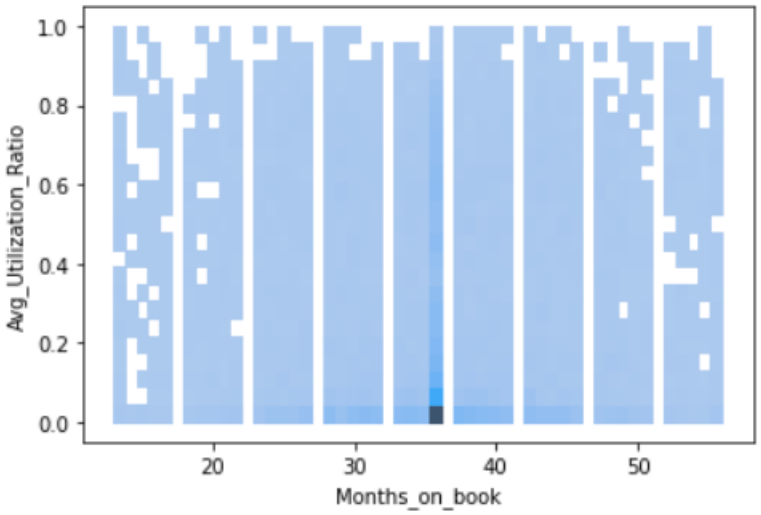
```
<AxesSubplot:xlabel='Total_Revolving_Bal', ylabel='Avg_Utilization_Ratio'>
```



The number of months a customer has held their credit card doesn't seem to have a major impact on their card utilization.

```
sns.histplot(data = ccData, x = 'Months_on_book', y = 'Avg_Utilization_Ratio')
```

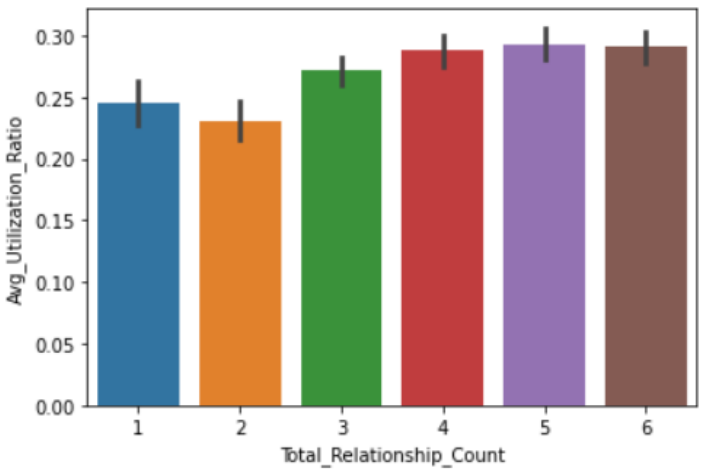
```
<AxesSubplot:xlabel='Months_on_book', ylabel='Avg_Utilization_Ratio'>
```



When it comes to the amount of product and services a customer has through this bank, card usage is on average lowest for people with two, and average usage maxes out at four.

```
sns.barplot(data = ccData, x = 'Total_Relationship_Count', y = 'Avg_Utilization_Ratio')
```

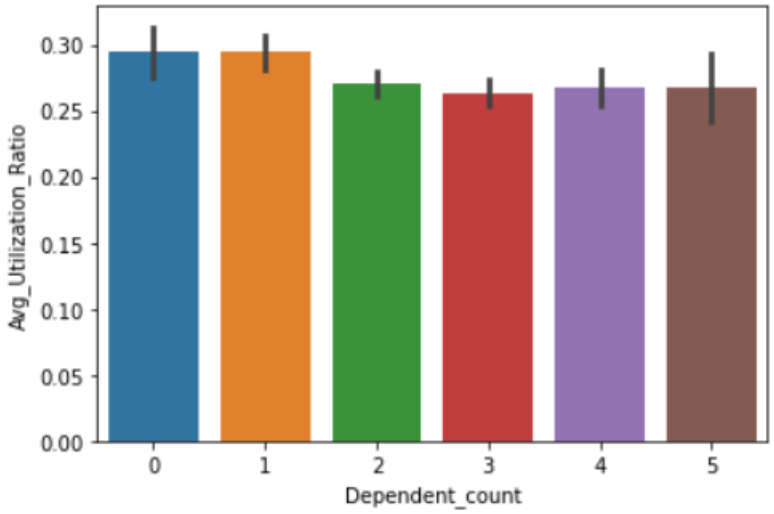
```
<AxesSubplot:xlabel='Total_Relationship_Count', ylabel='Avg_Utilization_Ratio'>
```



Customers with more than one dependent have a lower average utilization than customers with one or less dependents.

```
sns.barplot(data = ccData, x = 'Dependent_count', y = 'Avg_Utilization_Ratio')
```

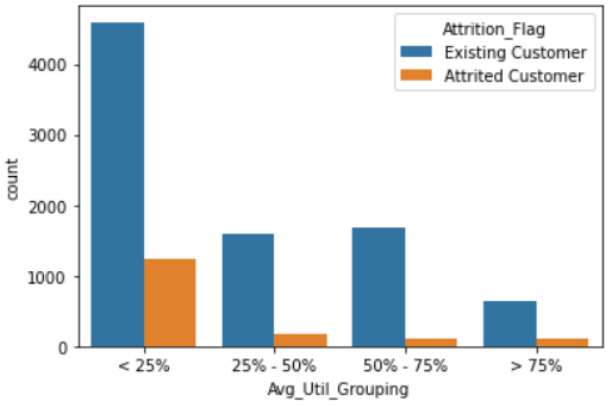
```
<AxesSubplot:xlabel='Dependent_count', ylabel='Avg_Utilization_Ratio'>
```



When it comes to customer churn, there is less churn per utilization group, but there are also very few customers in high utilization groups which makes conclusions difficult.

```
sns.countplot(data = ccData, x = 'Avg_Util_Grouping', hue = 'Attrition_Flag')
```

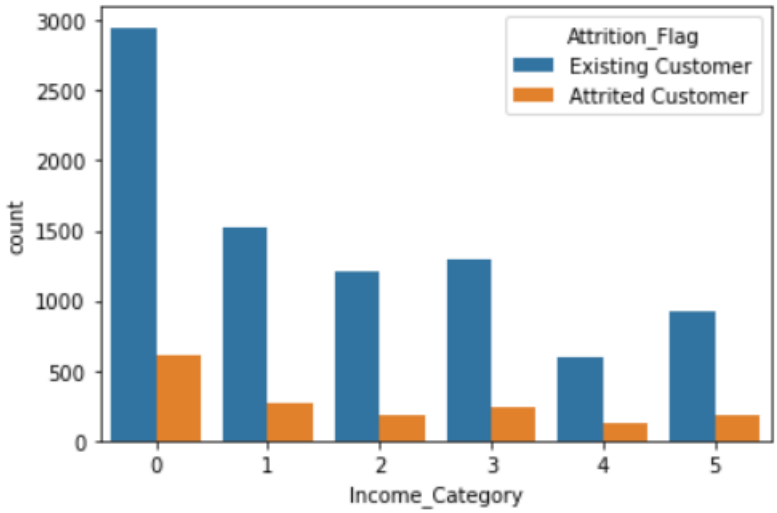
```
<AxesSubplot:xlabel='Avg_Util_Grouping', ylabel='count'>
```



This churn rate also decreases with customers who have higher income levels.

```
sns.countplot(data = ccData, x = 'Income_Category', hue = 'Attrition_Flag')
```

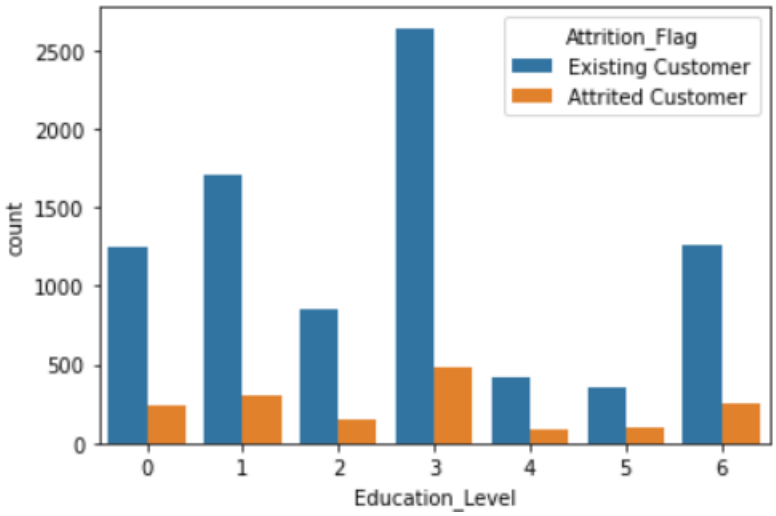
```
<AxesSubplot:xlabel='Income_Category', ylabel='count'>
```



There is a very high degree of variability for customer churn in relation to the education level of that customer.

```
sns.countplot(data = ccData, x = 'Education_Level', hue = 'Attrition_Flag')
```

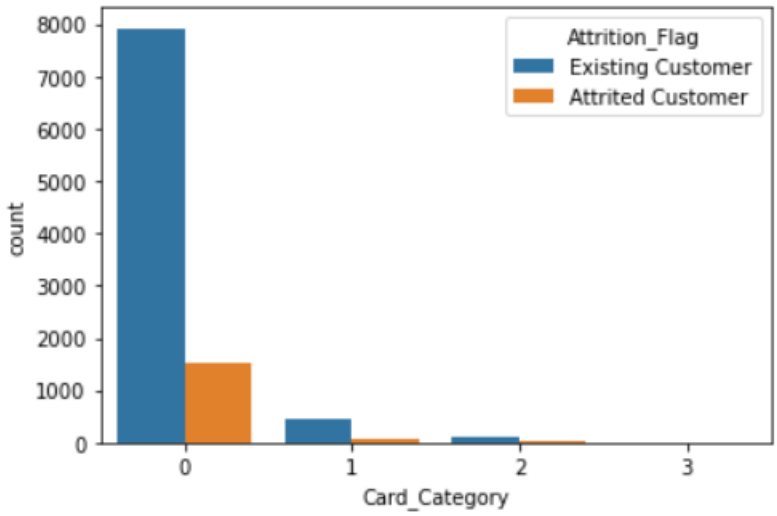
```
<AxesSubplot:xlabel='Education_Level', ylabel='count'>
```



Because of the low number of customers with premium cards, very little can be concluded about customer churn from their card category.

```
sns.countplot(data = ccData, x = 'Card_Category', hue = 'Attrition_Flag')
```

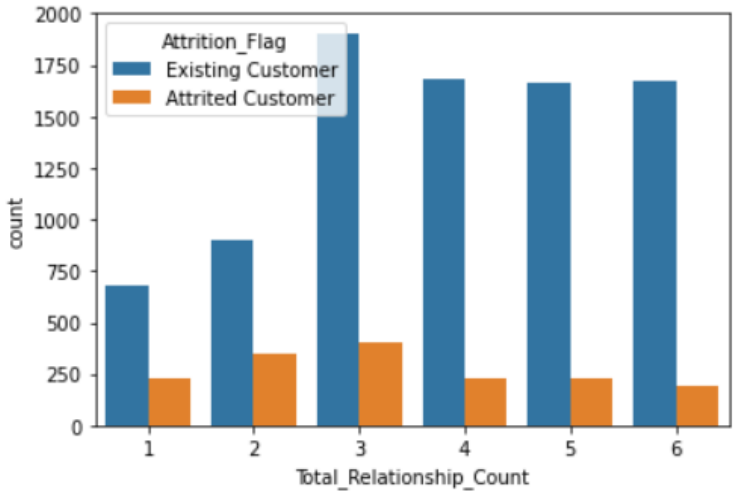
```
<AxesSubplot:xlabel='Card_Category', ylabel='count'>
```



Customer churn also seems to stay very consistent once a customer has more than three products or services. This could be because in addition to the checking and savings accounts that could be in the two to three product groupings, having four or more products or services could be the result of having loans, mortgages, and other products or services that would require a customer to have good credit standing.

```
sns.countplot(data = ccData, x = 'Total_Relationship_Count', hue = 'Attrition_Flag')
```

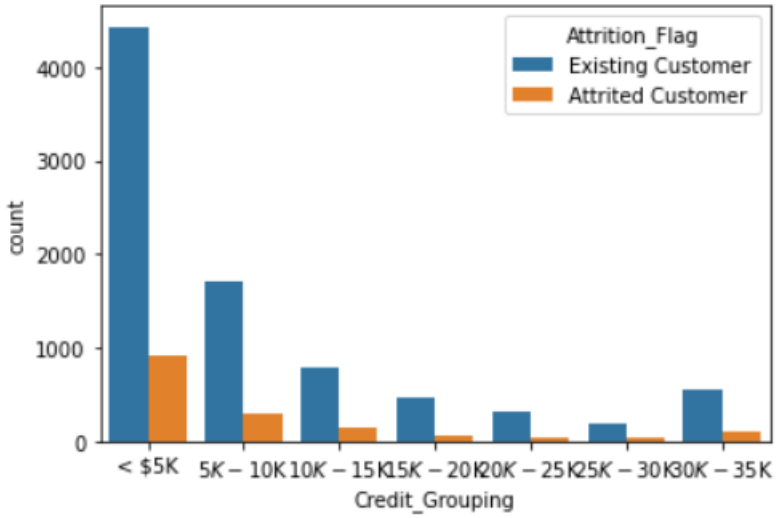
```
<AxesSubplot:xlabel='Total_Relationship_Count', ylabel='count'>
```



This seems to be backed up by looking at credit grouping versus churn, which decreases as credit increases. The lower number of customers in higher credit groups does provide some limitation to the category.

```
sns.countplot(data = ccData, x = 'Credit_Grouping', hue = 'Attrition_Flag')
```

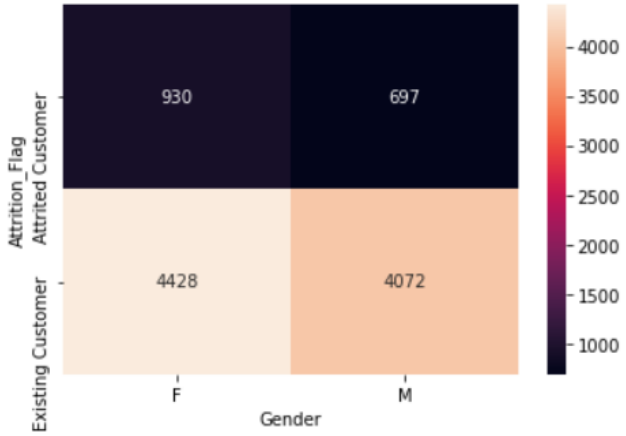
```
<AxesSubplot:xlabel='Credit_Grouping', ylabel='count'>
```



The heatmap below compares gender to customer churn. For both males and females there is less churn than retention, but female customers are about 2% more likely to churn than male customers.

```
sns.heatmap(pd.crosstab(ccData['Attrition_Flag'], ccData['Gender']), annot = True, fmt = 'd')
```

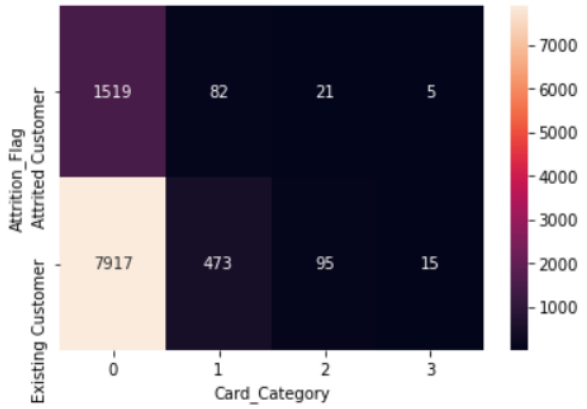
```
<AxesSubplot:xlabel='Gender', ylabel='Attrition_Flag'>
```



Lastly, I looked again at card category versus customer churn. The percentages were that 16% of Blue customers, 15% of Silver customers, 18% of Gold customers, and 25% of Platinum customers churned.

```
sns.heatmap(pd.crosstab(ccData['Attrition_Flag'], ccData['Card_Category']), annot = True, fmt = 'd')
```

```
<AxesSubplot:xlabel='Card_Category', ylabel='Attrition_Flag'>
```



Conclusions

I included many of my conclusions in the experiments section of this report, so here I will focus on what I considered the main takeaways of the EDA, what limitations were present, and how this could have been expanded done differently.

When it comes to card utilization, the credit limit has a highly negative relationship, while the revolving balance of the customer has a slightly positive relationship. When combined with the fact that the majority of customers make less than $40,000 in yearly income, it is possible that this relationship is from customers who are poorer are more likely to need to draw from their limited credit lines, and this lower income is decreasing their ability to pay off debt and increasing their revolving balance. Since this is a very negative scenario for the customers, it would be in poor taste for the bank in question to try and encourage this, even though it dramatically increases card utilization. Customers who use many bank services and customers who are the main providers for large families tend to use their cards more.

Regarding customer churn, the impact that each individual feature has on detecting this is minimal. Customers who are more affluent and have high card utilization are less likely to cancel their card, as are customers with at least four products or services with their bank.

Like was said in the dataset source, this dataset proved somewhat challenging to make conclusions with. The low number of customers who had churned combined with the very low number of customers in various categories of the data were a limiting factor in the analysis. As an example, I found it hard to consider if being a Platinum card holder led to a 25% chance of cancelling their card since there were only 20 customers with that kind of card. Having a more diversified set of customers for the categories would have been very helpful, but since that was not available, I could have normalized the data in certain categories as well as searched for outliers in the data to mitigate this.

If I were to continue working on this project, I would have liked to add more complex visualizations and compared more data features against each other. Additionally, I was hoping to be able to implement a machine learning model to perform predictions on customer churn, as a model would be able to more easily look through the large number of features to come to a conclusion as long as the data is prepared for it.