Insights on New Credit Card Product Opportunities

Group Assignment - Group 29

Group members contributing:

- Radmila Levin
- Haresh Gopiani
- Robert Fournier

Submitted to Data Science 1: Foundations of Data Science
April 11th, 2021

1. Objectives

The objectives of this report are to analyze the credit card data for 10,127 bank customers in order to draw insights into their spending and use of credit card products in order to better segment and sell to credit card users.

The data used in this report are all sourced from this Kaggle dataset called Credit Card customers.[1] According to the Kaggle data description, this data was initially sourced from Leaps (https://leaps.analyttica.com/home). Leaps is an online learning platform and directory for data science. In this scenario, the data is from an anonymous bank containing anonymous credit card customer data points.

As businesses move to adopt data science into decision making, it is critical to leverage the data available to businesses and draw insights that can lead to real action and insights for business leaders.

This report will aim to do so for DataBank, the fictional bank that we've received a contract to analyze their existing products in order to advise on what new products they should focus on creating. Through leveraging data analysis, logistic regression, and clustering, this report will make final recommendations to DataBank leadership on what new products are most viable.

The null hypotheses at the start of this paper are as follows:

- There are no correlating attributes to credit card spending
- There are no correlating attributes to attrition
- There are no marketable demographics for credit cards
- The market is fully seized by existing products

This paper will challenge each hypothesis through the various sections contained within. To begin, a thorough analysis of attributes and attrition by Haresh. Then, a logistic classifier by Radmila focusing on attrition. Finally, clustering of credit card users by Robert.

At the end of the paper, it is our objective to present our findings and insights on new credit card opportunities to the executives of DataBank.

[1] Goyal, S. (2020, November 19). Credit card customers. Retrieved April 11, 2021, from https://www.kaggle.com/sakshigoyal7/credit-card-customers

2. Data Preparation

The data used for this report was sourced from Kaggle, under the dataset “Credit Card customers” available at this url: https://www.kaggle.com/sakshigoyal7/credit-card-customers. Kaggle is a well known repository of datasets and data science notebooks and challenges by industry, and it was chosen for its renown and broad repository.

The Kaggle dataset author indicates that the data was initially sourced from Leaps Analyticca (https://leaps.analyttica.com/). Leaps is an online data science learning platform that aims to host webinars by data science experts to those learning the field. This particular dataset is from a course on Naive Bayes, and is sourced from an anonymous bank.

This dataset was chosen over other options for its similarity to business datasets that our team had encountered and to best simulate a real business data problem.

The data was fairly well cleaned for analysis at the start, available to download from the Kaggle website as a CSV file or importing to the notebook from Kaggle API commands. Our team leveraged a download of the CSV file given its relatively small file size (1.5 mb).

Team members worked on distributed notebooks at their own pace on subsets of the problem and sections for analysis. The various team member notebooks are included as separate appendices at the end of this report.

3. Analysis

3.1. Data Analysis - Haresh

Data analysis is the process of collecting, modelling, and analyzing data to extract insights that support decision-making. There are several methods and techniques to perform analysis depending on the industry and the aim of the analysis.

Customers are perhaps the most critical element in any business. By using data analysis to get a broader perspective of all aspects related to customers, it is easy to understand their demographics, interests, habits, purchasing behaviors, and more. Analyzing data is beneficial, as it helps to make business decisions based on facts and not simple intuition.

In the long run, it will drive success to the marketing strategies, allow identifying new potential customers, and avoid wasting resources on targeting the wrong people or sending the wrong message.

The descriptive analysis method is essential, as it allows presentation of data in a meaningful way. It does this by ordering, manipulating, and interpreting raw data from various sources to turn it into valuable insights to the business.

Preliminary Observation:

1. There are 1,627 Attrited and 8,500 Existing clients.
2. Attrited clients consist of 57% of females and 43% of males.
3. Existing clients consist of 52% of females and 48% of males.
4. Education level of card holders has been categorized in 7 different areas. (College, Graduate, Post-graduate, Doctorate, High School, Uneducated, Unknown).
5. In population of Attrited clients, highest proportion were Graduates (29%) and the lowest proportion were Post-graduate (5.2%).
6. In population of Existing clients, highest proportion were Graduates (31%) and the lowest proportion were Doctorate.
7. Most Attrited clients and Existing clients (38% and 35%) fall in less than $40K income group
8. All clients are divided into four card categories (Blue, Gold, Platinum, Silver). Majority of the clients have Blue cards in both Attrited and Existing client groups.
9. Highest Transaction Amount (90%) among card categories, was from Blue card holders in both Attrited and Existing clients.

10. Highest percentage (Attrited - 25% and Existing - 24%) of female customers are married from both Attrited and Existing customer groups.
11. Minimum credit limit varies (1,400 to 15,000) amongst different type of cards. However, the maximum limit is same for all cards (35,000)
12. The highest transaction values have been observed in elderly males and females (Age group of 46 to 55 years) with Blue credit cards.

3.2. Logistic Regression - Rada

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. Logistic regression has become an important tool in the discipline of machine learning. The approach allows an algorithm being used in a machine learning application to classify incoming data based on historical data. As more relevant data comes in, the algorithm should get better at predicting classifications within data sets.

Logistic regression is very similar to linear regression (that is y = x), although it is for solving for classes (classification) and not for actual regression (continuous values). Further details on logistic regression is available from sci-kit learn documentation.[2]

Credit Card Ownership

Visualization of what type of people hold credit cards based on different criteria is included in Appendix B.

We have the following observations:

- Most credit card holders are Graduates and High school students and least credit card holders are Doctorate.
- Most credit card holders are Married people and least cc holders are Divorcees.
- People with an income less $40,000 holds the maximum number of credit cards.

Credit Card Churn

Visualization of churn rate by categories is included in Appendix B.

We have the following observations:

- Doctorate is more likely to churn.
- Married customers are less likely to churn.
- Customers with an income less $40,000 and greater than $120,000 are more likely to churn.
- Customers with **Platinum** card category has the highest churning rate.

---

[2] *1.1. Linear Models — scikit-learn 0.24.1 documentation*. (n.d.). Scikit Learn. Retrieved April 11, 2021, from https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

After analyzing the data, we split training and test data into 80% and 20% and used Logistic Regression for prediction accuracy.

After calculating metrics for performance evaluation of the model we received:

```
Recall: 0.818
Accuracy: 0.849
```

Recall is the ability of the model to find all positive samples.[3] In our case, the recall means that the model found 81% of all attrition flags. Accuracy refers to the amount of correct guesses by the model compared to incorrect guesses.

We received very high accuracy and recall rate on the Test dataset. So we can predict the churn customer with an accuracy of 84.9%.

[3] *3.3. Metrics and scoring: quantifying the quality of predictions — scikit-learn 0.24.1 documentation*. (n.d.). Scikit Learn. Retrieved April 11, 2021, from https://scikit-learn.org/stable/modules/model_evaluation.html#precision-recall-f-measure-metrics

3.3. Clustering and User Profiles - Robert

In this section of the analysis, credit card data is clustered based on key features to build user profiles and segments. Based on that data, further insights on demographics to target or user profiles to target can be determined.

This section is broken down into three subsections:

1. Methodology
2. Iterative clustering
3. Clustering analysis

Methodology

Analyzing 10,127 data points can be difficult at the minute, micro-level for leaders. Often times it is easier to take a wider approach, gathering like items together and leveraging insights that can affect more than the individual - impacting groups of individuals at a time. But how to group such diverse and numerically dense information such as credit card data?

That is the business case for clustering: impacting swathes of datapoints in the aggregate with the power of machine learning algorithms.

This analysis leverages one of the most famous clustering algorithms: k-means. As per scikit-learn's documentation, k-means divides a set of samples into disjoint clusters.[4] It aims to minimize inertia, otherwise known as within-cluster-sum-of-squares criterion. That formula is defined as:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

K-means was selected as the clustering algorithm in part due to its widespread usage, but also its applicability to a broad array of use-cases as it's known for its general-purpose nature. With 10,127 samples, we have a sufficiently large dataset for k-means as well.[5]

---

[4] *sklearn.cluster.KMeans*. (n.d.). Scikit Learn. Retrieved April 11, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[5] *Choosing the right estimator — scikit-learn 0.24.1 documentation*. (n.d.). Scitkit Learn. Retrieved April 11, 2021, from https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

From a data perspective, the dataset was normalized before each iteration along the features identified for that version. That was accomplished through subtracting the mean of the feature, and dividing by the standard deviation. This was accomplished to reduce the effect of outliers on the data and to allow for simpler and faster clustering processing.

Importantly, a decision was made at this point to exclude categorical features from the clustering data. Categorical features have been known to cause clustering algorithms to “cheat” through creating binary feature spaces. For example, a gender feature of [1] for Male or [0] for Female would create an exaggerated pull in the multi-dimensional space compared to non-categorical (continuous) features.

This decision aligned with the objective to understand credit card customer behaviours and to then map categorical and sociological variables onto that analysis.

The approach taken to building the k-means clustering model was to build iteratively, leveraging a test and learn approach on what features should be included and parameter optimization for the number of clusters.

Iterative Clustering

The approach taken to clustering was to build successive clustering k-means models based on a test and learn approach. The focus was to build light and quick, deploying new lessons learned onto each successive iteration.

Each iteration would test a set of features to include in the analysis as well as various parameters to optimize for the number of clusters to initiate in the scikit-learn k-means object. This approach allowed for a wide and shallow optimization to determine the best possible clustering approach based on the numerical data on hand.

The following table (3.3.A) provides a breakdown of the results off each iteration, along with the key differences, silhouette scores and silhouette plot for the best result.

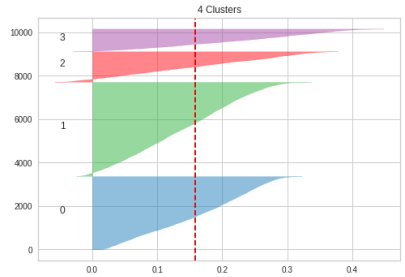
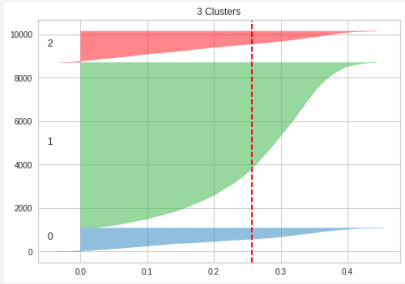
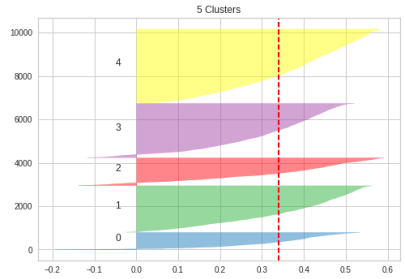
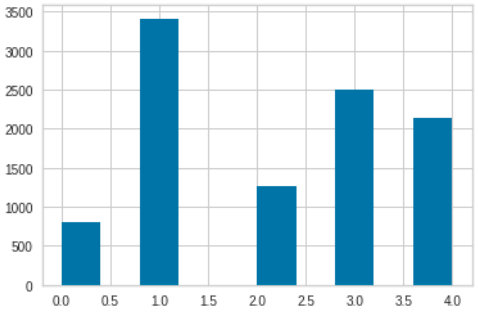
	Features	Best silhouette score (clusters number)	Silhouette Plot of best cluster number
Iteration 1	‘Months_on_book' ‘Total_Relationship_Count’ ‘Months_Inactive_12_mon' ‘Contacts_Count_12_mon’ ‘Credit_Limit' ‘Total_Revolving_Bal' ‘Avg_Open_To_Buy' ‘Total_Amt_Chng_Q4_Q1’ ‘Total_Trans_Amt' ‘Total_Trans_Ct’ ‘Total_Ct_Chng_Q4_Q1 'Avg_Utilization_Ratio'	0.158 (4 clusters)	
Iteration 2	‘Months_on_book' ‘Total_Relationship_Count’ ‘Months_Inactive_12_mon' ‘Contacts_Count_12_mon’ ‘Credit_Limit' ‘Total_Revolving_Bal' ‘Avg_Open_To_Buy' ‘Total_Trans_Amt' ‘Total_Trans_Ct’	0.257 (3 clusters)	
Iteration 3	‘Credit_Limit' ‘Total_Revolving_Bal' ‘Avg_Open_To_Buy' ‘Total_Trans_Amt' ‘Total_Trans_Ct’	0.339 (5 clusters)	

Table 3.3.A

The results of the iterative approach was that Iteration 3’s feature set and parameter set to 5 clusters resulted in the best combination of cluster sizes and silhouettes. Given time parameters, further iterations were not attempted, although certainly there is the chance that more feature reduction or simplification may have resulted in better clustering results.

Analysis

To begin, a breakdown of the five clusters was performed based on the distribution of users across each group. As hinted in the above silhouette chart, the distribution is fairly even, although Cluster 2 and 3 are smaller in size.



Cluster Distribution, Chart 3.3.B.

Next, a pairplot was conducted to consider the distribution of clusters across variables and the interconnectedness of the variables themselves. This is available in Appendix B under Chart 3.3.C.

As a result of the clustering algorithm and analysis, five credit card user profiles have been created. Leveraging insights from the features used to cluster, the following key assumptions can be made for each group.

- **High Volume Users:** They spend lots on many transactions with a range of credit limits. These are ideal users for transactional fees.
- **Low Credit Limit Users:** They spend more and in more transactions, but are differentiated with low credit limits. Due to the low credit limits, these customers have lower potential margins on fees, but do carry more of a revolving balance than other groups. They also have the lowest open to buy averages, meaning they spend closest to their limits.
- **High Credit Limit Users:** They keep a high credit limit, while spending more and in more transactions than Pay it Off or Low Volume Users. They also have the second smallest revolving balance, meaning they aren't likely to be caught in fees. But with a high credit limit, they are likely to take on new products and expand their limits.
- **Pay It Off Users:** They are mostly uniquely characterized by a low revolving balance, low credit limit, low transactions in count - although they do make many purchases on the card. These are customers with low revenue potential. These are unlikely to default.

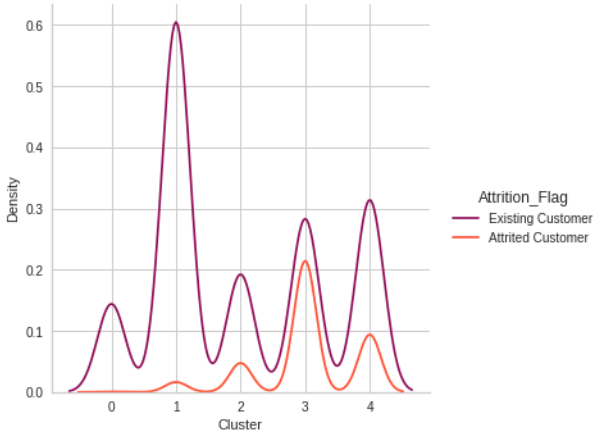


Chart 3.3.H

- **Low Volume Users:** They barely use the credit card (low transaction amounts and counts), although they carry the most revolving balances than any other user. Given that, they are likely to incur fees and generate revenue.

The following graphs provide a review of the distribution of various socio-economic and other categorical features for each cluster:

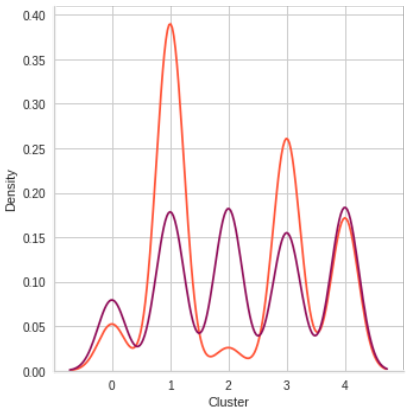


Chart 3.3.D

From the charts above,

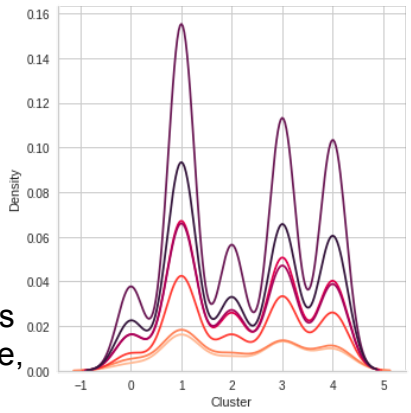


Chart 3.3.E

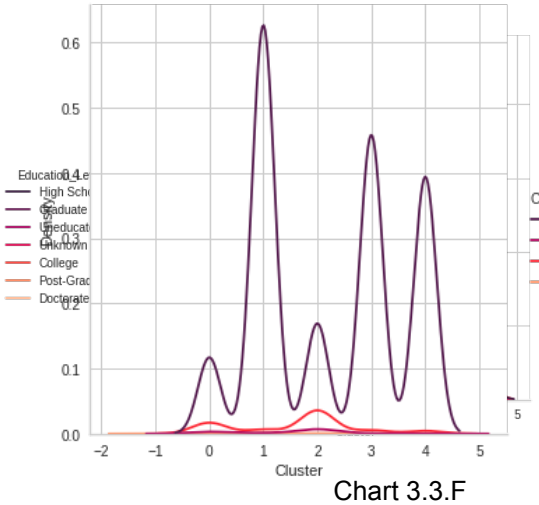


Chart 3.3.F

Chart 3.3.G

the following assumptions can be made:

- Women outnumber men in the Low Credit Limit and Pay It Off profiles.
- There is not a significant income category for high volume users, nor education level, though they are more likely to men.
- The biggest gender gap in favour of men is in Low Volume Users, while High Credit Limit Users are the most evenly split between women and men.

Chart 3.3.G to the right displays the current distribution of credit card products at DataBank to current users. As seen, High Volume Users and High Credit Limit Users are likely to have the most diverse card types and the least skew to Blue Cards. There is an opportunity to upsell Low Credit Limit Users (otherwise frequent ideal users) to more products with proper marketing. Low Volume Users are also a targetable market for new products, given the high revolving balance. These are also likely to be users to focus on retaining and growing services to satisfy.

Chart 3.3.H to the right displays attrition across user profiles as well. The noteworthy insights from this attrition analysis suggests that Low Credit Limit Users are very unlikely to exit, while Pay It Off Users are more likely to churn. Given both Low Credit Limit Users reliance on Blue Cards and their loyalty, they make a key demographic to focus on.

4. Conclusions

This paper has concluded the following observations for DataBank's leaders in determining the best new credit card products:

Blue Card customers can be diversified into more targeted products

The gross majority of customers have Blue Cards. These cards are prevalent in all user profiles, as well as across genders. These Blue Card users are generally high transaction users as well. The potential exists for this credit card user segment to be further targeted and more niche products used to better augment user experience and bank revenues. These products may be new, or into existing credit card products that are under-utilized. Specifically Gold and Silver Cards. Niche targeting to certain demographics may help in designing these products, including elderly men and women with Blue Cards who have more transactions than other segments.

Opportunity lies in focusing on non-traditional user segments

The adage in banking is that high spenders are the ideal customers to focus on given the margins made on their transactions and their potential investments to be leveraged into capital. However, users with current low credit limits are critical to DataBanks success as they hold low attrition and are mostly on the basic product (Blue Cards). This non-traditional user segment should be targeted for new products and marketing.

High value segments and customers are exiting

While the above non-traditional approaches may work to expand business, there is concerns on the attrition for high-end users with Platinum Cards, users with Doctorate degrees, and users with greater than $120,000 in annual income. Further analysis should be conducted, including user research, on these demographics to determine pain points and exit ramps for targeted solutions. Machine learning models, such as the logistic regression model used in this paper, should be developed and deployed to create early warning flags to address attrition and deploy solutions or support.

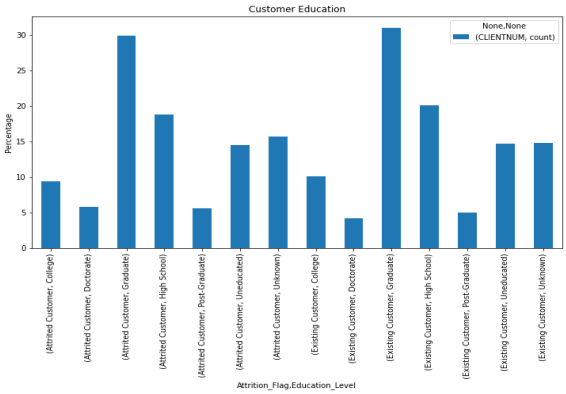
Appendix A - Python Notebooks

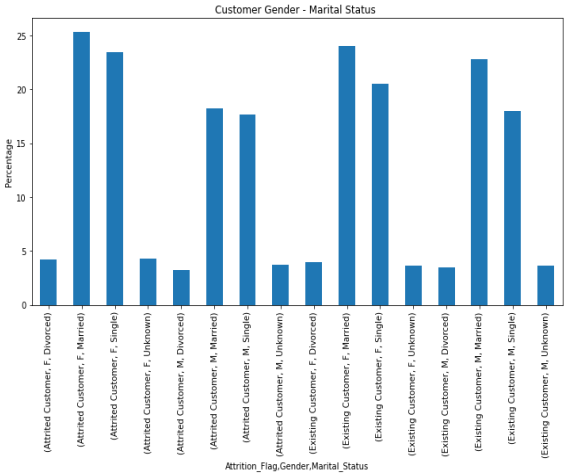
Python notebooks attached:

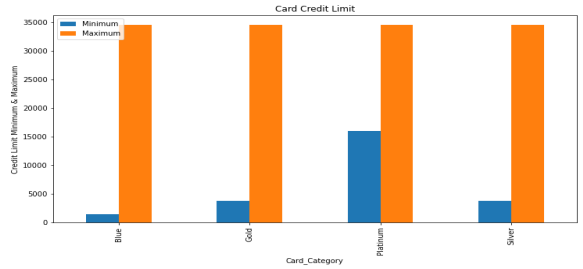
Appendix A.1. - Analysis Notebook
Appendix A.2. - Logistic Regression Notebook
Appendix A.3. - Clustering Notebook

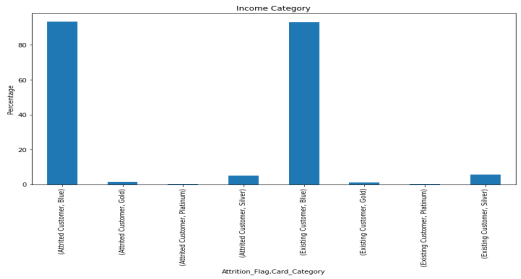
Appendix B - Charts

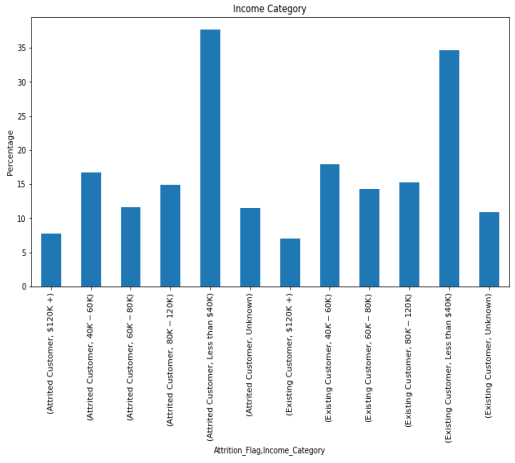
Section 3.1. Charts

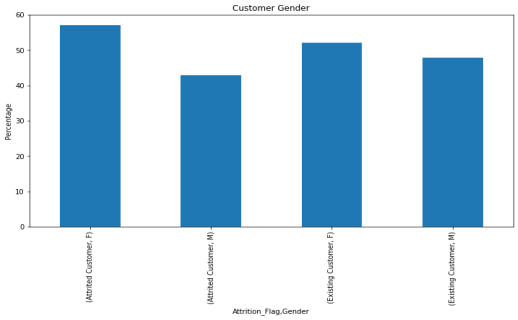


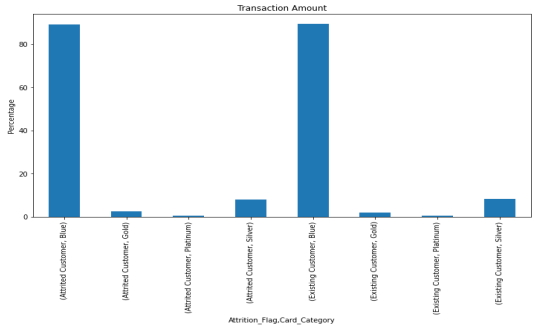


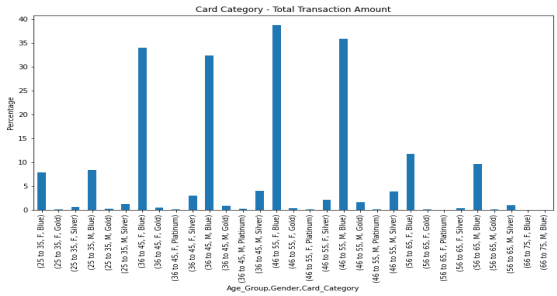




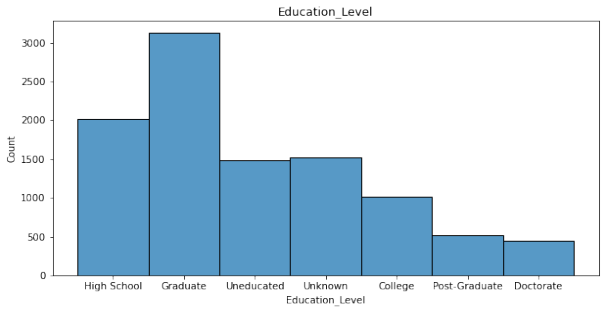


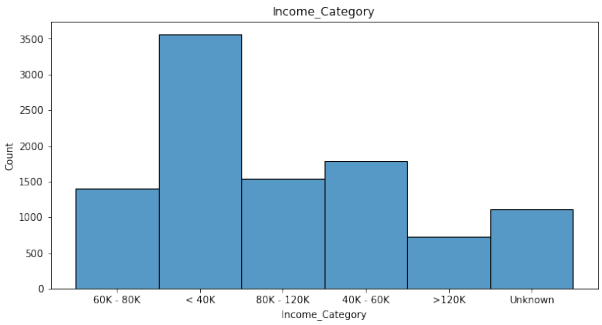


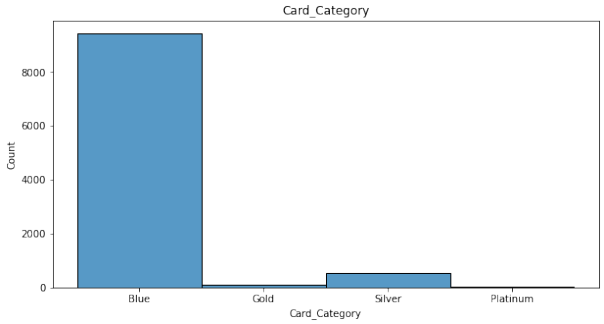


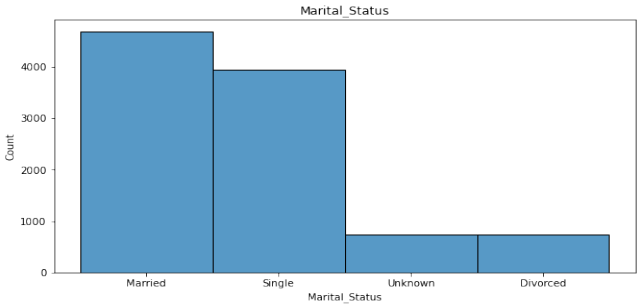


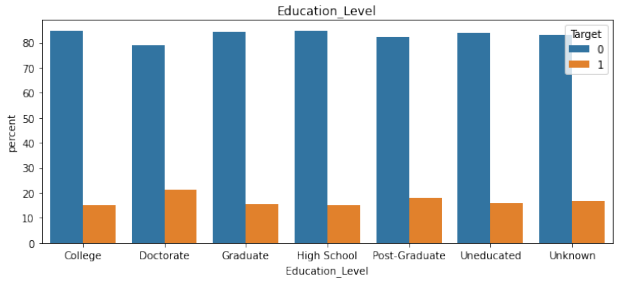
Section 3.2. Charts

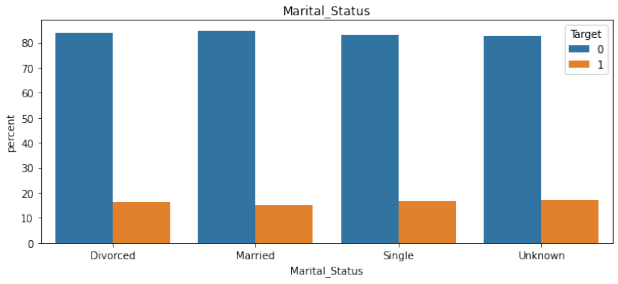


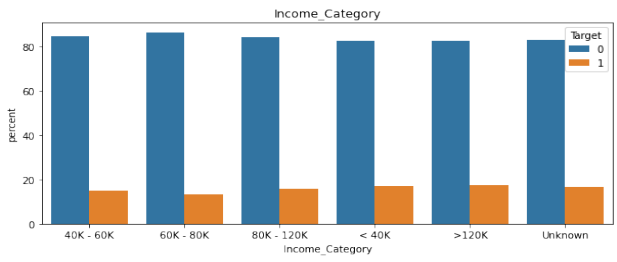


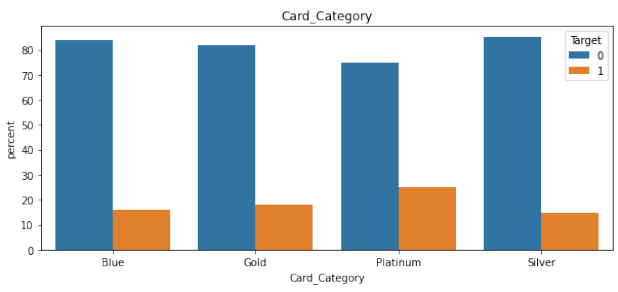












Section 3.3. Charts

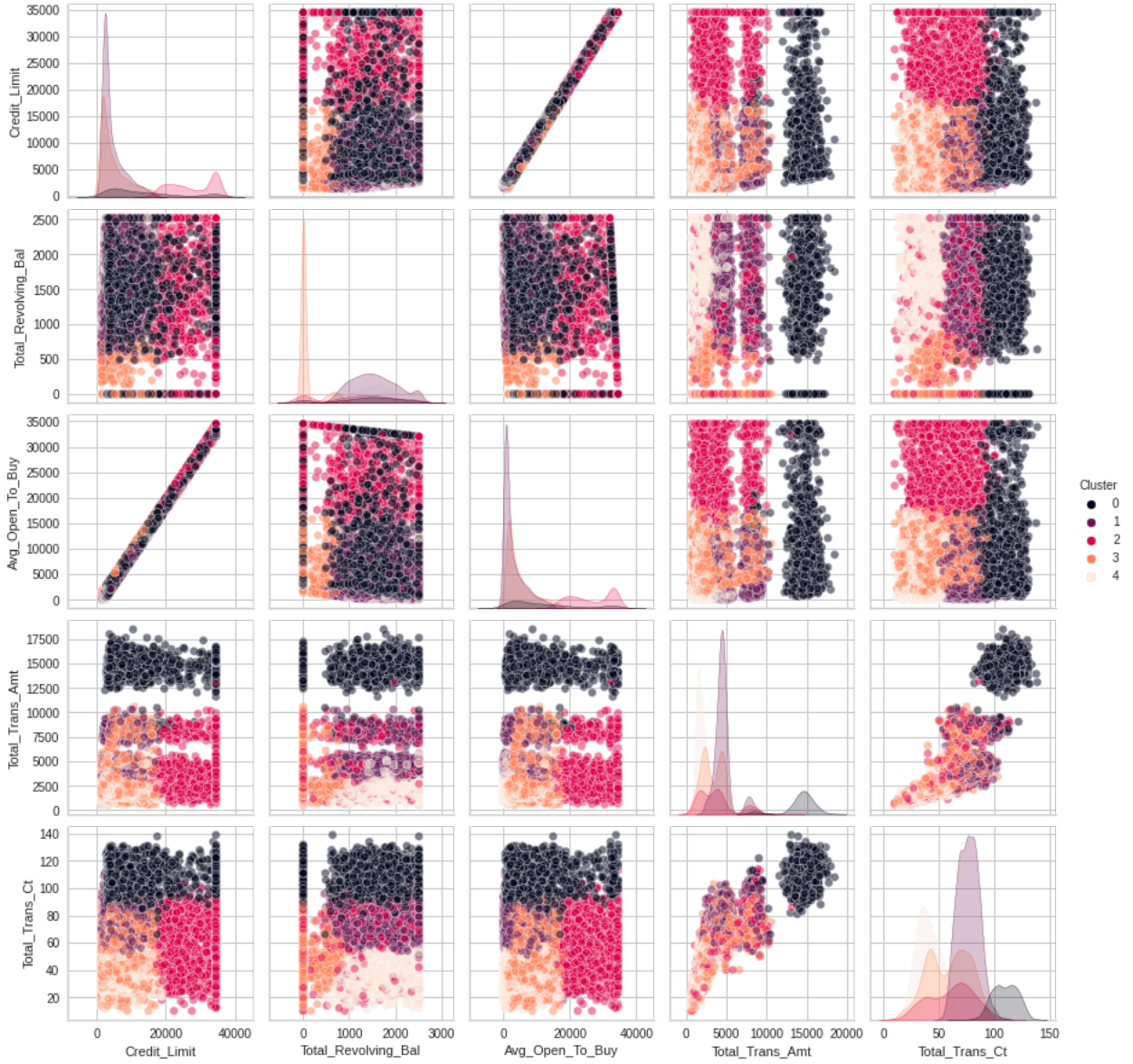


Chart 3.3.C.

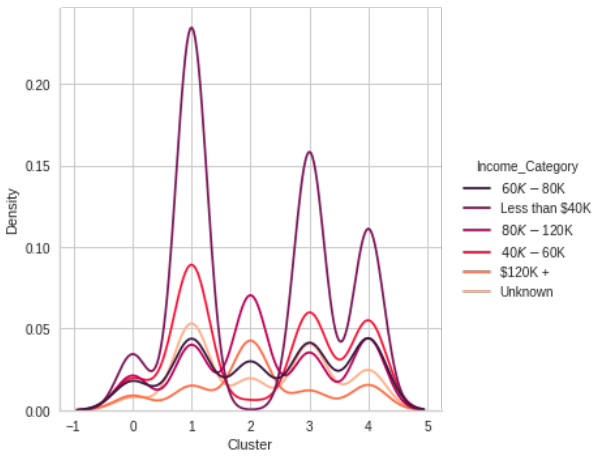


Chart 3.3.F

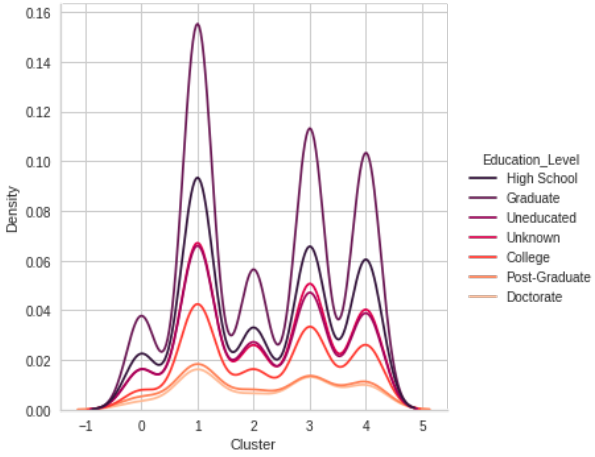


Chart 3.3.E

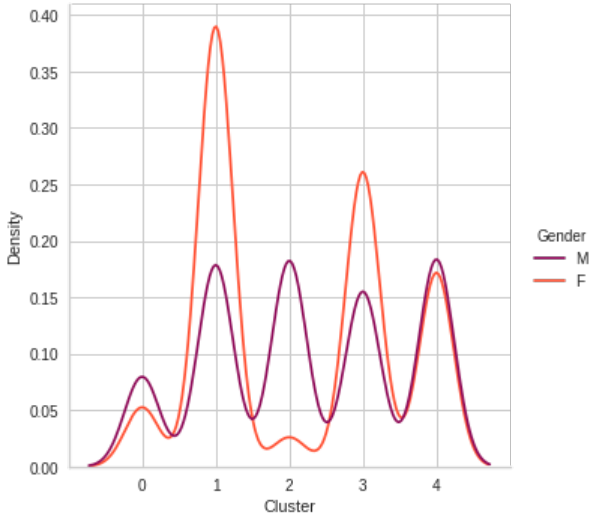


Chart 3.3.D

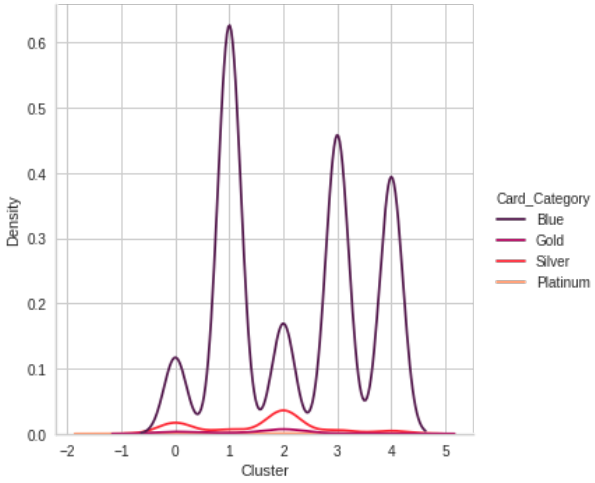


Chart 3.3.G

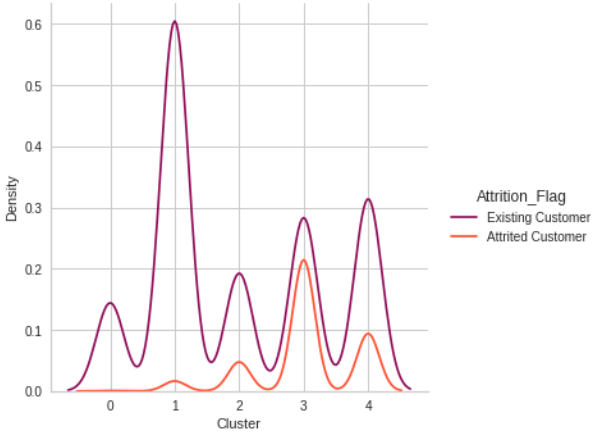


Chart 3.3.H