

**School of InfoComm Technology**

**Data Exploration & Analysis Assignment**

Diploma in DS

April 2023 Semester

**ASSIGNMENT 1**

(30% of DEA Module)

7th May 2023 – 4th June 2023

**Submission Deadline:**

**Presentation: 4th Jun 2023 (Sunday), 11:59PM**

**Excel files: 4th Jun 2023 (Sunday), 11:59PM**

|  |  |  |
| --- | --- | --- |
| **Tutorial Group** | **:** | **T01 / T02 / T03** |
| **Student Name** | **:** | Clayton Thia |
| **Student Number** | **:** | S10243162 |

**Penalty for late submission:**

10% of the marks will be deducted every calendar day after the deadline.

**NO** submission will be accepted after 11th Jun 2023 (Sunday), 11:59PM.

Table of Contents

[1](#_Toc136802619)

[Abstract/ Overview 3](#_Toc136802620)

[Problem Statement 4](#_Toc136802621)

[Data Exploration 5](#_Toc136802622)

[Univariate analysis 5](#_Toc136802623)

[***Age groups*** 5](#_Toc136802624)

[***Ratio of gender*** 6](#_Toc136802625)

[***Dependents analysis*** 7](#_Toc136802626)

[***Education Level Analysis*** 8](#_Toc136802627)

[***Marital Status distribution*** 9](#_Toc136802628)

[***Income Group trend*** 10](#_Toc136802629)

[***Proportion of churners*** 10](#_Toc136802630)

[***Card Categories distribution*** 11](#_Toc136802631)

[***Months on Book*** 12](#_Toc136802632)

[***Credit Limit*** 12](#_Toc136802633)

[***Relationship Count analysis*** 13](#_Toc136802634)

[***Inactive months*** 13](#_Toc136802635)

[***Number of Contacts*** 14](#_Toc136802636)

[***Card Utilization ratio*** 14](#_Toc136802637)

[***Count of transactions*** 15](#_Toc136802638)

[***Total Transaction Amount*** 15](#_Toc136802639)

[Bivariate analysis 16](#_Toc136802640)

[***Months on Book vs Customer Age Group*** 17](#_Toc136802641)

[***Total Transaction Amount vs Relationship Count*** 18](#_Toc136802642)

[***Total Transaction Count vs Relationship Count*** 19](#_Toc136802643)

[***Credit Limit vs Utilization Ratio*** 20](#_Toc136802644)

[***Utilization ratio vs age group*** 21](#_Toc136802645)

[Multivariate analysis 22](#_Toc136802646)

[***Credit Limit vs Utilization ratio vs Gender*** 24](#_Toc136802647)

[***Utilization ratio vs Age Group vs Marital Status*** 25](#_Toc136802648)

[***Total Transaction Amount vs Relationship Count vs Card category*** 26](#_Toc136802649)

[Summary 27](#_Toc136802650)

[Findings 27](#_Toc136802651)

[Reflection 28](#_Toc136802652)

# Abstract/ Overview

The practice of repeatedly opening and closing credit card accounts to take advantage of sign-up bonuses and rewards is called credit card churning, and it has gained a significant amount of attention in recent years from financial institutions. The role data exploration plays in the combat against this practice is crucial – it reveals valuable insights and patterns within datasets.

The goal of this report is to discuss the data from the given dataset in univariate, bivariate, and multivariate perspectives using data exploration techniques. During the data exploration process, findings will be backed up with supporting visualizations. Moreover, the choice of visualizations and critical findings will also be explored and explained.

Starting off with defining the problem statement, the importance of data exploration is emphasized in tackling churning effectively.

Data exploration will be performed in Excel and Jupyter Notebook, and all resources including websites and tools will be included.

# Problem Statement

Credit card churning, the process of opening and cancelling credit card accounts frequently in order to take advantage of sign-up incentives and perks, has grown to be a major worry for financial institutions all over the world. Numerous issues are raised by this practice, including higher client acquisition costs, decreased profitability, and potential reputational harm. It is essential to fully comprehend the dynamics and underlying reasons of credit card churning in order to address this issue, as well as to create practical mitigation measures.

Credit card churning and its negative consequences on financial institutions are the issue at hand. Credit card churners take advantage of enticing promotional offers to open new accounts, accrue rewards, and then close the accounts right away before incurring any significant charges. The planned customer lifecycle is disrupted by this practice, which also hurts profitability and the capacity of financial institutions to draw in and keep valuable clients.

Out of all the challenges in addressing the issue of credit card churning, the one this report is focusing on is the identification of churn patterns – namely understanding the patterns and hallmarks associated with credit card churning. Identifying specific characteristics of individuals that partake in churning will require analysis of historical data.

The primary objective of this study is to explore the data given other resources and discuss the data in univariate, bivariate, and multivariate perspectives so as to study the traits of customers who had churned.

Financial organizations can create targeted measures to retain clients, optimize profitability, and maintain regulatory compliance by studying the underlying patterns and dynamics of credit card churn. This study aims to provide useful information to help the industry avoid the negative impact of credit card churning.

# Data Exploration

*Before beginning my data exploration, I first created two extra columns – Age group and Years on Book. Age group is created with a VLOOKUP table that is included in the excel file while Years on Book is a function where I divided the value under Months on Book by 12. Doing this also helped me to reduce the impact that any outliers had on those variables.*

## Univariate analysis

### ***Age groups***

A picture containing text, screenshot, number, font

Description automatically generated

**Blue: Non-churned Orange: Churned**

Using this bar chart, we can see that most of the customers are from the age range of 30 – 59 years old. The distribution for this category is slightly left skewed, and there are no churned customers belonging into the age group of 70-79. This could be because sign-up bonuses aren’t as enticing to the elderly as they are to the adults, or it could also be because the bank just doesn’t have enough customers in that age group.

*I used a bar chart as it allows for comparison between categories. In this case, I wanted to see which age group had the highest quantity of customers.*

### ***Ratio of gender***

A blue and pink pie chart

Description automatically generated with medium confidence

The pie chart above shows the gender ratio of customers who did not churn.

A blue and pink pie chart

Description automatically generated with medium confidence

The pie chart above shows the gender ratio of customers who churned.

We can see that amongst the customers who didn’t churn, there is a higher percentage of males while in amongst those that ended up churning there is a higher percentage of females. This may suggest that females are more likely to end up churning.

*I used a pie chart as it allows for comparison between categories and the total. In this case, I wanted to see what percentage of the total each gender took up. Since there are only two genders in this dataset, it also allowed me to easily compare them like a bar chart.*

### ***Dependents analysis***

This pie chart above shows us the proportion of dependents that non-churned customers have.

The pie chart above shows us the proportion of dependents that churned customers have.

We can see that for both churners and non-churners the proportion of the number of dependents is similar – most of them have 1-4 dependents. This may be due to the fact that people with 5 dependents may be tight on cash and cannot afford to spare any money for their bank account.

*I used a pie chart as it allows me to see which categories take up most of the total. In this case, I can see the range of dependents for majority of the customers.*

### ***Education Level Analysis***

A picture containing line, text, screenshot, plot

Description automatically generated

**Blue: Non-churned Orange: Churned**

The combo chart with secondary axis above shows the number of customers for both churners and non-churners. As we can see, the line representing churners follows the distribution of the bars representing non-churners quite closely. It would be safe to say that the percentage of churners is about the same for all education levels, except for Post-graduates and Doctorates which have higher percentages.

*I used a combo chart as lets me overlay the distribution of the churners against the distribution of the non-churners for the education level category.*

### ***Marital Status distribution***

The pie chart above shows the proportion of marital status amongst the churners.

The pie chart above shows the proportion of marital status amongst the non-churners.

Between non-churners and churners, the proportion of marital status remains about the same.

*I used a pie chart to see the proportion of the marital status amongst the customers.*

***.***

### ***Income Group trend***

This line chart shows us that there is a general downwards trend of the number of customers as the income group gets higher for both churned and non-churned customers.

*A line chart was used as they are excellent for displaying trends and patterns in data over a series of chronological categories or periods of time.*

### ***Proportion of churners***

Here we can see that only a small percentage of the data belongs to churners. What this means is that we’ll be generalizing the features of the churners to predict whether the rest of the population will churn. This means that some of the insights gained from this exploration may be inaccurate.

*I used a pie chart as it allows me to easily compare the proportions of features.*

### ***Card Categories distribution***

The chart above shows the proportion of card categories amongst churned customers.

The chart above shows the proportion of card categories amongst non-churned customers.

For both types of customers, there is the same proportion- an overwhelming amount of the chart represents the blue cards, while the rest of the chart represents the silver, gold and platinum cards in decreasing amount respectively.

*I used a donut chart as like a pie chart, they are good for comparing values with respect to the total. However, I did not use a pie chart as I wanted to add a bit of variety of the univariate visuals.*

### ***Months on Book***

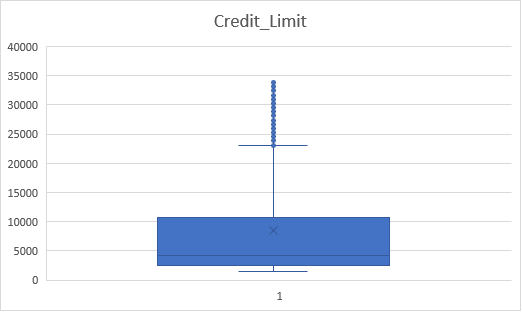
A picture containing text, screenshot, line, plot

Description automatically generated

This box plot tells us that the months on book variable assumes a somewhat normal distribution as the mean is extremely close to the median.

*For the rest of the continuous data, I use a box plot as it is one of the only ways to visualize continuous data for univariate analysis.*

### ***Credit Limit***



We can see that most of the data lies below 25000. The mean is above the median which means the distribution will have a right skew.

### ***Relationship Count analysis***

Relationship count is the number of products the customer has. This visual shows us that most churned customers have 2 or 3 products with a sharp drop in the number of customers who have 4 and above products.

*For the rest of the univariate analysis, I use combo charts because they allow me to overlay the churned data over the non-churned data for easy comparison.*

### ***Inactive months***

We can see from this visual that most of the inactive months for regular customers are between 1-3. However, 2-3 months is when there is a risk of a customer churning as that is where the line representing churned customers spikes.

### ***Number of Contacts***

Contacts refers to the number of times that the customer has reached out to the bank in the last 12 months. We can see from this that the distribution of contacts for churners is nearly normal, if not for the slight right skew while the non-churners don’t follow any obvious patterns. Additionally, no non-churner has contacted the bank any more than 5 times.

*I swapped the type of chart for the type of customer as I wouldn’t be able to emphasize on the distribution of churners for this category otherwise.*

### ***Card Utilization ratio***

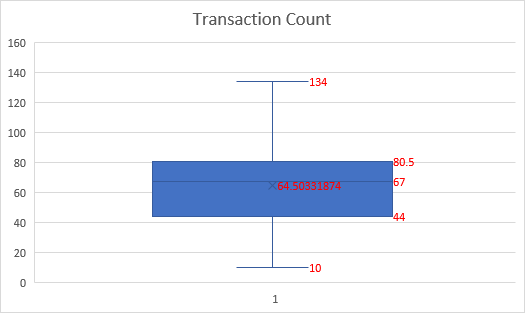
A picture containing text, screenshot, line, plot

Description automatically generated

Card utilization ratio is the amount of money you owe divided by your credit limit; it essentially represents how much of your credit limit you’ve spent. From this visual we can see that most customers use between 2.5 to 50% of their limit, with the highest being 99%.

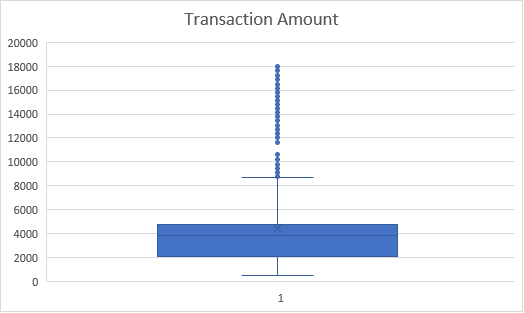
The mean is above the median, which means the distribution will have a right skew.

### ***Count of transactions***



The numerical data being represented here is the number of transactions made within the last 12 months. We can see that most customers make between 44 and 80 transactions per 12 months. The mean is below the median, which means the distribution will have a slight left skew.

### ***Total Transaction Amount***



We can see that while there are many outliers, majority of the data lies in the 0<X<9000 range. Most of the data in that range comes from the 2000<X<4500. The mean is above the median which means the distribution will have a right skew.

## Bivariate analysis

*I first made a correlation heatmap in Jupyter Notebook for existing and churned customers.*

**Churned**

A screenshot of a graph

Description automatically generated with low confidence

**Non-Churned**

**A screenshot of a computer

Description automatically generated with low confidence**

*Using the heatmaps, I decided to make visuals based on the correlations with colours that contrasted the most as that meant that they were the variables closest to having a somewhat linear relationship.*

### ***Months on Book vs Customer Age Group***

The line chart above shows the average years on for each age group for both churners and non -churners.

As the correlation heatmap earlier has shown, there is a strong positive linear relationship between the Age Group and Years on Book column. We can also see that for the non-churners and churners, their average years on book are also very similar. The exception in this case is the 70-79 age group which breaks the linear relationship. However, this may be because there isn’t enough sample data for that age group.

*I used a line chart as they are good for tracking changes over categories that appear in a series (for example, time). In this case, I am tracking the change in the average years on book as the customers get older.*

### ***Total Transaction Amount vs Relationship Count***

The area chart above shows the average total transaction amount for the past 12 months against the count of relationship. A relationship is the total amount of products held by the customer.

Using this area chart, we can see that for relationship counts 1 through 3, the average total transaction amount for non-churners is noticeably bigger than churners. For relationship counts 4 and beyond, the line between churner and non-churner gets harder to identify.

Additionally, the average total transaction amount for churners tends to stay around the 3k to 3.5k mark except for relationship count 3 which may have been skewed by an outlier.

Lastly there is a general downwards trend of transaction amount for non-churners as the relationship count increases.

The takeaway here is that once the average transaction amount stays below 4000, there is a higher chance of the customer churning.

*I used an area chart instead of a line chart for this visual as I felt that it better visualized the range of which a customer is at risk of being a churner.*

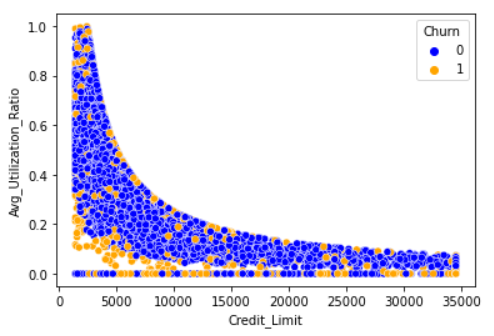
### ***Total Transaction Count vs Relationship Count***

The bar chart above shows the average count of transactions as the customer has more products for both types of customers. We can see that while the non-churners start to spend less as they get more products, the churners’ spending stays at around the same range of 40-50 regardless of how many products they have.

This may indicate that churners do not have a reason to go above 50 transactions, or they are able to get the bonuses they want within 50 transactions.

*I used a bar chart as I wanted to compare the transaction counts between the relationship counts. Bar charts also show the distribution and allow for easy comparison.*

### ***Credit Limit vs Utilization Ratio***



This scatter plot shows us the trend of utilization ratio as the credit limit increases for both types of customers where blue represents non-churners. The smooth curve shows us that as the credit limit increases, the utilization ratio starts to decrease at an increasing rate. This curve suggests that customers’ spending does not increase proportionally to their credit limit.

*I used a scatter plot to see the distribution churners and non-churners alike.*

### ***Utilization ratio vs age group***

This line chart shows the average utilization ratio as the age group gets older. As we can see, the utilization ratio for churners drops steeply from the 20-29 to 30-39 group, followed by a plateau at around 0.16. Meanwhile for non-churners, their average utilization rate stayed above 0.3 for all age groups except for the 40-49 group, which is still above the utilization ratio of all the churners above the age of 29.

This may be a critical finding as this shows that the utilization ratio may be used to predict the churn potential of a customer. For example, if a customer’s utilization ratio is hovering at around 0.16, it may mean they are going to churn.

*I used a double line chart as they are good for tracking changes over a period of time, or in this case periods of years.*

## Multivariate analysis

For my multivariate analysis, I decided to first start off by plotting a pair plot on Jupyter Notebook using seaborn and continue by expanding on my bivariate visuals. I colour coded to the low cardinality categorical columns. However, the gender colour coded plots were the only plots that were readable.

**Churned customers by gender**

A picture containing text, diagram, purple

Description automatically generated

**Non-churn customers by gender**

**A picture containing text, purple, majorelle blue

Description automatically generated**

Here, we can see that most of the customers with high credit limits are males for both churners and non-churners alike.

*I used a pair plot as it allows me to gain a quick and brief overview of the patterns between each continuous variables.*

### ***Credit Limit vs Utilization ratio vs Gender***

**Churners**

A picture containing text, screenshot, diagram, plot

Description automatically generated

**Non-churners**

A picture containing screenshot, text

Description automatically generated

From the scatter plot we can see that for churners and non-churners alike, the females tend to have a higher utilization ratio than the males. We can also see that the distribution of the churners as compared to non-churners is about the same. This shows us that looking at just the average utilization ratio and credit limit of a customer isn’t useful when trying to predict whether the customer would churn.

*I used a scatter plot as it lets me easily see the distribution of charts that contain two continuous and 1 discrete data.*

### ***Utilization ratio vs Age Group vs Marital Status***

*Expanding on bivariate*

This is a bar chart that shows the average utilization ratio of the different marital statuses in each age group, which is colour coded by the churn status. To extract the important information from this graph, we look at the difference between the orange and blue pairs.

For example,

1. 20-29 age group
   1. Single churners have about the same utilization ratio as single non-churners.
   2. We can see that for the married couples in the 20-29 age group, churners have a utilization ratio nearly twice that of the non-churners.
   3. Divorced non-churners have about 4x the utilization ratio of their churn counterparts.
2. Other than the 20-29 age group, churners in general tend to have about half the utilization ratio of non-churners.

By referring to the data, the bank could see what they could tweak for the sign-on bonuses and rewards for the corresponding groups of people to make it less rewarding to churn.

*I used a bar chart here as with these many variables, things could get messy with other types of charts. Bar chart also allows for easy comparison.*

### ***Total Transaction Amount vs Relationship Count vs Card category***

This is a 3D area chart that shows the total transaction amount of the count of relationships, divided into the card categories. Looking at the visual, we notice that there is a noticeable difference between the transaction amounts for the churners and non-churners’ card categories except for gold cards.

Even as the count of relationships increase, the transaction amount for churners and non-churners holding the gold card does not have a very noticeable difference. This means that the conclusion in my bivariate analysis earlier will have to be tweaked a bit. Previously, I had said that once the transaction amount stays below 4000, there is a higher risk of the customer churning.

After looking at this chart, we can tell that while that may be applied to the other card categories, it certainly does not apply to the gold cards as we can easily tell that the average transaction amount for churners’ gold cards are well above 4000.

*A 3D area chart was used as I was expanding on my visual for bivariate analysis. I already had a variable each in my X and Y axis, so naturally the third variable would go into the Y axis.*

# Summary

## Findings

After performing univariate, bivariate and multivariate analyses on the data, I have gained some insights and will now summarize them.

***Univariate***

1. The characteristics of churners in comparison to the rest of the customers are quite similar, with the only noticeable differences being that churners contained more females than males.
2. Only 16% of the data belongs to churners, so trying to use the features of this small subset to predict the rest of the 84% may make some insights inaccurate.

***Bivariate***

1. The average time on book for each age group increases almost linearly as the age increases.
2. Average total transaction amount for each relationship counts stays below 4000 for churners.
3. Average transaction count for each relationship counts stays below 50.
4. Average utilization ratio decreases as credit limit increases.
5. The average utilization ratio of churners in the 20-29 age group is similar to that of non-churners, but as they get older their utilization ratio decreases to below 0.2 whereas the average utilization ratio of non-churners never reaches below 0.2.

***Multivariate***

1. Looking at just utilization ratio or credit limit of a customer is not enough to predict if they will churn.
2. Churners in general tend to have about half the utilization ratio of non-churners
   1. In the 20-29 age group, single churners have about the same average utilization ratio as their non-churner counterparts.
   2. In the 20-29 age group, married churners have nearly twice the average utilization ratio of non-churners
   3. In the 20-29 age group, divorced churners have about half the average utilization ratio of non-churners.
3. In addition to point 2 of bivariate findings above, it is only applicable to card categories that are not gold. Churners with gold cards have about the same average total transaction amount as their non-churner counterparts.

I, as a bank manager, can follow-up on these findings and test them to see if they hold true. These findings may also help me to identify some of the existing customers who are planning to churn.

## Reflection

The process of completing my first Data Exploration and Analysis assignment has been a rewarding experience as a Data Science student. Owing the opportunity to this assignment, I had the chance to delve into the given dataset to extract meaningful insights. This is my reflection on the overall process and the things I had learned.

When performing univariate analysis, I was able to grasp the distribution of each variable by using visualization techniques. Throughout the process, I felt like it was very tedious, but I pushed on as I knew that if I cut corners there was a chance that my findings may be affected negatively.

The bivariate analysis phase allowed me to begin exploring the relationship between pairs of variables. I started off with a heatmap to give me leads on what charts to plot. Through bivariate analysis, I gained a deeper understanding of how each variable was related to one another and whether any correlations existed. Any significant relationships uncovered at this stage of data exploration would allow me to narrow down characteristics that would influence my findings.

During multivariate analysis, I charted pair plots for both the existing set of churned customers and colour coded them by low cardinality features in order to see if there were any obvious correlations. I then expanded on some of my bivariate visuals in order to explore my bivariate findings further. This portion of the assignment was challenging in comparison to the rest, as I had to find the right features and visual type that was able to make the data make sense. Most chart types would become unreadable when there were 3 or more variables included.

Throughout the assignment, there were times I were sure that a feature would have a relationship or correlation with another only to be met with non-meaningful visuals. This taught me that it was important to have an open mind and be unbiased when approaching the data.

In conclusion, this assignment has strengthened my understanding of the whole data exploration and analysis process from start to finish. It has provided me with valuable hands-on experience and equipped me with skills to take on future explorations.