NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

MH3511 Data Analysis With Computer Group Project

Bank Churners: Relationship Between Attrition of Credit Card Users and Other Variables

Name	Matriculation Number
Amanda Loh Wai Mun	U1940583K
Deng Jinyang	U1820375G
Neo Shun Xian Nicholas	U1820539F
Widad Binte Ahmad Sharif	U1640884B

Content Page

Abstract:

Over the last few years, consumer interest in credit cards appears to be declining, with the return on assets in the credit card business decreasing by around 40% between 2011 and 2018. (Wood, 2020) Due to this, the cost of acquiring new users has increased. Therefore, banks are now working on improving the retention rate of their current customers instead, as it costs less to maintain an existing customer than to seek out new ones. To accomplish this, banks first try to identify customers who have a high risk of being attrited. They then take appropriate measures to lower the risk, and hopefully by doing so, the customer chooses to continue utilizing their credit card services.

1. Introduction

In our project, a dataset containing the credit card attrition of a bank is analysed, with other variables such as average credit card utilisation ratio, total transaction amount, customer age, etc. Based on this dataset, we seek to answer the following popular questions around the attrition of credit card customers:

1. Does attrition depend on the customer's average credit card utilization ratio?
2. Does attrition depend on the total revolving balance of the customer's credit card?
3. Is the total credit card transaction amount and count dependent on the attrition status?
4. How does the age of the customer affect their attrition status?
5. Does the total number of products held by the customer influence attrition?

This report will cover the data descriptions and analysis using R language. For each of our research objectives, we performed statistical analysis and drew conclusions in the most appropriate approach, together with explanations and elaborations.

2. Data Description

The dataset, titled "Credit Card Customers", is obtained from the online data science community Kaggle. The original data consists of 1 csv data frame, titled "BankChurners.csv". The dataset was originally posted on https://leaps.analyttica.com, an innovative experiential data science platform which is open to the public for study and research.

Before proceeding to data analysis, we first performed a preliminary data cleaning to ensure that:

- Irrelevant columns are eliminated, e.g. "CLIENTNUM", etc;
- Rows with null entries are eliminated;
- The zero values in variables Total_Revolving_Balance and Avg_Utilisation_Ratio are excluded;

After all the preparation, 6923 observations (customers) with 18 variables were retained for analysis:

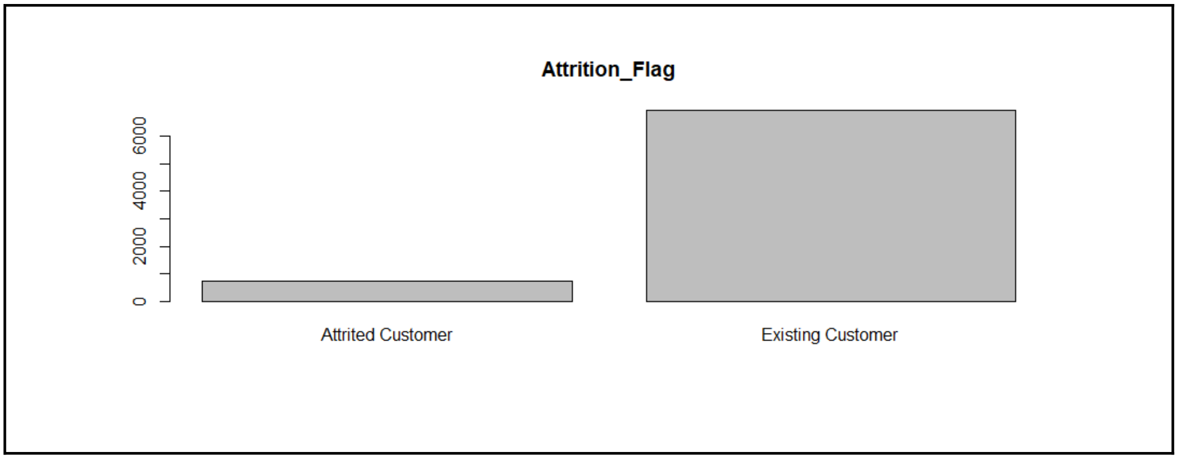
1. Customer_Age : Customer's age in years
2. Gender : Gender of the customer
3. Dependent_count : Number of dependents
4. Education_Level : Educational qualification of the account holder
5. Marital_Status : Marital status of the customer
6. Income_Category : Annual income category of the account holder
7. Card_Category : Type of card
8. Months_on_book : Period of relationship with bank
9. Total_Relationship_Count : Total no. of products held by the customer
10. Months_Inactive_12_mon : No. of months inactive in the last 12 months
11. Contacts Count_12_mon : No. of contacts in the last 12 months
12. Credit_Limit : Credit limit on the credit card
13. Total_Revolving_Bal : Total revolving balance on the credit card
14. Avg_Open_To_Buy : Open to buy credit line (average of last 12 months)
15. Total_Trans_Amt : Total transaction amount (last 12 months)
16. Total_Trans_Ct : Total transaction count (last 12 months)
17. Avg_Utilization_Ratio : Average credit card utilization ratio
18. Attrition_Flag: Attrition status of the customer (attrited or existing)

3. Description and Cleaning of Dataset

In this section, we shall look into the data in more detail. Each variable is investigated individually to look for possible outliers, and/or to perform a transformation to avoid highly skewed data.

3.1 Summary statistics for the main variable of interest, Attrition_Flag

The following plots show the overall distribution of the variable Attrition_Flag:



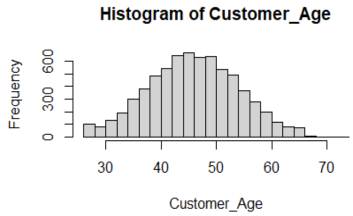
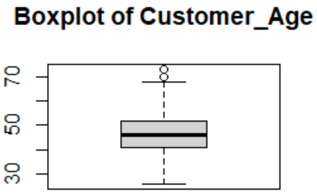
The summary of the variable Attrition_Flag is as follows:

Attrited Customers	Existing Customers
734	6923

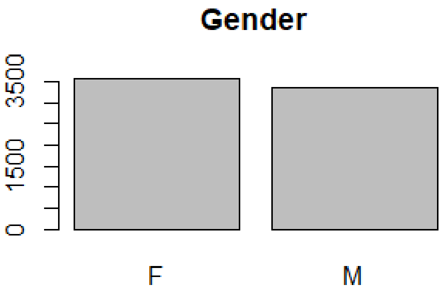
3.2 Summary statistics for other variables

The histogram, the boxplot, the transformation applied and the outliers removed from the variables are tabulated in the following subsections.

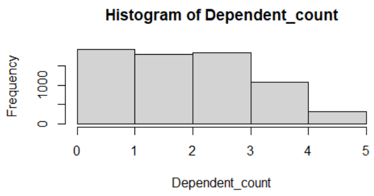
3.2.1 Customer's age in years, Customer_Age

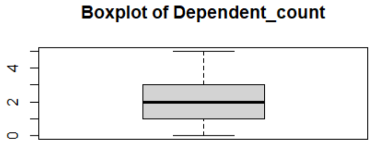
		- No outlying value of Customer_Age is removed.

3.2.2 Gender of the customer, Gender

	- No outlying value of Gender is removed. - The number of male and female customers are almost equal.

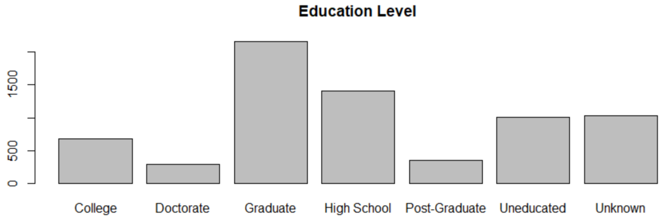
3.2.3 Number of dependents, Dependent_count





- No outlying value of Dependent_count is removed.

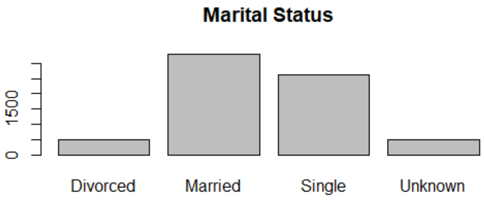
3.2.4 Educational qualification of the account holder, Education_Level



- No outlying value of Education Level is removed.

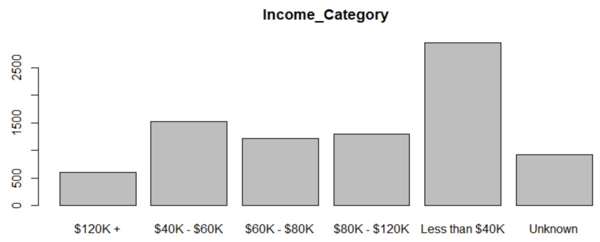
- It appears that most customers have a graduate or high-school education level.

3.2.5 Marital status of the customer, Marital_Status

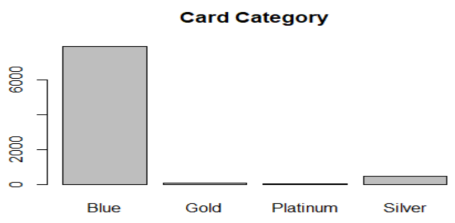


- Most customers are either married or single.
- No outlying value is removed

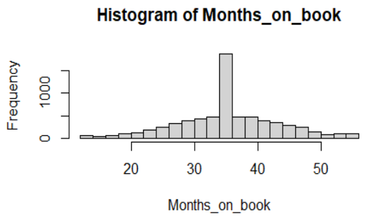
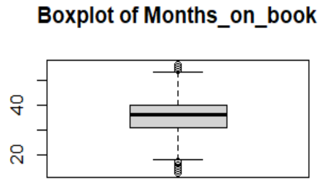
3.2.6 Annual income category of the account holder, Income_Category

	● No outlying value of Income_Category is removed. ● Majority of the customers fall under the ‘Less than $40K’ category in terms of income earned.

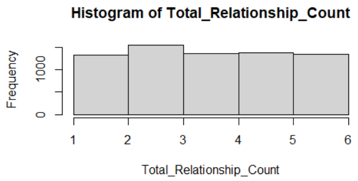
3.2.7 Type of card, Card_Category

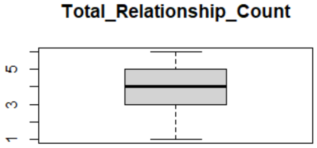
	● No outlying value is removed. ● Majority of the customers use the blue type card.

3.2.8 Period of relationship with bank, Months_on_book

		● Most of the customers have a 34-36 month relationship with the bank. ● No outlying value is removed

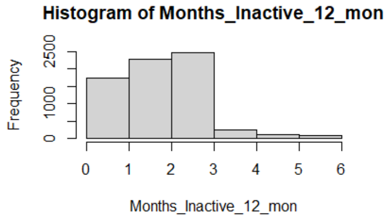
3.2.9 Total no. of products held by the customer, Total_Relationship_Count

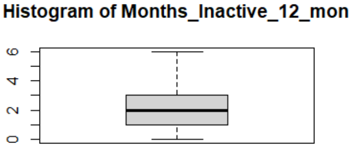




- The total no. of products held by customers is spread quite evenly.
- No outlying value is removed

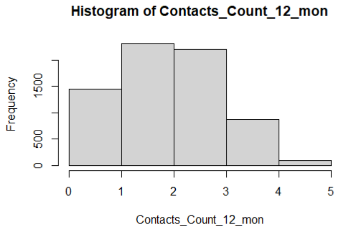
3.2.10 No. of months inactive in the last 12 months, Months_Inactive_12_mon

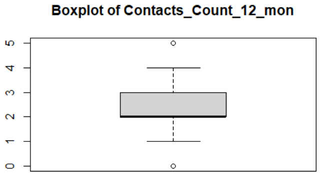




- Most customers spend at most 3 months inactive in regards to their credit card usage.
- No outlying value is removed

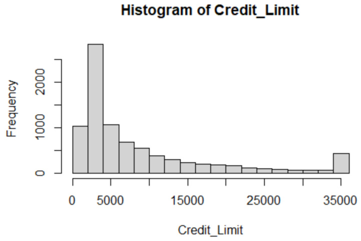
3.2.11 No. of contacts in the last 12 months, Contacts Count_12_mon

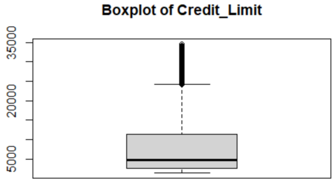




- No outlying value is removed

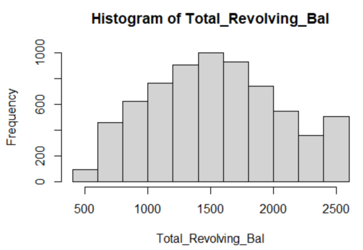
3.2.12 Credit limit on the credit card, Credit_Limit

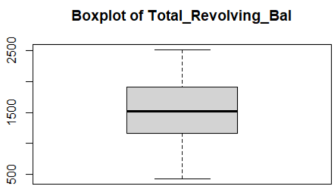




- No outlying value is removed
- The histogram of the credit limit is very right-skewed. With the majority around the 2500-6000 group.
- The log-transformation (base e) is applied

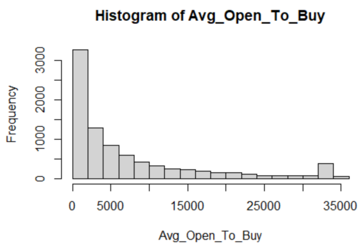
3.2.13 Total revolving balance on the credit card, Total_Revolving_Bal

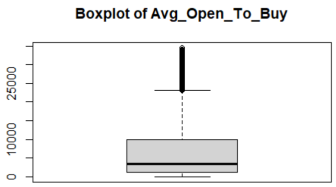




- Most customers have a total revolving balance of $1500 on their credit card.
- The zero value has been removed in the preliminary data cleaning.
- No other outlying value is removed

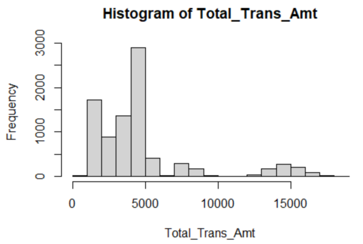
3.2.14 Open to buy credit line (average of last 12 months), Avg_Open_To_Buy

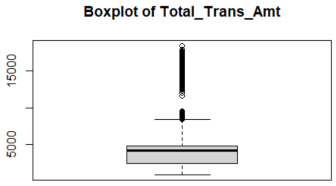




- The histogram of the average open to buy credit line is very right skewed. With the majority around the 0 to 2500 group.
- The log-transformation (base e is applied)
- No outlying value is removed.

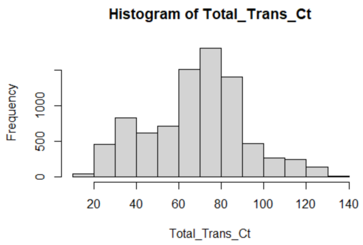
3.2.15 Total transaction amount (last 12 months), Total_Trans_Amt

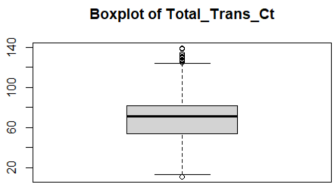




- No outlying value of Total_Trans_Amt is removed.
- The total transaction amount appears to be separated into 3 groups, with majority falling in the $0 to $6000 group.

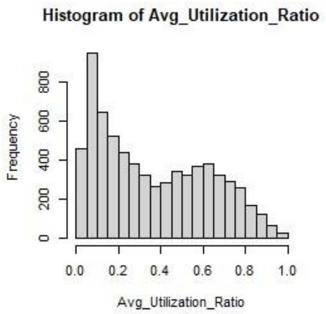
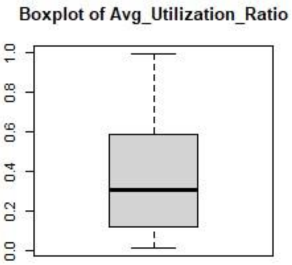
3.2.16 Total transaction count (last 12 months), Total_Trans_Ct





- No outlying value of Total_Trans_Ct is removed.
- Most customers made a total of 70 to 80 transactions in the last 12 months.

3.2.17 Average credit card utilisation ratio, Avg_Utilisation_Ratio

		● The zero value has been removed in the preliminary data cleaning. ● No other outlying value is removed ● Most of the customers average credit card utilization lie in the 0.05 to 0.15 ratio range

3.3 Final Dataset for Analysis

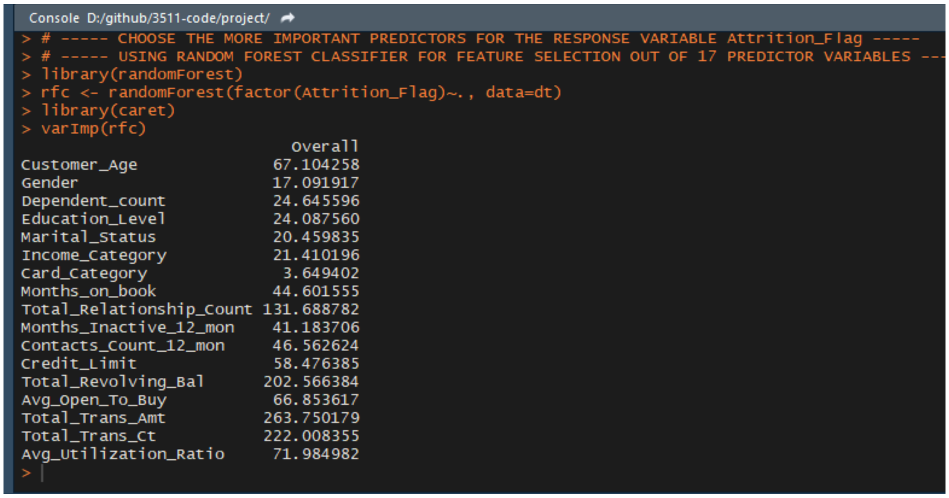
Based on the above analysis, the dataset need not be further reduced and will still be at 6923 observations with the suggested transformations. Namely, log-transformation (base e) to be applied to Credit_Limit, and Avg_Open_To_Buy.

4. Statistical Analysis

In this section we will look at the important variables to do statistical analysis.

4.1 Using Random Forest Classifier for feature selection out of the 17 predictor variables

As we have 17 predictor variables, there may be some variables that do not contribute much to the prediction of the attrition flag. As such, we use the randomForest library to build a random forest classifier model, set the attrition flag as our response variable and the rest of the variables as predictors. Then, we pass the model into the varImp class of the caret package to determine the importance value of the predictor variables against the attrition flag. The figure below shows the code snippet and the output of the importance value.



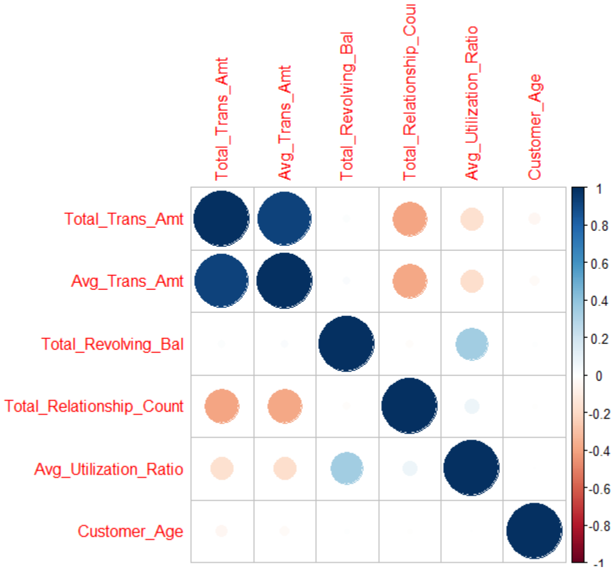
```
Console D:/github/3511-code/project/
> # ----- CHOOSE THE MORE IMPORTANT PREDICTORS FOR THE RESPONSE VARIABLE Attrition_Flag -----
> # ----- USING RANDOM FOREST CLASSIFIER FOR FEATURE SELECTION OUT OF 17 PREDICTOR VARIABLES ---
> library(randomForest)
> rfc <- randomForest(factor(Attrition_Flag)~., data=dt)
> library(caret)
> varImp(rfc)
                            Overall
Customer_Age              67.104258
Gender                    17.091917
Dependent_count           24.645596
Education_Level           24.087560
Marital_Status            20.459835
Income_Category           21.410196
Card_Category              3.649402
Months_on_book            44.601555
Total_Relationship_Count 131.688782
Months_Inactive_12_mon    41.183706
Contacts_Count_12_mon     46.562624
Credit_Limit              58.476385
Total_Revolving_Bal      202.566384
Avg_Open_To_Buy           66.853617
Total_Trans_Amt          263.750179
Total_Trans_Ct           222.008355
Avg_Utilization_Ratio     71.984982
> 
```

The larger the overall value, the more important the predictor variable is towards the attrition flag. Hence, we selected the six most important predictor variables for predictions. These six variables are:

1. Total_Trans_Amt (numerical)
2. Total_Trans_Ct (numerical)
3. Total_Revolving_Bal (numerical)
4. Total_Relationship_Count (numerical)
5. Avg_Utilization_Ratio (numerical)
6. Customer_Age (numerical)

We will proceed to do statistical analysis on the average utilization ratio based on the six predictor variables as stated above.

4.2 Correlations between all the numerical variables

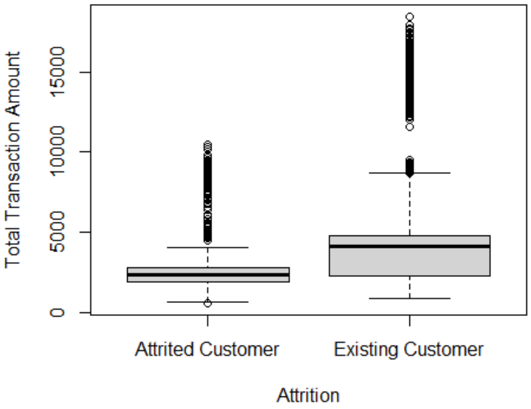


Scatter plots and correlation coefficients are useful in studying the possible linear relationships between the numerical variables. From the plot, it seems that the Avg_Trans_Amt is strongly correlated to the Total_Trans_Amt as compared to other numerical variables.

4.3 Relation between the response variable and the individual predictor variable

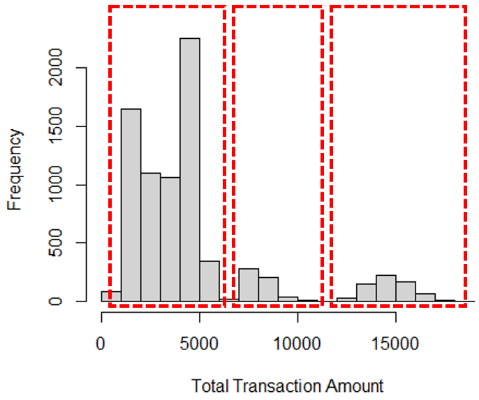
4.3.1 Relation between attrition and total transaction amount

Total transaction amount refers to the total amount that the customer has spent using the credit card. This is a steady and recurring revenue stream for any bank, as they can charge the customer a processing fee, proportionate to the transaction amount, for using the bank's payment infrastructure and network.



The boxplot above clearly shows a significant difference in the distribution of total transaction amount between attrited and existing customers.

To determine the difference in total transaction amount between attrited and existing customers, we use a t-test approach to construct a 95% confidence interval (CI). We first determine that the variances of total transaction amount are different in the two groups (Attrited and Existing), with p-value smaller than 2.2e-16. Hence, a Welch two-sample t-test is used to construct a 95% CI for the difference in mean total transaction amount: [-1723.131, -1341.483]. This means that we are 95% confident that the total transaction amount for attrited customers is on average, between $1341.48 and $1723.13, lower than that of existing customers.



Interestingly, we noted during data exploration that the total transaction amount can be classified into three major groups (see above histogram):

1. Between $0 and $6,000 (inclusive)
2. Between $6,000 (exclusive) and $12,000 (inclusive)
3. Greater than $12,000 (exclusive)

Hence, we created an additional variable named total_transaction_amount_grp using the above-mentioned logic.

By comparing the actual table against the expected table:

<Actual>	0 to 6000	6000 to 12000	> 12000
Attrited Customer	628	106	0
Existing Customer	5853	431	639

<Expected>	0 to 6000	6000 to 12000	> 12000
Attrited Customer	621	51	61
Existing Customer	5860	486	578

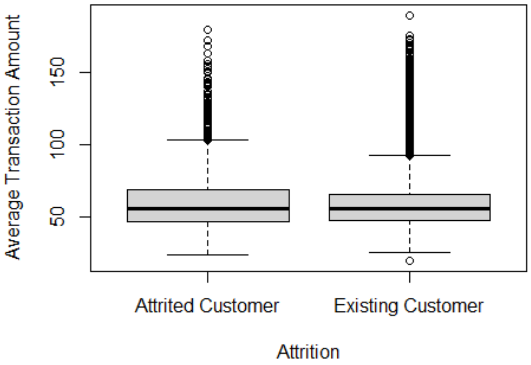
We noted that despite the fact that attrition_flag and total_transaction_amount_grp are not independent (chisq test yields a p-value of < 2.2e-16), it still does appear to be fairly independent under the “0 to 6000” category, what does appear to be the case is that attrited customers are not having huge total transaction amount (there is 0 attrited customer under the “> 12000” category in the actual table), but they are ‘compacting’ their transactions into the “6000 to 12000” category, resulting in the number of attrited customer under the “6000 to 12000” category being noticeably higher in reality than expected.

4.3.2 Relation between attrition and total transaction count

We observe a similar pattern attrition and total transaction count.

Since total transaction amount can be thought as total transaction count multiplied by average transaction amount, and we are seeing similar patterns in the first two variables (in the sense that they are different between attritied and existing customers), we want to understand if the remaining variable, average transaction amount, are similar or different between attrited and existing customers.

Hence, we decided to create a new variable named avg_trans_amt by dividing total_trans_amt by total_trans_ct to further investigate.



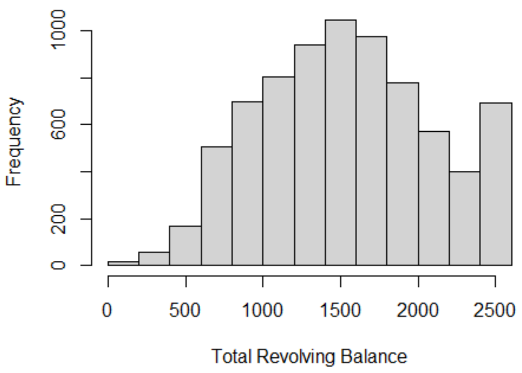
The boxplot above certainly shows similar distributions of average transaction amount between attrited and existing customers.

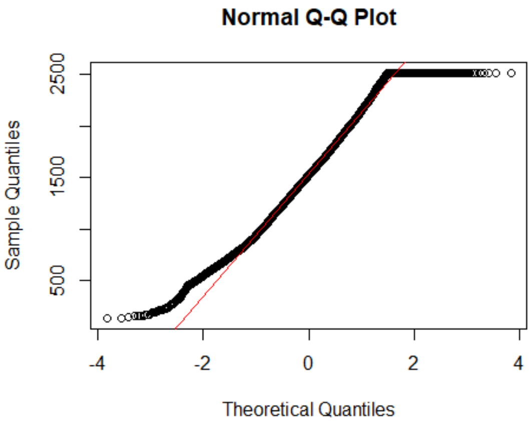
To determine the difference in total transaction amount between attrited and existing customers, we use a t-test approach to construct a 95% confidence interval (CI). We first determine that the variances of total transaction amount are different in the two groups (Attrited and Existing), with p-value equal to 0.03097, which is actually quite close to 5%. Nonetheless, a Welch two-sample t-test is used to construct a 95% CI for the difference in mean average transaction amount: [-0.4615595, 3.8509370]. This means that we are 95% confident that the average transaction amount for attrited customers is on average the same as that of existing customers, both hovering around the $60+ range.

With this in mind, we note that on average, the status of being an attrited or an existing customer does not affect the average transaction amount, but rather, it affects the total transaction count, which in turn will affect the total transaction amount.

4.3.3 Relation between attrition and total revolving balance

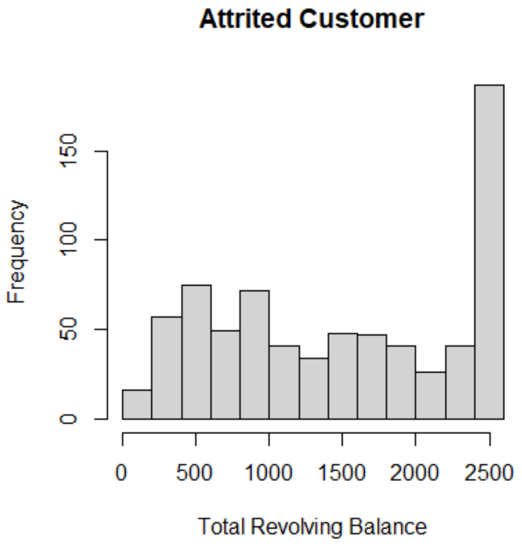
In credit card terms, the revolving balance is the portion of credit card spending that goes unpaid at the end of the month, following which the bank can charge the customer interests as costs of borrowing, till the balance is resolved in full. As compared to secured loans, credit card loans are unsecured and thus often incur extremely high interest rates. A quick search on DBS' website shows a prevailing interest rate of 26.80% per annum for credit card loans.

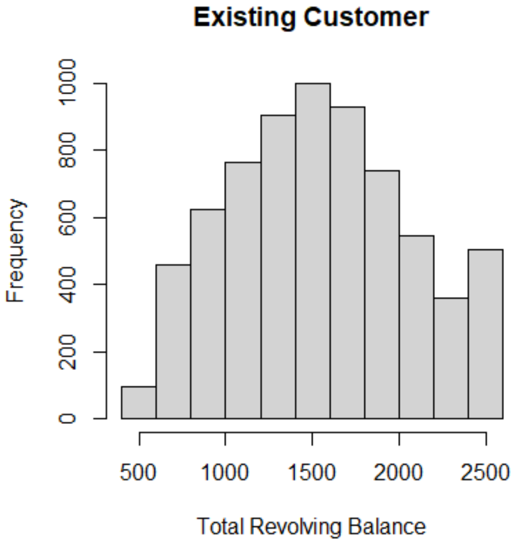




From the histogram (above, left), we see that the total revolving balance seems to resemble a normal distribution, but it does not follow one. Initially, we attempted to conduct a Shapiro-Wilk test for normality, but the sample size has exceeded 5000, which is the upper limit imposed in R. Hence, we switched to using QQPlot (above, right), where we can see that the distribution of total revolving balance has a very fat right tail.

We then examine the total revolving balance for attrited and existing customers separately:

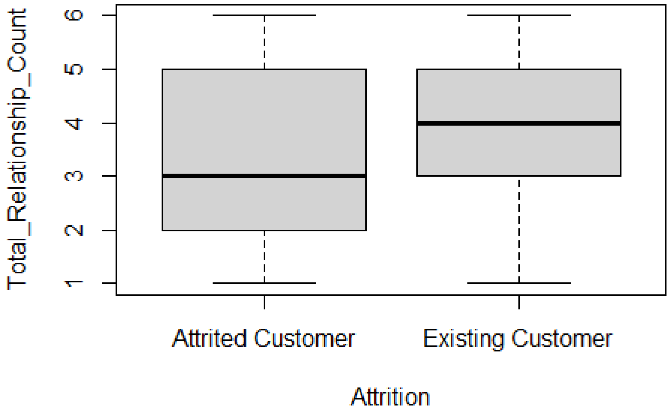




While the above two distributions may look drastically different, we can conduct a t-test if they have the same mean. We first determine that the variances of total revolving balance are different in the two groups (Attrited and Existing), with p-value < 2.2e-16. Next, a Welch two-sample t-test is used to construct a 95% CI for the difference in mean total revolving balance: [-111.579781, 8.671739]. We conclude that the average total revolving balance for attrited customers is not different from that of existing customers.

4.3.4 Relation between attrition and total relationship count

Total relationship count refers to the total number of financial products held by the customers. We are interested in looking at the relationship between attrition and the total relationship count.



The boxplot above clearly shows a significant difference in the distribution of total relationship count between attrited and existing customers.

To determine the difference in the between attrited and existing customers, we use a t-test approach to construct a 95% confidence interval (CI). We take the level of significance to be 10%. We first determine that the variances of total relationship count are different in the two groups (Attrited and Existing), with p-value smaller than 0.06824. Hence, a Welch two-sample t-test is used to construct a 95% CI for the difference in mean total relationship count: [-0.73870, -0.49345]. This means that we are 95% confident that the total relationship count for attrited customers is on average, between 0.49345 and 0.73870, lower than that of existing customers. After constructing the t-test, the p-value is smaller than 2.2x10e-16, implying that the average total relationship count is different for attrition and existing customers.

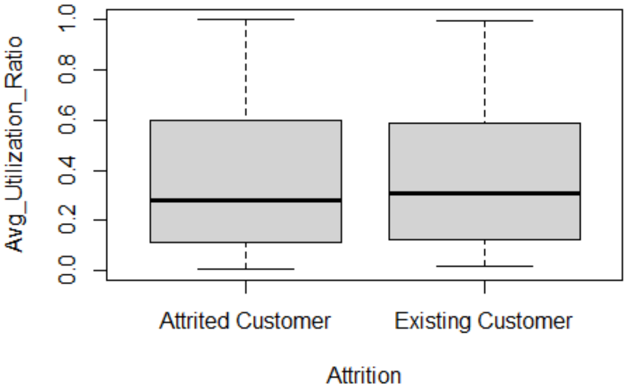
We then plot the contingency table between attrition and the total relationship count. The table is as shown below

	1	2	3	4	5	6
Attrited Customer	117	149	168	106	99	95
Existing Customer	586	732	1545	1348	1372	1340
Proportion (Attrited/Existing)	0.1996	0.2036	0.108	0.0786	0.0722	0.0708

From the table above, we can see that the proportion of the attrited customer to the existing customer is larger (about 20%) for total relationship count of 1 and 2, whereas for total relationship count of 3,4,5 and 6, the proportion is smaller (about 10%). This may imply that compared to the existing customers, there are relatively more attrited customers when the number of financial products held by the customers are low.

4.3.5 Relation between attrition and average utilization ratio

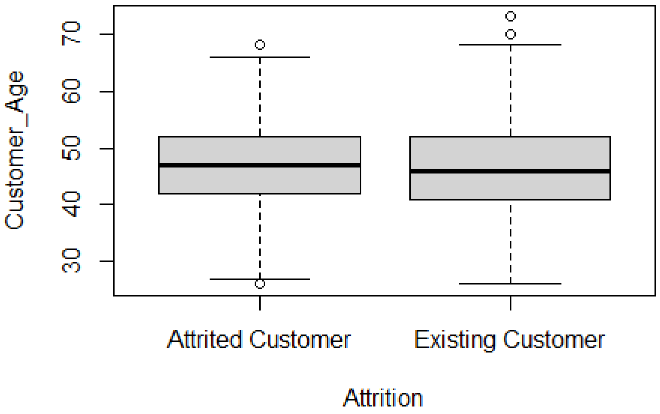
Average utilization ratio is how much the credit cards are being used by the customers. This is a metric used by the banks to detect how much the customers have used their cards for transactions, thereby giving the banks insights on how to further promote their services to the customers.



From the boxplot, we can see similar distributions of average utilization ratio between attrited and existing customers. To test whether the variance of their average utilization ratio are the same, we first have to construct a variance test to make our decision. We realised that the p-value is 1.503e-05, which is a very small value. At 5% level of significance, we conclude that the variances of average utilization ratio among attrited and existing customers are different.

To determine the difference in total transaction amount between attrited and existing customers, we use a t-test approach to construct a 95% confidence interval (CI). A Welch two-sample t-test is used to construct a 95% CI for the difference in the average utilization ratio: [-0.025630, 0.01806]. The p-value is 0.7338 which implies that there is insufficient evidence that the average utilization ratio is different for the attrited customers and the existing customers.

4.3.6 Relation between attrition and customer age



From the boxplot, we can see similar distributions of the age between attrited and existing customers. To test whether the variances of the age are the same, we first construct a variance test to decide. We realised that the p-value is 0.01923, which is a small value. At 5% level of significance, we conclude that the variance of the average age is different for attrited and existing customers.

To determine the difference in the age between attrited and existing customers, we use a t-test approach to construct a 95% confidence interval (CI). A Welch two-sample t-test is used to construct a 95% CI for the difference in the average utilization ratio: [-0.10162, 1.06036]. The p-value is 0.1056 which is larger than the significance value of 0.10, implying that there is insufficient evidence that the average age is different for the attrited customers and the existing customers. As such, we further classify the age into different age groups to see if there are any potential trends.

<Actual>	Below 30	30 to 39	40 to 49	50 to 59	Above 60
Attrited Customer	14	109	342	239	30
Existing Customer	184	1241	3050	2074	374

<Expected>	Below 30	30 to 39	40 to 49	50 to 59	Above 60
Attrited Customer	19	129	325	222	39
Existing Customer	179	1221	3067	2091	365

From the two tables above, attrition_flag and age group are not independent (chisq test yields a p-value of 0.04705) at the 5% level of significance. The proportion of attrited customers against the existing customer is similar in both the expectation and the actual table.

5. Conclusion and Discussion

Customers are the most important bedrock to every business. In the banking industry, retaining customers is more profitable than building new relationships, as attracting a new customer is about five times more costly than retaining an existing customer. Many banks have begun to realise the importance of this, and have attempted to lower customer attrition by paying greater attention to Customer Relationship Management. In our project, we studied the relationship between customer attrition and the their underlying characteristics and behavioral patterns, and made the following observations:

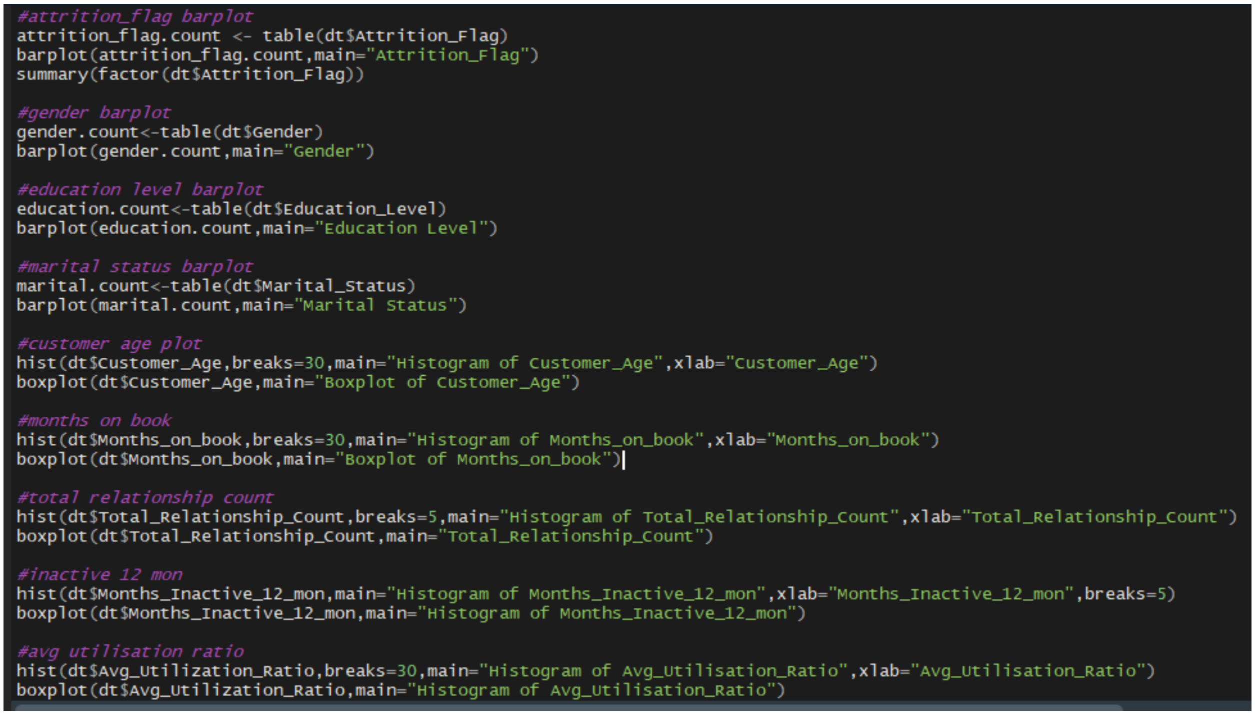
- The total transaction amount for attrited customers is on average, between $1341.48 and $1723.13, lower than that of existing customers
- This is not because attrited customers are spending less per transaction (their average transaction amount are actually the same), but rather because they participate in lesser transactions than existing customers
- Average total revolving balance for attrited customers is not different from that of existing customers
- When the customer relationship is not very strong (i.e. the customer is only holding on to a few financial products), these customers are more likely to churn.
- The average utilization ratio is different for the attrited customers as compared to the existing customers.
- When classifying the age into different age groups, the attrited/existing customers and the age groups are not independent of one another

Additionally, we note that between the numerical variables, there is a strong positive linear correlation between Total_Trans_Amt and Avg_Trans_Amt, but does not seem to have a strong linear correlation with the other numerical variables.

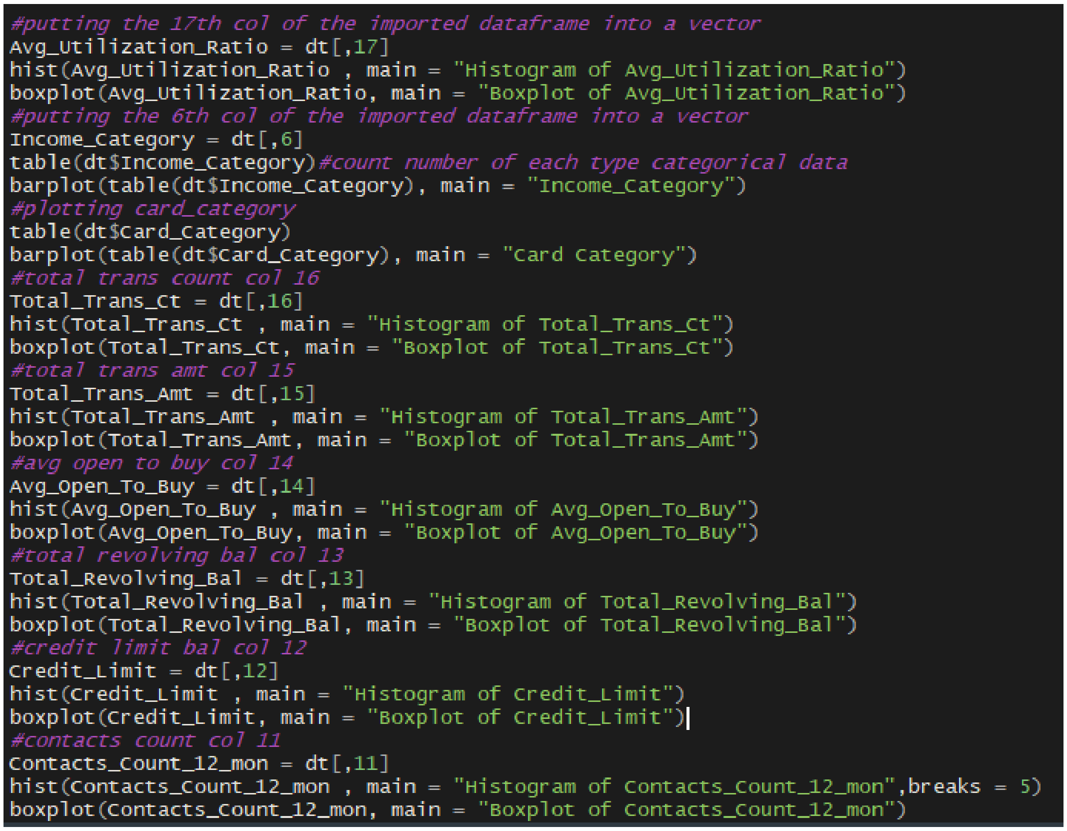
Although the content of this report is interesting and the insights definitely promising, it should be noted that this report is only based on one single source of data published on the internet. Furthermore, with the advancement of data mining / behavioral tracking techniques, many banks have been able to capture much more sophisticated relationships than what we have considered. Last but not least, it should not be forgotten that maintaining customer relationships is both an art and a science, and thus insights on customer attrition should be combined with targeted strategies that seek to generate and deliver long-term value and convenience for the customer.

6. Appendix

3. Description and Cleaning of Dataset



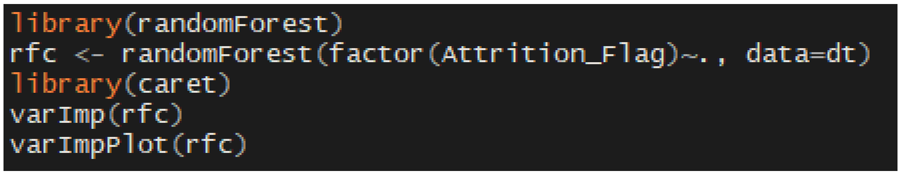
```
#attrition_flag barplot
attrition_flag.count <- table(dt$Attrition_Flag)
barplot(attrition_flag.count,main="Attrition_Flag")
summary(factor(dt$Attrition_Flag))

#gender barplot
gender.count<-table(dt$Gender)
barplot(gender.count,main="Gender")

#education level barplot
education.count<-table(dt$Education_Level)
barplot(education.count,main="Education Level")

#marital status barplot
marital.count<-table(dt$Marital_Status)
barplot(marital.count,main="Marital Status")

#customer age plot
hist(dt$Customer_Age,breaks=30,main="Histogram of Customer_Age",xlab="Customer_Age")
boxplot(dt$Customer_Age,main="Boxplot of Customer_Age")

#months on book
hist(dt$Months_on_book,breaks=30,main="Histogram of Months_on_book",xlab="Months_on_book")
boxplot(dt$Months_on_book,main="Boxplot of Months_on_book")

#total relationship count
hist(dt$Total_Relationship_Count,breaks=5,main="Histogram of Total_Relationship_Count",xlab="Total_Relationship_Count")
boxplot(dt$Total_Relationship_Count,main="Total_Relationship_Count")

#inactive 12 mon
hist(dt$Months_Inactive_12_mon,main="Histogram of Months_Inactive_12_mon",xlab="Months_Inactive_12_mon",breaks=5)
boxplot(dt$Months_Inactive_12_mon,main="Histogram of Months_Inactive_12_mon")

#avg utilisation ratio
hist(dt$Avg_Utilization_Ratio,breaks=30,main="Histogram of Avg_Utilisation_Ratio",xlab="Avg_Utilisation_Ratio")
boxplot(dt$Avg_Utilization_Ratio,main="Histogram of Avg_Utilisation_Ratio")
```



```
#putting the 17th col of the imported dataframe into a vector
Avg_Utilization_Ratio = dt[,17]
hist(Avg_Utilization_Ratio , main = "Histogram of Avg_Utilization_Ratio")
boxplot(Avg_Utilization_Ratio, main = "Boxplot of Avg_Utilization_Ratio")
#putting the 6th col of the imported dataframe into a vector
Income_Category = dt[,6]
table(dt$Income_Category)#count number of each type categorical data
barplot(table(dt$Income_Category), main = "Income_Category")
#plotting card_category
table(dt$Card_Category)
barplot(table(dt$Card_Category), main = "Card Category")
#total trans count col 16
Total_Trans_Ct = dt[,16]
hist(Total_Trans_Ct , main = "Histogram of Total_Trans_Ct")
boxplot(Total_Trans_Ct, main = "Boxplot of Total_Trans_Ct")
#total trans amt col 15
Total_Trans_Amt = dt[,15]
hist(Total_Trans_Amt , main = "Histogram of Total_Trans_Amt")
boxplot(Total_Trans_Amt, main = "Boxplot of Total_Trans_Amt")
#avg open to buy col 14
Avg_Open_To_Buy = dt[,14]
hist(Avg_Open_To_Buy , main = "Histogram of Avg_Open_To_Buy")
boxplot(Avg_Open_To_Buy, main = "Boxplot of Avg_Open_To_Buy")
#total revolving bal col 13
Total_Revolving_Bal = dt[,13]
hist(Total_Revolving_Bal , main = "Histogram of Total_Revolving_Bal")
boxplot(Total_Revolving_Bal, main = "Boxplot of Total_Revolving_Bal")
#credit limit bal col 12
Credit_Limit = dt[,12]
hist(Credit_Limit , main = "Histogram of Credit_Limit")
boxplot(Credit_Limit, main = "Boxplot of Credit_Limit")
#contacts count col 11
Contacts_Count_12_mon = dt[,11]
hist(Contacts_Count_12_mon , main = "Histogram of Contacts_Count_12_mon",breaks = 5)
boxplot(Contacts_Count_12_mon, main = "Boxplot of Contacts_Count_12_mon")
```

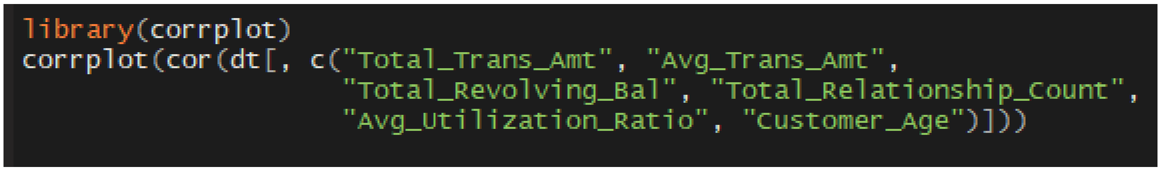
4. Statistical Analysis

4.1 Using Random Forest Classifier for feature selection out of the 17 predictor variables



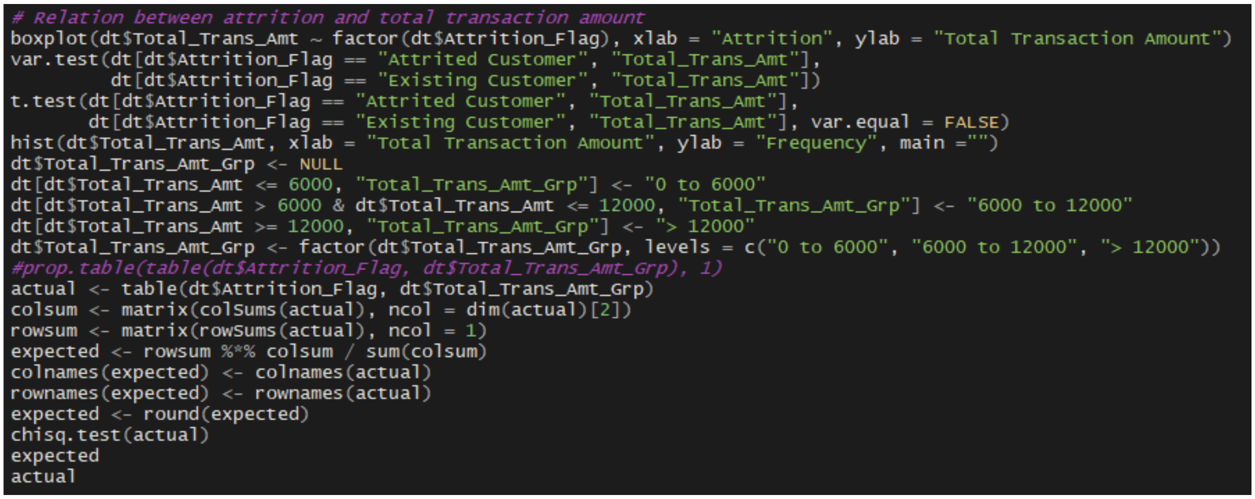
```
library(randomForest)
rfc <- randomForest(factor(Attrition_Flag)~., data=dt)
library(caret)
varImp(rfc)
varImpPlot(rfc)
```

4.2 Correlations between all the numerical variables



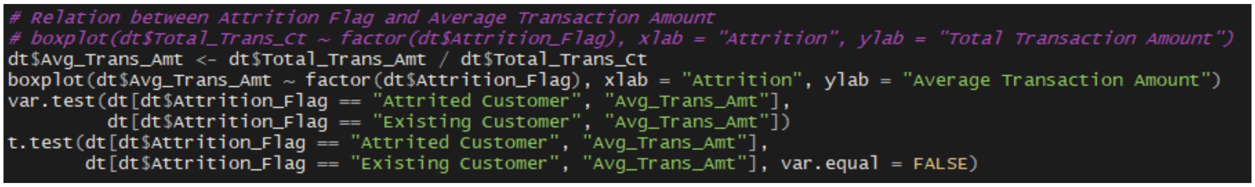
```
library(corrplot)
corrplot(cor(dt[, c("Total_Trans_Amt", "Avg_Trans_Amt",
                    "Total_Revolving_Bal", "Total_Relationship_Count",
                    "Avg_Utilization_Ratio", "Customer_Age")]))
```

4.3.1 Relation between attrition and total transaction amount



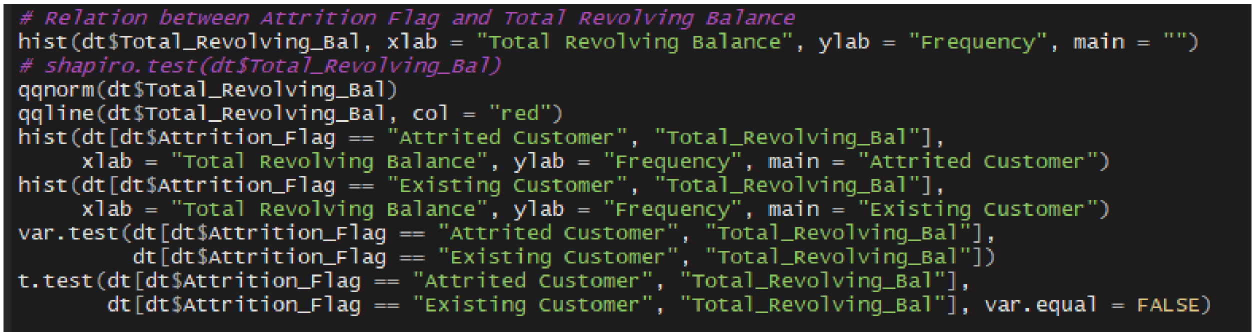
```
# Relation between attrition and total transaction amount
boxplot(dt$Total_Trans_Amt ~ factor(dt$Attrition_Flag), xlab = "Attrition", ylab = "Total Transaction Amount")
var.test(dt[dt$Attrition_Flag == "Attrited Customer", "Total_Trans_Amt"],
         dt[dt$Attrition_Flag == "Existing Customer", "Total_Trans_Amt"])
t.test(dt[dt$Attrition_Flag == "Attrited Customer", "Total_Trans_Amt"],
       dt[dt$Attrition_Flag == "Existing Customer", "Total_Trans_Amt"], var.equal = FALSE)
hist(dt$Total_Trans_Amt, xlab = "Total Transaction Amount", ylab = "Frequency", main ="")
dt$Total_Trans_Amt_Grp <- NULL
dt[dt$Total_Trans_Amt <= 6000, "Total_Trans_Amt_Grp"] <- "0 to 6000"
dt[dt$Total_Trans_Amt > 6000 & dt$Total_Trans_Amt <= 12000, "Total_Trans_Amt_Grp"] <- "6000 to 12000"
dt[dt$Total_Trans_Amt >= 12000, "Total_Trans_Amt_Grp"] <- "> 12000"
dt$Total_Trans_Amt_Grp <- factor(dt$Total_Trans_Amt_Grp, levels = c("0 to 6000", "6000 to 12000", "> 12000"))
#prop.table(table(dt$Attrition_Flag, dt$Total_Trans_Amt_Grp), 1)
actual <- table(dt$Attrition_Flag, dt$Total_Trans_Amt_Grp)
colsum <- matrix(colSums(actual), ncol = dim(actual)[2])
rowsum <- matrix(rowSums(actual), ncol = 1)
expected <- rowsum %*% colsum / sum(colsum)
colnames(expected) <- colnames(actual)
rownames(expected) <- rownames(actual)
expected <- round(expected)
chisq.test(actual)
expected
actual
```

4.3.2 Relation between attrition and total transaction count



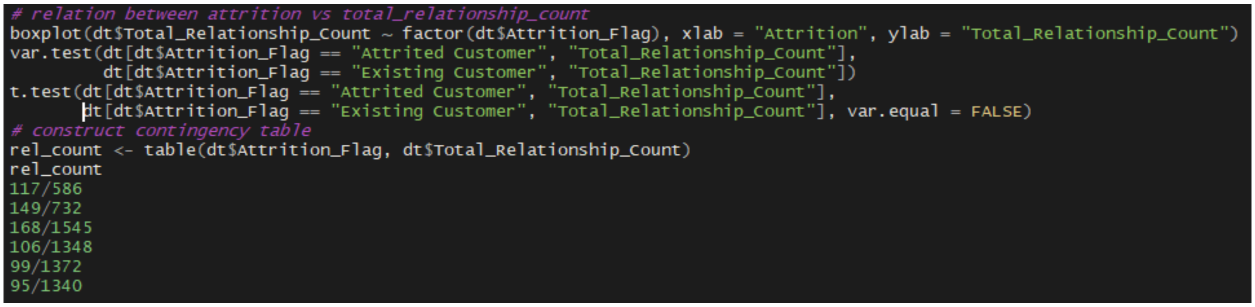
```
# Relation between Attrition Flag and Average Transaction Amount
# boxplot(dt$Total_Trans_Ct ~ factor(dt$Attrition_Flag), xlab = "Attrition", ylab = "Total Transaction Amount")
dt$Avg_Trans_Amt <- dt$Total_Trans_Amt / dt$Total_Trans_Ct
boxplot(dt$Avg_Trans_Amt ~ factor(dt$Attrition_Flag), xlab = "Attrition", ylab = "Average Transaction Amount")
var.test(dt[dt$Attrition_Flag == "Attrited Customer", "Avg_Trans_Amt"],
         dt[dt$Attrition_Flag == "Existing Customer", "Avg_Trans_Amt"])
t.test(dt[dt$Attrition_Flag == "Attrited Customer", "Avg_Trans_Amt"],
       dt[dt$Attrition_Flag == "Existing Customer", "Avg_Trans_Amt"], var.equal = FALSE)
```

4.3.3 Relation between attrition and total revolving balance



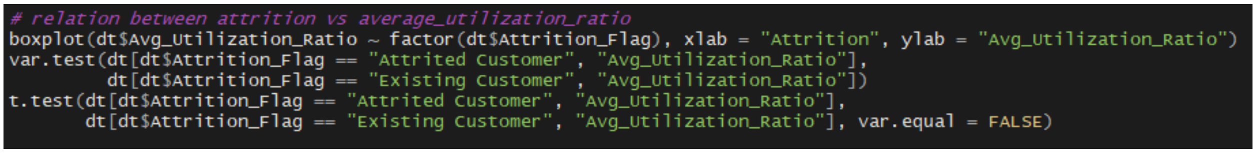
```
# Relation between Attrition Flag and Total Revolving Balance
hist(dt$Total_Revolving_Bal, xlab = "Total Revolving Balance", ylab = "Frequency", main = "")
# shapiro.test(dt$Total_Revolving_Bal)
qqnorm(dt$Total_Revolving_Bal)
qqline(dt$Total_Revolving_Bal, col = "red")
hist(dt[dt$Attrition_Flag == "Attrited Customer", "Total_Revolving_Bal"],
     xlab = "Total Revolving Balance", ylab = "Frequency", main = "Attrited Customer")
hist(dt[dt$Attrition_Flag == "Existing Customer", "Total_Revolving_Bal"],
     xlab = "Total Revolving Balance", ylab = "Frequency", main = "Existing Customer")
var.test(dt[dt$Attrition_Flag == "Attrited Customer", "Total_Revolving_Bal"],
         dt[dt$Attrition_Flag == "Existing Customer", "Total_Revolving_Bal"])
t.test(dt[dt$Attrition_Flag == "Attrited Customer", "Total_Revolving_Bal"],
       dt[dt$Attrition_Flag == "Existing Customer", "Total_Revolving_Bal"], var.equal = FALSE)
```

4.3.4 Relation between attrition and total relationship count



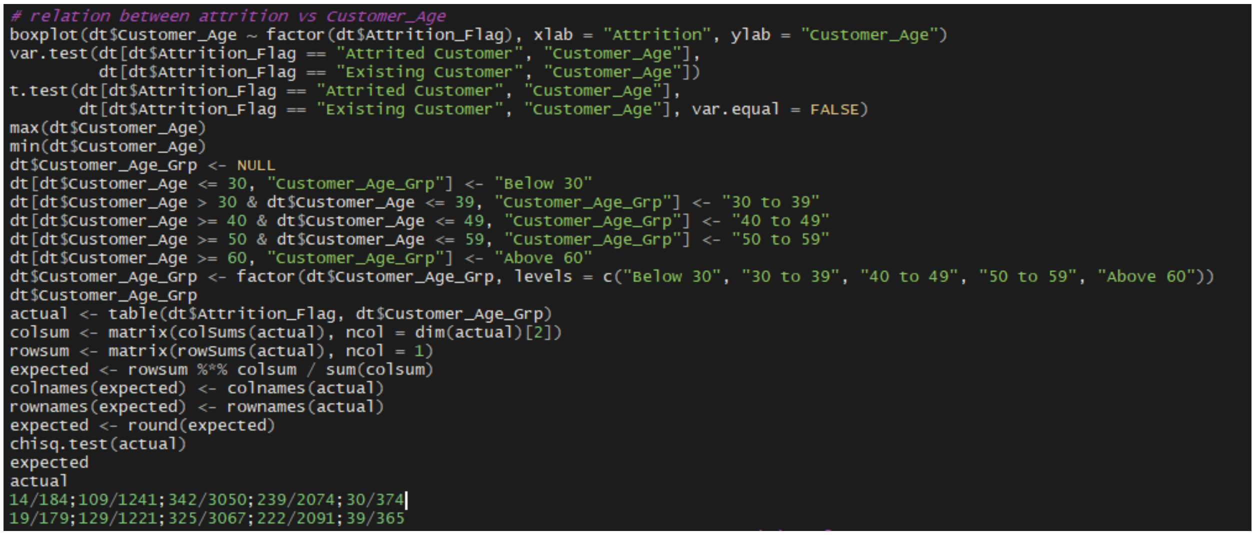
```
# relation between attrition vs total_relationship_count
boxplot(dt$Total_Relationship_Count ~ factor(dt$Attrition_Flag), xlab = "Attrition", ylab = "Total_Relationship_Count")
var.test(dt[dt$Attrition_Flag == "Attrited Customer", "Total_Relationship_Count"],
         dt[dt$Attrition_Flag == "Existing Customer", "Total_Relationship_Count"])
t.test(dt[dt$Attrition_Flag == "Attrited Customer", "Total_Relationship_Count"],
       dt[dt$Attrition_Flag == "Existing Customer", "Total_Relationship_Count"], var.equal = FALSE)
# construct contingency table
rel_count <- table(dt$Attrition_Flag, dt$Total_Relationship_Count)
rel_count
117/586
149/732
168/1545
106/1348
99/1372
95/1340
```

4.3.5 Relation between attrition and average utilization ratio



```
# relation between attrition vs average_utilization_ratio
boxplot(dt$Avg_Utilization_Ratio ~ factor(dt$Attrition_Flag), xlab = "Attrition", ylab = "Avg_Utilization_Ratio")
var.test(dt[dt$Attrition_Flag == "Attrited Customer", "Avg_Utilization_Ratio"],
         dt[dt$Attrition_Flag == "Existing Customer", "Avg_Utilization_Ratio"])
t.test(dt[dt$Attrition_Flag == "Attrited Customer", "Avg_Utilization_Ratio"],
       dt[dt$Attrition_Flag == "Existing Customer", "Avg_Utilization_Ratio"], var.equal = FALSE)
```

4.3.6 Relation between attrition and customer age



```
# relation between attrition vs Customer_Age
boxplot(dt$Customer_Age ~ factor(dt$Attrition_Flag), xlab = "Attrition", ylab = "Customer_Age")
var.test(dt[dt$Attrition_Flag == "Attrited Customer", "Customer_Age"],
         dt[dt$Attrition_Flag == "Existing Customer", "Customer_Age"])
t.test(dt[dt$Attrition_Flag == "Attrited Customer", "Customer_Age"],
       dt[dt$Attrition_Flag == "Existing Customer", "Customer_Age"], var.equal = FALSE)
max(dt$Customer_Age)
min(dt$Customer_Age)
dt$Customer_Age_Grp <- NULL
dt[dt$Customer_Age <= 30, "Customer_Age_Grp"] <- "Below 30"
dt[dt$Customer_Age > 30 & dt$Customer_Age <= 39, "Customer_Age_Grp"] <- "30 to 39"
dt[dt$Customer_Age >= 40 & dt$Customer_Age <= 49, "Customer_Age_Grp"] <- "40 to 49"
dt[dt$Customer_Age >= 50 & dt$Customer_Age <= 59, "Customer_Age_Grp"] <- "50 to 59"
dt[dt$Customer_Age >= 60, "Customer_Age_Grp"] <- "Above 60"
dt$Customer_Age_Grp <- factor(dt$Customer_Age_Grp, levels = c("Below 30", "30 to 39", "40 to 49", "50 to 59", "Above 60"))
dt$Customer_Age_Grp
actual <- table(dt$Attrition_Flag, dt$Customer_Age_Grp)
colsum <- matrix(colSums(actual), ncol = dim(actual)[2])
rowsum <- matrix(rowSums(actual), ncol = 1)
expected <- rowsum %*% colsum / sum(colsum)
colnames(expected) <- colnames(actual)
rownames(expected) <- rownames(actual)
expected <- round(expected)
chisq.test(actual)
expected
actual
14/184;109/1241;342/3050;239/2074;30/374
19/179;129/1221;325/3067;222/2091;39/365
```

7. References

Goyal S. (2020, December). *Credit Card customers.* Kaggle. https://www.kaggle.com/sakshigoyal7/credit-card-customers

Wood, J. (2020, August 10). *Research note: Consumer interest in the credit card is declining.* Payments Cards & Mobile. https://www.paymentscardsandmobile.com/research-note-consumer-interest-in-the-credit-card-is-declining/