AMS 572 DATA ANALYSIS I

Course Project Group 12

M V D Satya Swaroop, Akshay Kurup, Sameer Kyalkond, Bipasha Ray

1.Introduction

This project analyzes factors for Attrition in Credit Card Customers. The Credit Card Customers Prediction Dataset measures various variables regarding Customer attributes such as Credit limit, Education Level, Income Level, Transaction Amount, etc which includes around 22 variables with both continuous and categorical values. Initially, exploratory data analysis was conducted and we arrived with two questions of Interest.
Hypothesis tests were then conducted to test :

1) Effect of Credit limit on Attrition Flag of Customer.
2) Which factors impact Attrition.

Wilcoxon Signed-Rank Test and Multiple Logistic Regression were used to test the aforementioned hypotheses respectively.

Data Definition:

➔ **Numerical Variables:**
Personal Information: Customer_Age
Income : Credit_limit, TotalRevolvingBalance, AvgOpenTo_Buy, Total_Amt_Chng_Q4_Q1, Total_Trans_Ct Total_Trans_Amt, Total_Ct_Chn_Q4_Q1, Avg_Utilisation_Ratio.
Miscellaneous:
Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1,
Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2

➔ **Categorical Variables:**
Personal Information: Gender, Dependent_count, Education_Level, Marital Status.
Income: Income_Category, Card_Category, Months_on_Book
Miscellaneous:
Total_Relationship_count, Months_Inactive_12_mon, Contacts_Count_12_mon, Attrition_Flag

2.Exploratory Data Analysis:

We start our project by performing initial investigations on our data so as to spot anomalies and discover patterns which will help us test our hypothesis and also check assumptions with the help of summary statistics and graphical representations.In this project, we will use the R language and environment to do statistical computing and graphics work.

Distinct Values in Categorical Variables:

a. Income_Category

```
> (distinct(data, Income_Category))
  Income_Category
1     $60K - $80K
2  Less than $40K
3    $80K - $120K
4     $40K - $60K
5         $120K +
6         Unknown
```

b. Marital_Status

```
> (distinct(data, Marital_Status))
  Marital_Status
1        Married
2         Single
3        Unknown
4       Divorced
```

c. Education_Level

```
> (distinct(data, Education_Level))
  Education_Level
1     High School
2        Graduate
3      Uneducated
4         Unknown
5         College
6   Post-Graduate
7       Doctorate
```

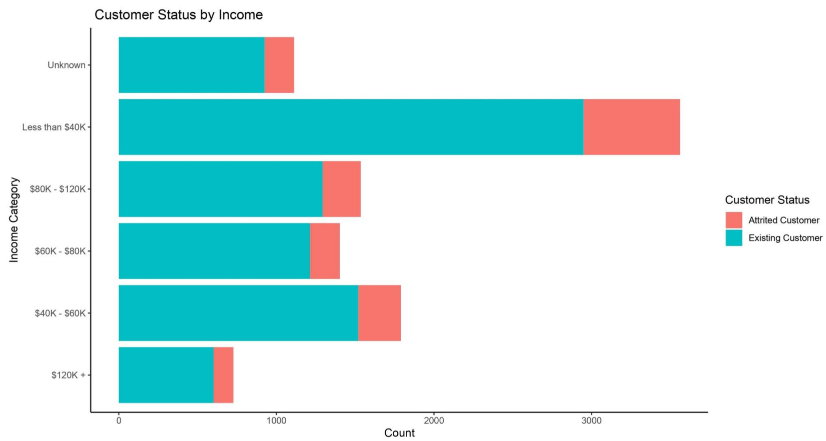
d. Card_Category

```
> (distinct(data, Card_Category))
  Card_Category
1          Blue
2          Gold
3        Silver
4      Platinum
```

Distribution of Categorical Variables Based on Attrition_Flag

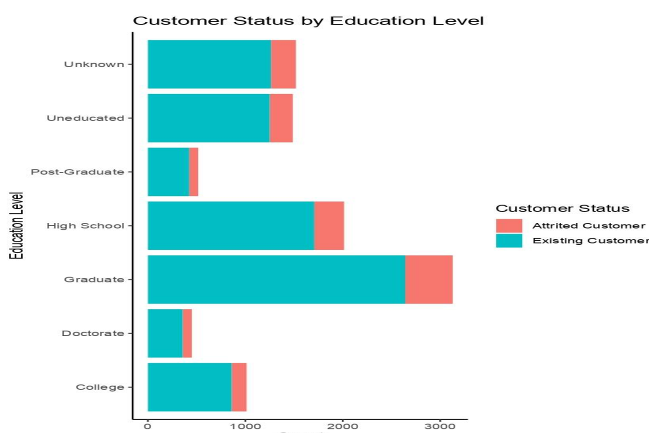
- Income_Category

```
> ggplot(data , aes(y = Income_Category)) +
+  geom_bar(aes(fill = Attrition_Flag), position = position_stack(reverse =
FALSE)) +theme(legend.position = "top") + theme_classic() + xlab("Count") +
ylab("Income Category") + ggtitle(" Customer Status by Income" )+  labs(fill
= "Customer Status")
```



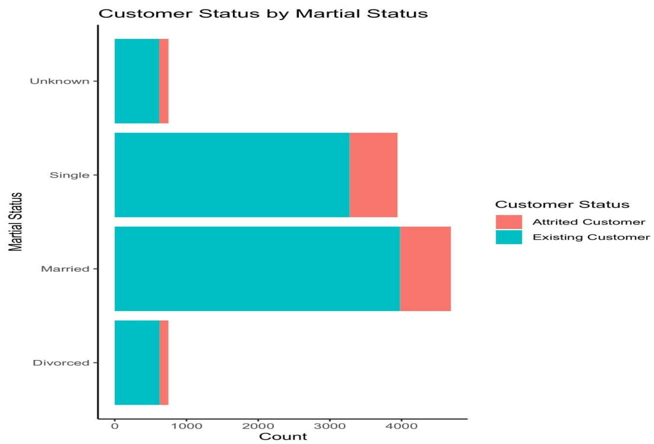
- Education_Level

```
> ggplot(data , aes(y = Education_Level)) +
+  geom_bar(aes(fill = Attrition_Flag), position = position_stack(reverse =
FALSE)) +
+  theme(legend.position = "top") + theme_classic() + xlab("Count") +
ylab("Education Level") + ggtitle("Customer Status by Education Level" ) +
labs(fill = "Customer Status")
```



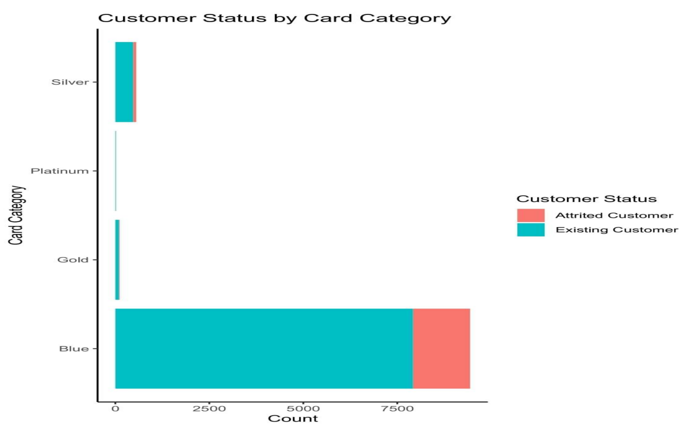
- Marital_Status

```
> ggplot(data , aes(y = Marital_Status)) +
+  geom_bar(aes(fill = Attrition_Flag), position = position_stack(reverse =
FALSE)) +
+  theme(legend.position = "top") + theme_classic() + xlab("Count") +
ylab("Martial Status") + ggtitle("Customer Status by Martial Status" )+
labs(fill = "Customer Status")
```



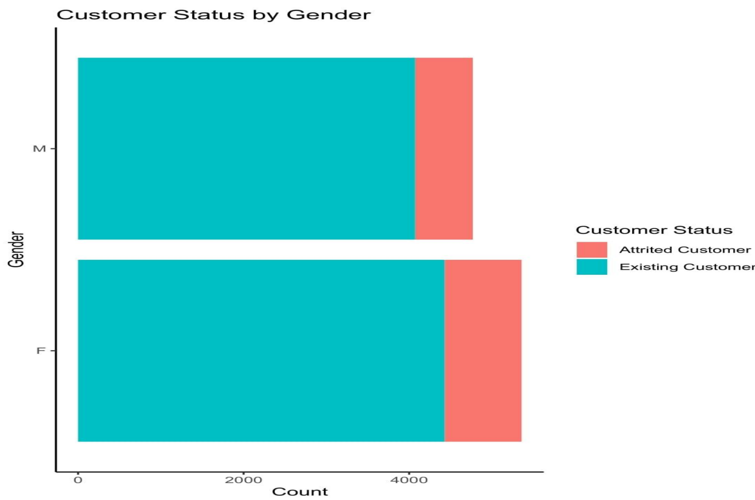
- Card_Category

```
> ggplot(data , aes(y = Card_Category)) +
+  geom_bar(aes(fill =  Attrition_Flag), position = position_stack(reverse =
FALSE)) +
+  theme(legend.position = "top") + theme_classic() + xlab("Count") +
ylab("Card Category") + ggtitle("Customer Status by Card Category" )+
labs(fill = "Customer Status")
```



- Gender

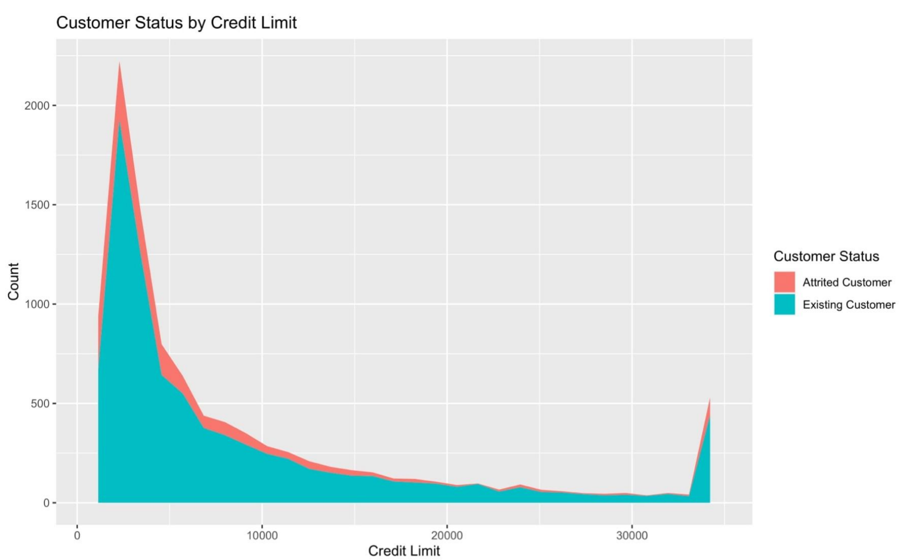
```
> ggplot(data , aes(y = Gender)) +
+  geom_bar(aes(fill =  Attrition_Flag), position = position_stack(reverse =
FALSE)) +
+  theme(legend.position = "top") + theme_classic() + xlab("Count") +
ylab("Gender") + ggtitle("Customer Status by Gender" )+  labs(fill =
"Customer Status")
```



Distribution of Continuous Variables Based on Attrition_Flag:

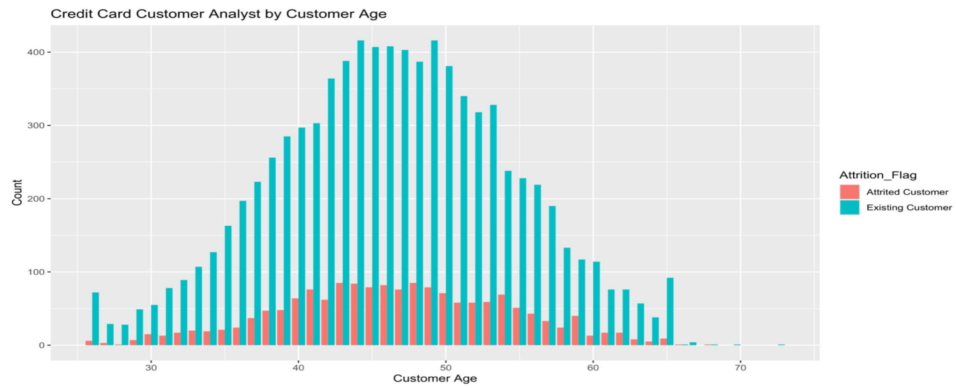
➢ Credit_Limit

```
> ggplot(data, aes(x=Credit_Limit, fill=Attrition_Flag)) +
+   geom_area(stat ="bin") + xlab("Credit Limit")+ylab("Count")
+ggtitle("Customer Status by Credit Limit " ) +  labs(fill = "Customer
Status")
```



➢ Customer_Age

```
> l6<-ggplot(data = data,
+             aes(x = Customer_Age,
+                 fill = Attrition_Flag)) +
+   geom_bar(position = position_dodge(preserve = "single"))+
+   labs(x = "Customer Age",
+        y = "Count", title = "Credit Card Customer Analyst by Customer
Age")
> l6
```



Summary Statistics on Data:

- Mean Values based on Attrition_Flag

```
> (data  %>% group_by(Attrition_Flag) %>% summarize(meanAge=
mean(Customer_Age), meanDepdent= mean(Dependent_count), meanCreditLim=
mean(Credit_Limit)))
# A tibble: 2 x 4
  Attrition_Flag    meanAge meanDepdent meanCreditLim
  <chr>               <dbl>       <dbl>         <dbl>
1 Attrited Customer    46.7        2.40         8136.
2 Existing Customer    46.3        2.34         8727.
```

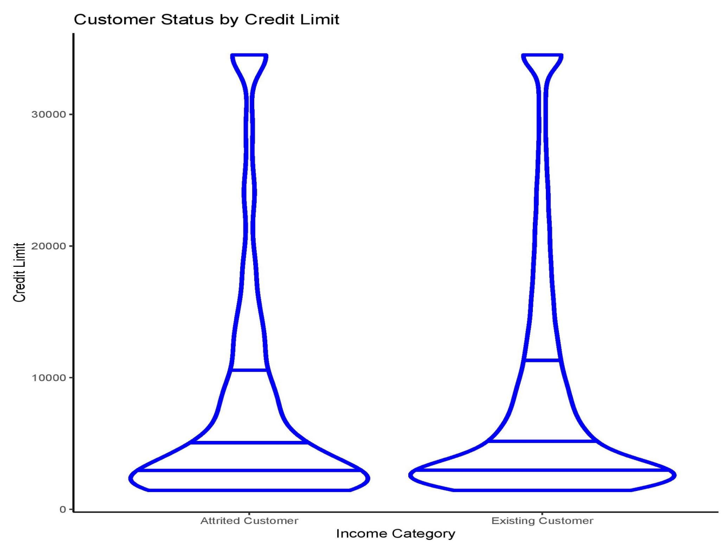
- Summary of All Numeric Variables

```
> summary(numericData)
  Customer_Age   Dependent_count 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
 Min.   :26.00   Min.   :0.000   Min.   :13.00   Min.   :1.000           Min.   :0.000          Min.   :0.000         Min.   : 1438   Min.   :   0
Min.   :    3   Min.   :0.0000      Min.   :  510
 1st Qu.:41.00   1st Qu.:1.000   1st Qu.:31.00   1st Qu.:3.000           1st Qu.:2.000          1st Qu.:2.000         1st Qu.: 2555   1st Qu.: 359
1st Qu.: 1324   1st Qu.:0.6310      1st Qu.: 2156
 Median :46.00   Median :2.000   Median :36.00   Median :4.000           Median :2.000          Median :2.000         Median : 4549   Median :1276
Median : 3474   Median :0.7360      Median : 3899
 Mean   :46.33   Mean   :2.346   Mean   :35.93   Mean   :3.813           Mean   :2.341          Mean   :2.455         Mean   : 8632   Mean   :1163
Mean   : 7469   Mean   :0.7599      Mean   : 4404
 3rd Qu.:52.00   3rd Qu.:3.000   3rd Qu.:40.00   3rd Qu.:5.000           3rd Qu.:3.000          3rd Qu.:3.000         3rd Qu.:11068   3rd Qu.:1784
3rd Qu.: 9859   3rd Qu.:0.8590      3rd Qu.: 4741
 Max.   :73.00   Max.   :5.000   Max.   :56.00   Max.   :6.000           Max.   :6.000          Max.   :6.000         Max.   :34516   Max.   :2517
Max.   :34516   Max.   :3.3970      Max.   :18484
 Total_Trans_Ct   Total_Ct_Chng_Q4_Q1
 Min.   : 10.00   Min.   :0.0000
 1st Qu.: 45.00   1st Qu.:0.5820
 Median : 67.00   Median :0.7020
 Mean   : 64.86   Mean   :0.7122
 3rd Qu.: 81.00   3rd Qu.:0.8180
 Max.   :139.00   Max.   :3.7140
```

Distribution of other continuous variables with Discrete Variables:

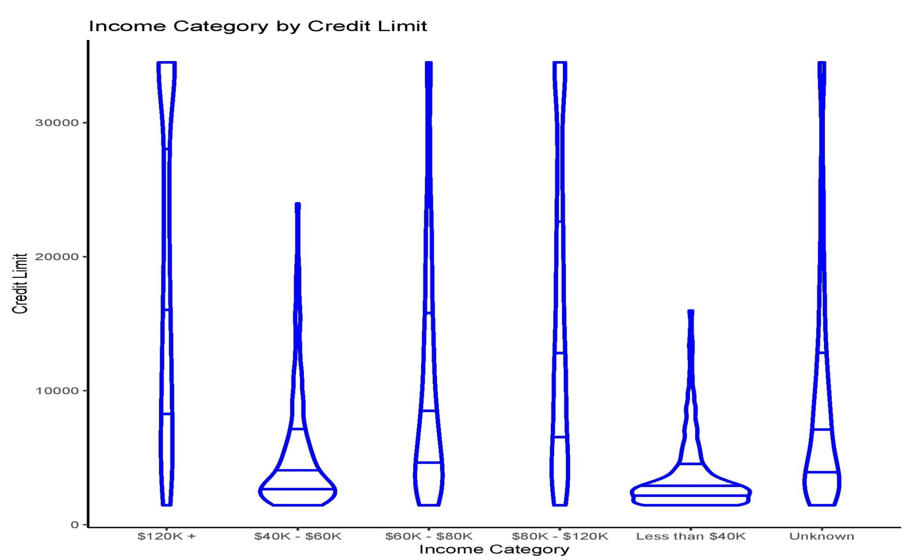
❖ Credit Limit with Attrition Flag

```
> ggplot(data , aes(Attrition_Flag,Credit_Limit,color= Credit_Limit)) +
geom_violin(draw_quantiles = c(0.25,0.5,0.75),colour="blue",size=1.4) +
theme_classic() +xlab("Income Category") + ylab("Credit Limit") + ggtitle("Customer
Status by Credit Limit" ) +   labs(fill = "Customer Status")
```



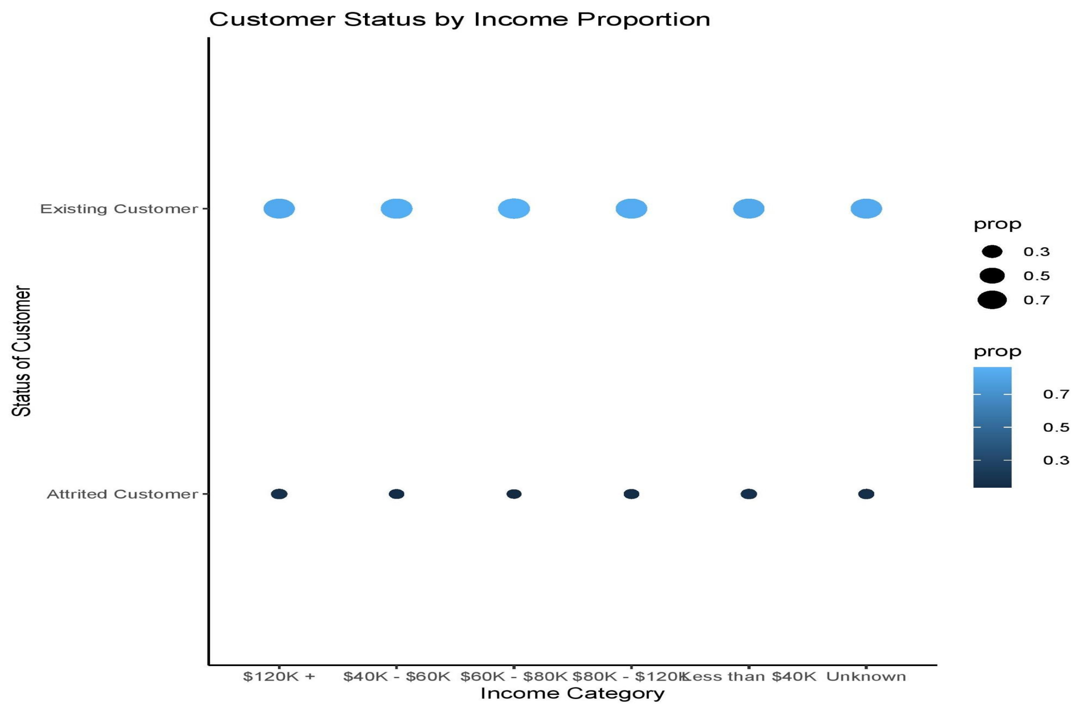
❖ Credit Limit with Income Category

```
> ggplot(data , aes(Income_Category,Credit_Limit,color= Credit_Limit)) +
geom_violin(draw_quantiles = c(0.25,0.5,0.75),colour="blue",size=1) +
theme_classic() +xlab("Income Category") + ylab("Credit Limit") + ggtitle("Income
Category by Credit Limit" )
```



❖ Income Category with Attrition Flag

```
> ggplot(data , aes(Income_Category,Attrition_Flag, colour= after_stat(prop), size =
after_stat(prop), group = Income_Category)) + geom_count() +  scale_size_area() +
theme_classic() +xlab("Income Category") + ylab("Status of Customer") +
ggtitle("Customer Status by Income Proportion" )
```



Conclusion from Exploratory Data Analysis

Based on our analysis in R program on our data as shown above,we can observe the following:

- Current Customers have higher mean credit limits than Attrited Customers.
- Majority of Attrited Customers fall in the less than $40K Income Category, but also the majority of Customers fall in this category.
- Majority of our Current Customers are having Graduate and High School degrees.
- Distribution of Current and Attrited Customers seems to be even. With less total of Customers being divorced.
- Blue Card is the most significant Card Category among Customers.
- Gender and Age is not a significant factor in determining Attrition Status of Customers.
- As per the Violin Plot between Income Category and Credit Limit, we see a wider spread for Current Customers than Attrited Customers.
- As assumed, the Higher Income Category correlates with Higher Credit Limit.
- Majority of our Data is of Current Customers than Attrited Customers.

3.HYPOTHESIS OF INTERESTS

HYPOTHESIS 1

We conducted a hypothesis test for two sample means - we took the mean of the Credit_Limit for both attrited customers and existing customers to test for significant differences. Let us consider the mean of the credit limit of the attrited customers to be μ_1 and the mean of the credit limit of the existing customers to be μ_2.We perform the hypothesis test at the 5% level of significanceThe null and two-sided research hypotheses for the nonparametric test are stated as follows:

$$\textbf{Ho} = \mu_1 - \mu_2 = 0 \text{ vs } \textbf{H1} = \mu_1 - \mu_2 \neq 0$$

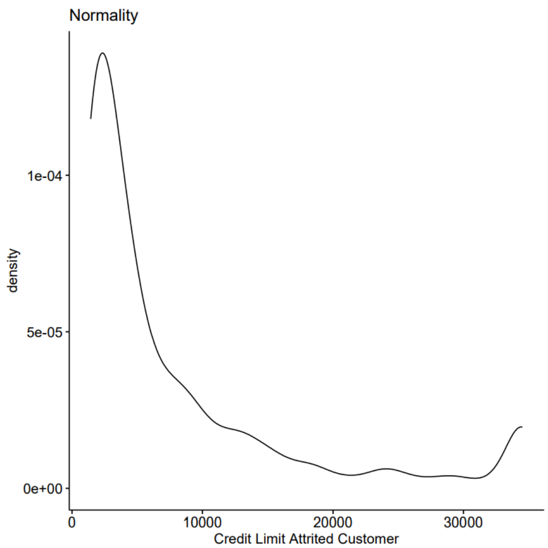
Let us read the data into our R program and subset the data based on the attrition flag and checking the normality of our data. There are two main methods of assessing normality are graphically and numerically.We also check our normality assumption using the Shapiro-Wilk statistical test and Anderson-Darling test.To perform the above mentioned tests, the R function **shapiro.test()** and **ad.test()** can used as shown

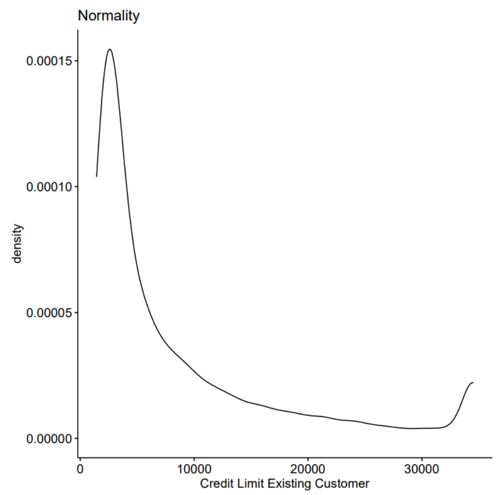
```
> #To Read the given data into R program
> data<-read.csv("C:/Users/skyalkond/Desktop/572 project/data.txt")
> #Subsetting our data based on attrition flag
> attrited<-subset(data,data$Attrition_Flag=="Attrited Customer")
> existing<-subset(data,data$Attrition_Flag=="Existing Customer")
> #Checking Normality of data
> shapiro.test(attrited$Credit_Limit)

        Shapiro-Wilk normality test

data:  attrited$Credit_Limit
W = 0.71476, p-value < 2.2e-16

> ad.test(existing$Credit_Limit)

        Anderson-Darling normality test

data:  existing$Credit_Limit
A = 779.17, p-value < 2.2e-16
```

We have used two different tests to check for normality of our data as the size of 2 subsets of data is different.For the Shapiro–Wilk's tests at 5% level of significance, because the p-value < 0.05 ,we conclude that the given data set does not following a normal distribution and for Anderson-Darling test at the same level of significance, because the p-value<0.05, we conclude that the given data set does not following a normal distribution.

To confirm we have visualized the normality using the R function **ggdensity()** as shown below.

```
> #Now we visualize the Credit Limit of attrited and existing customers
> ggdensity(attrited$Credit_Limit,main="Normality",xlab="Credit Limit Attrited Customer")
> ggdensity(existing$Credit_Limit,main="Normality",xlab="Credit Limit Existing Customer")
```





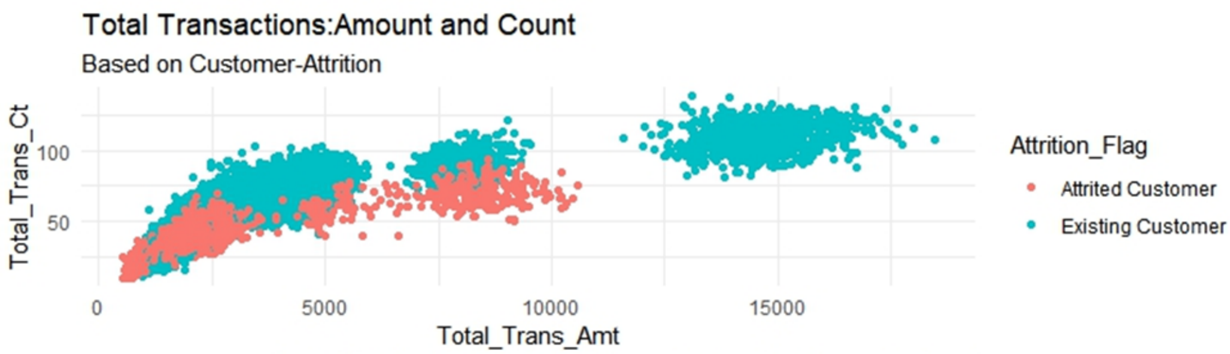
The data is very large and not normally distributed, therefore a Wilcoxon Rank-Sum Test is used to test the hypothesis on mean.To perform two-samples Wilcoxon test comparing the means, the R function **wilcox.test**() can be used as below

```
> wilcox.test(attrited$Credit_Limit,existing$Credit_Limit)

        Wilcoxon rank sum test with continuity correction

data:  attrited$Credit_Limit and existing$Credit_Limit
W = 6361348, p-value = 3.008e-07
alternative hypothesis: true location shift is not equal to 0
```

CONCLUSION

The p-value of **$3.008e^{-07}$** is smaller than the alpha value of 0.05%(5% significance level),thus we reject the null hypothesis. Therefore we can conclude that there is a major significant difference between attrited customers and existing customers when it comes to the credit limit of the customer.Now,we move on to more advanced statistical analysis and hypothesis testing using Multiple Logistic Regression.

HYPOTHESIS 2



Scatter-Plot of Individual Customers' Transaction Amount vs Transaction Count

The scatterplot of individual customers' total transaction amount (measured by variable Total_Trans_Amt) versus total transaction counts (measured by variable Total_Trans_Ct) reveals some interesting facts regarding the attrition trend among customers.

I. There is a general positive relationship visible between transaction amount and transaction counts for both types of customers.

II. However, for very high transaction amounts (roughly above the $12000 level), there are no attrited customers. In the lower transaction amount range, ($0-$12000), for a given level of transaction amount, the existing customers have transacted more frequently than the attrited customers. In other words, the leaving customers show a pattern of having spent higher amounts than the loyal customers for similar frequency of card usage.

We formed our hypothesis 2 based on the above two observations. It is only intuitive to expect that

a. The probability of attrition should be lower with higher transaction counts;

b. The probability of attrition should be higher with higher transaction amounts;

c. The change in the attrition-probability from an increase in transaction amount will be lower for higher transaction counts.

We transform our dependent variable *Attrition_Flag* (a categorical variable) into a dummy variable *Dummy_Attrition* which takes a value 1 when *Attrition_Flag* indicates an 'Attrited Customer' and 0 for an 'Existing Customer'. Compatible with our hypotheses, we are specifically interested in measuring the effects of the independent variables *Total_Trans_Ct, Total_Trans_Amt* and also an interaction of these two variables *Total_Trans_Ct* * *Total_Trans_Amt* on the probability that a customer is an Attrited Customer i.e., Prob{Dummy_Attrition = 1} = p (say).

We formalize our hypothesis by using the mathematical notation as:

$$H_0: \beta_i = 0 \; vs. H_0: \beta_i \neq 0; \;\; i = 1, 2, 3, \;\; \alpha = 0.05$$

To test our hypotheses, we perform a baseline logistic regression (REG-1) quantifying the below relationship.

$$\text{logit p} = \ln\left(\frac{p}{1-p}\right)$$

$$= \beta_0 + \beta_1 \times \text{Total\_Trans\_Ct} + \beta_2 \times \text{Total\_Trans\_Amt} + \beta_3 \times \text{Total\_Trans\_Ct} \cdot \text{Total\_Trans\_Amt}$$

The results from REG-1 are tabulated in Table 2.1. Not only the coefficient estimates are non-zero and statistically significant, the signs of $\beta_1, \beta_2, \beta_3$ are as expected (i.e., negative, positive, and negative respectively).

However, this is our baseline model with no control variables. To make our model more robust, we incorporate all the other variables in our dataset as independent variables and determine which variables have statistically significant effects on the attrition probability. REG-2 in Table 2.1 shows the result from this regression and separates the variables which are significant from those which are not.

We now use these variables as the predictor variables in our final regression model (REG-3). We test our hypotheses using REG-3 which quantifies the model as:

$$\text{logit p} = \ln\left(\frac{p}{1-p}\right)$$
$$= \beta_0 + \beta_1 \times \text{Total\_Trans\_Ct} + \beta_2 \times \text{Total\_Trans\_Amt} + \beta_3 \times \text{Total\_Trans\_Ct} \cdot \text{Total\_Trans\_Amt} + \boldsymbol{\beta}\mathbf{x}$$

; where **x** is a vector of the control variables we extracted from REG-2.

The R-command and the R-output of the regression are included subsequently. For our data, the logistic regression model is estimated as:

$$\widehat{\text{logit}}\text{ p} = 3.190 + -0.09926 \times \text{Total\_Trans\_Ct} + 0.00297 \times \text{Total\_Trans\_Amt} +$$
$$-0.00003 \times \text{Total\_Trans\_Ct} \cdot \text{Total\_Trans\_Amt} + \hat{\boldsymbol{\beta}}\mathbf{x}$$

We use the following command in R to execute our final regression:

R-command for logistic Regression:

```
> model_final <- glm(data= data, formula = Dummy_Attrition ~ Gender +
+                      Dependent_count + Income_Category  +
+                      Total_Relationship_Count +
+                      Months_Inactive_12_mon + Contacts_Count_12_mon +
+                      Total_Revolving_Bal + Total_Amt_Chng_Q4_Q1 +
+                      Total_Trans_Amt + Total_Trans_Ct + Total_Ct_Chng_Q4_Q1 + Total_Trans_
Ct * Total_Trans_Amt ,
+                  family = "binomial" (link = "logit"))
> |
```

The R-output of the final regression results is obtained as:

R-Output for Summary results:

```
> summary(model_final)

Call:
glm(formula = Dummy_Attrition ~ Gender + Dependent_count + Income_Category +
    Total_Relationship_Count + Months_Inactive_12_mon + Contacts_Count_12_mon +
    Total_Revolving_Bal + Total_Amt_Chng_Q4_Q1 + Total_Trans_Amt +
    Total_Trans_Ct + Total_Ct_Chng_Q4_Q1 + Total_Trans_Ct * Total_Trans_Amt,
    family = binomial(link = "logit"), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4151  -0.3142  -0.1134  -0.0227   3.9381

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    3.190e+00  3.747e-01   8.515  < 2e-16 ***
GenderM                       -8.580e-01  1.531e-01  -5.603 2.10e-08 ***
Dependent_count                1.284e-01  3.141e-02   4.087 4.36e-05 ***
Income_Category$40K - $60K    -6.568e-01  1.965e-01  -3.342 0.000832 ***
Income_Category$60K - $80K    -4.496e-01  1.816e-01  -2.476 0.013299 *
Income_Category$80K - $120K   -2.810e-01  1.761e-01  -1.596 0.110576
Income_CategoryLess than $40K -5.655e-01  2.118e-01  -2.670 0.007590 **
Income_CategoryUnknown        -7.003e-01  2.382e-01  -2.940 0.003286 **
Total_Relationship_Count      -4.864e-01  2.892e-02 -16.822  < 2e-16 ***
Months_Inactive_12_mon         4.880e-01  4.045e-02  12.063  < 2e-16 ***
Contacts_Count_12_mon          4.936e-01  3.869e-02  12.758  < 2e-16 ***
Total_Revolving_Bal           -9.191e-04  4.869e-05 -18.876  < 2e-16 ***
Total_Amt_Chng_Q4_Q1          -1.193e+00  2.066e-01  -5.773 7.78e-09 ***
Total_Trans_Amt                2.972e-03  1.281e-04  23.204  < 2e-16 ***
Total_Trans_Ct                -9.926e-02  4.557e-03 -21.781  < 2e-16 ***
Total_Ct_Chng_Q4_Q1           -3.098e+00  2.084e-01 -14.870  < 2e-16 ***
Total_Trans_Amt:Total_Trans_Ct -2.729e-05  1.478e-06 -18.468  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8927.2  on 10126  degrees of freedom
Residual deviance: 4132.3  on 10110  degrees of freedom
AIC: 4166.3

Number of Fisher Scoring iterations: 8
```

The p-values of the individual hypotheses are given in the R-output of the summary results. Since the p-values of $\widehat{\beta_1}$, $\widehat{\beta_2}$, $\widehat{\beta_3}$, are all in the range of 0 and 0.001, therefore less than $\alpha = 0.05$, we can reject the null-hypothesis that,

$$\beta_i = 0 \; ; \; i = 1, 2, 3$$

in favor of the alternate hypothesis with 95% confidence level. Also, the signs conform to our intuitive expectation. We can conclude that, the transaction count, transaction amount as well as their interaction being statistically significant, are valuable predictors for the attrition probability. For an individual customer, as the transaction count and amount both increase, a net negative effect will be on attrition but this effect will be lower for higher levels of transaction counts.

Table 2.1: Summary of Regression Results

Dependent Variable for the regressions: $\ln\left(\frac{p}{1-p}\right)$; p = Prob{*Dummy_Attrition* = 1}			
Predictor Variables	REG-1	REG-2	REG-3
Intercept	-0.66724*** (0.17087)	3.94636 *** (0.52611)	3.19033 *** (0.37469)
Total_Trans_Amt	0.00257*** (0.0001)	0.00291 *** (0.00013)	0.00297 *** (0.00013)
Total_Trans_Ct	-0.08245*** (0.00363)	-0.10314 *** (0.00465)	-0.09926 *** (0.00456)
Total_Trans_Ct:Total_Trans_Amt	-0.00002*** (0.00000126)	-0.00003 *** (0.000001474)	-0.00003 *** (0.000001478)
Dummy_Gender_M		-0.87689 *** (0.15472)	-0.85797 *** (0.15312)
Dependent_count		0.12936 *** (0.03215)	0.12839 *** (0.03141)
Dummy_Income_40-60		-0.87529 *** (0.2171)	-0.65679 *** (0.19652)
Dummy_Income_60-80		-0.60786 ** (0.19249)	-0.44957 . (0.1816)
Dummy_Income_80-120		-0.35081 . (0.17922)	-0.281 (0.17611)
Dummy_Income_Below40		-0.77865 *** (0.23435)	-0.56552 * (0.21182)
Total_Relationship_Count		-0.47862 *** (0.02909)	-0.48642 *** (0.02892)
Months_Inactive_12_mon		0.49472 *** (0.04108)	0.48801 *** (0.04045)
Contacts_Count_12_mon		0.50118 *** (0.03908)	0.49357 *** (0.03869)
Total_Revolving_Bal		-0.00082 *** (0.00008)	-0.00092 *** (0.00005)
Total_Amt_Chng_Q4_Q1		-1.16892 *** (0.21021)	-1.19278 *** (0.20661)
Total_Ct_Chng_Q4_Q1		-3.06271 *** (0.20912)	-3.09827 *** (0.20836)
Credit_Limit		-0.00002 * (0.00001)	
Customer_Age		0.00091 (0.00832)	
Dummy_Education_Doctorate		0.33871 (0.22124)	
Dummy_Education_Graduate		-0.01715 (0.14771)	
Dummy_Education_HighSchool		0.01946 (0.15755)	
Dummy_Education_PostGraduate		0.17603 (0.22469)	
Dummy_Education_Uneducated		0.0722 (0.16682)	
Dummy_Education_Unknown		0.09249 (0.16586)	

3.MECHANISMS TO COUNTER MISSING DATA

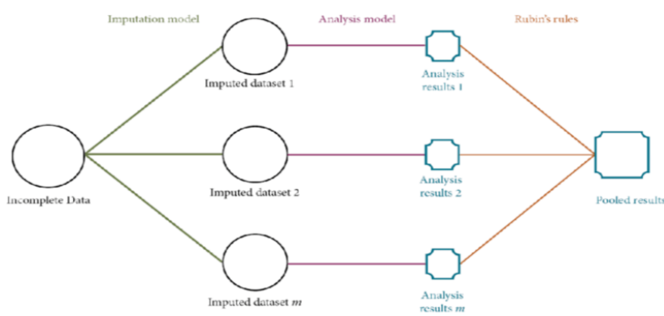
Commonly, data is explained based on the reasons for missing data. We assume TWO kinds of missing data, based on the missingness mechanisms:
`

Missing Completely at Random (MCAR):

MCAR is defined as the probability that the missingness of the data is not related to either the specific value which is supposed to be obtained or the set of observed responses.

Handling of MCAR data:

This function uses the generated simulated matrix and generates missing data points in a missing-completely-at-random pattern for each variable, considering the fraction of missingness for each variable, so potential missing data fraction imbalances between variables in the original data will be retained.In our data set, we have chosen the Credit_Limit column and decided to generate missing values We tried to remove 20% of the data randomly using the MCAR function. In order to impute the missing values here in our data set, we have used the mice package and we imputed the missing value.



Missing Not at Random (MNAR)

The concept of MNAR is the most complicated of the assumed natures of missing data. MNAR suggests that the probability of a value being missing fluctuates for reasons unknown to us. When the characteristics of missing data do not meet those of MCAR, they are categorized into data that is MNAR. A case of MNAR assumes that the missingness is directly related to what is missing

Handling of MNAR data

This function uses the generated simulate`d matrix and generates missing data points in a missing-not-at-random pattern for each variable ,considering the fraction of missingness in the original dataset and the original missingness pattern. The characteristic of the MNAR pattern is that the missingness in a variable is dependent on its own distribution.Since there are no missing values in the data, we removed 20% of the data from the column "Credit_Limit " and imputed the missing data using the mice package as shown below.

MCAR for Hypothesis 1:

Initially,we check for the existing null values in our data

```
> apply(is.na(data),2,sum)
                                                                                                 CLIENTNUM
                                                                                                         0
                                                                                            Attrition_Flag
                                                                                                         0
                                                                                              Customer_Age
                                                                                                         0
                                                                                                    Gender
                                                                                                         0
                                                                                           Dependent_count
                                                                                                         0
                                                                                           Education_Level
                                                                                                         0
                                                                                            Marital_Status
                                                                                                         0
                                                                                           Income_Category
                                                                                                         0
                                                                                             Card_Category
                                                                                                         0
                                                                                            Months_on_book
                                                                                                         0
                                                                                  Total_Relationship_Count
                                                                                                         0
                                                                                    Months_Inactive_12_mon
                                                                                                         0
                                                                                     Contacts_Count_12_mon
                                                                                                         0
                                                                                              Credit_Limit
                                                                                                         0
                                                                                       Total_Revolving_Bal
                                                                                                         0
                                                                                           Avg_Open_To_Buy
                                                                                                         0
                                                                                      Total_Amt_Chng_Q4_Q1
                                                                                                         0
                                                                                           Total_Trans_Amt
                                                                                                         0
                                                                                            Total_Trans_Ct
                                                                                                         0
                                                                                       Total_Ct_Chng_Q4_Q1
                                                                                                         0
                                                                                     Avg_Utilization_Ratio
                                                                                                         0
Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1
                                                                                                         0
Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2
                                                                                                         0
```

We can conclude that the data does not contain any duplicates.Now, we randomly generate NA data using functions from the missMethods package in R.Now, we create missing data for credit limit and then impute the missing values using the functions from mice library.

```
#MCAR
data_mcar1- delete_MCAR(data, 0.2,"Credit_Limit")
> sapply(data_mcar1,function(x) sum(is.na(x)))
```

Now, we can check the number of missing data using the R function **sapply**() and impute values into our new missing dataset using the mice library.

```
> imputed_Data <- mice(data_mcar1, m=5, maxit = 50, method = 'cart', seed = 500)
> completeData1 <- complete(imputed_Data,2)
> sapply(completeData1,function(x) sum(is.na(x)))
```

We use the Classification and regression trees (CART) method to impute the given missing values into the table.This is generally used when we have continuous or categorical dependent variables and we have confirmed our data to contain no missing values after we imputed the values and confirmed the mean of the credit limit hasn't changed much compared to the original dataset.

```
> summary(completeData1$Credit_Limit)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1438    2554    4555    8637   11094   34516 
```

Now, we carry out the **Wilcoxon test** as earlier,

```
> attrited3<-subset(completeData1,completeData1$Attrition_Flag=="Attrited Customer")
> existing3<-subset(completeData1,completeData1$Attrition_Flag=="Existing Customer")
> wilcox.test(attrited3$Credit_Limit,existing3$Credit_Limit)

	Wilcoxon rank sum test with continuity correction

data:  attrited3$Credit_Limit and existing3$Credit_Limit
W = 6266213, p-value = 3.595e-07
alternative hypothesis: true location shift is not equal to 0
```

We have now compared between imputed data predictions and original data predictions as the P-value of our imputed data is closely equivalent to the P-Value of our original data as shown in hypothesis 1.

MNAR for Hypothesis 1:

For MNAR, the only change we made to the previous code is to change how missing data is generated. We generated missing data here by removing 20% of the smallest values under Credit_Limit. Our assumption was that data may knowingly or unknowingly provide a sample of data which contain smaller credit limits for the attrited customers.

```
MNAR
> data_mnar1<-delete_MNAR_censoring(data,0.2,"Credit_Limit")
> sapply(data_mnar1,function(x) sum(is.na(x)))
```

Now,we impute the missing values using the mice package and

```
> imputed_Data <- mice(data_mnar1, m=5, maxit = 50, method = 'rf', seed = 500)
> completeData2 <- complete(imputed_Data,2)
> sapply(completeData2,function(x) sum(is.na(x)))
```

We employ Imputation by random forests(rf) method to impute values into the missing values since they perform for non-normally distributed data or when there are non-linear relationships or interactions without assuming normality or require specification of parametric models.

Now, we carry out the **Wilcoxon test**.

```
> wilcox.test(attrited2$Credit_Limit,existing2$Credit_Limit)

Wilcoxon rank sum test with continuity correction

data:  attrited2$Credit_Limit and existing2$Credit_Limit
W = 7397592, p-value = 7.839e-06
alternative hypothesis: true location shift is not equal to 0
```

We executed the Wilcoxon rank sum test and got a negligible P-value. We have now compared between imputed data predictions and original data predictions as the P-value of our imputed data is more or less in line with the P-Value of our original data as shown in hypothesis 1.

Original data P-value-> $3.008e^{-07}$
Imputed data P-value(MCAR)->$3.595e^{-07}$
Imputed data P-value(MNAR)->$7.839e^{-06}$

MCAR for Hypothesis 2

Let us create missing data for below mentioned columns as mentioned in R program

```
> data_mcar2<- delete_MCAR(data, 0.2,
c("Total_Trans_Amt","Months_Inactive_12_mon","Total_Revolving_Bal"))
> summary(data_mcar2)
```

We have created missing values in three columns at Random, "Total_Trans_Amt", "Months_Inactive_12_mon" and "Total_Revolving_Bal". As we can see these are essential columns in our Logistic Regression Model and are not dependent on each other. We remove 20% of this data which is around 2056 rows in each of the three columns. We Impute the data using mice package for the very same reason. We use a random forest method for imputation.

```
> imputed_data_mcar2 <- mice(data_mcar2,m=5,maxit=50,method='rf',seed=500)

 iter imp variable
  1   1  Months_Inactive_12_mon  Total_Revolving_Bal  Total_Trans_Amt
  1   2  Months_Inactive_12_mon  Total_Revolving_Bal  Total_Trans_Amt
  1   3  Months_Inactive_12_mon  Total_Revolving_Bal  Total_Trans_Amt
  1   4  Months_Inactive_12_mon  Total_Revolving_Bal  Total_Trans_Amt
  1   5  Months_Inactive_12_mon  Total_Revolving_Bal  Total_Trans_Amt
  2   1  Months_Inactive_12_mon  Total_Revolving_Bal  Total_Trans_Amt
  2   2  Months_Inactive_12_mon  Total_Revolving_Bal  Total_Trans_Amt
  2   3  Months_Inactive_12_mon  Total_Revolving_Bal  Total_Trans_Amt
  2   4  Months_Inactive_12_mon  Total_Revolving_Bal  Total_Trans_Amt
  2   5  Months_Inactive_12_mon  Total_Revolving_Bal  Total_Trans_Amt
  3   1  Months_Inactive_12_mon  Total_Revolving_Bal  Total_Trans_Amt
  3   2  Months_Inactive_12_mon  Total_Revolving_Bal  Total_Trans_Amt
  3   3  Months_Inactive_12_mon  Total_Revolving_Bal  Total_Trans_Amt
```

Now,since we have imputed the miss values, the summary of the data before and after imputation is shown below.

```
> summary(data_mcar2)
   CLIENTNUM         Attrition_Flag      Customer_Age      Gender          Dependent_count Education_Level
 Min.   :708082083   Length:10127       Min.   :26.00   Length:10127       Min.   :0.000   Length:10127
 1st Qu.:713036770   Class :character   1st Qu.:41.00   Class :character   1st Qu.:1.000   Class :character
 Median :717926358   Mode  :character   Median :46.00   Mode  :character   Median :2.000   Mode  :character
 Mean   :739177606                      Mean   :46.33                      Mean   :2.346
 3rd Qu.:773143533                      3rd Qu.:52.00                      3rd Qu.:3.000
 Max.   :828343083                      Max.   :73.00                      Max.   :5.000

 Marital_Status     Income_Category    Card_Category      Months_on_book  Total_Relationship_Count Months_Inactive_12_mon
 Length:10127       Length:10127       Length:10127       Min.   :13.00   Min.   :1.000            Min.   :0.000
 Class :character   Class :character   Class :character   1st Qu.:31.00   1st Qu.:3.000            1st Qu.:2.000
 Mode  :character   Mode  :character   Mode  :character   Median :36.00   Median :4.000            Median :2.000
                                                          Mean   :35.93   Mean   :3.813            Mean   :2.339
                                                          3rd Qu.:40.00   3rd Qu.:5.000            3rd Qu.:3.000
                                                          Max.   :56.00   Max.   :6.000            Max.   :6.000
                                                                                                   NA's   :2025
 Contacts_Count_12_mon  Credit_Limit   Total_Revolving_Bal Avg_Open_To_Buy Total_Amt_Chng_Q4_Q1 Total_Trans_Amt
 Min.   :0.000         Min.   : 1438   Min.   :   0.0      Min.   :    3   Min.   :0.0000       Min.   :  563
 1st Qu.:2.000         1st Qu.: 2555   1st Qu.: 397.8      1st Qu.: 1324   1st Qu.:0.6310       1st Qu.: 2148
 Median :2.000         Median : 4549   Median :1285.0      Median : 3474   Median :0.7360       Median : 3889
 Mean   :2.455         Mean   : 8632   Mean   :1166.5      Mean   : 7469   Mean   :0.7599       Mean   : 4390
 3rd Qu.:3.000         3rd Qu.:11068   3rd Qu.:1788.0      3rd Qu.: 9859   3rd Qu.:0.8590       3rd Qu.: 4732
 Max.   :6.000         Max.   :34516   Max.   :2517.0      Max.   :34516   Max.   :3.3970       Max.   :18484
                                       NA's   :2025                                             NA's   :2025
 Total_Trans_Ct   Total_Ct_Chng_Q4_Q1 Avg_Utilization_Ratio
 Min.   : 10.00   Min.   :0.0000      Min.   :0.0000
 1st Qu.: 45.00   1st Qu.:0.5820      1st Qu.:0.0230
 Median : 67.00   Median :0.7020      Median :0.1760
 Mean   : 64.86   Mean   :0.7122      Mean   :0.2749
 3rd Qu.: 81.00   3rd Qu.:0.8180      3rd Qu.:0.5030
 Max.   :139.00   Max.   :3.7140      Max.   :0.9990

 Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1
 Min.   :0.0000077
 1st Qu.:0.0000990
 Median :0.0001815
 Mean   :0.1599975
 3rd Qu.:0.0003373
 Max.   :0.9995800

 Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2
 Min.   :0.00042
 1st Qu.:0.99966
 Median :0.99982
 Mean   :0.84000
 3rd Qu.:0.99990
 Max.   :0.99999
```

```
> completeData1 <- complete(imputed_data_mcar2,2)
> summary(completeData1)
   CLIENTNUM          Attrition_Flag      Customer_Age       Gender          Dependent_count Education_Level
 Min.   :708082083   Length:10127        Min.   :26.00   Length:10127       Min.   :0.000   Length:10127
 1st Qu.:713036770   Class :character    1st Qu.:41.00   Class :character   1st Qu.:1.000   Class :character
 Median :717926358   Mode  :character    Median :46.00   Mode  :character   Median :2.000   Mode  :character
 Mean   :739177606                       Mean   :46.33                      Mean   :2.346
 3rd Qu.:773143533                       3rd Qu.:52.00                      3rd Qu.:3.000
 Max.   :828343083                       Max.   :73.00                      Max.   :5.000
 Marital_Status      Income_Category     Card_Category      Months_on_book  Total_Relationship_Count Months_Inactive_12_mon
 Length:10127        Length:10127        Length:10127        Min.   :13.00   Min.   :1.000            Min.   :0.000
 Class :character    Class :character    Class :character    1st Qu.:31.00   1st Qu.:3.000            1st Qu.:2.000
 Mode  :character    Mode  :character    Mode  :character    Median :36.00   Median :4.000            Median :2.000
                                                              Mean   :35.93   Mean   :3.813            Mean   :2.329
                                                              3rd Qu.:40.00   3rd Qu.:5.000            3rd Qu.:3.000
                                                              Max.   :56.00   Max.   :6.000            Max.   :6.000
 Contacts_Count_12_mon  Credit_Limit    Total_Revolving_Bal Avg_Open_To_Buy Total_Amt_Chng_Q4_Q1 Total_Trans_Amt
 Min.   :0.000          Min.   : 1438   Min.   :   0        Min.   :    3   Min.   :0.0000       Min.   :  563
 1st Qu.:2.000          1st Qu.: 2555   1st Qu.: 243        1st Qu.: 1324   1st Qu.:0.6310       1st Qu.: 2148
 Median :2.000          Median : 4549   Median :1276        Median : 3474   Median :0.7360       Median : 3894
 Mean   :2.455          Mean   : 8632   Mean   :1161        Mean   : 7469   Mean   :0.7599       Mean   : 4403
 3rd Qu.:3.000          3rd Qu.:11068   3rd Qu.:1786        3rd Qu.: 9859   3rd Qu.:0.8590       3rd Qu.: 4739
 Max.   :6.000          Max.   :34516   Max.   :2517        Max.   :34516   Max.   :3.3970       Max.   :18484
 Total_Trans_Ct   Total_Ct_Chng_Q4_Q1 Avg_Utilization_Ratio
 Min.   : 10.00   Min.   :0.0000      Min.   :0.0000
 1st Qu.: 45.00   1st Qu.:0.5820      1st Qu.:0.0230
 Median : 67.00   Median :0.7020      Median :0.1760
 Mean   : 64.86   Mean   :0.7122      Mean   :0.2749
 3rd Qu.: 81.00   3rd Qu.:0.8180      3rd Qu.:0.5030
 Max.   :139.00   Max.   :3.7140      Max.   :0.9990
```

Now let us run the model on the Imputed Data, and check if our Logistic Regression model is performing as expected for Missing Values Completely at Random.

```
> completeData1 <- completeData1 %>%
+ +   mutate(Dummy_Attrition = ifelse(Attrition_Flag == "Attrited Customer", 1, 0))
Error in ifelse(Attrition_Flag == "Attrited Customer", 1, 0) :
  object 'Attrition_Flag' not found
> completeData1 <- completeData1 %>% mutate(Dummy_Attrition = ifelse(Attrition_Flag == "Attrited Customer", 1, 0))
> model_final <- glm(data= completeData1, formula = Dummy_Attrition ~ Gender +
+                      Dependent_count + Income_Category  +
+                      Total_Relationship_Count +
+                      Months_Inactive_12_mon + Contacts_Count_12_mon +
+                      Total_Revolving_Bal + Total_Amt_Chng_Q4_Q1 +
+                      Total_Trans_Amt + Total_Trans_Ct + Total_Ct_Chng_Q4_Q1 + Total_Trans_Ct * Total_Trans_Amt ,
+                    family = "binomial" (link = "logit"))
```

We first convert the Dependent Variable Attrition Flag into Dummy Attrition with 0|1 value. Then we fit the Logistic Regression model on the data. Let's check the Model Summary.

```
> summary(model_final)

Call:
glm(formula = Dummy_Attrition ~ Gender + Dependent_count + Income_Category +
    Total_Relationship_Count + Months_Inactive_12_mon + Contacts_Count_12_mon +
    Total_Revolving_Bal + Total_Amt_Chng_Q4_Q1 + Total_Trans_Amt +
    Total_Trans_Ct + Total_Ct_Chng_Q4_Q1 + Total_Trans_Ct * Total_Trans_Amt,
    family = binomial(link = "logit"), data = completeData1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.5245  -0.3365  -0.1374  -0.0341   3.8170

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   3.121e+00  3.643e-01   8.566  < 2e-16 ***
GenderM                      -8.692e-01  1.504e-01  -5.780 7.49e-09 ***
Dependent_count               1.186e-01  3.058e-02   3.878 0.000105 ***
Income_Category$40K - $60K   -6.241e-01  1.920e-01  -3.250 0.001155 **
Income_Category$60K - $80K   -3.632e-01  1.773e-01  -2.048 0.040517 *
Income_Category$80K - $120K  -2.282e-01  1.724e-01  -1.323 0.185670
Income_CategoryLess than $40K -5.018e-01  2.080e-01  -2.412 0.015858 *
Income_CategoryUnknown       -6.941e-01  2.338e-01  -2.969 0.002992 **
Total_Relationship_Count     -4.537e-01  2.795e-02 -16.233  < 2e-16 ***
Months_Inactive_12_mon        5.439e-01  3.984e-02  13.654  < 2e-16 ***
Contacts_Count_12_mon         4.749e-01  3.730e-02  12.730  < 2e-16 ***
Total_Revolving_Bal          -9.284e-04  4.723e-05 -19.657  < 2e-16 ***
Total_Amt_Chng_Q4_Q1         -9.749e-01  1.984e-01  -4.914 8.94e-07 ***
Total_Trans_Amt               2.272e-03  1.103e-04  20.594  < 2e-16 ***
Total_Trans_Ct               -8.721e-02  4.213e-03 -20.703  < 2e-16 ***
Total_Ct_Chng_Q4_Q1          -3.016e+00  2.027e-01 -14.875  < 2e-16 ***
Total_Trans_Amt:Total_Trans_Ct -2.075e-05  1.290e-06 -16.082  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8927.2  on 10126  degrees of freedom
Residual deviance: 4387.7  on 10110  degrees of freedom
AIC: 4421.7

Number of Fisher Scoring iterations: 8
```

We can Infer that our Model performs well looking at the Summary of Model fit for Original Data and Imputed Data

MNAR for hypothesis 2

```
> data.mis<-delete_MNAR_censoring(data,0.2,"Total_Revolving_Bal")
```

Here, we create Missing Values Not at Random, From correlation matrix it could be clearly seen that, we have no relationship between Total Revolving Bal and other columns. Hence we remove 20% of that Column. Therefore we are using the MICE package for imputation.

```
> imputed_Data <- mice(data.mis, m=5, maxit = 50, method = 'rf', seed = 500)
```

We are using Random Forest, as it's a Non parametric Imputation algorithm and also it does not make any assumption of the underlying data. It can also handle correlation between input variables. Hence we employ Random Forest for Imputation. As we can see as per Summary of before the After imputation the MICE package was able to accurately plot the missing data closer to the original data.

```
> completeData1 <- complete(imputed_Data,2)
> summary(completeData1$Total_Revolving_Bal)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0     458    1277    1165    1784    2517 
> summary(data$Total_Revolving_Bal)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0     359    1276    1163    1784    2517 
```

We first convert the Dependent Variable Attrition Flag into Dummy Attrition with 0|1 value. Then we fit the Logistic Regression model on the data. Let's check the Model Summary.

```
> completeData1 <- completeData1 %>%
+ +   mutate(Dummy_Attrition = ifelse(Attrition_Flag == "Attrited Customer", 1, 0))
Error in ifelse(Attrition_Flag == "Attrited Customer", 1, 0) : 
  object 'Attrition_Flag' not found
> completeData1 <- completeData1 %>% mutate(Dummy_Attrition = ifelse(Attrition_Flag == "Attrited Customer", 1, 0))
> model_final <- glm(data= completeData1, formula = Dummy_Attrition ~ Gender +
+                      Dependent_count + Income_Category  +
+                      Total_Relationship_Count +
+                      Months_Inactive_12_mon + Contacts_Count_12_mon +
+                      Total_Revolving_Bal + Total_Amt_Chng_Q4_Q1 +
+                      Total_Trans_Amt + Total_Trans_Ct + Total_Ct_Chng_Q4_Q1 + Total_Trans_Ct * Total_Trans_Amt ,
+                    family = "binomial" (link = "logit"))
> summary(model_final)

Call:
glm(formula = Dummy_Attrition ~ Gender + Dependent_count + Income_Category + 
    Total_Relationship_Count + Months_Inactive_12_mon + Contacts_Count_12_mon + 
    Total_Revolving_Bal + Total_Amt_Chng_Q4_Q1 + Total_Trans_Amt + 
    Total_Trans_Ct + Total_Ct_Chng_Q4_Q1 + Total_Trans_Ct * Total_Trans_Amt, 
    family = binomial(link = "logit"), data = completeData1)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-6.5245  -0.3365  -0.1374  -0.0341   3.8170  

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)                     3.121e+00  3.643e-01   8.566  < 2e-16 ***
GenderM                        -8.692e-01  1.504e-01  -5.780 7.49e-09 ***
Dependent_count                 1.186e-01  3.058e-02   3.878 0.000105 ***
Income_Category$40K - $60K     -6.241e-01  1.920e-01  -3.250 0.001155 ** 
Income_Category$60K - $80K     -3.632e-01  1.773e-01  -2.048 0.040517 *  
Income_Category$80K - $120K    -2.282e-01  1.724e-01  -1.323 0.185670    
Income_CategoryLess than $40K  -5.018e-01  2.080e-01  -2.412 0.015858 *  
Income_CategoryUnknown         -6.941e-01  2.338e-01  -2.969 0.002992 ** 
Total_Relationship_Count       -4.537e-01  2.795e-02 -16.233  < 2e-16 ***
Months_Inactive_12_mon          5.439e-01  3.984e-02  13.654  < 2e-16 ***
Contacts_Count_12_mon           4.749e-01  3.730e-02  12.730  < 2e-16 ***
Total_Revolving_Bal            -9.284e-04  4.723e-05 -19.657  < 2e-16 ***
Total_Amt_Chng_Q4_Q1           -9.749e-01  1.984e-01  -4.914 8.94e-07 ***
Total_Trans_Amt                 2.272e-03  1.103e-04  20.594  < 2e-16 ***
Total_Trans_Ct                 -8.721e-02  4.213e-03 -20.703  < 2e-16 ***
Total_Ct_Chng_Q4_Q1            -3.016e+00  2.027e-01 -14.875  < 2e-16 ***
Total_Trans_Amt:Total_Trans_Ct -2.075e-05  1.290e-06 -16.082  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8927.2  on 10126  degrees of freedom
Residual deviance: 4387.7  on 10110  degrees of freedom
AIC: 4421.7

Number of Fisher Scoring iterations: 8
```

CONCLUSION

In our first hypothesis we were able to conclude through the Wilcoxon Rank-Sum Test that we have a significant difference between the mean credit limit of attrited and existing customers. Through Logistic Regression, we were able to identify the variables which impact Attrition. We tried out both our Hypothesis on Original, MCAR and MNAR data.

REFERENCES

[1] Buuren, S. V., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software [2]Kropko, Jonathan, Ben Goodrich, Andrew Gelman, and Jennifer Hill. 2014. "Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches." Political Analysis 22, no. 4.
[3]https://socialsciences.mcmaster.ca/jfox/Courses/soc740/Missing-data-notes.pdf
[4]https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputingmissing-values/
[5]Rubin DB. (1976). Inference and missing data Biometrika,63(3), 581–592. [6]Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83(404), 1198-1202.
[7]https://www.researchgate.net/publication/275499071_Exploratory_Data_Analysis
[8]https://www.statisticssolutions.com/logistic-regression/
[9]https://www.statology.org/assumptions-of-logistic-regression/
[10]https://medium.com/geekculture/essential-guide-to-handle-outliers-for-your-logistic-r
egression-model-63c97690a84