# GROUP PROJECT

## WQD 7004 PROGRAMMING FOR DATA SCIENCE

## SEM 1

| NAME | MATRIC NO |
| --- | --- |
| FATIMAH BINTI MOHD NIZAM | 17218825 |
| ILANI DAYANA BINTI NOOR AZMAN | S2003292 |

# Introduction

- Customer Churn or known as customer attrition happens when a company loss clients or its customers.
- Bank, insurance and telecommunication companies use customer churn analysis to analyze which existing customers more likely have the potential to leave the companies' services.
- To retain existing customers cost lesser than finding new ones.
- This analysis focuses on analyzing the behavior of bank customers who are more likely to close their accounts according to the other customers'

# Dataset

https://www.kaggle.com/sakshigoyal7/credit-card-customers?select=BankChurners.csv

# Problem Statement

i. Does the proportion of income category on education level difference?

i. Is there any significant difference between customer age and income category?

i. Which classification algorithms performance in the Customer Churn Prediction has a better accuracy?

# Objectives

i. To determine the proportion of income category on education level

i. To determine is there are any significance difference between customer age and income category using ANOVA table.

i. To conduct the Customer Churn Prediction by using classification algorithms such as Classification Tree and Extreme Gradient Boosting

# 1. Introduction to Dataset

- There are 10127 rows and 23 columns in the dataset.

| | CLIENTNUM <int> | Attrition_Flag <chr> | Customer_Age <int> | Gender <chr> | Dependent_count <int> | Education_Level <chr> | Marital_Status <chr> | Income_Category <chr> |
|---|---|---|---|---|---|---|---|---|
| 1 | 768805383 | Existing Customer | 45 | M | 3 | High School | Married | $60K - $80K |
| 2 | 818770008 | Existing Customer | 49 | F | 5 | Graduate | Single | Less than $40K |
| 3 | 713982108 | Existing Customer | 51 | M | 3 | Graduate | Married | $80K - $120K |
| 4 | 769911858 | Existing Customer | 40 | F | 4 | High School | Unknown | Less than $40K |
| 5 | 709106358 | Existing Customer | 40 | M | 3 | Uneducated | Married | $60K - $80K |
| 6 | 713061558 | Existing Customer | 44 | M | 2 | Graduate | Married | $40K - $60K |

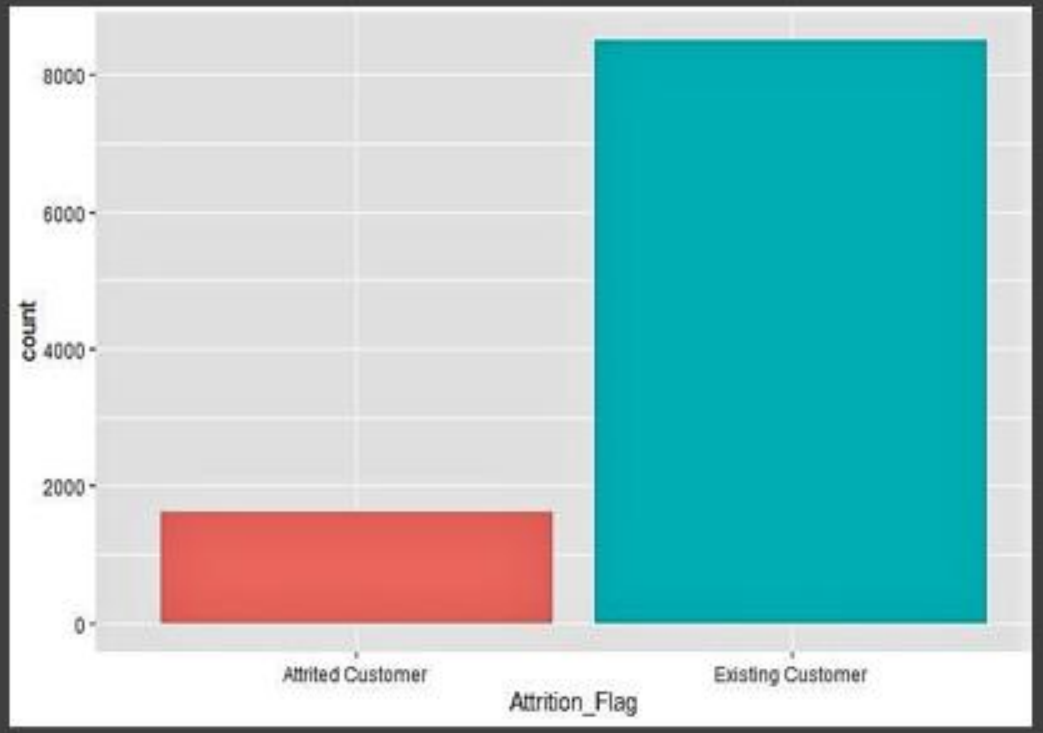
| Card_Category <chr> | Months_on_book <int> | Total_Relationship_Count <int> | Months_Inactive_12_mon <int> | Contacts_Count_12_mon <int> | Credit_Limit <dbl> |
|---|---|---|---|---|---|
| Blue | 39 | 5 | 1 | 3 | 12691 |
| Blue | 44 | 6 | 1 | 2 | 8256 |
| Blue | 36 | 4 | 1 | 0 | 3418 |
| Blue | 34 | 3 | 4 | 1 | 3313 |
| Blue | 21 | 5 | 1 | 0 | 4716 |
| Blue | 36 | 3 | 1 | 2 | 4010 |

# 2. Data Extraction & 3. Data Cleaning

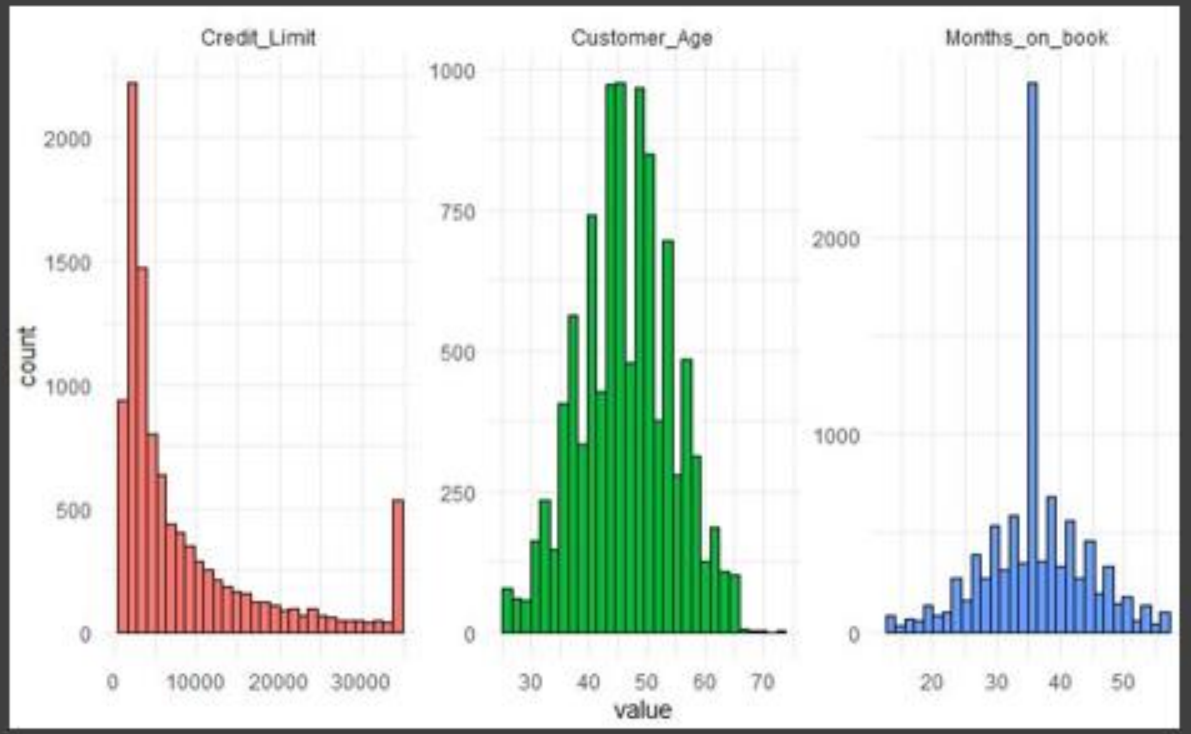
- Remove the features that are not used for data analysis and convert the necessary features to factors. Remove the NA or null values.

| Attrition_Flag <fctr> | Customer_Age <int> | Gender <fctr> | Education_Level <fctr> | Income_Category <fctr> | Card_Category <fctr> | Months_on_book <int> |
|---|---|---|---|---|---|---|
| Existing Customer | 45 | M | High School | $60K - $80K | Blue | 39 |
| Existing Customer | 49 | F | Graduate | Less than $40K | Blue | 44 |
| Existing Customer | 51 | M | Graduate | $80K - $120K | Blue | 36 |
| Existing Customer | 40 | F | High School | Less than $40K | Blue | 34 |
| Existing Customer | 40 | M | Uneducated | $60K - $80K | Blue | 21 |
| Existing Customer | 44 | M | Graduate | $40K - $60K | Blue | 36 |
| Existing Customer | 51 | M | Unknown | $120K + | Gold | 46 |
| Existing Customer | 32 | M | High School | $60K - $80K | Silver | 27 |
| Existing Customer | 37 | M | Uneducated | $60K - $80K | Blue | 36 |
| Existing Customer | 48 | M | Graduate | $80K - $120K | Blue | 36 |

# 4. EDA




The Number of Existing Customers are larger than Churning Customers



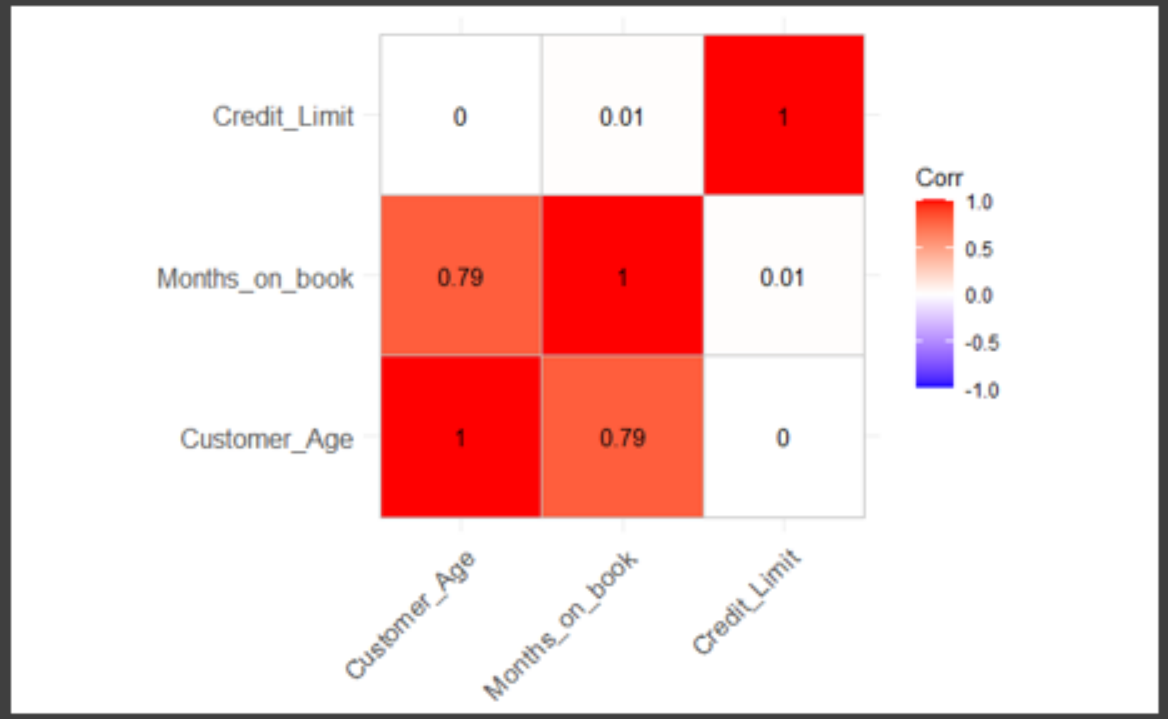

**CONTINUOUS VARIABLE DISTRIBUTION**

- The credit limit is skewed to the right

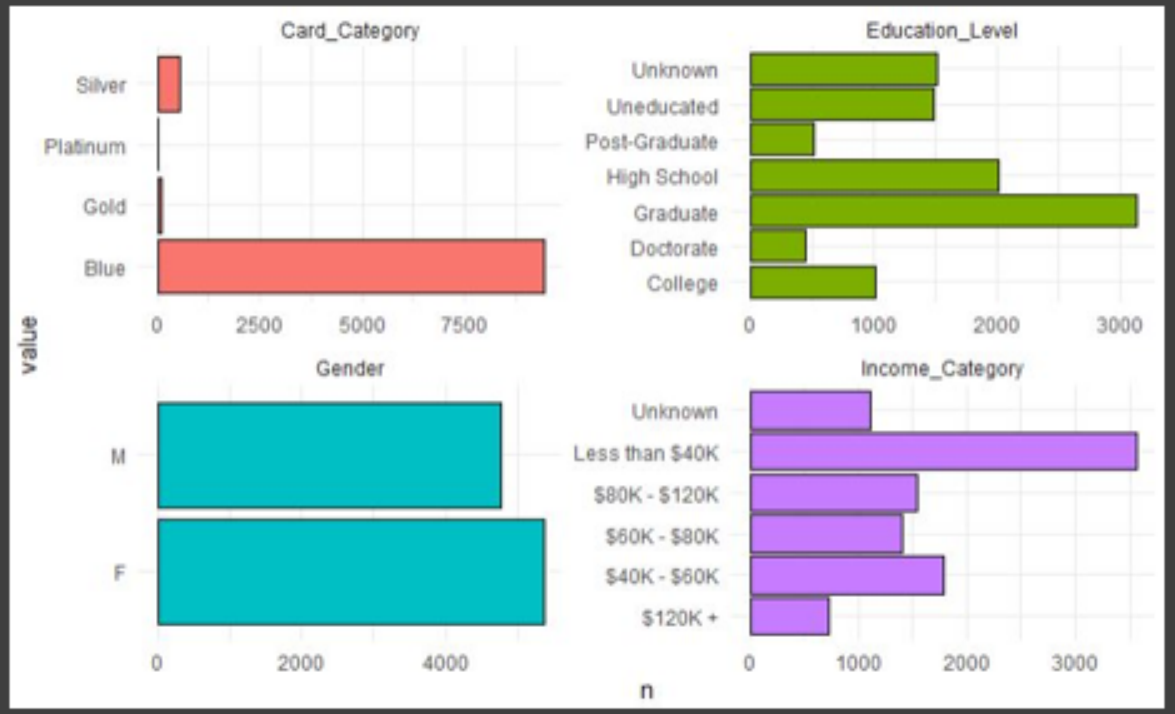
-The customer age almost have normal distribution

- The Highest Period of relationship with the Bank would be around 35 months

# 4. EDA




There is a quite high correlation between the Customer Age and the months on book (period of relationship with the bank). This is an example of multicollinearity.




**CATEGORICAL VARIABLE DISTRIBUTION**

- Most of the bank customers purchased the blue Card.
- Most of the Bank Customers are graduates.
- Most of the Bank Customers are female.
- Most of the Bank Customers income are less than $40k

# 5. Modelling

- The dataset is split into train and test data. 70% of the data are set as a train data and the others are concluded as the test data.

```
set.seed(1234)
sample_set <- BankData %>%
  pull(.) %>%
  sample.split(SplitRatio = .7)

bankTrain <- subset(BankData, sample_set == TRUE)
bankTest <- subset(BankData, sample_set == FALSE)
```

```
```{r}
round(prop.table(table(bankTest$Attrition_Flag)),3)
```
```

# 6. Regression

**A. Selected variable to use for ANOVA and chi-test**.

**To see the structure of the data**

```
>glimpse(BankData)
## Rows: 10,127
## Columns: 8
$ Attrition_Flag  <fct> Existing Customer, Existing Customer, Existing Cust...
$ Customer_Age    <int> 45, 49, 51, 40, 40, 44, 51, 32, 37, 48, 42, 65, 56,...
$ Gender          <fct> M, F, M, F, M, M, M, M, M, M, M, M, M, M, F, M, M, ...
$ Education_Level <fct> High School, Graduate, Graduate, High School, Unedu...
$ Income_Category <fct> $60K - $80K, Less than $40K, $80K - $120K, Less tha...
$ Card_Category   <fct> Blue, Blue, Blue, Blue, Blue, Blue, Gold, Silver, B...
$ Months_on_book  <int> 39, 44, 36, 34, 21, 36, 46, 27, 36, 36, 31, 54, 36,...
$ Credit_Limit    <dbl> 12691.0, 8256.0, 3418.0, 3313.0, 4716.0, 4010.0, 34...
```

**Select the certain variables needed**

```
>BankData<-select(BankData, Attrition_Flag, Customer_Age, Gender, Education_Level,Income_Category,
Card_Category)
>dim(BankData)
[1] 10127     6
```

**Double check variable needed**

```
>str(BankData)
'data.frame':    10127 obs. of  6 variables:
$ Attrition_Flag : Factor w/ 2 levels "Attrited Customer",..: 2 2 2 2 2 2 2 2 2 2 ...
$ Customer_Age   : int  45 49 51 40 40 44 51 32 37 48 ...
$ Gender         : Factor w/ 2 levels "F","M": 2 1 2 1 2 2 2 2 2 2 ...
$ Education_Level: Factor w/ 7 levels "College","Doctorate",..: 4 3 3 4 6 3 7 4 6 3 ...
$ Income_Category: Factor w/ 6 levels "$120K +","$40K - $60K",..: 3 5 4 5 3 2 1 3 3 4 ...
$ Card_Category  : Factor w/ 4 levels "Blue","Gold",..: 1 1 1 1 1 1 2 4 1 1 ...
```

**B. Create a sort frequency table uses the tabyl function from the janitor package along with pipe operator which is loaded with the dplyr package.**

*This operator allows to string together a series of commands without having to create intermediate output objects.*

```
>library(janitor)
>BankData %>% tabyl(Gender,
Education_Level,Income_Category)
```

```
$`$120K +`
##  Gender College Doctorate Graduate High School Post-Graduate Uneducated Unknown
##      F      0       0      0           0              0       0      0
##      M     70      37    204         147             30     119    120
##
## $`$40K - $60K`
##  Gender College Doctorate Graduate High School Post-Graduate Uneducated Unknown
##      F    108      43    319         195             61     133    155
##      M     75      27    234         160             50     116    114
##
## $`$60K - $80K`
##  Gender College Doctorate Graduate High School Post-Graduate Uneducated Unknown
##      F      0       0      0           0              0       0      0
##      M    132      59    422         307             77     195    210
##
## $`$80K - $120K`
##  Gender College Doctorate Graduate High School Post-Graduate Uneducated Unknown
##      F      0       0      0           0              0       0      0
##      M    175      57    478         308             81     217    219
##
## $`Less than $40K`
##  Gender College Doctorate Graduate High School Post-Graduate Uneducated Unknown
##      F    319     147   1039         617            160     485    517
##      M     26      11    100          54             10      37     39
##
## $Unknown
##  Gender College Doctorate Graduate High School Post-Graduate Uneducated Unknown
##      F    105      67    312         216             42     178    140
##      M      3       3     20           9              5       7      5
```

## C. Count the number of observation within each level of factor variable.

i) Attribution Flag

```
>table(BankData$Attrition_Flag)
[1] Attrited Customer Existing Customer
               1627              8500
```

ii) Education Level

```
>table(BankData$Education_Level)
[1] College     Doctorate     Graduate   High School   Post-Graduate    Uneducated     Unknown
        1013           451         3128          2013            516            1487            1519
```

iii) Gender

```
>table(BankData$Gender)
[1]     F    M
    5358 4769
```

iv) Card Category

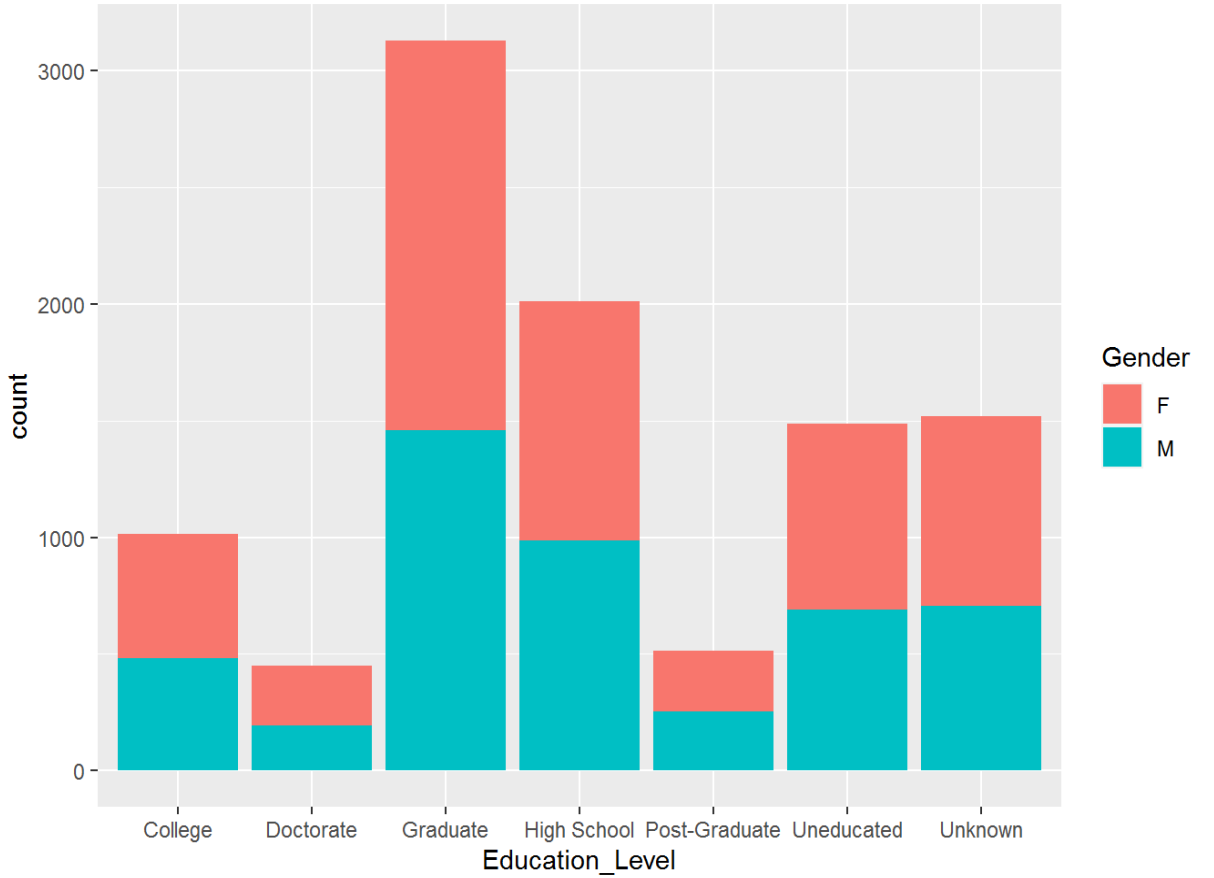
```
> table(BankData$Card_Category)
[1] Blue     Gold Platinum   Silver
       9436      116       20      555
```

v) Income Category

```
>table(BankData$Income_Category)
[1]  $120K +    $40K - $60K     $60K - $80K     $80K - $120K     Less than $40K    Unknown
        727           1790            1402             1535             3561          1112
```

Which in educational qualification of the account holder have the potential to leave the companies' services based on gender?

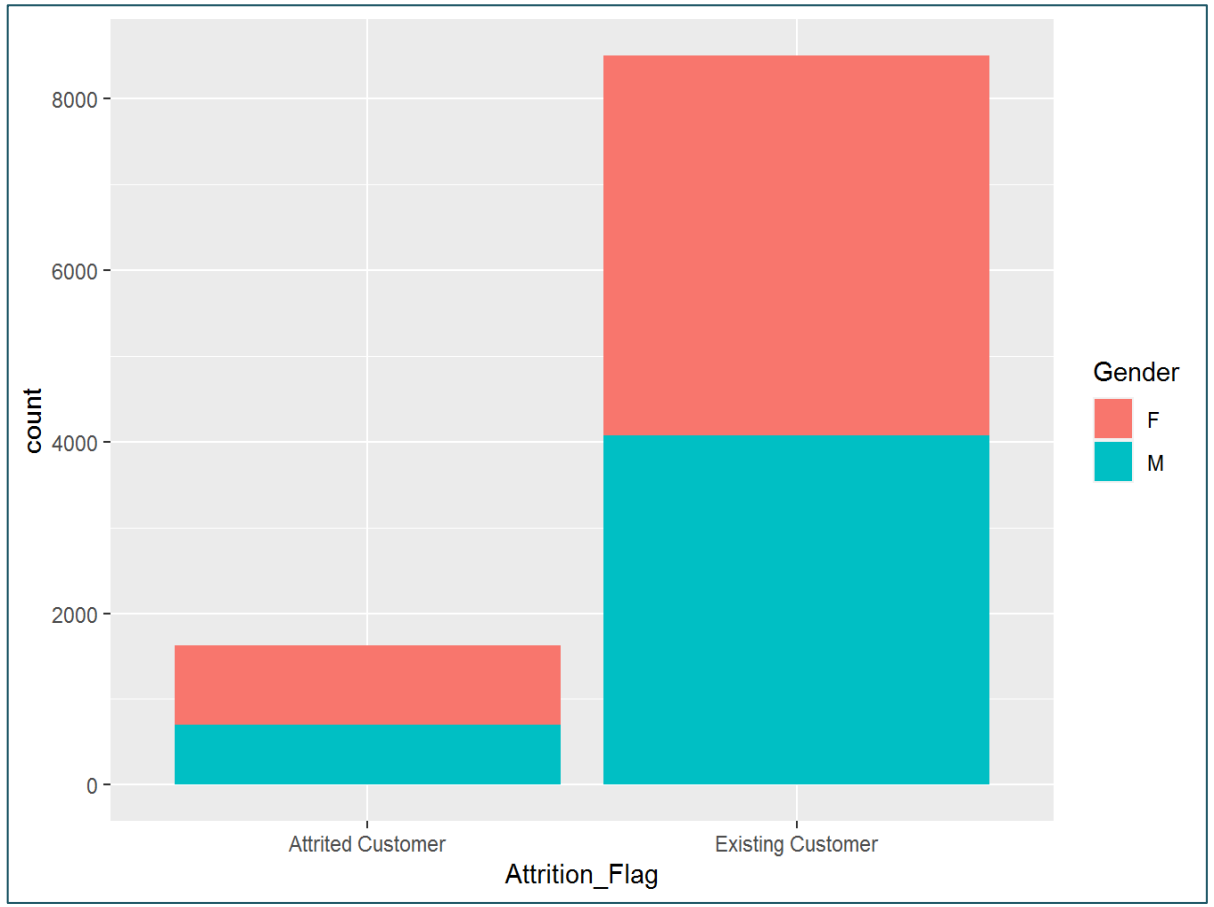
```
>p1<-ggplot(BankData, aes(x=Education_Level, fill= Gender))+
geom_bar()
>p1
```




Result: In education level for **graduate** has highest count **for both gender** which educational qualification of the account holder have the potential to leave the companies' services

Which gender for customer activity in attrition flag have the potential to leave the companies' services?

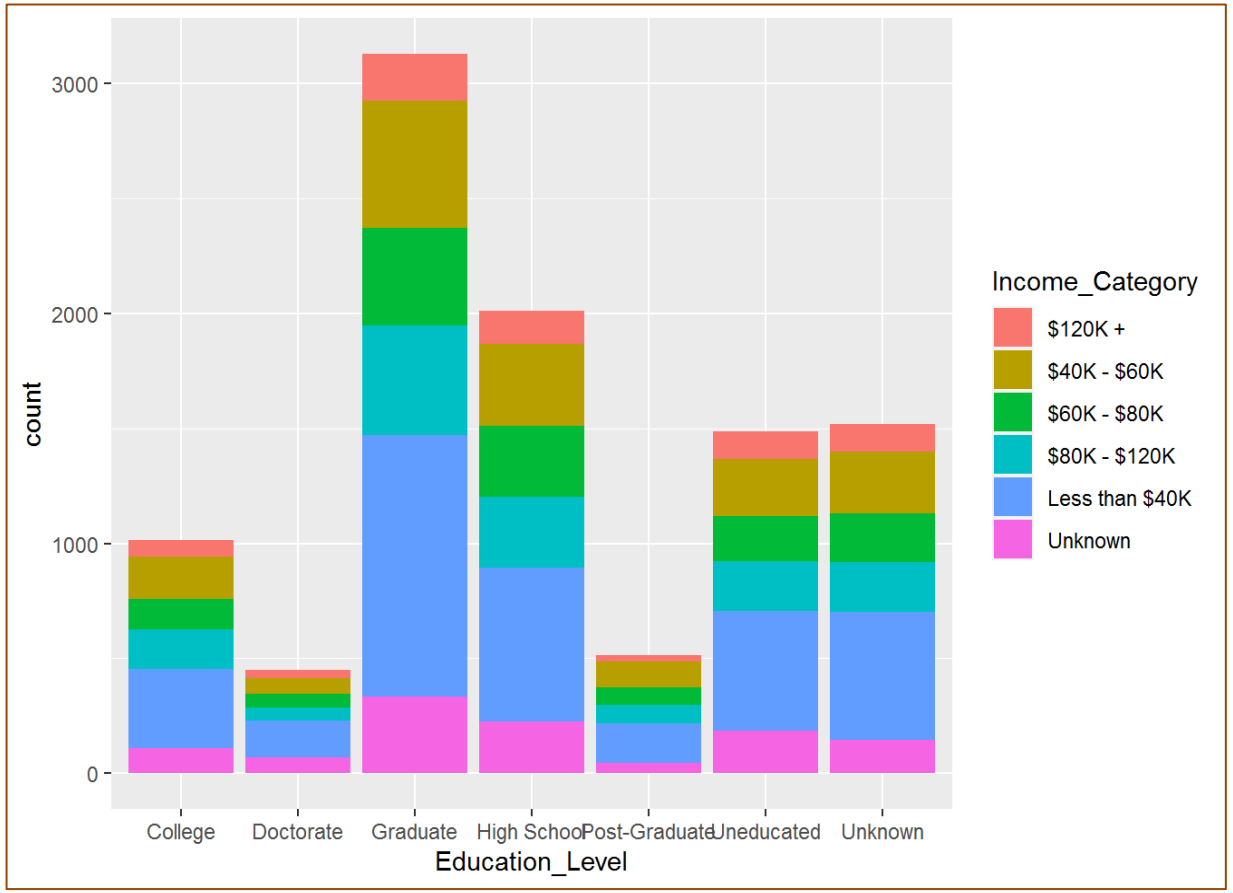
```
>p2<-ggplot(BankData, aes(x=Attrition_Flag, fill= Gender))+
geom_bar()
>p2
```




Result: **Both gender** has highest frequency in **existing customer** for customer activity likely have the potential to leave the companies' services

Which income category has the highest likely have the potential to leave the companies' services based on education level?

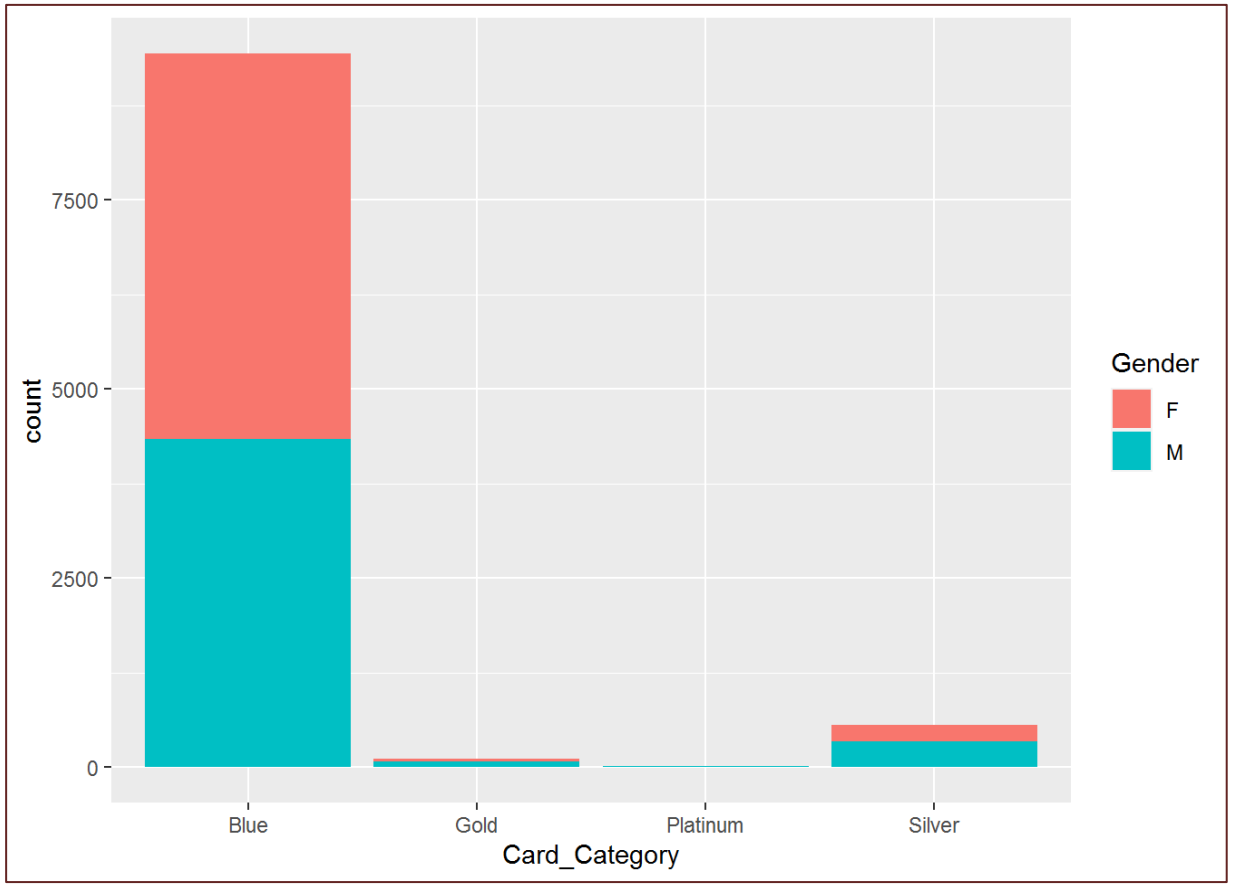
```
>p3< ggplot(BankData, aes(x=Education_Level , fill= Income_Category))+ geom_bar()
>p3
```




Result: The most likely to leave the companies for income category is **less than $40k** with education level is **graduate.**

What type of card category has the highest count use by customer based on gender?

```
p4<-ggplot(BankData, aes(x=Card_Category , fill= Gender))+ geom_bar()
p4
```




Result: **Blue** card hast the highest use for **both gender**

## D. To determine if there any significant differences between customer age on income category

i) ANOVA to test is there any significant differences between customer age on income category

```
>library(ggpubr)
>res.aov<- aov(Customer_Age~Income_Category, BankData)
>summary(res.aov)
```

```
Df Sum Sq Mean Sq F value   Pr(>F)
## Income_Category     5   1490  298.01   4.645 0.000309 ***
## Residuals       10121 649301   64.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

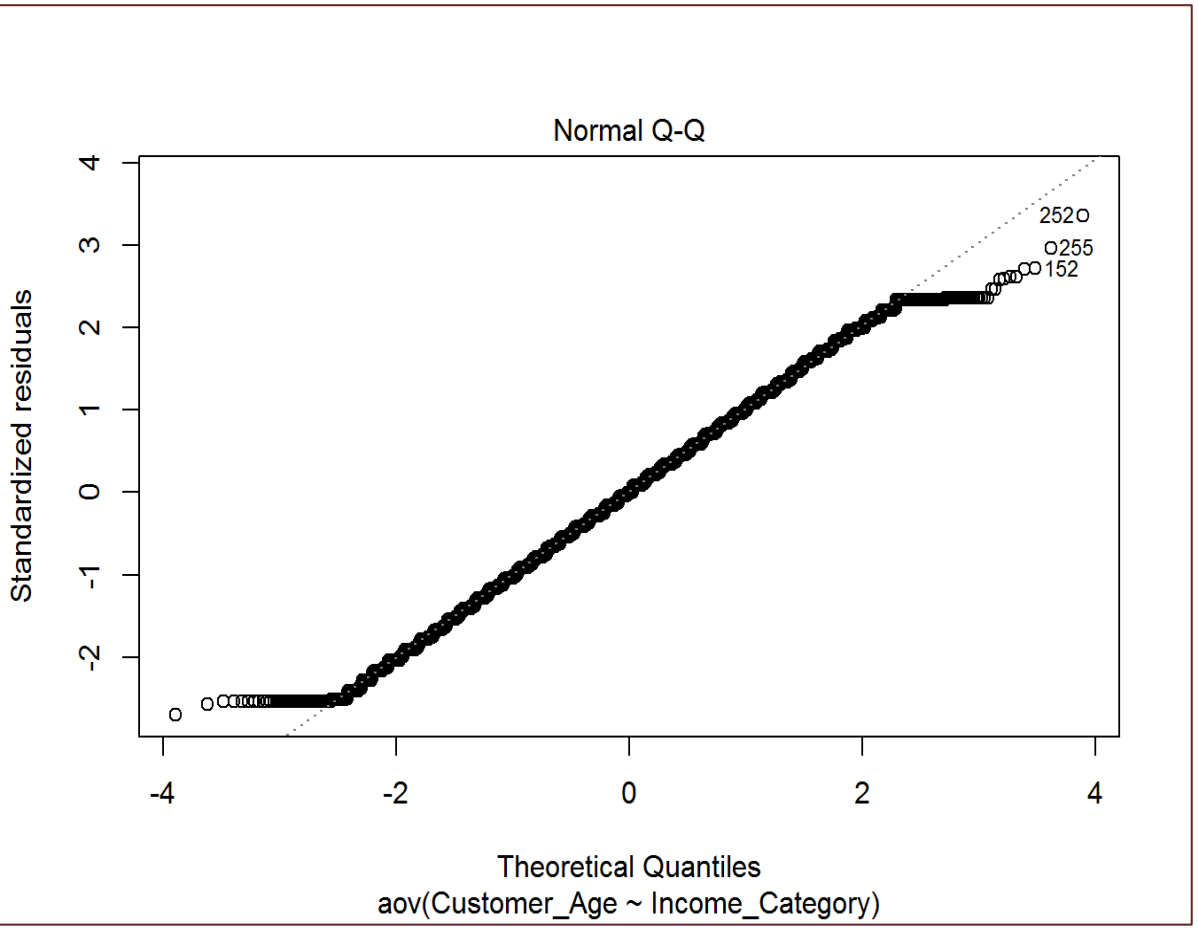
**Hypothesis Test:**
H0: There is no significant difference between customer age on income category
H1: There is significant difference between customer age on income category

p-value=0.000309<0.05, reject H0.

**At alpha =.0.05, there is enough evidence to reject H0. Therefore, there is significant difference between customer age on income category**

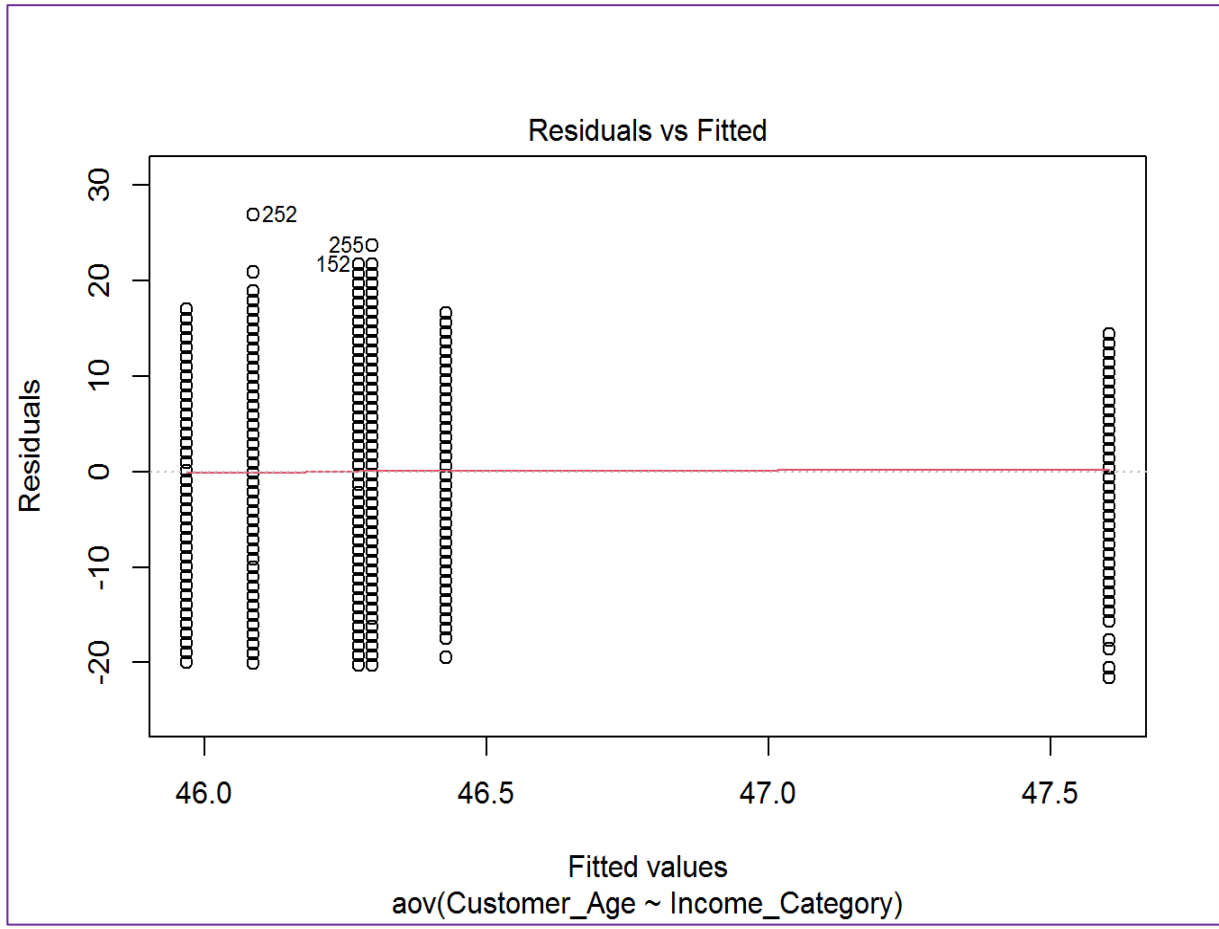
## E. Normality

>plot(res.aov, 2)




## F. Homogeneity of variances

>plot(res.aov, 1)




**Result: Has 3 outliers which are 152, 252, and 255**

```
> library(car)
>leveneTest(Customer_Age~Income_Category, BankData)

[1] Levene's Test for Homogeneity of Variance (center = median)
Df F value    Pr(>F)
group     5  22.091 < 2.2e-16 ***
10121
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Hypothesis Test**
H0: The variance among all groups is equal
H1: The variance among all groups is not equal

p-value = 2.2e-16 *** < 0.05

At alpha =.0.05, there is enough evidence to reject H0.
Therefore, not all variance among all groups is equal.

## G. To determine the proportion of income category on education level

### ii)Finding proportion For Both Categorical Variables

```
>chisq.test(BankData$Income_Category, BankData$Education_Level)

[1]Pearson's Chi-squared test

data:  BankData$Income_Category and BankData$Education_Level
X-squared = 45.254, df = 30, p-value = 0.03655
```

**Hypothesis Test**
H0: Income category are independent on the education level H1: Income category are dependent on the education level

$\chi 2 = 45.254 < 0.05$, reject H0.

At alpha =.0.05, there is enough evidence to reject H0. Therefore, income category are dependent on the education level.

# 7. Classification

- **DECISION TREE**

**Decision tree splits the data into multiple sets and each set is further split into subsets until a decision is made.**

```
Confusion Matrix and Statistics

                    Reference
Prediction           Attrited Customer Existing Customer
  Attrited Customer                 30                73
  Existing Customer                472              2464

               Accuracy : 0.8207
                 95% CI : (0.8066, 0.8342)
    No Information Rate : 0.8348
    P-Value [Acc > NIR] : 0.9824

                  Kappa : 0.0455

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.059761
            Specificity : 0.971226
         Pos Pred Value : 0.291262
         Neg Pred Value : 0.839237
             Prevalence : 0.165186
         Detection Rate : 0.009872
   Detection Prevalence : 0.033893
      Balanced Accuracy : 0.515493

       'Positive' Class : Attrited Customer
```

# 7. Classification

- **EXTREME GRADIENT BOOSTING**

**Extreme Gradient Boosting algorithm can be used for supervised learning task such as regression, classification and ranking.**

**It produces a prediction model in the form of an ensemble of weak decision trees.**

**It is a more regularized model formalization to control over-fitting, which gives a better performance and completes in a high speed.**

```
Confusion Matrix and Statistics

                    Reference
Prediction           Attrited Customer Existing Customer
  Attrited Customer                 73                22
  Existing Customer                429              2515

               Accuracy : 0.8516
                 95% CI : (0.8385, 0.8641)
    No Information Rate : 0.8348
    P-Value [Acc > NIR] : 0.006279

                  Kappa : 0.2026

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.14542
            Specificity : 0.99133
         Pos Pred Value : 0.76842
         Neg Pred Value : 0.85428
             Prevalence : 0.16519
         Detection Rate : 0.02402
   Detection Prevalence : 0.03126
      Balanced Accuracy : 0.56837

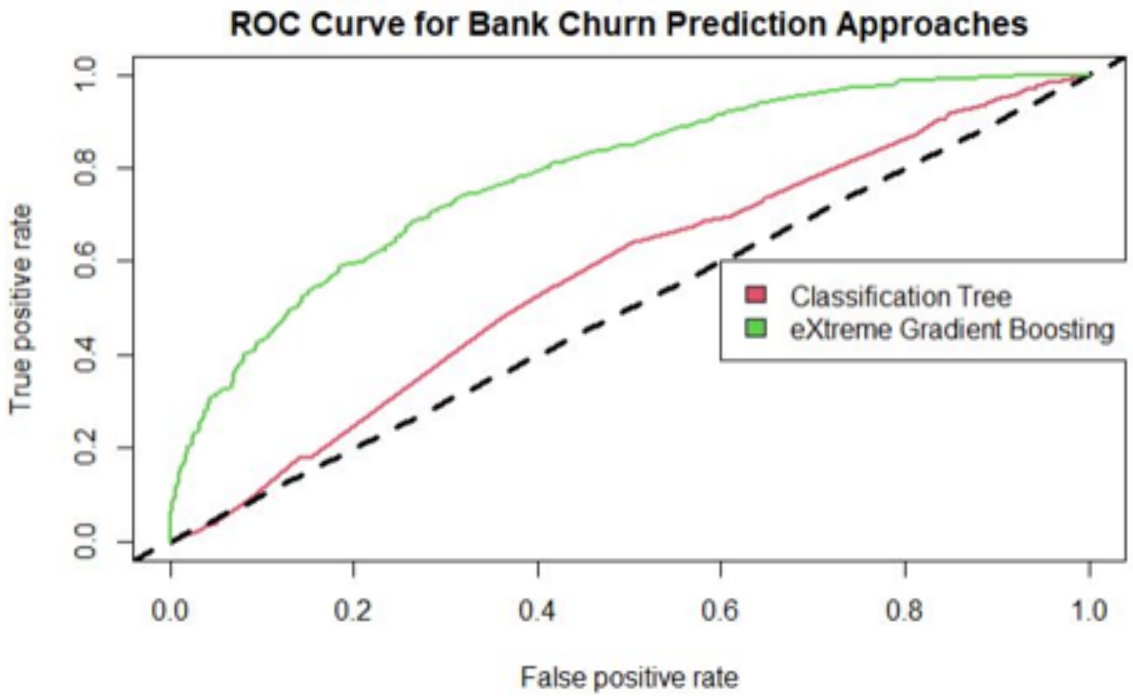
       'Positive' Class : Attrited Customer
```

# 8. Comparing the Classification Model Performance

- Receiver Operating Characteristics (ROC) curve is used as an evaluation metrics for checking classification model's performance.
- It tells how much a model can distinguish between the classes.

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{FPR} = 1 - \text{Specificity} = \frac{FP}{TN + FP}$$



# 8. Comparing the Classification Model Performance

As a conclusion, we can see that:

1. EXtreme Gradient Boosting provides a **better percentage accuracy** that Classification Tree.
2. The classification accuracy or **kappa** for eXtreme Gradient Boosting is **0.16 higher** than Classification Tree.
3. The **AUC Curve** (measure the usefulness of a test) results is better in extreme Gradient Boosting than Classification Tree.
4. The **fmeasure (F-Score)** performs better in eXtreme Gradient Boosting than Classification Tree.

| approach | accuracy | fmeasure | kappa | auc |
|---|---|---|---|---|
| Classification Tree | 0.8206647 | 0.0991736 | 0.0454868 | 0.5709727 |
| eXtreme Gradient Boosting | 0.8515959 | 0.2445561 | 0.2026374 | 0.7839419 |

- AUC curve 0.7 means there is 70% chance that the model can distinguish between the (+) and (-) classes.
- AUC curve 0.5 means the model has no discrimination to distinguish the classes.

# 9. Conclusion and Future Work

- The result shows the customer age and income category having p-value less than 0.05 using ANOVA.
- The p-value of income category and education level is less than 0.05 using chi square test.
- The accuracy of Extreme Gradient Boosting is higher than Classification tree.
- The future work can be improved by implementing more machine learning algorithms to predict with a better accuracy result.