BankChurners Report

Wuji Shan

12/01/2021

Abstract

Credit card has taken up a significant part of people's lives in the current society. The "Credit Card Customers" dataset is data of the consumer credit card portfolio of a bank, including 10,000 customers with their age, gender, income, marital status, education level, and etc. The manager wants to know the reason behind customer attrition.

To deal with this problem, I made exploratory data analysis and built a multilevel model. This report are consisted of 5 main parts: Introduction, Method, Result and Discussion. Other explorations besides what in Method part are all put in Appendix.

Introduction

Customer churning is a serious situation faced by many corporations and organizations. It has become a significant point that how to deal with this issue and keep customers. To achieve that goal, one of the main focuses of corporations may be detecting the reasons why customers have made the churning decision via analyzing past data, so that they could take effective actions to prevent the customer leaving situation better.

The dataset I choose for this project is published on Kaggle: Credit Card customers – BankChurners Dataset, which includes 10127 observations and 23 variables. After data cleaning and processing, I will make exploratory data analysis from various angles, and use a multilevel model to see what and how attributes may influence the choice of customers to churn from the bank credit card service.

Method

Data Cleaning and Processing

Look into the data and finished the following processing steps:

1. Subset down to 10 out of 23 variables;
2. Removed customer observations with "unknown" answer of categorical variables;
3. Split the data into 2 groups based on customer type: Existing and Attrited Customer.

After my data cleaning and processing, the dataset utilized later has 7081 observations and 10 variables besides customer type.

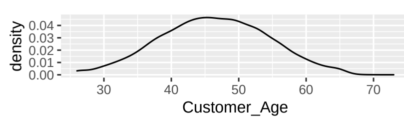
10 Variables Utilized:

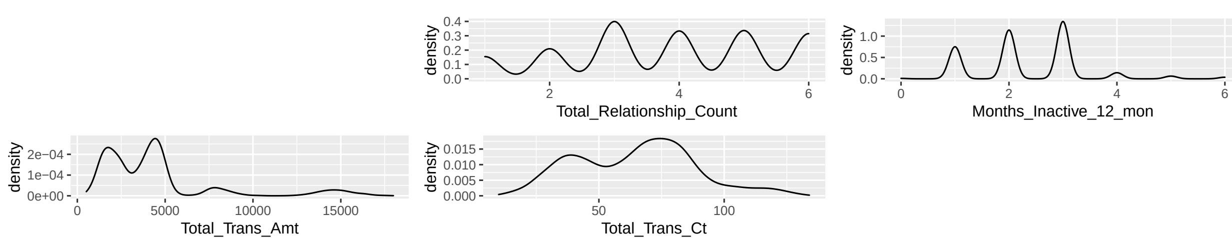
- Customer_Age: Customer's Age in Years
- Gender: M = Male, F = Female
- Education_Level: Educational Qualification of the account holder
- Marital_Status: Married, Single, Divorced, Unknown
- Income_Category: Annual Income Category of the account holder
- Card_Category: Type of Card (Blue, Silver, Gold, Platinum)
- Total_Relationship_Count: Total no. of products held by the customer
- Months_Inactive_12_mon: Number of months inactive in the last 12 months
- Total_Trans_Amt: Total Transaction Amount (Last 12 months)
- Total_Trans_Ct: Total Transaction Count (Last 12 months)

Exploratory Data Analysis

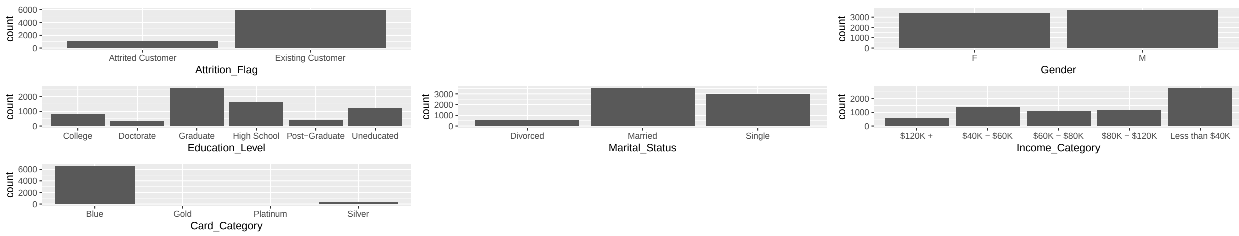
```r
# Create theme parameters
theme <- theme_bw() +
      theme(plot.title = element_text(face = "bold", color = "black", size=14),
        plot.subtitle = element_text(face = "italic", color = "black", size=12),
        axis.text = element_text(color = "black"), legend.text = element_text(size=10),
        legend.title = element_text(size = 12), legend.position = "none",
        strip.background =element_rect(fill="#666666"), strip.text = element_text(color="white", face="
        plot.caption = element_text(face = "italic"))
```

Distributions of all numeric variables





Distributions of all categorical variables



Pie Charts of Attrition_Flag

a. Pie charts of Gender Proportion comparison for Existing and Attrited customers

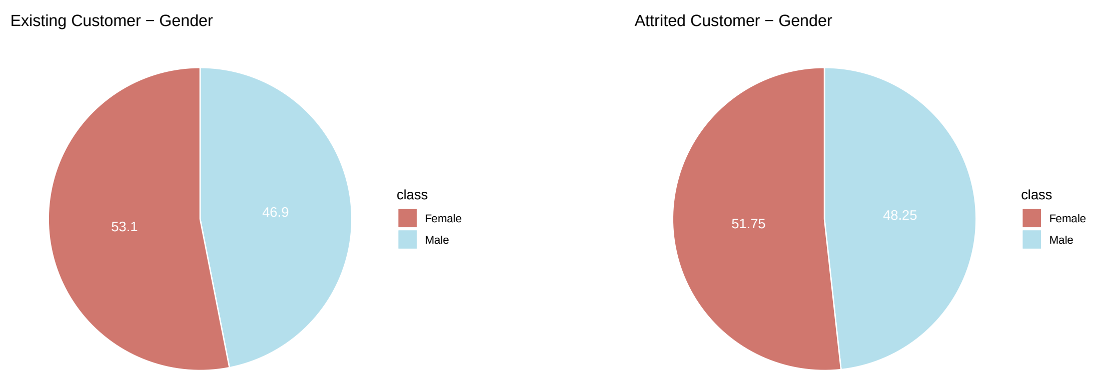


Figure 1: Gender Proportion Comparison

b. Pie charts of Education Level Proportion comparison for Existing and Attrited customers

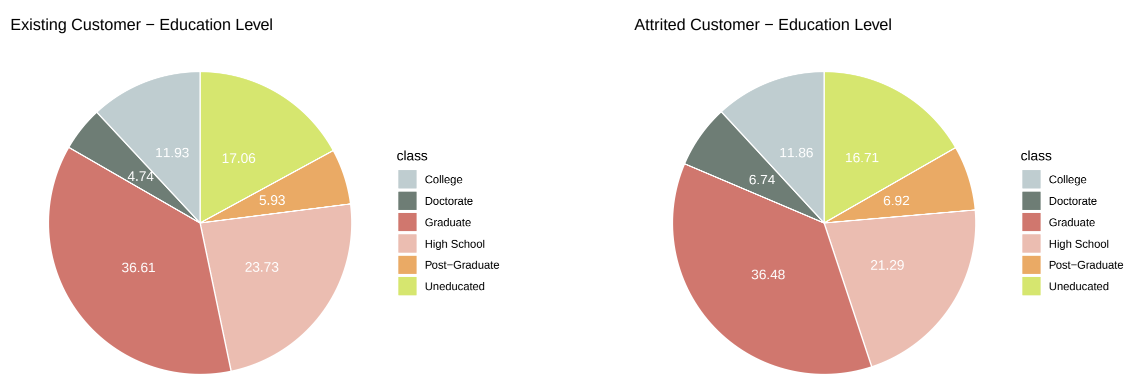


Figure 2: Education Level Proportion Comparison

Figure 2:

Assuming that customers with “unknown” education level did not receive any education, we can observe that more than 70% of the customers have a formal education level for both existing and attrited customers. Moreover, about 40% have a higher level of education for two groups.

c. Pie charts of Marital Status Proportion comparison for Existing and Attrited customers

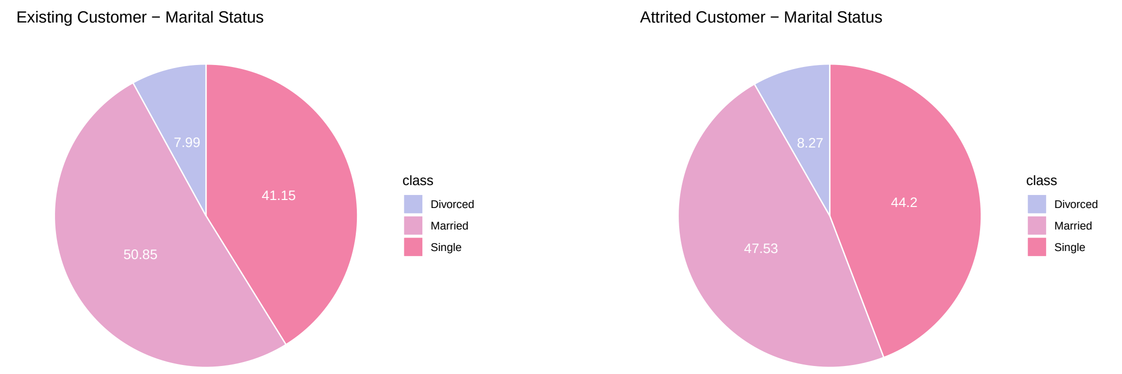


Figure 3: Marital Status Proportion Comparison

Figure 3:

“Almost half of the bank customers are married, and interestingly enough, almost the entire other half are single customers. only about 7% of the customers are divorced, which is surprising considering the worldwide divorce rate statistics!”

The proportion of married status in attrited customers is slightly smaller than that in existing customers; correspondingly, the proportion of single status in attrited customers is slightly larger than that in existing customers.

d. Pie charts of Income Category Proportion comparison for Existing and Attrited customers

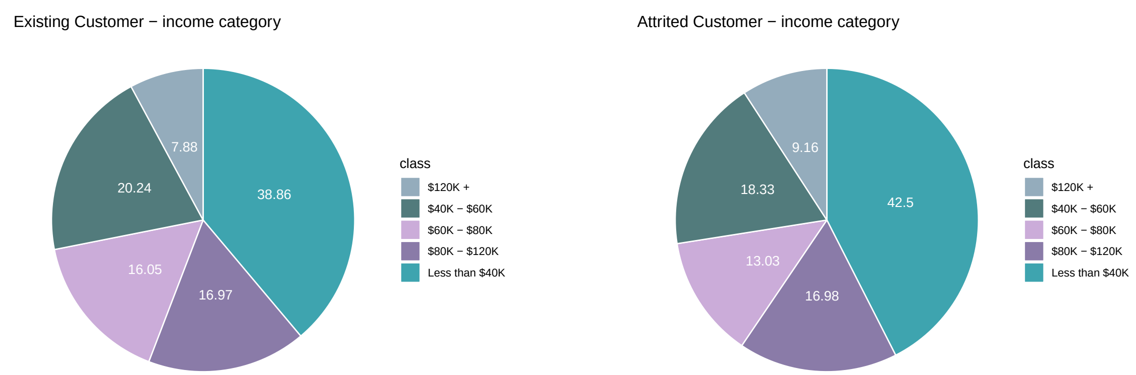


Figure 4: Income Category Proportion Comparison

Income Category

Card Category

“Platinum cards have the highest attrition rates.”

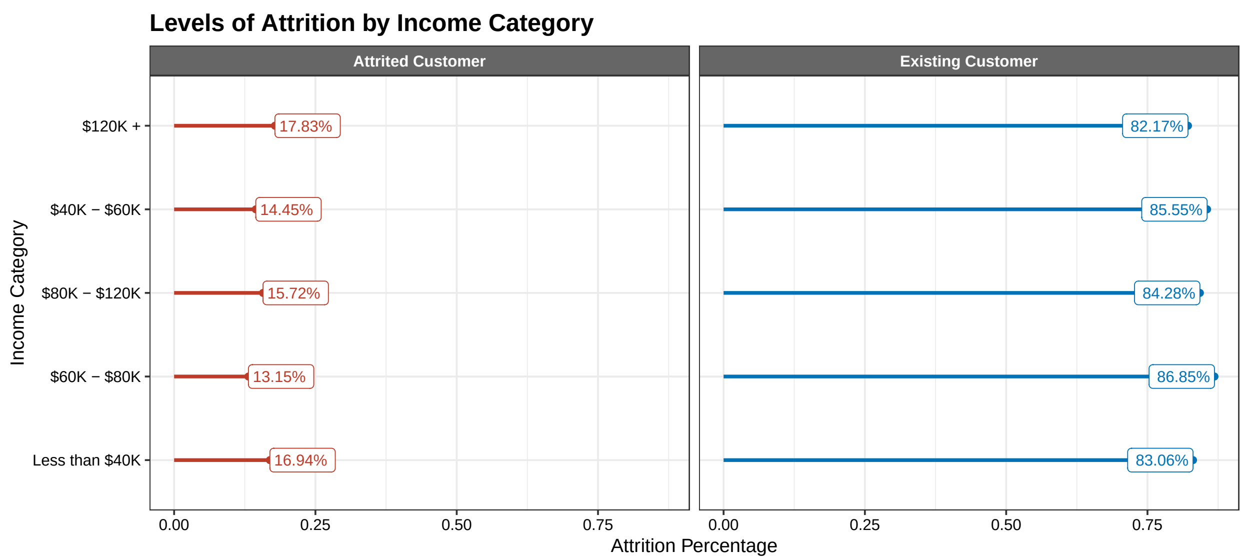


Figure 5: Income Category Proportion Analysis

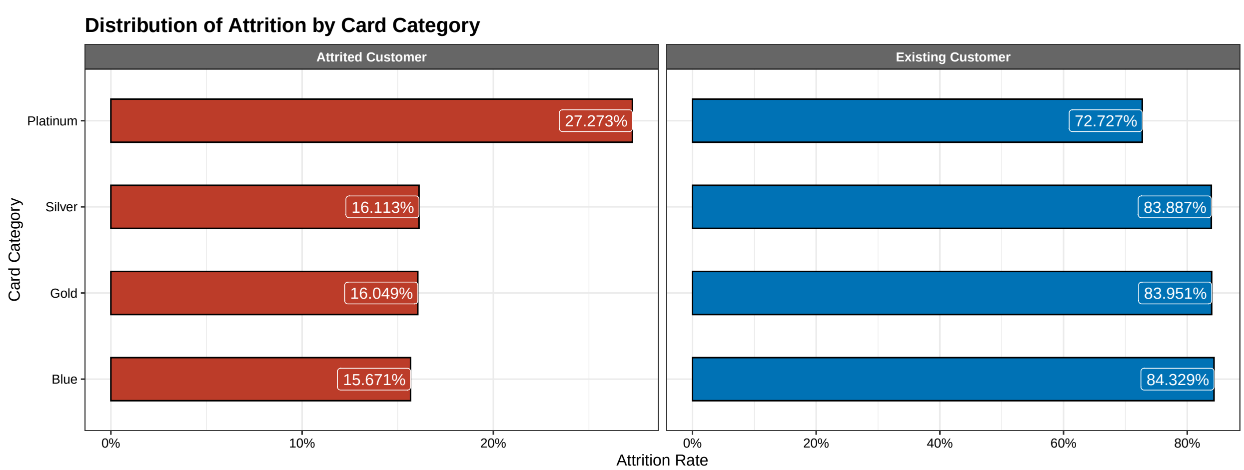


Figure 6: Card Category Proportion Analysis

Level of Inactivity

```
## Picking joint bandwidth of 0.142
```

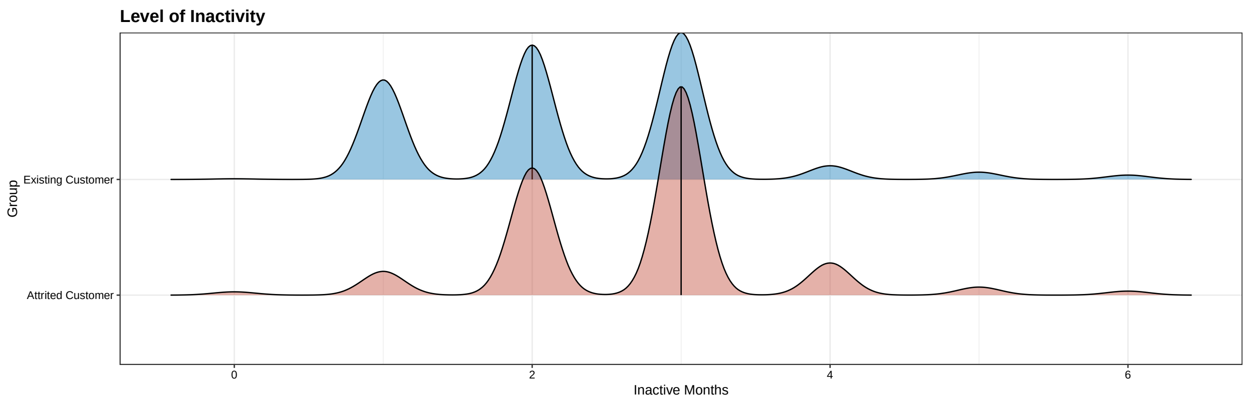


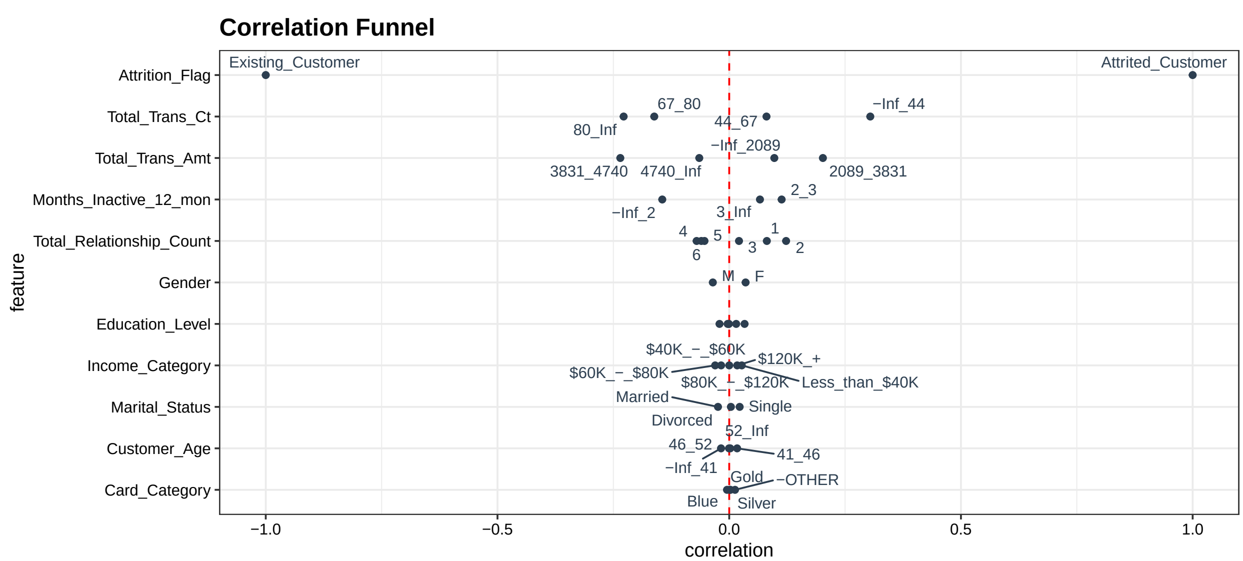
Figure 7: Level of Inactivity

"Attrition Customers have higher levels of Inactivity (3 months vs 2 months Median)."

Total Customer Transactions

Correlation Analysis

```
## Warning: ggrepel: 6 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



The most important features are towards the top. We can investigate these.

```
## Warning: ggrepel: 2 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

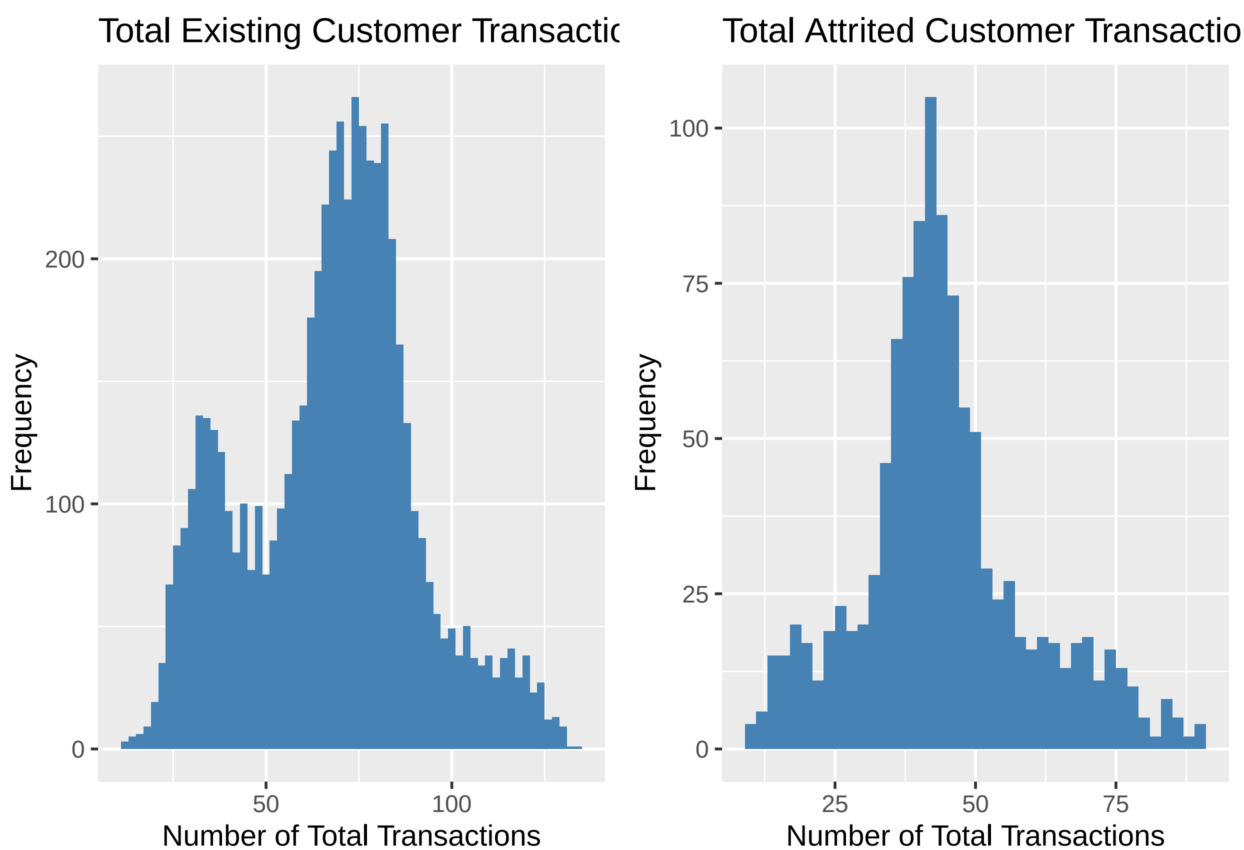
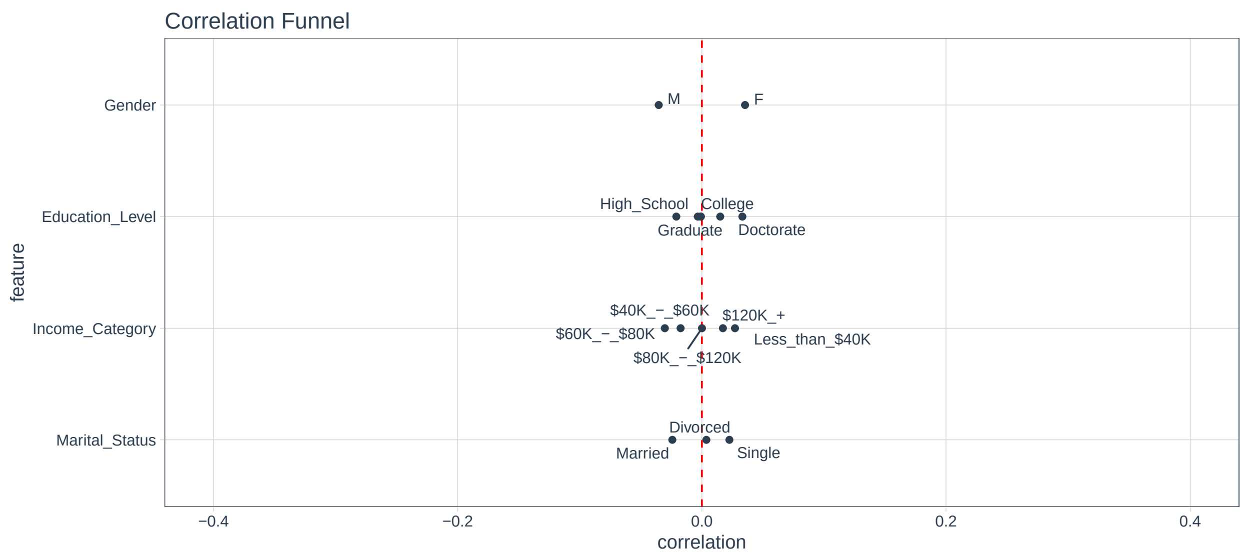


Figure 8: Total Customer Transactions Comparison



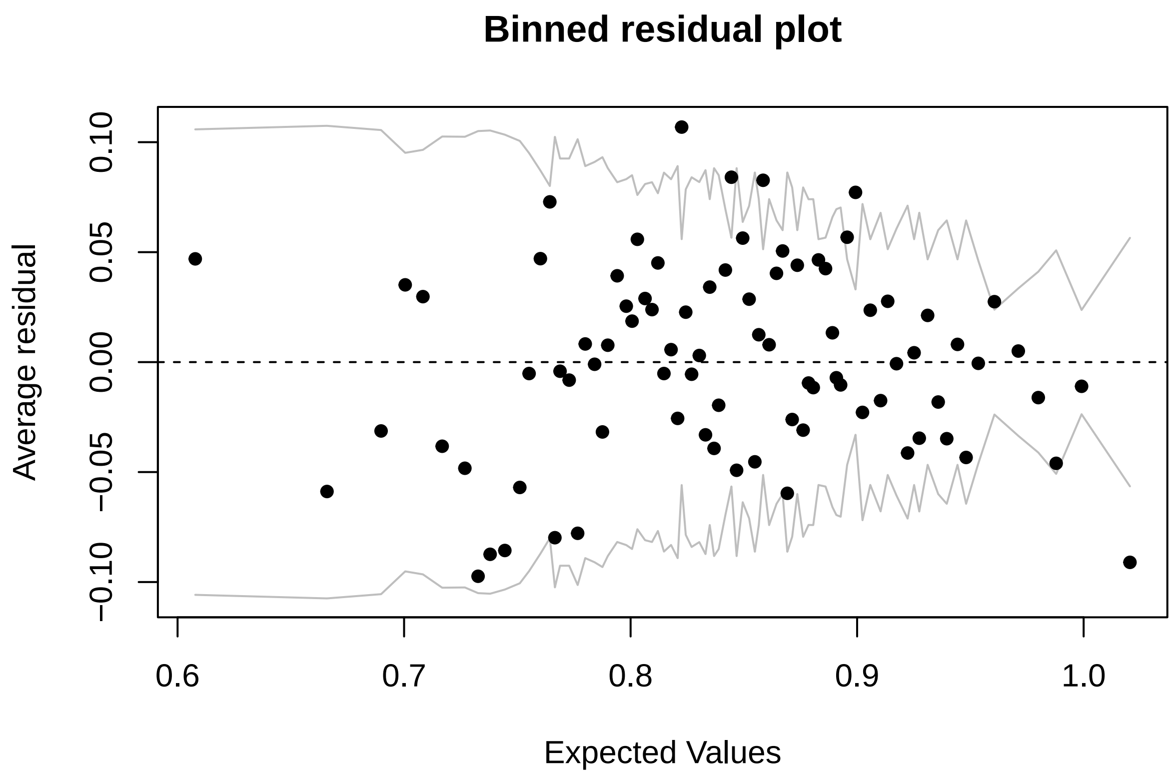
Correlation Funnel
feature
Gender
Education_Level
Income_Category
Marital_Status
M
F
High_School
College
Graduate
Doctorate
$40K_−_$60K
$120K_+
$60K_−_$80K
Less_than_$40K
$80K_−_$120K
Divorced
Married
Single
−0.4
−0.2
0.0
0.2
0.4
correlation

Model Fitting

```
## Warning in glmer(Attrition_Flag ~ Customer_Age + Education_Level +
## Income_Category + : calling glmer() with family=gaussian (identity link) as a
## shortcut to lmer() is deprecated; please call lmer() directly

## boundary (singular) fit: see ?isSingular

## Linear mixed model fit by REML ['lmerMod']
## Formula: Attrition_Flag ~ Customer_Age + Education_Level + Income_Category +
##     Card_Category + Total_Relationship_Count + Months_Inactive_12_mon +
##     (1 + Gender | Marital_Status)
##    Data: BankChurners
##
## REML criterion at convergence: 5546.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9637  0.1515  0.3606  0.5308  1.4682
##
## Random effects:
##  Groups         Name        Variance Std.Dev. Corr
##  Marital_Status (Intercept) 0.000000 0.00000
##                 GenderM     0.000832 0.02884   NaN
##  Residual                   0.126264 0.35534
## Number of obs: 7081, groups:  Marital_Status, 3
##
## Fixed effects:
##                                Estimate Std. Error t value
## (Intercept)                   0.7858415  0.0362490  21.679
## Customer_Age                  0.0002542  0.0005280   0.481
## Education_LevelDoctorate     -0.0519156  0.0224449  -2.313
## Education_LevelGraduate      -0.0019473  0.0140908  -0.138
## Education_LevelHigh School    0.0086636  0.0150448   0.576
## Education_LevelPost-Graduate -0.0286569  0.0210645  -1.360
## Education_LevelUneducated    -0.0003882  0.0159652  -0.024
## Income_Category$40K - $60K    0.0466172  0.0188472   2.473
## Income_Category$60K - $80K    0.0450547  0.0183545   2.455
## Income_Category$80K - $120K   0.0181771  0.0180916   1.005
## Income_CategoryLess than $40K 0.0348467  0.0194935   1.788
## Card_CategoryGold             0.0224438  0.0398957   0.563
## Card_CategoryPlatinum        -0.0486505  0.1074790  -0.453
## Card_CategorySilver           0.0054833  0.0186161   0.295
## Total_Relationship_Count      0.0345164  0.0027464  12.568
## Months_Inactive_12_mon       -0.0560731  0.0042547 -13.179

##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it

## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
## (Intercept)
##       0.786
```

```
##          (Intercept) GenderM
## Divorced           0   0.011
## Married            0   0.037
## Single             0   0.018
```

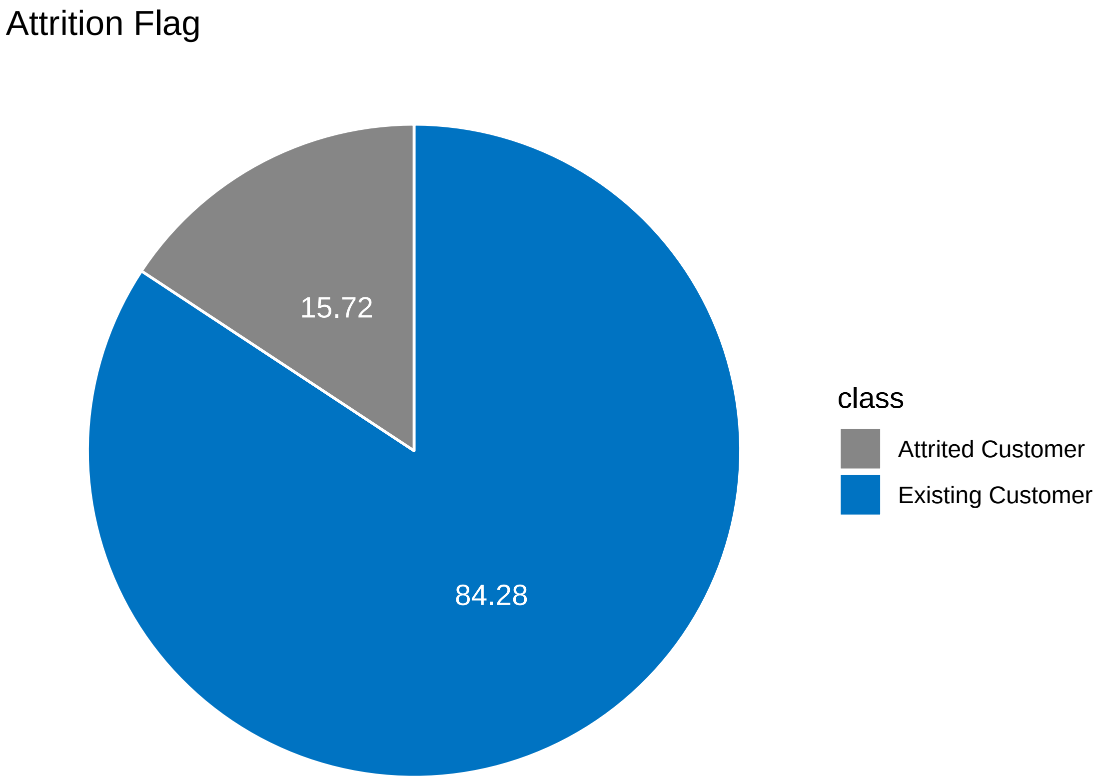


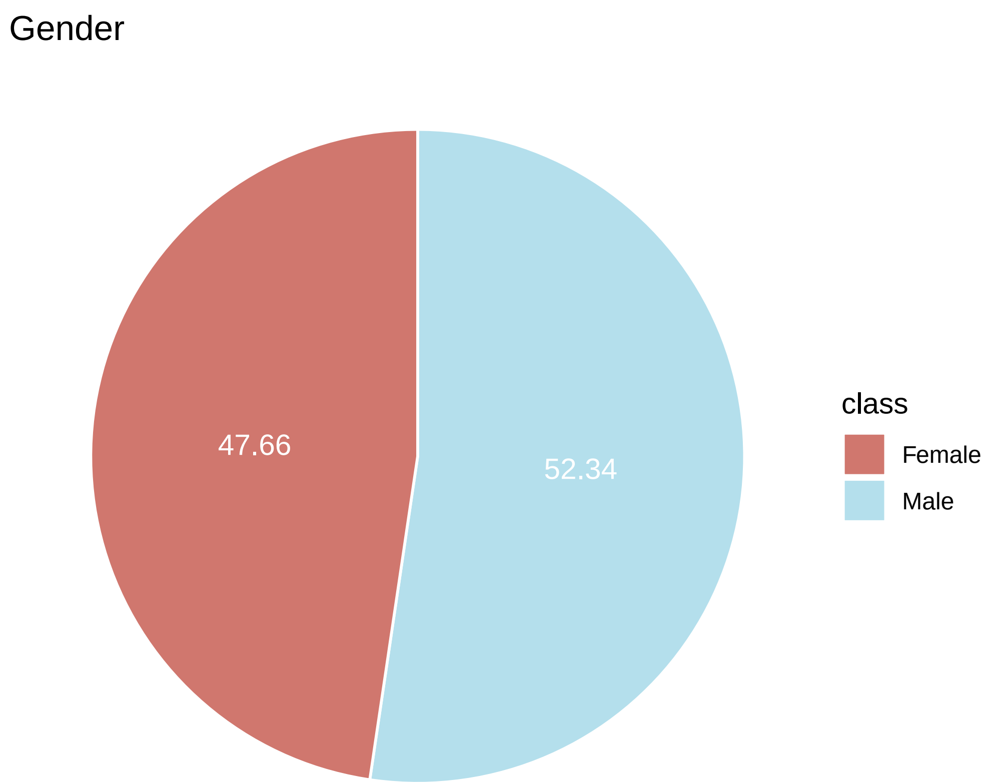
Reference

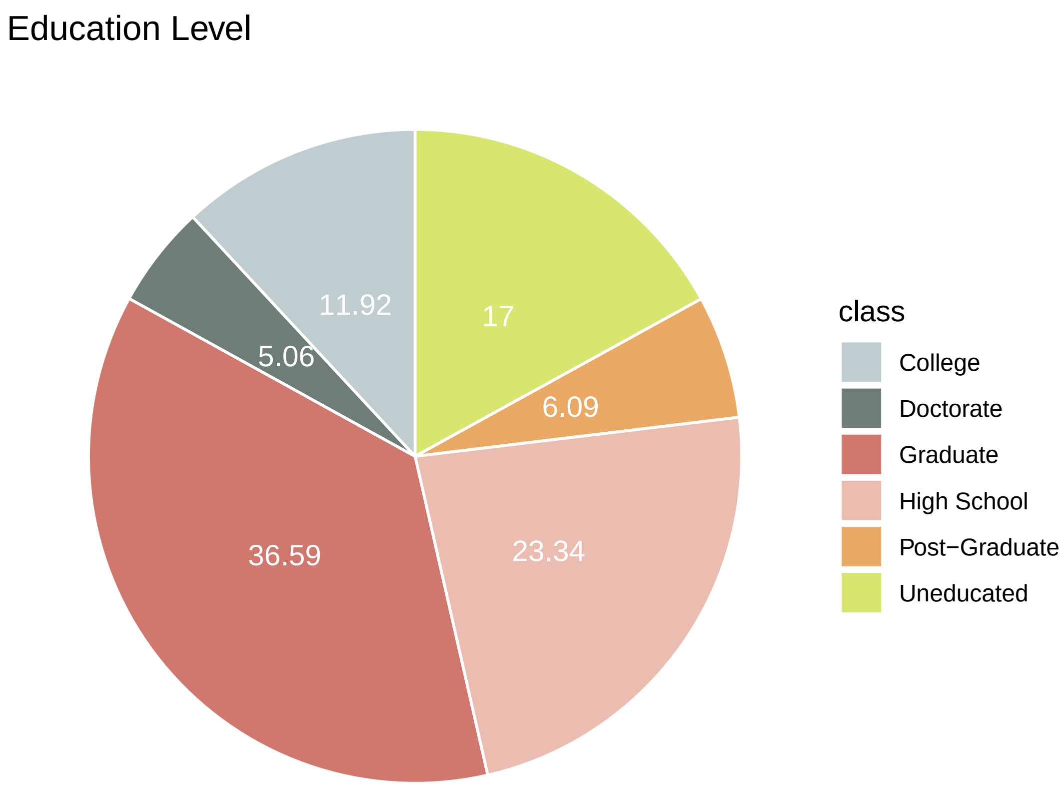
1. https://cran.r-project.org/web/packages/correlationfunnel/vignettes/introducing_correlation_funnel.html

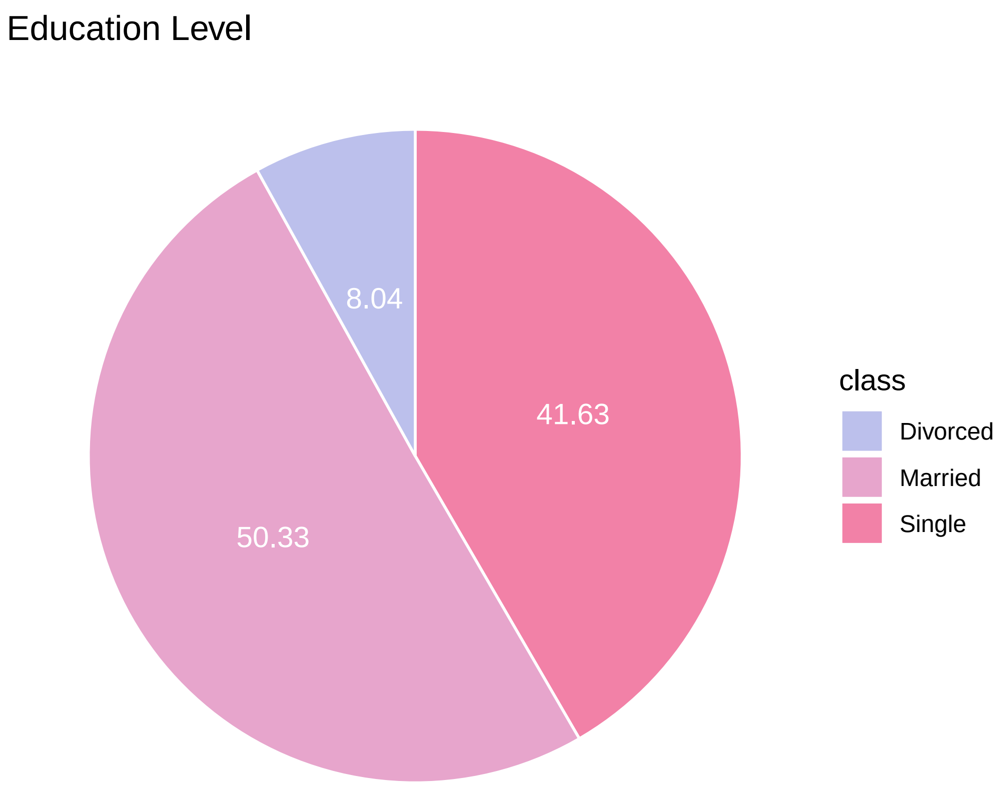
Appendix

Showing pie charts of overall analysis of the whole data.




Gender
47.66
52.34
class
Female
Male


Education Level
11.92
17
5.06
6.09
36.59
23.34
class
College
Doctorate
Graduate
High School
Post−Graduate
Uneducated


Education Level
8.04
41.63
50.33
class
Divorced
Married
Single