EDA for Final Poster (Team 8)

Team Members: Joseph Gallegos, Ajay Kallepalli, Lyndon Liang, Charles Liu, Anshuman Mahalley

2/27/2021

Questions to Answer:

1. What is the issue or overal topic that interests you? [Issue (broad)]
2. What is a specific research question that you want to investigate [Research Question (narrower)]
3. What is the problem that makes your question worth investigating? Is this an actual problem or an assumed problem? [Underlying Problem(s)]
4. Does your problem have social significance? [Social Significance]
5. What is your proposal for addressing this problem? Is your proposal both arguable & feasible? [Proposal/Solution]

Answers to Questions:

1. Data, demographics, and churn of credit card users
2. Is there a way for banks to minimize their customer churn and if so what specific areas or demographics should the company focus on.
3. Banks will from time to time have customers/clients leave their credit company for another credit company. The problem is trying to figure what are some causations/correlations that would make these customers/clients leave the credit company. Companies can then use this information to better focus their limited resources on clients that appear likely to churn and minimize churn rate (actual problem).
4. There are many areas of social significance that may benefit from such an analysis. One example is the ability to identify underserved demographics that have high churn in order to help improve their relationship with credit. In general this analysis would be useful in the areas of consumer debt, finance, and customs.
5. We propose using the variables given to us in this dataset to possibly narrow down the reason(s) and/or probability of customers leaving a credit company. We will look into it by using statistical methods. This proposal is extremely wide ranging and does leave room to be argued, but it is an overall feasible solution.

Note: 16.07% of customers churned in this data, with 1627 attrited customers and 8500 existing customers.

Loading Necessary Packages & Setting Working Directory

```
setwd(getwd())
library(readr)
library(tidyverse)
```

```r
library(DT)
library(knitr)
library(lubridate)
library(ggthemes)
library(tidytext)
library(wordcloud)
library(RColorBrewer)
```

Loading the Dataset

Credit Card Cancellation

```r
bank <- read_csv("BankChurners.csv")[,-22:-23]
```

We are going to ignore the last 2 columns of the csv dataset since they are not relevant. They both use Naive Bayes Classifier, which is not what we need for our analysis.

Preliminary Analysis (Application)

```r
# Take an initial look at the data
glimpse(bank)
```

```
## Rows: 10,127
## Columns: 21
## $ CLIENTNUM                <dbl> 768805383, 818770008, 713982108, 769911858...
## $ Attrition_Flag           <chr> "Existing Customer", "Existing Customer", ...
## $ Customer_Age             <dbl> 45, 49, 51, 40, 40, 44, 51, 32, 37, 48, 42...
## $ Gender                   <chr> "M", "F", "M", "F", "M", "M", "M", "M", "M...
## $ Dependent_count          <dbl> 3, 5, 3, 4, 3, 2, 4, 0, 3, 2, 5, 1, 1, 3, ...
## $ Education_Level          <chr> "High School", "Graduate", "Graduate", "Hi...
## $ Marital_Status           <chr> "Married", "Single", "Married", "Unknown",...
## $ Income_Category          <chr> "$60K - $80K", "Less than $40K", "$80K - $...
## $ Card_Category            <chr> "Blue", "Blue", "Blue", "Blue", "Blue", "B...
## $ Months_on_book           <dbl> 39, 44, 36, 34, 21, 36, 46, 27, 36, 36, 31...
## $ Total_Relationship_Count <dbl> 5, 6, 4, 3, 5, 3, 6, 2, 5, 6, 5, 6, 3, 5, ...
## $ Months_Inactive_12_mon   <dbl> 1, 1, 1, 4, 1, 1, 1, 2, 2, 3, 3, 2, 6, 1, ...
## $ Contacts_Count_12_mon    <dbl> 3, 2, 0, 1, 0, 2, 3, 2, 0, 3, 2, 3, 0, 3, ...
## $ Credit_Limit             <dbl> 12691.0, 8256.0, 3418.0, 3313.0, 4716.0, 4...
## $ Total_Revolving_Bal      <dbl> 777, 864, 0, 2517, 0, 1247, 2264, 1396, 25...
## $ Avg_Open_To_Buy          <dbl> 11914.0, 7392.0, 3418.0, 796.0, 4716.0, 27...
## $ Total_Amt_Chng_Q4_Q1     <dbl> 1.335, 1.541, 2.594, 1.405, 2.175, 1.376, ...
## $ Total_Trans_Amt          <dbl> 1144, 1291, 1887, 1171, 816, 1088, 1330, 1...
## $ Total_Trans_Ct           <dbl> 42, 33, 20, 20, 28, 24, 31, 36, 24, 32, 42...
## $ Total_Ct_Chng_Q4_Q1      <dbl> 1.625, 3.714, 2.333, 2.333, 2.500, 0.846, ...
## $ Avg_Utilization_Ratio    <dbl> 0.061, 0.105, 0.000, 0.760, 0.000, 0.311, ...
```

```r
summary(bank)
```

```
##    CLIENTNUM         Attrition_Flag      Customer_Age      Gender
##  Min.   :708082083   Length:10127       Min.   :26.00   Length:10127
##  1st Qu.:713036770   Class :character   1st Qu.:41.00   Class :character
##  Median :717926358   Mode  :character   Median :46.00   Mode  :character
##  Mean   :739177606                      Mean   :46.33
##  3rd Qu.:773143533                      3rd Qu.:52.00
##  Max.   :828343083                      Max.   :73.00
##  Dependent_count Education_Level    Marital_Status     Income_Category
##  Min.   :0.000   Length:10127       Length:10127       Length:10127
##  1st Qu.:1.000   Class :character   Class :character   Class :character
##  Median :2.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2.346
##  3rd Qu.:3.000
##  Max.   :5.000
##  Card_Category      Months_on_book  Total_Relationship_Count
##  Length:10127       Min.   :13.00   Min.   :1.000
##  Class :character   1st Qu.:31.00   1st Qu.:3.000
##  Mode  :character   Median :36.00   Median :4.000
##                     Mean   :35.93   Mean   :3.813
##                     3rd Qu.:40.00   3rd Qu.:5.000
##                     Max.   :56.00   Max.   :6.000
##  Months_Inactive_12_mon Contacts_Count_12_mon  Credit_Limit
##  Min.   :0.000          Min.   :0.000         Min.   : 1438
##  1st Qu.:2.000          1st Qu.:2.000         1st Qu.: 2555
##  Median :2.000          Median :2.000         Median : 4549
##  Mean   :2.341          Mean   :2.455         Mean   : 8632
##  3rd Qu.:3.000          3rd Qu.:3.000         3rd Qu.:11068
##  Max.   :6.000          Max.   :6.000         Max.   :34516
##  Total_Revolving_Bal Avg_Open_To_Buy Total_Amt_Chng_Q4_Q1 Total_Trans_Amt
##  Min.   :   0        Min.   :    3   Min.   :0.0000       Min.   :  510
##  1st Qu.: 359        1st Qu.: 1324   1st Qu.:0.6310       1st Qu.: 2156
##  Median :1276        Median : 3474   Median :0.7360       Median : 3899
##  Mean   :1163        Mean   : 7469   Mean   :0.7599       Mean   : 4404
##  3rd Qu.:1784        3rd Qu.: 9859   3rd Qu.:0.8590       3rd Qu.: 4741
##  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
```

```r
# Check for NA's
bank %>%
  summarise_all(funs(is.na(.) %>% sum()))
```

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
```

```
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
## # A tibble: 1 x 21
##   CLIENTNUM Attrition_Flag Customer_Age Gender Dependent_count Education_Level
##       <int>          <int>        <int>  <int>           <int>           <int>
## 1         0              0            0      0               0               0
## # ... with 15 more variables: Marital_Status <int>, Income_Category <int>,
## #   Card_Category <int>, Months_on_book <int>, Total_Relationship_Count <int>,
## #   Months_Inactive_12_mon <int>, Contacts_Count_12_mon <int>,
## #   Credit_Limit <int>, Total_Revolving_Bal <int>, Avg_Open_To_Buy <int>,
## #   Total_Amt_Chng_Q4_Q1 <int>, Total_Trans_Amt <int>, Total_Trans_Ct <int>,
## #   Total_Ct_Chng_Q4_Q1 <int>, Avg_Utilization_Ratio <int>
```

Good news is that we have no NA's to deal with.

Perform EDA using Plots and Tables

Dimension Reduction for "Income_Category"

```r
# Categorization of Incomes
low_income <- c("Less than $40K")
mid_income <- c("$40K - $60K", "$60K - $80K")
high_income <- c("$80K - $120K", "$120K +")

# Implement the categories for the economic statuses
bank$Economic_Class <- ifelse(bank$Income_Category %in% low_income,
                              "Low Income",
                              bank$Income_Category)
bank$Economic_Class <- ifelse(bank$Income_Category %in% mid_income,
                              "Mid Income",
                              bank$Economic_Class)
bank$Economic_Class <- ifelse(bank$Income_Category %in% high_income,
                              "High Income",
                              bank$Economic_Class)

# Check
unique(bank$Economic_Class)
```

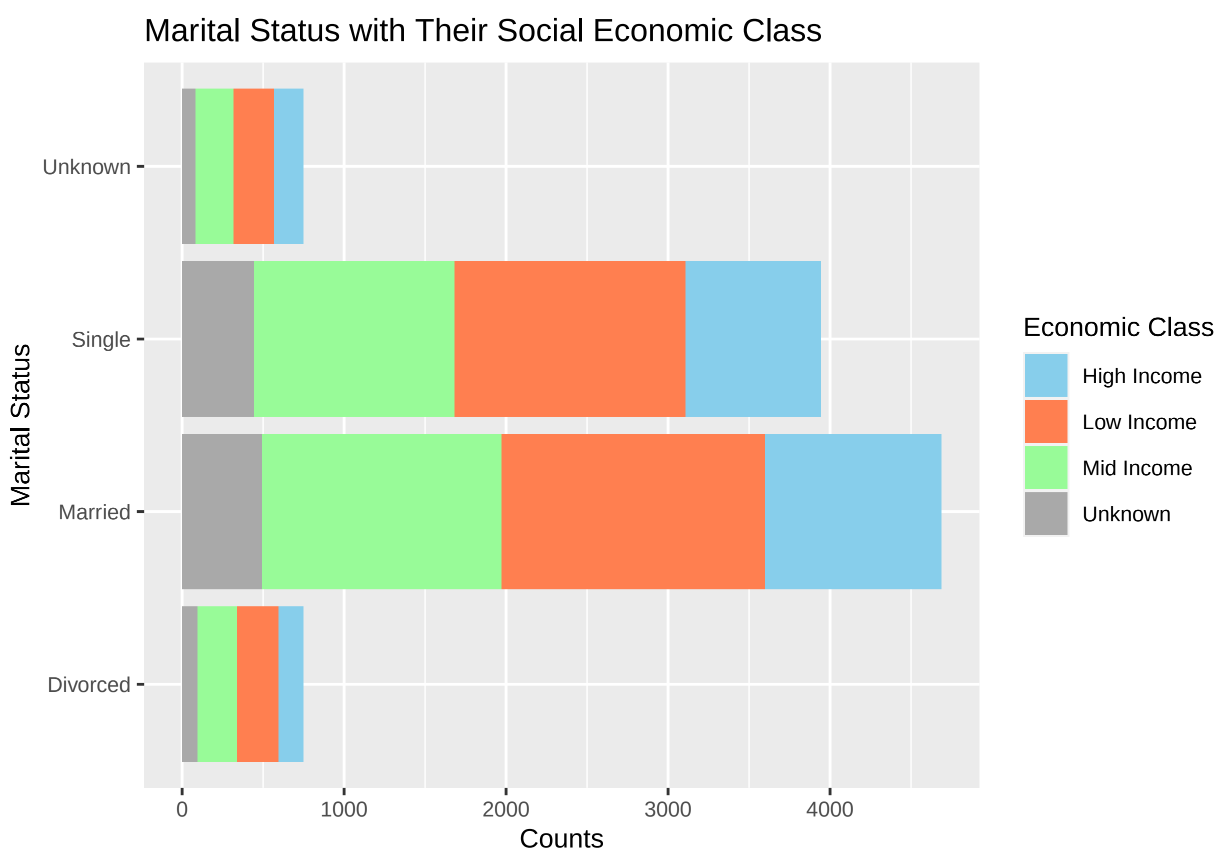
```
## [1] "Mid Income"  "Low Income"  "High Income" "Unknown"
```

I decided to determine that lower income status would be "Less than $40K", middle income would be "$40K - $60K" & "$60K - $80K", and high income would be "$80K - $120K" & "$120K +". I did this to simplify the process of determining people's Social Economic Status. Using this Dimension Reduction, we have to keep in mind that there will be an increase in bias and reduction in variance for our models.

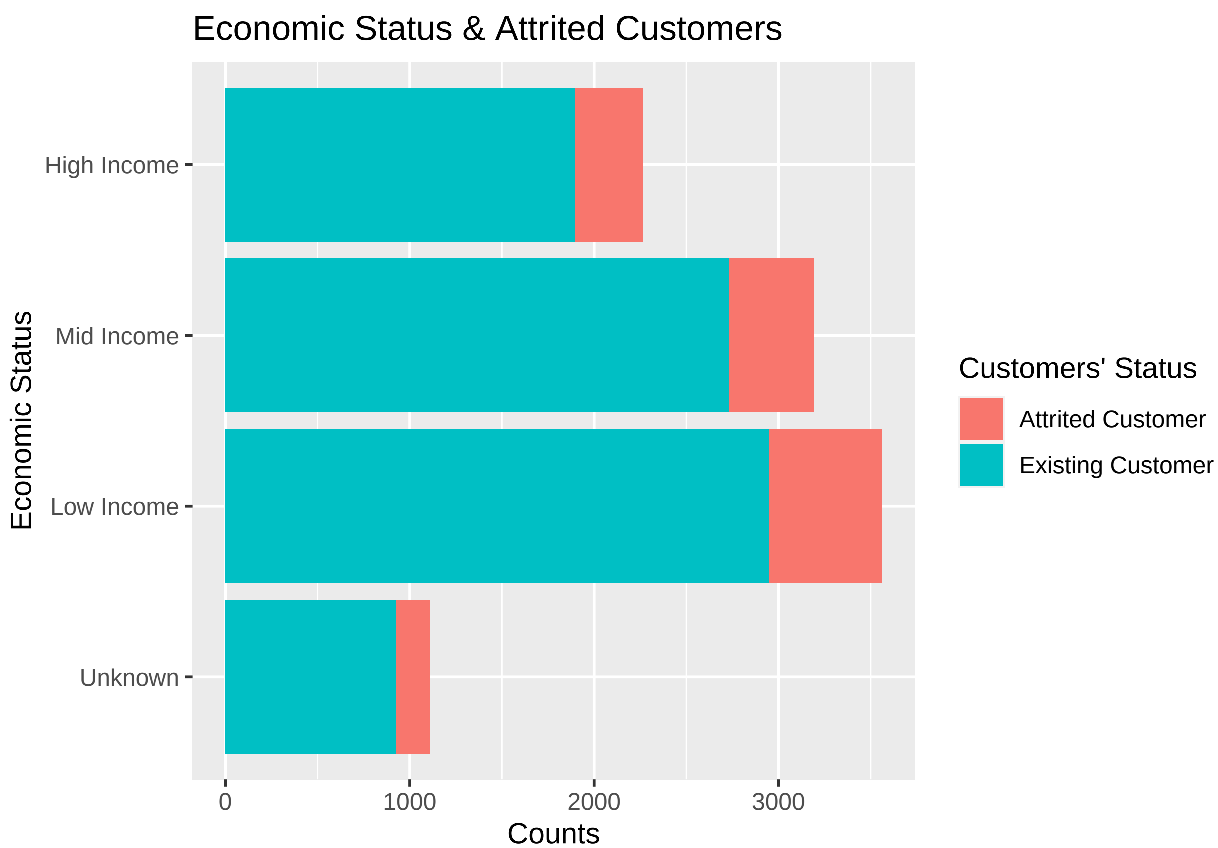
```r
# Checking variables
names(bank)
```

```
##  [1] "CLIENTNUM"                "Attrition_Flag"
##  [3] "Customer_Age"             "Gender"
##  [5] "Dependent_count"          "Education_Level"
##  [7] "Marital_Status"           "Income_Category"
##  [9] "Card_Category"            "Months_on_book"
## [11] "Total_Relationship_Count" "Months_Inactive_12_mon"
## [13] "Contacts_Count_12_mon"    "Credit_Limit"
## [15] "Total_Revolving_Bal"      "Avg_Open_To_Buy"
## [17] "Total_Amt_Chng_Q4_Q1"     "Total_Trans_Amt"
## [19] "Total_Trans_Ct"           "Total_Ct_Chng_Q4_Q1"
## [21] "Avg_Utilization_Ratio"    "Economic_Class"
```

```r
# Count of Marital Statuses
g1 <- ggplot(bank, aes(y=Marital_Status))
g1 +
  geom_histogram(stat="count", aes(fill=Economic_Class)) +
  ggtitle("Marital Status with Their Social Economic Class") +
  labs(x="Counts", y="Marital Status", fill="Economic Class") +
  theme(text = element_text(size=10)) +
  scale_fill_manual(values=c("skyblue", "coral", "palegreen", "darkgrey"))
```



```r
# Income Categories vs. Attrition (cancelled)
g2 <- ggplot(bank, aes(y=Economic_Class, fill=Attrition_Flag))
g2 +
  geom_bar() +
  labs(x="Counts", fill="Customers' Status") +
       ggtitle("Economic Status & Attrited Customers") +
  scale_y_discrete(name="Economic Status",
                   limits=c("Unknown", "Low Income",
                            "Mid Income", "High Income"))
```



For people of single and married statuses, we can see that the Middle Income and Low Income economic statuses are relatively equal to each other. What's interesting is that the High Income is smaller for usage of credit for all categories. There happens to be more married people using lines of credit. With these observations, we can further explore why this might be the case and may help us determine why these people of the following categories become attrited customers.

As for the second plot, we are trying to figure out where there are more "attrited customers". It would appear that lower income customers have more churning than the other income statuses, and lower income individuals have the most frequency of people who hold some line of credit.

Tabling Customers' Statuses w/ Education and Economic Class

```r
table(bank$Attrition_Flag, bank$Education_Level, bank$Economic_Class) %>%
  prop.table(c(1,3)) %>%
  round(4)
```

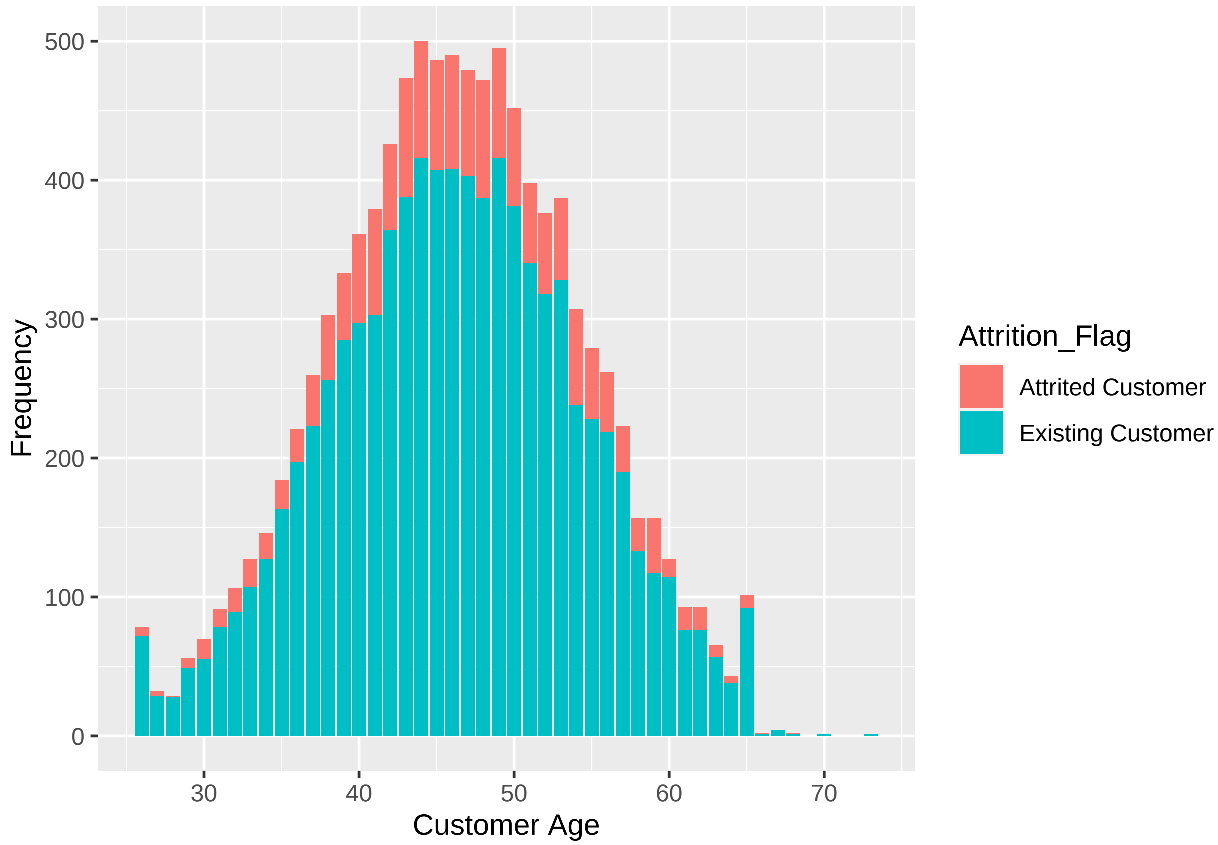
```
## , ,  = High Income
##
##
##                     College Doctorate Graduate High School Post-Graduate
##   Attrited Customer  0.1304    0.0516   0.2690      0.1875        0.0571
##   Existing Customer  0.1040    0.0396   0.3078      0.2038        0.0475
##
##                     Uneducated Unknown
##   Attrited Customer     0.1630  0.1413
##   Existing Customer     0.1457  0.1515
##
## , ,  = Low Income
##
##
##                     College Doctorate Graduate High School Post-Graduate
##   Attrited Customer  0.0801    0.0686   0.3431      0.1650        0.0556
##   Existing Customer  0.1004    0.0393   0.3150      0.1933        0.0461
##
##                     Uneducated Unknown
##   Attrited Customer     0.1307  0.1569
##   Existing Customer     0.1499  0.1560
##
## , ,  = Mid Income
##
##
##                     College Doctorate Graduate High School Post-Graduate
##   Attrited Customer  0.0891    0.0391   0.2804      0.2022        0.0717
##   Existing Customer  0.1003    0.0406   0.3097      0.2083        0.0567
##
##                     Uneducated Unknown
##   Attrited Customer     0.1413  0.1761
##   Existing Customer     0.1387  0.1457
##
## , ,  = Unknown
##
##
##                     College Doctorate Graduate High School Post-Graduate
##   Attrited Customer  0.0856    0.0856   0.2620      0.2299        0.0214
##   Existing Customer  0.0995    0.0584   0.3059      0.1968        0.0465
##
##                     Uneducated Unknown
##   Attrited Customer     0.1711  0.1444
##   Existing Customer     0.1654  0.1276
```

For all the Economic Statuses, there is a higher concentration of graduates and second to that is high school education. Our team will further explore as to why this might be the case and how they might contribute to churning.

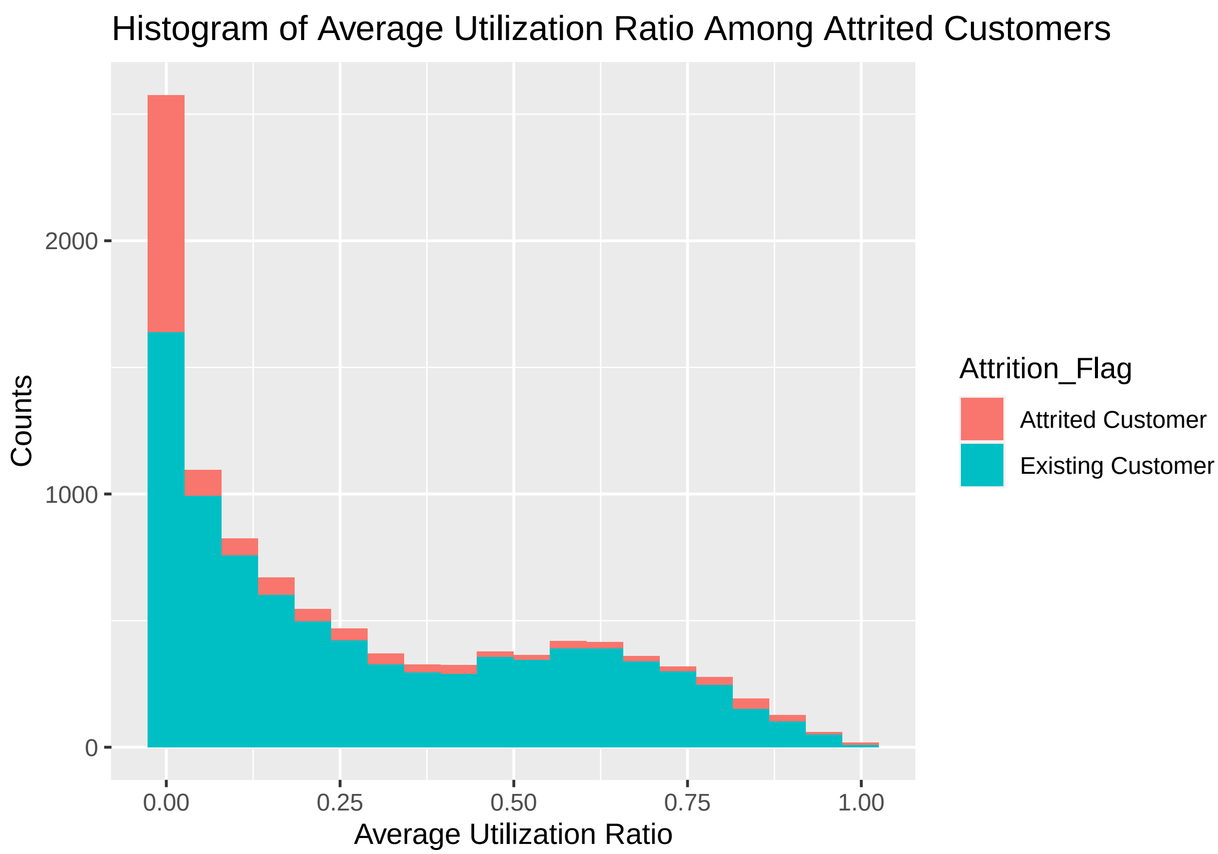
Further Exploration and More Plots

```
# Customer Age
ggplot(bank, aes(x=Customer_Age, fill=Attrition_Flag)) +
```

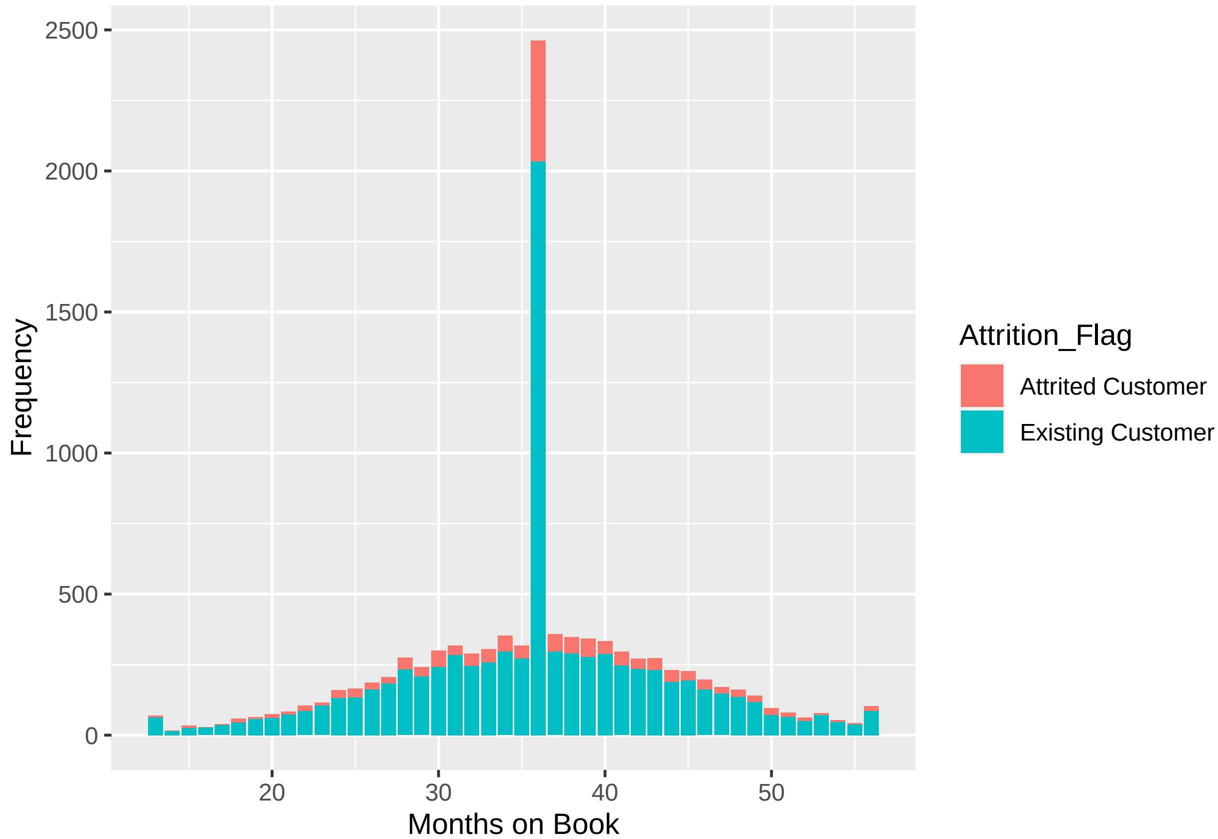
```
  geom_bar() +
  xlab("Customer Age") +
  ylab("Frequency")
```



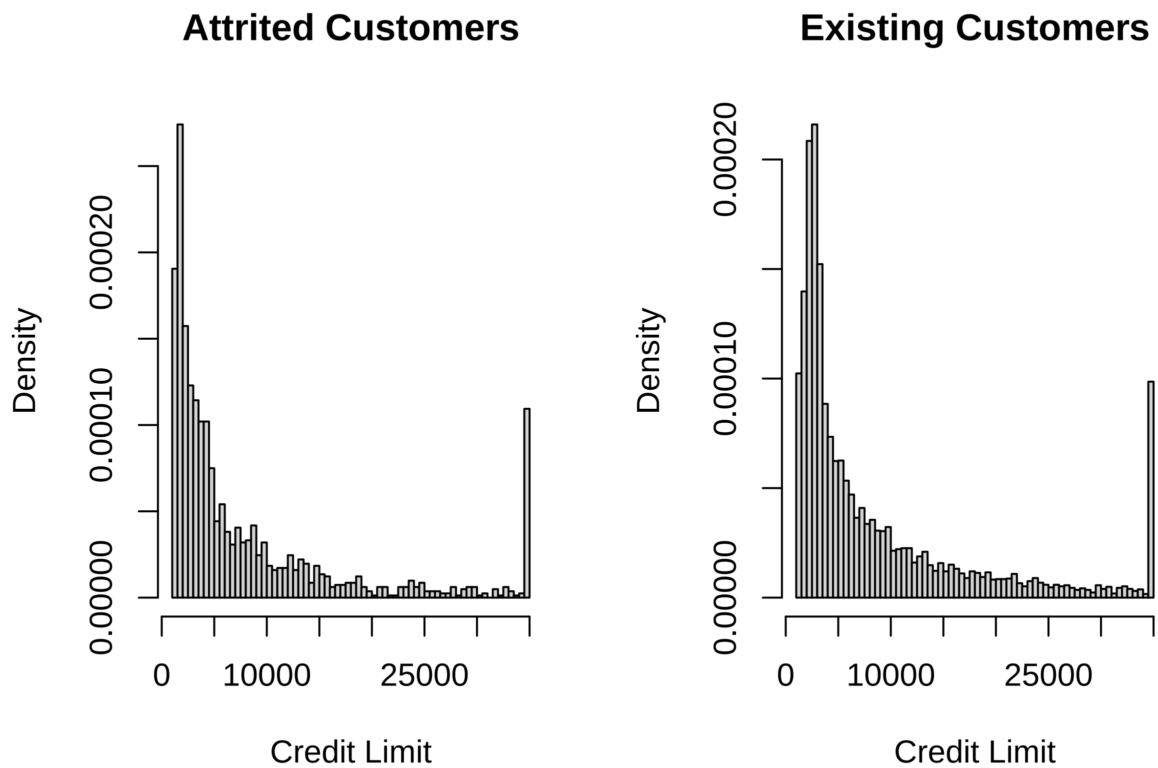

```
# Average Utilization Ratio
ggplot(bank, aes(x=Avg_Utilization_Ratio, fill=Attrition_Flag)) +
  geom_histogram(bins=20) +
  ggtitle("Histogram of Average Utilization Ratio Among Attrited Customers") +
  ylab("Counts") +
  xlab("Average Utilization Ratio")
```




```r
# Months on Book
ggplot(bank, aes(x=Months_on_book, fill=Attrition_Flag)) +
  geom_bar() +
  xlab("Months on Book") +
  ylab("Frequency")
```




```
# Credit Limit
par(mfrow=c(1, 2))
hist(bank$Credit_Limit[bank$Attrition_Flag == "Attrited Customer"], main = "Attrited Customers", xlab =
hist(bank$Credit_Limit[bank$Attrition_Flag == "Existing Customer"], main = "Existing Customers", xlab =
```



Age does not seem to have a large impact on customer retention.

Attrited customers tended to use less credit than existing customers did. Is there a causal relationship between credit usage and customer churning?

One would suspect that customers who left would have a shorter relationship with the bank. Surprisingly, there does not seem to exist much difference between the length of the period of relationship with the bank with respect to churning.

The credit limits of existing customers tend to be greater than those of attrited customers.

Categorical Variables

```
table(bank$Attrition_Flag, bank$Education_Level)/c(1627, 8500)
```

```
##
##                        College  Doctorate   Graduate High School Post-Graduate
##   Attrited Customer 0.09465274 0.05838967 0.29932391  0.18807621    0.05654579
##   Existing Customer 0.10105882 0.04188235 0.31070588  0.20082353    0.04988235
##
##                      Uneducated    Unknown
##   Attrited Customer 0.14566687 0.15734481
##   Existing Customer 0.14705882 0.14858824
```

```r
table(bank$Attrition_Flag, bank$Income_Category)/c(1627, 8500)
```

```
##
##                        $120K + $40K - $60K $60K - $80K $80K - $120K
##   Attrited Customer 0.07744315  0.16656423  0.11616472   0.14874001
##   Existing Customer 0.07070588  0.17870588  0.14270588   0.15211765
##
##                     Less than $40K    Unknown
##   Attrited Customer     0.37615243 0.11493546
##   Existing Customer     0.34694118 0.10882353
```

```r
table(bank$Attrition_Flag, bank$Marital_Status)/c(1627, 8500)
```

```
##
##                       Divorced    Married     Single    Unknown
##   Attrited Customer 0.07437001 0.43577136 0.41057160 0.07928703
##   Existing Customer 0.07376471 0.46800000 0.38529412 0.07294118
```

There doesn't seem to be a large difference between the educations, marital status, and incomes of churned and existing customers.

Potential Questions to Ask When Modeling

Which variables should the bank optimize to retain customers? Could we estimate the probability that a certain customer will leave or stay? What kind of relationship should a bank have with its customers to minimize churning? What types of customers does the bank have the most issues with?

Final Thoughts

Since there are a plethora of variables to look at, we should look at variable selection (such as backwards, forwards, stepwise selection, etc.). We need to utilize the general linear model to see what we can predict from the model with the best given variables. Once we have all our statistical numbers, we can start drawing possible solutions to our questions above.