Analytical Plan

Introduction

This project will work through the process of using predictive modeling to help the ABC Corporation identify customers with high risk attrition. In the first part, we will present a comprehensive analytical plan that addresses ABC Corporation's challenges, outlines the target and potential predictor variables, provides a feature engineering strategy, and details the proposed modeling approach.

1 Business Problem

ABC Corporation aims to identify customers at high risk of attrition - that is, customers likely to leave their services. The risk of attrition will be quantified as a probability between 0 and 1, reflecting the likelihood of a customer leaving. By identifying high-risk customers, ABC Corporation can proactively target these individuals with tailored advertisements or special offers in an effort to retain them, ultimately enhancing customer loyalty, increasing customer lifetime value, and reducing the cost of acquiring new customers.

My analysis will focus on examining the data to determine which variables are most predictive of attrition. With data from over 10,000 customers and insights into whether each customer has been retained or left, we will conduct a detailed analysis to inform the modeling strategy. The target, or response, variable is based on 'Attrition Flag', a factor with two values - 'Existing Customer' and 'Attrited Customer'. The target is the risk of attrition (discussed previously), a value between 0 and 1 that can be interpreted as the probability of customer attrition.

2 Customer Data Exploration

2.1 Dataset Structure

The customer dataset provided by ABC Corporation contains 10,127 rows, each representing an individual customer, and 21 columns corresponding to distinct variables. These columns include the customer ID, the target variable (e.g., 'Attrition Flag'), and 19 predictive variables. A detailed breakdown of these variables, including their types and possible values, is provided in Table 1 in the Appendix.Exploratory Data Analysis

2.1.1 Demographic Variables

I began by analyzing the demographic variables in the dataset to identify potential patterns related to customer attrition. Table 2 in the Appendix shows the proportion of customers who have attrited ("Attrited Customers") versus those who remain ("Existing Customers") for each of the qualitative demographic variables. Graphical representations of these variables are presented in the Appendix (indicated by figure numbers), to visualize any noticeable trends or imbalances in the data.

'Education_Level' emerged as the most relevant demographic variable for predicting attrition. There are clear differences in the proportions of attrited versus existing customers across different educational levels. For instance, the lowest proportion of attrited customers have a postgraduate degree, while the lowest proportion of existing customers hold a doctorate degree. This suggests that customers with higher education levels may be less likely to attrite, which could be valuable when targeting high-risk customers.

A comparison of ages between attrited and existing customers in the boxplots (Figure 1 in the Appendix) revealed little overall difference. However, when segmented by Card_Category (Figure 26 in the Appendix), the age distributions varied. Customers with a 'Silver' card showed more variation in age, with existing customers generally being younger than attrited customers. For 'Gold' cardholders, attrited customers were younger than existing ones, with less age

variation. Notably, for 'Platinum' cardholders, at least 75% of attrited customers were in their late forties or early fifties, while 50% existing customers were under 45 years old.

While other demographic variables, such as gender, marital status and income category, did not show differing trends between attrited and existing customers, it is possible that combining these variables with others (e.g., 'Credit_Limit') might reveal subtle patterns. Further analysis may explore these interactions further.

2.1.2 Business-Related Variables

Next, I analyzed business-related variables to explore how customer behavior might be linked to attrition. The distributions of the variables 'Credit_Limit' (Figure 7 in the Appendix), 'Avg_Open_To_Buy' (Figure 8 in the Appendix), 'Avg_Utilization_Ratio' (Figure 9 in the Appendix), and 'Months_on_book' (Figure 10 in the Appendix) for attrited and existing customers, highlight the following notable trends:

- 'Credit_Limit': Both histograms are right-skewed, with many customers at the maximum credit limit for both attrited and existing customers.
- 'Avg_Open_To_Buy': The histograms are similarly right-skewed, with most customers having an average open-to-buy credit line between $1,700 and $3,500 over the past 12 months. Both histograms also show a slight increase in customers with open-to-buy credit lines exceeding $30,000.
- 'Avg_Utilization_Ratio': The histograms for both groups are right-skewed, with a large number of customers reporting an average utilization ratio close to zero.
- 'Months_on_book': Parallel boxplots comparing the number of months 'Attrited Customers' and 'Existing Customers' have been with ABC Corporation show little difference. Both distributions are roughly symmetrical, with a median of around 36 months and a few outliers at both extremes.

Table 3 in the Appendix further breaks down the proportions of attrited versus existing customers across each of the qualitative business-related variables. Bar graphs in the Appendix (indicated by figure numbers provided in the table) show the distribution of the data and the differences between the 'Attrition_Flag' values. The key business-related variables explored in this part of the analysis are:

- 'Card_Category': No clear trend was observed for 'Card_Category' in terms of attrition, as there was little difference between attrited and existing customers. However, when used to segment the data for ages, there were noteworthy differences between attrited and existing customers (as discussed in Section 2.2.1).
- 'Total_Relationship_Count': This variable showed a more uniform distribution between attrited and existing customers, but existing customers had fewer individuals with a relationship count of less than three. This suggests that customers who attrite tend to hold fewer products than those who remain.
- 'Months_Inactive_12_mon': The distribution indicates that the majority of attrited customers were inactive for 3 months or more, whereas the majority of existing customers were inactive for 3 months or less. This finding suggests that customers who have been inactive longer are more likely to leave ABC Corporation.

Summary statistics are examined across several quantitative variables in Table 4 in the Appendix. Graphical representations of these variables are provided in the Appendix (indicated

by figure numbers), showing noticeable trends or imbalances in the data. The key findings from these comparisons are shared here:

- 'Total_Trans_Amt' and 'Total_Trans_Ct': Both of these variables show right-skewed distributions with existing customers exhibiting a higher total transaction amount and count compared to attrited customers. This suggests that smaller and fewer transactions may be associated with an increased risk of attrition. Summary statistics also show that the values are smaller for attrited customers.
- 'Total_Revolving_Balance': The distribution of revolving credit card balances is informative. While both histograms show a large number of customers with balances near zero, more existing customers report balances above $1,000. The median revolving balance for attrited customers is $0, whereas for existing customers it is $1,364. This suggests that customers with low or no revolving balance may be more likely to attrite.
- 'Total_Amt_Chng_Q4_Q1' and 'Total_Ct_Chng_Q4_Q1': Both variables show right-skewed distributions for both attrited and existing customers. However, the interquartile range for existing customers is narrower, with many high outliers. Summary statistics confirm this, with higher values for existing customers. This suggests that increases in transaction amounts and counts between Q1 and Q4 are more prominent among customers who remain with ABC Corporation.
- "Avg_Utilization_Ratio': Summary statistics reveal that at least 50% of attrited customers have an average card utilization ratio of 0, compared to existing customers, for whom the majority have a ratio greater than 0.2. The histogram for existing customers shows more data concentrated on the right side, further suggesting that customers who do not leave tend to have higher average utilization ratios.

2.1.3 Variable Relationships

I explored the relationships between key variables using a scatterplot matrix (Figure 20 in the Appendix). The following associations were found to noteworthy:

- 'Credit_Limit' versus 'Avg_Open_To_Buy': This relationship shows a strong positive linear relationship (r=0.9660), suggesting that higher credit limits correspond with higher average open-to-buy amounts. The scatterplots for both 'Attrited Customer' and 'Existing Customer' (Figure 21 in the Appendix) indicate that this relationship appears to be consistent across both groups. However, the absence of a significant difference between the two groups suggests that this relationship may not have a strong impact on customer attrition.
- 'Credit_Limit' versus 'Avg_Utilization_Ratio' and 'Avg_Open_To_Buy' versus 'Avg_Utilization_Ratio': Both relationships display negative, curved associations (Figure 22 and Figure 24 in the Appendix). This suggests an exponential decay pattern for both 'Attrited Customer' and 'Existing Customer'. Notably, the relationships differ between the two groups: customers with lower utilization ratios, relative to their credit limit and average open-to-buy amounts, tend to be more likely to attrite. This indicates that customers who underutilize their credit are at a higher risk of attrition.
- 'Total_Trans_Amt' versus 'Total_Trans_Ct': The scatterplots for these variables (Figure 25 in the Appendix) reveal a positive relationship, with more spread-out data points for 'Existing Customers'. There appear to be three distinct groups in the 'Existing Customer' plot, indicating variability in transaction behavior. In contrast, the 'Attrited

Customer' plot shows a more concentrated grouping with fewer distinct clusters, and both the total transaction amounts and the number of transactions are generally lower - totaling under $10,000 and 100 transactions, respectively. This difference suggests that the frequency and magnitude of transactions may play a role in customer retention, and further exploration of this relationship could provide valuable insights.

2.1.4 Initial Variable Selection

Based on the exploratory data analysis findings, I propose the following predictors for inclusion in the model:

- 'Customer_Age'
- 'Education_Level'
- 'Total_Relationship_Count'
- 'Card_Category'
- 'Months_Inactive_12_mon'
- 'Credit_Limit', 'Total_Revolving_Bal'
- 'Avg_Open_To_Buy', 'Total_Amt_Chng_Q4_Q1'
- 'Total_Trans_Amt', 'Total_Trans_Ct'
- 'Total_Ct_Chng_Q4_Q1'
- 'Avg_Utilization_Ratio'.

As an identifier value, 'CLIENTNUM' will be excluded as it does not provide predictive value. The additional variables listed below will also be excluded as they did not show strong associations with attrition:

- 'Gender'
- 'Dependent_count'
- 'Marital_Status'
- 'Income_Category'
- 'Months_on_book'
- 'Contacts_Count_12_mon'

Because there is a significant class imbalance between the attrited customers (minority class) and existing customers (majority class), strategies will need to be implemented to mitigate this imbalance. Undersampling the majority class will be used to initially reduce this imbalance. This will involve randomly removing samples from the 'Existing Customer' class to create a more balanced dataset. While this can lead to loss of information, it is a simple and effective strategy, particularly given the size of the dataset.

3 Feature Engineering Plan

3.1 Indicators

Indicators will be created for qualitative predictors in order to use the variables as quantitative predictors if necessary. By creating these indicators, there will be more flexibility around how these predictors can be used in the model, allowing for their use in a model that typically requires numeric inputs.

The indicators will be created for 'Education_Level' and 'Card_Category' and defined as followings:

- 'Education_Level':

  0 if 'Unknown'
  1 if 'Uneducated'
  2 if 'High School'
  3 if 'College'
  4 if 'Graduate'
  5 if 'Post-Graduate'
  6 if 'Doctorate'

- 'Card_Category'

  0 if 'Blue'
  1 if 'Silver'
  2 if 'Gold'
  3 if 'Platinum'

By encoding these variables as dummies, the model will be able to account for each category's distinct impact on the outcome variable. Furthermore, because both of these variables are ordinal in nature, their dummy values have been given with intent. The higher the value of the indicator for 'Education_Level', the more years of education the customer likely has. The more premium the 'Card_Category', the higher the value assigned to the dummy variable. This could potentially enhance the model's ability to understand the relationship between education level or card category and attrition.

3.2 Data Transformations

Based on the curved relationships observed between 'Credit_Limit' versus 'Avg_Utilization_Ratio' and 'Avg_Open_To_Buy' versus 'Avg_Utilization_Ratio', I propose applying a logarithmic transformation to these variables. This transformation is expected to linearize the relationships and reduce skewness, improving model performance. Logarithmic transformations are particularly useful when data follows an exponential or multiplicative pattern, as is the case here. A small constant will be added to 'Avg_Utilization_Ratio' prior to the log transformation to avoid errors when taking the logarithm of zero.

4 Next Steps

4.1 Modeling Strategy

ABC Corporation aims to predict customer attrition using classification techniques. The target variable is 'Attrition Flag', a binary variable indicating whether the customer is an 'Attrited Customer' or an 'Existing Customer'. I have considered multiple algorithms and evaluation methods in order to build a robust predictive model for customer attrition.

4.1.1 Logistic Regression

Logistic Regression is a strong candidate for binary classification problems, such as predicting customer attrition. It will model the probability of customer attrition based on predictor variables. One of its key advantages is interpretability; it provides coefficients that show how each feature affects the odds of customer attrition.

Logistic Regression assumes a linear relationship between the predictors and the log-odds of the outcome, a requirement that we will address through the feature engineering process. Regularization techniques will be applied to reduce the risk of overfitting, especially given the large number of predictors in the dataset.

4.1.2 k-Nearest Neighbors (kNN)

k-Nearest Neighbors (kNN) is a flexible, non-parametric algorithm that does not assume any specific distribution of the data. This makes it suitable for capturing complex or nonlinear relationships between customer attributes. Given the combination of numerical and categorical features in the ABC Corporation's customer dataset, kNN is a viable option to model customer attrition.

One challenge of kNN is its sensitivity to class imbalance. To mitigate this, weighted kNN or resampling techniques (e.g., undersampling the 'Existing Customer' class) will be implemented. Additionally, feature scaling will be applied to ensure that distance metrics used in kNN are meaningful.

While kNN is less interpretable than Logistic Regression, we can explore the nearest neighbors for specific predictions to provide some level of explanation for the model's decisions.

4.1.3 Excluded Algorithms

Other potential algorithms have been considered but are ultimately excluded due to the assumptions that are not met by this data. Because Linear Discriminant Analysis (LDA) assumes normally distributed predictors, it may not be an ideal method for this dataset given the number of predictors with skewed distributions. Naive Bayes (NB) assumes conditional independence between predictor variables, which is unlikely to hold in this dataset due to correlations between many features (e.g., 'Credit_Limit' and 'Avg_Utilization_Ratio'). Therefore, it has been determined that LDA and NB are not suitable choices.

4.2 Evaluation Metrics

Model evaluation will be performed based on the following metrics:

- Accuracy will be measured as the proportion of correctly classified instances out of the total predictions made. While accuracy is a useful metric, it is important to consider its limitations in imbalanced datasets, where the majority class (e.g., 'Existing Customer') could dominate the results.
- Area Under the ROC Curve (AUC) will be used to assess the ability of each model to discriminate between the positive (attrited) and negative (existing) classes across various decision thresholds. AUC is particularly useful for imbalanced datasets, as it evaluates model performance independently of the classification threshold.

- A confusion matrix will provide insight into the number of true positives, false positives, true negatives, and false negatives, to help understand where each model is making errors. This analysis will also inform business decisions about the relative importance of types of errors. For example, missing an attrited customer might be more costly than falsely predicting attrition.
- To avoid overfitting and ensure the model performs well on test data, k-fold cross-validation will be used. This will allow the assessment of stability and generalizability of the model's performance.

4.3 Anticipated Challenges and Solutions

Many of the challenges that may be encountered have been addressed previously in the analytical plan. These challenges include:

- Class imbalance: With existing customers outnumbering attrited customers in the training data, class imbalance will need to be addressed using techniques such as resampling (e.g., undersampling the majority class) or weighted algorithms. This ensures that the model effectively predicts the minority class.
- Overfitting: With a large number of predictors, there is a risk of overfitting with the Logistic Regression and kNN models. To mitigate overfitting, regularization and cross-validation will be used.
- Feature Engineering and Data Transformation: Proper data preprocessing, including handling missing values, encoding categorical variables, and feature transformations, are needed for optimal model performance. Feature transformations (e.g., log transformations) will be tested to ensure the models meet the assumptions necessary for accurate predictions.
- Model Interpretability: While Logistic Regression offers high interpretability, kNN may be harder to explain. In this case, nearest neighbor analysis will be used to understand specific predictions.

5 Conclusion

ABC Corporation's objective to identify customers at high risk of attrition is both a strategic and data-driven initiative. Through the comprehensive exploratory data analysis and initial variable selection process, I have identified several key demographic and business-related features that evidently contribute to predicting attrition risk. The proposed approach, using a combination of predictive models such as Logistic Regression and k-Nearest Neighbors (kNN), provides a solid foundation for understanding customer characteristics and behaviors to accurately predict the likelihood of attrition.

By leveraging insights and methodologies discussed and proposed in the analytical plan, ABC Corporation can transform its customer retention strategy into a data-driven process, yielding benefits in both customer satisfaction and business profitability.

6 Appendix

Table 1

Variable Name	Variable Type	Description	Possible Values
CLIENTNUM	quantitative, integer	unique customer identifier	positive integer values
Attrition_Flag	qualitative, factor	indicates whether or not the customer has attrited	"Existing customer" "Attrited Customer"
Customer_Age	quantitative, integer, demographic	customer's age in years	integer values betwen 26 and 73
Gender	qualitative, factor, demographic	customer's gender	M' = Male 'F' = Female
Dependent_count	quantitative, integer, demographic variable	number of dependents the customer has	integer values between 0 and 5
Education_level	qualitative, factor, demographic	customer's education level	"Uneducated", "High School", "College", "Graduate", "Post-Graduate", "Doctorate", "Unknown"
Marital_Status	qualitative, factor, demographic	customer's marital status	"Single", "Married", "Divorced", "Unknown"
Income_Category	qualitative, factor, demographic	annual income range of customer	"Less than $40K", "$40K-$60K", "$60K-$80K", "80K-$120K", "$120K +"
Card_Category	qualitative, factor, demographic	type of card the customer has	"Blue", "Silver", "Gold", "Platinum"

Variable Name	**Variable Type**	**Description**	**Possible Values**
Months_on_book	quantitative, integer	period of customer's relationship with ABC in months	integer values between 13 and 56
Total_Relationship_Count	quantitative, integer	number of products held by the customer	integer values between 1 and 6
Months_Inactive_12_mon	quantitative, integer	number of months customer has been inactive in the last 12 months	integer values between 0 and 6
Contacts_Count_12_mon	quantitative, integer	number of times the customer has been in contact with ABC over the last 12 months	integer values between 0 and 6
Credit_Limit	quantitative	credit limit on the credit card	values between $1438 and $34,516
Total_Revolving_Balance	quantitative, integer	revolving balance on the credit card	integer values between $0 and $2,517
Avg_Open_To_Buy	quantitative	average open to buy credit line from the last 12 months	values between $3 to $34,516
Total_Amt_Chng_Q4_Q1	quantitative	change in transaction amount (Q4 over Q1)	values between 0.000 and 3.397
Total_Trans_Amt	quantitative, integer	total transaction amount from the last 12 months	integer values between $510 and $18,484
Total_Trans_Ct	quantitative, integer	total transaction count from the last 12 months	integer values between 10 and 139
Total_Ct_Chng_Q4_Q1	quantitative	change in number of transactions (Q4 over Q1)	values between 0.000 and 3.714
Avg_Utilization_Ratio	quantitative	average card utilization ratio	values between 0.000 and 0.999

Table 2

Variable Name	Attrited Customer Proportions	Existing Customer Proportions	Graphical Representation
Gender	F = 0.5716 M = 0.4284	F = 0.5209 M = 0.4791	Figure 2
Dependent_count	0 = 0.0830 1 = 0.1653 2 = 0.2563 3 = 0.2963 4 = 0.1598 5 = 0.0393	0 = 0.0905 1 = 0.1846 2 = 0.2633 3 = 0.2647 4 = 0.1546 5 =0.0423	Figure 3
Education_Level	College = 0.0947 Doctorate = 0.0584 Graduate = 0.2993 High School = 0.1881 Post-Graduate = 0.0566 Uneducated = 0.1457 Unknown = 0.1573	College = 0.1011 Doctorate = 0.0419 Graduate = 0.3107 High School = 0.2008 Post-Graduate = 0.0499 Uneducated = 0.1471 Unknown = 0.1486	Figure 4
Marital_Status	Divorced = 0.0744 Married = 0.4358 Single = 0.4106 Unknown = 0.0793	Divorced = 0.0738 Married = 0.4680 Single = 0.3853 Unknown = 0.0729	Figure 5
Income_Category	Less than $40K = 0.3762 $40K-$60K = 0.1666 $60K-$80K = 0.1162 $80K - $120K = 0.1487 $120K + = 0.0774 Unknown = 0.1149	Less than $40K = 0.3469 $40K-$60K = 0.1787 $60K-$80K = 0.1427 $80K - $120K = 0.1521 $120K + = 0.0707 Unknown = 0.1088	Figure 6

Table 3

Variable Name	Attrited Customer Proportions	Existing Customer Proportions	Graphical Representation
Card_Category	Blue = 0.9336 Silver = 0.0504 Gold = 0.0129 Platinum = 0.0031	Blue = 0.9314 Silver = 0.0557 Gold = 0.0112 Platinum = 0.0018	Figure 11
Total_Relationship_Count	1 = 0.1432 2 = 0.2127 3 = 0.2459 4 = 0.1383 5 = 0.1395 6 = 0.1205	1 = 0.0797 2 = 0.1055 3 = 0.2241 4 = 0.1985 5 = 0.1959 6 = 0.1965	Figure 12
Months_Inactive_12_mon	0 = 0.0092 1 = 0.0615 2 = 0.3104 3 = 0.5077 4 = 0.0799 5 = 0.0197 6 = 0.0117	0 = 0.0017 1 = 0.2509 2 = 0.3267 3 = 0.3553 4 = 0.0359 5 = 0.0172 6 = 0.0124	Figure 13

Table 4

Variable Name	Attrited Customer Proportions	Existing Customer Proportions	Graphical Representation
Total_Trans_Amt	Min = 510 Q1 = 2156 Median = 2329 Mean = 3095 Q3 = 2772 Max = 10583	Min = 816 Q1 = 2385 Median = 4100 Mean = 4655 Q3 = 4781 Max = 18484	Figure 14
Total_Trans_Ct	Min = 10 Q1 = 37 Median = 43 Mean = 44.93 Q3 = 51 Max = 94	Min = 11 Q1 = 54 Median = 71 Mean = 68.67 Q3 = 82 Max = 139	Figure 15
Total_Revolving_Bal	Min = 0 Q1 = 0 Median = 0 Mean = 672.8 Q3 = 1303.5 Max = 2517.0	Min = 0 Q1 = 80 Median = 1364 Mean = 1257 Q3 = 1807 Max = 2517	Figure 16
Total_Amt_Chng_Q4_Q1	Min = 0 Q1 = 0.5445 Median = 0.7010 Mean = 0.6943 Q3 = 0.8560 Max = 1.4920	Min = 0.2560 Q1 = 0.6430 Median = 0.7430 Mean = 0.7725 Q3 = 0.8600 Max = 3.3970	Figure 17
Total_Ct_Chng_Q4_Q1	Min = 0.0000 Q1 = 0.4000 Median = 0.5310 Mean = 0.5544 Q3 = 0.6920 Max = 2.5000	Min = 0.0280 Q1 = 0.6170 Median = 0.7210 Mean = 0.7424 Q3 = 0.8330 Max = 3.7140	Figure 18
Avg_Utilization_Ratio	Min = 0.0000 Q1 = 0.0000 Median = 0.0000 Mean = 0.1625 Q3 = 0.2310 Max = 0.9990	Min = 0.0000 Q1 = 0.0550 Median = 0.2110 Mean = 0.2964 Q3 = 0.5292 Max = 0.9940	Figure 19

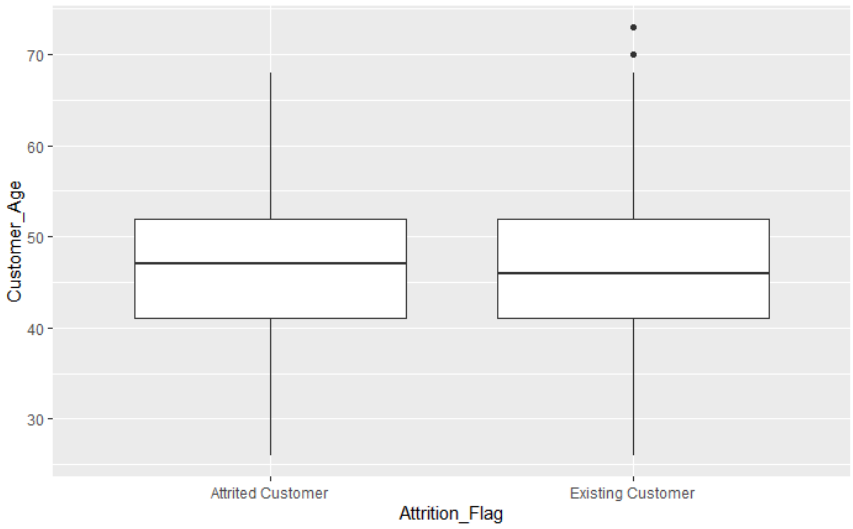

Figure 1 *These parallel boxplots compare the ages of attrited customers and existing customers.*

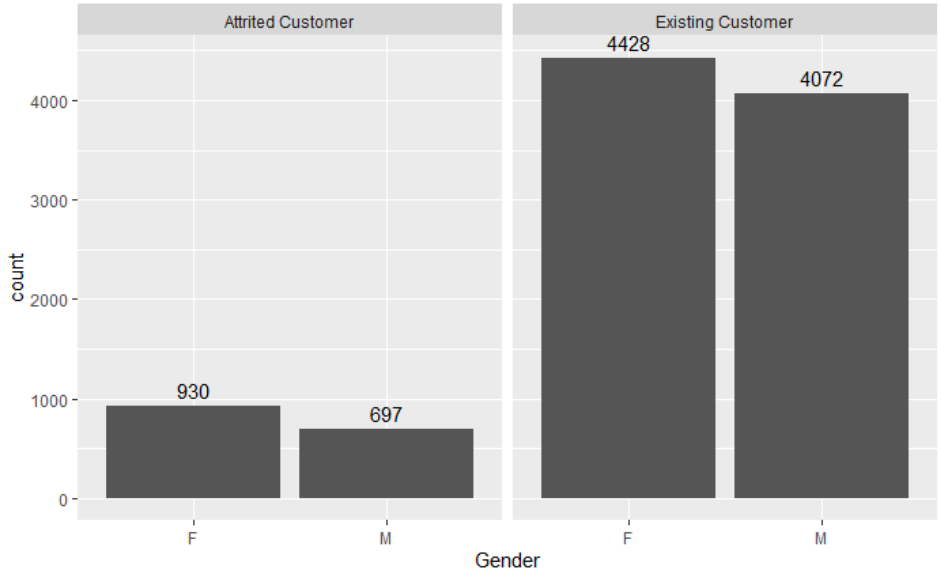

Figure 2 *The bar graphs shown compare the breakdown of customer's gender between attrited customers and existing customers.*

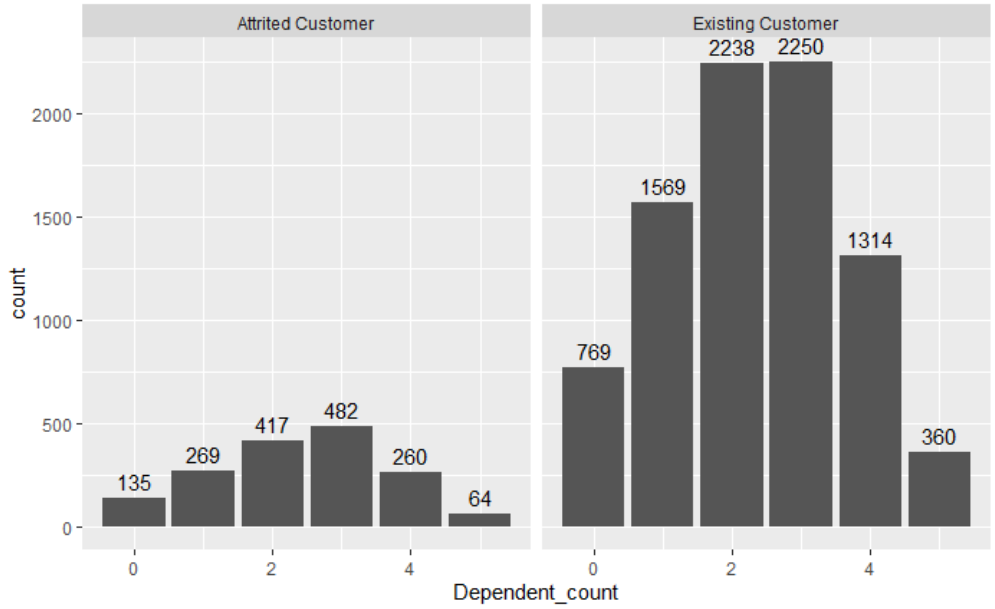

Figure 3 *The bar graphs shown compare the number of dependents a customer has between attrited customers and existing customers.*

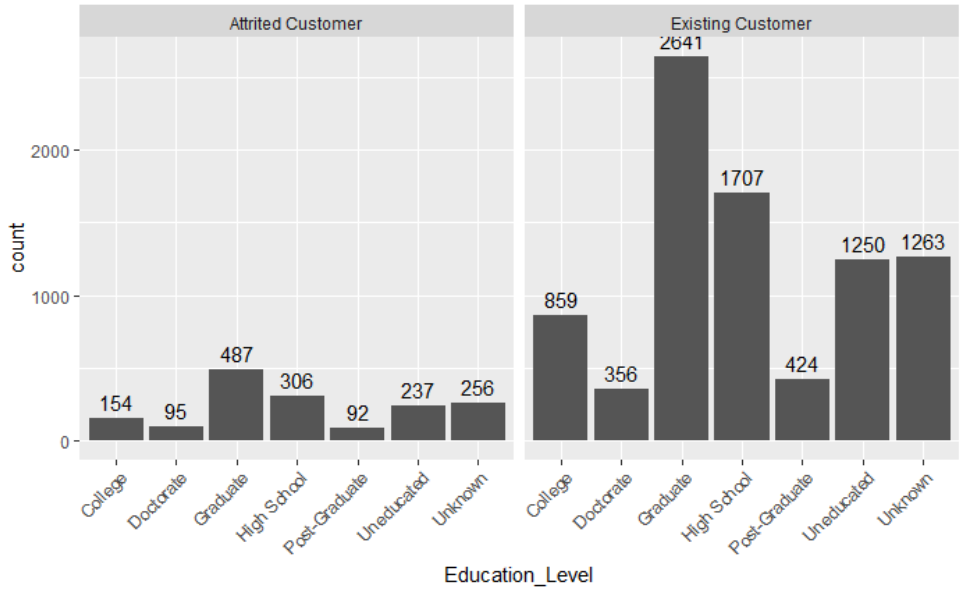

Figure 4 *The bar graphs shown compare the breakdown of customer's education level between attrited customers and existing customers.*

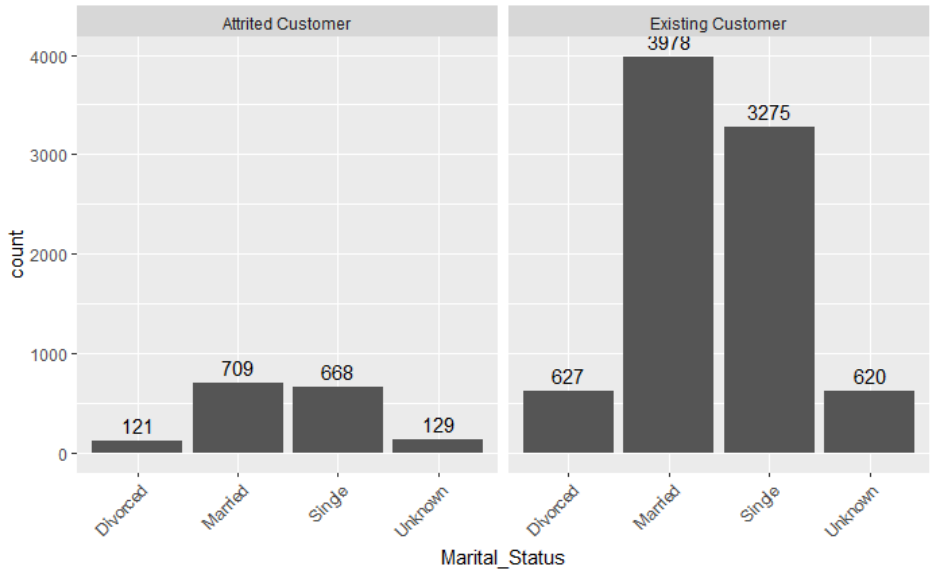

Figure 5 *The bar graphs shown compare the marital status between attrited customers and existing customers.*

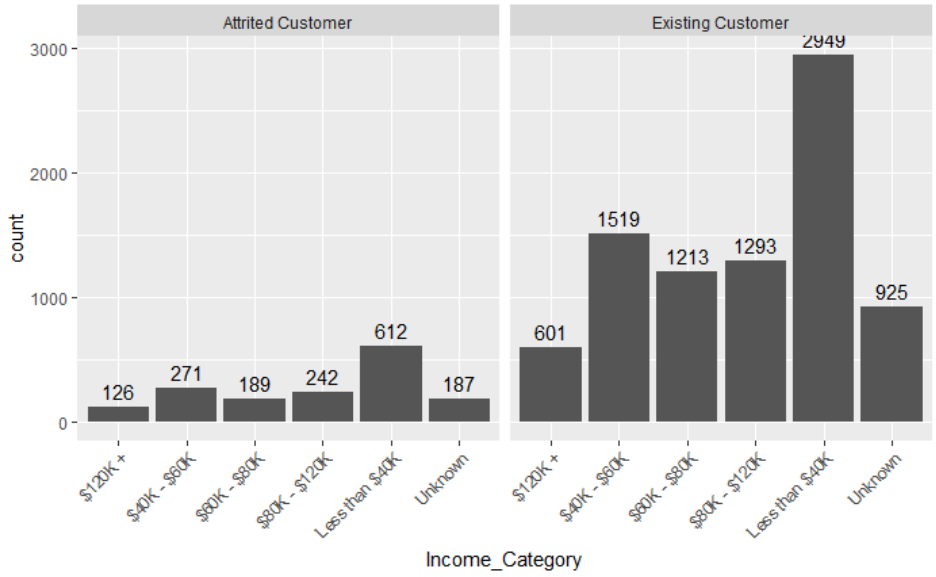

Figure 6 *The bar graphs shown compare the level of income of attrited customers versus existing customers.*

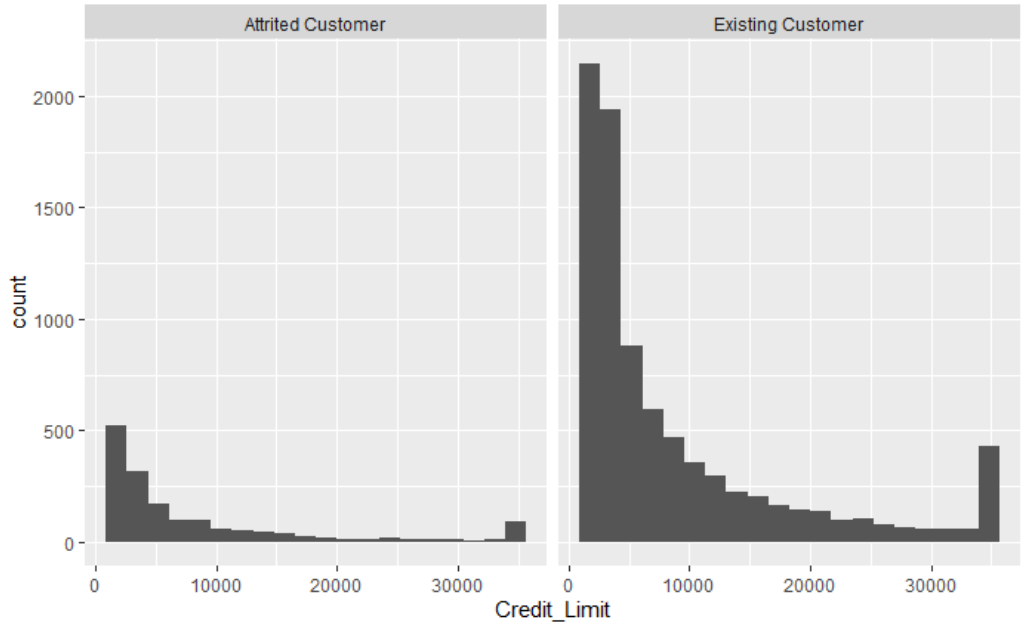

Figure 7 *The histograms show the distribution of attrited and existing customers' credit limit.*

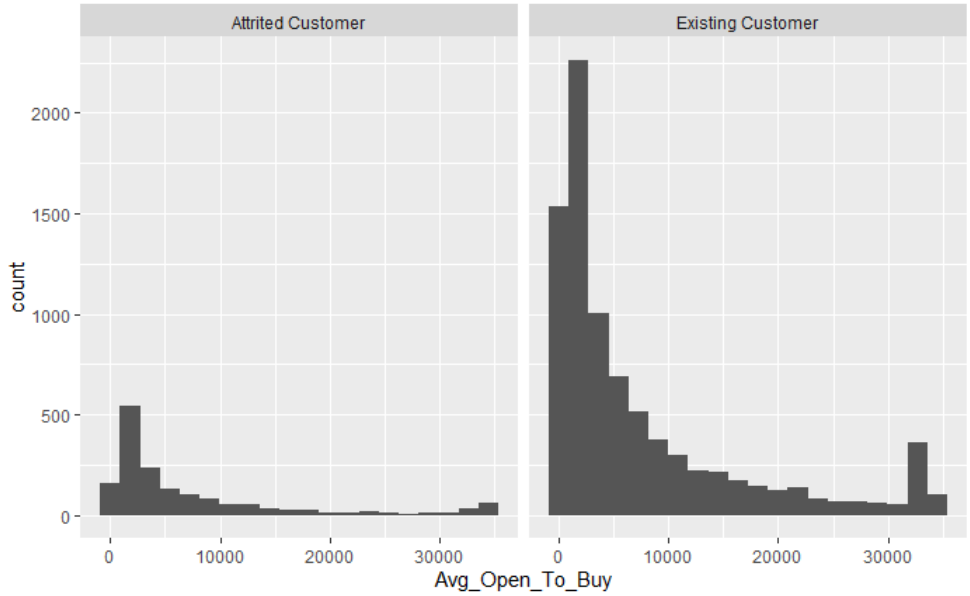

Figure 8 *The histograms compare the average value open to buy credit line between attrited customers and existing customers.*

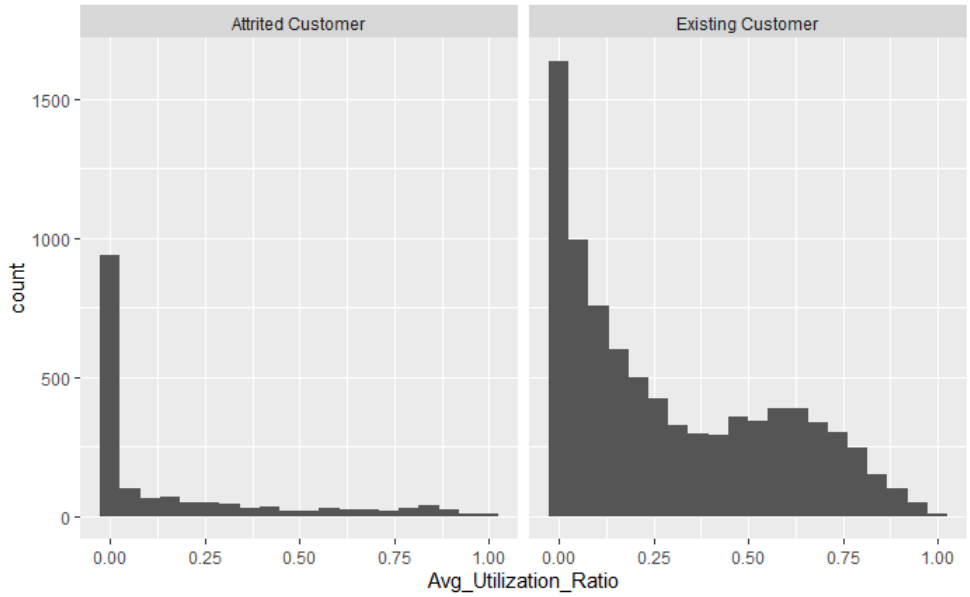

Figure 9 *The histograms compare the average utilization ratio between attrited customers and existing customers.*

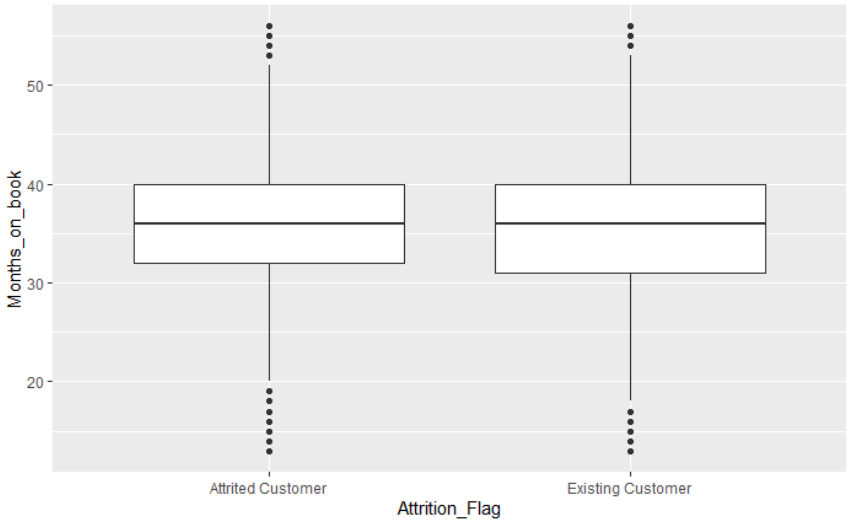

Figure 10 *The parallel boxplots compare the number of months "Attrited Customers" versus "Existing Customers" have been at ABC Corporation.*

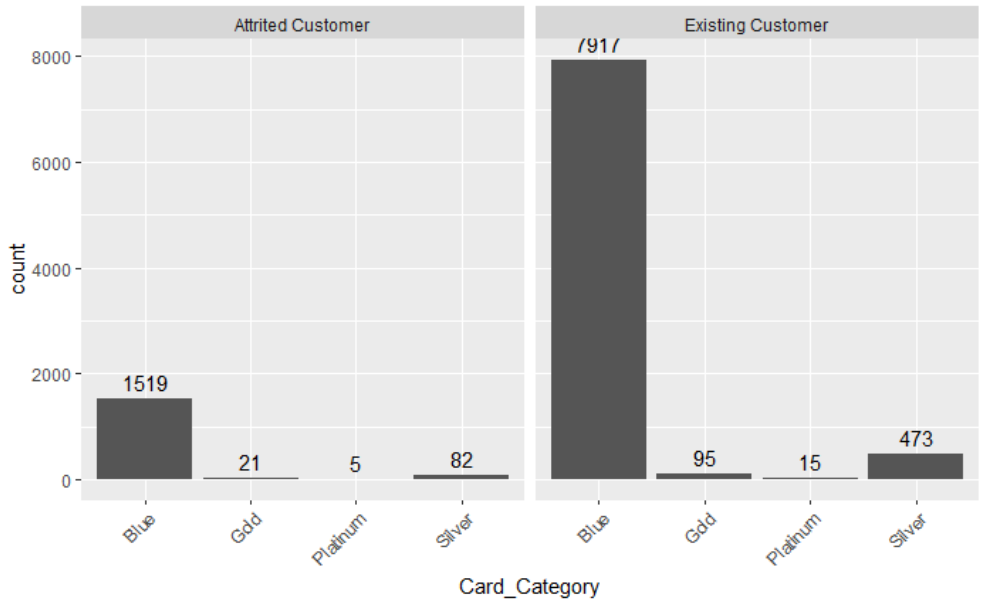

Figure 11 *These bar graphs compare the number of the "Attrited Customers" versus "Existing Customers" who fall in each credit card category at ABC Corporation.*

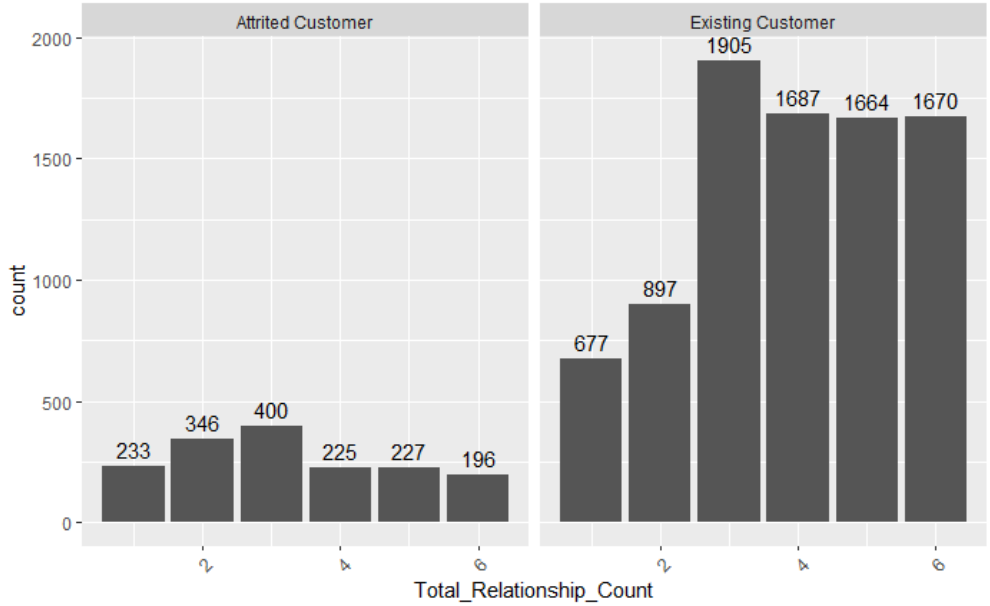

Figure 12 *This plot shows bar graphs comparing the number of the products held by "Attrited Customers" versus "Existing Customers".*

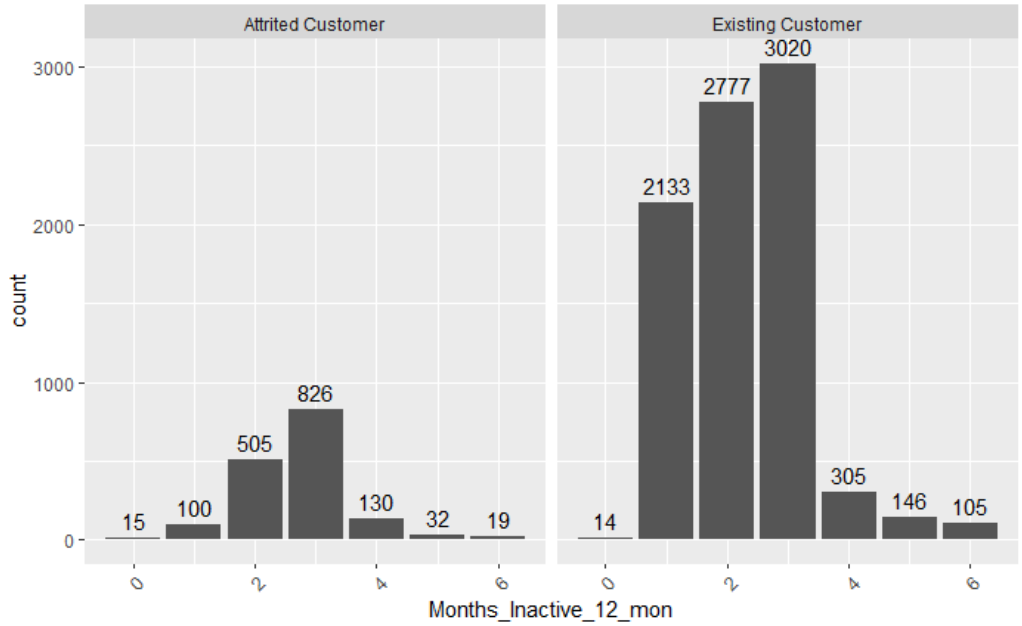

Figure 13 *The bar graphs show the number of the months customers were inactive over the past 12 months for "Attrited Customers" versus "Existing Customers".*

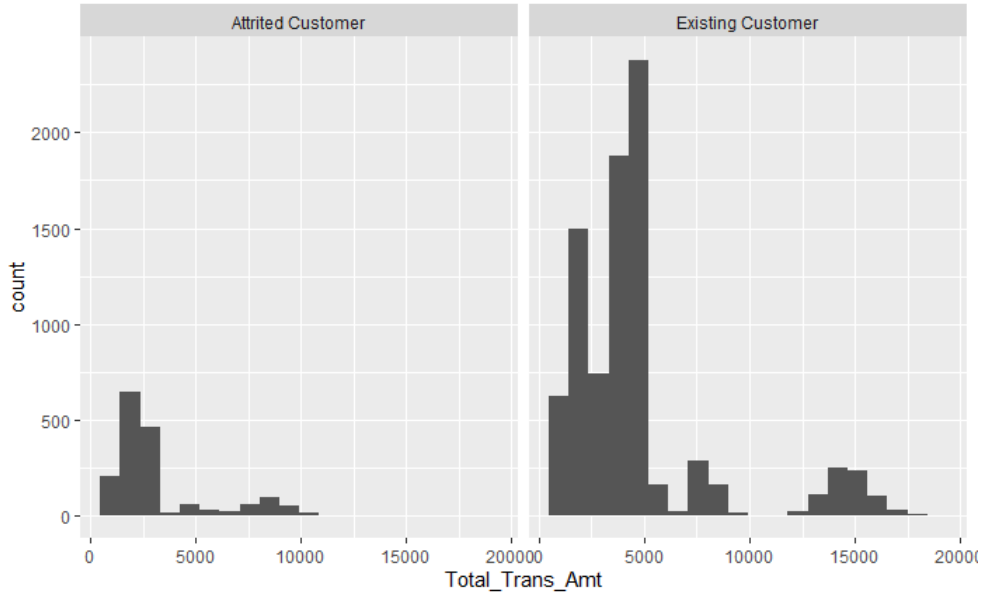

Figure 14 *The histograms compare the total transaction amounts from the last 12 months of "Attrited Customers" and "Existing Customers".*

Figure 15 *The histograms compare the total number of transactions from the last 12 months of "Attrited Customers" and "Existing Customers".*

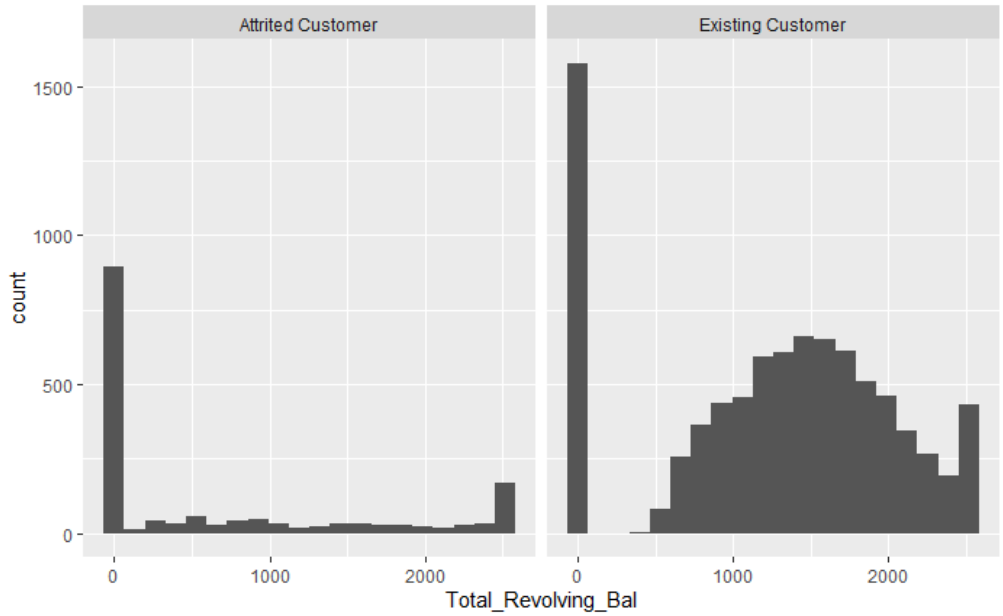

Figure 16 *The histograms show the revolving credit card balances of "Attrited Customers" versus"Existing Customers".*

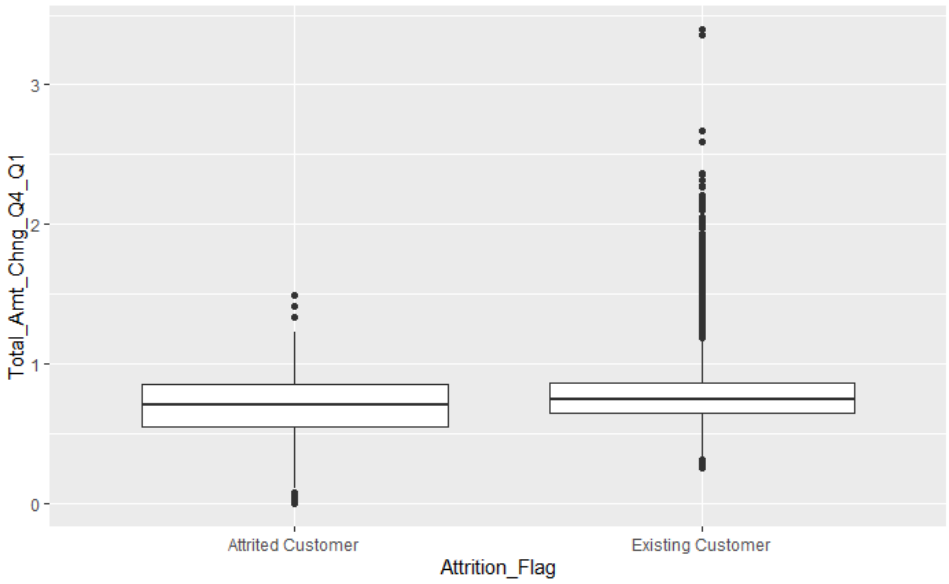

Figure 17 *The parallel boxplots show the total transaction amount change (Q4 over Q1) for "Attrited Customers" and "Existing Customers".*

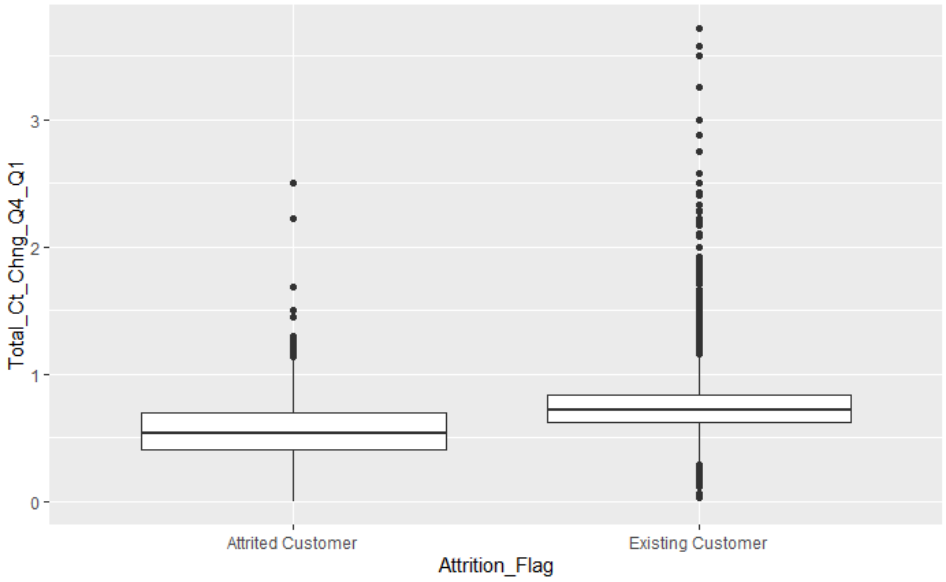

Figure 18 *The parallel boxplots show the change in the number of transactions (Q4 over Q1) for "Attrited Customers" and "Existing Customers".*

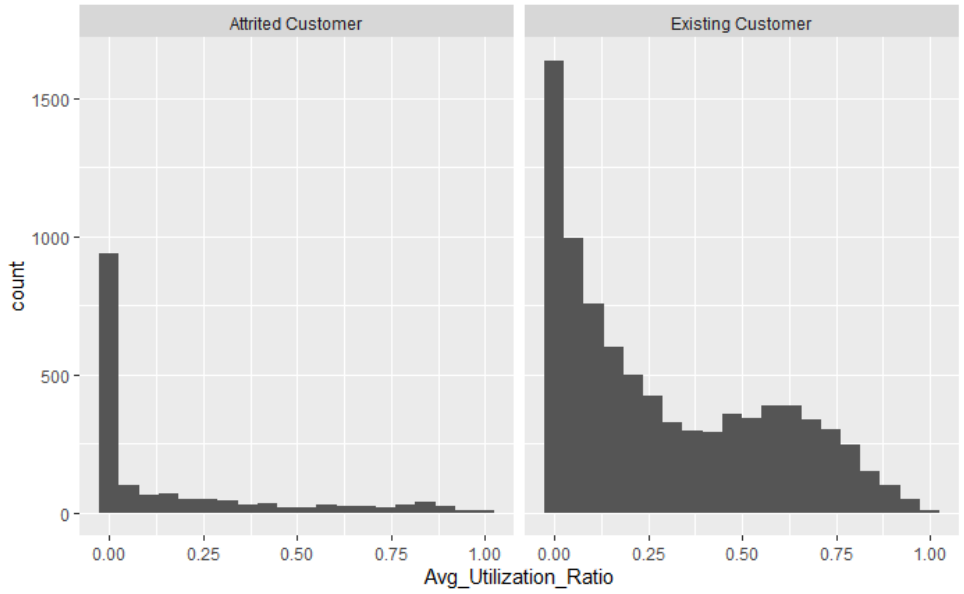

Figure 19 *The histograms compare the average utilization ratio for "Attrited Customers" versus"Existing Customers".*

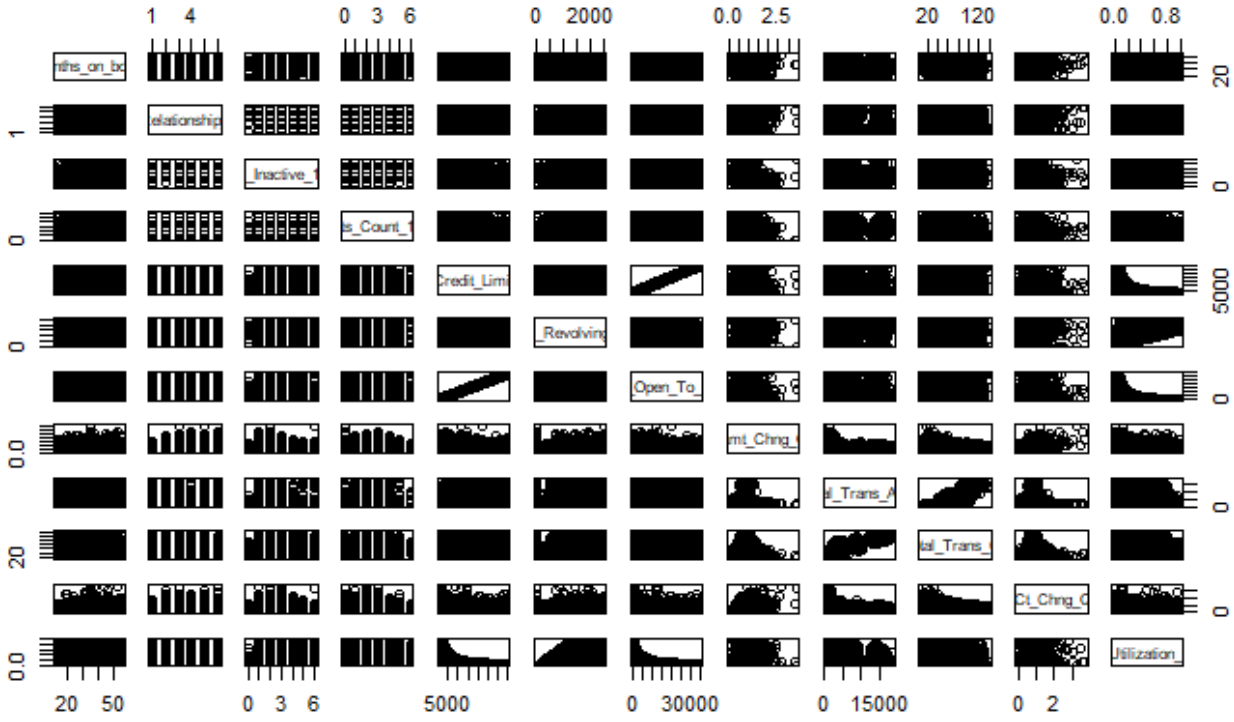

Figure 20 *This scatterplot matrix shows the relationships between each of the quantitative, business-related variables, including: 'Months_on_book', 'Total_Relationship_Count', 'Months_Inactive_12_mon', 'Contacts_Count_12_mon', 'Credit_Limit', 'Total_Revolving_Balance', 'Avg_Open_To_Buy', 'Total_Amt_Chng_Q4_Q1', 'Total_Trans_Amt', 'Total_Trans_Ct', 'Total_Ct_Chng_Q4_Q1', and 'Avg_Utilization_Ratio'.*

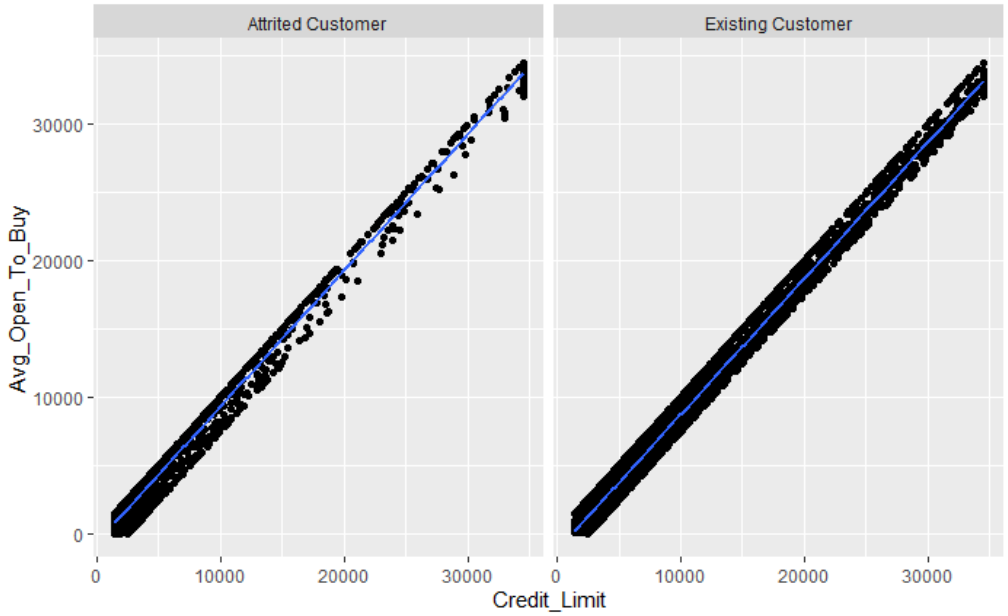

Figure 21 *The scatterplots show the association between "Credit_Limit" and "Avg_Open_To_Buy" with fitted linear models.*

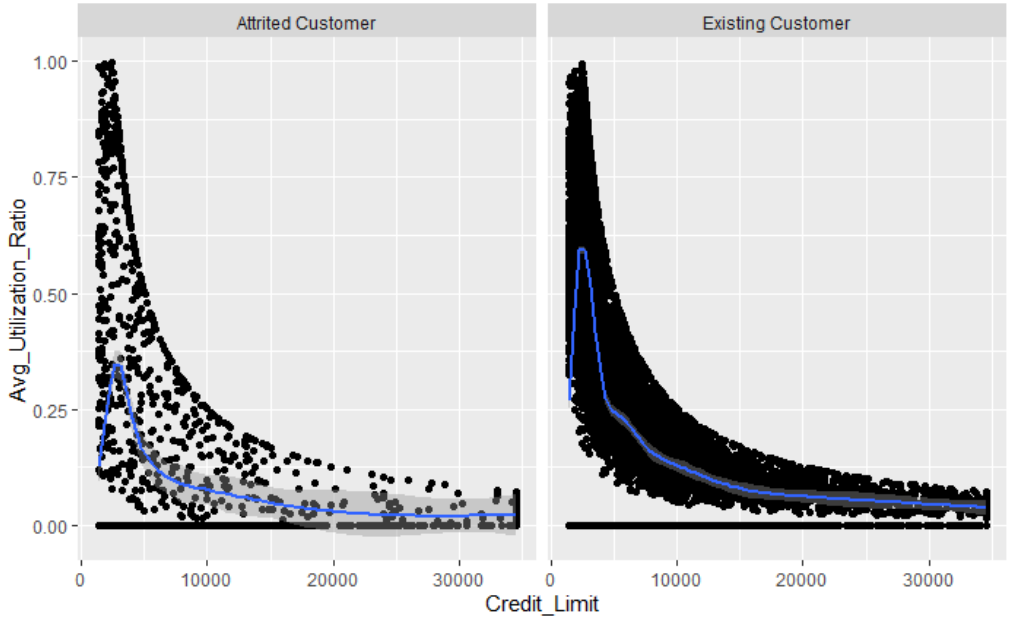

Figure 22 *The scatterplots show the association between "Credit_Limit" and "Avg_Utilization_Ratio" with fitted generalized additive models.*

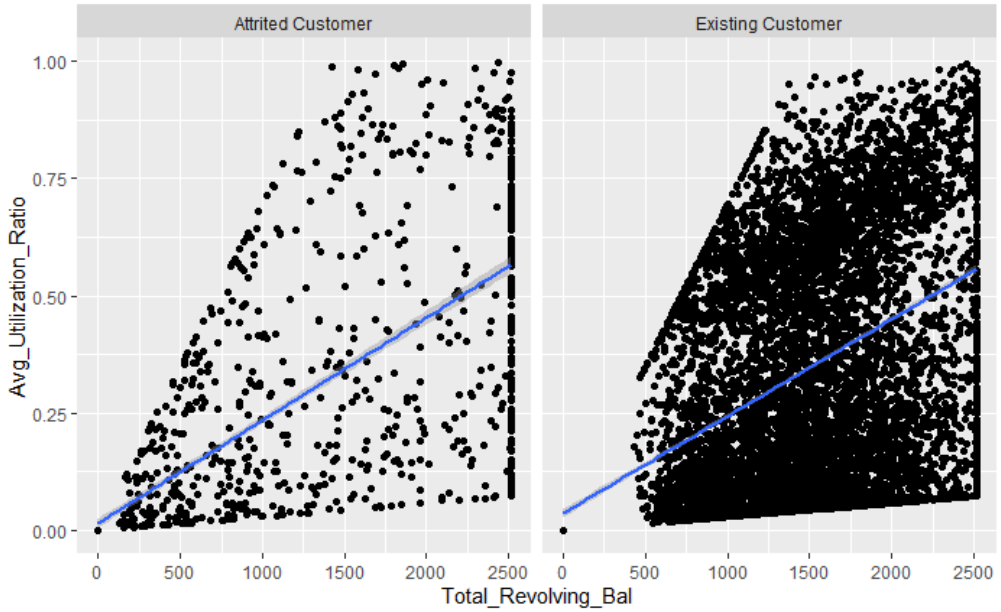

Figure 23 *The scatterplots show the association between "Total_Revolving_Bal" and "Avg_Utilization_Ratio" with fitted linear models.*

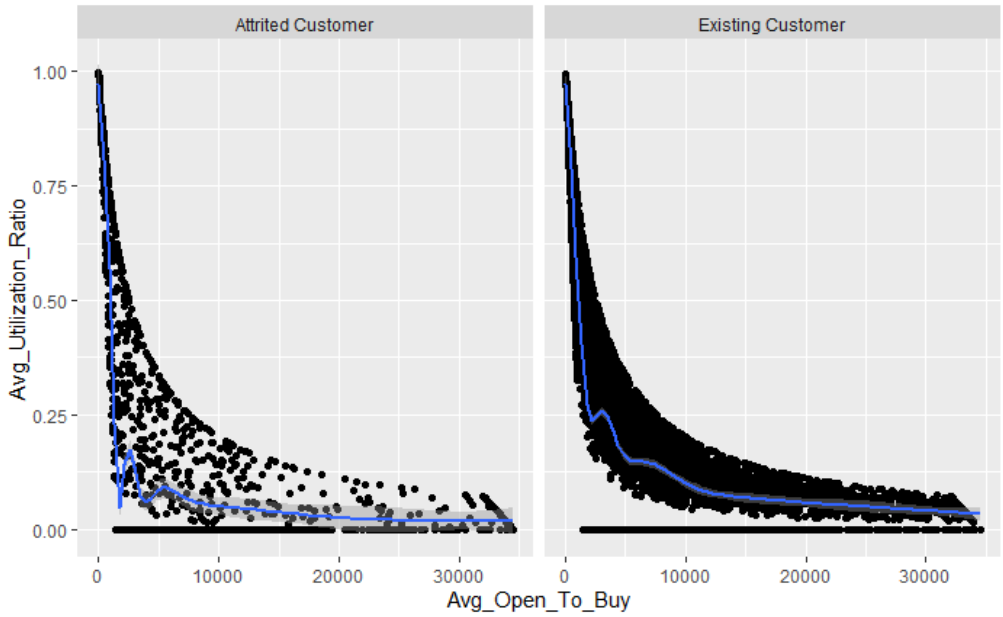

Figure 24 *The scatterplots show the association between "Avg_Open_To_Buy" and "Avg_Utilization_Ratio" with fitted generalized additive models.*

Figure 25 *The scatterplots show the association between "Total_Trans_Amt" and "Total_Trans_Ct" with fitted linear models.*

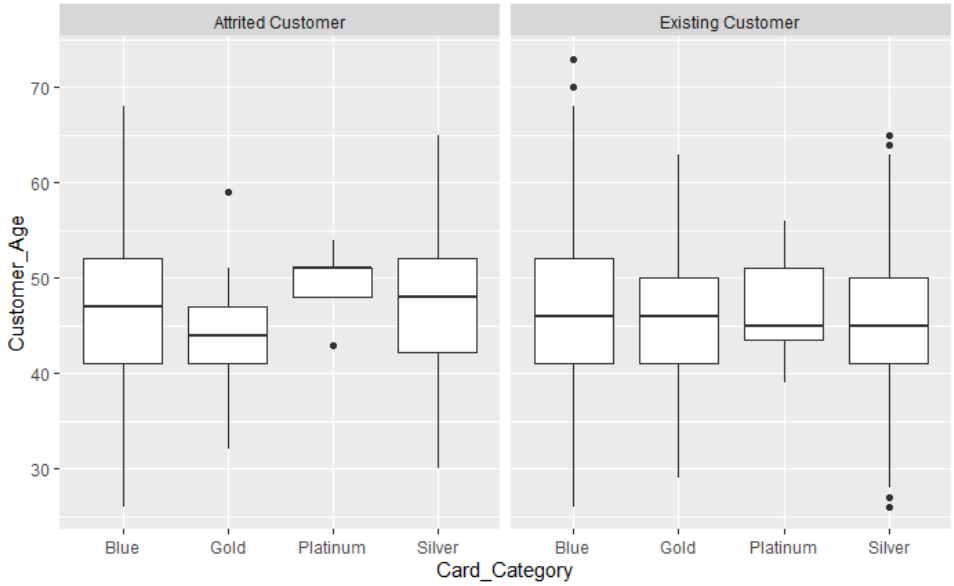

Figure 26 *This shows a comparison between customers' age for the card categories for 'Attrited Customer' versus 'Existing Customer'.*

7 References

James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R.* Springer.

Tang, Travis. (2023, April 24). Class Imbalance Strategies - A Visual Guide with Code. *Medium.* https://medium.com/data-science/class-imbalance-strategies-a-visual-guide-with-code-8bc8fae71e1a#:~:text=Class%20imbalance%20occurs%20when%20one,anomaly%20detection%2C%20and%20medical%20diagnosis.