DALT7016 Data Visualisations Assignment

Submitted By: Ashwini Pant

Student ID: 19212510

Word Count: 1500

Table of Contents

1 Dataset Selection and justification

1.1 Description of the dataset

The dataset I have chosen for this project is the BankChurners dataset (Rastogi, 2022), available on Kaggle. The dataset contains 21 columns and 10,000 rows of data, each representing a bank customer. The variables in the dataset include customer demographics such as age, gender, education level, marital status, and income category, as well as financial variables such as credit limit, total transaction amount, and total revolving balance. The dataset also includes a binary variable indicating whether or not the customer has churned, i.e., closed their account with the bank.

1.2 Justification of its selection

The BankChurners dataset was chosen for this project because it provides an opportunity to explore the factors that contribute to customer churn in the banking industry. Customer churn is a critical issue for banks as it can lead to a loss of revenue and a decrease in customer loyalty. By analysing this dataset, we can gain insights into the factors that influence customer churn and develop models that can predict which customers are likely to churn. This information can then be used to develop strategies to reduce churn rates and improve customer retention.

2 Formulating the brief

2.1 Purpose of the visualisation

The purpose of this project is to analyse the BankChurners dataset and develop visualizations that can help us understand the factors that contribute to customer churn in the banking industry. The audience for this project includes banking executives and data analysts who are interested in understanding the factors that influence customer churn and developing strategies to reduce churn rates. The context for this project is the banking industry, and the goal is to develop insights that can be used to improve customer retention and reduce churn rates.

2.2 Questions that want to be answered

The main questions that I want to answer through this project include:

- What are the demographic and financial characteristics of customers who are most likely to churn?
- Are there any trends or patterns in the data that can help us understand the factors that contribute to customer churn?

- Can I develop models that can predict which customers are likely to churn based on their demographic and financial characteristics?

2.3 Purposed Map

The purpose map for this project is to develop visualizations that can help us understand the factors that contribute to customer churn and develop models that can predict which customers are likely to churn based on their demographic and financial characteristics. By answering these questions, we hope to provide insights that can be used to develop strategies to reduce churn rates and improve customer retention in the banking industry.

3 Analysis of the Data

3.1 Examination and Transformation

The context for this visualization is the banking industry, where customer churn is a critical issue that can impact the profitability and long-term sustainability of banks.

Before conducting the exploratory data analysis, I needed to clean and transform the data. This involved filtering out missing data, eliminating outliers, and creating new variables that could help us answer our research questions. I have performed these transformations using R, and the code is available upon r script and since there were no missing values, no imputation or removal was required for the rows.

```
> #to handle missing values
> sapply(data, function(x) sum(is.na(x)))
```

CLIENTNUM	Avg_Open_To_Buy	Income_Category
0	0	0
Attrition_Flag	Total_Amt_Chng_Q4_Q1	Card_Category
0	0	0
Customer_Age	Total_Trans_Amt	Months_on_book
0	0	0
Gender	Total_Trans_Ct	tal_Relationship_Count
0	0	0
Dependent_count	Total_Ct_Chng_Q4_Q1	Months_Inactive_12_mon
0	0	0
Education_Level	Avg_Utilization_Ratio	Contacts_Count_12_mon
	0	0
		Credit_Limit
		0

Figure 1: No missing values.

Since, the dataset has more than 10 columns, among them I have used only some specific column to plot the visualisation since using that column the visualisation would be meaningful. The columns that I have used are *Attrition_Flag, Customer_Age, Income_Category, Credit_Limit, Total_Trans_Amt and Avg_Utilization_Ratio*. Once I had transformed the data, I have performed exploratory data analysis to understand the patterns and trends in the data. I have created various visualizations to help us explore the data, including bar charts.

Trends or patterns in the data

- An interesting trend observed was that the average age of customers who churned was lower than the average age of customers who did not churn
- Churned customers had a higher average credit utilization ratio and lower credit limit compared to non-churned customers.

4 Editorial thinking

4.1 Prototype of the visualization

- After examining the data and exploring different variables, I have developed five different visualizations to gain insights into the dataset. The visualization aims to identify patterns and insights that can be used to inform customer retention strategies and to build predictive models for identifying customers who are at risk of churning.

1. **Bar Graph** showing the total attrited customer and existing customer. Similarly, the distribution of churn by income category: The purpose of this chart is to understand the relationship between churn and income category. The chart clearly shows that the customers with lower income are more likely to churn than those with higher income.

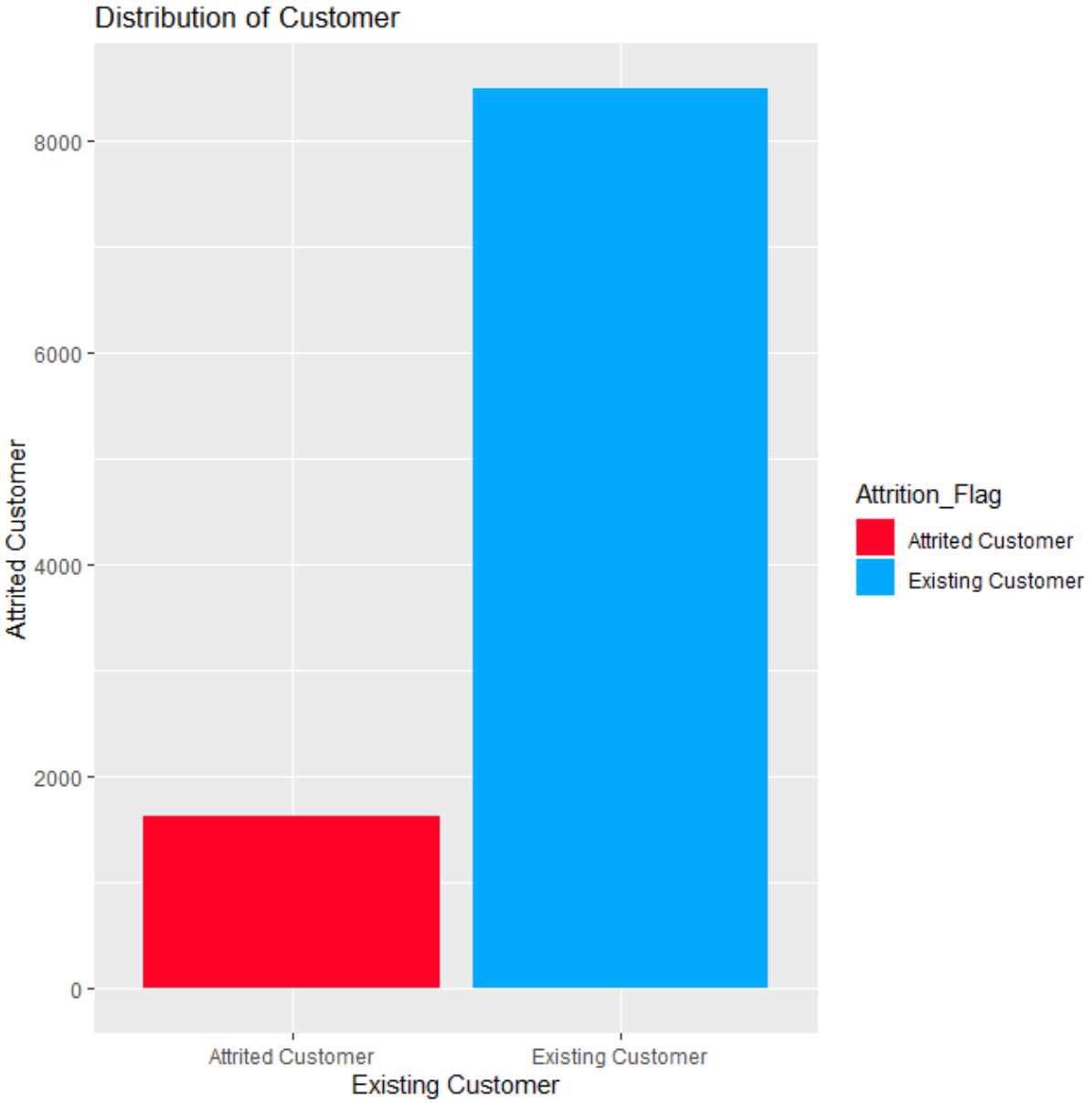

Figure 2: Distribution by Customer.

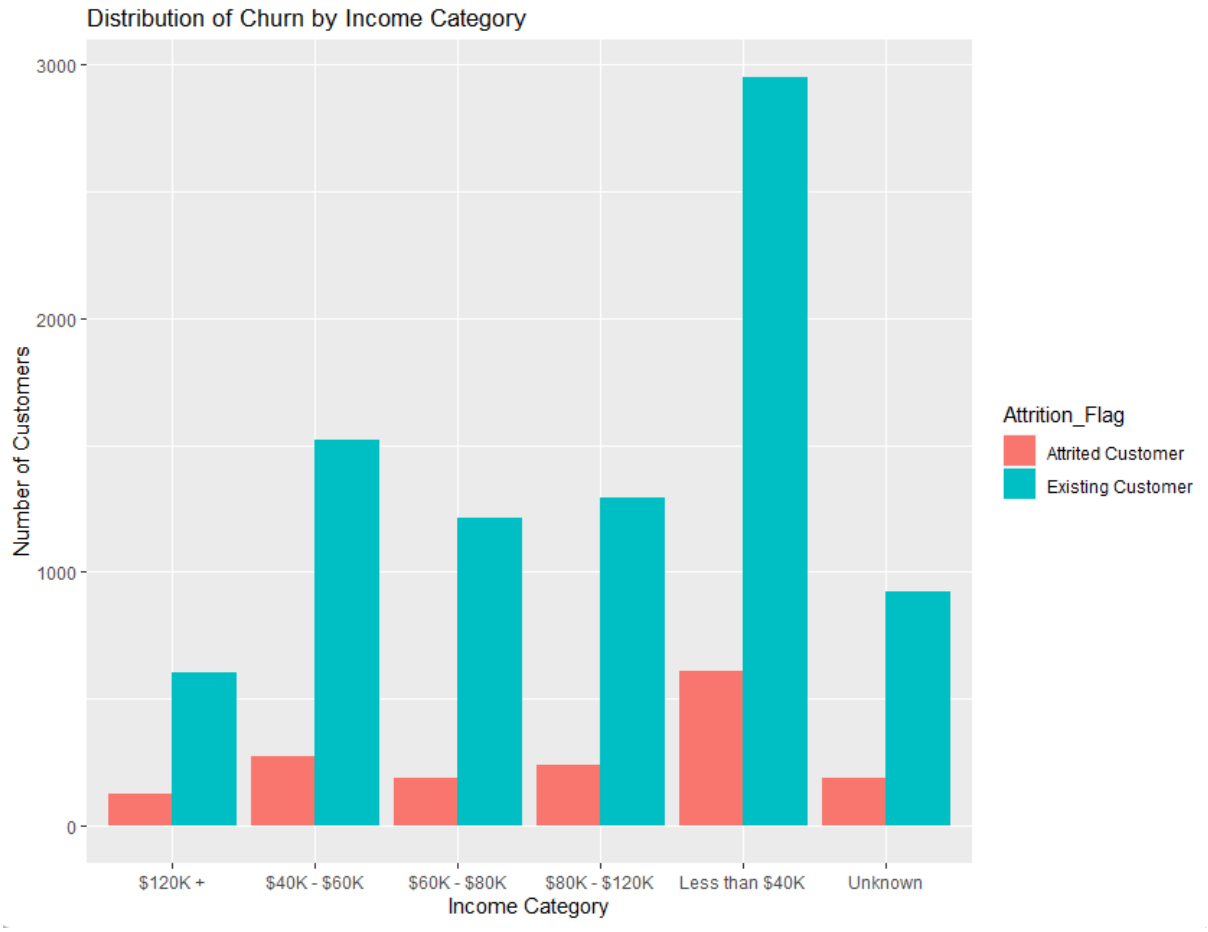

Figure 3: Distribution by income category.

2. **Histogram** showing the distribution of credit limit: The purpose of this chart is to understand the distribution of credit limit among bank customers. The x-axis represents the credit limit range, and the y-axis represents the number of customers falling into that range. The chart indicates that the majority of customers have a credit limit between 0 and 10,000, while only a few customers have a credit limit greater than 20,000.

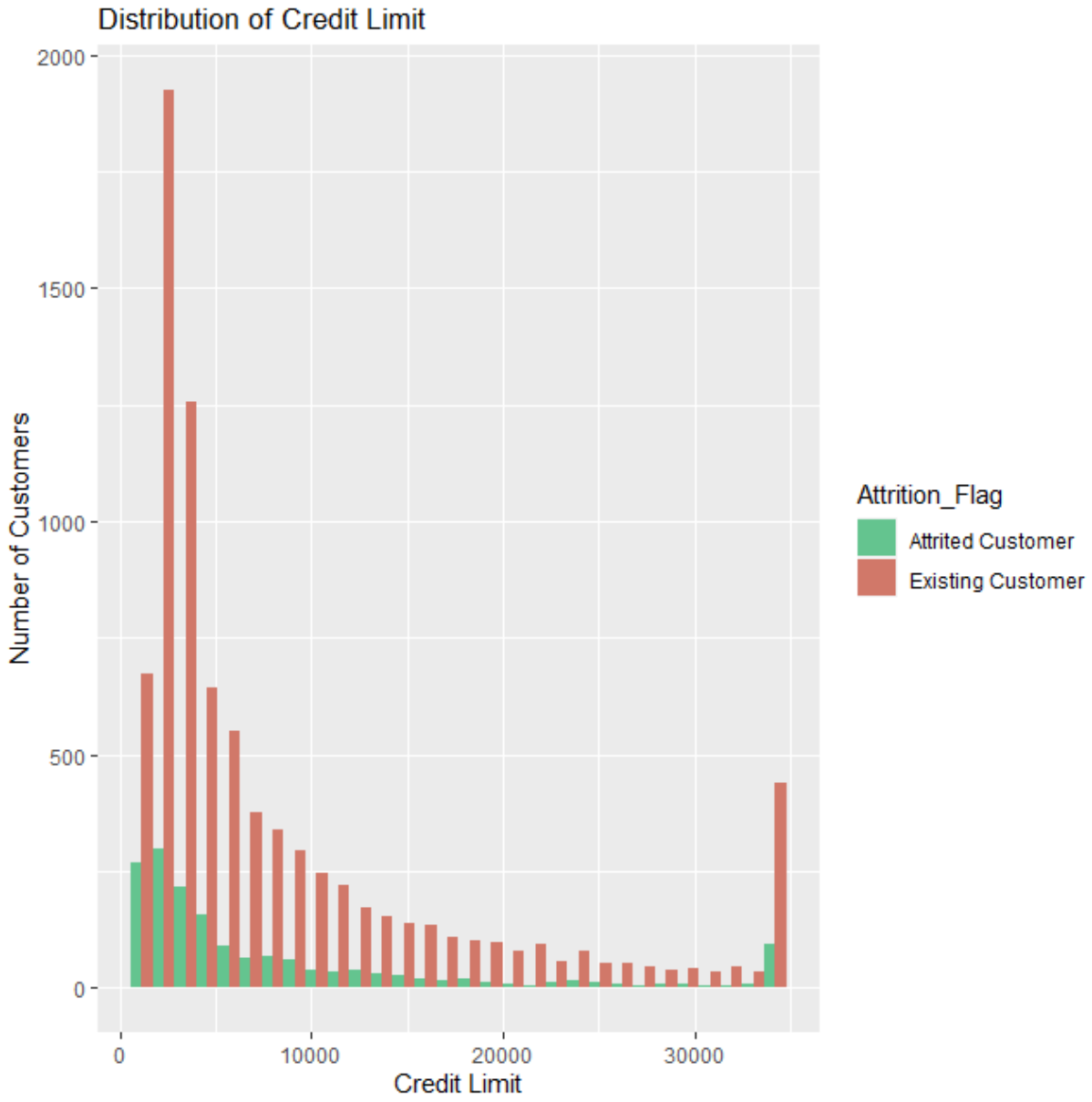

Figure 4: Distribution by credit limit.

3. **Bar Graph** showing the distribution of customers by age: The purpose of this chart is to understand the age distribution of bank customers. The x-axis represents different age groups, and the y-axis represents the number of customers falling into that age group. The chart indicates that the majority of customers fall into the age group of 40-55, while very few customers are younger than 20 or older than 70.

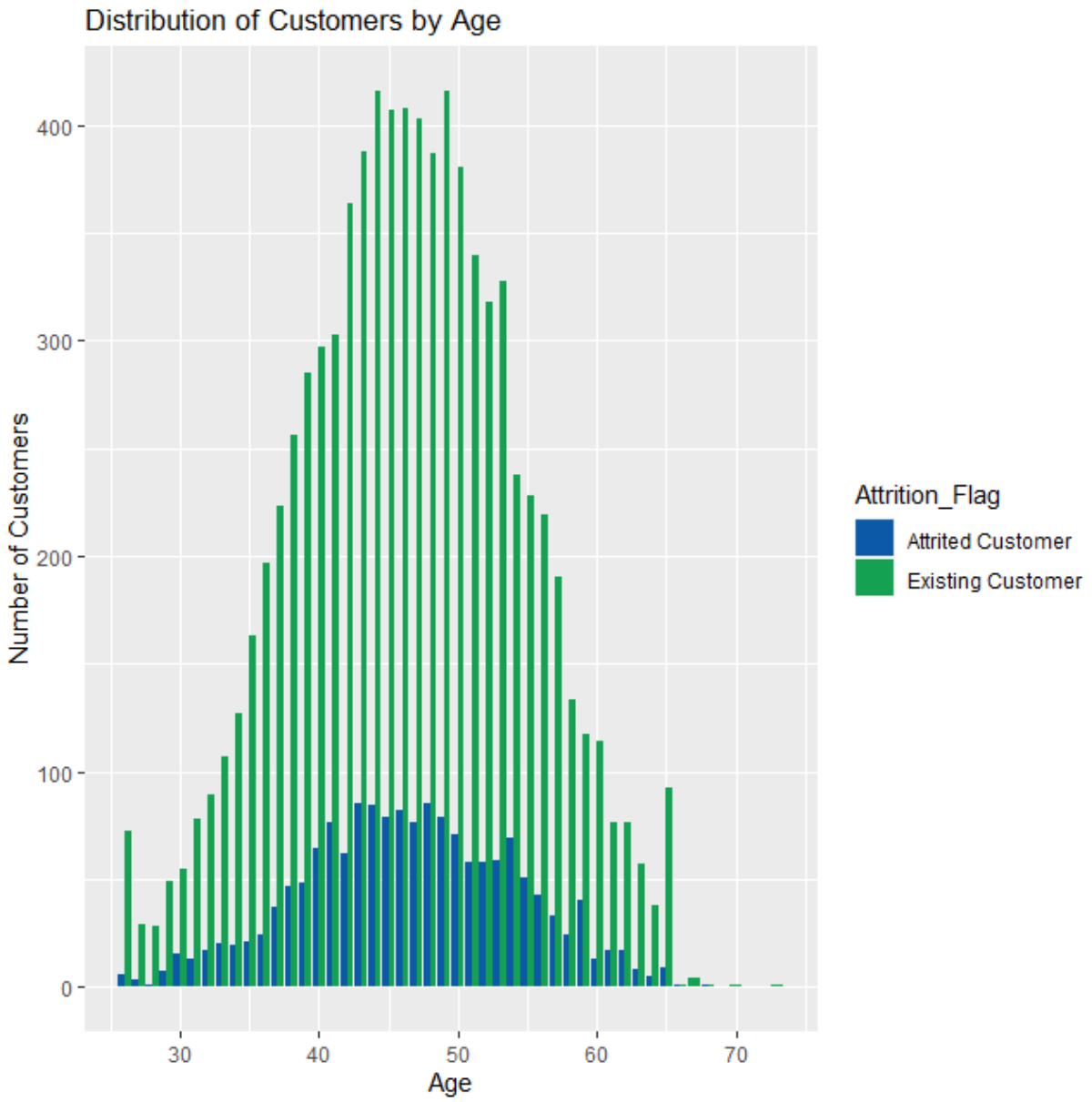

Figure 5: Distribution by Age.

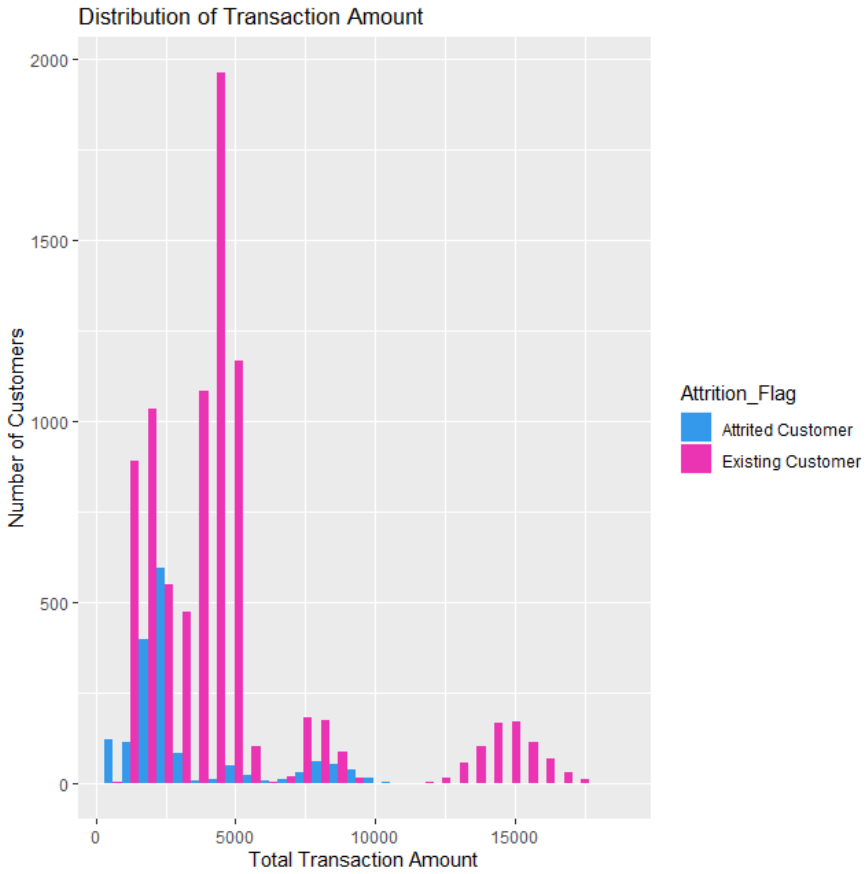

Figure6: Distribution by Total Transaction Amount.

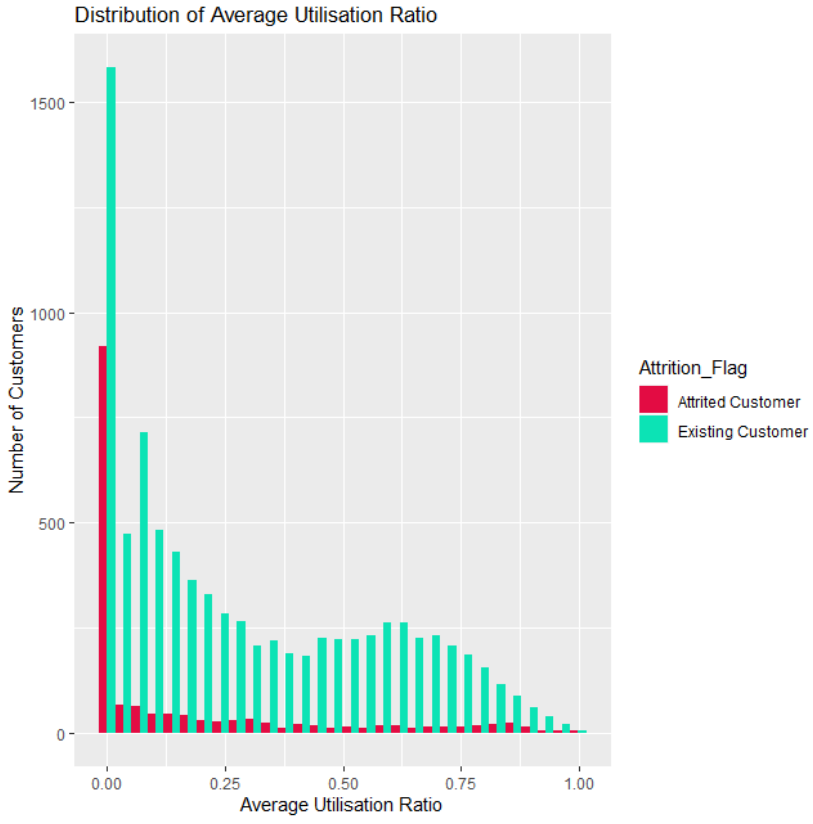

Figure 7: Distribution by Average utilization ratio.

Hence, by looking at the graph we can predict Avg_Utilization_Ratio and Total_Trans_Amt, seems to have effect in customer leaving the service or not.

4.2 Description of the chart types and the main features in the term of colours, composition

The first chart is a simple bar chart that uses two colours pink and blue. Pink colour represents the attrited customer whereas blue colour represents existing customer to represent different income categories. The percentage of customers who churned is shown on the y-axis, and the x-axis represents the different income categories. The chart is easy to understand and shows a clear relationship between income category and churn.

The second chart is a histogram that uses different shades of colours to represent different credit limit ranges. This colours is used for colour blind people. The x-axis represents the credit limit range, and the y-axis represents the number of customers falling into that range. The chart suggests us that the majority of the customers have a credit limit between 0 and 10,000.

The third chart is a simple bar chart represent different age groups. The number of customers in each age group is shown on the y-axis, and the x-axis represents the different age groups. The chart is easy to understand and shows the majority of customers fall into the age group of 40-55.

5 Design Solution

5.1 Code of the final Charts

I have included the R code used to generate each of the final charts in the data_visualization.R script.

5.2 Conclusions and reflections

Overall, the BankChurners dataset provides valuable insights into the factors that contribute to customer churn in the banking industry. Through exploratory data analysis and data visualization, I was able to gain insights into the age distribution, credit limit distribution, and income categories of customers who churned. The visualizations showed that the customers with lower income are more likely to churn than those with higher income. The majority of customers have a credit limit between 0 and 10,000, and the majority of customers fall into the age group of 40-50. However, there are some limitations to this dataset, including the lack of information about the customers' reasons for churning. Future research could investigate the reasons behind customer churn to develop more accurate models that can predict if a customer is likely to churn.

Insights gained from the analysis can be used in real-world scenarios to identify customers who are at a higher risk of churning and take proactive measures to retain them. For example,

the bank can offer personalized promotions and services to customers who have a higher likelihood of churning, such as increasing their credit limit or offering them incentives to use their card more frequently. Additionally, the bank can target marketing efforts towards younger customers to increase engagement and reduce churn rates.

Overall, this project highlights the importance of data visualization in gaining insights into complex datasets and informing decision-making in the banking industry.

6 References

Rastogi, S. (2022) 'Bank Customer Churn Prediction'. Available at: https://www.kaggle.com/code/sudhanshu2198/bank-customer-churn-prediction/input (Accessed.