BU51037 - Data Visualisation for Business ( SEM 1 24/25 )

A blue and white rectangular sign with a shield and crown

Description automatically generated  
  
Name: Shreeja Bhattacharya  
Student ID: 2617228

**Visualizing Customer behavior and its impact on revenue**

Introduction  
Visualization is the easiest and fastest way to convey data and information to an individual. With an objective and current understanding of the market, visualization can be a key process to change the trajectory of any calculated step benefitting an organization. The purpose of the report is to utilize visualization to its optimum capacity to ease the understanding of the dataset we have.  
In the below report, we will be analyzing **“Customer behavior and its impact on revenue”**, leveraging data visualization and exploratory techniques. The dataset chosen for this report, *Bank Transactions*, offers ample columns of data citing the account and their holder details. It includes key factors determining their dynamics and their spending pattern. Through the report, we will be identifying the scope of improvement for the bank.  
The dataset consists of structured data with mixed data types, including numerical values and categorical fields. Each record represents an individual customer, providing a granular perspective on financial engagement. Pre-processing steps included identifying and handling missing or irrelevant data, and ensuring the dataset is optimized for analysis. This step is crucial to mitigate any biases or inaccuracies in insights.  
This analysis will consider the first to explore summary statistics to establish a foundational understanding of the data. Measures such as the mean, median, and standard deviation of key variables will be calculated, complemented by visualizations like histograms and box plots to illustrate distributions. Following this, relationships between variables will be examined using scatter plots, heatmaps, and other visual tools to uncover potential causality or patterns, such as the relationship between credit limits and spending behavior.  
The report aims to narrate meaningful insights derived from the dataset, emphasizing factors that influence customer behavior and revenue generation.

Exploratory Data Analysis  
  
First, we will begin with sorting the data set. The first step of sorting any dataset is figuring out the missing values. Next, eliminate any duplicate values.

A computer screen shot of a computer

Description automatically generated with medium confidence  
*Figure 1: Identifying the missing values and removing the duplicate values*

We will delete the unnecessary data once we verify the above two conditions. As mentioned above the objective, “**Analyze the impact of revenue with respect to customer behavior”,** columns like “CLIENTNUM”, “Education\_Level”, “Marital\_Status” are not required. We will remove them. However, one might argue that “Marital\_Status” can impact certain schemes like home loan, but our primary object is not that for the time being.   
Out of the required columns, we will convert some of them into binary values. This increases efficiency by taking less run time. Additionally, we will be converting all the categorical variables into factors. This is crucial for efficient storage and processing as well as it enables statistical analysis. Going further, we will group variables for faster and cleaner analysis as well as to plot trend lines. For our requirement further, we will be sorting the data by “Credit\_Limit\_Range” and “Total\_Trans\_Amt”. “Credit\_Limit\_Range” is a bin that we have created for simplifying analysis since it transforms continuous variables like “Credit\_Limit” into categorical bins making the data easier to analyze and interpret.

A screenshot of a computer

Description automatically generated

*Figure 2: Cleaned dataset*

For the mean and median, summarizing the clean data will give us the desired output.A screenshot of a computer

Description automatically generated *Figure 3: Mean and Median of the data*  
  
Visualization and Interpretation

Having a clean dataset will make it feasible to visualize the data and identify the points of improvement from the dataset. All the colours used are carefully chosen from the colourblind palette and the visualization is done in RStudio.  
We will be vividly analyzing the below five scenarios:  
  
**1. Age VS Total Transaction Amount  
2. Card Type preference by Income category  
3. Credit Utilization VS Attrition Rate  
4. Total Transaction Amount by Credit Limit Range**  
**5. Average Transaction Amount by Credit Limit Range and Attrition Status**  
  
To analyze our objective first behavioural pattern **“**Age VS Total Transaction Amount**”,** we will consider “Customer\_Age” and “Total\_Trans\_Amt”. In this case, we will be using a Scatter Plot.   
By definition, the **scatter plot** is a graph that displays the relationship between two numerical variables. Each point represents an observation, positioned based on its variable values. It helps identify trends, correlations, and outliers.   
A graph of different colored lines

Description automatically generated  
*Figure 3: Age vs Total Transaction Amount*

The scatter plot shows that total transaction amounts are the highest for the middle age group. Especially, peak in customers aged 40–50 and decline with age additionally, with younger consumers (age <30) contributing less. Hence, we can identify middle-aged customers are the primary contributors, emphasizing the need to prioritize retention strategies for this group, given this is the most profitable. Efforts, like new schemes, should also address declining transactions in older customers (age >60) and enhance engagement among younger consumers to boost transaction volumes. Digitalization for the younger audience can be a great opportunity for a performance boost.

## Card Type Preference by Income Category

A **Horizontal bar chart** illustrates the distribution of card type preferences across different income categories. Each bar's colour represents an income group, with the count displayed along the x-axis.  
According to the graphic, the "Blue" card is the most widely used card type and is overwhelmingly chosen by consumers in all income brackets. This bolsters the observation that most consumers, irrespective of their financial status, choose "Blue" by default. Premium cards like Gold and Platinum are rarely used by higher-income groups. This implies that premium card products may be underutilized among higher-income demographics. Even in higher income brackets, the Silver card is still much less popular than the "Blue" card, although being used more frequently than Gold or Platinum cards.  
Suggestions would be focussing on boosting Premium Card adoption by developing targeted marketing campaigns highlighting the exclusive features, rewards, and benefits of Gold and Platinum cards for high-income consumers. Enhance “Blue” card features by improving its value proposition by offering better incentives as well as placing the “Silver” card better amongst the consumers for better engagement among customers.

A graph with colorful bars

Description automatically generated

*Figure 3: Card Type preference by Income category*

## Credit Utilization VS Attrition Status

**Boxplot:** A statistical graph that displays the distribution of data across different categories using five-number summaries: minimum, first quartile, median, third quartile, and maximum. Outliers are shown as individual points.

The below **boxplot** visualizes the distribution of the average credit utilization ratio across two categories of customer attrition status: attrited customers and existing customers. It summarizes the data using minimum, first quartile, median, third quartile, maximum, and outliers.

The x-axis represents the binary attrition status (0 for attrited and 1 for existing customers), while the y-axis represents the average credit utilization ratio. Each boxplot shows the spread and central tendency of utilization ratios for the two groups, with the median line and diamond-shaped mean indicating key trends.  
Existing customers have a broader spread of credit utilization, with a higher mean and median compared to attrited customers. Attrited customers show lower credit utilization ratios with minimal variation. This suggests that higher utilization ratios may correlate with customer retention. Optimizing credit policies or offering incentives for increased utilization could improve retention rates among lower-utilization customers.

A chart of a credit utility status

Description automatically generated  
*Figure 4: Credit Utilization VS Attrition Status*

## Total Transaction Amount by Credit Limit Range

The boxplot illustrates the relationship between credit limit ranges and total transaction amounts. Median transaction amounts increase with higher credit limits, peaking for "Very High" credit limits, which also show the widest range of transactions. Lower credit limit categories have similar interquartile ranges but include notable outliers, suggesting opportunities for targeted credit limit increases. High-limit customers contribute significantly to revenue, warranting retention strategies and tailored incentives. For outliers in lower credit limits, promotional offers or limit increases could encourage higher engagement. Medium credit limit customers may also benefit from personalized incentives to boost transactions.

A graph of different colored squares

Description automatically generated  
*Figure 5: Total Transaction amount by credit limit range*

## Average Transaction Amount by Credit Limit Range and Attrition Status

**Heatmap:** A heatmap is a graphical representation of data where individual values are depicted using colour gradients, highlighting patterns, trends, and relationships within the dataset.

The heatmap visualizes the average transaction amount segmented by customer attrition status (Existing vs. Attrited) and credit limit ranges. Each tile's colour represents the average transaction amount, as indicated by the colour gradient legend. Lighter colours denote higher average transactions, while darker colours represent lower amounts.  
  
The aforementioned data emphasizes the profitability of customer retention by showing that average transaction quantities for current customers continuously exceed those of attrited customers across all credit limit limits. Retaining high-credit-limit customers is essential to maintaining income because they are the most lucrative and generate the largest transaction amounts. Lower credit limit attrited clients, on the other hand, make very little contribution, indicating that efforts to retain this market may not be very successful. In order to address these findings, companies should concentrate on reducing attrition in medium to high credit ranges while giving priority to keeping high-credit-limit clients through loyalty programs, tailored offers, and credit upgrades. Furthermore, knowing the causes of high-credit-limit customer attrition might help improve retention tactics for maximum effect.

A chart of different shades of blue

Description automatically generated  
*Figure 6: Representation of the Average Transaction amount*

# Conclusion

The analysis identifies key tactics to maximize revenue creation and consumer engagement. Middle-aged clients are prioritized for retention efforts since they account for the largest portion of total transaction volumes, especially those between the ages of 40 and 50. Revenue can be increased by using digital solutions to improve engagement with younger demographics and address the decline in transactions with older clients.   
Even among high-income groups, card type preferences reveal an excessive reliance on the "Blue" card, while premium cards like Gold and Platinum continue to be neglected. Enhancing the value of the "Blue" card guarantees ongoing engagement, while targeted marketing campaigns highlighting premium card perks can encourage acceptance in affluent demographics.

The analysis identifies key tactics to maximize revenue creation and consumer engagement. Middle-aged clients are prioritized for retention efforts since they account for the largest portion of total transaction volumes, especially those between the ages of 40 and 50. Revenue can be increased by using digital solutions to improve engagement with younger demographics and address the decline in transactions with older clients.  
Even among high-income groups, card type preferences reveal an excessive reliance on the "Blue" card, while premium cards like Gold and Platinum continue to be neglected. Enhancing the value of the "Blue" card guarantees ongoing engagement, while targeted marketing campaigns highlighting premium card perks can encourage acceptance in affluent demographics.

Appendix:  
---

title: "Final\_assignment\_2617228"

author: "Shreeja\_2617228"

date: "`r Sys.Date()`"

output: word\_document

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

```{r}

# Load required libraries

library(readxl)

library(ggplot2)

library(dplyr)

library(tidyverse)

# Read the Excel file

file\_path <- "C:/Users/shree/Downloads/Data Visualization/BankData/BankChurners.xlsx"

bank\_data <- read\_excel(file\_path, sheet = "BankChurners")

# Check for missing values

missing\_values <- colSums(is.na(bank\_data))

print("Missing Values in Each Column:")

print(missing\_values)

```

```{r}

# Clean and Prepare Data

# Step 1: Remove duplicates (if any)

bank\_data <- bank\_data %>% distinct()

print("Duplicate Values in Each Column:")

print(bank\_data)

```

```{r}

# Example: Convert "Attrition\_Flag" column to binary

# Existing Customer = 1, Attrited Customer = 0

bank\_data$Attrition\_Flag\_Binary <- ifelse(bank\_data$Attrition\_Flag == "Existing Customer", 1, 0)

```

```{r}

# Create Credit Limit Range bins

bank\_data <- bank\_data %>%

mutate(

Credit\_Limit\_Range = cut(

Credit\_Limit,

breaks = c(0, 5000, 10000, 15000, 20000, Inf),

labels = c("$0-$5K", "$5K-$10K", "$10K-$15K", "$15K-$20K", ">$20K"),

right = FALSE

)

)

```

```{r}

# Ensure all required columns are in proper format

# Convert categorical variables to factors

bank\_data$Attrition\_Flag <- as.factor(bank\_data$Attrition\_Flag)

bank\_data$Card\_Category <- as.factor(bank\_data$Card\_Category)

bank\_data$Income\_Category <- as.factor(bank\_data$Income\_Category)

```

```{r}

#Sort data by Credit Limit Range and Total Transaction Amount

bank\_data <- bank\_data %>%

arrange(Credit\_Limit\_Range, desc(Total\_Trans\_Amt))

# Remove unnecessary columns

bank\_data <- bank\_data[ , !names(bank\_data) %in% c("CLIENTNUM", "Marital\_Status", "Education\_Level")]

```

```{r}

# Display cleaned and sorted data

print("Cleaned and Sorted Data Sample:")

print(head(bank\_data))

view(bank\_data)

```

```{r}

summary(bank\_data)

```

```{r}

# Load necessary libraries

library(ggplot2)

library(readxl)

library(viridis) # For colorblind-friendly palettes

# Convert 'Attrition\_Flag\_Binary'

bank\_data$Attrition\_Flag\_Binary <- as.factor(bank\_data$Attrition\_Flag\_Binary)

# Create the boxplot with colorblind-friendly colors and a legend

ggplot(bank\_data, aes(x = Attrition\_Flag\_Binary, y = Avg\_Utilization\_Ratio, fill = Attrition\_Flag)) +

geom\_boxplot(outlier.shape = NA, alpha = 0.7) + # Remove outliers for clarity and add transparency

stat\_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "white") + # Add mean points

scale\_fill\_viridis(discrete = TRUE, option = "D", name = "Attrition Status") + # Use Viridis palette and add legend title

labs(

title = "Credit Utilization vs Attrition Status",

x = "Attrition Status (Binary)",

y = "Average Utilization Ratio"

) +

theme\_minimal() +

theme(

legend.position = "right", # Position the legend on the right

plot.title = element\_text(hjust = 0.5) # Center-align title

)

```

```{r}

# Create age group bins for categorization

bank\_data$Age\_Group <- cut(

bank\_data$Customer\_Age,

breaks = c(18, 30, 40, 50, 60, 70, Inf),

labels = c("18-30", "31-40", "41-50", "51-60", "61-70", "71+"),

right = FALSE

)

# Create the scatter plot with colorblind-friendly colors

ggplot(bank\_data, aes(x = Customer\_Age, y = Total\_Trans\_Amt, color = Age\_Group)) +

geom\_point(alpha = 0.7) + # Semi-transparent points

geom\_smooth(method = "loess", color = "black", se = FALSE) + # Black trend line without confidence interval

scale\_color\_viridis(discrete = TRUE, option = "D", name = "Age Groups") + # Colorblind-friendly color palette

labs(

title = "Age vs Total Transaction Amount",

x = "Age",

y = "Total Transaction Amount"

) +

theme\_minimal() +

theme(

plot.title = element\_text(hjust = 0.5), # Center-align title

legend.position = "right" # Position legend on the right

)

```

```{r}

# Create Credit Limit Range bins for grouping

bank\_data <- bank\_data %>%

mutate(

Credit\_Limit\_Bins = cut(

Credit\_Limit,

breaks = c(0, 5000, 10000, 15000, 20000, Inf),

labels = c("$0-$5K", "$5K-$10K", "$10K-$15K", "$15K-$20K", ">$20K"),

right = FALSE

)

)

# Calculate the average transaction amount for each Credit Limit Range and Attrition Status

trans\_heatmap <- bank\_data %>%

group\_by(Credit\_Limit\_Bins, Attrition\_Flag) %>%

summarise(Avg\_Trans\_Amt = mean(Total\_Trans\_Amt, na.rm = TRUE)) %>%

ungroup()

# Create the heatmap

ggplot(trans\_heatmap, aes(x = Credit\_Limit\_Bins, y = Attrition\_Flag, fill = Avg\_Trans\_Amt)) +

geom\_tile(color = "white") + # Add white gridlines between tiles

scale\_fill\_viridis(option = "G", name = "Avg Transaction Amount") + # Use colorblind-friendly palette

labs(

title = "Average Transaction Amount by Credit Limit Range and Attrition Status",

x = "Credit Limit Range",

y = "Attrition Status"

) +

theme\_minimal() +

theme(

plot.title = element\_text(hjust = 0.5), # Center-align the title

)

```

```{r}

ggplot(data\_summary, aes(x = Count, y = reorder(Card\_Category, Count), fill = Income\_Category)) +

geom\_bar(stat = "identity", position = "dodge") +

scale\_fill\_viridis(discrete = TRUE, name = "Income Category") +

labs(

title = "Card Type Preferences by Income Category",

x = "Count",

y = "Card Category"

) +

theme\_minimal() +

theme(

legend.position = "bottom"

)

```

```{r}

# Create bins for Credit\_Limit

data$Credit\_Limit\_Bins <- cut(

data$Credit\_Limit,

breaks = quantile(data$Credit\_Limit, probs = seq(0, 1, 0.2), na.rm = TRUE), # Use quintiles for balanced bins

labels = c("Very Low", "Low", "Medium", "High", "Very High"),

include.lowest = TRUE

)

# Boxplot: Total Transaction Amount by Credit Limit Range

ggplot(data, aes(x = Credit\_Limit\_Bins, y = Total\_Trans\_Amt, fill = Credit\_Limit\_Bins)) +

geom\_boxplot(color = "black", outlier.color = "red", outlier.size = 1.5) + # Boxplot with colored outliers

scale\_fill\_viridis\_d(option = "viridis") + # Apply colorblind-friendly palette

labs(

title = "Total Transaction Amount by Credit Limit Range",

x = "Credit Limit Range",

y = "Total Transaction Amount",

fill = "Credit Limit Range"

) +

theme\_minimal() +

theme(

plot.title = element\_text(size = 14, hjust = 0.5), # Center-align and increase title size

axis.text.x = element\_text(size = 10), # Increase x-axis text size

axis.text.y = element\_text(size = 10) # Increase y-axis text size

)

```

```{r}

```