ALY6000 Introduction to Analytics

Northeastern University

Final Project Summary

Date: 05/19/2023

Submitted To: Richard He

Submitted By: Shree Vipulbhai Tejani

Credit Card Data Analysis

Part 1

Business Question:

What is the distribution of cardholders among various brands like Visa, MasterCard, Amex and Discover?

Calculating frequency count for cards having a sensor chip and not having a chip?

What is the average credit limit across all users?

Analysis:

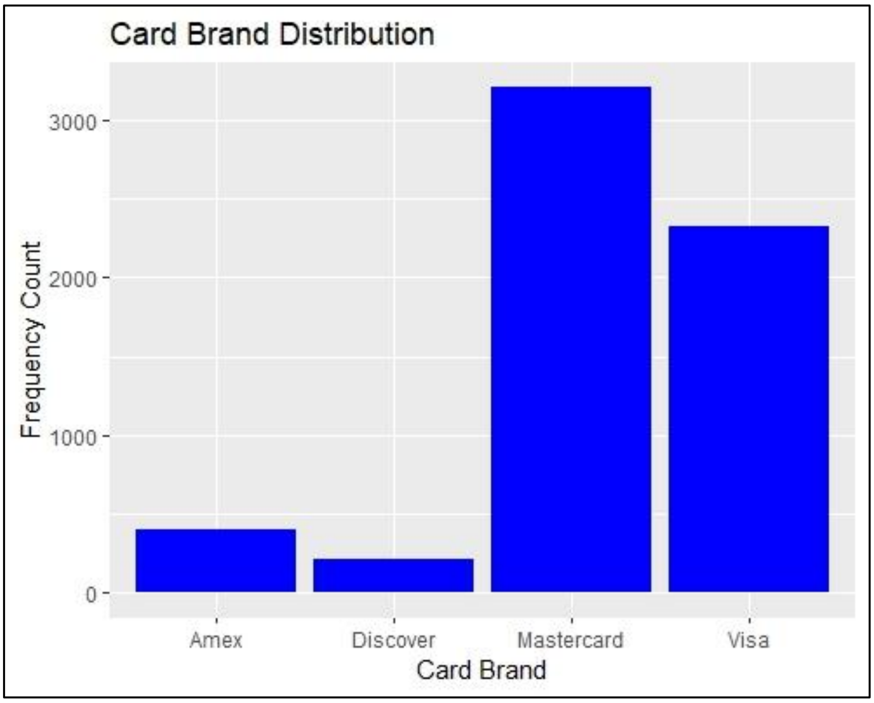
```
> struc <- str(sd254_cards)
'data.frame':	6146 obs. of  13 variables:
 $ User                : int  0 0 0 0 0 1 1 1 1 1 ...
 $ CARD.INDEX          : int  0 1 2 3 4 0 1 2 3 4 ...
 $ Card.Brand          : chr  "Visa" "Visa" "Visa" "Visa" ...
 $ Card.Type           : chr  "Debit" "Debit" "Debit" "Credit" ...
 $ Card.Number         : num  4.34e+15 4.96e+15 4.58e+15 4.88e+15 5.72e+1
5 ...
 $ Expires             : chr  "12/2022" "12/2020" "02/2024" "08/2024" ...
 $ CVV                 : int  623 393 719 693 75 736 972 48 722 908 ...
 $ Has.Chip            : chr  "YES" "YES" "YES" "NO" ...
 $ Cards.Issued        : int  2 2 2 1 1 1 2 2 2 1 ...
 $ Credit.Limit        : chr  "$24295" "$21968" "$46414" "$12400" ...
 $ Acct.Open.Date      : chr  "09/2002" "04/2014" "07/2003" "01/2003" ...
 $ Year.PIN.last.Changed: int  2008 2014 2004 2012 2009 2012 2011 2015 201
5 2012 ...
 $ Card.on.Dark.Web    : chr  "No" "No" "No" "No" ...
```

Table 1: Summary of the numerical values dataset.

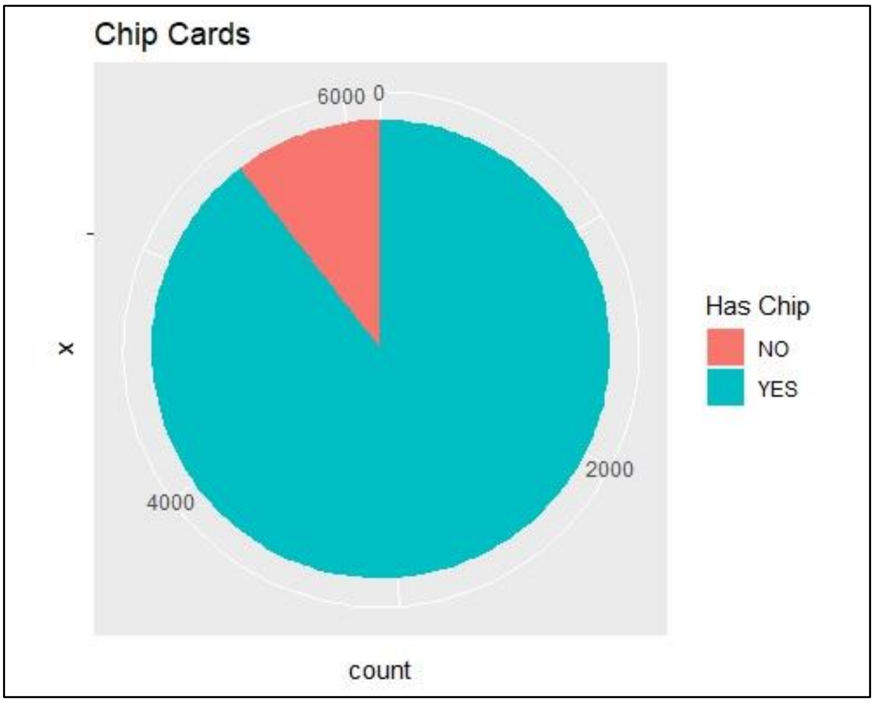
For example, the summary table provides summary statistics for numerical variables, such as count, mean, minimum, maximum, and quartiles.

```
> summary(sd254_cards$Card.Number)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
3.001e+14 4.486e+15 5.109e+15 4.820e+15 5.585e+15 6.997e+15 

> summary( sd254_cards$CVV)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0   257.0   516.5   506.2   756.0   999.0 

> summary(sd254_cards$Cards.Issued)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   1.000   1.503   2.000   3.000 

> summary(sd254_cards$Credit.Limit)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0    7043   12592   14347   19157  151223 

> summary(sd254_cards$Acct.Open.Date)
   Length     Class      Mode 
     6146 character character 

> summary(sd254_cards$sd254_cards$Year.PIN.last.Changed)
Length  Class   Mode 
     0   NULL   NULL 
```

Graph 1: The bar chart illustrates the frequency of different card brands.



Graph 2: The pie chart shows the proportion of cards with and without chips.



The above-mentioned descriptive analyses offer information and solutions to numerous business queries using credit card data.

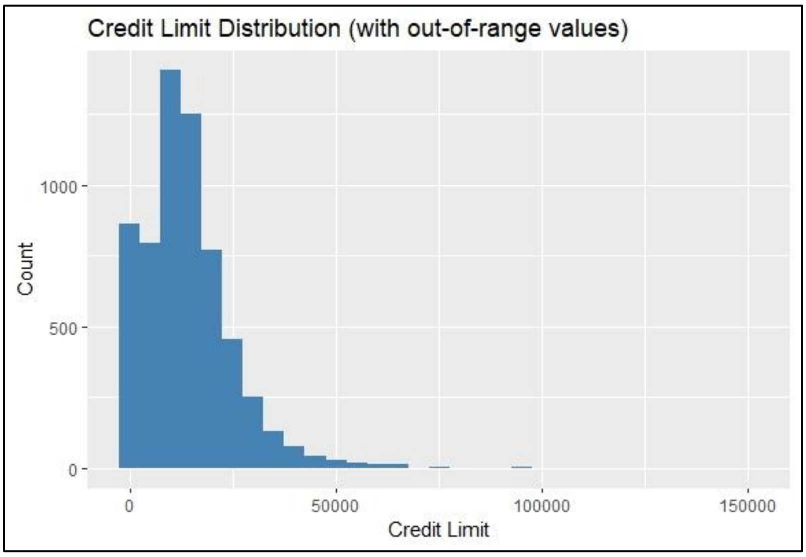
The variables in the dataset are summarized in the summary table. For numerical variables, it covers count, mean, minimum, maximum, and quartiles. We can comprehend the range and distribution of numerical variables like Credit Limit, Cards Issued, CVV, and Year PIN last changed thanks to this study. It aids in locating any odd numbers or probable outliers in the dataset.

The bar chart shows how frequently various card brands are used. It gives a general overview of the acceptance and popularity of various card brands among consumers like MasterCard followed by Visa. It is possible to tell which brands are more and less common (Amex & Discover) by looking at the chart. Understanding market share, brand loyalty, and future alliances or collaborations with particular card brands can all be aided by this data.

The percentage of cards with and without chips is depicted in a pie chart. It makes the adoption of chip technology by credit card customers easier to understand. The percentage of cards that have chip technology—a crucial security feature—implemented can be found by examining the chart. The ability to evaluate the security level offered by the cards and comprehend the market trend toward chip-enabled cards depends on this information.

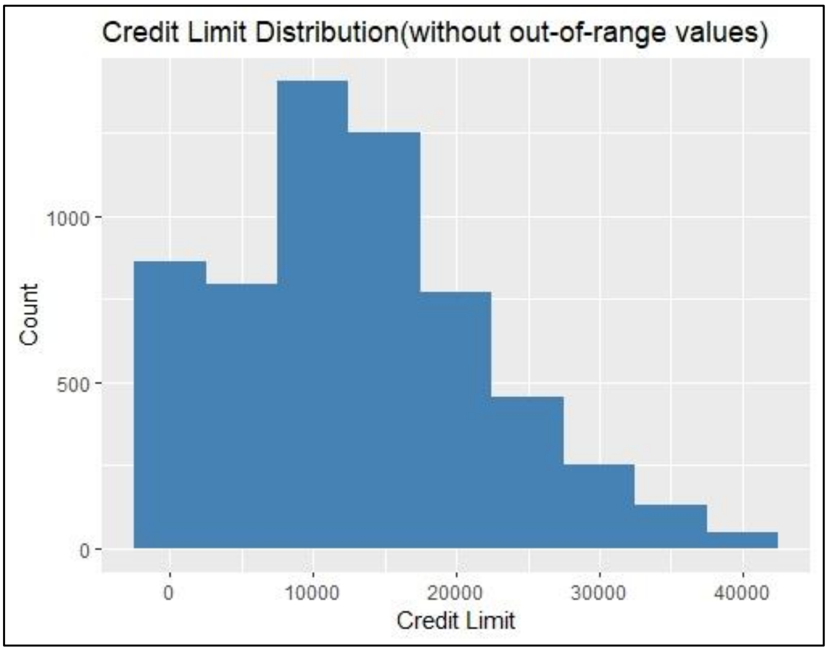
With Out-Of-Range Values (Original data)

```
"Summary Table (with out-of-range values):"
print(summary_table)
  Average_Cards_Issued Average_Credit_Limit
1             1.503091             14347.49
```



Without Out-Of-Range Values (Filtered data)

```
"Summary Table (without out-of-range values):”
  Average_Cards_Issued  Average_Credit_Limit
              1.504526              13056.42
```



We can also filter the Credit.Limit and display required information by removing the ‘out-of-range’ values. The above data displays both ‘out-of-range’ and ‘without out-of-range’ values. This helps in providing better visualization and customizing the output results based on the client’s requirement.

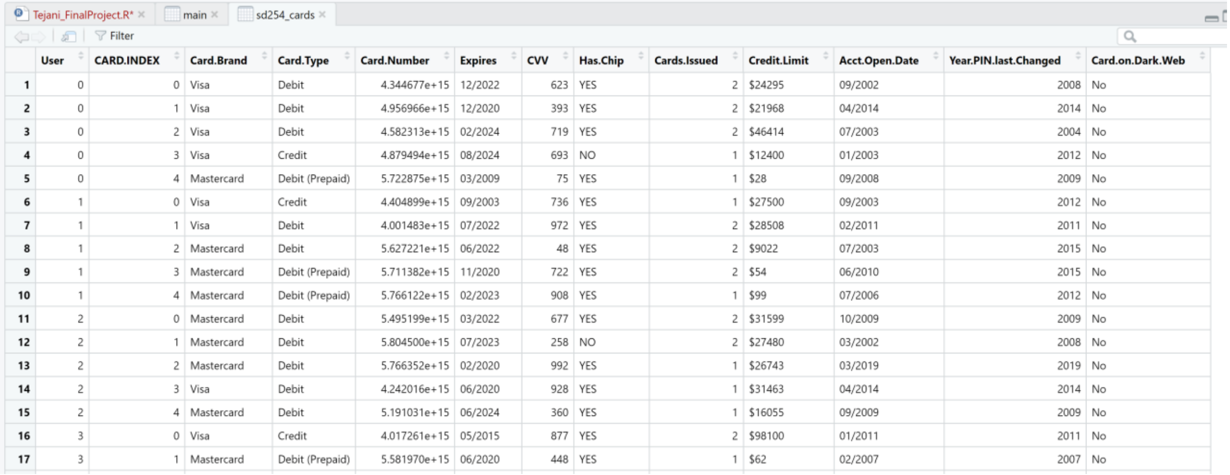
Additional Questions that can be asked by looking at this dataset’s analysis are:

Is there a correlation between the number of cards issued and the credit limit? How has the number of cards issued per user changed over time? Is there a relationship between the card brand and the presence on the dark web?

More details or information regarding user demographics, transaction history, or security measures may be required in order to respond to these inquiries. Such information would enable a more thorough analysis of credit card activity and associated risk factors.

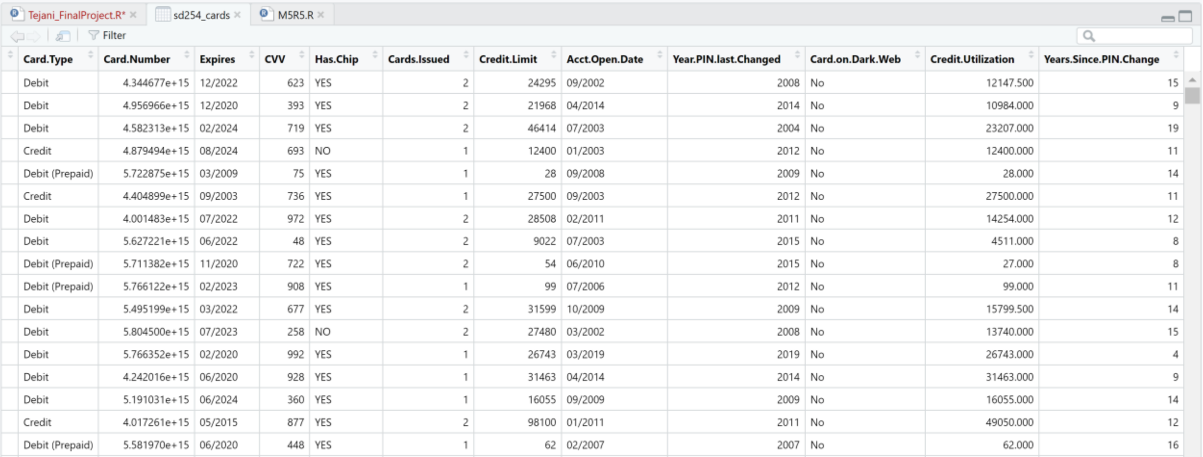
Part 2

Original Dataset



	User	CARD.INDEX	Card.Brand	Card.Type	Card.Number	Expires	CVV	Has.Chip	Cards.Issued	Credit.Limit	Acct.Open.Date	Year.PIN.last.Changed	Card.on.Dark.Web
1	0	0	Visa	Debit	4.344677e+15	12/2022	623	YES	2	$24295	09/2002	2008	No
2	0	1	Visa	Debit	4.956966e+15	12/2020	393	YES	2	$21968	04/2014	2014	No
3	0	2	Visa	Debit	4.582313e+15	02/2024	719	YES	2	$46414	07/2003	2004	No
4	0	3	Visa	Credit	4.879494e+15	08/2024	693	NO	1	$12400	01/2003	2012	No
5	0	4	Mastercard	Debit (Prepaid)	5.722875e+15	03/2009	75	YES	1	$28	09/2008	2009	No
6	1	0	Visa	Credit	4.404899e+15	09/2003	736	YES	1	$27500	09/2003	2012	No
7	1	1	Visa	Debit	4.001483e+15	07/2022	972	YES	2	$28508	02/2011	2011	No
8	1	2	Mastercard	Debit	5.627221e+15	06/2022	48	YES	2	$9022	07/2003	2015	No
9	1	3	Mastercard	Debit (Prepaid)	5.711382e+15	11/2020	722	YES	2	$54	06/2010	2015	No
10	1	4	Mastercard	Debit (Prepaid)	5.766122e+15	02/2023	908	YES	1	$99	07/2006	2012	No
11	2	0	Mastercard	Debit	5.495199e+15	03/2022	677	YES	2	$31599	10/2009	2009	No
12	2	1	Mastercard	Debit	5.804500e+15	07/2023	258	NO	2	$27480	03/2002	2008	No
13	2	2	Mastercard	Debit	5.766352e+15	02/2020	992	YES	1	$26743	03/2019	2019	No
14	2	3	Visa	Debit	4.242016e+15	06/2020	928	YES	1	$31463	04/2014	2014	No
15	2	4	Mastercard	Debit	5.191031e+15	06/2024	360	YES	1	$16055	09/2009	2009	No
16	3	0	Visa	Credit	4.017261e+15	05/2015	877	YES	2	$98100	01/2011	2011	No
17	3	1	Mastercard	Debit (Prepaid)	5.581970e+15	06/2020	448	YES	1	$62	02/2007	2007	No

With New Variables added



Card.Type	Card.Number	Expires	CVV	Has.Chip	Cards.Issued	Credit.Limit	Acct.Open.Date	Year.PIN.last.Changed	Card.on.Dark.Web	Credit.Utilization	Years.Since.PIN.Change
Debit	4.344677e+15	12/2022	623	YES	2	24295	09/2002	2008	No	12147.500	15
Debit	4.956966e+15	12/2020	393	YES	2	21968	04/2014	2014	No	10984.000	9
Debit	4.582313e+15	02/2024	719	YES	2	46414	07/2003	2004	No	23207.000	19
Credit	4.879494e+15	08/2024	693	NO	1	12400	01/2003	2012	No	12400.000	11
Debit (Prepaid)	5.722875e+15	03/2009	75	YES	1	28	09/2008	2009	No	28.000	14
Credit	4.404899e+15	09/2003	736	YES	1	27500	09/2003	2012	No	27500.000	11
Debit	4.001483e+15	07/2022	972	YES	2	28508	02/2011	2011	No	14254.000	12
Debit	5.627221e+15	06/2022	48	YES	2	9022	07/2003	2015	No	4511.000	8
Debit (Prepaid)	5.711382e+15	11/2020	722	YES	2	54	06/2010	2015	No	27.000	8
Debit (Prepaid)	5.766122e+15	02/2023	908	YES	1	99	07/2006	2012	No	99.000	11
Debit	5.495199e+15	03/2022	677	YES	2	31599	10/2009	2009	No	15799.500	14
Debit	5.804500e+15	07/2023	258	NO	2	27480	03/2002	2008	No	13740.000	15
Debit	5.766352e+15	02/2020	992	YES	1	26743	03/2019	2019	No	26743.000	4
Debit	4.242016e+15	06/2020	928	YES	1	31463	04/2014	2014	No	31463.000	9
Debit	5.191031e+15	06/2024	360	YES	1	16055	09/2009	2009	No	16055.000	14
Credit	4.017261e+15	05/2015	877	YES	2	98100	01/2011	2011	No	49050.000	12
Debit (Prepaid)	5.581970e+15	06/2020	448	YES	1	62	02/2007	2007	No	62.000	16

Credit Utilization (New Variable 1)

```
> cat("Mean Credit Utilization:", mean_credit_utilization, "\n")
Mean Credit Utilization: 11035.43

> cat("Median Credit Utilization:", median_credit_utilization, "\n")
Median Credit Utilization: 8700
```

Mean Years Since PIN Change (New Variable 2)

```
> cat("Mean Years Since PIN Change:", mean_years_since_pin_change, "\n")
Mean Years Since PIN Change: 9.563293

> cat("Median Years Since PIN Change:", median_years_since_pin_change, "\n
")
Median Years Since PIN Change: 10
```

The average amount of credit being used by users, as calculated by the mean credit utilization, is 11,035.43. The midpoint value in the distribution, the median Credit Utilization is 8,700. Given that the mean is greater than the median, we can infer that the distribution of credit use is probably skewed to the right. This implies that there are certain instances of high credit consumption that are affecting the average.

The average number of years since a PIN change is 9.563293, meaning that customers last updated their PIN on average 9.6 years ago. Ten represents the median number of years since a PIN change. The modest discrepancy between the mean and median suggests that the variable has a fairly symmetrical distribution.

The mean and median values give information about the central tendency of the variables overall. However, it is crucial to take the distribution shape into account, especially when the mean and median diverge. This can aid in locating any potential outliers or skewed data points that might affect how the analysis is interpreted.

Part 3

Numerous observations and follow-up inquiries might be made in light of the data analysis.

The distribution of card brands is shown graphically by a bar chart, and it reveals that the dataset's most prevalent card brand is MasterCard. Follow Up Question's can be, What are the causes of Master cards' greater popularity compared to other brands? Are there any special features or collaborations that increase the user attractiveness of Master?

According to the histogram showing the distribution of credit limits, there is a wide variety of credit limitations, with some users having large credit limits. Do credit limitations and other factors, such as card type or the number of years from account inception, have any correlations?

The median number of years since the last PIN change is 10, with a mean of roughly 9.6 years. This suggests that, generally speaking, people haven't updated their PIN in a while. What steps may be taken to encourage more frequent PIN changes? Are there any security dangers connected with infrequent PIN changes?

It is feasible to learn more about the traits, their relationships, and potential ramifications for users' financial stability by digging deeper into these findings and follow-up inquiries.

Conclusion:

In conclusion, the dataset analysis revealed certain important characteristics of credit cards. With some individuals having high credit use rates, the data showed that the average credit utilization was $11,035.43. Additionally, it has been 9.6 years on average since the last PIN change, which suggests that PIN security procedures may need to be strengthened. Visa was the most common brand of card in the dataset, according to the distribution of card brands, and a wide range of credit limits were emphasized by the distribution of credit limits. Concerns concerning card security were raised by the existence of cards that were detected as being on the dark web. Further research into these characteristics and follow-up inquiries can help reveal deeper trends and guide initiatives to enhance cardholders' financial security and well-being.

Bibliography:

Kabacoff, R. I. (2015). *R in action: Data analysis and graphics with R*. Manning.

ERIK ALTMAN, Apoorva Nitsure IBM, Youssef Mroueh. (n.d.). *Credit Card Transactions*. Kaggle. https://www.kaggle.com/datasets/ealtman2019/credit-card-transactions

Appendix:

```
library(dplyr)
library(ggplot2)

#import data
sd254_cards <- read.csv("D:\\MPS_Quater 1\\ALY6000_Intro to Analytics\\sd254_cards.csv", header=TRUE, stringsAsFactors=FALSE)
sd254_cards

#Part 1
#dataset structure
struc <- str(sd254_cards)

# Remove the dollar sign ($) from 'Credit.Limit' variable
sd254_cards$Credit.Limit <- as.numeric(gsub("\\$", "", sd254_cards$Credit.Limit))
# Print the updated 'Credit.Limit' values
print(sd254_cards$Credit.Limit)

#summary statistics
summary(sd254_cards$Card.Number)
summary(sd254_cards$Expires)
summary( sd254_cards$CVV)
summary(sd254_cards$Cards.Issued)
summary(sd254_cards$Credit.Limit)
summary(sd254_cards$Acct.Open.Date)
summary(sd254_cards$sd254_cards$Year.PIN.last.Changed)

# Summarize the data in a table
summary_table <- summary(sd254_cards)
print(summary_table)
```

```
# Graphs to visualize the data
# Example bar chart to give frequency count of cards brand
bar_chart <- ggplot(sd254_cards, aes(x = Card.Brand)) +
  geom_bar(fill = "blue") +
  labs(title = "Card Brand Distribution", x = "Card Brand", y = "Frequency Count")
print(bar_chart)
# Example pie chart for displaying if card has sensor chips
pie_chart <- ggplot(sd254_cards, aes(x = "", fill = Has.Chip)) +
  geom_bar(width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Chip Cards", fill = "Has Chip")
print(pie_chart)

#check unsual values
# Clean the data by removing out-of-range values for Credit.Limit
data_clean <- sd254_cards %>% filter(Credit.Limit >= 0 & Credit.Limit <= 40000)
data_clean
#(with out-of-range values) ORIGINAL Data
# 1. Summarize the data in a table
summary_table <- sd254_cards %>%
  summarise(
    Average_Cards_Issued = mean(Cards.Issued),
    Average_Credit_Limit = mean(Credit.Limit)
  )
print("Summary Table (with out-of-range values):")
print(summary_table)
# Histogram: Credit Limit Distribution
credit_limit_chart <- ggplot(sd254_cards, aes(x = Credit.Limit)) +
  geom_histogram(binwidth = 5000, fill = "steelblue") +
  labs(title = "Credit Limit Distribution (with out-of-range values)", x = "Credit Limit", y = "Count")
print(credit_limit_chart)


# (without out-of-range values) FILTERED DATA CREDIT.LIMIT OF $40000
# Summarize the data in a table
summary_table_clean <- data_clean %>%
  summarise(
    Average_Cards_Issued = mean(Cards.Issued),
    Average_Credit_Limit = mean(Credit.Limit)
  )
print("Summary Table (without out-of-range values):")
print(summary_table_clean)
# Histogram: Credit Limit Distribution
credit_limit_chart_clean <- ggplot(data_clean, aes(x = Credit.Limit)) +
  geom_histogram(binwidth = 5000, fill = "steelblue") +
  labs(title = "Credit Limit Distribution(without out-of-range values)", x = "Credit Limit", y = "Count")
print(credit_limit_chart_clean)


#Part 2
#New variable 1
# Calculate difference in credit limit utilization
sd254_cards$Credit.Utilization <- sd254_cards$Credit.Limit / sd254_cards$Cards.Issued
head(sd254_cards,5)
#New Variable 2
# Calculate difference in years since the PIN was last changed
current_year <- 2023  # Assuming the current year is 2023
sd254_cards$Years.Since.PIN.Change <- current_year - sd254_cards$Year.PIN.last.Changed
```

```
# Compute mean and median for the new variables
mean_credit_utilization <- mean(sd254_cards$Credit.Utilization)
median_credit_utilization <- median(sd254_cards$Credit.Utilization)
cat("Mean Credit Utilization:", mean_credit_utilization, "\n")
cat("Median Credit Utilization:", median_credit_utilization, "\n")

mean_years_since_pin_change <- mean(sd254_cards$Years.Since.PIN.Change)
median_years_since_pin_change <- median(sd254_cards$Years.Since.PIN.Change)
cat("Mean Years Since PIN Change:", mean_years_since_pin_change, "\n")
cat("Median Years Since PIN Change:", median_years_since_pin_change, "\n")
```