Term: Fall 2023

Course Title: Statistics for Data Analysts (DSE 598)

Project Report on

Analysis of Credit Card Customer Information Dataset

Submitted to:

Dr. Rong Pan

Submitted by:

	Name	ASU ID	Email ID
1.	**Nayan Bhiwapurkar**	1229979233	nbhiwapu@asu.edu
2.	**Kaumudi Patil**	1230036771	kspatil1@asu.edu
3.	**Shreyas Hingmire**	1230415136	shingmir@asu.edu
4.	**Sharan Patel**	1230596772	spate294@asu.edu

TABLE OF CONTENTS

Abstract:

Credit cards have become a necessary tool in today's financial environment, influencing the direction of consumer finance. This emphasizes how important it is for financial institutions to study credit card user behavior in detail because it has a direct impact on strategic decision-making, profitability, and customer satisfaction. The goal of this study is to better understand credit card customer dynamics by finding hidden patterns, trends, and correlations in the given dataset.

The significance of this analysis resides in its ability to identify important factors influencing consumer behavior, separate variables influencing credit card use, and highlight patterns that could predict customer attrition. Financial institutions can optimize services and customize offerings to meet the changing needs and expectations of their clientele by utilizing data-driven approaches.

Keywords:

Customer Attrition, Utilization Ratio, Credit Card Analysis, Logistic Regression, Customer Segmentation, Financial Services, Retention Strategies.

Objective:

To analyze and derive insights from a dataset containing credit card customer information.

Introduction:

The widespread use of credit cards has fundamentally reshaped the financial landscape, serving not only as a practical payment method but also as a key determinant of consumer behavior and financial decision-making. In this dynamic context, analyzing credit card customer data has become essential for financial institutions seeking to enhance their understanding of consumer behavior, streamline operations, and make informed strategic choices. Recognizing credit cards as a vital component of consumer finance, financial institutions must grasp the nuances of user behavior to make strategic decisions, ensure profitability, and satisfy customers. This research aims to provide valuable insights to stakeholders in the financial sector by identifying underlying patterns, trends, and correlations within the dataset.

A credit card customer information dataset covering a wide range of parameters, including client demographics, income, credit limits, card kinds, and transaction history, is thoroughly examined in this data analysis project. The main objective is to derive significant insights that go beyond simple statistical synopses and offer a sophisticated comprehension of consumer behavior and inclinations concerning credit card utilization. This analysis will reveal important aspects that influence consumer behavior, pinpoint variables that affect credit card usage, and reveal trends that could lead to customer attrition.

Dataset Collection and Characteristics:

The dataset for this credit card customer churn prediction analysis was obtained from a source is known for its reliability and relevance to the credit card industry, making it suitable for investigating customer churn patterns. The dataset provides a subset of credit card customer information in 2018 and 2019. The dataset provides a comprehensive view of customer information in a banking context, encompassing both numerical and categorical variables. Key demographic details include client numbers, ages, genders, and marital statuses, while financial aspects are represented by annual incomes, credit limits, and card categories. Additionally, the dataset captures customer behavior, such as transaction counts, months of inactivity, and contact frequency with the bank. Notably, the "Attrition_Flag" variable serves as a critical indicator, categorizing customers as either "Attrited" or "Existing," shedding light on account closure activity. The inclusion of attrition-related time stamps, including quarters and years, facilitates a temporal analysis of customer churn.

This dataset's richness extends beyond mere customer profiles, allowing for a nuanced exploration of factors influencing attrition. The variety of categorical variables, such as education levels and income categories, enables segmentation analysis to discern patterns among different customer groups. Numerical variables, including credit limits and transaction counts, provide a quantitative basis for examining customer financial behavior. The three columns which we are going to focus more on are Credit_Limit, Attrition_Flag and Avg_Utilization_Ratio. With these attributes, the dataset is well-suited for predictive modeling and exploratory data analysis aimed at understanding the dynamics of customer attrition within the banking context.

Problems found with Dataset and how we tackled them:

Data Imbalance: The Credit_Limit column contained outliers. We removed these outliers as these could influence the Hypothesis.

Feature Selection: Analyzed correlations and distributions of features. We removed some features which did not contribute much to the hypothesis testing.

Explorative Data Analysis:

Column	Description	Variable Type	mean	std	min	max
CLIENTNUM	Unique number for the customer	Numerical				
Attrition_Flag	customer activity variable - if the account is closed then "Attrited Customer" else "Existing Customer"	Categorical				
Customer_Age	Age in Years	Numerical	46	8	24	65
Gender	Gender of the account holder - M / F	Categorical				
Dependent_count	Number of dependents	Numerical	2	1	0	5
Education_Level	Educational Qualification of the account holder - College, Doctorate, Graduate, High School, Post-Graduate, Uneducated	Categorical				
Marital_Status	Marital Status of the account holder - Divorced, Married, Single	Categorical				
Income_Category	Annual Income Category of the account holder - Less than $40K, $40K - $60K, $60K - $80K, $80K - $120K, $120K +	Categorical				
Card_Category	Type of Card - Blue, Silver, Gold, Platinum	Categorical				
Months_on_book	Period of relationship with the bank	Numerical	41	10	13	68
Total_Relationship_Count	Total no. of products held by the customer	Numerical	4	2	1	6
Months_Inactive_12_mon	No. of months inactive in the last 12 months	Numerical	3	2	0	6
Contacts_Count_12_mon	No. of Contacts between the customer and bank in the last 12 months	Numerical	2	1	0	6
Credit_Limit	Credit Limit on the Credit Card	Numerical	8637	9084	1400	35000
Total_Revolving_Bal	The balance that carries over from one month to the next is the revolving balance	Numerical	1011	658	0	12080
Avg_Open_To_Buy	Open to Buy refers to the amount left on the credit card to use (Average of last 12 months)	Numerical	7480	9103	3	34516
Total_Trans_Ct	Total Transaction Count (Last 12 months)	Numerical	68	27	10	139
Avg_Utilization_Ratio	Represents how much of the available credit the customer spent	Numerical	0	0	0	1
Quarter	Attrition Quarter - none, Q1, Q2, Q3, Q4	Categorical				
Year	Attrition Year - 2018, 2019	Categorical				
Date_Leave	Attrition Date - Quarter, Year	Categorical				

Table name: Summary of Statistics of the Dataset

Data Visualization:

The visualizations above provide insights into key attributes of the dataset, such as age distribution, credit limit variability, and transaction behavior. They help in understanding the range and patterns within these variables, which is essential for data exploration and analysis.

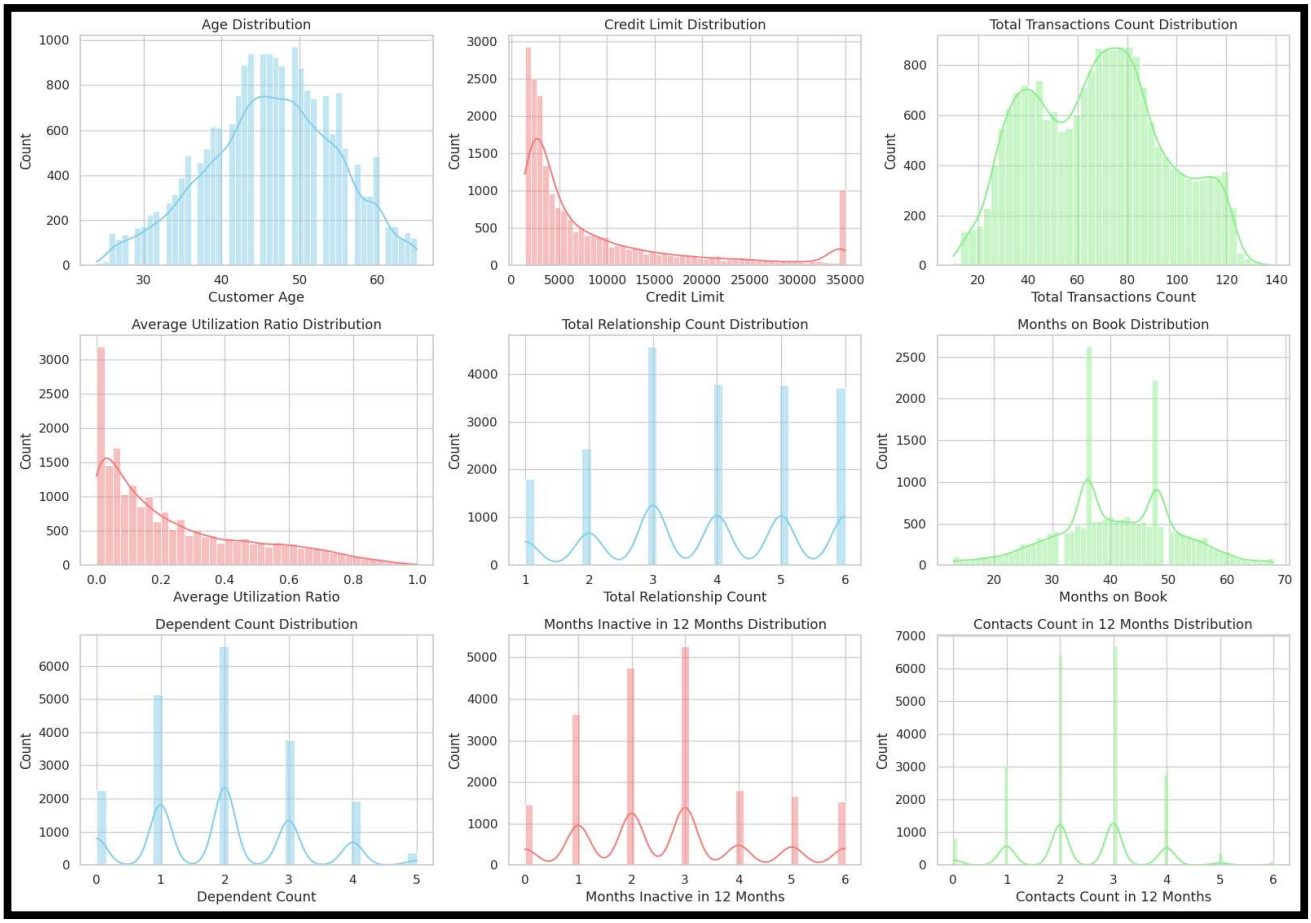

Fig: EDA (Exploratory Data Analysis)

Problem / Hypothesis of interest:

First Hypothesis investigates whether there is a significant gender disparity among individuals with high credit limits or is there any other factor which influences credit Limit.

Second hypothesis explores the potential relationship between credit card utilization and the likelihood of customers experiencing attrition, suggesting that a lower utilization ratio may be associated with a higher probability of attrition.

Importance of the solution:

- Strategic Decision-Making

This data-driven investigation holds immense significance for stakeholders in the financial sector, serving as a compass for strategic decision-making. The comprehensive analysis delves beyond surface-level statistics, providing a nuanced understanding of consumer behavior within the credit card services ecosystem. Armed with these insights, financial institutions can make informed decisions on product development, marketing strategies, and customer engagement initiatives. By discerning patterns and trends, stakeholders can proactively address potential challenges and capitalize on emerging opportunities, fostering a proactive approach that ensures the sustained relevance and competitiveness of credit card services in a rapidly evolving market.

- Risk Mitigation and Customer Satisfaction:

Beyond strategic advantages, the solution plays a pivotal role in risk mitigation and enhancing customer satisfaction. Through meticulous examination of factors influencing attrition, financial institutions can identify early warning signs and implement targeted retention strategies. This proactive approach not only minimizes the financial impact of customer churn but also contributes to a positive customer experience. Armed with a deep understanding of customer preferences and behaviors, institutions can tailor services to meet evolving expectations, fostering long-term loyalty. Ultimately, the project's insights have the potential to create a win-win scenario, reducing risks for financial institutions while enhancing the overall satisfaction and loyalty of credit card customers.

Methodology of Hypothesis Test - 1:

Objective:

To test whether the percentage of males within the population having a credit limit exceeding $25,000 surpasses 90%.

1. Data Collection and Preparation:

- Collect a representative sample of individuals with a credit limit greater than 25,000 from the dataset.
- Extract relevant columns, 'Gender' and 'Credit_Limit', from the dataset.
- Subset the data to include only those with credit limits exceeding 25,000.

2. Define Hypotheses:

- Null Hypothesis (H_0): H_0: $p \geq 0.90$
  - the percentage of males within the population having a credit limit exceeding $25,000 25,000 is greater than 90%

- Alternative Hypothesis (H_1): H_1: $p < 0.90$
  - the percentage of males within the population having a credit limit exceeding $25,000 is less than 90%.

3. Data Analysis:

- Count the number of men (num_men) in the subset.
- Count the total number of individuals (total_individuals) in the subset.
- Calculate the proportion of men (proportion_men).

4. Set Expected Proportion:

- Set the expected proportion under the null hypothesis (expected_proportion = 0.90).

5. Perform Z-Test:

- Used the 'proportions_ztest' function from 'statsmodels' to perform a one-sided proportion test.
- Set the null hypothesis value, alternative hypothesis, and calculate the Z-statistic and p-value.

6. Interpret Results:

- Print the results, including the proportion of men, Z-statistic, and p-value.
- Compare the p-value with the chosen significance level ($\alpha = 0.05$).
- If the p-value is less than α, reject the null hypothesis.

7. Decision:

- Proportion of Men: 0.903954802259887
- Z-Statistic: 0.5646771591959807
- P-Value: 0.7138533136322209

Fail to reject the null hypothesis.

There is enough evidence to suggest that the percentage of males within the population having a credit limit exceeding $25,000 is MORE than 90%.

z test

- Null Hypothesis (H0): The proportion of men among individuals with a credit limit greater than 25,000 is greater than or equal to 90%.
- Alternative Hypothesis (H1): The proportion of men among individuals with a credit limit greater than 25,000 is less than 90%.

```
from statsmodels.stats.proportion import proportions_ztest

# We have a DataFrame called 'credit_data' with columns 'Gender' and 'Credit_Limit'

# Extract a subset of data with credit limit > 25000
high_credit_data = credit_data[credit_data['Credit_Limit'] > 25000]

# Count the number of men in the subset
num_men = high_credit_data[high_credit_data['Gender'] == 'M'].shape[0]

# Count the total number of individuals in the subset
total_individuals = high_credit_data.shape[0]

# Calculate the proportion of men
proportion_men = num_men / total_individuals

# Set the expected proportion under the null hypothesis
expected_proportion = 0.9

# Perform a one-sided proportion test
z_statistic, p_value = proportions_ztest(num_men, total_individuals, value=expected_proportion, alternative='smaller')

# Print the results
print("Proportion of Men:", proportion_men)
print("Z-Statistic:", z_statistic)
print("P-Value:", p_value)

# Interpret the results
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis - There is enough evidence to suggest that the proportion of men is less than 90%.")
else:
    print("Fail to reject the null hypothesis - There is enough evidence to suggest that the proportion of men is MORE than 90%.")
```

```
Proportion of Men: 0.903954802259887
Z-Statistic: 0.5646771591959807
P-Value: 0.7138533136322209
Fail to reject the null hypothesis - There is enough evidence to suggest that the proportion of men is MORE than 90%.
```

Fig: Code snippet for z-test

- While exploring the factors influencing credit limits, we investigated various variables. We wanted to find if 'Gender' really affects credit limit or not, and is 'Gender' the only factor impacting credit limits.
- Instead, we observed a strong correlation between credit limits and two key factors: Education Level and Income Category.
- Specifically, individuals with higher education levels tend to belong to higher income categories, and those in higher income categories generally have higher credit limits.
- This suggests that Education Level and Income Category play pivotal roles in determining credit limits, overshadowing any discernible influence from gender.

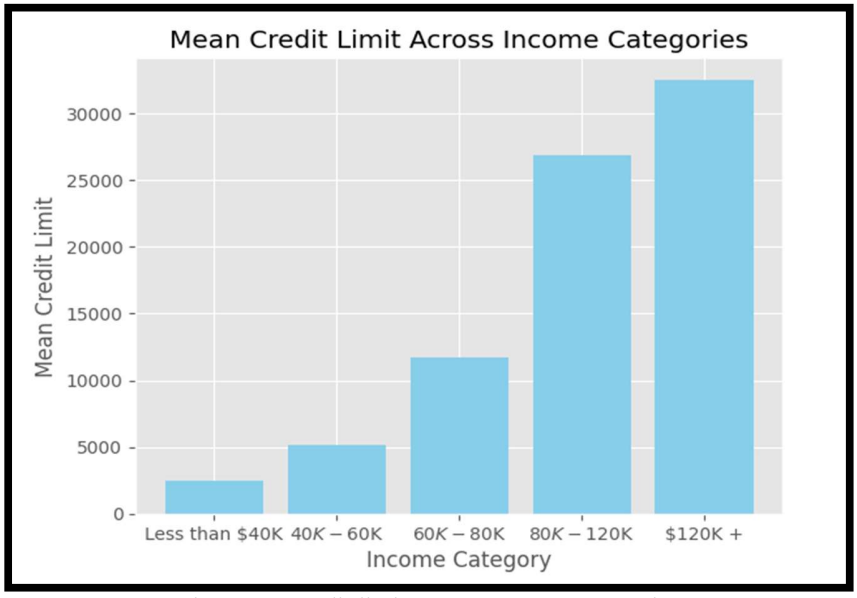

Fig: Mean credit limit across Income Categories

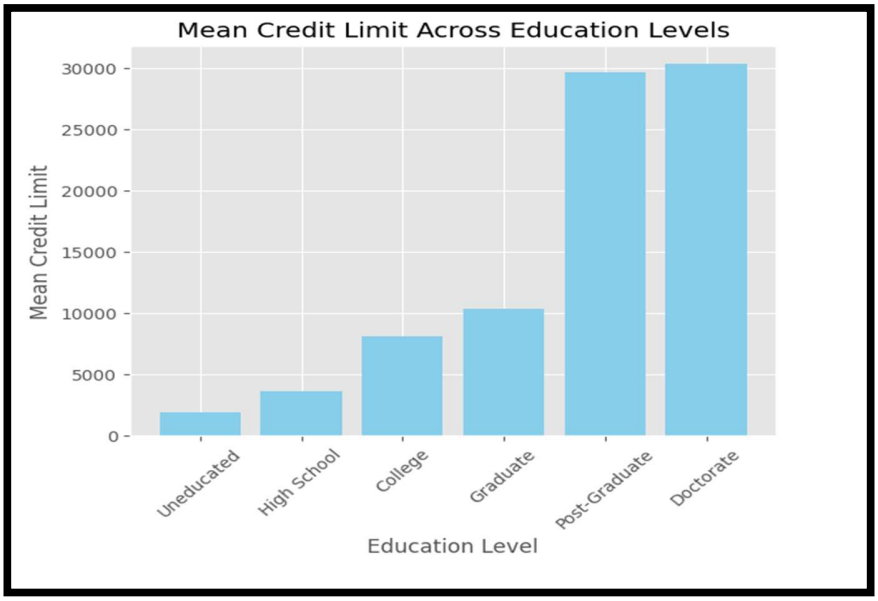

Fig: Mean credit limit across Education Levels

Next, we wanted to find whether 'Gender' directly affects the 'Credit Limit.'

So, we decided to perform **Chi-square test** to find out if there is any relation between gender and credit limit.

Chi-square test:

Methodology of Hypothesis Testing using Chi-Square Test:

Objective:

To test whether there is an association between gender and credit limit among individuals with a credit limit greater than 25,000.

1. Data Collection and Preparation:

- Collect a representative sample of individuals with a credit limit greater than 25,000 from the dataset.
- Extract relevant columns, 'Gender' and 'Credit_Limit', from the dataset.
- Subset the data to include only those with credit limits exceeding 25,000.

2. Define Hypotheses:

- Null Hypothesis (H_0): H_0: There is no association between gender and credit limit.
- Alternative Hypothesis (H_1): H_1: There is an association between gender and credit limit.
- If the p-value is less than your chosen significance level (e.g., 0.05), you would reject the null hypothesis.

3. Data Analysis:

- Create a contingency table (2 × 1) summarizing the counts of individuals for each gender category.

4. Perform Chi-Square Test:

- Use the 'chi2_contingency' function from 'scipy.stats' to perform the chi-square test.
- Calculate the chi-square statistic chi-square), p-value, degrees of freedom (dof), and expected frequencies.

5. Interpret Results:

- Print the results, including the chi-square statistic and p-value.
- Compare the p-value with the chosen significance level (e.g., $\alpha = 0.05$).
- If the p-value is less than α, reject the null hypothesis.

6. Decision:

- P-Value: 1.0
- Fail to reject the null hypothesis.

So Null Hypothesis is true. There is not enough evidence to suggest an association between gender and credit limit.

Hence, gender does not play pivotal role in deciding the credit limit, but education level and salary surely affects the credit limit.

chi-square test

- to prove there is no relation between gender and credit limit

In this test:

- Null Hypothesis (H0): There is no association between gender and credit limit.
- Alternative Hypothesis (H1): There is an association between gender and credit limit.

```
[ ]  from scipy.stats import chi2_contingency

     # We have the DataFrame 'credit_data' with columns 'Gender' and 'Credit_Limit'

     # Extract a subset of data with credit limit > 25000
     high_credit_data = credit_data[credit_data['Credit_Limit'] > 25000]

     # Create a contingency table
     contingency_table = pd.crosstab(high_credit_data['Gender'], columns='count')

     # Perform the chi-square test
     chi2_stat, p_value, dof, expected = chi2_contingency(contingency_table)

     # Print the results
     print("Chi-Square Statistic:", chi2_stat)
     print("P-Value:", p_value)

     # Interpret the results
     alpha = 0.05
     if p_value < alpha:
         print("Reject the null hypothesis. There is enough evidence to suggest an association between gender and credit limit.")
     else:
         print("Fail to reject the null hypothesis. So Null Hypothesis is true. There is no association between gender and credit limit.")
```

```
Chi-Square Statistic: 0.0
P-Value: 1.0
Fail to reject the null hypothesis. So Null Hypothesis is true. There is no association between gender and credit limit.
```

Fig: Code snippet for Chi-Square test

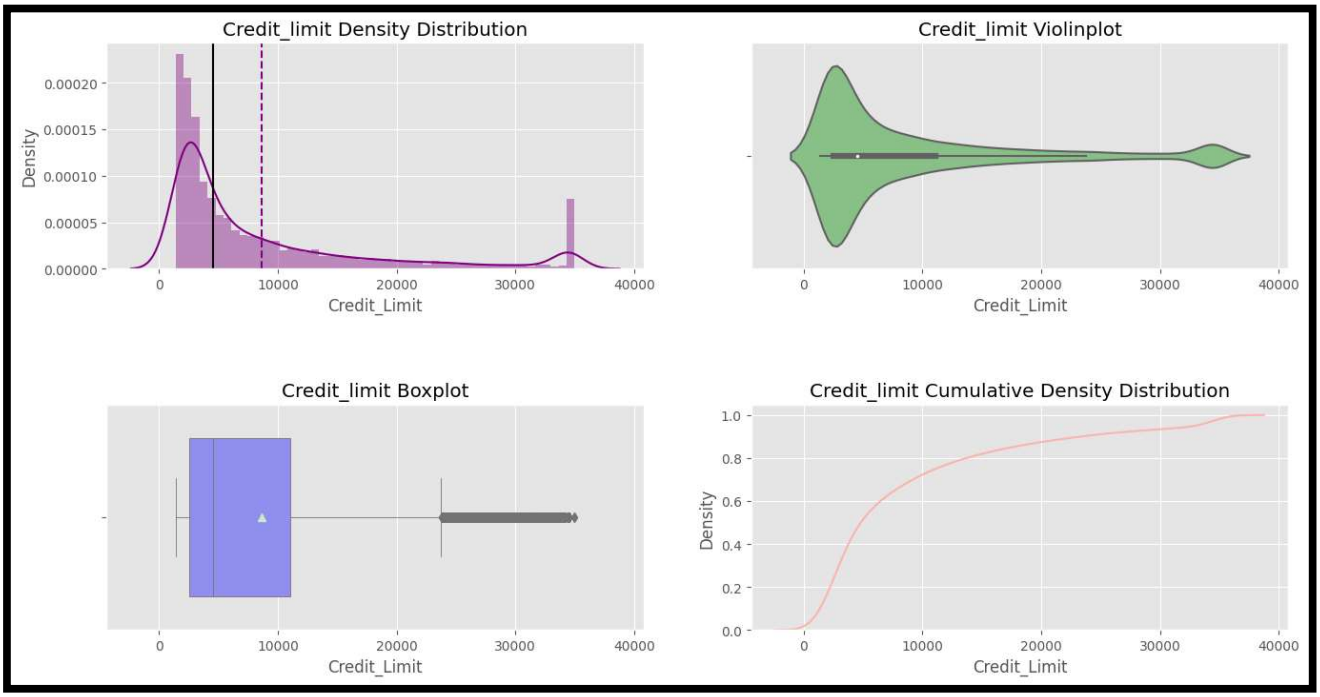

Fig: Data Visualization for hypothesis 1

Methodology of Hypothesis Test - 2:

Logistic Regression:

The Credit card dataset has column 'Attrition_flag' (with values Existing Customer and Attritional Customer), utilization ratio (with values from 0 to 0.99)

Objective:

To assess whether the utilization ratio has a significant effect on the likelihood of a customer being classified as attrited.

1. **Logistic Regression:**
   - Use logistic regression to model the relationship between the binary response variable ('Attrition_Flag') and the predictor variable ('utilization_less_than_0.5').
   - Include a constant term in the model using `sm.add_constant`.
   - Fit the logistic regression model using the `sm.Logit` class.

2. **Interpret Logistic Regression Results:**
   - Print the logistic regression summary using `result.summary()`.
   - Review coefficients, odds ratios, Wald statistics, and p-values.

3. **Conduct Hypothesis Test:**
   - Null Hypothesis (H0): The utilization ratio has no effect on the likelihood of being an attrition customer.
   - Alternative Hypothesis (H1): If a person has a utilization ratio less than 0.5, then the chances of being an attrition customer are higher.

4. **Get P-value:**
   - Extract the p-value associated with the 'utilization_less_than_0.5' coefficient from the logistic regression results.
   - P-value for utilization_less_than_0.5: 1.1463332660313418e-12

5. **Decision:**
   - **Reject the null hypothesis**.

There is a significant effect of utilization ratio on the likelihood of being an attrition customer.

Hypothesis 2

LOGISTIC REGRESSION

We have a column Attrition_flag(with values Existing Customer and Attrited Customer) , and utilization ratio(with values from 0 to 0.99)

- Null Hypothesis (H0): The utilization ratio has no effect on the likelihood of a person being an attritional customer.
- Alternative Hypothesis (H1): If a person has a utilization ratio less than 0.5, then the chances of that person being an attritional customer are higher.

```
import pandas as pd
import statsmodels.api as sm

# Encode 'Attrition_Flag' as 0 for 'Existing Customer' and 1 for 'Attrited Customer'
credit_data['Attrition_Flag'] = credit_data['Attrition_Flag'].map({'Existing Customer': 0, 'Attrited Customer': 1})

# Create a binary variable for utilization ratio less than 0.5
credit_data['utilization_less_than_0.5'] = (credit_data['Avg_Utilization_Ratio'] < 0.5).astype(int)

# Logistic regression
X = sm.add_constant(credit_data['utilization_less_than_0.5'])
y = credit_data['Attrition_Flag']

model = sm.Logit(y, X)
result = model.fit()

# Print the logistic regression summary
print(result.summary())

# Conduct a hypothesis test on the utilization_less_than_0.5 coefficient
print("\nHypothesis Test Results:")
print("Null hypothesis (H0): The utilization ratio has no effect on the likelihood of being an attrition customer.")
print("Alternative hypothesis (H1): If a person has a utilization ratio less than 0.5, then the chances of being an attrition customer are higher.")

# Get the p-value for the utilization_less_than_0.5 coefficient
p_value = result.pvalues['utilization_less_than_0.5']

print(f"\nP-value for utilization_less_than_0.5: {p_value}")

# Check if the p-value is less than the significance level (e.g., 0.05)
alpha = 0.05
if p_value < alpha:
    print("\nReject the null hypothesis. There is a significant effect of utilization ratio on the likelihood of being an attrition customer.")
else:
    print("\nFail to reject the null hypothesis. There is no significant effect of utilization ratio on the likelihood of being an attrition customer.")
```

```
Optimization terminated successfully.
         Current function value: 0.381469
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:         Attrition_Flag   No. Observations:                20071
Model:                          Logit   Df Residuals:                    20069
Method:                           MLE   Df Model:                            1
Date:                Thu, 07 Dec 2023   Pseudo R-squ.:               0.003594
Time:                        13:45:41   Log-Likelihood:                -7656.5
converged:                       True   LL-Null:                       -7684.1
Covariance Type:            nonrobust   LLR p-value:                 1.068e-13
==============================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         -2.2982      0.059    -39.067      0.000      -2.413      -2.183
utilization_less_than_0.5      0.4483      0.063      7.112      0.000       0.325       0.572
==============================================================================================

Hypothesis Test Results:
Null hypothesis (H0): The utilization ratio has no effect on the likelihood of being an attrition customer.
Alternative hypothesis (H1): If a person has a utilization ratio less than 0.5, then the chances of being an attrition customer are higher.

P-value for utilization_less_than_0.5: 1.1463332660313418e-12

Reject the null hypothesis. There is a significant effect of utilization ratio on the likelihood of being an attrition customer.
```

Fig: Code snippet for logistic regression in hypothesis-2

We compared our logistic regression model with another model (Chi-Square test).

The Chi-Square test also gave us the same results which we obtained in the logistic regression.

Below is the code snippet for its implementation.

CHI-SQAURE TEST

- proves same thing

```
[ ]  import pandas as pd
     from scipy.stats import chi2_contingency

     # We have a DataFrame named 'credit_data' with columns 'Attrition_flag' and 'utilization_ratio'
     # We make sure 'utilization_ratio' is converted into a categorical variable based on your thresholds

     # For example, we create a new column 'utilization_category' based on the threshold 0.5
     credit_data['utilization_category'] = pd.cut(credit_data['Avg_Utilization_Ratio'], bins=[-float('inf'), 0.5, float('inf')],
                                                  labels=['Less than 0.5', '0.5 and above'])

     # Create a contingency table
     contingency_table = pd.crosstab(credit_data['Attrition_Flag'], credit_data['utilization_category'])

     # Perform the chi-squared test
     chi2, p, dof, expected  = chi2_contingency(contingency_table)

     # Print the result
     print(f"Chi-squared value: {chi2}")
     print(f"P-value: {p}")

     # Check if the p-value is less than your chosen significance level (e.g., 0.05)
     if p < 0.05:
         print("Reject the null hypothesis: There is a significant relationship between utilization ratio and attrition.")
     else:
         print("Fail to reject the null hypothesis: There is no significant relationship between utilization ratio and attrition.")
```

```
Chi-squared value: 50.77752694816439
P-value: 1.0345145597359718e-12
Reject the null hypothesis: There is a significant relationship between utilization ratio and attrition.
```

Fig: Code snippet for Chi-Square test in hypothesis-2

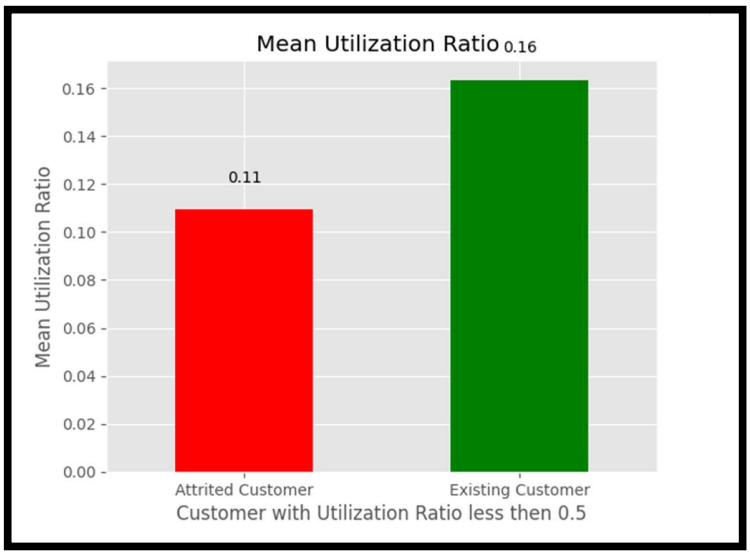

Fig: Data Visualization for hypothesis 2

Conclusion:

In summary, while the initial hypothesis suggests a higher representation of males among individuals with a credit limit exceeding $25,000, the available evidence does not strongly support a significant correlation between gender and credit limits. This implies that gender might not be a decisive factor in determining credit limits. Instead, the findings suggest that variables such as education level and income play more pivotal roles in shaping credit limits. This nuanced perspective emphasizes the intricate nature of credit assessment processes, warranting further exploration into the multifaceted factors influencing credit limit determinations.

The examination of the credit card customer information dataset has provided valuable insights into the interplay between the utilization ratio and customer attrition. The logistic regression model reveals a substantial impact of the utilization ratio on the probability of a customer being classified as attrited. Specifically, the coefficient associated with the binary variable 'utilization_less_than_0.5' demonstrates statistical significance, indicating that customers with a utilization ratio less than 0.5 are more prone to experiencing attrition. Furthermore, the hypothesis test on the utilization ratio coefficient yields compelling evidence to reject the null hypothesis, underscoring the critical role of the utilization ratio as a key factor influencing customer attrition.

Recommendations for Business Strategies:

1. Targeted Engagement Strategies:

Tailor marketing and engagement strategies to customers with utilization ratios below 0.5. Implement targeted communication campaigns highlighting benefits, rewards, or promotions to encourage increased card usage among this segment.

2. Personalized Offerings:

Leverage the understanding of customer behavior to create personalized offerings for individuals with low utilization ratios. This could include customized credit limit adjustments, promotional interest rates, or exclusive rewards to incentivize increased card utilization.

3. Proactive Customer Retention:

Implement proactive customer retention programs for those identified as having a utilization ratio below the critical threshold. This may involve timely communication, special loyalty programs, or personalized customer service to address any potential concerns or issues.

4. Utilization Education:

Launch educational initiatives to inform customers about the benefits of utilizing a higher proportion of their credit limits. Educate them on responsible credit card usage and showcase the advantages of maximizing available credit while maintaining financial discipline.

5. Continuous Monitoring and Adaptation:

Establish a system for continuous monitoring of customer utilization patterns. Regularly reassess the effectiveness of implemented strategies and adapt them based on evolving customer behaviors and market dynamics.

To sum up, this analysis's conclusions offer a solid basis for making strategic decisions regarding customer attrition. Financial institutions can lessen customer attrition and increase customer loyalty by proactively addressing the link between attrition and utilization ratios. When incorporated into the business plan, these suggestions could improve client relations and boost the general success of the provided credit card services.

Limitations and Future Work:

1. Representativeness and Quality of Data:

The quality of the dataset has a major impact on the results' representativeness and correctness. The results might not be entirely representative of the larger customer base if the data is biased or skewed.

2. Temporal Dynamics:

The analysis is predicated on a moment in time in which customer data was captured. Customer behavior shifts over time, impacted by outside variables or the state of the economy, are not fully recorded. A more dynamic understanding could be obtained from longitudinal data.

3. Correlation versus Causation:

Although the analysis shows a correlation between utilization ratios and attrition, more research is needed to determine the cause. The relationship could be muddled by additional unobserved factors.

Ideas for Further Study or Additional Data Gathering:

1. Analysis of Longitudinal Data:

Undertake a long-term investigation to monitor alterations in consumer conduct. This would offer a more thorough comprehension of how utilization ratios change over time and how they affect attrition in various contexts.

2. Multivariate Evaluation:

Extend the analysis to encompass a larger range of factors affecting customer attrition. A more comprehensive knowledge could be aided by variables like economic indicators, customer satisfaction ratings, and usage of additional banking products.

3. External Verification:

Verify the results by working with other financial institutions or by comparing them to external datasets. This would strengthen the basis for strategic decision-making and improve the results' generalizability.

4.Models for Machine Learning:

Use cutting-edge machine learning models to find interactions and nonlinear relationships between variables. This might make hidden patterns visible that are missed by conventional statistical models.

REFERENCES:

1. Exploring Customer Behavior: A Data Analysis of Credit Card Dataset (https://www.linkedin.com/pulse/exploring-customer-behavior-data-analysis-credit-card-gordon-kwok/)

2. BUREAU OF CONSUMER FINANCIAL PROTECTION | SEPTEMBER 2021 The Consumer Credit Card Market (https://files.consumerfinance.gov/f/documents/cfpb_consumer-credit-card-market-report_2021.pdf)

3. OpenAI. (2023). *ChatGPT* [Large language model]. https://chat.openai.com

4. Analysis and prediction for credit card fraud detection dataset using data mining approaches (https://www.researchgate.net/publication/362098155_Analysis_and_prediction_for_credit_card_fraud_detection_dataset_using_data_mining_approaches)

5. The effect of feature extraction and data sampling on credit card fraud detection (https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00684-w)