Analysis of Credit Card Customer Information Dataset

By:
Nayan Bhiwapurkar,
Kaumudi Patil,
Shreyas Hingmire,
Sharan Patel

Introduction

- **Significance of Credit Cards:**
  - Fundamental in today's financial landscape.
  - Key determinant of consumer behavior and financial decision-making.

- **Importance for Financial Institutions:**
  - Essential to study credit card user behavior for strategic decision-making, profitability, and customer satisfaction.

- **Research Objective:**
  - Derive insights from a comprehensive credit card customer information dataset.
  - Financial institutions can optimize services and offerings with the help off objective.

Data Collection and Characteristics

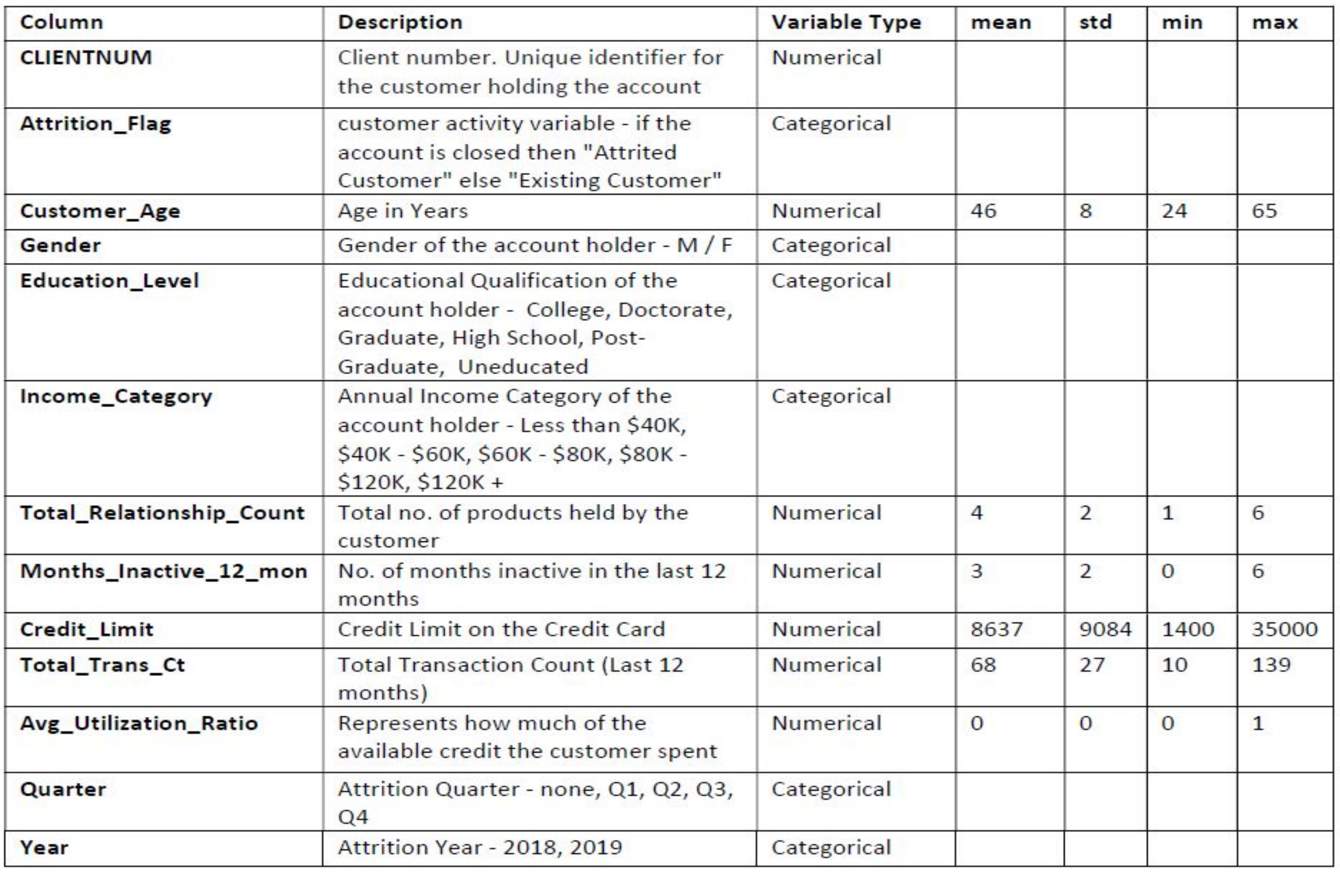
- Spans 2018 and 2019, providing a comprehensive snapshot.
- **Customer Information:** Encompasses demographics (client numbers, ages, genders) and financials (incomes, credit limits, card categories).
- **Focus Variables for Analysis:**
  - **Credit_Limit:** Reflects credit extended to customers.
  - **Attrition_Flag:** Crucial for identifying customer attrition.
  - **Avg_Utilization_Ratio:** Average card utilization, an essential metric for analysis.
- **Behavior Metrics:**
  - Transaction counts, months of inactivity, and contact frequency captured.
  - Facilitates a holistic view of customer interactions.

Problems in Dataset

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

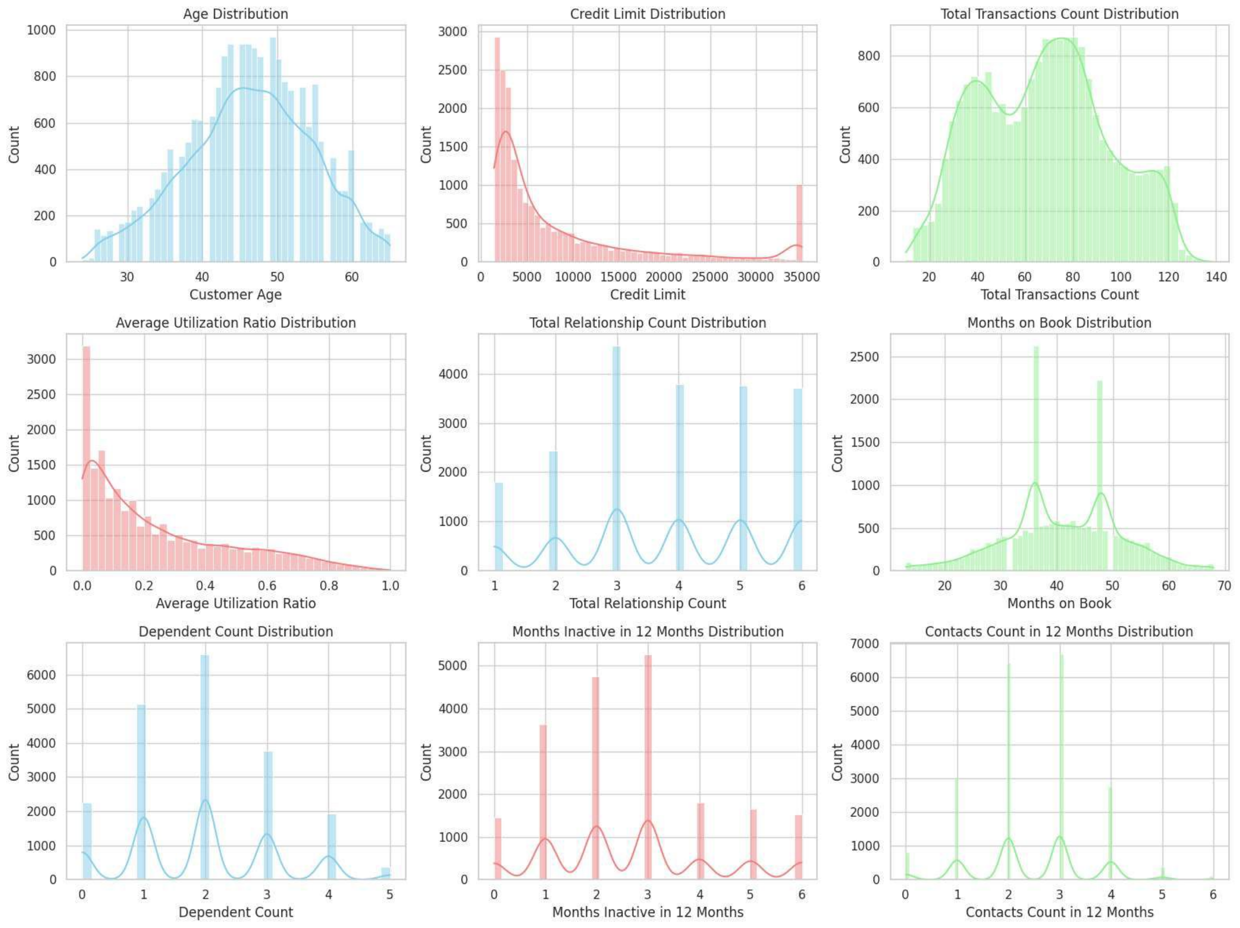
- **Feature Selection:**
  - Analyzed correlations and distributions of features.
  - We removed some features which did not contribute much to the hypothesis testing.

EDA and Attribute understanding



Column	Description	Variable Type	mean	std	min	max
CLIENTNUM	Client number. Unique identifier for the customer holding the account	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				
Education_Level	Educational Qualification of the account holder - College, Doctorate, Graduate, High School, Post-Graduate, Uneducated	Categorical				
Income_Category	Annual Income Category of the account holder - Less than $40K, $40K - $60K, $60K - $80K, $80K - $120K, $120K +	Categorical				
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
Credit_Limit	Credit Limit on the Credit Card	Numerical	8637	9084	1400	35000
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				

Data Visualization



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

Importance of Solution

Strategic Decision Making:

1. Serves as a crucial guide for strategic decision-making in the financial sector
2. Goes beyond surface-level statistics, providing a nuanced understanding of consumer behavior in credit card services.
3. Enables stakeholders to proactively address challenges and capitalize on emerging opportunities.
4. Fosters a proactive approach, ensuring the sustained relevance and competitiveness of credit card services in a dynamic market.

Importance of Solution

Risk Mitigation and Customer Satisfaction

1. Plays a pivotal role in identifying early warning signs of attrition.
2. Enables financial institutions to implement targeted retention strategies for minimizing financial impact.
3. Tailors services based on a deep understanding of customer preferences and behaviors.
4. Fosters long-term loyalty by meeting evolving customer expectations.

Hypothesis Test - 1

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

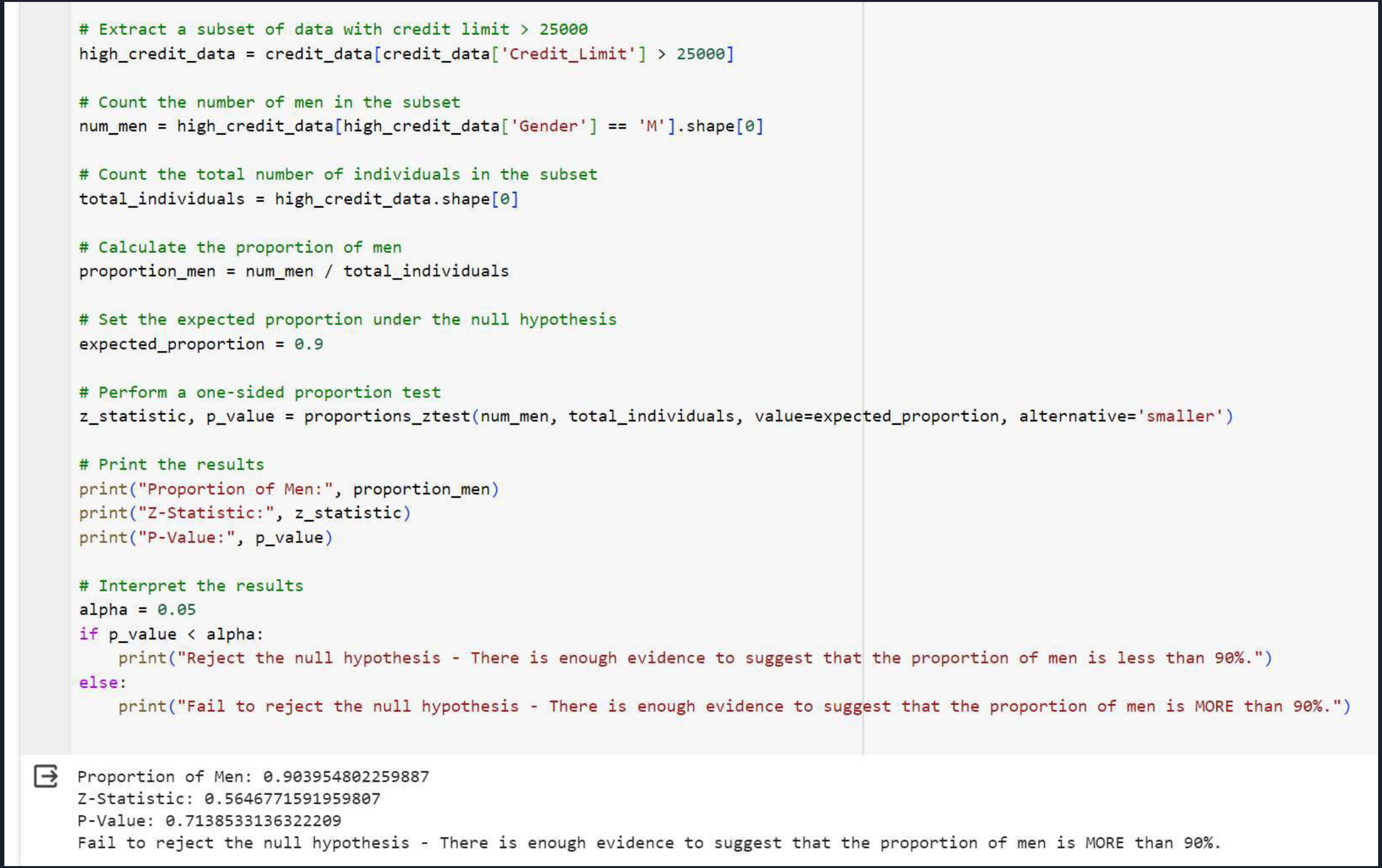
- Null Hypothesis (H0): $H0: p \geq 0.90$

  The percentage of males within the population having a credit limit exceeding $25,000 is **greater** than 90%

- Alternative Hypothesis (H1): $H1: p < 0.90$

  The percentage of males within the population having a credit limit exceeding $25,000 is **less** than 90%.

Z - Test for our hypothesis testing.



```
# 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%.
```

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%.**

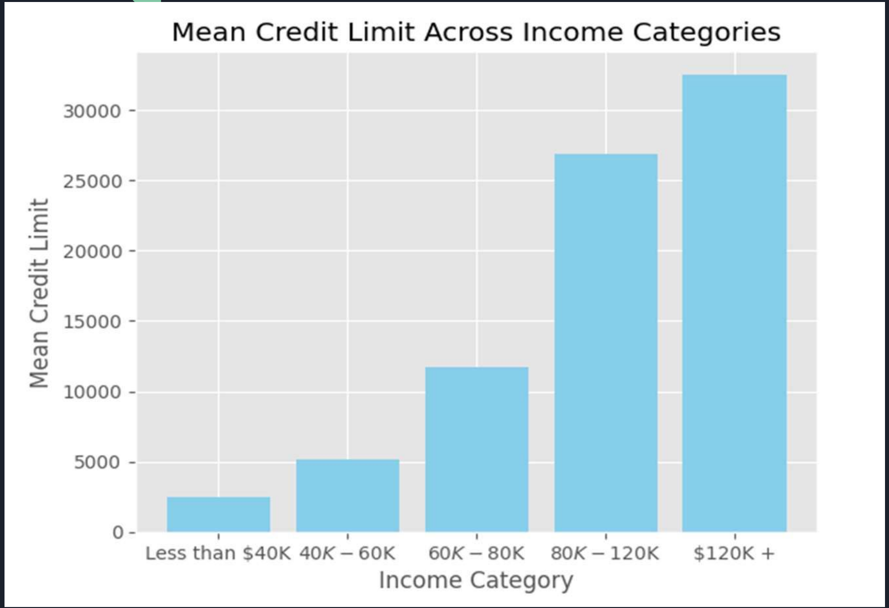
Deeper Analysis

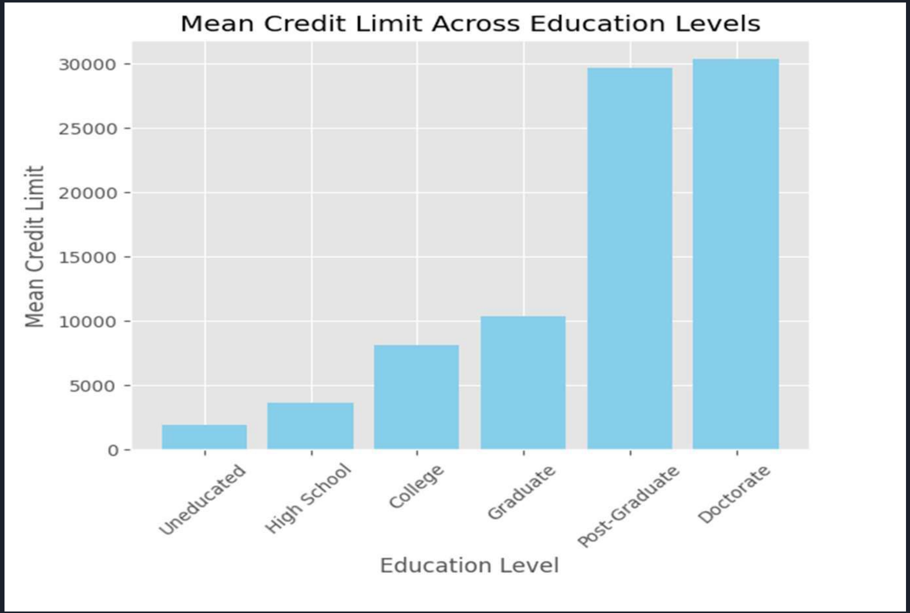
❖ While exploring the factors influencing credit limits, we investigated various variables. We wanted to find if ‘Gender’ really affects credit limit or not, and is ‘Gender’ the only factor impacting credit limits.

❖ 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.

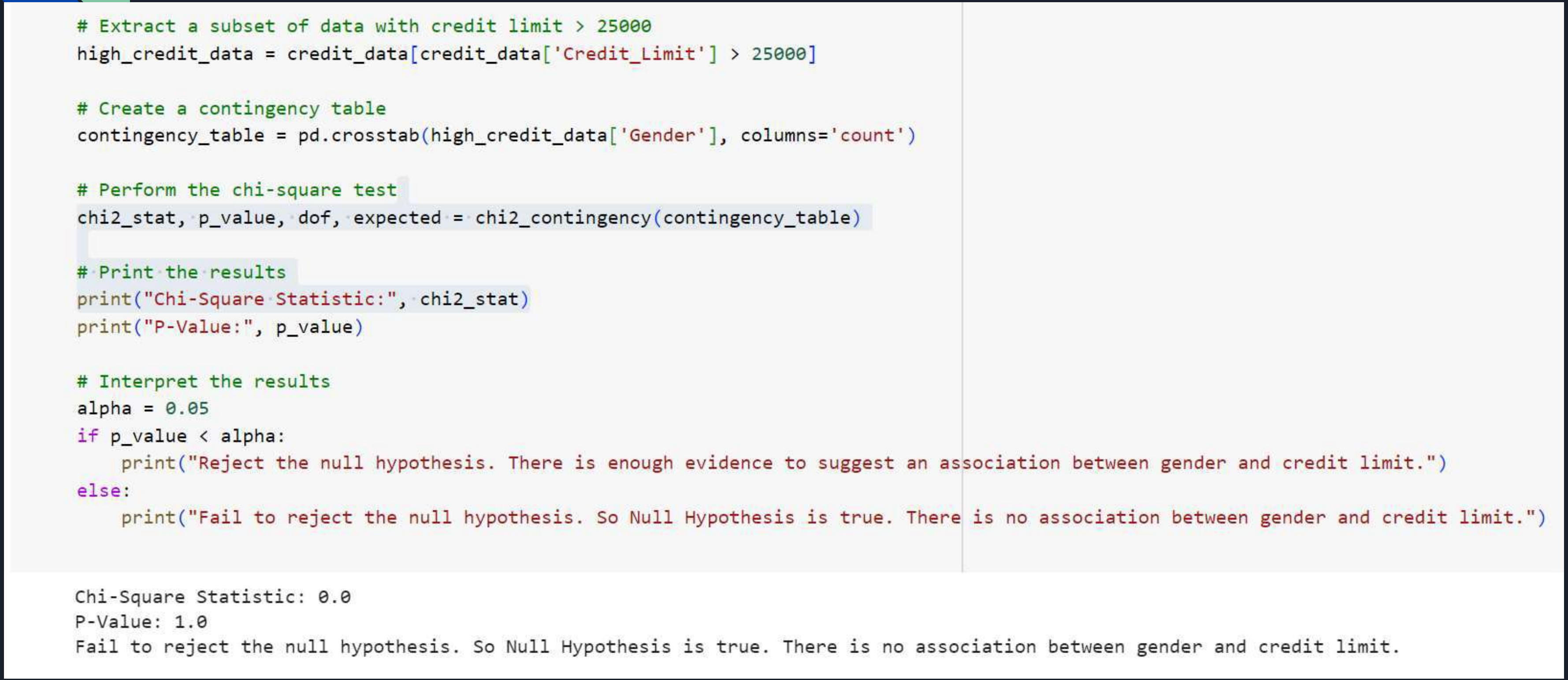

Mean Credit Limit Across Income Categories
Mean Credit Limit
0
5000
10000
15000
20000
25000
30000
Less than $40K
40K − 60K
60K − 80K
80K − 120K
$120K +
Income Category


Mean Credit Limit Across Education Levels
Mean Credit Limit
0
5000
10000
15000
20000
25000
30000
Uneducated
High School
College
Graduate
Post-Graduate
Doctorate
Education Level

Chi-square test:

- ❖ 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.
- ❖ Objective:
  - ➢ To test whether there is an association between gender and credit limit among individuals with a credit limit greater than 25,000.
- ❖ Null Hypothesis (H0):
  - ➢ H0: There is no association between gender and credit limit.
- ❖ Alternative Hypothesis (H1):
  - ➢ H1: There is an association between gender and credit limit.

Code Snippet for Chi-Square test in hypothesis-1



```python
# 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.
```

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.

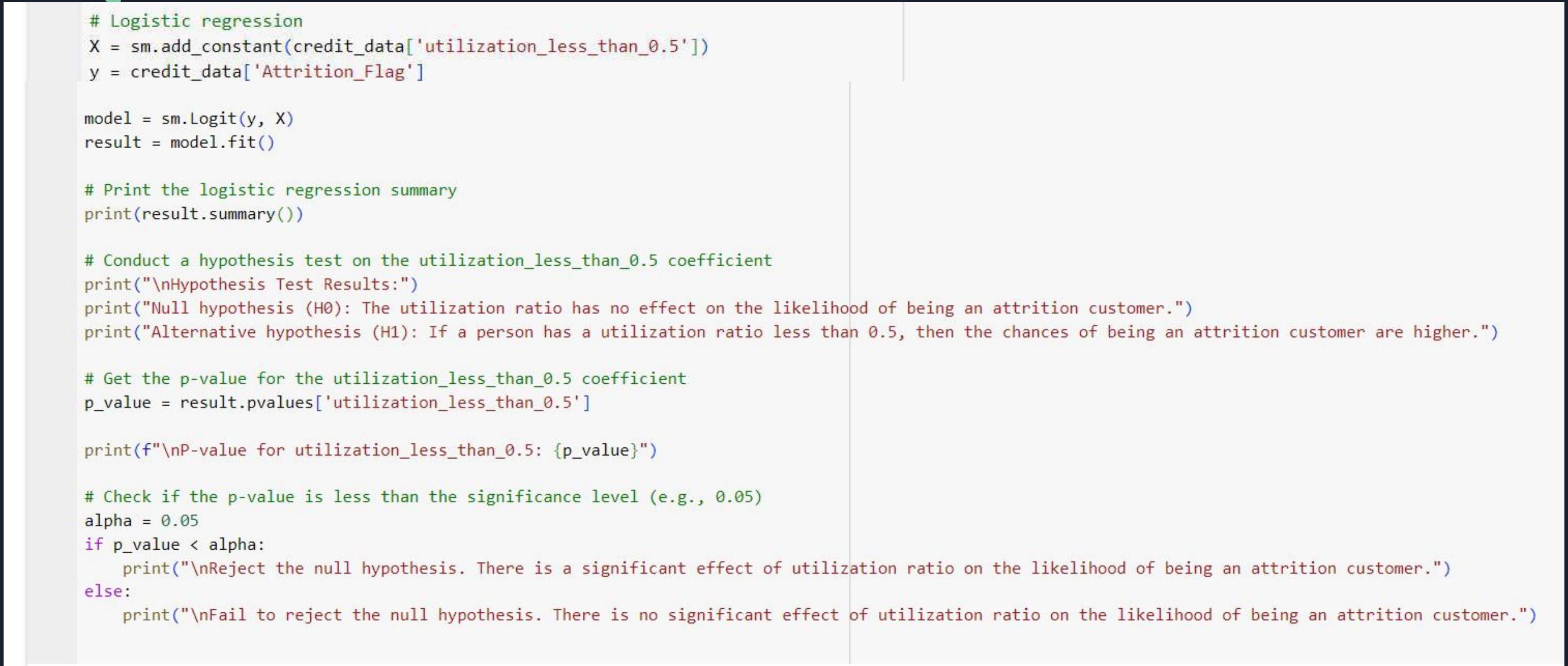
Hypothesis Test - 2

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

- **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.

Logistic Regression to test our hypothesis-2

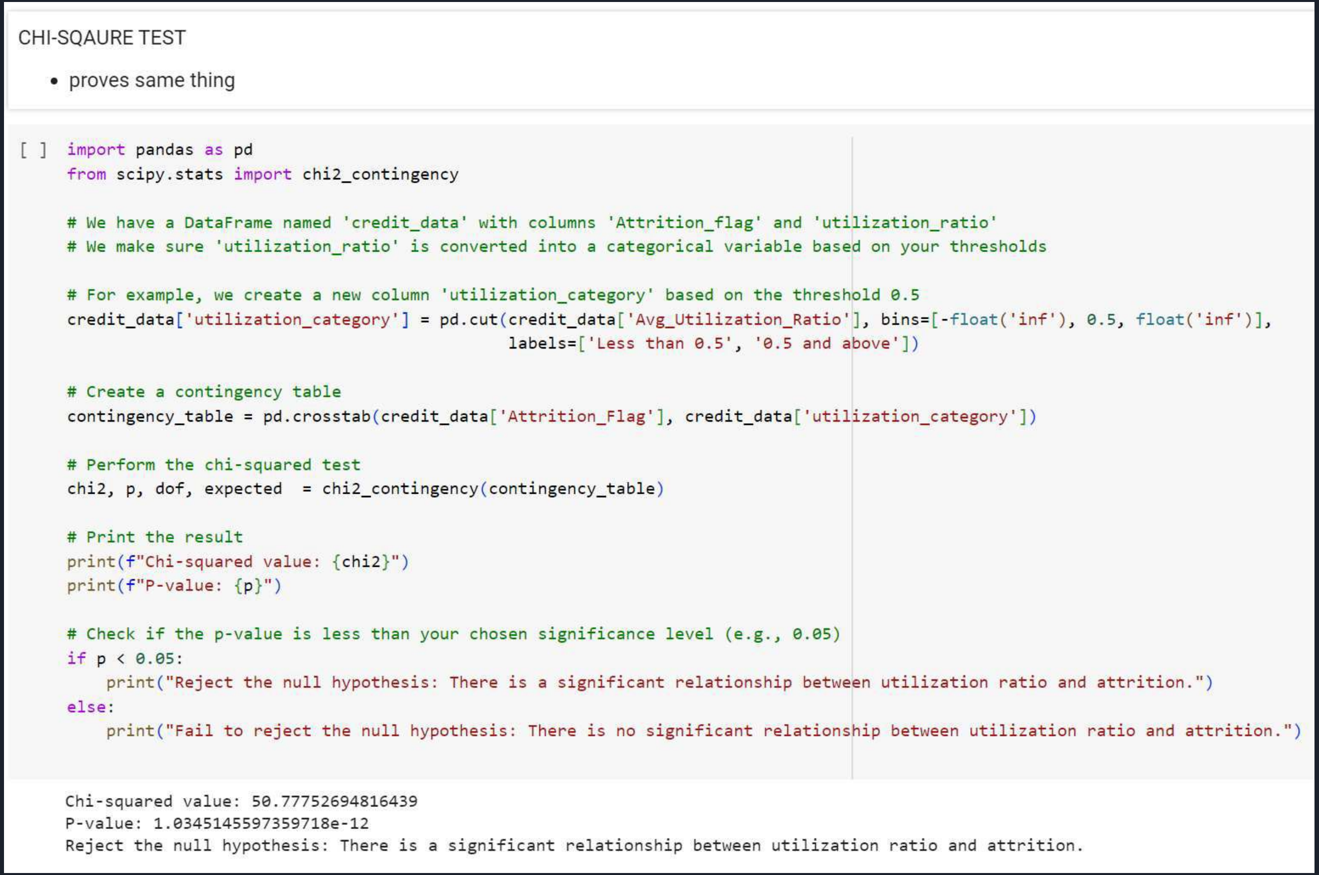


```
# Logistic regression
X = sm.add_constant(credit_data['utilization_less_than_0.5'])
y = credit_data['Attrition_Flag']

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.")
```

Decision

- 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.

We used **CHI-square test** as well to test our hypothesis and it also gave the **same** result.



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.
```

Conclusion

- **Gender and Credit Limits:**
  - Initial hypothesis of higher male representation in credit limits > $25,000.
  - Evidence suggests no strong correlation; gender might not be a decisive factor.
  - Emphasis on variables like education level and income influencing credit limits.

- **Utilization Ratio and Customer Attrition:**
  - Logistic regression model highlights substantial impact of utilization ratio on attrition probability.
  - Coefficient for 'utilization_less_than_0.5' is statistically significant, indicating higher attrition for ratios < 0.5.

Recommendations for Business Strategies:

1. Targeted Engagement Strategies
2. Personalized Offerings
3. Proactive Customer Retention
4. Utilization Education
5. Continuous Monitoring and Adaptation

Limitations and Future Work:

1. **Representativeness and Quality of Data:**

   The results might not be entirely representative of the larger customer base.

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.

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 / Additional Data Gathering:

1. Analysis of Longitudinal Data

2. Multivariate Evaluation

3. External Verification

4. Models for Machine Learning

Thank you!