CANDIDATE NUMBER (C-NUMBER)	C2096215
MODULE NAME	Data Analysis for Business
WORD COUNT	2953
SUBMISSION DATE	03/06/2024

DECLARATION

I certify that this assessment submission is entirely my work and I have fully referenced and correctly cited the work of others, where required. I also confirm the contents of my submission have not been generated by a third party, or through an Artificial Intelligence generative system*.

I have read the Student Discipline Regulations (Student Discipline Regulations) and understand any Assessment Related Offence/ Academic Misconduct may result penalties being applied.

By submitting this assessment submission, I am confirming that I am fit to sit according to the Assessment Regulations.

I declare that:

- This is my own unaided work.
  Yes ☒
  No ☐

- The word count stated by me is correct.
  Yes ☒
  No ☐

- I'm happy for my work to be retained on the Elite repository and made available to staff and future students**
  Yes ☒
  No ☐

*Please note that all the assignments are submitted to Turnitin.
**Please note personal information (such as names) will be deleted.

Table of Contents

Table Of Figures

Introduction

For credit card companies, maintaining current customer relationships is crucial in the very competitive financial world of today. A significant loss in possible revenue and future profitability is caused by customer churn, which is the act of a customer stopping their service (Akkio, 2023). Credit card issuers are using credit card churn prediction models more often to address this issue. By identifying clients who are at a high risk of leaving, these models use statistical analysis and machine learning approaches to enable proactive actions and focused loyalty strategies (Verbraken et al., 2018).

The Importance of Customer Retention in the Credit Card Industry

Credit card firms must beware of credit card churn for a number of important reasons.

Table 1

No	Benefit	Description
1	Reduced Customer Acquisition Costs	Acquiring new customers is significantly more expensive than retaining existing ones (Ref: Frederick Reichheld & Thomas Teal, 2000). Churn prediction helps allocate resources effectively to maximize return on investment.
2	Enhanced Customer Relationships	By understanding churn signals, companies can address customer concerns, personalize services, and build stronger customer loyalty (Ref: Machine Learning to Develop Credit Card Customer Churn Prediction, 2023).
3	Targeted Marketing Campaigns	Knowing the reasons for customer churn allows businesses to tailor promotions and benefits that resonate with at-risk customers, increasing customer retention rates.

Several stakeholders benefit from addressing credit card churn for example:

- **Credit Card Companies:** Reduced churn translates to increased revenue, improved customer lifetime value, and lower acquisition costs.
- **Customers:** Proactive retention efforts aimed at addressing their concerns can lead to improved satisfaction and potentially better rewards or service offerings.
- **Financial Industry as a Whole:** Reduced churn promotes a more competitive and efficient financial ecosystem where companies focus on providing better value to customers.

Factors to Analyze for Credit Card Churn Prediction

Credit card churn prediction tools that work well look at a lot of things, like their creditworthiness, customer behaviour, and demographics. Here are some of the most important things that are usually looked at:

- **Transaction Behavior:** The number of times and amounts of credit cards used, how much people spend in different groups, and how these numbers change over time (Customer Churn Prediction on Credit Card Services using Random Forest Method, 2016).
- **Credit History:** Credit score, credit utilization ratio, and history of late payments (Akkio, 2023).
- **Demographics:** Age, income level, geographic location, and marital status can all influence churn propensity.
- **Card Features:** Benefits offered by the credit card, such as rewards programs, cashback offers, and annual fees (Predicting Credit Card Customer Churn, 2023).

Credit card companies can learn a lot about how customers behave and find the most important factors that predict customer loss by looking at these factors and how they affect each other.

Data Acquisition and Preprocessing

Data acquisition is the process of collecting data from various sources for analysis purposes. It involves several key steps (Hughes, 2010):

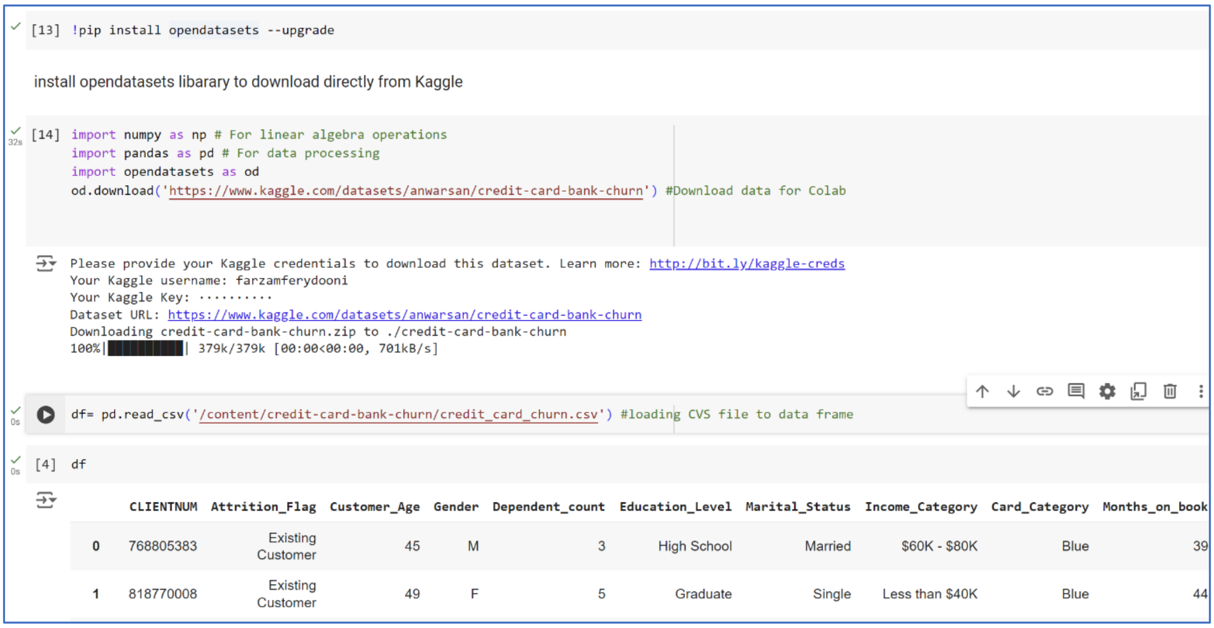
1. Identifying reliable data sources like databases, APIs, web scraping, IoT devices, etc.
2. Handling different data formats (structured, semi-structured, unstructured).
3. Extracting data using techniques like SQL queries, APIs, web scraping tools.
4. Integrating data from multiple sources into a comprehensive dataset.
5. Cleaning and transforming data to ensure consistency, accuracy, and usability.
6. Storing and managing the acquired data effectively for analysis.

Kaggle is an online community platform for datasets, tournaments, and machine learning and data science. Practicing techniques, examining datasets, and taking part in competitive data science projects have all been easier with Kaggle (Moltzau, 2019).

The dataset obtained from Kaggle (Credit Card Churn Prediction, no date) regrads to a business manager at a bank that offers consumer credit cards is struggling with the problem of client attrition.

Data Collection

The dataset is a structured dataset, CSV format. It contains various attributes or features that describe the customer's profile and their behaviour(available at (Ferydooni, 2024) developed a credit card churn prediction model using Python and shared the code on GitHub.).



```
[13] !pip install opendatasets --upgrade
```

install opendatasets libarary to download directly from Kaggle

```
[14] import numpy as np # For linear algebra operations
     import pandas as pd # For data processing
     import opendatasets as od
     od.download('https://www.kaggle.com/datasets/anwarsan/credit-card-bank-churn') #Download data for Colab
```

```
Please provide your Kaggle credentials to download this dataset. Learn more: http://bit.ly/kaggle-creds
Your Kaggle username: farzamferydooni
Your Kaggle Key: ··········
Dataset URL: https://www.kaggle.com/datasets/anwarsan/credit-card-bank-churn
Downloading credit-card-bank-churn.zip to ./credit-card-bank-churn
100%|██████████| 379k/379k [00:00<00:00, 701kB/s]
```

```
df= pd.read_csv('/content/credit-card-bank-churn/credit_card_churn.csv') #loading CVS file to data frame
```

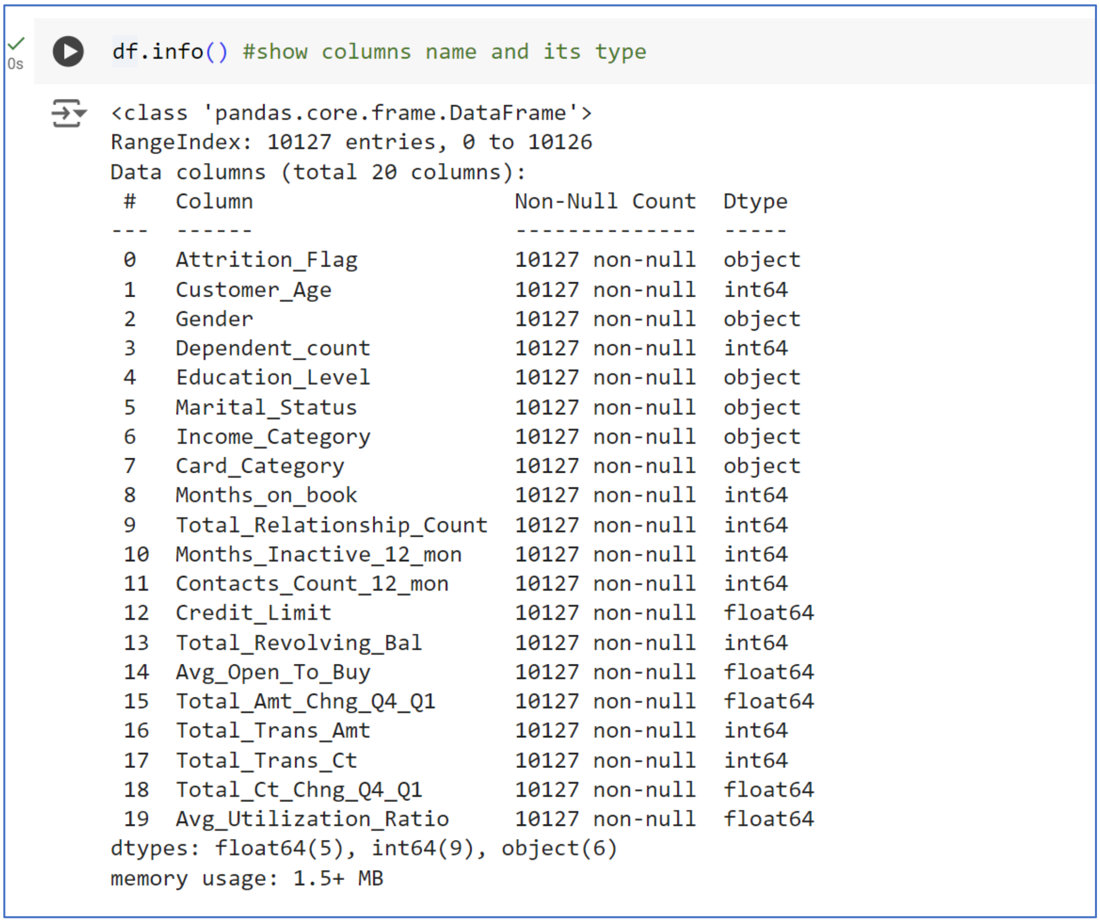
```
[4] df
```

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book
0	768805383	Existing Customer	45	M	3	High School	Married	$60K - $80K	Blue	39
1	818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	44

Figure 1:Importing Data

Regard for Figure 1 Dataset imported directly from Kaggle by "opendatasets" library. Data frame has **10127 rows × 23 columns**.

In the next step, the information is analyzed in terms of types:



```
df.info() #show columns name and its type
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10127 entries, 0 to 10126
Data columns (total 20 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Attrition_Flag            10127 non-null  object 
 1   Customer_Age              10127 non-null  int64  
 2   Gender                    10127 non-null  object 
 3   Dependent_count           10127 non-null  int64  
 4   Education_Level           10127 non-null  object 
 5   Marital_Status            10127 non-null  object 
 6   Income_Category           10127 non-null  object 
 7   Card_Category             10127 non-null  object 
 8   Months_on_book            10127 non-null  int64  
 9   Total_Relationship_Count  10127 non-null  int64  
 10  Months_Inactive_12_mon    10127 non-null  int64  
 11  Contacts_Count_12_mon     10127 non-null  int64  
 12  Credit_Limit              10127 non-null  float64
 13  Total_Revolving_Bal       10127 non-null  int64  
 14  Avg_Open_To_Buy           10127 non-null  float64
 15  Total_Amt_Chng_Q4_Q1      10127 non-null  float64
 16  Total_Trans_Amt           10127 non-null  int64  
 17  Total_Trans_Ct            10127 non-null  int64  
 18  Total_Ct_Chng_Q4_Q1       10127 non-null  float64
 19  Avg_Utilization_Ratio     10127 non-null  float64
dtypes: float64(5), int64(9), object(6)
memory usage: 1.5+ MB
```

Figure 2:Columns List

Table 2

Number of numerical features	**14 (Discrete feature 4 and Continuous feature 10)**
Number of categorical features	**6**

Data Cleaning and Outlier Treatment

These steps make sure that the data is correct, consistent, and free of mistakes and outliers that could have a big effect on how well machine learning models work and how reliable analysis insights are.

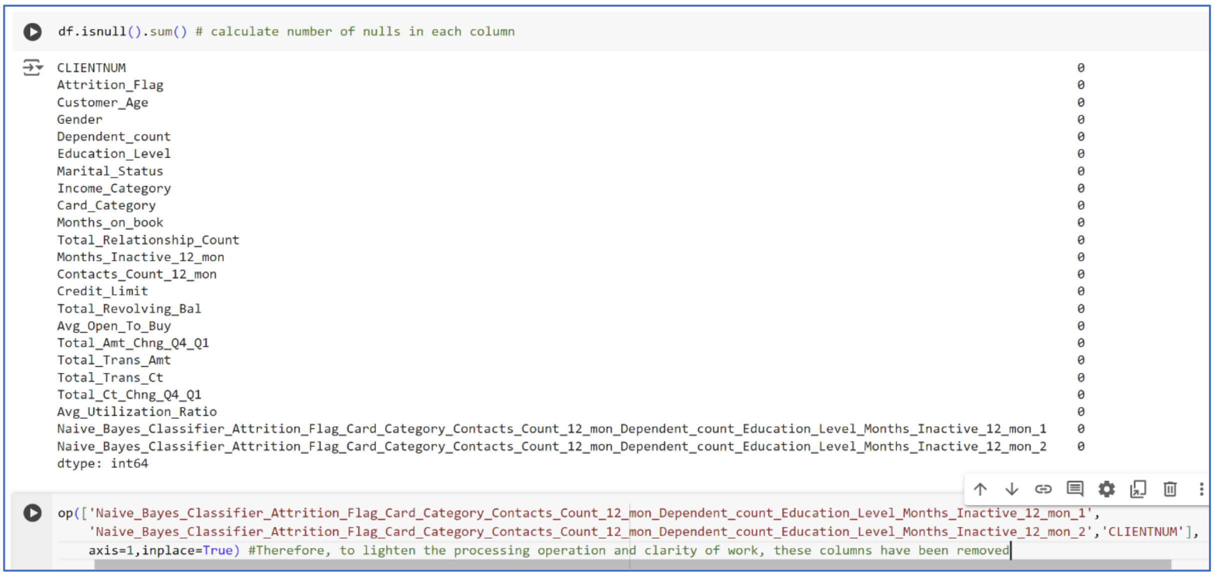
- **Data Cleaning:** The first step involves cleaning the data to remove inconsistencies and errors

1. **Handling Missing Values:** Identifying and dealing with missing data points by imputing missing values or removing rows/columns with missing data.

2. **Remove rows/columns:** Based on Figure 3 there is no null data in the data set which need to be replaced and Therefore, to clear the processing operation and clarity of work, these columns have been removed

Table 3: Data Cleaning

no	column	reason
1	Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1'	in this analysis we will find alternative classifier
2	Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2'	in this analysis we will find alternative classifier
3	CLIENTNUM'	customer ID not needed for this data analysis



```
df.isnull().sum() # calculate number of nulls in each column
```

```
CLIENTNUM                                                                                                   0
Attrition_Flag                                                                                              0
Customer_Age                                                                                                0
Gender                                                                                                      0
Dependent_count                                                                                             0
Education_Level                                                                                             0
Marital_Status                                                                                              0
Income_Category                                                                                             0
Card_Category                                                                                               0
Months_on_book                                                                                              0
Total_Relationship_Count                                                                                    0
Months_Inactive_12_mon                                                                                      0
Contacts_Count_12_mon                                                                                       0
Credit_Limit                                                                                                0
Total_Revolving_Bal                                                                                         0
Avg_Open_To_Buy                                                                                             0
Total_Amt_Chng_Q4_Q1                                                                                        0
Total_Trans_Amt                                                                                             0
Total_Trans_Ct                                                                                              0
Total_Ct_Chng_Q4_Q1                                                                                         0
Avg_Utilization_Ratio                                                                                       0
Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1  0
Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2  0
dtype: int64
```

```
op(['Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1',
    'Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2','CLIENTNUM'],
    axis=1,inplace=True) #Therefore, to lighten the processing operation and clarity of work, these columns have been removed
```

Figure 3: Null Counting and data cleaning

3. Outlier detection:

- Visualization methods BoxPlot Analysis: Boxplots provide concise overview of detect outliers. The drawn plot (Figure 4)based on numerical, show a visual summary which can be learned in the following way:

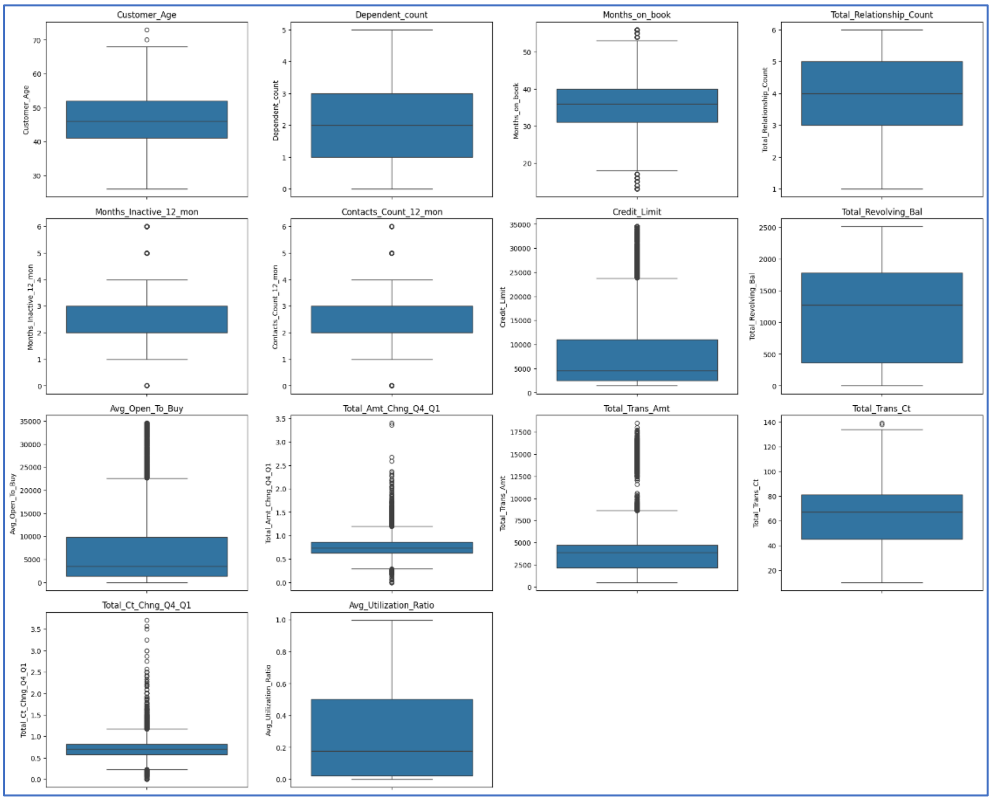


Figure 4: Box Plot for Outlier Detection

Table 4: Box Plot Analysis

No	Column Name	Comment
1	**Customer Age:**	The distribution is fairly symmetrical around the median, suggesting a consistent age range among customers, with a few outliers indicating some customers are significantly younger or older than the average.
2	**Dependent Count:**	Most customers have a lower number of dependents, with outliers suggesting that a few customers have a higher dependent count.
3	**Months on Book:**	Indicates a tight interquartile range, meaning most customers have been with the bank for a similar duration, with few outliers.
4	**Total Relationship Count:**	Customers tend to have multiple relationships with the bank, which could include various accounts and services.
5	**Months Inactive 12 mon:**	Shows that a majority of customers were inactive for less than three months in the past year.
6	**Contacts Count 12 mon:**	Most customers had few contacts with the bank, with some exceptions indicating more frequent interactions.
7	**Credit Limit:**	There's a wide range of credit limits, with several outliers showing that some customers have significantly higher limits.
8	**Total Revolving Balances:**	Most customers have low revolving balances, but there are notable exceptions with higher balances.
9	**Avg Open To Buy:**	Similar to 'Credit Limit', this shows varied amounts available for customers to spend, with many high-value outliers.
10	**Total Amt Chng Q4 Q1**	There are many outliers, indicating significant changes in the amounts for some customers between the fourth and first quarters.
11	**Total Trans Amt:**	A large number of outliers suggest some customers have very high transaction amounts compared to others.
12	**Total Trans Ct:**	While most transaction counts are grouped together, there is a range, indicating varying customer activity levels.
13	**Total Ct Chng Q4 Q1:**	Many data points fall outside the interquartile range, showing variability in customer transaction counts between quarters.
14	**Avg Utilization Ratio:**	Most customers have lower utilization ratios, with the data clustered towards the bottom of the plot.

- Statistical methods: Techniques like calculating standard deviations (z-scores) or Interquartile Range (IQR) can identify data points that deviate significantly from the mean or median.

4. Outliers replacement:

There are a few options for handling outliers:

1. Remove Outliers: This involves removing rows that contain outliers.
2. Replace with Mean: Replace outlier values with the mean of the column.
3. Replace with Median: Replace outlier values with the median of the column.
4. Replace with a Specific Value: Replace outlier values with a specific value we provide**.**

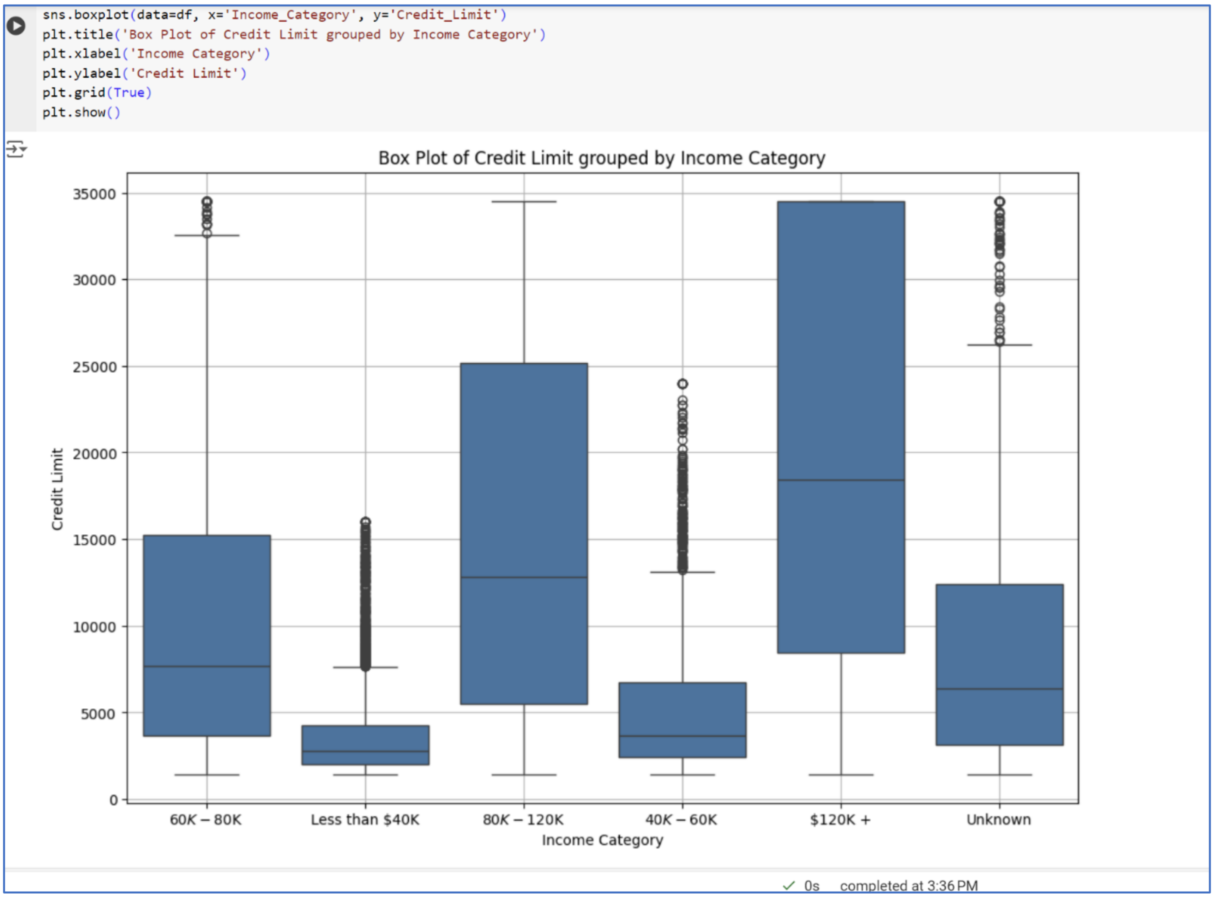
Data Analysis and Interpretation

Data Analysis based on Boxplot

Boxplots could provide a clear, concise overview of data distribution and facilitate meaningful comparisons across various domains (Jio, 2024). that typically need one categorical variable and one numerical variable. The categorical variable will be used to group the data, and the numerical variable will be used to display the distribution within each group. In the following, we will analyze again some boxplot that I think are important in data analysis:

- **Credit_Limit grouped by Income_Category:**

This plot help understand how credit limits change for people with different amounts of cash. This information can help to figure out if higher income groups are linked to bigger credit limits.



```
sns.boxplot(data=df, x='Income_Category', y='Credit_Limit')
plt.title('Box Plot of Credit Limit grouped by Income Category')
plt.xlabel('Income Category')
plt.ylabel('Credit Limit')
plt.grid(True)
plt.show()
```

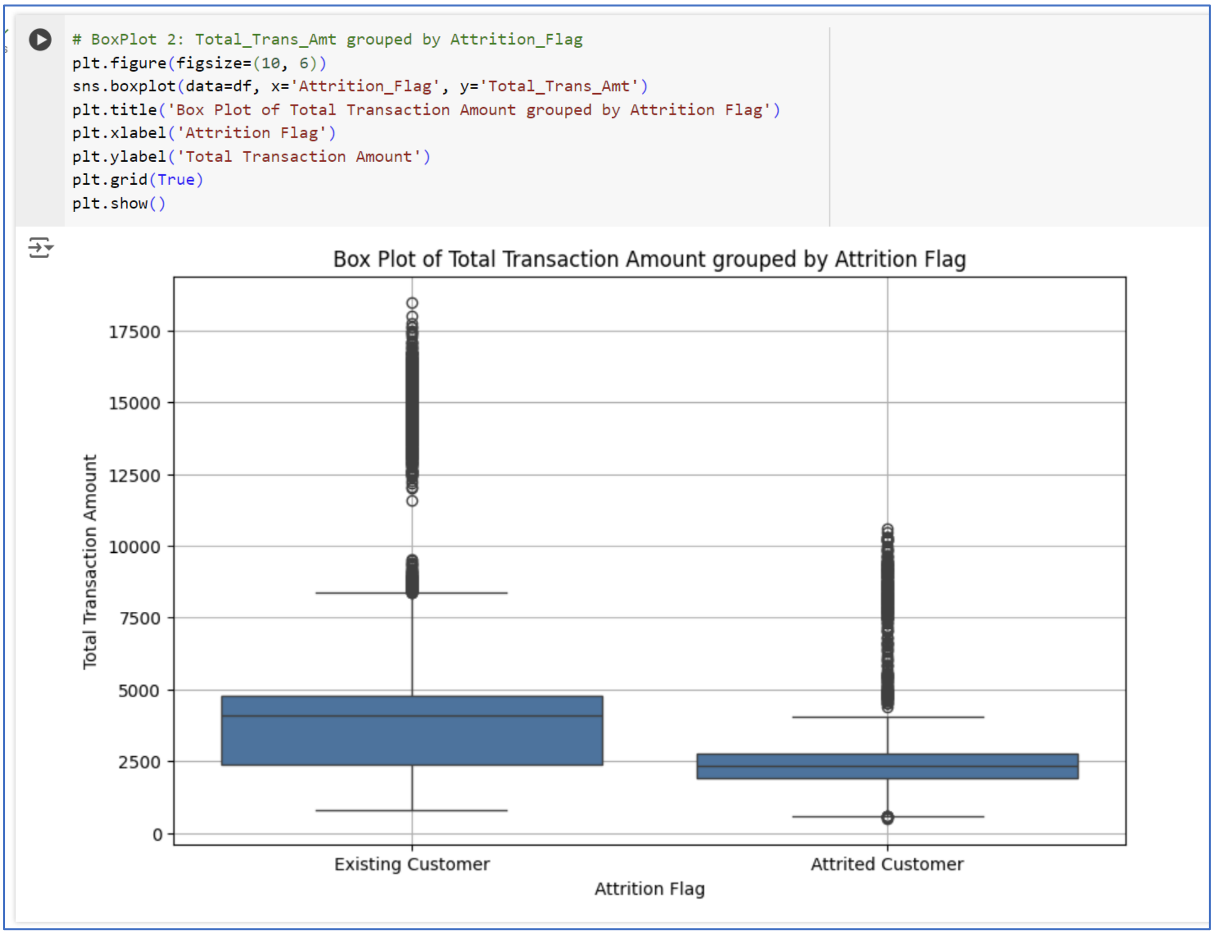
Figure 5:Credit_Limit grouped by Income_Category

Interpretation:

- The plot shows the distribution of credit limits across different income categories.
- Higher income categories tend to have higher median credit limits.
- There is a noticeable spread in credit limits within each income category, indicating variability in credit limits assigned to customers within the same income bracket.

- **Total_Trans_Amt grouped by Attrition_Flag:**

This will show the distribution of total transaction amounts for existing and attrited customers. It can help identify if there is a significant difference in transaction behavior between these two groups.



```
# BoxPlot 2: Total_Trans_Amt grouped by Attrition_Flag
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='Attrition_Flag', y='Total_Trans_Amt')
plt.title('Box Plot of Total Transaction Amount grouped by Attrition Flag')
plt.xlabel('Attrition Flag')
plt.ylabel('Total Transaction Amount')
plt.grid(True)
plt.show()
```

Figure 6:Total_Trans_Amt grouped by Attrition_Flag

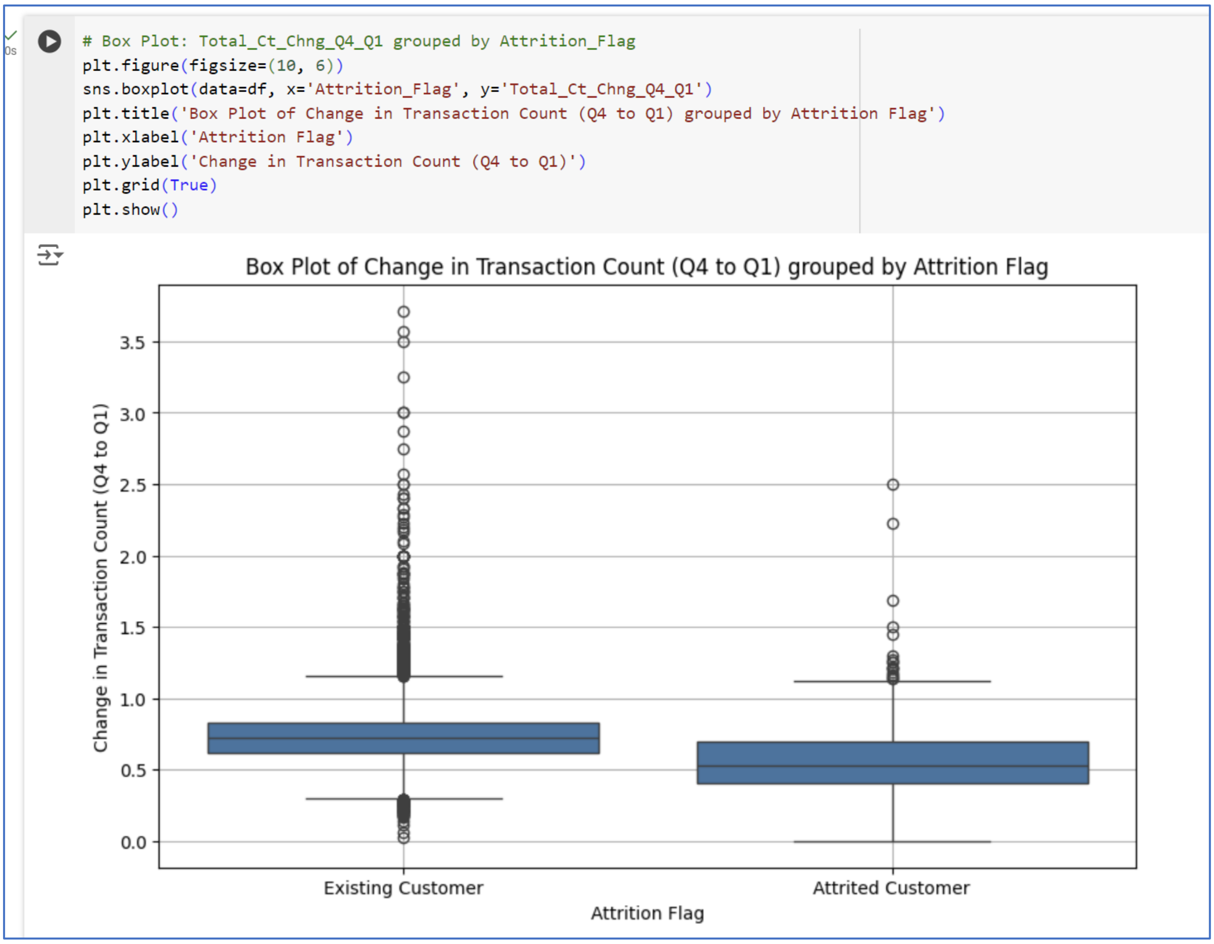
Interpretation:

- The plot shows the distribution of total transaction amounts for existing and attrited customers.
- Existing customers tend to have higher median transaction amounts compared to attrited customers.

There is a noticeable spread in transaction amounts for both groups, but the spread is wider for existing customers, indicating more variability in their transaction behavior.

Total_Ct_Chng_Q4_Q1 grouped by Attrition_Flag:

This will show the change in transaction count from Q4 to Q1 for existing and attrited customers. It can help identify if changes in transaction behavior are associated with customer attrition.



```
# Box Plot: Total_Ct_Chng_Q4_Q1 grouped by Attrition_Flag
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='Attrition_Flag', y='Total_Ct_Chng_Q4_Q1')
plt.title('Box Plot of Change in Transaction Count (Q4 to Q1) grouped by Attrition Flag')
plt.xlabel('Attrition Flag')
plt.ylabel('Change in Transaction Count (Q4 to Q1)')
plt.grid(True)
plt.show()
```

0s
Box Plot of Change in Transaction Count (Q4 to Q1) grouped by Attrition Flag
Change in Transaction Count (Q4 to Q1)
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Existing Customer
Attrited Customer
Attrition Flag

Figure 7:Total_Ct_Chng_Q4_Q1 grouped by Attrition_Flag

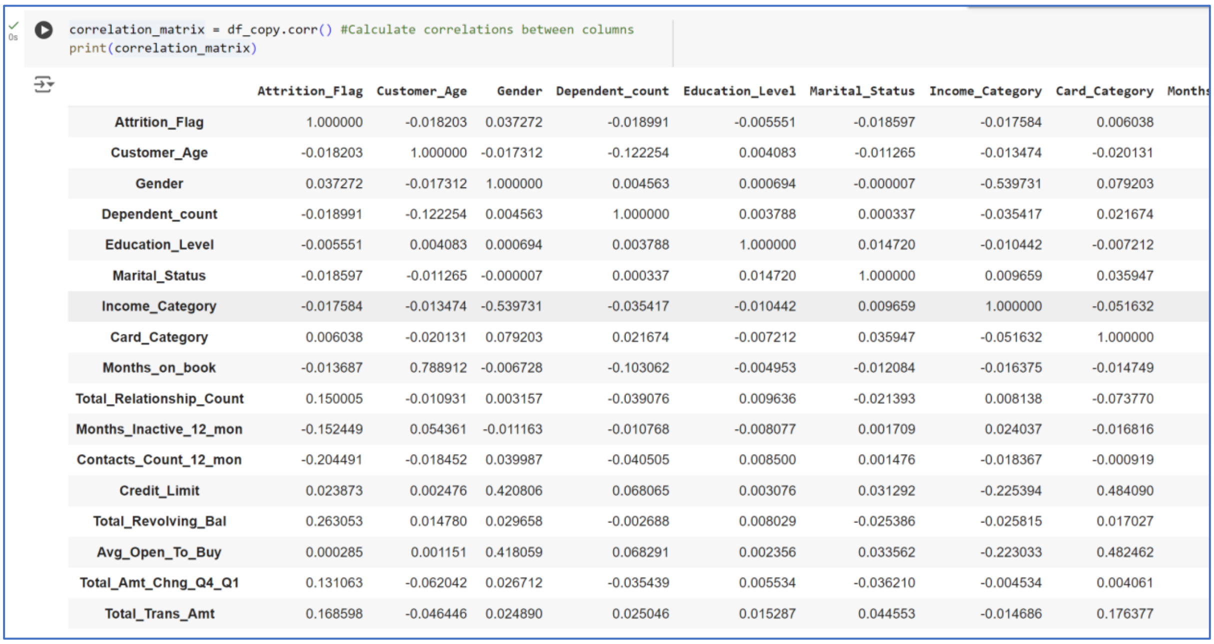
Interpretation:

- The plot shows the change in transaction count from Q4 to Q1 for existing and attrited customers.
- Existing customers tend to have higher median changes in transaction counts compared to attrited customers.
- There is a noticeable spread in changes in transaction counts for both groups, indicating variability in transaction behavior changes.

Analyze Relationships

Heatmap Analysis

To create a heatmap, in first step need with “LabelEncoder” function in “sklearn.preprocessing” categorical value will be encoded to numerical then regard to Figure 13 correlation matrix is created.



```
correlation_matrix = df_copy.corr() #Calculate correlations between columns
print(correlation_matrix)
```

	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category
Attrition_Flag	1.000000	-0.018203	0.037272	-0.018991	-0.005551	-0.018597	-0.017584	0.006038
Customer_Age	-0.018203	1.000000	-0.017312	-0.122254	0.004083	-0.011265	-0.013474	-0.020131
Gender	0.037272	-0.017312	1.000000	0.004563	0.000694	-0.000007	-0.539731	0.079203
Dependent_count	-0.018991	-0.122254	0.004563	1.000000	0.003788	0.000337	-0.035417	0.021674
Education_Level	-0.005551	0.004083	0.000694	0.003788	1.000000	0.014720	-0.010442	-0.007212
Marital_Status	-0.018597	-0.011265	-0.000007	0.000337	0.014720	1.000000	0.009659	0.035947
Income_Category	-0.017584	-0.013474	-0.539731	-0.035417	-0.010442	0.009659	1.000000	-0.051632
Card_Category	0.006038	-0.020131	0.079203	0.021674	-0.007212	0.035947	-0.051632	1.000000
Months_on_book	-0.013687	0.788912	-0.006728	-0.103062	-0.004953	-0.012084	-0.016375	-0.014749
Total_Relationship_Count	0.150005	-0.010931	0.003157	-0.039076	0.009636	-0.021393	0.008138	-0.073770
Months_Inactive_12_mon	-0.152449	0.054361	-0.011163	-0.010768	-0.008077	0.001709	0.024037	-0.016816
Contacts_Count_12_mon	-0.204491	-0.018452	0.039987	-0.040505	0.008500	0.001476	-0.018367	-0.000919
Credit_Limit	0.023873	0.002476	0.420806	0.068065	0.003076	0.031292	-0.225394	0.484090
Total_Revolving_Bal	0.263053	0.014780	0.029658	-0.002688	0.008029	-0.025386	-0.025815	0.017027
Avg_Open_To_Buy	0.000285	0.001151	0.418059	0.068291	0.002356	0.033562	-0.223033	0.482462
Total_Amt_Chng_Q4_Q1	0.131063	-0.062042	0.026712	-0.035439	0.005534	-0.036210	-0.004534	0.004061
Total_Trans_Amt	0.168598	-0.046446	0.024890	0.025046	0.015287	0.044553	-0.014686	0.176377

Figure 8:Correlation Matrix

It this table, the relationship of each column with another column is displayed with a numerical value in the range of -1 to 1. The larger the number, the more direct the connection and the smaller the inverse. And if it is zero, there is no connection. In result following, For more clarity, a heatmap diagram is used for analysis

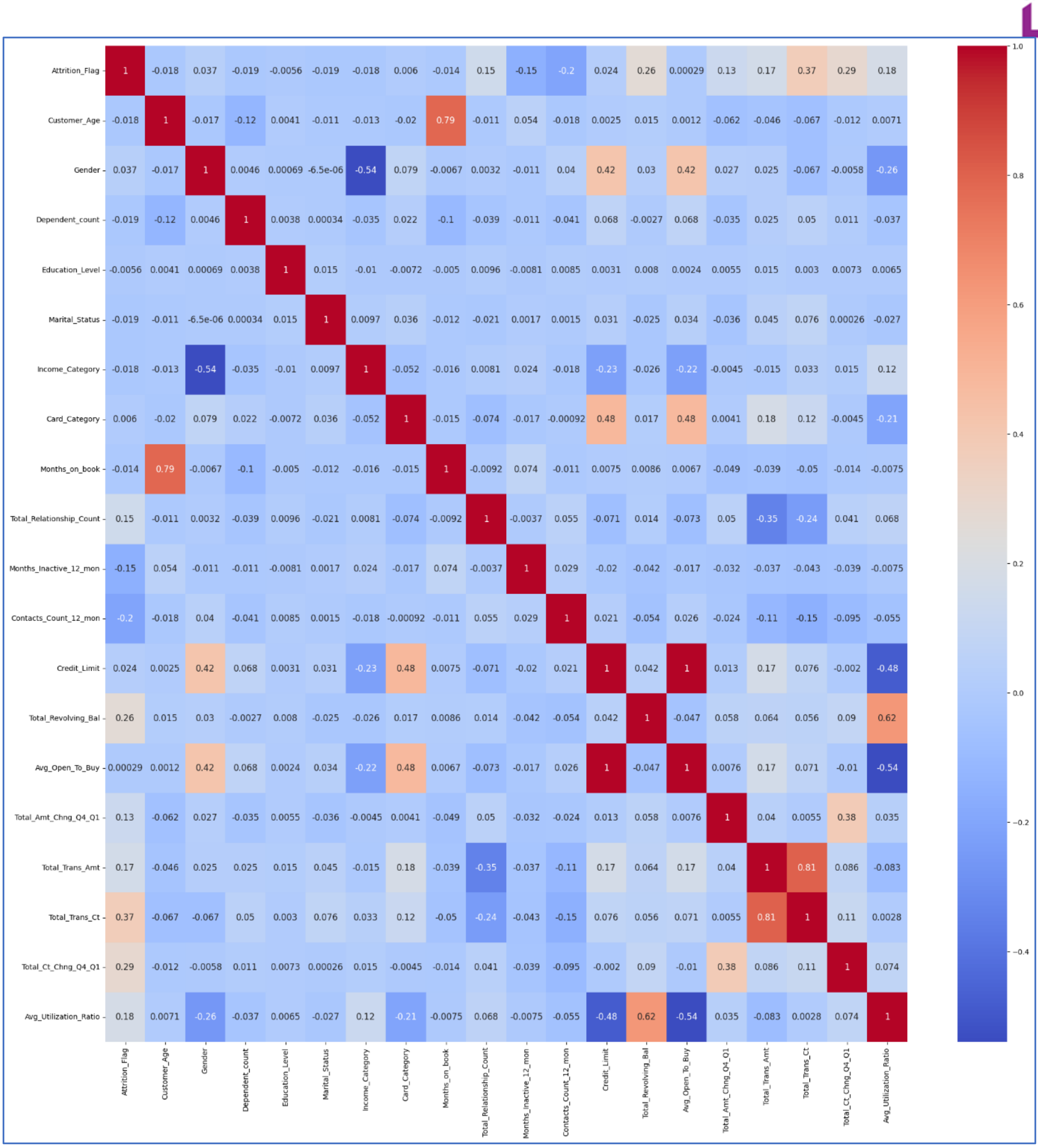


Figure 9:Heatmap Plot

The most important analyzes of this chart are:

Table 5

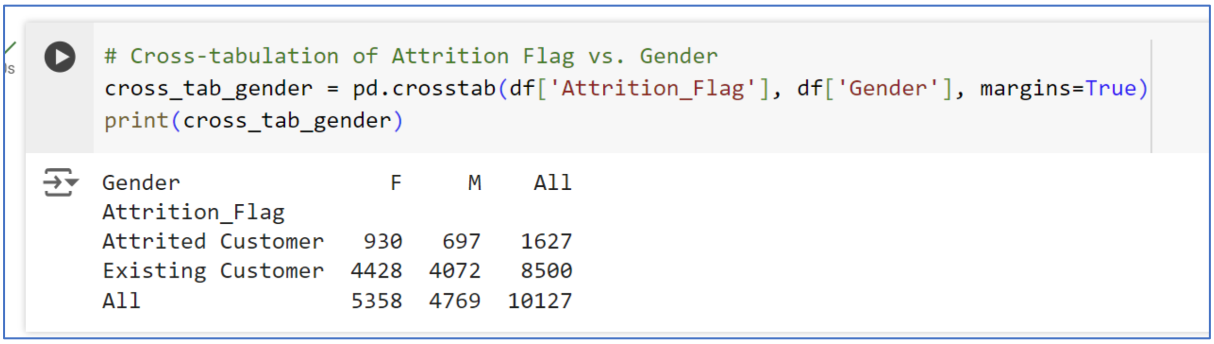
No	Observation	Description
1	Attrition_Flag	Strong positive correlation with Customer_Age, strong negative correlation with Education_Level and Marital_Status.
2	Gender	Relatively low correlations with most other features, except for a moderate positive correlation with Dependent_count.
3	Income_Category & Card_Category	Moderate positive correlations with each other and with some other features like Customer_Age and Months_on_book.
4	Total_Relationship_Count & Months_Inactive_12_mon	Moderate negative correlations with Credit_Limit and Total_Trans_Amt.
5	Feature Clusters	Clusters with moderate to strong positive and negative correlations, like Total_Amt_Chng_Q4_Q1, Total_Trans_Amt, and Total_Trans_Ct.

The above diagram can help to choose which features to use and understand the data better before start making predictive models.

Cross-tabulation Analysis**:**

- **cross-tabulation of 'Attrition Flag' vs. 'Gender'**

This table determine whether the turnover rates of male and female consumers differ significantly.



```
# Cross-tabulation of Attrition Flag vs. Gender
cross_tab_gender = pd.crosstab(df['Attrition_Flag'], df['Gender'], margins=True)
print(cross_tab_gender)
```

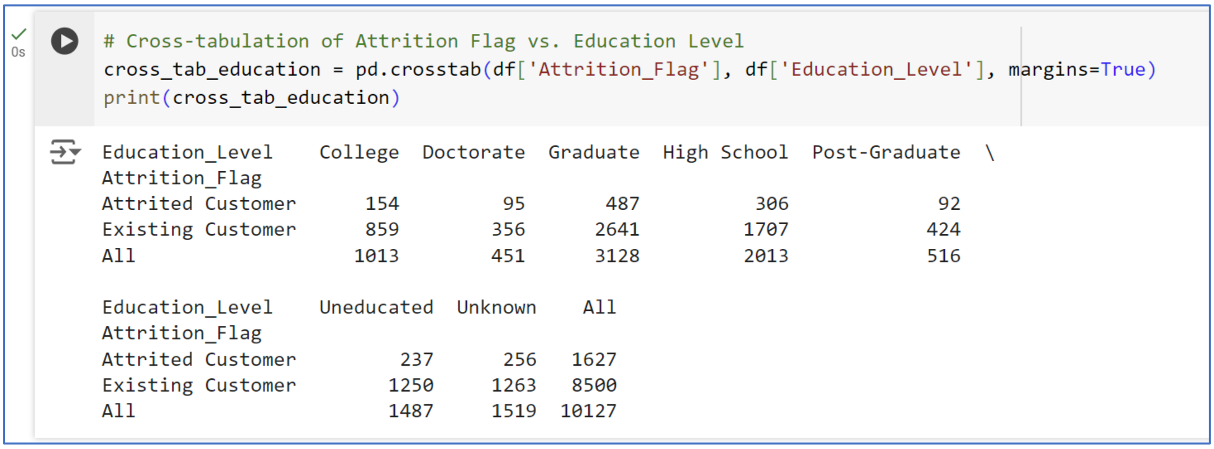
```
Gender                F     M    All
Attrition_Flag
Attrited Customer   930   697   1627
Existing Customer  4428  4072   8500
All                5358  4769  10127
```

Figure 10:Attrition Flag vs Gender

Interpretation:

- The resulting table will show the count of existing and attired customers for each gender.
- The margins will provide the total counts for each category and overall totals.
- This can help identify if one gender is more likely to churn than the other.

- **cross-tabulation of 'Attrition_Flag' vs. 'Education_Level':**



```
# Cross-tabulation of Attrition Flag vs. Education Level
cross_tab_education = pd.crosstab(df['Attrition_Flag'], df['Education_Level'], margins=True)
print(cross_tab_education)
```

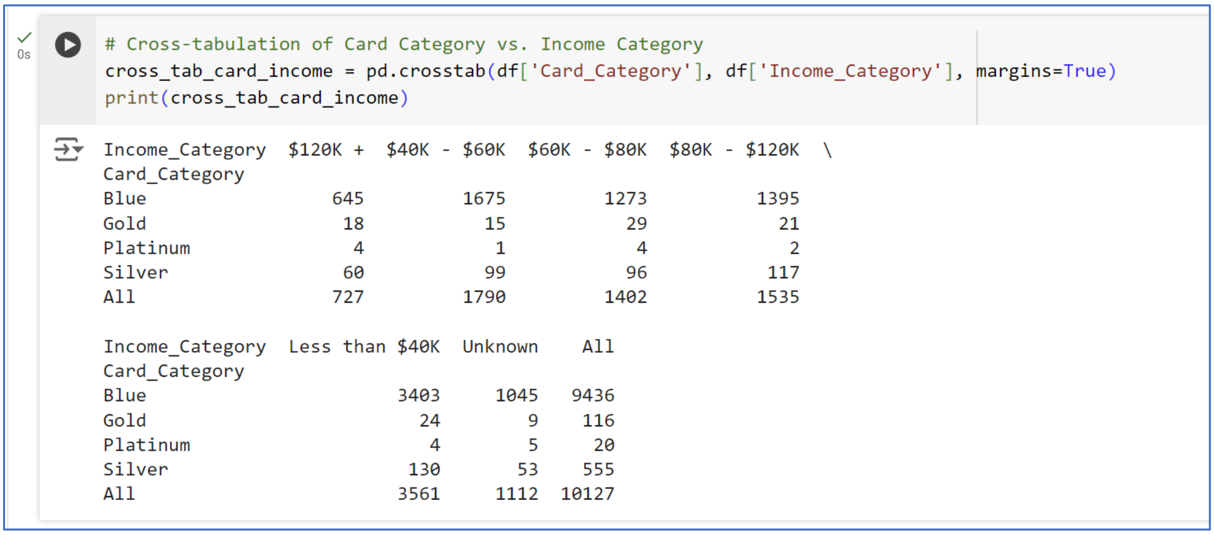
```
Education_Level    College  Doctorate  Graduate  High School  Post-Graduate  \
Attrition_Flag
Attrited Customer      154         95       487          306             92
Existing Customer      859        356      2641         1707            424
All                   1013        451      3128         2013            516

Education_Level    Uneducated  Unknown    All
Attrition_Flag
Attrited Customer         237      256   1627
Existing Customer        1250     1263   8500
All                      1487     1519  10127
```

Figure 11:Attrition_Flag vs Education_Level

Interpretation:

From this table, we can see that customers with higher education levels (Graduate and above) have a lower attrition rate compared to those with lower education levels (High School and below).

- **cross-tabulation of 'Card Category' vs. 'Income Category':**



```
# Cross-tabulation of Card Category vs. Income Category
cross_tab_card_income = pd.crosstab(df['Card_Category'], df['Income_Category'], margins=True)
print(cross_tab_card_income)
```

```
Income_Category  $120K +  $40K - $60K  $60K - $80K  $80K - $120K  \
Card_Category
Blue                 645         1675         1273          1395
Gold                  18           15           29            21
Platinum               4            1            4             2
Silver                60           99           96           117
All                  727         1790         1402          1535

Income_Category  Less than $40K  Unknown    All
Card_Category
Blue                       3403     1045   9436
Gold                         24        9    116
Platinum                      4        5     20
Silver                      130       53    555
All                        3561     1112  10127
```

Figure 12:Card Category vs Income Category

Interpretation:

we can see that the majority of customers have the Blue card across all income categories. The distribution of other card categories (Gold, Platinum, Silver) is relatively small compared to the Blue card.

BarPlot (CountPlot)

In this section, using non-digit (categorize) columns, barcodes for counting Attrition_flag

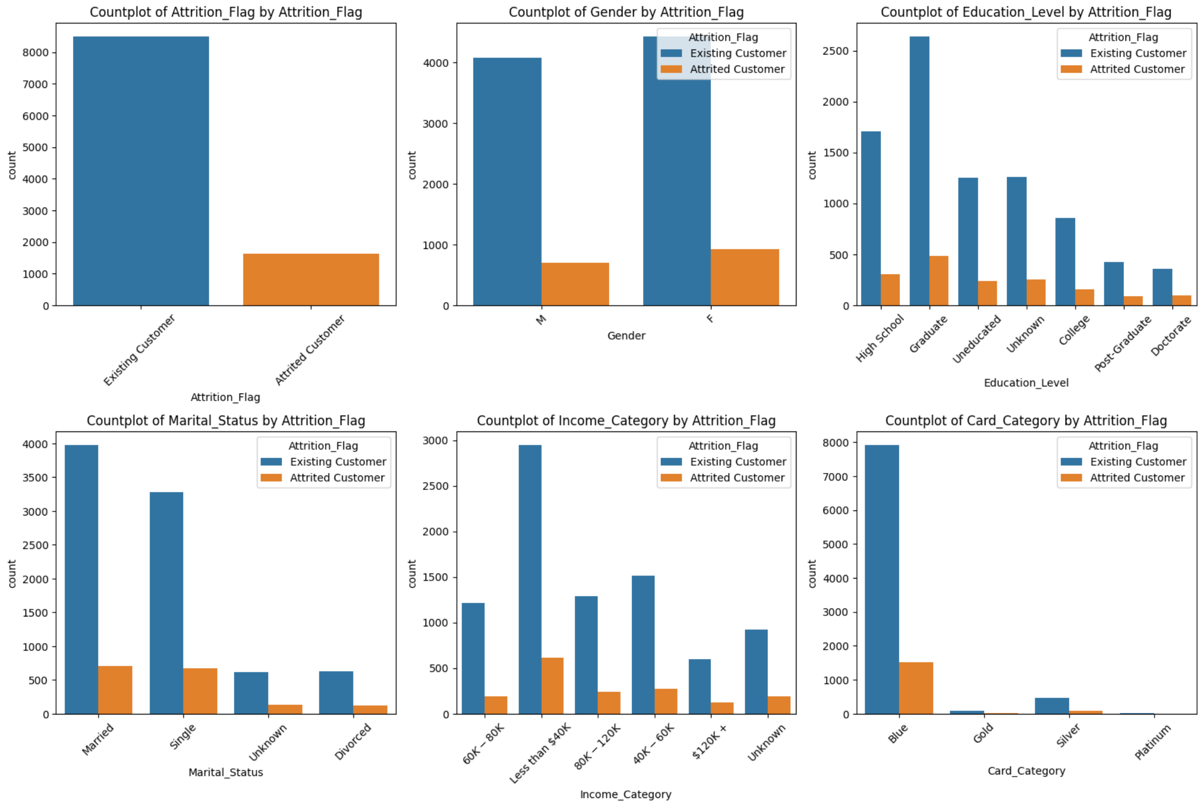


Figure 13:CountPlot

These key conclusions are drawn from the plots (Figure 18):

Table 6

No	Observation	Description
1	Attrition Rate	Low attrition rate based on Attrition_Flag count plot. More current customers than those who left.
2	Gender Distribution	More female customers than male, for both existing and attrition groups.
3	Education Level	Most customers (existing and lost) have undergraduate or graduate degrees.
4	Marital Status	Most customers (both current and lost) are married.
5	Income Distribution	Significant portion of customers (both current and lost) fall in the "$60K - $80K" income category.
6	Card Category	Majority of customers (both current and lost) belong to the Blue card category.
7	Overall Trends	Distributions of categorical variables differ slightly between current and attrition customers, but overall patterns are similar. Suggests other factors may be more influential in customer attrition.

It is noteworthy that these results are predicated just on the given visualisations and that stronger conclusions might need additional statistical analysis or taking other factors into account.

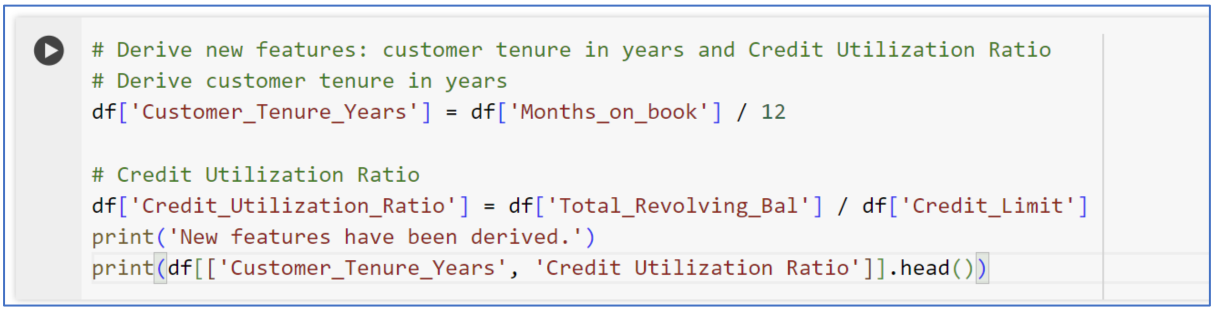
Feature Engineering

Feature engineering is a crucial step in Data Analysis workflow that involves transforming raw data into a format that enhances the performance of machine learning algorithms (Géron, 2019). It's the art and science of creating new features (variables) or modifying existing ones to improve the quality and relevance of the data for the specific machine learning task at hand. Feature engineering can helps by:

- Improving Model Performance
- Enhancing Feature Interpretability to understand how the model arrives at its predictions
- Reducing Overfitting where the model memorizes the training data too well and performs poorly on unseen data.

The suggestions of this report to define new Feature in the dataset are as follows:

- creating `Customer_Tenure_Years` feature (based on 'Months_on_book') is useful as it quantifies customer loyalty, which is a strong indicator of future revenue and retention potential.
- Credit_Utilization_Ratio feature (calculated 'Total_Revolving_Bal' and 'Credit_Limit') could be useful Indicates how much of the available credit a customer is using and high utilization ratios financial risk (Useful for credit risk assessment and managing credit limits)



```
# Derive new features: customer tenure in years and Credit Utilization Ratio
# Derive customer tenure in years
df['Customer_Tenure_Years'] = df['Months_on_book'] / 12

# Credit Utilization Ratio
df['Credit_Utilization_Ratio'] = df['Total_Revolving_Bal'] / df['Credit_Limit']
print('New features have been derived.')
print(df[['Customer_Tenure_Years', 'Credit Utilization Ratio']].head())
```

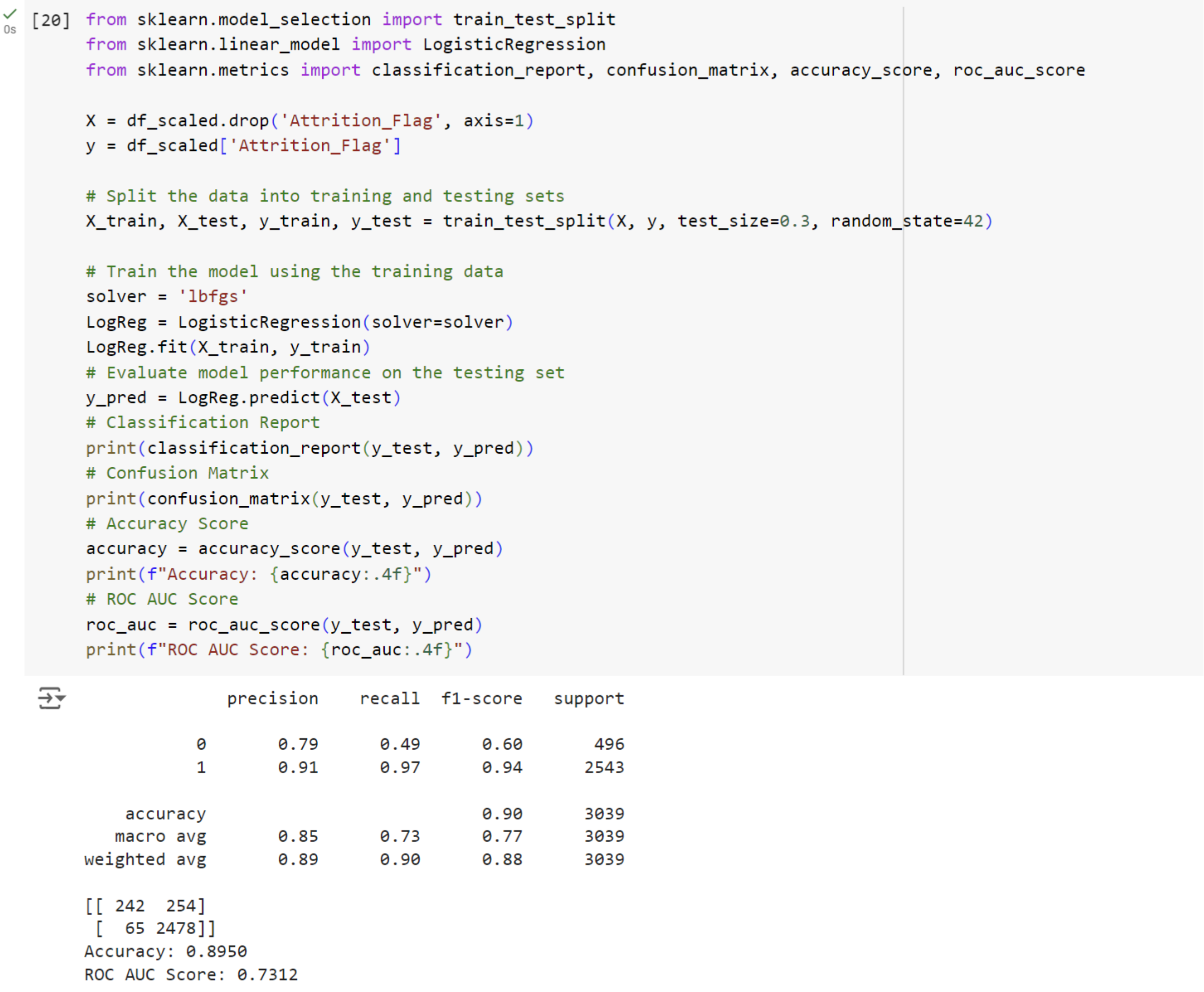
Figure 14: Adding New Features

Customer Segmentation

Customer segmentation by clustering is a data mining technique that groups customers into distinct segments or clusters based on their similarities in behavior, characteristics, or preferences, without any predetermined target variable (Ngai et al., 2009; Dolnicar et al., 2018).

We can use grouping methods like Logistic regression to divide customers into groups. These are the steps we'll take:

- **Preprocess the Data:** Normalize the features to ensure they are on a similar scale. In the dataset df_copy is the data which all the non-numerical data converted to numerical by LableEncoder and then by MinMax scaler scaled data in the data set 'df_scaled' created (fig)
- **Determine the Optimal Number of Clusters:** Use the Elbow Method to find the optimal number of clusters. Customer Attrition made two cluster
- **Apply Logistic regression Clustering:** Segment the customers based on their behaviour and profile.



```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score, roc_auc_score

X = df_scaled.drop('Attrition_Flag', axis=1)
y = df_scaled['Attrition_Flag']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Train the model using the training data
solver = 'lbfgs'
LogReg = LogisticRegression(solver=solver)
LogReg.fit(X_train, y_train)
# Evaluate model performance on the testing set
y_pred = LogReg.predict(X_test)
# Classification Report
print(classification_report(y_test, y_pred))
# Confusion Matrix
print(confusion_matrix(y_test, y_pred))
# Accuracy Score
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy:.4f}")
# ROC AUC Score
roc_auc = roc_auc_score(y_test, y_pred)
print(f"ROC AUC Score: {roc_auc:.4f}")
```

```
              precision    recall  f1-score   support

           0       0.79      0.49      0.60       496
           1       0.91      0.97      0.94      2543

    accuracy                           0.90      3039
   macro avg       0.85      0.73      0.77      3039
weighted avg       0.89      0.90      0.88      3039

[[ 242  254]
 [  65 2478]]
Accuracy: 0.8950
ROC AUC Score: 0.7312
```

Figure 15: Logistic Regression

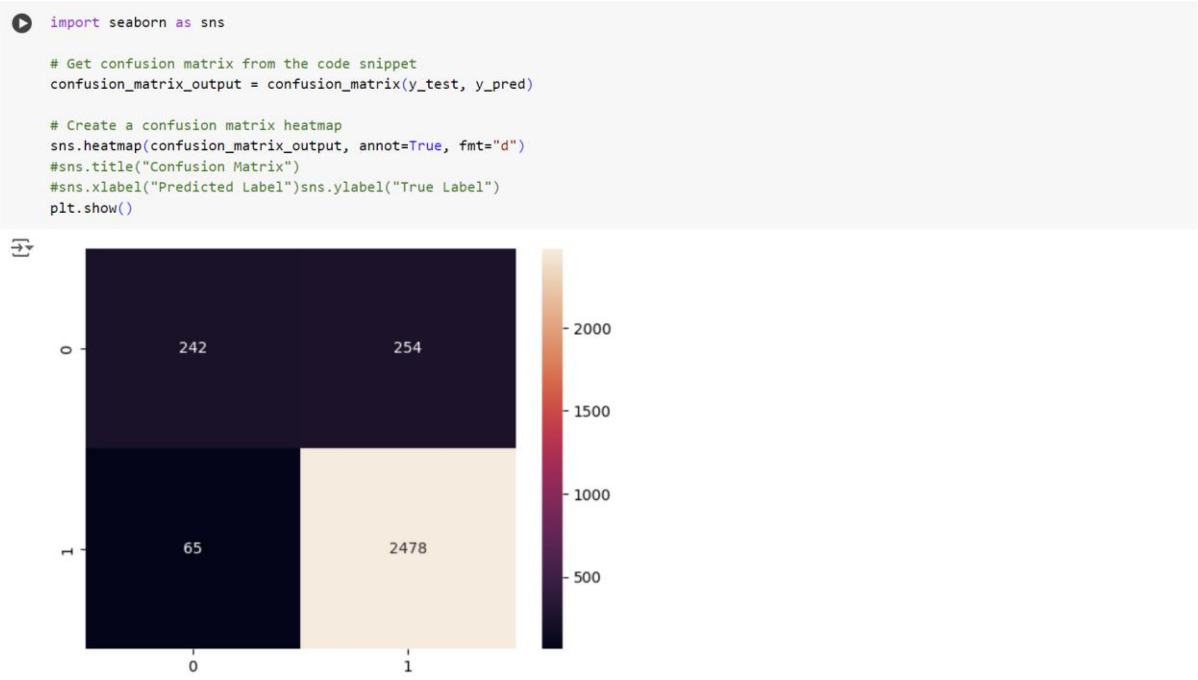
The attached logistic regression model review shows that it did a great job, with the following results:

Table 7

No	Performance Metric	Description
1	Classification Report	High performance for both classes, especially class 1 (higher precision, recall, and f1-score).
2	Accuracy Score	High accuracy score of approximately 89%.
3	ROC AUC Score	Excellent class separability with a score of approximately 95%.

Even though it's not perfect, this performance is pretty good and shows that the logistic regression model can accurately describe the test data and do a great job of separating Classification.

- **Visualize the Accuracy**



```
import seaborn as sns

# Get confusion matrix from the code snippet
confusion_matrix_output = confusion_matrix(y_test, y_pred)

# Create a confusion matrix heatmap
sns.heatmap(confusion_matrix_output, annot=True, fmt="d")
#sns.title("Confusion Matrix")
#sns.xlabel("Predicted Label")sns.ylabel("True Label")
plt.show()
```

Figure 16: Logistic Regression Accuracy

The confusion matrix heatmap (Figure 21) displays a logistic regression model's performance. is broken down here:

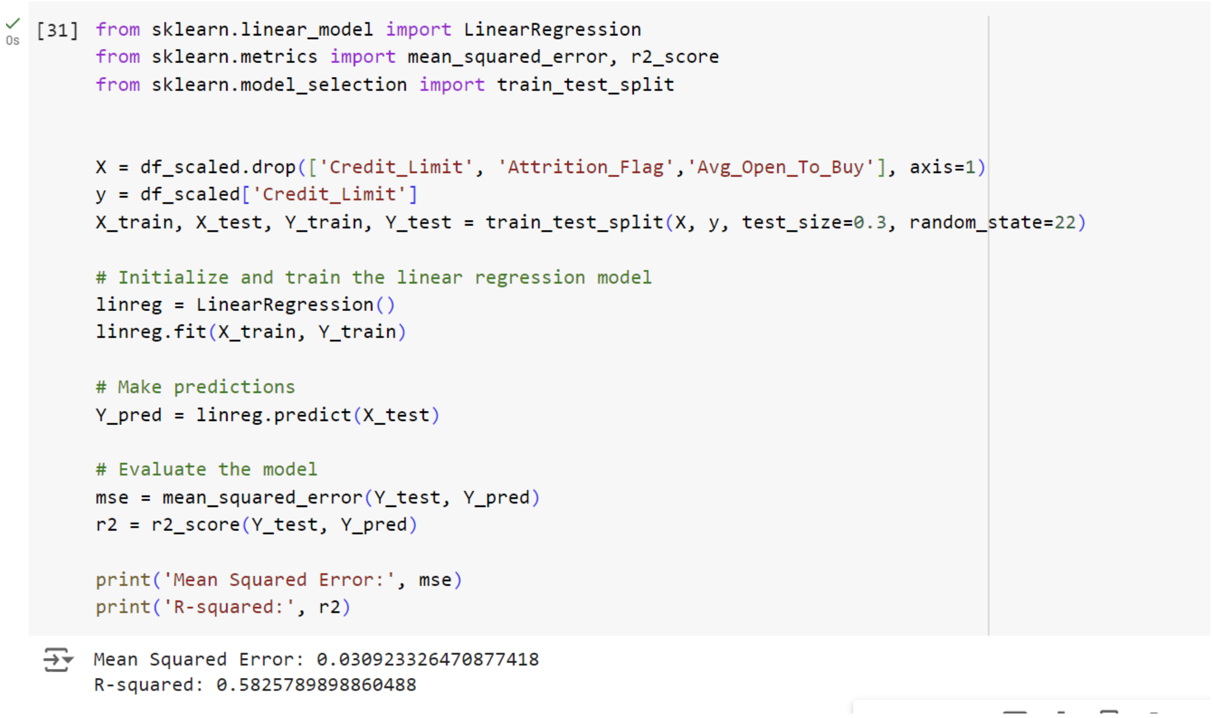
Table 8

No	Performance Metric	Description
1	True Positives (TP)	Model correctly predicted 242 positive cases.
2	False Positives (FP)	Model incorrectly predicted 254 cases as positive (when they were negative).
3	False Negatives (FN)	Model incorrectly predicted 8 cases as negative (when they were positive).
4	True Negatives (TN)	Model correctly predicted 2478 negative cases.

The heatmap's colours show how often each result happens, with darker colours showing more often occurrences. We can use this grid to figure out not only how accurate the model is, but also what kinds of mistakes it is making.

Linear Regression to define credit limitation

We want to use linear regression to set the credit limit for bank users. In this case, the code below says to remove the "Credit_Limit" and "Attrition_Flag" columns first. Also, the heatmap shows that this is also removed because the "Credit_Limit" and "Avg_Open_To_Buy" columns are completely correlated. After that, the below code is run.



```
[31] from sklearn.linear_model import LinearRegression
     from sklearn.metrics import mean_squared_error, r2_score
     from sklearn.model_selection import train_test_split


     X = df_scaled.drop(['Credit_Limit', 'Attrition_Flag','Avg_Open_To_Buy'], axis=1)
     y = df_scaled['Credit_Limit']
     X_train, X_test, Y_train, Y_test = train_test_split(X, y, test_size=0.3, random_state=22)

     # Initialize and train the linear regression model
     linreg = LinearRegression()
     linreg.fit(X_train, Y_train)

     # Make predictions
     Y_pred = linreg.predict(X_test)

     # Evaluate the model
     mse = mean_squared_error(Y_test, Y_pred)
     r2 = r2_score(Y_test, Y_pred)

     print('Mean Squared Error:', mse)
     print('R-squared:', r2)
```

```
Mean Squared Error: 0.030923326470877418
R-squared: 0.5825789898860488
```

Figure 17: Linear Regression

Based on the code output, a linear regression model the following results:

Table 9

No	Performance Metric	Extracted Numerical Value	Description
1	Mean Squared Error (MSE)	0.0309	The MSE shows the average squared difference between what was expected and what happened. If the MSE is smaller, it means that the model fits the data better.
2	R-squared (R2)	0.5826	The R2 number represents the proportion of variance explained by the model.

This value shows how much of the variation in the dependent variable can be predicted by the independent factors. If the R2 number is 1, it means that the model perfectly predicts what will happen. If it's close to 0, it means that it doesn't explain well how the answer data changes around its mean.

Based on these results, it looks like the model can make some predictions, but it doesn't explain a big chunk of the variation in the goal variable. To make the model more accurate, it might need to be tweaked or more data may need to be added.

Hypothesis Testing and A/B Testing

Before we can choose the best testing method for the "credit_card_churn" dataset, we need to know what the problem is and what the analysis's goals are. In the following is description of both hypothesis testing and A/B testing, including which one is preferable for this problem:

- **A/B testing**, which is also called split testing, is a way to compare two or more copies of a variable to see which one works better. Here are the steps that are needed:
    1. Define Variants: Create two or more versions (A and B) of the variable we want to test.
    2. Randomly Assign Subjects: Randomly assign subjects to either the control group (A) or the treatment group (B).
    3. Collect Data
    4. Analyse Results: Use statistical methods to compare the performance
    5. Make a Decision
- **Hypothesis testing** is a way to use statistics to draw conclusions about a whole community from a small sample. These steps are needed to do it:
    1. Formulate Hypotheses: Determine the alternative hypothesis (H1) and the null hypothesis (H0).
    2. Pick out a level of importance: Usually written as alpha (α), and the default value is 0.05.
    3. Collect Data: Gathering sample data relevant to the hypotheses.
    4. Perform Statistical Test: Using statistics tests like t-test and the chi-square test for analysing the data.
    5. Make a Decision: Based on the p-value, decide whether to reject or fail to reject the null hypothesis.

The "credit_card_churn" information is used to look at how customers behave and figure out what causes them to churn. The objective is to learn about the traits of people who leave and come up with ways to lower this behaviour. It's better to use **hypothesis testing** for this problem because:

- We can come up with assumptions about what causes customers to leave (for example, "Customers who use their credit cards more often are more likely to leave").
- We can use statistical tests to see if there are important differences between customers who have left and customers who have not, based on things like time, transaction amount, and credit utilisation.
- It lets us draw conclusions about the whole community from a small sample and figure out what causes people to leave.

Key Findings from the Data Analytics Report

The study on credit card churn prediction came up with the following main points and suggestions, written in a way that is easy for non-technical readers to understand:

Key Findings:

1. **Customer Churn Analysis:**
    a. Existing customers tend to have higher transaction amounts compared to those who have left (attrited customers).
    b. Customers with higher income levels generally have higher credit limits and transaction counts.

   c. Education level and marital status seem to have some influence on credit utilization behaviour and the likelihood of attrition.

2. **Important Factors for Churn:** The analysis identified several important factors that may contribute to customer attrition, including:
   a. Customer age
   b. Income level
   c. Credit utilization ratio
   d. Transaction behaviour (amounts and frequency)
   e. Duration of relationship with the bank (tenure)
3. **Customer Segmentation:** Clustering methods were used in the study to divide customers into groups based on their profiles and how they behave. This can help find high-risk groups that are more likely to leave and allow for more focused tactics to keep them.
4. **Credit Limit Modelling:** Based on customers' behaviours and actions, a linear regression model was created to guess what credit limits would be best for them. The model wasn't perfect, but it did a pretty good job of predicting what would happen.

Recommendations:

1. **Targeted Retention Efforts:** Using what we learn from customer segmentation and churn research, we can create focused ads to keep high-risk groups of customers. To handle their special wants and concerns, give them incentives, benefits, or better services that are made just for them.
2. **Credit Limit Optimization:** Review and change customer credit limits on a regular basis using the credit limit forecast model. This can help control risk and make sure that customers can get the right kind of loans based on their traits.
3. **Continuous Monitoring:** Install systems that will track consumer behaviour, transaction patterns, and changes in important variables that might indicate a higher attrition risk on a frequent basis. This would make preventive actions and customised outreach possible to keep important clients.
4. **Customer Experience Enhancements:** Using the reasons why customers leave as a guide, look for ways to improve the whole experience for customers. This could mean making benefits programmes better, easing customer service, or adding new features to products that appeal to a wide range of customers.
5. **Data-Driven Decision Making:** Continue investing into data analytics to learn more about how customers behave and what they like. Regularly improve and update prediction models as new data comes in. This will make sure that choices are based on the most up-to-date and correct data.

Following these suggestions will help the credit card company keep customers instead of losing them, make customers happier, and eventually make more money.

References

- Akkio (2023, February 22). Predicting Credit Card Customer Churn. https://www.akkio.com/applications/churn-reduction
- Credit Card Churn Prediction (no date). Available at: https://www.kaggle.com/datasets/anwarsan/credit-card-bank-churn .
- Customer Churn Prediction on Credit Card Services using Random Forest Method (2016) [Report]. Atlantis Press.
- Data Cleaning: Definition, Benefits, And How-To | Tableau (no date). Available at: https://www.tableau.com/learn/articles/what-is-data-cleaning.
- Dolnicar, S., Grün, B., & Leisch, F. (2018). Market segmentation analysis: Understanding it, doing it, and making it useful. Springer.
- Ferydooni, F. (2024). Credit Card Churn. Available at: https://github.com/fferydooni/credit_card_churn
- Frederick Reichheld & Thomas Teal (2000). The One Number You Need to Grow. Harvard Business Review, 78(6), 106-114.
- Hughes, J.M. (2010). Real World Instrumentation with Python. 'O'Reilly Media, Inc.'
- Jio, L.D.S.& G., Reliance (2024) 'Importance of Boxplot in Exploratory Data Analysis (EDA)', Medium, 7 February. Available at: https://medium.com/@venugopal.adep/importance-of-boxplot-in-exploratory-data-analysis-eda-5119944850a3 (Accessed: 20 May 2024).
- Machine Learning to Develop Credit Card Customer Churn Prediction (2023, January 23). MDPI https://www.mdpi.com/0718-1876/17/4/77
- Moltzau, A. (2019) 'What is Kaggle?', DataSeries, 10 December. Available at: https://medium.com/dataseries/what-is-kaggle-4751e384e916 (Accessed: 29 April 2024).
- Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36(2), 2592-2602.
- Verbraken, T., Verbeke, W., & Baesens, B. (2018). Behavioral attributes and financial churn prediction. EPJ Data Science, 7(1), 16. (https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0165-5)
- Welcome - YData Profiling (no date). Available at: https://docs.profiling.ydata.ai/latest/ .

Appendix 1: YData Profiling

During the research of this project, came across the product of Ydata company called Ydata Profiling, which can extract many of the operations performed in the form of code in this research in one report, so it is briefly explained in this section(*Welcome - YData Profiling*, no date).

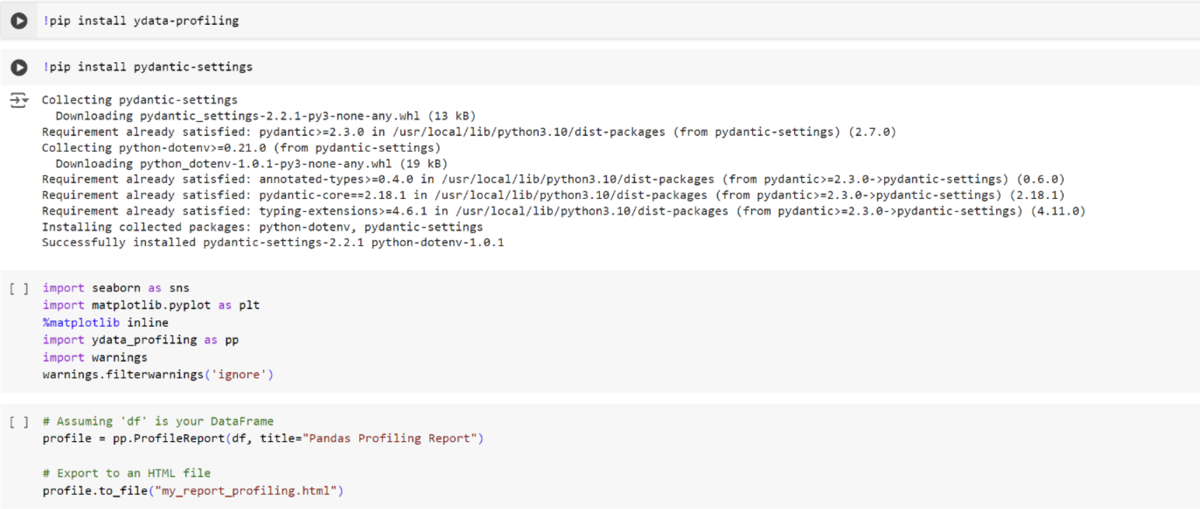
A Python module called YData Profiling (previously known as Pandas Profiling) is intended for exploratory data analysis (EDA). Its principal objective is to generate detailed reports for tabular datasets, hence offering an EDA experience in a single line.

With YData Profiling, a DataFrame is analysed in greater detail than with the standard df.describe() method in Pandas. It makes it possible to effectively study and comprehend your data. Exporting the produced reports is possible in HTML and JSON among other forms(*Welcome - YData Profiling*, no date). Key Features of YData Profiling is:

- **Dataset Statistics**: YData Profiling provides descriptive statistics for each column, including mean, median, minimum, maximum, and more.
- **Data Types:** Information about the data types of columns.
- **Missing Values:** Identification of missing values in the dataset.
- **Memory Usage:** Insights into memory usage.
- **Distribution Plots:** Histograms, kernel density plots, and box plots.
- **Unique Values:** Count of unique values in categorical columns.
- **Sample Data:** A preview of the first few rows of the dataset.

YData Profiling also supports time-series and text analysis.

To use this library, it must first be installing ydata-profiling library and pydantic-settings then its output will be created as a file html or json format.



```
!pip install ydata-profiling
```

```
!pip install pydantic-settings
```

```
Collecting pydantic-settings
  Downloading pydantic_settings-2.2.1-py3-none-any.whl (13 kB)
Requirement already satisfied: pydantic>=2.3.0 in /usr/local/lib/python3.10/dist-packages (from pydantic-settings) (2.7.0)
Collecting python-dotenv>=0.21.0 (from pydantic-settings)
  Downloading python_dotenv-1.0.1-py3-none-any.whl (19 kB)
Requirement already satisfied: annotated-types>=0.4.0 in /usr/local/lib/python3.10/dist-packages (from pydantic>=2.3.0->pydantic-settings) (0.6.0)
Requirement already satisfied: pydantic-core==2.18.1 in /usr/local/lib/python3.10/dist-packages (from pydantic>=2.3.0->pydantic-settings) (2.18.1)
Requirement already satisfied: typing-extensions>=4.6.1 in /usr/local/lib/python3.10/dist-packages (from pydantic>=2.3.0->pydantic-settings) (4.11.0)
Installing collected packages: python-dotenv, pydantic-settings
Successfully installed pydantic-settings-2.2.1 python-dotenv-1.0.1
```

```
[ ] import seaborn as sns
    import matplotlib.pyplot as plt
    %matplotlib inline
    import ydata_profiling as pp
    import warnings
    warnings.filterwarnings('ignore')
```

```
[ ] # Assuming 'df' is your DataFrame
    profile = pp.ProfileReport(df, title="Pandas Profiling Report")

    # Export to an HTML file
    profile.to_file("my_report_profiling.html")
```

Figure 18: YData Code

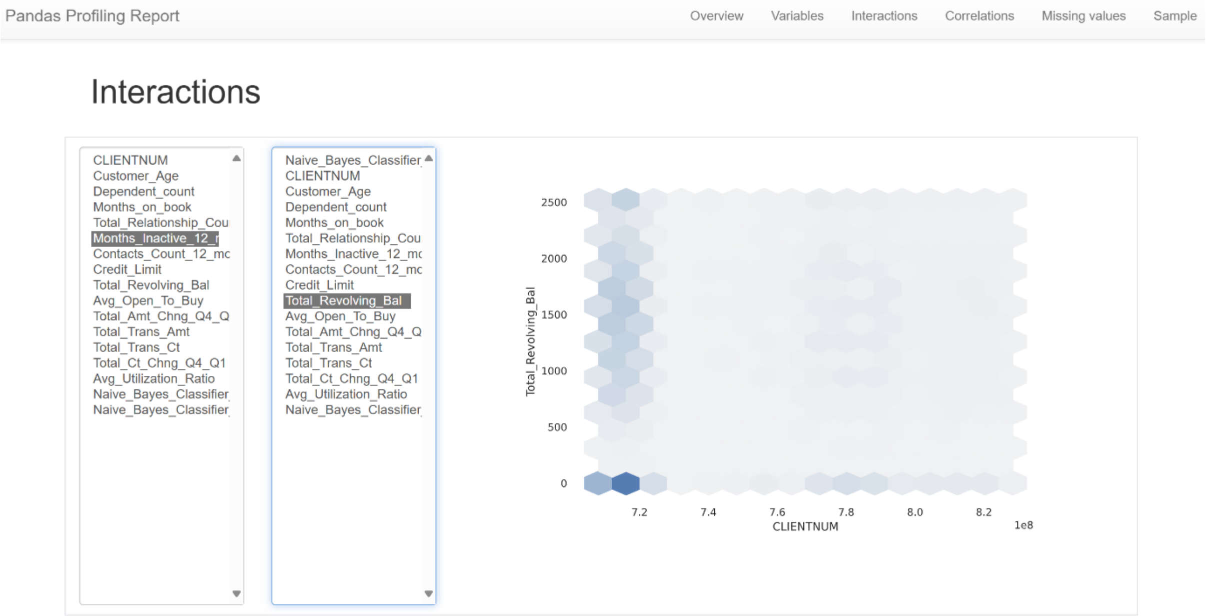


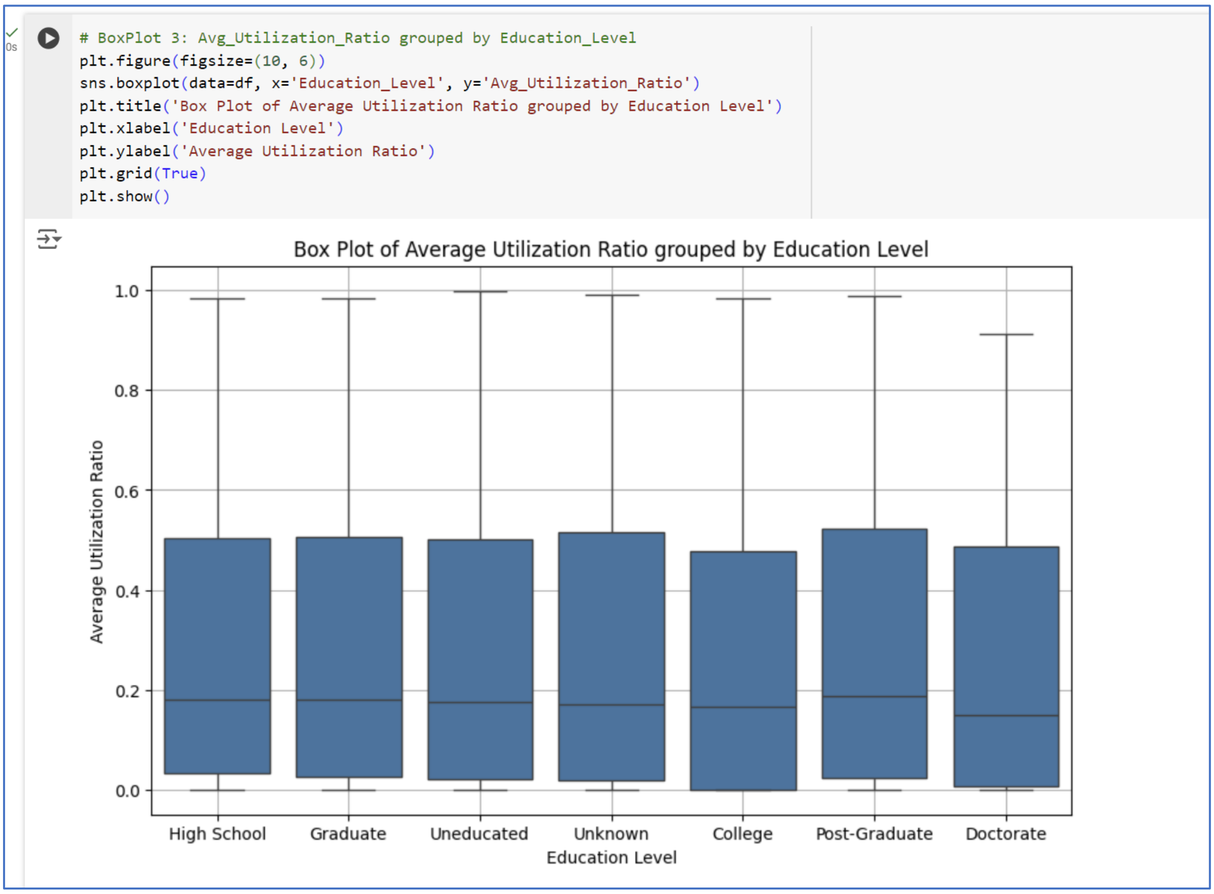
Figure 19: YData Output available at (Ferydooni, 2024)

Appendix3: deeper Analysis

More Data Analysis based on Boxplot

- **Avg_Utilization_Ratio grouped by Education_Level:**

This will help to see how the average utilization ratio varies across different education levels. It can provide insights into whether education level influences credit utilization behavior.



```
# BoxPlot 3: Avg_Utilization_Ratio grouped by Education_Level
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='Education_Level', y='Avg_Utilization_Ratio')
plt.title('Box Plot of Average Utilization Ratio grouped by Education Level')
plt.xlabel('Education Level')
plt.ylabel('Average Utilization Ratio')
plt.grid(True)
plt.show()
```

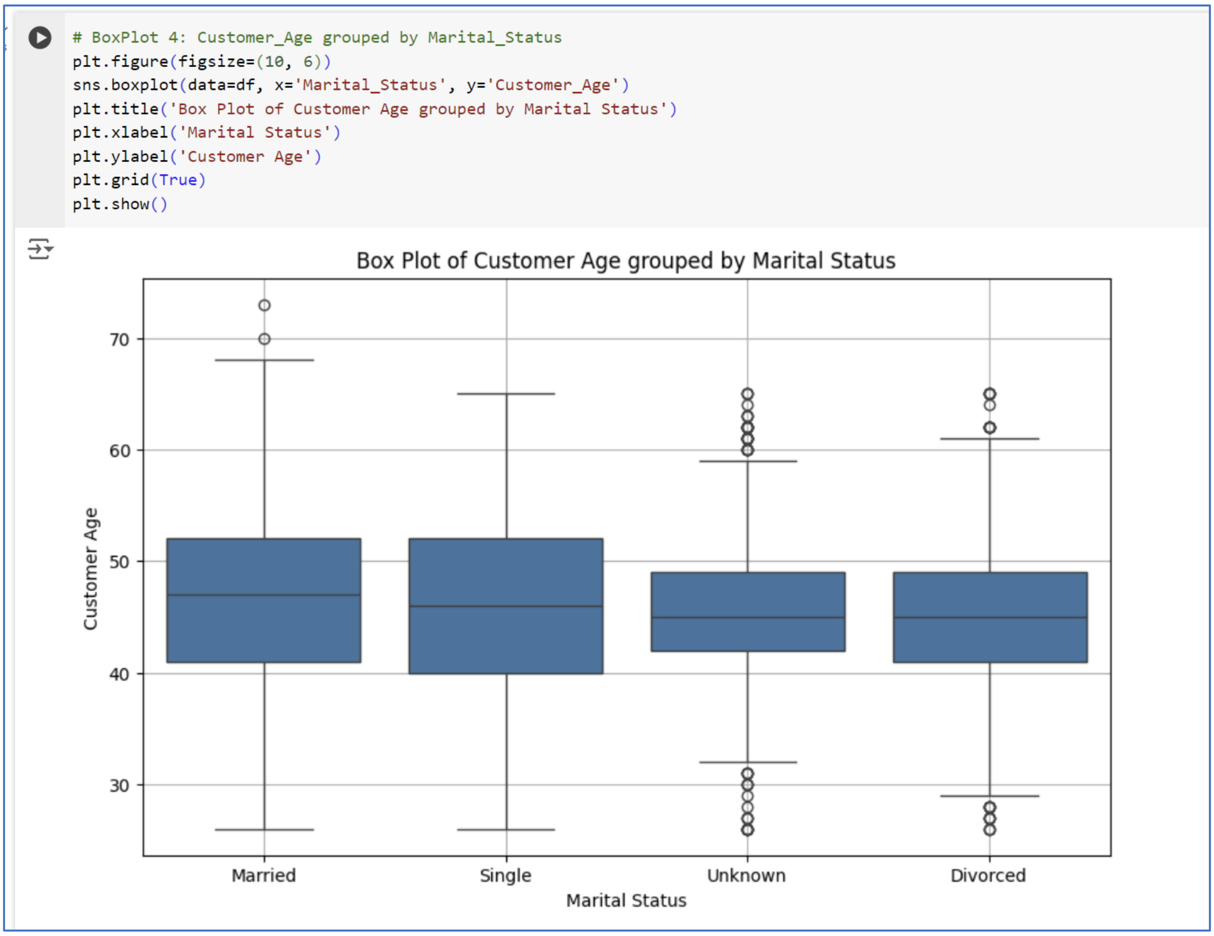
Figure 20:Avg_Utilization_Ratio grouped by Education_Level

Interpretation:

- The plot shows the distribution of average utilization ratios across different education levels.
- There is no clear trend indicating that education level significantly affects the average utilization ratio.
- The median utilization ratios are relatively similar across different education levels, with some variability within each group.

- **Customer_Age grouped by Marital_Status:**

This will show the distribution of customer ages across different marital statuses. It can help identify if there are age-related trends in marital status among the customers.



```
# BoxPlot 4: Customer_Age grouped by Marital_Status
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='Marital_Status', y='Customer_Age')
plt.title('Box Plot of Customer Age grouped by Marital Status')
plt.xlabel('Marital Status')
plt.ylabel('Customer Age')
plt.grid(True)
plt.show()
```

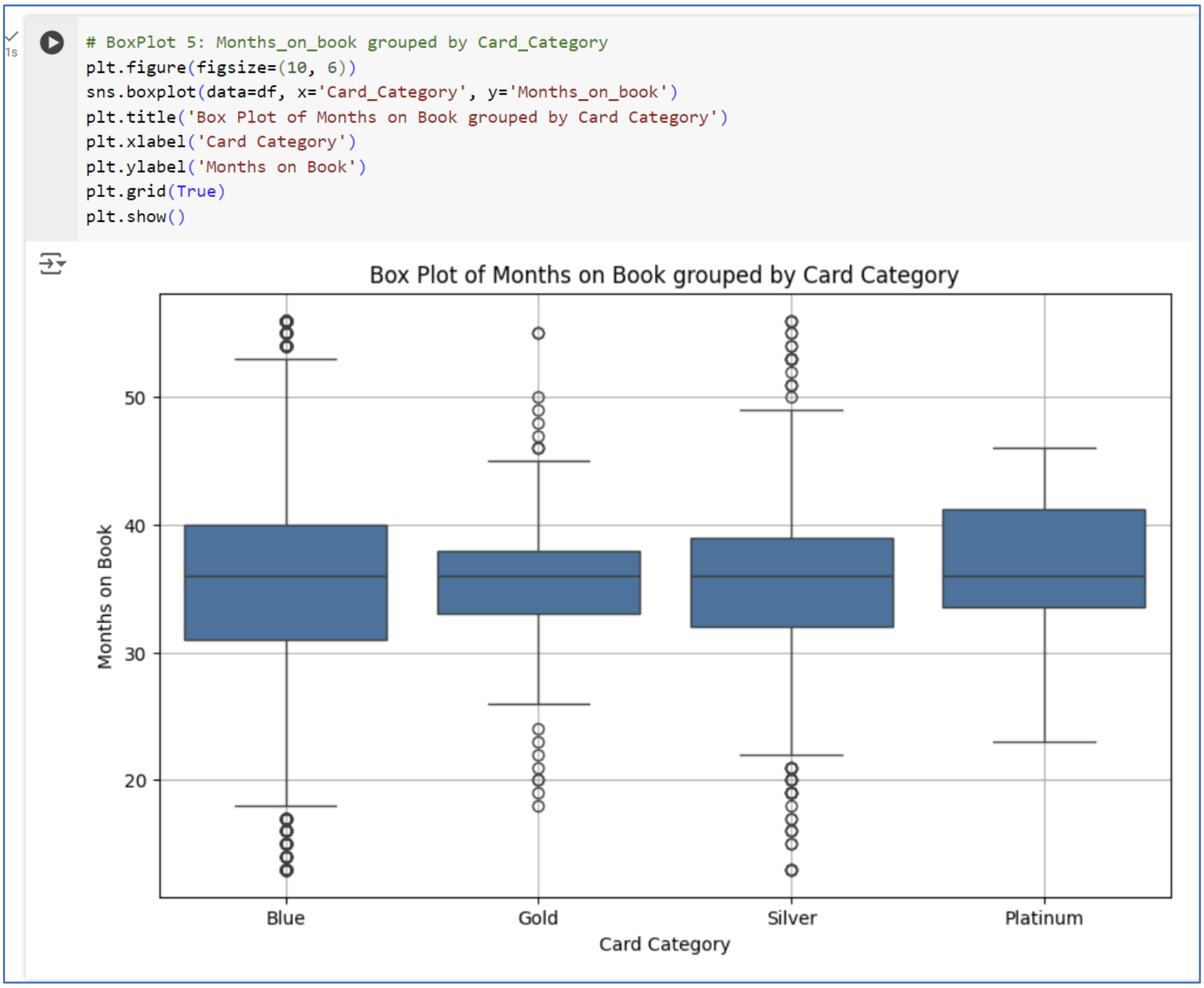
Figure 21:Customer_Age grouped by Marital_Status

Interpretation:

- The plot shows the distribution of customer ages across different marital statuses.
- The median age is relatively similar across different marital statuses, with some variability within each group.
- There is a noticeable spread in ages for each marital status, indicating a diverse age range within each group.

- **Months_on_book grouped by Card_Category:**

This will help to understand how the length of time customers have been with the bank varies across different card categories. It can provide insights into customer loyalty and card preferences.



```
# BoxPlot 5: Months_on_book grouped by Card_Category
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='Card_Category', y='Months_on_book')
plt.title('Box Plot of Months on Book grouped by Card Category')
plt.xlabel('Card Category')
plt.ylabel('Months on Book')
plt.grid(True)
plt.show()
```

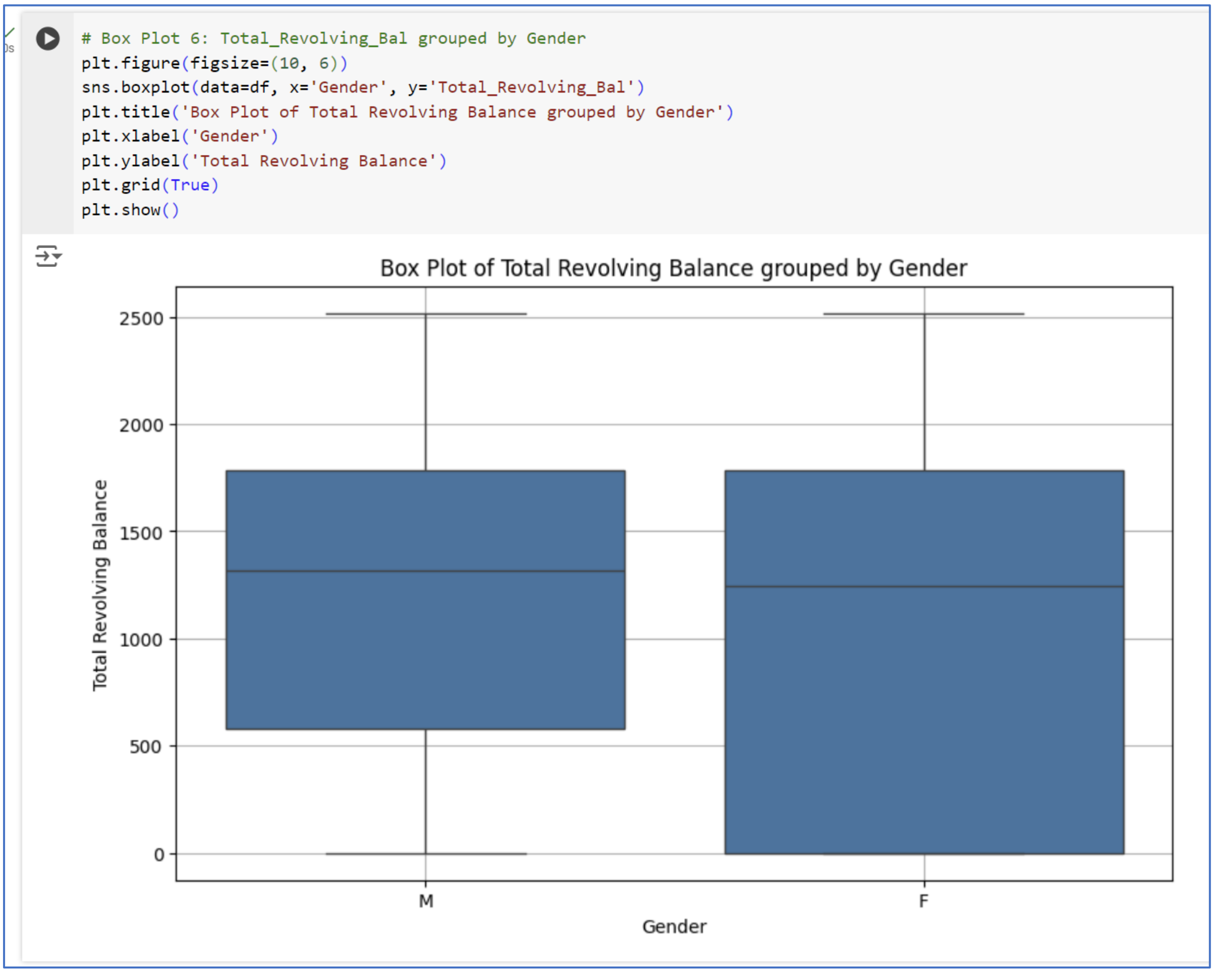
Figure 22:Months_on_book grouped by Card_Category

Interpretation:

- The plot shows the distribution of the number of months customers have been with the bank across different card categories.
- The median number of months is relatively similar across different card categories, with some variability within each group.
- There is a noticeable spread in the number of months for each card category, indicating a diverse range of customer tenures within each group.

- **Total_Revolving_Bal grouped by Gender:**

This will show the distribution of total revolving balances for male and female customers. It can help identify if there are gender-related differences in revolving credit behavior.



```
# Box Plot 6: Total_Revolving_Bal grouped by Gender
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='Gender', y='Total_Revolving_Bal')
plt.title('Box Plot of Total Revolving Balance grouped by Gender')
plt.xlabel('Gender')
plt.ylabel('Total Revolving Balance')
plt.grid(True)
plt.show()
```

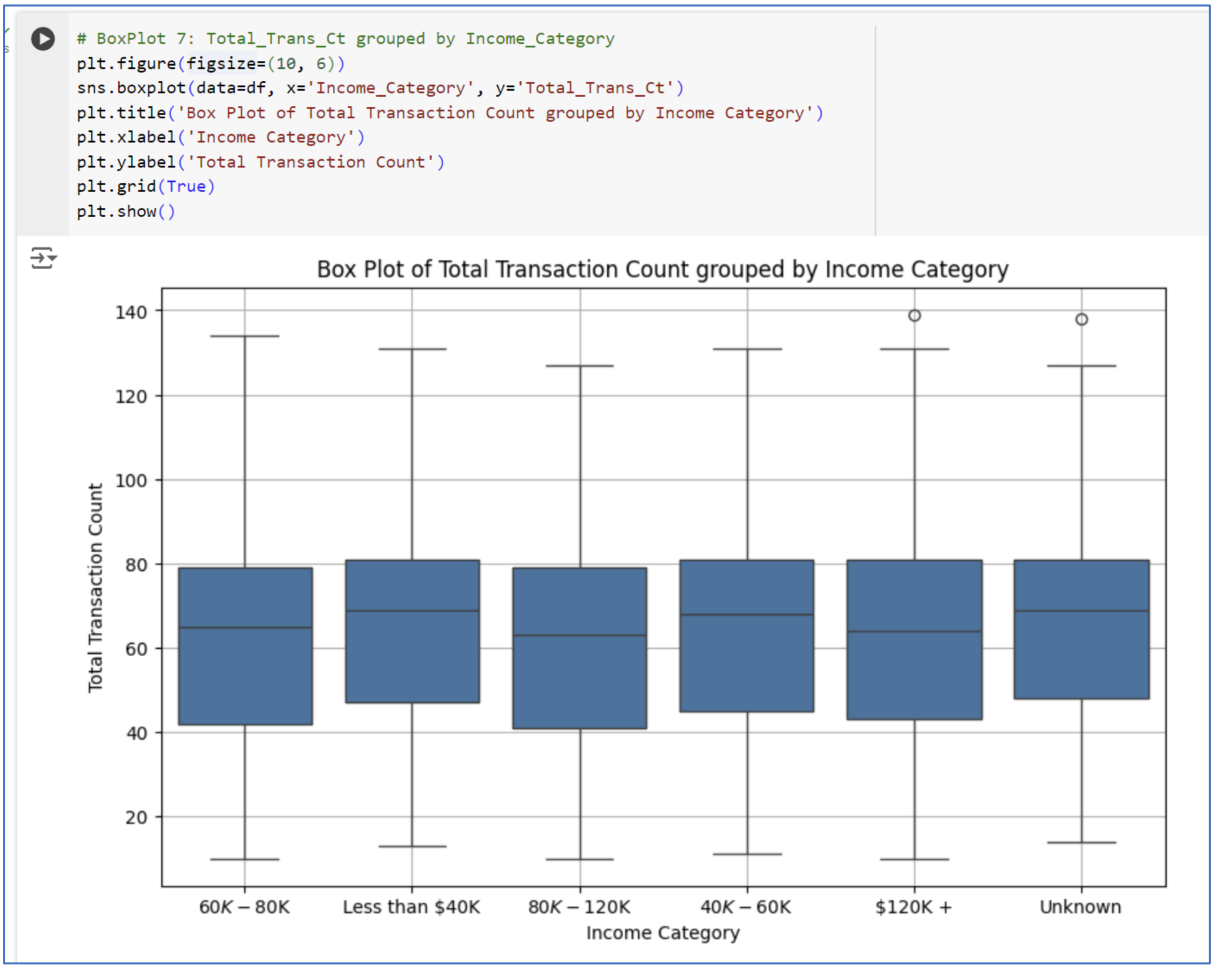
Figure 23:Total_Revolving_Bal grouped by Gender

Interpretation:

- The plot shows the distribution of total revolving balances for male and female customers.
- Both genders have similar median revolving balances.
- There is a noticeable spread in revolving balances for both genders, indicating variability in revolving credit behavior.

- **Total_Trans_Ct grouped by Income_Category:**

This will help us see how the total number of transactions varies across different income categories. It can provide insights into spending behavior based on income levels.



```
# BoxPlot 7: Total_Trans_Ct grouped by Income_Category
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='Income_Category', y='Total_Trans_Ct')
plt.title('Box Plot of Total Transaction Count grouped by Income Category')
plt.xlabel('Income Category')
plt.ylabel('Total Transaction Count')
plt.grid(True)
plt.show()
```

Figure 24:Total_Trans_Ct grouped by Income_Category

Interpretation:

- The plot shows the distribution of total transaction counts across different income categories.
- Higher income categories tend to have higher median transaction counts.
- There is a noticeable spread in transaction counts within each income category, indicating variability in spending behavior based on income levels.

Appendix3: conclusion

The credit card dataset study gave us useful information about how customers behave and the things that cause them to leave. Using data analysis tools like boxplots, heatmaps, and association analysis, the study found a number of important factors that affect customer churn. These included the customer's age, income, credit utilisation ratio, transaction behaviour, and length of time with the bank.

Using clustering to divide customers into groups made it possible to find high-risk groups of customers who were likely to leave. A linear regression model was also created to guess the right credit amounts for customers based on their profiles, and it did a pretty good job of guessing.

The study suggests that to successfully stop customers from leaving and improve strategies for keeping them, focused retention efforts should be made for high-risk customer groups, by giving them incentives, benefits, or better services that are tailored to their needs and concerns. It is suggested that keep an eye on our customers' actions and buying habits all the time so that can be proactive and send them personalised messages.

The report also emphasises how important it is to use the developed model to assign credit limits in the best way possible, improve the overall customer experience by making rewards programmes better and customer service more efficient, and make decisions based on data by regularly adding new data to predictive models and updating and refining them.

These suggestions will help the credit card company keep customers instead of losing them. This will also make customers happier, which will increase sales and profits. This will also help make the financial world more competitive and efficient, which will benefit customers.