This report is based on dataset from Bati bank it explain about Credit Risk and Credit Scoring first then it will try to show some EDA

Week-6 Report

Interim

Yodahe Teshome
https://github.com/jodahe1/Bati-Bank.git

Summary Report on Credit Risk and Credit Scoring Approaches

1. Introduction to Credit Risk

Credit risk refers to the potential that a borrower or counterparty will fail to meet its obligations in accordance with agreed terms. This risk is inherent in any financial institution's lending and investment activities and is managed to ensure the institution's stability and profitability.

2. Credit Risk Management Framework

According to the Basel III framework, the objective of credit risk management is to maximize a bank’s risk-adjusted rate of return by maintaining credit risk exposure within acceptable parameters. This involves managing the credit risk in individual credits or transactions, as well as the entire portfolio.

Key elements of a robust credit risk management framework include:

- **Credit Risk Assessment:** Evaluating the creditworthiness of counterparties.
- **Credit Risk Monitoring:** Ongoing tracking of credit exposures.
- **Credit Risk Mitigation:** Utilizing collateral, guarantees, and other measures.
- **Stress Testing:** Assessing the impact of adverse scenarios on credit risk exposure.

3. Credit Scoring and Modeling Approaches

Credit scoring models are essential tools for assessing the credit risk of individual borrowers. These models use statistical techniques to predict the probability of default (PD) based on historical data and borrower characteristics.

Traditional Credit Scoring Models

Traditional models, such as logistic regression, use factors like credit history, income, and employment status to predict default risk. These models are widely used due to their simplicity and interpretability.

Alternative Credit Scoring Models

Emerging approaches leverage alternative data sources, such as social media activity, utility payments, and mobile phone usage. These models employ machine learning techniques to enhance prediction accuracy, especially for individuals with limited credit histories.

4. Developing a Credit Risk Model and Scorecard (Towards Data Science)

A credit risk model and scorecard can be developed using the following steps:

- **Data Collection:** Gather relevant data on borrowers, including traditional financial metrics and alternative data.
- **Data Preprocessing:** Clean and preprocess the data to handle missing values and outliers.
- **Feature Selection:** Identify the most relevant features for predicting default.
- **Model Building:** Use statistical or machine learning models to develop the credit risk model.
- **Model Validation:** Validate the model using out-of-sample testing to ensure its accuracy and robustness.
- **Scorecard Development:** Convert the model output into a scorecard format that can be used for decision-making.

5. Case Studies and Guidelines

World Bank Guidelines

The World Bank provides comprehensive guidelines on credit scoring approaches, emphasizing the importance of using both traditional and alternative data sources. These guidelines highlight best practices in model development, validation, and implementation to ensure the models are fair, accurate, and robust.

Hong Kong Monetary Authority (HKMA)

The HKMA discusses the potential of alternative credit scoring models in financial inclusion. These models can help extend credit to underserved populations by utilizing non-traditional data sources. The HKMA emphasizes the need for regulatory oversight to ensure these models are used responsibly.

Statistica Sinica Study

A study published in Statistica Sinica explores advanced statistical methods for credit risk modeling. The paper highlights the use of complex algorithms and machine learning techniques to improve prediction accuracy and model performance.

6. Conclusion

Effective credit risk management and credit scoring are critical for the stability and profitability of financial institutions. By leveraging traditional methods and embracing alternative data and advanced modeling techniques, institutions

can enhance their credit risk assessment capabilities, improve decision-making, and promote financial inclusion.

7. References

- Towards Data Science: How to Develop a Credit Risk Model and Scorecard
- Corporate Finance Institute: Credit Risk
- Risk Officer: What is Credit Risk?
- World Bank: Credit Scoring Approaches Guidelines
- HKMA: Alternative Credit Scoring
- Statistica Sinica: Credit Risk Modeling Study

Exploratory Data Analysis (EDA)

First let's see the rows and columns we got to understand the data better. As we see on the image, we have 95662 rows and 16 columns.

Metric	Value
Number of rows	95662
Number of columns	16

Then I try to remove non numeric values from the columns mentioned on the image .

Extract Only Numeric Parts From Columns

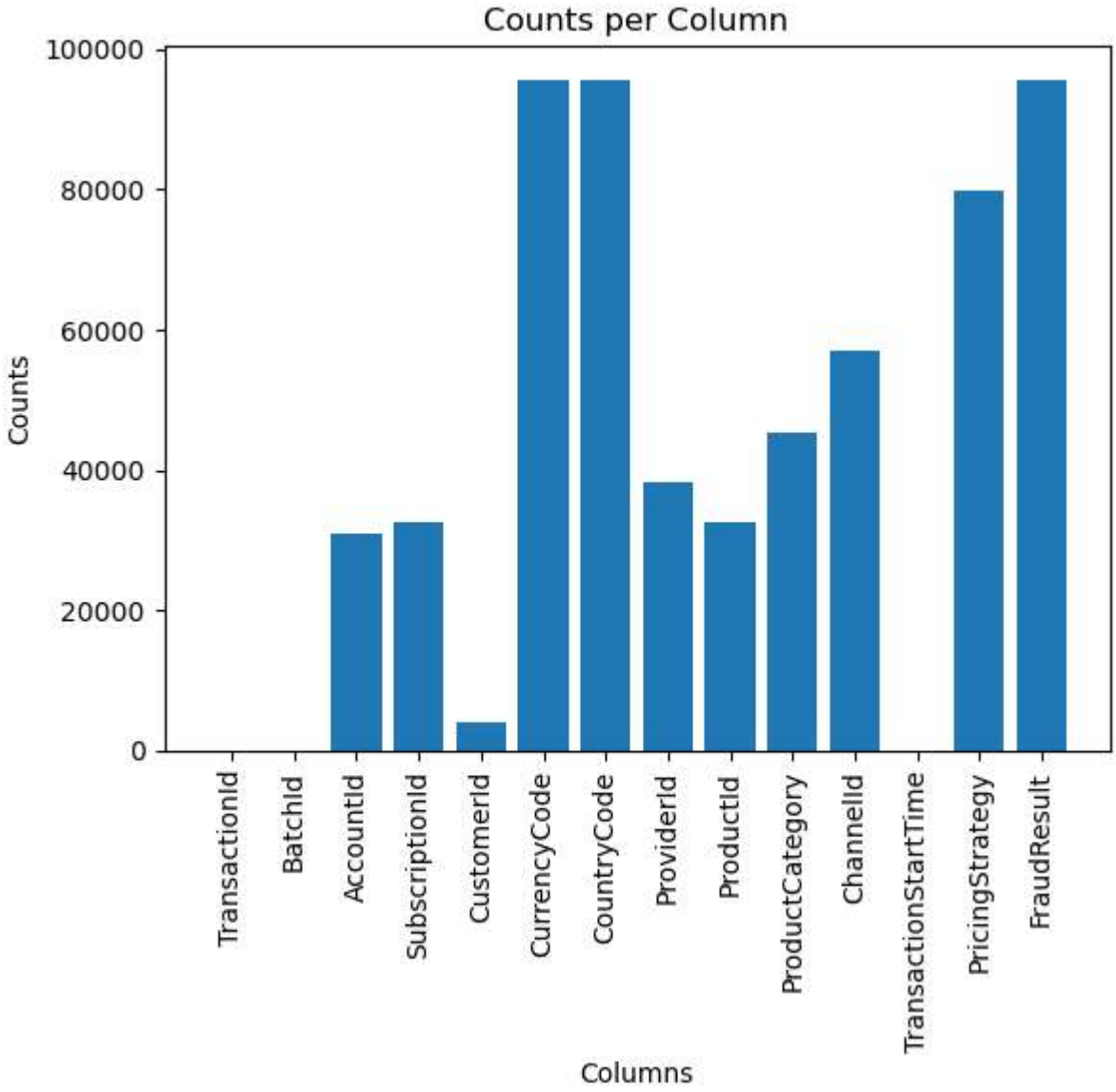
+ Code + Markdown

```
df['TransactionId'] = df['TransactionId'].str.extract('(\d+)') # Extract only numeric part
df['BatchId'] = df['BatchId'].str.extract('(\d+)') # Extract only numeric part
df['AccountId'] = df['AccountId'].str.extract('(\d+)') # Extract only numeric part
df['SubscriptionId'] = df['SubscriptionId'].str.extract('(\d+)') # Extract only numeric part
df['CustomerId'] = df['CustomerId'].str.extract('(\d+)') # Extract only numeric part
df['ProviderId'] = df['ProviderId'].str.extract('(\d+)') # Extract only numeric part
df['ProductId'] = df['ProductId'].str.extract('(\d+)') # Extract only numeric part
df['ChannelId'] = df['ChannelId'].str.extract('(\d+)') # Extract only numeric part
```

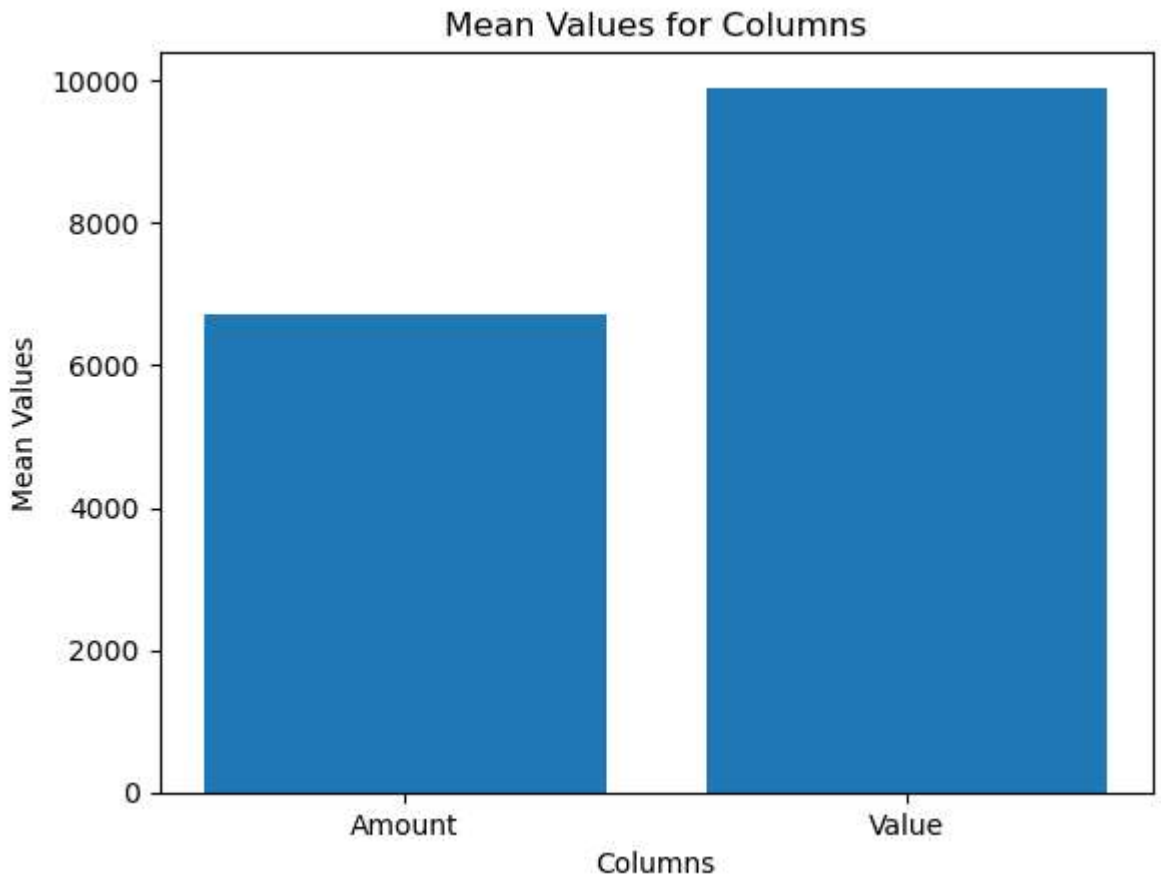
5]

Python

I use bar chart to show most frequent items as we saw in the chart transaction id mode(frequent is null) which tell us it's the unique or primary key for the data set.



The above chart was for non-numeric columns or categorical data to analyze numeric data's I calculate mean.



So, from the above two chart's we can learn: -

TransactionId: There is only one transaction with a TransactionId of 1. This indicates that the TransactionId might be a unique identifier for each transaction, ensuring data integrity and avoiding duplicate entries.

BatchId: There are 28 transactions with a BatchId of 67019. This suggests that these transactions might be grouped together for some specific processing or reporting purposes.

AccountId: There are 30,893 transactions associated with the AccountId 4841. This indicates that AccountId 4841 is quite active and has a significant number of transactions, possibly representing a regular customer or an important account.

SubscriptionId: There are 32,630 transactions with the SubscriptionId 3829. This suggests that SubscriptionId 3829 might correspond to a popular or widely used subscription service.

CustomerId: There are 4,091 transactions associated with the CustomerId 7343. This indicates that CustomerId 7343 might be a relatively active customer, engaging in multiple transactions.

CurrencyCode and **CountryCode**: The currency code UGX and country code 256.0 appear in 95,662 transactions, indicating that these transactions are likely from Uganda, as the currency code UGX represents the Ugandan shilling so we can say most people participate from UGANDA.

ProviderId: There are 38,189 transactions associated with the ProviderId 4. This suggests that ProviderId 4 is a significant provider and is involved in a large number of transactions.

ProductId and ProductCategory: There are 32,635 transactions associated with the ProductId 6 and the product category "financial_services" appears in 45,405 transactions. This indicates that financial services (represented by the product category) and ProductId 6 are popular among customers, potentially indicating a high demand for financial products.

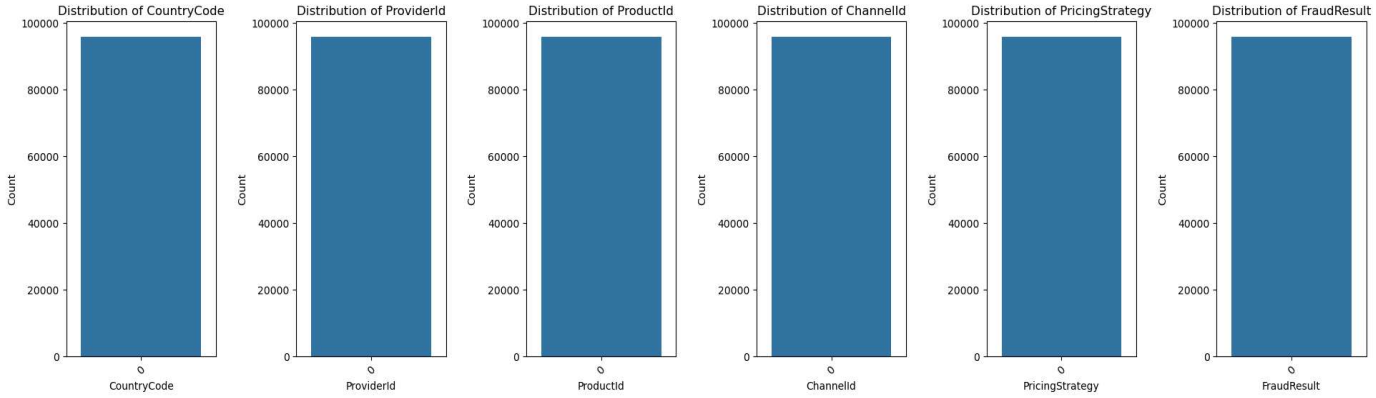
ChannelId: There are 56,935 transactions associated with the ChannelId 3. This suggests that ChannelId 3 is a predominant channel through which transactions are conducted, possibly representing a specific distribution channel or platform.

TransactionStartTime: There are 17 transactions with a TransactionStartTime of 2018-12-24T16:30:13Z. This specific timestamp may indicate a particular event, such as a system update or a batch processing time.

PricingStrategy: There are 79,848 transactions with a PricingStrategy of 2. This implies that PricingStrategy 2 is commonly used, potentially indicating a specific pricing model or strategy.

FraudResult: There are 95,469 transactions classified as non-fraudulent (FraudResult = 0). This suggests that the majority of the transactions in the dataset are classified as non-fraudulent, highlighting the importance of fraud detection and prevention measures.

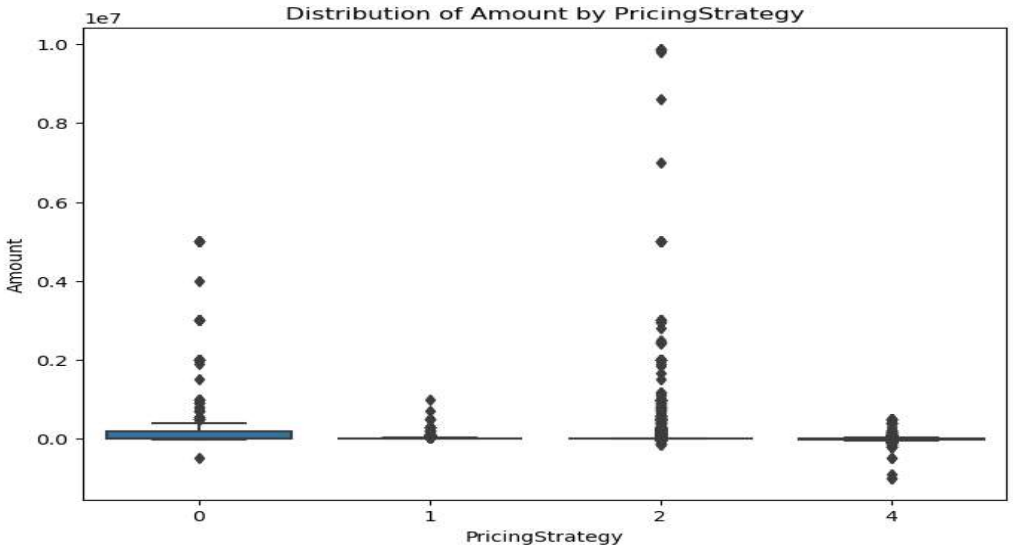
Here another chart to show distribution of categorical data

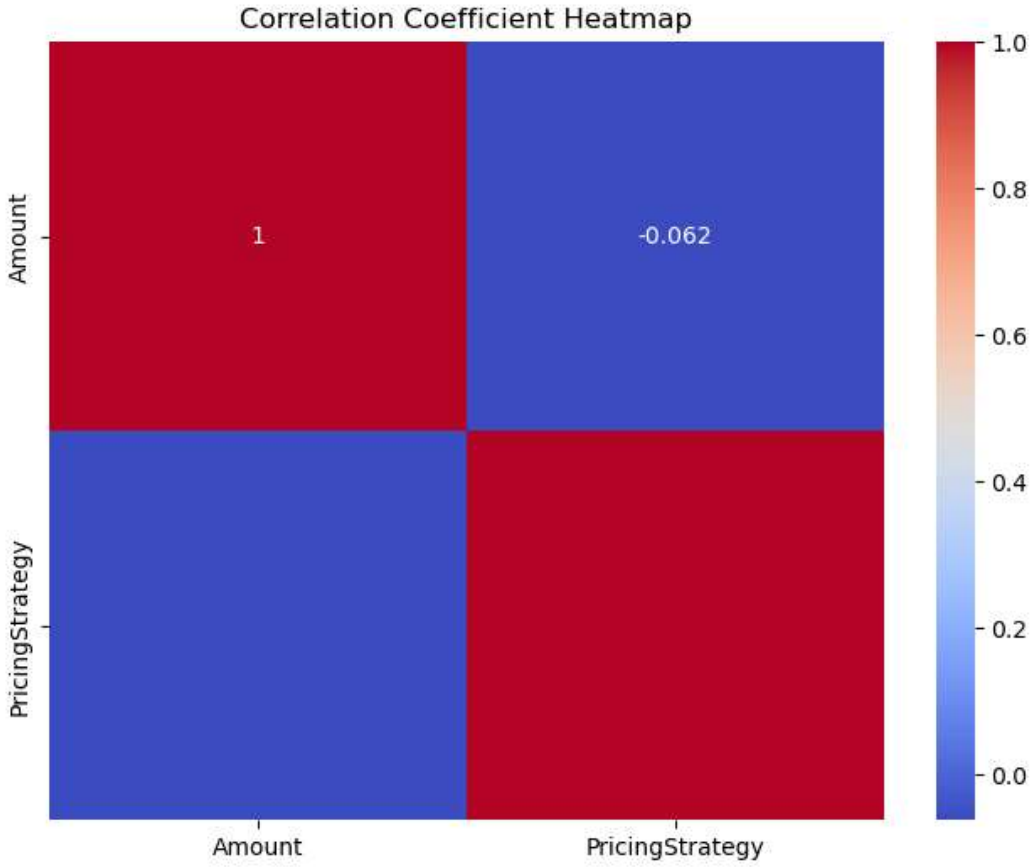


Correlation Analysis

I tried to show correlation among Amount vs Price strategy And Amount vs Fraud Reult.

Amount vs Price strategy





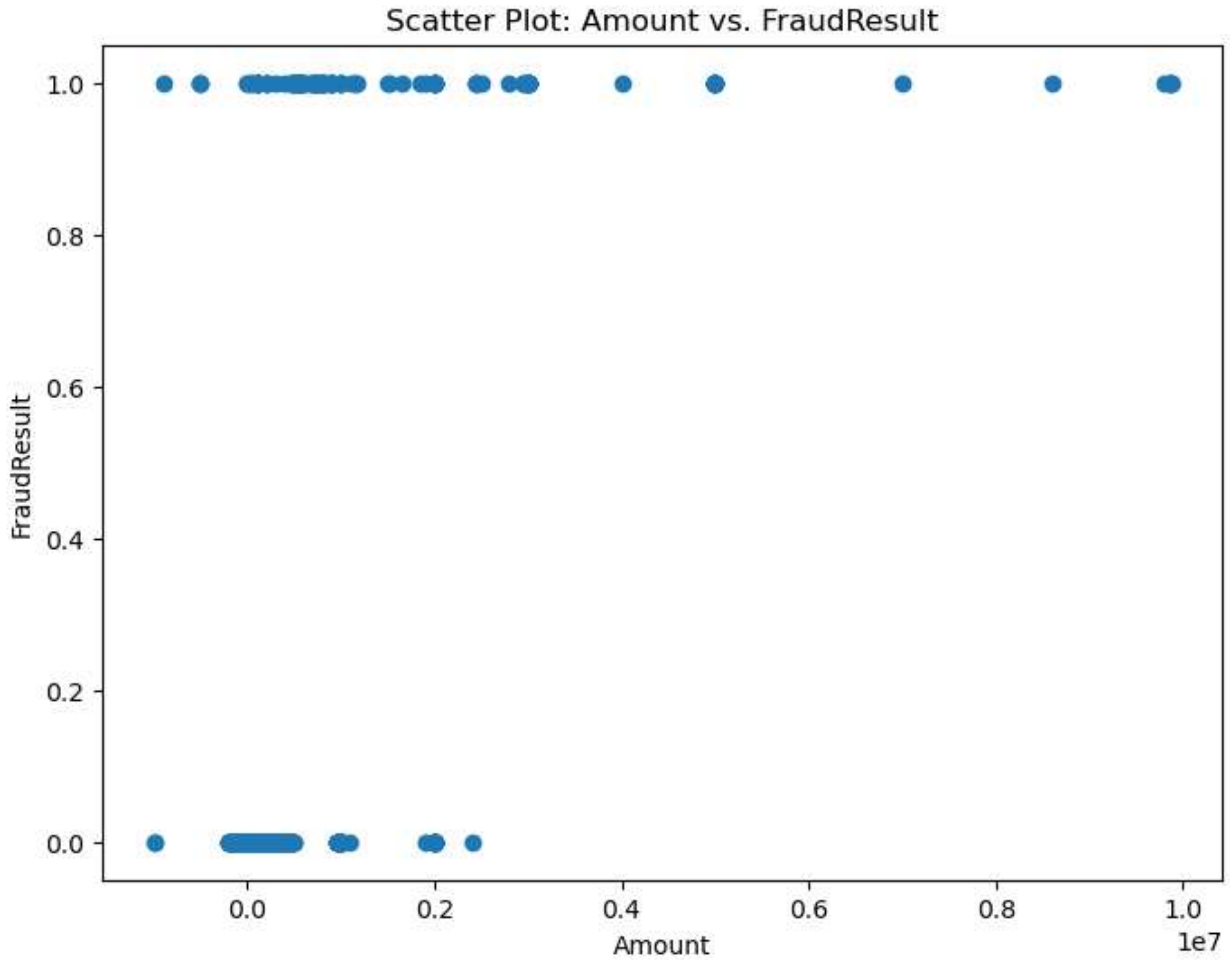
The correlation coefficient of -0.061930792420875715 between the 'Amount' and 'PricingStrategy' from a statistical perspective.

The correlation coefficient measures the strength and direction of the linear relationship between two variables. In this case, the correlation coefficient of -0.0619 suggests a weak negative correlation between 'Amount' and 'PricingStrategy'.

A negative correlation means that as the 'PricingStrategy' increases, the 'Amount' tends to decrease slightly, although the relationship is weak. However, it's essential to note that the correlation coefficient is close to zero, indicating a very weak relationship.

In practical terms, this means that there might be a subtle tendency for lower pricing strategies to have slightly higher transaction amounts.

Amount vs Fraud Result



The correlation coefficient of 0.5573700909352298 between the 'Amount' and 'FraudResult' features, there is a moderate positive relationship between these two variables. This indicates that higher transaction amounts are more likely to be associated with fraudulent transactions

Outlier Detection

I tried to use Box Plot to show outliers