I have completed

- Project structure setup
- .github/workflows/ci.yml created
- Docker and API working
- Placeholder test added

Summary of What You Found

- **Rows/Columns**: 95,662 rows × 16 columns → large and realistic transactional dataset.
- **No missing values** (good news □).
- **Data types**:
    - Most are object → likely categorical IDs or strings.
    - Amount is a float → could contain cents/fractions.
    - Value, CountryCode, PricingStrategy, and FraudResult are integers.

Key Observations from df.describe(include='all')

1. Categorical Features

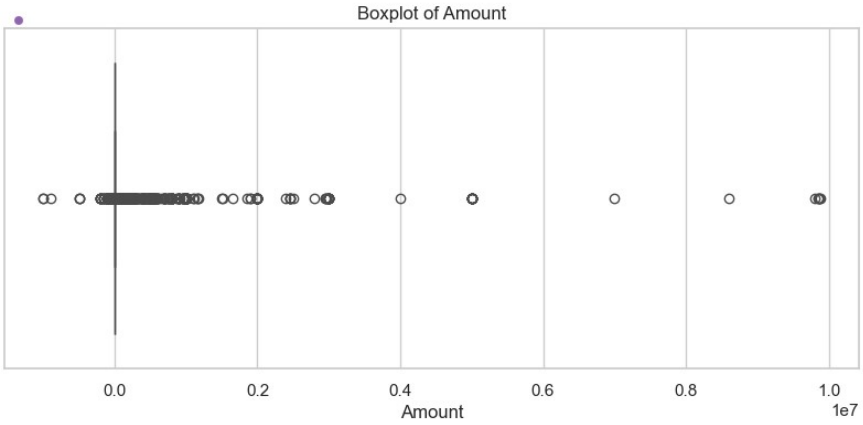
Column	Unique	Top (Most Frequent)	Count	Insight
CurrencyCode	1	UGX	95,662	Single currency → drop it (no predictive value).
CountryCode	1 (256)	N/A	95,662	Constant → drop it.
CustomerId, AccountId, SubscriptionId	3,600+	skewed	High cardinality → might need encoding or grouping later.	
ProviderId, ChannelId, ProductCategory	4–9 unique	Present	Useful for modeling.	
TransactionStartTime	94,556 unique	Skewed	Some duplicated timestamps → maybe not useful directly. Extract time features.	

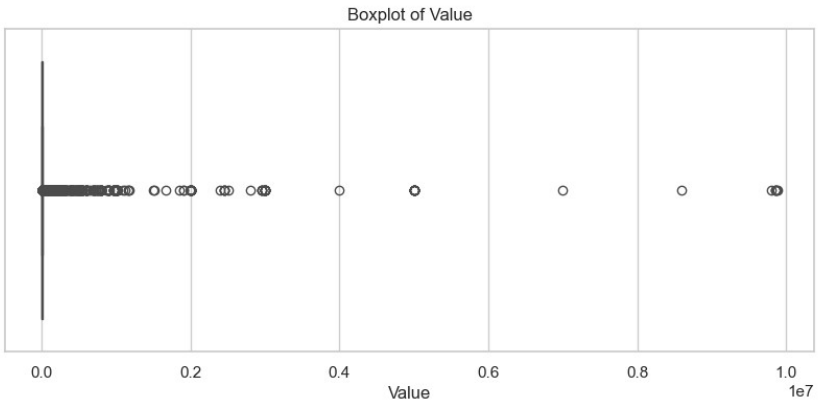
---

2. Numerical Features

Column	Mean	Std Dev	Min	Max	Insight

Column	Mean	Std Dev	Min	Max	Insight
Amount	6,718	123,306	-1,000,000	9,880,000	Extreme outliers. Needs clipping or log-scaling.
Value	9,900	123,122	2	9,880,000	Same as above.
PricingStrategy	Mode = 2	0 to 4	0.73 std dev	Categorical encoded as int. Consider converting to string for clarity.	
FraudResult	0.002 → ≈0.2% fraud	0–1	Highly imbalanced!	Might require stratified split or resampling.	





The **boxplots** show a long line of individual dots, which represent **outliers**.

The boxplots get "compressed" and we lose detail on most of the data, because of the `Amount` and `Value` have **extreme values** (e.g., up to **9.8 million**),

Therefore we make detecting of Outliers Using IQR Method

Outliers in Amount (IQR): 24441

Outliers in Value (IQR): 9021

Detecting the Outliers Using Z-Score Method

Z-score works better on normal (bell-shaped) distributions:

Outliers in Amount (Z-Score): 269

Outliers in Value (Z-Score): 269

Then Cleaned data saved to: C:/Users/ayedr/week-5-credit-risk-model/data/processed/cleaned_data.csv

□ Exploratory Data Analysis (EDA) Summary

Data Overview

- **Total Rows**: 95,662
- **Total Columns**: 16
- All columns are complete (no missing values).
- Dropped constant columns: `CurrencyCode`, `CountryCode`.

Summary Statistics

- **`Amount`** and **`Value`** are skewed with wide ranges (min = -1,000,000, max = 9,880,000).
- Most transactions are non-fraudulent (`FraudResult = 0` for ~99.8%).
- Top categories:
    - `ProductCategory`: `financial_services`, `airtime`, `utility_bill`
    - `ChannelId`: `ChannelId_3` (most common)

Outlier Detection

We used **two methods** to identify outliers in `Amount` and `Value`:

1. **IQR Method**:
    - Outliers in `Amount`: **24,441**
    - Outliers in `Value`: **9,021**
2. **Z-Score Method** (`|z| > 3`):
    - Outliers in `Amount`: **269**
    - Outliers in `Value`: **269**

We chose to **remove outliers using Z-Score** for a conservative approach.

Cleaned Dataset

- Removed extreme outliers using Z-Score flags.
- Dropped non-informative features.
- Saved cleaned data to:

`data/processed/cleaned_data.csv`