Case Study: Building a Financial Data Validation and Analysis Pipeline

Background

You are a data engineer at a financial institution. Each day, transaction data is received from various systems. These records include numeric financial metrics, transaction IDs, dates, and descriptive text (often referencing invoice numbers and payment codes). Data quality issues are common—there are inconsistencies in formats, missing entries, and occasional suspicious or invalid data.

Objectives

After completing this case study, you should be able to:

1. Efficiently process and analyze large numeric datasets using **NumPy**.
2. Manipulate, clean, and extract insights from transactional data using **Pandas**.
3. Validate transaction codes and extract key text features using **regular expressions**.
4. Address data quality issues (missing, inconsistent, suspicious values).
5. Prepare high-quality datasets and summary reports for downstream analytics.

Sample Data Description

The dataset represents daily financial transactions. Main columns include:

Field	**Description**	**Example**
transaction_id	Unique transaction code (should have fixed format)	TXN-20250821-1234
account_id	Customer account number	987654
date	Transaction date	2025-08-21
amount	Transaction amount (float)	2500.75
balance	Account balance after transaction	10000.50
description	Textual description with invoice ref/code	Payment for INV-2025-ABC123

Analysis & Pipeline Engineering Tasks

Participants are invited to address the following:

1. **Data Loading & Exploration**

- Load the transactions data using Pandas.
- Summarize column types, missing values, and basic statistics.

2. **Numerical Analysis with NumPy**
    - Extract the amount column as a NumPy array; calculate mean, standard deviation, min, and max.
    - Compute rolling averages and analyze day-over-day transaction changes.
3. **Pandas DataFrame Manipulation**
    - Convert transaction dates to pandas datetime.
    - Create new features (e.g., day of week).
    - Filter and profile high-value transactions; aggregate totals by account/week/day.
    - Systematically handle missing and inconsistent values for all columns.
4. **Regular Expression Validation & Extraction**
    - Define regex patterns for valid transaction IDs (TXN-YYYYMMDD-XXXX).
    - Validate transaction IDs, flag or correct invalid ones.
    - Extract invoice numbers from free-text description fields.
    - Identify missing or incorrectly formatted invoice numbers.
5. **Data Cleaning & Quality Checks**
    - Detect and flag suspicious transactions (negative amounts, negative balances).
    - Standardize/correct common formatting errors (especially in IDs).
    - Fill missing values for invoices, amounts, balances with suitable defaults or flags.
6. **Reporting & Summary**
    - Generate a report/table showing transaction volume by weekday.
    - Identify accounts or transactions with suspicious / invalid characteristics.
    - Export the cleaned dataset for further analysis or reporting.

Financial Transaction Analysis: Questions for Exploration

1. What are the basic size and structure characteristics of the dataset?
2. How many missing values exist in each column, and which columns have the most missing data?
3. Are all records' transaction_id fields formatted correctly according to the pattern TXN-YYYYMMDD-XXXX? How many are invalid?

4. How many transactions include valid invoice numbers in the description field, and how many are missing or malformed?
5. What is the distribution of transaction amounts? Are there any extreme outliers or suspiciously negative values?
6. How do transaction amounts vary by weekday? Are there certain days with consistently higher or lower transaction volumes?
7. Which accounts have the highest number of transactions or the greatest cumulative transaction amount?
8. Are there accounts or transactions flagged as suspicious due to negative balances or amounts? How many and which ones?
9. What is the average and median transaction amount, and what is the range from minimum to maximum?
10. What percentage of transactions fall into various amount brackets (e.g., <$500, $501-$1000, etc.)?
11. How do daily total transaction amounts trend over time? Are there periods with spikes or drops requiring investigation?
12. What is the frequency distribution of transactions by hour or day part if timestamps are available?
13. Are there relationships or correlations between transaction amount and account balance?
14. Do the descriptions contain recurring or repeated invoice numbers? What are the most frequently appearing invoices?
15. What proportion of transactions have missing or inconsistent data that could impact downstream processing?
16. How many transactions per account deviate significantly from typical metrics (e.g., very large or very small amounts)?
17. What sources or methods (if available) are associated with higher amounts or suspicious transactions?
18. Can you identify patterns in invalid transaction IDs to propose correction rules or validation enhancements?
19. How do rolling average and daily difference statistics on amounts help detect anomalies or trends?
20. Are any accounts disproportionately responsible for suspicious transactions, and what are their characteristics?
