Venmito Data Engineering Project

José A. Megret Bonilla
3/2/2025

Project Overview

Objective: Build a robust data analysis and visualization system for Venmito by transforming raw transactional data into structured insights.

✅ **Data Ingestion & Cleaning** – Process multi-source data into structured format.

✅ **Database Design & SQL Optimization** – Efficient storage, indexing & querying.

✅ **Visualization & Dashboard** – Deliver insights using **Jupyter Notebooks & Streamlit**.

Project Structure

Key Components:

- **/data** → **Raw Data Files** (CSV, JSON, YAML, XML)
- **/dashboard** → **Streamlit App** (For Non-Technical Users)
- **/notebooks** → **Jupyter Notebooks** (For Technical Analysis)
- **/processing** → **Python ETL Scripts** (ETL = extract, transform, load)
- **/db** → SQL **Schema** & **Queries**

Data Sources & Processing

Data Sources:

- Files: JSON, CSV, XML, and YAML (client, transaction, promotion data).
- Database: PostgreSQL for structured storage and querying.

Process:

- Read multi-format files using Pandas.
- Standardize data (normalize names, format dates, remove duplicates).
- Load cleaned data into PostgreSQL tables.

SQL Workflow

Workflow:

1. Extract raw data from CSV, JSON, YAML, and XML.
2. Transform and clean data (normalization, type conversions).
3. Load structured data into PostgreSQL.
4. Generate reports and visualizations using SQL queries.

Data Flow Architecture

Data Flow Overview:

1. **Extract**: Read raw data from multiple file formats (CSV, JSON, XML, YAML).
2. **Transform**: Process & clean using Python & Pandas.
3. **Load**: Store structured data into PostgreSQL.
4. **Analyze**: Query & extract meaningful insights.
5. **Visualize**: Present insights via Streamlit Dashboard & Jupyter Notebook

Technologies Used:

✅ PostgreSQL – Data storage & querying.

✅ Python (Pandas, SQLAlchemy) – Data processing & transformation.

✅ Docker – Containerized environment for deployment.

API & Integration

How Data is Integrated:

- Used SQLAlchemy in Python to interact with PostgreSQL.
- Queried structured data & transformed for visualization.
- Integrated SQL queries directly into Streamlit Dashboard for real-time insights.

API Call Example:

```
engine =
create_engine(DATABASE_URL)
with engine.connect() as conn:
        df = pd.read_sql("SELECT *
FROM transactions", conn)
```

Client Visualization Analyzed

📊 1. Which Clients Have What Type of Promotion?

Summary: This stacked bar chart illustrates the distribution of promotions across the top 10 clients.

- **What it shows:** It categorizes clients by the number and type of promotions they received.The chart helps identify which clients are being targeted the most and whether promotion distribution is balanced.

- **Why it matters:** Businesses can determine if promotions are being allocated effectively to high-value clients. Identifies clients who may be oversaturated with promotions or those who might need better targeting strategies.

📊 2. Analyzing ‘No’ Responses in Promotions

Summary: This bar chart visualizes the percentage of “No” responses per promotion type.

- **What it shows:** It highlights which promotions receive the most rejections from clients. The data is sorted from least to most rejected, making it easier to spot ineffective promotions.
- **Why it matters:** Helps optimize promotional strategies by eliminating or adjusting underperforming promotions. Saves marketing resources by focusing on offers that have a higher acceptance rate.

Client Visualization Analyzed Cont.

📊 3. Promotion Rejections by Contact Method

Summary: This bar chart displays the rejection rate for promotions sent via email, phone, or both.

- **What it shows**: It segments “No” responses based on the contact method used for the promotion. Helps in understanding whether clients prefer certain communication channels over others.
- **Why it matters:** Enables businesses to refine outreach strategies by favoring the most effective communication methods. Can improve customer engagement by reducing unwanted or ineffective outreach.

📊 4. Promotion Effectiveness Analysis

Summary: This bar chart compares response rates across different promotions.

- **What it shows**: Highlights which promotions generate the most engagement. Provides a percentage-based comparison of promotion success.
- **Why it matters**: Ensures marketing efforts are focused on promotions with the highest return. Helps eliminate ineffective promotions that do not engage clients.

Client Visualization Analyzed Cont.

📊 5. Most Profitable Stores

Summary: This bar chart ranks stores by total revenue generated.

- **What it shows:** Identifies the top-performing stores based on their revenue contributions. Highlights disparities between locations, which can be used for benchmarking.
- **Why it matters**: Helps businesses allocate resources more effectively to high-revenue stores. Provides insights for potential expansion or restructuring strategies.

📊 6. Best-Selling Items

Summary: This horizontal bar chart ranks the best-selling items based on the total quantity sold.

- **What it shows**: It provides a clear ranking of items with the highest sales volume. Highlights demand trends by showcasing which products consistently perform well.
- **Why it matters**: Helps optimize inventory management by prioritizing high-demand items. Can be used to adjust marketing efforts towards best-sellers or underperforming products.

Client Visualization Analyzed

📊 7. Client Distribution by Country & City

Summary: A sunburst chart breaking down clients by country and city.

- **What it shows**: Provides a hierarchical view of customer distribution. Shows where the largest customer bases are located.
- **Why it matters**: Essential for geographic expansion planning and localized marketing efforts. Helps tailor promotions and services to high-density customer areas.

📊 8. Device Usage Breakdown

Summary: This pie chart displays the percentage of users on Android, iPhone, or Desktop.

- **What it shows**: Segments customers based on the device they use for transactions. Clearly visualizes whether mobile or desktop platforms dominate user behavior.
- **Why it matters:** Helps optimize digital experiences by prioritizing the most commonly used platforms. Supports app development decisions by indicating which platforms require more attention.

Client Visualization Analyzed

📊 9. Transfer Analysis (Sent vs. Received)

Summary: This bar chart compares the amount of money sent vs. received per client.

- **What it shows**: Displays clients with high transfer activity. Identifies patterns in client transactions, including potential high-volume senders or receivers.
- **Why it matters**: Helps detect unusual transaction behavior that may indicate fraud. Provides insights into customer liquidity and potential upsell opportunities.

Project Demo

Questions?