E-Commerce Data Analyst Project (SQL + Python)

Project Overview

This project demonstrates **end-to-end data analysis** on an e-commerce database using **MySQL, SQL analytics, and Python (Pandas, Matplotlib, Seaborn)**

The analysis answers real-world business questions related to **customers, orders, revenue, products, sellers, retention, and growth trends**.

Tech Stack

- **Database:** MySQL
- **Language:** Python
- **Libraries:**
    - pandas
    - numpy
    - matplotlib
    - seaborn
    - mysql-connector-python
- **Environment:** Jupyter Notebook

Database Schema (Used Tables)

- `customers`
- `orders`
- `order_items`
- `payments`
- `products`
- `sellers`

Dataset resembles a real-world Brazilian e-commerce platform (similar to Olist).

Business Questions & Analysis

1. List all unique cities where customers are located

- Used `DISTINCT` on `customer_city`
- Helps understand **geographic reach**

---

2. Count total orders placed in 2017

- Filtered orders using `YEAR(order_purchase_timestamp)`
- Useful for **yearly performance tracking**

---

3. Total sales per product category

- Joined `products`, `order_items`, and `payments`
- Aggregated revenue using `SUM(payment_value)`
- Identifies **top-performing categories**

---

4. Percentage of orders paid in installments

- Used conditional aggregation
- Business insight into **customer payment behavior**

---

5. Number of customers from each state

- Grouped customers by `customer_state`
- Visualized using bar chart
- Helps in **regional demand analysis**

---

6. Monthly order count in 2018

- Extracted month from timestamp
- Trend analysis using bar plots
- Shows **seasonality and demand patterns**

---

7. Average number of products per order by customer city

- Used **CTE (WITH clause)**

- Shows cities with **bulk or frequent buyers**

---

8. Revenue percentage contribution by each product category

- Calculated category revenue / total revenue
- Identifies **revenue-driving segments**

---

9. Correlation between product price and purchase frequency

- Calculated correlation using NumPy
- Insight: **Weak negative correlation**, indicating price is not the only buying factor

---

. Seller revenue ranking

- Used `DENSE_RANK()` window function
- Ranked sellers based on total revenue
- Helps identify **top-performing sellers**

---

11. Moving average of order values per customer

- Window function with `ROWS BETWEEN`
- Used for **customer spending trend analysis**

---

12. Cumulative sales per month for each year

- Used window functions with `SUM() OVER()`
- Helps track **progressive revenue growth**

---

13. Year-over-Year (YoY) sales growth

- Used `LAG()` window function
- Measures **business growth rate**

---

14. Customer retention rate (within 6 months)

- Compared first and next purchase dates
- Identifies **repeat customer behavior**

---

15. Top 3 customers by spending per year

- Used `DENSE_RANK()` partitioned by year
- Helps identify **high-value customers**

Visualizations

- Bar charts for:
    - Customers by state
    - Orders per month
    - Top sellers revenue
- All plots created using **Matplotlib & Seaborn**

Key Insights

- A small number of states contribute the majority of customers
- Certain product categories dominate total revenue
- Installment payments are widely used
- Seller revenue is highly skewed
- Strong YoY growth observed in 2017–2018
