e-Commerce Store Analysis Target

Isaac Silva
Scalero Data eng. Candidate

1. Introduction

Data Overview

- Database Structure:
  - Connected to MySQL database (**EcommerceDB**)
  - **Tables**: Customers, Geolocation, Orders, Order Items, Payments, Products, Sellers
- Data Cleaning Steps:
  - Checked for missing values, duplicates, and inconsistencies
  - Ensured categorical consistency (e.g., payment types, order status)

Objectives:

- Analyze store performance metrics
- Identify best- and worst-performing stores
- Forecast future sales using machine learning

2.1 ETL Pipeline Proposal for Store Dataset

Proposed ETL Pipeline

1. **Extract**
   - Source: **CSV files / API / Database**
   - Automated data ingestion using **Python (Pandas, SQLAlchemy) or Apache Airflow**
   - Handle missing values and inconsistencies during extraction
2. **Transform**
   - Data Cleaning: Remove duplicates, handle null values, standardize formats
   - Feature Engineering: Compute **Total Revenue, AOV, Unique Customers**
   - Normalize categorical data (e.g., seller locations)
   - Store cleaned data in **a structured format (e.g., DataFrames, SQL tables)**
3. **Load**
   - Load into a **MySQL/PostgreSQL database** for structured querying
   - Optionally, use **Parquet/BigQuery** for large-scale analytics
   - Ensure optimized indexing for fast retrieval

2.1 ETL Pipeline Proposal for Store Dataset

Future Enhancements

✅ Implement **incremental loading** to handle new data

✅ Schedule automated ETL runs using **Apache Airflow**

✅ Leverage **AWS/GCP for cloud-based storage and processing**

LINK TO CODE

2.2 Data cleaning

Missing Values:

- Most tables (customers, geolocation, sellers, products, order_items) have **no missing values** (
- **Orders table** has some missing values for the following columns:
  - order_approved_at: 23 missing values.
  - order_delivered_carrier_date: 236 missing values.
  - order_delivered_customer_date: 364 missing values.

Action: Review why these columns are missing values (e.g., incomplete order records) and whether they should be handled (e.g., filling missing values, removing rows, or leaving them as is based on business requirements).

2.2 Data cleaning

Duplicates:

- There are **no duplicate rows** in the dataset.

Action: No further steps are required here.

Inconsistencies: We checked just for 2 table columns as an example.

1. For the **payment_type** in the payments table, there are **no inconsistencies** (i.e., all values are part of the valid payment types list: credit_card, UPI, voucher, debit_card, not_defined).
2. For **order_status** in the orders table, there are **no inconsistencies**, meaning all statuses match the expected list: delivered, processing, shipped, invoiced, canceled, unavailable.

Action: We need to define which table columns to check for inconsistencies like **product.product_category**

LINK TO CODE

3. Store Performance Analysis

1

3.1 Total Sales Per Store

- Ranking of stores by revenue
- Insights into regional performance (city/state)

2

3.2 Top 10 Best-Performing Stores

- High revenue stores
- Factors influencing success

3

3.3 Bottom 10 Performing Stores

- Struggling stores
- Potential causes: low demand, high shipping costs, poor visibility

4

3.4 Customer Distribution Per Store

- Number of unique customers per store
- Relation between customer base and sales performance

5

3.5 Average Order Value (AOV) per Store

- Higher AOV suggests premium pricing or bundling
- Identified stores with the highest AOV

LINK TO CODE

1. Top-Performing Sellers (3.1 & 3.2)

1. The seller in **Mogi Guaçu, SP (Seller ID: 0)** is the top performer, generating the highest **total revenue (R$ 4,969,925.88)** and **total orders (36,224)**.
2. Other high-performing cities include **Tubarao, Tabatinga, Penapolis, and São Paulo**, with significant sales volumes but lower revenue compared to Mogi Guaçu.
3. **São Paulo sellers have the highest AOV (Avg Order Value), reaching R$ 171.14**, indicating that fewer but higher-value orders are placed in this city.

2. Low-Performing Sellers (3.3)

1. Some sellers have **only 1 order with extremely low revenue**, such as in **Gaspar (R$ 10.90), Sorocaba (R$ 8.49), and São Paulo (R$ 6.50)**.
2. These could be new sellers, inactive sellers, or those selling low-cost items.

3. Customer Base & Market Reach (3.4)

1. The **largest unique customer base belongs to Mogi Guaçu (1,203 customers)**, reinforcing its status as the leading seller in terms of both revenue and orders.
2. **Other key cities include Tubarao (292), Penapolis (265), and Tabatinga (252),** showing strong customer engagement

LINK TO CODE

4. High AOV Sellers (3.5)

1. Some sellers, despite having a **low total order count, have an exceptionally high AOV**, such as:
2. **São Paulo (Seller ID: 39,000,000) – AOV: R$ 2,585.00**
3. **Rio de Janeiro (Seller ID: 961) – AOV: R$ 2,486.50**
4. **Limeira (Seller ID: 59,417) – AOV: R$ 2,025.50**
5. These sellers may specialize in high-value products, such as luxury items or bulk sales.

Key Takeaways:

✅ **Mogi Guaçu** dominates in both revenue and orders, making it a critical market for sellers.

✅ **São Paulo** has fewer orders but the highest AOV, indicating high-ticket sales.

✅ **Certain sellers** generate minimal revenue, suggesting opportunities for improvement or re-evaluation of their business model.

✅ **AOV analysis** highlights sellers focusing on premium products, which might require different marketing strategies compared to high-volume, low-ticket sellers.

LINK TO CODE

PROPHET

Sales Forecasting

Prophet was chosen because it's **easy to implement, interpretable, and robust** against missing values and seasonality.

Other options:

- ➔ If data had stronger short-term autocorrelation, **ARIMA** or **SARIMA** could be better.
- ➔ For deep learning-based insights, **LSTM** could be explored.
- ➔ If lot of extra data is available (e.g., promotions, competitor activity), **XGBoost** could be powerful.

LINK TO CODE

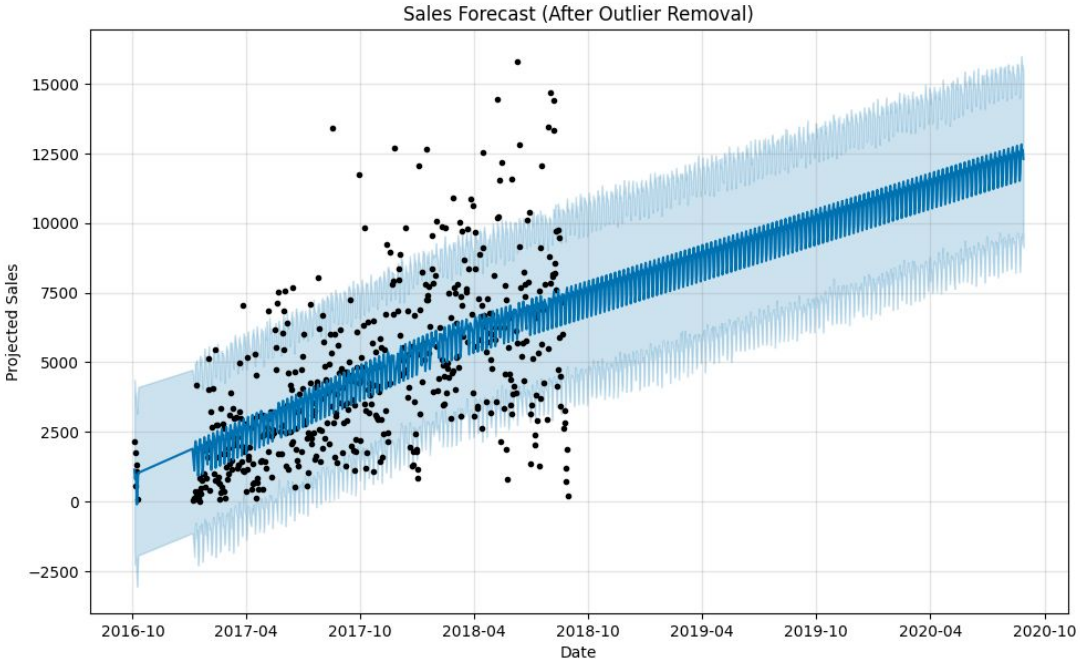
Sales Forecasting

Steps taken:

- Aggregated daily sales data
- Handled outliers using the IQR method
- Forecasted sales for 2 years

Forecasted Monthly & Yearly Sales

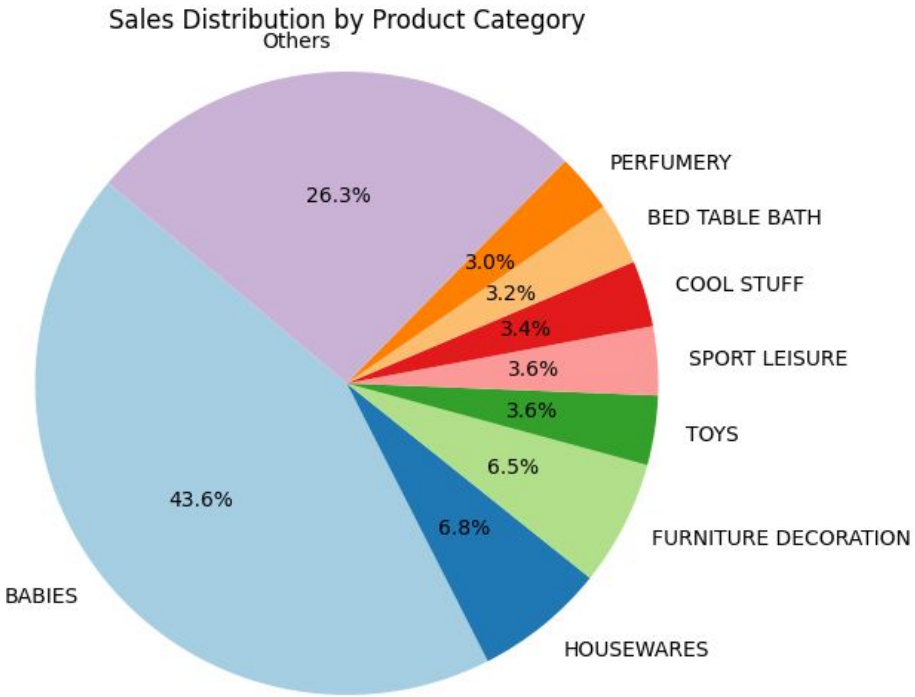
- Expected revenue trends
- Seasonal variations & growth insights

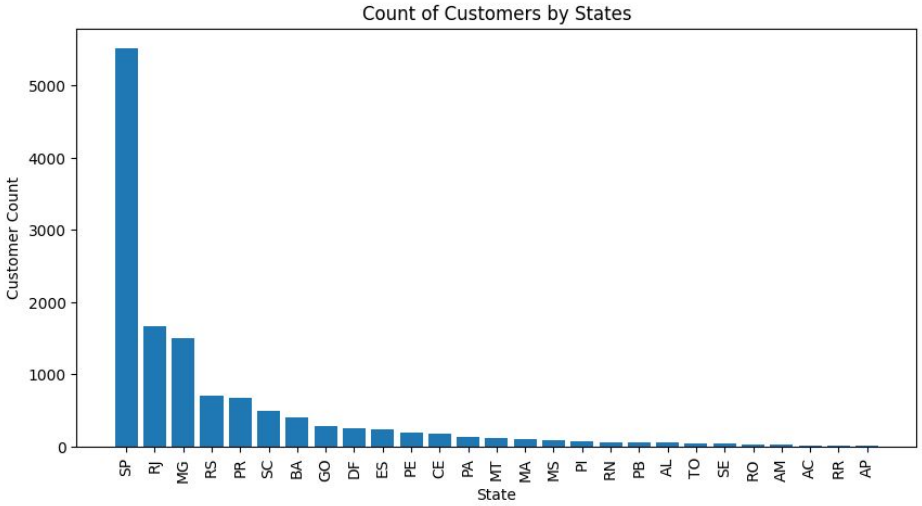


Year	Predicted_Sales
0 2019	3,407,661
1 2020	2,765,862

LINK TO CODE

Additional Insights





⏪ **Sales by product**

Customer Count ⏩

State	Customer Count
SP	5511
RJ	1673
MG	1502
RS	701
PR	682

Thank you.