OLIST E-COMMERCE DATA ANALYSIS PROJECT
TOOL USED—Power Bi

PHASE A: DATA CLEANING & PREPARATION

OBJECTIVE

To ensure data accuracy, consistency, and reliability before performing exploratory and business analysis.

DATA IMPORT & INITIAL ASSESSMENT

All Olist dataset tables were imported into Power BI. An initial review was conducted to understand table structure, key fields, and data types. Each table was inspected for missing values, duplicates, and inconsistencies.

HANDLING MISSING VALUES

- Orders without delivery dates were identified and excluded from delivery performance analysis, as delivery time could not be calculated.
- Missing review scores were retained but excluded from review-related metrics to avoid skewing averages.
- Null values in optional fields (e.g., seller or product attributes) were left unchanged when they did not affect core metrics.

WHY:

Removing or excluding incomplete delivery records prevents inaccurate operational metrics.

REMOVING DUPLICATES

Duplicate records were checked using primary identifiers such as

- Order_id
- Customer_unique_id
- Seller_id

No duplicate order IDs were found in the orders table.

Distinct counts were used in measures to prevent duplication across related tables.

WHY:

Ensures that orders, customers, and sellers are not double-counted in aggregations.

FIXING DATA INCONSISTENCIES

- All date-time columns were standardized to a consistent format.
- Categorical text fields (e.g., product categories) were normalized using the category translation table.
- Numeric fields were validated to ensure correct data types (e.g., prices, payment values, review scores).

WHY:

Standardized formats improve accuracy in time-based analysis and category comparisons.

DATA MODELING & RELATIONSHIPS

A star-schema-style data model was created with olist_orders_dataset as the central fact table

Key Relationships Established:

- Orders → Customers (customer_id)
- Orders → Order Items (order_id)
- Order Items → Products (product_id)
- Orders → Payments (order_id)
- Orders → Reviews (order_id)
- Order Items → Sellers (seller_id)

All relationships were configured as one-to-many, with single-directional filtering to prevent ambiguity and double counting.

WHY:

A clean data model ensures accurate aggregation and performance-efficient analysis.

FEATURE ENGINEERING

New calculated columns and measures were created to support business analysis:

- Delivery Time (Days): Difference between order purchase date and delivery date.
- Delivery Delay (Days): Difference between estimated and actual delivery dates.
- Late Delivery Rate: Percentage of orders delivered after the estimated date.
- Total Revenue: Sum of actual payment values.
- Average Order Value (AOV): Total revenue divided by total orders.
- Average Spend per Customer: Total spend divided by unique customers.

WHY:

Derived metrics enable deeper operational, financial, and customer behavior insights.

INSIGHTS

Data quality issues were primarily related to missing delivery and review data

Proper modeling was essential to prevent overcounting revenue and orders.

RECOMMENDATIONS

- Enforce stricter data validation at data capture (especially delivery dates)
- Monitor missing review data to improve feedback collection
- Maintain relational integrity to support scalable analytics

PHASE B: EXPLORATORY DATA ANALYSIS(EDA)

BUSINESS PERFORMANCE

Insights

- Revenue and order volume show clear seasonality.
- A small number of product categories drive a large portion of total revenue.
- AOV remains relatively stable over time.

Why It Matters

- Seasonal demand affects inventory planning and logistics.
- Revenue concentration increases business risk.

Recommendations

- Plan promotions and logistics capacity ahead of peak seasons.
- Diversify revenue streams by promoting mid-tier categories.

CUSTOMER ANALYSIS

Insights

- Customers are heavily concentrated in a few states.
- One-time customers dominate the customer base.
- Repeat customers have higher average spend.

Why It Matters

- Retention is more cost-effective than acquisition.
- Geographic concentration highlights regional growth opportunities.

Recommendations

- Introduce loyalty and retention programs.
- Focus marketing efforts on high-value regions.
- Target repeat customers with personalized offers.

PRODUCT ANALYSIS

Insights

- Sales are dominated by a small group of top-selling products.
- Certain categories consistently generate higher revenue.
- Price distribution is skewed toward lower-priced products.

Why It Matters

- Product dependency increases revenue volatility.
- Pricing strategy affects demand and profitability.

Recommendations

- Promote mid-performing products to balance revenue.
- Bundle products or introduce dynamic pricing strategies.
- Optimize inventory for high-demand categories.

SELLER ANALYSIS

Insights

- A small percentage of sellers generate most orders and revenue.
- Seller performance varies significantly by region.
- Seller activity fluctuates over time.

Why It Matters

- Over-dependence on top sellers poses operational risk.
- Seller churn can affect order fulfillment.

Recommendations

- Implement seller performance monitoring.
- Support underperforming sellers with training or incentives.
- Encourage seller retention through better onboarding and SLAs.

OPERATIONAL METRICS

Insights

- Most deliveries occur within acceptable time frames, but delays exist.
- Late deliveries are higher in specific product categories and seller regions.
- Freight cost increases with distance but not always proportionally.

Why It Matters

- Delivery delays directly impact customer satisfaction.
- Inefficient logistics increase operational costs.

Recommendations

- Optimize logistics routes and warehouse allocation.
- Enforce delivery SLAs for high-delay sellers.
- Use data to renegotiate freight contracts.

REVIEW ANALYSIS

Insights

- Average review score is moderately high.
- Review scores decrease as delivery time increases.
- Certain product categories consistently receive lower ratings.

Why It Matters

- Delivery reliability strongly influences customer satisfaction.
- Poor reviews affect repeat purchases and brand perception.

Recommendations

- Prioritize delivery performance improvements.
- Proactively communicate delays to customers.
- Improve quality control in poorly reviewed categories.

PHASE C: RECOMMENDATIONS & DASHBOARD DELIVERY

OBJECTIVE

To translate analytical insights into actionable business recommendations and deliver an interactive dashboard that enables stakeholders to monitor performance and support decision-making.

INSIGHT SYNTHESIS

Insights generated during Phase B were consolidated across key business dimensions revenue, customers, products, sellers, operations, and reviews. Patterns and relationships identified during analysis were reviewed to determine their strategic impact on growth, customer satisfaction, and operational efficiency.

Outcome:

A clear understanding of the key drivers and constraints affecting marketplace performance.

BUSINESS RECOMMENDATIONS

Revenue & Growth

Recommendation: Leverage seasonal demand patterns by planning promotions, inventory, and logistics capacity ahead of peak periods.

Why : Revenue and order volumes show strong seasonality, indicating predictable demand cycles.

Customer Retention

Recommendation: Implement loyalty programs, targeted discounts, and post-purchase engagement to convert one-time buyers into repeat customers.

Why : Repeat customers demonstrate higher average spend and long-term value.

Product Strategy

Recommendation: Promote mid-performing products and expand high-demand categories to reduce reliance on a small number of top-selling product.

Why : Revenue concentration increases risk and limits scalability.

Seller Performance

Recommendation: Introduce seller performance monitoring and support programs to improve delivery reliability and revenue contribution among mid-tier sellers.

Why : A small group of sellers dominates revenue, creating dependency risk.

Operational Efficiency

Recommendation: Identify high-delay categories and seller locations and optimize logistics routes or enforce stricter delivery SLAs.
Why : Late deliveries are a major contributor to poor customer reviews and dissatisfaction.

Customer Experience
Recommendation: Improve delivery communication and proactive delay notifications to manage customer expectations.
Why : Review scores decline significantly as delivery delays increase.

DASHBOARD DESIGN & DELIVERY

An interactive Power BI dashboard was developed to present key metrics and insights in a clear and intuitive format.
Dashboard Components:

- Executive KPIs: Total Revenue, Total Orders, AOV, Average Spend per Customer
- Trends: Monthly Revenue and Order Volume
- Customer Insights: Geographic distribution, repeat vs one-time customers
- Product & Seller Performance: Top products, revenue by category, top sellers
- Operations: Delivery time distribution, late delivery rates
- Customer Feedback: Average review score and delivery impact

Filters and slicers were added for time period, product category, seller location, and customer region.
Visuals were created for all six EDA categories.
Metrics were designed to update dynamically based on user selections.

INSIGHT FROM VISUALIZATION:
Stakeholders can explore insights independently and drill down into specific areas of interest.
Relationships between delivery delays and reviews are visually clear.
Regional and category based performance differences are easily explored.

FINAL CONCLUSION

This project transformed analytical findings into strategic recommendations and delivered a decision-ready dashboard. The final output enables stakeholders to monitor performance, identify issues, and take informed actions to improve revenue growth, customer retention, and operational efficiency.