E-commerce Data Analysis Report

This report summarizes the data analysis performed on an e-commerce dataset, covering basic, intermediate, and advanced business questions. The analysis uses MySQL for data querying and manipulation, and Python (Pandas, Matplotlib, Seaborn) for data presentation and visualization.

Basic Queries :

Customer Geographic Distribution

The customer base shows a high degree of geographic diversity and a strong concentration in a single state.

- The total number of **unique cities** where customers are located is **4,119**.
- When grouping customers by state, the state with the highest number of customers is **SP (São Paulo)**, with **41,746** customers.

Order Volume and Trends

Order volume in 2017 provides a measure of annual activity.

- The total number of orders placed in the year **2017** was **45,101**.

Sales and Payment Analysis

Sales by Product Category

Total sales revenue was aggregated and calculated for all 74 product categories.

- The category generating the **highest total sales** is **BED TABLE BATH**, with revenue of **$1,712,553.67**.
- Other high-revenue categories include **FURNITURE DECORATION** ($1,430,176.39) and **AUTOMOTIVE** ($852,294.33).
- Categories with the lowest sales include **FASHION CHILDREN'S CLOTHING** ($785.67) and **INSURANCE AND SERVICES** ($324.51).

Payment Method

The preference for instalment payments is nearly universal in the dataset.

- The percentage of orders that were paid using one or more instalments (payment_installments >= 1) is **99.9981%**. This suggests that instalment payment options are the default or heavily preferred method for transactions.

Intermediate Queries :

Sales and Product Metrics

1. Orders Per Month in 2018

The total number of orders placed was calculated for each month in 2018.

- The month with the **highest order count** was **August**, with **6,608** orders.
- The data covers orders placed from **January** through **October**.

- Order counts generally trended upward through the summer months, a trend easily observable in the generated bar chart.

2. Average Products Per Order by City

The analysis determined the average number of products contained in each order, grouped by the customer's city.

- This was achieved by first creating a **Common Table Expression (CTE)** to count the items in every individual order, and then calculating the average of these counts for each city.
- For the city of **Sao Paulo**, the average was **1.16 products per order**.
- Some smaller cities, like **buriti**, showed a much higher average of **3.00 products per order**.

Revenue and Correlation Analysis

3. Percentage of Total Revenue by Product Category

The contribution of each product category to the total overall revenue was calculated.

- The top-contributing categories are:
  - **BED TABLE BATH** with **10.70%**.
  - **HEALTH BEAUTY** with **10.35%**.
  - **COMPUTER ACCESSORIES** with **9.90%**.
- The least contributing categories, at **0.00%**, are **FASHION CHILDREN'S CLOTHING** and **INSURANCE AND SERVICES**.

4. Correlation: Price vs. Purchase Frequency

A correlation analysis was performed to identify the relationship between the average price of a product category and the total number of times products in that category were purchased.

- The calculated **correlation coefficient is -0.106** (rounded).
- This indicates a **weak negative correlation**, suggesting that as the average price for a product category increases, the number of times it's purchased slightly decreases.

5. Total Revenue and Rank by Seller

The total revenue generated by each seller was calculated, and sellers were then ranked by this metric.

- The analysis used the **DENSE_RANK()** window function on the total revenue grouped by seller_id to determine the ranking.
- The **Top-Ranked Seller** (Rank 1) generated **$227,092.36** in total revenue.
- The total revenue for the top 5 sellers is presented in the bar chart, highlighting the concentration of sales among the leading sellers.

Advance Queries :

Time-Series and Trend Analysis

1. Moving Average of Order Values

The **moving average** of order payment values was calculated for each customer over their order history.

- The analysis used a **window function** (AVG() OVER(...)) that partitioned the data by customer_id and ordered it by order_purchase_timestamp.
- The window size was set to the **current row and the 2 preceding rows** (a 3-order moving average).
- This metric is used to **smooth out short-term fluctuations** in individual customer spending to identify underlying trends in their monetary value over time. The output shows the calculated moving average for each sequential order by a customer.

2. Cumulative Sales Per Month

Cumulative sales were calculated for each month across all years in the dataset.

- This was achieved using the **SUM() OVER(ORDER BY year, month)** window function.
- The running total sales show continuous growth, reaching a cumulative total of **$16,008,872.12** by October 2018.
- For example, total cumulative sales grew from **$7,309,109.07** at the end of **2017** to **$8,424,113.25** by **January 2018**.

3. Year-over-Year (YoY) Growth Rate of Total Sales

The YoY percentage growth rate of total sales was calculated to assess the business's annual revenue performance.

- The calculation used the **LAG()** window function to retrieve the sales from the previous year for the growth rate formula.
- **2017 Growth:** Sales in 2017 showed a substantial growth of **12,112.70%** over the preceding partial year (2016). This large figure is likely due to the limited number of orders in 2016.
- **2018 Growth:** Sales in 2018 grew by **20.00%** compared to 2017.

Customer Performance Metrics

4. Customer Retention Rate

The **customer retention rate** was defined and calculated as the percentage of customers who make **another purchase within 6 months** of their first purchase.

- The query used **Common Table Expressions (CTEs)** and date arithmetic (DATE_ADD(first_order, INTERVAL 6 MONTH)) to find customers who placed a second order within the defined 6-month window.
- The query result was **[(None,)]**. This indicates that the specific retention calculation, based on the customer base in the available data, yielded a null or zero result, suggesting **very few or no customers** met the criteria of making a subsequent purchase within 6 months of their initial order.

5. Top 3 Customers by Annual Spending

The **top 3 customers** who spent the most money (total payment value) in **each year** were identified and ranked.

- The analysis employed the **DENSE_RANK()** window function, partitioned by year and ordered by total payment descending, to assign ranks.
- The visualization highlights the spending of these top customers across the years.

Summarized E-commerce Data Analysis & Business Intelligence Report

Core Business Profile & Distribution (Basic Findings)

The initial analysis established the basic parameters of the customer base and sales volume:

- **Geographic Diversity:** The customer base is highly diverse, spanning **4,119 unique cities**.
- **Customer Concentration:** Despite the wide reach, customers are heavily concentrated in the state of **SP (São Paulo)**, which accounts for **41,746** customers.
- **Order Volume:** A total of **45,101 orders** were placed in **2017**.
- **Top Sales Category:** The **BED TABLE BATH** category leads in total sales revenue, generating **$1,712,553.67**.
- **Payment Method:** Instalment payments are the overwhelming preference, used in nearly **99.9981%** of all orders.

Intermediate Performance Metrics

These findings provide granular detail on monthly performance, product analysis, and seller ranking:

- **Monthly Order Trends (2018):** In 2018, the month with the **highest order count** was **August**, with **6,608** orders.

- **Product Consumption:** The average number of products per order is relatively low, though some smaller cities like **buriti** show higher averages (**3.00**).
- **Revenue Contribution:** The top three categories contributing to total revenue are **BED TABLE BATH (10.70%)**, **HEALTH BEAUTY (10.35%)**, and **COMPUTER ACCESSORIES (9.90%)**.
- **Price-Purchase Correlation:** There is a **weak negative correlation** (≈ -0.106) between a product category's average price and its purchase frequency.
- **Top Seller:** The top-ranked seller generated **$227,092.36** in revenue.

Advanced Business Intelligence

The advanced analysis revealed critical growth and customer lifetime value metrics:

- **Cumulative Sales:** Cumulative sales show continuous growth, reaching a total of over **$16 million** by October 2018.
- **Year-over-Year (YoY) Growth:** The YoY sales growth rate from **2017 to 2018 was 20.00%**.
- **Moving Average:** A **3-order moving average** of customer spending was calculated to identify stabilized trends in individual customer monetary value, smoothing out transactional volatility.
- **Top Spenders:** The **top 3 customers** who spent the most money in each year were identified using a dense ranking function.
- **Customer Retention:** The retention rate, defined as the percentage of customers who make a second purchase within 6 months of their first order, was found to be effectively **zero** ([(None,)]). This suggests a significant opportunity for strategies focused on increasing repeat purchases and customer loyalty.