**Olist Store Sales Analysis**

**Report**

**Prepared by:**

* **Mohamed Abdeltawab Mohamed Hassan (Data Cleaning & preprocessing, Presentation)**
* **Abdulaziz Suleiman Abdulaziz Suleiman (Dashboard, Insights)**
* **Ahmed Gamal Muhammad Alazhary (Analysis Questions, Insights)**
* **Doaa Khaled Abdelwahab (Analysis Questions, Insights)**
* **Abdulkareem ramadan Ali (Data Cleaning & preprocessing, presentation)**
* **Omar Ayman Refaat (Dashboard, Insights)**

**And Under Supervision of :**

* **Ministry of Communication and information Technology (DEPI)**
* **AST (Trainer : Mennat-allah Selim , Manager : Mansour El-Husseiny)**

1. **Overview:**

Olist is a Brazilian e-commerce platform founded in 2015, aiming to empower small and medium-sized businesses to sell their products online.

Olist provides integrated solutions including inventory management, logistics, and customer service, making it easier for sellers to showcase their products on major e-commerce websites such as Mercado Livre, Amazon, and Americanas12.

1. **Objective:**

The purpose of this analysis is to provide actionable insights into e-commerce performance over a specified time frame (September 2016 to October 2018).

**The key questions this analysis seeks to answer are:**

* Overall sales trend over time.
* Top- and worst-performing product categories.
* Regional revenue contributions.
* Customer behavior patterns related to payment preferences, shipping times, and order statuses.

1. **Data Cleaning Process: (Mohamed Abdeltawab and Abdulkareem Ramadan)**

**Tool: Python, Power Query**

**1. Initial Data Inspection**:

* + Load dataset and inspect data types, column names, and structure.
  + Use info(), describe(), and head() functions for basic data overview.

\*\* in Products table🡪 we decide to drop (Product name length, product length, height, width) as no need for this data in our proposed analysis

\*\*in Customer table🡪 we drop customer unique ID column, as already ID presents.

\*\*in order reviews table 🡪 we drop comment title, comment message columns, as no need in the final model.

\*\*in order Items table 🡪 we categorize the products to only 24 sections (categories) to align with the aim of analysis. Using power Query

\*\*create List of factitious names to Sellers and merged to the original table to easily identify and represent the sellers in Dashboard.

**2. Handling Missing Values**:

* + Identify columns with missing values using isnull().sum().
  + Investigate causes of missing data (e.g., order\_status influences nulls in order\_approved\_at, order\_delivered\_carrier\_date, and order\_delivered\_customer\_date).

**Reason for Retaining Nulls**: In cases where order\_status is "cancelled", it's expected that delivery-related timestamps would be null, as the order was never shipped or delivered. These null values are intrinsic to the data's nature and do not need to be imputed or removed.

* + Decide on appropriate actions: drop, fill, or retain null values based on their significance.

\*\*in products table 🡪 we decide to keep the 610 NUILLS values in Product category name AND replaced it by Others using Python.

\*\*we replace the NULLS in Product weight with Median as not normally distributed data.

**3. Duplicates Removal**:

* + Check for duplicates using duplicated() and identify the extent of duplication.
  + Remove duplicates to ensure accurate analysis (e.g., 261,831 duplicates found in df\_geolocationdataset).

**4. Data Type Conversion**:

* + Convert columns to suitable data types (e.g., datetime for order timestamps).
  + Ensure numeric, categorical, and text data are correctly formatted for analysis.

**5. Data Consistency Checks**:

* + Standardize text fields (e.g., product categories).
  + Align column formats across datasets to allow seamless merging and analysis.

**6. Final Data Validation**:

* + Recheck for any remaining issues post-cleaning.
  + Save the cleaned dataset for further analysis.

**Analysis and Insights: (Doaa Khaled and Ahmed El Azhary )**

**Tool: SQL**

* A table with numbers and numbers

  Description automatically generated**Sales Performance:**

1. **What is the overall sales trend (revenue) over time?**

We found that during the period from September 2016 to October 2018, the revenues fluctuated significantly, peaking at **$1,194,882.80** in November 2017. There was a significant growth in annual revenues from **$59,362.34** in 2016 to **$8,699,763.00** in 2018. However, a sharp decline was observed in September and October 2018.

**A screenshot of a computer

Description automatically generated**

1. **What are the top-selling product categories? (Top 5)**

We found that the **furniture\_decor** category was the highest selling category, with total sales of **$481,915,077.43.** The **sports** category was the fifth best-selling category, with total sales of **$138,332,663.99.**

**A screenshot of a computer

Description automatically generated**

1. **What are the worst-selling categories?**

"Kitchen" was the worst-selling category, grossing **$224,124,210.**

1. **Who are the top 10 sellers based on sales?**

**Donald Evans** topped the list with the highest sales of **$507,166.91**. Sales varied widely among the top ten sellers.

And the last one in the list is **Paul Garcia with $185,134**

1. **Which regions (based on customer geolocation) generate the most revenue?**

A screenshot of a computer

Description automatically generatedHighest Revenue: Sao Paulo had the highest revenue of **$2,203,373.09**.

Other cities: Rio de Janeiro **(1,161,927.36)**, Belo Horizonte **(421,765.12)**, Brasilia **(354,216.78)**, and Curitiba **(247,392.48)**.

Revenue Comparison: Sao Paulo far outperforms other cities in total revenue.

1. **A screenshot of a computer screen

   Description automatically generated What is the average order value (AOV) and how has it changed over time?--(From 09-2016 to 10 - 2018)**

**Between 2016 and 2018, the average order value has seen significant changes. In 2016, the average order value was 171.57, while in 2017 it decreased to 152.55. However, in 2018, the average order value increased slightly to 155.31. These figures reflect fluctuations in financial performance over the three years.**

* A table with black and white text

  Description automatically generated**Customer Behavior:**

1. **What are the top 10 review categories?**

The following categories all received an average rating of 4: \*\*Books & Music, Kids, Automotive, Arts & Crafts, Home Appliances, Health & Beauty, Gifts & Toys, Food & Drink, Fashion, and Luggage

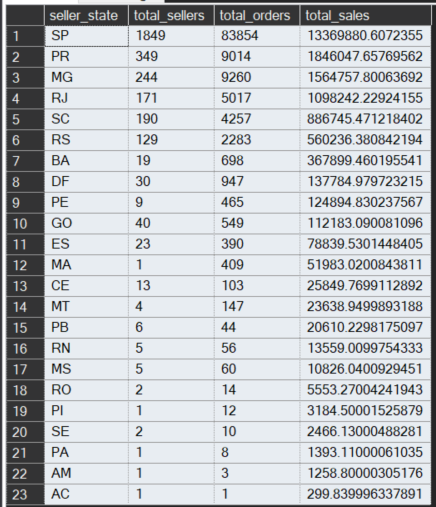
1. **What is the contribution of each payment type to overall sales? (The preferred payment method for our customers)**

**A screenshot of a computer screen

Description automatically generated**

**Credit card is** the preferred payment method for our customers at **78.34%.**

* **Seller Performance and Geographical Insights:**

1. **How many sellers per state and how many orders and sales?**

Sales and orders were distributed across several states, with the highest numbers being recorded by **São Paulo (SP) with 1,849 sellers, 83,854 orders and sales of $13,369,880.61. Paraná (PR) followed with 349 sellers, 9,014 orders and sales of R$1,846,047.66.** While the lowest numbers were recorded by states such as **Acre (AC) and Amazonas (AM) with only one seller and sales of less than R$1,300.**

A screenshot of a computer

Description automatically generated

1. **What is the most common order status (e.g., delivered, canceled, etc.)?**

The most common order status is **"Delivered",** with a total of **96,478** orders, far surpassing all other statuses. This indicates that a significant majority of orders were successfully delivered to customers.

* **A table with numbers and letters

  Description automatically generatedLogistics and Time Analysis:**

1. **What is the average shipping time, and how does it vary by region?**

**Acre (AC) and Ceará (CE)** have the longest average shipping times, with 21 days each.

**São Paulo (SP)** has the shortest average shipping time at 8 days, followed by **Paraná (PR) and Distrito Federal (DF)** with 11 days each.

1. **What is the time period which we receive the highest number of orders? (hour/day/month)?**

A table with numbers and numbers

Description automatically generated**The day with the highest number of requests was found to be Monday, with 16,196 requests. The most requested hour was the first hour of the day, with 6,695 requests. As for the month, January was the most requested, with 7,544 requests. On the other hand, the least requested day was Saturday, with only 10,887 requests.**

A table with numbers and a black and white background

Description automatically generatedA screenshot of a computer

Description automatically generated

* **Delivery Timelines and Customer Satisfaction:**

1. **What is the percentage of the on time delivered orders and what is the percentage orders that are not delivered on time?**

On time delivered percentage **(91.89%)**

Not on time delivery percentage **(8.11%)**

1. **A screenshot of a computer screen

   Description automatically generatedHow do early, on-time and late orders affect the customer reviews?**

The average review score for early orders was **4,** on-time was **4** and the late orders average review was **2**

* **Forecasting Questions**

1. **What are the expected earnings for the upcoming quarter, based on historical sales trends? (Dashboard)**
2. **What will be the total number of orders for the next 12 months? (Dashboard)**
3. **How will order cancellations and returns change over the next 12 months? (Dashboard)**