



Business Intelligence

Brazilian

E-Commerce Sales



Topic and Data Selection

Datasets Overview:

Topic: Brazilian E-Commerce Public Dataset by Olist

Content: This dataset features 100,000 orders from 2016 to 2018 made at multiple marketplaces in Brazil. It constitutes of 8 datasets that include out information about customer location, order, purchase timestamps, and price.

For our analytical assignment questions, we only required the utilization of **five datasets: Orders, Order Items, Products, Customers, and Sellers.**

Link: <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>

BI Justification and Focus

1. Top Selling Regions(Sales Champion Regions)

Description: Identify which region generate the highest revenue.

Analysis Approach: Aggregate total sales by state and sort them to find the top selling region.

Benefits: Strategic Marketing: Understanding top performing states can help tailor marketing strategies that leverage successful sales tactics.

2. Top Category According to Revenue (Revenue Leading Category)

Description: Determine which product categories yield the highest total revenue.

Analysis Approach: Sum total sales for each product category and rank them to see which categories are the most lucrative.

Benefits: Promotional Focus: Target promotions and discounts more effectively by understanding which categories drive the most revenue.

BI Justification and Focus

3. Top 5 Products According to Revenue (Best selling Products Analysis)

Description: Identify the 5 individual products that have generated the highest revenues.

Analysis Approach: Calculate total revenue for each product and identify the top earner 5 products.

Benefits:

Market Trends: Gain insights into consumer preferences and trends by understanding which products are most popular.

4. Number of Orders Per Month (Peak Demand Insights)

Description: Analyze the monthly order volume to identify seasonal trends and peak order months.

Analysis Approach: Group data by month and year, then count the number of orders for each period.

Benefits:

Marketing Campaigns: Time promotions and marketing campaigns to coincide with months of historically high demand.

Revenue Forecasting: Improve accuracy in revenue forecasts by understanding seasonal impacts on sales.

Data Modeling (Source Data)

Entities:

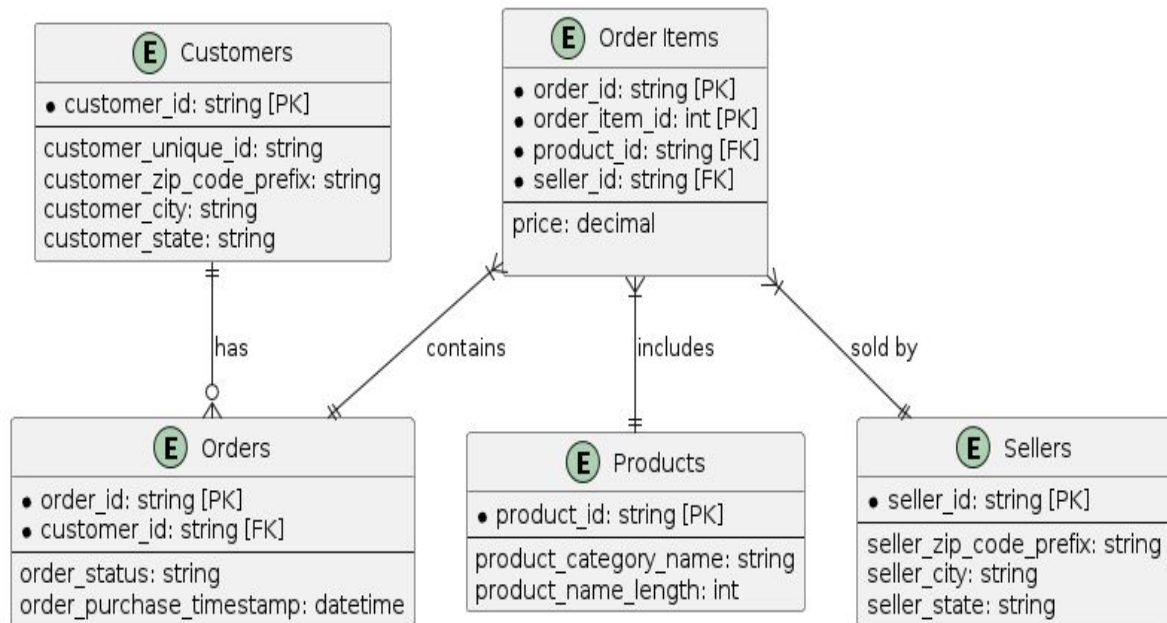
Customers, Orders, Order Items, Products, Sellers

Key Relationships:

Customers to Orders: One-to-many, each customer may have multiple orders.

Orders to Order Items: One-to-many, each order can consist of multiple items.

Order Items to Products and Sellers:
Many-to-one, connecting each item to its
respective product and seller.



Data Modeling (BI Data)

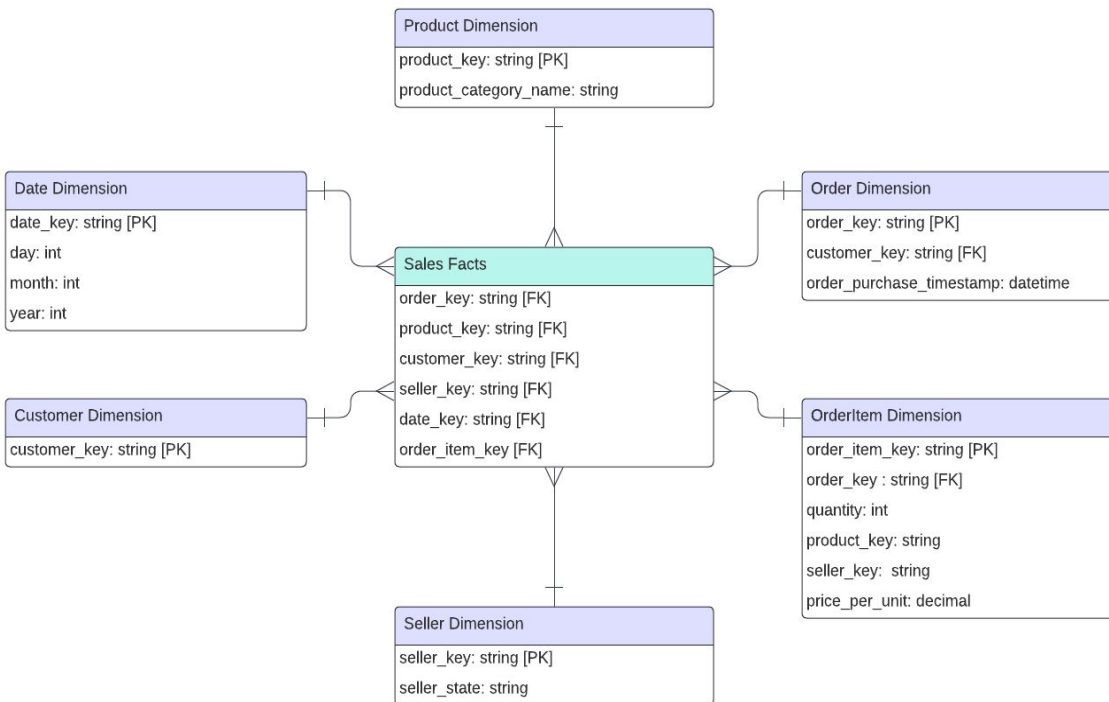
This star schema is designed to optimize query performance and simplify reporting by organizing data around a central fact table linked to several dimension tables.

Sales Facts

Primary Elements: order_key, product_key, customer_key, seller_key, date_key, order_item_key

Dimension Tables:

Date Dimension, Customer Dimension, Product Dimension, Order Dimension, Order Item Dimension, Seller Dimension



Data Processing

- Data processing is a crucial step to get structured data for efficient analysis in Tableau.
- Steps Involved:

1- Data Import

- Imported the source datasets including:

olist_orders_dataset.csv

olist_order_items_dataset.csv

olist_products_dataset.csv

olist_customers_dataset.csv

olist_sellers_dataset.csv

2- Data Transformation:

- The data transformation process involved restructuring the raw data so that we can use it in a star schema format.
- Key steps included creating separate DataFrames for fact and dimensions tables and deriving relevant attributes for the date dimension.

Steps:

- Created a date dimension with attributes such as year, month, day, and month name.
- Established separate dimensions for products, sellers, customers, orders, and order items.
- Due to the lack of uniqueness in the order_item_id within the source dataset, we assumed that it represented the quantity rather than a unique id. This assumption was based on the observance range of values, which varied from 1 to 4.
- Subsequently, we generated a unique id called 'order_item_unique_id' for the order_items dataframe.

2- Data Transformation: (cont.)

- Updated the column name 'order_item_id' to 'quantity' in the order_items dataframe.
- Additionally, upon examination of the source dataset of order_items, we identified duplicate rows in the order_items. These duplicates shared identical entries except for the quantity, which varied. To address this, we retained only the rows containing the maximum quantity value and discarded the duplicates.
- Converted the order date (order_purchase_timestamp) in the order_dimension to year-month-day format.

3- Data Integration:

- Data integration involved connecting datasets through shared attributes and merging DataFrames of the dimensions to create the fact table (sales_fact).
- The integration process additionally formed connections between the fact table and dimension tables by utilizing the primary keys of dimensions, which serve as foreign keys in the fact table.

Steps:

- Merged the order_items dimension with the orders dimension based on the order_id and subsequently with the customers (based on the customer_id), products (based on the product_id), sellers (based on the seller_id), and date (based on order_purchase_timestamp) dimensions.

4- Data Cleaning:

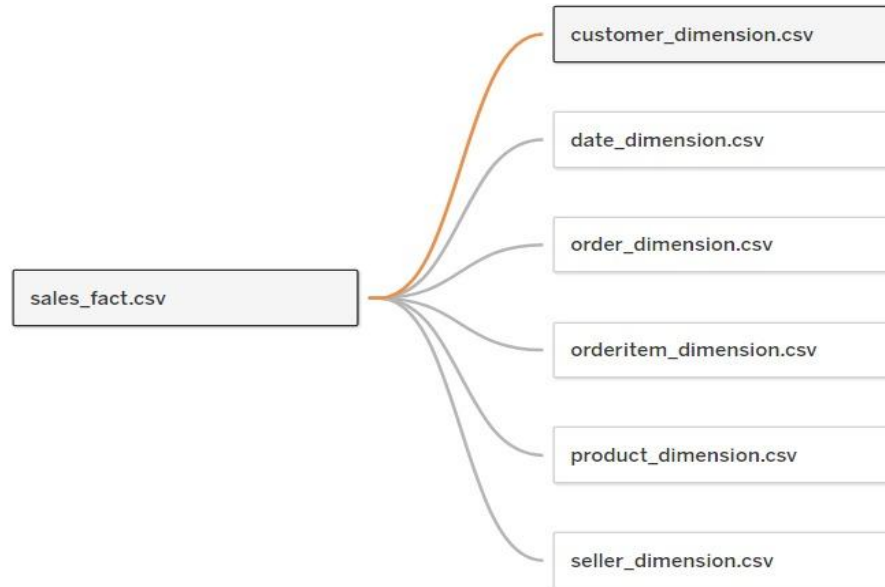
- Data cleaning focused on assessing data quality issues and resolving them to ensure data integrity.
- The cleaning process included checking for empty entries, removing rows with null values, and filtering out duplicate rows.

Steps:

- Checked for null values in all columns of the fact table and dimension tables and removed rows with null values.
- Identified and filtered out duplicate rows based on specified key identifiers (order_key, order_item_key, customer_key, and product_key).

5- Data Export: Output

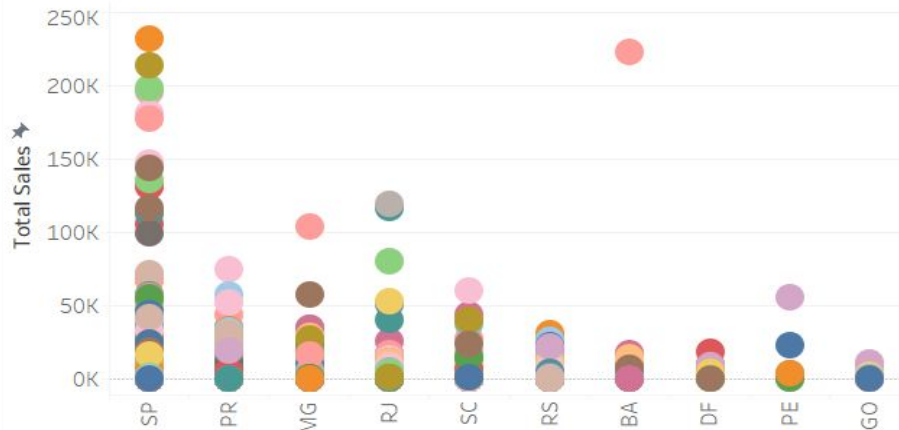
- Exported the transformed, integrated and cleaned data to CSV files for each dimension and the fact table as described in the Star Schema.



Data Analytics

Brazilian E-Commerce Sales Dashboard

Top Selling States



Product Categories



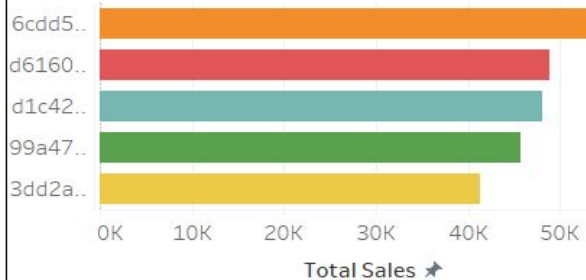
Year of Date

- ☒ (All)
- ☒ Null
- ☒ 2016
- ☒ 2017
- ☒ 2018

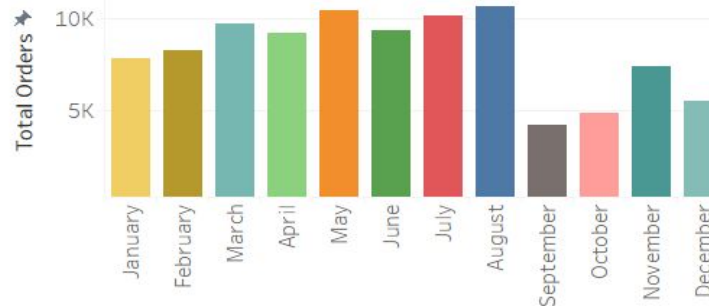
Product Category Name

- beleza_saude
- cama_mesa_banho
- esporte_lazer
- informatica_acessorios
- moveis_decoracao
- relogios_presentes

Top 5 Products



Peak Orders Per Month



Product Key

- 6cdd53843498f92890...
- d6160fb7873f184099...
- d1c427060a0f73f6b8...
- 99a4788cb24856965...
- 3dd2a17168ec895c7...

Dashboard Overview

This dashboard is designed to address four key analytical questions based on our Brazilian sales dataset.

View 1 - Sales Champion Regions: We identified the **top selling regions** generating the geographical regions that have the highest revenue over the whole period time of 2016-2018. We utilized states and total sales to derive these insights.

View 2 - Best Selling Product: We identified individual 5 products that have generated the most revenue over the whole period time of 2016-2018. While we didn't have the product names, we used product keys and total sales to determine which **top 5 products** in terms of generating revenue.

View 3 - Revenue Leading Categories: We have identified the product categories with the highest total revenue over the whole period time of 2016-2018. The names of the **top product categories** are displayed in Brazilian Portuguese, as the data was originally in this language.

View 4 - Peak Demand Insight: We examined the number of **orders per month** to find out the sales peak. By calculating the total number of orders for each month over the whole period time of 2016-2018, we found out the months with the highest number of sales.

Dashboard Interaction

- Our dashboard features four views that are synchronized with a filter for the years 2016 to 2018 on the right-hand side.
- This filter allows users to interact seamlessly with the data across all views.
- Additionally, the filter panel includes other filters related to month name and product category name because we did not displayed the names of the categories on the individual circle to keep the views simple.
- We have also included an extra filter to get the total sales per year that does not act as year filter but interacts with all of the four views, just to have an idea about the total sales over the years.

Dashboard Enhancement

- We explored alternative graph types for visual appeal, but in our case bar graphs were best suited for effective data visualization of answers to our analytical questions.
- We considered adding visual elements and KPIs but omitted to avoid clutter on a single dashboard.
- We tried to enhance the design and effectiveness of dashboard with simplistic views.
- We prioritized the core functionality of the dashboard and focused on seamless interaction.