Brazilian E-Commerce Data Analysis

Data Source:

Brazilian E-Commerce Public Dataset by Olist

The Brazilian E-Commerce Dataset by Olist includes information on orders from 2016 – 2018. It contains multiple features across different tables including, order status, price, payment, customer location, review, and more. The data is authentic that has been anonymized for security.

Olist is the largest department store in Brazilian marketplaces. They connect small buisnesses throughout Brazil in a single hub for ease. I chose this data set because E-Commerce has been on the rise and it is a good to get familiar with the layout of the sales. Also, I liked that this data was divided into different tables for understanding that could be merged. Going to allow for good practice merging data sets with Python.

Data Profile

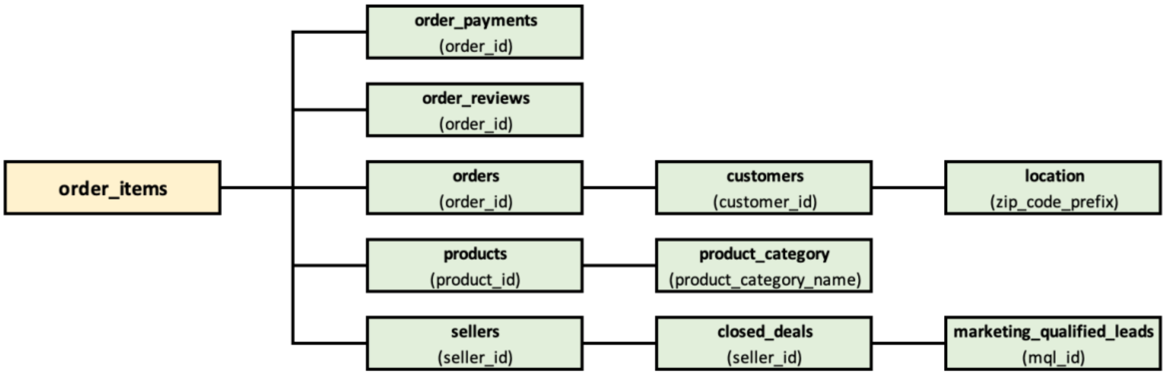
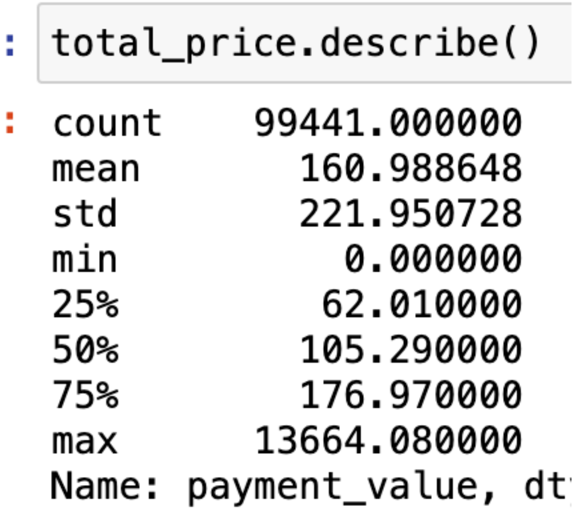


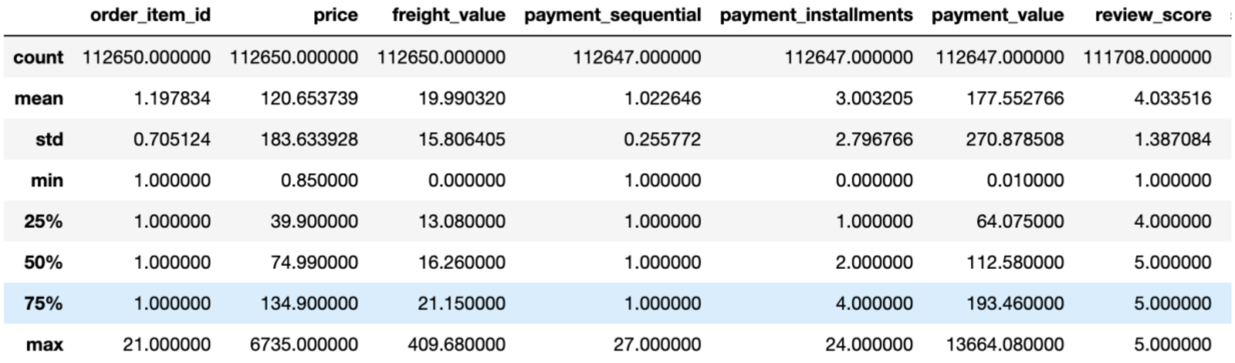
Table	Variable	Cleaned
Location	Zip_code_prefix	Removed geolocation_lat and _lng from data as that is more specific than I need.
Order_items	Order_id	Order_id can have duplicates because of multiple items being purchased in a single order.
Products	Product_category Product_name_length Product_description_length Product_photos_qty	Removed null entries (610)
Products	Product_name_lenght Product_description_lenght	Rename columns for correct spelling

- Checked every table for null values and duplicates. Some tables such as reviews, nulls were expected because not every customer will give a review.
- Merge tables into 1 dataframe.
- Group data by order_id to get descriptive statistics on price.



```
total_price.describe()
```

```
count    99441.000000
mean       160.988648
std        221.950728
min          0.000000
25%         62.010000
50%        105.290000
75%        176.970000
max      13664.080000
Name: payment_value, dt
```



	order_item_id	price	freight_value	payment_sequential	payment_installments	payment_value	review_score
count	112650.000000	112650.000000	112650.000000	112647.000000	112647.000000	112647.000000	111708.000000
mean	1.197834	120.653739	19.990320	1.022646	3.003205	177.552766	4.033516
std	0.705124	183.633928	15.806405	0.255772	2.796766	270.878508	1.387084
min	1.000000	0.850000	0.000000	1.000000	0.000000	0.010000	1.000000
25%	1.000000	39.900000	13.080000	1.000000	1.000000	64.075000	4.000000
50%	1.000000	74.990000	16.260000	1.000000	2.000000	112.580000	5.000000
75%	1.000000	134.900000	21.150000	1.000000	4.000000	193.460000	5.000000
max	21.000000	6735.000000	409.680000	27.000000	24.000000	13664.080000	5.000000

The data was collected by Olist. Customers were sent an email after they received their product to write a review and rate the product/experience. Not every customer filled out a review for the product so there could be bias related to who wrote a review and who didn't.

Questions To Explore:

What products are the most popular? Is there a reason they are popular or is it based on necessity?

Which products received the highest/lowest review rating. Why is this?

What is the delivery time for each product? Do certain products take longer to deliver than others? Is the review rating effected by the length of the delivery?