Chiao-Hsin Lo

Enhancing E-commerce Decision

Making Through Analysis and Visualization of Customer Shopping Trends

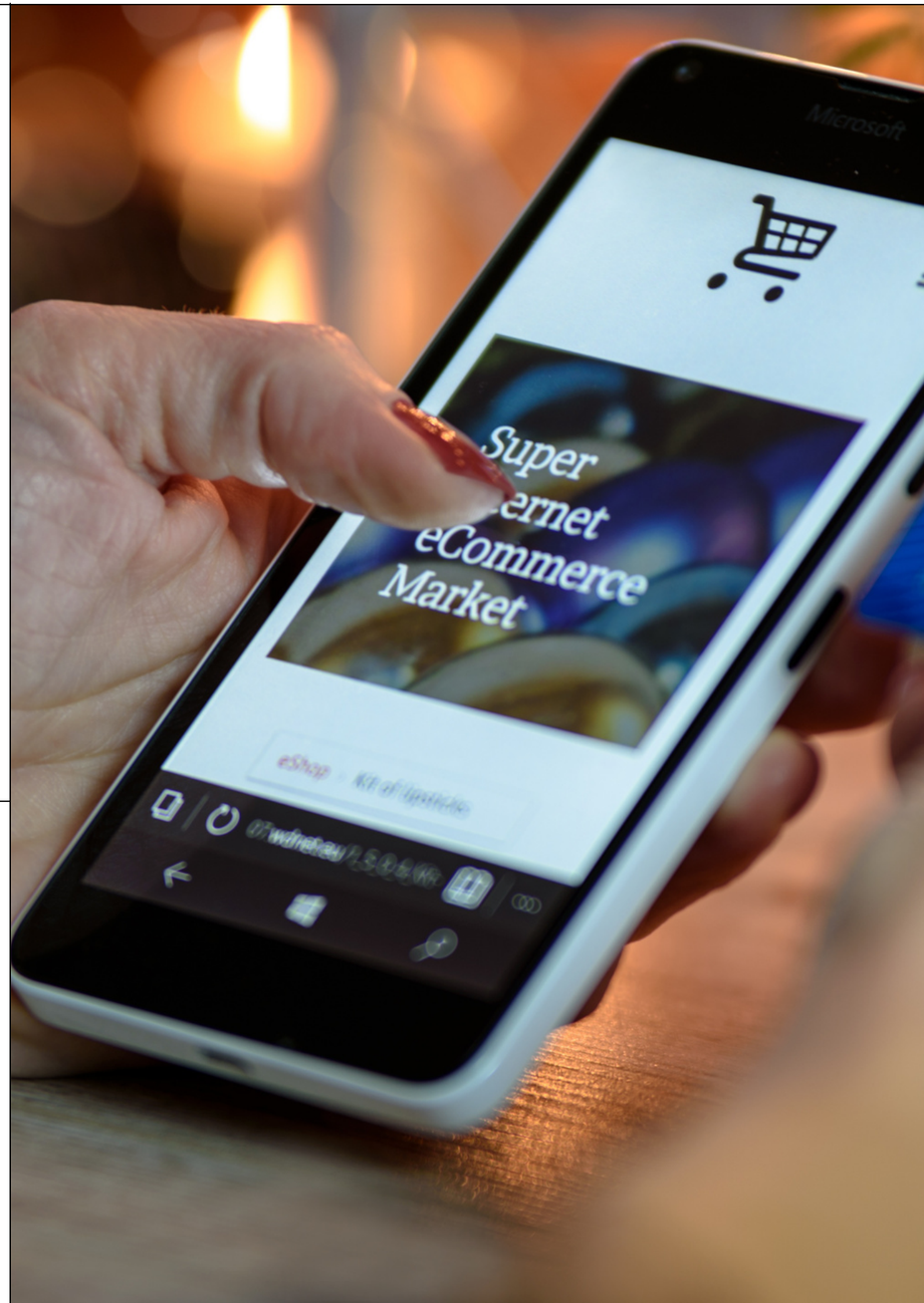
Content Table

Introduction

Background & Analytical Question

Background

This analysis utilizes an E-commerce platform's database to understand consumer buying patterns and develop strategies to enhance sales. The datasets are derived from the Brazilian E-commerce platform, Olist.

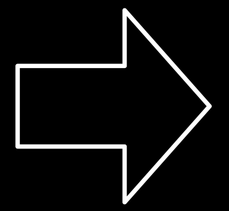


Analytical Question

- How did the performance fare?
- What marketing strategy recommendations could be implemented to enhance performance?

Exploratory Data Analysis

Datasets and Relationships & EDA



DATASETS & RELATIONSHIPS

- Olist E-commerce datasets are from Kaggle.
- There are 9 different datasets.

DATASET INFO

Datasets	Column Names	Column Names	Column Names	Column Names	Column Names	Column Names	Column Names	Column Names	Column Names
olist_customers_dataset	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state				
olist_geolocation_dataset	geolocation_zip_code_prefix	geolocation_lat	geolocation_lng	geolocation_city	geolocation_state				
olist_order_items_dataset	order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value		
olist_order_payments_dataset	order_id	payment_sequential	payment_type	payment_installments	payment_value				
olist_order_reviews_dataset	review_id	order_id	review_score	review_comment_title	review_comment_message	review_creation_date	review_answer_timestamp		
olist_orders_dataset	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date	
olist_products_dataset	product_id	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm
olist_sellers_dataset	seller_id	seller_zip_code_prefix	seller_city	seller_state					
product_category_name_translation	product_category_name	product_category_name_english							

DATASET EXPLORATORY DATA ANALYSIS (EDA)

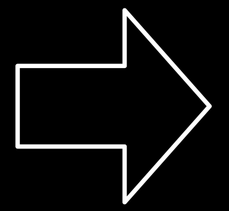
Exploratory Data Analysis (EDA)

```
# Collections for each dataset
# datasets = table['customers_dataset'], table['order_items_dataset'], table['order_payments_dataset'], table['order
datasets = [customers, order_items, order_payments, orders, order_reviews, products, sellers, product_category_name_

names = ['customers_dataset', 'order_items_dataset', 'order_payments_dataset', 'orders_dataset', 'order_reviews_data

# Creating a DataFrame with useful information about all datasets
data_info = pd.DataFrame({})
data_info['dataset'] = names
data_info['n_rows'] = [df.shape[0] for df in datasets]
data_info['n_cols'] = [df.shape[1] for df in datasets]
data_info['null_amount'] = [df.isnull().sum().sum() for df in datasets]
data_info['qty_null_columns'] = [len([col for col, null in df.isnull().sum().items() if null > 0]) for df in dataset
data_info['null_columns'] = [', '.join([col for col, null in df.isnull().sum().items() if null > 0]) for df in datas

data_info.style.background_gradient()
```

	dataset	n_rows	n_cols	null_amount	qty_null_columns	null_columns
0	customers_dataset	99441	5	0	0	
1	order_items_dataset	112650	7	0	0	
2	order_payments_dataset	103886	5	0	0	
3	orders_dataset	99441	8	4908	3	order_approved_at, order_delivered_carrier_date, order_delivered_customer_date
4	order_reviews_dataset	99224	7	145903	2	review_comment_title, review_comment_message
5	products_dataset	32951	9	2448	8	product_category_name, product_name_lenght, product_description_lenght, product_photos_qty, product_weight_g, product_length_cm, product_height_cm, product_width_cm
6	sellers_dataset	3095	4	0	0	
7	product_category_name_translation	71	2	0	0	
8	geolocation_dataset	1000163	5	0	0	

Data Analysis
&
Data Visualization

Order Analysis, Review Analysis, Geolocation Analysis,
Payment Analysis, Delivery Analysis, & Sales Analysis



Order Analysis

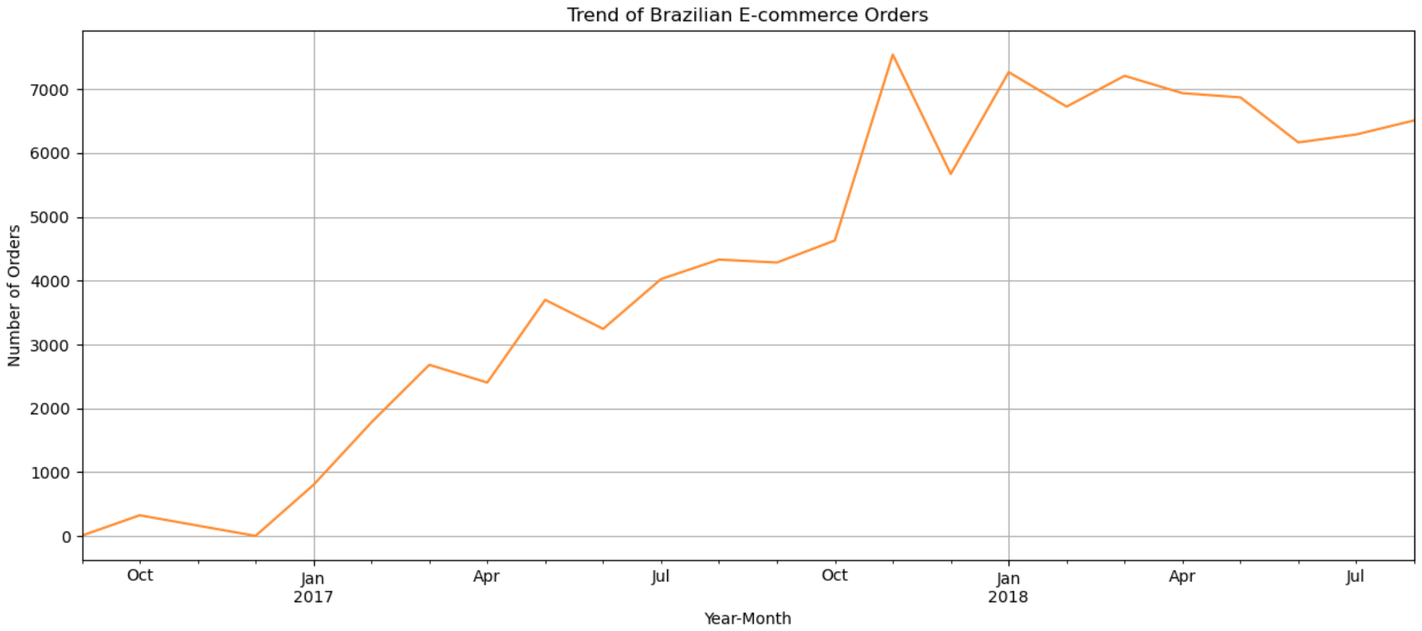
- Most orders were delivered, but still have around 3% of the orders were undelivered.
- The trend showed that people were more relying on E-commerce from 2016 to 2018.

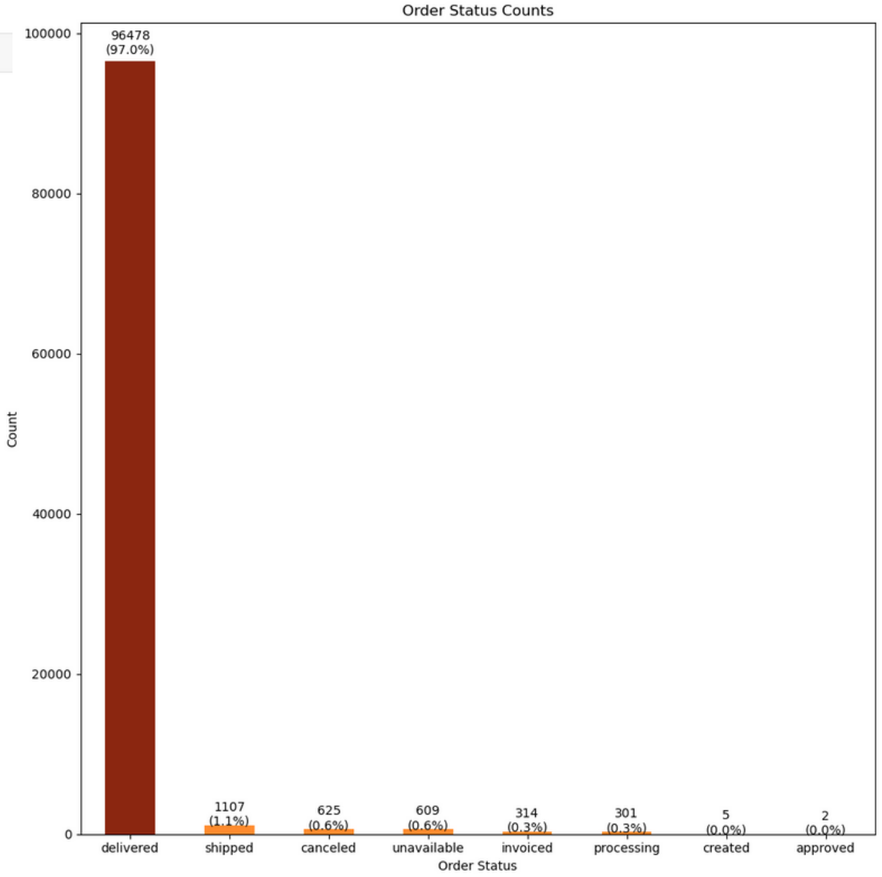
1. Orders Status Distribution

How many orders do we have for each status?

```
order_status_counts = orders['order_status'].value_counts()
order_status_counts
```

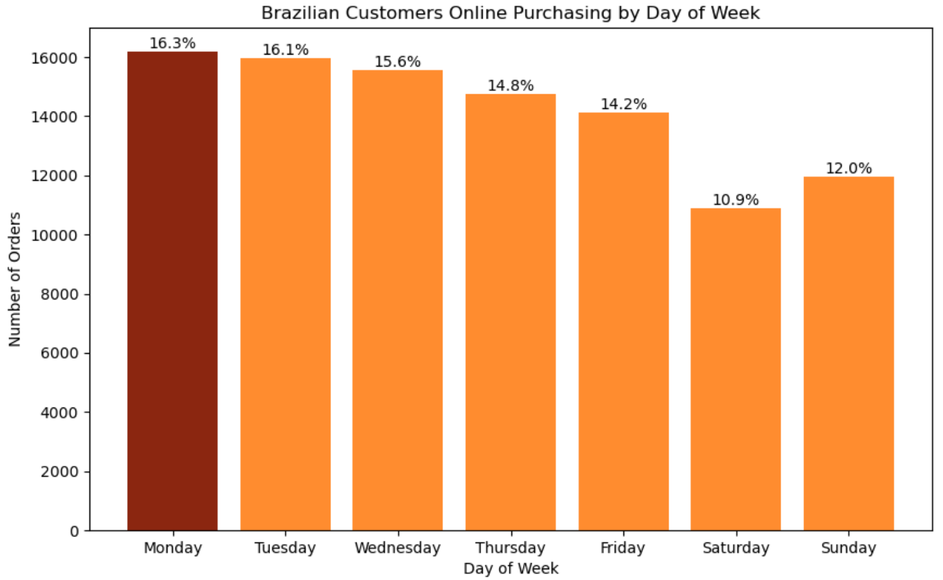
```
order_status
delivered      96478
shipped         1107
canceled         625
unavailable      609
invoiced         314
processing       301
created            5
approved           2
Name: count, dtype: int64
```

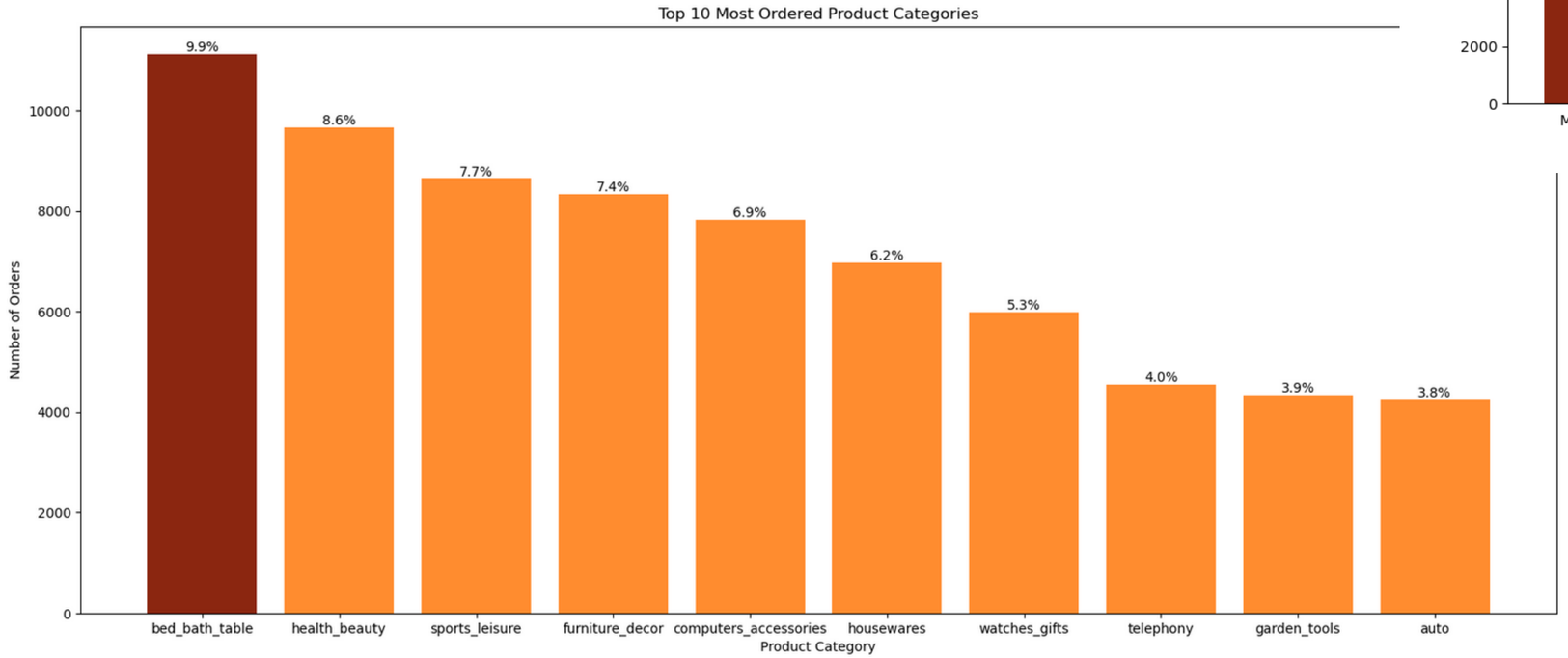


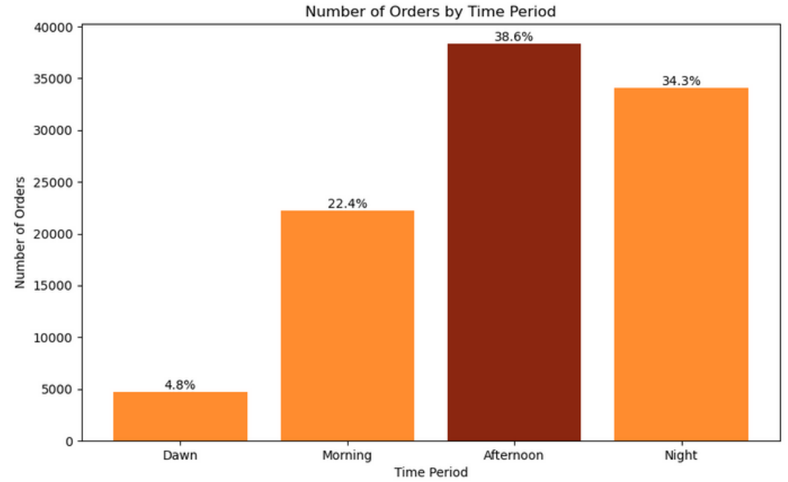


Order Analysis

People's purchase habits: online shopping the most on Monday, in the afternoon, and Bed, Bath, and Table category is the top puchase product category.





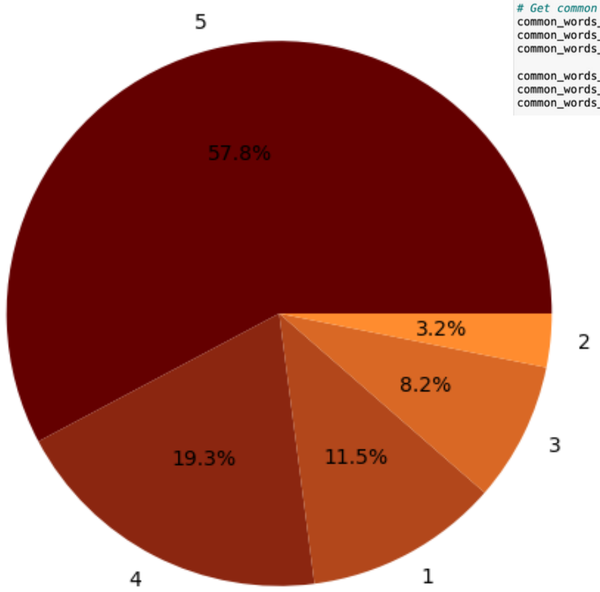


Review Analysis

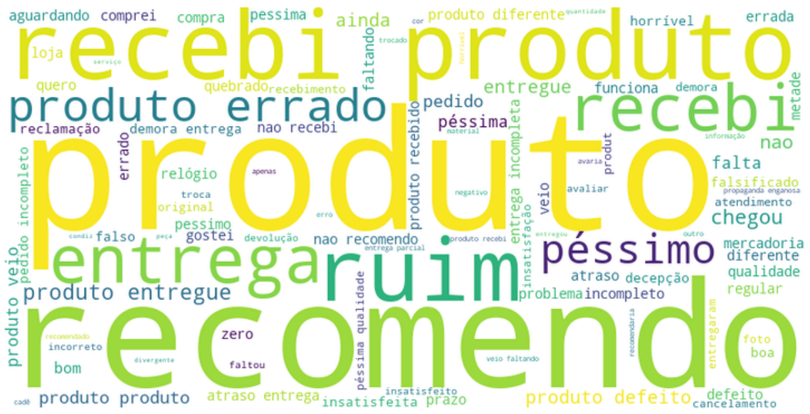
```python
def analyze_comments(score_values, comment_column):
    # Filter dataframe based on score values
    df_filtered = order_reviews[order_reviews['review_score'].isin(score_values)]

    # Get comments from the specified column
    comments = df_filtered[comment_column].dropna()

    # Convert all comments to lower case, split into words
    words = comments.str.lower().str.split()

    # Flatten the list of lists of words
    words = [word for sublist in words for word in sublist]

    # Remove stopwords
    words = [word for word in words if word not in stopwords.words('portuguese')]

    # Get 15 most common words
    common_words = Counter(words).most_common(15)

    return common_words

# Get common words for different review scores and different columns
common_words_high_scores_title = analyze_comments([5, 4], 'review_comment_title')
common_words_medium_scores_title = analyze_comments([3], 'review_comment_title')
common_words_low_scores_title = analyze_comments([1, 2], 'review_comment_title')

common_words_high_scores_message = analyze_comments([5, 4], 'review_comment_message')
common_words_medium_scores_message = analyze_comments([3], 'review_comment_message')
common_words_low_scores_message = analyze_comments([1, 2], 'review_comment_message')
```

- Most orders had high review score (4 & 5) ; however, low review score (1 & 2) still had 14.7%.
- In the reviews, although there were a lot good keywords, negative keywords were basically about delivery or wrong and defect product.
- Some negative keywords include 'entrega', 'errado', 'defeito', and 'péssimo', which means 'delivery', 'wrong', 'defect', and 'terrible' in English.

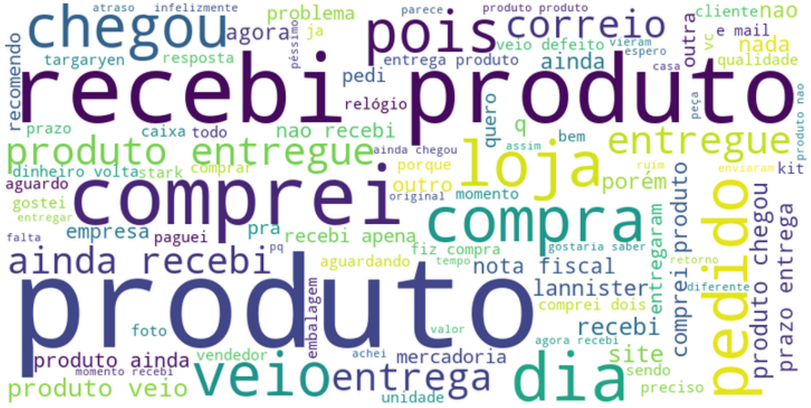
Distribution of Review Scores



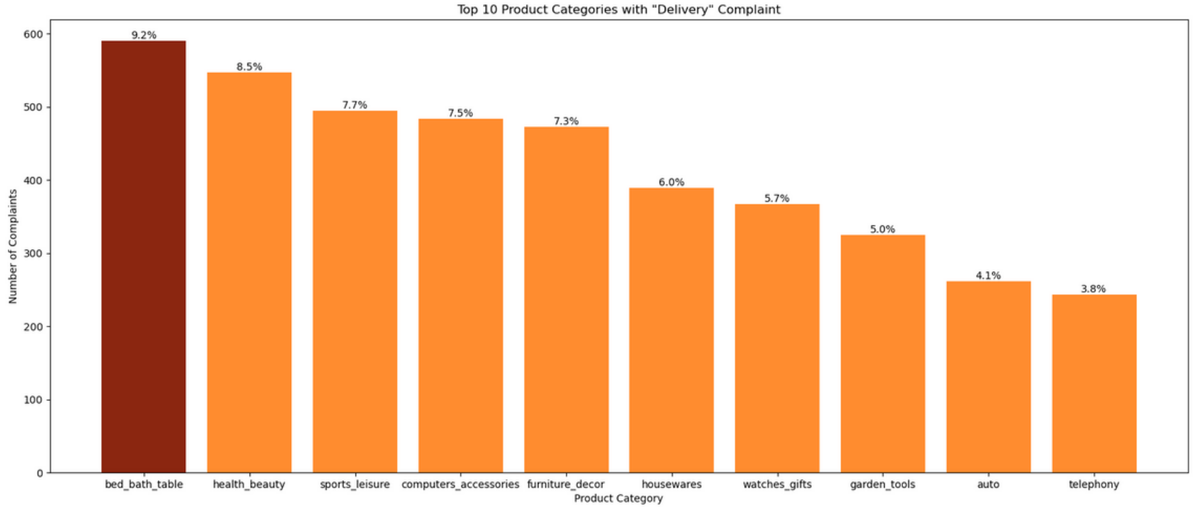
Common Words For Low Scores in Review Titles



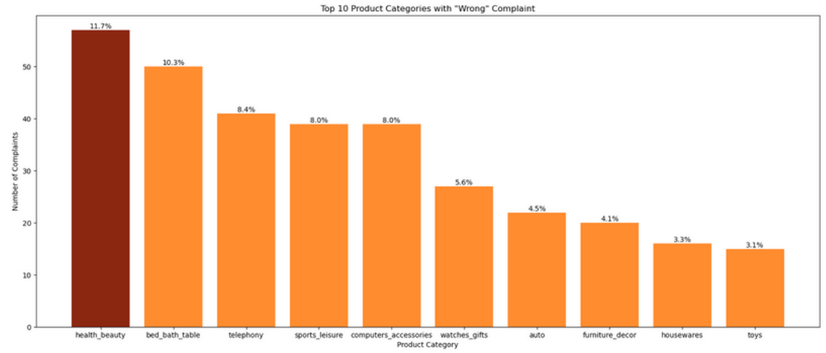
Common Words For Low Scores in Review Messages



Review Analysis

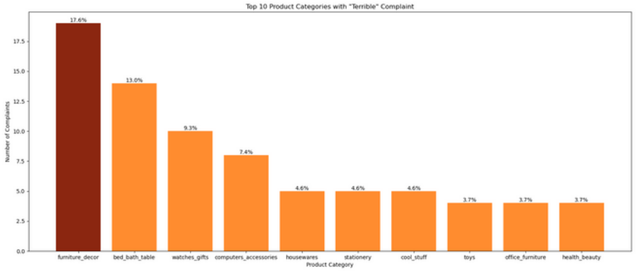


- Top category that was complained the most about 'delivery' is "bed_bath_table."

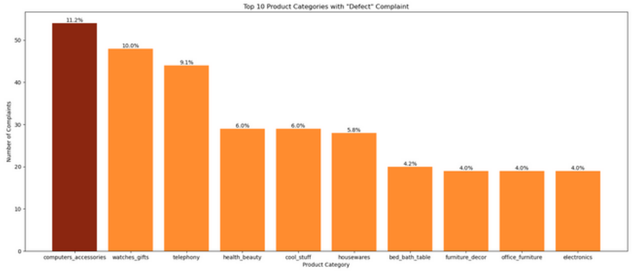


- Top category that was complained the most about 'wrong' is "health_beauty."

- Top category that was complained the most about 'defect' is "computers_accessories."



- Top category that was complained the most about 'terrible' is "furniture_decor."



Geolocation Analysis

SP (São Paulo) is the state with the highest payment value distribution and where most sellers locate in.

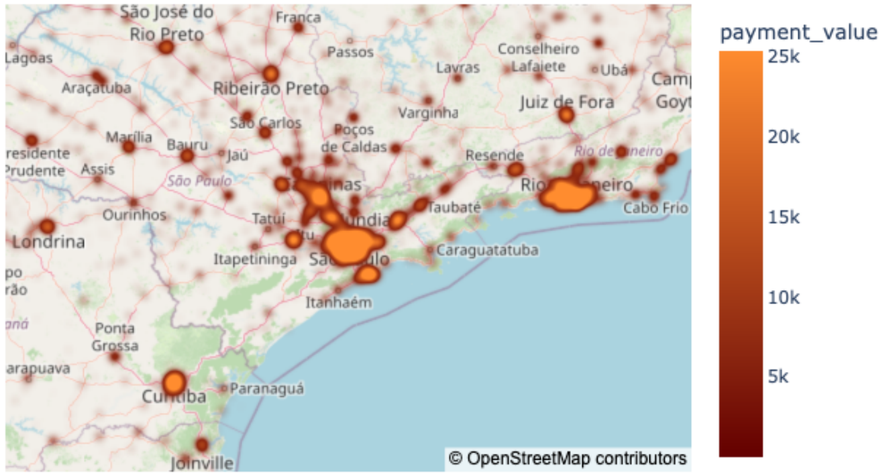
About 1.82% of all orders did not care about the freight, and SP (São Paulo) is also the state that has the most highest frieght ratio.

```
# Define outliers as any "freight ratio" over 1
outliers_over_1 = freight_value_ratio_new[freight_value_ratio_new['freight ratio'] > 1]

# Count the number of these outliers
num_outliers_over_1 = outliers_over_1.shape[0]

print(f"Number of 'freight ratio' values over 1: {num_outliers_over_1}")
print(f"Percentage of 'freight ratio' values over 1: {100 * num_outliers_over_1 / len(freight_value_ratio_new):.2f}%
```

```
Number of 'freight ratio' values over 1: 269
Percentage of 'freight ratio' values over 1: 1.82%
```

Since those outliers means that people would pay for high freight no matter how much the freight is, so would want to show heatmap that those outliers also mean 1.

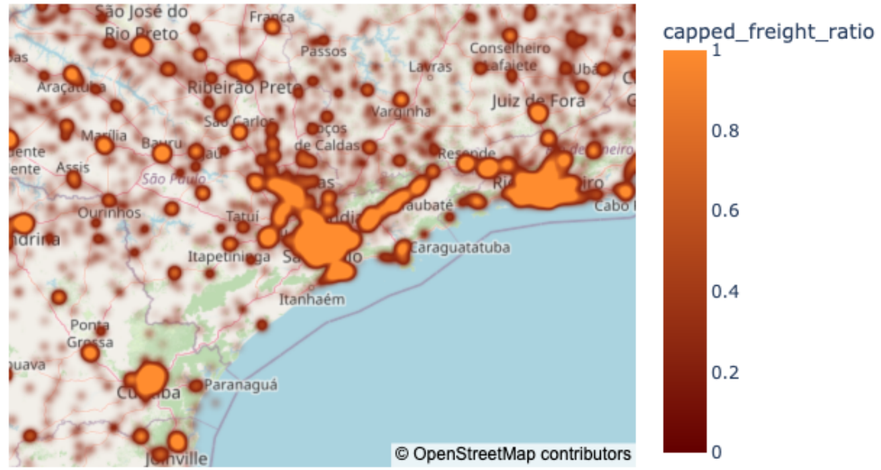
Payment Value Distribution across Geolocations




```
#seller by region
sellers_geo['seller_state'].value_counts(dropna=False)
```

```
seller_state
SP    1847
PR     348
MG     243
SC     190
RJ     171
RS     128
GO      40
DF      28
ES      23
BA      19
CE      13
PE       9
PB       6
RN       5
MS       5
MT       4
RO       2
SE       2
AC       1
PI       1
PA       1
MA       1
AM       1
```

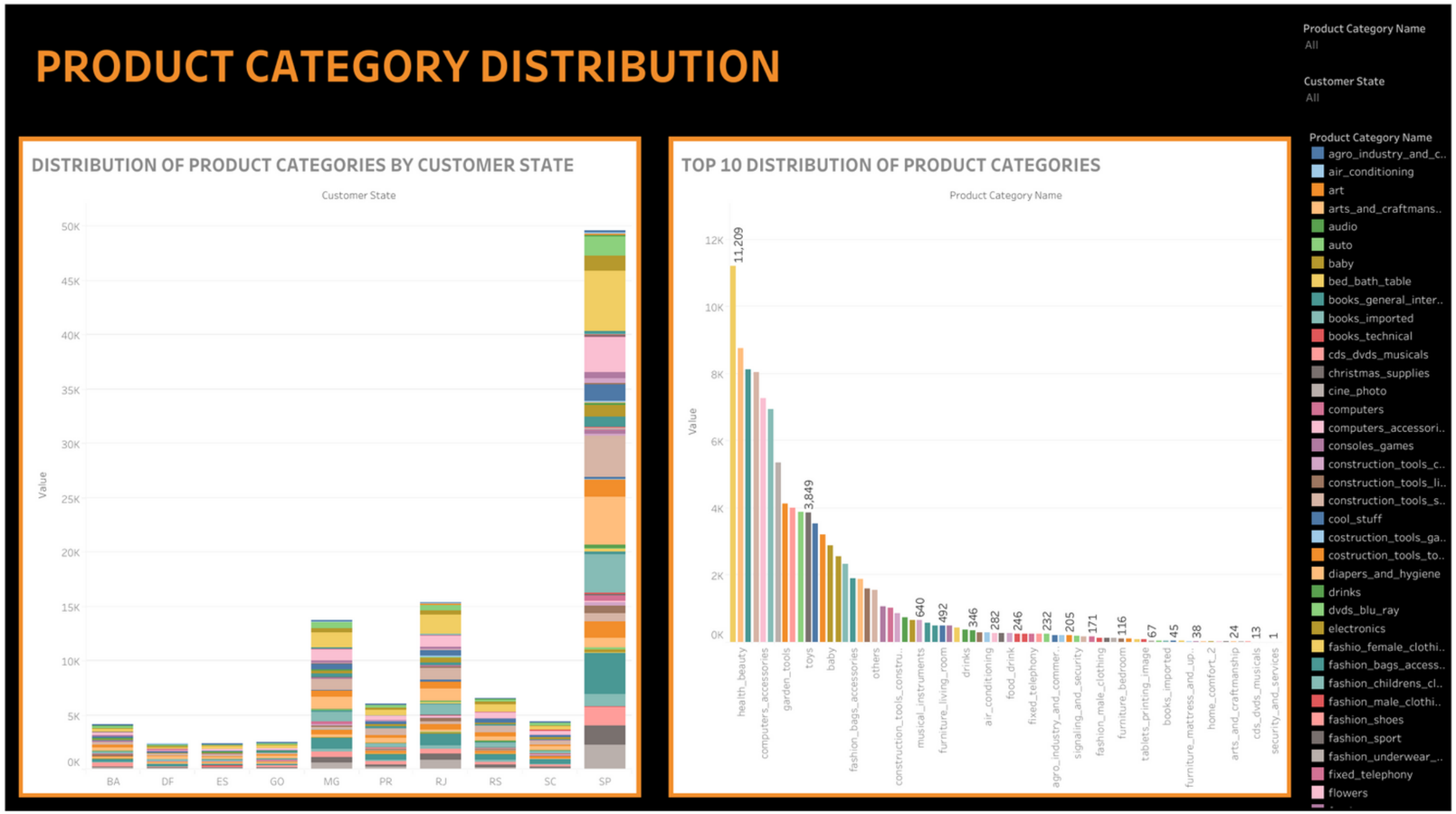
Freight Ratio Distribution across Geolocations



Geolocation Analysis

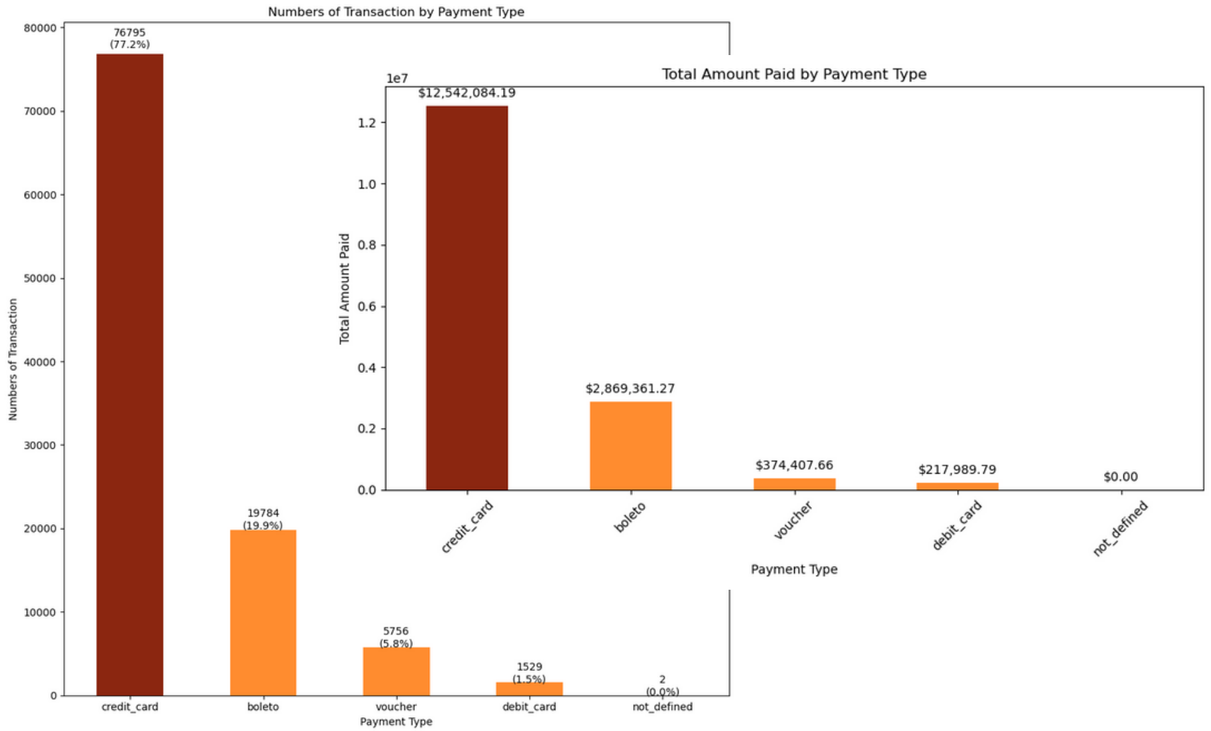
To look into product category distribution in each state, I created a dashbroad via Tableau for better GUI.

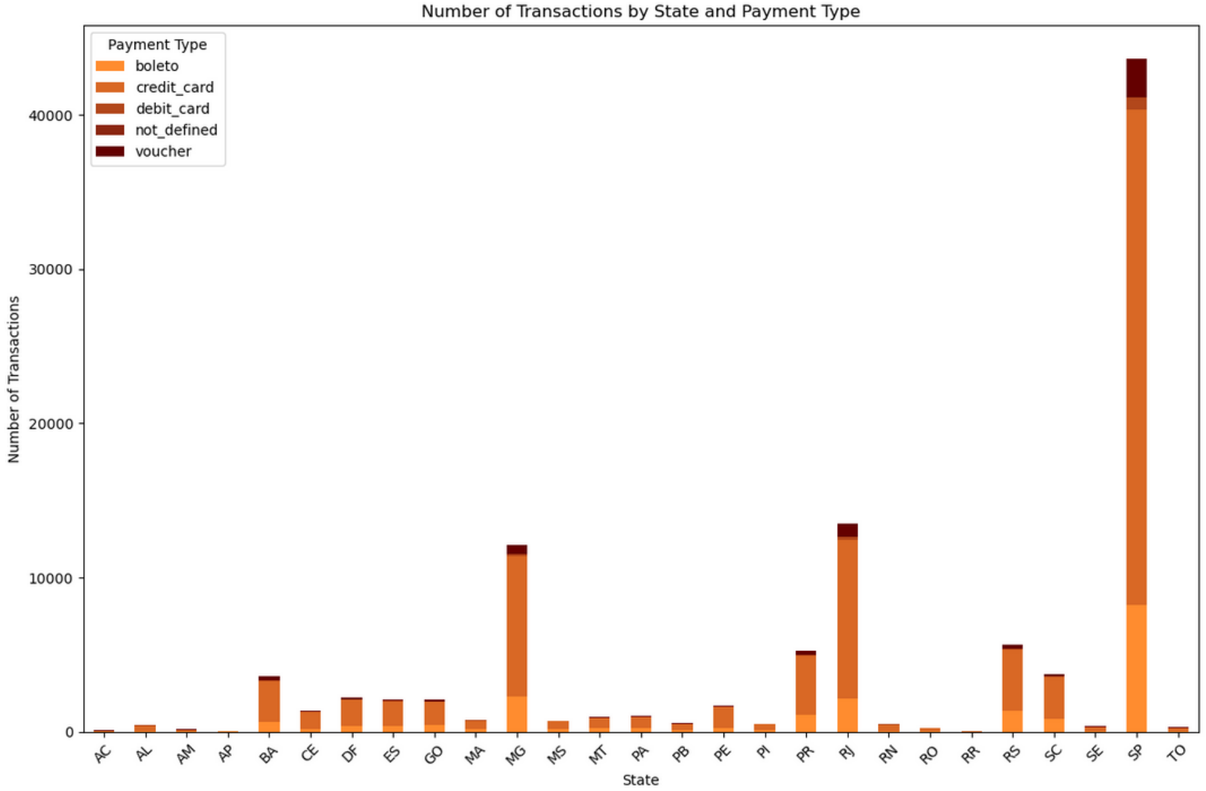
Click here to the “Product Category Distribrition” Link



Payment Analysis

- The most popular payment type is credit card and boleto* got the second-highest user.
- For the total amount paid in different payment types, credit card is still the highest.
- SP (São Paulo) had the highest number of transactions with most of them using credit card, following by boleto*.





*Boleto is a popular payment method in Brazil, it's a printed or virtual voucher with a barcode.

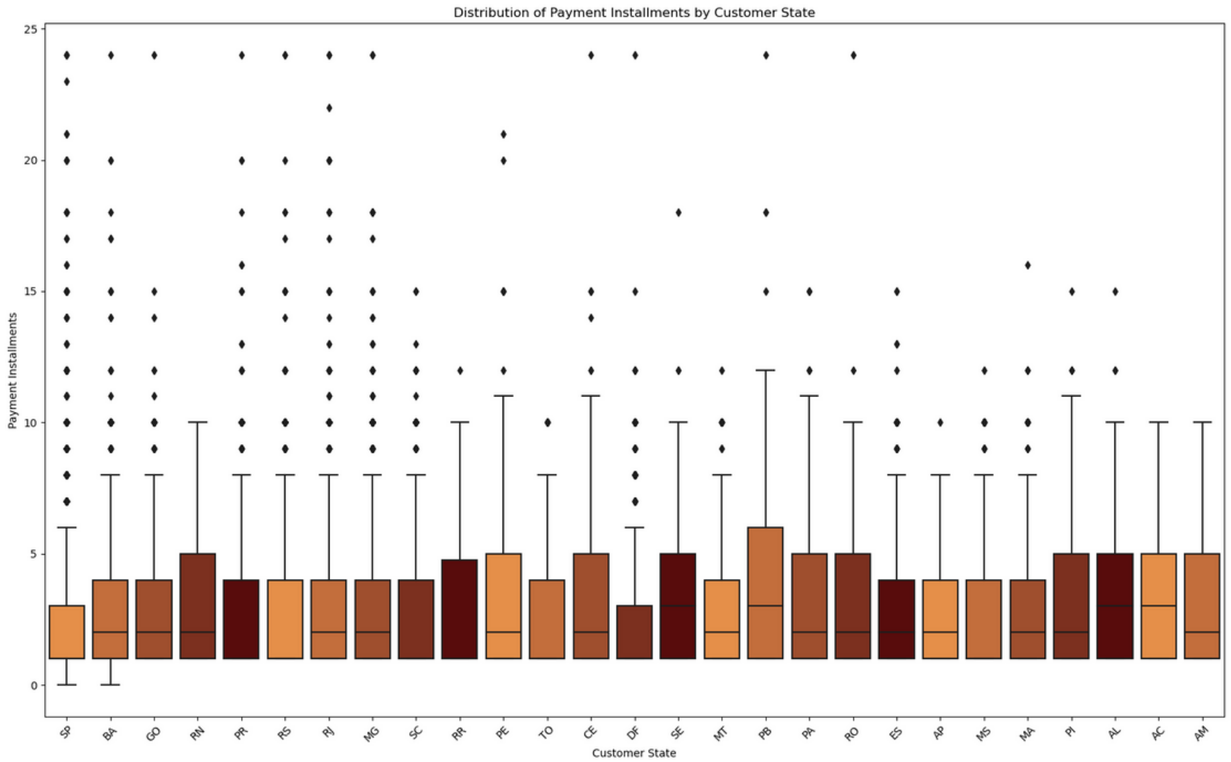
Payment Analysis

- Only credit card have more than one installment, with a mean around 3.5.
- There were numerous outliers for each state. The distribution of installment numbers in SP is right-skewed.

a. Installment count per payment

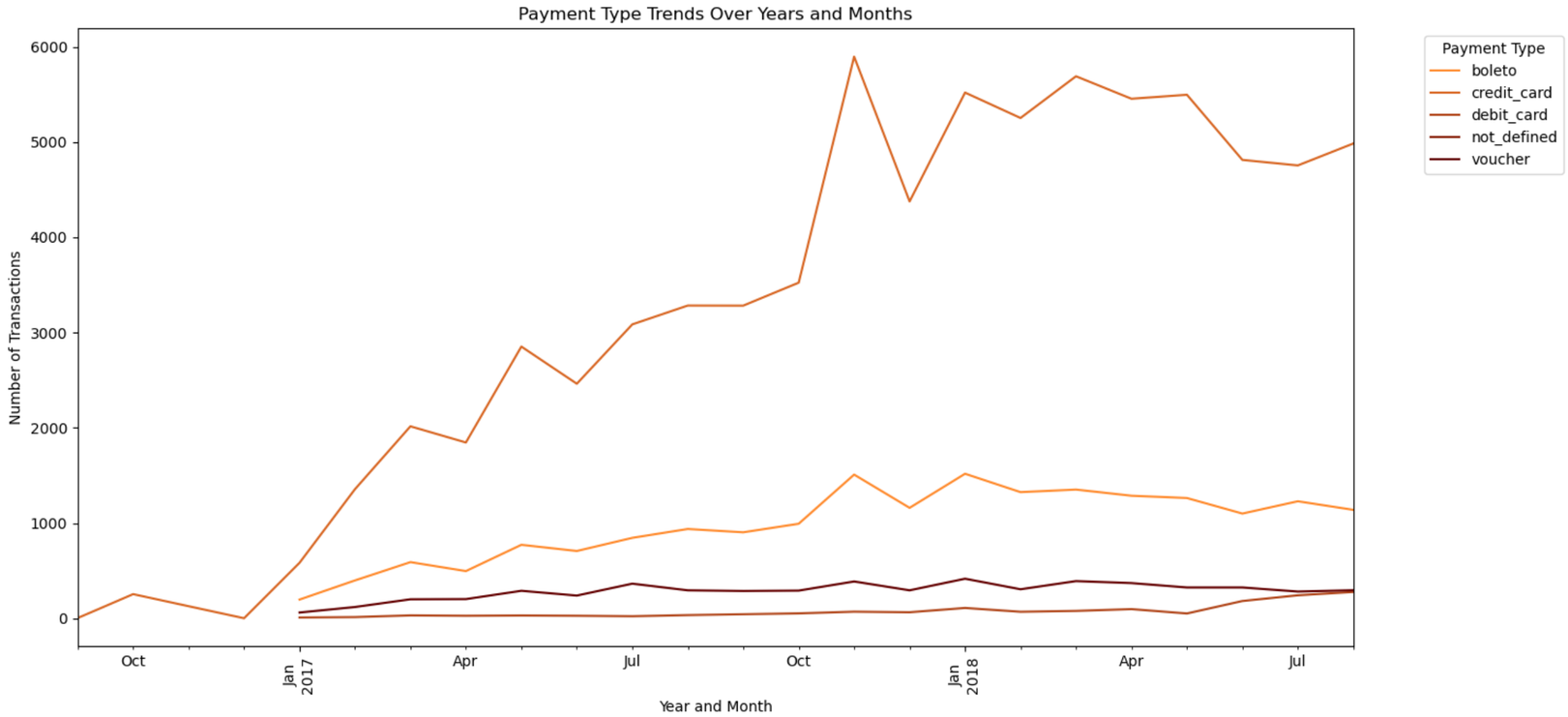
```
# Filter transactions with more than one installment
more_than_one_installment = customers_orders_payments[customers_orders_payments['payment_installments'] > 1]

# Check which payment types have more than one installment
payment_types_with_more_than_one_installment = more_than_one_installment['payment_type'].unique()

# Calculate the mean of installments for each payment type
mean_installments_by_payment_type = customers_orders_payments.groupby('payment_type')['payment_installments'].mean()

# Display the results
print("Payment Types with more than one installment:", payment_types_with_more_than_one_installment)
print("\nMean of installments by payment type:")
print(mean_installments_by_payment_type)
```

```
Payment Types with more than one installment: ['credit_card']

Mean of installments by payment type:
payment_type
boleto         1.000000
credit_card    3.507155
debit_card     1.000000
not_defined    1.000000
voucher        1.000000
```

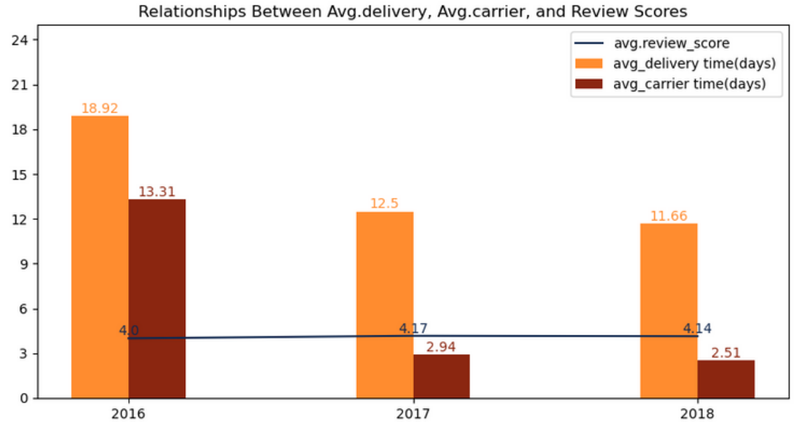


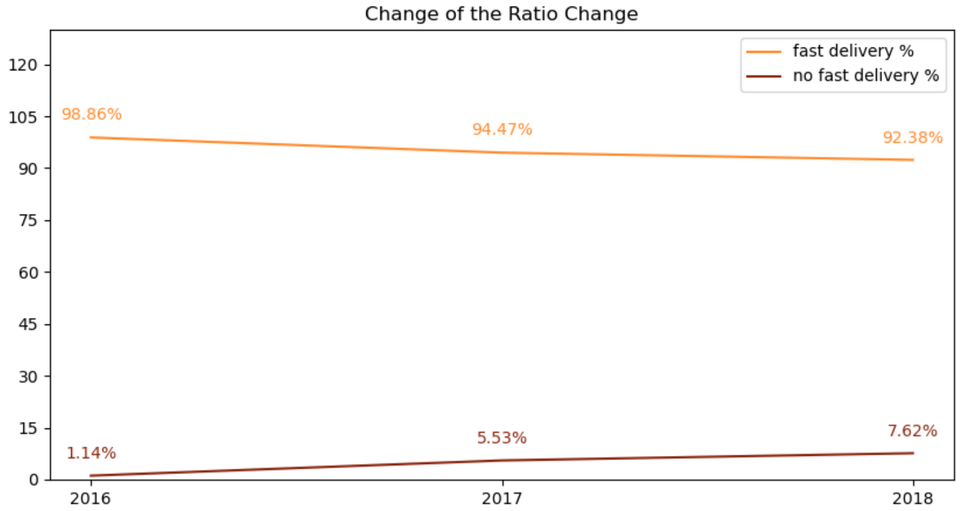
Payment Analysis

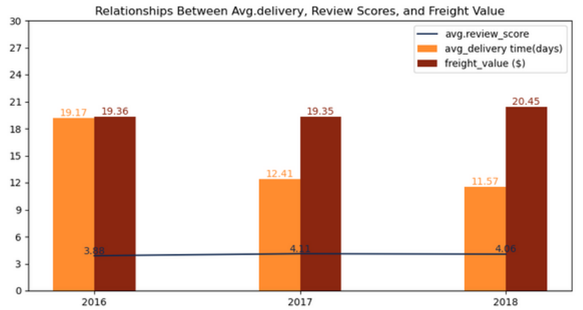
Credit card transactions dominate and have a significant fluctuation in the trend.



Delivery Analysis







- Avg. delivery time and Avg. carrier time have become faster. However, the delivery time has shortened, it didn't affect the Avg. review score.
- The ratio of Fast delivery** has declined. The review score of orders that have Fast Delivery is generally higher than those that don't have Fast Delivery.
- Freight Value does not affect Avg. delivery time and Avg. review score.

**Fast Delivery: comparing "Delivered Date" and "Estimated Date". If Delivered Date is faster than Estimated Date, it'll be identified as Fast Delivery.

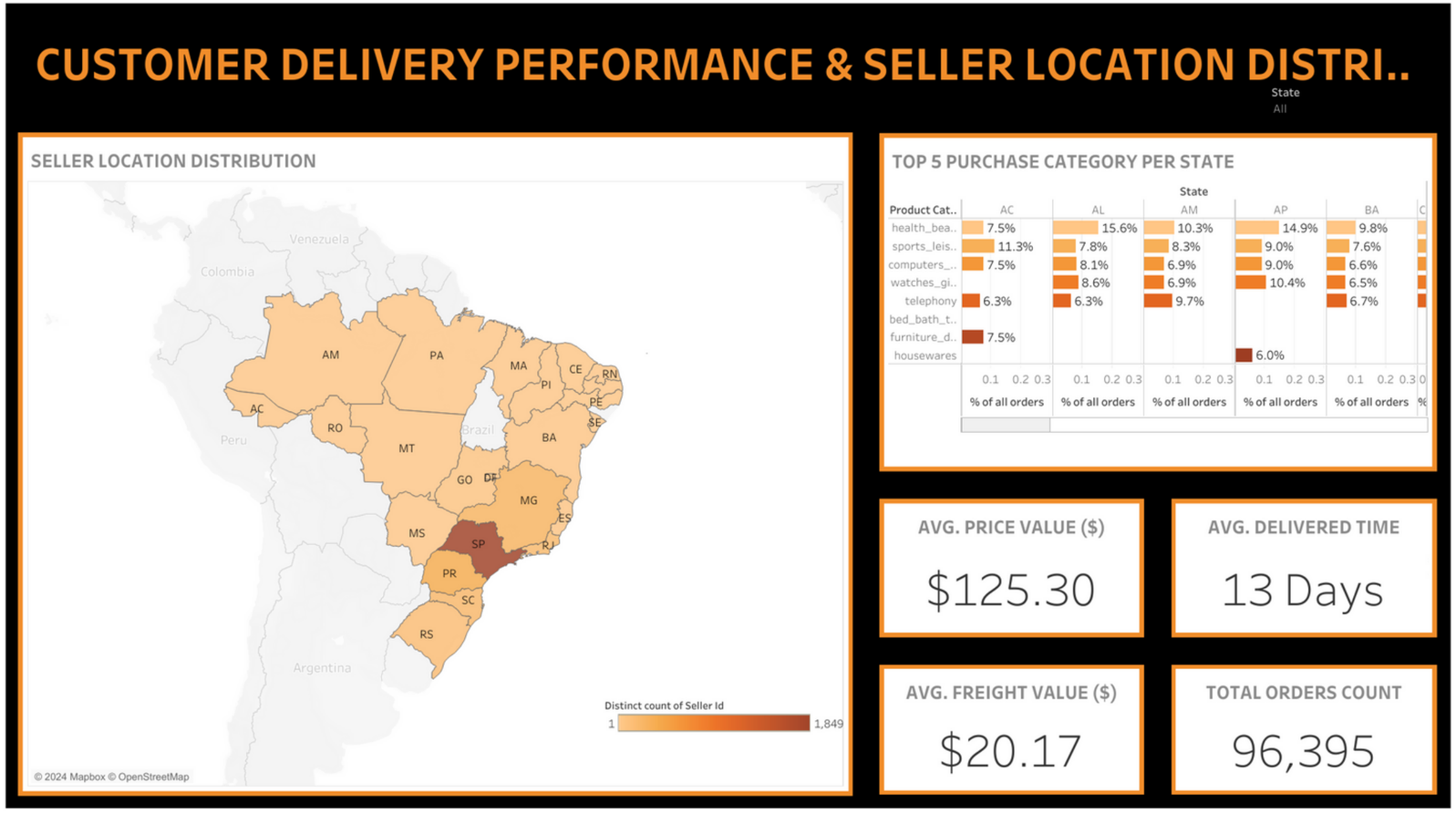
Delivery Analysis

I created a dashbroad via Tableau for better GUI.

Click here to the “Customer Delivery Data & Seller Location Distribution” Link

The number of orders spreads around SP (São Paulo), the farther away the area, the longer the Avg.delivery time and the higher the Avg.freight value, and vice versa.

The farther away from SP (São Paulo), the less amount the number of Seller locations.



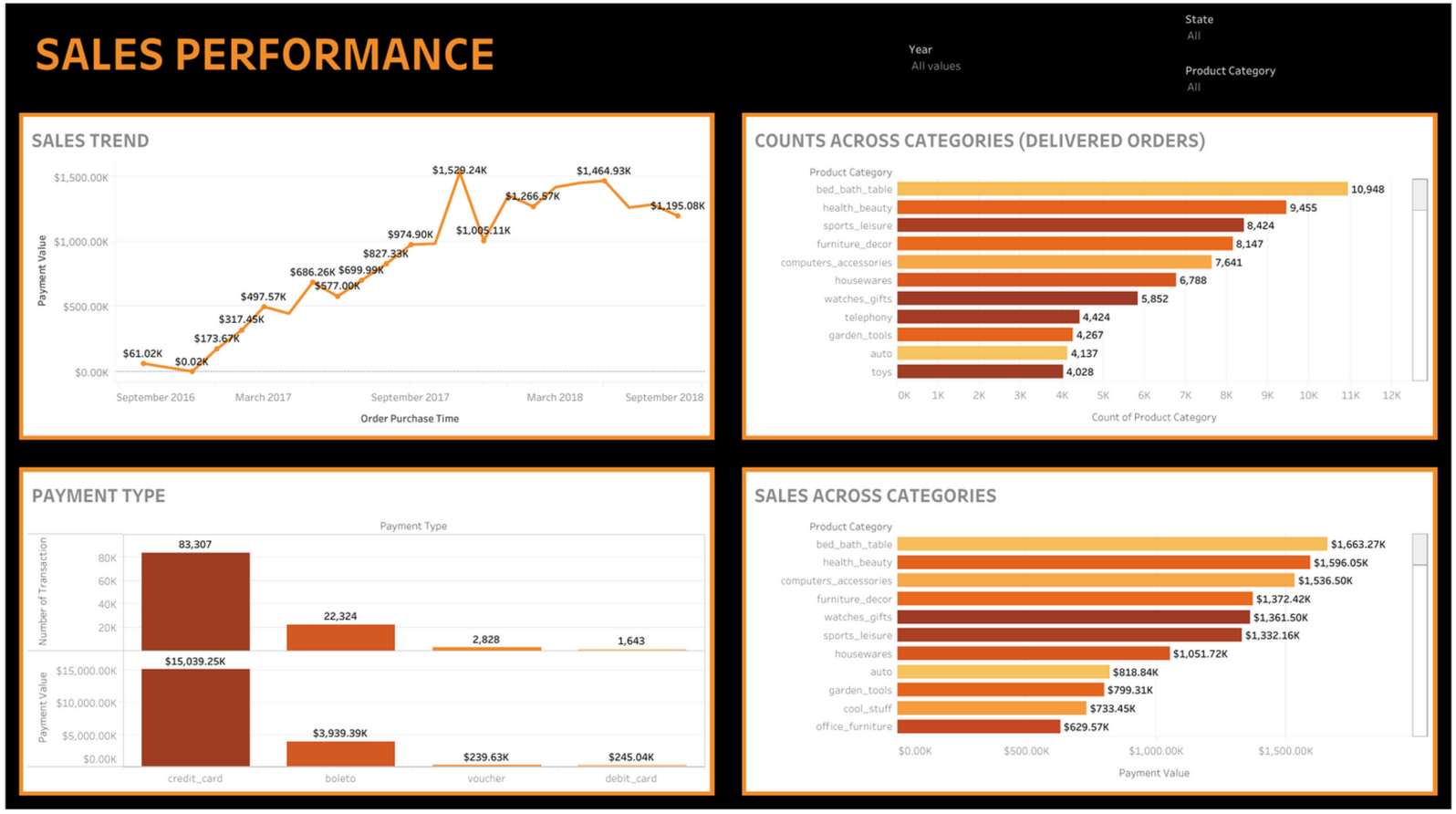
Sales Analysis

I created a dashbroad via Tableau for better GUI.

Click here to the “Sales Performance” Link

Among all, Bed, Bath, & Table > Health & Beauty > Computers Acsessories are the top 3 categories for sale from 2016 to 2018.

In terms of states, most states' sales are growing from 2016 to 2018.



Conclusions

Marketing Strategy Recommendations & Challenges

Marketing Strategy Recommendations

1

Host promotional events in the afternoon on Monday to maximize engagement based on customers' purchase habits and order

2

Set up a consolidated terminal for sellers to store the goods in the warehouse by the Olist to save the freight cost, and speed up the delivery time based on reviews, sales, geolocation, and delivery

Thank you!