# ANALYTIC TOOLS AND DECISION MAKING

# PROJECT SUBMISSION

# OLIST STORE INSIGHTS

*Hackslash*

Dig deep

**<u>Group Members</u>**

**Manoj Kumar Ravindranath**

**Sai Sowmith Komakula**

**Jerin Thomas**

**Elda Maria Thomas**

**Siva Krishna Unnam**

**Srilekhya Byreddi**

# CONTENT

# PROJECT PROPOSAL

## Introduction:

Olist is a Brazilian Startup that operates in the e-commerce segment, mainly through the marketplace. Olist connects small businesses to larger product marketplaces to help entrepreneurs sell their products to a larger customer base. On the one hand, the olist concentrates sellers who want to advertise on marketplaces such as Mercado Livre , B2W , Via Varejo and Amazon . On the other hand, it concentrates the products of all sellers in a single store that is visible to the final consumer. The company is headquartered in Curitiba, Paraná, and has an office in São Paulo. Currently, the business has 300 employees and more than 9,000 shopkeepers, in addition to 2 million unique consumers.

## Dataset Description:

Brazilian E-Commerce Public Dataset by Olist was obtained from Kaggle that consist of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. This also includes multiple Csv files based on region, customer information, seller information.
https://www.kaggle.com/olistbr/brazilian-ecommerce?select=olist_customers_dataset.csv

## Problem Statement:

The Dataset contains the order and customer information of 98K orders and reviews of those orders. Out of these orders, around 22000 orders have least review scores of 1 and 2 and this will be the problem to be solved. 22k is almost 21 percent of the orders received in the given time period and this could have a toll on the reputation this E-commerce business. This data also has information of orders that had got good reviews and with this support what worked well among customers can be found.

These are some of the questions to be analyzed:

What is the average review score across states based on delivery delay?
Does the payment value influence review score?
What is the review score based on Price and freight value?
Does the product from a particular seller state affect the review score?
Does the price range of a particular product have an impact on customer behavior?

## Background:

Online shopping refers to buying goods and services from merchants on the Internet. Online consumers are equally distributed between men and women and are more educated, younger, and affluent than those living in general. E-Commerce or Electronic commerce is a process of buying, selling, transferring, or exchanging products, services, and/or information via electronic networks and computers. It uses the internet and other online services to be engaged in buying and selling of digital and non digital products and services which require digital transportation or physical transportation.

The three technologies used in E-commerce are: Electronic Markets, Electronic Data Interchange Internet Commerce. E-commerce market is growing at a high rate. The online market is expected to grow by 56% in the next two years. retail e-commerce sales worldwide amounted to 2.3 trillion US dollars and e-retail revenues are projected to grow to 4.891 trillion US dollars while the growth of traditional market is gradually declining. Many larger retailers can maintain a presence offline and online by linking physical and online offerings to stay in the competition. There are different strategies for conducting business in online and traditional markets.

## Insight:

In the contemporary world number of customers for any business is proportional to the number of good reviews provided. This data has good proportion of good and bad review score its well balanced.

It is very apparent that the E-commerce will provide next generation shopping experience and customers are leaning towards this internet experience and studies show that the customers are content with online shopping and most likely to prefer only this in future at least with regards to some products. The more people prefer online shopping the more the expectations, Hence E-commerce business must be on point when it comes to providing a smooth shopping experience.

**Motivation:**

As we all know data is one of the biggest drivers of making an analysis and predictions. With this amount of data, and the kind of data, it is feasible to deliver valuable insights just by utilizing the open resources.

Earlier it was an expensive task to analyze a data because that mean using a lot of software resources and manpower. Today with the growth in software and other valuable resources. With regards to this data, all the necessary information that can be used to perform different analysis and with the time and date data it is also possible to predict the trend.

**Target Audience:**

This analysis will be focused on the needs of customers and small business owners. Since the problem to solved is based on review score, customers will be someone getting benefited. Hence, they are the target audience.

**Proposal:**

As a part of this project, a descriptive analysis will be conducted of the dataset and correlation will be found between the Review scores and other variables. The past trend will be thoroughly examined and will be used to understand the customer behavior. In the end these insights could be used improve the business performance.

**Approach:**

- Finding the relationships among the files will be the first step towards this project.
- Each and every file will be investigated.
- All the missing values, outliers and any other discrepancy will be addressed.
- Merging tables will be performed based on the needs.
- Excel and Python will be used for plotting the graphs.
- Tableau will be used to create dashboards and relationships.

## EXPLORATORY DATA ANALYSIS

**Dataset Description:**

This is a Brazilian ecommerce public dataset of orders made at Olist Store. The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews scores given by customers.

This dataset has 9 csv files.

**olist_customers_dataset.csv**

This dataset has information about the customer and its location. Use it to identify unique customers in the orders dataset and to find the orders delivery location.

| Column Name | Data Type | Description |
|---|---|---|
| customer_id | Varchar | Each order has a unique customer_id |
| customer_unique_id | Varchar | Unique identifier of a customer |
| customer_zip_code_prefix | Categorical | First five digits of customer zip code |

| | | |
|---|---|---|
| Customer_city | Categorical | Customer city name |
| customer_state | Categorical | Customer state name |

- There are 99441 unique customer ids.
- There are 4119 unique customer cities and most of the customers is from Sao Paulo.
- There are 27 states and most common state is SP.

### olist_geolocation_dataset.csv

This dataset has information Brazilian zip codes and its lat/lng coordinates. Use it to plot maps and find distances between sellers and customers.

| Column Name | Data Type | Description |
|---|---|---|
| geolocation_zip_code_prefix | Categorical | First 5 digits of zip code |
| Geolocation_lat | Numerical | Latitude |
| geolocation_lng | Numerical | Longitude |
| geolocation_city | Categorical | City name |
| geolocation_state | Categorical | State |

### Olist_order_items_dataset.csv

This dataset includes data about the items purchased within each order.

| Column Name | Data Type | Description |
|---|---|---|
| Order_id | Varchar | Order unique identifier |
| Order_item_id | Numerical | Number identifying number of items included in the same order. |
| product_id | Varchar | Product unique identifier |
| seller_id | Varchar | Seller unique identifier |
| Shipping_limit_date | Date | Shows the seller shipping limit date for handling the order over to the logistic partner. |
| price | Numerical | Item price |
| Freight_value | Numerical | Item freight value item |

### Olist_order_payments_dataset.csv

This dataset includes data about the orders payment options.

| Column Name | Data Type | Description |
|---|---|---|
| order_id | Varchar | Unique identifier of an order. |
| payment_sequential | Numerical | Customer may pay an order with more than one payment method. If he does so, a sequence will be created to accommodate all payments. |
| payment_type | Categorical | Method of payment chosen by the customer. |
| payment_installments | Numerical | Number of installments chosen by the customer. |
| payment_value | Numercial | Transaction value. |

- Mean and standard deviation of payment_value is 154 and 217 respectively.
- There are some outliers found in payment_installments and Payment_value.

**Olist_order_reviews_dataset.csv**

This dataset includes data about the reviews made by the customers.

| Column Name | Data Type | Description |
|---|---|---|
| review_id | Varchar | Unique review identifier |
| order_id | Varchar | Unique order identifier |
| review_score | Categorical | Note ranging from 1 to 5 given by the customer on a satisfaction survey. |
| review_creation_date | Date | Shows the date in which the satisfaction survey was sent to the customer. |
| review_answer_timestamp | Date | Shows satisfaction survey answer timestamp. |

**Olist_orders_dataset.csv**

This is the core dataset. From each order you might find all other information.

| Column Name | Data Type | Description |
|---|---|---|
| order_id | Varchar | Unique identifier of the order. |
| customer_id | Varchar | Each order has a unique customer_id. |
| order_status | Categorical | Reference to the order status (delivered, shipped, etc.) |
| order_purchase_timestamp | Date | Shows the purchase timestamp. |
| order_approved_at | Date | Shows the payment approval timestamp. |
| order_delivered_carrier_date | Date | Shows the order posting timestamp. When it was handled to the logistic partner. |
| order_delivered_customer_date | Date | Shows the actual order delivery date to the customer. |
| order_estimated_delivery_date | Date | Shows the estimated delivery date that was informed to customer at the purchase moment. |

- There are missing values found in order_approved_at (160 values), order_delivered_carrier_date (1783 values) and order_delivered_customer_date (2965 values).
- No outliers found in this dataset.

**Olist_products_dataset.csv**

This dataset includes data about the products sold by Olist.

| Column Name | Data Type | Description |
|---|---|---|
| product_id | Varchar | Unique product identifier |
| product_category_name | Categorical | Root category of product, in Portuguese. |
| product_name_length | Numerical | Number of characters extracted from the product name. |
| product_description_length | Numerical | Number of characters extracted from the product description. |
| product_photos_qty | Numerical | Number of product published photos |
| product_weight_g | Numerical | Product weight measured in grams. |
| product_length_cm | Numerical | Product length measured in centimeters. |
| product_height_cm | Numerical | Product height measured in centimeters. |

| | | |
|---|---|---|
| product_width_cm | Numerical | Product width measured in centimeters |

- Except for product_id, 610 missing values are found in rest of each columns.
- There are outliers found in product dimensions.

**Olist_seller_dataset.csv:**

This dataset includes data about the sellers that fulfilled orders made at Olist. Use it to find the seller location and to identify which seller fulfilled each product.

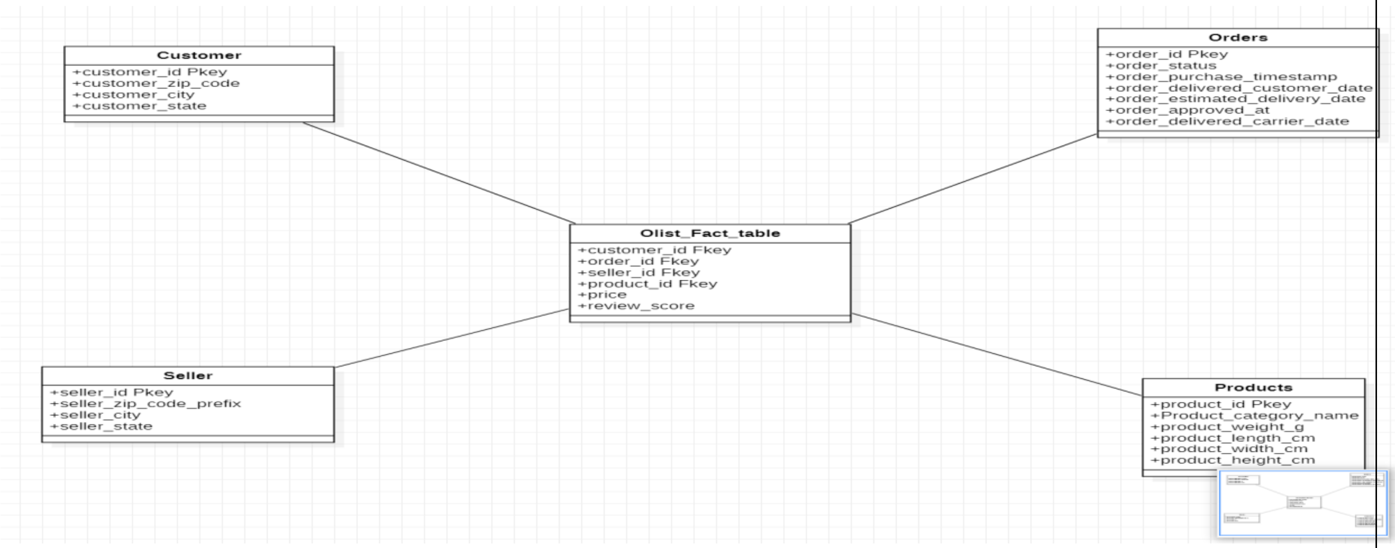
| Column Name | Data Type | Description |
|---|---|---|
| seller_id | Varchar | Seller unique identifier |
| seller_zip_code_prefix | Categorical | First 5 digits of seller zip code |
| seller_city | Categorical | Seller city name |
| seller_state | Categorical | Seller state |

**product_category_name_translation.csv**

Translates the product *category* name to English.

| Column Name | Data Type | Description |
|---|---|---|
| product_category_name | Categorical | Category name in Portuguese |
| product_category_name_english | Categorical | Category name in English |

**Schema:** The original dataset is modified by replacing the columns using the power query. Below is the star schema of the dataset. We have most of the ids in fact table as foreign keys and dimension table contains data of customers, orders, products, and sellers.



**Data Segmentation:**

Since the dataset has large verities of information it is necessary to subset the data into groups of product, customers, and sellers. Each segment holds appropriate information with respect to the data of various subsets.

**Data Imbalance:**

Review score column in our fact table looks imbalanced with 57% of the score being 5 and remaining 43% attributes to other score, but this shouldn’t be a problem as the data aligns with real world data.

Mostly customers preferred payment method is credit cards compare to other methods. More than 75% customer orders from olist are paid through credit cards.

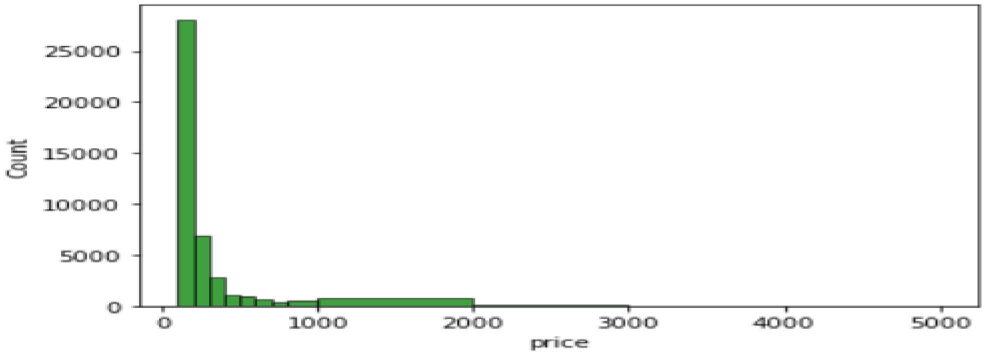
**Inclusion Criteria:**

Few custom columns need to be included that calculates number of days between estimated_delivery_date column and order_delivered_customer_date. Key features that are to be used in this analysis are order price, payment value, freight value, customer state, product category and payment installments.

**Preliminary Visualizations:**

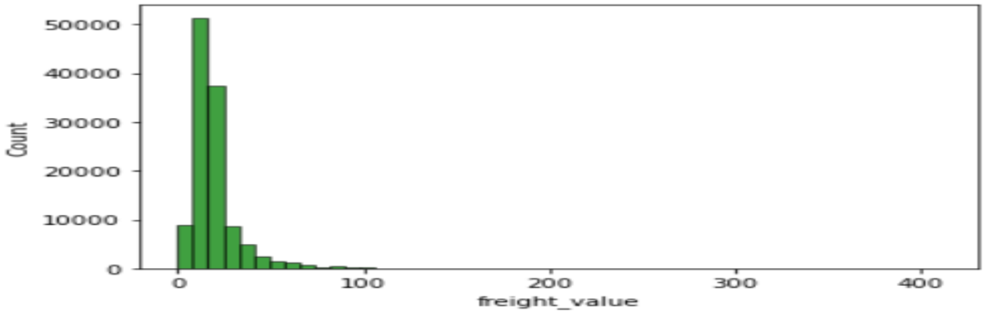
**Univariate Analysis**

**1. Price distribution of products. (Price of product vs Count)**



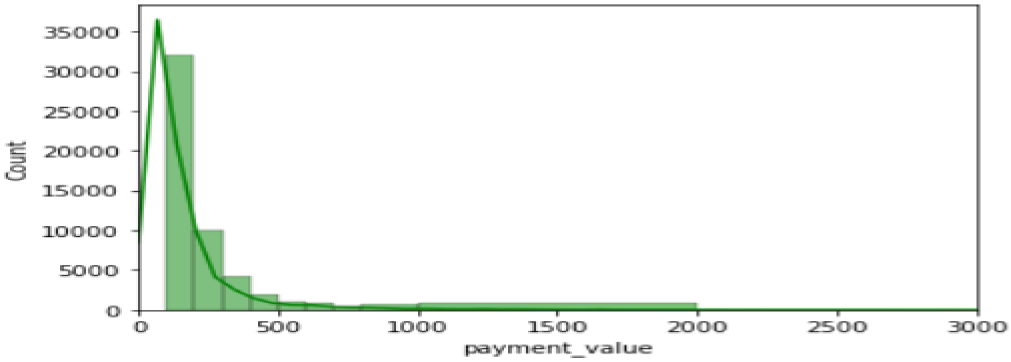
On X-axis we have price of each product and on Y-axis we have counts of products which have same price.Price of orders anything above 1000 is considered as an outlier.

**2. Distribution of freight value. (Freight value of product vs count)**



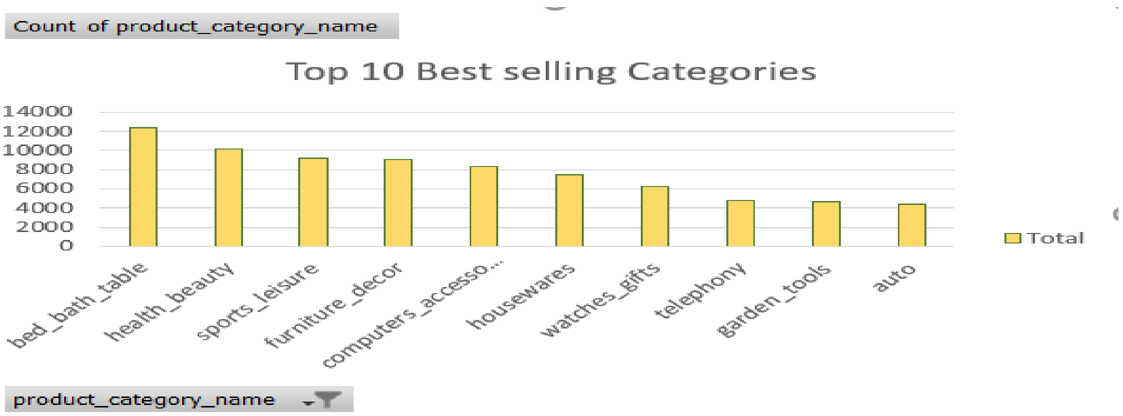
The graph shows the distribution of fright value among orders. if an order has more than one item the freight value is splitted between items. The data from the graph is right skewed.

**3. Distribution of payment value (Payment Value of product vs count)**



This graph shows distribution of payment value. Almost most of the payment is under 500 and this distribution clearly has some outliers. So, anything above 1000 could be considered as an outlier.

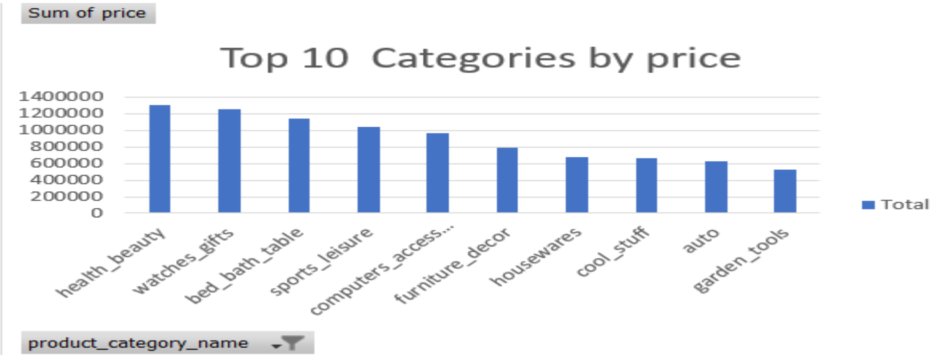
**4. Top 10 best selling product categories (Category of products vs count)**



Among all the categories these products are in top 10 mostly purchased by customers. We can observe the products of bed_bath_table is on the top 1 which are sold in more than double of the auto products which are in top 10.

**Bivariate Analysis**

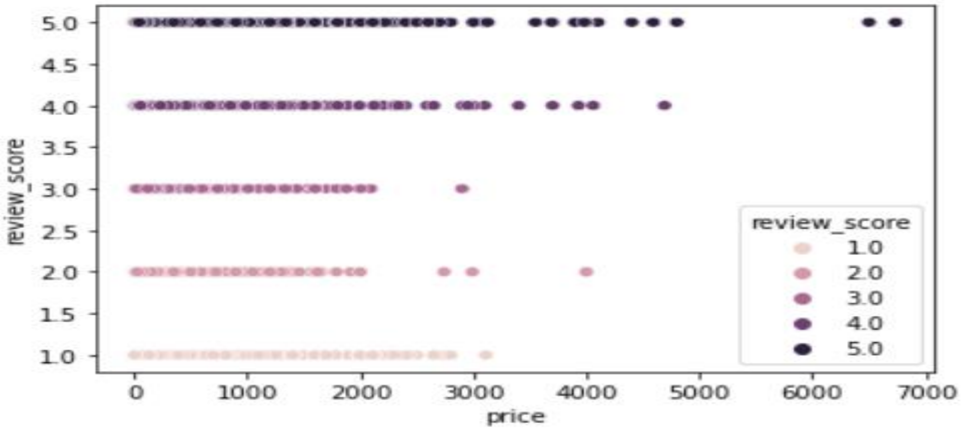
**1. Top 10 categories by sum of their price (category of product vs price of product)**



The total price of Health_beauty and watches_gifts purchased by customers in olist store is over 2.3M which generates highest revenue for the store.

**Correlations:**

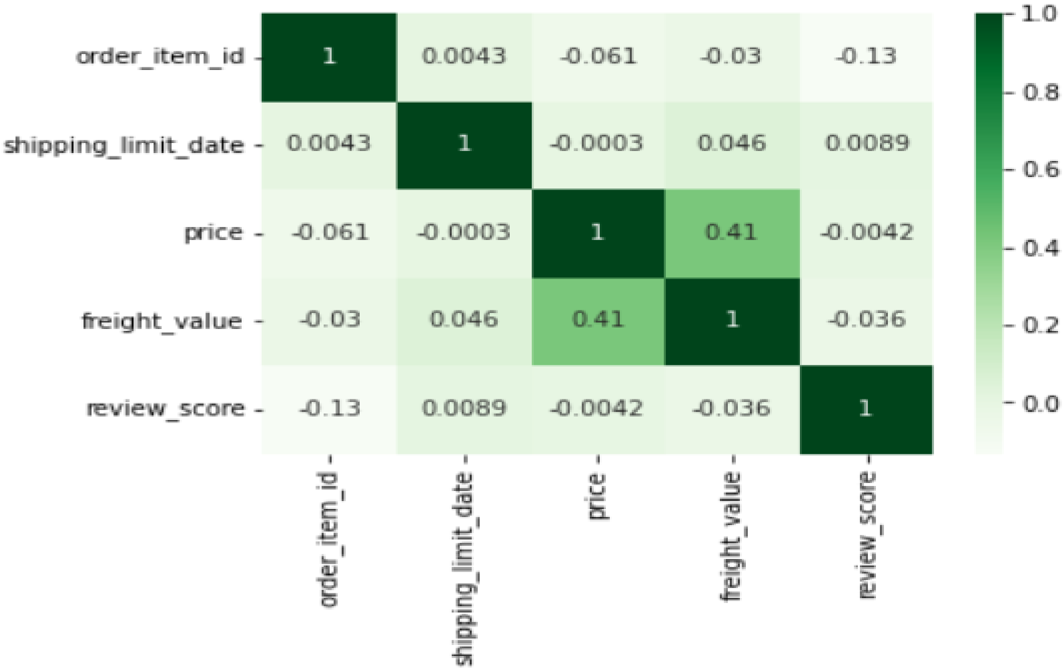
**1. Correlation between price and review score**



Most of the customers who had given their review score are the people who order value is less than 3000.Anything above 3000 maybe considered as outliers.

### 2. Correlation matrix of fact table

Correlation values ranges between -1 to +1 where -1 and +1 represents the strong correlation.



The above matrix shows correlation between key features that will be used in our analysis. It is clear from the matrix few variables are positively correlated and there is a negative correlation between few variables. The variable price and freight value has the second-best correlation.

## Data Transformation

**Merging Tables**: Data Merging was performed in Excel, Fact table was taken as a primary table to initiate merging. Second table is **olist_customers_dataset.csv**

**Step 1:**

First set of tables are fact table and olist_customers_dataset.csv.

Merging of the Fact table and the customer table was performed based on “customer id” column.

**Step 2:** The newly merged table is having the fact table information along with the customer details.

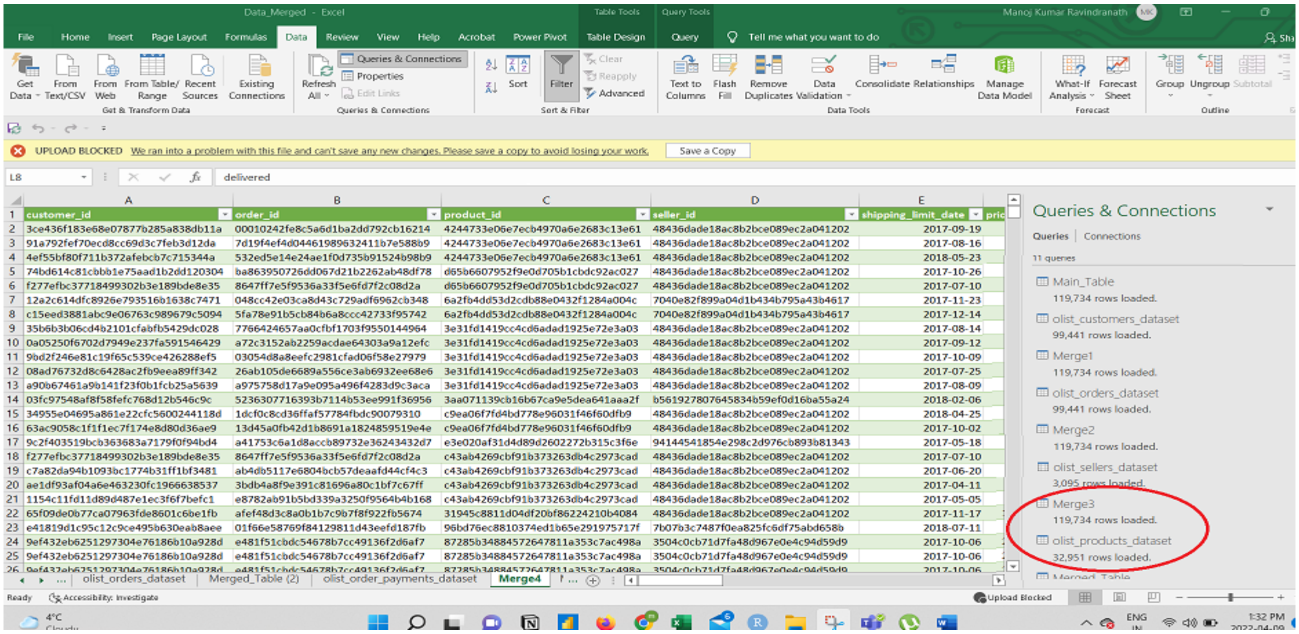
This table is now merged with the **Olist_orders_dataset.csv** based on order id column.

**Step 3:**

Naming the newly merged table as Merge2. This table is further merged with Olist_seller_dataset.csv based on “Seller id” Column.

**Step 4:**

The new table “Merge 3” is merged with Olist_products_dataset.csv based on the “product id“ column.



UPLOAD BLOCKED We ran into a problem with this file and can't save any new changes. Please save a copy to avoid losing your work. Save a Copy

L8 delivered

| | customer_id | order_id | product_id | seller_id | shipping_limit_date |
|---|---|---|---|---|---|
| 2 | 3ce436f183e68e07877b285a838db11a | 00010242fe8c5a6d1ba2dd792cb16214 | 4244733e06e7ecb4970a6e2683c13e61 | 48436dade18ac8b2bce089ec2a041202 | 2017-09-19 |
| 3 | 91a792fef70ecd8cc69d3c7feb3d12da | 7d19f4ef4d04461989632411b7e588b9 | 4244733e06e7ecb4970a6e2683c13e61 | 48436dade18ac8b2bce089ec2a041202 | 2017-08-16 |
| 4 | 4ef55bf80f711b372afebcb7c715344a | 532ed5e14e24ae1f0d735b91524b98b9 | 4244733e06e7ecb4970a6e2683c13e61 | 48436dade18ac8b2bce089ec2a041202 | 2018-05-23 |
| 5 | 74bd614c81cbbb1e75aad1b2dd120304 | ba863950726dd067d21b2262ab48df78 | d65b6607952f9e0d705b1cbdc92ac027 | 48436dade18ac8b2bce089ec2a041202 | 2017-10-26 |
| 6 | f277efbc37718499302b3e189bde8e35 | 8647ff7e5f9536a33f5e6fd7f2c08d2a | d65b6607952f9e0d705b1cbdc92ac027 | 48436dade18ac8b2bce089ec2a041202 | 2017-07-10 |
| 7 | 12a2c614dfc8926e793516b1638c7471 | 048cc42e03ca8d43c729adf6962cb348 | 6a2fb4dd53d2cdb88e0432f1284a004c | 7040e82f899a04d1b434b795a43b4617 | 2017-11-23 |
| 8 | c15eed3881abc9e06763c989679c5094 | 5fa78e91b5cb84b6a8ccc42733f95742 | 6a2fb4dd53d2cdb88e0432f1284a004c | 7040e82f899a04d1b434b795a43b4617 | 2017-12-14 |
| 9 | 35b6b3b06cd4b2101cfabfb5429dc028 | 7766424657aa0cfbf1703f9550144964 | 3e31fd1419cc4cd6adad1925e72e3a03 | 48436dade18ac8b2bce089ec2a041202 | 2017-08-14 |
| 10 | 0a05250f6702d7949e237fa591546429 | a72c3152ab2259acdae64303a9a12efc | 3e31fd1419cc4cd6adad1925e72e3a03 | 48436dade18ac8b2bce089ec2a041202 | 2017-09-12 |
| 11 | 9bd2f246e81c19f65c539ce426288ef5 | 03054d8a8eefc2981cfad06f58e27979 | 3e31fd1419cc4cd6adad1925e72e3a03 | 48436dade18ac8b2bce089ec2a041202 | 2017-10-09 |
| 12 | 08ad76732d8c6428ac2fb9eea89ff342 | 26ab105de6689a556ce3ab6932ee68e6 | 3e31fd1419cc4cd6adad1925e72e3a03 | 48436dade18ac8b2bce089ec2a041202 | 2017-07-25 |
| 13 | a90b67461a9b141f23f0b1fcb25a5639 | a975758d17a9e095a496f4283d9c3aca | 3e31fd1419cc4cd6adad1925e72e3a03 | 48436dade18ac8b2bce089ec2a041202 | 2017-08-09 |
| 14 | 03fc97548af8f58fefc768d12b546c9c | 5236307716393b7114b53ee991f36956 | 3aa071139cb16b67ca9e5dea641aaa2f | b561927807645834b59ef0d16ba55a24 | 2018-02-06 |
| 15 | 34955e04695a861e22cfc5600244118d | 1dcf0c8cd36ffaf57784fbdc90079310 | c9ea06f7fd4bd778e96031f46f60dfb9 | 48436dade18ac8b2bce089ec2a041202 | 2018-04-25 |
| 16 | 63ac9058c1f1f1ec7f174e8d80d36ae9 | 13d45a0fb42d1b8691a1824859519e4e | c9ea06f7fd4bd778e96031f46f60dfb9 | 48436dade18ac8b2bce089ec2a041202 | 2017-10-02 |
| 17 | 9c2f403519bcb363683a7179f0f94bd4 | a41753c6a1d8accb89732e36243432d7 | e3e020af31d4d89d2602272b315c3f6e | 94144541854e298c2d976cb893b81343 | 2017-05-18 |
| 18 | f277efbc37718499302b3e189bde8e35 | 8647ff7e5f9536a33f5e6fd7f2c08d2a | c43ab4269cbf91b373263db4c2973cad | 48436dade18ac8b2bce089ec2a041202 | 2017-07-10 |
| 19 | c7a82da94b1093bc1774b31ff1bf3481 | ab4db5117e6804bcb57deaafd44cf4c3 | c43ab4269cbf91b373263db4c2973cad | 48436dade18ac8b2bce089ec2a041202 | 2017-06-20 |
| 20 | ae1df93af04a6e463230fc1966638537 | 3bdb4a8f9e391c81696a80c1bf7c67ff | c43ab4269cbf91b373263db4c2973cad | 48436dade18ac8b2bce089ec2a041202 | 2017-04-11 |
| 21 | 1154c11fd11d89d487e1ec3f6f7befc1 | e8782ab91b5bd339a3250f9564b4b168 | c43ab4269cbf91b373263db4c2973cad | 48436dade18ac8b2bce089ec2a041202 | 2017-05-05 |
| 22 | 65f09de0b77ca07963fde8601c6be1fb | afef48d3c8a0b1b7c9b7f8f922fb5674 | 31945c8811d04df20bf86224210b4084 | 48436dade18ac8b2bce089ec2a041202 | 2017-11-17 |
| 23 | e41819d1c95c12c9ce495b630eab8aee | 01f66e58769f84129811d43eefd187fb | 96bd76ec8810374ed1b65e291975717f | 7b07b3c7487f0ea825fc6df75abd658b | 2018-07-11 |
| 24 | 9ef432eb6251297304e76186b10a928d | e481f51cbdc54678b7cc49136f2d6af7 | 87285b34884572647811a353c7ac498a | 3504c0cb71d7fa48d967e0e4c94d59d9 | 2017-10-06 |
| 25 | 9ef432eb6251297304e76186b10a928d | e481f51cbdc54678b7cc49136f2d6af7 | 87285b34884572647811a353c7ac498a | 3504c0cb71d7fa48d967e0e4c94d59d9 | 2017-10-06 |
| 26 | 9ef432eb6251297304e76186b10a928d | e481f51cbdc54678b7cc49136f2d6af7 | 87285b34884572647811a353c7ac498a | 3504c0cb71d7fa48d967e0e4c94d59d9 | 2017-10-06 |

Queries & Connections

Queries | Connections

11 queries

Main_Table
119,734 rows loaded.

olist_customers_dataset
99,441 rows loaded.

Merge1
119,734 rows loaded.

olist_orders_dataset
99,441 rows loaded.

Merge2
119,734 rows loaded.

olist_sellers_dataset
3,095 rows loaded.

Merge3
119,734 rows loaded.

olist_products_dataset
32,951 rows loaded.

Merged_Table

**Step 5:** Finally, to get the product name translation from the product_category_name_translation.csv the final table is merged based on product name column.

Now the Merged tables has all the columns from the fact table and the other tables.

Additionally, to get the payment information from the Payments table the existing merged table is further merged with the Payment table.

The final table has all the information available in the olist_customers_dataset.csv, Olist_order_items_dataset.csv, Olist_order_payments_dataset.csv, Olist_orders_dataset.csv, Olist_products_dataset.csv,Olist_seller_dataset.csv, product_category_name_translation.csv.

## Data Cleaning Using Python

### Step 1: Loading necessary packages

Pandas and numpy is imported into python. A new python file was created in the same location where the csv file is located.

### Step 2: Importing the csv file

Import dataset using pd.read_csv and save it as df1.Check all the columns have the appropriate data type.

### Step 3: Checking the null values

Check the null values in the dataset. Since the null values account to less than 5% of the data set it doesn't affect the final analysis, Hence, dropping all the null values.

The following columns has null values:

- review_score
- olist_orders_dataset.order_approved_at
- olist_orders_dataset.order_delivered_carrier_date
- olist_orders_dataset.order_delivered_customer_date
- olist_orders_dataset.order_estimated_delivery_date
- product_category_name
- product_name_lenght

**Step 4: Checking duplicate observation**

All the duplicate observation were found and removed.

**Step 5: Dropping columns**

Columns like product_length_cm', 'product_height_cm', 'product_width_cm', product_name_lenght were dropped because they have the least correlation with our target variable.

**Step 6: Renaming the columns**

olist_orders_dataset.order_status”: “order_status”,
olist_orders_dataset.order_purchase_timestamp”: “order_purchase_Time”,
olist_orders_dataset.order_approved_at”:”order_approved_at”,
olist_orders_dataset.order_delivered_carrier_date”:”order_delivered_carrier_date”
olist_orders_dataset.order_delivered_customer_date”:”order_delivered_customer_date”,
olist_orders_dataset.order_estimated_delivery_date”:”order_estimated_delivery_date”}

**Step 7: Creating custom columns**

Based on the analysis questions few custom columns like “**delivery_delay”, ”Days_to_Deliver”, “delayed”** have been created.
Where “delivery delay” has the information on difference between estimated delivery and the actual delivery.
Days to deliver: Difference between delivered date and order approved date.
Delayed: This column has information on delayed orders. This is a categorical column that has only two values called “Delayed” and “Not delayed”

**Step 8: Outlier Treatment**

Since we have only few and valid outliers in the numerical columns like price, freight_value, Payment_Value It doesn’t affect the output of the Analysis. Hence it doesn’t require any transformation

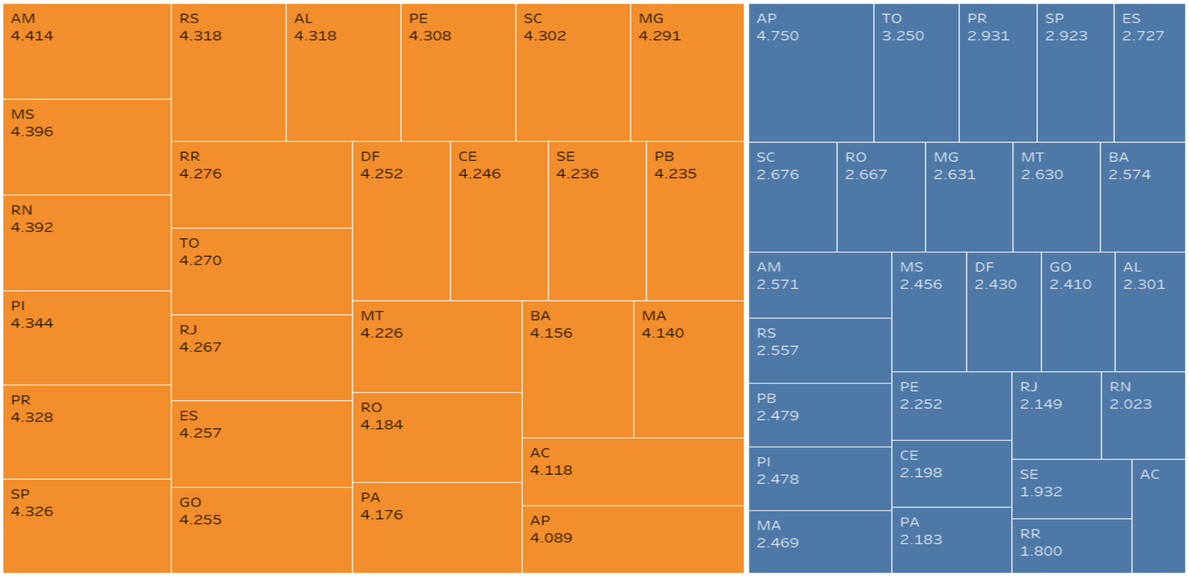
**Step 9: Exporting the into CSV file**

The cleaned data is exported as a csv file into the working location.

## Visualizations Using Tableau

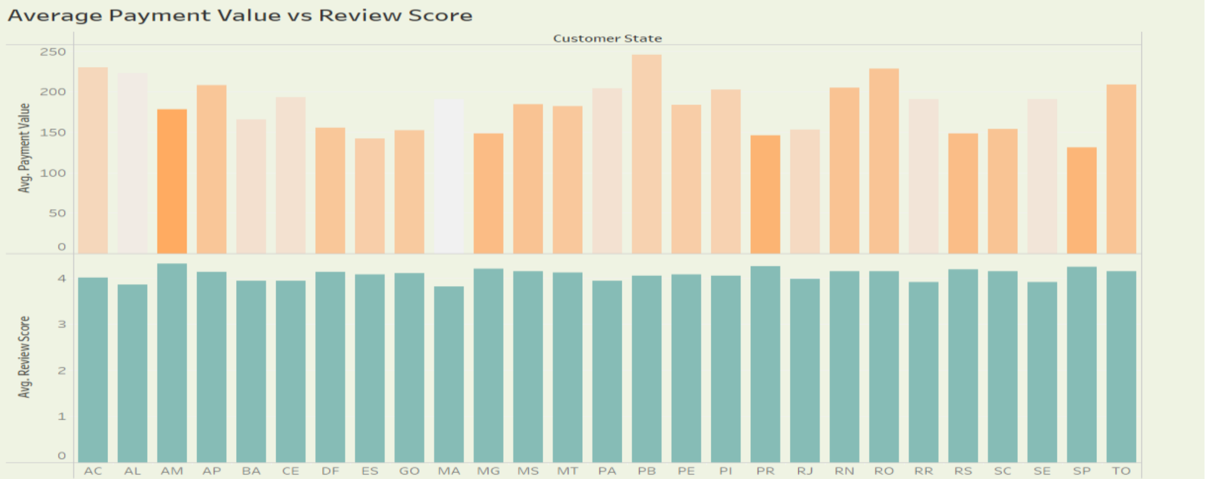
Answering the questions form problem statement using Tableau visualizations.

**What is the average review score based on delivery delay?**



The above graph was plotted to find if the delay in order influences the review score. The average review score for each state has been plotted based on whether the order was delivered on time or not. It is evident from the graph that the review score has been affected by the delivery delays. The orange portion of the graph which shows the order delivered on time has an average review score of more than 5 whereas the blue portion that represents the delayed order has an average review score of 2. this clearly shows the customer is not happy with the delayed orders.

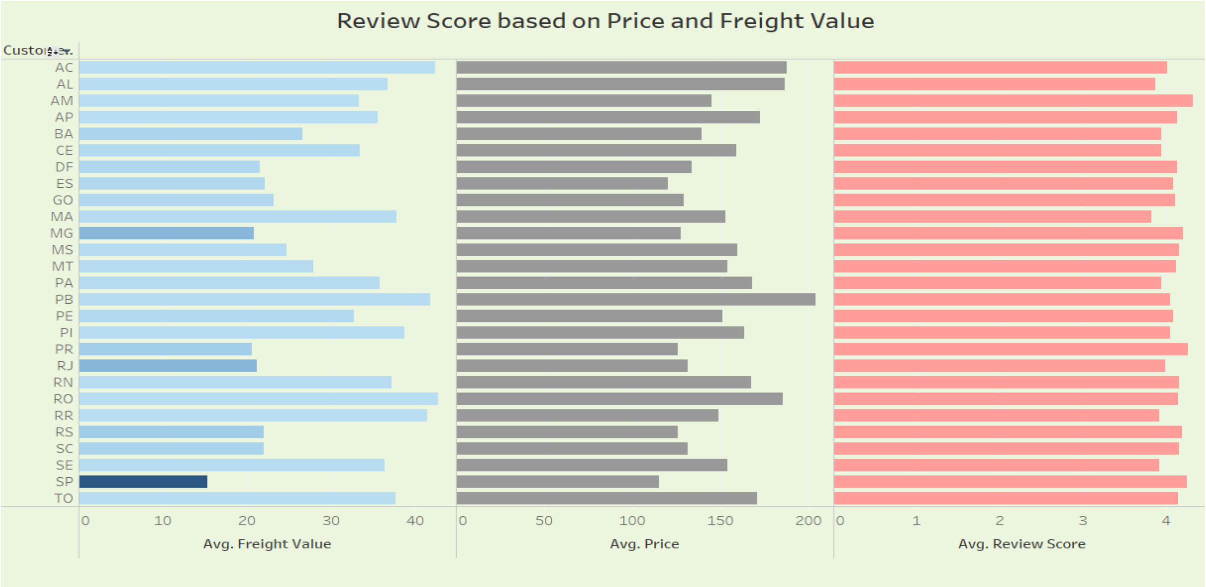
**Does the payment value affect the review score?**



The above graph was made with the intention to know if the payment value has any effect on the review score. The average payment value vs the review score has been plotted on province level.

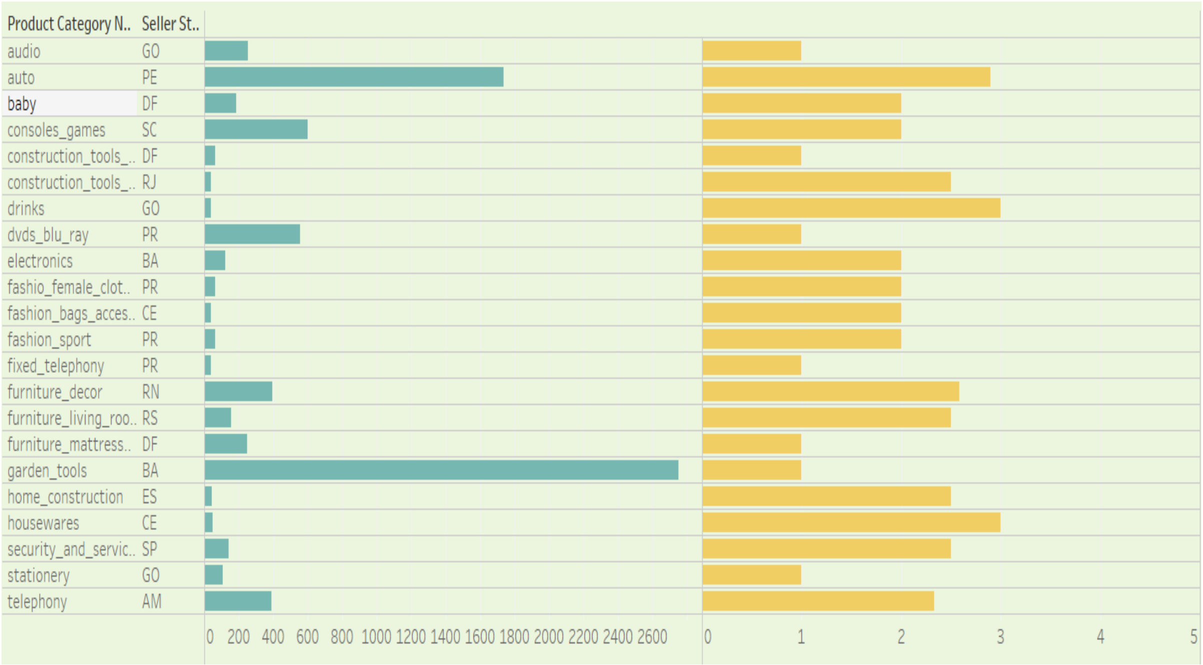
While almost all provinces have average review score over 4, states like AL, BA, MA have review score less than 4. By looking at the graph the payment value has no significant effect because there is no big correlation between these two variables. There are different averages for the same payment range.

**What is the review score based on Price and freight value?**



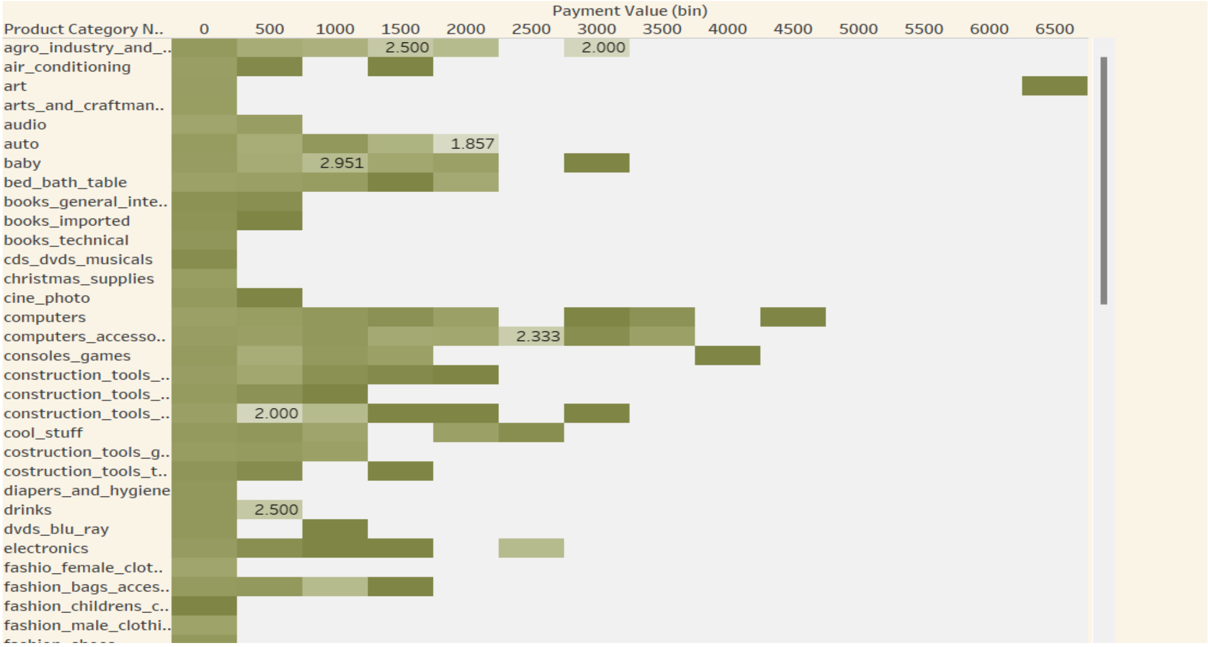
Breaking down the payment value in the previous graph into freight value and average price and plotting against review score. When trying to find the relationship between freight value and the review score it was found that the state with the least freight value has the highest average. And the review score is dropped below 4 in the states where the freight value is relatively higher. The “price” variable doesn’t seem to have much influence on the review score.

**Does the product from a particular seller state affect the review score?**



A visual for product category from a particular seller state is plotted against the average review score. Out of all product category and seller state combination around 20 combinations have review score less than 3. It is found that certain products from a state didn’t meet the minimum expectations of the customers. And one more interesting find is the state “PR” and “GO” have been mentioned many times in this graph. Hence products from these states needs a check.

**Does the price range of a particular product have an impact on customer behavior?**



The above graph shows the review score based on the product category and the amount spend by the customers. This graph gives a specific insight on the customer behaviour based on the amount they spend on a particular product. The darker the shade the higher the review. The light shades represent low review scores. Customers that spent around 1500 and 3000 on Argo industry and commerce product have given very low review scores. Similarly, Auto products that cost around 2000 didn't work well among customers. And Computer accessories that cost around 2500, Construction tools around 500, furniture décor around 1500 have got poor reviews.

**Conclusion:**

Customers are clearly not happy with the delayed orders, And the orders that were delivered on time is rewarded with high reviews.

Payment value has no direct influence in the review scores

The freight value which is included inside the payment value has an impact on the review scores. If the logistic partner has a way reduce this value, there can be a good change in the customer behavior.

There are states like "GO", "PR" that are selling products that are disliked by customers, a special attention has to be given to the product from these states.

A product in general has high review scores but drilling down with respect to the prices, products like agro, auto, baby, computer accessories have very poor review score , Hence this must be taken care of.

**References:**

Plotting of correlation matrix

How to Create a Correlation Matrix using Pandas - Data to Fish

Link to Dataset

https://www.kaggle.com/olistbr/brazilian-ecommerce?select=olist_customers_dataset.csv

Finding Relationships in Data

Python Guide: Finding Data Relationships with a Correlation Matrix | Pluralsight

Link to GitHub repository

https://github.com/Hackslash-DA/DB103-Data_Analysis_Project

Tableau Public Dashboard link

https://public.tableau.com/app/profile/elda3801/viz/103Final/Dashboard1