**ANALYTIC TOOLS AND DECISION MAKING**

**PROJECT SUBMISSION: 1**

**TITLE: OLIST STORE INSIGHTS**

A picture containing icon

Description automatically generated

**Group Members**

**Manoj Kumar Ravindranath**

**Sai Sowmith Komakula**

**Jerin Thomas**

**Elda Maria Thomas**

**Siva Krishna Unnam**

**Srilekhya Byreddi**

**CONTENT**

1. **PROPOSAL**

Introduction

Dataset Description

Problem Statement

Background

Insight

Motivation

Target Audience

Proposal

Approach

Analysis Questions

1. **EXPLORATORY DATA ANALYSIS**

Dataset Description

Data Segmentation

Dataset Collection

Data Imbalance

Inclusion Criteria

Preliminary Visualizations

Correlations

Schema

**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](https://pt.wikipedia.org/wiki/Mercado_Livre_(empresa)) , [B2W](https://pt.wikipedia.org/wiki/B2W) , [Via Varejo](https://pt.wikipedia.org/wiki/Via_Varejo) and [Amazon](https://pt.wikipedia.org/wiki/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 we will be trying to solve. 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 we should to able to figure out what worked well among customers.

Failing to retain the customers is one of the worst nightmares for any business owner. In today’s world for someone to choose a product there must be ample number of good reviews. Even though Olist dataset has a lot of good reviews it also has some bad review score. And this problem needs to address because that way the business can be improved and not take those bad reviews for granted.

**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. Online retailers hold no inventory and send orders directly to manufacturers, whereas traditional retailer stock fewer products due to shelf space. Traditional and online retailers also have different pricing strategies. Traditional retailers base their prices on traffic in their stores and inventory costs. Online retailers base their prices on delivery time.

Businesses can conduct e-commerce in one of two ways: fully online or online with a brick-and-mortar store. Online marketers can offer low prices, greater products selections, and high conversion rates. Customers prefer online markets if the products they buy are accessible online.

**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 and with this support we can draw insights.

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. They are always under constant pressure for coming up with new strategies especially with the competitors around them. Also, they should know one simply cannot write off the In- store experience without a genuine reason and so any E-commerce business should always give a customer a reason as to why one must choose online over offline.

**Motivation:**

As we all know data is one of the biggest drivers of making an analysis and predictions. With the amount of data, we have 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, we have all the necessary information that can be used to perform different analysis and with the time and date data we can also extend by predicting the trend.

**Target Audience:**

In this analysis, we are focusing on the needs of customers and small business owners. Since we are trying to solve a problem based on review score, customers will be someone getting benefited. Hence, they are the target audience.

**Proposal:**

As a part of this project, we will conduct a descriptive analysis of the dataset and find the correlation 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:**

* Since we have multiple csv files, 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.

**Analysis Questions:**

Is there any delay between the Estimated shipping date and the actual shipping date? If yes How may days? and how it is affecting the review?

What are the top 5 best selling categories?

What is the most popular payment method?

How are customers distributed across states?

What are all the variables that affect the customer review score?

What is the Total order amount by states?

What is the average cart price by states?

What is the Average delivery time by states?

What is the Seller order dispatch time across states?

**EXPLORATORY DATA ANALYSIS**

**Dataset Description:**

This is a Brazilian ecommerce public dataset of orders made at [Olist Store](http://www.olist.com/). 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.
* There are no missing values in the customers dataset.

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 (if an order has more than one item the freight value is splitted between items) |

* Mean of the price is 121 and standard deviation of price column is 184.
* Mean of freight\_value is 20 and standard deviation is 15.8.
* There are no missing values in any of the columns in the dataset.
* The outliers are found in price and freight\_value columns.

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 no missing values in this dataset.
* 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. |

* Mean and Standard deviation of the column review\_score is 4.09 and 1.35 respectively.
* There are no missing values in the dataset.
* There are no outliers in this dataset.

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 the each columns.
* Mean and standard deviation for product\_length\_cm is 30.8 and 16.9 respectively.
* Mean and standard deviation for product\_height\_cm is 16.9 and 13.6 respectively.
* Mean and standard deviation for product\_width\_cm is 23.2 and 12.1 respectively.
* 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 |

* There are 3095 unique sellers and most of them are from Sau Paulo.
* There are no missing values and outliers in this dataset.

product\_category\_name\_translation.csv

Translates the product categor*y* 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 |

* There are no missing values and outliers in this dataset.

**Note:** The outliers in the dataset are found by finding out the first and third quartile, then calculating the interquartile range (IQR), lower bound and upper bound.

**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.

**Dataset Collection:**

This dataset was originally provided by Olist, the largest department store in Brazilian marketplaces. Olist connects small businesses from all over Brazil to channels without hassle and with a single contract. Those merchants are able to sell their products through the Olist Store and ship them directly to the customers using Olist logistics partners. See more on our website: [www.olist.com](https://www.olist.com/).We found this dataset on Kaggle.com.

**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.**

Chart, histogram

Description automatically generated

Price of orders anything above 1000 is considered as an outlier. Each order can have multiple products to be shipped. The data from the graph is right skewed.

1. **Distribution of fright value.**

Chart, histogram

Description automatically generated

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.

1. **Distribution of payment value**

Chart

Description automatically generated

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.

1. **Top 10 best selling product categories**

Chart, bar chart

Description automatically generated

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.

1. **Most widely used payment method by customers**

Chart

Description automatically generated

Customers are mostly interested to pay their purchases by credit card.

1. **Payment installments**

Chart

Description automatically generated

Most of the customers choose to pay their payments for products in 0-3 installments by credit card.

**Bivariate Analysis**

1. **Top 10 categories by sum of their price**Chart

   Description automatically generated

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.

1. **Total number of orders per state**

Chart

Description automatically generated

It is clear to our eyes that most of the purchases done in olist store are by customers who lives in the state of Sao Paulo.

1. **Total payment by cities**

A picture containing text

Description automatically generated

More than 3 million of payment is done by the cities of Sao Paulo and Rio de Janeiro.

1. **Review score.**

Graphical user interface, chart, application

Description automatically generated

Customers who purchased their products from olist store are fulfilled and satisfied with the products and delivery time and gave mostly 5 and 4 as their review score.

**Correlations:**

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

Chart, scatter chart

Description automatically generated

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.

1. **Correlation matrix of fact table**

Chart

Description automatically generated

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.

**Schema:** We had modified our original dataset 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.

**Diagram

Description automatically generated**