Text

Description automatically generated

*OLIST E-commerce Report*

1. Assessing Summary
2. Clean Summary
3. Analysis & Visualization Summary
4. Clustering Summary
5. Clustering Analysis

**Graphical user interface, application

Description automatically generated1-Assessing Summary:**

* Number of tables used equal 7.
* There is a translation table for categories names.
* Data has no duplicates.
* Data have some missing values
  1. **Customers Table:**
* No missing values.
* Unuseful column (customer zip code prefix).
* About 99k records.
  1. **Items Table:**
* No missing values.
* About 112,650 records.
  1. **Payment Table:**
* About 103,886 records.
* No missing values.
  1. **Reviews Table:**
* 100k record
* There are missing values in 2 columns, but they are not useful columns.
  1. **Orders Table:**
* 99,441 records.
* Dates in this table has Wrong Data types.
* Some columns have missing values.
  1. **Products Table:**
* 32,951 records
* Include missing values (product\_category\_name, product\_photos\_qty, product\_weight\_g  
  ,product\_length\_cm, product\_height\_cm, product\_width\_cm)
  1. **Sellers Table:**
* No missing values.
* 3095 records.
  1. **Translation Table:**
* This table include the Brazilian categories names translated to English.
* 71 categories translated.
* There are 2 categories not translated.

**2-Cleaning Summary:**

This section includes the steps followed in cleaning the dataset

**2.1 Drop unuseful columns:**

There is no need to some columns in the dataset so Drop FUNCTION was used to delete them here are columns dropped names:

* customer\_zip\_code\_prefix
* review\_comment\_title
* review\_comment\_message
* review\_creation\_date
* review\_answer\_timestamp
* order\_delivered\_carrier\_date
* product\_description\_lenght
* product\_name\_lenght
* seller\_zip\_code\_prefix

**2.2 Join Tables:**

In this step the seven tables are joined using Merge function.

**2.3 Drop Null values:**

In this step the seven tables are joined using Merge function.

**2.4 Fix missing in translation table:**

In this step we added the two categories missed and translated it manually Then merge the translated categories with the entire dataset.

**2.5 Fix wrong data types:**

In this step we fixed wrong data types in our data.

**2.5.1 From object to datetime:**

* order\_purchase\_timestamp, order\_approved\_at, order\_delivered\_customer\_date, order\_estimated\_delivery\_date, shipping\_limit\_date.

**2.5.1 From object to int:**

* order\_item\_id, product\_photos\_qty , payment\_sequential , payment\_installments.

The result of cleaning step is a dataset called **olist\_data** consist of30 columns with about 114k row and with no missing data and right data types, so data now is ready for analysis and visualization.

**3-Data analysis & visualization Summary:**

* We have 114074 customer they are from 4073 city and 27 state.
* Number of customers increased over the years from 2016 to 2018 and it reached it's maximum at 2018.
* most popular city in our data is sao paulo.
* most popular state in our data is SP.
* We have 95129 reviews have score from 1 to 5.
* It seems that Olist products have good quality because of its high reviews.
* We have 95129 orders in our data.
* Most of customers order just one item per order.
* Max number of items in the same order was 21.
* Only 7 orders canceled, and others delivered.
* We have 31626 products divided into 73 categories.
* There are 4 payment methods about of 73% of people pay using credit card.
* Total sales are about 19.6 M.
* Most of customers choose 1 payment sequential.
* Most of customers pay freight less than 100.
* Payment sequential between 1 to 22.

**4-Clustering Summary:**

* cluster 3 have smallest recency but did not spend much money they may be new customers >> new\_customer.
* cluster 0 have largest recency and smallest monetary value I think they churned >> Risk to churn.
* cluster 1 have intermediate recency and largest frequency and largest monetary value >> active\_customer.
* cluster 2 have large recency , small frequency and small monetary value >> not\_active\_customer.
* cluster 4 have intermediate recency and intermediate frequency and intermediate monetary value >> intermediate\_customer..

**5-Clustering Analysis:**

* a lot of customers are risk to churn.
* new customers started purchase in 2017.
* in all clusters count of customers increase over time
* intermediate and active customers increased from 2016 to 2017 and did not change from 2017 to 2018
* risk to churn customers and not active decreased from 2017 to 2018
* not\_active customers have the largest payment value.
* risk to churn customers have the smallest payment value.
* new customers start pay from 2017 and increased in 2018.
* active customers have large payment value increased from 2017 to 2018.
* intermediate customers have fixed payment value from 2017 to 2018.
* most of customers pay using credit\_card.
* intermediate customers and active customers are the most who use debit\_card.