**Group – 3**

**Capstone Project- Interim Report**

**Brazilian Ecommerce Dataset**

**Industry Review:**

* Industry Review: Businesses have always tried to keep their customer base engaged and satisfied with the services provided by them. For remaining relevant in the industry, they need to incorporate the latest technological advances into their services.

More than a decade back, it was the internet which was completely new and various industries tried to leverage the capabilities of this technology that effortlessly acted as a medium of communication between various businesses and their customers.

In this decade, industries have started to provide services that are catered towards each client’s individual needs. For such services, they are required to leverage the power of artificial intelligence.

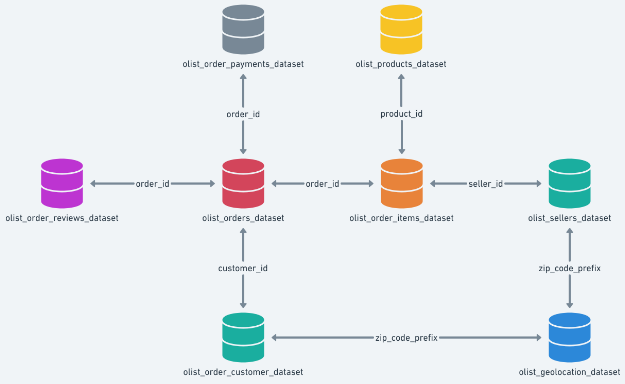
Comprehending the reviews of customers is very crucial for a business to be successful. Analyzing the reviews helps to properly discern the customer different preferences, likes, dislikes, etc. These extracted insights can then be used to improve customer service and experience.

* Business Problem: Here we are working on a Brazilian E-commerce dataset where we would perform some exploratory data analysis (EDA) and derive meaningful insights from our analysis. This project has multiple datasets containing different fields such as orders, payments, geolocation, products, products\_category, etc. but we would be mainly focusing on the customer orders dataset for our analysis.

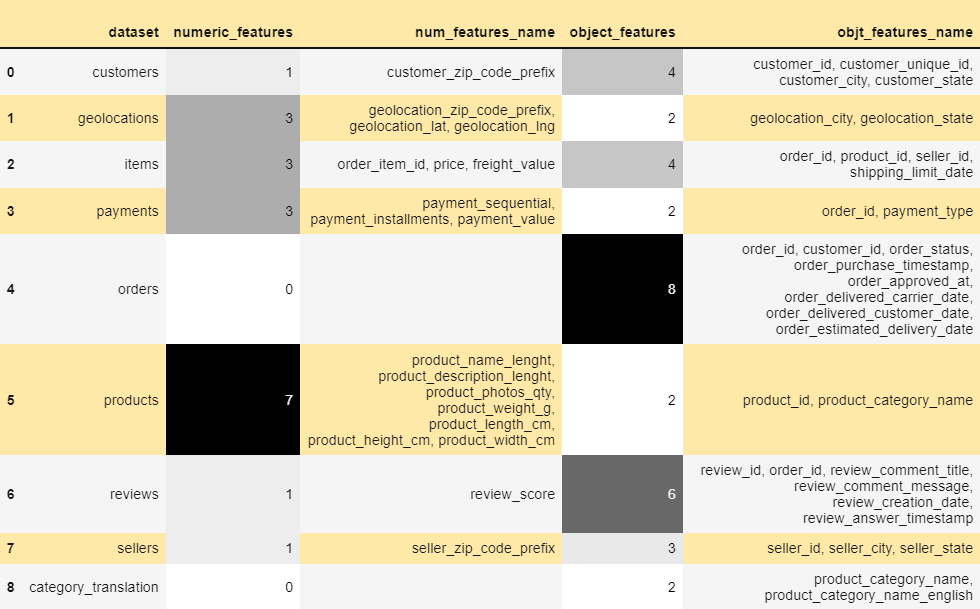
**Dataset and Domain:**

* Data Description:

**olist\_orders\_dataset**: This table is connected to 4 other tables. It is used to connect all the details related to an order.  
2) **olist\_order\_items\_dataset**: It contains the details of an item that had been purchased such as shipping date, price and so on.  
3) **olist\_order\_reviews\_dataset**: It contains details related to any reviews posted by the customer on a particular product that he had purchased.  
4) **olist\_products\_dataset**: It contains related to a product such as the ID, category name and measurements.  
5) **olist\_order\_payments\_dataset**: The information contained in this table is related to the payment details associated with a particular order.  
6) **olist\_customers\_dataset**: Details the customer base information of this firm.  
7) **olist\_geolocation\_dataset**: It contains geographical information related to both the sellers and customers.  
8) **olist\_sellers\_dataset**: This table contains the information related to all the sellers who have registered with this firm.



* **Variable categorization:**
* Numeric and Categorical

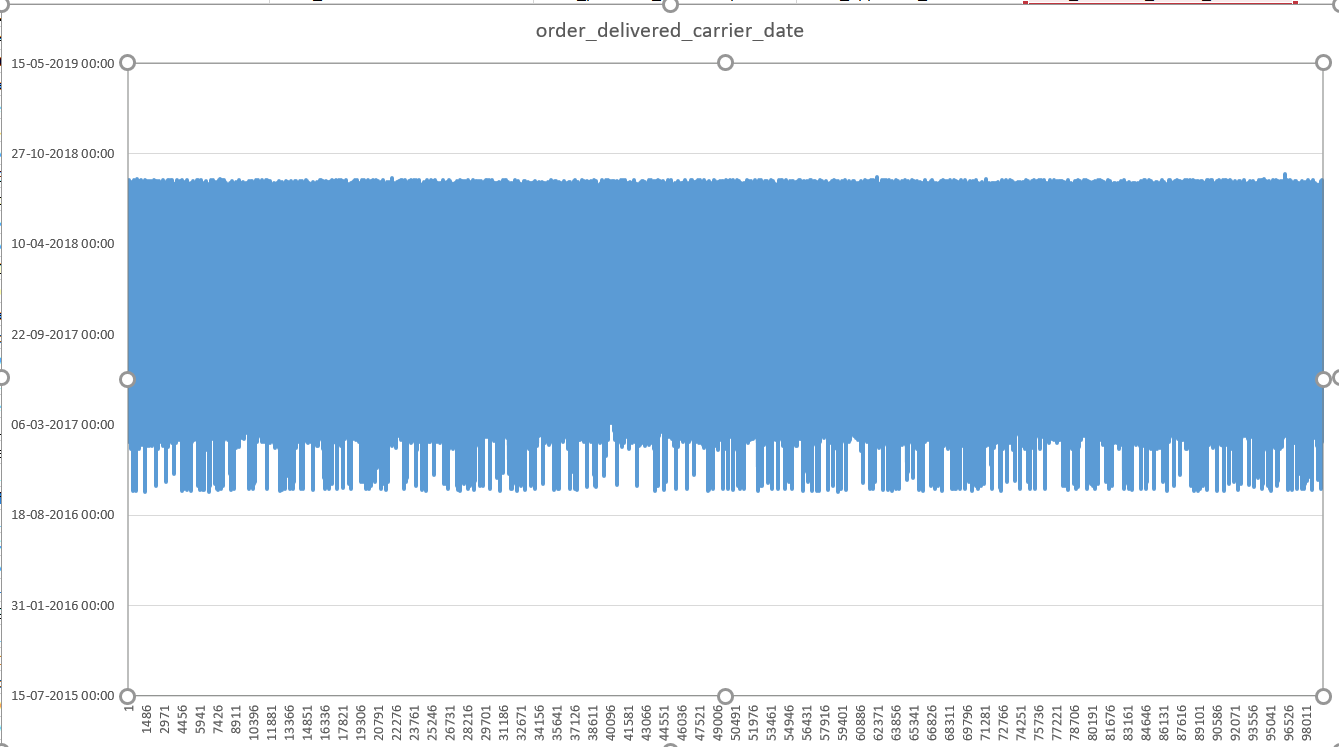


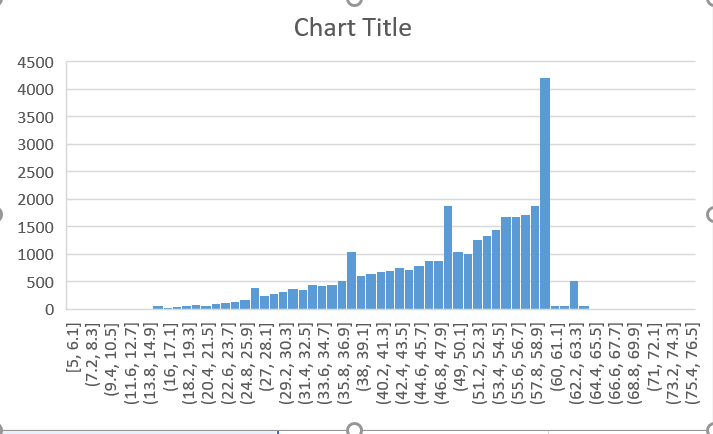
**Observations:**

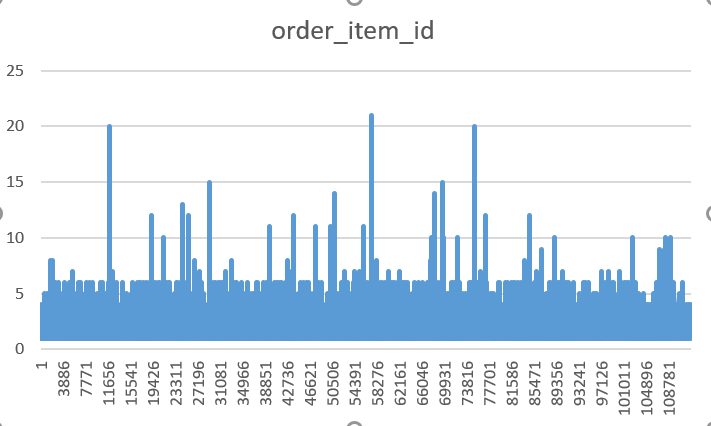
* Products dataset has maximum number of numeric features(i.e dtype: 'int16', 'int32', 'int64', 'float16', 'float32', 'float64')
* Orders dataset has maximum number of features of object dtype.
* We can also observe that all the timestamps are in object datatypes .So, we have to convert it into datetime type to do analysis on these features.
* **Pre processing of Data:**

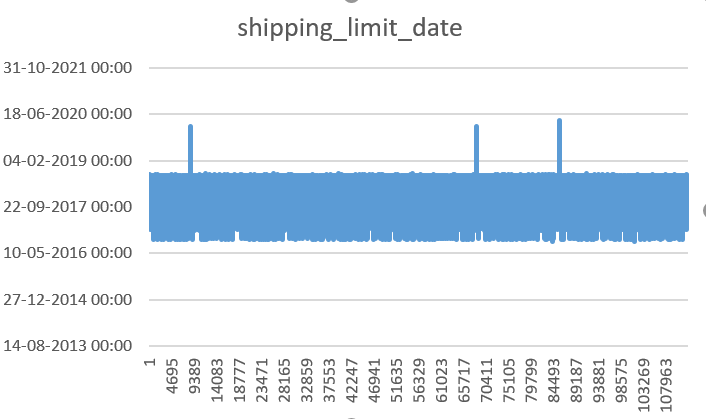
**Missing values**

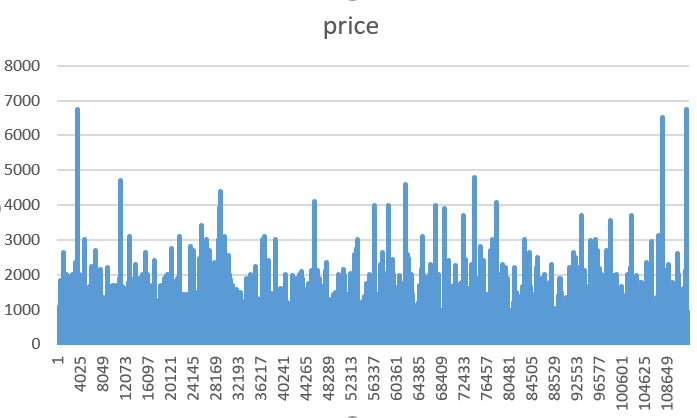
Different graphs showing missing values in the table.

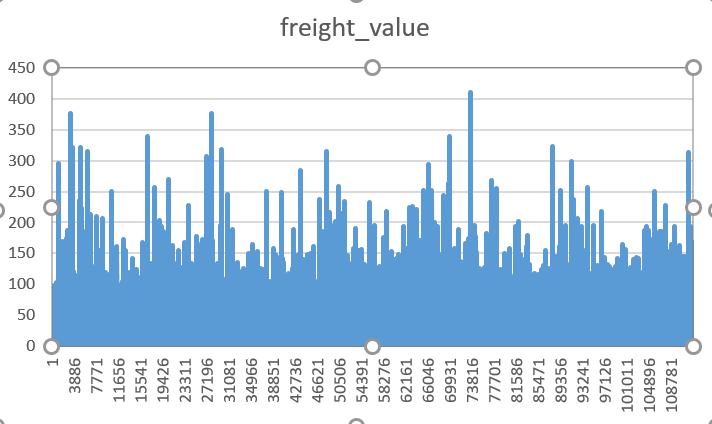












* **Project Justification:**

**Problem statement:**

To understand the product placement as per the consumer needs that would help the retailers understand the consumer demands.

**Project Outcome:**

After the completion of the project, we would be able to understand how the market basket analysis works. Also we will be predicting customer scores based on customer segmentation (recency, frequency, monetary value) for the company.

**Data Exploration (EDA):**

* **Univariate analysis:**









**Word Cloud :**



Initially the language of the review was translated from Portuguese to English. Then all the null values were dropped.

Finally, word cloud was created which shows the words in order of their frequency.

The word with highest frequency is shown bigger as compared to the word the lowest frequency which is shown smaller.

#### **RFM –Analysis**

**What is RFM?**

Behavioral segmentation by 3 important features:

Recency — number of days since the last purchase

Frequency — number of transactions made over a given period

Monetary — amount spent over a given period of time

**Observations:**

* There are three density plots of recency, frequency and monetary are plotted.From the first plot of recency we can observe that most of the users stayed with olist for long duration which is positive thing but order frequency is less.
* From the second plot of frequency most number of transaction or order is less than 5. from the third plot of monetary the maximum amount spend over the given very period is seems to less than 1500 approximately.

**REFERENCES**

[**www.kaggle.com**](http://www.kaggle.com)**/olistbr/brazilian-ecommerce**

[**www.nairaproject.com**](http://www.nairaproject.com)

[**www.theconversation.com**](http://www.theconversation.com)

[**www.papers.ssrn.com**](http://www.papers.ssrn.com)