O-List Ecommerce - Insight Report

### **HYPOTHESIS**

* **Firstly**, come up with a parameter to distinguish between the Good Reviews and Bad Reviews given by the customers.
* **Secondly,** from all the given features segregate the most important factors effecting the customers while giving Reviews.

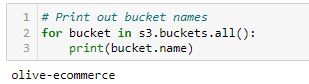
### **EXPLORATORY DATA ANALYSIS**

* **Extraction of dataset**

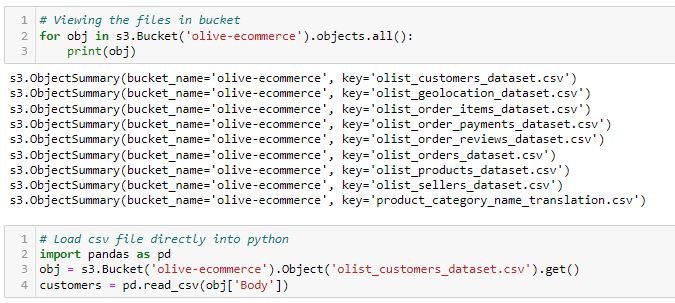
All the datasets were extracted from **Amazon Simple Storage Service** (Amazon S3) which is an object storage commonly used for data analytics application. To extract the datasets we had to install **Boto3** which is the Python Software Development Kit (SDK) that allows to manage AWS services in a programmatic way from Jupyter Notebook.



In Amazon S3 the datasets are stored in Buckets having unique names.



By using Boto3 library we can view the files in any bucket, load the files in jupyter notebook and also download the files.



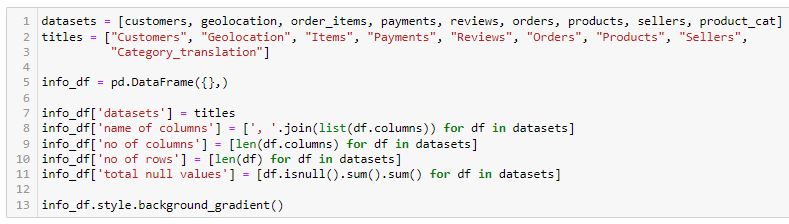


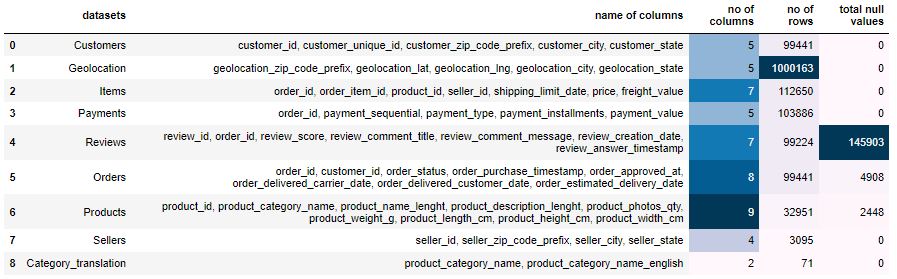
* **Data preparation**

The first step to be followed is to **read** the all the datasets.

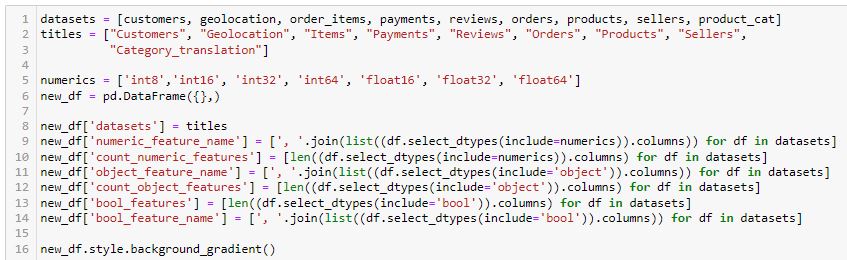


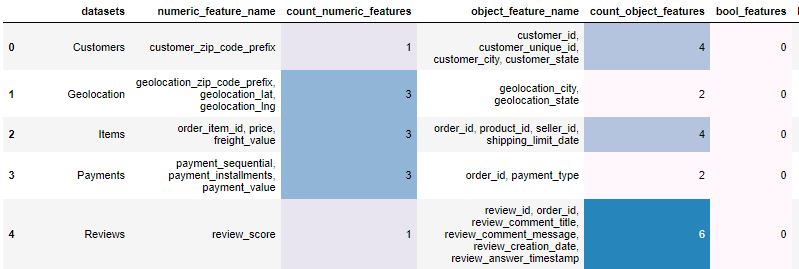
After reading the datasets, the next step is to check the distinct column names, no of columns, no of rows and the any null values if present in the columns.



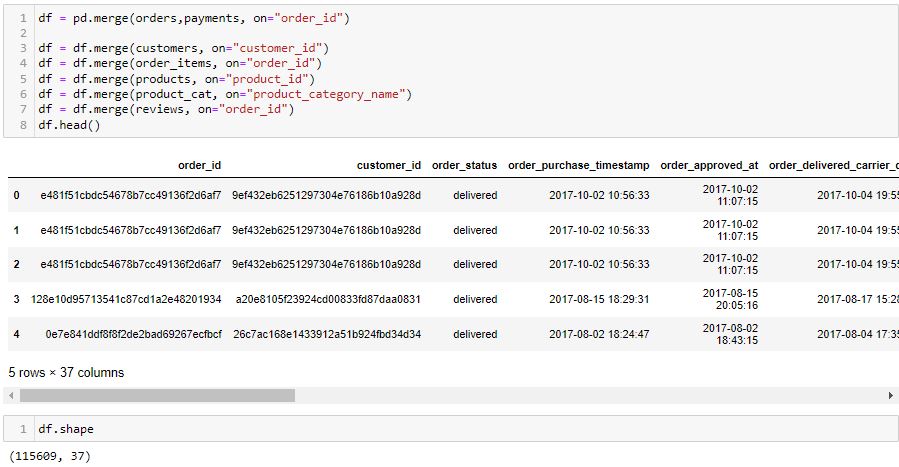


Checking the data types.





Merging all the datasets for the ease of working and checking the total no of rows and columns of the final merged dataset.



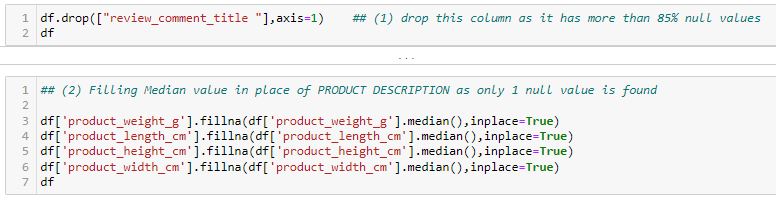
* **Handling missing data**

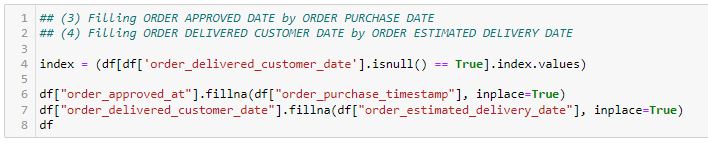
The merged final data have many null values. The maximum number of null values are present in the column *review\_comment\_title* which of object dtype.

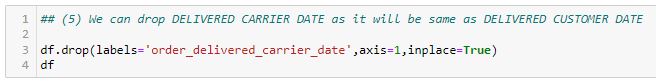
Columns like *order\_delivered\_carrier\_date*, *order\_approved\_at*, *order\_delivered\_customer\_*

*date* and *review\_comment\_message* also have missing values which are either replaced or

dropped. The following codes are shown below.





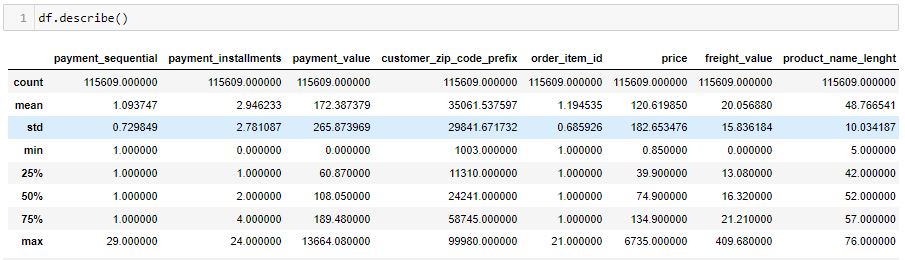




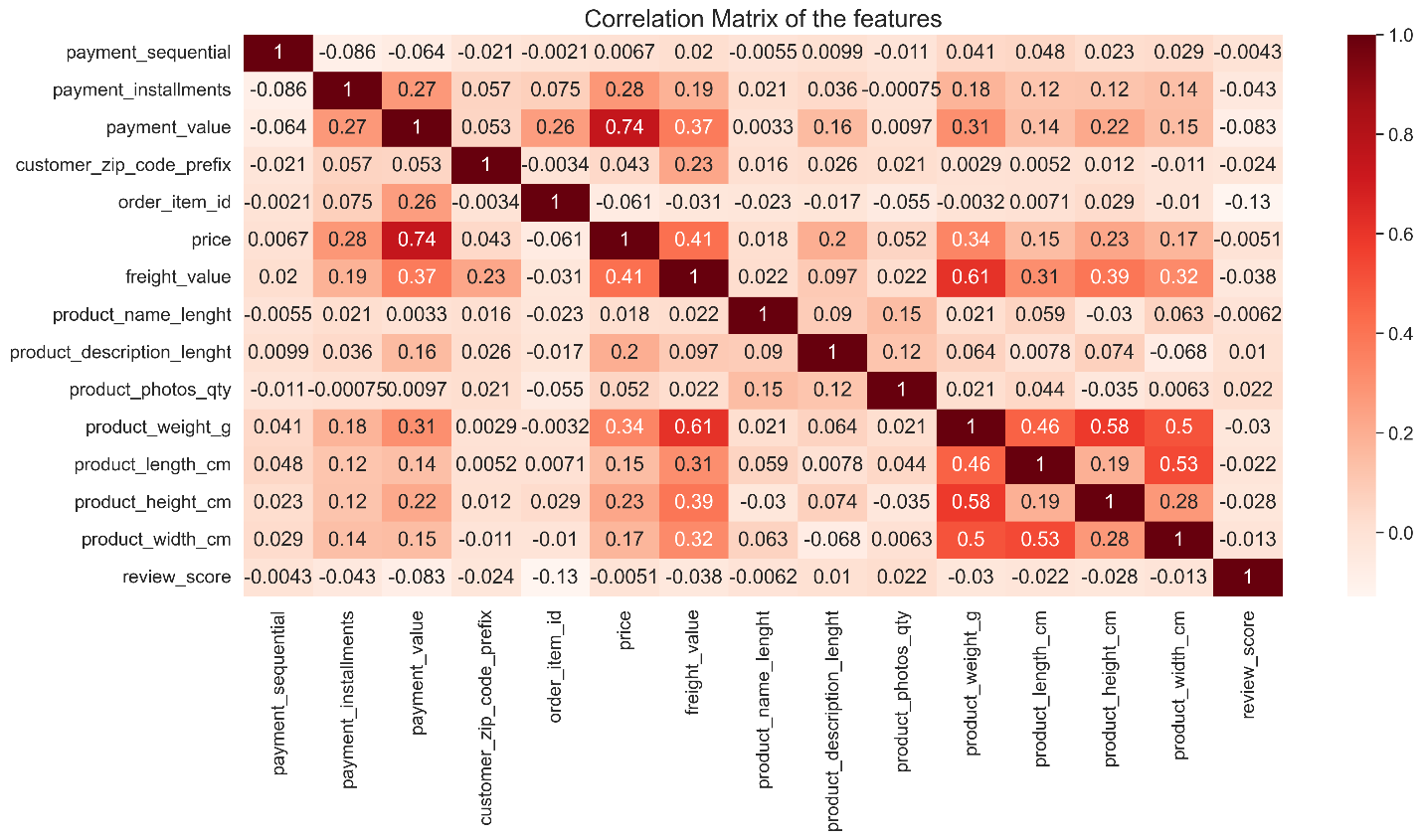
* **Initial Statistical data insights**

The final dataset after merging and cleaning has the following features:

* The merged dataset has 32 columns
* It has categorical features like *order\_status*, *customer\_state*, *product\_category\_name\_english* etc.
* One column name *review\_comment\_message* has text data that is in Portuguese.
* It has numerical features as well like *price*, *freight\_value*, *review\_score* etc.
* The maximum price of an order is 6735 and the max freight value is 410 Brazilian currency and the minimum price is 0.85 Brazilian currency.
* The maximum and average payement\_value made is 13664 and 173 Brazilian currency.
* Also we can observe the statistical values of other features like, mean, standard deviation, percentile etc.



* Correlation matrix



From this heat map we can observe:

* There is a strong positive correlation between ***price*** and ***payment\_value*** i.e. of 74%
* 61% correlation between *freight\_value* and *product\_weight\_g*