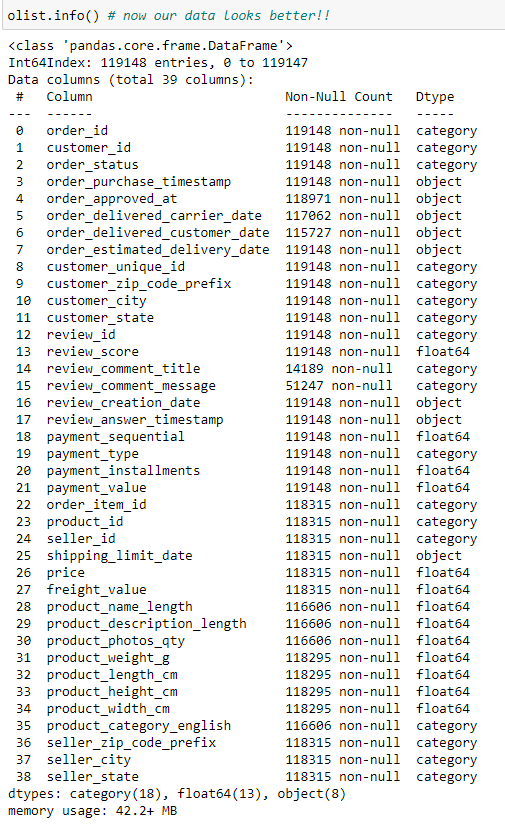
**Business Understanding [10 points]**

Describe the meaning and type of data (scale, values, etc.) for each attribute in the data file.



|  |  |  |
| --- | --- | --- |
| **Attribute** | **Value\_type** | **Description** |
| order\_id | category | order unique identifier (99,441 unique) |
| customer\_id | category | key to the orders dataset - each order has a unique customer\_id (99,431 unique) |
| order\_status | category | order status, 7-levels (shipped, canceled, invoiced, processing, approved, unavailable, delivered) |
| order\_purchase\_timestamp | datetime64[ns] | purchase initiation timestamp (9/4/16 – 10/17/18) |
| order\_approved\_at | datetime64[ns] | payment approval timestamp (9/15/16-9/3/18) |
| order\_delivered\_carrier\_date | datetime64[ns] | order posting timestamp when it was handed to the logistic partner (10/8/16-9/11/18) |
| order\_delivered\_customer\_date | datetime64[ns] | actual order delivery date to the customer (10/11/16 – 10/17/18) |
| order\_estimated\_delivery\_date | datetime64[ns] | estimated delivery date provided to the customer at the time of purchase initiation (9/29/16 – 11/11/18) |
| customer\_unique\_id | category | unique identifier of a customer (96,096) |
| customer\_zip\_code\_prefix | category | first five digits of customer zip code (14,994 unique) |
| customer\_city | category | customer city name (4,119 unique) |
| customer\_state | category | customer state name (27 unique) |
| order\_item\_id | category | sequential number identifying number of items included in the same order (1-21) |
| product\_id | category | product unique identifier (32,951 unique) |
| seller\_id | category | seller unique identifier (3,095 unique) |
| shipping\_limit\_date | datetime64[ns] | seller shipping limit date for handing the order off to the logistic partner (9/18/16-4/9/20) |
| price | float64 | item price (0.85-6,735) |
| freight\_value | float64 | item freight value (if an order has more than one item, the freight value is split between the items, scale: 0-409.68) |
| payment\_sequential | float64 | number of payment methods used by the customer (1-26) |
| payment\_type | category | method of payment by customer [4 levels: credit\_card, boleto, voucher, debit\_card] |
| payment\_installments | float64 | number of payment installments by customer (0-24) |
| payment\_value | float64 | transaction value (0-13664.08, note vouchers don’t count towards payment value) |
| seller\_zip\_code\_prefix | category | first five digits of seller zip code (2246 unique) |
| seller\_city | category | seller city name (611 unique) |
| seller\_state | category | seller state name (23 unique) |
| product\_category\_name | category | root category of product in Portuguese (73 levels) |
| product\_name\_lenght | float64 | number of characters extracted from the product name (5-76) |
| product\_description\_lenght | float64 | number of characters extracted from the product description (4-3992) |
| product\_photos\_qty | float64 | number of product photos published (1-20) |
| product\_weight\_g | float64 | product weight measured in grams (0-40425) |
| product\_length\_cm | float64 | product length measured in centimeters (7-105) |
| product\_height\_cm | float64 | product height measured in cemitmeters (2-105) |
| product\_width\_cm | float64 | product width measured in centimeters (6-118) |
| product\_category\_name\_english | category | product category name in English (71 levels – 2 need to imputed) |
| review\_id | category | review unique identifier (99,173 unique0 |
| review\_score | float64 | 1 to 5 rating given by the customer on a satisfaction survey (1-5) |
| review\_comment\_title | category | comment titles from the review left by the customer (4600 unique) |
| review\_comment\_message | category | comment message from the review left by the customer [note: 58% missing] (36,921 unique) |
| review\_creation\_date | datetime64[ns] | date satisfaction survey sent to customer (10/10/16-8/30/18 |
| review\_answer\_timestamp | datetime64[ns] | satisfaction survey answer timestamp (10/7/16 – 10/29/18) |

**Data Quality [15 points]**

Verify data quality: Explain any missing values, duplicate data, and outliers. Are those mistakes? How do you deal with these problems? Be specific.

**Missing Values:**

Rename the misspelled product\_name\_lenght and product\_description\_lenght to correct the spelling on length.

Initially our dataset had quite a few thousand missing values, but we made the business decision to only focus on orders that were completed based on 2 factors, (1) that the order status was ‘delivered’ and (2) there was a timestamp showing that the order was delivered to the customer. Once we reduced the initial dataset to only completed orders we found there were still some missing values:

|  |  |
| --- | --- |
| order\_delivered\_carrier\_date | 1 |
| order\_approved\_at | 15 |
| product\_weight\_g | 20 |
| product\_length\_cm | 20 |
| product\_height\_cm | 20 |
| product\_width\_cm | 20 |
| product\_name\_lenght | 1638 |
| product\_description\_lenght | 1638 |
| product\_photos\_qty | 1638 |
| product\_category\_english | 1638 |
| review\_comment\_message | 66764 |
| review\_comment\_title | 101973 |

1. **Order\_delivered\_carrier\_date:** The 1 missing value for the delivered to carrier attribute appears to be python not recognizing the datetime value in the cell, after much trial and error, it is just easier to delete the one record.
2. **Order\_approved\_at:** For the 15 order\_approved\_at missing values, we imputed those values by adding the difference between the average of the order\_purchase\_timestamp and the order\_approved\_at timestamp from the rest of the dataframe, that was about 10.5 hours.
3. **Product dimension information:** For the 20 values missing for weight and product dimensions, 19 of them had the same item number and the other 1 did not exist anywhere else in the dataset, so we deleted those since there was no way to impute them
4. **Product identity information:** For the 1,638 products missing product information such as name length (misspelled in the dataset from Kaggle), description length (also misspelled in the dataset from Kaggle), photo quantity and category, we performed a search of those same products across the entire dataset and did not find any other product\_id’s that match those same product\_id’s, in addition when we compiled the dataset, these records did not have category’s assigned to them. With this information, it made more sense to remove them from the dataset since we have so much data to work with already. Note: This can have an adverse effect if/when we predict pricing and we may need to add those records back in at a later date.
5. **Review Information:** due to the nature of shoppers leaving comments, based on the [*Online Review Statistics for 2021(Editor’s Choice)*](https://websitebuilder.org/blog/online-review-statistics/#:~:text=About%205%25%E2%80%9310%25%20of,shoppers%20specifically%20seek%20negative%20reviews.&text=About%205%25%E2%80%9310%25%20of,shoppers%20specifically%20seek%20negative%20reviews.), about 5-10% of consumers write reviews of e-commerce purchases, so having the bulk of records missing reviews is not a surprise, but we want to keep those records in the dataframe.

**Duplicate Values:**

* **order\_ids** and **customer\_ids:** 
  + Each order has an associated order\_id and customer\_id
  + An order may contain more than one item, so these order\_id and customer\_id will be duplicated to show the association of those items with the corresponding order
  + All values associated with an order\_id will be duplicated as well, this includes timestamps, unique\_customer\_id’s, etc.
* **unique\_customer\_id**: each customer has a unique\_customer\_id and that will be duplicated to show associated order\_id and customer\_id’s for customers with more than one order
* **timestamps**: timestamps will be duplicated everytime an order\_id/ customer\_id is duplicated due to an order having multiple items
* **freight\_value**: for orders with more than one item, the freight\_value is evenly (as possible) distributed across each item in the order
* **sellers and seller information:** sellers will be duplicated when they have more than one transaction or transactions with more than one item within the dataset

**Outliers:**

* 0 payment installments, that means the whole payment was made at the time of purchase
* $0 payment\_value is due to someone paying with a voucher; likewise, if someone pays a portion of their order with voucher and the rest with another payment form, only the other payment forms will show up under payment\_value
* 0g,2g –
  + there were 8 records with 0g weight and none of the product id’s matched to any other product\_id’s in the dataset, we can see that they have a product category of bed\_bath\_table and we can see the 4 smallest values in weight are: 0, 2, 25, 50, 53 grams, so 0 seems like an anomalie.
  + We also looked at the freight costs associated with these