# 

# 

Machine Language

Lab I- Visualization and Data Preprocessing

Submitted By

Justin Ehly

John Olanipekun

Helene Barrera

Feby Thomas Cheruvathoor

[**Business Understanding**](#_hhcou7ri4sif) **3**

[**Data Meaning Type**](#_3jvehw1zxhu1) **4**

[**Data Quality**](#_n2nk3cqvssoi) **8**

[Missing Values](#_7wvwbwchn9sz) 8

[Duplicate Data](#_puuwx6df31tb) 9

[Outliers Continuous Variables](#_j5b9rwrro7qw) 11

[**Simple Statistics**](#_qxbqhronq4nc) **16**

[**Visualize Attributes**](#_6qfbl1ns5ngl) **16**

[Broad, high level visualization of the dataset](#_6dooms4thldw) 16

[**Explore Joint Attributes**](#_yltfoi3cn4fh) **19**

[**Explore Attributes and Class**](#_yfa1utbb36a2) **22**

[Delivery Performance](#_ucgwsnh02uph) 22

[Average Delay in Delivery State Level](#_se5uo8qci3l4) 24

[Key Findings](#_4uvyh2p7h5lv) 25

[Approach to modeling delivery estimate](#_a6t3viy250d6) 27

[Predict Review Score](#_ul79kblus267) 30

[Approach](#_nel0omzbcdxy) 30

[**New Features**](#_rhjv94xf0xt9) **30**

[**References**](#_cmay31ceiafr) **32**

[**Appendix**](#_csohom6x7izl) **33**

[Data Meaning Type](#_65wyh81i607) 33

[Modeling](#_riljqqwn9q2w) 33

[Evaluation](#_vlibzlopgb6o) 34

[Success Criteria](#_vprmqqrqwo66) 34

# Business Understanding

Olist is the largest department store in Brazilian marketplaces, although it is relatively new to the eCommerce space having been founded in 2016. Much like Amazon and Etsy, Olist provides an integrated platform which connects the small,medium sellers to reach out to international marketplaces. Merchants are able to sell their products through the Olist Store and ship them directly to customers using Olist logistics partners. Olist provides a unique sales experience while improving their logistics performance, specifically when it comes to fulfillment options and Last Mile Delivery. It is important for Olist to ensure their logistics partners are performing effectively to improve customer review scores and gaining customer confidence in Olist eCommerce platform.

Olist contributed its past years sales order data to kaggle([link](https://www.kaggle.com/olistbr/brazilian-ecommerce)), dataset consists of roughly 100,000 orders from 2016 to 2018 and is multidimensional covering order information, consumer information, seller information, geolocation information, product attributes and customer reviews. The dataset will allow us to meet our stated business objectives. We will process the data using a combination of Python for data cleaning, mining, wrangling, exploration, feature selection and data modeling and will possibly employ cloud services for tasks such as running sentiment analysis. We are intended to bring out data insights which will help the Olist platform to the next level.

**Logistic Advancement:** There are many factors that contribute to a customer’s review score, but the majority of our data set is focused around the logistics side of eCommerce. We have detailed time stamps through each stage of purchase and delivery, as well as geographical information which will allow us to dig into how to make improvements to shipping time and estimated shipping time. Our success can be measured by improvements to the estimated delivery time and a decrease in bad reviews that mention shipping.

**Improving Customer Satisfaction:** Based on the attributes available in the data set, we decided to approach the dataset from the standpoint of working for Olist to help improve customer satisfaction, which in turn likely has a strong relationship with various logistics attributes. To that end, we will focus on understanding how factors like price, freight cost and estimated delivery time influence each other.

**Accurate Price Predictions:** Along the same lines, we will choose a specific product category and look at trends for products in that category for the past 3 years to understand seasonal influences on price and total sales. We can also identify patterns in customer purchases based on the day of the week and major holidays and festivals observed in Brazil. This will help ensure that Olist is prepared for fluctuations in the use of their site, allowing them to adjust their logistics partners and stock accordingly. A major success factor will be to obtain accuracy of at least 85%, precision of at least 80%, sensitivity of at least 85%. These values are subject to review, contingent upon exploratory data analysis.

# Data Meaning Type

Olist supplemented 120MB of data and the high level data schema is depicted below. (Figure 2A)

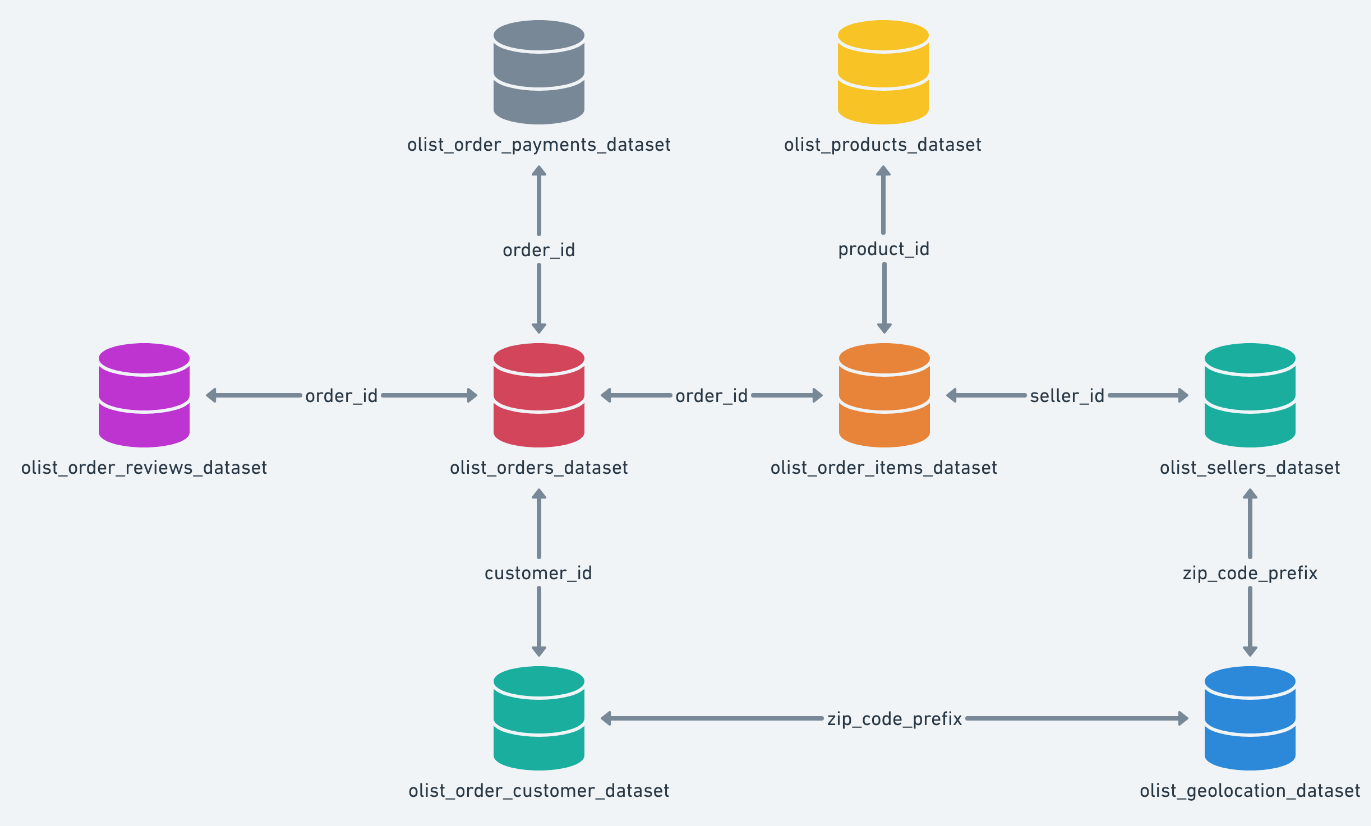


Figure 2A

Dataframe Info and the attribute description is listed below (Figure 2B)

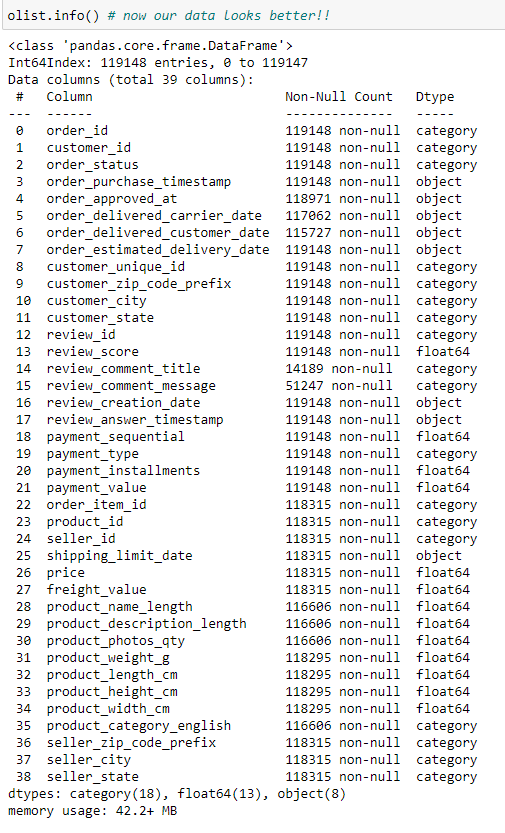


Figure 2B. Dataframe Info and the attribute description

More detailed description of the dataset.

|  |  |  |
| --- | --- | --- |
| **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

## Missing Values

* Rename the column names 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. Order status was ‘delivered’
2. 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 categories 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 Data

* **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 every time 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

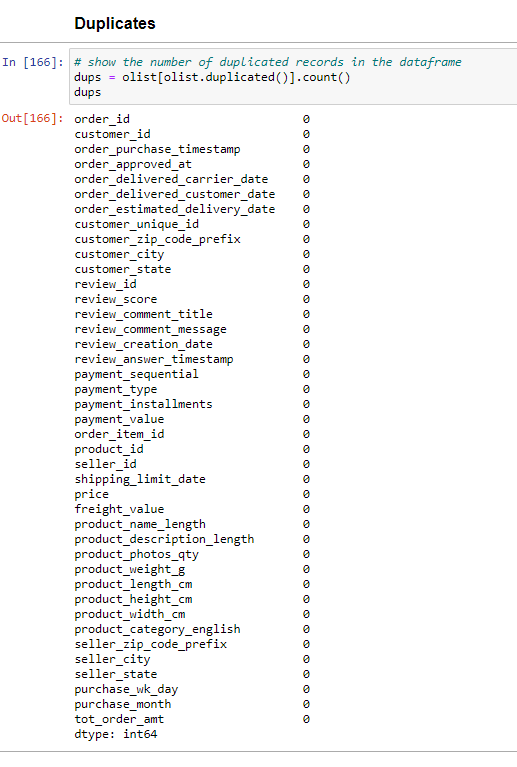


Figure 3A. Chart showing no duplicates in the dataset.

## Outliers Continuous Variables

* We will start exploring outliers by visually inspecting boxplots that can quickly show us what may appear to be outliers or anomalies with the continuous variables. Then we will address each variable in order below starting with review\_score and finishing with tot\_order\_amt. (Figure 3B)

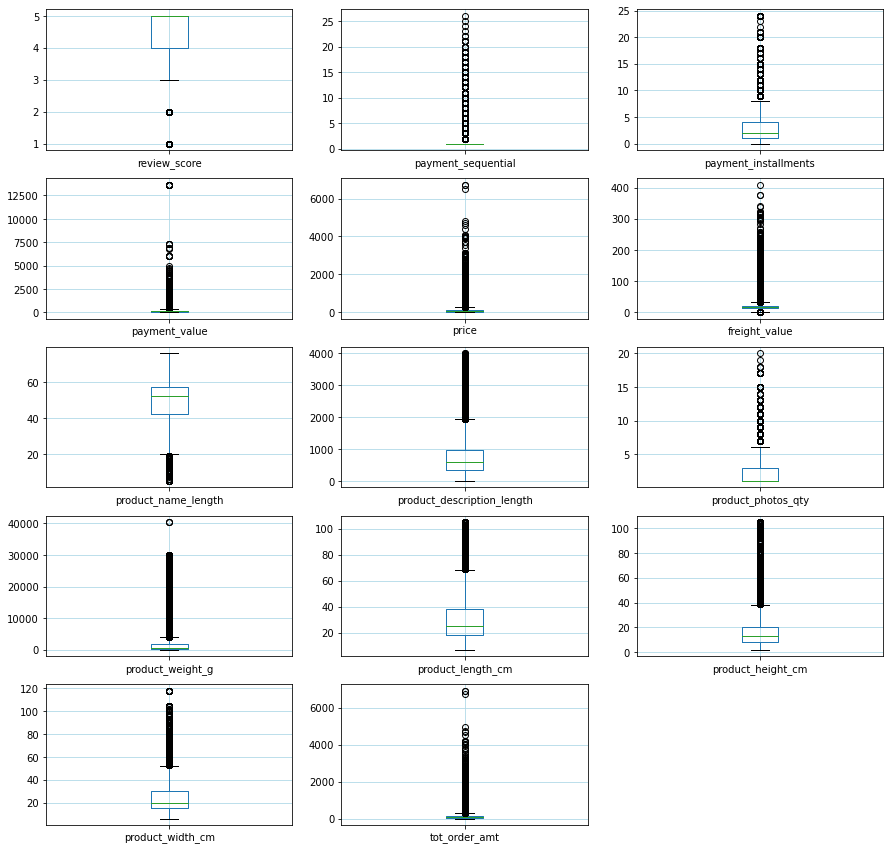
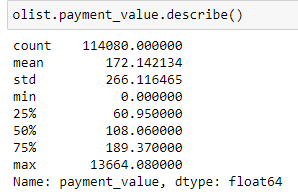


Figure 3B. Boxplots to visually show where they may be outliers in the continuous attributes.

* **Review Score:** Knowing that this range is from 1-5, we don’t consider these low scores to be outliers
* **Payment Sequential:** this means how many forms of payment a customer used, the range is 1:26, sounds crazy, but there are 11 customers that used more than 20 payment types - looking at an individual sample, one customer used vouchers or small payments less than $2 each that added up to the total amount (price + shipping) of $40.85
* **Payment Installments:** 425 customers made more than 10 payment installments based on an average price of $251.39
* **Payment Value:** Payment Value = price+shipping-(any vouchers). (Figure 3C)



**Figure 3C. Payment Value statistics**

* **Price:** Price has a range of R$0.85 : R$6,735 so it appears there are some expensive things for sale in the marketplace (Figure 3D)

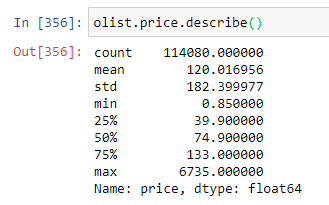


Figure 3D. Price Statistics

* **Freight Value:** 
  + It makes sense that freight value has a relatively low 25% (R$13.08) : 75% (R$21.19) range when you compare it to the distribution of the price variable.
  + There are some items that cost a lot more to ship, but aren’t shipped with the same frequency as the lower priced items and they fall into categories such as health and beauty, construction tools, housewares, baby and industry commerce and business
  + The last friday of April is free shipping day in Brazil, so that is why there is a minimum of R$0.

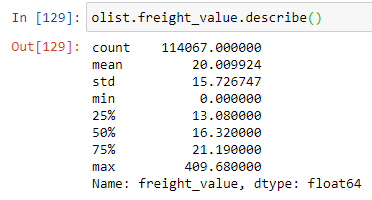


Figure 3F. Freight Value statistics

* **Freight Value** and **Product Weight** are going be correlated just based on general knowledge of shipping: (Figure 3G)

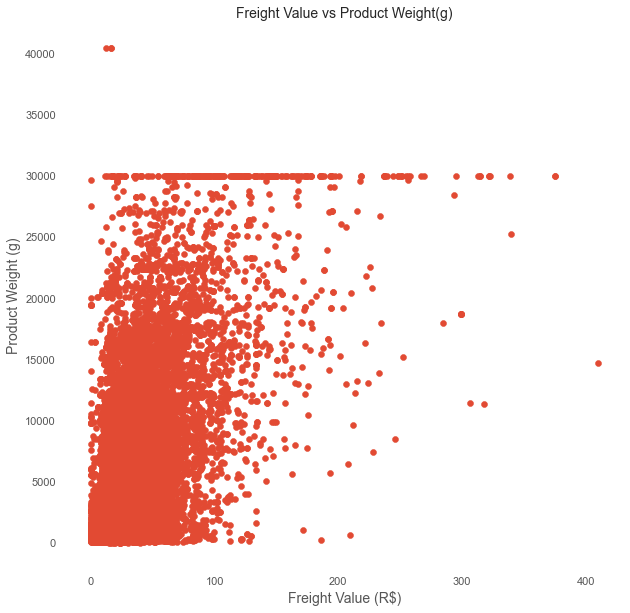


Figure 3G. Correlation between Product Weight and Freight Value

* **Product Height:** there are over 6,500 records with product height above 40cm, and 134 records with product height above or equal to 105cm, so we are confident these are correct
* **Product Width:** is very similar to product length in terms of stats (Figure 3H)

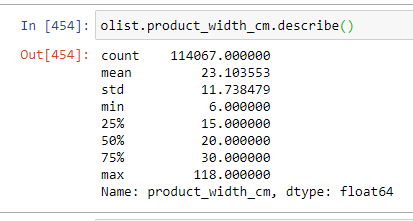


Figure 3H. Product Wight statistics

* + - Based on the boxplot where outliers seem to start above 50cm, there are nearly 2,900 records with widths over 50cm so they aren’t considered outliers.
    - At the max of 105cm width we show 13 records, so again we are confident these measurements are accurate
* **Product Name Length** and **Product Description Length** are descriptors of the products and populated by the sellers, they can be as detailed as they want or not
* **Product Photos Qt:y** is the number of photos associated with a product, again this is entirely up the seller depending on how descriptive they want to be
* **Product Length:** shows some possible outliers above 60cm, but it turns out there more than 6,000 records with length over 50cm. Also a the longest length of 105cm there are 311 rows, so we are confident these measurements are accurate (Figure 3I)

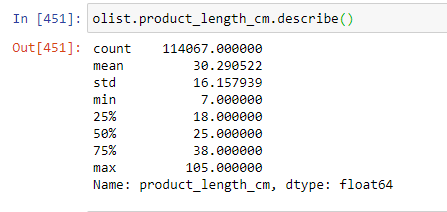


Figure 3I. Product Length statistics

* **Product Weight:** as charted above in the boxplot matrix does something interesting around 30k grams, it just seems to cutoff
  + The smallest weights are: 0, 2, 25, 50 and 53 grams
  + There are 13 records with a weight of 0 or 2 grams and when we look at the means of the items we get
    - Price: R$135.62
    - Freight Value: R$22.92
    - Height: 19.62cm
    - Width: 38.46cm
    - Length: 22.69cm
    - Categories: best\_bath\_table, funiture\_decor, stationary
      * Note: while stationary may weigh 2g, we are suspicious that all of these are mistakes so we are going to remove them from the dataset (Figure 3J)

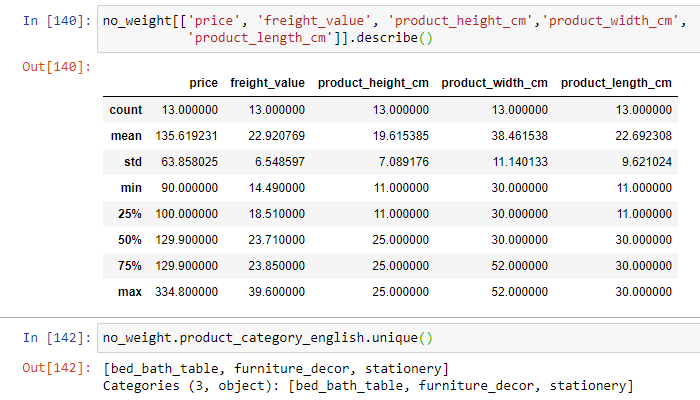


Figure 3J. Statistics for items that weigh 0 or 2grams.

* + The other numbers that jump out are the gap between about 30kg and 40kg,
    - There are 312 records with product weights above 29kg
    - There are 3 records with product weights above 40kg, they are all in the bed\_bath\_table category and that sounds reasonable
* **Total Order Amount:** total order amount = price + freight value (shipping), so it is not surprising that this column follows the price column very closely and based on the findings that there aren’t any outliers in price we are confident this attribute is ok as is.

# Simple Statistics

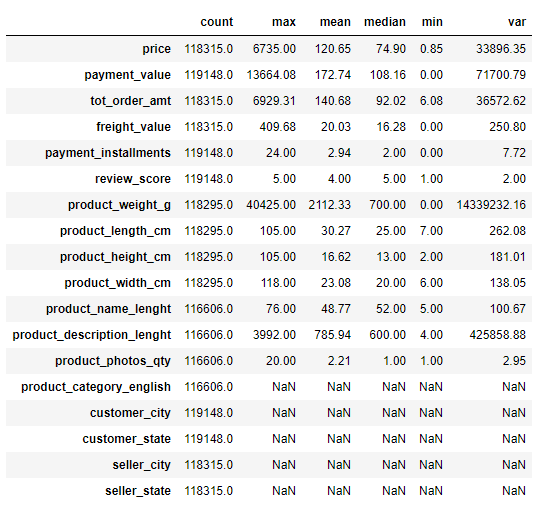


Figure 4A. Simple statistics for the dataset.

# Visualize Attributes

## Broad, high level visualization of the dataset

Below is another image of the boxplots of the data distribution for continuous attributes in the data set as we find that most are heavily skewed towards the top. (Figure 5A.)

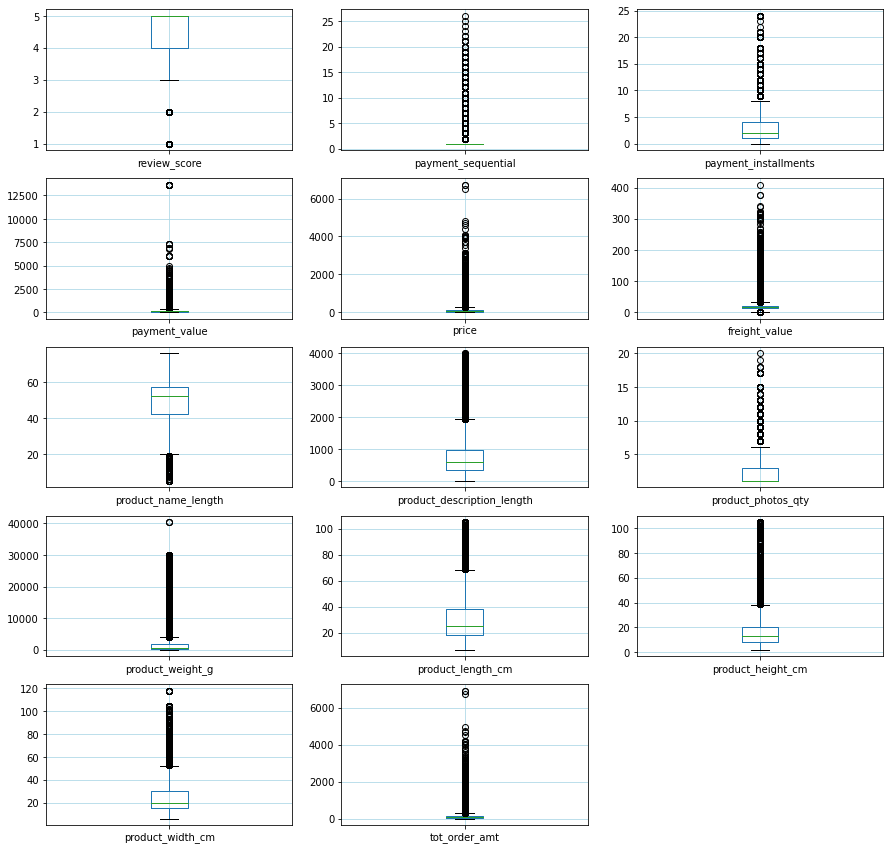


Figure 5A. Boxplots of the continuous attributes of the dataset.

Of particular interest are the skewed distribution shown for price related attributes (i.e. payment\_value, price, feight\_value). Figure 5B further displays the distribution as a histogram, showing strong skewness in the price data. The modeling tasks will strive to minimize the effect of this non-normality by scaling and/or transformation.

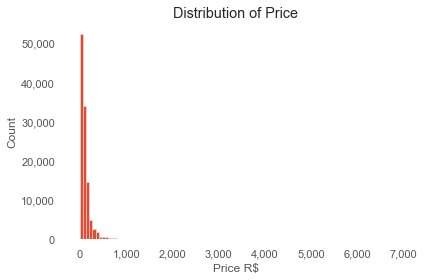


Figure 5B. Left skew of the price distribution.

A bar chart plot of review scores (Figure 5C) indicates that most of the sampled customers provided very good reviews (5). This is a key finding for understanding customer satisfaction and how to improve on it because we can drill into what constitutes a high score vs a low score.

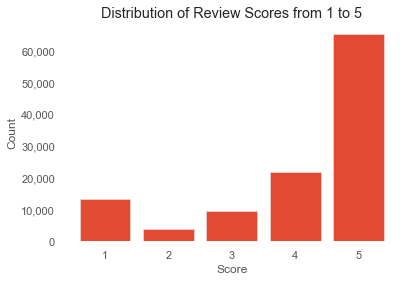


Figure 5C. Distribution of Review Scores

We see in Figure 5D. That the distribution of Freight Value is very similar to the distribution of price with a heavy left skew, suggesting that we should consider transforming it.

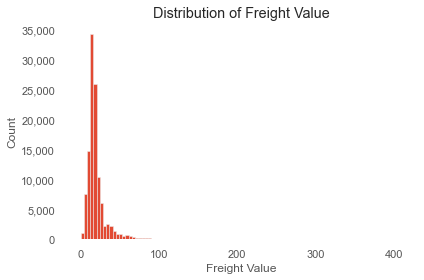


Figure 5D. Distribution of Freight Value.

We see that the majority of payments made from 2016-2018 on olist were by credit card, followed by boleto (a direct payment method like a bank wire), then by voucher and debit card.

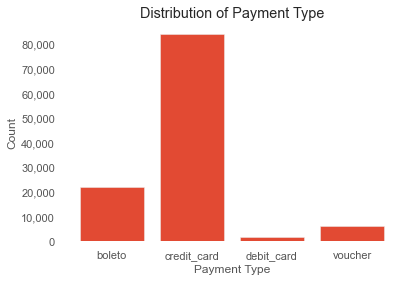


Figure 5F. Count of Payment Types

We also want to explore how many products are in each category to give us an idea of what the biggest or possibly more diverse categories are. This will help us later on determine the best category to work with for our accurate price predictions. The largest category is bed\_bath\_table with 11,808 products, followed by health\_beauty with 9,814 products and third is sports\_leisure with 8,791 products. The smallest 3 categories in order from smallest to largest are security\_and\_services (2), fashion\_childrens\_clothes (7) and pc\_gamer (9).

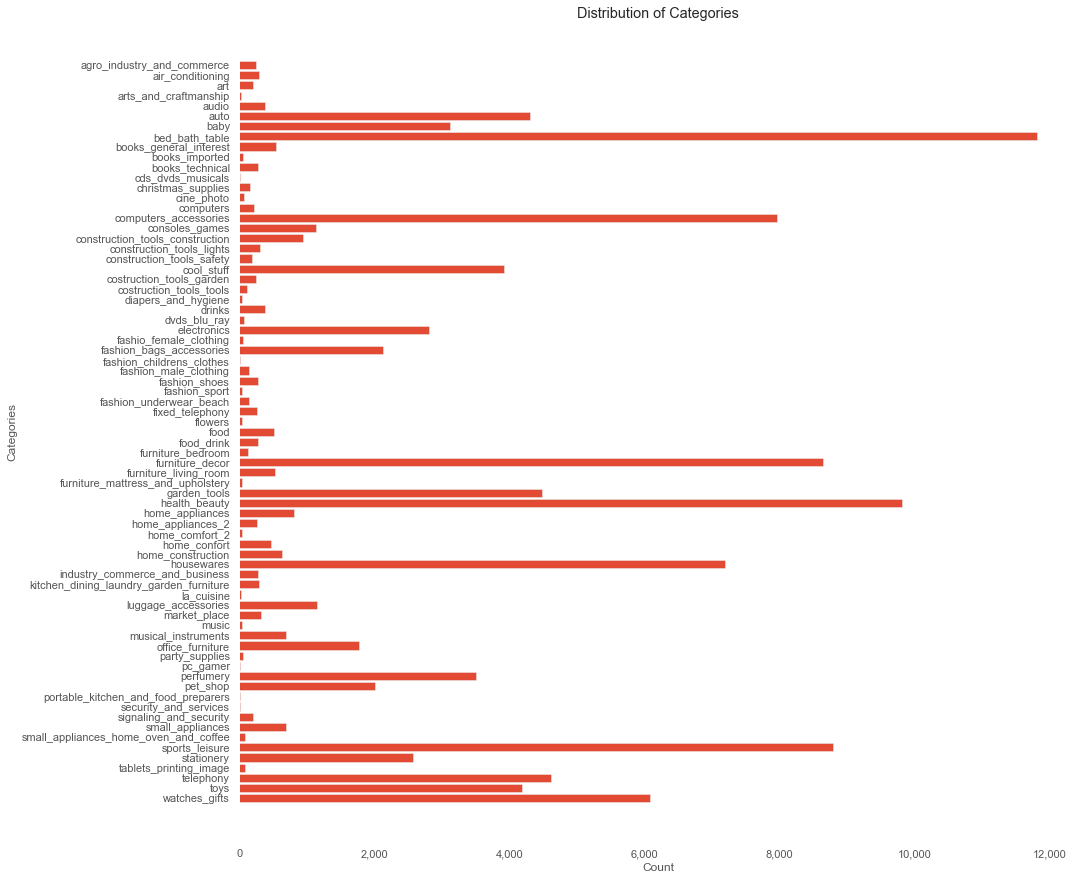


Figure 5G. The number of products by category in the dataset.

# Explore Joint Attributes

Figure 6A below shows a correlation matrix for the raw numerical attributes in the Olist dataset. The best relationships that can be useful (i.e. correlation coefficient > 0.15) for the project objectives are:

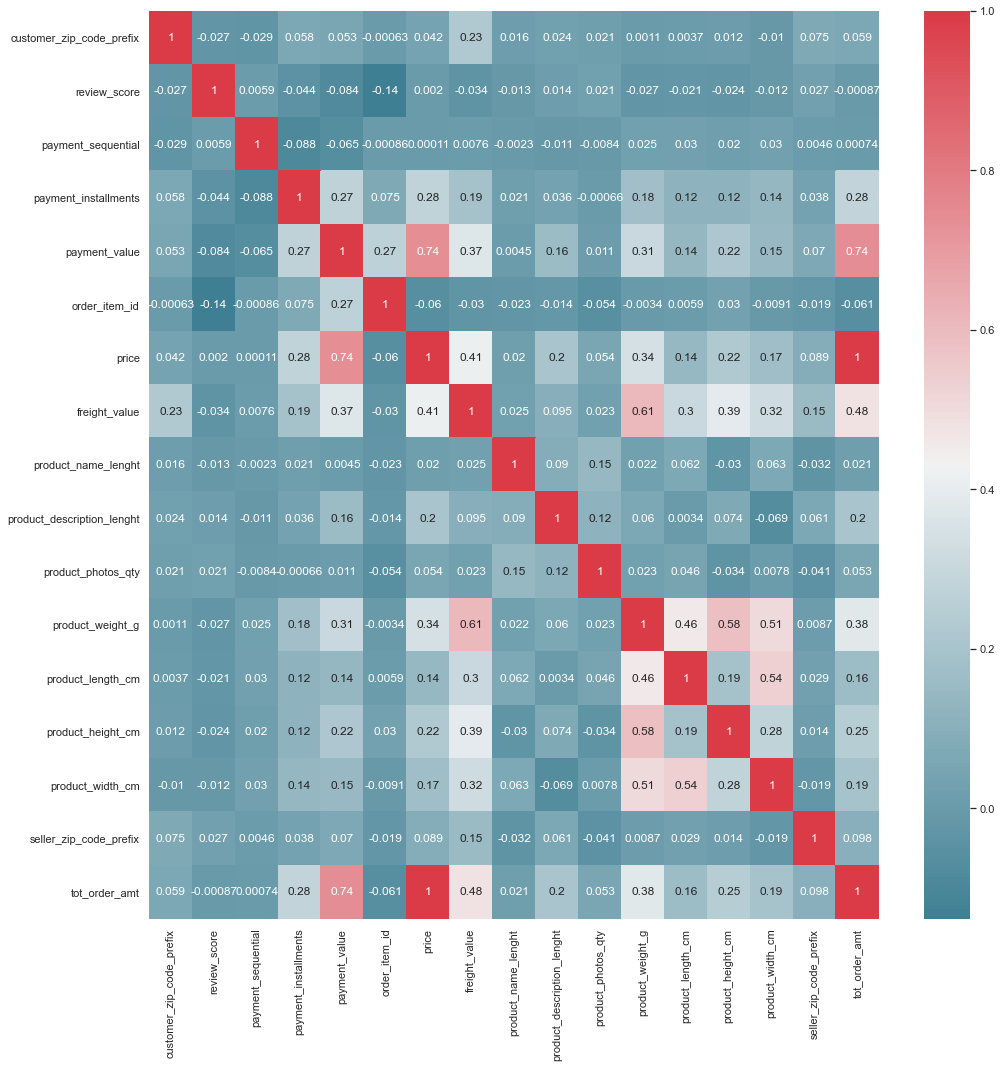


Figure 6A. Correlation heatmap for the numerical attributes in the Olist dataset.

* Review Score score does not appear to be highly correlated with any of the other variables
* Payment Value and Price are highly correlated because Payment Value = Price + Freight - Voucher
* Payment Value and Total Order Amount are also highly correlated, Payment Value is a portion of Total Order Amount or the entire amount if no vouchers are used in the payment process.
* Price and Freight Value 0.42: not too highly correlated so cost doesn’t necessarily dictate shipping cost
* Product Weight and Freight Value are highly correlated at 0.61, that makes sense because shipping costs are based on weight, package size, distance and type of shipping if that’s any option.
* Interesting that the Product Length, Width and Height are not highly correlated with Freight Value, but they are correlated with Product Weight
* Product Length and Width also appear to be correlated with each other, we may consider merging the H+W+L into one attribute called Volume to reduce the overall number of attributes
* Total order amount and product description(s): There does not appear to be a strong association between price and the product descriptions (title length, description length, quantity of photos), but we suspect that there might be a correlation between the number of items sold by sellers and the length of the product title, description and the number of photos associated with a listing
* Price and payment instalments do not appear to be highly correlated either, that could be due to most people paying for entire orders up front in one payment vs electing to make multiple payments.
* Price and Payment Sequential similarly are not highly correlated and that makes it seem like more people are paying with a single payment method.

A bar chart plot of monthly sales (Figure 6B) shows that sales volume are highest in May, August and July and lowest in September and October.

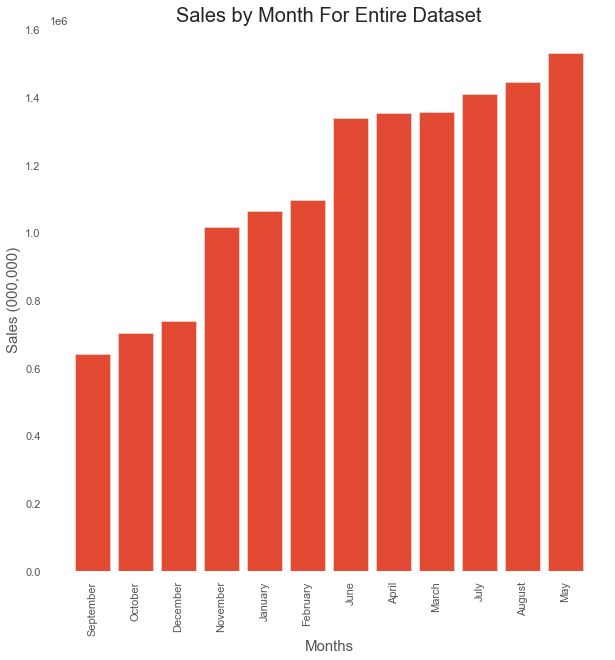
****

Figure 6B. Histogram of Sales volume versus Month.

Based on an article we found from [*Latin America Business Stories*](https://labsnews.com/en/articles/ecommerce/brazil-ecommerce-shopping-dates/) written on July 31, 2018, towards the end of our dataset, these trends can be explained as follows:Important dates: Consumers Day March 15th

* April: Free Shipping Day - last Friday in April, but in 2017 (that our dataset covers) some sellers gave customers an extra day of free shipping…this is important because according to research (not documented in the article), up to 90% of Brazilian consumers did not make an online purchase because of shipping costs
* May: Mother’s Day - the 2nd Sunday in May (also contributes to free shipping day sales) - ranked as the 2nd busiest ecommerce day in the country
* June: June 12th ‘Dia do Namorados’, Brazil’s Valentine’s Day (in 2018 the most popular categories were Beauty and Health, Fashion and Technology
* August: Father’s Day - 2nd Sunday in August, most popular categories are Fashion and Accessories (31%), Electronics (22%), Books (14%), Fragrances (9%) and the rest are Travel, Shoes, CDs, DVDs and others.
* October: Children’s Day - October 12th - a lot of companies release new items around Children’s Day, this gives both kids and sellers a preview of what the Christmas season will look like
* November: Black Friday/ Cyber Monday or Black Week events are increasingly popular in November with big discounts on things like home appliances (-15.28%), electronics (-11.26%), cosmetics (-10.38%) and fashion (-10.35%)
* November: Chinese Double Eleven - November 11th. (basically celebrates Single’s Day), its a shopping day with deep discounts
* December: Christmas - the 13th paycheck is sort of a bonus that all registered workers are entitled to

# 

Figure 6C. Plot of log transformed total order amount vs log of freight value

Figure 6A (correlation heatmap) shows that total order amount and freight values have a relatively good association (correlation coefficient = 0.48), however the shape of the scatterplot of their raw values did not show this relationship. Therefore, these attributes were log transformed (Figure 6C). The log transformed plot depicts a positive linear association from log of freight value 0 and 2. Log of freight value is relatively invariant for lower values. The plot does not show an apparent trend for reviews score in total amount and freight value.

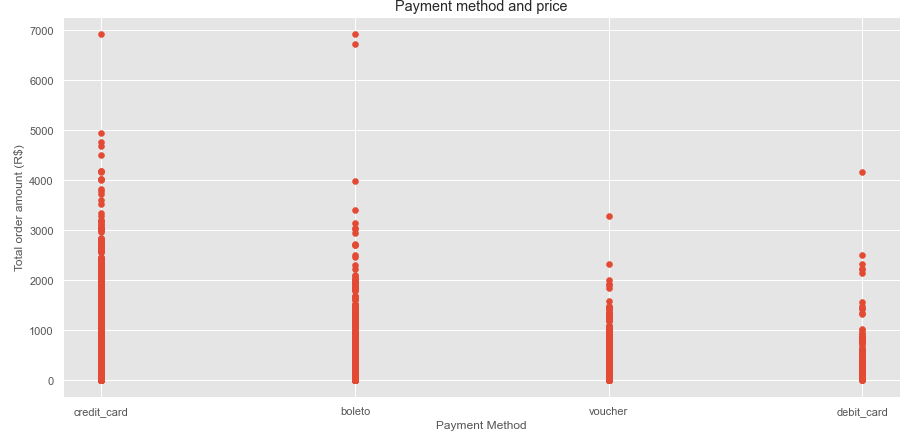


Figure 6D. Plot of Payment Method vs total order amount

A plot of payment method versus total amount paid for the order indicates that credit card payment is the most popular method for amounts higher than R$ 3000 and as noted earlier, credit cards are the most popular payment method.

# 

# Explore Attributes and Class

## **Delivery Performance (classification task)**

One of the main objectives of this project is to understand factors that are related to customer satisfaction as measured by review attributes (e.g., review\_score, review\_comments).

As shown in Figure 7A, a relatively good correlation between delivery\_estimate\_discrepancy and review\_score implies an association. The positive correlation suggests that customers that received their order earlier than the estimated delivery date tend to provide a high/good review\_score. This relationship will be further explored during the modeling phase.

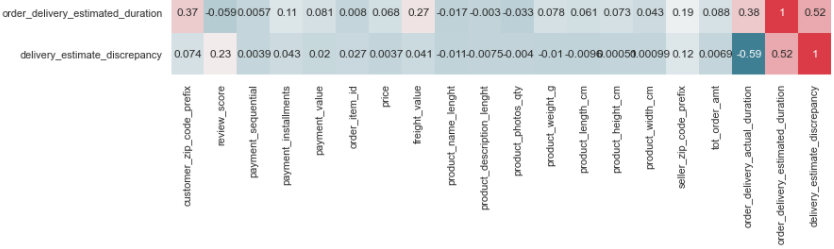


Figure 7A. Correlation heatmap that shows linear relationship between *delivery\_estimate\_discrepancy* and other numerical attributes in the Olist dataset.

The data distribution for delivery estimate discrepancy is not normally distributed as it shows a long-tailed distribution shape(Figure 5). This shape is due to the occurrence of outliers on both sides of the data distribution.

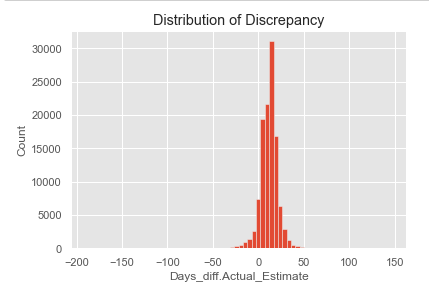


Figure 7B: Distribution of Delivery estimate discrepancy.

In this case, setting a threshold value of 95% quantile of the delivery estimate discrepancy values was able to exclude these outliers and resulted in a more representative data distribution that is less skewed (Figure 7C and 7D).

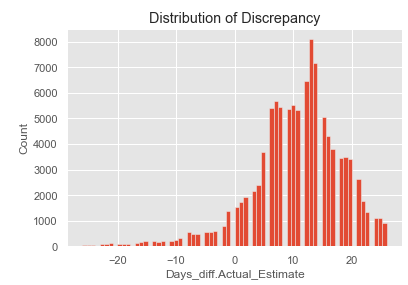


Figure 7C. Histogram of 95% percentile of delivery estimate discrepancy data.

In Figure 7D, each side of the gray line is the kernel density estimation which shows the distribution shape of the delivery estimate discrepancy. Looking at the wider sections of the violin plot, it can be said that there is a higher probability for delivery discrepancy values to fall within 5 and 25, unlike the skinnier sections which represent a lower probability.

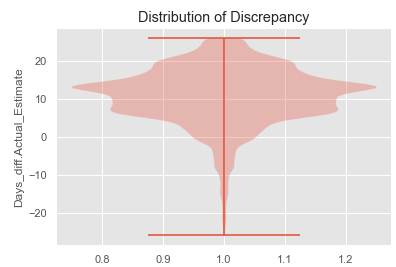


Figure 7D: Violin plot for 95% percentile of distribution of delivery estimate discrepancy.

### **Key Findings**

One notable observation from figure 6 below was that customers that received their orders earlier than the estimated dates provided good review scores and vice versa for customers that received their orders late. This association deserves exploration to understand the nature of the relationship while noting any confounding variable.

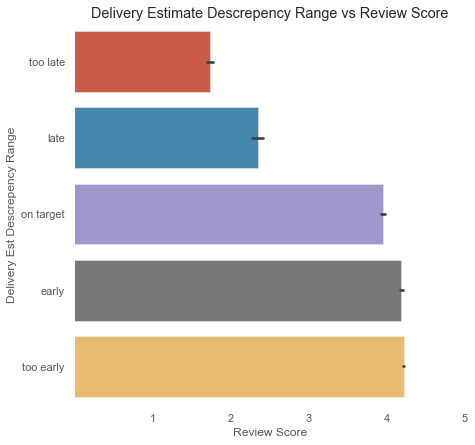


Figure 7E. Bar plot for levels of delivery estimate discrepancy range and their corresponding mean review scores to give us an idea of the relationship between the actual review score and the discrepancy between an estimated time of delivery and an actual time of delivery. More information for us to be able to help determine how to improve customer satisfaction in regards to logistic advancement.

Table 1 shows the distribution of delivery estimate discrepancy for respective months and corresponding review scores. It is normalized to show row proportions of the review scores.

From this table most of the processed orders are being delivered too early across all months of the year and even for some customers that provided low review scores. Despite the identified positive impact of early delivery, a major limitation to the practice is loss of sales from potential customers who view an estimated delivery date to be too long. These customers are not aware that, if placed, such orders might arrive 10 to 20 days earlier. For this reason, it is more informative to provide an estimated delivery date that is within 4 days of actual delivery date.

*Table 1. Generated using pandas crosstab(). Delivery estimate discrepancy for respective months and corresponding review scores.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Month** | **Review\_score** | **too late** | **Late** | **Target** | **Early** | **Too Early** |
| April | 1 | 0.15 | 0.04 | 0.09 | 0.06 | 0.65 |
|  | 2 | 0.06 | 0.03 | 0.08 | 0.13 | 0.70 |
|  | 3 | 0.03 | 0.02 | 0.10 | 0.13 | 0.72 |
|  | 4 | 0.01 | 0.00 | 0.09 | 0.13 | 0.77 |
|  | 5 | 0.00 | 0.00 | 0.05 | 0.11 | 0.83 |
| August | 1 | 0.10 | 0.07 | 0.19 | 0.12 | 0.52 |
|  | 2 | 0.07 | 0.06 | 0.16 | 0.17 | 0.54 |
|  | 3 | 0.02 | 0.03 | 0.24 | 0.15 | 0.56 |
|  | 4 | 0.00 | 0.01 | 0.23 | 0.19 | 0.57 |
|  | 5 | 0.00 | 0.00 | 0.24 | 0.17 | 0.59 |
| December | 1 | 0.23 | 0.06 | 0.10 | 0.09 | 0.54 |
|  | 2 | 0.10 | 0.03 | 0.11 | 0.13 | 0.62 |
|  | 3 | 0.03 | 0.01 | 0.12 | 0.12 | 0.72 |
|  | 4 | 0.01 | 0.00 | 0.09 | 0.09 | 0.80 |
|  | 5 | 0.01 | 0.01 | 0.05 | 0.07 | 0.86 |
| February | 1 | 0.34 | 0.10 | 0.12 | 0.09 | 0.35 |
|  | 2 | 0.14 | 0.03 | 0.17 | 0.16 | 0.50 |
|  | 3 | 0.07 | 0.02 | 0.17 | 0.13 | 0.61 |
|  | 4 | 0.02 | 0.01 | 0.12 | 0.13 | 0.72 |
|  | 5 | 0.01 | 0.01 | 0.09 | 0.12 | 0.78 |
| January | 1 | 0.19 | 0.05 | 0.09 | 0.04 | 0.63 |
|  | 2 | 0.08 | 0.03 | 0.09 | 0.10 | 0.71 |
|  | 3 | 0.02 | 0.01 | 0.14 | 0.10 | 0.73 |
|  | 4 | 0.00 | 0.00 | 0.07 | 0.10 | 0.82 |
|  | 5 | 0.00 | 0.00 | 0.05 | 0.07 | 0.88 |
| July | 1 | 0.11 | 0.03 | 0.12 | 0.11 | 0.63 |
|  | 2 | 0.04 | 0.03 | 0.08 | 0.13 | 0.72 |
|  | 3 | 0.02 | 0.01 | 0.12 | 0.16 | 0.70 |
|  | 4 | 0.01 | 0.01 | 0.10 | 0.17 | 0.72 |
|  | 5 | 0.00 | 0.00 | 0.08 | 0.16 | 0.75 |
| June | 1 | 0.07 | 0.01 | 0.05 | 0.07 | 0.80 |
|  | 2 | 0.02 | 0.01 | 0.05 | 0.09 | 0.84 |
|  | 3 | 0.01 | 0.00 | 0.05 | 0.09 | 0.83 |
|  | 4 | 0.01 | 0.00 | 0.04 | 0.09 | 0.86 |
|  | 5 | 0.00 | 0.00 | 0.03 | 0.08 | 0.89 |
| March | 1 | 0.36 | 0.10 | 0.12 | 0.08 | 0.34 |
|  | 2 | 0.22 | 0.05 | 0.22 | 0.11 | 0.40 |
|  | 3 | 0.08 | 0.03 | 0.22 | 0.15 | 0.52 |
|  | 4 | 0.02 | 0.02 | 0.17 | 0.18 | 0.61 |
|  | 5 | 0.01 | 0.01 | 0.13 | 0.19 | 0.66 |
| May | 1 | 0.14 | 0.06 | 0.09 | 0.10 | 0.61 |
|  | 2 | 0.09 | 0.03 | 0.14 | 0.16 | 0.58 |
|  | 3 | 0.04 | 0.02 | 0.19 | 0.11 | 0.64 |
|  | 4 | 0.01 | 0.01 | 0.13 | 0.14 | 0.70 |
|  | 5 | 0.01 | 0.01 | 0.10 | 0.13 | 0.76 |
| November | 1 | 0.31 | 0.08 | 0.13 | 0.10 | 0.37 |
|  | 2 | 0.16 | 0.06 | 0.15 | 0.15 | 0.48 |
|  | 3 | 0.06 | 0.02 | 0.21 | 0.16 | 0.55 |
|  | 4 | 0.02 | 0.01 | 0.17 | 0.20 | 0.61 |
|  | 5 | 0.01 | 0.01 | 0.13 | 0.18 | 0.67 |
| October | 1 | 0.11 | 0.03 | 0.09 | 0.10 | 0.67 |
|  | 2 | 0.04 | 0.02 | 0.15 | 0.11 | 0.69 |
|  | 3 | 0.02 | 0.03 | 0.15 | 0.13 | 0.67 |
|  | 4 | 0.00 | 0.01 | 0.09 | 0.15 | 0.75 |
|  | 5 | 0.00 | 0.00 | 0.08 | 0.13 | 0.78 |
| September | 1 | 0.16 | 0.03 | 0.13 | 0.14 | 0.54 |
|  | 2 | 0.11 | 0.03 | 0.10 | 0.16 | 0.59 |
|  | 3 | 0.03 | 0.01 | 0.12 | 0.15 | 0.68 |
|  | 4 | 0.01 | 0.01 | 0.11 | 0.18 | 0.69 |
|  | 5 | 0.00 | 0.00 | 0.09 | 0.15 | 0.75 |

#### **Average Delay in Delivery - State Level**

Another key finding revealed higher delivery delay (days) for the states such as AC (20), RO (18.9), AM(~ 19 days) while lowest delivery delays were measured for AL (8 days) and MA (9 days) respectively (Figure 7). Interestingly, states with highest delivery delay have the lowest population density which can negatively impact delivery network/supply chain efficiency in such states. (Figure 7G.) We should also point out that AM is the Amazonas state in Brazil, a state that is covered almost entirely by the Amazon rainforest.

Figure 7G. Bar plot showing average delay in delivering products to remote states is plotted.

### **Approach to modeling delivery estimate**

Modeling the delivery estimate will involve predicting actual delivery dates for the test set.

Then 4 days will be added to that prediction to obtain the estimated delivery date. The resulting estimated delivery dates will be a refinement to the currently reported estimated delivery dates.

## **Predict Review Score (Classification task)**

From the correlation coefficient heatmap presented in figure 6, review score is associated with item\_id. But this relationship needs to be vetted. As previously stated, review\_score is associated with delivery attributes.

### **Approach to modeling review score**

The review score attribute is grouped into 3 classes viz: Bad (1 to 2), Fair (3) and Good (4 to 5) (Figure 8). A review score class will be predicted for the test set instances and the enumerated number of ‘good’ reviews will be used as a key performance index for customer satisfaction.

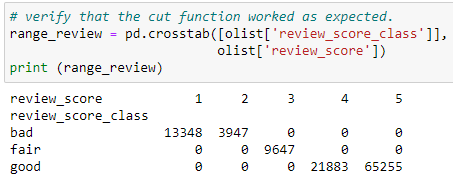


Figure 8. Output showing the counts of review score classes generated from the review score attribute.

# **New Features**

New features were created from their related counterparts. These are highlighted below.

* Product\_dimension was formed from a product of product length, product width and product height.
* Total order amount was derived from price and freight\_value attributes and it represents the uniform actual order amount because it excludes payment portions made by using vouchers and boleto payment methods.
* .purchase\_month and purchase\_wk\_day were derived by extracting the month and day (respectively) portion of the order\_purchase\_timestamp converted datetime datatype.
* Actual duration of order delivery was derived from time duration between order approved date and actual date the customer received the order.
* Estimated duration of order delivery was obtained by calculating the time duration between order approved date and estimated date projected for delivery.
* Delivery estimate discrepancy is the difference between estimated duration of order delivery and actual duration of order delivery.**This feature is key to optimizing delivery performance**
* Delivery estimate range is binned levels of delivery estimate discrepancy set as too late, late, on-target, early, too early.
  + On-target delivery estimates are those delivered within +- 4 days (arbitrarily set to account for non-working days e.g. weekends) of the actual delivery date.
* Review\_score\_class is the grouped levels of review score attribute.

The next phase will add a new column for the total length of time it took between each stage of delivery could prove helpful in identifying steps in the logistics process where improvements could be made. Specifically, a factor called Processing, which is the time between the order\_approved\_at time and the order\_delivered\_carrier\_date. This would show any slowdowns between the customer’s payment being approved and it getting to the shipping/logistics partner. Another factor called Delivery could be created between order\_delivered\_carrier\_date to order\_delivered\_customer\_date, to show any delays with shipping times.

# 

# 

# References

1. Olist dataset in Kaggle : <https://www.kaggle.com/olistbr/brazilian-ecommerce?select=olist_order_items_dataset.csv>
2. <https://olist.com/>
3. <https://pt.wikipedia.org/wiki/Olist>
4. Violin Plots 101: Visualizing Distribution and Probability Density: <https://mode.com/blog/violin-plot-examples/>. Accessed on 05/15/2021.
5. [*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.), March 20, 2021
6. [*Latin America Business Stories*](https://labsnews.com/en/articles/ecommerce/brazil-ecommerce-shopping-dates/), July 31, 2018
7. <https://en.wikipedia.org/wiki/Amazonas_(Brazilian_state)>

# 

# Appendix

Olist, a Brazilian E-commerce site provided a robust dataset of orders made at the Olist Store. The data set consists of roughly 100,000 orders from 2016 to 2018 and is multidimensional covering order information, consumer information, seller information, geolocation information, product attributes and customer reviews. The dataset will allow us to meet our stated business objectives. We will process the data using a combination of Python for data cleaning, mining, wrangling, exploration, feature selection and data modeling. and will possibly employ cloud services for tasks such as running sentiment analysis.

## Data Meaning Type

There are 2,515 orders missing a customer delivery date. Of those missing delivery dates it appears there were 564 cancelled orders, 7 orders that were unavailable, 8 orders that show as delivered but are missing delivery dates and 364 orders that were invoiced but there is no record of the orders being delivered. Other reasons we don't have delivery data is because at the time the data was pulled, an order could've been shipped, processing or approved or otherwise still in the process of making its way to the customer. These are just good things to note, we may consider just removing all 2,515 lines since we have so much data already. There were 7 orders that were cancelled, but still delivered and more than likely that's because the order was cancelled after it shipped from the seller.

## Modeling

Main objective is to determine the optimal delivery duration with the aim that this prediction will be better than the current estimated delivery. We will then recommend our best model to management to use for estimating delivery date for future orders.

Approach:

* Regression problem (can be changed to classification problem by transforming target to *‘Late delivery’, ‘Early Delivery’* or *‘Precise delivery’*)
* Duration\_Actual =
  + Actual delivery date (order\_delivered\_customer\_date) - order purchase date (order\_purchase\_timestamp)
* Duration\_Estimated =
  + Estimated delivery date (order\_estimated\_delivery\_date) - order purchase date(order\_purchase\_timestamp)
* Split to train-test
* Predict ‘actual duration’ for the test.
  + Set performance threshold (e.g., accuracy of 85%)
* Further compare the predicted result to ‘Estimated delivery’
* Modeling techniques:
  + Start with explainable models so we can determine the important features
    - E.g. linear or logistic regression, single trees model
  + For prediction, apply complex models. E.g. xgboost(), SVM(), ensemble learning techniques.
* Assess model:
  + Keep improving the model till the threshold is surpassed

## Evaluation

* Evaluate results
  + Compare with the original project requirement.
* Review the process:
  + Review the steps for any mistake or misstep.
* Determine next steps: Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

## Success Criteria

**Outcome Measurement:**

Measured by improvements to accuracy of estimated delivery times and less bad reviews that mention shipping.

Estimated delivery time is within 1 day of the customer receiving their ordered package.

**Data Importance:**

Strengthen their competitiveness in the market by

**Prediction Effectiveness:**

A major success factor will be to obtain accuracy of at least 85%, precision of at least 80%, sensitivity of at least 85%. These values are subject to review, contingent upon exploratory data analyses.

Python Code: added via pdf merge