Table of Contents

[Data Analysis – Olist 1](#_Toc102468283)

[Data source and Dataset Description 1](#_Toc102468284)

[Research Questions 3](#_Toc102468285)

[Descriptive Statistics 3](#_Toc102468286)

[Correlation Analysis 7](#_Toc102468287)

[Data Visualisation – Histogram 9](#_Toc102468288)

[Multiple Linear Regression Analysis and Its Interpretation 11](#_Toc102468289)

[Recommendations for Improvement 13](#_Toc102468290)

[References 13](#_Toc102468291)

[Appendix 14](#_Toc102468292)

# Data Analysis – Olist

The dataset of Olist was analysed in this section in line with the research questions outline for the report, followed by the recommendations made on how Olist can improve on its financial performance.

## Data source and Dataset Description

Olist is the largest department store operating in Brazil. It connects small businesses that sell their products through the Olist store (Olist, 2021). The data was made available in separate files, each containing key information about the company’s activities. 9 csv files precisely were present in the data. The data covered the periods of 2016 to 2018. Five datasets from the Olist data namely olist\_orders\_dataset; olist\_orders\_reviews\_dataset, olist\_customers\_dataset, olist\_order\_items\_dataset, and olist\_order\_payments\_dataset were merged in the RStudio environment, using the left\_join function in the **dplyr** library to enable us achieve our aims in this report. A new column named purchase\_title\_year was added to the data. This is to enable us work with years in the analysis. The data was renamed **Olist\_data\_combined**, and was used for analysis in this section. **Olist\_data\_combined** dataset has 119,143 rows and 29 columns. Below is the description of the dataset, and it includes the column names, the description of the columns and the data type of each column.

|  |  |  |
| --- | --- | --- |
| **Column** | **Column Description** | **Data type** |
| Order\_id |  | character |
| Customer\_id |  | character |
| Order\_status |  | character |
| Order\_purchase\_timestamp |  | datetime |
| Order\_approved\_at |  | datetime |
| Order\_delivered\_carrier\_date |  | datetime |
| Order\_delivered\_customer\_date |  | datetime |
| Order\_estimated\_delivery\_date |  | datetime |
| Review\_id |  | character |
| Review\_score |  | integer |
| Review\_comment\_title |  | character |
| Review\_comment\_message |  | character |
| review\_creation\_date |  | datetime |
| Review\_answer\_timestamp |  | datetime |
| payment\_sequential |  | integer |
| payment\_type |  | character |
| payment\_installments |  | integer |
| payment\_value |  | integer |
| Order\_item\_id |  | integer |
| product\_id |  | character |
| seller\_id |  | character |
| shipping\_limit\_date |  | datetime |
| price |  | integer |
| freight\_value |  | integer |
| customer\_unique\_id |  | character |
| customer\_zip\_code\_prefix |  | character |
| customer\_city |  | character |
| customer\_state |  | character |
| Order\_purchase\_year |  | integer |

## Research Questions

The research questions highlighted what we aimed to focus on in this analysis. The following research questions were outlined for the report:

2. What is the average price of an item on Olist?

1. What percentage of the total customer reviews is 5-star?

3. Which payment type generates the most payment value on Olist?

4. In which year was most orders made?

5. Is there a growth in the purchase on Olist over the years?

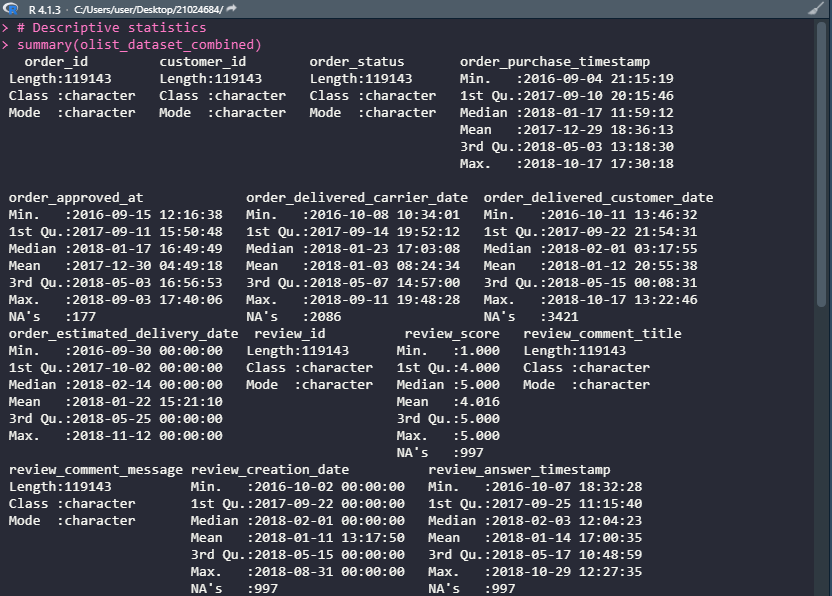
6. What impact do price, freight value, and payment installments have on payment value?

7. How can the level of customer satisfaction be increased?

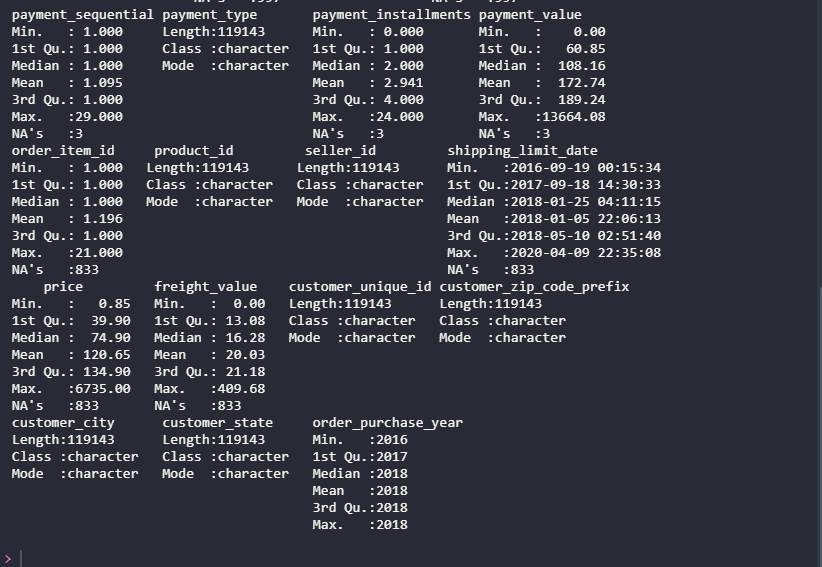
8. How can Olist boost its annual revenue?

## Descriptive Statistics

The descriptive statistics of the variables in the Olist\_dataset\_combined DataFrame was conducted, with the result presented in this section. It revealed important summary about the dataset. It is only the descriptive statistics of the numeric variables that are important to the analysis. For example, the mean of freight\_value was 20.03, and it has median, minimum, maximum values of 16.28, 0.00 and 409.68 respectively. It was revealed from the descriptive statistics also that the average price of an item on Olist was 120.65. This means over the periods of 3 years (2016 to 2018), an item’s price averaged 120.65.

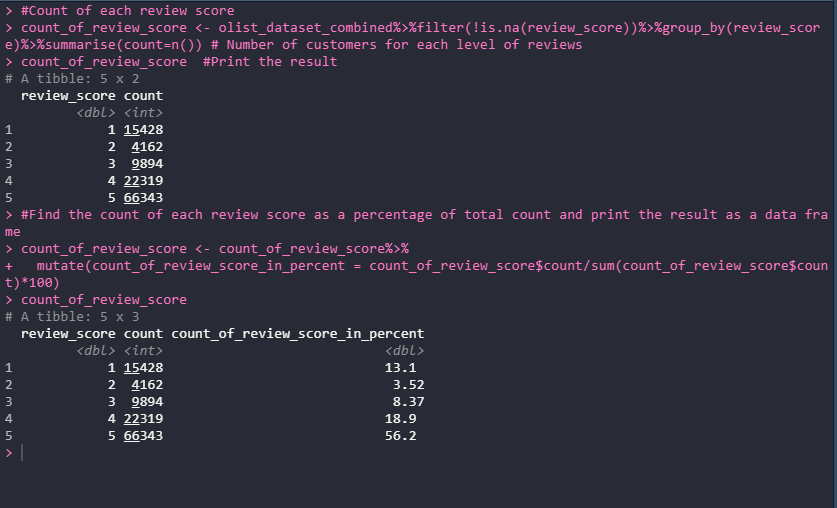


**Figure 1a: Descriptive Statistics of the Olist Dataset Variables**



**Figure 1b: Descriptive Statistics of the Olist Dataset Variables continued**

The percentage of the total customer reviews that is 5-star is 56.2%, which was more than half of the total of all the possible review scores, as revealed in Figure 2.



**Figure 2: Percentage of the Total Customer Reviews that is 5-star**

The total payment value by total payment was presented in Figure 3. It was revealed that payment type which generates the most payment value on Olist was credit card. A total of 15,775,451 was paid via credit card.

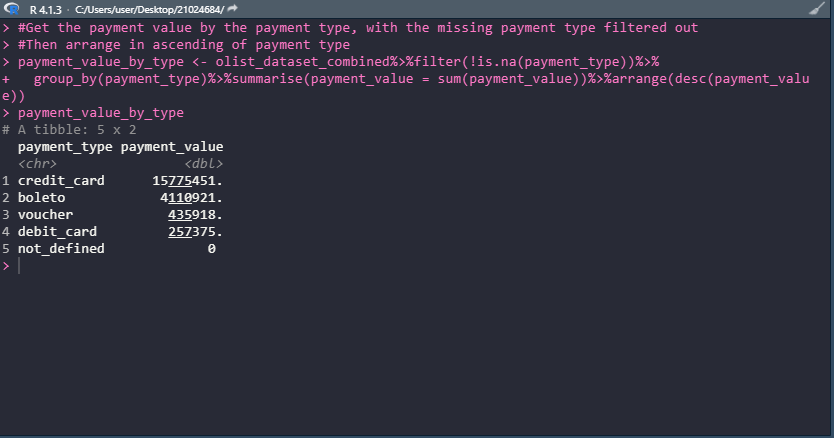


Figure 3: Payment Value by Payment Type

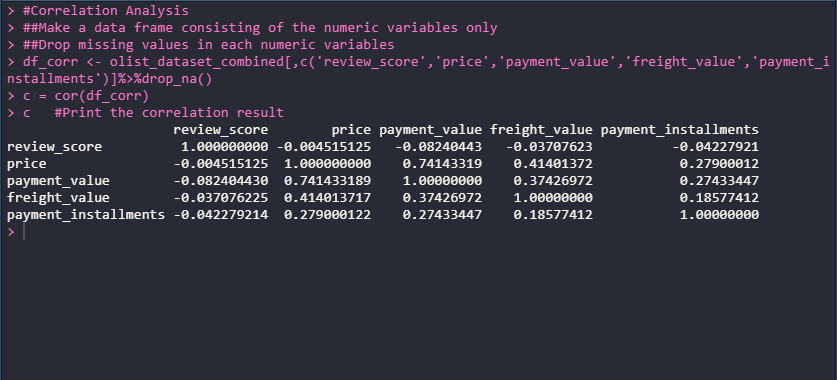
As regards the year in which most orders were made, it was revealed in the analysis that most order purchase was made in the year 2018. The total order purchase this year was 64,185. Also, it is clear that there is a growth in the purchase on Olist over the years, as the purchase has continues to rise over the years. The total purchase on Olist in 2016 was 409, it was 54,549 in 2017 and it rose to 64,185 by the year 2018. The result of the analysis was presented in Figure 4.



**Figure 4: Order purchase by Purchase Year**

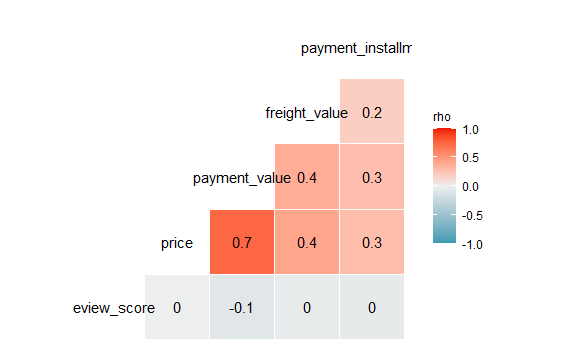
## Correlation Analysis

With correlation, it is easy to find the degree of association among numerical variables in a dataset. The values of a correlation can range between 1 and -1. The closer the value is to 1, the higher the positive correlation, and the closer the value is to -1, the higher the negative correlation between any two variables. The result of the correlation analysis was presented in Figure 5. The correlation between some of the variables is positive, while the correlation between some other variables is negative. For example, the correlation between review\_score and payment\_value was -0.08, signifying a low negative association between the two variables. Also, the correlation between price and payment\_installments was 0.28, which means there is a positive bit low association between the two variables. Correlation, however, does not imply causation; it only shows how two variables are associated with each other.



**Figure 5a: Correlation Analysis Result**

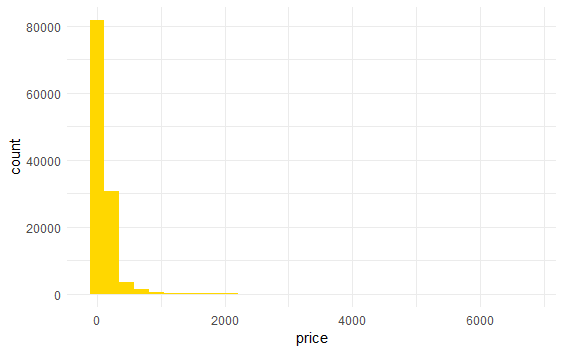
The correlation result was presented visually also in Figure 5b. The very bright red color indicates that there is a high positive correlation among the variables; the very bright blue color indicates that there is a negative correlation among the variables. So, how low the correlation among the variables is, depends on how faint the red and blue colors of the boxes are in the correlation matrix.



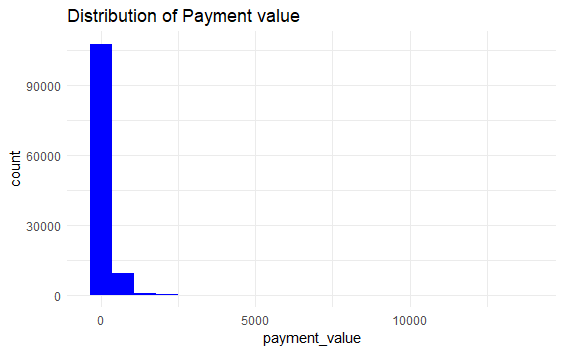
**Figure 5b: Correlation Result Presented Visually**

## Data Visualisation – Histogram

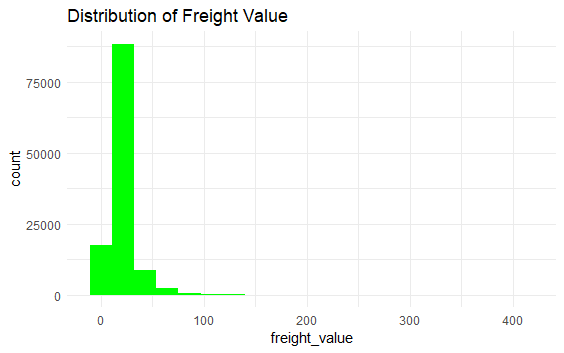
The distribution of key numeric variables was examined in this section, using histogram. A histogram is a plot that is used to explore the distribution of a continuous variable. It gives an insight into the underlying distribution, e.g. normal distribution (Roshan, 2020). The data visualization library namely ggplot2, present in the Tidyverse R package was employed to construct histograms for each key numeric variable. According to the figures presented in Figures 6, price is highly skewed to the left as most of its values are concentrated on the left side of the distribution. This is an indication that most items listed on the Olist website are items that are fairly low in price. Also, the other variables which include payment values, freight amount and installments were skewed to the left. This indicated that most values fell on the left side of the distribution in the case of the variables also.

****

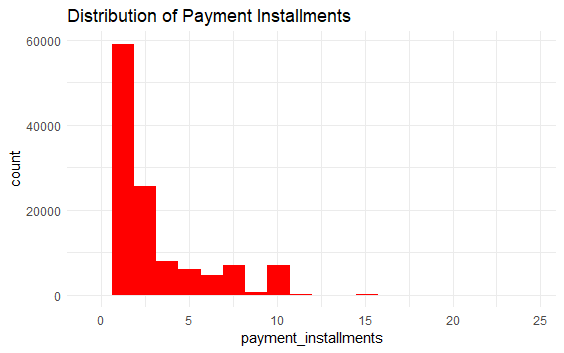
**Figure 6a: Distribution of Price**

****

**Figure 6b: Distribution of Payment Value**

****

**Figure 6c: Distribution of Freight value**

****

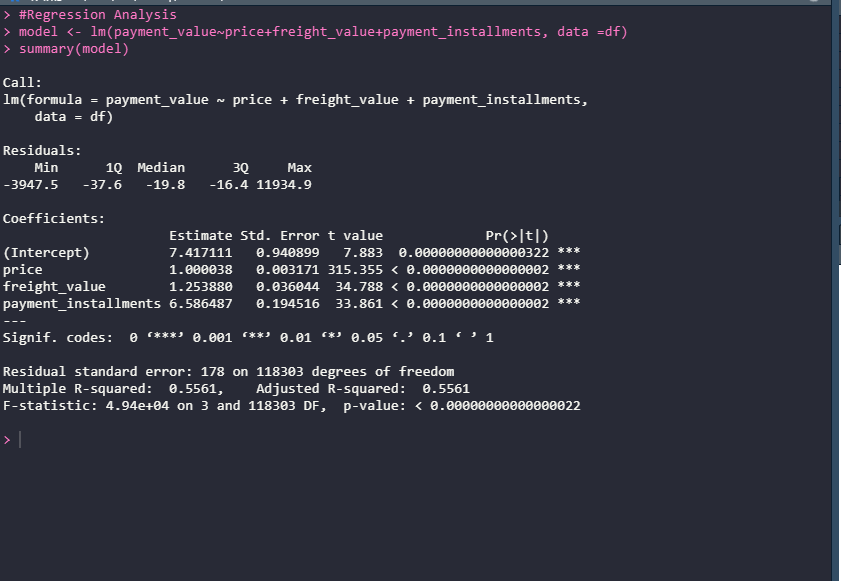
**Figure 6d: Distribution of Payment Installments**

## Multiple Linear Regression Analysis and Its Interpretation

In this section, a multiple linear regression was employed to investigate the relationship between payment value of an item and the independent variables which includes price, freight value and payment installments. The **lm()** function allowed to estimate parameters for the independent variables. According to the result which was presented in Figure 7, a unit increase in price will increase the payment values by 1.00, while holding the rest of the independent variables constant. A unit increase in freight value will increase payment value by 1.2539, while holding the rest of the independent variables constant. A unit increase in payment installments will lead to a 6.5865 units increase in payment values, while holding the rest of the independent variables constant. The constant term in the linear regression model is 7.4171. The p-values generated with the regression result enable us to evaluate the significance or otherwise of the independent variables. The p-value of price was 0.00. It is less than 0.05, which implies that price is statistically significant to the model. The p-value of freight value was 0.00, a value less than 005. This implies that freight value is statistically significant to the model. The p-value of payment installments was 0.00, which is also less than 0.05. This implied that payment installments were statistically significant to the model. Analysis of variance (ANOVA), otherwise known as F-test is used to test the overall significance of a model. The ANOVA result was also generated with the regression output. According to the result, the F-statistic was 494000, with a corresponding of 0.00. The ANOVA p-value is less than 0.05, which implied that the model is statistically significant. Also, the adjusted R-squared was 0.55, implying that the proportion in payment value that is explained jointly by price, freight\_value and payment\_installments was 55%.

The estimated regression model is given as:

7.4171 + 1.00price + 1.2539freight\_value + 6.5865payment\_installments



**Figure 7: Multiple Linear Regression result**

## Recommendations for Improvement

Olist can improve on the level of customer satisfaction by ensuring only quality and genuine products are listed on its ecommerce website. With this, the company will earn the trust of customers who shop on the platform. Also, customers will leave a good review when they get items that provide value for their money.

To increase its revenue, and consequently, proft, Olist should ensure they minimize the rate at which orders are cancelled from the part of the sellers on its platform. The company should also ensure the prices of the items sold on the platforms are affordable to buyers. With this, more customers will be encouraged to buy items via Olist.

# References

Lamorte, W., 2016. *Basic statistical analysis using the R statistical package: Saving an R dataframe as as a .csv file.* [Online]   
Available at: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R-Manual/R-Manual5.html  
[Accessed 1 May 2022].

Olist, 2021. *Brazilian E-commerce Public Dataset by Olist.* [Online]   
Available at: kaggle.com/datasets/olistbr/brazilian-ecommerce  
[Accessed 1 May 2022].

Roshan, B., 2020. *Data Visualisation with R.* [Online]   
Available at: https://www.kaggle.com/code/benroshan/part-4-data-visualization-with-r/notebook  
[Accessed 1 May 2022].

Stack Overflow, 2015. *how to perform multiple left joins using dplyr in R [duplicate].* [Online]   
Available at: https://stackoverflow.com/questions/32066402/how-to-perform-multiple-left-joins-using-dplyr-in-r  
[Accessed 1 May 2022].

## Appendix

