Final Project Team 29

- Ximena Rios Cotazo
- Ramon Manuel Sandoval
- Luisa Maria Carabali
- Alejandro Camargo Garcia
- Cristian Sarmiento
- Hector Melo

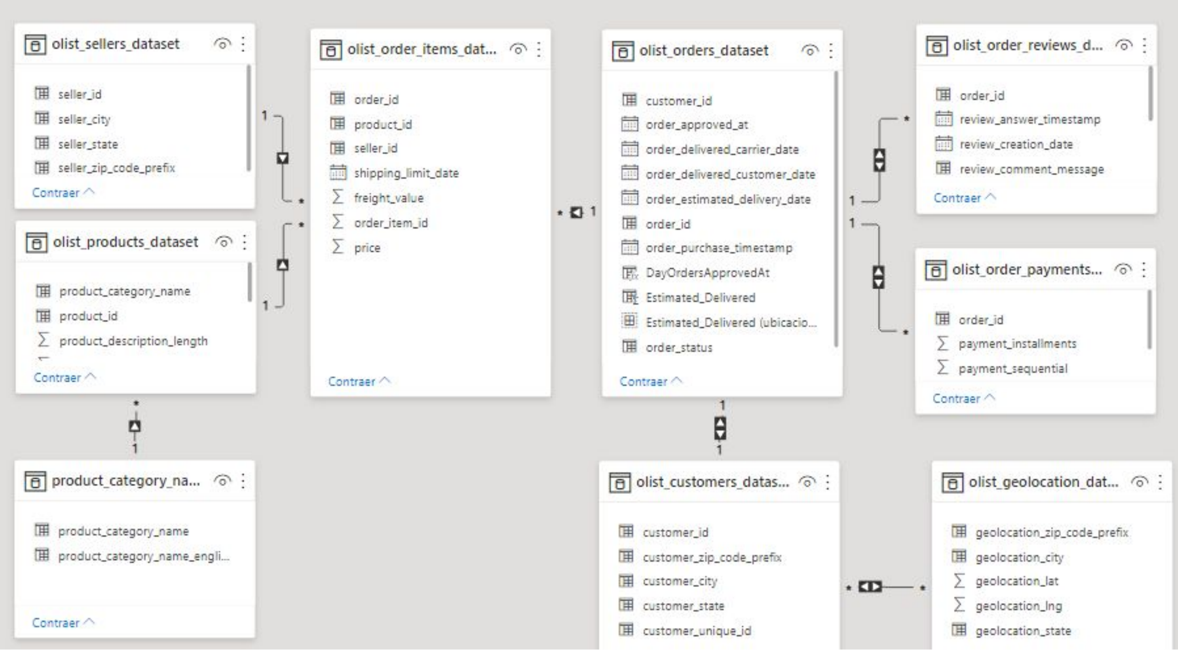
E-COMMERCE

How the consumer behavior could be described for the Olist Platform?

This is a Brazilian ecommerce public dataset of orders made at Olist Store. The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers.

We also released a geolocation dataset that relates Brazilian zip codes to lat/log coordinates. This is real commercial data, it has been anonymized, and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses.

Below is a diagram to understand the structure of the information provided.



2.2 **Analytical Context.**

This dataset was generously provided by Olist, the largest department store in Brazilian marketplaces. Olist connects small businesses from all over Brazil to channels without hassle and with a single contract. Those merchants are able to sell their products through the Olist Store and ship them directly to the customers using Olist logistics partners. See more on their website: www.olist.com

After a customer purchases the product from Olist Store a seller gets notified to fulfill that order. Once the customer receives the product, or the estimated delivery date is due, the customer gets a satisfaction survey by email where he can give a note for the purchase experience and write down some comments. Taken from: Kaggle

In this project we will proceed as follows:

1) EDA using the seven datasets provided by Correlation one.

2) Built a few a predictive model using linear regression,

3) Discussed the challenges of feature engineering in order to implement regression in real-world contexts.

4) Looked at neural networks as an alternative to improve the results of the linear regression model.

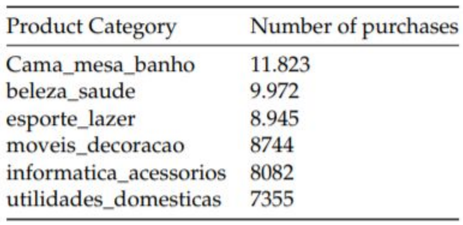
5) Make recommendations to Olist according to the insights.

Business Problem. What aspects of online consumers behavior are useful to have a better understanding of e-commerce and to predict consumer demand?

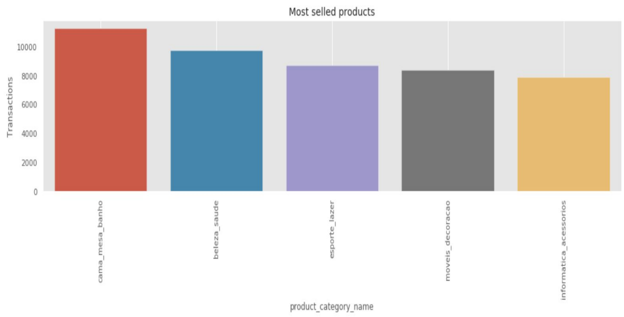
EXPLORATORY DATA ANALYSIS.

From the EDA we found out the states that make the most purchases are Sao Paulo and Rio de Janeiro, the two main states in the country.

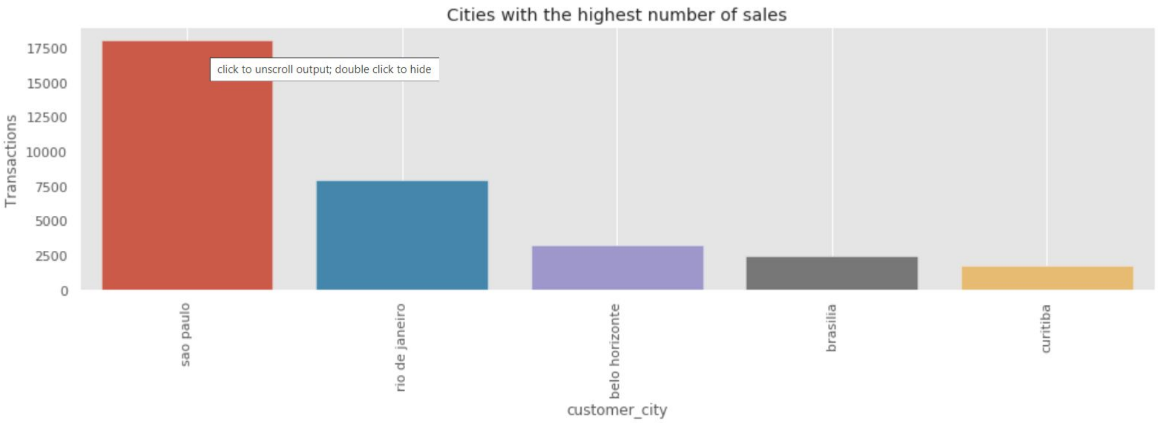
The data set has 74 categories of products and the top 5 most sell categories are:



Product Category	Number of purchases
Cama_mesa_banho	11.823
beleza_saude	9.972
esporte_lazer	8.945
moveis_decoracao	8744
informatica_acessorios	8082
utilidades_domesticas	7355

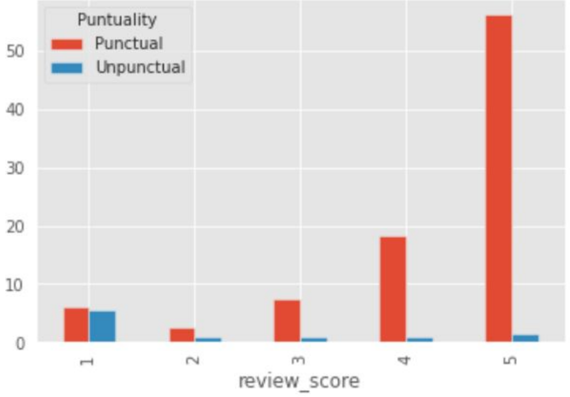


There are 4119 cities which Olist has delivered orders, the top five Cities with the highest number of sales are Sao Paulo, Rio de Janeiro, Bello horizonte, Brasilia and Curitiba.

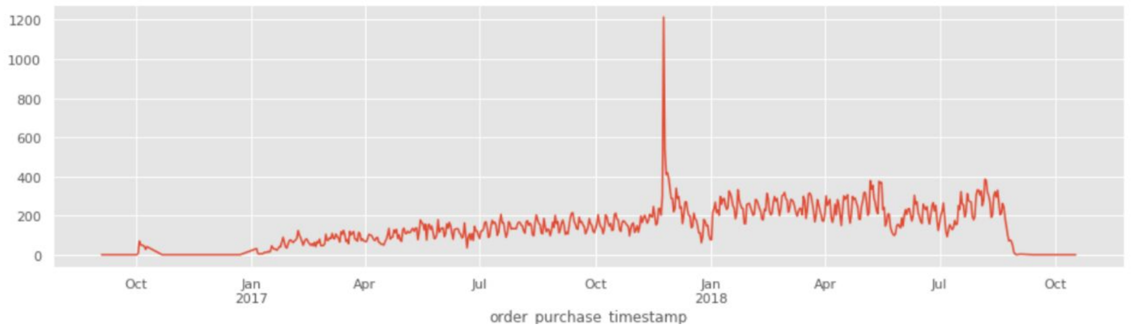


On the other hand the highest incomes for the platform come from the category health/ beauty.

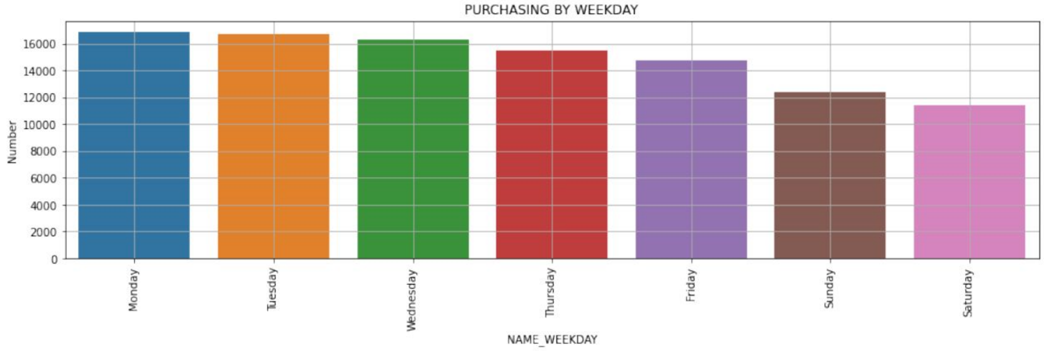
We found some differences between delivered time and delivered estimated time vs reviews, we realize that there is an inverse correlation between the score of the review and the punctuality of the delivery. This relationship is inverse. When the company is punctual in his delivery, then the customer tends to make a good review with a good score and vice versa.



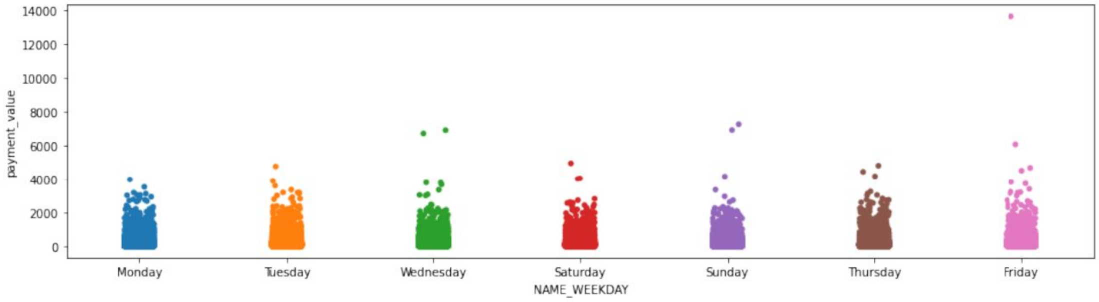
Analyzing time stamps for each order we see the number of orders has been increasing over the 2017, till around November-December 2017 where we can see a decrease, and around January 2018 continue fluctuating but increasing till October when the data end.



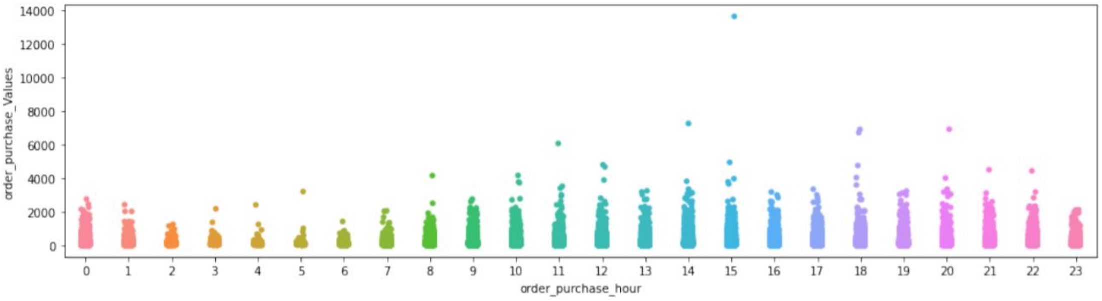
When we evaluated the number of purchases per week we can see how, during the week, the largest number of purchases is found on Mondays followed closely with the other days of the week and lastly we find the weekends, with Saturday being the day with the least amount of purchases.



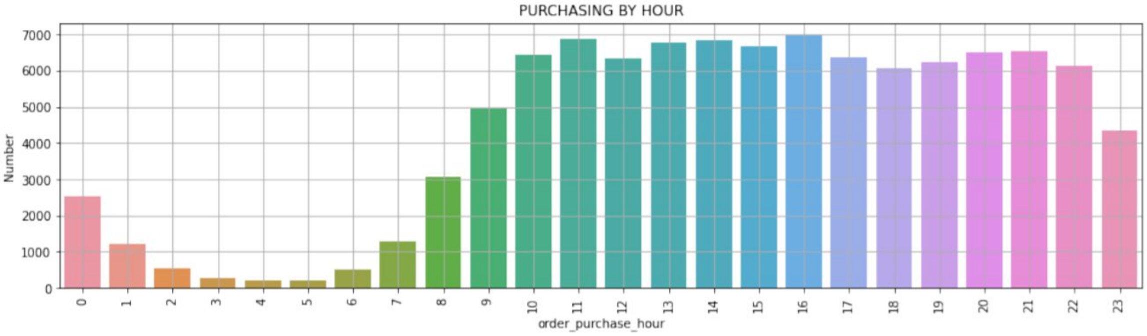
As shown in the graph, the sales values are more or less grouped below 3000, only a few values exceeded 3000 and are located below 6000, another four sales above 8000 and a single one for more of 13000 reais.



Here in the Stripplox we can see a better look of how the purchasings are accumulated over the hours with more of the values around 3000 and less over the 3000 and 9000 and just one value below the 14000.

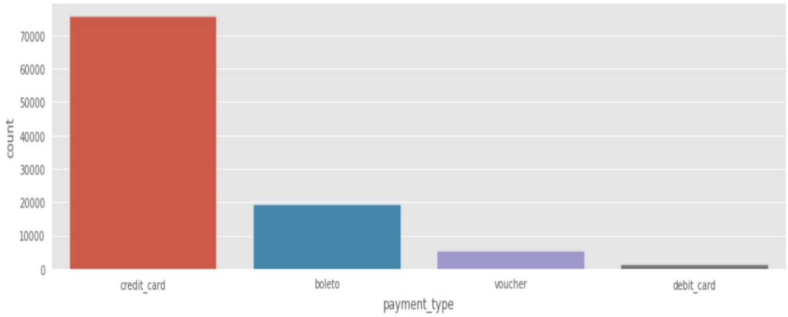


The purchasing by hour is low from 12:00AM, and starts decreasing till 5:00 AM where it is the lowest purchasing hour and from there it starts to increase till 12:00PM.



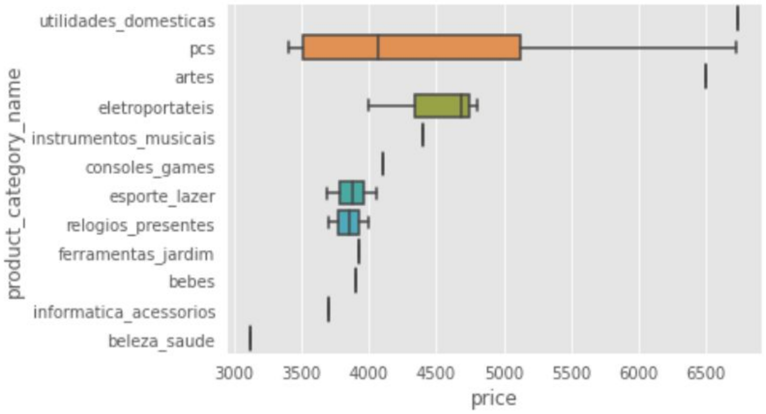
Payment Method Behavior

In Monthly bases Credit card payment has shown changes in September, payments with boleto show some peak in May, July and November, debit-card is more used in Jun to August. Also, Thursday and Saturday shows peaks of buyers using credit card, boletos and vouchers are dispersed along week.



Looking at the top 50 of purchases, they have more than 4 purchases in the period, with a couple of exceptions that they make more

We find that the most expensive products are related to the categories of Household utilities, Computers, Arts and electrical appliances. Within these, the highest price range refers to computers.



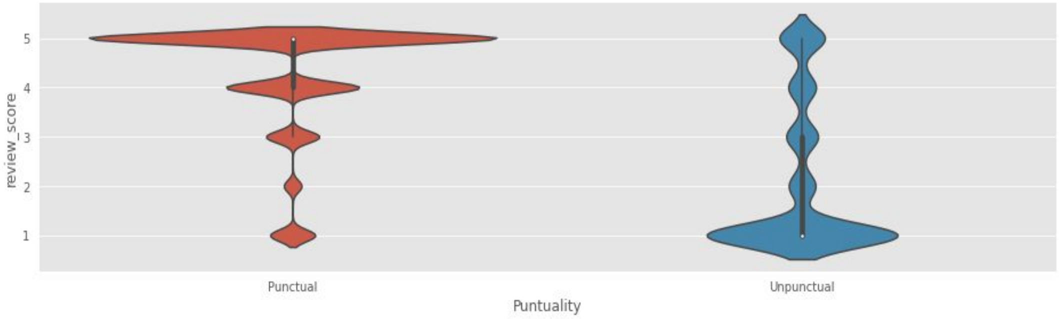
LINEAR REGRESSIONS

We used different methodologies to predict customer behavior. One of these was **Linear Regression Models**. We tried to explain the relationship between several variables in the dataset. We focused on total sales, the average price and the reviews score, and category product name, as relevant variables for demand.

In the **Exploratory Data Analysis (EDA)**, we did several simple model regressions. In these we highlight the following: Price against time, reviews against late deliveries, total sales against time, average prices against time, among others. The results of the built models are presented below.

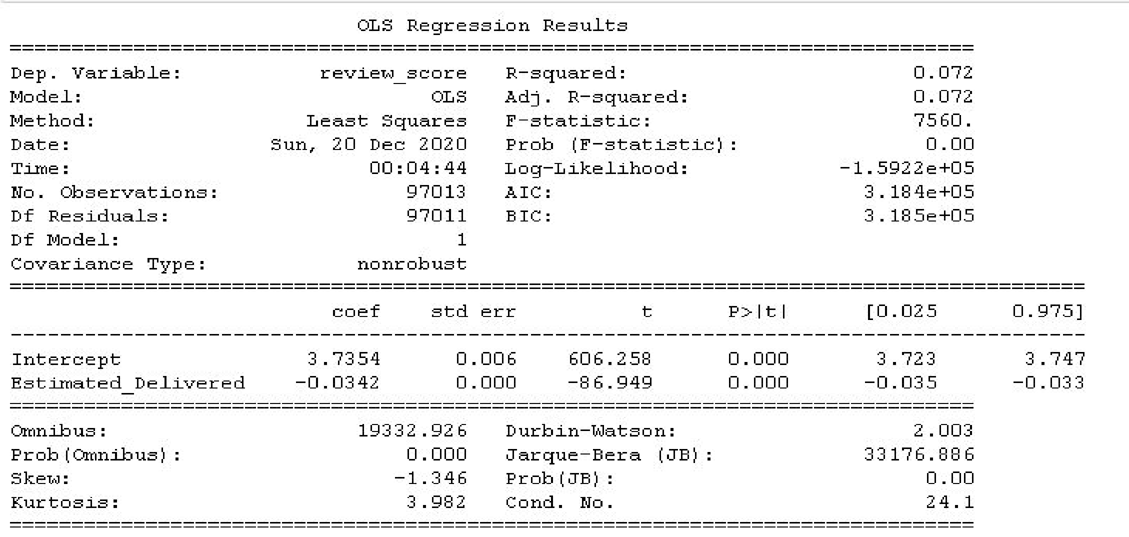
- **Reviews vs Late Deliveries**

The EDA showed that there is an inverse relationship between Review Scores, and Late Deliveries. Common sense tells us that the longer it takes me to deliver the order, the lower the score will be. We used the review score as the dependent variable, and late delivery as the independent variable. To calculate the late delivery, we used the estimated time of delivery that Olist gives their buyers, and make a difference with the real time delivery when customers receive his orders.



The previous graph shows the inverse relationship between the two variables. The longer it takes to deliver the order, the lower the rating of the review. Seeing this relationship, we do a simple model regression to try to measure how the late deliveries impact the reviews score.

The variable is indeed significant for the regression. However, we have a rather low R-squared. This may be because although late deliveries are an important factor, they are not the only one. This is why we may have model specification errors.

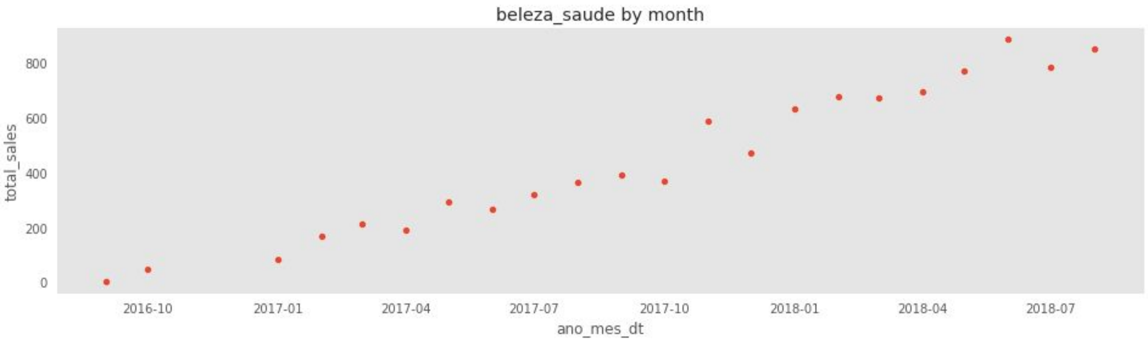


```
                            OLS Regression Results
==============================================================================
Dep. Variable:           review_score   R-squared:                       0.072
Model:                            OLS   Adj. R-squared:                  0.072
Method:                 Least Squares   F-statistic:                     7560.
Date:                Sun, 20 Dec 2020   Prob (F-statistic):               0.00
Time:                        00:04:44   Log-Likelihood:            -1.5922e+05
No. Observations:               97013   AIC:                         3.184e+05
Df Residuals:                   97011   BIC:                         3.185e+05
Df Model:                           1
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept               3.7354      0.006    606.258      0.000       3.723       3.747
Estimated_Delivered    -0.0342      0.000    -86.949      0.000      -0.035      -0.033
==============================================================================
Omnibus:                    19332.926   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            33176.886
Skew:                          -1.346   Prob(JB):                         0.00
Kurtosis:                       3.982   Cond. No.                         24.1
==============================================================================
```

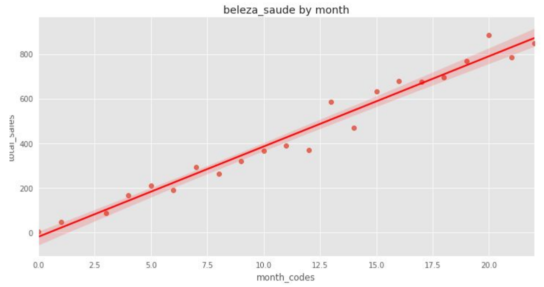
From the figure above, we can conclude that for a day's delay of the product estimate, we get a 0.03 variation in the rating that the customer may give us. As we had proposed, the model gives us an inverse relationship. However, it is still not a powerful tool to be able to perform analysis on consumers

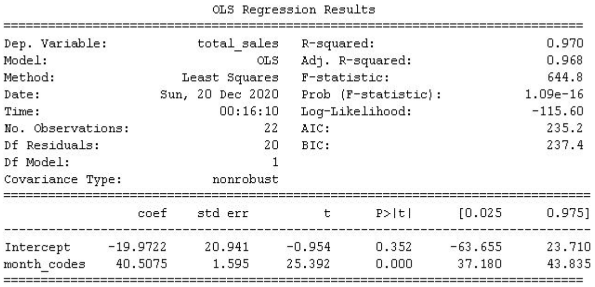
- **Sales Count by Month for Category products**

In this section, we will explore a little bit the best product on sales in the month. We are going to review its behavior over time and regress against the month codes.



Our case of study is going to be “beleza_saude”, which corresponds to the category of health and beauty products. Here we observe that the more time passes, the better the sales of the product are, to have a directly proportional relationship. Seeing it this way, we proceed to perform a linear regression to try to explain the consumption behavior of products in this category over time.



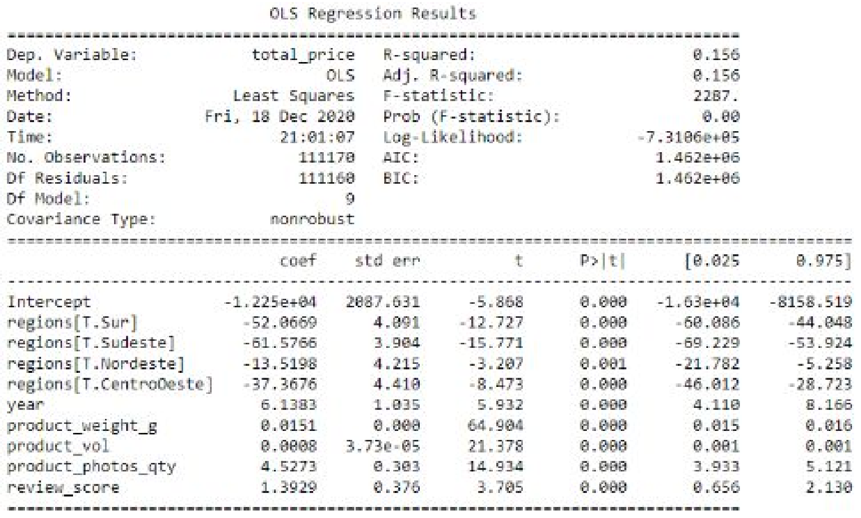


```
                            OLS Regression Results
==============================================================================
Dep. Variable:            total_sales   R-squared:                       0.970
Model:                            OLS   Adj. R-squared:                  0.968
Method:                 Least Squares   F-statistic:                     644.8
Date:                Sun, 20 Dec 2020   Prob (F-statistic):           1.09e-16
Time:                        00:16:10   Log-Likelihood:                -115.60
No. Observations:                  22   AIC:                             235.2
Df Residuals:                      20   BIC:                             237.4
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept    -19.9722     20.941     -0.954      0.352     -63.655      23.710
month_codes   40.5075      1.595     25.392      0.000      37.180      43.835
==============================================================================
```

What we get is an almost perfect relationship, with an R-square of 98% explanation of the model. With a significance of the variable demonstrated statistically. However, although the model is very good, data are left out and specification is needed to describe sales behavior on the product category. For the analysis it is not significant to know that only over time our sales will increase, we must found other significant factors.

- **What does affect the total price?**

Associated with the information from our previous simple model we went on to add more information. A multiple regression was performed to try to explain the behavior of average prices. This depends on the region, some physical characteristics of the product and the review score.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            total_price   R-squared:                       0.156
Model:                            OLS   Adj. R-squared:                  0.156
Method:                 Least Squares   F-statistic:                     2287.
Date:                Fri, 18 Dec 2020   Prob (F-statistic):               0.00
Time:                        21:01:07   Log-Likelihood:            -7.3106e+05
No. Observations:              111170   AIC:                         1.462e+06
Df Residuals:                  111160   BIC:                         1.462e+06
Df Model:                           9
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
Intercept             -1.225e+04   2087.631     -5.868      0.000   -1.63e+04   -8158.519
regions[T.Sur]          -52.0669      4.091    -12.727      0.000     -60.086     -44.048
regions[T.Sudeste]      -61.5766      3.904    -15.771      0.000     -69.229     -53.924
regions[T.Nordeste]     -13.5198      4.215     -3.207      0.001     -21.782      -5.258
regions[T.CentroOeste]  -37.3676      4.410     -8.473      0.000     -46.012     -28.723
year                      6.1383      1.035      5.932      0.000       4.110       8.166
product_weight_g          0.0151      0.000     64.904      0.000       0.015       0.016
product_vol               0.0008   3.73e-05     21.378      0.000       0.001       0.001
product_photos_qty        4.5273      0.303     14.934      0.000       3.933       5.121
review_score              1.3929      0.376      3.705      0.000       0.656       2.130
==============================================================================
```

As the summary shows, we obtain an r-square of 15%, much lower than what was obtained in the previous exercise. However, we obtain relationships that are interesting for interpretation and that can become important for the design of policies and strategies to make the price sustainable over time.

Regionally, we have strong variations in where customers are located. This variable is statistically significant for making decisions about a focal point for better sales. Our reference region gets quite high prices and attention compared to the other four regions in the analysis.

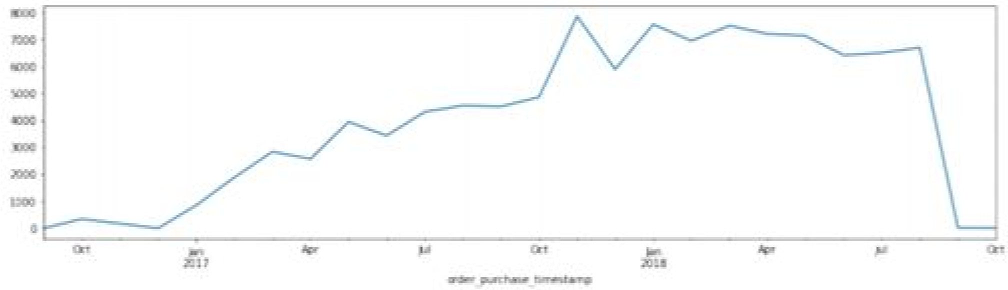
The characteristics of the product, such as volume, weight and photos, have significance for the model. However, it is striking that the more photos you have, the higher the price you will charge. This can be used to strategize about the way products are listed.

Last but not least, reviews have an important weight in this regression. Again, we see an interesting relationship with the average price. An increase in the reviews will increase the price of our product by 1.3 monetary units on average and keep the rest of the model constant.

FEATURE ENGINEERING

In this process of becoming a Data Scientist we are aware that our models are only ever as good as the data we train it on. Here is when the feature engineering comes in, feature engineering refers to a process of selecting and transforming variables when creating a model. The process involves a combination of data analysis, applying rules of thumb, and judgement.

We used the decomposing the timestamp variable from the order_purchase_timestamp into day of the year, day of the week, month, hour of the day, and year to have a better understanding of the data and discovery some relationships, it helped us to represent the structure or the data seasonality in the data.

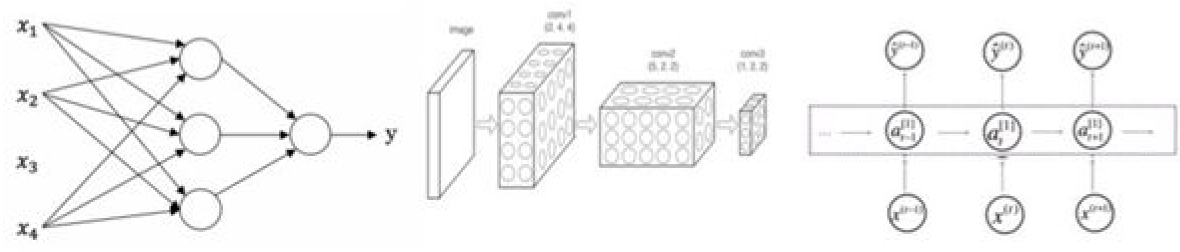


After observing the graphic we notice there are some days where there was not an order, so we have cero values for that period of time, we decided to use imputation and drop the value where there was not a purchase, we delete the dates with no values between the beginning of 2016 and around October of 2018.

In addition, to create our recurrent network models we use the logarithm transformations to handle skewed values for them to adjust more approximate to the normal distribution, also to decrease the effect of the outliers and to make the model more robust.

RECURRENT NEURAL NETWORKS

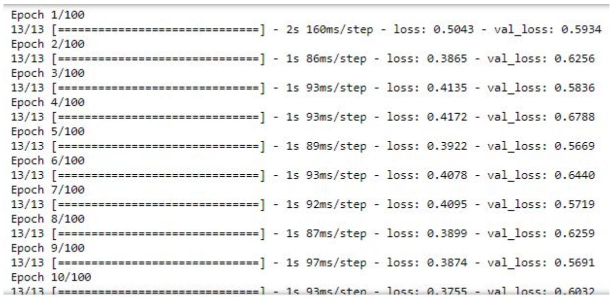
“A neural network is a type of Machine Learning Model. This particular type of model derives its name from how we believe our own neurons work: as a series of simple components, that together can achieve amazing things” taken fromNotebook neural_networks_bootcamp.ipynb



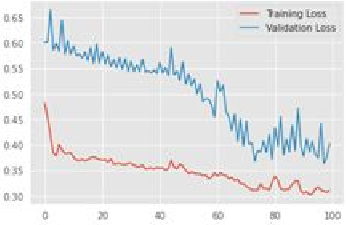
Standard NN Convolutional NN Recurrent NN

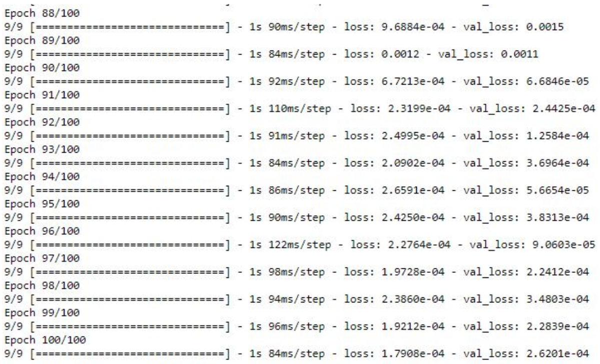
Although Standard NN is commonly used for this kind of models, in this case we use Recurrent NN architecture which is designed for time series, in our case we have sales, prices and reviews on a time series and want to predict future sales from that information.

We define this models to help predtict total sales.

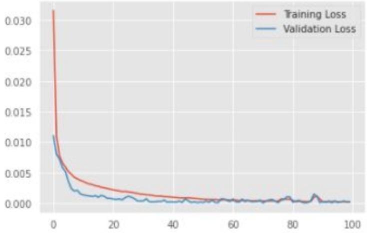


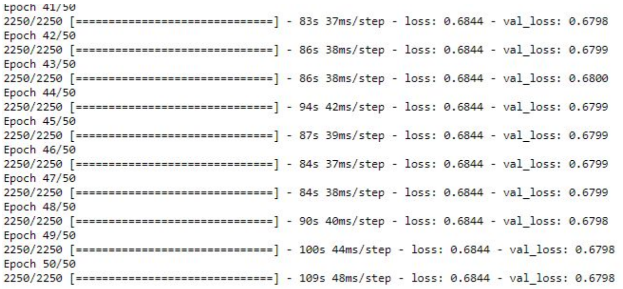
```
Epoch 1/100
13/13 [==============================] - 2s 160ms/step - loss: 0.5043 - val_loss: 0.5934
Epoch 2/100
13/13 [==============================] - 1s 86ms/step - loss: 0.3865 - val_loss: 0.6256
Epoch 3/100
13/13 [==============================] - 1s 93ms/step - loss: 0.4135 - val_loss: 0.5836
Epoch 4/100
13/13 [==============================] - 1s 93ms/step - loss: 0.4172 - val_loss: 0.6788
Epoch 5/100
13/13 [==============================] - 1s 89ms/step - loss: 0.3922 - val_loss: 0.5669
Epoch 6/100
13/13 [==============================] - 1s 93ms/step - loss: 0.4078 - val_loss: 0.6440
Epoch 7/100
13/13 [==============================] - 1s 92ms/step - loss: 0.4095 - val_loss: 0.5719
Epoch 8/100
13/13 [==============================] - 1s 87ms/step - loss: 0.3899 - val_loss: 0.6259
Epoch 9/100
13/13 [==============================] - 1s 97ms/step - loss: 0.3874 - val_loss: 0.5691
Epoch 10/100
13/13 [==============================] - 1s 93ms/step - loss: 0.3755 - val_loss: 0.6032
```



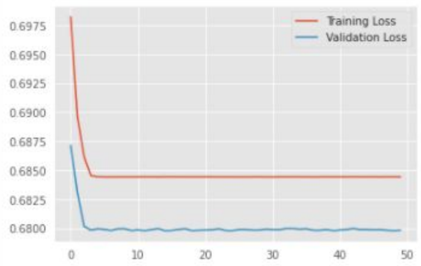


```
Epoch 88/100
9/9 [==============================] - 1s 90ms/step - loss: 9.6884e-04 - val_loss: 0.0015
Epoch 89/100
9/9 [==============================] - 1s 84ms/step - loss: 0.0012 - val_loss: 0.0011
Epoch 90/100
9/9 [==============================] - 1s 92ms/step - loss: 6.7213e-04 - val_loss: 6.6846e-05
Epoch 91/100
9/9 [==============================] - 1s 110ms/step - loss: 2.3199e-04 - val_loss: 2.4425e-04
Epoch 92/100
9/9 [==============================] - 1s 91ms/step - loss: 2.4995e-04 - val_loss: 1.2584e-04
Epoch 93/100
9/9 [==============================] - 1s 84ms/step - loss: 2.0902e-04 - val_loss: 3.6964e-04
Epoch 94/100
9/9 [==============================] - 1s 86ms/step - loss: 2.6591e-04 - val_loss: 5.6654e-05
Epoch 95/100
9/9 [==============================] - 1s 90ms/step - loss: 2.4250e-04 - val_loss: 3.8313e-04
Epoch 96/100
9/9 [==============================] - 1s 122ms/step - loss: 2.2764e-04 - val_loss: 9.0603e-05
Epoch 97/100
9/9 [==============================] - 1s 98ms/step - loss: 1.9728e-04 - val_loss: 2.2412e-04
Epoch 98/100
9/9 [==============================] - 1s 94ms/step - loss: 2.3860e-04 - val_loss: 3.4803e-04
Epoch 99/100
9/9 [==============================] - 1s 96ms/step - loss: 1.9212e-04 - val_loss: 2.2839e-04
Epoch 100/100
9/9 [==============================] - 1s 84ms/step - loss: 1.7908e-04 - val_loss: 2.6201e-04
```





```
Epoch 41/50
2250/2250 [==============================] - 83s 37ms/step - loss: 0.6844 - val_loss: 0.6798
Epoch 42/50
2250/2250 [==============================] - 86s 38ms/step - loss: 0.6844 - val_loss: 0.6799
Epoch 43/50
2250/2250 [==============================] - 86s 38ms/step - loss: 0.6844 - val_loss: 0.6800
Epoch 44/50
2250/2250 [==============================] - 94s 42ms/step - loss: 0.6844 - val_loss: 0.6799
Epoch 45/50
2250/2250 [==============================] - 87s 39ms/step - loss: 0.6844 - val_loss: 0.6799
Epoch 46/50
2250/2250 [==============================] - 84s 37ms/step - loss: 0.6844 - val_loss: 0.6799
Epoch 47/50
2250/2250 [==============================] - 84s 38ms/step - loss: 0.6844 - val_loss: 0.6799
Epoch 48/50
2250/2250 [==============================] - 90s 40ms/step - loss: 0.6844 - val_loss: 0.6798
Epoch 49/50
2250/2250 [==============================] - 100s 44ms/step - loss: 0.6844 - val_loss: 0.6798
Epoch 50/50
2250/2250 [==============================] - 109s 48ms/step - loss: 0.6844 - val_loss: 0.6798
```



CONCLUSIONS

Our study is composed of a vast analysis on the dataset provided by Olist, which included but is not limited to a rigorous descriptive analysis, analysis in linear regression models and analysis on recurrent neural networks for time series. After all this rigorous examination, we can conclude:

One of the most importants, almost 60 % of orders with late deliveries got a 1 star review. This can be a huge improvement opportunity, since depending on a seller's reputation, given the reviews, they may receive more purchases. This is because the more reviews you have, it is possible that a customer will decide to buy from you, as indicated by the marketing strategies.

Working on late deliveries for different products can help achieve the goal of better reviews. A trend was seen in the descriptive analysis, which could then be statistically proven, but which may leave steps to be able to perform more specific studies on it in order to generate the desired impact.

In another analysis on the reviews variable, we again found statistical significance for the explanation of the model. In this case, we got that the reviews have a directly proportional relationship to the average price. Since we have better reviews about the product, on average the price tends to increase. Again, we emphasize the strategy of taking care of each client's review, proposing different strategies.

In one of the models we review, we include all the detailed sales and calculate the model with the variables: price mean by month, product volume, freight value, review score, product photos quantity, and product weight. We can conclude that:

- R-squared in most of linear models give low values, represents a simple model
- Promoting quality and quantity of product photos with sellers could be a good strategy to improve sales, according to models any changes in this variable have a positive impact on the total sales.

Regions can be another sensitive starting point. Our analysis is based on the Northern region, which has higher average prices than the other regions. One of the possible causes of this price increase over this region may be the freight value, which will be higher for this region. One of the possible shock plans may be to negotiate with the delivery companies, to achieve a decrease in these fixed costs and to appropriate this payment from the customer in profits

Recurrent Neural Networks are designed to time series modeling, for our case we build a model to predict sales according to the historical data. Those information with a good defined strategy of the company could help to rise sales and get the best fit to customer needs.