# Capstone Project Proposal

Olist is a SaaS type of company in Brazil that provides software for centralised order management , provides one exclusive contract to have access to all marketplaces and also acts as a large department store through which shopkeepers can sell and hence can share the reputation that Olist has built.

Their business model works with by active subscription of the sellers with Olist and also increasing their brand image at various marketplaces by providing better customer service, product satisfaction and timely delivery. Olist would like to understand their customers and sellers better so as to provide better services to both customers and sellers.

Olist has provided information regarding 100 k orders that were made using Olist platform at various marketplaces between 2016 and 2018. The data includes customer order amounts, geo location of customers, customer reviews , product lists, payment method, order information and seller information.

Analysis of the data can be done from various viewpoints:

1. Understand the customer segments that use Olist using clustering so we can target the customers better
2. Understanding the Customer reviews in order to improve the customer experience
3. Using sales data to predict future trends
4. Improving delivery performance by optimizing delivery times
5. Understanding products which leads to more customer satisfaction
6. Understanding Customer purchase patterns and conducting a behaviour analysis
7. Predicting product demand based on population density, products sold in various geographies and ecommerce friendly geographies

Deliverables

1. Code
2. Slide Deck

# Data Wrangling Process

The Olist Data consists of 9 separate datasets:

1. Customer Dataset

2. Orders Dataset

3. Geolocation Dataset

4. Order Payment Dataset

5. Order Reviews Dataset

6. Order Item Dataset

7. Products Dataset

8. Sellers Dataset

9. Product Category Name Translation Dataset

In the customer dataset, there were 96096 unique customers (Customer\_Unique\_Id) with 3345 customers having repeat purchases. As there is a customer id created for each order placed, there are 99441 customer ids. There was no missing value in this data set.

In the orders dataset, we can link with customer dataset using the Customer\_Id as key. As there are different order status like Delivered, Cancelled, Shipped, Unavailable etc.. only Delivered has been considered as analysis is going to be done on orders that have been delivered(97% of the data). That reduced the data points from 99441 to 96478.

There were further 65 missing values in various rows which were removed from orders dataset to bring final number to 96455.

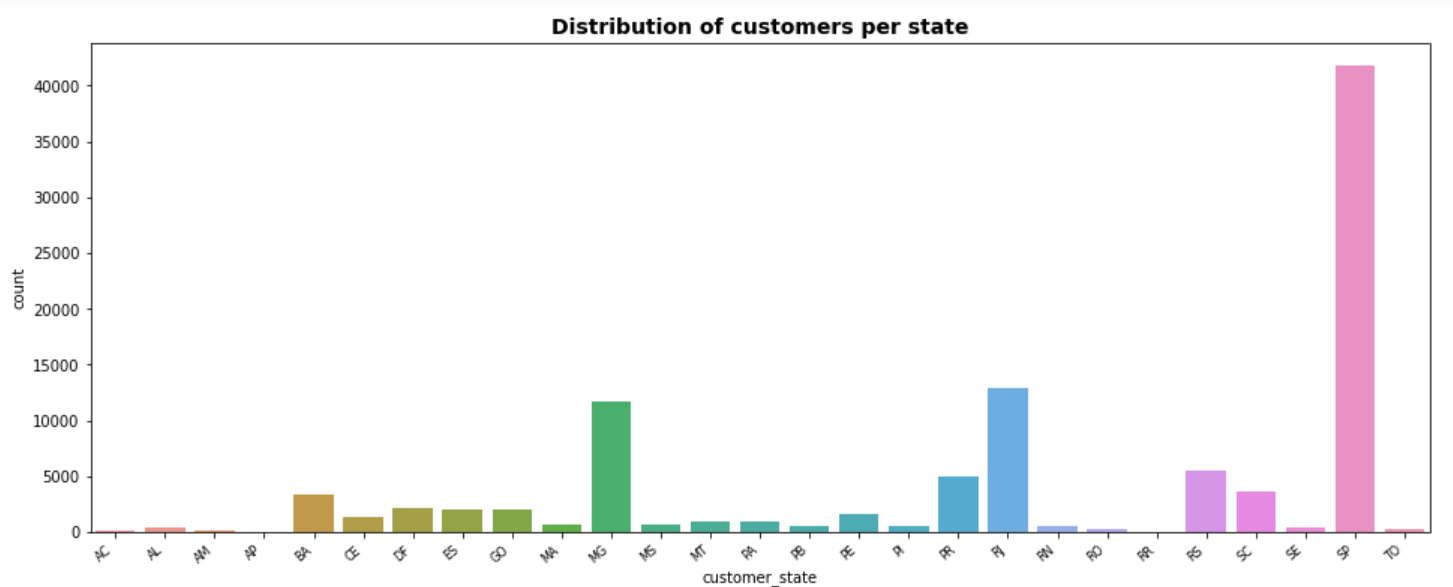
There were no missing values in Order Payments Data and also Order Items data. But in the Order Reviews data, there were many data points missing but that was because most customers had not given a review. As the number of missing values is large, all that can be done is ignore the reviews columns and only consider the product rating column.

Products data consists of weight and dimensions of product as well a product description in Portuguese that be converted to English by joining the data set with the poduct category English translation dataset. Initial data wrangling does not show any alarming outliers.

# Data Story Telling and Exploratory Analysis

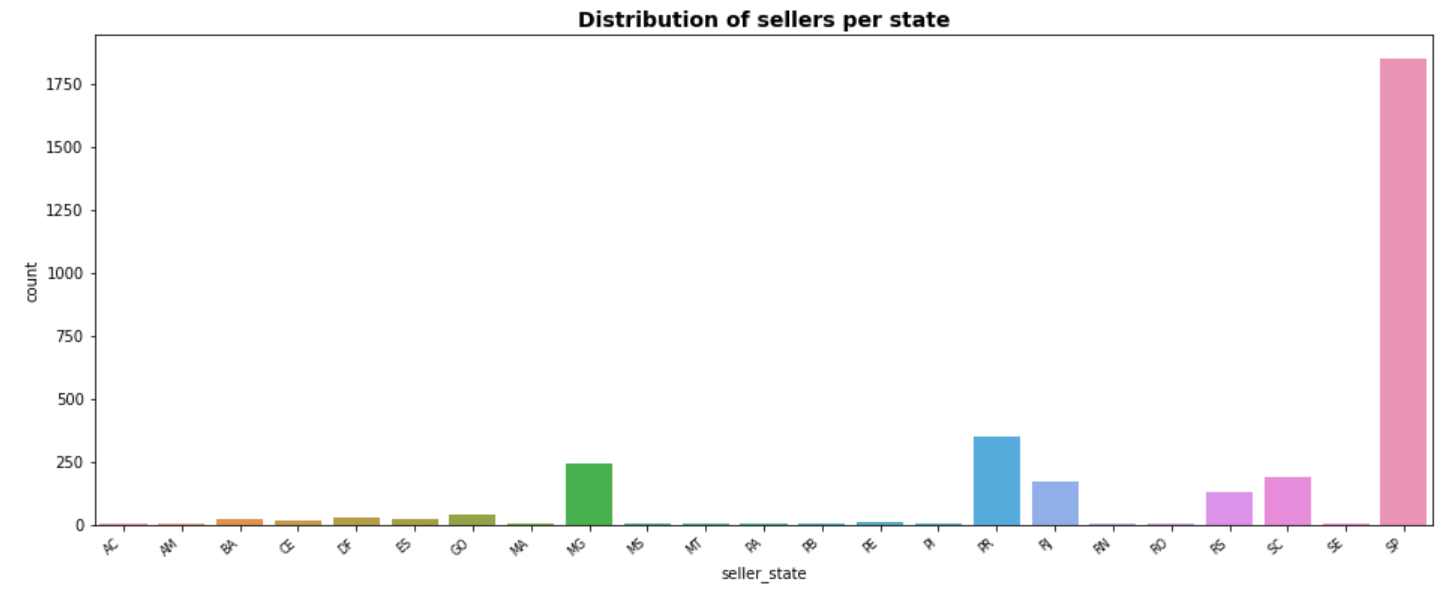
Olist is an e-commerce platform that has tapped into the Brazil market and has many sellers registered on the platform from which the customers can chose the best price from. From the data that is available, there are many questions that can be analysed:

1. Where so we have the most customers from?

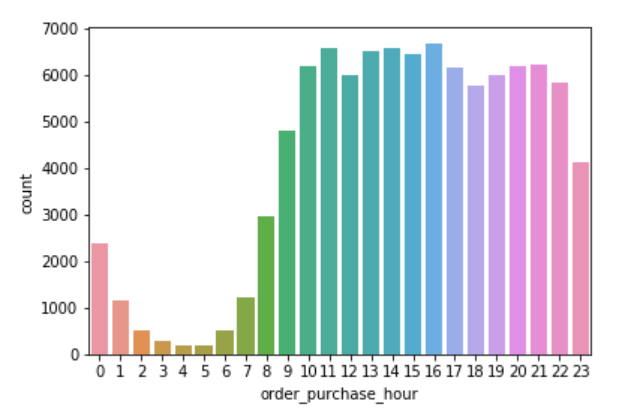
As we can see, almost 42% of the customer base is from Sao Paulo, 13% from Rio Di Janerio, 12% from Minas Gerais and the rest 33% of the customers are distributed among the other states of Brazil

1. Where are most of the sellers located?

Out of 3095 sellers, approximately 60 % of the sellers are also from Sao Paulo.

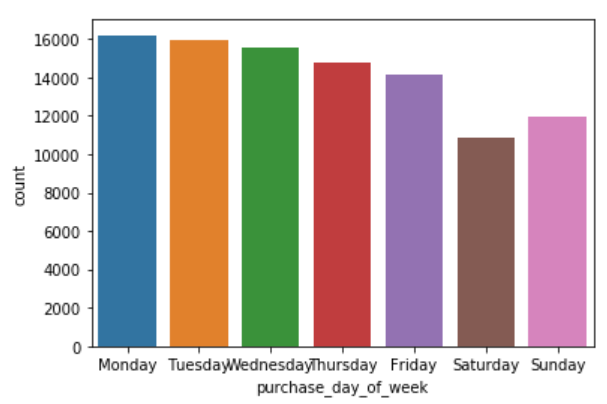


1. What time of the day do we see the customers making the highest number of purchases?

As seen, most purchases happen after 9 am and slowly starts reducing after 4 pm.

1. Which day of the week have the highest purchases?

As seen, we see customers shopping more on the start of the week and lesser during the weekend.

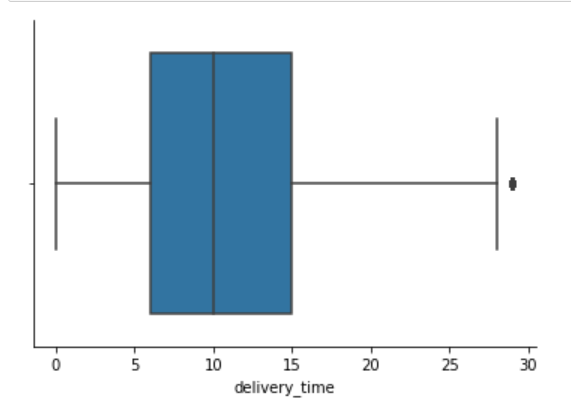


1. Which months see the highest purchases?



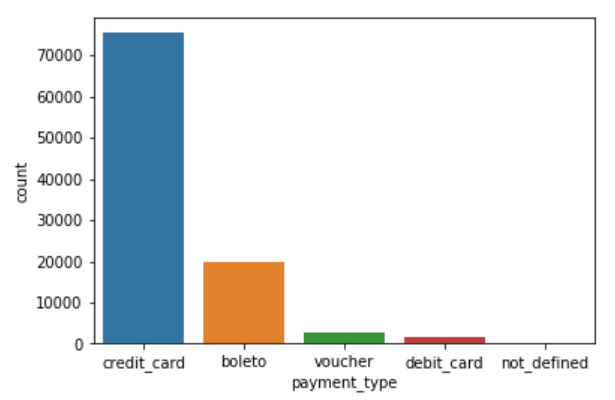
1. What is the average delivery time of a product?

The median Delivery time of orders is 10 days with a standard deviation of around 6 days. This is after removing 4% of the outliers.

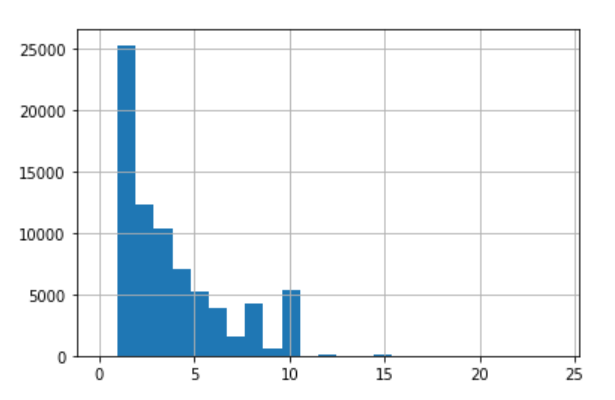


1. What is the most common method used for payments?

Around 74% of customers use Credit Card which is the most common form of payment followed by Boletos that are used for 19% of the purchases.

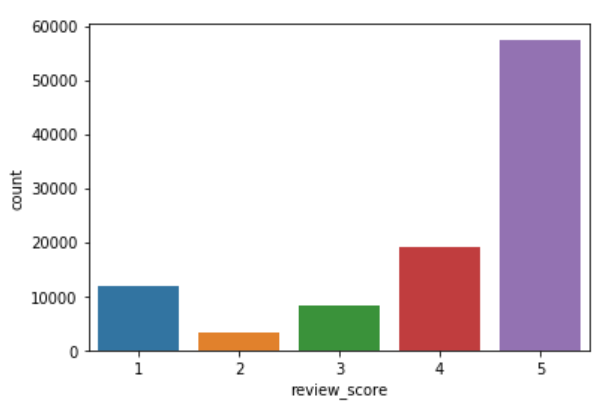


1. In Credit Card payments, how many instalments do customers generally pay?



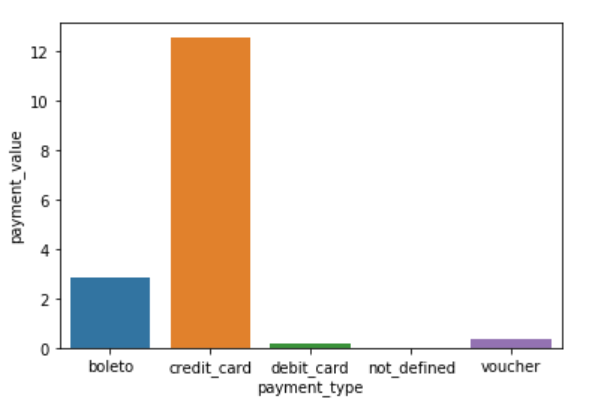
1. What is the most ratings given by customers?

As seen, only 15% percent of customers gave below average ratings and around 77% of customers gave above average ratings with 57% of customers giving 5 star ratings



1. What is the revenue from different payment methods?

Of the overall 16 million revenue over 2 years, credit card payments contribute to 12 million(75%) of the revenue



# In-depth Analysis (Machine Learning)

Key questions to be analysed using the machine learning:

1. Does the mode of payment impact the purchasing power of the customer?

For understanding this relationship, as payment type was a categorical variable, it was initially converted into dummy variables after which a OLS Regression was done to see if there was any relation between Payment Value of customer and Payment method.

**Code Used:**

from statsmodels.formula.api import ols

fit = ols('payment\_value ~ C(payment\_type\_boleto) + C(payment\_type\_credit\_card) + C(payment\_type\_debit\_card)+C(payment\_type\_voucher)', data=df\_f).fit()

fit.summary()

**Results:**



As seen, there is no relation between payment method and purchasing behaviour. Also, as noted previously 75% of the purchases are done on credit and it also accounts to 75% of the sales as well.

1. Is Delivery time affected by:
   1. Distance between customer and seller
   2. Volume of order
   3. Weight of order
   4. Freight value

To check this, I initially used OLS regression but saw a very small R squared value of 0.2 and also did Lasso regression to pick up the important features for the regression which also showed that none of the features are important. From this we can conclude, none of these factors affect the delivery time and hence delivery time is something that could be standardised and SLAs could be set.

**OLS Regression Code Used:**

import statsmodels.api as sm

X = sm.add\_constant(X)

est = sm.OLS(Y, X).fit()

est.summary()

**Result:**



**Lasso Regression Code:**

from sklearn.linear\_model import Lasso

names=df\_g.drop(['order\_id'],axis=1).columns

len(names)

# Instantiate a lasso regressor: lasso

lasso = Lasso(alpha=0.4,normalize=True)

# Fit the regressor to the data

lasso\_coef=lasso.fit(X,Y).coef\_

# Compute and print the coefficients

lasso\_coef = lasso.fit(X,Y).coef\_

print(lasso\_coef)

# Plot the coefficients

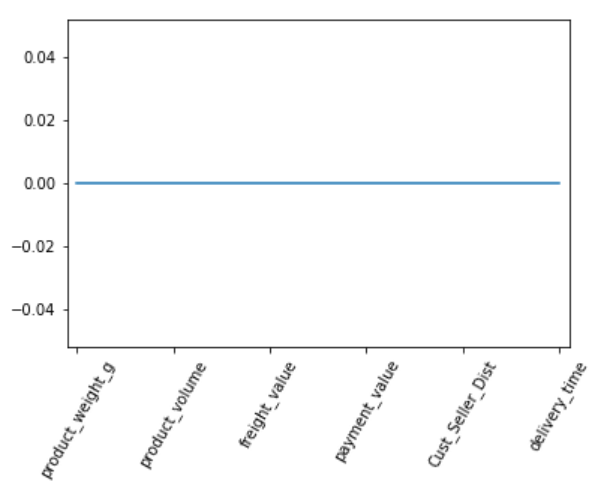
plt.plot(range(len(names)), lasso\_coef)

plt.xticks(range(len(names)), names.values, rotation=60)

plt.margins(0.02)

plt.show()

**Lasso Regression Results:**



As you can see, none of the features are significant to affect delivery time.

1. What factors do the customer rating depend upon?
   1. The delivery time
   2. Difference between estimated and actual delivery time

**Code Used:**

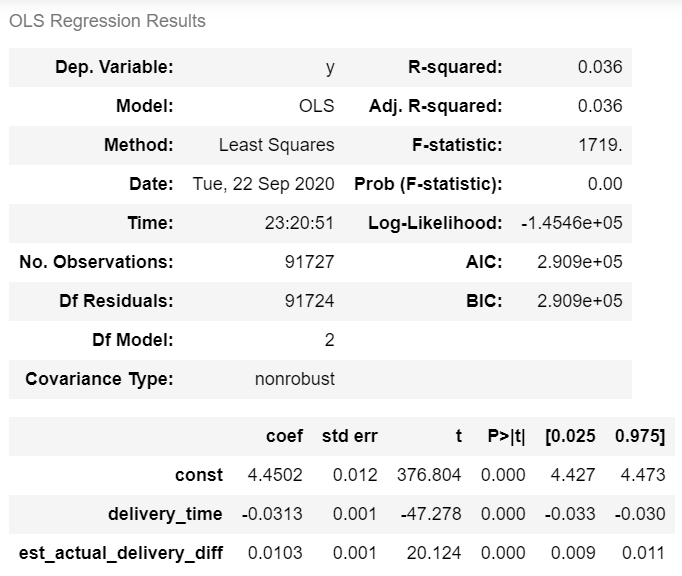
import statsmodels.api as sm

X1 = sm.add\_constant(X1)

est = sm.OLS(Y1, X1).fit()

est.summary()

**Result:**



As seen , there is no significant relationship between the delivery time or even anticipated delivery time- actual delivery time and the ratings of the customer.