# DSC 630 – PROJECT PRELIMINARY ANALYSIS

Project Milestone - 3

JULY 17, 2022

KARTHIKEYAN CHELLAMUTHU, SUBHASHINI NATARAJAN

# Table of Contents

**Project Milestone 2: Data selection and planning**



**Project Milestone 3: Data selection and planning**

**Project Milestone 2: Data selection and planning**

**Objective:**

The business problem that we plan to build the model for is, customer segmentation, for the Brazilian e-commerce company, Olist. Customer Segmentation is an important step in marketing for an organization. The process helps in customizing marketing campaigns, prevent customer churn, prioritizing product development or services.

**Data Source:**

Data for the project is sourced from Kaggle. This is a public dataset of orders made at Olist store. This dataset has several dimensions including, customer, geolocation, order, payment, reviews, product, product category and sellers. For the purposes of customer segmentation, not all dimensions will be considered, but only the following will be used -

- Customer,
- Order
- Customer review

**Link to the dataset:**

https://www.kaggle.com/code/marianakralco/brazilian-ecommerce-clustering-rfm-and-kmeans/data

**Planned Process:**

The process to be followed for the project is laid out below -

1. Data ingestion – Load the multiple data files into panda dataframes.
2. Exploratory Data Analysis – Perform EDA to understand the dataset. Plot essential graphs for the understanding.
3. Data transformation – Based on EDA results, perform necessary data transformation required for modeling.
4. Feature selection – Identify and select the features required for the model building.
5. Build the model.
6. Evaluate the model.
7. Determine if the model can be deployed for solving the business problem.

**Planned model selection:**

The model that we plan to build for customer segmentation is k-means clustering. The model helps in classifying the customer groups based on purchase history and customer review. We chose K-means clustering due to the following reasons –

- Scalability for large datasets – The data set we chose has three years sales data.
- Familiarity and relatively simpler to implement.
- Possibility for generalization

**Planned model evaluation:**

While clustering algorithms have multiple evaluation metrics like, homogeneity score, v-measure, silhouette coefficient, we intend to use, Adjusted Rand Index (ARI), due to the possibility of larger number of clusters with three years data.

**Learning objective:**

Through our prior projects we have learnt, building linear regression for trend analysis, logistic regression for classification problems. By performing this project, we aim to learn the process involved in customer segmentation through clustering. We aim to learn to use elbow method to determine the optimal number of clusters (optimal k-value), determine the clusters, evaluate the model and interpret the various clusters. Also, will learn to plot the clusters for better visualization.

**Risks:**

The three-year dataset, if has data quality issues or has imbalanced classes will be a risk to the project. Secondly, if the clustering score is not good, then it implies that k-means will not be the correct clustering algorithm to use for segmentation.

**Contingency plan:**

If the classes are imbalanced then will use SMOTE to balance them. If ARI has close to 0 value, it implies that k-means is not the correct clustering algorithm. In this case will manually determine RFM (Recency, Frequency, Monetary value) and cluster the customers based on RFM score.

# Project Milestone - 3

**Preliminary Analysis:**

The data for the project includes customer data, order, order reviews, order items, product data in separate csv files. The individual csv files are imported in separate python dataframes. The individual data frames are then merged into a single data frame, called olist_df, with required fields.

The final data frame is inspected using .info() method and the data frame data types were listed as below –

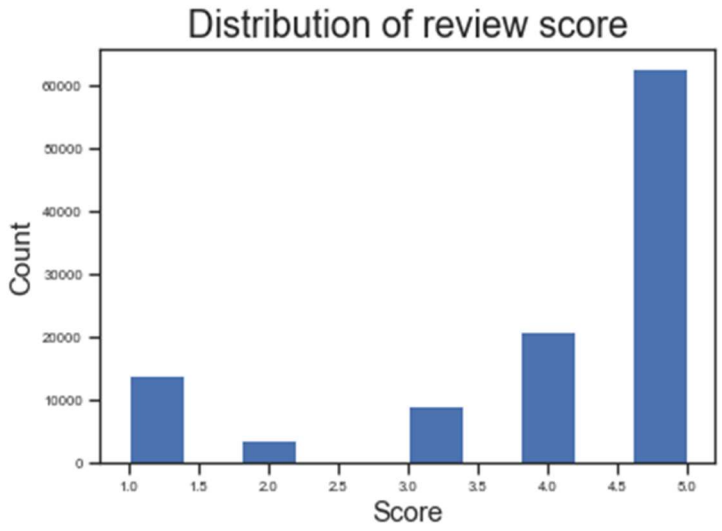
```
customer_id                 110774 non-null object
customer_unique_id          110774 non-null object
order_id                    110774 non-null object
product_id                  110774 non-null object
review_id                   110774 non-null object
order_purchase_timestamp    110774 non-null object
customer_city               110774 non-null object
product_category_name       110774 non-null object
review_score                110774 non-null int64
order_item_id               110774 non-null int64
review_creation_date        110774 non-null object
price                       110774 non-null float64
freight_value               110774 non-null float64
```

It is observed from the above list that order_purchase_timestamp is populated as an object. In order to obtain trends, the data type of the field has to be converted into date using to_datetime() method.

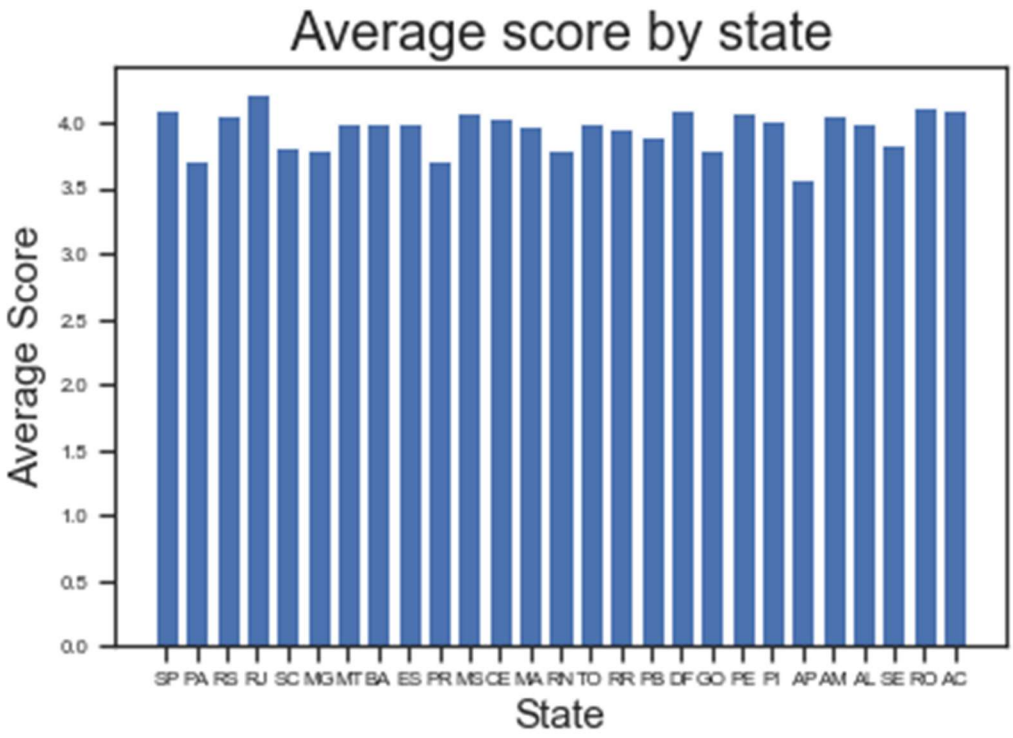
**Data explorations:**

The following data explorations were performed to understand the data more.

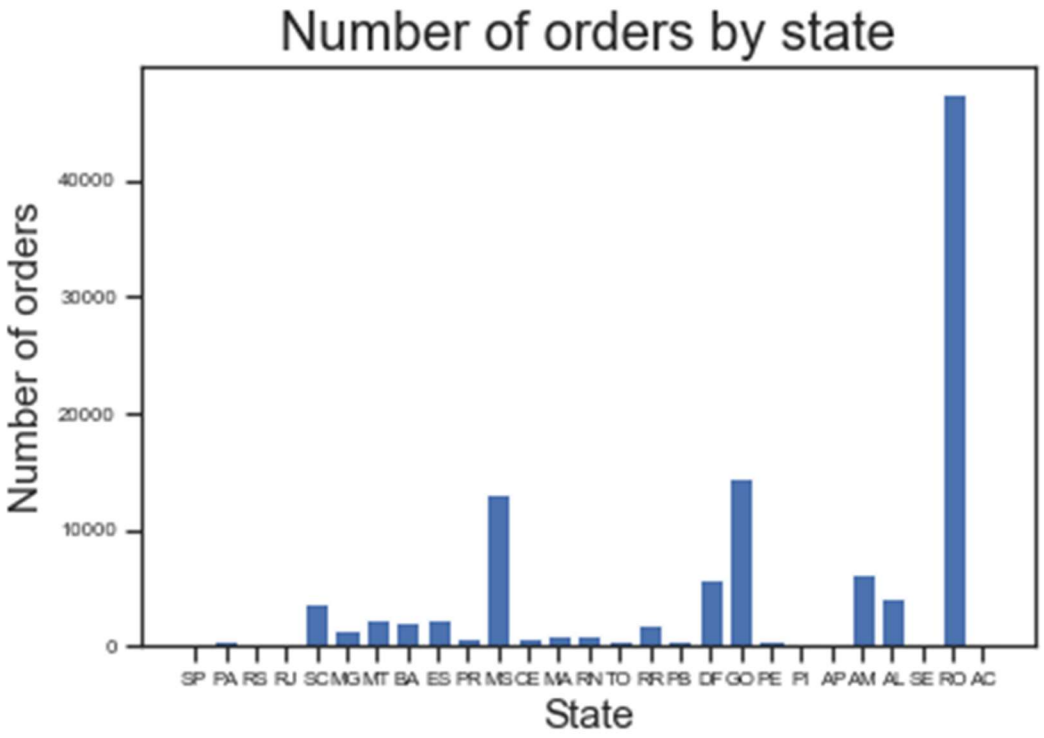
1. Obtain the distribution of review score – The below given distribution shows that number of records with higher ratings are quite high. This implies good quality products and service from Olist.



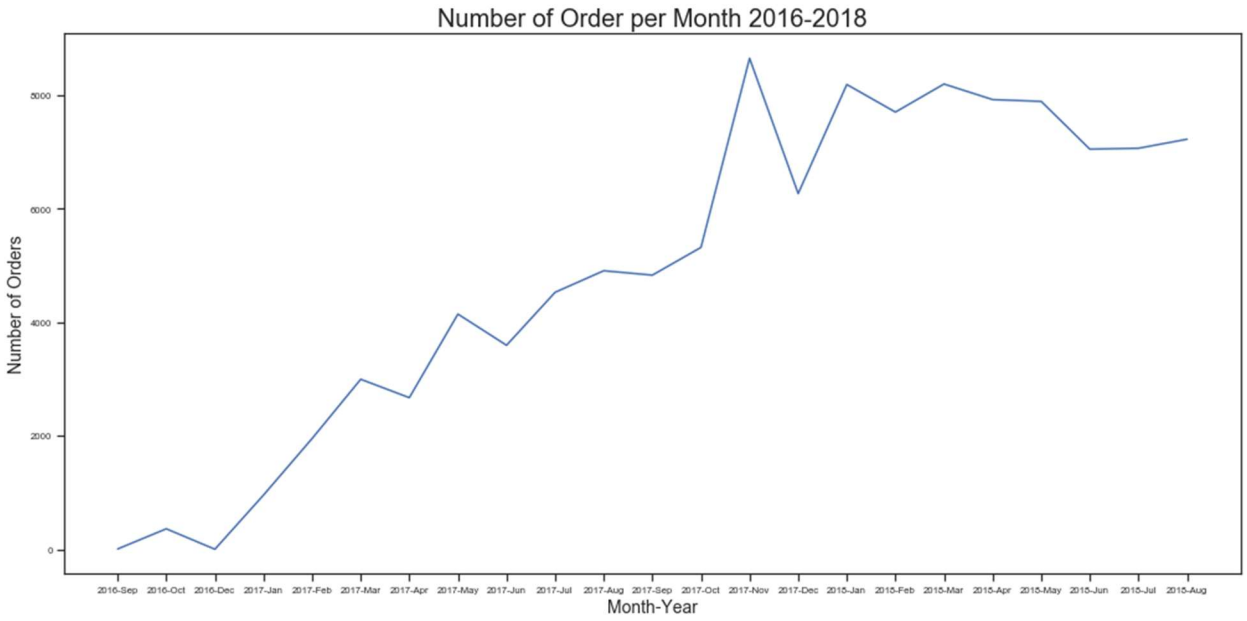

2. To ascertain any regional preferences, did a bar chart of mean review score by state. The plot below shows no strong preferences by state.



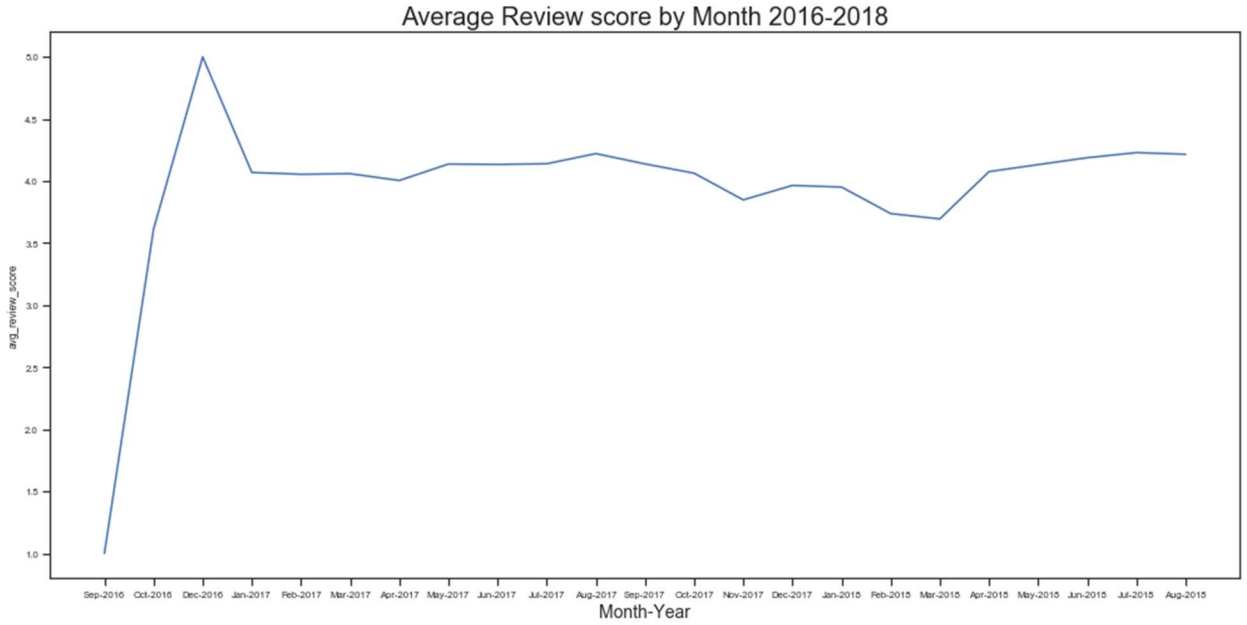

3. To ascertain the number of orders by state, did a bar chart of order count by state. The visualization below indicates that Olist is popular in 5 states and the most popular in Rondonia (RO).



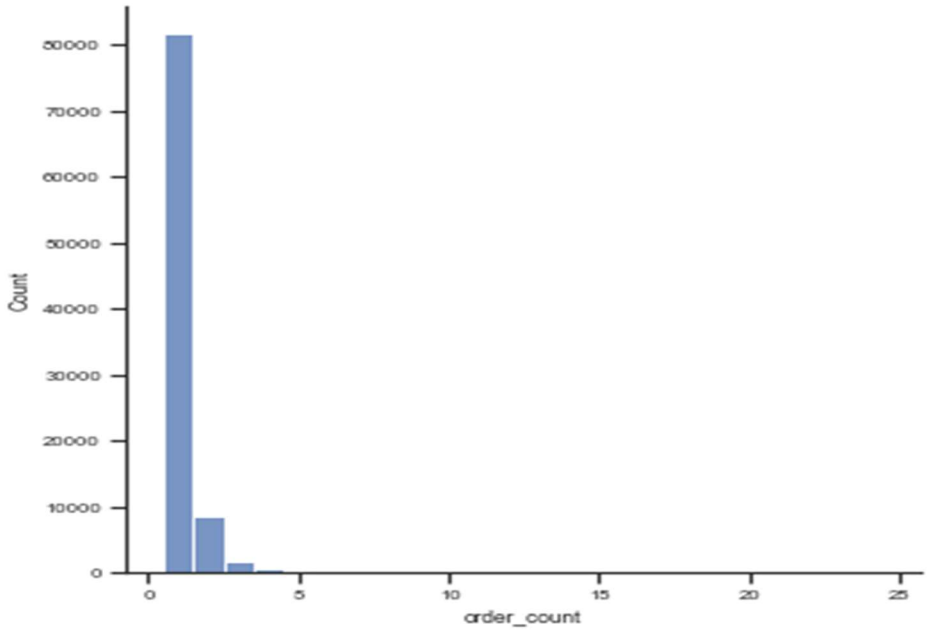

4. In order to plot the trend, grouped the data set based on year – month of order timestamp and plotted the number of orders against the year-month dataset. The trend is shown in the chart below. The number of orders from the plot above can be seen with steady increase apart from few occasional dips and the latest data at the end of 2018 is well above 70000 orders.



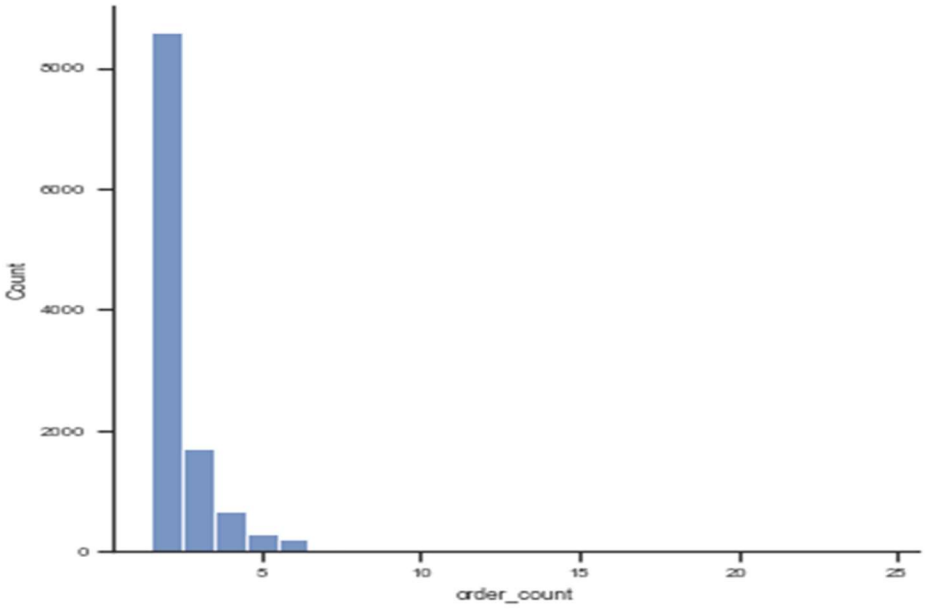

5. Mean review score was also plotted against Month – Year. The average review score for the products from the plot below can be seen as pretty constant and above 4, except for a brief dip between Jan 2018 - April 2018.



6. In order to check the distribution of orders placed by the same customer, performed a distplot for number of orders. The number of one-time orders is the highest, followed by twice and thrice.




7. In order to check the distribution of multiple orders placed by the same customer, performed a distplot for number of orders, where the number of orders is more than 1. Number of three orders and more, by the same customer is lesser than 20,000.




**Milestone 3 Summary:**

By exploring the data, it is observed that we can perform the customer segmentation for the customers provided in the dataset. The plots that were helpful in understanding the dataset were histograms/distribution plot, trend line and bar charts. The data didn't require significant adjustments, but only the require data type change, summarization at required levels. For the month of Sep-2018, there was only order. We considered this as an outlier and removed the data set as part of data preparation. With the required attributes, correctly available, RFM model and k-means clustering model still hold good. Our original expectations still hold good as well.

**Reference:**

https://www.kaggle.com/code/marianakralco/brazilian-ecommerce-clustering-rfm-and-kmeans/data