Order Ratings, Sales, and Customer Relationship on a Brazilian E-Commerce Website

Adam Foley, Mann Purohit,
Nhat Pham, Sarvagna Shukla

olist
empowering commerce

Agenda

❏ Background

❏ Data Collection, Cleaning, and Loading

❏ Objectives

❏ Data Processing, Modeling and Results

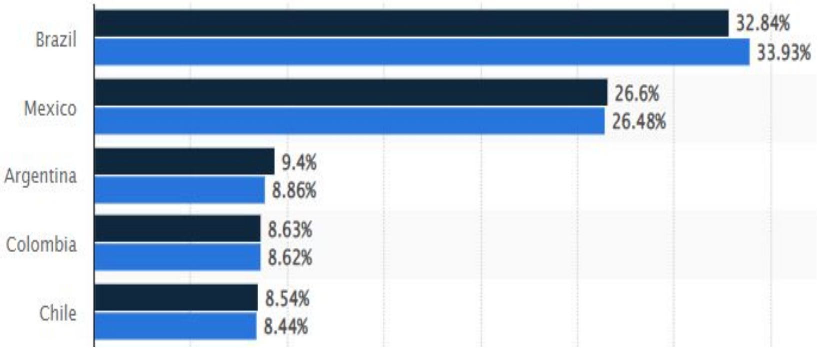
❏ Discussions

❏ Future Work

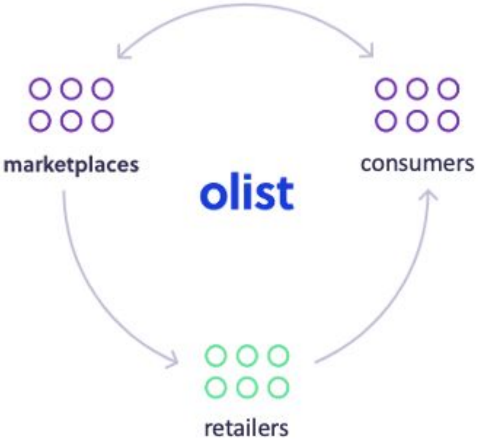
Background

Brazil has the biggest and fastest growing e-commerce market in Latin America

Market is dominated by several large-size marketplaces rather than by a few like in the US (Amazon, Ebay etc.)



Top 5 e-commerce market in Latin America (2020-21) (Statistica)

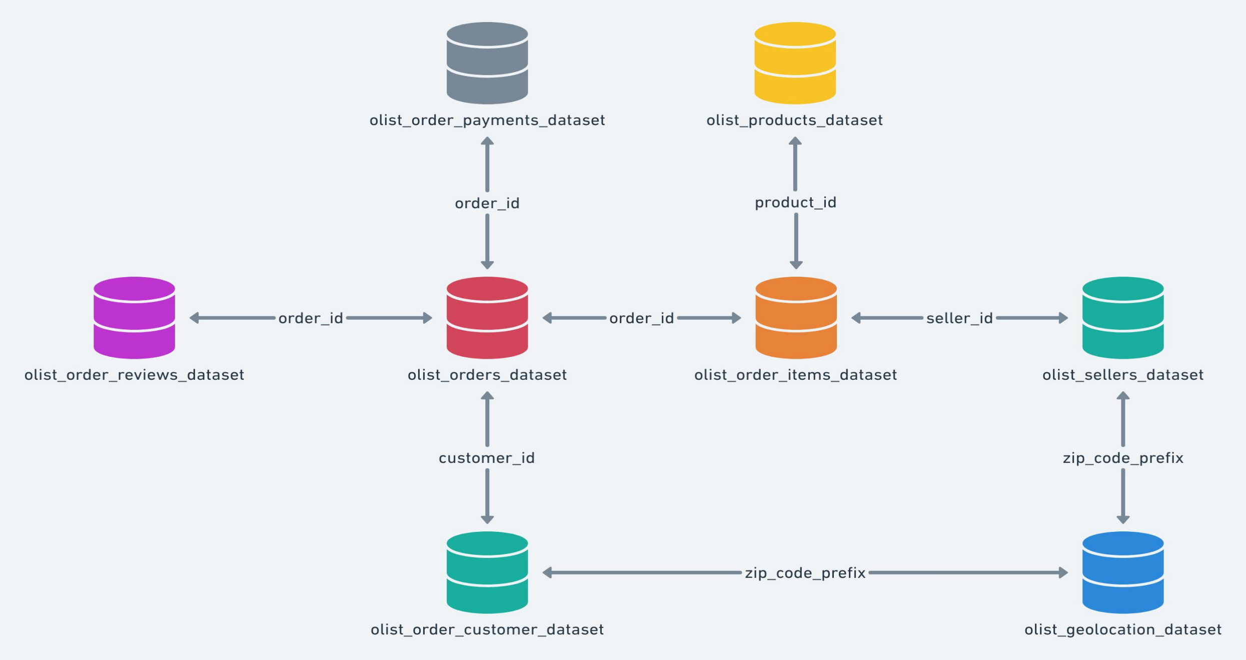


Olist: *"Marketplace of Marketplaces"*

Ecommerce marketplace integrator: provides full stack operational support to merchants (inventory, pricing, fulfillment, customer service, payments), connecting them to larger product marketplaces

Data Collection

The Olist data was stored as tables in csv format in the following Schema



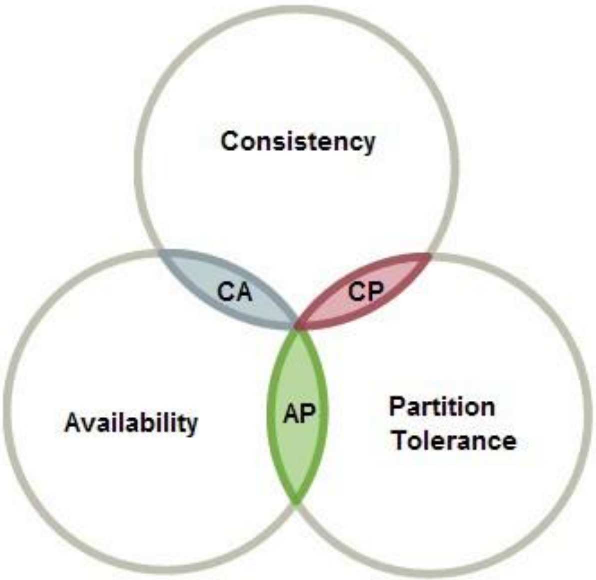
Data Cleaning Example

order_id	payment_sequential	payment_type	payment_installments	payment_value
1	1	voucher	1	89.46
1	2	voucher	1	23.99
1	3	voucher	1	4.78

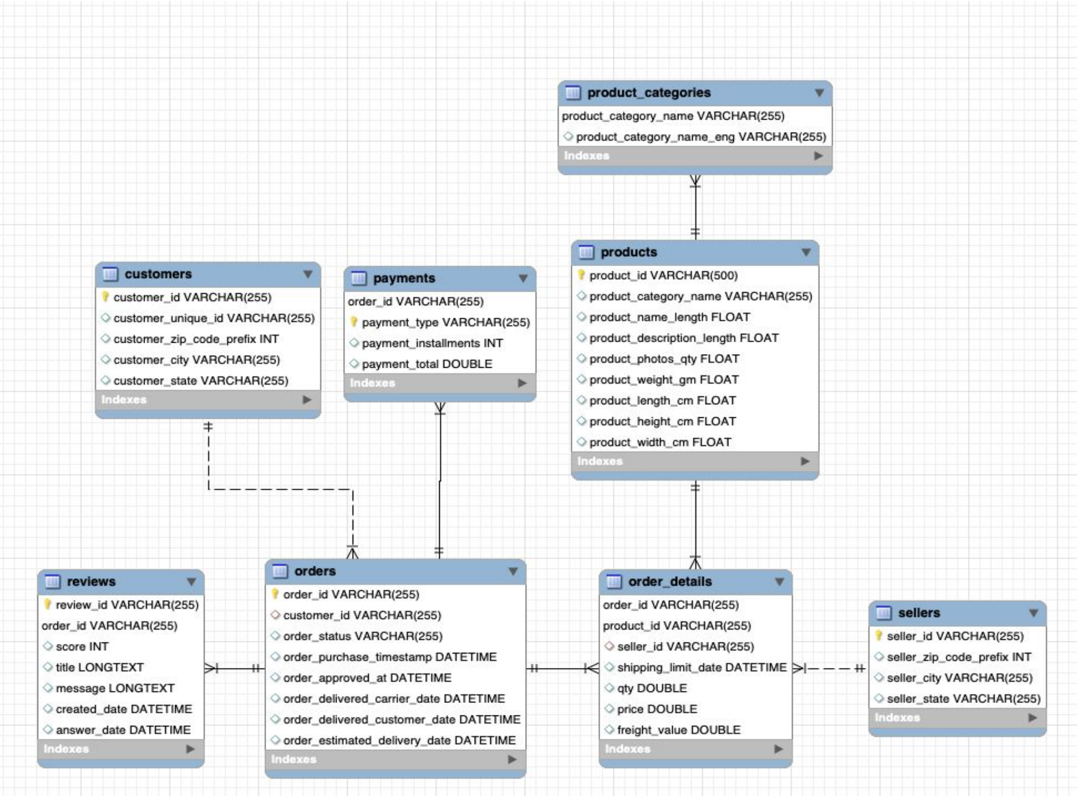
order_id	payment_type	payment_installments	payment_value
1	voucher	3	118.23

Leveraged a Relational Database to ensure Consistency and Availability

Loaded 400k rows of data into an AWS MySQL Instance



Our Final Schema



product_categories
product_category_name VARCHAR(255)
product_category_name_eng VARCHAR(255)
Indexes

customers
customer_id VARCHAR(255)
customer_unique_id VARCHAR(255)
customer_zip_code_prefix INT
customer_city VARCHAR(255)
customer_state VARCHAR(255)
Indexes

payments
order_id VARCHAR(255)
payment_type VARCHAR(255)
payment_installments INT
payment_total DOUBLE
Indexes

products
product_id VARCHAR(500)
product_category_name VARCHAR(255)
product_name_length FLOAT
product_description_length FLOAT
product_photos_qty FLOAT
product_weight_gm FLOAT
product_length_cm FLOAT
product_height_cm FLOAT
product_width_cm FLOAT
Indexes

reviews
review_id VARCHAR(255)
order_id VARCHAR(255)
score INT
title LONGTEXT
message LONGTEXT
created_date DATETIME
answer_date DATETIME
Indexes

orders
order_id VARCHAR(255)
customer_id VARCHAR(255)
order_status VARCHAR(255)
order_purchase_timestamp DATETIME
order_approved_at DATETIME
order_delivered_carrier_date DATETIME
order_delivered_customer_date DATETIME
order_estimated_delivery_date DATETIME
Indexes

order_details
order_id VARCHAR(255)
product_id VARCHAR(255)
seller_id VARCHAR(255)
shipping_limit_date DATETIME
qty DOUBLE
price DOUBLE
freight_value DOUBLE
Indexes

sellers
seller_id VARCHAR(255)
seller_zip_code_prefix INT
seller_city VARCHAR(255)
seller_state VARCHAR(255)
Indexes

Learning Objectives

Analysis of customer satisfaction based on order reviews
Exploratory analysis of sales

Customer segmentation analysis to classify customers into categories based on important features
Sentiment analysis of product reviews (textual data) to discern customer's preferences, likes and dislikes

Data Processing, Modeling, and Results

olist

empowering commerce

Learning Object 1: Understanding Review Scores

Review Score Classification

🛒 Online **reviews highly influence customer buying decisions**, hence product sales and revenue.

❓ How can Olist leverage their order review data to attract more shopkeepers through **enhancing its services**, and attract more consumers through **providing higher satisfaction** ?

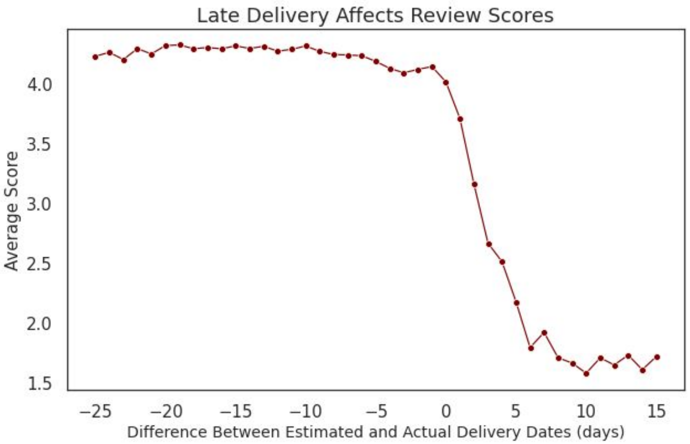
💡 Through EDA and machine learning models for classification, we are interested in identifying key features that affect a good or bad review score, using **reviews, orders, customers, sellers, products, payments** information

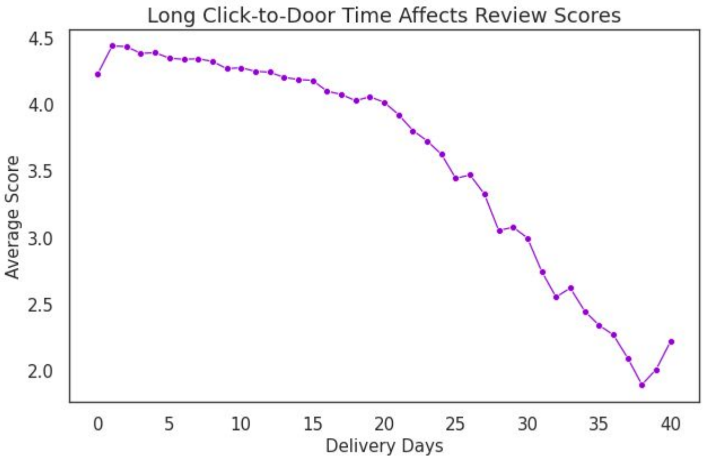
Data Exploration: Orders

Issue with deliveries

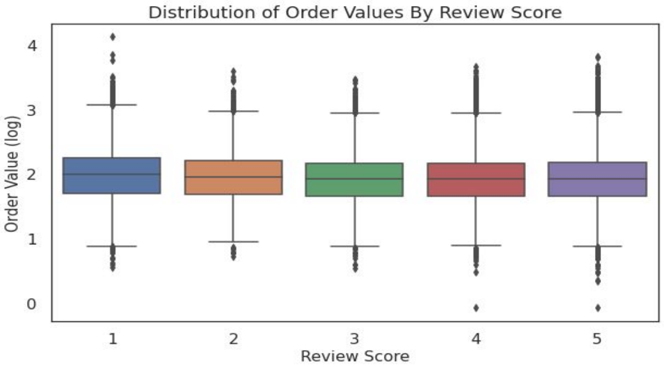
=

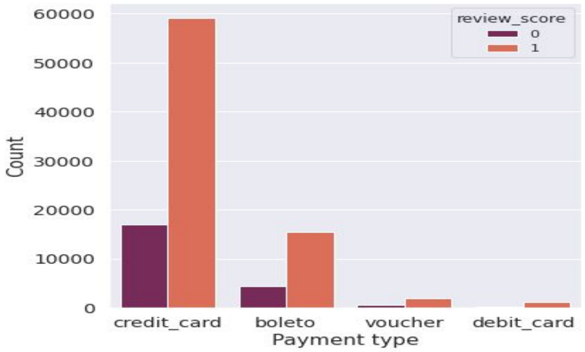
Lower review score





Order properties (photos, description, name, payment type) do not seem to affect the scores

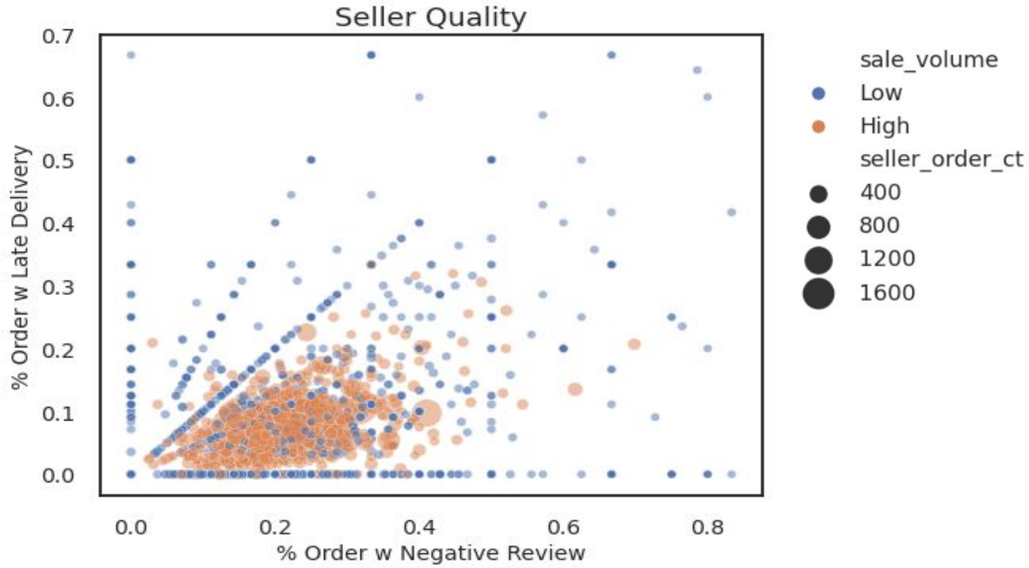


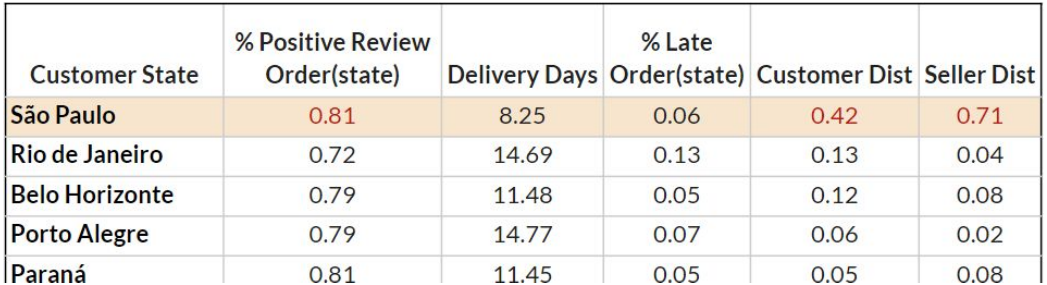


Data Exploration: Seller & Customer Profiles

Compared to small sellers, high volume sellers tend to have lower % of delayed order and relatively lower % of bad reviews

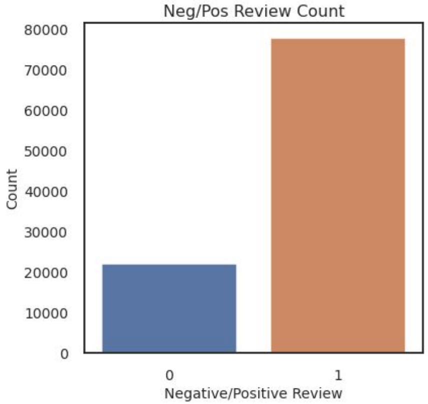
Majority of sellers and customers are from Sao Paulo, which may explain faster delivery and hence better review scores

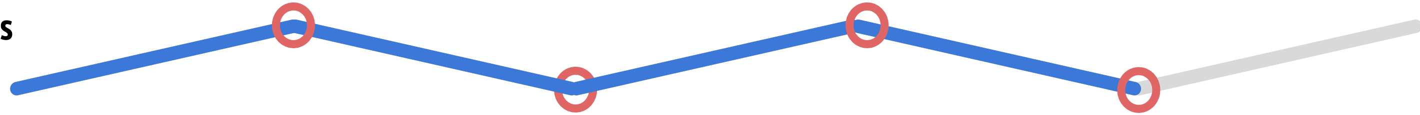




Customer State	% Positive Review Order(state)	Delivery Days	% Late Order(state)	Customer Dist	Seller Dist
São Paulo	0.81	8.25	0.06	0.42	0.71
Rio de Janeiro	0.72	14.69	0.13	0.13	0.04
Belo Horizonte	0.79	11.48	0.05	0.12	0.08
Porto Alegre	0.79	14.77	0.07	0.06	0.02
Paraná	0.81	11.45	0.05	0.05	0.08

Modeling

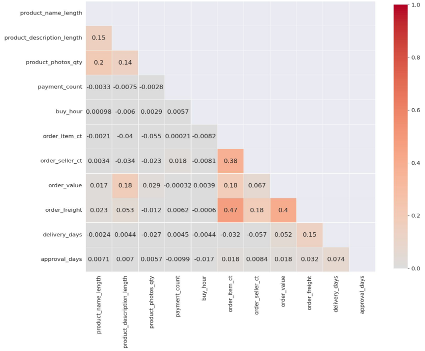


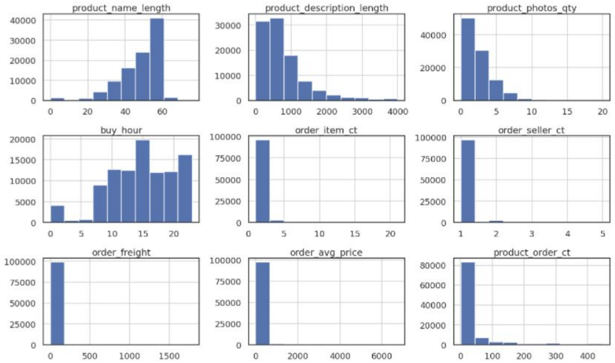


Logistic Regression,
Random Forest

Tidy Data:
~99,500 orders
37 features

Results



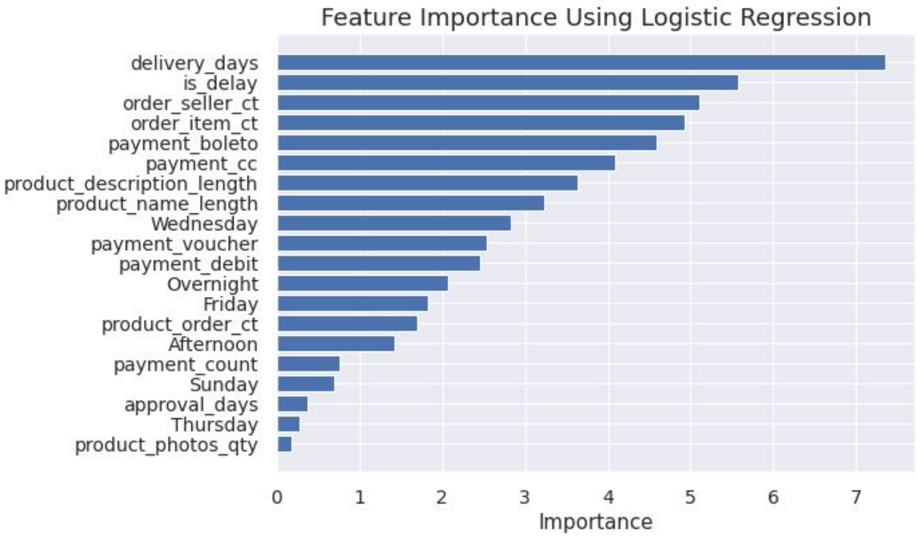


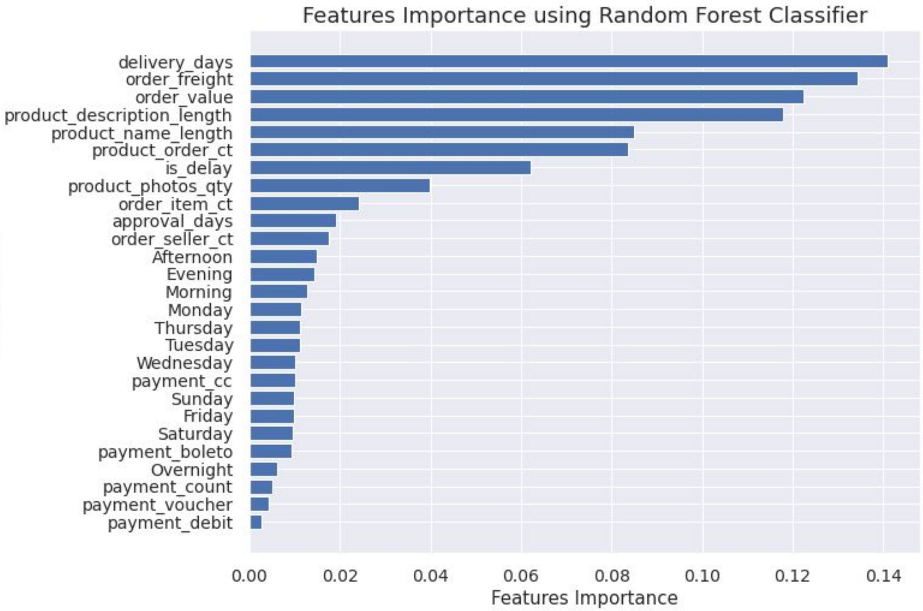
Results

LR: Logistic Regression RF: Random Forest	Original (LR)	Log-transformed (LR)	Oversampling (LR)	Undersampling (LR)	Original (RF)	Oversampling (RF)	Undersampling (RF)
Accuracy	0.82	0.81	0.76	0.76	**0.82**	**0.82**	0.68
Sensitivity	0.82	0.83	0.86	0.85	**0.83**	**0.84**	0.86
Specificity	0.68	0.63	0.45	0.44	**0.73**	**0.65**	0.36
Balanced	0.75	0.73	0.65	0.65	**0.78**	**0.74**	0.61
Precision (0)	0.68	0.63	0.45	0.44	**0.73**	**0.65**	0.36
Recall (0)	0.29	0.28	0.48	0.47	**0.32**	**0.35**	0.59
F1 (0)	0.41	0.39	0.47	0.46	**0.44**	**0.46**	0.44
Precision (1)	0.83	0.83	0.86	0.85	**0.83**	**0.84**	0.86
Recall (1)	0.96	0.95	0.84	0.83	**0.97**	**0.95**	0.70
F1 (1)	0.89	0.89	0.85	0.84	**0.90**	**0.89**	0.77

	Original (LR)	Original (RF)
Precision (0)	0.68	**0.73**
Recall (0)	0.29	**0.32**

(need to change)


Feature Importance Using Logistic Regression


Features Importance using Random Forest Classifier

Learning Object 2: Sales Forecasting

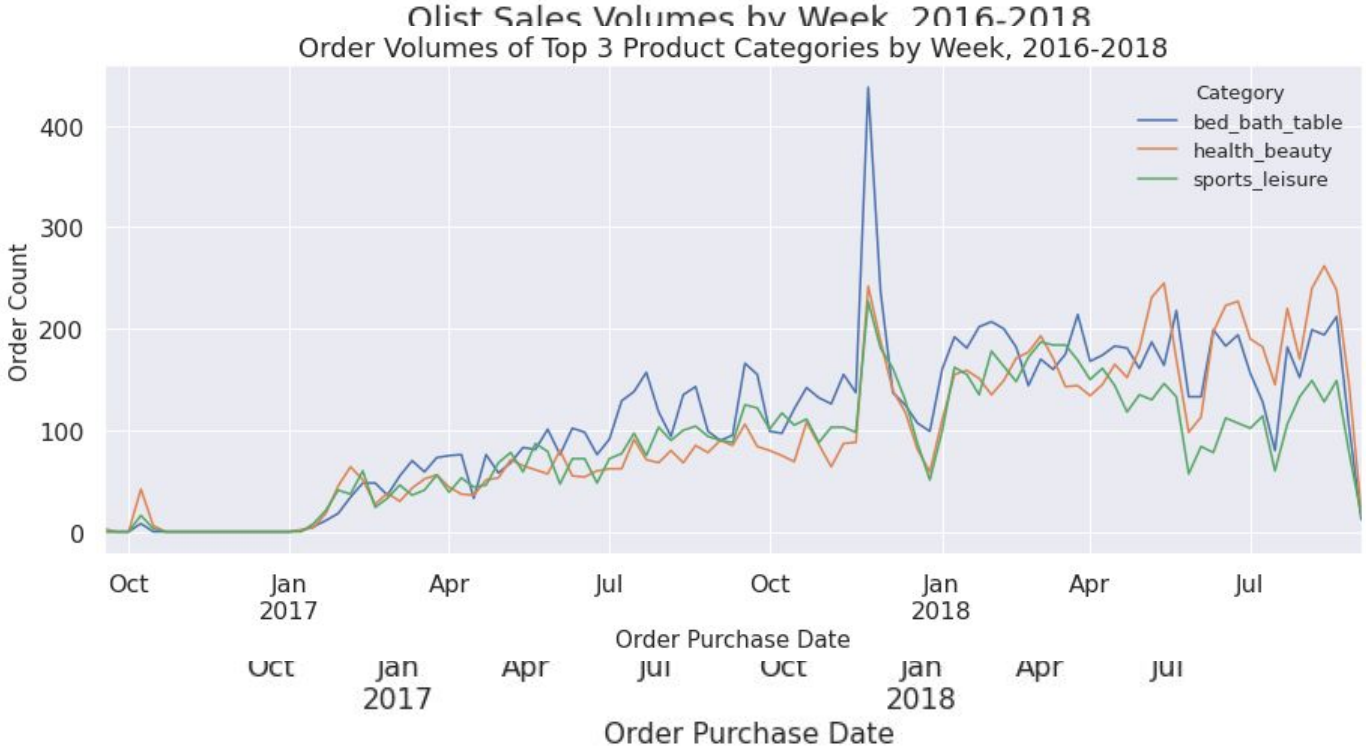
Sales Analysis

🛒 Olist serves a wide market with no targeted segment. Understanding sales will help them strategize business decisions, maximize growth and revenue

❓ What is the general **business trend?** What are the **best-performed sectors?**

💡 Using EDA as a means to understand these trends

Sales Analysis: Trend & Seasonality

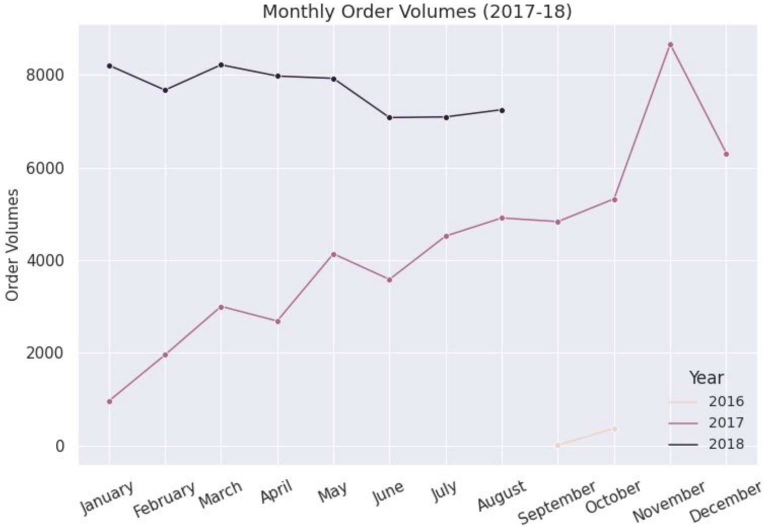


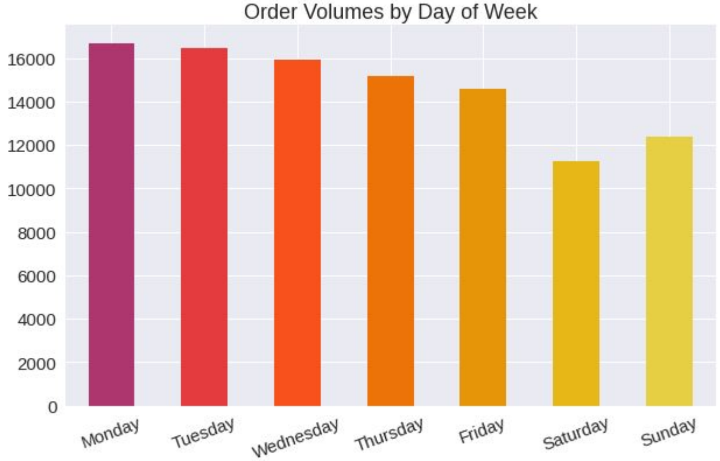
Upward trend with an extreme jump on **Black Friday** and other holidays

Top 3 product categories (by sales) have different seasonality and peak effect

Limited historical data and **extreme seasonality** makes it difficult to perform forecasting

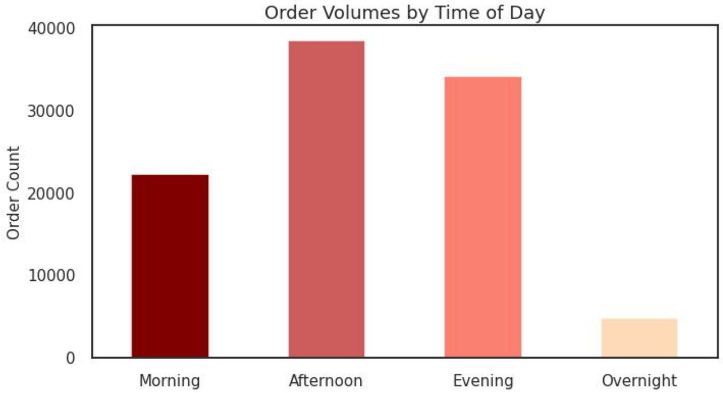
Sales Analysis: Daily and Hourly Seasonality



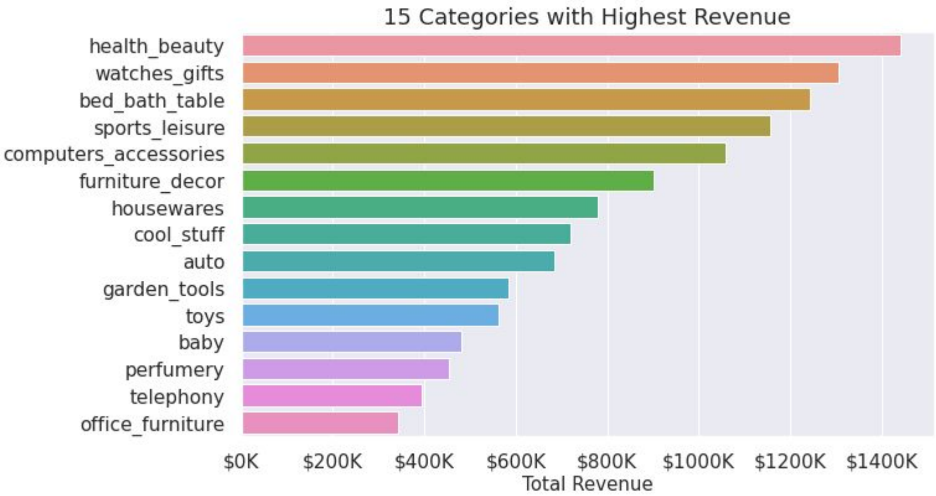


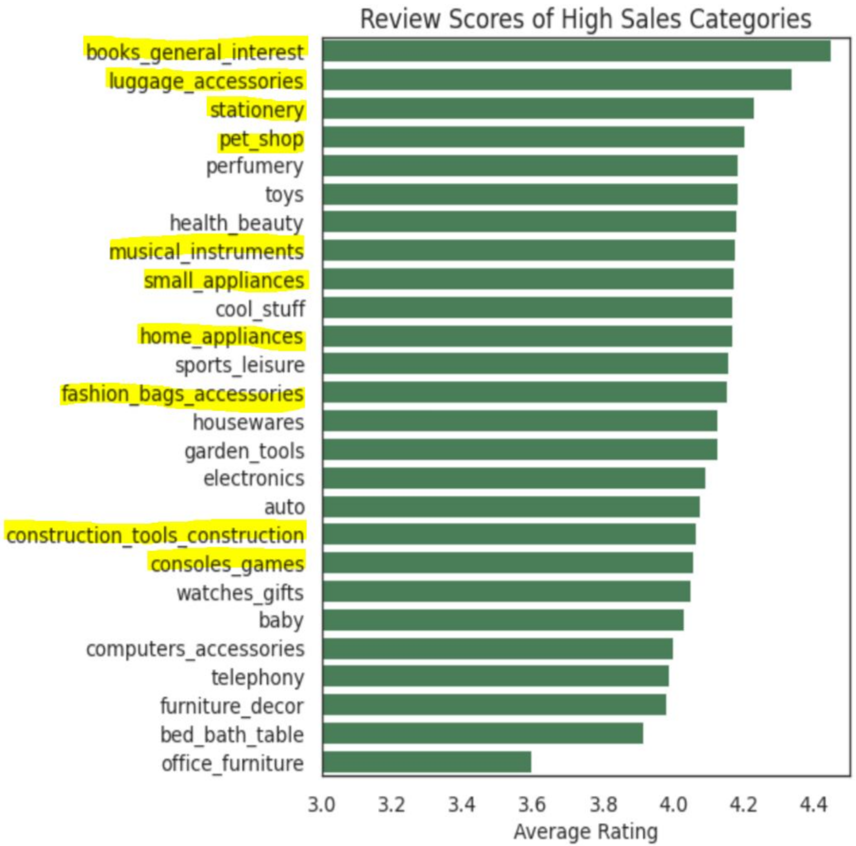
Monday blues is showing effect on purchases.

Afternoon boredom is making customers buy more items then any other time of the day.



Sales Analysis: Product Categories





Top product categories are sold mostly and have the highest turnover

Look into popular categories that have lower ratings *(office_furniture)*

Future: sales by region

Learning Object 3: Customer Segmentation

Customer Clustering using Unsupervised Machine learning

- ❑ We set out to **cluster customers based on similarities** in key features such as sales, delivery times, and review scores.

- ❑ The goal is to use these clusters to **provide a better customer experience** through enhanced product suggestions and pain points

Feature Collection

- **Orders:** Order count, Days since last order, Product count, Sales

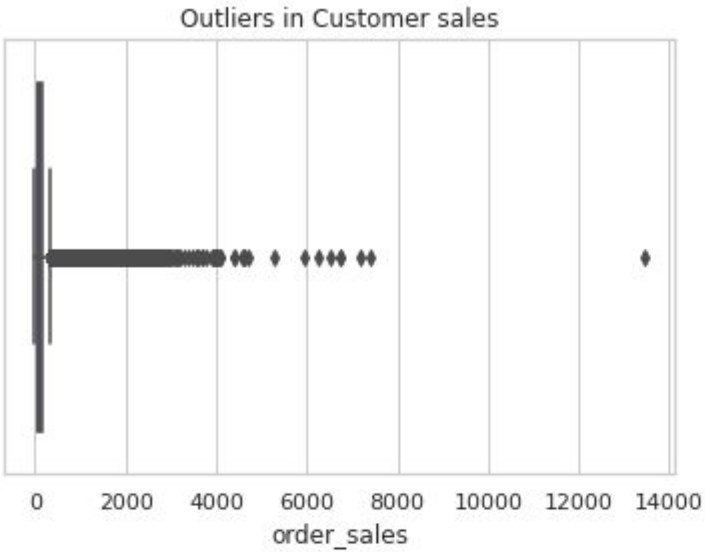
- **Delivery:** Early and Late order count

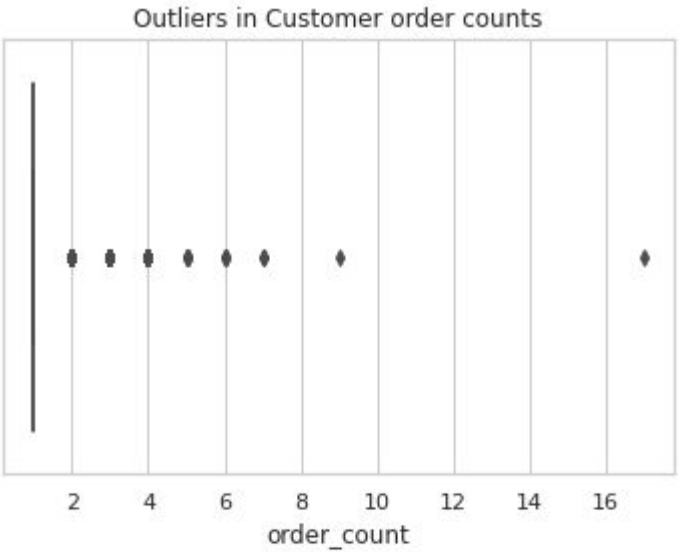
- **Payments:** Average payment installments

- **Reviews:** Review count and Average review score

Outlier Removal

We removed all customers with **more than $2000 in sales** or **more than 3 orders**





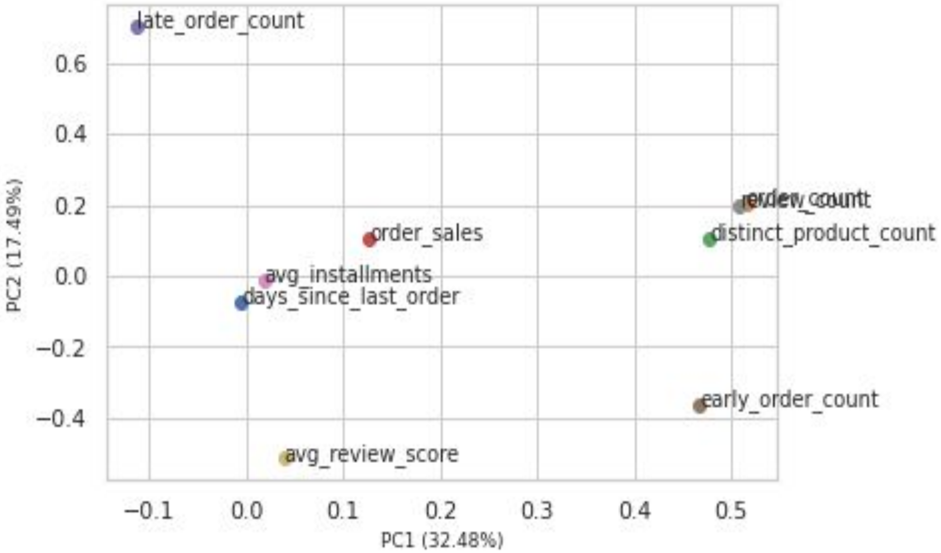
Data Preparation and PCA

- ❑ Scaled all features using a standard scaler

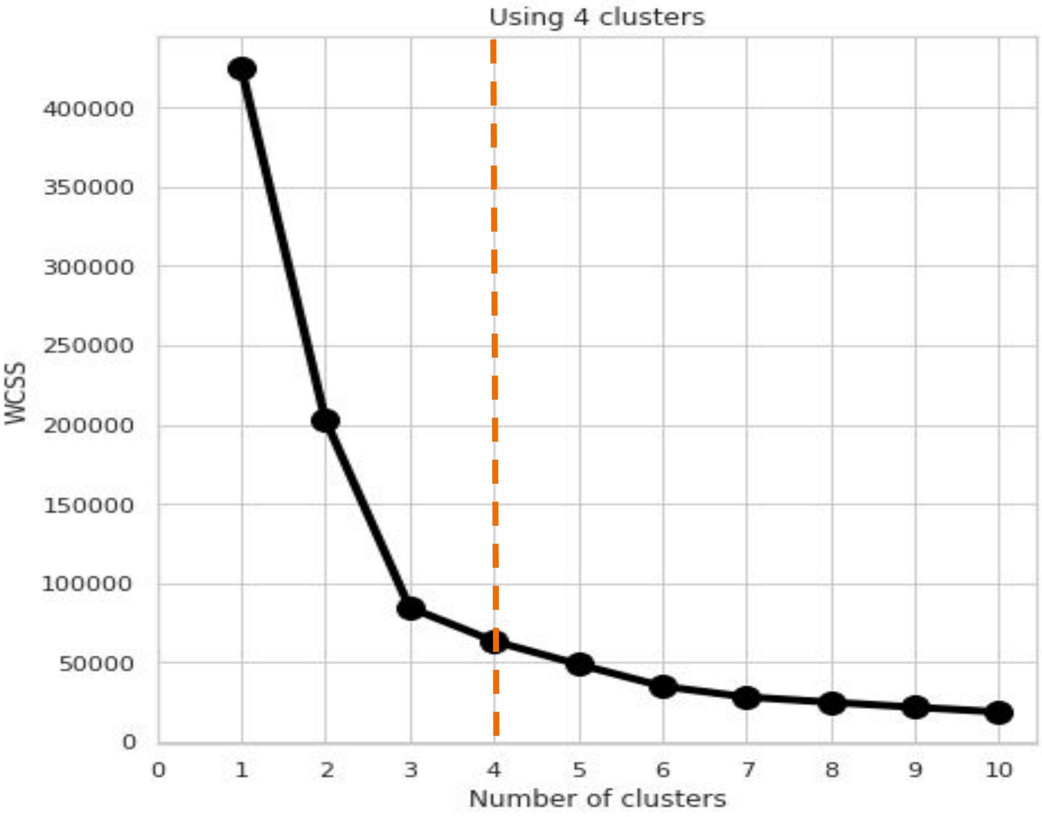
- ❑ Created 2 features using PCA

PC	Explained Variance
PC1	32.5%
PC2	17.5%

PC Loadings

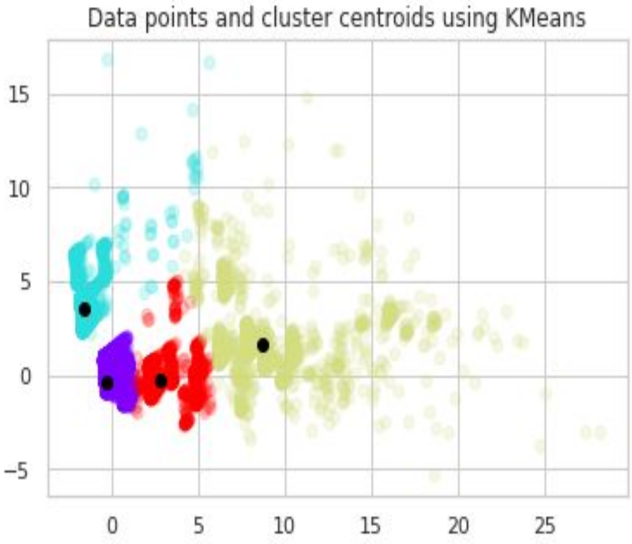


Using Within Cluster Sums of Squares to determine amount of clusters

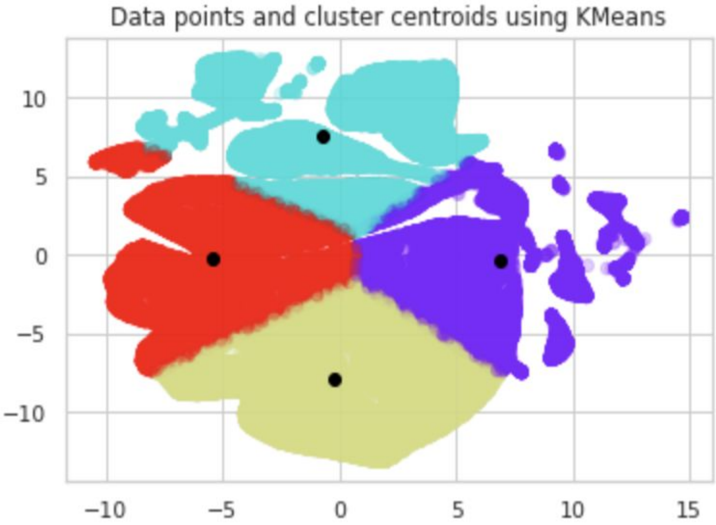


K-Means and PCA provides the best clustering results

K-Means and PCA

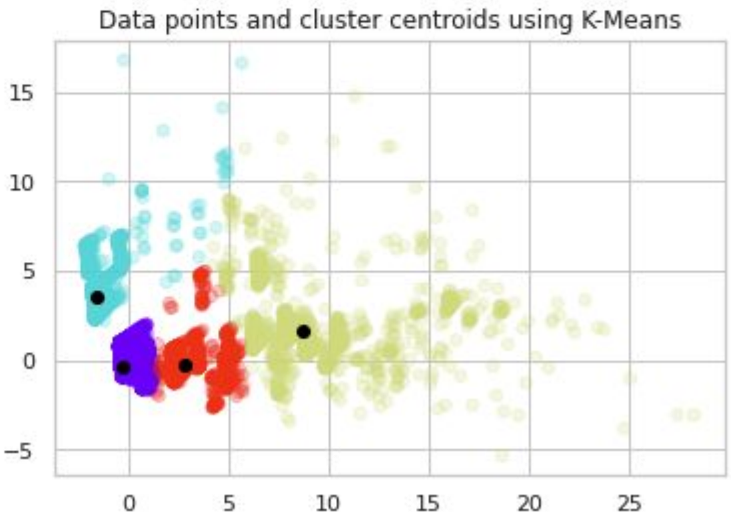


K-Means and t-sne

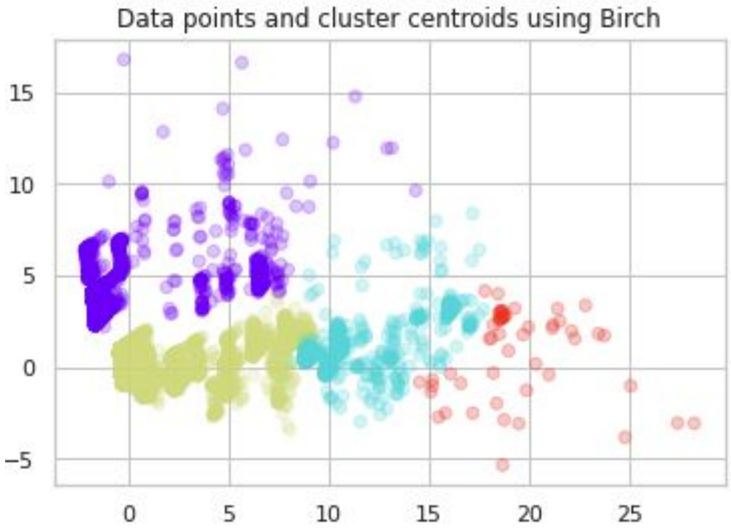


K-Means and PCA provides the best clustering results

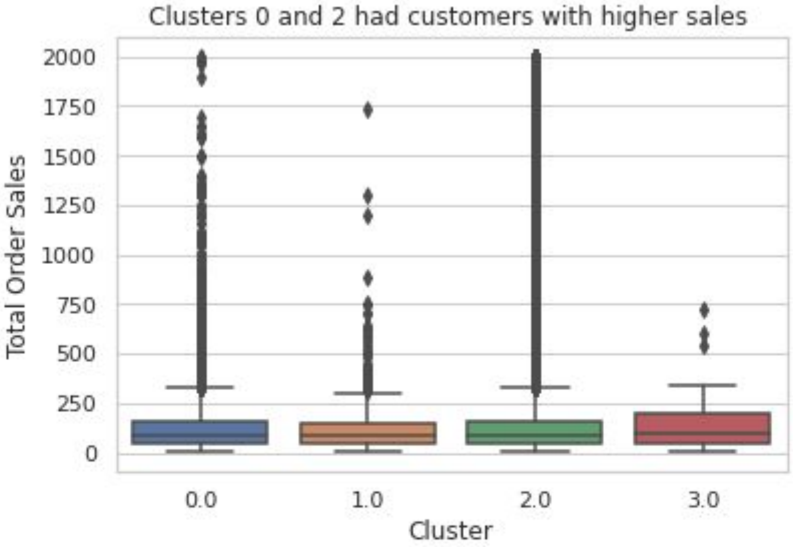
K-Means and PCA

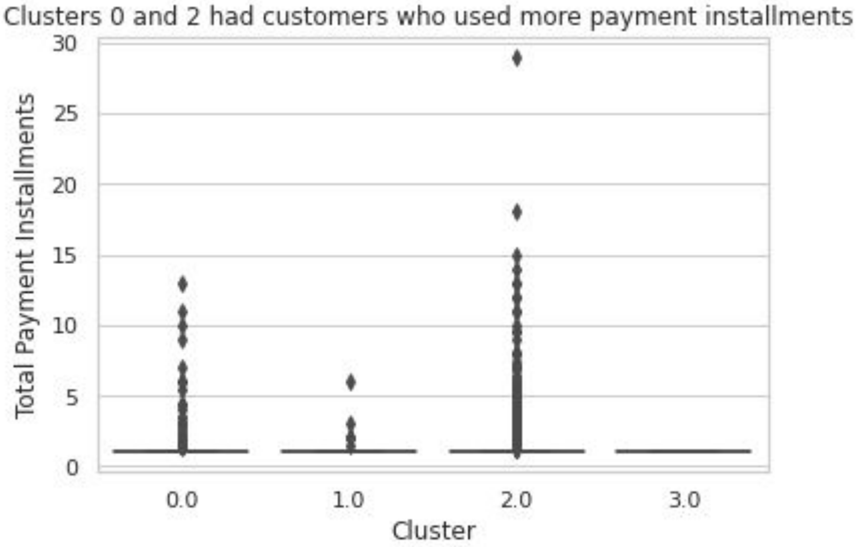


Birch and PCA

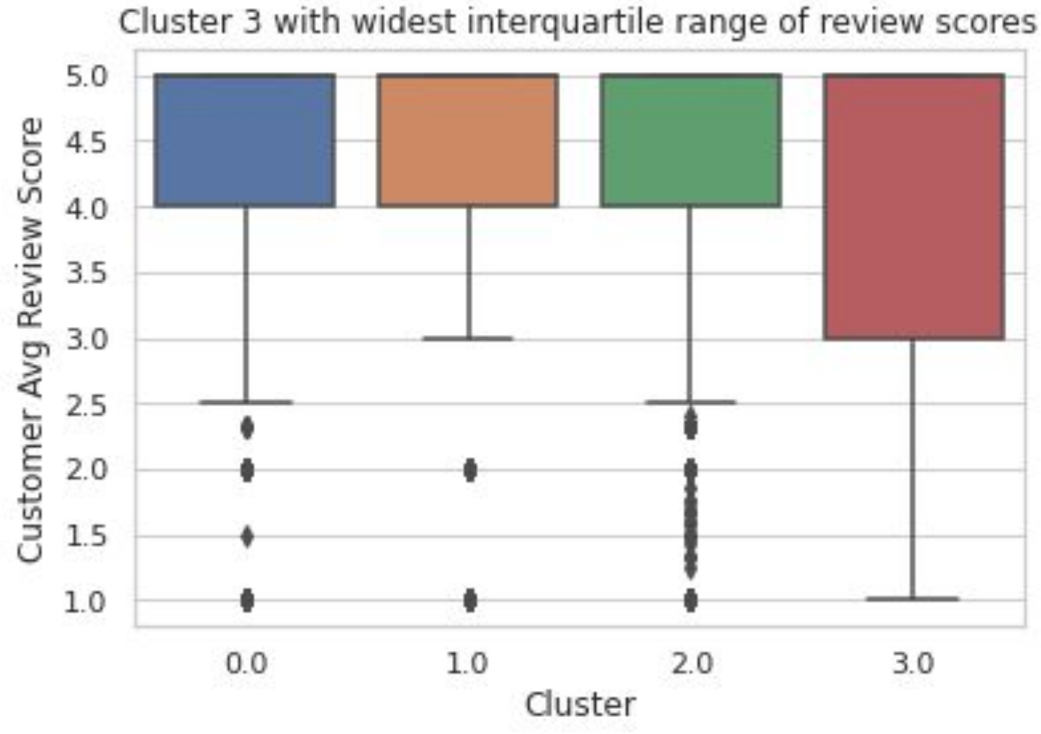


Customer Sales and Payment Installments By Cluster





Customer Average Reviews By Cluster



Learning Object 4: Review Sentiment Analysis

Sentiment Analysis on Text Data using NLP.

- We have a corpus of text data, where **each review text is a customer's message** for the purchased product.
- We transform the score from 1 to 5.
    - **Negative sentiment** = Score of 3 and Below
    - **Positive sentiment** = Score of 4 and Above
- We then **extract the most frequently occuring words** within each category
- Goal is to **predict a review's sentiment** given the textual message using Machine Learning Models

Data Cleaning and Preprocessing steps

We use **Regular Expressions and Stemming** to clean and preprocess the corpus.

- ❏ Keep words which consists of alphabets only and remove all others
- ❏ Convert alphabets to lowercase and **remove single alphabet words** like a, and I.
- ❏ **Remove stopwords** like the, from the textual review, and replace extra spaces with a single space.
- ❏ Use **Stemming to reduce related word forms** and derive a common base form (a root word) for them.
    - ❏ Ex. { organize, organizes, organized } => { organize }

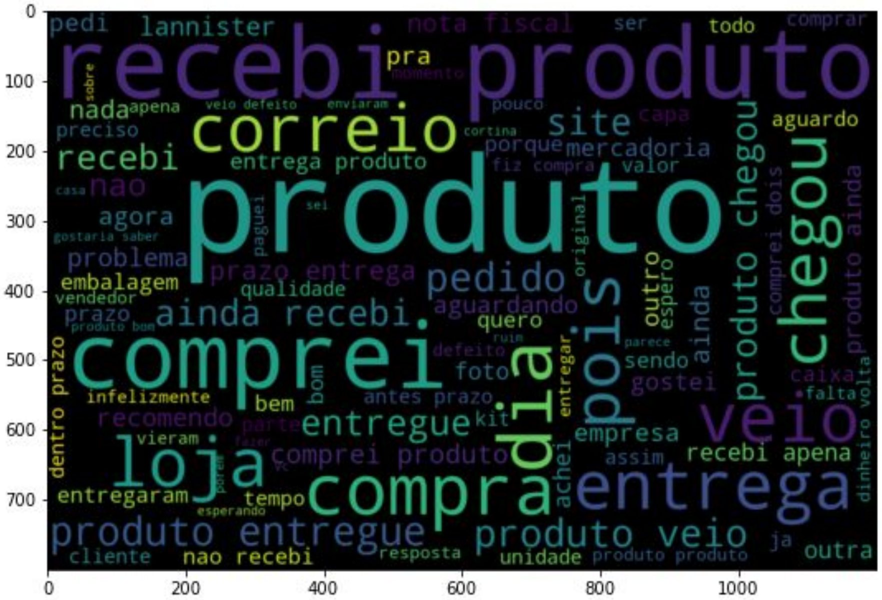
Data Visualization

❑ After cleaning, we plot the most frequently occuring words using **Word Clouds**, where words with high count are displayed in larger fonts.

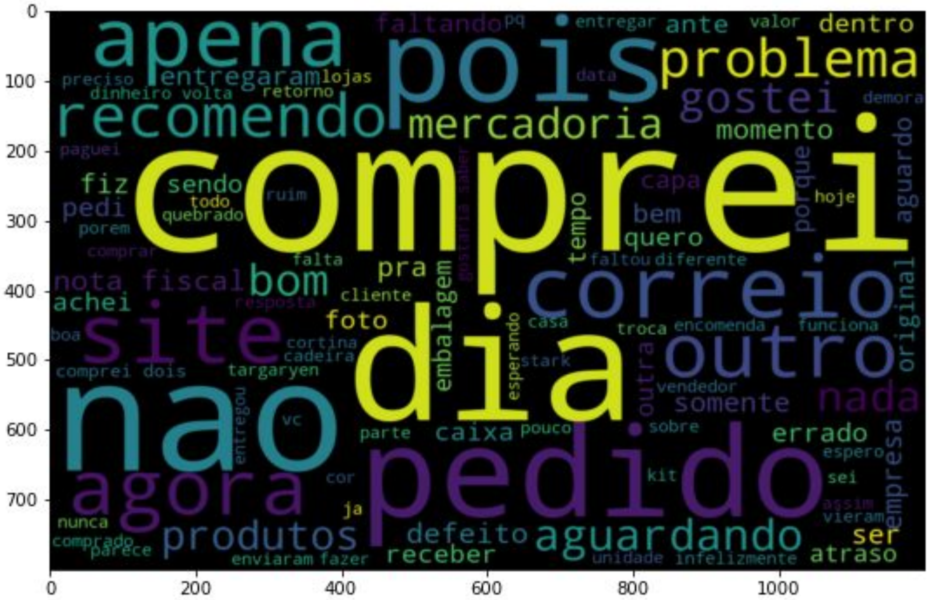
Note : There will be some overlapping frequently occuring words for both the categories, and hence they are not valuable as they won’t help in making predictions for review sentiment.

❑ Therefore, we plot a new **Word Cloud** that only contains **unique top common words** to its category.

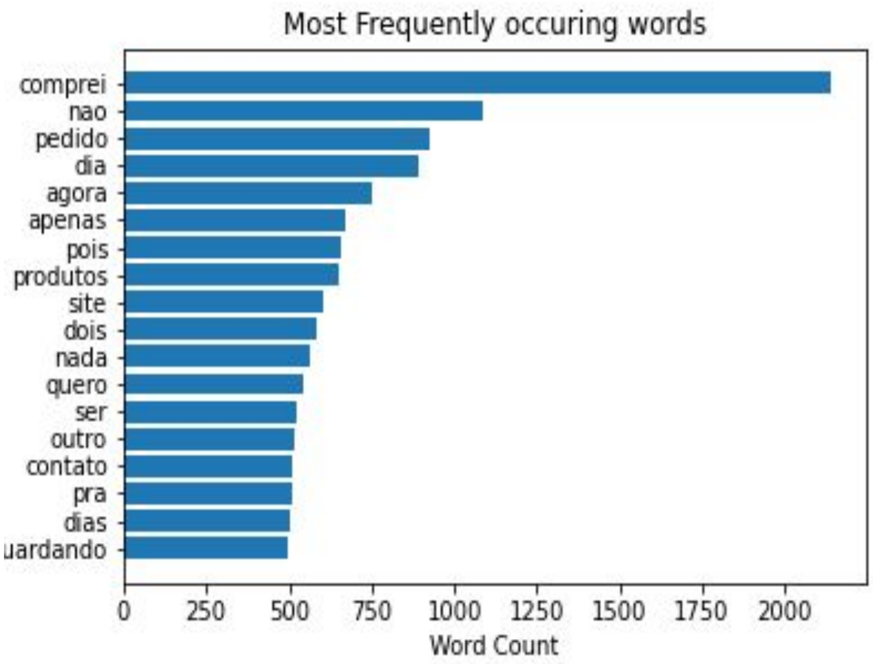
Word Cloud for **negative sentiment** **before** removing overlapping common words.



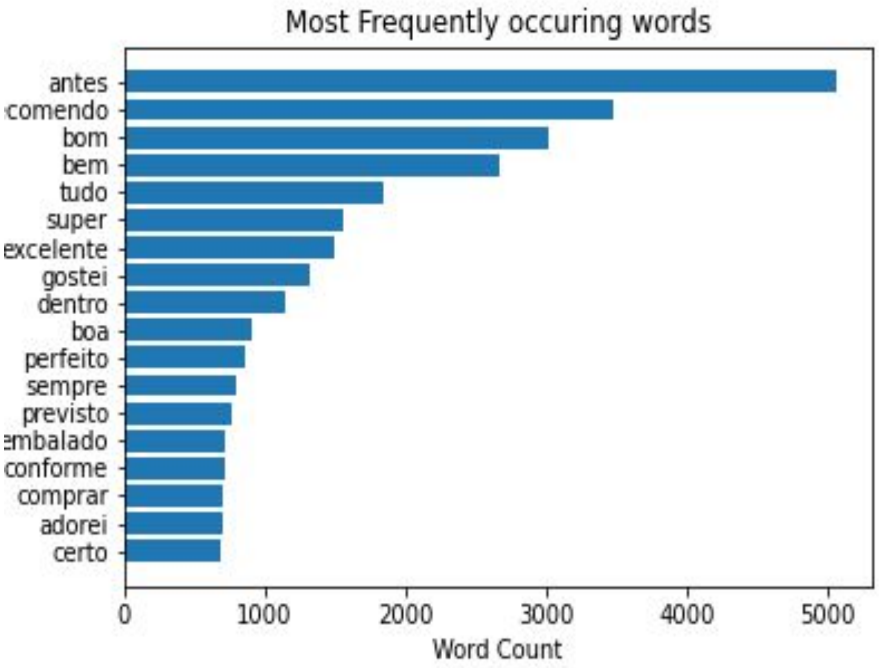
Word Cloud for **negative sentiment** **after** removing overlapping common words.



Most frequently occuring unique words for **Negative** sentiment.



Most frequently occurring unique words for **Positive** sentiment.



Machine Learning

- ❑ Using **Term Frequency-Inverse Document Frequency (TF-IDF)**, we first convert the textual data into numeric form.

- ❑ We use a TF-IDF vectorizer that **gives each text word a score**, calculated on the basis of **how frequently** it occurs in a document and in **how many** documents does it occur.

- ❑ We limit the maximum number of words used (vocabulary of the TF-IDF vectorizer) to 10,000 and hence, it will only score top 10,000 words and other words will be denoted as UNK or 0.

Training different ML models

We train 2 machine learning models and score them on the validation data:

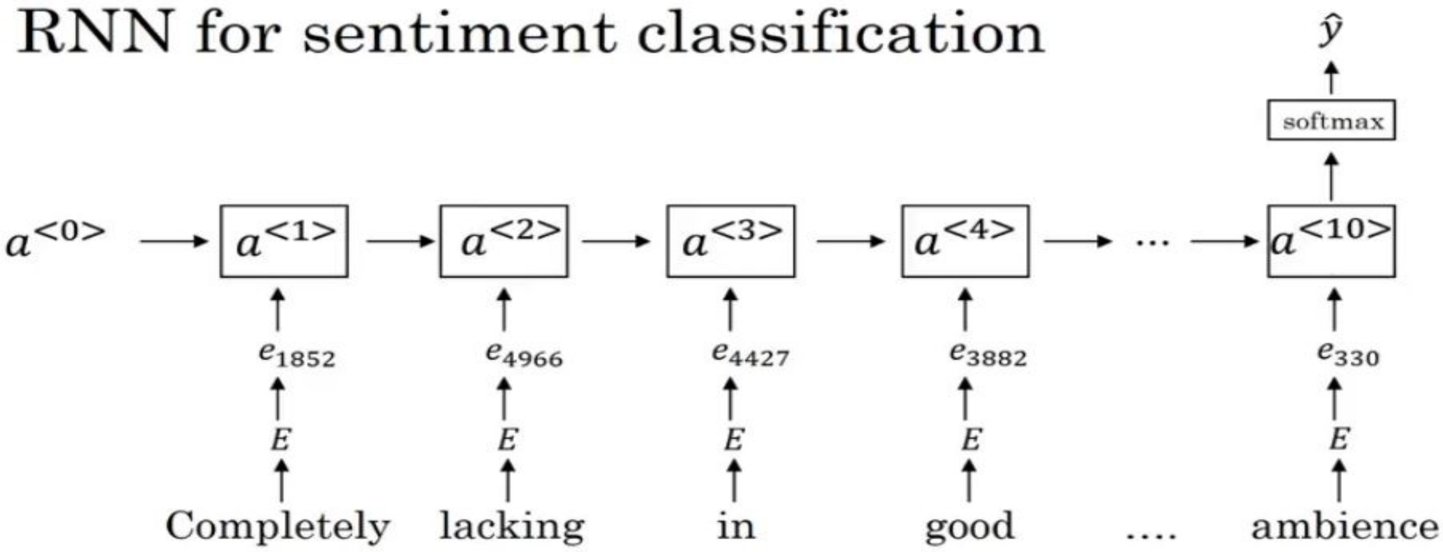
- **Logistic Regression** - **86.98%**
- **Random Forest Classification** - **76.65%**

Logistic regression outperforms the more complex Random Forest, as the later overfits on the training data with a high variance, train accuracy of **99.78%.**

Training a Recurrent Neural Network (RNN)

Recurrent Neural Networks inputs words it uses predictions from previous inputs to compute output for the next node/word.

Below is the architecture of a vanilla RNN.



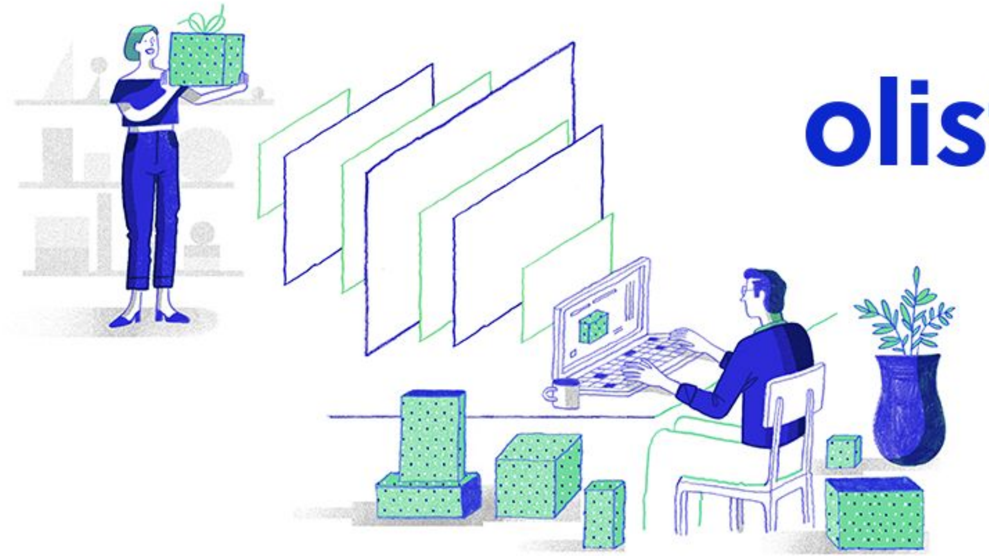
We construct a Recurrent Neural Network model as follows:

- First layer is a vanilla RNN with 128 nodes
- Next layer is a fully connected layer with 64 nodes with dropout of 20%(fo regularization).
- The last layer contains a single node with sigmoid activation.

The validation accuracy obtained on fitting a **RNN** is **88.9%,** only a little more than the accuracy of the Logistic Regressor.

Hence, the **accuracy** for the simplest classifier and an advanced Neural Network is the same, which is quite interesting.

Discussions & Future Work



olist

What we learned

1. **Delivery has a high impact on review score.** To ensure customer satisfaction Olist should focus on improving their logistics and supply chain.

2. Built a model to understand customer sentiment. This can be used by Olist to **quickly identify and address areas of poor customer experience** through classifying good and bad review text

Limitations

1. Due to the nature of the **Imbalanced Data**, we are unable to build trustworth forecasting.
   - ❑ Only 2 years of order data
   - ❑ High percentage of good reviews

     Many observations have only 1 review

2. With no customer labels and limited experience in the retail industry, we are unable to best classify and cluster customers

Future Work

1. With **access to more data**, possibly 5 years, we could better forecast things such as seasonality and customer churn rates.

2. **Work with retail experts** to better understand how to classify customers based on features in the data

Questions