BRAZILIAN E-COMMERCE

CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING

Fauziah Habibah

TABLE OF CONTENTS

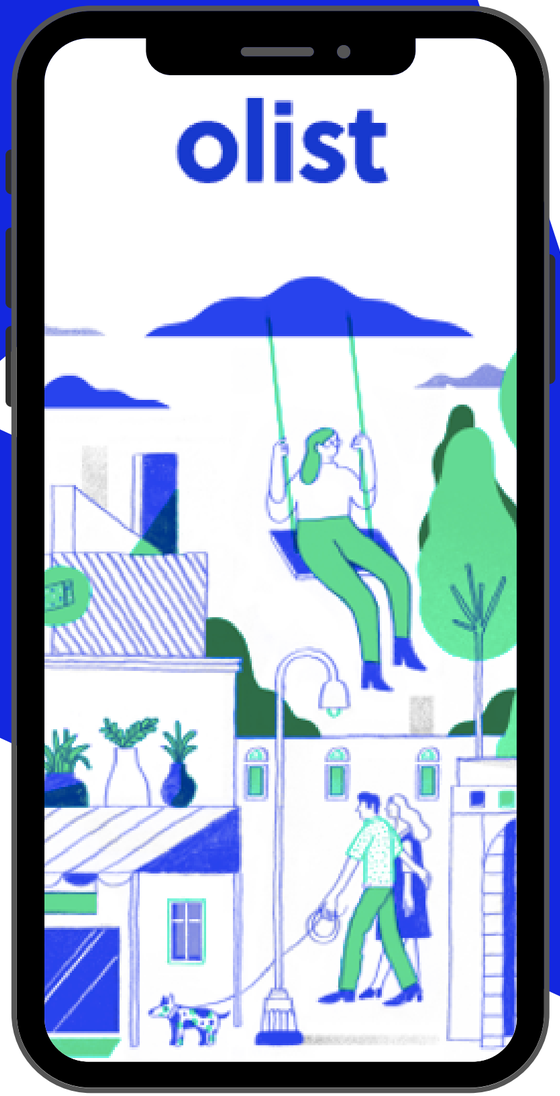
BACKGROUND

BACKGROUND

Revenue from the marketplace business model can come from paid features, advertisements, payment gateways, and partnerships. The point is that the more users or traffic on a marketplace, the more revenue you will get. One way to **increase users or traffic** on a marketplace is to do a **marketing campaign**, such as giving discounts, copywriting, etc.

Companies prepare a budget of 5% — 12% of the total revenue to carry out marketing campaigns. However, the problem is that the **marketing campaign is not right on target**, so the company suffers a loss. A marketing campaign can be measured by how many new users manage to get and how many old users retain to keep using the marketplace.

One way to solve the problem is to segment customers or commonly referred to as **Customer Segmentation**. Customer Segmentation helps marketplace owners to group customers with similar characteristics.



ABOUT THE COMPANY

Olist is a Brazilian e-commerce platform that connects small and medium-sized businesses to customers across Brazil. The platform operates as a marketplace, where merchants can list their products and services and customers can browse and purchase them online.

BUSINESS PROBLEM

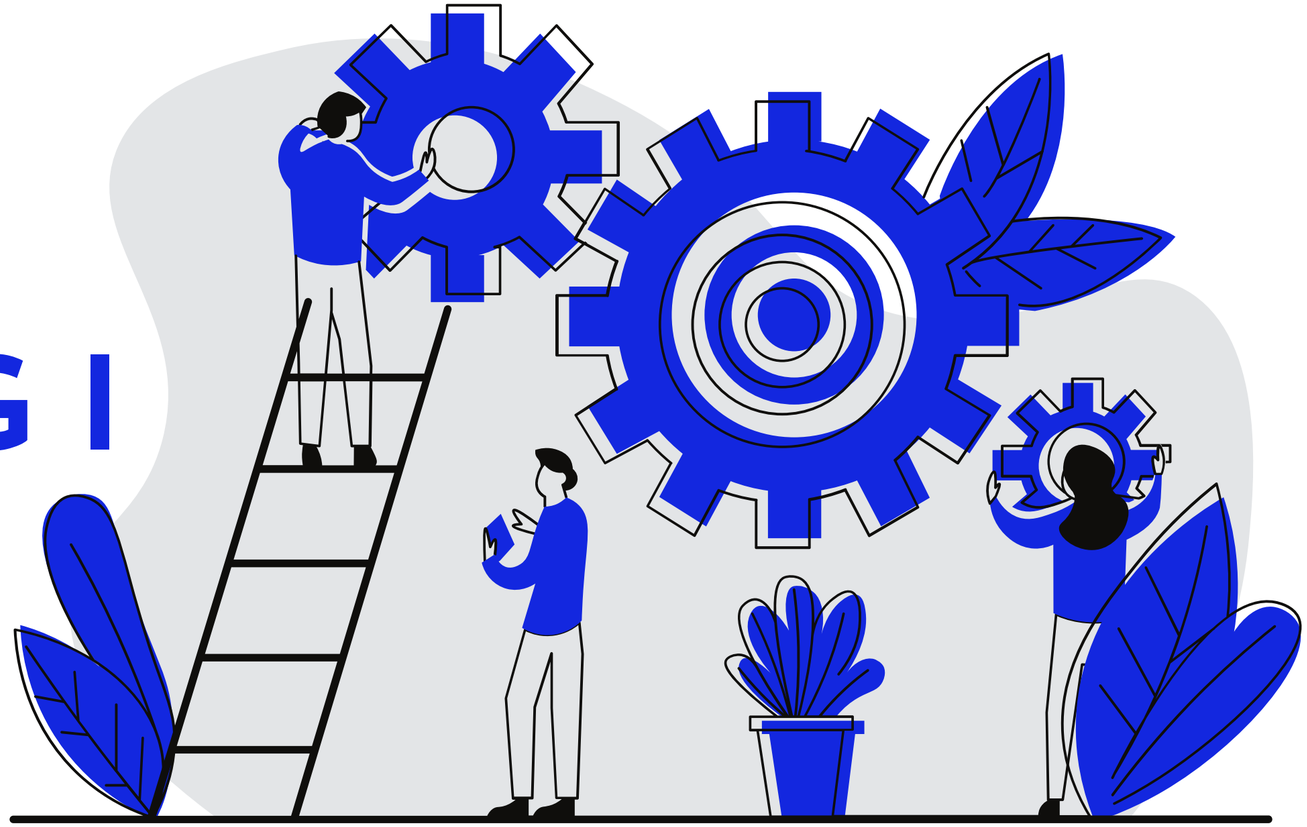
1. **How to segment the customers at Olist marketplace so we can divide customers based on their shopping behavior?**
2. **What kind of treatment for each cluster to increase customer retention rate?**

OBJECTIVES

To provide the right marketing campaign to the right segment in order to increase customer retention rate and engagement.

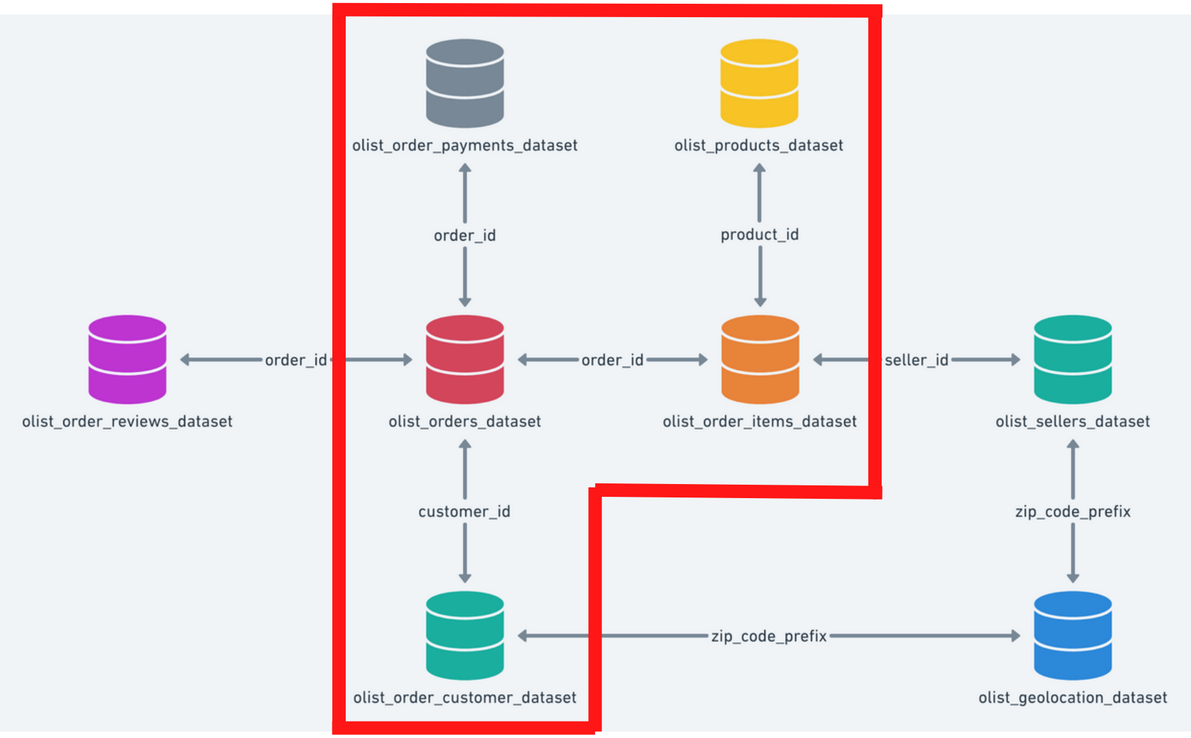
DATA PRE-PROCESSING I

For Data Exploration



ABOUT THE DATASET

Data Source: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data



JOIN INTO 1 TABLE

```
Int64Index: 113367 entries, 0 to 113366
Data columns (total 31 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   order_id                       113367 non-null  object
 1   order_item_id                  113367 non-null  int64
 2   product_id                     113367 non-null  object
 3   seller_id                      113367 non-null  object
 4   shipping_limit_date            113367 non-null  object
 5   price                          113367 non-null  float64
 6   freight_value                  113367 non-null  float64
 7   payment_sequential             113367 non-null  int64
 8   payment_type                   113367 non-null  object
 9   payment_installments           113367 non-null  int64
 10  payment_value                  113367 non-null  float64
 11  product_category_name          113367 non-null  object
 12  product_name_lenght            113367 non-null  float64
 13  product_description_lenght     113367 non-null  float64
 14  product_photos_qty             113367 non-null  float64
 15  product_weight_g               113367 non-null  float64
 16  product_length_cm              113367 non-null  float64
 17  product_height_cm              113367 non-null  float64
 18  product_width_cm               113367 non-null  float64
 19  customer_id                    113367 non-null  object
 20  order_status                   113367 non-null  object
 21  order_purchase_timestamp       113367 non-null  object
 22  order_approved_at              113367 non-null  object
 23  order_delivered_carrier_date   113367 non-null  object
 24  order_delivered_customer_date  113367 non-null  object
 25  order_estimated_delivery_date  113367 non-null  object
 26  product_category_name_english  113367 non-null  object
 27  customer_unique_id             113367 non-null  object
 28  customer_zip_code_prefix       113367 non-null  int64
 29  customer_city                  113367 non-null  object
 30  customer_state                 113367 non-null  object
dtypes: float64(10), int64(4), object(17)
memory usage: 27.7+ MB
```

This is Olist data from 2016 - 2018

DATA CLEANING

MISSING VALUES HANDLING

7,44% Missing values in Products Table

	null_counts	%
product_category_name	610	1.85
product_name_lenght	610	1.85
product_description_lenght	610	1.85
product_photos_qty	610	1.85
product_weight_g	2	0.01
product_length_cm	2	0.01
product_height_cm	2	0.01
product_width_cm	2	0.01
product_id	0	0.00

DROP

4,93% Missing values in Products Table

	null_counts	%
order_delivered_customer_date	2965	2.98
order_delivered_carrier_date	1783	1.79
order_approved_at	160	0.16
order_id	0	0.00
customer_id	0	0.00
order_status	0	0.00
order_purchase_timestamp	0	0.00
order_estimated_delivery_date	0	0.00

FEATURE ENGINEERING

CHANGING WRONG DATA TYPE

```
order_purchase_timestamp       113367 non-null  object
order_approved_at              113367 non-null  object
order_delivered_carrier_date   113367 non-null  object
order_delivered_customer_date  113367 non-null  object
order_estimated_delivery_date  113367 non-null  object
```

```
order_purchase_timestamp       113367 non-null  datetime64[ns]
order_approved_at              113367 non-null  datetime64[ns]
order_delivered_carrier_date   113367 non-null  datetime64[ns]
order_delivered_customer_date  113367 non-null  datetime64[ns]
order_estimated_delivery_date  113367 non-null  datetime64[ns]
```

Change Object Data Type to Date Time

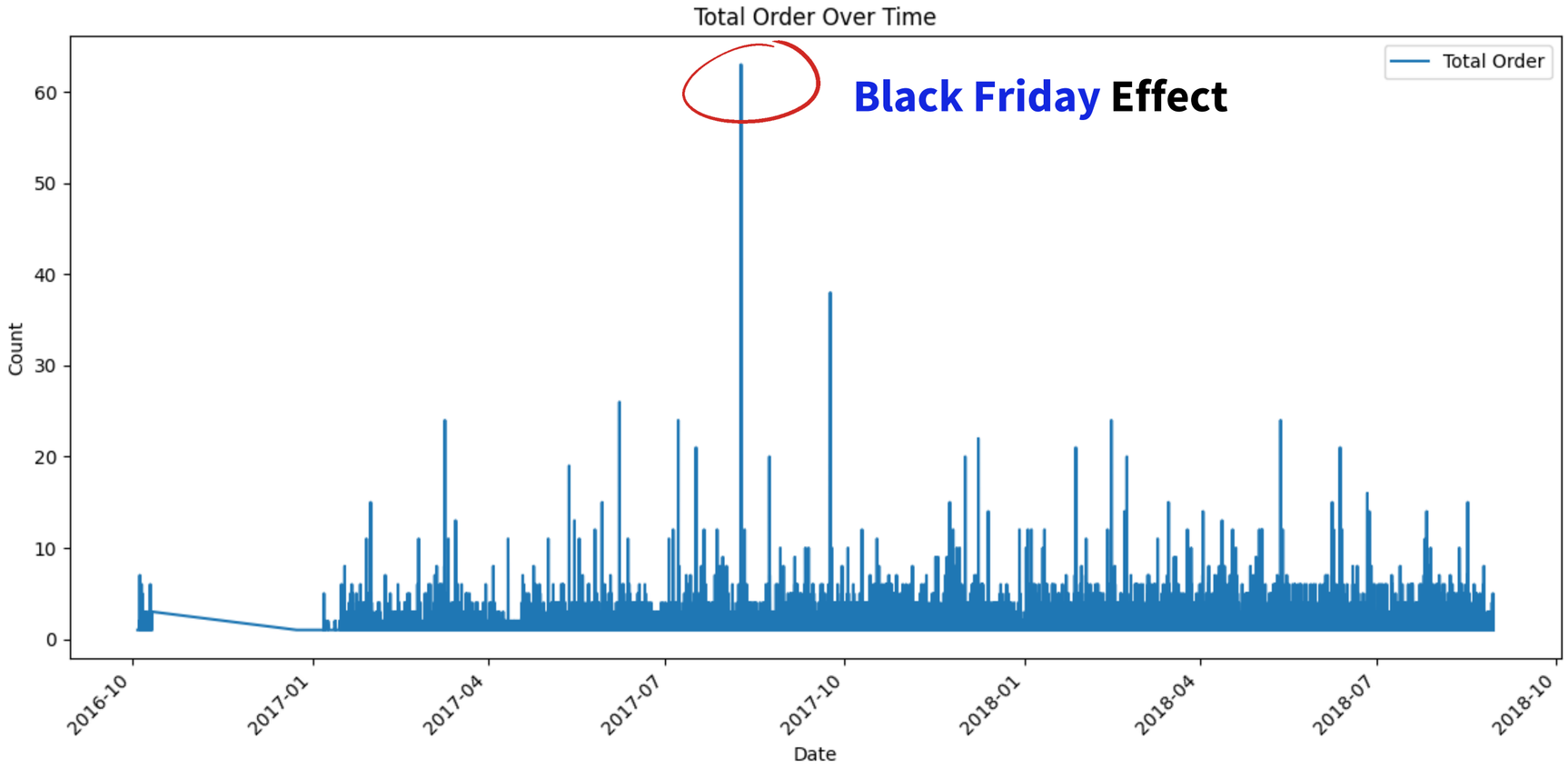
FOR DEEP DIVE ANALYSIS:

- **Add Revenue Column**
- **Extract Order Date (DD-MM-YYYY HH:MM) Into Day, Month, Year, Hour**

EXPLORATORY DATA ANALYSIS

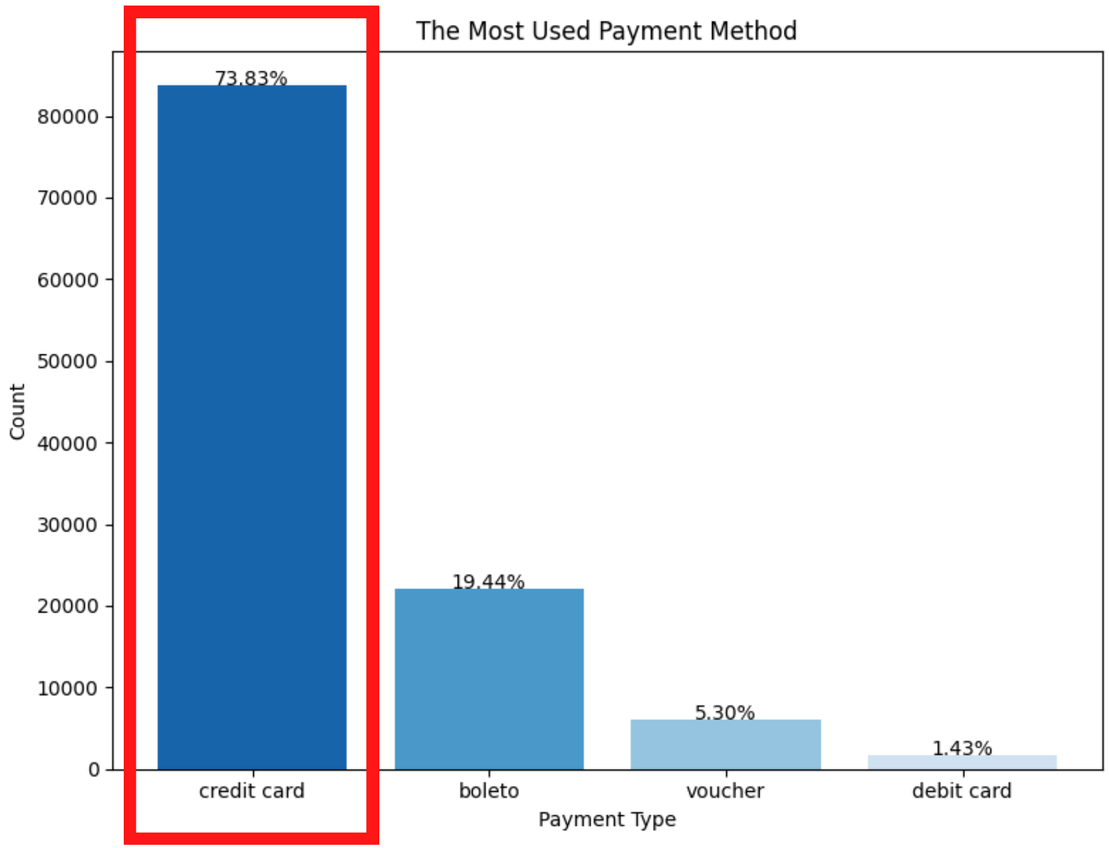
- How is total order over time?
- What is the most used payment method by customers?
- On what day and time do customers tend to make transactions?
- What products category generate most revenue?
- What is the most popular products category?
- How is customer retention rate?

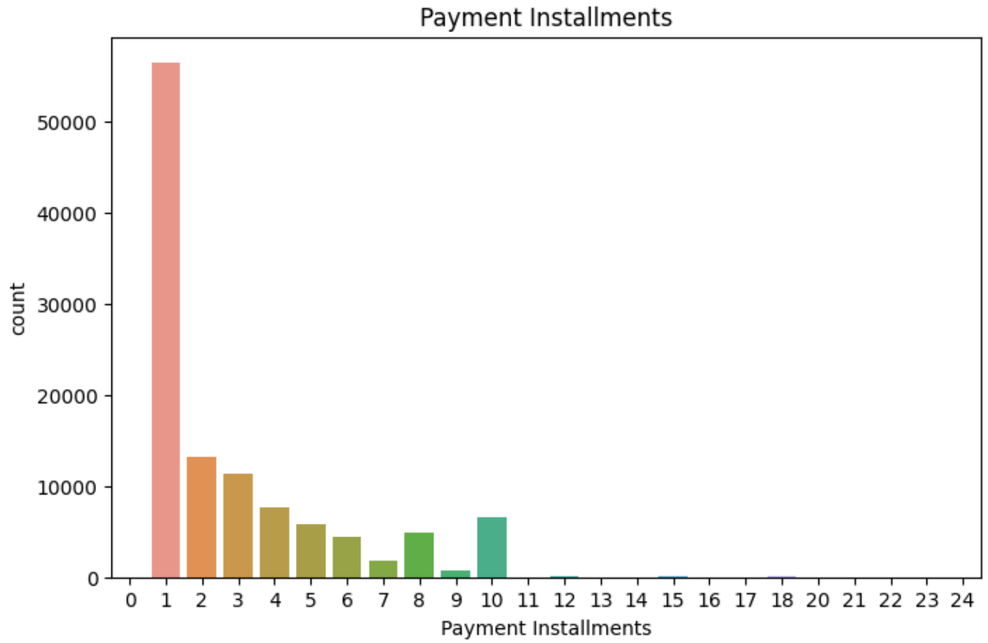
HOW IS TOTAL ORDER OVER TIME?



Highest total order occured on Black Friday

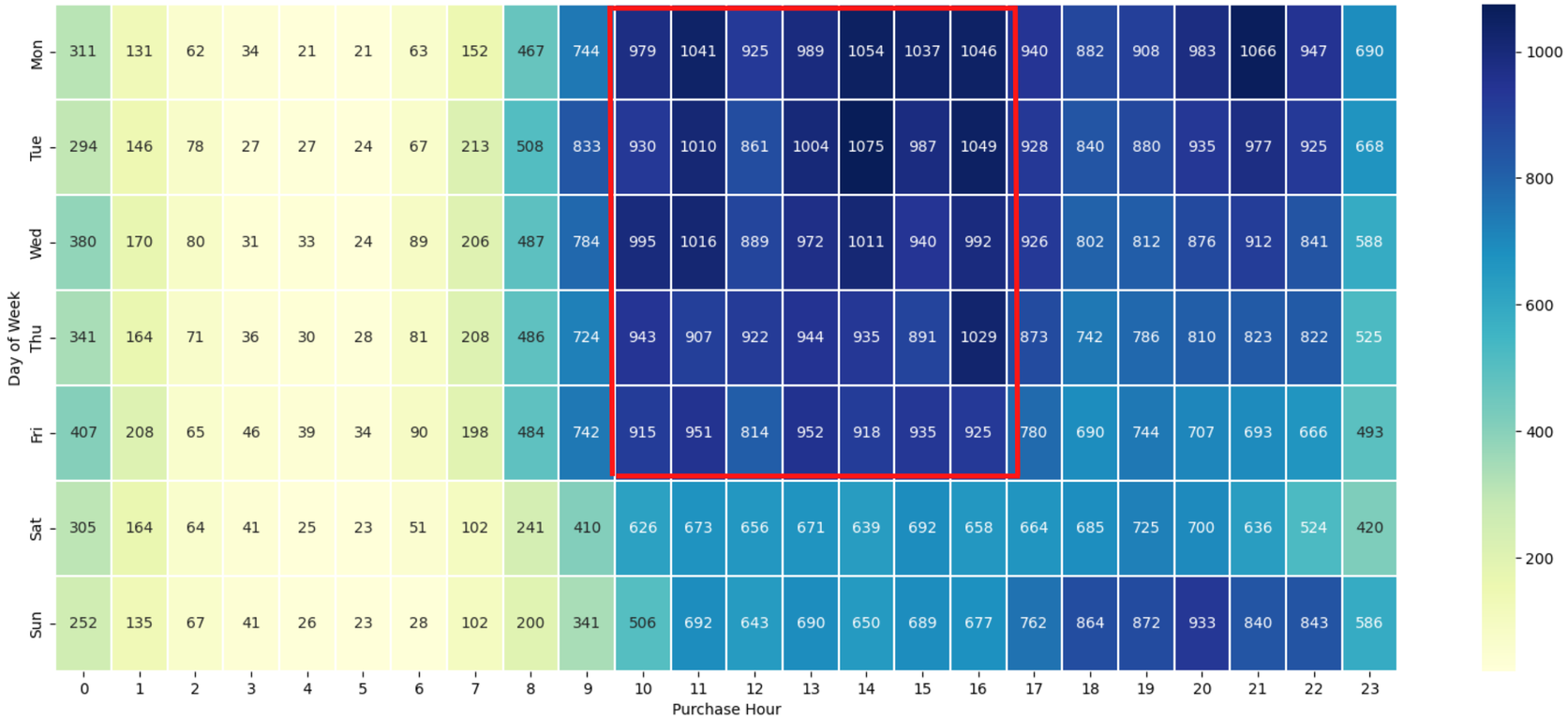
73,83% CUSTOMERS USE CREDIT CARD





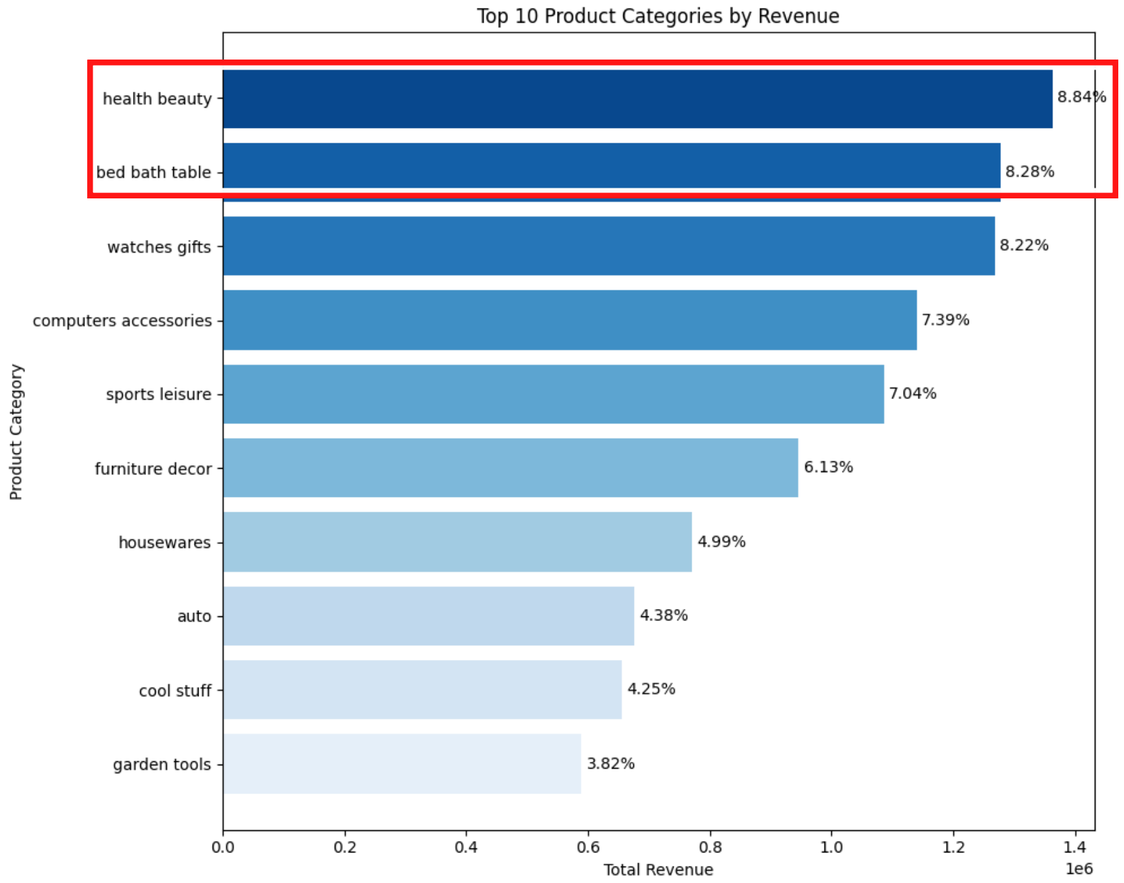
Olist can cooperate with credit card company to create campaign to increase more sales & new customers

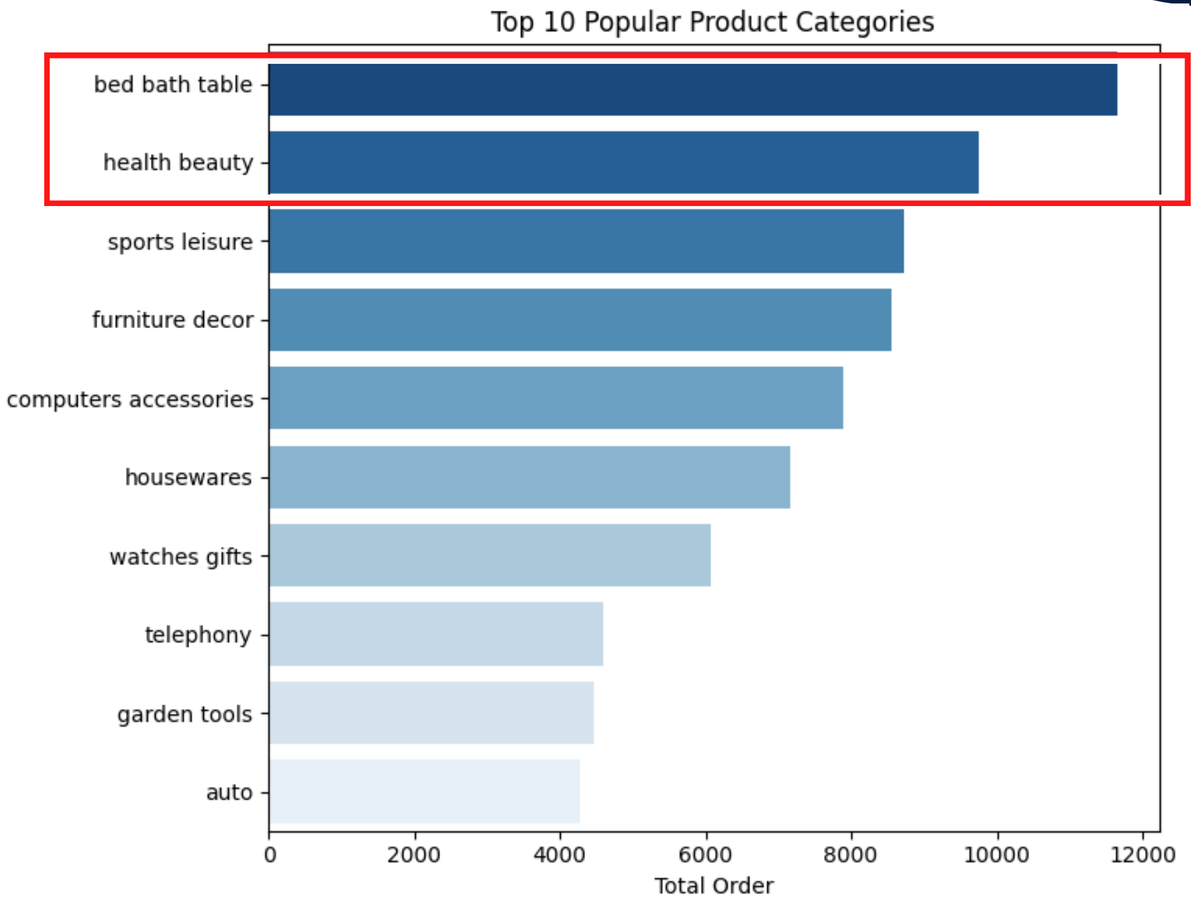
TRANSACTIONS ARE HIGH ON WEEKDAYS



Highest transactions occurred on weekdays especially from 10:00 - 16:00

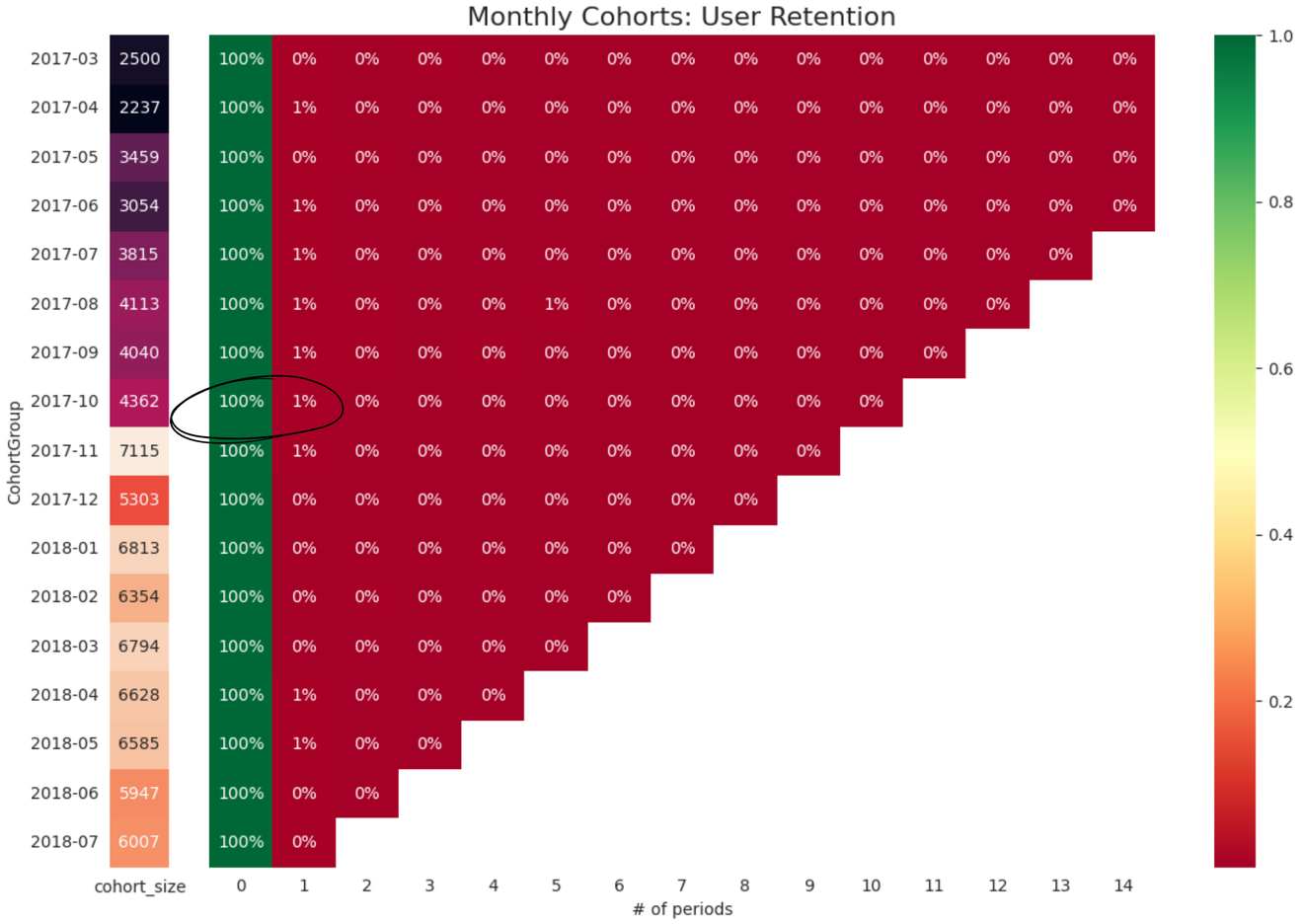
HEALTH BEAUTY & BED BATH TABLE GENERATES MOST REVENUE & BEING THE MOST POPULAR CATEGORY



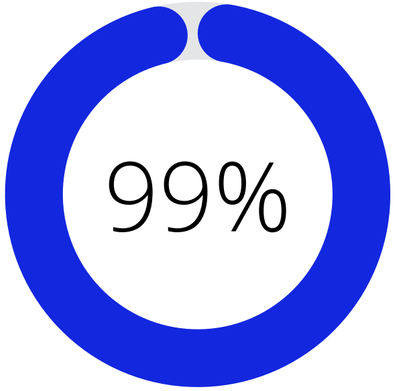


Olist can create campaign focusing on health beauty & bed bath table brands to generate more income/sales

HOW IS CUSTOMER RETENTION RATE?



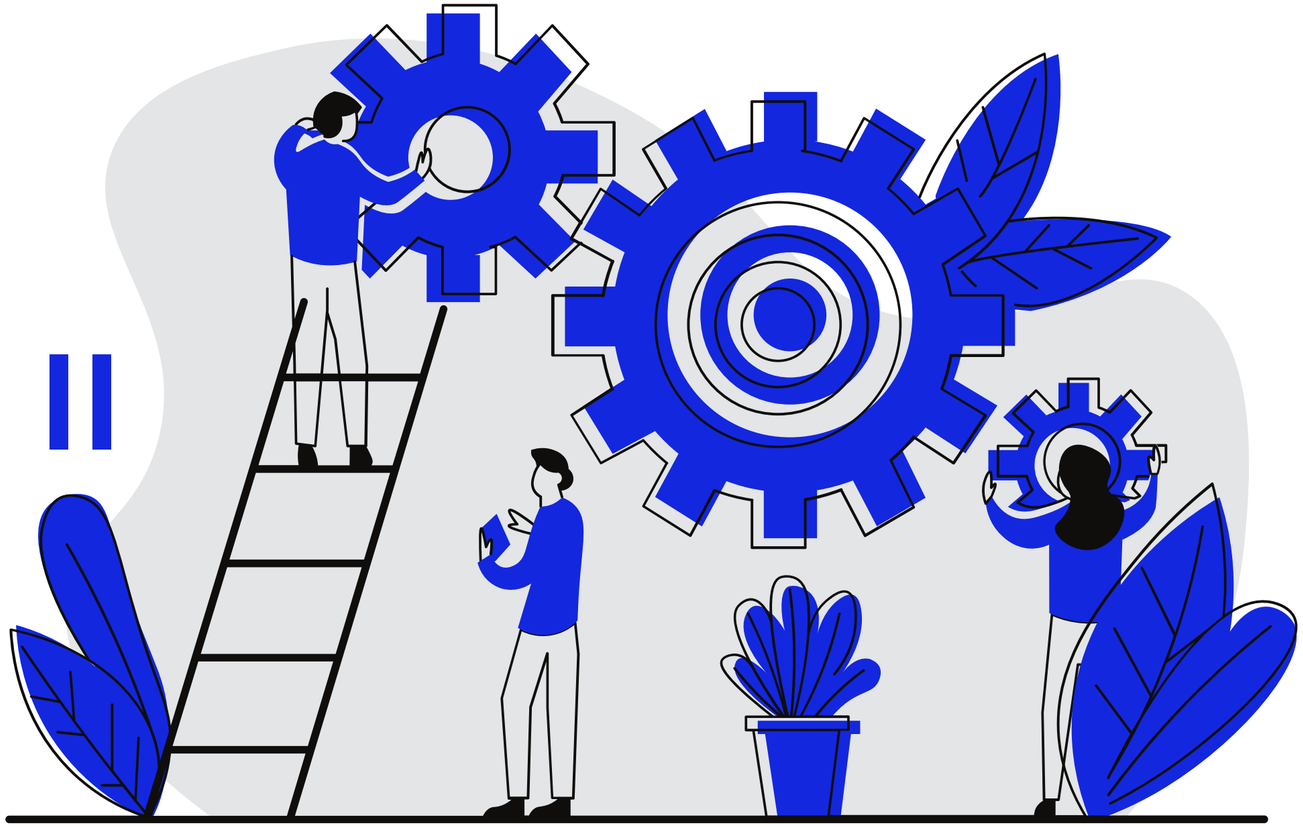
In 2017 & 2018,



of new customers did not make a transaction within the next month

DATA PRE-PROCESSING II

For Data Modelling



FEATURE SELECTION

Using RFM Model

The RFM explanation is as follows:

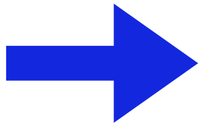
- **R (RECENCY) : Number of days since the last purchase**
- **F (FREQUENCY) : Number of transactions made over a given period**
- **M (MONETERY) : Amount spent over a given period of time**

So the features taken are: Order Purchase Date (R), Order Unique ID (F), Payment Value (M)

FEATURE ENGINEERING

- **Change Order Purchase Date into Customer Last Transactions (in Days) for Recency Column**

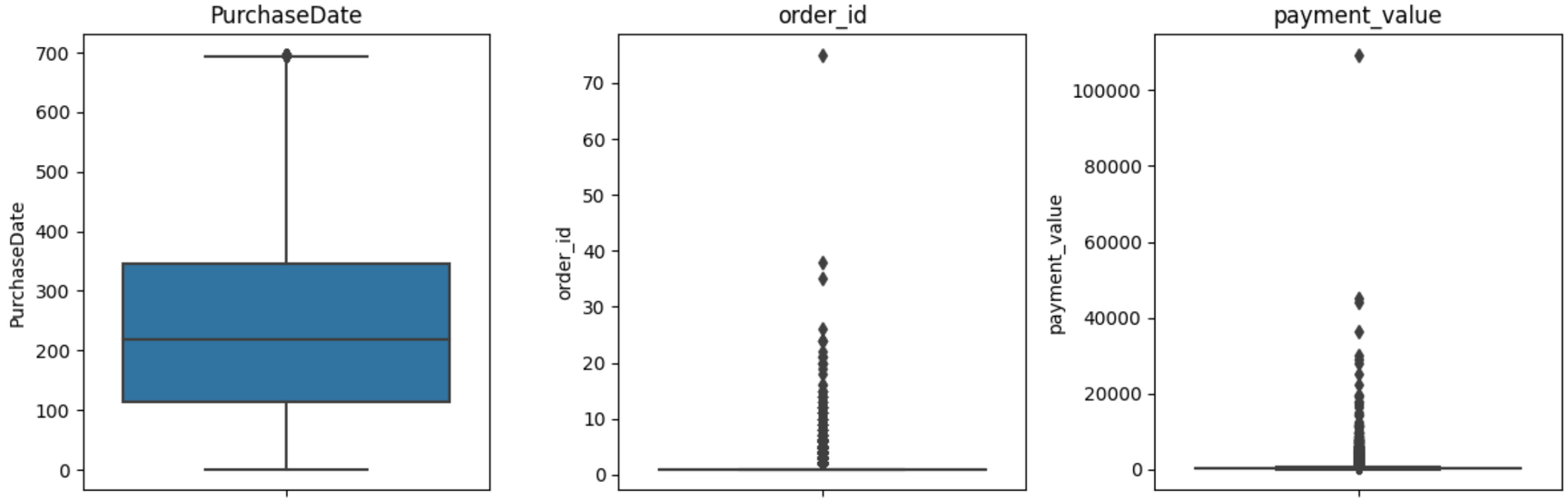
PurchaseDate
2017-09-13
2017-06-28
2018-05-18
2017-08-01
2017-08-10



PurchaseDate
112
115
538
322
289

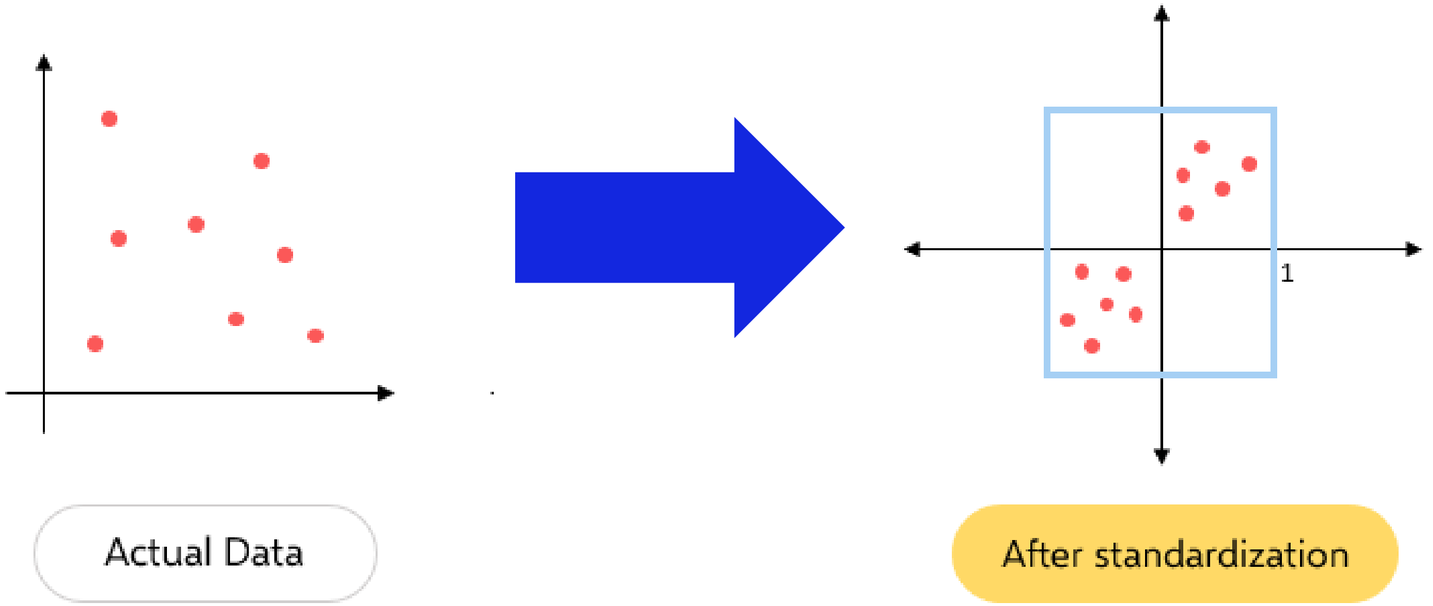
MISSING VALUES, DUPLICATED DATA, & OUTLIER CHECK

- **0 Missing Values**
- **0 Duplicated Data**



The data in reasonable amount, so I decided not dropping the outlier

FEATURE SCALLING



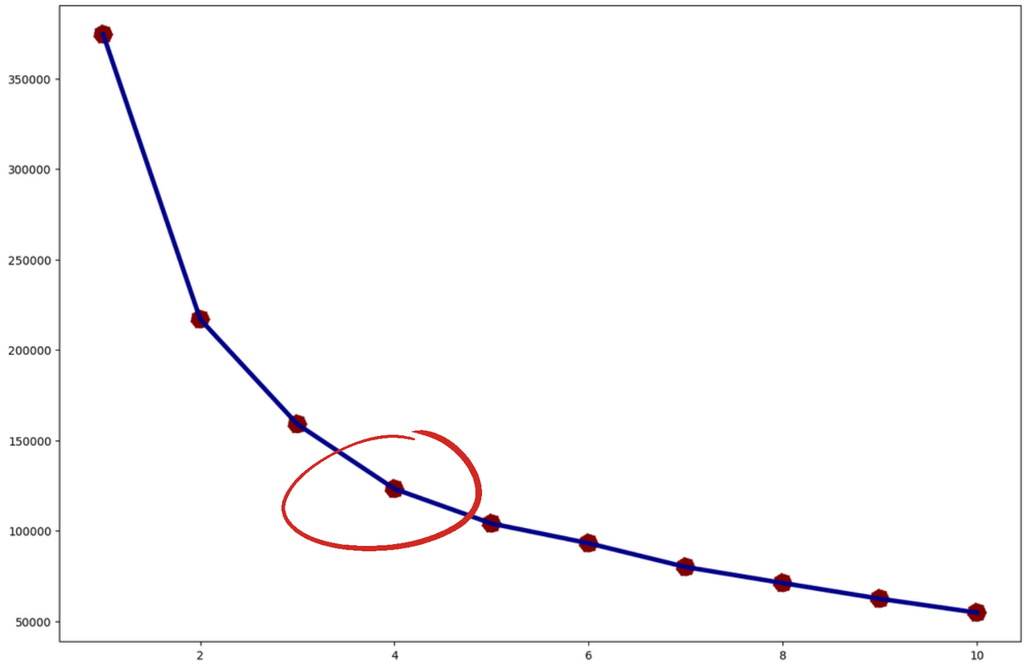
- **This technique used to re-scale features value so that the data used does not have large deviations**
- **Since clustering algorithms including k-means use distance-based measurements, this technique help improve the k-means performance**

DATA MODELLING

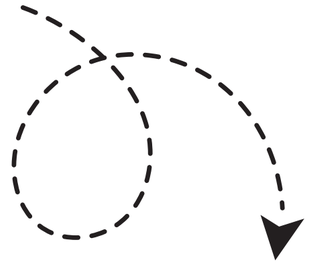
Using K-Means Clustering

DETERMINING THE NUMBER OF CLUSTERS & FIT THE MODEL

Elbow Method



Fit The Model

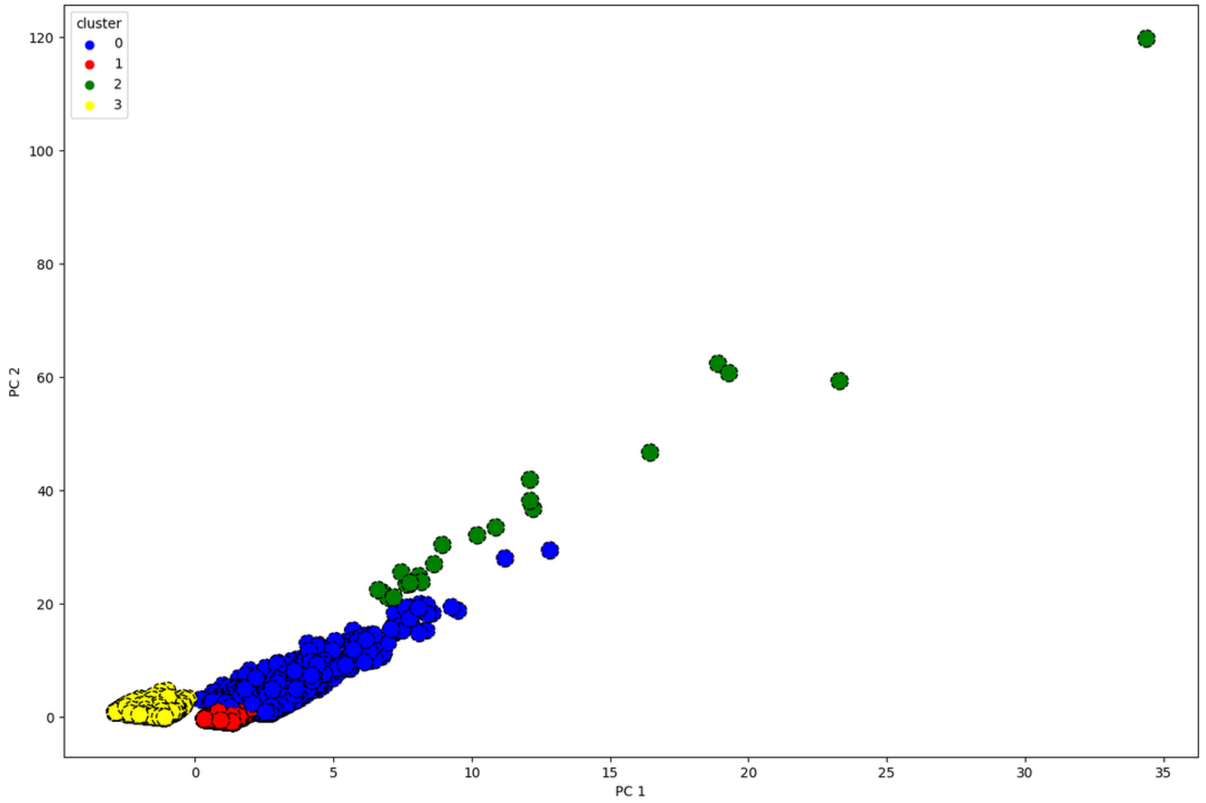


```
45] from sklearn.cluster import KMeans
    kmeans = KMeans(n_clusters=4, random_state=0)
    kmeans.fit(dfs.values)
```

```
KMeans
KMeans(n_clusters=4, random_state=0)
```

The best K is 4, means there will be 4 clusters for customer segmentation

K-MEANS CLUSTERING DISTRIBUTION



From the results of this customer clustering and visualized with a scatterplot as shown below.

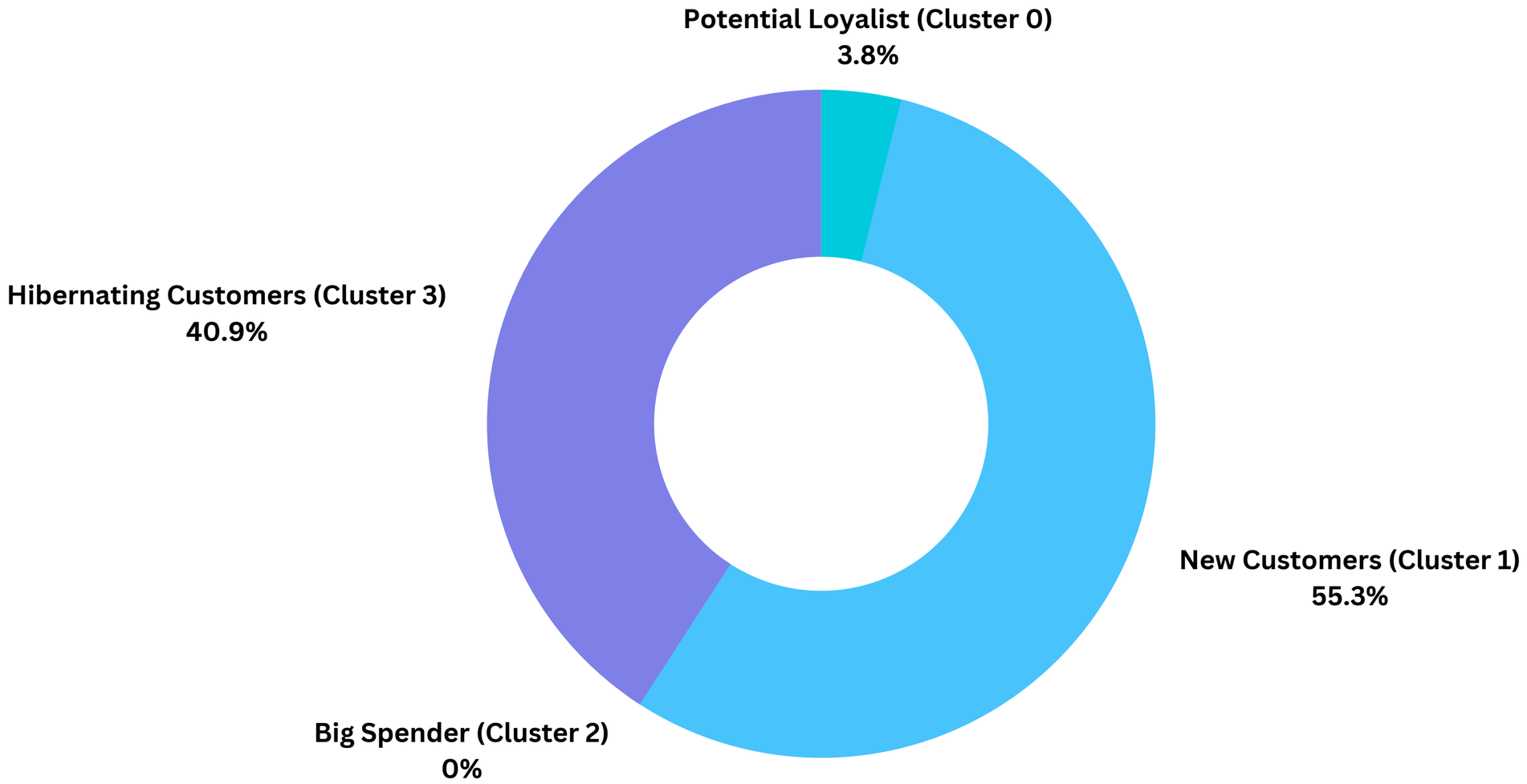
This diagram shows the distribution of customer data which is divided into clusters according to the K-Means Clustering algorithm.

CUSTOMERS SEGMENTATION BASED ON SCORE

	PurchaseDate	order_id	payment_value	
	median	median	median	count
cluster				
0	211.0	3.0	640.485	3542
1	129.0	1.0	109.420	50880
2	231.0	10.0	19342.260	21
3	374.0	1.0	106.970	37613

- **Cluster 0: Medium Recency, Medium Frequency, High Monetary -> *Potential Loyalist***
- **Cluster 1: High Recency, Low Frequency, Medium Monetary -> *New Customers***
- **Cluster 2: Medium Recency, High Frequency, High Monetary -> *Big Spender***
- **Cluster 3: Low Recency, Low Frequency, Medium Monetary -> *Hibernating customers***

CUSTOMER SEGMENTATION
PERCENTAGE



BUSINESS RECOMMENDATIONS

RECOMMENDATIONS

MARKETING APPROACH BASED ON PRIORITY

New Customers 55,3%

- **Provide onboarding support, and brand awareness campaigns to increase their visits**
- **Mobile App Engagement: utilize push notifications to inform customers about new arrivals, discounts, or abandoned cart reminders**
- **Offer them discounts for an additional product to see whether they are the kinds of the customer to whom you can upsell**
- **Send them satisfaction surveys concerning their recent order**

Hibernating 40,9%

- **Re-Engagement Campaign: Give frequent product updates, newsletter on email, & offering exclusive/limited-time discounts**
- **Reactivation Surveys: Send surveys to understand the reasons behind their inactivity. Gather feedback about their experience and reasons for not returning. Use this information to address any issue**

RECOMMENDATIONS

MARKETING APPROACH BASED ON PRIORITY

- Motivate them to increase the number of items in their cart by showing them cross-selling recommendations
- Offer them loyalty programs
- Use gamification rewards to increaser engagement (for example get reward/voucher after completing certain order)

Potentital Loyalist 3,8%

- Offer them premium products, membership or loyalty programs
- Surprise Upgrades and Gifts (Send gifts to customer in their birthday)
- Create a sense of community by inviting them to join a VIP club or a social media group where they can interact with the company and other big spender customers
- Don't use discount pricing to generate sales

Big Spender 0,02%

REFERENCES

The following are references for working on this project:

https://clevertap.com/blog/rfm-analysis/

https://documentation.bloomreach.com/engagement/docs/rfm-segmentation-business-use

https://www.putler.com/rfm-segmentation/

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data

https://becominghuman.ai/what-does-feature-scaling-mean-when-to-normalize-data-and-when-to-standardize-data-c3de654405ed

https://ariqmuh.medium.com/customer-segmentation-for-e-commerce-e5ea4c2d630a

THANK YOU

LINKEDIN

https://www.linkedin.com/in/fauziah-habibah/