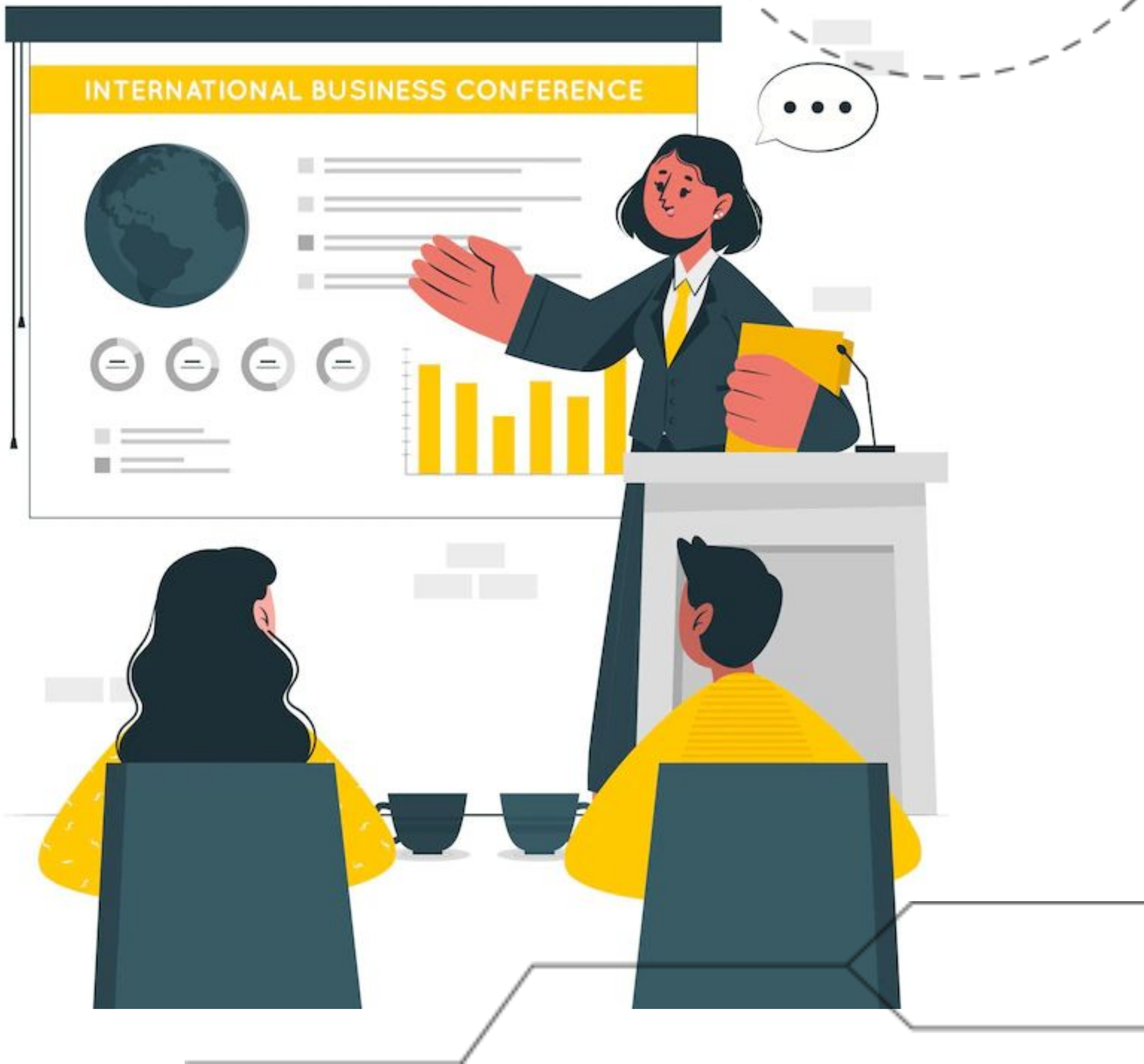
# E-Commerce Customer Segmentation

Gilang Putra Bahana

# Outline

**Topics Covered:**

- Business Problem
- Data Source
- Data Overview/Insights
- Clustering Analysis
- Recommendation



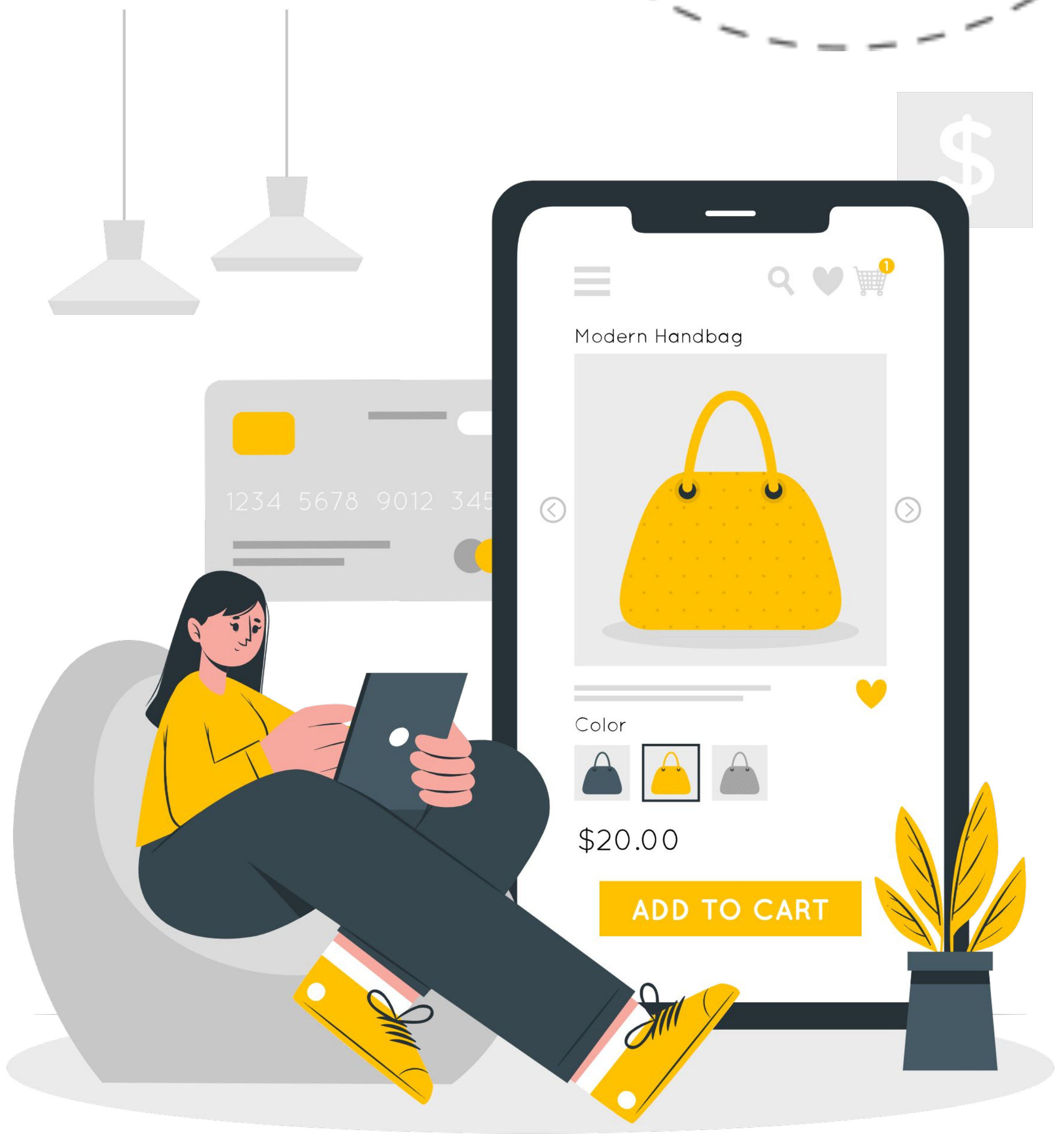
# Business Problem

## Introduction & Objective

- An e-commerce startup launched its website when the covid-19 hits, and it makes the company get its fortune.

- The CEO, Gustavo, believes that **the company has to improve the marketing strategy**, because they are not using targeted marketing and it hurts their marketing budget, as only a fraction of their customer back to their website.

- The CEO ask our help as a data analyst, **on how to increase the marketing conversion rate by doing more targeted marketing using customer segmentation.**



# Data Source

The data is provided in 3 (three) google sheet **customer dataset**, **order dataset**, and **order payment dataset**, which consist of customers that have already bought some item from 2016 to 2018. The data can be seen through the entity relationship database below:

**order_payments_dataset**

| | | |
|---|---|---|
| FK1 | order_id | nvarchar(128) |
| | payment_sequential | tinyint |
| | payment_type | nvarchar(25) |
| | payment_installments | tinyint |
| | payment_value | float(64) |

**customers_dataset**

| | | |
|---|---|---|
| PK | customer_id | nvarchar(128) |
| | customer_unique_id | nvarchar(128) |
| | customer_zip_code_prefix | int |
| | customer_city | nvarchar(25) |
| | customer_state | nvarchar(25) |

**orders_dataset**

| | | |
|---|---|---|
| PK | order_id | nvarchar(128) |
| FK1 | customer_id | nvarchar(128) |
| | order_status | nvarchar(25) |
| | order_purchase_timestamp | datetime |
| | order_approved_at | datetime |
| | order_delivered_carrier_date | datetime |
| | order_delivered_customer_date | datetime |
| | order_estimated_delivery_date | datetime |

# Data Source

## Data Dictionary

**customers_dataset**

- **customer_id:** customer unique identifier (PK)
- **customer_unique_id:** similar with customer id
- **customer_zip_code_prefix:** zip code
- **customer_city:** city where customer live
- **customer_state:** state where customer live

**orders_dataset**

- **order_id:** order unique identifier (PK)
- **customer_id:** customer unique identifier (FK)
- **order_status:** status of customer's order
- **order_purchase_timestamp:** date and time when the customer purchase goods
- **order_approved_at:** date and time when the e-commerce approved order
- **order_delivered_carrier_date:** date and time when the goods with courier
- **order_delivered_customer_date:** date and time when the goods arrived to customer
- **order_estimated_delivery_date:** estimated delivery date

**order_payments_dataset**

- **order_id:** customer unique identifier (PK)
- **payment_sequential:** the way customer pay on their goods (sequential or once, start from 1 to 26.
- **payment_type:** credit card, boleto, debit card, voucher
- **payment_installments:** the number of installments, start from 0 to 24 months
- **payment_value:** the money spent to buy goods

# Data Source

## Data Frame Preview

| | customer_id | customer_city | customer_state | order_id | order_status | order_purchase_timestamp | payment_sequential | payment_type | payment_installments | payment_value |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 06b8999e2fba1a1fbc88172c00ba8bc7 | franca | SP | 00e7ee1b050b8499577073aeb2a297a1 | delivered | 2017-05-16 15:05:35 | 1.0 | credit_card | 2.0 | 146.87 |
| 1 | 18955e83d337fd6b2def6b18a428ac77 | sao bernardo do campo | SP | 29150127e6685892b6eab3eec79f59c7 | delivered | 2018-01-12 20:48:24 | 1.0 | credit_card | 8.0 | 335.48 |
| 2 | 4e7b3e00288586ebd08712fdd0374a03 | sao paulo | SP | b2059ed67ce144a36e2aa97d2c9e9ad2 | delivered | 2018-05-19 16:07:45 | 1.0 | credit_card | 7.0 | 157.73 |
| 3 | b2b6027bc5c5109e529d4dc6358b12c3 | mogi das cruzes | SP | 951670f92359f4fe4a63112aa7306eba | delivered | 2018-03-13 16:06:38 | 1.0 | credit_card | 1.0 | 173.30 |
| 4 | 4f2d8ab171c80ec8364f7c12e35b23ad | campinas | SP | 6b7d50bd145f6fc7f33cebabd7e49d0f | delivered | 2018-07-29 09:51:30 | 1.0 | credit_card | 8.0 | 252.25 |

## Data Frame Properties

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 93120 entries, 0 to 103743
Data columns (total 10 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   customer_id               93120 non-null  object
 1   customer_city             93120 non-null  object
 2   customer_state            93120 non-null  object
 3   order_id                  93120 non-null  object
 4   order_status              93120 non-null  object
 5   order_purchase_timestamp  93120 non-null  datetime64[ns]
 6   payment_sequential        93120 non-null  float64
 7   payment_type              93120 non-null  object
 8   payment_installments      93120 non-null  float64
 9   payment_value             93120 non-null  float64
```

# Data Overview/ Insights

*The data has been cleaned and merged before analysis is conducted.*

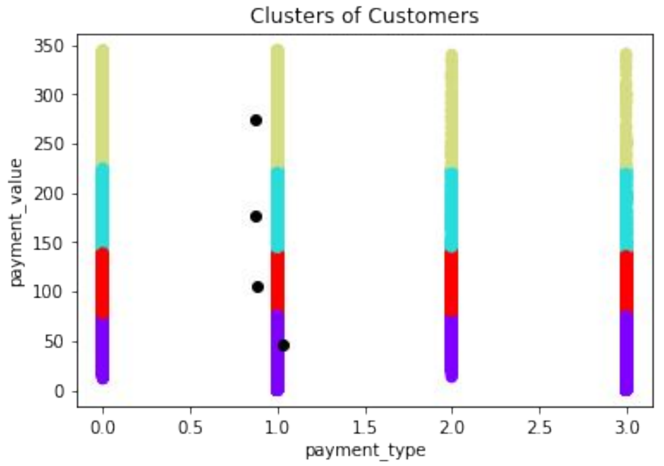
- There are **88.975** unique customer in the clean dataset, who buys **93.120** times in the e-commerce.
- Most of the customers **come from Sao Paulo city**, whose state, SP, **comprises 43% of the total customers**.
- Most widely used payment type is credit card. **73% of success order paid by credit card**.
- **Customer on average spend $110.** Minimum payment is $0, the customer make payment using voucher. **The largest payment made is $344.**

# Clustering Analysis

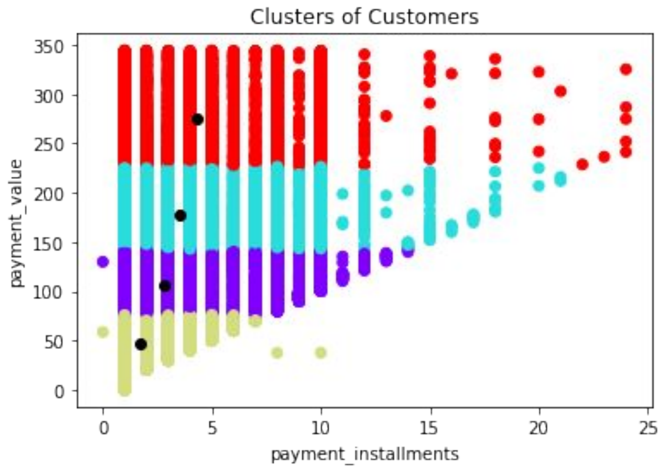
Only 4 (four) columns used in order to see the relationship and create cluster in the data. The columns used are:

- payment_value
- payment_type
- payment_installments
- customer_state

**payment value vs payment type**

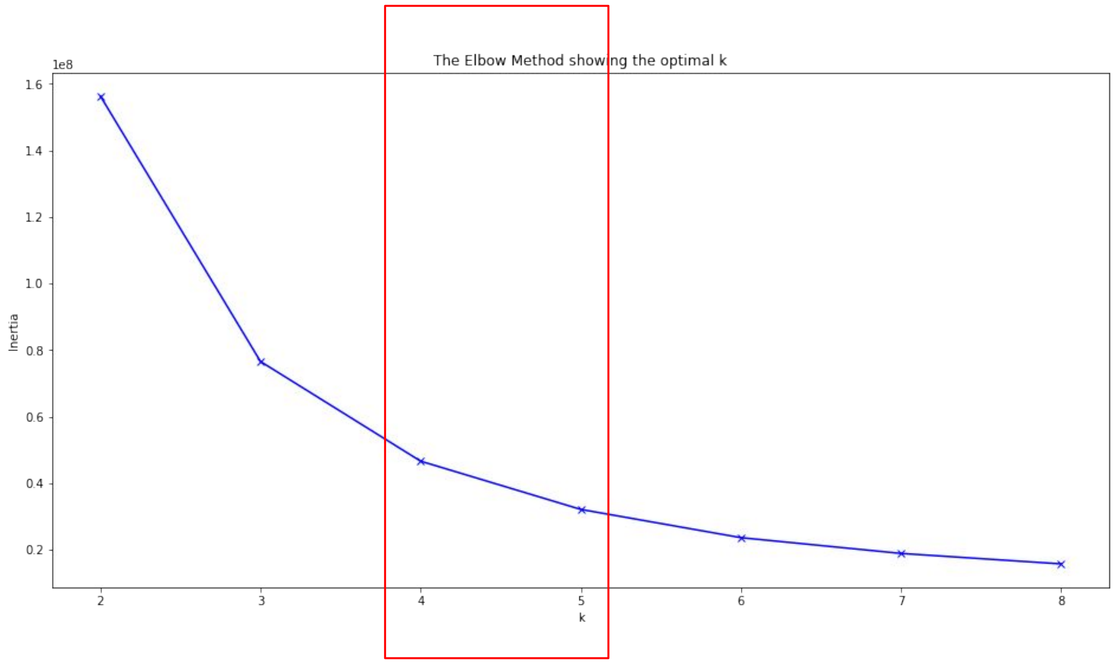



**payment value vs payment installments**




**payment value vs customer state**


**Cluster using all data**




Based on elbow analysis, we can infer the cluster within the data is 4 or 5. **Based on business judgment, we will make only 4 clusters.**

# Clustering Analysis

## Why 4 clusters? Not 5?

- Based on elbow method, **the most suitable cluster would be either 4 or 5**. I have assesed both of the options, and choose 4 cluster as the segmentation of the e-commerce customer, **because we can still differentiate the characteristics of the customer when there's only 4 clusters.** When we go more than 4 clusters, the differentiation within each cluster is vague.
- With this 4 cluster, we can give specific recommendations to each cluster, and make more targeted marketing, and hopefully, make more sales.

# Clustering Characteristics

## Descriptive Statistics each cluster:

| cluster | payment_type | payment_installments_cat | count | mean | median | max | min |
|---|---|---|---|---|---|---|---|
| 0 | credit_card | medium term | 359 | 73.032061 | 73.340 | 76.39 | 37.58 |
| | debit_card | no debt-debt club | 636 | 47.781934 | 46.555 | 76.39 | 13.38 |
| | voucher | no debt-debt club | 4091 | 29.615031 | 25.000 | 76.49 | 0.00 |
| | boleto | no debt-debt club | 7776 | 49.600837 | 49.535 | 76.51 | 11.62 |
| | credit_card | pinjol friendly | 11386 | 52.907541 | 54.100 | 76.51 | 20.03 |
| | | no debt-debt club | 14070 | 45.064644 | 45.130 | 76.52 | 0.01 |
| 1 | credit_card | long commitment | 63 | 177.073968 | 173.290 | 225.06 | 142.11 |
| | debit_card | no debt-debt club | 251 | 177.705498 | 174.160 | 225.70 | 141.62 |
| | voucher | no debt-debt club | 315 | 177.914603 | 175.490 | 225.71 | 141.72 |
| | credit_card | no debt-debt club | 2463 | 174.524633 | 170.280 | 225.96 | 141.60 |
| | | medium term | 3101 | 180.846301 | 179.190 | 225.96 | 141.51 |
| | boleto | no debt-debt club | 3249 | 176.294503 | 172.520 | 225.97 | 141.63 |
| | credit_card | pinjol friendly | 8798 | 176.810631 | 173.255 | 225.99 | 141.60 |
| 2 | credit_card | long commitment | 38 | 277.793947 | 273.615 | 338.77 | 228.71 |
| | debit_card | no debt-debt club | 102 | 273.716863 | 269.250 | 340.07 | 226.06 |
| | voucher | no debt-debt club | 112 | 269.712054 | 253.580 | 341.02 | 226.88 |
| | credit_card | no debt-debt club | 832 | 274.410300 | 268.340 | 343.91 | 226.11 |
| | boleto | no debt-debt club | 1337 | 274.435669 | 267.790 | 344.33 | 226.04 |
| | credit_card | medium term | 1901 | 276.402593 | 271.580 | 344.32 | 226.12 |
| | | pinjol friendly | 3579 | 274.713954 | 268.450 | 344.34 | 226.05 |
| 3 | credit_card | long commitment | 13 | 135.566923 | 135.520 | 139.98 | 130.57 |
| | debit_card | no debt-debt club | 417 | 105.225156 | 102.350 | 140.91 | 76.57 |
| | voucher | no debt-debt club | 904 | 102.199226 | 100.000 | 141.13 | 76.66 |
| | credit_card | medium term | 3019 | 106.300119 | 105.280 | 141.50 | 76.44 |
| | boleto | no debt-debt club | 5603 | 105.936682 | 104.690 | 141.58 | 76.55 |
| | credit_card | no debt-debt club | 6825 | 102.846399 | 98.800 | 141.53 | 76.54 |
| | | pinjol friendly | 11880 | 108.308539 | 107.780 | 141.58 | 76.50 |

| cluster | total | mean | median | max | min |
|---|---|---|---|---|---|
| 0 | 38318 | 46.973324 | 47.84 | 76.52 | 0.00 |
| 1 | 18240 | 177.128408 | 173.60 | 225.99 | 141.51 |
| 2 | 7901 | 274.972215 | 268.84 | 344.34 | 226.04 |
| 3 | 28661 | 106.107420 | 104.73 | 141.58 | 76.44 |

# Customer Segmentation

## Summary:

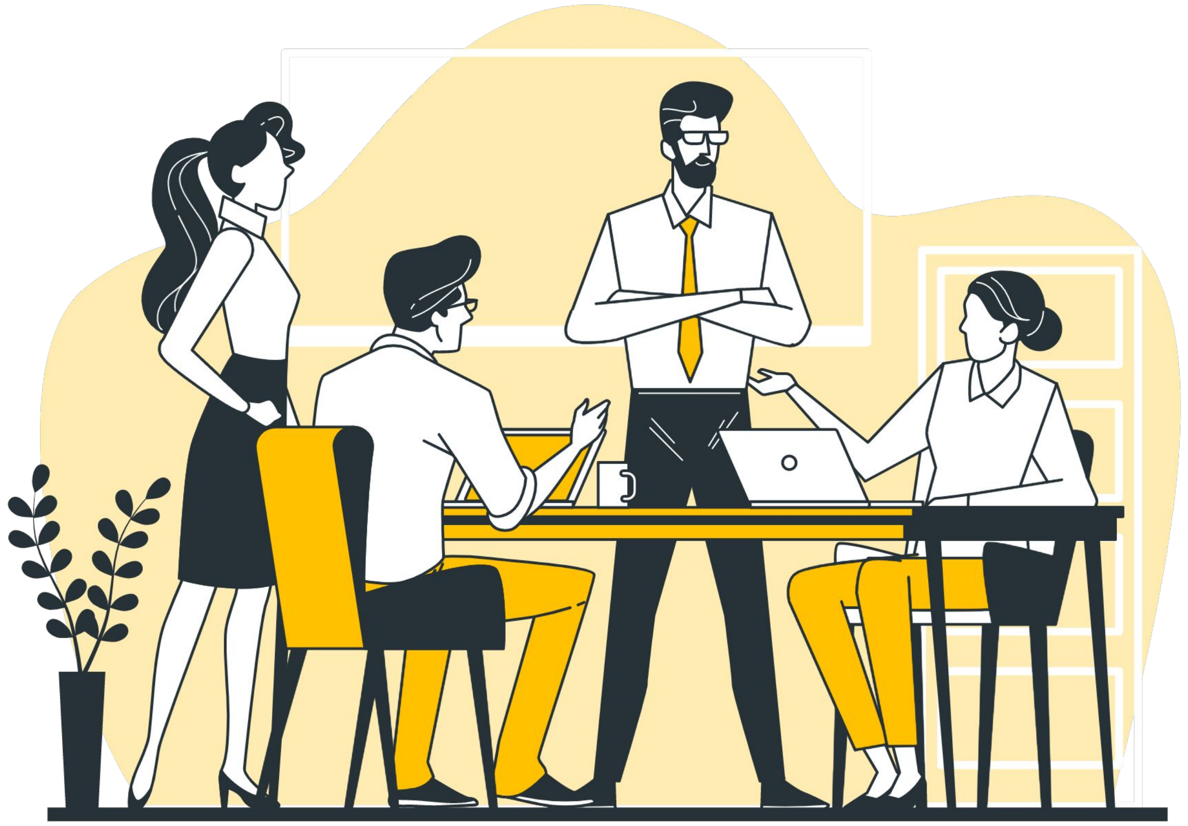
The distribution of data on each cluster quite good (no cluster with small count).
Cluster characteristic:

a. **cluster 0: team lunas**
characteristics:
- Average spending is around $46.91.
- Most of the customers using credit card without installments.

b. **cluster 1: team cc & offline payment**
characteristics:
- Average spending is around $176.84.
- Most of the customers using credit card with <= 6 months installments. Beside credit card, they are also using boleto (some kind of offline payment, common in Brazil, in Indonesia very similar with pay in Indomaret/Alfamart).

c. **cluster 2: team buy now, pay later**
characteristics:
- Average spending is around $274.67.
- Most of the customers pay using credit card within 1 year installments (medium term).

d. **cluster 3: team cc short term**
characteristics:
- Average spending is around $105.92.
- Most of the customers pay using credit card with <= 6 months installments, meaning that they are pinjol friendly 🤑.



# Customer Segmentation

## Recommendation:

a. **cluster 0: team lunas**
- Cluster 0 named as team lunas, because they tend to purchase goods directly with credit card without installments.
- We can make bundling promotions on the items they usually buy, and expect their spending behaviour will improve.

b. **cluster 1: team cc & offline payment**
- This is our second biggest spender in our e-commerce. In order to target this cluster, we have to improve our boleto payment to be more seamless, because this cluster rely on credit card and boleto.
- We can also give specialized discount if people pay the items using boleto.

c. **cluster 2: team buy now, pay later**
- This is our highest spender in our e-commerce. In order to target this cluster, we have to give more incentive for people paying with credit card + installments.
- We can partner with bank, and give various bank-related discounts. We can also give 0% interest for the installments

d. **cluster 3: team cc short term**
- We can use similar strategy to attract customer from this cluster with strategy from cluster 2. We can give incentive by offering more bank-related discounts, and give 0% interest for the installments.

# Thank you!

**Full Python Script:**

https://tinyurl.com/cust-segmentation

**Any question, comment, or feedback?**

**Feel free to reach me at:**

https://www.linkedin.com/in/gilangbahana/