Predict Customer Personality to Boost Marketing Campaign by Using Machine Learning

By: Joni Syofian

Supported by:
Rakamin Academy

About Me

Joni is a fresh graduate student from Bandung Institute of Technology. He is interested in data science, data analytics, and ocean issues. To improve his skills in the field of data, he took several courses and just completed a data science bootcamp with a good grade.

Table of Content

01
Business Understanding

02
Data Overview

03
Exploratory Data Analysis

04
Insights

05
Data Preprocessing

06
Modeling

07
Cluster Interpretation

08
Business Strategy Recomendations

BUSINESS UNDERSTANDING

Background

A company can develop rapidly when it knows its customers' personalities and behaviors, so it can provide better services and benefits to customers who have the potential to become loyal customers. Our focus is to create a cluster prediction model to make it easier for companies to make decisions by processing historical marketing campaign data to improve performance and target the right customers so they can transact on the company's platform.

Goals

Enhance the effectiveness of marketing initiatives and target the appropriate customers to encourage them to transact on the company's platform.

Objective

Create a cluster prediction model using unsupervised learning to facilitate decision-making for businesses.

DATA OVERVIEW

```
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Unnamed: 0           2240 non-null   int64
 1   ID                   2240 non-null   int64
 2   Year_Birth           2240 non-null   int64
 3   Education            2240 non-null   object
 4   Marital_Status       2240 non-null   object
 5   Income               2216 non-null   float64
 6   Kidhome              2240 non-null   int64
 7   Teenhome             2240 non-null   int64
 8   Dt_Customer          2240 non-null   object
 9   Recency              2240 non-null   int64
 10  MntCoke              2240 non-null   int64
 11  MntFruits            2240 non-null   int64
 12  MntMeatProducts      2240 non-null   int64
 13  MntFishProducts      2240 non-null   int64
 14  MntSweetProducts     2240 non-null   int64
 15  MntGoldProds         2240 non-null   int64
 16  NumDealsPurchases    2240 non-null   int64
 17  NumWebPurchases      2240 non-null   int64
 18  NumCatalogPurchases  2240 non-null   int64
 19  NumStorePurchases    2240 non-null   int64
 20  NumWebVisitsMonth    2240 non-null   int64
 21  AcceptedCmp3         2240 non-null   int64
 22  AcceptedCmp4         2240 non-null   int64
 23  AcceptedCmp5         2240 non-null   int64
 24  AcceptedCmp1         2240 non-null   int64
 25  AcceptedCmp2         2240 non-null   int64
 26  Complain             2240 non-null   int64
 27  Z_CostContact        2240 non-null   int64
 28  Z_Revenue            2240 non-null   int64
 29  Response             2240 non-null   int64
dtypes: float64(1), int64(26), object(3)
```

Description

The data set contains behavior features of customers who's made transaction and interaction in our platform.

Rows and Columns

There are 2240 rows and 30 columns

Data Types

1 Float, 26 Integer, and 3 Object

Data Duplicates

There is no duplicate

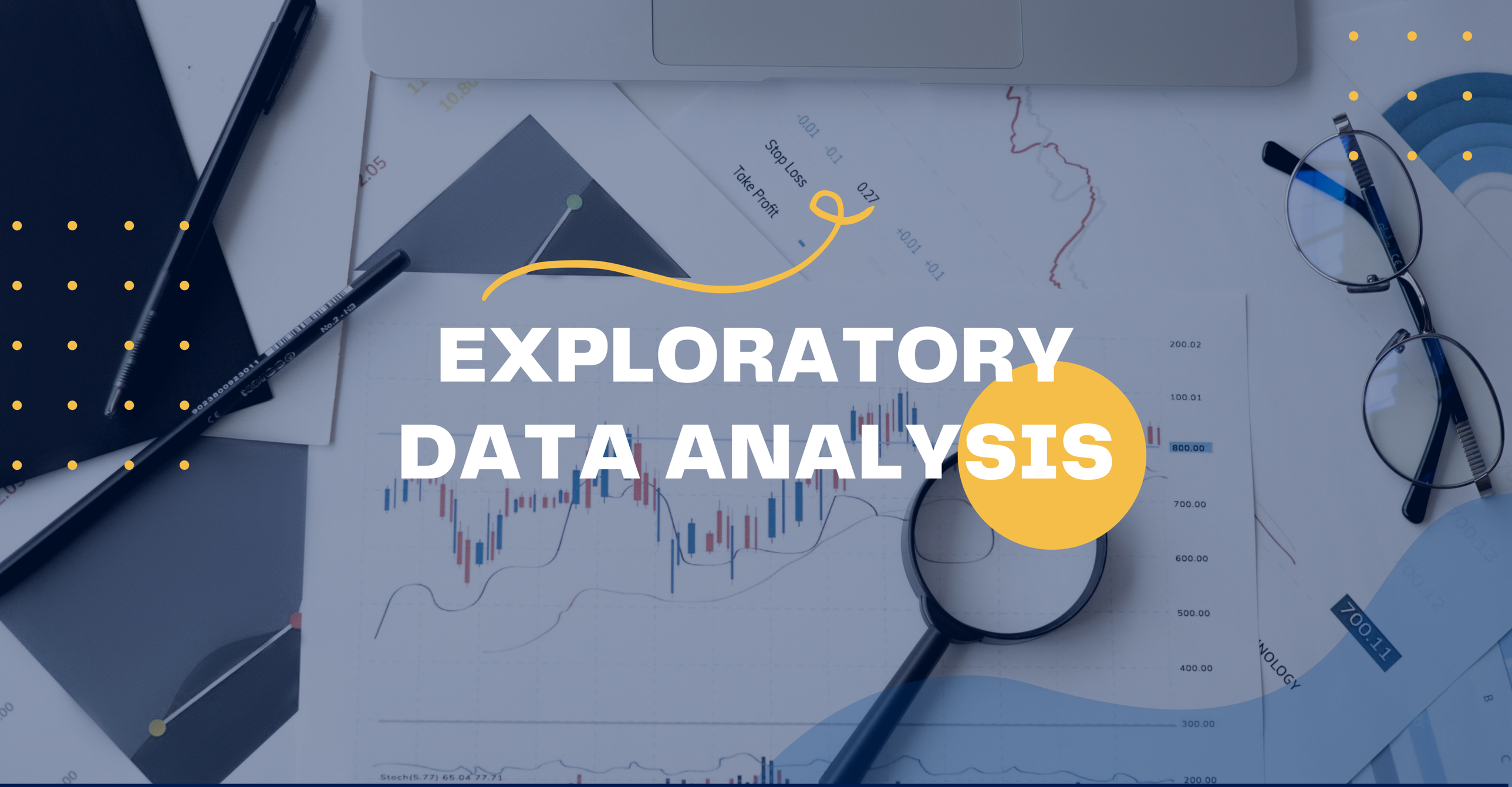
Missing Value

income columns (24 rows)

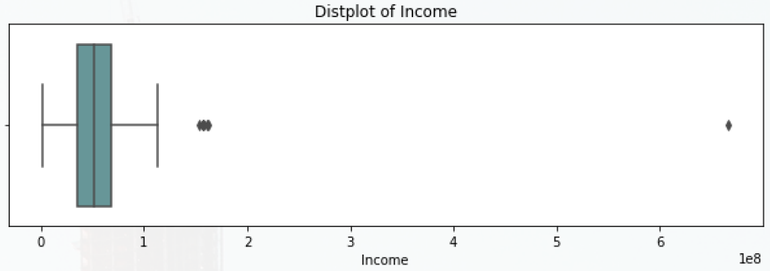
Feature Engineering

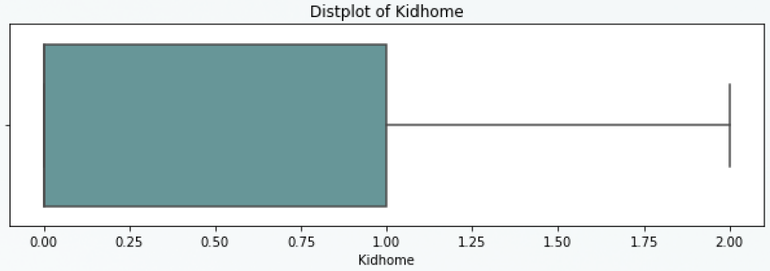
From the owned data, several columns are then created, namely:

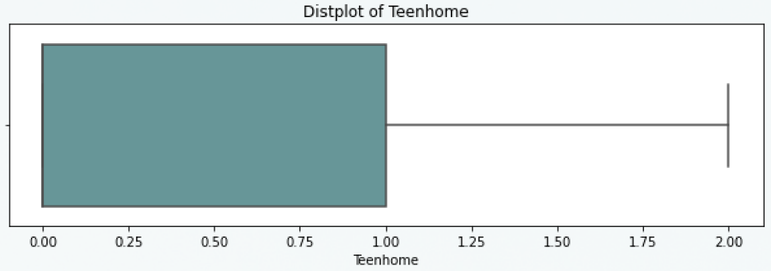
- **conversion_rate**: (responses / visit) customer history
- **age:** `Year_Birth` - now.
- **number of children (noc):** 'Kidhome` + `Teenhome`.
- **Total monenumber of children (noc):** from `Kidhome` + `Teenhome`.
- **spent `total_spent`:** `MntCoke` + `MntFruits` + `MntMeatProducts` + `MntFishProducts` + ` MntSweetProducts` + `MntGoldProds`
- **Total transaction (total_transaction)** : from `NumDealsPurchases ` + `NumWebPurchases` + `NumCatalogPurchases` + `NumStorePurchases`
- **Total of days joined (days) :** now - `Dt_Customer`

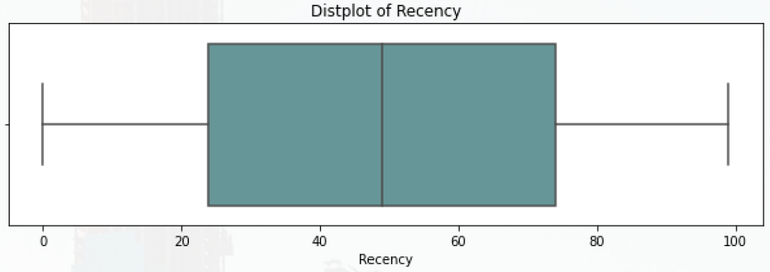


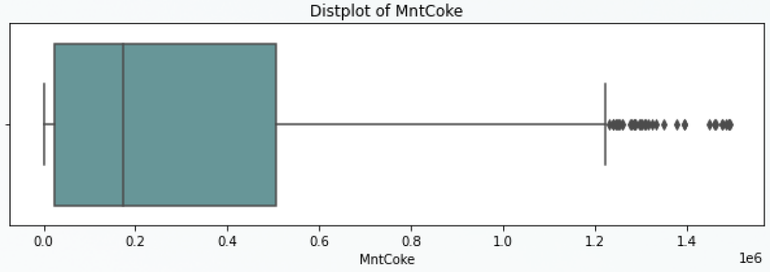
EXPLORATORY DATA ANALYSIS

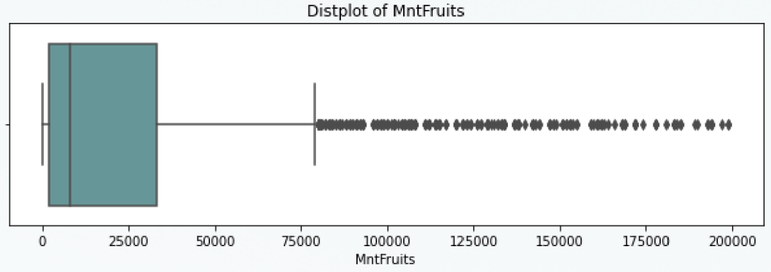


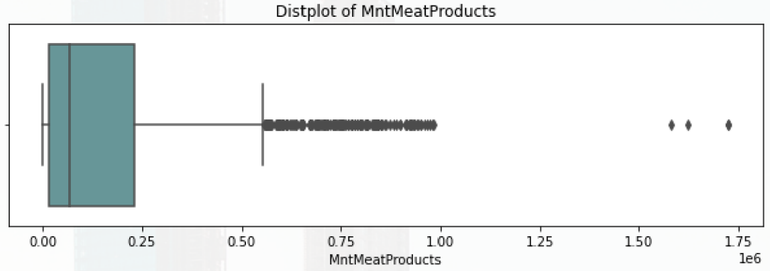


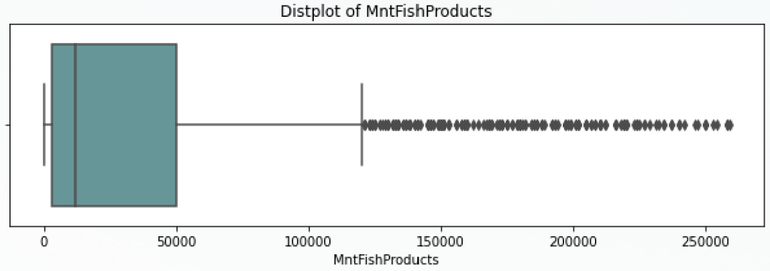


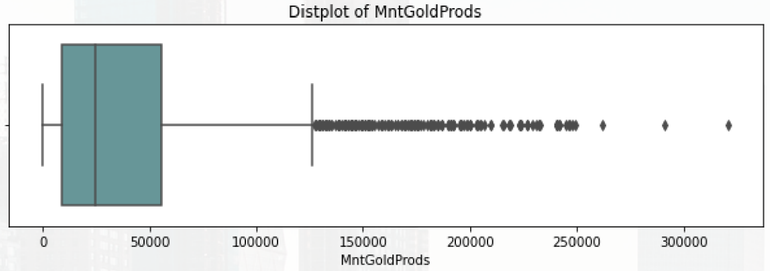


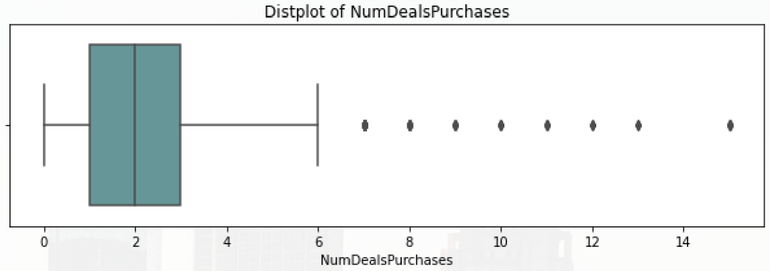


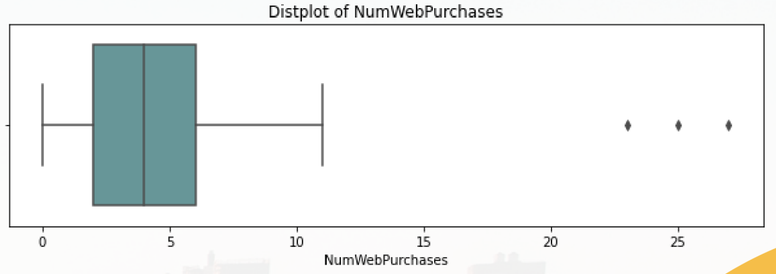


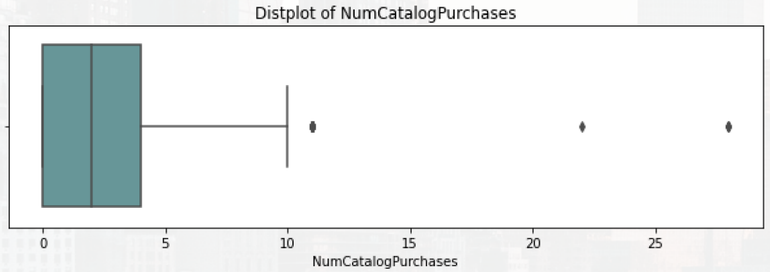


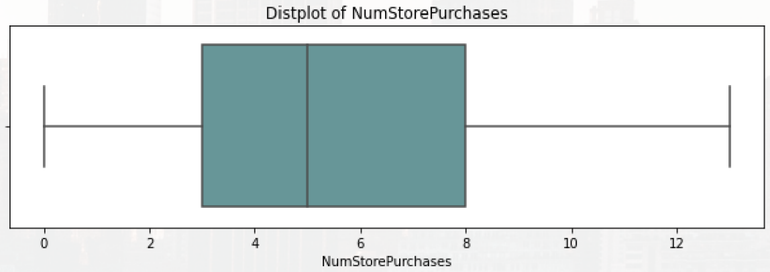




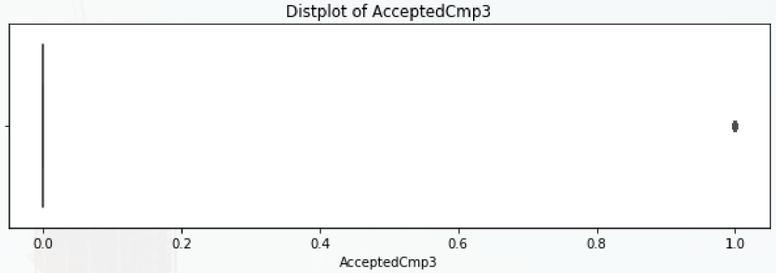


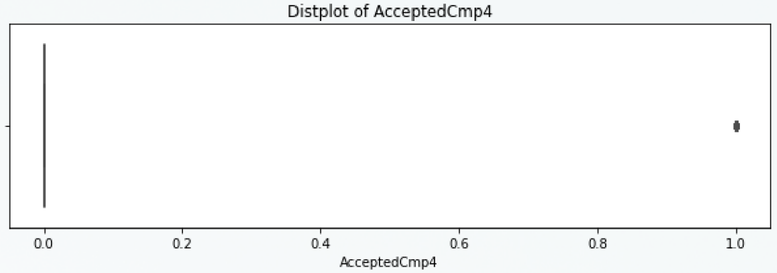


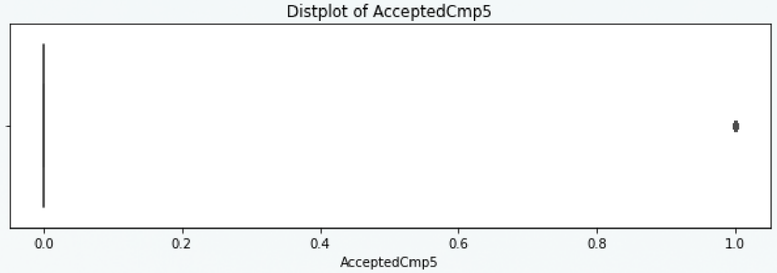


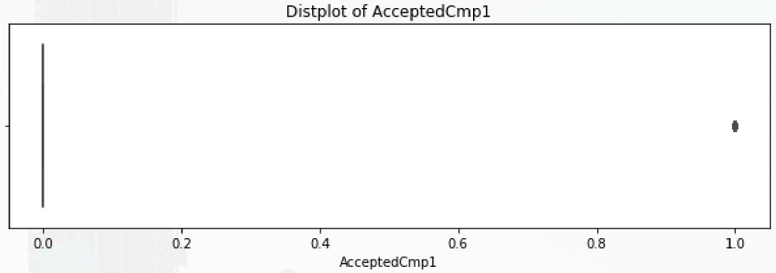


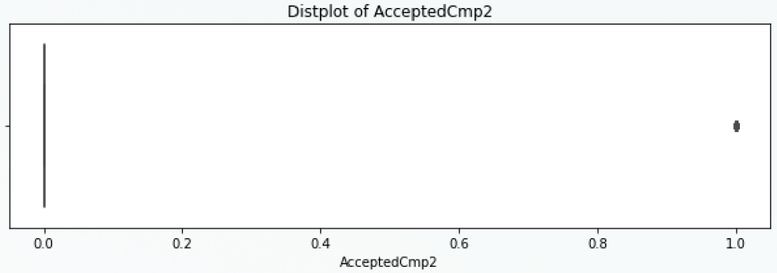
There are outliers in our data set

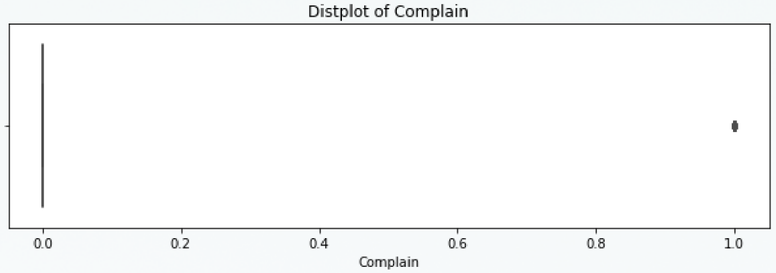


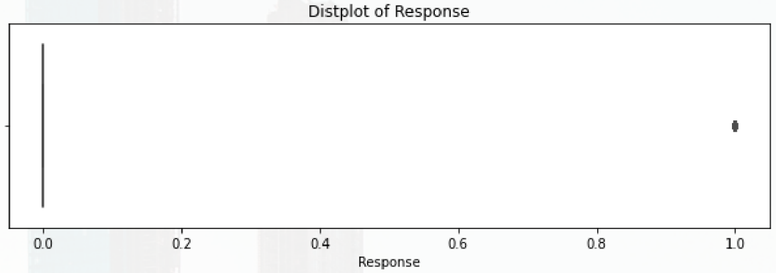


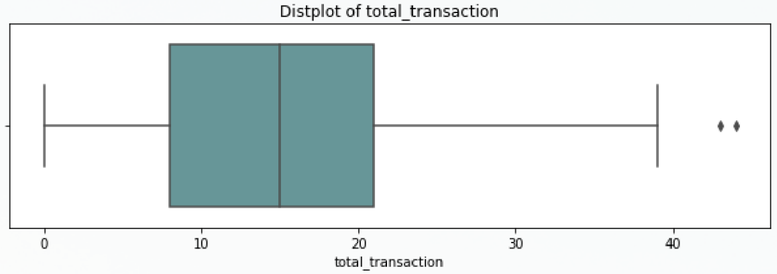


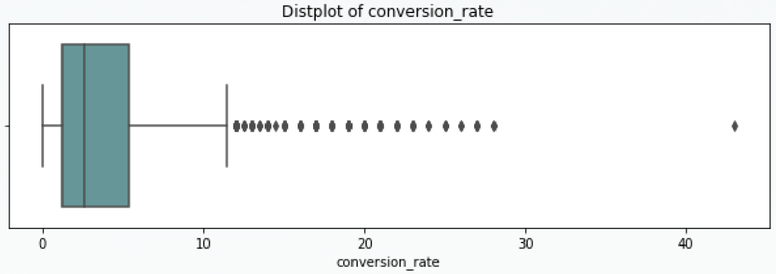


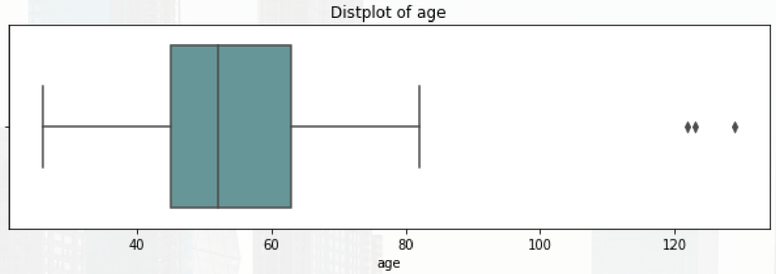


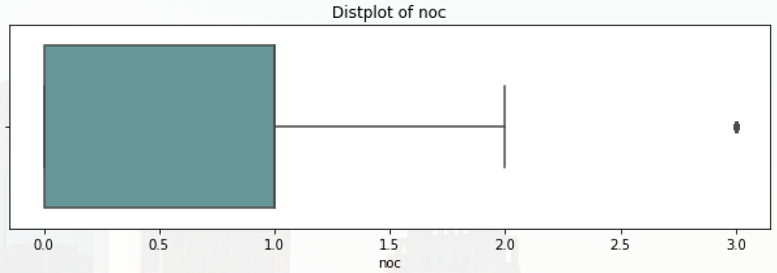


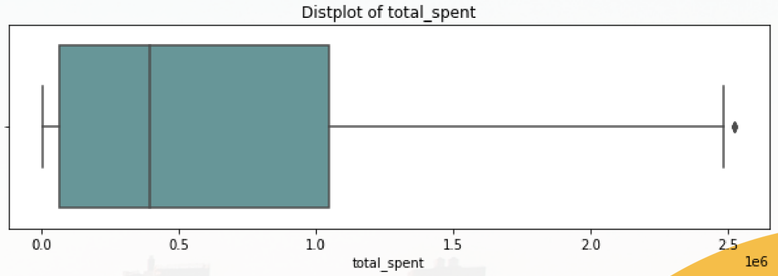


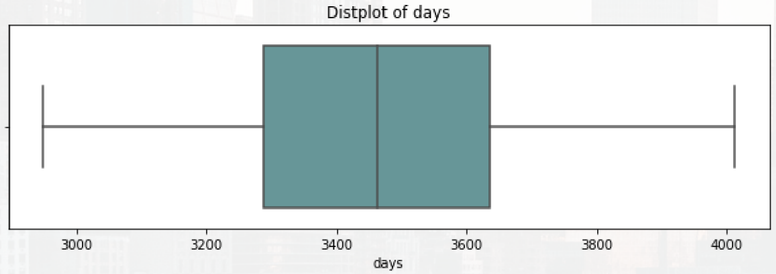












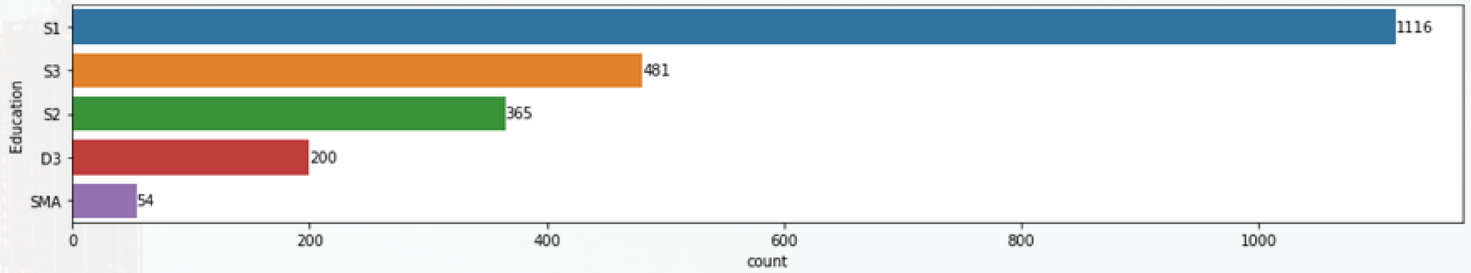
There are outliers in our data set

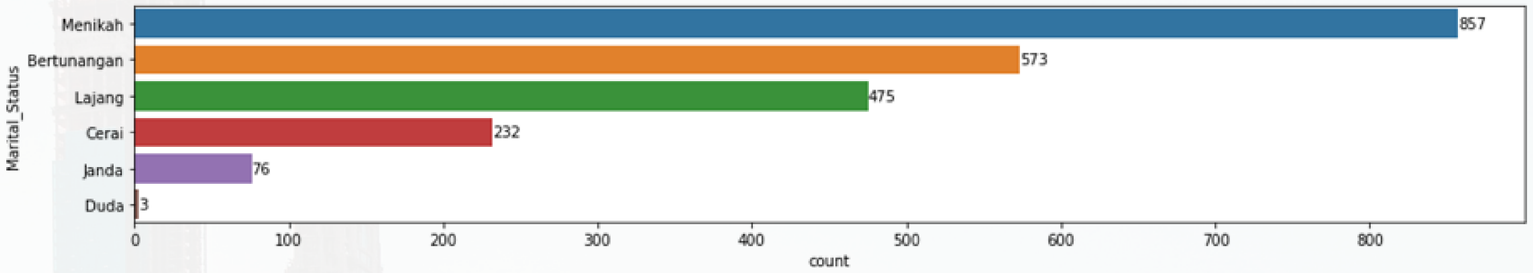
SKEWNESS

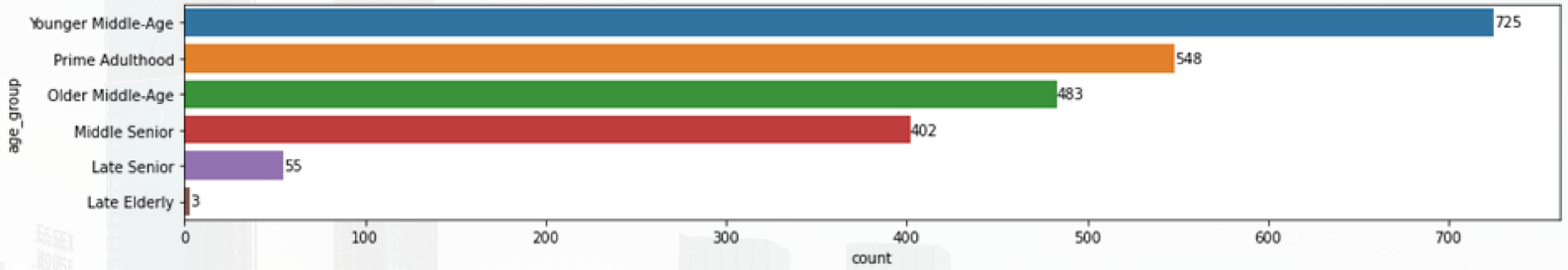
```
Skewness Income : 6.763487372811117
Skewness Kidhome : 0.6356100646634026
Skewness Teenhome : 0.40755280239070707
Skewness Recency : 0.0016477067463847978
Skewness MntCoke : 1.1707200955543913
Skewness MntFruits : 2.101657525150595
Skewness MntMeatProducts : 2.0255768067844637
Skewness MntFishProducts : 1.916368950232168
Skewness MntSweetProducts : 2.1033275863706797
Skewness MntGoldProds : 1.839230936129571
Skewness NumDealsPurchases : 2.415271762315824
Skewness NumWebPurchases : 1.1970370130708343
Skewness NumCatalogPurchases : 1.8810750511540515
Skewness NumStorePurchases : 0.7018262973284631
Skewness NumWebVisitsMonth : 0.2180430456390019
Skewness AcceptedCmp3 : 3.2693969568782
Skewness AcceptedCmp4 : 3.256758193853604
Skewness AcceptedCmp5 : 3.2821432492590605
Skewness AcceptedCmp1 : 3.5624821800168314
Skewness AcceptedCmp2 : 8.424753003647503
Skewness Complain : 10.132736682249801
Skewness Response : 1.9587479194384483
Skewness total_transaction : 0.25093607541663454
Skewness conversion_rate : 2.325809490591629
Skewness age : 0.35366147002882264
Skewness noc : 0.4087482263114007
Skewness total_spent : 0.8580547969469019
Skewness days : 0.006096201015139509
```

- **Category features have a positive skew**. However, the skew value of these features is not so big; it's just that the `Complain` feature has a skew value that reaches 10.

- From the boxplot, It can be seen that there are outliers in the data. Then the `AcceptedCmpx` data includes the `Complain` feature, which has a strange plot because basically the feature is a boolean that is worth 0 or 1. As a result, the distribution plot and box plot appear strange.

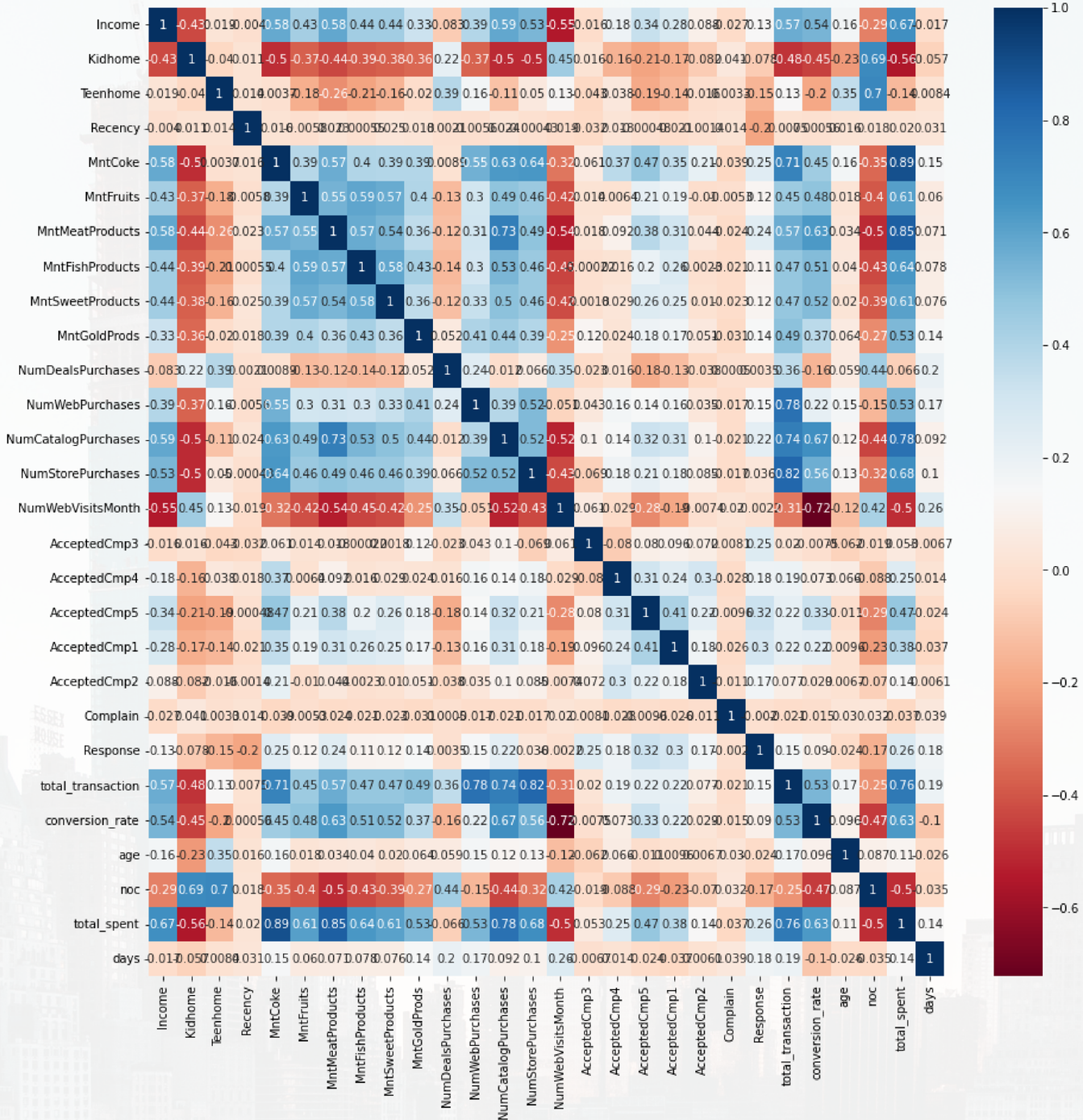






	Dt_Customer
count	2216
unique	662
top	2012-08-31 00:00:00
freq	12
first	2012-01-08 00:00:00
last	2014-12-06 00:00:00

- Based on the level of **education**, most customers are **S1 graduates.**
- Based on **marital status,** most customers are **married**.
- The majority of customers are **Young Middle Age (45-54 years old).**
- Most customers **joined** on **September 31, 2012.**
- The **start date of the first customer** joining is **January 8, 2012.**
- **The last customer** to join was on **December 6, 2014.**



CORRELATION

Correlation > 0.7

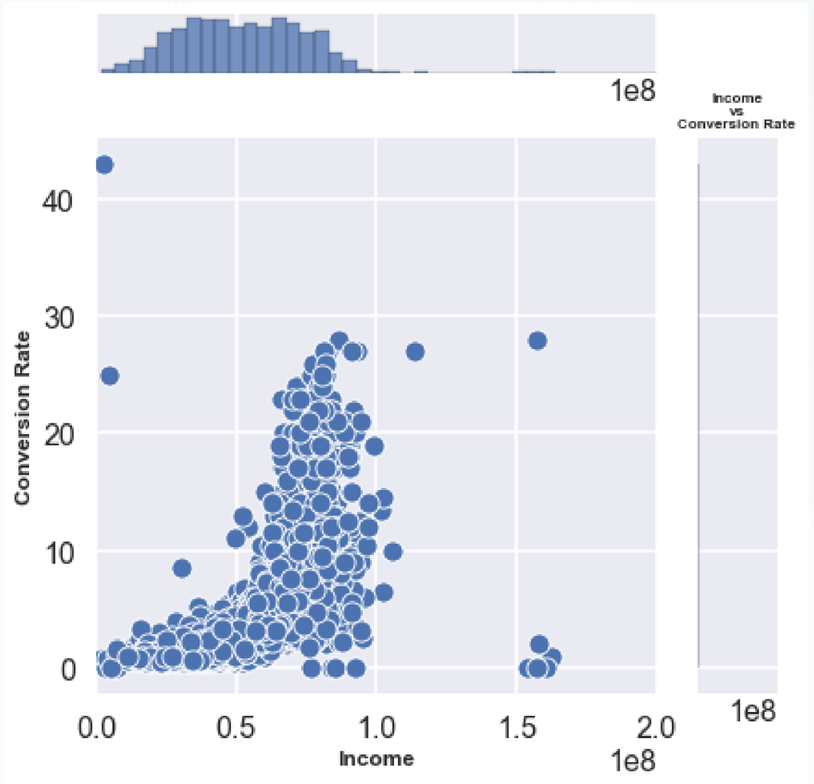
	MntCoke	MntMeatProducts	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	total_transaction	total_spent
MntCoke	1.000000	0.568860	0.553786	0.634753	0.640012	0.713508	0.893136
MntMeatProducts	0.568860	1.000000	0.307090	0.734127	0.486006	0.565360	0.845884
NumWebPurchases	0.553786	0.307090	1.000000	0.386868	0.516240	0.784238	0.528973
NumCatalogPurchases	0.634753	0.734127	0.386868	1.000000	0.517840	0.736179	0.780482
NumStorePurchases	0.640012	0.486006	0.516240	0.517840	1.000000	0.822391	0.675181
total_transaction	0.713508	0.565360	0.784238	0.736179	0.822391	1.000000	0.756403
total_spent	0.893136	0.845884	0.528973	0.780482	0.675181	0.756403	1.000000

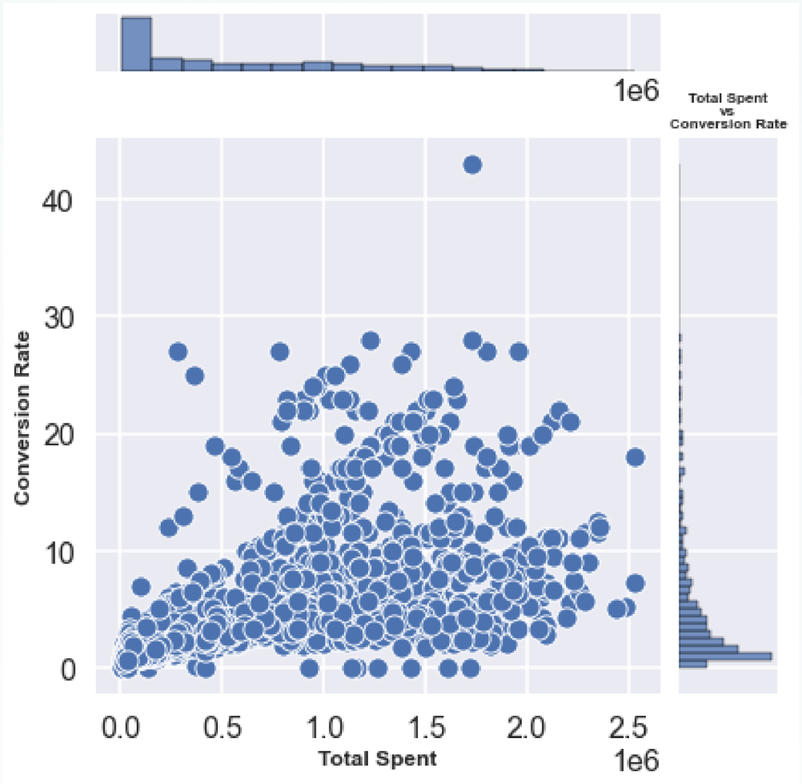
Correlation < - 0.7

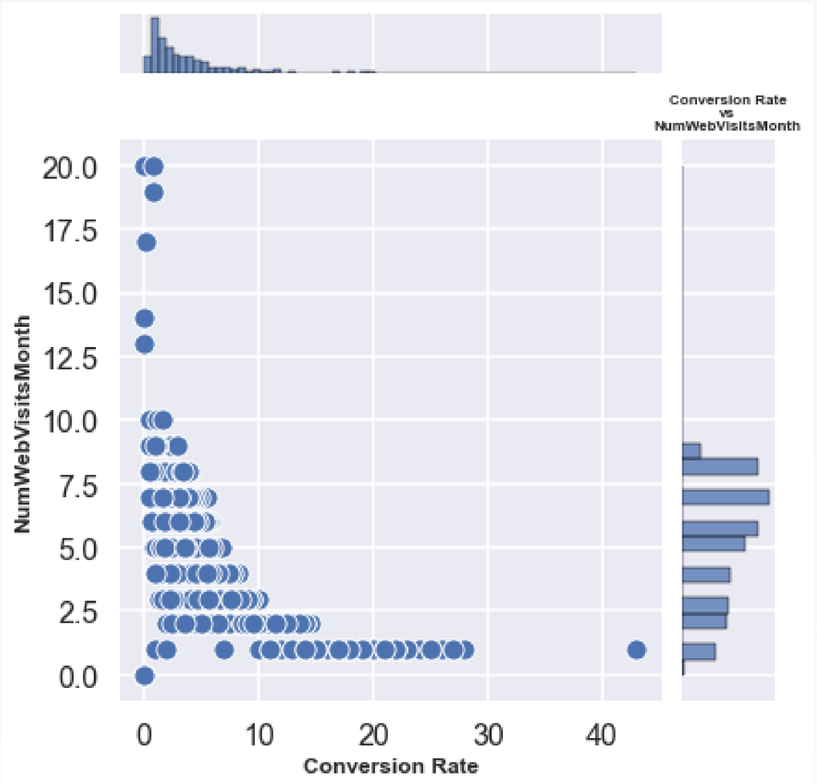
	NumWebVisitsMonth	conversion_rate
NumWebVisitsMonth	1.000000	-0.722706
conversion_rate	-0.722706	1.000000

From the information above, it can be seen that there are several features that have a correlation value of more than or equal to 0.7 and -0.7. One thing is certain: the more purchases made, the greater the `total_spent`. The interesting thing here is that the number of website visits is inversely proportional to the conversion rate. It is possible that the customer just looks at the product without making a purchase.


INSIGHTS



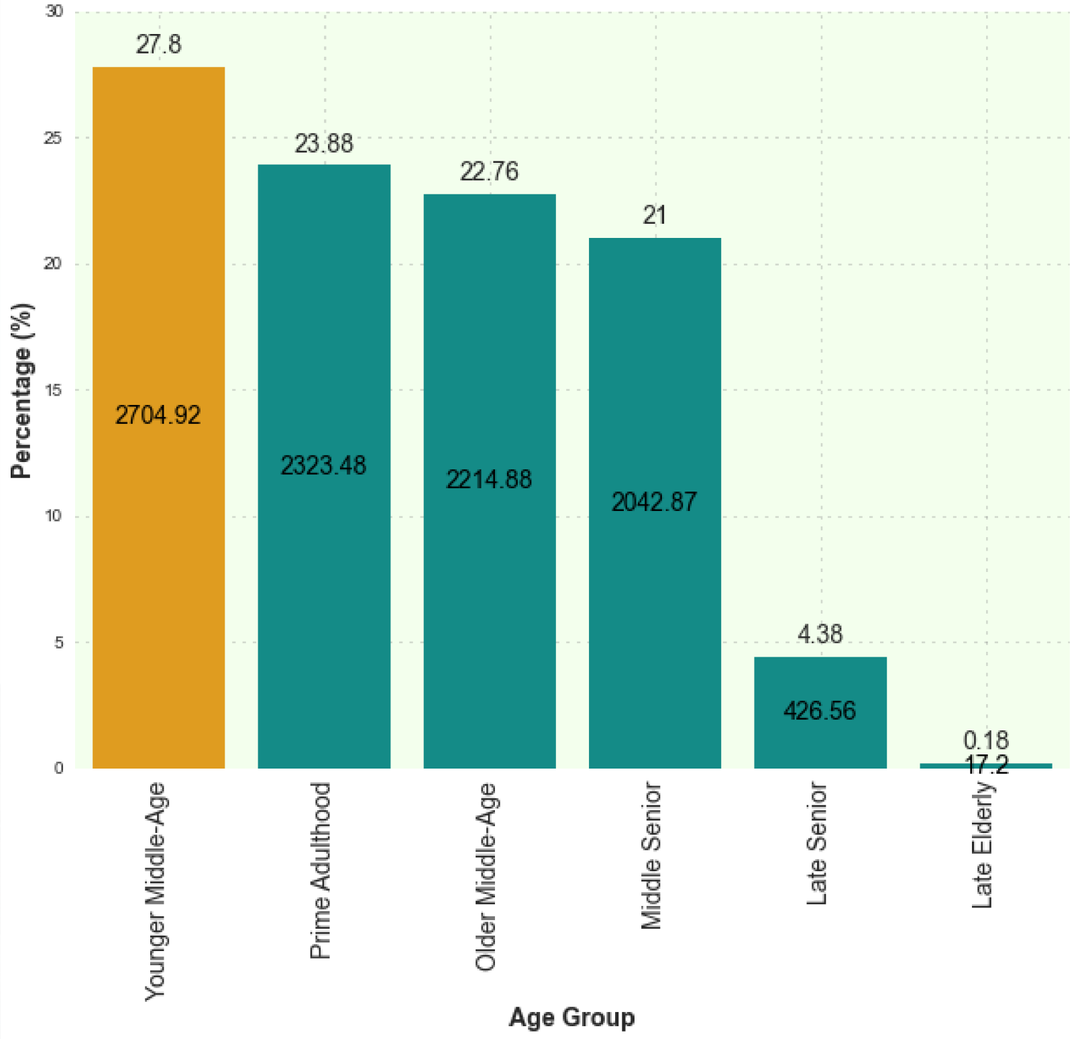


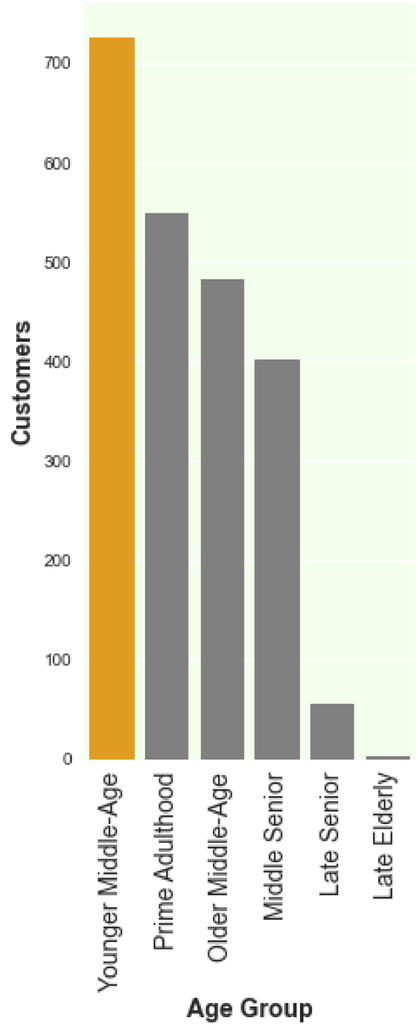


- The relationship between **conversion rate** and customer **income** has a **positive correlation**.
- The relationship between **conversion rate** and **total spent** is **positively correlated**.
- The relationship between **conversion rate** and the **number of website visits** is **inversely proportional.** This means that customers who often visit the website tend to have low conversions.

The Age Group 'Young Middle-Age'
is the age group with the most total conversion rate

Further analysis needs to be done to make this age group in marketing/supply targets
Because this age group is the group with the most customers



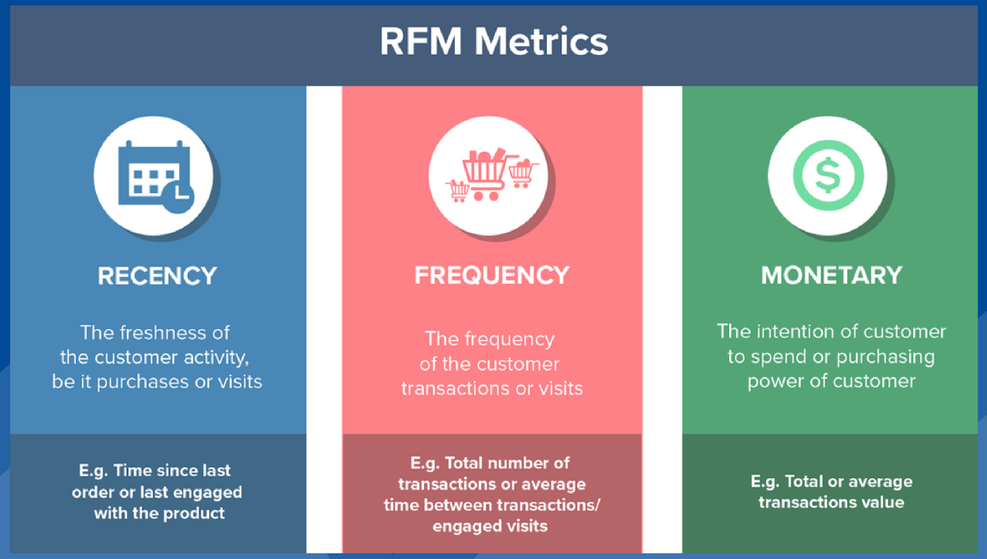


25–44: Prime Adulthood
45–54: Younger Middle-Age
55–64: Older Middle-Age/Early Senior Years
65–74: Middle Senior/Early Elderly Years
75–84: Late Senior/Middle Elderly Years
85+: Late Elderly/Geriatric Years

Customers with an age range of **45–54** are t**he largest number of customers,** as well as the customers with **the most total conversions.** So that customers in this age range can be used as marketing targets or offers by first conducting further analysis.


DATA
PREPROCESSING

Feature Selection

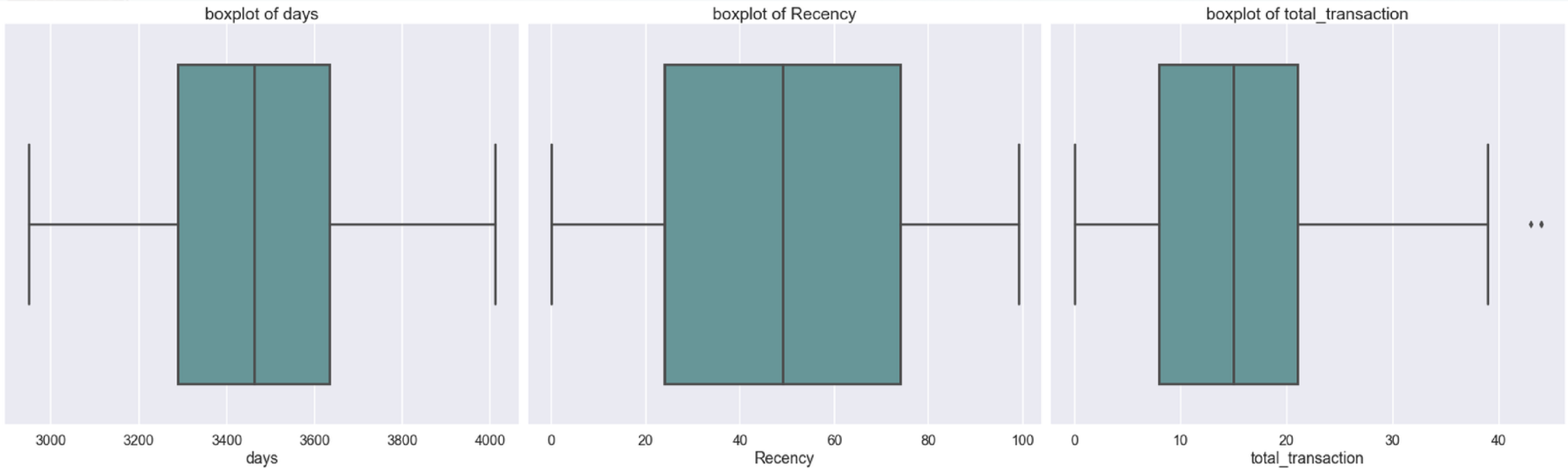


Reference: https://babel.sg/digitalizing-the-events-industry-with-artificial-intelligence-machine-learning/

To segment customers, the RFM (Recency, Frequency, Monetory) method is used. However, in this case, `days` is also used to see whether there are new users or not (L) and age (C) are also used as a feature for customer segmentation. The RFM metric is an important indicator of customer behavior because frequency and monetary value affect the customer's lifetime value, and recency affects retention.

LRFMC	Kolom
Loyalty	L = days
Recency	R = Recency
Frequency	F = total_transaction
Monetary	M = total_spent
Counts	C = age

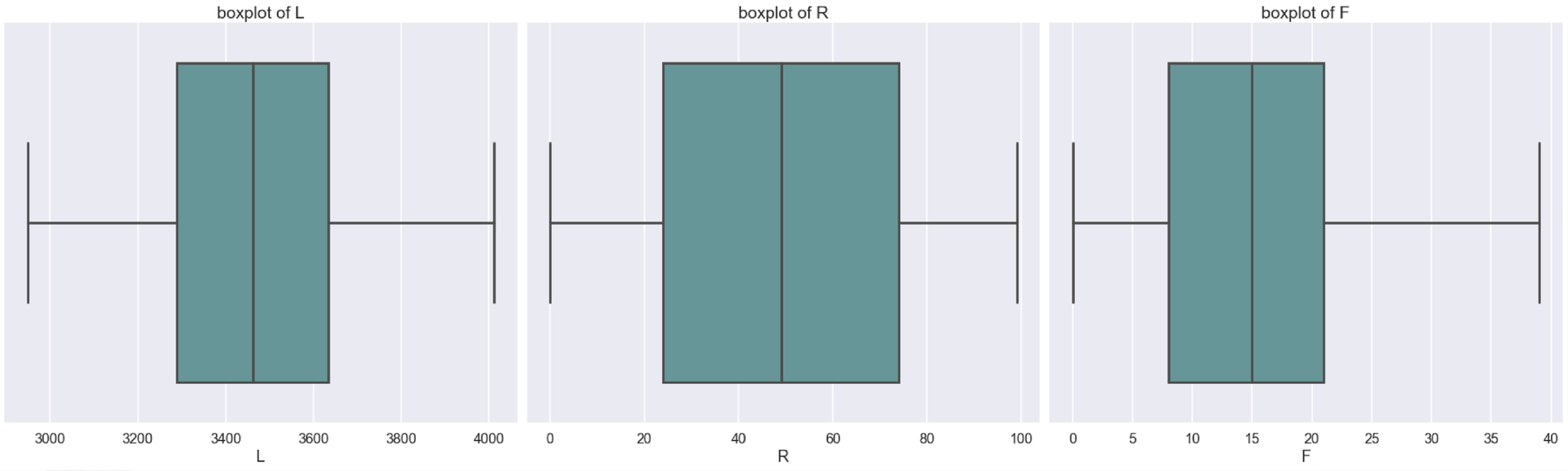
Handling Outliers

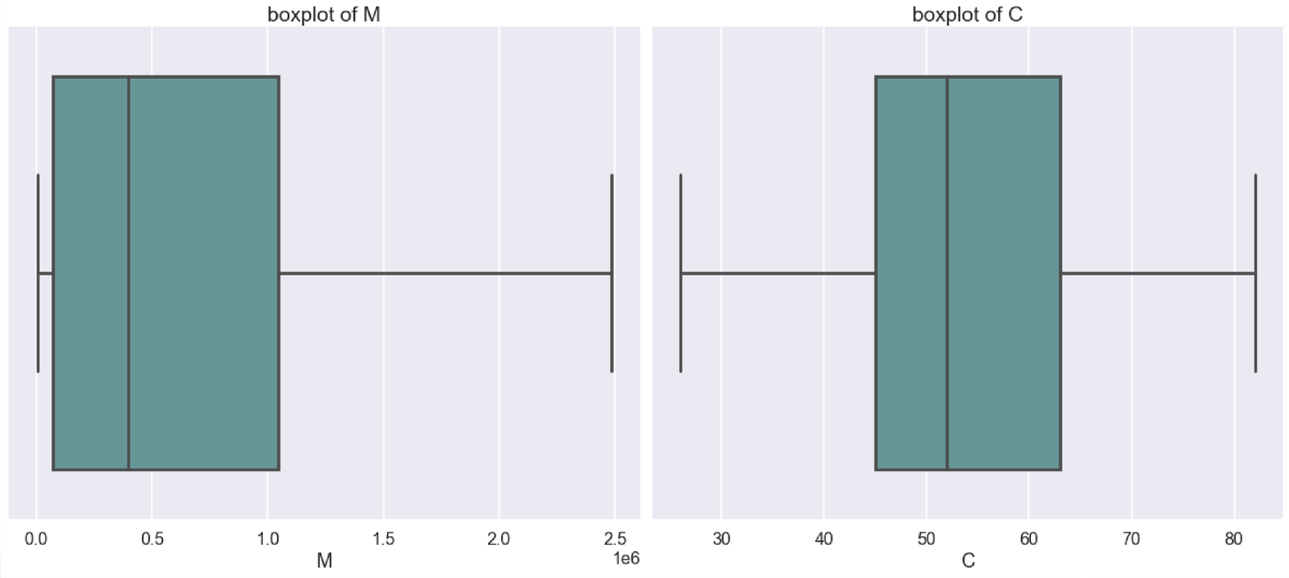


Removing outliers using the IQR method

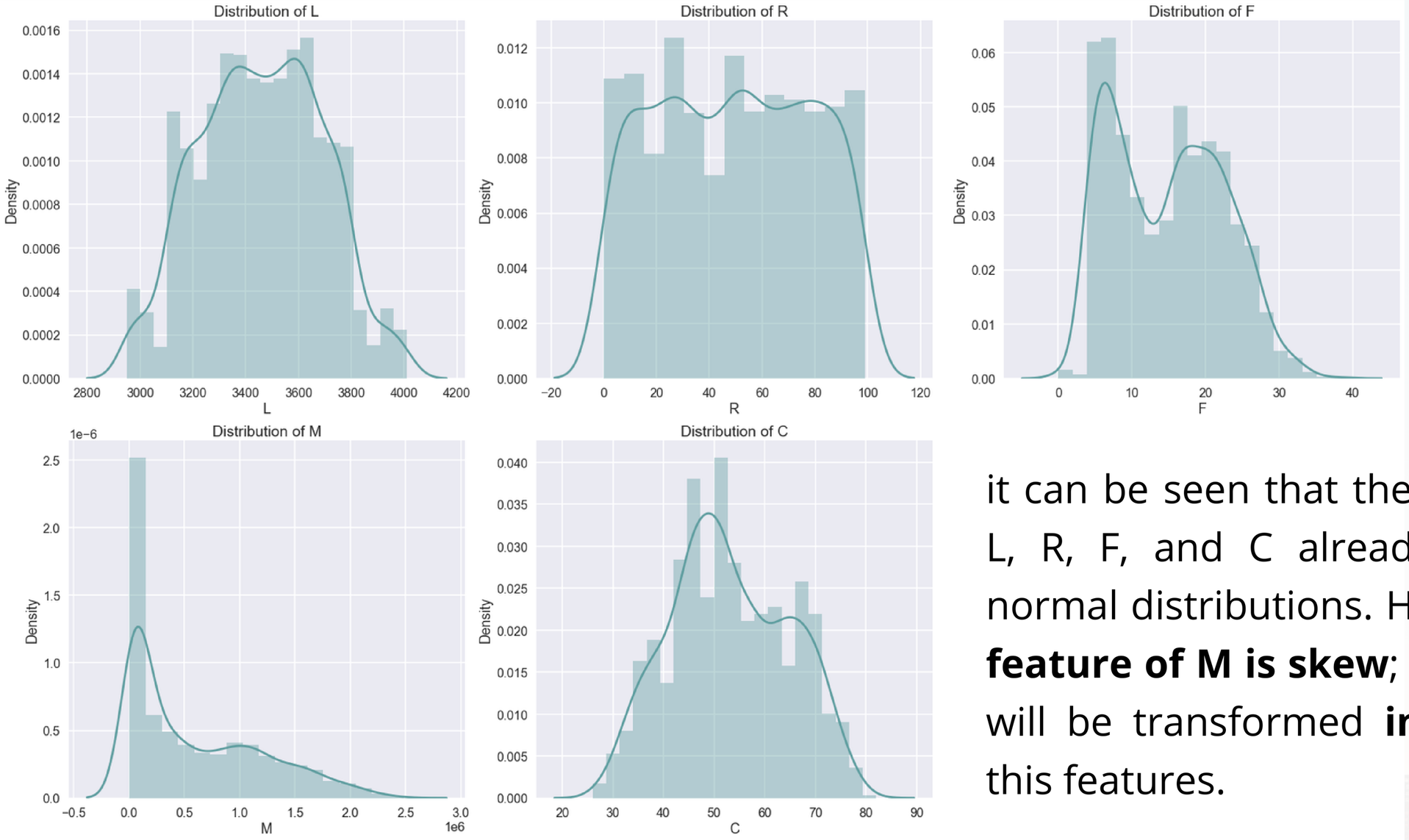
- Number of rows before filtering outliers: 2216.
- Number of rows after filtering outliers: 2208.
- Number of filtered data: 0.36%

Results


boxplot of L
3000
3200
3400
3600
3800
4000
L
boxplot of R
0
20
40
60
80
100
R
boxplot of F
0
5
10
15
20
25
30
35
40
F

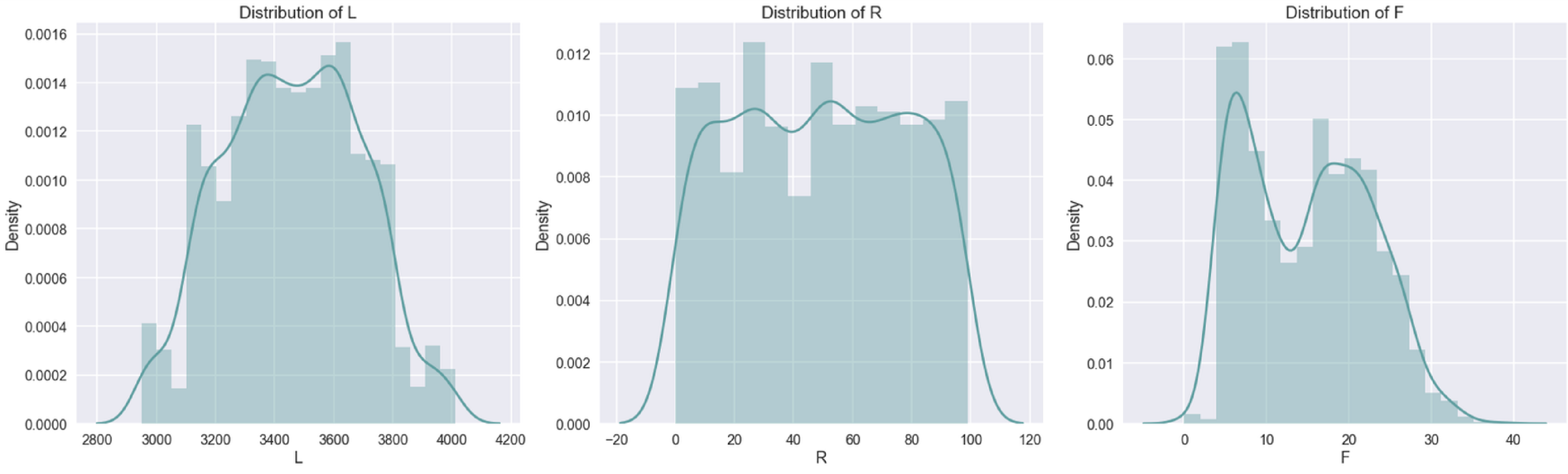

boxplot of M
0.0
0.5
1.0
1.5
2.0
2.5
1e6
M
boxplot of C
30
40
50
60
70
80
C

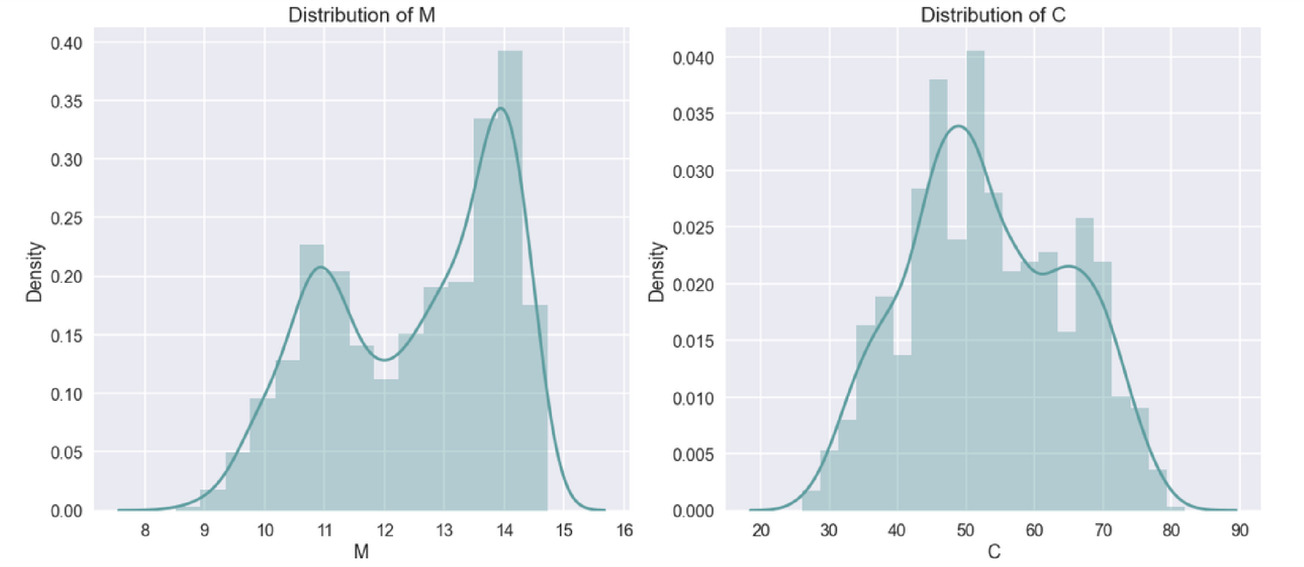
Feature Transformation



it can be seen that the features of L, R, F, and C already resemble normal distributions. However, the **feature of M is skew**; therefore, it will be transformed **into logs** on this features.

Results


Distribution of L
Density
0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
2800
3000
3200
3400
3600
3800
4000
4200
L
Distribution of R
Density
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−20
0
20
40
60
80
100
120
R
Distribution of F
Density
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
F

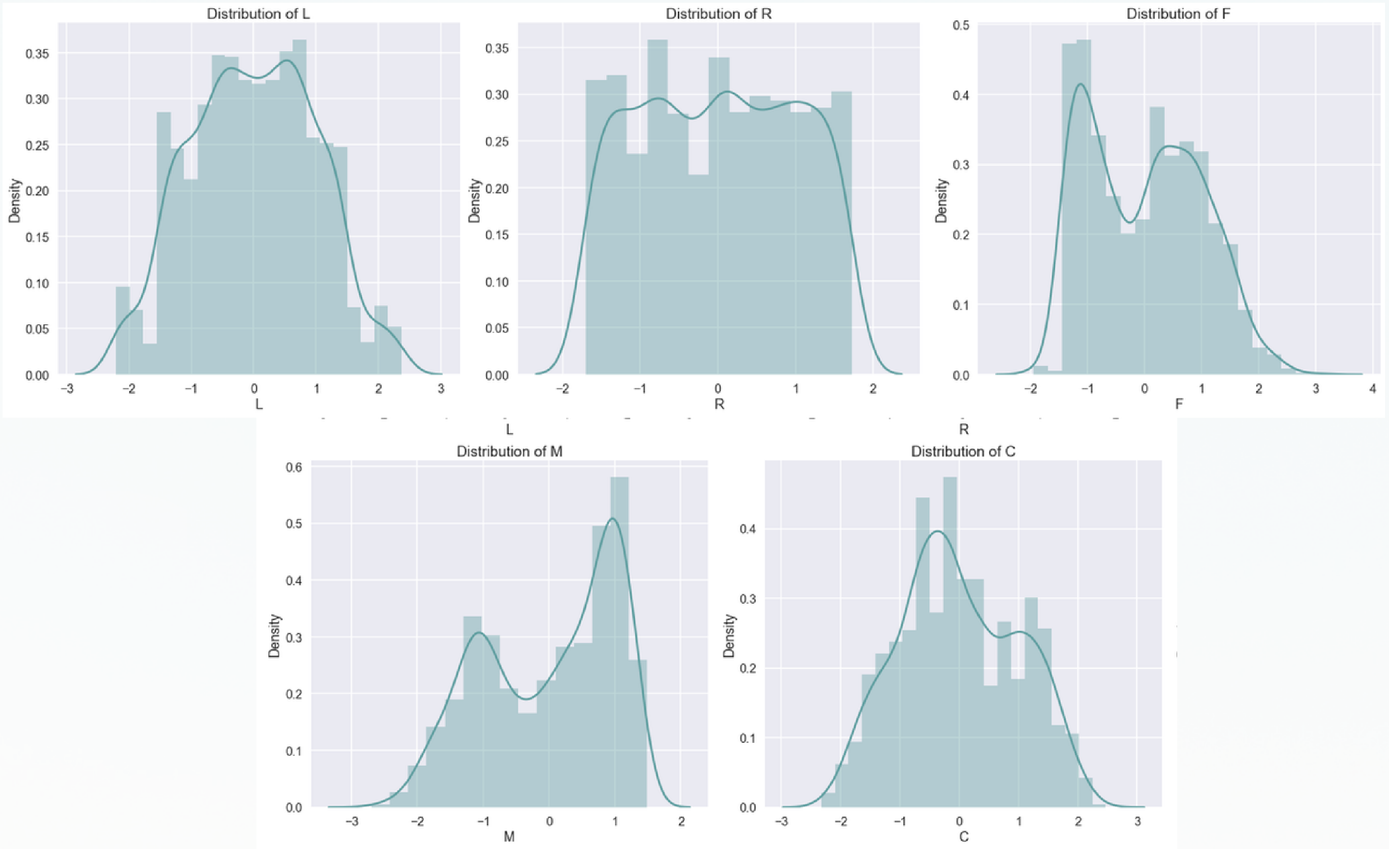

Distribution of M
Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
8
9
10
11
12
13
14
15
16
M
Distribution of C
Density
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
20
30
40
50
60
70
80
90
C

Feature Scaling

```
# Standardize data
std = StandardScaler().fit_transform(df_LRFMC_log)
df_LRFMC_std = pd.DataFrame(std, columns = list(df_LRFMC_log))
df_LRFMC_std
```

	L	R	F	M	C
0	1.973931	0.309603	1.331020	1.201455	1.017008
1	-1.665807	-0.381149	-1.161679	-1.561868	1.273868
2	-0.172912	-0.795600	0.806241	0.705727	0.332049
3	-1.923944	-0.795600	-0.899289	-1.106464	-1.294730
4	-0.822558	1.552956	0.543852	0.294420	-1.037870
...	...	...	...	...	...
2203	0.123946	-0.104849	0.412657	1.075083	0.160809
2204	-1.941153	0.240527	0.937436	0.328734	1.958827
2205	-0.848372	1.449343	0.543852	1.022755	-1.037870
2206	-0.844069	-1.417277	1.068630	0.761644	1.102628
2207	1.160798	-0.312074	-0.505705	-0.311596	1.273868

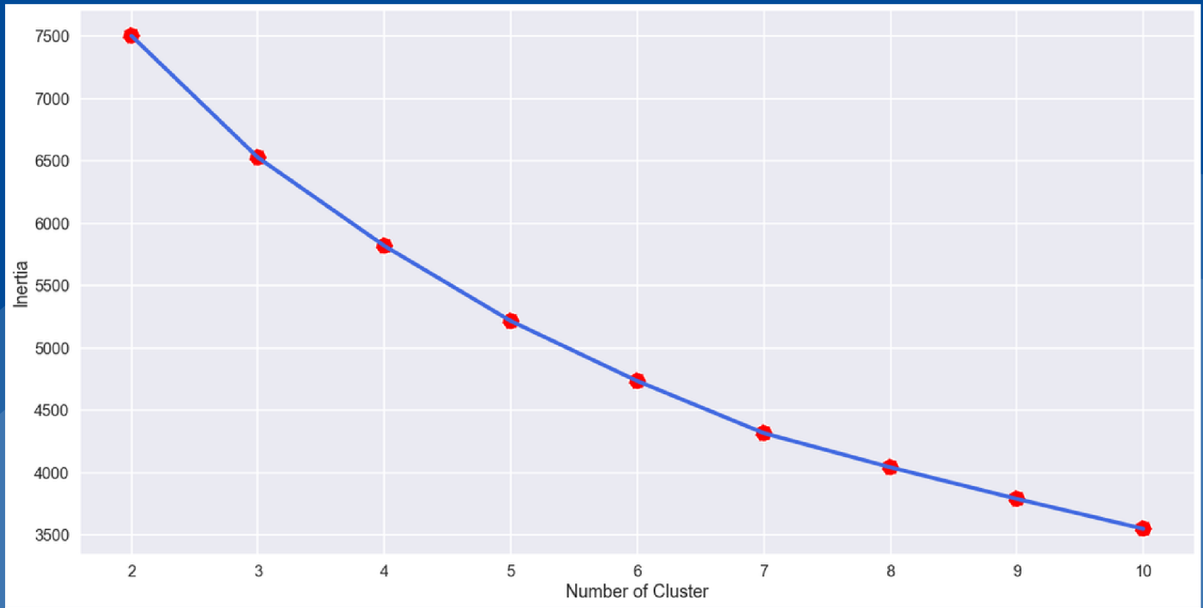
2208 rows × 5 columns



K-Means will be used, which is machine learning based on distance, so that the scaling data used is standardized.

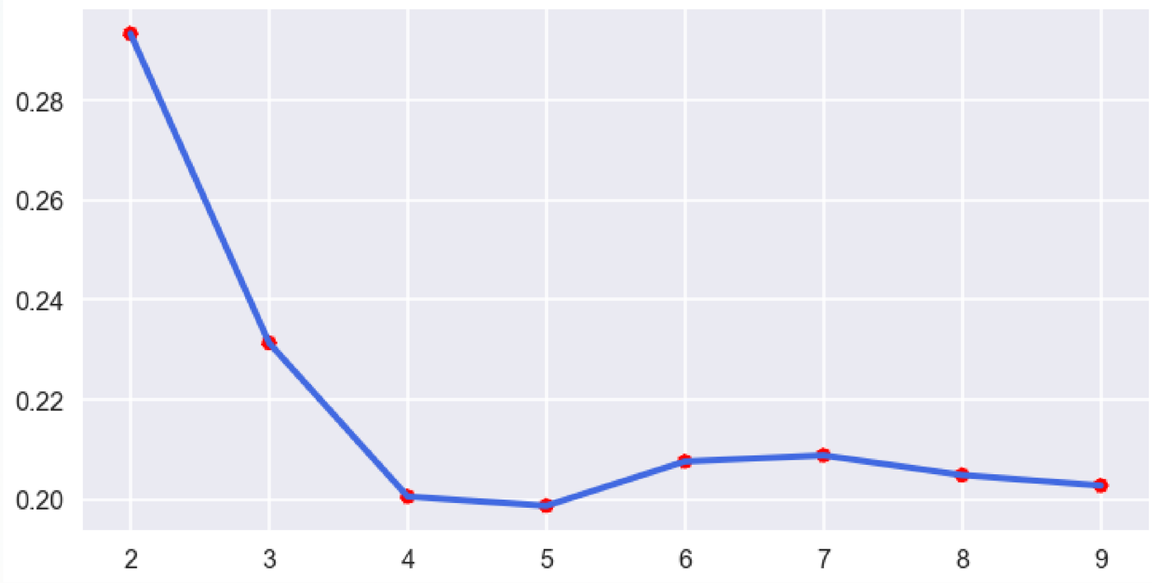
Modeling

Elbow Methods



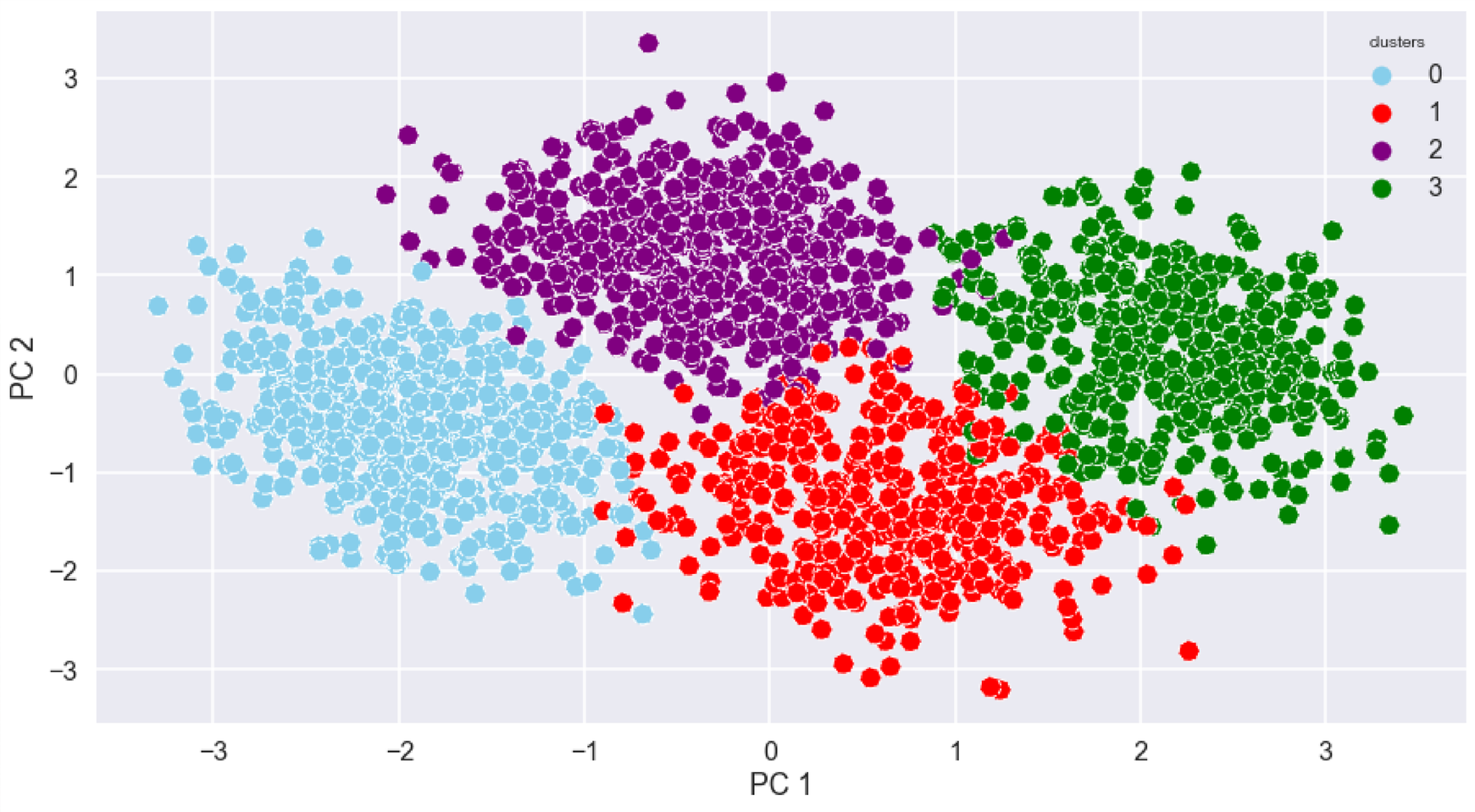
From the graph obtained, it can be seen that the optimal **number of clusters is 4.** However, it is necessary to evaluate the results of the cluster. Therefore, a validation of cluster results with the Sihouette Score is carried out.

Silhouette Score



Based on the silhouette score graph, it can be seen that **the optimal number of clusters is 4**. So the results of elbow methods show the same result as validation.

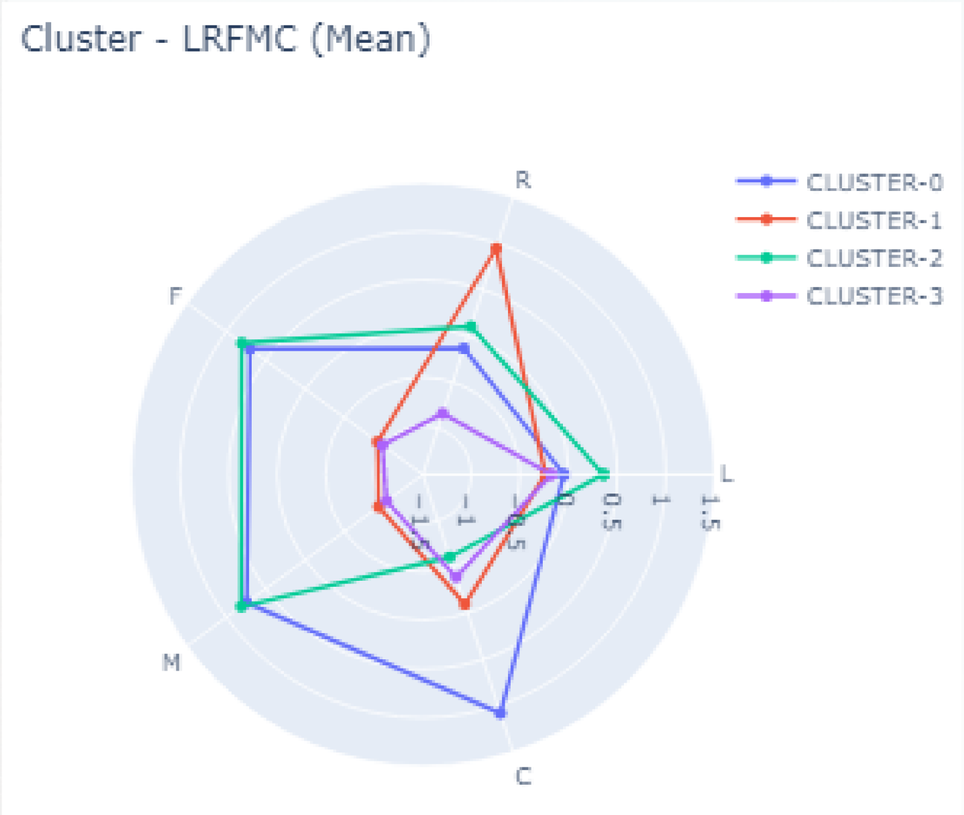
Visualization with PCA



From the results of the scatterplot above, it can be said that the **number of clusters equal to 4 is the right number of clusters.** where it can be seen that there is a fairly clear segmentation between the clusters.

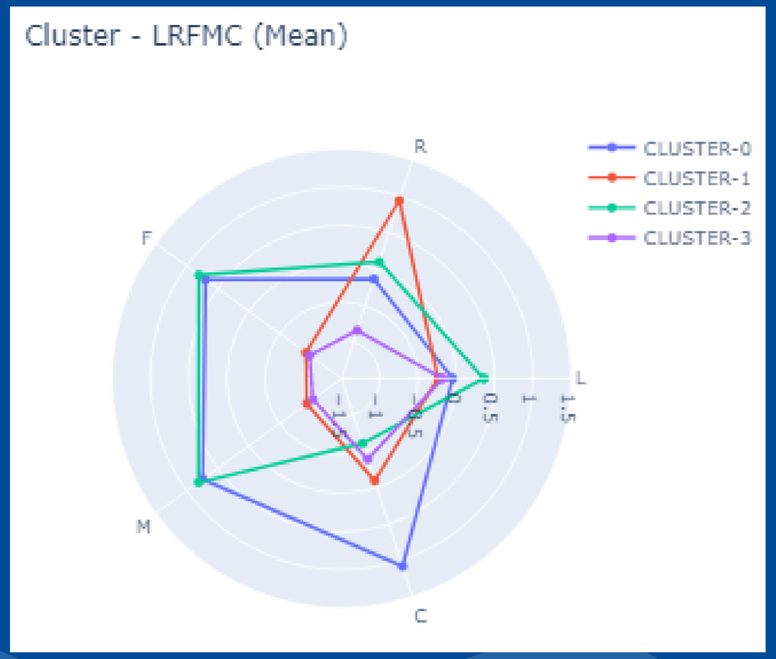
Cluster Interpretation

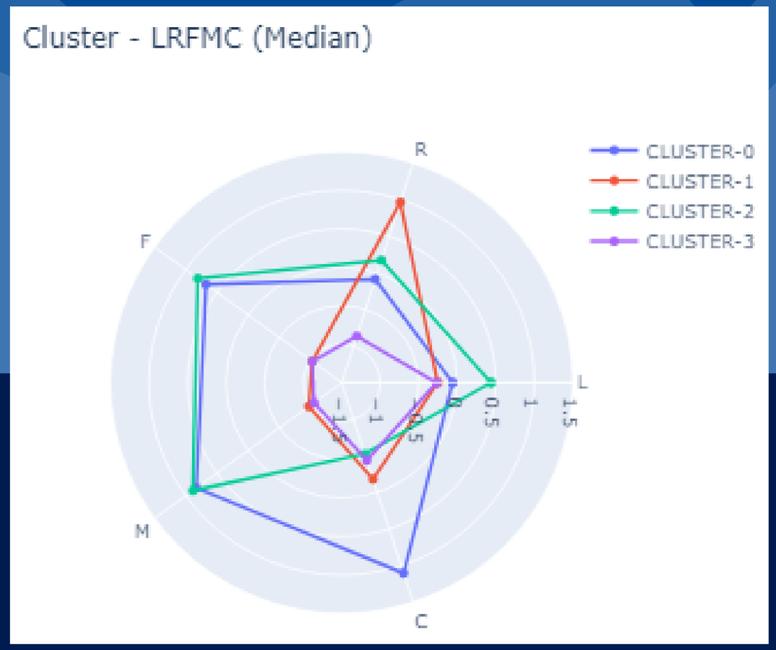
Cluster Interpretation



From the above information, it can be argued that:

- Cluster 0 : High C, F, M ; Middle L, R
- Cluster 1 : Middle C; Low L, F, M, R
- Cluster 2 : High L, F, M ; Middle R; Low C
- Cluster 3 : High R; LOW L, F, M, C





Cluster 0

Cluster 0 is a group of customers with an average age of 65 years, **an average of many transactions or purchases, and also a cluster with a fairly high total spent.** The average customer is neither a new nor an old customer, but rather somewhere in the middle. In addition to the recency, this cluster has an averagely short purchase distance from the last purchase. These clusters are referred to as **"Potential Loyalists".**

Cluster 1

Cluster 1 is a cluster with an average age of 51 years, **average newcomers, small average transactions, a small total spent**, and making purchases from the **last purchase over a long period of time.** This cluster can be referred to as **"Hibernating"**.

Cluster 2

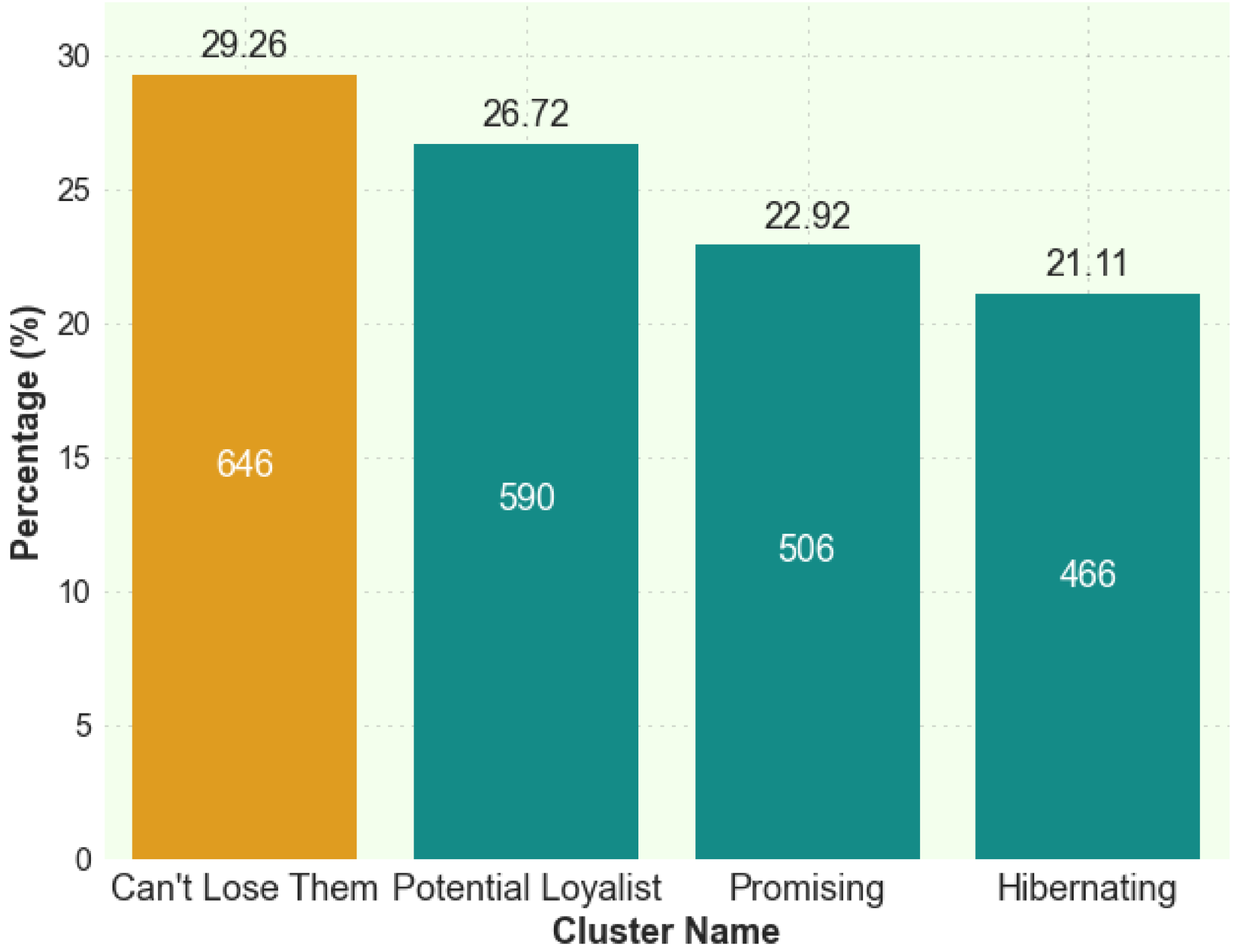
Cluster 2 is a cluster with an average age of 46 years. This cluster is the cluster with the **most average purchases and total spent**. From the perspective of recency, this cluster includes those that make purchases from the last purchase at a considerable distance. In addition, the average cluster consists of old customers. These clusters are referred to as **"Can't Lose Them".**

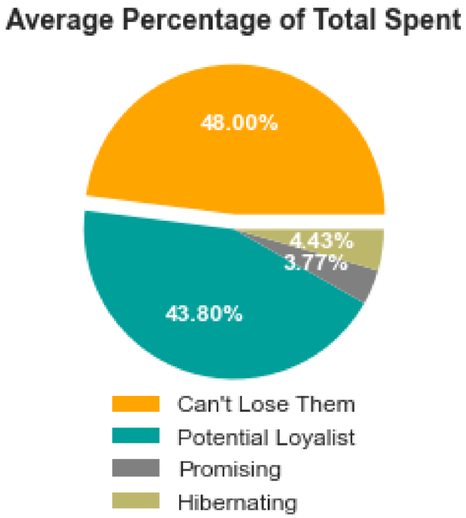
Cluster 3

Cluster 3 is a cluster with an average age of 48 years and consists of an average number of new customers. This cluster is the cluster with **the smallest average recency,** meaning that this cluster makes purchases from the last purchase with an average adjacent time distance. Nevertheless, this cluster is the one with **the least number of purchases and the lowest total spent.** These clusters are referred to as **"Promising."**

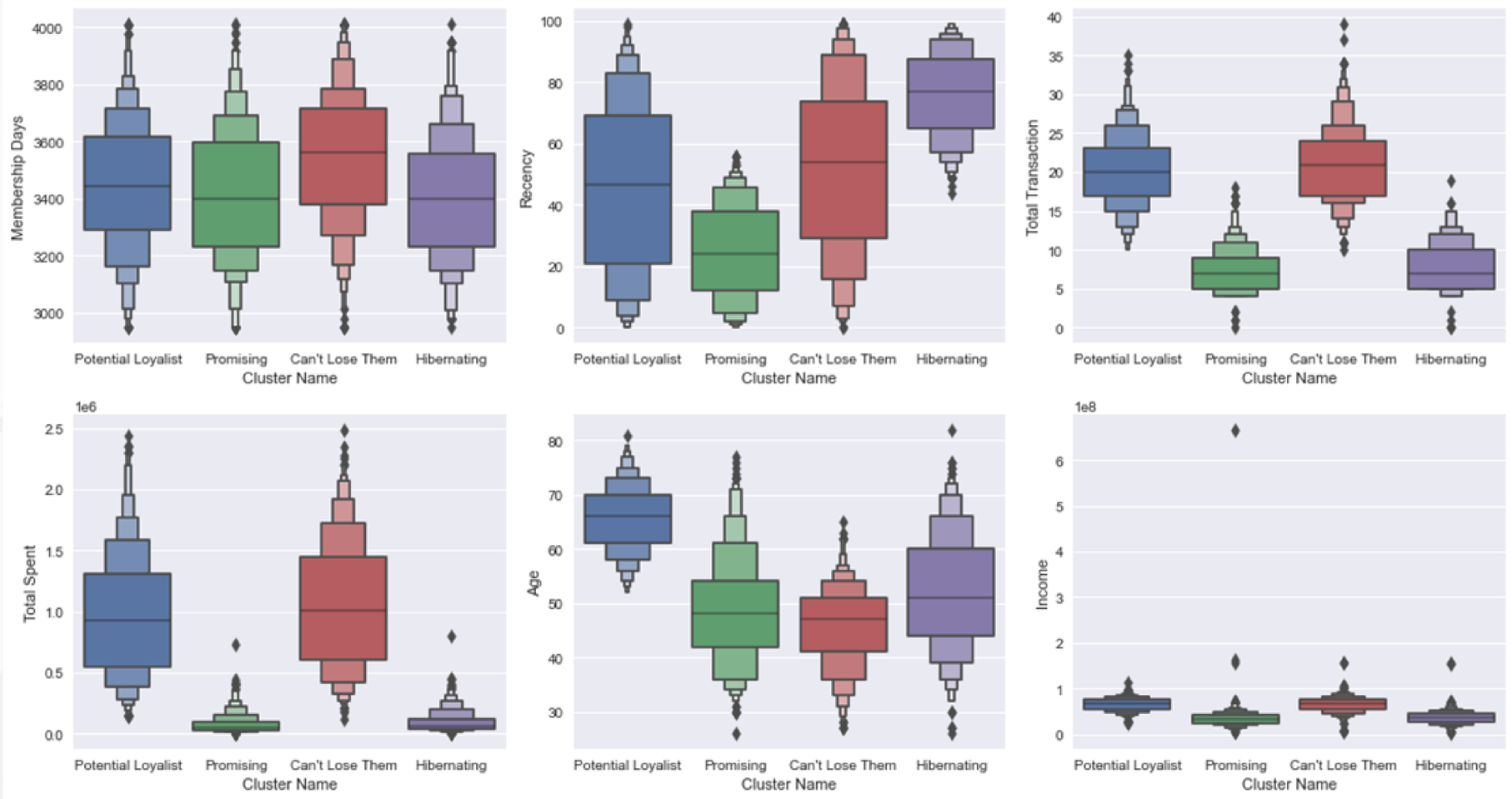
Cluster Interpretation

The Cluster Name 'Cant Lose Them'
is the cluster name with the most customers
Although there is no statistically significant difference in customer numbers between clusters,
this cluster has the highest average purchase and monetory. So, this cluster can be used as a priority cluster





Cluster Interpretation



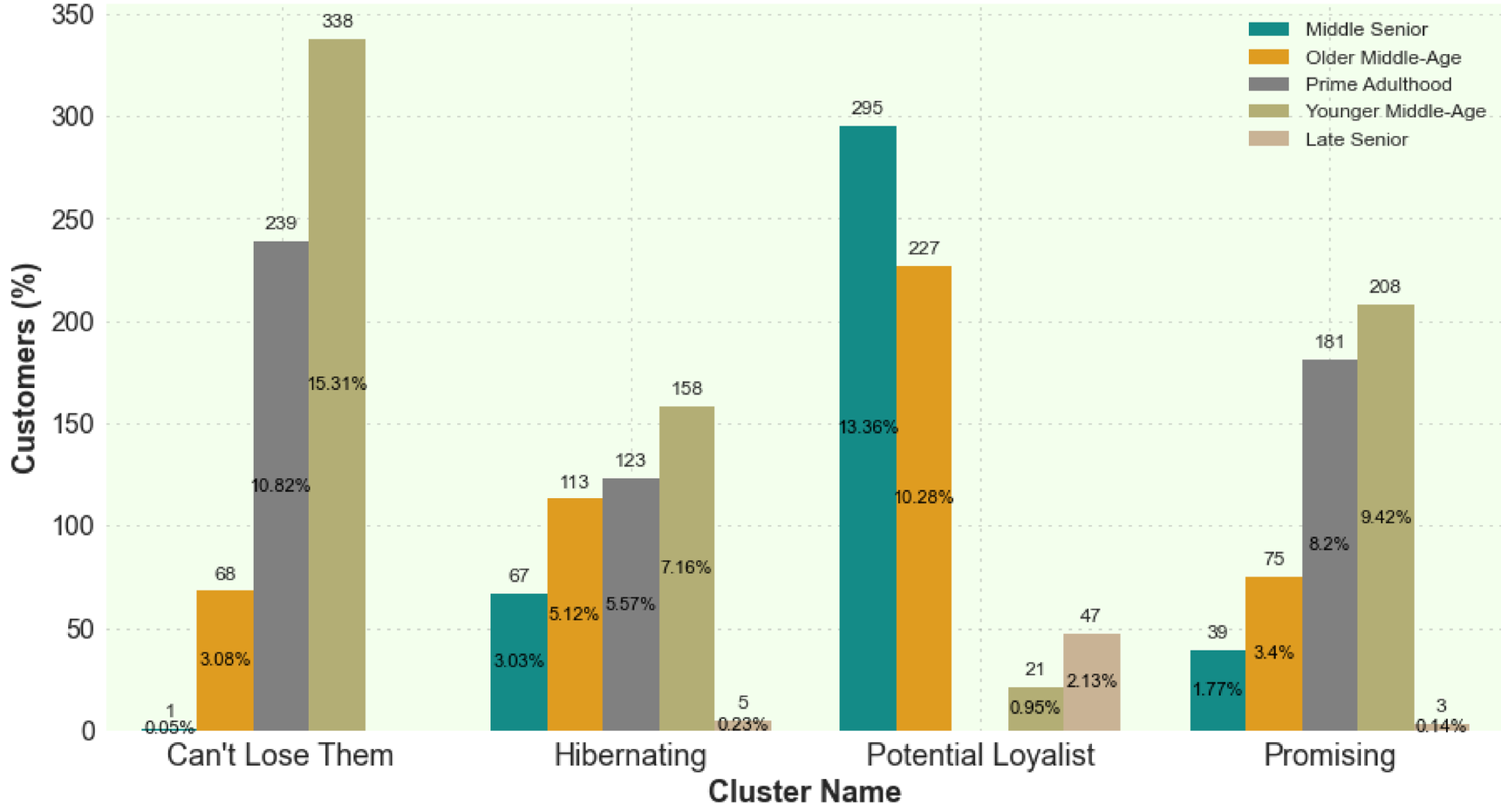
It can be seen that **the "Can't Lose Them"** cluster also has a **high average income** compared to other clusters.

Cluster Interpretation

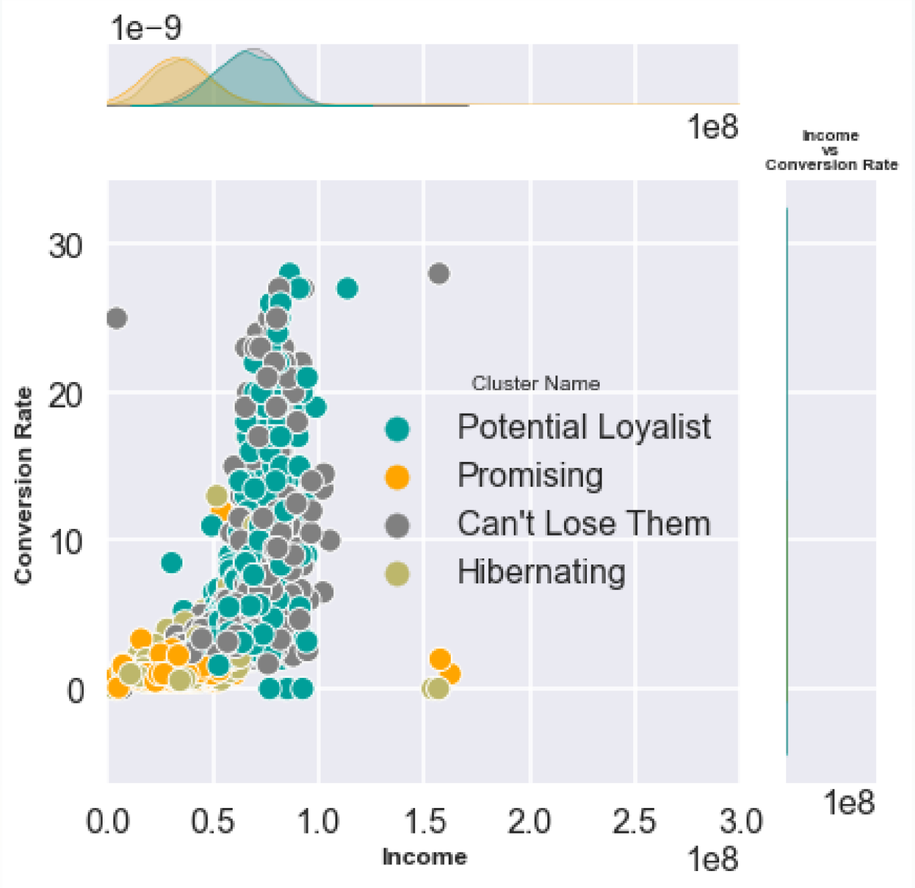
The Age Group 'Young Middle-Age'

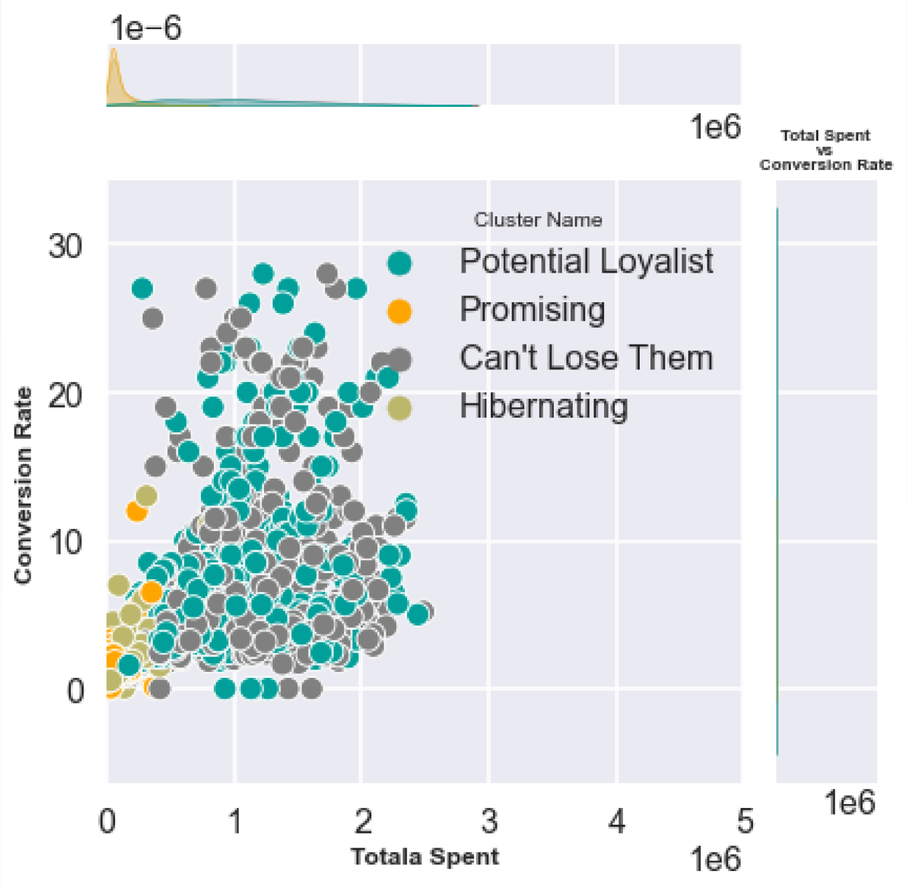
is the dominating age group

More than 30% of customers belong to this age group, and 15% of them belong to the Can't lose them cluster, which is the cluster with the highest average number of product purchases and a high average total spent.



Cluster Interpretation





It can be seen that the cluster **"Can't Lose Them and Potential Loyalists**" have a **large income and total spending.** These two clusters can be used as priority customers so as not to lose income or profits from the activities they do.

Business Strategy Recomendations

Business Strategy Recomendations

1. Loyalist Potential Cluster

Because the customers in this cluster are clusters that make quite a lot of purchases, the total amount of money spent is also quite large. So in this cluster, customers can be given loyalty programs such as discounts. Furthermore, because customers in this cluster have a short distance to the next purchase, they can offer products to this cluster in order to increase recency activities.

2. Promising Cluster

Because the customers in this cluster are clusters of new entrants and make quite small purchases, the strategy that can be applied is to provide opportunities such as offers so that these customers are interested in making purchases (brand awareness).

3. Can't Lose Them Cluster

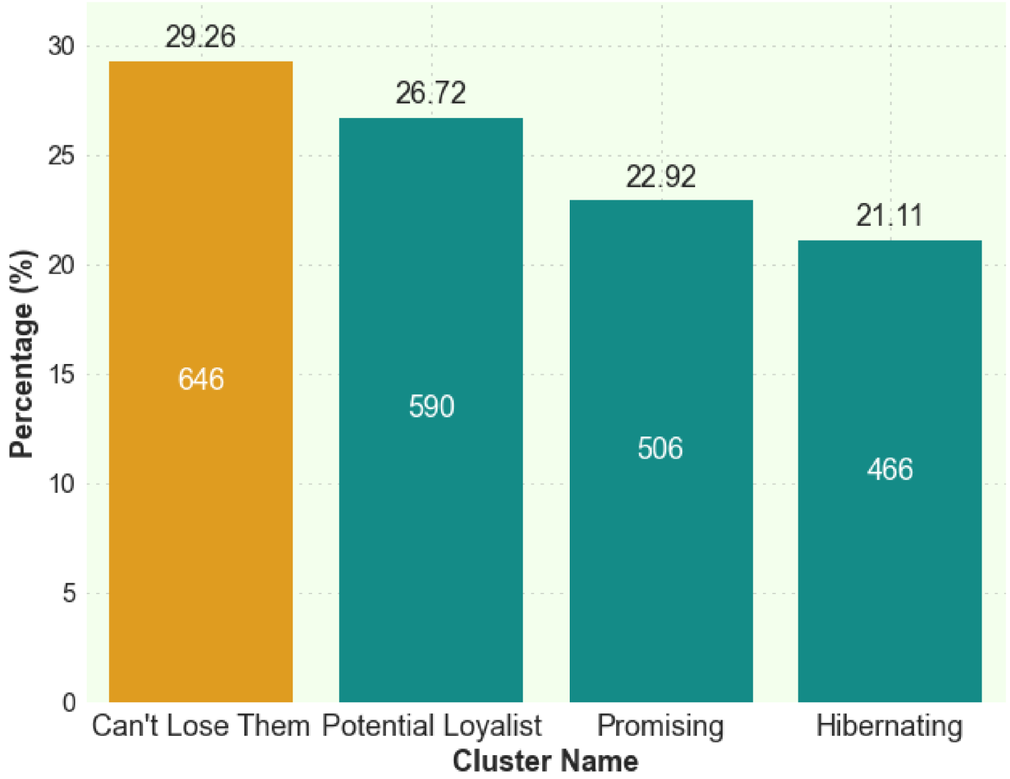
This cluster purchases the most products, and the total amount spent is also substantial. Customers in this cluster are the most potential customers and provide high profits. In addition, the number of customers in this cluster is also the largest. However, customers in this cluster make purchases after the purchase is quite far away.Therefore, the strategy that can be done is to contact and provide offers to this cluster so that it can attract them to make purchases that can be in the form of products that suit most of the ages in this cluster and also provide discounts.

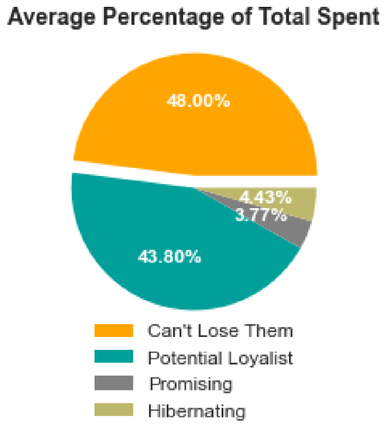
4. Hibernating Cluster

This cluster contains small purchases and total money spent. In addition, the distance from the last purchase to the next purchase is also far. The strategy that can be implemented is to provide special offers and products that are relevant to the customers in this cluster (according to the age of the customers in this cluster).

Business Strategy Recomendations

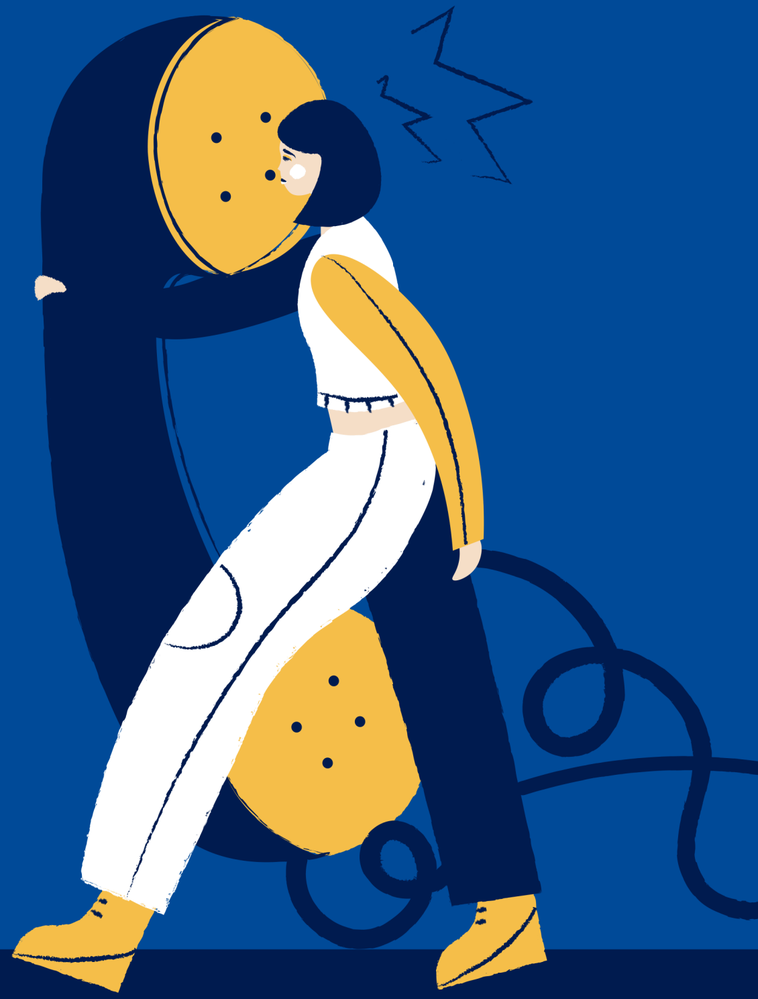
The Cluster Name 'Cant Lose Them'
is the cluster name with the most customers
Although there is no statistically significant difference in customer numbers between clusters,
this cluster has the highest average purchase and monetory. So, this cluster can be used as a priority cluster





Two clusters that can be **prioritized** are **"potential loyalists" and "can't lose them"**. These two clusters make t**he most purchases and spend the most money in total.** With the total customers of these two clusters being more than **55% of the total customers and the total spent being more than 80%,** these two clusters can be used as the main focus. The potential impact of focusing on these two clusters is that we will obtain a **GMV (Gross merchandise value)**
of IDR 566 million from the Potential Loyalist cluster and **IDR 679 million from the Can't Lose Them cluster.**

Thank You



Email
jonisyofian14@gmail.com

LinkedIn
https://www.linkedin.com/in/jonisyofian/