Predict Customer Personality to boost marketing campaign by using Machine Learning

Created by:
Nur Cahyanti
nurcahyanti3152@gmail.com
https://www.linkedin.com/in/nur-cahyanti/
https://github.com/nrchyanti

"Bachelor of Mathematics from Bandung Institute of Technology who is interested in data field. Graduated from data science bootcamp with excellent grade predicate. Love to solve problems related to data analysis and data science using SQL, Python, and Tableau."

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Overview

"A company can grow rapidly when it knows the behavior of its customer personality. Because by knowing customer behavior, companies can provide better services and benefits to customers who have the potential to become loyal customers. By processing historical marketing campaign data, companies can improve performance and target the right customers so they can transact on the company's platform. From this data insight, our focus is to create a cluster prediction model to make it easier for companies to make decisions."

Conversion Rate Analysis Based on Income, Spending and Age

Rakamin Academy

Feature Engineering

- **Age** column obtained by subtracting the current year from the Year_Birth column.
- **GroupAge** column obtained by categorizing Age into 4 groups. Babies, teens, young adults, and adults.
- **child** column obtained by adding up the Kidhome and Teenhome columns.
- **is_parent** column is obtained by categorizing customers who are married and have children with a value of 'Yes', customers who are married but do not have children with a value of 'No', and customers who are not married with a value of 'Single'.
- **total_spent** is the total purchase price obtained by adding up the MntCoke, MntFruits, MntMeatProducts, MntFishProducts, MntGoldProds, and MntSweetProducts columns.

Click here for jupyter notebook code

Conversion Rate Analysis Based on Income, Spending and Age

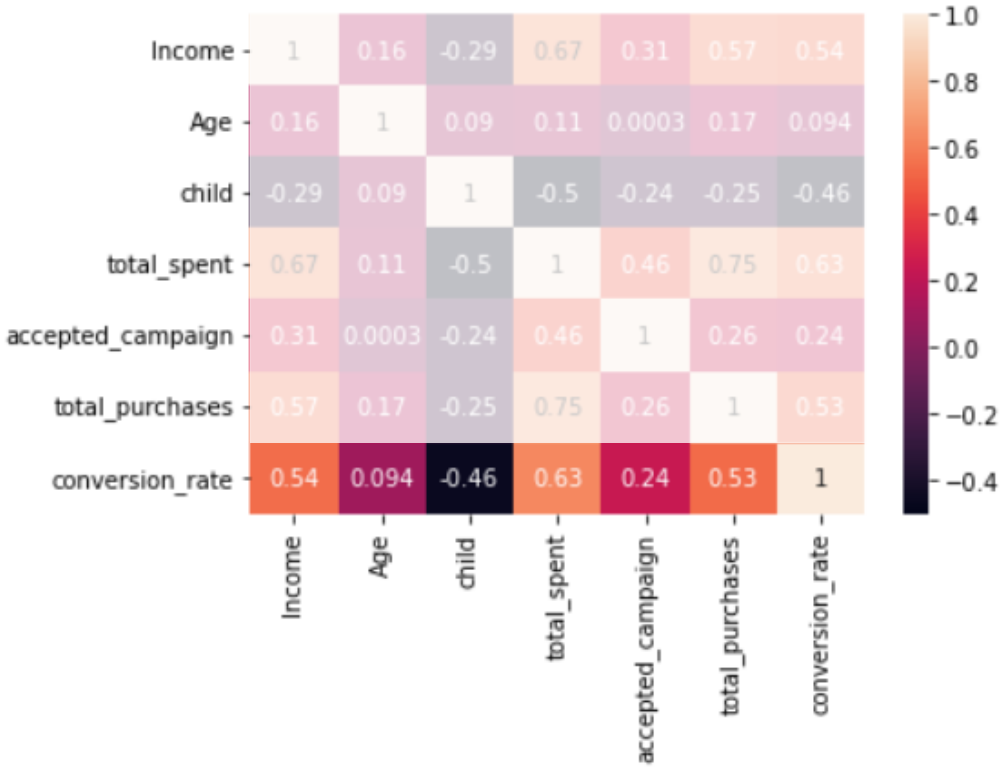
Rakamin Academy

Feature Engineering

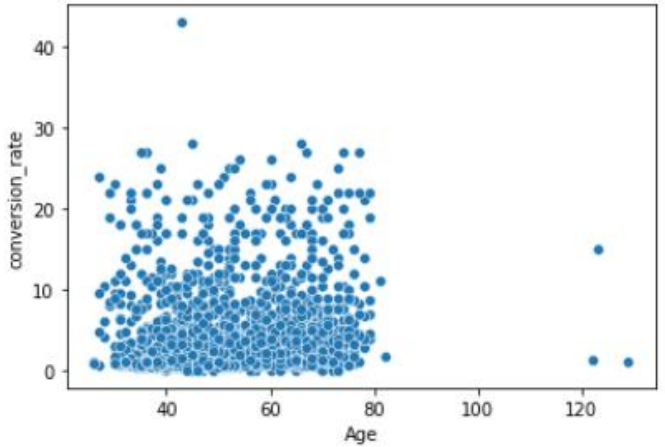
- **accepted_campaign** is the total accepted campaigns obtained by adding up the AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, and AcceptedCmp2 columns.
- **total_purchases** is the total product purchases obtained by adding up the NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, and NumStorePurchases columns.
- **conversion_rate** is obtained from the comparison of the total_purchases column with the NumWebVisitsMonth column, to analyze customer behavior towards the product.

Click here for jupyter notebook code

Conversion Rate Analysis Based on Income, Spending and Age

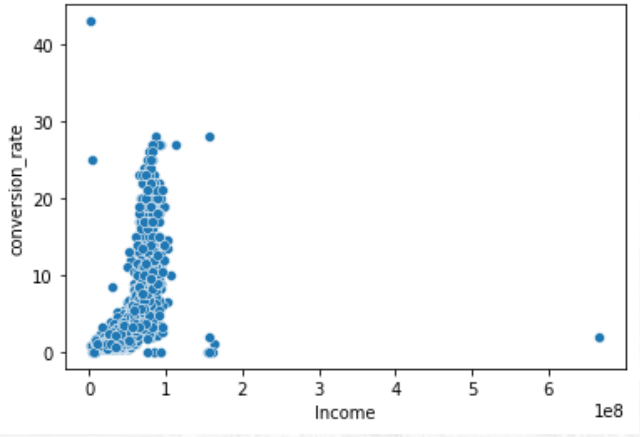


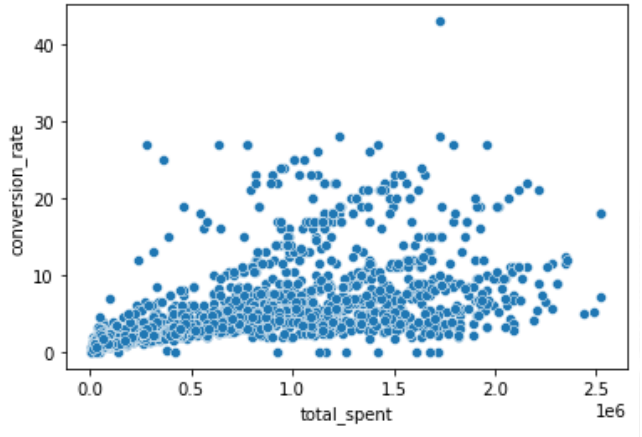
The correlation between numeric columns can be seen on the heatmap. We focus on seeing the correlation of conversion_rate with other columns.The correlation of conversion_rate with age is the weakest. This means that age and conversion_rate do not affect each other.

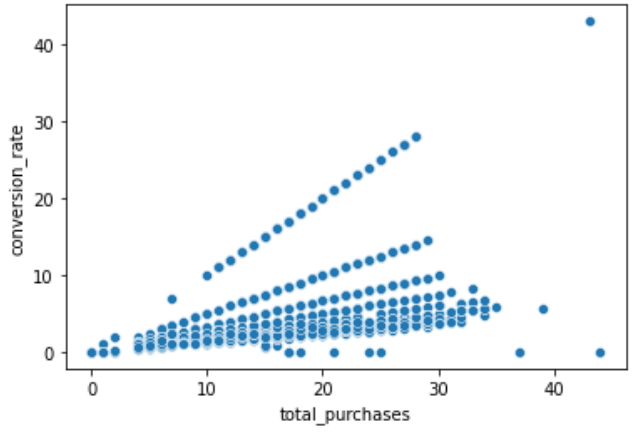


Click here for jupyter notebook code

Conversion Rate Analysis Based on Income, Spending and Age



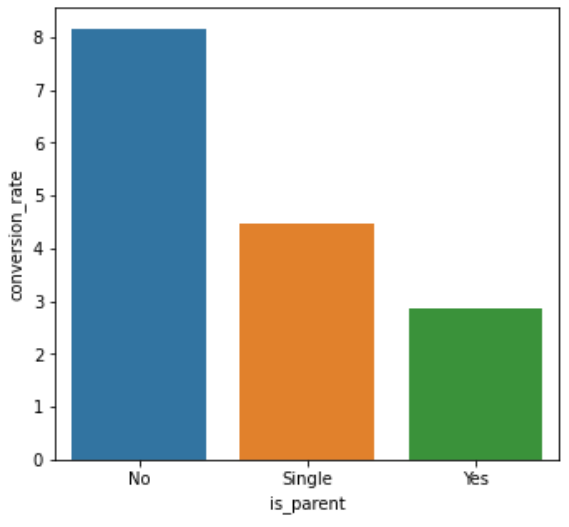


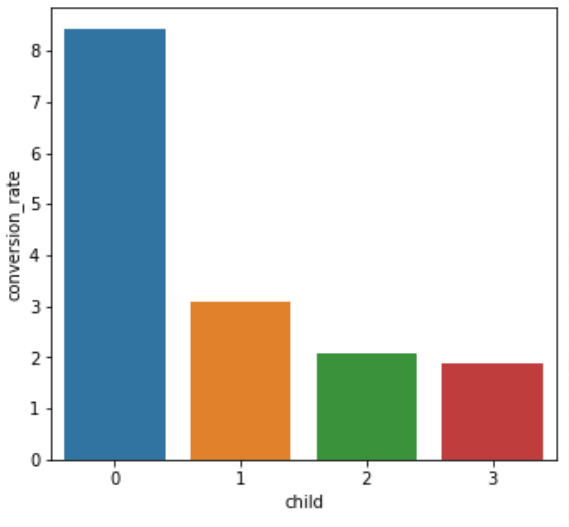


Other features that correlate quite strongly with conversion_rate are Income, total_spent, and total_purcases. All of them are positively correlated, which means that the larger the value of these columns, the higher the conversion_rate. The distribution of Income, total_spent, and total_purcases data compared to conversion_rate can be seen in the scatter plot above.

Click here for jupyter notebook code

Conversion Rate Analysis Based on Income, Spending and Age

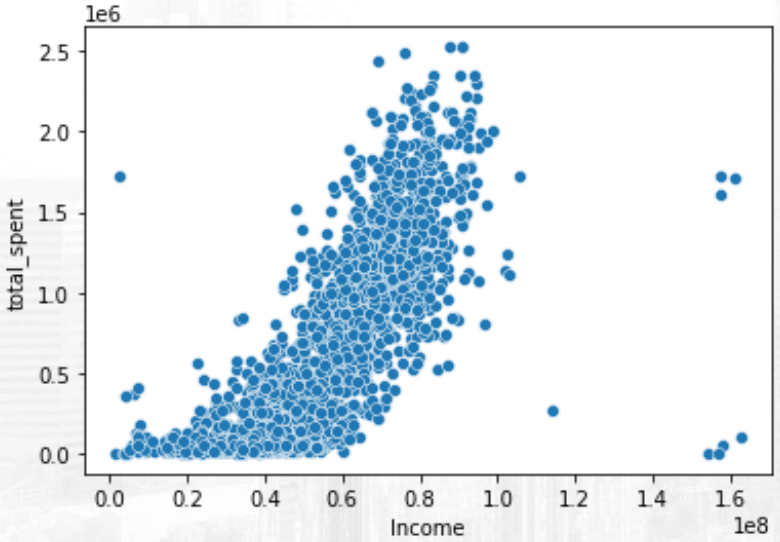




If we look at the type of customer whether they are parents or not, it turns out that **customers who are married and have children have a lower average conversion_rate** compared to customers who are not married or who are married but do not have children. This is consistent with the data that customers who do not have children have the highest average conversion_rate compared to customers who have 1 or more children.

Click here for jupyter notebook code

Conversion Rate Analysis Based on Income, Spending and Age



The greater the income, there is a tendency to have more expenses and have a larger total expenditure on our platform.

Recommendation

We can choose customers who have income of 60 million and above for marketing campaigns by targeting many people who have CVR above 10%.

But what about the other groups, can we do something about them?

Click here for jupyter notebook code

Data Cleaning & Preprocessing

```
data.isna().sum()
✓ 0.1s
```

Income	24
Age	0
GroupAge	0
child	0
total_spent	0
accepted_campaign	0
total_purchases	0
conversion_rate	0
is_parent	0

dtype: int64

```
data = data.dropna()
✓ 0.2s
```

```
data.isna().sum().sum()
✓ 0.2s
0
```

There are **24 missing values** in the Income column. Because the number is relatively small compared to the entire data, the data rows that have this missing value will be deleted.

```
df.duplicated().sum()
✓ 0.2s
0
```

There are **no duplicate rows** in the data.

Click here for jupyter notebook code

Data Cleaning & Preprocessing

Rakamin Academy

Feature Encoding

Performed on the GroupAge and is_parent columns using the **label encoding**. Especially for is_parent column, the 'No' and 'Single' are given the 0 value. While the 'Yes' data is 1.

Handling Outliers

All columns are given z-score treatment to remove outliers. The z-score value used is less than 3.

Data Used

Columns 'Income', 'Age', 'GroupAge', 'child', 'total_spent', 'accepted_campaign', 'total_purchases', 'conversion_rate', 'is_parent'

Standard Scaler

Columns 'Income', 'total_spent', and 'conversion_rate' to be used for clustering

Click here for jupyter notebook code

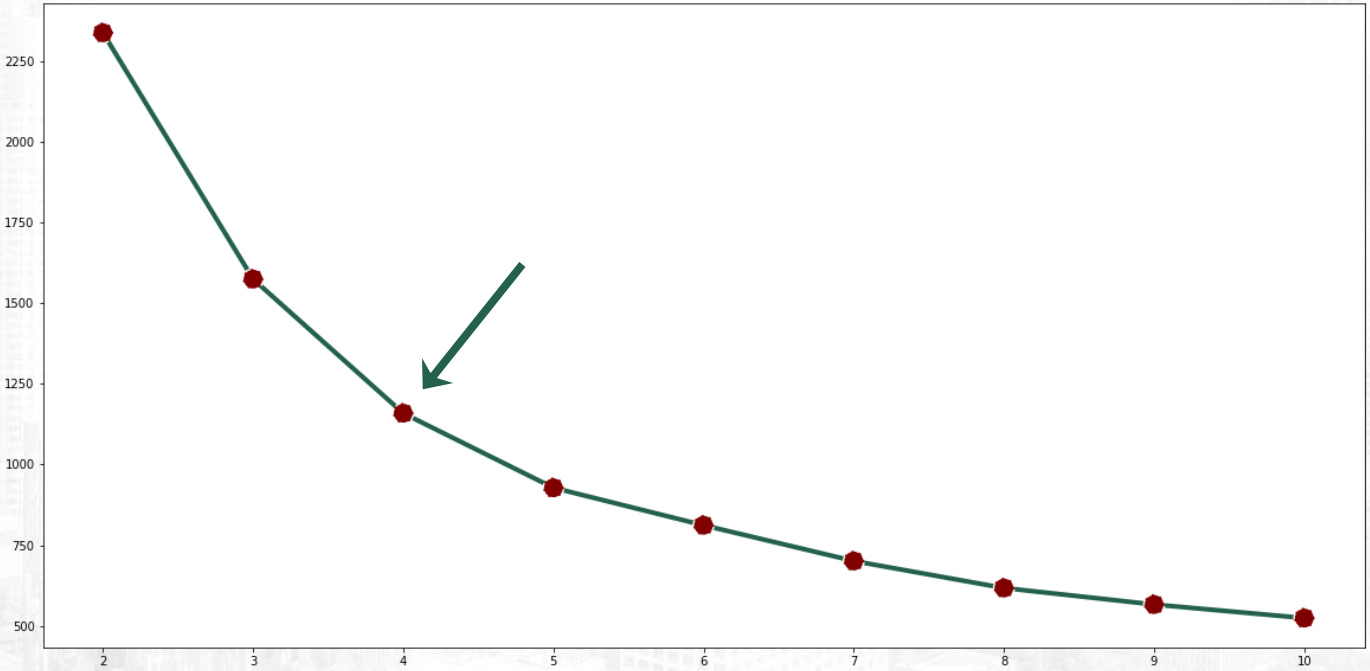
Data Snippets Ready to Modelling

	Income	Age	GroupAge	child	total_spent	accepted_campaign	total_purchases	conversion_rate	is_parent
0	58138000.0	65	3	0	1617000	0	25	3.571429	0
1	46344000.0	68	3	2	27000	0	6	1.200000	0
2	71613000.0	57	3	0	776000	0	21	5.250000	0
3	26646000.0	38	3	1	53000	0	8	1.333333	0
4	58293000.0	41	3	1	422000	0	19	3.800000	1
...	...	...	...	...	...	...	...	...	...
2235	61223000.0	55	3	1	1341000	0	18	3.600000	1
2236	64014000.0	76	3	3	444000	1	22	3.142857	0
2237	56981000.0	41	3	0	1241000	1	19	3.166667	0
2238	69245000.0	66	3	1	843000	0	23	7.666667	0
2239	52869000.0	68	3	2	172000	0	11	1.571429	1

2068 rows × 9 columns

Click here for jupyter notebook code

Data Modeling



The results curve for the **elbow method** is shown in the image above. It can be seen that the slope begins to decrease from 4 to 5. So that **4 clusters** will be selected to perform the k-means clustering model.

Click here for jupyter notebook code

Data Modeling

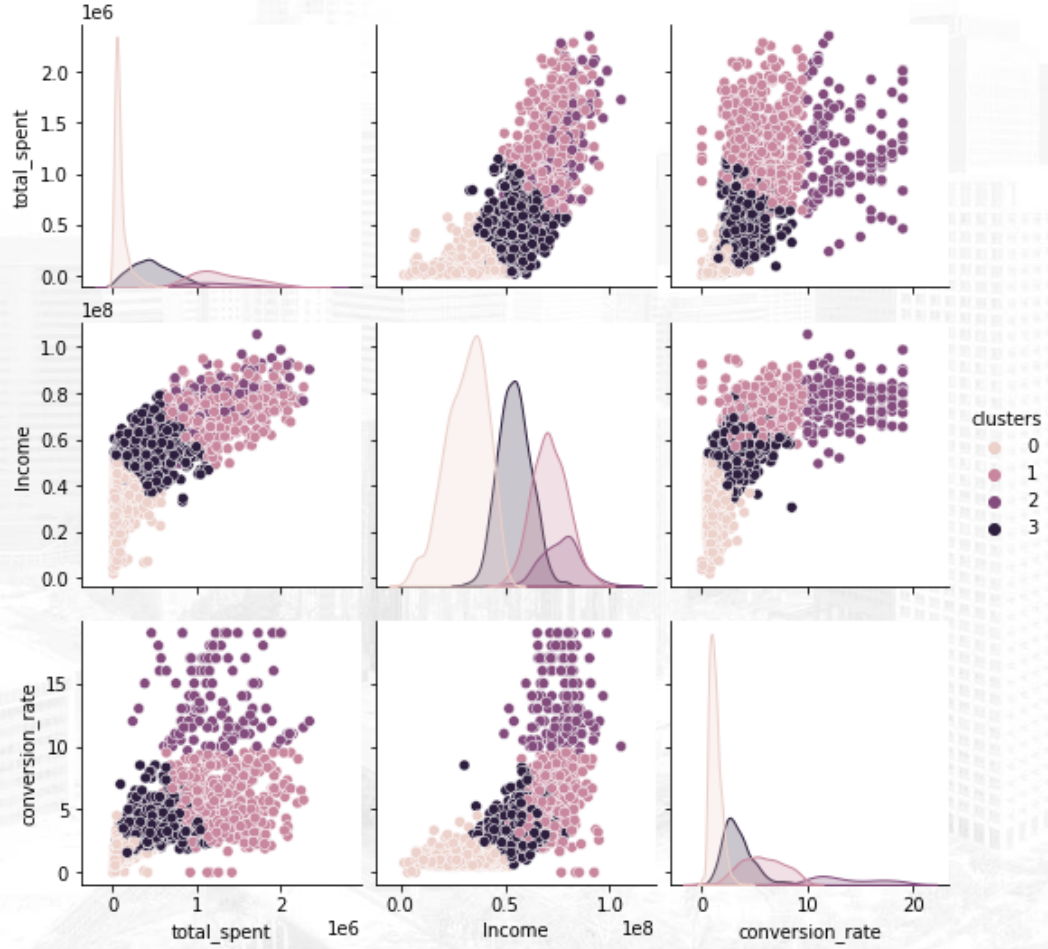
The optimal number of clusters is checked again using a silhouette score. Below you can see a snippet of data that has been grouped into 4 clusters. With a new column, namely the clusters column.

```
For n_clusters = 2, silhouette score is 0.5553597812501511)
For n_clusters = 3, silhouette score is 0.5043305700155443)
For n_clusters = 4, silhouette score is 0.4319634422899811)
For n_clusters = 5, silhouette score is 0.38989582312913834)
For n_clusters = 6, silhouette score is 0.3540694264011356)
For n_clusters = 7, silhouette score is 0.3704987069062408)
For n_clusters = 8, silhouette score is 0.3539720992104933)
```

	Income	Age	GroupAge	child	total_spent	accepted_campaign	total_purchases	conversion_rate	is_parent	clusters
0	58138000.0	65	3	0	1617000	0	25	3.571429	0	1
1	46344000.0	68	3	2	27000	0	6	1.200000	0	0
2	71613000.0	57	3	0	776000	0	21	5.250000	0	1
3	26646000.0	38	3	1	53000	0	8	1.333333	0	0
4	58293000.0	41	3	1	422000	0	19	3.800000	1	3

Click here for jupyter notebook code

Data Modeling

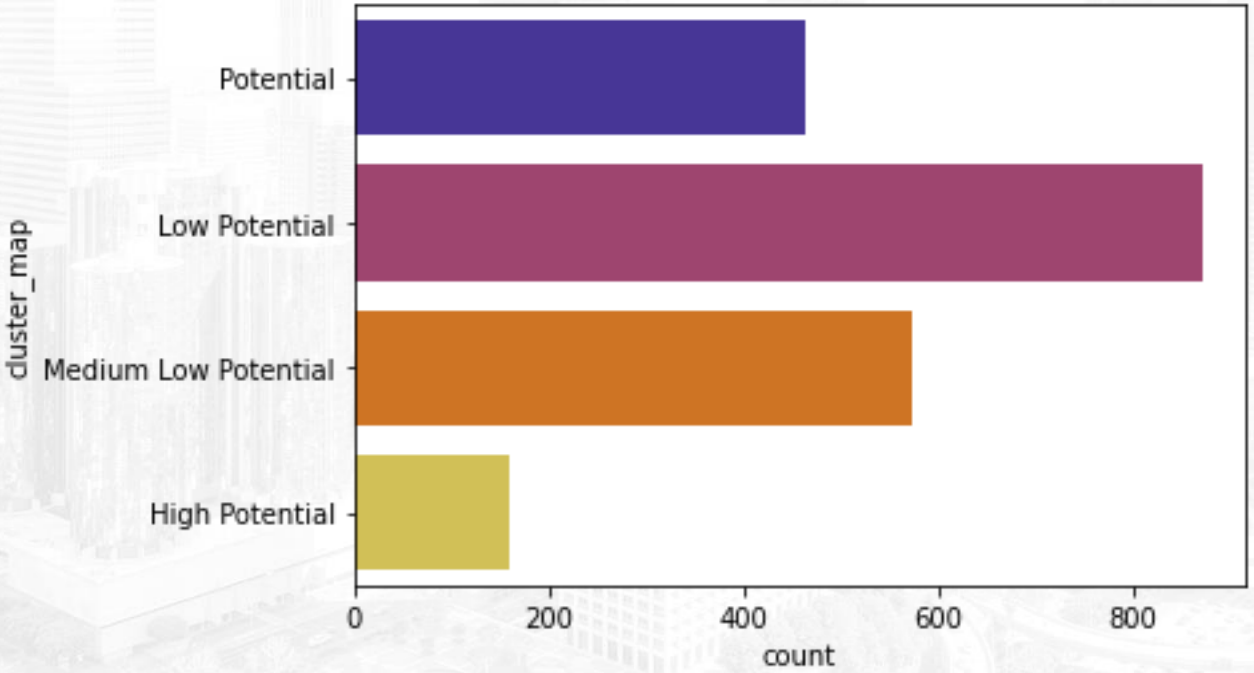


Using the **k-means algorithm**, the data is divided into 4 clusters. The distribution of the data can be seen in the scatter plot on the side. The data is divided into 4 clusters based on the features of Income, total_spent, and conversion_rate.

Click here for jupyter notebook code

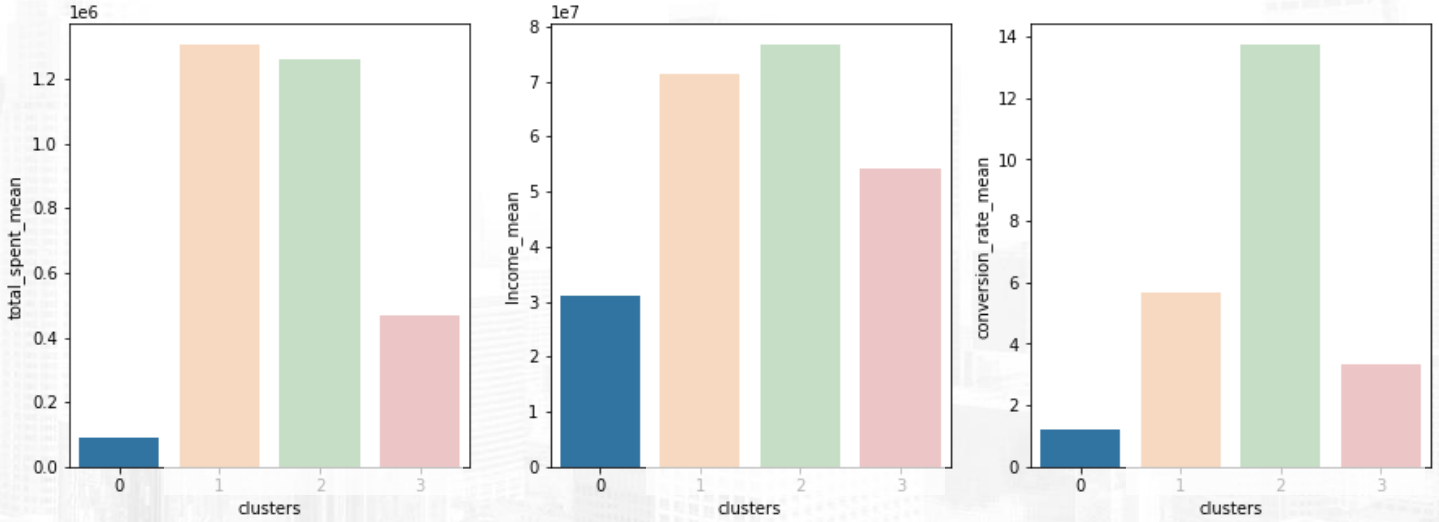
Customer Personality Analysis for Marketing Retargeting

The results of customer clustering which are divided into 4 groups based on the average total spent, average income, and average conversion rate can be seen in the following diagram.



Click here for jupyter notebook code

Customer Personality Analysis for Marketing Retargeting

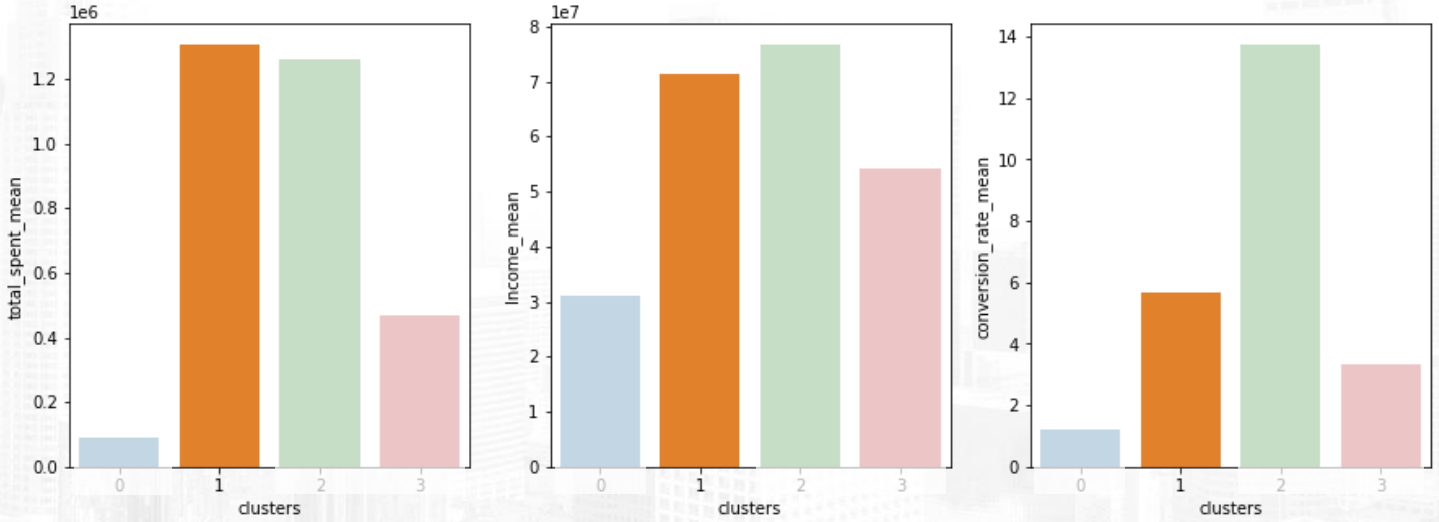


Cluster 0 → Low Potential Customer

Customers of this cluster have low income with very low total product purchases. The conversion rate is also very low, which means these customers are not interested in visiting the store and are not interested in buying. This treatment for customers can be done by providing regular advertisements to find out customer interest in the company's products.

Click here for jupyter notebook code

Customer Personality Analysis for Marketing Retargeting

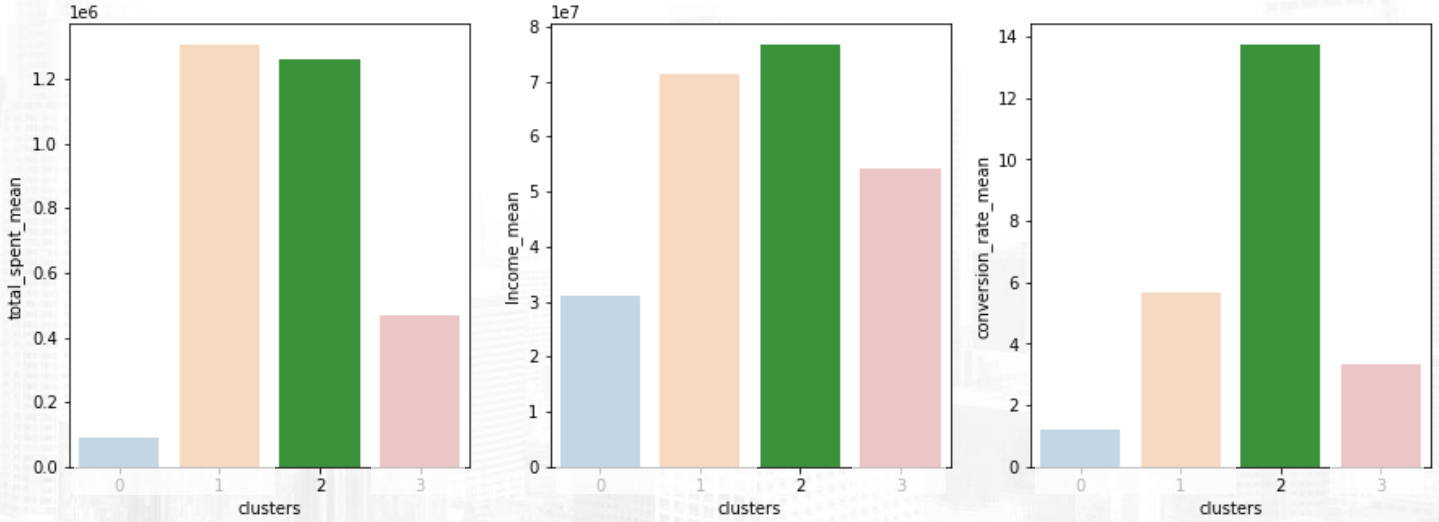


Cluster 1 → Potential Customer

This cluster customer has a fairly high income with a very high total product purchase. However, the conversion rate tends to be low, which means that these customers often make purchases without having to visit the store many times or make large purchases for each transaction. This treatment for customers can be done by providing special promos for purchases with a large minimum nominal.

Click here for jupyter notebook code

Customer Personality Analysis for Marketing Retargeting

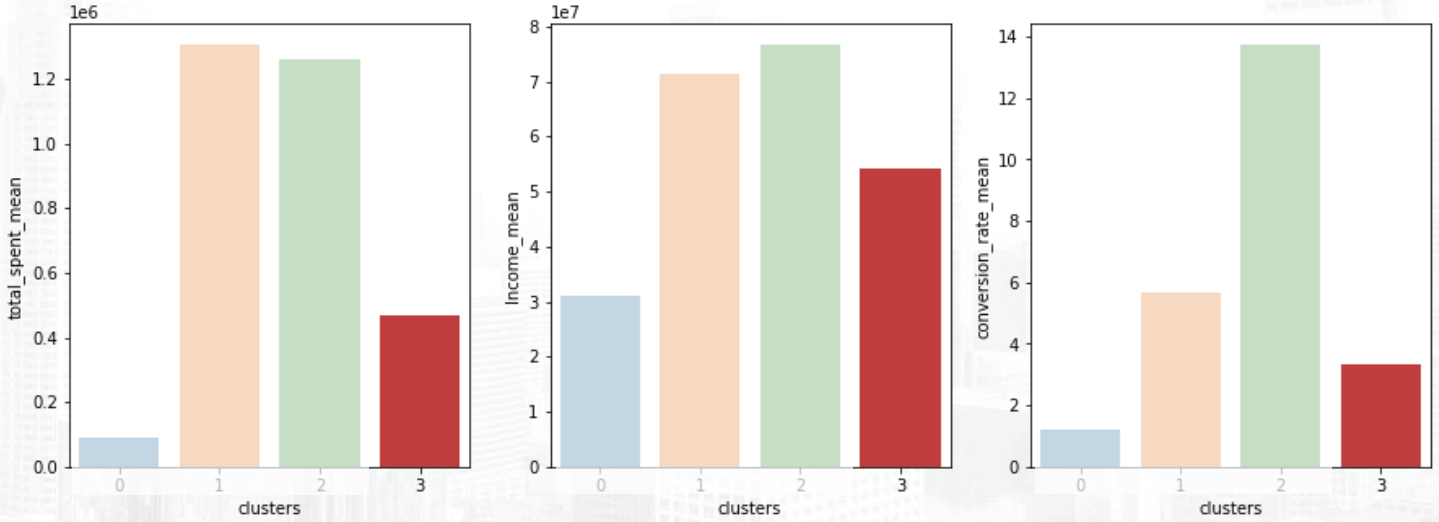


Cluster 2 → High Potential Customer

This cluster customer has the highest income followed by a very high total product purchase. The conversion rate is also very high, which means that this customer is a loyal customer and makes the most frequent transactions. This treatment for customers can be done by prioritizing customers and maintaining service quality so that customers are not disappointed. You can also be rewarded if you make a certain number of transactions.

Click here for jupyter notebook code

Customer Personality Analysis for Marketing Retargeting



Cluster 3 → Medium Low Potential Customer

Customers of this cluster have sufficient income with low total product purchases. The conversion rate is also low, which means these customers are not very interested in buying the product. This treatment for customers can be done by providing regular advertisements and providing attractive promos for items that are of interest to this cluster customer.

Click here for jupyter notebook code