Rakamin Academy X KALBE Nutritionals

[Task 5] Final Project Data Modelling

Kalbe Nutritionals Data Scientist Project Based Internship Program

Presented by
Dicky Maulana Rozi

Dicky Maulana Rozi

Hello,

I am a fresh graduate from the Institut Teknologi Sepuluh Nopember with a degree in Informatics Engineering. I am a passionate data enthusiast who dedicates my time to learning about data science, particularly data analysis, on a daily basis.

Rakamin Academy

My Experience

- IT Support Intern
  At Badan Pertanahan Nasional Surabaya
  In September 2019 - December 2019
- UI/UX Designer Intern
  At SKILVUL
  In August 2021 - January 2022
- Bootcamp Data Science
  At Rakamin Academy
  In January 2023 - June 2023

I am a Data Scientist at Kalbe Nutritionals and currently working on a new project involving both the inventory and marketing teams.

From the inventory team, I have been tasked with predicting the quantity of products sold for the entire Kalbe product range. The goal of this project is to estimate the quantity of products that will be sold on a daily basis. This prediction is essential for the inventory team to ensure they maintain sufficient daily stock levels.

From the marketing team, I have been asked to create customer clusters or segments based on various criteria. The objective of this project is to create customer segments that can be used by the marketing team to provide personalized promotions and sales treatments to different customer groups.

Challenge:

1. Exploratory data analysis in dbeaver
2. Create a dashboard in tableau
3. Create predictive models using regression
4. Create clustering models

1. Exploratory data analysis in dbeaver

Rakamin Academy

- query 1: What is the average age of the customer in terms of their marital status?

	Marital Status	avg_umur
1		31
2	Married	43
3	Single	29

- query 2: What is the average age of the customer in terms of gender?

	gender	avg_age
1	0	40
2	1	39

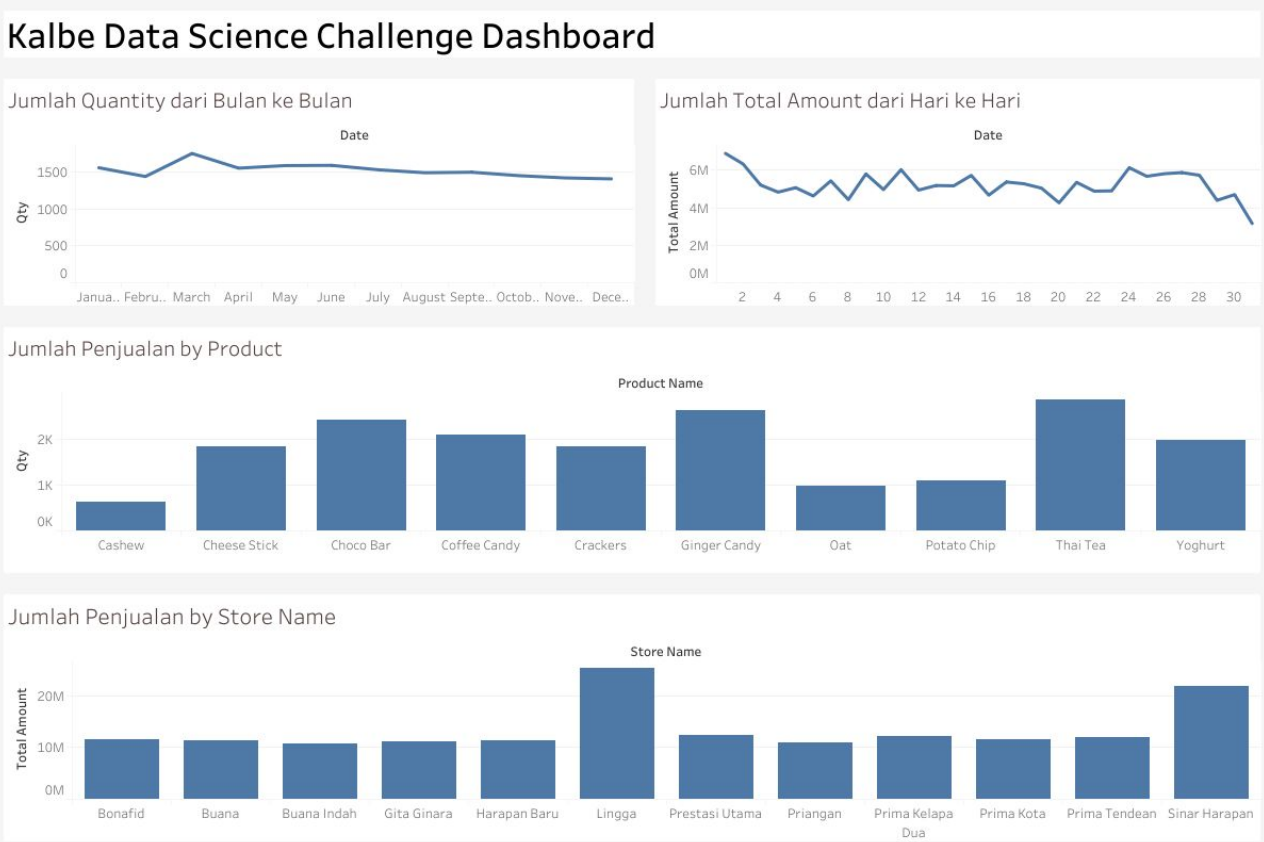
- query 3: Determine the name of the store with the highest total quantity!

	storename	total_qty
1	Lingga	2,777

- query 4: Determine the name of the best-selling product with the highest total amount!

	Product Name	total_amount
1	Cheese Stick	27,615,000

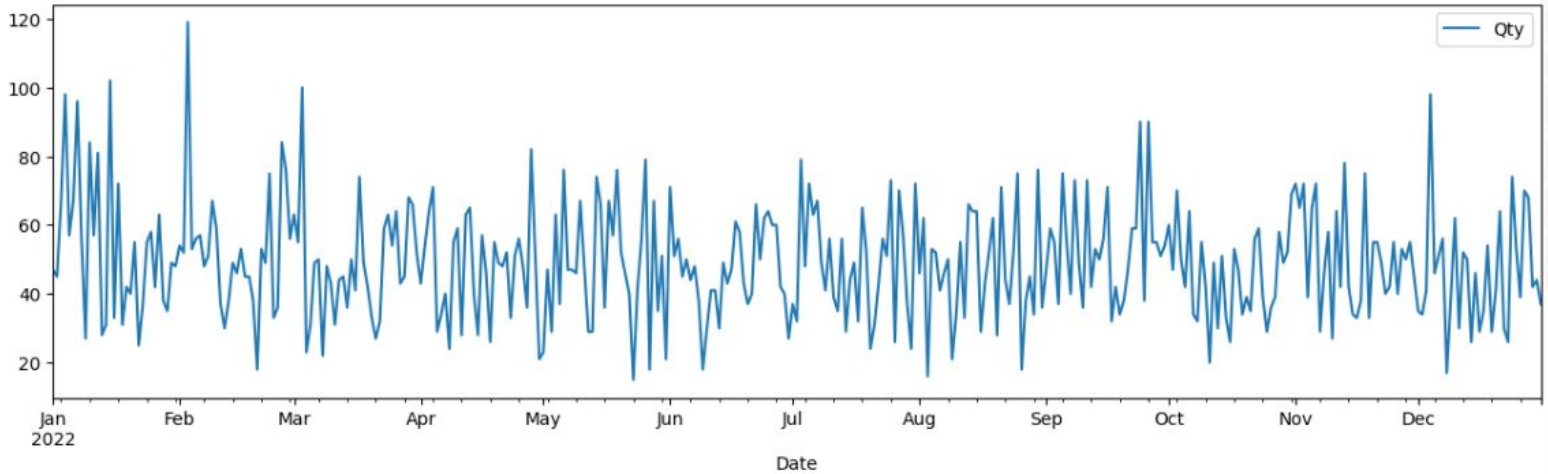
2. Create a dashboard in tableau



- Worksheet 1
  Total qty from month to month
- Worksheet 2
  Total amount from day to day
- Worksheet 3
  Total sales (qty) by product
- Worksheet 4
  Sales amount (total amount) by store name

3. Create predictive models using regression

- After cleaning the data and combining the 4 datasets, aggregate the data based on date (column 'Date') and produce a new DataFrame containing the total quantity (Qty) sold on each date.



3. Create predictive models using regression

Rakamin Academy

- To use the Arima model, check the ADF test and KPSS test and I also added an ACF plot to see the Lag

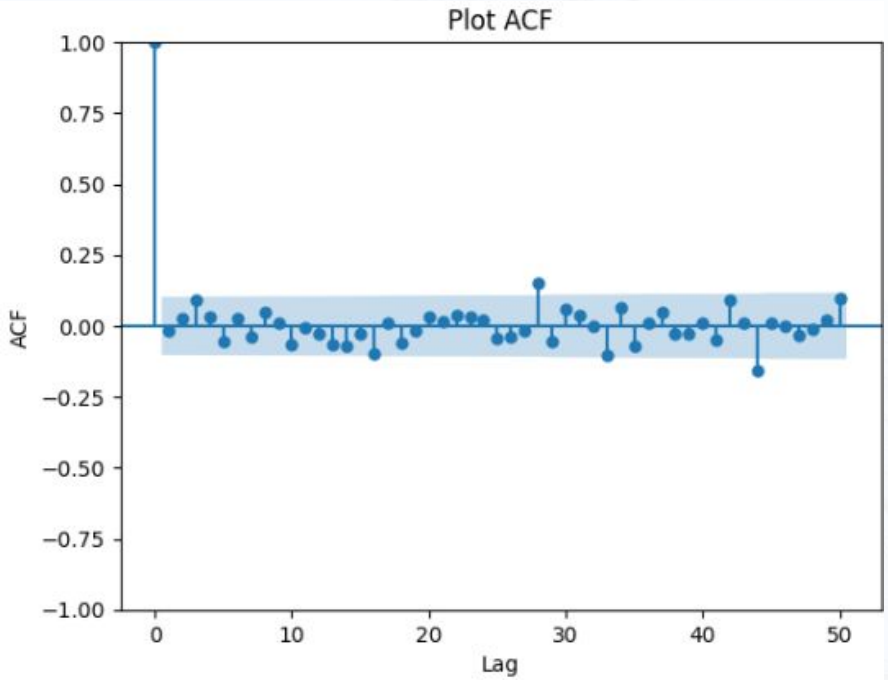
```
Hasil ADF Test:
ADF Statistic: -19.323189851289346
p-value: 0.0
Critical Values: {'1%': -3.4484434475193777, '5%': -2.869513170510808, '10%': -2.571017574266393}

Hasil KPSS Test:
KPSS Statistic: 0.2073642420009865
p-value: 0.1
Critical Values: {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739}
```

The results of the ADF Test show: The ADF Statistics is very negative (-19.32), which indicates that the data tends to be stationary. In general, the more negative the ADF statistic, the stronger the evidence that the data is stationary.

The results of the KPSS Test show: the p-value is 0.1, which is greater than the normally used significance level (eg 0.05). Means that the data can be considered as stationary at the 0.1 significance level.

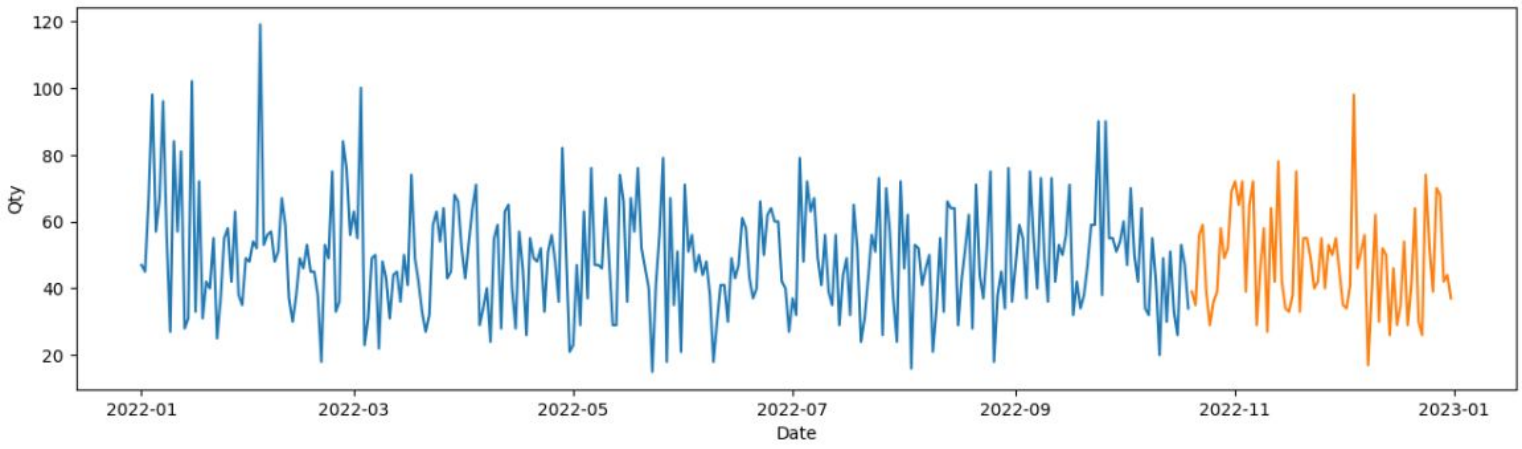
3. Create predictive models using regression

- From the results of Check Stationary Data, it can be concluded that the data tends to be stationary



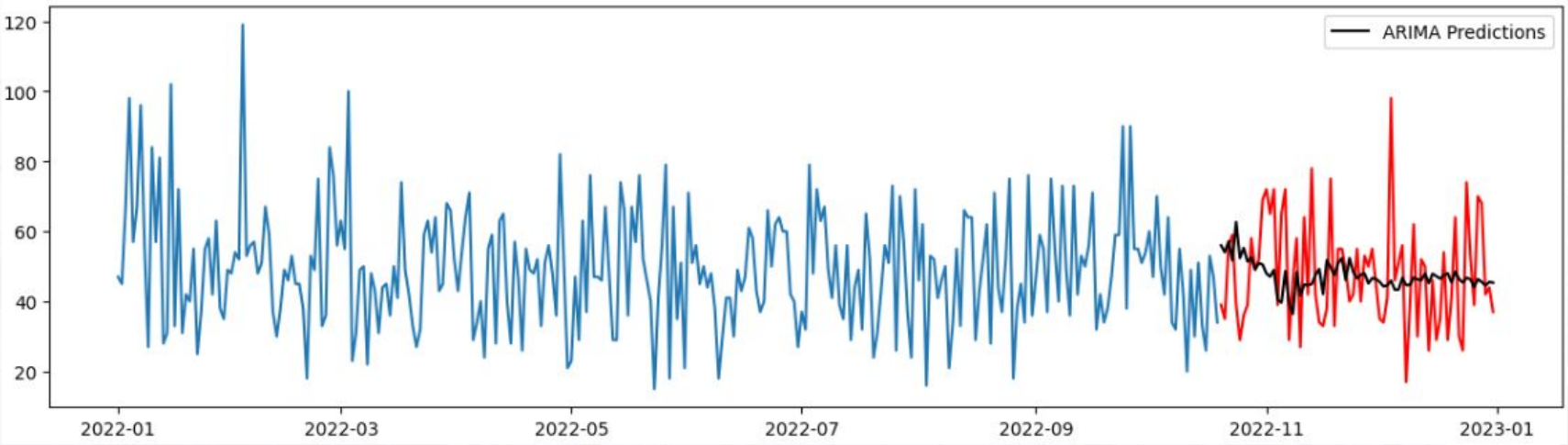
3. Create predictive models using regression

- Splitting the dataset into two distinct subsets: one for training the model (train set = 80%) and the other for testing the model (test set = 20%)



3. Create predictive models using regression

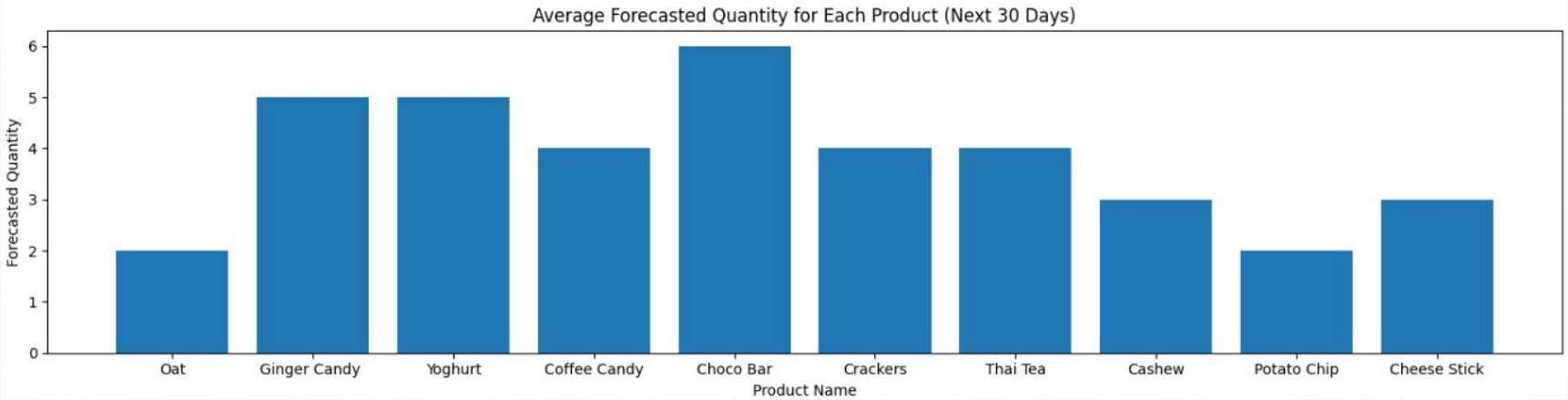
- After that, manually optimize the parameters used in the ARIMA model to get better results or higher performance. Based on the evaluation metrics used (RMSE, MAE, MAPE) the choice of parameters (p=40, d=2, q=1) provides the best performance.



3. Create predictive models using regression

Rakamin Academy

- Because I have found the right parameters, I try to predict the number of sales or the quantity of a certain product for the next 30 days



4. Create clustering models

- From the previously merged data, group customer transaction data, calculate the number of transactions, total quantity, and total amount for each customer. Then, the data was normalized to prepare for clustering analysis.

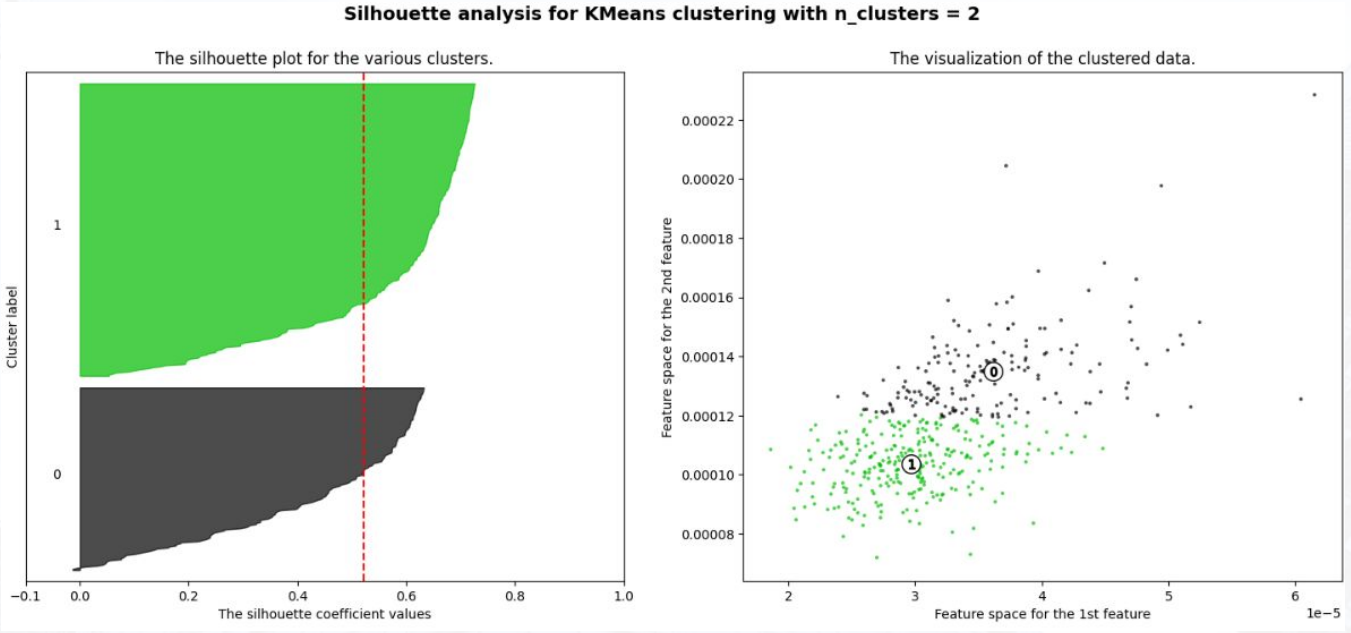
	CustomerID	TransactionID	Qty	TotalAmount
0	1	17	60	623300
1	2	12	56	382300
2	3	15	56	446200
3	4	10	46	302500
4	5	7	27	268600

```
array([[2.72741856e-05, 9.62618317e-05, 9.99999995e-01],
       [3.13889612e-05, 1.46481819e-04, 9.99999989e-01],
       [3.36172117e-05, 1.25504257e-04, 9.99999992e-01],
       ...,
       [3.20271286e-05, 1.16804822e-04, 9.99999993e-01],
       [2.59862980e-05, 9.92204105e-05, 9.99999995e-01],
       [2.98656046e-05, 9.20856143e-05, 9.99999995e-01]])
```

4. Create clustering models

Perform Silhouette Analysis

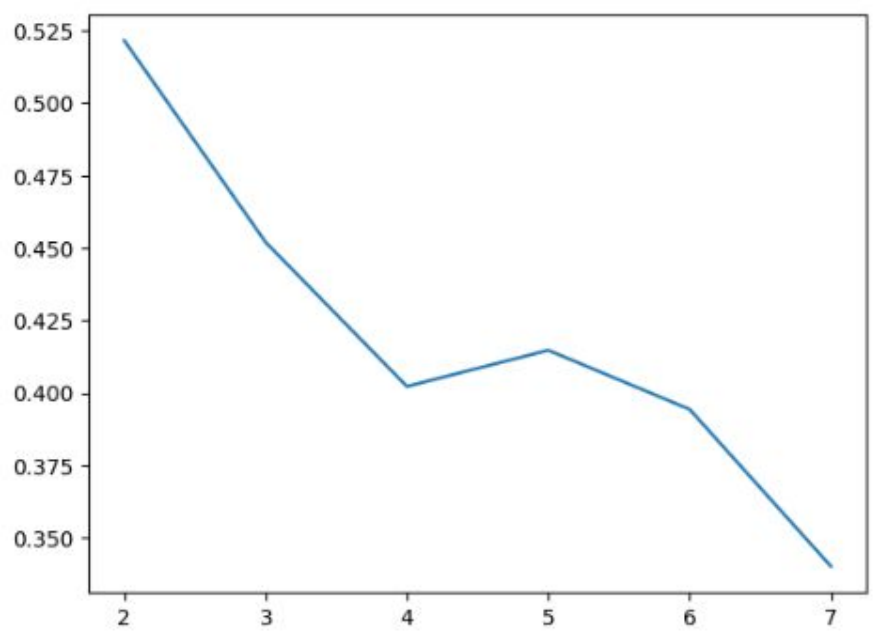
The highest Silhouette Score usually indicates a better number of clusters. In this case, n_clusters = 2 has the highest Silhouette Score, which is around 0.522.




```
For n_clusters = 2 The average silhouette_score is : 0.5215559322569366
For n_clusters = 3 The average silhouette_score is : 0.4512782628208685
For n_clusters = 4 The average silhouette_score is : 0.4571296597676697
For n_clusters = 5 The average silhouette_score is : 0.41467433964189043
For n_clusters = 6 The average silhouette_score is : 0.398563459340275
```

4. Create clustering models

- Because I have used Silhouette Analysis previously and got the information that the best cluster is n_cluster = 2, I will continue with the Elbow Method



4. Create clustering models

Rakamin Academy

cluster_label	CustomerID	TransactionID	Qty	TotalAmount
0	126	11.396825	39.079365	409910.317460
1	27	8.703704	33.370370	212385.185185
2	121	10.578512	39.826446	301657.024793
3	173	11.312139	41.693642	371978.612717

Group the data by label and calculate statistics on the average number of customers, average number of transactions, average quantity, and average total purchases in each group.

And the **business recommendation** results were obtained as follows:

- Group 0 (Label Cluster 0): Focus on promoting products that are already in demand.
- Group 1 (Label Cluster 1): Encourage increased transaction frequency.
- Group 2 (Label Cluster 2): Increase promotion of products of interest.
- Group 3 (Label Cluster 3): Maintain high service and try cross-selling.

Rakamin Academy

Link Github

Rakamin Academy

Video

Thank You