Rakamin Academy X KALBE Nutritionals

Predictive Model Using Regression and Clustering

Kalbe Nutritionals Data Scientist

Presented by
Philipus Dima Wira Pratomo

Philipus Dima Wira Pratomo

About You

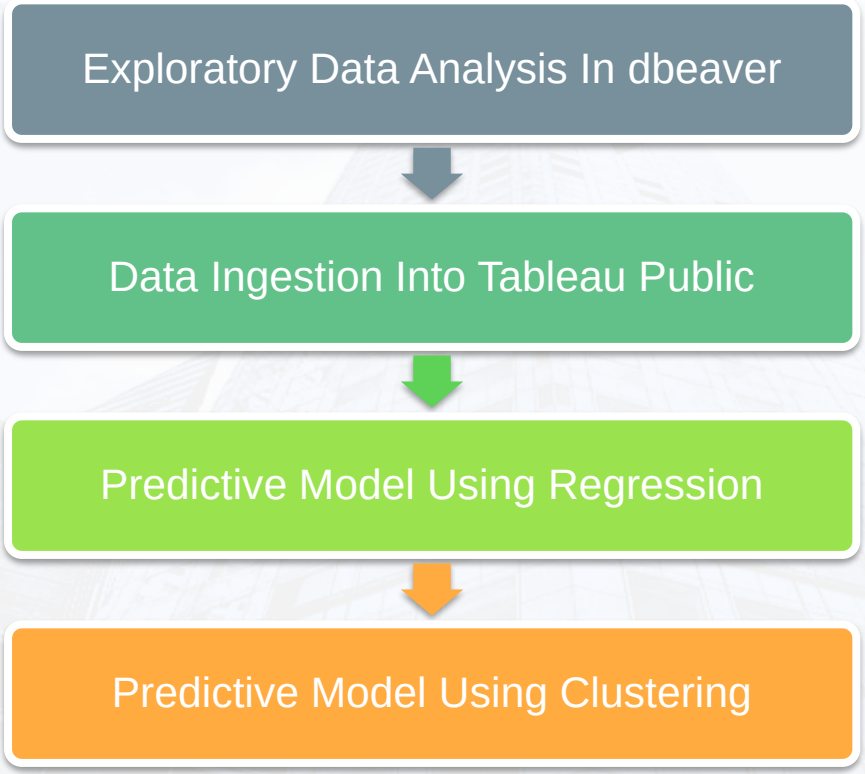
Graduated from Pertamina University majoring in petroleum engineering with a GPA of 3.78. Currently I want to shift career to data scientist by attending a data science bootcamp. where I am developing expertise using data visualization tools such as Tableau and Looker Studio, SQL and several Python libraries to create machine learning. My objective is to become a professional data scientist and leverage my skills to contribute to data-driven decision-making in diverse industries.

Job Experience

- PT Global Digital Niaga Tbk (blibli.com)
  Merchandising Junior Officer
- Warung Pintar Group
  Marketplace Executive
- PT. Pertamina Persero (Dit.Hulu)
  Strategic Planning & Risk Management Intern

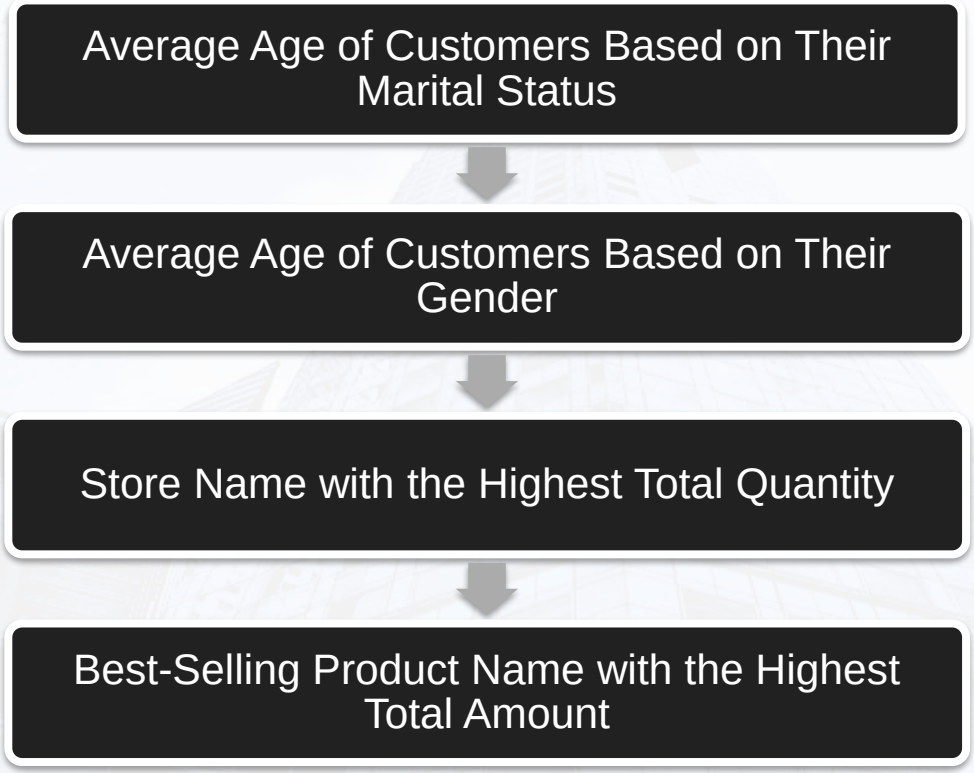
Challenge

Rakamin Academy



Exploratory Data Analysis In dbeaver

Rakamin Academy



Average Age of Customers Based on Their Marital Status

Rakamin Academy

```
select "Marital Status" , avg(age) from customer group by 1
```

ner 1 ×

"Marital Status" , avg(age

Enter a SQL expression to filter results (use Ctrl+

ABC Marital Status	123 avg
	31.3333333333
Married	43.0382352941
Single	29.3846153846

Average Age of Customers Based on Their Gender

Rakamin Academy

```
select gender, avg(age) from customer group by 1
```

ner 1 ×

gender, avg(age) from cu | Enter a SQL expression to filter result

123 gender	123 avg
0	40.326446281
1	39.1414634146

Store Name with the Highest Total Quantity

Rakamin Academy

```
select t.storeid,s.storename,sum(t.qty) total_qty from "transaction" t left join store s
on t.storeid = s.storeid group by 1,2 order by 3 desc
```

tion(+) 1

Enter a SQL expression to filter results (use Ctrl+Space)

storeid	storename	total_qty
9	Lingga	1,439
12	Prestasi Utama	1,395
3	Prima Kota	1,358
6	Lingga	1,338
11	Sinar Harapan	1,331
13	Buana	1,320
1	Prima Tendean	1,310
2	Prima Kelapa Dua	1,296
10	Harapan Baru	1,286
5	Bonafid	1,283
8	Sinar Harapan	1,257
14	Priangan	1,239
4	Gita Ginara	1,236
7	Buana Indah	1,208

Value

9

Best-Selling Product Name with the Highest Total Amount

Rakamin Academy

```
select t.productid,p."Product Name",sum(t.totalamount) total_amount from "transaction" t left join product p
on t.productid = p.productid group by 1,2 order by 3 desc
```

ction(+) 1 ×

Enter a SQL expression to filter results (use Ctrl+Space)

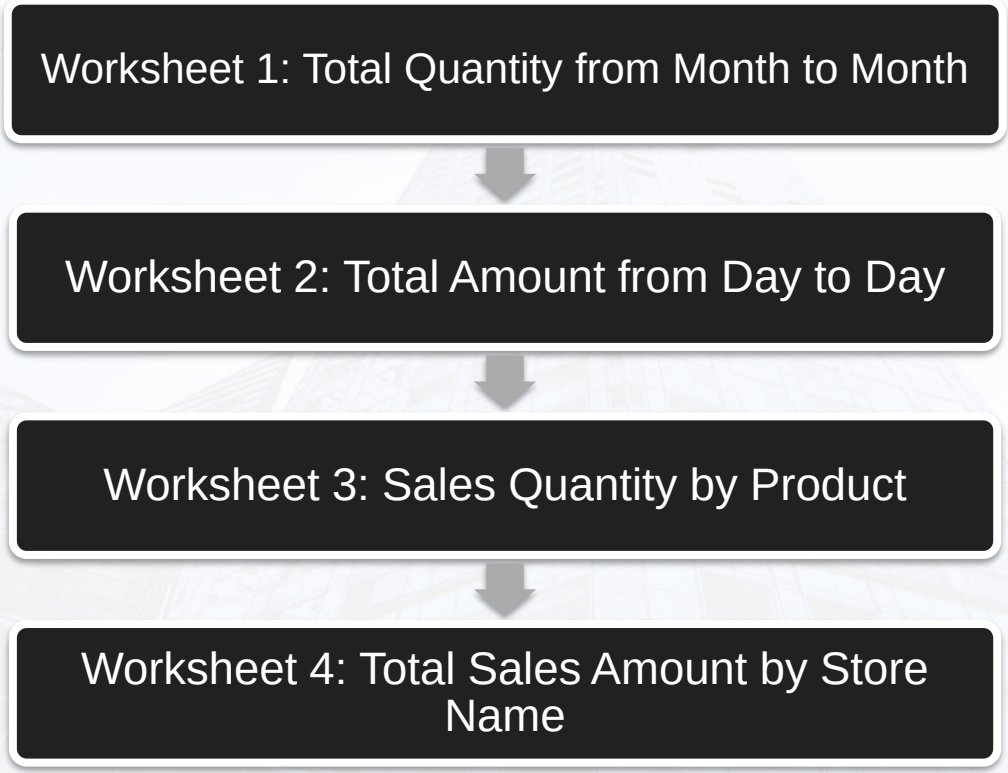
productid	Product Name	total_amount
P10	Cheese Stick	27,615,000
P1	Choco Bar	21,190,400
P7	Coffee Candy	19,711,800
P9	Yoghurt	19,630,000
P8	Oat	15,440,000
P3	Crackers	13,680,000
P4	Potato Chip	13,104,000
P5	Thai Tea	11,982,600
P6	Cashew	11,286,000
P2	Ginger Candy	8,403,200

Value ×

P10

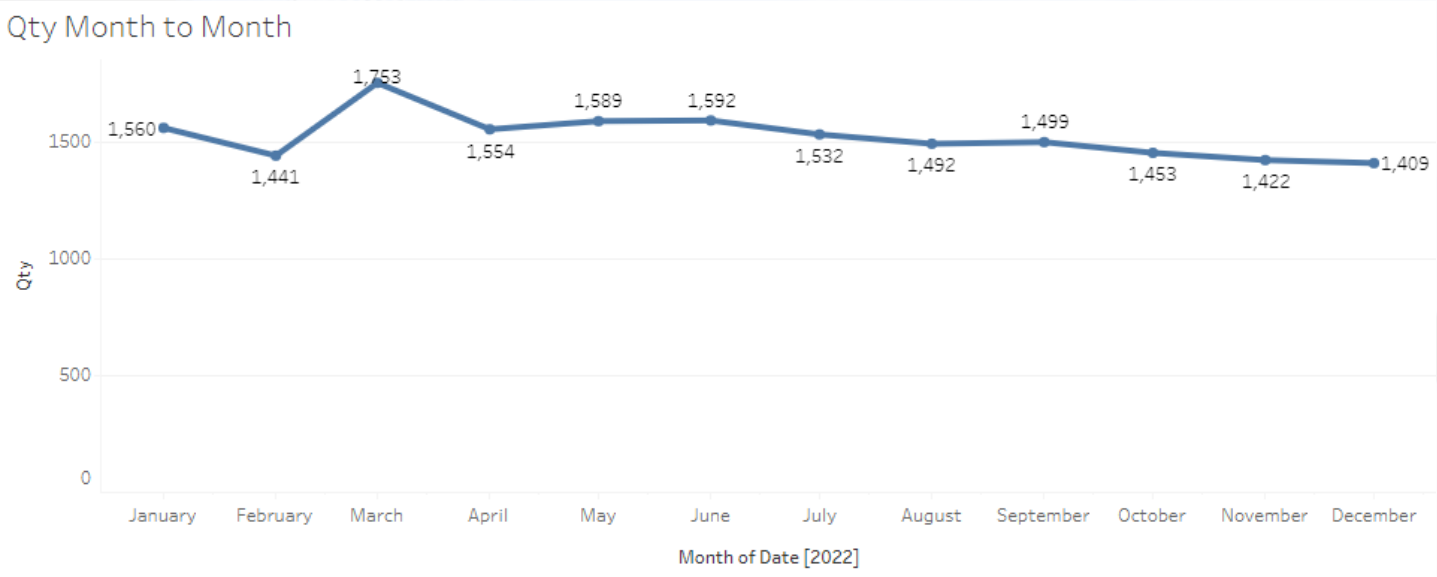
Data Ingestion Into Tableau Public

Rakamin Academy



For details, you can see here

Worksheet 1: Total Quantity from Month to Month

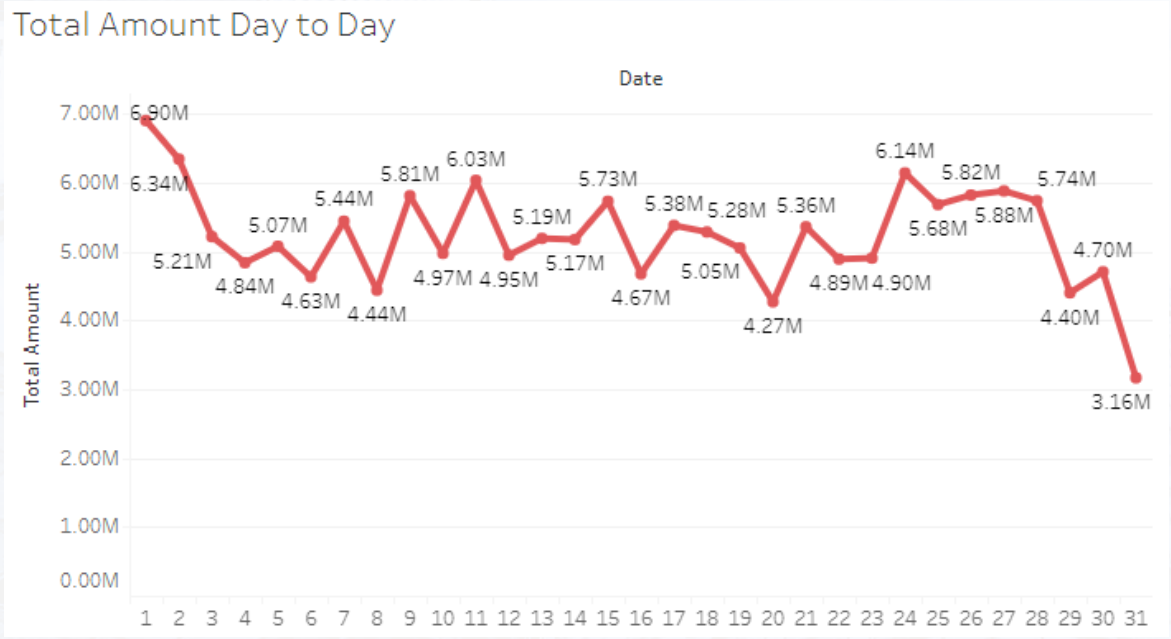


INSIGHT

- When viewed as a whole over each month, there are no significant increases or decreases in Quantity.
- The highest Quantity of sales was recorded in March.
- The lowest Quantity of sales was recorded in February.

For details, you can see here

Worksheet 2: Total Amount from Day to Day

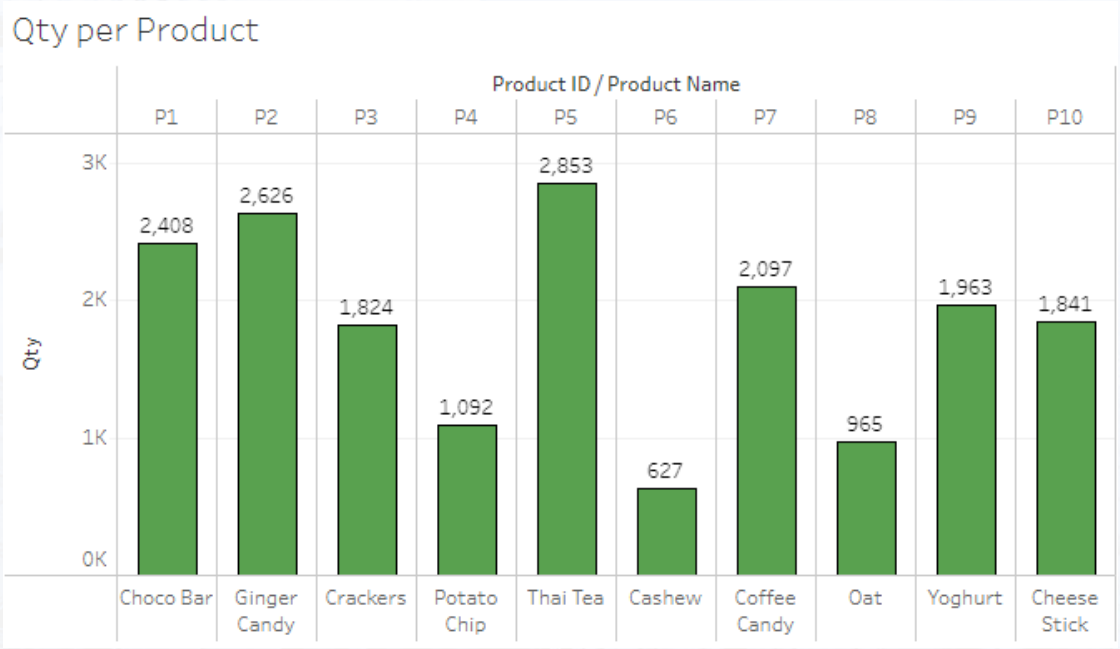


INSIGHT

- there was a decrease in the Total Amount on the last date
- The highest total was on the 1st
- And the lowest total number was on the 31st

For details, you can see here

Worksheet 3: Sales Quantity by Product



INSIGHT

- The following is a table of Sales Quantity for each product.
- It is evident that the highest total is in Thai Tea, with a total of 2,853.
- The lowest total is for Cashew, with 627.

For details, you can see here

Worksheet 4: Total Sales Amount by Store Name

INSIGHT

- The following is a graph displaying the total sales amount for each store name.
- It is apparent that there isn't a significant difference in the total sales amount among the various store names.
- However, the lowest is Buana Indah with a total of 10.63 million, while the highest is Lingga with a total of 13.11 million.

For details, you can see here

Predictive Model

For details, you can see here

Data Understanding

Rakamin Academy

Checking for Null Values, Data Types, Value in every Columns and Unique Value in ID

- There are some null values in the "Marital Status" column of the "df_customer" dataset.
- The values in the "income" feature in the "df_customer" dataset are currently using a comma as a decimal separator; these should be replaced with a period and converted to float data type.
- The values in the "Latitude" and "Longitude" features in the "df_store" dataset are currently using a comma as a decimal separator; these should be replaced with a period.
- The "date" feature in the "df_transaction" dataset should be converted to a datetime data type.
- Many values in the "TransactionID" feature in the "df_transaction" dataset are duplicated, even though "TransactionID" should be unique. In this case, we will select the records with the latest date.

For details, you can see here

Data Cleaning & Data Preprocessing

Rakamin Academy

Handling Type Data & Correct Value

```
[ ] df_customer['Income'] = df_customer['Income'].replace(',','.',regex=True).astype('float')

[ ] df_store['Latitude'] = df_store['Latitude'].replace(',','.',regex=True)
    df_store['Longitude'] = df_store['Longitude'].replace(',','.',regex=True)

[ ] df_transaction['Date'] = pd.to_datetime(df_transaction['Date'])
```

Handling Duplicate Data

```
df_transaction = df_transaction.sort_values(by=['TransactionID', 'Date'], ascending=[True, False])
df_transaction = df_transaction.drop_duplicates(subset='TransactionID', keep='first')
```

Handling Missing Value

Missing values in Marital Status are filled in with "Other"

```
df_customer['Marital Status'] = df_customer['Marital Status'].fillna("Others")
```

For details, you can see here

Combine All Data

Rakamin Academy

```
[24] df_merge = pd.merge(df_transaction,df_customer,on=["CustomerID"])
```

```
[25] df_merge = pd.merge(df_merge, df_product.drop(columns=['Price']), on=["ProductID"])
```

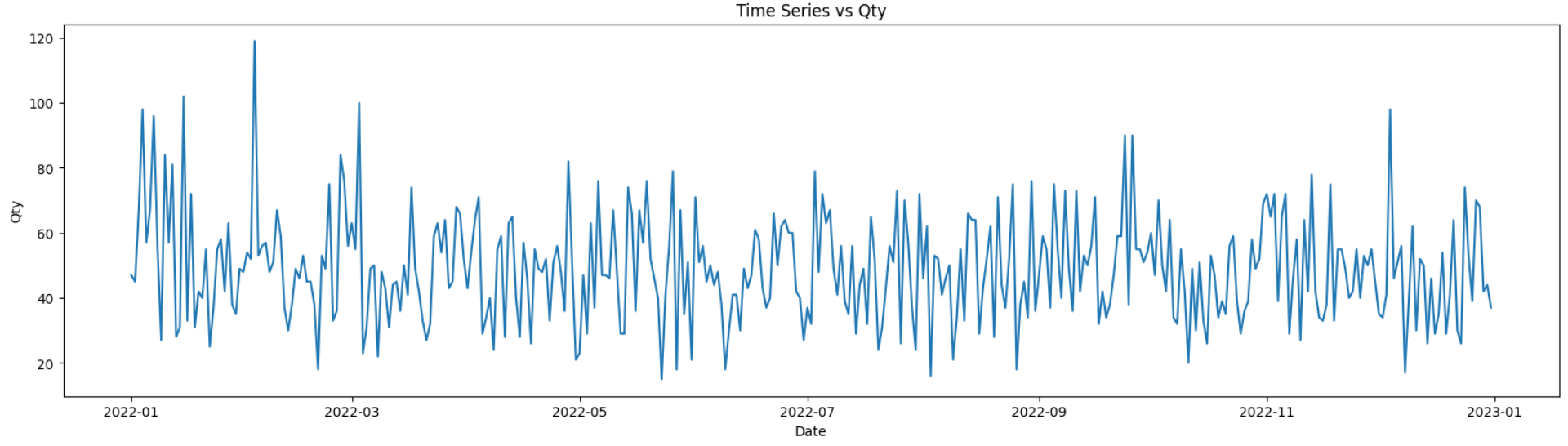
```
[26] df_merge = pd.merge(df_merge,df_store,on=["StoreID"])
```

For details, you can see here

Machine Learning Regression (Time Series)

Rakamin Academy

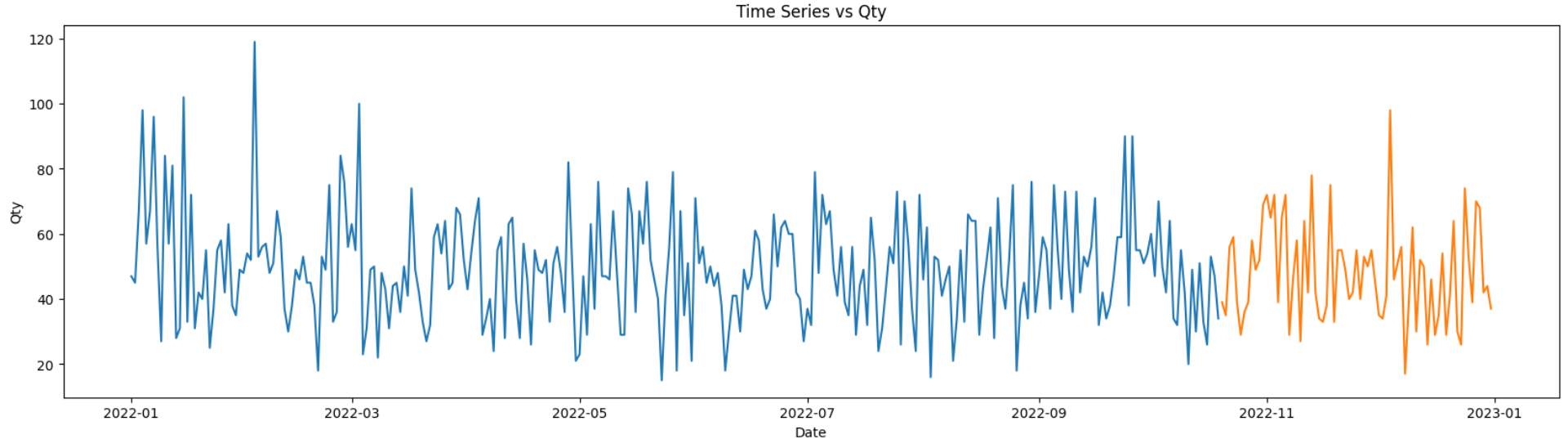
All Data



For details, you can see here

Machine Learning Regression (Time Series)

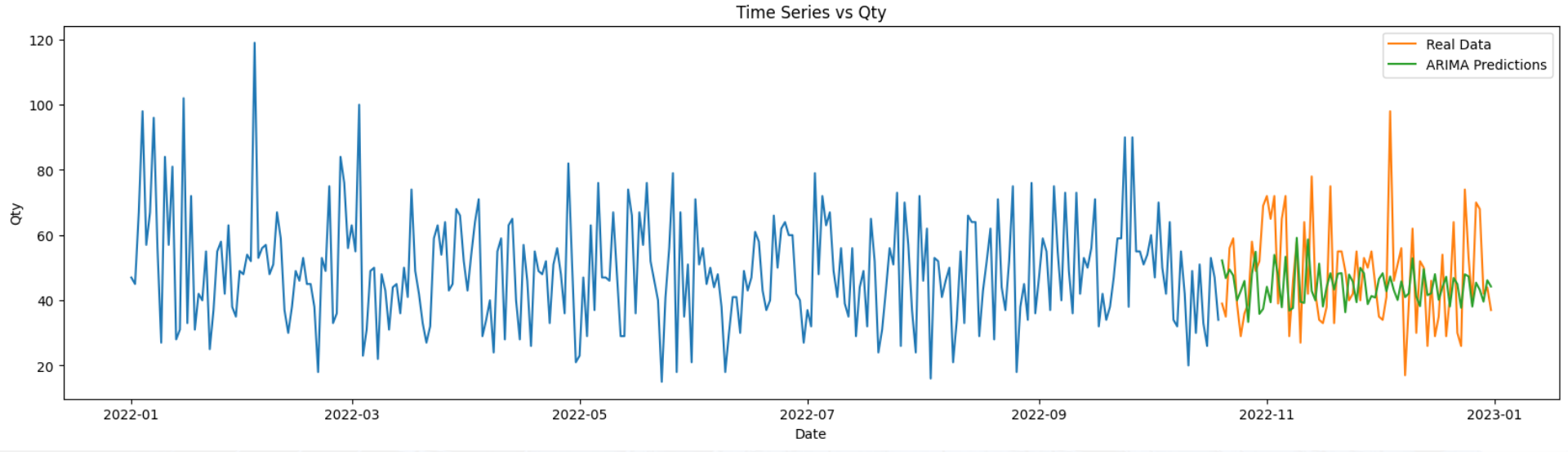
Split into Data Training 80% and Data Testing 20%



For details, you can see here

ARIMA

Parameter	Value
P	48
D	2
Q	1



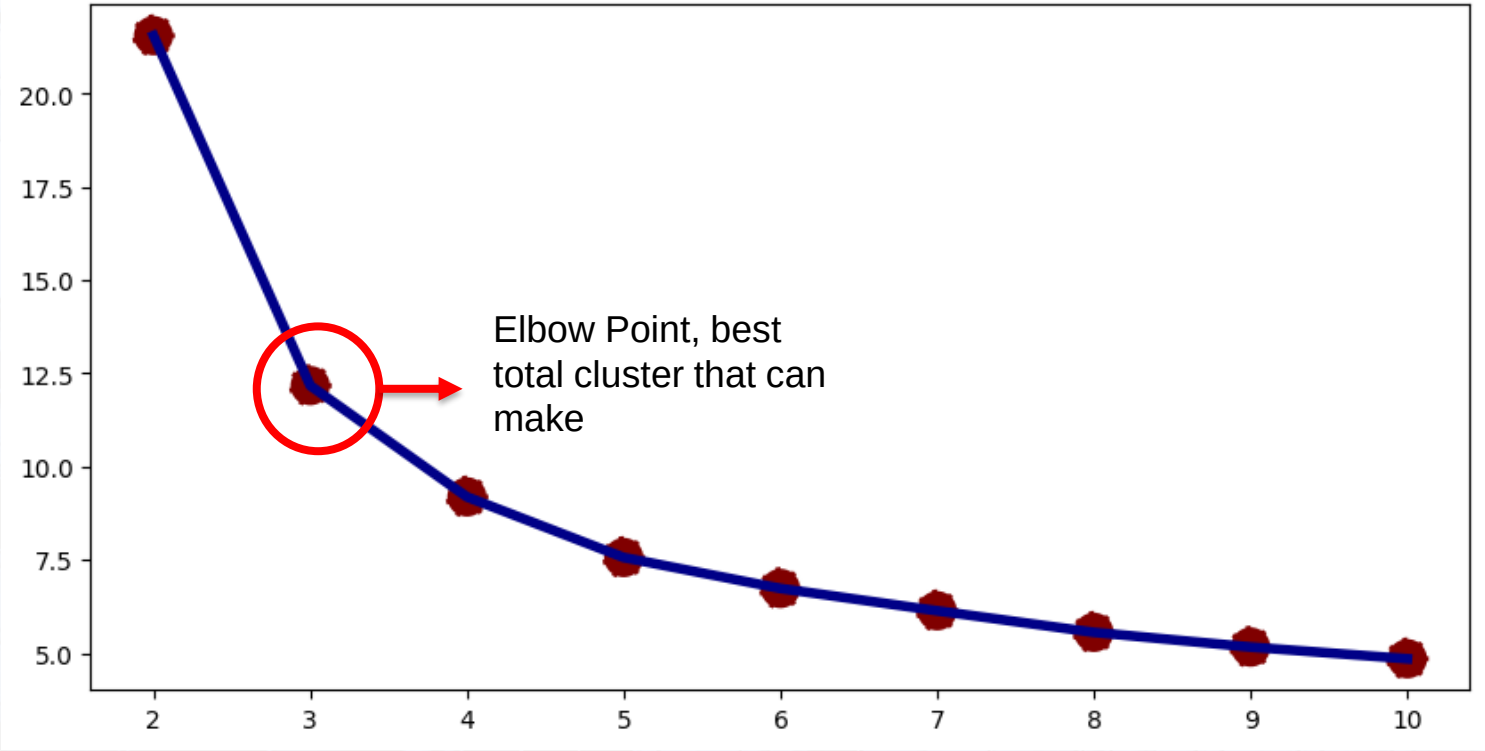
Machine Learning Clustering

Rakamin Academy

Feature Engineering	• There will be Feature Engineering 1. total_trx (count TransactionID) 2. total_qty (sum Qty) 3. total_amount (sum TotalAmount)
Outlier Handling	Will be Outlier Handling in new Feature because Clustering very influential on outliers
Normalize	Normalization helps eliminate the effects of scale differences, ensuring that attributes with a larger range of values do not dominate attributes with a smaller range of values.

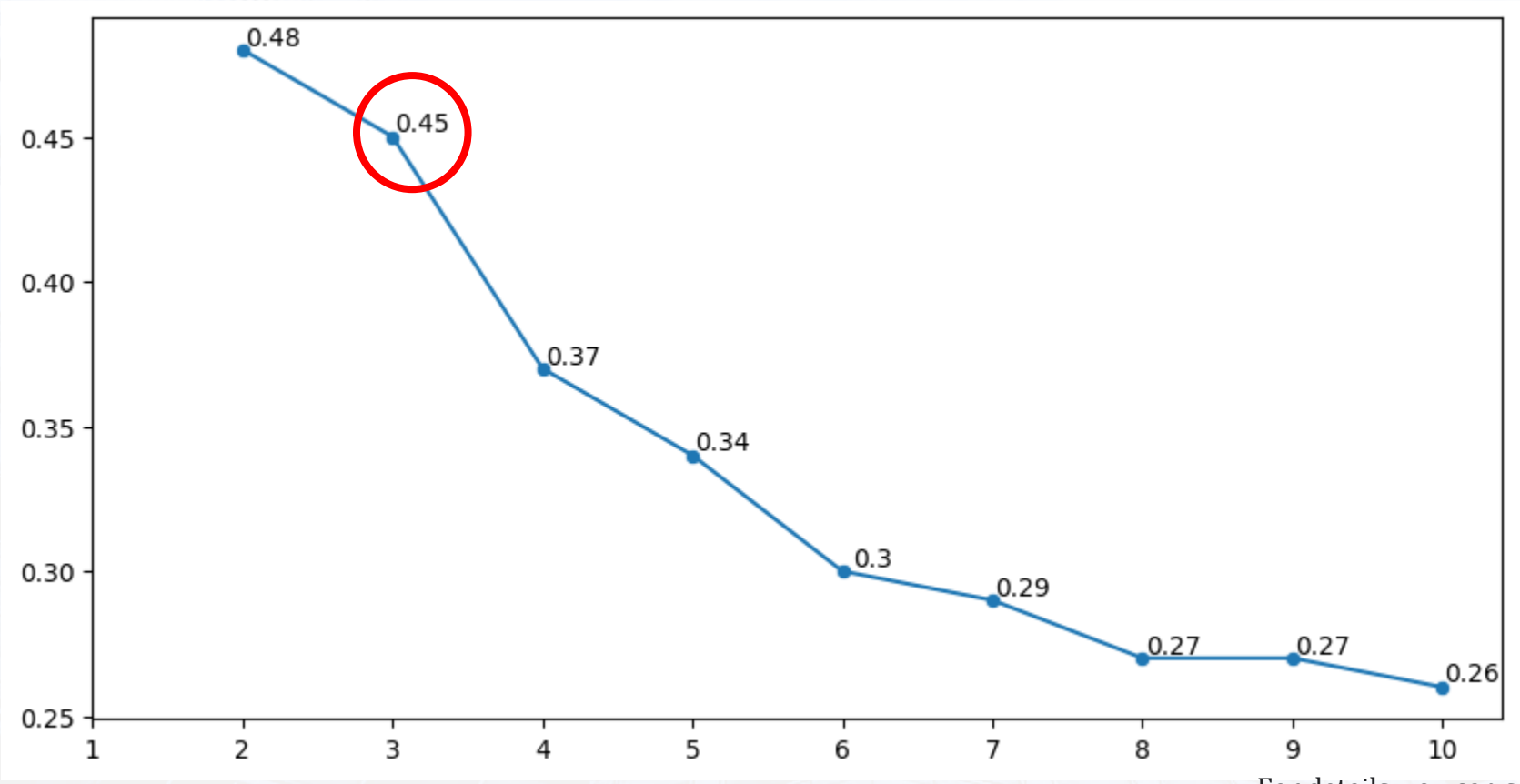
For details, you can see here

Elbow Method



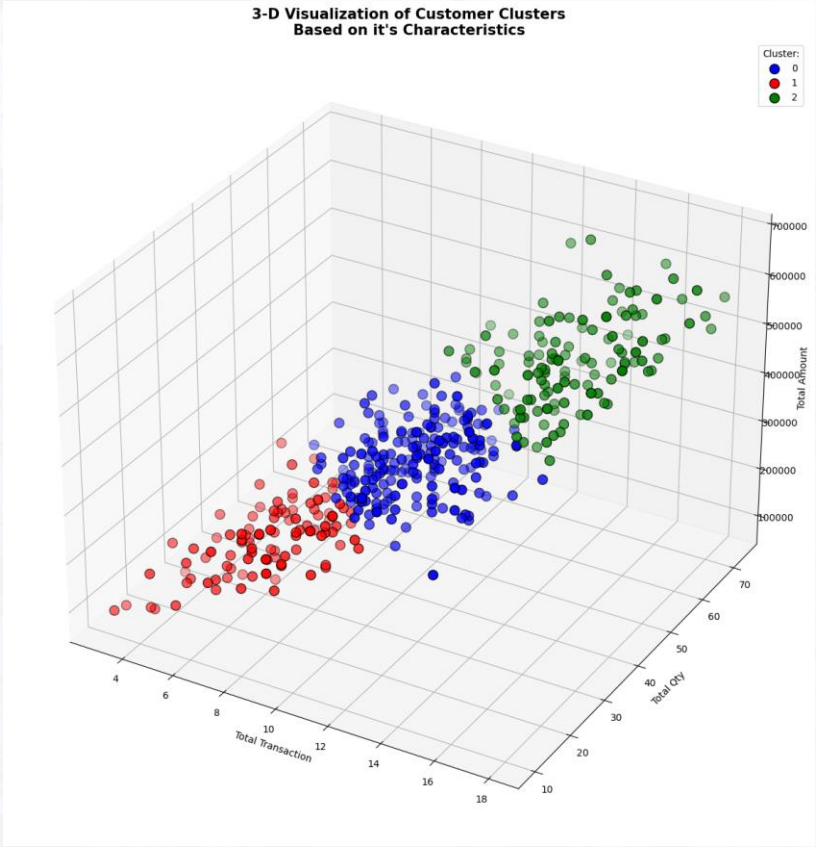
For details, you can see here

Silhouette score



For details, you can see here

3-D Visualization



Cluster	Total Customer	Total Transaction			Total Quantity			Total Amount		
		mean	median	max	mean	median	max	mean	median	max
0	211	11	11	15	38	38	53	337,079	335,000	476,200
1	108	7	7	10	25	25	37	208,935	212,950	322,600
2	124	15	15	18	55	54	73	496,818	490,100	676,200

INSIGHTS

1. Cluster 1 represents customers with the lowest total count, characterized by average transaction values, average purchased quantity, and the lowest total amount compared to the other clusters.
2. Cluster 2 represents customers with a moderate total count, but they exhibit characteristics of higher average transaction values, average purchased quantity, and the highest total amount compared to the other clusters.
3. Cluster 0 is the largest cluster in terms of customer count.

For details, you can see here

Video Presentation Here

link video here

Thank You

Rakamin Academy X KALBE Nutritionals