Rakamin Academy X KALBE Nutritionals

FINAL PROJECT

Project Based Internship - Data Scientist

Presented by : Fikri Alinfijar

Fikri Alinfijar

About Me

An undergraduate Data Science student at Telkom University. Very interested in the field of data analysis and machine learning. Have experience in working on projects in this field. A hard worker and adapt quickly to new environments. Currently, developing technical skills that support my future career. Have skills in SQL, python, and visualization using Tableau, Power BI, and Looker Studio.

Rakamin Academy

Experience

Project-Based Virtual Intern : Data Scientist Kalbe Nutritionals x Rakamin Academy

- Developing an ARIMA and SARIMAX model to predict the daily total quantity of products sold and Utilizing the KMeans method to cluster similar customers.

Awardee of Data Analyst SMKDEV Scholarship 2023

- Deepen understanding of the basic concepts of data analysis, data processing, and implementation in various business contexts.
- Engage in projects involving data analysis, which provides insight into how to interpret data to generate valuable insights.

Rakamin Academy

DATASET

Transaction

Trasaction ID : Unique ID for each transaction

Customer ID : Unique ID for each customer

Date : Date the transaction was made

Product ID : Unique ID for each product

Price : Product price

Qty : Number of products purchased

Total Amount : Total amount to be paid

Store ID : Unique ID for each store

Customer

Customer ID : Unique ID for each customer

Age : Age of customer

Gender : Gender of customer

Marital Status : Customer's marital status

Income : Customer income

Store

Store ID : Unique ID for each store

Store Name: Store name

Group Store : Store category

Type : Store type

Latitidue : Store location latitude coordinates

Longitude : Store location longitude coordinates

Product

Product ID : Unique ID for each product

Product Name: Name of product

Price : Product price

Task : Display the average age of customers based on their marital status

Query :

```
SELECT
    "Marital Status",
    AVG(c."Age") AS Average_Age
FROM
    kalbe."Customer" c
GROUP BY
    "Marital Status"
```

Result :

Marital Status	average_age
	31.3333333333
Married	43.0382352941
Single	29.3846153846

Explanation

It was found that the average age of customers with the status "Married" is 43.038, while the customer with the status "Single" is 29.384.

Task : Display the average age of customers by gender

Query :

```
SELECT
    "Gender",
    AVG(c."Age") AS Average_Age
FROM
    kalbe."Customer" c
GROUP BY
    "Gender"
```

Result :

123 Gender	123 average_age
0	40.326446281
1	39.1414634146

Explanation

It was found that the average age of male customers is 40.326, while the female customer is 39.141.

Task : Display the store name with the highest total quantity

Query :

```
SELECT
    s."StoreName",
    SUM(t."Qty") AS Total_Quantity
FROM
    kalbe."Store" s
JOIN
    kalbe."Transaction" t ON s."StoreID"  = t."StoreID"
GROUP BY
    s."StoreName"
ORDER BY
    Total_Quantity DESC
LIMIT 1;
```

Result :

StoreName	total_quantity
Lingga	2,777

Explanation

The store with the highest total quantity is the Lingga store, with a total quantity of 2,777.

Task : Displays the best-selling product name with the highest total amount

Query :

```sql
SELECT
    p."Product Name",
    SUM(t."TotalAmount") AS Total_Amount
FROM
    kalbe."Product" p
JOIN
    kalbe."Transaction" t ON p."ProductID" = t."ProductID"
GROUP BY
    p."Product Name"
ORDER BY
    Total_Amount DESC
LIMIT 1;
```

Result :

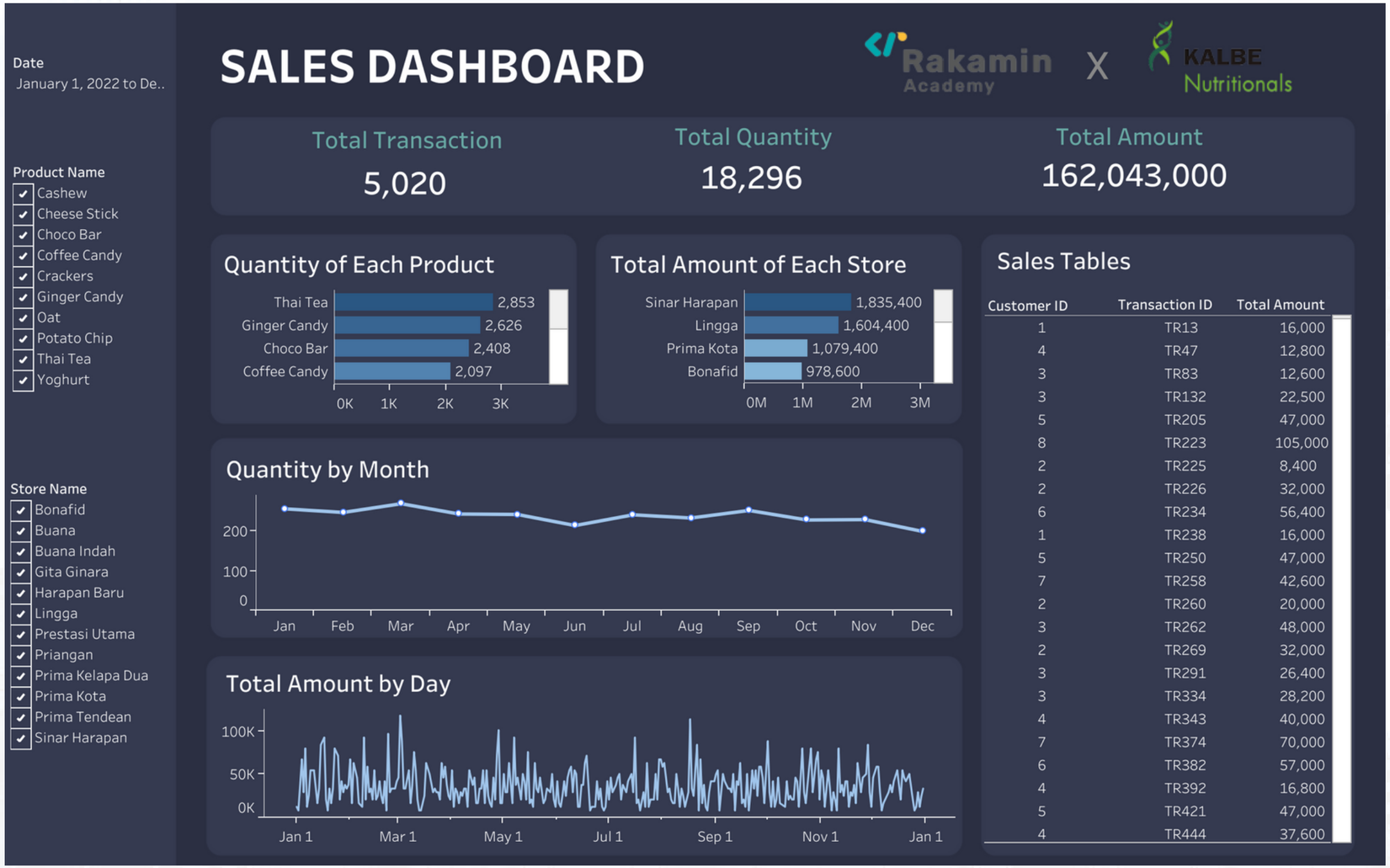
ABC Product Name	123 total_amount
Cheese Stick	27,615,000

Explanation

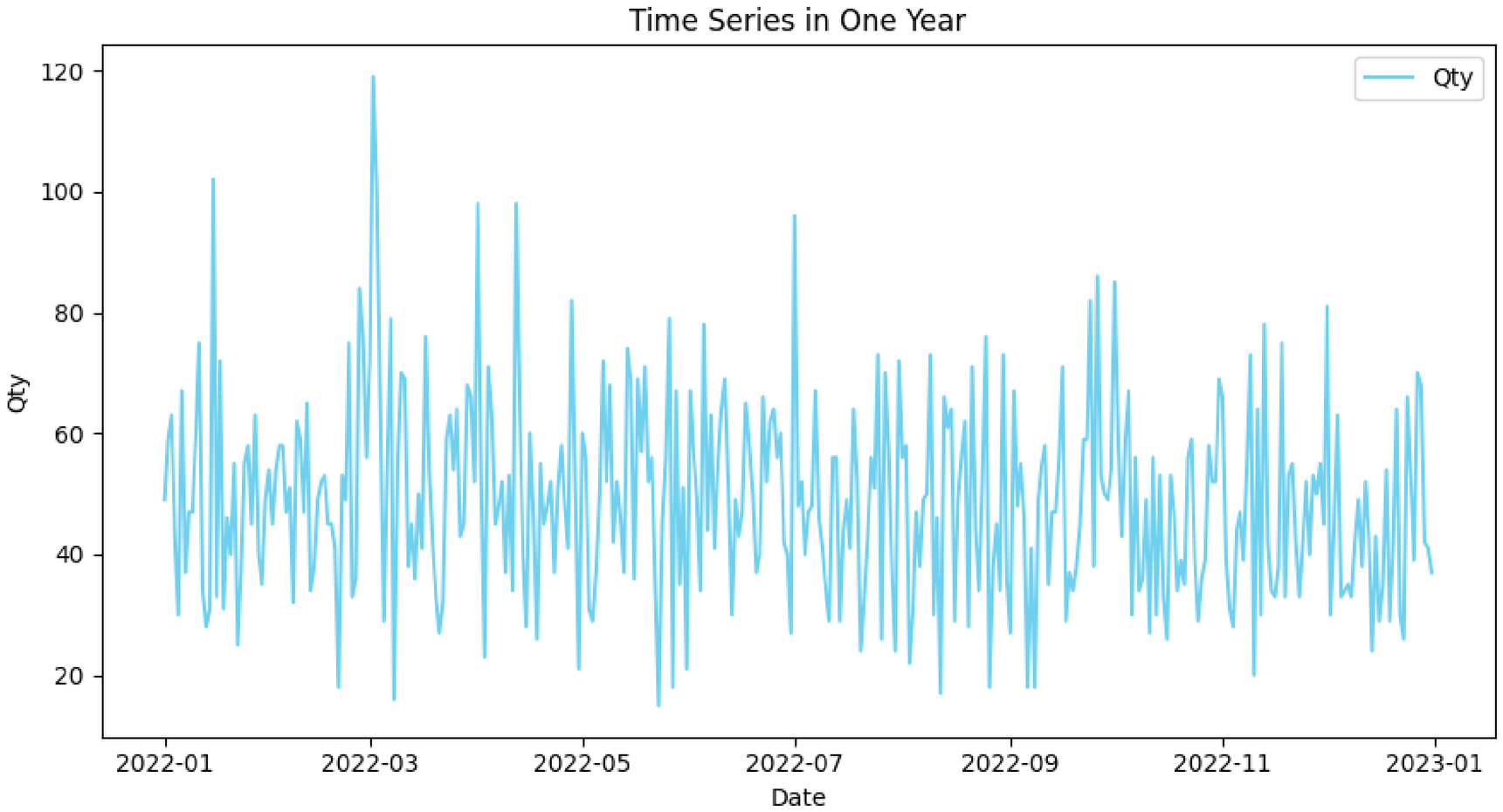
The product with the highest total amount is Cheese Stick, with a total amout of 27.6 million

Dashboard

Rakamin Academy



Plot Data :


Time Series in One Year
Qty
120
100
80
60
40
20
Qty
2022-01
2022-03
2022-05
2022-07
2022-09
2022-11
2023-01
Date

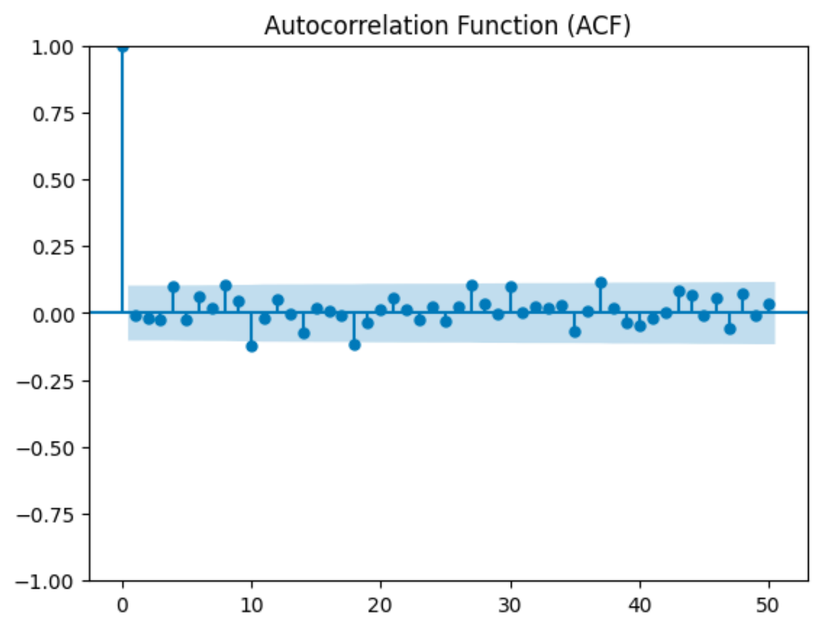
```python
# Melakukan uji stasioneritas menggunakan Augmented Dickey-Fuller (ADF) test
result = adfuller(df_regression['Qty'])
print('ADF Statistic:', result[0])
print('p-value:', result[1])
print('Critical Values:')
for key, value in result[4].items():
    print(f'\t{key}: {value}')

if result[1] <= 0.05:
  print("\nThe data is stationary")
else:
  print("\nThe data is non-stationary")
```

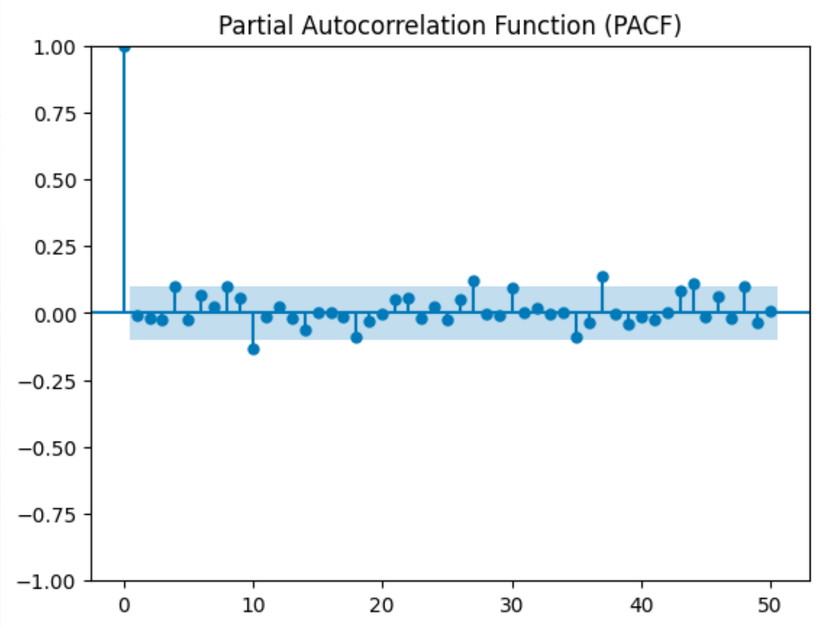
```
ADF Statistic: -19.131655296406304
p-value: 0.0
Critical Values:
        1%: -3.4484434475193777
        5%: -2.869513170510808
        10%: -2.571017574266393

The data is stationary
```

Explanation

The code is used to check whether the data is stationary or not.

It is found that the p-value < alpha, so the data is stationary.





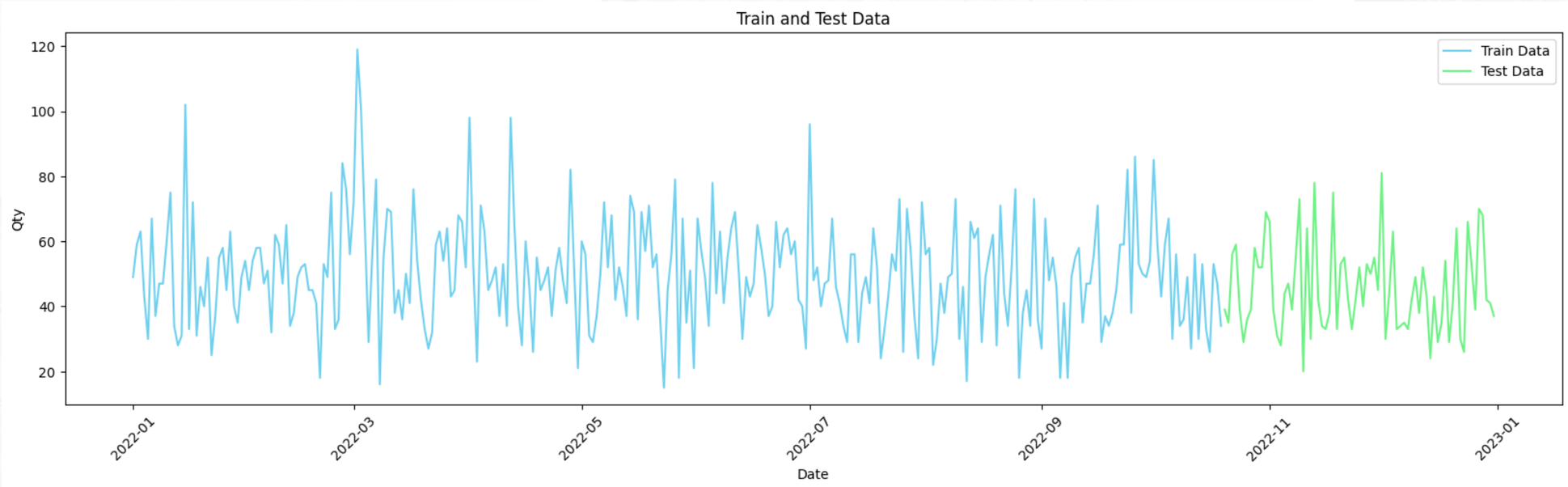
Explanation

ACF and PACF plots are used to determine the ARIMA model. After testing to determine the best model, it was found that AR order = 4, MA order = 8, while I order = 0 because the data was stationary so differencing was not done.

```
# Membagi dataset menjadi data train dan data test
cut_off = round(df_regression.shape[0] * 0.8)
df_train = df_regression[:cut_off]
df_test = df_regression[cut_off:].reset_index(drop=True)
```

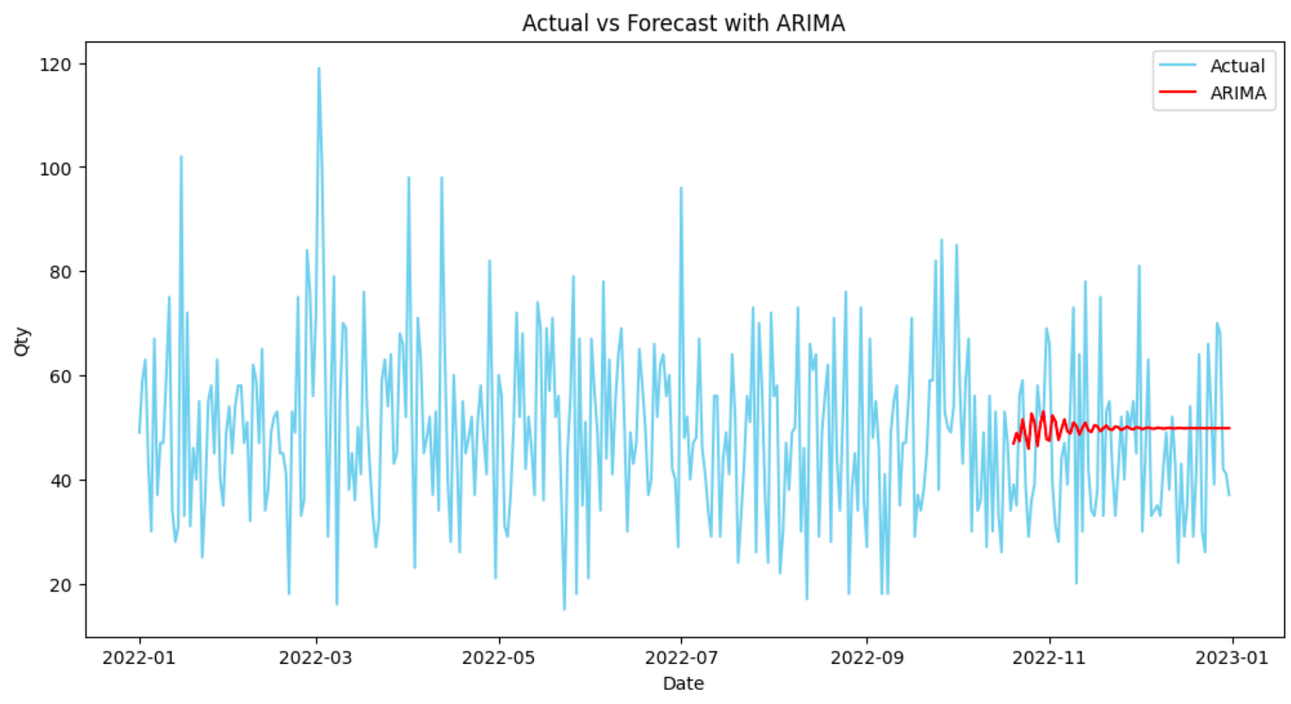
Explanation

Split the dataset into train data and test data



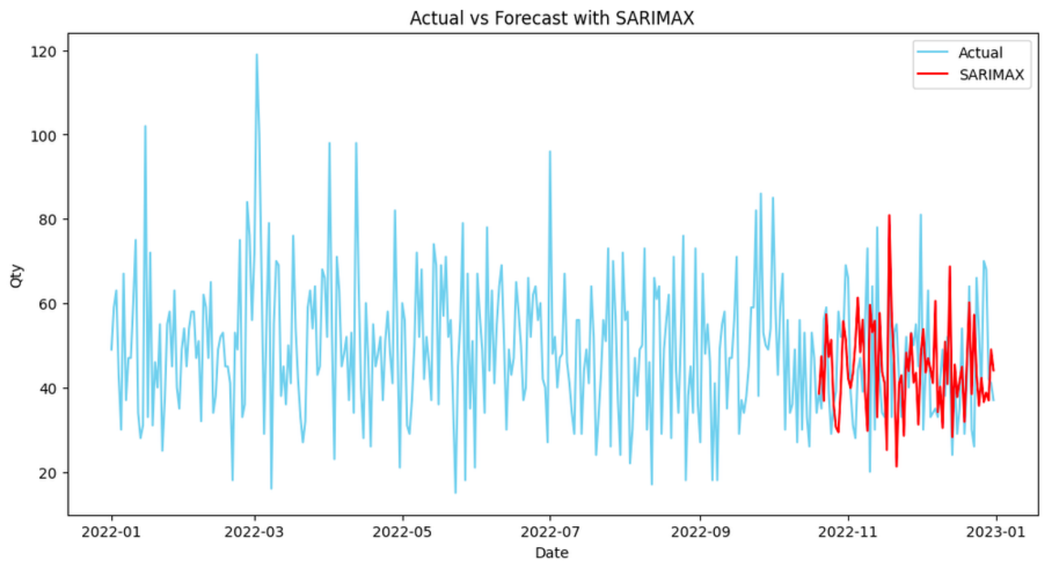
```python
# Membangun model ARIMA dengan parameter yang telah ditetapkan
model = sm.tsa.ARIMA(df_train['Qty'], order=(4, 0, 8))
model_fit = model.fit()

# Menampilkan ringkasan model
print(model_fit.summary())
```

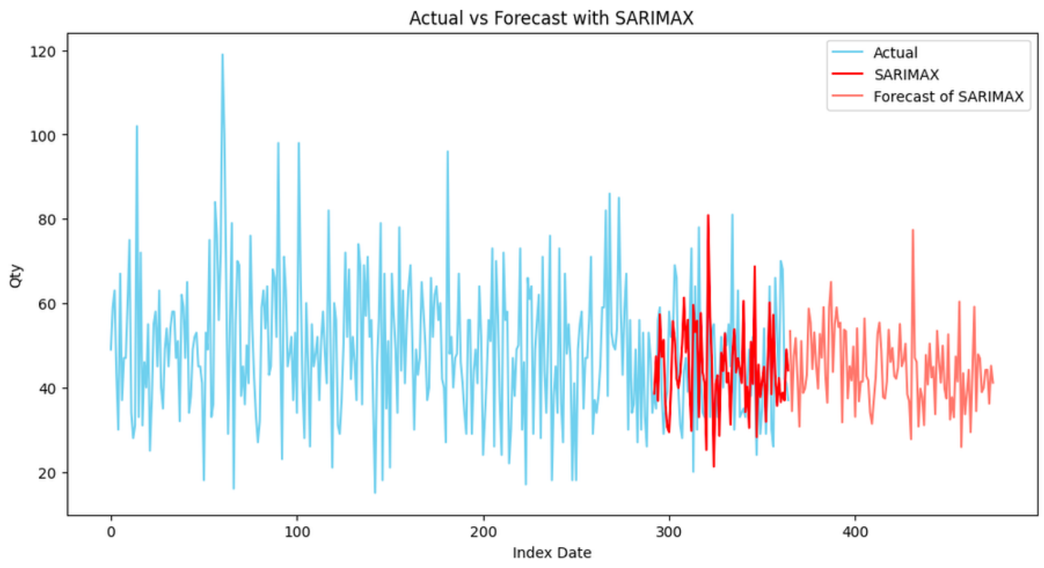
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                    Qty   No. Observations:                  292
Model:                 ARIMA(4, 0, 8)   Log Likelihood               -1223.869
Date:                Sat, 28 Oct 2023   AIC                           2475.738
Time:                        23:47:37   BIC                           2527.213
Sample:                             0   HQIC                          2496.357
                                  - 292
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         49.8212      1.173     42.487      0.000      47.523      52.119
ar.L1          0.1115      0.133      0.839      0.402      -0.149       0.372
ar.L2         -1.3907      0.133    -10.456      0.000      -1.651      -1.130
ar.L3          0.1269      0.128      0.990      0.322      -0.124       0.378
ar.L4         -0.6357      0.127     -5.005      0.000      -0.885      -0.387
ma.L1         -0.1126      0.137     -0.819      0.413      -0.382       0.157
ma.L2          1.3969      0.136     10.290      0.000       1.131       1.663
ma.L3         -0.1946      0.154     -1.261      0.207      -0.497       0.108
ma.L4          0.7353      0.155      4.741      0.000       0.431       1.039
ma.L5         -0.0804      0.104     -0.771      0.441      -0.285       0.124
ma.L6          0.2345      0.106      2.217      0.027       0.027       0.442
ma.L7         -0.0577      0.068     -0.845      0.398      -0.192       0.076
ma.L8          0.2492      0.063      3.939      0.000       0.125       0.373
```



Explanation

Even though we have used the best ARIMA mode, the forecasting results are still poor because the data is seasonal.

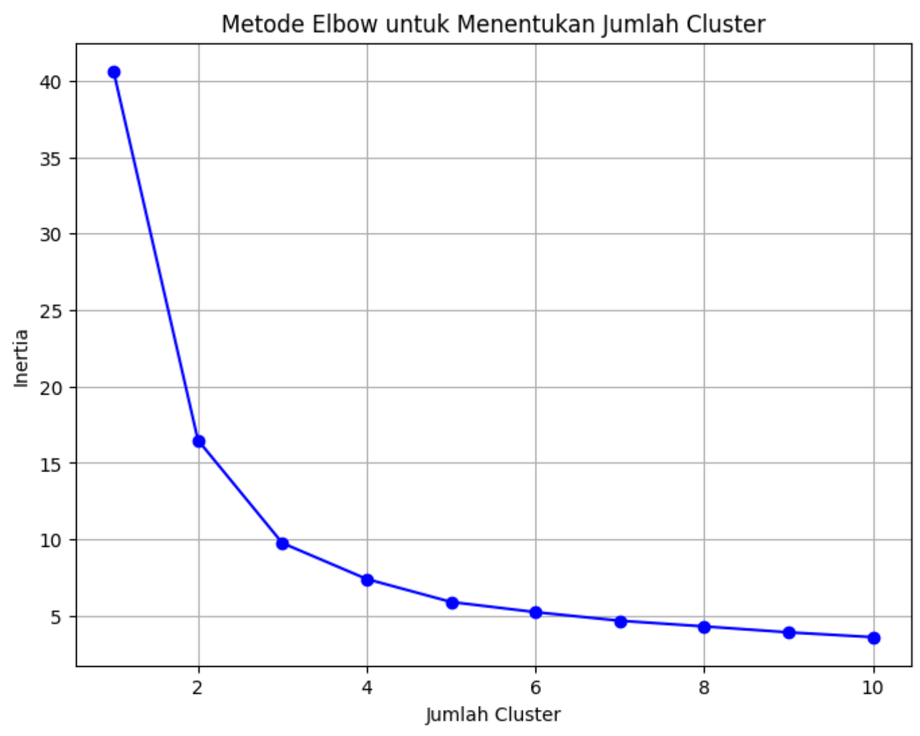




Explanation

It can be seen that by using the SARIMAX model, the prediction becomes more accurate because the data is seasonal. In the right image, I made a prediction for the next 110 days

```python
# Menentukan jumlah cluster (gunakan metode Elbow untuk memilih jumlah cluster yang optimal)
inertia = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, random_state=0, n_init=10)
    kmeans.fit(normalized_data)
    inertia.append(kmeans.inertia_)

# Plot Elbow Method untuk menentukan jumlah cluster yang optimal
plt.figure(figsize=(8, 6))
plt.plot(range(1, 11), inertia, marker='o', linestyle='-', color='b')
plt.xlabel('Jumlah Cluster')
plt.ylabel('Inertia')
plt.title('Metode Elbow untuk Menentukan Jumlah Cluster')
plt.grid(True)
plt.show()
```



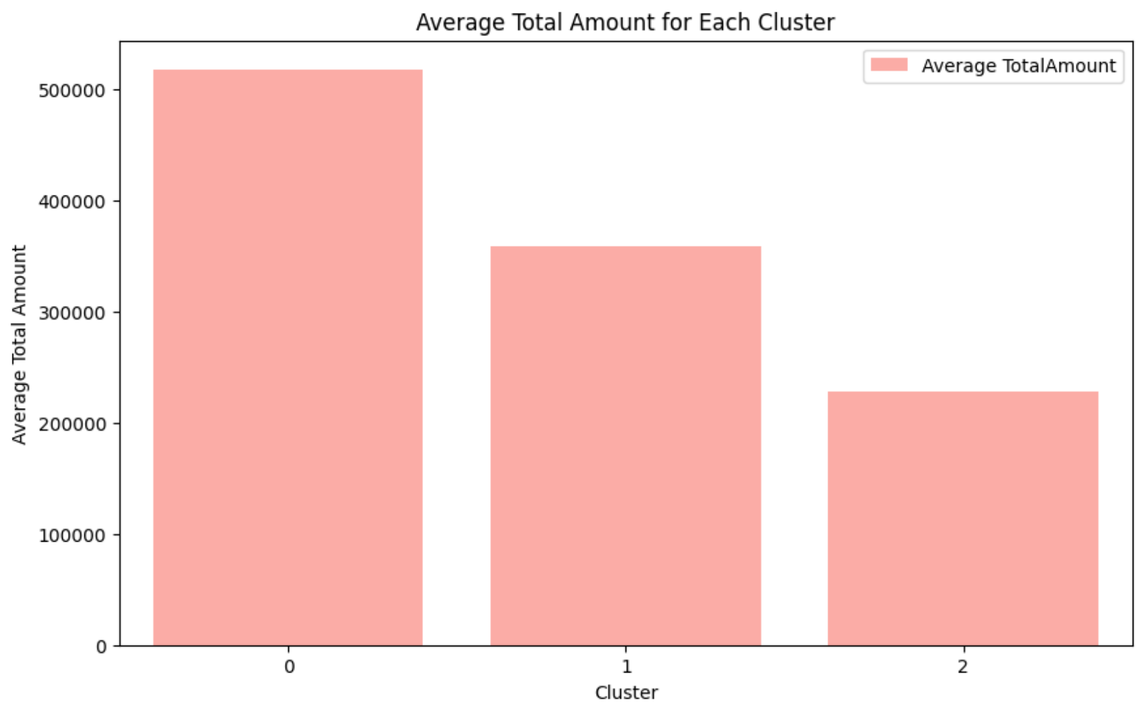

```python
# Menentukan cluster yang paling optimal menggunakan KneeLocator
kl = KneeLocator(range(1, 11), inertia, curve="convex", direction="decreasing" )
kl.elbow
```

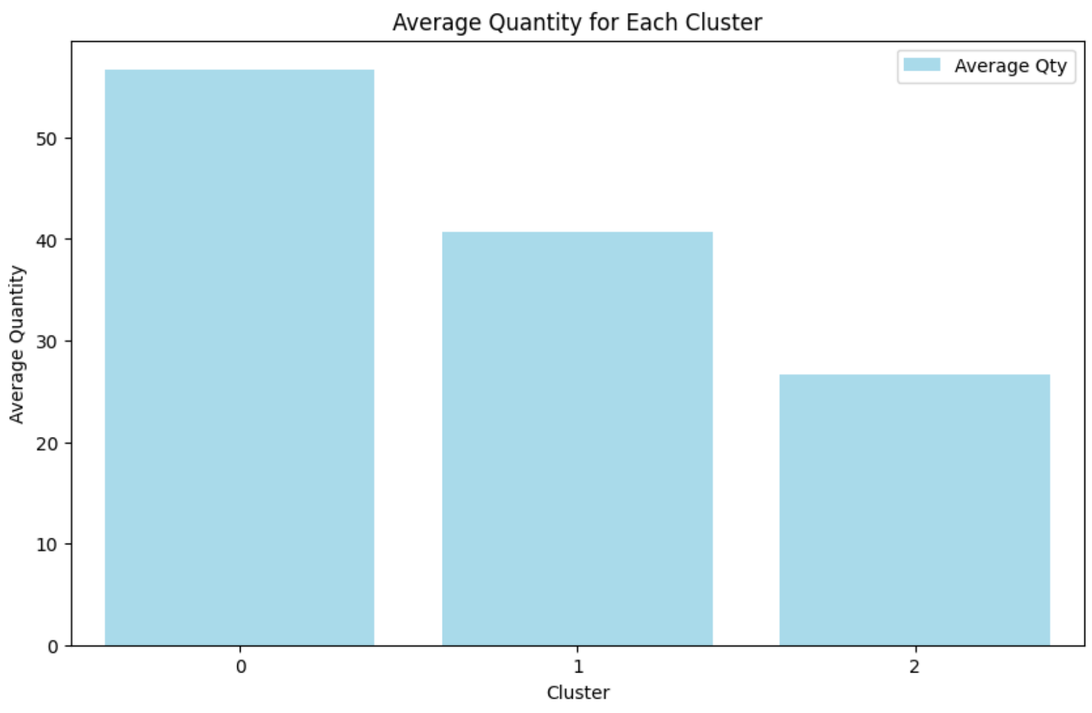
```
3
```

Explanation

By using the elbow method and KneeLocator, it is found that the most optimal number of clusters is 3 clusters.

Plot Clustering :


Average Total Amount for Each Cluster
Average TotalAmount
Average Total Amount
500000
400000
300000
200000
100000
0
0
1
2
Cluster


Average Quantity for Each Cluster
Average Qty
Average Quantity
50
40
30
20
10
0
0
1
2
Cluster

Customer Characteristics

1. Cluster 0 is a "Loyal Customer", this is because the average number of purchases of goods and total expenditure is the highest.

2. Cluster 1 is a "Potential Loyal Customer, this is because the average number of purchases of goods and total expenditure is medium.

3. Cluster 2 is a "New Customer", this is because the average number of purchases of goods and total expenditure is the lowest.

Business Recomendation

Cluster 0

- Develop an attractive loyalty program.
- Conduct regular communication with customers.
- Conduct regular customer satisfaction surveys to understand their preferences and ensure satisfaction is maintained.

Cluster 1

- Provide relevant and attractive offers.
- Offer discounts based on previous purchases or reward programs to encourage repeat purchases and build customer loyalty.
- Ensure responsive and friendly customer service, responding to their queries or feedback quickly and efficiently.

Cluster 2

- Provide a special discount or gift upon first purchase.
- Offers a loyalty program that provides many benefits.

Dashboard

https://public.tableau.com/views/KalbeDashboard_16981621414690/Dashboard4?:language=en-US&:display_count=n&:origin=viz_share_link

Github

https://github.com/fikrialnfjr/DS-KalbeNutritionals

Gdrive (Video Presentasi)

https://drive.google.com/file/d/1ED4WL-l3LppAc8oN39hZlbMiYUyDKcsn/view?usp=sharing

Thank You

Rakamin Academy X KALBE Nutritionals