Rakamin Academy X KALBE Nutritionals

Total Sales Prediction and Customer Segmentation Using ARIMA and K-Means Clustering Algorithm

Project Based Virtual Intern:
Data Scientist Kalbe Nutritionals X Rakamin Academy

Presented by
Jaelani (Jay)

Jaelani (Jay)

Jay has been working since he was 18. Jay has 11 years of industry experience;

- 9 years in the Pharmaceutical Sector as a Quality Control Laboratory Technician
- 2 years with an NGO (Philanthropy) as a Risk Management and Data Analyst.

He has now decided to pursue his dream job as a Data Analyst.

Jay participated in various online learning courses after graduating from the Data Science Bootcamp in 2022 to hone his Data Analysis skills.

Rakamin Academy

EXPERIENCES

- Data Analyst at BAZNAS (Badan Amil Zakat Nasional
  March 2023 – June 2023
- Risk Management at Rumah Zakat Indonesia
  September 2021 – March 2023
- Quality Control Laboratory Technician at PT. Lucas Djaja
  April 2018 – February 2021
- Quality Control Laboratory Technician at PT. Kimia Farma
  April 2017 – November 2017
- Quality Control Laboratory Technician at PT. Takeda Indonesia
  December 2012 – October 2016

PRESENTATION OUTLINE

1. Background Story
2. Exploratory Data Analysis
3. Sales Dasboard
4. Sales Prediction
5. Customer Segmentation

BACKGROUND STORY

Rakamin Academy

Case Study:

- Requests from the Inventory Team that you assist in predicting the quantity of total Kalbe product sold so that the Inventory Team can make sufficient daily inventory.
- Requests from the Marketing Team to establish customer segmentation for the Marketing Team to use in providing personalized promotion and sales treatment.

EXPLORATORY DATA ANALYSIS (Page 1)

Rakamin Academy

Using

Challenge:

1. What is the average age of a customer based on marital status?

```
#query1
select "Marital Status", avg(age) as "Rata - Rata Umur"
from customer
group by "Marital Status"
```

Result:

	Marital Status	Rata - Rata Umur
1		31.3333333333
2	Married	43.0382352941
3	Single	29.3846153846

2. What is the average age of a customer based on gender?

```
#query2
select "gender", avg(age) as "Rata - Rata Umur"
from customer
group by "gender"
```

Result:

	gender	Rata - Rata Umur
1	0	40.326446281
2	1	39.1414634146

*P.S: 0 stand for Female & 1 stand for Male

EXPLORATORY DATA ANALYSIS (Page 2)

Rakamin Academy

Using

Challenge:

3. Specify the store name with the total quantity!

```
#query3
select store.storename,sum(transaction.qty) as "Total Quantity"
from store
inner join transaction
on store.storeid = transaction.storeid
group by store.storename
order by "Total Quantity" desc
limit 1;
```

Result:

	storename	Total Quantity
1	Lingga	2,777

4. Specify the name of the best-selling product with the total amount!

```
#Query4
select case_study."Product Name",sum(transaction.totalamount) as "Total Amount"
from case_study
inner join transaction
on case_study.productid = transaction.productid
group by case_study."Product Name"
order by "Total Amount" desc
limit 1;
```

Result:

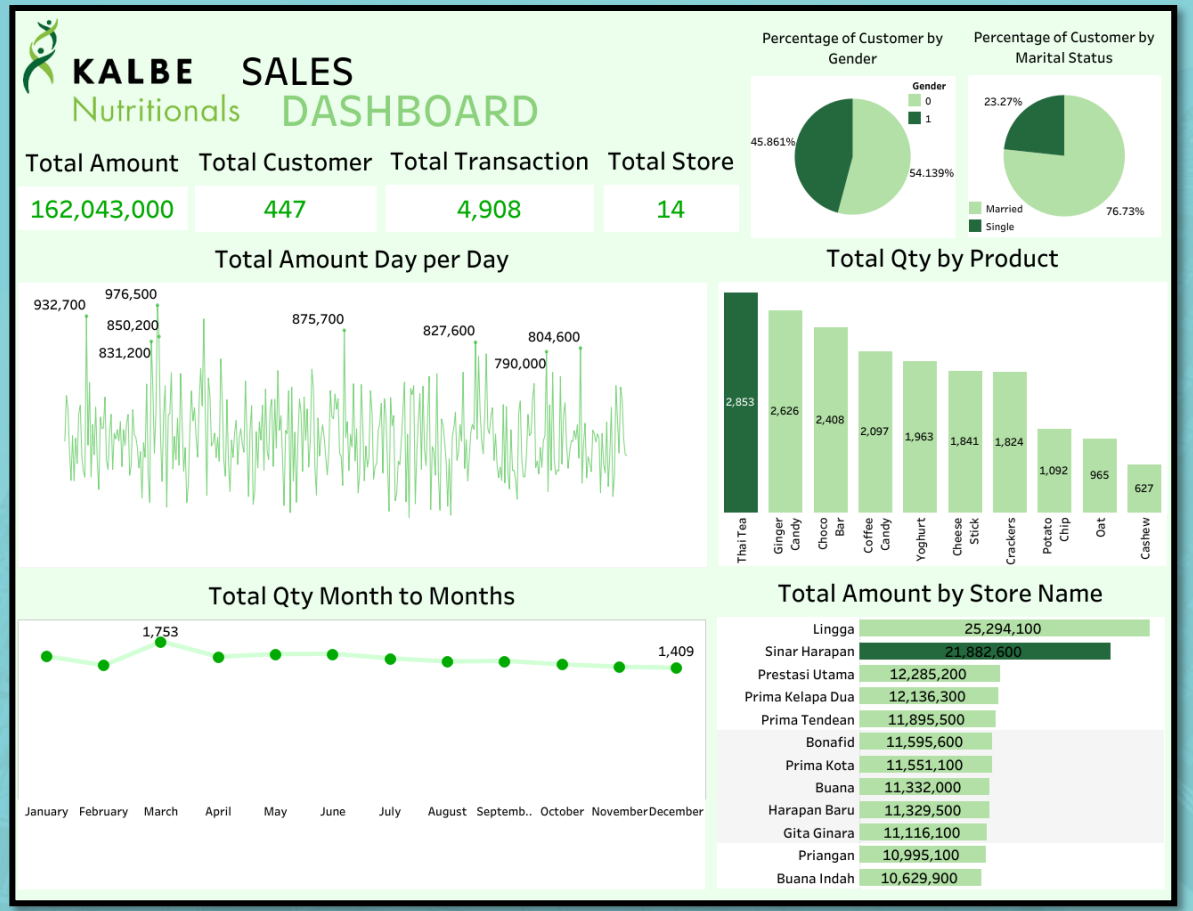
	Product Name	Total Amount
1	Cheese Stick	27,615,000

SALES DASHBOARD

Using

tableau

Rakamin Academy



Challenge:

1. Make Chart about Total Amount Day per Day.
2. Make Chart about Total Qty Month to Month.
3. Make Chart about Total Qty by Product.
4. Make Chart Total Amount by Store Name.

Link to Dashboard:
https://public.tableau.com/views/KalbeNutritionalSalesDashboardPBI-RakaminAcademy/SalesDashboard?:language=en-US&:display_count=n&:origin=viz_share_link

TOTAL SALES PREDICTION USING ARIMA

Rakamin Academy

STEPS:

1. Data Understanding
2. Data Cleaning
3. Built Model Machine Learning Using Time Series Analysis & Forecasting
4. Model Evaluation

Link to Script Code:
https://colab.research.google.com/drive/16jsbtcpw5jAZV_eihqouU8FIc9iCk1ku?usp=sharing

TOTAL SALES PREDICTION USING ARIMA (Page 1)

Rakamin Academy

1. Data Understanding

a Dataset Informastion

- Dataset ini terdiri dari 4 csv file yaitu customer, store, product dan transaction.
- Merupakan dummy data untuk studi kasus FMCG dalam kurun waktu 1 tahun yang diambil melalui program membership.

b Attribute Information

- Customer
  - `CustomerID`: No Unik Customer
  - `Age`: Usia Customer
  - `Gender`: 0 Wanita, 1 Pria
  - `Marital Status`: Married, Single (Blm menikah/Pernah menikah)
  - `Income` : Pendapatan per bulan dalam jutaan rupiah
- Store
  - `StoreID`: Kode Unik Store
  - `StoreName`: Nama Toko
  - `GroupStore`: Nama group
  - `Type`: Modern Trade, General Trade
  - `Latitude`: Kode Latitude
  - `Longitude`: Kode Longitude
- Product
  - `ProductID`: Kode Unik Product
  - `Product Name`: Nama Product
  - `Price`: Harga dlm rupiah
- Transaction
  - `TransactionID`: Kode Unik Transaksi
  - `Date`: Tanggal transaksi
  - `Qty`: Jumlah item yang dibeli
  - `Total Amount`: Price x Qty

c

Company Goals

- Kamu adalah seorang Data Scientist di Kalbe Nutritionals dan sedang mendapatkan project baru dari tim inventory.
- Dari tim inventory, kamu diminta untuk dapat membantu memprediksi jumlah penjualan (quantity) dari total keseluruhan product Kalbe

Objectives

- Untuk mengetahui perkiraan quantity product yang terjual sehingga tim inventory dapat membuat stock persediaan harian yang cukup.
- Prediksi yang dilakukan harus harian

d Data Shape

```
customer.shape, product.shape, store.shape, transaction.shape
```

```
((447, 5), (10, 3), (14, 6), (5020, 8))
```

TOTAL SALES PREDICTION USING ARIMA (Page 2)

Rakamin Academy

2. Data Cleaning

4.1. Data Merge

```
[ ] df = pd.merge(transaction,customer,on='CustomerID', how='inner')
    df = pd.merge(df,product,on='ProductID',how='inner')
    df = pd.merge(df,store,on='StoreID', how='inner')
    df.head()
```

	TransactionID	CustomerID	Date	ProductID	Price_x	Qty
0	TR11389	328	01/01/2022	P3	7500	4
1	TR89318	183	17/07/2022	P3	7500	1
2	TR9106	123	26/09/2022	P3	7500	4
3	TR4331	335	08/01/2022	P3	7500	3
4	TR8445	181	10/01/2022	P3	7500	4

4.2. Data Shape

```
[ ] df.shape
    (5020, 19)
```

Data Type Checking

```
df.dtypes
```

```
TransactionID     object
CustomerID         int64
Date              object
ProductID         object
Price_x            int64
Qty                int64
TotalAmount        int64
StoreID            int64
Age                int64
Gender             int64
Marital Status    object
Income            object
Product Name      object
Price_y            int64
StoreName         object
GroupStore        object
Type              object
Latitude          object
Longitude         object
dtype: object
```

Data Duplicates Checking

```
df.duplicated().sum()
```

```
0
```

Missing Value Checking

```
df.isna().sum()
```

```
TransactionID      0
CustomerID         0
Date               0
ProductID          0
Price_x            0
Qty                0
TotalAmount        0
StoreID            0
Age                0
Gender             0
Marital Status    44
Income             0
Product Name       0
Price_y            0
StoreName          0
GroupStore         0
Type               0
Latitude           0
Longitude          0
dtype: int64
```

4.2. Data Imputation with Mode

```
df['Marital Status'] = df['Marital Status'].fillna(df['Marital Status'].mode()[0])
```

```
df['Marital Status'].unique()
```

```
array(['Married', 'Single'], dtype=object)
```

```
df.isna().sum()
```

```
TransactionID     0
CustomerID        0
Date              0
ProductID         0
Price_x           0
Qty               0
TotalAmount       0
StoreID           0
Age               0
Gender            0
Marital Status    0
Income            0
Product Name      0
Price_y           0
StoreName         0
GroupStore        0
Type              0
Latitude          0
Longitude         0
dtype: int64
```

TOTAL SALES PREDICTION USING ARIMA (Page 3)

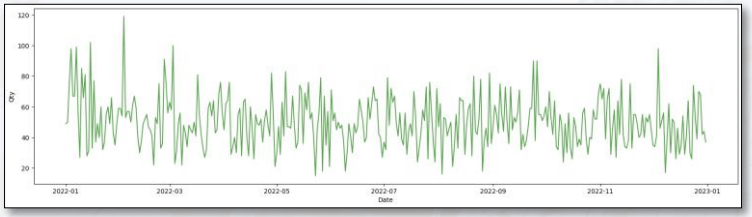
Rakamin Academy

3. Built Model Machine Learning Using Time Series Analysis & Forecasting

1. Data Preparation

```
data_tsa = df.groupby(['Date']).agg({
    'Qty' : 'sum'
}).reset_index()
```

a Data preparation by using 2 features which is Date and Qty



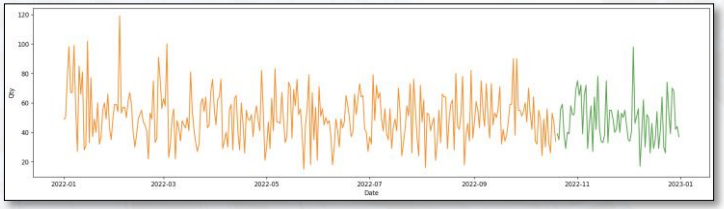
b Time Series Plot

. Stationarity Test

```
result = adfuller(data_tsa['Qty'])
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])
print('Critical Values:')
for key, value in result[4].items():
    print('\t%s: %.3f' % (key, value))
```

```
ADF Statistic: -19.018783
p-value: 0.000000
Critical Values:
        1%: -3.448
        5%: -2.870
        10%: -2.571
```

c Doing Stationary Test (Augmented-Dicky Fuller Test), the results said that P-Value < 0,05 so that the data can be used.



d Train-Test Split

```
# train-test split
cut_off = round(data_tsa.shape[0]*0.8)
df_train = data_tsa[:cut_off]
df_test = data_tsa[cut_off:].reset_index(drop=True)
df_train.shape, df_test.shape
```

```
((292, 2), (73, 2))
```

TOTAL SALES PREDICTION USING ARIMA (Page 4)

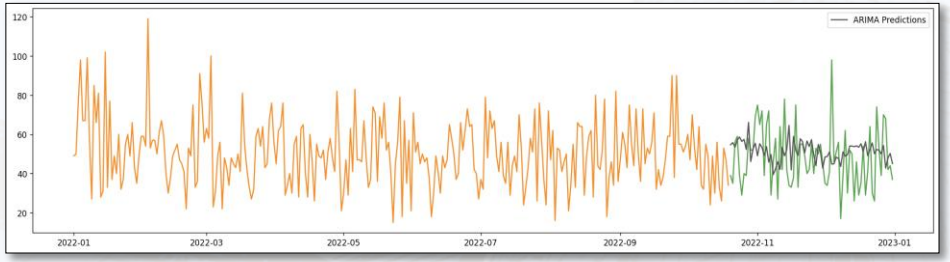
Rakamin Academy

4. Model Evaluation

a Using RMSE and MAE for the Model Evaluation, the results show →

```
RMSE value 16.80722771118357
MAE value 13.760085004301152
```

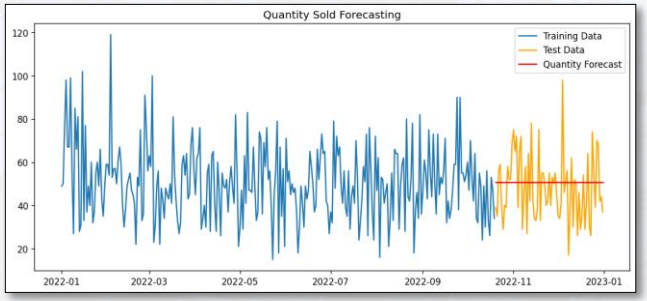
The result of Forecasting Plot



b Auto arima best model and prediction results

```
# auto arima best model and prediction results

model = auto_arima(df_train, trace=True, error_action='ignore', suppress_warnings=True)
model.fit(df_train)
forecast = model.predict(n_periods=len(df_test))
forecast = pd.DataFrame(forecast, index=df_test.index, columns=['Predictions'])
print(forecast)

Performing stepwise search to minimize aic
 ARIMA(2,0,2)(0,0,0)[0] intercept   : AIC=2492.660, Time=3.50 sec
 ARIMA(0,0,0)(0,0,0)[0] intercept   : AIC=2486.299, Time=0.11 sec
 ARIMA(1,0,0)(0,0,0)[0] intercept   : AIC=2488.299, Time=0.25 sec
 ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=2488.299, Time=0.60 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=3153.727, Time=0.09 sec
 ARIMA(1,0,1)(0,0,0)[0] intercept   : AIC=2490.294, Time=0.56 sec

Best model:  ARIMA(0,0,0)(0,0,0)[0] intercept
Total fit time: 5.173 seconds
            Predictions
Date
2022-10-20    50.633562
2022-10-21    50.633562
2022-10-22    50.633562
2022-10-23    50.633562
2022-10-24    50.633562
...                 ...
2022-12-27    50.633562
2022-12-28    50.633562
2022-12-29    50.633562
2022-12-30    50.633562
2022-12-31    50.633562

[73 rows x 1 columns]
```

c Forecasting Results




```
forecast.mean()

Predictions    50.633562
dtype: float64
```

CUSTOMER SEMENTATION USING K-MEANS CLUSTERING ALGORITHM

Rakamin Academy

STEPS:

1. Data Understanding
2. Data Correlation Using Heatmap
3. Built Model Machine Learning Using K-Means Clustering
4. The Results of Customer Segmentation

Link to Script Code:
https://colab.research.google.com/drive/1avqbcqsOpBFCjaKVOOuqAyZH_6SpY-2e?usp=sharing

CUSTOMER SEMENTATION USING K-MEANS CLUSTERING ALGORITHM (Page 1)

Rakamin Academy

1. Data Understanding

a Dataset Informastion

- Dataset ini terdiri dari 4 csv file yaitu customer, store, product dan transaction.
- Merupakan dummy data untuk studi kasus FMCG dalam kurun waktu 1 tahun yang diambil melalui program membership.

b Attribute Information

- Customer
  - `CustomerID`: No Unik Customer
  - `Age`: Usia Customer
  - `Gender`: 0 Wanita, 1 Pria
  - `Marital Status`: Married, Single (Blm menikah/Pernah menikah)
  - `Income` : Pendapatan per bulan dalam jutaan rupiah
- Store
  - `StoreID`: Kode Unik Store
  - `StoreName`: Nama Toko
  - `GroupStore`: Nama group
  - `Type`: Modern Trade, General Trade
  - `Latitude`: Kode Latitude
  - `Longitude`: Kode Longitude
- Product
  - `ProductID`: Kode Unik Product
  - `Product Name`: Nama Product
  - `Price`: Harga dlm rupiah
- Transaction
  - `TransactionID`: Kode Unik Transaksi
  - `Date`: Tanggal transaksi
  - `Qty`: Jumlah item yang dibeli
  - `Total Amount`: Price x Qty

c

Company Goals

- Kamu adalah seorang Data Scientist di Kalbe Nutritionals dan sedang mendapatkan project baru dari tim inventory.
- Dari tim inventory, kamu diminta untuk dapat membantu memprediksi jumlah penjualan (quantity) dari total keseluruhan product Kalbe

Objectives

- Untuk mengetahui perkiraan quantity product yang terjual sehingga tim inventory dapat membuat stock persediaan harian yang cukup.
- Prediksi yang dilakukan harus harian

d Data Shape

```
customer.shape, product.shape, store.shape, transaction.shape
```

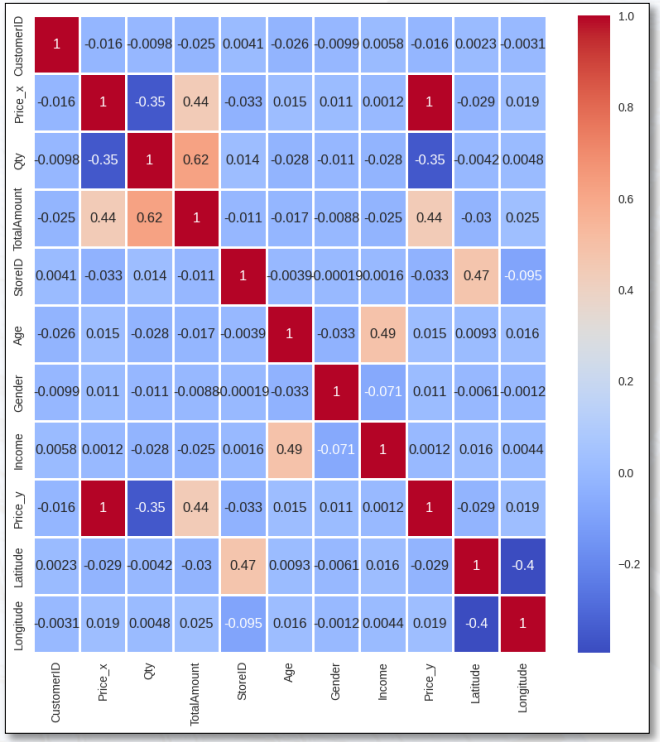
```
((447, 5), (10, 3), (14, 6), (5020, 8))
```

CUSTOMER SEMENTATION USING K-MEANS CLUSTERING ALGORITHM (Page 2)

2. Data Correlation Using Heatmap

```
df.corr()
```

	CustomerID	Price_x	Qty	TotalAmount	StoreID	Age	Gender	Income	Price_y	Latitude	Longitude
CustomerID	1.000000	-0.016423	-0.009755	-0.024915	0.004129	-0.025952	-0.009947	0.005783	-0.016423	0.002278	-0.003122
Price_x	-0.016423	1.000000	-0.353640	0.440632	-0.032863	0.014693	0.010705	0.001196	1.000000	-0.029008	0.018652
Qty	-0.009755	-0.353640	1.000000	0.621129	0.014365	-0.027768	-0.010542	-0.028425	-0.353640	-0.004170	0.004807
TotalAmount	-0.024915	0.440632	0.621129	1.000000	-0.010722	-0.016900	-0.008774	-0.025350	0.440632	-0.029938	0.025437
StoreID	0.004129	-0.032863	0.014365	-0.010722	1.000000	-0.003872	-0.000189	0.001613	-0.032863	0.471852	-0.094943
Age	-0.025952	0.014693	-0.027768	-0.016900	-0.003872	1.000000	-0.033183	0.486692	0.014693	0.009266	0.015951
Gender	-0.009947	0.010705	-0.010542	-0.008774	-0.000189	-0.033183	1.000000	-0.071443	0.010705	-0.006051	-0.001183
Income	0.005783	0.001196	-0.028425	-0.025350	0.001613	0.486692	-0.071443	1.000000	0.001196	0.015518	0.004385
Price_y	-0.016423	1.000000	-0.353640	0.440632	-0.032863	0.014693	0.010705	0.001196	1.000000	-0.029008	0.018652
Latitude	0.002278	-0.029008	-0.004170	-0.029938	0.471852	0.009266	-0.006051	0.015518	-0.029008	1.000000	-0.395995
Longitude	-0.003122	0.018652	0.004807	0.025437	-0.094943	0.015951	-0.001183	0.004385	0.018652	-0.395995	1.000000



CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING ALGORITHM (Page 3)

Rakamin Academy

3. Built Model Machine Learning Using K-Means Clustering

a

3.4. Data Aggregation (Data Clustering)

Membuat data baru untuk clustering, yaitu groupby by customerID lalu yang di aggregasi adalah :

- Transaction id count
- Qty sum
- Total amount sum

```
df = df.groupby(['CustomerID']).agg({'TransactionID' : 'count','Qty' : 'sum','TotalAm

df.head()
```

	CustomerID	TransactionID	Qty	TotalAmount
0	1	17	60	623300
1	2	13	57	392300
2	3	15	56	446200
3	4	10	46	302500
4	5	7	27	268600

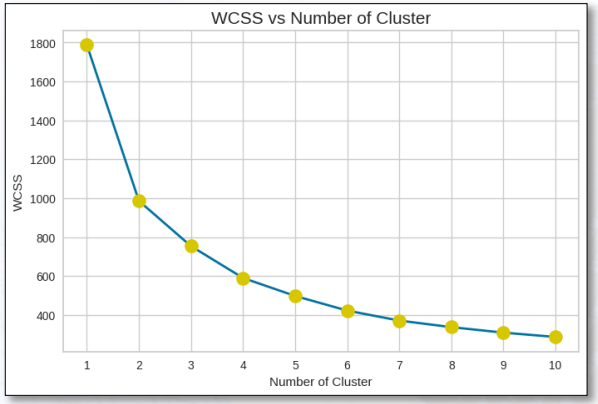
b

4.1. Standarize and Normalize Data

	CustomerID	TransactionID	Qty	TotalAmount
0	-1.728180	1.779816	1.496527	2.094768
1	-1.720431	0.545884	1.261093	0.239269
2	-1.712681	1.162850	1.182615	0.672218
3	-1.704931	-0.379565	0.397833	-0.482047
4	-1.697182	-1.305014	-1.093251	-0.754347
...	...	...	...	...
442	1.697182	1.471333	1.418049	0.984681
443	1.704931	2.088298	1.653484	1.728488
444	1.712681	2.088298	2.124352	1.804796
445	1.720431	-0.071082	0.083921	0.488275
446	1.728180	0.545884	0.083921	0.616794

447 rows × 4 columns

c Determining the number of cluster
Elbow Method



d Sillhoutte Method, the results show that 2 Cuslters show the most efficient method but I am choosing Customer to be 3 Clusters to see the data variation.

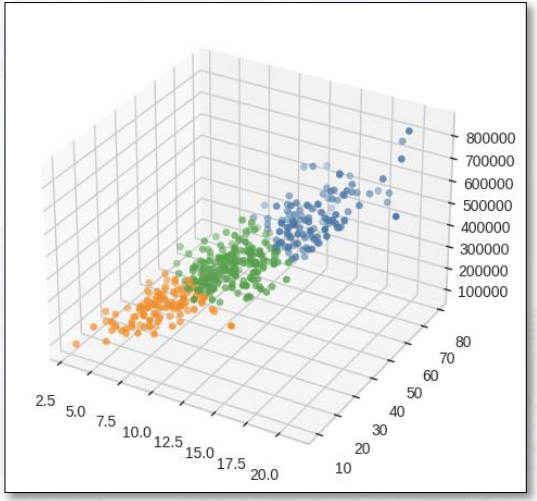
3. Sillhouette Method

```
# Silhouette Analysis
range_n_clusters = [2, 3, 4, 5, 6, 7, 8]
for num_clusters in range_n_clusters:

    # Initialise kmeans
    kmeans = KMeans(n_clusters=num_clusters, max_iter=500)
    kmeans.fit(df)
    cluster_labels = kmeans.labels_

    # Silhouette Score
    silhouette_avg = silhouette_score(df, cluster_labels)
    print("For n_clusters={0}, the silhouette score is {1}".for
```

```
For n_clusters=2, the silhouette score is 0.5634425108616966
For n_clusters=3, the silhouette score is 0.5408006989949187
For n_clusters=4, the silhouette score is 0.5201142404920729
For n_clusters=5, the silhouette score is 0.5388948343949318
For n_clusters=6, the silhouette score is 0.5281138669883647
For n_clusters=7, the silhouette score is 0.5371695505890094
For n_clusters=8, the silhouette score is 0.5375429562500206
```

CUSTOMER SEMENTATION USING K-MEANS CLUSTERING ALGORITHM (Page 4)

4. The Results of Customer Segmentation



cluster	CustomerID	TransactionID	Qty	TotalAmount
1	217	11.023041	40.073733	353001.382488
2	117	15.239316	57.145299	518861.538462
0	113	7.477876	25.787611	218892.920354

Based on the results of Customer Segmentation using K-Means that customer segmentation is divided into 3 Clusters, ie;

1. Cluster 2: Customers with the highest amount of product purchases and transaction value.

Suggestion treatment:

- Give Exclusive promo like giving a reward loyalty by going extra mile, Offer special incentives to your VIPs, Give premium quality, and Give special treatment like fast respons.

2. Cluster 1: Customers with average amount of product purchases and transaction value.

Suggestion treatment:

- Improve your loyalty programs (like: providing early access to products and services, free or expedited shipping, two-for-one discounts, even the occasional free product or service, provide experiences to loyalty program members that leave them feeling emotionally satisfied)
- Listen to customer feedback and act on it

3. Cluster 0: Customers with the lowest amount of product purchases and transaction value

Suggestion treatment:

- Give a Discounts
- Give a Gifts

Link Github

https://github.com/jaelaniuwahyu/Kalbe-Total-Sales-Prediction-and-Customer-Segmentation-Using-ARIMA-and-K-Means-Clustering-Algorithm

Rakamin Academy

Video Presentation

https://drive.google.com/file/d/1jY-KCn71myAbNNNISy2t4u1XWeWNPsGD/view?usp=sharing

Thank You