Rakamin Academy X KALBE Nutritionals

Customer Segmentation and Forecasting Sales Quantity

Kalbe Nutritionals Data Science

Presented by
Tito

Case Study

You are currently getting a new project from the **inventory team** and **marketing team**.

- From the **marketing team** you are asked to create customer clusters/segments based on several criteria.
- From the **inventory team**, you are asked to help predicting the number of sales (quantity) from the total all Kalbe products.

Dataset

Two types of customer segmentation are conducted, **Behavioral Segmentation** and **Demographic Segmentation**.

CustomerID	Frequency	Total_Quantity	Total_Amount_Sales	Recency	Age	Gender	Marital Status	Income
1	17	60	623300	53	55	Man	Married	5.12
2	13	57	392300	97	60	Man	Married	6.23
3	15	56	446200	10	32	Man	Married	9.17
4	10	46	302500	5	31	Man	Married	4.87
5	7	27	268600	91	58	Man	Married	3.57
...	...	...	...	...	...	...	...	...
443	16	59	485100	63	33	Man	Married	9.28
444	18	62	577700	38	53	Woman	Married	15.31
445	18	68	587200	42	51	Woman	Married	14.48
446	11	42	423300	11	57	Woman	Married	7.81
447	13	42	439300	26	54	Man	Married	20.37

Behavioral data
- Number of transactions.
- Total Quantity.
- Total Amount Sales.
- Number of days since last transaction.

KMeans

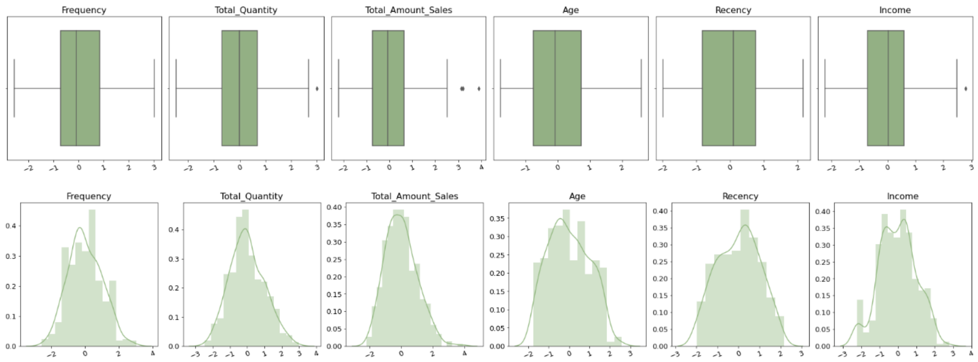
Demographic data
- Age
- Gender
- Marital Status
- Income

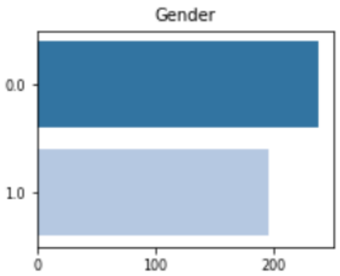
KPrototypes

Data Cleaning and Preprocessing

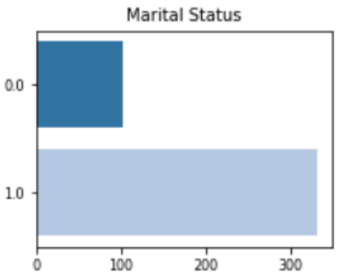
- Removing outliers
- Data transformation to make the distribution close to normal
- Standardization to make the features space having the same scale
- Label Encoding for categorical variables

	Frequency	Total_Quantity	Total_Amount_Sales	Age	Recency	Income	Marital Status	Gender
0	1.773511	1.482019	2.083088	1.218149	0.989894	-0.492632	1.0	1.0
1	0.528691	1.245034	0.213660	1.621540	1.783630	-0.245929	1.0	1.0
2	1.151101	1.166039	0.649860	-0.637451	-0.703350	0.323560	1.0	1.0
3	-0.404924	0.376089	-0.513070	-0.718129	-1.214207	-0.551401	1.0	1.0
4	-1.338539	-1.124817	-0.787415	1.460184	1.694733	-0.880941	1.0	1.0
...	...	...	...	...	...	...	...	...
430	1.462306	1.403024	0.964668	-0.556773	1.206197	0.343043	1.0	1.0
431	2.084716	1.640010	1.714058	1.056792	0.596408	1.278902	1.0	0.0
432	2.084716	2.113980	1.790939	0.895436	0.711674	1.162541	1.0	0.0
433	-0.093719	0.060109	0.464536	1.379505	-0.624566	0.072791	1.0	0.0
434	0.528691	0.060109	0.594020	1.137471	0.182779	1.928694	1.0	1.0





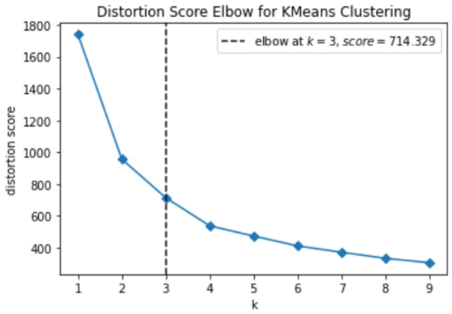
1 is Man. 0 is Woman

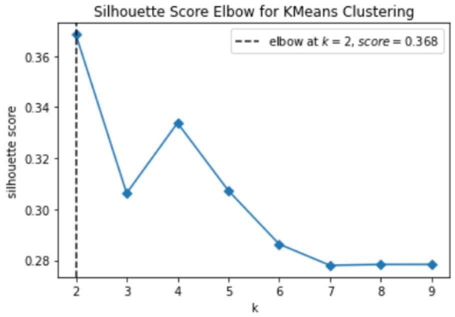


1 is Married. 0 is Single

Modeling – Elbow Method

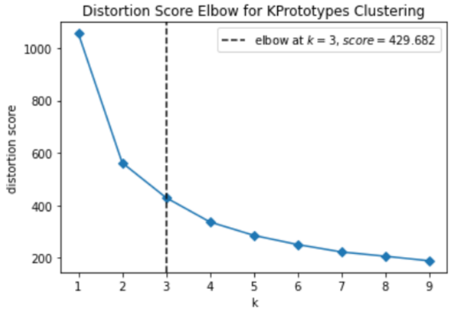
Behavioral segmentation (Kmeans)

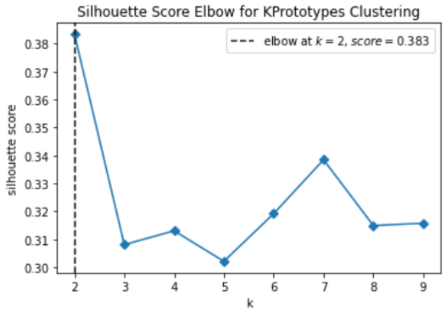




- The recommended number of clusters are 3 or 4 for both models based on the the elbow method.
- 4 clusters are chosen since the Silhouette score is higher compared to 3 clusters.

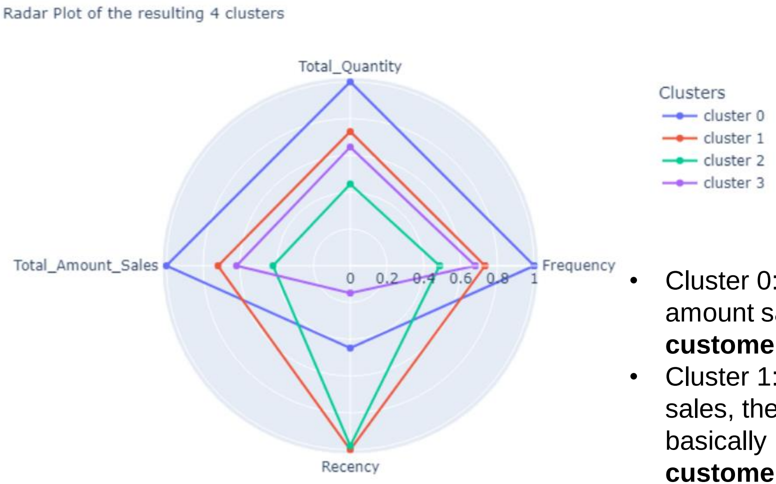
Demographic segmentation (KPrototypes)





Modeling – Interpretation

Behavioral segmentation (Kmeans)



	Recency	Frequency	Total_Quantity	Total_Amount_Sales	
	mean	mean	mean	mean	count
Behavioral_Segmentation					
Diamond	20.42	15.48	57.92	524915.38	104
Gold	45.63	11.34	42.26	377484.96	133
Potential	6.78	10.52	37.38	324564.08	103
Lost	44.65	7.52	25.74	220411.58	95

- Cluster 0: Diamond customer segment. They have spent the most in term of quantity and total amount sales, the most frequent buyers, and have purchased quite recently. **This is the best customers so far.**
- Cluster 1: Gold customer segment. They have spent the second most quantity and total amount sales, the second most frequent buyers, BUT have purchased a long time ago. They are basically near Diamond customers level but haven't purchased anything for a long time. **This customers almost got churned.**
- Cluster 2: Lost customer segment. They have spent the least in term of quantity and total amount sales. The least frequent buyers, and also haven't purchased anything long time ago. **This is most likely the churned customers**.
- Cluster 3: Potential customer segment. They have quite good spending in term of quantity and total amount sales, quite purchase frequently, and the most recent buyers. **This customers have a lot of potential to improve.**

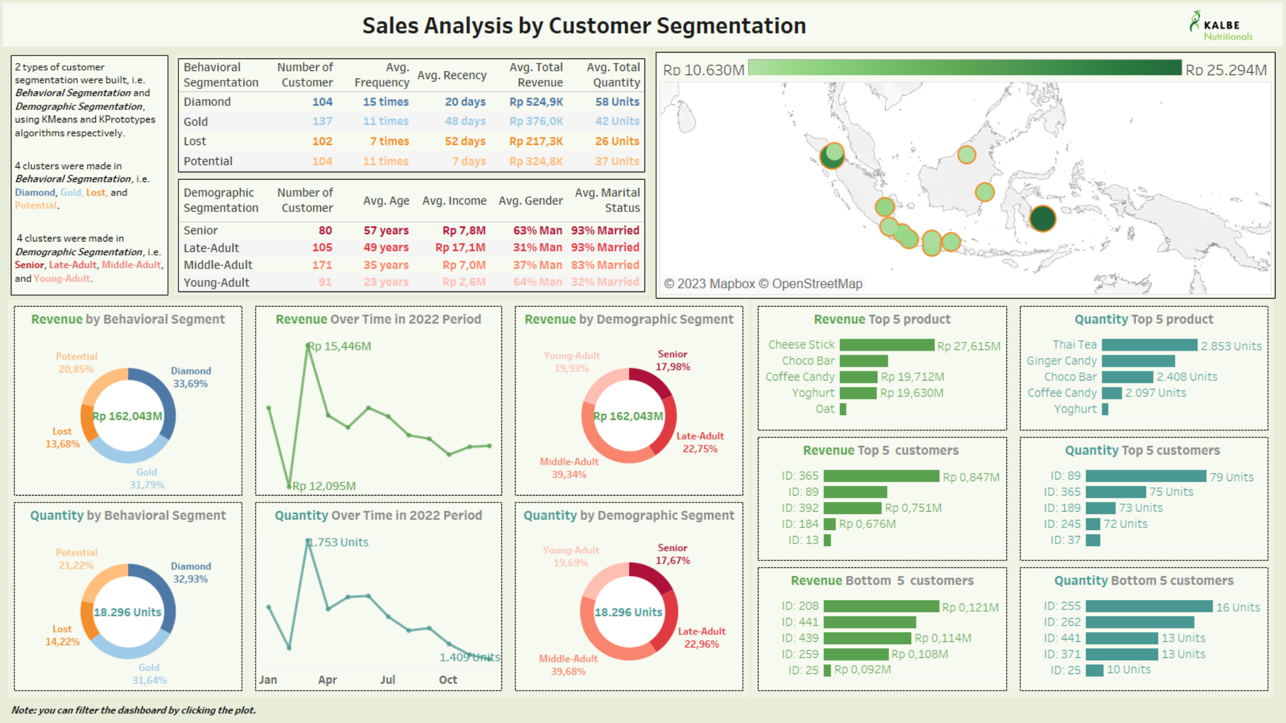
Modeling – Interpretation

Demographic segmentation (KPrototypes)

Demographic_Segmentation	Age mean	Income mean	Marital Status mean	Gender mean	Gender count
Senior	57.03	7.82	0.92	0.62	78
Late-Adult	48.51	15.86	0.94	0.30	99
Middle-Adult	35.16	6.96	0.83	0.38	172
Young-Adult	23.95	2.46	0.29	0.62	86

- Cluster 0: Senior customer segment. They are elder customers having been married with medium income. Most of them are Man. Most likely they are retired worker.
- Cluster 1: Middle-Adult customer segment. Their age is within 30-40 years old, having been married with medium income. Most of them are Woman.
- Cluster 2: Young-Adult customer segment. They are most likely teenager with range 20-30 years old, which are most likely single and low income. Most of them are Man.
- Cluster 3: Late-Adult customer segment. They are in mature age within 40-55 years old, having been married with high income. Most of them are Woman.

Dashboard



You can explore further using a dashboard created on Tableau Public.

https://public.tableau.com/app/profile/tito5892/viz/shared/CG5HZHQRK

Strategy

Behavioral Segmentation

Diamond customer segment. (Focus on increasing frequency and retention)
- Loyalty programs and give rewards/promo to make them feel respected,
- Market most expensive products,
- Offer new products, and cross-selling/up-selling strategy.

Gold customer segment. (Focus on maintaining their loyalty and improve their value)
- Offer them a discount, free trial, or another incentive.
- Make limited time offers.

Lost customer segment. (Focus on reactivating the customer)
- Reactivation strategy such as send them reactivation emails and ask them for feedback.
- Provide support.

Potential customer segment. (Focus on increasing their value)
- Cross-selling/up-selling strategy, give price incentives and new products recomendation.

Demographic Segmentation

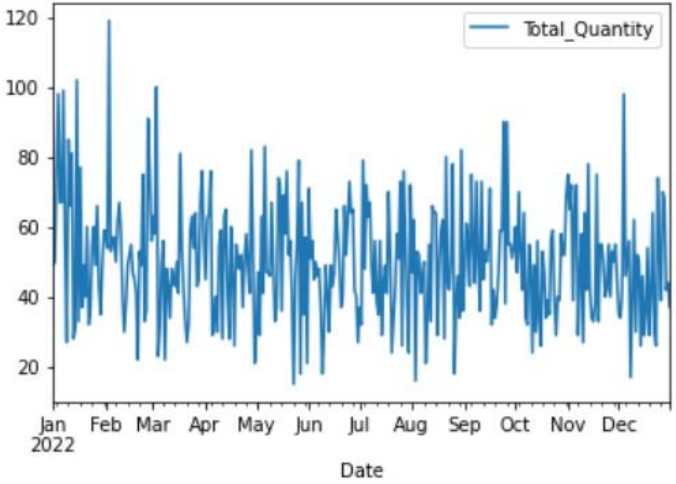
- Adjust pricing strategies based on income levels.
- Adapt products or services to satisfy to the needs and preferences of different demographic segments.

Dataset

The transaction dataset (5020 rows) are grouped by day and the quantity are summed for each group resulting daily total quantity of all products sold in 2022 within 365 days (365 rows).

Our time-series is already stationary

Date	Total_Quantity
2022-01-01	49
2022-01-02	50
2022-01-03	76
2022-01-04	98
2022-01-05	67
...	...
2022-12-27	70
2022-12-28	68
2022-12-29	42
2022-12-30	44
2022-12-31	37

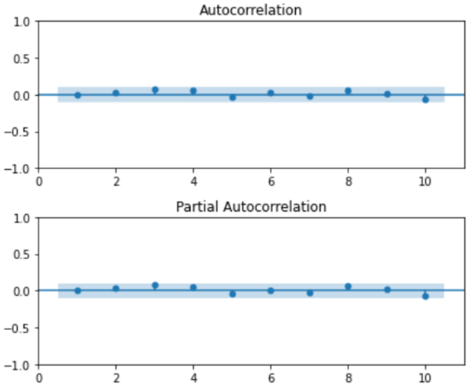



```
adfuller test:
p-values: 0.0
Reject H0 ---> time series is stationary

KPSS test:
p-values: 0.06622747821677664
Accept H0 ---> time series is stationary
```

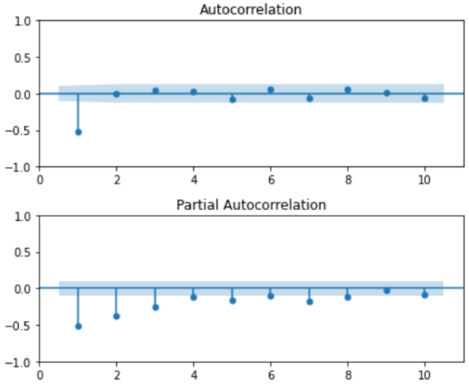
ACF and PACF

Original Time Series



Using the original data, There is no significant autocorrelation. This is called “**White Noise**”. In this case our time-series is stationary, yet zero autocorrelations at all lags are found.

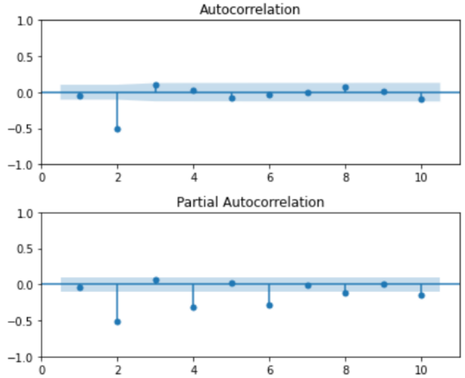
d = 1




```
adfuller test:
p-values: 6.148467265279569e-21
Reject H0 ---> time series is stationary

KPSS test:
p-values: 0.1
Accept H0 ---> time series is stationary
```

d = 2




```
adfuller test:
p-values: 3.4821335991476044e-11
Reject H0 ---> time series is stationary

KPSS test:
p-values: 0.1
Accept H0 ---> time series is stationary
```

- With **d** = 1 or **d** = 2, our time-series now have autocorrelation greater than zero.
- From ACF and PACF, It suggest that the candidates for parameters **p, q,** and **d** as follows:

1. **d** can be 1 or 2,
2. **p** can be 1 to 6,
3. **q** can be 1 to 2

Build ARIMA Model

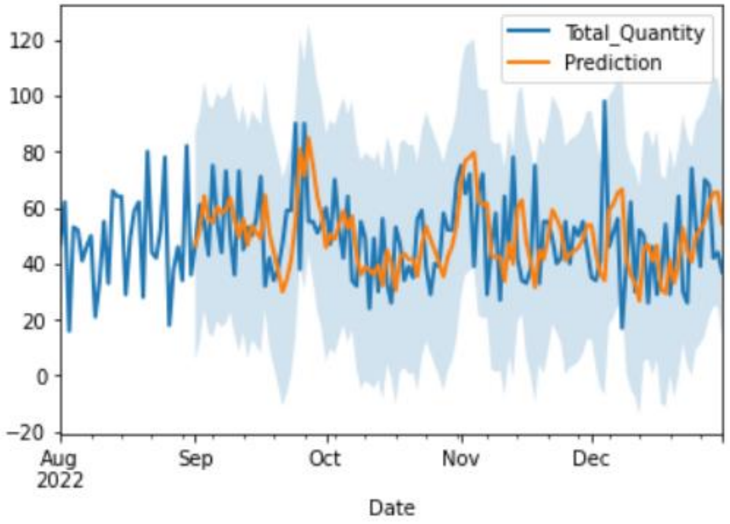
```
Performing stepwise search to minimize aic
 ARIMA(2,2,2)(0,0,0)[0]             : AIC=inf, Time=0.55 sec
 ARIMA(0,2,0)(0,0,0)[0]             : AIC=3731.707, Time=0.02 sec
 ARIMA(1,2,0)(0,0,0)[0]             : AIC=3518.903, Time=0.03 sec
 ARIMA(0,2,1)(0,0,0)[0]             : AIC=inf, Time=0.09 sec
 ARIMA(2,2,0)(0,0,0)[0]             : AIC=3401.324, Time=0.05 sec
 ARIMA(3,2,0)(0,0,0)[0]             : AIC=3319.933, Time=0.07 sec
 ARIMA(4,2,0)(0,0,0)[0]             : AIC=3287.880, Time=0.11 sec
 ARIMA(5,2,0)(0,0,0)[0]             : AIC=3259.857, Time=0.10 sec
 ARIMA(6,2,0)(0,0,0)[0]             : AIC=3249.721, Time=0.19 sec
 ARIMA(6,2,1)(0,0,0)[0]             : AIC=inf, Time=0.82 sec
 ARIMA(5,2,1)(0,0,0)[0]             : AIC=inf, Time=0.78 sec
 ARIMA(6,2,0)(0,0,0)[0] intercept   : AIC=3251.685, Time=0.43 sec

Best model:  ARIMA(6,2,0)(0,0,0)[0]
Total fit time: 3.226 seconds
```

```
In [50]: # build ARIMA model
         from statsmodels.tsa.arima.model import ARIMA
         model = ARIMA(df_daily, order = (6,2,0))
         result = model.fit()
```

Using pm.auto_arima library we should use ARIMA(6, 2, 0) in which gives the optimize evaluation metrics

Model Evaluation



In [79]:

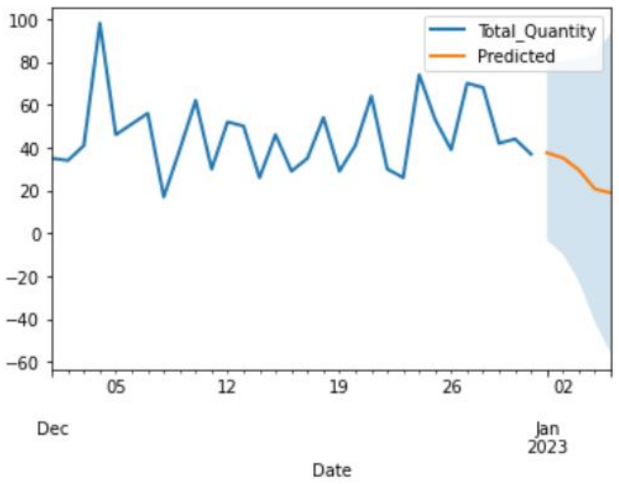
```
# calculate performance metric
from sklearn.metrics import mean_squared_error, mean_absolute_error
y_true = df_daily.values
y_pred = result.get_prediction(start = df_daily.index[0], dynamic = False).predicted_mean.values

MAE = mean_absolute_error(y_true, y_pred)
RMSE = mean_squared_error(y_true, y_pred, squared = False)
print('MAE: {:.2f} units'.format(MAE))
print('RMSE: {:.2f} units'.format(RMSE))
```

```
MAE: 16.84 units
RMSE: 21.01 units
```

The model is far from good, but it can follow general upward and downward trends

Forecast The Future



	Predicted
2023-01-01	37.606836
2023-01-02	35.212744
2023-01-03	29.564290
2023-01-04	20.710638
2023-01-05	18.881172

The model said that the total quantity of products will be declining
5 days later from the end of December

https://github.com/dstito/Customer-Segmentation-and-Forecasting-Sales-Quantity

Thank You

Rakamin Academy X KALBE Nutritionals