Rakamin Academy X KALBE

Kalbe Nutritionals Data Scientist

Project Based Internship Program

Machine Learning Project Using ARIMA and K-Means

Presented By Radisha Fanni Sianti

About Me

“ I am a fresh graduate of Master of Statistics and Bachelor of Mathematics from Institut Teknologi Sepuluh Nopember. I have a passion for continuing to learn and develop myself by constantly trying to broaden my knowledge in statistics and other related disciplines. I’m enthusiastic about deriving valuable insights from data and leveraging them to facilitate well-informed decision-making. ”

Table of contents

01 **Overview**

Background & Project Description

02 **Visualization**

Create a dashboard.

03 **Data Preprocessing**

Handling Data Type, Duplicated, Missing Value.

04 **EDA**

Exploratory Data Analysis.

05 **Machine Learning**

Using ARIMA and K-Means

06 **Insight & Recommendations**

Provide Recommendations based on analysis

01

Overview

Overview

This project comes from the inventory team and marketing team of Kalbe Nutritionals.

The inventory team wants to know the estimated quantity of products sold so that the inventory team can create sufficient daily inventory. To overcome this problem, we can use the machine learning method to forecast product quantity time series using the ARIMA model.

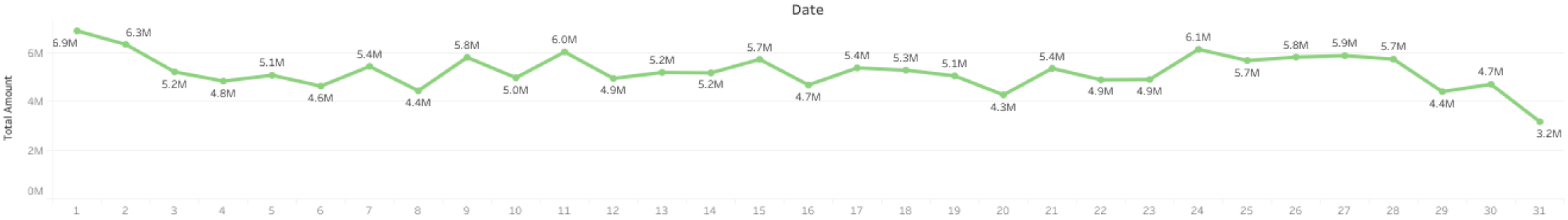
The marketing team wants to segment customers based on several criteria, which will later be used to provide personalized promotion and sales treatment. To overcome this problem, we can utilize the machine learning clustering method using the K-Means model.

02

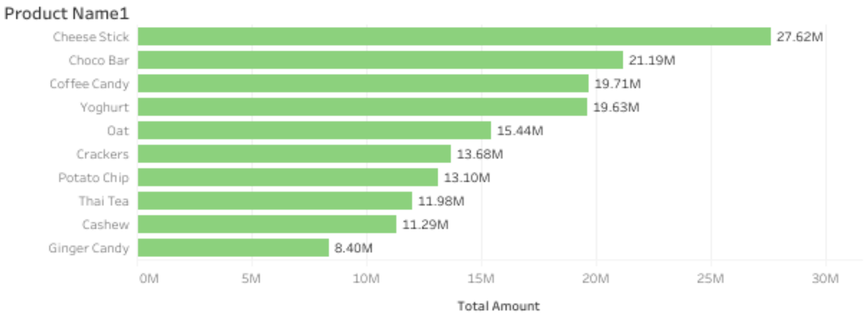
Visualization

Sales Performance Dashboard

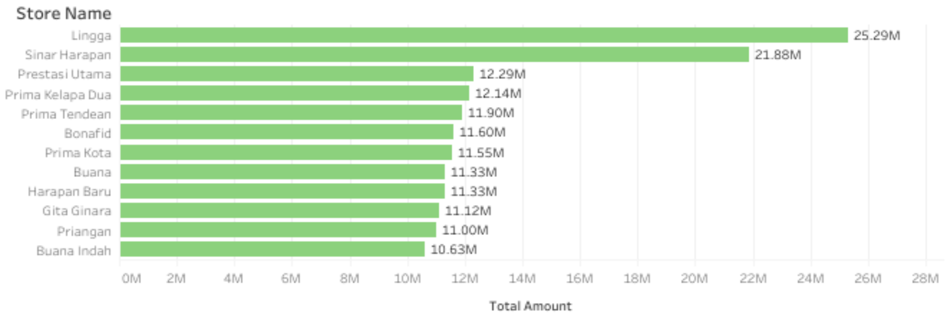
Total Amount Per Day



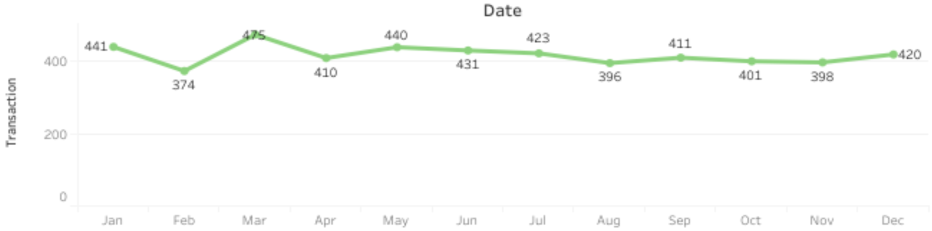
Total Amount By Product



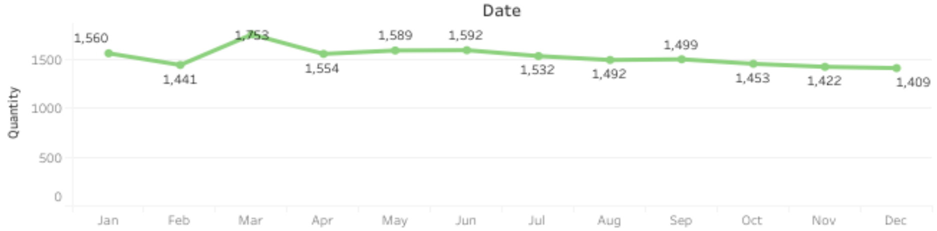
Total Amount By Store Name



Total Transaction By Month



Total Quantity Per Month



03

Data Preprocessing

Data Preprocessing

1. **Check Data Type**

   “Date” column should be of the datetime data type not object.

2. **Check Duplicated Data**

   There is no duplicated data.

3. **Check Missing Value**

   There is no missing value.

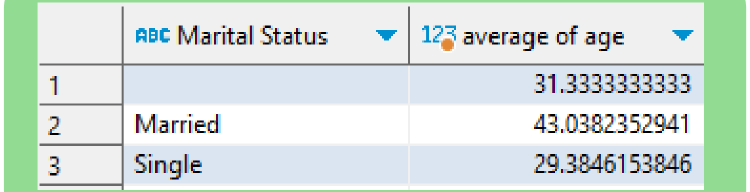
4. **Preparing Data for Machine Learning Modeling**

   - ARIMA Modeling using df_inventory (Date, Qty).
   - K-Means Modeling using df_marketing (CustomerID, Total_Transaction, Total_Quantity, Total_Amount)

04

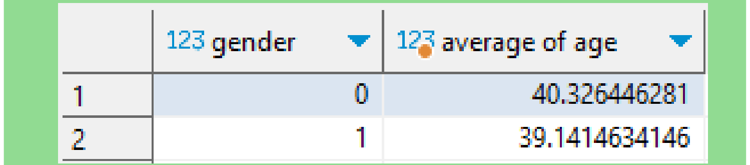
EDA

Average Customer Age Based on Marital Status



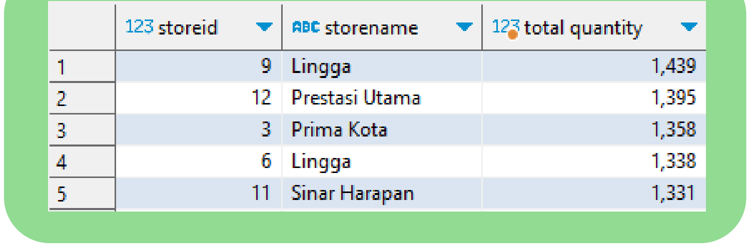
	Marital Status	average of age
1		31.3333333333
2	Married	43.0382352941
3	Single	29.3846153846

Average Customer Age Based on Gender



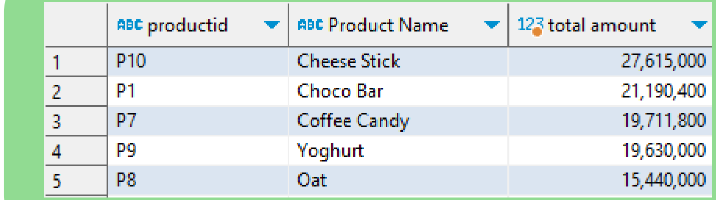
	gender	average of age
1	0	40.326446281
2	1	39.1414634146

Total Quantity Based On Store Name



	storeid	storename	total quantity
1	9	Lingga	1,439
2	12	Prestasi Utama	1,395
3	3	Prima Kota	1,358
4	6	Lingga	1,338
5	11	Sinar Harapan	1,331

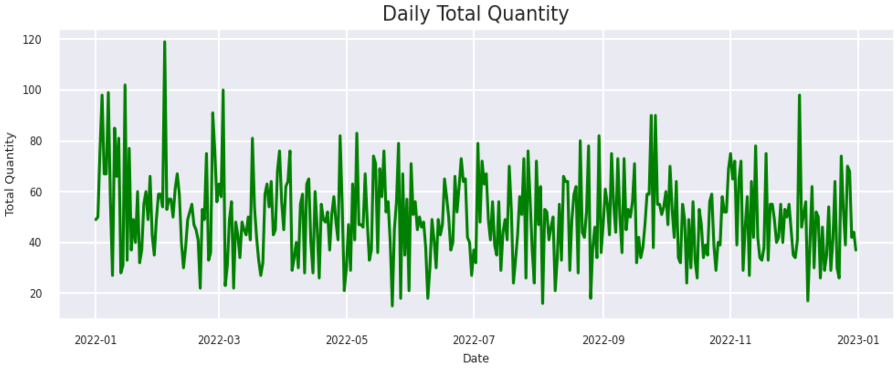
Total Amount Based On Product Name



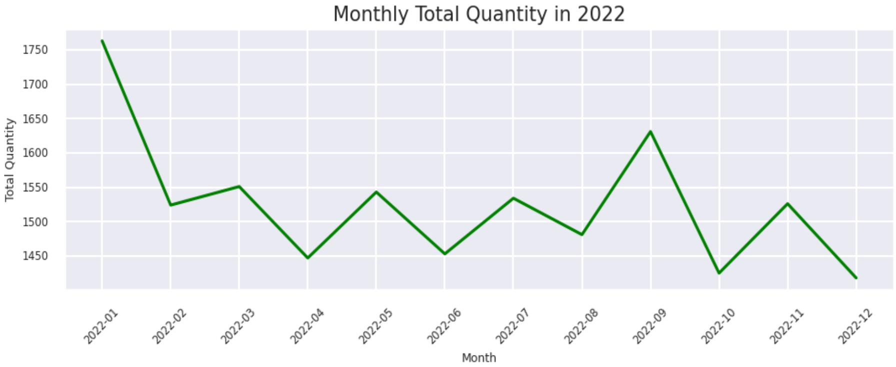
	productid	Product Name	total amount
1	P10	Cheese Stick	27,615,000
2	P1	Choco Bar	21,190,400
3	P7	Coffee Candy	19,711,800
4	P9	Yoghurt	19,630,000
5	P8	Oat	15,440,000

EDA Inventory Data

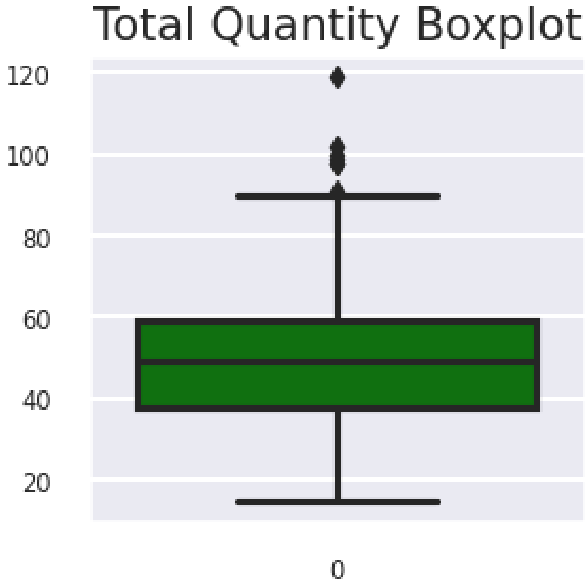
Daily



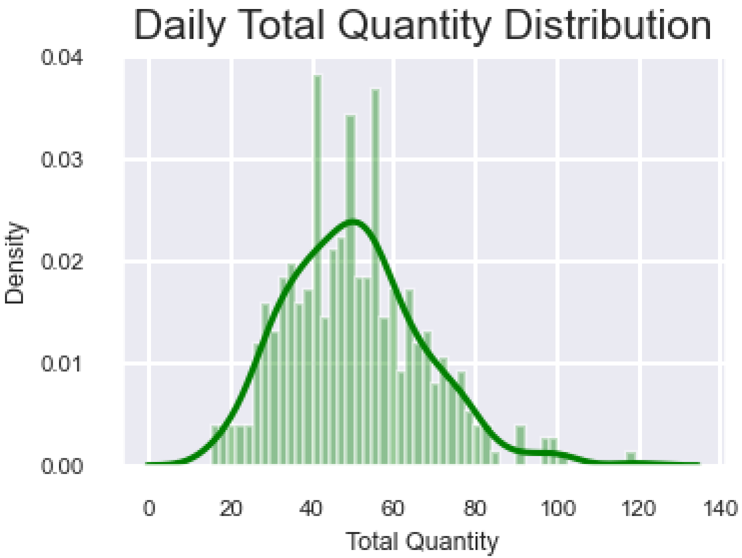
Monthly



EDA Inventory Data



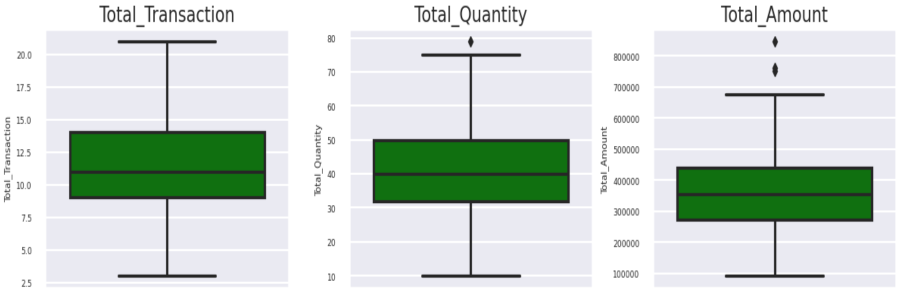
There are several outliers



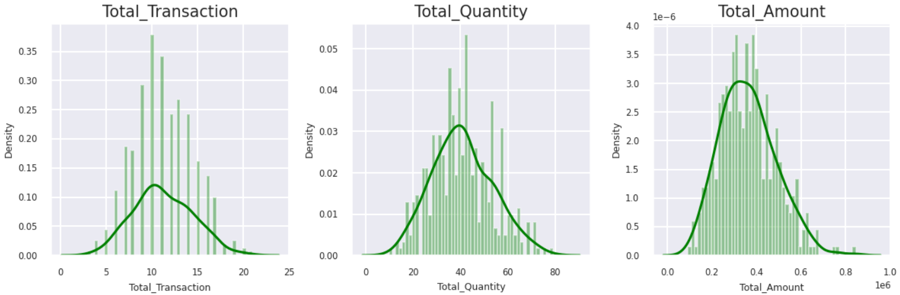
Approximately Symmetrical

EDA Marketing Data

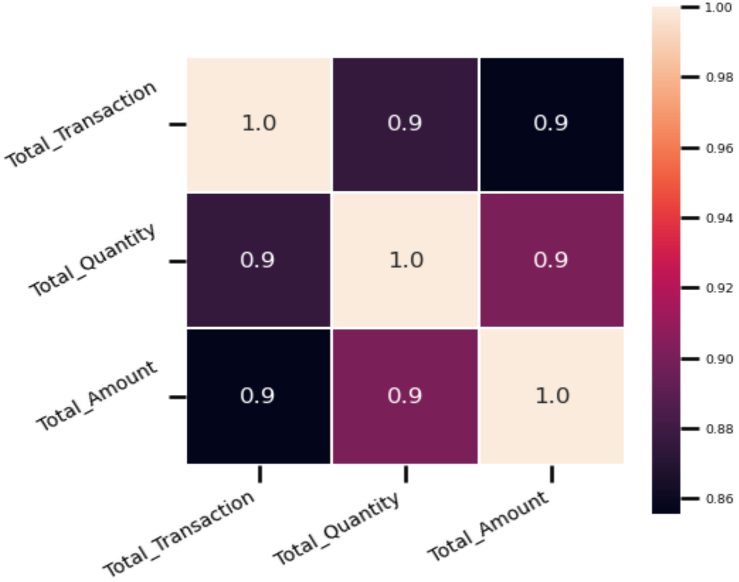
Boxplot



Distribution



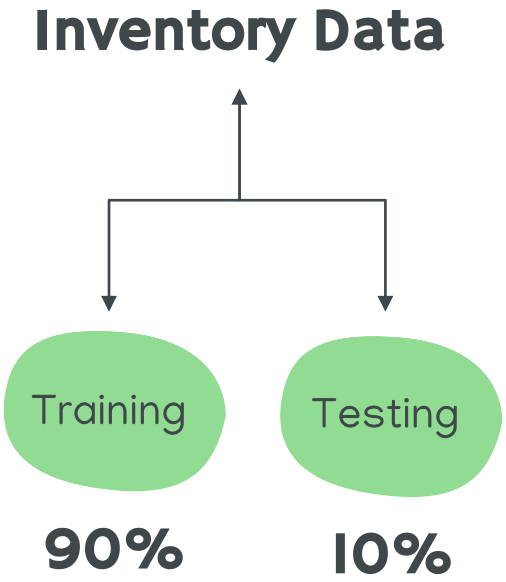
EDA Marketing Data



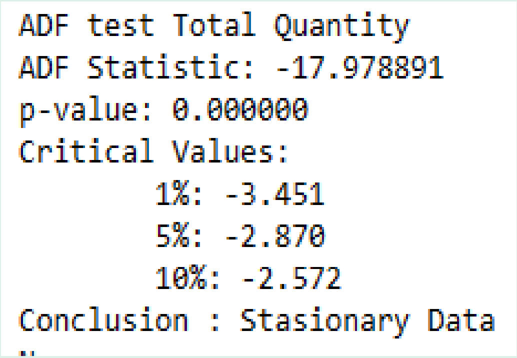
There is a very strong correlation between these three features, with a correlation of 0.9. It indicates that there is a close relationship between these variables.

05 Machine Learning

I. ARIMA Modeling



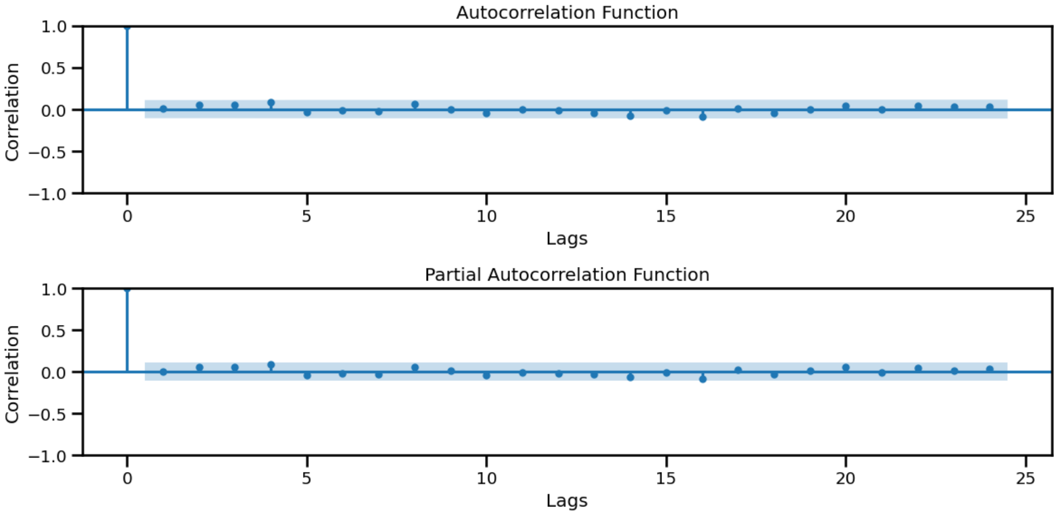
ADF Test



```
ADF test Total Quantity
ADF Statistic: -17.978891
p-value: 0.000000
Critical Values:
        1%: -3.451
        5%: -2.870
        10%: -2.572
Conclusion : Stasionary Data
```

I. ARIMA Modeling

ACF and PACF Inventory Data



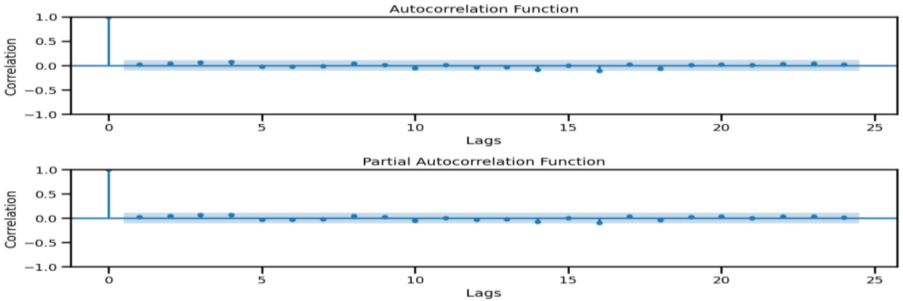
I. ARIMA Modeling

Model	Significance of Parameter
ARIMA (1,0,0)	There are parameters that are not significant
ARIMA (0,0,1)	There are parameters that are not significant
ARIMA (1,0,1)	All parameters are significant
ARIMA (2,0,2)	There are parameters that are not significant
ARIMA (3,0,3)	There are parameters that are not significant
ARIMA (4,0,4)	There are parameters that are not significant

I. ARIMA Modeling

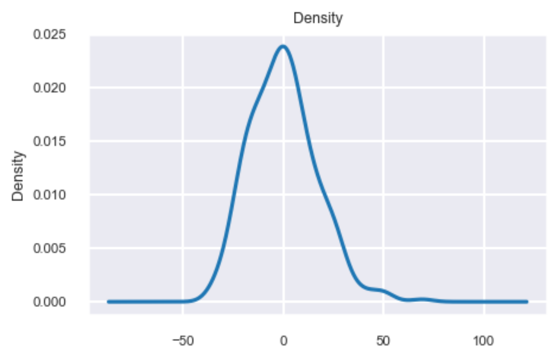
White Noise

ACF and PACF Residuals



Lower lags barely show any significant 'spikes'. This indicates that the residuals are close to white noise. It can be concluded that ARIMA (4,0,4) meets the white noise assumption.

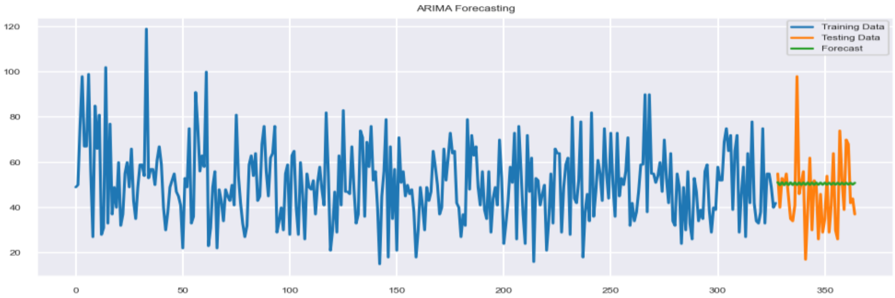
Normality



The normality test results show that the residuals are not normally distributed

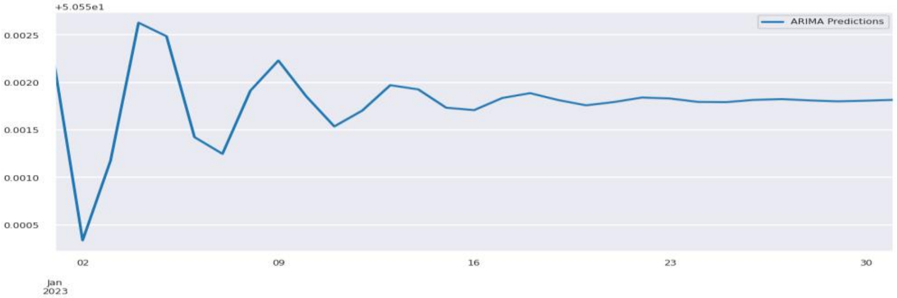
I. ARIMA Modeling

Forecasting Testing Data



```
MAE  : 12.675885484467079
MAPE : 0.3416714532952726
RMSE : 16.21730311234308
```

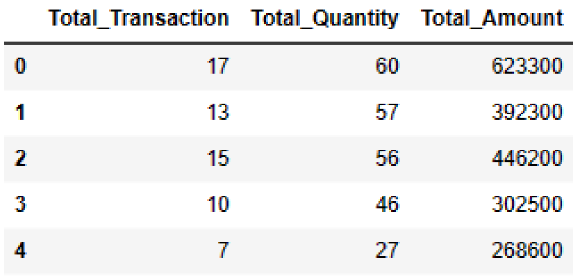
Forecasting for The next 31 Days



The forecasting results show that for the next period the total quantity per day will be around 50.

2. K-Means Modeling

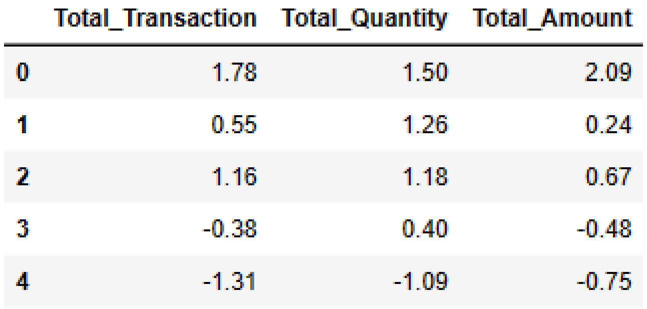
Marketing Data



	Total_Transaction	Total_Quantity	Total_Amount
0	17	60	623300
1	13	57	392300
2	15	56	446200
3	10	46	302500
4	7	27	268600

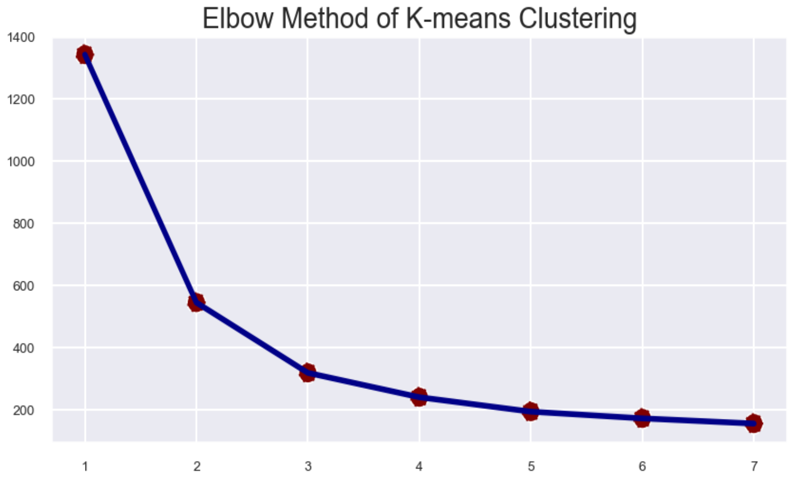
Data must be standardized because the data scale is different

Marketing data that has been standardized

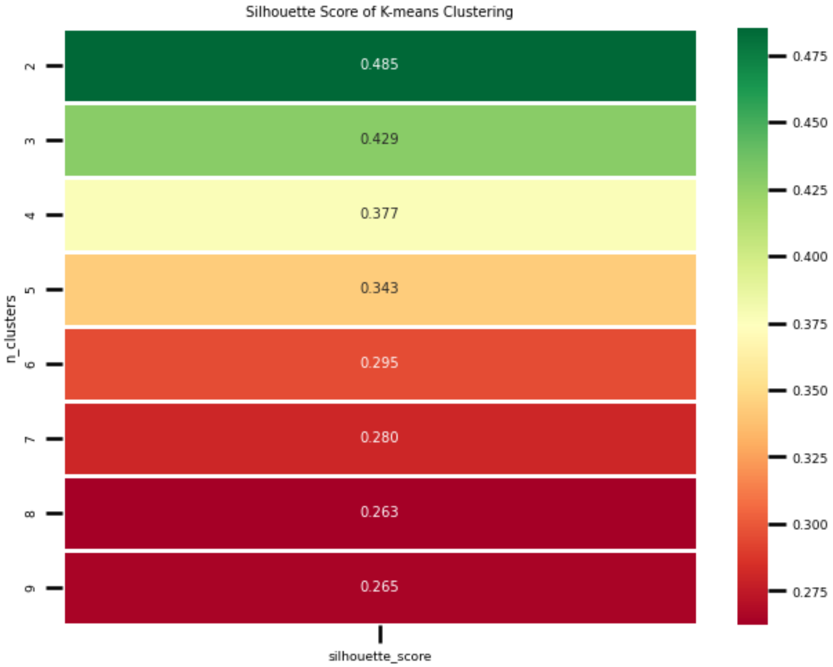


	Total_Transaction	Total_Quantity	Total_Amount
0	1.78	1.50	2.09
1	0.55	1.26	0.24
2	1.16	1.18	0.67
3	-0.38	0.40	-0.48
4	-1.31	-1.09	-0.75

2. K-Means Modeling

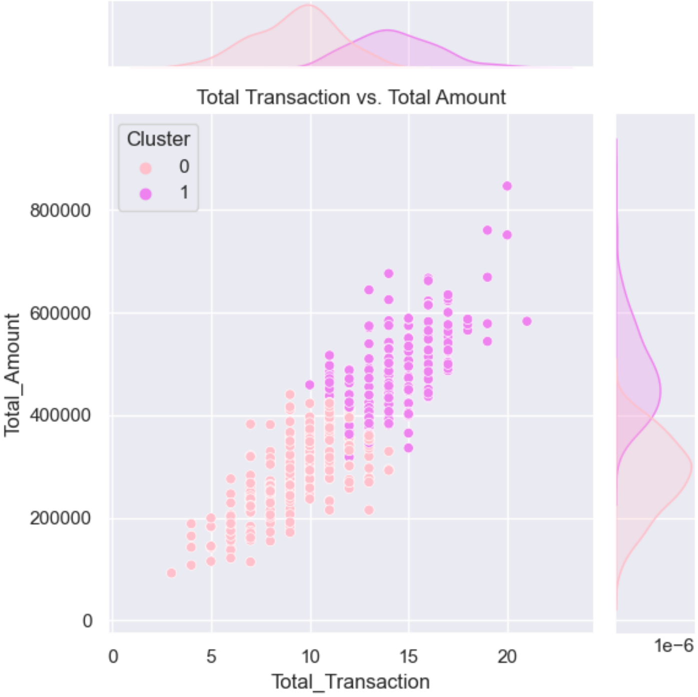


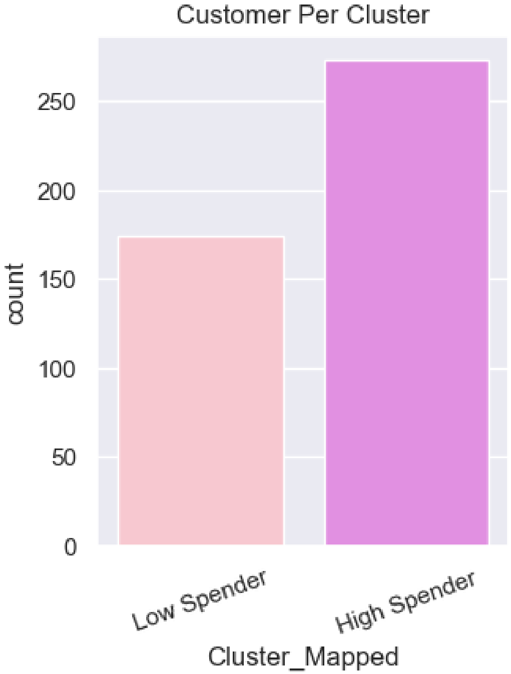
In the Elbow graph, it can be seen that the elbow fracture occurs when the clusters are equal to 2 and 3.

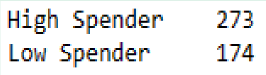


2. K-Means Modeling

K-Means 2 Clusters



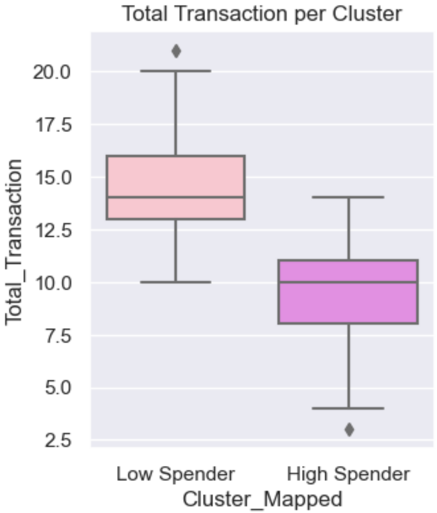


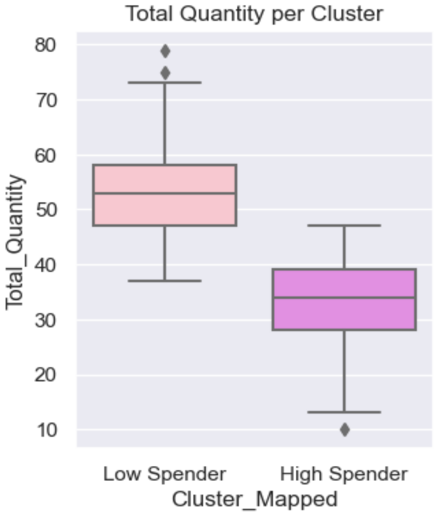


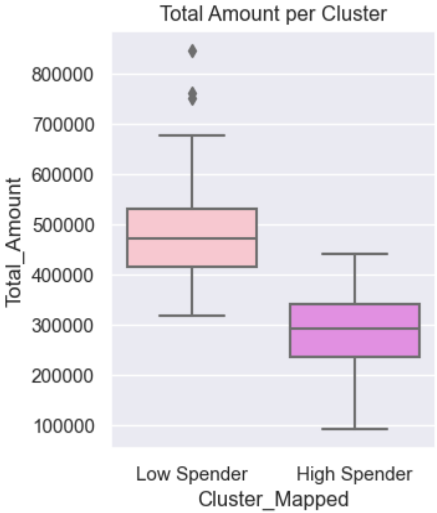
```
High Spender    273
Low Spender     174
```

2. K-Means Modeling

Boxplot of Each Features







06

Insight & Recommendations

Insight

Inventory Data

1. There is interesting pattern in monthly time series plot. This pattern may reflect a seasonal pattern or annual trend that deserves further scrutiny.
2. The best ARIMA modeling result for total quantity is ARIMA (1,0,1). The forecasting results show that for the next period the total quantity per day will be around 50.

Marketing Data

1. The results from the Heatmap show that there is a very strong correlation between these three features, with a correlation of 0.9. It indicates that there is a close relationship between these variables.
2. The cluster analysis results formed 2 clusters, namely low spenders and high spenders. High spenders have a greater number of customers who are classified as high spenders than the number of customers who are classified as low spenders.

Recommendations

Inventory Data

1. The inventory team can perform more in-depth analysis to identify the cause of the pattern.
2. The existence of outliers in the data is important to pay attention to. The inventory team needs to conduct further analysis of these outliers to understand why.
3. ARIMA (1,0,1) The model has residuals that meet the white noise assumption but do not meet the normality assumption

Marketing Data

1. Marketing team can design different marketing strategies for each group
1. Marketing team can develop more premium products or services for the high spender. More affordable products or bundle packages may tempt low spenders.
2. Marketing team can provide reward and loyalty programs to motivate low spender customers to increase their spending.

Thanks

linkedin.com/in/radishafannis

radishafs@gmail.com