Rakamin Academy X KALBE Nutritionals

Machine Learning Project

Kalbe Nutritionals Data Scientist Virtual Internship

Presented by
Ruth Yohanna Banjarnahor

About Rakamin Academy

Rakamin Academy is an end-to-end career development platform that offers a comprehensive range of services. These services include psychological assessments, technical assessments, intensive training, career guidance and consultation, portfolio enhancement programs, virtual internships, and job placements. Rakamin Academy aims to provide inclusive and impactful education access to the Indonesian community, assisting them in starting careers in the field of technology.

About Kalbe Nutritionals

Kalbe Nutritionals is a subsidiary of PT Kalbe Farma Tbk, a leading pharmaceutical company in Indonesia. It specializes in health and nutrition products, offering a wide range of food, supplements, and beverages that cater to nutritional needs. Kalbe Nutritionals is actively involved in research, development, and educational campaigns to promote a healthy lifestyle. They are a market leader in the Indonesian health and nutrition industry.

Ruth Yohanna

KALBE Nutritionals

Rakamin Academy

About You

I am a passionate graduate in Bioengineering from Bandung Institute of Technology. I have a strong inclination towards learning new things every day. Although data science is a recent interest, I am dedicated to understanding it. With a solid foundation in mathematics and analytical thinking from my engineering background, I am excited about the potential of data science to make a significant impact in various industries. My goal is to continuously grow and contribute to solving real-world challenges through data-driven approaches.

My Experience

Research and Development Intern

PT Djarum

Developed and executed a research project using IBM SPSS Statistics 20 to investigate the correlation between the weight of water hyacinth plants and their effectiveness in treating wastewater from Djarum Oasis Kretek Factory.

in Ruth Yohanna Banjarnahor

✉ ruthyohanna11banjarnahor@gmail.com

OUTLINE

- 01 Background story
- 02 Tools
- 03 Challenges
- 04 Business Recomendation
- 05 Repository

Background Story

I am a Data Scientist at Kalbe Nutritionals and have recently received a new project from the inventory and marketing teams.

From the inventory team, I have been tasked with assisting in predicting the sales quantity for all Kalbe products.

- ❑ The objective of this project is to determine the estimated quantity of products sold so that the inventory team can maintain sufficient daily stock levels.
- ❑ The predictions made should be on a daily basis.

From the marketing team, I have been requested to create customer clusters/segments based on several criteria.

- ❑ The goal of this project is to develop customer segments.
- ❑ These customer segments will be utilized by the marketing team to provide personalized promotions and sales treatments.

Tools

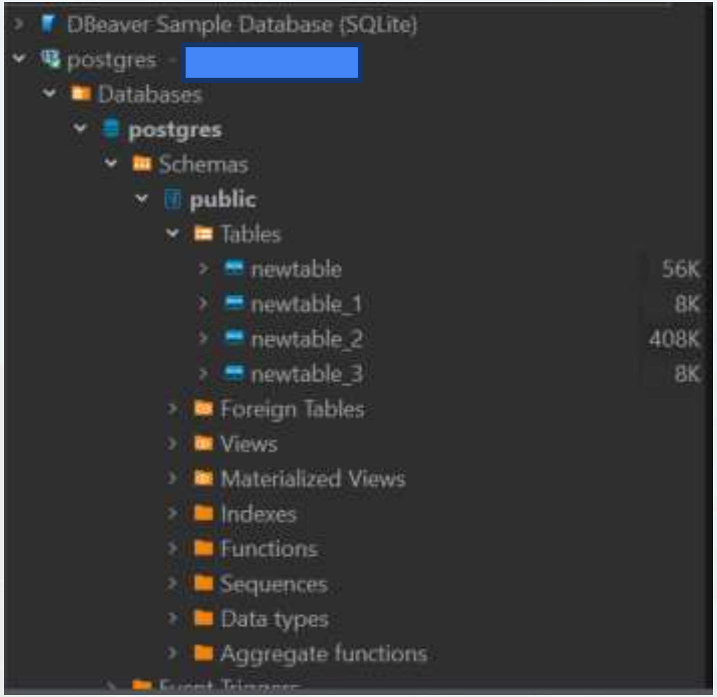
- □ Python
- □ Jupiter Notebook
- □ Tableau
- □ Dbeaver
- □ PostgreSQL

Rakamin Academy

Challenge 1

SQL, Dbeaver Connection with PostgreSQL

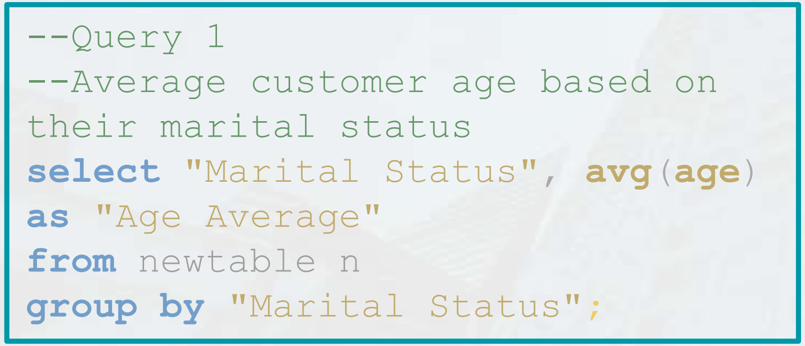
Data ingestion into DBeaver



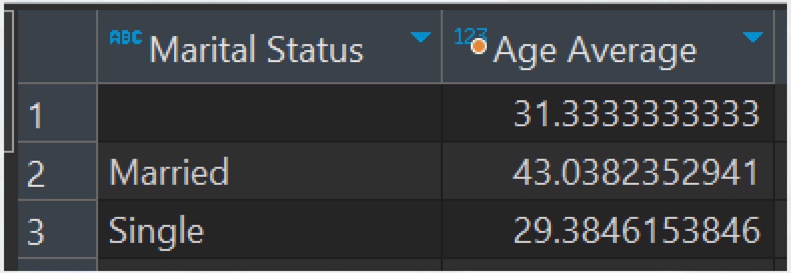
```
> DBeaver Sample Database (SQLite)
v postgres - [redacted]
  v Databases
    v postgres
      v Schemas
        v public
          v Tables
            > newtable                56K
            > newtable_1               8K
            > newtable_2             408K
            > newtable_3               8K
          > Foreign Tables
          > Views
          > Materialized Views
          > Indexes
          > Functions
          > Sequences
          > Data types
          > Aggregate functions
      > Event Triggers
```

Exploratory data analysis in DBeaver

Question 1: What is the average age of customers based on their marital status?



```sql
--Query 1
--Average customer age based on
their marital status
select "Marital Status", avg(age)
as "Age Average"
from newtable n
group by "Marital Status";
```



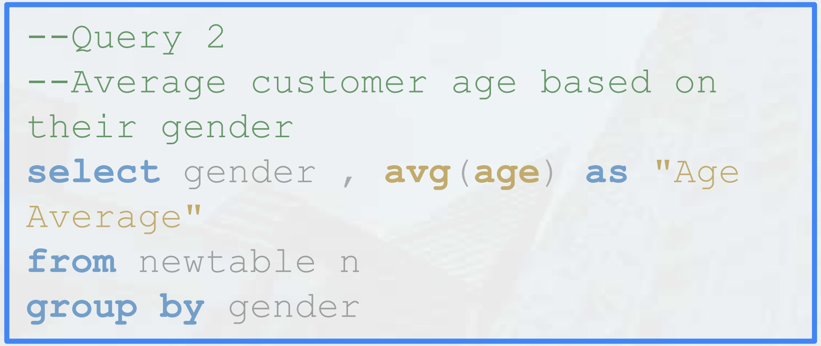
	Marital Status	Age Average
1		31.3333333333
2	Married	43.0382352941
3	Single	29.3846153846

The average age of customers based on their marital status = 31 years

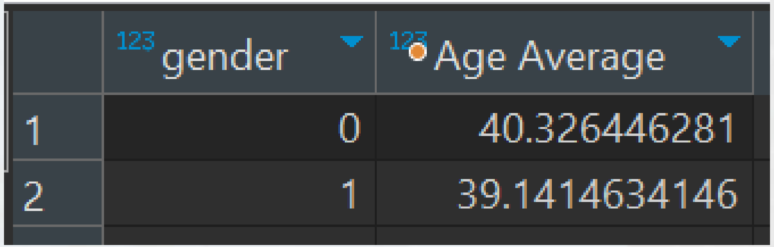
- The average age of married customers = 43 years
- The average age of single customers = 29 years

Exploratory data analysis in DBeaver

Question 2: What is the average age of customers based on their gender?



```
--Query 2
--Average customer age based on their gender
select gender , avg(age) as "Age Average"
from newtable n
group by gender
```



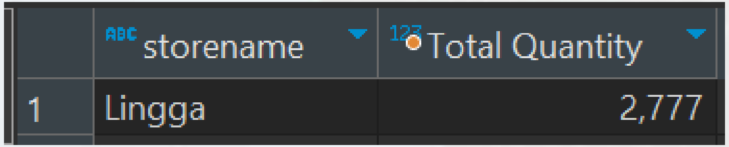
	gender	Age Average
1	0	40.326446281
2	1	39.1414634146

- The average age of female customers (0) = 40 years
- The average age of male customers (1) = 39 years

Exploratory data analysis in DBeaver

Question 3: Determine the name of the store with the highest total quantity!

```sql
--Query 3
--Store with the highest total quantity sold
select n.storename, sum(n2.qty) as "Total Quantity"
from newtable_1 n inner join newtable_2 n2
on n.storeid = n2.storeid
group by n.storename
order by "Total Quantity" desc
limit 1;
```



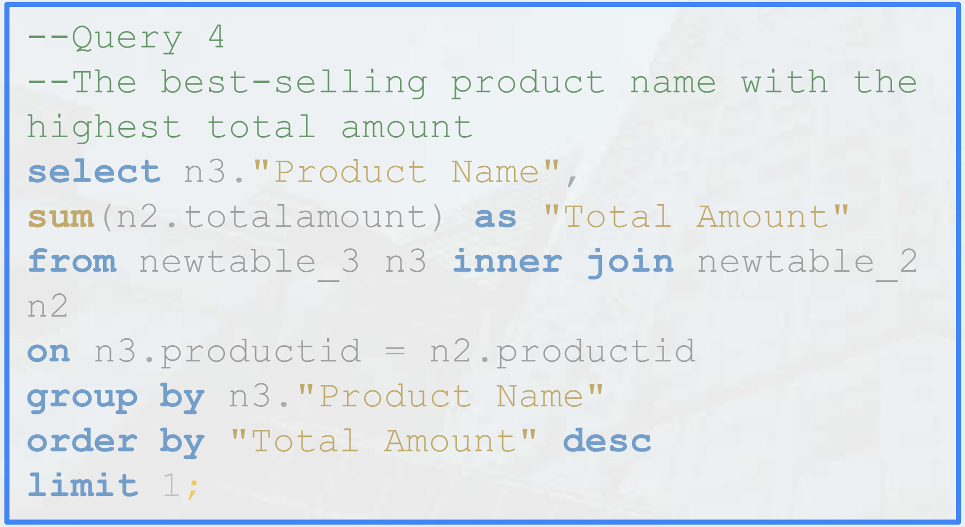
	storename	Total Quantity
1	Lingga	2,777

The name of the store with the highest total quantity is “Lingga”

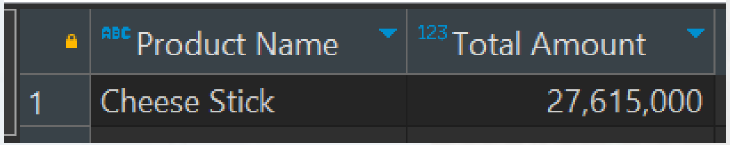
KALBE Nutritionals Rakamin Academy

Exploratory data analysis in DBeaver

Question 3: Determine the name of the best-selling product with the highest total amount!



```sql
--Query 4
--The best-selling product name with the
highest total amount
select n3."Product Name",
sum(n2.totalamount) as "Total Amount"
from newtable_3 n3 inner join newtable_2
n2
on n3.productid = n2.productid
group by n3."Product Name"
order by "Total Amount" desc
limit 1;
```



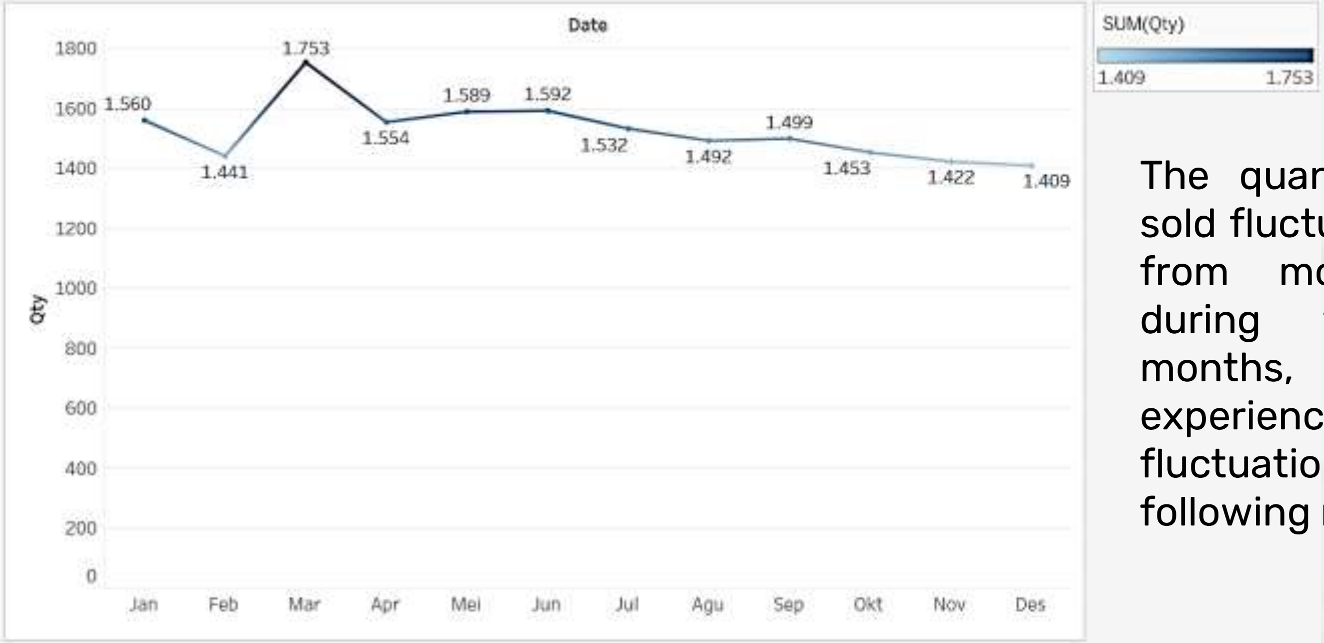
	Product Name	Total Amount
1	Cheese Stick	27,615,000

The name of the product with the highest total amount is "Cheese Stick"

Challenge 2

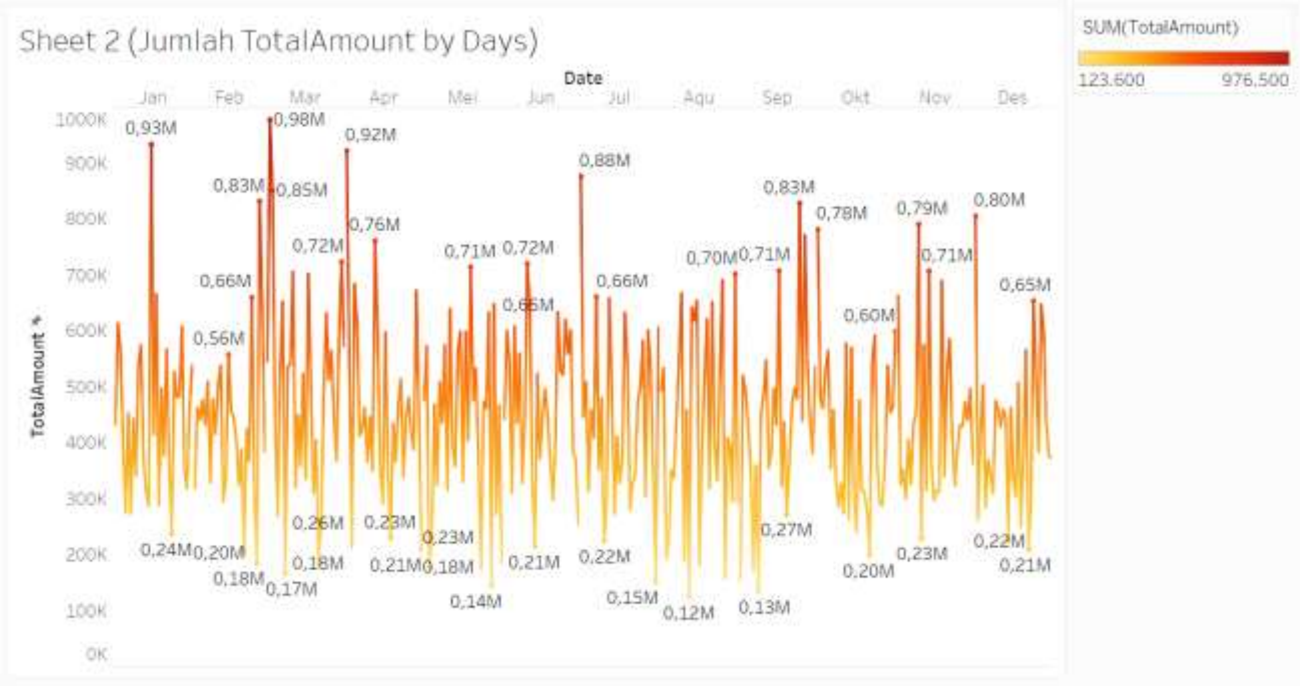
Data Visualization & Dashboard using Tableau

The Quantity (Qty) from Month to Month



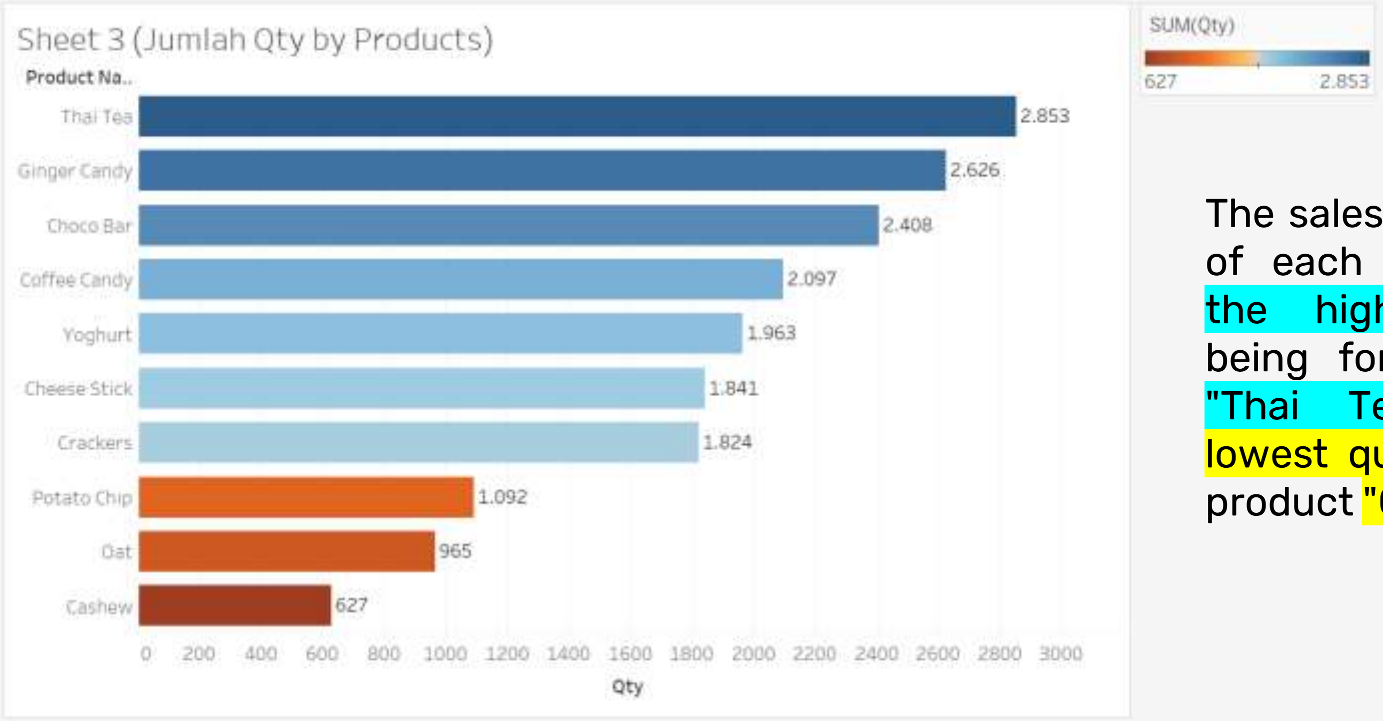
The quantity of products sold fluctuates significantly from month to month during the first four months, and then experiences moderate fluctuations in the following months.

The Total Amount from Day to Day



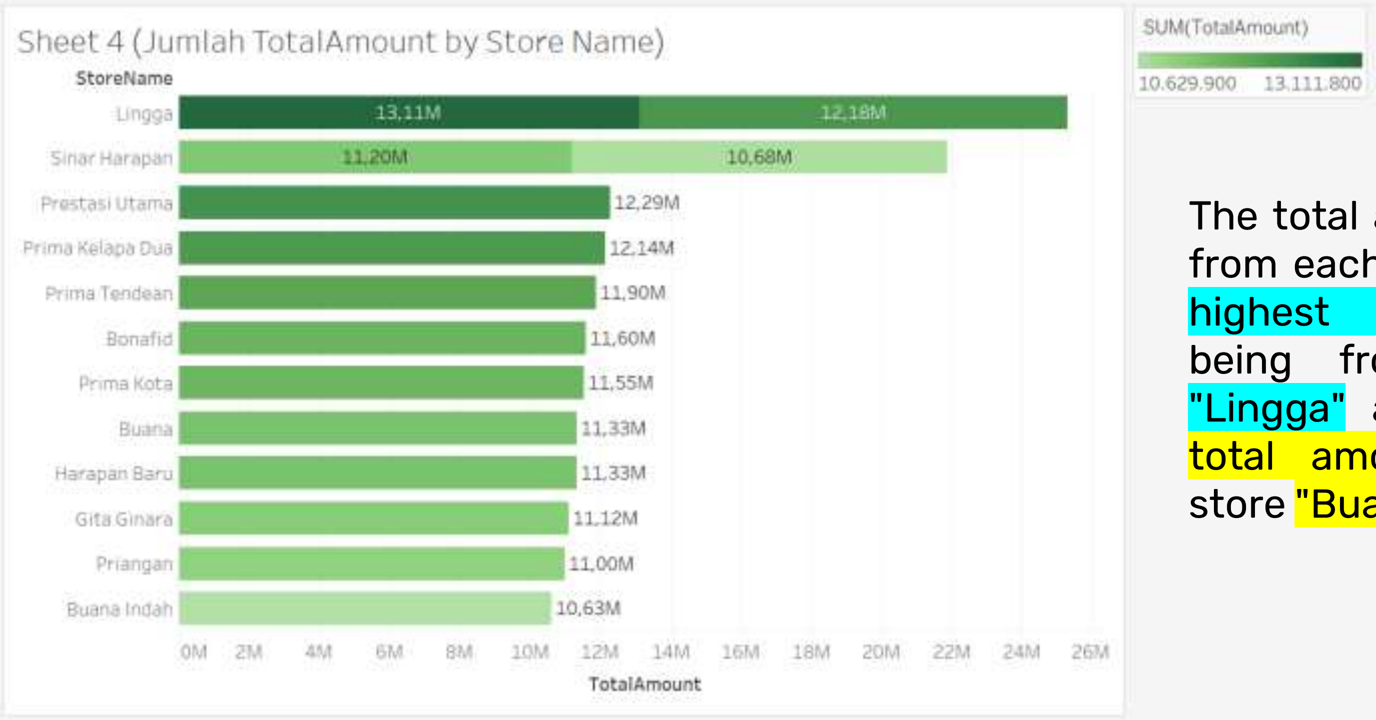
The total amount fluctuates significantly from day to day, with significant variations observed on a daily basis

The Sales Quantity (Qty) by Product

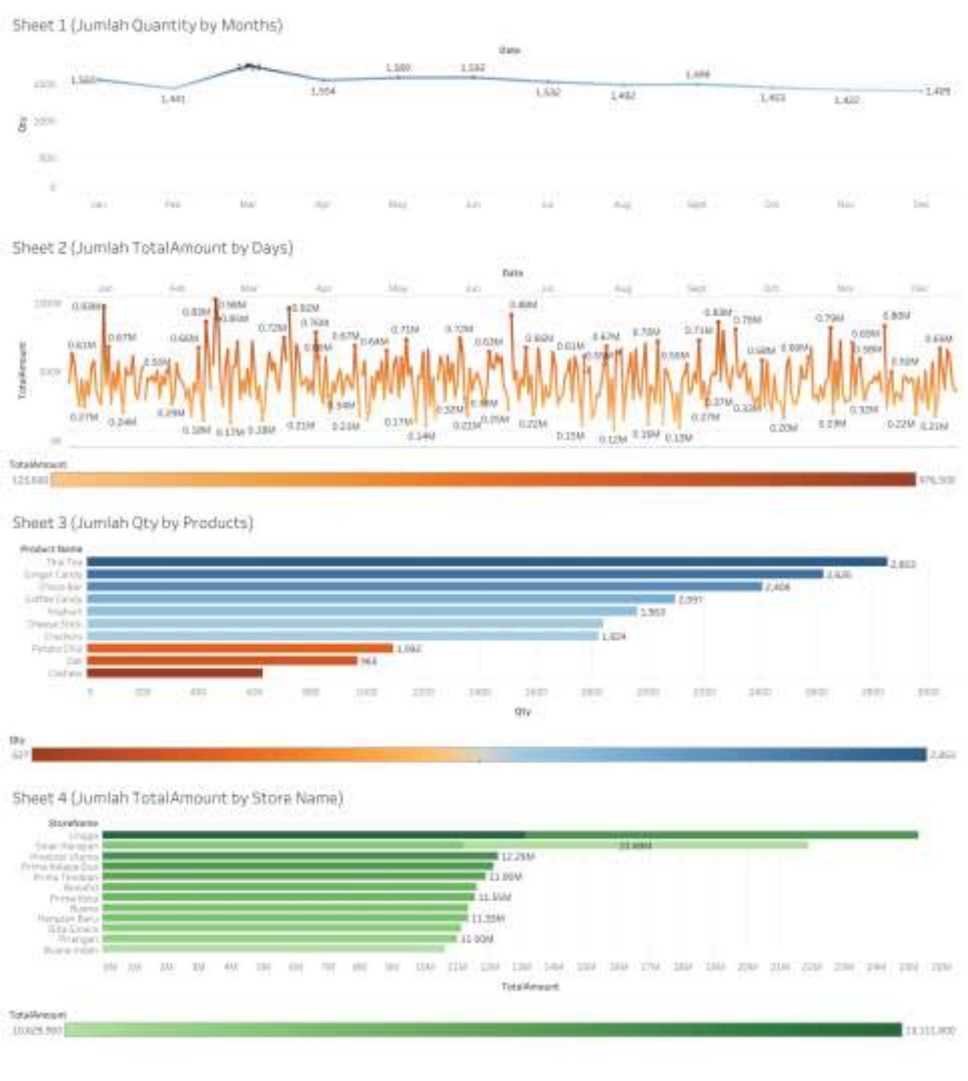


The sales quantity (Qty) of each product, with the highest quantity being for the product "Thai Tea" and the lowest quantity for the product "Cashew".

The Total Amount by Store Name



The total amount of sales from each store, with the highest total amount being from the store "Lingga" and the lowest total amount from the store "Buana Indah"



Dashboard

[Kalbe Nutritionals Sales Dashboard]

Rakamin Academy

Challenge 3

Machine Learning Regression (Time Series)

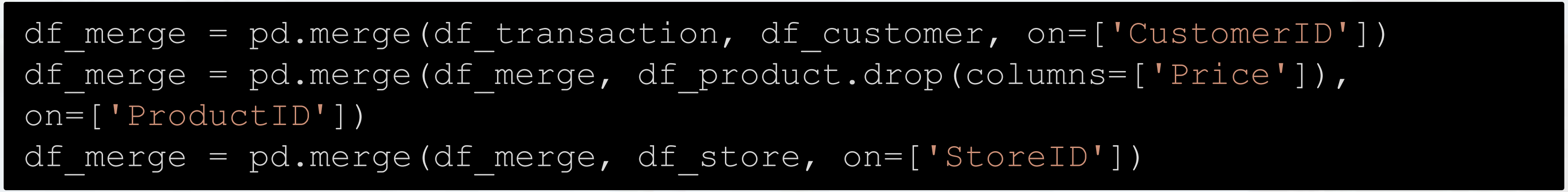
Data Cleansing

```
#Data cleansing df customer
df_customer['Income'] = df_customer['Income'].replace('[,]','.',
regex=True).astype('float')
```

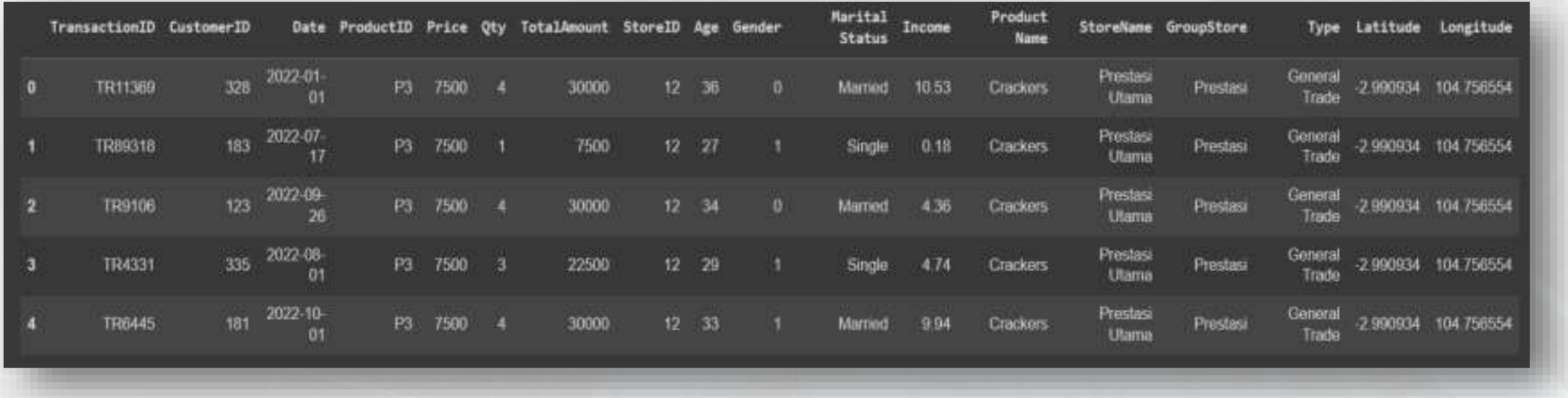
```
#Data cleansing df store
df_store['Latitude'] = df_store['Latitude'].replace('[,]','.',
regex=True).astype('float')
df_store['Longitude'] = df_store['Longitude'].replace('[,]','.',
regex=True).astype('float')
```

```
#Data cleansing df transaction
df_transaction['Date'] = pd.to_datetime(df_transaction['Date'])
```

Data Merging



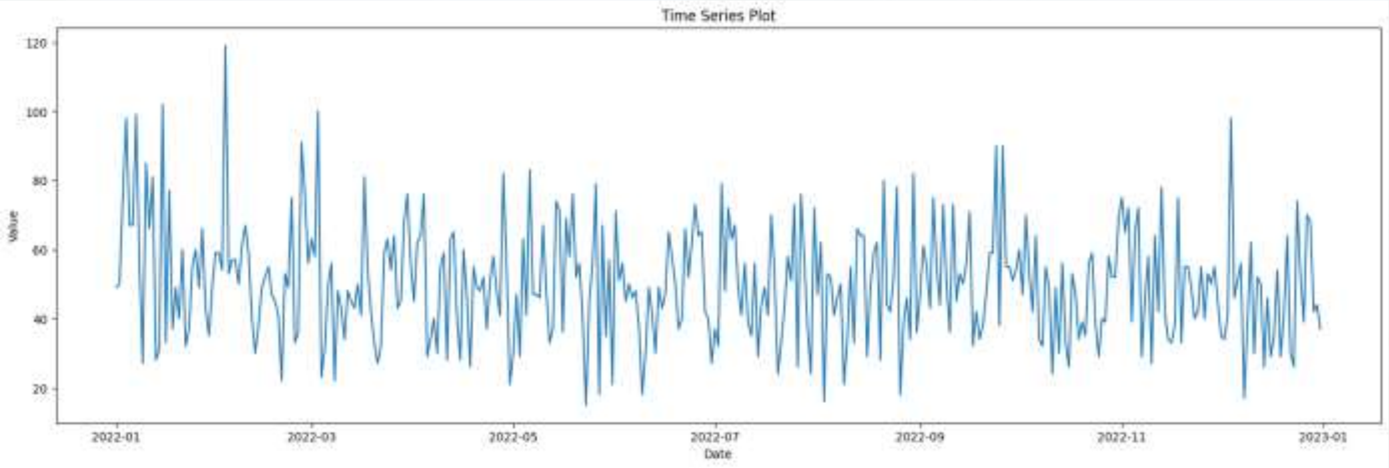
```
df_merge = pd.merge(df_transaction, df_customer, on=['CustomerID'])
df_merge = pd.merge(df_merge, df_product.drop(columns=['Price']),
on=['ProductID'])
df_merge = pd.merge(df_merge, df_store, on=['StoreID'])
```

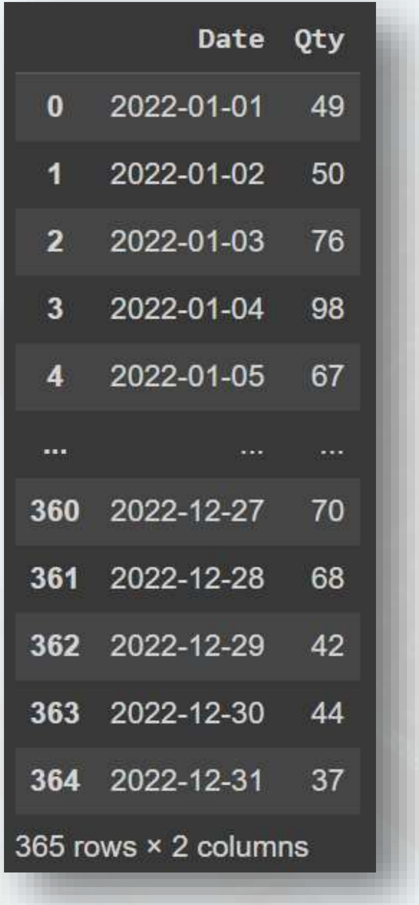


	TransactionID	CustomerID	Date	ProductID	Price	Qty	TotalAmount	StoreID	Age	Gender	Marital Status	Income	Product Name	StoreName	GroupStore	Type	Latitude	Longitude
0	TR11369	328	2022-01-01	P3	7500	4	30000	12	36	0	Married	10.53	Crackers	Prestasi Utama	Prestasi	General Trade	-2.990934	104.756554
1	TR89318	183	2022-07-17	P3	7500	1	7500	12	27	1	Single	0.18	Crackers	Prestasi Utama	Prestasi	General Trade	-2.990934	104.756554
2	TR9106	123	2022-09-26	P3	7500	4	30000	12	34	0	Married	4.36	Crackers	Prestasi Utama	Prestasi	General Trade	-2.990934	104.756554
3	TR4331	335	2022-08-01	P3	7500	3	22500	12	29	1	Single	4.74	Crackers	Prestasi Utama	Prestasi	General Trade	-2.990934	104.756554
4	TR6445	181	2022-10-01	P3	7500	4	30000	12	33	1	Married	9.94	Crackers	Prestasi Utama	Prestasi	General Trade	-2.990934	104.756554

Data Preprocessing

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

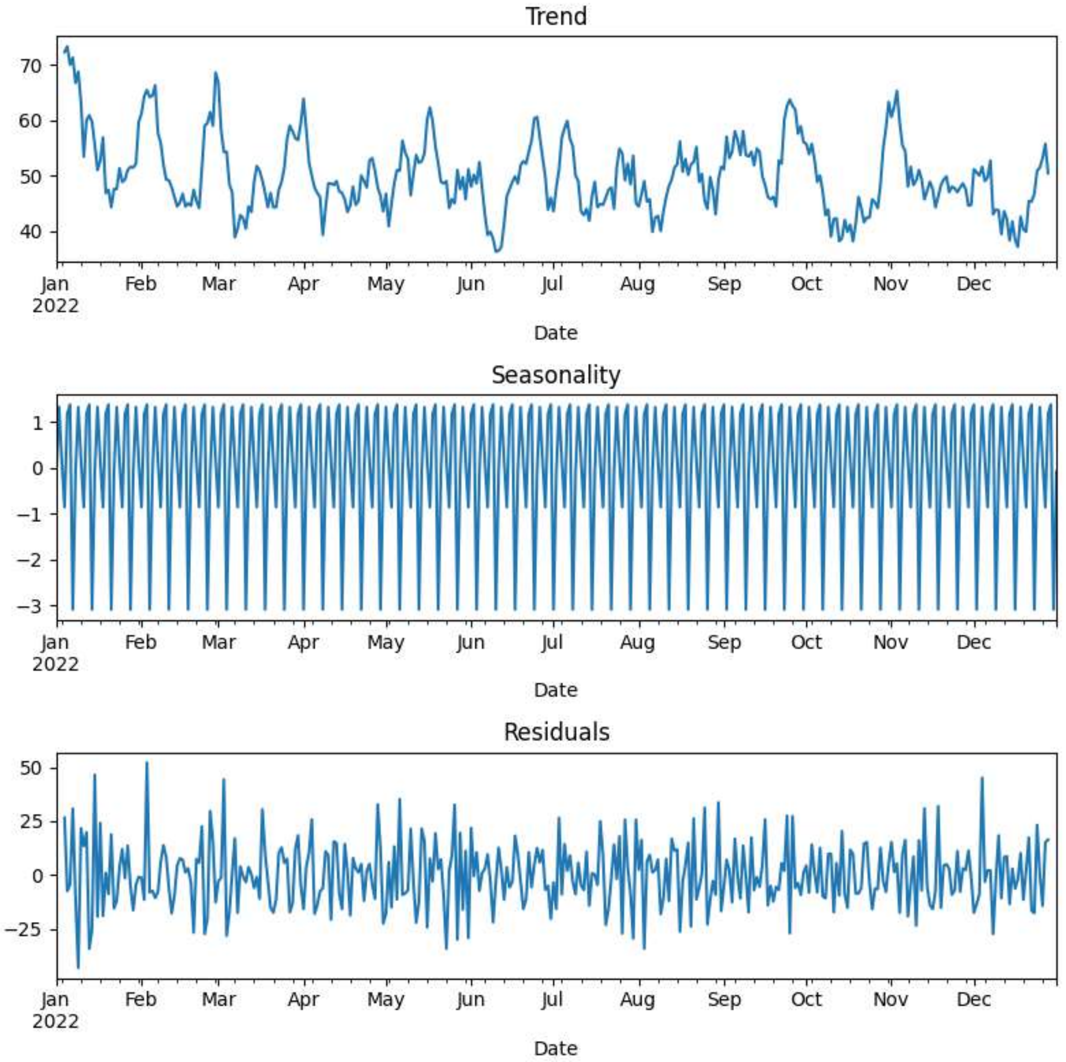


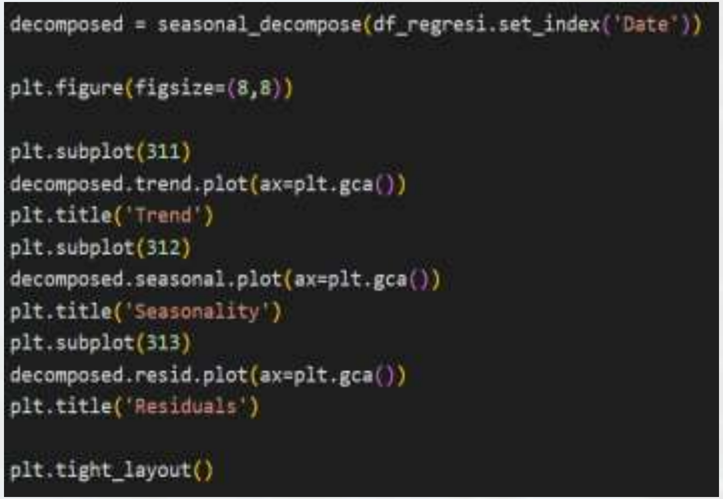


	Date	Qty
0	2022-01-01	49
1	2022-01-02	50
2	2022-01-03	76
3	2022-01-04	98
4	2022-01-05	67
...	...	...
360	2022-12-27	70
361	2022-12-28	68
362	2022-12-29	42
363	2022-12-30	44
364	2022-12-31	37

365 rows × 2 columns

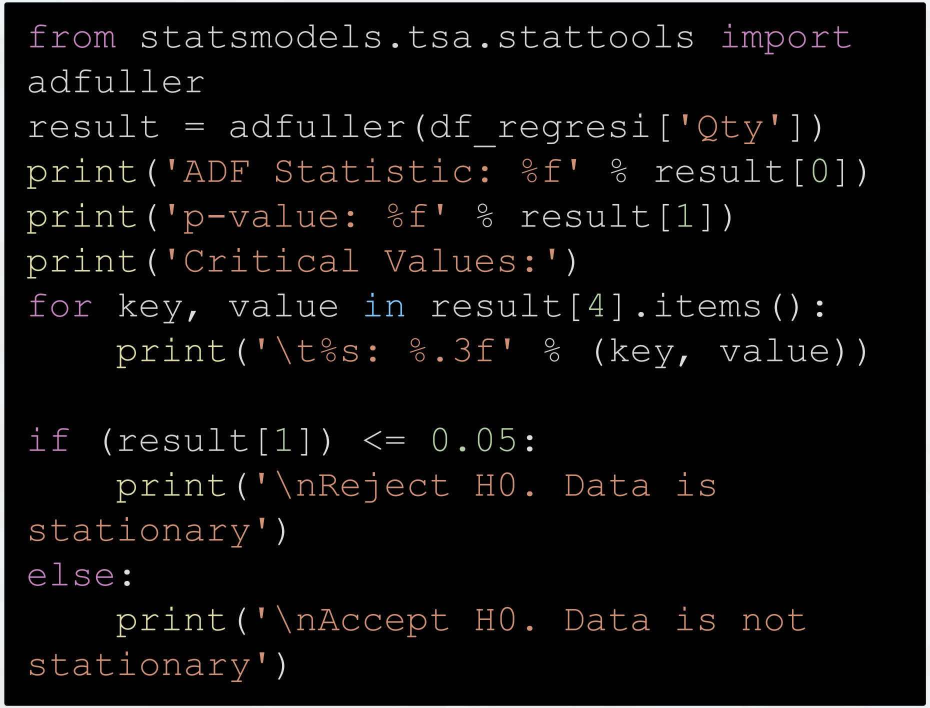
Data Trend & Seasonality



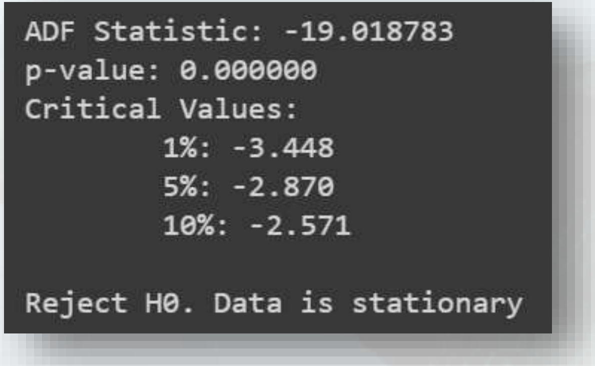


```
decomposed = seasonal_decompose(df_regresi.set_index('Date'))

plt.figure(figsize=(8,8))

plt.subplot(311)
decomposed.trend.plot(ax=plt.gca())
plt.title('Trend')
plt.subplot(312)
decomposed.seasonal.plot(ax=plt.gca())
plt.title('Seasonality')
plt.subplot(313)
decomposed.resid.plot(ax=plt.gca())
plt.title('Residuals')

plt.tight_layout()
```

Stationary Data Check



```
from statsmodels.tsa.stattools import
adfuller
result = adfuller(df_regresi['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))

if (result[1]) <= 0.05:
    print('\nReject H0. Data is
stationary')
else:
    print('\nAccept H0. Data is not
stationary')
```

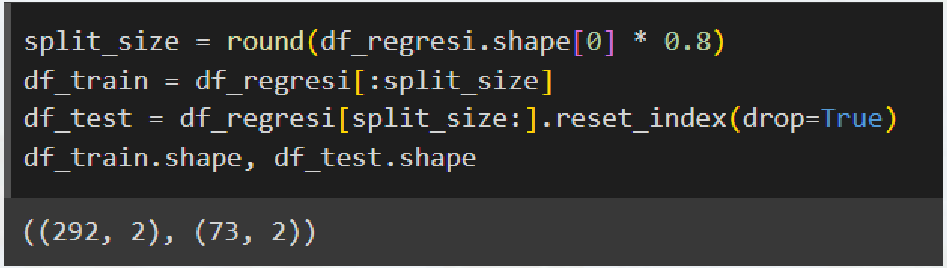


```
ADF Statistic: -19.018783
p-value: 0.000000
Critical Values:
        1%: -3.448
        5%: -2.870
        10%: -2.571

Reject H0. Data is stationary
```

H0 = Data is not stationary
H1 = Data is stationary
if p-value < 0.05, we will reject the H0 hypothesis

Based on Augmented Dicky-Fuller test, the p-value is 0.00, which is lower than 0.05. Therefore, we will reject H0 and accept H1. So, data is stationary (d=0)

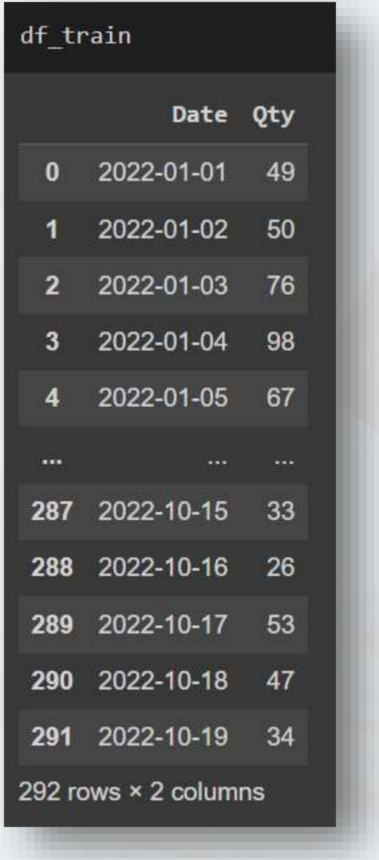
Data Splitting and Training



```
split_size = round(df_regresi.shape[0] * 0.8)
df_train = df_regresi[:split_size]
df_test = df_regresi[split_size:].reset_index(drop=True)
df_train.shape, df_test.shape
```

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

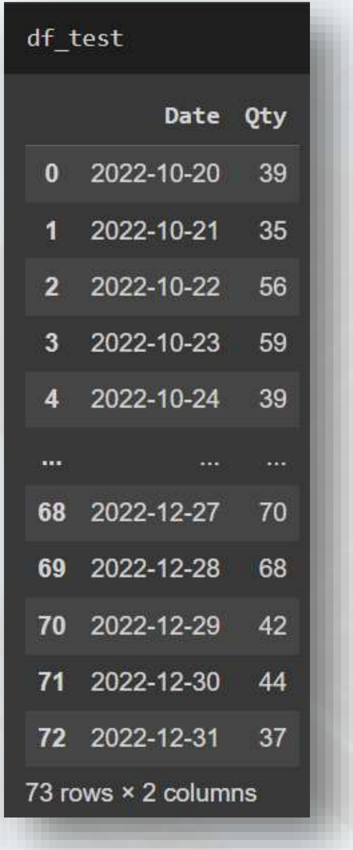
80% for training and 20% for testing



df_train

	Date	Qty
0	2022-01-01	49
1	2022-01-02	50
2	2022-01-03	76
3	2022-01-04	98
4	2022-01-05	67
...	...	...
287	2022-10-15	33
288	2022-10-16	26
289	2022-10-17	53
290	2022-10-18	47
291	2022-10-19	34

292 rows × 2 columns

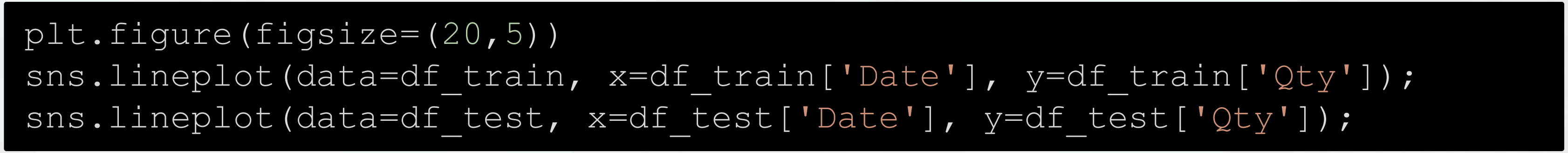


df_test

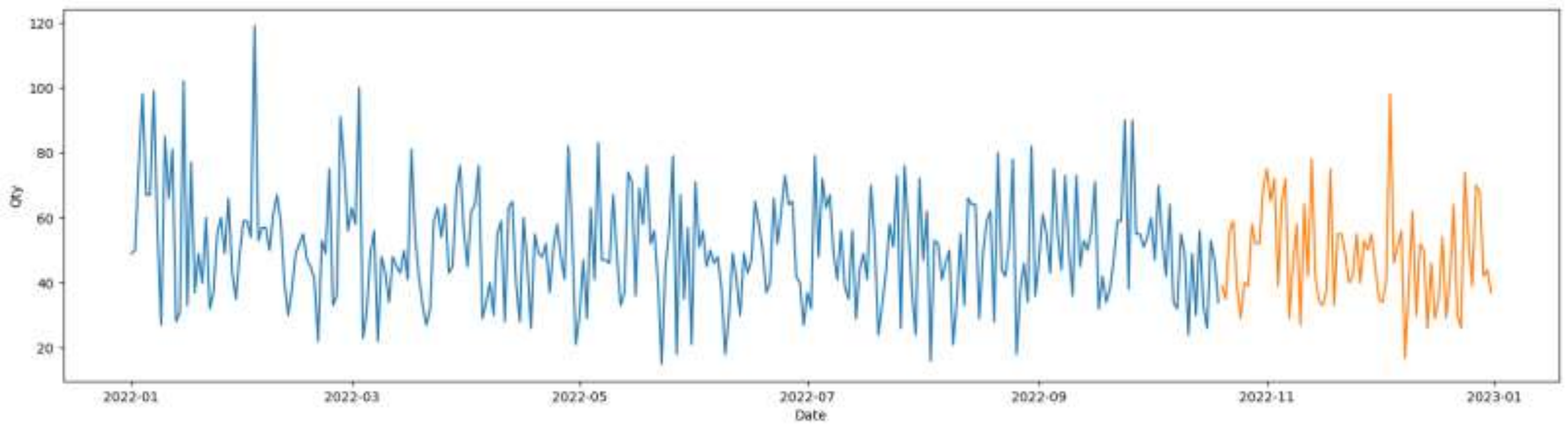
	Date	Qty
0	2022-10-20	39
1	2022-10-21	35
2	2022-10-22	56
3	2022-10-23	59
4	2022-10-24	39
...	...	...
68	2022-12-27	70
69	2022-12-28	68
70	2022-12-29	42
71	2022-12-30	44
72	2022-12-31	37

73 rows × 2 columns

Data Splitting and Training



```
plt.figure(figsize=(20,5))
sns.lineplot(data=df_train, x=df_train['Date'], y=df_train['Qty']);
sns.lineplot(data=df_test, x=df_test['Date'], y=df_test['Qty']);
```



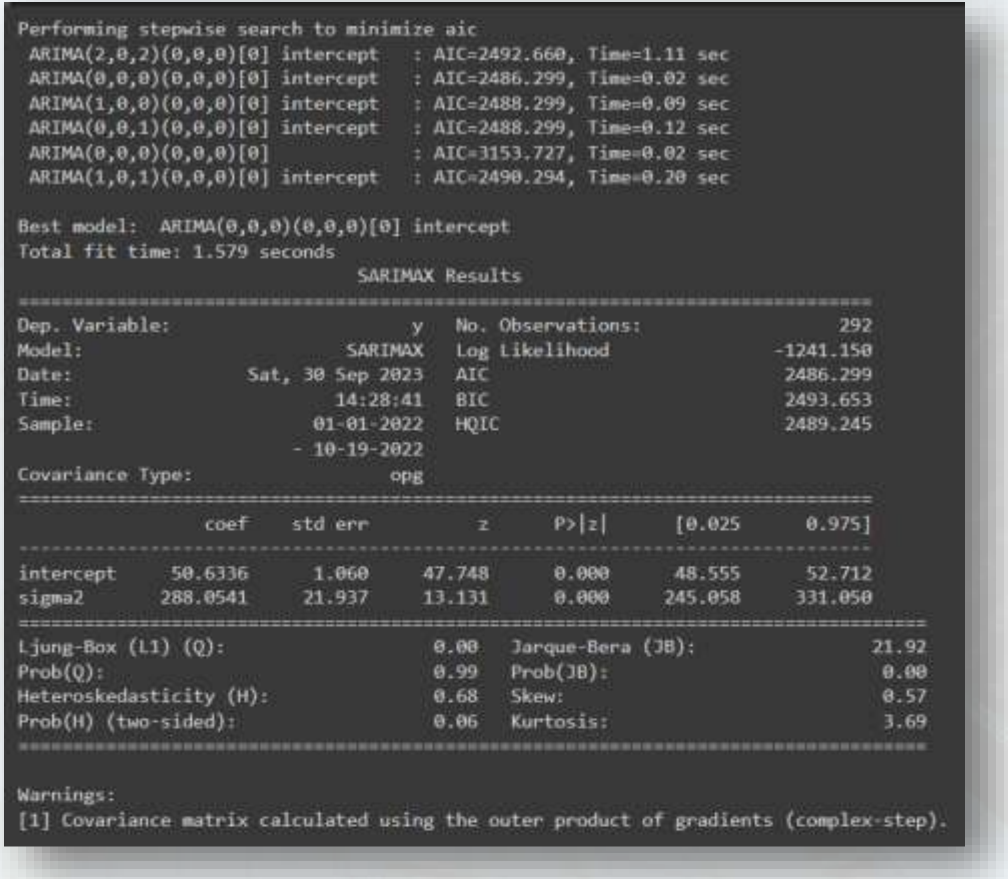
Finding Optimum Parameter (p, d, q)

Model 1: Auto Arima

p, d, q = (0, 0, 0)

AIC = 2486.299

When using the Auto Arima function to find p, d, q values, if it returns 0 for each parameter, it means that the automatic selection process could not find an appropriate ARIMA model that fits the data well within the explored search space. This could be due to two reasons: either the data is already stationary, indicating d = 0 and no differentiation is required, or the data does not require any AR or MA components, leading to p = q = 0.

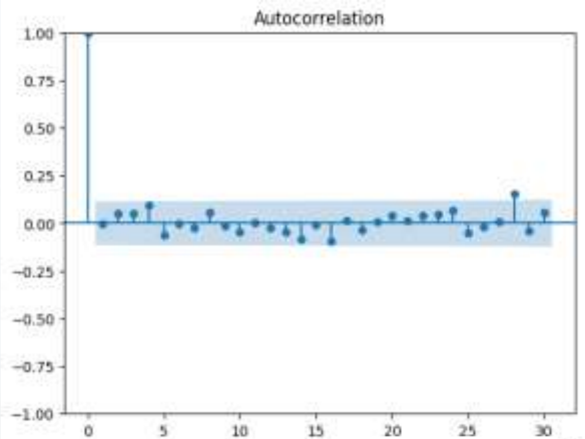


```
Performing stepwise search to minimize aic
 ARIMA(2,0,2)(0,0,0)[0] intercept   : AIC=2492.660, Time=1.11 sec
 ARIMA(0,0,0)(0,0,0)[0] intercept   : AIC=2486.299, Time=0.02 sec
 ARIMA(1,0,0)(0,0,0)[0] intercept   : AIC=2488.299, Time=0.09 sec
 ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=2488.299, Time=0.12 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=3153.727, Time=0.02 sec
 ARIMA(1,0,1)(0,0,0)[0] intercept   : AIC=2490.294, Time=0.20 sec

Best model:  ARIMA(0,0,0)(0,0,0)[0] intercept
Total fit time: 1.579 seconds
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  292
Model:                        SARIMAX   Log Likelihood               -1241.150
Date:                Sat, 30 Sep 2023   AIC                           2486.299
Time:                        14:28:41   BIC                           2493.653
Sample:                    01-01-2022   HQIC                          2489.245
                         - 10-19-2022
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     50.6336      1.060     47.748      0.000      48.555      52.712
sigma2       288.0541     21.937     13.131      0.000     245.058     331.050
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                21.92
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):               0.68   Skew:                             0.57
Prob(H) (two-sided):                  0.06   Kurtosis:                         3.69
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Finding Optimum Parameter (p, d, q)

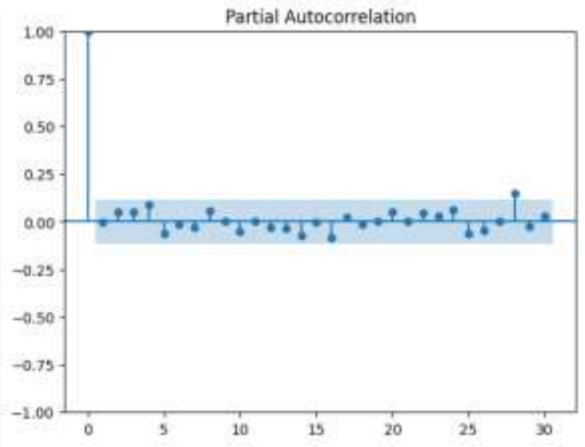
Model 2: ACF & PCF Plot

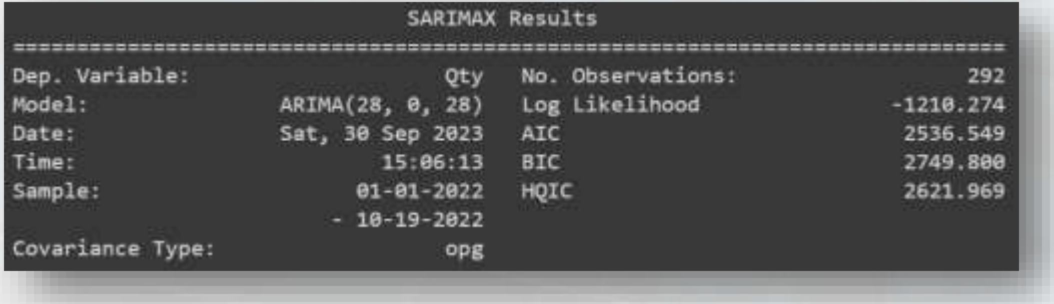
By analyzing the ACF and PACF plots, we can observe that only the 28th lag exceeds the significance limit in both plots. Hence, for the second model, we will select a value of 28 for both p and q to capture the significant correlation at that lag.



p, d, q = (28, 0, 28)

AIC = 2536.549





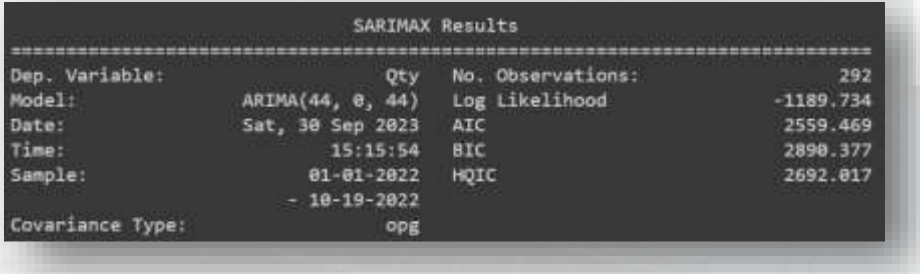
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                    Qty   No. Observations:                  292
Model:                ARIMA(28, 0, 28)   Log Likelihood               -1210.274
Date:                Sat, 30 Sep 2023   AIC                           2536.549
Time:                        15:06:13   BIC                           2749.800
Sample:                    01-01-2022   HQIC                          2621.969
                         - 10-19-2022
Covariance Type:                  opg
```

Finding Optimum Parameter (p, d, q)

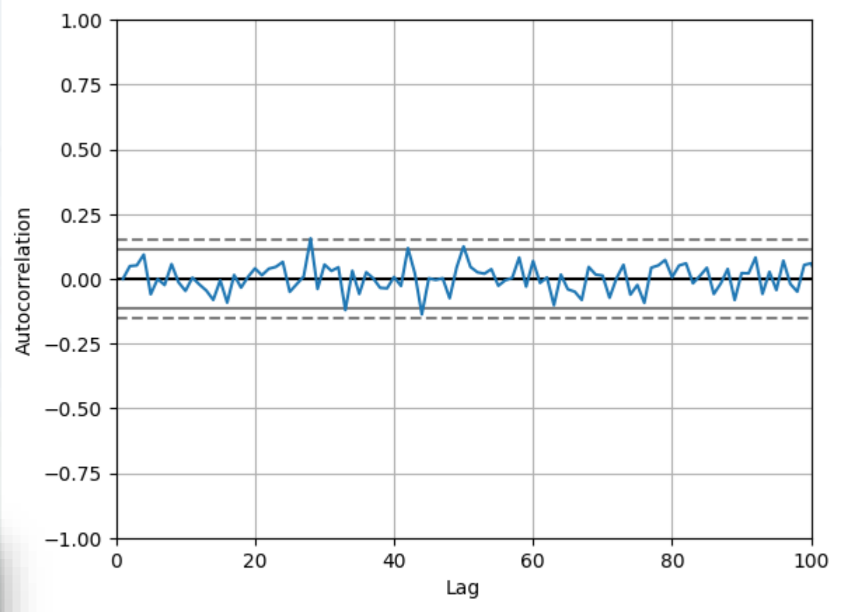
Model 3: Pandas Autocorrelation Plot

p, d, q = (44, 0, 44)

AIC = 2486.299

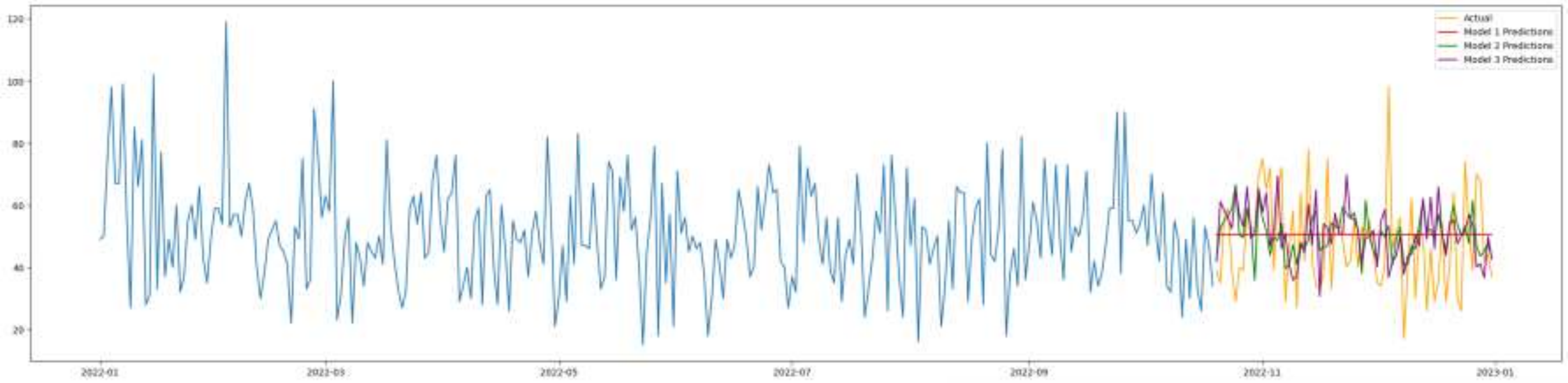


```
                                      SARIMAX Results
==============================================================================
Dep. Variable:                    Qty   No. Observations:                  292
Model:               ARIMA(44, 0, 44)   Log Likelihood               -1189.734
Date:                Sat, 30 Sep 2023   AIC                           2559.469
Time:                        15:15:54   BIC                           2890.377
Sample:                    01-01-2022   HQIC                          2692.017
                         - 10-19-2022
Covariance Type:                  opg
```

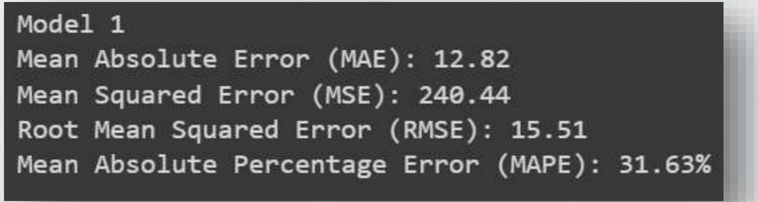


When analyzing the autocorrelation plot generated by pandas to determine the parameters, it becomes apparent that one of the values exceeds the 95% confidence interval, specifically at lag 44.

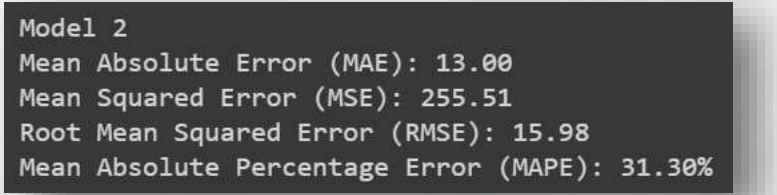
ARIMA Modeling



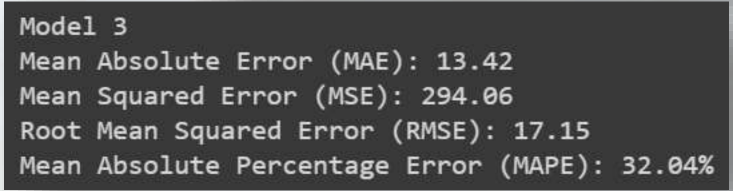
Model 1 : Auto Arima
Model 2 : ACF & PCF Plot
Model 3 : Pandas Autocorrelation Plot



```
Model 1
Mean Absolute Error (MAE): 12.82
Mean Squared Error (MSE): 240.44
Root Mean Squared Error (RMSE): 15.51
Mean Absolute Percentage Error (MAPE): 31.63%
```

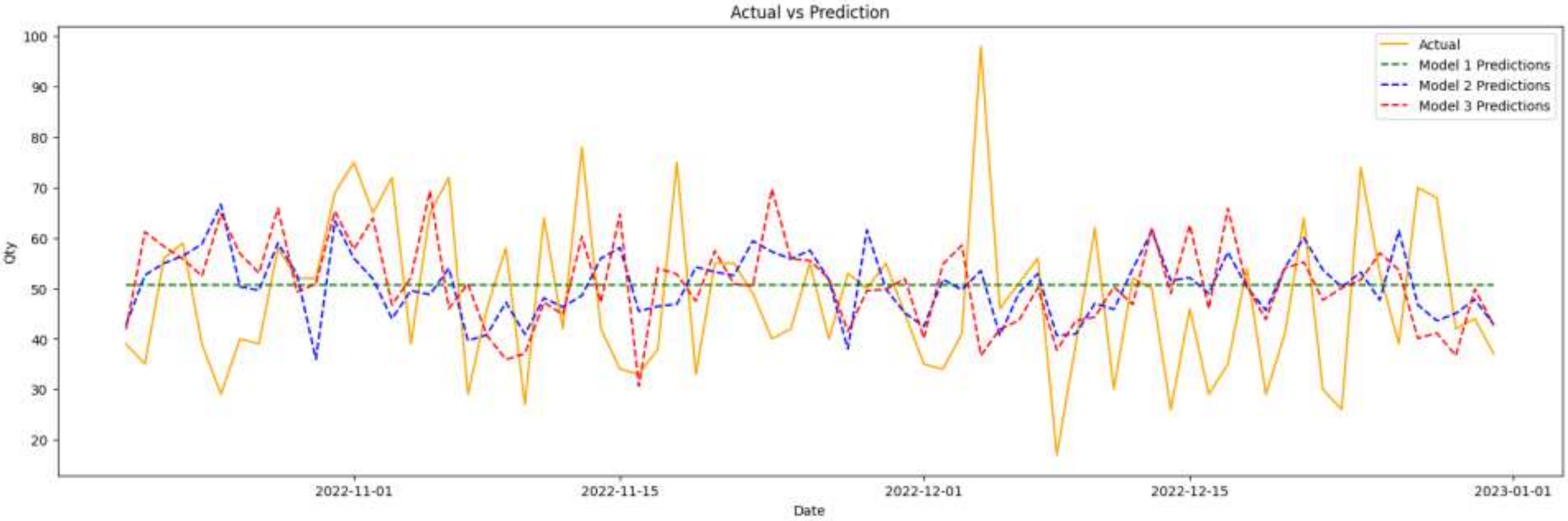


```
Model 2
Mean Absolute Error (MAE): 13.00
Mean Squared Error (MSE): 255.51
Root Mean Squared Error (RMSE): 15.98
Mean Absolute Percentage Error (MAPE): 31.30%
```



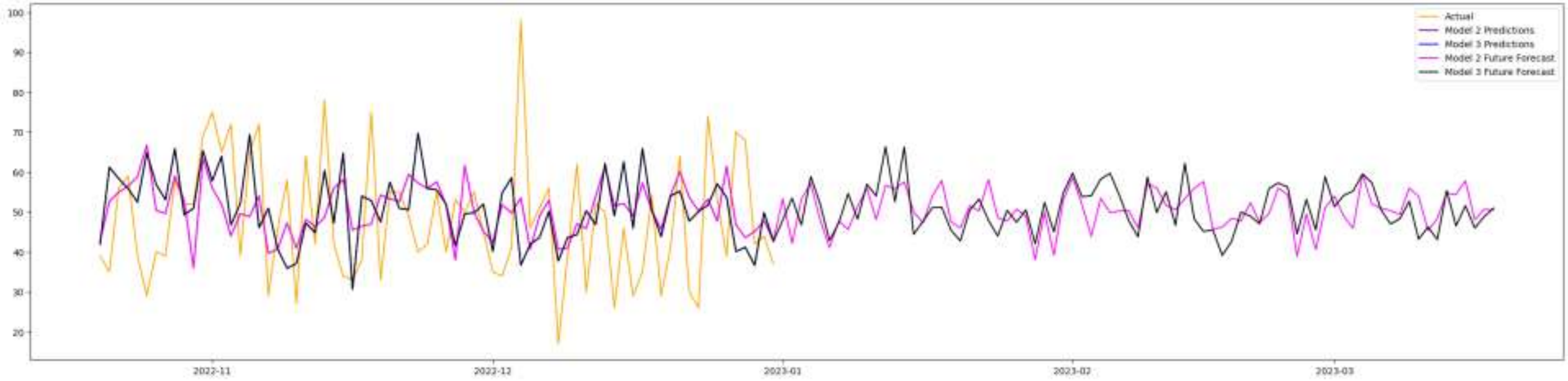
```
Model 3
Mean Absolute Error (MAE): 13.42
Mean Squared Error (MSE): 294.06
Root Mean Squared Error (RMSE): 17.15
Mean Absolute Percentage Error (MAPE): 32.04%
```

Actual vs Prediction



Model 1 : Auto Arima
Model 2 : ACF & PCF Plot
Model 3 : Pandas Autocorrelation Plot

ARIMA Modeling



Model 2 Future Forecast
Model 3 Future Forecast

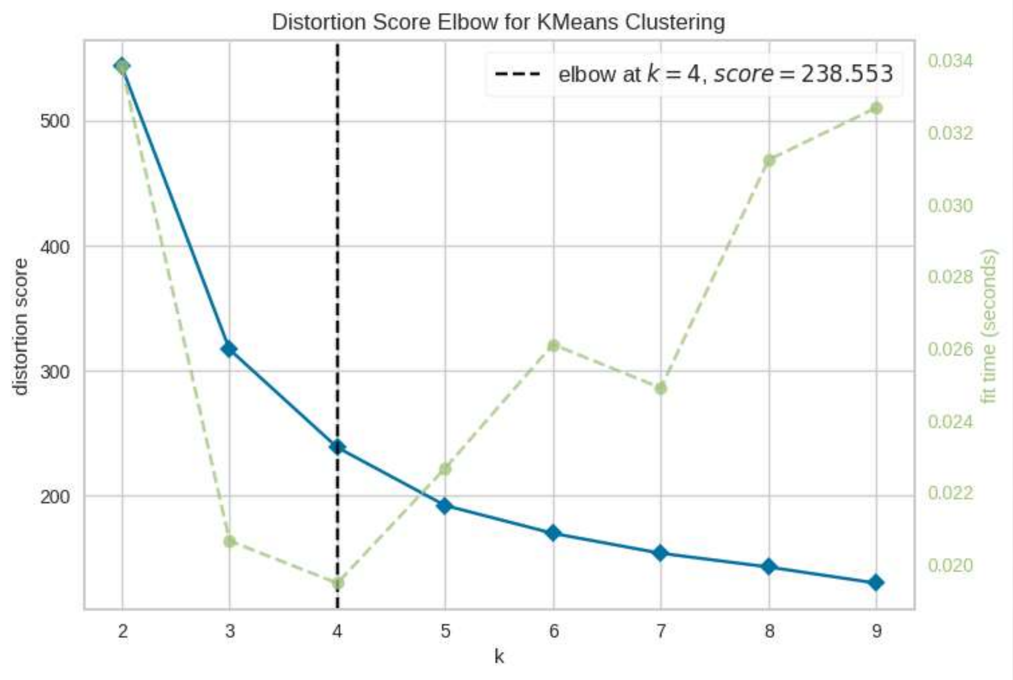
Challenge 4

Customer Segmentations Using K-Means Clustering

Finding K-Value

Model 1: Elbow Method

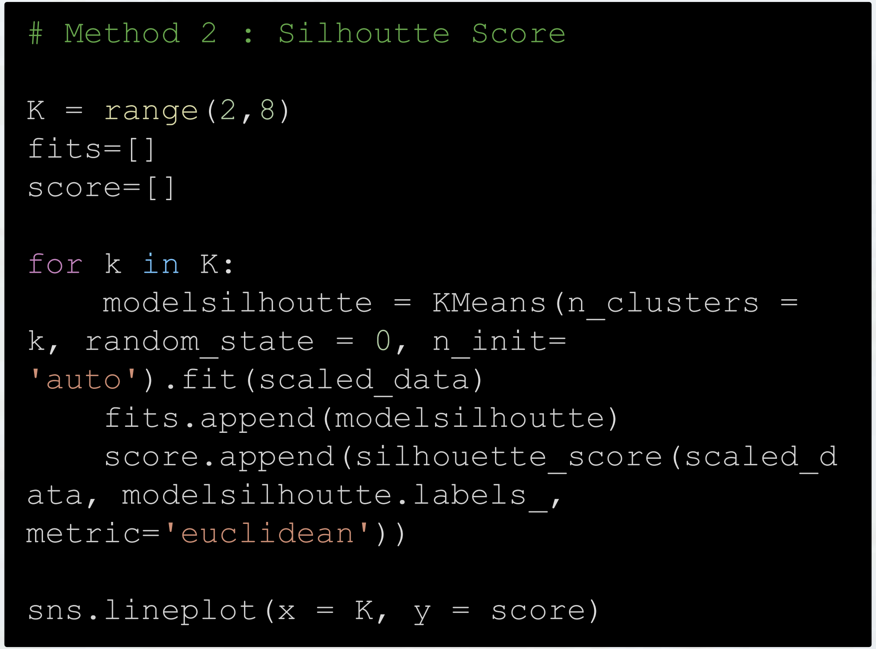
```
from yellowbrick.cluster.elbow import KElbowVisualizer

#Elbow Method with yellowbrick library
visualizer = KElbowVisualizer(kmeanModel, k=(2,10))
visualizer.fit(X_std)
visualizer.show()
```



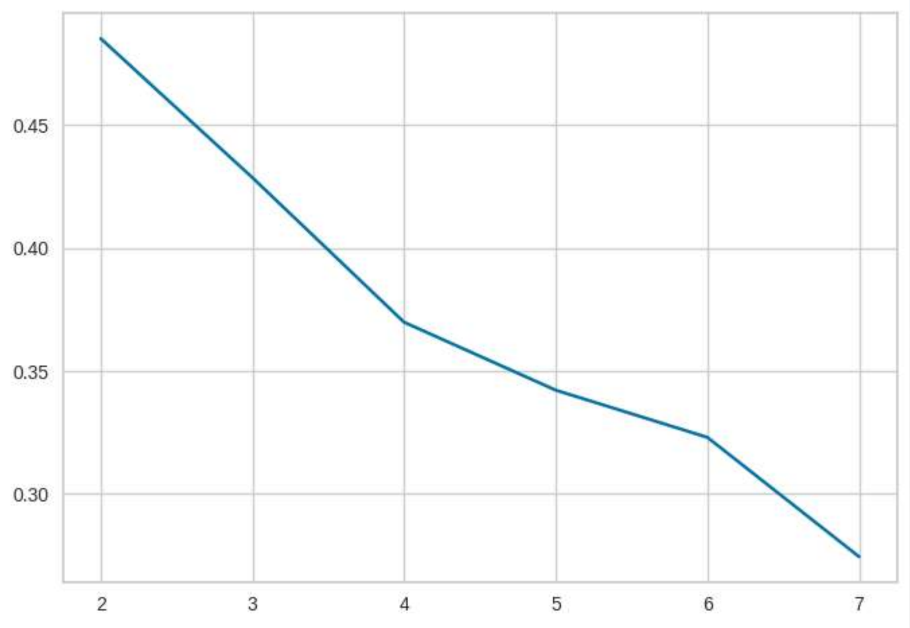
K = 4

Finding K-Value

Model 2: Silhoutte Score



```
# Method 2 : Silhoutte Score

K = range(2,8)
fits=[]
score=[]

for k in K:
    modelsilhoutte = KMeans(n_clusters =
k, random_state = 0, n_init=
'auto').fit(scaled_data)
    fits.append(modelsilhoutte)
    score.append(silhouette_score(scaled_d
ata, modelsilhoutte.labels_,
metric='euclidean'))

sns.lineplot(x = K, y = score)
```

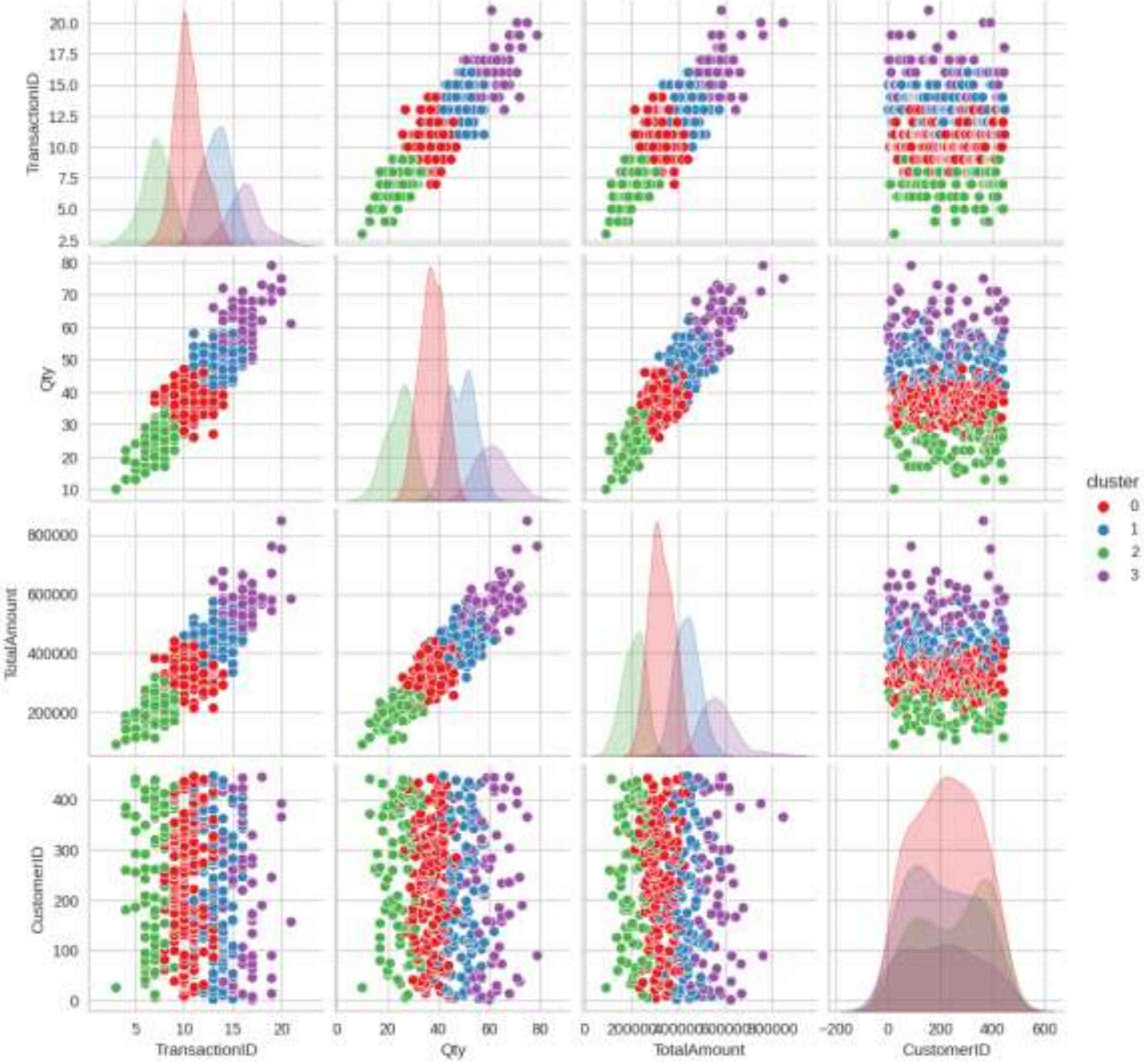


K = 4

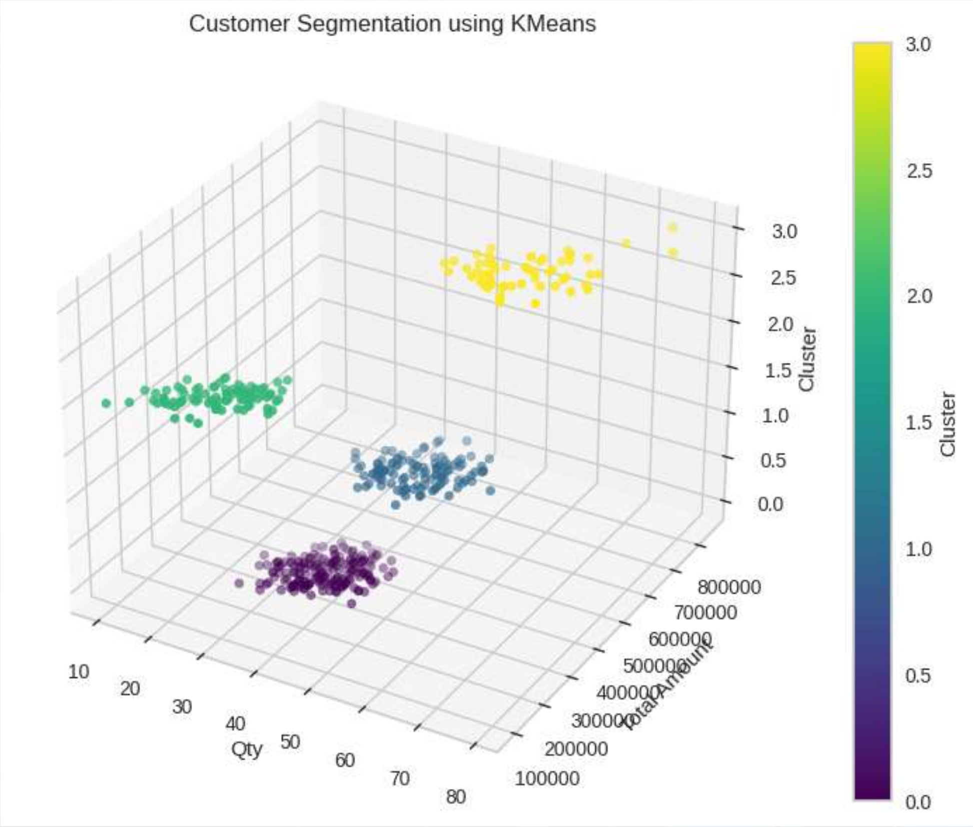
KALBE Nutritionals

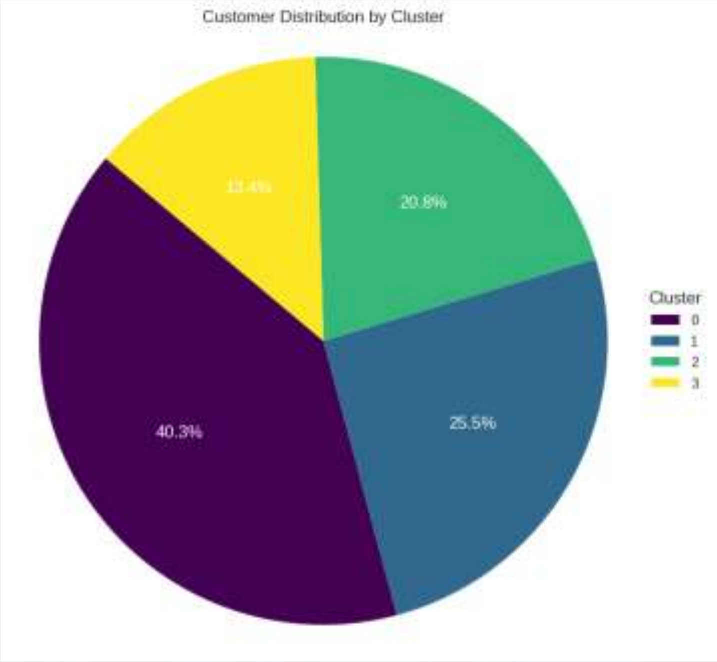
Rakamin Academy

Customer Segmenting



Customer Segmenting





Customer Result

- **Cluster 0** = Customers who make moderate to large purchases in terms of both quantity and total amount. These are customers who spend significant sums of money.
- **Cluster 1** = customers with smaller purchases and relatively lower spending. These are customers who make smaller purchases on a more affordable scale.
- **Cluster 2** = customers with large purchases and significant spending. This cluster signifies customers with a higher need or preference for larger quantities and higher spending.
- **Cluster 3** = customers with moderate purchases and moderate spending. These customers have moderate needs or preferences in terms of quantity and spending.

Business Recomendation

Cluster 1: "Budget Shoppers"

These are customers with smaller purchases and relatively lower spending.

- Implement targeted promotional campaigns for Kalbe Nutritionals' affordable nutritional products that cater to the needs of customers in Cluster 1. Emphasize the value and affordability of Kalbe Nutritionals products.
- Increase awareness about the benefits of Kalbe Nutritionals' nutritional products through educational programs highlighting the importance of good nutrition for overall health, particularly during specific life stages. This will help customers understand the significance of consuming high-quality nutrition.
- Provide special discounts or shopping vouchers as incentives for Cluster 1 customers who make repeat purchases or buy multiple products at once.

Business Recomendation

Cluster 2: "Nutrition Enthusiasts"

These are customers with large purchases and significant spending.

- Focus on developing specialized and nutrient-rich nutritional products for Kalbe Nutritionals that cater to the specific needs of customers in Cluster 2, particularly for adults with higher requirements and preferences in terms of quantity and spending. Highlight that Kalbe Nutritionals products in Cluster 2 are the optimal choice for meeting specific nutritional needs.
- Establish relationships with doctors or prominent clinics that serve clients with elevated nutritional needs. This way, Kalbe Nutritionals products can be recommended to patients by healthcare professionals.
- Organize special events involving healthcare professionals to educate customers in Cluster 2 about the benefits and advantages of Kalbe Nutritionals' nutritional products in fulfilling advanced nutritional requirements.

Business Recomendation

Cluster 3: "Quality Seekers"

These are customers with moderate purchases and moderate spending.

- Enhance the availability of Kalbe Nutritionals' most popular nutritional products that are highly sought after by customers in Cluster 3. Ensure these products are consistently well-stocked.
- Increase communication about the quality and latest technology utilized in the production of Kalbe Nutritionals nutritional products. This will instill greater confidence in customers from Cluster 3 regarding the quality of the products they purchase.
- Concentrate on developing nutritional products for Kalbe Nutritionals that contribute to strengthening the immune system and overall health. Given that customers in Cluster 3 are seeking reliable products for their well-being, emphasize the significance of these offerings.

KALBE Nutritionals

Rakamin Academy

Repository

https://github.com/Ruthyohanna11/VIX-Kalbe-Nutritionals

Thank You

Rakamin Academy X KALBE Nutritionals