Rakamin Academy x KALBE Nutritionals

Machine Learning Project (Regression and Clustering)

Kalbe Nutritionals Data Scientist Project Based Internship Program

Presented by
Vicky Tanamal

Vicky Tanamal

About Me

Just graduated from Data Science Bootcamp at Rakamin Academy and ready to switch career to become Data Scientist.

Experience

- Project-Based Virtual Intern : Data Scientist Home Credit Indonesia x Rakamin Academy
- Project-Based Virtual Intern : Data Scientist id/x partners x Rakamin Academy
- Sales Engineer at PT Sinergi Giat Perkasa

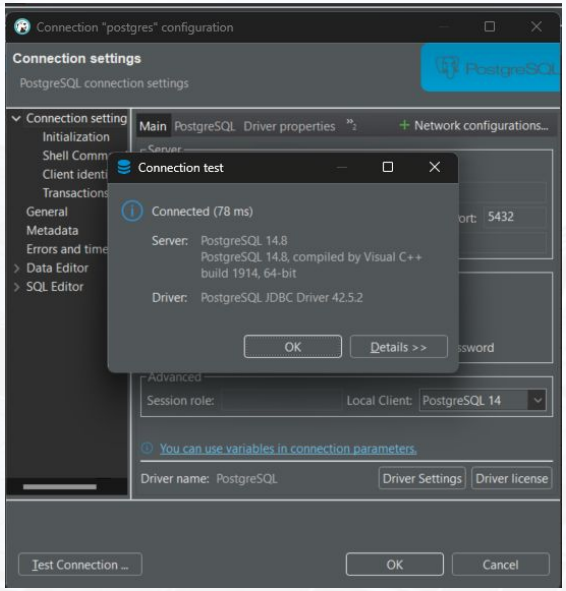
Rakamin Academy

EDA on DBeaver

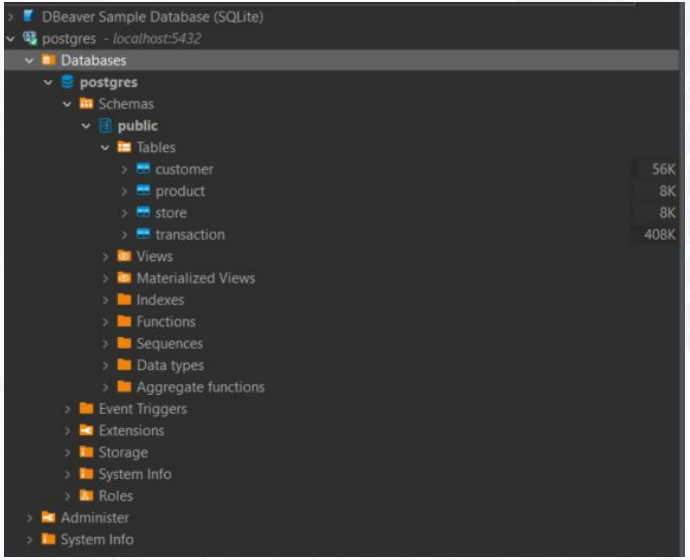
Connect to the database and get the insight from it.

Rakamin Academy

Connect DBeaver to PostgreSQL and test the connection

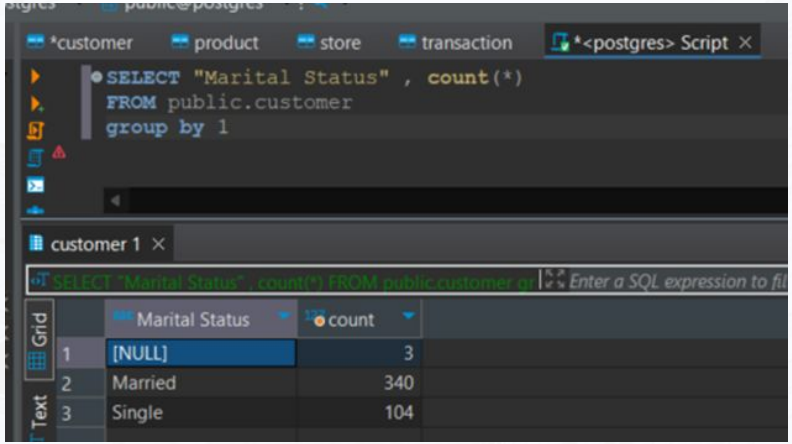


After the connection has connected, then import the data

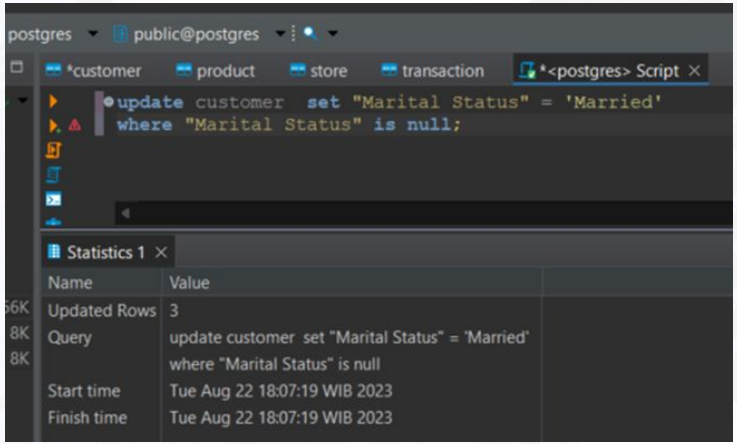


Rakamin Academy

There are NULL values in Marital Status

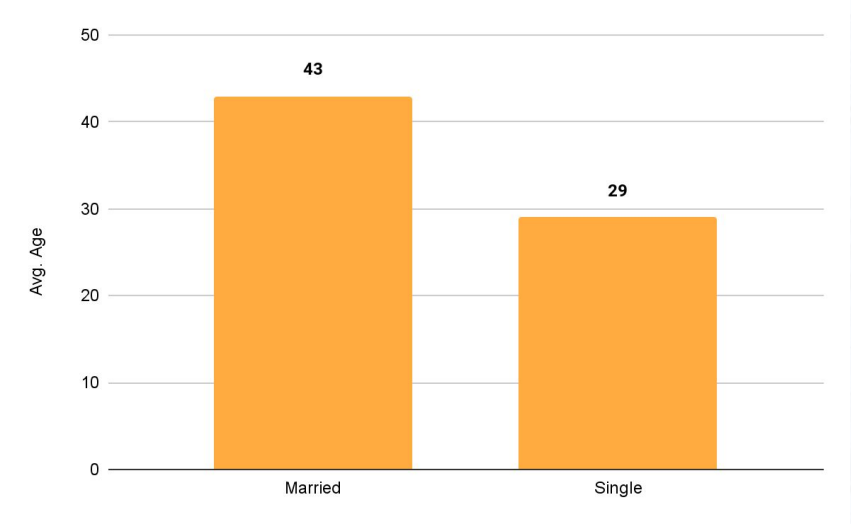


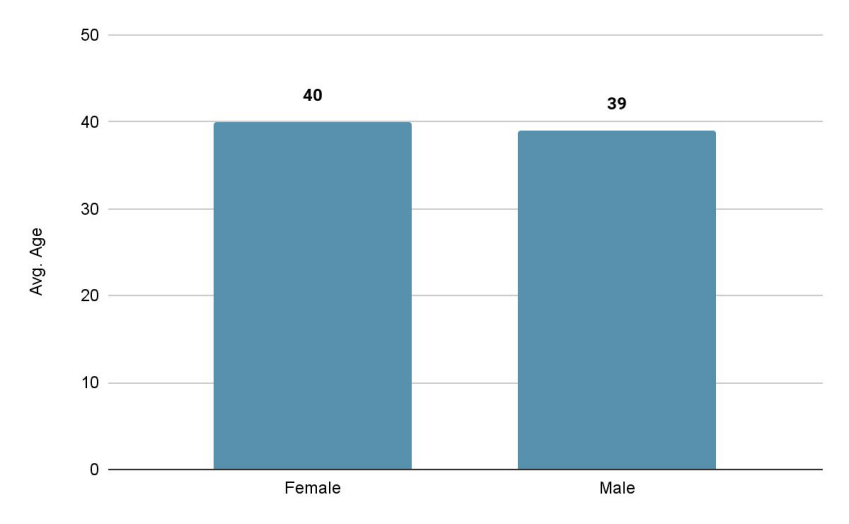
Fill the NULL values with Mode



Because Marital Status is categorical, we can fill the NULL value using **mode** which is in this case the mode is **Married.**

Customer's Average Age based on Marital Status and Gender





Average age based on Marital Status for Married is 43 and for Single is 29

Average age based on Gender for Female is 40 and for Male is 39

Store with The Most Total Sales Quantity

Store Name	Lingga
Total Qty	738

Product with The Most Sold Total Amount

Product Name	Cheese Stick
Total Amount	27.615.000

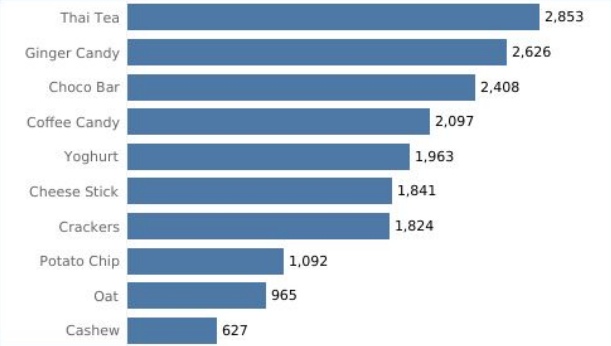
Rakamin Academy

Dashboard on Tableau

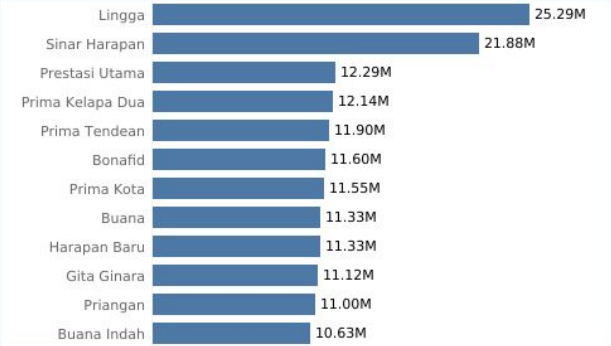
Create dashboard for Sales Report on Tableau.

Sales Report Kalbe Nutritionals

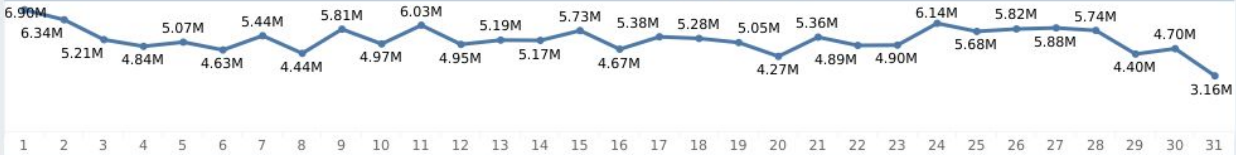
Total Sales Quantity by Product



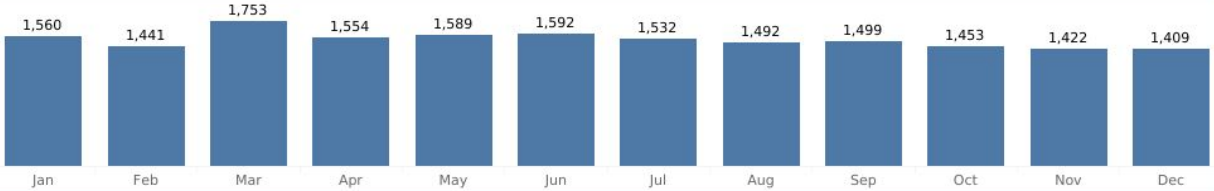
Total Sales Amount by Store Name



Total Amount Day by Day



Sales Quantity Over Months



Rakamin Academy

Sales Report Dashboard

Click here to view on Tableau Public

Rakamin Academy

Predictive Modelling

Create Machine Learning Modelling Regression (Time Series)

Load Dataset

```
df_transaction = pd.read_csv('Case Study - Transaction.csv', sep=';')
df_store = pd.read_csv('Case Study - Store.csv', sep=';')
df_product = pd.read_csv('Case Study - Product.csv', sep=';')
df_customer = pd.read_csv('Case Study - Customer.csv', sep=';')
```

Handling Missing Values

```
df_customer.isnull().sum()
✓ 0.0s

CustomerID        0
Age               0
Gender            0
Marital Status    3
Income            0
dtype: int64
```

```
df_customer = df_customer.apply(lambda x: x.fillna(x.mode()[0]))
df_customer.isnull().sum()
✓ 0.0s

CustomerID        0
Age               0
Gender            0
Marital Status    0
Income            0
dtype: int64
```

Changing Data Type

```
# Change data type
df_transaction['Date'] = pd.to_datetime(df_transaction['Date'])
df_store['Latitude'] = df_store['Latitude'].str.replace(',','.').astype(float)
df_store['Longitude'] = df_store['Longitude'].str.replace(',','.').astype(float)
df_customer['Income'] = df_customer['Income'].str.replace(',','.').astype(float)
✓ 0.0s
```

Merge Data

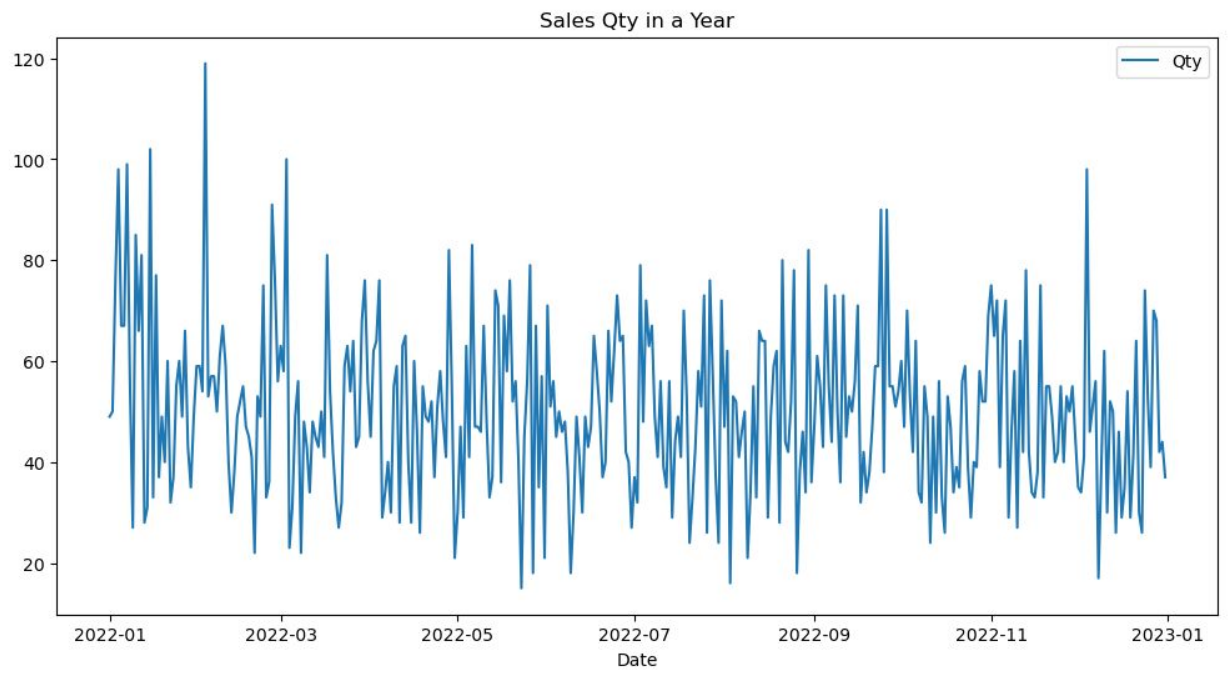
```
# Merge data
merged_df = pd.merge(df_transaction, df_store, on='StoreID', how='inner')
merged_df = pd.merge(merged_df, df_product, on='ProductID', how='inner')
merged_df = pd.merge(merged_df, df_customer, on='CustomerID', how='inner')
✓ 0.0s
```

Overview Data

	TransactionID	CustomerID	Date	ProductID	Price_x	Qty	TotalAmount	StoreID	StoreName	GroupStore	Type	Latitude	Longitude	Product Name	Age	Gender	Marital Status	Income
0	TR11369	328	2022-01-01	P3	7500	4	30000	12	Prestasi Utama	Prestasi	General Trade	-2.990934	104.756554	Crackers	36	0	Married	10.53
1	TR57013	328	2022-09-15	P7	9400	6	56400	1	Prima Tendean	Prima	Modern Trade	-6.200000	106.816666	Coffee Candy	36	0	Married	10.53
2	TR97172	328	2022-05-21	P1	8800	5	44000	1	Prima Tendean	Prima	Modern Trade	-6.200000	106.816666	Choco Bar	36	0	Married	10.53
3	TR67395	328	2022-01-22	P8	16000	3	48000	11	Sinar Harapan	Prestasi	General Trade	0.533505	101.447403	Oat	36	0	Married	10.53
4	TR45738	328	2022-12-29	P2	3200	3	9600	11	Sinar Harapan	Prestasi	General Trade	0.533505	101.447403	Ginger Candy	36	0	Married	10.53
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5015	TR37670	193	2022-09-26	P5	4200	2	8400	7	Buana Indah	Buana	General Trade	3.316694	114.590111	Thai Tea	42	0	Married	20.64
5016	TR98043	385	2022-06-27	P2	3200	7	22400	11	Sinar Harapan	Prestasi	General Trade	0.533505	101.447403	Ginger Candy	41	1	Married	15.84
5017	TR91332	385	2022-09-01	P9	10000	1	10000	10	Harapan Baru	Harapan Baru	General Trade	3.597031	98.678513	Yoghurt	41	1	Married	15.84
5018	TR88968	385	2022-08-21	P9	10000	6	60000	9	Lingga	Lingga	Modern Trade	-3.654703	128.190643	Yoghurt	41	1	Married	15.84
5019	TR90487	385	2022-12-24	P9	10000	5	50000	8	Sinar Harapan	Harapan Baru	General Trade	5.548290	95.323753	Yoghurt	41	1	Married	15.84

5020 rows × 18 columns

Sales Qty Data Overview



Sales Qty is very cyclical, the changes day by day is quite significant.

Train and Test Data

```python
# Splitting data train and test
print(df_ts.shape)
ts_train = df_ts.iloc[:-92] # First 9 months for training
ts_test = df_ts.iloc[-92:] # Last 3 months for testing
print(ts_train.shape,ts_test.shape)
```

✓ 0.0s

```
(365, 1)
(273, 1) (92, 1)
```

```python
ts_train.head()
```

✓ 0.0s

	Qty
Date	
2022-01-01	49
2022-01-02	50
2022-01-03	76
2022-01-04	98
2022-01-05	67

```python
ts_test.head()
```

✓ 0.0s

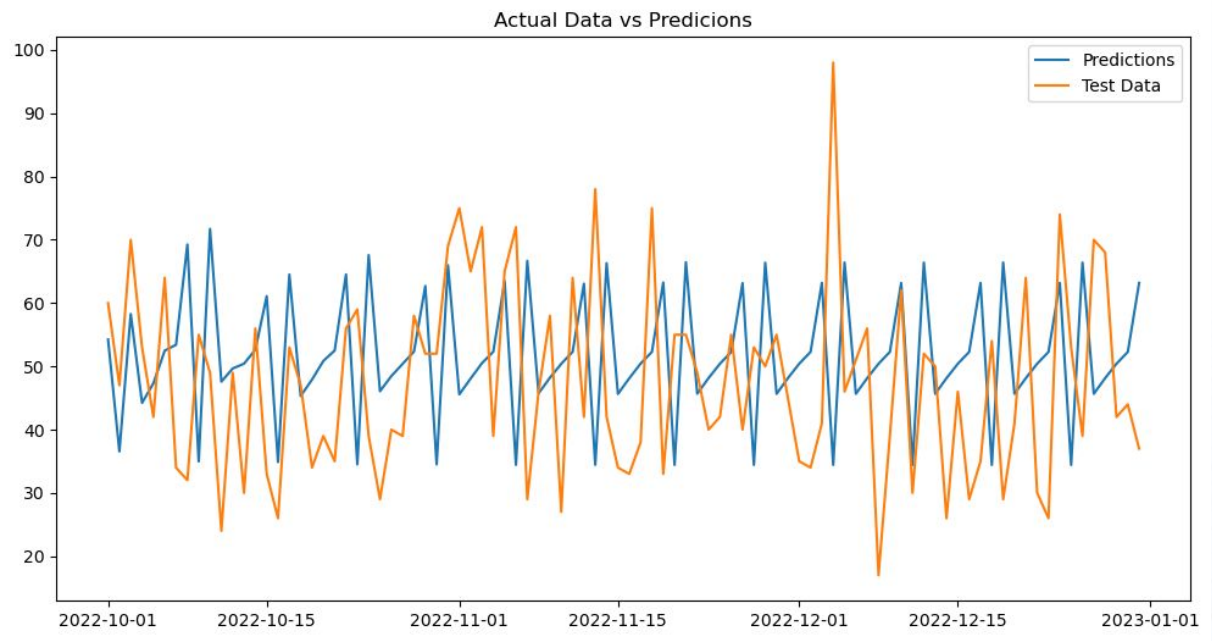
	Qty
Date	
2022-10-01	60
2022-10-02	47
2022-10-03	70
2022-10-04	53
2022-10-05	42

Modelling

```python
model2=sm.tsa.statespace.SARIMAX(ts_train,order=(3, 0, 2),
                                 seasonal_order=(1,1,0,7))
results=model2.fit()
results.summary()
```

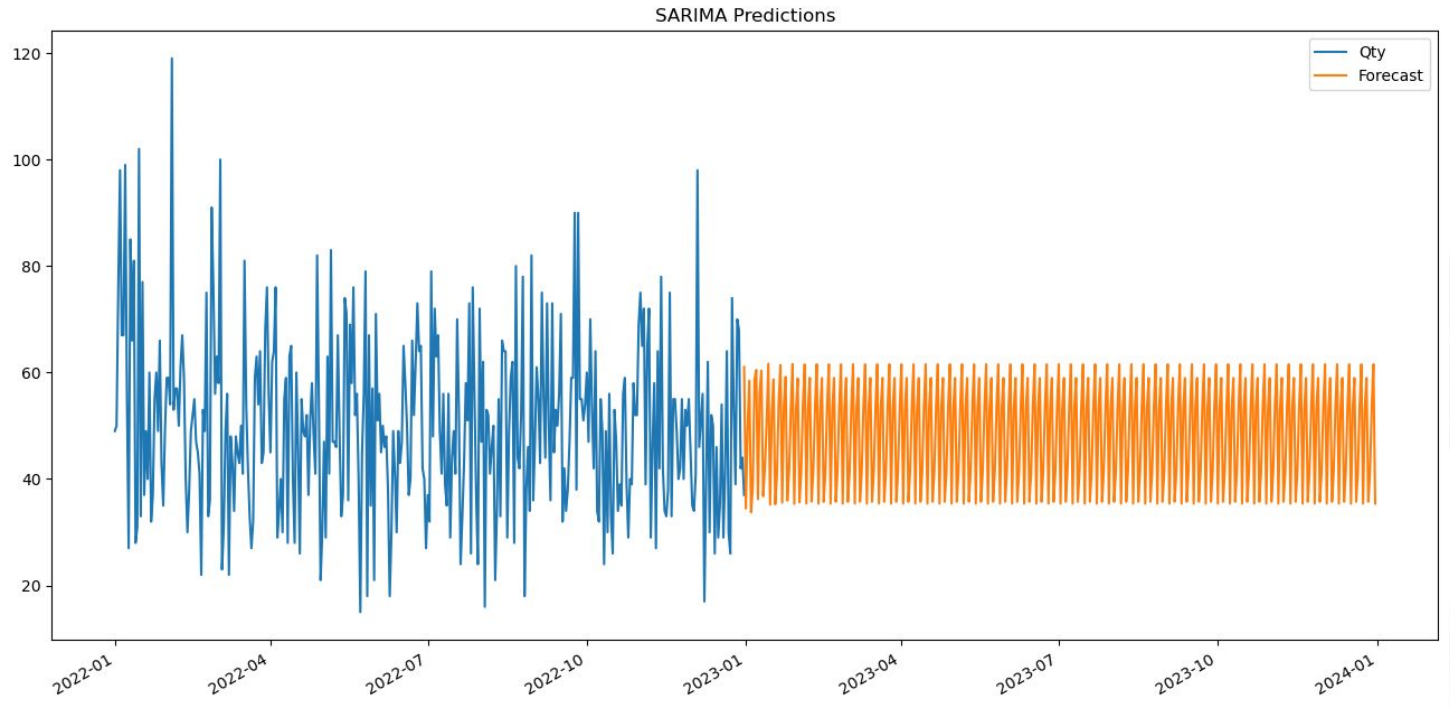
✓ 1.2s

Result from Modelling



RMSE Value: 19.69852579129112
R-squared Value: -0.7496914917708499
MAE Value: 16.405871447185074

Forecasting for Next Year



For the forecasting data it's not really similar with the Sales actual data, this is because the actual data is very cyclical and the forecasting is just take the mean of the sales data.

Rakamin Academy

Clustering Modelling

Create Machine Learning Modelling Clustering using K-Means

Rakamin Academy

Data Preparation

	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
...	...	...	...	...
442	443	16	59	485100
443	444	18	62	577700
444	445	18	68	587200
445	446	11	42	423300
446	447	13	42	439300

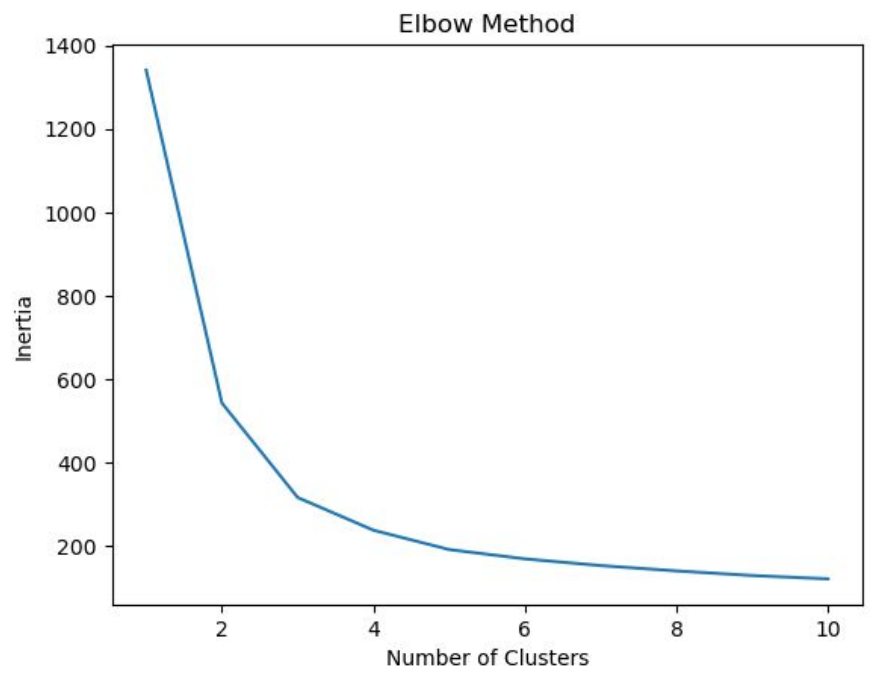
447 rows × 4 columns

Standardization Data

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

447 rows × 3 columns

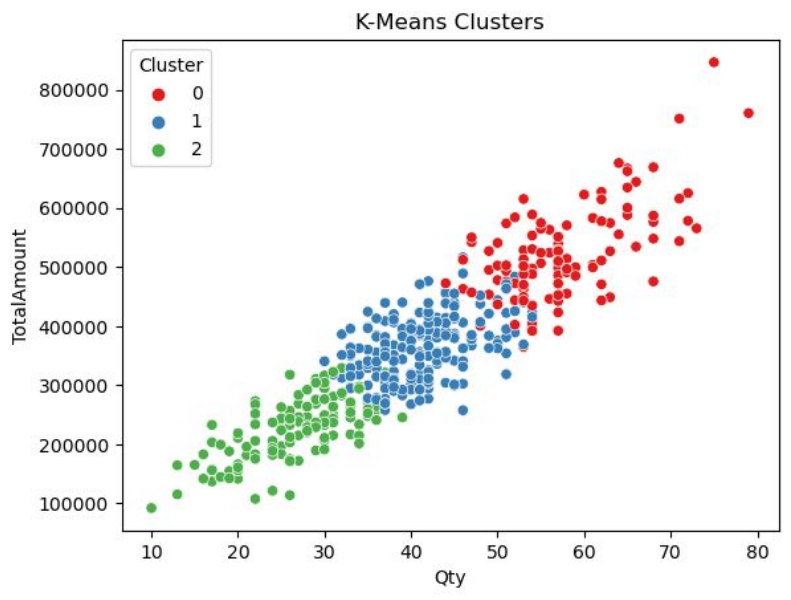
Determine Number of Clusters

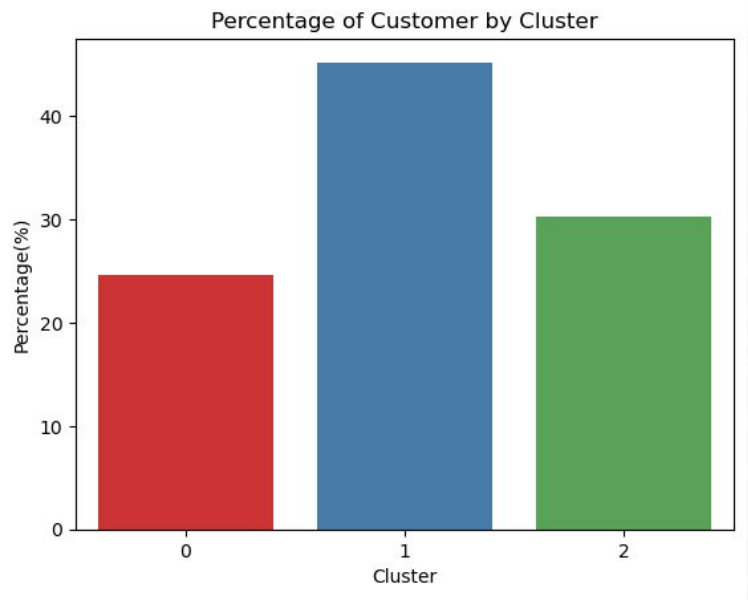



```
Number of Clusters = 4
Silhouette Score: 0.3177554917332133

Number of Clusters = 3
Silhouette Score: 0.40585912730645
```

Clustering Result





Cluster 1 have the most customer, but Cluster 0 has higher Qty, Total Amount and Transaction. Meanwhile Cluster 2 has the lowest for Qty, Total Amount and Transaction.

Business Recommendation

1. Cluster 0

We must keep this customer, because this Cluster has high value. We can give them like Loyalty Programs for repeat purchases or buy product by passing the limit of shopping and on that program we can give some points and the points can be exchanged with our another product for free.

2. Cluster 1

Most of customer in this Cluster, so we must increase buying rate of the customer. We can give them discount voucher after they bought product, so they consider to buy another product using that voucher.

3. Cluster 2

We must do some campaigns that can make our products become their top of mind to increase the buying rate. We must give them knowledge our product, why must choose and buy our product and we can highlight the good review for our product to proof that our product is good and worth to buy.

For More Information

GitHub :
https://github.com/vickytanamal/Machine-Learning-Time-Series-and-Regression-Kalbe-Nutritionals
Tableau:
https://public.tableau.com/views/KalbeNutritionalsVIX/Dashboard1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

Thank You

Rakamin Academy X KALBE Nutritionals

Video Presentation Here

https://youtu.be/WWwAkN8gATM