Project-based Internship

Kalbe Nutritionals X Rakamin Academy

Muhammad Ifzal Asril

Table of contents

Tools to Use

The following is a list of some of the tools that were used in the project :

1. Dbeaver
2. PostgreSQL
3. Jupyter Notebook
4. Tableau

01

About Dataset

This dataset consists of 4 csv files namely customer, store, product and transaction. It is a dummy data for FMCG case study within 1 year taken through membership program.

Explanation

1. Customer
- CustomerID : unique number of the customer
- Age : Age of the customer
- Gender : 0 Female, 1 Male
- Marital status : Married, Single (Not married / Formerly married)
- Income : Monthly income in million rupiah

2. Store
- StoreID : Unique store code
- StoreName : Store name
- GroupStore : Group name
- Type : Modern Trade, General Trade
- Latitude : Latitude Code
- Longitude : Longitude Code

3. Product
- ProductID : Product unique code
- Product Name : Product name
- Price: Price in Rupiah

4. Transaction
- TransactionID : Unique transaction code
- Date: Date of the transaction
- Qty : Number of items purchased
- Total Amount : Price x Qty

02

SQL Analysis

1. What is the average age of customers when in terms of marital status?

2. What is the average age of customers when by gender?

3. Determine the name of the store with the highest total quantity!

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

Task 1

Marital Status	average_age
Single	29.3846153846
Married	43.0382352941

Task 2

gender	average_age
0	40.326446281
1	39.1414634146

Task 3

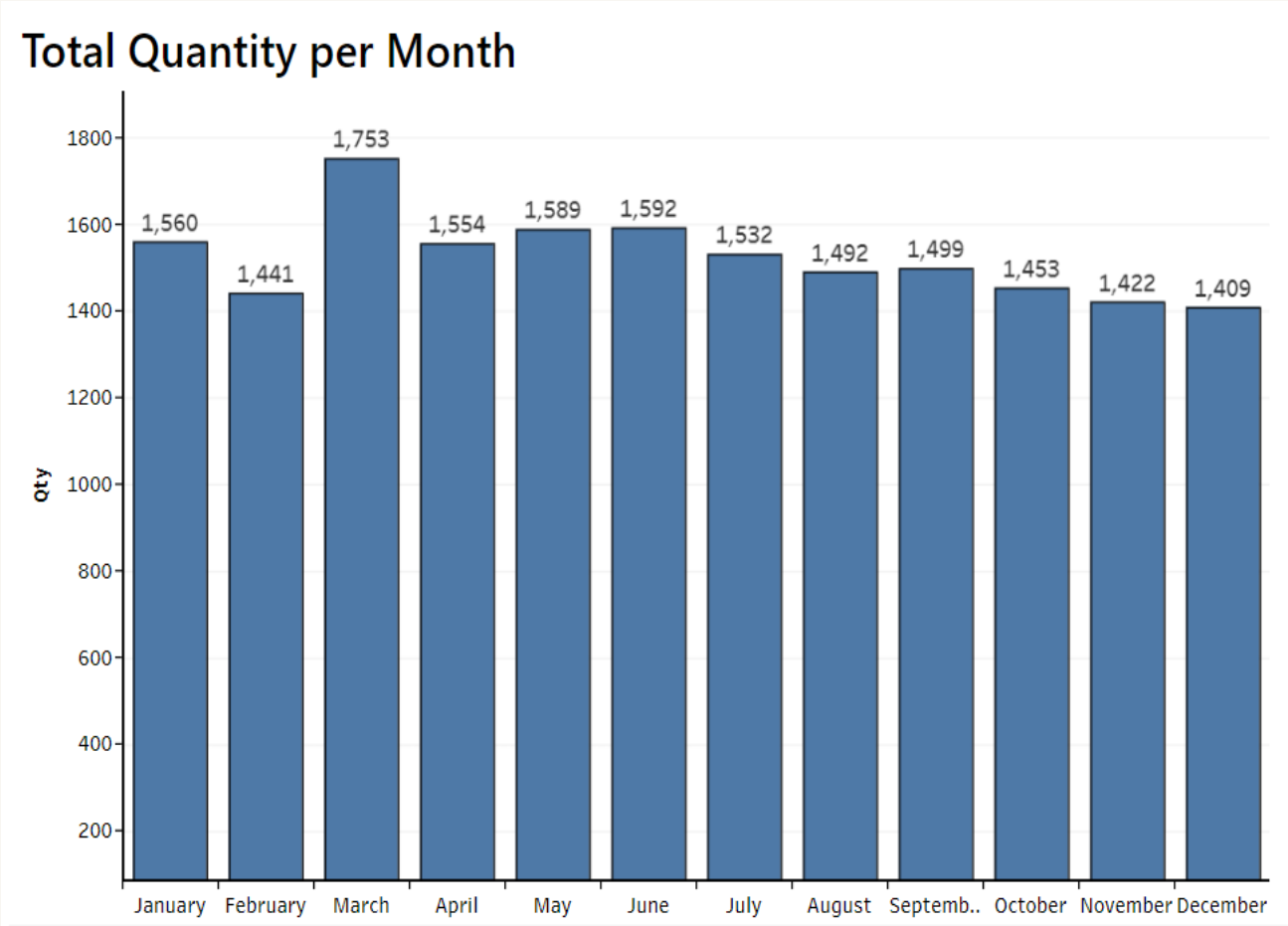
total_quantity	storename
5,324	Sinar Harapan
5,280	Buana
5,240	Prima Tendean
5,184	Prima Kelapa Dua
5,144	Harapan Baru
5,132	Bonafid
5,028	Sinar Harapan
4,956	Priangan
4,944	Gita Ginara
4,832	Buana Indah

Task 4

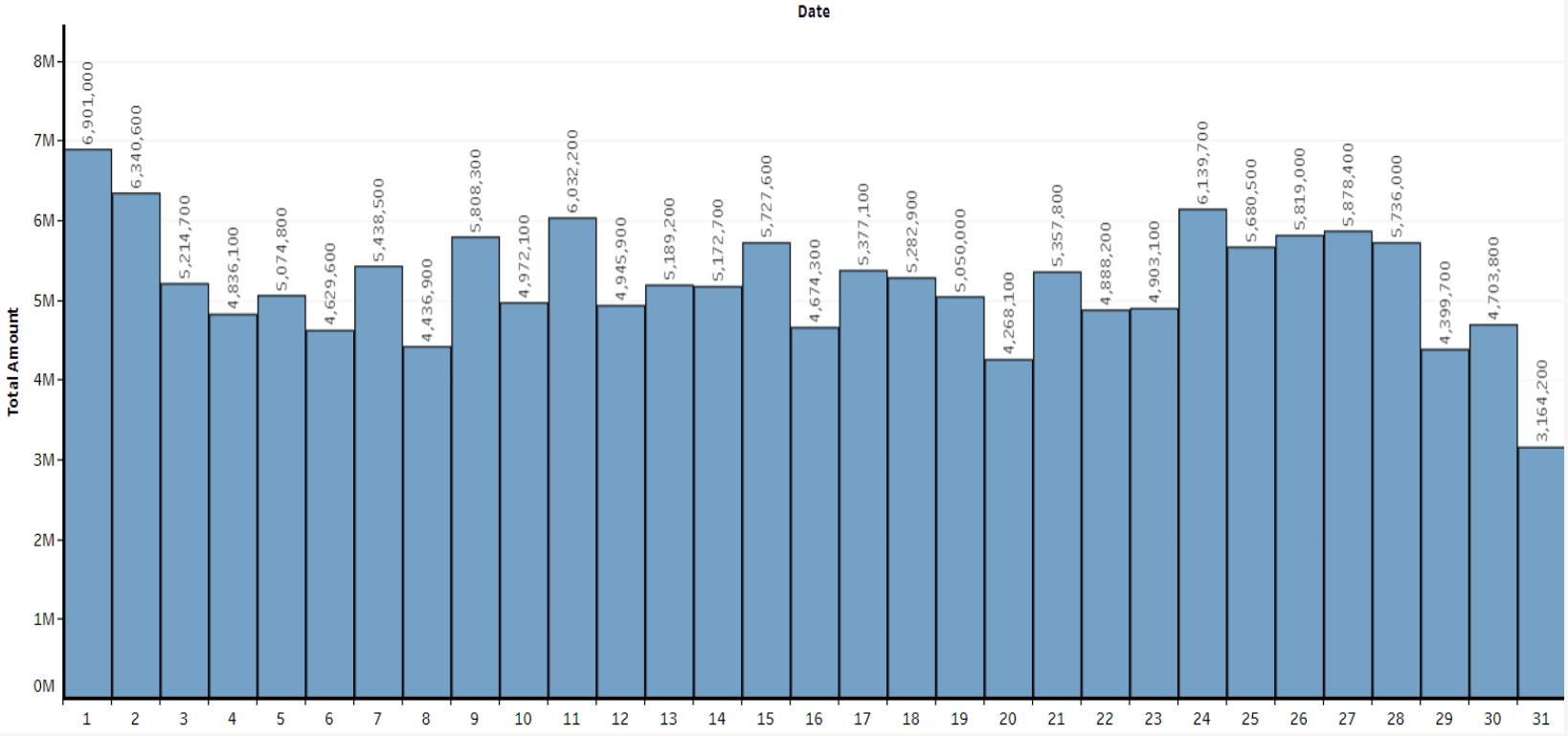
productid	Product Name	Total Amount
P10	Cheese Stick	110,460,000
P1	Choco Bar	84,761,600
P7	Coffee Candy	78,847,200
P9	Yoghurt	78,520,000
P8	Oat	61,760,000
P3	Crackers	54,720,000
P4	Potato Chip	52,416,000
P5	Thai Tea	47,930,400
P6	Cashew	45,144,000
P2	Ginger Candy	33,612,800

03

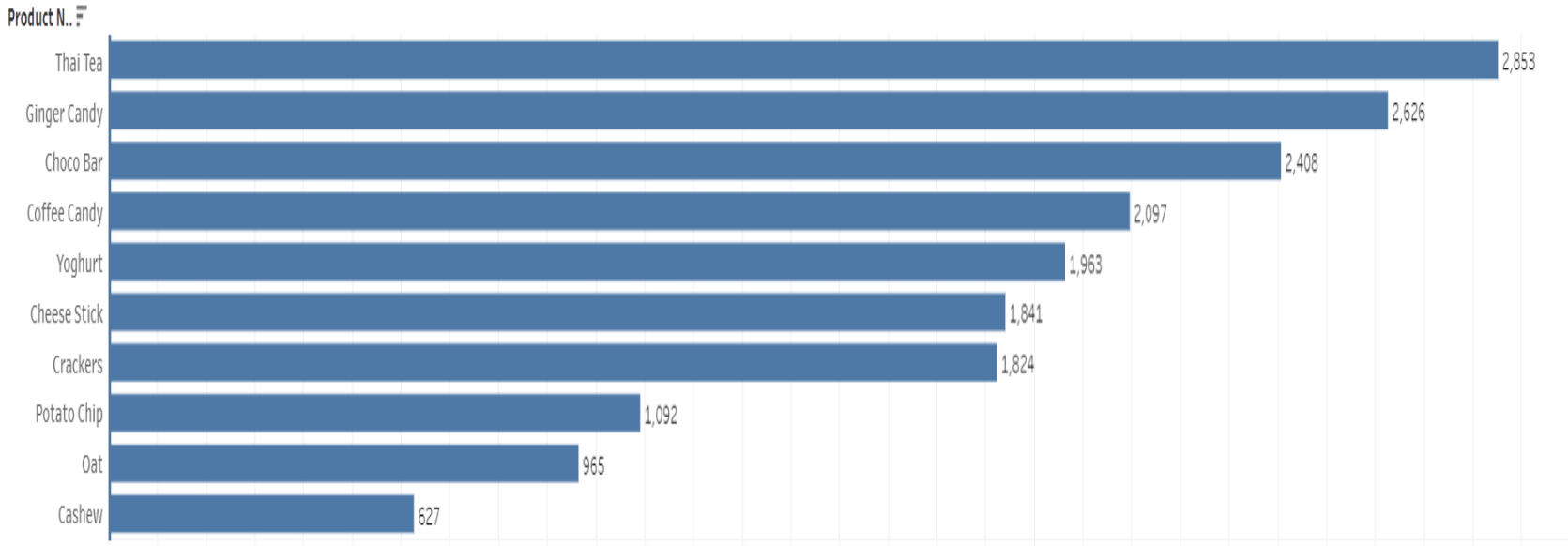
Analysis with Tableau

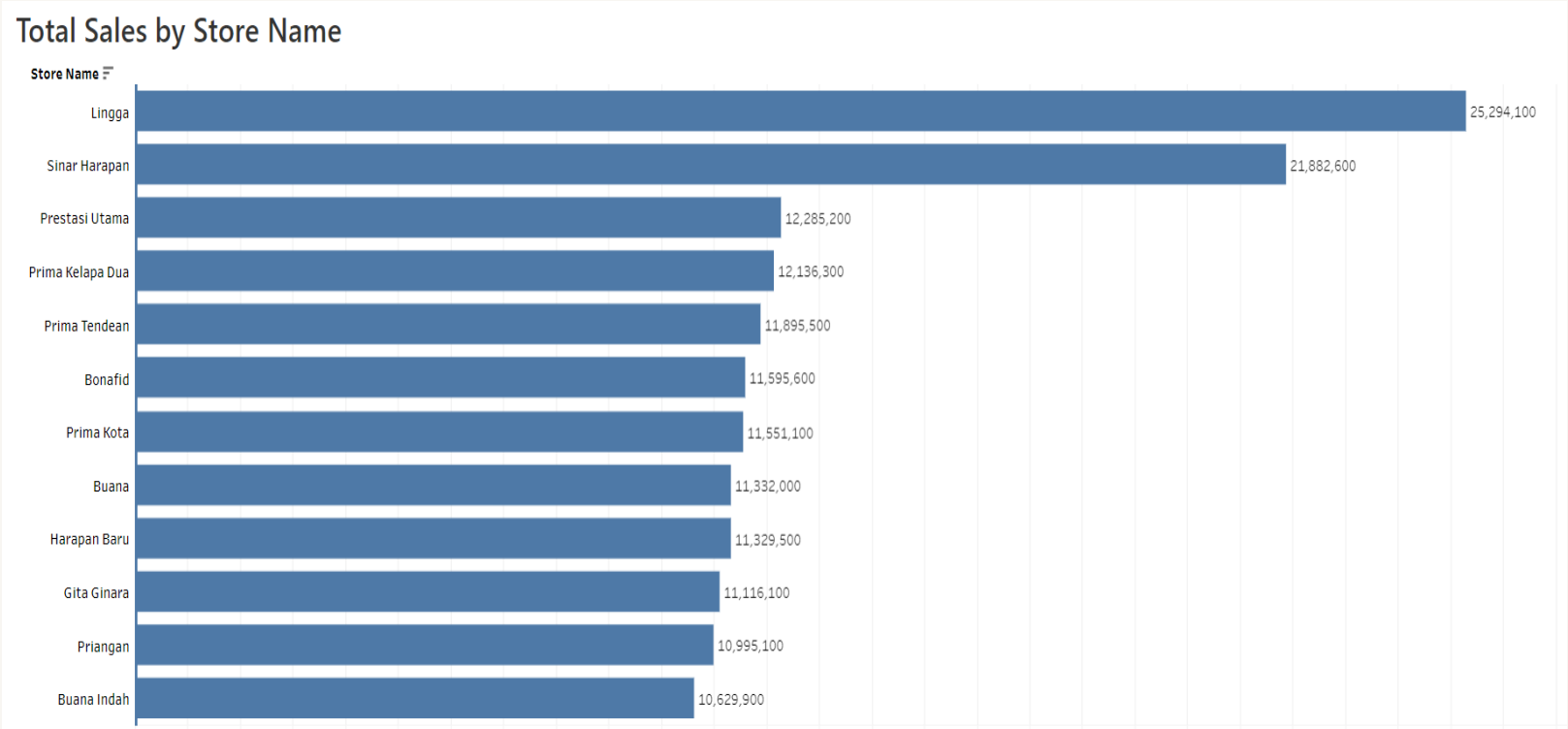

Total Quantity per Month
Qty
1800
1600
1400
1200
1000
800
600
400
200
1,560
1,441
1,753
1,554
1,589
1,592
1,532
1,492
1,499
1,453
1,422
1,409
January
February
March
April
May
June
July
August
Septemb..
October
November
December

Total Amount Day to Day



Number of Sales (qty) by Product Category

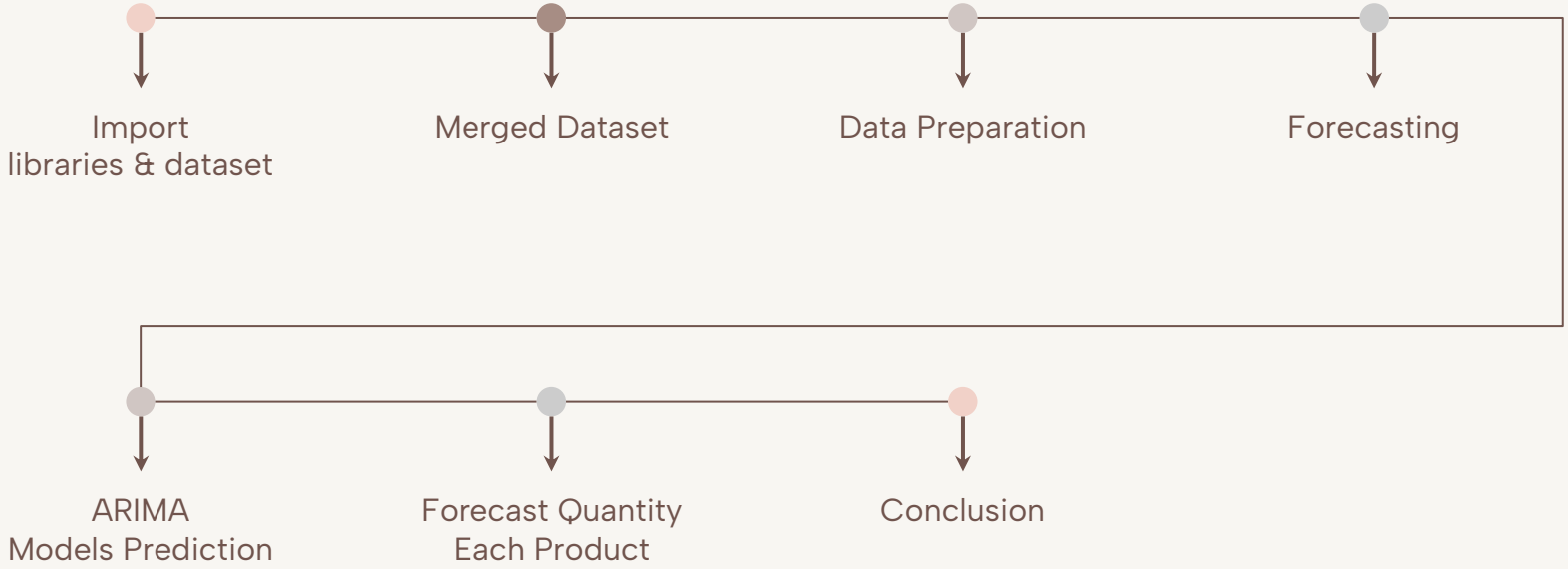



Total Sales by Store Name
Store Name
Lingga
25,294,100
Sinar Harapan
21,882,600
Prestasi Utama
12,285,200
Prima Kelapa Dua
12,136,300
Prima Tendean
11,895,500
Bonafid
11,595,600
Prima Kota
11,551,100
Buana
11,332,000
Harapan Baru
11,329,500
Gita Ginara
11,116,100
Priangan
10,995,100
Buana Indah
10,629,900

04

Model Machine Learning Regression

Regression Modeling Timeline



Merged Dataset

	transactionid	customerid	date	productid	qty	totalamount	storeid	product_name	price_product	storename	groupstore	type	age	gender	marital_status	income
492	TR39272	176	2022-12-21	P3	5	37500	12	Crackers	7500	Prestasi Utama	Prestasi	General Trade	54	1	Married	18,32
4481	TR95242	426	2022-05-25	P2	4	12800	7	Ginger Candy	3200	Buana Indah	Buana	General Trade	19	0	Single	2,16
3288	TR40987	69	2022-10-01	P1	10	88000	11	Choco Bar	8800	Sinar Harapan	Prestasi	General Trade	26	0	Single	2,84
4393	TR10854	49	2022-04-19	P8	1	16000	2	Oat	16000	Prima Kelapa Dua	Prima	Modern Trade	44	1	Married	13,48
3434	TR5234	127	2022-07-28	P7	3	28200	3	Coffee Candy	9400	Prima Kota	Prima	Modern Trade	27	0	Married	2,07

Data Preparation

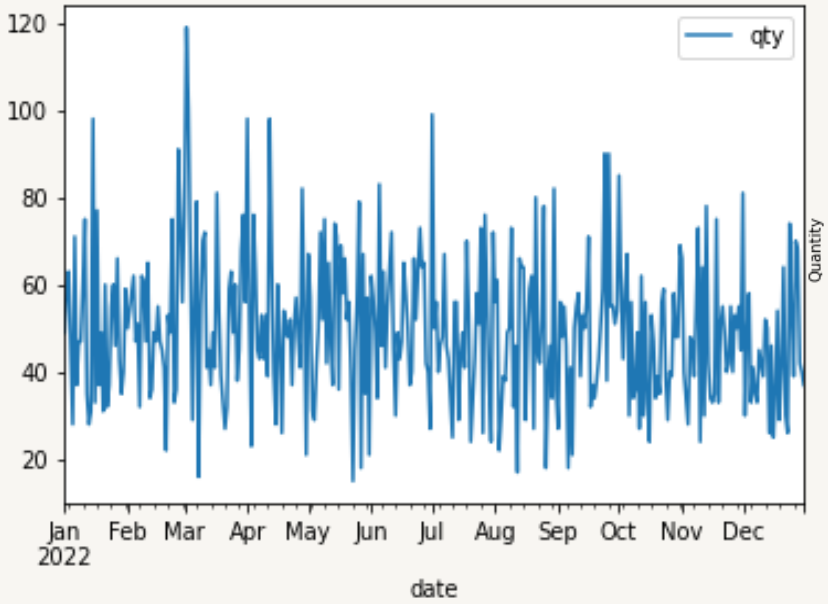
```
# checking null values
df.isna().sum()
```

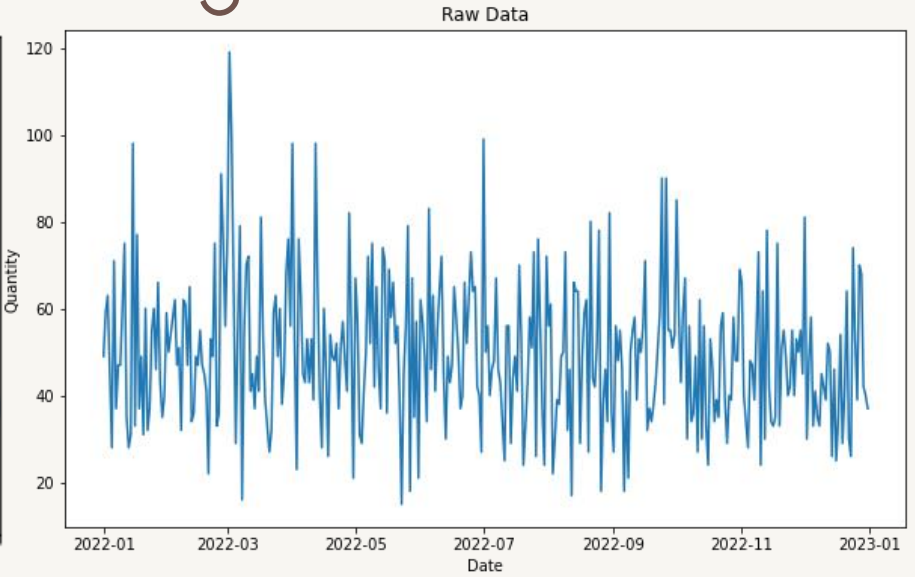
```
transactionid      0
customerid         0
date               0
productid          0
qty                0
totalamount        0
storeid            0
product_name       0
price_product      0
storename          0
groupstore         0
type               0
age                0
gender             0
marital_status    44
income             0
dtype: int64
```

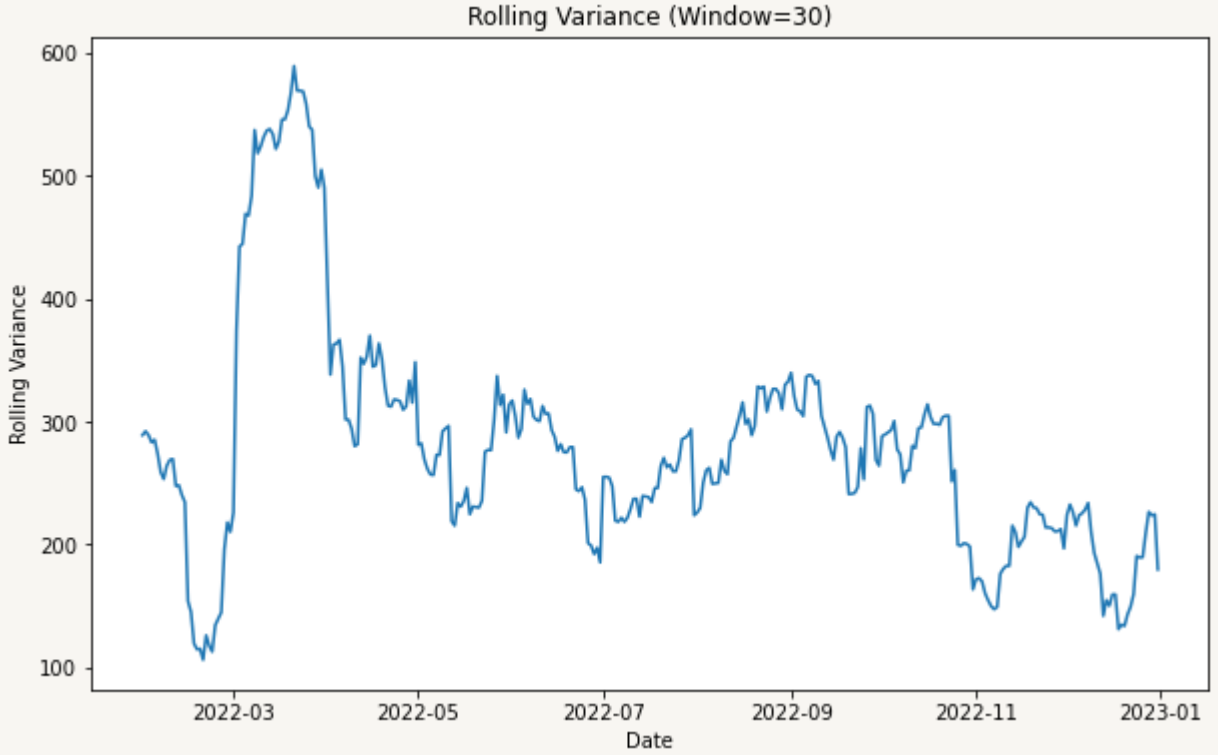
```
df.dropna(inplace=True)
df.isna().sum()
```

```
transactionid      0
customerid         0
date               0
productid          0
qty                0
totalamount        0
storeid            0
product_name       0
price_product      0
storename          0
groupstore         0
type               0
age                0
gender             0
marital_status     0
income             0
dtype: int64
```

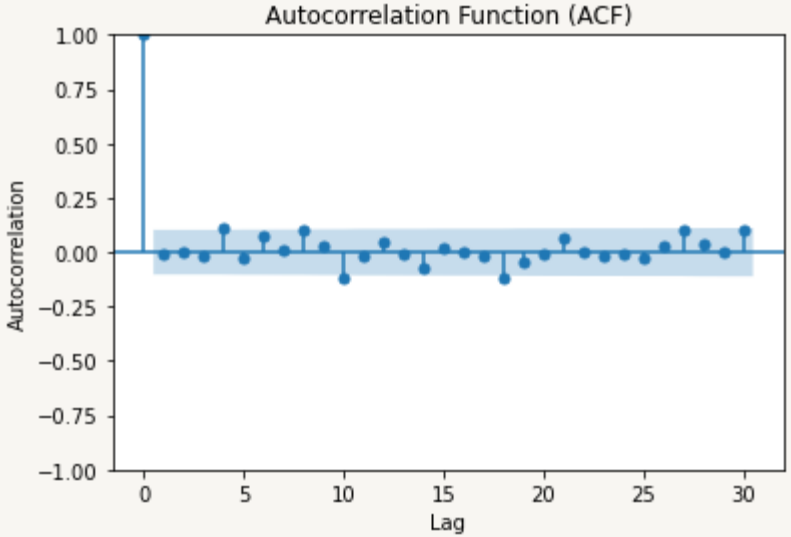
Forecasting

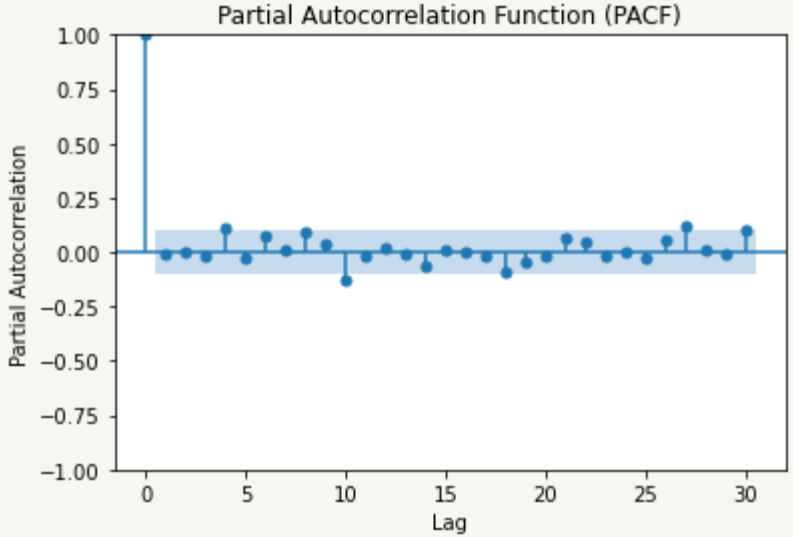




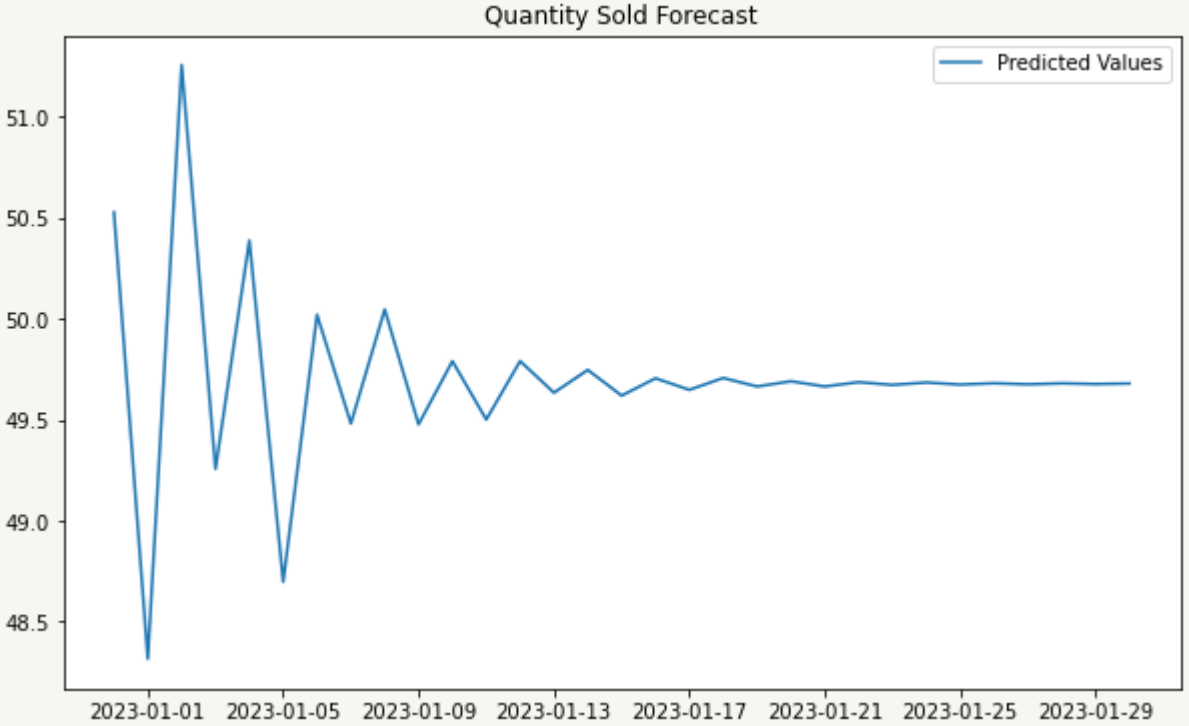

Rolling Variance (Window=30)
Rolling Variance
600
500
400
300
200
100
2022-03
2022-05
2022-07
2022-09
2022-11
2023-01
Date

Plot ACF & Plot PACF





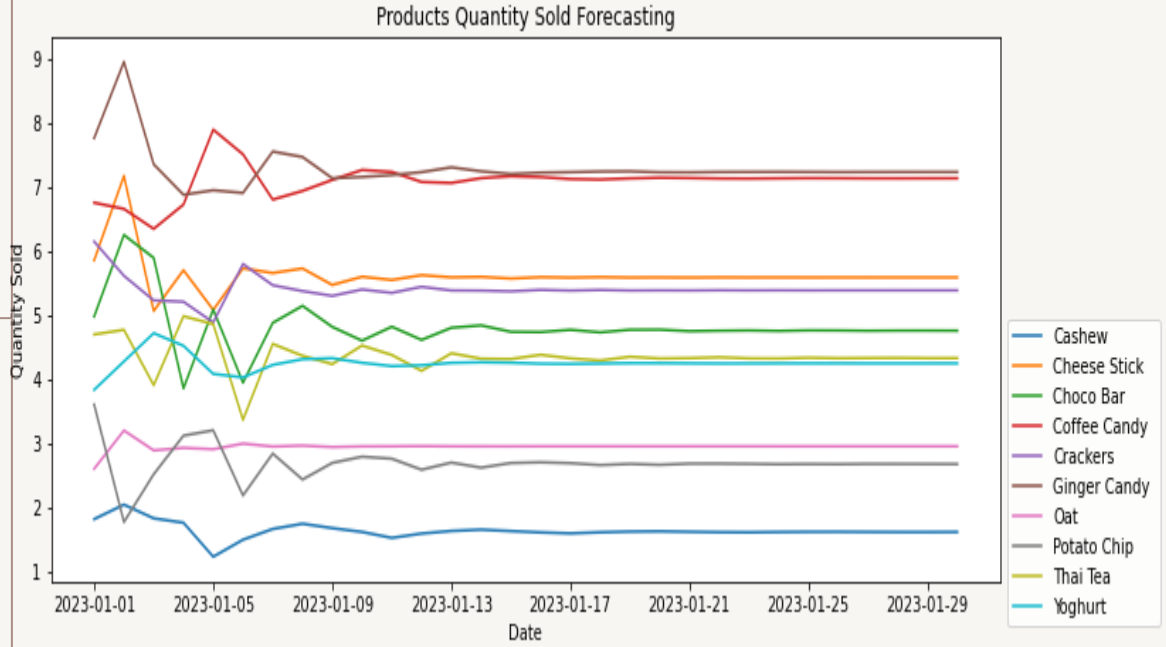
ARIMA Models Prediction



Forecast Quantity Each Product

date	Cashew	Cheese Stick	Choco Bar	Coffee Candy	Crackers	Ginger Candy	Oat	Potato Chip	Thai Tea	Yoghurt
2023-01-01	1.828707	5.865455	4.988922	6.762355	6.156308	7.770985	2.614091	3.614042	4.708479	3.845464
2023-01-02	2.054528	7.181429	6.259827	6.665999	5.623186	8.959238	3.210881	1.782759	4.780916	4.287678
2023-01-03	1.840936	5.074704	5.902763	6.355107	5.239667	7.361602	2.900499	2.529006	3.916482	4.728293
2023-01-04	1.772704	5.710972	3.867328	6.733553	5.218634	6.886524	2.939847	3.130321	4.992573	4.531827
2023-01-05	1.242324	5.091668	5.097367	7.901850	4.896050	6.956514	2.918980	3.213889	4.869936	4.093337

Conclusion

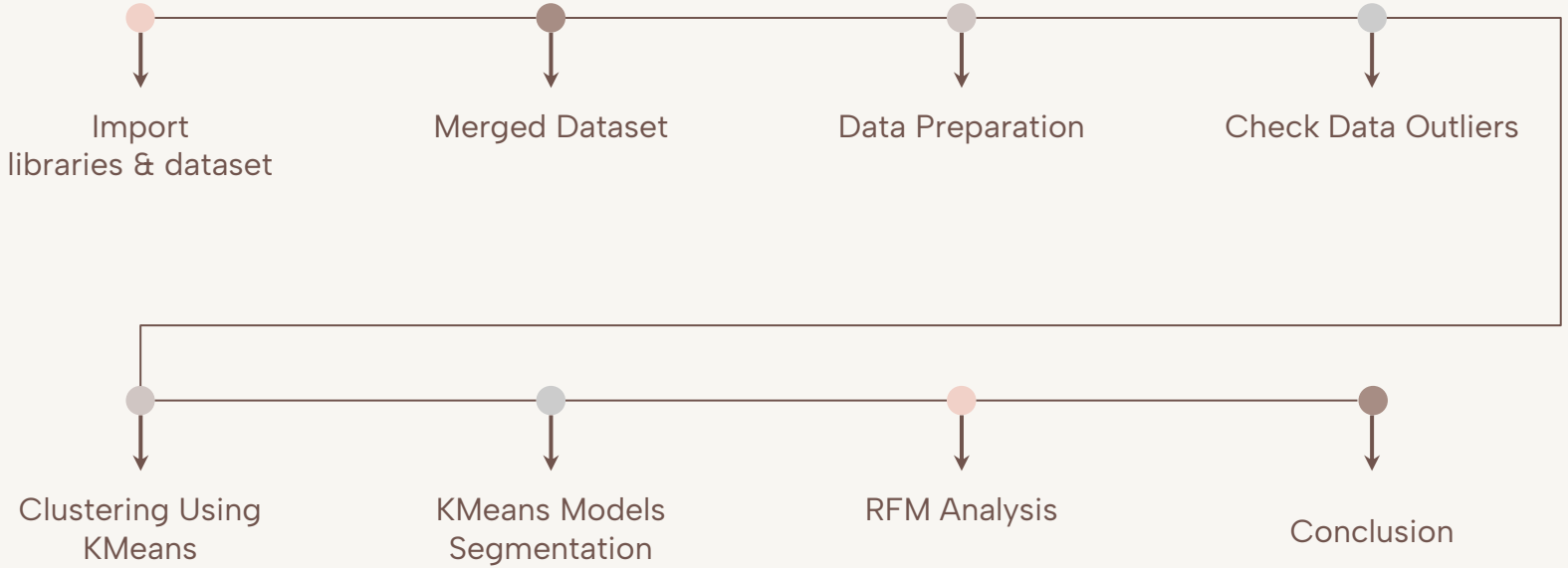


Product Name	Mean Quantity Sold Daily
Cashew	2
Cheese Stick	3
Choco Bar	6
Coffee Candy	4
Crackers	4
Ginger Candy	5
Oat	2
Potato Chip	3
Thai Tea	4
Yoghurt	4

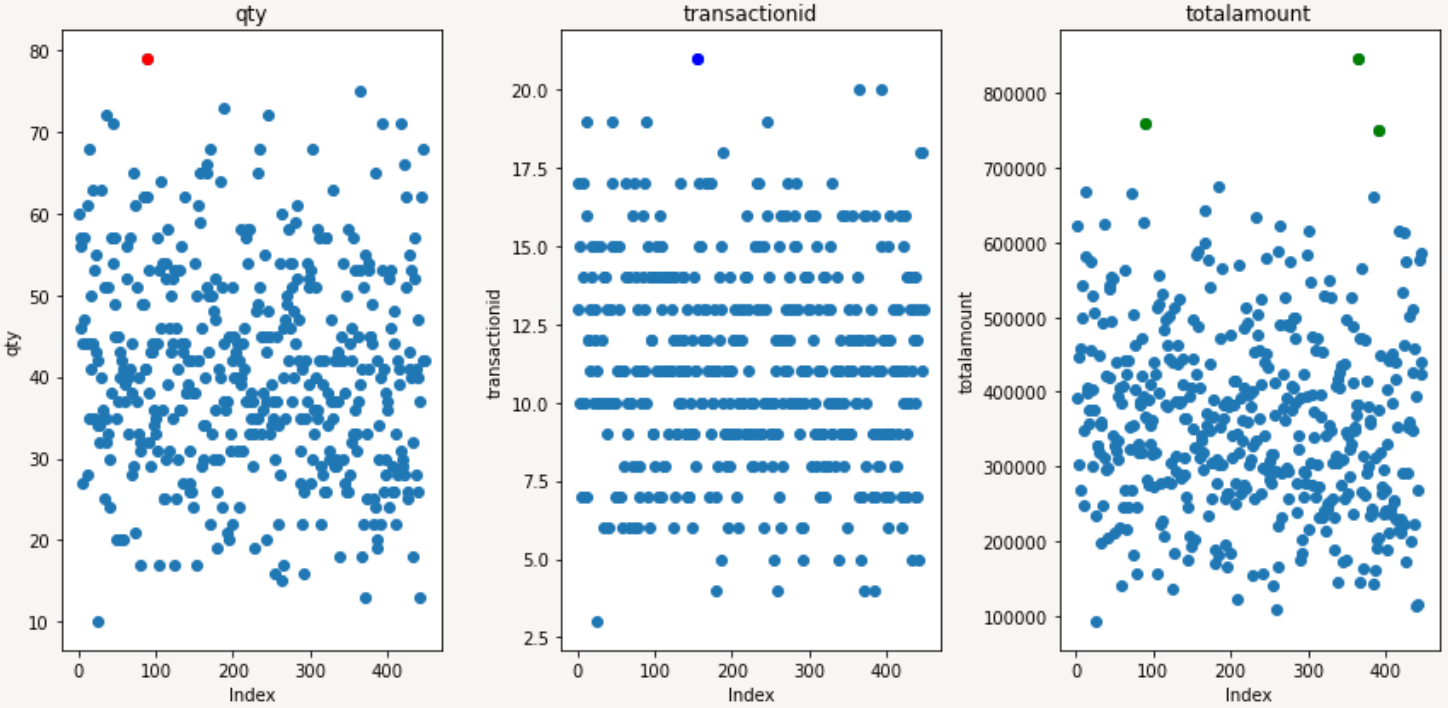
05

Model Machine Learning Clustering

Clustering Modeling Timeline



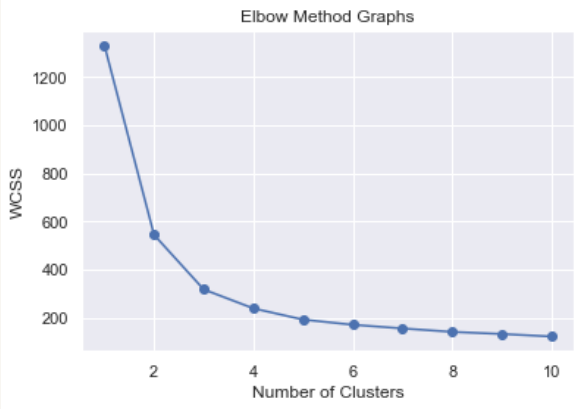
Outliers


qty
qty
Index
transactionid
transactionid
Index
totalamount
totalamount
Index

Clustering Using KMeans

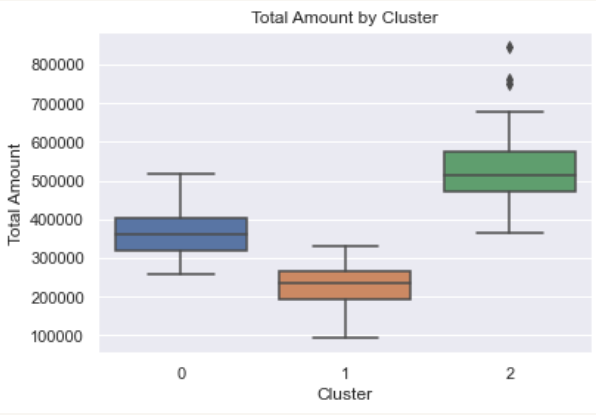
customerid	transactionid	qty	totalamount
1	17	60	623300
2	13	57	392300
3	15	56	446200
4	10	46	302500
5	7	27	268600

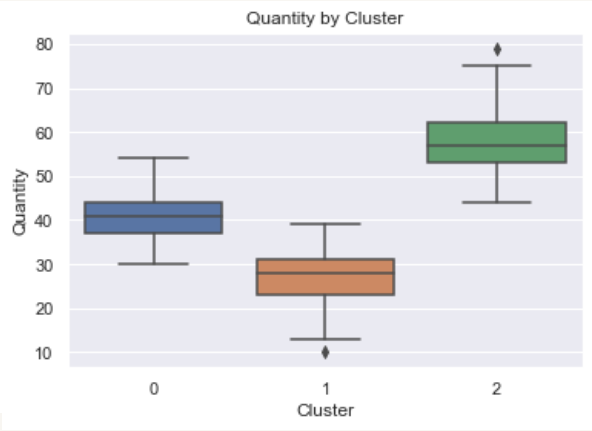
	transactionid	qty	totalamount
0	1.788282	1.508934	2.102424
1	0.553450	1.272891	0.246343
2	1.170866	1.194211	0.679428
3	-0.372675	0.407403	-0.475199
4	-1.298799	-1.087531	-0.747585

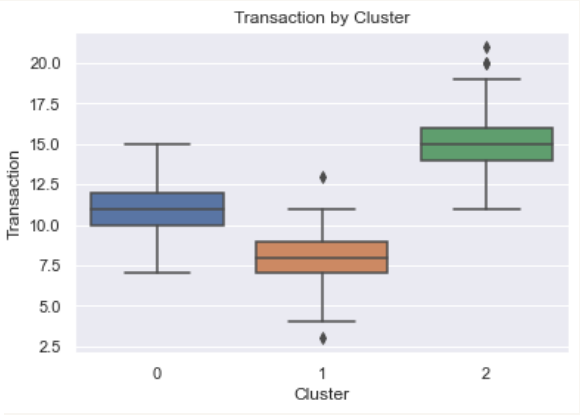


KMeans Models Segmentation

customerid	transactionid	qty	totalamount	cluster
1	17	60	623300	2
2	13	57	392300	2
3	15	56	446200	2
4	10	46	302500	0
5	7	27	268600	1

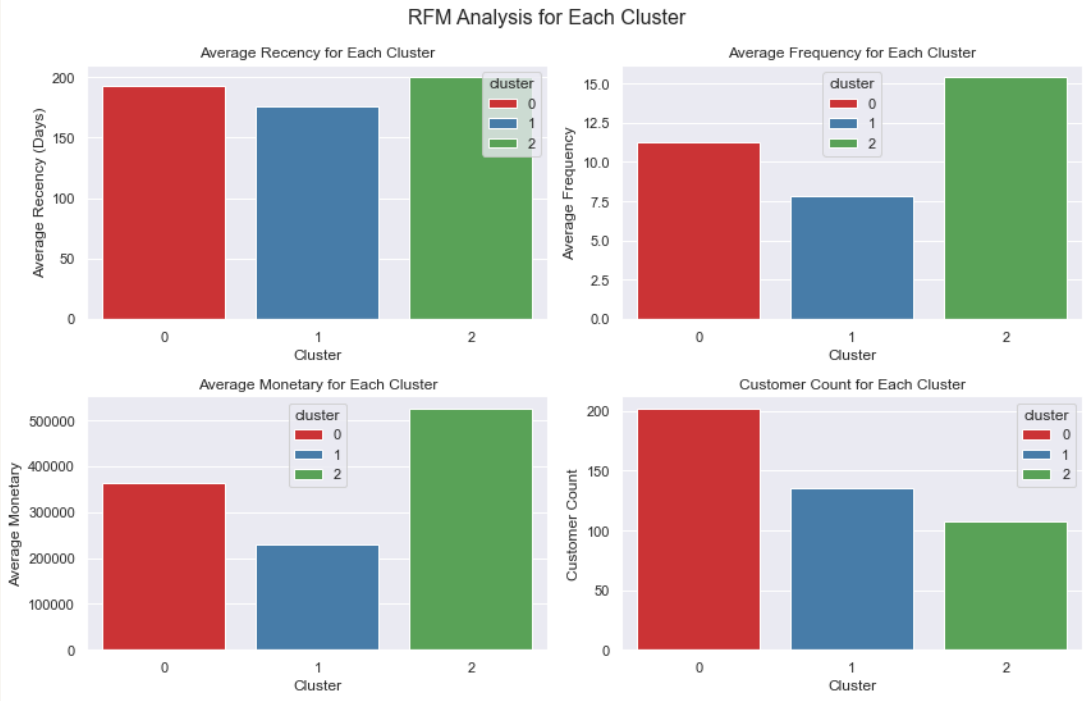

Total Amount by Cluster
Total Amount
800000
700000
600000
500000
400000
300000
200000
100000
0
1
2
Cluster


Quantity by Cluster
Quantity
80
70
60
50
40
30
20
10
0
1
2
Cluster


Transaction by Cluster
Transaction
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
1
2
Cluster

RFM Analysis

customerid	transactionid	qty	totalamount	date
1	17	60	623300	2022-02-23
2	13	57	392300	2022-01-15
3	15	56	446200	2022-01-25
4	10	46	302500	2022-02-18
5	7	27	268600	2022-02-10



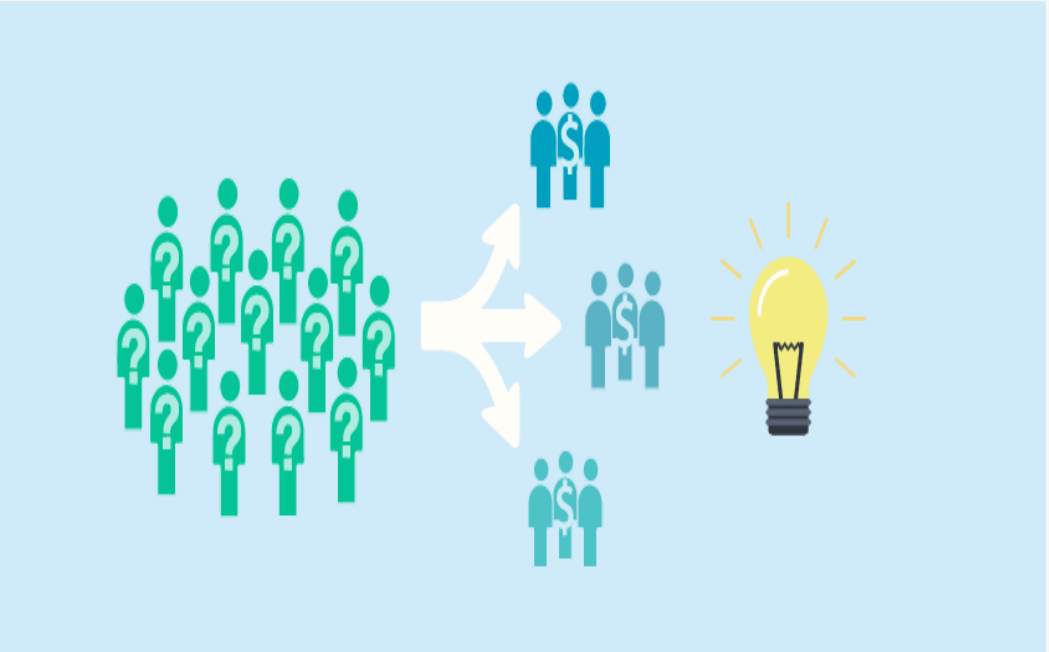
Conclusion

Dengan menggunakan Analisis RFM untuk setiap cluster, dapat disimpulkan bahwa

* Cluster 0
  * Lowest Average Recency
  * Lowest Average Frequency
  * Lowest Average Monetary
  * Medium Customer Count

* Cluster 1
  * Highest Average Recency
  * Highest Average Frequency
  * Highest Average Monetary
  * Lowest Customer Count

* Cluster 2
  * Medium Average Recency
  * Medium Average Frequency
  * Medium Average Monetary
  * Highest Customer Count



Thank You