Rakamin Academy X KALBE Nutritionals

Machine Learning Project

Data Science Project-Based Internship

Presented by
Dinda Galuh Guminta

Dinda Galuh

Statistics graduate with passion in data visualization, data analysis, and reporting. Dedicated and hard working person.

I joined the data science bootcamp because I am more interested and have a passion for learning and working in the data field. I like doing data analysis and data visualization. I want to apply my knowledge in the real world.

Rakamin Academy

2023
Rakamin Academy

2020-2022
Master of Statistics
Institut Teknologi Sepuluh Nopember

2015-2019
Bachelor of Statistics
Institut Teknologi Sepuluh Nopember

Rakamin Academy

Kalbe Case Study Outline

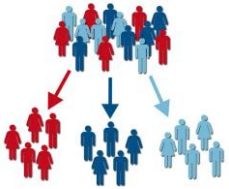
- EDA in DBeaver
- Create dashboard using tableau
- Time series analysis using Python
- Clustering using Python

Business Understanding

Problem Statement

Kalbe Nutritionals needs to optimize operational efficiency

1. **Inventory Team Needs**:
   - Predict daily sales quantity for all products.
   - Ensure there is always sufficient stock available based on predictions.
2. **Marketing Team Needs**:
   - Segment customers effectively.
   - Deliver personalized promotions and sales treatments based on segments.
3. **Role of Data Scientist**:
   - Address the above challenges using relevant data.
   - Employ analytical techniques to provide solutions.



Goals

- Estimate the quantity of product sold so that the inventory team can make sufficient daily inventory stock
- Increase the **effectiveness of marketing campaign** by targeting the right customers so that **sales increase**.

Objective

Predict the daily sales quantity for all products of Kalbe Nutritionals

Data Overview

Transaction Data

	TransactionID	CustomerID	Date	ProductID	Price	Qty	TotalAmount	StoreID
0	TR11369	328	01/01/2022	P3	7500	4	30000	12
1	TR16356	165	01/01/2022	P9	10000	7	70000	1
2	TR1984	183	01/01/2022	P1	8800	4	35200	4
3	TR35256	160	01/01/2022	P1	8800	7	61600	4
4	TR41231	386	01/01/2022	P9	10000	1	10000	4

Customer Data

	CustomerID	Age	Gender	Marital Status	Income
0	1	55	1	Married	5,12
1	2	60	1	Married	6,23
2	3	32	1	Married	9,17
3	4	31	1	Married	4,87
4	5	58	1	Married	3,57

Store Data

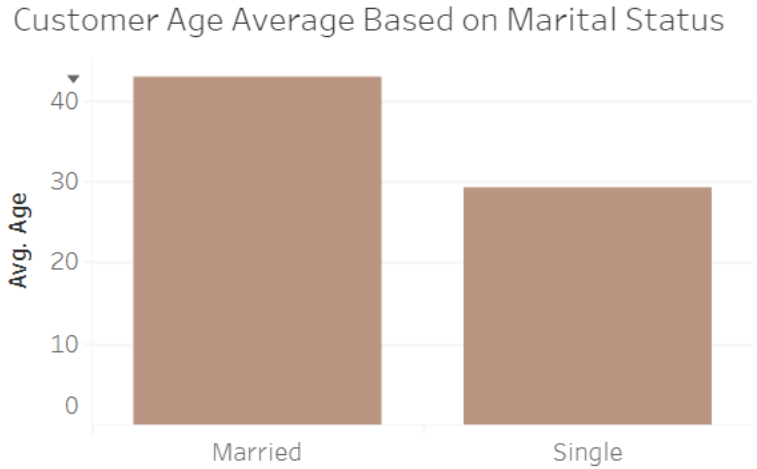
	StoreID	StoreName	GroupStore	Type	Latitude	Longitude
0	1	Prima Tendean	Prima	Modern Trade	-6,2	106,816666
1	2	Prima Kelapa Dua	Prima	Modern Trade	-6,914864	107,608238
2	3	Prima Kota	Prima	Modern Trade	-7,797068	110,370529
3	4	Gita Ginara	Gita	General Trade	-6,966667	110,416664
4	5	Bonafid	Gita	General Trade	-7,250445	112,768845

Product Data

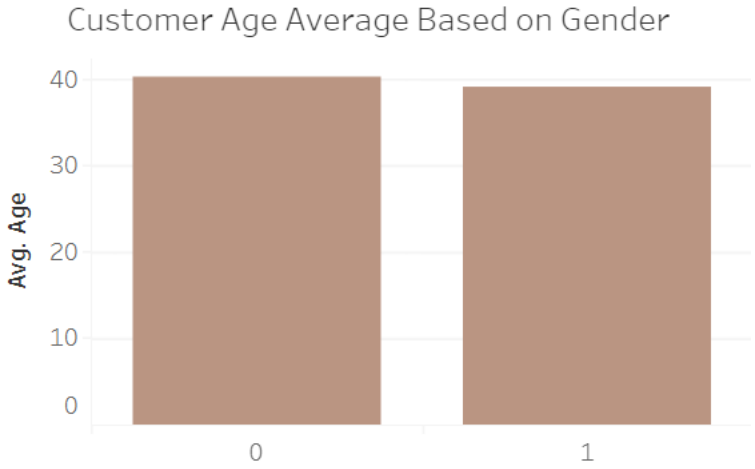
	ProductID	Product Name	Price
0	P1	Choco Bar	8800
1	P2	Ginger Candy	3200
2	P3	Crackers	7500
3	P4	Potato Chip	12000
4	P5	Thai Tea	4200

Exploratory Data Analysis in DBeaver

```
select `Marital Status`, avg(Age)
from kn_customer kc
where `Marital Status` <> ""
group by 1
```




```
select Gender, avg(Age)
from kn_customer kc
group by 1
```

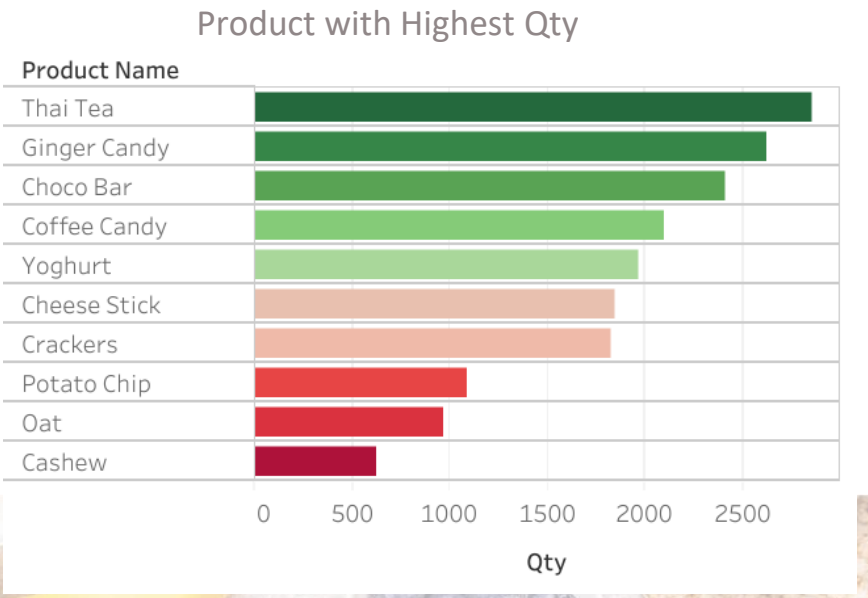


Exploratory Data Analysis in DBeaver

```
select kp.ProductID , kp.`Product Name`,  sum(kt.TotalAmount) as 'Total Amount'
from kn_product kp
left join kn_transaction kt
on kp.ProductID = kt.ProductID
group by 1, 2
order by 3 desc
```

product with the highest sales amount

Product ID	Product Name	
P10	Cheese Stick	27,615,000
P1	Choco Bar	21,190,400
P7	Coffee Candy	19,711,800
P9	Yoghurt	19,630,000
P8	Oat	15,440,000
P3	Crackers	13,680,000
P4	Potato Chip	13,104,000
P5	Thai Tea	11,982,600
P6	Cashew	11,286,000
P2	Ginger Candy	8,403,200

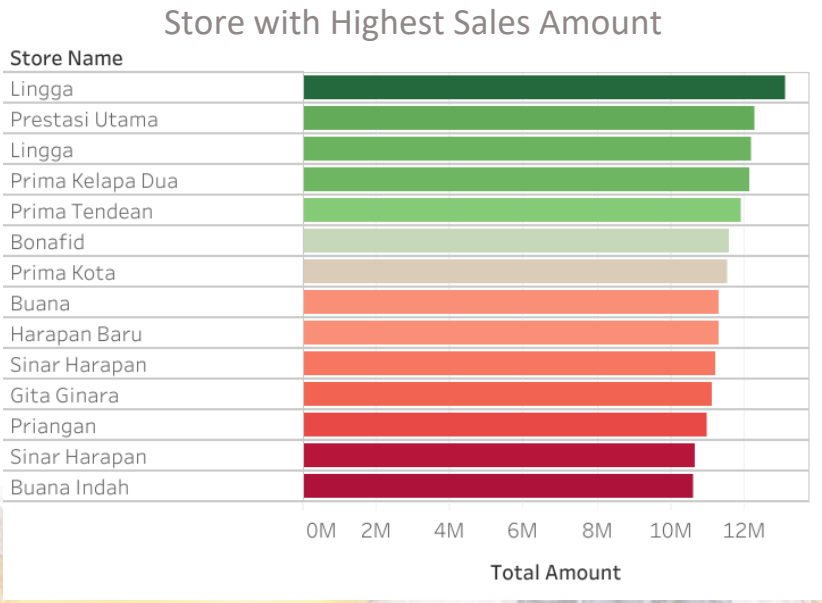


Exploratory Data Analysis in Dbeaver & Tableau

```sql
select ks.StoreID,ks.StoreName, sum(kt.Qty) as 'Total Transaksi'
from kn_store ks
left join kn_transaction kt
on ks.StoreID =kt.StoreID
group by 1, 2
order by 3 desc
```

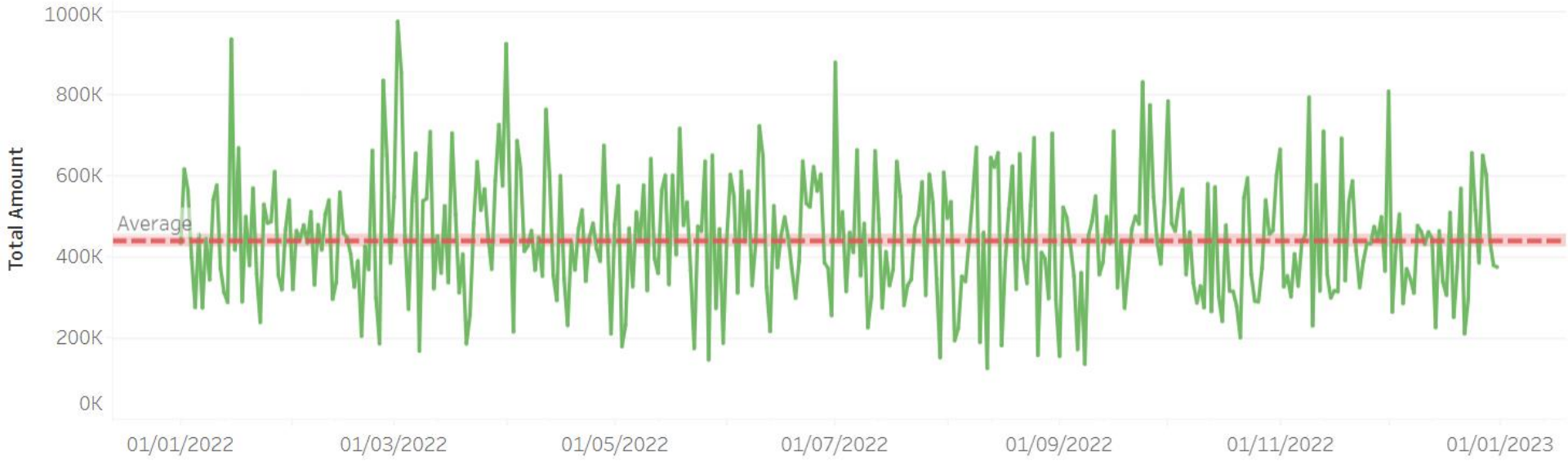
store with the highest total quantity

Store ID	Store Name	
9	Lingga	1,439
12	Prestasi Utama	1,395
3	Prima Kota	1,358
6	Lingga	1,338
11	Sinar Harapan	1,331
13	Buana	1,320
1	Prima Tendean	1,310
2	Prima Kelapa Dua	1,296
10	Harapan Baru	1,286
5	Bonafid	1,283
8	Sinar Harapan	1,257
14	Priangan	1,239
4	Gita Ginara	1,236
7	Buana Indah	1,208

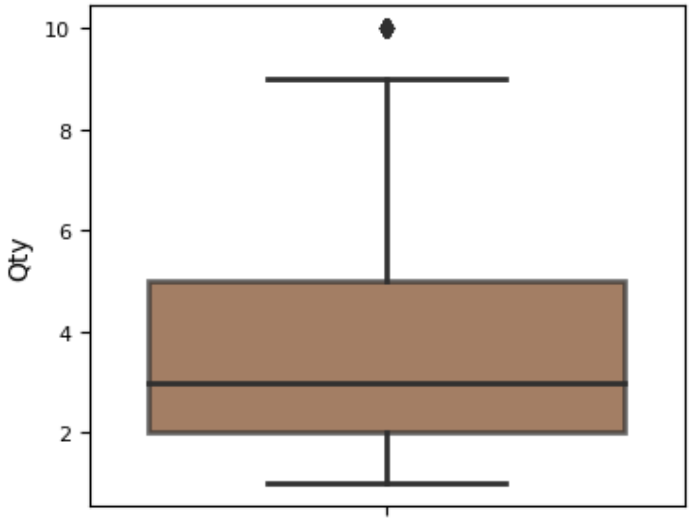


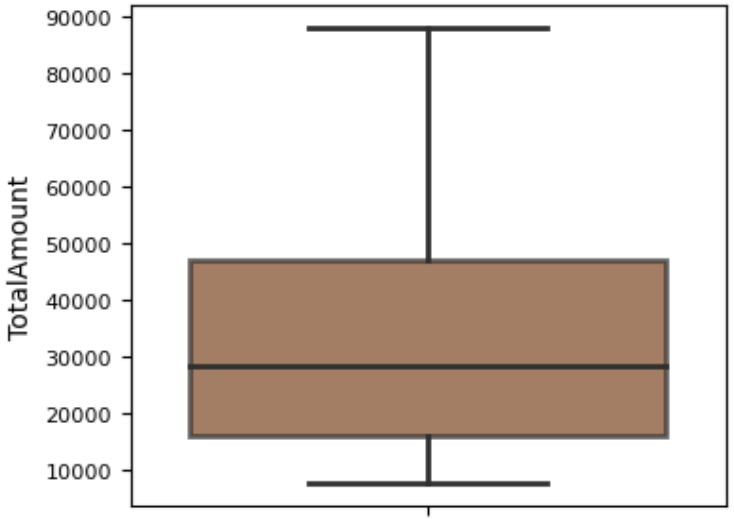
Monthly Total Quantity

Daily Sales Amount



Data Characteristics





Data Preprocessing

Merge all data frame → Check missing value → Check duplicated row → Change date data type → aggregate daily transaction data for time series / perform aggregation of transaction in each customer for clustering

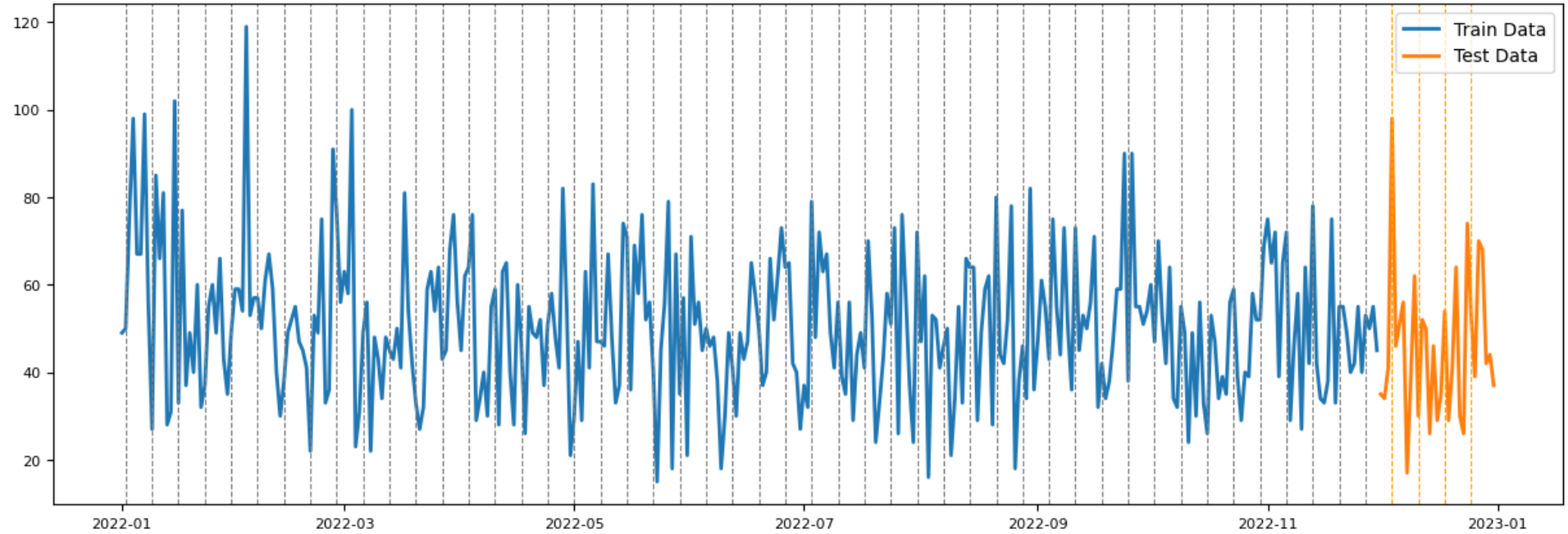
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5020 entries, 0 to 5019
Data columns (total 18 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   TransactionID   5020 non-null   object
 1   CustomerID      5020 non-null   int64
 2   Date            5020 non-null   object
 3   ProductID       5020 non-null   object
 4   Price           5020 non-null   int64
 5   Qty             5020 non-null   int64
 6   TotalAmount     5020 non-null   int64
 7   StoreID         5020 non-null   int64
 8   Age             5020 non-null   int64
 9   Gender          5020 non-null   int64
 10  Marital Status  4976 non-null   object
 11  Income          5020 non-null   object
 12  Product Name    5020 non-null   object
 13  StoreName       5020 non-null   object
 14  GroupStore      5020 non-null   object
 15  Type            5020 non-null   object
 16  Latitude        5020 non-null   object
 17  Longitude       5020 non-null   object
```

- 5020 row (transaction)
- 18 column (feature)
- There are 0,87% missing value in marital status
- Change data type in **Date**, **Income**

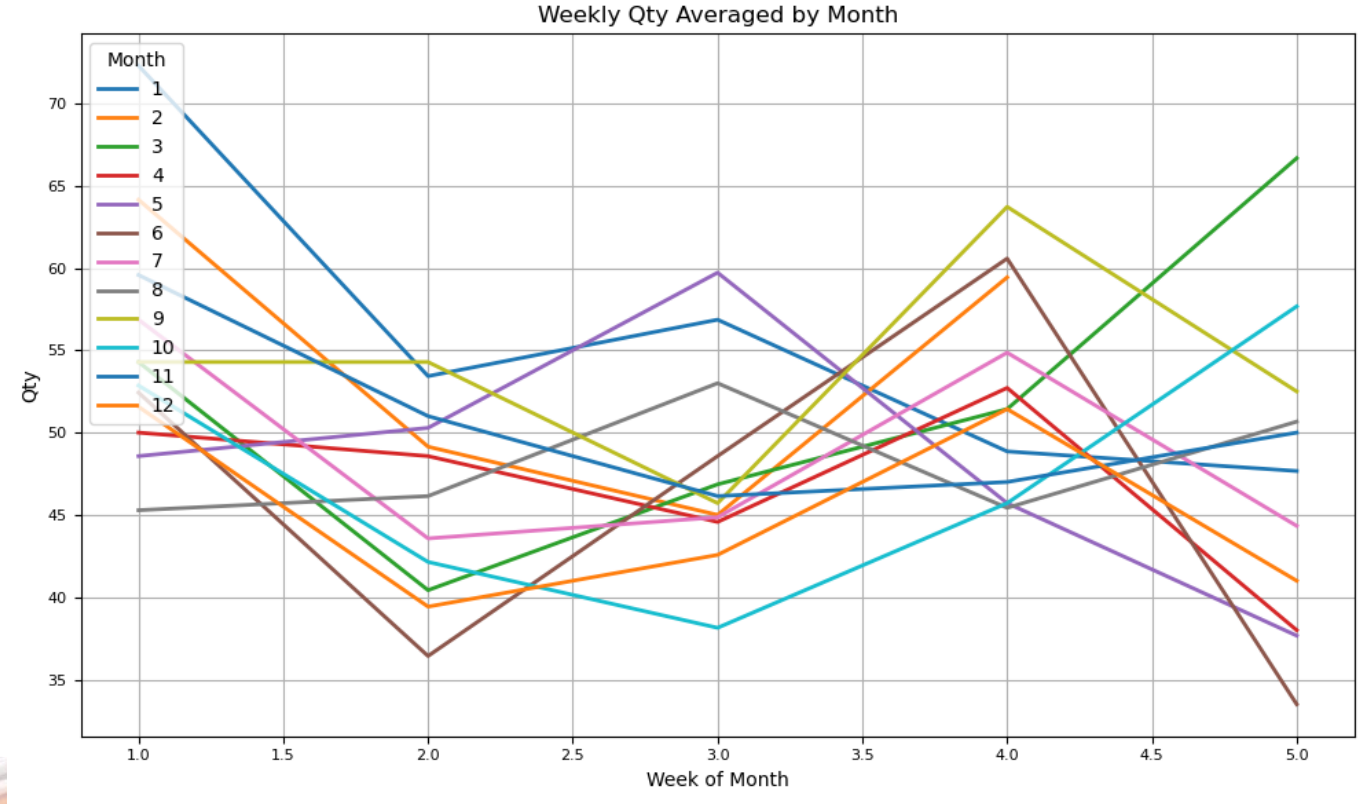
ARIMA Time Series

Split data into train and test

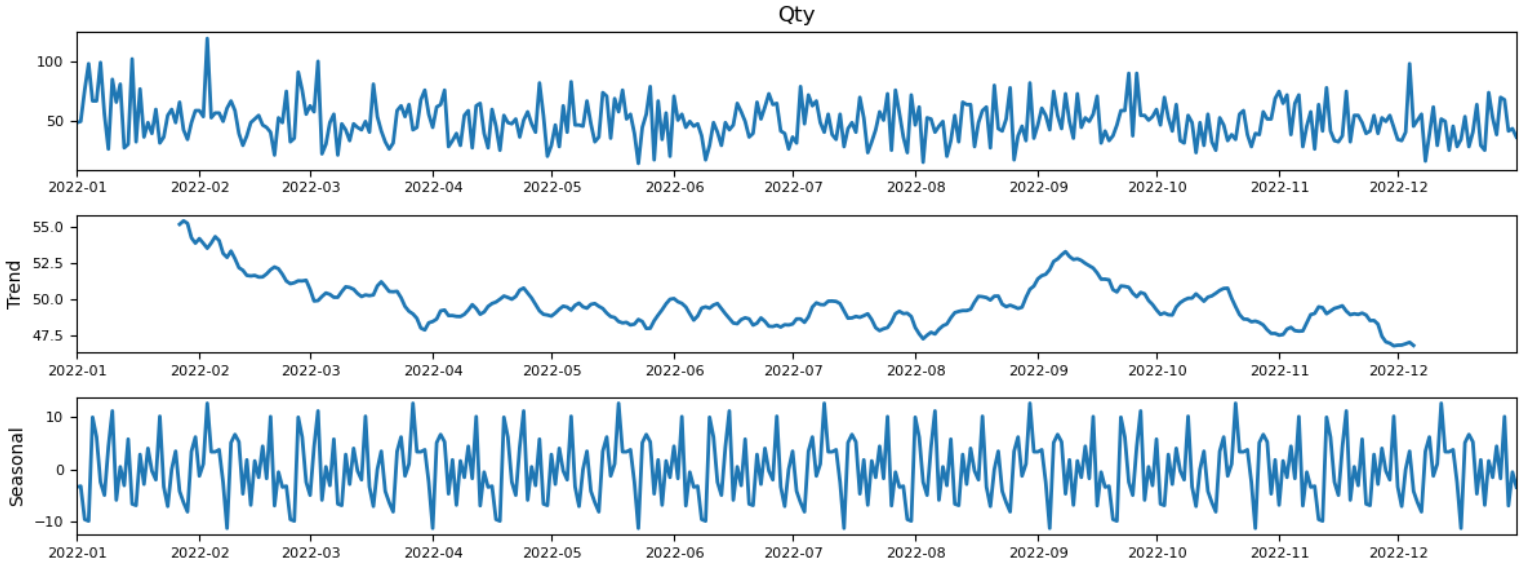
- Train (in sample) : 1 Jan 2022 – 30 Nov 2022
- Test (out sample) : 1 Dec 2022 – 31 Dec 2022



ARIMA Time Series



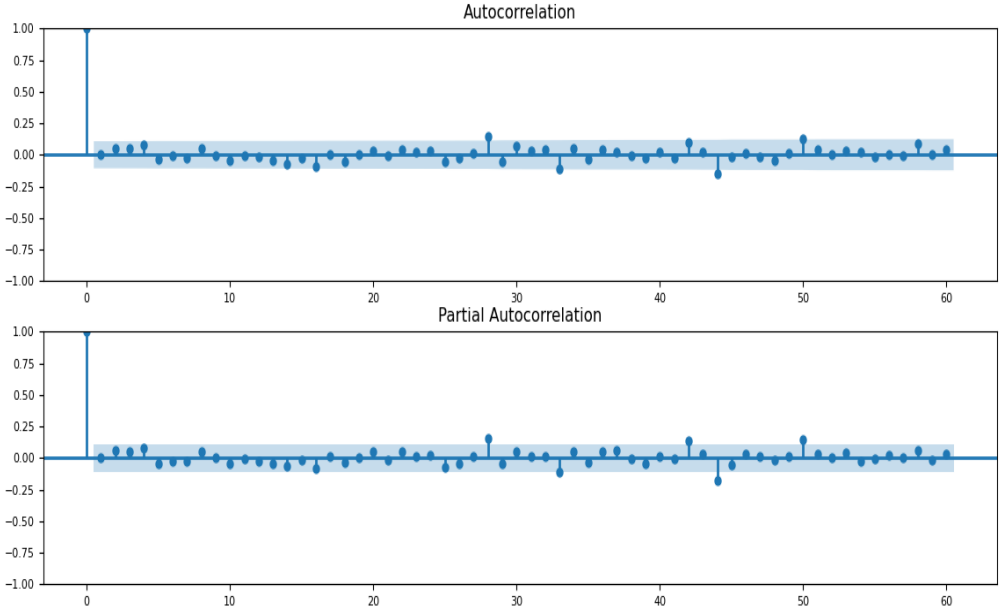
Regression – ARIMA Time Series



Based on the visualizations, it's evident that our data exhibits seasonality and trend. Therefore, we'll employ prediction models that account for both these characteristics.

Check Stationarity

```
ADF Statistic: -18.175873
p-value: 0.000000
Critical Test Statistics Values:
        1%: -3.450
        5%: -2.870
        10%: -2.571
```



Given the data, and considering that we want to set the trend and seasonality differentials manually:

- d=1 for differencing once for the trend.
- D=1 for differencing once for seasonality.
- m=7 / 12 / 52 since there's a weekly seasonality for daily data.
- trend='c' to include a constant.
- seasonal=True to fit a seasonal ARIMA.

Model Performance

model	MSE	RMSE	MAPE
ARIMA(1,1,0)(2,1,0)[52]	293,7716	17,1398	33,6774
ARIMA(0,1,1)(2,1,0)[52]	306,1370	17,4968	38,7011
ARIMA(5,1,0)(2,1,1)[7]	308,0584	17,5516	37,7732
ARIMA(5,1,0)(2,1,0)[12]	385,7396	19,6403	39,4329

ARIMA(1,1,0)(2,1,0)[52] has the lowest MSE, RMSE, and MAPE values.

ARIMA

Results of SARIMAX on train

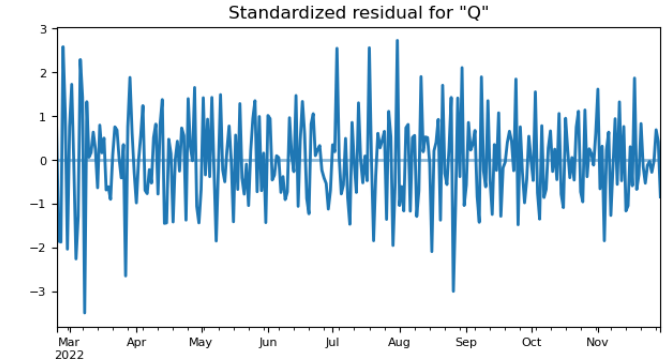
SARIMAX Results

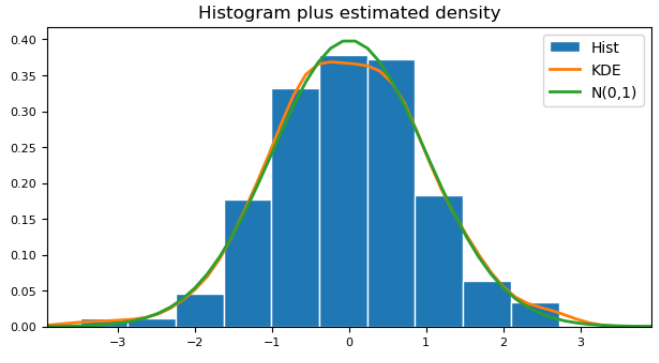
Dep. Variable:	Qty	No. Observations:	334
Model:	SARIMAX(1, 1, 0)x(2, 1, 0, 52)	Log Likelihood	-1304.933
Date:	Sat, 02 Sep 2023	AIC	2617.865
Time:	12:04:44	BIC	2632.419
Sample:	01-01-2022	HQIC	2623.702
	- 11-30-2022		
Covariance Type:	opg		

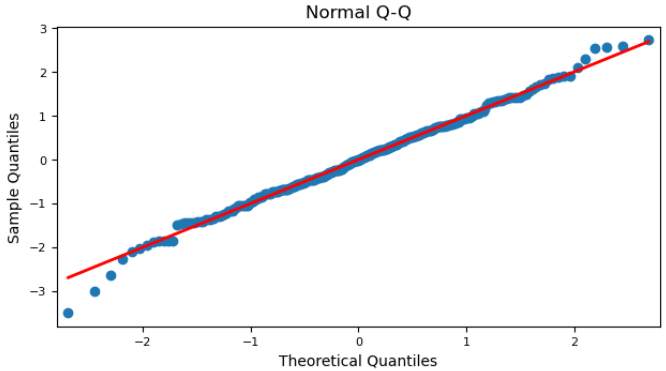
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5334	0.053	-10.018	0.000	-0.638	-0.429
ar.S.L52	-0.6747	0.064	-10.492	0.000	-0.801	-0.549
ar.S.L104	-0.3336	0.072	-4.605	0.000	-0.476	-0.192
sigma2	572.8368	47.962	11.944	0.000	478.833	666.840

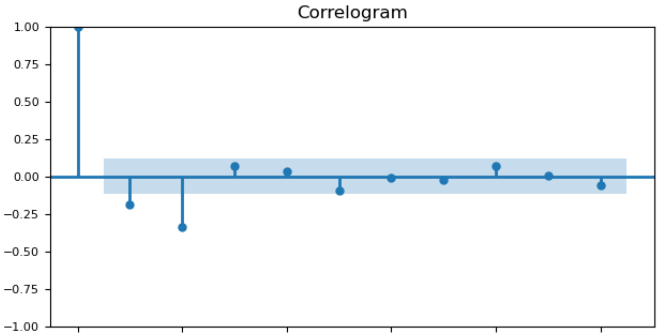
Ljung-Box (L1) (Q):	9.80	Jarque-Bera (JB):	1.40
Prob(Q):	0.00	Prob(JB):	0.50
Heteroskedasticity (H):	0.59	Skew:	-0.06
Prob(H) (two-sided):	0.01	Kurtosis:	3.32

Check Residual

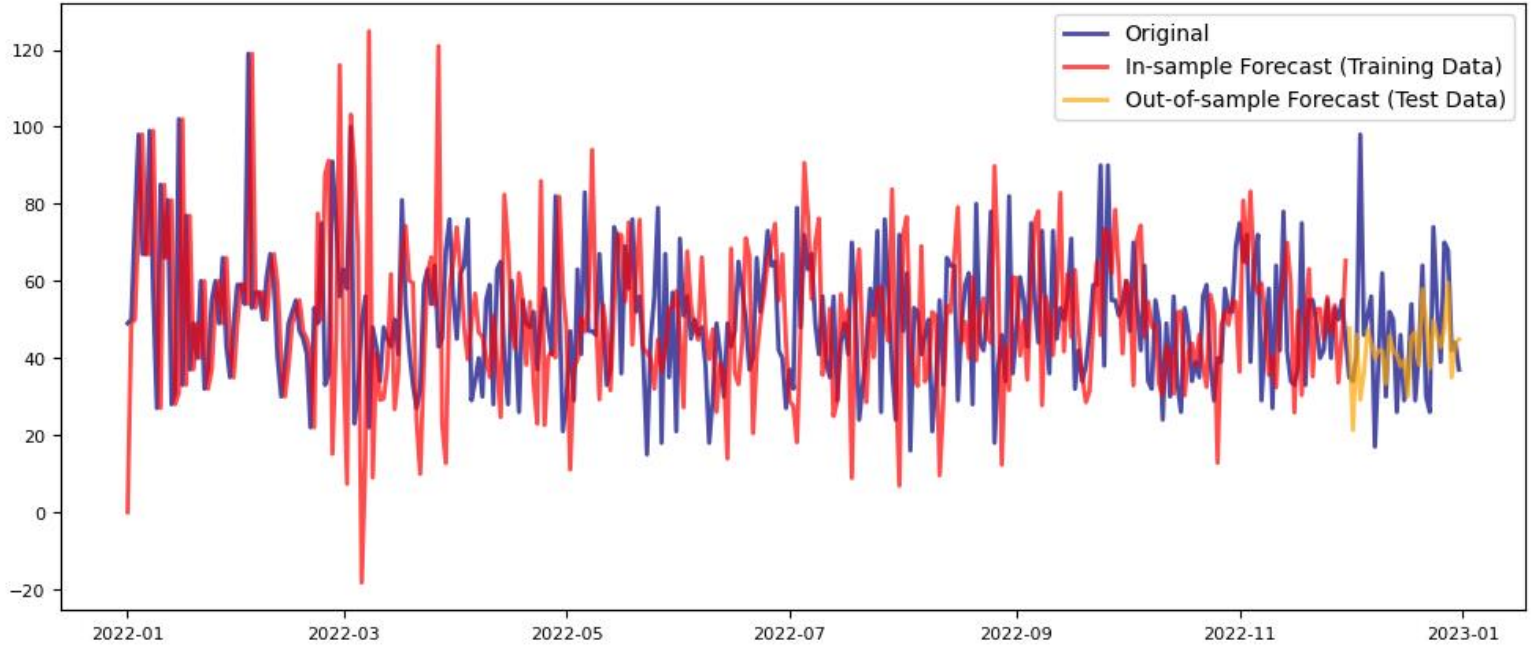




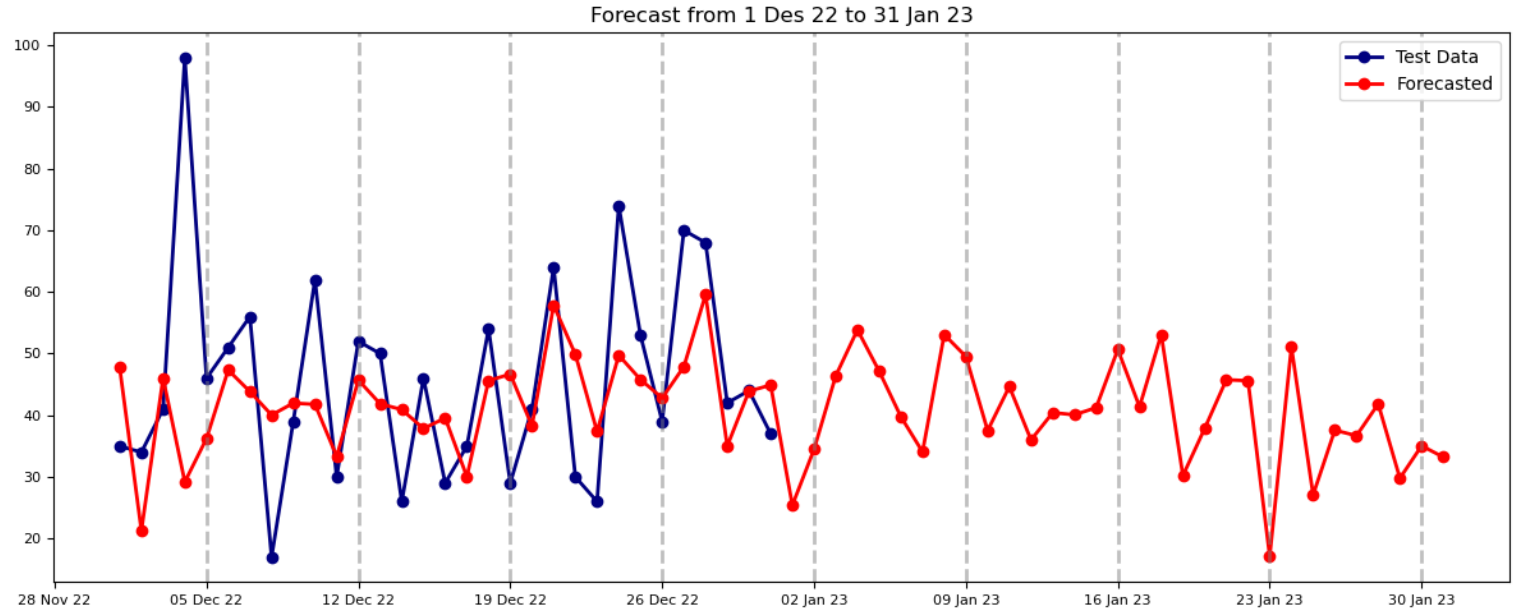




Model Validation



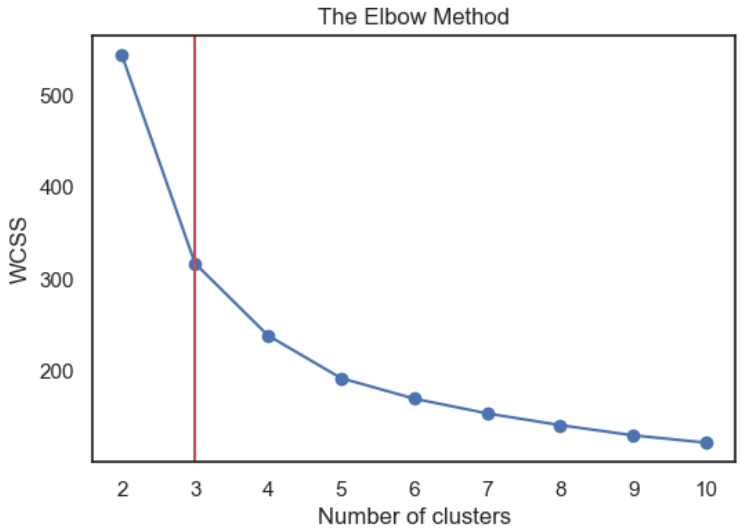
Forecast 1 Month Ahead (1 Jan 23 – 31 Jan 23)



The forecasted data suggests a pattern of high demand at the start and mid-month. The forecast provides valuable insights for the inventory team to ensure that the stock is adequately maintained during peak demand days, ensuring continuous availability of snack products for consumers.

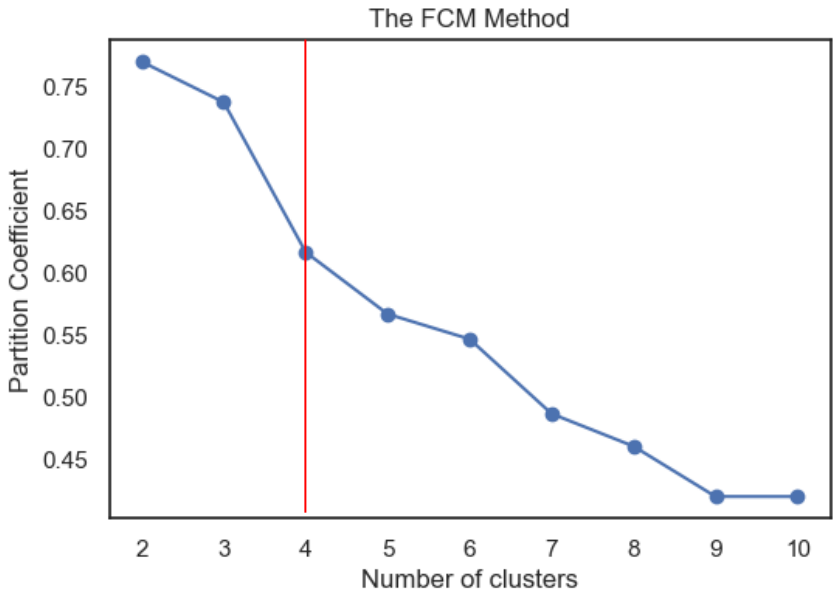
Clustering – K Means

Number of Cluster	silhouette	CH index	Davies-Bouldin	ICD Rate
3	0,1318	717,2068	0,7495	0,2364
4	0,0936	682,4251	0,8300	0,1779
5	0,0626	660,6022	0,9202	0,1433
6	0,0524	608,0578	1,0512	0,1267
7	0,0245	565,8685	1,0849	0,1147
8	0,0403	533,5413	1,0976	0,1052
9	0,0226	510,0196	1,1213	0,0969
10	0,0390	485,3105	1,0534	0,0910



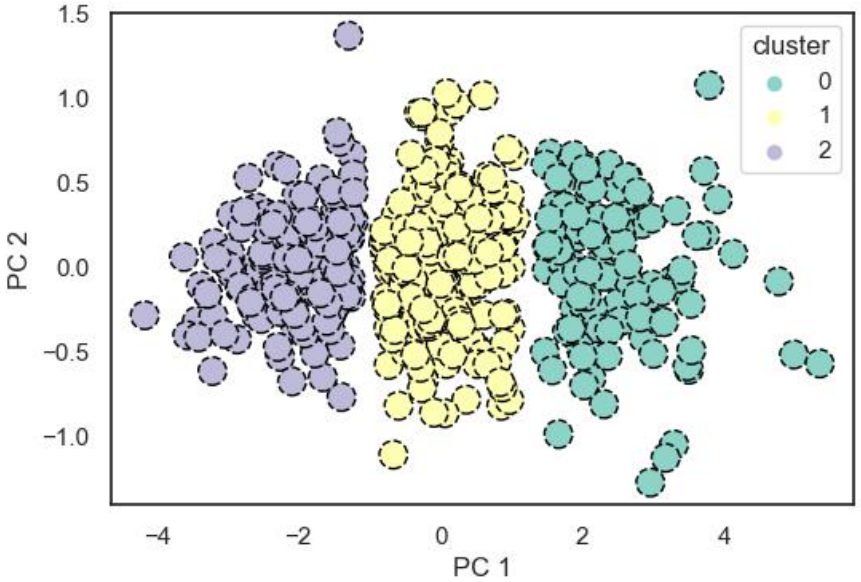
Clustering – Fuzzy C-Means

Number of Cluster	silhouette	CH index	Davies-Bouldin	ICD Rate
3	0,1503	887,8202	0,6691	0,2000
4	0,1176	703,2862	0,9502	0,1735
5	0,1255	670,2717	1,0096	0,1415
6	0,1611	723,2879	0,9942	0,1087
7	0,1387	642,4926	1,1509	0,1024
8	0,1290	596,8273	1,0959	0,0951
9	0,1335	555,6635	1,1288	0,0912
10	0,1303	547,1646	1,1738	0,0815

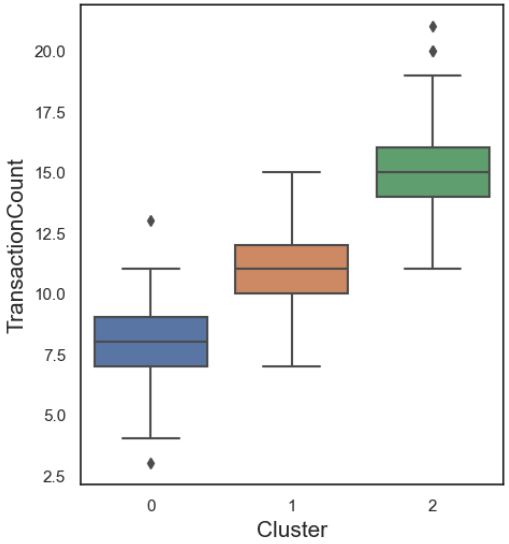


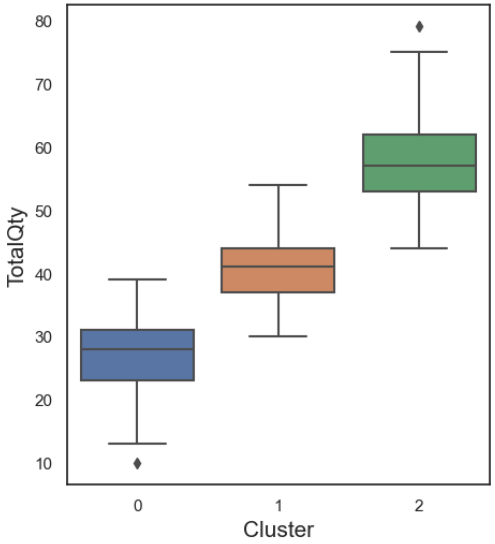
Clustering Method Evaluation

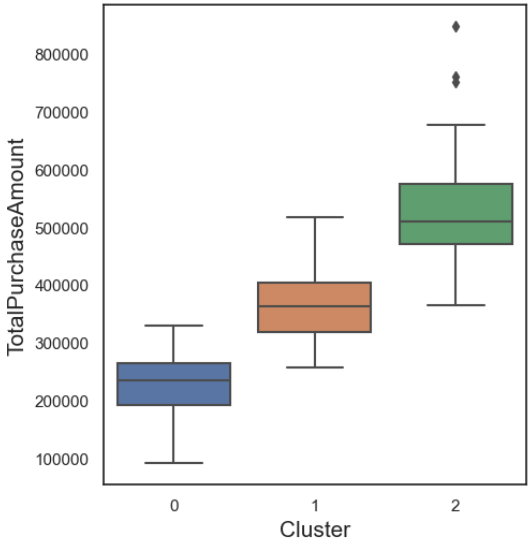
Method	ICD Rate
K-Means	0,2364
Fuzzy C-Means	0,2000



Clustering Result

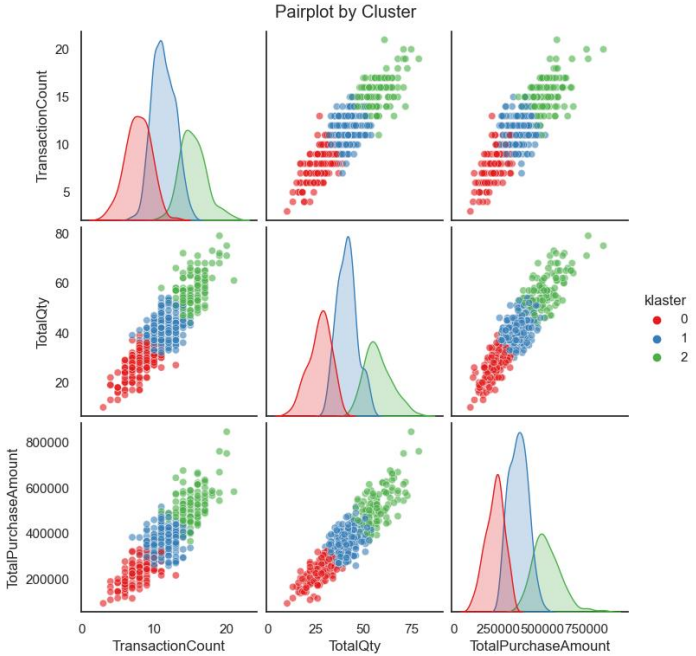






Cluster 0: **Casual Buyers**
Cluster 1: **Moderate Shoppers / Mid Value**
Cluster 2: **Loyal High-Spenders / High Value**

Cluster Characteristic



- **Cluster 0** customers who seldom shop or perhaps only buy when there's a specific need or promotion. They might be more budget-conscious customers or those who purchase based on necessities.
- **Cluster 1** customers with a moderate shopping frequency and total expenditure. They might be purchasing for personal consumption or a small family.
- **Cluster 2** "loyal" customers who frequently purchase and spend more money on snacks. They might be regular patrons or those buying for business purposes or large families.

Casual Buyers:

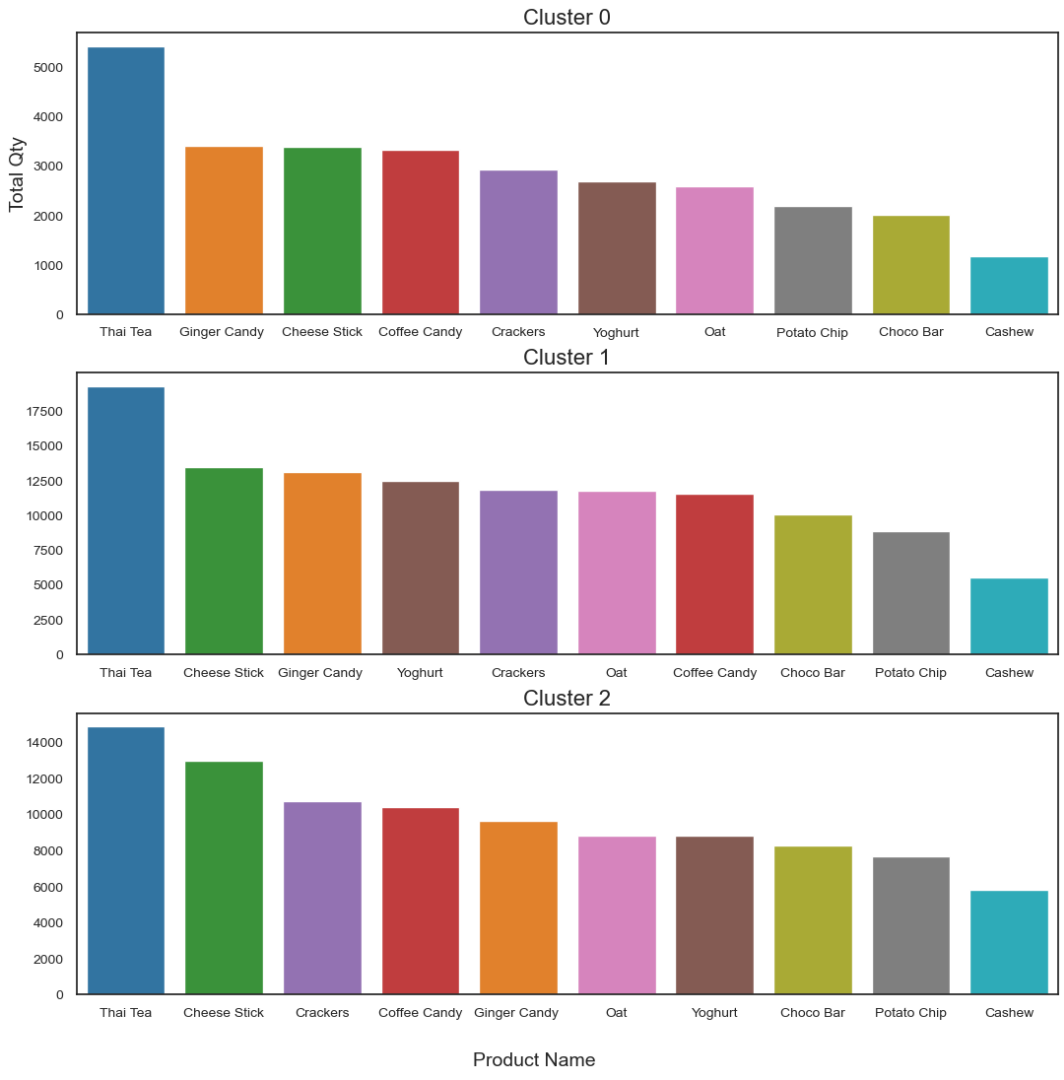
- Flash Sales
- Discount Coupons
- Product Recommendations

Moderate Shoppers:

- Bundling Deals
- Seasonal Promotions
- Newsletter

Loyal High-Spenders

- Special Offers
- Loyalty Programs
- Feedback



Top Product in Each Cluster

Recommendation for Increasing Sales:

1. Promote Underperforming Products
2. Bundle Products
3. Loyalty Programs
4. Feedback & Improvements

Conclusion & Recommendation

Leveraging the January Forecast:

- The forecast indicates certain peak sales days in January. Planning inventory based on the buying patterns of the three clusters can ensure optimal stock levels.
- The peaks in January can be used to introduce new products or flavors, especially targeting the High-Value and Mid-Value clusters. Exclusive pre-launch access or discounts can encourage purchases.
- Utilize the insights from clustering to create targeted advertising campaigns. For example, on predicted peak sales days, target Cluster 2 with premium products, Cluster 1 with bundle offers, and Cluster 0 with discounted items.
- The forecast can help in optimizing staffing in retail outlets, ensuring high-service levels during peak days, especially for Cluster 2 customers.

Conclusion & Recommendation

1. **Loyal High-Spenders / High Value**
   - Prioritize stock availability for products popular (thai tea, cheese stick, crackers) within this segment,
   - Special loyalty programs or early-access sales can be considered for this segment to further boost their purchase frequency in January.
2. **Moderate Shoppers / Mid Value**
   - Since we anticipate a rise in sales in January 2023, it's essential to engage this group with targeted marketing campaigns, offering bundle deals or limited-time discounts. This can potentially shift their buying behavior closer to Cluster 2's pattern.
3. **Casual Buyers**
   - Given the forecasted sales for January, consider introducing special promotional offers to attract this segment. Since post-holiday sales often attract bargain hunters, this cluster might be more active in looking for deals. Make sure to have attractive entry-level offers or deals to draw them in.

Tools Used

Acknowledgments

- https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b
- https://ejurnal.its.ac.id/index.php/sains_seni/article/download/58349/6451
- https://medium.com/data-science-business/time-series-arima-sarima-10f5b6a528d5

Dbeaver Output

	Marital Status	avg(Age)
1	Married	43.0382
2	Single	29.3846

	Gender	avg(Age)
1	1	39.1415
2	0	40.3264

	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
6	P3	Crackers	13,680,000
7	P4	Potato Chip	13,104,000
8	P5	Thai Tea	11,982,600
9	P6	Cashew	11,286,000
10	P2	Ginger Candy	8,403,200

	StoreID	StoreName	Total Transaksi
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
6	13	Buana	1,320
7	1	Prima Tendean	1,310
8	2	Prima Kelapa Dua	1,296
9	10	Harapan Baru	1,286
10	5	Bonafid	1,283
11	8	Sinar Harapan	1,257
12	14	Priangan	1,239
13	4	Gita Ginara	1,236
14	7	Buana Indah	1,208

Rakamin Academy

See Github Here

See Tableau Dashboard Here

See Video Presentation Here


Thank You
Rakamin Academy
X
KALBE Nutritionals