Rakamin Academy X KALBE Nutritionals

PROJECT-BASED INTERN :

DATA SCIENTIST KALBE NUTRITIONALS X RAKAMIN ACADEMY

MACHINE LEARNING PROJECT

By :
Moh. Harwin Prayoga

Moh. Harwin Prayoga

About Me

Yoga is an Engineering Physics graduate at Institut Teknologi Sepuluh Nopember that passionate about Data Science and Data Analysis. He joined a data science bootcamp to forge his programming skill in Python and RDBMS querying with SQL and gratefully his final project team got awarded as The Winner of Best Final Project Team. With Data Science skills and experience, he believes that he is able to deliver strategic insights and recommendations through data to achieve company goals.

Check my details

LinkedIn | Github | Medium

OUTLINE

- EXPLORATORY DATA ANALYSIS BY POSTGRESQL WITH DBEAVER
- DASHBOARD VISUALIZATION WITH TABLEAU
- MACHINE LEARNING REGRESSION (TIME SERIES)
- MACHINE LEARNING CLUSTERING

BUSINESS UNDERSTANDING

PROBLEM STATEMENT

As Data Scientist support stakeholders optimize operational efficiency based on each division below:

1. Inventory Team

To find out the estimated quantity of products sold so that the inventory team can create sufficient daily.

2. Marketing Team
    a. Segment customers effectively.
    b. Deliver proper personalized promotions and sales treatments based on segments

BUSINESS UNDERSTANDING

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
- Create customer segment clustering

DATA OVERVIEW

Rakamin Academy

KALBE Nutritionals

Transaction.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5020 entries, 0 to 5019
Data columns (total 8 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
dtypes: int64(5), object(3)
memory usage: 313.9+ KB
```

Customer.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 447 entries, 0 to 446
Data columns (total 5 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   CustomerID      447 non-null    int64
 1   Age             447 non-null    int64
 2   Gender          447 non-null    int64
 3   Marital Status  444 non-null    object
 4   Income          447 non-null    object
dtypes: int64(3), object(2)
memory usage: 17.6+ KB
```

Store.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14 entries, 0 to 13
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   StoreID     14 non-null     int64
 1   StoreName   14 non-null     object
 2   GroupStore  14 non-null     object
 3   Type        14 non-null     object
 4   Latitude    14 non-null     object
 5   Longitude   14 non-null     object
dtypes: int64(1), object(5)
memory usage: 800.0+ bytes
```

Product.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 3 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   ProductID     10 non-null     object
 1   Product Name  10 non-null     object
 2   Price         10 non-null     int64
dtypes: int64(1), object(2)
memory usage: 368.0+ bytes
```

EXPLORATORY DATA ANALYSIS BY DBEAVER(POSTGRESQL)

Rakamin Academy

KALBE Nutritionals

Average Age by Marital Status

marital_status	avg_umur
	31.3333333333
Married	43.0382352941
Single	29.3846153846

Average Age by Gender

gender	avg_umur
0	40.33
1	39.14

Quantity by Store

storename	quantity
Lingga	2,777
Sinar Harapan	2,588
Prestasi Utama	1,395
Prima Kota	1,358
Buana	1,320
Prima Tendean	1,310
Prima Kelapa Dua	1,296
Harapan Baru	1,286
Bonafid	1,283
Priangan	1,239
Gita Ginara	1,236
Buana Indah	1,208

Total Amount by Product

product_name	total_amount
Cheese Stick	27,615,000
Choco Bar	21,190,400
Coffee Candy	19,711,800
Yoghurt	19,630,000
Oat	15,440,000
Crackers	13,680,000
Potato Chip	13,104,000
Thai Tea	11,982,600
Cashew	11,286,000
Ginger Candy	8,403,200

DATA PREPROCESSING

Rakamin Academy

KALBE Nutritionals

Handling Duplicated Values

- Found duplicated values in **TransactionID**

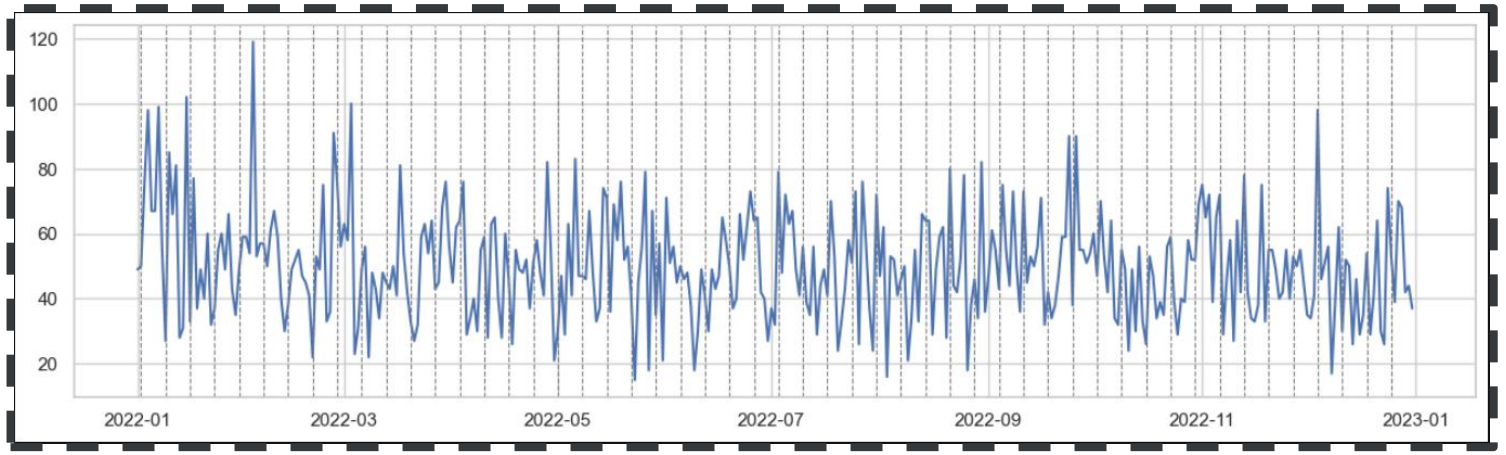
Handling Invalid Values

- Replacing "," with "."
- Changing certain columns to datetime and float

Handling Null Values

- Imputing Null Values with KNNImputer

DATA PREPROCESSING FOR TIME SERIES



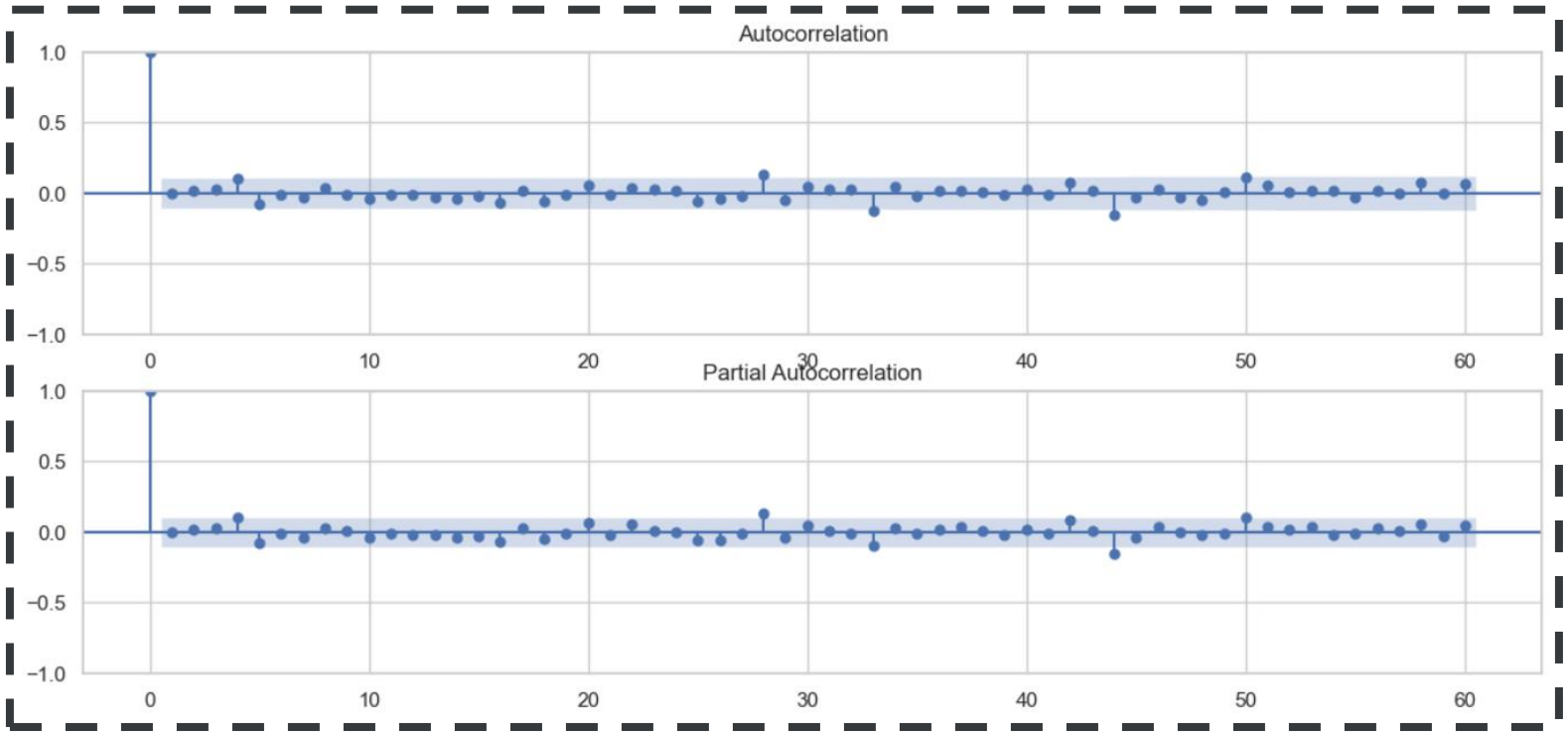
Data Transformation

- Transforming **qty** feature with Log Transformation due Right-Skewed distribution.

Data Splitting

Using last latest month as test set, and the rest are train set

CHECKING DATA STATIONARY



Augmented Dickey-Fuller test

```
1  from statsmodels.tsa.stattools import adfuller
2
3  #ADF test
4  adf_test = adfuller(df_train)
5  print(f'p-value: {adf_test[1]}')
```

```
p-value: 2.4374177848749832e-30
```

The ACF/PACF and ADF plots show the data is stationary and **can be used for the ARIMA model.**

ARIMA MODELING

Rakamin Academy

KALBE Nutritionals

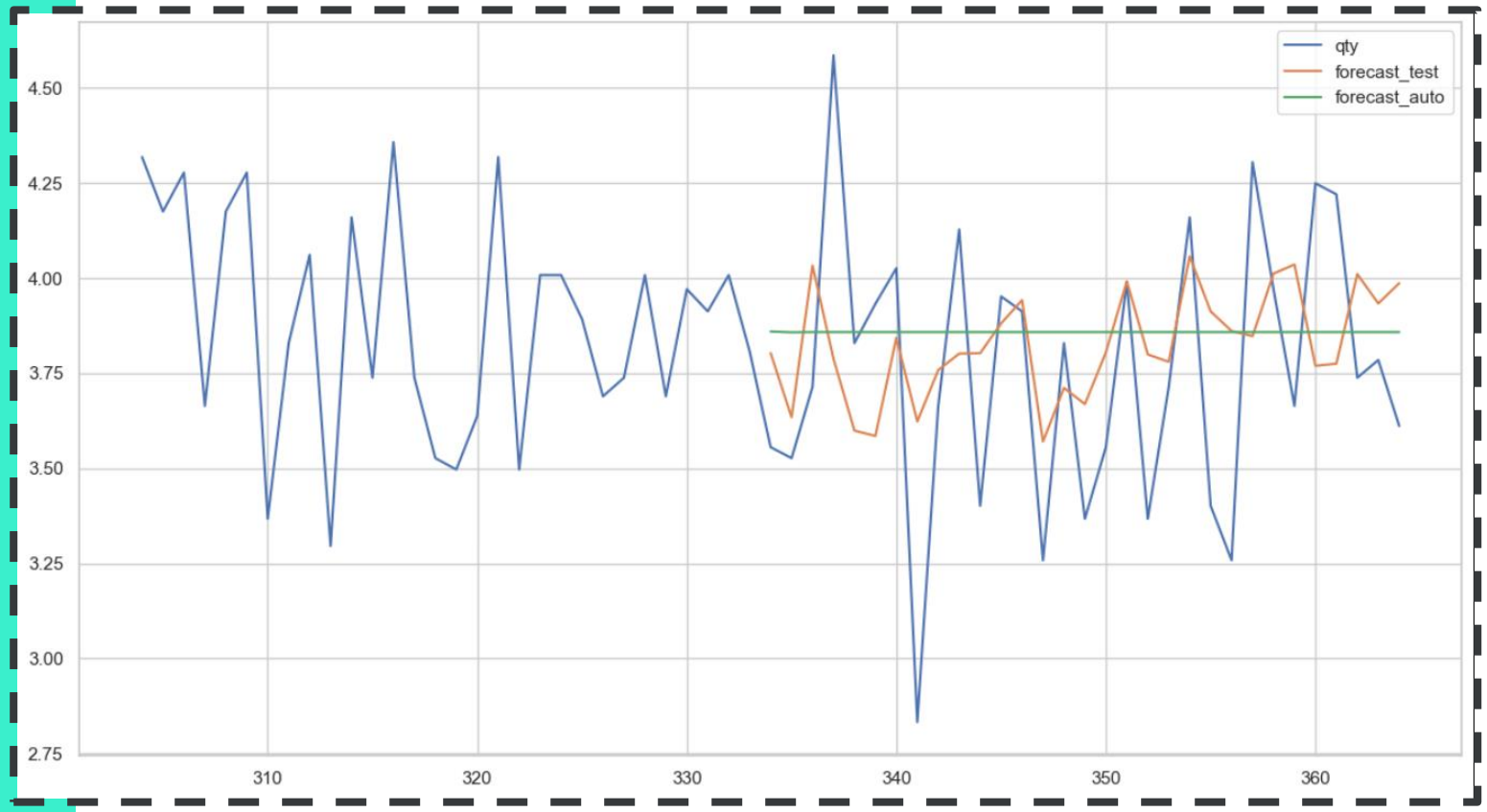
Auto

```
mae - auto: 0.3019
mape - auto: 0.0842
rmse - auto: 0.3771
```

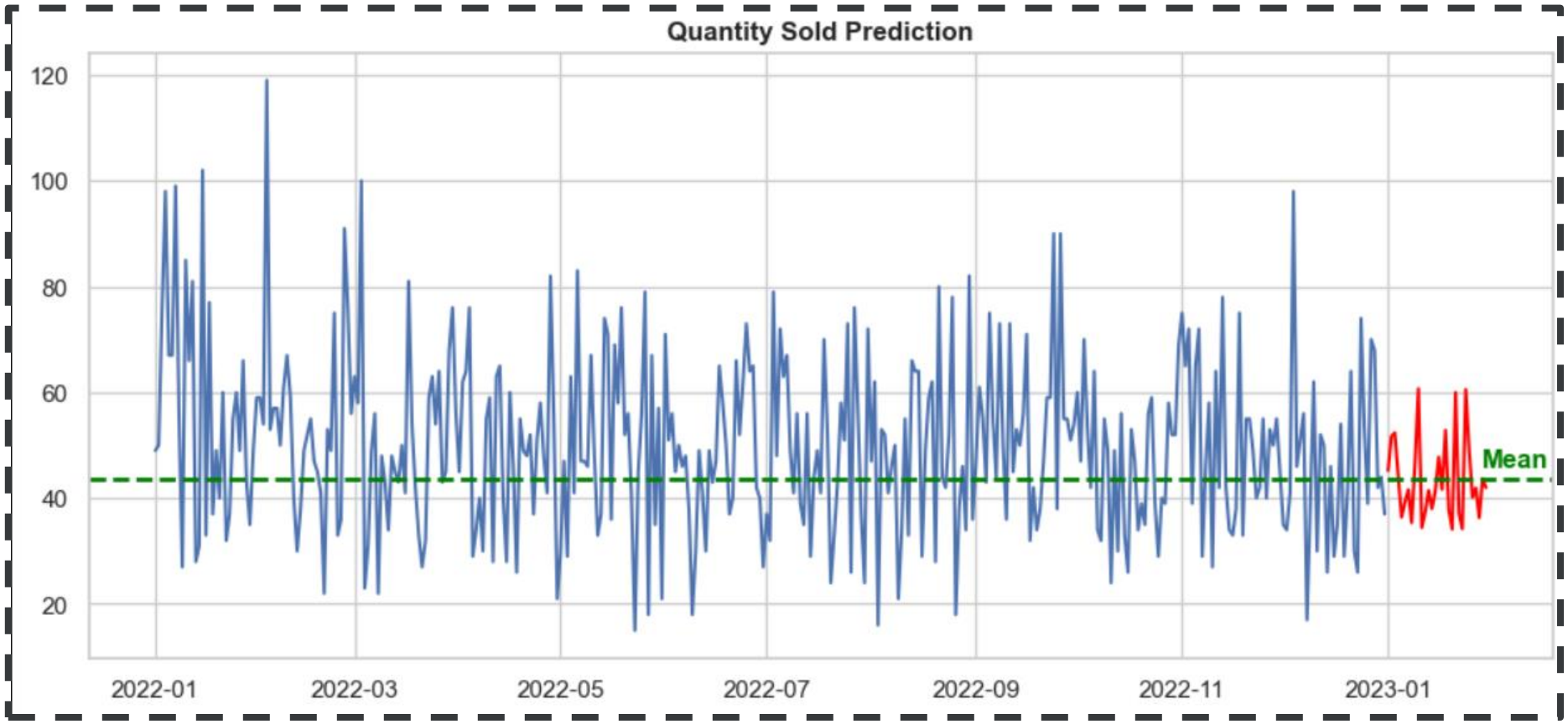
Manual

```
mae - manual: 0.2976
mape - manual: 0.0814
rmse - manual: 0.3603
```

From metric above, chosen parameters of ARIMA model are **(70, 2, 1)**

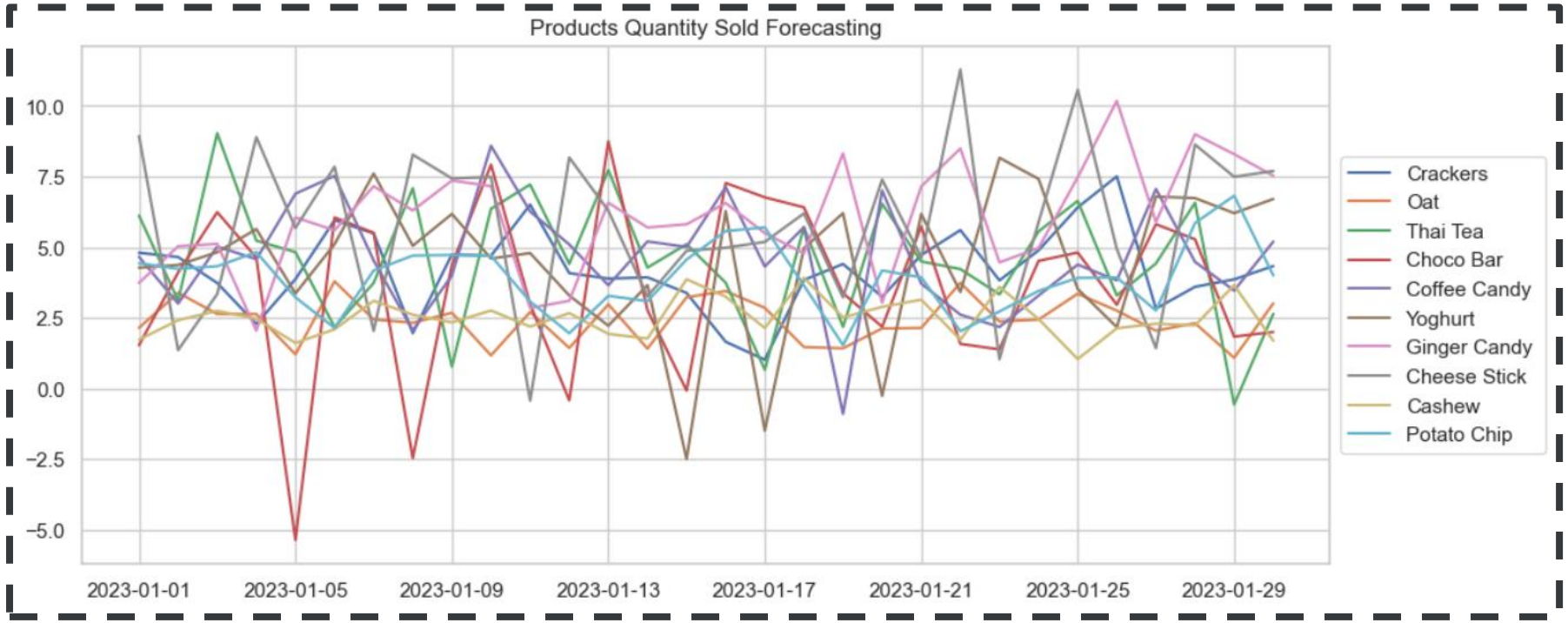


PREDICTING OVERALL QUANTITY



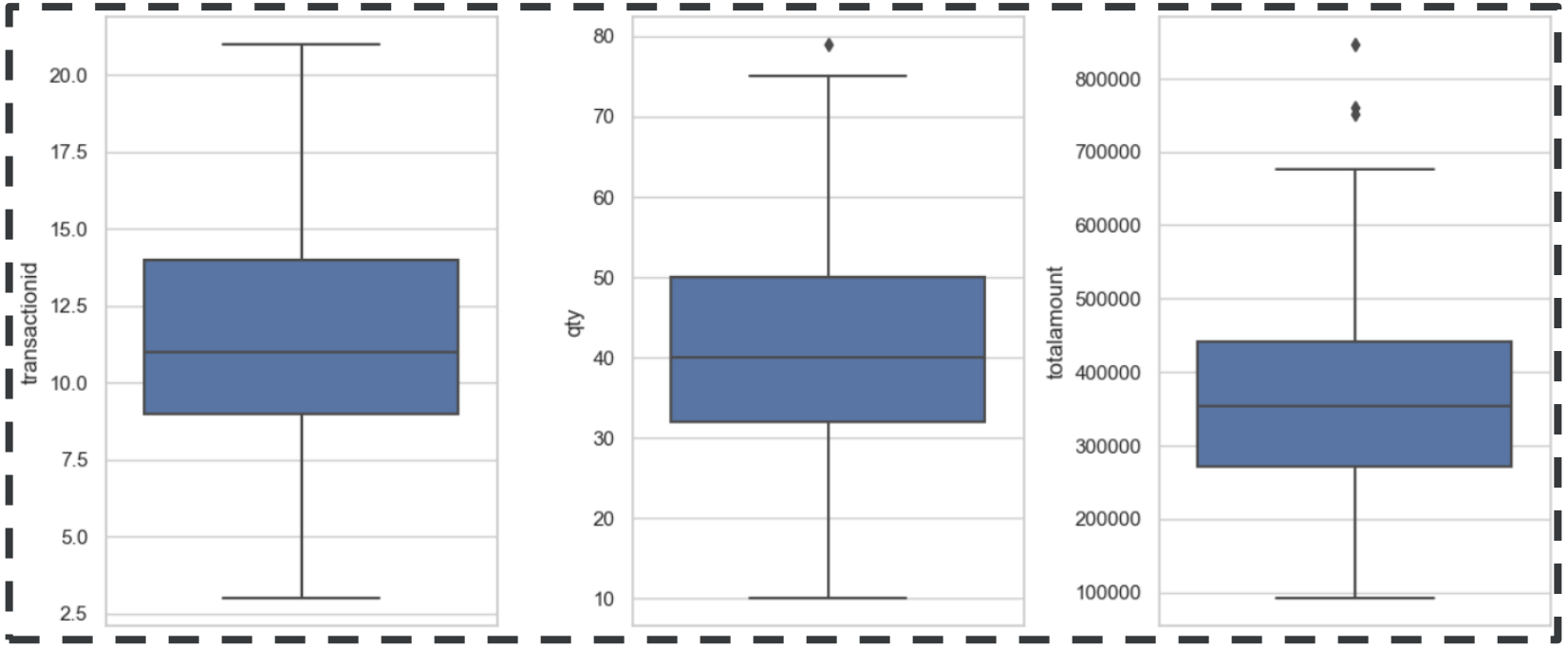
Mean = **43.53150997613 63**

PREDICTING EACH PRODUCT SOLD




```
Crackers          4.0
Oat               2.0
Thai Tea          5.0
Choco Bar         4.0
Coffee Candy      5.0
Yoghurt           5.0
Ginger Candy      6.0
Cheese Stick      6.0
Cashew            2.0
Potato Chip       4.0
Name: mean, dtype: float64
```

DATA PREPROCESSING FOR CLUSTERING

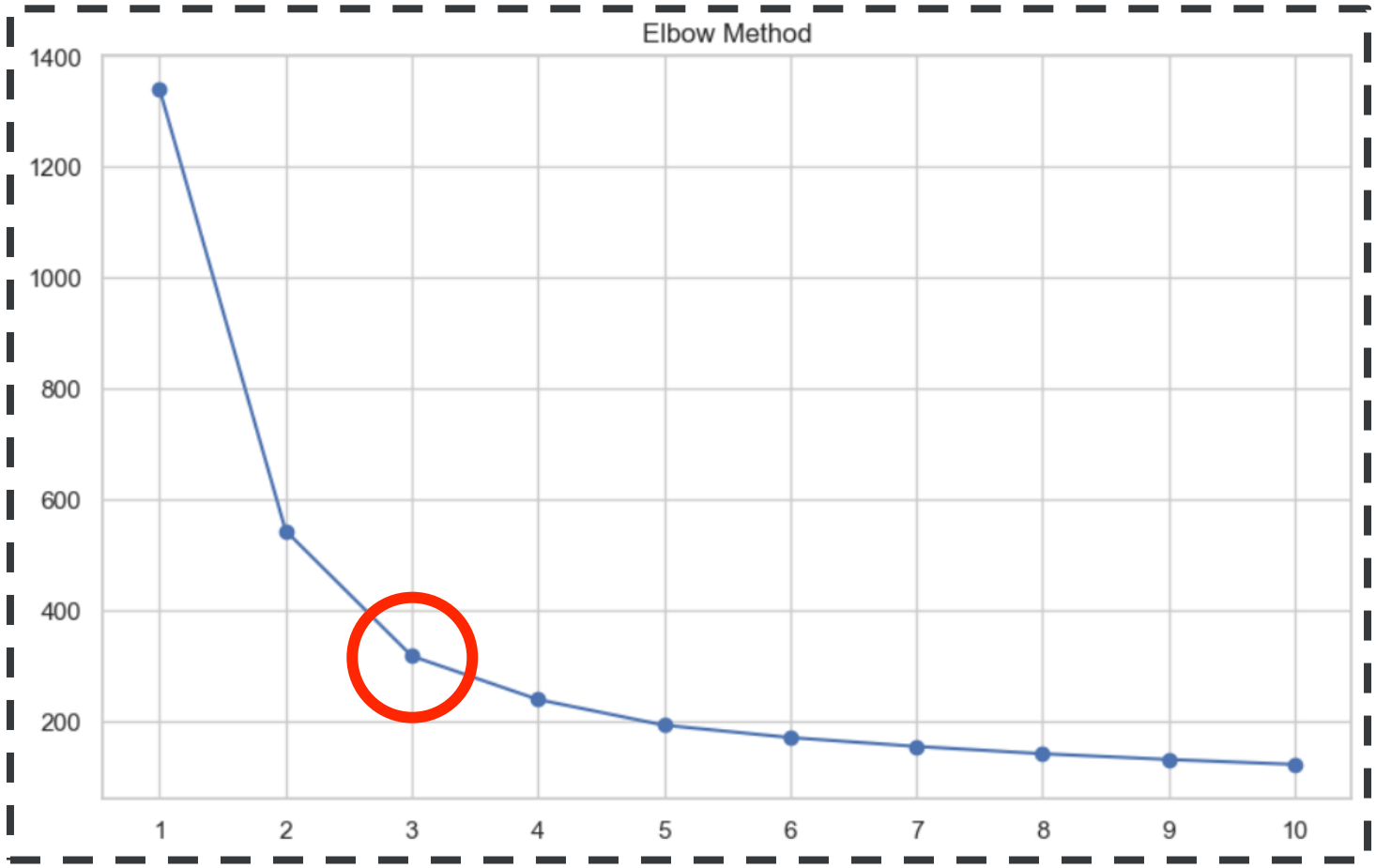


Data Transformation

- Transforming **qty** feature with Log Transformation due Right-Skewed distribution.

CLUSTERING MODELING

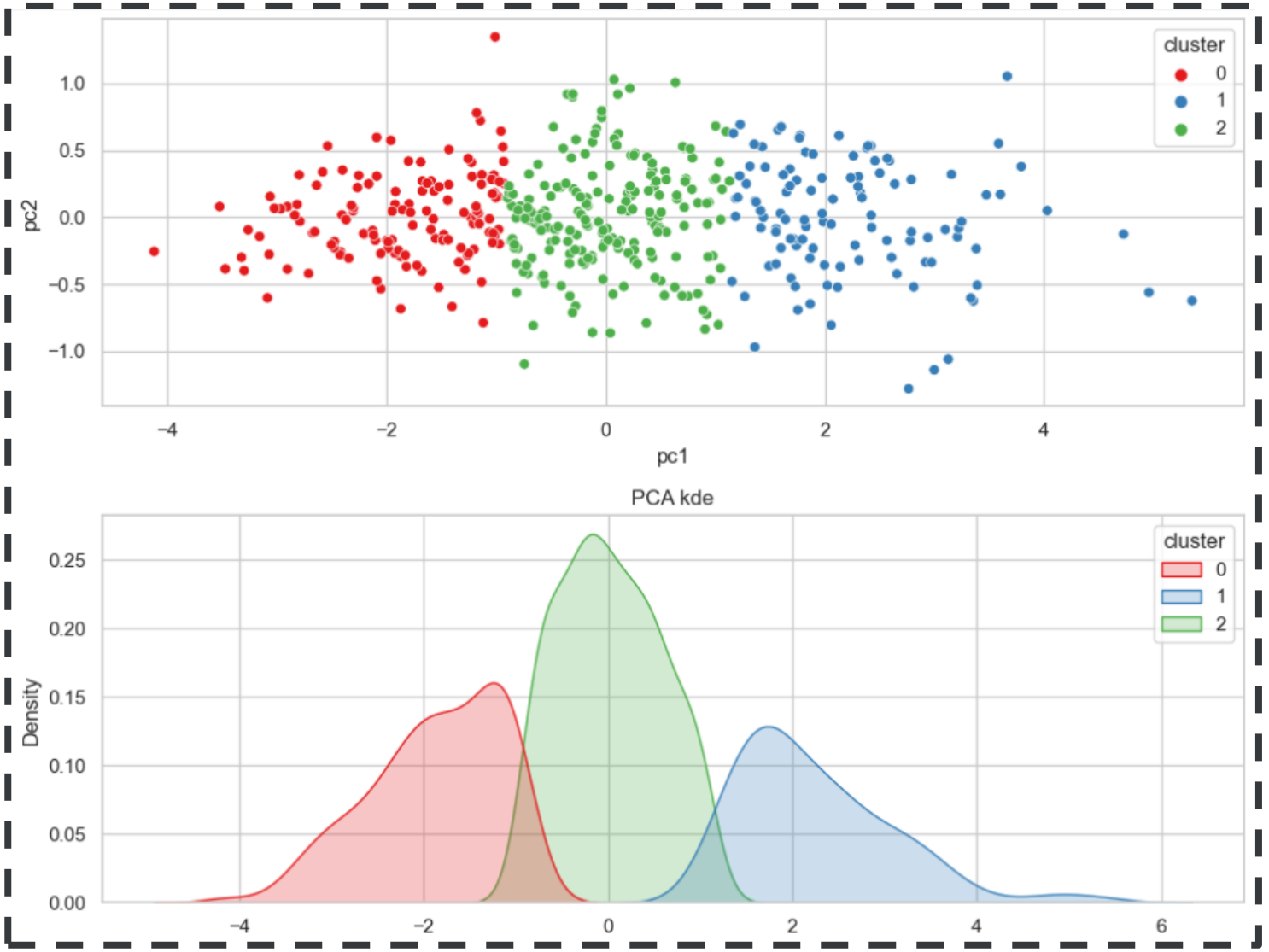
ELBOW METHOD



According to the elbow method above, we decided to choose **n_clusters=3** as the number of clusters

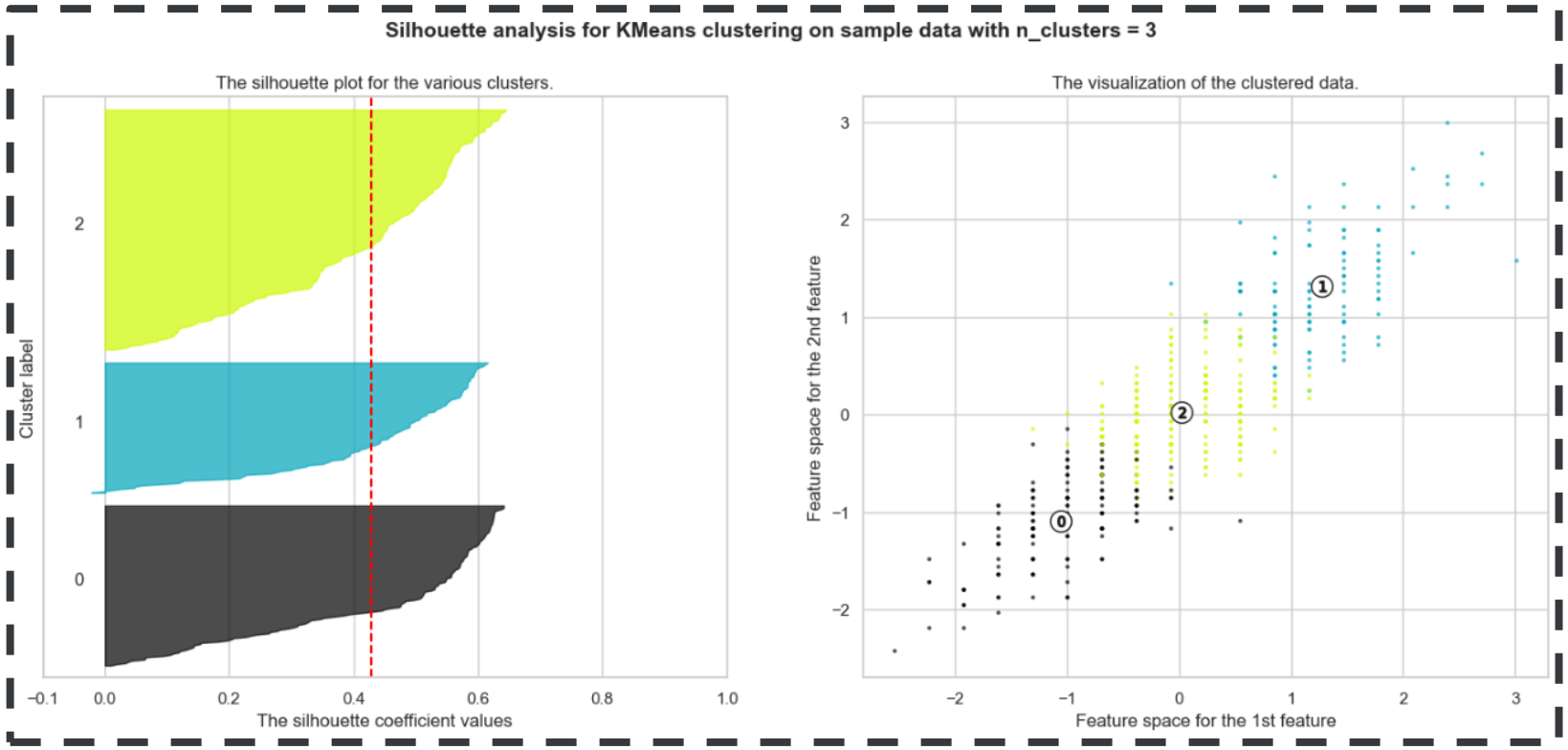
CLUSTERING MODELING

PRINCIPAL COMPONENT ANALYSIS (PCA)



CLUSTERING MODELING

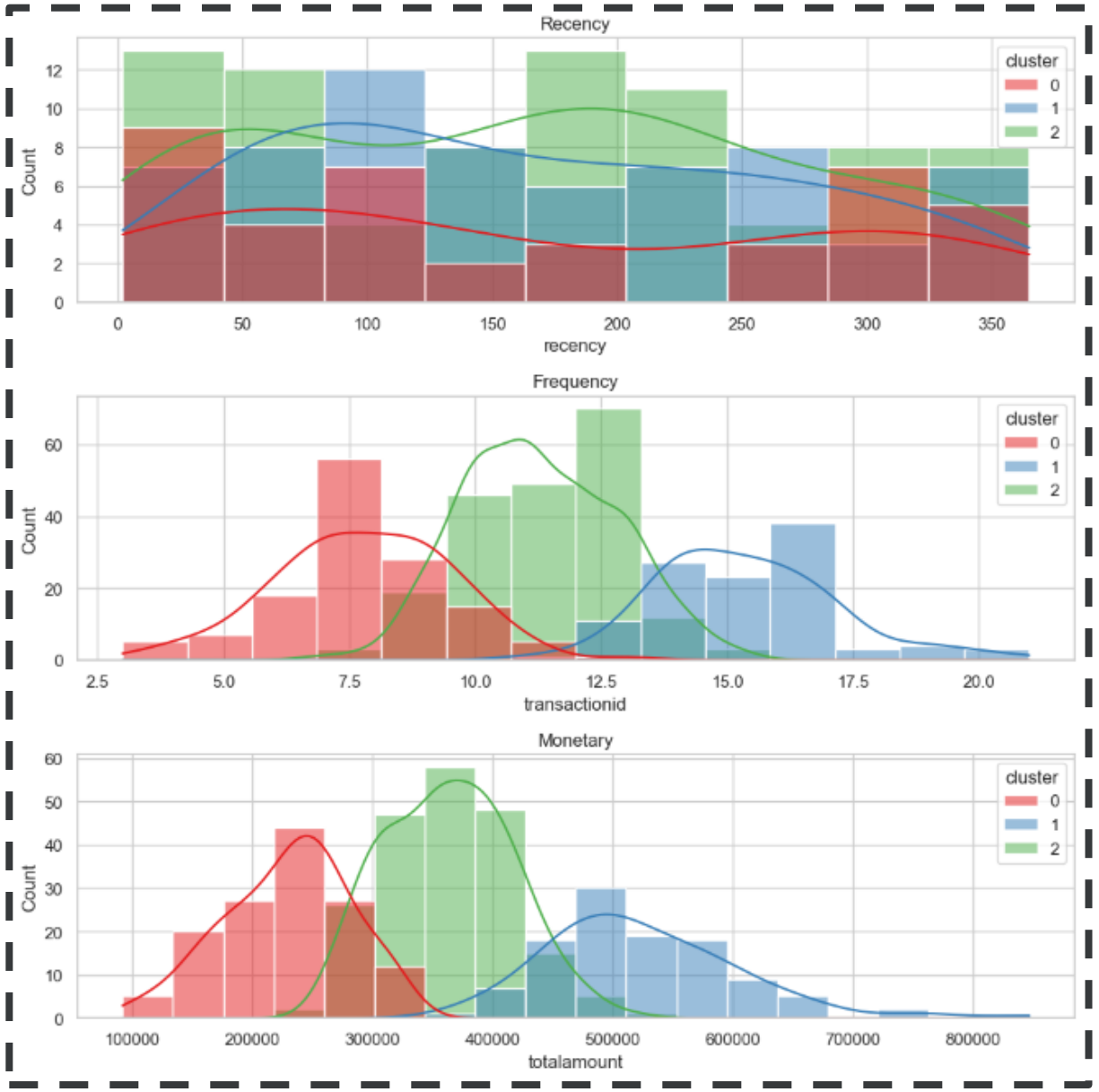
SILHOUETTE ANALYSIS



By silhouette analysis above, **n_cluster = 3** is fairly well divided with silhoette coefficient is **0.429**

INTERPRETATION

RFM (RECENCY, FREQUENCY, MONETARY) ANALYSIS



CLUSTER OBSERVATIONS

Cluster 0 (New Customer):

- ❑ Mostly having higher recency
- ❑ Have Lowest Frequency and Monetary
- ❑ Strategies Recommendations :
  - Provide support
  - Gift discount
  - Build Relationship

Cluster 1 (Potential Loyalist):

- ❑ Have medium recency
- ❑ Have Highest Frequency and Monetary
- ❑ Strategies Recommendations :
  - Offer loyalty program
  - Run contest
  - Make them feel special

Cluster 2 (Loyal Customer):

- ❑ Have Highest low recency
- ❑ Have Medium Frequency and Monetary
- ❑ Strategies Recommendations :
  - Take feedback and surveys
  - Upsell product
  - Present bonus

THANK YOU!

Check GitHub here

Check Tableau Dashboard here

Check Video Presentation here