Time Series & Clustering

Machine Learning Project

Presented by **Arif Styawan**

ABOUT

I am a mathematics student with a strong interest in pursuing a career in the field of data, focusing on **data analysis** and **machine learning** skills. I have proficiency in analytical skills and tools such as **Excel/Spreadsheets**, **Python**, **SQL**, and **Tableau**. I have a strong passion for learning and enjoy exploring new things.

EXPERIENCE

Bangkit Academy 2023
Machine Learning Student (Feb - July)

Pacmann Academy Scholarship
Analytics & Data Science Awardee (Feb - present)

TOOLS

linkedin.com/in/**arifstyawan**

github.com/**arifstyawan**

arifstyawaan@gmail.com

Overview

Saya berperan sebagai seorang **Data Scientist** di Kalbe Nutritionals dan sedang mendapatkan project baru dari tim inventory dan tim marketing.
Tujuan dari project ini yaitu:

1. Mengetahui **perkiraan** quantity product yang terjual sehingga tim inventory dapat membuat stock persediaan harian yang cukup.
2. Membuat **cluster/segment customer** berdasarkan beberapa kriteria.

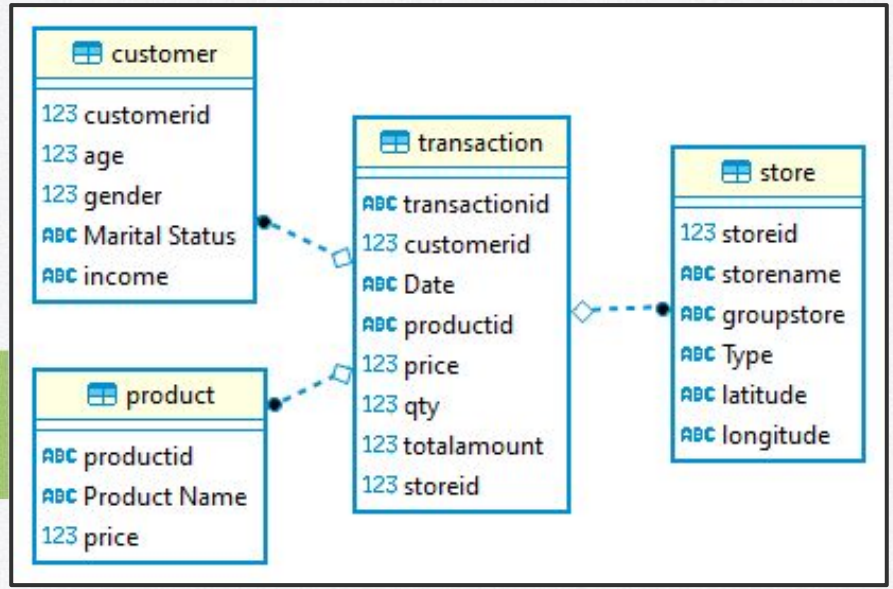
Challenge

Objektif yang harus dicapai antara lain:

1. Melakukan **data ingestion** dan **exploratory data analysis** ke dalam dbeaver, dengan query sebagai berikut.
   a. Berapa rata-rata umur customer jika dilihat dari marital statusnya?
   b. Berapa rata-rata umur customer jika dilihat dari gender nya?
   c. Tentukan nama store dengan total quantity terbanyak!
   d. Tentukan nama produk terlaris dengan total amount terbanyak!
2. Melakukan **data ingestion** dan membuat **dashboard** di Tableau
3. Membuat model machine learning **time series ARIMA**
4. Membuat model machine learning clustering dengan **K-Means**

1 Melakukan **data ingestion** dan **exploratory data analysis** ke dalam dbeaver

Dataset yang digunakan berupa 4 file dengan format .csv yaitu:
Case Study - Customer.csv, **Case Study - Product.csv**, **Case Study - Transaction.csv**, dan **Case Study - Store.csv**.
Kolom untuk setiap tabel beserta hubungan antar tabel sebagai berikut.



- Berapa **rata-rata umur customer** jika dilihat dari **marital status**nya?

	Marital Status	average_age
1		31.3333333333
2	Married	43.0382352941
3	Single	29.3846153846

Rata-rata umur untuk customer dengan status **Married** adalah **43 tahun** dan rata-rata dengan status **Single** adalah **29,4 tahun**.
Terdapat NULL data dengan rata-rata 31,3 tahun

- Berapa **rata-rata umur customer** jika dilihat dari **gender**nya?

	gender	average_age
1	0	40.326446281
2	1	39.1414634146

Rata-rata umur untuk customer **wanita** adalah **40 tahun** dan customer **pria** adalah **39 tahun**.

- Tentukan **nama store** dengan **total quantity** terbanyak!

	storename	total_qty
1	Lingga	2,777
2	Sinar Harapan	2,588

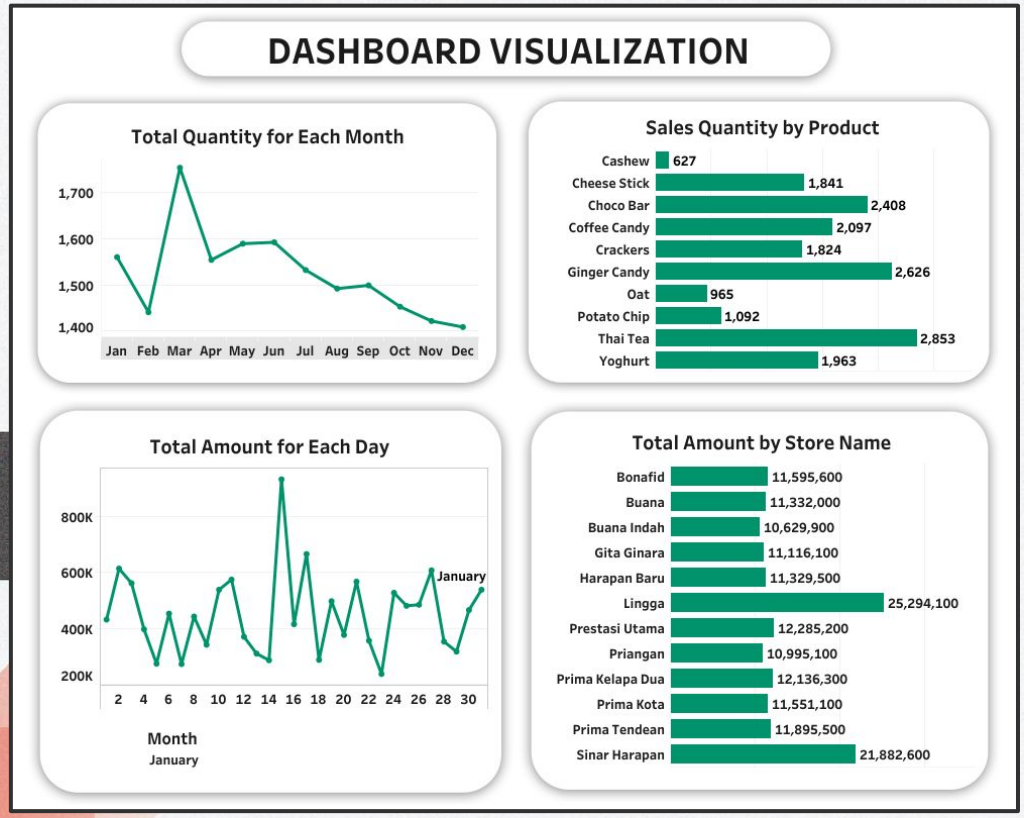
Nama store dengan total quantity terbanyak adalah **Lingga** dengan total quantity sebanyak **2.777**. Disusul oleh Store Sinar Harapan.

- Tentukan **nama produk** terlaris dengan **total amount** terbanyak!

	Product Name	total_amount
1	Cheese Stick	27,615,000
2	Choco Bar	21,190,400

Nama produk terlaris dengan total amount terbanyak adalah **Cheese Stick** dengan total amount sebesar **27.615.000**. Disusul oleh produk Choco Bar.

2 Melakukan **data ingestion** dan membuat **dashboard** di Tableau



Dashboard di samping terdiri dari 4 worksheet, sebagai berikut :

- **Worksheet 1**
  Jumlah qty dari bulan ke bulan
- **Worksheet 2**
  Jumlah total amount dari hari ke hari
- **Worksheet 3**
  Jumlah penjualan (qty) by product
- **Worksheet 4**
  Jumlah penjualan (total amount) by store name

3 Membuat model machine learning **time series ARIMA**

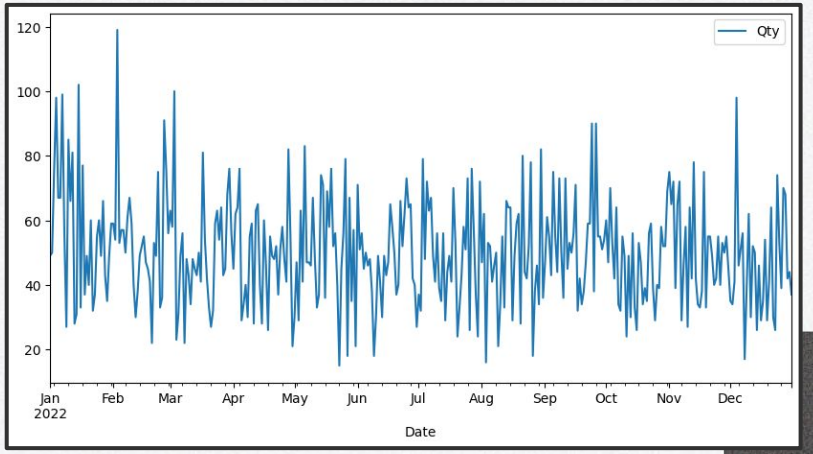
- **IMPORT LIBRARY**
  Library yang diimport berupa pandas, datetime, matplotlib, scikit-learn, dan library lainnya.

- **LOAD DATASET**
  Dataset yang digunakan yaitu ‘Case Study - Transaction.csv’.

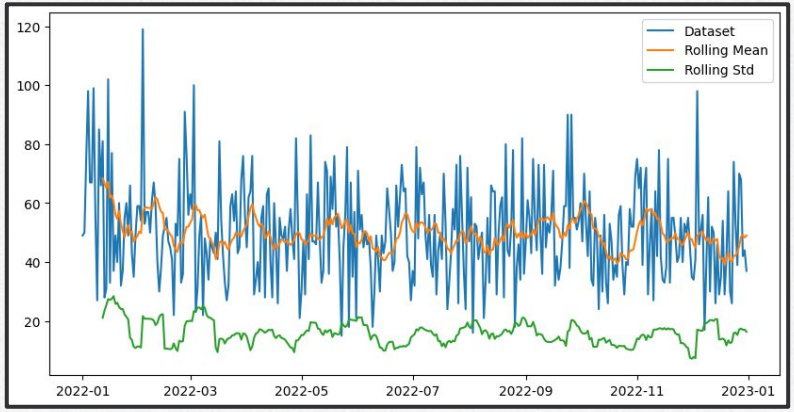
- **DATA CLEANSING & PREPROCESSING**
  a. Mengecek **missing values** (tidak ada missing values)
  b. Mengecek **tipe data** kemudian mengubah kolom 'Date' menjadi tipe **datetime**
  c. Melakukan **grouping** berdasarkan 'Date' dan menjumlahkan 'Qty'.
  d. Mengubah kolom 'Date' menjadi **indeks** dan membuat **plot** pada 'Qty'.

Date	Qty
2022-01-01	49
2022-01-02	50
2022-01-03	76
2022-01-04	98
2022-01-05	67
...	...
2022-12-27	70
2022-12-28	68
2022-12-29	42
2022-12-30	44
2022-12-31	37

[365 rows x 1 columns]



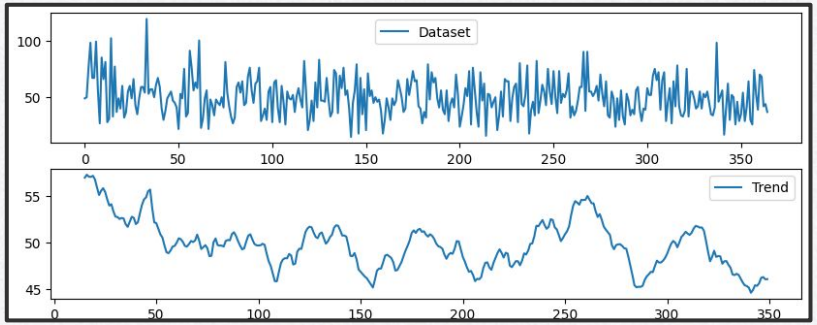
- **CHECK STATIONARITY**

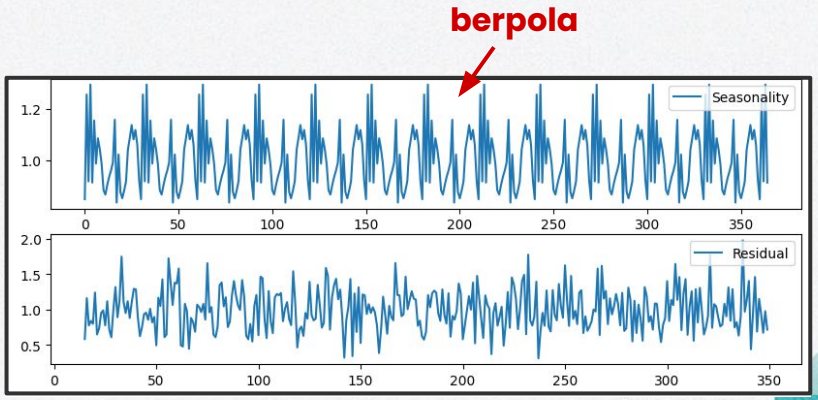


Mengecek stationarity menggunakan **rolling mean** dan **rolling standard deviation**.

Karena grafik rolling mean dan rolling std cenderung **konstan** dan tidak memiliki tren, maka data bersifat **stasioner**

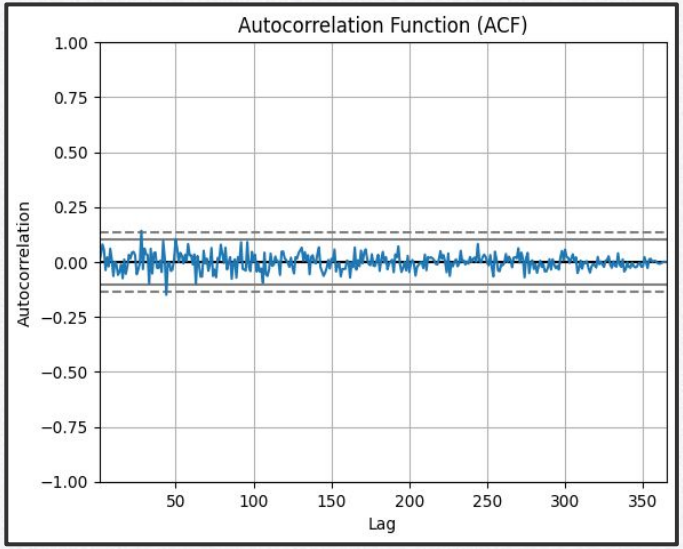
- **CHECK SEASONALITY**

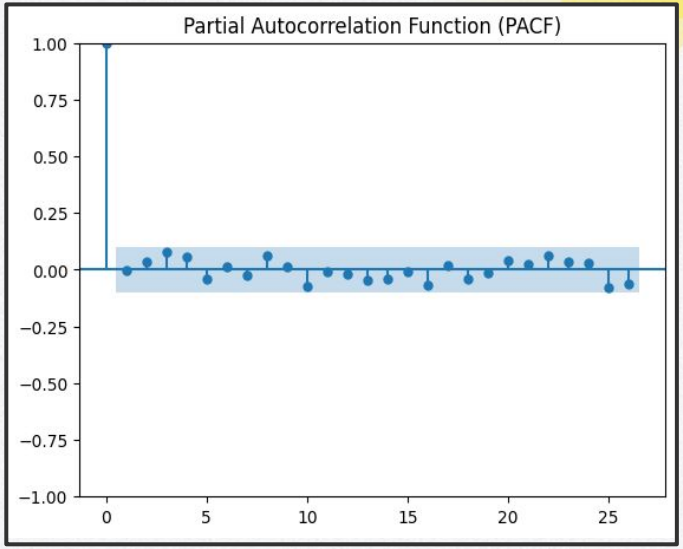




Terlihat bahwa data bersifat **seasonal (musiman)** karena memiliki pola setiap periode tertentu. Karena bersifat seasonal, maka nilai **d = 1**

- **CHECK ACF(p) AND PACF(q)**





Dari grafik **ACF** terlihat bahwa garis biru memotong 5 garis horizontal, sehingga **p = 5**
Dari grafik **PACF** terlihat bahwa terdapat 1 titik terjauh/mendekati 1, sehingga **q = 1**

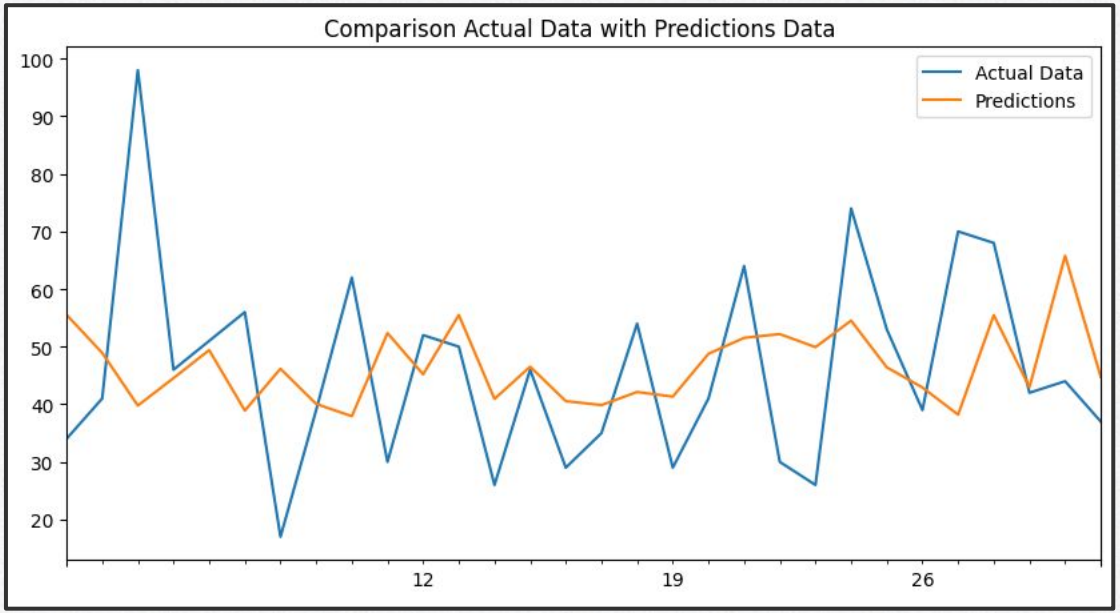
Jadi, dari check seasonality, ACF, dan PACF didapatkan **d = 1, p = 5, q = 1**

- **SPLIT DATASET**
  Split dataset dengan pembagian **train set = 335 data** dan **test set = 30 data**.

- **BUILD MODEL**
  Model machine learning yang dibuat merupakan **Seasonal ARIMA (SARIMA)**. Training model menggunakan data training yang telah di split. Berikut model summary nya:

```
==========================================================================================
Dep. Variable:                                 Qty   No. Observations:                  335
Model:             SARIMAX(5, 1, 1)x(5, 1, 1, 30)   Log Likelihood               -1311.517
Date:                            Wed, 26 Jul 2023   AIC                           2649.035
Time:                                    07:27:30   BIC                           2697.356
Sample:                                01-01-2022   HQIC                          2668.365
                                     - 12-01-2022
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0091      0.065     -0.140      0.888      -0.136       0.118
ar.L2         -0.0206      0.064     -0.319      0.749      -0.147       0.106
ar.L3          0.0703      0.061      1.152      0.250      -0.049       0.190
ar.L4          0.0964      0.061      1.577      0.115      -0.023       0.216
ar.L5         -0.0178      0.062     -0.287      0.774      -0.139       0.104
ma.L1         -0.9867      0.033    -30.266      0.000      -1.051      -0.923
ar.S.L30      -0.0276      0.167     -0.165      0.869      -0.354       0.299
ar.S.L60      -0.0288      0.143     -0.202      0.840      -0.308       0.251
ar.S.L90      -0.0805      0.127     -0.633      0.527      -0.330       0.169
ar.S.L120     -0.0705      0.114     -0.616      0.538      -0.295       0.154
ar.S.L150     -0.1235      0.111     -1.112      0.266      -0.341       0.094
ma.S.L30      -0.9944      6.695     -0.149      0.882     -14.117      12.128
sigma2       242.5074   1586.077      0.153      0.878   -2866.146    3351.161
==========================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                 0.98
Prob(Q):                              0.85   Prob(JB):                         0.61
Heteroskedasticity (H):               0.75   Skew:                            -0.09
Prob(H) (two-sided):                  0.15   Kurtosis:                         3.21
==========================================================================================
```

- **MODEL EVALUATION**
  Testing data menggunakan data testing. Selanjutnya dilakukan prediksi model untuk 30 data terakhir.
  Grafik di samping adalah perbandingan antara **actual data** dengan **prediksi model**.

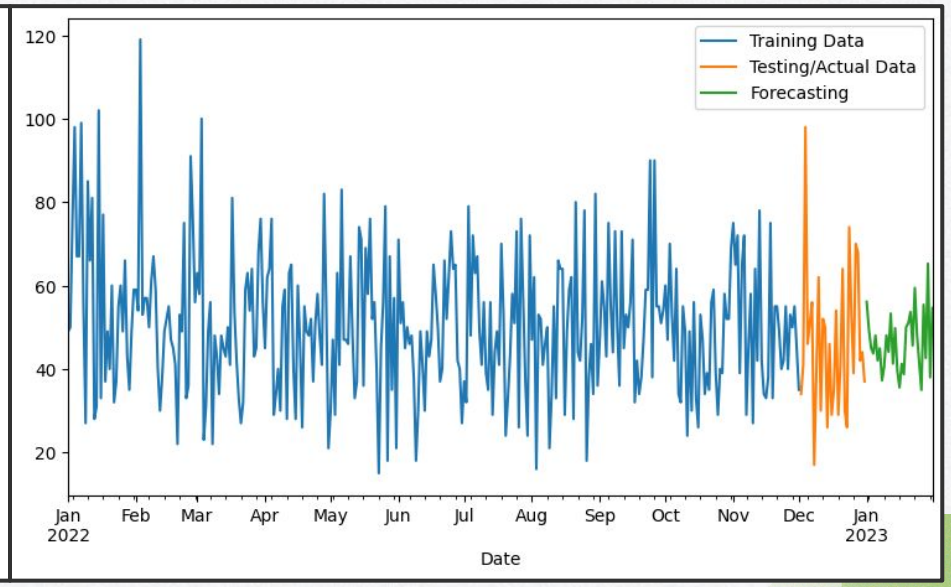


- **TES RMSE**
  Didapatkan RMSE sebesar **18,544**
  Angka ini menandakan bahwa model sudah berjalan cukup baik.

- **FORECASTING FOR NEXT MONTH**
  Berikut merupakan hasil prediksi jumlah quantity untuk **31 hari ke depan** dari hari ke hari. Berikut terdapat juga visualisasi training data, actual data, dan prediksi 31 hari kedepan.

2023-01-01	56.098186
2023-01-02	49.546683
2023-01-03	45.055179
2023-01-04	43.692128
2023-01-05	48.016123
2023-01-06	42.033181
2023-01-07	44.946107
2023-01-08	37.256635
2023-01-09	40.647687
2023-01-10	48.043476
2023-01-11	44.107906
2023-01-12	53.294111
2023-01-13	41.294605
2023-01-14	49.787157
2023-01-15	39.550210

2023-01-16	35.546540
2023-01-17	41.218269
2023-01-18	38.742441
2023-01-19	50.008640
2023-01-20	50.975427
2023-01-21	53.648984
2023-01-22	45.616540
2023-01-23	59.438613
2023-01-24	48.081410
2023-01-25	41.508262
2023-01-26	34.955224
2023-01-27	55.409811
2023-01-28	42.621646
2023-01-29	65.269550
2023-01-30	37.982738
2023-01-31	54.615704



4 Membuat model machine learning clustering **K-Means**

- **IMPORT LIBRARY**
  Library yang diimport berupa pandas, numpy, matplotlib, scikit-learn, dan seaborn.

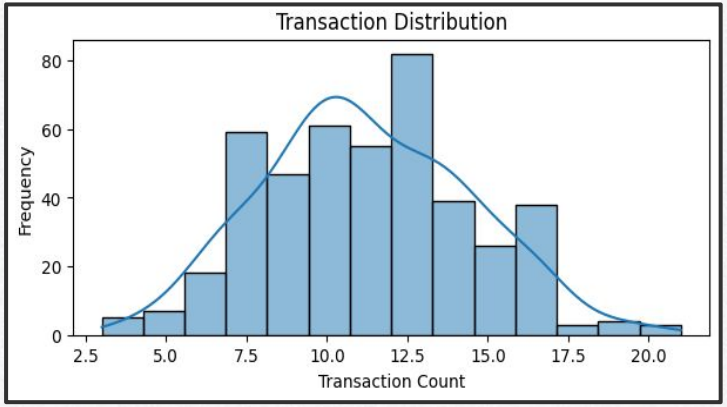
- **LOAD DATASET**
  Dataset yang digunakan yaitu 'Case Study - Transaction.csv'.

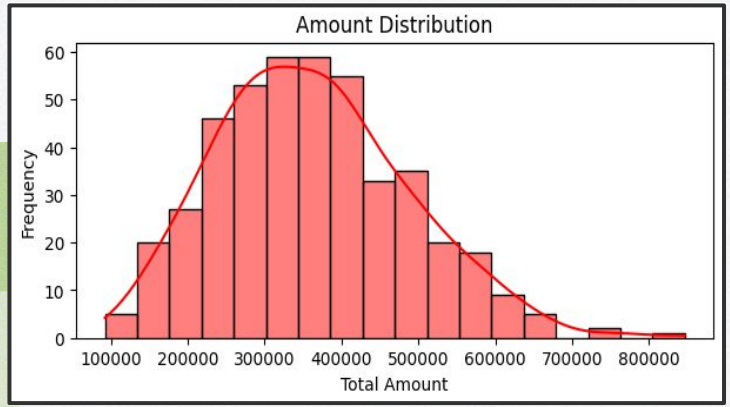
- **DATA CLEANSING & PREPROCESSING**
  a. Mengecek **missing values** (tidak ada missing values)
  b. Membuat data baru dengan **groupby** berdasarkan customerID dan di**aggregasi** dengan Transaction id (count), Qty (sum), Total amount (sum)
  c. Didapat data baru dengan 447 baris dan 4 kolom, diantaranya **CustomerID**, **TransactionCount**, **TotalQty**, dan **TotalAmount**
  d. Mengubah kolom CustomerID menjadi indeks kolom
  e. Mengecek konsistensi format/tipe data

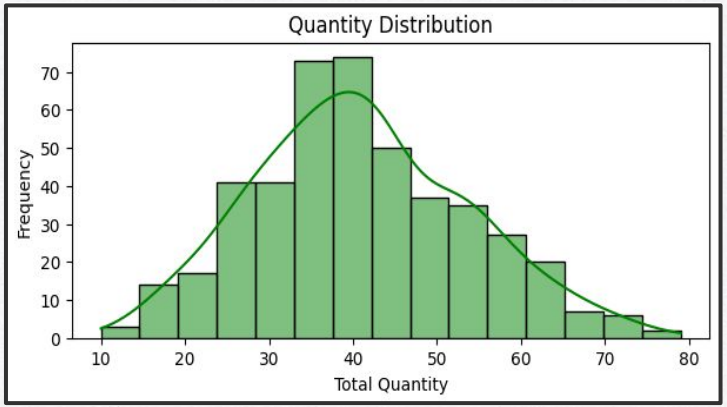
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 447 entries, 1 to 447
Data columns (total 3 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   TransactionCount  447 non-null    int64
 1   TotalQty          447 non-null    int64
 2   TotalAmount       447 non-null    int64
dtypes: int64(3)
memory usage: 14.0 KB
```

TransactionCount = Tipe integer
TotalQty = Tipe integer
TotalAmount = Tipe integer

- **CHECK DATA DISTRIBUTION**
  Berikut visualisasi persebaran dari tiap data, yaitu Transaction Count, Total Quantity, dan Total Amount.



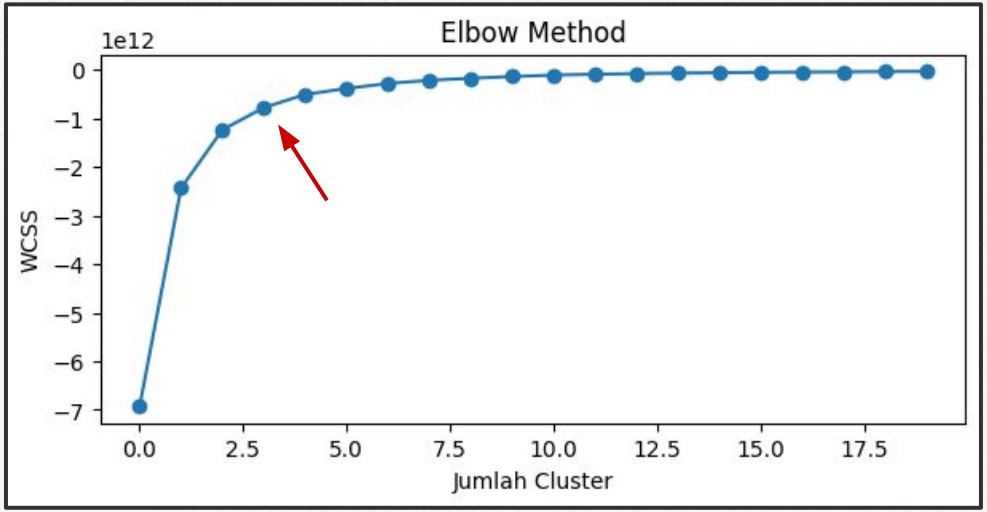




- **BUILD K-MEANS MODEL CLUSTERING**

  Langkah pembuatan K-Means model sebagai berikut

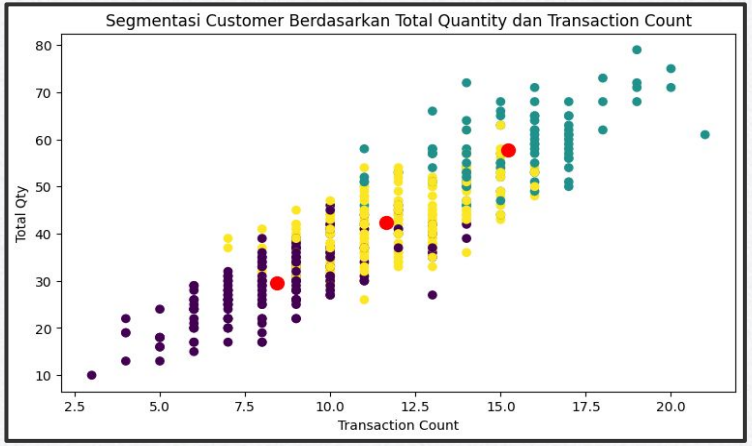
  1. Mencari nilai jumlah **n cluster** yang paling optimal.
  2. Menggunakan looping dengan **n = 1** sampai **n = 20**
  3. Visualisasikan hasilnya kemudian analisis menggunakan **Elbow Method**

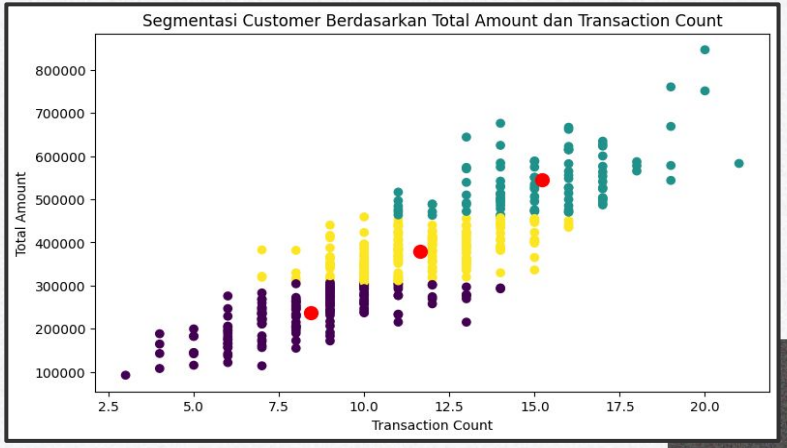


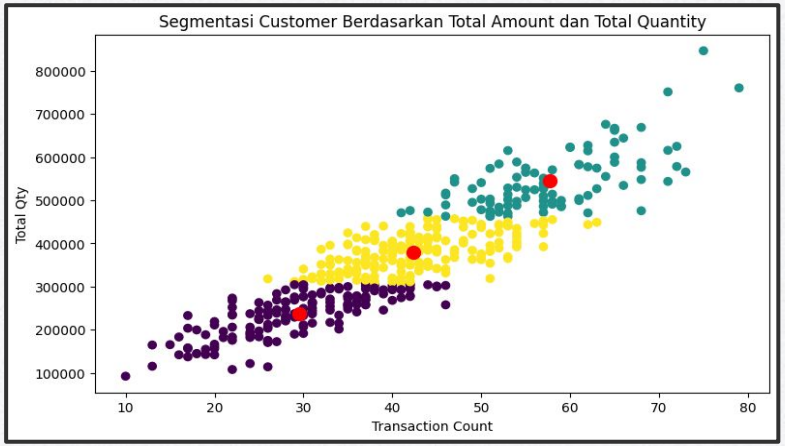
  4. Dengan Elbow Method, didapatkan **n = 3** yang paling optimal
  5. Build model K-Means menggunakan **n = 3**

```
model = KMeans(n_clusters=3)
model.fit(df)
```

- **CLUSTERING**
  Berikut visualisasi persebaran data berdasarkan segmentasi/clustering







DASHBOARD VISUALIZATION :
https://public.tableau.com/views/SalesDashboardVisualization/Dashboard1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

GITHUB REPOSITORY :
https://github.com/arifstyawan/TimeSeries-Clustering

VIDEO PRESENTATION :
https://youtu.be/3KAVt50YO_U

Thank You