

# **Machine Learning Project Membuat model Regression dan Clustering**

## **Kalbe Nutritionals Data Scientist Project Based Internship Program**

Presented by  
Putri Kirey Eki Yogaswari



# Putri Kirey Eki Yogaswari

## About You

I am a fresh graduate from Public Health with a focus on Health Statistics. During college, I have knowledge of situation analysis, designing health programs, leadership skills, systems thinking and data analysis. I have a career interest in data management and proficient in using SPSS and Ms. Excel. Has managed data up to 2000+.

## My Experiences

-  Project-Based Intern: Big Data Analytics  
Virtual Internship  
PT. KIMIA FARMA, TBK
-  Data Team Lead  
Universitas Muhammadiyah Prof. Dr. Hamka Jakarta
-  Data Management Staf  
Pusat Data dan Informasi, Kemenkes RI

# Case Study

- 1. Data ingestion ke dalam dbeaver**
- 2. Melakukan exploratory data analysis di dbeaver**
- 3. Data ingestion ke dalam tableau public**
- 4. Membuat dashboard di tableau**
- 5. Membuat model machine learning regression (Time Series)**
- 6. Membuat model machine learning clustering (K-Means)**

# Melakukan Exploratory Data Analysis di Dbeaver

Rata-rata umur customer dilihat dari marital statusnya

```
-- rata-rata umur customer jika dilihat dari marital statusnya
SELECT
    "Marital Status",
    round(avg("Age"),2) AS average_age
FROM
    customer_2 c
WHERE "Marital Status" != ''
GROUP BY
    "Marital Status";
```

Marital Status	average_age
Married	43.04
Single	29.38

Nama store dengan total quantity terbanyak

```
-- Nama Store dengan Total Quantity Terbanyak
SELECT s."StoreName", SUM(t."Qty") AS total_qty
FROM "store_2" s
JOIN "transaction_2" t ON s."StoreID" = t."S"
GROUP BY s."StoreName"
ORDER BY total_qty DESC
LIMIT 1;
```

StoreName	total_qty
Lingga	2,777

Rata-rata umur customer dilihat dari gender

```
-- Rata-rata Umur Customer Jika Dilihat dari Gender nya
SELECT
    "Gender",
    ROUND(AVG(c."Age"), 2) AS average_age
FROM
    customer_2 c
GROUP BY
    "Gender";
```

Gender	average_age
0	40.33
1	39.14

Nama produk terlaris dengan total amount terbanyak

```
-- Nama Produk Terlaris dengan Total Amount Terbanyak
SELECT
    p."Product Name",
    SUM(t."TotalAmount") AS total_sales_amount
FROM "product_2" p
JOIN "transaction_2" t ON p."ProductID" = t."ProductID"
GROUP BY p."Product Name"
ORDER BY total_sales_amount DESC
LIMIT 1;
```

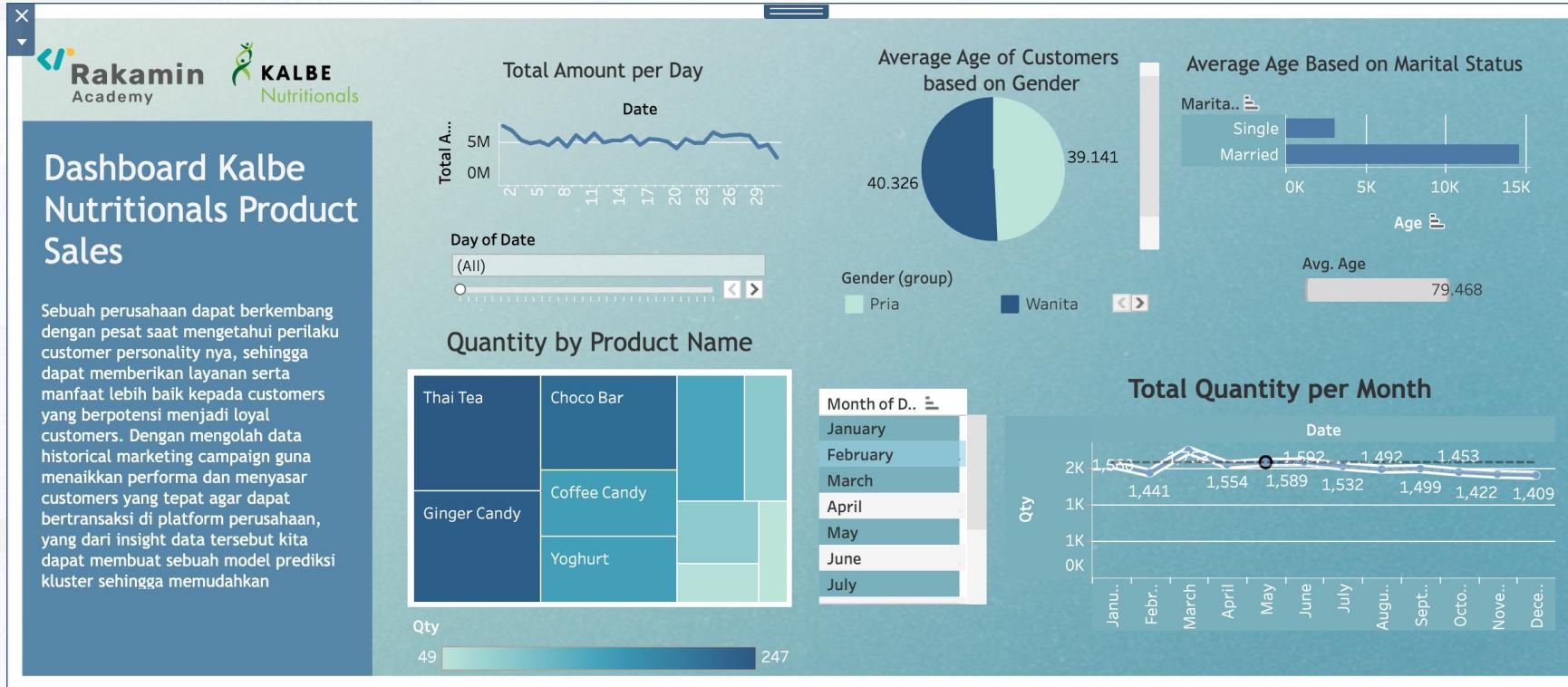
Product Name	total_sales_amount
Cheese Stick	27,615,000,-

## Insight:

- Customer dengan status menikah memiliki rata-rata usia 43 tahun, lebih tua dibandingkan dengan status single dengan rata-rata 29 tahun.
- Customer berjenis kelamin Wanita memiliki rata-rata usia 40 tahun, sedangkan Pria memiliki rata-rata usia 39 tahun.
- Nama toko dengan total quantity terbanyak adalah toko Lingga dengan jumlah quantity yaitu 2.777 item.
- Nama produk terlaris dengan total amount terbanyak adalah Cheese Stick dengan total amount terbanyak yaitu 27.615.000,-.

Selengkapnya lihat [disini](#)

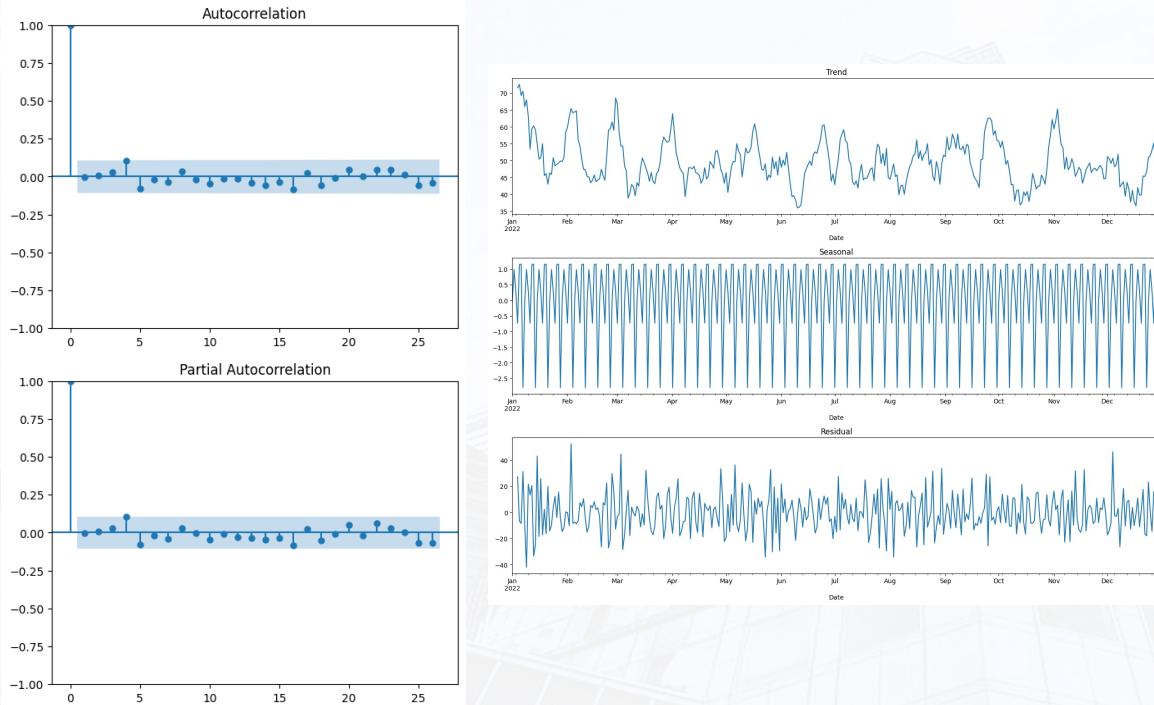
# Membuat dashboard di tableau



Selengkapnya lihat [disini](#)

# Model prediktif menggunakan regresi

## Plot Data Time series



## Observasi:

- Dari ADF Test kita dapat melihat bahwa nilai p lebih dari 0,05 yang berarti hipotesis nol kita akan ditolak dan deret ini dianggap sudah stasioner.
- Dari plot ACF dan PACF, data tersebut sudah stationary dan bisa digunakan untuk ARIMA model.

# Model prediktif menggunakan regresi

2 metode untuk mendapatkan parameter ( $p, d, q$ ) untuk menghasilkan forecast yang akurat yaitu dengan melakukan auto-fit ARIMA dan manual parameter tuning

## Auto fit ARIMA

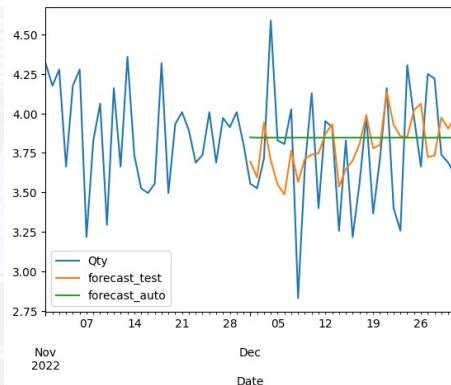
```
#auto-fit ARIMA
auto_arima = pm.auto_arima(df_train, stepwise=False, seasonal=False)
auto_arima
```

```
▼ ARIMA
ARIMA(1,0,3)(0,0,0)[0]
```

## Manual Parameter Tuning

	param	mae	mape	rsme
2	(70, 2, 1)	0.309134	0.082776	0.395608
1	(60, 2, 1)	0.328327	0.088083	0.415141
0	(50, 2, 1)	0.380785	0.105433	0.455044

mae - manual: 0.3019  
 mape - manual: 0.0825  
 rmse - manual: 0.37



## Observasi:

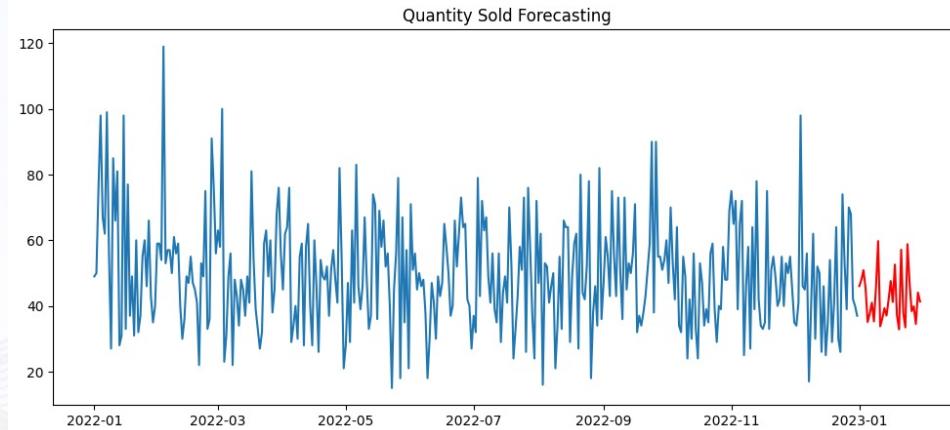
Dari kedua metrics kita akan pilih ARIMA model dengan Manual Parameter Tuning (70, 2, 1)

# Model prediktif menggunakan regresi

## Forecasting Overall Quantity

```
forecast.mean()
```

```
42.57925225656516
```

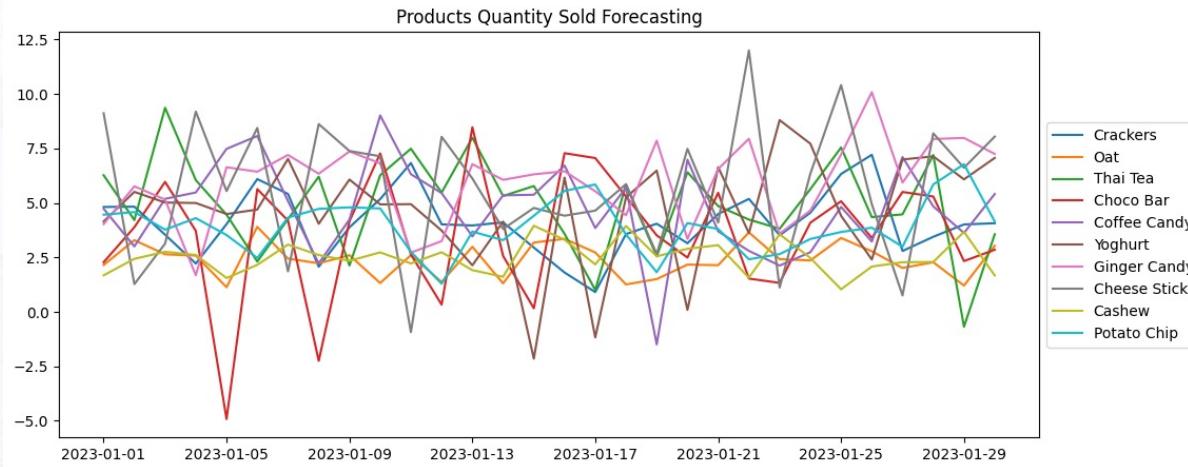


### Observasi:

Dari forecasting di atas dapat disimpulkan bahwa untuk quantity penjualan bulan depan adalah sekitar rata-rata 43 pcs per harinya.

# Model prediktif menggunakan regresi

## Forecasting Overall Quantity



```
#Products  Quantity forecast  
round(dfp.describe().T['mean'],0)
```

Products	Quantity forecast
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

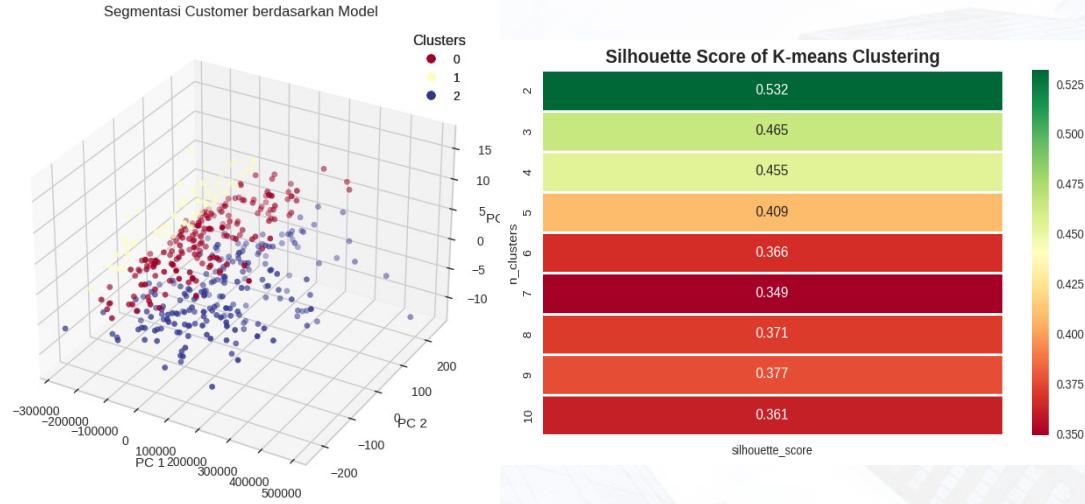
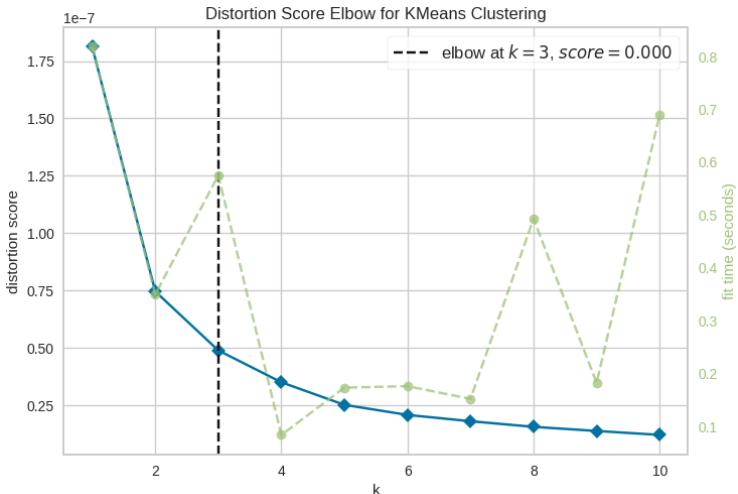
Name: mean, dtype: float64

### Observasi:

Dari data forecasting product diatas kita bisa mendapatkan perkiraan rata-rata quantity produk yang terjual setiap harinya. Jenis produk yang paling banyak terjual adalah Cheese Stick dan Ginger Candy yaitu sebanyak 6 pcs per harinya. Sedangkan jenis produk yang kurang terjual adalah Oat dan Cashew yaitu 2 pcs per harinya.

# Machine Learning

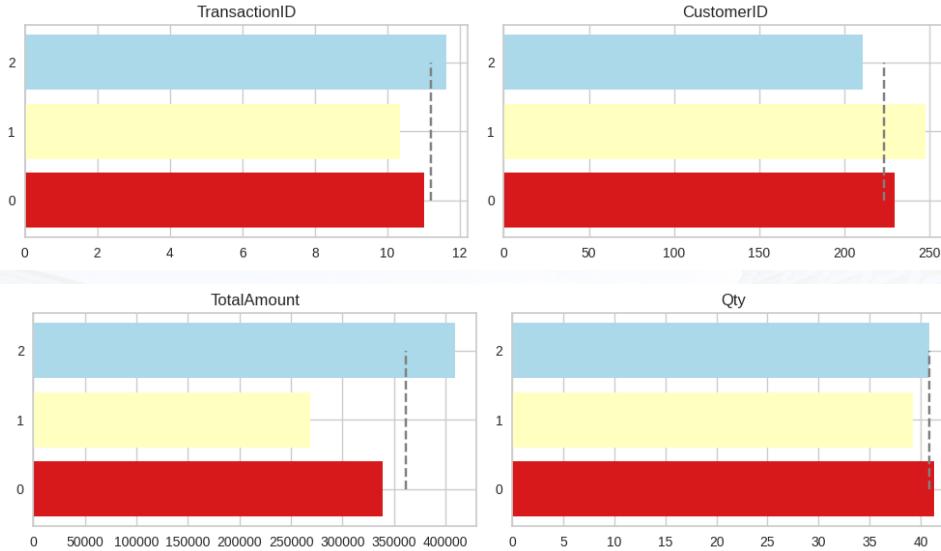
## Clustering (K-Means)



### Observasi:

- Berdasarkan grafik diatas, parameter distortion menunjukkan nilai optimal k = 3
- Silhouette Score n\_clusters = 3 adalah 0.465

# Segmentation Customer Analysis



## Characteristic Customer:

### 1. Cluster 0 (Loyalist Customer):

Jumlah transaksi stabil, jumlah customer cukup banyak, total belanja yang dikeluarkan cukup besar dan paling banyak item produk kalbe yang dibelanjakan.

### 2. Cluster 1 (New Customer):

Jumlah transaksi sedikit, jumlah customer paling banyak, total belanja yang dikeluarkan sangat sedikit dan kurang membelanjakan item produk kalbe dibandingkan kelompok lain.

### 3. Cluster 2 (Potential Loyalist):

Jumlah transaksi paling banyak, jumlah customer paling sedikit, total belanja yang dikeluarkan sangat besar, cukup banyak item produk kalbe yang dibelanjakan.

# Business Recommendation:

## 1. Cluster 0 (Loyalist Customer):

- **Program Loyalty:** Tingkatkan loyalitas pelanggan dengan mengembangkan program loyalitas yang memberikan insentif kepada pelanggan setia. Ini bisa berupa diskon khusus, hadiah, atau penawaran eksklusif untuk pelanggan dalam cluster ini.
- **Ekspansi Produk:** Tawarkan lebih banyak produk Kalbe yang relevan kepada pelanggan dalam cluster ini. Mungkin ada produk baru yang dapat menarik minat mereka, atau variasi produk yang dapat meningkatkan nilai belanja mereka.
- **Pelayanan Pelanggan Terbaik:** Pastikan pelayanan pelanggan yang sangat baik, seperti layanan pengiriman cepat dan responsif terhadap pertanyaan dan masalah pelanggan, untuk mempertahankan pelanggan dalam cluster ini.

## 2. Cluster 1 (New Customer):

- **Program Pemasaran Target:** Gunakan strategi pemasaran yang ditargetkan untuk meningkatkan kesadaran tentang produk Kalbe di antara pelanggan dalam cluster ini. Ini bisa mencakup kampanye iklan online, konten sosial media, atau promosi khusus.
- **Penawaran Khusus untuk Pelanggan Baru:** Tawarkan penawaran khusus, diskon, atau paket bundel produk kepada pelanggan baru untuk mendorong mereka untuk melakukan lebih banyak transaksi.
- **Program Penghargaan untuk Mengundang Teman:** Buat program referensi di mana pelanggan dalam cluster ini dapat mendapatkan insentif jika mereka mengundang teman-teman mereka untuk berbelanja produk Kalbe.

## 3. Cluster 2 (Potential Loyalist):

- **Fokus pada Retensi:** Meskipun jumlah pelanggan dalam cluster ini sedikit, mereka memiliki potensi besar untuk menjadi pelanggan setia. Berfokus pada mempertahankan dan meningkatkan kepuasan pelanggan dalam cluster ini.
- **Program Eksklusif:** Tawarkan program eksklusif seperti keanggotaan premium yang memberikan manfaat khusus kepada pelanggan dalam cluster ini, seperti akses terhadap produk terbaru atau penawaran eksklusif.
- **Up-selling dan Cross-selling:** Identifikasi produk-produk Kalbe yang paling diminati oleh pelanggan dalam cluster ini dan tawarkan produk-produk terkait atau produk-produk dengan nilai tambah yang lebih tinggi.



**Link SQL, Python dan dataset selengkapnya [disini](#)**



**Video presentasi penjelasan project dapat dilihat [disini](#)**

# Thank You



**Rakamin**  
Academy



**KALBE**  
Nutritionals