Rakamin Academy X KALBE Nutritionals

Machine Learning Project

Data Science VIX

Presented by
Sakinah Nurul Ramadhani

Sakinah Nurul Ramadhani

About You

As a recent graduate with a specialization in health statistics, I am currently seeking roles as a data analyst or data scientist. My background equips me with strong skills in situational analysis, systematic thinking, and effective data analysis. I am experienced in data collection, data cleaning, and predictive modeling.

Rakamin Academy

Experience :

Experience 1

BKKBN

- Making an Alokon Obygn Bed & IUD KIT report on the Ms. Excel and do data visualization with Nitro Pro application
- Create an achievement data dashboard vaccination family year 2021 with application Tableu

Experience 2

Kimia Farma X Rakamin Academy

- *Project-Based Intern : Big Data Analytics Virtual Internship Experience Program*

Experience 3

Rakamin Academy

- Bootcamp Data Science Batch 32 with Excellent Grade

Case Study

Rakamin Academy

1. **Dapat melakukan data ingestion ke dalam dbeaver**
2. **Dapat melakukan exploratory data analysis di dbeaver :**

- query 1 : Berapa rata-rata umur customer jika dilihat dari marital statusnya ?
- query 2 : Berapa rata-rata umur customer jika dilihat dari gender nya ?
- query 3 : Tentukan nama store dengan total quantity terbanyak!
- Query 4 : Tentukan nama produk terlaris dengan total amount terbanyak!

3. **Dapat melakukan data ingestion ke dalam tableau public**

4. **Dapat membuat dashboard di tableau :**

- 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

5. **Dapat membuat model machine learning regression (Time Series)**
6. **Dapat membuat model machine learning clustering (K-Means)**

Rakamin Academy

Dbeaver (SQL)

Rata-rata Umur Customer Jika Dilihat Dari Marital Statusnya

```
select
    "Marital Status" ,
    round(avg(age),2) as average_age
from
    customer c
where "Marital Status" != ''
group by
    "Marital Status" ;
```

customer 1

	Marital Status	average_age
1	Married	43.04
2	Single	29.38

Rata-rata Umur Customer Jika Dilihat dari Gender nya

```
select
    gender,
    round(avg(age),2) AS average_age
from
    customer
group by
    gender;
```

customer 1

	gender	average_age
1	0	40.33
2	1	39.14

Nama Store dengan Total Quantity Terbanyak

```
select s.storename, sum(t.qty) as total_qty
from store s
join transaction t
on s.storeid = t.storeid
group by s.storename
order by total_qty desc
limit 1;
```

store 1

	storename	total_qty
1	Lingga	2,777

Nama Produk Terlaris dengan Total Amount Terbanyak

```
select
    p."Product Name" ,
    sum(t.totalamount) as total_sales_amount
from
    product p
join transaction t
on p.productid = t.productid
group by p."Product Name"
order by total_sales_amount desc
limit 1;
```

product 1

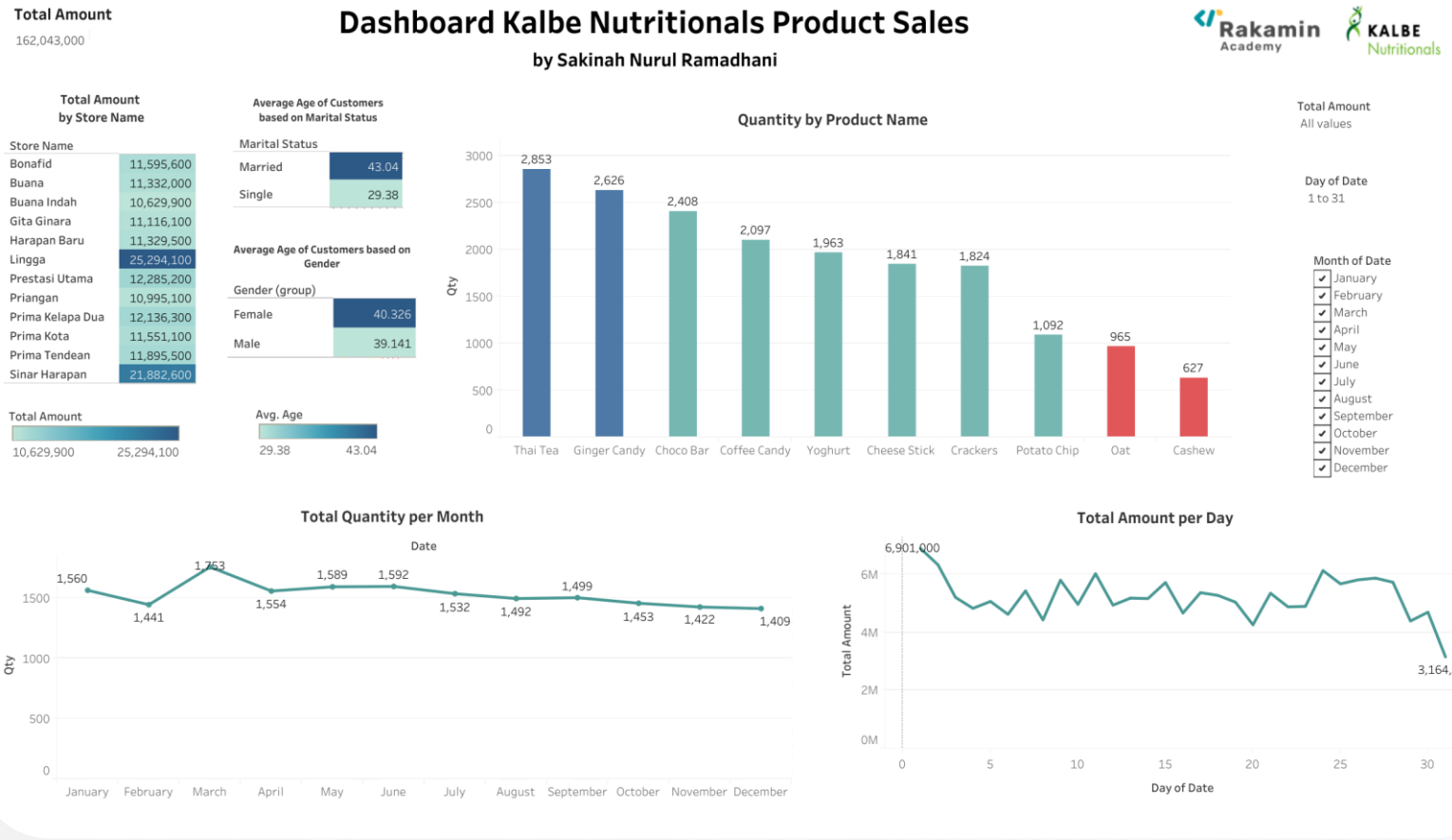
	Product Name	total_sales_amount
1	Cheese Stick	27,615,000

Insight:

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

Klik link disini

Tableau



Klik link disini

Machine Learning Regression (Time Series)

Rakamin Academy

Data Cleaning :

Data Cleaning

```
#Data cleaning pada df_customer
df_customer['Income'] = df_customer['Income'].replace('[,]','.', regex=True).astype('float')

#Data cleaning pada df_store
df_store['Latitude'] = df_store['Latitude'].replace('[,]','.', regex=True).astype('float')
df_store['Longitude'] = df_store['Longitude'].replace('[,]','.', regex=True).astype('float')

#Data cleaning pada df_transaction
df_transaction['Date'] = pd.to_datetime(df_transaction['Date'])

#Data cleaning Menghapus baris Null pada Marital Status
df_customer.isnull().sum()
df_customer = df_customer.dropna(subset=['Marital Status'])
```

Merge Data :

Merge Data

```
df_merge = pd.merge(df_transaction, df_customer, on=['CustomerID'])
df_merge = pd.merge(df_merge, df_product.drop(columns=['Price']), on=['ProductID'])
df_merge = pd.merge(df_merge, df_store, on=['StoreID'])
```

Merge Data Forecast

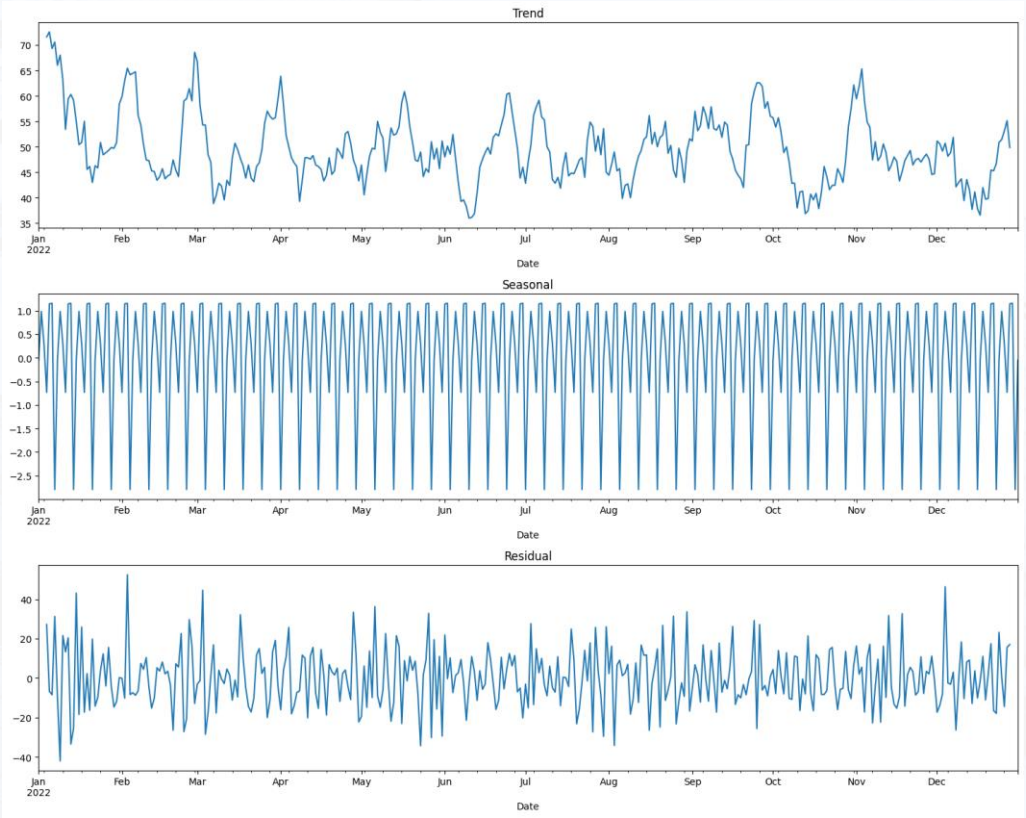
```
#Forecast data
df_fore = df_merge[['Date','Qty']]
df_fore= df_fore.groupby('Date')[['Qty']].sum()
df_fore.head(3)
```

Splitting Data :

Split Data Train and Test

```
#Split data train and test
df_train = df_fore.iloc[:-31]
df_test = df_fore.iloc[-31:]
```

Klik link disini



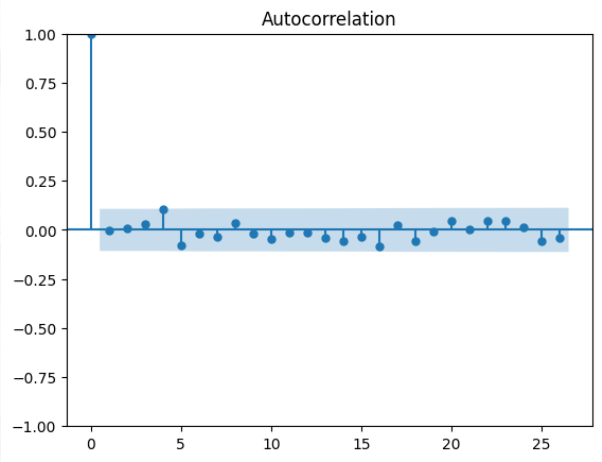
Grafik Trend, Seasonal, Residual

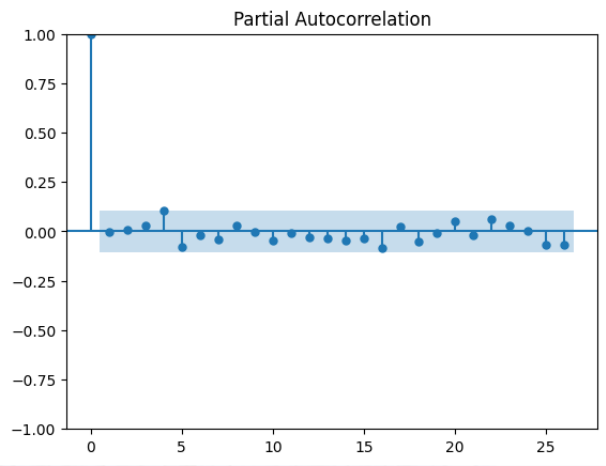
```
decompose = seasonal_decompose(df_fore)

fig,ax = plt.subplots(3,1,figsize=(15,12))
decompose.trend.plot(ax=ax[0])
ax[0].set_title('Trend')
decompose.seasonal.plot(ax=ax[1])
ax[1].set_title('Seasonal')
decompose.resid.plot(ax=ax[2])
ax[2].set_title('Residual')

plt.tight_layout()
plt.show()
```

Klik link disini






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

```
p-value: 2.308849689470346e-30
```

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.

Klik link disini

MODEL ARIMA

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

1. 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]

```
#Auto-fit ARIMA metrics

mae = mean_absolute_error(df_test, forecast_auto)
mape = mean_absolute_percentage_error(df_test, forecast_auto)
rmse = np.sqrt(mean_squared_error(df_test, forecast_auto))

print(f'mae - auto: {round(mae,4)}')
print(f'mape - auto: {round(mape,4)}')
print(f'rmse - auto: {round(rmse,4)}')
```

```
mae - auto: 0.3059
mape - auto: 0.0855
rmse - auto: 0.3821
```

Klik link disini

2. Manual Parameter Tuning

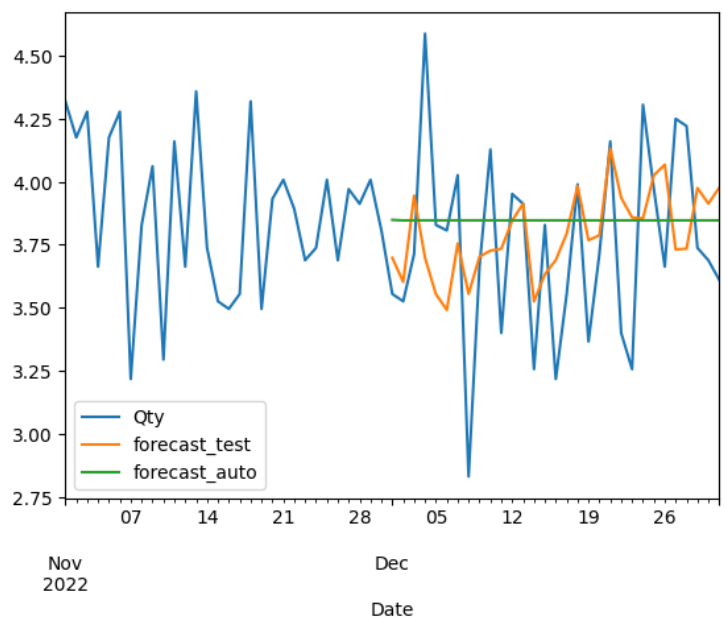
	param	mae	mape	rsme
2	(70, 2, 1)	0.309134	0.082776	0.395608
1	(60, 2, 1)	0.328327	0.088083	0.415141
0	(50, 2, 1)	0.380785	0.105433	0.455044

```
#Manual parameter tuning metrics

mae = mean_absolute_error(df_test, forecast_test)
mape = mean_absolute_percentage_error(df_test, forecast_test)
rmse = np.sqrt(mean_squared_error(df_test, forecast_test))

print(f'mae - manual: {round(mae,4)}')
print(f'mape - manual: {round(mape,4)}')
print(f'rmse - manual: {round(rmse,4)}')
```

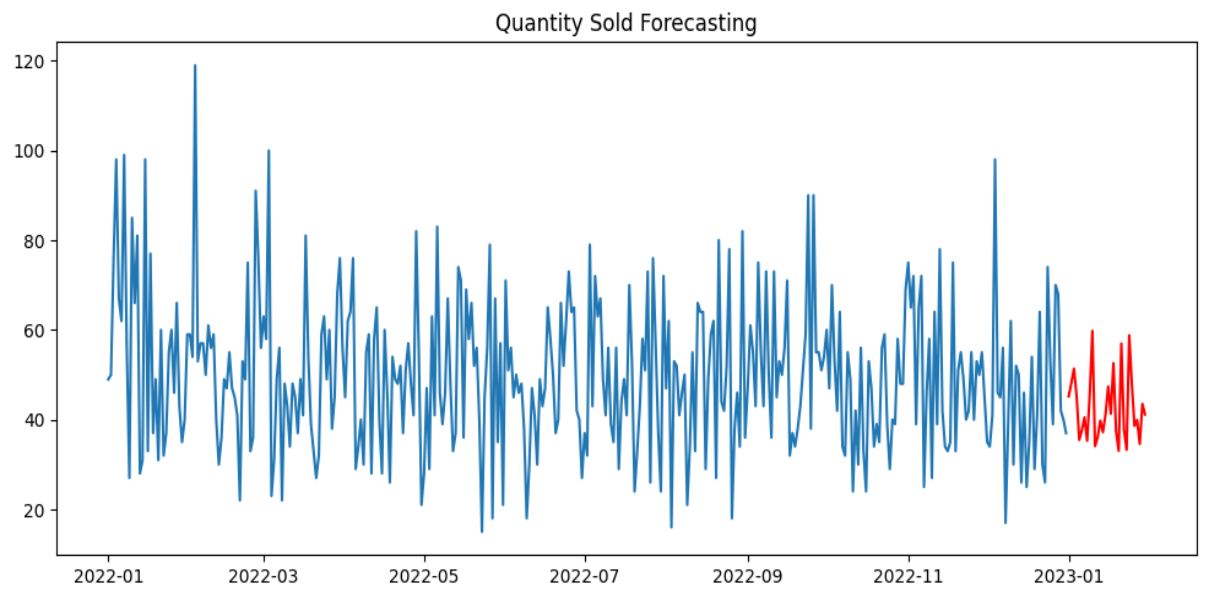
```
mae - manual: 0.3016
mape - manual: 0.0823
rmse - manual: 0.3694
```



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

Klik link disini

Forecasting Overall Quantity



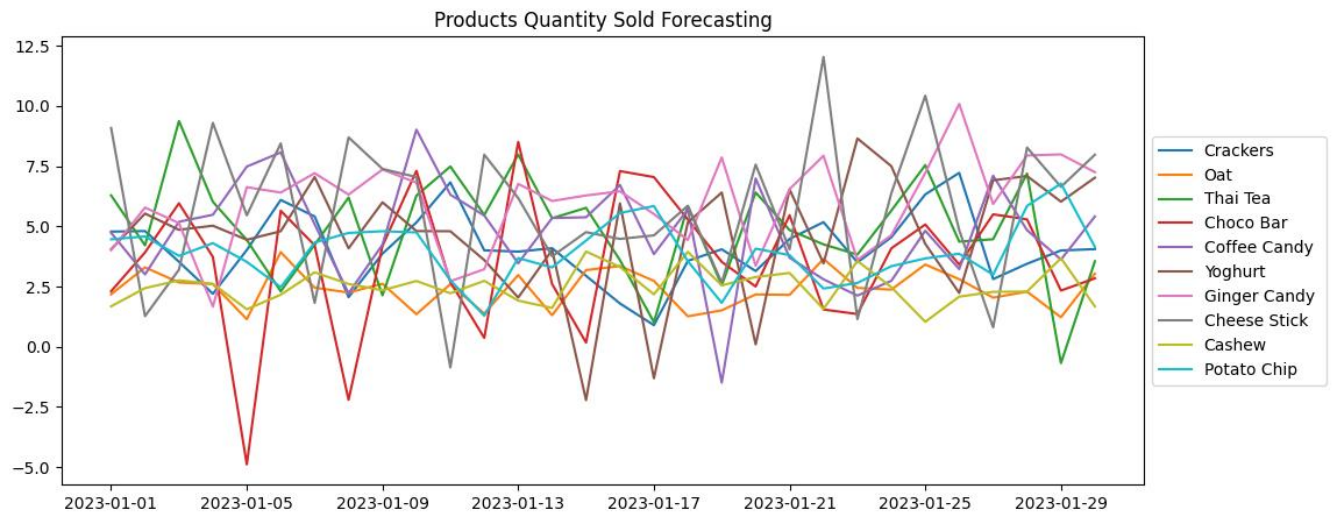

```
forecast.mean()
```

```
42.564966762859505
```

Observasi:
Dari forecasting diatas dapat disimpulkan bahwa untuk rata-rata quantity penjualan bulan depan adalah sekitar 43 pcs per harinya.

Klik link disini

Forecasting Each Product




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

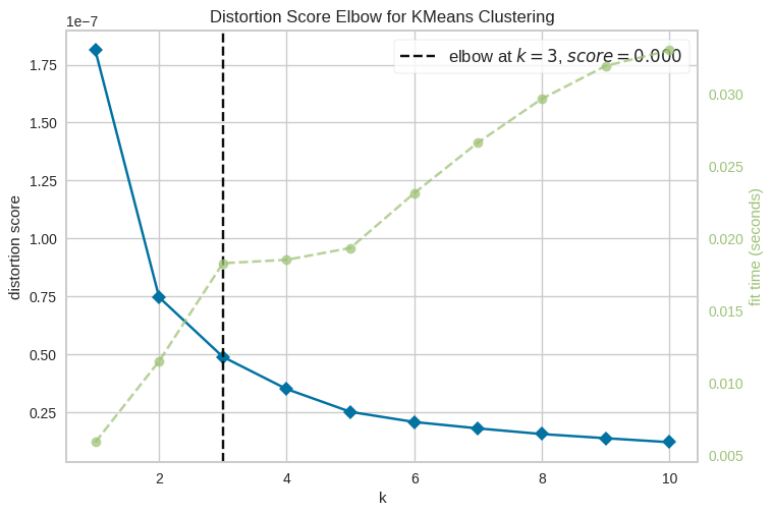
```
Crackers          4.0
Oat               2.0
Thai Tea          5.0
Choco Bar         4.0
Coffee Candy      5.0
Yoghurt           5.0
Ginger Candy      6.0
Cheese Stick      6.0
Cashew            2.0
Potato Chip       4.0
Name: mean, dtype: float64
```

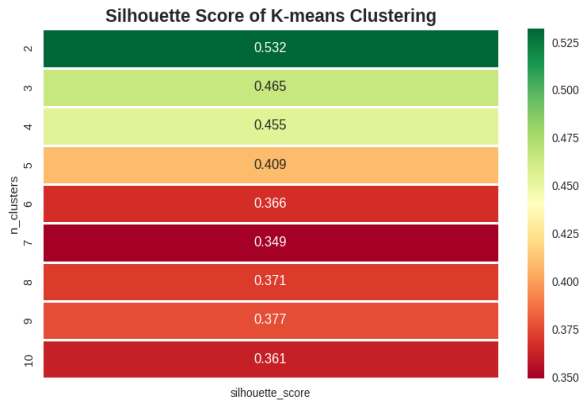
Observasi:

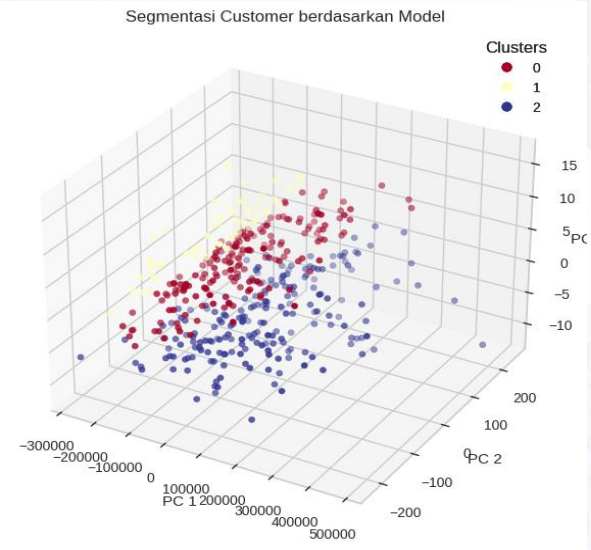
Dari data forecasting product diatas kita bisa mendapatkan perkiraan rata-rata quantity produk yang terjual setiap harinya. Jenis produk yang paling banyak terjual adalah Cheese Stick dan Ginger Candy yaitu sebanyak 6 pcs per harinya. Sedangkan jenis produk yang kurang terjual adalah Oat dan Cashew yaitu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

Klik link disini

Machine Learning Clustering (K-Means)





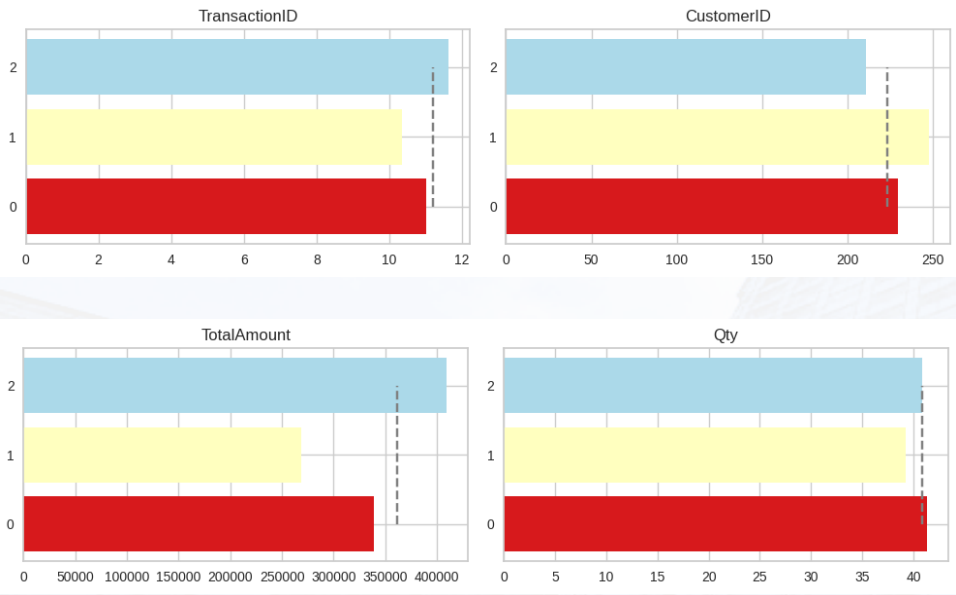


Observasi:

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

Klik link disini

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.

Klik link disini

Business Recommendation:

Rakamin Academy

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.

Rakamin Academy

You can see more about my **project** on **GitHub**

Rakamin Academy

You can see my **video presentation** about this project in **Google Drive**

Thank You