Predict Customer Personality to boost marketing campaign by using Machine Learning

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Rakamin Academy

Created by:
Rajudin Ali Slamet
rajudinalislamet@gmail.com
www.linkedin.com/in/rajudin-ali-s

Industrial Engineering graduate with strong analytical and problem-solving skills, specializing in data analytics and data-driven decision making. Skilled in data cleaning, exploratory data analysis (EDA), and visualization using Python, SQL, Excel, and Power BI. Experienced in transforming raw data into actionable insights through academic projects and professional training. Passionate about leveraging data to optimize processes, improve business efficiency, and support strategic decision-making.

Overview

“Sebuah perusahaan dapat berkembang dengan pesat saat mengetahui perilaku customer personality nya, sehingga dapat memberikan layanan serta manfaat lebih baik kepada customers yang berpotensi menjadi loyal customers. Dengan mengolah data historical marketing campaign guna menaikkan performa dan menyasar customers yang tepat agar dapat bertransaksi di platform perusahaan, dari insight data tersebut fokus kita adalah membuat sebuah model prediksi kluster sehingga memudahkan perusahaan dalam membuat keputusan ”

Link Jupyter Notebook

Overview

Background

Sebuah perusahaan dapat mencapai pertumbuhan yang signifikan dengan memahami perilaku dan kepribadian pelanggan. Pengetahuan ini memungkinkan bisnis untuk memberikan layanan yang lebih baik dan nilai yang lebih tinggi kepada pelanggan, terutama mereka yang memiliki potensi besar untuk menjadi pelanggan loyal. Dengan memanfaatkan data historis dari kampanye pemasaran, perusahaan dapat meningkatkan kinerja, menargetkan segmen pelanggan yang tepat, dan meningkatkan jumlah transaksi di platformnya

Goals

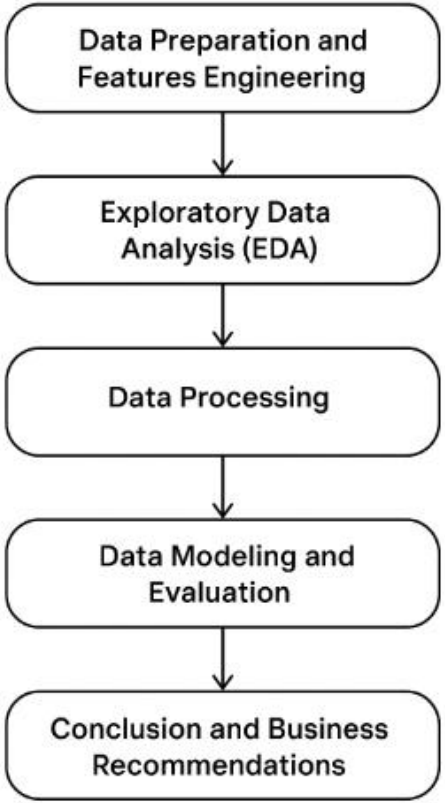
Tujuan dari proyek ini adalah memanfaatkan data pelanggan dan kampanye untuk mengembangkan wawasan yang dapat memperkuat strategi bisnis. Secara lebih spesifik, proyek ini bertujuan untuk membangun model *predictive clustering* yang memungkinkan perusahaan mengidentifikasi, mengelompokkan, dan melibatkan pelanggan dengan lebih efektif.

Objectives

- Menganalisis data historis kampanye pemasaran untuk menemukan pola utama perilaku pelanggan.
- Mengidentifikasi dan mengelompokkan pelanggan dengan potensi tinggi untuk loyalitas dan keterlibatan.
- Mengembangkan model predictive clustering untuk meningkatkan efektivitas penargetan pelanggan.
- Memberikan wawasan berbasis data yang mendukung pengambilan keputusan strategis dan meningkatkan kinerja perusahaan.

Link Jupyter Notebook

Workflow Project



Dataset Information

```
df.info()
```

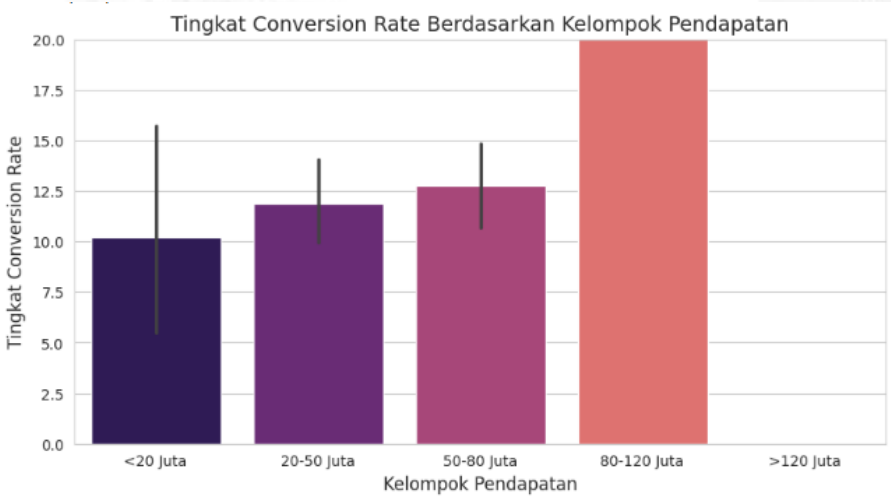
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 30 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   Unnamed: 0           2240 non-null   int64  
 1   ID                   2240 non-null   int64  
 2   Year_Birth           2240 non-null   int64  
 3   Education            2240 non-null   object 
 4   Marital_Status       2240 non-null   object 
 5   Income               2216 non-null   float64
 6   Kidhome              2240 non-null   int64  
 7   Teenhome             2240 non-null   int64  
 8   Dt_Customer          2240 non-null   object 
 9   Recency              2240 non-null   int64  
 10  MntCoke              2240 non-null   int64  
 11  MntFruits            2240 non-null   int64  
 12  MntMeatProducts      2240 non-null   int64  
 13  MntFishProducts      2240 non-null   int64  
 14  MntSweetProducts     2240 non-null   int64  
 15  MntGoldProds         2240 non-null   int64  
 16  NumDealsPurchases    2240 non-null   int64  
 17  NumWebPurchases      2240 non-null   int64  
 18  NumCatalogPurchases  2240 non-null   int64  
 19  NumStorePurchases    2240 non-null   int64  
 20  NumWebVisitsMonth    2240 non-null   int64  
 21  AcceptedCmp3         2240 non-null   int64  
 22  AcceptedCmp4         2240 non-null   int64  
 23  AcceptedCmp5         2240 non-null   int64  
 24  AcceptedCmp1         2240 non-null   int64  
 25  AcceptedCmp2         2240 non-null   int64  
 26  Complain             2240 non-null   int64  
 27  Z_CostContact        2240 non-null   int64  
 28  Z_Revenue            2240 non-null   int64  
 29  Response             2240 non-null   int64  
dtypes: float64(1), int64(26), object(3)
memory usage: 525.1+ KB
```

- **Dataset Overview**: Terdiri dari 2.240 entri dengan 30 kolom, sebagian besar berupa data numerik (26 integer, 1 float) dan 3 kategorikal.
- **Demographics & Behavior**: Berisi data demografis (Year_Birth, Education, Marital_Status, Income, Kidhome, Teenhome) serta perilaku pelanggan (Recency, pengeluaran produk, saluran pembelian, keluhan, respon kampanye).
- **Data Issues**: Terdapat nilai yang hilang pada kolom Income, kolom yang tidak relevan (Unnamed: 0, Z_CostContact, Z_Revenue), serta ketidaksesuaian tipe data (kolom Dt_Customer perlu dikonversi menjadi datetime, dan variabel kategorikal perlu dilakukan encoding)

EXPLORATORY DATA ANALYSIS

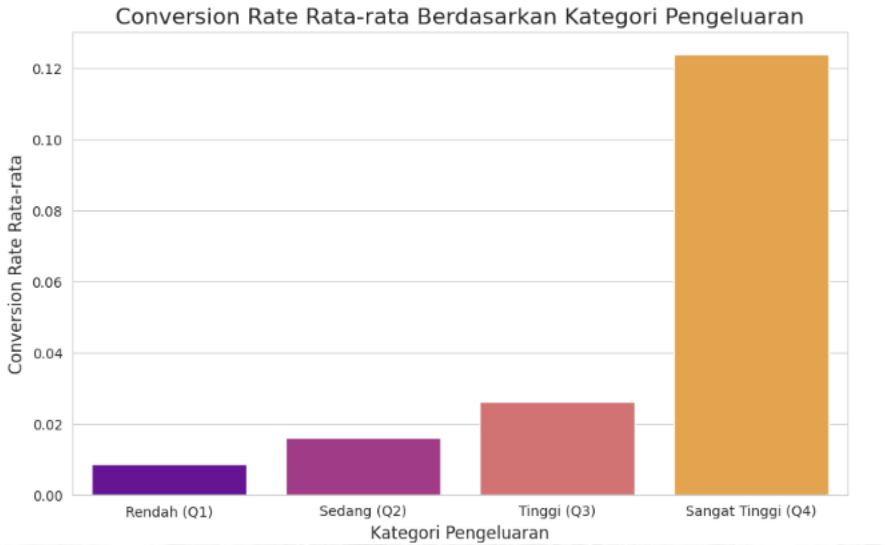
Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Based on Income



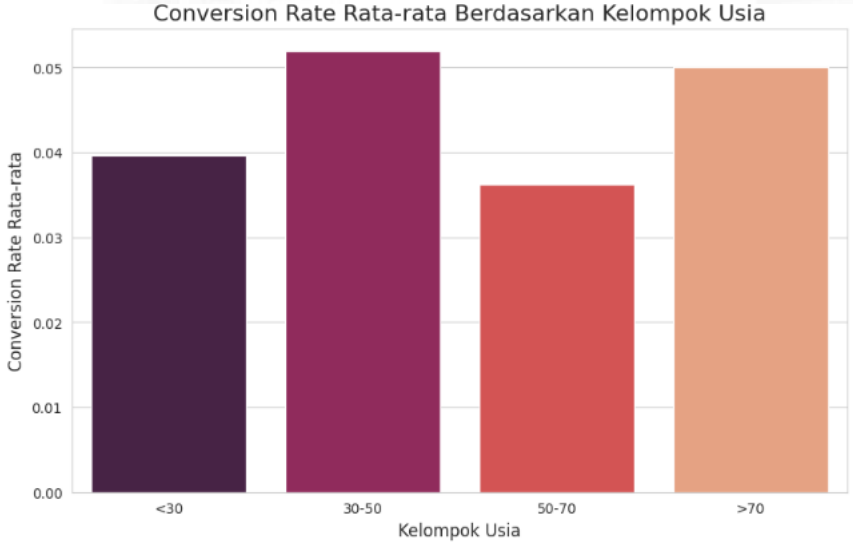
- Pelanggan dengan pendapatan antara **80–120 juta** menunjukkan tingkat konversi tertinggi (sekitar 18–19%), secara signifikan lebih tinggi dibandingkan kelompok berpendapatan rendah (<20 juta IDR dengan sekitar 10%).
- Hal ini menunjukkan bahwa pelanggan dengan pendapatan menengah ke atas lebih dominan dibandingkan kelompok lainnya

Conversion Rate Based on Spending



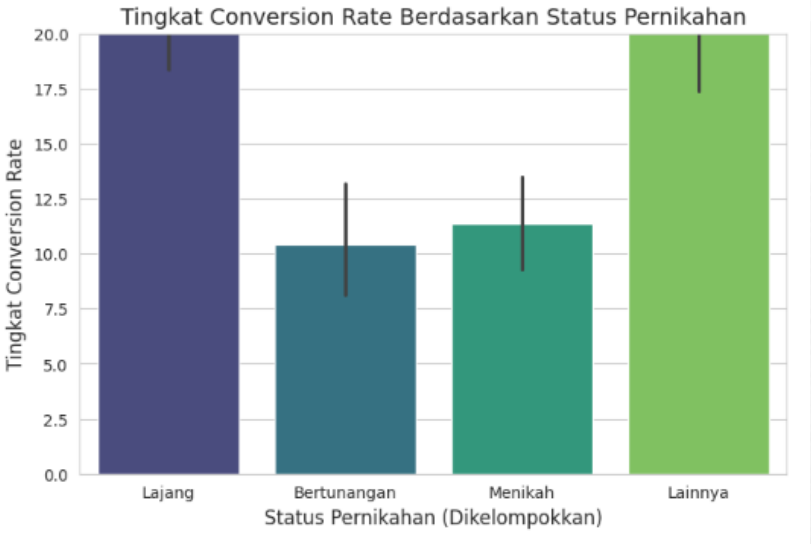
- Tingkat konversi meningkat secara signifikan seiring dengan kategori pengeluaran. Pelanggan dalam kelompok pengeluaran **"Sangat Tinggi" (Q4) memiliki tingkat konversi lebih dari 12%,**
- Hal jauh lebih tinggi dibandingkan kelompok dengan pengeluaran rendah (<3%), ini menunjukkan bahwa pelanggan dengan pengeluaran tinggi merupakan segmen paling bernilai dalam mendorong konversi.

Conversion Rate Based on Age



- Kelompok **usia 30–50** tahun menunjukkan tingkat konversi tertinggi (5,2%), menjadikannya segmen paling potensial.
- Kelompok usia >70 tahun juga menunjukkan kinerja yang kuat (5%), menandakan adanya peluang signifikan pada segmen lansia.
- Kelompok usia 50–70 tahun memiliki tingkat konversi terendah (3,6%), yang mungkin memerlukan pendekatan pemasaran khusus untuk meningkatkan konversi.Sementara itu,
- kelompok usia <30 tahun berada pada tingkat sedang (4%), menunjukkan potensi namun tidak setinggi kelompok usia yang lebih tua.

Conversion Rate Based on Marital Status



- Kategori **Single dan "Other"** mencatat tingkat konversi tertinggi (~19–20%), jauh di atas kategori Engaged (~10,5%) dan Married (~11%).
- Hal ini dapat mengindikasikan bahwa individu lajang atau mereka yang berada di luar kategori pernikahan lebih responsif terhadap konversi dibandingkan dengan mereka yang memiliki komitmen keluarga yang stabil.

DATA CLEANING AND PROCESSING

Data Cleaning dan Processing

Missing Value

```
df.isna().sum()[df.isna().sum() > 10]

0
Income 24

dtype: int64
```

→

```
#Handle Missing Value
df["Income"] = df["Income"].fillna(df["Income"].median())
```

→

```
df.isna().sum()[df.isna().sum() > 10]

0

dtype: int64
```

Missing Value digantikan dengan nilai median karena ketiga kolom ini memiliki distribusi data yang sangat condong (skewed). Penggunaan median sebagai nilai imputasi lebih tepat dibandingkan mean, yang lebih sensitif terhadap keberadaan outlier.

Duplicate Data

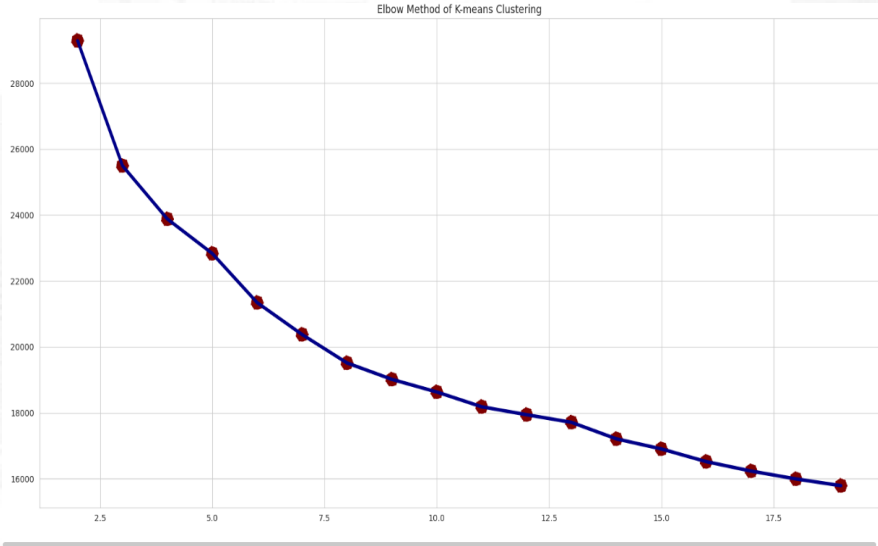
```
duplicate_rows = df.duplicated().sum()
print(f"\nJumlah baris duplikat: {duplicate_rows}")

Jumlah baris duplikat: 0
```

Tidak ada data duplikat sehingg bisa dilanjutkan ke proses selanjutnya

Untuk selengkapnya, dapat melihat jupyter notebook disini

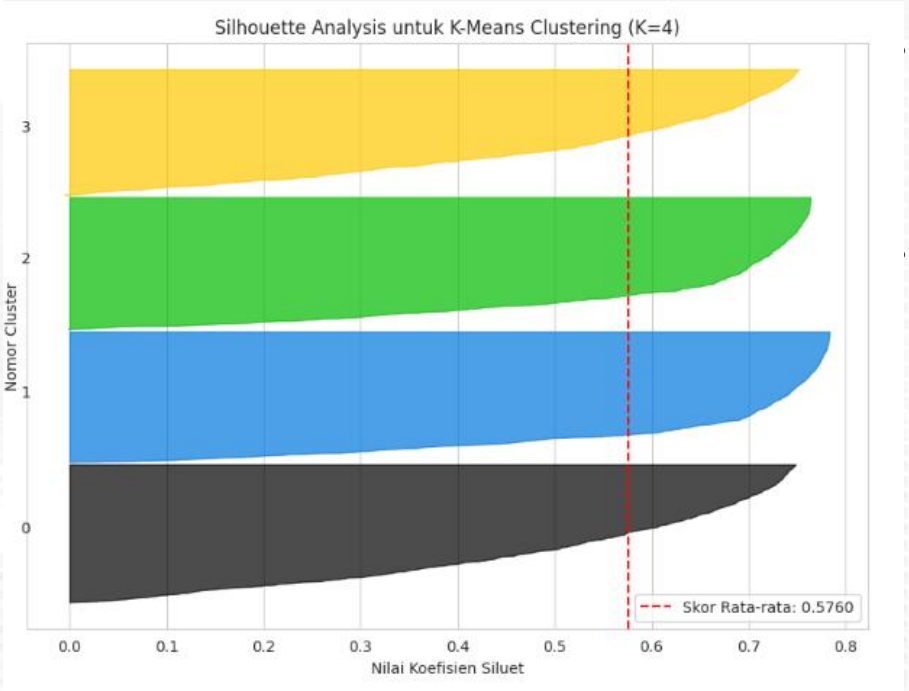
DATA MODELING

Elbow Method dan Silhouette Score



- Terjadi penurunan tajam antara K=1 hingga K=3, yang berarti penambahan jumlah klaster pada tahap ini secara signifikan meningkatkan kualitas pengelompokan.
- Antara K=4 hingga K=6, kurva masih menurun tetapi dengan laju yang lebih lambat.
- Mulai dari K=7 ke atas, kurva menjadi datar, menunjukkan penurunan manfaat (diminishing returns) saat menambah lebih banyak klaster.
- **Titik “elbow” terlihat muncul di sekitar K=4.**

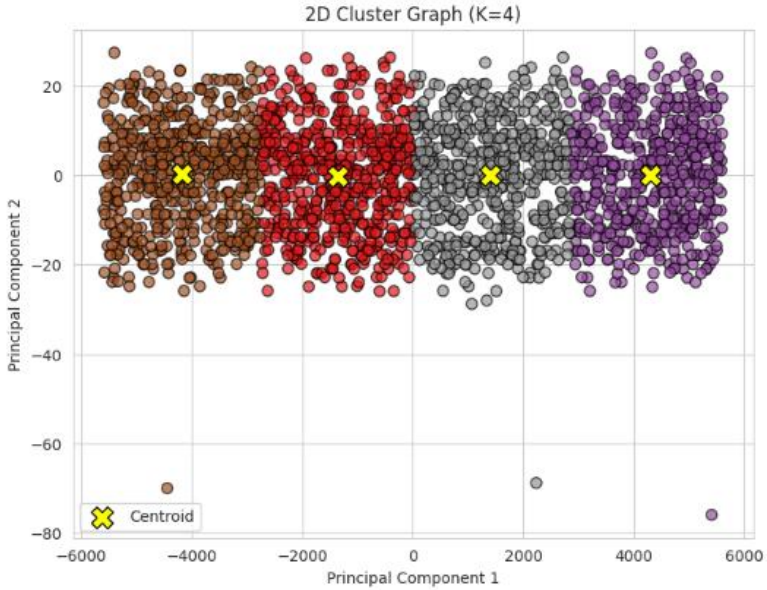
Silhouette Analysis



Good clustering quality – Nilai average silhouette score sebesar 0,576 menunjukkan bahwa klaster yang terbentuk tergolong cukup kuat, terpisah dengan baik, dan bermakna.

Cluster consistency – ebagian besar klaster (biru, hijau, kuning) menunjukkan nilai silhouette yang tinggi secara konsisten, sementara Klaster 0 (hitam) memiliki beberapa nilai yang lebih rendah, menandakan adanya beberapa titik batas (borderline points).

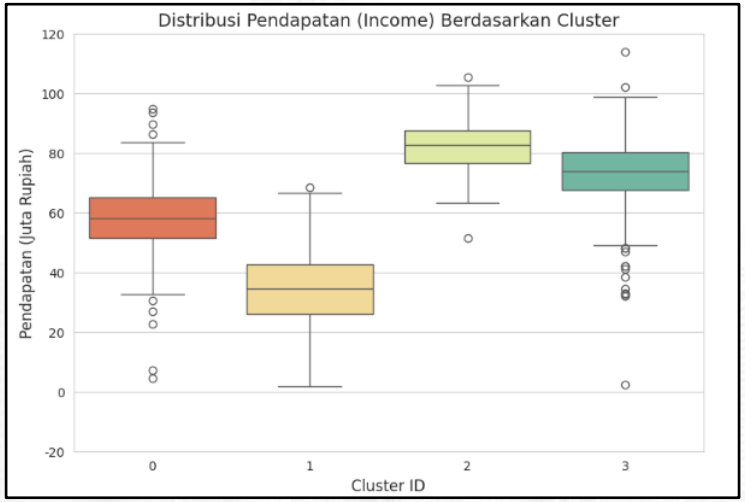
Cluster Graph 2D

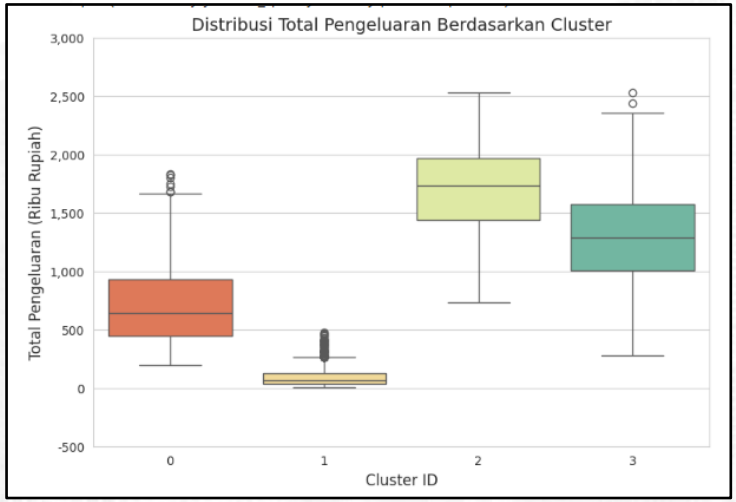


- **Pemisahan klaster yang jelas** – Empat klaster (cokelat, merah, abu-abu, ungu) terpisah dengan baik pada komponen utama pertama (sumbu x), menunjukkan bahwa data telah terbagi secara efektif ke dalam kelompok yang berbeda.
- **Distribusi yang seimbang** – Setiap klaster memiliki ukuran dan kepadatan yang relatif serupa, menandakan bahwa K=4 memberikan segmentasi yang seimbang tanpa ada satu klaster yang mendominasi.
- **Titik pusat sebagai representasi klaster** – Tanda X berwarna kuning menunjukkan titik pusat (centroid) dari masing-masing klaster, yang terletak di tengah kelompoknya. Hal ini berarti titik pusat tersebut merupakan representasi yang baik dari karakteristik setiap klaster.

EDA AFTER CLUSTERING

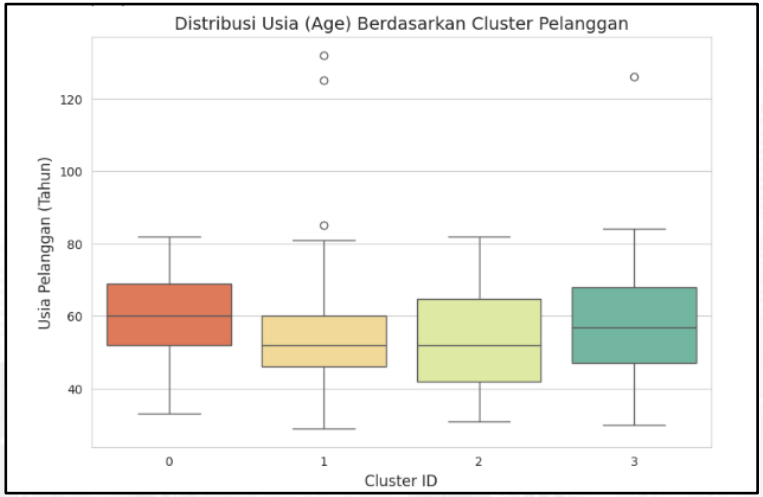
Visualisasi Analisis dari EDA dengan Menggunakan *Hasil cluster*





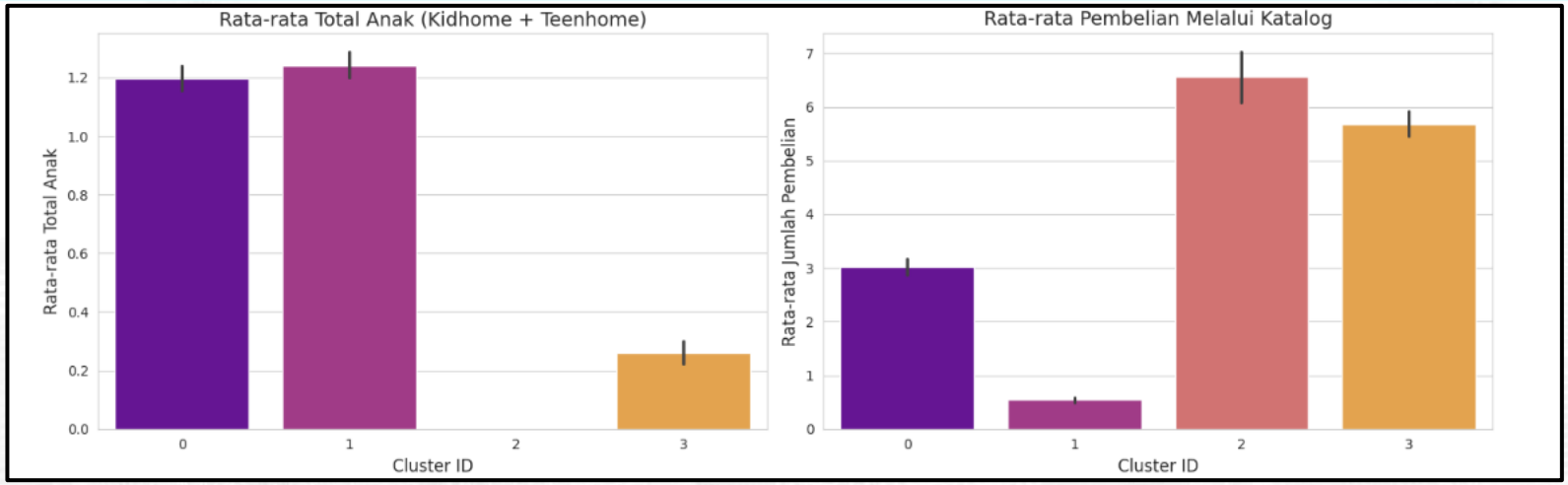
Cluster ID	Nama Segmen	Rata-rata Pendapatan (Income)	Rata-rata Pengeluaran (Total Spent)
2	High-Value, Highly Responsive	Rp 87.705.000	Rp 1.365.000
3	Budget-Conscious, Older	Rp 52.042.000	Rp 526.000
0	Mid-Value, Sleepers	Rp 52.190.000	Rp 319.000
1	Low-Value, Low-Involvement	Rp 24.019.000	Rp 48.000

Visualisasi Analisis dari EDA dengan Menggunakan *Hasil cluster*



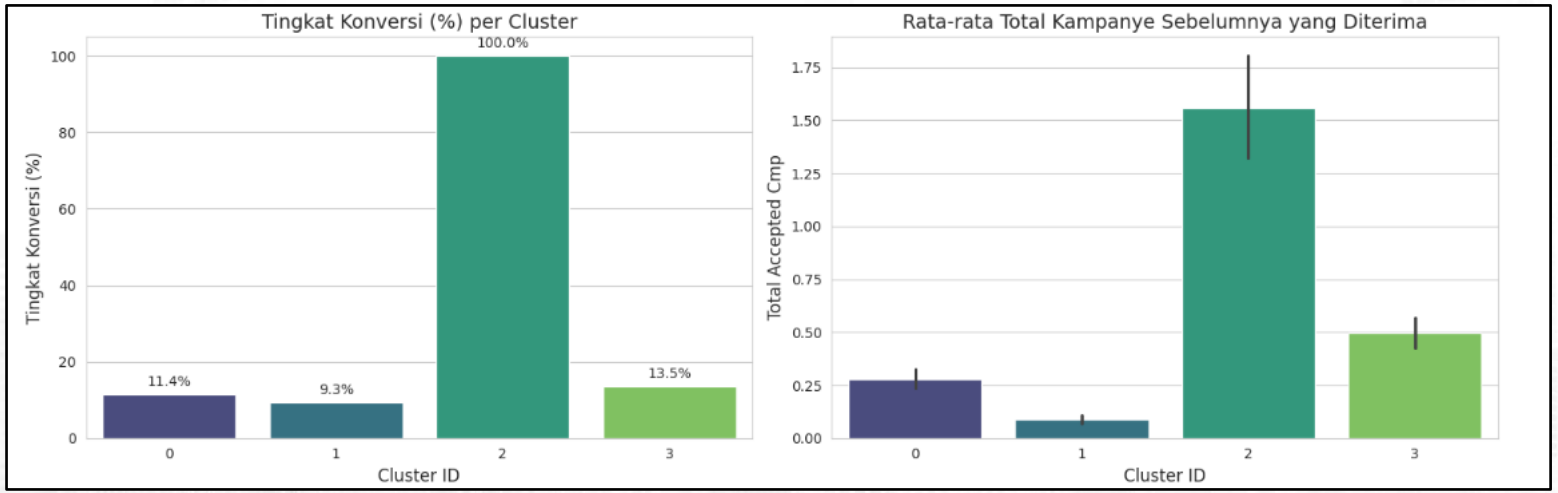
Cluster ID	Nama Segmen	Rentang Usia	Keterangan
2	High-Value, Highly Responsive	39 - 68 tahun	Usia **Puncak Karir**
3	Budget-Conscious, Older	>56 tahun	Segmen **TERTUA**. Pelanggan mendekati/sudah pensiun.
0	Mid-Value, Sleepers	31 - 67 tahun	Rentang Usia **Paruh Baya** yang luas, tidak terlalu membedakan segmen ini.
1	Low-Value, Low-Involvement	30 - 68 tahun	Rentang Usia **Terluas/Termuda** (termasuk pelanggan paling muda).

Visualisasi Analisis dari EDA dengan Menggunakan *Hasil cluster*



Cluster ID	Nama Segmen	Rata-rata Total Anak (Kidhome + Teenhome)	Rata-rata Pembelian Katalog (NumCatalogPurchases)
2	**High-Value, Highly Responsive**	**≈0,44 (TERENDAH)**	**≈6,97 (TERTINGGI)**
3	**Budget-Conscious, Older**	≈0,81 (Menengah)	≈2,42 (Menengah)
0	**Mid-Value, Sleepers**	≈1,02 (Tinggi)	≈2,34 (Menengah)
1	**Low-Value, Low-Involvement**	**≈1,32 (TERTINGGI)**	≈0,40 (TERENDAH)

Visualisasi Analisis dari EDA dengan Menggunakan *Hasil cluster*



Cluster	Nama Segmen	Tingkat Konversi (Response Rate)	Rata-rata Kampanye yang Diterima Sebelumnya
2	**High-Value, Highly Responsive**	**≈22,7% (TERTINGGI)**	**≈1,76** (TERTINGGI)
3	**Budget-Conscious, Older**	**≈10,6% (Tinggi Kedua)**	**≈0,72** (Tinggi Kedua)
0	**Mid-Value, Sleepers**	**≈3,9% (Terendah)**	≈0,42 (Rendah)
1	**Low-Value, Low-Involvement**	≈2,4% (Terendah)	≈0,16 (TERENDAH)

Bussiness Recomendations

Cluster	Nama Segmen	Karakteristik Kunci Kepribadian (Profil)	Rekomendasi Strategi
1	Low-Value	Berpenghasilan dan berbelanja terendah. Responsivitas sangat rendah (< 5%).. Pelanggan ini memiliki kebutuhan pengeluaran rumah tangga yang tinggi karena presentase jumlah anak yang tinggi. Memiliki rentang usia yang paling lebar dan bervariasi (mulai dari awal 30-an)	Promosikan produk kebutuhan dasar atau produk yang berorientasi keluarga/anak dengan harga sangat kompetitif. Fokus pada volume daripada margin awal.
2	High-Value	Berpenghasilan tinggi daya belanja tertinggi, usia matang (39-68 tahun), sangat responsif terhadap kampanye (CR > 20%), dan anak sedikit. Menyukai saluran pembelian Katalog	Targetkan dengan produk Premium/Mewah (High-End). Pertahankan loyalitas melalui program VIP dan personalisasi layanan maksimal.

Bussiness Recomendations

Cluster	Nama Segmen	Karakteristik Kunci Kepribadian (Profil)	Rekomendasi Strategi
3	**Budget-Conscious (Sadar anggaran dan cenderung belanja saat diskon)**	Berpenghasilan dan berbelanja menengah-rendah, usia cenderung lebih tua (56-68 tahun), tetapi cukup responsif terhadap penawaran (CR ≈ 10%). Mereka sensitif terhadap harga.	Tawarkan kupon, diskon, dan *deal* yang menekankan penghematan. Gunakan saluran pemasaran tradisional (katalog, email) karena faktor usia.
0	**Mid-Value**	Berpenghasilan dan berbelanja menengah yang stabil. Tingkat Konversi Rendah (< 5%), dan Recency TERTINGGI (sudah lama tidak aktif). Memiliki rentang usia yang paling lebar dan bervariasi (mulai dari awal 30-an) sama seperti cluster 0	Kirimkan kampanye atau penawaran yang mendesak (*scarcity*) seperti penawaran yang berbatas waktu dan nilai tinggi (misalnya, diskon 30% hanya untuk 48 jam).