Predict Customer Personality to boost marketing campaign by using Machine Learning

Created by:
Danu Satria Wiratama
danusatria06@gmail.com
https://www.linkedin.com/in/danusatria/

"I am a computer science graduate from Satya Wacana Christian University with a concentration in data science. I completed an online bootcamp at Rakamin Academy with a topic of data science, which further enhance my skills in data analysis and machine learning. I have mastered several skills such as SQL, python, data visualization, and additional proficiency in machine learning for further modeling and analysis.. "

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

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 ”

Feature Engineering

- Membuat kolom ‘Conversion_Rate’ berdasarkan ‘Total_Transaction’ / ‘NumWebVisitsMonth’

```
# Hitung total jumlah pembelian
df['Total_Transaction'] = df['NumDealsPurchases'] + df['NumWebPurchases'] + df['NumCatalogPurchases'] + df['NumStorePurchases']
```

```
# Hitung conversion rate
df['Conversion_Rate'] = df['Total_Transaction'] / df['NumWebVisitsMonth']

# Handle infinite values (jika ada pembagian dengan nol)
df['Conversion_Rate'].replace([np.inf, -np.inf], 0, inplace=True)
```

- Membuat kolom total campaign, dan total pengeluaran.

```
# Hitung total campain
df['Total_Acc_Cmp'] = df['AcceptedCmp1']+df['AcceptedCmp2']+df['AcceptedCmp3']+df['AcceptedCmp4']+df['AcceptedCmp5']

# Hitung total pengeluaran
df['Total_Spending'] = df['MntCoke'] + df['MntFruits'] + df['MntMeatProducts'] + df['MntFishProducts'] + df['MntSweetProducts'] + df['MntGoldProds']
```

- Membuat kolom usia berdasarkan kolom ‘Year_Birth’, dan mengelompokan berdasarkan kategori.

```
#Menambahkan kolom usia
df['Age'] = 2024 - df['Year_Birth']
df.sample(5)
```

```
# mengkategorikan berdasarkan usia
age_list = []
for i in df['Age']:
  if i < 36:
    age_list.append('Young Adult')
  elif i >= 36 and i < 56:
    age_list.append('Middle-Aged Adults')
  else:
    age_list.append('Seniors')

df['Age_Category'] = age_list

df["Age_Category"].value_counts()
```

Age_Category	
Middle-Aged Adults	1132
Seniors	994
Young Adult	90

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Engineering

- Membuat kolom durasi membership

```
# Mengubah kolom 'Dt_Customer' ke Datetime
df['Dt_Customer'] = pd.to_datetime(df['Dt_Customer'], format='%d-%m-%Y')

# Membuat kolom durasi membership
df['Membership_Duration'] = 2024 - df['Dt_Customer'].dt.year

df.value_counts('Membership_Duration')
```

Membership_Duration	
11	1173
10	553
12	490

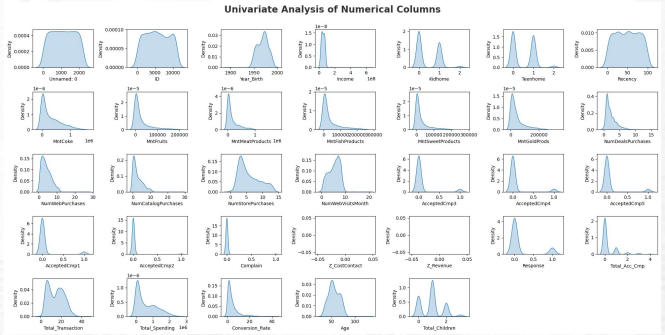
- Membuat kolom total anak

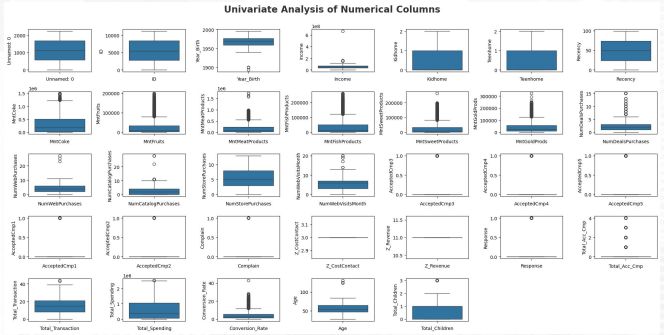
```
# total anak
df['Total_Children'] = df['Kidhome'] + df['Teenhome']

df['Total_Children'].value_counts()
```

Total_Children	
1	1117
0	633
2	416
3	50

Exploratory Data Analysis

- Univariate Analysis Numerical Columns



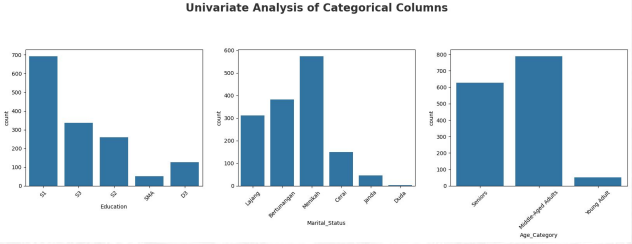


Terdapat outlier pada beberapa kolom diantaranya :

['Year_birth', 'Income', 'MntCoke', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumWebVisitsMonth', 'Total_Transaction', 'Total_Spending', 'Conversion_Rate', 'Age']. Yang selanjutnya kolom kolom tersebut akan dilakukan penghapusan outliers.

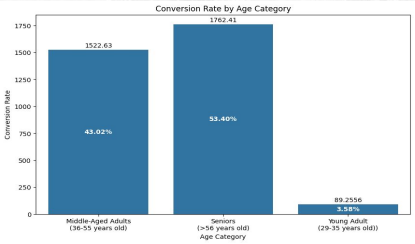
Exploratory Data Analysis

- Univariate Analysis Categorical Columns



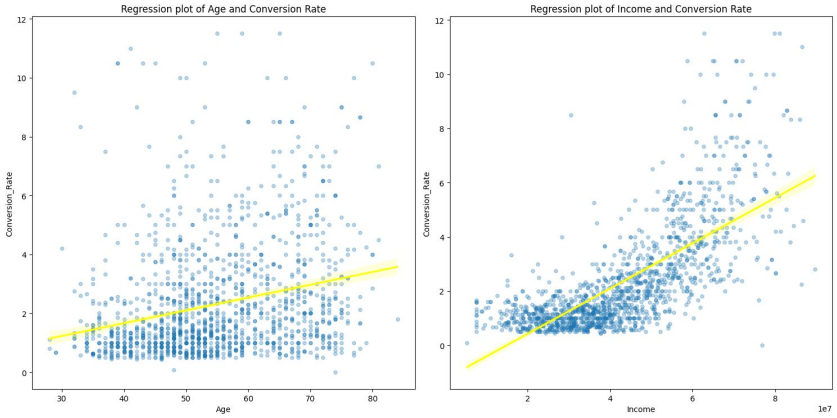
Customer didominasi dengan education S1, status pernikahan menikah, dan kategori usia Middle-Aged Adults.

- Bivariate Analysis



Berdasarkan plot diatas terdapat hubungan yang signifikan antara conversion rate dengan age category. Dimana kategori young adult memilki convertion rate yang sangat rendah (3.58%) jika dibandingkan dengan kategori Midle-Aged Adults (43.02%) dan Senior(53.40%).

Exploratory Data Analysis



Conclusion :

1. **Age dan conversion rate** :

   Terlihat korelasi positif lemah antara umur dan tingkat konversi. Semakin bertambah usia, tingkat konversi cenderung meningkat, tetapi hubungan ini tidak terlalu kuat.Ini bisa berarti bahwa usia yang lebih tua sedikit lebih cenderung untuk terlibat atau berkonversi, mungkin karena stabilitas finansial atau preferensi belanja tertentu.

2. **Income dan conversion rate** :

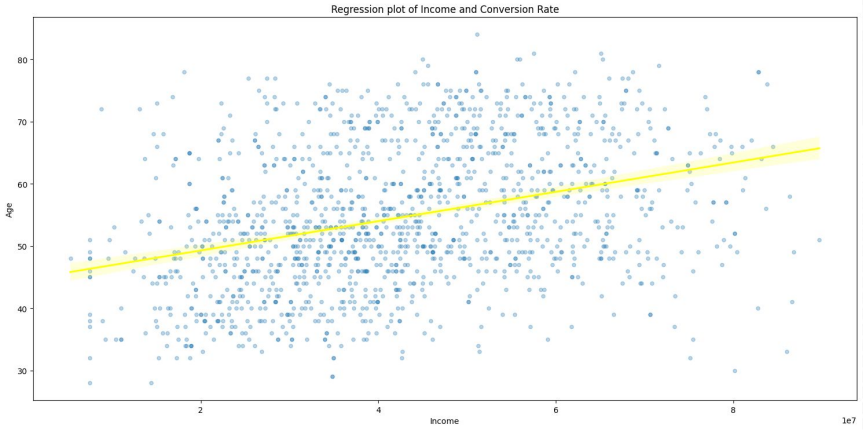
   Hubungan korelasi positif kuat antara pendapatan (Income) dan tingkat konversi. Hal ini menunjukan bahwa semakin tinggi pendapatan, tingkat konversi cenderung meningkat. Ini menunjukkan bahwa pelanggan dengan pendapatan lebih tinggi lebih cenderung terlibat atau melakukan pembelian.

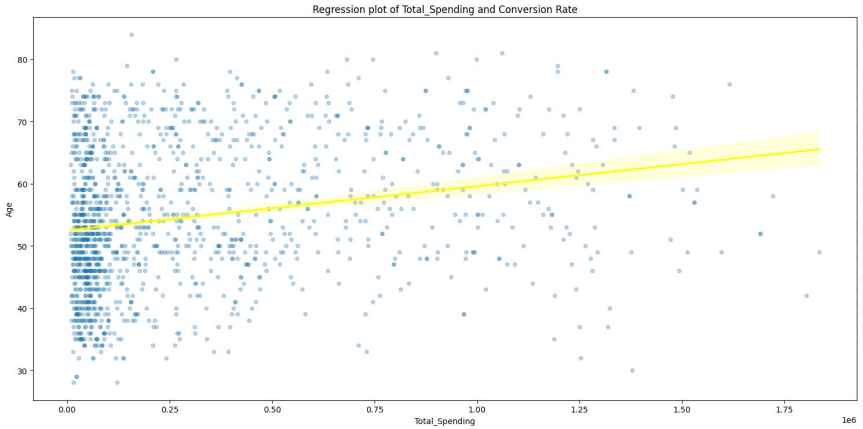
3. **Total spending dan conversion rate** :

   Korelasi positif yang kuat antara pengeluaran total (Total Spending) dan tingkat konversi. Hal ini menunjukan bahwa pelanggan dengan pengeluaran lebih tinggi juga menunjukkan tingkat konversi yang lebih tinggi. Hubungan ini bisa menunjukkan bahwa pelanggan dengan keterlibatan yang lebih besar pada kampanye pemasaran memiliki pengeluaran yang lebih besar, sehingga meningkatkan Conversion Rate.

Strategi pemasaran bisa difokuskan pada pelanggan dengan pendapatan dan pengeluaran tinggi, karena mereka memiliki peluang konversi lebih besar.

Exploratory Data Analysis





Conclusion :

1. **Age dan Income** :

   Terdapat hubungan positif antara pendapatan (Income) dan usia (Age). Korelasi ini tidak sangat kuat, mengindikasikan faktor lain seperti tingkat pendidikan atau pekerjaan juga berpengaruh.

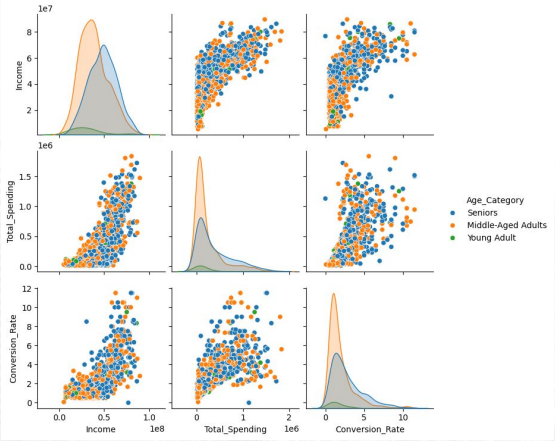
2. **Age dan Total spending** :

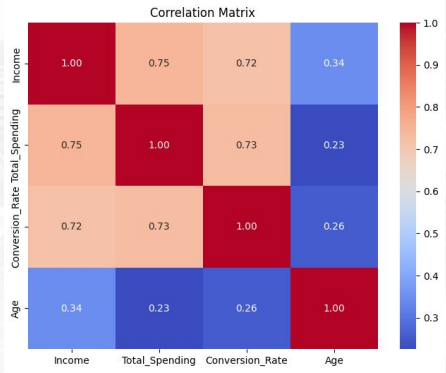
   Ada hubungan positif lemah antara total pengeluaran (Total Spending) dan usia (Age). Usia yang lebih tua cenderung dihubungkan dengan pengeluaran yang sedikit lebih tinggi. Namun, hubungan ini lebih lemah dibandingkan dengan "Income vs Age", yang menunjukkan bahwa pengeluaran tidak selalu sebanding dengan usia.

Strategi pemasaran bisa diarahkan pada segmen usia yang lebih tua (50 ke atas) dengan pendapatan tinggi, karena mereka memiliki potensi untuk pengeluaran lebih besar.

Exploratory Data Analysis

- Multivariate Analysis





Conclusion :

- Middle-Aged Adults adalah kelompok dominan dalam pendapatan, pengeluaran, dan tingkat konversi.
- Ada hubungan positif yang kuat antara pendapatan, pengeluaran, dan tingkat konversi, tetapi dengan variasi yang lebih tinggi di beberapa kategori usia.
- Young Adults dan Seniors cenderung memiliki nilai yang lebih rendah dibandingkan Middle-Aged Adults.

Suggestion :

1. Fokus pada Middle_Age Adult sebagai segmen utama :
   - Buat kampanye pemasaran yang ditargetkan pada segmen ini.
   - Tawarkan produk atau layanan premium sesuai daya beli mereka.
   - Perkuat loyalitas dengan program member, diskon eksklusif, penawaran berbasis langganan.
2. Membuat strategi khusus untuk segmen senior :
   - Kembangkan layanan yang memprioritaskan kenyamanan dan kebutuhan mereka(misalnya, layanan berbasis kesehatan atau kenyamanan).
   - Gunakan pendekatan pemasaran yang lebih personal, seperti konsultasi langsung atau pemasaran offline.
3. Optimalkan hubungan pendapatan, tingkat pembelian, dan konversi :
   - Buat program upselling dan cross-selling untuk mendorong pengeluaran pelanggan.
   - Tawarkan insentif berbasis pengeluaran (contoh: "Dapatkan diskon jika belanja lebih dari X").
   - Gunakan data analitik untuk mengidentifikasi pelanggan bernilai tinggi dan tingkatkan konversi mereka.

Data Cleaning & Preprocessing

- Tidak ada data yang duplikat.

```
[33] df_cln = df.copy()

     df_cln.duplicated().sum()

     0
```

- Tidak ada data yang kosong.

```
df_cln.isnull().sum()
```

	0
Unnamed: 0	0
ID	0
Year_Birth	0
Education	0
Marital_Status	0
Income	0
Kidhome	0
Teenhome	0
Dt_Customer	0
Recency	0
MntCoke	0
MntFruits	0
MntMeatProducts	0
MntFishProducts	0
MntSweetProducts	0
MntGoldProds	0
NumDealsPurchases	0
NumWebPurchases	0
NumCatalogPurchases	0
NumStorePurchases	0
NumWebVisitsMonth	0
AcceptedCmp3	0
AcceptedCmp4	0
AcceptedCmp5	0
AcceptedCmp1	0
AcceptedCmp2	0
Complain	0
Z_CostContact	0
Z_Revenue	0
Response	0
Total_Acc_Cmp	0
Total_Spending	0
Total_Transaction	0
Conversion_Rate	0
Age	0
Age_Category	0
Membership_Duration	0
Total_Children	0

Feature encoding & Feature standardisation

Rakamin Academy

- Melakukan encoding pada kolom education.

```
# Feature Encoding
df['Education'].value_counts()

edu = {'SMA' : 0, 'D3' : 1, 'S1' : 2, 'S2' : 3, 'S3' : 4}

df_cln['Education'] = df_cln['Education'].map(edu)
```

- Menghapus kolom kolom yang tidak diperlukan.

```
df_clean = df_cln.copy()
df_clean = df_clean.drop(['Unnamed: 0', 'ID', 'Year_Birth', 'Education', 'Marital_Status', 'Age_Category', 'Dt_Customer'], axis=1)
```

- Melakukan scaling feature

```
from sklearn.preprocessing import MinMaxScaler, StandardScaler

scaler = StandardScaler()
df_clean= pd.DataFrame(scaler.fit_transform(df_clean), columns=df_clean.columns)
```

Data Modeling

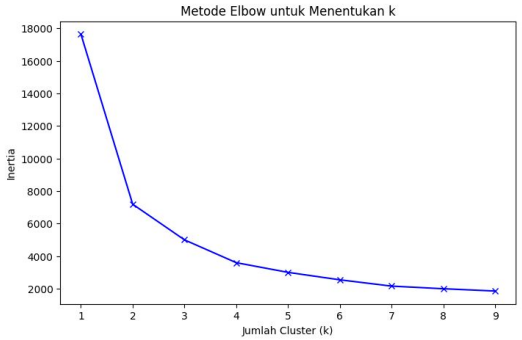
- Sebelum melakukan modelling dilakukan teknik PCA untuk mereduksi dimensi data sambil tetap mempertahankan sebanyak mungkin informasi penting.

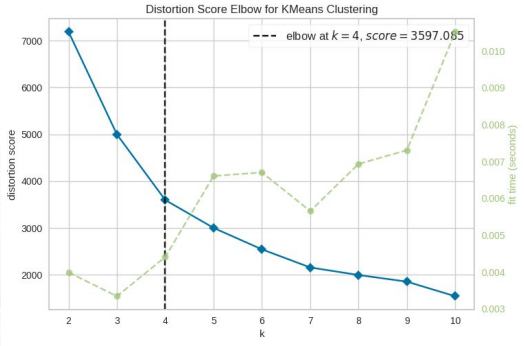
```
from sklearn.decomposition import PCA

# Inisialisasi PCA dengan jumlah komponen yang diinginkan
pca = PCA(n_components=2)
pca.fit(df_model)

# Transformasikan data ke ruang fitur baru
pca_data = pca.transform(df_model)

pca_df = pd.DataFrame(data=pca_data, columns=['PC1', 'PC2'])
```

- Melakukan modeling dengan metode K-Means Clustering

```
from sklearn.cluster import KMeans

distortions = []
K = range(1, 10)  # Jumlah cluster dari 1 hingga 10

for k in K:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(pca_df)
    distortions.append(kmeans.inertia_)
```

- Visualisasi dengan elbow method

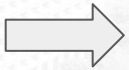




Berdasarkan Distortion Score dan Elbow metode didapatkan jumlah cluster terbaik adalah 4.

- Clustering dengan menggunakan K-Means

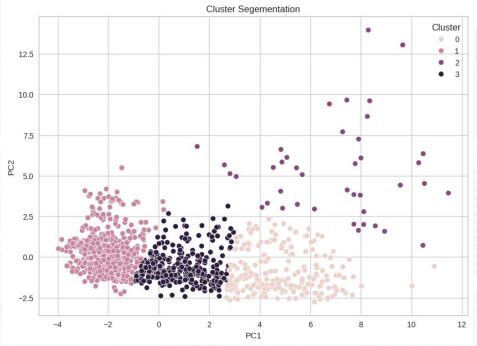
```
kmeans = KMeans(n_clusters=4, random_state=42)
kmeans.fit(pca_df.values)

# Menambahkan label cluster ke dataset
pca_df['Cluster'] = kmeans.labels_
```



	PC1	PC2	Cluster
0	-2.190879	-0.990998	1
1	5.380192	-1.023377	0
2	-2.357484	0.365173	1
3	2.368557	-1.226815	3
4	3.952642	-1.360936	0

Data Modeling

- Visualisasi hasil segmentasi.



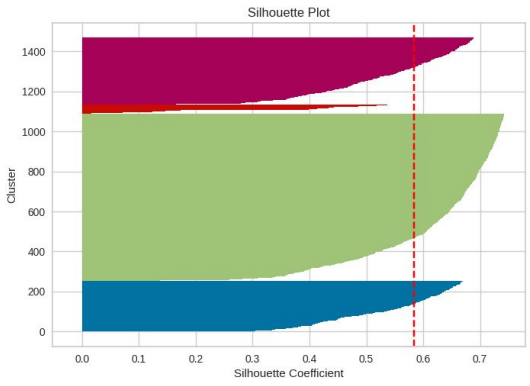
- Evaluasi dengan silhouette score.

```
from sklearn.metrics import silhouette_score

silhouette_avg = silhouette_score(pca_df, kmeans.labels_)
print("Silhouette Score:", silhouette_avg)
```

```
Silhouette Score: 0.5832898419362629
```

Hasil evaluasi dengan silhouette score menunjukan hasil 0.58. Nilai ini terbilang cukup baik.

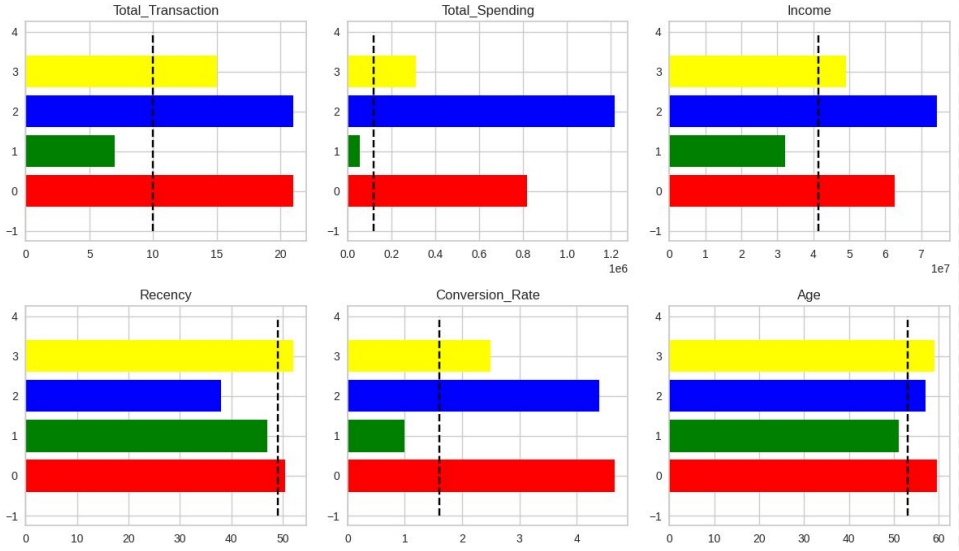
- Silhouette Plot



Kesimpulan

1. Berdasarkan Silhouette Score, jumlah cluster terbaik yang direkomendasikan adalah 4.
2. Hasil Clustering: Sebagian besar data dikelompokkan dengan baik (silhouette coefficient positif), meskipun ada beberapa data dengan nilai negatif.
3. Rata-rata Silhouette: Garis merah menunjukkan hasil clustering cukup baik (mendekati atau melebihi 0.5).

Customer Personality Analysis for Marketing Retargeting



Berdasarkan hasil clustering, diketahui bahwa :

Cluster 0

- Rata rata transaksi di cluster 0 dan cluster 2 berada di angka yang sama diangka 21 transaksi.
- Total spending cukup sedang diangka Rp 816.000/bulan.
- Income berada di nilai yang cukup tinggi diangka Rp 62.551.500/tahun.
- Convension rate tertinggi, senilai 5%.
- Mayoritas usia 60 tahun.

Cluster 1

- Total transaksi terendah, diangka 7 transaksi.
- Total spending terendah, diangka Rp 54.000/bualn.
- Income terendah diangka Rp 32.233.000/tahun.
- Conversion rate terendah, senilai 1%.
- Mayoritas usia 51 tahun.

Cluster 2

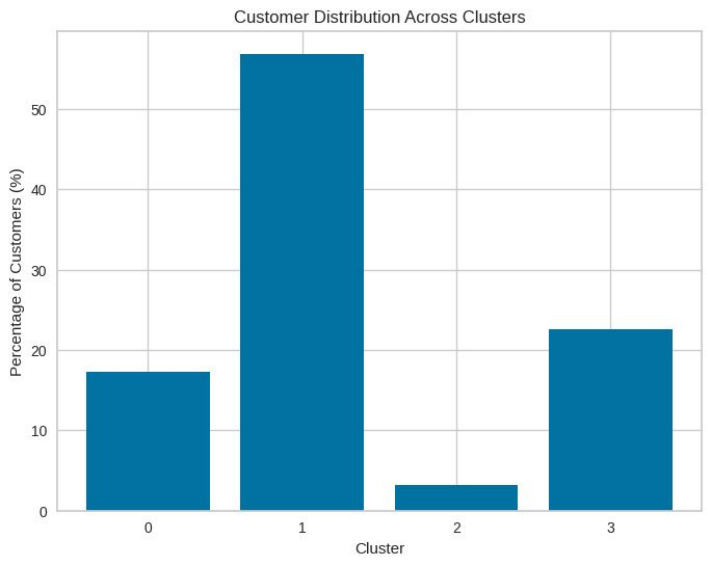
- Rata rata transaksi di cluser 2 sama dengan cluster 0 diangka 21 transaksi.
- Total spending dan total income tertinggi, dengan total spending senilai Rp 1.217.000/bulan dan income Rp 74.290.000/tahun.
- Convesion rate cukup tinggi diangka 4%.
- Mayoritas usia 57 tahun.

Cluster 3

- Rata rata transaksi cukup sedang diangka 15 transaksi.
- Total spending cukup sedang cinderung rendah diangka Rp 311.000/bulan.
- Total Income cukup sedang cinderung rendah diangka Rp 49.176.500/tahun.
- Conversion rate cinderung rendah diangka 2%.
- Mayoritas usia 59 tahun.

Customer Personality Analysis for Marketing Retargeting

- Persentase setiap cluster

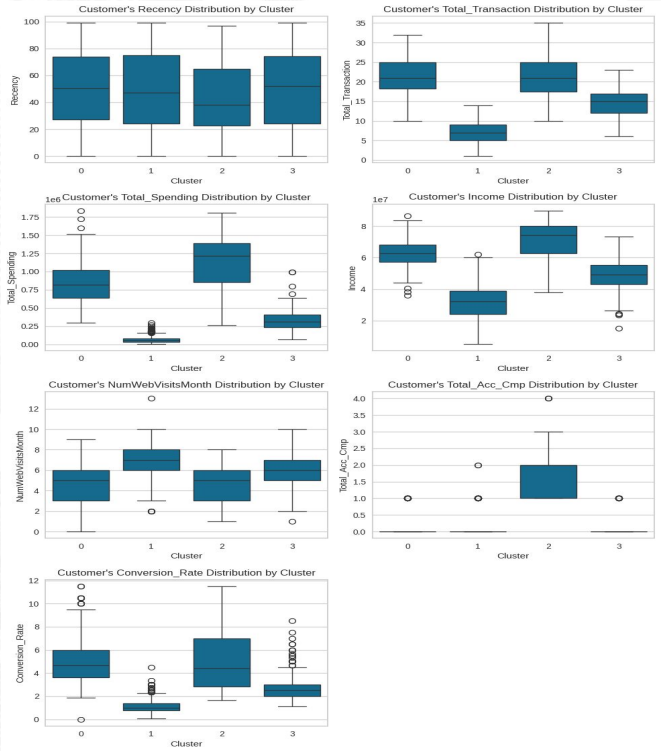


Lebih dari 50% customer berada di cluster low, dengan total transaksi yang rendah, total spending rendah, serta pendapatan rendah. Meskipun demikian, perusahaan harus memperhatikan cluster ini karena tingginya populasi.

Cluster 0 dan 2 termasuk dalam high cluster, namun populasinya rendah sehingga perusahaan harus mempertimbangkan strategi pemasaran untuk mempertahankan loyalitas cluster ini.

Customer Personality Analysis for Marketing Retargeting

- Univariate Analysis



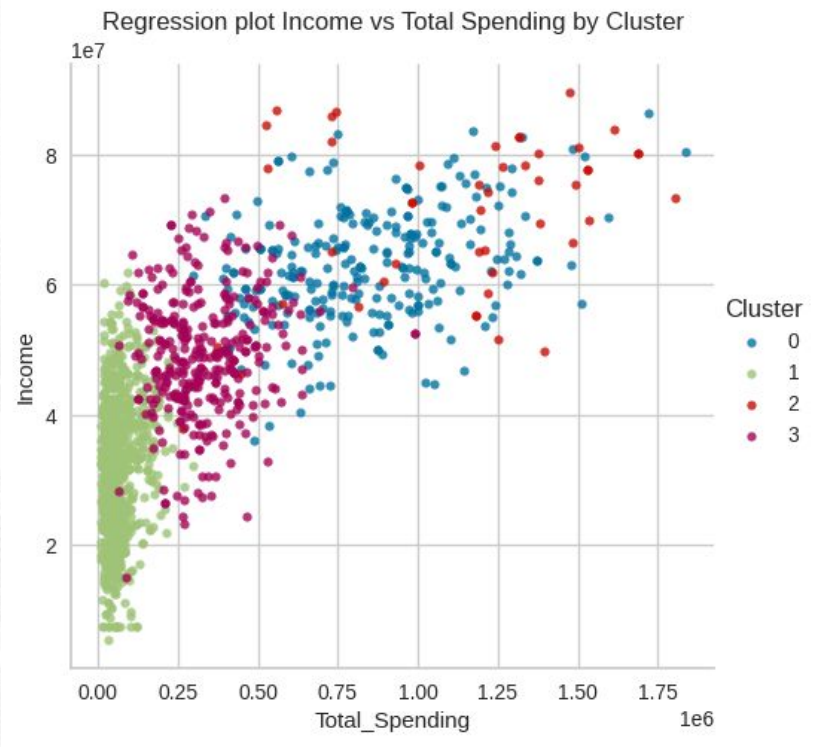
Cluster 1 memiliki NumVisitsMonth yang paling tinggi namun dengan TotalSpending terendah. Hal ini menunujkan bahwa mereka sangat sering mengunjungi website namun tidak melakukan transaksi. Perusahaan harus memperhatikan fenomena ini mengingat cluster ini adalah cluster dengan populasi yang paling besar sehingga perusahaan perlu mengembangkan strategi untuk menarik perhatian mereka.

Cluster 2 memiliki tingkat accept campaign yang paling besar, cluster ini juga memilki total spending yang paling besar. Hal ini menunjukan bahwa mereka sangat sensitif terhadap campain yang ditawarkan namun terlihat juga bahwa mereka memiliki NumVisitsMonth yang rendah dengan kata lain mereka jarang membuka website. Dengan ini perusahaan dapat memaksimalkan promosi dengan mengubungi melalui email, whatsapp, atau media lainnya agar mampu menjangkau segmen ini.

Cluster 3 memiliki spending, visit month, dan conversion rate yang semua berada di nilai rata-rata. Pelanggan ini berpotensi untuk ditingkatkan kontribusinya (potential growth) dengan cara analisis tambahan untuk memahami apa yang mendorong perilaku belanja mereka seperti preferensi produk, memberikan penawaran yang relevan secara personal seperti memberi diskon tambahan untuk pembelian berulang, dan memberikan program loyalitas yang menarik.

Customer Personality Analysis for Marketing Retargeting

- Multivariate Analysis



- Secara keseluruhan, terdapat korelasi positif antara pendapatan dan pengeluaran
- Semakin tinggi pendapatan, semakin tinggi pengeluaran.
- Terlihat bahwa High cluster 0 dan 3 cenderung berada dalam satu kelompok, yaitu dalam kategori high customer. Dengan mengetahui pola korelasi ini, perusahaan dapat mengoptimalkan strategi pemasaran dan penawaran produk mereka.

Business Recommendation

Customer Personality Analysis for Marketing Retargeting

Rakamin Academy

Dari hasil analisis, kita dapat mengenali karakteristik atau profil pelanggan berdasarkan kluster yang ada. Memahami karakteristik ini sangat penting untuk merancang strategi pemasaran yang lebih tepat sasaran. Dengan mengetahui preferensi, kebutuhan, dan perilaku konsumen di setiap kluster, perusahaan dapat menciptakan kampanye yang lebih relevan dan menarik bagi masing-masing kelompok pelanggan.

- **Cluster 1 (Low Customer - Low Transaction, Low Spending, Low Income):**
  - Memiliki proporsi pelanggan terbesar (lebih dari 50%).
  - Mayoritas pelanggan berada dalam kategori ini, yang berarti sebagian besar basis pelanggan memiliki kemampuan dan aktivitas ekonomi yang rendah.
  - Insight Bisnis:
    - Strategi pemasaran dan promosi yang terjangkau dapat diutamakan untuk kelompok ini.
    - Berikan edukasi atau penawaran khusus untuk meningkatkan transaksi dan loyalitas pelanggan.
    - Identifikasi peluang untuk meningkatkan pendapatan mereka (misalnya, melalui layanan kredit kecil atau bundling produk).

- **Cluster 0 (High Customer 1 - High Transaction, High Spending, High Income):**
  - Proporsi pelanggan ini lebih kecil dibandingkan Cluster 1, tetapi lebih signifikan dibandingkan Cluster 2.
  - Insight Bisnis:
    - Fokus pada layanan premium atau eksklusif untuk mempertahankan pelanggan ini.
    - Peluang untuk cross-selling atau up-selling produk dengan margin lebih tinggi.
    - Perhatikan loyalitas mereka melalui program VIP atau reward khusus.

- **Cluster 2 (High Customer 2 - High Transaction, High Spending, High Income):**
  - Proporsi pelanggan sangat kecil (di bawah 5%).
  - Insight Bisnis:
    - Kelompok ini merupakan segmen potensial dengan nilai tinggi; perlu dilakukan analisis lebih mendalam untuk memahami kebutuhan mereka.
    - Investasi dalam mempertahankan pelanggan ini sangat penting karena kontribusi mereka terhadap pendapatan perusahaan bisa signifikan.

- **Cluster 3 (Moderate Customer - Moderate Transaction, Moderate Spending, Moderate Income):**
  - Memiliki proporsi pelanggan yang cukup besar setelah Cluster 1 (sekitar 20%-30%).
  - Insight Bisnis:
    - Segmen ini dapat diarahkan untuk meningkatkan pengeluaran mereka melalui promosi produk atau layanan yang sesuai dengan daya beli mereka.
    - Berikan pengalaman yang lebih personal untuk menarik mereka menuju Cluster 0 atau 2.

Terimakasih