Predict Customer Personality to boost marketing campaign by using Machine Learning

Created by:
Limatan Luviar
Limatan.junior@gmail.com
Limatan Luviar

Graduated from Universitas Sriwijaya majoring in Informatics Management in 2018 and interned at a government institution. Enrolling in and completing the Data Science Bootcamp at Rakamin Academy in 2024 was driven by my interest in data analysis. Currently, I am actively seeking job opportunities and have completed various data-driven projects at Rakamin. My focus is on Data Analyst roles, where I can utilize my analytical skills and business understanding to make a significant impact

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 ”

Exploratory Data Analysis

Dari dataset yang diberikan kesimpulan sebagai berikut:

1. Dataset terdiri dari 2240 baris dan 30 kolom.
2. Terdapat beberapa kolom dengan data yang beragam, seperti 'Education' (5 unique values), 'Marital_Status' (6 unique values), dan 'Dt_Customer' (663 unique values).
3. Kolom 'Income' memiliki 24% data yang hilang (NaN).
4. Mayoritas pelanggan memiliki 0 anak di rumah (Kidhome) dan 0 remaja di rumah (Teenhome).
5. Mayoritas pelanggan memiliki 0 komplain (Complain).
6. Kolom 'Z_CostContact' dan 'Z_Revenue' memiliki nilai yang konstan (3 dan 11), sehingga mungkin tidak memberikan informasi yang berguna untuk analisis.
7. Kolom 'Unnamed: 0' dan 'ID' dianggap tidak relevan dan dapat dibuang
8. Kolom 'Response' merupakan target variabel yang akan dijadikan acuan dalam model prediksi atau analisis selanjutnya.

	Feature	Data Type	Null	Null (%)	Unique	Unique Sample
0	Unnamed: 0	int64	0	0.000000	2240	[0, 1, 2, 3, 4]
1	ID	int64	0	0.000000	2240	[5524, 2174, 4141, 6182, 5324]
2	Year_Birth	int64	0	0.000000	59	[1957, 1954, 1965, 1984, 1981]
3	Education	object	0	0.000000	5	[S1, S3, S2, SMA, D3]
4	Marital_Status	object	0	0.000000	6	[Lajang, Bertunangan, Menikah, Cerai, Janda]
5	Income	float64	24	1.071429	1974	[58138000.0, 46344000.0, 71613000.0, 26646000...
6	Kidhome	int64	0	0.000000	3	[0, 1, 2]
7	Teenhome	int64	0	0.000000	3	[0, 1, 2]
8	Dt_Customer	object	0	0.000000	663	[04-09-2012, 08-03-2014, 21-08-2013, 10-02-201...
9	Recency	int64	0	0.000000	100	[58, 38, 26, 94, 16]
10	MntCoke	int64	0	0.000000	776	[635000, 11000, 426000, 173000, 520000]
11	MntFruits	int64	0	0.000000	158	[88000, 1000, 49000, 4000, 43000]
12	MntMeatProducts	int64	0	0.000000	558	[546000, 6000, 127000, 20000, 118000]
13	MntFishProducts	int64	0	0.000000	182	[172000, 2000, 111000, 10000, 46000]
14	MntSweetProducts	int64	0	0.000000	177	[88000, 1000, 21000, 3000, 27000]
15	MntGoldProds	int64	0	0.000000	213	[88000, 6000, 42000, 5000, 15000]
16	NumDealsPurchases	int64	0	0.000000	15	[3, 2, 1, 5, 4]
17	NumWebPurchases	int64	0	0.000000	15	[8, 1, 2, 5, 6]
18	NumCatalogPurchases	int64	0	0.000000	14	[10, 1, 2, 0, 3]
19	NumStorePurchases	int64	0	0.000000	14	[4, 2, 10, 6, 7]
20	NumWebVisitsMonth	int64	0	0.000000	16	[7, 5, 4, 6, 8]
21	AcceptedCmp3	int64	0	0.000000	2	[0, 1]
22	AcceptedCmp4	int64	0	0.000000	2	[0, 1]
23	AcceptedCmp5	int64	0	0.000000	2	[0, 1]
24	AcceptedCmp1	int64	0	0.000000	2	[0, 1]
25	AcceptedCmp2	int64	0	0.000000	2	[0, 1]
26	Complain	int64	0	0.000000	2	[0, 1]
27	Z_CostContact	int64	0	0.000000	1	[3]
28	Z_Revenue	int64	0	0.000000	1	[11]
29	Response	int64	0	0.000000	2	[1, 0]

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Engineering

1. Age: Kurangkan tahun ini (menggunakan datetime) dari tahun kelahiran pelanggan
2. Total Children: Jumlahkan jumlah anak dari kolom 'Kidhome' dan 'Teenhome'.
3. Age Group: Kategorikan usia menjadi 'Young adult' jika kurang dari 35, 'Adult' jika antara 35 dan 65, dan 'Senior adult' jika lebih dari 65.4. Total Accepted Campaign: Jumlahkan kampanye yang diterima dari kolom-kolom yang mengandung 'Acceptedcmp' dalam namanya.
5. Total Spent: Menambahkan kolom 'total_spent' yang merupakan total pengeluaran pelanggan untuk produk-produk tertentu
6. Total transaksi: Menambahkan kolom 'total_transaksi' yang merupakan total transaksi pembelian pelanggan kecuali 'NumWebVisitsMonth'.
7. Conversion Rate: Bagi total transaksi dengan jumlah kunjungan 'NumWebVisitsMonth.
8. Membership_duration: Hitung total hari bergabung, lalu bagikan dengan 365 untuk mendapatkan total tahun bergabung.
9. Dan terakhir menghapus kolom-kolom yang tidak diperlukan dari DataFrame.

```python
# date
df_prep['Dt_Customer'] = pd.to_datetime(df_prep['Dt_Customer'])

# age
df_prep['Age'] = 2022 - df_prep['Year_Birth']

# children
df_prep['children'] = df_prep['Kidhome'] + df_prep['Teenhome']

# is parent
df_prep['is_parent'] = np.where(df_prep['children']>0, 1, 0)

# Total of accepted campaign
df_prep['total_accepted_campaign']  = df_prep.loc[:, df_prep.columns.str.contains('AcceptedCmp')].sum(axis=1)

# Total Spending
df_prep['total_spent'] = df_prep[['MntCoke', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds']].sum(axis=1)

# Total Purchase Order except NumWebVisitsMonth
df_prep['total_transaksi'] = df_prep[['NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases', 'NumDealsPurchases']].sum(axis=1)

# Age Group
age_group = []
for i in df_prep['Age']:
    if i < 35:
        group = 'Young Adult'
    elif 35 <= i < 65:
        group = 'Adult'
    else:
        group = ' Senior Adult'
    age_group.append(group)

df_prep['age_group'] = age_group
```

```python
# Conversion rate
def cvr(x,y):
    if y == 0:
        return 0
    return x / y
df_prep['conversion_rate'] = df_prep.apply(lambda x: cvr(x['total_transaksi'],x['NumWebVisitsMonth']), axis=1)


# Total Years joined
df_prep['membership_duration'] = 2022 - df_prep['Dt_Customer'].dt.year
```

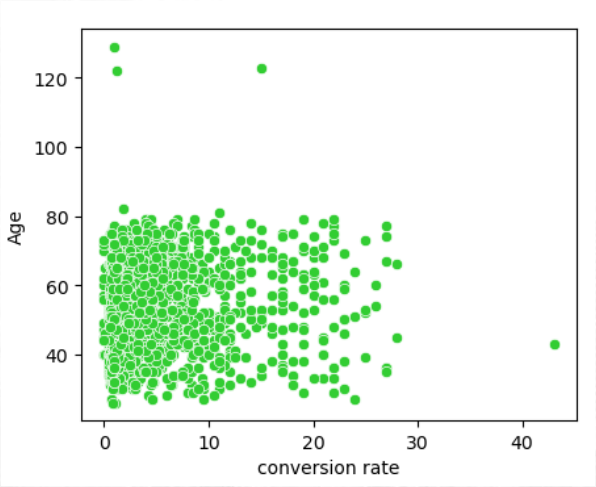
1.1s

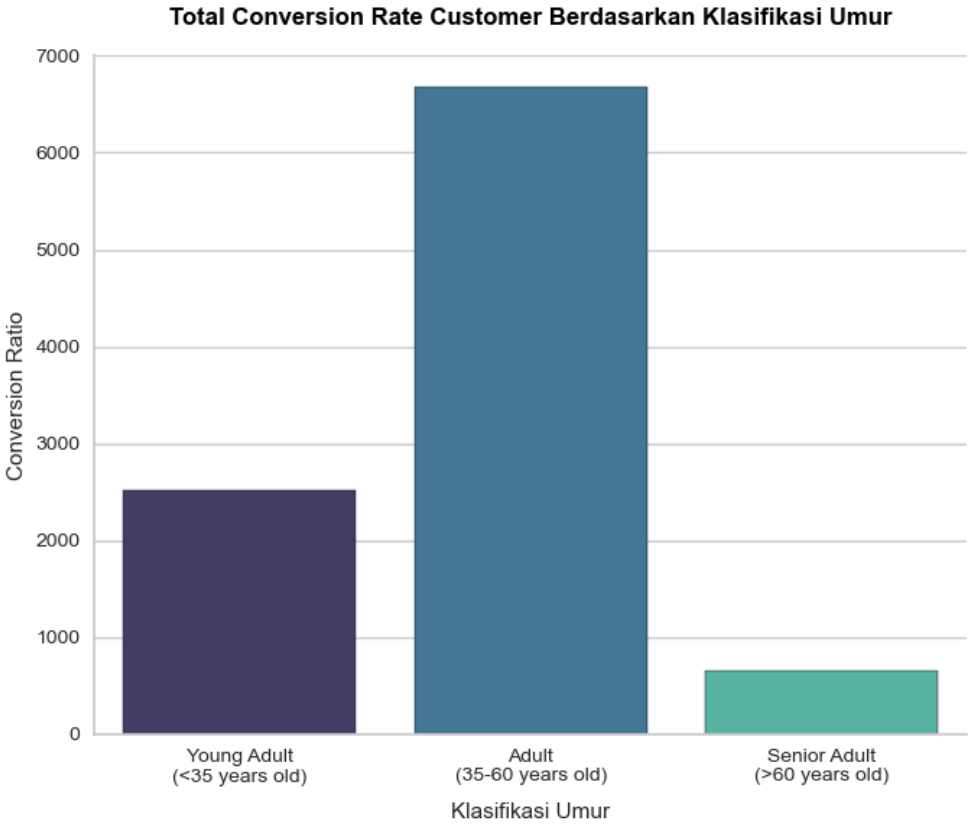
```python
df_prep.drop(['Unnamed: 0', 'ID', 'Year_Birth', 'Z_CostContact', 'Z_Revenue'], inplace=True, axis=1)
```

0.0s

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

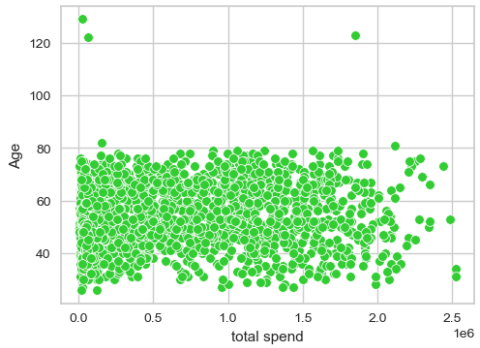


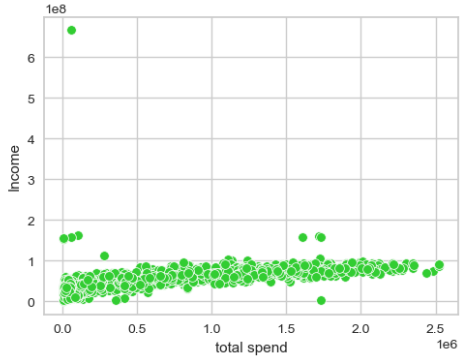


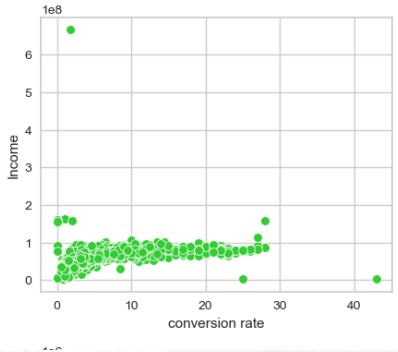
- Korelasi antara age dan conversion rate tidak begitu kuat / lemah yang artinya age dan conversion rate tidak saling mempengaruhi
- Berdasrkan visualisasi di sebelah kanan ini, kelompok Adult (usia 35-60 tahun) memiliki kontribusi tertinggi dalam conversion rate sebesar 67.79%, diikuti oleh kelompok Senior Adult (>60 tahun) dengan kontribusi 25.58%, dan kelompok Young Adult (<35 tahun) memiliki kontribusi terendah sebesar 6.63%.

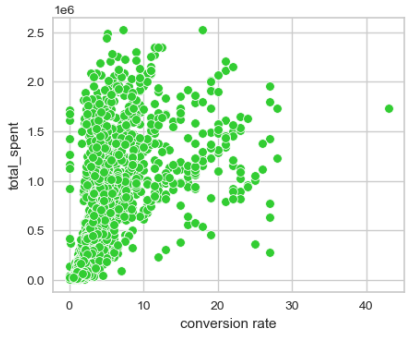
Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age









Ada hubungan positif linier antara tingkat konversi dan total pengeluaran serta antara tingkat konversi dan pendapatan. Semakin tinggi jumlah yang dibelanjakan atau pendapatan, semakin tinggi tingkat konversi. Namun, hubungan antara tingkat konversi dan usia tidak begitu signifikan karena distribusi tingkat konversi relatif merata di berbagai rentang usia. Selain itu, terdapat hubungan positif linier antara total pengeluaran dan pendapatan, yang menunjukkan bahwa semakin tinggi pendapatan, semakin tinggi total pengeluaran.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Cleaning & Preprocessing

```
missing_count = df_clean.isnull().sum()

# Menghitung persentase missing value pada setiap kolom
missing_percentage = round((df_clean.isnull().sum() / len(df_clean)) * 100,2)

# Menggabungkan kedua Series ke dalam DataFrame
missing_df = pd.concat([missing_count, missing_percentage], axis=1)
missing_df.columns = ['Jumlah Missing Value', 'Persentase (%)']

# Menampilkan kolom-kolom yang memiliki nilai null beserta persentase missing valuenya
missing_df[missing_df['Jumlah Missing Value'] > 0]
```

✓ 0.0s

	Jumlah Missing Value	Persentase (%)
Income	24	1.07

```
# Mencari Nilai Null dengan Whitespace dalam DataFrame
white_space = []
for col in df_clean.columns:
    for val in df_clean[col]:
        if isinstance(val, str) and '  ' in val:
            white_space.append(val)

# Output
print(white_space)
```

✓ 0.0s

```
[]
```

```
df.duplicated().sum()
```

✓ 0.0s

```
0
```

•Tidak ditemukan nilai null dengan whitespace dalam DataFrame, karena output dari pencarian tersebut (print(white_space)) menunjukkan list kosong ([]).

•Sebelumnya, telah dihitung bahwa kolom 'Income' memiliki 24 nilai null, yang merupakan 1.07% dari total data.

•tidak ada baris yang merupakan duplikat dalam DataFrame. Dengan kata lain, setiap baris dalam DataFrame adalah unik.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Cleaning & Preprocessing

```
from scipy import stats

# Define the list of features
features = ['Income', 'Age']

print(f'Total Baris sebelum melakukan Handling Outlier = {len(df_cast)}')

handling_outlier = np.array([True] * len(df_cast))

for col in features:
    zscore = abs(stats.zscore(df_cast[col]))
    handling_outlier = (zscore < 3) & handling_outlier

df_cast = df_cast[handling_outlier]

print(f'Total Baris setelah melakukan Handling Outlier = {len(df_cast)}')
```

✓ 0.0s

```
Total Baris sebelum melakukan Handling Outlier = 2216
Total Baris setelah melakukan Handling Outlier = 2205
```

```
# Label Encoding
mapping_education = {
    'SMA': 0,
    'D3' : 1,
    'S1' : 2,
    'S2' : 3,
    'S3' : 4
}
df_cast['mapp_education'] = df_cast['Education'].map(mapping_education)
```

✓ 0.0s

```
# One Hot Encoding

for category in ['age_group', 'Marital_Status']:
    onehots = pd.get_dummies(df_cast[category], prefix=category)
    df_cast = df_cast.join(onehots)
df_cast.head()
```

✓ 0.1s

```
from sklearn.preprocessing import StandardScaler
ss = StandardScaler()
df_scaled = df_cast.copy()

for col in numerical:
    df_scaled[col] = ss.fit_transform(df_scaled[[col]])

display(df_scaled.shape, df_scaled.head())
```

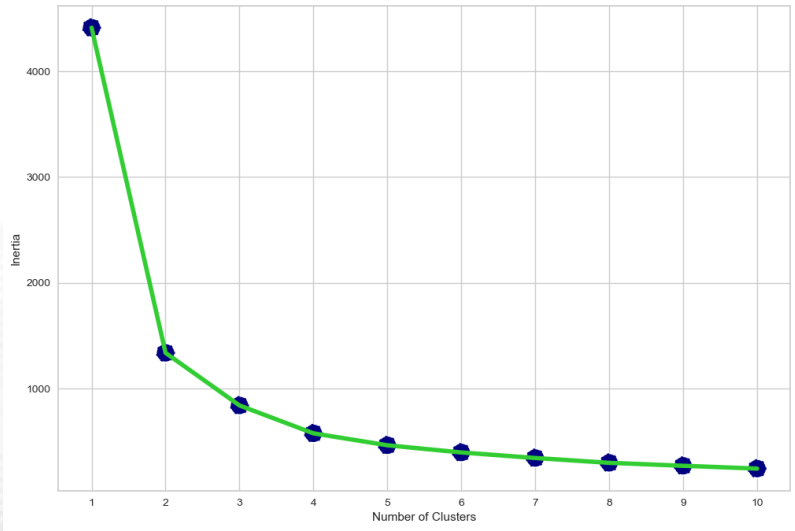
✓ 0.5s

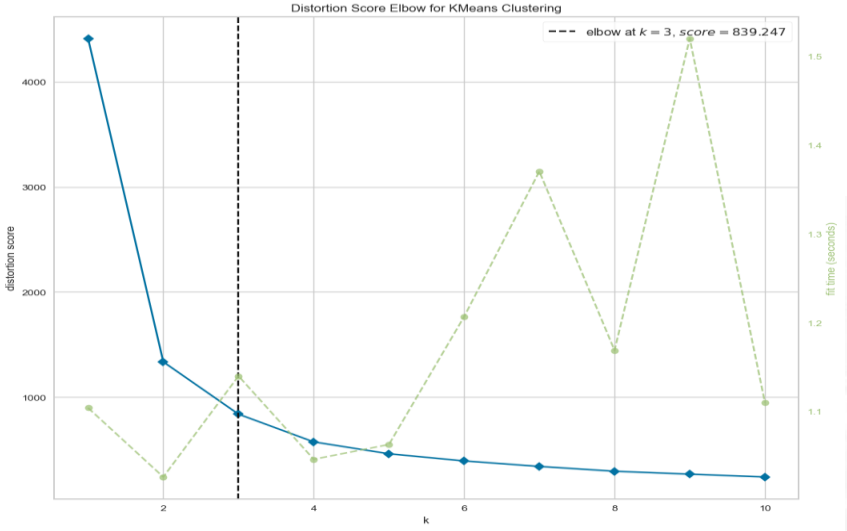
```
(2205, 43)
```

1. handling outlier menggunakan zscore, terdapat 11 baris yang dihapus dari DataFrame karena dianggap sebagai outlier berdasarkan kriteria Z-score kurang dari 3 untuk fitur 'Income' dan 'Age'.
2. Feature encoding menggunakan 2 cara yaitu dengan label encoding dan one hot encoding, Setelah proses encoding, DataFrame memiliki kolom baru 'mapp_education' yang berisi nilai numerik untuk kolom 'Education' berdasarkan mapping yang telah ditentukan. Kolom-kolom baru hasil One Hot Encoding untuk 'age_group' dan 'Marital_Status' telah ditambahkan ke DataFrame.
3. Standarization menggunakan standarscaller, Dengan normalisasi, kita dapat memastikan bahwa seluruh fitur memiliki pengaruh yang seimbang terhadap hasil prediksi model.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling





Dari ke 2 grafik tersebut, kita bisa melihat bahwa penurunan yang signifikan terjadi dari 1 cluster hingga 3 cluster, kemudian penurunan masih terjadi tapi dengan laju yang lebih lambat. Hal ini bisa menunjukkan bahwa memilih 3 cluster merupakan pilihan yang baik, karena penambahan cluster setelah itu memberikan penurunan inertia yang lebih sedikit.

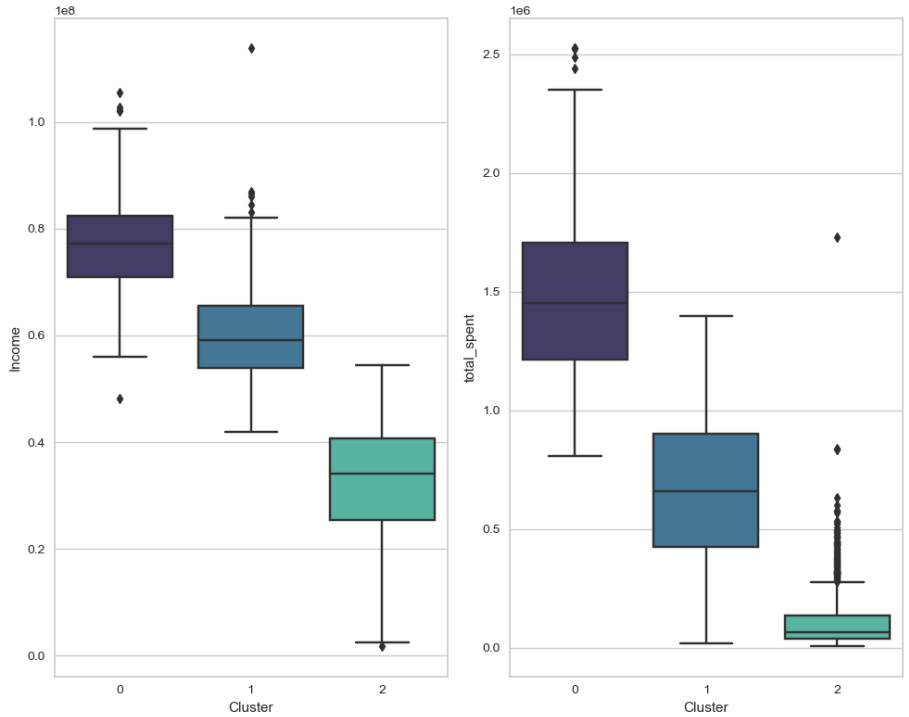
Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling

Berdasarkan hasil Silhouette score, jumlah cluster yang optimal untuk data tersebut adalah 2 atau 3.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

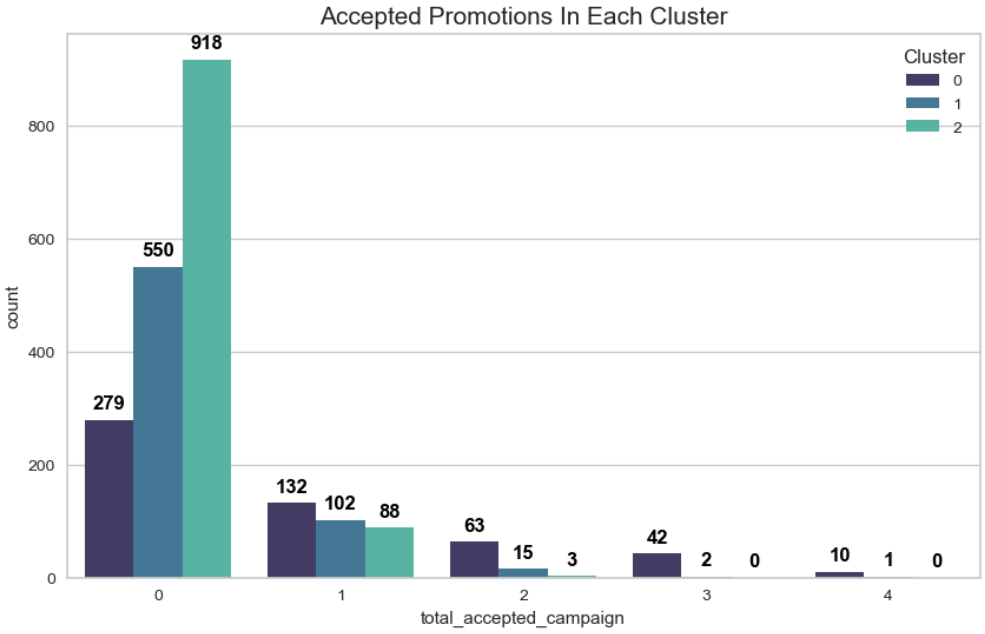


Kesimpulan:

- Cluster 0 dapat dianggap sebagai "High Spender" karena memiliki rata-rata pendapatan yang tinggi ($77,072,150) dan total belanja ($1,494,454) tertinggi di antara ketiga cluster.
- Cluster 1 dapat dianggap sebagai "Mid Spender" karena memiliki rata-rata pendapatan ($59,966,660) dan total belanja ($656,528) di antara Cluster 0 dan 2.
- Cluster 2 dapat dianggap sebagai "Low Spender" karena memiliki rata-rata pendapatan ($32,813,780) dan total belanja ($111,085) terendah di antara ketiga cluster.

Untuk selengkapnya, dapat melihat jupyter notebook disini

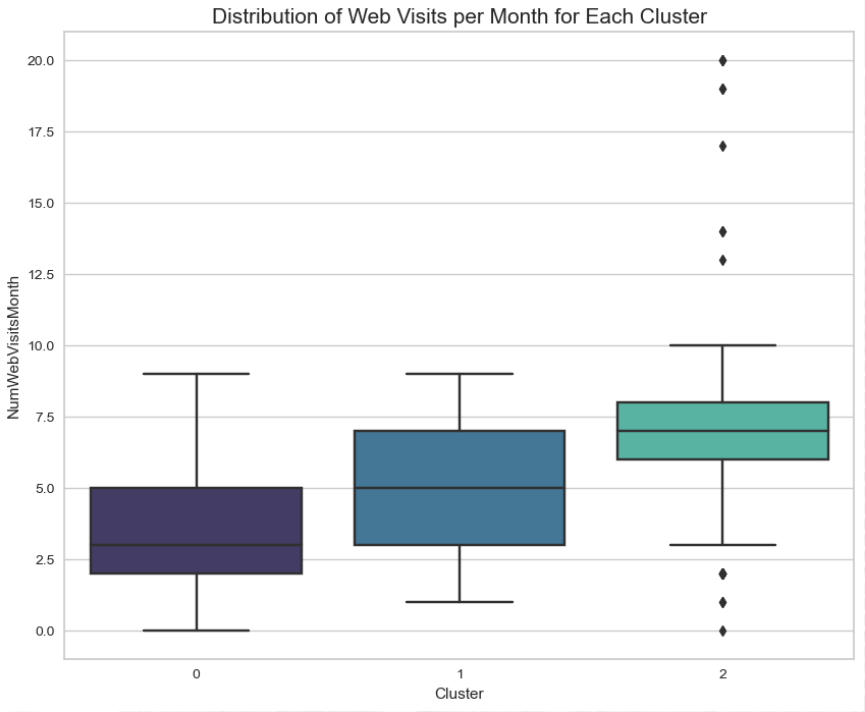
Customer Personality Analysis for Marketing Retargeting



- Cluster 0 memiliki sebagian besar pelanggan yang tidak menerima promosi (279 pelanggan), diikuti oleh 132 pelanggan yang menerima 1 promosi, 63 pelanggan yang menerima 2 promosi, dan seterusnya.
- Cluster 1 memiliki pola yang mirip dengan Cluster 0, namun dengan jumlah yang lebih tinggi. Mayoritas pelanggan Cluster 1 tidak menerima promosi (550 pelanggan), diikuti oleh 102 pelanggan yang menerima 1 promosi, dan jumlah yang semakin sedikit untuk promosi yang lebih banyak.
- Cluster 2 memiliki pola yang berbeda, dengan mayoritas pelanggan menerima 0 promosi (918 pelanggan), diikuti oleh 88 pelanggan yang menerima 1 promosi, dan hanya sedikit yang menerima 2 promosi.

Untuk selengkapnya, dapat melihat jupyter notebook disini

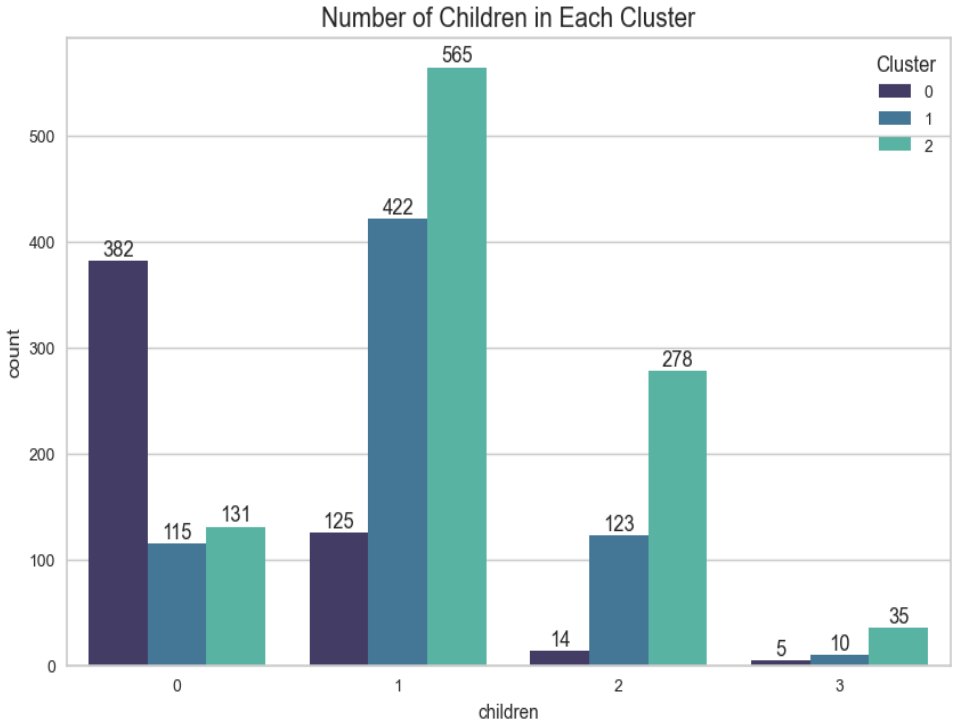
Customer Personality Analysis for Marketing Retargeting



- Cluster 2 adalah kelompok dengan jumlah kunjungan web per bulan tertinggi di antara ketiga Cluster. memiliki rata-rata kunjungan web per bulan sekitar 3, dengan sebagian besar data berada di rentang 2 hingga 5 kunjungan, menunjukkan tingkat aktivitas online yang lebih tinggi.
- Cluster 1 berada di posisi tengah dengan jumlah kunjungan web per bulan yang lebih rendah dari Cluster , memiliki rata-rata kunjungan web per bulan sekitar 5, dengan sebagian besar data berada di rentang 3 hingga 7 kunjungan, namun lebih tinggi dari Cluster 0.
- Cluster 0 adalah kelompok dengan jumlah kunjungan web per bulan terendah, menunjukkan tingkat aktivitas online yang lebih rendah dibandingkan dengan kedua Cluster lainnya, yaitu memiliki rata-rata kunjungan web per bulan sekitar 7, dengan sebagian besar data berada di rentang 6 hingga 8 kunjungan. Terdapat beberapa nilai outlier yang mencapai 20

Untuk selengkapnya, dapat melihat jupyter notebook disini

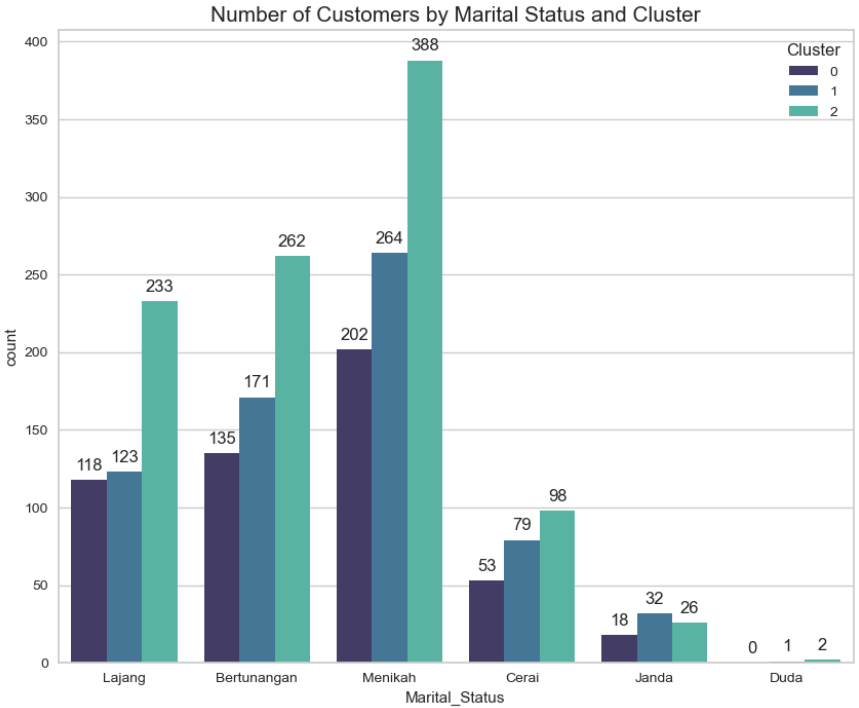
Customer Personality Analysis for Marketing Retargeting



- Cluster 0 cenderung memiliki anggota tanpa anak atau dengan 1 anak.
- Cluster 1 cenderung memiliki anggota dengan 1 anak, tetapi juga memiliki sejumlah anggota tanpa anak atau dengan 2 anak.
- Cluster 2 cenderung memiliki anggota dengan 1 anak atau 2 anak, dengan jumlah yang cukup signifikan.

Untuk selengkapnya, dapat melihat jupyter notebook disini

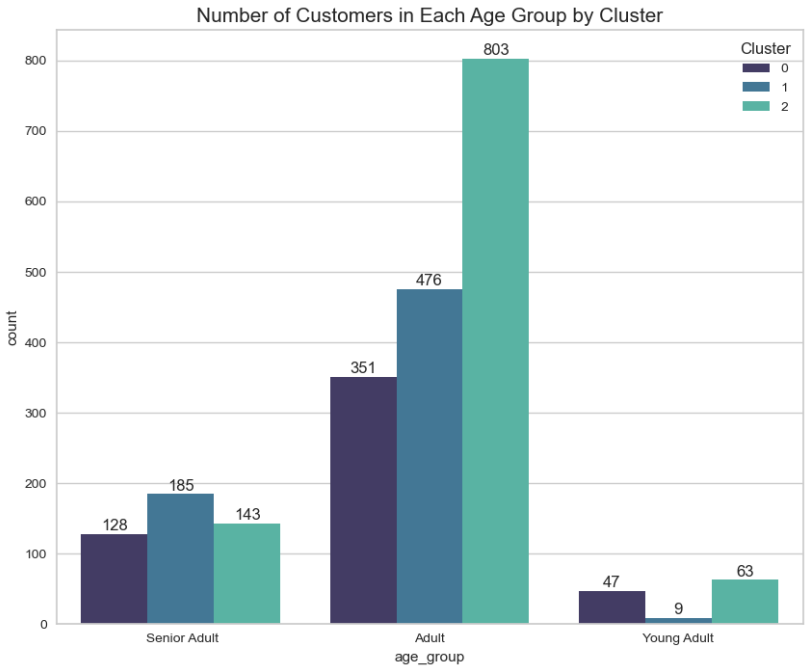
Customer Personality Analysis for Marketing Retargeting



- Cluster 0: Mayoritas pelanggan dalam cluster ini adalah yang sudah menikah (Menikah), diikuti oleh pelanggan yang masih lajang (Lajang). Jumlah pelanggan yang bertunangan (Bertunangan) dan cerai (Cerai) juga signifikan, namun jumlah pelanggan janda (Janda) relatif lebih sedikit.
- Cluster 1: Juga didominasi oleh pelanggan yang sudah menikah (Menikah), dengan jumlah yang lebih sedikit untuk status pernikahan lainnya. Terdapat satu pelanggan dengan status duda (Duda).
- Cluster 2: Sama seperti Cluster 0, cluster ini juga didominasi oleh pelanggan yang sudah menikah (Menikah), diikuti oleh pelanggan yang masih lajang (Lajang) dan bertunangan (Bertunangan). Jumlah pelanggan dengan status pernikahan lainnya (Cerai, Duda, Janda) relatif lebih sedikit.

Untuk selengkapnya, dapat melihat jupyter notebook disini

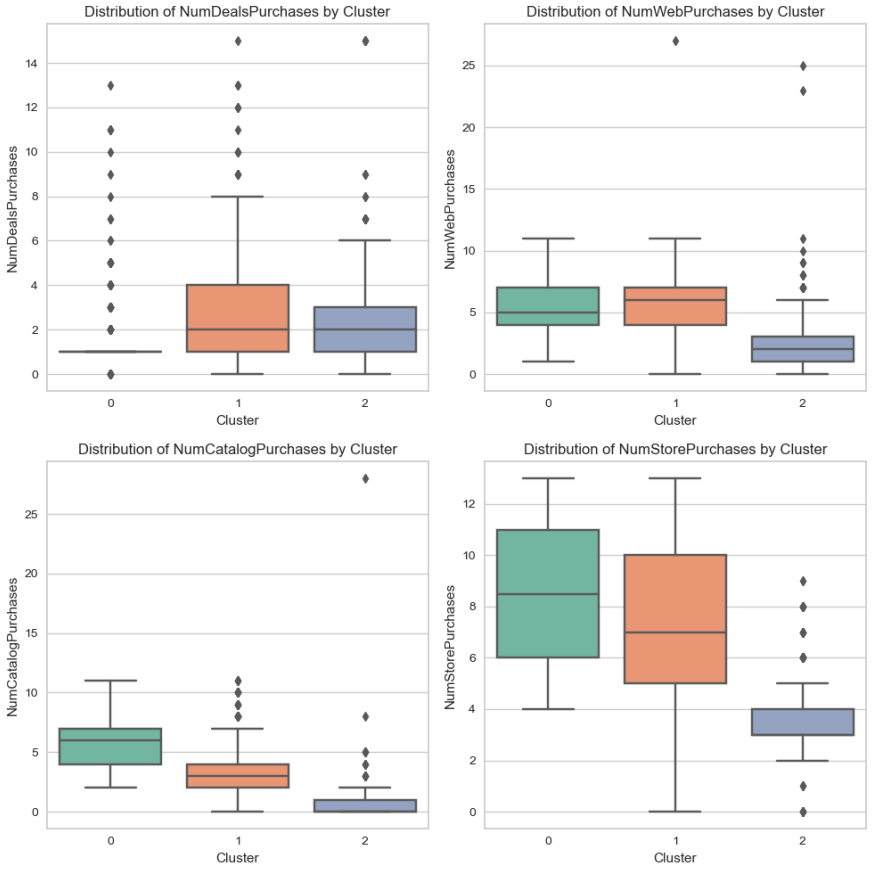
Customer Personality Analysis for Marketing Retargeting



- Cluster 0 memiliki sebagian besar pelanggan dewasa (Adult) dengan jumlah yang signifikan dari kelompok usia Senior Adult, tetapi hanya sedikit pelanggan dari kelompok usia Young Adult.
- Cluster 1 juga didominasi oleh pelanggan dewasa (Adult), tetapi memiliki jumlah pelanggan Senior Adult yang lebih sedikit dibandingkan dengan Cluster 0. Cluster ini hampir tidak memiliki pelanggan dari kelompok usia Young Adult.
- Cluster 2 memiliki sebagian besar pelanggan dewasa (Adult) dengan jumlah yang signifikan dari kelompok usia Senior Adult, tetapi memiliki proporsi pelanggan dari kelompok usia Young Adult yang lebih tinggi dibandingkan dengan Cluster 0 dan 1.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting



- Cluster 0 cenderung menjadi cluster dengan pola pembelian paling tinggi untuk hampir semua kategori pembelian.
- Cluster 1 memiliki pola pembelian yang cukup tinggi, terutama dalam pembelian melalui web dan katalog.
- Cluster 2 cenderung menjadi cluster dengan pola pembelian paling rendah untuk hampir semua kategori pembelian, terutama dalam pembelian melalui web dan katalog.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

Summary

1. High Spender (Cluster 0

- Memiliki total belanja dan pendapatan tertinggi di antara Cluster.
- Cenderung kurang aktif secara online namun lebih suka berbelanja langsung di toko atau melalui katalog.
- Mayoritas tidak tertarik pada promosi.
- Didominasi oleh pelanggan yang sudah menikah dengan satu atau tanpa anak.

2. Mid Spender (Cluster 1)

- Memiliki total belanja dan pendapatan sedang.
- Aktivitas online sedang, dengan jumlah kunjungan web dan pembelian melalui web yang cukup.
- Kurang tertarik pada promosi.
- Didominasi oleh pelanggan yang sudah menikah dengan satu anak, tetapi juga termasuk yang tanpa anak atau dengan dua anak.

3. Low Spender (Cluster 2)

- Memiliki total belanja dan pendapatan terendah di antara Cluster.
- Aktivitas online tinggi, dengan rata-rata kunjungan web per bulan tertinggi.
- Tertarik pada promosi, dengan sebagian besar menerima sedikit atau tidak ada promosi.
- Didominasi oleh pelanggan yang sudah menikah dengan satu atau dua anak.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Recommendation

1. Untuk meningkatkan kinerja bisnis dan mengurangi potensi churn, perusahaan dapat menerapkan beberapa strategi berdasarkan analisis klaster pelanggan. Pertama, untuk pelanggan High Spender (Cluster 0), fokus pada pengembangan produk unggulan dan layanan eksklusif, serta meningkatkan promosi di toko fisik dan katalog. Implementasikan juga membership program dengan manfaat eksklusif untuk meningkatkan loyalitas pelanggan.
2. Kedua, untuk pelanggan Mid Spender (Cluster 1), perbaiki promosi online yang lebih terarah, tawarkan paket produk atau penawaran khusus, dan sediakan membership program dengan poin reward atau diskon khusus untuk mempertahankan pelanggan.
3. Terakhir, untuk pelanggan Low Spender (Cluster 2), berikan promosi yang menarik, tingkatkan interaksi melalui platform online, dan perbaiki layanan pelanggan online. Evaluasi juga potensi implementasi membership program dengan manfaat tambahan untuk meningkatkan keterlibatan pelanggan. Dengan strategi ini, diharapkan dapat meningkatkan retensi pelanggan, meningkatkan nilai transaksi, dan mengurangi potensi churn secara efektif.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

	Cluster	sum
0	0	786083000
1	1	439874000
2	2	112085000

```
# Jumlah potensial penghematan dari optimalisasi biaya promosi (dengan asumsi pengurangan target 50%)
potensi_penghematan = (df_cast[df_cast.Cluster == 2]['total_spent'].sum() / df_cast[df_cast.Cluster == 2]
                       ['total_transaksi'].sum()) * df_cast[df_cast.Cluster == 2]['NumDealsPurchases'].sum() * 0.5

print("Potensi Penghematan:", potensi_penghematan)
```

✓ 0.0s

```
Potensi Penghematan: 14835558.938283283
```

Hasilnya menunjukkan total pengeluaran untuk masing-masing cluster:

- Cluster 0 memiliki total pengeluaran sebesar 786.083.000.
- Cluster 1 memiliki total pengeluaran sebesar 439.874.000.
- Cluster 2 memiliki total pengeluaran sebesar 112.085.000.

Dari hasil potensi penghematan sebesar 14.835.558,94, ini berarti perusahaan dapat menghemat biaya promosi sebesar itu. Jika asumsi pengurangan target 50%, berarti potensi penghematan tersebut adalah separuh dari biaya promosi yang biasanya dikeluarkan. Dengan demikian, total biaya promosi yang dapat disave adalah sekitar 29.671.117,88.

Untuk selengkapnya, dapat melihat jupyter notebook disini