Predict Customer Personality to boost marketing campaign by using Machine Learning

Created by:
Muhammad Oky Hariawan
okyhariawan@gmail.com
https://www.linkedin.com/in/okyhariawan/

"A bachelor's degree in business management with 2 year's expertise in Customer Service and Workforce Management. A data-driven & tech-savvy person who has huge interest in data analytics and currently learning data science. Skilled at analyse data, keen attention to detail, ability to effectively prioritize and execute tasks in a high-pressure environment. Proficient using python, SQL, Microsoft Power BI, Excel VBA, & other statistical tools."

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

“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 ”

Exploration Data Analysis (EDA)

	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]

Insight:

- Data ini memiliki 2240 baris dan 30 kolom
- Terdapat missing value pada kolom Income sebanyak 24 baris
- Data ini memiliki tipe data int64 (26), float64 (1), object(3)

Untuk selengkapnya, dapat melihat jupyter notebook disini

FEATURE ENGINEERING

1. Age

```
# age
df_clean['age'] = 2022 - df_clean['Year_Birth']
```

4. Member Duration

```
# Membership duration
df_clean['member_duration']= 2022 - df_clean['Dt_Customer'].dt.year
```

3. Is Parent

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

2. Children

```
# children
df_clean['children'] = df_clean['Kidhome'] + df_clean['Teenhome']
```

5. Conversion Rate

```
#create conversion rate feature
def cvr(x,y):
    if y == 0:
        return 0
    return x / y
df_clean['conversion_rate'] = df_clean.apply(lambda x: cvr(x['totaltransaction'],x['NumWebVisitsMonth']), axis=1)
```

FEATURE ENGINEERING

6. Total Spending

```
# total spending
df_clean['totalspending'] =  df_clean['MntCoke'] \
                        + df_clean['MntFruits'] \
                        + df_clean['MntMeatProducts'] \
                        + df_clean['MntFishProducts'] \
                        + df_clean['MntSweetProducts'] \
                        + df_clean['MntGoldProds']
```

7. Total Transaction

```
# total transaction
df_clean['totaltransaction'] = df_clean['NumWebPurchases'] \
                           + df_clean['NumCatalogPurchases'] \
                           + df_clean['NumStorePurchases'] \
                           + df_clean['NumDealsPurchases']
```

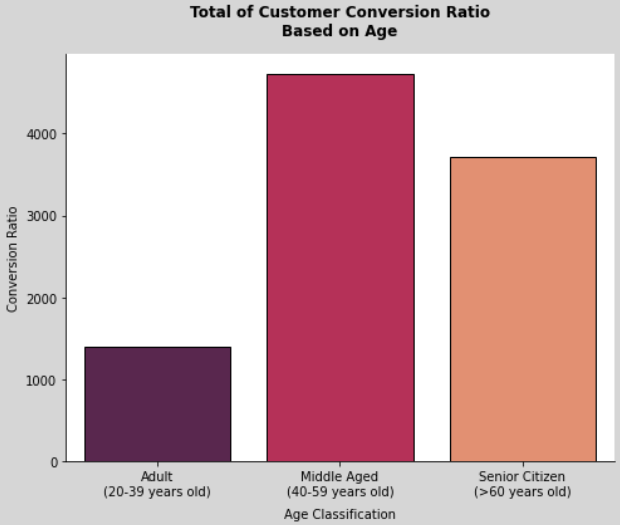
8. Total Accepted Campaign

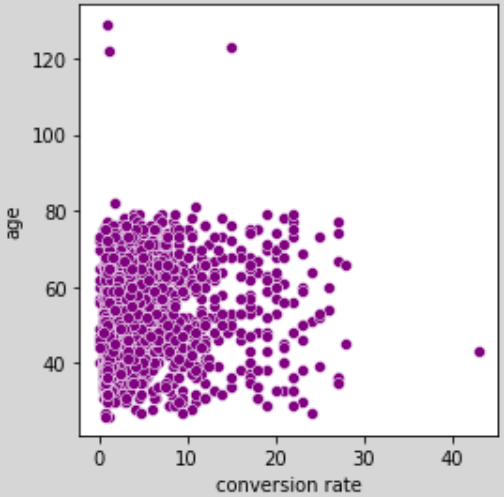
```
# total accepted campaign
df_clean['acceptcampaign'] = df_clean['AcceptedCmp3'] \
                        + df_clean['AcceptedCmp4'] \
                        + df_clean['AcceptedCmp5'] \
                        + df_clean['AcceptedCmp1'] \
                        + df_clean['AcceptedCmp2']
```

9. Age Group

```
#create age group
age_list=[]
for i in df_clean['age']:
    if i >= 0 and i <= 1:
        group = 'Infant'
    elif i >= 2 and i <= 4:
        group = 'Toddler'
    elif i >= 5 and i <= 12:
        group = 'Child'
    elif i >= 13 and i <= 19:
        group = 'Teen'
    elif i >= 20 and i <= 39:
        group = 'Adult'
    elif i >= 40 and i <= 59:
        group = 'Middle Aged'
    else:
        group = 'Senior Citizen'
    age_list.append(group)
df_clean['Age_Group'] = age_list
```

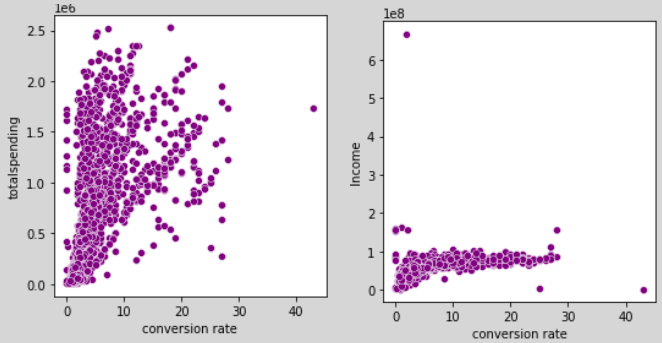
Conversion Rate Analysis Based on Income, Spending and Age

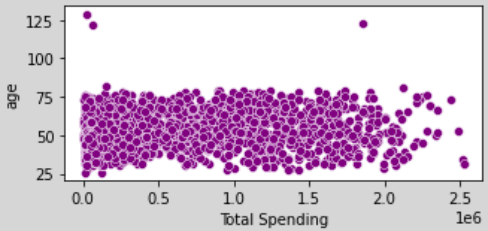


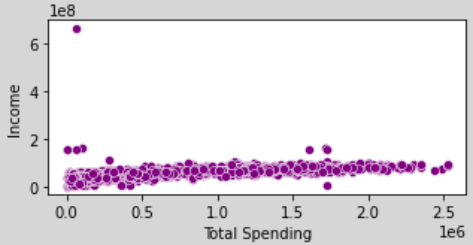


- Dari visualisasi diatas, distribusi customer berdasarkan conversion rate di dominasi oleh **Middle Aged** (48.05%), di ikuti oleh **Senior Citizen** (37.76%), dan **Adult** (14.19%)
- Korelasi antara conversion rate dan umur kurang signifikan atau lemah yang berarti umur dan conversion rate tidak saling mempengaruhi

Conversion Rate Analysis Based on Income, Spending and Age







- Terdapat korelasi positif linier antara variabel conversion rate dan variabel total spending. Semakin banyak jumlah yang dibelanjakan, semakin tinggi conversion rate.
- Terdapat korelasi positif linier antara variabel conversion rate dan variabel income. Semakin tinggi income, semakin tinggi conversion rate.
- Korelasi antara variabel conversion rate dan age kurang signifikan karena distribusi conversion rate pada variabel age cenderung rata-rata.
- Terdapat korelasi positif linier antara variabel total spending dan variabel income. Berarti semakin tinggi income maka semakin tinggi pula total spending

DATA PREPROCESSING

Handle Null Value

```
data_missing_value = df_clean.isnull().sum().reset_index()
data_missing_value.columns = ['feature','missing_value']
data_missing_value['percentage'] = round((data_missing_value['missing_value']/len(df_clean))*100,3)
data_missing_value = data_missing_value.sort_values('percentage', ascending=False).reset_index(drop=True)
data_missing_value = data_missing_value[data_missing_value['percentage']>0]
data_missing_value
```

	feature	missing_value	percentage
0	Income	24	1.071

- Terdapat *null*/value pada kolom income sebanya 24 baris
- Baris ini akan di drop karena persentasenya hanya 1%

```
#Remove rows that have no income data
df_clean.dropna(subset=['Income'], inplace=True)
```

Handle Duplicated Value

```
# Check duplicated value
df.duplicated().sum()
```

```
0
```

- Tidak terdapat duplicated value dalam dataset ini

DATA PREPROCESSING

Feature Encoding

- Label encoding: Education
- One Hot Encoding: Age_Group & Marital_Status

```
# label encoder
mapping_education = {
    'SMA' : 0,
    'D3' : 1,
    'S1' : 2,
    'S2' : 3,
    'S3' : 4}

df_clean['education_mapped'] = df_clean['Education'].map(mapping_education)
```

```
# handle dengan one hot encoding
for cat in ['Age_Group', 'Marital_Status']:
    onehots = pd.get_dummies(df_clean[cat], prefix=cat)
    df_clean = df_clean.join(onehots)
```

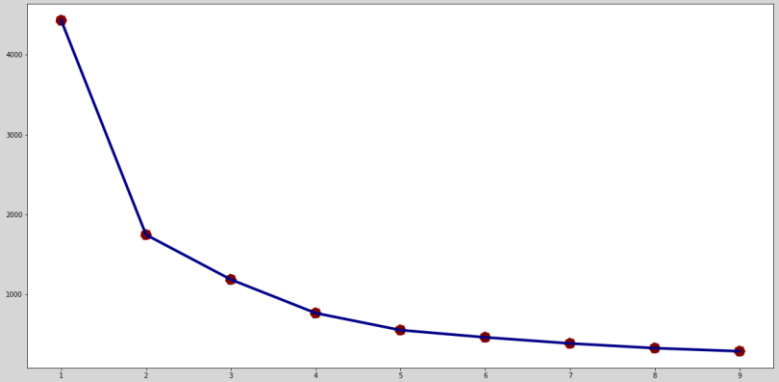
Standardization

Standarisasi kolom numerical dalam dataset

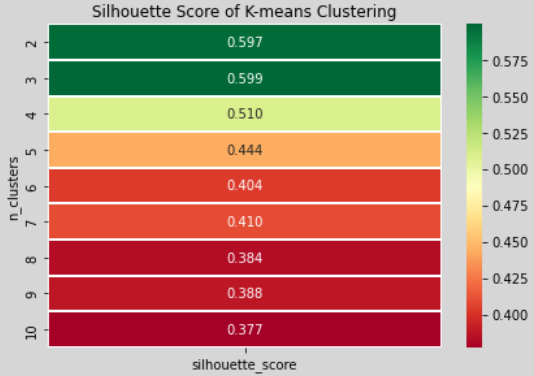
```
from sklearn.preprocessing import StandardScaler
ss = StandardScaler()
df2_scaled = df_clean.copy()

for col in numerical:
    df2_scaled[col] = ss.fit_transform(df2_scaled[[col]])
```

DATA MODELING

Elbow Method



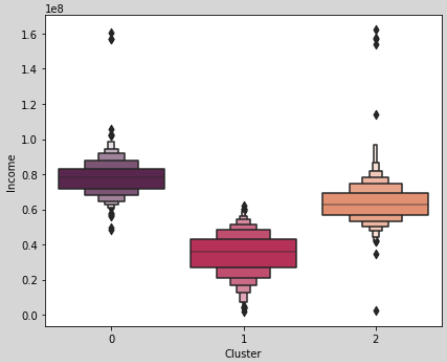
Silhouette Score

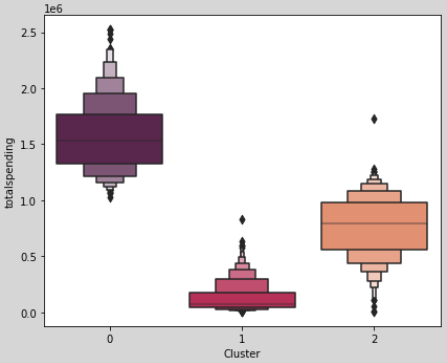


- Berdasarkan elbow method, cluster terbaik ada pada antara 3 & 4
- Kemudian berdasarkan Silhouette Score kita menentukan untuk menggunakan 3 cluster

CLUSTER IDENTIFICATION

BASED ON INCOME AND TOTAL SPEND





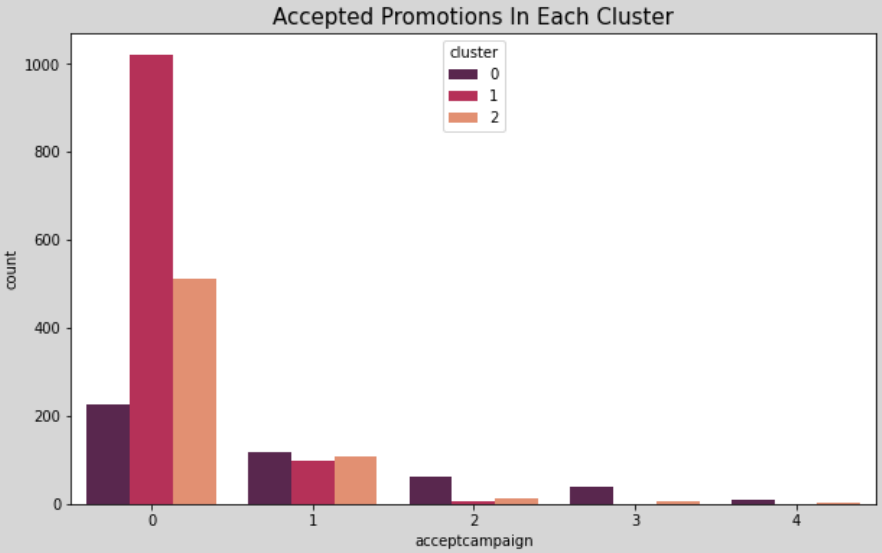
- **Cluster 0**
  - Memiliki income tertinggi dengan rata-rata Rp. 77.972.000
  - Memiliki total spending atau pengeluaran yang paling tinggi dengan Rp. 15.28.000
  - High Spender cluster

- **Cluster 1**
  - Memiliki income paling rendah dengan rata-rata Rp. 35.683.000
  - Memiliki total spending atau pengeluaran yang paling rendah dengan Rp. 70.000
  - Low Spender cluster

- **Cluster 2**
  - Memiliki income yang berada ditengah ketiga cluster dengan Rp. 62.559.500
  - Memiliki total spending atau pengeluaran yang berada ditengah ketiga cluster dengan Rp. 794.000
  - Mid Spender Cluster

CLUSTER IDENTIFICATION

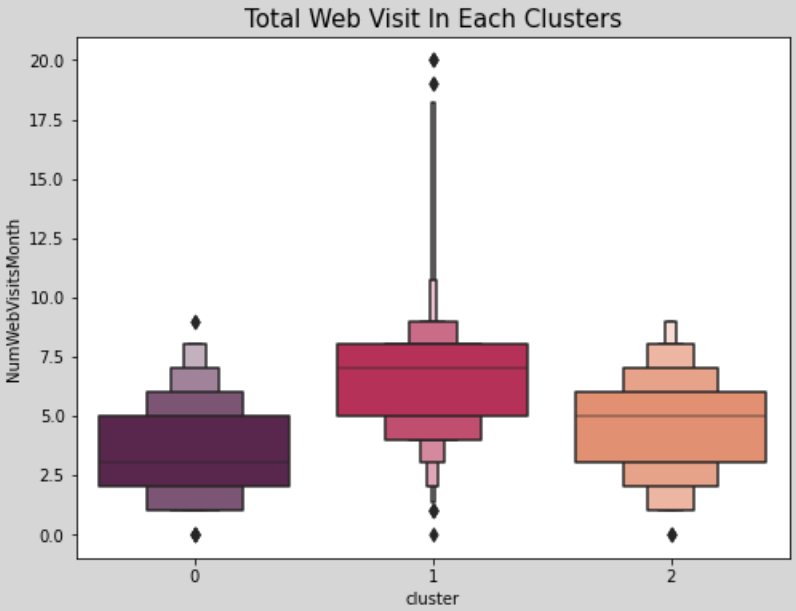
BASED ON ACCEPTED CAMPAIGN



- **Cluster 0**
  - Setidaknya dapat menerima 1 hingga 4 campaign
- **Cluster 1**
  - Cenderung tidak memiliki ketertarikan dalam menerima campaign
- **Cluster 2**
  - Memiliki ketertarikan yang rendah dalam menerima campaign setidaknya 1 campaign

CLUSTER IDENTIFICATION

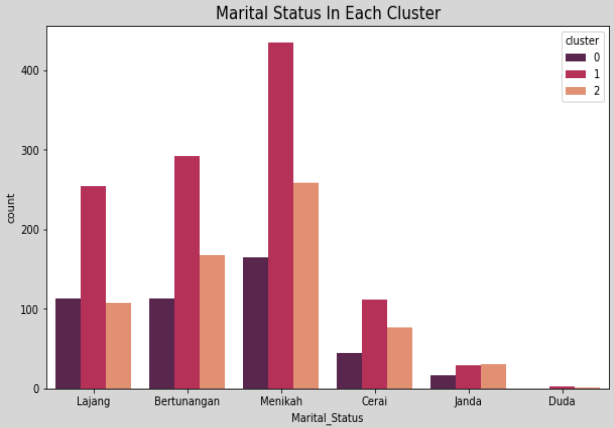
BASED ON WEB VISIT

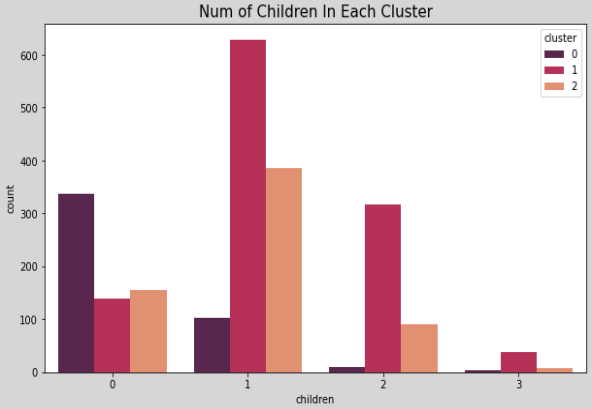


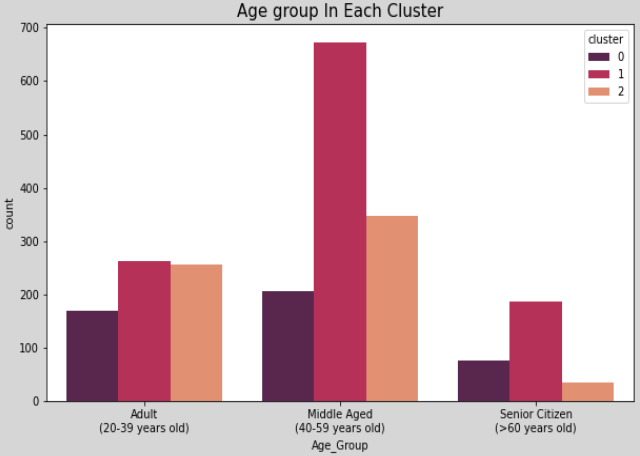
Cluster 1 memiliki jumlah web visit terbanyak sekitar 6-7 kali perbulan. Sedangkan cluster 0 menjadi yang terendah dengan rata-rata 3 kali perbulan

CLUSTER IDENTIFICATION

CUSTOMER PROFILE



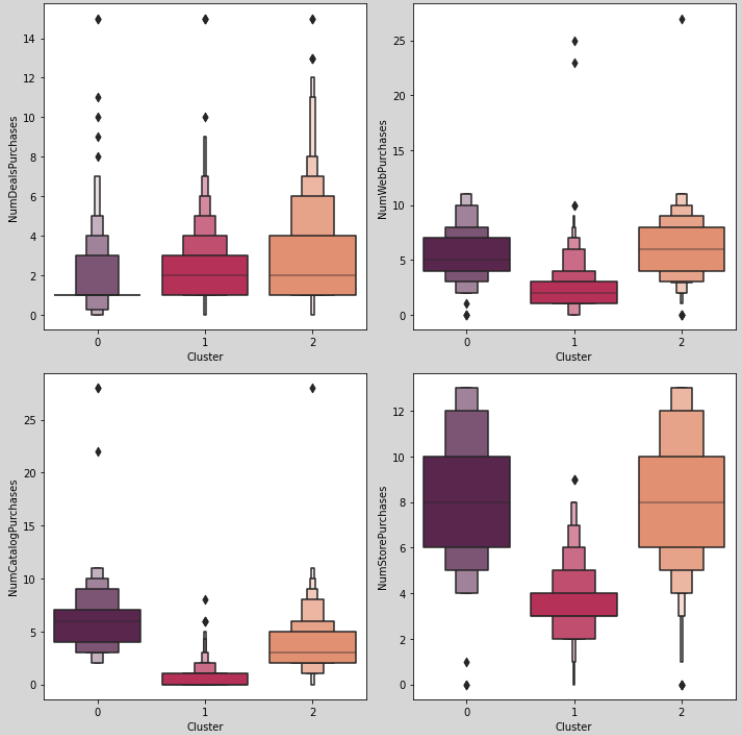




- **Cluster 0**
  - Didominasi oleh Middle Aged berusia 40-59 tahun berstatus menikah (namun memiliki angka tinggi dalam status lajang dan bertungan) dan mayoritas tidak memiliki anak
- **Cluster 1**
  - Didominasi oleh Middle Aged berusia 40-59 tahun berstatus menikah dan mayoritas memiliki 1 anak atau lebih
- **Cluster 2**
  - Didominasi oleh Middle Aged berusia 40-59 tahun berstatus menikah dan setidaknya memiliki 1 anak.

CLUSTER IDENTIFICATION

BASED ON PURCHASING HISTORY



- Cluster 1 & 2 melakukan pembelian deals sekitar 2 kali per bulan
- Cluster 2 memiliki NumWebPurchases tertinggi kemudian disusul oleh cluster 0
- Cluster 0 tampak lebih melakukan pembelian melalui katalog (NumCatalogPurchases) sedangkan cluster 1 terendah dengan average 0 pembelian
- Cluster 0 & 2 tampak memiliki pembelian tertinggi melalui store (NumStorePurchases) dengan average 8 kali pembelian

SUMMARY

Cluster 0 (High Spender)

- Didominasi oleh Middle Aged berusia 40-59 tahun berstatus menikah (namun memiliki angka tinggi dalam status lajang dan bertungan) dan mayoritas tidak memiliki anak
- Setidaknya dapat menerima 1 hingga 4 campaign
- Jumlah pembelian melalui web sekitar 5 kali walaupun dengan jumlah visit web terendah
- Memiliki history pembelian terbanyak pada catalog & store sekitar 8 kali pembelian
- Memiliki income tertinggi dan total spending tertinggi.

Cluster 1 (Low Spender)

- Didominasi oleh Middle Aged berusia 40-59 tahun berstatus menikah dan mayoritas memiliki 1 anak atau lebih
- Cenderung tidak memiliki ketertarikan dalam menerima campaign
- Memiliki angka jumlah visit web terbanyak (6 kali) namun history pembelian melalui web hanya sekitar 2 kali
- Memiliki history pembelian menggunakan deals setidaknya 2 kali perbulan
- Memiliki income dan total spending terendah

Cluster 2 (Mid Spender)

- Didominasi oleh Middle Aged berusia 40-59 tahun berstatus menikah dan setidaknya memiliki 1 anak.
- Memiliki ketertarikan yang rendah dalam menerima campaign setidaknya 1 campaign
- Memiliki history pembelian terbanyak pada web sekitar 6 kali pembelian walaupun hanya memiliki angka web visit sekitar 4-5 kali per bulan
- Menggunakan deals setidaknya 2 kali dalam sebulan
- Memiliki income dan total spending diantara ketiga cluster

BUSINESS RECOMMENDATION

- Meningkatkan pelayanan pada store karena kebnyakan customer melaukan pembelian melalui store

- Cluster low spender dan paling beresiko untuk churn. Dapat dilakukan analisis lebih lanjut bagaimana meningkatkan rasio konversi visit to transaction. Mereka mempunyai jumlah visit yang paling tinggi tapi jarang melakukan transaksi. Hal ini dapat disebabkan oleh category product yang kurang, harga yang kurang cocok, biaya jasa kirim atau biaya sevice di platform yang tinggi, dll.'

- Karena cluster 1 (low spender) memiliki pengeluaran yang kecil, kita perlu membuat personalize ads, promosi atau campaign dengan produk yang murah/memiliki harga rendah melalui web yang dapat menarik cluster ini berbelanja di platform kita.

- Membuat membership program dengan 3 tingkatan berdasarkan cluster (Gold= High spender, Silver=Mid spender, Bronze= Low Spender) dengan previlages yang berbeda ditiap tingkatannya. Harapannya program ini dapa tmembuat customer lebih sering berbelanja.

POTENTIAL IMPACT

- Berdasarkan hasil analisis target akan lebih focus pada high spender dan mid spender yang memiliki pendapat tinggi sehingga lebih potensial melakukan pembelian lebih banyak
- High spender memiliki potensi GMV Rp. 711.356.000
- Mid spender memiliki potensi GMV Rp. 492.692.000
- jumlah cost yang dapat di save jika kita melakukan optimisasi promo di mid spender (asumsi: target reduce 50%) adalah Rp. 70.405.646