Predict Customer Personality to Boost Marketing Campaign by Using Machine Learning

Created by:
Indra Maulidin
indramaulidin@gmail.com
linkedin.com/in/indra-maulidin

I'm a Data Enthusiast and have a desire to create a career in Data Field especially as Data Scientist or Data Analyst. I have skills to use programming language such as Python and SQL. I have project experience doing Exploratory Data Analysis, Data Pre-Processing and creating Machine Learning Model.

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 ”

Conversion Rate Analysis Based on Income, Spending and Age

Data 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
```

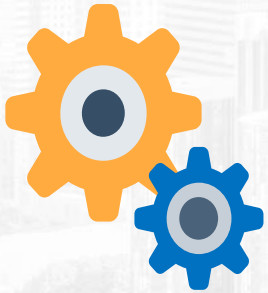
- Keseluruhan terdapat **2240 Baris** dan **30 Kolom**.
- Untuk kolom **Dt_Customer** perlu diubah tipe datanya menjadi ***datetime***.
- Pada kolom **Income** terdapat **24 Missing Values**.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

Rakamin Academy

Feature Engineering

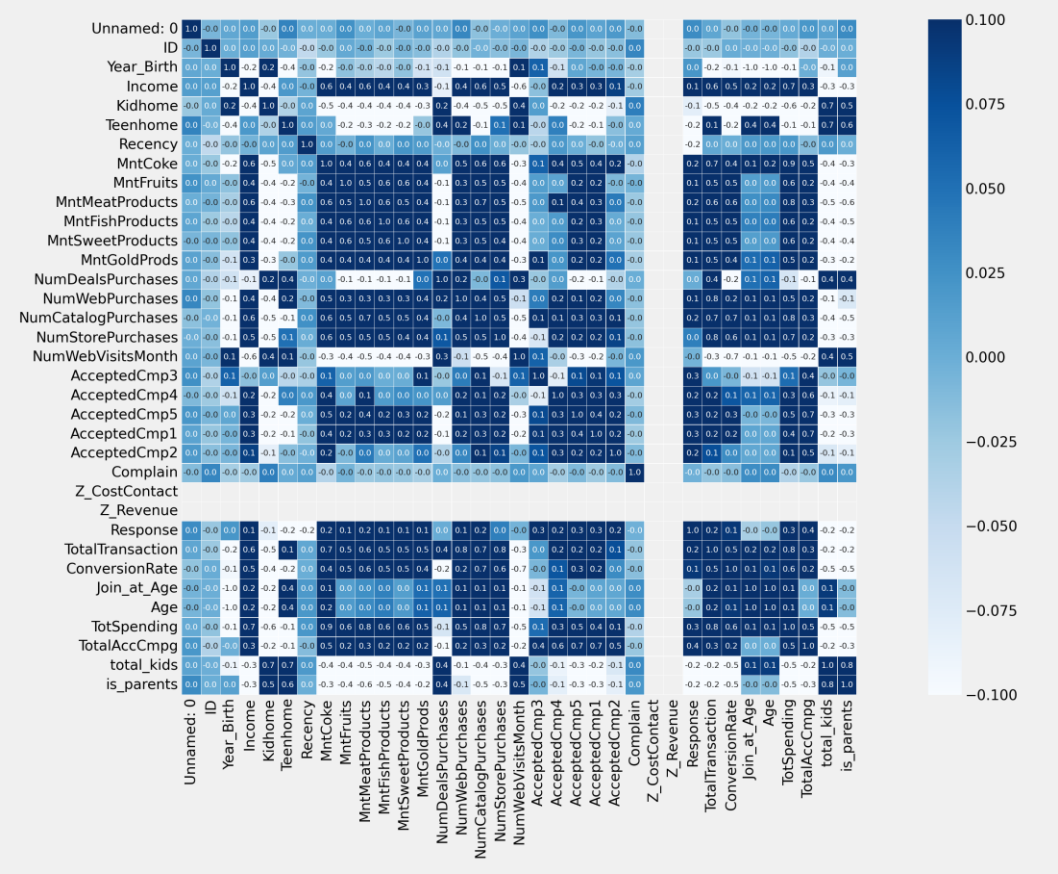


- **Conversion Rate**
- **Age & Age Category**
- **Total Spending**
- **Total Accepted Campaign**
- **Total Kids**
- **Is Parents**

Untuk selengkapnya, dapat melihat jupyter notebook disini

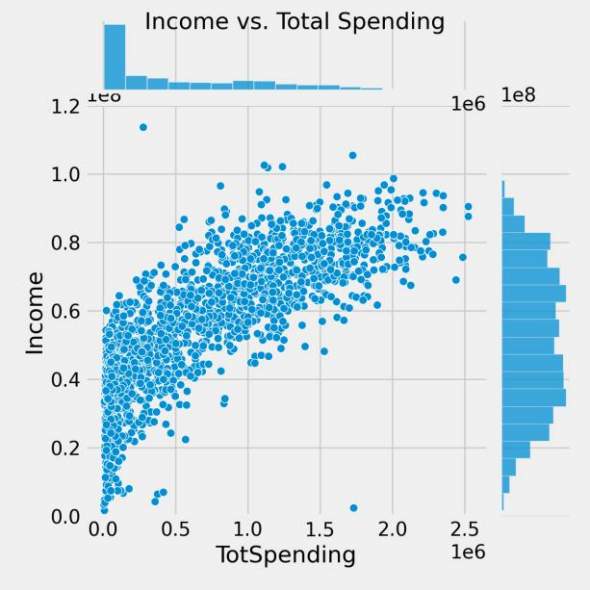
Conversion Rate Analysis Based on Income,

Multivariate Analysis

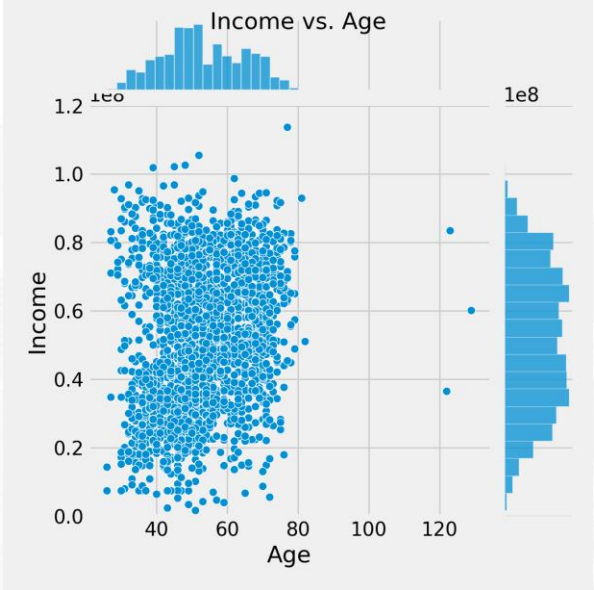


- Income dan Total Spending memiliki korelasi tinggi (0,7).
- Conversion Rate memiliki korelasi dengan Total Transaction dan Total Spending masing - masing sebesar 0,5 dan 0,6.
- Total Transaction dan Total Spending memiliki korelasi tinggi (0,8).
- Untuk Umur tidak memiliki korelasi yang tinggi dengan kolom apapun.

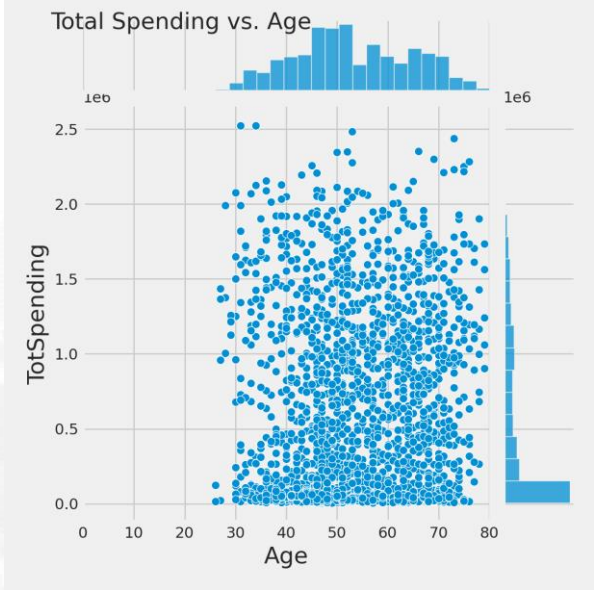
Conversion Rate Analysis Based on Income, Spending and Age



Grafik diatas menunjukkan bahwa untuk Income dan Total Spending saling berkorelasi (nilai korelasi 0.8).

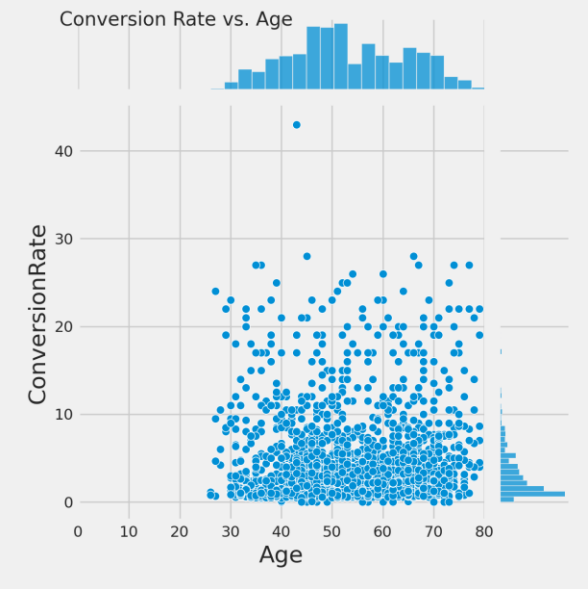


Grafik diatas menunjukkan bahwa untuk Income dan Age tidak saling berkorelasi (nilai korelasi 0.2).

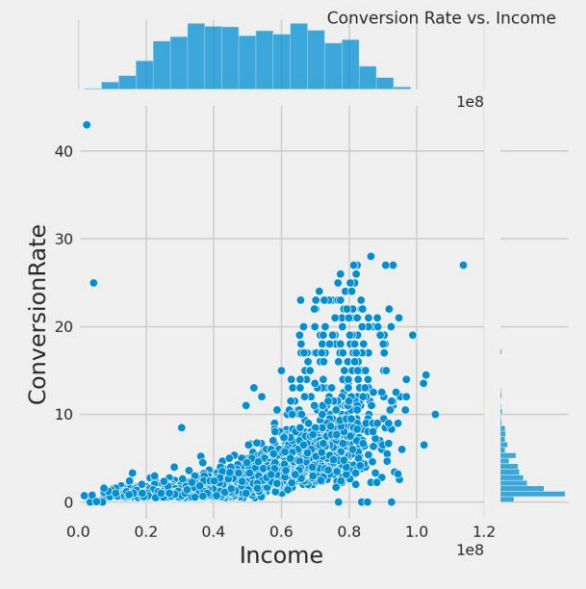


Grafik diatas menunjukkan bahwa untuk Total Spending dan Age tidak saling berkorelasi (nilai korelasi 0.1).

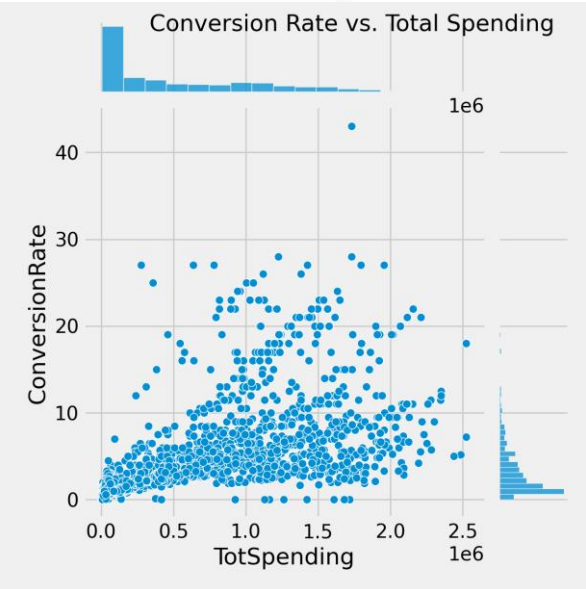
Conversion Rate Analysis Based on Income, Spending and Age



Grafik diatas menunjukkan bahwa untuk Conversion Rate dan Age tidak saling berkorelasi (nilai korelasi 0.1).



Grafik diatas menunjukkan bahwa untuk Conversion Rate dan Income saling berkorelasi (nilai korelasi 0.5).



Grafik diatas menunjukkan bahwa untuk Conversion Rate dan Total Spending saling berkorelasi (nilai korelasi 0.6).

Data Cleaning & Preprocessing

Handling Missing Value

```
df.isnull().sum()

Unnamed: 0              0
ID                      0
Year_Birth              0
Education               0
Marital_Status          0
Income                 24
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
TotalTransaction        0
ConversionRate          0
Join_at_Age             0
Age                     0
AgeCtg                  0
TotSpending             0
TotalAccCmpg            0
total_kids              0
is_parents              0
dtype: int64
```

- Untuk Kolom **Income** terdapat **24 Missing Values (1,07% Data).**
- Karena jumlahnya yang rendah maka kita bisa **menghilangkan (drop) Data Missing Value** tersebut.

Handling Duplicated Data

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

0
```

Pada Dataset **Tidak Terdapat Duplicated Data**.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Encoding

- **Label Encoding** → Feature **Education**
- **One Hot Encoding**
  - Feature **Marital Status**
  - Feature **Age Category**

Untuk selengkapnya, dapat melihat jupyter notebook disini

Scaling Data

```
df_scaled = df_preprocessing.drop(columns=['Unnamed: 0','ID','Year_Birth','Dt_Customer','Marital_Status',
                                           'Kidhome','Teenhome','Complain', 'Z_CostContact','Z_Revenue','Response','AgeCtg'])
```

```
nums = ['Income','Recency','MntCoke','MntFruits','MntMeatProducts','MntFishProducts','MntSweetProducts','MntGoldProds',
        'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth',
        'TotalTransaction', 'ConversionRate', 'Join_at_Age', 'Age', 'TotSpending', 'TotalAccCmpg', 'total_kids']

for col in nums:
  df_scaled[col] = StandardScaler().fit_transform(df_scaled[[col]])
```

Scalling Data menggunakan **Standardization**.

```
df_scaled.sample(5)
```

	Education	Income	Recency	MntCoke	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	Nu
751	4	-0.028976	1.485303	0.133160	-0.461382	-0.450407	-0.377014	-0.560821	-0.771475	-0.168231	-0.396043	
1145	3	-1.231444	0.241428	-0.818651	-0.662463	-0.633253	-0.651038	-0.609527	-0.790778	0.871658	-0.760962	
887	2	-0.397508	-0.518718	-0.762314	-0.662463	-0.735825	-0.687574	-0.658233	-0.752171	-0.168231	-0.760962	
295	3	0.147165	0.448740	2.775103	-0.662463	-0.387972	-0.687574	-0.341645	-0.076544	0.351713	-0.396043	
1934	2	0.939936	0.034115	0.103508	2.052132	0.080292	1.687300	0.267179	1.081674	-0.168231	0.698715	

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Selection

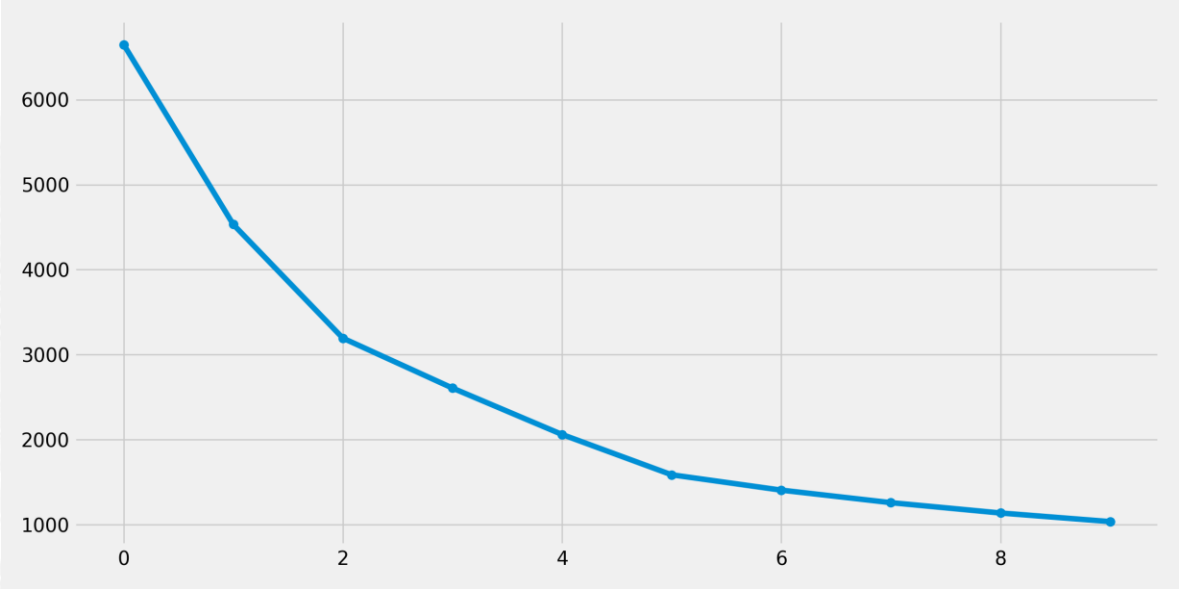
Akan dilakukan Feature Selection berdasarkan RFM Analysis dengan kolom yang akan dipilih yaitu :

- (Recency) Recency : Waktu terakhir Customer melakukan Transaksi.
- (Frequency) Conversion Rate : Frekuensi Transaksi/Kunjungan(Visits) yang dilakukan Customer.
- (Monetary) Income : Pendapatan Customer.

```
new_df_scaled = df_scaled[['Recency', 'Income', 'ConversionRate']]
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

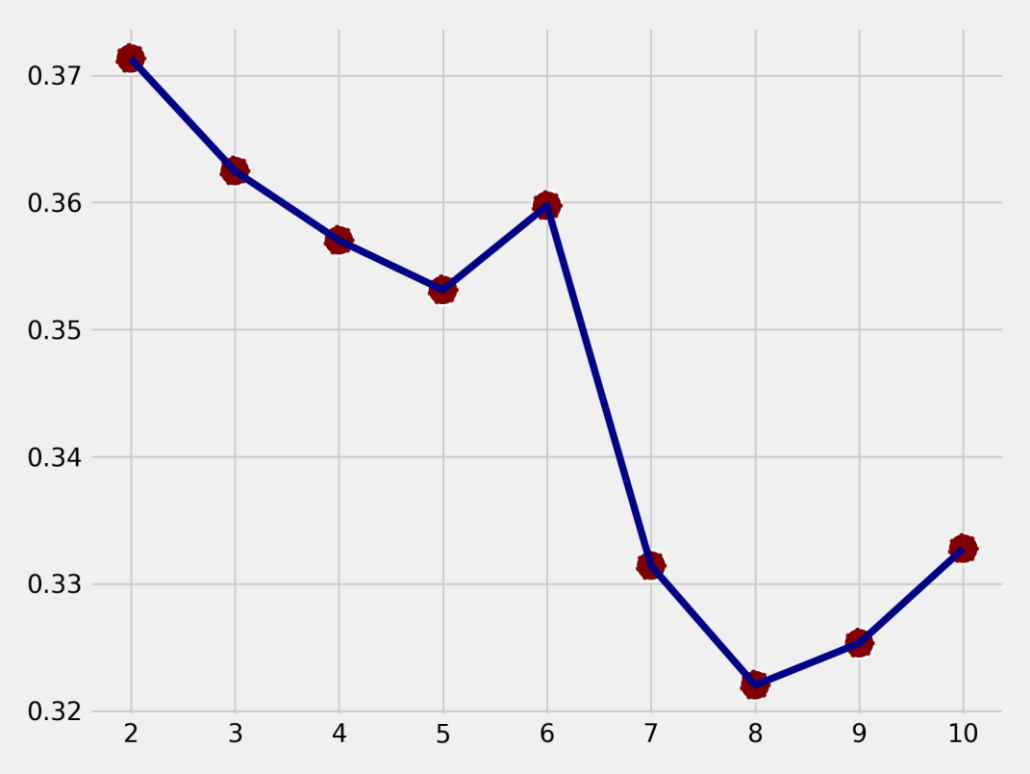
Elbow Method



- Berdasarkan grafik Elbow Method diatas kita dapat menggunakan Jumlah Cluster antara 4 s.d. 6.
- Pada Model ini akan digunakan Jumlah Cluster = 4.

Untuk selengkapnya, dapat melihat jupyter notebook disini

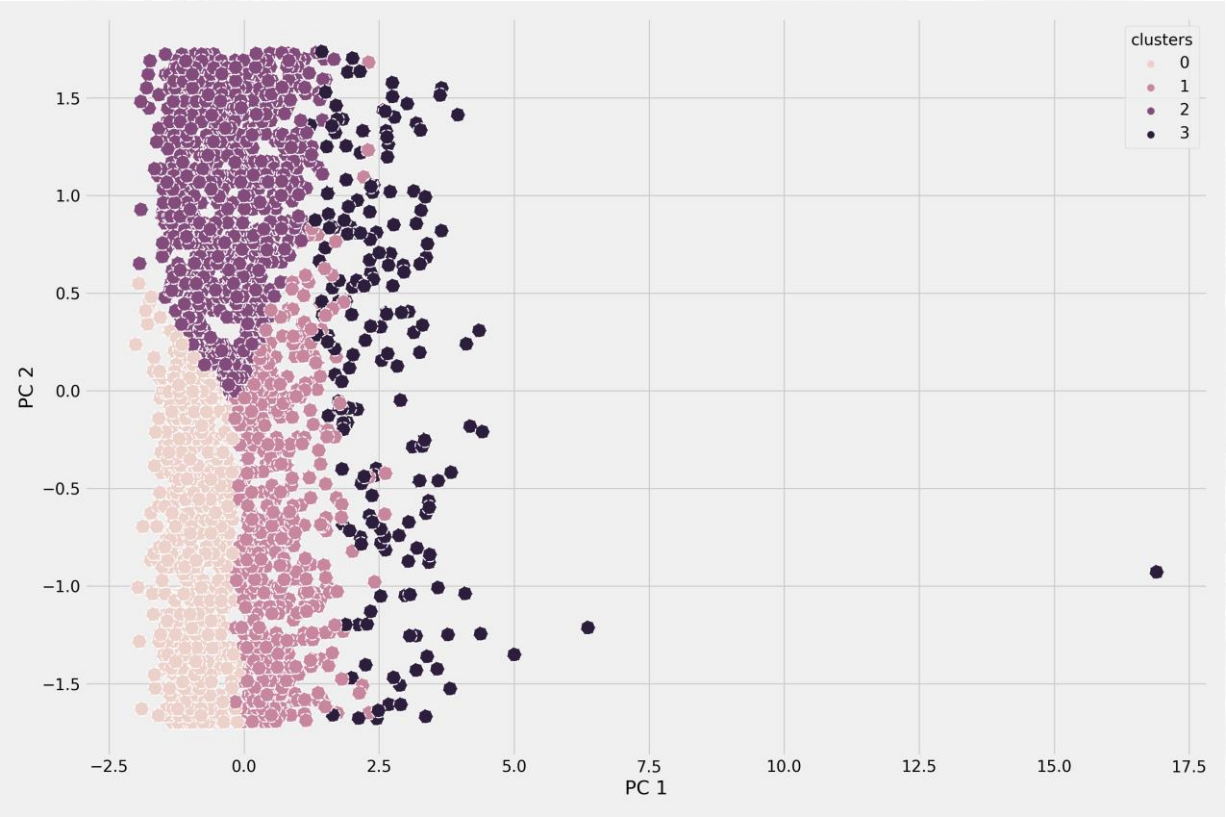
Silhouette Score



- Meskipun Jumlah Cluster 6 memiliki Score yg sedikit tinggi dibandingkan Jumlah Cluster 4, Jumlah Cluster 4 akan lebih menghasilkan Cluster yang lebih baik dalam memberikan Insight bagi Tim Bisnis/Marketing untuk meningkatkan Bisnis Perusahaan.
- Hal Tersebut akan diperlihatkan dalam Grafik Cluster.

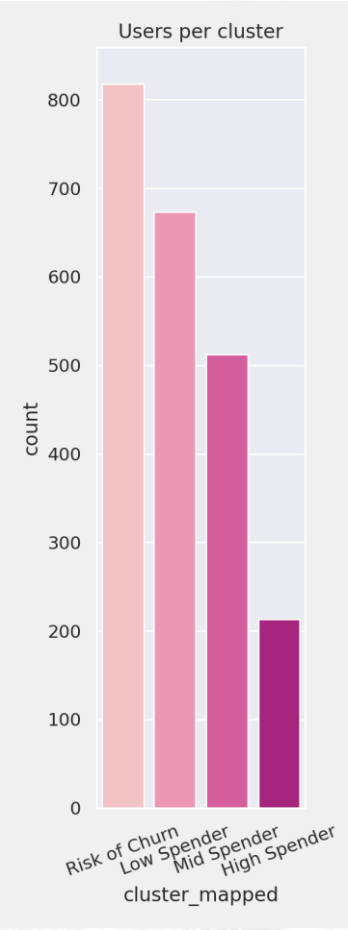
Untuk selengkapnya, dapat melihat jupyter notebook disini

Clustering



- Terlihat pada Grafik bahwa Customer dikelompokkan menjadi 4 Cluster.
- Jika kita perhatikan terdapat **1 Titik** yang **terletak jauh dari Titik lainnya**.
- **Titik tersebut** akan mempengaruhi Clustering yang akan kita buat jika jumlah Cluster diubah menjadi **lebih dari 4 Cluster**, dimana akan terdapat **sebuah Cluster dengan anggota Cluster hanya satu titik saja (titik terjauh tersebut)**.
- Hal tersebut tentu tidak akan bisa membantu Tim Bisnis/Marketing dalam memberikan Insight bagi mereka untuk menyusun strategi bisnis berdasarkan Cluster kedepannya.

Customer Personality Analysis for Marketing Retargeting



	Income		Recency		ConversionRate		TotSpending		NumDealsPurchases		NumWebVisitsMonth	
	mean	count	mean	count	mean	count	mean	count	mean	count	mean	count
clusters												
0	3.426563e+07	673	27.369985	673	1.469960	673	1.510327e+05	673	2.369985	673	6.717682	673
1	7.057009e+07	512	29.564453	512	5.629227	512	1.095934e+06	512	2.433594	512	4.232422	512
2	4.784832e+07	818	77.654034	818	2.834513	818	4.858557e+05	818	2.536675	818	5.881418	818
3	8.191233e+07	213	54.150235	213	16.618936	213	1.338432e+06	213	1.093897	213	1.352113	213

Untuk selengkapnya, dapat melihat jupyter notebook disini

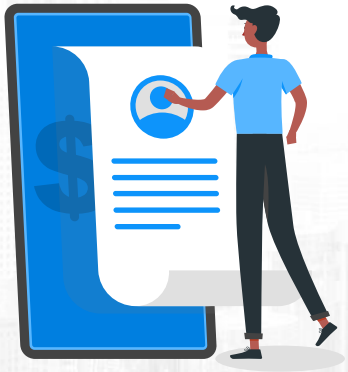
Customer Personality Analysis for Marketing Retargeting

Rakamin Academy

	Cluster 3	Cluster 1	Cluster 0	Cluster 2
Category	High Spender	Mid Spender	Low Spender	Risk of Churn
Description	• Merupakan Cluster dengan jumlah Customer **terendah yaitu 213 Orang** didominasi oleh **Customer Senior Adults (>55 tahun) dan Middle Adults (36-55 tahun)**, yang didominasi **telah/belum menikah** dan **belum mempunyai anak**. • Merupakan Customer dengan **Income & Total Spending tertinggi**, yang masing - masing sebesar **IDR 81 juta** untuk total **Income setahun**, dan **IDR 1,3 juta** untuk **total spending dalam setahun**. • Memiliki **Nilai Recency tertinggi kedua**. • Memiliki **Nilai Conversion Rate tertinggi** dan **Jumlah Kunjungan Web terendah**. • Merupakan Customer dengan **pembelian menggunakan Deals/Promo terendah**. • Merupakan **Customer Champions**.	• Merupakan Cluster dengan jumlah Customer **terendah kedua yaitu 512 Orang** didominasi oleh **Customer Senior Adults (>55 tahun) dan Middle Adults (36-55 tahun)**, yang didominasi **telah menikah** dan **sudah mempunyai anak**. • Merupakan Customer dengan **Total Spending & Income tertinggi kedua**, yang masing - masing sebesar **IDR 70 juta** untuk **total Income setahun**, dan **IDR 1 juta** untuk **total spending dalam setahun**. • Memiliki **Nilai Recency terendah kedua**. • Memiliki **Nilai Conversion Rate tertinggi kedua** dan **Jumlah Kunjungan Web terendah kedua**. • Merupakan Customer dengan **pembelian menggunakan Deals/Promo tertinggi kedua**. • Merupakan **Customer Potential**.	• Merupakan Cluster dengan jumlah Customer **terbanyak kedua yaitu 673 Orang** didominasi oleh **Customer Middle Adults (36-55 tahun)**, yang didominasi **telah menikah** dan **sudah mempunyai anak**. • Merupakan Customer dengan **Total Spending & Income terendah**, yang masing - masing sebesar **IDR 34 juta** untuk total **Income setahun**, dan **IDR 150K** untuk **total spending dalam setahun**. • Memiliki **Nilai Recency terendah**. • Memiliki **Nilai Conversion Rate terendah** dan **Jumlah Kunjungan Web tertinggi**. • Merupakan Customer dengan **pembelian menggunakan Deals/Promo tertinggi ketiga**. • Merupakan **Customer Need Attention**.	• Merupakan Cluster dengan jumlah Customer **terbanyak yaitu 818 Orang** didominasi oleh Customer **Middle Adults (36-55 tahun)**, yang didominasi **telah menikah** dan **sudah mempunyai anak**. • Merupakan Customer dengan **Total Spending & Income terendah kedua**, yang masing - masing sebesar **IDR 47 juta** untuk total **Income setahun**, dan **IDR 480K** untuk **total spending dalam setahun**. • Memiliki **Nilai Recency tertinggi**. • Memiliki Nilai **Conversion Rate terendah kedua** dan Jumlah **Kunjungan Web tertinggi kedua**. • Merupakan Customer dengan **pembelian menggunakan Deals/Promo tertinggi**. • Merupakan **Customer At Risk**.

Customer Personality Analysis for Marketing Retargeting

Recommendation



Untuk kelompok **High Spender** harus kita **pertahankan**, kita dapat berupaya **meningkatkan pelayanan** dan **membuat strategi bisnis khusus** untuk kelompok ini agar mereka tidak churn.

Untuk kelompok **Mid Spender** cukup sering mengunjungi web dan bertransaksi menggunakan deals/promo. Kita dapat menganalisis lebih lanjut agar dapat **mengoptimasikan deals/promo** yang diberikan dengan **mengurangi cost** yang dikeluarkan saat memberikan deals/promo untuk kelompok ini dengan **tetap membuat mereka berbelanja di platform kita**.

Untuk kelompok **Low Spender** dan **Risk of Churn** merupakan customer - customer yang **paling sering mengunjungi web tetapi tidak melakukan transaksi**. Hal tersebut dapat disebabkan oleh **Harga atau Promo yang kurang cocok** bagi kelompok - kelompok ini. Maka kita dapat melakukan **analisis lebih lanjut** untuk **meningkatkan Conversion Rate** dari kelompok - kelompok ini.

Potential Impact

Cost untuk Promo

40%

Lebih Rendah

Jika kita dapat **mempertahankan Kelompok High Spender** dan **Mid Spender** kita berpotensi mendapatkan **GMV** masing - masing sebesar **IDR 285 juta** dan **IDR 561 juta**.

Jika kita berhasil **mengoptimasikan promo** yang diberikan kepada kelompok **Mid Spender** (dengan **asumsi reduksi sebesar 40%),** maka kita dapat **mengurangi cost** sebesar **IDR 25 juta**.