Predict Customer Personality to boost marketing campaign by using Machine Learning

Created by:
M TOHAR SAGARA
Email : toharsagara@gmail.com
LinkedIn : M Tohar Sagara

A data enthusiast who graduated of Electrical Engineering majoring in Informatics and Computer Engineering. More than 5 (five) years in community empowerment, especially in rural communities. Understanding of data cleaning, exploratory data analysis, data visualization, and data storytelling. Fluent in SQL, Python, and C#, as well as familiar with Power BI and Looker Studio.

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

Feature Engineering

1. Age (Umur)

```
#Membuat kolom "Umur"

df['Age'] = 2022-df['Year_Birth']
```

2. Age_Group (Pengelompokkan Umur)

Dari kolom `Age`, kita akan mengelompokkan lagi umur ke beberapa grup sebagai berikut:

- **Children** = di bawah 12 tahun
- **Teen** = 12-16 tahun
- **Young Adult** = 17-25 tahun
- **Adult** = 26-44 tahun
- **Middle Age** = 45-59 tahun
- **Elderly** = 60 tahun ke atas

```
df['Age'].describe()
```

```
count   2240.00
mean      53.19
std       11.98
min       26.00
25%       45.00
50%       52.00
75%       63.00
max      129.00
Name: Age, dtype: float64
```

```
Age_Group = []

for i,kolom in df.iterrows():
  if kolom['Age'] < 45:
    Group = 'Adult'
  elif kolom['Age'] < 60:
    Group = 'Middle Age'
  else:
    Group = 'Elderly'
  Age_Group.append(Group)

df['Age_Group'] = Age_Group
```

*karena range umur dari 26-129,
maka pengelompokkan akan dimulai dari adult saja

3. Minorhome (Jumlah anak-anak di rumah)

```
df['Minorhome'] = (df['Kidhome'])+(df['Teenhome'])
```

*walaupun ada yang belum menikah, kita tetap hitung jumlah anak di rumah yang dapat menjadi potensi tanggungan dari customer (seperti adik, keponakan, dll)

4. Total_Spent (Total Pengeluaran Customer)

```
df['Total_Spent'] = df['MntCoke'] + df['MntFruits'] + df['MntMeat
Products'] + df['MntFishProducts'] + df['MntSweetProducts'] + df[
'MntGoldProds']
```

5. Total_Transactions (Total Transaksi)

```
df['Total_Transactions'] = df['NumDealsPurchases'] + df['NumWebPurch
ases'] + df['NumCatalogPurchases'] + df['NumStorePurchases']
```

6. Total_AcceptedCmp (Total Campaign yang diterima Cust)

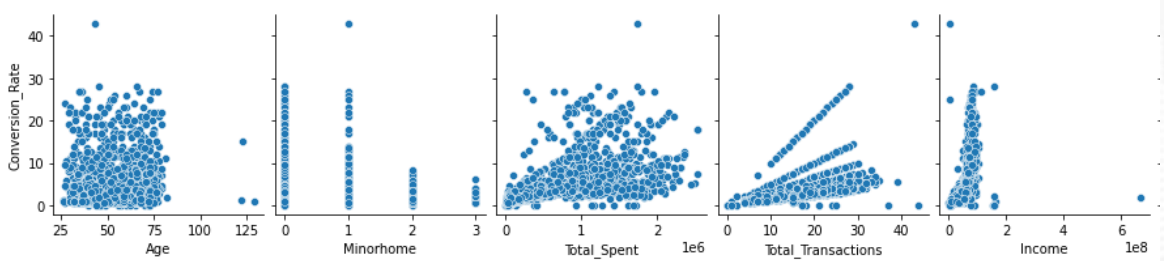
```
df['Total_AcceptedCmp'] = df['AcceptedCmp3'] + df['AcceptedCmp4']
+ df['AcceptedCmp5'] + df['AcceptedCmp1'] + df['AcceptedCmp2']
```

*asumsi total transaksi adalah pada bulan sebelumnya

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

7. Conversion_Rate (Jumlah Transaksi yang dilakukan Customer setiap Web Visit)

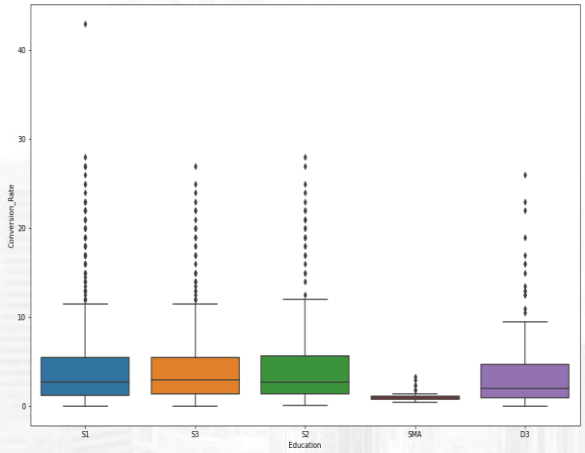


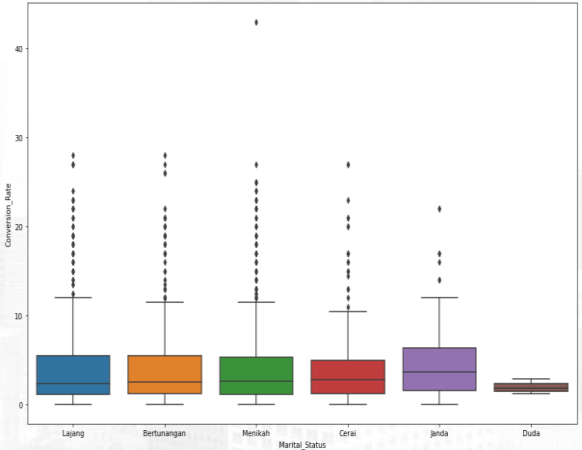
Hasil pengamatan hubungan kolom lainnya terhadap **Conversion_Rate** yaitu:
- Umur tidak berkorelasi atau tidak berpengaruh terhadap Conversion Rate.
- Customer dengan jumlah anak-anak serumah yang lebih banyak cenderung memmiliki Conversion Rate yang lebih kecil dibanding yang tidak memiliki anak/hanya memiliki 1 anak dalam serumah.
- Total_Spent dan Total_Transactions memiliki kecenderungan berbanding lurus atau pengaruh terhadap Conversion_Rate.
- Income memiliki korelasi (berbanding lurus) terhadap Conversion Rate.

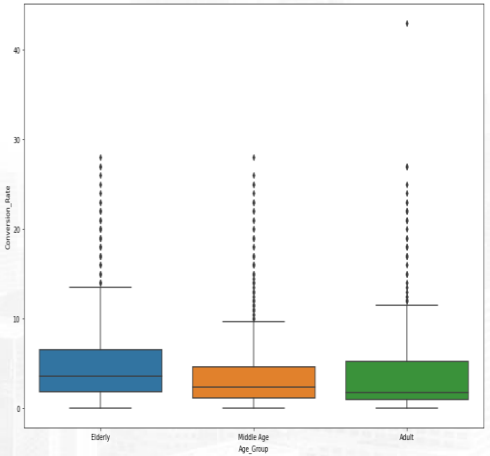
Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

Categorical Columns VS Conversion Rate







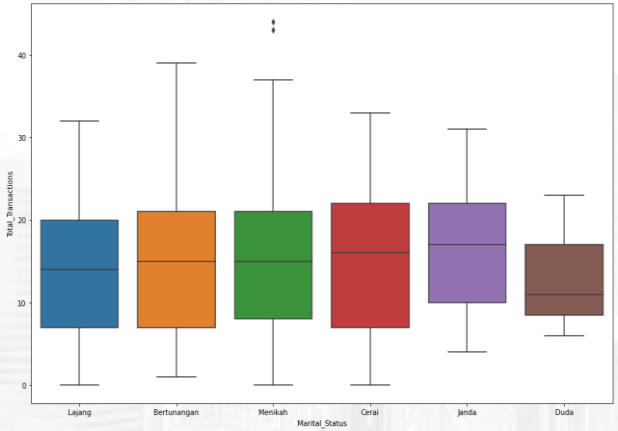
Hasil Pengamatan:

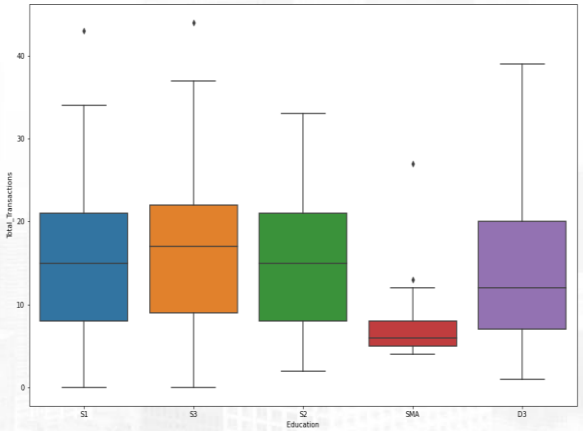
- Sebaran Conversion Rate pada pendidikan S1 , S2 dan S3 cukup merata,sedangkan sebaran untuk pendidikan D3 lebih sempit dan SMA sangat sempit
- Conversion Rate pada customer dengan status sebagai janda cenderung sedikit lebih tinggi dibanding yang lainnya. Untuk duda tidak dapat dijadikan sebagai perbandingan terhadap yang lainnya karena hanya berjumlah 3 orang.
- Sebaran Conversion rate pada customer kelompok Elderly (lanjut usia) sedikit lebih tinggi dibanding Adult dan kelompok Middle Age adalah kelompok yang sebaran conversion ratenya paling rendah.

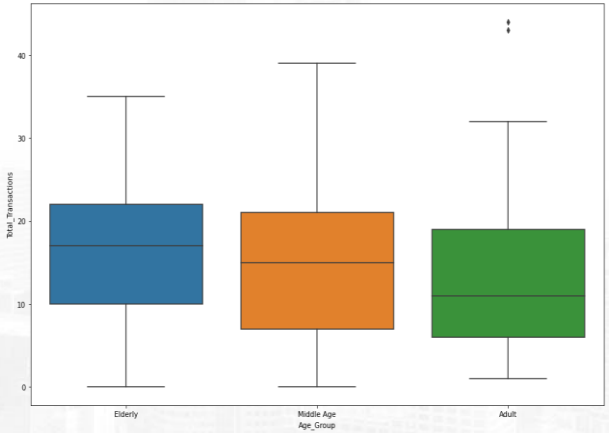
Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

Categorical Columns VS Total Transactions







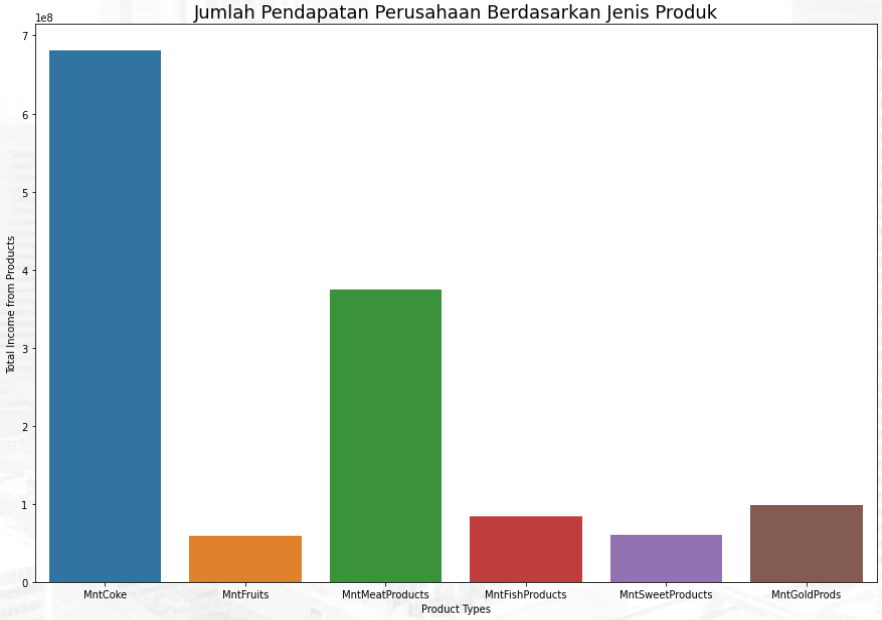
Hasil Pengamatan:

- Customer dengan status Bertunangan cenderung memiliki total transaksi yang paling tinggi dibanding customer dengan status lainnya.
- Customer dengan latar pendidikan SMA memiliki total transaksi yang cenderung lebih kecil sedangkan customer dengan latar pendidikan D3 dan S3 cenderung memiliki total transaksi sedikit lebih tinggi dibanding S1 dan S2.
- Customer Middle Age memiliki total transaksi yang cenderung lebih tinggi dibanding yang lainnya.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

Products Type

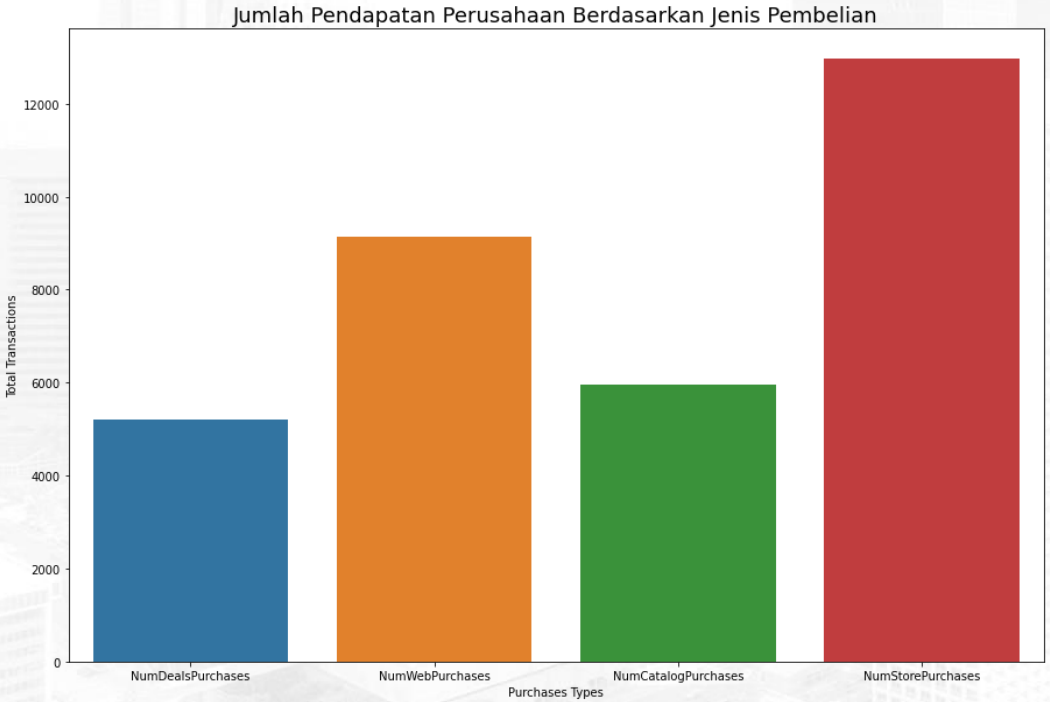


Jumlah pendapatan dari produk tipe cola adalah yang terbanyak dari semua produk yang ada, sedangkan jumlah pendapatan terkecil ada pada produk tipe buah dan manisan. Untuk marketing campaign selanjutnya, disarankan untuk melakukan campaign terhadap produk buah/manisan/ikan/emas karena keempat tipe produk tersebut jauh di bawah produk cola dan daging sehingga dalam rangka meningkatan revenue perusahaan, perlu adanya keseimbangan revenue masuk dari semua jenis produk yang ada. Untuk produk tipe cola dan daging tidak perlu menjadi fokus dari marketing campaign selanjutnya karena sudah cukup banyak penghasilan dari kedua produk tersebut dan sebaiknya lebih fokus kepada produk yang masih sedikit dalam memberikan revenue kepada perusahaan.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

Purchases Type

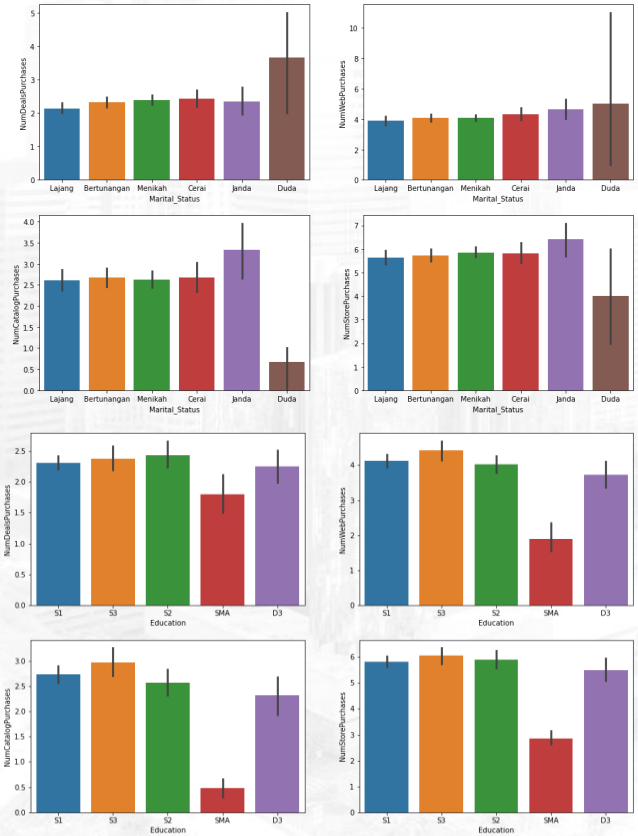


Jumlah pembelian terbanyak adalah melalui store dan web. Dikarenakan kedua metode pembelian tersebut sudah cukup terkenal, maka sebaiknya diberi perhatian lebih untuk pembelian tipe Deals dan Catalog yang masih di bawah tipe Web dan Store. Rekomendasi yang dapat diberikan yaitu seperti produk Deals yang ditawarkan adalah tipe produk yang bagaimana, apakah memang dibutuhkan orang banyak atau tidak. Jika merupakan bundle, apakah isi bundle tersebut cukup menarik customer atau tidak, hal seperti itu harus dipelajari lebih lanjut. Untuk pembelian tipe Catalog, dapat diperhatikan seperti design dari catalog apakah cukup menarik customer atau tidak.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

Categorical Columns VS Purchases Type



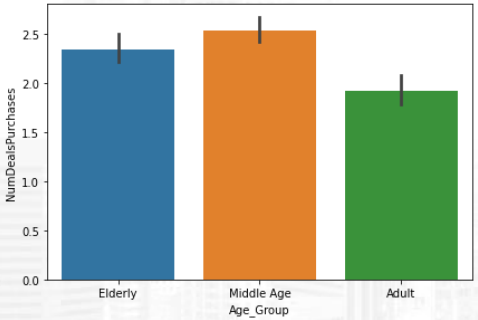
- Distribusi jumlah transaksi dari tiap metode pembelian berdasarkan Marital_Status (Status Pernikahan) cukup merata kecuali Duda.

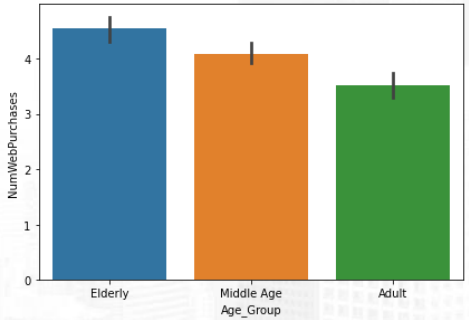
- Dapat dilihat bahwa Tingkat Pendidikan SMA selalu menjadi yang terendah dari yang lain. .

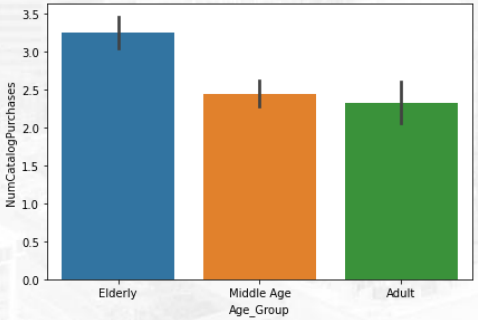
Untuk selengkapnya, dapat melihat jupyter notebook disini

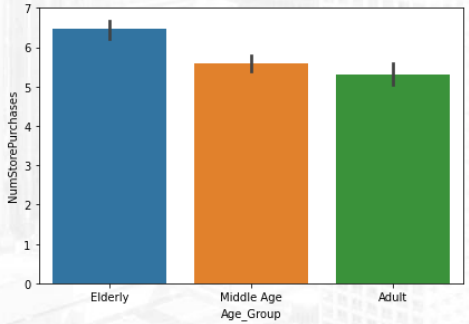
Conversion Rate Analysis Based on Income, Spending and Age

Categorical Columns VS Purchases Type







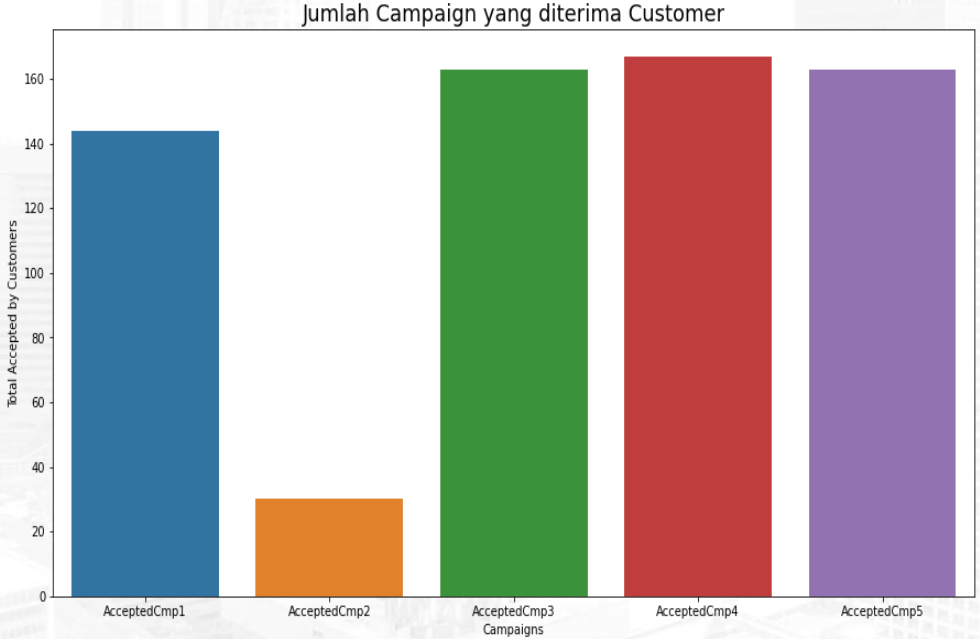


Distribusi jumlah transaksi dari tiap metode pembelian berdasarkan kelompok usia customer tidak jauh berbeda di setiap kelompok usia.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

Total Accepted Campaigns



- Jumlah customer yang menerima pada campaign kedua jauh menurun dibanding pada campaign pertama.
- Jumlah customer yang menerima campaign ketiga naik signfikan dari campaign kedua bahkan melebihi campaign pertama.
- Jumlah customer yang menerima campaign ketiga ,keempat dan kelima cenderung cenderung sama

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Cleaning & Preprocessing

1. Handling Missing Value

```
Income                1.07
ID                    0.00
Z_Revenue             0.00
AcceptedCmp4          0.00
AcceptedCmp5          0.00
AcceptedCmp1          0.00
AcceptedCmp2          0.00
Complain              0.00
Z_CostContact         0.00
Response              0.00
NumWebVisitsMonth     0.00
Age                   0.00
Age_Group             0.00
Minorhome             0.00
Total_Spent           0.00
Total_Transactions    0.00
Total_AcceptedCmp     0.00
AcceptedCmp3          0.00
NumStorePurchases     0.00
Year_Birth            0.00
MntCoke               0.00
Education             0.00
Marital_Status        0.00
Kidhome               0.00
Teenhome              0.00
Dt_Customer           0.00
Recency               0.00
MntFruits             0.00
NumCatalogPurchases   0.00
MntMeatProducts       0.00
MntFishProducts       0.00
MntSweetProducts      0.00
MntGoldProds          0.00
NumDealsPurchases     0.00
NumWebPurchases       0.00
Conversion_Rate       0.00
dtype: float64
```

Kolom yang memiliki nilai null hanya pada kolom **Income** sebesar 1.07%.
maka nilai null pada **Income** di **drop**

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Cleaning & Preprocessing

2. Handling Duplicated Value

```
print('Jumlah data duplikat =',df_1.duplicated().sum())
```

```
Jumlah data duplikat = 0
```

Tidak ada baris data yang duplikat.

3. Drop Rows with Odd Values

```
df_1[df_1['Age']>=100]
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntCoke	...	Z_CostContact	Z_Revenue	Response	Age	Age_Group
192	7829	1900	D3	Cerai	36640000.00	1	0	26-09-2013	99	15000	...	3	11	0	122	Elderly
239	11004	1893	D3	Lajang	60182000.00	0	1	17-05-2014	23	8000	...	3	11	0	129	Elderly
339	1150	1899	S3	Bertunangan	83532000.00	0	0	26-09-2013	36	755000	...	3	11	0	123	Elderly

3 rows × 36 columns

Ketiga baris data tersebut di drop karena kemungkinan terjadi kesalahan input data umur.

Untuk selengkapnya, dapat melihat jupyter notebook disini

4. Feature Encoding

`Education` -> Label Encoding

`Marital_Status` -> One Hot Encoding

`Age_Group` -> One Hot Encoding

a. Label Encoding

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

Dilakukan Label Encoding pada kolom **Education**

Untuk selengkapnya, dapat melihat jupyter notebook disini

b. One Hot Encoding

```
34  Marital_Status_Bertunangan   2213 non-null   uint8
35  Marital_Status_Cerai         2213 non-null   uint8
36  Marital_Status_Duda          2213 non-null   uint8
37  Marital_Status_Janda         2213 non-null   uint8
38  Marital_Status_Lajang        2213 non-null   uint8
39  Marital_Status_Menikah       2213 non-null   uint8
40  Age_Group_Adult              2213 non-null   uint8
41  Age_Group_Elderly            2213 non-null   uint8
42  Age_Group_Middle Age         2213 non-null   uint8
```

Dilakukan OHE untuk kolom **Marital_Status** dan **Age_Group**

Untuk selengkapnya, dapat melihat jupyter notebook disini

Rakamin Academy

5. Feature Transformation (Standardization)

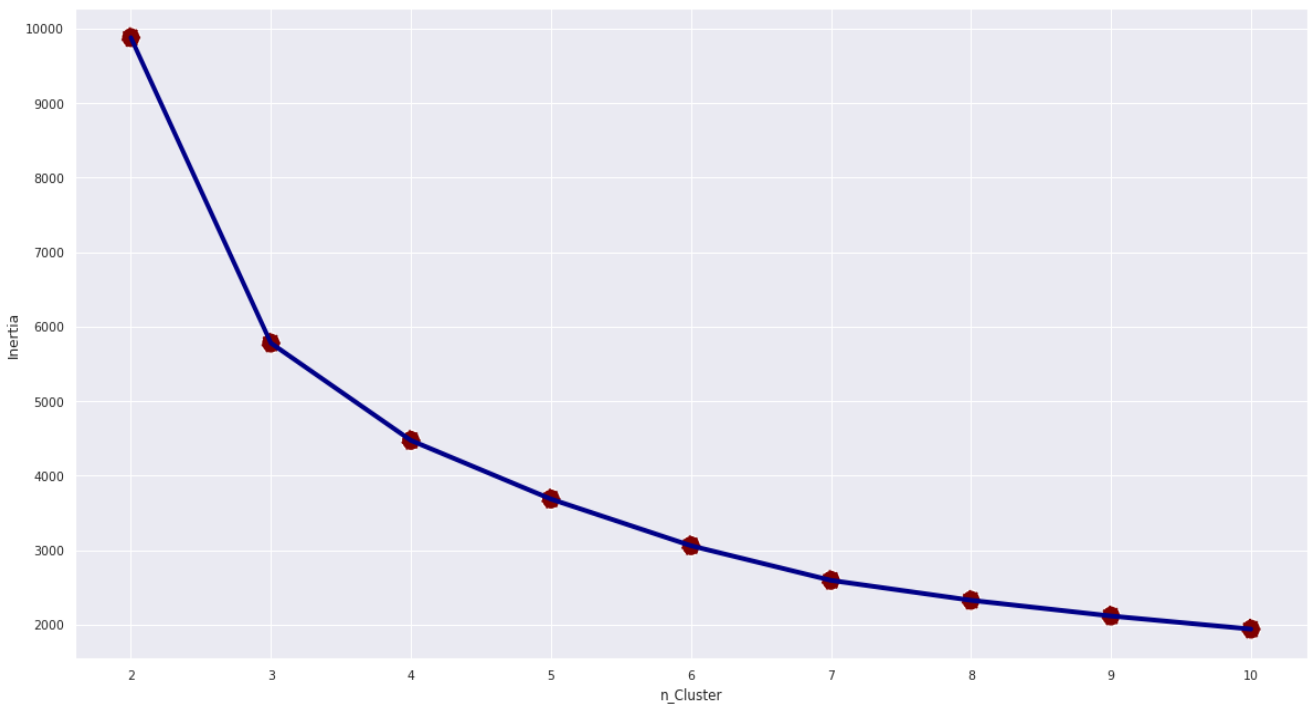
Feature yang distandardisasi yaitu:

- Income
- Kidhome
- Teenhome
- Recency
- MntCoke
- MntFruit
- MntMeatProducts
- MntFishProducts
- MntSweetproducts
- MntGoldProds
- NumDealsPurchases
- NumWebPurchases
- NumCatalogPurchases
- NumStorePurchases,
- NumWebVisitsMonth
- Age, Minorhome
- Total_Spent
- Total_Transactions
- Total_AcceptedCmp
- Conversion_Rate

	Income	Kidhome	Teenhome	Recency	MntCoke	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	...
0	0.23	-0.82	-0.93	0.31	0.98	1.55	1.69	2.45	1.48	0.85	...
1	-0.23	1.04	0.91	-0.38	-0.87	-0.64	-0.72	-0.65	-0.63	-0.73	...
2	0.77	-0.82	-0.93	-0.80	0.36	0.57	-0.18	1.34	-0.15	-0.04	...
3	-1.02	1.04	-0.93	-0.80	-0.87	-0.56	-0.66	-0.50	-0.59	-0.75	...
4	0.24	1.04	-0.93	1.55	-0.39	0.42	-0.22	0.15	-0.00	-0.56	...

Untuk selengkapnya, dapat melihat jupyter notebook disini

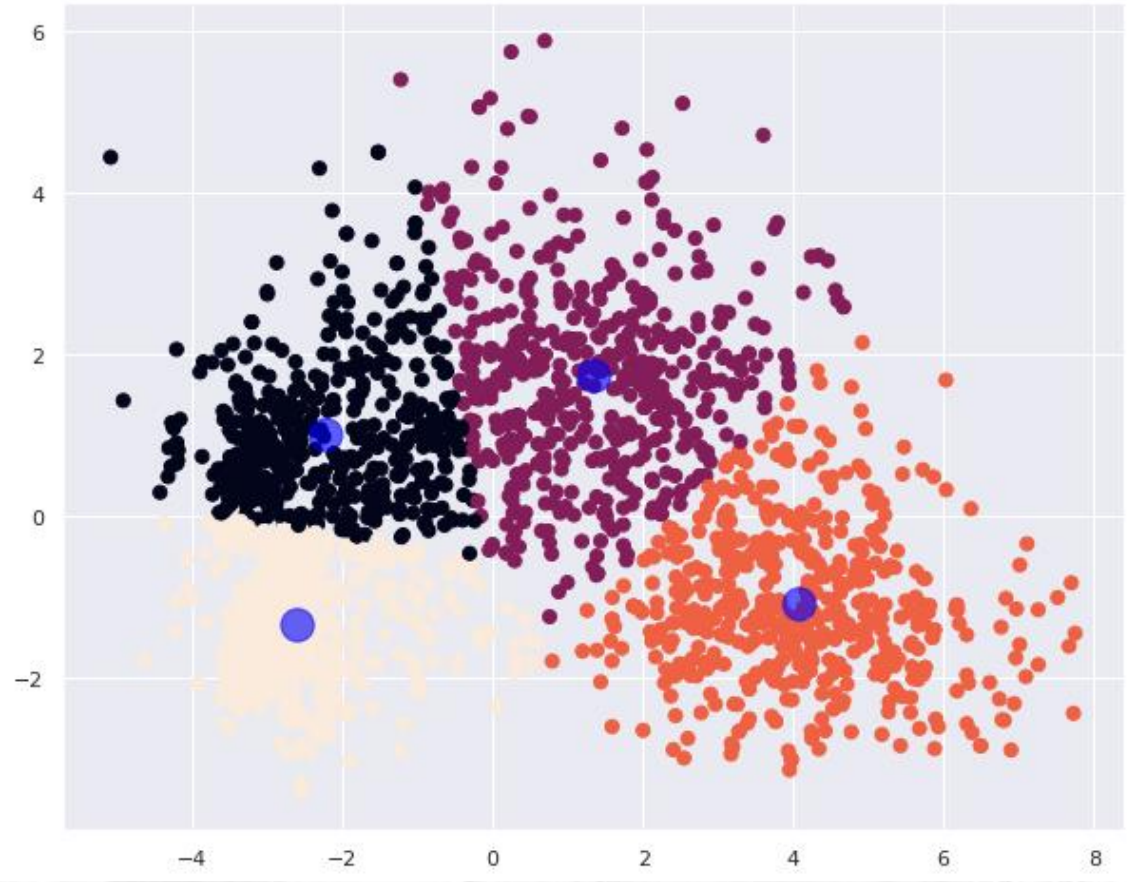
1. Elbow Method



Jumlah cluster yang akan digunakan adalah 5
karena dari 6 sampai 10 nilai inertia turun hanya 1000 dari rentan 3000 - 2000.

Untuk selengkapnya, dapat melihat jupyter notebook disini

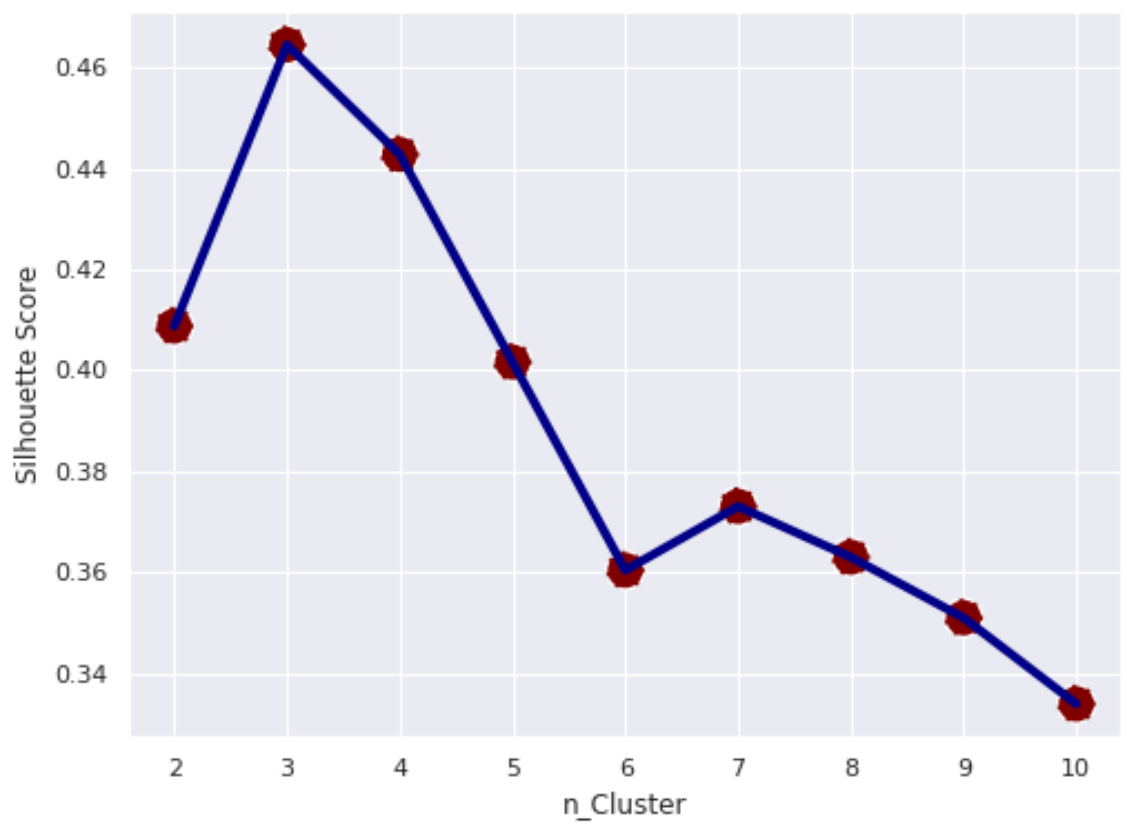
2. Clustering



Dalam melakukan clustering,
dilakukan PCA terlebih dahulu karena
terdapat beberapa kolom dengan korelasi kuat.

Untuk selengkapnya, dapat melihat jupyter notebook disini

3. Silhouette Score



Pada perhitungan silhouette score dengan metode pengukuran **Mahalanobis distance** karena masih terdapat beberapa kolom kategorikal yang memiliki nilai range yang berbeda dengan nilai numerik yang telah distandarisasi.
Dapat terlihat bahwa untuk jumlah **cluster 5 memiliki nilai score diatas 0.4.**

Untuk selengkapnya, dapat melihat jupyter notebook disini

4. Implementasi Clustering

	K_Cluster	Income	Kidhome	Teenhome	Minorhome	Total_Spent	Total_Transactions	Total_AcceptedCmp	NumDealsPurchases
0	0	-0.37	0.59	0.84	1.03	-0.70	-0.47	-0.29	0.44
1	1	0.37	-0.49	0.78	0.21	0.45	0.96	0.02	0.63
2	2	1.01	-0.77	-0.75	-1.10	1.30	0.75	0.65	-0.60
3	3	-0.82	0.56	-0.69	-0.10	-0.86	-1.01	-0.32	-0.37

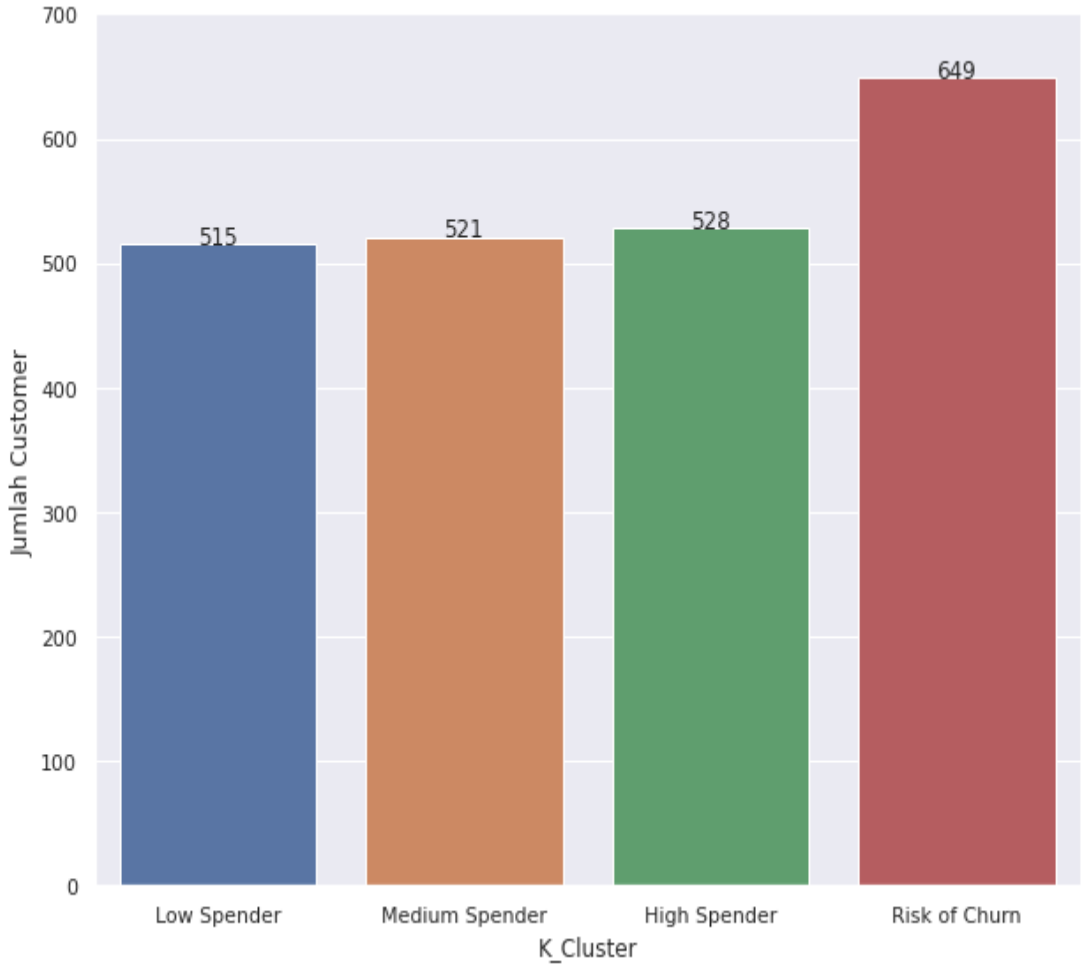
Berdasarkan rata-rata Income, Minorhome, Total_Spent, Total_Transactions maka keempat cluster kelompok tersebut dapat dibagi menjadi:

- 0 -> Low Income, Most Likely to Have Minors at Home (Kids and Teens), Low Spending, Low Transactions -> **Low Spender**
- 1 -> Medium Income, Likely to Have Minors at Home (Mostly Teens), Medium Spending, High -> **Medium Spender**
- 2 -> High Income, Most Unikely to Have Minors at Home, High Spending, High Transactions -> **High Spender**
- 3 -> Very Low Income, Likely to Have Minors at Home (Mostly Kids), Very Low Spending, Very Low Transactions -> **Risk of Churn**

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

1. Jumlah Customer



Marketing campaign harus berfokus pada Risk of Churn karena kelompok ini yang terbanyak dibandingkan kelompok lainnya yaitu **649 customer.**

Business Recommendations :

- **Risk of Churn** -> Fokus memberikan promo (baik diskon ataupun cashback) kepada kelompok ini dan dapat lebih didalami lagi karakteristik customer di cluster ini karena kelompok ini yang akan menjadi target marketing campaign.
- **High, Medium dan Low Spender** -> Untuk kelompok ini, promo diberikan untuk kondisi-kondisi tertentu saja, sebagai contoh perayaan hari tertentu.

Untuk selengkapnya, dapat melihat jupyter notebook disini

2. Income vs Conversion Rate

Income berkorelasi dengan Conversion Rate.

Kelompok High Spender memiliki Income dan Conversion rate yang cenderung paling besar, disusul oleh Medium Spender, Low Spender dan paling rendah adalah yang berada pada kelompok Risk of Churn.

Business Recommendations :
Dapat memberikan promo tanggal tertentu seperti akhir bulan atau awal bulan saat customer umumnya gajian.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

3. Income vs Conversion Rate

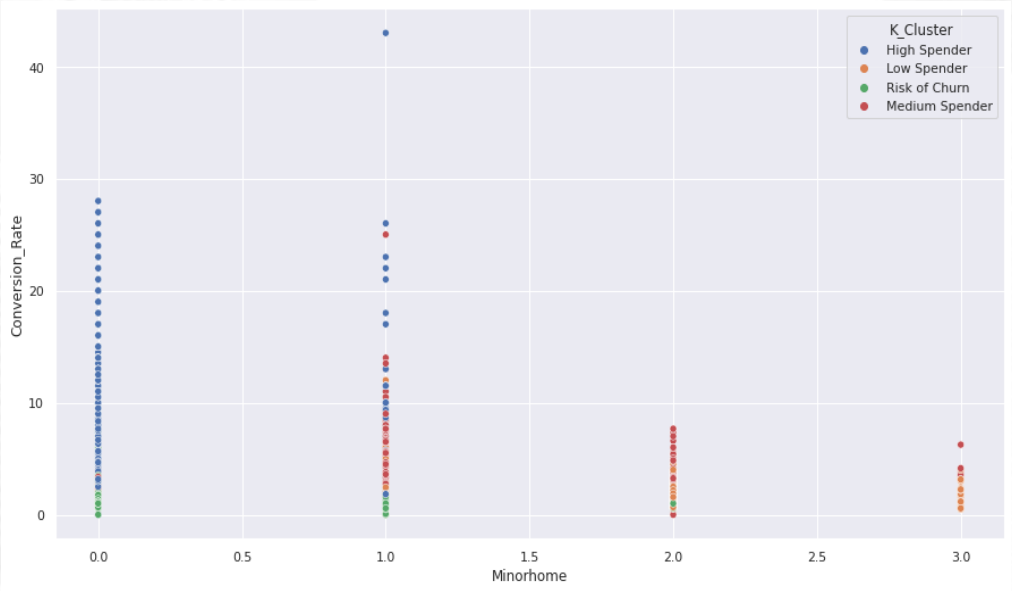
Income berkorelasi dengan Total Spend

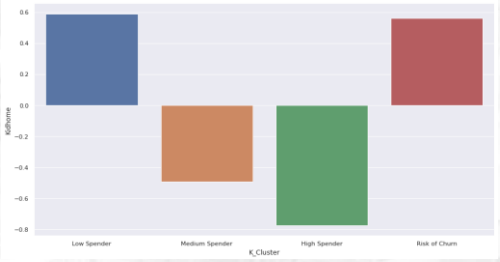
Kelompok High Spender dan medium Spender memiliki Income dan Total Spend yang cenderung paling besar, disusul Low Spender dan paling rendah adalah yang berada pada kelompok Risk of Churn.

Business Recommendations :
Dapat memberikan promo tanggal tertentu seperti akhir bulan atau awal bulan saat customer umumnya gajian.

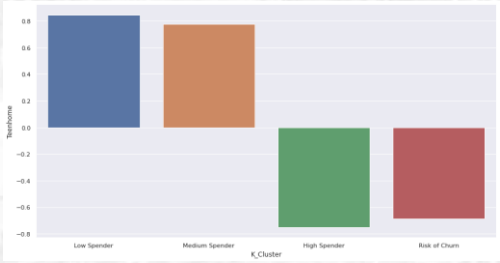
Untuk selengkapnya, dapat melihat jupyter notebook disini

4. Total Minors vs Conversion Rate





Total Anak dari setiap Cluster



Total Remaja dari setiap Cluster

Semakin sedikit minors (anak) di rumah, maka conversion rate customer akan semakin tinggi

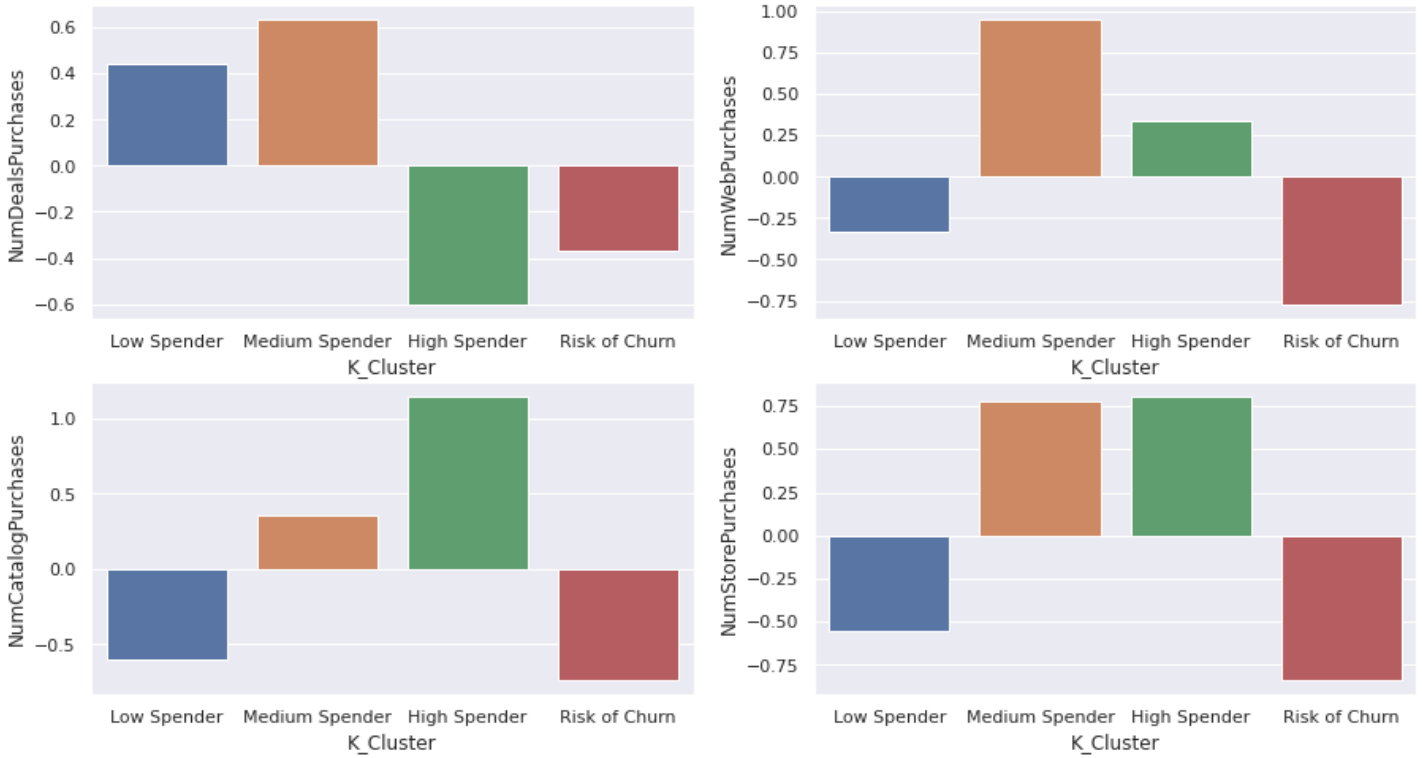
kebanyakan dari kelompok High Spender tidak memiliki atau hanya memiliki 1 anak. Rekomendasi bisnis yang dapat diberikan yaitu karena minimnya conversion rate untuk customer yang memiliki 2 anak atau lebih,

Business Recommendations :
menambah produk-produk seperti perlengkapan anak atau mainan untuk menarik perhatian customer yang memiliki anak.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

5. Tipe Purchase dari setiap Cluster



Business Recommendations :
Membuat skema bundle product untuk produk-produk yang sering dibeli atau dicari oleh customer

Untuk selengkapnya, dapat melihat jupyter notebook disini