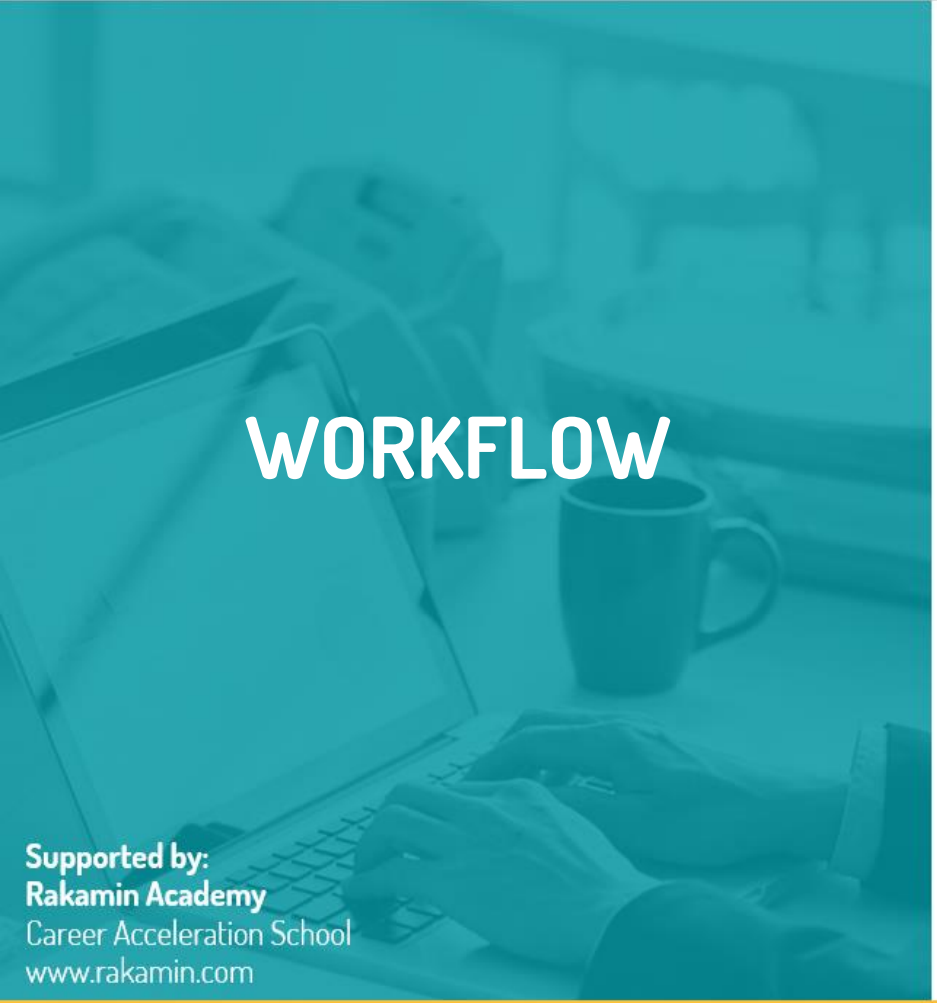
Predictive Customer Personality to Boost Marketing Campaign by Using Machine Learning

Created by:
Fikri Diva Sambasri
fikri.sambasri@gmail.com
http://www.linkedin.com/in/fikridivasambasri

I'm final year Informatics Engineering student at Dian Nuswantoro University, Semarang with experience in working on several projects related to Data Science using Python and several data science tools. After i finish my studies, I aspire to work as a data scientist or data analyst.

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



WORKFLOW

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

Rakamin Academy

- EDA
- Data Cleaning & Preprocessing
- Modelling
- Interpretasi Model

Rakamin Academy

EDA

- Info Dataset
- Feature Engineering
- Multivariate Analysis (Corelation Features)
- Conversion Rate Analysis Based on Income, Spending, and Age

Info Dataset

Informasi Dataset

Terdapat 2240 baris data pada dataset dengan 27 data numerik dan 3 data kategorikal. Terdapat missing value pada variable / fitur Income sebanyak 24 baris.

```
# Read data
df = pd.read_csv('/content/drive/MyDrive/Mini Project 3/marketing_campaign_data.csv', sep=',')

# Cek informasi dataset
print(df.info())

# Cek nilai missing value
print(df.isna().sum())
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Engineering

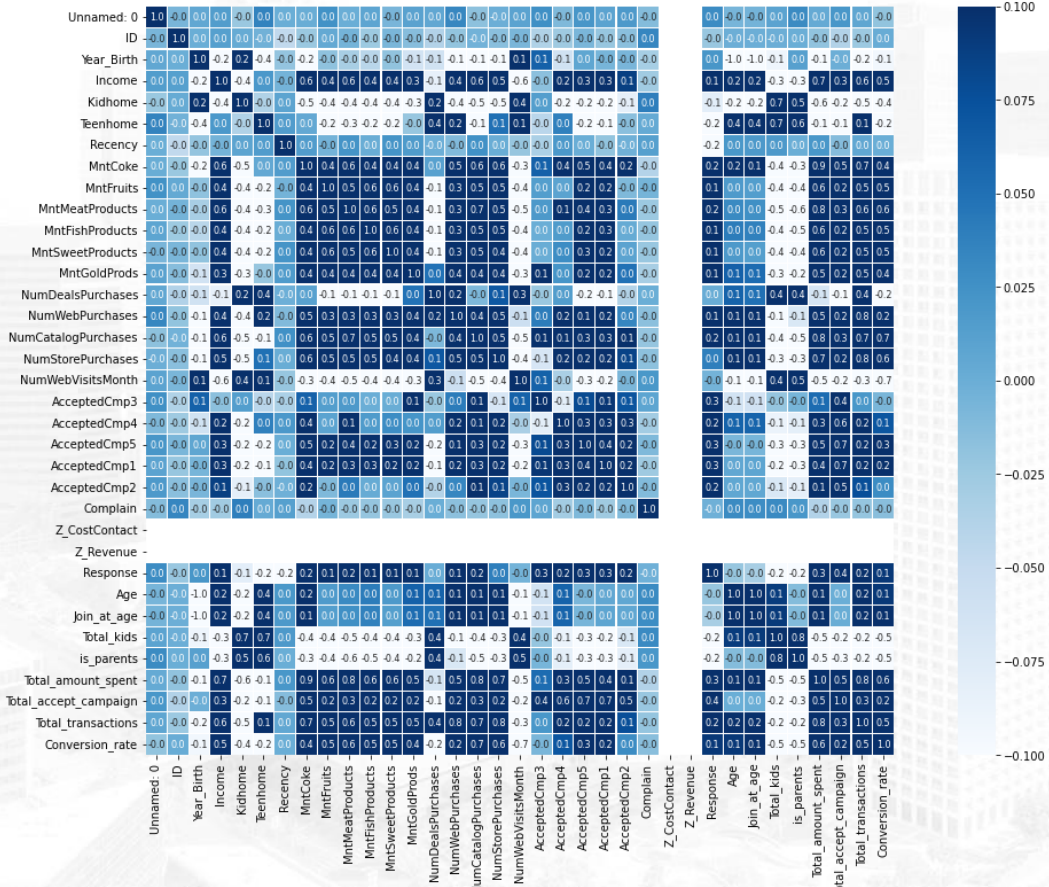
Feature Engineering

Melakukan beberapa feature engineering sebagai berikut:

1. Menghitung umur customer
2. Menghitung umur diwaktu customer bergabung
3. Menghitung total anak
4. Menentukan apakah customer sudah menjadi orang tua atau belum
5. Mengelompokkan umur customer ke beberapa kategori seperti anak-anak, remaja, dewasa, dan lain-lain
6. Menghitung total pengeluaran yang dilakukan customer
7. Menghitung total campaign yang di accept customer
8. Menghitung total transaksi yang dilakukan oleh customer
9. Menghitung conversion rate customer

Untuk selengkapnya, dapat melihat jupyter notebook disini

Corelation Features



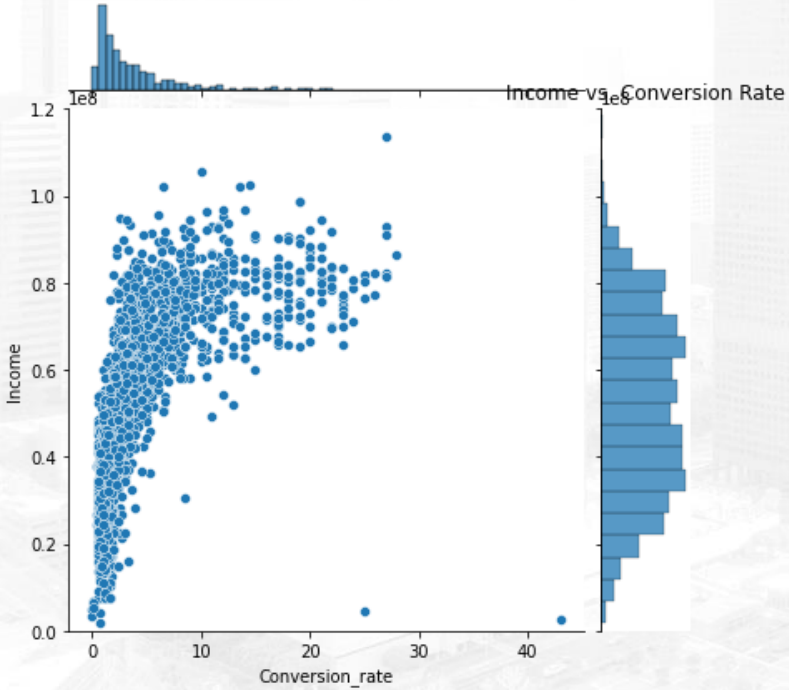
Pada fitur convertion rate terdapat nilai bersifat korelasi positif dengan fitur income, total amount spending, dan umur. Sehingga nantinya akan dilakukan analisis convertion rate terhadap ketiga fitur tersebut.

Untuk selengkapnya, dapat melihat jupyter notebook disini

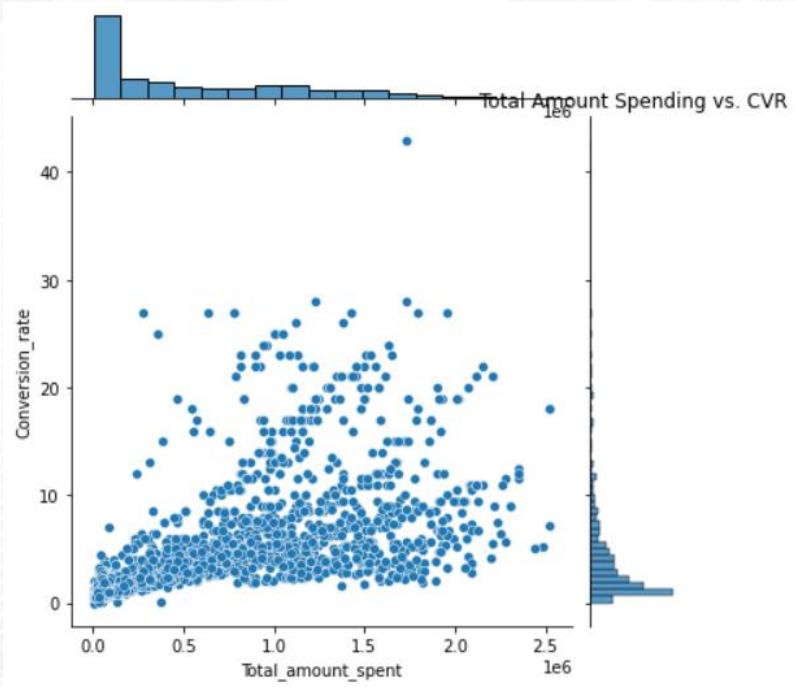
Conversion rate analysis based on income, spending, and age

Rakamin Academy

Income and Conversion Rate



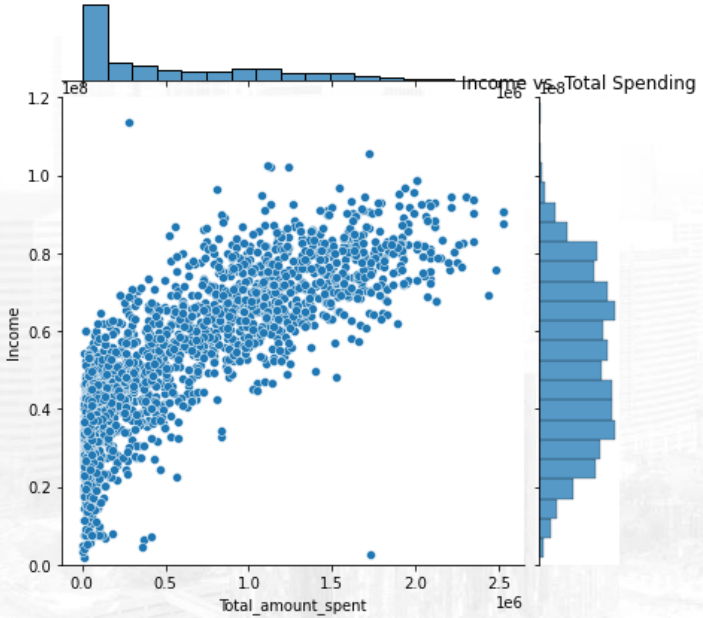
Total Amount Spending and Conversion Rate



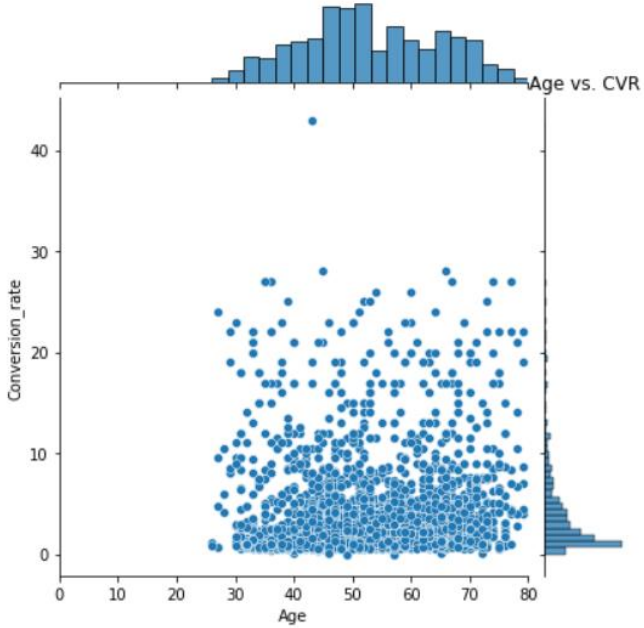
Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion rate analysis based on income, spending, and age

Income and Total Amount Spending



Age and Conversion Rate



Insight / Analisis:

- Hasil analisis di atas didapatkan semakin besar income yang dimiliki customer, terdapat kecenderungan untuk mempunyai pengeluaran lebih banyak dan mempunyai total pengeluaran yang lebih besar pula di platform kita. Hal ini tidak berlaku untuk fitur umur.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Rakamin Academy

Data Cleaning & Preprocessing

- Cek Missing Value
- Cek Duplicated Data
- Membuang Data yang Tidak Perlu
- Feature Encoding
- Standarisasi

Cek Missing Value

Missing Value

Terdapat 1 kolom yang memiliki nilai missing value, dengan jumlahnya 24 baris pada kolom Income. Kita dapat menghapus baris tersebut karena jumlahnya sedikit.

```
df.isna().sum()
```

```
Unnamed: 0          0
ID                  0
Year_Birth          0
Education           0
Marital_Status      0
Income             24
```

```
df = df.dropna()
```

```
df.isna().sum()
```

```
Unnamed: 0          0
ID                  0
Year_Birth          0
Education           0
Marital_Status      0
Income              0
Kidhome             0
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Cek Duplicated Data

Duplicated Data

Pada proses data cleaning dilakukan deteksi data duplikat, jika terdapat data duplikat maka akan dihapus. Untuk dataset ini tidak terdapat data duplicated.

```
df.duplicated().any()
```

```
False
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Membuat Data yang Tidak Perlu

Drop Data

Pada tahap ini dilakukan drop / menghapus data yang tidak diperlukan untuk proses modelling. Ada beberapa data yang dihapus diantaranya adalah Unnamed: 0, Recency, Year_Birth, z_CostContact, z_Revenue, dan Dt_Customer.

```
df = df.drop(columns=['Unnamed: 0', 'Recency', 'Year_Birth', 'Z_CostContact', 'Z_Revenue', 'Dt_Customer'])
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Encoding

Feature Encoder

Untuk fitur kategorikal seperti Education, Marital status, age_range, dan is_parents dapat dilakukan encoding sesuai dengan jenisnya. Fitur Education dapat dilakukan label encoding, sedangkan sisanya dapat dilakukan proses one hot encoding.

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

df['Education_mapped'] = df['Education'].map(mapping_education)
```

```
# One hot encoder
for cat in ['Marital_Status', 'Age_range', 'is_parents']:
    onehots = pd.get_dummies(df[cat], prefix=cat)
    dataset_categorical = dataset_categorical.join(onehots)
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Standarization

Standarisasi

Untuk fitur yang bersifat numerikal / angka dilakukan proses standarisasi untuk menyamakan skala nilai pada masing masing fitur numerik.

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

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

display(df_scaled.shape, df_scaled.head(3))
```

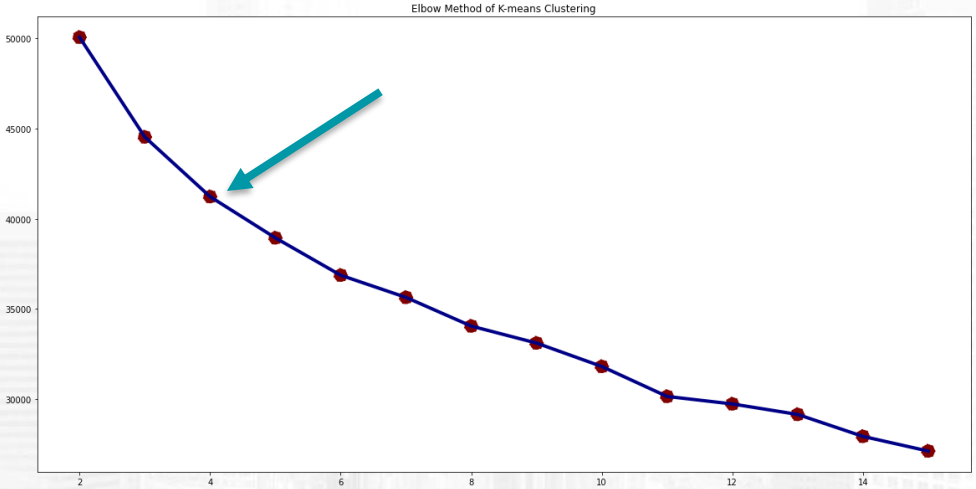
Untuk selengkapnya, dapat melihat jupyter notebook disini

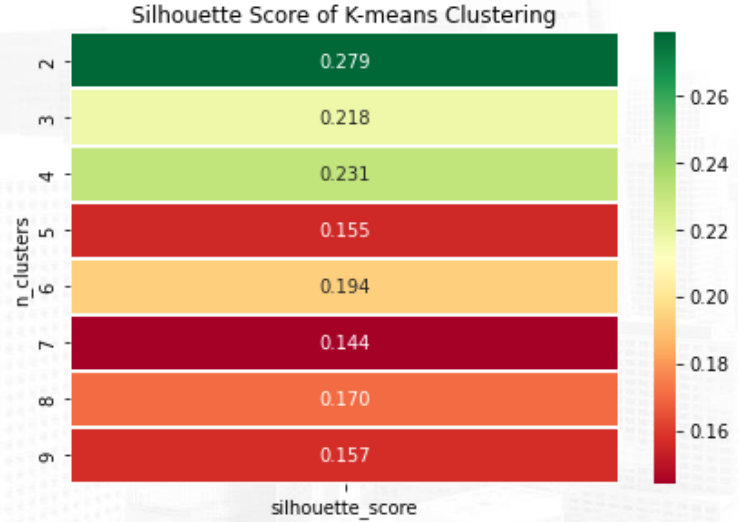
Rakamin Academy

Modelling

- Data Modelling K-Means

Data Modelling K-Means





Pada proses modelling ini dilakukan Elbow Method dan Silhouette Score Algoritma K-Means untuk menentukkan nilai K atau nilai cluster.

Pada grafik di atas didapatkan jumlah cluster = 4 berdasarkan Elbow Method dan Silhouette Score.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Interpretasi Model

- Cluster Analysis

Cluster Analysis

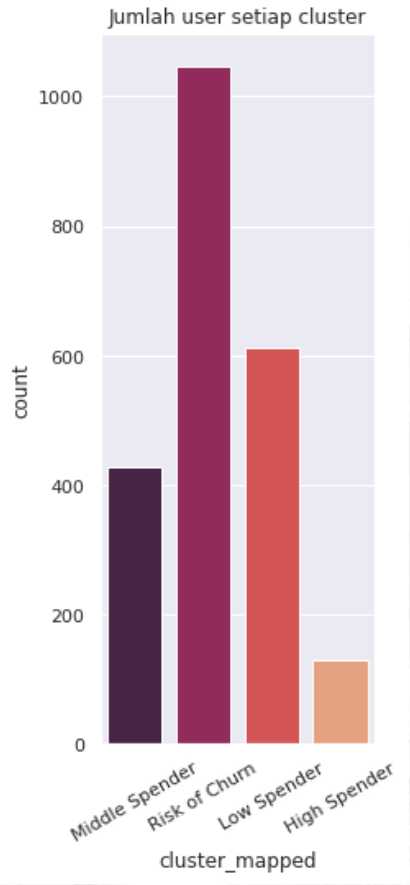
Cluster Analysis

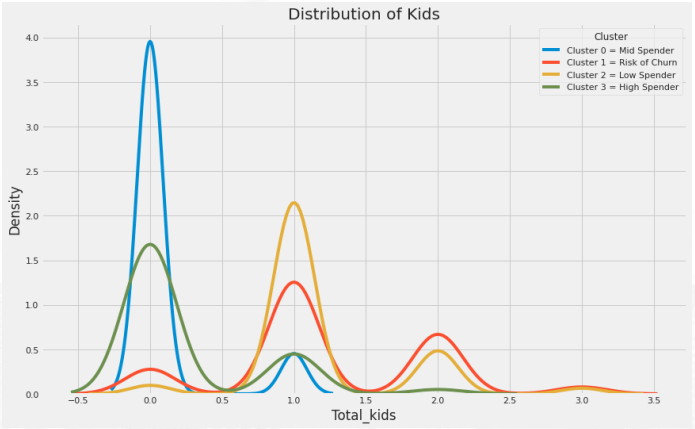
Dari hasil statistik rata-rata fitur spending didapatkan cluster 0 merupakan customer dengan kategori cusmer middle spender. Cluster 1 merupakan customer dengan kategori customer risk of churn. Cluster 2 merupakan customer dengan kategori customer low spender. Dan cluster 4 merupakan customer dengan kategori high spender.

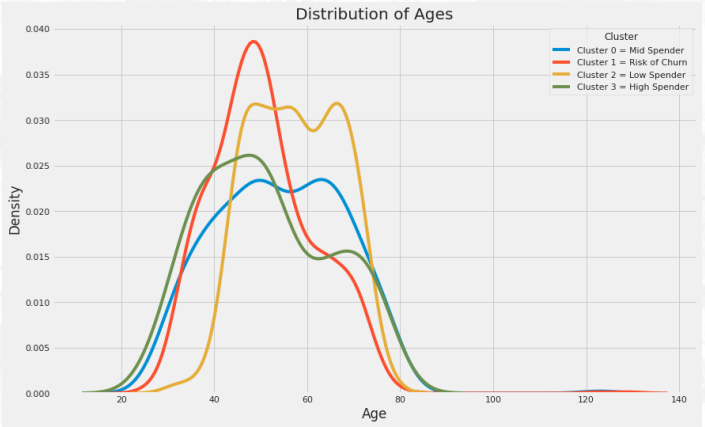
	Total_amount_spent					Income				
	count	mean	median	min	max	count	mean	median	min	max
cluster										
0	428	1.296965e+06	1281500.0	62000	2525000	428	7.557575e+07	74888500.0	2447000.0	666666000.0
1	1046	1.008614e+05	65000.0	5000	473000	1046	3.512499e+07	34616500.0	1730000.0	162397000.0
2	611	7.812488e+05	725000.0	232000	2092000	611	5.923169e+07	59462000.0	4428000.0	157243000.0
3	131	1.582702e+06	1631000.0	265000	2524000	131	8.016937e+07	81929000.0	37929000.0	105471000.0

Untuk selengkapnya, dapat melihat jupyter notebook disini

Cluster Analysis

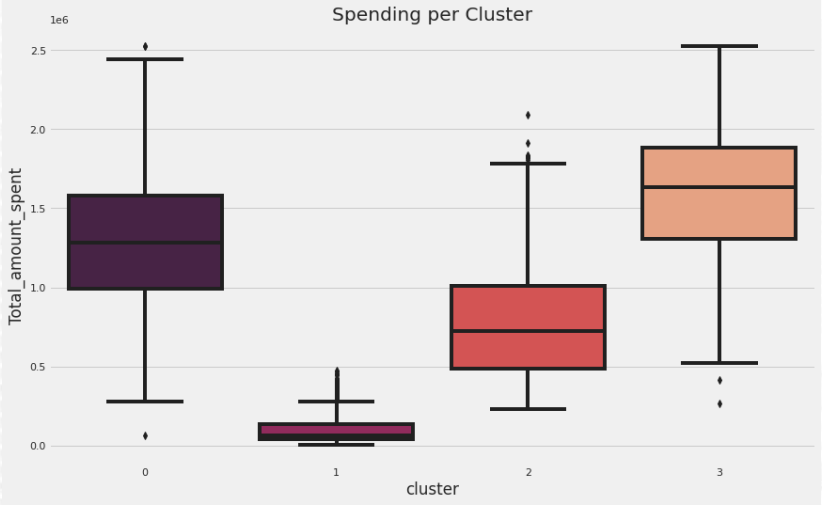






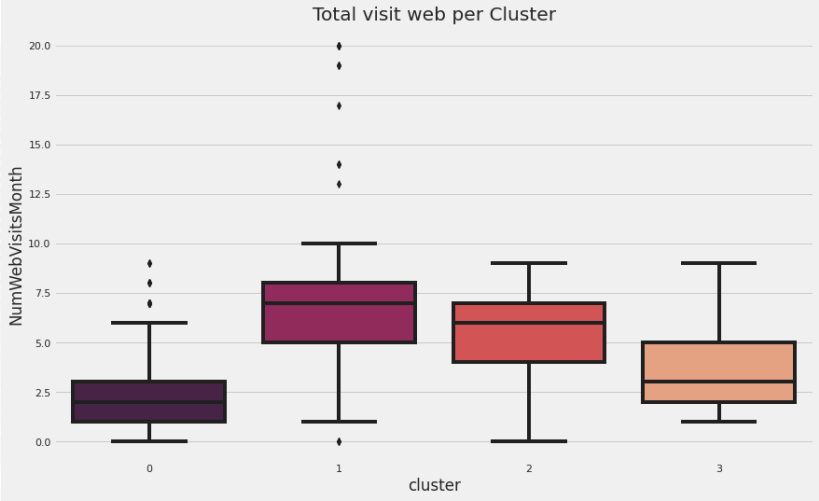
Untuk selengkapnya, dapat melihat jupyter notebook disini

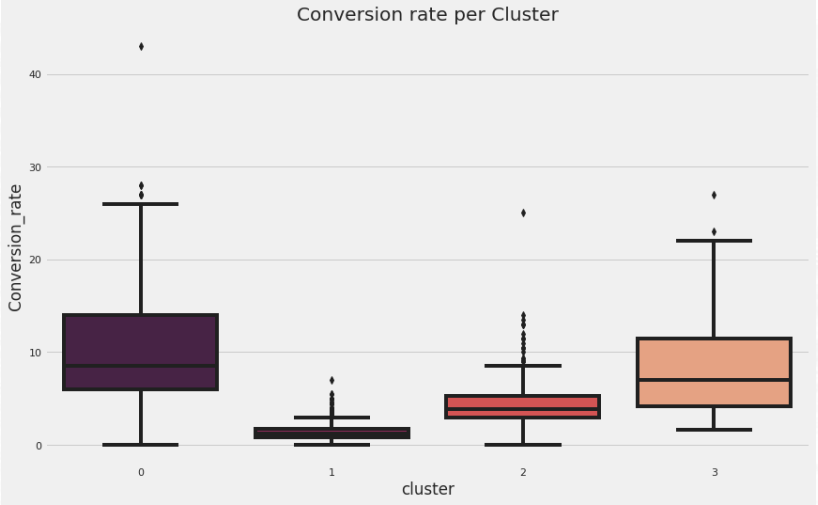
Cluster Analysis



Untuk selengkapnya, dapat melihat jupyter notebook disini

Cluster Analysis





Untuk selengkapnya, dapat melihat jupyter notebook disini

Cluster Analysis

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conclusion

1. **Low Spender (Cluster = 2)**

- Kelompok ini didominasi oleh older adults (>=55 tahun) dan middle adults (36 tahun – 55 tahun) yang dominan memiliki anak 1.
- Kelompok ini cukup sering mengujungi website setelah cluster 1 (risk of churn) dengan median 5 kali tiap bulannya.
- Kelompok ini memiliki total pendapatan dan total pengeluaran terkecil kedua setelah cluster 1 (rosk of churn).

2. **Risk of churn (Cluster = 1)**

- Kelompok ini didominasi middle adults (36-55) dan older adults (>= 55 tahun) yang dominan memiliki anak dengan jumlah 1 - 2.
- Kelompok ini sering mengujungi website dengan median lebih dari 5 kali tiap bulannya.
- Kelompok ini memiliki total pendapatan dan total pengeluaran terkecil diantara cluster yang lain.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conclusion

3. Middle Spender (Cluster = 0)

- Kelompok ini didominasi oleh older adults (>=55 tahun) dan middle adults (36 tahun – 55 tahun) yang dominan tidak memiliki anak.
- Kelompok ini tidak cukup sering mengunjungi website tiap bulannya.
- Kelompok ini memiliki total pendapatan dan total pengeluaran terbesar setelah cluster 3 (High Spender).

4. High Spender (Cluster = 3)

- Kelompok ini didominasi middle adults (36-55) yang dominan tidak memiliki anak.
- Kelompok ini cukup sering mengujungi website setelah cluster 0 (Middle Spender) dengan median 3 kali tiap bulannya.
- Kelompok ini memiliki total pendapatan dan total pengeluaran paling tinggi diantara cluster yang lainnya.

Untuk selengkapnya, dapat melihat jupyter notebook disini