Predict Customer Personality to boost marketing campaign by using Machine Learning

Created by:
Yehezkiel Novianto Aryasena
hezkyaryasena@gmail.com
www.linkedin.com/in/yehezkielnov

"**Saya Yehezkiel Novianto Aryasena**. Saya merupakan fresh graduate dari Institut Teknologi Sepuluh Nopember Surabaya. Saat menjalani masa perkuliahan, saya memiliki pengalaman organisasi dan kepanitiaan yang membuat saya mampu bekerja mandiri maupun dalam tim. Saya memiliki ketertarikan untuk mempelajari hal baru terutama pada bidang data science dan saat ini sedang mendalami pengetahuan saya dalam hal tersebut dengan mengikuti course yang diselenggarakan oleh Rakamin. "

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

Overview

Rakamin Academy

“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

Rakamin Academy

- Lakukan Feature Engineering dengan menghitung conversion rate dengan definisi (#response / #visit). Tidak hanya conversion rate, namun juga cari feature lain yang representatif, contohnya seperti umur, jumlah anak, total pengeluaran, total transaksi, dll.

- Tulislah ***Exploration Data Analysis*** (EDA) yang sudah kamu lakukan, mulai dari plot yang kamu buat hingga analisis interpretasinya. Tuliskan pula insight yang dapat dijadikan rekomendasi (jika ada).

- **Source code** yang sudah kamu buat, dapat ditampilkan dan berikan link untuk mengakses file tersebut. Contohnya seperti di pojok kanan bawah.

https://colab.research.google.com/drive/1G2gPwbwEoIgBJc_WpoKLVY475EtrwS73?usp=sharing

Feature Engineering (Feature Extraction) I

- Membuat Feature Conversion Rate

```
df['Conversion_Rate']=df['Response']/df['NumWebVisitsMonth']*100
```

```
df['Conversion_Rate'].value_counts()
```

```
0.000000      1895
12.500000       57
16.666667       48
14.285714       45
50.000000       40
33.333333       33
100.000000      30
11.111111       29
25.000000       26
20.000000       25
10.000000        1
Name: Conversion_Rate, dtype: int64
```

Feature Engineering (Feature Extraction) II

- Membuat Feature Customer Age Year (df[‘Dt_Customer’] –df[‘ Year_Birth’])

```
df['Customer_Age_Year'].value_counts().sort_index()
```

```
16.0     1
17.0     2
18.0     3
19.0     4
20.0     7
        ..
72.0     1
73.0     1
113.0    1
114.0    1
121.0    1
```

- Membuat Feature Customer Age Group untuk mengelompokkan umur customer

```
df['Customer_Age_Group']= np.where((df['Customer_Age_Year'] >= 0) & (df['Customer_Age_Year'] < 15),'Children',
                          np.where((df['Customer_Age_Year'] >= 15) & (df['Customer_Age_Year'] < 25),'Youth',
                                   np.where((df['Customer_Age_Year'] >= 25) & (df['Customer_Age_Year'] < 45),'Young Adults',
                                            np.where((df['Customer_Age_Year'] >= 45) & (df['Customer_Age_Year'] < 65),'Middle-aged Adults','Old-aged Adults'))))
```

```
df['Customer_Age_Group'].value_counts()
```

```
Young Adults          1140
Middle-aged Adults     930
Youth                   89
Old-aged Adults         81
```

- Membuat Feature Total Kids (df['Kidhome']+df['Teenhome'])

```
df['Total_Kids']=df['Kidhome']+df['Teenhome']
```

```
df['Total_Kids'].value_counts()
```

```
1    1128
0     638
2     421
3      53
Name: Total_Kids, dtype: int64
```

Feature Engineering (Feature Extraction) III

Rakamin Academy

- Membuat Feature Is Parents?

```
df['Is_Parents']=np.where((df['Kidhome'] >= 1) & (df['Teenhome'] >= 1),'Yes','No')
```

```
df['Is_Parents'].value_counts()
```

```
No     1813
Yes     427
Name: Is_Parents, dtype: int64
```

- Membuat Feature Total Purchases (df['NumDealsPurchases']+df['NumWebPurchases']+df['NumCatalogPurchases'])

```
df['Total_Purchases']=df['NumDealsPurchases']+df['NumWebPurchases']+df['NumCatalogPurchases']
```

```
df['Total_Purchases']
```

```
0       21
1        4
2       11
3        4
4       13
        ..
2235    14
2236    17
2237     6
2238    13
2239     7
```

- Membuat Feature Total Outcomes (df['MntCoke']+df['MntFruits']+df['MntMeatProducts']+df['MntFishProducts']+df['MntSweetProducts']+df['MntGoldProds'])

```
df['Total_Outcomes']=df['MntCoke']+df['MntFruits']+df['MntMeatProducts']+df['MntFishProducts']+df['MntSweetProducts']+df['MntGoldProds']
```

```
df['Total_Outcomes']
```

```
0       1617000
1         27000
2        776000
3         53000
4        422000
         ...
2235    1341000
2236     444000
2237    1241000
2238     843000
2239     172000
Name: Total_Outcomes, Length: 2240, dtype: int64
```

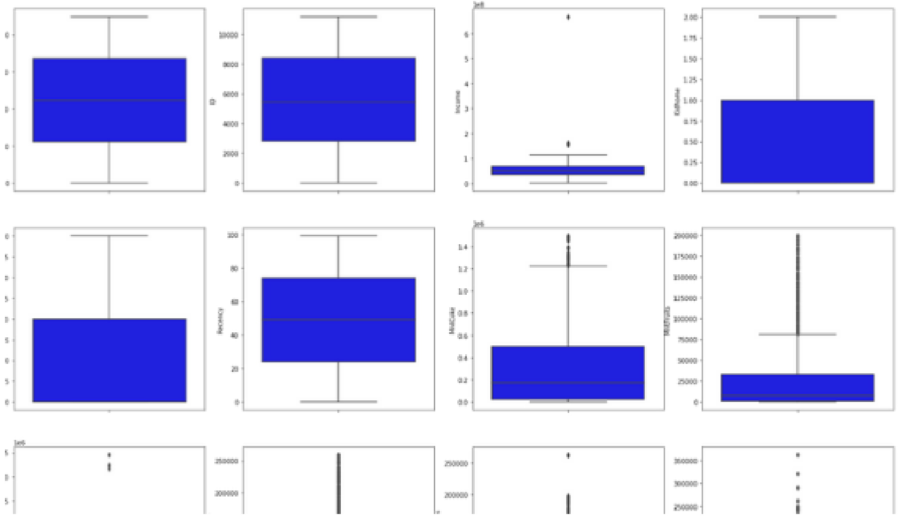
Exploratory Data Analysis I

Insight:

- There are still Outliers in several features such as Income, Mntcoke, Mntfruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, NumDealPurchase, NumWebPurchase and so on.

- Although there are still many features that have outliers, there are also some features that are free from outliers such as ID, KidHome, TeenHome, Recency, NumStorePurchases

- Box Plot Numerical



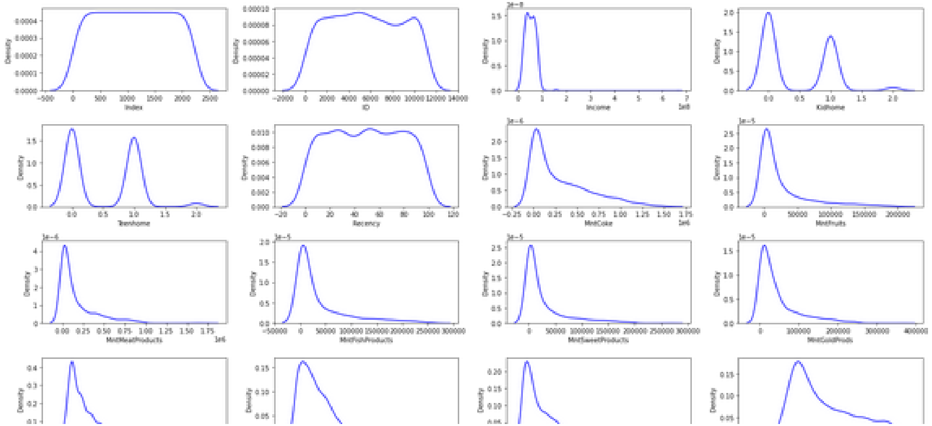
Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Exploratory Data Analysis II

Insight:

- From the visualization beside, there are still many features that are not normally distributed and are dominated by features that have positive skewness (Right Skewness).
- Features that are close to normal distribution are ID and Recency

- Dist Plot Numerical

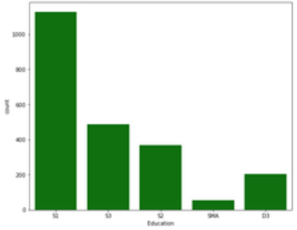


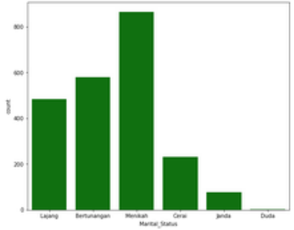
Exploratory Data Analysis III

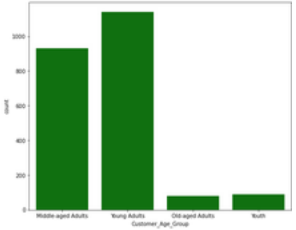
Insight:

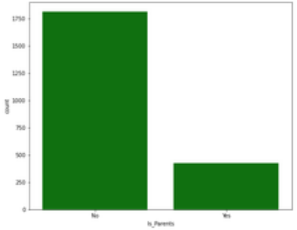
- There is an imbalance of one value with another value in a feature. This can be seen in the Customer_Age_Group feature with a Young Adults value of almost 1000+.
- This also applied at Marital Status feature with a married value reaching 800+ while a widower value that does not reach 50
- Feature Education is dominated by customers with a final education level of S1
- Most customers in the dataset are not parents because they don't have children yet

- Bar Plot Categorical







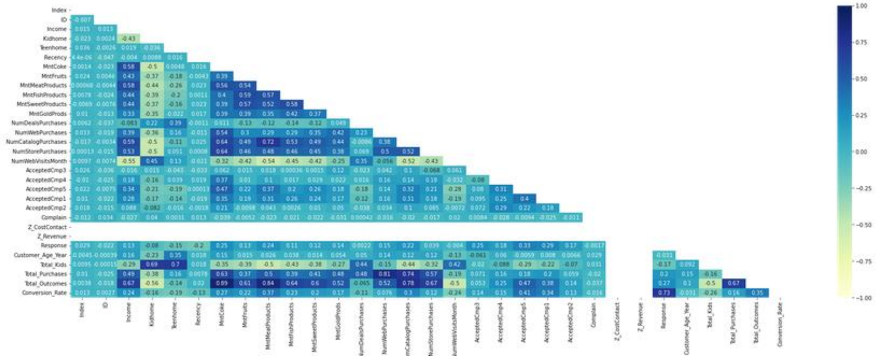


Exploratory Data Analysis IV

Insight:

- Some features that have a strong correlation (0.7) include: Total_Kids with Kidhome and Teenhome NumCatalogPurchase with MntMeatProducts Total_Outcomes with MntCoke, MntMeatProducts, NumCatalogPurchase Conversion_Rate with Response
- Conversion_Rate with Customer_Age_Year has a negative correlation. This indicates that there is no strong relationship between the age of the customer and the level of visitor interest in becoming a customer

- Bar Plot Categorical



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

Exploratory Data Analysis V

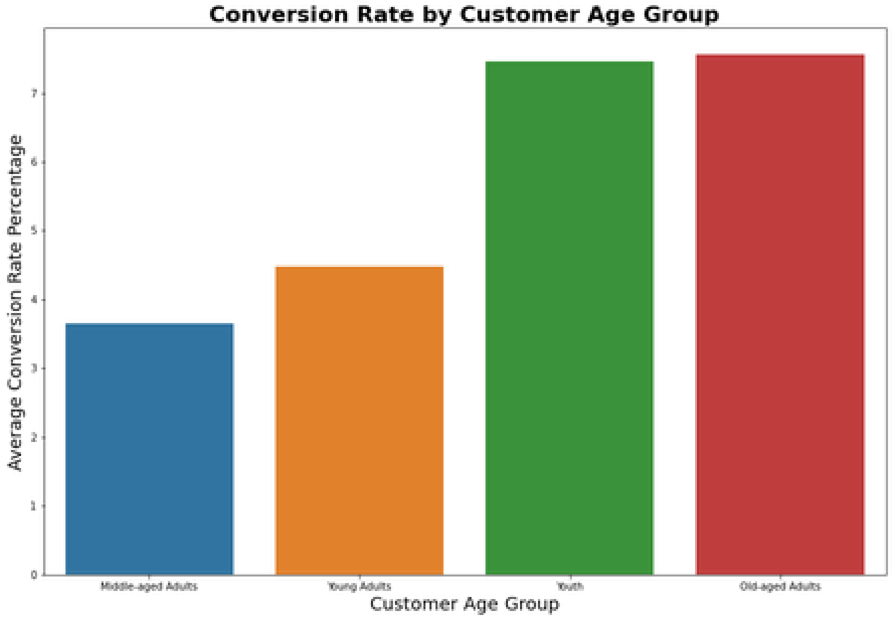
Insight:

- The average conversion rate for each age group, namely Youth, Middle Adults, Young Adults, and Old Aged Adults, does not reach 10%.
- The average Conversion Rate for Middle-aged Adults and Young Adu groups is only 3% and 4%, while Youth and Old-Aged Adults reach 7.5%. Although there is a slight difference in the Conversion Rate level in the age groups, the difference is not too significant.

Business Recommendation:

- The recommended steps that can be implemented in the future is to make the campaign more personalized to certain age groups. Example: Collaborating with well-known bands or famous artists to conduct campaigns on age groups that are relevant to that band/artist.

- Conversion Rate & Customer Age Group



Data Cleaning & Preprocessing

- Pada tahap ***cleaning data***, tunjukan ***null*** atau ***missing value*** serta ***duplicated value*** pada dataset, serta cara penyelesaiannya.

- Selanjutnya untuk data preprocessing, tunjukan bahwa data sudah dilakukan proses ***feature encoding*** dan ***feature standardisation***.

- **Source code** yang sudah kamu buat, dapat ditampilkan dan berikan link untuk mengakses file tersebut. Contohnya seperti di pojok kanan bawah.

https://colab.research.google.com/drive/1G2gPwbwEoIgBJc_WpoKLVY475EtrwS73?usp=sharing

Data Pre-Processing I

Treatment:

- Because the missing values in both the Income and Conversion Rate columns have very small proportions compared to the overall data, the data rows containing the missing values will be dropped.

```
df_clean.dropna(subset=['Income','Conversion_Rate'], inplace=True)
```

```
df_clean.isnull().sum().sort_values(ascending=False)
```

```
Index               0
Z_Revenue           0
AcceptedCmp3        0
AcceptedCmp4        0
AcceptedCmp5        0
AcceptedCmp1        0
AcceptedCmp2        0
Complain            0
Z_CostContact       0
Response            0
ID                  0
```

- Problem: Missing Value

```
df_clean.isnull().sum().sort_values(ascending=False)
```

```
Income                 24
Conversion_Rate        11
Response                0
AcceptedCmp4            0
AcceptedCmp5            0
AcceptedCmp1            0
AcceptedCmp2            0
Complain                0
Z_CostContact           0
Z_Revenue               0
Customer_Age            0
NumWebVisitsMonth       0
Customer_Age_Year       0
Customer_Age_Group      0
Total_Kids              0
Is_Parents              0
Total_Purchases         0
Total_Outcomes          0
AcceptedCmp3            0
Index                   0
ID                      0
NumCatalogPurchases     0
Year_Birth              0
Education               0
Marital_Status          0
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
NumStorePurchases       0
dtype: int64
```

Data Pre-Processing II

Treatment:

- No duplicate data found

- Problem: Duplicated Value

```
]: df_clean.duplicated().sum()

]: 0
```

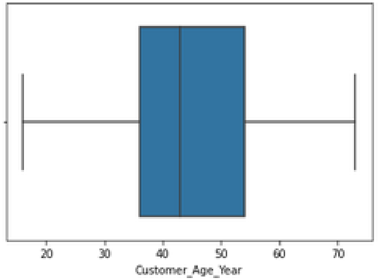
Data Pre-Processing III

Treatment:

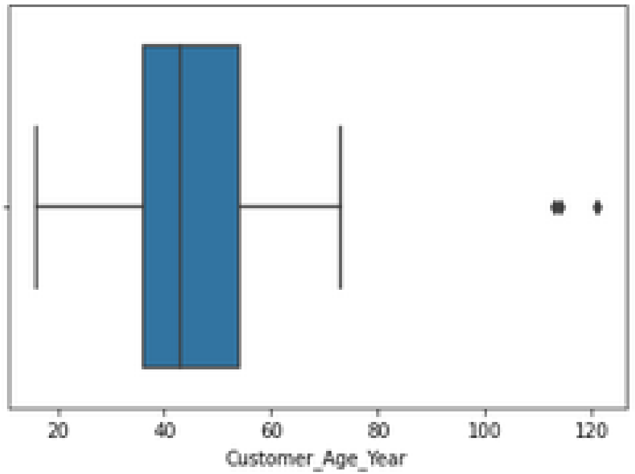
```
df_clean=df_clean[(df_clean['Customer_Age_Year']>=low_limit) & (df_clean['Customer_Age_Year']<=high_limit)]
```

```
plt.figure(figsize=(6,4))
sns.boxplot(x = df_clean['Customer_Age_Year'])
```

```
<AxesSubplot:xlabel='Customer_Age_Year'>
```



- Problem: Customer Age Year Outlier



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

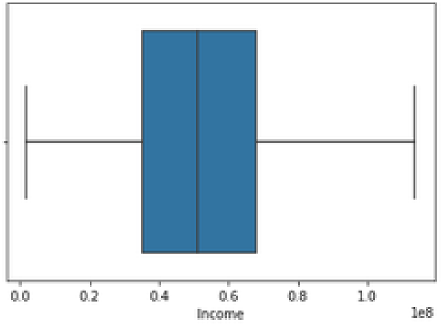
Data Pre-Processing IV

Treatment:

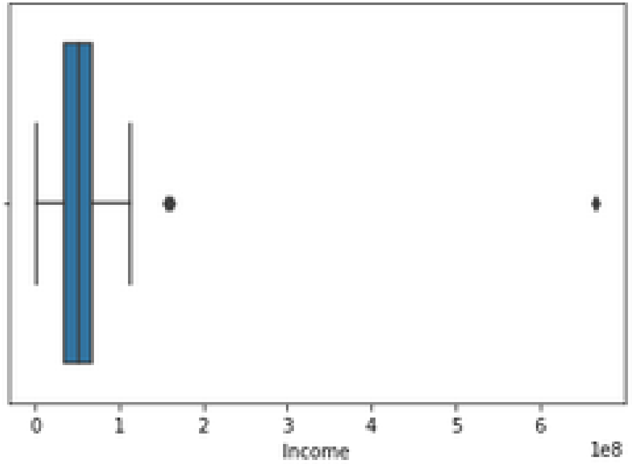
```
df_clean=df_clean[(df_clean['Income']>=low_limit) & (df_clean['Income']<=high_limit)]
```

```
plt.figure(figsize=(6,4))
sns.boxplot(x = df_clean['Income'])
```

```
<AxesSubplot:xlabel='Income'>
```



- Problem: Income Outlier



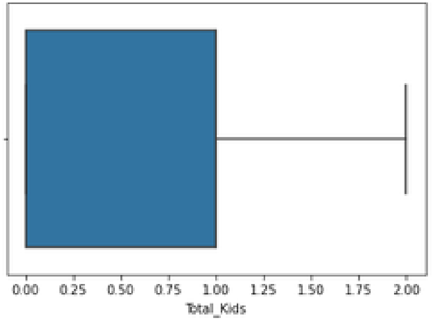
Data Pre-Processing V

Treatment:

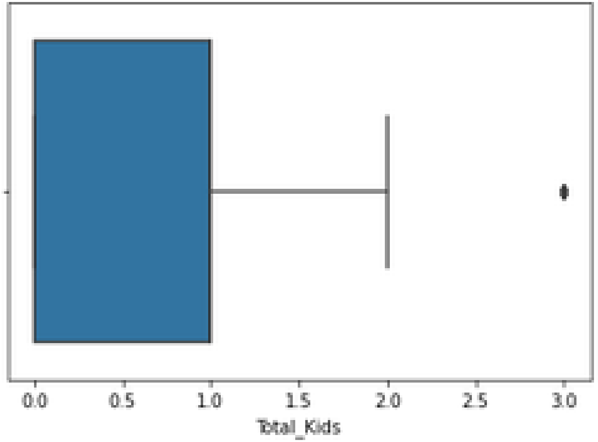
```
df_clean=df_clean[(df_clean['Total_Kids']>=low_limit) & (df_clean['Total_Kids']<=high_limit)]
```

```
plt.figure(figsize=(6,4))
sns.boxplot(x = df_clean['Total_Kids'])
```

<AxesSubplot:xlabel='Total_Kids'>



- Problem: Total_Kids Outlier



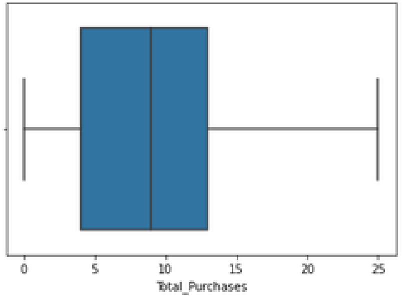
Data Pre-Processing VI

Treatment:

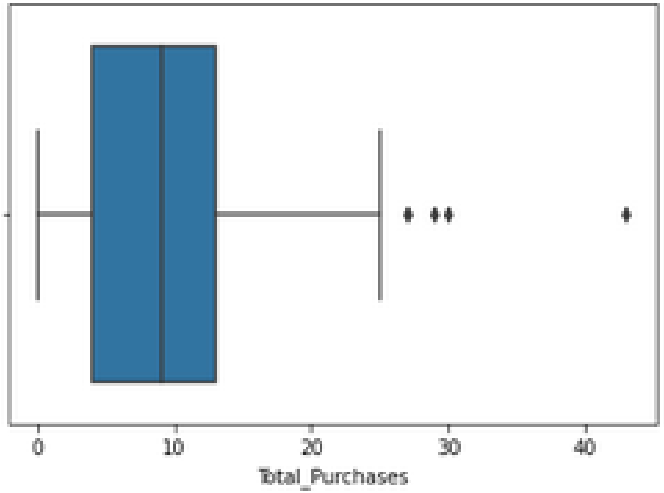
```
df_clean=df_clean[(df_clean['Total_Purchases']>=low_limit) & (df_clean['Total_Purchases']<=high_limit)]
```

```
plt.figure(figsize=(6,4))
sns.boxplot(x = df_clean['Total_Purchases'])
```

<AxesSubplot:xlabel='Total_Purchases'>



- Problem: Total_Purchases Outlier



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

Data Pre-Processing VII

Treatment: Label Encoding to Numerical

```
df_clean['Is_Parents'].value_counts()
```

```
0    1778
1     367
Name: Is_Parents, dtype: int64
```

```
df_clean['Education'].value_counts()
```

```
2    1086
4     456
3     356
1     193
0      54
Name: Education, dtype: int64
```

- Problem: Feature Is_Parents & Education Categorical

```
mapping_parents = {
    'No' : 0,
    'Yes' : 1
}
df_clean['Is_Parents']=df_clean['Is_Parents'].map(mapping_parents)

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

Data Pre-Processing VIII

Treatment: OHE to Numerical

Status_Bertunangan	Status_Cerai	Status_Duda	Status_Janda	Status_Lajang	Status_Menikah	GolUmur_Middle-aged Adults	GolUmur_Old-aged Adults	GolUmur_Young Adults	GolUmur_Youth
0	0	0	0	1	0	1	0	0	0
0	0	0	0	1	0	1	0	0	0
1	0	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	1	0	0	1	0
...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	1	0	0	1	0
0	0	0	0	0	1	1	0	0	0
0	1	0	0	0	0	0	0	1	0
1	0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	1	0	0	0

- Problem: Feature Marital_Status & Customer_Age_Group Categorical

```
status_kawin=pd.get_dummies(df_clean['Marital_Status'], prefix='Status')
```

```
grup_umur=pd.get_dummies(df_clean['Customer_Age_Group'], prefix='GolUmur')
```

```
df_clean=pd.concat([df_clean,status_kawin,grup_umur], axis=1)
```

Data Pre-Processing VIII

Treatment: Choose features based on RFM and standardized it.

```
df_clean2['Recency_std'] = StandardScaler().fit_transform(df_clean2['Recency'].values.reshape(len(df_clean2),1))
df_clean2['Total_Purchases_std'] = StandardScaler().fit_transform(df_clean2['Total_Purchases'].values.reshape(len(df_clean2),1))
df_clean2['Total_Outcomes_std'] = StandardScaler().fit_transform(df_clean2['Total_Outcomes'].values.reshape(len(df_clean2),1))
df_clean2.describe()
```

	Recency	Total_Purchases	Total_Outcomes	Recency_std	Total_Purchases_std	Total_Outcomes_std
count	2145.000000	2145.000000	2.145000e+03	2.145000e+03	2.145000e+03	2.145000e+03
mean	48.847552	9.051748	6.126014e+05	8.035530e-17	-9.821204e-17	4.917072e-17
std	28.852891	5.100997	6.026071e+05	1.000233e+00	1.000233e+00	1.000233e+00
min	0.000000	0.000000	5.000000e+03	-1.693381e+00	-1.774919e+00	-1.008523e+00
25%	24.000000	4.000000	7.000000e+04	-8.613815e-01	-9.905761e-01	-9.006331e-01
50%	49.000000	9.000000	4.050000e+05	5.284846e-03	-1.014710e-02	-3.445857e-01
75%	74.000000	13.000000	1.049000e+06	8.719512e-01	7.741961e-01	7.243531e-01
max	99.000000	25.000000	2.525000e+06	1.738618e+00	3.127226e+00	3.174281e+00

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

- Problem: Too many irrelevant features for clustering

```
from sklearn.preprocessing import MinMaxScaler, StandardScaler
```

```
#Mengambil feature yang berhubungan dengan RFM
df_clean2=df_clean[['Recency','Total_Purchases','Total_Outcomes']]
```

```
df_clean2
```

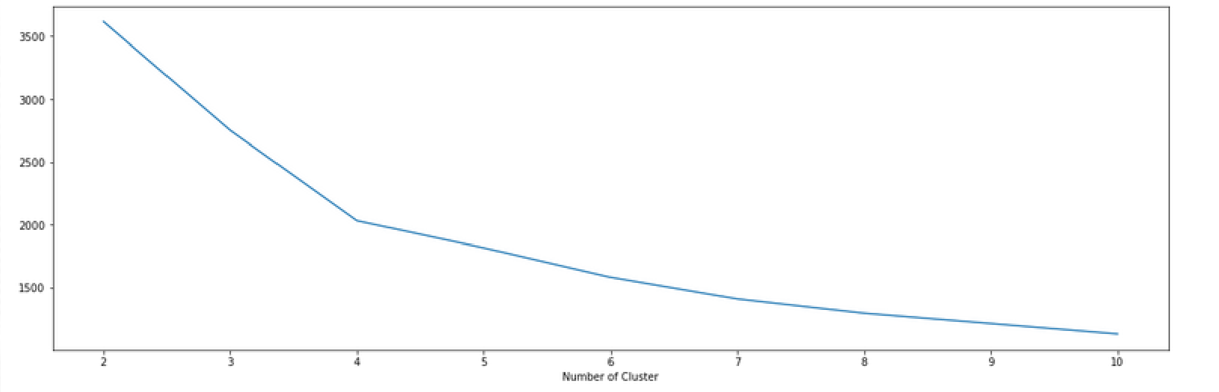
	Recency	Total_Purchases	Total_Outcomes
0	58	21	1617000
1	38	4	27000
2	26	11	776000
3	26	4	53000
4	94	13	422000
...	...	...	...
2234	81	2	30000
2235	46	14	1341000
2237	91	6	1241000
2238	8	13	843000
2239	40	7	172000

Data Modeling

- Tunjukan visualisasi ***Elbow Method*** menggunakan ***K-Means Clustering*** dan hasil evaluasinya menggunakan ***Silhouette Score***, serta buatkan lah hasil interpretasinya.

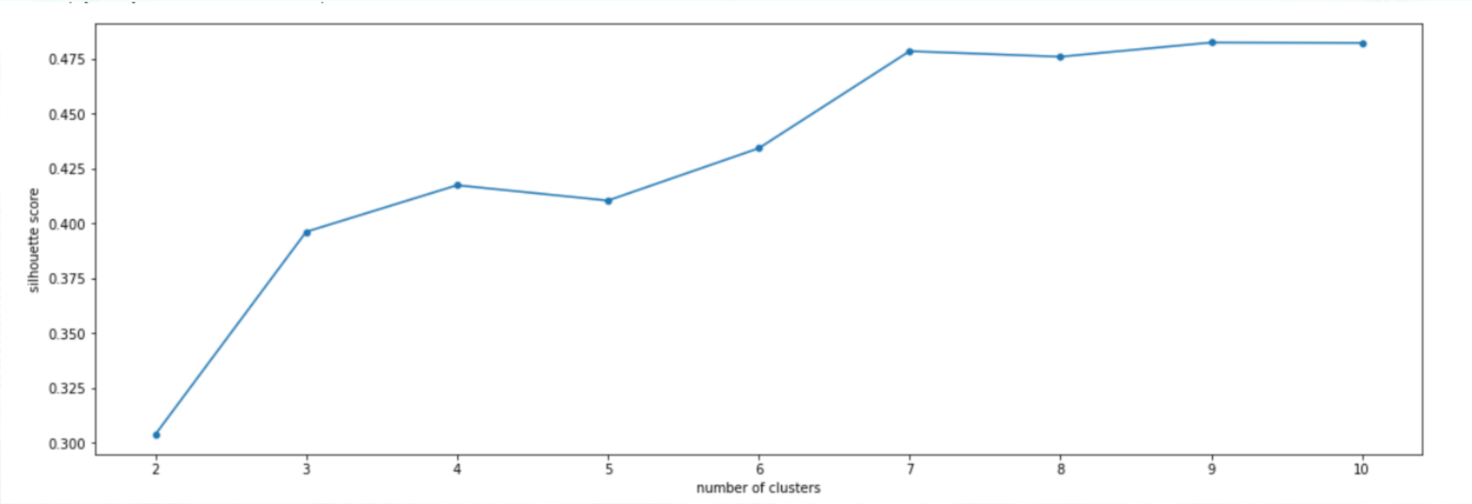
https://colab.research.google.com/drive/1G2gPwbwEoIgBJc_WpoKLVY475EtrwS73?usp=sharing

Elbow Method



According to the Elbow Method above, there is no more significant decrease when the number of clusters becomes more than 4. So the most suitable number of clusters to be used later in clustering using KMeans is 4 clusters.

Silhouette Score



After getting the right number of clusters, it can be seen from the graph above that the Silhouette Score for the number of n_cluster = 4 is 0.41. Because the Silhouette Score is 0, the distance between the visualized clusters will not be so significant.

Customer Personality Analysis for Marketing Retargeting

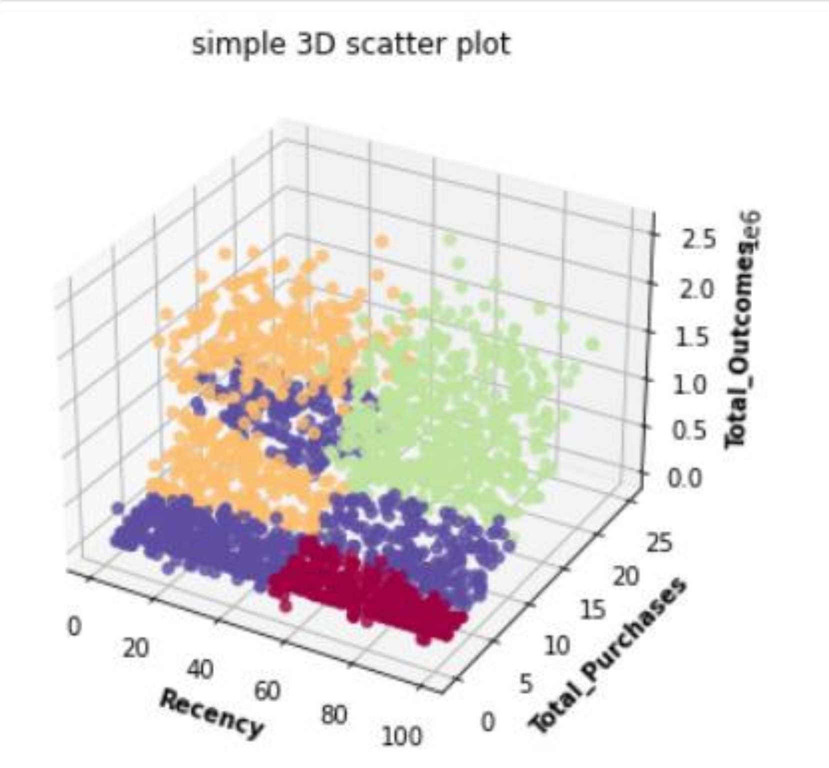
- Tunjukan visualisasi analisis dari EDA dengan menggunakan ***hasil cluster*** yang sudah didapat. Buatlah rekomendasi bisnis yang dapat dilakukan dari analisis tersebut.

Untuk selengkapnya, dapat melihat jupyter notebook disini :
https://colab.research.google.com/drive/1G2gPwbwEoIgBJc_WpoKLVY475EtrwS73?usp=sharing

Clustering Visualization

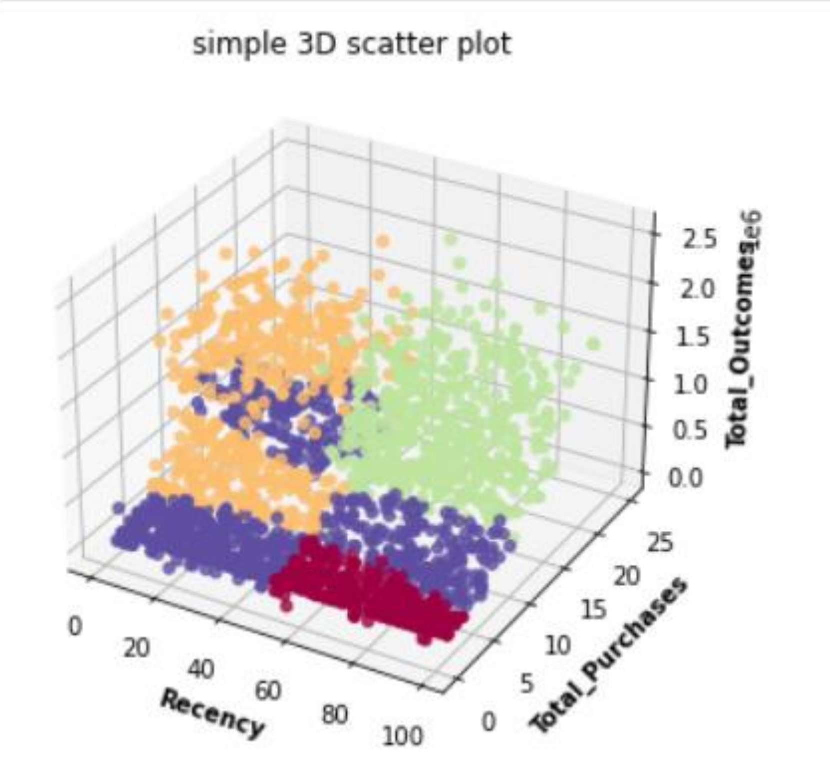
Conclusion:

The use of the Kmeans algorithm with features based on **RFM** (Recency, Frequency, Monetary) namely the number of days since the customer's last purchase, the total number of customer purchases, and the total number of customer expenditures along with the number of **clusters = 4** (the best cluster results using the **Elbow Method**) will produce a visualization in the form of a Scatterplot which can be seen in the following image.



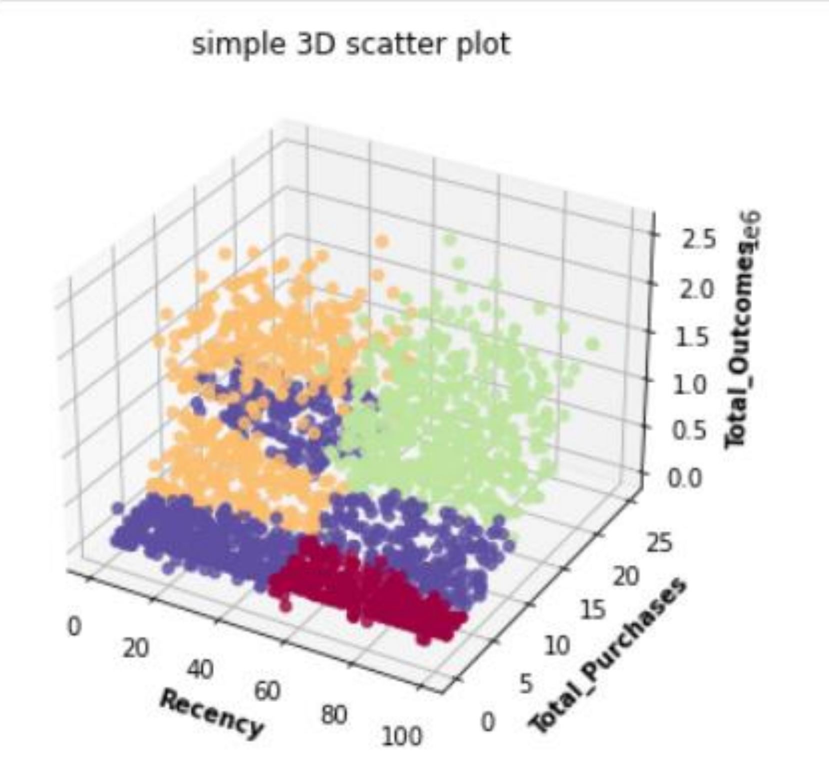
Cluster 1

Cluster 1 is a customer who has low recency with total purchases that vary from the range of 3-25+ purchases, but the total customer spending in **Cluster 1** is relatively low because it is below 500000. This may happen when the customer makes a purchase only when there is a promo



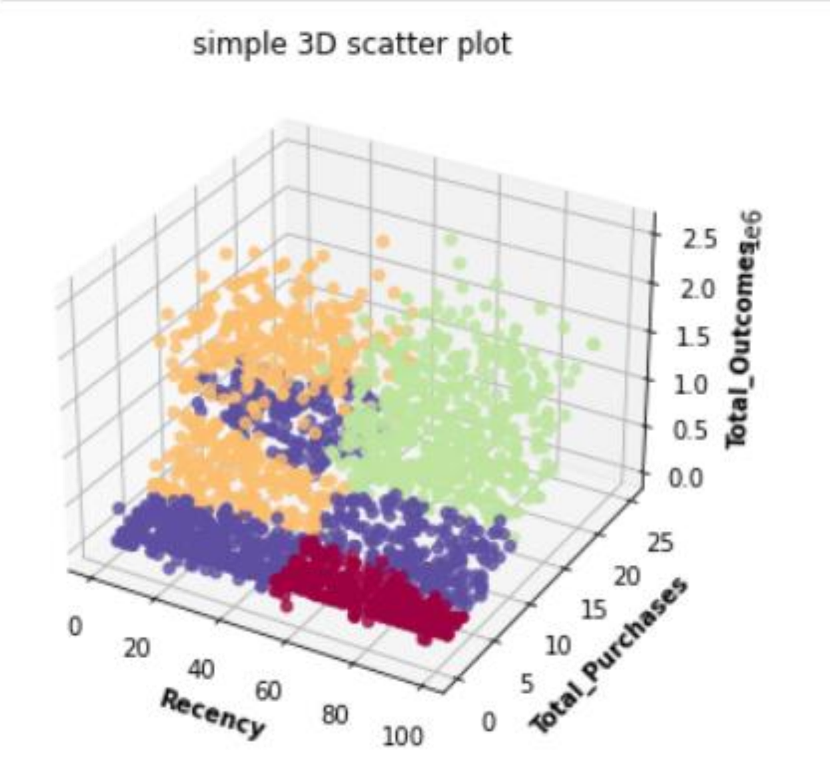
Cluster 2

Cluster 2 is a customer who has high recency with a low total purchase and expenditure, which is below 10 purchases and an expenditure below 500000. This may occur when the customer is not really interested in the products offered or the campaign has not been successful.



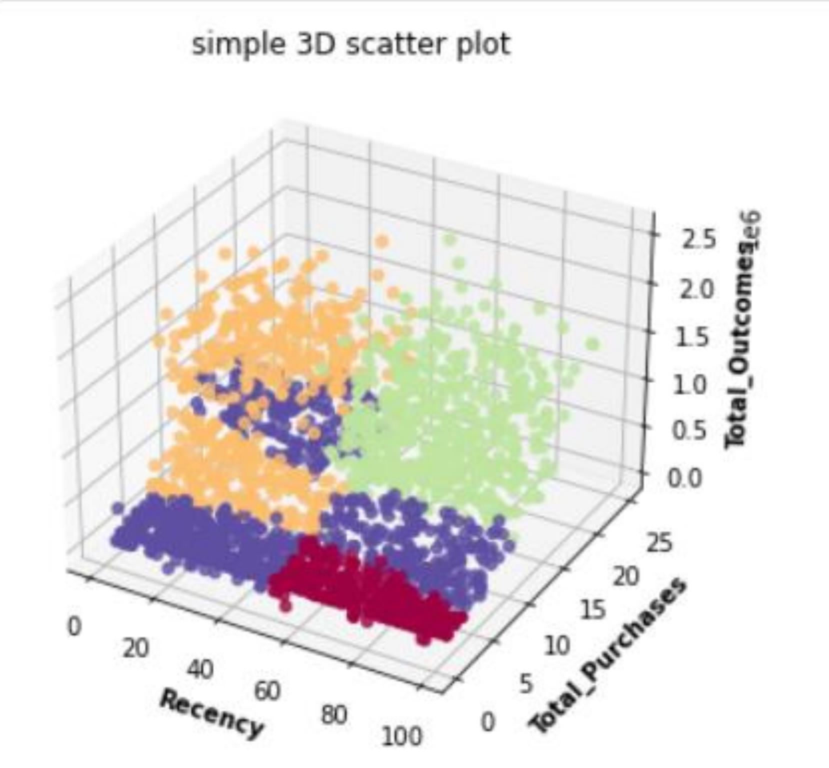
Cluster 3

Cluster 3 are customers who have low recency with total purchases classified as medium to large, namely from 15-25 purchases and total expenses from 0 to 1500000. In this Cluster 3, the campaign can be said to be successful because it can invite customers to make purchases. Customers in this Cluster 3 can also be considered loyal because they will still make purchases even though there is no promo. This is evidenced by the gap in the total expenditure in this cluster 3



Cluster 4

Cluster 4 is a customer who has a high level of recency but is balanced with a high total purchase as well. However, in cluster 4, total customer spending is classified slightly to medium, namely from 0 to 1500000. Customers who enter this **Cluster 4** can be said to only buy products for a long period of time but in large quantities.



Business Recommendation

1. Companies can focus on paying attention to customers who are in **Cluster 3** because these customers are loyal customers. What companies can do is to provide promos and can create innovations such as membership cards with special benefits for customers who have them.
2. Trying to do a campaign that is more in line with the customer persona in order to reduce the customer recency level in **Cluster 4**