PREDICT CUSTOMER PERSONALITY TO BOOST MARKETING CAMPAIGN USING MACHINE LEARNING

Created by:
Putrini Nur Amalina H.
putrininur@gmail.com
linkedin.com/in/putrininur
github.com/putrini

A Diponegoro University graduate who experienced working in the Finance and Accounting Department at a Food Distribution company. A data-driven and tech-savvy person who has huge interest in data analytics who is skilled in SQL, Python, and data visualization using Google Data Studio. Highly skilled in Microsoft Excel and able to actively communicate in English fluently.

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

OVERVIEW

Machine learning is a method of data analysis that automates analytical model building. For organizations overflowing with data but struggling to turn it into useful insights, machine learning can provide the solution to analyze and make data-driven recommendations and decisions. Businesses use machine learning to recognize patterns and then make predictions about what will appeal to customers, improve operations, or help make a product better.

One of the applications of machine learning for enhancing business is customer segmentation. Customer segmentation is the process by which you divide your customers up based on common characteristics such as behaviors, so you can market to those customers more effectively. By better understanding the customer, and therefore being able to target them more effectively to generate insights that will boost marketing campaign. In this paper we will implement the machine learning model to identify customer profile and behavior.

GOAL

The goal is to identify customer profile and behavior such as total spend amount, income, purchasing activities based on its cluster using classification model to generate data driven business recommendation which objectives is to boost marketing campaign.

TOOLS

Python as
Programming language

Google colab as
notebook

Python library used are :
- Pandas
- Numpy
- Matplotlib
- Seaborn
- StandardScaler
- Kmeans
- Silhouette score

EXPLORATORY DATA ANALYSIS

```
<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
```

Insight :

- Data consist of 2240 rows and 30 columns
- There is missing value in Income column
- Data types are int64 (26), float64 (1), object(3) data types.

click to show the complete query

FEATURE ENGINEERING

1. Age

```
# age
df1['age'] = 2022 - df1['Year_Birth']
```

2. Children

```
# children
df1['children'] = df1['Kidhome'] + df1['Teenhome']
```

3. Total spending

```
# total spending
df1['totalspending'] =  df1['MntCoke'] \
                      + df1['MntFruits'] \
                      + df1['MntMeatProducts'] \
                      + df1['MntFishProducts'] \
                      + df1['MntSweetProducts'] \
                      + df1['MntGoldProds']
```

4. Total transaction

```
# total transaction
df1['totaltransaction'] = df1['NumWebPurchases'] \
                        + df1['NumCatalogPurchases'] \
                        + df1['NumStorePurchases'] \
                        + df1['NumDealsPurchases']
```

5. Conversion rate

```
# conversion rate
df1['conversionrate'] = df1['totaltransaction'] * 100 / (df1['NumWebVisitsMonth'])
```

6. Marital situation

```
# marital situation
df1['maritalsituation'] = np.where(df1['Marital_Status'].isin(['Menikah','Bertunangan']),
                                   'in_couple', 'alone')
```

FEATURE ENGINEERING

7. Is parent

```
# is parent
df1['is_parent'] = np.where(df1['children']>0, 1, 0)
```

8. Total accepted campaign

```
# total accepted campaign
df1['acceptcampaign'] = df1['AcceptedCmp3'] \
                       + df1['AcceptedCmp4'] \
                       + df1['AcceptedCmp5'] \
                       + df1['AcceptedCmp1'] \
                       + df1['AcceptedCmp2']
```

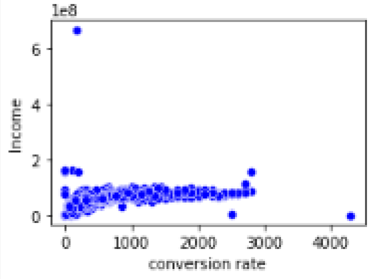
9. Year join

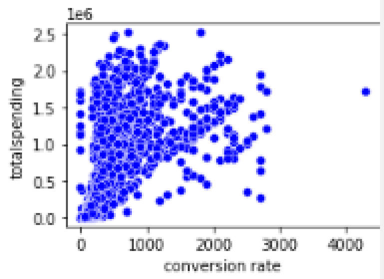
```
# year_join
df1['Dt_Customer'] = pd.to_datetime(df1['Dt_Customer'])
df1['year_join']= 2022 - df1['Dt_Customer'].dt.year
```

10. Age group

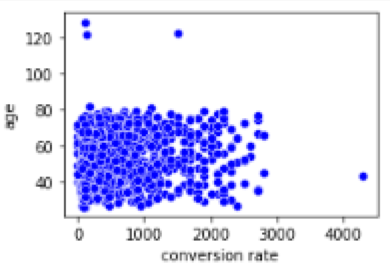
```
# age_range
df1.loc[(df1['age'] >= 0) & (df1['age'] < 12), 'age_range'] = "child"
df1.loc[(df1['age'] >= 12) & (df1['age'] < 18), 'age_range'] = "teens"
df1.loc[(df1['age'] >= 18) & (df1['age'] < 36), 'age_range'] = "young_adults"
df1.loc[(df1['age'] >= 36) & (df1['age'] < 55), 'age_range'] = "middle_aged_adults"
df1.loc[(df1['age'] >= 55), 'age_range'] = "older_adults"
```

EXPLORATORY DATA ANALYSIS



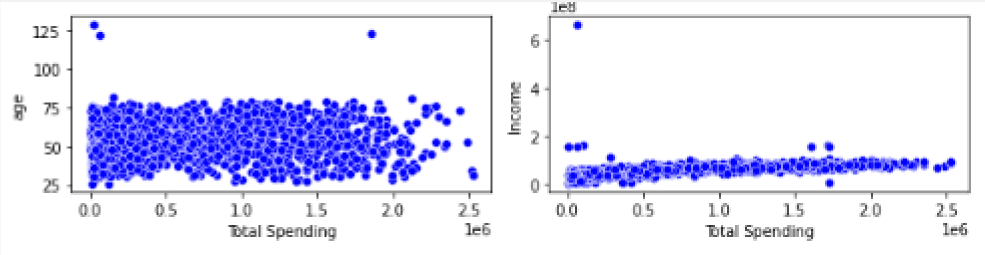


- There is linear positive correlation between conversion rate variable and income variable. The higher the income, the higher the conversion rate
- There is linear positive correlation between conversion rate variable and total spending variable. The more amount spend, the higher the conversion rate



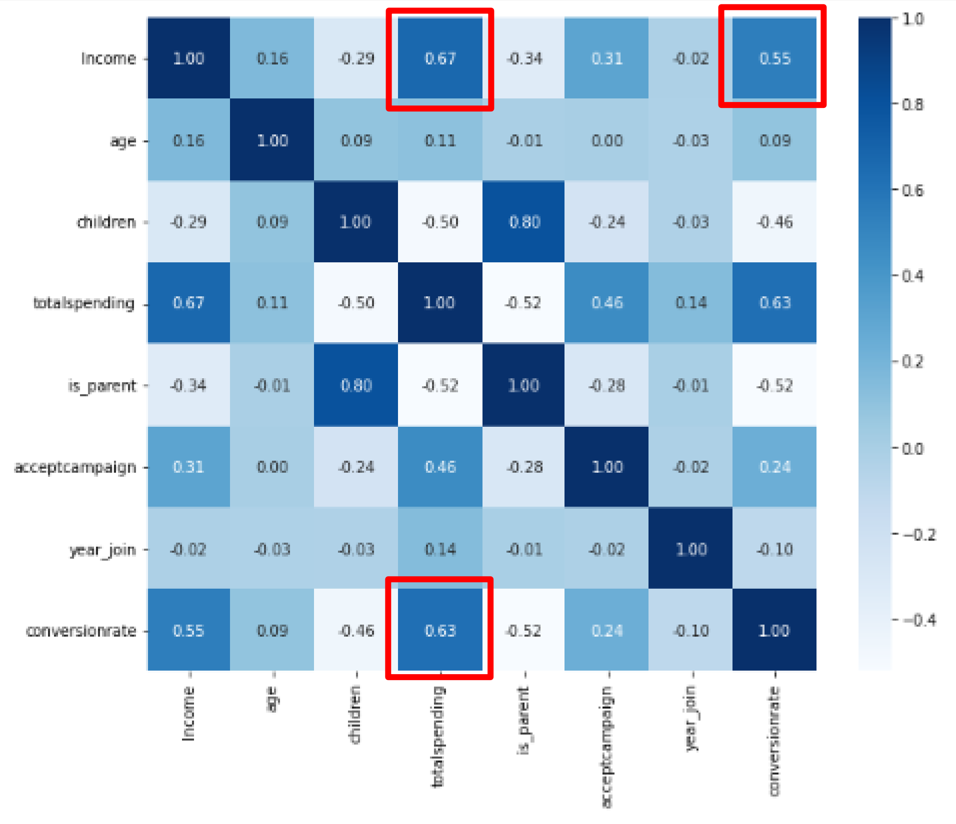
- The correlation between conversion rate and age variable is less significant because the conversion rate distribution in age variable tend to be average.

EXPLORATORY DATA ANALYSIS



- There is also less significant correlation between total spending and age variable because the distribution of total spending in age variable tend to be average
- There is linear positive correlation between total spending variable and income variable. The more income amount, the more spend amount

EXPLORATORY DATA ANALYSIS



- There is strong correlation between total spending and income variable
- There is strong correlation between total spending and conversion rate variable
- There is strong correlation between income and conversion rate variable

DATA PREPROCESSING

HANDLE NULL VALUES

```
Unnamed: 0            0
ID                    0
Year_Birth            0
Education             0
Marital_Status        0
Income               24
conversionrate        2
maritalsituation      0
```

- There are null values in income and conversion rate
- How to handle : drop NA because the amount is not significant (1%)

```
#drop missing value pada kolom income dan conversion rate
df2=df1.copy()
df2=df2.dropna(subset=['Income','conversionrate'])
```

HANDLE DUPLICATED VALUE

```
# drop duplicated rows
print(f'jumlah row duplicated adalah {df2.duplicated().sum()}')
```

```
jumlah row duplicated adalah 0
```

There is no duplicated values in dataset

DATA PREPROCESSING

FEATURE ENCODING

How to handle each column:

- Marital situation, age group, Marital Status : one hot encoding
- education : label encoding

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

df2['education_mapped'] = df2['Education'].map(mapping_education)
```

```
# handle dengan one hot encoding
for cat in ['maritalsituation', 'age_range', 'Marital_Status']:
    onehots = pd.get_dummies(df2[cat], prefix=cat)
    df2 = df2.join(onehots)
```

FEATURE SELECTION

Drop unused columns :

- Unnamed (have high variation)
- ID (have high variation)
- AcceptedCmp5, AcceptedCmp4, AcceptedCmp3 AcceptedCmp2 AcceptedCmp1

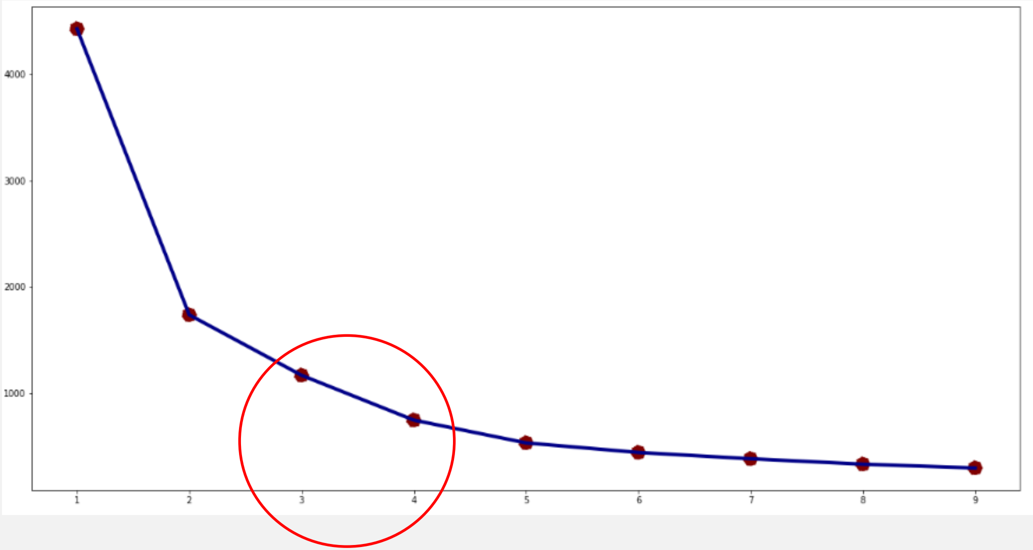
STANDARDIZATION

Standardize the numerical columns in dataset

```
# Subtask 4 : melakukan standarisasi

from sklearn.preprocessing import StandardScaler
df2_scaled = df2.copy()
ss = StandardScaler()

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

display(df2_scaled.shape, df2_scaled.head(3))
```

DATA MODELING

1. Find the number of cluster using elbow method

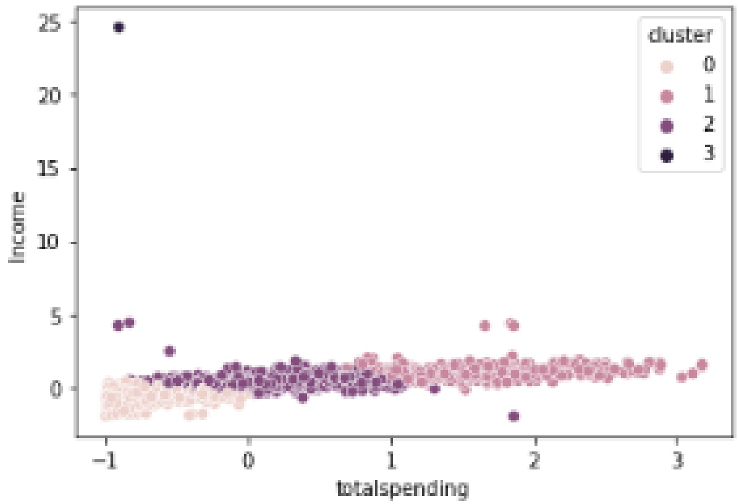


Based on elbow method, the best number of cluster is between 3 and 4

DATA MODELING

2. Data segmentation using Kmeans clustering

<matplotlib.axes._subplots.AxesSubplot at 0x7fc69c8f8bd0>

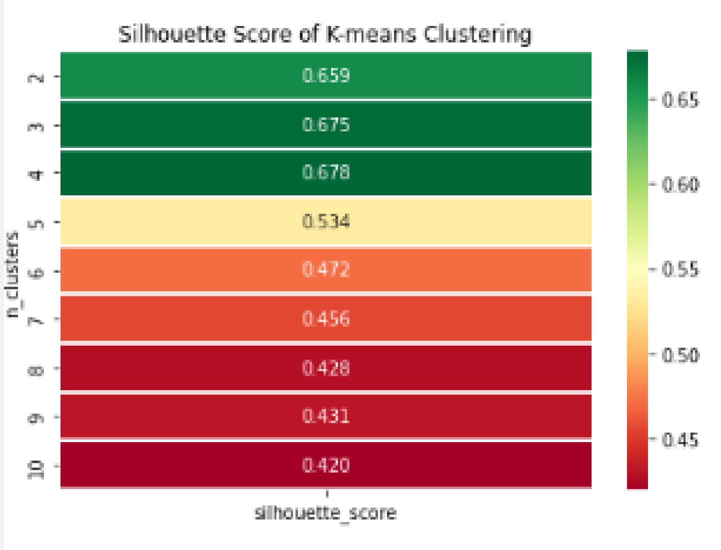


Lets try the data segmentation for 4 clusters. The result shows that the distribution of data is well segmented in each cluster. But cluster no 3 is most likely to be an outlier because of its distribution in plot and the number of data in this cluster is only 1

	totalspending		Income	
	count	mean	count	mean
cluster				
0	1122	1.258191e+05	1122	3.478553e+07
1	449	1.573203e+06	449	7.828739e+07
2	642	7.751807e+05	642	6.327413e+07
3	1	6.200000e+04	1	6.666660e+08

DATA MODELING

3. Generate silhouette score to evaluate model

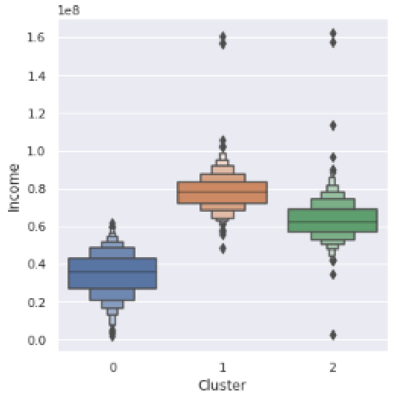


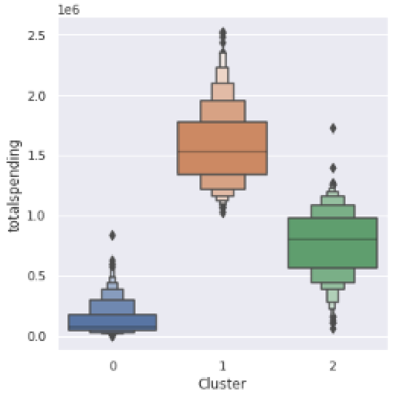
	totalspending		Income	
	count	mean	count	mean
cluster				
0	1122	1.258191e+05	1122	3.478553e+07
1	449	1.573203e+06	449	7.828739e+07
2	642	7.751807e+05	642	6.327413e+07
3	1	6.200000e+04	1	6.666660e+08

Based on plot, the highest silhouette score is for 3 clusters and 4 clusters. After further analysis, there is 1 cluster with unsuitable value and distribution which most likely an outlier. Therefore, we decided to drop the 4th cluster so the final number of cluster is 3.

```
# drop row untuk nilai cluster = 3
df2 = df2[df2.cluster != 3]
```

CLUSTER IDENTIFICATION
BASED ON INCOME AND TOTAL SPEND





Cluster identification

1. Cluster 0
- Have the lowest income based on its cluster which is by avg IDR 35.683.000 per year
- Also the lowest total amount spend by which is by avg IDR 70.000
- Then categorized as low spender

2. Cluster 1
- Have the highest income which is by avg IDR 78.093.000 per year
- Have the highest amount of total spending which is IDR 1.529.000
- Then categorized as high spender

3. Cluster 2
- Have the second highest income which is by avg IDR 62.559.500 per year
- Also the second highest total spending which is by avg IDR 795.000
- Then categorized as mid spender

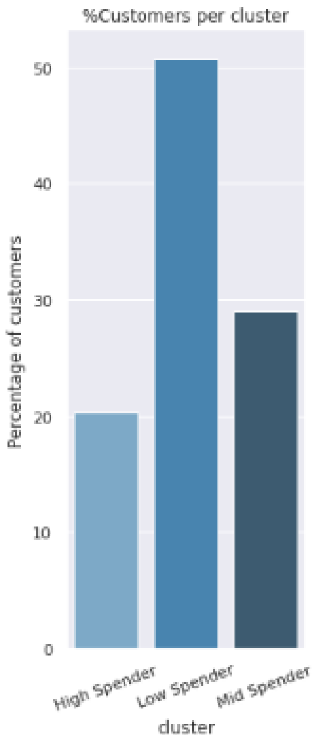
```
map_cluster = {
    0 : 'Low Spender',
    1 : 'High Spender',
    2 : 'Mid Spender'
}

df2['cluster_mapped'] = df2['cluster'].map(map_cluster)
```

CUSTOMER PROFILE

	age_range				Marital_Status			
	count	unique	top	freq	count	unique	top	freq
cluster_mapped								
High Spender	449	3	older_adults	214	449	5	Menikah	164
Low Spender	1122	3	middle_aged_adults	662	1122	6	Menikah	434
Mid Spender	642	3	older_adults	344	642	6	Menikah	258

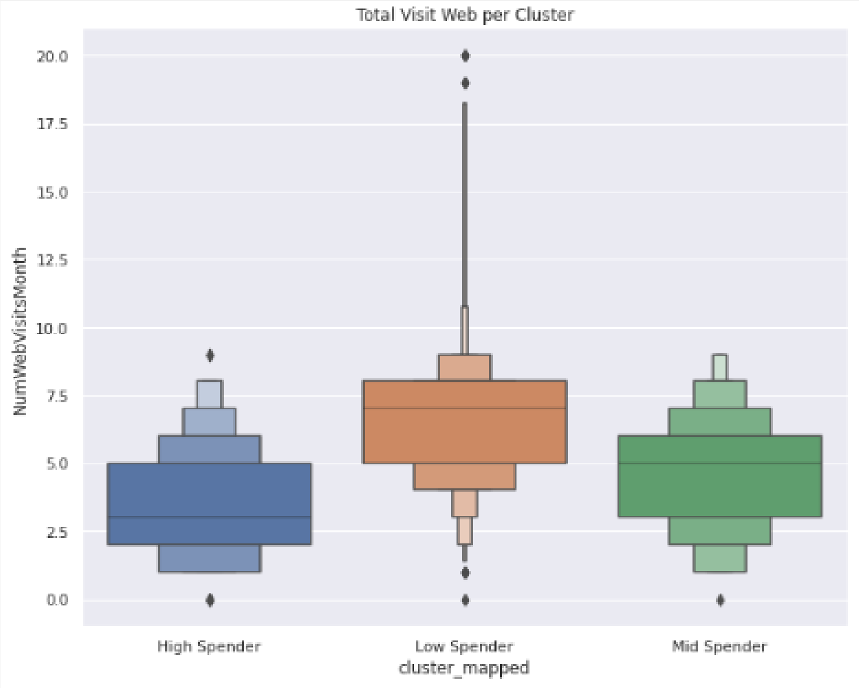
	cluster	avg kids	median kids
0	High Spender	0.293987	0.0
1	Low Spender	1.224599	1.0
2	Mid Spender	0.922118	1.0

- Low spender category mostly are middle aged adults (36-55 years old), have married and have 1 kids.
- Mid spender category are in older adults (>55 years old), have married and have 1 kids
- High spender are older adults (>55 years old), have married and have no kids.



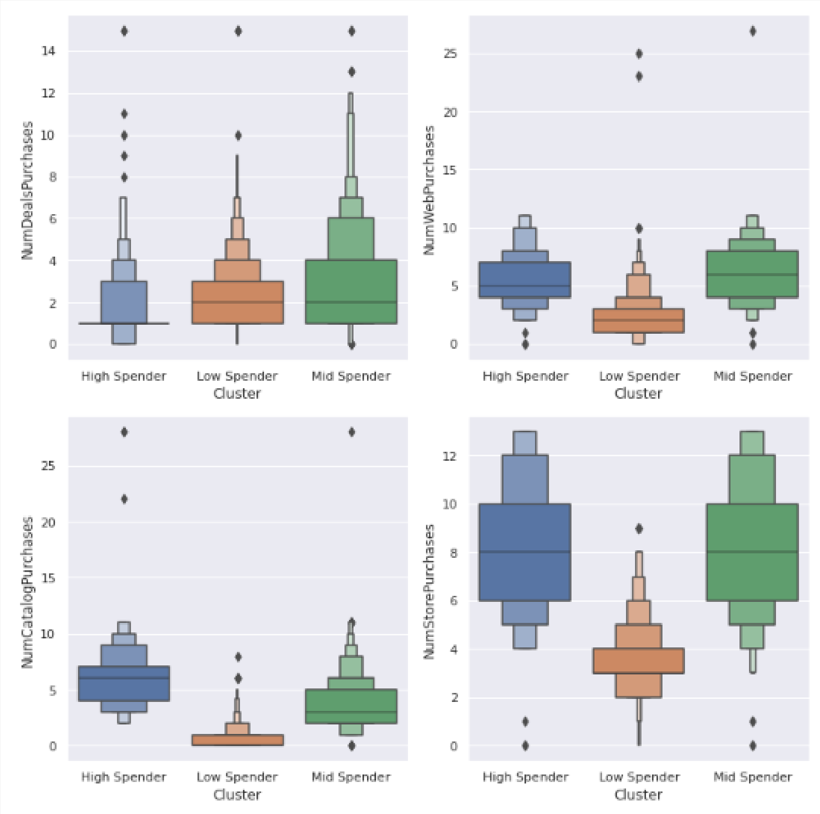
Most customers which is half of the population (50,7%) are categorized as low spender who have the lowest average amount of income and total spend

CUSTOMER BASED ON TOTAL VISIT WEB



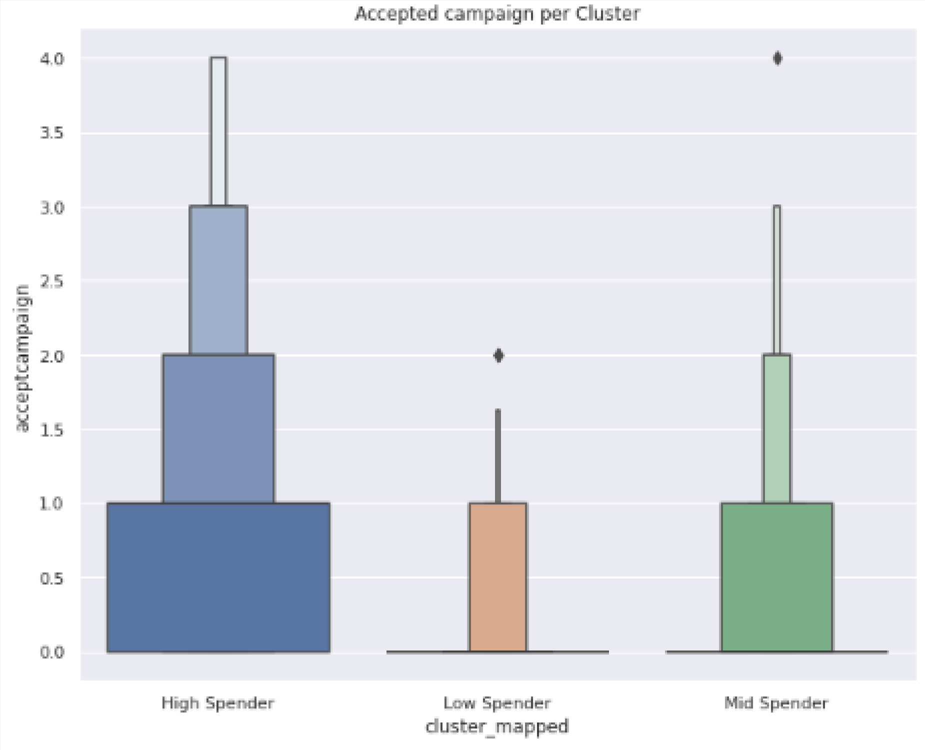
Low spender category have the highest amount of total web visit which is 6-7 times per month. Meanwhile high spender have the lowest amount which is by average 3 times per month.

CUSTOMER BASED ON PURCHASING HISTORY



- Most of the deals purchase user are low spender and mid spender category who approximately using promo for 2 times per month
- Mid spender have highest number of web purchase then followed by high spender
- High spender category are most likely to purchase by catalog purchase meanwhile low spender by avg have 0 purchase by catalog
- Both high spender and mid spender have the most purchase by store which by avg 8 times per month

CUSTOMER BASED ON ACCEPTED CAMPAIGN



High spender are most likely to accept campaign where at least accepting 1 campaign. Meanwhile most users in low spender and mid spender are not following any campaign (avg 0 campaign)

SUMMARY
ON CUSTOMER PROFILE

Low Spender

- Most of the customers in this category are middle aged adults (36-55 years old), have married and have 1 kids. Half of the customers are low spender customers
- Have the most web visit which approx 6-7 times per month but have the least amount of purchase by web (2 times)
- Most of the customer are not accepting any campaign
- Using deals promo at least 2 times per month
- Have the lowest rate of income and total spending amount

Mid Spender

- Most of the customers are in older adults age (>55 years old), have married and have 1 kids
- They have the highest number of web purchase (6 times) eventhough visited web 4-5 times per month
- Also have high amount of purchase by store at least 8 times
- Mid spender are using deals promo at least 2 times per month and doesnt accept any campaign
- Have high amount of income around 62 mios but have low total spending amount which only 795.000 or 1.27% percent of their income

High Spender

- Most of the customers are an older adults (>55 years old), have married and have no kids.
- Have the highest amount of purchasing history by store and catalog. Both at least 8 times
- They have high number of web purchase (5 times) despite the low number of web visits which only 3 times per month.
- Accepting at least 1 campaign and used deals promo at least 1 times.
- Have the highest rate of income and total spending amount

BUSINESS RECOMMENDATION

Identify why low spender category have low purchase in web eventhough they have the highest number of web visit.

This could probably happened because of uncomplete product category, unsuitable product price, high shipping cost, high service cost, etc.

BUSINESS RECOMMENDATION

Increase the purchasing service by store, where most of customers from three categories are most likely to shop by store.

BUSINESS RECOMMENDATION

Develop features in web purchase to increase the number of web visits particularly for high spender category who have high amount of web purchase but lowest amount of web visit and low spender category who have highest web visit but low web purchase.

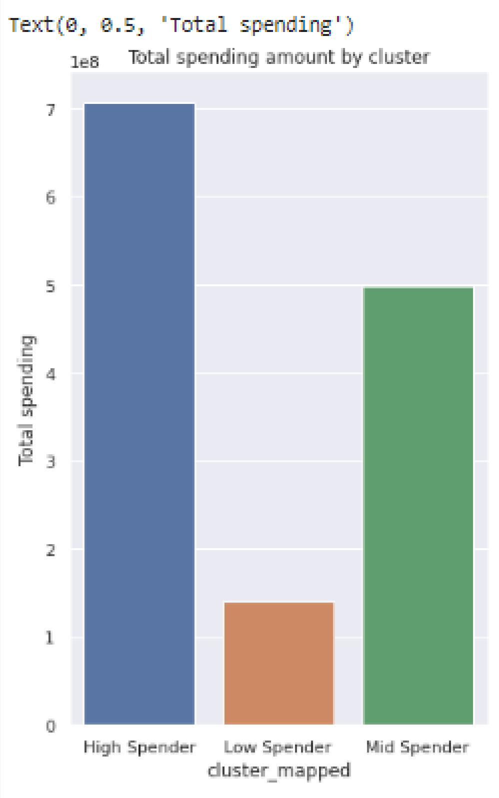
Feature that are recommended to be added are :

- Clicked ads to increase the number of web visits
- Product recommender to increase purchasing service using web
- Abandoned cart email to increase web purchase

BUSINESS RECOMMENDATION

Give rewards for certain amount of spending by giving coupon or voucher to customers especially for mid spender category who have high number of deals purchased but low amount of total spending.

CONCLUSION



- Based on analyses, the market retargeting is focused to high spender and mid spender category who have high amount of income so they have more potential to do more spending compared to other category.
- High spender have potential GMV for IDR 706.368.000
- Mid spender have potential GMV for IDR 497.666.000
- The amount of potential reduction cost to do promo optimization for mid spender category (assume the 50% reduction) is IDR 70.978.836