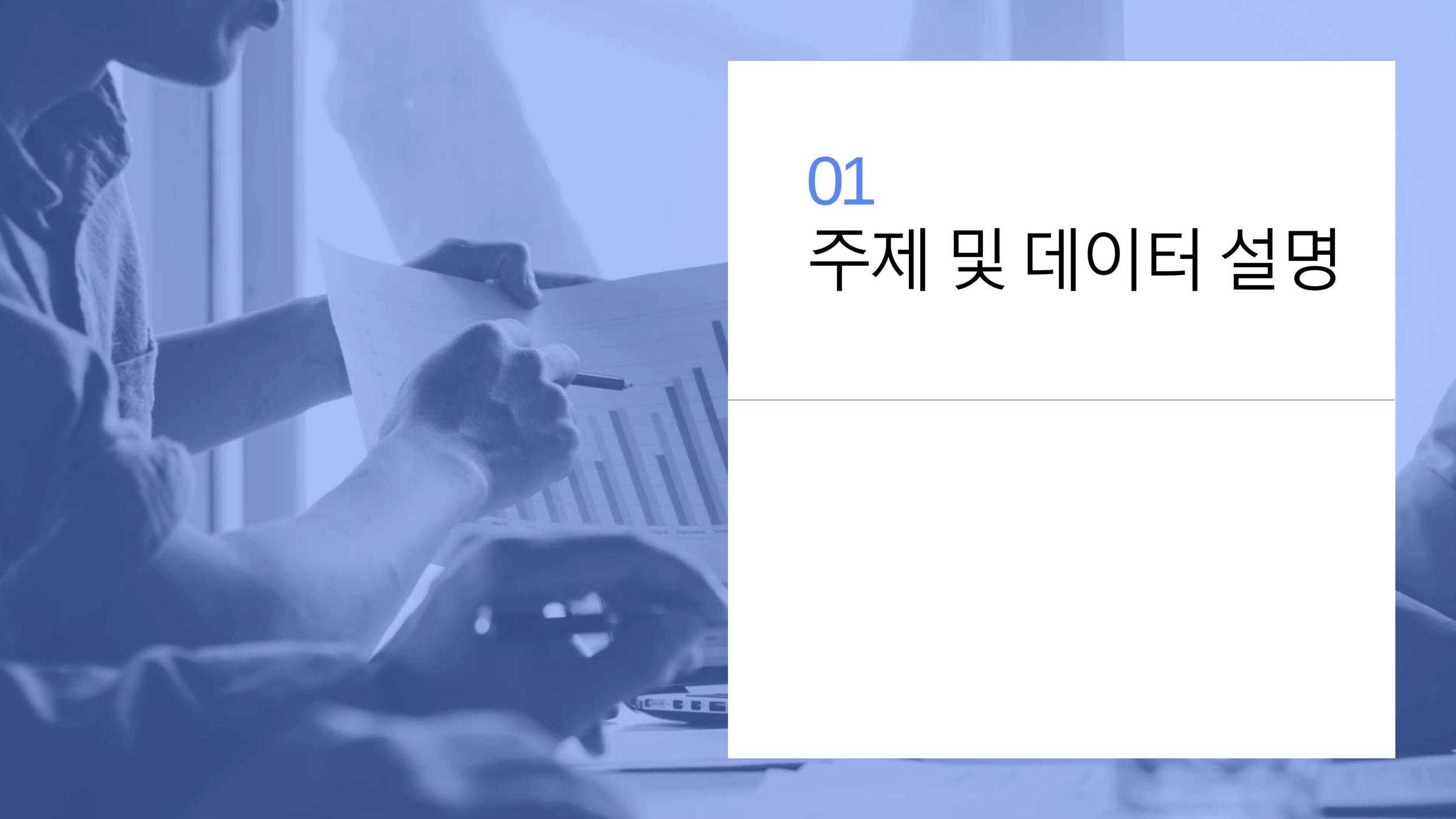
Customer Personality Analysis

22-2 SookTat Kaggle Project

Team 2
고나경, 임주영, 조민영,
정재원, 최호경



01

주제 및 데이터 설명

Customer Personality Analysis - Subject

■ 'marketing_campaign.csv'

```
data = pd.read_csv("marketing_campaign.csv", sep="₩t")
data
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	...	7	0
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	...	5	0
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	...	4	0
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	...	6	0
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	...	5	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2235	10870	1967	Graduation	Married	61223.0	0	1	13-06-2013	46	709	...	5	0
2236	4001	1946	PhD	Together	64014.0	2	1	10-06-2014	56	406	...	7	0
2237	7270	1981	Graduation	Divorced	56981.0	0	0	25-01-2014	91	908	...	6	0
2238	8235	1956	Master	Together	69245.0	0	1	24-01-2014	8	428	...	3	0
2239	9405	1954	PhD	Married	52869.0	1	1	15-10-2012	40	84	...	7	0

2240 rows × 29 columns

Customer Personality Analysis - Data

People

- ID : Customer's unique identifier
- Year_Birth : Customer's birth year
- Education : Customer's education level
- Marital_Status : Customer's marital status
- Income: Customer's yearly household income
- Kidhome : Number of children in customer's household
- Teenhome : Number of teenagers in customer's household
- Dt_Customer : Date of customer's enrollment with the company
- Recency : Number of days since customer's last purchase
- Complain : 1 if the customer complained in the last 2 years, 0 otherwise

Product

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Promotion

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place

- NumWebPurchases: Number of purchases made through the company’s website
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company’s website in the last month

02
데이터 전처리 및 EDA

데이터 확인

이상치, 결측치 처리

추가변수, 파생변수 생성

변수간 상관관계 파악

PCA

Customer Personality Analysis - 결측치

■ 'Income' 변수 결측치

```
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Complain             2240 non-null   int64
 26  Z_CostContact        2240 non-null   int64
 27  Z_Revenue            2240 non-null   int64
 28  Response             2240 non-null   int64
dtypes: float64(1), int64(25), object(3)
```

→

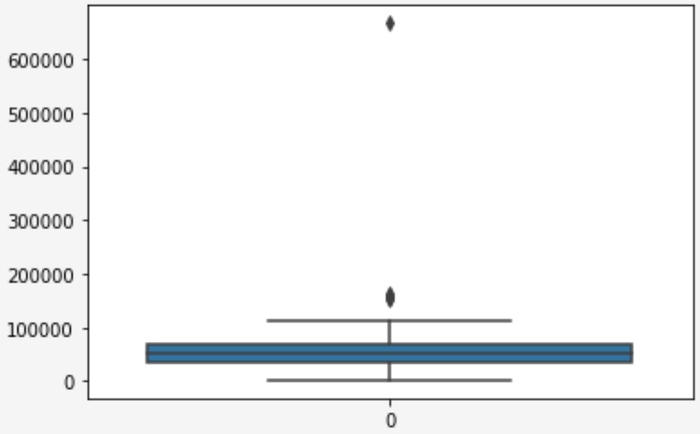
```
customer_df.dropna(inplace=True)
```

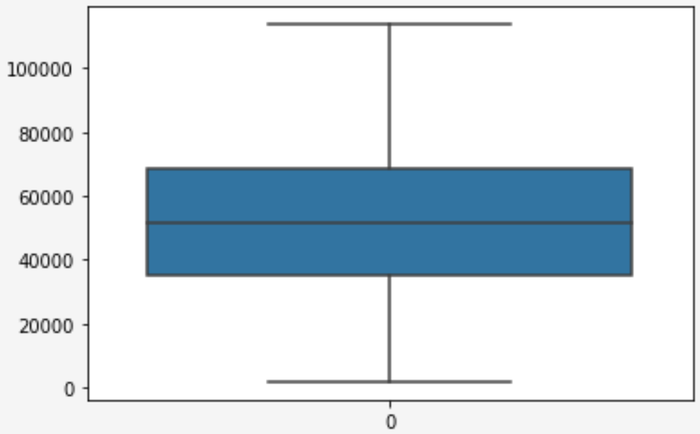
```
customer_df.isnull().sum()
```

```
ID                   0
Year_Birth           0
Education            0
Marital_Status       0
Income               0
Kidhome              0
Teenhome             0
Dt_Customer          0
Recency              0
MntWines             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
dtype: int64
```

Customer Personality Analysis - 이상치

■ 'Income' 변수 이상치






```
# 4분위수, IQR 계산

q1 = customer_df['Income'].quantile(0.25)
q2 = customer_df['Income'].quantile(0.5)
q3 = customer_df['Income'].quantile(0.75)

iqr = q3 - q1
iqr
```

```
33219.0
```

```
# outlier cutoff, lower/upper bound 계산

cut_off = iqr * 1.5

lower = q1 - cut_off
upper = q3 + cut_off

print(lower)
print(upper)
```

```
-14525.5
118350.5
```

```
# 1사분위와 4사분위에 속해있는 데이터 각각 저장

data1 = customer_df[customer_df['Income'] > upper]
data2 = customer_df[customer_df['Income'] < lower]
```

```
# 이상치 개수

data1.shape[0] + data2.shape[0]
```

```
8
```

```
data1['Income']
```

```
164     157243.0
617     162397.0
655     153924.0
687     160803.0
1300    157733.0
1653    157146.0
2132    156924.0
2233    666666.0
Name: Income, dtype: float64
```

```
customer_df[customer_df['Income'] >= 153924]
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome
164	8475	1973	PhD	Married	157243.0	0
617	1503	1976	PhD	Together	162397.0	1
655	5555	1975	Graduation	Divorced	153924.0	0
687	1501	1982	PhD	Married	160803.0	0
1300	5336	1971	Master	Together	157733.0	1
1653	4931	1977	Graduation	Together	157146.0	0
2132	11181	1949	PhD	Married	156924.0	0
2233	9432	1977	Graduation	Together	666666.0	1

Customer Personality Analysis - 이상치

■ 'Age', 'MntSpent' 변수 이상치

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

```
customer_df[customer_df['Age'] > 100]
```

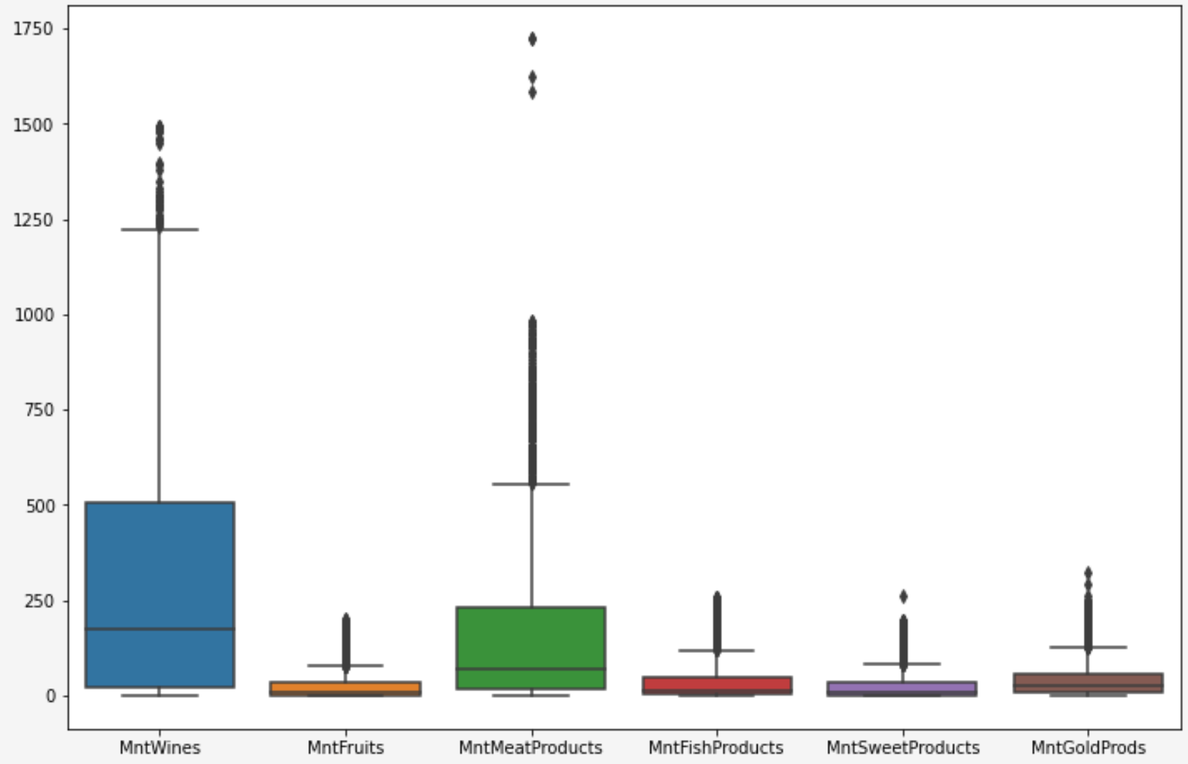
eptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response	Age
0	0	0	1	3	11	0	122
0	0	0	0	3	11	0	129
1	0	0	0	3	11	0	123

```
outlier_idx2 = customer_df[customer_df['Age'] > 100].index
outlier_idx2
```

```
Int64Index([192, 239, 339], dtype='int64')
```

```
customer_df.drop(outlier_idx2, inplace=True)
```

→ 100초과인 데이터 제거



→ 지출 금액 관련 변수에 대한 이상치는 제거하거나 따로 처리하지 않고 그대로 사용함

Customer Personality Analysis - 추가변수

■ 추가변수, 파생변수 생성

1 People

가구 내 총 자녀 수

('Children' = 'Kidhome' + Teenhome')

2 Product

2년간 지출 총액

('TotalMntSpent' = 'MntWines' + 'MntFruits' + 'MntFruits' + 'MntMeatProducts'
+ 'MntFishProducts' + 'MntSweetProducts' + 'MntGoldProds')

3 Promotion

승인된 총 캠페인 수

('Total_Acc_Cmp' = 'AcceptedCmp1' + AcceptedCmp2' + AcceptedCmp3'
+ AcceptedCmp4' + 'AcceptedCmp5' + 'Response')

4 Place

총 구매 건수

('TotalNumPurchases' = 'NumWebPurchases' + 'NumCatalogPurchases'
+ 'NumStorePurchases' + 'NumDealsPurchases')

Customer Personality Analysis - 추가변수

■ 추가변수, 파생변수 생성

	Marital_Status	Partner
0	Single	No
1	Single	No
2	Together	Yes
3	Together	Yes
4	Married	Yes
...	...	...
2200	Married	Yes
2201	Together	Yes
2202	Divorced	No
2203	Together	Yes
2204	Married	Yes

파트너 유무

	Family_Size	Partner	Children
0	1	No	0
1	3	No	2
2	2	Yes	0
3	3	Yes	1
4	3	Yes	1
...	...	...	...
2200	3	Yes	1
2201	5	Yes	3
2202	1	No	0
2203	3	Yes	1
2204	4	Yes	2

총 가족 구성원수

	Children	Is_Parent
0	0	0
1	2	1
2	0	0
3	1	1
4	1	1
...	...	...
2200	1	1
2201	3	1
2202	0	0
2203	1	1
2204	2	1

자녀 여부

	Education	Education_Level
0	Graduation	Graduate
1	Graduation	Graduate
2	Graduation	Graduate
3	Graduation	Graduate
4	PhD	Postgraduate
...	...	...
2200	Graduation	Graduate
2201	PhD	Postgraduate
2202	Graduation	Graduate
2203	Master	Postgraduate
2204	PhD	Postgraduate

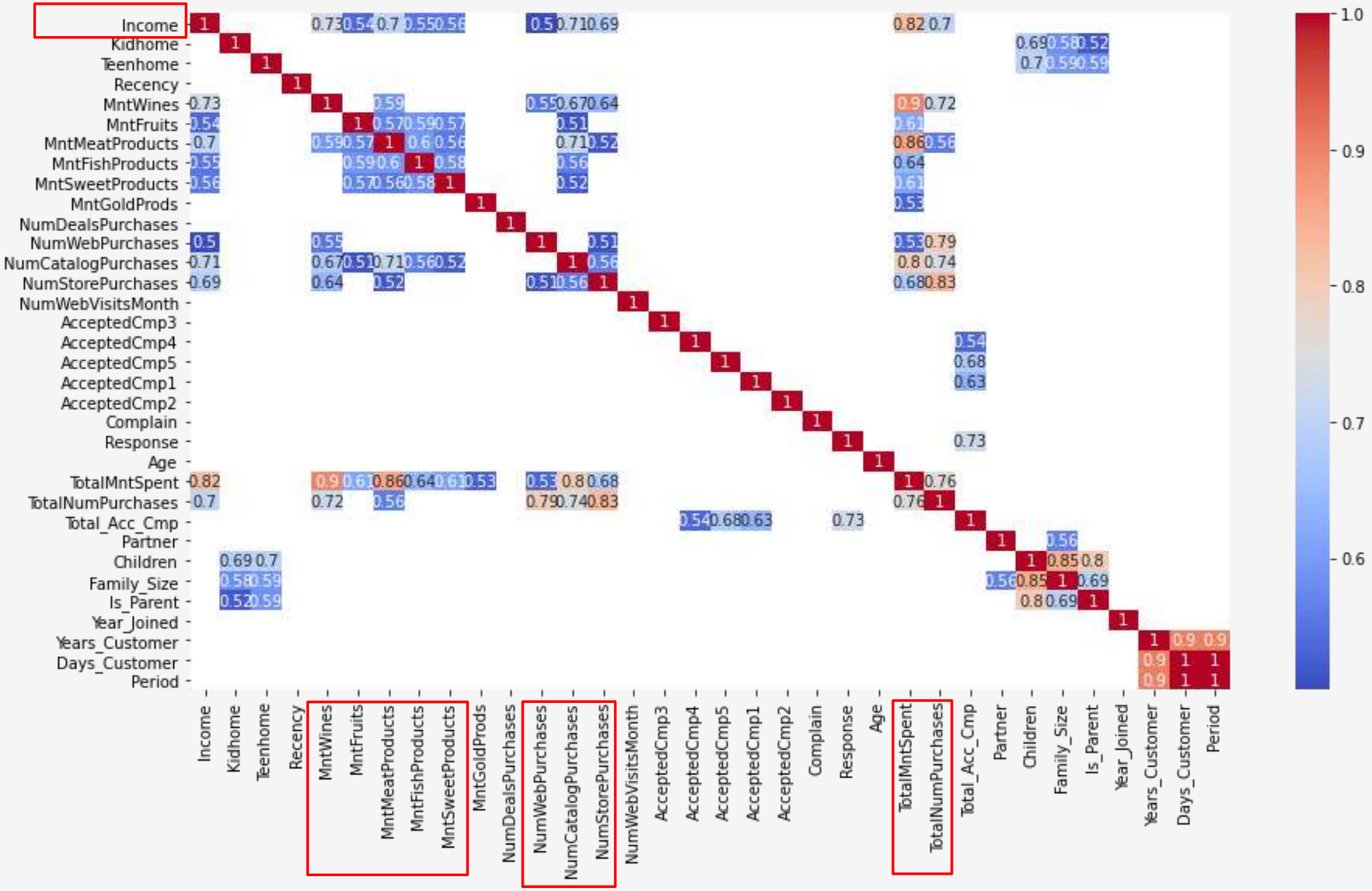
교육 수준 그룹화

	Dt_Customer	Period
0	2012-04-09	3751
1	2014-08-03	2905
2	2013-08-21	3252
3	2014-10-02	2845
4	2014-01-19	3101
...	...	...
2200	2013-06-13	3321
2201	2014-10-06	2841
2202	2014-01-25	3095
2203	2014-01-24	3096
2204	2012-10-15	3562

가입 일수

Customer Personality Analysis - 상관분석

■ 상관관계 분석



Customer Personality Analysis - PCA

1 PCA 실행

1. 연속형 변수 사용

'AcceptedCmp1','AcceptedCmp2','AcceptedCmp3','AcceptedCmp4','AcceptedCmp5',
"Complain","Response","Partner","Education_Level"

2. StandardScaler()로 데이터 정규화 (단위 상이)

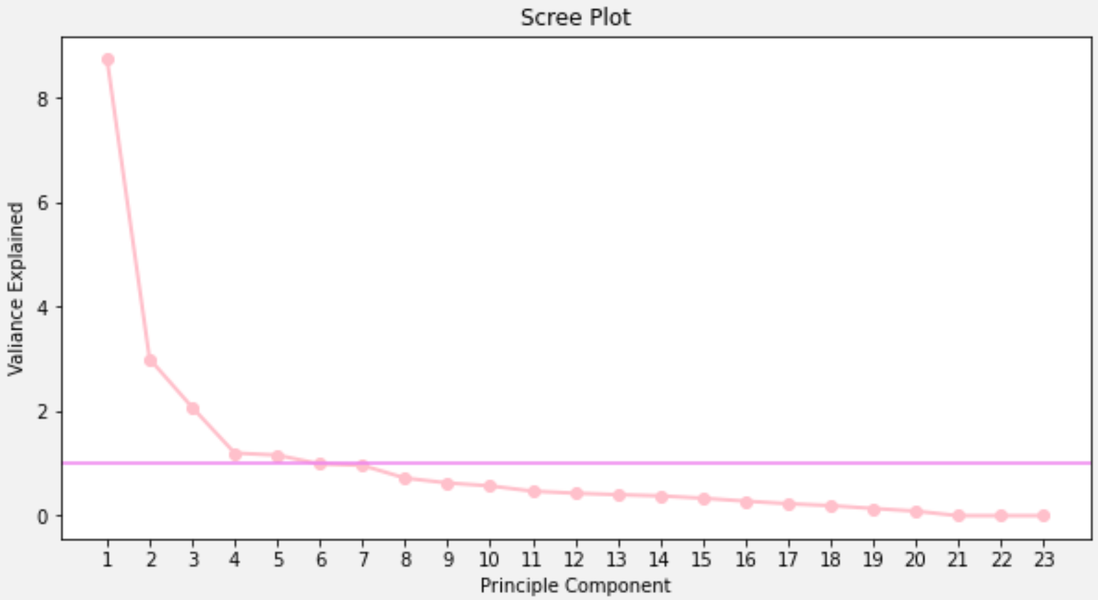
3. PCA()로 PCA실행

PCA 결과 (각 성분 개수에 따른 표준편차, 분산비율, 누적 분산비율)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21
Standard deviation	2.9578	1.7332	1.4413	1.0955	1.0775	0.9954	0.9822	0.8490	0.7919	0.7558	0.6853	0.6564	0.6348	0.6159	0.5788	0.5281	0.4794	0.4389	0.3722	0.2957	0.0
Proportion of variance	0.3802	0.1306	0.0903	0.0522	0.0505	0.0431	0.0419	0.0313	0.0273	0.0248	0.0204	0.0187	0.0175	0.0165	0.0146	0.0121	0.0100	0.0084	0.0060	0.0038	0.0
Cumulative proportion	0.3802	0.5107	0.6010	0.6532	0.7036	0.7467	0.7886	0.8199	0.8472	0.8720	0.8924	0.9111	0.9287	0.9451	0.9597	0.9718	0.9818	0.9902	0.9962	1.0000	1.0

Customer Personality Analysis - PCA

2 성분 개수 결정



Scree Plot

x축을 주성분 개수, y축을 고윳값(설명가능한 분산 값)으로 하는 line graph

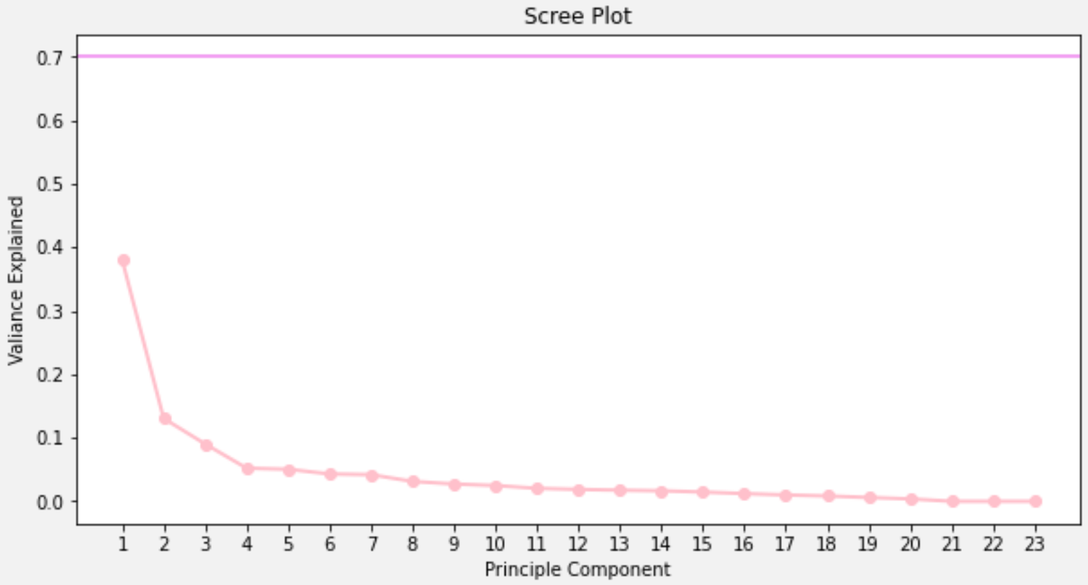
Kaiser's Rule

고유값이 1.0 이상인 주성분들을 선택

성분의 개수 6 or 7개가 적당

Customer Personality Analysis - PCA

2 성분 개수 결정



분산비율 선택법

전체 데이터에서 최소 70% 이상의 설명력을 가지는 주성분들을 선택

PC1도 0.4 이하인 상태

데이터가 굉장히 각 차원마다 분산되어 있는 형태로 예상

따라서, 되도록 많은 변수를 선택하는 것이 적절하다고 판단

성분의 개수를 7개로 결정

03
Modeling

K-Means Clustering Algorithm

최적의 k값 탐색

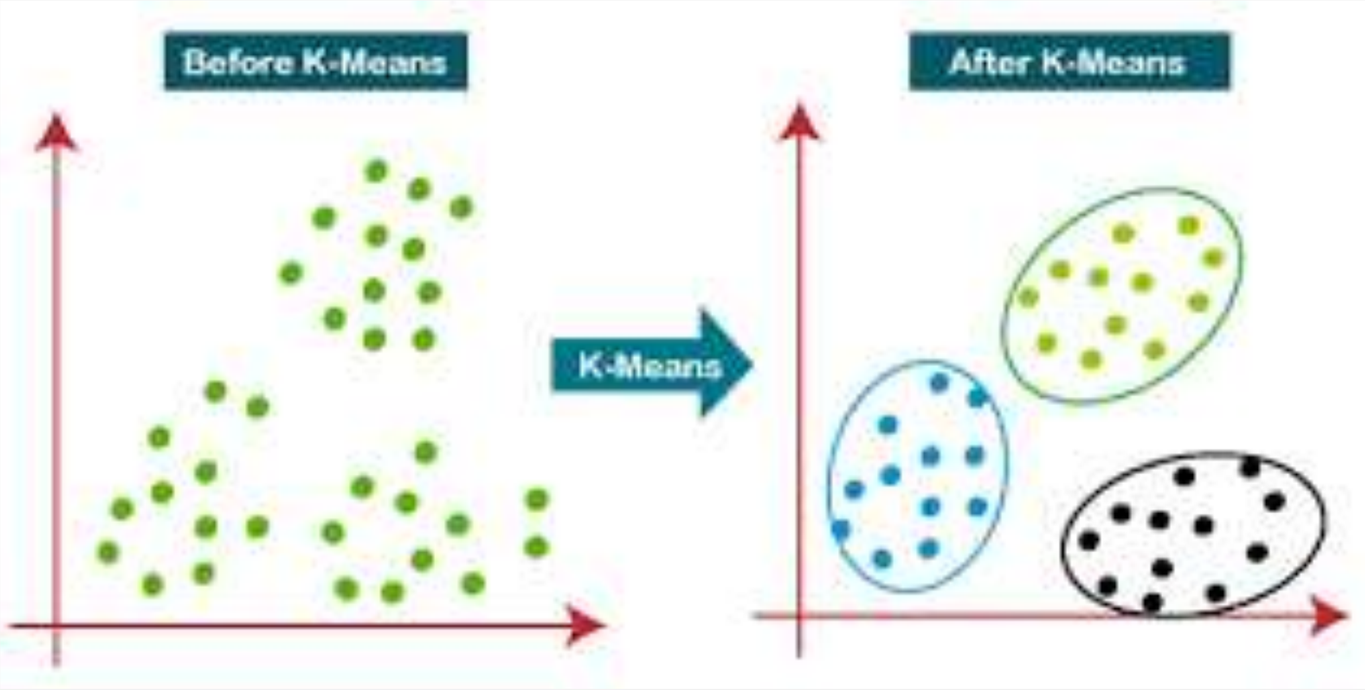
k 별 군집 특성 파악

Modeling

K-Means Clustering Algorithm

K – Means Clustering

주어진 데이터를 k개의 클러스터로 묶는 알고리즘
각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작



Modeling

K-Means Clustering Algorithm

K – Means Clustering

1. get_dummies( )를 이용하여 범주형 변수 가변수 생성
2. StandardScaler( )로 연속형 변수 스케일링
3. Kmeans( )로 k-means 모델 생성 및 데이터에 fit

label
2
1
0
1
0

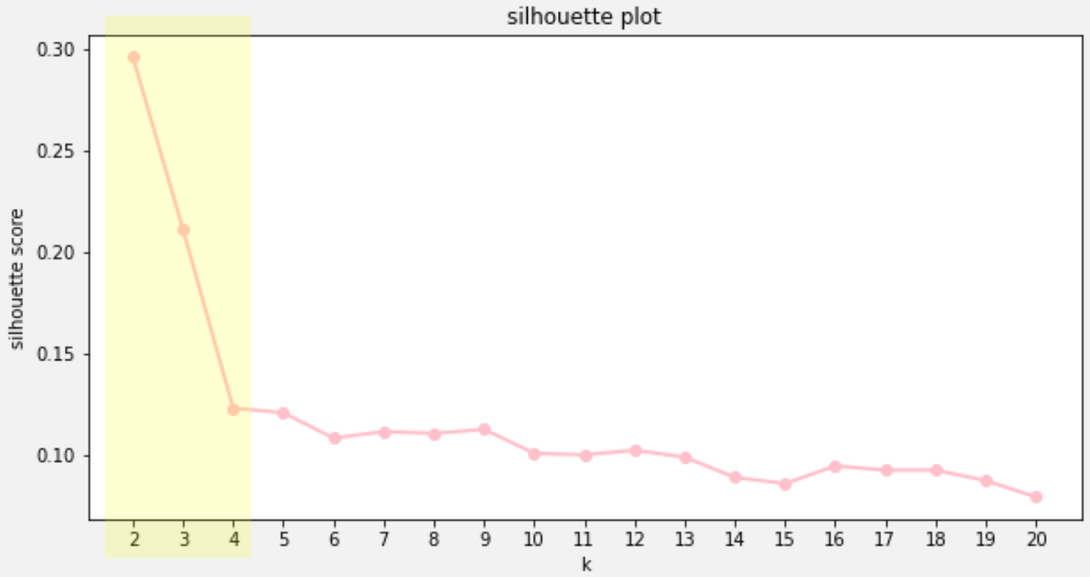
K-Means로 Clustering시킨 결과

Modeling

최적의 k값 탐색 – 기존 데이터

Silhouette Method

한 점이 다른 군집에 비해 자신의 군집과 얼마나 유사한지를 측정하여 (Silhouettee Score), 그 값이 가장 높은 k를 선택하는 방법



실루엣 스코어가 가장 높았던 k는 2 였으며, k=4를 기점으로 급격한 감소를 보임

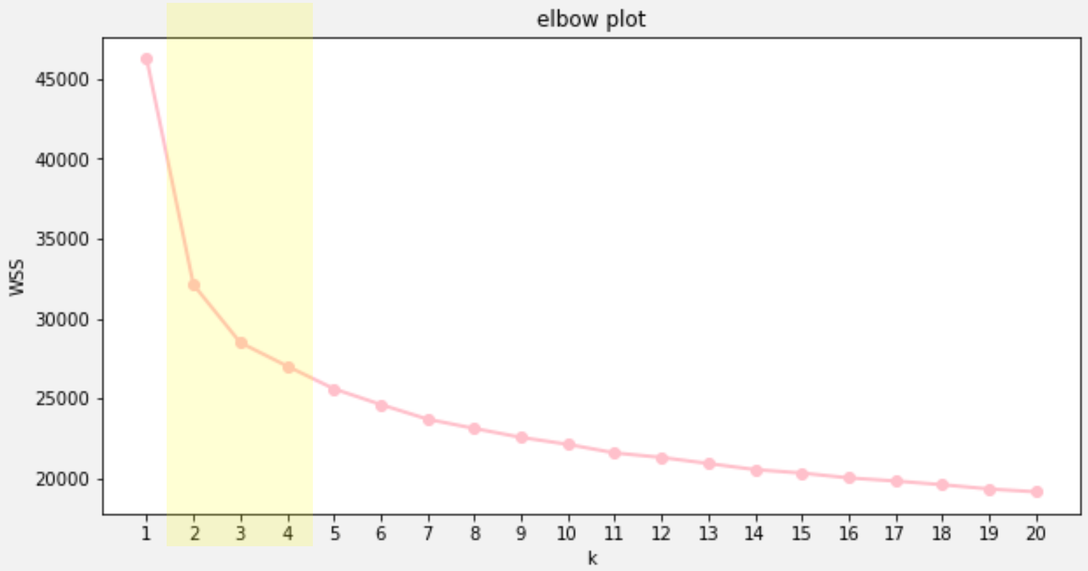
Modeling

최적의 k값 탐색 – 기존 데이터

Elbow Method

각 k개의 군집의 *WSS거리를 구하여, 짧은 WSS 값이 급격히 떨어지는 구간을 k를 선택한다.

* WSS(Within-Cluster-Sum of Squared) : 군집의 중심점에서 각 군집에 속한 데이터들의 거리를 다 합친 것



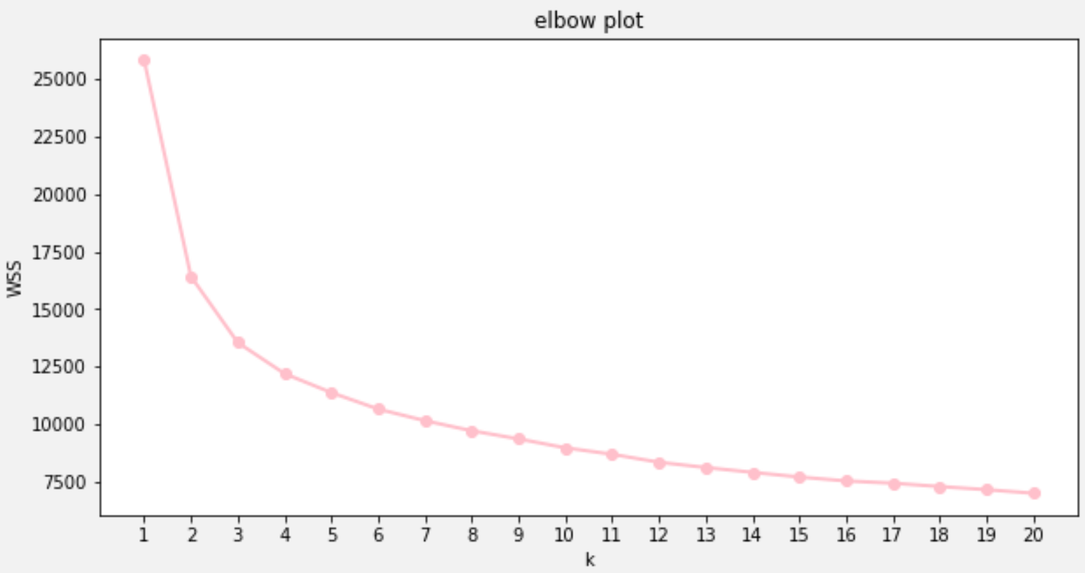
급격히 떨어지는 구간이 완만하여 (k=2 ~ k=4),
K값을 하나로 결정 짓지 못하였음.

K의 값을 2,3,4로 간주려 분석을 시행

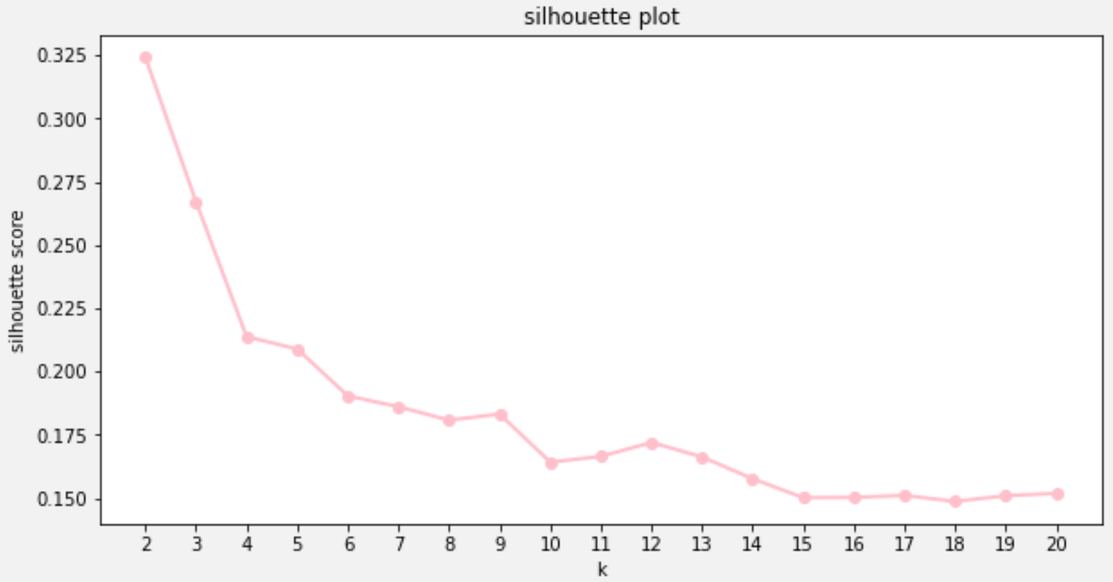
Modeling

최적의 k값 탐색 – PCA 데이터 (참고)

Elbow Method



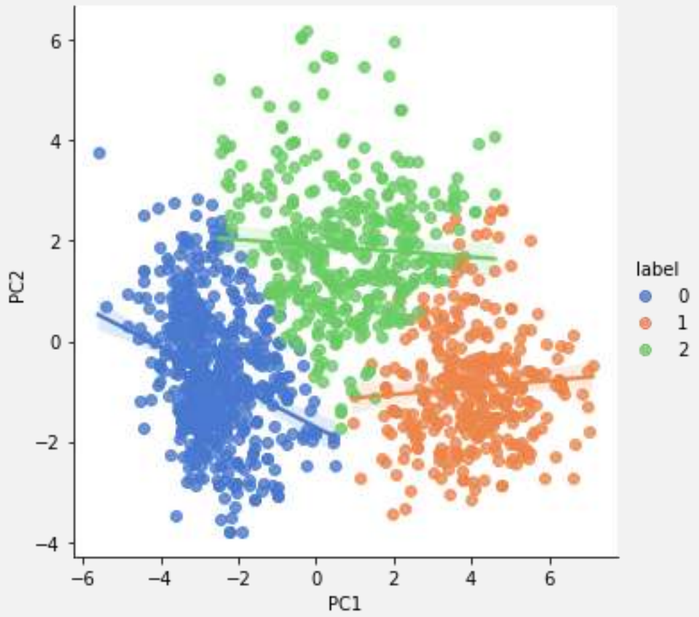
Silhouette Method



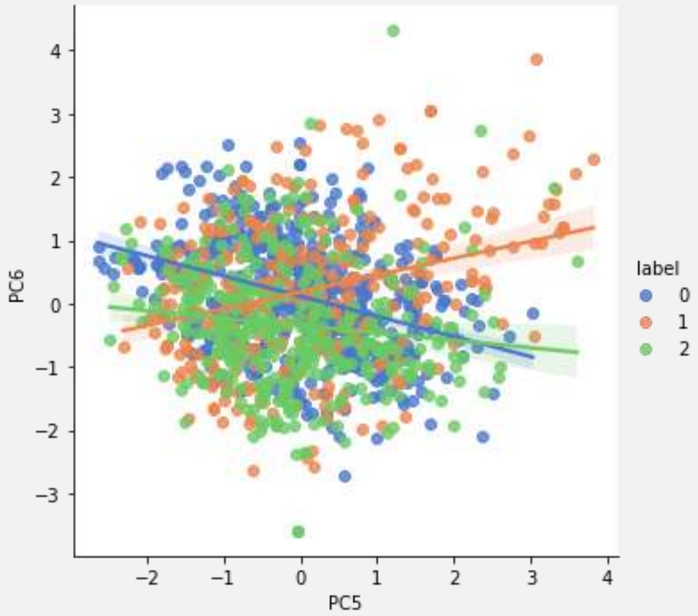
기존 데이터의 결과와 비슷한 모습을 띠었음

Modeling

PCA 데이터로 k-means 실행 결과 (참고)



저차원 성분 (PC1, PC2)로 분석을 하면
군집이 굉장히 잘 나뉘어진 걸 볼 수 있으나,



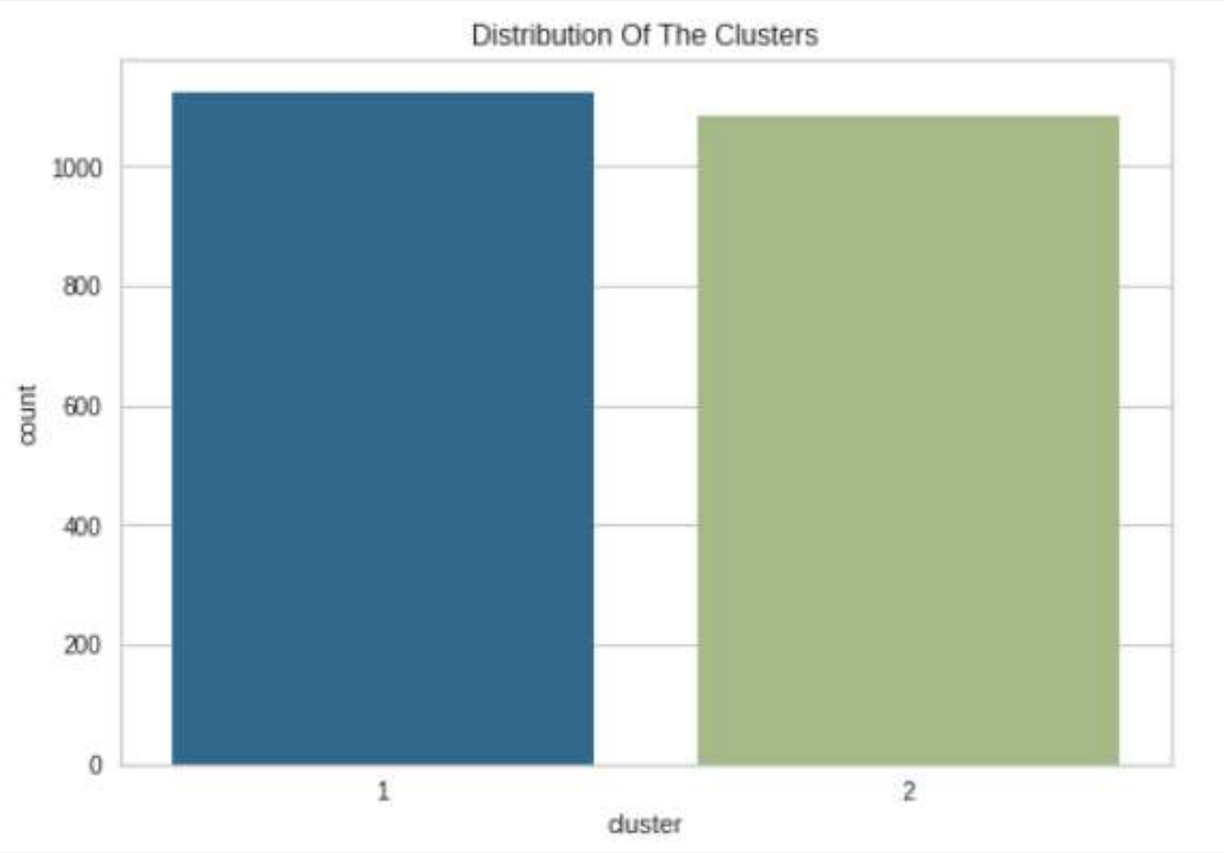
고차원 성분 (PC5, PC6) 로 분석을 할수록
군집이 잘 나뉘지 않는 양상을 보임

또한, PCA를 시킨 변수들로는
군집화 결과의 해석이 어렵다는 한계점이 있었음

Modeling

k별 군집 특성 ; k=2

1. 그룹별 분포




```
final_df['cluster'].value_counts()
```

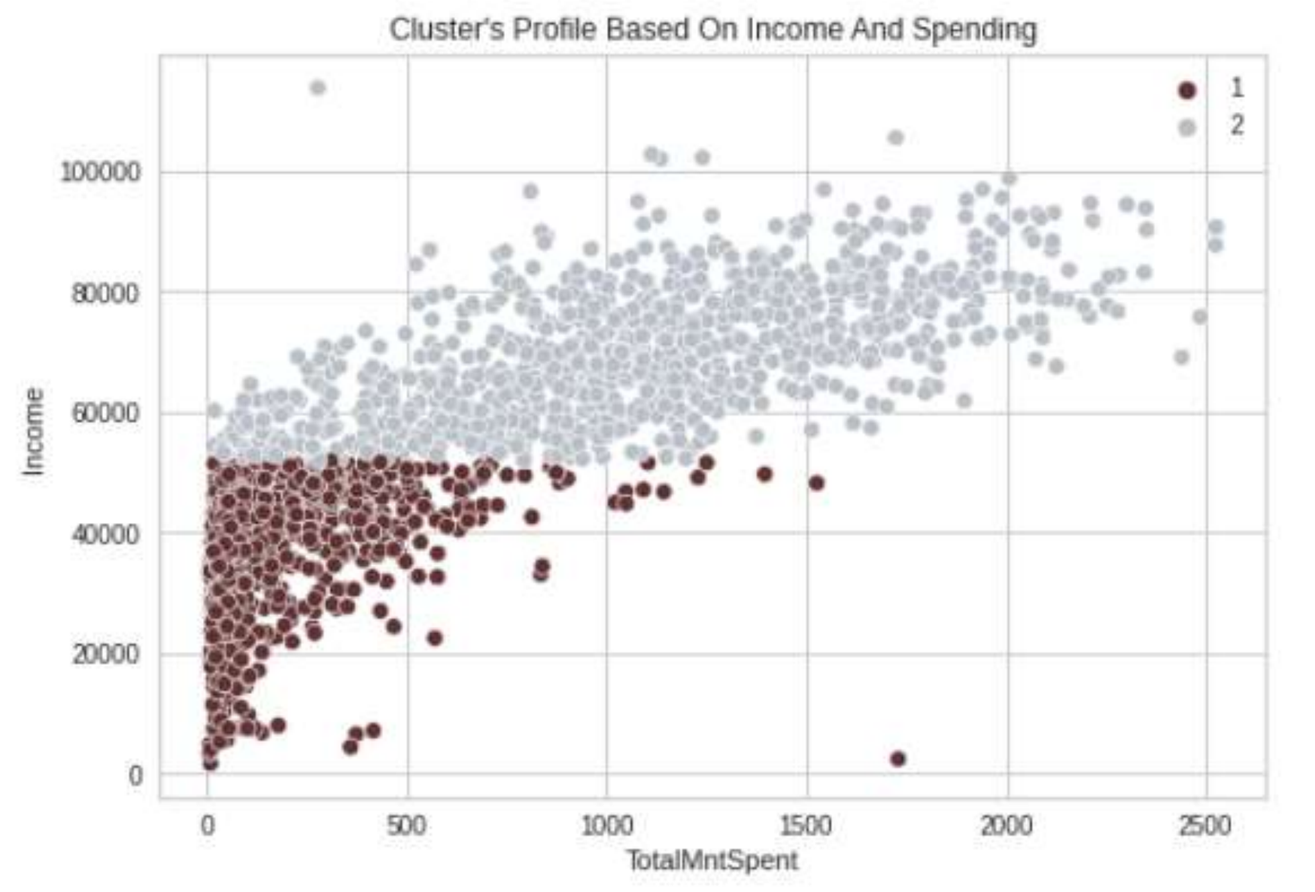
```
1    1123
2    1082
```

그룹 1 : 1123명
그룹 2 : 1082명

Modeling

k별 군집 특성 ; k=2

2. TotalMntSpent에 따른 Income



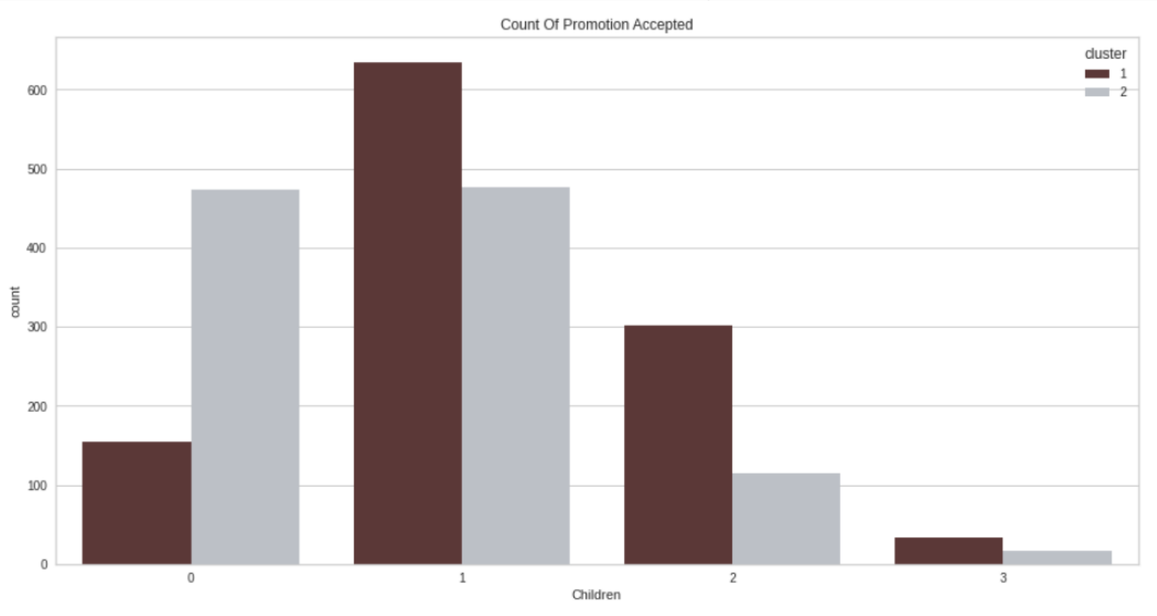
Income < 50000 : 그룹 1
Income > 50000 : 그룹 2.

0 < totalmntspent < 800 : 그룹1
0 < totalmntspent < 2500: 그룹2

Modeling

k별 군집 특성 ; k=2

3. Children, Age



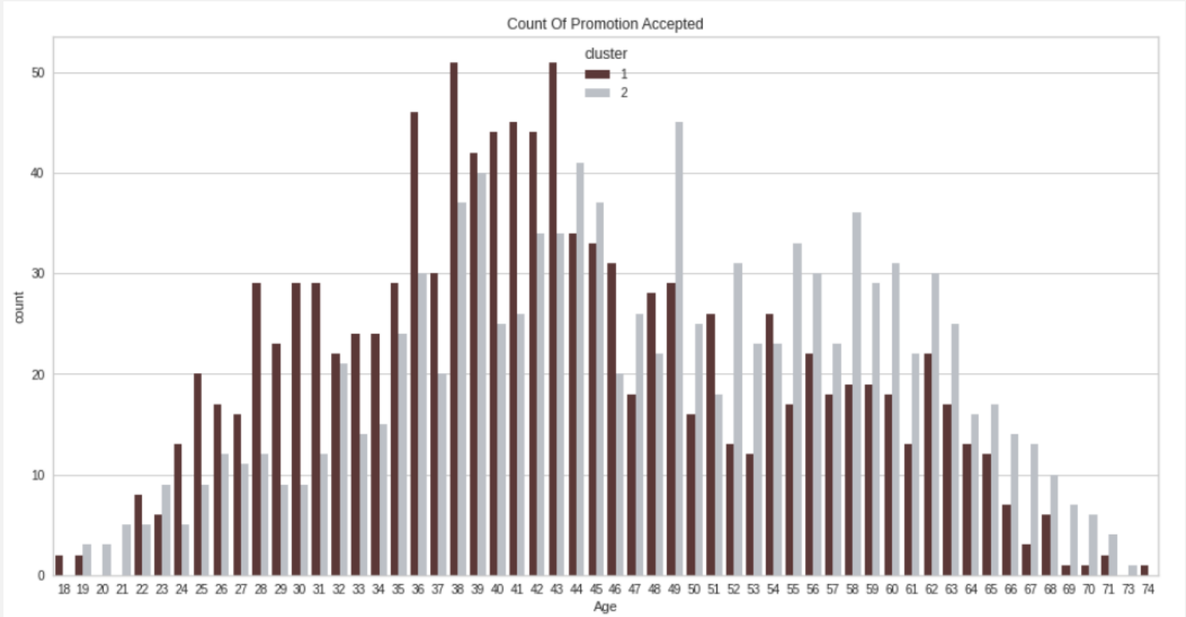

```
group1['Children'].mean()
```

```
1.1896705253784505
```

```
group2['Children'].mean()
```

```
0.6987060998151571
```

그룹을 나누는데
Children은 큰 영향을 미치지 않았다.




```
group1['Age'].mean()
```

```
42.97328584149599
```

```
group2['Age'].mean()
```

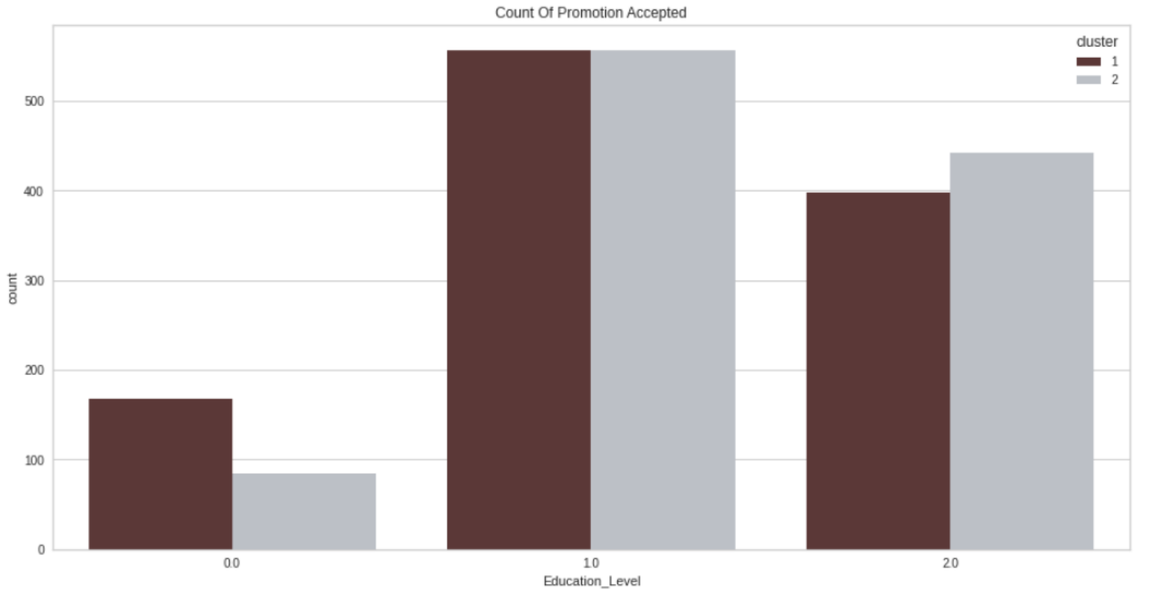
```
47.29852125693161
```

그룹을 나누는데
Age는 큰 영향을 미치지 않았다.

Modeling

k별 군집 특성 ; k=2

4. Education_Level, Days_Customerren



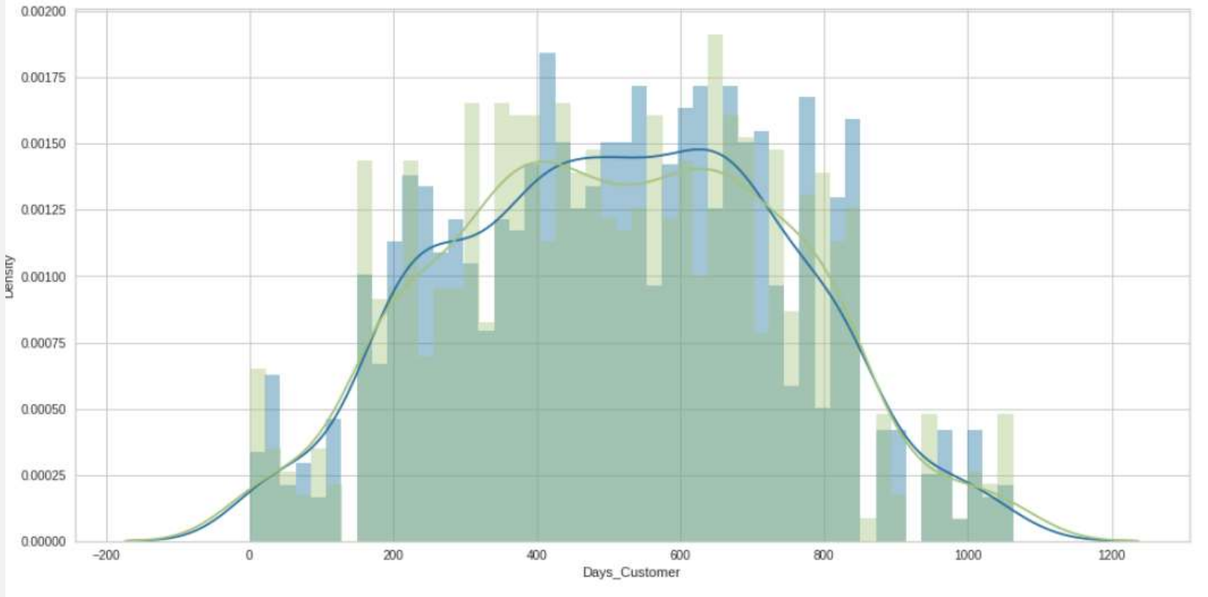

```
group1['Education_Level'].mean()
```

```
1.2048085485307214
```

```
group2['Education_Level'].mean()
```

```
1.3308687615526802
```

그룹을 나누는데
Education_Level은 큰 영향을 미치지 않았다.




```
group1['Days_Customer'].mean()
```

```
512.3722172751558
```

```
group2['Days_Customer'].mean()
```

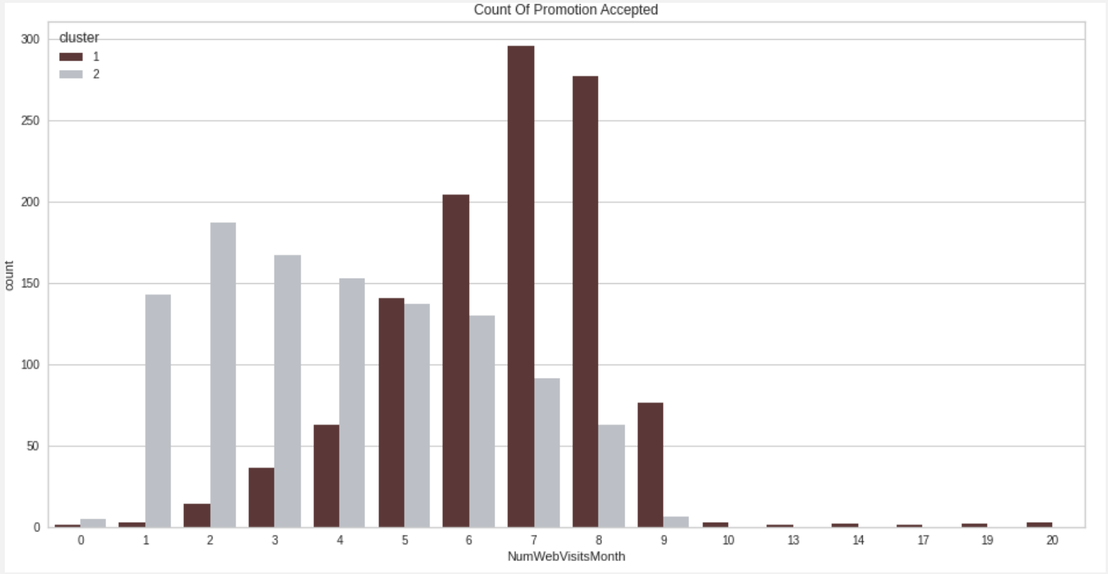
```
511.7412199630314
```

그룹을 나누는데
Days_Customer는 큰 영향을 미치지 않았다.

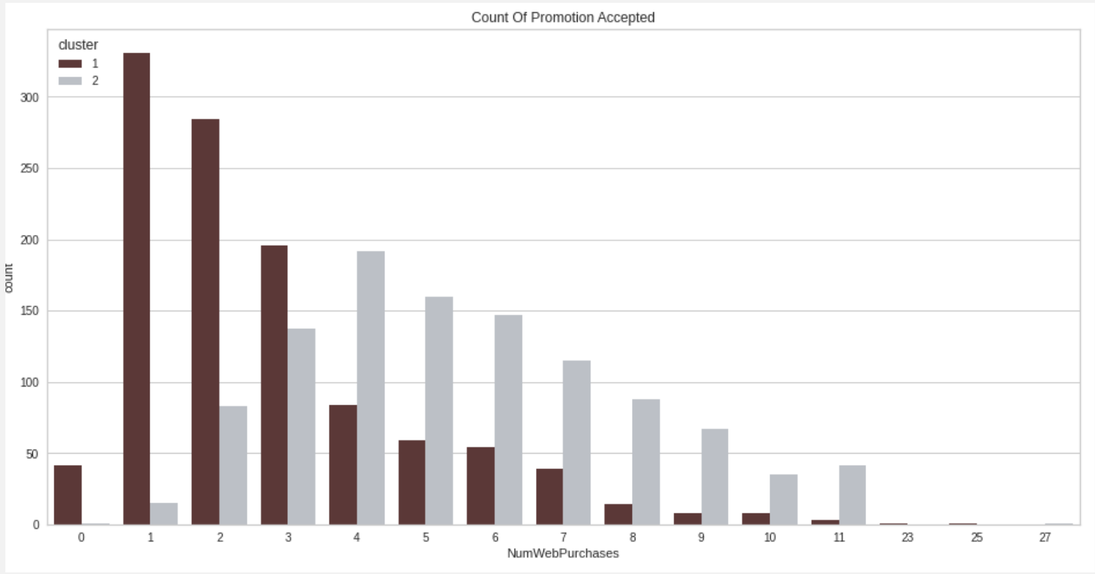
Modeling

k별 군집 특성 ; k=2

5. 구매한 방법에 따른 그룹 차이



NumWebVisitsMonth

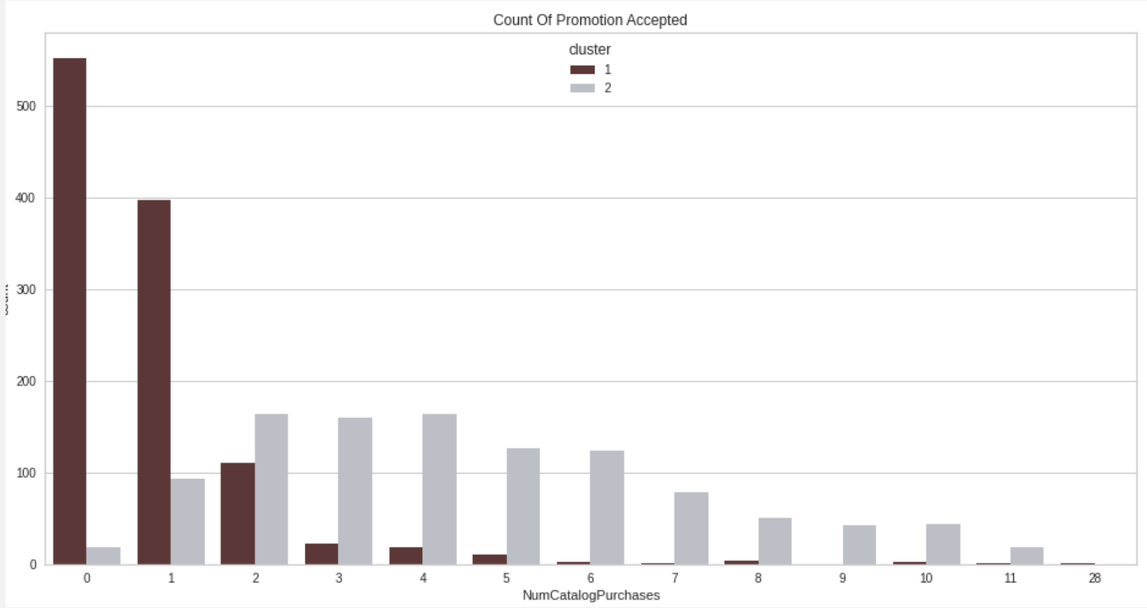


NumWebPurchases

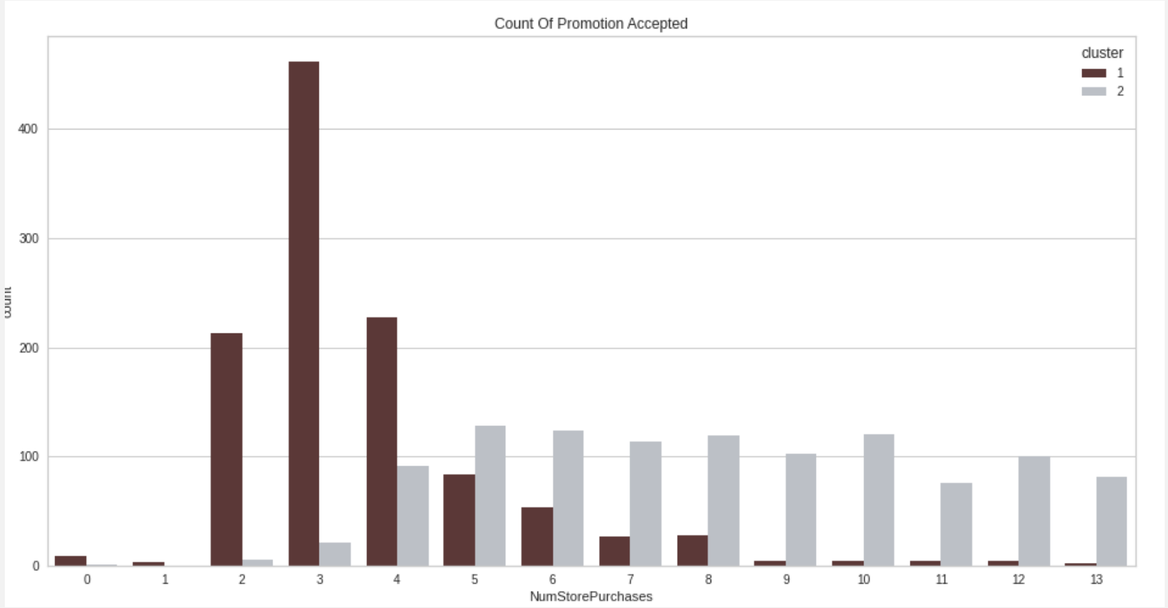
Modeling

k별 군집 특성 ; k=2

5. 구매한 방법에 따른 그룹 차이



NumCatalogPurchases



NumStorePurchases

Modeling

k별 군집 특성 ; k=2

5. 구매한 방법에 따른 그룹 차이

```
group1's NumWebPurchases:  2.7248441674087265
group2's NumWebPurchases:  5.528650646950092 5
group1's NumCatalogPurchases':  0.8334817453250223
group2's NumCatalogPurchases':  4.525878003696857 6
group1's NumStorePurchases:  3.617987533392698 3
group2's NumStorePurchases:  8.112754158964884
group1's NumWebVisitsMonth:  6.658949243098842
group2's NumWebVisitsMonth:  3.964879852125693 4
```

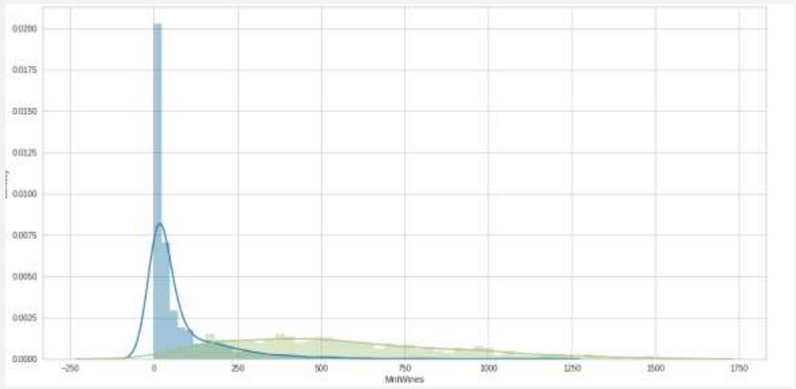
0 < totalmntspent < 800 : 그룹1
0 < totalmntspent < 2500: 그룹2

이라는 결과를 봤을 때 구매한 방법에 따른 구매 횟수도 모두 group2가 높다.
다만 지난 한달동안 웹사이트에 방문한 수는 그룹1이 더 높다는 걸 알 수 있다.

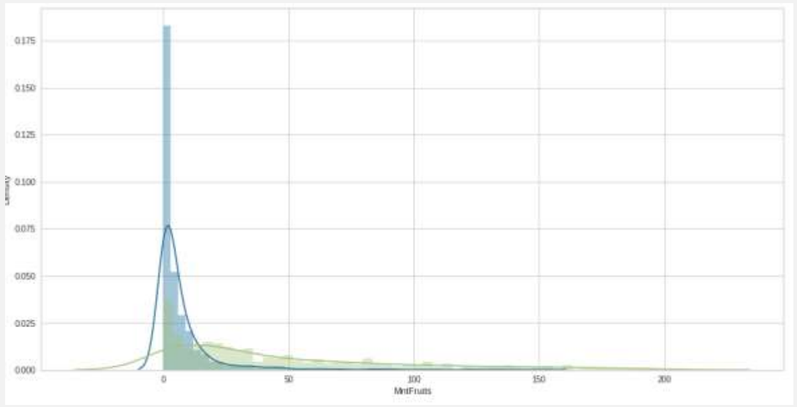
Modeling

k별 군집 특성 ; k=2

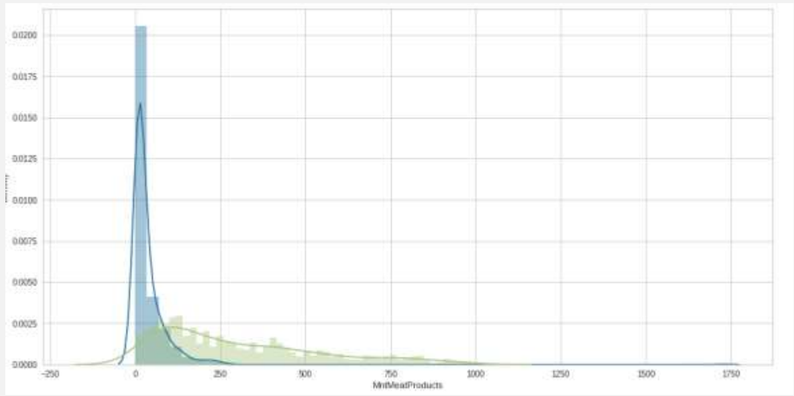
6. 구매한 제품에 따른 그룹 차이



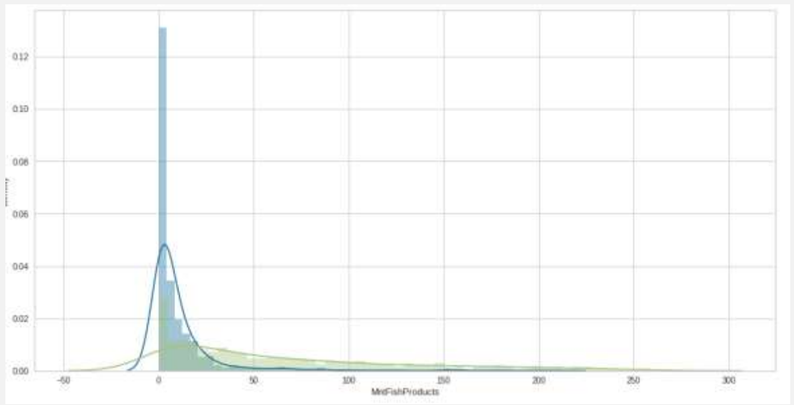
Wine



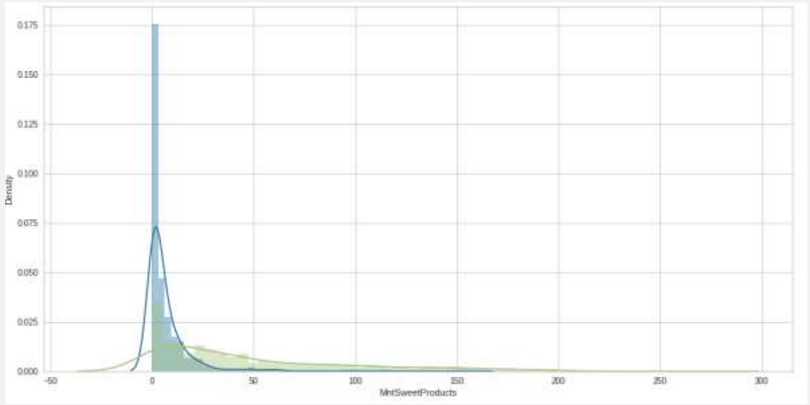
Fruits



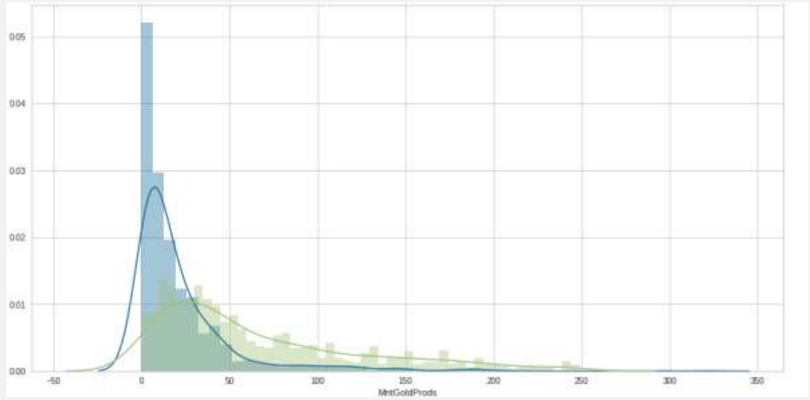
Meat



Fish



Sweet

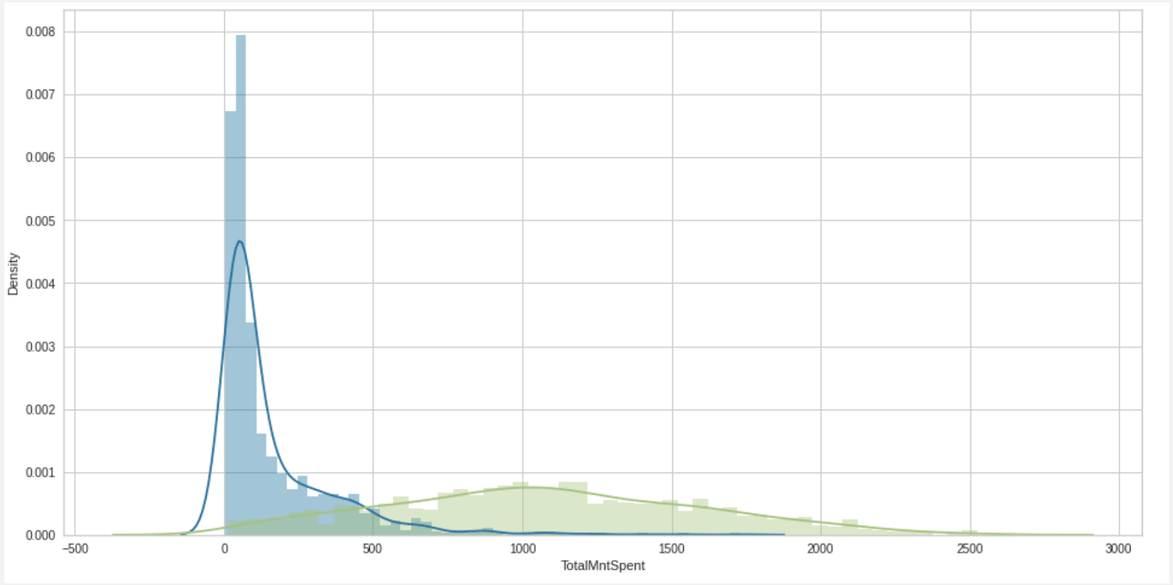


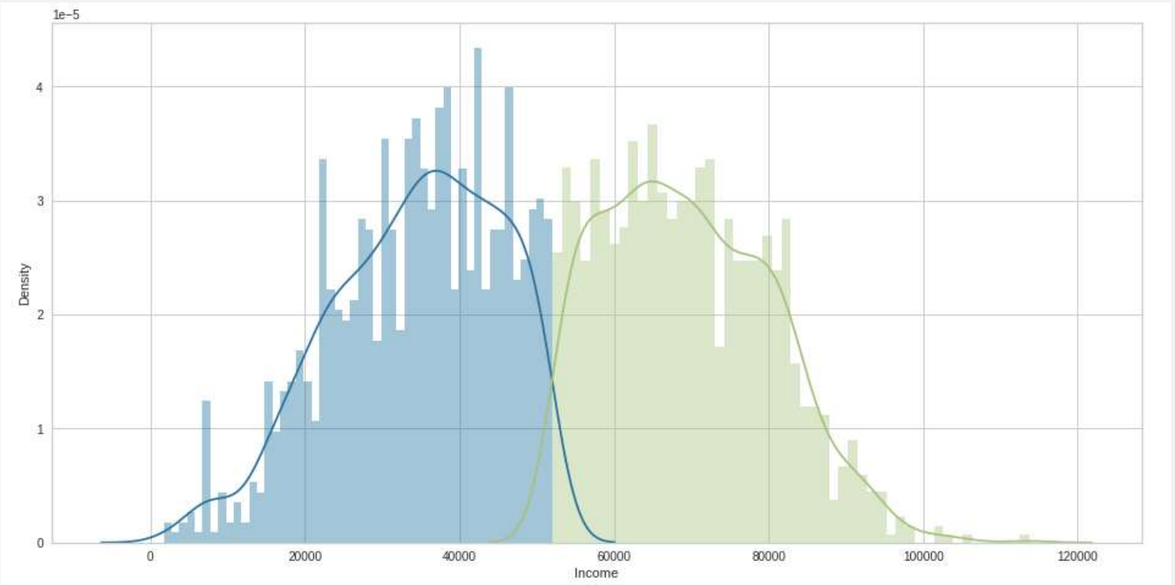
Gold

Modeling

k별 군집 특성 ; k=2

7. 결론





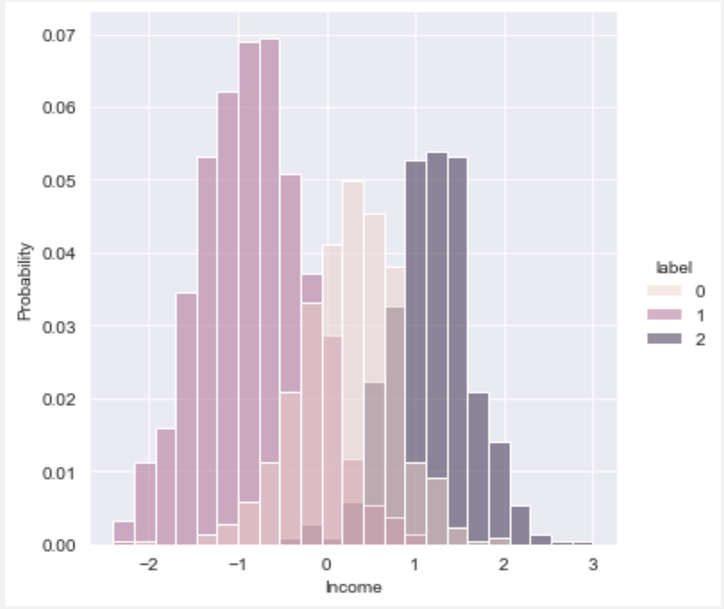
그룹 1과 2를 나누는 가장 큰 기준은 Income과 TotalMntSpent가 되고 사실 Income이 높으니 TotalMntSpent값이 높은건 당연해지고,TotalMntSpent값이 높은 그룹이 더 많은 제품을 구매하고 더 많은 횟수를 구매하기 때문에 실질적인 기준은 Income인 것 같다.
나이, 아이의 여부, 교육수준, 가입날짜는 그룹을 나누는데 영향을 미치지 않았다.

Modeling

k별 군집 특성 ; k=3

1차원 분석 - BY 수익, 총 구매건수, 총 지출액수

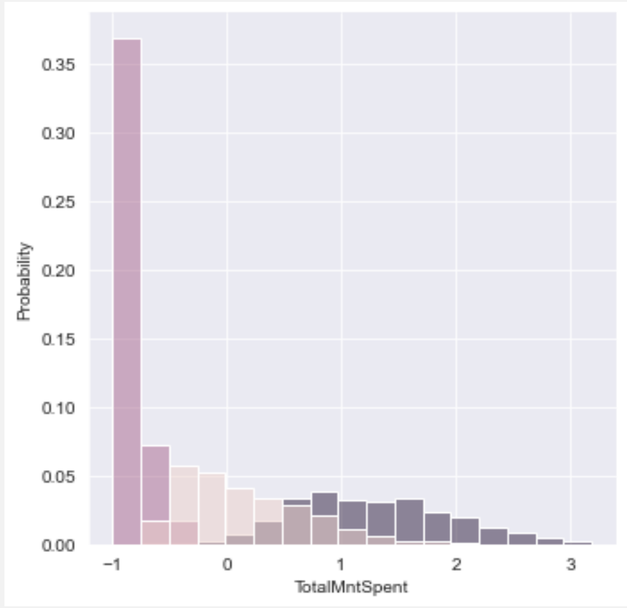
< Income : 총 수입 >



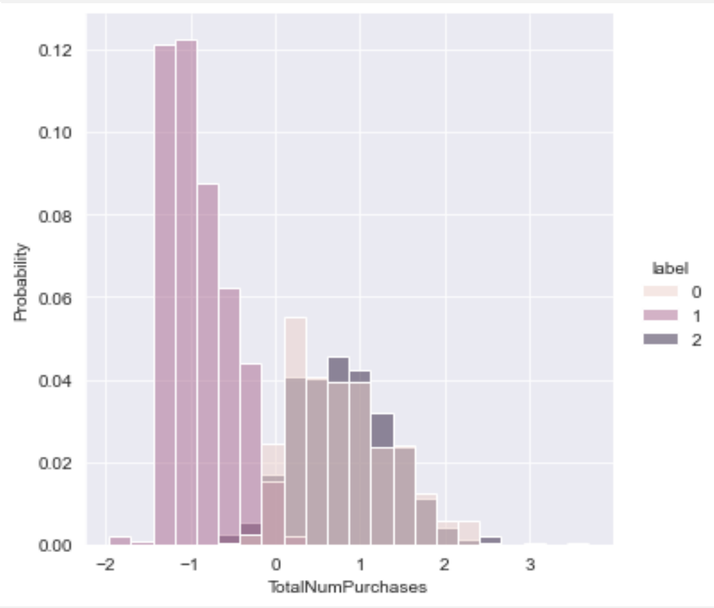
Income을 기준으로,

● 군집 1: 저소득층, ● 군집 0: 중산층, ● 군집 2:고소득층

< TotalMntSpent : 지출총액>



< TotalNumPurchased : 총 구매 건수>



지출총액에선 ● 군집 1 < ● 군집 0 < ● 군집 2 가 확연이 보임

하지만 총 구매건수를 보면 ● 군집 0 과 ● 군집 2 가 겹침.

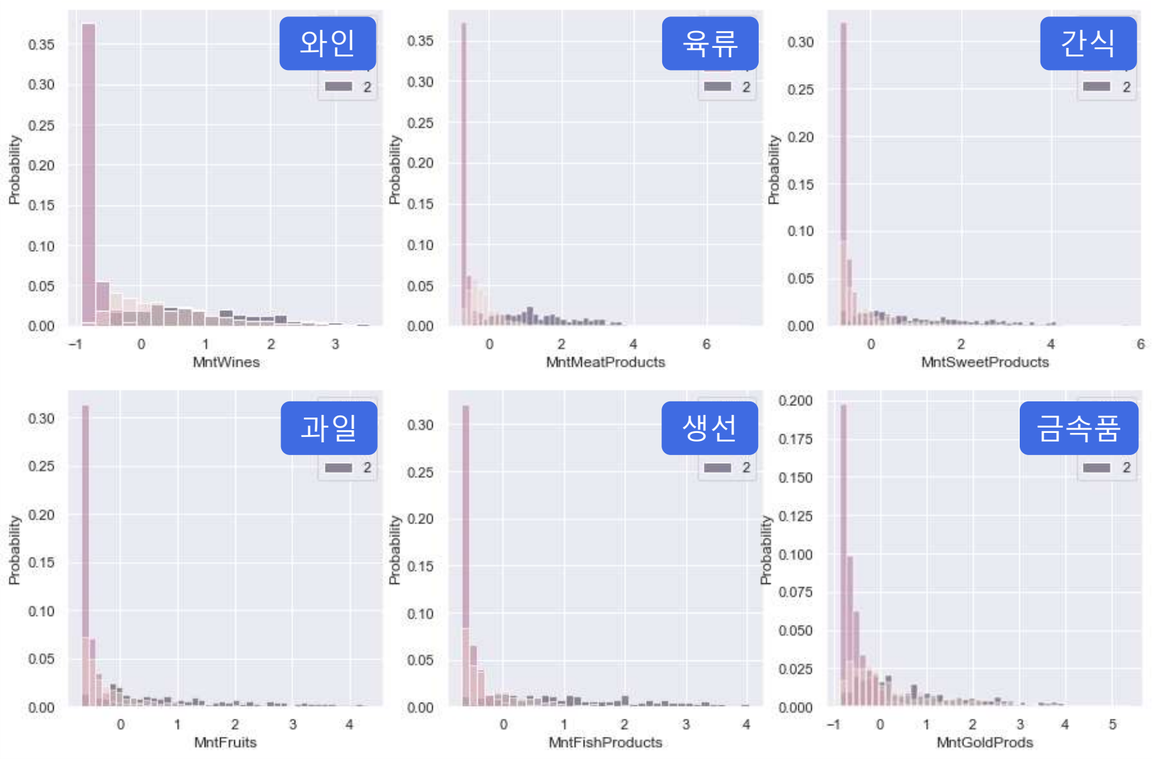
즉, ● 군집 0 과 ● 군집 2 는 구매 횟수는 비슷하지만 ● 군집 2가 한 번 살 때

더 많이, 더 비싼 품목을 구매할 가능성이 높다.

Modeling

k별 군집 특성 ; k=3

1차원 분석 - BY 품목별 구매



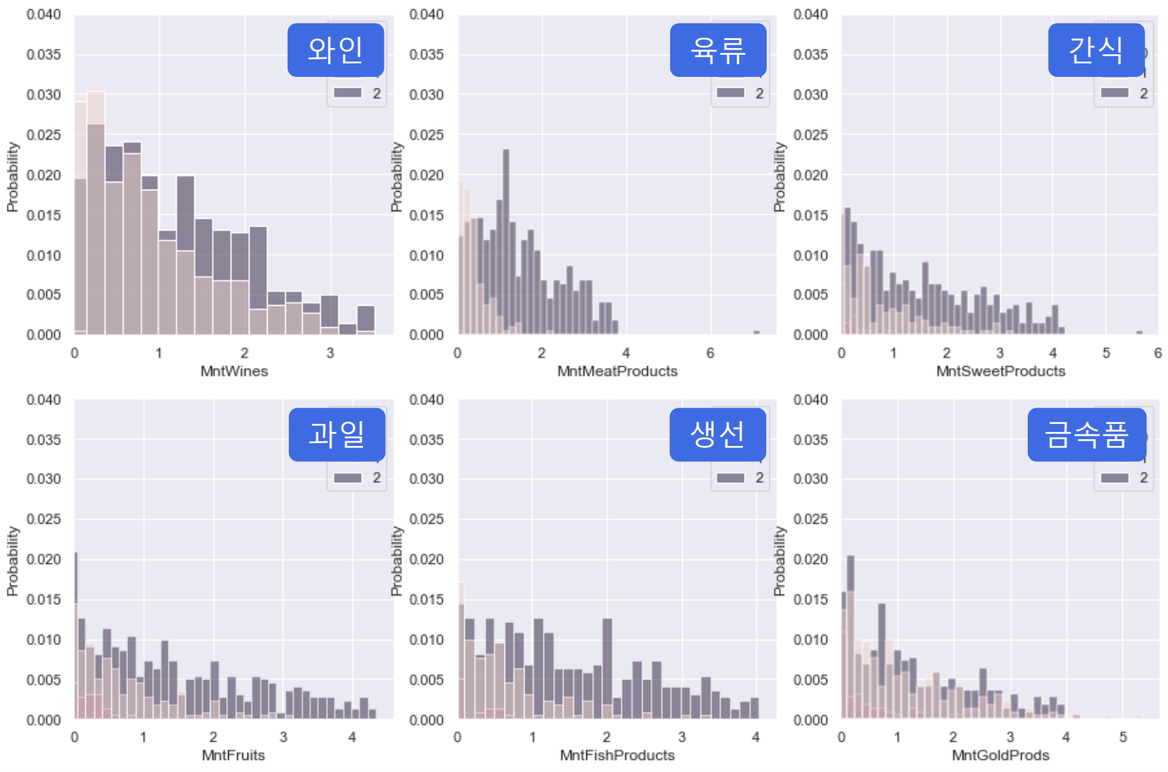
저소득층이었던 ● 군집 1이 확연하게 제품들을 적게 구매함.

● 군집 0 , ● 군집 2 는 비슷하지만 고소득층인 ● 군집 2 가 더 많이 구매

Modeling

k별 군집 특성 ; k=3

1차원 분석 - BY 품목별 구매



< ● 군집 0, ● 군집 2 의 분포를 자세히 보기 위해 확대 >

Meat, Sweet, Fruit, Fish에서

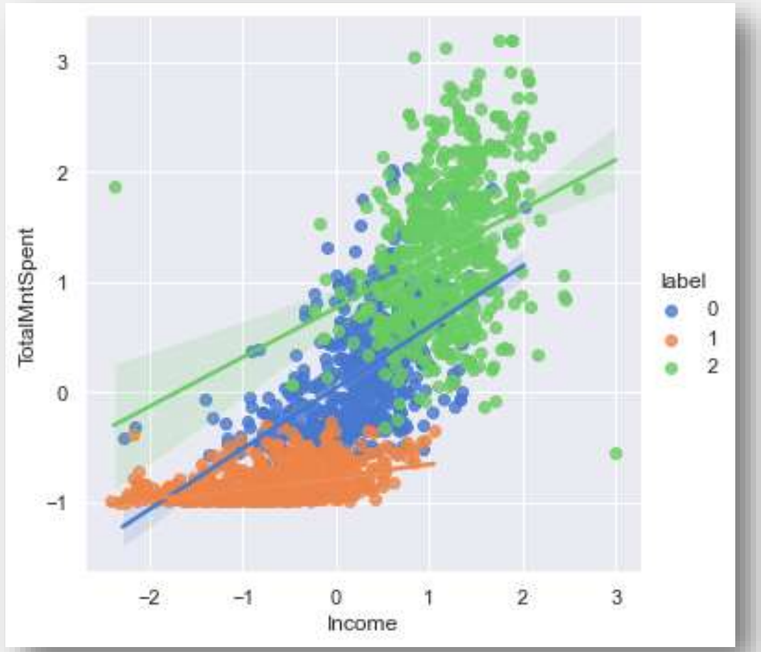
● 군집 0, ● 군집 2 의 구매 차이가 꽤 남

Modeling

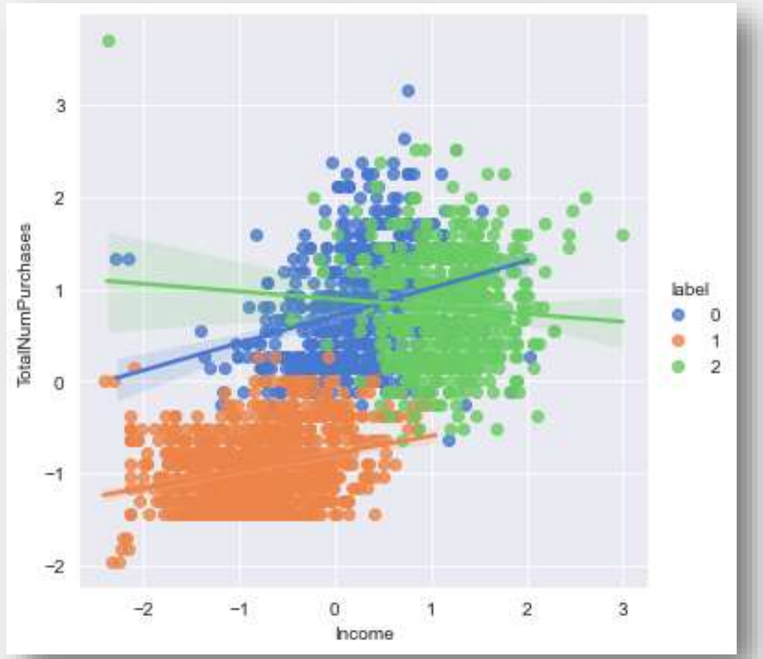
k별 군집 특성 ; k=3

- 상관관계 분석에서 다른 변수들과 가장 **높은 상관계수**를 가지는 변수가 Income, 그 다음으로 TotalMntSpent임
- Income과 TotalMntSpent와의 **산점도 분석** 결과 가장 군집이 뚜렷하게 나타남
- Income이 증가할수록 TotalMntSpent 역시 증가하는 경향을 보임

2차원 분석



⇒ 군집화가 잘 나타나는 변수 사용



Modeling

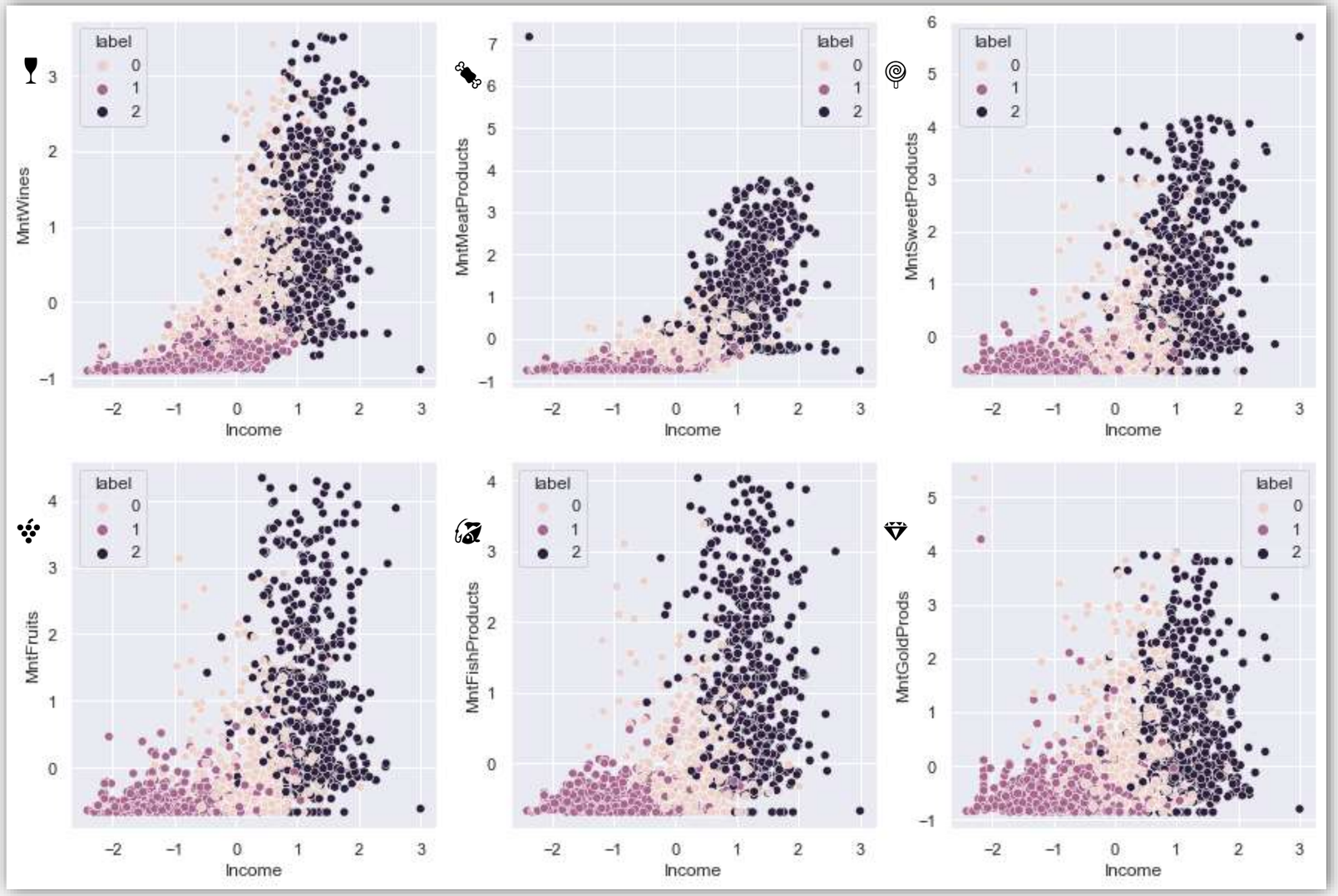
k별 군집 특성 ; k=3

2차원 분석 - BY 품목별 구매

Income 별 각 품목 지출액

- 품목 1: 와인
- 품목 2: 육류
- 품목 3: 당류
- 품목 4: 과일
- 품목 5: 어류
- 품목 6: 금

- ● 군집 1은 저소득 / 모든 품목에서 가장 적게 구매
- ● 군집 0은 중간층 / 모든 품목에서 중간 수준으로 구매
- ● 군집 2는 고소득 / 모든 품목에서 가장 많이 구매



Modeling

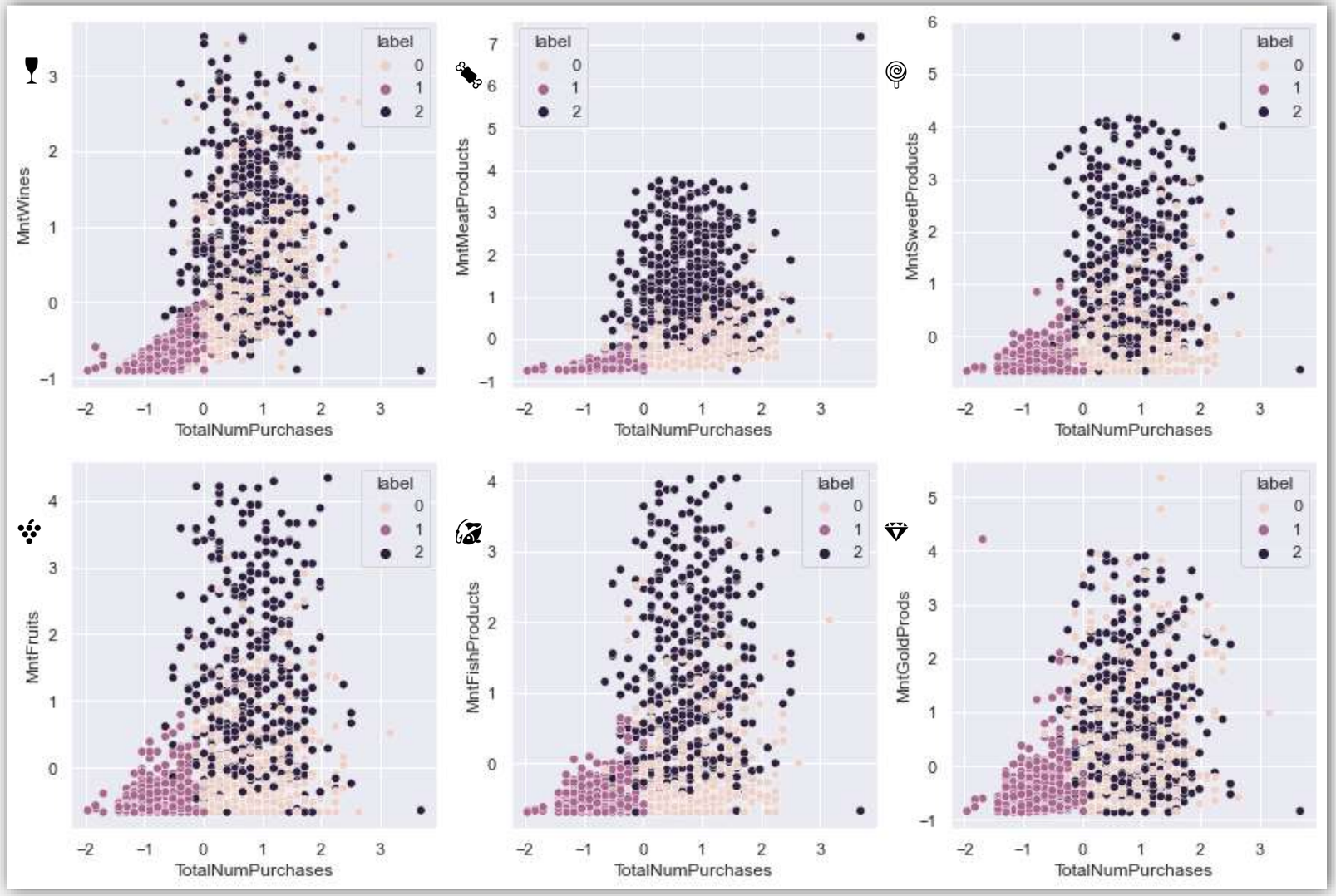
k별 군집 특성 ; k=3

2차원 분석 - BY 품목별 구매

총 구매건수 별 각 품목 지출액

- 품목 1: 와인
- 품목 2: 육류
- 품목 3: 당류
- 품목 4: 과일
- 품목 5: 어류
- 품목 6: 금속류

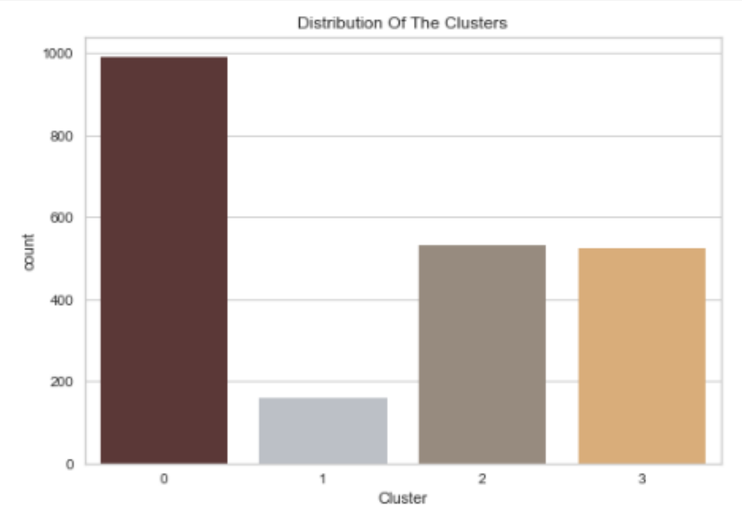
- ● 군집 1은 저소득 / 모든 품목에서 가장 적게 구매
- 와인, 금 품목에서는 ● 군집 0와 ● 군집 2의 차이가 크지 않음
- 육류, 당류, 과일, 어류 품목에서는 ● 군집 2가 ● 군집 0보다 더 많이 구매했으나 크게 뚜렷하지 않음



Modeling

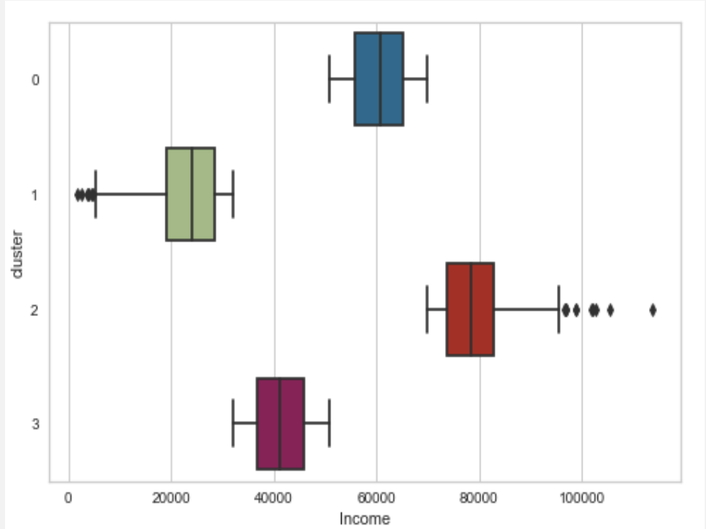
k별 군집 특성 ; k=4

< 모델 분포 >



- 군집1의 경우 적게 분포해있음

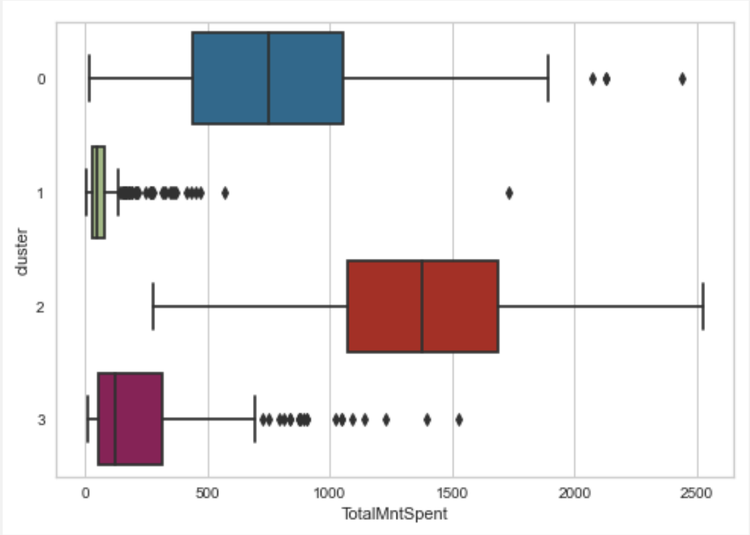
< Income : 총 수입 >



Income을 기준으로,

● 군집 1 : 저소득층, ● 군집 0: 중산층,
● 군집 3: 중산층, ● 군집 2 :고소득층

< TotalMntSpent : 지출총액>



지출총액의 경우 ● 군집 1 < ● 군집 3 < ● 군집 0 < ● 군집 2

즉, 수입이 많을수록 지출총액이 증가함을 알 수 있음

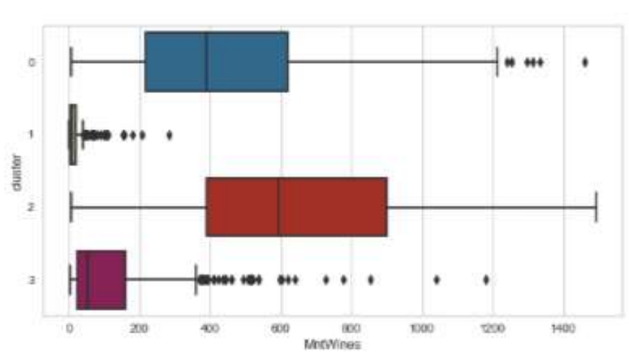
(∵지출총액은 수입에 비례)

Modeling

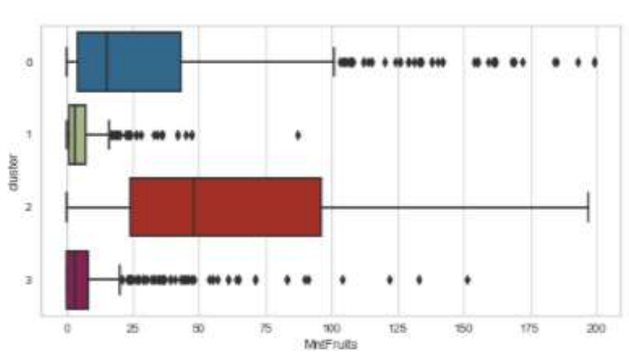
k별 군집 특성 ; k=4

■ 군집 별 특성 확인 : 품목 별 지출비용

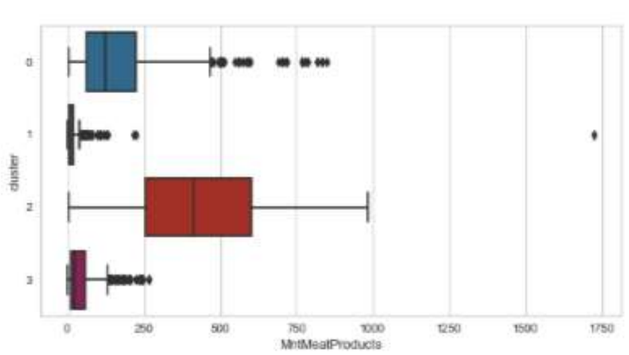
< MntWines 🍷 >



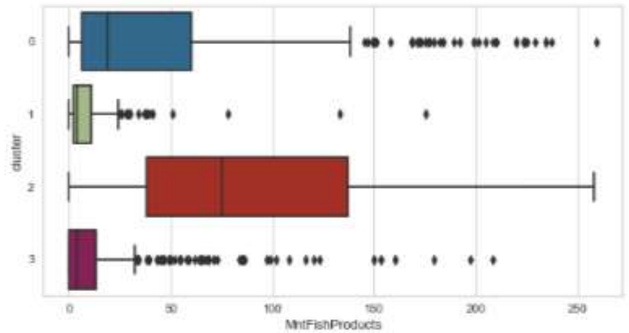
< MntFruits 🍇 >



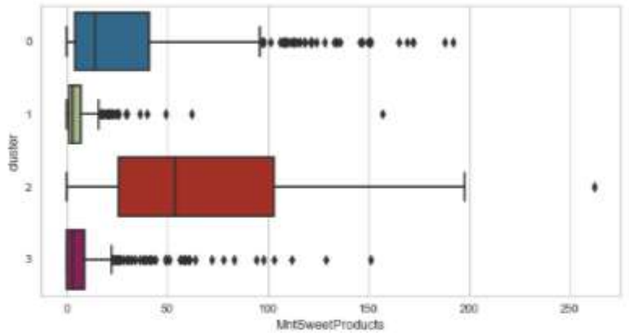
< MntMeatProducts🍖>



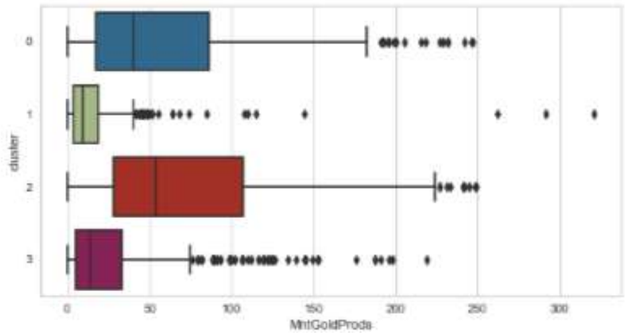
< MntFishProducts 🐠>



< MntSweetProducts🍭>



< MntGoldProds 💎>



지출 비용의 경우

● 군집 1 , ● 군집 3 < ● 군집 0 , ● 군집 2

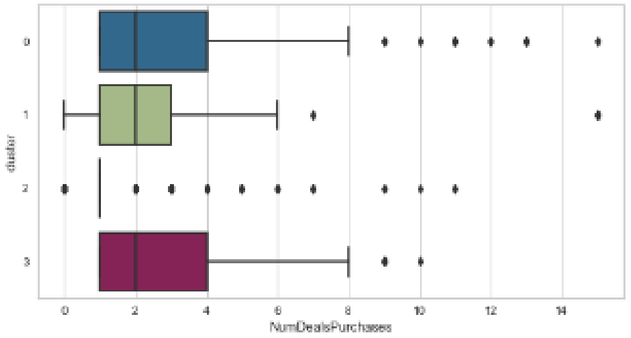
즉, 수입이 많을수록 각 품목 별 지출 비용도 증가함

Modeling

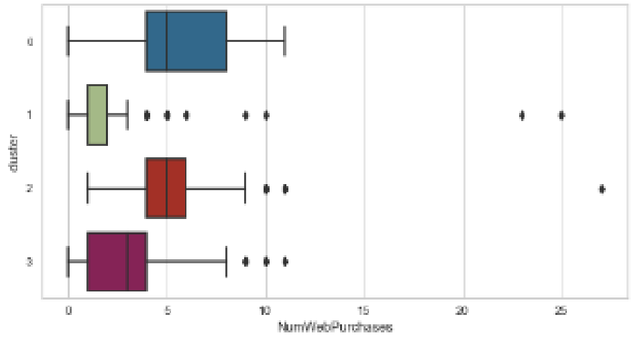
k별 군집 특성 ; k=4

■ 군집 별 특성 확인 : 구매 횟수

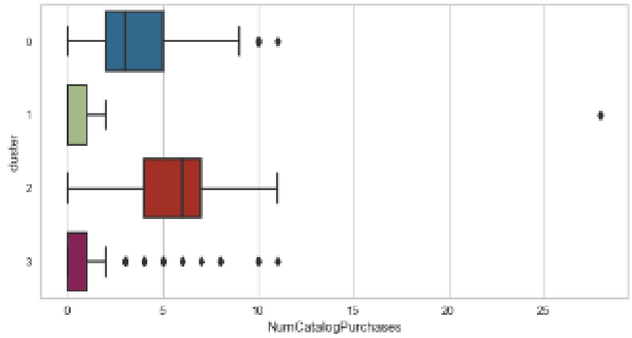
< NumDealsPurchases >

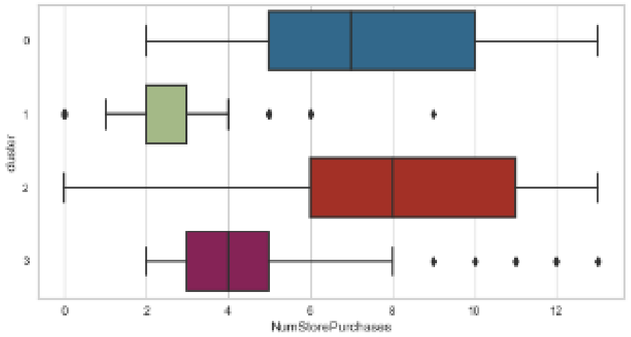


< NumWebPurchases >

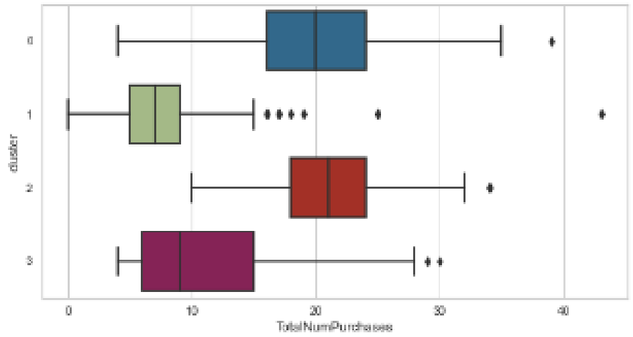


< NumCatalogPurchases >





< NumStorePurchases >



< TotalNumPurchases >

구매 횟수의 경우

● 군집 1, ● 군집 3 < ● 군집 0, ● 군집 2

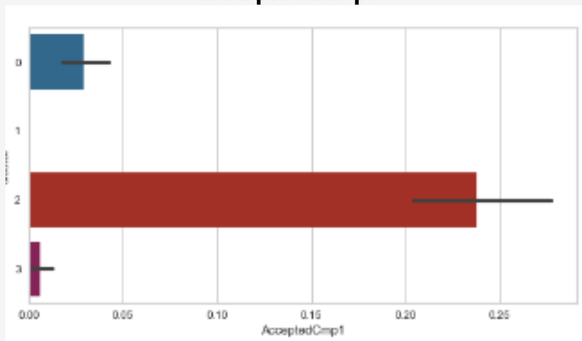
즉, 수입이 많을수록 구매 횟수도 증가함

Modeling

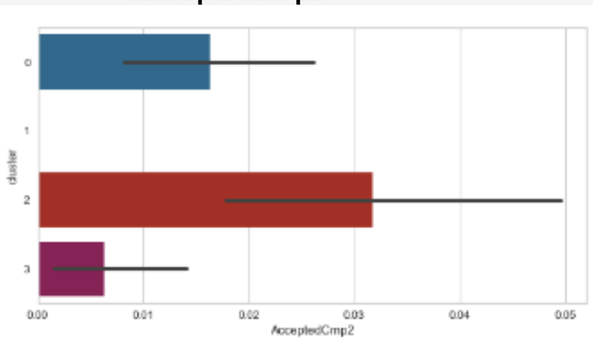
k별 군집 특성 ; k=4

■ 군집 별 특성 확인 : 캠페인 수락 여부

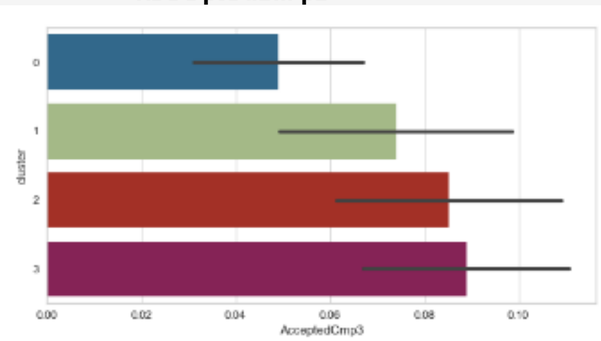
< AcceptedCmp1 >



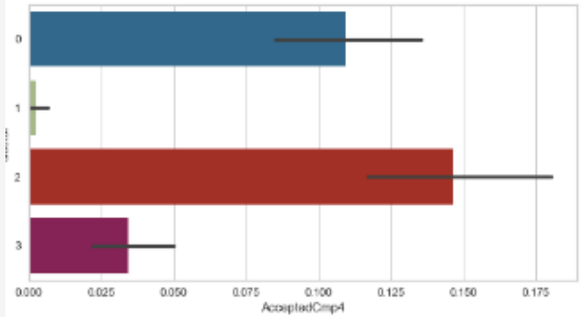
< AcceptedCmp2 >



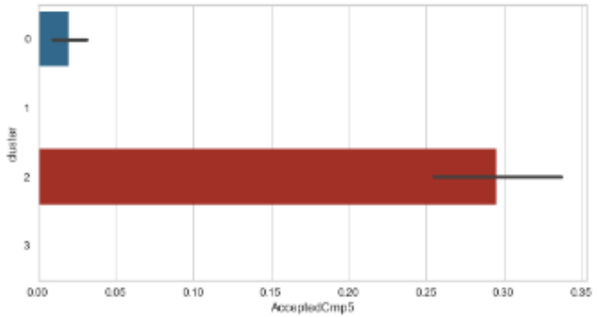
< AcceptedCmp3 >



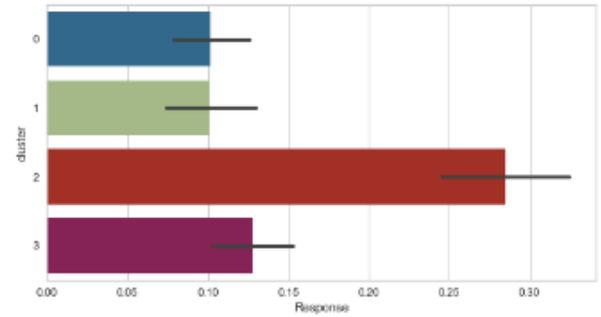
< AcceptedCmp4 >



< AcceptedCmp5 >



< AcceptedCmp6 >



캠페인 수락 여부의 경우

● 군집 1, ● 군집 3 < ● 군집 0, ● 군집 2

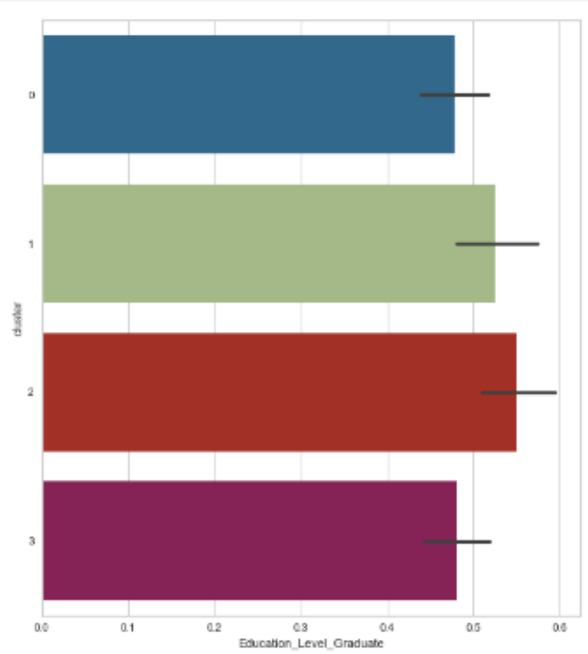
즉, 수입이 많을수록 캠페인을 수락할 확률이 높음

Modeling

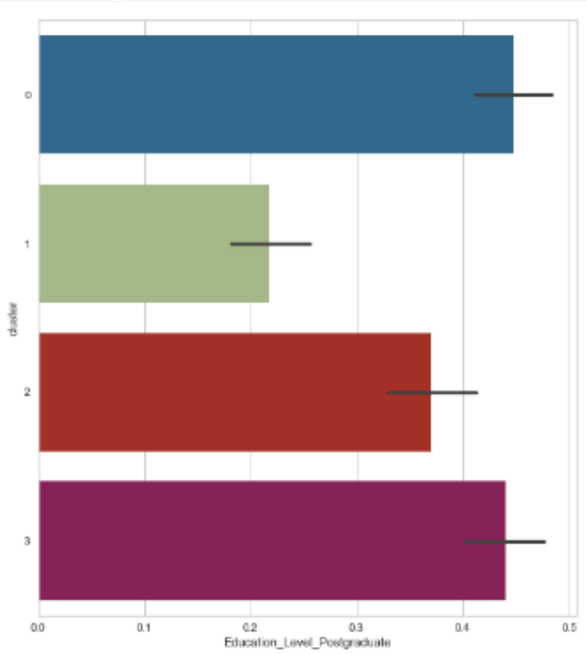
k별 군집 특성 ; k=4

■ 군집 별 특성 확인 : 교육

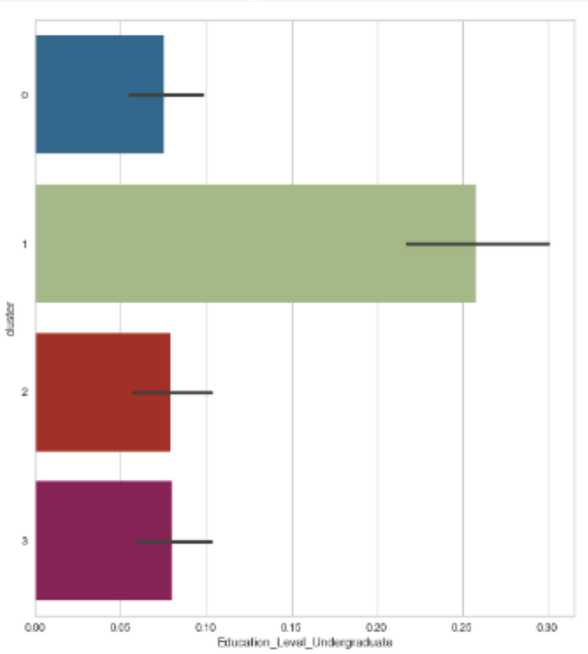
< Graduate >



< Postgraduate: Master, PhD >



< Undergraduate >



교육 수준의 경우

● 군집 1 에서 Undergraduate의 비율이 높음

● 군집 0 , ● 군집 2 에서 Postgraduate의 비율이 높음

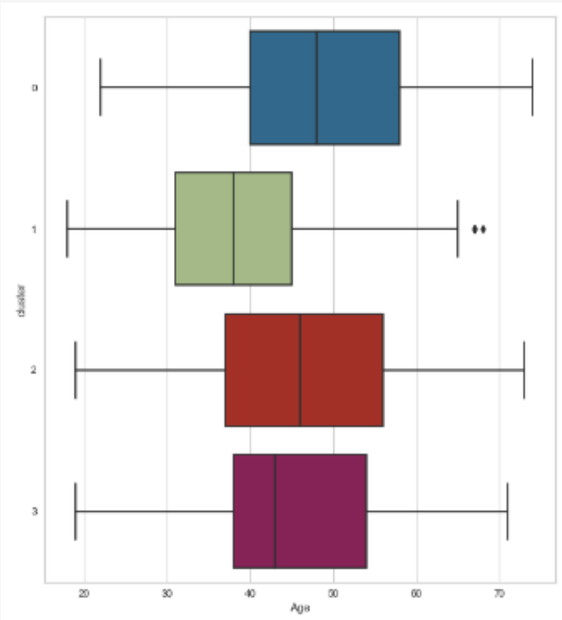
즉, 수입이 많을수록 높은 교육수준을 갖고 있음

Modeling

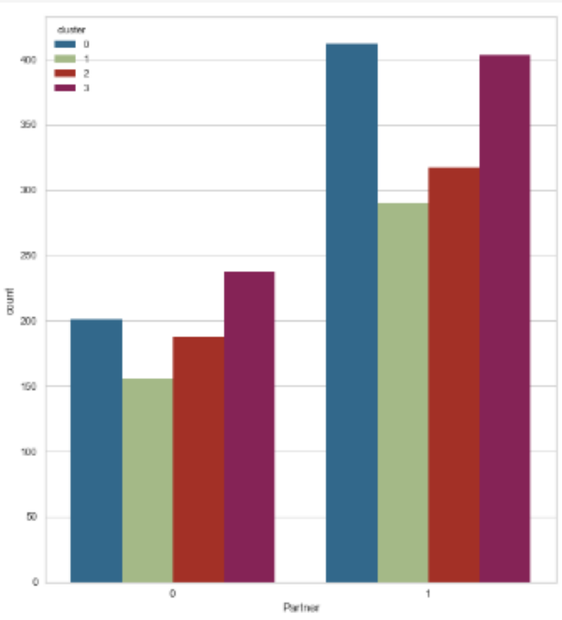
k별 군집 특성 ; k=4

■ 군집 별 특성 확인 : Age, Partner, Children

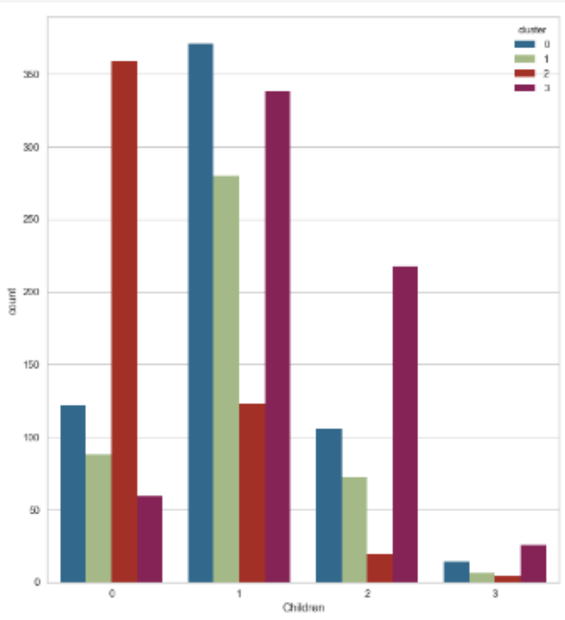
< Age: 고객 나이 >



< Partner: 파트너 유무>



< Children: 자녀 수>



고객 나이와 파트너 유무의 경우

군집 분류에 큰 영향을 미치지 않음

자녀 수의 경우

● 군집 1 , ● 군집 3 은 대부분 자녀가 1명 이상

● 군집 0 , ● 군집 2 는 대부분 자녀가 없거나 1명

Modeling

k별 군집 특성 ; k=4

■ 변수 간 관계

- 각 군집 별 변수들의 평균, 표준편차, 최솟값, 최댓값, 분위수 등을 확인해봄

```
cluster1.describe()
```

	Income	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldP
count	613.000000	613.000000	613.000000	613.000000	613.000000	613.000000	613.000000	613.00(
mean	60455.445351	48.924959	440.869494	30.993475	169.164763	41.324633	28.778140	58.19
std	5428.303443	28.142046	290.329815	40.185680	156.448024	52.706060	36.170751	53.53
min	50870.000000	0.000000	5.000000	0.000000	3.000000	0.000000	0.000000	0.00(
25%	55686.000000	25.000000	215.000000	4.000000	61.000000	6.000000	4.000000	17.00(
50%	60631.000000	50.000000	389.000000	15.000000	124.000000	19.000000	14.000000	40.00(
75%	65196.000000	73.000000	620.000000	43.000000	224.000000	60.000000	41.000000	86.00(
max	69805.000000	99.000000	1459.000000	199.000000	849.000000	259.000000	192.000000	247.00(

```
cluster3.describe()
```

● 군집 2 : 고소득층

	Income	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldI
count	505.000000	505.000000	505.000000	505.000000	505.000000	505.000000	505.000000	505.0(
mean	79243.243564	49.746535	647.635644	62.796040	441.481188	91.021782	68.221782	74.5(
std	6953.608252	29.335483	329.012951	49.079948	234.630705	65.797862	53.283145	61.7
min	69867.000000	0.000000	6.000000	0.000000	3.000000	0.000000	0.000000	0.0(
25%	73691.000000	25.000000	390.000000	24.000000	254.000000	38.000000	26.000000	28.0(
50%	78416.000000	52.000000	594.000000	48.000000	414.000000	75.000000	54.000000	54.0(
75%	82800.000000	74.000000	899.000000	96.000000	601.000000	137.000000	103.000000	107.0(
max	113734.000000	99.000000	1493.000000	197.000000	984.000000	258.000000	262.000000	249.0(

```
cluster2.describe()
```

● 군집 1 : 저소득층

	Income	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldPr
count	446.000000	446.000000	446.000000	446.000000	446.000000	446.000000	446.000000	446.000(
mean	23097.224215	48.959641	16.674888	5.508969	20.899103	8.051570	5.778027	16.262
std	6738.097812	29.011244	27.521508	8.356047	84.000060	13.118288	10.042049	28.091
min	1730.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000(
25%	19017.750000	24.000000	4.000000	1.000000	6.000000	2.000000	1.000000	4.000(
50%	24184.500000	49.000000	8.000000	3.000000	11.000000	4.000000	3.000000	9.500(
75%	28436.750000	75.750000	18.000000	7.000000	19.000000	11.000000	7.000000	19.000(
max	32146.000000	99.000000	284.000000	87.000000	1725.000000	175.000000	157.000000	321.000(

```
cluster4.describe()
```

	Income	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldP
count	641.000000	641.000000	641.000000	641.000000	641.000000	641.000000	641.000000	641.00(
mean	41261.046802	48.542902	109.745710	7.879875	44.533541	13.048362	8.031201	25.89
std	5326.707200	29.355714	139.446067	15.745027	48.309444	26.235993	15.838313	33.71(
min	32173.000000	0.000000	2.000000	0.000000	1.000000	0.000000	0.000000	0.00(
25%	36778.000000	24.000000	22.000000	0.000000	11.000000	0.000000	0.000000	5.00(
50%	41039.000000	49.000000	53.000000	3.000000	25.000000	4.000000	3.000000	14.00(
75%	45989.000000	73.000000	158.000000	8.000000	60.000000	13.000000	9.000000	33.00(
max	50785.000000	99.000000	1181.000000	151.000000	267.000000	208.000000	151.000000	219.00(

Modeling

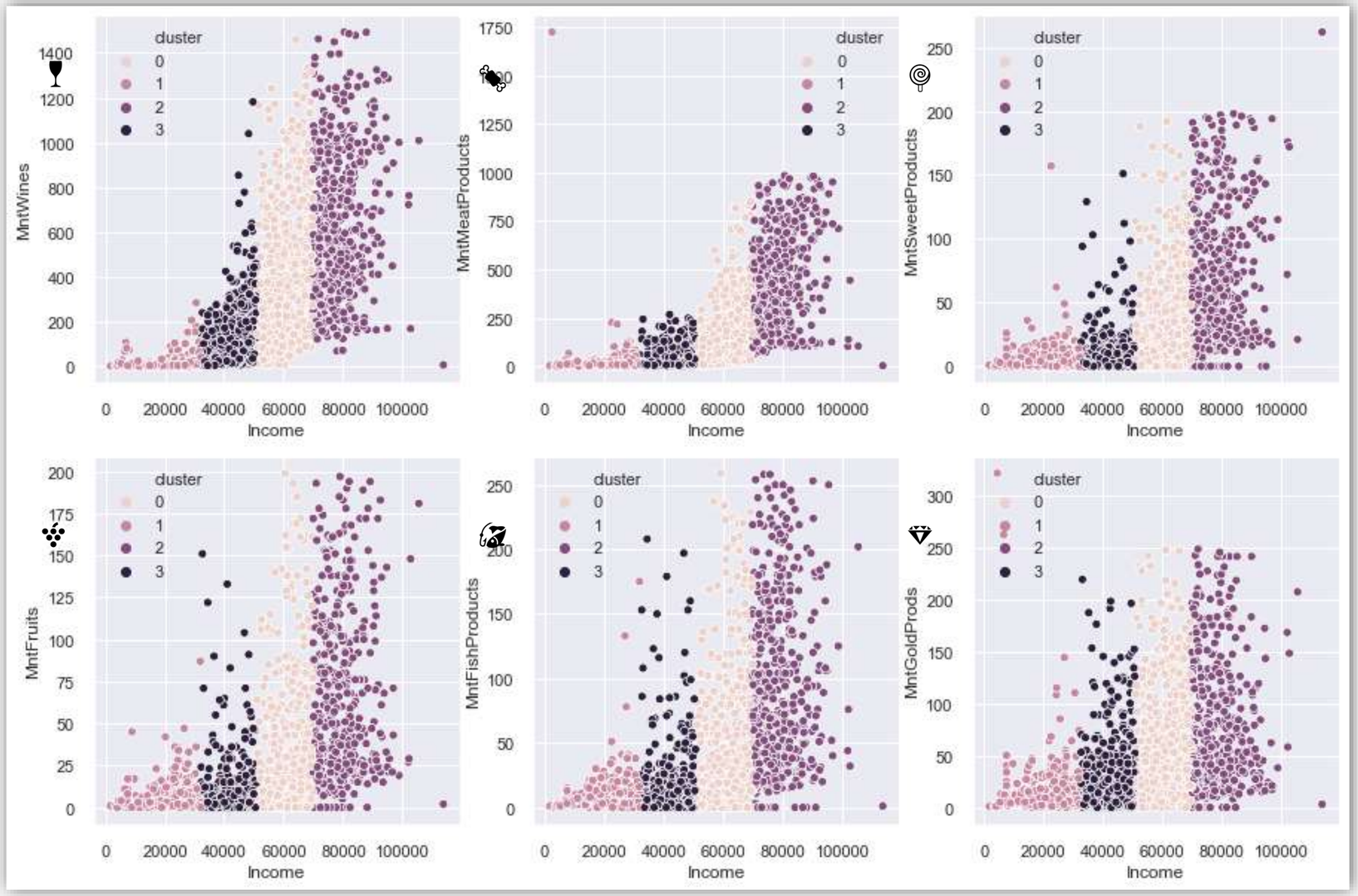
k별 군집 특성 ; k=4

■ 2차원분석

Income 별 각 품목 지출액

- 품목 1; 와인
- 품목 2; 육류
- 품목 3; 당류
- 품목 4; 과일
- 품목 5; 어류
- 품목 6; 금

- ● 군집 1은 저소득 / 모든 품목에서 가장 적게 구매
- ● 군집 0은 중간층 / 모든 품목에서 중간 수준으로 구매
- ● 군집 2는 고소득 / 모든 품목에서 가장 많이 구매
- ● 군집 3은 중간층 / 모든 품목에서 중간 수준으로 구매



Modeling

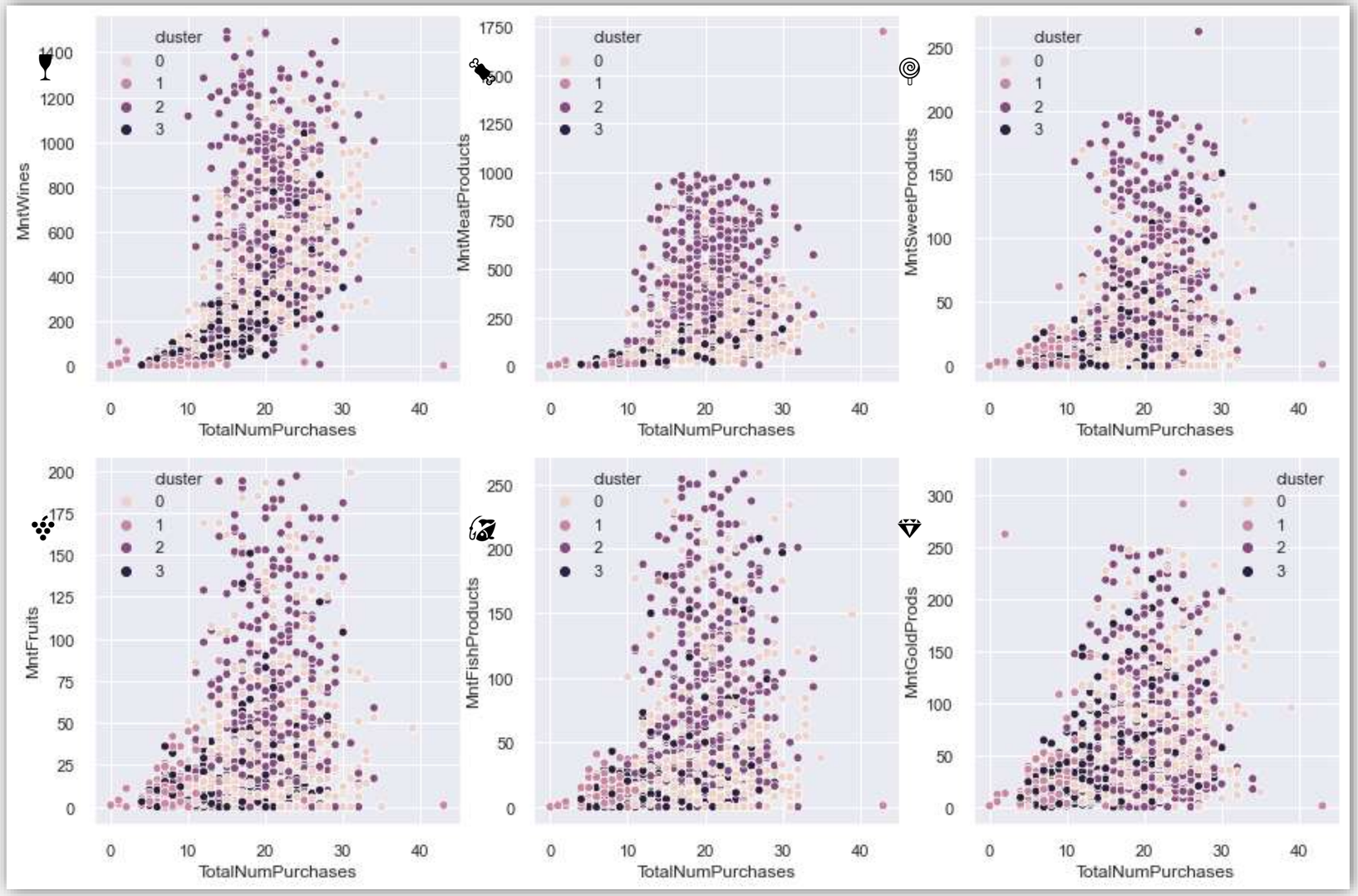
k별 군집 특성 ; k=4

■ 2차원분석

총 구매건수 별 각 품목 지출액

- 품목 1: 와인
- 품목 2: 육류
- 품목 3: 당류
- 품목 4: 과일
- 품목 5: 어류
- 품목 6: 금

- ● 군집 1은 저소득 / 모든 품목에서 가장 적게 구매
- 모든 품목에서는 ● 군집 0와 ● 군집 2 의 차이가 크지 않음
- 와인, 육류, 당류 품목에서는 ● 군집 1과 ● 군집 3의 지출액이 다른 그룹보다 적음



THANK YOU

https://github.com/jyoung19/Customer_Personality_Analysis