Consumer Behavior Analysis

5230 Midterm Presentation

Serena(Jiahui) Zeng. zeng.jiahu@northeastern.edu
Bingyang Ke. ke.bi@northeastern.edu
Xingxuan Xi xi.xing@northeastern.edu
Le Ju ju.l@northeastern.edu

Why?

- Analyze consumer behavior
- Explore relationship between different features(consumer profile, buying habits, etc)
- Help identify/cluster groups of consumer to gain profits
- Help future business entrepreneurs to identify target customers and business strategy

Methods

1. Data cleaning
2. EDA-Feature selections
3. Models to explore relationships and cluster customers
    - K_means
    - k_means with scaled data
    - k_means with bucket data
    - gaussian mixture model
    - apriori models

Dataset Description

1. Data Source: Kaggle Comsumer Personality Analysis
   Link:
   https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

2. Data Description:
   2240 rows x 29 columns, Group (profile, MNT(product comsuption), Promotion, Purchase places

Data cleaning

- Handling missing values in 'Income' column by filling them with the median value
- Converting 'Dt_Customer' to datetime format
- Replacing non-standard responses ('Absurd', 'YOLO') in 'Marital_Status' with 'Other'
- Capping outliers at the 1st and 99th percentiles

Datasets we had/created for modeling:

1. df: raw data (29 columns)
2. df_cleaned: data after cleaning (27 columns)
3. df_featured: after feature concatenation (11 columns)
4. df_combined: after encoding df_featured
5. df_encoded: after grouping and encoding df_cleaned
6. df_scaled: after scaling df_combined

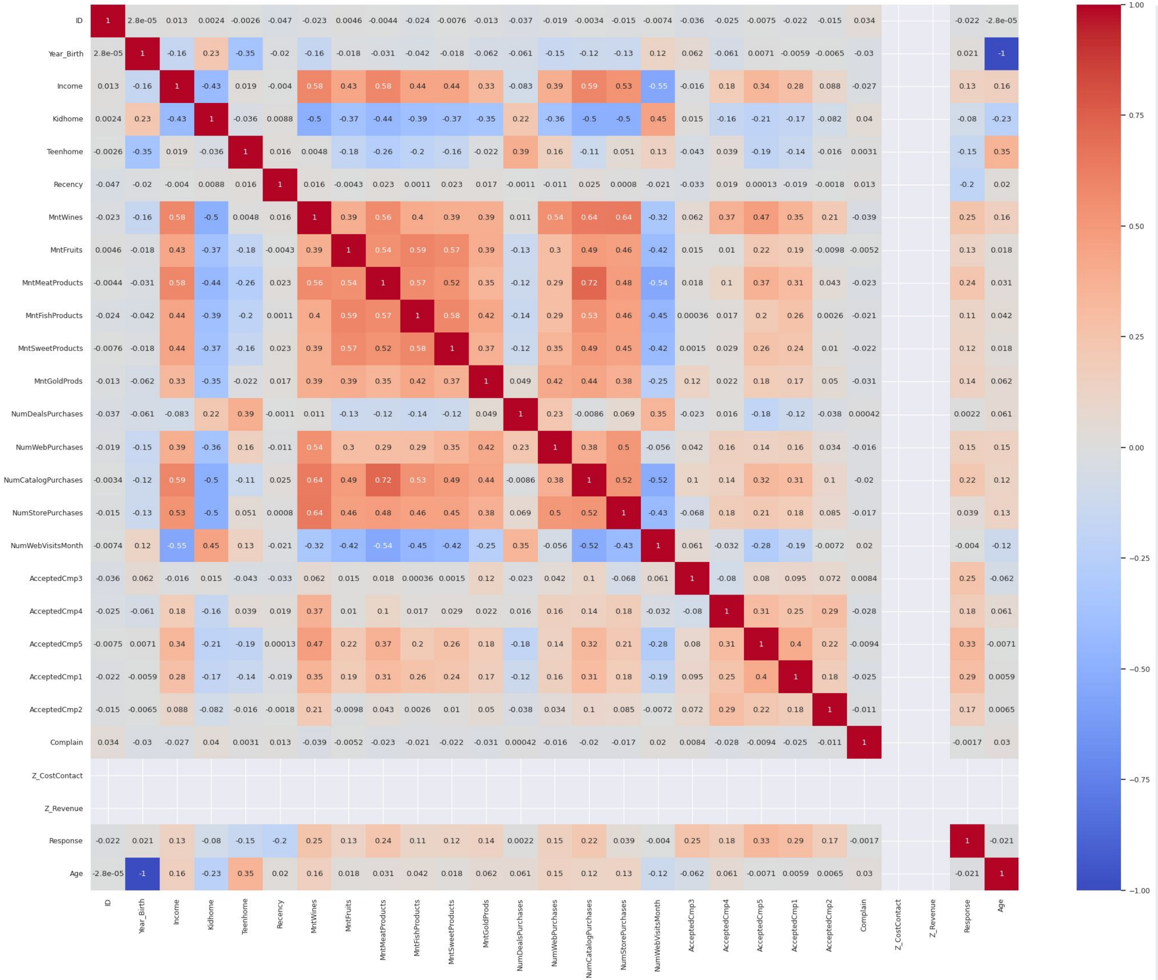
Graphs.EDA

Sub findings :

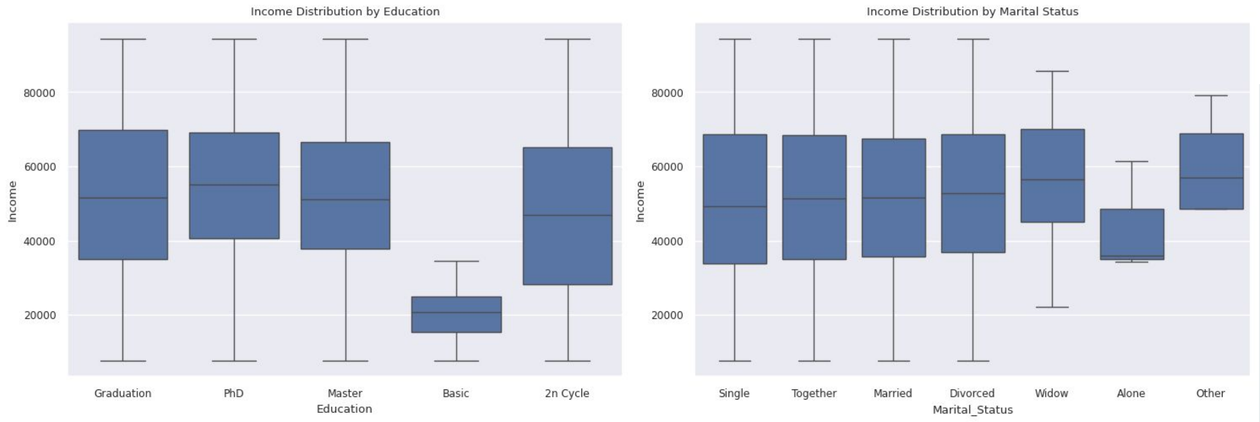
products correlation ----
1. income-income_bins_frequency
2. number of wine --number of catalog/store purchases. (0.64)
3. number of fish --number of fruits. (0.59) number of sweet products(0.58)
4. number of sweet product --number of fish, number of fruit.
5. number of gold product -- number of catalog purchases(0.44). number of web purchases, fish products.

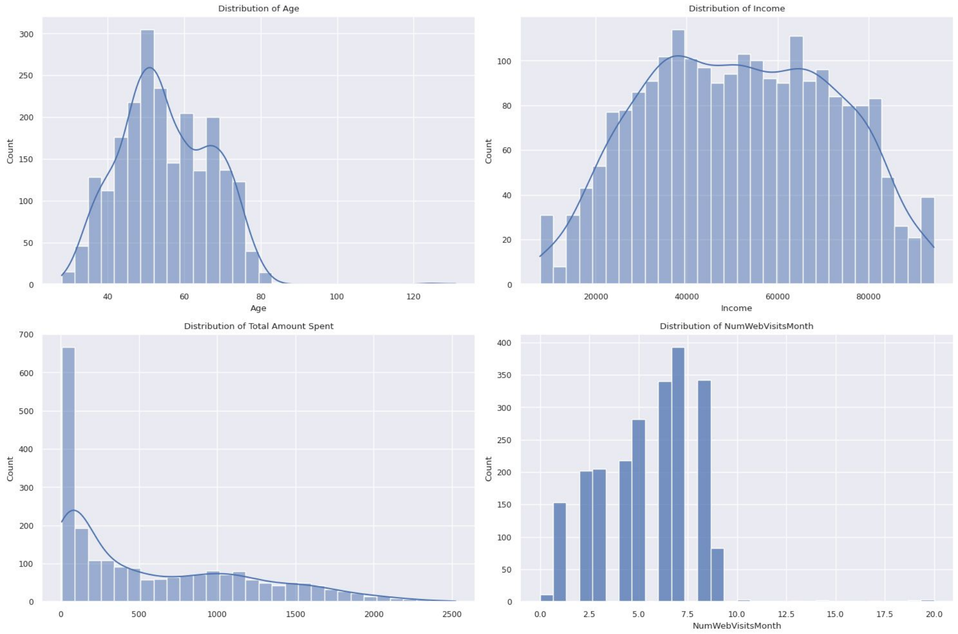
Types of purchases
1. number of web purchase---mintwines(0.54), number of store purchases(0.5).
2. number of catalog purchases --mintmeat(0.72), mintwints(0.64). income(0.5)
3. number of store purchases -- mintwines(0.64), income(0.53), number of catalog purchases(0.52)

4. number of web visits per month --- negative correlated to meat(-0.54), number of fish(-0.44), income(-0.55), general mint(purchases are low)
highly correlated with kidshome(0.45), a little correlated with number of deals purchased.



Graphs.EDA





- Higher education, especially those who have PhD degree, correlates with higher income
- Being married, divorced, or widow may be associated with higher income compared to single or alone status.

Graphs.EDA

	Income	Income_Bucket
0	58138.0	Middle Income
1	46344.0	Middle Income
2	71613.0	High Income
3	26646.0	Low Income
4	58293.0	Middle Income
...	...	...
2235	61223.0	Middle Income
2236	64014.0	Middle Income
2237	56981.0	Middle Income
2238	69245.0	Middle Income
2239	52869.0	Middle Income

income_bins = [0, 30000, 70000, 1000000]

income_labels = ['Low Income', 'Middle Income', 'High Income']

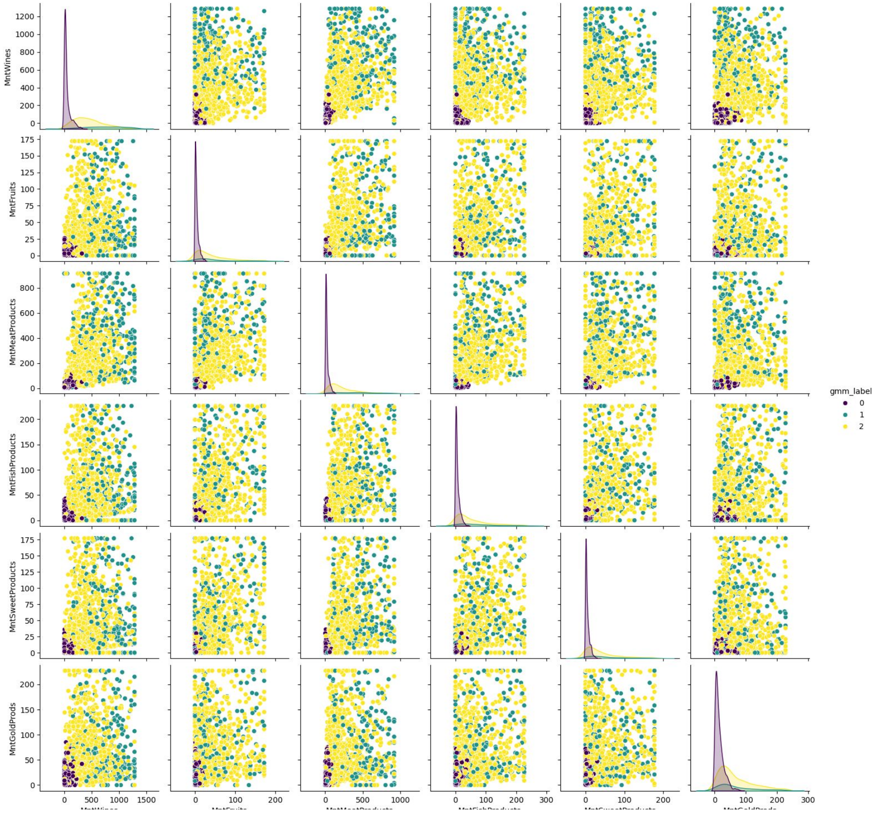
Findings/questions to explore with models

Based on EDA, we decide to explore more into the questions with different models

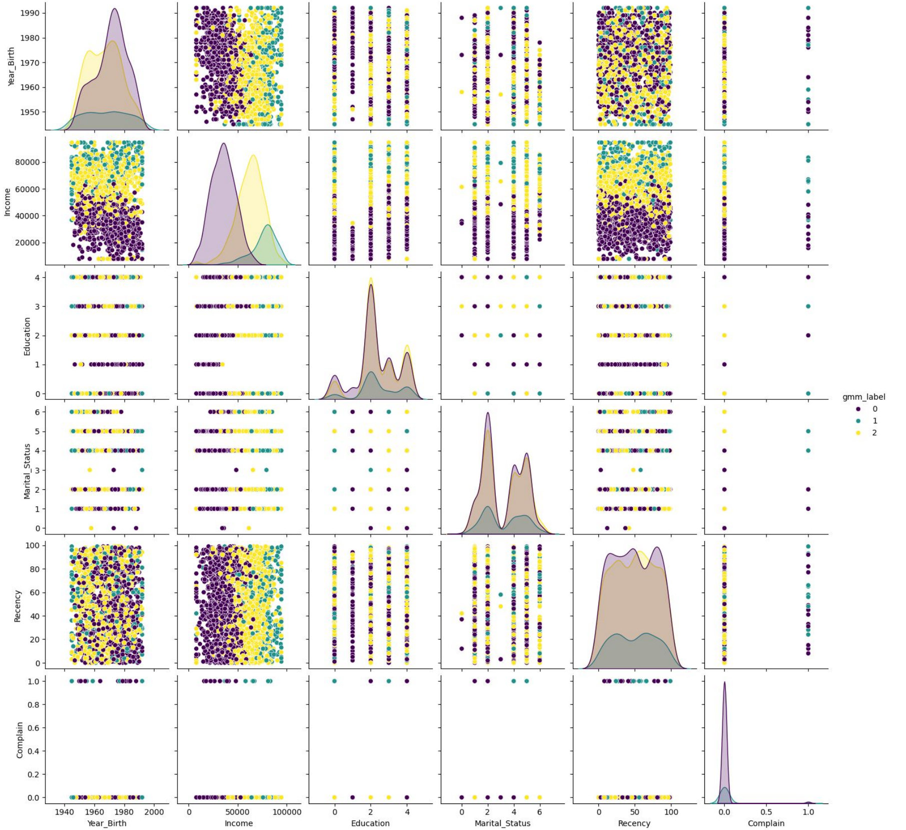
1, gaussian mixture models– explore —clean data, categorize customers based on purchase behaviors and promotions types
2, k_means -df_clean, df combine
3, bucket
4. Apriori - df_cleaned, df_encoded, split data to see if customer behavior change

Gaussian Mixture Model –df_clean

Types of products purchased

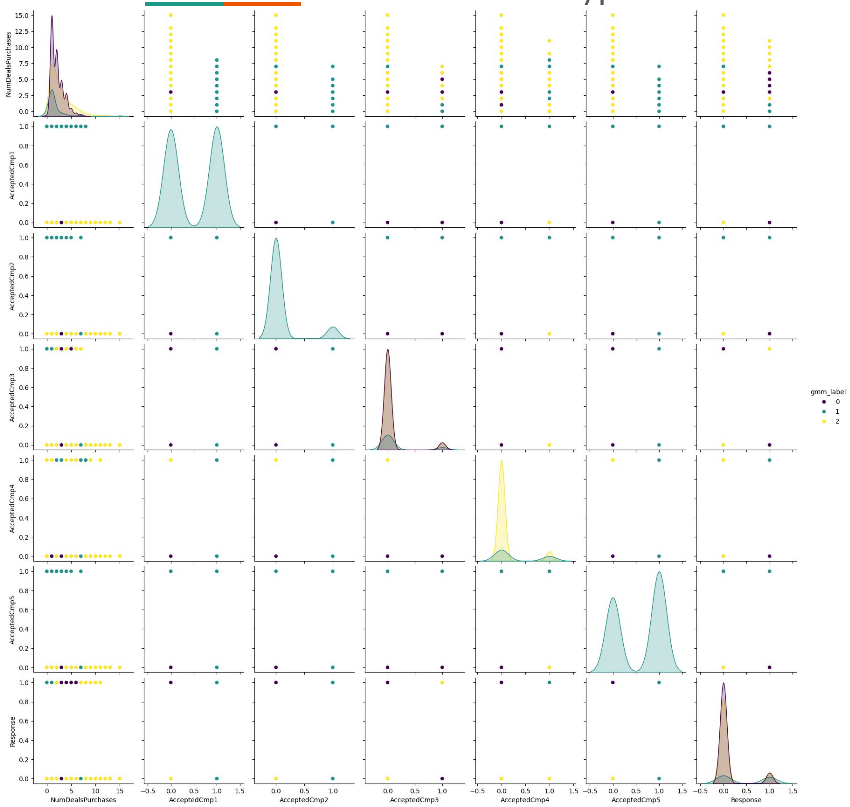


Categorize customers based on profiles

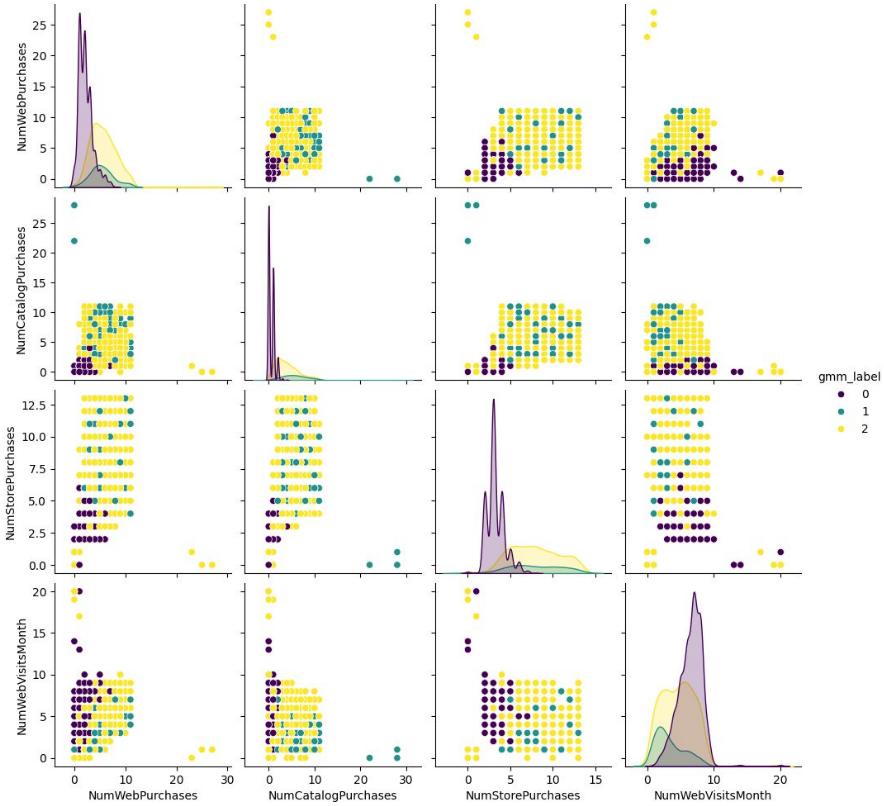


findings(corre/gaussian mixture)

Promotion types



Places to buy



Findings from gmm graph

1. Mint wines, mint meat can be mostly categorized by 3 types of customers, while the others can be categorized into two types
2. Income seems to be one of the most obvious feature that separate the customers apart

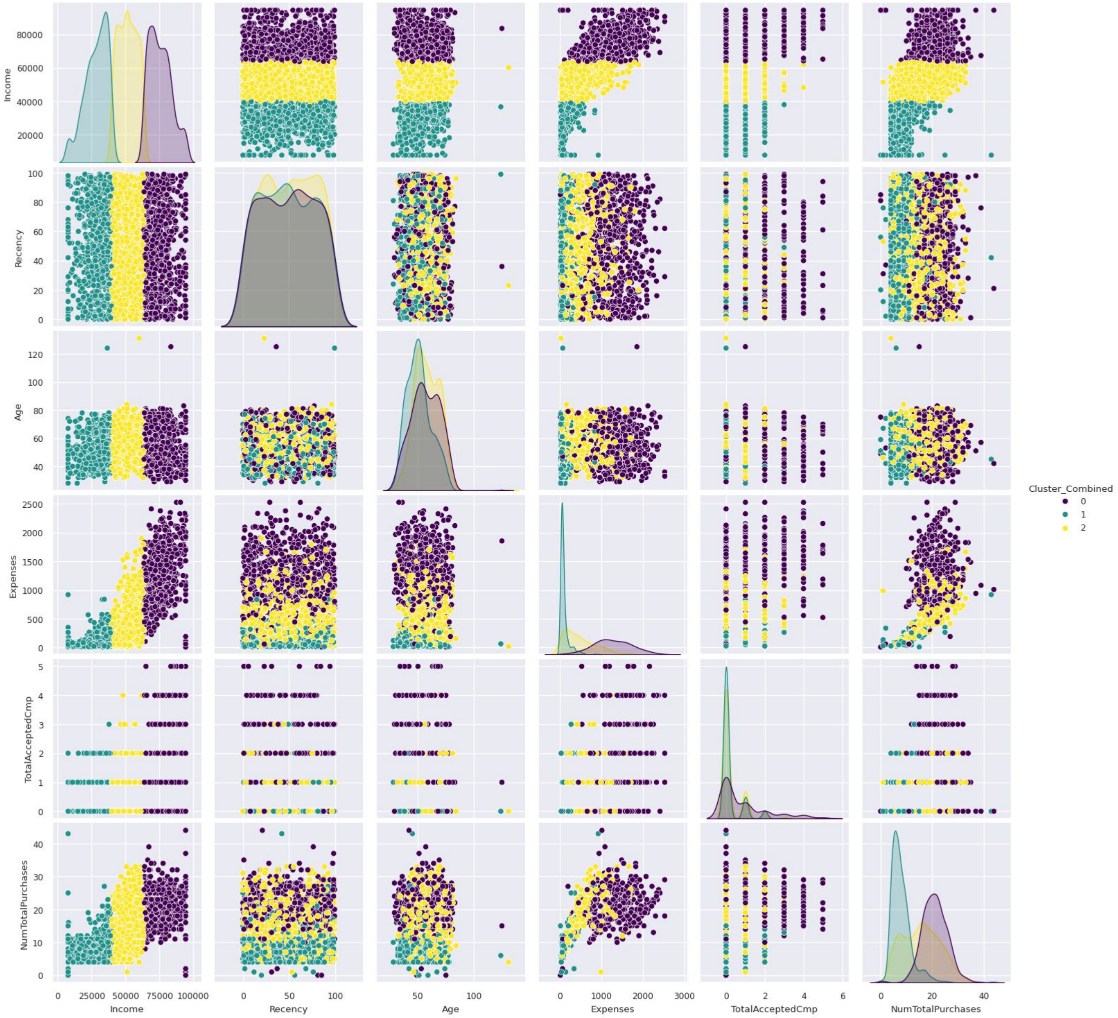
Applications and future questions?
— Does income have an effect on consumer purchase ability?
— recommend stock and sales based how many people fit into those customers group.

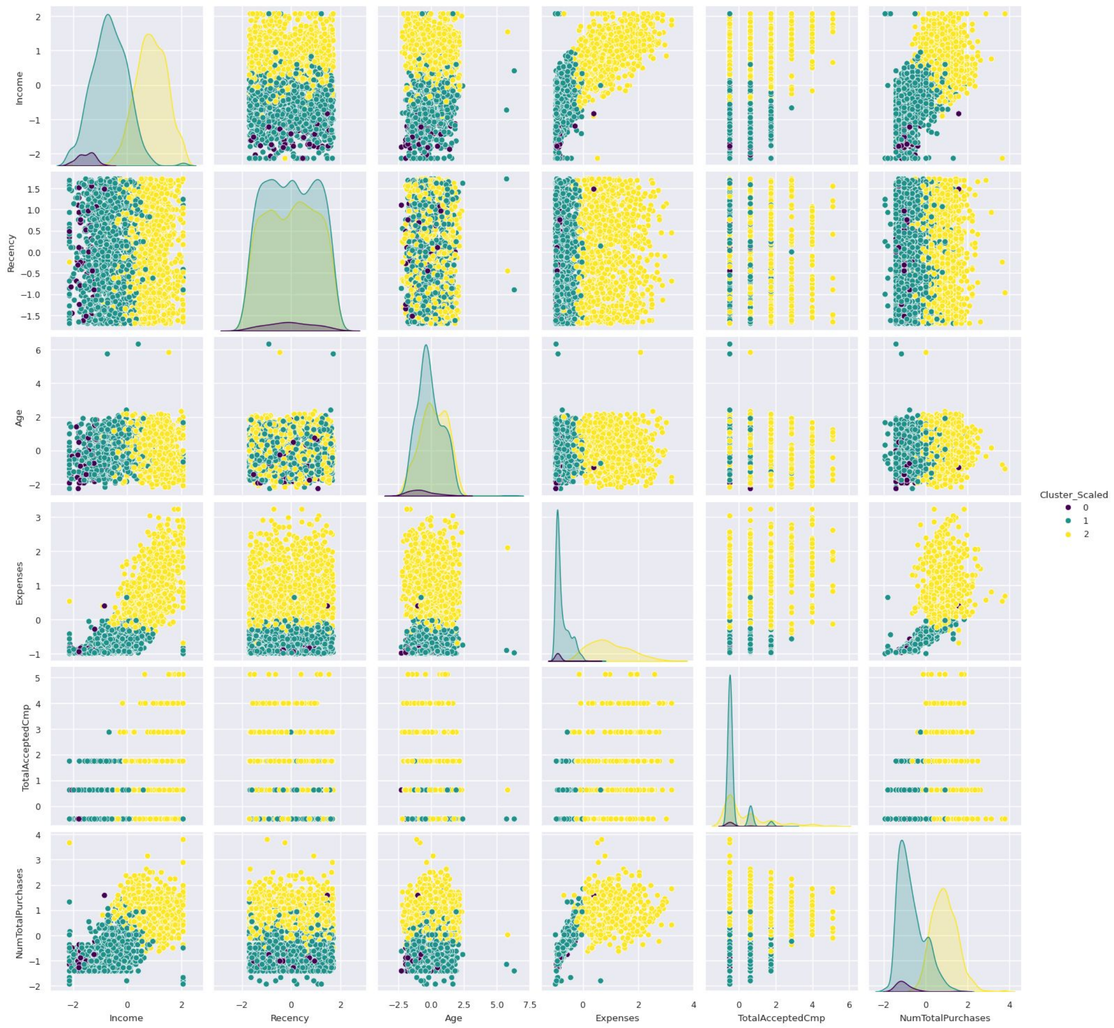
K-Means

- Feature selection and concatenation
- Elbow Method / Silhouette_score to find best “k”
- Applying K-Means for both unscaled and scaled data and finding clusters
- Clusters interpretation with scatter plots

Findings: k_means before scaled

- Each scatter plot shows the relationship between two variables, with different colors representing different clusters.
- The diagonal shows histograms for each variable, letting us see how they are distributed on their own.
- Customers with high income tend to have higher overall expenses.

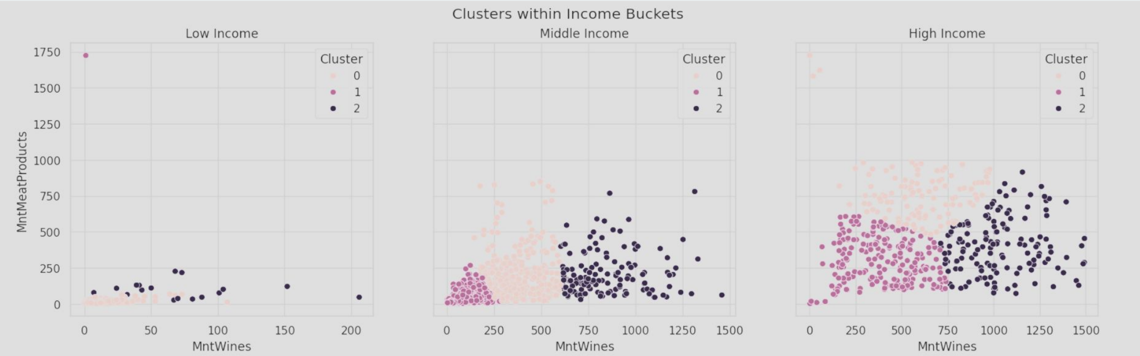




Findings: k_means after scaled

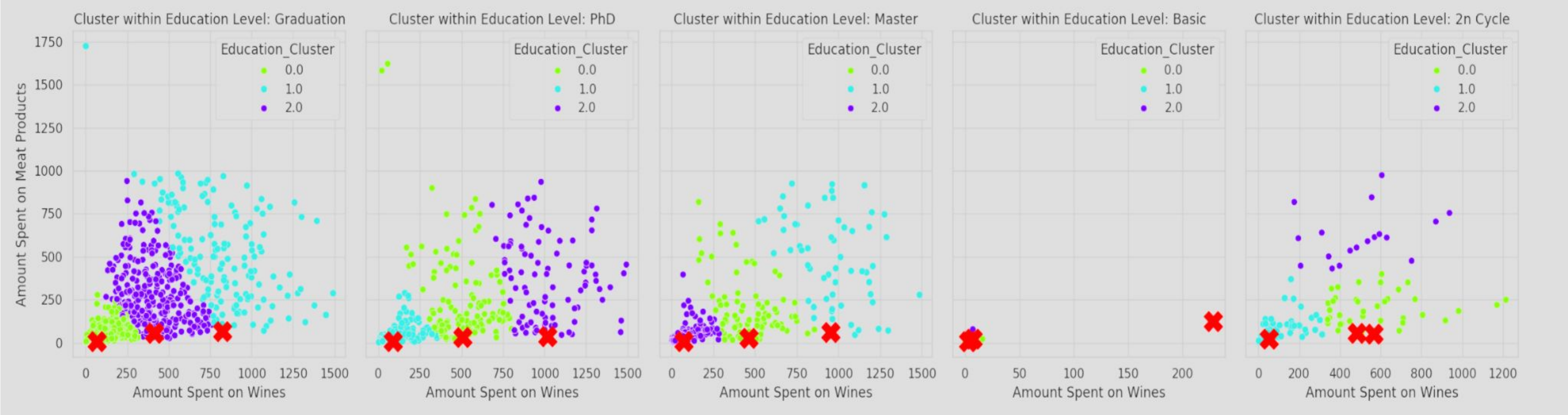
- Customers with high income tend to have higher total expenses/spends and purchases.
- The relationship between ‘income’ and ‘total_accepted_camp’ also indicates that they are more likely to accept campaign offers.

Findings(k_means bucket: Income)



- In the 'Low Income' bracket, there are very few points, indicating less data or less variability in spending. Most of the spending on wines and meat products is low
- The 'Middle Income' bracket shows a larger spread of points, with distinct clustering visible. There's a concentration of data points at the lower end of spending on wines, but a more distributed pattern on meat products spending.
- In the 'High Income' bracket, there is a much higher density of data points, especially in the mid to high range of wine spending. The clusters here appear to be more dispersed, indicating more variation in spending habits within this income group.

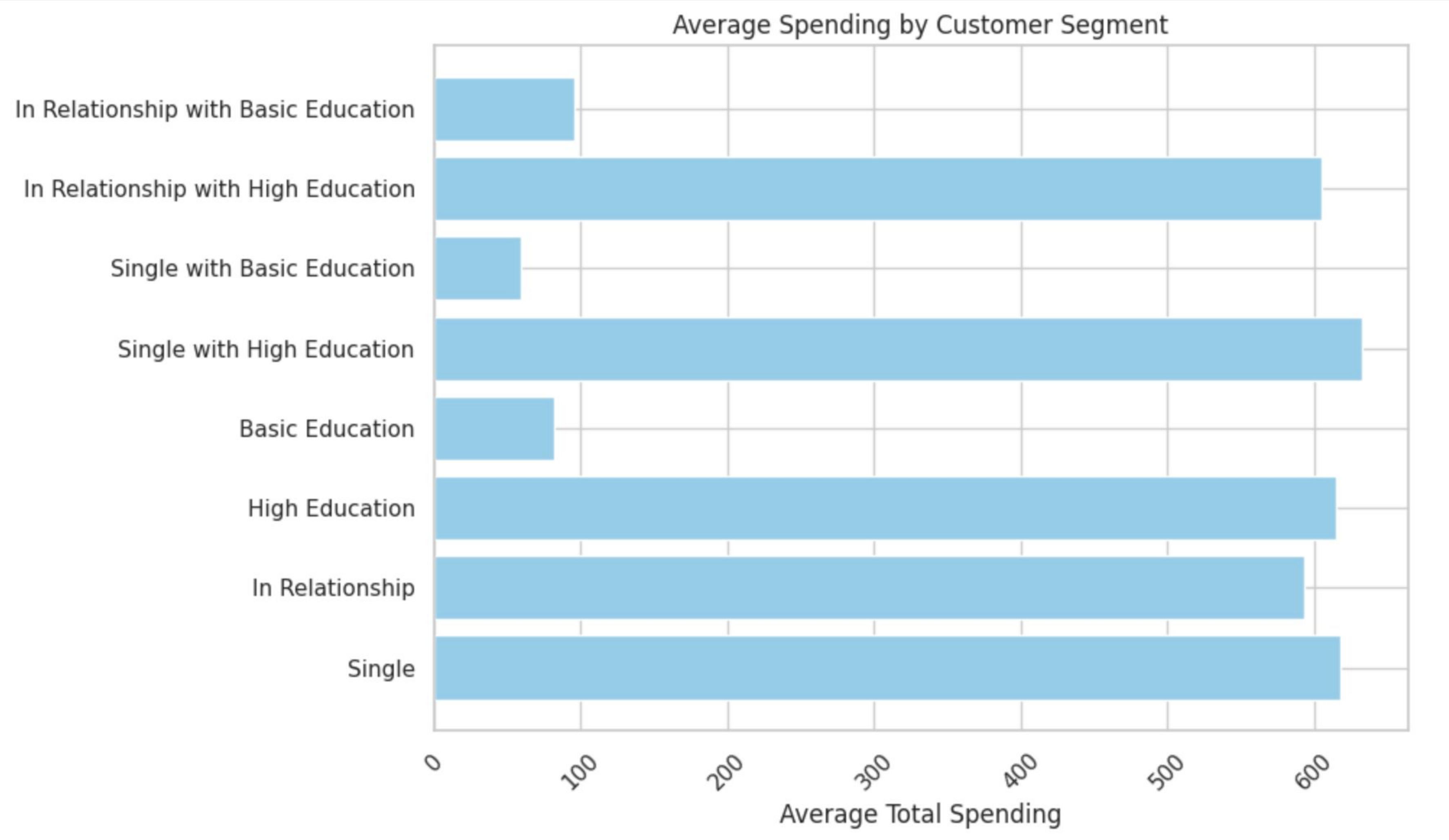
Findings(k_means buckets: Education)



- Graduation: This education level exhibits a broad range of spending patterns, encompassing both low and high spenders.
- PhD: Consumers with a PhD level of education tend to have a concentrated spending pattern, with most falling into a moderate spending range.
- Master: Consumers with a Master's level of education generally have lower overall spending, particularly on meat products.
- Basic: With fewer data points, this graph indicates a smaller consumer group at the basic education level.
- 2n Cycle: Consumers at the 2n Cycle education level show concentrated spending on wines with lower expenditure on meat products.

Findings (Apriori algorithm)

- df_cleaned
  - Most likely to buy wine product with meet (fruit with wine, fish with wine and meet, sweet with wine and meet, gold with wine and meet)
  - Rank product buying frequency
    1. Wine
    2. fruits
    3. meat
    4. fish
    5. sweet
    6. gold
- df_encoded
  - Split customers into different groups by education level, marriage status (no effect)


Average Spending by Customer Segment
In Relationship with Basic Education
In Relationship with High Education
Single with Basic Education
Single with High Education
Basic Education
High Education
In Relationship
Single
0
100
200
300
400
500
600
Average Total Spending

Conclusions & Business Suggestion

- Conclusions:
  - Higher educated/Income people intended to spend more
  - Wine/meat purchase can be grouped into 3 groups while other products only two.
  - Promotions 1,2 are generally more effective
- Business Suggestion:
  - Increase wine stock and decrease gold stock
  - Target on customers who have higher education level/higher income
  - Send wine related promotions to stimulate customer spend
  - Give product advertisements based on customer groups. To reduce promotion repetition, only give multiple promotions to the groups that are likely to respond.

Future steps

- ➔ Implement more algorithms and models
- ➔ Find more relative datasets to compare with
- ➔ Include more features to generalize the result

Thank you!