DATASCI207 Final Project

Customer Personality Analysis

Catherine Liao
Chloe Nguyen
Phoebe Yueh

Berkeley SCHOOL OF INFORMATION

Introduction

Project objective:

Use unsupervised algorithms to identify and explore customer personas

Why?

- Allow companies to better understand their customer base
- Improve customer experience
- Guide strategic decision-making for enhanced business performance

Data Source

Primary Dataset: Customer Personality Analysis
https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

Variable	Definition
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
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

Variable	Definition
MntWines / MntFruits/ MntMeatProducts/ MntFishProducts/ MntSweetProducts/ MntGoldProds	Amount spent on wine/fruits/meat/fish/sweet/gold in the last 2 years

Variable	Definition
NumWebPurchases	# of purchases made through websites
NumCatalogPurchases	# of purchases made through catalogues
NumStorePurchases	# of purchases made through stores
NumWebVisitsMonth	# of visits to company websites in the last month

Variable	Definition
NumDealsPurchase	Number of purchases made with a discount
AcceptedCmp 1 -5	1 if the customer accepted the nth campaign offer, 0 otherwise

Data Cleaning Process & Assumptions

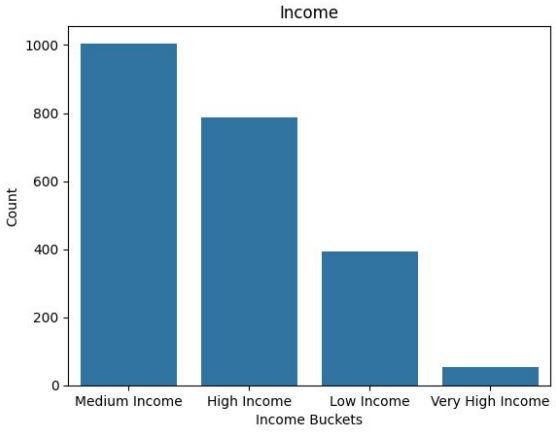
- 2240 records, 29 columns
- Replaced **24 null values** in the ‘Income’ column with 0s

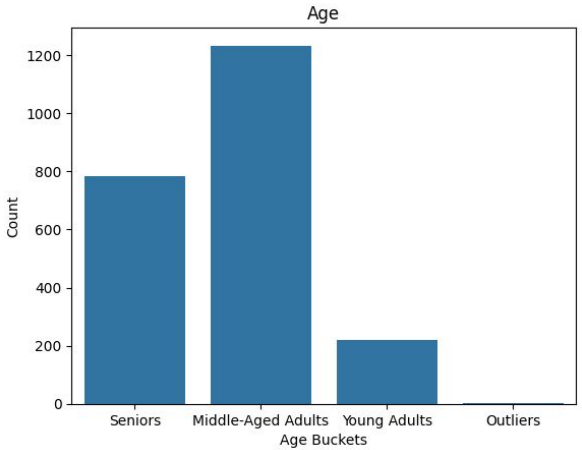
Feature Engineering

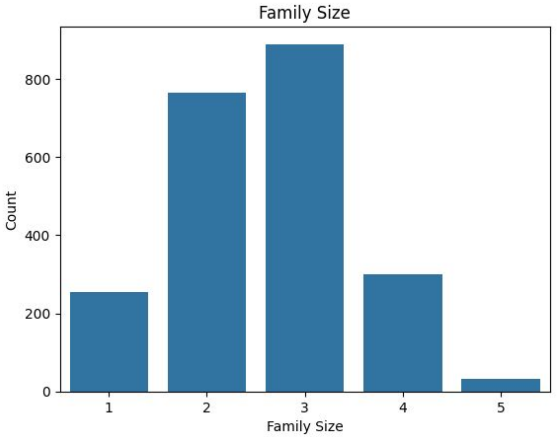
Variable	Definition
Year_Birth	Customer's birth year
Marital Status	Customer's marital status
Kidhome	Number of children
Teenhome	Number of teenagers
Dt_Customer	Date of customer's enrollment with the company
AcceptedCmp 1 -5	1 if the customer accepted the nth campaign offer, 0 otherwise
MntWines / MntFruits...	Amount spent on wine/fruits/meat/fish/sweet/gold in the last 2 years

- Age: Transformed from Year_Birth
- Family Size: Created a Family Size column using Marital Status, Kidhome, and Teenhome
- Days_Enrolled (Total Number of days a customer is enrolled with the company): today() - Dt_Customer
- TotalAcceptedCmp (Total Number of offers a customer accepted) : AcceptedCmp1 +... + AcceptedCmp5
- TotalAmtSpent (Total Amount Spent on all Products) : MntWines + MntFruits + MntGoldProds + ...

Exploratory Data Analysis (EDA)

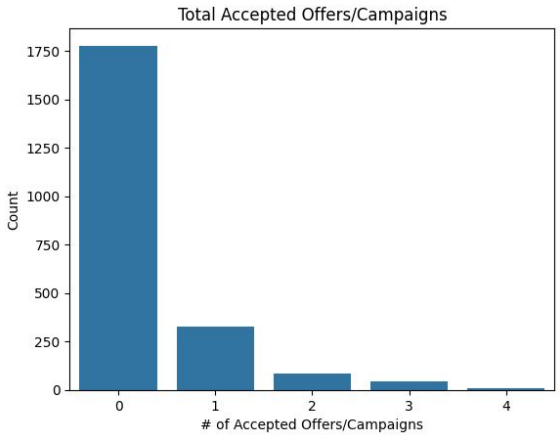


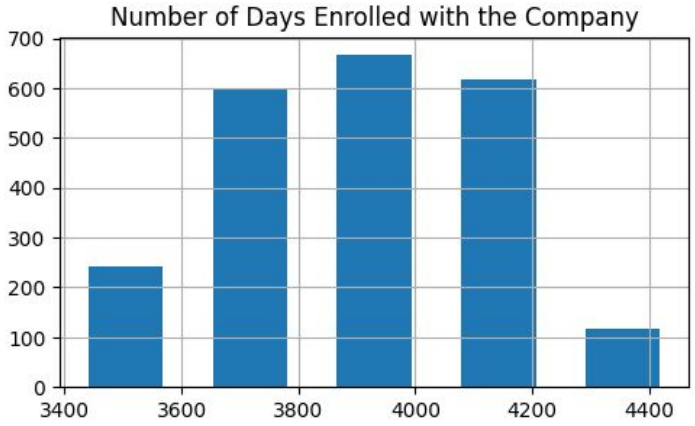


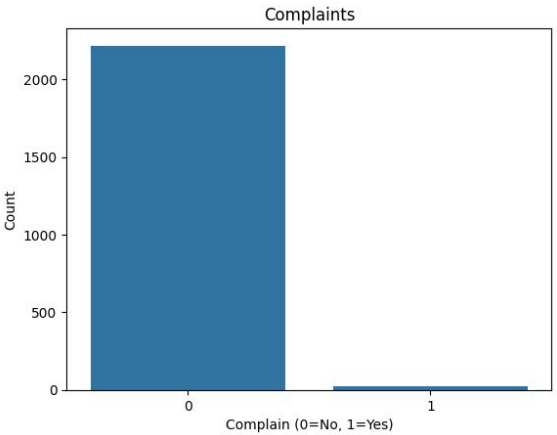


- **~ 55%** of the total population consists of individuals aged between 40 and 59, categorizing them as middle-aged adults.
- **~45%** of the total population falls within the income bracket ranging from $30,001 to $60,000, indicating a medium income range.
- **~40%** of the total population consists of families with a size of 3 members.

Exploratory Data Analysis (EDA)







- **~ 79%** of the total population declines all campaign offers presented to them.
- On average, customers have been enrolled with the company for **10.7 years**.
- Only **0.9%** of the customers registered a complaint within the last two years.

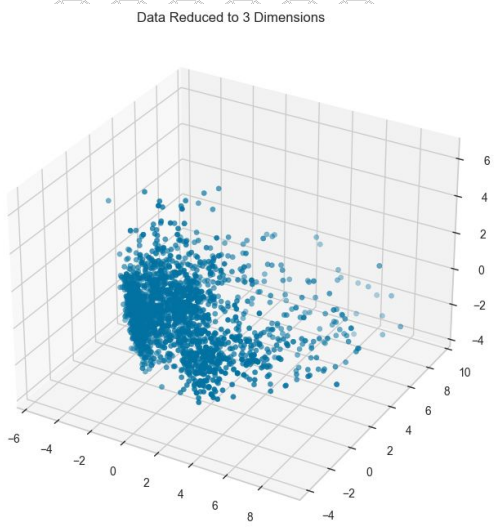
Pre-Processing

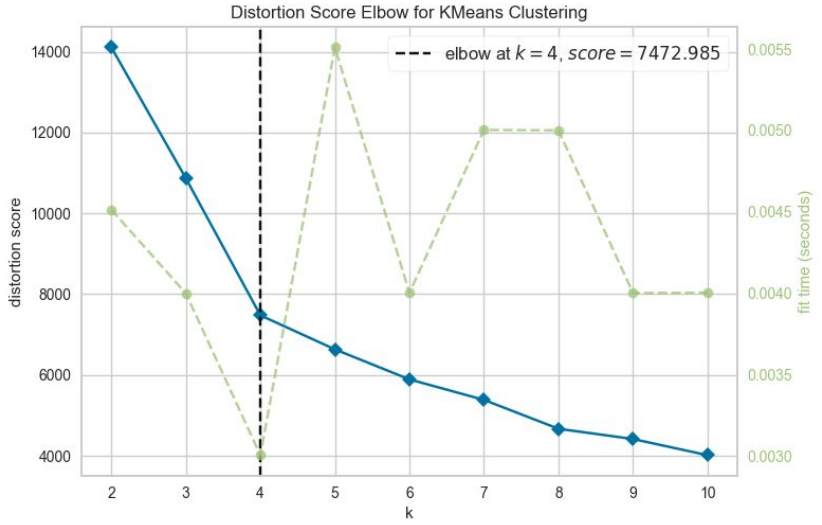
Step 1: Convert Categorical labels to Numerical labels
- Label Encoder (sklearn library)

Step 2: Standard scale all features
- StandardScaler (sklearn library)

Step 3: Reduce to 3 dimensions
- Principal Component Analysis (sklearn library)

Step 4: Determine optimal cluster count
- Elbow method (yellowbrick library)





Model 1: MiniBatch K-Means

Cluster 0

- Younger families with 1-2 kids
- Low income levels ($34758)
- Lowest spending habits
- Lowest response to marketing campaigns
- Most web visits

Cluster 1

- Middle-Aged couples with no children
- Moderate income levels ($73530)
- High spending habits across most categories (Fish, meat, fruits & sweets)
- Most store purchases
- Very Satisfied (complain rate ~0%)

Cluster 2

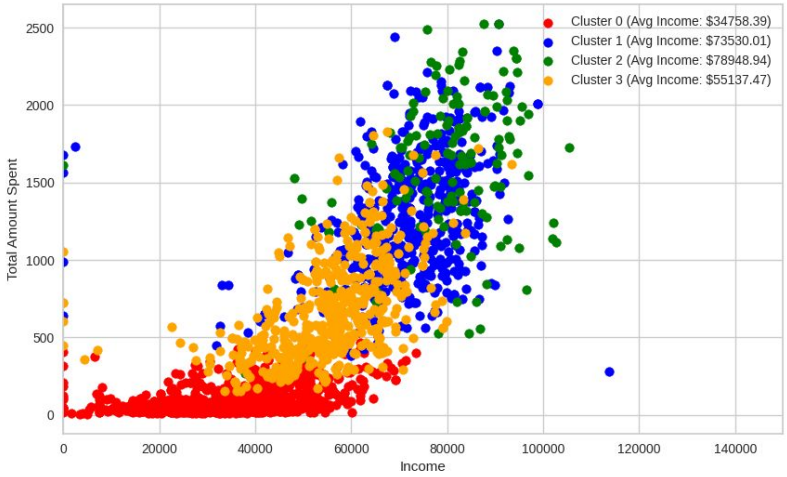
- Younger couples with no children
- Highest income levels ($78949)
- High spending habits (especially wine & gold)
- **Highest response to marketing campaigns**
- Most catalog purchase

Cluster 3

- Middle-aged families with 1-2 kids
- Average income levels ($55137)
- Moderate spending habits
- Most deal purchases
- Most web purchases

Euclidean Distance

- Minimum Distance: between centroid 0 and centroid 3 (~3.24)
- Maximum Distance: between centroid 0 and centroid 2 (~8.33)



Model 2: Agglomerative Clustering

Cluster 0	Cluster 1	Cluster 2	Cluster 3
• Younger families with 1-2 kids • Low income levels ($34005) • Lowest spending habits • Lowest response to marketing campaigns • Most web visits	• Older families with 1-2 kids • Average income levels ($58442) • Moderate to low spending habits across all categories • Most deal purchases • Most store and most web purchases • High web visits	• Singles and couples with no children • Highest income levels ($78875) • High spending habits- esp wine • Highest response to marketing campaigns by far	• Singles and couples with no children • Moderate income levels ($73395) • Highest spending habits across all categories except wine and gold • Moderate response to marketing campaigns

Model 3: DBSCAN

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
• Mean Income of 28701.05 • Family Size 2-3 with 1 child • Most recent customers • Most responsive	• Mean Income of 47324.94 • Family Size 2-3 with 1 child • Most Deals purchased	• Mean Income of 60820.00 • Family Size 2-3 with 1 child • Has highest gold sales • Highest web and store purchases	• Mean Income of 67968.00 • Family Size 1-2 without child • Highest wine sales • Highest web visits	• Mean Income of 77086.04 • Family Size 1-2 without child • Has highest fruits, meat, fish, and sweets sales • Highest catalog purchases

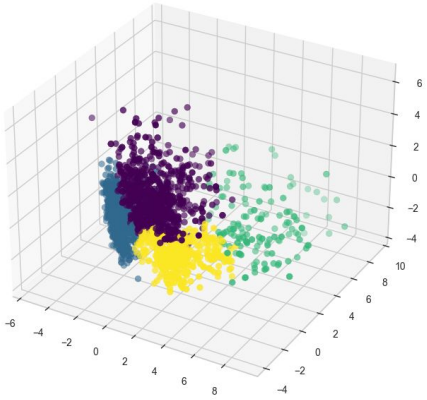
Model 4: Gaussian Mixture Model

Cluster 0	Cluster 1	Cluster 2	Cluster 3
• Mean Income of 35060.47 • Family Size 2-3 with 1 child • Most website visits	• Mean Income of 53594.37 • Family Size 2-3 with 1 child • Most deals purchased	• Mean Income of 71247.44 • Family Size 2-3 with 1 child • Most wine sales • Most web purchases • Most responsive	• Mean Income of 76298.20 • Family Size 1-2 without children • Most fruit, meat, fish, sweets, gold Sales • Most catalog, store purchases

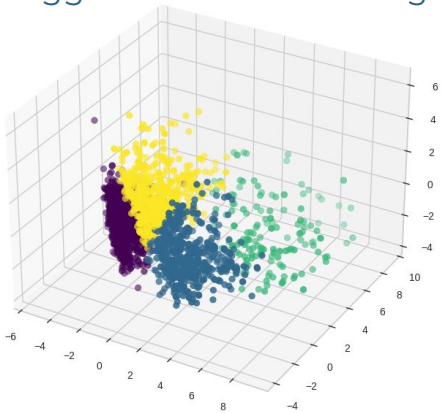
Conclusions & Model Comparisons

- All models showed similar results
- Common personas among the models:
  - Customer persona 1: Lowest income families (1-2 kids) with low spending habits across all products and all purchasing methods
  - Customer persona 2: Moderate income families (1-2 kids) with moderate spending habits and high deals, stores and web purchases
  - Customer persona 3: High income singles/ small families (0-1 kid) with high spending habits across all product categories- especially wine sales

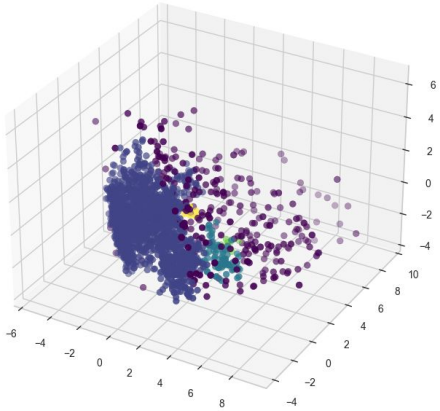
K-Means



Agglomerative Clustering



DBSCAN



GMM