Customer Personality Analysis Using Clustering Algorithms

The problem

- Customer personality analysis helps businesses to change their products and marketing services based on their customers behaviours.

- This implies cost reductions and improvements in customer satisfaction in different ways:
  - identify customers close to churn by tracking their behaviour over time
  - offering specific add-ones to the right customers
  - Analyzing business processes in terms of improving overall profit

Dataset: Customer Personality Analysis

The dataset includes customers and their attributes:

Year_Birth

Education: education level

Marital_Status

Income: yearly income

Kidhome: Number of children

Teenhome: Number of teenagers

Recency: last purchase

Complain: in the last 2 years

MntWines: Amount spent on wine in last 2 years

MntFruits: Amount spent on fruits in last 2 years

MntGoldProds: Amount spent on gold in last 2 years

NumDealsPurchases: Number of purchases made with a discount

AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise

NumWebPurchases: Number of purchases made through the company's website

NumStorePurchases: Number of purchases made directly in stores

Preprocessing
the Dataset

- Original dataset has **2240 rows and 29 columns**.
- After applying preprocessing techniques we have **(2216, 28)** .

Remove null values

remove 24 rows that have "income"= null

Feature Engineering

Dimension Reduction using PCA

Scaling values using RobustScaler

Feature Engineering

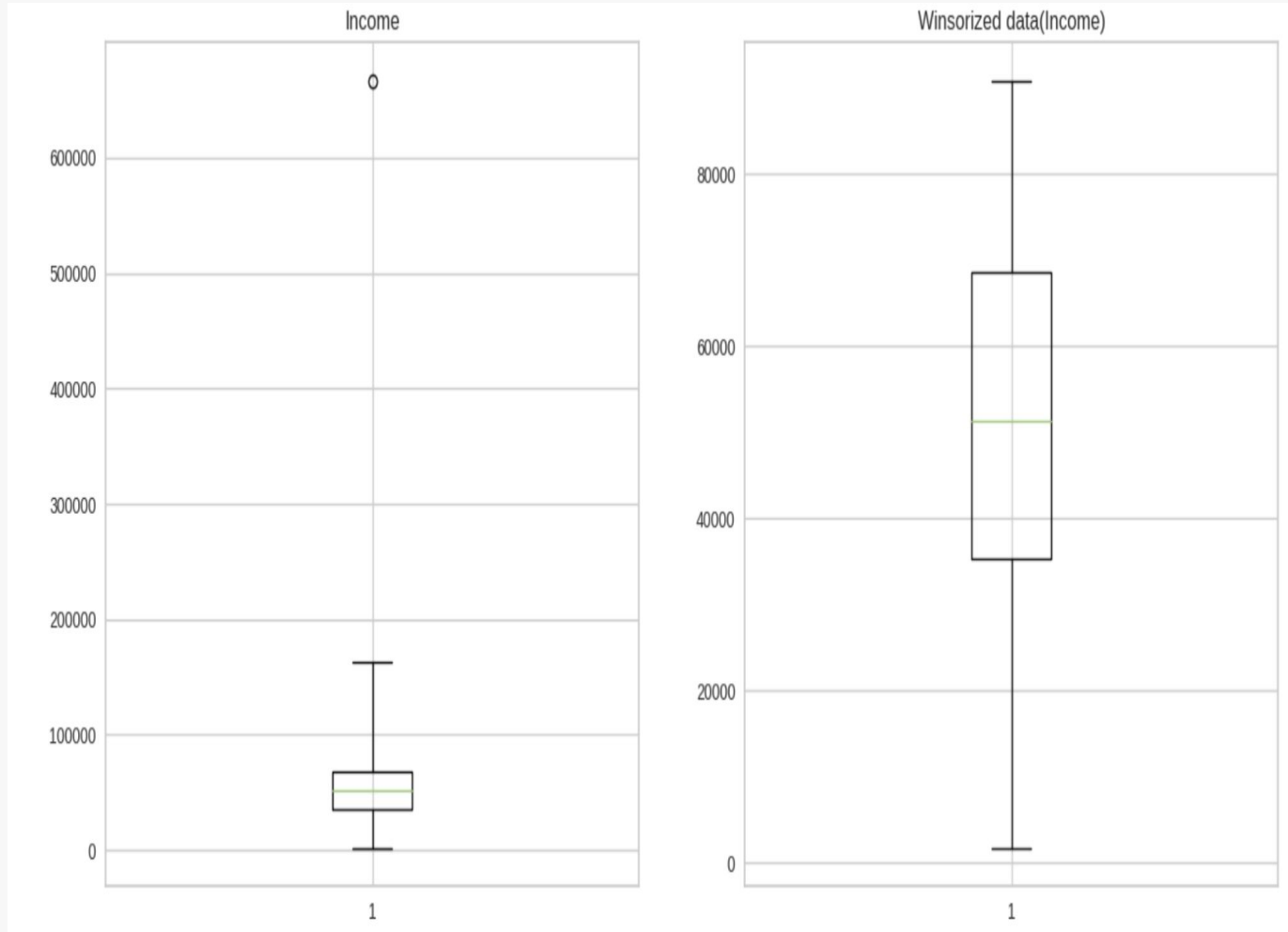Reduce **Marital status** to "Partner" and "Single" and Convert **Education** into "Undergraduate", "Graduate", "Postgraduate", and Birth date to **age.**

Add new attribute such as "**Day_with_market**", "**Total_spent**", "**Household_size**", "**NumTotalPurchases**" and "**Is_parent**".
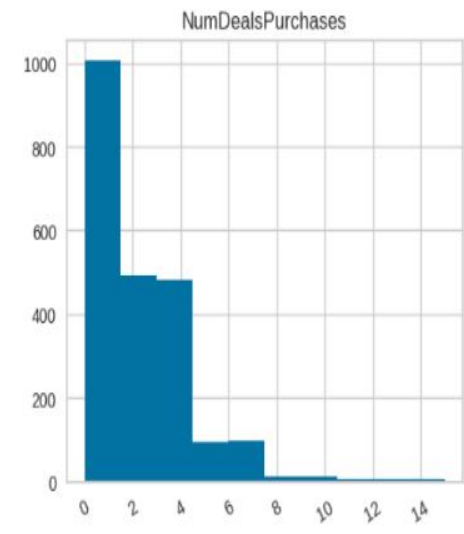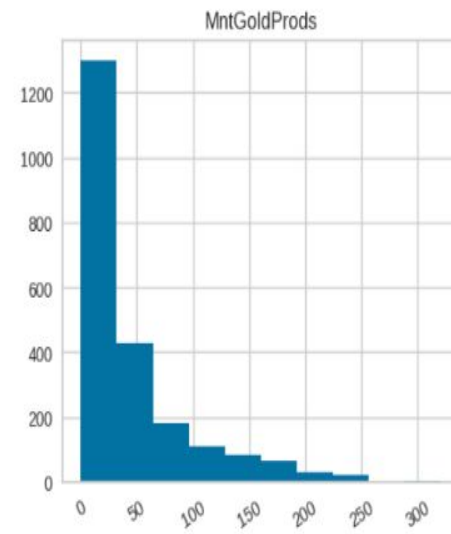
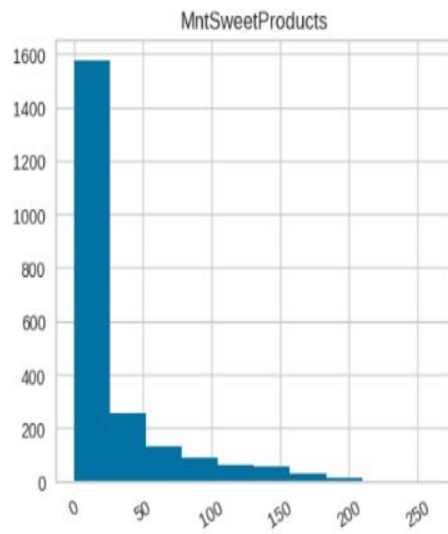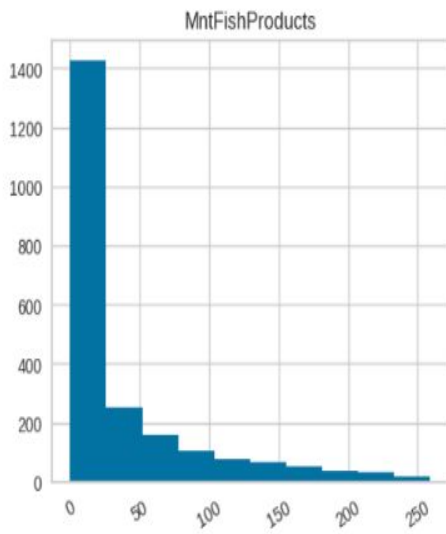Winsorizing some attributes to reduce the effectiveness of outliers.
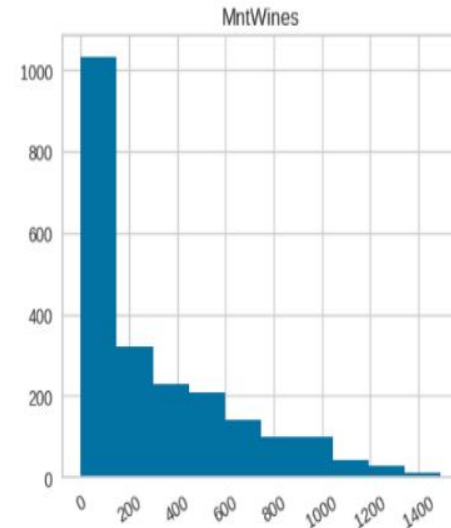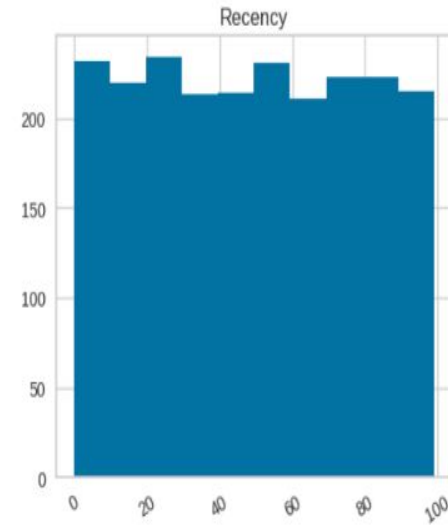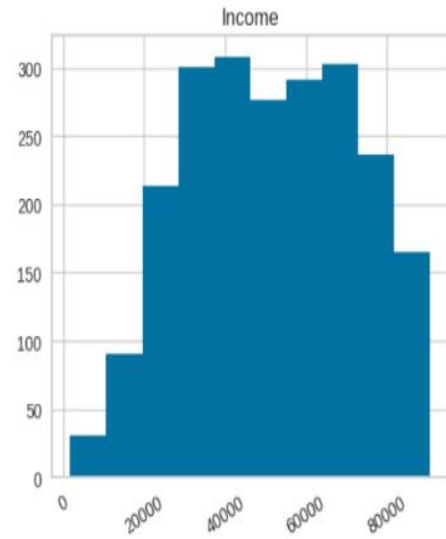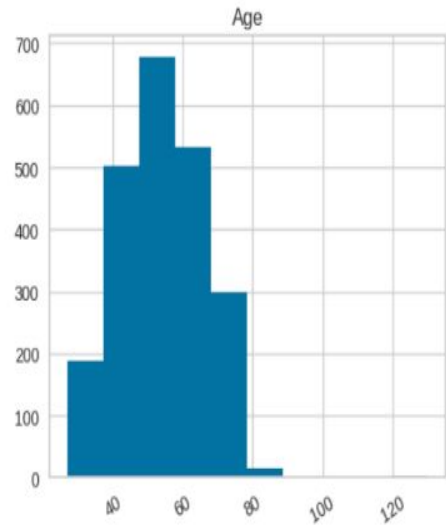
Winsorizing the unbalanced attribute

Exploratory **Data Analysis**
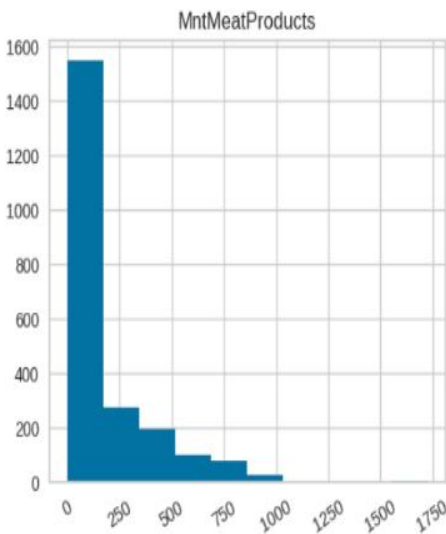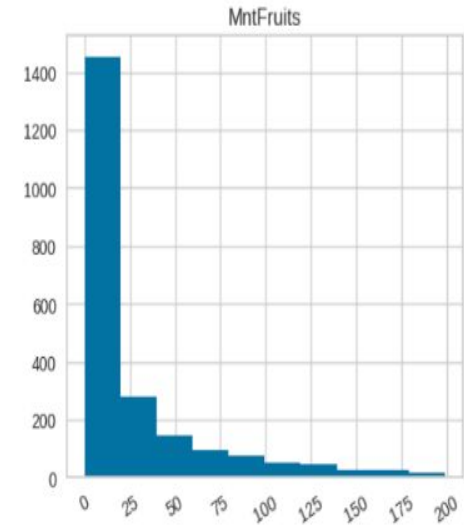
Customer's Average Spent on Products by Age Groups

Wines

Fruits

Meats
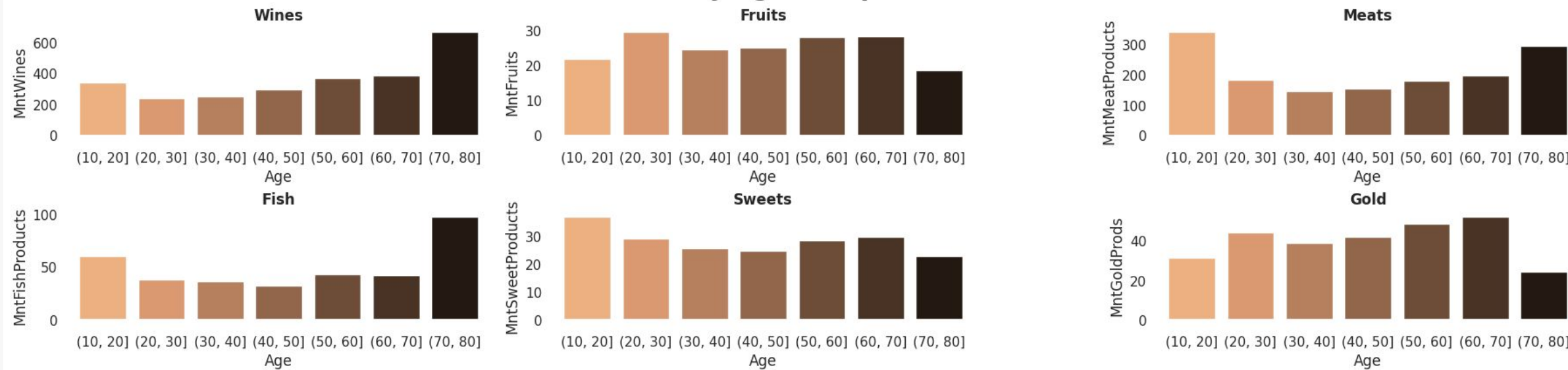
Fish

Sweets

Gold

Percentage of Company's Profit from Products

50.26% MntWines
27.51% MntMeatProducts
7.24% MntGoldProds
6.20% MntFishProducts
4.45% MntSweetProducts
4.34% MntFruits
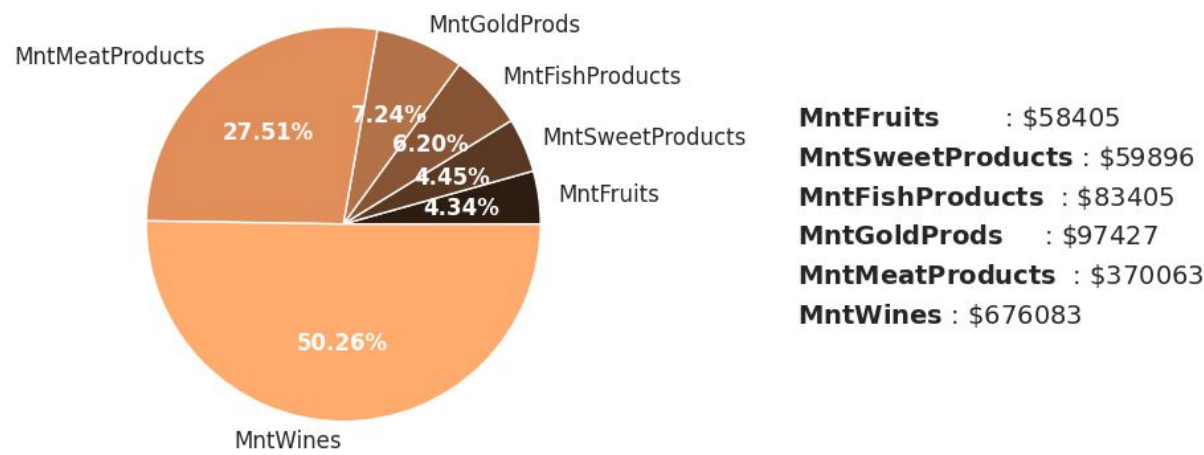
MntFruits   : $58405
MntSweetProducts : $59896
MntFishProducts : $83405
MntGoldProds   : $97427
MntMeatProducts : $370063
MntWines : $676083

Average Number of Purchases Made with a Discount by Age Groups
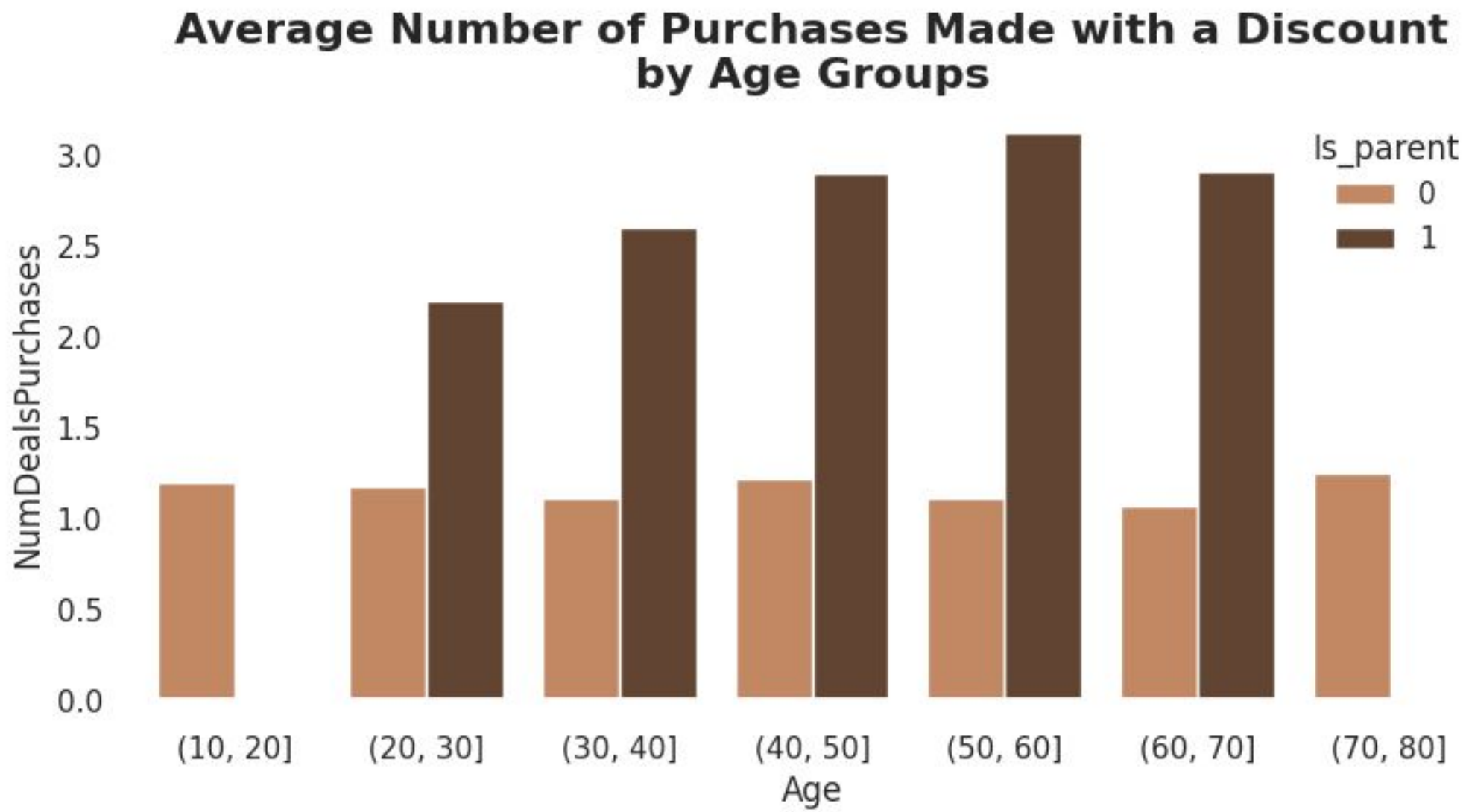
High Correlation

- **Total_spent**
- **Income**
- **MntWines**
- **MntMeat**
- **NumTotalPurchases**
- **Catalog**

- **Web**
- **Cataloge**
- **Store**
- **NumTotalPurchases**
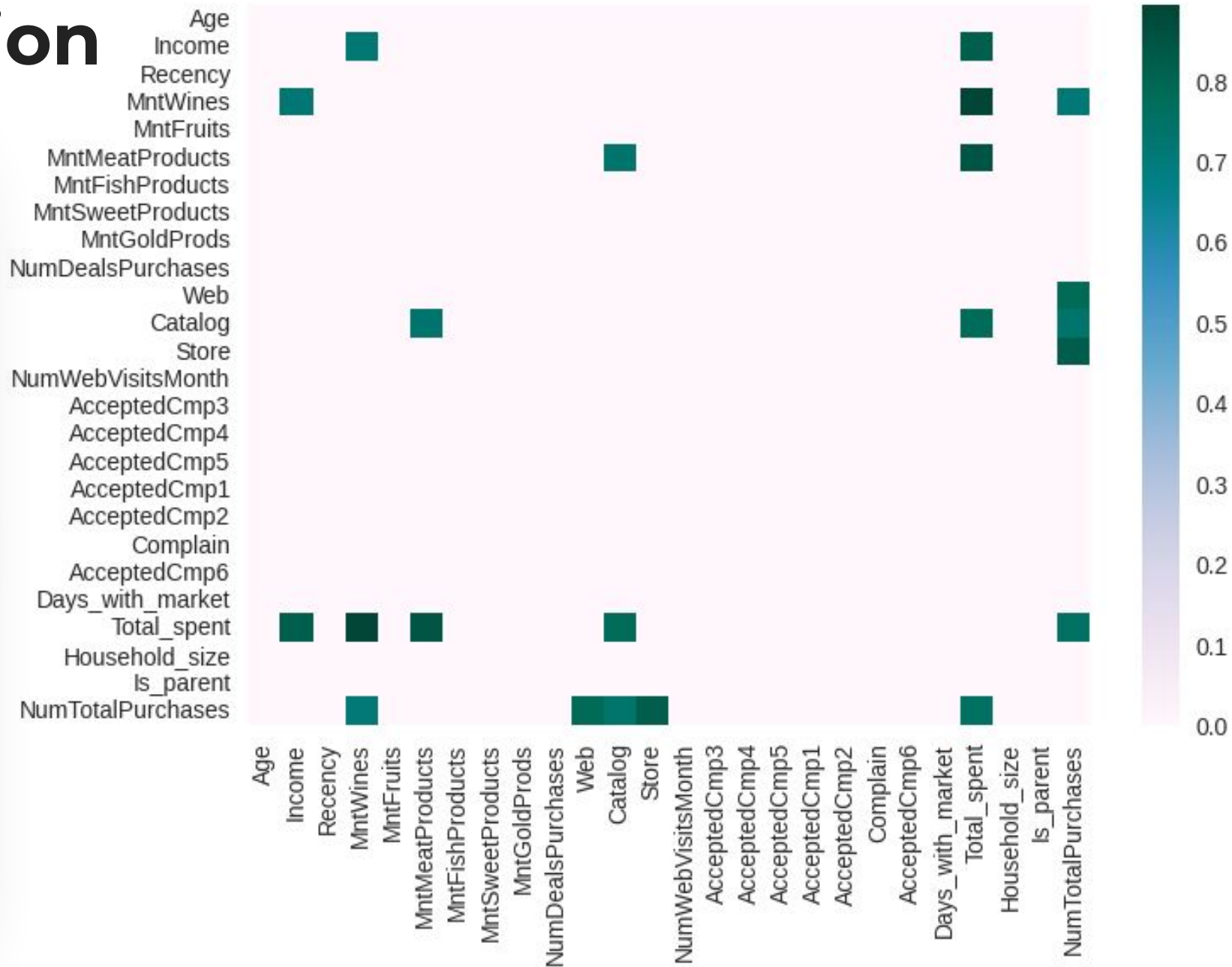
Dimension Reduction

We performed clustering on a dataset using dimensionality reduction techniques, including **correlation-based** dimension reduction and **PCA-based** dimension reduction. Additionally, we conducted clustering on a **subset of the dataset**, focusing on three specific features that we selected.

Feature correlation on DF_PCA
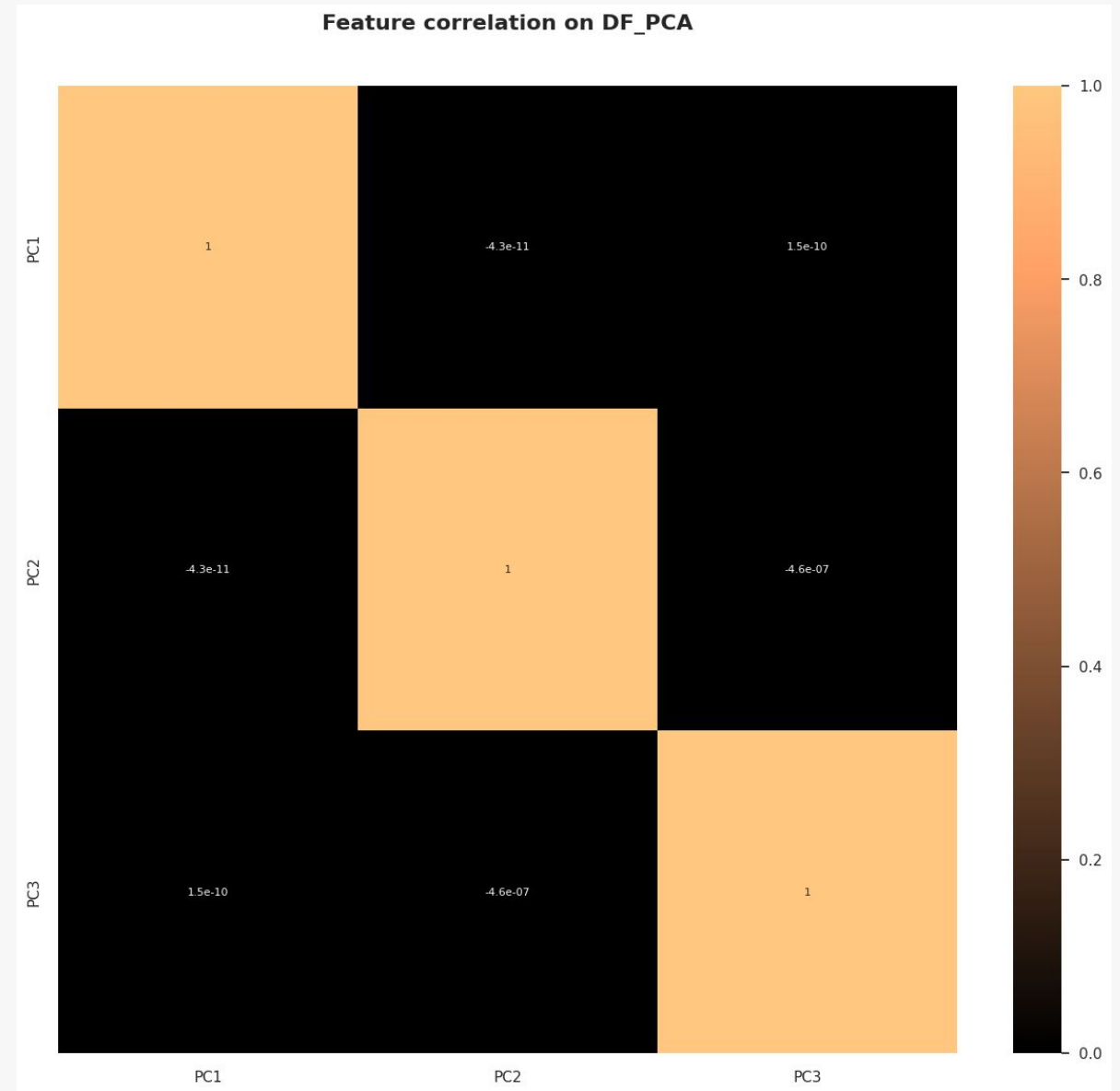
Choosing Parameters
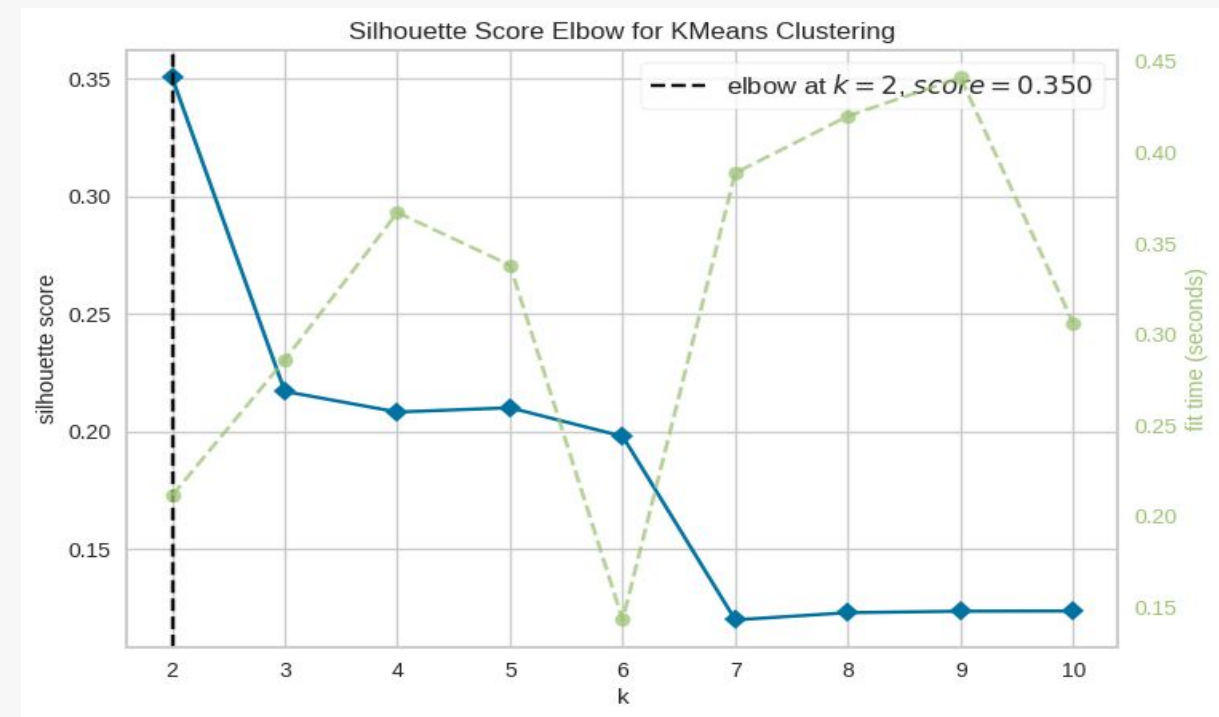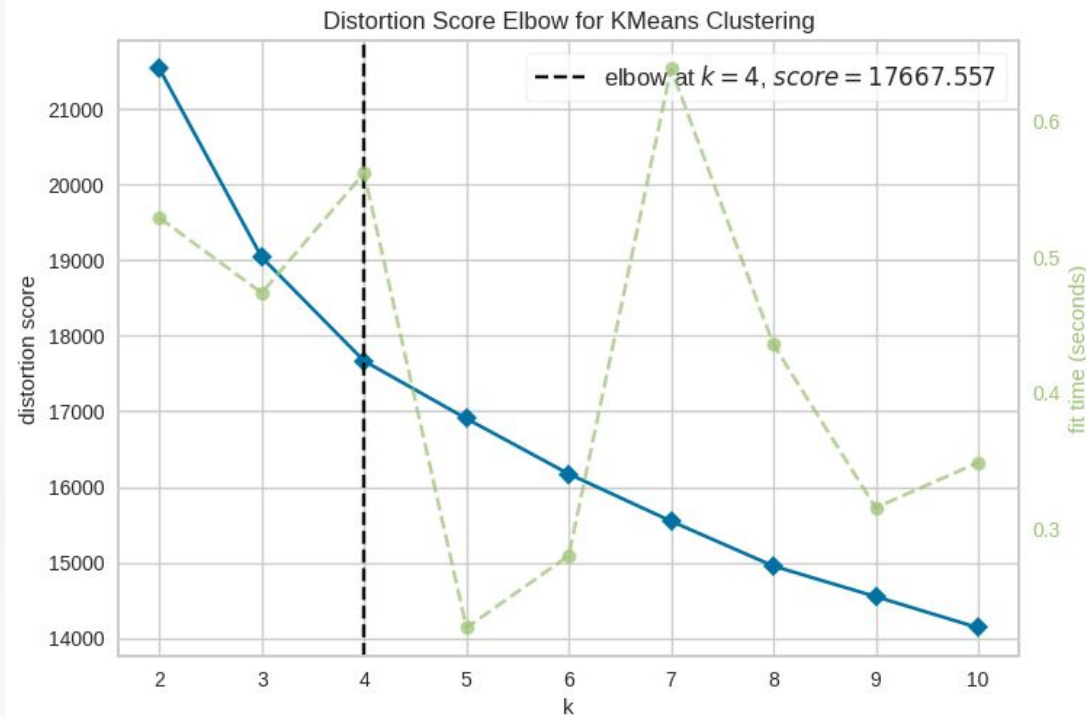
We choose k = 4 for KMeans and agglomerative clustering algorithms.

Experimental Results

Our clustering evaluation measure is the silhouette score, which quantifies cluster cohesion and separation. Higher scores indicate better clustering, with values close to 1 indicating well-defined and separated clusters, and values close to -1 indicating overlapping or poorly separated clusters.

- **Agglomerative Clustering**
- **KMeans Clustering**
- **GaussianMixture**

Algorithms	Silhouette Score
PCA + Agg.	0.409
PCA + KM	**0.426**
PCA + GaussianMix	0.388
CorrReduction + Agg.	0.110
CorrReduction + KM	0.199
CorrReduction + GaussianMix	0.199
FeatureSelection + Agg.	0.274
FeatureSelection + KM	0.333
FeatureSelection + GaussianMix	0.332

Cluster Analysis

The Plot Of The Clusters with agglomerative clustering
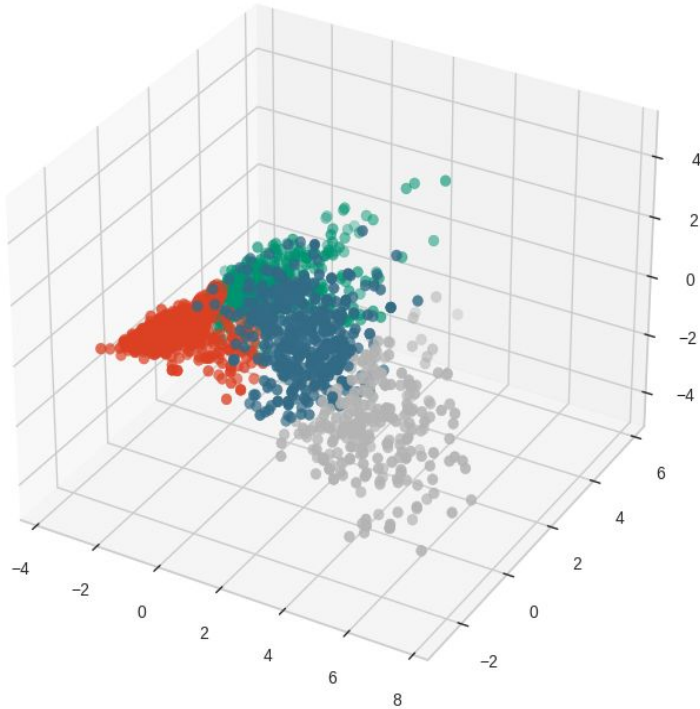
The Plot Of The Clusters with KMeans
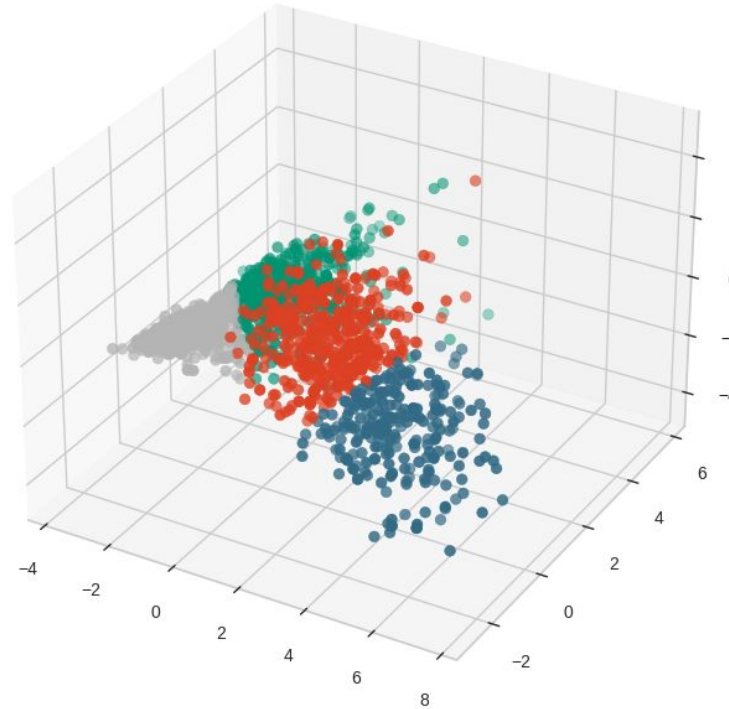
The Plot Of The Clusters with Gaussian Mixture
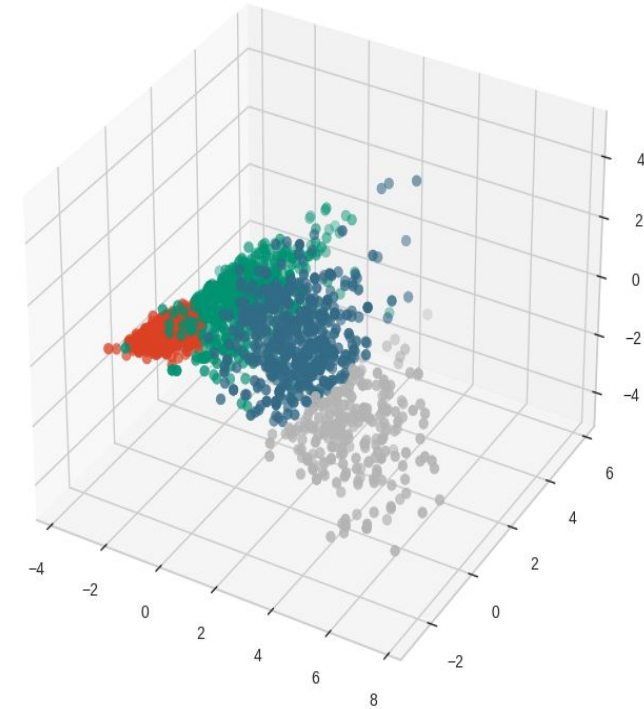
Distributions of clustering
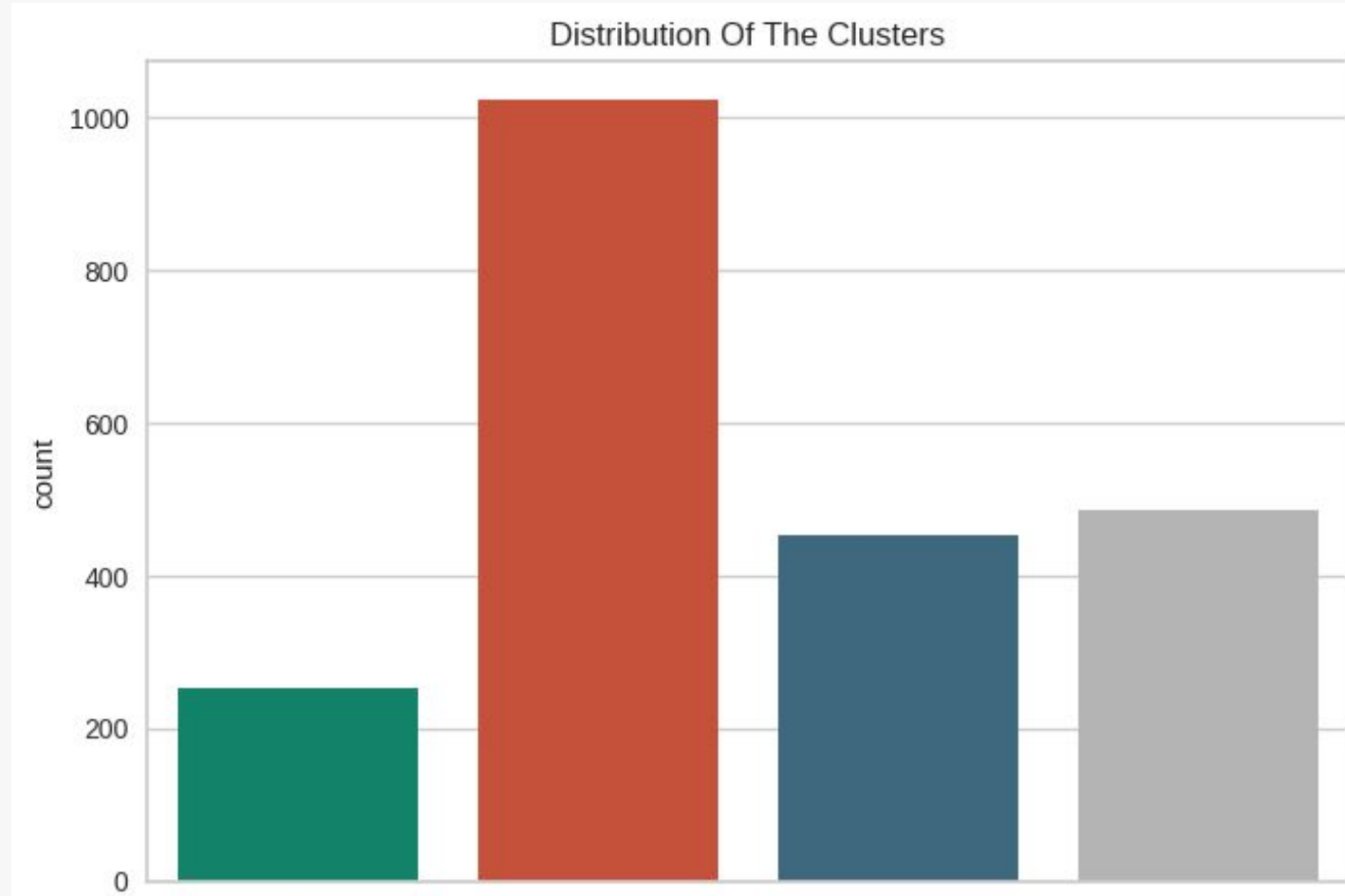
Segment Analysis Based On Income and Spending

Cluster's Profile Based On Income And Spending with KMeans clustering
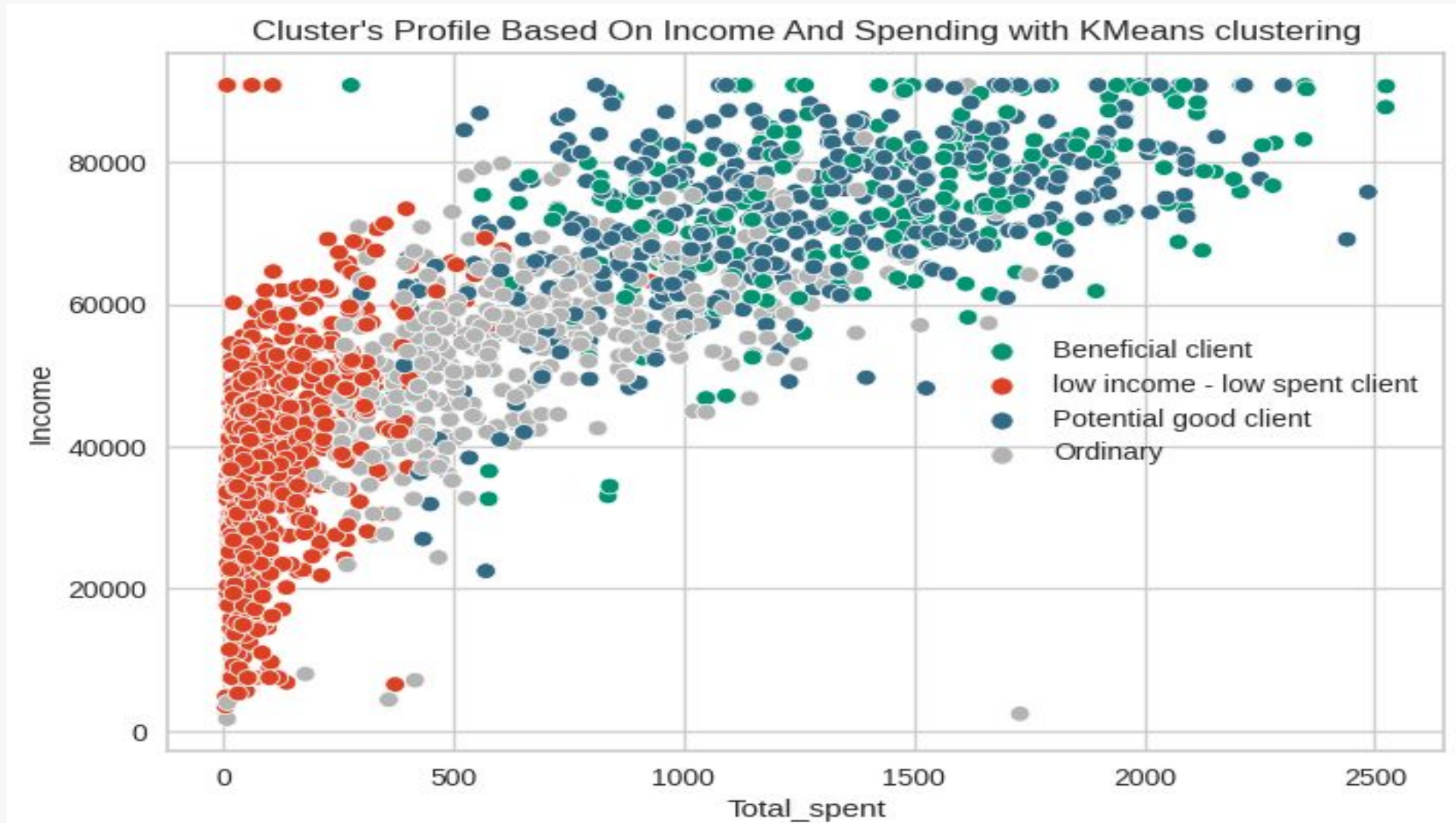
Customer Segment Analysis based on Education and spending
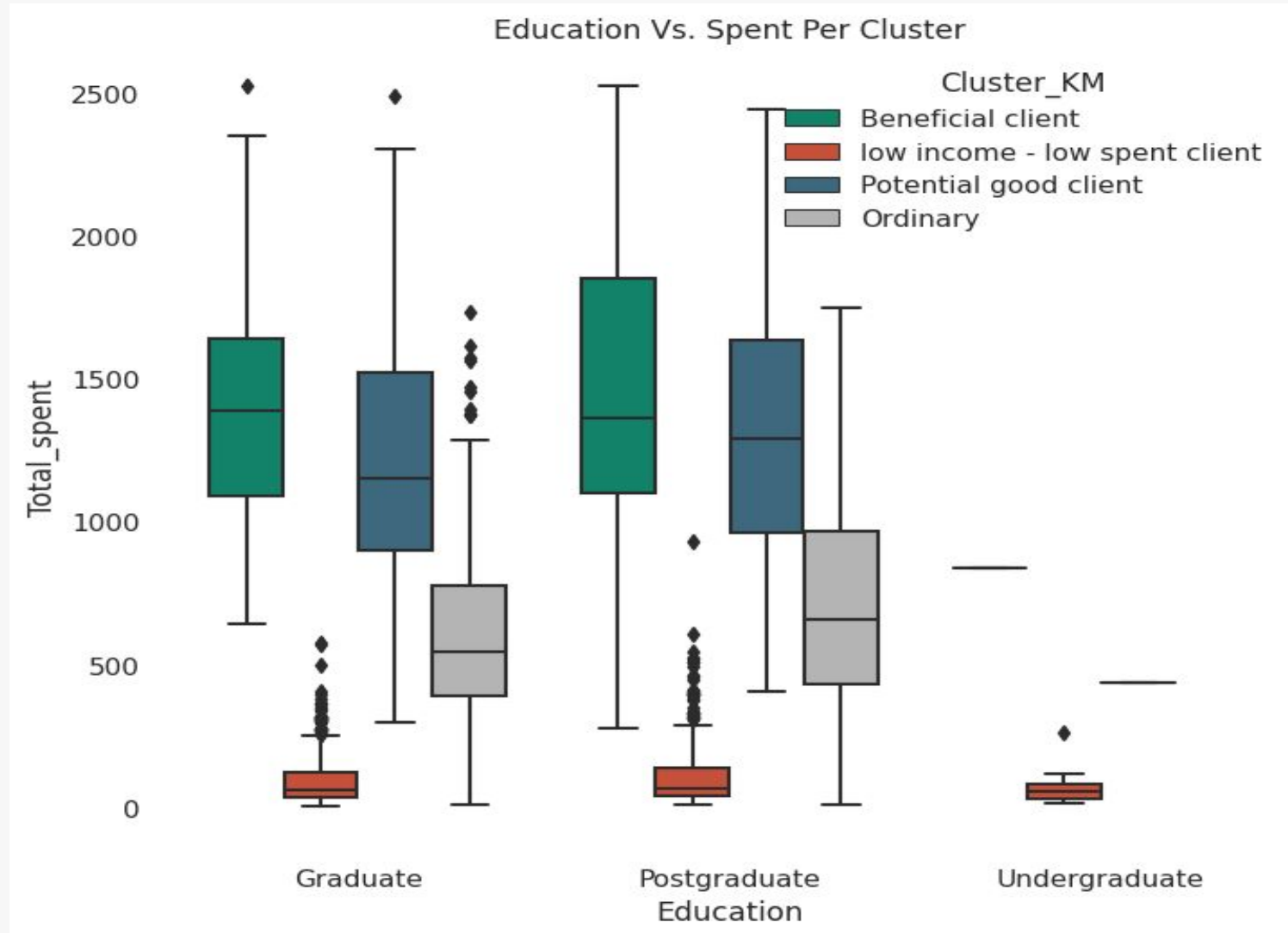
Customer Segment Analysis based on Household size and spending

Household_size Vs. Spent Per Cluster

What is the mean of household size for each cluster?

2.0   3.0   1.9   2.9
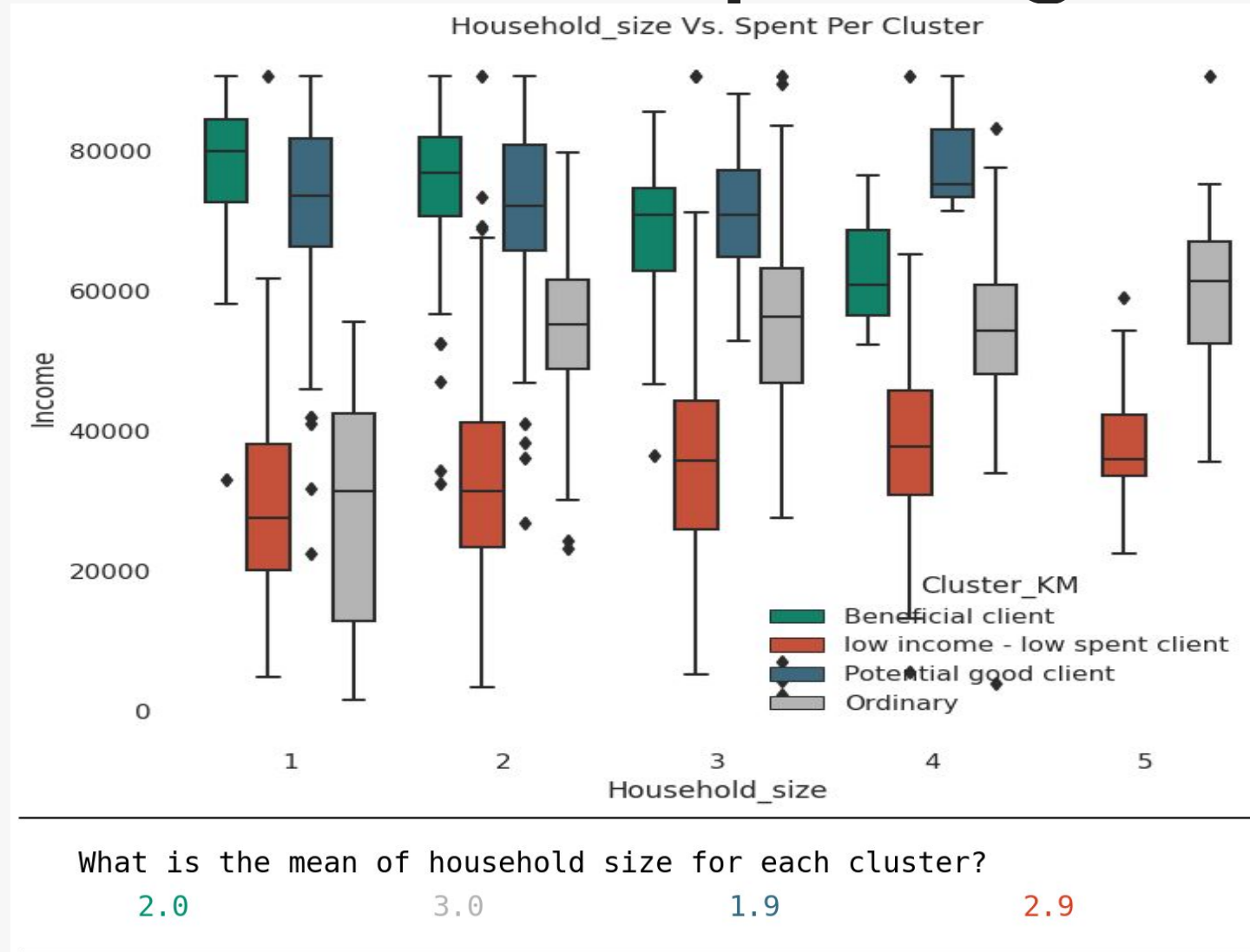
Customer Segment Analysis on different products

Potential good client
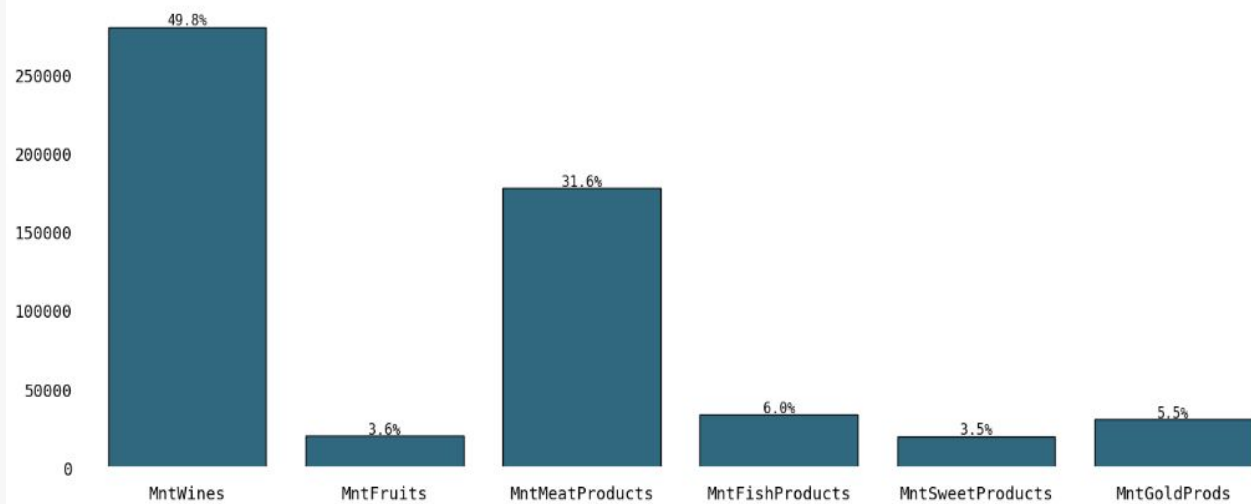
Beneficial client
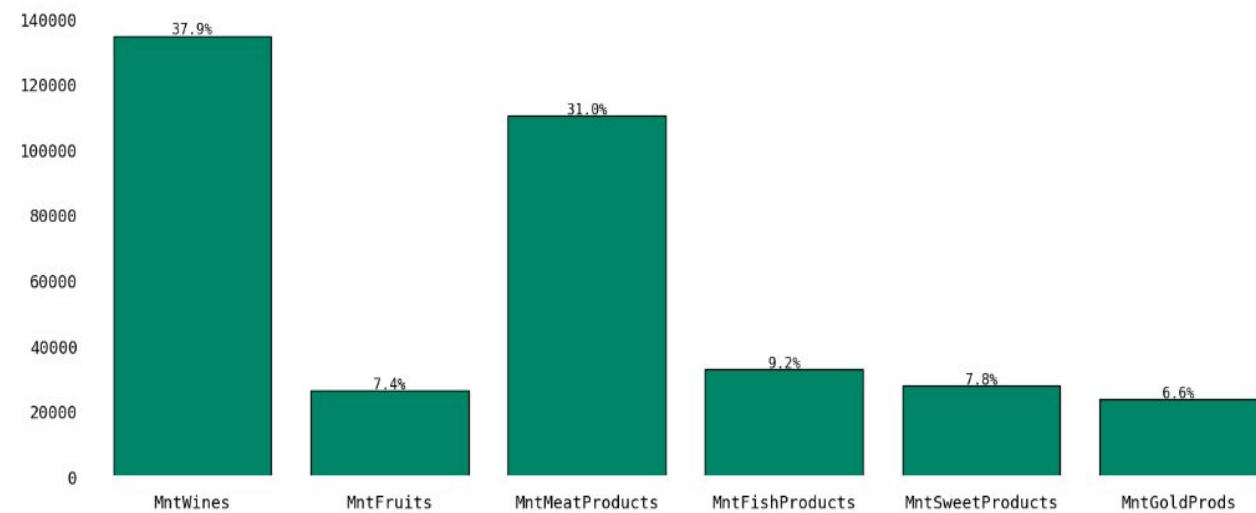
low income - low spent client
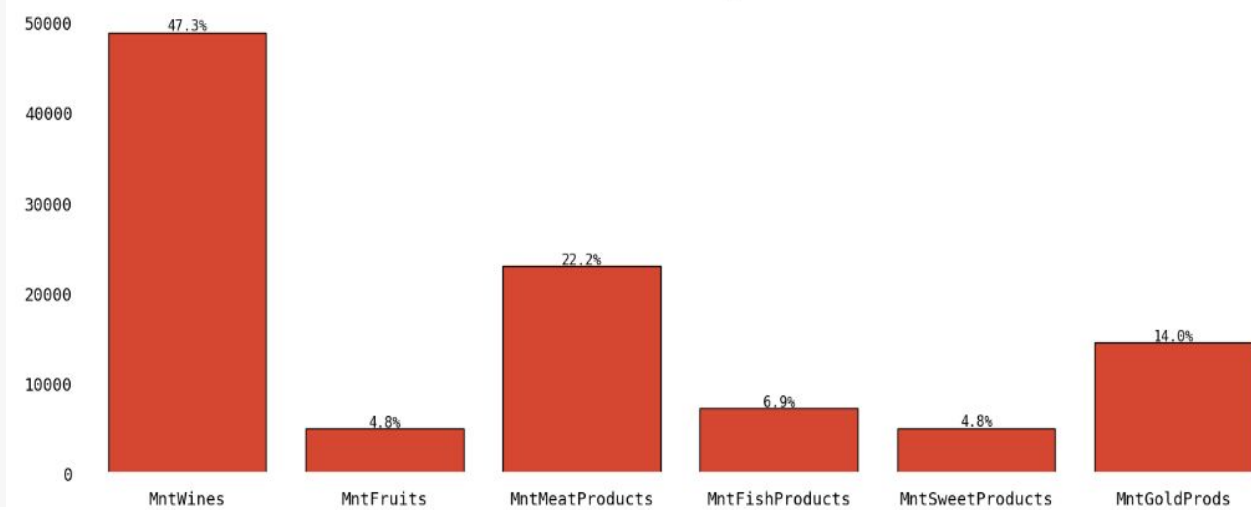
Ordinary
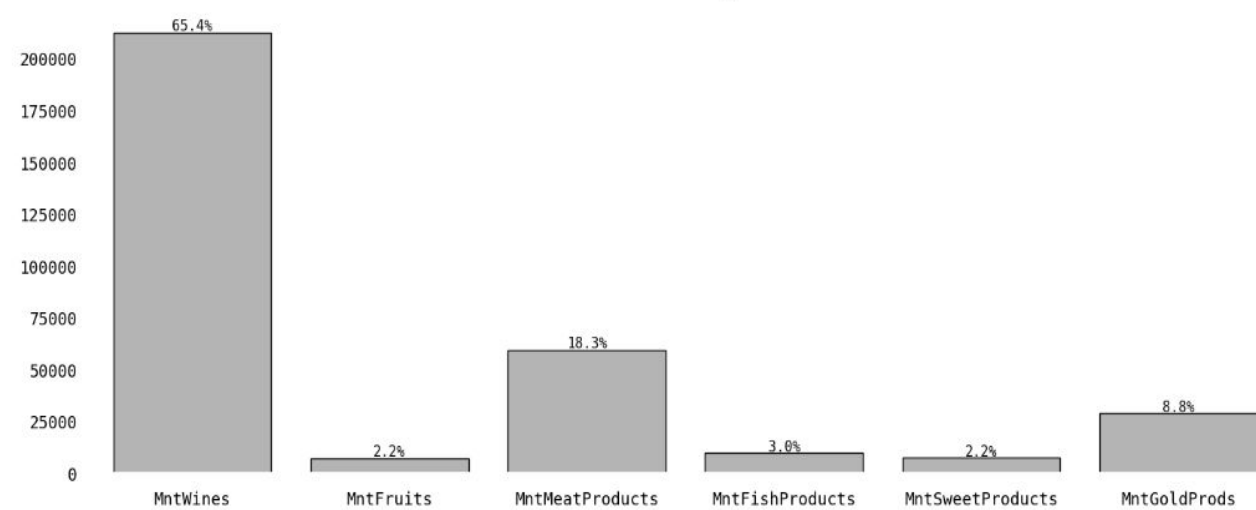
Customer Segment Analysis based on Complains

Who most often complains about the service (based on K-means clustering)?

Segment	Value
Beneficial client	2
Ordinary	3
Potential good client	4
low income - low spent client	12
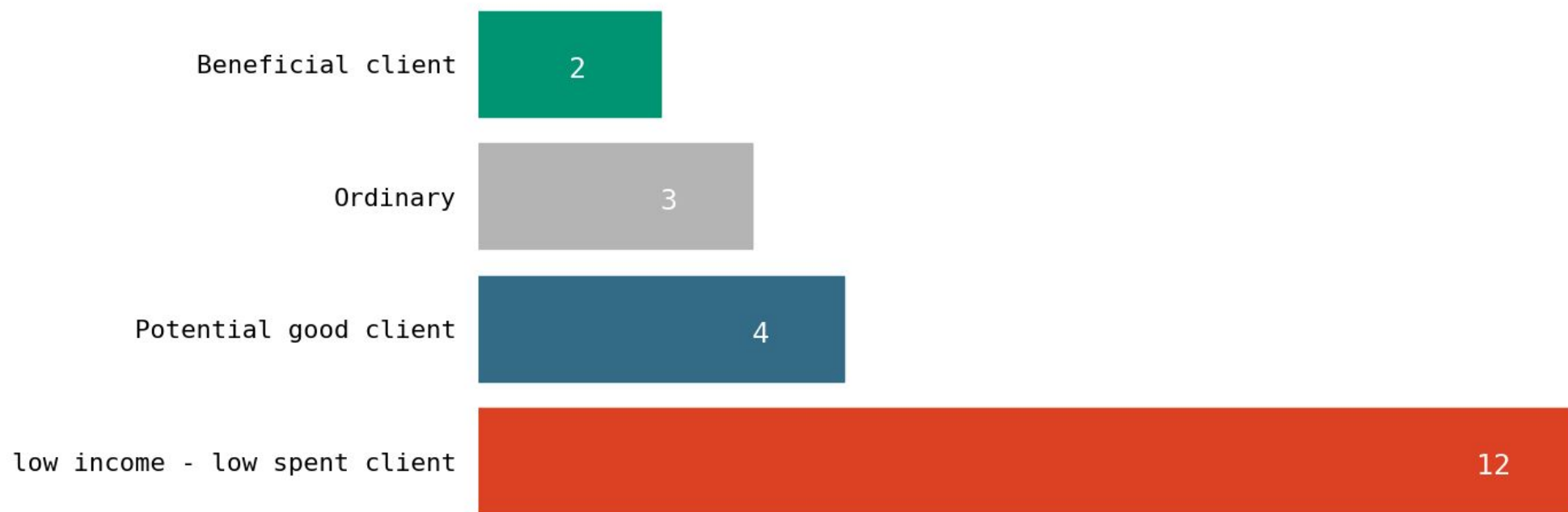
What is average number of complains for each cluster?

| 0.008 | 0.006 | 0.009 | 0.012 |

Customer Segment Analysis based on the number of deal purchases

Who most often Number of deals Purchases about the service (based on the K-means clustering)?

Segment	Value
Beneficial client	390
Ordinary	2180
Potential good client	673
low income - low spent client	1906
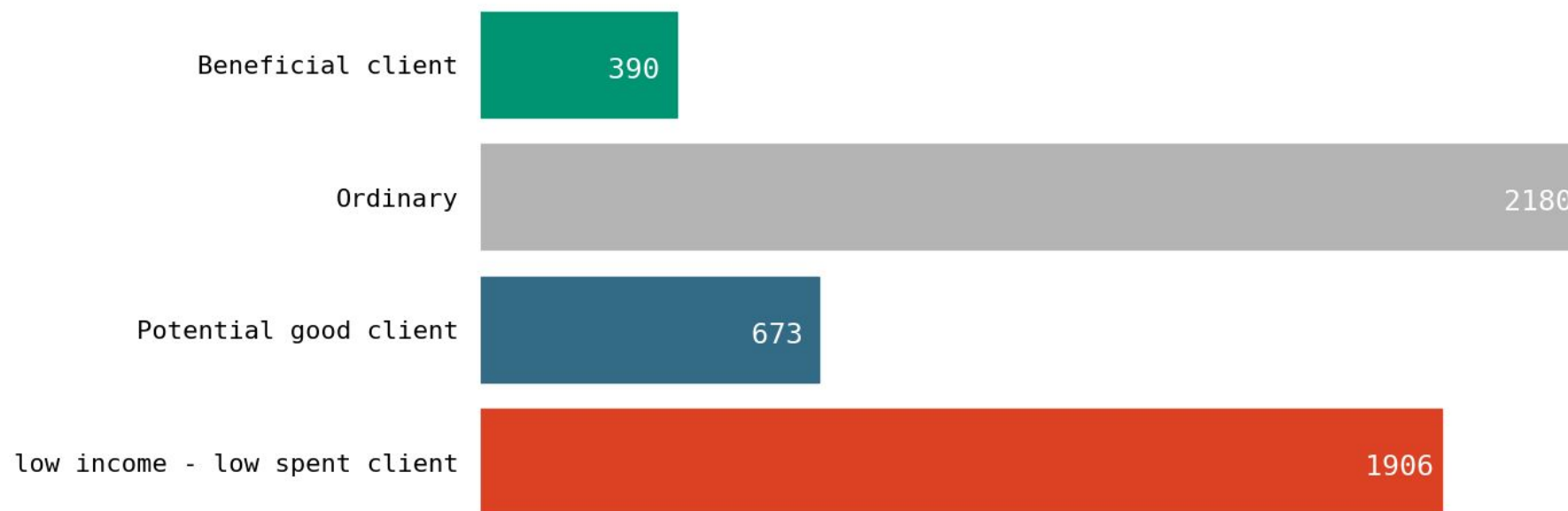
What is the average number of deals that accepted by for each cluster?

1.6   4.4   1.5   1.9

voodoo

Customer Segment Analysis based on the number of web purchases

Who most often purchased by web (based on K-means clustering)?

Segment	Value
Beneficial client	1415
Ordinary	3000
Potential good client	2470
low income - low spent client	2168

What is the percentage of Number of web Purchases customers for each cluster?

5.6	6.2	5.4	2.1

Customer Segment Analysis based on the number of store purchases

Who most often purchased by store (based on K-means clustering)?

Segment	Value
Beneficial client	2116
Ordinary	3497
Potential good client	3860
low income - low spent client	3382

What is the percentage of Number of store Purchases customers for each cluster?

8.4	7.2	8.5	3.3

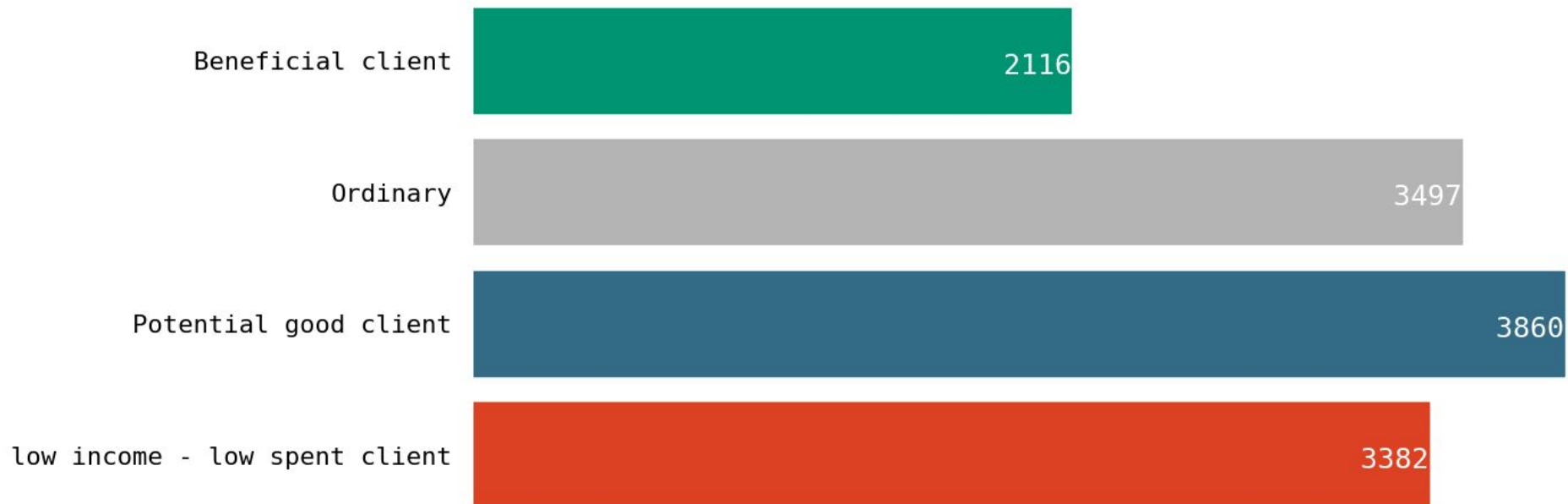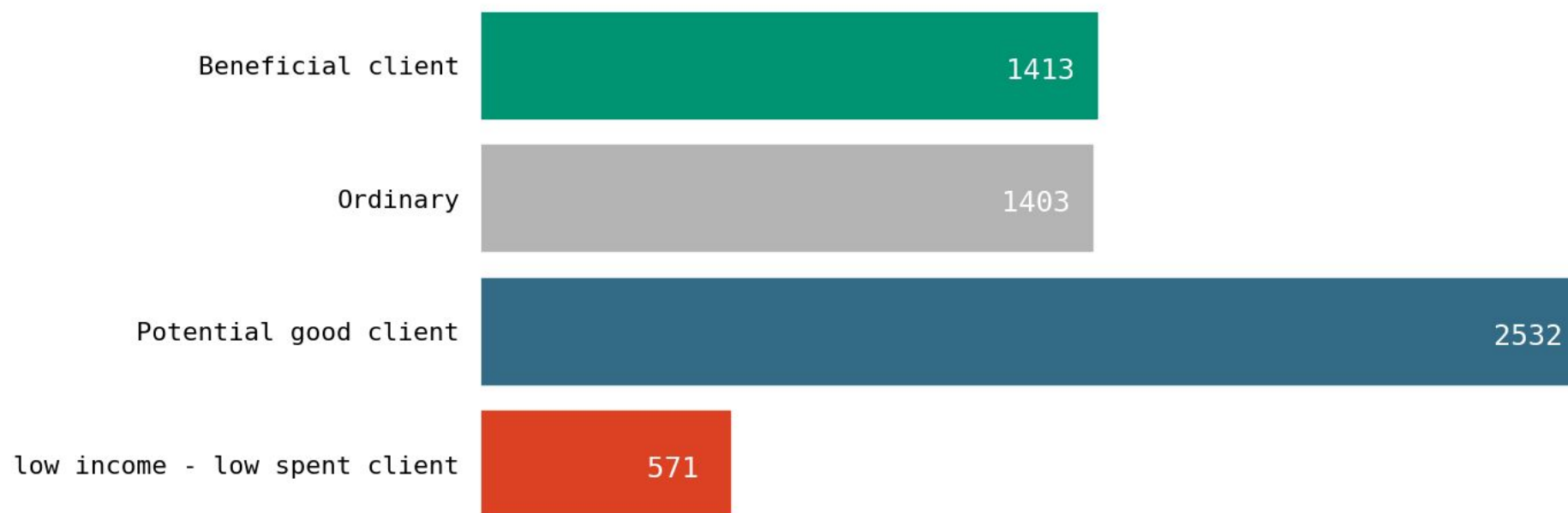# **Customer Segment Analysis** based on the number of catalog purchases

Who most often purchased by catalog (based on K-means clustering)?

Segment	Value
Beneficial client	1413
Ordinary	1403
Potential good client	2532
low income - low spent client	571

What is the percentage of Number of Catalog Purchases customers for each cluster?

5.6   2.9   5.6   0.6

The **customer behaviour**

Beneficial clients

- **High** Income: High purchasing power.
- **High** Spending: Known for high spending habits.
- Average Household Size: Average **household size of 2.**
- Education Level: Majority have **graduate/postgraduate** education.
- Parenthood Status: Most are **not parents,** may impact spending habits.

Potential good clients

- **High** Income: High purchasing power.
- **Average** Spending: Spending higher than ordinary people, but lower than beneficial clients.
- Household Size: **Mostly 2 and 3.**
- Purchase Preference: Tend to purchase **in store**.
- Education Level: Mostly **graduate/postgraduate** education.
- Parenthood Status: Mostly **not parents.**

The **customer behaviour**

voodoo

Ordinary clients

- **Low** income
- **Average** spending
- mostly their number of household size is **2 and 3.**
- tend to buy products by **web**.
- They mostly are **graduate or postgraduate** in education level.
- Mostly they are **parent**.
- seeking for **discounts**

Low-income low-spending clients

- **Low** income
- **Low** spending
- The average number of household size is **3**.
- tend to buy products **in store.**
- They are included people with **all level of education.**
- **Mostly parents.**