Customer Segmentation

Unsupervised Learning Final

Name: Alyson Riley

Date: Spring 1-2024

Project Overview

Customer segmentation analysis is an important tool for businesses to align their strategy and improve their targets for current and future customers

Problem Statement

Performing customer segmentation is the first step in the process of creating marketing personas. Targetable user stories from personas allows businesses to tailor the strategy to drive sales.

Primary focus: Mining transactional data and obtaining a segmentation model through:

- Exploratory Data Analysis
- Data Cleaning & Feature Engineering
- Clustering

Related Work

Customer segmentation analysis is a common application of unsupervised learning – lots of examples in the literature

Many methods I will be using are well documented online with many examples spanning applications:

- PCA
- K-means clustering
- Agglomerative clustering
- DBSCAN

For this data:

- Univariate analysis
- Bivariate analysis

Ref: Singh, S. (2021) *Customer-Personality Analysis + Segmentation*. Available at: https://www.kaggle.com/code/sonalisingh1411/customer-personality-analysis-segmentation (Accessed: 18 September 2023).

Project Work

Data Processing

Check data types

Deal with null values

```
 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)
memory usage: 507.6+ KB
```

Project Work

Data Processing

Pair plot of select variables

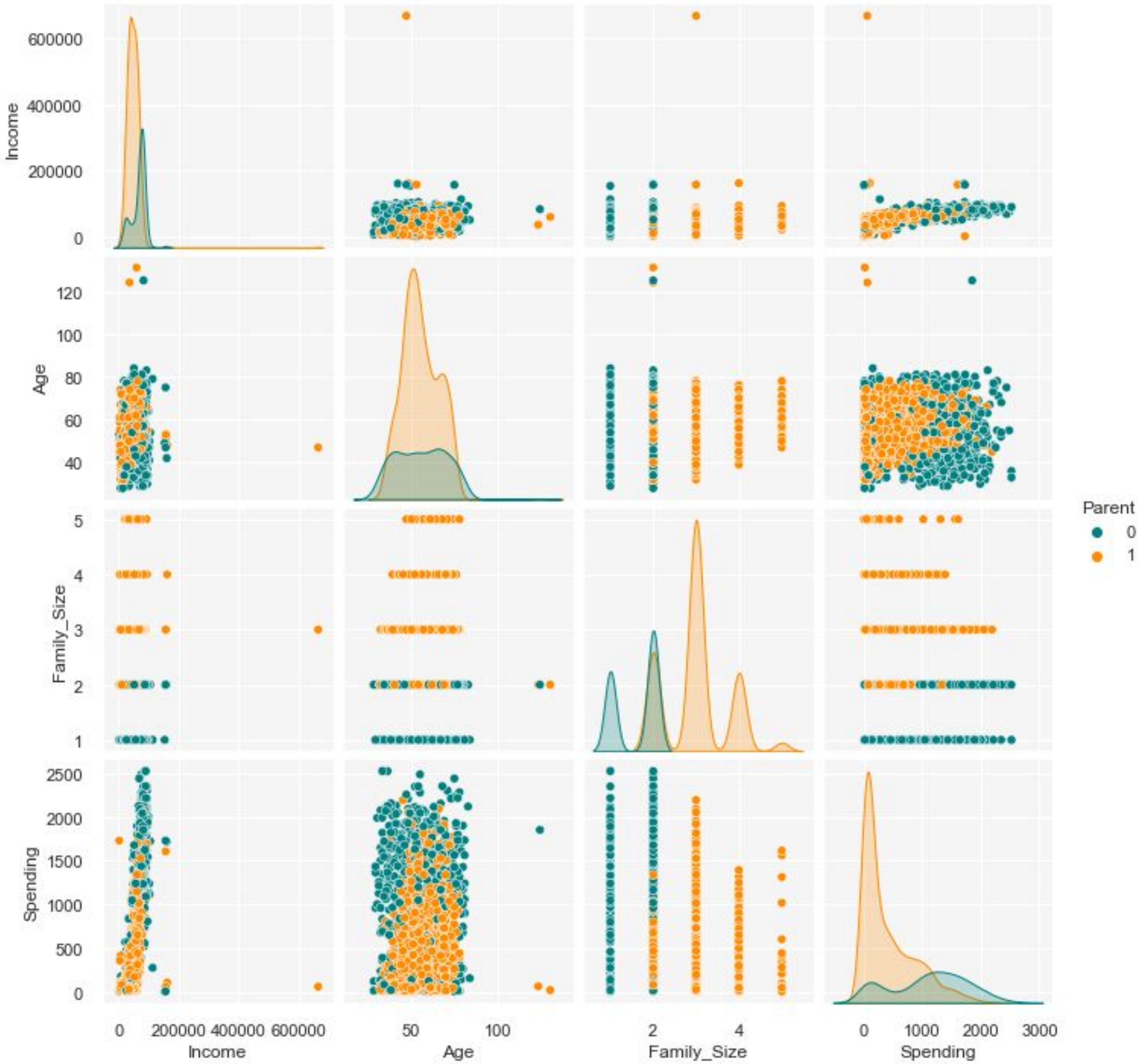
There are some outliers in income & age

Remove the outliers:

```
# drop outliers by setting a cap on Age and Income
df = df[(df['Age'] < 100)]
df = df[(df['Income'] < 250000)]
print("Total number of data points after dropping outliers:", len(df))
```

✓ 0.0s

```
Total number of data points after dropping outliers: 2236
```



Project Work

Data Processing

Feature engineering

Combine several variables into ones that will be easier to model or more informative

```
# create a variable for age
df['Age'] = 2024-df['Year_Birth']

# create 'Single' and 'Not_Single' from Marital_Status
df['Relationship']=df['Marital_Status'].replace({'Married':'not_single' , 'Together':'not_single' , 'Single':'single' , 'Divorced':'single', 'YOLO':'single' , 'Absurd':'single'

# create 'Children' indicating total children living in the household
df['Children']=df['Kidhome']+df['Teenhome']

# create 'Family_Size' for total members in the householde
df['Family_Size'] = df['Relationship'].replace({'single': 1, 'not_single':2})+ df['Children']

# create 'Parent' pertaining parenthood
df['Parent'] = np.where(df.Children> 0, 1, 0)

# create 'Spending' indicating total spent on different categories
df['Spending'] = df['MntWines'] + df['MntFruits'] + df['MntMeatProducts'] + df['MntFishProducts'] + df['MntSweetProducts'] + df['MntGoldProds']

# create 'Num_Purchases' indicating total purchases across locations
df['Num_Purchases'] = df['NumWebPurchases'] + df['NumCatalogPurchases'] + df['NumStorePurchases'] + df['NumDealsPurchases']

# create 'AcceptedCmp' totaling the number of accepted promotionals
df['AcceptedCmp'] = df['AcceptedCmp1'] + df['AcceptedCmp2'] + df['AcceptedCmp3'] + df['AcceptedCmp4'] + df['AcceptedCmp5'] + df['Response']

# change the names of the levels of education
df['Education']=df['Education'].replace({'Basic':'Undergraduate','2n Cycle':'Undergraduate', 'Graduation':'Graduate', 'Master':'Postgraduate', 'PhD':'Postgraduate'})
```

Project Work

Data Processing

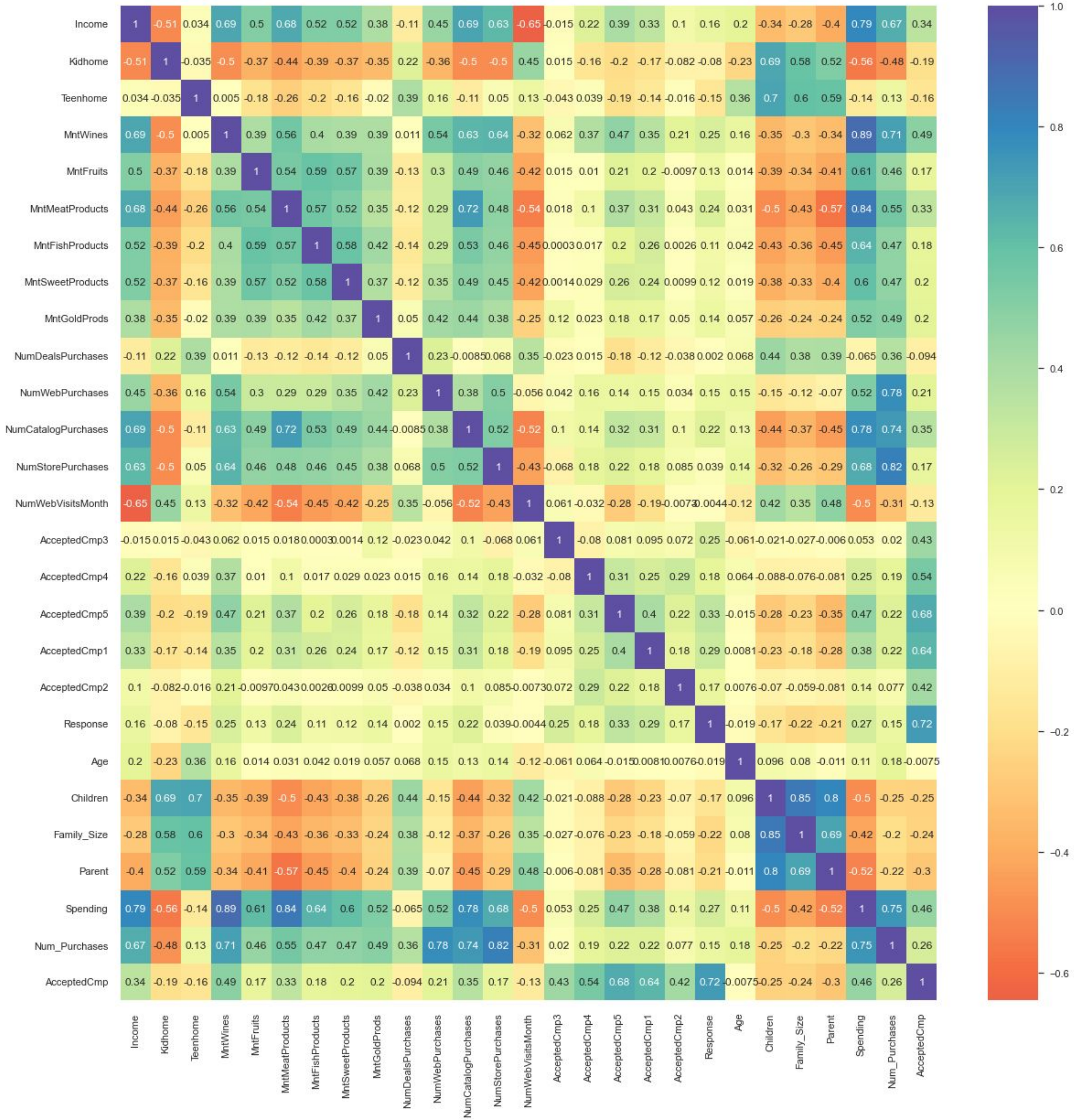
Correlation matrix

Amount spent & income

Amount spent & wines

Amount spent & meats

***Wines** & **meat** might be more expensive than other categories*



Project Work

Data Processing

Principal Component Analysis

- Make all data numeric
- Scale data
- Create a subset of the data removing the promotional and deal data
- Perform PCA

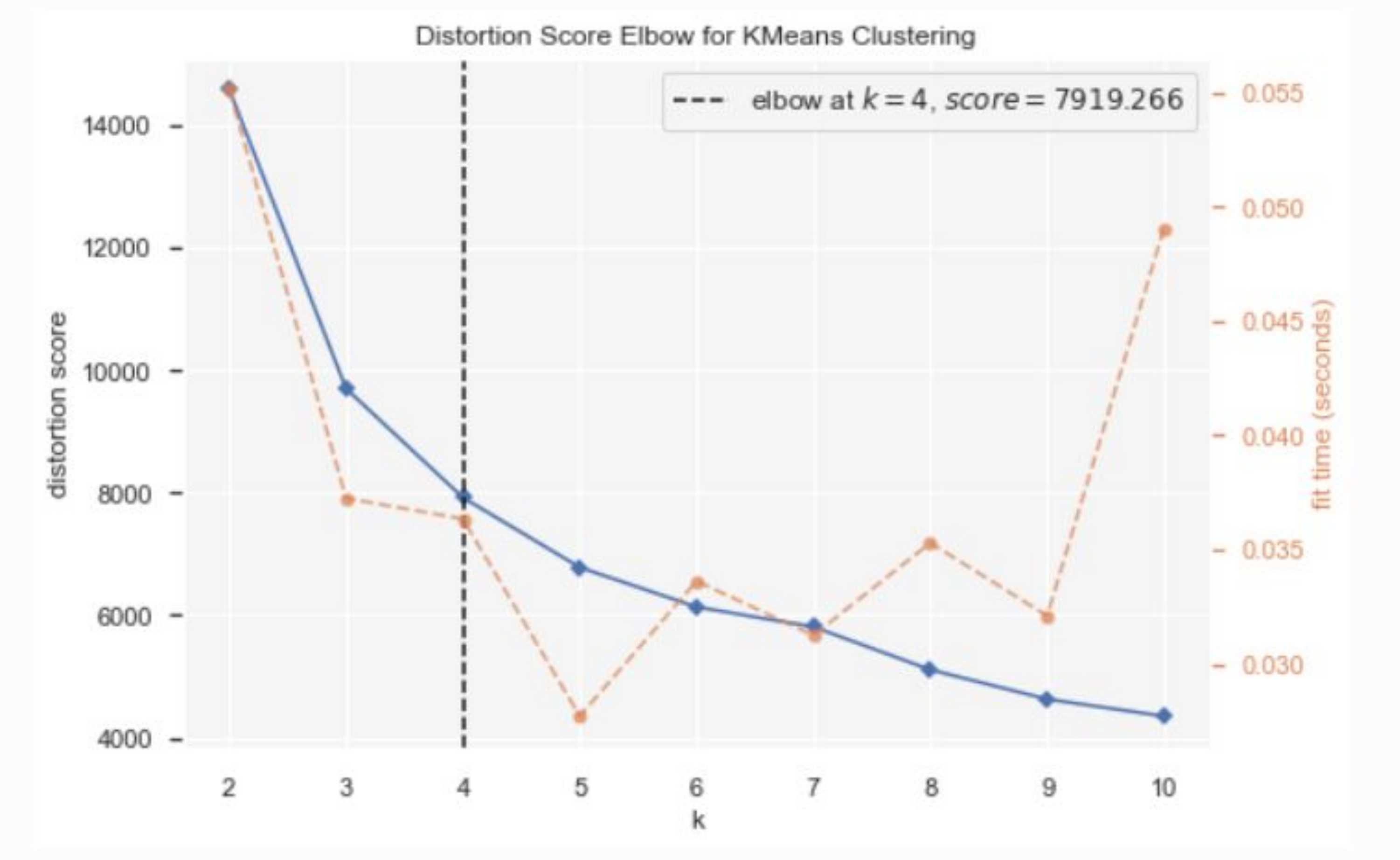
PCA reduces the dimensionality of the dataset by projecting it onto a lower-dimensional subspace. This subspace is defined by the principal components, which are orthogonal to each other and capture the maximum variance in the data

Project Work

Data Processing

Elbow test to determine the best number of clusters to use

4 clusters determined to be best



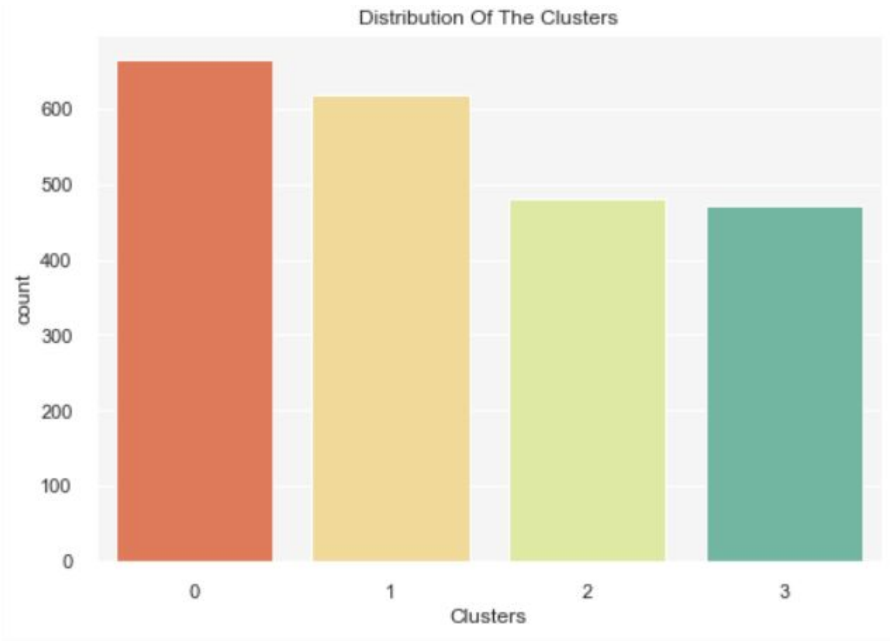
Results

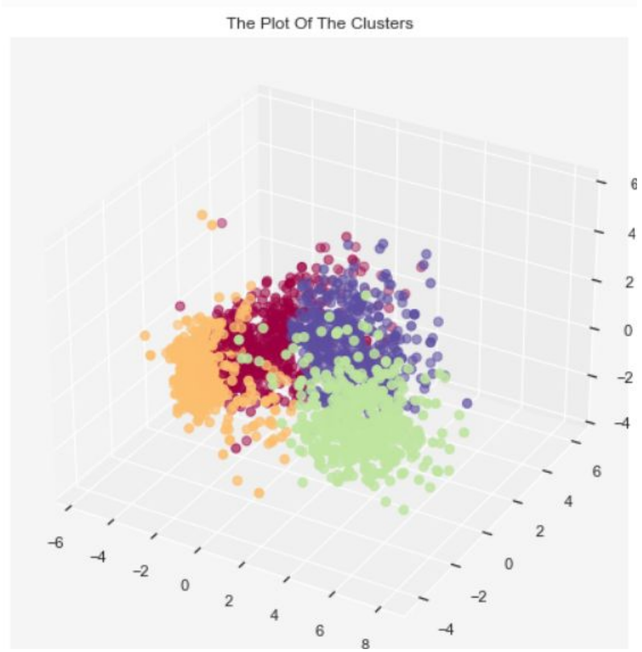
Agglomerative Clustering

4 clusters used – determined by elbow test

Evenly distributed

Clusters appear to have high intra-cluster similarity and low inter-cluster similarity

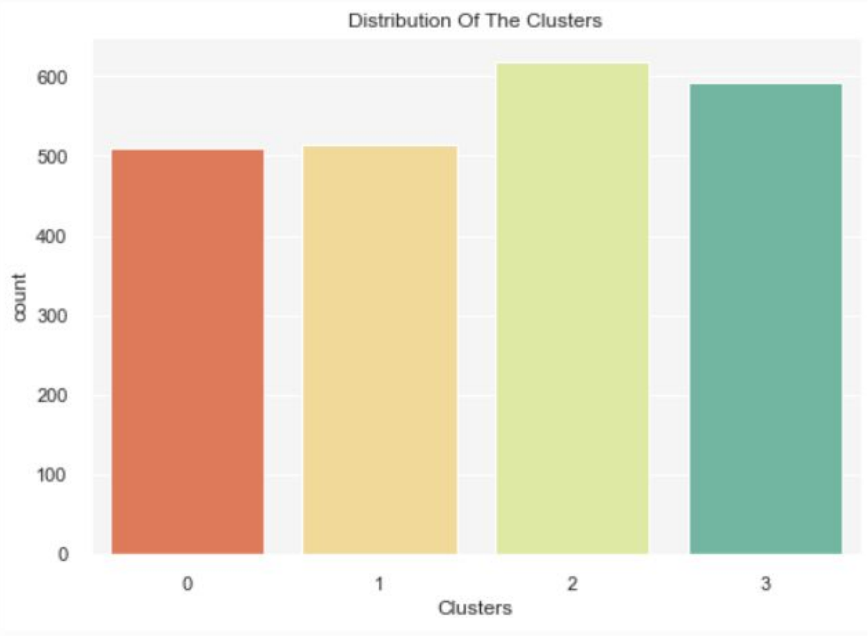


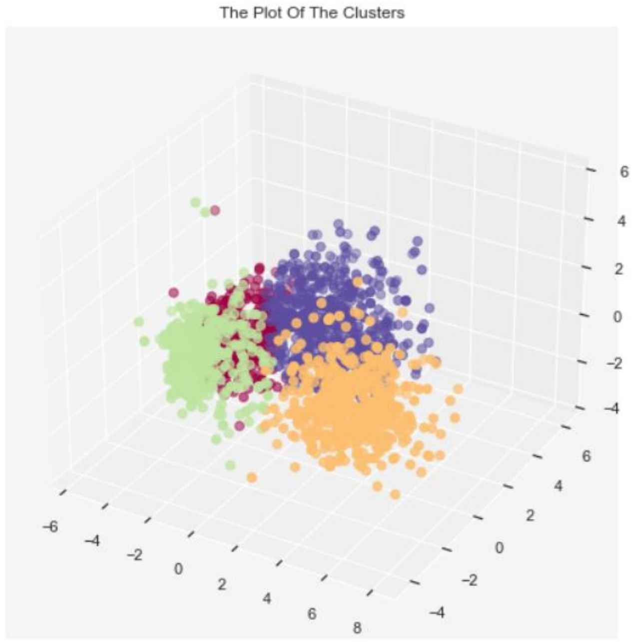


K-Means Clustering

Clusters look evenly distributed

Clusters appear to have high intra-cluster similarity and low inter-cluster similarity



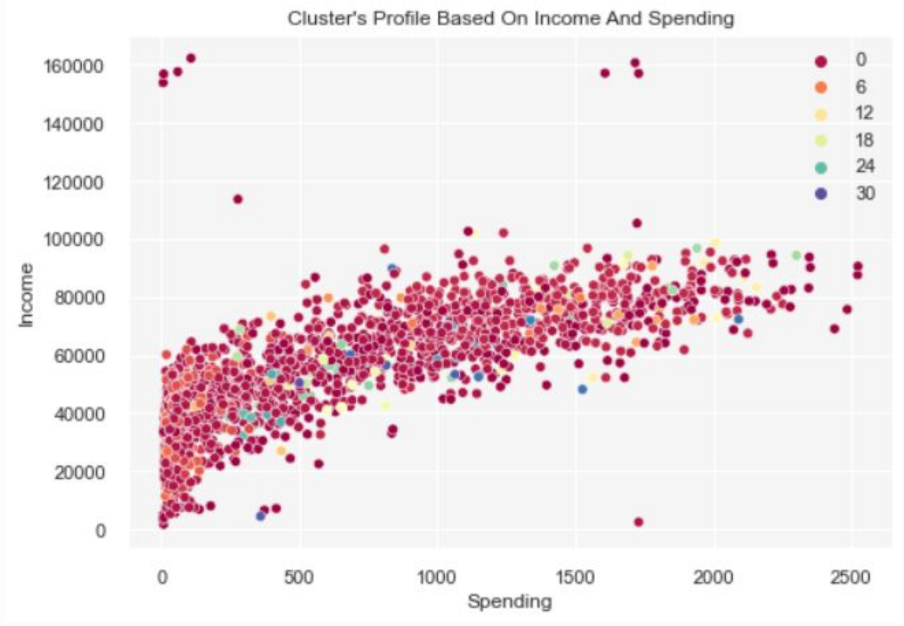


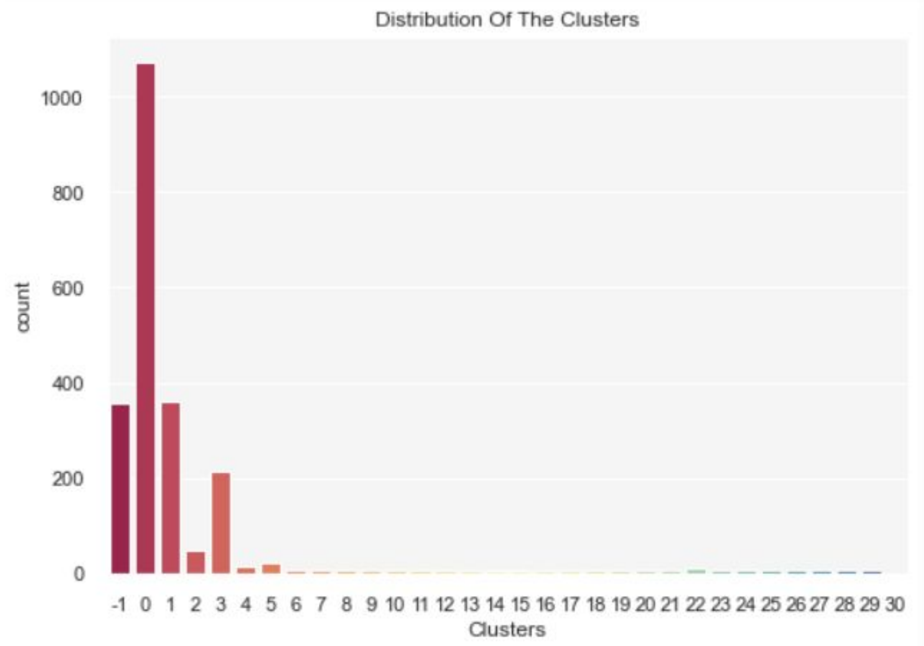
Results

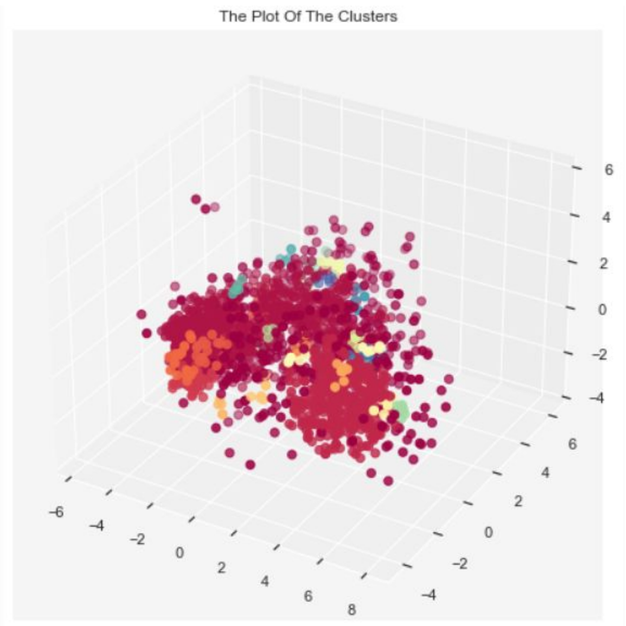
Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Not an even distribution of clusters

Will not proceed with investigating these clusters



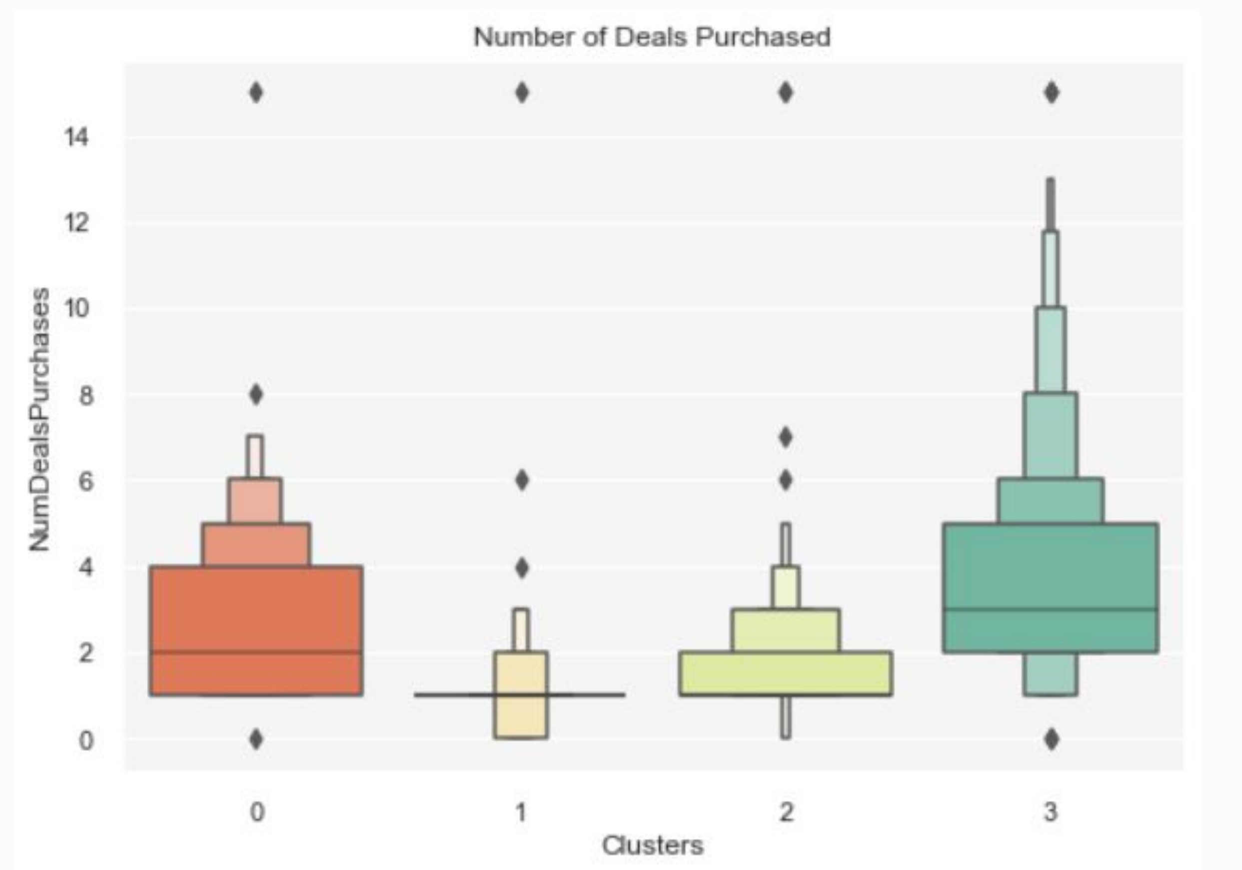


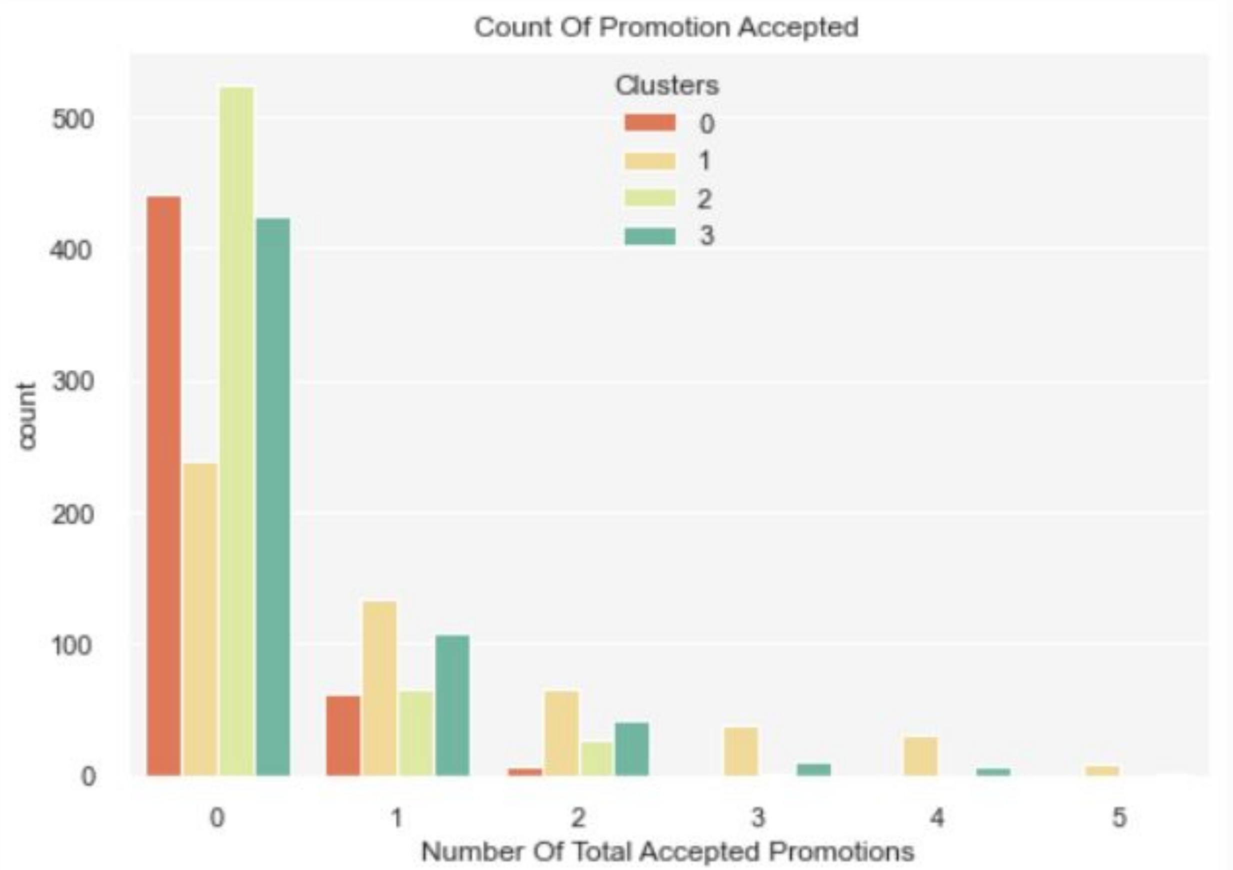


Similarity Metrics

Follow up performed on K-Means Clustering results

Explore distribution of accepted promotions across clusters



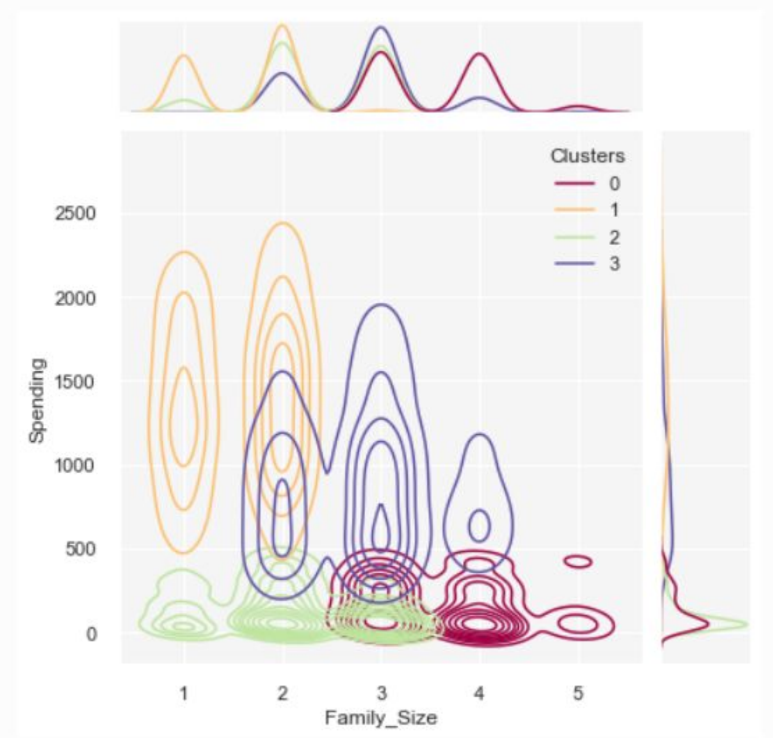


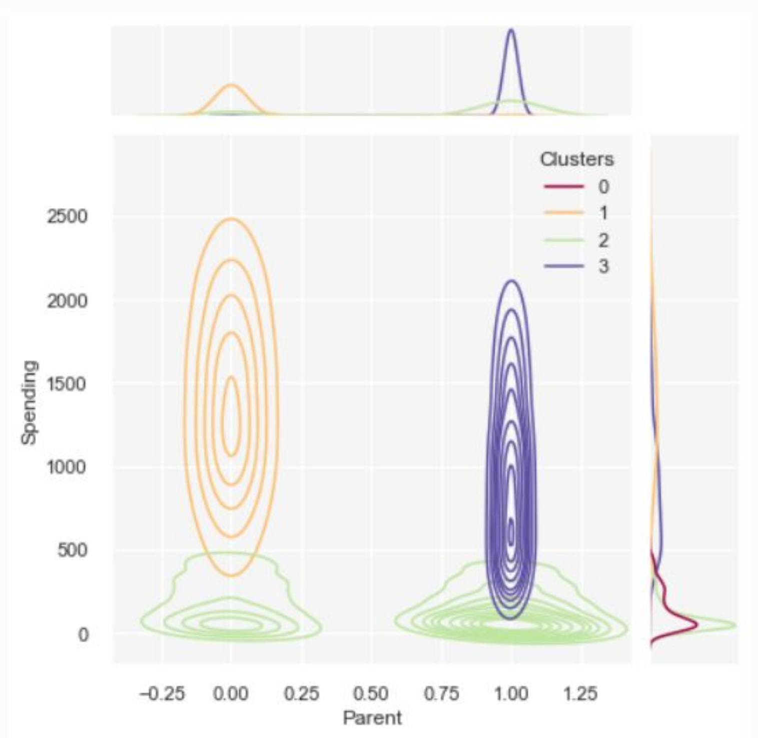
Evaluation

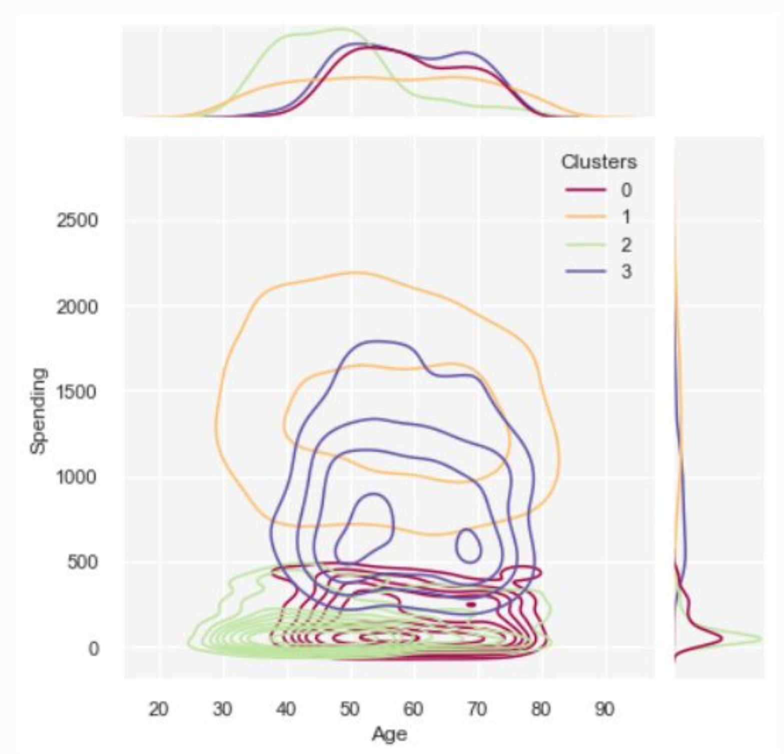
KDE Plots

Plotted several variables vs spending to find relationships and patterns

Some patterns quite clear, others distributed across clusters







Evaluation

Cluster 0

- Is a parent
- Between 2-5 members in the household
- Married and unmarried – some single parents
- Most have a teenager in the home
- Skew older

Cluster 1

- Not a parent
- 1 or 2 people in the household
- Married and unmarried
- Higher income
- Span all ages

Cluster 2

- Majority are parents
- 1-3 members in the household
- Skew younger
- Most have younger children (not teens)

Cluster 3

- Is a parent
- Between 2-5 members in the household
- Majority have a teenager
- Lower income group