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 data mining – 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
- K-means clustering

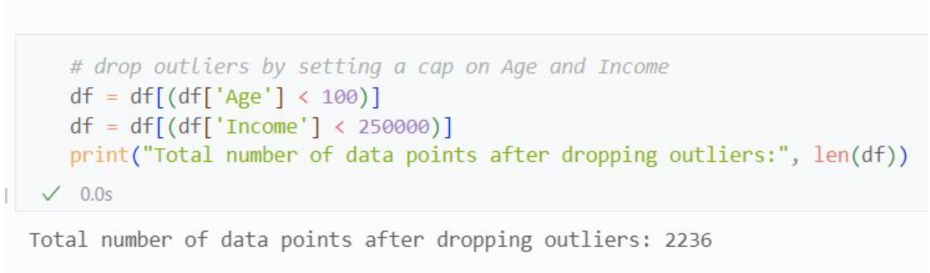
Ref: Singh, S. (2021) *Customer-Personality Analysis + Segmentation*. Available at: https://www.kaggle.com/code/sonalisingh1411/customer-personality-analysis-segmentation (Accessed: 18 September 2023).

Data Processing

Pair plot of select variables

There are some outliers in income & age

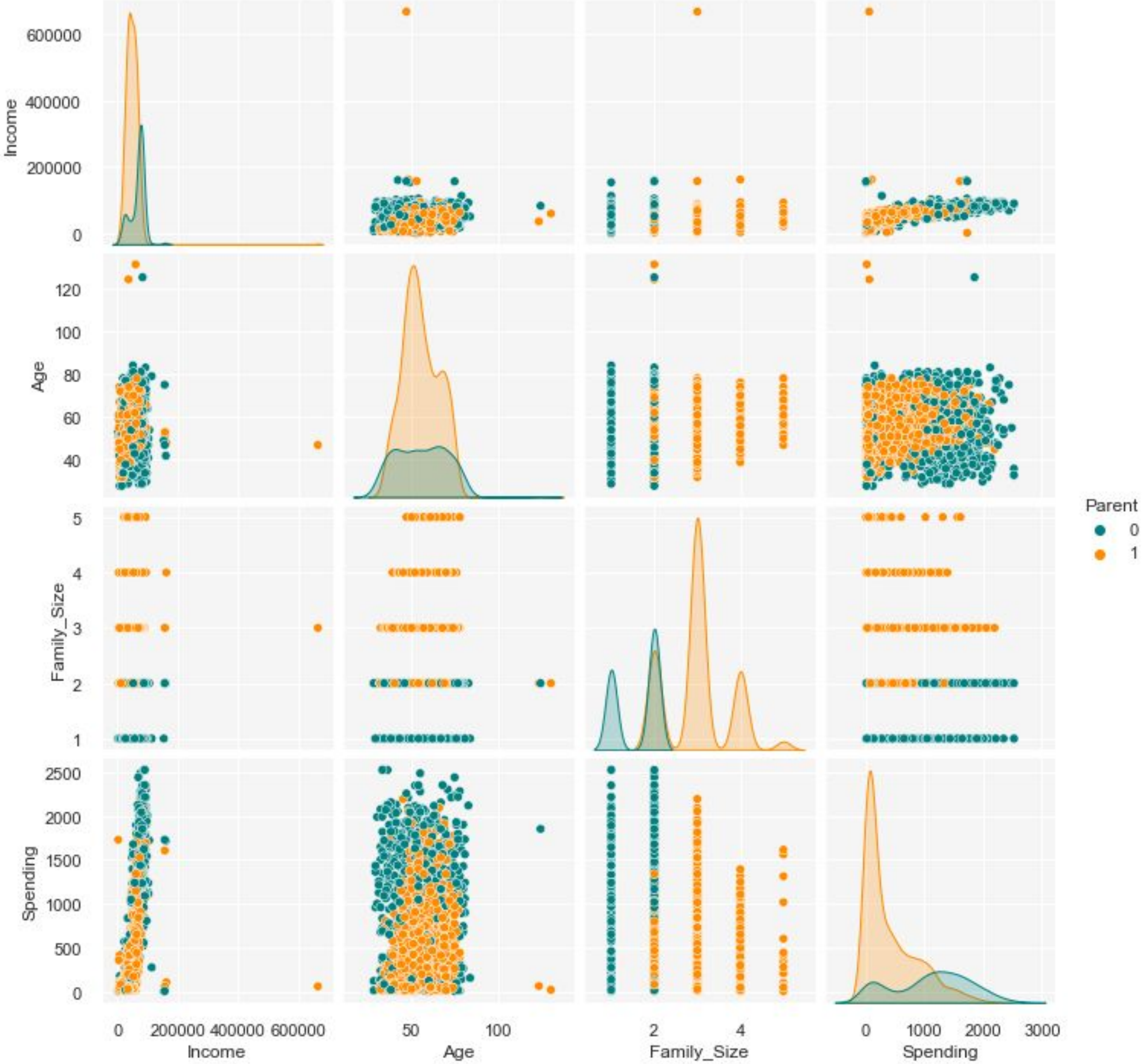
Remove the outliers:



```
# 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
```

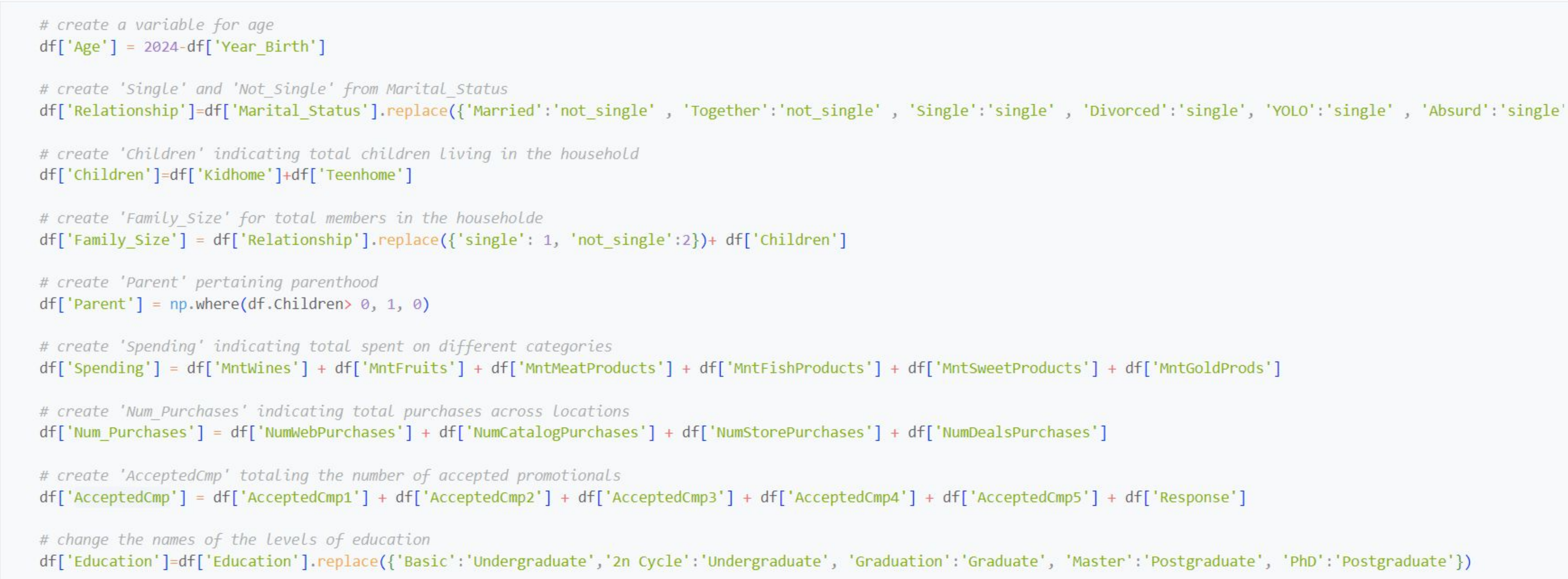


Proposed Work

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'})
```

Proposed Work

Data Processing

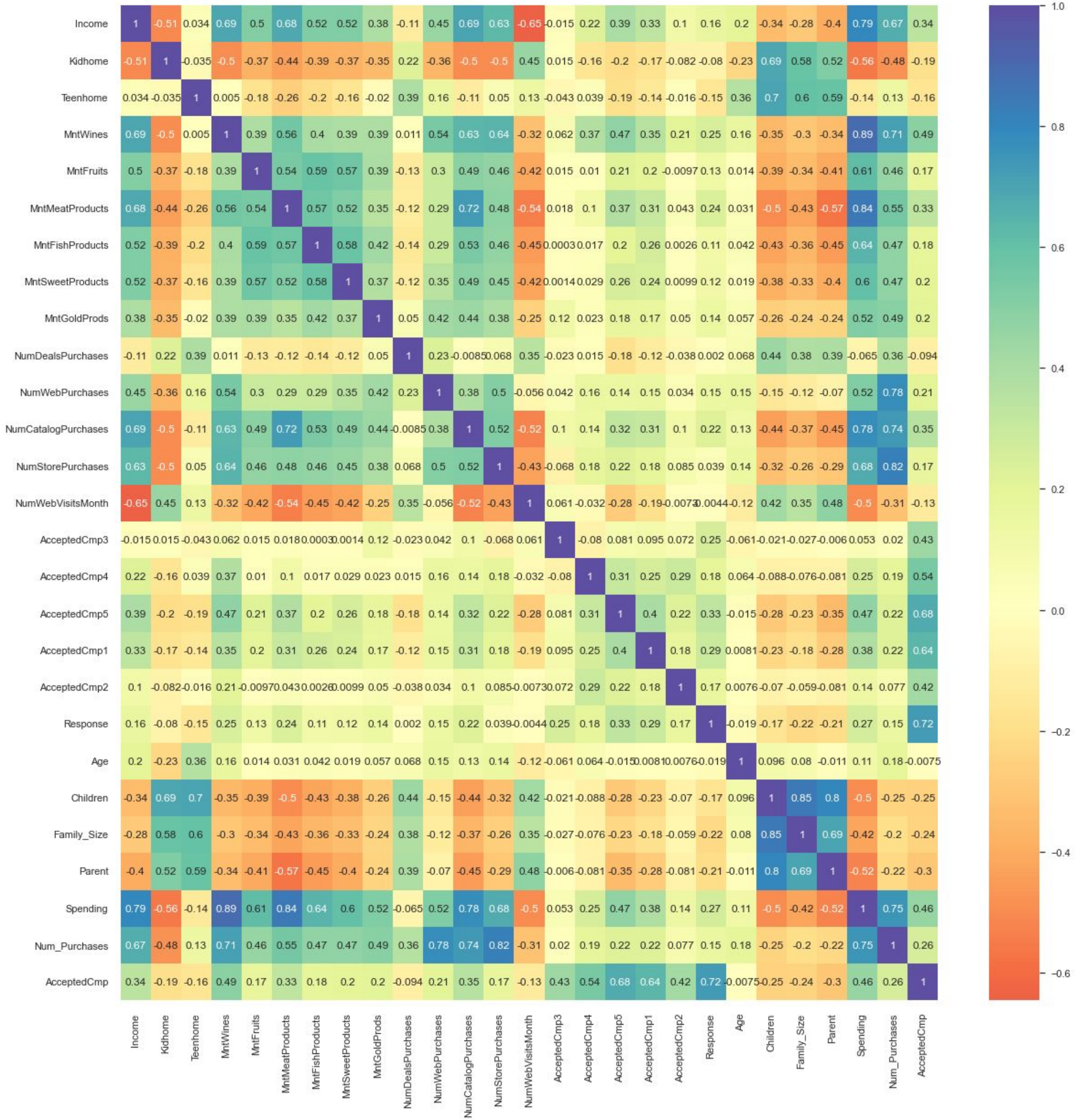
Correlation matrix

Amount spent & income

Amount spent & wines

Amount spent & meats

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



	Income	Kidhome	Teenhome	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Response	Age	Children	Family_Size	Parent	Spending	Num_Purchases	AcceptedCmp
Income	1	-0.51	0.034	0.69	0.5	0.68	0.52	0.52	0.38	-0.11	0.45	0.69	0.63	-0.65	-0.015	0.22	0.39	0.33	0.1	0.16	0.2	-0.34	-0.28	-0.4	0.79	0.67	0.34
Kidhome	-0.51	1	-0.035	-0.5	-0.37	-0.44	-0.39	-0.37	-0.35	0.22	-0.36	-0.5	-0.5	0.45	0.015	-0.16	-0.2	-0.17	-0.082	-0.08	-0.23	0.69	0.58	0.52	-0.56	-0.48	-0.19
Teenhome	0.034	-0.035	1	0.005	-0.18	-0.26	-0.2	-0.16	-0.02	0.39	0.16	-0.11	0.05	0.13	-0.043	0.039	-0.19	-0.14	-0.016	-0.15	0.36	0.7	0.6	0.59	-0.14	0.13	-0.16
MntWines	0.69	-0.5	0.005	1	0.39	0.56	0.4	0.39	0.39	0.011	0.54	0.63	0.64	-0.32	0.062	0.37	0.47	0.35	0.21	0.25	0.16	-0.35	-0.3	-0.34	0.89	0.71	0.49
MntFruits	0.5	-0.37	-0.18	0.39	1	0.54	0.59	0.57	0.39	-0.13	0.3	0.49	0.46	-0.42	0.015	0.01	0.21	0.2	-0.0097	0.13	0.014	-0.39	-0.34	-0.41	0.61	0.46	0.17
MntMeatProducts	0.68	-0.44	-0.26	0.56	0.54	1	0.57	0.52	0.35	-0.12	0.29	0.72	0.48	-0.54	0.018	0.1	0.37	0.31	0.043	0.24	0.031	-0.5	-0.43	-0.57	0.84	0.55	0.33
MntFishProducts	0.52	-0.39	-0.2	0.4	0.59	0.57	1	0.58	0.42	-0.14	0.29	0.53	0.46	-0.45	0.0003	0.017	0.2	0.26	0.0026	0.11	0.042	-0.43	-0.36	-0.45	0.64	0.47	0.18
MntSweetProducts	0.52	-0.37	-0.16	0.39	0.57	0.52	0.58	1	0.37	-0.12	0.35	0.49	0.45	-0.42	0.0014	0.029	0.26	0.24	0.0099	0.12	0.019	-0.38	-0.33	-0.4	0.6	0.47	0.2
MntGoldProds	0.38	-0.35	-0.02	0.39	0.39	0.35	0.42	0.37	1	0.05	0.42	0.44	0.38	-0.25	0.12	0.023	0.18	0.17	0.05	0.14	0.057	-0.26	-0.24	-0.24	0.52	0.49	0.2
NumDealsPurchases	-0.11	0.22	0.39	0.011	-0.13	-0.12	-0.14	-0.12	0.05	1	0.23	-0.0085	0.068	0.35	-0.023	0.015	-0.18	-0.12	-0.038	0.002	0.068	0.44	0.38	0.39	-0.065	0.36	-0.094
NumWebPurchases	0.45	-0.36	0.16	0.54	0.3	0.29	0.29	0.35	0.42	0.23	1	0.38	0.5	-0.056	0.042	0.16	0.14	0.15	0.034	0.15	0.15	-0.15	-0.12	-0.07	0.52	0.78	0.21
NumCatalogPurchases	0.69	-0.5	-0.11	0.63	0.49	0.72	0.53	0.49	0.44	-0.0085	0.38	1	0.52	-0.52	0.1	0.14	0.32	0.31	0.1	0.22	0.13	-0.44	-0.37	-0.45	0.78	0.74	0.35
NumStorePurchases	0.63	-0.5	0.05	0.64	0.46	0.48	0.46	0.45	0.38	0.068	0.5	0.52	1	-0.43	-0.068	0.18	0.22	0.18	0.085	0.039	0.14	-0.32	-0.26	-0.29	0.68	0.82	0.17
NumWebVisitsMonth	-0.65	0.45	0.13	-0.32	-0.42	-0.54	-0.45	-0.42	-0.25	0.35	-0.056	-0.52	-0.43	1	0.061	-0.032	-0.28	-0.19	-0.0073	-0.0044	-0.12	0.42	0.35	0.48	-0.5	-0.31	-0.13
AcceptedCmp3	-0.015	0.015	-0.043	0.062	0.015	0.018	0.0003	0.0014	0.12	-0.023	0.042	0.1	-0.068	0.061	1	-0.08	0.081	0.095	0.072	0.25	-0.061	-0.021	-0.027	-0.006	0.053	0.02	0.43
AcceptedCmp4	0.22	-0.16	0.039	0.37	0.01	0.1	0.017	0.029	0.023	0.015	0.16	0.14	0.18	-0.032	-0.08	1	0.31	0.25	0.29	0.18	0.064	-0.088	-0.076	-0.081	0.25	0.19	0.54
AcceptedCmp5	0.39	-0.2	-0.19	0.47	0.21	0.37	0.2	0.26	0.18	-0.18	0.14	0.32	0.22	-0.28	0.081	0.31	1	0.4	0.22	0.33	-0.015	-0.28	-0.23	-0.35	0.47	0.22	0.68
AcceptedCmp1	0.33	-0.17	-0.14	0.35	0.2	0.31	0.26	0.24	0.17	-0.12	0.15	0.31	0.18	-0.19	0.095	0.25	0.4	1	0.18	0.29	0.0081	-0.23	-0.18	-0.28	0.38	0.22	0.64
AcceptedCmp2	0.1	-0.082	-0.016	0.21	-0.0097	0.043	0.0026	0.0099	0.05	-0.038	0.034	0.1	0.085	-0.0073	0.072	0.29	0.22	0.18	1	0.17	0.0076	-0.07	-0.059	-0.081	0.14	0.077	0.42
Response	0.16	-0.08	-0.15	0.25	0.13	0.24	0.11	0.12	0.14	0.002	0.15	0.22	0.039	-0.0044	0.25	0.18	0.33	0.29	0.17	1	-0.019	-0.17	-0.22	-0.21	0.27	0.15	0.72
Age	0.2	-0.23	0.36	0.16	0.014	0.031	0.042	0.019	0.057	0.068	0.15	0.13	0.14	-0.12	-0.061	0.064	-0.015	0.0081	0.0076	-0.019	1	0.096	0.08	-0.011	0.11	0.18	-0.0075
Children	-0.34	0.69	0.7	-0.35	-0.39	-0.5	-0.43	-0.38	-0.26	0.44	-0.15	-0.44	-0.32	0.42	-0.021	-0.088	-0.28	-0.23	-0.07	-0.17	0.096	1	0.85	0.8	-0.5	-0.25	-0.25
Family_Size	-0.28	0.58	0.6	-0.3	-0.34	-0.43	-0.36	-0.33	-0.24	0.38	-0.12	-0.37	-0.26	0.35	-0.027	-0.076	-0.23	-0.18	-0.059	-0.22	0.08	0.85	1	0.69	-0.42	-0.2	-0.24
Parent	-0.4	0.52	0.59	-0.34	-0.41	-0.57	-0.45	-0.4	-0.24	0.39	-0.07	-0.45	-0.29	0.48	-0.006	-0.081	-0.35	-0.28	-0.081	-0.21	-0.011	0.8	0.69	1	-0.52	-0.22	-0.3
Spending	0.79	-0.56	-0.14	0.89	0.61	0.84	0.64	0.6	0.52	-0.065	0.52	0.78	0.68	-0.5	0.053	0.25	0.47	0.38	0.14	0.27	0.11	-0.5	-0.42	-0.52	1	0.75	0.46
Num_Purchases	0.67	-0.48	0.13	0.71	0.46	0.55	0.47	0.47	0.49	0.36	0.78	0.74	0.82	-0.31	0.02	0.19	0.22	0.22	0.077	0.15	0.18	-0.25	-0.2	-0.22	0.75	1	0.26
AcceptedCmp	0.34	-0.19	-0.16	0.49	0.17	0.33	0.18	0.2	0.2	-0.094	0.21	0.35	0.17	-0.13	0.43	0.54	0.68	0.64	0.42	0.72	-0.0075	-0.25	-0.24	-0.3	0.46	0.26	1

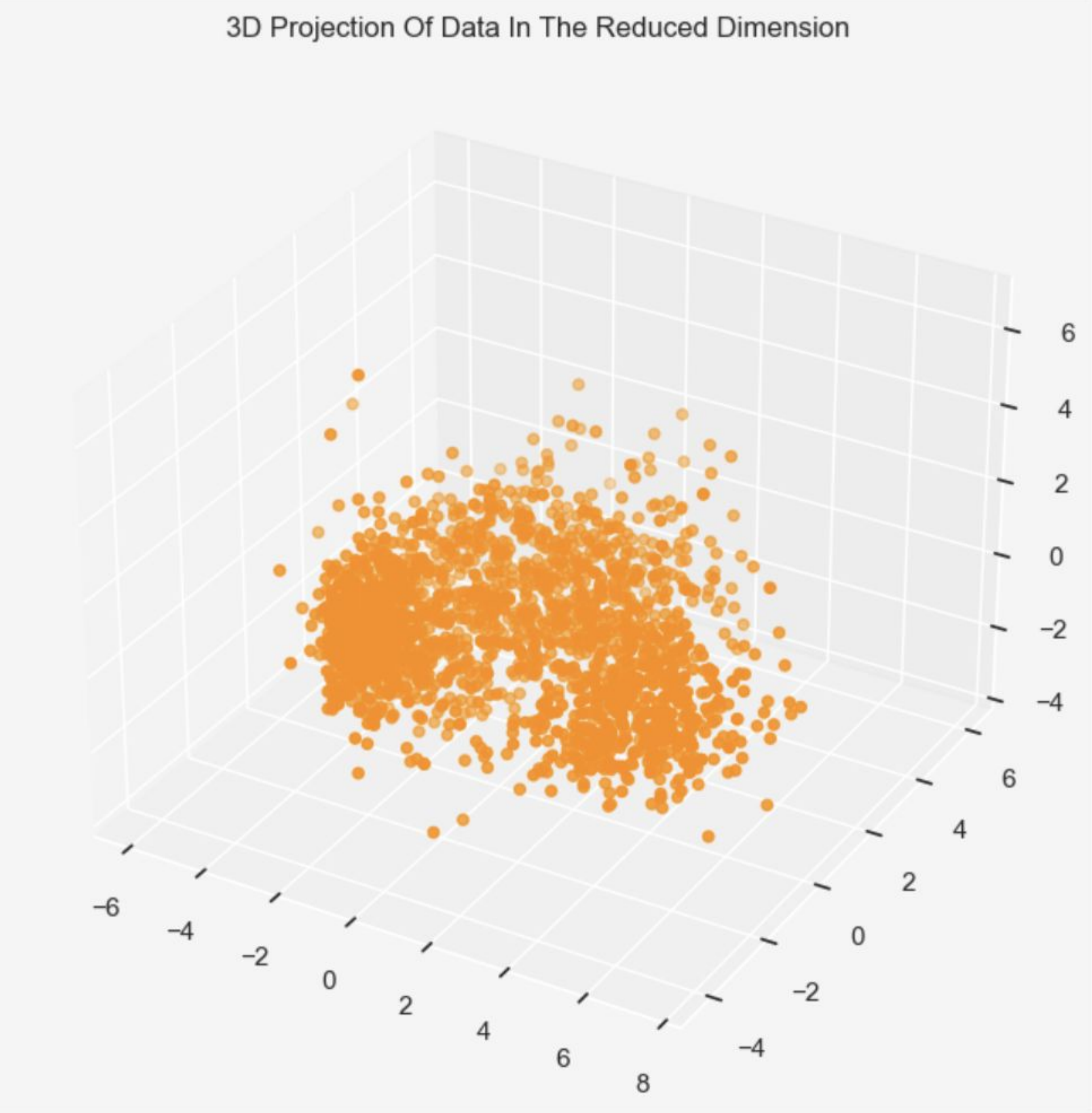
Proposed Work

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

Use the PCA columns to plot the clusters against

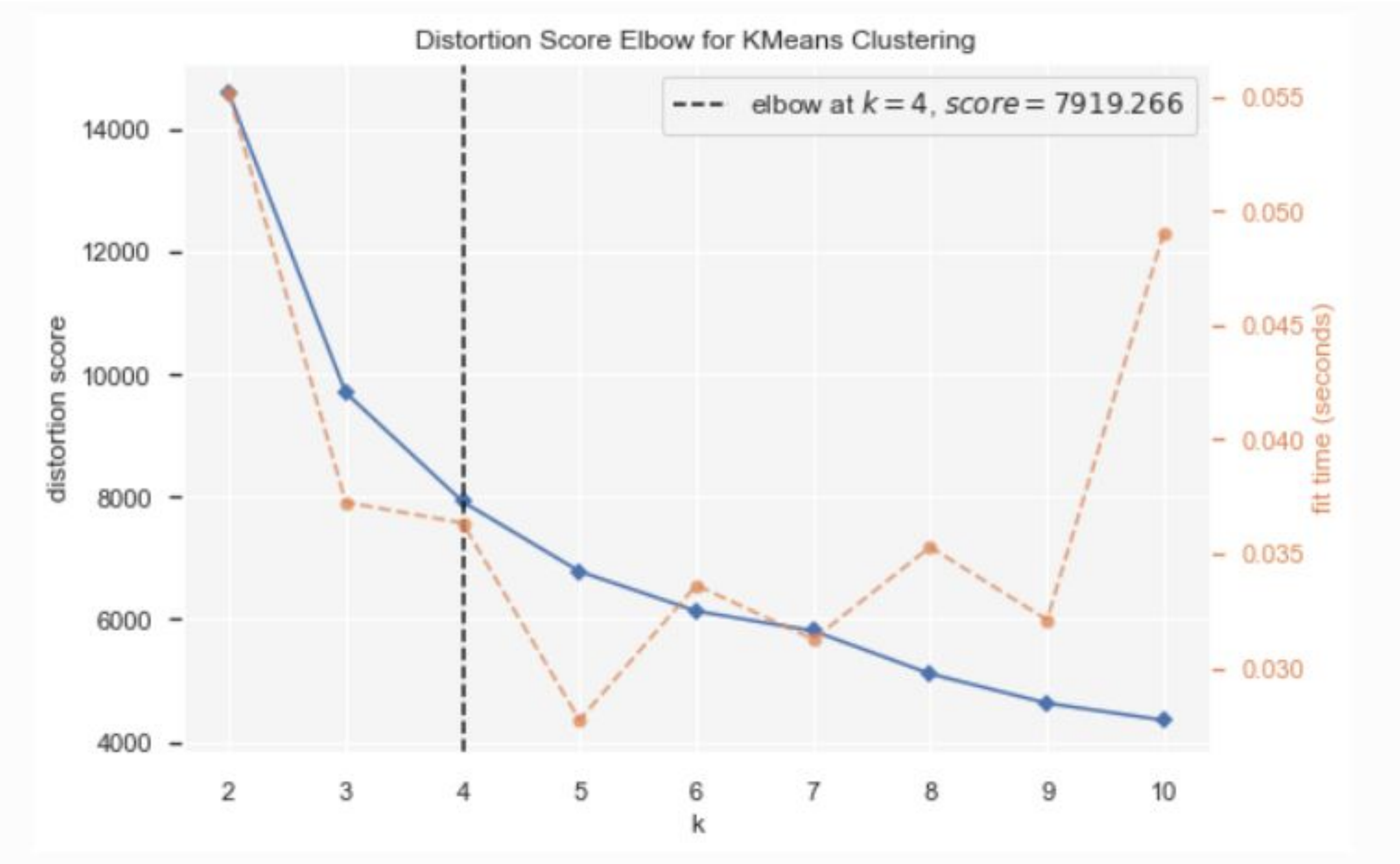


Proposed Work

Data Processing

Elbow test to determine the best number of clusters to use

4 clusters determined to be best



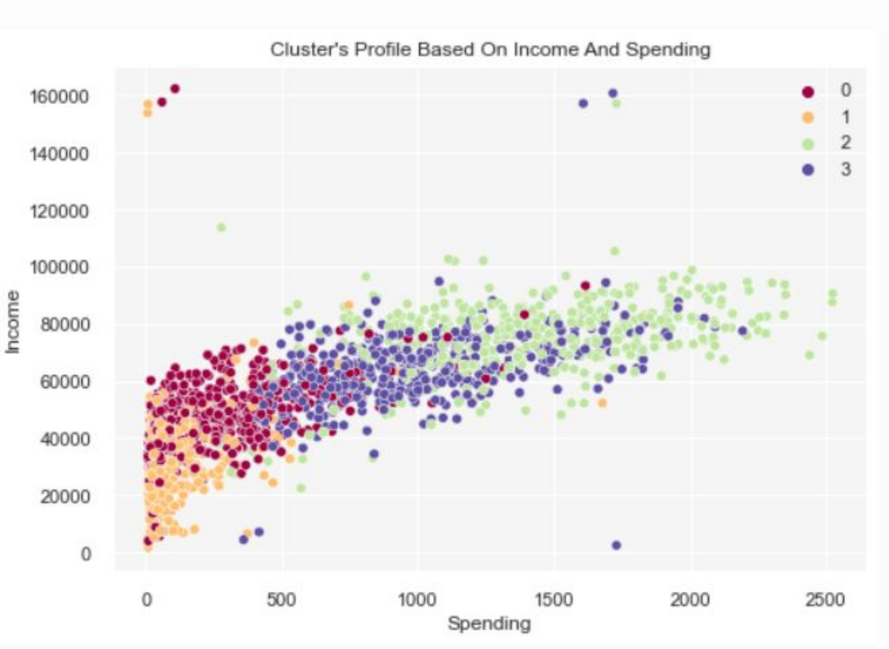
Proposed Work

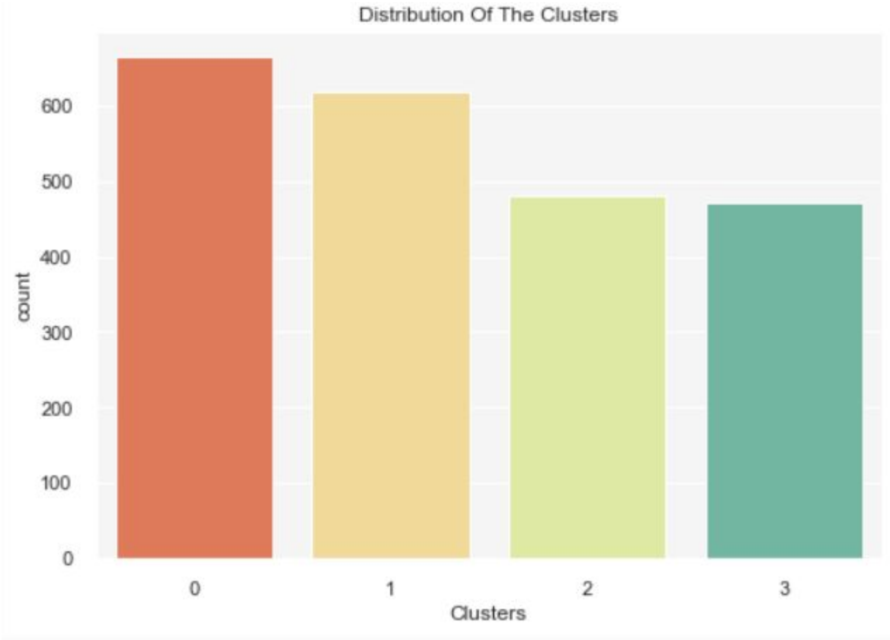
Agglomerative Clustering

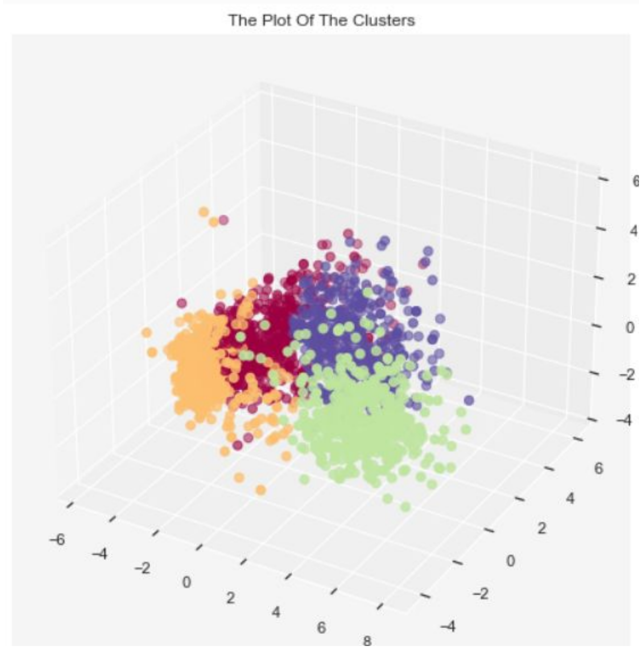
4 clusters used – determined by elbow test

Evenly distributed

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





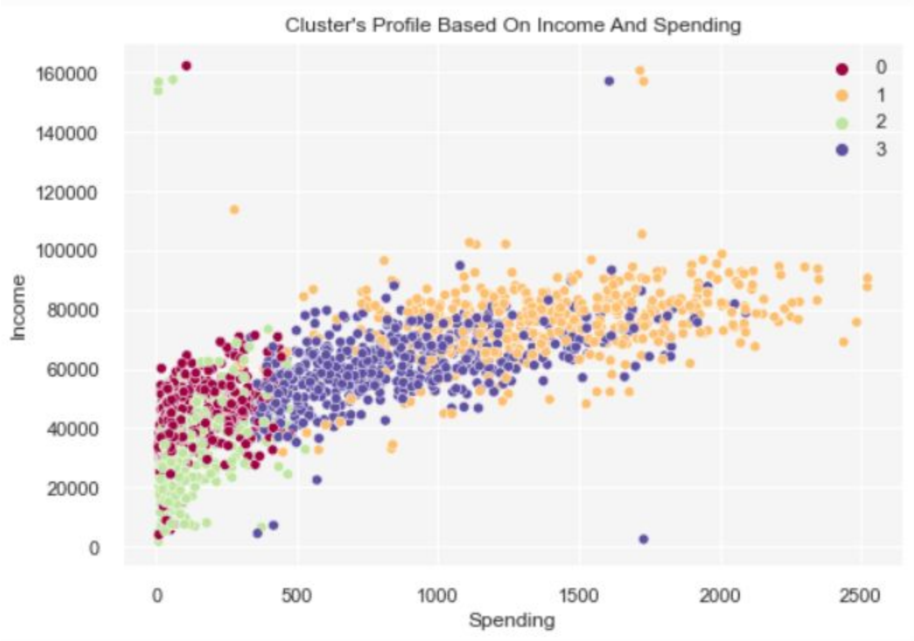


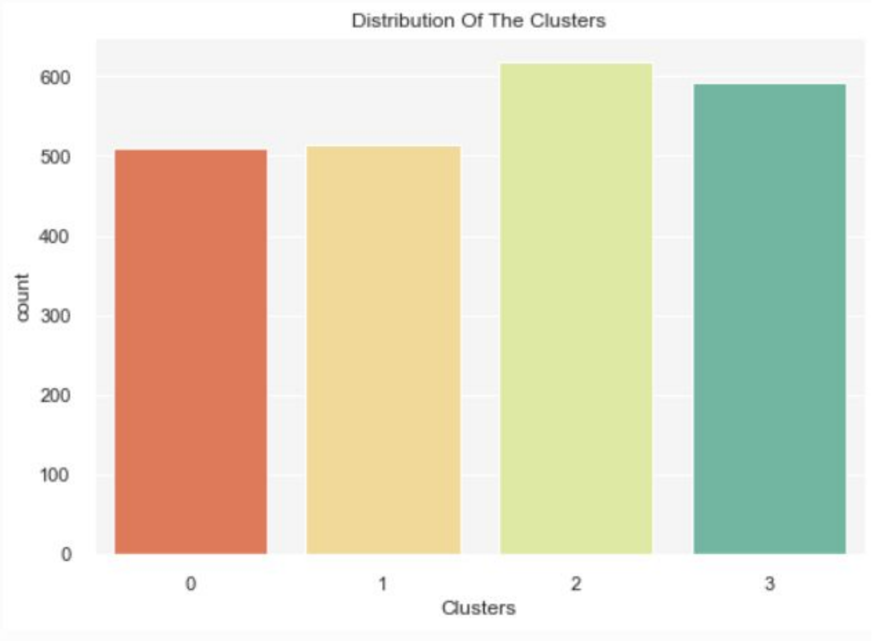
Proposed Work

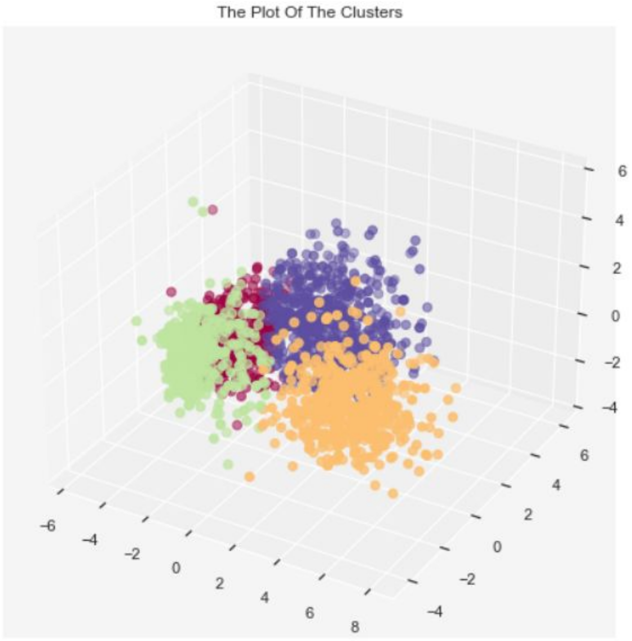
K-Means Clustering

Clusters look evenly distributed

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


Cluster's Profile Based On Income And Spending


Distribution Of The Clusters

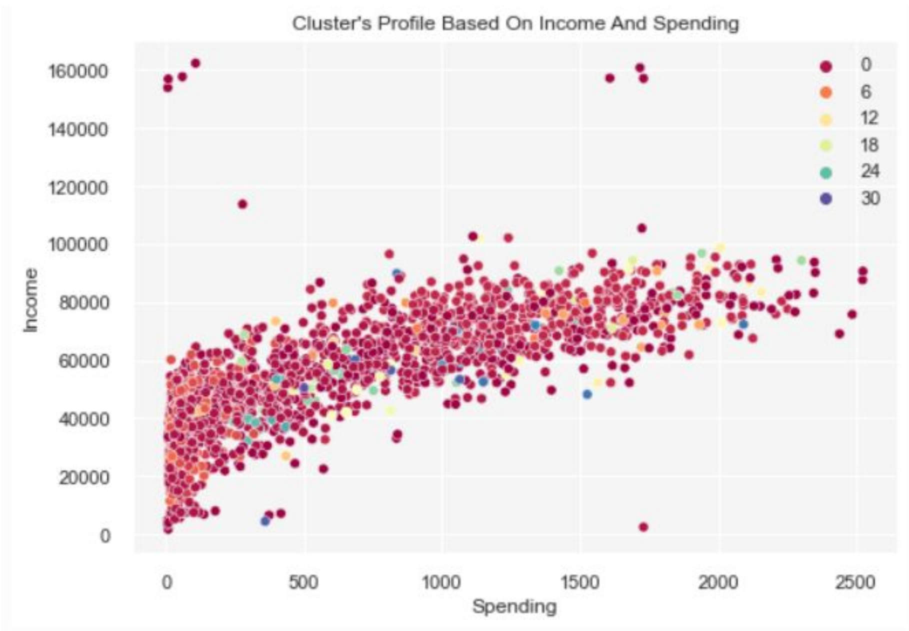

The Plot Of The Clusters

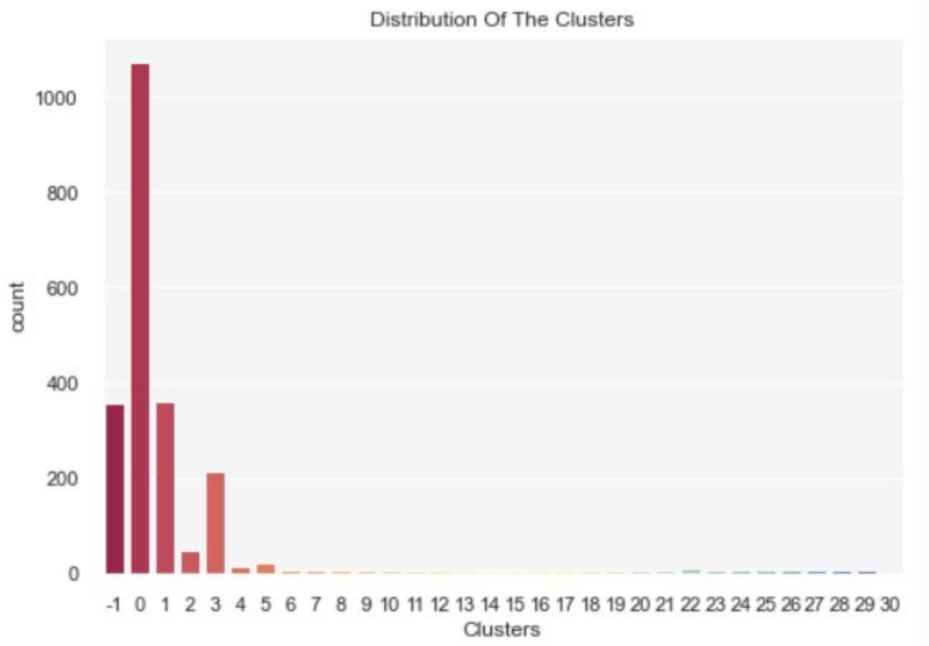
Proposed Work

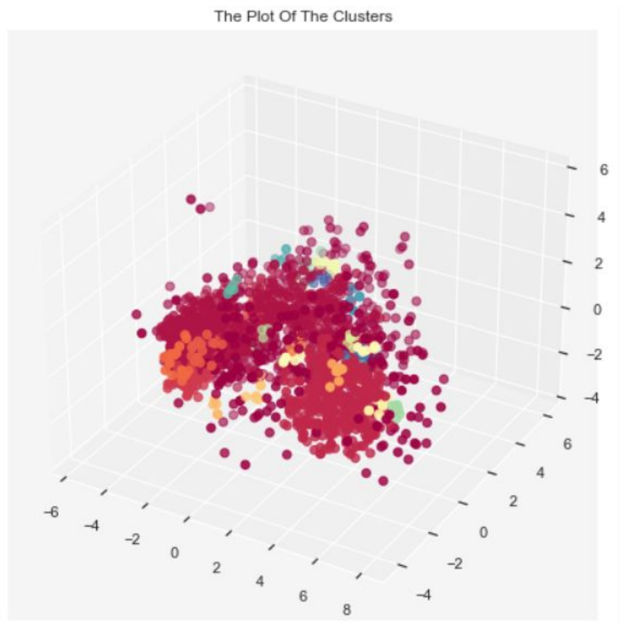
Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Not an even distribution of clusters

Will not proceed with investigating these clusters



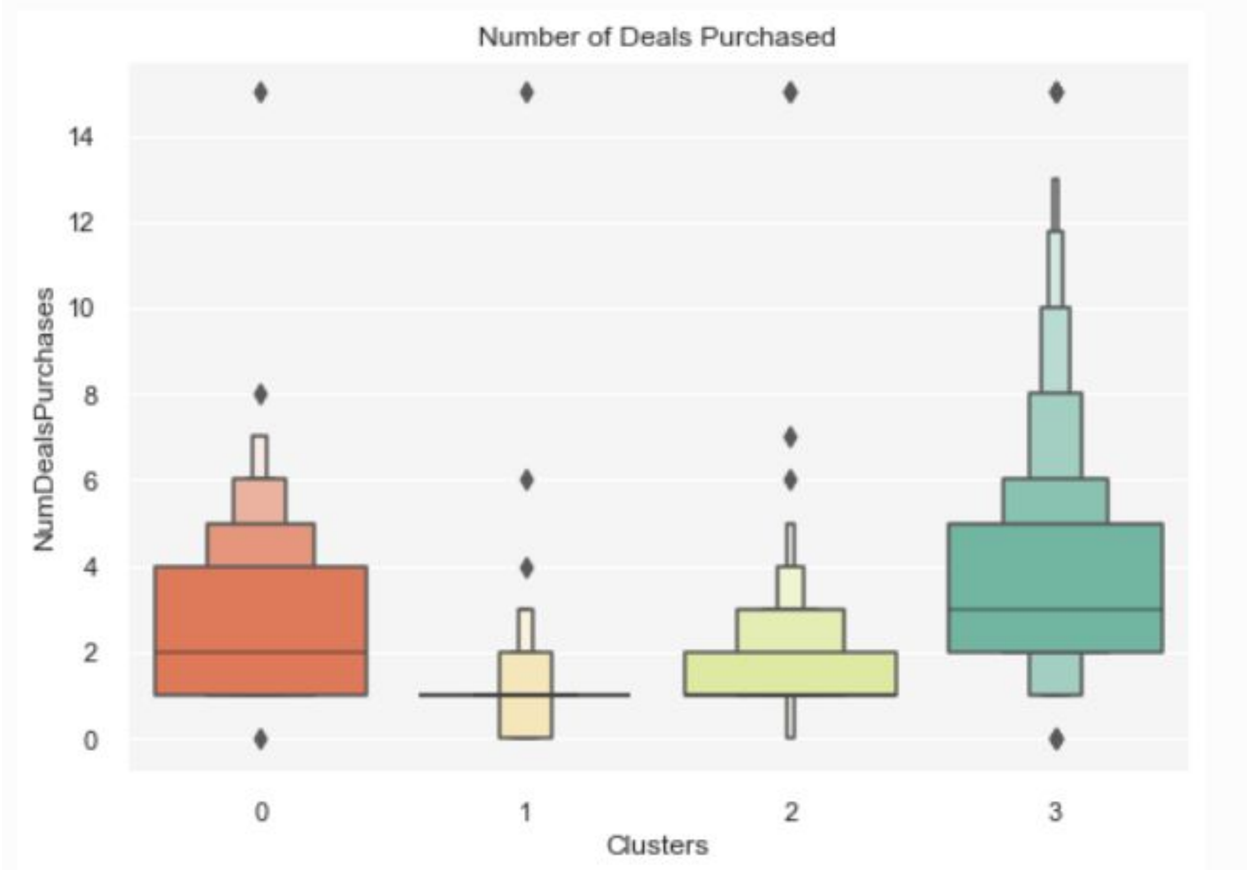




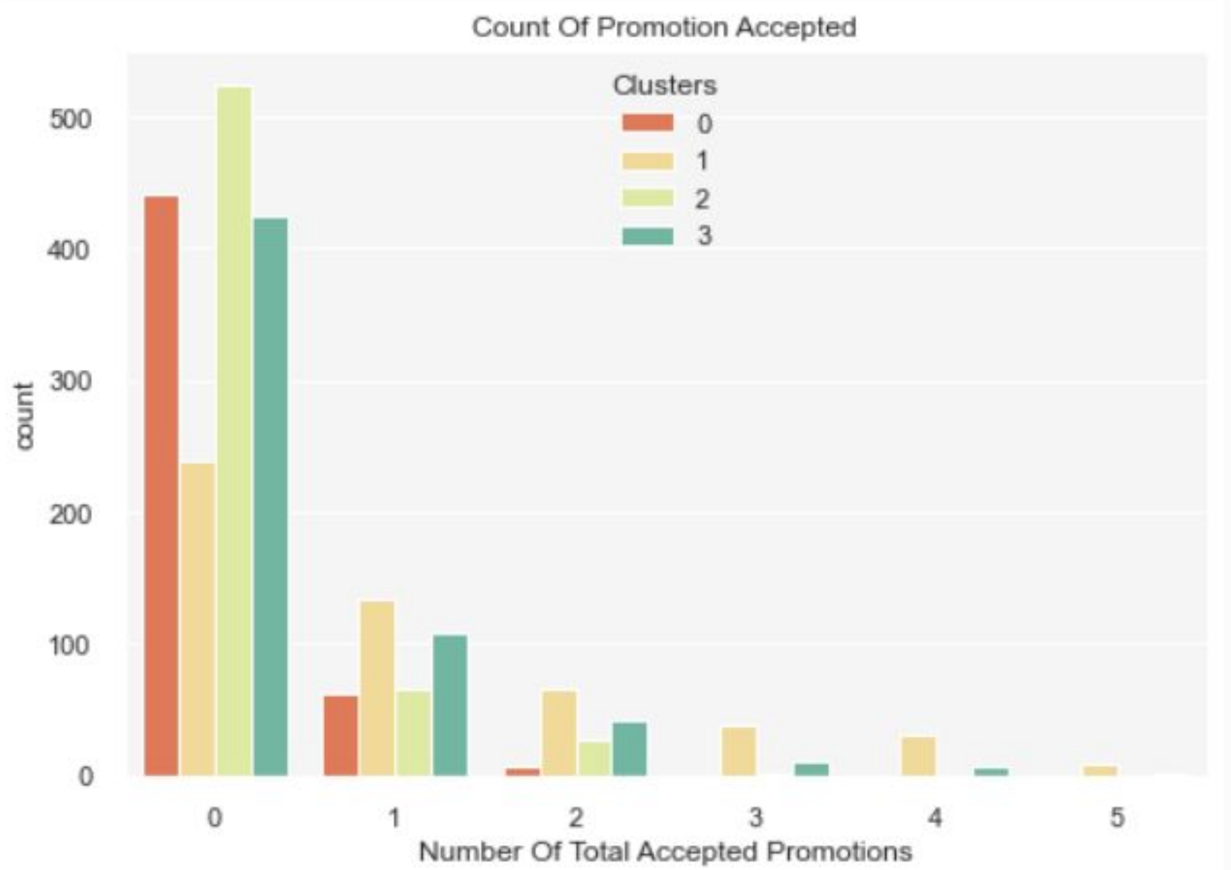
Similarity Metrics

Follow up performed on K-Means Clustering results

Explore distribution of accepted promotions across clusters



Number of Deals Purchased



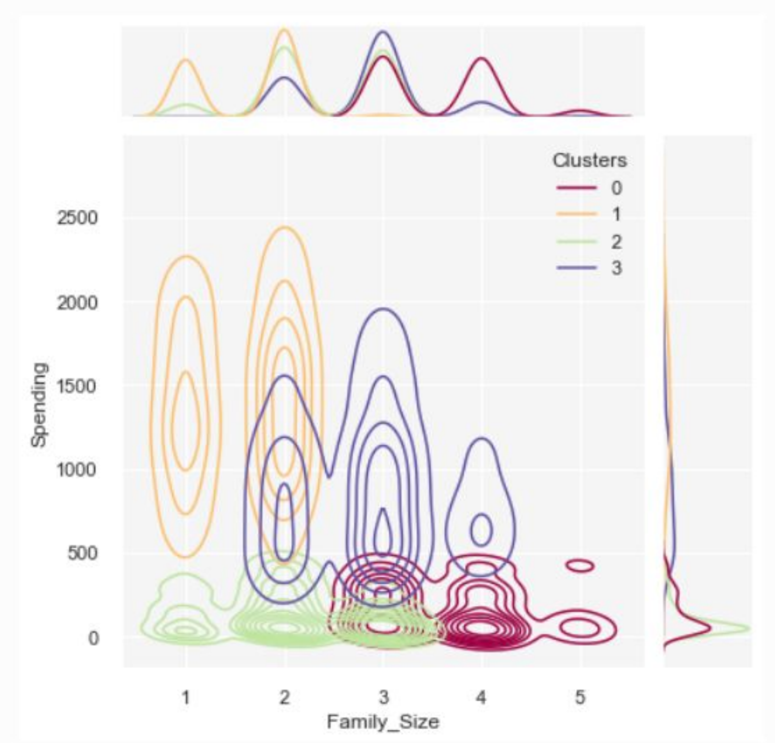
Count Of Promotion Accepted

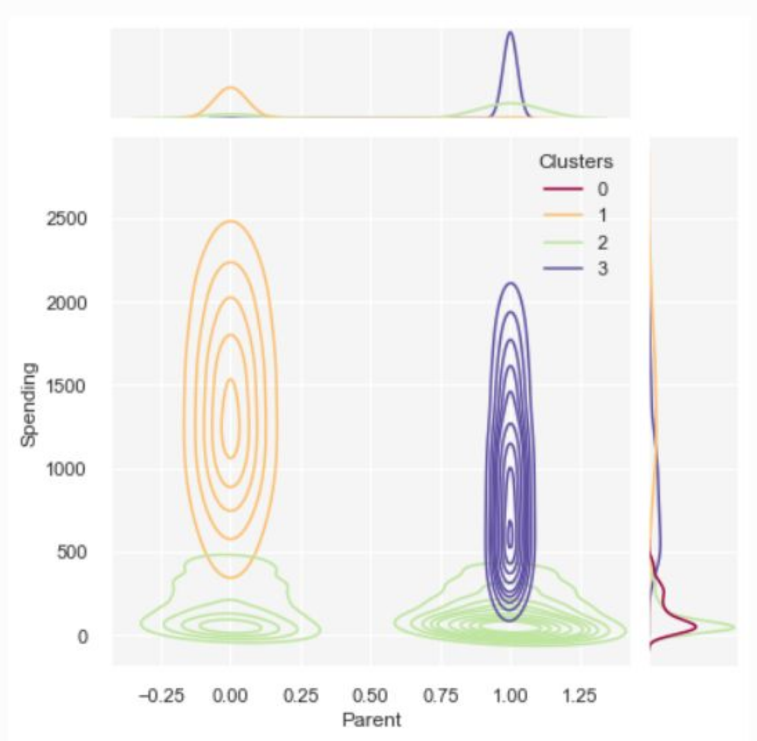
Evaluation

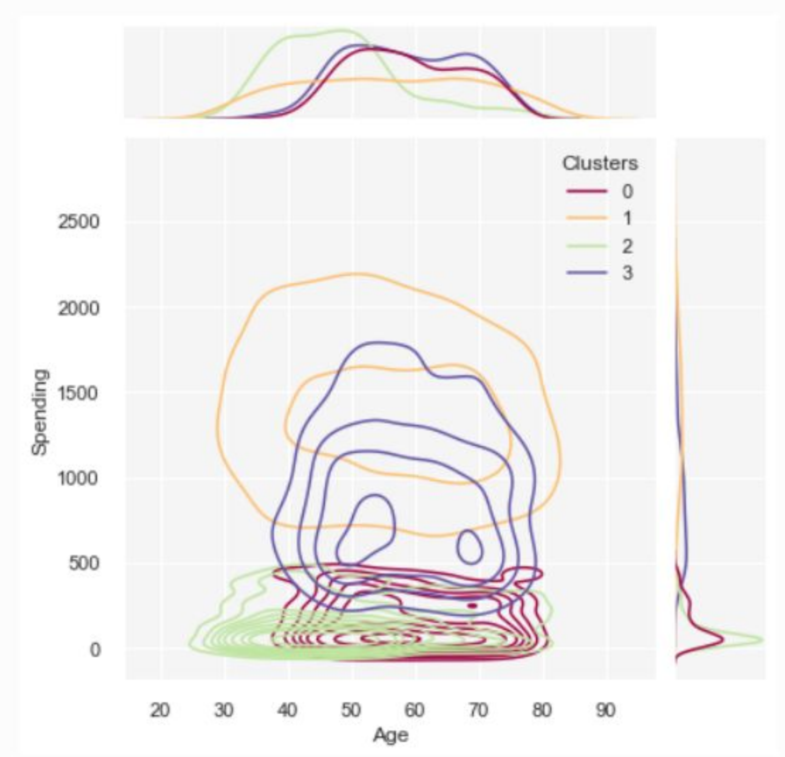
KDE Plots

Plotted several variables vs spending to find relationships and patterns

Some patterns quite clear, others distributed across clusters







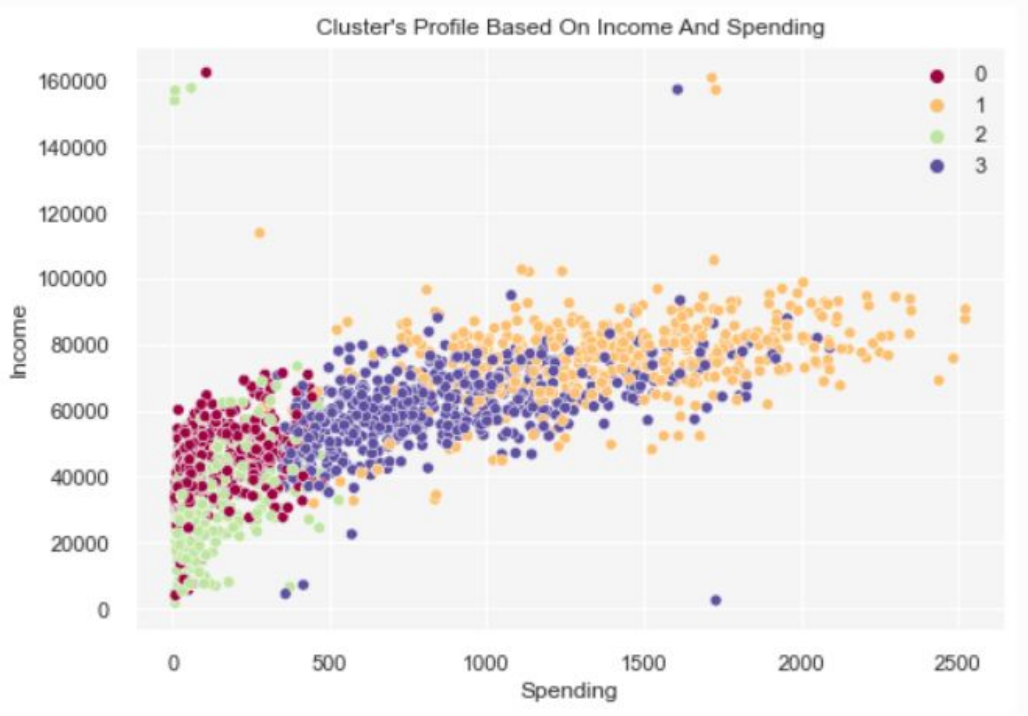
Evaluation

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