Customer Segmentation

Areeb Shafqat

6th December 2021

Aims & Objectives

01

Perform customer personality analysis with records from a grocery store's database

02

Explore the data using visualizations, feature engineering and dimension reduction (PCA) techniques

03

Segment customers using unsupervised learning into suitable clusters where each cluster reflects unique and similar set of characteristics and behaviours of customers.

04

Assist the store stakeholders in marketing their various products to specific types of customer segments after evaluating the clusters and customer traits of each cluster.

Data Description

The dataset contains 2240 observations with 29 different features where some of them include along with some statistics of the data:

- Customer ID
- Birth Year
- Education such as School (Basic) or University (Undergraduate, Masters, PhD)
- Yearly Income of Customer
- Number of kids and teenagers in customer's household
- Customer enrolment date with the grocery store
- Money spent on products such as wine, meat, sweets, gold etc.
- Did the customer complain?
- Customer purchases made in-store, on website and catalogue

	Education	Marital Status	Dt Customer
count	2240	2240	2240
unique	5	8	663
top	Graduation	Married	31-08-2012
freq	1127	864	12

	Year Birth	Income	Teen home	Gold Prods	Store Purchases
count	2109.000000	2109.000000	2109.000000	2109.000000	2109.000000
mean	1968.961119	51229.983404	0.504979	42.898530	5.774301
std	11.669281	20377.282998	0.541975	50.476469	3.204492
min	1940.000000	3502.000000	0.000000	0.000000	0.000000
25%	1960.000000	34853.000000	0.000000	9.000000	3.000000
50%	1970.000000	50616.000000	0.000000	24.000000	5.000000
75%	1977.000000	67605.000000	1.000000	54.000000	8.000000
max	1996.000000	105471.000000	2.000000	249.000000	13.000000

Feature Engineering & Data Analysis

i. Drop all null values from the data (only 24) and check for duplicates (none)

ii. Convert customer's date of enrolment into datetime object, subtract current date from that and select the number of days the customer has been buying from the grocery shop

iii. Calculate customer's age by subtracting current year with year of birth given

iv. Compute Total Spent by a customer by adding all the product purchase columns such as amount spent on wine, sweets, gold etc.

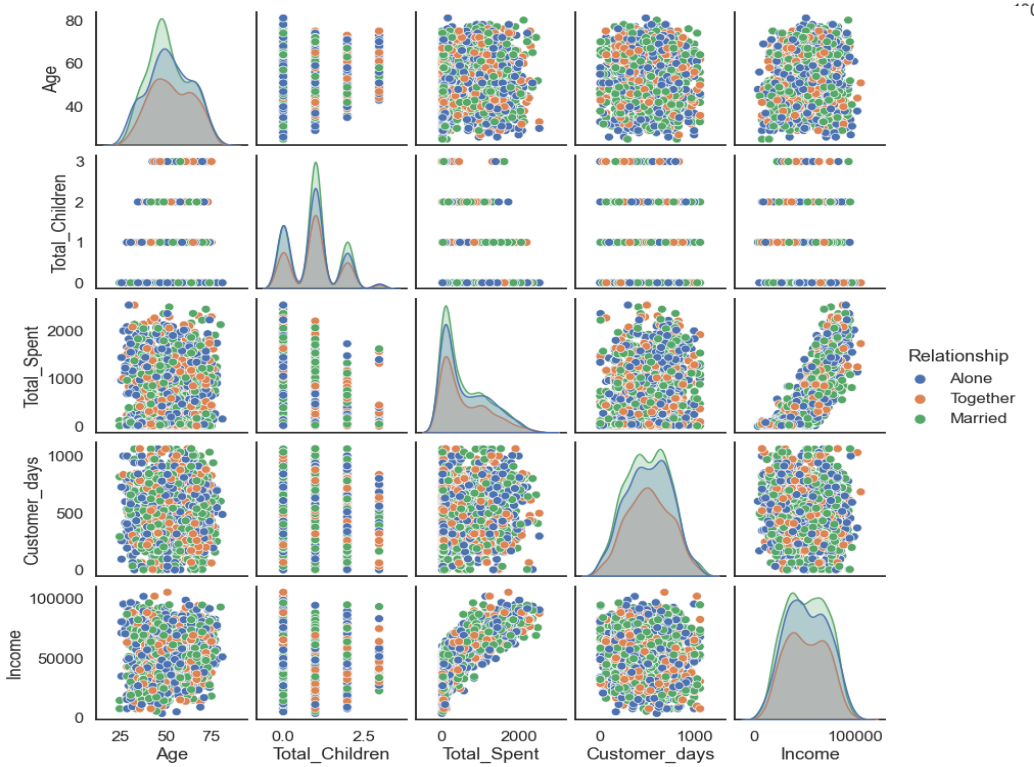
v. Replace observations such as divorced, alone, YOLO in the marital status all by the 'Single' string

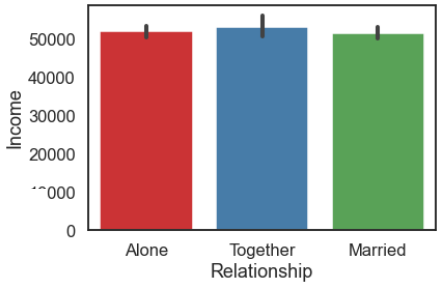
vi. Remove all outliers with z score above 4 standard deviations

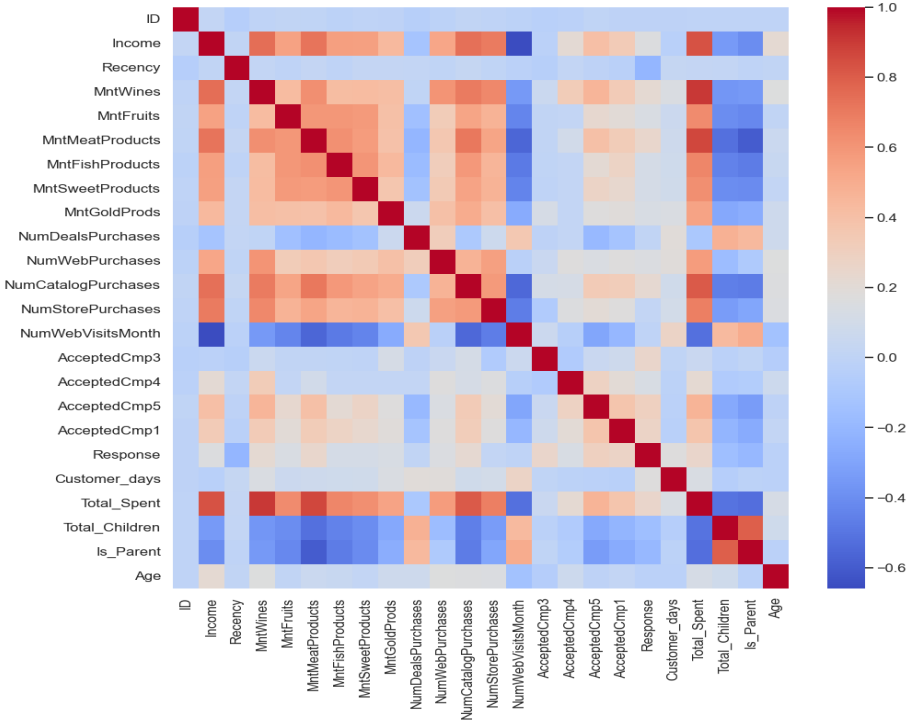
vii. Encode the categorical columns, namely Education and Marital Status, drop Customer ID and other redundant columns.

viii. Use a log transformation on columns with > 80% skew and standardized the data, ready to be used for Principal Component Analysis

Feature Engineering & Data Analysis

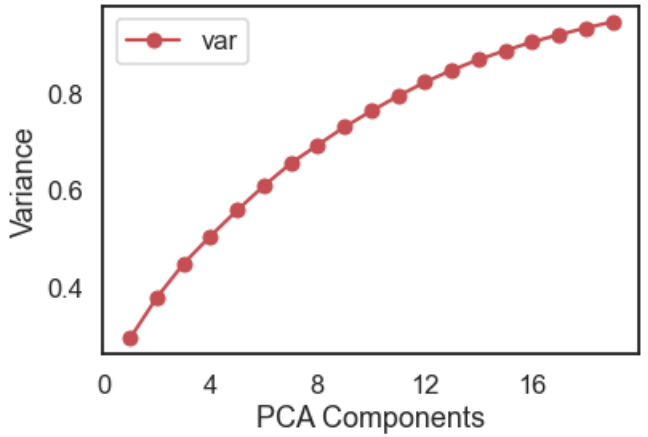


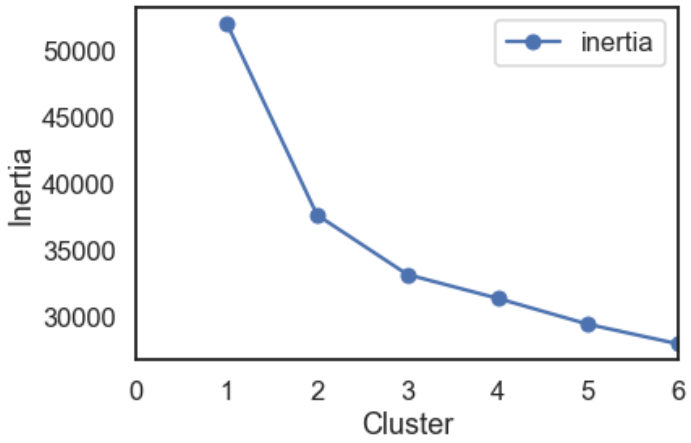




- ✓ After removal of outliers, there does not seem to be abnormal points in the pair plot shown.
- ✓ Yearly Income and Total Spent on the grocery store appear to have strong correlation unlike Income and Relationship
- ✓ There seem to be a lot of strong correlations in the dataset whereby using PCA, we can eliminate redundant columns

Unsupervised Learning Methods

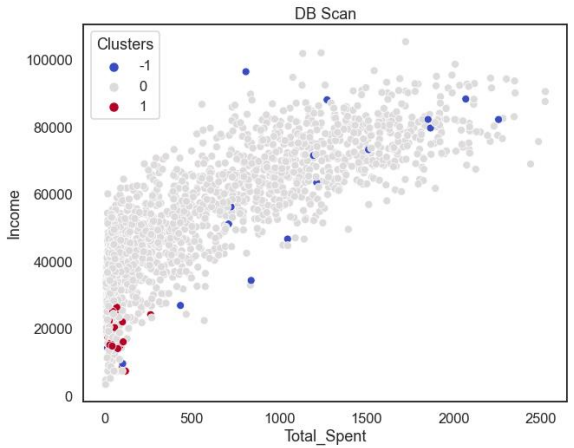


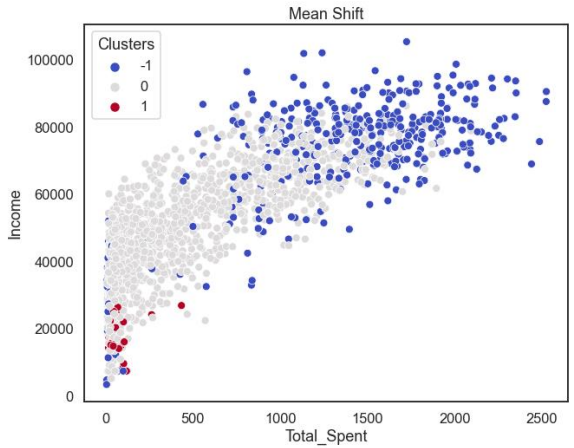


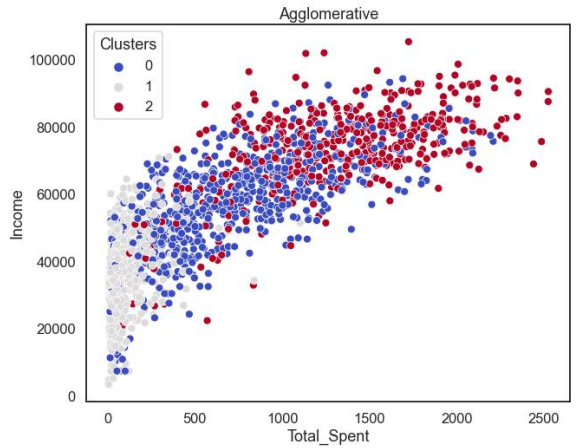
- ✓ After performing dimensionality reduction, the number of components reduced to 12 from 32 features
- ✓ Using K-means and the elbow methods, the rate of decrease in inertia looks to be the same after 3 clusters and hence, we will choose that value for K-Means and Agglomerative Hierarchical Clustering
- ✓ Other algorithms we will use are Mean Shift and DB Scan. The plots between Income and Total Spent is shown with the hue being the Cluster numbers and after evaluating how effectively the clusters are split, we will choose the most suitable algorithm and clusters.

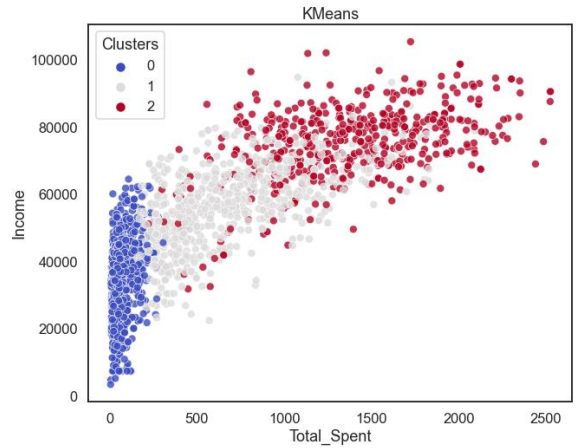
Unsupervised Learning Methods

- With K-Means and Agglomerative Clustering, there appears a much clearer split between Income and Total Spent which makes sense realistically as there can rich, working class and low-income customer types.
- Additionally, K-Means does a superior job at clustering the low income and average income customers compared with Agglomerative and hence, we will select K-Means
- From these figures, we can see that Cluster 0 represents the low income, low total spend community, Cluster 1, average income and spending and lastly, Cluster 2 being high income and high spending although there may be outliers such as people with high income and lower relative spending

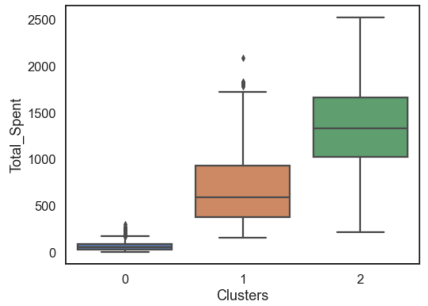


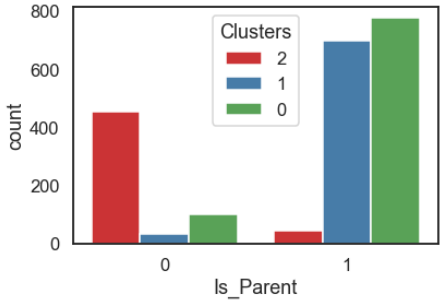


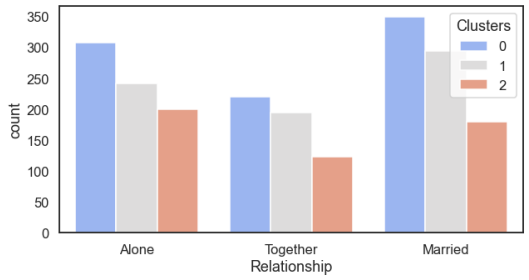


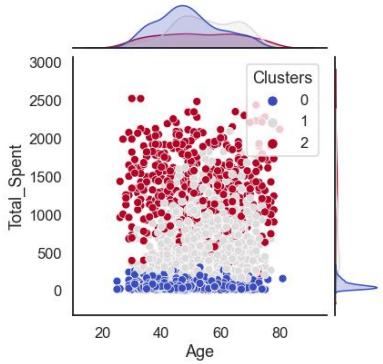


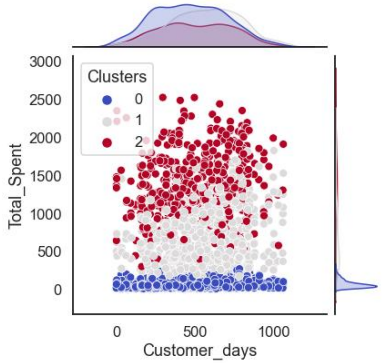
Segmentation Insights

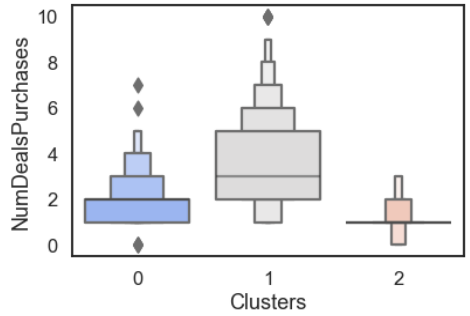


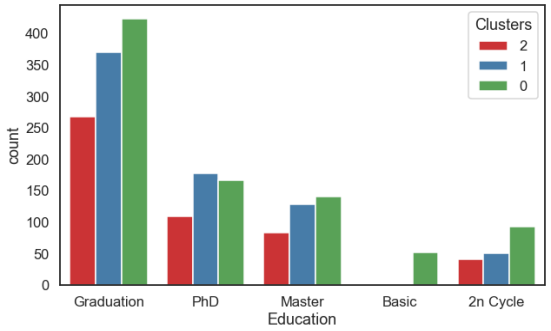


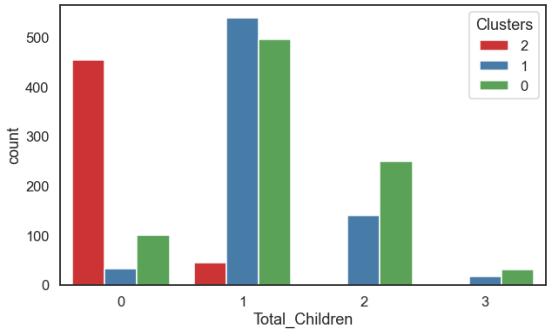


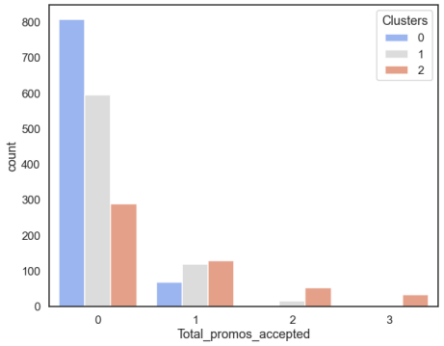












- Cluster 0 – Low spending, mainly parents and 1 child, do not engage with promos, could have high school education only unlike other clusters
- Cluster 1 – average spending, parents, can have 1-2 children, engages with deals
- Cluster 2 – Accept campaign promos, not deal purchases, not parents, high spending, relatively newer that cluster 0,1.

Segmentation Insights

Clusters	Wines	Fruits	Meat Products	Fish Products	Sweets	Gold	Web Purchases	Catalog Purchases	Store Purchases	Web Visits/ Month
0	29.925884	3.891676	16.567845	5.484607	4.010262	12.222349	1.851767	0.417332	3.023945	6.454960
1	406.463014	25.213699	138.490411	33.239726	25.790411	57.952055	6.031507	2.972603	7.326027	5.764384
2	605.515936	63.918327	449.316733	96.852590	64.388446	74.599602	4.982072	5.764940	8.322709	2.695219

- ✓ Cluster 0: These customers although spend less, like wine, meat and gold where they mainly use in store purchases after searching products on the web
- ✓ Cluster 1: Customers spend much on Wine and Meat and primarily shop either on the web or in store
- ✓ Cluster 2: Huge spending on Wine and Meat and do like Fish. They do not visit the web much and use store purchases largely and frequently engage in Catalog and Web purchases.

Flaws and Future work

- The clustering model was selected based off only two features (Income and Total Spent) and therefore, further exploration with hyperparameters and clustering may prove models such as Agglomerative clustering or Mean Shift to be more effective

- Revisiting this problem with further data of customer behaviors such as customer interests like travelling or sports could result in better clustering with regards to education levels, income and other features.

- Moreover, perhaps experimenting further with columns kept after PCA, using 4 clusters instead and the removal of further outliers (z score being 2 or 3 standard deviations above the mean) should give varied yet insightful clusters.

Thank You!