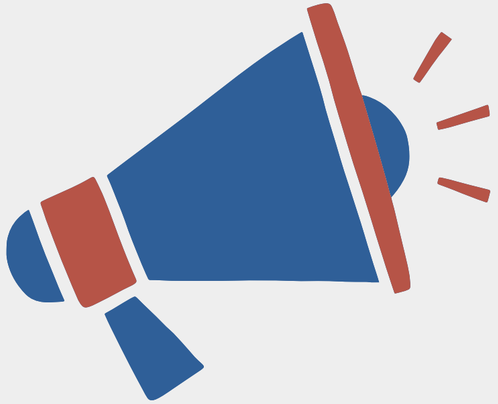
SC1015 MINI PROJECT

CUSTOMER PROFILE ANALYSIS

Lab Group: Z136_Team 3
Wong Yu Fei
Ng Li Lin Evonne

PROBLEM MOTIVATION



Supermarkets often find it **difficult to advertise** to customers as different customers have **different preferences and needs.**

From supermarket data, we can identify separate groups of customers with **similar attributes**

Supermarkets can **tailor advertising** towards these groups which will help increase their sales.

PROBLEM DEFINITION

How can supermarkets leverage machine learning to **identify customer segments** based on customer attributes?

With many clustering algorithms available, can we identify a model that can **best segment** the supermarket's customers?

DATASET USED

- Public dataset from Kaggle, titled "Customer Personality Analysis"
- Number of entries: 2240
- Number of columns: 29

Category	Columns
People	ID, Year_Birth, Education, Marital_Status, income, Kidhome, Teenhome, Dt_Customer, Recency, Complain
Product	MntWine, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds
Promotion	NumDealsPurchases, AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, Response
Place	NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth

EXPLORATORY DATA ANALYSIS

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 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
```

Result:

- Identified column with null values: Income
- Identified non-numerical columns: Education, Marital_Status, Dt_Customer

DATA PREPROCESSING

Data Preprocessing Steps:

- Modify Columns
  - *Membership Duration*: 1/1/2021 - Dt_Customer
  - *Age*: 2021 - Year_Birth
  - *RelationshipStatus*: single or duo (from Marital_Status)
  - *Education*: First, Second or Third Cycle

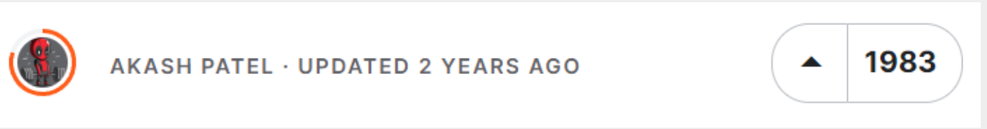
- New Columns
  - *PurchaseQty* = Sum of Purchases Columns
  - *NumOfChildren* = kidhome + Teenhome
  - *FamilySize* = NumOfChildren + RelationshipStatus (1 for single, 2 for duo)
  - *Expenditure* = Sum of Product Columns
  - *AcceptedCmpAll* = combine accepted campaigns 1-5 and Response

- Replacement of null values with the column's median
  - Columns with null values: Income

- One hot encode Categorical variables
  - Columns: Education, RelationshipStatus

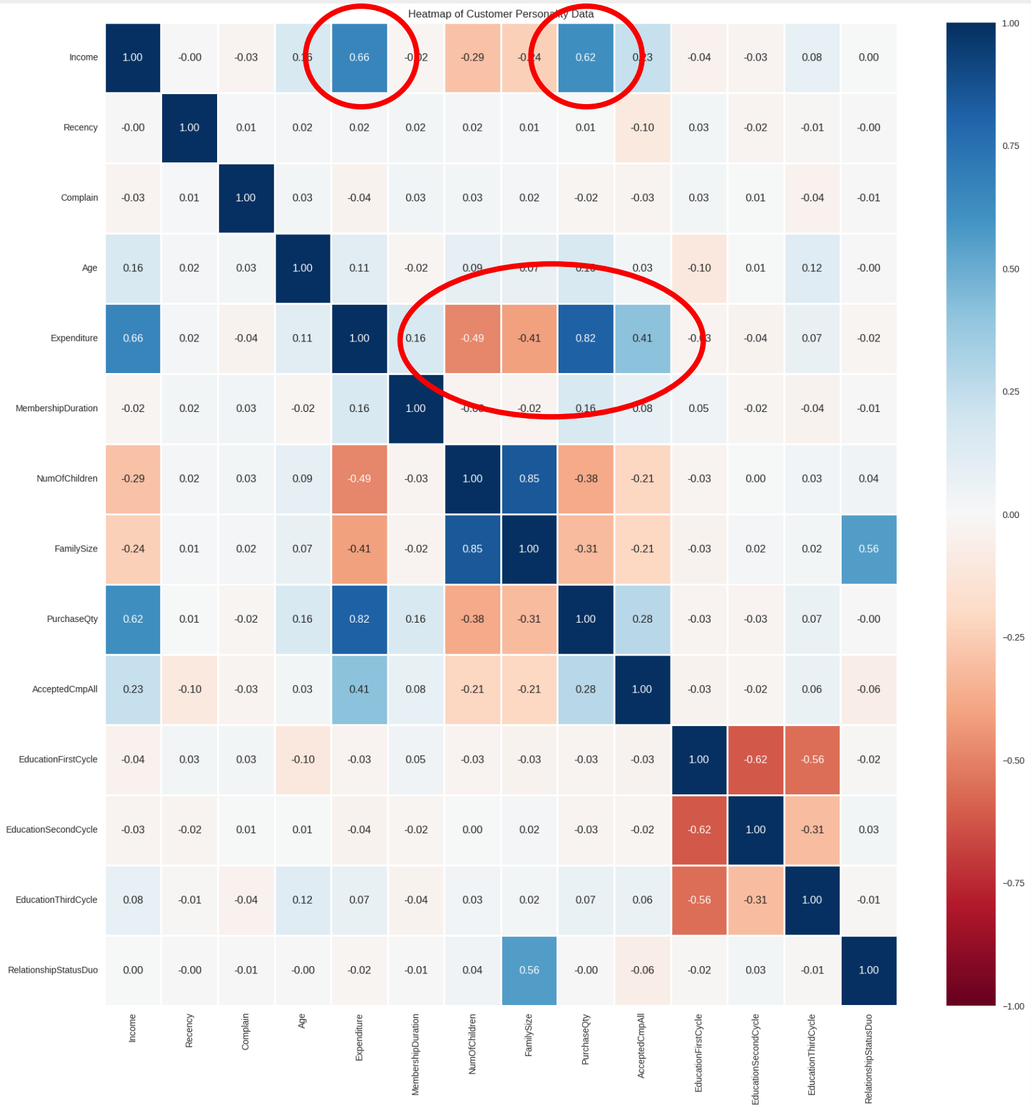
Assumption:

- Exact data collection date is unspecified and hence it is assumed to be 1 January 2021




```
14  MembershipDuration  2240 non-null  int64
15  Age                 2240 non-null  int64
16  Expenditure         2240 non-null  int64
17  RelationshipStatus  2240 non-null  object
18  PurchaseQty         2240 non-null  int64
19  NumOfChildren       2240 non-null  int64
20  FamilySize          2240 non-null  int64
21  AcceptedCmpAll      2240 non-null  int64
```

DATA VISUALISATION - HEAT MAP



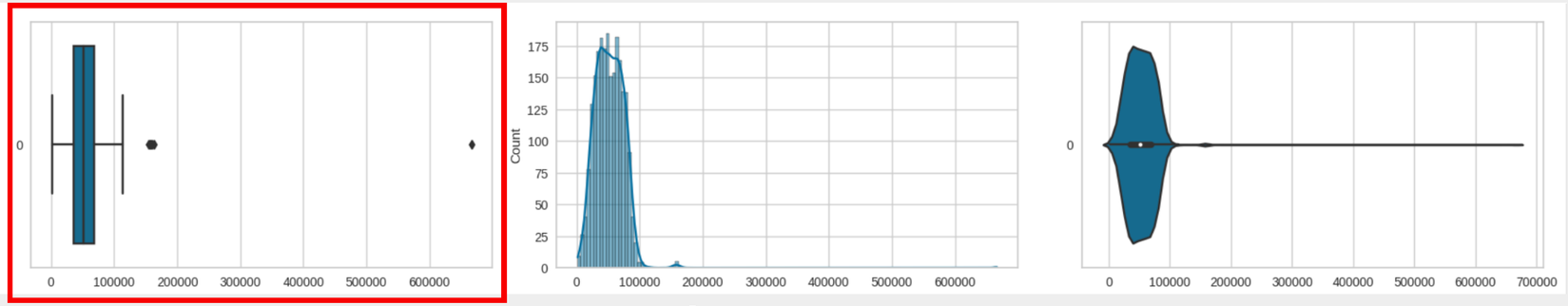
Purpose:

Show the relationships between variables

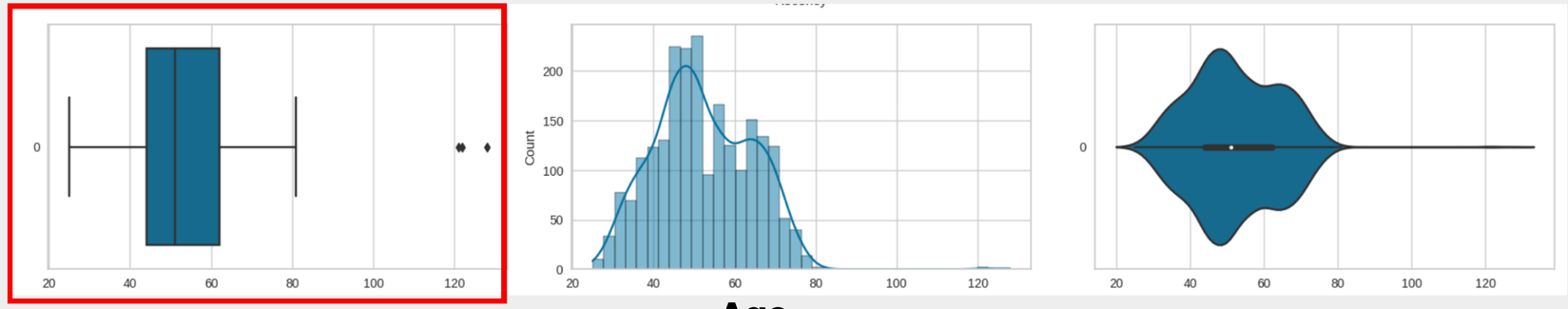
Observations:

- Income and PurchaseQty have a high positive correlation
- Income and Expenditure have a high positive correlation
- Expenditure and AcceptedCmpAll have a moderate positive correlation
- Expenditure and FamilySize/ NumOfChildren have a moderate negative correlation

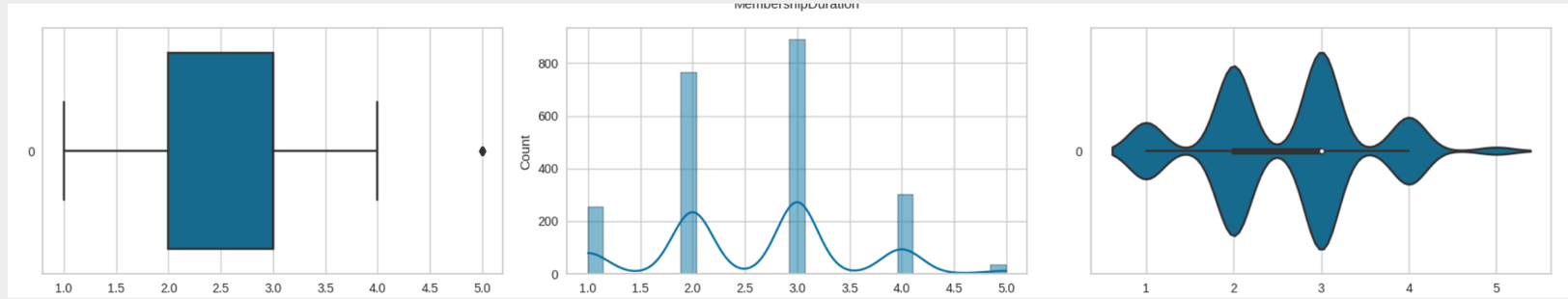
DATA VISUALISATION - BOX PLOT, KDE PLOT, VIOLIN PLOT



Income



Age



FamilySize

Purpose:
Visualise the distribution of numerical variables

Observations:

- Presence of significant outliers in the Income and Age columns
- Vastly different magnitude of values between Columns (Income ranging from 0-600000 vs FamilySize ranging from 1-5)

Data Preprocessing Steps (cont):

- Removal of numerical outliers
  - Outliers are values outside the range (q1-1.5 x IQR,q3+1.5 x IQR)
  - Exception for Family Size and NumOfChildren as values are small; 1-5 and 0-3 respectively

- Standard Scaled Numerical variables

Results:

- Removal of outliers dropped 171 rows (7.125%)
- Number of Columns dropped from 29 to 25

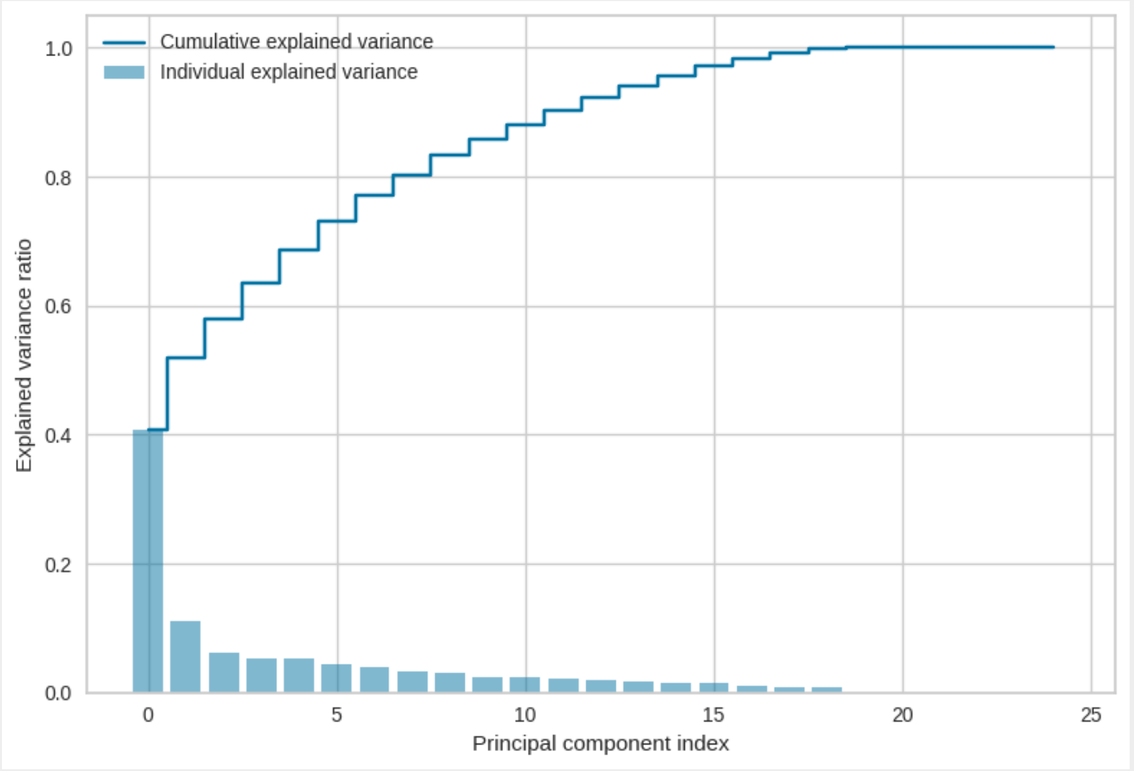
13	MembershipDuration	2229	non-null	float64
14	Age	2229	non-null	float64
15	Expenditure	2229	non-null	float64
16	PurchaseQty	2229	non-null	float64
17	NumOfChildren	2229	non-null	float64
18	FamilySize	2229	non-null	float64
19	AcceptedCmpAll	2229	non-null	float64
20	EducationFirstCycle	2229	non-null	float64
21	EducationSecondCycle	2229	non-null	float64
22	EducationThirdCycle	2229	non-null	float64
23	RelationshipStatusDuo	2229	non-null	float64
24	RelationshipStatusSingle	2229	non-null	float64

DIMENSIONALITY REDUCTION

Dimensionality Reduction:

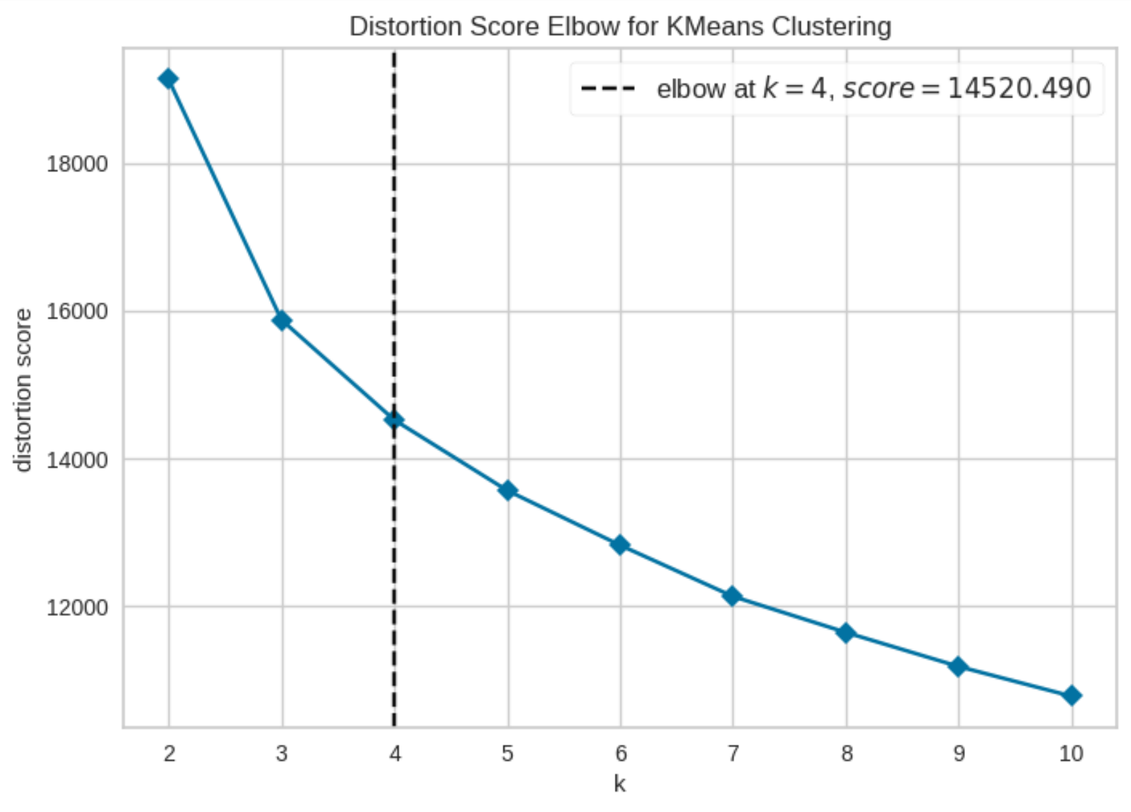
- Transformed variables into principal components and plotted them against explained variance ratio
- Kept principal components that contributed more than 1 variable's worth of information (4%)
- Top 7 PCs account for 77% of dataset's explained variance

PC No.	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Percentage/ %	40.7	11.2	6.1	5.4	5.2	4.3	4.0



FINDING OPTIMAL CLUSTER NUMBER

Method 1: Elbow Method



How it works:

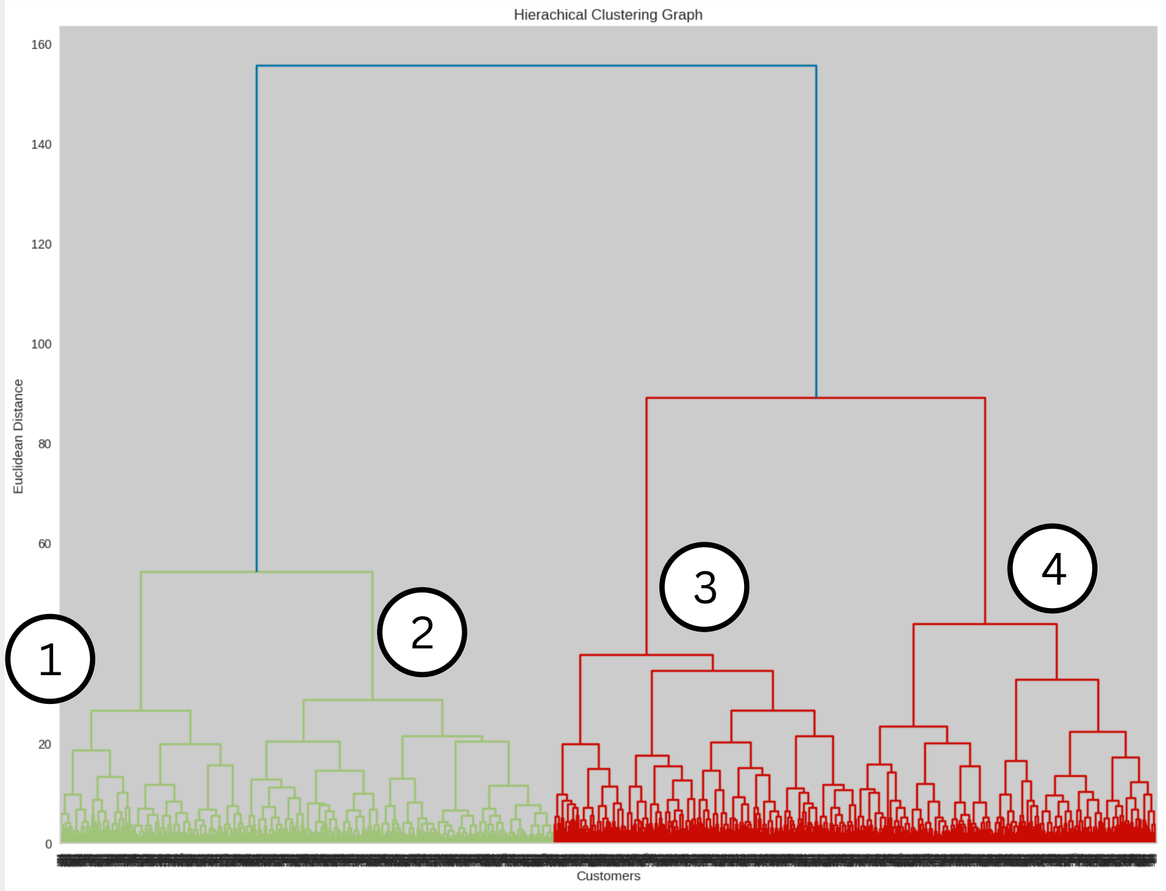
For each value of cluster number K, the Within-Cluster Sum of Square (WCSS) is calculated and compared.

WCSS is the sum of the squared distance between each point and the centroid in the cluster.

Optimal cluster number: 4

FINDING OPTIMAL CLUSTER NUMBER

Method 2: Hierarchical Graph



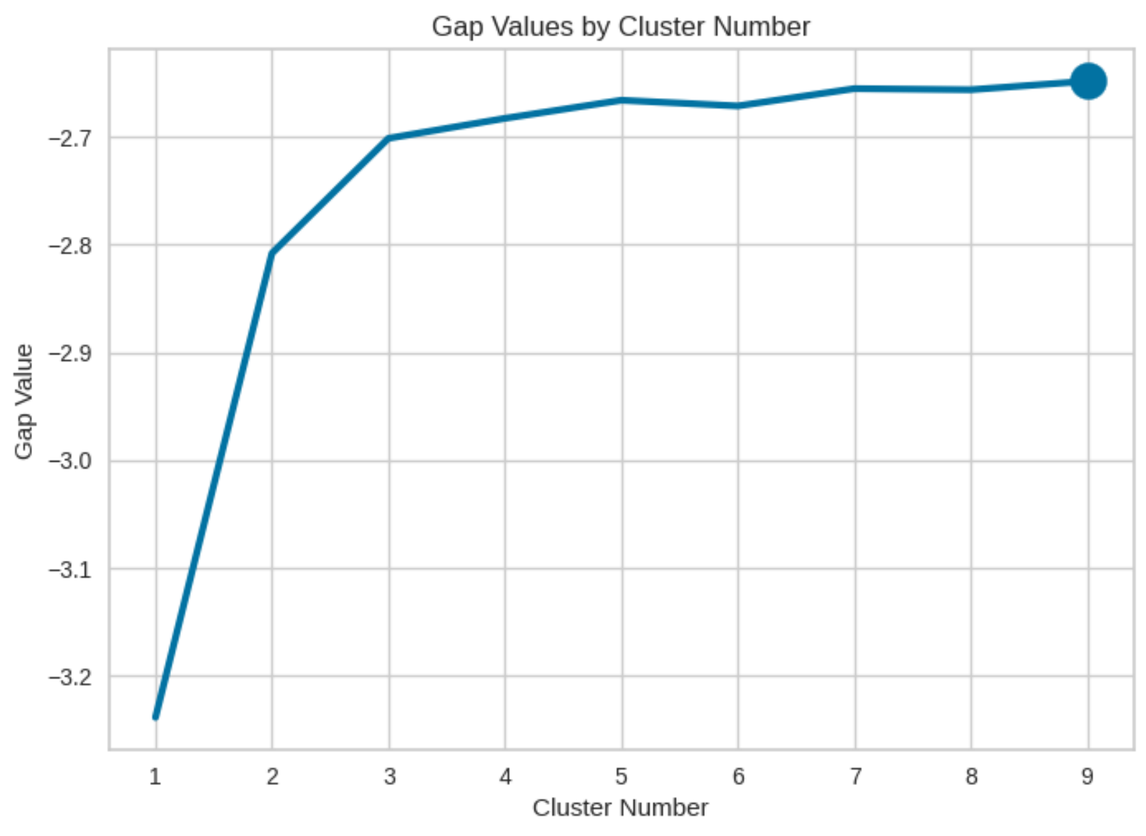
How it works:

Points are joined to other points to form clusters based on their euclidean distances. A dendrogram is used to show the sequences of merges or splits of clusters.

Optimal cluster number: 4

FINDING OPTIMAL CLUSTER NUMBER

Method 3: Gap Statistic



How it works:

Cluster compactness is compared with a null reference distribution of data. The cluster number for which cluster compactness falls the farthest below the reference curve is the optimal.

Optimal cluster number: 3

FINDING OPTIMAL CLUSTER NUMBER

Optimal Cluster Number:

4

6 clustering algorithms across 5 clustering methods

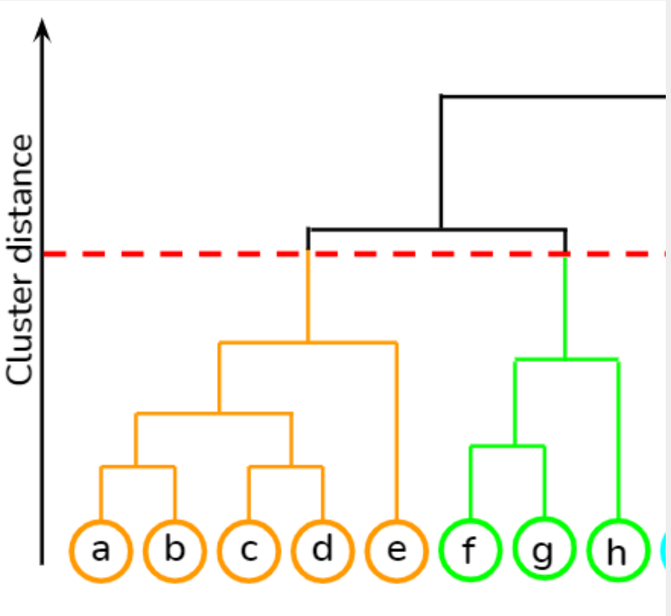
CLUSTERING ALGORITHMS

Connectivity/ Hierarchical Clustering

- Agglomerative Clustering Model

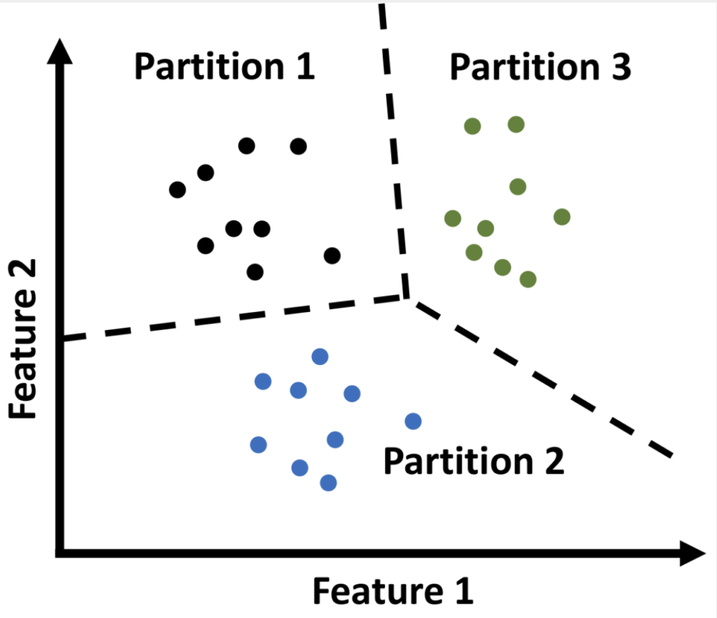
How it works:

Based on distance connectivity between points to cluster them. Works on the concept that nearby objects are more related than further objects



Centroid/ Partition Clustering

- K-Means
- Mean Shift



How it works:

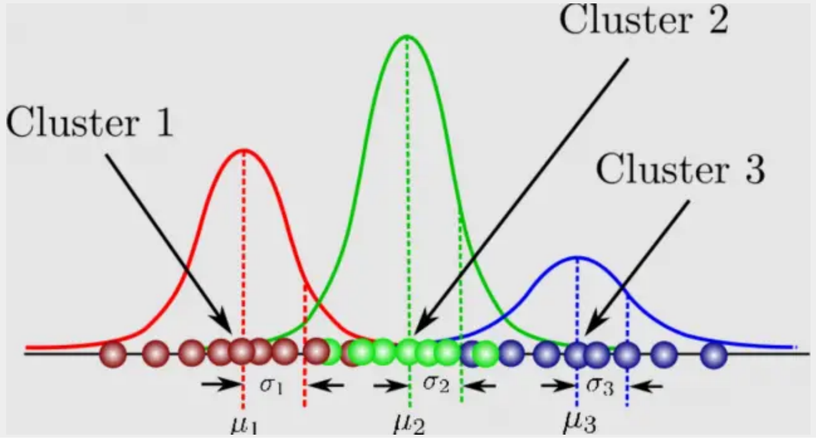
Each cluster is represented by a single vector. Algo finds the optimal cluster positions such that the squared distance object to other clusters is maximized.

6 clustering algorithms across 5 clustering methods

CLUSTERING ALGORITHMS

Distribution Model

- Gaussian Mixture Model

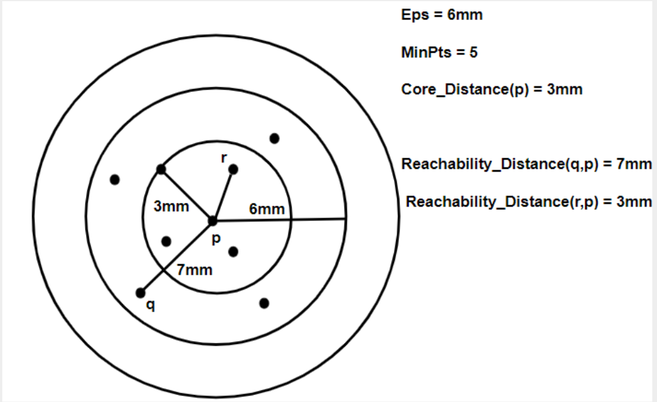


How it works:

Clusters are defined as objects belonging to the same distribution. As distance from the distribution increases, the probability that the point belongs to the distribution decreases

Density Model

- Ordering Points To Identify the Clustering Structure (OPTICS)



How it works:

Connects areas of high example density into clusters allowing for arbitrary shape distributions as long as dense regions are connected.

CLUSTERING ALGORITHMS

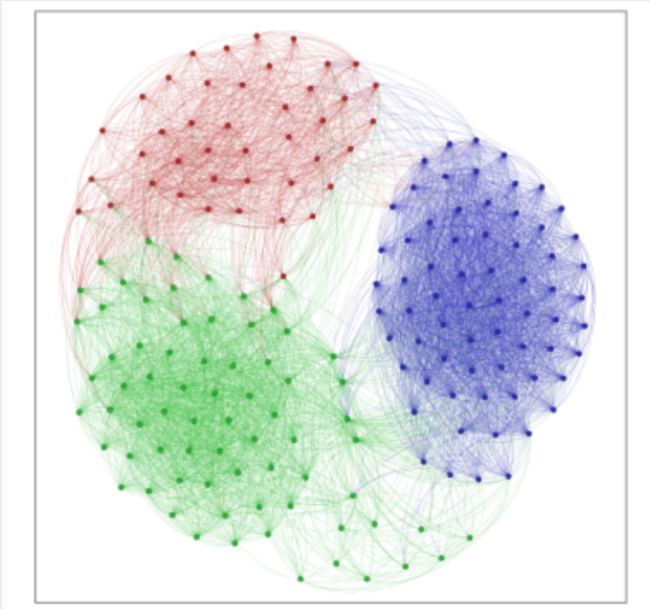
6 clustering algorithms across 5 clustering methods

Graph-based Method

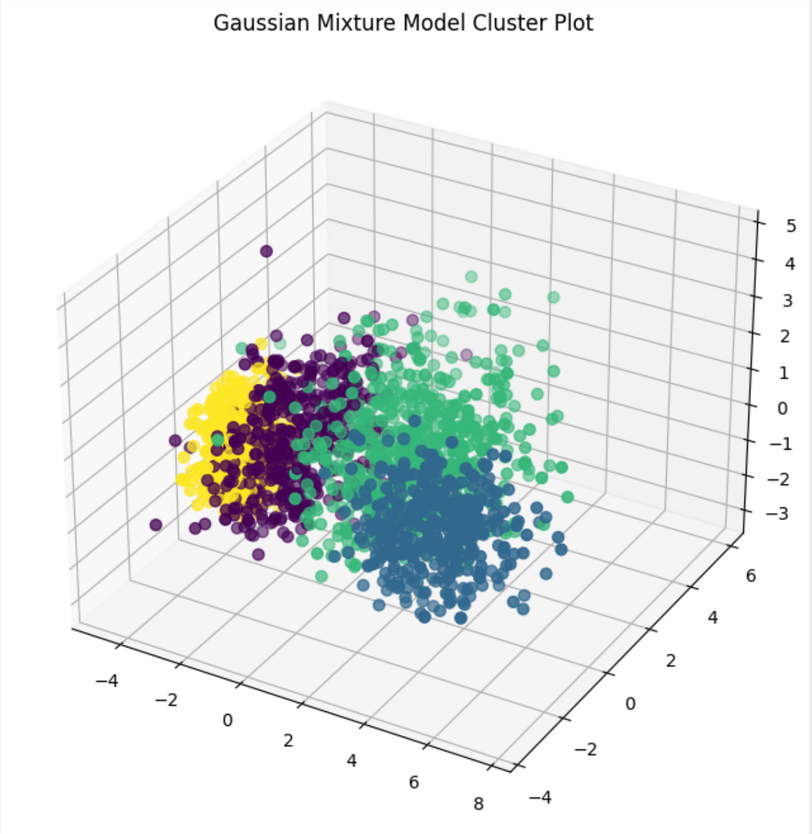
- Spectral Clustering

How it works:

Data is represented as a graph where each object is represented as a node and the distance between two objects is connected by a weighted edge



Visualisation of Clusters

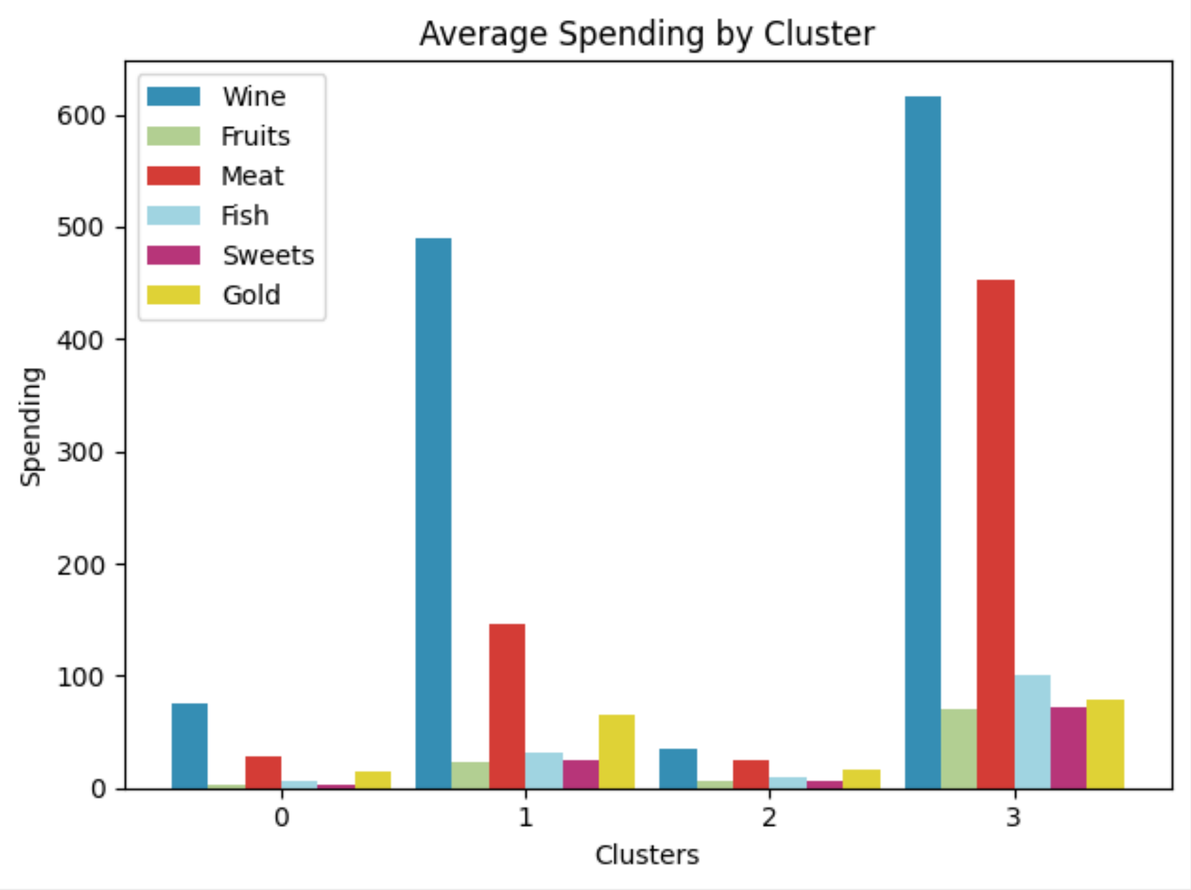


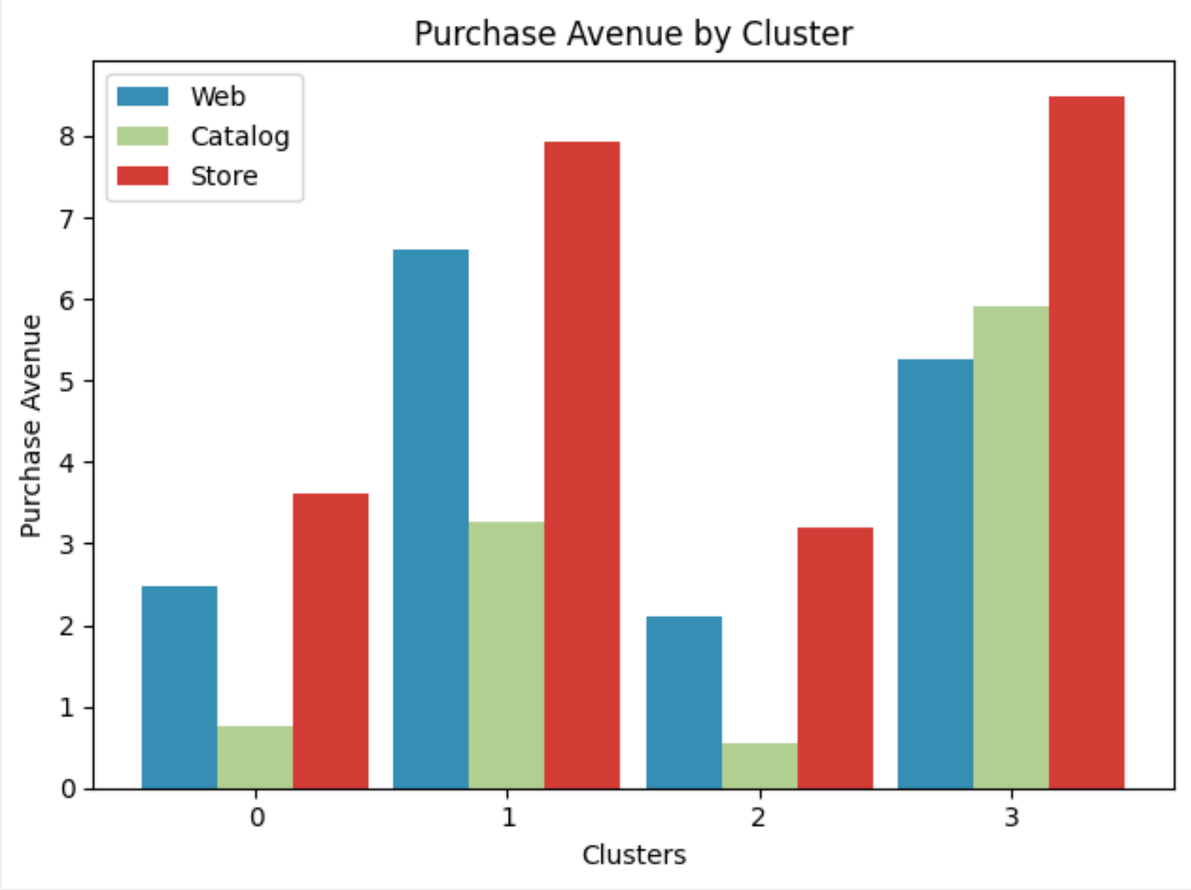
EVALUATION OF MODELS

Metric/ Model	KMeans	Mean Shift	OPTICS	Spectral Clustering	Gaussian Mixture	Agglomerative Clustering
Silhouette score	2 0.196	4 0.136	6 -0.366	1 0.217	5 0.155	3 0.178
Calinski Harabaz Index	1 948.7	4 62.4	5 9.68	6 5.13	3 816.1	2 836.9
Davies Bouldin Index	4 1.63	3 1.49	2 1.41	1 0.68	6 2.03	4 1.63

CUSTOMER PROFILING

Interpreting Customer Cluster Spending Behaviour





CUSTOMER PROFILING

Cluster No.	0	1
Size	447 (20.1%)	548 (25.8%)
Traits	• Medium-Low Income: ~42K • Oldest Age Group: ~57 years old • Highest Number of Children: 1.9	• Medium-High Income: ~59K • Medium-High Age Group: ~56 years old • Highest Third-Cycle Education: 27.7%
Behavioural Characteristics	• Top 3 Product Categories with Highest Expenditure: Wine, Meat, Gold • Ranking of Purchase Avenues: Store, Web, Catalog	• Highest number of purchases made with a discount • Percentage of Complaints: 1.28% (Highest) • Top 3 Product Categories with Highest Expenditure: Wine, Meat, Gold • Ranking of Purchase Avenue: Store, Web, Catalog

CUSTOMER PROFILING

Cluster No.	2	3
Size	659 (29.6%)	575 (25.8%)
Traits	• Lowest Income: ~32K • Youngest Age Group: ~45 years old • Lowest Third-Cycle Education: 14.7%	• Highest Income: ~75K • Medium-Low Age Group: ~53 years old • Lowest Number of Children: 0.2
Behavioural Characteristics	• Smallest Spender • Lowest Percentage of Campaign Acceptance: 14.4% • Top 3 Product Categories with Highest Expenditure: Wine, Meat, Gold • Ranking of Purchase Avenue: Store, Web, Catalog	• Largest Spender • Smallest number of purchases made with a discount • Highest Percentage of Campaign Acceptance: 49.6% • Top 3 Product Categories with Highest Expenditure: Wine, Meat, Fish • Ranking of Purchase Avenues: Store, Catalog, Web

RESULTS

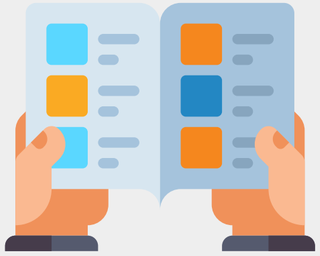
Most valuable customers are Cluster 3 due to their highest expenditure

Campaigns should be targeted towards customers of Cluster 3

Deals should be targeted towards customers of Cluster 1

Top Product Categories to promote: Wine, Meat, Fish and Gold

On the web, supermarkets should target customers of the Clusters 0, 1, 2



In catalogs, supermarkets should target customers of Cluster 3

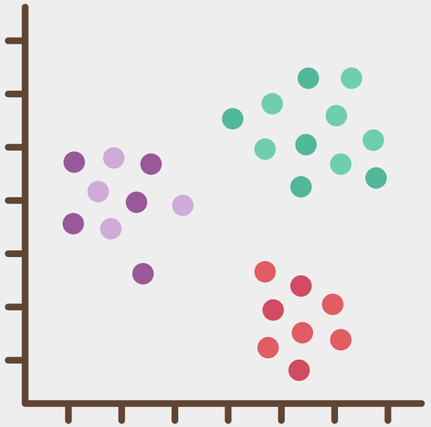
WHAT WE LEARNED

Data Preprocessing Techniques:

- Data Scaling- Sklearn StandardScaler
- Dimensionality Reduction through Principal Component Analysis

Understanding Dataset:

- Modifying columns
- Interpreting Clusters
- Drawing recommendations



Clustering Framework:

- *Identifying the optimal number of clusters:* Elbow, Hierarchical Graph, Gap Statistics
- *Applying Clustering algorithms:* Agglomerative Clustering, K-Means, Mean Shift, Gaussian Mixture, OPTICS, Spectral
- *Evaluating Clustering algorithms:* Silhouette Scorem Calinski Harabaz Index, Davies-Bouldin Index


THANK
YOU!