Enhancing Marketing Strategies through Data-Driven Customer Profiling

Team members:

Jithendra Bojedla - jbojedla@pdx.edu

Dileep Kumar Boyapati - dileepkb@pdx.edu

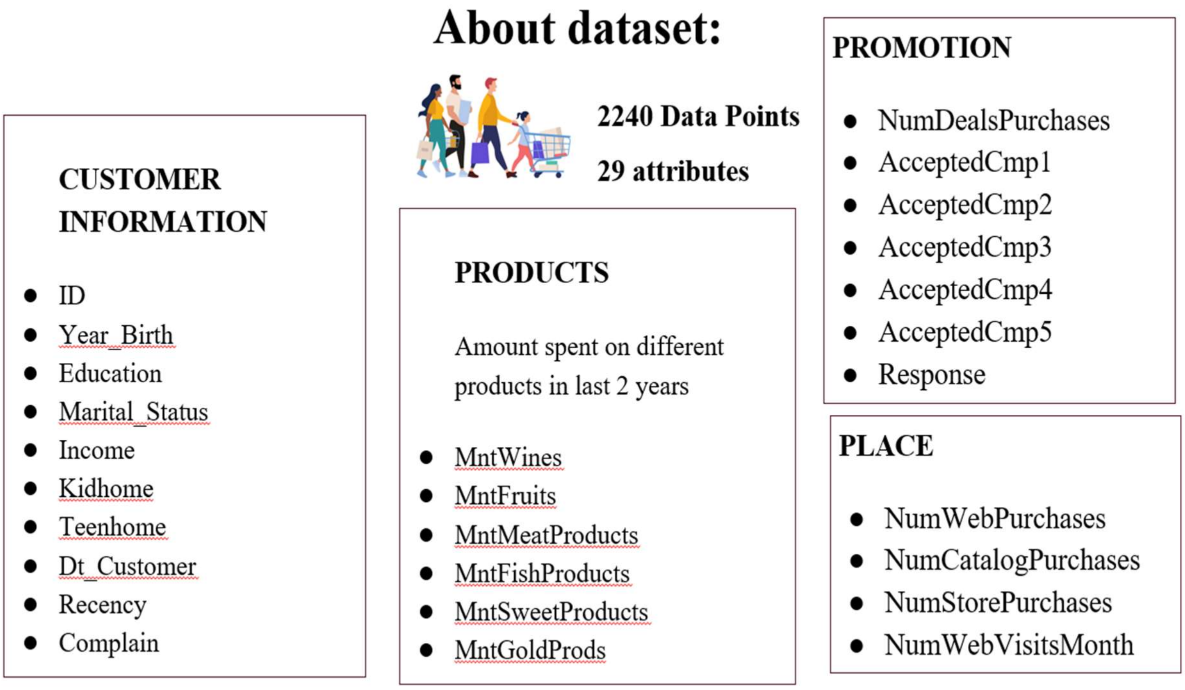
Lakshmi Sai Prasanna Addagalla - lakshmia@pdx.edu

Project Description

- This project aims to utilize unsupervised clustering techniques on customer records obtained from a groceries firm's database.
- Customer segmentation involves categorizing customers into groups based on similarities within each cluster.
- By dividing customers into segments, the goal is to enhance the relevance of each customer to the business.
- This enables customized product modifications to address the specific needs and behaviors of different customer groups, facilitating the business in addressing a diverse range of customer concerns effectively.

Dataset Description:

For this project, we have used the publicly available dataset from the Internet.



About dataset:

2240 Data Points

29 attributes

CUSTOMER INFORMATION

- ID
- Year_Birth
- Education
- Marital_Status
- Income
- Kidhome
- Teenhome
- Dt_Customer
- Recency
- Complain

PRODUCTS

Amount spent on different products in last 2 years

- MntWines
- MntFruits
- MntMeatProducts
- MntFishProducts
- MntSweetProducts
- MntGoldProds

PROMOTION

- NumDealsPurchases
- AcceptedCmp1
- AcceptedCmp2
- AcceptedCmp3
- AcceptedCmp4
- AcceptedCmp5
- Response

PLACE

- NumWebPurchases
- NumCatalogPurchases
- NumStorePurchases
- NumWebVisitsMonth

Implementation Details:

Run the source code preferably in Google Colabaratory.

Here is the breakdown of our code into various steps in order to complete the project:

1. Importing libraries
2. Loading data
3. Data cleaning
4. Data Preprocessing
5. Dimensionality Reduction
6. Clustering
7. Evaluating models
8. Profiling

Stepwise Description:

Importing libraries

We've imported various libraries necessary for data analysis and visualization, including:

- numpy and pandas for data manipulation.
- datetime for handling dates and times.
- matplotlib and seaborn for data visualization.
- LabelEncoder, StandardScaler, PCA, KMeans, AgglomerativeClustering from sklearn for machine learning tasks.
- KElbowVisualizer from yellowbrick for visualizing clustering.
- Necessary modules for 3D plotting and color mapping.
- metrics from sklearn for evaluating clustering performance.

Data Cleaning:

Initial Inspection:

- The dataset contains 2240 entries and 29 columns.
- Notable features include missing values in the Income column, and Dt_Customer not being parsed as datetime.

Handling Missing Values:

- Rows with missing Income values were dropped, reducing the dataset to 2216 entries.

Parsing Dates:

- The Dt_Customer column was converted to datetime.
- The range of customer registration dates was identified, from 2012-01-08 to 2014-12-06.

Creating New Features:

- Customer_For: Number of days a customer has been registered, calculated relative to the most recent customer.
- Age: Calculated from Year_Birth.
- Spent: Total spending across various product categories.
- Living_With: Simplified marital status to indicate if a customer is living with a partner or alone.
- Children: Total number of children in the household, combining Kidhome and Teenhome.
- Family_Size: Total members in the household.
- Is_Parent: Binary feature indicating parenthood.
- Simplified Education into three categories: Undergraduate, Graduate, Postgraduate.

Renaming Columns:

- Renamed columns for better readability (e.g., MntWines to Wines).

Dropping Redundant Features:

- Removed features that are no longer needed or redundant: Marital_Status, Dt_Customer, Z_CostContact, Z_Revenue, Year_Birth, ID.

Exploratory Data Analysis (EDA)

Descriptive Statistics:

- Examined the statistics of the cleaned data, noting some discrepancies in mean and maximum values for Income and Age.
- This involves computing summary statistics such as mean, median, mode, standard deviation, range, and percentiles for numerical variables in the dataset.
- These statistics help you understand the central tendency, dispersion, and shape of the data distribution.

Visualizing Data:

- Plotted selected features (Income, Recency, Customer_For, Age, Spent, Is_Parent) to get a broader view of the data.
- Visualization techniques such as histograms, box plots, scatter plots, and heatmaps are used to gain insights into the distribution, patterns, and relationships within the data.
- Visualizations make it easier to identify trends, outliers, and potential correlations.

Handling Outliers:

- Removed outliers by capping Age at 90 and Income at 600,000, reducing the dataset to 2212 entries.
- Outliers are data points that deviate significantly from the rest of the data. They can skew statistical measures and affect the performance of machine learning models.

Correlation Analysis:

- Generated a correlation matrix to study the relationships among numerical features.
- Correlation analysis examines the strength and direction of the relationship between pairs of variables.

Data Preprocessing:

Identify Categorical Values:

- Categorical values are those that represent categories or labels, rather than numerical values.
- Identifying which columns in the dataset contain categorical values is important for further processing.

Label Encode Categorical Variables:

- Label encoding is a process of converting categorical values into numerical labels.
- Each unique category is assigned a unique integer.
- This step is necessary for many machine learning algorithms that require numerical input.

Create a Copy of the Data:

- Creating a copy of the original dataset is a good practice before making any modifications.
- This ensures that we have a backup of the original data in case we need to refer back to it or compare the changes.

Create a Subset by Dropping Specific Features:

- Sometimes, we might want to work with a subset of your data by removing specific features that are not relevant to your analysis or modelling task.
- Dropping these features can simplify the dataset and improve computational efficiency.

Scale the Features:

- Feature scaling is a pre-processing step that standardizes the range of independent variables or features in your dataset.
- Common techniques include Min-Max scaling (scaling features to a range of [0, 1]) and Standardization (scaling features to have mean 0 and standard deviation 1).

Dimensionality Reduction:

In this project, we have numerous features influencing the final classification, making it challenging to handle high-dimensional data. Many of these features are correlated and redundant. To address this, we performed dimensionality reduction using Principal Component Analysis (PCA). This technique extracts a smaller set of principal variables while retaining as much relevant information as possible, simplifying the dataset and improving interpretability for classification tasks.

Initialize PCA:

PCA which specifies the desired number of principal components. This means we want to reduce the dimensionality of the dataset to 3 dimensions.

Fit PCA to Scaled Data:

The PCA model is fitted to the scaled dataset. This step calculates the principal components based on the correlation structure of the input features. Scaling the data before applying PCA is important to ensure that features with larger scales do not dominate the principal components.

Transform the Data:

The fitted PCA model is used to transform the scaled dataset into the reduced-dimensional space defined by the principal components. This step projects the data onto the new basis vectors formed by the principal components.

Descriptive Statistics:

Descriptive statistics are computed for the transformed data. This includes statistics such as count, mean, standard deviation, minimum, 25th percentile, median 75th percentile, and maximum for each of the three principal components. These statistics provide insights into the distribution and variability of the new features obtained after dimensionality reduction.

3D Projection Plot:

The reduced dataset is visualized in a 3D projection plot. Each point in the plot represents an observation in the reduced dataset, and its coordinates are determined by the values of the three principal components. This plot helps in understanding the spatial distribution of data points in the reduced-dimensional space and can reveal potential patterns or clusters.

Clustering:

- Determine the optimal number of clusters using the Elbow Method with K Means.
- Initialize the Agglomerative Clustering model with optimal number of clusters obtained from Elbow Method.
- Add the predicted cluster labels to the PCA-transformed dataset and the original dataset.
- Visualize the clusters in a 3D scatter plot.

K-means Clustering with Elbow Method:

Objective: Identify the optimal number of clusters to use in our clustering algorithm.

Process:

- Utilized the Elbow Method in conjunction with the K-means clustering algorithm.
- Plotted the distortion scores (sum of squared distances from each point to its assigned centroid) for a range of cluster counts (k = 1 to 10).
- Determined the optimal number of clusters (k) where the elbow point occurs, indicating a balance between minimizing within-cluster variance and avoiding overfitting.
- The Elbow Method suggested that 4 clusters would be optimal.

Agglomerative Clustering:

Objective: Group the data points into clusters based on their similarities.

Process:

- Implemented Agglomerative Clustering with the number of clusters set to 4 (as determined by the Elbow Method).
- Fitted the model to the PCA-transformed data and predicted the cluster assignments for each data point.

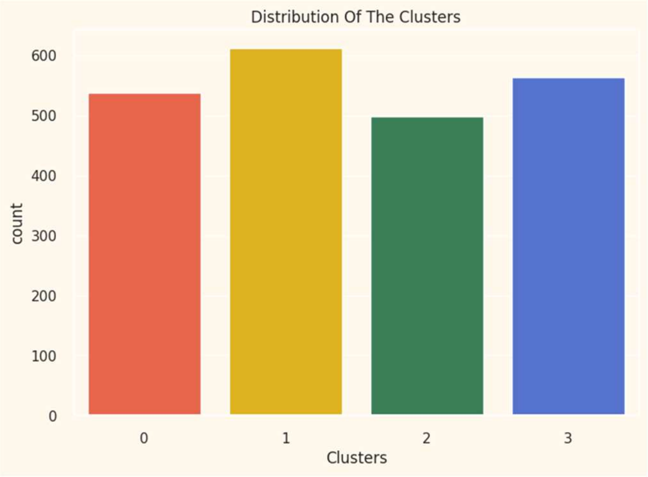
- This hierarchical clustering method progressively merged smaller clusters into larger ones, ultimately forming 4 distinct clusters.

Evaluating Models:

To explore patterns in clusters formed by unsupervised clustering and draw meaningful conclusions without labelled data.

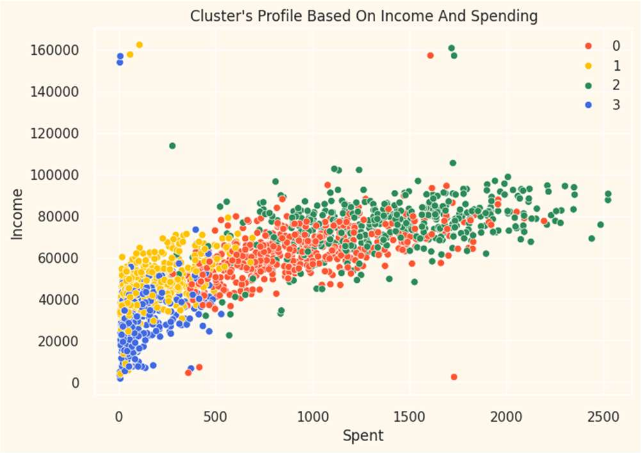
Cluster Distribution:

Clusters are fairly distributed, indicating a reasonable segmentation of data points.



Income vs. Spending Profile:

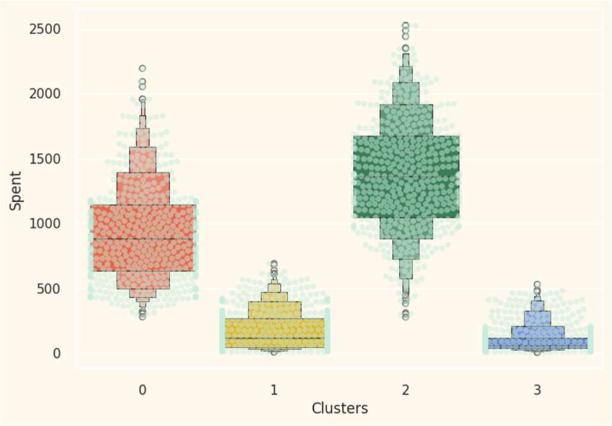
Analyzed the income and spending patterns within each cluster, helping to profile each cluster.



- Group 0: High spending, average income
- Group 1: High spending, low income
- Group 2: High spending, High income
- Group 3: Low spending, low income

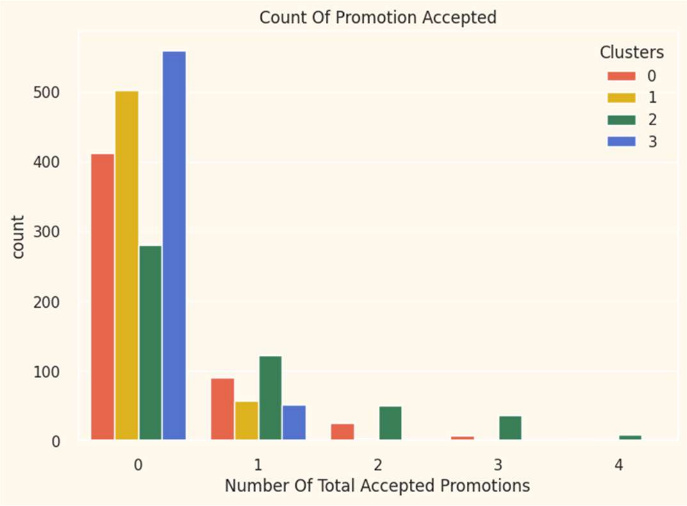
Spending Distribution:

Cluster 2 has the highest spending, closely followed by Cluster 0. This insight is useful for targeted marketing strategies.



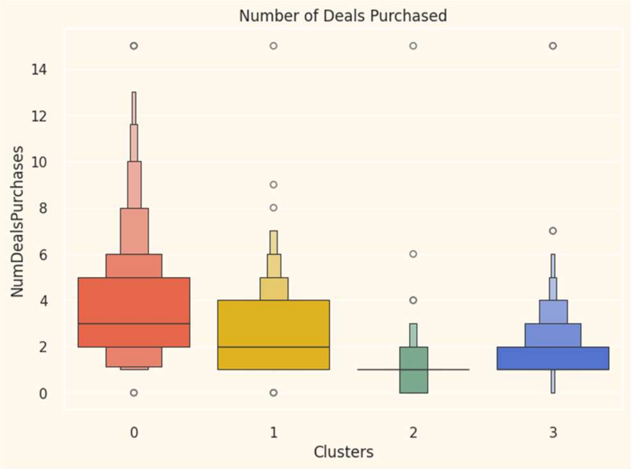
Campaign Response Analysis:

Low overall campaign response, with no cluster participating in all five campaigns. Suggests a need for better-targeted campaigns.



Deals Purchased Analysis:

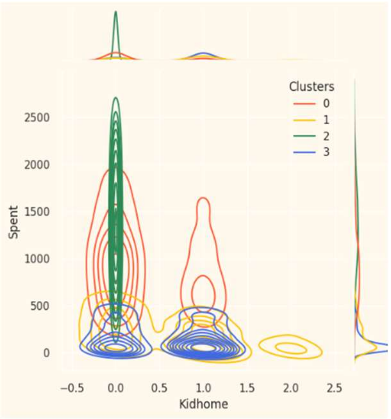
Deals performed well with Clusters 0 and 1, while Cluster 2 (the highest spenders) show little interest in deals. Cluster 3 has minimal interest in deals.

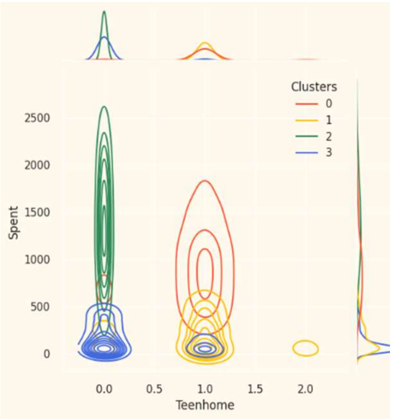


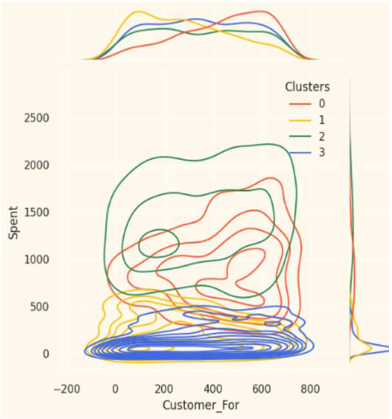
Profiling:

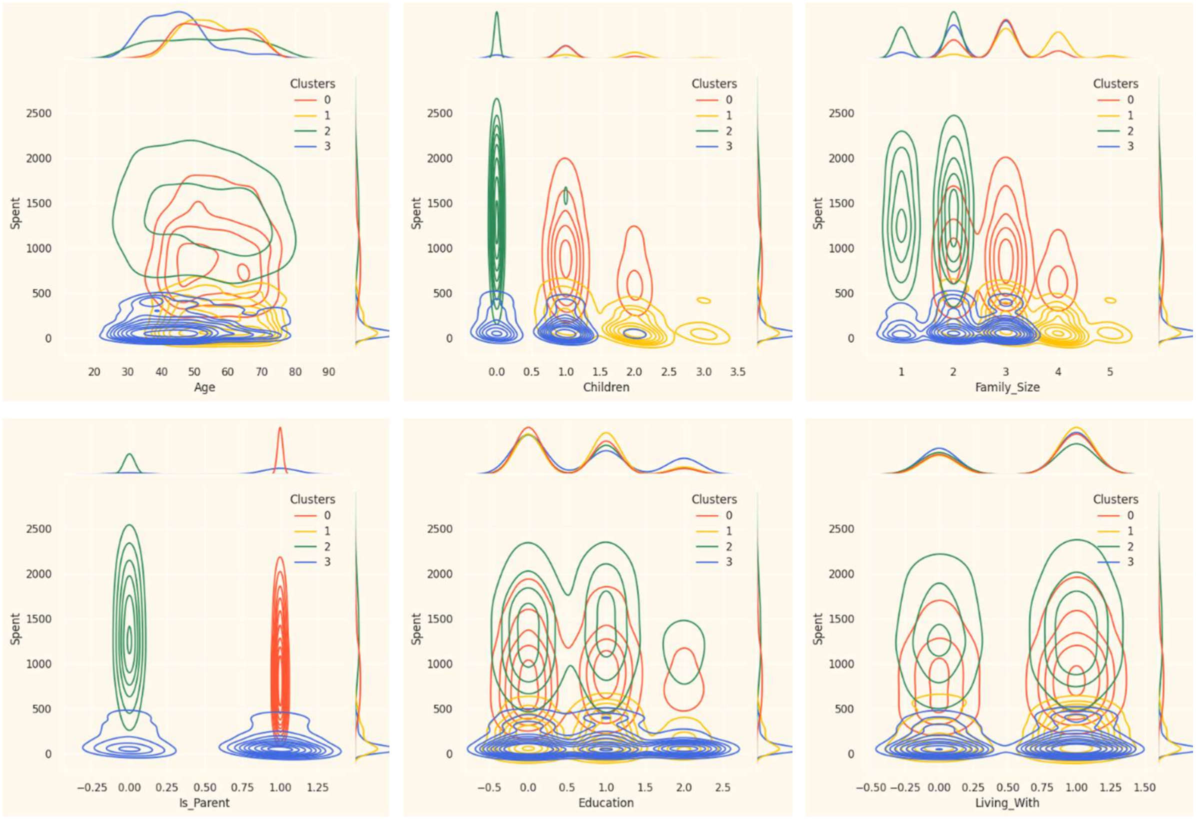
From the clusters and purchase habits. To look who is present in each of these clusters to determine who requires greater attention from the retail store's marketing team, we will be profiling the clusters that have formed.

To determine the customer's personal attributes based on the cluster they belong to; we plotted few features and drawn the conclusions considering the results.

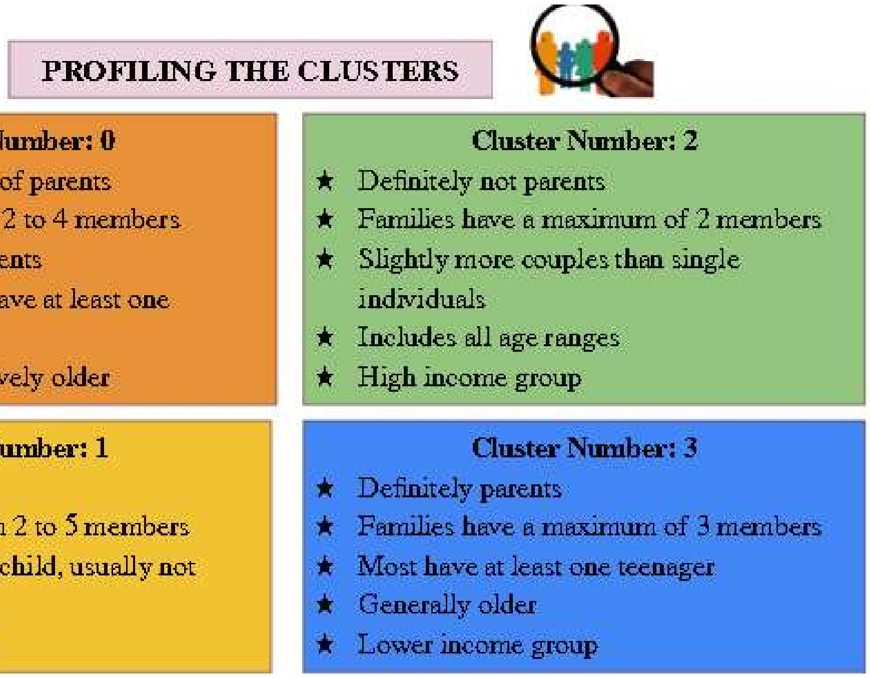








The following details can be inferred about the customers in 4 clusters:



PROFILING THE CLUSTERS

Cluster Number: 0

★ Definitely consists of parents
★ Family range from 2 to 4 members
★ Includes single parents
★ Most households have at least one teenager
★ Members are relatively older

Cluster Number: 2

★ Definitely not parents
★ Families have a maximum of 2 members
★ Slightly more couples than single individuals
★ Includes all age ranges
★ High income group

Cluster Number: 1

★ Mostly parents
★ Families range from 2 to 5 members
★ Typically have one child, usually not teenagers
★ Generally younger

Cluster Number: 3

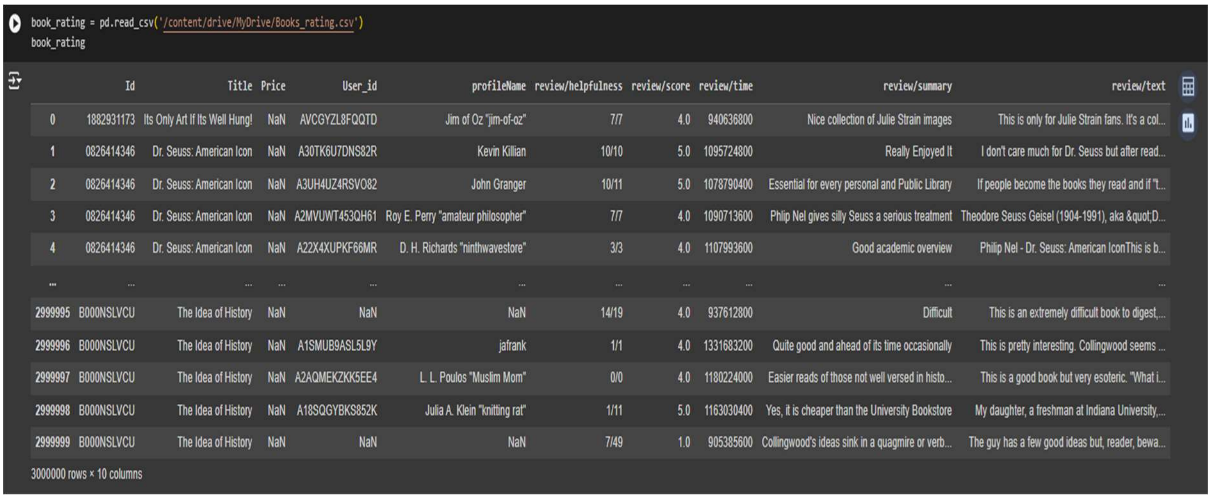
★ Definitely parents
★ Families have a maximum of 3 members
★ Most have at least one teenager
★ Generally older
★ Lower income group

Conclusion:

In this project, we analysed customer data using unsupervised clustering techniques and the clusters were then profiled based on factors like family structure, income, and spending habits.

This analysis can help businesses tailor their marketing strategies to different customer segments, ultimately improving customer engagement and driving better results.

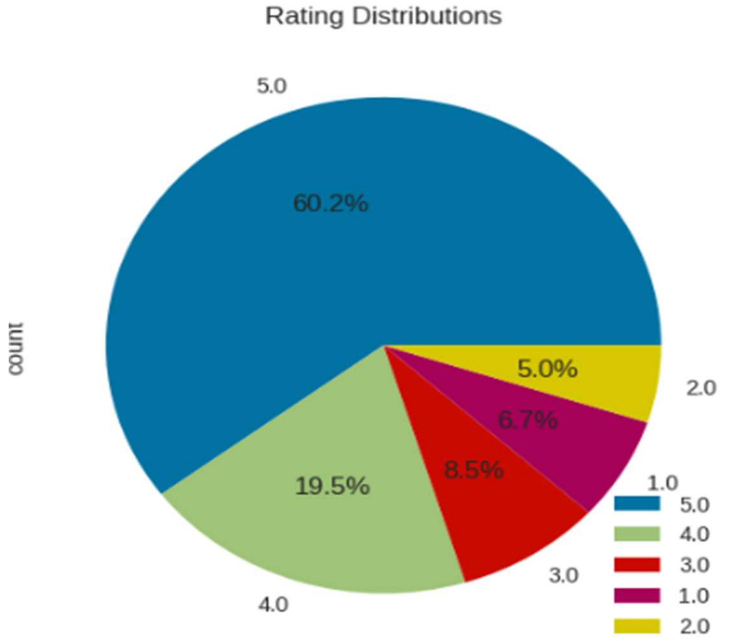
Extension: Amazon Review Dataset



```
book_rating = pd.read_csv('/content/drive/MyDrive/Books_rating.csv')
book_rating
```

	Id	Title	Price	User_id	profileName	review/helpfulness	review/score	review/time	review/summary	review/text
0	1882931173	Its Only Art If Its Well Hung!	NaN	AVCGYZL8FQQTD	Jim of Oz "jim-of-oz"	7/7	4.0	940636800	Nice collection of Julie Strain images	This is only for Julie Strain fans. It's a col...
1	0826414346	Dr. Seuss: American Icon	NaN	A30TK6U7DNS82R	Kevin Killian	10/10	5.0	1095724800	Really Enjoyed It	I don't care much for Dr. Seuss but after read...
2	0826414346	Dr. Seuss: American Icon	NaN	A3UH4UZ4RSVO82	John Granger	10/11	5.0	1078790400	Essential for every personal and Public Library	If people become the books they read and if "t...
3	0826414346	Dr. Seuss: American Icon	NaN	A2MVUWT453QH61	Roy E. Perry "amateur philosopher"	7/7	4.0	1090713600	Phlip Nel gives silly Seuss a serious treatment	Theodore Seuss Geisel (1904-1991), aka &quot;D...
4	0826414346	Dr. Seuss: American Icon	NaN	A22X4XUPKF66MR	D. H. Richards "ninthwavestore"	3/3	4.0	1107993600	Good academic overview	Philip Nel - Dr. Seuss: American IconThis is b...
...	...	...	...	...	...	...	...	...	...	...
2999995	B000NSLVCU	The Idea of History	NaN	NaN	NaN	14/19	4.0	937612800	Difficult	This is an extremely difficult book to digest,...
2999996	B000NSLVCU	The Idea of History	NaN	A1SMUB9ASL5L9Y	jafrank	1/1	4.0	1331683200	Quite good and ahead of its time occasionally	This is pretty interesting. Collingwood seems ...
2999997	B000NSLVCU	The Idea of History	NaN	A2AQMEKZKK5EE4	L. L. Poulos "Muslim Mom"	0/0	4.0	1180224000	Easier reads of those not well versed in histo...	This is a good book but very esoteric. "What i...
2999998	B000NSLVCU	The Idea of History	NaN	A18SQGYBKS852K	Julia A. Klein "knitting rat"	1/11	5.0	1163030400	Yes, it is cheaper than the University Bookstore	My daughter, a freshman at Indiana University,...
2999999	B000NSLVCU	The Idea of History	NaN	NaN	NaN	7/49	1.0	905385600	Collingwood's ideas sink in a quagmire or verb...	The guy has a few good ideas but, reader, bewa...

3000000 rows × 10 columns



Text Preprocessing:

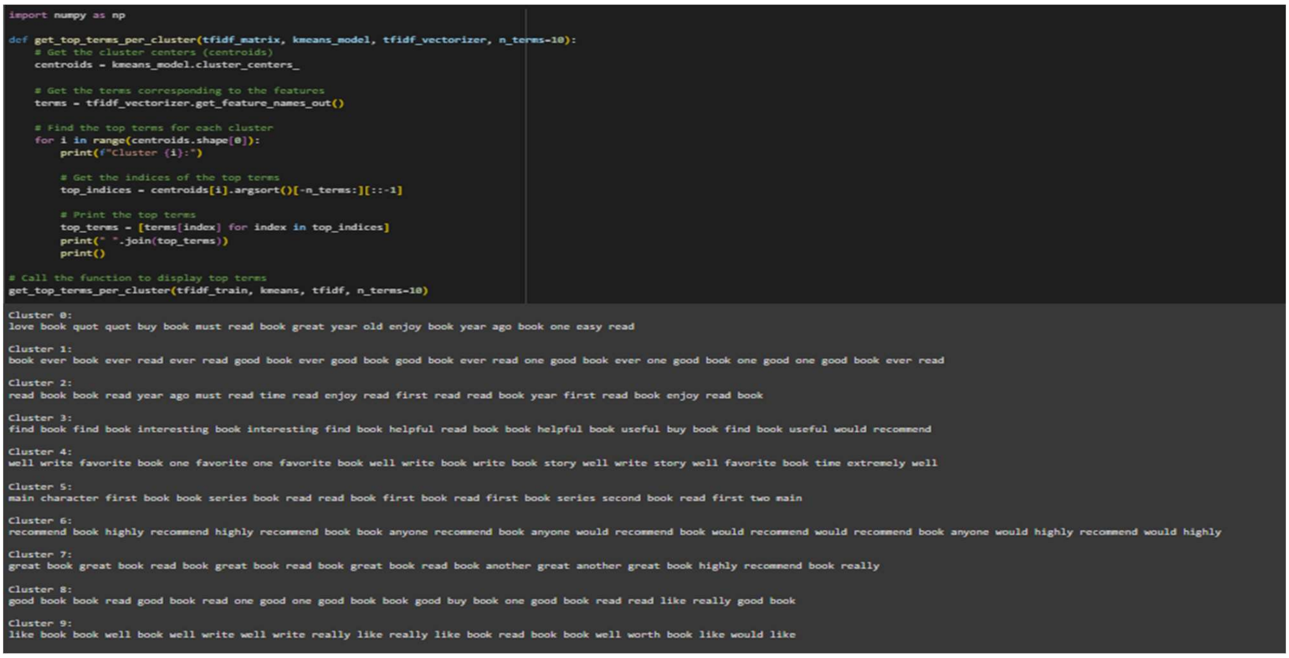
Preprocessing Steps:

- Lowercase Text: Ensures uniformity.
- Handled Accented Characters: Used unidecode to normalize text.
- Expanded Contractions: Used the contractions library to expand words like "can't" to "cannot".
- Tokenized Text and Removed Punctuation: Break text into words and remove unnecessary punctuation.
- Removed Stopwords: Excluded common words like "the", "and" that do not carry significant meaning.
- Lemmatization: Reduced words to their base form using spacy.

TF-IDF Vectorization:

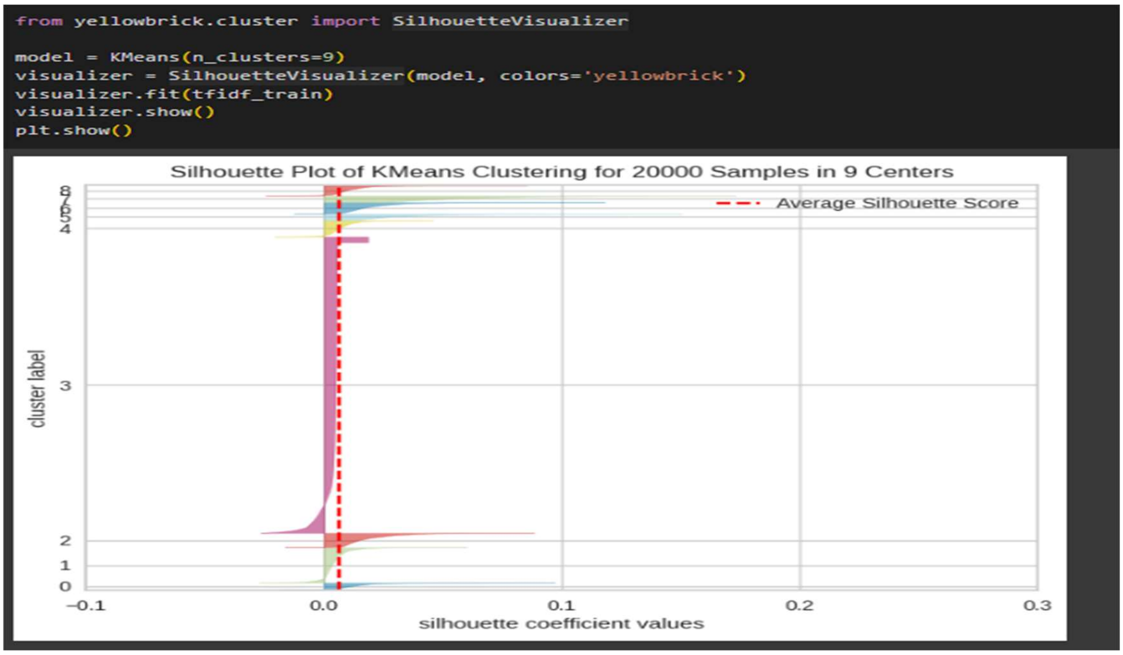
- TF-IDF (Term Frequency-Inverse Document Frequency): A technique to convert text data into numerical features.
- Focuses on the importance of words based on their frequency in the document and across all documents.

Results:



```
import numpy as np

def get_top_terms_per_cluster(tfidf_matrix, kmeans_model, tfidf_vectorizer, n_terms=10):
    # Get the cluster centers (centroids)
    centroids = kmeans_model.cluster_centers_

    # Get the terms corresponding to the features
    terms = tfidf_vectorizer.get_feature_names_out()

    # Find the top terms for each cluster
    for i in range(centroids.shape[0]):
        print(f"Cluster {i}:")

        # Get the indices of the top terms
        top_indices = centroids[i].argsort()[-n_terms:][::-1]

        # Print the top terms
        top_terms = [terms[index] for index in top_indices]
        print(" ".join(top_terms))
        print()

# Call the function to display top terms
get_top_terms_per_cluster(tfidf_train, kmeans, tfidf, n_terms=10)
```

```
Cluster 0:
love book quot quot buy book must read book great year old enjoy book year ago book one easy read

Cluster 1:
book ever book ever read ever read good book ever good book good book ever read one good book ever one good book one good one good book ever read

Cluster 2:
read book book read year ago must read time read enjoy read first read read book year first read book enjoy read book

Cluster 3:
find book find book interesting book interesting find book helpful read book book helpful book useful buy book find book useful would recommend

Cluster 4:
well write favorite book one favorite one favorite book well write book write book story well write story well favorite book time extremely well

Cluster 5:
main character first book book series book read read book first book read first book series second book read first two main

Cluster 6:
recommend book highly recommend highly recommend book book anyone recommend book anyone would recommend book would recommend would recommend book anyone would highly recommend would highly

Cluster 7:
great book great book read book great book read book great book read book another great another great book highly recommend book really

Cluster 8:
good book book read good book read one good one good book book good buy book one good book read read like really good book

Cluster 9:
like book book well book well write well write really like really like book read book book well worth book like would like
```



```
from yellowbrick.cluster import SilhouetteVisualizer

model = KMeans(n_clusters=9)
visualizer = SilhouetteVisualizer(model, colors='yellowbrick')
visualizer.fit(tfidf_train)
visualizer.show()
plt.show()
```

Silhouette Plot of KMeans Clustering for 20000 Samples in 9 Centers
Average Silhouette Score
cluster label
8
7
6
5
4
3
2
1
0
−0.1
0.0
0.1
0.2
0.3
silhouette coefficient values