SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

CUSTOMER SEGMENTATION AND PROFILING USING CLUSTERING TECHNIQUES

PRESENTED TO
Dr. A. Shobanadevi
Vaskar Deka

PRESENTED BY
Pijush Pathak
RA2011027010152

A DATA ANALYSIS AND VISUALIZATION PROJECT

11-May-2023

Agenda

ABSTRACT

The project 'Customer Segmentation and Profiling using Clustering Technique' focuses on utilizing clustering techniques to group customers based on their shared characteristics. The project follows a systematic approach, starting with data collection from various sources, including demographic information, purchase history, and online behavior. The collected data is then preprocessed, including handling missing values, removing outliers, and standardizing variables. Next, relevant variables are selected to capture the key aspects of customer behavior. Clustering techniques, such as k-means, hierarchical clustering, or DBSCAN, are applied to identify distinct customer segments. Finally, customer profiles are created based on the characteristics of each segment, enabling businesses to tailor their marketing efforts, enhance customer experiences, and improve overall business performance.

Objectives

1/ Importing Libraries

2/ Loading Data

3/ Data Cleaning

4/ Data Preprocessing

5/ Dimensionality Reduction

6/ Clustering

7/ Evaluating Models

8/ Profiling

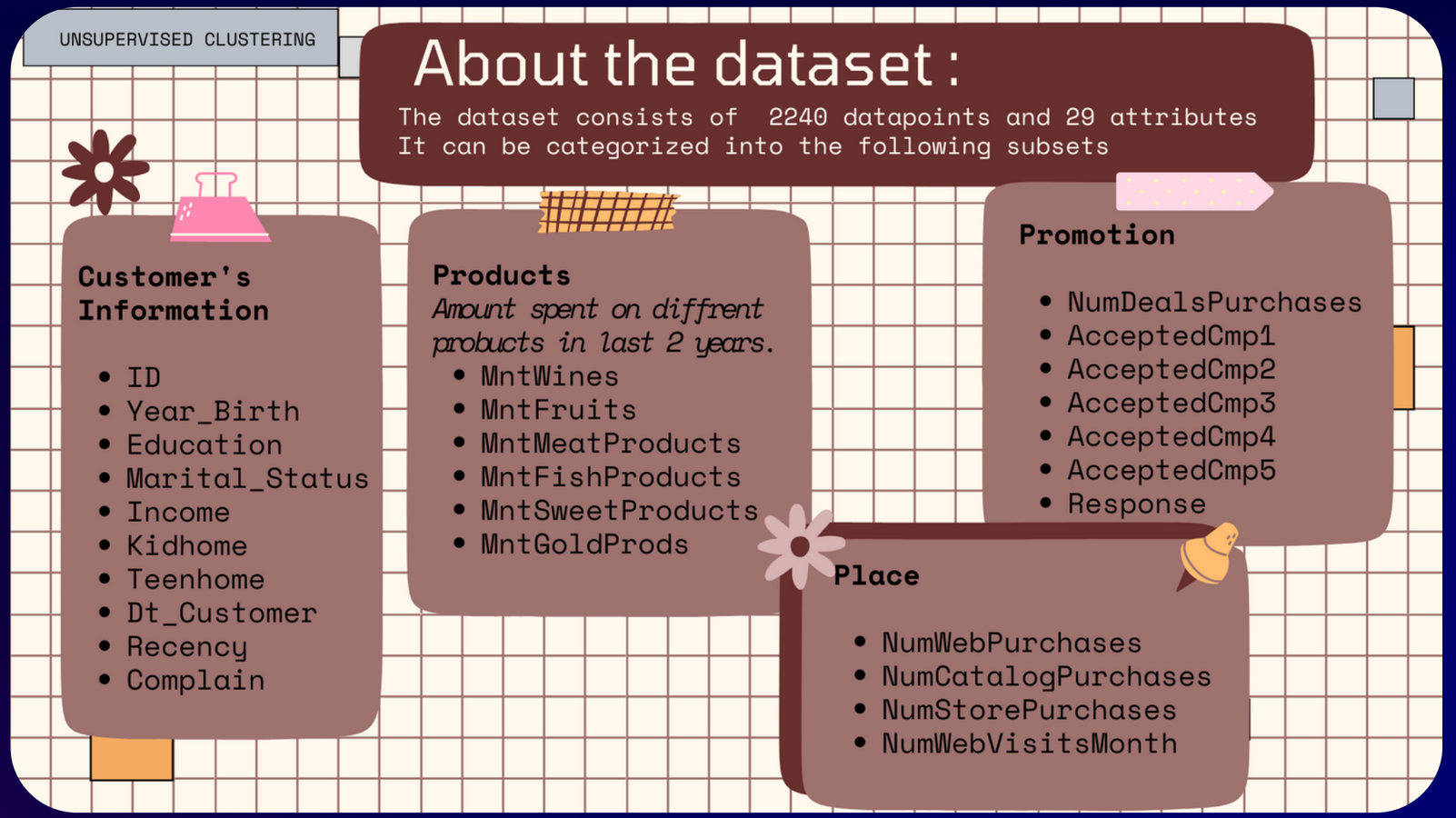
Importing Libraries

1	**Numpy**: It is a library for numerical computing in Python.
2	**Pandas**: It is a library for data manipulation and analysis
3	**Matplotlib**: It is a library for data visualization in Python
4	**Seaborn**: It is a library for creating more visually appealing and informative statistical graphics in Python.
5	**Label Encoder, Standard Library**: They are preprocessing techniques from the scikit-learn library for preparing data for machine learning models.
6	**KElbow Visualizer**: It helps to determine the optimal number of clusters for KMeans clustering algorithm.
7	**Axes3D**: It is a 3D plotting tool from matplotlib.

Loading Data

We shall load the dataset and find the number of datapoints.

The number of Data Points in the dataset is 2240



Data Cleaning

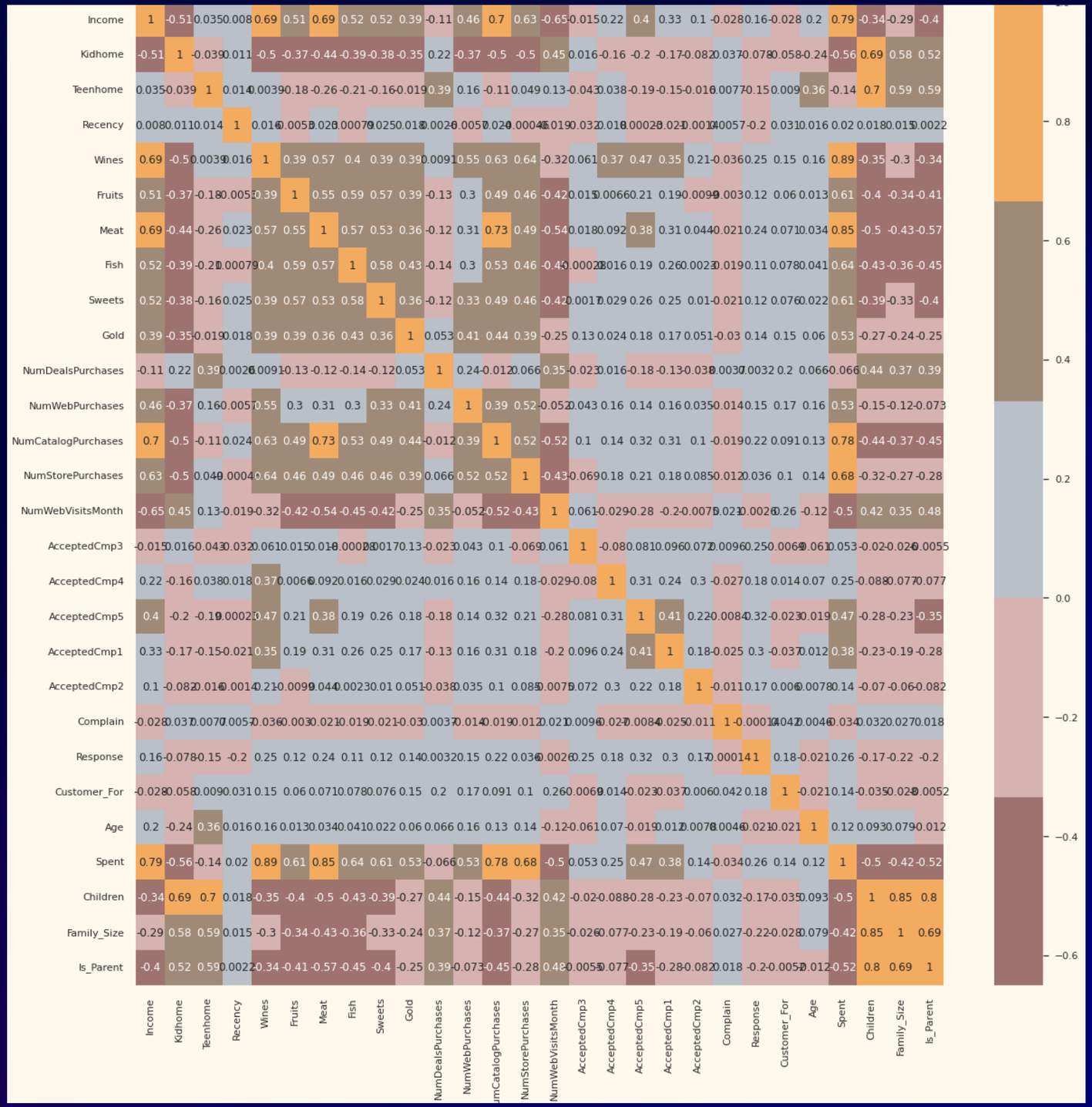
In order to, get a full grasp of what steps should be taking to clean the dataset.We shall have a look at the information in data. We found that there are missing values in income. First of all, for the missing values, I had dropped the rows that have missing income values.

In the next bit, I had performed the following steps to engineer some new features:

1. Extracted the "Age" of a customer by the "Year_Birth" that indicated the birth year of the respective person.
2. Created another feature "Spent" indicating the total amount spent by the customer in various categories over the span of two years.
3. Created another feature "Living_With" out of "Marital_Status" to extract the living situation of couples.
4. Created a feature "Children" to indicate total children in a household that is, kids and teenagers.
5. To get further clarity of household, I had created a feature indicating "Family_Size"
6. Created a feature "Is_Parent" to indicate parenthood status
7. Lastly, I had created three categories in the "Education" by simplifying its value counts. Dropping some of the redundant features

Correlation Matrix

It is the measure of strength and direction of the linear relationship between two variable, they range from -1 to 1



Data Preprocessing

Label Encoding the categorial features

Label encoding is a technique used to transform categorical data into numerical data. This is done to represent the data in a way that can be easily used by machine learning algorithms.

Scaling the features using the standard scalar

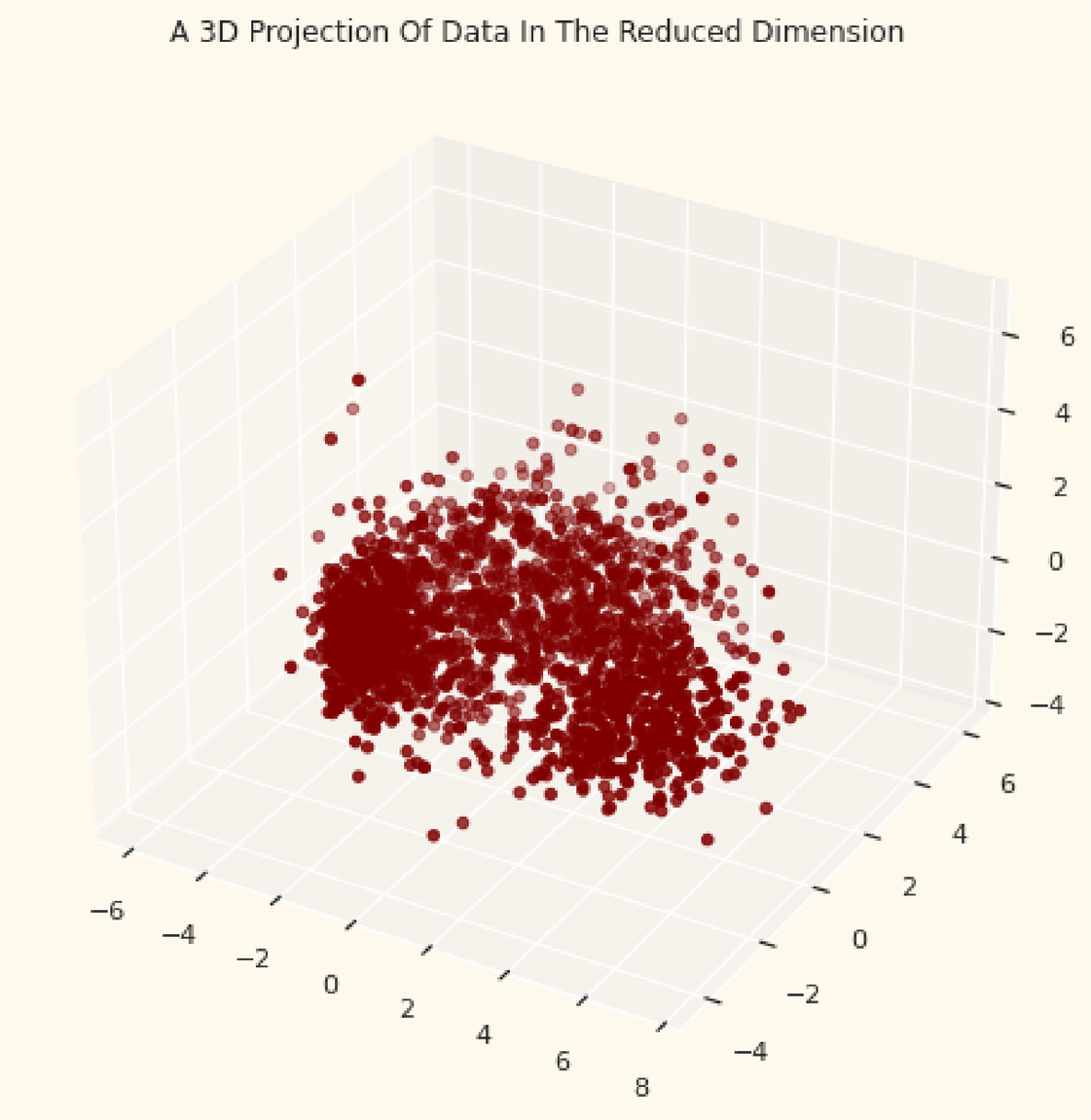
Scaling is a common data preprocessing step in machine learning that involves transforming features to have the same scale, typically with the goal of improving the performance of machine learning algorithms.

Creating a subset dataframe for dimentionality reduction

We create a subset of a dataset for dimensionality reduction when we want to reduce the number of features in the dataset while retaining as much of the original information as possible.

Dimentionality Reduction

We shall apply dimensionality reduction using PCA(Principle Component Analysis)
After applying PCA, we can visualize the data is the new reduces-dimension shape to gain insights into the structure of the data. In PCA, the original high dimension dataset is projected into the lower - dimension sub space



Data Preprocessing

Elbow method to determine the number of clusters to be formed

Elbow method is a common technique that is used to determine the optimal number of cluster for a given dataset. As the number of cluster increases, the within cluster Sum of Squared Error(SSE) decrease

Clustering via Agglomerative Clustering

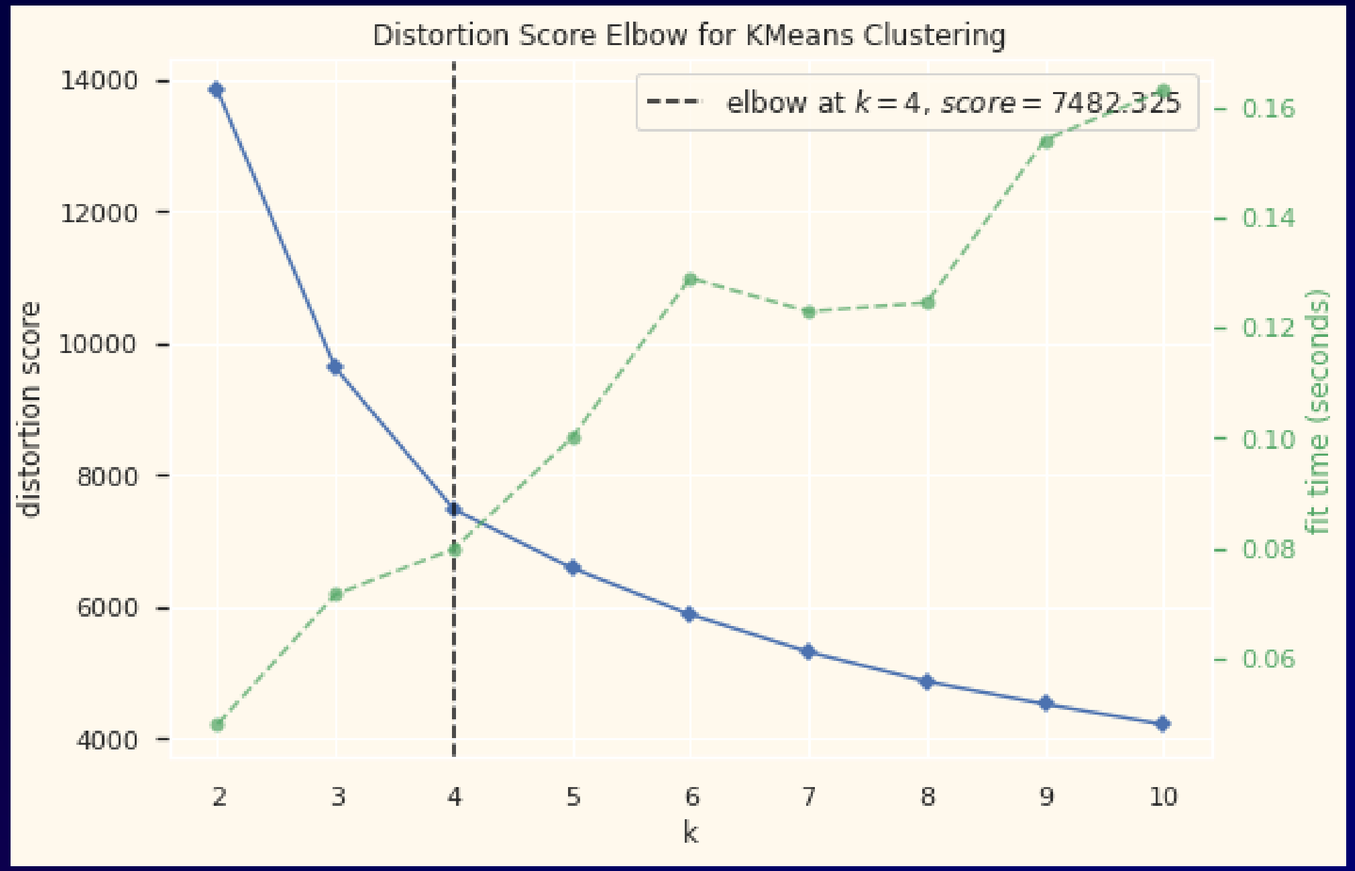
Agglomerative Clustering is a hierarchical clustering algorithm the starts by treating each data point as a separate cluster and then iteratively merges the closest pairs of clusters until all the data points belong to a single cluster.

Examining the cluster formed via Scatter Plot

I had examined the clusters formed by the 3-D distribution of the clusters.

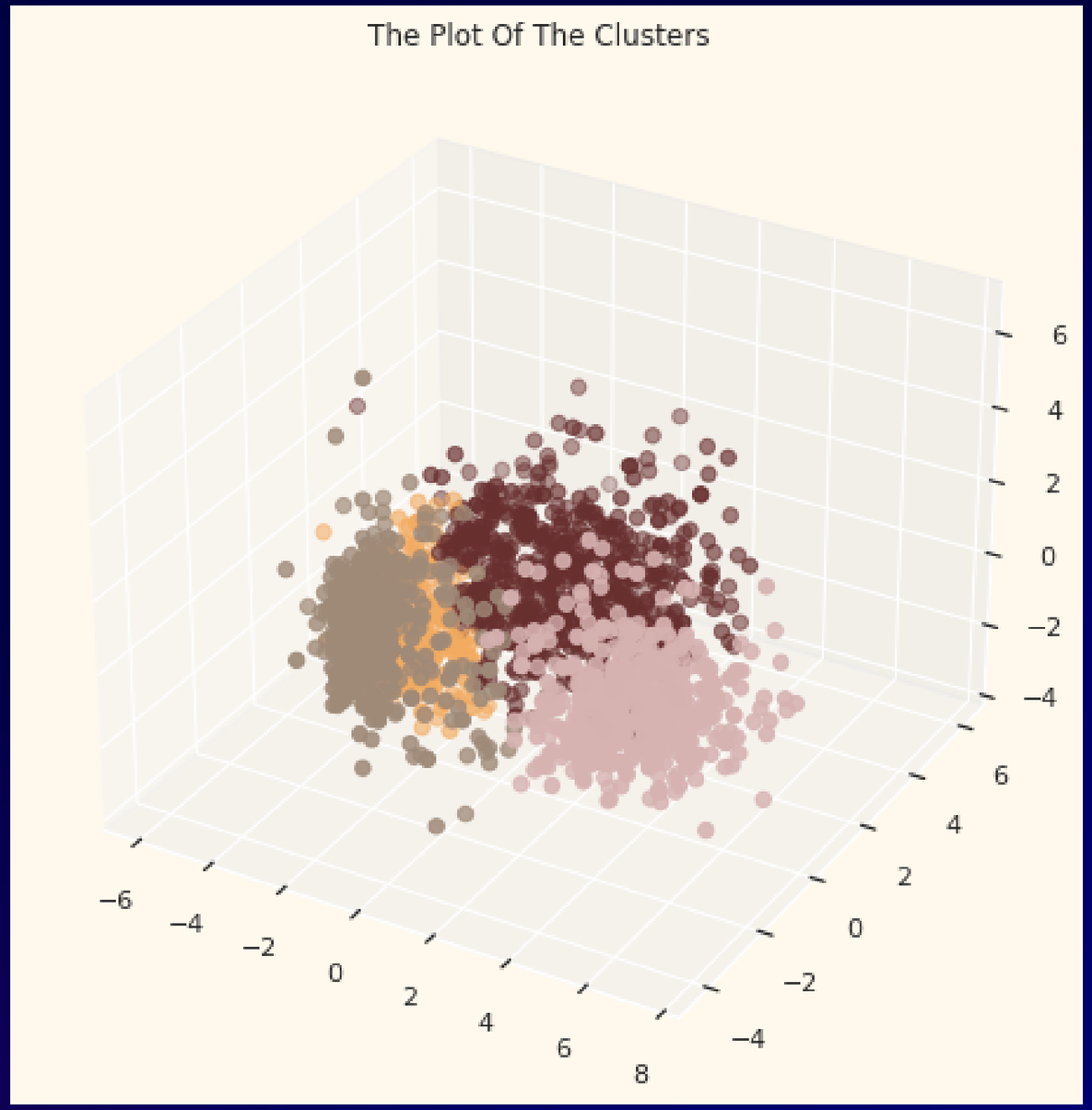
KElbowVisualizer

Elbow Method to determine the number of clusters to be formed.



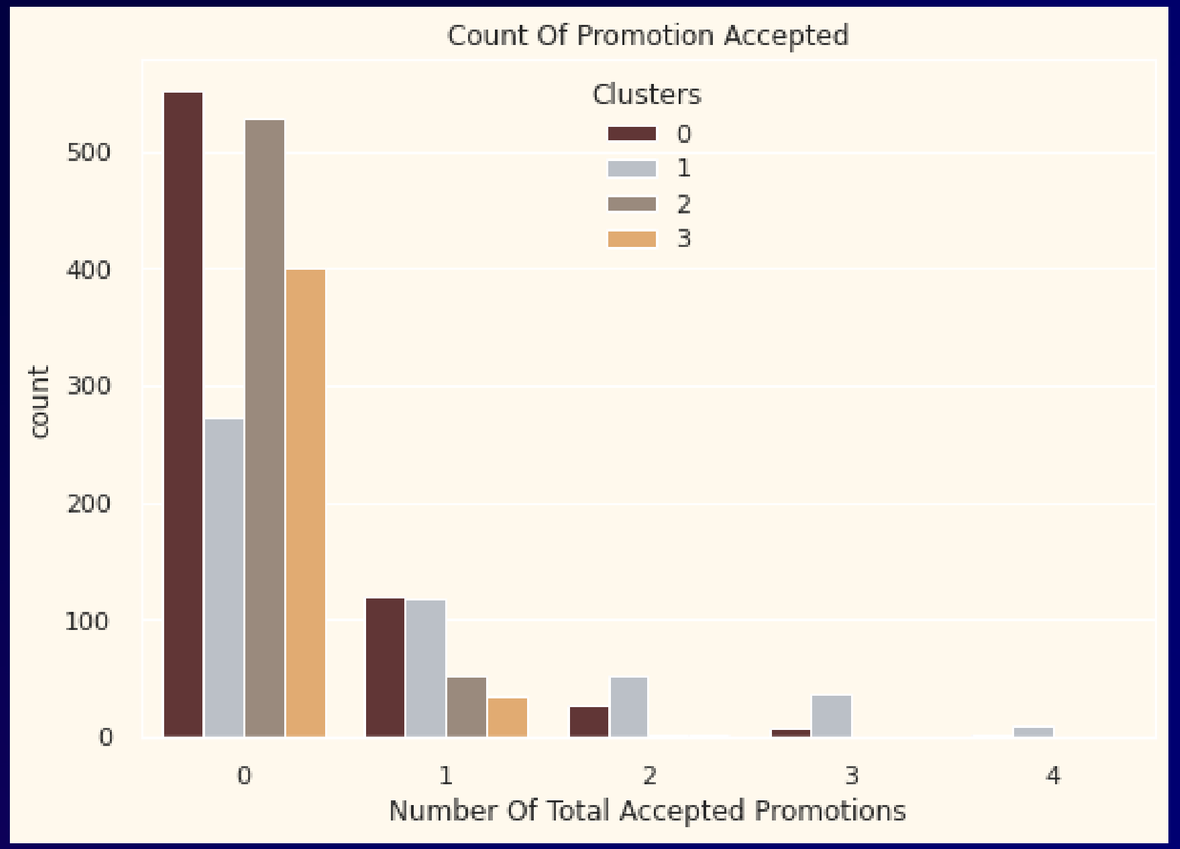
Scatter Plot

After applying clustering algorithms to a dataset, it is common to visualize the resulting clusters using scatter plots. A scatter plot can help to visualize how well the data points are separated into different clusters.



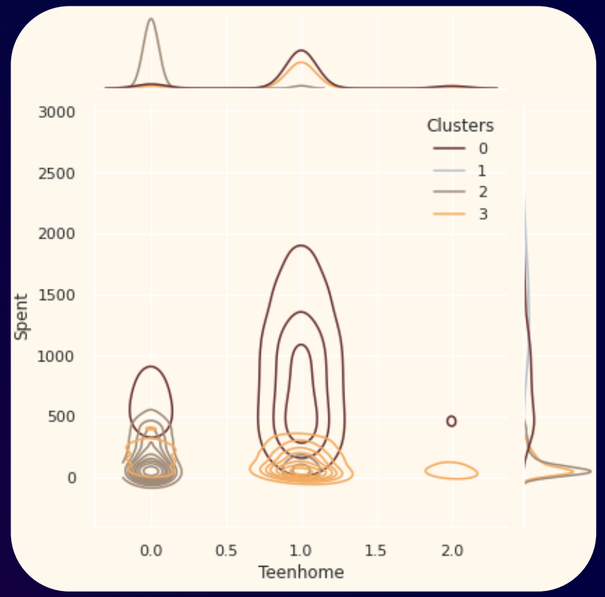
Evaluating Models

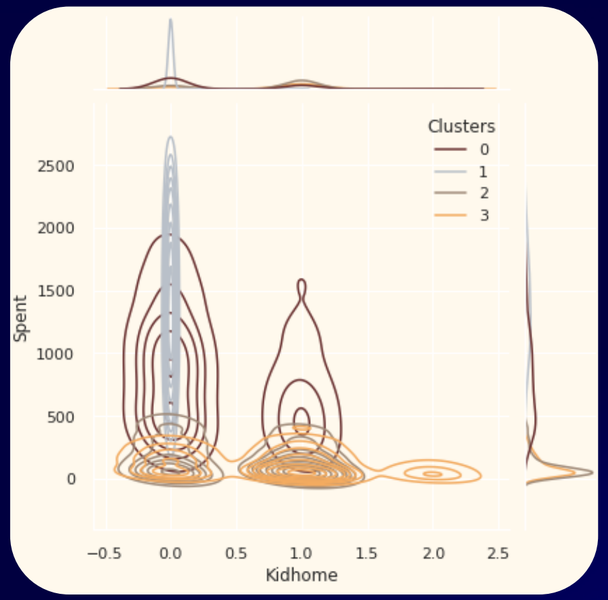
Since this is an unsupervised clustering. We do not have a tagged feature to evaluate or score our model. The purpose of this section is to study the patterns in the clusters formed and determine the nature of the clusters' patterns.

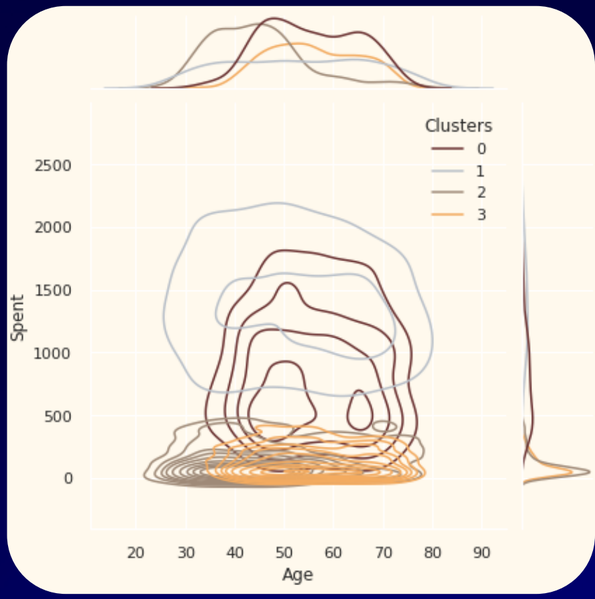


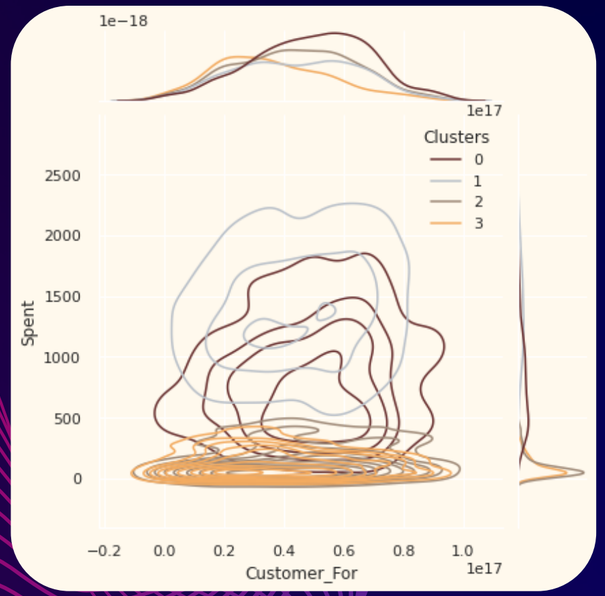
Profiling

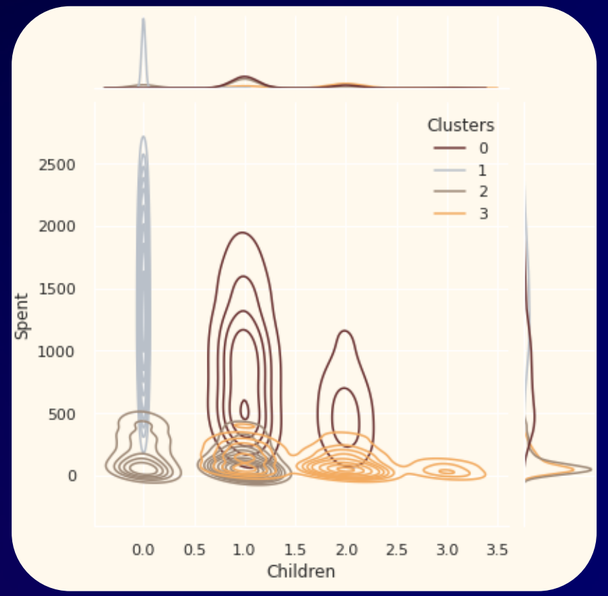
To decide the conclusion of the clustering I will be plotting some of the features that are indicative of the customer's personal traits in light of the cluster they are in. On the basis of the outcomes, I will be arriving at the conclusions.

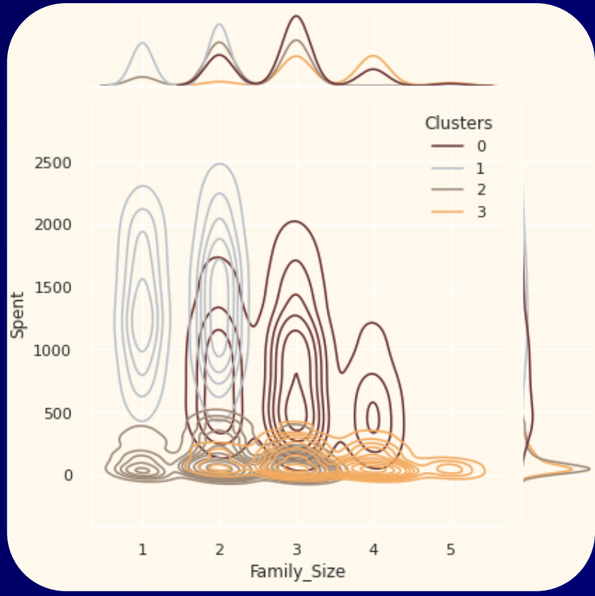












Points to be noted after profiling

Profiling The Clusters

0

About Cluster Number : 0

- Are a definitely a parent
- At the max have 4 members in the family and at least 2
- Single parents are a subset of this group
- Most have a teenager at home
- Relatively older

1

About Cluster Number : 1

- Are a definitely not a parent
- At the max are only 2 members in the family
- A slight majority of couples over single people
- Span all ages
- A high income group

2

About Cluster Number : 2

- The majority of these people are parents
- At the max are 3 members in the family
- They majorly have one kid (and not teenagers, typically)
- Relatively younger

3

About Cluster Number : 3

- They are definitely a parent
- At the max are 5 members in the family and at least 2
- Majority of them have a teenager at home
- Relatively older
- A lower-income group

UNSUPERVISED CLUSTERING

Conclusion

In conclusion, the project 'Customer Segmentation and Profiling using Clustering Technique' is a valuable approach for businesses to gain insights into their customer base. By using unsupervised machine learning to group customers based on shared characteristics, businesses can tailor their marketing strategies and product offerings to better meet customer demands. This can lead to improved customer satisfaction and retention, which ultimately translates to increased revenue and a competitive advantage in the market. Overall, this project highlights the importance of utilizing data-driven techniques to understand and cater to customer needs in today's competitive business landscape.

Future Enhancement

1	One possibility is to incorporate additional data sources, such as social media activity, to gain a more comprehensive understanding of customer behavior and preferences. This could provide more nuanced insights into customer needs and allow for even more targeted marketing strategies.
2	Another potential enhancement is to explore more advanced clustering techniques, such as hierarchical clustering or density-based clustering. These techniques may be better suited for datasets with complex or non-linear relationships.
3	Additionally, incorporating predictive modeling techniques could help businesses anticipate future customer behavior and tailor their strategies accordingly. For example, businesses could use clustering to identify customers who are likely to churn and develop retention strategies to prevent this.