Customer segmentation from Customer Personality Analysis Dataset Using Unsupervised Learning

Capstone Project Spring 2023 by:
Tejal Deshmukh

Appendix

Introduction

The process of finding and comprehending each customer's distinctive qualities and attributes is known as customer personality analysis. Companies can utilize this data to customize their marketing and sales strategies to better target and meet the individual needs and preferences of each consumer.

Customers' personalities have traditionally been analyzed manually by marketing and sales teams, who would utilize their knowledge and experience to spot recurring patterns and trends. But now that data mining and machine learning have been developed, it is possible to automate this procedure by employing algorithms that can examine massive volumes of data and find recurring patterns and characteristics among clients.

Unsupervised learning is one kind of machine learning technique that can be applied to consumer personality analysis. Unsupervised learning algorithms can automatically find patterns and similarities among clients without being explicitly taught what to look for because they are trained on a lot of data. They are therefore especially well-suited for client personality assessments since they can spot small variances and patterns that people would not notice right away.

The use of unsupervised learning algorithms for customer personality analysis and how businesses can utilize them to better understand and serve their consumers will be covered in this project.

Customer Personality Analysis

Customer personality analysis refers to the process of analyzing the personality traits and characteristics of customers to better understand their behaviors, needs, and preferences. This analysis is typically done by examining customer data such as purchase history, browsing patterns, and social media activity, and using various techniques such as data mining, machine learning, and psychometric profiling to identify common personality traits among customers.

The goal of customer personality analysis is to gain insights into customer characteristics and preferences to better target marketing efforts, personalize customer experiences, and ultimately increase customer satisfaction and loyalty. By understanding the personality traits and characteristics of their customers, businesses can tailor their products, services, and marketing messages to better resonate with their target audience, resulting in increased sales and customer loyalty.

Overview of Dataset

The dataset utilized for this project is a public dataset from Kaggle provided by Dr. Omar Romero-Hernandez. This dataset has 2,240 rows of observations and 29 attributes in the dataset. Among the variables, there are 3-categorical variables and 26 numerical variables. The dataset has been collected from a grocery firm database over the course of years to record the data against certain features that could help us study the behavior of customers and better plan and strategize the marketing policies of the company.

The attributes are grouped into following 4 distinct categories for our understanding namely:

- Customer Information - These attributes have been used to gather data about Customers
  The attributes are : ID, Birth Year, Education, Marital Status, Income, Kid home or Teen home, Date of registration with the grocer firm for Customer, Recency, Complain

- Product Information - These attributes have been used to gather data about the Products on which a customer spent his/her money in the past 2 years.
  The product attributes are : Wines, Fruits, Meat, Fish, Sweet, Gold

- Promotion Information - These attributes have been used to present information about promotions conducted for the customer and the response of the customer towards the promotions
  The Promotion attributes are : Number of Deals Purchased with a Discount, Accepted Campaign 1, Accepted Campaign 2, Accepted Campaign 3, Accepted Campaigns 4, Accepted Campaign 5, Response

- Place Information - These attributes have been used to specify the mode of purchase/visits by a Customer to the company's website, catalog and stores.
  The Place attributes are : Web Purchases, Catalog Purchases, Store Purchases and Website Visits in the last Month

ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases
5524	1957	Graduation	Single	58138	0	0	04-09-2012	58	635	88	546	172	88	88	3	8	10	4
2174	1954	Graduation	Single	46344	1	1	08-03-2014	38	11	1	6	2	1	6	2	1	1	2
4141	1965	Graduation	Together	71613	0	0	21-08-2013	26	426	49	127	111	21	42	1	8	2	10
6182	1984	Graduation	Together	26646	1	0	10-02-2014	26	11	4	20	10	3	5	2	2	0	4
5324	1981	PhD	Married	58293	1	0	19-01-2014	94	173	43	118	46	27	15	5	5	3	6
7446	1967	Master	Together	62513	0	1	09-09-2013	16	520	42	98	0	42	14	2	6	4	10
965	1971	Graduation	Divorced	55635	0	1	13-11-2012	34	235	65	164	50	49	27	4	7	3	7
6177	1985	PhD	Married	33454	1	0	08-05-2013	32	76	10	56	3	1	23	2	4	0	4

Snippet of some variables of Raw Data

Implementation

The project has been implemented in below stages respectively.

1. Loading and Cleaning the dataset

Loading a dataset refers to the process of importing data from an external file or source into a program or tool for further analysis. This process is a crucial first step in data analysis and machine learning projects, as it allows researchers and practitioners to work with the data in a format that can be manipulated easily and analyzed.

Cleaning a dataset refers to the process of identifying and correcting or removing inaccurate, incomplete, or irrelevant data from a dataset. This process is important because it can significantly impact the results and insights obtained from the dataset. Cleaning the raw datasets an aid in avoiding bias which can skew the results leading to inaccurate conclusions. By removing the unwanted and incomplete data, the risk of bias can be minimize ensuring that the analysis is objective and unbiased.

Post loading marketing campaign dataset using pandas, data cleaning was performed to prepare it for analysis. The shape, first few rows and column names in the dataset have been printed to get an idea of how the data looks like. Next, a check for missing values was implemented and it was found that there are some missing values in the dataset for the attribute income. Since income plays a major role in customer segmentation as we progress in the project, an absence of income needs to be handled in this step of implementation. The rows with the missing values have been dropped using the dropna() function. Additionally, attributes related to product information in the dataset have been renamed for better readability and understanding using the rename() function.

2. Exploratory Data Analysis

Exploratory Data Analysis (EDA) in data analysis involves examining and understanding the characteristics of a dataset. The goal of EDA is to discover patterns, relationships, and anomalies in the data that can help inform subsequent modeling and analysis steps. EDA is a vital step in data analysis, as it can help identify potential issues or biases in the data, as well as highlight important relationships and patterns that can be used to inform further analysis. By understanding the characteristics of the data through EDA, it is possible to develop more accurate and effective models that can be used to make predictions or inform decision-making processes.

To help us understand the data in detail, Exploratory Data Analysis (EDA) was performed on the dataset. Analysis was performed on the income variable by plotting a histogram and it was found

that most customers have income in the range of 30000–800000 and mean income was around 53000. Further, a bar graph was plotted for every product purchased to check the expenditure by product category. It was found that customers spent the most on wine purchases followed by Meat. Attributes related to campaigns were also studied to find the impact of campaigns on customers. It was observed that more people accepted or responded to campaigns 3rd, 4th and 5th and lower numbers of people responded to campaign 1 and 2.

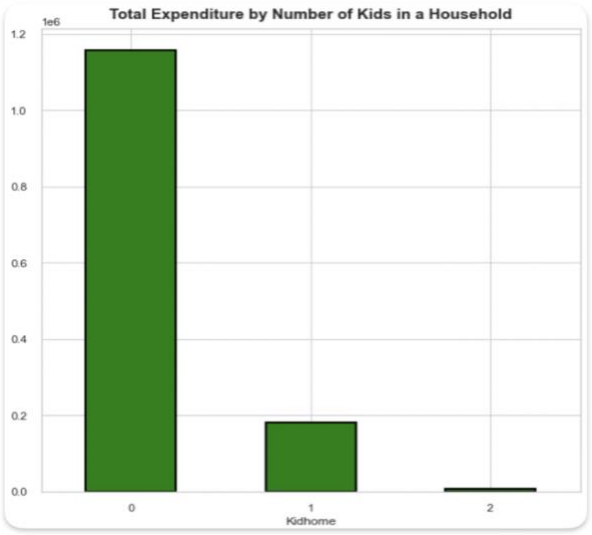
Further, the impact of customer education, marital status, teen, or kid household on the total expenditure of all the products purchased by a customer was also studied. Customers with a graduate degree had higher expense history followed by consumers with post graduate degree. Customers without a proper education in the form of graduate or postgraduate degrees had very lower expenses.

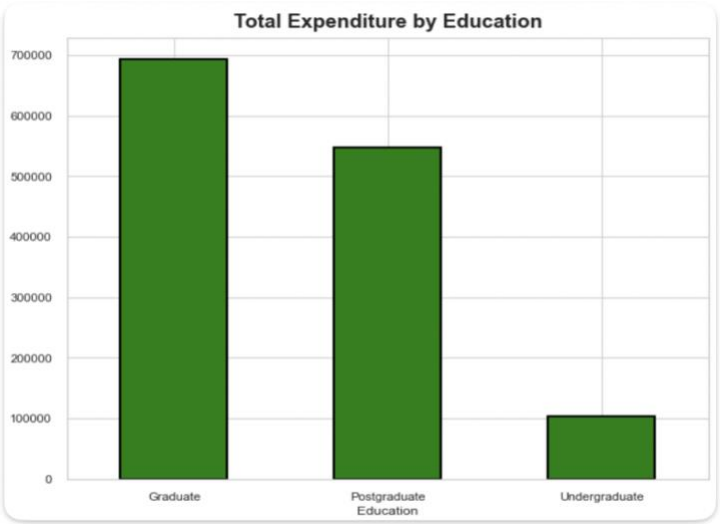
It was also observed that households without kids and teens had the highest expenditure compared to households with 1 or 2 kids or teens. Marital status also had a direct effect on customer spending. Customers who are married spend the most money followed by people who are engaged/living together/committed to a partner. Customers who are single/widowed/divorced had much lower expense histories as compared to married/committed couples.

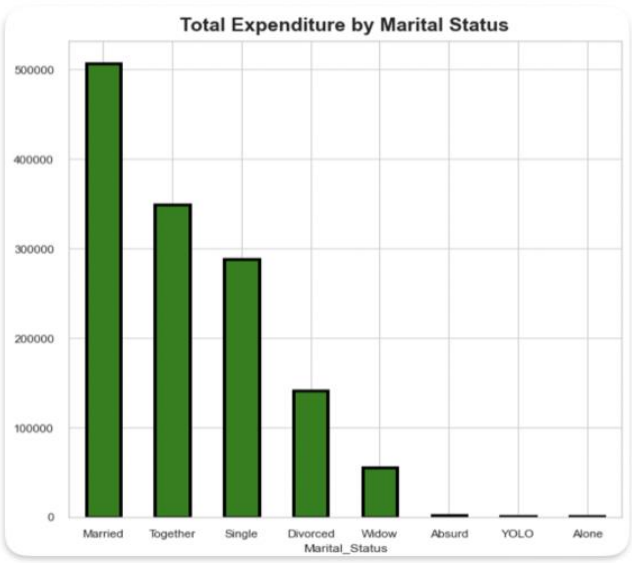
A study of the effect of customer education, marital status and teen households on campaign acceptance was also conducted to help understand the behavior of customers towards the marketing campaigns. It was observed that graduate degree holders had the highest campaign acceptance rate followed by postgraduate customers.

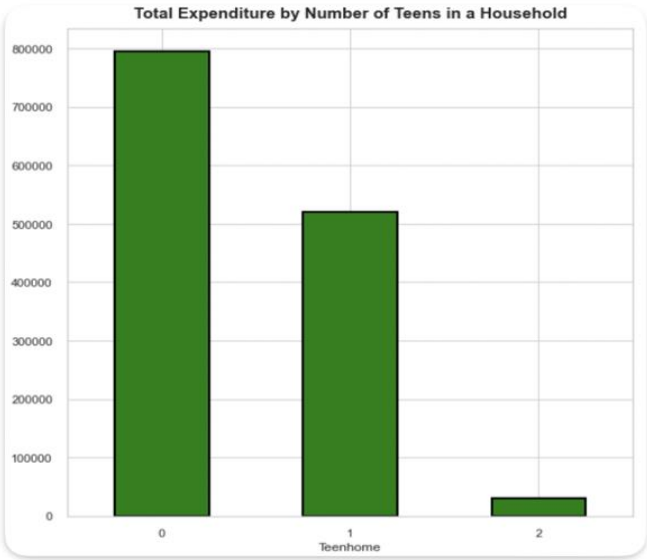
Additionally, it was found that, in comparison to households with one or two children or teenagers, households without children or teenagers spent the most money. Customer spending was also directly impacted by marital status. The customers with most expenses are those who are married, followed by those who are engaged, cohabiting, or devoted to a partner. In comparison to married or committed couples, customers who are single, widowed, or divorced had substantially lower expense histories.

The visualization plots have been created using a combination of two python libraries namely Seaborn and Matplotlib. Customization of the various attributes of both the libraries has been implemented to create the plot.









Snippets of EDA performed on raw dataset.

3. Feature Engineering

Feature engineering is an important step in the process of creating and training machine learning models. It involves the process of selecting, transforming, and extracting relevant features from raw data to improve the performance of the model. The goal of feature engineering is to identify the most informative and relevant features from the available data and transform them in a way that the machine learning algorithm can understand and utilize them effectively. This can include various techniques such as dimensionality reduction, data normalization, data imputation, and feature scaling which will be accentuated in the later part of the report.

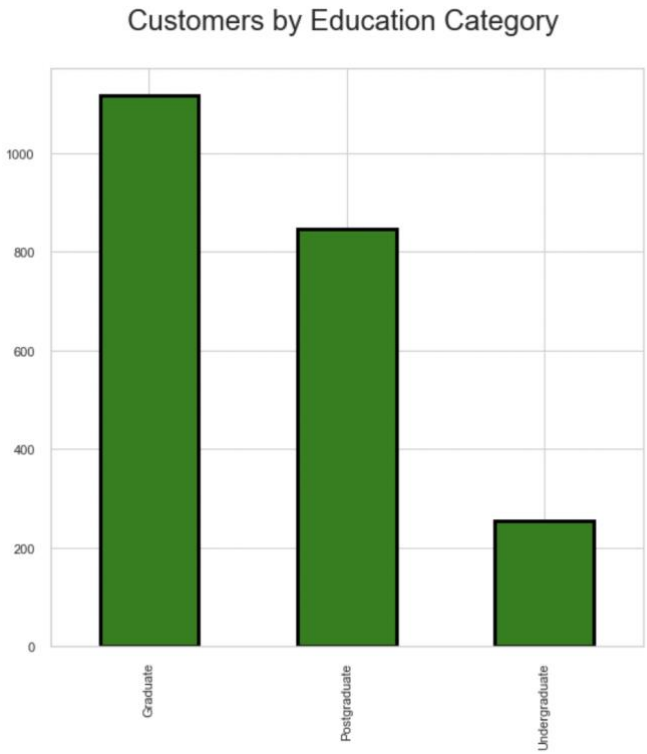
Feature engineering in this project aims to create new features from the existing attributes as well as explore the attributes from the dataset containing customer information to understand the customer better. The first feature was created to understand how long the customer has been shopping with the grocery firm with respect to the oldest existing customer entry in the database. Next a feature was created from the existing year of birth attribute to estimate the age of the customer. The age feature will help us understand the age group in which most of the shopping audience lies which in turn can help us make better predictions. Upon plotting the Age feature we could see from the age graph; most customers lie in the age range of 39 – 56 years old.

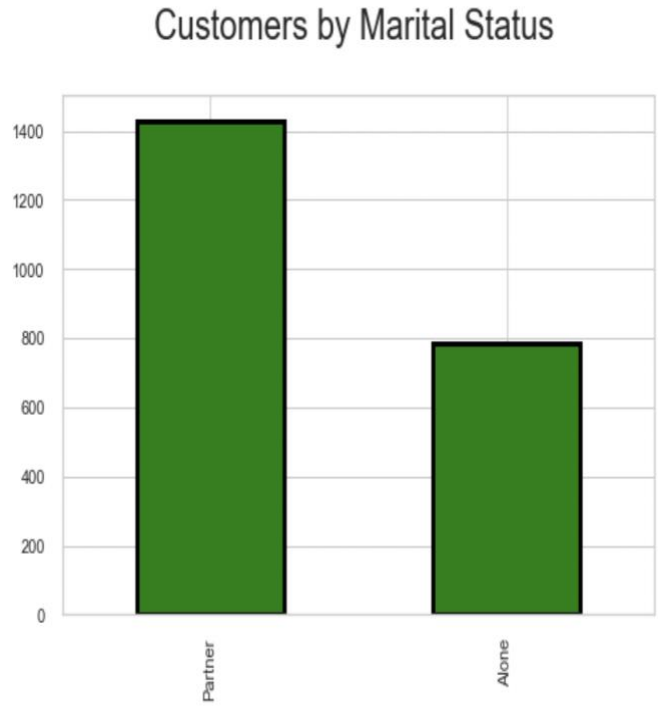
We observed the values counts of every education category and manipulated the variable Education to create three distinctive categories namely 'Postgraduate', 'Graduate' and 'Undergraduate' to simplify its value counts. We further went on to visualize the newly cleaned variable Education and found that most of the grocery firm's buyers are indeed graduates followed by the postgraduates' customers. We then focused on the variable Marital_Status and followed a similar approach that of the Education variable of first analyzing the counts and then categorizing the variable into 2 distinct categories 'Alone' and 'Partner'. This helped us to group and narrow down the similar categories into 2 types. Visualization graphs for Marital Status variable showed partner households had more customers buying for the firm than Single households.

Further, a new feature 'Children' was created by adding the two variables Kidhome and Teenhome to indicate the total number of children in a household including kids and teenagers.
Similarly, feature 'Family_Members' was created by combining the variables pertaining to marital status and number of children in a household. Another feature was generated which helped us to check whether a customer is a parent based on the presence or absence of children in the household.

Towards the end of the feature engineering process, features from raw data which had been used to extract meaningful features were dropped and the dataset with newly created features was visualized to check how the new variables look like. The visualization depicted that our data consisted of some outliers which can affect model building in the later stages of the project and that they need to be handled at this stage.

The cap to limit customers above 90 years of age and Income falling more than 600000 helped to tackle the outlier data points. A correlation matrix was created to study the relationship between the features in the dataset.





Snippet of New Engineered Features

Gold	...	AcceptedCmp2	Complain	Response	Expenditure	Total_acc	Customer_Days	Age	Children	Family_Members	Parent_Status
2216.000000	...	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2.216000e+03	2216.000000	2216.000000	2216.000000	2216.000000
43.965253	...	0.013538	0.009477	0.150271	607.075361	0.298285	4.423735e+16	54.179603	0.947202	2.592509	0.714350
51.815414	...	0.115588	0.096907	0.357417	602.900476	0.679209	2.008532e+16	11.985554	0.749062	0.905722	0.451825
0.000000	...	0.000000	0.000000	0.000000	5.000000	0.000000	0.000000e+00	27.000000	0.000000	1.000000	0.000000
9.000000	...	0.000000	0.000000	0.000000	69.000000	0.000000	2.937600e+16	46.000000	0.000000	2.000000	0.000000
24.500000	...	0.000000	0.000000	0.000000	396.500000	0.000000	4.432320e+16	53.000000	1.000000	3.000000	1.000000
56.000000	...	0.000000	0.000000	0.000000	1048.000000	0.000000	5.927040e+16	64.000000	1.000000	3.000000	1.000000

A snippet of data after Feature Engineering

4. Preprocessing Data

4.1 Converting Categorical variables - Label Encoding technique

In data analysis, it is often necessary to convert categorical variables into numeric variables to enable further analysis and modeling. This is because many statistical algorithms and machine learning models require numeric data as inputs and cannot directly handle categorical data. Categorical variables are those that have discrete values that represent different categories or groups, such as gender, occupation, or education level. These variables are typically represented as text labels or codes and cannot be used directly in most statistical analyses or machine learning models. To convert categorical variables into numeric variables, one common technique is to use a process called encoding.

As discussed in the overview of the data section, the dataset consisted of some categorical variables such as Education and Living status of the customers. These categorical variables needed to be converted to numerical values so that they can be utilized in model generation. Here for the purpose of this project, label encoding technique from Python's scikit-learn library has been used to perform the task of categorical i.e object data type to numeric conversion of variables.

Label Encoding is a popular technique of encoding object type variables. A unique integer is assigned to each label based on alphabetical ordering. It is possible to preserve the information contained in the categorical variables by using appropriate encoding techniques and enable more accurate analysis and modeling.

4.2 Scaling the variables - StandardScaler Method

After successful conversion of categorical variables, the next step was to scale the variables in the dataset. Scaling of features is an important stage in modeling algorithms since the data obtained encompasses features of various dimensions and scales which can have an impact on modeling. This can cause the model results to be skewed if not scaled. In data analysis, scaling is the process of transforming the numerical values of variables to a standardized scale, typically to bring them all to a common range. The purpose of scaling is to ensure that all variables have an equal impact on the analysis or model, regardless of their original values or units of measurement.

Scaling is important because many statistical algorithms and machine learning models are sensitive to the scale of the input variables. If the variables have widely varying scales, then some variables may dominate the analysis or model, while others may have little impact. This can result in biased or inaccurate results.

Standardization is a scaling approach that makes data scale-free by changing the statistical distribution of the data into mean (0) and standard deviation (1). The StandardScaler () method in the Python sklearn module has been used to standardize the data values into a standard format. A copy of the dataset was created, followed by dropping the attributes related to campaigns and then scaling the rest of the attributes in the dataset.

4.3 Preparing Data for Dimensionality Reduction

To prep the data for Dimensionality Reduction, we inspected the dataset to check and validate the scaled features. This helped to gain an understanding of whether we can proceed with the reduction process. Since we created a subset of data to be scaled and considered for dimensionality reduction it was necessary to perform a check before moving ahead.

5. Dimensionality Reduction - Principal Component Analysis (PCA)

The features or attributes in a dataset help to determine the final classification in the model building stages. Many of these features are connected and so can also be redundant. Hence, before running the selected features through a classifier, it is important to reduce their dimensionality. The technique of lowering the number of input features under consideration by generating a set of primary variables is known as dimensionality reduction. Principal component analysis (PCA) is one such technique used to reduce the dimensionality of datasets which boosts interpretability while maintaining the information prior to reduction.

Principal Component Analysis (PCA) is a popular data analysis statistical technique for reducing the dimensionality of huge data sets while maintaining as much information as feasible. It works by identifying the most relevant patterns and correlations in the data and reducing it to a smaller set of uncorrelated variables known as principal components.

The principal components are nothing but linear combinations of the original variables that are organized in descending order of importance, with the first component accounting for the greatest variance in the data, the second component accounting for the second greatest variance, and so on. Overall, PCA is an effective approach for studying huge data sets and can provide useful insights into complicated systems by reducing data dimensionality.

PCA was implemented to reduce the dimensionality to 3 features by fitting the scaled data from the scaling step. The reduced data was described to view the statistics such as mean, max,count and std deviation.

6. Unsupervised Learning - Clustering

6.1 Finding K - Elbow Method

The Elbow Method is a technique used in data analysis to determine the optimal number of clusters to use in a clustering algorithm. The Elbow Method involves plotting the number of clusters on the x-axis and the sum of squared distances of the data points to their closest cluster center on the y-axis. The sum of squared distances is a measure of how spread out the data points are within each cluster. As the number of clusters increases, the sum of squared distances will generally decrease, because smaller clusters can better fit tightly grouped data points.

However, at some point, adding more clusters will not significantly reduce the sum of squared distances, and the improvement in clustering performance will become minimal. This point is known as the "elbow" of the curve, and it represents the optimal number of clusters to use.
By identifying the elbow point, the appropriate number of clusters has been identified to use in the clustering algorithm. To find the optimal number of clusters in which the data from our dataset can be clustered, the Elbow method was initialized and implemented.

As discussed above, to identify the best number of clusters, we must choose the value of k at the "elbow," or the point at which the distortion/inertia begins to decrease linearly. As a result, we found that the optimal cluster for the given data is four. We visualized ‘K’ using the Elbow Visualizer method to confirm the number of clusters to be formed.

6.2 Agglomerative Clustering

Agglomerative Clustering also called as agglomerative nesting is a type of hierarchical clustering used to group data points in clusters based on similarities. It works in a bottom-up approach by starting off with smaller clusters and then merging them together to form larger clusters closest clusters together until a stopping criterion is met. A major advantage of agglomerative clustering is that it can be used with any distance metric, making it a versatile technique for a wide range of applications. It is also easy to interpret and can be visualized using a dendrogram.

However, a disadvantage of agglomerative clustering is that for large datasets it can be computationally expensive, as it involves computing the distance between each pair of clusters at each iteration.

For implementation in this project, the Agglomerative Clustering method was used to form 4 clusters. Dimensionally reduced data using PCA was used to fit and classify the data points into these clusters. The cluster data was then appended to the original data frame to be further used for visualization. Through 3D visualization, we could clearly see 4 distinct clusters of data points.

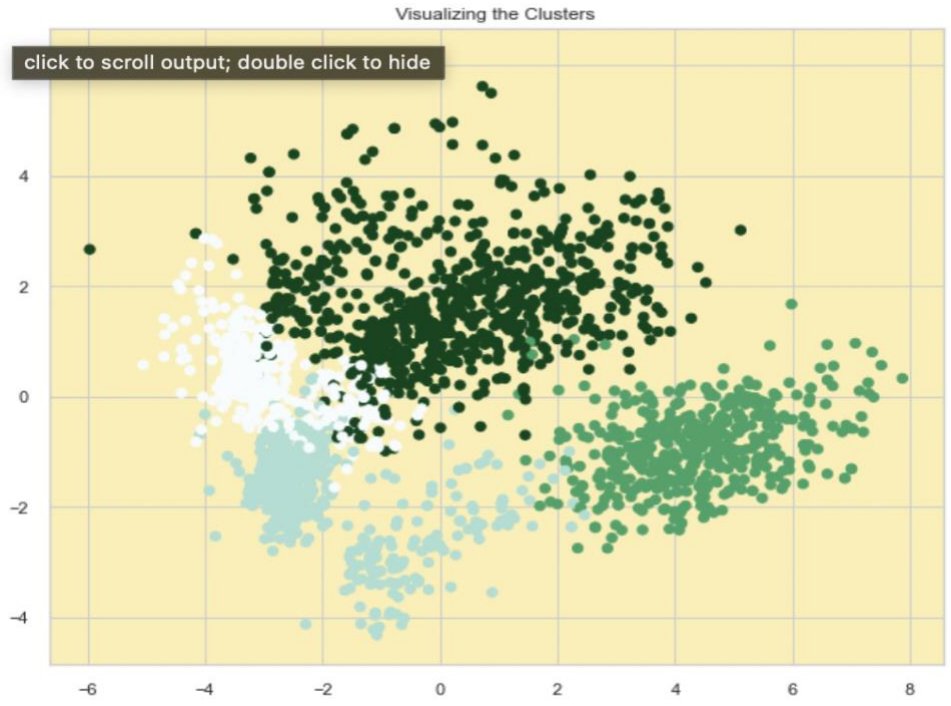
6.3 K-means Clustering

K-means clustering is a kind of unsupervised algorithm that groups similar data points into clusters. It is popularly used in data analysis, pattern recognition, and image segmentation. The k-means algorithm can be used with different distance metrics, although the Euclidean distance is most used. The choice of distance metric can affect the clustering results, and it is important to choose an appropriate metric based on the data and the problem at hand.

K-means clustering has several advantages, including, it is fast and scalable, making it suitable for large datasets. It is easy to implement and interpret and it can be used with any distance metrics. However, k-means clustering also has some limitations and potential drawbacks, including requiring the user to choose the value of k, which can be difficult to determine It assumes the clusters to be spherical and of equal size, which may not be true in all cases.

K-means clustering is a useful and widely accepted technique for clustering analysis, particularly for large datasets with many variables. Its simplicity and scalability make it a popular choice for exploratory data analysis and unsupervised machine learning. K-means is a simple clustering algorithm, which groups data points based on similarities to find hidden patterns. The *K* in k-means refers to the number of centroids (imaginary or real location which represents the center of a cluster) needed to form the clusters.

Here, based on the output of Elbow method *K* was found to be 4 and the same was fed to the k-means algorithm. The result set produced by fitting and then training the dimensionally reduced data was found similar to agglomerative clustering and no changes in the cluster were noted amongst the 2 methods.



7. Model Evaluation

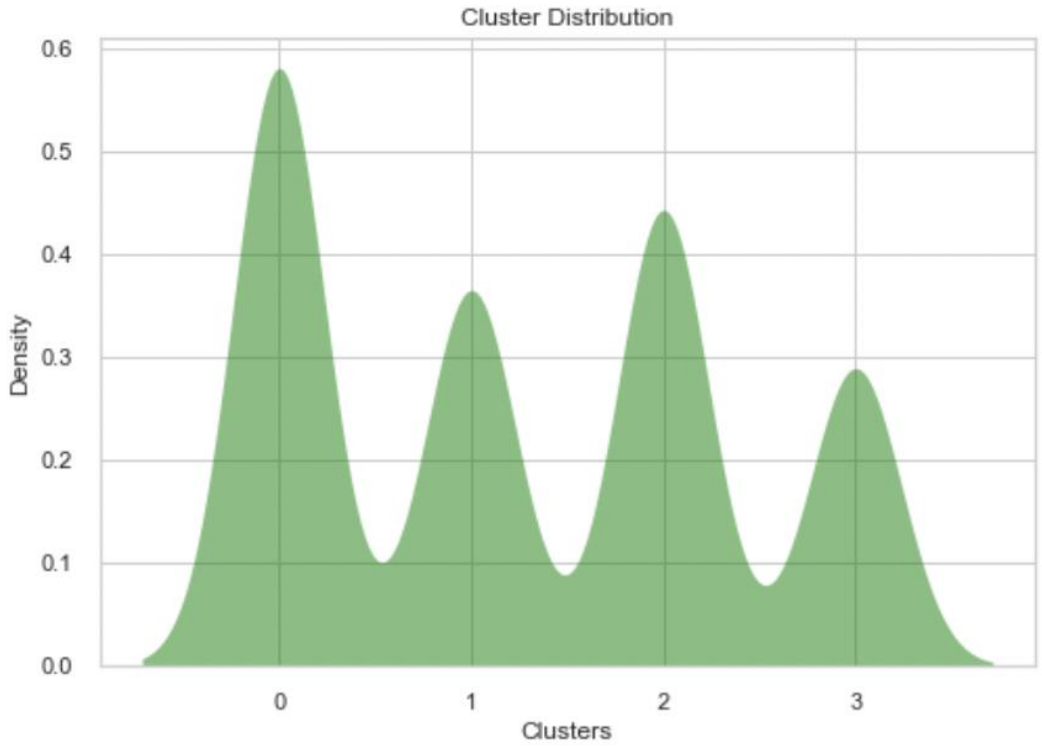
Model evaluation for clustering algorithms is the process of assessing the performance of a clustering model on a given dataset. Unlike supervised learning, clustering is an unsupervised learning technique, which means there are no pre-labeled target values to measure the accuracy of the clustering model. However, there are several evaluation metrics that can be used to assess the performance of a clustering algorithm. One such common method of evaluation metrics include:

Visual inspection: Visualization techniques, such as scatter plots or dendrograms, can be used to visually inspect the clustering results and identify any obvious patterns or anomalies.

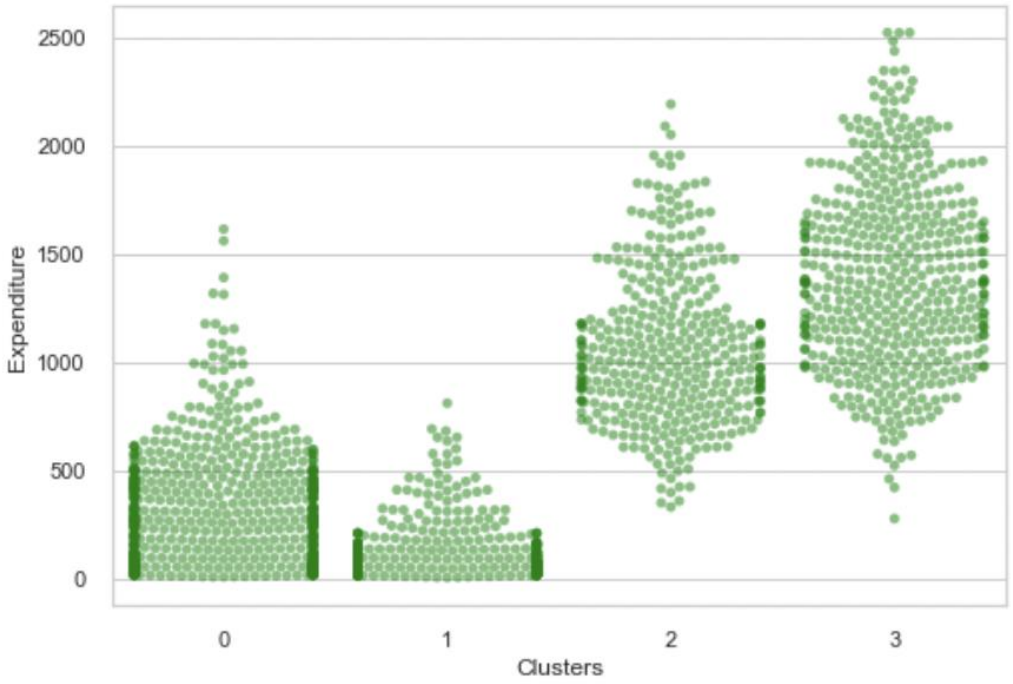
Since clustering is an unsupervised learning method, we do not have a target variable to predict data points for. The model has been evaluated using visualization to see the behavior of customers based on the clusters formed.

Key observations from post evaluation are:

1. Highest number of customers lie in clusters 0 and 2
2. Customers in clusters 0 and 1 had higher incomes and customers in cluster 1 had the highest expenditure followed by cluster 0 customers. Clusters 1 and 2 had lower incomes and lower expenditures.
3. Cluster 1 had high expenditure customers followed by cluster 0.
4. Customers in cluster 0 had accepted/responded to most campaigns and no group of customers responded to all the campaigns lauded so far. This paves way for a more targeted approach to marketing to improve sales.
5. Cluster 0 had the most number of deals purchased followed by cluster 3. No insignificant effect was seen on cluster 1 and 2.



Snippet of Distribution of Clusters



Snippet of Swarmplot to show the Distribution of Clusters by Expenditure

8. Profiling the Customers

Consumer profiling is a marketing approach that employs data to construct an image of the ideal consumer in the minds of the analysts who will engage with your product or service. A good customer profile, when done correctly, can serve as a guide for your marketing and advertising to target your ideal clients.

Customer profiling is based on the behaviors and experiences of the consumers. It investigates pain sites and touchpoints. In other words, customer profiling is concerned with the personalities of your consumers. Its goal is to help you understand your consumers so you can provide a better experience, product, or service to the individuals who use your product.

Please note that the assessment discussed in this section is based on the facts provided and implementation performed on the features in the dataset.

Cluster Number: 0

This group being discussed is a group of parents who are older in age and have a family of at at least two people and at most four members. The majority of parents in this category have a teen at home, and parents who are single make up a sub-part of this group. Cluster 0 has the highest number of customers.

Cluster Number: 1

It can be concluded that the customers grouped in this cluster are non-parents with a maximum of two member families. The majority of this group consists of couples, rather than single people. The members of this group span a wide range of ages and are part of a high-income group.

Cluster Number: 2

It has been observed that this cluster of clusters are mostly parents who are relatively younger and have families of more than three people. The majority of these parents have only one child, who is usually not a teenager.

Cluster Number: 3

As per the results of implementation, the customers grouped in this cluster are a group of parents that are older and have a family size of at least two and at most five members. The majority of these families are teen households and belong to lower income.

4. Conclusion

The goal of an unsupervised learning project on customer personality analysis would be to create and implement a machine learning model capable of uncovering hidden trends and insights in customer behavior data. The project's purpose is to use this model to further comprehend consumer personalities and preferences to improve the entire consumer experience and foster corporate success. This could entail training the model on a big collection of customer behavior data and identifying patterns and trends in the data using approaches like clustering or dimensionality reduction. The model developed could be utilized to assess new data and predict client personalities.

Unsupervised learning can be an excellent way for analyzing client behaviors. It is possible to uncover patterns in customer data and forecast consumer personality traits based on their behavior and features by utilizing clustering algorithms and dimensionality reduction techniques. This can give you vital insights into your consumer base and allow you to bolster and enhance the company's marketing strategies and customer care efforts to better meet the customers' requirements and preferences. The use of online assistants and chatbots could be a potential use case for unsupervised learning in consumer personality profiling. These systems could enable more individualistic and successful interactions with customers by incorporating the results of customer personality research into their algorithms, potentially leading to increased customer satisfaction and loyalty.

There are some setbacks to the process of customer personality analysis. The most critical factor is the quality and reliability of data. Accurate and dependable consumer personality forecasts necessitate high-quality data devoid of errors. The process of predicting the customer personality analysis is heavily dependent on the quality and reliability of data available for study. In some circumstances, there might be too little data or incomplete data, limiting the algorithm's capacity to produce accurate predictions. However, the future of unsupervised learning-based customer personality analysis looks promising and will be increasingly adopted by companies around the globe to uncover the patterns and relationships among different buyer behaviors as well utilize the derived insights into a plethora of business applications.

5. Future Scope

The dataset features could update, change as per company policies and company resizing as well as change in customer behavior in the near future. This would need th organization to find a more elaborate storage of attributes related to Product purchasing since wine, meat would not be the only products the customers continue to buy. Also, there would also be a potential change in the way the data for households has been impacting the customer behavior.

All the characteristics would need to be investigated and the many associations with not only the existing features but also all the other features included in the dataset, using various algorithms, clustering techniques, and more plots.

In that case, the newly added features would need to be cleaned, analyzed, and visualized. The emergence of new features in the future could also give rise to new engineered features. Also, other methods of unsupervised clustering algorithms along with the different ways to estimate the number of clusters could be explored.

6. References

Julien Ah-Pine (2018). *An Efficient and Effective Generic Agglomerative Hierarchical Clustering Approach*, 19(42):1–43, 2018. URL: An Efficient and Effective Generic Agglomerative Hierarchical Clustering Approach

Python 3: https://www.python.org/

PCA: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Elbow: https://www.scikit-yb.org/en/latest/

Label Encoding: sklearn.preprocessing.LabelEncoder — scikit-learn 1.2.2 documentation

Standard Scaler: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Pandas: https://pandas.pydata.org/

Numpy: https://numpy.org/

Seaborn: https://seaborn.pydata.org/

Matplotlib: https://matplotlib.org/

K-Means Clustering: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Hierarchical Clustering: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

Data Set: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis?datasetId=1546318&sortBy=voteCount