# Customer Personality Analysis: An End To End Clustering Project

Abdulla Nabeel
Department of Computer Science
Khalifa University
Abu Dhabi, UAE
Email: Abdullahuni@outlook.com

***Abstract*— This paper presents a comprehensive analysis of customer spending patterns and their impact on marketing strategies within a company. By examining customer demographics, purchasing behavior, and response to marketing campaigns, the study identifies key insights and opportunities for optimization. The findings highlight the importance of targeting older, higher-income customers and focusing on profitable categories such as wine and meat products. The analysis also emphasizes the need for improved online marketing strategies to convert website visits into sales and enhance the effectiveness of deals and promotions. By segmenting customers and tailoring marketing efforts to specific clusters, particularly the low-class segment, the company can tap into market expansion opportunities. Continuous data monitoring is emphasized to adapt strategies in real-time and maintain a competitive edge. Overall, the study provides valuable recommendations for driving customer engagement, increasing sales, and expanding market presence.**

***Keywords—Customer Segmentation, Purchasing Behavior, Marketing Strategies, Clusters.***

## I. INTRODUCTION

A thorough examination of a company's ideal customers is known as Customer Personality Analysis. Its purpose is to provide a deeper understanding of customers, enabling businesses to tailor their products to meet the distinct requirements, behaviors, and concerns of various customer groups. By conducting a customer personality analysis, businesses can adapt their products to appeal to specific customer segments. For instance, rather than advertising a new product to every customer on their list, a company can identify which segment of customers is most likely to purchase the product and concentrate its marketing efforts on that particular segment.

Clustering is a widely-used machine learning technique that groups similar data points together based on their characteristics. It has been applied to various fields, including marketing, to gain insights into customer behavior. In recent years, clustering has become an increasingly popular method for analyzing customer personality traits and preferences. By clustering customers based on their shared characteristics, businesses can gain a better understanding of their target audience, personalize their marketing strategies, and improve customer satisfaction. In this paper, we will explore the use of clustering in personality analysis to gain insights into customer segments and provide a practical guide for businesses to implement this technique in their marketing strategies.

The summary of the customer segmentation analysis resulted in the identification of three distinct customer clusters based on their demographics and spending patterns. The results of the Exploratory Data Analysis suggest that older, highly educated customers with high income are the biggest spenders, and that the Meat and Drinks products generate the highest profit.

## II. RELATED WORK

### *A. Analysis done by Raghavendra and Singenahalli (2021)*

The paper discusses a customer personality analysis conducted using unsupervised learning techniques. The dataset used contains 2240 entries with 29 features. Data preprocessing involves handling null values and removing features with no variance. Feature selection is performed using a correlation matrix and heatmap. An ensemble model combining multiple algorithms is used for prediction, resulting in a high accuracy of approximately 99%. The study concludes that understanding customer preferences based on personality traits can aid companies in targeting customers effectively. Future scope includes testing the algorithm on different datasets and exploring regression models and neural networks. The findings offer insights for implementing customer personality analysis in business strategies [1].

### *B. Analysis done by Soumica, M., Varma, C. S., & Krishna (2021)*

The paper discusses a customer personality analysis conducted using unsupervised learning techniques. The dataset used contains 2240 entries with 29 features. Data preprocessing involves handling null values and removing features with no variance. Feature selection is performed using a correlation matrix and heatmap. An ensemble model combining multiple algorithms is used for prediction, resulting in a high accuracy of approximately 99%. The study concludes that understanding customer preferences based on personality traits can aid companies in targeting customers effectively. Future scope includes testing the algorithm on different datasets and exploring regression models and neural networks. The findings offer insights for implementing customer personality analysis in business strategies [2].

### *C. Analysis done by Andhika Widyadwatmaja (2021)*

The Author focuses on applying clustering algorithms, specifically K-means and hierarchical clustering, to group similar customers based on their attributes or behavior. The author emphasizes the importance of feature selection and dimensionality reduction techniques in preparing the data for clustering. They discuss methods like PCA and t-SNE for visualizing the clustered data effectively. Evaluation of clustering results is covered, with metrics such as silhouette score

and WCSS being highlighted. Techniques like the elbow method are suggested to determine the optimal number of clusters. The clustering analysis primarily considered factors such as income, expenses, the number of purchases by category, and the total accepted campaign. Interestingly, education level, marital status, and age did not significantly impact the clustering process. Based on the model, the analysis suggests that there are two distinct customer segments [3].

### *D. Analysis done by Snehi Nainesh Pachchigar (2021)*

The study explores customer personality analysis using data analysis techniques. It involves cleaning, observing, and enhancing the data to prepare it for clustering algorithms. K-means and Agglomerative clustering, along with the Elbow method, are utilized to segment customers into clusters. The analysis includes examining plots and graphs to understand cluster distribution, spending patterns in relation to income, the impact of deals and campaigns on different clusters, and spending trends based on various attributes such as children at home, teenagers at home, family size, and income. The findings indicate that customers in groups 0 and 1, characterized by higher or average income and children, are more confident in retail. Suggestions are made to offer more products and deals specifically targeting customers in groups 2 and 3. The study proposes future work with larger datasets, the potential expansion to a recommendation system, and the exploration of newer models for data analysis [4].

### *E. Analysis done by Anindhita Dewabharata (2022)*

The study aims to address the risk of overstocking in online shops by focusing on customer segmentation to understand their purchase interests. Using the k-means algorithm, the study analyzes data from an online shop in Indonesia to divide prospective customers into different segments. By identifying the top three customer segments, the study recommends an inventory strategy for online shops to target these segments and prioritize their preferences. This approach helps reduce the risk of overstocking and allows online shops to optimize their inventory management based on customer preferences [5].

### *F. Analysis done by Alicja Rachwał et al. (2023)*

The research aimed to find effective ways to analyze and verify the quality of datasets for segmenting customer observations. The paper suggests a new approach for handling datasets that contain both categorical and continuous variables in customer segmentation tasks. To accomplish this, a unique unsupervised model based on an autoencoder was used to embed categorical variables. Then, different clustering algorithms, including k-means, DBSCAN, Louvain algorithm, greedy algorithm, and label propagation algorithm, were applied to divide the customers into groups based on similarity. Two datasets were used for the research, one consisting of retail customers and the other of wholesale customers. Various metrics such as the Calinski-Harabasz index, Davies-Bouldins index, NMI index, Fowlkes-Mallows index, and silhouette score were employed to evaluate the quality of the clustering results. The study found that the modularity parameter in graph methods was a useful indicator to determine whether a particular dataset could be meaningfully divided into distinct groups [6].

## III. METHODS

The life cycle of this project involves several stages that are essential for its success. Below are the main followed stages:

1. Problem Statement: In this stage, the project's objectives and goals are defined, and the problem to be solved is identified. This stage helps to set clear expectations for the project's outcome and ensures that the project is focused and on track.
2. Data Collection: The next stage involves gathering data from various sources. Data collection involves determining the relevant data sources, such as databases, APIs, or web scraping, and then obtaining the data in a structured format.
3. Data Cleaning: Data cleaning involves removing any irrelevant or incorrect data and correcting any inaccuracies or inconsistencies in the data. This stage is crucial in ensuring that the data is accurate and reliable for analysis.
4. Exploratory Data Analysis: In this stage, the data is analyzed to gain insights and identify patterns or trends. Data visualization techniques are used to help understand the data and identify any correlations.
5. Data Pre-Processing: This stage involves transforming the data into a format suitable for modeling. This may involve feature engineering, where new features are created, or feature selection, where irrelevant features are removed.
6. Model Training: This stage involves selecting a suitable machine learning algorithm and training the model on the pre-processed data. The model is evaluated using various metrics to determine its accuracy.
7. Model Selection: Based on the model's evaluation, the best model is selected, and profiling is carried out to assess the analysis.
8. Recommendations: In the final stage, recommendations are made based on the insights gained from the data analysis and the modeling. The recommendations may include specific actions to address the problem statement or suggestions for further research.

## IV. EXPERIMENTAL RESULTS

### *A. Description of Dataset*

The dataset contains information about customers and their purchasing behavior. It includes attributes such as customer demographics (ID, Year_Birth, Education, Marital_Status, Income), household composition (Kidhome, Teenhome), enrollment date (Dt_Customer), recency of last purchase (Recency), and customer complaints (Complain).

The dataset also provides details about the amount spent on different product categories over the last 2 years, including Drinks (MntDrinks), fruits (MntFruits), meat

(MntMeatProducts), fish (MntFishProducts), sweets (MntSweetProducts), and gold (MntGoldProds).

Information about promotional activities is available, such as the number of purchases made with a discount (NumDealsPurchases) and whether customers accepted offers in specific campaigns (AcceptedCmp1-5). The response of customers to the last campaign is indicated by the Response attribute.

Additionally, the dataset includes information about the place of purchase, including the number of purchases made through the company's website (NumWebPurchases), using a catalog (NumCatalogPurchases), directly in stores (NumStorePurchases), and the number of visits to the website in the last month (NumWebVisitsMonth).

These attributes provide valuable insights into customer behavior, preferences, and responses to promotions, which can be utilized for customer segmentation, targeted marketing strategies, and business decision-making. Below is a table representing these attributes.

| *Attribute* | *Description* |
|---|---|
| ID | Customer's unique identifier |
| Year_Birth | Customer's birth year |
| Education | Customer's education level |
| Marital_Status | Customer's marital status |
| Income | Customer's yearly household income |
| Kidhome | Number of children in customer's household |
| Teenhome | Number of teenagers in customer's household |
| Dt_Customer | Date of customer's enrollment with the company |
| Recency | Number of days since customer's last purchase |
| Complain | 1 if the customer complained in the last 2 years, 0 otherwise |
| MntDrinks | Amount spent on drinks in last 2 years |
| MntFruits | Amount spent on fruits in last 2 years |
| MntMeatProducts | Amount spent on meat in last 2 years |
| MntFishProducts | Amount spent on fish in last 2 years |
| MntSweetProducts | Amount spent on sweets in last 2 years |
| MntGoldProds | Amount spent on gold in last 2 years |
| NumDealsPurchases | Number of purchases made with a discount |
| AcceptedCmp1 | 1 if customer accepted the offer in the 1st campaign, 0 otherwise |
| AcceptedCmp2 | 1 if customer accepted the offer in the 2nd campaign, 0 otherwise |
| AcceptedCmp3 | 1 if customer accepted the offer in the 3rd campaign, 0 otherwise |
| AcceptedCmp4 | 1 if customer accepted the offer in the 4th campaign, 0 otherwise |
| AcceptedCmp5 | 1 if customer accepted the offer in the 5th campaign, 0 otherwise |
| Response | 1 if customer accepted the offer in the last campaign, 0 otherwise |
| NumWebPurchases | Number of purchases made through the company's website |
| NumCatalogPurchases | Number of purchases made using a catalogue |
| NumStorePurchases | Number of purchases made directly in stores |
| NumWebVisitsMonth | Number of visits to company's website in the last month |

*Table A: Summary of Attributes*

## *B. Data Cleaning*

### *B.I: Missing Values*

The dataset has missing values in the Income attribute. The type of missing data is classified as Missing At Random. This means that the missing values are not related to the values of other variables but may depend on some observed data.

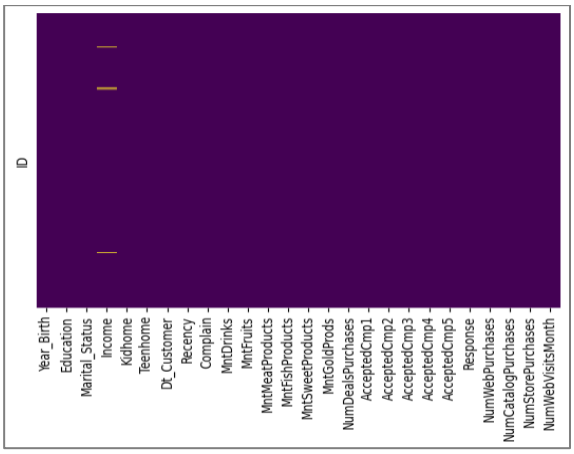



*Figure A: Heatmap of missing values*

To handle the missing values in the Income attribute, we will utilize an iterative imputer. This method uses the numerical features in the dataset to impute the null values in the Income feature. Since the missing values are of the type Missing At Random, a supervised learning technique is chosen for imputation.

| Type of missing data | Imputation method |
|---|---|
| Missing Completely At Random | Mean, Median, Mode, or any other imputation method |
| Missing At Random | Multiple imputation, Regression imputation |
| Missing Not At Random | Pattern Substitution, Maximum Likelihood estimation |

*Table B: Summary of Imputation methods*

Since the Income feature in the dataset contains outliers, it is crucial to choose an algorithm that is robust and less influenced by outliers. Tree-based algorithms are known to be less sensitive to outliers compared to other algorithms. Therefore, in this case, we will utilize the Gradient Boosting Regressor (GBR) as our model.

The Gradient Boosting Regressor is a machine learning algorithm that belongs to the ensemble learning family. It combines multiple decision trees as base estimators and gradually improves the model's predictive power through boosting. Boosting is a technique where weak learners (individual decision trees) are sequentially trained, and their predictions are combined to create a strong predictive model.

The advantage of using GBR in this scenario is that decision trees, which are the fixed base estimators in GBR, are inherently robust to outliers. Decision trees make splits based on thresholds and do not assume any particular distribution of the data. This makes them less affected by extreme values or outliers present in the Income feature.

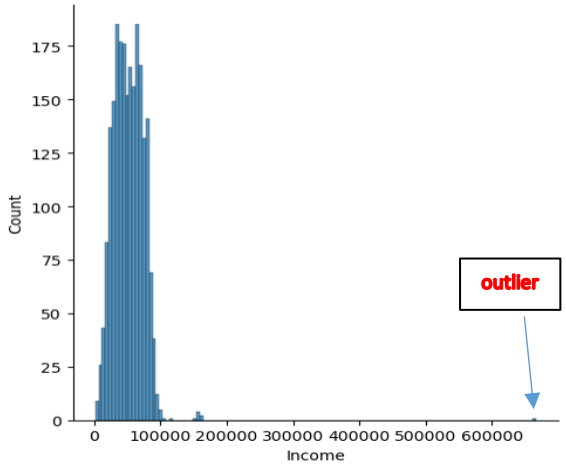



*Figure B: Income Attribute Distribution*

We compare the iterative imputation using GBR model with mean and median imputation:

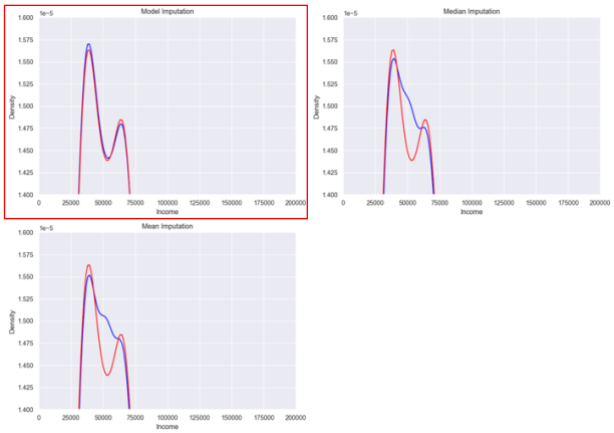



*Figure C: Comparison of 3 imputation techniques*

Based on the observation that the Iterative imputation method yields the closest distribution to the original data, we have decided to utilize Iterative imputation to fill the null values in the Income feature.

*B.II: Duplicated Samples*

Upon examining the dataset, it has been observed that there are a total of 182 duplicated samples present. To handle this issue, the duplicates will be addressed by dropping them from the dataset. By removing the duplicated samples, we ensure that each observation in the dataset is unique, avoiding any potential biases or distortions in the analysis. Dropping the duplicates allows us to work with a clean and non-redundant dataset, ensuring the accuracy and integrity of the subsequent analysis, modeling, and interpretation of the results.

*B.III: Handling Categorical Features*

- Education attribute: There are 5 categories, but some of them are considered unrealistic. The categories will be encoded into two main classes: "undergraduate" and "graduate" to simplify the classification.
- Marital_Status attribute: There are 8 categories, but some of them are not realistic. The categories will be encoded into two main classes: "Together" and "Alone" to provide a more meaningful and interpretable classification.

*B.IV: Feature Engineering*

1. The 'Dt_Customer' attribute will be used to create a new feature called 'DaysJoin', representing the number of days since each customer joined the company.
2. New column named 'TotalSpent' by summing up the values from various columns: 'MntDrinks', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'NumDealsPurchases', and 'MntGoldProds'. The resulting sum represents the total amount spent by each customer on different products.
3. New column named 'Children' by summing up the values from two columns: 'Kidhome' and 'Teenhome'. This column represents the total number of children and teenagers in each customer's household.
4. New column named 'Age' by subtracting the values in the 'Year_Birth' column.
5. 'TotalMembers': It calculates the total number of members in each customer's household by summing the 'Children' column and the encoded values of the 'Marital_Status' column. This column provides information about the household size.
6. 'Parent': It determines whether a customer is a parent or not based on the value in the 'TotalMembers' column. If the total members are greater than 2, the 'Parent' column is assigned a value of 1, indicating the presence of a parent in the household; otherwise, it is assigned a value of 0.
7. 'TotalPurchase': It computes the total number of purchases made by each customer across different channels, including web, catalog, store, and deals. The values from corresponding columns are summed to obtain the total purchase count.
8. 'TotalAcceptedCamps': It calculates the total number of marketing campaigns accepted by each customer by summing the values from individual campaign acceptance columns (AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5).

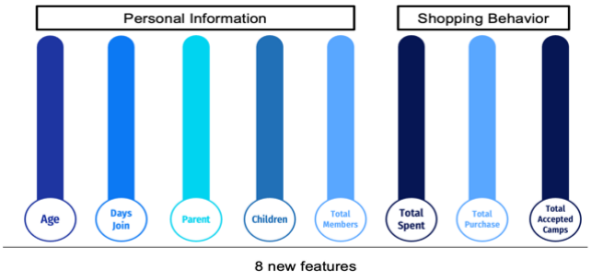



*Figure D: Summary of New Features*

*B.V: Handling Outliers*

1. Income: There is an outlier in the dataset with an income value of 666,666. This value is significantly higher than the other income values and may be considered an anomaly.

2. Age: There are three individuals in the dataset with ages exceeding 110 years. Their ages are recorded as 115, 116, and 122. These values seem unrealistic and may be erroneous.

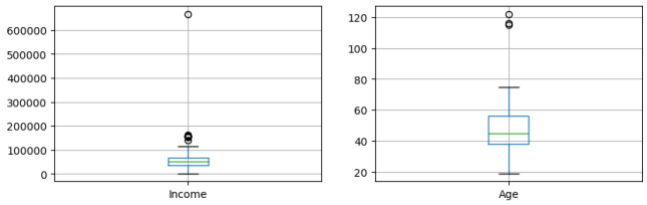



*Figure E: Boxplots of Income and Age*

To address the outliers, a filter will be applied to exclude data points that fall outside a reasonable range or distribution. This filtering process will help remove extreme values that could negatively impact the analysis or modeling tasks.

### *C. Exploratory Data Analysis*

The goal of Exploratory Data Analysis (EDA) is to gain insights and understanding from the data by examining its characteristics, patterns, and relationships. In this context, we are aiming to find some characteristics of customers who spend the most.

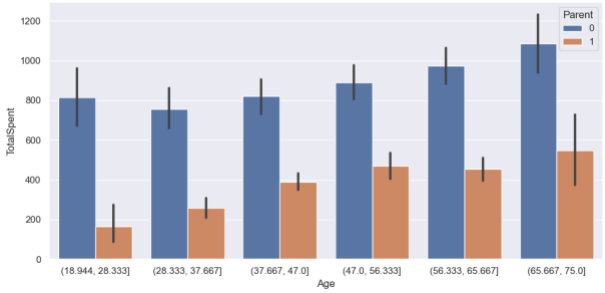
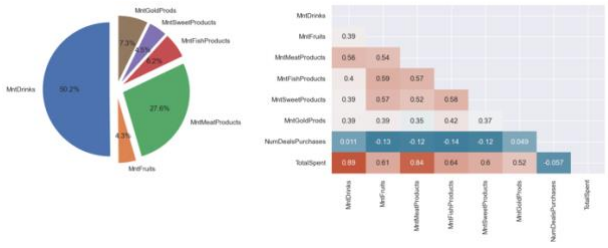



*Figure F: Age vs. TotalSpent*



*G: Percent Spending on products Figure*

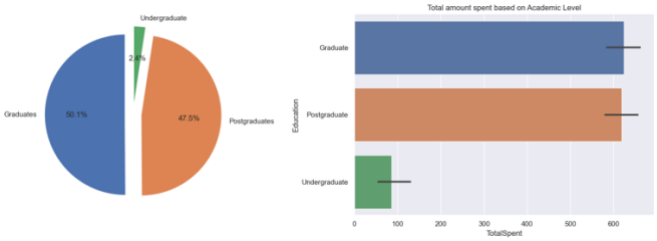



*Figure H: Percent Spending relative to academic level*

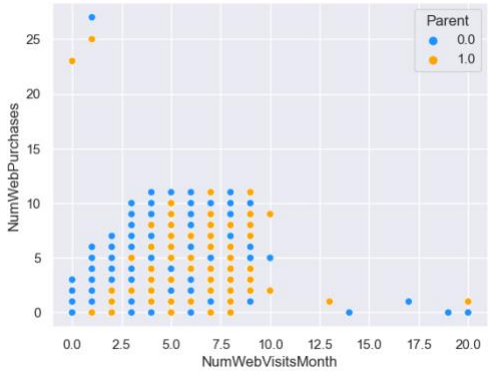



*Figure I: Web purchases vs. Web Visits*

1. Customer Characteristics: Older customers, those with higher education levels (graduates), and higher incomes tend to spend the most. These customer segments may be the target audience for marketing campaigns aimed at high-spending customers.
2. Profitable Product Categories: Drinks and Meat products are the top contributors to profit generation. This information can guide the company in allocating resources and optimizing marketing efforts towards these product categories.
3. Online Purchases: There is no significant correlation observed between the number of website visits in the last month and online purchases. To increase online purchases, the company should consider offering daily online deals and promotions to attract customers and encourage online transactions.

### *D. Data Preprocessing*

To prepare the data for further analysis and modeling, the following preprocessing steps were performed:

1. Drop Features: Features that were used to generate new, meaningful features will be dropped from the dataset. Additionally, features that do not provide information spanning over 2 years, such as NumWebVisitsMonth, will also be dropped.
2. Label Encoding: Categorical variables will be encoded using label encoding. This process assigns a unique numeric label to each category, allowing the categorical data to be represented in a numerical format that can be utilized by machine learning algorithms.
3. Standardization: The numerical features will be standardized using a standardization technique, such as Z-score standardization. This process scales the data to have a mean of 0 and a standard deviation of 1, ensuring that all features are on a similar scale. Standardization is useful for models that are sensitive to the scale of the input features.
4. Low Variance Feature Removal: Features with very low variance that do not contribute much to the overall variation of the data may be dropped. This step helps reduce the dimensionality of the dataset and removes features that are less informative or redundant.

By performing these data preprocessing steps, the dataset is prepared for subsequent analysis and modeling tasks, ensuring that the data is in a suitable format and optimized for accurate and effective modeling results.

### *E. Model Training*

To determine the optimal number of clusters in a dataset we use the elbow method.

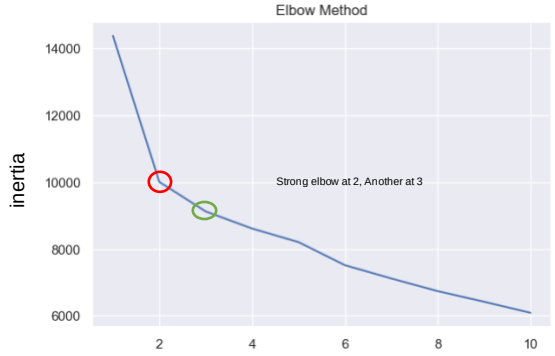



*Figure J: Elbow Method*

To utilize the full potential of unsupervised machine learning, we are going to choose three clusters for our analysis. Choosing two clusters gives a good separation of clusters, as shown in figure above, however it is easy to do the separation using a rule-based approach. On the other hand, choosing three clusters helps find that separation of clusters which cannot be done manually. Even though the silhouette score clearly suggests the prior option -as shown below- we will try to find a model suitable for the latter option.

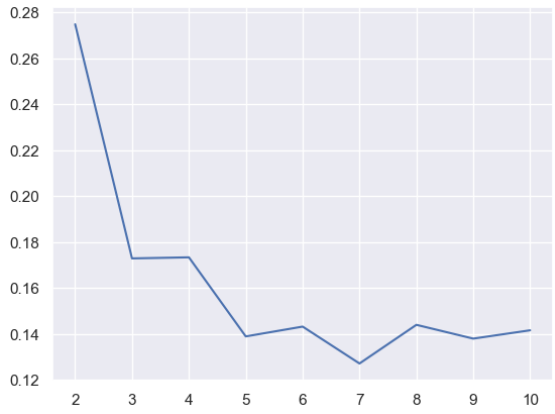



*Figure K: Silhouette Score*

For this analysis, we are going to use three unsupervised clustering algorithms:

1. K-Means assumes spherical clusters with equal variance and assigns data points to the nearest centroid based on mean calculations. It is simple and efficient, suitable for large datasets and known cluster numbers, but sensitive to initial centroid placement and not suitable for non-linear clusters.
2. K-Medoids allows for non-spherical clusters and uses actual data points as medoids. It minimizes dissimilarities between data points and medoids. It is robust to outliers, suitable for non-linear clusters, but computationally more expensive and requires specifying the number of clusters.
3. Gaussian Mixture Models (GMM) assumes data points are generated from a mixture of Gaussian distributions. It estimates parameters (mean, covariance) of Gaussian components and assigns data points based on their likelihood. It captures complex cluster shapes and allows for soft assignment but is computationally more expensive and sensitive to the number of components.

### *F. Model Selection*

Below are the results of each model on the plot Income vs. TotalSpent:


*Figure L: Results of K-means*

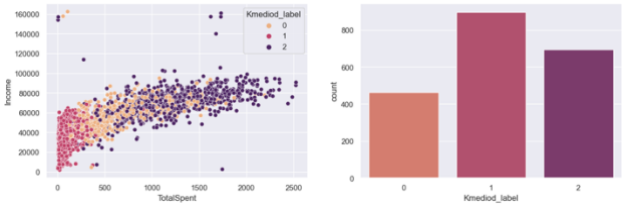



*Figure M: Results of K-medoids*

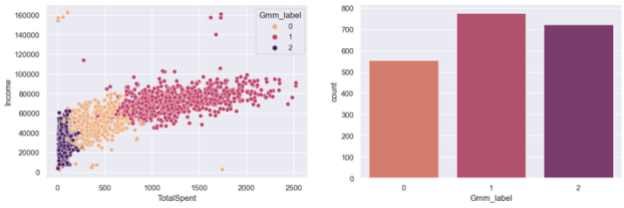



*Figure N: Results of GMM*

After evaluating the K-Means, K-Medoids, and Gaussian Mixture Models (GMM) for clustering the data, we have decided to choose the GMM model. The GMM model produced clusters that were relatively equal in size, indicating a more balanced distribution of data points among the clusters compared to the other models. This balance can be beneficial in terms of representation and interpretation of the clusters. Furthermore, the GMM model, being a probabilistic model, considers the covariance structure of the data, allowing for more flexible cluster shapes and accommodating potential overlaps between clusters. This can be advantageous when dealing with complex data distributions. Next, we will plot for profiling using joint and scatter plots.

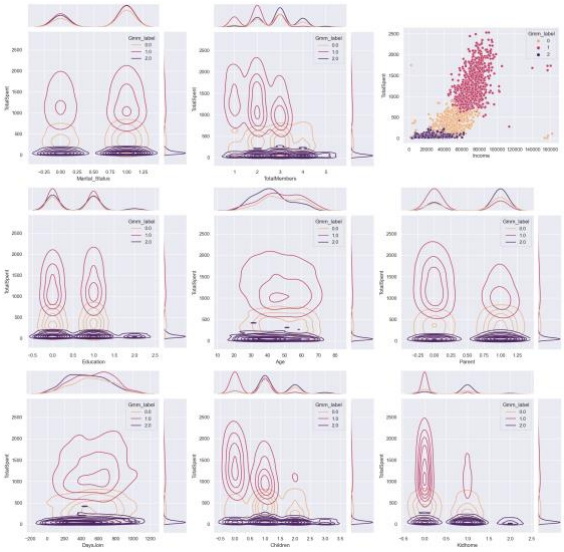


*Figure O: Profile plots*

From the above plots of joint plots and a single scatter plot, we can derive the profile for each cluster:

Cluster 0: The Middle Class

- Middle class payers
- Have at most 2 children at home.
- Family size does not exceed 4 members
- Age between 25 and 70.
- Amongst the cluster, customers who are parent spend a little more.
- Customers are graduate and postgraduate only.
- Moderate spendings, moderate to high incomes.

Cluster 1: High Class

- High class payers.
- Have at most 1 child at home. Mostly a teen.
- Family size does not exceed 3 members
- Age between 25 and 75.
- Amongst the cluster, customers who are not parent spend the most.
- Customers are graduate and postgraduate only.
- High spendings, moderate to high incomes.

Cluster 2: Low Class

- Low class payers.
- Have at most 3 children at home.
- Family size reaches 5 members.
- Age between 19 and 75.
- Amongst the cluster, customers who are parent spend a little more.
- The only cluster that includes undergraduate customers.
- low income, low spendings.

*G. Recommendations*

Based on the analysis conducted using the GMM clustering model and given plot, we can derive several recommendations to inform decision-making and marketing strategies:

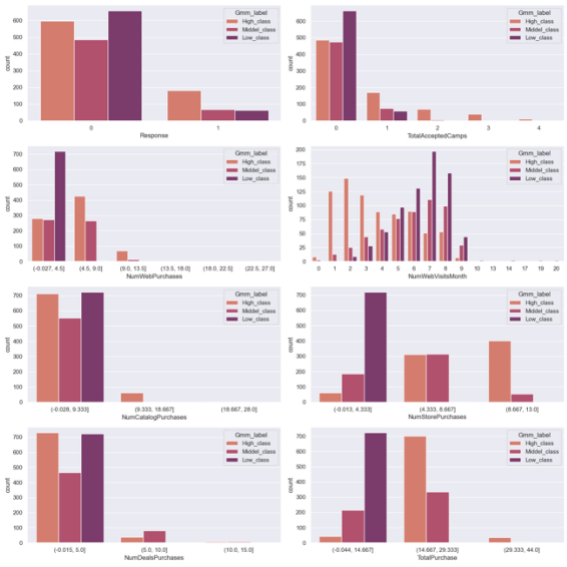


*Figure P: Customer behavior per cluster*

1. Marketing Strategy Review: The company should conduct a thorough review of its marketing strategies and campaigns to identify potential areas of improvement. Despite the clustering analysis, the company's marketing campaigns have not been successful across all clusters, including the response to the last campaign. It is crucial to reassess the messaging, targeting, and channels used to reach different customer segments.
2. Online Marketing Optimization: The analysis revealed that low-class customers visit the website the most but make the least website purchases. This indicates a potential gap in the company's online marketing strategy. The company should investigate the reasons behind this disparity and focus on optimizing its online promotions and user experience to convert website visits into actual purchases. Implementing targeted online promotions, personalized recommendations, and improving the overall website functionality can help attract and engage low-class customers.
3. Enhance Deals and Promotions: The analysis indicates that deals are not as effective in attracting customers, particularly among the low-class cluster. Given that the low-class cluster represents a significant portion of the customer base, the company should pay special attention to this segment's preferences and needs. Increasing the frequency and attractiveness of deals specifically tailored to the low-class cluster can help improve customer engagement and drive sales. It is essential to understand the underlying reasons why the current deals are not resonating with this segment and adjust the offers accordingly.

By implementing these recommendations, the company can improve customer targeting, increase customer satisfaction and loyalty, optimize marketing efforts, and ultimately drive business growth. It is important to regularly evaluate and adapt these strategies based on ongoing data analysis and customer feedback to stay competitive in the market.

## V. Conclusion

This analysis aimed to gain insights into customer spending patterns and improve the marketing strategies of a company. The analysis revealed that older customers with higher education and income tend to spend the most. Drinks and meat products were identified as profitable categories, while deals purchases were lower. The company's marketing campaigns were not successful across all customer segments, highlighting the need for a comprehensive review of strategies. Low-class customers visited the website frequently but made fewer purchases, indicating the need to optimize online marketing. Deals were not effectively attracting customers, especially in the low-class segment. Recommendations included enhancing deals and promotions, targeting the low-class cluster, and continuously monitoring data to make data-driven decisions. By implementing these recommendations, the company can drive customer engagement, increase sales, and expand its market presence.

## VI. References


[1] Varun, W. R., & Prabhu, P. S. (n.d.). CUSTOMER PERSONALITY ANALYSIS USING UNSUPERVISED LEARNING. https://bpb-us-w2.wpmucdn.com/sites.umassd.edu/dist/1/1282/files/2022/12/Customer_Personality_Analysis_using_unsupervised_learning_report.pdf

[2] Soumica, M., Varma, C. S., & Krishna, B. S. R. (2021, December 8). *Customer personality prediction using the ensemble technique*. International Journal of Engineering Research & Technology. https://www.ijert.org/customer-personality-prediction-using-the-ensemble-technique

[3] Widyadwatmaja, A. (2021, November 23). *Customer personality analysis segmentation (clustering)*. Medium. https://medium.com/@andhikaw.789/customer-personality-analysis-segmentation-clustering-1b68a62a61a2

[4] Pachchigar, S. N. (2021). *Customer Personality Exploratory Data Analysis* (thesis). CALIFORNIA STATE UNIVERSITY, CALIFORNIA.

[5] Dewabharata, A. (2022). Customer segmentation using the K-means clustering as a strategy to avoid overstock in online shop inventory. *Proceedings of the 1st International Conference on Contemporary Risk Studies, ICONIC-RS 2022, 31 March-1 April 2022, South Jakarta, DKI Jakarta, Indonesia*. https://doi.org/10.4108/eai.31-3-2022.2320688

[6] Rachwał, A., Popławska, E., Gorgol, I., Cieplak, T., Pliszczuk, D., Skowron, Ł., & Rymarczyk, T. (2023). Determining the quality of a dataset in clustering terms. Applied Sciences, 13(5), 2942. https://doi.org/10.3390/app13052942

[7] Geron, Aurelien. (2017). Hands-On Machine Learning with Scikit-Learn & TensorFlow : concepts, tools, and techniques to build intelligent systems . Beijing: O`Reilly.

[8] G Swamynathan M (2017) Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-2866-1