Cluster Analysis on Customer Data

Daeyoung In

Table of Contents

1. Introduction

Customer analysis is one of the most important topics of data analysis within the industry. According to a "DataMatics" survey done by McKinsey & Company, companies that extensively utilize customer analytics are more likely to be ahead of their competitors in terms of profit, sales, and sales growth (A. Bokman, et.al., 2014). Companies need to establish relationships from the data they have collected about their customers in order to maintain a healthy and well-functioning business model. The main purpose of this project is to find meaningful insight from the provided data set to help make informed decisions. In particular, we will be conducting various types of cluster analysis techniques that may provide insight for companies based on the needs and behavior of their customers. Companies can optimize their business model by analyzing the impact of certain types of customers through customer segmentation. In addition, a company may benefit from uncovering which modes of marketing are making the most impact on their company (D. Najjar, 2022). In this paper, we will conduct principal component analysis, one of the most common dimensionality reduction techniques in data analysis (I. Jolliffe, et.al., 2016). By projection onto some of its principal components, the data set will become more suitable for visualization and analysis. Furthermore, widely-used clustering methods, i.e., hierarchical clustering and K-means clustering (E. Oti, et.al., 2021), will be implemented in order to extract meaningful insight from the data. We evaluate the clustering results using silhouette scores and then compare the data points using a confusion matrix. Additional cluster analyses are implemented to investigate characteristics of the clusters and possible factors influencing customer behavior.

2. Research Plan

The first main objective of this study is to successfully form clusters in order to better understand the target of the company's product. Secondly, we want to find relationships between important variables for insight on how to improve on the current business model. For the first objective, clustering methods such as hierarchical clustering and K-means clustering will be implemented. Then we will compare the two results using a confusion matrix. In addition, we will use silhouette scores in order to evaluate how well the clusters are formed for each method (O. Arbelaitz, et.al., 2013). The silhouette score is calculated by the formula given in Figure 1 (Platform.ai, 2020). This score gives a quantitative measure of how close together the data points of a given cluster are. The best possible score is when *s(i)=1* and the worst score is when *s(i)=-1*. For the second objective, we need to begin with identifying which variables are important. One method for achieving this is to observe the correlation matrix. By analyzing the correlations between variables, we can check how different aspects of the data set are related to each other. Principal component analysis is a useful dimensionality reduction method for better understanding the data set. It can be used for visualizing the results of cluster analysis as well as identifying significant underlying variables that are initially difficult to observe.

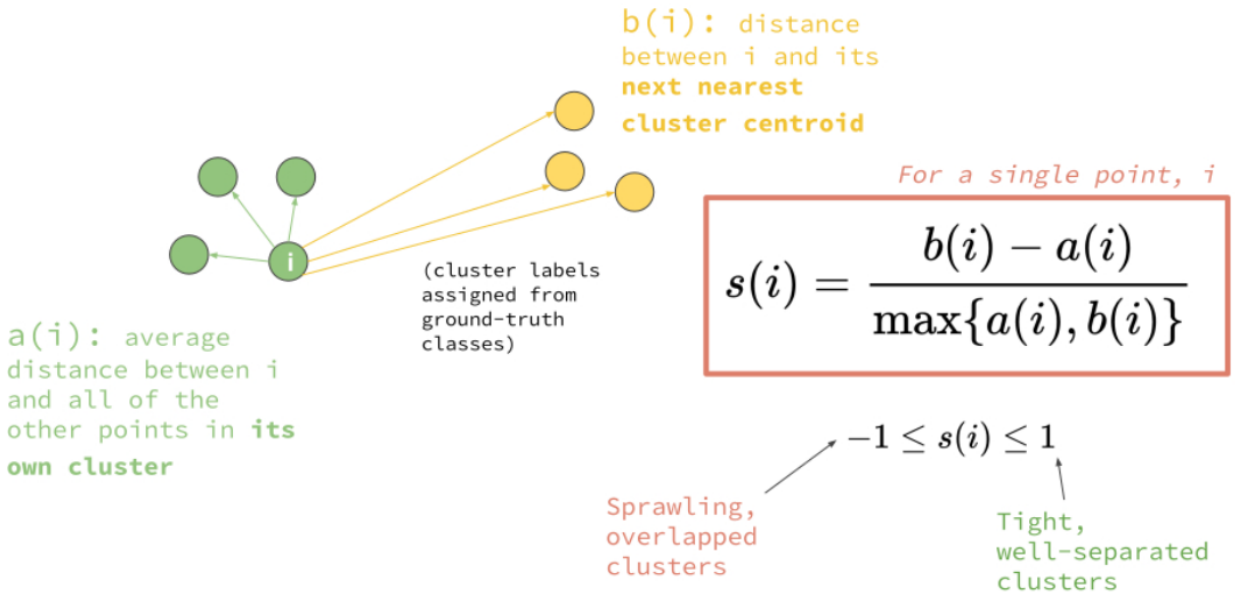


Figure 1. The silhouette score calculates how close together a group of data points are within a cluster. A silhouette score closer to 1 translates to having well-formed clusters.

Once we have reduced the data set to its principal components, we can visualize the clusters obtained by hierarchical and K-means clustering, respectively. Hierarchical clustering is based on calculating the dissimilarity of different data points using pairwise distances. There are various types of distances that can be used such as single linkage, average linkage, Ward's method, or centroid method. The method chosen for this analysis is Ward's method which minimizes the

variance of the clusters (R. Johnson, et.al., 2007). The algorithm of hierarchical clustering begins with a distance matrix that contains the pairwise distances of all observations. If we consider each observation to be a cluster, the most similar clusters are merged into one, thereby reducing the size of the distance matrix. This process is repeated until all observations are merged into a single cluster. It is advantageous in that it does not require initial determination of the number of clusters, since the algorithm requires all observations to be merged until a single cluster is left. The user is free to choose the number of clusters or they can set a threshold value on the distance between clusters. Although not implemented in this study, hierarchical clusters can also be visualized using a dendrogram which depicts the distances between different data points. However, since the algorithm relies on the pairwise distances, there is a relatively high amount of computation required, i.e. $O(n^2)$, and thus it may not be suitable for large sets of data. Furthermore, misclassification of an observation cannot be reversed, so one should especially be careful of outliers before implementing this method. K-means clustering, on the other hand, is better-suited for large data sets. The algorithm requires pre-determining the number of clusters K, hence the name 'K-means'. K randomly selected observations are designated as the initial centroids, and depending on distance, the other observations are assigned to one of these centroids, thus forming K initial clusters. A new centroid can be calculated for each of these clusters and each observation can be reassigned to a different cluster according to the new centroids. This process is repeated until the centroid converges to a single point. K-means clustering is O(n) so it is computationally faster than hierarchical clustering. However, it can sometimes be difficult to draw a meaningful conclusion from the result obtained by K-means clustering.

3. Data Pre-Processing

The given data indicates various information on customers regarding their socioeconomic status and living conditions, as well as interactions with the company such as purchases, engaged marketing channels, and number of promotions accepted. The data provides information on a total of 2,240 customers with their respective ID numbers as well as 28 other variables. We implement some data pre-processing methods to better fit the data for customer analysis (Thecleverprogrammer, 2021). Some of the variables such as 'Marital_Status' and 'Education' have values we can merge together. For example, we merged responses for 'Marital_Status' from 'Single', 'Widow', 'Divorced', 'Absurd', and 'YOLO' into a single category, 'Alone'. In addition, not all of the variables are usable for customer analysis so such redundant variables are removed. Furthermore, we can extract new variables from existing variables so that they can be in a more usable form. 'Age' is extracted from the variable 'Year_Birth'. 'Spent' is the total amount spent on each category of products. 'Marital_Status' is changed to a new variable 'Living_With' as we merge the data as described previously. 'Children' variable is created by summing 'Kidhome' and 'Teenhome' variables. Finally, 'Family_Size' and 'Is_Parent' variables are obtained from 'Living_With' and 'Children' variables. In addition, 'Education' data is integrated and then

divided into three categories: ‘Undergraduate’, ‘Graduate’, and ‘Postgraduate’. A total of 30 variables are left and the complete list after pre-processing is shown in Table 1. The variables in bold denote features extracted from the raw data.

Variable Name	Description
Education	Customer’s education level
Income	Customer’s yearly household income
Kidhome	Number of kids in customer’s household
Teenhome	Number of teenagers in customer’s household
Recency	Number of days since customer’s last purchase
Wines	Amount spent on wine in last 2 years
Fruits	Amount spent on fruits in last 2 years
Meat	Amount spent on meat in last 2 years
Fish	Amount spent on fish in last 2 years
Sweets	Amount spent on sweets in last 2 years
Gold	Amount spent on gold in last 2 years
NumDealsPurchases	Number of purchases made with a discount
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
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
Complain	1 if the customer complained in the last 2 years, 0 otherwise
Customer_For	Number of days the customers started to shop in the store relative to the last recorded date
Age	Calculated from the birth year of the customer
Spent	Total amount spent in all categories in last 2 years
Living_With	Living situation of couples extracted from marital status
Children	Total number of children (kids and teenagers) in customer’s household
Family_Size	Total number of people in customer’s household
Is_Parent	Indicates parenthood status

Table 1. The variables for customer data after pre-processing. Variables in bold denote features extracted from the raw data.

There are some missing values for the ‘Income’ variable. The missing values are filled in using imputation by regression. From plotting some of the variables, we can observe some outliers in our data as seen in Figure 2(a). The outliers can be removed by setting caps of 600,000 on income and 90 on age. In addition, the categorical variables, ‘Education’ and ‘Living_With’ are transformed into numerical variables using one-hot encoding. Then the data is standardized using ‘StandardScaler’ from the ‘sklearn’ module. Furthermore, by observing some of the correlations between variables, we can get an overall idea of our data as well as to check for excessively high

correlations. A color map is used to help visualize the correlation strengths as shown in Figure 2(c). We extracted variables with a correlation factor greater than or equal to 0.8. High correlations were found between 'Spent', 'Meat', and 'Wines' variables, as well as between 'Children' and 'Family_Size' variables. This is expected as we created the variables. 'Spent' and 'Family_Size' using these other variables. Overall, the data is clean and does not have an excessive number of highly correlated variables, so this is the finalized data that we proceed with for analysis.

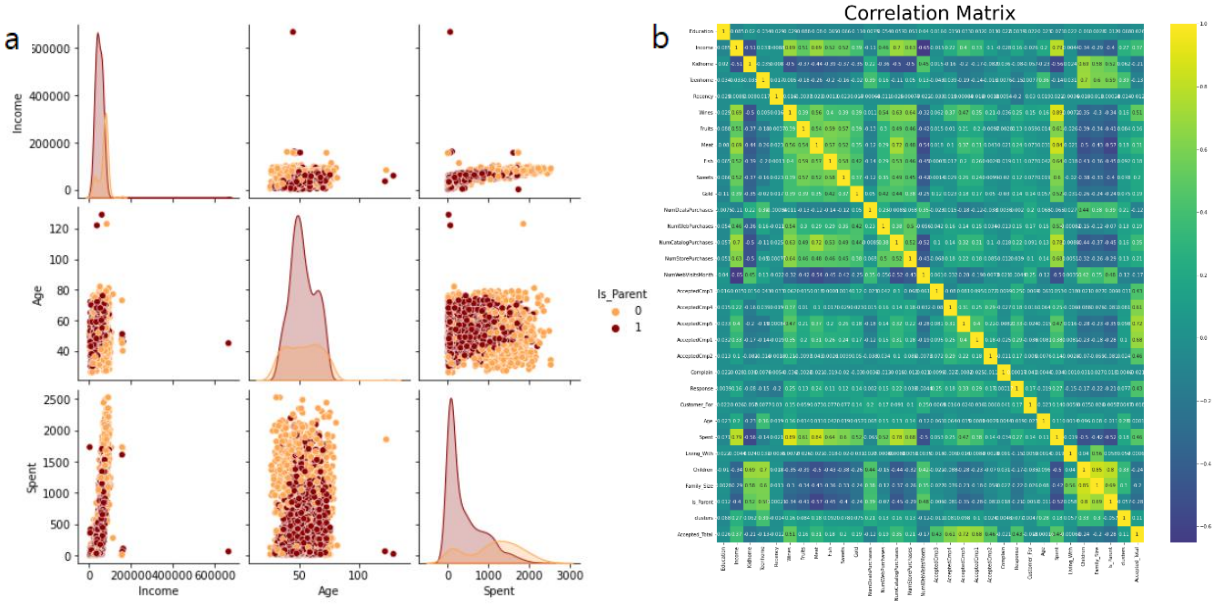
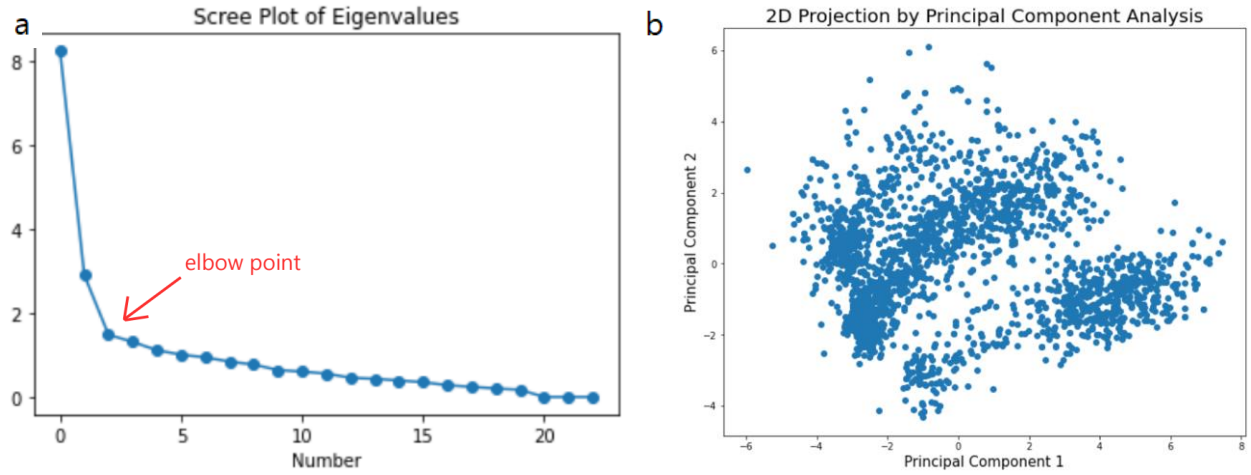


Figure 2. (a) The scatter plots between different variables color-coded by parenthood. From top to bottom (left to right), the variables are 'Income', 'Age', and 'Spent'. **(b)** The correlation matrix of all variables of the customer data after pre-processing. The values of the color legend are from -0.6 to 1.

4. Data Analysis and Visualization

By reducing the dimensions of our data using principal component analysis, we can achieve a better picture of the clusters. Before doing so, we need to choose an appropriate number of dimensions. One of the most common methods for determining the number of dimensions is using a scree plot. The variables of the original data can be expressed as a linear combination of some new variables. By spectral decomposition, the linear combination can be expressed in such a way that the eigenvectors of the correlation matrix are the coefficients and each term is a principal component. Then the eigenvalues of the correlation matrix account for the variance of each principal component. By computing the correlation between the first two principal components and the original variables, it can be observed that the first principal component is highly correlated with variables related to the financial aspects of customers such as income and amount spent. The second principal component is highly correlated with variables related to the number of household members, but the underlying variables are not as distinct as the first principal component.



e

K-means	0	1	2	3	Total
Hierarchical					
0	584	6	18	82	690
1	0	499	24	2	525
2	0	1	471	0	472
3	24	81	0	444	549
Total	608	587	513	528	2236

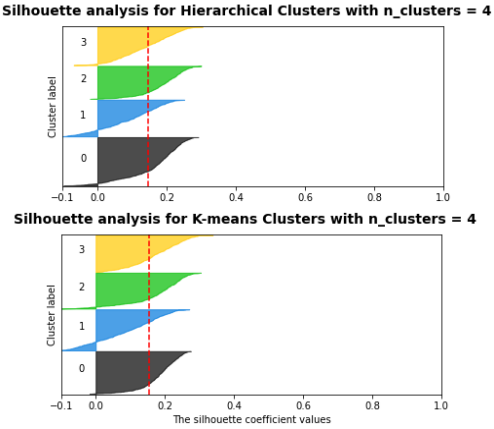


Figure 3. (a) Scree plot of the eigenvalues of the correlation matrix. The slope of the scree plot is steep up until the elbow point. **(b)** 2-dimensional projection of the data set on the first and second principal components. 2-dimensional projection with **(c)** hierarchical clusters and **(d)** K-means clusters. **(e)** Confusion matrix and **(f)** silhouette plots of the clusters obtained from hierarchical and K-means clustering.

The number of principal components can be determined using a scree plot which depicts the rate at which the eigenvalues decrease. Selecting a small number of principal components that

account for a large amount of variance is desirable because information loss should be minimized. Thus, we chose the number of principal components such that adding another principal component does not significantly increase the explained variance. The corresponding point in the scree plot is called the elbow point. From the elbow point in Figure 3(a), it is appropriate to reduce our data to 2 dimensions. The data points are projected to a 2-dimensional space comprised of the first two principal components as shown in Figure 3(b). Clustering methods can be implemented to distinguish the data points projected on the 2-dimensional space. First, hierarchical clustering was conducted using the 'sklearn' module (Figure 3(c)). K-means clustering was also implemented using the same module (Figure 3(d)). From the two cluster results, the hierarchical clusters have more well-defined borders, whereas, the K-means clusters have some overlapping points between clusters. If we compare the clustering results using a confusion matrix as shown in Figure 3(e), the majority of the data points are similar except for some observations, especially those of clusters 1 and 3, that are not matching. The similarity of the two clustering results can easily be calculated as 89.36% from the diagonal components of the confusion matrix. Finally, the silhouette scores can be computed to evaluate how well the clusters are formed. The silhouette plots in Figure 3(f) show that the two clustering methods have similar scores. The average silhouette scores are approximately 0.1461 and 0.1547 for hierarchical clustering and K-means clustering, respectively.

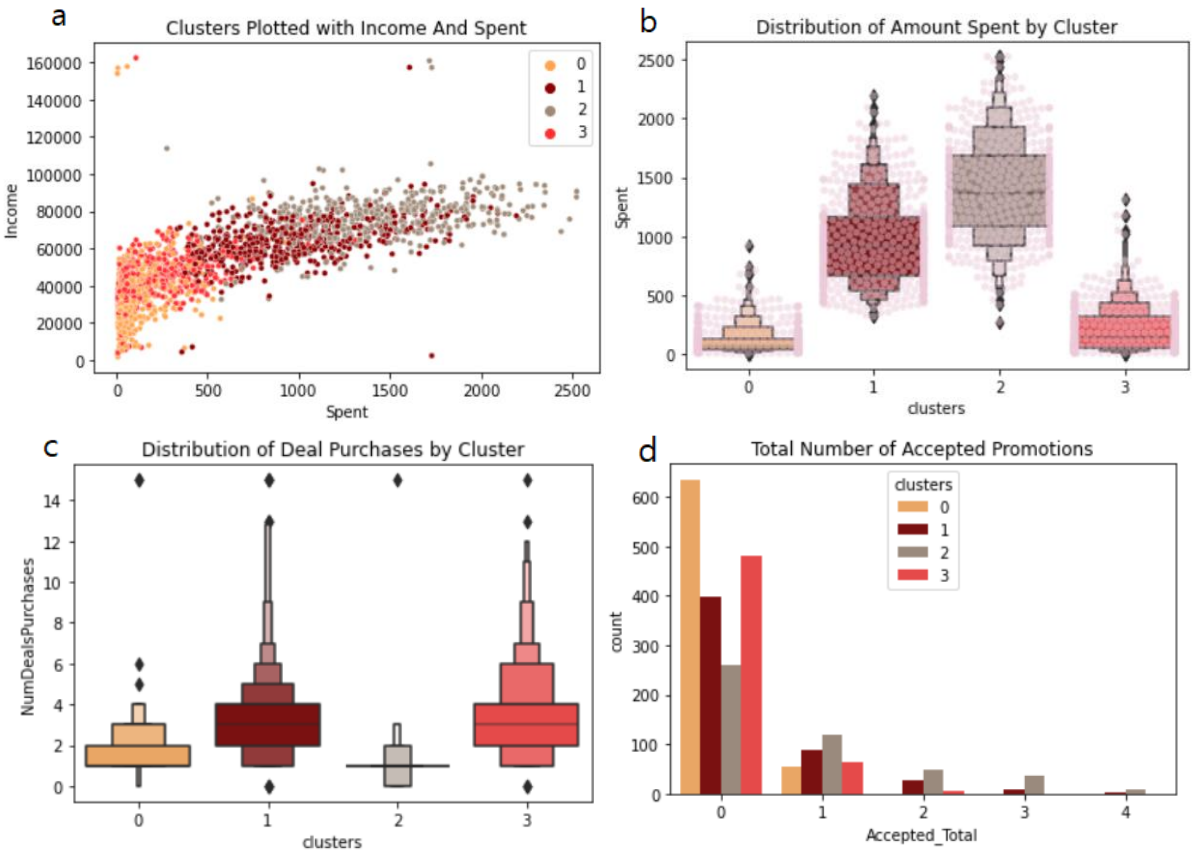


Figure 4. (a) Scatter plot of 'Income' and 'Spent' variables. **(b)** 'Spent' variable of each cluster depicted by swarm plot and letter-value plot. **(c)** Letter-value plot depicting the distribution of the number of deals purchased for each cluster. **(d)** Bar plot of the total number of accepted promotions for each cluster.

There is a trend in the observations where the clusters can be categorized based on their income and amount spent as seen in Figure 4(a). The high spending clusters are clusters 1 and 2, with cluster 1 being the lower income group. The low spending clusters are clusters 0 and 3, with cluster 0 being the lower income group. From the swarm plot in Figure 4(b), it is clear that cluster 2 spends the most, thus they should be prioritized. In addition, it can be deduced from the letter-value plot in Figure 4(c) that the deals offered to customers were relatively successful. However, the number of deal purchases was the worst for cluster 2. The company should benefit from focusing their marketing on cluster 2 which is the priority target group. Furthermore, not many of the five promotion campaigns have been very successful as shown in Figure 4(d), so this is one area that the company may need to improve. Finally, we observed bivariate relationships from some of the variables using a joint plot with Gaussian kernels as shown in Figure 5. The variables chosen for this analysis are those related to the personal traits of the customers such as (a) education, (b) age, (c) family size, and (c) number of children. The distribution of education was quite similar for all clusters with cluster 0 having a slightly higher number of postgraduates. There were slight age differences between clusters with cluster 0 being relatively younger, and clusters 1 and 2 being relatively older. In addition, clusters 0 and 2 have smaller family sizes than the other two clusters. Lastly, customers in cluster 0 have one or no children, whereas, customers in cluster 3 have at least one child. It may be useful to take these different characteristics into account when planning new products and marketing strategies that are targeted to specific cluster groups.

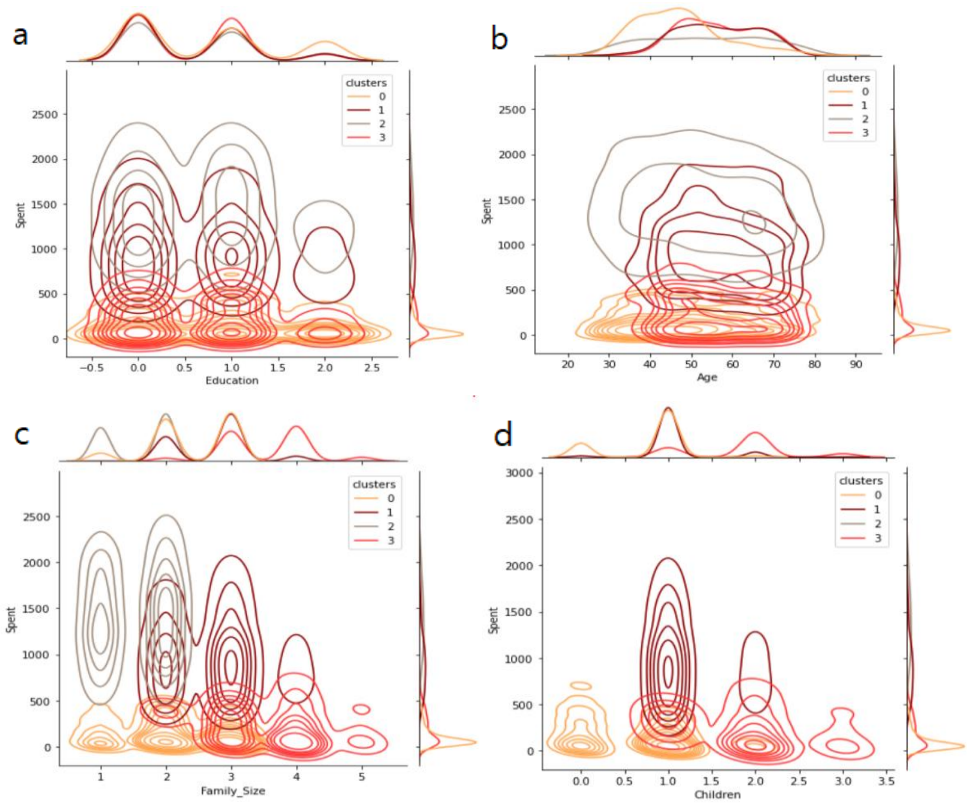


Figure 5. Joint plots of 'Spent' variable plotted against **(a)** education, **(b)** age, **(c)** family size, and **(d)** number of children using Gaussian kernel densities.

5. Conclusion

Principal component analysis was implemented in order to visualize potential clusters in the data set. By using a scree plot, we were able to determine the number of principal components while minimizing information loss. The correlations between the principal components and the original variables revealed some of the underlying factors within the data set. However, it was difficult to pinpoint the underlying variables of each principal component, especially for the second principal component. This is probably due to the large number of variables and correlations existing between them which may have accounted for some noise. For reference in future studies, removing some more of the highly correlated variables may improve the clustering results. After projection onto the first two principal components, we utilized hierarchical and K-means clustering to detect cluster patterns within the data set. A confusion matrix was used to measure the similarity of the cluster results, and silhouette plots were used to evaluate how well the clusters were formed. The clusters were found to be approximately 89.36% similar by the confusion matrix. Although the silhouette scores were quite similar, the hierarchical clusters are slightly more preferable because there is less overlapping between different clusters. In addition, various visualization tools were used to analyze the characteristics of the different clusters. In particular, the joint plots provided insight on the personal traits of the customers, thus shining a light on how to plan marketing strategies. For example, advertising specifically focused on grabbing the attention of customer households with children may be effective if the goal is targeting customers belonging to cluster 3. Furthermore, cluster 2 was found to be the highest spending group and thus should be set as the top priority group in order to maximize profit. In conclusion, we performed principal component analysis and clustering techniques to extract some useful information from the customer data. These methods proved to be useful for providing insight on the current state of a company's business model in relation to its customers. Thus, this approach can be implemented for further industrial applications on similar types of data sets.

6. References

1. Dataset: Akash Patel, August 2021. marketing_campaign.csv, ver. 1. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

2. Alec Bokman, Lars Fiedler, Jesko Perrey, Andrew Pickersgill, **Five facts: How customer analytics boosts corporate performance** (2014) , McKinsey & Company, https://www.mckinsey.com/business-functions/growth-marketing-and-sales/our-insights/five-facts-how-customer-analytics-boosts-corporate-performance

3. Dennis Najjar, **What is Customer Analysis and How Can it Help You?** AccountingDepartment.com LLC, https://www.accountingdepartment.com/blog/what-is-customer-analysis-how-does-it-benefit-business-analytics

4. Ian T. Jolliffe, Jorge Cadima, **Principal component analysis: a review and recent developments.** "Philosophical Transactions Royal Society Publishing A" (2016), pp. 374-377

5. Eric U. Oti, Micahel O. Olusola, Francis C. Eze, Samuel U. Enogwe, **Comprehensive Review of K-Means Clustering Algorithms.** "International Journal of Advances in Scientific Research and Engineering" (2021), pp. 64-65

6. Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesus M. Perez, Inigo Perona, **An extensive comparative study of cluster validity indices.** "Pattern Recognition" (2013), pp. 245-255

7. Platform.ai, **The Silhouette Loss Function: Metric Learning with a Cluster Validity Index** (2020), https://www.platform.ai/post/the-silhouette-loss-function-metric-learning-with-a-cluster-validity-index

8. Richard A. Johnson, Dean W. Wichern, **Applied Multivariate Statistical Analysis** (2007), pp. 692-694

9. Thecleverprogrammer, **Customer Personality Analysis with Python** (2021), https://thecleverprogrammer.com/2021/02/08/customer-personality-analysis-with-python/

10. Karnika Kapoor, **Customer Segmentation: Clustering** (2021), https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering