MIC 2023
Management International Conference
Trieste, Italy • 7–10 June 2023

Preliminary Study in Using Large Language Models as Interpretation Assistants in Cluster Analysis for Customer Segmentation

Ante Panjkota
University of Zadar, Department of Information Sciences, Croatia
ante.panjkotat@unizd.hr

Aleksandra Krajnović
University of Zadar, Department of Economics, Croatia
aleksandra.krajnovic@unizd.hr

Josipa Perkov
University of Zadar, Department of Economics, Croatia
josipa.perkov@unizd.hr

Abstract. Marketing experts have used clustering methods in customer segmentation tasks for over two decades. Besides the problem of defining a number of clusters and appropriate evaluation metrics to obtain high intra-cluster similarity and high inter-cluster dissimilarity, one of the main concerns in this field is an understanding of final cluster grouping by domain experts. This paper explores the possibility of using the Large Language Model (ChatGPT) as a cluster interpretation assistant for marketing experts in the field of customer segmentation using K-means, DBSCAN, Agglomerative clustering, and Fuzzy C-means algorithms on three publicly available datasets. The datasets differ in size, feature number (dimensions), level of knowledge related to the datasets, and underlying phenomenon that generates data. In that sense, the focus is on the clustering results that reasonably suit the best interpretation of ChatGPT fed with additional knowledge on datasets. The obtained results confirmed the initial hypothesis that the quality of ChatGPT interpretations related to the final clusters is more dependent on additional knowledge than the success of the clustering procedure.

Keywords: customer segmentation, clustering, large language model, ChatGPT, interpretation assistant

1 Introduction

Customer segmentation is a well-known and extensively studied phenomenon in marketing that employs different approaches, from statistical to machine learning and deep learning (Singh 2021; Alves Gomes and Meisen 2023) Various clustering techniques have been successfully used for the same task, including K-means (Kansal et al. 2018), DBSCAN (Sembiring Brahmana, Mohammed, and Chairuang 2020), Hierarchical clustering (Abdulhafedh 2021), Fuzzy C-means (Munusamy and Murugesan 2020), and many others (Kaur and Kiranbir 2017). Besides standard problems like data preprocessing, coping with the curse of dimensionality with dimensionality reduction techniques (Khalid, Khalil, and Nasreen 2014), and determining the number of clusters and perform clusters validity check (Alves Gomes and Meisen 2023), applications of clustering techniques in customer segmentation require appropriate interpretation of obtained clusters. With the latest development of large language models (Li et al. 2021), investigating the possibilities of such advanced systems in different areas of human endeavours is worth researching direction.

This paper presents a preliminary study on the possibilities of utilizing a large language model (e.g., ChatGPT) to support marketing experts in cluster analysis for customer segmentation and, more precisely, for assistance in interpreting obtained clusters. The goal was to test this research assumption through experiments on publicly available real datasets by employing different clustering techniques.

The datasets differ in size, description details, and underlying market phenomena. On the other hand, the possible impact of the clustering method was tested by employing well-known clustering methods, K-means, DBSCAN, Aggregative clustering, and fuzzy C-means. The rest of the paper is organized as follows. The theoretical background section provides a short overview of customer segmentation problem, the use of clustering methods in that task, and evidence for the basis to investigate possible role of Large Language Models as interpretation assistants in the customer segmentation problems. The third section describes the developed methodology approach in dept. The main results are presented and discussed in the fourth section, while the last section summarizes the findings and emphasizes the shortcomings of this preliminary study in the conclusion section. The discussion also provides some directions for possible future improvements.

2 Theoretical backgrounds

Customer segmentation is a task where all company's customers with legally obtained records can be divided into distinct groups based on similarities described by relevant features (demographics, geographics, psychographic factors, behavioural factors, etc.). The obtained groups need to adhere to real-world logical explanations and descriptions.

Segmentation goals can vary from customer retention, which directly improves company revenue (Dawane, Waghodekar, and Pagare 2021), to the development of efficient target marketing strategies (Guoan ZHU and Xue GAO 2019), or even toward creating new products or services (Nasiopoulos et al. 2015). So, the importance of customer segmentation is well established, which implies the necessity of developing and using efficient and reliable methods for segmentation.

2.1 Clustering in customer segmentation

Among widely used methods for customer segmentation, clustering methods have an essential role. Moreover, the simple K-means algorithm is frequently used (Kansal et al. 2018; Alves Gomes and Meisen 2023) due to its simple implementation and ease of use. The main problem with this algorithm is that it can adequately work only with convex cluster structures. Density-based clustering algorithm DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and its derivatives can detect convex and non-convex shapes (Kamran et al. 2014) and have been used in the problems of customer segmentation (Wang et al. 2020).

Hierarchical clustering methods generally suffer from early mistakenly formed clusters that propagate until the end of the clustering procedure. Their main advantage to partition clustering methods (e.g., mentioned K-means and DBSCAN) is the level of provided details, on average shorter execution time, and sound visualization principle in the form of dendrograms. Agglomerative Clustering is a representative of hierarchical clustering algorithms with frequent application in customer segmentation (Hung, Thuy Lien, and Ngoc 2019).

These methods fall into so-called crisp clustering algorithms where a customer can only belong to one cluster. In real life, the belonging principle is not defined so strictly, especially in the systems that involve humans. Soft clustering methods are developed to model the uncertainty principle of belonging to some groups. A Fuzzy C-means algorithm is the first successful implementation of the soft clustering principle and can be employed for customer segmentation (Pradipta et al. 2018).

Besides the mentioned clustering methods used in customer segmentation differing in underlined properties, there is also diversity in the association with the final clusters, even when the number of clusters is the same. This phenomenon can also be seen in the same clustering method, e.g. K-means with different initialization methods will end up with different final cluster associations (Steinley and

Brusco 2007), or using various linkage measures in Agglomerative Clustering (Emmendorfer 2019). So, this implies that the clustering method's effect needs to be considered in the customer segmentation task.

All clustering algorithms will produce some grouping or final cluster association of customers. Still, the crucial problem, besides cluster verification, is an adequate interpretation of obtained clusters and appropriate meaning in the context of underlying business data generation principle.

2.2 Clusters Interpretability Problem

In the last years, there have been tremendous efforts in the research community to add transparency to AI systems besides their proven powerful prediction abilities. As a result of that effort, the new field known as Explainable AI is established (Adadi and Berrada 2018). Two main aspects are of the utmost importance in that field: model explainability and model interpretability. In many studies, authors used the mentioned terms interchangeably, but in the essence, they are distinct. Explainability is related to humans understanding the whole or at least the main part of models or systems process in generating solutions based on input data. At the same time, interpretability is the ability of humans to understand the meanings the outputs of the models in the context of domain knowledge (Broniatowski 2021). In this paper, we adhere to these definitions.

Clustering, one of the essential research parts of Machine Learning, suffers from the same problems addressed in the previous paragraph. Most clustering algorithms produce final cluster assignments that are complex to understand due to the overall dependence of all features. In general, all possible interpretations of cluster membership may become intractable, and sometimes, it is impossible to grasp the most crucial dependencies (Dasgupta et al. 2020). Additional problems to the interpretability are cluster overlapping and subcluster structures in the main clusters (Parsons et al. 2004). These problems are tackled with different strategies: using visualization techniques (Mucha, Hans-Joachim, Hans-Georg Bartel 2012) or developing more explainable and interpretable clustering models (Basak and Krishnapuram 2005; Fraiman, Ghattas, and Svarc 2013; Esfandiari, Mirrokni, and Narayanan 2022; Makarychev and Shan 2022; Saisubramanian, Galhotra, and Zilberstein 2020).

With the rapid development in the field of Large Language Models and their potential as assistants in many tasks, it is worth exploring the possibility of using these models as interpretation assistants in clustering problems as complementary tools to established techniques and procedures.

2.3 Large Language Models

A Large Language Model (LLM) is an artificial intelligence system trained on vast amounts of textual data. Its primary goal is interacting with humans by understanding textual inputs and generating relevant and optimal responses based on learned patterns. The breakthrough paper that led to the development of these popular intelligent systems is "The Attention Is All You Need" (Vaswani et al. 2017), where the authors introduced the Transformer architecture, which represents the backbone of all contemporary LLMs. The second important paper that brings the so-called alignment principle of LLM to prevent or reduce hallucination, unhelpful, untruthful, or even toxic outputs is based on Reinforcement Learning with Human Feedback (Ouyang et al. 2022).

Considering the nature of LLMs, it is natural to explore applications as domain assistants in different fields. In that sense, LLMs are successfully used as programming assistants (Moradi Dakhel et al. 2023). Some researchers focused on the investigation of LLMs as writing assistants in higher education (Imran and Almusharraf 2023), and writing assistants in science in general (Salvagno, Taccone, and Gerli 2023), as assistants in healthcare and medicine (Dave, Athaluri, and Singh 2023), or as assistants in improving online marketing (Mutoffar et al. 2023).

There is evident broad interest in the scientific community in the possibility of using LLMs as assistants for different tasks. That gives an alibi to investigate possibilities of such systems in customer segmentation via clustering methods or, more precisely, as an assistant in interpreting obtained clusters.

3 Methodology

To test the possibility of LLMs as interpretation assistants in customer segmentation via a clustering approach, the focus of this preliminary study is on the following research questions:

[RQ-1]: Does the LLMs interpretation ability of the final clusters depend on the employed clustering method?
[RQ-2]: Does the quality of LLMs' interpretations of final clusters in customer segmentation problems vary based on prior knowledge of data and the underlying process that generates data?

As the LLMs representative, we decided to use ChatGPT ("OpenAI" 2023) as one of the most advanced LLMs available for free or as a Plus version via a monthly subscription. We used the ChatGPT Plus version with the Noteable plugin in this study. Clustering is performed with four clustering algorithms as follows:

- K-means (sklearn.cluster.KMeans),
- Agglomerative clustering (sklearn.cluster.AgglomerativeClustering),
- DBSCAN (sklearn.cluster.DBSCAN),
- Fuzzy C-means (skfuzzy.cluster.cmeans).

As the input parameter for K-means, Agglomerative clustering, and Fuzzy C-means algorithms, the number of clusters is estimated with repeated silhouette method changing the random seed.

Figure 1: Repeated silhouette method for cluster estimation in the first dataset with random seeds changes

From Figure 1 estimated number of clusters is 5 for the first dataset. The same approach is repeated for all datasets. DBSCAN has an integrated cluster estimation procedure in the algorithm, so it does not require this input parameter. Three clustering experiments in customer segmentation were carried out with different datasets chosen to investigate the second research question. Table 1 summarizes basic information related to used datasets. This experimental design implies that all three experiments also investigate the first research question of interest.

Table 1: Datasets information

Domain	Detailed description	Features description	Features	Missing values	Size	Reference (access)
Fictional BMW dealership	No	Basic	8 numeric	No	100	(IBM 2010)
Credit risk - finance	No	Yes	5 categorical 4 numeric	Yes	1000	(KAGGLE, n.d.)
Customer personality analysis	Yes	Yes	25 numeric 3 categorical	Yes	2240	(Romero-Hernandez, n.d.)

The first dataset is the simplest fictional dataset with a small size and superficial knowledge about the underlying problem with a basic description of the features. The second dataset is a simplified dataset obtained from the well-known German Credit Data from the UCI Irvine Machine Learning Repository. German credit risk problem has a rudimental description and solid features descriptions. There are substantial missing values in the categorical attributes, *Saving accounts,* and *Checking account,* so they are replaced with the most frequent values.

The last dataset comes with a detailed problem description and detailed knowledge of features. Missing values in this dataset are rare, and instances with missing values can easily be dropped without significant loss of information. After dropping those instances, the dataset size was 2216. Categorical features in the second and third datasets were encoded into numerical values using MinMaxScaler from the scikit-learn library. Fundamental feature engineering on the third dataset gives:

- Age feature – the age of the customer calculated as the difference between the current year and Year_Birth
- Total_Spending – as the sum of all spending on particular products
- Relationship_Status – all values are transferred to Single or Couple
- Family_Size – as the sum of Relationship_Status (Single=1, Couple=2) and value of Children feature
- Is_Parent – encoded as 0 or 1 based on Children's values.
- Education – transferred to UnderGraduate, Graduate, PostGraduate

Some features are renamed for the sake of clarity (MntWines: Wines, MntFruits: Fruits, MntMeatProducts: Meat, MntFishProducts: Fish, MntSweetProducts: Sweets, MntGoldProds: Gold), and redundant features are dropped:

- Marital_Status,
- Dt_Customer,
- Z_CostContact,
- Z_Revenue,
- Year_Birth,
- ID

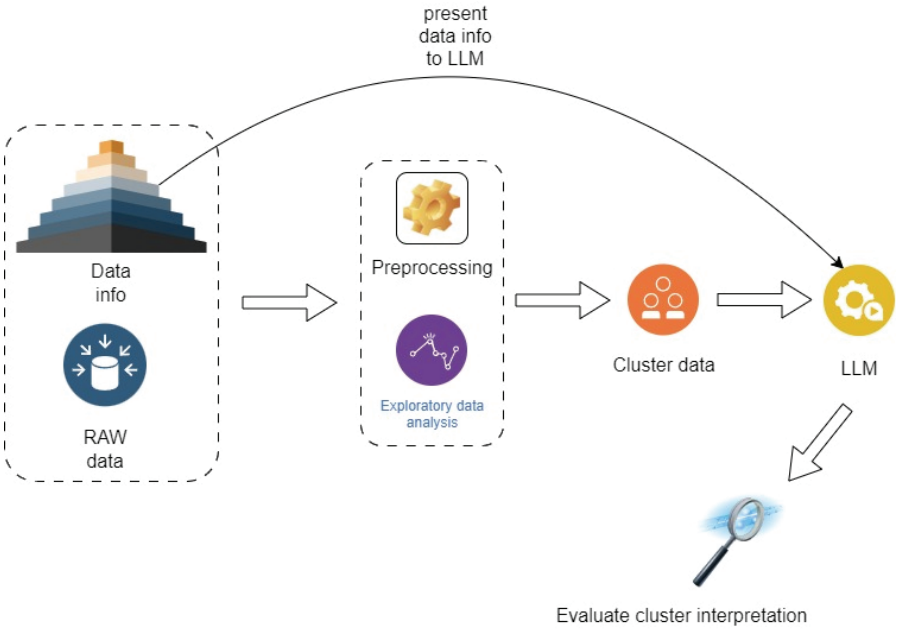
Exploratory analysis shows that features Age and Income contain some outliers and can be removed by thresholding values (dataset size is 2012). Features that deal with acceptance and promotions are also dropped.

The final dataset is comprised of 19 features*: Education, Income, Kidhome, Teenhome, Recency, Wines, Fruits, Meat, Fish, Sweets, Gold, Complain, Customer_for_days, Age, Total_Spending, Relationship_Status, Children, Family_Size, and IS_Parent*. This number of features within 2012 instances represents a problem due to the curse of dimensionality for all clustering algorithms of interest. To resolve this problem, we can employ a feature selection or feature extraction approach (Zebari et al. 2020). In this study, we have implemented RRFS (Relevance Redundancy Feature Selection Method) (Ferreira and Figueiredo 2012), which was repeated for different values of parameter eta in the recommended range [0.4, 0.9] with the step of 0.05. Finally, obtained selected features from all rounds are counted and sorted according to a number of occurrences, which gives the list of the N best features for a particular dataset. With this approach, selected features are: *Relationship_Status, Family_Size, Education, Total_Spending, Customer_for_days, Wines, Fish, Fruits, Meat, Sweets*.

For the second dataset, exploratory data analysis shows no significant difference between men and women in credit debt data; the same can be inferred for housing categories.

Clustering was performed on pre-processed, and, in general, analyzed datasets, and clustering results were presented to ChatGPT as final centroids. There is no such notion as cluster centroids for Agglomerative clustering, but implicitly, we can assume them through the mean of each obtained cluster. ChatGPT interprets clustering results in combination with previously fed dataset info on the presented level shown in Table 1. Finally, ChatGPT interpretation results were evaluated by three marketing experts. A diagram of the developed methodology is depicted on Figure 2:

Figure 2: Proposed methodology for testing possibility of LLMs as interpretation assistant in customer segmentation[1]



[1] Diagram was prepared with the free online tool available at https://app.diagrams.net/

4 Results and Discussion

In this section we will present only most interesting results. The data was fed into the chatGPT using the same strategy:

- We delivered the dataset information,
- The clustering results in the form of final centroids were presented,
- Prompting to interpret of obtained results based on all available information is performed.

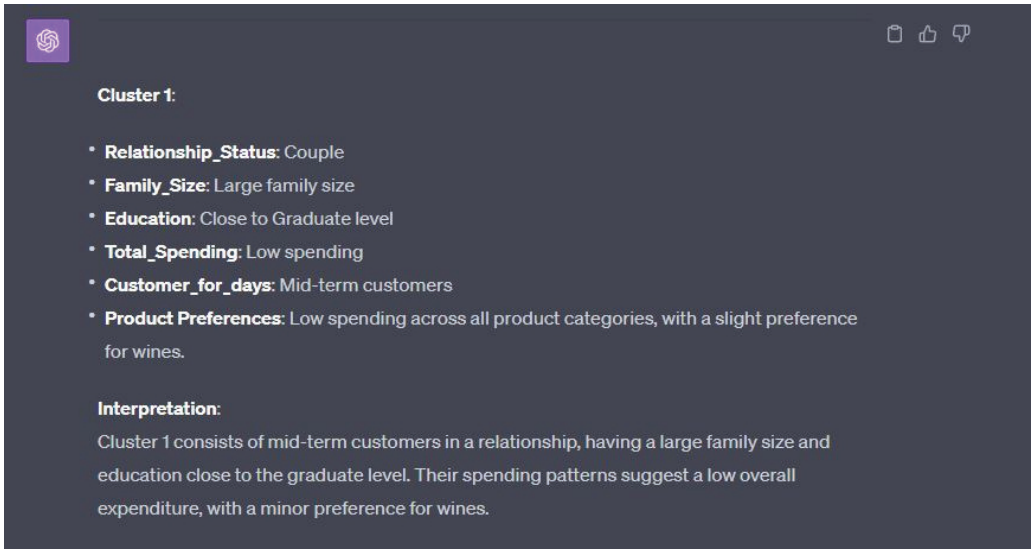
For the third dataset K-means algorithm produces final centroids presented in Table 2:

Table 2 K-means final centroids for the third dataset

R_S	F_S	Education	T_S	C_f_D	Wines	Fish	Fruits	Meat	Sweets	Cluster
0.0	0.3004	0.7678	0.0902	0.4584	0.0845	0.0418	0.0423	0.0289	0.0343	0
1.0	0.5680	0.8625	0.0913	0.4558	0.0896	0.0412	0.0382	0.0282	0.0320	1
0.0	0.0908	0.7897	0.5290	0.5228	0.4302	0.3491	0.3171	0.2326	0.2388	2
1.0	0.3695	0.8616	0.5297	0.5114	0.4494	0.3163	0.2918	0.2299	0.2236	3
1.0	0.4626	0.0000	0.1358	0.5302	0.0862	0.1619	0.1214	0.0489	0.1057	4

Chat GPT interpretation for the cluster with index one is shown in Figure 3, and the summary for all clusters is in Figure 4. Representation in the form of the final centroids has substantial limitations. Still, there must be some form in which data are presented to LLMs, in this case to the ChatGPT, considering the limitations of interactions throughout chat prompts. Later, we will talk about the possibilities that currently offer ChatGPT plugins. This is more important, knowing that final centroids have no real meanings in the Agglomerative clustering and DBSCAN and, for this study, are imposed in the form of mean values for each cluster.

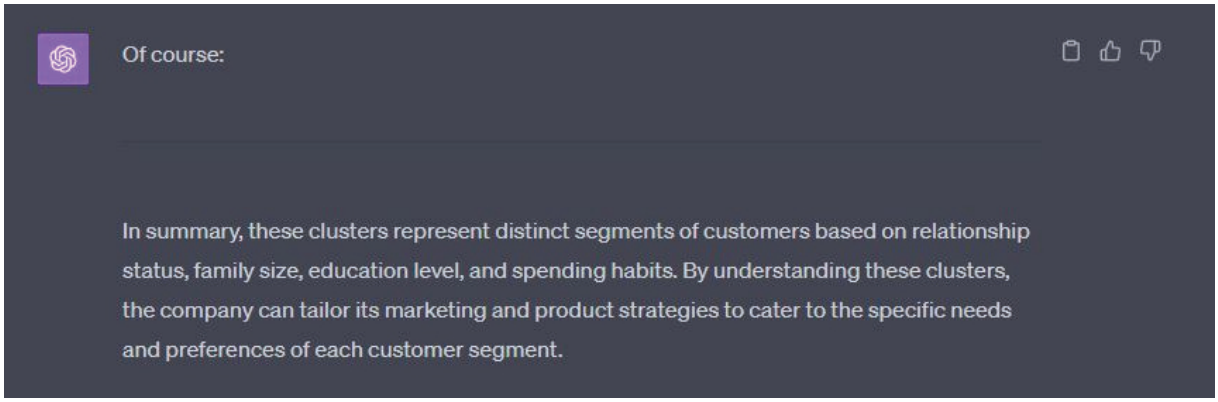
Figure 3: Interpretation results of ChatGPT for the second cluster (index position 1) of the third dataset clustered with the K-means algorithm



Results of all used clustering algorithms are presented to the three marketing experts, who then evaluate the LLM interpretation results of obtained clustering results. K-means results are the base case for comparison of three other used clustering algorithms in the task of customer segmentation. Evaluations of the quality of the generated interpretation of the LLM are subjective and are scaled with values very

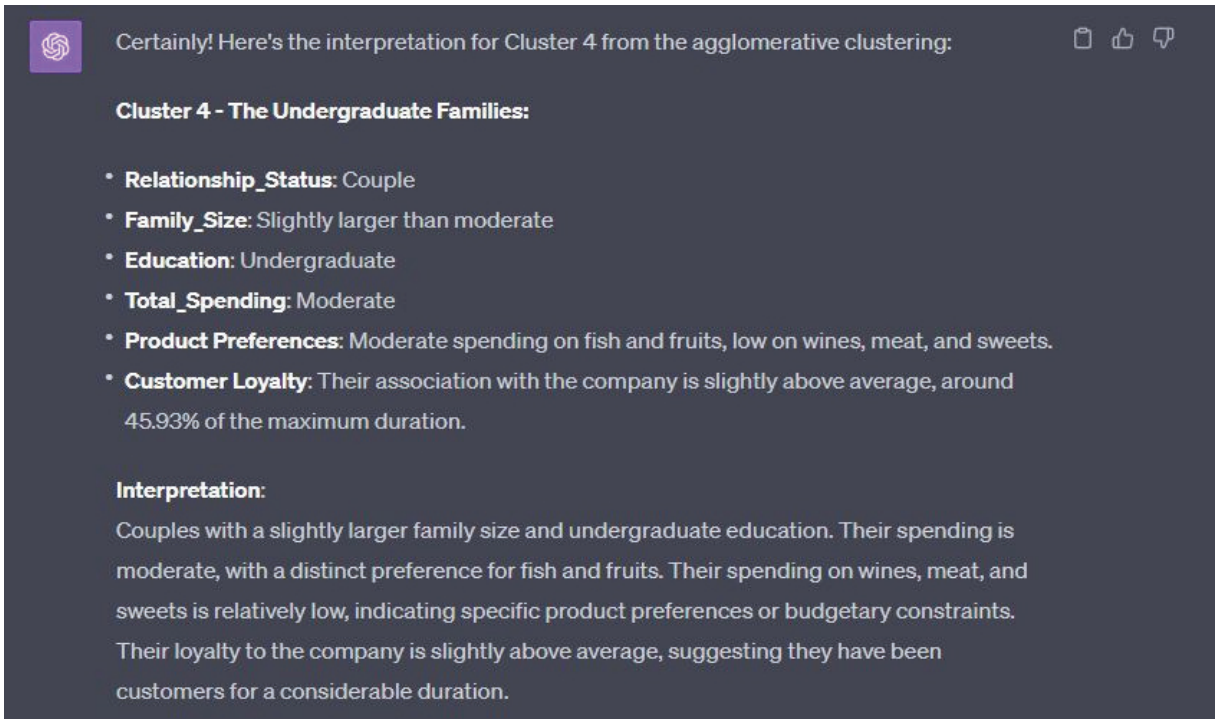
poor, poor, satisfactory, good, very good, and excellent. Scale for comparison of the impact of the used algorithm for clustering is no difference, slight difference, medium difference, and significant difference again compared with the base case of the results of the K-means algorithm. Besides a small number of evaluators, this evaluation procedure is a primary limitation of this preliminary study.

Figure 4: Summary tailored by ChatGPT based on fed information related to dataset and performed K-means clustering



Results of the Agglomerative clustering for the same dataset are depicted in Figure 5 (for the last cluster), and the corresponding summary is almost the same as for the K-means. If we compare ChatGPT results of K-means and Agglomerative clustering and the other two clustering results, we can find some similarities but also nuanced differences.

Figure 5: Interpretation results of ChatGPT for the fifth cluster (index position 4) of the third dataset clustered with the Agglomerative clustering algorithm



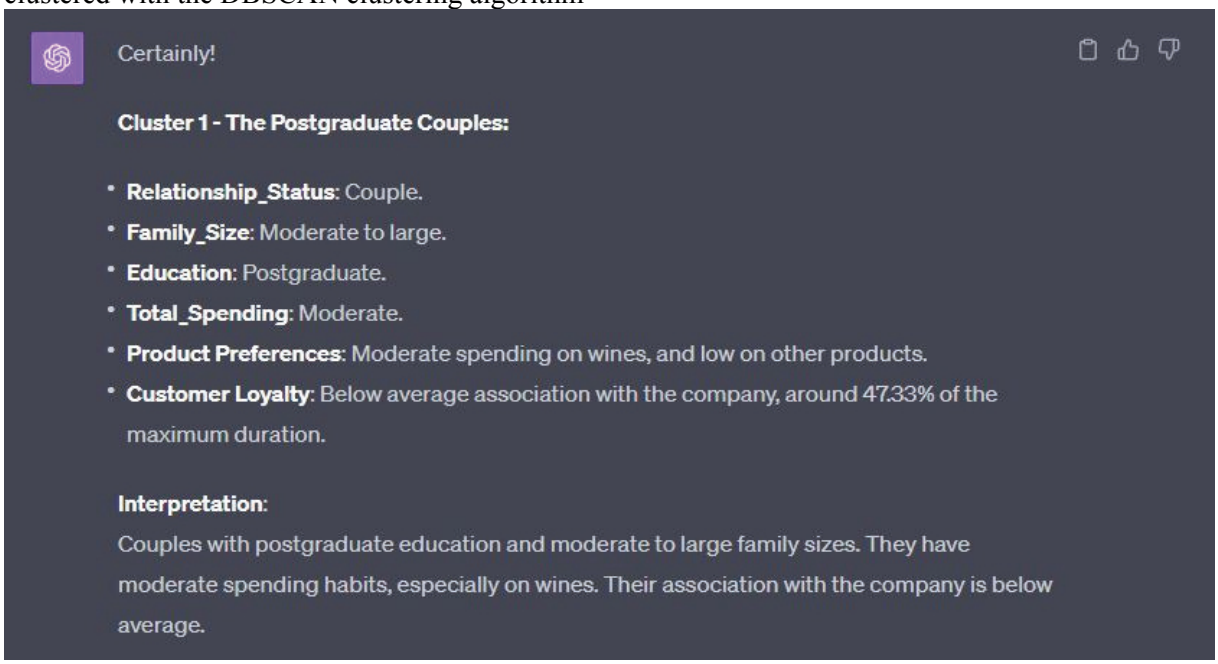
Results for the DBSCAN that are presented to ChatGPT are shown in Table 3.

Table 3 DBSCAN final results presented to ChatGPT

R_S	F_S	Education	T_S	C_f_D	Wines	Fish	Fruits	Meat	Sweets	Cluster
0.54	0.2929	0.3570	0.5765	0.5165	0.3590	0.4874	0.3625	0.3251	0.3278	-1
0.00	0.2270	0.8668	0.2487	0.4791	0.2129	0.1495	0.1384	0.1015	0.1016	0
1.00	0.4982	0.8643	0.2445	0.4733	0.2179	0.1362	0.1278	0.0963	0.0994	1
1.00	0.4609	0.0000	0.1259	0.5429	0.0753	0.1475	0.1028	0.0507	0.0960	2
0.00	0.2326	0.0000	0.1374	0.5021	0.1024	0.0937	0.1058	0.0535	0.0929	3

Interpretations for the larger DBSCAN cluster are depicted in Figure 6, where we can detect differences in interpretation compared to interpretation results of obtained clusters with K-means and Agglomerative clustering.

Figure 6: Interpretation results of ChatGPT for the third cluster (index position 1) of the third dataset clustered with the DBSCAN clustering algorithm



All obtained ChatGPT interpretation results on the three datasets for all four clustering algorithms are presented to marketing experts which than evaluate them according to the evaluation principle described in this section. Results of findings are summarized in the Table 4, 5, and 6.

Table 4 Evaluation results for ChatGPT interpretability of obtained clusters on BMW-Browsers dataset

BMW – Browsers dataset	ChatGPT Interpretation	Difference (base is K-means)
K - means	satisfactory	-
Agglomerative clustering	satisfactory	Slight difference
Fuzzy C - means	satisfactory	Medium difference
DBSCAN	satisfactory	Significant difference

Table 5 Evaluation results for ChatGPT interpretability of obtained clusters on German Credit dataset

German Credit - bank risk	ChatGPT Interpretation	Difference (base is K-means)
K - means	good	-
Agglomerative clustering	good	Slight difference
Fuzzy C - means	good	Slight difference
DBSCAN	satisfactory	Medium difference

Table 6 Evaluation results for ChatGPT interpretability of obtained clusters on Customer personality analysis

Customer personality analysis	ChatGPT Interpretation	Difference (base is K-means)
K - means	very good	-
Agglomerative clustering	very good	Slight difference
Fuzzy C - means	very good	Slight difference
DBSCAN	very good	Slight difference

The final decisions of the experts are formed by employing a majority voting policy. The results presented in Tables 4 – 6 show that the lowest interpretation level assigned to the ChatGPT interpretation is satisfactory. All experts agreed that with limited information fed to ChatGPT, we cannot expect spectacular results; in other words, they all agreed that with such an amount of data, LLM provides at least satisfactory interpretation. Analyzing the results, we can conclude that the interpretation quality of ChatGPT almost does not depend on the clustering algorithm.
Considering obtained interpretations of the used clustering algorithm, experts recommend employing more than one clustering algorithm to get fine nuances that enrich interpretation space. Further, results clearly show that ChatGPT interpretation ability in customer segmentation tasks based on clustering strongly depends on the a priori level of information related to the dataset, its feature description, and the knowledge of the underlying business process that generates data. With these observations on the experimental results, we have answered both research questions in this study.

The main limitations of this preliminary research study are threefold:

1. There needs to be more experts that evaluate LLM interpretation ability.
2. Evaluation procedures require a stricter, standardized approach with objective, clear, and meaningful metrics.
3. A more diverse, larger real-world dataset must be exploited for more scientifically valuable results.

To be precise, there is one more limitation, and it is related to the way of presenting information to the LLM. We already researched that direction and have a candidate to resolve that problem. Namely, with the plugin Notable, it is possible to load raw, engineered, and pre-processed datasets on this cloud analytic platform. The plugin enables direct connection of ChatGPT with the platform. In that way, one can lead ChatGPT through prompt instructions in detail clustering analysis. The first experiment on the third dataset shows positive results, which gives good promises of the real potential of the ChatGPT as the interpretation assistant in the task of customer segmentation using the clustering approach. This promising direction needs to be verified with appropriate scientific study.

5 Conclusion

This preliminary study explored the possibility of using Large Language Models as interpretation assistants to marketing experts for better understanding the final cluster association in customer segmentation tasks. ChatGPT Plus was picked as the representative of the LLMs and tested on three datasets that differ in size, number, types of features, underlying business problem, and, most importantly, level of knowledge related to the description of data. Data clustering was performed with K-means, Agglomerative clustering, DBSCAN, and fuzzy C-means to investigate the impact of the employed clustering algorithm on interpretation results. Interpretation outcomes were presented to three marketing experts, and with a majority voting principle on their evaluations, final results were obtained. Based on the final experts' evaluations, the interpretation quality of the ChatGPT in this problem almost does not depend on the engaged clustering method. It is significantly related to the level of additional knowledge of data fed to ChatGPT. That means if more relevant and detailed dataset information is provided to ChatGPT, it will produce interpretations of higher quality. In future work, the main shortcomings of this study need to be resolved by employing more experts in the evaluation process, providing more extensive and more versatile real datasets on customer segmentation problems, and developing objective, standardized evaluation procedures with clear and meaningful metrics. Also, it would be a good idea to investigate the prompt-based directed lead of ChatGPT in this task in combination with adequate plugins. Overall, this preliminary study showed positive results of the potential of the LLMs as the interpretation assistants in cluster analysis for the customer segmentation problem.

References

Abdulhafedh, Azad. 2021. “Incorporating K-Means, Hierarchical Clustering and PCA in Customer Segmentation.” *Journal of City and Development* 3 (1): 12–30. https://doi.org/10.12691/jcd-3-1-3.

Adadi, Amina, and Mohammed Berrada. 2018. “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI).” *IEEE Access* 6: 52138–60. https://doi.org/10.1109/ACCESS.2018.2870052.

Alves Gomes, Miguel, and Tobias Meisen. 2023. *A Review on Customer Segmentation Methods for Personalized Customer Targeting in E-Commerce Use Cases*. *Information Systems and E-Business Management*. Springer Berlin Heidelberg. https://doi.org/10.1007/s10257-023-00640-4.

Basak, Jayanta, and Raghu Krishnapuram. 2005. “Interpretable Hierarchical Clustering by Constructing an Unsupervised Decision Tree.” *IEEE Transactions on Knowledge and Data Engineering* 17 (1): 121–32. https://doi.org/10.1109/TKDE.2005.11.

Broniatowski, David A. 2021. “NISTIR 8367 Psychological Foundations of Explainability and Interpretability in Artificial Intelligence.” *Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8367*, 56.

Dasgupta, Sanjoy, Nave Frost, Michal Moshkovitz, and Cyrus Rashtchian. 2020. “Explainable k -Means and k -Medians Clustering.” In *ICML 2020 In XXAI Workshop*.

Dave, Tirth, Sai Anirudh Athaluri, and Satyam Singh. 2023. “ChatGPT in Medicine: An Overview of Its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations.” *Frontiers in Artificial Intelligence* 6. https://doi.org/10.3389/frai.2023.1169595.

Dawane, Vinit, Prajakta Waghodekar, and Jayshri Pagare. 2021. “RFM Analysis Using K-Means Clustering to Improve Revenue and Customer Retention.” *SSRN Electronic Journal*, no. Icsmdi. https://doi.org/10.2139/ssrn.3852887.

Emmendorfer, Leonardo Ramos. 2019. “An Empirical Evaluation of Two Novel Linkage Criteria for Hierarchical Agglomerative Clustering.” *Proceedings - 2019 Brazilian Conference on Intelligent Systems, BRACIS 2019*, no. 3: 622–26. https://doi.org/10.1109/BRACIS.2019.00114.

Esfandiari, Hossein, Vahab Mirrokni, and Shyam Narayanan. 2022. “Almost Tight Approximation

Algorithms for Explainable Clustering." *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms* 2022-Janua: 2641–63. https://doi.org/10.1137/1.9781611977073.103.

Fraiman, Ricardo, Badih Ghattas, and Marcela Svarc. 2013. "Interpretable Clustering Using Unsupervised Binary Trees." *Advances in Data Analysis and Classification* 7 (2): 125–45. https://doi.org/10.1007/s11634-013-0129-3.

Guoan ZHU, and Xue GAO. 2019. "The Digital Sales Transformation Featured ByPrecise Retail Marketing Strategy." *Expert Journal of Marketing*.

Hung, Phan Duy, Nguyen Thi Thuy Lien, and Nguyen Duc Ngoc. 2019. "Customer Segmentation Using Hierarchical Agglomerative Clustering." *ACM International Conference Proceeding Series* Part F1483: 33–37. https://doi.org/10.1145/3322645.3322677.

IBM. 2010. "BMW Browers Dataset." https://developer.ibm.com/articles/os-weka2/.

Imran, Muhammad, and Norah Almusharraf. 2023. "Analyzing the Role of ChatGPT as a Writing Assistant at Higher Education Level: A Systematic Review of the Literature." *Contemporary Educational Technology* 15 (4): ep464. https://doi.org/10.30935/cedtech/13605.

KAGGLE. n.d. "German Credit Risk." https://www.kaggle.com/datasets/uciml/german-credit.

Kamran, Khan, Saif ur Rehman, Sohail Asghar, Simon Fong, and Sababady Sarasvady. 2014. "DBSCAN : Past , Present and Future." In *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014) IEEE*, 232–38.

Kansal, Tushar, Suraj Bahuguna, Vishal Singh, and Tanupriya Choudhury. 2018. "Customer Segmentation Using K-Means Clustering." *Proceedings of the International Conference on Computational Techniques, Electronics and Mechanical Systems, CTEMS 2018*, 135–39. https://doi.org/10.1109/CTEMS.2018.8769171.

Kaur, Er Rupampreet, and Er Kiranbir Kiranbir. 2017. "Data Mining on Customer Segmentation: A Review." *International Journal of Advanced Research in Computer Science* 8 (5): 857–61.

Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. 2014. "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning." *Proceedings of 2014 Science and Information Conference, SAI 2014*, 372–78. https://doi.org/10.1109/SAI.2014.6918213.

Li, Junyi, Tianyi Tang, Wayne Xin Zhao, and Ji Rong Wen. 2021. "Pretrained Language Models for Text Generation: A Survey." *IJCAI International Joint Conference on Artificial Intelligence*, 4492–99. https://doi.org/10.24963/ijcai.2021/612.

Makarychev, Konstantin, and Liren Shan. 2022. "Explainable K-Means: Don't Be Greedy, Plant Bigger Trees!" *Proceedings of the Annual ACM Symposium on Theory of Computing*, 1629–42. https://doi.org/10.1145/3519935.3520056.

Moradi Dakhel, Arghavan, Vahid Majdinasab, Amin Nikanjam, Foutse Khomh, Michel C. Desmarais, and Zhen Ming (Jack) Jiang. 2023. "GitHub Copilot AI Pair Programmer: Asset or Liability?" *Journal of Systems and Software* 203. https://doi.org/10.1016/j.jss.2023.111734.

Mucha, Hans-Joachim, Hans-Georg Bartel, and Carlos Morales-Merino. 2012. "Visualisation of Cluster Analysis Results." In *Studies in Classification, Data Analysis, and Knowledge Organization - Classification and Data Mining*, edited by Antonio Giusti, Gunter Ritter, and Maurizio Vichi, 261–70. Springer Heidelberg. http://www.springerlink.com/index/10.1007/978-3-642-76307-6.

Munusamy, Sivaguru, and Punniyamoorthy Murugesan. 2020. "Modified Dynamic Fuzzy C-Means Clustering Algorithm – Application in Dynamic Customer Segmentation." *Applied Intelligence* 50 (6): 1922–42. https://doi.org/10.1007/s10489-019-01626-x.

Mutoffar, Muhamad Malik, Sri Kuswayati, Fetty Tri Anggraeny, and Tarsinah Sumarni. 2023. "Exploring the Potential of ChatGPT in Improving Online Marketing and Promotion of MSMEs." *Jurnal Minfo Polgan* 12 (1): 480–89. https://doi.org/10.33395/jmp.v12i1.12440.

Nasiopoulos, Dimitrios K., Damianos P. Sakas, D. S. Vlachos, and Amanda Mavrogianni. 2015. "Modeling of Market Segmentation for New IT Product Development." *AIP Conference Proceedings* 1644 (February): 51–59. https://doi.org/10.1063/1.4907817.

"OpenAI." 2023. 2023. https://openai.com/chatgpt.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong

Zhang, et al. 2022. “Training Language Models to Follow Instructions with Human Feedback.” *Advances in Neural Information Processing Systems* 35 (NeurIPS).

Parsons, Lance, Lance Parsons, Ehtesham Haque, Ehtesham Haque, Huan Liu, and Huan Liu. 2004. “Subspace Clustering for High Dimensional Data.” *ACM SIGKDD Explorations Newsletter* 6 (1): 90–105. https://doi.org/10.1145/1007730.1007731.

Pradipta, I Made Dhanan, Agus Eka, Anwar Wahyudi, and Sri Aryani. 2018. “Fuzzy C-Means Clustering for Customer Segmentation.” *International Journal of Engineering and Emerging Technology* 3 (1): 18–22. https://ojs.unud.ac.id/index.php/ijeet/article/download/41251/25103.

Romero-Hernandez, Dr. Omar. n.d. “Customer Personality Analysis.” https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis.

Saisubramanian, Sandhya, Sainyam Galhotra, and Shlomo Zilberstein. 2020. “Balancing the Tradeoff between Clustering Value and Interpretability.” *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 351–57. https://doi.org/10.1145/3375627.3375843.

Salvagno, Michele, Fabio Silvio Taccone, and Alberto Giovanni Gerli. 2023. “Can Artificial Intelligence Help for Scientific Writing?” *Critical Care* 27 (1): 1–5. https://doi.org/10.1186/s13054-023-04380-2.

Sembiring Brahmana, Rahma Wati, Fahd Agodzo Mohammed, and Kankamol Chairuang. 2020. “Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods.” *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi* 11 (1): 32. https://doi.org/10.24843/lkjiti.2020.v11.i01.p04.

Singh, Ritika. 2021. “Exploratory Data Analysis and Customer Segmentation for Smartphones Analysis and Simulation of COVID-19 View Project,” 1–5. https://www.researchgate.net/publication/351351474.

Steinley, Douglas, and Michael J. Brusco. 2007. “Initializing K-Means Batch Clustering: A Critical Evaluation of Several Techniques.” *Journal of Classification* 24 (1): 99–121. https://doi.org/10.1007/s00357-007-0003-0.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” *Advances in Neural Information Processing Systems* 2017-Decem (Nips): 5999–6009.

Wang, Xuejin, Chongdong Zhou, Yijing Yang, Yueyong Yang, Tianyao Ji, Jifei Wang, Jingrui Chen, and Ying Zheng. 2020. “Electricity Market Customer Segmentation Based on DBSCAN and K-Means : - A Case on Yunnan Electricity Market.” *2020 Asia Energy and Electrical Engineering Symposium, AEEES 2020*, 869–74. https://doi.org/10.1109/AEEES48850.2020.9121413.