Empowering Persona Creation in Small Organizations

Evaluating ChatGPT 4o for Clustering and Analysis using PersonaCraft

Jefferson Lewis Velasco
Graphic Expression Department
Federal University of Santa Catarina
Florianópolis, Brazil[1]
e-mail: jeffvelasco.crm@gmail.com

Gustavo Modena
Knowledge Engineering Department
Federal University of Santa Catarina
Florianópolis, Brazil
e-mail: gustavoomodena2@gmail.com

Melise Peruchini
Network and IT department
Federal University of Santa Catarina
Florianópolis, Brazil
e-mail: meliseperuchini@gmail.com

Julio Monteiro Teixeira
Graphic Expression Department
Federal University of Santa Catarina
Florianópolis, Brazil
e-mail: juliomontex@gmail.com

Abstract— **This paper explores the use of ChatGPT 4o to assist small organizations in creating data-driven personas by applying an adapted version of the PersonaCraft methodology. Using a retail demographic dataset from Kaggle, the approach demonstrates how the LLM aids in data cleaning, clustering, and statistical testing, culminating in five meaningful customer segments. The k-prototypes algorithm was chosen based on PersonaCraft, and the Kruskal-Wallis test confirmed that numeric variables (especially purchase frequency and age) most effectively differentiate clusters. By producing insights with minimal user effort, ChatGPT 4o underscores the viability of employing LLM-based tools for persona creation and other advanced analytical tasks in resource-constrained contexts.**

Keywords—Generative AI, Artificial Intelligence, Personas, Customer Segmentation.

I. INTRODUCTION

Small organizations often face challenges in creating personas due to the complexity of clustering and other advanced statistical methods needed for accurate segmentation [1]. While data-driven persona generation has traditionally required specialized expertise [2], the rising capabilities of Large Language Models (LLMs) offer a promising alternative [3].

This paper evaluates how Chat Generative Pre-trained Transformer (ChatGPT) 4o can operationalize the PersonaCraft methodology for persona creation by automating key steps—such as clustering and statistical testing—through a natural language interface. This adaptation shows how prompt engineering enables non-experts to segment customers effectively, enabling users without data science expertise to produce robust, data-driven personas. This bridges a key accessibility gap in small organizations.

We apply the PersonaCraft [4] methodology to build personas using a demographic dataset from a bike and bike accessories retail business in Australia, publicly available on Kaggle [5]. GPT executes the Stages and Steps defined by the method, enabling users with minimal background in statistics or coding to engage in persona creation, specifically segmentation tasks.

By systematically incorporating prompts aligned with the PersonaCraft framework, we assess GPT's capability in assisting small organizations in clustering tasks and broadening access to advanced analytical techniques. It is important to note that this work does not aim to evaluate the PersonaCraft methodology itself. Rather, the goal is to examine GPT's capabilities as a data analysis assistant.

This paper is structured as follows: Section II presents the theoretical background. Section III details the employed methodology, including adaptations made to PersonaCraft. Section IV provides results and analysis. Section V offers conclusions and future work. This study addresses the following research question: **How effectively can ChatGPT 4o assist non-expert users in executing the core stages of PersonaCraft for the creation of data-driven personas?**

II. THEORY

Data-driven personas, derived from statistical analyses of real data, ensure credibility by reflecting actual behavioral and demographic patterns rather than assumptions [6]. Their development through clustering algorithms has shown promising applications in the literature, helping identify patterns and create representative audience segments [7].

Recent studies suggest that LLMs can enable non-experts to perform data analysis and code generation using natural language prompts [3], including clustering and concept identification [8]. Some initiatives use LLMs to build

[1] Although all authors are affiliated with Brazilian institutions, the dataset used was chosen for its accessibility, completeness, and relevance to segmentation tasks—not due to a connection with the Australian retail market.

personas—either fully automatically or in combination with human refinement. Combining human input with LLMs often improves results [9].

Despite these advances, challenges remain regarding representativeness, bias, and stereotypes in AI-generated personas [10]. LLMs may reinforce existing patterns or overlook outliers that don't align with dominant clusters. These risks underscore the need for bias mitigation strategies—such as diverse datasets, interpretability checks, or human review. In this study, bias was partially addressed by retaining numeric granularity and reporting statistical significance, though more safeguards are needed for socially sensitive contexts.

III. Methodology

A publicly available dataset was sourced from Kaggle, containing 3,908 rows of simulated customer demographic data modeled after an Australian retail business [5]. Originally intended for educational use, it was selected for its simplicity and similarity to datasets typically available in small organizations. While suitable for this initial application, future studies should explore datasets with richer behavioral or psychographic features to further test GPT's adaptability.

To fit the PersonaCraft framework, minor adjustments were made to accommodate the dataset's simplicity, without altering the methodology itself. For example, "questions" in the original method were mapped to "variables" in our data.

Prompt adaptations were also necessary to better align with GPT's functionality. Compared to the original PersonaCraft prompts, the revised versions included clearer instructions about formatting and handling variables. Numericac columns were kept in continuous form and described using centrality measures (mean, median, and mode).

The study was conducted using the ChatGPT 4o interface via OpenAI's web platform (chat.openai.com). Prompts were submitted iteratively by a non-expert user simulating the role of a typical small-organization user. Prompt structure followed PersonaCraft stages and typically included: (1) context and general objective, (2) task description (e.g., "cluster this dataset using k-prototypes"), and (3) output expectations (e.g., "summarize each cluster using centrality metrics").

Basic understanding of PersonaCraft and statistical concepts—such as clustering techniques and descriptive statistics—was required to guide prompt refinement effectively.

Given the study's focus on cluster generation and LLM support, certain PersonaCraft components were excluded. Specifically, Step 4 of Stage 4 was omitted for scope reasons, while all earlier stages related to segmentation were executed in full. Also, this study did not involve human participants and relied exclusively on anonymized, synthetic data. Therefore, no IRB approval was necessary. Care was taken to comply with privacy standards and research ethics.

IV. Results

In applying the PersonaCraft methodology to the chosen dataset, four major stages were conducted with GPT acting as an assistant. Throughout the process, the LLM guided the user in performing clustering, statistical analysis, and persona generation.

Stage 1: Preparing the Dataset involved removing irrelevant variables (e.g., names, addresses, and other identifying attributes). The retained variables included: gender, age, number of bike-related purchases in the past 3 years, job industry category, wealth segment, car ownership, customer tenure, and state of residence. This streamlined view highlighted the demographic and behavioral patterns relevant to persona creation.

Stage 2: Mapping Variable Types classified each variable according to PersonaCraft categories (e.g., Demographic, Purchasing Behavior, Assets & Ownership). Numeric data—including age, tenure, and purchase frequency—was kept in its original scale to allow accurate calculation of centrality measures (mean, median, mode) during analysis.

TABLE I. Relevant Variable Classification

Variable	*Description*	*Classification*	*PersonaCraft Type*
customer_id	Customer ID index	Numeric	Not applicable
gender	Customer gender (female or male)	Categoric	Demographic
age	Customer age in years	Numeric	Demographic
past_3_years_bike_related_purchases	Number of bike related purchases in the last 3 years	Numeric	Other
job_industry_category	Job Industry the customer is in	Categoric	Demographic
wealth_segment	Wealth segment to which the customer belongs	Categoric	Demographic
tenure	Tenure of the customer in months	Numeric	Other
Owns_Car	If customer owns a car	Categoric	Demographic
State	State of residence	Categoric	Demographic

Stage 3: Data Pre-Processing, involved refining prompt inputs for GPT, classifying the variables into coherent "groups", and creating updated headings.

TABLE II. Variable Groups and Headings

Variable Group	*Description*	*Variables*	*Headings*
Personal Identification	Used to uniquely identify or describe a person.	customer_id	Customer ID
Demographics	Attributes related to socioeconomic status, age,	AGE	Age

	and personal traits.	gender	Gender
		job_industry_category	Job Industry Category
		wealth_segment	Wealth Segment
Purchasing Behavior	Data on purchases or transaction-related behavior.	past_3_years_bike_related_purchases	Bike Purchases Last 3 Years
Customer Tenure	Duration of the customer's relationship.	tenure	Customer Tenure
Assets & Ownership	Indicators of asset ownership	Owns_Car	Car Ownership
Location Information	Geographic location of the customer.	State	State

Stage 4: Persona Generation involves four steps. In Step 1: Clustering, GPT was prompted to select the most suitable clustering method from PersonaCraft for our dataset. Given the mix of numeric and categorical variables, it recommended k-prototypes, a method well-suited for handling both types. Since the algorithm requires a predefined number of clusters [5], GPT applied the Elbow Method to determine the optimal value. As shown in Figure 1, the analysis indicated that k=5 provided best balance between cohesion and interpretability.

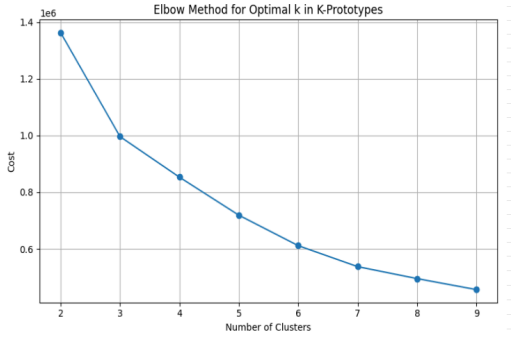

Figure 1. Elbow Method Results

The model proceeded with clustering using this parameter, generating an Excel file with cluster assignments for each record and a visualization of the resulting segments, as shown in Figure 2.

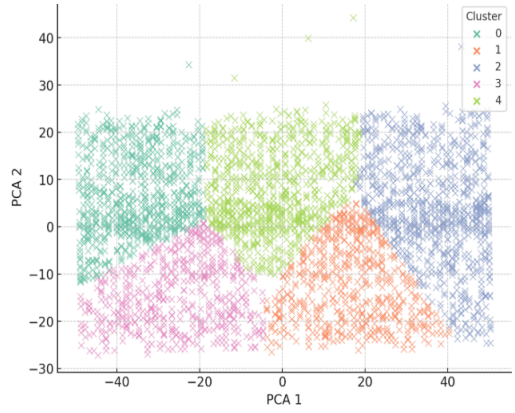

Figure 2. K-Prototypes Clustering PCA Projection

On Step 2: Statistical Analysis, GPT was prompted to calculate the Kruskal-Wallis test across the clusters to identify the variables most effective for differentiating personas. The test revealed that numeric variables (Bike Purchases in the Last 3 Years, Age, and Customer Tenure) had statistically significant variation, while Gender, State, Wealth Segment, Car Ownership, and Job Industry Category showed no significant differences across clusters.

TABLE III. KRUSKAL-WALLIS RESULTS

Variable	*H-statistic*	*p-value*	*Significant*
Bike Purchases Last 3 Years	34.767.060.696.875.200	0.0	Yes
Age	19.713.746.810.310.800	0.0	Yes
Customer Tenure	7.365.761.161.085.200	4,19E-143	Yes
Gender	56.158.533.280.999	22.973.201.404.749.700	No
State	5.162.671.440.601.860	27.100.991.196.740.200	No
Wealth Segment	33.190.520.564.661.200	5.059.193.012.085.890	No
Car Ownership	24.488.311.071.079.600	653.821.209.414.484	No
Job Industry Category	25.407.845.117.565.400	9.925.824.609.188.990	No

By not transforming numeric variables into discrete categories, GPT was able to generate and interpret measures of centrality within each cluster, providing insights into how age, purchase frequency, and tenure shape each persona. Cluster summaries are available in Table IV.

TABLE IV. CLUSTER DESCRIPTIONS

Variable	*Values*	*C1*	*C2*	*C3*	*C4*	*C5*
Gender	Female	468	328	484	295	462
	Male	373	317	471	281	429
Job Industry Category	Manufacturing	173	144	185	107	187
	Financial Services	162	131	179	122	172
	n/a	142	102	165	96	150
	Health	122	86	161	83	144
	Retail	86	66	85	52	69
	Property	46	35	66	46	73
	Entertainment	35	21	30	21	29
	IT	29	26	39	27	29
	Agriculture	25	25	29	10	24
	Telecommunications	21	9	16	12	14
Wealth Segment	Mass Customer	401	328	498	275	449
	High Net Worth	229	162	227	157	220

	Affluent Customer	211	155	230	144	222
Car Ownership	Yes	432	338	472	280	449
	No	409	307	483	296	442
State	New South Wales	456	330	530	316	457
	Victoria	214	155	229	147	253
	Queensland	171	160	196	113	181
Age	Mean	52,11	32,95	47,86	31,25	52,55
	Median	51	32	47	30	51
	Mode	45	27	44	28	44
Bike Purchases Last 3 Years	Mean	14,06	65,38	85,90	23,45	47,22
	Median	14	66	87	25	47
	Mode	2	68	98	27	53
Customer Tenure	Mean	12,34	7,45	11,41	6,82	13,05
	Median	12	6	12	5	13
	Mode	11	1	12	2	18

All underlying data and resultant analyses from this research are available for review upon request.

V. Conclusions and Future Work

Overall, the results demonstrate that ChatGPT 4o can be of significant help in data exploration and clustering tasks, even for users with limited analytic expertise. By adapting PersonaCraft to a simpler retail demographic dataset and employing the model's prompt-driven guidance, viable segments emerged that underline the importance of numeric variables (particularly purchase frequency and age) when distinguishing among diverse customer groups.

These findings suggest that the LLM can be a reliant assistant for small organizations lacking deep data analysis expertise. By guiding the user through data preparation, variable mapping, and clustering steps, the model demonstrated its capacity to bridge knowledge gaps.

Although this research was limited by the simplicity of the dataset and by the absence of direct human validation, the results demonstrate the method's feasibility. Future studies should validate this approach using larger and more complex datasets, as well as through participatory evaluation—where human experts and stakeholders assess the relevance, coherence, and representativeness of the generated personas.

Adopting PersonaCraft's prompt-driven approach as a framework to segment data through LLM allowed for a more accessible and intuitive workflow, enabling persona creation without extensive statistical background or programming experience. These findings point toward broader opportunities for LLM-based tools to support diverse, data-centric tasks within smaller organizations, enabling tailored marketing insights.

References

[1] E. L. Melnic, "How to strengthen Customer Loyalty, using Customer Segmentation?", Bulletin of the Transilvania University of Brasov. Series V: Economic Sciences, pp. 51–60, dez. 2016. [retrieved: March, 2025].

[2] J. Brickey, S. Walczak, and T. Burgess, "Comparing Semi-Automated Clustering Methods for Persona Development", IEEE Transactions on Software Engineering, vol. 38, nº 3, pp. 537–546, May 2012, doi: 10.1109/TSE.2011.60. [retrieved: March, 2025].

[3] J. A. Jansen, A. Manukyan, N. A. Khoury, and A. Akalin, "Leveraging large language models for data analysis automation," PLOS ONE, vol. 20, no. 2, p. e0317084, Feb. 2025, doi: 10.1371/journal.pone.0317084. [retrieved: March, 2025].

[4] S.-G. Jung, J. Salminen, K. K. Aldous, and B. J. Jansen, "PersonaCraft: Leveraging language models for data-driven persona development", International Journal of Human-Computer Studies, vol. 197, p. 103445, mar. 2025, doi: 10.1016/j.ijhcs.2025.103445. [retrieved: March, 2025].

[5] E. Harish, "KPMG Customer Demography Cleaned Dataset", Kaggle. Accessed on: March 19, 2025. [Online]. Available at: https://www.kaggle.com/datasets/harishedison/kpmg-customer-demography-cleaned-dataset. [retrieved: March, 2025].

[6] J. (Jen) McGinn and N. Kotamraju, "Data-driven persona development," in Proceedings of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08, Florence, Italy: ACM Press, 2008, pp. 1521–1524. doi: 10.1145/1357054.1357292. [retrieved: March, 2025].

[7] E. Ditton, A. Swinbourne, and T. Myers, "Selecting a clustering algorithm: A semi-automated hyperparameter tuning framework for effective persona development," Array, vol. 14, p. 100186, Jul. 2022, doi: 10.1016/j.array.2022.100186. [retrieved: March, 2025].

[8] F. Lanfermann, T. Rios, and S. Menzel, "Large Language Model-assisted Clustering and Concept Identification of Engineering Design Data," in 2024 IEEE Conference on Artificial Intelligence (CAI), Singapore, Singapore: IEEE, Jun. 2024, pp. 788–795. doi: 10.1109/CAI59869.2024.00150. [retrieved: March, 2025].

[9] N. Arora, I. Chakraborty, and Y. Nishimura, "AI–Human Hybrids for Marketing Research: Leveraging Large Language Models (LLMs) as Collaborators," Journal of Marketing, vol. 89, no. 2, pp. 43–70, Mar. 2025, doi: 10.1177/00222429241276529. [retrieved: March, 2025].

[10] T. Goel, O. Shaer, C. Delcourt, Q. Gu, and A. Cooper, "Preparing Future Designers for Human-AI Collaboration in Persona Creation," in Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work, Oldenburg Germany: ACM, Jun. 2023, pp. 1–14. doi: 10.1145/3596671.3598574. [retrieved: March, 2025].