PRAGUE UNIVERSITY OF ECONOMICS AND BUSINESS

Acta Informatica Pragensia
2023, Volume 12, Issue 2, pp. 357–378
https://doi.org/10.18267/j.aip.220

Article Open Access

Segmenting Customers with Data Analytics Tools: Understanding and Engaging Target Audiences

Tomáš Pitka, Jozef Bucko

Department of Applied Mathematics and Business Informatics, Faculty of Economics, Technical University of Kosice, Kosice, Slovakia

Corresponding author: Jozef Bucko (jozef.bucko@tuke.sk)

Abstract

This paper presents a decision support system for identifying customer typology using cluster analysis to segment relevant customers. The approach is demonstrated using data from a company selling nutritional supplements, consisting of approximately 130,000 records from six Central European countries. The analysis results in distinct groups of customers, which are proposed for more effective management of customer relationships. The findings have implications for retailers, helping them focus on the most profitable customer segments to increase sales and profits and build lasting relationships. Furthermore, cluster analysis proves to be an appropriate statistical method for classification and provides valuable insights into patterns and trends in the analysed data. Overall, this paper contributes to development and comparison of methods for customer segmentation and demonstrates their potential for improving economic efficiency and building long-term customer relationships.

Keywords

Cluster analysis; Consumer behaviour; Two-step analysis; Mixed data; Customer segmentation.

Citation: Pitka, T., & Bucko, J. (2023). Segmenting Customers with Data Analytics Tools: Understanding and Engaging Target Audiences. *Acta Informatica Pragensia*, 12(2), 357–378. https://doi.org/10.18267/j.aip.220

Academic Editor: Zdenek Smutny, Prague University of Economics and Business, Czech Republic

1 Introduction

Customer segmentation is one of the most critical aspects of customer relationship management (Chocarro et al., 2015). Studying different types of customers is essential to the success of retail enterprises. It means clustering customers reasonably and effectively to improve the economic efficiency of enterprises (Alves Gomes & Meisen, 2023). Direct marketers use segmentation based on cluster analysis to target specific groups of customers. Cluster analysis is a data mining tool that allows organisations to identify distinct groups of customers, sales transactions or other types of behaviours and objects (Güçdemir & Selim, 2015). Classification is a statistical method for dividing a large group into smaller subgroups based on the characteristics of their members. This technique can help organisations to gain a deeper understanding of their data and make informed decisions based on that knowledge. In addition, this approach can provide valuable insights into patterns and trends in the analysed data. Customer relationship management (CRM) based on data mining is a strategy and process that uses customer knowledge to improve the business (Sokol & Holý, 2021). Choosing the correct clustering algorithm is essential, especially when the data are mixed.

Clustering procedures assign each subject to a single class, assuming that issues are indistinguishable within that class. The classes may be organised into a hierarchy, with some types divided into subcategories (Fraley et al., 1998). Statistical software programs offer a variety of algorithms that can be used to analyse data in this manner. Data analytics tools possess both advantages and disadvantages when it comes to comprehending customer behaviour (Rushi & Pradha, 2022). Data analytics tools offer valuable insights into customer behaviour, which can assist organisations in making informed business decisions. However, organisations must confront obstacles such as data quality, privacy, and technical expertise to exploit these tools entirely. Therefore, organisations must prioritise addressing these challenges to maximise the benefits of data analytics tools (Anshari et al., 2019).

This paper aims to identify customer groups with common characteristics and their subsequent integration into customer groups. These findings help retailers focus their business on the most profitable segments of customers, increase sales and profits, and build lasting relationships with customers. The article aims to educate readers on the benefits, best practices and limitations of using these tools to understand customer behaviour better.

2 State of the Art

In recent years, competition among companies has increased significantly, making it challenging for firms to remain successful in their field. A company's profitability can be enhanced by strategically implementing a customer segmentation model. This approach underscores the significance of prioritising the retention of current customers over the pursuit of new ones. By identifying and catering to the unique needs and preferences of distinct customer segments, an organisation can bolster its bottom line and foster lasting customer loyalty, solidifying its position in the market and sustaining long-term growth (Butt et al., 2012).

Businesses have increasingly leveraged data analysis software to gain insights into customer behaviour. Customer segmentation divides a customer base into groups based on shared characteristics such as psychographics, demographics, or purchase patterns. Data exploration tools can help businesses effectively segment their customers and uncover actionable patterns. This information enables companies to tailor their marketing campaigns to specific customer groups and improve the overall customer experience (Chen et al., 2012; Wan et al., 2022).

Hicham et al. (2022) and Grandhi et al. (2021) emphasise the significance of understanding customer behaviour in today's dynamic business landscape. Traditional cluster analysis approaches were previously employed on retail databases. However, these traditional methods face challenges due to the

diverse nature of customers and the vastness of modern databases. The rise of data sources, especially social networks, has amplified the importance of data-driven marketing, a key component in customer segmentation. For instance, an integrated strategy using the Apriori algorithm and the CRM method with associated mining has been utilised for this purpose. Another method highlighted is the LRFM (Length, Recency, Frequency, and Monetary) model and its extended version, the LRFM-Average Item (AI) model (Grandhi et al., 2021; Hicham et al., 2022). Pradana and Ha (2021) highlight the critical importance of customer segmentation in marketing. The study aims to address the challenges of large datasets in traditional segmentation techniques. As described in the article, the goal of clustering is to optimise the similarity within a cluster while maximising the dissimilarity between clusters. The authors employ the K-means clustering algorithm as the foundational method for segmentation and conclude that targeting customers with high income and high spending scores using the K-means clustering method can lead to more effective and efficient marketing strategies, reducing the wastage of resources (Pradana & Ha, 2021). Singh and Mittal (2021) aim to address the challenge of traditional clustering algorithms unsuitable for clustering Boolean and categorical attributes. The work focuses on grouping items based on their purchase dependencies. For instance, grocery items are frequently purchased, and electronic items are often bought together. By understanding these dependencies, items can be grouped, and attractive offers can be made to customers, thereby increasing the organisation's overall profit. The authors present a method to analyse customer behaviour by mining historical transactional databases. The paper highlights the limitation of the K-means algorithm, which works only on numeric data. In real-world datasets, objects often have categorical attributes or a mix of numeric and categorical values. According to the authors, association rules relate to frequently purchased data items. With enough association rules, customer purchases can be predicted. Decision-makers can use this knowledge to predict future purchases and make strategic decisions to improve customer relationships (Singh & Mittal, 2021).

When choosing data-driven tools for customer segmentation, it is essential to consider the following factors:

- **data integration capabilities:** Ensure the tool can seamlessly integrate with existing data sources, such as CRM systems, e-commerce platforms, and other databases (Yan et al., 2018);
- **scalability:** The tool should handle increasing data as the business grows (Feng et al., 2020);
- **cost and ROI:** Evaluate the cost of the tool against the potential return on investment. Consider both the immediate costs and long-term value (Coston et al., 2022);
- **analytical power:** The tool should offer advanced analytics capabilities, such as predictive modelling, to help forecast future trends and behaviours (Godinho et al., 2021);
- **customisation and personalisation:** The ability to customise segmentation criteria and generate personalised reports can provide more relevant insights into specific business needs (Godinho et al., 2021);
- **data visualisation features:** Tools that offer clear data visualisation can help better understand and interpret the segmentation results (Coston et al., 2022);
- **segmentation flexibility:** Look for tools to segment customers based on various criteria, including demographics, purchase behaviour, and engagement metrics (Yan et al., 2018).

After selecting the correct tool, the next step is to initiate the customer segmentation process. The ideal method for segmenting customers will vary based on business objectives and the data available to the company. Nonetheless, there are several typical approaches to customer segmentation, including:

- Demographic segmentation (Gajanova et al., 2019);
- Behavioural segmentation (Griva et al., 2022);
- Psychographic segmentation (Gajanova et al., 2019);
- Geographic segmentation (Griva et al., 2022).

Customer understanding is critical in online commerce, as it allows businesses to tailor their strategies and offerings to meet the needs and preferences of their target audience. This understanding has become increasingly important in recent years, as e-commerce has made it easier for customers to shop around and switch between brands. Mkpojiogu and Hashim (2016) discuss the significance of customer satisfaction concerning product quality, viability, and profitability. The study emphasises that understanding customers' requirements is essential to increasing loyalty and profitability. The research also delves into how well these requirements are met. It suggests that meeting requirements influences customers' satisfaction with a product (Mkpojiogu & Hashim, 2016). Alrumiah and Hadwan (2021) explore the benefits of using Big Data Analytics (BDA) in e-commerce for sellers and buyers. Researchers investigate the influence of deep data analysis on e-commerce and its potential advantages in today's digital age. The authors also look at the challenges e-commerce faces due to the large amount of data that needs complex analysis. The goal is to see how BDA can improve decision-making by analysing large datasets. The study also highlights how BDA can improve e-commerce systems, boosting seller profits and attracting more customers (Alrumiah & Hadwan, 2021).

Data mining helps organisations gain valuable insights into customers' behaviour by analysing their data. Several algorithms used in data analytics tools assist in understanding customer behaviour. The most common ones are:

- **Association rules:** represent a data mining methodology employed for detecting and examining patterns and associations among items within extensive datasets. In online customer segmentation, these rules facilitate the comprehension of shared attributes and distinct characteristics within customer cohorts' consumption patterns. This comprehension is paramount in informing and guiding judicious marketing decisions and strategic planning (Xiao & Paio, 2022).
- **Clustering:** is a data-driven technique for grouping online customers into discrete groups based on shared characteristics and behaviours (Hallishma, 2023). Through the application of algorithms such as K-Means and hierarchical clustering, organisations can discern and comprehend diverse customer segments, thereby enabling the customisation of marketing strategies, optimisation of product offerings, and the enhancement of customer relationship management (Kovács et al., 2021).
- **Decision trees**: are machine learning algorithms which enable the segmentation of online customers by considering various criteria, providing a methodical framework for analysing and forecasting customer behaviour, particularly within the online marketplace. By examining patterns in customer data, decision trees assist enterprises in pinpointing distinct market segments and devising tailored strategies accordingly (Vaca et al., 2020).

Machine learning tools use advanced algorithms to analyse large amounts of data and identify patterns for making predictions. In online customer categorisation, these tools are essential for understanding and grouping customers based on their online behaviour, likes, and purchase habits. Traditional marketing methods have been improved using data analysis, especially when segmenting online shoppers by their recent activity, how often they shop, and how much they spend (Yi & Liu, 2020).

As mentioned above, data analytics is crucial in today's digital world, allowing businesses to gain valuable insights from large data sets. This approach is beneficial in spotting unmet customer needs. By combining data from various sources, including social media, companies can better understand customer preferences, leading to better resource use and increased customer satisfaction (Alsayat, 2023).

3 Objectives and Methods

The fundamental objective of this paper is to attain a deeper understanding of decisions and conduct exhibited by consumers. The investigation concentrates on two discrete variables: consumers' gender and

their country of origin. The following section will examine differences between behaviours manifested by consumers from Slovakia and those from abroad. By exploring these disparities in consumer conduct, we aspire to derive valuable insights into preferences and motivations of distinct cohorts of consumers.

This paper is dedicated to understanding customer behaviour using cluster analysis and data analytics tools created for several potential objectives, including:

- **Exploring the importance of customer behaviour**: The article emphasises the importance of comprehending customer behaviour for businesses and the role of analytics tools in accomplishing this goal.
- **Identifying key customer behaviour metrics**: The article recognises crucial metrics and elucidates how data analysis can monitor and scrutinise these metrics.
- **Providing real-world examples**: The article provides tangible examples of companies that have effectively employed data analytics tools to comprehend customer behaviour and enhance business results.
- **Offering practical advice**: The article offers actionable recommendations on how businesses can exploit reporting and visualisation tools to obtain insights into customer behaviour and act upon those insights.
- **Discussing challenges and limitations**: The article additionally deliberates on obstacles and restrictions linked to utilising data mining to comprehend customer behaviour, such as data quality predicaments and the necessity for adept analysts.

3.1 Methodological procedure based on CRISP-DM

The CRISP-DM (Cross-industry standard process for data mining) procedure is a popular methodology used in data mining to understand customer behaviour (Huber et al., 2019; Saltz, 2021; Schröer et al., 2021). The process involves the following steps:

1) Business understanding

This phase involved gaining a comprehensive understanding of the business objectives and requirements of the project (Schröer et al., 2021). As the study uses actual company data, one of its primary objectives during this phase was establishing effective communication with the relevant company departments. The process involved regular online sessions to consult the data and brainstorm ideas about the direction and potential outcomes of the project. Through these discussions, the team established vital objectives and deliverables for the project, which can be found in Section 3. Unfortunately, researchers often disregard the current phase of the CRISP-DM framework and, at times, entirely omit it, which is deemed a significant oversight. A lack of comprehensive understanding of the research direction frequently leads to a need for explicitly articulated objectives and established pathways to achieve them (Saltz, 2021). Moreover, objectives are often identified without proper communication with the relevant company, resulting in inconsequential findings for the organisation or conclusions already drawn by the company in the past. As such, it is imperative to recognise the importance of this phase in research and respect it.

2) Data understanding

In the CRISP-DM methodology, an equally important stage involves understanding the input data. The initial stage is intrinsically linked to this phase, as understanding the input data typically delimits the potential research outcomes (Huber et al., 2019). In our case, this phase necessitated acclimatising ourselves with the provided input datasets, encompassing details relating to the customers, orders, and the company's products. Given that our input data were sourced from 10 distinct databases, a meticulous understanding of their interrelations and interactions was imperative for advancing the research. Analogous to the first phase, our collaborative meetings with representatives of diverse departments within the company featured a thorough discourse on the input databases. Comprehensive research into

customer behaviour was feasible only with this knowledge. Therefore, as with the first phase, neglect, or exclusion of this phase from the overall CRISP-DM procedure can have fatal consequences for the results and their credibility.

In understanding the input data, the most important thing was to understand the correctness of the linkage identifier of each database. In the initial phase, before holding joint discussions with company representatives, we tried to link the input databases based on identifiers such as: order_id, customer_id, customer_email, entity_id, product_id, and so on. These variables were represented in each input database, and we intended to create one central database containing all the essential and relevant variables suitable for our research. We succeeded by using the identifier entity_id in each input database. However, after our first meetings with the company's responsible persons, we realised that the identifier entity_id represents a different parameter in each input database. Thus, we could not use it as a general identifier for the interconnection of our databases. The results would be significantly skewed and inaccurate without adequately understanding the input data based on communication with the company. Therefore, based on the findings and the documentation provided by the company, in the next step, we interlinked the output databases based on multiple identifiers so as not to incorrectly assign attributes and variables in the central database that we worked with in the subsequent phases of CRISP-DM.

3) Data preparation

In this phase, we focused on basic input data operations such as cleaning, reduction, transformation, integration and creating new variables. After understanding the input data, we proceeded to analysis in the form of descriptive statistics, including the distribution of each variable, description of attributes, and the overall data structure. This exploration allowed us to understand the input data even more profoundly and facilitated the next phase of the CRISP-DM methodology, which focused on the modelling phase. By exploiting these fundamental insights, we optimised our data preparation efforts and ultimately improved the quality of our modelling results. Selected outputs of this phase can be seen in Section 4.2.

As mentioned, new variables such as day_period, day_name, quarter or delivery_method were created during the input data processing phase. The description of each variable is described in more detail in Section 4.2.

4) Modelling

This step involved applying various models to the data to identify the relationships and patterns in customer behaviour. It included tasks such as selecting the appropriate model and validating its performance.

In the context of examining customer online shopping behaviour based on cluster analysis, our modelling approach was based on the following steps:

a) Data preparation

In this step, selecting the relevant elements and variables to be used in the cluster analysis was necessary. The data must always be checked and formatted to ensure that they are suitable for cluster analysis. Since carried out thoroughly the data understanding and preparation phase, this step involved only a checking of the actions performed in the previous stages.

b) Selection of clustering technique

The proper clustering technique was selected by comparing the three selected algorithms: the TwoStep algorithm, K-means clustering and Kohonen clustering. The evaluation procedure of each approach and more details can be found in Section 4.3.

c) Model building

We used a clustering algorithm to group customers based on their behaviour. We selected the most accurate and suitable algorithm based on our data and objectives. As mentioned in the previous section, the actual creation of the model was based on the TwoStep clustering principle, the results of which can be seen in Section 5. While building the final model, we also tried to work carefully with the different input attributes. In the initial phase, we focused on building a generic model without our new specific variables, primarily created from the time stamp of the order. In the next step, we focused on creating a more detailed model that included all newly available attributes about customers and their demands. The results of the general model were very homogeneous, the presentation of which would not be of any use within the scope of this paper. On the other hand, when building a more specific model, we found more heterogeneous clusters of customers and their behavioural patterns. Thus, the final model includes the following variables: order_value, gender, delivery_method, country, day_period, day_name and quarter.

d) Model evaluation

In evaluating the results of the individual models, we worked primarily with the silhouette score validation parameter. An overview of the values of the models based on TwoStep clustering, along with a comparison of this parameter with the use of K-means and Kohonen clustering, can be seen in Section 4.3. Furthermore, among other criteria of the clustering results, we can consider the actual distribution of customers into different customer segments, which can be seen in sections 5.1, 5.2, 5.3 and 5.4.

e) Evaluation and deployment

We evaluated the model performance in this phase to determine whether it meets the objectives defined in the business understanding phase. The main point of discussion in this phase was the differences found in the general and specific views of customer behaviour. The simple creation of new parameters (as an example from the time stamp of an order) greatly enriched the results of the different models. Companies often have large amounts of data that they do not even know they have. The order timestamp is one of them. Every online order contains a timestamp in the form of the date and time when it was created. Manually assigning parameters such as day of the week, hour of the day, day of the year or similar is very laborious and time-consuming. However, we can extract several new attributes from this timestamp using data mining tools, such as IBM's SPSS Modeler to help us understand and predict customer behaviour over time.

To better understand customer behaviour, we can create new variables and include them in our models so that cluster analysis can provide new insights. These temporal attributes can help us achieve our objectives and make recommendations to the company for improving their marketing activities, as outlined in Section 6.

4 Research Location and Data Description

In this section, we will delve into a comprehensive depiction of the research site and provide an in-depth analysis of the distinctive features of the input data. We aim to give readers a detailed understanding of our study's contextual backdrop and elucidate the pertinent factors that have a bearing on our research outcomes.

4.1 Research location

Gymbeam s.r.o. is a private limited liability company based in Slovakia that specialises in producing and distributing sports nutrition supplements and accessories. Gymbeam offers a wide range of products, including protein powders, amino acids, vitamins and workout accessories. The company was founded

in 2010 and has since become one of the leading providers of sports nutrition in Central Europe. Gymbeam sells its products online and through a network of retail partners in various countries. The company strongly emphasises quality, and all its products are manufactured following strict international standards such as ISO 22000, HACCP and GMP. In addition to its brand of products, the company also offers a wide selection of products from other well-known sports nutrition brands, making it a one-stop shop for customers. Gymbeam's online store ships to over 30 countries worldwide. Furthermore, the company has local e-commerce websites tailored to customers in various regions, including Slovakia, the Czech Republic, Hungary and Poland. The company has won several awards for its products and services, including the "Best E-Shop of the Year" award in Slovakia in 2020.

4.2 Data description and preparation

In the Table 1, we can see a basic overview of the input data from Gymbeam s.r.o. The data examined comprise 107,228 records representing individual orders executed between 2018 and 2019.

Table 1. *Basic overview of input data.*

Field	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
order_value_eur	Continuous	2.000	299.940	36.038	30.601	2.193	--	107228
gender	Flag	--	--	--	--	--	2	107228
delivery_method	Flag	--	--	--	--	--	2	107228
day_name	Nominal	--	--	--	--	--	7	107228
day_period	Nominal	--	--	--	--	--	5	107228
quarter	Nominal	--	--	--	--	--	4	107228
country	Nominal	--	--	--	--	--	4	107228

The data processing began with an initial familiarisation with the data. Since the raw data comprised a relatively large dataset with dozens of parameters, understanding the data was one of the primary initial goals of the study. The company provided a set of databases relating to orders placed in the relevant period under review and the essential characteristics of individual customers. Therefore, we focused on understanding the variables and data structure in the initial processing phase. Selected outputs of this phase can be seen in the following tables and figures.

Table 2. *Distribution of individual input database parameters.*

Attribute	Frequency	Percentage	Valid percentage	Cumulative percentage
Female	36,603	34.1	34.1	34.1
Male	70,625	65.9	65.9	100
delivery_to_address	72,508	67.6	67.6	67.6
pick_up_delivery	34,720	32.4	32.4	100
Friday	12,112	11.3	11.3	11.3
Monday	19,607	18.3	18.3	29.6
Saturday	8,862	8.3	8.3	37.8
Sunday	13,473	12.6	12.6	50.4
Thursday	16,222	15.1	15.1	65.5
Tuesday	18,648	17.4	17.4	82.9
Wednesday	18,304	17.1	17.1	100
Afternoon	13,070	12.2	12.2	12.2
Evening	20,949	19.5	19.5	31.7
Noon	23,812	22.2	22.2	53.9

Attribute	Frequency	Percentage	Valid percentage	Cumulative percentage
Morning	19,197	17.9	17.9	71.8
Night	30,200	28.2	28.2	100
Q1	34,188	31.9	31.9	31.9
Q2	24,231	22.6	22.6	54.5
Q3	21,685	20.2	20.2	74.7
Q4	27,124	25.3	25.3	100
Czech Republic	4,368	4.1	4.1	4.1
Hungary	1,024	1	1	5
Romania	6,945	6.5	6.5	11.5
Slovakia	94,891	88.5	88.5	100

Table 3. *Crosstabulation of gender and country.*

		Czech Republic	Hungary	Romania	Slovakia	
Gender	**Female**	1,659	290	1,232	33,422	36,603
	Male	2,709	734	5,713	61,469	70,625
Total		4,368	1,024	6,945	94,891	107,228

Table 4. *Crosstabulation of quarter and country.*

		Czech Republic	Hungary	Romania	Slovakia	
Quarter	**Q1**	1,538	376	2,315	29,959	34,188
	Q2	935	201	1,323	21,772	24,231
	Q3	830	183	1,427	19,245	21,685
	Q4	1,065	264	1,880	23,915	27,124
Total		4,368	1,024	6,945	94,891	107,228

Table 5. *Crosstabulation of gender and delivery_method.*

		delivery_to_address	pick_up_delivery	
Gender	**Female**	25,102	11,501	36,603
	Male	47,406	23,219	70,625
Total		72,508	34,720	107,228

Table 6. *Descriptive statistics of order value.*

	N	Minimum	Maximum	Mean	Std. deviation
order_value_eur	107,228	2.00	299.94	36.0377155289122	30.6011917769
Valid N (listwise)	107,228				

Table 7. *Descriptives of order value.*

			Statistic	Std. error
order_value_eur	**Mean**		36.037715528912464	0.093451061268056
	95% confidence interval for mean	**Lower bound**	35.854552804986300	
		Upper bound	36.220878252838630	
	5% trimmed mean		32.688569131616880	
	Median		26.560000000000000	
	Variance		936.433	
	Std. deviation		30.601191776984226	
	Minimum		2.000000000000000	
	Maximum		299.94000000000000	
	Range		297.94000000000000	
	Interquartile range		32.709959632935510	
	Skewness		2.193	0.007
	Kurtosis		7.856	0.015

Table 8. *Extreme values of order value.*

			Cases	Value
order_value_eur	**Highest**	**1**	56,323	299.94
		2	77,067	299.89
		3	62,373	299.24
		4	41,977	298.60
		5	38,843	297.92
	Lowest	**1**	101,280	2.00
		2	77,028	2.00
		3	76,953	2.00
		4	76,947	2.00
		5	76,827	2.00 a

a. Only a partial list of cases with the value 2.00 is shown in the table of lower extremes.

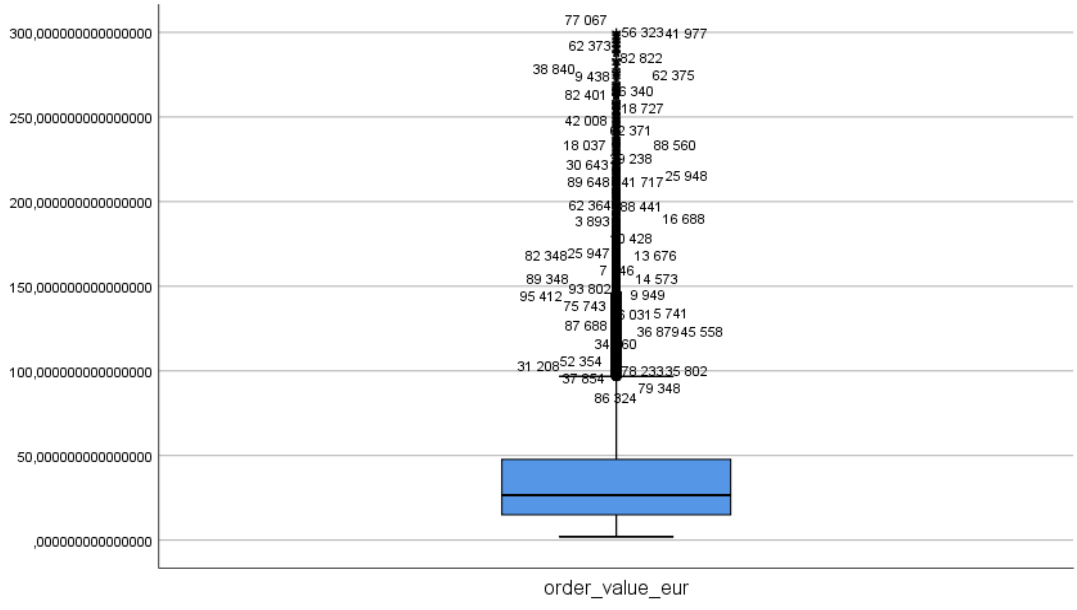

Figure 1. *Graphical representation of outlying values of order value parameter.*

From the point of view of the data structure, we can state the following. First, the model input data were created by combining two initial datasets: Customer Entity and Order Sale. These two datasets comprise a set of data that, in the case of Customer Entity, represent basic information about individual customers (specifically selected variables will be described in the next section). In the case of the Order Sale database, we can talk about identification parameters associated with a specific executed order.

Description of individual parameters of the input dataset:

- **customer_email**: This parameter represents the customer's unique identifier in the form of an email address, which the customer entered during his registration into the company registration system.
- **order_value**: As the parameter name suggests, it represents the total value of a specific order in euros.
- **gender**: This parameter reflects the gender of a particular customer:
    - male
    - female
- **delivery_method**: This parameter describes the method used to deliver the order. Specifically, we divided the input data into the following two groups:
    - delivery_to_address
    - pick_up_delivery
- **country**: The aspect of the customer's country of origin is expressed by the parameter "country". We can divide customers into several groups:
    - Slovakia
    - Czech Republic
    - Hungary
    - Romania
    - …
- **status**: final status of executed order (completed, waiting, cancelled)
- **day_period**: Looking at the time aspects of our data, we chose the periods of the day in which the order was created as the first parameter. In practice, this means that the parameter "day_period" represents the division of the day into five-day intervals as follows:
    - Morning (from 6:00 am to 11:00 am)
    - Noon (from 11:01 am to 2:00 pm)
    - Afternoon (from 2:01 pm to 4:30 pm)
    - Evening (from 4:31 pm to 7:00 pm)
    - Nigh (from 7:01 pm to 5:59 am)

This parameter was created based on the date and time of the order creation. The time intervals were chosen subjectively by our party.

- **day_name**: This parameter represents the following type of time-based variable: the name of the day on which the order was completed. Naturally, we are talking about the following seven options:
    - Monday
    - Tuesday
    - Wednesday

- Thursday
- Friday
- Saturday
- Sunday

This parameter was created based on the date and time of the order.

- **quarter:** We can consider the division of the year into quarters as the third parameter connected with the time of the order creation. From our point of view, selection into quarters is necessary for logistics and procurement processes. We divided the year naturally into:
  - Q1
  - Q2
  - Q3
  - Q4

This parameter was created based on the date and time of the order.

4.3 Data analysis description

In the modelling process, as part of the CRISP-DM methodology, we chose the following algorithms as the three main options for cluster analysis:

- TwoStep clustering
- K-means clustering
- Kohonen clustering

We can consider the TwoStep clustering method as the most appropriate modelling method based on the tests performed. We concluded thus by comparing the silhouette scores of the different modelling approaches in the clusters of orders that we had created according to their country of origin and customer gender. An overview of the results of the other methods can be seen in the Table 9.

Table 9. *Overview of results of silhouette score parameter.*

	TwoStep	K-means	Kohonen
	Silhouette score		
Male (SK)	0.425	0.241	0.207
Male (global)	0.326	0.226	0.204
Female (SK)	0.425	0.222	0.212
Female (global)	0.423	0.271	0.213
All orders	0.399	0.246	0.159

In cluster analysis, determining an appropriate number of clusters is a crucial task that directly influences the quality and usefulness of the resulting groups. For example, inadequate clustering may result if the number of sets is too low, as the clusters may be overly broad and fail to capture relevant distinctions within the data.

Conversely, if the number of clusters is excessively high, the groups may be particular, leading to overfitting and possibly masking more significant patterns or trends within the data. As a result, identifying the optimal number of clusters is necessary to ensure that the collections accurately reflect the underlying structures and patterns in the data and provide valuable insights and information to facilitate further analysis and decision making.

We followed two approaches to determine the optimal number of clusters. The first approach was the Akaike Information Criterion (AIC) and its evolution for different numbers of resulting sets. The results for each order group can be seen in the Table 10.

Table 10. *Overview of AIC values.*

MALE (SK)				
Number of clusters	Akaike information criterion (AIC)	AIC change	Ratio of AIC changes	Ratio of distance measures
1	717,981.73			
2	659,925.45	-58,056.28	1.00	1.86
3	628,674.44	-31,251.01	0.54	1.12
4	600,858.69	-27,815.75	0.48	1.21
5	577,815.52	-23,043.17	0.40	1.03
6	555,406.71	-22,408.80	0.39	1.12
7	535,399.71	-20,007.00	0.34	1.32
8	520,232.09	-15,167.62	0.26	0.95
9	504,225.66	-16,006.43	0.28	1.02
10	488,522.72	-15,702.94	0.27	1.14
11	474,773.46	-13,749.25	0.24	1.19
12	463,206.03	-11,567.43	0.20	1.07
13	452,365.09	-10,840.95	0.19	1.00
14	441,568.03	-10,797.05	0.19	1.12
15	431,929.95	-9,638.09	0.17	1.00

MALE (GLOBAL)				
Number of clusters	Akaike information criterion (AIC)	AIC change	Ratio of AIC changes	Ratio of distance measures
1	86,247.03			
2	78,783.10	-7,463.94	1.00	1.22
3	72,653.13	-6,129.96	0.82	1.93
4	69,496.40	-3,156.73	0.42	1.14
5	66,739.00	-2,757.41	0.37	1.08
6	64,188.30	-2,550.70	0.34	1.26
7	62,173.76	-2,014.54	0.27	1.01
8	60,177.84	-1,995.92	0.27	1.15
9	58,453.06	-1,724.78	0.23	1.06
10	56,820.51	-1,632.55	0.22	1.03
11	55,232.41	-1,588.09	0.21	1.06
12	53,741.38	-1,491.03	0.20	1.09
13	52,379.31	-1,362.07	0.18	1.01
14	51,029.54	-1,349.76	0.18	1.25
15	49,959.62	-1,069.93	0.14	1.05

FEMALE (SK)				
Number of clusters	Akaike information criterion (AIC)	AIC change	Ratio of AIC changes	Ratio of distance measures
1	1,231,064.47			
2	1,112,810.59	-118,253.88	1.00	2.16
3	1,058,084.34	-54,726.25	0.46	1.01
4	1,004,115.51	-53,968.83	0.46	1.60
5	970,455.77	-33,659.73	0.28	1.10
6	939,944.04	-30,511.73	0.26	1.25
7	915,612.41	-24,331.64	0.21	1.03
8	891,942.41	-23,669.99	0.20	1.09
9	870,298.98	-21,643.43	0.18	1.18
10	851,897.83	-18,401.15	0.16	1.05
11	834,296.12	-17,601.71	0.15	1.02
12	817,060.34	-17,235.78	0.15	1.11
13	801,515.44	-15,544.90	0.13	1.12
14	787,634.74	-13,880.70	0.12	1.05
15	774,457.14	-13,177.60	0.11	1.06

FEMALE (GLOBAL)				
Number of clusters	Akaike information criterion (AIC)	AIC change	Ratio of AIC changes	Ratio of distance measures
1	37,095.08			
2	34,401.97	-2,693.11	1.00	1.45
3	32,560.38	-1,841.59	0.68	1.24
4	31,078.95	-1,481.43	0.55	1.15
5	29,792.24	-1,286.71	0.48	1.22
6	28,739.99	-1,052.25	0.39	1.15
7	27,832.25	-907.74	0.34	1.06
8	26,979.94	-852.31	0.32	1.09
9	26,201.13	-778.81	0.29	1.08
10	25,483.43	-717.70	0.27	1.04
11	24,791.49	-691.94	0.26	1.24
12	24,237.75	-553.74	0.21	1.11
13	23,744.28	-493.47	0.18	1.01
14	23,258.44	-485.85	0.18	1.04
15	22,790.63	-467.80	0.17	1.07

We also paid considerable attention to graphical representations as a valuable tool for cluster analysis. These visual methods offer a convenient and efficient way to determine the optimal number of clusters. This is particularly useful in cases where the number of groups cannot be resolved through other analytical approaches, such as statistical tests. Therefore, graphical representations can aid researchers in selecting the appropriate number of clusters and improve the overall quality of the cluster analysis results.

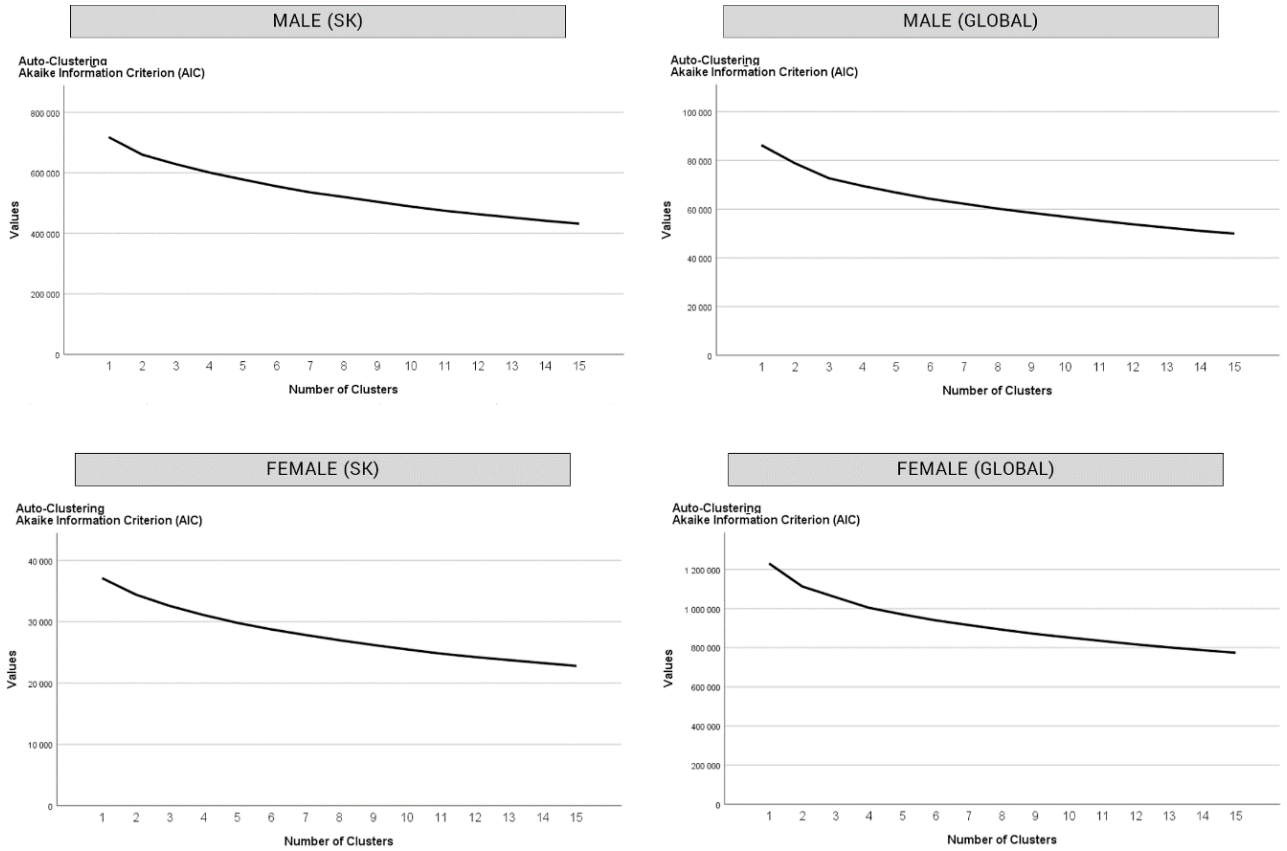

Figure 2. *Graphical overview of AIC parameter.*

In conducting our analysis, we recognised the importance of identifying the optimal number of clusters for the group and categorising data effectively. Selecting an inappropriate number of sets may result in the loss of basic patterns or groupings in the data or, conversely, create arbitrary collections that could have been more meaningful and useful.

Upon conducting the requisite steps, we determined that the optimal number of clusters for the analysis was three. Our conclusion was based on the AIC values of the individual cluster counts, which indicated that the most suitable outcome for each order group was either two or three clusters. Since the division into only two sets appeared overly simplistic, we proceeded with the three-cluster option.

5 Results and Discussion

Before the analysis and modelling, we assumed relatively significant differences in customer behavioural characteristics. As mentioned earlier, this assumption was based on the input data being split into two steps. The first step was dividing all orders according to their country of origin, specifically those from Slovakia and other countries (global). Orders from the Czech Republic, Romania and Hungary represented these other "global" countries. In the second step, since we are talking about a company that sells exercise and healthy lifestyle products, we assumed that the subjects' behaviour also differed between genders, mainly male and female. Therefore, we proceeded to the following division of orders:

- Male (Slovakia) – orders from Slovakia, made by a male,
- Male (global) – orders from non-Slovak countries made by a male,
- Female (Slovakia) – orders from Slovakia, made by a female,
- Female (global) – orders from non-Slovak countries made by a female.

The following subsections will discuss the different order groups and their typical characteristics.

5.1 Male (Slovakia) – Orders from Slovakia, made by a male

The first cluster represents men from Slovakia who shop on Mondays, at night in the first quarter of the year. The average value of the order made is equal to 28.92€. Moreover, these customers have chosen delivery to a pick-up point as the delivery method for their order. The second cluster shows men from Slovakia who shop on the same day and at time intervals during the day as customers from the first cluster. We observe a different behaviour compared to the first but equally to the third cluster for the variable "quarter," where the value Q4 defines the customers of this cluster. This value represents the fourth quarter of the year.

Regarding order value, we are talking about an amount of 39.30€, representing an increase of about 10€ compared to the first cluster. We also register differences in the delivery method, where in the second cluster, this parameter is defined by delivery directly to the customer's address. The last, third cluster represents a male from Slovakia who shops on Monday during night hours in the first quarter of the year, with an order value of 39.30€, by delivery to an address. A summary of the modelling results for the Male (Slovakia) group can be seen below.

Table 11. *Overview of TwoStep analysis results (male orders from Slovakia).*

		Inputs				
cluster	**size**	**delivery_method**	**order_value_eur**	**quarter**	**day_name**	**day_period**
cluster-1	19,991	pick_up_delivery	28.92	Q1	Monday	night
cluster-2	9,401	delivery_to_address	39.30	Q4	Monday	night
cluster-3	31,813	delivery_to_address	39.30	Q1	Monday	night

Table 12. *Overview of basic model characteristics.*

Model summary	
algorith	TwoStep
inpusts	5
clusters	3

Clusters description		
cluster name	**cluster percentage**	**number of records**
cluster-1	32.66%	19,991
cluster-2	15.36%	9,401
cluster-3	51.98%	31,813

Predictor importance	
input	**importance**
quarter	1
delivery_method	1
order_value_eur	1
day_name	0.2452
day_period	0.0465

5.2 Male (global) – Orders from non-Slovak countries made by a male

Looking at the group of orders that came from abroad and were made by men, we can state the following. The total number of records of this group of orders is 9,115. If we look at the distribution of the input records among the three clusters we defined, the second cluster is the most numerous, with 3,803 records representing a percentage share of 41.7%. The first cluster contains slightly fewer records, namely 3,148, and the third cluster is the smallest in size with only 2,164 records. When specifying the individual characteristics of these clusters, we can observe a similar homogeneity to that of the Slovak males. The first cluster represents men from abroad who buy in the first quarter of the year, and their order value is 25.30€.

Regarding time, we are talking about orders made on Mondays at night. These customers chose delivery to a pick-up point as the delivery method for their order, in contrast to the second and third clusters, where delivery directly to the customer's address was chosen. The second cluster similarly preferred night hours during the first quarter but chose Tuesday instead of Monday. In terms of order value, there is an increase of about 10€ compared to the first cluster. A similar value, 36.29€, is also recorded for the order value of the last, third cluster. A summary of the modelling results for the male (global) group can be seen below.

Table 13. *Overview of TwoStep analysis results (male orders from abroad).*

		Inputs				
cluster	**size**	**delivery_method**	**quarter**	**order_value_eur**	**day_name**	**day_period**
cluster-1	3,148	pick_up_delivery	Q1	25.30	Monday	night
cluster-2	3,803	delivery_to_address	Q1	34.13	Tuesday	night
cluster-3	2,164	delivery_to_address	Q2	36.29	Monday	night

Table 14. *Overview of essential model characteristics.*

Model summary	
algorith	TwoStep
inpusts	5
clusters	3

Clusters description		
cluster name	**cluster percentage**	**number of records**
cluster-1	34.54%	3,148
cluster-2	41.72%	3,803
cluster-3	23.74%	2,164

Predictor importance	
input	**importance**
delivery_method	1
quarter	1
order_value_eur	0.2114
day_name	0.0692
day_period	0.0444

5.3 Female (Slovakia) – Orders from Slovakia, made by a female

As in the previous clustering, we worked with setting the number of clusters to three. The total number of input records is 33,296. At first glance, we can see that compared to men from Slovakia, we register a more even distribution of individual narratives into clusters. Therefore, this distribution is almost uniform. Let us look at the characteristics of the respective clusters. The first cluster represents a woman from Slovakia who makes purchases in the first quarter of the year with an order value of 39.50€ on Tuesday night and delivered to the customer's address. For the second cluster, we record differences only on the day the order was created and the order value. Specifically, we are talking about the amount of 38.72€. However, this amount is very similar to the first cluster, so we cannot consider this result different. Thus, the only difference we register is that the order is executed on Monday. The last cluster can be defined as a woman from Slovakia who shops on Monday, during night hours in the first quarter of the year. However, the delivery method differs between the first and second clusters. Orders in the second group are delivered to a pick-up point, and the value of the shipment itself is about 10€ lower than in the previous clusters. The modelling results for the female (Slovakia) group can be summarised below.

Table 15. *Overview of TwoStep analysis results (female orders from Slovakia).*

		Inputs				
cluster	**size**	**day_name**	**delivery_method**	**order_value_eur**	**quarter**	**day_period**
cluster-1	10,444	Tuesday	delivery_to_address	39.50	Q1	night
cluster-2	12,368	Monday	pick_up_delivery	38.72	Q1	night
cluster-3	10,484	Monday	delivery_to_address	27.40	Q1	night

Table 16. *Overview of essential model characteristics.*

Model summary	
algorith	TwoStep
inpusts	5
clusters	3

Clusters description		
cluster name	**cluster percentage**	**number of records**
cluster-1	31.37%	10,444
cluster-2	37.15%	12,368
cluster-3	31.49%	10,484

Predictor importance	
input	**importance**
day_name	1
delivery_method	1
order_value_eur	0.8856
quarter	0.1666
day_period	0.0113

5.4 Female (global) – Orders from non-Slovak countries made by a female

The last group to be monitored are orders placed by women from abroad. As in the previous cases, the number of clusters was set to three. Looking at the analysis results and the variable characteristics, we can observe a similarity between the first and second clusters in the quarter of the year. In the same way, we can keep homogeneity in the case of the time interval of the day in which the order is executed. In all three clusters, this parameter takes the value of night hours. On the contrary, we can see differences in the result of the day of the week the order is executed. While for the first and third clusters, this day is Monday, for the second cluster, we are talking about Sunday.

Similarly, differences in the value of the order itself are also evident. The second and third clusters are around 35€, while the first cluster is only around 25€. The last variable compared was the order delivery method. Customers chose delivery to an address for the first and third clusters, whereas delivery to a pick-up point for the second cluster. A summary of the modelling results for the female (global) group can be seen below.

Table 17. *Overview of TwoStep analysis results (female orders from abroad).*

		Inputs				
cluster	**size**	**delivery_method**	**quarter**	**day_name**	**order_value_eur**	**day_period**
cluster-1	988	pick_up_delivery	Q1	Monday	25.26	night
cluster-2	1,404	delivery_to_address	Q1	Sunday	36.29	night
cluster-3	781	delivery_to_address	Q4	Monday	35.73	night

Table 18. *Overview of essential model characteristics.*

Model summary	
algorith	TwoStep
inpusts	5
clusters	3

Clusters description		
cluster name	**cluster percentage**	**number of records**
cluster-1	31.14%	988
cluster-2	44.25%	1,404
cluster-3	24.61%	781

Predictor importance	
input	**importance**
delivery_method	1
quarter	1
day_name	0.1075
order_value_eur	0.0779
day_period	0.0466

5.5 Comparison of characteristics of individual customer groups of results

The primary purpose of this paper was to gain a deeper understanding of actions and choices made by customers. This study focused on two distinct factors: the customers' gender and country of origin. The following section will delve into the differences between the behaviours exhibited by customers from Slovakia and those from other countries. By exploring these variations in customer behaviour, we hope to gain valuable insights into the preferences and motivations of different groups of customers.

5.5.1 Male (Slovakia) vs male (global)

We defined the similarities or differences within the group in the previous section. One of our other objectives was to compare the specification of clusters between groups of men from Slovakia and abroad. In the figure below, we can see a graphical representation of the cross-section of characteristics across the collections of the two groups.

In this section, we can compare males from Slovakia with global males. The main difference between these two groups is that Slovak male clusters shop only on Mondays in contrast to the global group with Monday and Tuesday representation. The next difference is in the quarter variable, where Slovak males

are represented in the first and fourth quarters of the year and global males in the first and second quarters. What is more, the importance of input variables is also different.

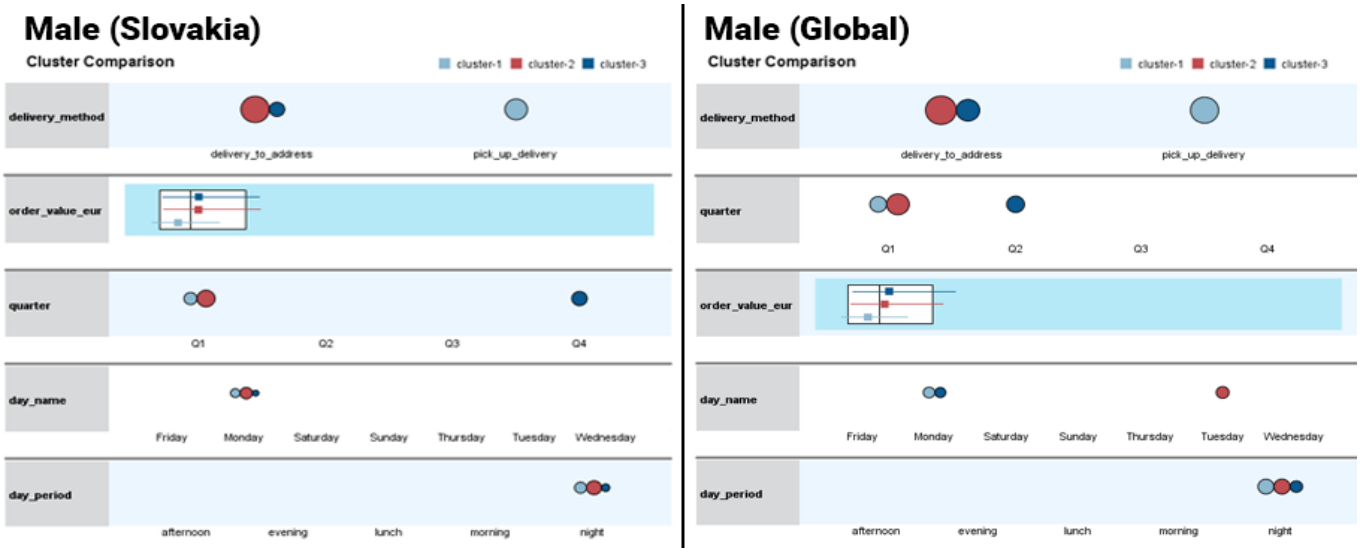

Figure 3. *Comparison of male (Slovakia) vs male (global).*

5.5.2 Female (Slovakia) vs female (global)

In the same way, let us look at the differences between the groups female (Slovakia) and female (global). At first glance, the differences will be more pronounced compared to the male groups. While the female (Slovakia) group shops on Mondays and Tuesdays, the female (global) group shops on Mondays and Sundays. Similarly, we also register differences in quarter conversion. Female (Slovakia) is represented purely by the first quarter of the year, while the female (global) group is represented by the fourth quarter of the year in addition to the first. Conversely, the same values are taken repeatedly by the period of the day, which in each case takes the value "night," represented by night hours.

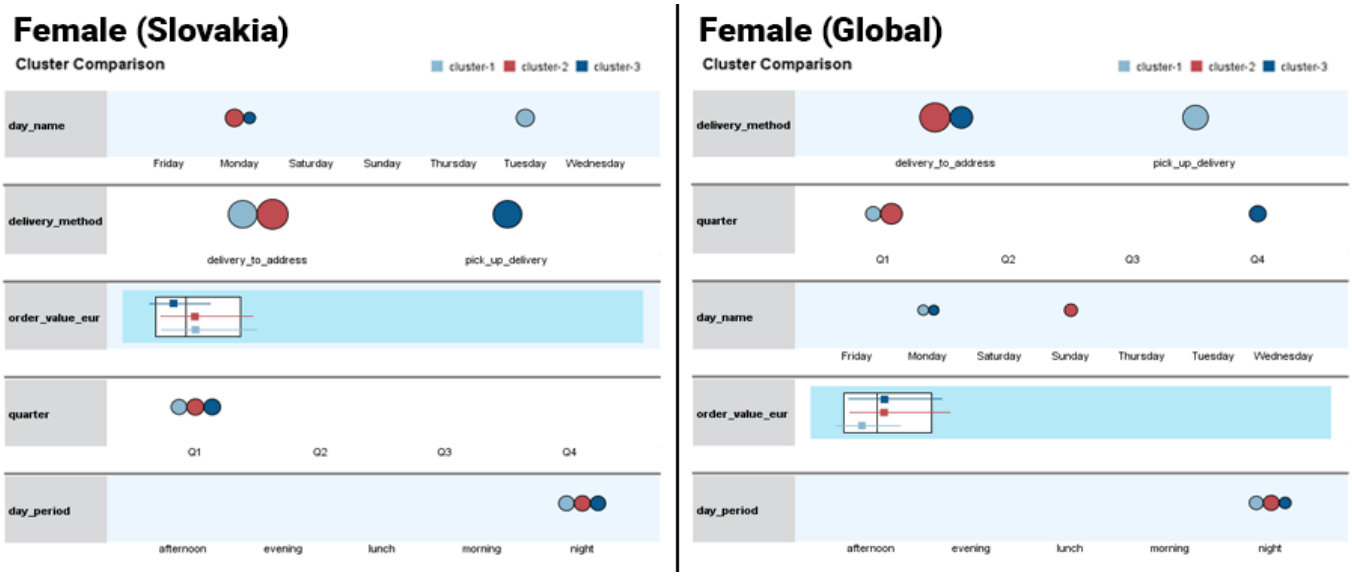

Figure 4. *Comparison of female (Slovakia) vs female (global).*

6 Conclusion and Limitations

Customer segmentation entails categorising customers into clusters based on analogous needs and their reactions to various marketing strategies and products (Alves Gomes & Meisen, 2023). It operates on the

foundational belief that diverse customer groups necessitate varying marketing approaches. As computer networks burgeoned, enterprises have accrued vast customer data in digital repositories (Chen et al., 2012). Considering the shortcomings inherent in conventional enterprise customer segmentation methodologies, novel segmentation techniques have been formulated specifically for entities engaged in the e-commerce sphere (Singh & Mittal, 2021). Segmenting a customer base enhances the depth of an organisation's relationship with its clientele. While the acquisition of new customers remains essential, the preservation of existing clientele is of paramount importance (Butt et al., 2012).

The study conducted in this article highlights the critical significance of understanding consumer behaviour for companies. The results demonstrate that comprehending customer behaviour is pivotal to the success of any business. By analysing customer behaviour patterns, companies can identify their preferences and motivations, allowing them to tailor their products and services accordingly. Through data-driven analysis, businesses can understand consumer behaviour comprehensively, leading to improved decision-making, product development and marketing strategies. (Anshari et al., 2019)

The research conducted in this article has successfully identified crucial customer behaviour metrics, which can provide valuable insights into the preferences and motivations of different groups of customers. Furthermore, the report has elucidated the potential utility of data analytics tools for the monitoring and analysis of these metrics. By exploiting these reporting and visualisation tools, companies can effectively track and scrutinise key customer behaviour metrics. Such metrics can be tracked over time and across various customer segments, allowing companies to discern patterns and trends in customer behaviour. In addition, the use of analytical tools can also facilitate the identification of potential problems or issues that may be affecting customer behaviour, enabling companies to proactively address these issues and improve overall customer experience.

As mentioned, we worked with categorical variables in the analysis. However, the only numeric variable was the order value. This limitation of the input data prevented us from using the most widely used clustering method – the K-means algorithm. However, we solved this problem by choosing a different, proper form of clustering that works with so-called mixed data (numeric and categorical variables). However, a problem arose in the graphical visualisation of the clustering results. When trying to plot the visualisation of the clusters, due to the categorical variables, we encountered the problem of displaying them in space since the maximum number of values of each variable was 7, specifically for the variable "day_name." Therefore, the graphical representation of clustering in the case of categorical variables did not have a meaningful or representative value from our point of view. The individual variables on the x-axis and y-axis represented only rows or columns of issues of personal values. For this reason, we decided not to include a graphical representation of the clusters in the visual content of the article.

Among the main limitations of the research, we can consider the shortcomings in determining the gender of each customer. This limitation is mainly due to the two different approaches to determining the variable "gender." While a particular part of the input data represents an exact definition of gender, in some circumstances, the company was unable to identify and obtain the customer's gender and, for this reason, used a prediction model to determine gender. This prediction results in the percentage probability of the gender of that given order. Even though the accuracy percentage is in the 85-95% range in 99% of the cases, we cannot 100% verify the validity of this data. However, for our research purposes, we have taken these prediction values to be the exact values of gender.

Another limitation of our research is the different types of names in each order. As mentioned, the company operates on several European markets. However, some countries where the company operates do not have the euro as a payment currency. Examples are Hungary, the Czech Republic and Romania. Therefore, we had to convert orders from such countries at the conversion rate against the euro for our research. However, it should be noted that this conversion created a slight distortion in the value of the order itself, as the exchange rate changes over time. For example, for data between 2014 and 2019, the

exchange rate may have differed significantly in specific periods. Therefore, we recommend that the company implement a module that would convert the non-eurozone sourced values to euros in real time based on the current exchange rate to remove this limitation.

Based on the results of our research, there are slight differences in the characteristics of the clusters. Our initial assumption, which considered significantly different behaviours of customers from Slovakia and other countries and between the genders, turned out wrong. On the contrary, our research uncovered that customer behaviour and traits are relatively homogeneous, despite minor differences. We can consider the time of day and the quarter when customers shop to be typical characteristics. We also observed similarities in the day of the week itself.

Customer segmentation is a crucial part of successful customer relationship management. Our cluster analysis shows differences in cluster specifications according to gender and country of origin. Still, the shopping behaviour is homogenous despite the customer's country of origin or gender across all clusters. Due to these facts, the company should focus on the specific days of the week and times of the day in its marketing strategy, according to research findings. These real-world examples can also inspire and motivate businesses to invest in data analytics tools, as they can see the tangible benefits that other companies have achieved. Ultimately, including real-world examples can help bridge the gap between theoretical concepts and practical applications, making the article more informative and engaging for readers. Moreover, we are working on an in-depth analysis of the company's customers to bring more specific research results based on expanding the input parameters.

Additional Information and Declarations

Conflict of Interests: The authors declare no conflict of interest.

Author Contributions: T.P.: Formal analysis, Data curation, Investigation, Resources, Software, Validation, Visualization, Writing – Original draft preparation, Writing – Reviewing and Editing. J.B.: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Writing – Original draft preparation, Writing – Reviewing and Editing.

Informed Consent Statement: The authors declared that they had obtained written informed consent from the subject used in the research to publish this article.

Data Availability: Due to the commercial nature of the research, supporting data is not available.

References

Alrumiah, S. S., & Hadwan, M. (2021). Implementing big data Analytics in E-Commerce: vendor and customer view. *IEEE Access*, 9, 37281–37286. https://doi.org/10.1109/access.2021.3063615

Alsayat, A. (2023). Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca. *Neural Computing and Applications*, 35(6), 4701–4722. https://doi.org/10.1007/s00521-022-07992-x

Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, (in press). https://doi.org/10.1007/s10257-023-00640-4

Anshari, M., Almunawar, M. N., Lim, S. A., & Al-Mudimigh, A. S. (2019). Customer relationship management and big data enabled: Personalization & customization of services. *Applied Computing and Informatics*, 15(2), 94–101. https://doi.org/10.1016/j.aci.2018.05.004

Butt, F.-D., Bhutto, N. A., & Laghari, M. K. (2012). Exploring the difference between stayers and switchers as corporate customers for life insurance companies in Sindh. *Asian Social Science*, 8(6), 233–239. https://doi.org/10.5539/ass.v8n6p233

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *The Journal of Database Marketing & Customer Strategy Management*, 19(3), 197–208. https://doi.org/10.1057/dbm.2012.17

Chocarro, R., Cortiñas, M., & Villanueva, M. L. (2015). Customer heterogeneity in the development of e-loyalty. *Journal of Research in Interactive Marketing*, 9(3), 190–213. https://doi.org/10.1108/jrim-07-2014-0044

Coston, A., Kawakami, A., Zhu, H., Holstein, K., & Heidari, H. (2023). A validity perspective on evaluating the justified use of data-driven decision-making algorithms. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE. https://doi.org/10.1109/satml54575.2023.00050

Feng, Y., Zhao, Y., Zheng, H., Li, Z., & Tan, J. (2020). Data-driven product design toward intelligent manufacturing: A review. *International Journal of Advanced Robotic Systems*, 17(2), 172988142091125. https://doi.org/10.1177/1729881420911257

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8), 578–588. https://doi.org/10.1093/comjnl/41.8.578

Gajanova, L., Nadanyiova, M., & Moravcikova, D. (2019). The use of demographic and psychographic segmentation to creating marketing strategy of brand loyalty. *Scientific Annals of Economics and Business*, 66, 65–84. https://doi.org/10.2478/saeb-2019-0005

Godinho, M. A., Borda, A., Kariotis, T., Molnar, A., Kostkova, P., & Liaw, S.-T. (2021). Knowledge co-creation in participatory policy and practice: Building community through data-driven direct democracy. *Big Data & Society*, 8(1). https://doi.org/10.1177/20539517211019430

Grandhi, B., Patwa, N., & Saleem, K. (2020). Data-driven marketing for growth and profitability. *Euromed Journal of Business*, 16(4), 381–398. https://doi.org/10.1108/emjb-09-2018-0054

Griva, A., Zampou, E., Stavrou, V., Papakiriakopoulos, D., & Doukidis, G. I. (2022). A two-stage business analytics approach to perform behavioural and geographic customer segmentation using e-commerce delivery data. *Journal of Decision Systems*, (in press), 1–29. https://doi.org/10.1080/12460125.2022.2151071

Güçdemir, H., & Selim, H. (2015). Integrating multi-criteria decision making and clustering for business customer segmentation. *Industrial Management & Data Systems*, 115(6), 1022–1040. https://doi.org/10.1108/imds-01-2015-0027

Hallishma, L. (2023). Customer Segmentation Based on RFM Analysis and Unsupervised Machine Learning Technique. In I. Woungang, S. K. Dhurandher, K. K. Pattanaik, A. Verma, & P. Verma (Eds.), *Advanced Network Technologies and Intelligent Computing* (pp. 46–55). Springer Nature. https://doi.org/10.1007/978-3-031-28183-9_4

Hicham, N., & Karim, S. (2022). Analysis of unsupervised machine learning techniques for an efficient customer segmentation using clustering ensemble and spectral clustering. *International Journal of Advanced Computer Science and Applications*, 13(10), 122–130. https://doi.org/10.14569/ijacsa.2022.0131016

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. *Procedia CIRP*, 79, 403–408. https://doi.org/10.1016/j.procir.2019.02.106

Kovács, T., Kő, A., & Asemi, A. (2021). Exploration of the investment patterns of potential retail banking customers using two-stage cluster analysis. *Journal of Big Data*, 8(1), Article 141. https://doi.org/10.1186/s40537-021-00529-4

Milligan, G. W., & Cooper, M. C. (1987). Methodology Review: Clustering Methods. *Applied Psychological Measurement*, 11(4), 329–354. https://doi.org/10.1177/014662168701100401

Mkpojiogu, E. O. C., & Hashim, N. L. (2016). Understanding the relationship between Kano model's customer satisfaction scores and self-stated requirements importance. *SpringerPlus*, 5, Article 197. https://doi.org/10.1186/s40064-016-1860-y

Pradana, M., & Ha, H. (2021). Maximizing strategy improvement in mall customer segmentation using K-means clustering. *Journal of Applied Data Sciences*, 2(1), 19–25. https://doi.org/10.47738/jads.v2i1.18

Rushi, W., & Pradhan, V. (2023). Factors influencing customer grocery shopping behaviour amid COVID-19 pandemic. *Cardiometry*, 25, 743–755. https://doi.org/10.18137/cardiometry.2022.25.743755

Saltz, J. S. (2021). CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. In *2021 IEEE International Conference on Big Data* (pp. 2337–2344). IEEE. https://doi.org/10.1109/BigData52589.2021.9671634

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 181, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Singh, J., & Mittal, M. (2021). Customer's purchase prediction using customer segmentation approach for clustering of categorical data. *Management and Production Engineering Review*, 12(2), 57–64. https://doi.org/10.24425/mper.2021.137678

Sokol, O., & Holý, V. (2020). The role of shopping mission in retail customer segmentation. *International Journal of Market Research*, 63(4), 454–470. https://doi.org/10.1177/1470785320921011

Vaca, C., Riofrío, D., Pérez, N., & Benítez, D. (2020). Buy & Sell Trends Analysis Using Decision Trees. In *2020 IEEE Colombian Conference on Applications of Computational Intelligence (IEEE ColCACI 2020)* (pp. 1–6). IEEE. https://doi.org/10.1109/ColCACI50549.2020.9247907

Wan, S., Chen, J., Qi, Z., Gan, W., & Tang, L. (2022). Fast RFM model for customer segmentation. In *Companion Proceedings of the Web Conference 2022 (WWW '22)* (pp. 965–972). Association for Computing Machinery. https://doi.org/10.1145/3487553.3524707

Xiao, B., & Piao, G. (2022). Analysis of influencing factors and enterprise strategy of online consumer behavior decision based on association rules and mobile computing. *Wireless Communications and Mobile Computing*, 2022, Article ID 6849017. https://doi.org/10.1155/2022/6849017

Yan, S., Kwan, Y. H., Tan, C. S., Thumboo, J., & Low, L. L. (2018). A systematic review of the clinical application of data-driven population segmentation analysis. *BMC Medical Research Methodology*, 18(1), Article 121. https://doi.org/10.1186/s12874-018-0584-9

Yi, S., & Liu, X. (2020). Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. *Complex & Intelligent Systems*, 6(3), 621–634. https://doi.org/10.1007/s40747-020-00155-2

Editorial record: The article has been peer-reviewed. First submission received on 22 December 2022. Revisions received on 5 March 2023, 4 May 2023, 10 June 2023 and 27 September 2023. Accepted for publication on 2 October 2023. The editor in charge of coordinating the peer-review of this manuscript and approving it for publication was Zdenek Smutny.

Acta Informatica Pragensia is published by Prague University of Economics and Business, Czech Republic.

ISSN: 1805-4951