Application of K-means Clustering Algorithm with RFM for Customer Segmentation in Logistics Sector

Author: Sergey Vichev

Abstract: In today's business landscape, understanding customer behavior is crucial for organizations to optimize their operations, improve profitability and increase retention. Customer segmentation, as a marketing analytical tool, plays an essential role in identifying distinct groups of customers based on their purchasing patterns. This work proposes the application of K-means clustering algorithm for customer segmentation in the logistics sector using an anonymous dataset from a real logistics company with a 10-year history and 1043 customer records. The study aims to enhance decision-making processes by delineating customer segments based on transactional behavior metrics such as total revenue, recency, and frequency of orders (RFM analysis). By employing this unsupervised machine learning technique, are identified distinct clusters that allow producing tailored business strategies and managerial decisions. The strategic significance of each customer segment is highlighted by their unique characteristics. This work underscores the power of K-means clustering in extracting meaningful patterns from complex datasets, providing a valuable tool for logistics enterprises to gain competitive advantage through nuanced customer insights.

Keywords: K-means clustering; customer segmentation; logistics sector; unsupervised machine learning; R; RStudio; RFM analysis

JEL: M42

1. INTRODUCTION

Understanding customer behavior is crucial for businesses in today's competitive landscape, particularly in sectors like logistics where efficient operations and profitability are key. Customer segmentation, as a marketing analytical tool, plays an essential role in identifying distinct groups of customers based on their purchasing patterns. This work aims to evaluate the application of unsupervised machine learning algorithms, specifically K-means clustering, for customer segmentation in the logistics sector using an anonymous dataset from a real logistics company with a 10-year history and 1043 unique customer records.

The objective of this study is to extract valuable insights from customers' historic and economic behavior patterns by applying clustering methods. By employing K-means clustering, the aim is to identify distinct clusters that allow for the production of tailored business strategies and managerial decisions based on transactional behavior metrics such as total revenue, recency, and frequency of orders (RFM analysis).

The strategic significance of each customer segment will be highlighted by their unique characteristics. This work underscores the power of K-means clustering in extracting meaningful patterns from complex datasets, providing a valuable tool for logistics enterprises to gain competitive advantage through nuanced customer insights.

2. RELATED WORK

Clustering is based on information contained within the data that describes the objects and their relationships. The term clustering was first introduced by Tryon (1939).

In the realm of data analysis, unsupervised machine learning algorithms like K-means clustering have proven effective for identifying distinct customer segments. This review explores the studies that applied K-means clustering and RFM model for customer segmentation in different domains, revealing valuable insights for organizations in these industries.

RFM model was used in combination with K-means clustering in across different industries.

In related work, Onur Doğan (2018) explored the application of clustering methods in conjunction with the RFM model for customer segmentation in the retail industry. The study analyzed 700032 customer records and Doğan proposed two different clustering models to segment customers based on their RFM values, aiming to provide better customer understanding, well-designed strategies, and more efficient decisions.

In public transporation Chen's (2022) study focused on analyzing the patterns of transit travellers in Taipei, Taiwan, to identify distinct clusters using K-means clustering. The RFM model was employed alongside K-means clustering to construct mode-switching traveller profiles on MRT and YouBike riders. The results revealed three distinct customer segments: potential, vulnerable, and loyal. These findings can help the Ministry of Transportation encourage green transportation usage and promote a sustainable environment by tailoring strategies for each cluster.

In banking sector Aliyev's (2020) study aimed to retain customers in the banking sector. The author applied these methods to real customer data from one of Azerbaijan's largest private banks. The results showed that segmenting customers based on their buying behavior and response to new services led to improved conversion rates.

This research aligns with our study as it also emphasizes the importance of customer segmentation using RFM and unsupervised machine learning techniques for gaining valuable insights into customer groups. It also proves relevance of the study as no information was found for similar work logistics sector.

3. DATASET

For the empirical basis of this study, the dataset was provided by company "XYZ". Company "XYZ" is a real-life company, but for confidentiality purposes its true name has been replaced. The database under investigation contains data on all transactions of clients

since the inception of the company. The analyzed data covers the period from 1 August 2008 to 1 April 2018.

For confidentiality reasons, real income and expense figures have been multiplied by several coefficients, while client names have been replaced with codes.

The initial data provided by the company consists of 31,109 observations (rows) and 24 variables (columns).

The table below describes the variables and their description. Every row in the data represents the characteristics of a single order or shipment made for a client.

Tab. 1 Data dictioanary of the initial data

Variable Name	Variable Description
Relation	Relation or Direction of Shipping
Client	Name of the Client Placing the Order
Type	Type of Legal Registration of Client
EIK	Client's Registration Number
VAT	Client's VAT Number
Loading Point	Shipping Origin, Sending Location
LP_Country	Country of Origin, Sending Country
LP_Postcode	Postal Code of Sending Country
LP_City	City of Origin, Sending City
LP_Address	Address of Shipping Origin
Unloading Point	Destination, Receiving Location
UP_Country	Country of Destination, Receiving Country
UP_Postcode	Postal Code of Receiving Country
UP_City	City of Destination, Receiving City
UP_Address	Address of Destination
Volume	Shipping Volume
Weight	Shipping Weight
TypeShipment	Type of Shipment
Departure	Date and Time of Shipping Departure
Delivery	Date and Time of Delivery Estimated
Days	Number of Days in Transit
Income	Revenue from Service Provided
Expense	Cost for Service Provided
Relation	Relation or Direction of Shipping

Source: Company ‘XYZ’ data

From a technical standpoint, the format of the data is crucial when working with it because software relies on each format in a specific way. The initial data format is presented in the RStudio output below. In the initial data, three types of data can be observed:

- "int" indicates that the data consist of single integer values (1, 2, 3, 4);

- "factor" means that software assigns a number to each variant in a given variable and can represent them on a nominal or ordinal scale. In this case, it is clear that the Client variable has 1044 unique values, meaning there are 1044 unique clients.
- "num" indicates that the variable accepts numerical values.

RStudio output for initial data formats:

```
'data.frame':  31109 obs. of  24 variables:
 $ Shipment      : int
 $ Relation      : Factor w/ 134 levels
 $ Client        : Factor w/ 1044 levels
 $ Type          : Factor w/ 12 levels
 $ EIK           : Factor w/ 829 levels
 $ VAT           : Factor w/ 913 levels
 $ LoadingPoint  : Factor w/ 4535 levels
 $ LP_Country    : Factor w/ 44 levels
 $ LP_Postcode   : Factor w/ 2799 levels
 $ LP_City       : Factor w/ 3134 levels
 $ LP_Address    : Factor w/ 6231 levels
 $ UnloadingPoint: Factor w/ 5063 levels
 $ UP_Country    : Factor w/ 46 levels
 $ UP_Postcode   : Factor w/ 1659 levels
 $ UP_City       : Factor w/ 2465 levels
 $ UP_Address    : Factor w/ 7351 levels
 $ Volume        : num
 $ Weight        : num
 $ TypeShipment  : Factor w/ 6 levels
 $ Departure     : Factor w/ 13879 levels
 $ Delivery      : Factor w/ 16403 levels
 $ Days          : Factor w/ 53 levels
 $ Income        : num
 $ Expense       : num
```

To answer the posed questions and tasks regarding knowledge extraction from the database, certain variables have been selected that are significant. These variables are those that can characterize clients rather than the shipments themselves, such as: client name, revenue from the service provided, cost for the service provided, etc.

4. METHODOLOGY

4.1. K-means clustering algorithm

Clustering with K-Means is a simple and elegant approach for dividing a set of data into K distinct, non-overlapping clusters. The fundamental idea behind K-Means clustering is to define clusters in such a way as to minimize the total variation within the cluster (known as the overall intracluster variation) (Kassambara, 2017)

The goal of cluster analysis is for objects in a group to be similar (or connected) to each other and different from (or not connected with) objects in other groups. The more similarity (or homogeneity) within a group and the larger the difference between groups, the better or clearer the clustering. (Tan, Steinbach, Kumar, Karpatne, 2018)

Through cluster analysis, the possibilities of prediction are expanded. Data is distributed into homogeneous or semi-homogeneous groups, allowing them to be classified based on common characteristics. The term "cluster" means a heap, bunch. Clusters are

non-overlapping sets, each containing objects that are similar to one another and distinct from objects in the neighboring cluster. (Agarwal, 2003)

If the data from clustering is represented as points in the feature space, then clustering can be visualized through determining the concentration of points around centers of density. (Ivanov, 2016)

K-means algorithm determines the optimal number of clusters and their respective centroids by iteratively minimizing the sum of squared errors (SSE) between each data point and its closest centroid, with the goal of producing a clustering that results in the smallest SSE and, consequently, better representation of points within their assigned clusters (Tan, 2018).

The SSE is formally defined as follows:

$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2 \quad (1)$$

where:

x: an object

C_i: the i^{th} cluster

c_i: the centroid of cluster C_i

K: the number of clusters.

Clustering is mentioned in literatures is a classical method of unsepurvised machine learning. Alpaydin (2004) and Sugiyama (2016) mention data clustering as typical tasks of unsupervised machine learning used for analyzing various social networks, system diagnosis and others.

4.2. Tools

The data has been submitted to the R Studio analysis tool, where the data preparation and clustering algorithm was performed, using the R programming language.

The main libraries used for modeling are lubridate, zoo, dplyr, and gapminder. For visualization, the RStudio environment was utilized with the help of the plotly and ggplot2 libraries.

4.3. Set up

For the purposes of this study, the K-means clustering technique has been chosen for data analysis. This method is part of unsupervised machine learning.

In the data preparation phase, the data were prepared for modeling. The modeling will be performed using two types of data:

- clustering A: clustering based on three parameters: total revenue from orders, total number of orders, days since last order;

- clustering B: clustering based on three parameters: average order value, total number of orders, days since last order.

The only difference between the two types of data is in the first parameter.

4.4. Evaluation

The evaluation criteria for the models where chosen business and statistical factors:

- from business perspective - the ability of creation of client segments that can provide relevant business information, identifying which clients are more important than others, and which clients should be proactively contacted to renew orders.

- statistical criterion for the obtained clusters is the ratio of within-cluster sum of squares and total distance. The goal is for this indicator to be lower than 85%.

5. IMPLEMENTATION

5.1. Data preparation

For the purpose of modeling client segments using cluster analysis, the data was aggregated at the client level. In other words, each new row in the data array corresponds to one unique client and contain their characteristics.

Tab. 2 Data dictionary after preprocessing

Variable name	Variable Description
Client	Client Name
Type	Type of Client Registration
SumOrder	Total Revenue from Orders
AvgOrder	Average Revenue per Order
SumExpense	Total Transport Expenses
AvgExpense	Average Transport Expenses per Order
LatestDepDate	Last Shipping Date
NumberOfOrders	Number of Parcels Shipped
AvgVolume	Average Volume of Shipped Goods
SumVolume	Total Volume of Shipped Goods
AvgWeight	Average Weight of Shipped Goods
SumWeight	Total Weight of Shipped Goods
AvgDays	Average Number of Days for Delivery
SumDays	Total Number of Days for Delivery
DaysFromLastOrder	Number of Days since Last Order
MonthsFromLastOrder	Number of Months since Last Order

Source: Company ‘XYZ’ data.

The criteria for selecting parameters will be those that can be aggregated: Client Name (Client), Income, Expenses, Volume, Weight. Initially, the data is grouped for the entire period for which there are available data.

Data was formatted accordingly and 21 records with missing data deleted.

After data aggregation by client the following variable fields where obtained.

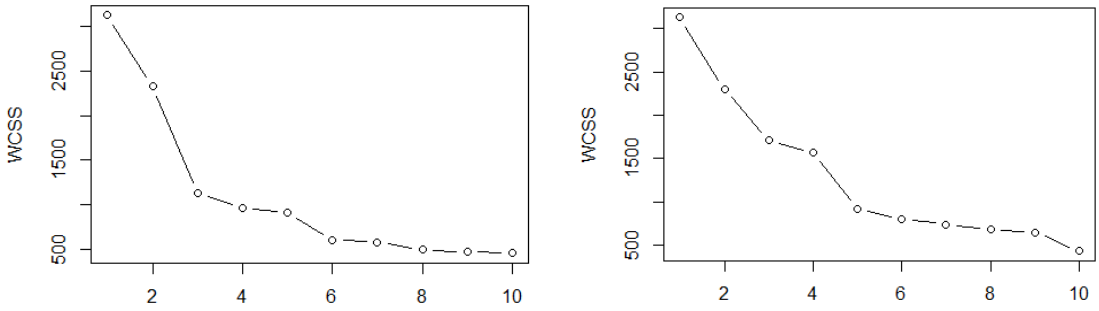
5.2. Modelling

After selecting the necessary parameters the data has been standardized, as such parameters as SumOrder has a different scale than the others, in order to prevent this parameter from dominating the others.

In the next step the optimal number of clusters was determined. This was done using the "Elbow Method." With this method, the within-cluster sum of squares (WCSS) is calculated for different numbers of clusters and visualized in a linear diagram. The optimal number of clusters is considered to be that after which the reduction in within-cluster sum of squares is much less than transitional.

For example, Figure 1 shows the reduction in within-cluster sum of squares for different numbers of clusters.

Fig. 1 WCSS for different number of clusters (left – clustering A, right – clustering B).



For clustering A, in the chart two distinct "elbow points" can be observed - at 3 clusters and 6 clusters. The reduction in within-cluster sum of squares between 3-4 and 6-7 clusters is not significant.

Given that there are a large number of clients involved and the data was collected over a period of 10 years, for K-means clustering with Euclidean distances, was chosen clustering with 6 clusters.

K-Means clustering was applied algorithm with 6 clusters, a maximum of 300 iterations, and 10 different starting positions for the centers. In other words, the algorithm will start from 10 different initial positions and repeat steps up to a maximum of 300 times.

Six client segments have been created with 264, 187, 5, 238, 25, and 324 clients respectively.

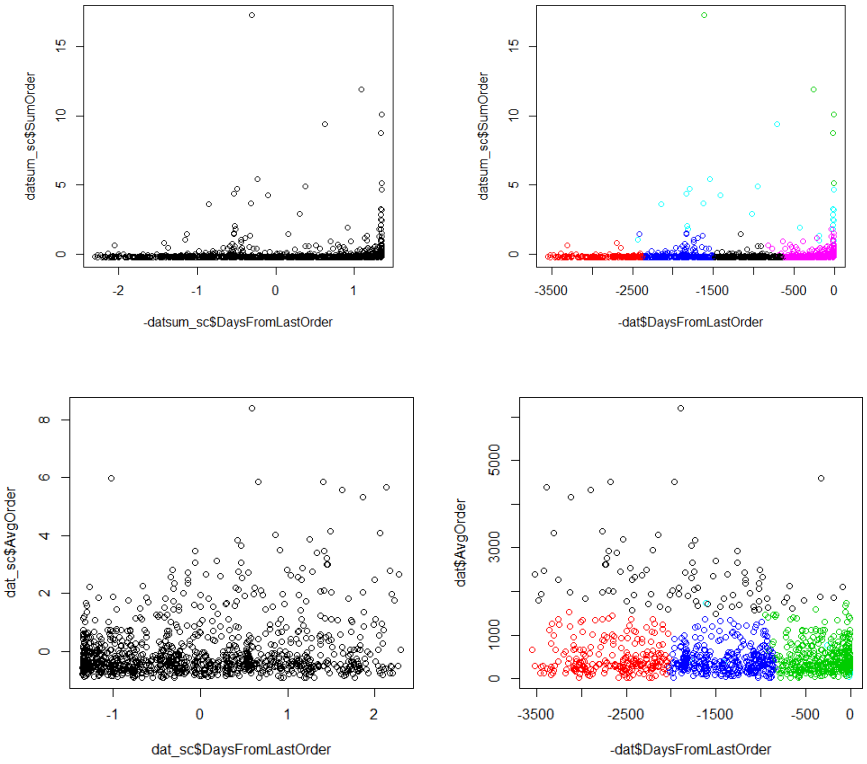
For clustering B, a "break" can be noticed in the graph, specifically at five clusters. The decrease in within-cluster sum of squares after fifth cluster is not significant. In other

words, if choosen K-means clustering algorithm with 6 or more clusters, it won't make a big difference.

K-means clustering algorithm was executed with 5 clusters, a maximum number of iterations of 300, and from ten different initial center positions for the algorithm to start from. This means that the algorithm will start anew from ten different positions and repeat the steps up to a maximum of 300 times.

Five client segments have been created with 89, 228, 374, 341, and 11 clients respectively. The figure below shows the results of the formed segments.

Fig. 2 Distribution of Clients Before and After K-Means Clustering (top – clustering A, bottom – clustering B).

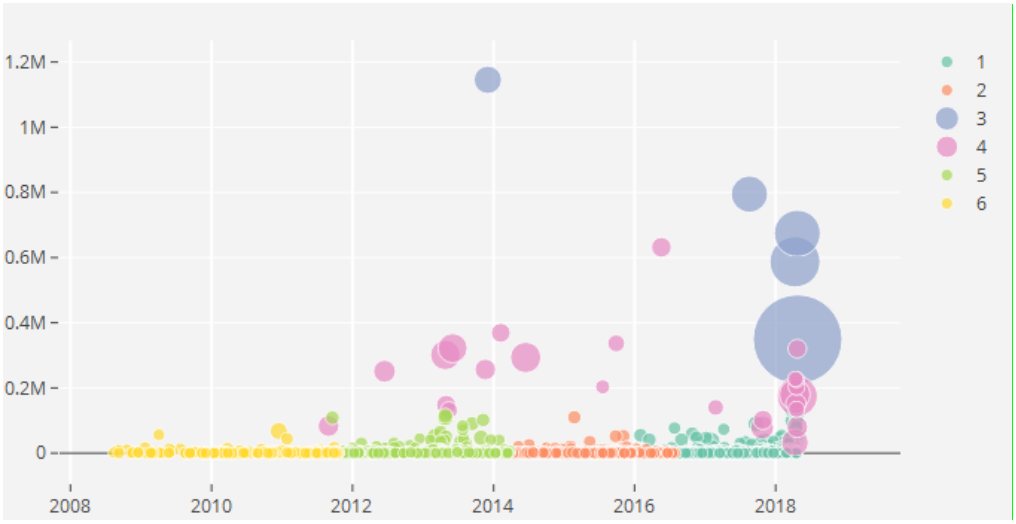


5.3. Evaluation

To evaluate the model, "labels" were applied to the received clusters against real data.

To better visualize the customer segments obtained from clustering A and B, we use a so-called “bubble chart” with 4 variables. For color distinction, we will use the order of clusters obtained. For the x-axis and y-axis, we will use real data, the date of the last order and total sales revenue (A) / average revenue per order (B), respectively. The size of the bar represents the total number of orders placed by the corresponding client.

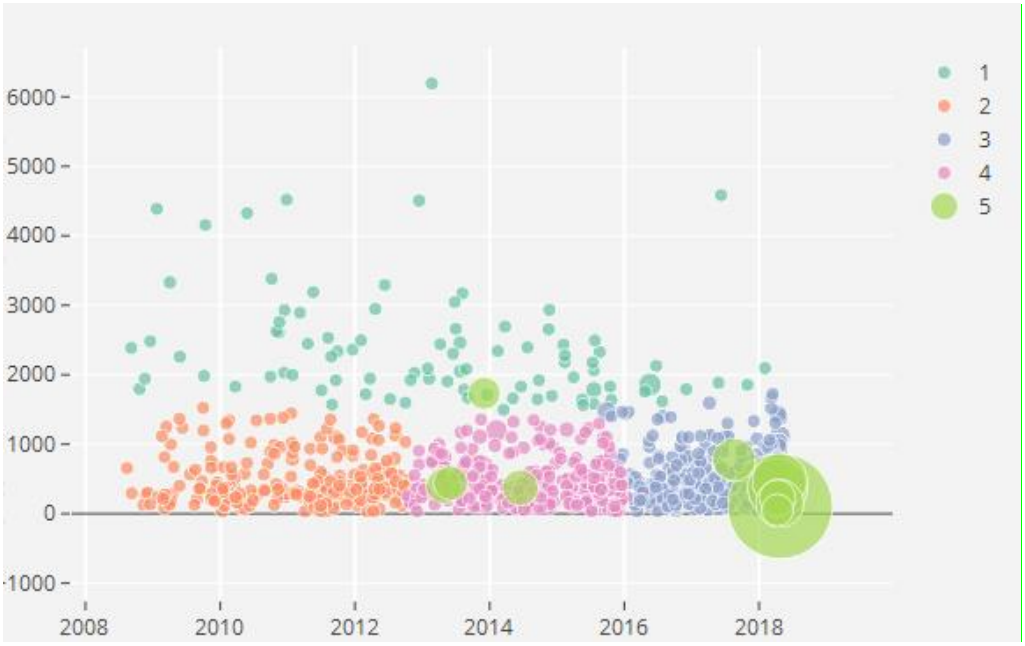
Fig. 3 Vizualized obtained customer segments (clustering A).



From this visualization, distinctive characteristics of each cluster can be identified, which will be done in the next phase.

Clearly visible customer segments indicate good performance of the model. The ratio of between-cluster square sums to the total square sum of distances is equal to 80.8%. This is a good indicator considering the chaotic distribution of clients.

Fig. 4 Vizualized obtained customer segments (clustering B).



The ratio of between-cluster square sums to the total square sum of distances is equal to 70.5%, which shows a better indicator than in clustering A and indicates that there are more clearly defined segments.

We can conclude that both models perform well and meet the initially set evaluation criteria.

6. RESULTS AND INTERPRETATION

6.1. Results evaluation

Validation of clustering models can be quite challenging as there is no clear metric for comparison. In other words, after the client segments have been formed, it cannot be determined whether they are correct or incorrect because there is no information about "correct segments."

Therefore, a criterion for evaluating the model has been chosen to be "business value," meaning that the formed segments will provide useful business information. The expected outcome is that the segments will separate important clients from less important ones and segment clients based on how long it has been since they placed an order.

Through formed customer segments, businesses can determine which clients hold key positions currently and which clients, with whom they have worked in the past, are of greatest importance. The business can then develop a strategy for each group of clients.

Customer segmentation was performed using two types of data: Clustering A and Clustering B. The data for both types of analysis included three parameters. The first two parameters were common to both types of analysis:

- Total number of clients (for Clustering A and B);
- Date of last order (for Clustering A and B).

The difference between the two types of data lies in the third parameter:

- Total sales revenue (for Clustering A);
- Average sales revenue (for Clustering B).

Evaluation of the two types of analysis is presented below.

6.1.1. Evaluation – Clustering A

Table 3 describes the six obtained clusters or segments in Clustering A, based on average order size, total number of orders, and last order date, which are visualized in Figure 3.

Most clients were segmented into clusters 1, 5, 4, and 2. The total for these clusters is relatively similar in terms of average revenue and number of orders. The main difference between them is the period in which these clients were last "active".

Clusters 6 and 3 stand out from others with significantly higher order revenue and cumulative number of orders. Clients from these segments can be identified as key for the business.

Tab. 3 Statistics of Obtained Customer Segments (Clustering A).

Cluster	Clients	MinSumOrder	AvgSumOrder	MaxSumOrder	MinOrders	AvgOrders	MaxOrders	From	Till
6	5	349 327	710 263	1145 519	661	1 581	3 197	2013-Dec-2,	2018-Apr-25,
3	25	32 485	215 199	631 688	114	414	1 187	2011-Aug-30,	2018-Apr-23,
1	324	55	10 691	132 700	1	23	296	2016-Feb-1,	2018-Apr-25,
5	264	35	3 119	110 116	1	5	91	2014-Apr-12,	2016-Jul-27,
4	238	35	8 486	114 812	1	13	181	2011-Sep-19,	2014-Mar-26,
2	187	35	3 075	67 909	1	5	236	2008-Aug-13,	2011-Oct-24,

Each obtained customer segment has distinctive characteristics that allow firm "XYZ" to apply a corresponding business strategy:

• Segment 1. Current clients – the core of the business

Key characteristic: Most recently active customers (2016-2018), average sales revenue.

Business strategy: The company should continue working actively with this group of clients and make efforts to «activate» those from 2016 and 2017.

- **Segment 2. "Lost" clients**

Key characteristic: This group of clients has a low total sales revenue and were last active between 2008 and 2011.

Business strategy: Contact clients in descending order by year, low priority.

- **Segment 3. "Silver" clients**

Key characteristic: Higher total sales revenue, more orders.

Business strategy: Retain and nurture current clients and take measures to win back lost clients.

- **Segment 4. "Forgotten" clients**

Key characteristic: Last active between 2011 and 2014, low sales revenue and fewer orders.

Business strategy: Contact clients in descending order by year, low priority.

- **Segment 5. "Seeking" clients**

Key characteristic: Low average sales revenue and fewer orders, last active between 2014 and 2016.

Business strategy: Re-evaluate potential for collaboration.

- **Segment 6. "Golden" clients**

Key characteristic: Highest average turnover;

Business strategy: Offer best working conditions to current clients, make efforts to win back lost clients.

6.1.1. Evaluation – Clustering B

Table 4 describes the five obtained segments or clusters in Clustering B, based on average order size, total number of orders, and last order date, which are visualized in Figure 4.

Tab. 4 Statistics of Obtained Customer Segments (Clustering B).

Cluster	Clients	MinAvgOrder	AvgAvgOrder	MaxAvgOrder	MinOrders	AvgOrders	MaxOrders	From	Till
1	89	1494	2375	6193	1	9	340	2008 Sep 5	2018 Feb 2
2	228	35	485	1524	1	10	432	2008 Aug 13	2012 Oct 16
3	374	35	485	1724	1	28	481	2015 Sep 28	2018 Apr 25
4	341	35	435	1355	1	13	382	2012 Oct 12	2016 Jan 13
5	11	52	464	1733	630	1162	3197	2013 Apr 26	2018 Apr 25

- **Segment 1. Disappearing Valuable Clients**

Key characteristic: Highest average revenue but few orders. Covers the entire period of the firm's operation, mainly in the beginning.

Business strategy: Analyze reasons for the decline of this client segment.

- **Segment 2. Long-Lost Clients**

Key characteristic: Low average revenue, last ordered between 2008 and 2012.

Business strategy: Contact clients in descending order by year, low priority.

- **Segment 3. Current Clients – the Core of the Business**

Key characteristic: Most recently active customers (2015-2018), low average revenue from orders.

Business strategy: The company should continue working actively with this group of clients and make efforts to «activate» clients who last ordered in 2016 and 2017.

- **Segment 4. Forgotten Clients.**

Key characteristic: Low average revenue, last ordered between 2012 and 2016 (Similar to Segment 3 but with a different period)

Business strategy: Re-evaluate potential for collaboration.

- **Segment 5. Golden Clients**

Key characteristic: Most orders.

Business strategy: Offer best working conditions to current clients, make efforts to win back lost clients.

As shown above, each customer segment has unique distinguishing key characteristics that allow businesses to apply different actions and business strategies.

Customer segments can be generated automatically using the built model and various input data.

However, a limitation of the model developed through K-means clustering is that this method requires the number of clusters K to be predefined.

6.2. Re-examination of the process

During the re-examination of the model, no correction points were identified. For future improvements to the model, it is suggested to adjust the weights of parameters based on business requirements. For instance, if for some businesses the average revenue per client is more important than other factors, then this factor should have a greater weight in forming customer segments.

7. CONCLUSION

This study was aiming to develop a customer segmentation model to enhance customer related decision-making processes for companies in the logistics sector. To achieve this, RFM technique which implies recency, frequency and monetary factors was adopted through using unsupervised machine learning K-means clustering algorithm, to produce meaningful data-driven customer segments. The obtained segments provided useful business insights into cluster characteristics and a specific strategy for each was suggested. This approach is beneficial to professionals performing customer segmentation, specifically in the logistics sector and for B2B customers, as RFM technique is focused more on the consumption factors, rather than on behavioristic ones. For future work, the use of other clustering algorithms along with additional attributes and weights, is proposed as they will help provide more insights into customer segments.

References

1. Agarwal, P., 2003. "Clustering & classification" in Algorithms in Computational Biology", Duke University, North Carolina.
2. Aliyev, M., Ahmadov, E., Gadirli, H., Mammadova, A., & Alasgarov, E., 2020. Segmenting Bank Customers via RFM Model and Unsupervised Machine Learning. arXiv:2008.08662.
3. Alpaydin, E. 2004. Introduction to Machine Learning, The MIT Press, London, 2004; ISBN: 0-262- 01211.
4. Chen, A., Liang, Y.-C., Chang, W.-J., & Siauw, H.-Y., 2022. RFM Model and K-Means Clustering Analysis of Transit Traveller Profiles: A Case Study. Journal of Advanced Transportation, 3, 1108105. DOI: 10.1155/2022/1108105. License: CC BY 4.0.
5. Doğan, O., Ayçin, E., & Bulut, Z. (2018). Customer segmentation using RFM model and clustering methods: A case study in retail industry. International Journal of Contemporary Economics and Administrative Sciences, 8(1), 1-19. ISSN: 1925–4423.
6. Kassambara. A., , 2017. Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (Multivariate Analysis) (Volume 1), STHDA.
7. Sugiyama M. 2016. Introduction to Statistical Machine Learning, Morgan Kaufmann, ISBN: 978-0-12-802121-7.
8. Tan, P., Steinbach M., Kumar V., Karpatne A., 2018. "Introduction to Data Mining, 2nd Edition", New York, NY, Pearson Education, ISBN 9780133128901.