IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Clustering Customers of an E-commerce Store using K-means Clustering and RFM Analysis

[1]Twarit Shah, [2]Vikhyat Singh, [3]Sumit Aggrawal, [4]Dr Mohd Shuaib

[4]Professor, [1,2,3]Students
[1,2,3,4]Department of Mechanical Engineering
[1,2,3,4]Delhi Technological University

Abstract: This research paper explores the application of K-means clustering and RFM analysis to segment customers in a Brazilian marketplace, aiming to enhance logistics and supply chain management and improve customer satisfaction. The study identifies four distinct customer segments: Relatively New Comers/Low Spenders, Loyal Customers, Lost/Low Spenders, and Big Spenders. Through the analysis, the marketplace can tailor its logistics and supply chain strategies based on the specific needs and preferences of each segment.

***IndexTerms* – Clustering, K-Means, RFM Analysis, Logistics, Supply-Chain**

1. INTRODUCTION

Organizations are constantly looking for new, creative methods to improve customer satisfaction and gain a competitive edge in today's extremely competitive business market. To provide a seamless customer experience in the context of e-commerce marketplaces, effective supply chain management and logistics are essential. Businesses can shorten delivery times, increase product availability, and raise customer satisfaction levels by optimizing these operations. Application of sophisticated analytics methods, like K-means clustering and RFM analysis, has shown promise in achieving these objectives. Smart supply chain solutions like K-means clustering, a prominent unsupervised machine learning technique, are required since it allows for the discovery of hidden patterns and linkages within vast amounts of consumer data.

Unsupervised machine learning technique K-means clustering is frequently used to find different groups or clusters within datasets. The three main aspects of customer behavior that RFM analysis focuses on are recency, frequency, and monetary value. Businesses can efficiently segment their client base and learn more about the various levels of client engagement and loyalty by analyzing these aspects. In order to cluster customers in a Brazilian marketplace, this research article will investigate the use of K-means clustering and RFM analysis. Through the identification of consumer categories with comparable traits and purchasing patterns, the market will be able to adjust its supply chain management and logistics procedures.

2. RELATED WORKS

2.1 Su Bu et al. "Logistics engineering optimization based on machine learning and artificial intelligence technology." J. Intell. Fuzzy Syst. (2021).

Smart logistics has become a crucial instrument for improving people's quality of life and everyday routines in the age of the Internet of Things. The current state of logistics engineering has a number of problems that are keeping it from being as effective as many had hoped. Based on these results, this paper proposes a logistics engineering optimisation approach based on artificial intelligence and machine learning. Additionally, this work suggests an enhanced multi-label chain learning technique for high-dimensional data based on the classifier chain and the combined classifier chain. In order to optimise logistics engineering and yield the optimum outcome using an artificial intelligence model, this study also combines the constraints of the logistics transportation process with the actual requirements of logistics transportation. The effectiveness of the model is examined by conducting a control experiment to measure the performance of the method proposed in this study. The study's findings reveal that the paper's proposed logistics engineering optimisation based on AI and machine learning technology has a unique practical impact.

2.2 Dino Knoll et al. "Predicting Future Inbound Logistics Processes Using Machine Learning." Procedia CIRP (2016).

The development of globalization and the rising dynamic of product life cycles, which leads to worldwide supply chain networks, have a significant impact on the Abstract Manufacturing business. The assembly line must receive a wide variety of parts from various sources and locations as part of inbound logistics. Planning for these inbound logistics procedures is based on continuously changing data from purchasing, assembly line planning, and product development. Currently, planning requires a significant amount of time spent acquiring data, and future planning rarely makes use of knowledge from earlier planning procedures. As a result, this study proposes a method for planning logistics for incoming goods. Machine learning can be used to extract general knowledge from logistical procedures and utilize that knowledge to forecast future events.

2.3 Pascal Wichmann et al. "Extracting supply chain maps from news articles using deep neural networks." International Journal of Production Research (2020).

Supply chains are becoming more multi-tiered, complicated, and global. As a result, businesses frequently struggle to keep full visibility of their supplier network. This is problematic since tasks like successfully controlling supply chain risk require visibility of the network structure. In this study, we utilize deep learning to automatically extract buyer-supplier relations from natural language text and propose automated supply chain mapping as a way to retain structural visibility of a company's supply chain. Early results indicate that organizations may be able to (a) automatically create basic supply chain maps, (b) confirm existing supply chain maps, or (c) enhance current maps with more supplier data utilizing supply chain mapping technologies that use Natural Language Processing and Deep Learning.

2.4 S. Ton et al. "Research on Supply Chain Risk Assessment Based on Support Vector Machines." Economic Survey (2014).

An essential component of supply chain risk management is supply chain risk assessment. The study employs questionnaires to develop a more scientific approach of supply chain risk assessment on the basis of literature reviews and research studies. The supply chain risk assessment model is created by using the support vector machine, a machine learning technique, to the evaluation of supply chain risk. The empirical study's findings suggest that the support vector machine model for supply chain risk assessment can fully identify actual risk and has higher training efficiency and accuracy, demonstrating the model's efficacy.

2.5 Nesma Mahmoud Taher et al. "Investigation in Customer Value Segmentation Quality under Different Preprocessing Types of RFM Attributes." Int. J. Recent Contributions Eng. Sci. IT (2016).

Consumer value segmentation aids merchants in comprehending various consumer sorts, forges lasting bonds with them, and subsequently raises their worth and loyalty. This study intends to assess the effectiveness of customer value segmentation using two preprocessing techniques for RFM variables. Based on the scored RFM and the actual value of RFM, the customer value segmentation is carried out using the K-means clustering algorithm. The Sum of Squared Error (SSE) is used to evaluate how well the clustering results are produced. The collected results demonstrate that using the real value of RFM rather than the scored RFM improves segmentation accuracy and decreases clustering error (SSE) in customer segmentation.

2.6 Sardjoeni Moedjiono et al. "Customer loyalty prediction in multimedia Service Provider Company with K-Means segmentation and C4.5 algorithm." 2016 International Conference on Informatics and Computing (ICIC) (2016).

The annual growth in the demand for internet and cable television entertainment has an impact on the emergence of numerous multimedia service provider businesses that provide a wide range of services in an effort to gain market share. Because of the numerous firm options available to them, customers are more demanding and can switch providers with ease because businesses are aware that keeping existing customers costs less than acquiring new ones. Therefore, it's crucial for a business to understand customer loyalty and to be able to estimate income for use as a benchmark in business development planning. Company wants an accurate model, therefore the researcher applies the C4.5 classification method and the K-means segmentation algorithm, which results in a model with an accuracy of 79.33% and an Area Under Curve (AUC) of 0.831. In order to improve the accuracy % in customer loyalty classification study, this research contribution uses linked data to categorize customer potential using the Recency Frequency Monetary (RFM) model.

3. DATASET

Olist, the biggest department store in Brazilian marketplaces, kindly shared this dataset. Olist effortlessly and with a single contract links small businesses from all over Brazil to channels. These business owners can use Olist logistics partners to sell their goods through the Olist Store and send them straight to customers.

Details on 100k orders placed between 2016 and 2018 on several Brazilian marketplaces are included in the dataset. Thanks to its features, you may view an order from a variety of perspectives, including customer location, product attributes, order status, pricing, payment, and freight performance. A seller is informed to complete the order after a buyer orders the product from Olist Store. After receiving the product or when the projected delivery date approaches, the consumer receives an email with a satisfaction survey where he can rate his shopping experience and provide some feedback.

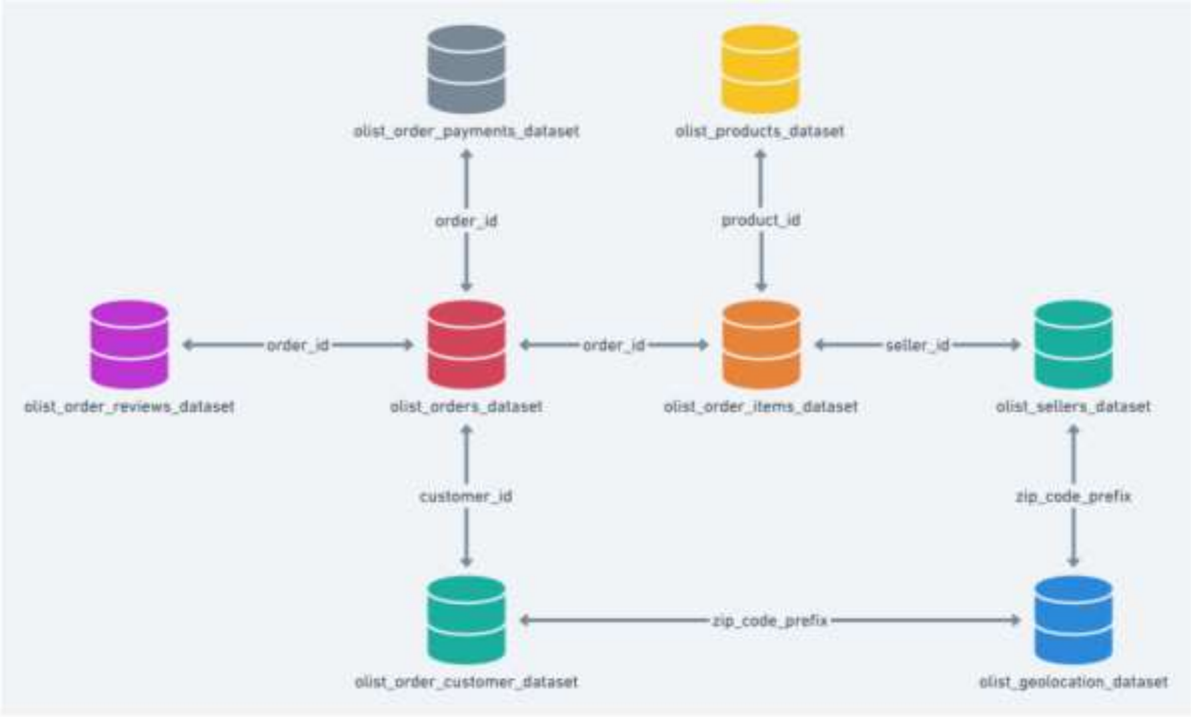


Figure 1: Data Schema

4. METHODOLOGY

4.1 Pre-processing

1. Change the datatype of "Order_status" and "product_category_name" columns from regular string to categorical variable
2. Normalize the text(city names, state names, customer reviews, product category names)
3. Create a column which tells us how quickly or slowly each item was delivered(estimated delivery date- actual delivery date).

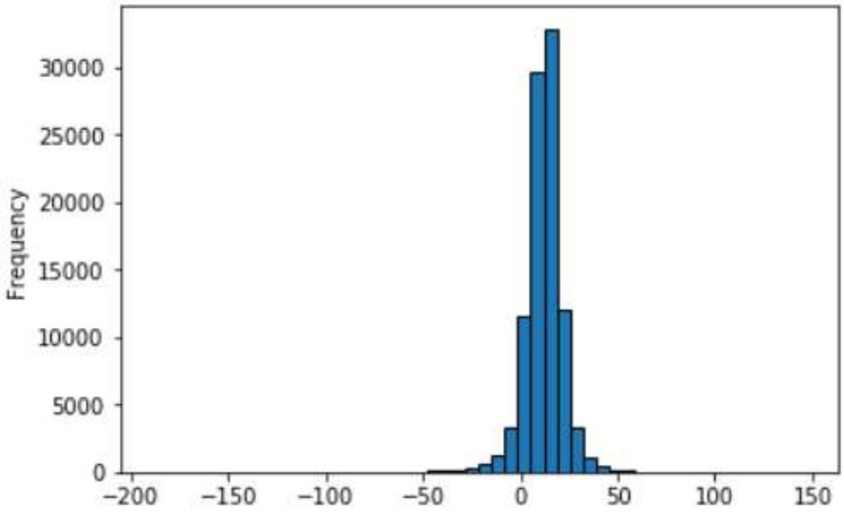


Figure 2: Number of orders delivered before/after estimated time

4.2 Exploratory Data Analysis/Visualization

The approach to ask questions about the dataset and answer questions in the form of a visualization was adopted

a. How many items are ordered at one transaction and how the total cost of each transaction increases as more quantity is added?

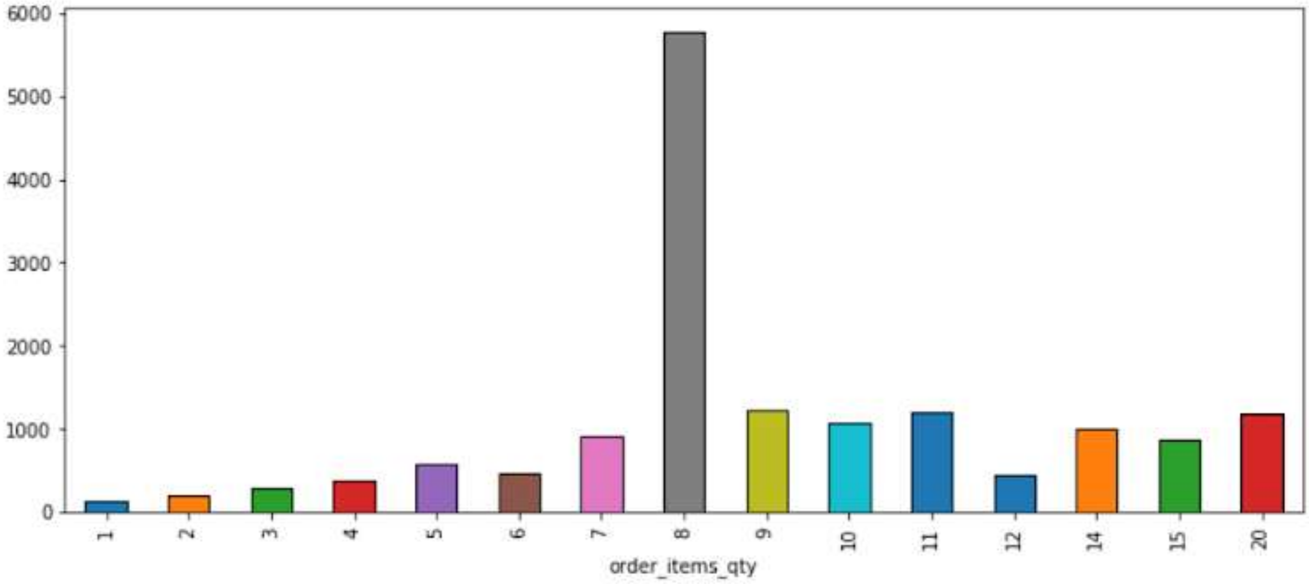


Figure 3: Distribution of Order Quantity vs Cost of transaction

Insight: 93% of transactions have only one ordered item. Moreover, more than 6 units of items were ordered in transaction only in 0.024% cases.

b. What is the average order value?

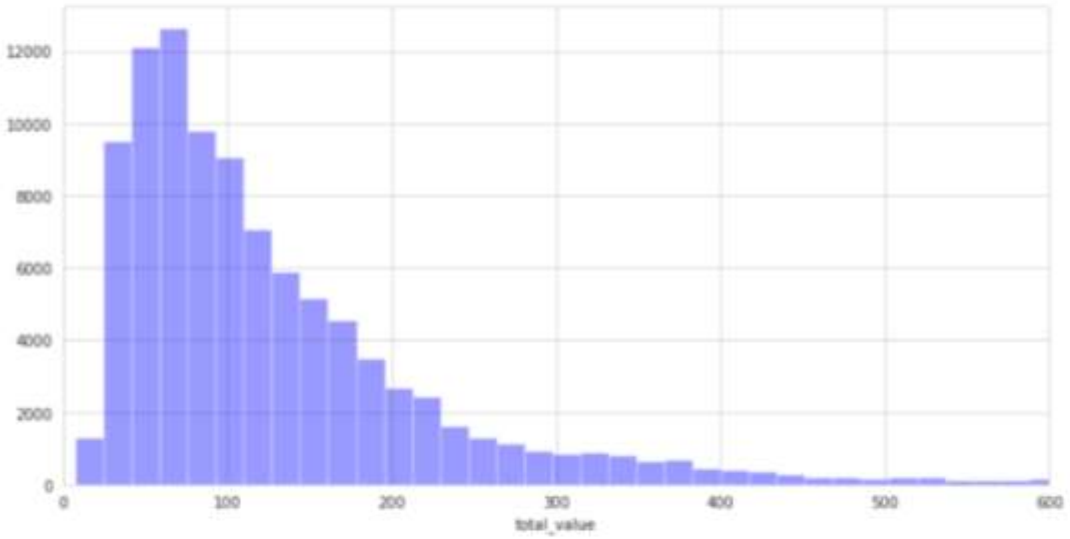


Figure 4: Average Order Value

Insight: As it was expected, we got right skewed histogram for total_value column - most of the times, people buy cheaply priced goods on Olist

c. Do some states tend to give harsher reviews compared to others?

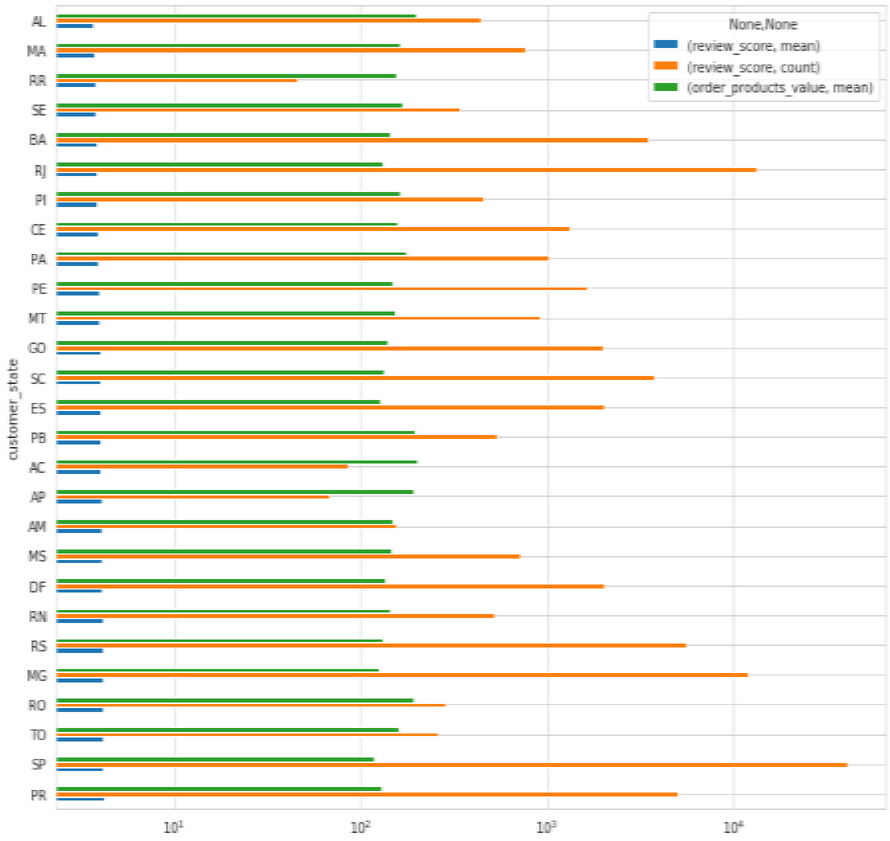


Figure 5: Reviews Distribution per State

Insight: Customer reviews don't change much from state to state (roughly, 0.6 point difference on average), but total volume of orders alongside with number count of reviews fluctuate quite dramatically. In short, some states inclined to shop more frequently than others but their experience doesn't differ much

d. What is the relation between time of the year v/s purchases made at Olist?

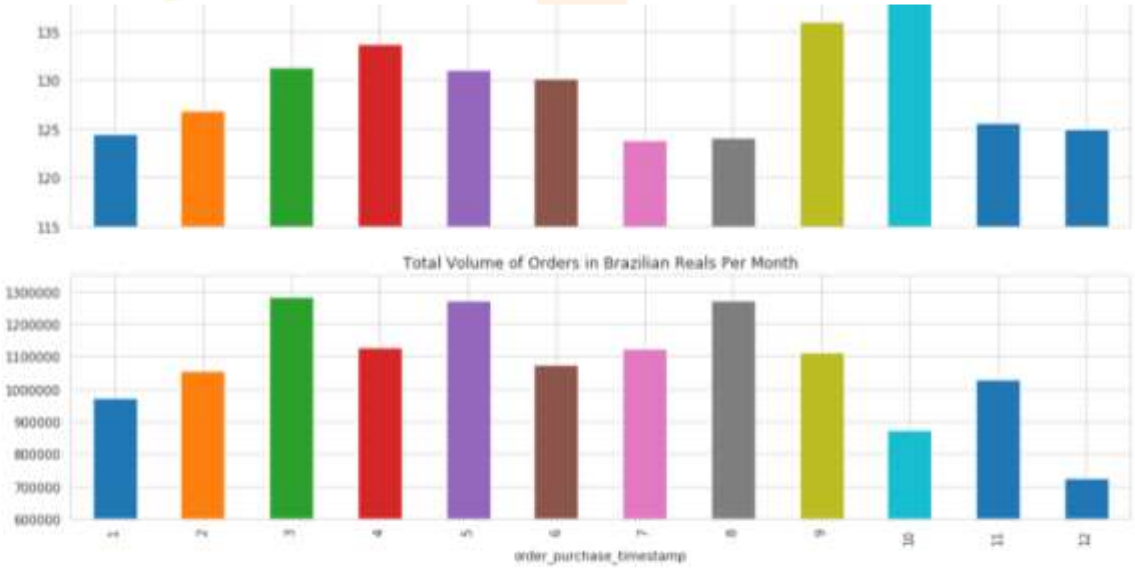


Figure 6: Order Volume per Month

Insight: Volume of goods ordered on Olist were lowest in December for some reason - Maybe people shop locally on holiday season!? While average prices of orders were relatively higher on Sept and Oct

e. Is there a particular hour, or a particular day of the week the customers tend to shop more?

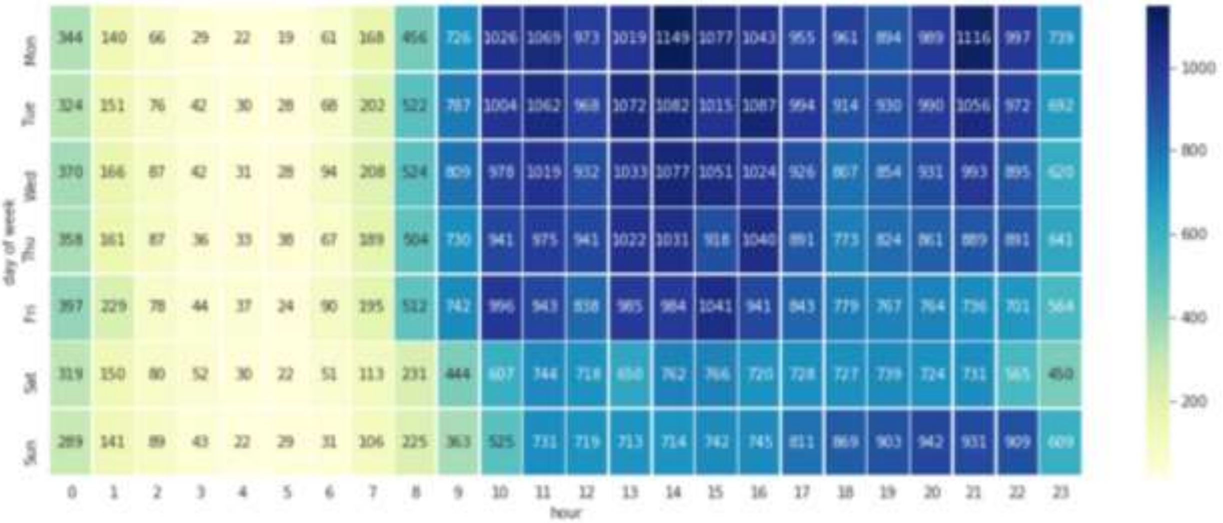


Figure 7: Distribution of Orders on Days of Week

Insight: Customers often purchase online on weekdays between 10 am and 4 pm. Intuitively, on Sunday nights (5-9pm), internet shoppers resume their purchasing habits following a relatively low Saturday. There are also unexpected increases around 8-9pm (Mon-Thu).

f. What is the most preferred hour for postal deliveries?

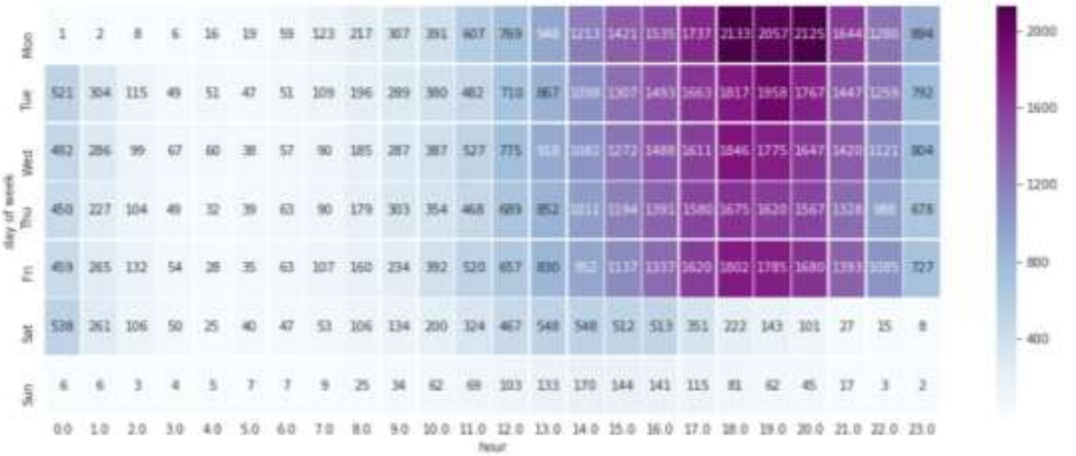


Figure 8: Preferred hour for Postal Deliveries

Insight: The story of delivery reveals that weekdays from 3-9pm are heaviest postal delivery truck operators

g. Delivery accuracy and Customer reviews should have a positive correlation, no?

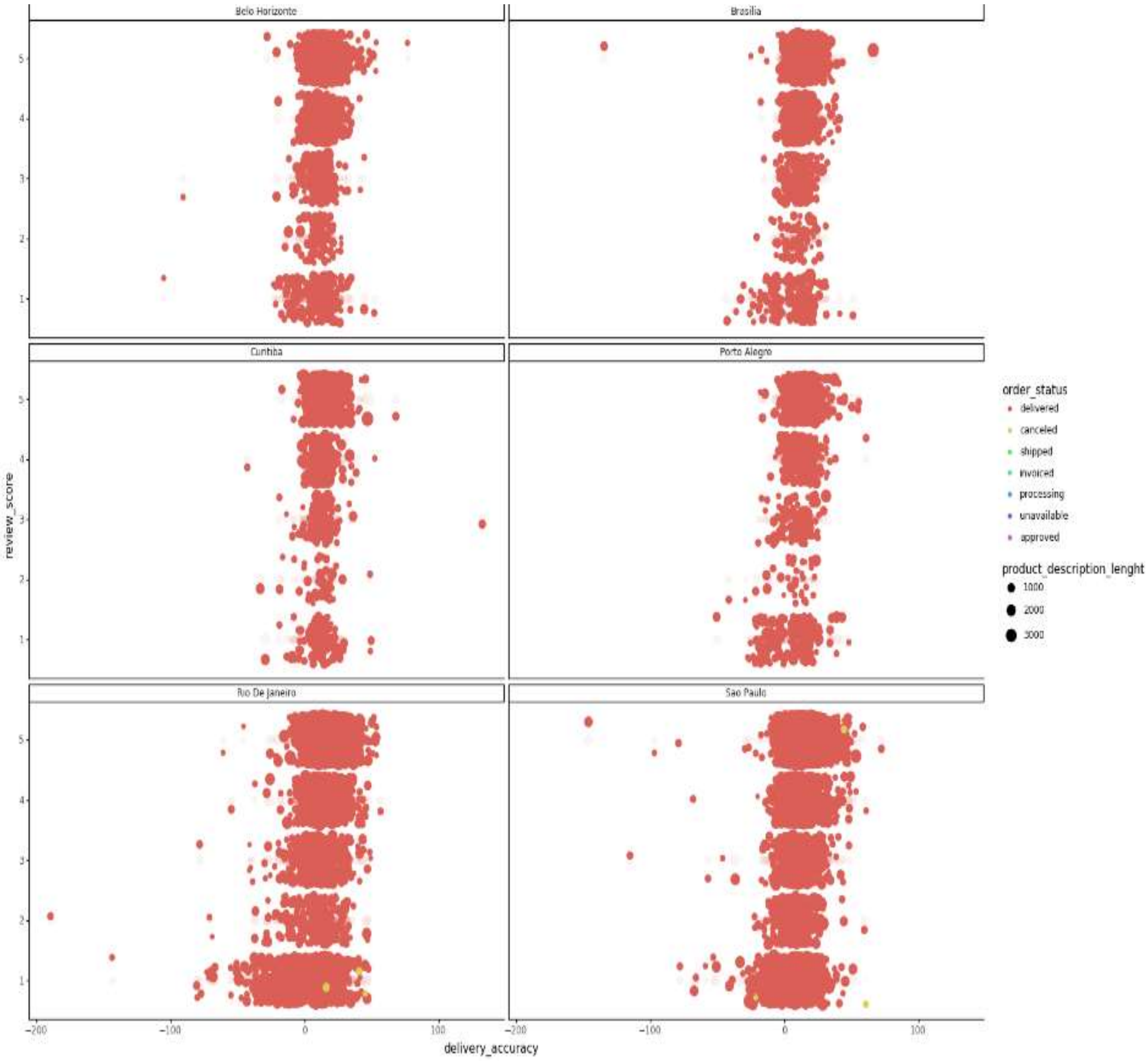


Figure 9: Correlation between Delivery Accuracy and Customer Reviews

Insight: We couldn't see a clear positive relationship between items being delivered earlier than promised and review score through our scatter plot.

h. What is the relationship of other important factors with review score?

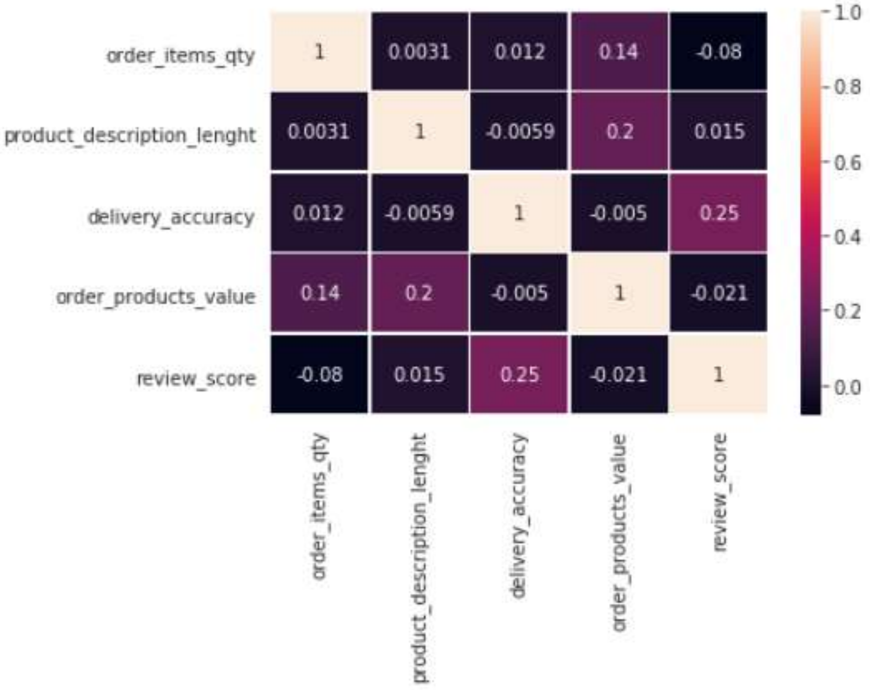


Figure 10: Correlation of Factors with Review Score
Insight: Most variables do have quite weak relationship with review score

i. Which product category has the lowest review score?

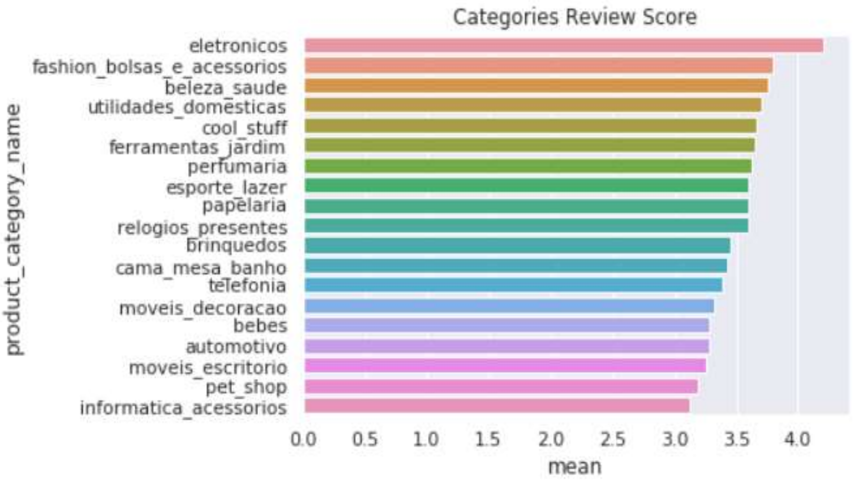


Figure 11: Categories Review Score

4.3 RFM Analysis

Recency-Frequency-Monetary analysis, often known as RFM analysis, is a popular marketing strategy that aids companies in segmenting and comprehending their client base. Recency, frequency, and monetary value are three crucial aspects of client behavior that must be examined.

Recency: This factor gauges how recently a customer has interacted with the brand or made a transaction. It displays the interval since the previous transaction or interaction. Customers who have interacted recently are regarded as being more responsive and engaged.

Frequency: The frequency of a customer's interactions with the firm over a given time period is measured by this dimension. It offers information about the degree of client loyalty and the likelihood of subsequent purchases. Customers who transact frequently are viewed as being more loyal.

Monetary Value: This dimension evaluates the monetary value of transactions or the amount of money spent by customers on purchases. It assists in locating valuable clients who make large financial contributions to the company.

Steps:

a. Creating 3 different datasets for calculating recency, frequency and monetary value.

b. Check if data is skewed first.

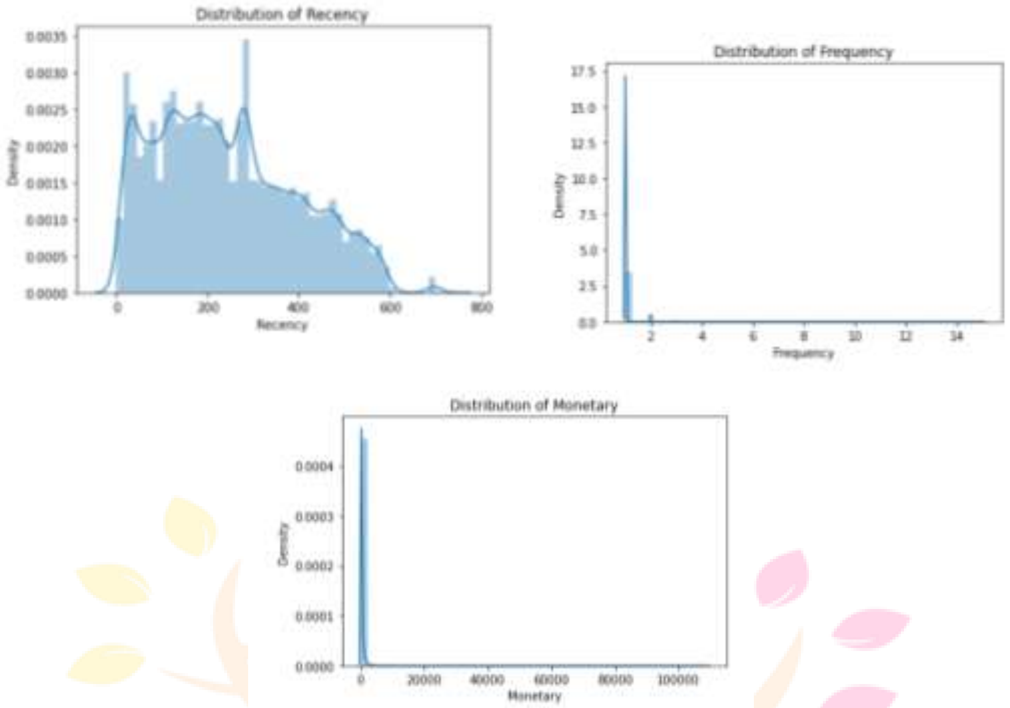


Figure 12: Recency Skew: 0.452, Frequency Skew: 10.990, Monetary Skew: 70.336

c. Log transforms Frequency data and Monetary data as they are highly skewed.

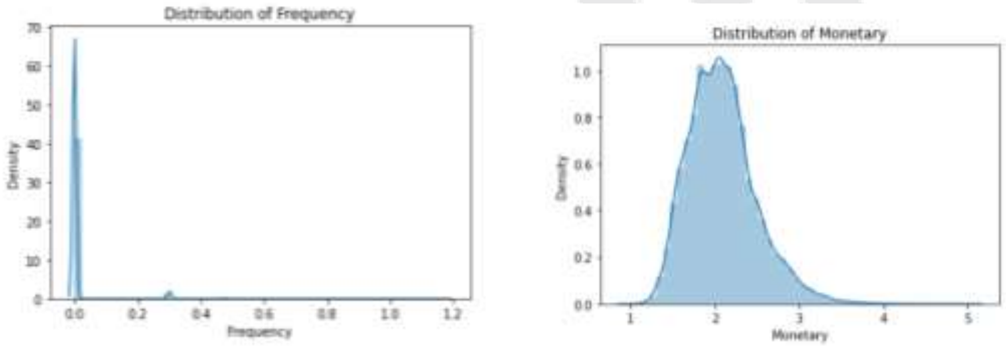


Figure 13: Frequency Skew: 6.068, Monetary Skew: 0.729

4.4 **Training K-Means Model**

a. Scaling the data with the Standard Scaler: Remove the mean and scale to unit variance to standardise features.

b. Using *sklearn.cluster.kmeans* to cluster the customers into groups

c. Determining the ideal number of clusters using the Elbow technique

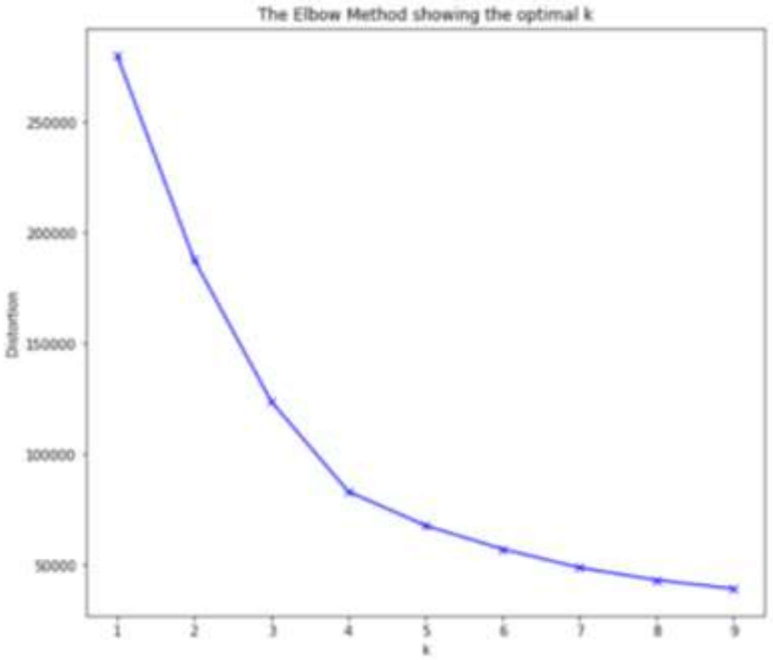
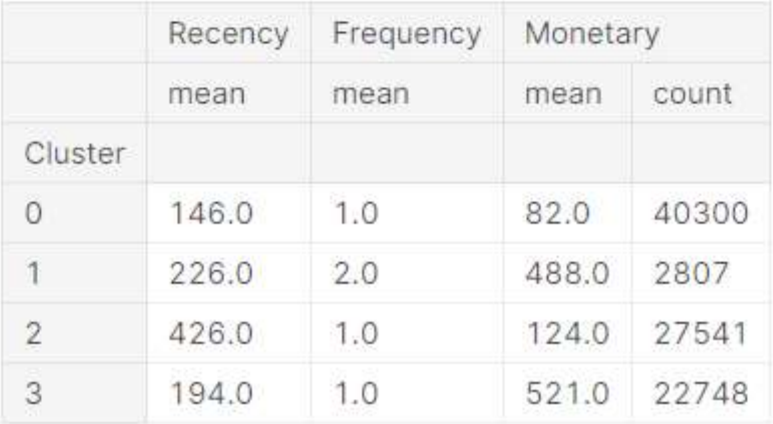


Figure 14: Elbow Method to obtain optimal k

d. Training the model on 4 clusters

5. EXPERIMENTAL RESULTS

Table 1: Recency, Frequency and Monetary Value Comparison of Clusters



	Recency	Frequency	Monetary	
	mean	mean	mean	count
Cluster				
0	146.0	1.0	82.0	40300
1	226.0	2.0	488.0	2807
2	426.0	1.0	124.0	27541
3	194.0	1.0	521.0	22748

- **Cluster 0** : This cluster can be interpreted as (*Relatively New Comers/Low spenders*) as their spending is the smallest among all clusters but have the lowest recency. This cluster is quite huge. So, some marketing effort could be advisable towards it in order to increase its monetary value.
- **Cluster 1** : can be considered as our best one (*Loyal customers*) since although customers in this cluster haven't on average ordered in a while, their frequency of orders is greater than all other clusters and the value of their orders is quite high.
- **Cluster 2** : (*Lost/Low spenders*) is the name we can give to this cluster. Customers in this cluster seem to have one of the highest recency and their monetary value is low. This cluster shouldn't be the focus of marketing effort.
- **Cluster 3** : This is our second best (*Big spenders*) cluster as its recency comes second and the average amount spent is largest. This cluster is also second in terms of number of customers.

6. CONCLUSION AND FUTURE WORK

In order to optimize logistics and supply chain management and increase customer satisfaction, we segmented customers in a Brazilian market using K-means clustering and RFM analysis in this research report. Relatively New Comers/Low Spenders, Loyal Customers, Lost/Low Spenders, and Big Spenders were the four unique consumer groupings that we were able to properly identify through the investigation. These segments offer insightful information on consumer behavior and can direct the market's strategic decision-making procedures. We were able to group clients based on their buying habits using K-means clustering, which gave the market the ability to customize logistics and supply chain strategies depending on the particular requirements and preferences of each segment.Big spenders, who have the potential to have a significant impact on revenue and should receive particular consideration in the form of individualized services and limited-time offers, were identified thanks to RFM analysis.

Our research has ramifications that go beyond supply chain and logistics management alone. Businesses may increase overall customer happiness, encourage client retention, and ultimately improve their bottom line by effectively targeting each consumer segment with specialized techniques.Future research in this field has a great deal of potential to improve supply chain management and logistics in online marketplaces. First off, broadening the scope of the analysis to take into account more demographic and consumer data could lead to a better knowledge of customer preferences and wants. Incorporating other data sources, such as location information or social media sentiment analysis, may potentially provide insightful information for targeted marketing and delivery optimisation.

Furthermore, Real-time data and predictive analytics can also be combined to improve supply chain management and promote proactive decision-making. Businesses may predict client requests, optimize inventory levels, and expedite last-mile delivery operations by utilizing technology like machine learning and artificial intelligence, which leads to higher customer satisfaction and lower costs. Additionally, researching how new technologies like blockchain and the Internet of Things (IoT) are affecting logistics and supply chain management may yield insightful results.

7. DECLARATION

9.1 Author contributions

All the authors participated equally in the analysis of the component.

9.2 Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

9.3 Availability of data and materials

The interested reader can contact the authors to access the analysis files.

9.4 Competing interests

The authors declare that they have no competing interests.

9.5 Acknowledgements

Not Applicable

8. REFERENCES

[1] SangJun Ahn, Mohammed Sadiq Altaf , SangUk Hanb , and Mohamed Al-Husseinc "Application of machine learning approach for logistics cost estimation in panelized construction." 2017 Modular and Offsite Construction Summit & the 2nd International Symposium on Industrialized Construction Technology, Shanghai, China(2017).

[2] Su Bu. "Logistics engineering optimization based on machine learning and artificial intelligence technology." Journal of Intelligent & Fuzzy Systems 40 (2021), 2505–2516 DOI:10.3233/JIFS-189244

[3] Sachin Gupta, Anurag Saxena. "Classification of Operational and Financial Variables Affecting the Bullwhip Effect in Indian Sectors: A Machine Learning Approach." Recent Patents on Computer Science (2019).

[4] Dino Knolla, Marco Prüglmeierb, Gunther Reinharta. "Predicting Future Inbound Logistics Processes Using Machine Learning." Changeable, Agile, Reconfi gurable & Virtual Production Conference 2016 , Procedia CIRP 52 (2016) 145 – 150.

[5] Nasution, A.A., Matondang, N. and Ishak, A. 2022. Inventory Optimization Model Design with Machine Learning Approach in Feed Mill Company. Jurnal Sistem Teknik Industri. 24, 2 (Jul. 2022), 254-272. DOI:https://doi.org/10.32734/jsti.v24i2.8637.

[6] Min, Hokey. “Artificial intelligence in supply chain management: theory and applications.” International Journal of Logistics Research and Applications Vol. 13 (2010): 13 - 39.

[7[Nesrin Ada, Yigit Kazancoglu Muruvvet Deniz Sezer ,Cigdem Ede-Senturk ,Idil Ozer and Mangey Ram "Analyzing Barriers of Circular Food Supply Chains and Proposing Industry 4.0 Solutions." Sustainability (2021),,13, 6812.https://doi.org/10.3390/su13126812Ac

[8] Pascal Wichmann, Alexandra Brintrup, Simon Baker, Philip Woodall &Duncan McFarlane. "Extracting supply chain maps from news articles using deep neural networks."International Journal of Production Research, 2020: https://doi.org/10.1080/00207543.2020.1720925 .

[9] S. Ton. "Research on Supply Chain Risk Assessment Based on Support Vector Machines." Economic Survey (2014)

[10] Gopal K. Kanji, Alfred Wong.. "Business Excellence model for supply chain management." Total Quality Management & Business Excellence (1999) :OL. 10, NO. 8, 1999, 1147-1168.

[11] Nesma mahmoud Taher1, Doaa Elzanfaly, Shaimaa Salama1"Investigation in Customer Value Segmentation Quality under Different Preprocessing Types of RFM Attributes." Int. J. Recent Contributions Eng. Sci. IT (2016).

[12] Sardjoeni Moedjiono, Yosianus Robertus Isak, Aries Kusdaryono. "Customer loyalty prediction in multimedia Service Provider Company with K-Means segmentation and C4.5 algorithm." 2016 International Conference on Informatics and Computing (ICIC) (2016).

[13] Peiman Alipour Sarvari, Alp Ustundag, Hidayet Takci ."Performance evaluation of differentcustomer segmentation approaches based on RFM and demographics analysis", Kybernetes(2016), Vol. 45Iss 7pp.1129-1157:document:http://dx.doi.org/10.1108/K-07-2015-0180

[14] Rachid Ait daoud, Abdellah Amine, Bouikhalene Belaid, Rachid Lbibb. "Combining RFM model and clustering techniques for customer value analysis of a company selling online." 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA) (2015).

[15] Tian J, Gao M., “Building Logistics Cost Forecast Based on Improved Simulated Annealing Neural Network”, Intelligent Computation Technology and Automation, 3:914-917, 2009

[16] Weston J, Watkins C., “Multi-class support vector machines”, Department of Computer Science, Royal Holloway, University of London, 1998.

[17] S. Sarkar, and S. Kumar, "A behavioral experiment on inventory management with supply chain disruption", Int. J. Prod. Econ., vol. 169, pp. 169-178, 2015.

[18] J.W. Hamister, and N.C. Suresh, "The impact of pricing policy on sales variability in a supermarket retail context", Int. J. Prod. Econ., vol. 111, pp. 441-455, 2008.

[19] K.B. Praveen, P. Kumar, J. Prateek, G. Pragathi. “Inventory Management using Machine Learning,” International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 6, pp. 866 – 871, 2020.

[20] K. Namira, H. Labriji, and E.H.B. Lahmar. “Decision Support Tool for Dynamic Inventory Management using Machine Learning, Time Series and Combinatorial Optimization,” Internasional Workshop of Information Sciences and Advanced Technologies: Edge Big Data – AI – IoT (ISAT 2021), Procedia Computer Science, vol. 198, pp. 423–428, 2022.

[21] Juan Liao1, Aman Bin Jantan1, Yunfei Ruan, Changmin Zhou. “Multi-behavior RFM Model Based on ImprovedSOM Neural Network Algorithm for CustomerSegmentation” in IEEE Access, vol. 10, pp. 122501-122512, 2022, doi: 10.1109/ACCESS.2022.3223361.

[22] Z. Tabaei, and M. Fathian, “Developing W-RFM Model for Customer Value: An Electronic Retailing Case Study”, IEEE,2011, p.305-306.

[23] M. Kim, J. Eun Park, A.J. Dubinsky, and S. Chaiy, “Frequency of CRM implementation activities: a customer-centric view”, Journal of Services Marketing, 2012m p. 84-85. https://doi.org/10.1108/08876041211215248

[24] M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh, “Estimating customer lifetime value based on RFM analysis of customer purchase behavior”, Elsevier, 2011, p. 60-62.

[25] Yousef Amer, Sang-Heon Lee. Azeem Ashraf, Lee Luong.”Optimizing order fulfillment using design for six sigma and fuzzy logic”,International Journal ofManagement Science and Engineering Management, 3 (2), 83–99.

[26] Jiju Antony, Rahul Swarnkar, Maneesh Kumar, Manoj Kumar Tiwari. Design of synchronized supply chain: a genetic algorithm based six sigma constrained approach.International Journal of Systems and Management, 2 (2), 120–141

[27]Coussement, K., Van den Bossche, F.A.M. and De Bock, K.W. (2014), “Data accuracy’s impact onsegmentation performance: benchmarking RFM analysis, logistic regression, and decisiontrees”, Journal of Business Research, Vol. 67 No. 1, pp. 2751-2758, doi: 10.1016/j.jbusres.2012.09.024.

[28] Dursun, A. and Caber, M. (2016), “Using data mining techniques for profiling profitable hotelcustomers: an application of RFM analysis”, Tourism Management Perspectives, Vol. 18,May, pp. 153-160, available at: http://doi.org/10.1016/j.tmp.2016.03.001

[29] Hu, Y.H. and Yeh, T.W. (2014), “Discovering valuable frequent patterns based on RFM analysiswithout customer identification information”, Knowledge-Based Systems, Vol. 61 No. 3,pp. 76-88, available at: http://doi.org/10.1016/j.knosys.2014.02.009

[30] Mannila, H., Toivonen, H. and Verkamo, A.I. (1994), "Efficient algorithms for discoveringassociation rules", AAAI Workshop on Knowledge Discovery in Databases KDD94, Vol. 118Nos 1-4, pp. 181-192, available at: http://ukpmc.ac.uk/abstract/CIT/21659.

[31] Yu, H.-H., Chen, C.-H. and Tseng, V.S. (2011), "Mining emerging patterns from time series datawith time gap constraint", International Journal of Innovative Computing, Information andControl, Vol. 7 No. 9, pp. 5515-5528.

[32] Awad, E. M. 1996. Building expert systems: principles, procedures, and applications, St. Paul, MN: West Publishing Company.

[33] Carbonneau, R., Laframboise, K. and Vahidov, R. 2008." Application of machine learning techniques for supply chain demand forecasting". European Journal of Operational Research, 184(3): 1140–1154.

[34] Dorigo, M., Maniezzo, V. and Colorni, A. 1996. "The ant systems: optimization by a colony of cooperative agents". IEEE Transactions in Man, Machine and Cybernetics Part B, 26(1): 29–41

[35] Gjerdrum, J., Shah, N. and Papageorgiou, L. G. 2001."A combined optimization and agent-based approach to supply chain modeling and performance assessment". Production Planning and Control, 12(1): 81–88.

[36] Lau, H. C.W., Pang, W. K. and Wong, C. W.Y. 2002. "Methodology for monitoring supply chain performance: a fuzzy logic approach. Logistics Information Systems", 15(4): 271–280.

[37] Nissen, M. E. and Sengupta, K. 2006. "Incorporating software agents into supply chains: experimental investigation with a procurement task". MIS Quarterly, 30(1): 145–166.

[38] Pawlak, Z. 1989. "Knowledge, reasoning and classification – a rough set perspective". Bulletin of the European Association for Theoretical Computer Science, 38: 199–210.

[39] Santos, E. Jr., Zhang, F. and Li, P. B. 2003." Intra-organizational logistics management through multi-agent systems". Electronic Commerce Research, 3: 337–364.

[40] Yandra, Y. 2007. "An integration of multi-objective genetic algorithm and fuzzy logic for optimization of agro-industry supply chain design". Proceedings of the 51st annual meeting of the International Society for the System Sciences. August5–102007. pp.1–15. Tokyo