Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol.4 No.4 (2025) pp: 83-88
P-ISSN: 2963-9298, e-ISSN: 2963-914X

K-Means Clustering Analysis For Identifying Product Purchase Patterns Based On Country On E-Commerce Platforms

Ryan Arya Pramudya[1*], Vinsent Brilian Adiguna[2]
[1]Management, Faculty of Economics and Business, Universitas 17 Agustus 1945 Semarang
[2]Digital Business, Faculty of Economics and Business, Universitas 17 Agustus 1945 Semarang
[1]ryanarya-pramudya@untagsmg.ac.id, [2]vinsentbrilian@untagsmg.ac.id

Abstract

E-commerce sites get a lot of transaction data from people in different countries who like different kinds of products. It is very important to know how people buy things based on their country and the type of product they buy in order to come up with better and more efficient marketing plans. This study seeks to discern product purchasing patterns by country through the application of the K-Means clustering algorithm on international e-commerce transaction data. This study utilized a dataset comprising 6,000 e-commerce transaction records, characterized by two primary variables: country and product category. Several methods were used in the preprocessing stage. For example, missing values were replaced to deal with missing data, nominal data was changed to numerical data to change categorical data into numerical data, and Z-transformation was used to normalize the data so that it was all on the same scale. We used the K-Means algorithm to group data into clusters with different k values, such as k=2, 5, 10, 15, 20, and 25. We then used the average within centroid distance metric and the elbow method to find the best number of clusters. The elbow method analysis showed that the best number of clusters was k=10, which showed a big drop in the average within centroid distance value. The ten clusters with algorithms K-Means that were made show very specific market segmentation, with each cluster having its own set of countries and product categories that are most popular.

Keywords: *K-Means, Clustering, E-commerce, Market Segmentation, Elbow Method*

1. Introduction

The evolution of information technology and the internet has revolutionized the shopping process for consumers by moving from conventional to online through e-commerce platforms. As per the latest figures, the volume of global e-commerce transactions is still rising significantly which generates a tremendous quantity of data [1]. Each transaction produces important information on consumer behavior, product likes, and buying patterns which can later on be analyzed for business purposes.

The huge and complicated data of e-commerce transactions necessitate the use of suitable analysis techniques to draw out useful information. One of the significant issues is to identify the purchasing behaviors of consumers from different countries who possess different characteristics and preferences [2]. The differences in geographical location, culture, and economy determine the categories of products that consumers in each country will be interested in.

Clustering is one of the most popular data mining techniques and it is applied to the data in order to create groups with similar features [3]. K-Means has become a favorite clustering algorithm, mainly because it is straightforward to implement, fast, and can work with large amounts of data [4]. It is the case that this algorithm has found its way through many domains such as segmentation of customers, analysis of market baskets, and identification of buying patterns.

K-Means has been frequently mentioned in the contexts of e-commerce data analysis in some previous studies. One paper that [5] has written introduced K-Means as a method for creating customer segments based on RFM (Recency, Frequency, Monetary) criteria, while another paper [6] employed K-Means as a tool for analyzing

shopping baskets. However, the number of studies that primarily focus on using categorical data to identify product purchasing patterns by country is limited.

Several previous studies have utilized K-Means for customer segmentation, RFM (Recency, Frequency, Monetary) analysis, and the optimization of recommendation systems. For example, Ramadhani et al (2023) utilized K-Means clustering to customize e-commerce recommendations, where as Hidayat et al (2023) employed it to examine fashion shopping trends. Even with these improvements, most current research still looks at numerical or transactional variables, like how often someone buys something and how much they spend. There aren't many studies that look at categorical attributes, like country and product category.

International e-commerce businesses can gain a strategic edge by knowing how people in different parts of the world buy things. It lets you make decisions based on data in areas like localizing marketing, managing product inventory, and expanding your business across borders. It also helps us learn more about the diversity of consumers around the world by showing us how culture and the economy affect how people use technology.

The issues mentioned above, this study intends to address. The first issue is about preprocessing categorical data for K-Means algorithm. Next, the second issue is about the optimal number of clusters for country-based purchasing patterns. Finally, the third issue is about the characteristics of each cluster that has been formed.

The objective of this study is to apply appropriate preprocessing procedure for e-commerce transaction data so that it can be analyzed using the K-Means algorithm. Furthermore, this study also attempts to determine the optimal number of clusters using the elbow method, identify product purchase patterns by country, and provide relevant business strategy recommendations based on the obtained cluster results.

This research enhances the academic discourse on data-driven market segmentation and provides practical insights for personalized marketing and product strategy optimization. By using clustering techniques, e-commerce businesses can learn more about their customers and be more flexible in how they respond to the changing global markets.

2. Research Methods

This look at uses a quantitative approach with an experimental method to apply the K-Means clustering algorithm to e-trade transaction information. The studies design follows the Knowledge Discovery in Databases (KDD) tiers, which consist of selection, preprocessing, transformation, facts mining, and interpretation [8].

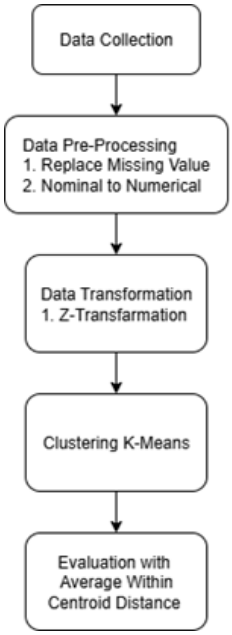

Figure 1. Research Methods

2.1. Data Collection

A overall of 6,000 e-commerce transaction records had been collected with variables: united states, containing 10 nations (Australia, Brazil, Canada, China, France, Germany, India, Japan, UK, USA), and often bought item class, containing Apparel, Books, Electronics, Home and Kitchen, Sports, and Toys.

DOI: https://doi.org/10.31004/riggs.v4i4.3296
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

2.2. Data Preprocessing

a. Replace Missing Values
   Handling missing values using the mode imputation method for express records. This method is crucial to ensure that no missing records affects the clustering outcomes.
b. Nominal to Numerical
   Transform specific facts into numerical statistics using encoding techniques. The country and category variables are transformed into numerical representations so they can be processed by using the K-Means algorithm, which requires numerical enter.
c. Data Normalization (Z-Transformation)
   Data normalization the usage of Z-score standardization to make sure that each variable has the same scale.

2.3. *K-Means Clustering*

K-Means clustering is implemented as an preliminary method because of its simplicity and velocity of execution on medium-sized datasets. K-Means is typically used for consumer segmentation in an e-trade context.

K-Means Clustering Parameters:

a. Number of clusters (K): 2, 5, 10, 15, 20, 25
b. Distance measure: Euclidean distance
c. Max iterations: 100

2.4. Clustering Evaluation

Average Within Centroid Distance (AWCD) This metric measures the average distance of every records point to its cluster centroid. A decrease AWCD value indicates a more compact and homogeneous cluster.

3. Results and Discussions

The dataset used consists of 6,000 e-commerce transaction facts with two categorical variables: usa and category. The facts indicates the distribution of transactions from 10 one of a kind countries (Canada, Brazil, China, Germany, France, UK, USA, India, Australia, Japan) with 6 product classes (Electronics, Apparel, Toys, Home.

Clustering turned into done with varying numbers of clusters from K=2 to K=25. Evaluation using the Average Within Centroid Distance (AWCD) metric yielded the following values:

Table 1. K-Means Cluster with Average within Centroid Distance

Number of Clusters (K)	Average Within Centroid Distance (AWCD)
2	14,878
5	11,792
10	6,960
15	5,384
20	4,755
25	4,129

Based on the table above, there may be a huge lower in the AWCD price from K=2 of the common inside centroid distance with a end result of 14,878 to K=10 which produces a median within centroid distance of 6,960.

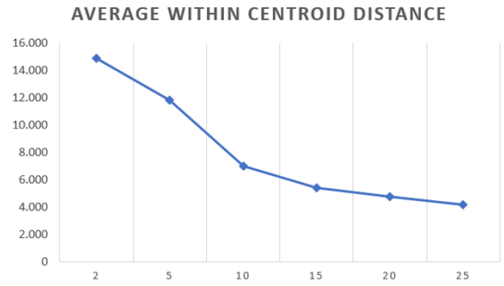

Figure 2. Elbow Method Analysis Results

DOI: https://doi.org/10.31004/riggs.v4i4.3296
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

The elbow method graph shows a clean "elbow" at K=10, indicating that including more than 10 clusters does now not offer a large improvement in clustering best. The most reliable quantity of clusters is K=10 based totally at the elbow method with an AWCD value of 6,960.

The clustering consequences with K=10 screen a very detailed market segmentation with specific purchasing styles for each u . S .-class aggregate. This better granularity offers deeper insight in comparison to segmentation with fewer clusters.

3.1. *Specific Country-Based Segmentation*

Nine out of ten clusters show the dominance of a particular u . S . (>eighty five%), indicating that geographical elements are very sturdy in figuring out shopping patterns. This is steady with studies [2] that determined a giant influence of local tradition and economic system on product choices. Only Cluster 10 shows a pass-border sample, indicating the life of a international area of interest phase for sure categories.

3.2. Clear Product Category Differentiation

Each cluster has a extraordinary dominant category, starting from Electronics (Clusters 1)

3.3. Specific Patterns in Developed Countries vs. Emerging Markets

a. **Developed Countries (USA, Japan, UK):** Focus on Electronics and premium categories (Clusters 1, 2, 3, 6)
b. **Emerging Markets (Brazil, India):** Focus on Home & Kitchen, Toys, and value products (Cluster 4, 5)
c. **Balanced Markets (Canada, Germany):** High diversification or a combination of smart home products (Clusters 8, 9)

3.4. Australia as a Unique Segment

Cluster 7, dominated via Australia with a focus on Sports.

3.5. Overview of Clustering Results

The dataset included 6.000 e-commerce transactions from 10 countries and 6 different types of products. The data was ready for K-Means numerical analysis after preprocessing, which included replacing missing values, encoding, and Z transformation. The clustering process with different K values (2,5,10,15,20,25) and the Average Within Centroid (AWCD) and the elbow method showed that K = 10 was the best choice, with AWCD = 6.960. This result shows that the clusters are both compact and easy to understand.

3.6. Interpretation of Cluster Patterns

Ten distinct clusters were generated, each exhibiting unique combinations of country and predominant product category. In most clusters, one country made up more than 85% of the total, which suggests that geography has a big effect on how people shop. For example, Cluster 1 showed that people in the US and Japan bought a lot of electronics, while Cluster 4 showed that people in India and Brazil liked home and kitchen items. The study says that national culture, income levels, and how digitally savvy consumers are the main factors that affect category preferences.

3.7. Comparative Analysis among Clusters

A comparative analysis reveals three primary market classifications : Developed nations, Emerging markets, and balanced economies.

Table 2. Comparative Analysis among Clusters

Country	Market classifications	Product
USA, UK, Japan	Developed nations	Technology – Oriented Product
Brazil, India	Emerging markets	Essential and family – oriented product
Germany, Canada	Balanced economies	Moderate interest in all categories

DOI: https://doi.org/10.31004/riggs.v4i4.3296
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

People in developed nations like the US, UK, and Japan really like technology-based products like electronics and smart home devices. Emerging markets like Brazil and India are mostly interested in basic family goods like toys and home goods. Balanced economies like Germany and Canada have a wide range of buying habits, with moderate interest in all categories.

3.8. Business Implications

The segmentation outcome gives e-commerce platforms useful information that they can use. Companies can make personalized recommendation systems by focusing on groups of people who live in the same area and act in the same way. For instance, send promotional bundles of electronics and digital accessories to Cluster 1 (Tech-oriented) and furniture and home tools to Cluster 4 (Home – oriented). Also, marketing teams can change their pricing and languange localization strategies to fit the demographic characteristics of each cluster.

3.9. Methodological Reflection

The K-Means algorithm worked well for medium-sized datasets and easy-to-understand segmentation. But it can't handle data that isn't spherical or has different types of data. Even though nominal-to-numeric encoding was used, some semantic connections between categorical values may have been lost. Future studies may integrate sophisticated clustering techniques, such as K-Prototypes or DBSCAN, alongside feature embedding to enhance clustering robustness and representation.

3.10. Limitations and Future Work

This study exclusively examines two categorical variables country and product category without incorporating temporal or behavioral attributes (e.g., frequency or expenditure). Future studies ought to integrate more comprehensive attributes and investigate temporal clustering to discern dynamic transformations in consumer behavior. Another option is to combine K-Means with recommendation systems that use machine learning to automatically divide up marketing groups.

4. Conclusion

This study effectively utilized the K-Means clustering algorithm to discern purchasing behavior patterns according to country and product category within e-commerce data. The preprocessing steps, which included filling in missing values, encoding categorical data, and normalizing Z-Scores, made sure that the data could be used for numerical analysis. The clustering process showed that the best number of clusters is K=10. This makes the clusters easy to understand and small enough to clearly show the differences between market segments in different countries. The findings indicated that geographical and cultural influences significantly affect consumer preferences. Countries like the US, UK, and Japan were more likely to buy electronics and high-end products, while countries like India and Brazil were more likely to buy toys and home and kitchen items. The segmentation showed that personalized marketing, better inventory management, and product recommendations based on location were all possible. From an academic standpoint, this research substantiates the relevance of unsupervised learning teachniques-especially clustering in the examination of behavioral diversity within global e-commerce. It builds on previous research by demonstrating that K-Means analysis can produce meaningful clusters from categorical data when it is properly preprocessed. The research adds to the growing body of work on using data to divide up markets and analyze customers. These results are very useful for making decisions and planning e-commerce strategies. Businesses can use clustering to make campaigns that are specific to each market segment's culture, economy, and preferences. Companies can use clustering to help them change their pricing strategies, advertising content, and product recommendations on the fly. Also adding these clustering models to Customer Relationship Management (CRM) systems lets you use adaptive segmentation that changes in real time as customer behavior changes. Based on the results of the clustering, e-commerce sites should : a. Add clustering analytics to CRM systems so that segmentation and targeting can happen all the time and in real time. b. Use K-Means clustering and predictive models together to better plan logistics, predictive models together to better plan logistics, predict product demand, and spot seasonal trends. c. Make marketing messages more personal by customizing the content, deals, and suggestions for each cluster based on its most important traits. d. Use marketing strategies that work in many languages and cultures to get more people interested in your brand around the world. By using these strategies, businesses can use clustering insights not just for marketing but also for planning their strategeis, coming up with new products, and improving the customer experience. This study has several limitations, even though it made some important contributions. The analyzed data comprised solely two categorical variables (country and product category), excluding purchase frequency, monetary value, or temporal patterns, subsequent research ought to integrate behavioral and temporal dimensions to examine the progression of customer trends. In addition, future research could test hybrid models like K-Prototypes, DBSCAN, or Gussian Mixture

DOI: https://doi.org/10.31004/riggs.v4i4.3296
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Models to work with data that is both linear and nonlinear. Another promising direction is to use machine learning – based recommendation system or deep clustering models to create automated, adaptive segmentation for personalized online shopping experiences.

Reference

[1] A Sharma, R., & Kumar, P. (2023). Digital Transformation in E-Commerce: Trends and Challenges in the Post-Pandemic Era. *International Journal of Electronic Commerce Studies*, *14*(2), 145–162.
[2] Chen, A., Wang, H., & Zhang, L. (2023). Cross-Cultural Consumer Behavior Analysis in Global E-Commerce Platforms. *Journal of Business Research*, *156*, 324–338.
[3] Jain, D., & Singh, S. (2023). A Comprehensive Review of Clustering Algorithms for Big Data Analytics. *Journal of King Saud University - Computer and Information Sciences*, *35*(4), 101–119.
[4] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, *9*(8), 1295. https://doi.org/10.3390/electronics9081295
[5] Ramadhani, P., Wibowo, A., & Prasasti, D. (2023). Customer Segmentation Using K-Means Clustering and RFM Analysis for E-Commerce Personalization Strategy. *Journal of Information Systems Engineering and Business Intelligence*, *9*(1), 85–94.
[6] Hidayat, N., Wardhani, S., & Rakhmawati, A. (2023). Implementation of K-Means Clustering for Customer Shopping Pattern Analysis in Fashion E-Commerce. *International Journal of Advanced Computer Science and Applications*, *14*(5), 234–242.
[7] Wang, L., Liu, Y., & Zhang, X. (2022). Big Data Analytics in E-Commerce: A Systematic Literature Review. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(3), 1044–1070.
[8] Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and Techniques* (4th ed.). Morgan Kaufmann.
[9] Kumar, S., & Sharma, D. (2020). Enhanced K-Means Clustering Algorithm with Improved Initial Centroids. *International Journal of Computer Sciences and Engineering*, *8*(6), 150–157.
[10] Murtagh, F., & Contreras, P. (2022). Algorithms for Hierarchical Clustering: An Overview and Recent Advances. *WIREs Data Mining and Knowledge Discovery*, *12*(1), e1430.
[11] Prasetyo, Y., Santoso, B., & Kurniawan, I. (2023). Comparative Analysis of Clustering Algorithms for E-Commerce Customer Segmentation: K-Means, DBSCAN, and Hierarchical Clustering. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *10*(2), 345–356.
[12] Ardiansyah, R., Handayani, F., & Safitri, N. (2023). Market Basket Analysis and Customer Clustering for Product Recommendation in Indonesian E-Commerce. *Journal of Big Data*, *10*, 78.
[13] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, *9*(8), 1295. https://doi.org/10.3390/electronics9081295
[14] Murtagh, F., & Contreras, P. (2011). Algorithms for hierarchical clustering: an overview. *WIREs Data Mining and Knowledge Discovery*, *2*(1), 86–97. https://doi.org/10.1002/widm.53
[15] Gelbrich, K., Müller, S., & Westjohn, S. (2023). Global trends in consumer behavior. In *Cross-Cultural Consumer Behavior* (pp. 54–67). Edward Elgar Publishing. https://doi.org/10.4337/9781803923192.00009
[16] Li, Y., Zhang, R., & Jiang, D. (2022). Order-Picking Efficiency in E-Commerce Warehouses: A Literature Review. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(4), 1812–1830. https://doi.org/10.3390/jtaer17040091
[17] Han, J., Kamber, M., & Pei, J. (2012). Advanced Pattern Mining. In *Data Mining* (pp. 279–325). Elsevier. https://doi.org/10.1016/b978-0-12-381479-1.00007-1
[18] Rahadiyan, H. A. (2023). Segmentation of Mentoring Customer Characteristics Using the K-Means Method and Hierarchical Clustering for Customer Relationship Management (CRM). *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer Dan Teknologi Informasi*, *9*(1), 64. https://doi.org/10.24014/coreit.v9i1.21567
[19] E-commerce Recommendation Algorithm Based on Big Data Analysis and Genetic Fuzzy Clustering. (2023). *Financial Engineering and Risk Management*, *6*(9). https://doi.org/10.23977/ferm.2023.060904
[20] Anam, S., Fitriah, Z., Hidayat, N., & Maulana, M. H. A. A. (2023). Classification Model for Diabetes Mellitus Diagnosis based on K-Means Clustering Algorithm Optimized with Bat Algorithm. *International Journal of Advanced Computer Science and Applications*, *14*(1). https://doi.org/10.14569/ijacsa.2023.0140172
[21] Wang, Y., Krishna Saraswat, S., & Elyasi Komari, I. (2023). Big data analysis using a parallel ensemble clustering architecture and an unsupervised feature selection approach. *Journal of King Saud University - Computer and Information Sciences*, *35*(1), 270–282. https://doi.org/10.1016/j.jksuci.2022.11.016
[22] Sharma, V. (2025). Law and Emerging Technologies in Global Commerce: Regulatory Challenges in the Digital Transformation Era. *Journal of International Commercial Law and Technology*, *6*(1), 549–554. https://doi.org/10.61336/jiclt/25-01-51
[23] Kaewpradit, T. (2025). *Optimizing Retail Strategy: A Data-Driven Approach to Customer Segmentation Using RFM Analysis and K-Means Clustering*. Elsevier BV. https://doi.org/10.2139/ssrn.5238097
[24] KumarSihag, V., & Kumar, S. (2013). Graph based Text Document Clustering by Detecting Initial Centroids for k-Means. *International Journal of Computer Applications*, *62*(19), 1–4. https://doi.org/10.5120/10185-5005

DOI: https://doi.org/10.31004/riggs.v4i4.3296
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)