PRODUCT CLUSTERING ANALYSIS ON THE MARKETPLACE USING K-MEANS APPROACH (CASE STUDY: SHOPEE)

Maria Arista Ulfa[1*], Selo[2], Muslikhin Hidayat[3]

[1]Master in Systems Engineering, Universitas Gadjah Mada
[2]Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada
[3]Department of Chemical Engineering, Universitas Gadjah Mada
*Corresspondence: mariarista3@gmail.com

Abstract

The business world has experienced a paradigm shift towards a more modern concept. Many business processes are carried out through the internet or commonly known as e-commerce, by utilizing a platform known as Marketplace. One of the marketplaces that are quite well-known and in great demand in Indonesia is Shopee. The high online shopping activity in the current marketplace indirectly encourages business actors to understand the online market. However, one of the obstacles that are quite often faced by sellers, especially new sellers who are starting to enter the digital realm, is the emergence of confusion in the selection of products to be sold due to a lack of information regarding the demand for what products are in demand in the market.

The process of searching for information related to the demand for products of interest is carried out through clustering analysis to find out the groups of products that are of interest to those that are less attractive to the public. The data used is product data from 6 categories in the Shopee market which was taken using web scraping techniques. The clustering processes used the K-means approach by determining the number of K and the optimal center point through the calculation of Sum Square Error (SSE) by looking at the elbow graph. The final results show the optimal number of K clusters that are different in each category, namely in category women's clothing, men's clothing, and electronics are at K=4 then for products in the category of Muslim fashion, care & beauty and household appliances are at K=3. Based on the validation results using the Davies Bouldin Index, values were obtained in6 categories, namely 0.391, 0.438, 0.414, 0.357, 0.387, and 0.377, which means that the cluster structure and the level of information formed in each category using the K-Means method is quite good.

History:

Received: September 18, 2021
Accepted: October 27, 2021
First published online:
October 31, 2021

Keywords:

Davies Bouldin Index
K-Means Clustering
Shopee Product
Web Scraping

1. Introduction

The need for the internet today seems to have become a primary need that is very influential in human life. Based on the latest report data from We Are Social and Hootsuite 2020, it is stated that the Indonesian population aged 16-64 years who use the internet reaches 175.4 million or 64% of the total population of Indonesia (We are social and Hootsuite, 2020). This condition is undoubtedly based on the role of the internet, which can provide convenience in various sectors of life without being limited by space and time, one of which is the business sector.

Business activities that are carried out through the internet are usually known as e-commerce or electronic commerce. Based on a report by Google, Temasek, and Bain, the E-commerce market is the most prominent digital economy sector in Southeast Asia based on total transaction value or gross merchandise value (GMV). The e-commerce market is projected to quadruple from US$38.2 billion in 2019 to US$153 billion in 2025.

The Majority of the market comes from Indonesia, where its value is also expected to increase from US$20.9 billion in 2019 to US$82 billion in 2025 (Google, Temasek, and Bain Company, 2020). An e-commerce business that involves a third party is usually called a marketplace. The marketplace is a platform that acts as a liaison between sellers and buyers. The existence of a marketplace is considered very profitable for business actors. This is because the marketplace provides a system for business people as their place of business so that the separate system will not be bothered.

In addition, it helps to intensively promote the product since it is sold virtually (Yustiani and Yunanto, 2017). One of the marketplaces that are currently chart-topping in Indonesia is Shopee. Shopee is an application that is engaged in buying and selling online. iPrice data in the third quarter of 2020 (July-September) shows that Shopee is the largest Indonesian marketplace based on the number of monthly visitors of 96.5 million people and ranks first for the most downloaded mobile marketplace application on the AppStore and PlayStore (iPrice, 2020).

Seeing that online shopping activities are increasing, especially during this coronavirus disease pandemic, it is inevitable to encourage business actors to understand the online market, especially for new business actors or new sellers who will enter the online business world. However, one of the obstacles that sellers, especially new sellers frequently face, is the emergence of confusion in selecting products to be sold due to a lack of information regarding the demand for what products are in high demand in the market. This condition leads them to poor product management, which ended up in the sales turnover.

Therefore, the researchers tried to analyze the products sold online in the marketplace, especially Shopee. The main focus of this research is to extract the product data through the Shopee website using a web scraping technique which is then stored in a file with CSV format. After that, K-Means clustering is analyzed to obtain cluster information of most in-demand products to less attractive consumers. This cluster can also analyze which areas sell the most products to the price range of products sold based on the level of products of interest to the public. Therefore, it can be used as a reference in decision-making when managing an online business to maximize sales going forward.

The potential of clustering is knowing the structure in the data, which is then represented into several groups with the same data resemblance to produce new information. The K-Means clustering method is quite popular and widely used because of its simplicity, easy implementation, handling large-scale data and outliers, and fast computation

time (Kantardzic, 2019). Therefore, in this study, the K-Means method is used in classifying products in the Shopee marketplace.

However, the K-Means method has not determined the optimal K value selection and is also quite sensitive to the initial centroid. If the initial centroid given is crummy, the clustering results will also be unsatisfactory (Primandana, 2019). Determination of the number of K and the optimal center point in the clustering process is overcome by looking at the elbow graph to calculate Sum Square Error (SSE). In the final stage, the cluster results are validated using the Davies Bouldin Index (DBI) to determine how well the resulting cluster is.

2. Methodology

A. Data Collection

In this study, the data used is product data taken through the Shopee marketplace site from April to May 2021. The product categories used are the categories of women's clothing, men's clothing, Muslim fashion, electronics, care and beauty and household appliances with several product variables i.e., URL, Product Name, Price, Product Category, Store Name, Rating, Sold, Stock and City. Data collection is done using the web scraping technique with the python programming language. Python is an interpretative and dynamic programming language used for numerical calculations, analysis, and data visualization (Budiharto, 2018).

Web scraping is a process of retrieving semi-structured documents from the internet, which are generally web pages in markup languages such as HTML or XHTML. Web scraping focuses on getting data employing retrieval and automatic extraction. The Use of web scraping in the business world is usually done to create precise business concepts in helping to promote and develop a business, such as knowing market conditions. So, that it is possible to lead the market, facilitate promotions, and retrieve important competitor information (Thomas et al., 2019). The data collection process is shown in Figure 1.

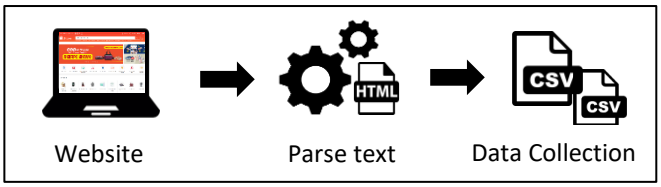


Figure 1. Process of scraping Shopee product

At the time of data collection, several products were not included in the category understudy, so filtering was necessary. The following details the amount of data in each category:

Table 1. Number of products

No.	Category	Before Filtering	After Filtering
1.	Women's clothing	35,000 data	34,739 data
2.	Men's clothing	20,000 data	19,575 data
3.	Muslim fashion	27,500 data	27,352 data
4.	Electronics	37,500 data	36,667 data
5.	Care and beauty	32,500 data	31,002 data
6.	Household appliances	35,000 data	34,785 data
Total		187,500 data	184,120 data

B. Initial Data Processing

The data taken is not entirely in an ideal condition for processing. Initial processing is needed to process the data format into the system or commonly referred to as pre-processing. The initial data processing process can be shown in Figure 2.

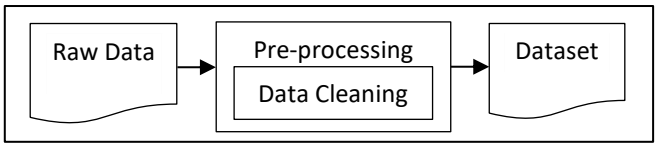


Figure 2. Pre-processing

C. K-Means Clustering

Clustering is one of the techniques in data mining that is unsupervised learning, or there are no attributes that can be used to guide the learning process so that all input is treated the same (Vulandari, 2017). K-Means is a clustering method included in partition grouping because it separates data into K separate parts. The partition approach means grouping data based on partitions where data are separated to be grouped so that similar data are in the same section and dissimilar data are in different parts. The purpose of clustering is to minimize variation within a cluster. The K-means method generally groups data based on distance and only works on numeric attributes. The stages in K-Means clustering are (Kantardzic, 2019):

1. Determine the number of k clusters.
2. Allocate data according to the number of clusters.
3. Calculate the centroid of each data in each cluster.
4. Allocate n data to each of the nearest centroids.
5. Repeat step 3 if there is still data transfer from one cluster to another.

Calculation of distance with Euclidean Distance can be calculated using equation 1.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

where:
d= Distance
x= Data in the i-th cluster
y= Centroid in cluster i

Calculation of the value of the new center point or centroid can be calculated using equation 2.

$$C_i = \left(\frac{1}{N_i}\right)\sum x_i \quad (2)$$

where:
C= new centroids
x= Data in the i-th cluster
N= The amount of data in the i-th cluster

The Elbow method is used to determine the number of clusters or groups in a data set. The elbow method provides an idea in choosing the best cluster by observing the value in each K cluster that is improved (Maheswari, 2019). The value observed and compared is the value of the Sum Square Error (SSE), where the results of the SSE calculation show several K values that have decreased drastically with an angle forming an indentation. In addition, the SSE calculation in the K-Means method can also help determine the most optimum centroid (Nainggolan et al., 2018). The SSE formula can be calculated using equation 3.

$$SSE = \sum_{i=1}^{n} (d)^2 \quad (3)$$

where:
SSE = Sum Square Error
d = Distance between data and centroid

D. Davies Bouldin Index

Clustering validation is a way of evaluating the results of a cluster. In the application of K-Means, there is no certainty in finding the accuracy of the clusters formed. Therefore, validation is needed in measuring how well or accurate the resulting clusters are. Davies Bouldin Index (DBI) is one method of validity in clustering. The DBI approach maximizes inter-cluster distance and minimizes intra-cluster distance (Singh, 2020). Calculation with DBI is based on the value of cohesion and separation. Cohesion in the grouping is the proximity of the data to the centroid in a cluster, while separation is the distance between the centroids and each cluster.

The cohesion value can be found using the Sum of Square Within Cluster (SSW) equation, and the separation value can be found using the Sum of Square Between Cluster (SSB) equation. The DBI value is in the interval (0, 1) where the minimum value indicates the optimal cluster. Based on research that carried out on 11 internal validation methods in 5 aspects to determine the best validation, it was found that DBI is in the 2nd place where DBI is not sensitive to the boundary points and can get the right number of cluster when clustering greater than 2 (Xiao et al., 2017).
DBI formula can be calculated using equations 4 to 7.

$$SSW = \frac{1}{m_i} \sum_{j=1}^{m_i} d(x_j, c_i) \quad (4)$$

where:
SSW = Sum of Square Within Cluster
m = Number of data in cluster i
d = Distance between data x and centroid c

$$SSB = d(c_i, c_j) \quad (5)$$

where:
SSB = Sum of Square Between Cluster
d = Distance between centroid i and centroid j

$$R_{ij} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (6)$$

where:
R = Ratio of cluster i and cluster j

$$DBI = \frac{1}{k} \sum_{j=1}^{k} max(R_{i,j}) \quad (7)$$

where:
DBI = Davies Bouldin Index
k = Total of clusters

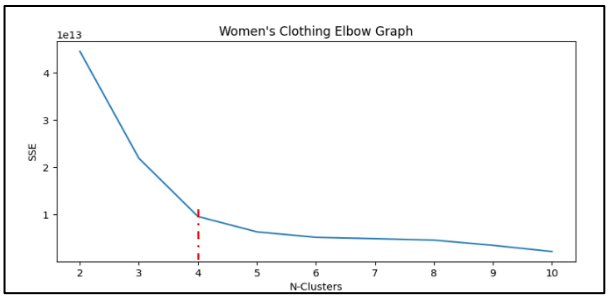
3. Results & Discussion

A. Clustering Implementation

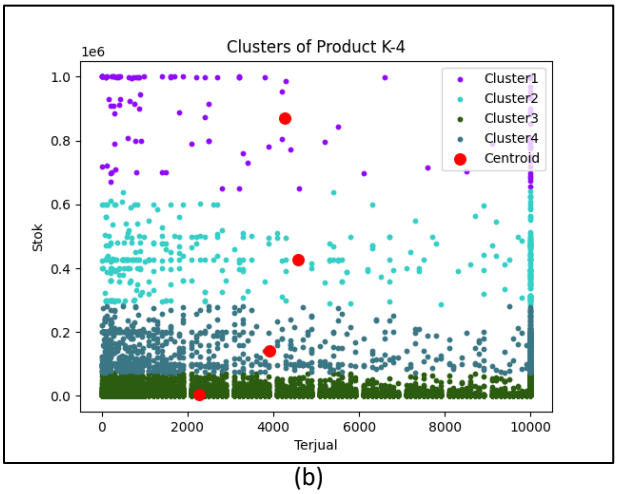
The application of data processing uses the K-Means method to determine the cluster of each product. The optimal center point or centroid is determined by randomly taking n center points and then iterating five times where the number of n centers observed is from 200 to 1000 center points. The total observed center points are 1000 to 5000 center points on each product category. In addition, K cluster calculations are also carried out to determine the optimal cluster where the observed k value is from 2 to 10 clusters.

1. Women's Clothing Product Category

Figure 3(a) shows the optimal cluster value for women's clothing products in 4 clusters with an SSE value of 9610355038184.62, while Figure 3(b) illustrates a scatter graph of clustering results with optimal centroid values, namely [4263.67826087, 870562.96521739], [4576.65469613, 427936.52762431], [2279.58722765, 3013.27882795], [3909.73454914, 140422.90780142] and with a total data of 34739 data which is divided into Cluster 1=115 data, Cluster 2=362 data, Cluster 3=33275 data, and Cluster 4=987 data.



(a)

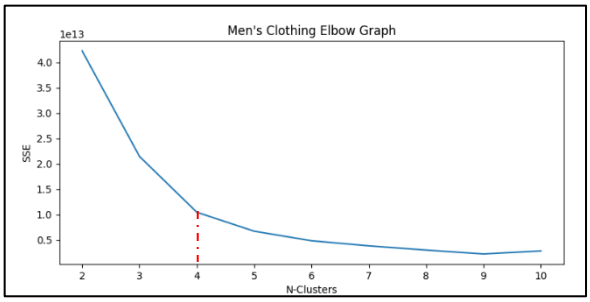


(b)

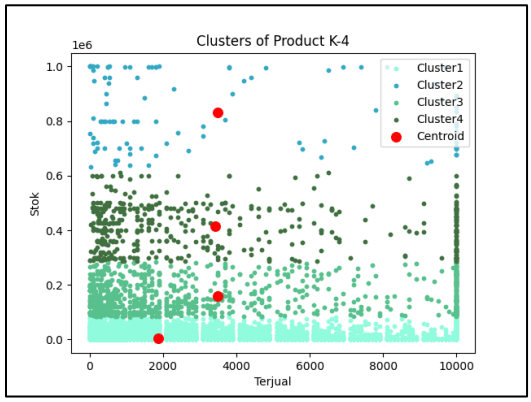
Figure 3. Women's clothing elbow graph (a) and Women's clothing scatter graph (b)

2. Men's Clothing Product Category

Figure 4(a) shows the optimal cluster value for men's clothing products in 4 clusters with n SSE value of 10435606667927,611, while Figure 4(b) depicts a scatter graph of the results of clustering with optimal centroid values, namely [1863.55820485, 555468138767], [3496.5, 832426.76], [3490.41367323, 159682.19698725], [3428.30309735, 413673.18141593] and with a total data of 19575 data which is divided into cluster 1=18162 data, cluster 2=100 data, cluster 3=861 data and cluster 4=452 data.



(a)

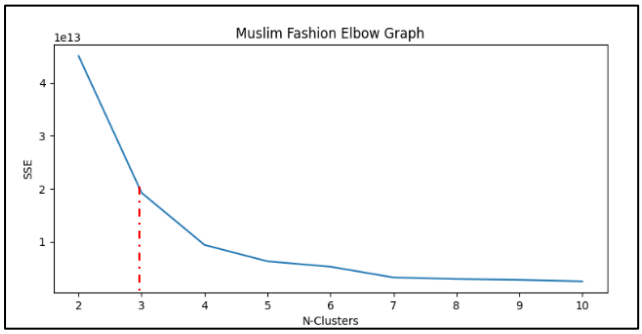


(b)

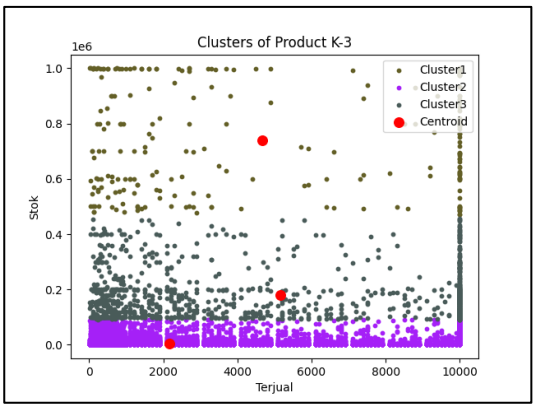
Figure 4. Men's clothing elbow graph (a) and Men's clothing scatter graph (b)

3. Muslim Fashion Product Category

Figure 5(a) shows the optimal cluster value for Muslim fashion products in 3 clusters with an SSE value of 19241826323912.76. In contrast, Figure 5(b) depicts a scatter graph of the results of clustering with optimal centroid values, namely [4663.51101322, 739603.72687225], [2156.96594178, 345579411652], [5160.48669202, 179731.44296578] and with a total of 27352 data which is divided into cluster 1=227 data, Cluster2=26073, and Cluster 3=1052 data.



(a)

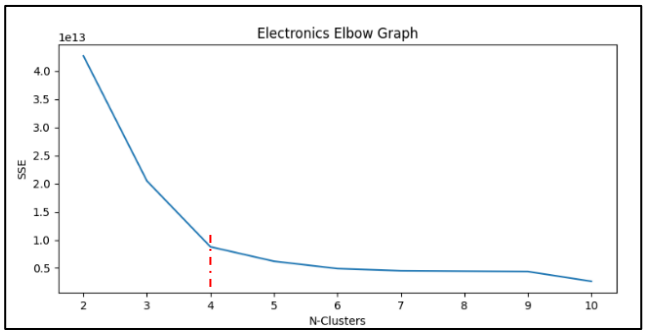


(b)

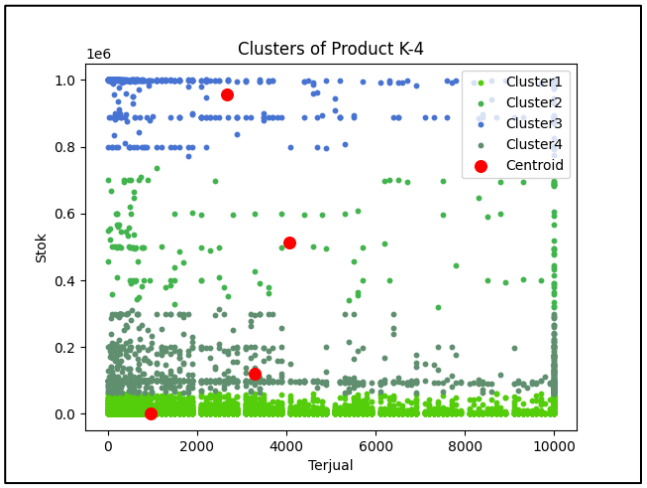
Figure 5. Muslim fashion elbow graph (a) and Muslim fashion scatter graph (b)

4. Electronics Product Category

In Figure 6(a) shows the optimal cluster value of electronic products is in 4 clusters with an SSE value of 8765871368900912 while Figure 6(b) depicts scatter graph of clustering results with optimal centroid values [967.4197403, 2275.323818], [4065.65333, 513847.78666], [2666,10238095, 956813.93571428], [3282,97259136, 121454.84551495] and with a total data of 36667 data which is divided into cluster 1= 34894 data, Cluster 2 = 150 data, Cluster 3 = 420 data and Cluster 4 = 1203 data.



(a)

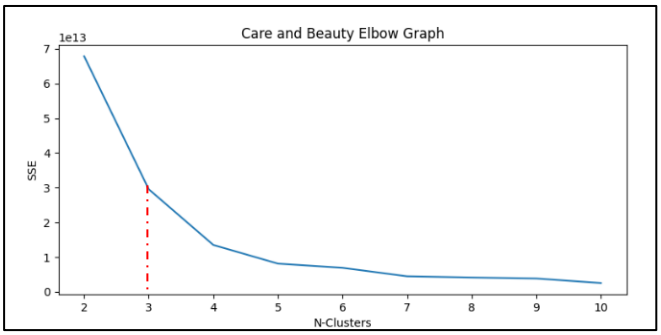


(b)

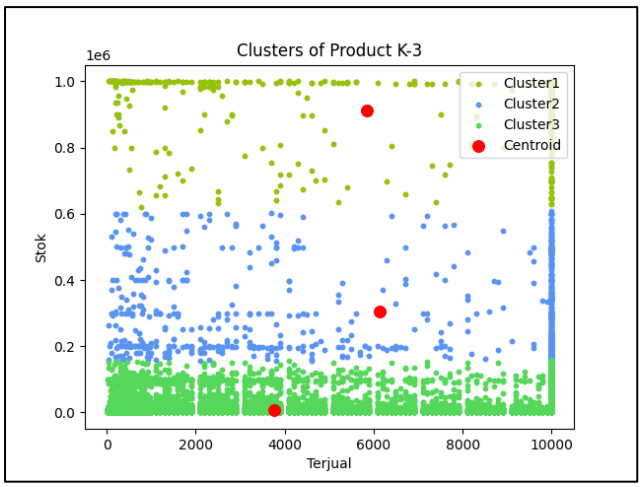
Figure 6. Electronics elbow graph (a) and Electronics scatter graph (b)

5. Care and Beauty Product Category

Figure 7(a) shows the optimal cluster value for care and beauty products in 3 clusters with an SSE value of 29548396052128.58, while Figure 7(b) illustrates a scatter graph of clustering results with optimal centroid values, namely [5853.451790, 912704.903581], [6132.73054755, 305827.69596542], [3766.22875271, 7140.71073635] and with a total of 31002 data which is divided into cluster 1 = 363 data, Cluster 2 = 694 data and Cluster 3 = 29945 data.



(a)



(b)

Figure 7. Care and beauty elbow graph (a) and Care and beauty scatter graph (b)

6. Household Appliances Product Category

Figure 8(a) shows the optimal cluster value of household appliances products in 3 clusters with an SSE value of 43042310431608.69, while Figure 8(b) illustrates a scatter graph of clustering results with optimal centroid values, as follows: [6096.33757062, 935488.25], [6991.70987039, 307449.9501496], [3558.93841084, 9441.6876096] and with a total of 34785 data which is divided into cluster 1 = 708 data, cluster 2 = 1003 data and cluster 3 = 33074 data.

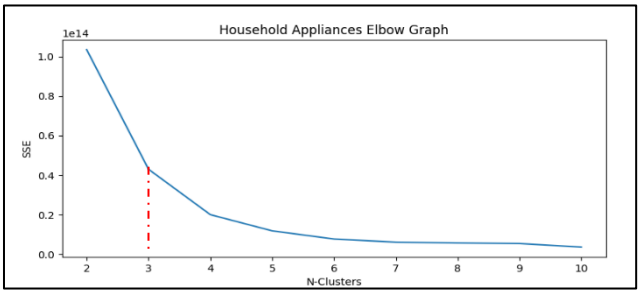


Figure 8(a). Household appliances elbow graph

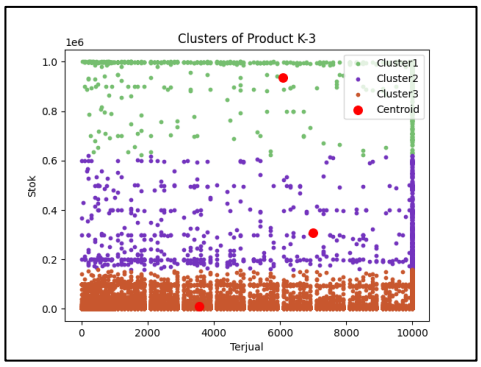


Figure 8(b) Household appliances scatter graph

Based on the results of the SSE calculation from each cluster in each product category with different N centroid values (1000, 2000, 3000, 4000, and 5000). It was found that the 5000 centroid value showed a smaller SSE value than the other random centroids. This is because the selection of the random centroid value increases according to the specified iteration, namely five iterations, which causes a large opportunity for the value with the smallest error to be included. To produce a final value that is smaller than the number of other centroid samples. Following the smallest SSE value obtained, this value is used to determine the optimal cluster and centroid, represented through the elbow graph where the optimal K value (2 to 10) is obtained on a graph that decreases by forming an elbow or bend. It can also be stated that determining the optimal centroid using N iterations takes time since the larger the number of datasets used, the more random centroids are needed.

B. Clustering Validation

Clustering validation using DBI is based on the degree of similarity in the same cluster and differences in different clusters. The validation results for each product category can be seen in Figure 9 and Table 2.

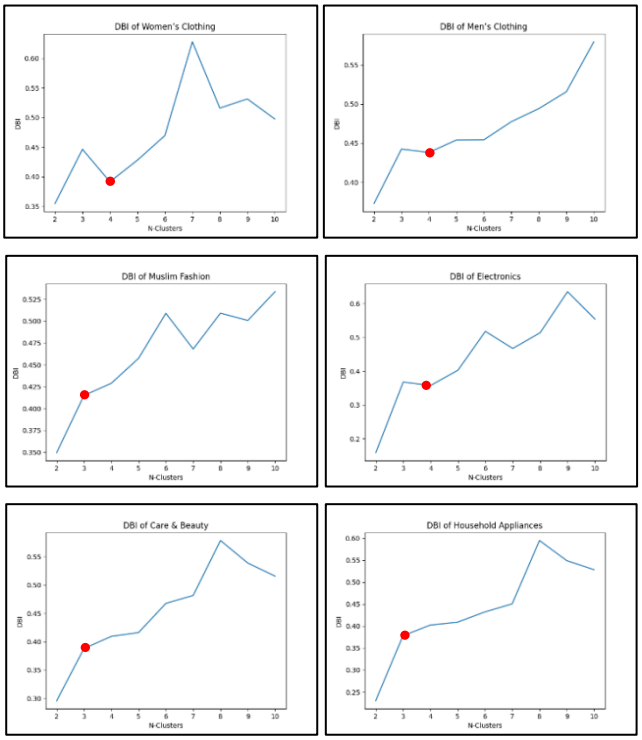


Figure 9. DBI Graph of each product category

Table 2. DBI Value of each product category

No	Category	Optimal Cluster	DBI
1	Women's clothing	4	0.3914523886
2	Men's clothing	4	0.4380935970
3	Muslim fashion	3	0.4149511797
4	Electronics	4	0.3574902070
5	Care and beauty	3	0.3879534769
6	Household appliances	3	0.3779623549

Based on the DBI principle, which is an internal cluster validation, the number of good clusters is indicated by a smaller DBI value (non-negative>=0) or a minimum intra-cluster value, which means that each object in the cluster has a similar level of characteristics. From Table 2, it can be said that the results of the cluster structure formed in each product category using the K-Means method are pretty good.

C. Clustering Results Analysis

Analysis of the results was carried out to figure out the information contained in the cluster after the data processing was carried. Therefore, it can be used as a reference in conducting online sales activities.

1. Women's Clothing Product Category

Based on the scatter graph in Figure 3(b), the information obtained for the 4 clusters are:

Table 3. Women's clothing product cluster information

Cluster	Product	Average Price	Store Rating
Most in demand	Trousers	Rp.35,644	4.8
	Dress	Rp.96,500	
	Tops	Rp.44,965	
	Pajamas	Rp.47,094	
Quite in demand	Fabric	Rp.20,358	4.8
	Outerwear	Rp.53,097	
	Skirt	Rp.35,578	
	Batik & kebaya	Rp.120,233	
In demand	Jeans	Rp.62,962	4.8
	Jumpsuit & overall	Rp.65,834	
	Suits	Rp.90,588	
Less in demand	Underwear	Rp.23,249	4.8
	Maternity clothes	Rp.64,150	
	Costume	Rp.236,176	

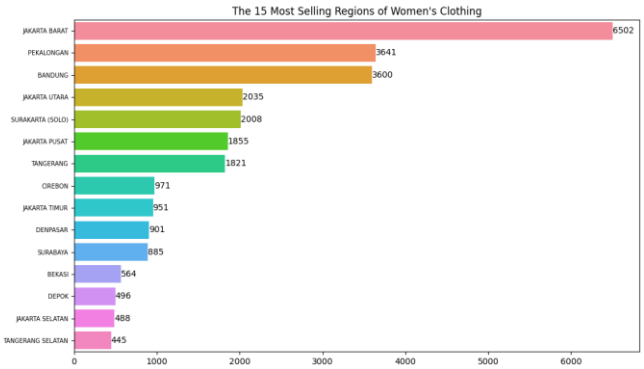


Figure 10. The 15 most selling regions of women's clothing

2. Men's Clothing Product Category

Based on the scatter graph in Figure 4(b), the information obtained for the 4 clusters are:

Table 4. Men's clothing product cluster information

Cluster	Product	Average Price	Store Rating
Most in demand	Shorts	Rp.55,976	4.8
	Batik	Rp.103,336	
	Tops	Rp.62,727	
Quite in demand	Long pants	Rp.67,724	4.7
	Jeans	Rp.127,992	
In demand	Outerwear	Rp.94,571	4.7
	Underwear	Rp.26,728	
Less in demand	Pajamas	Rp.82,771	4.7

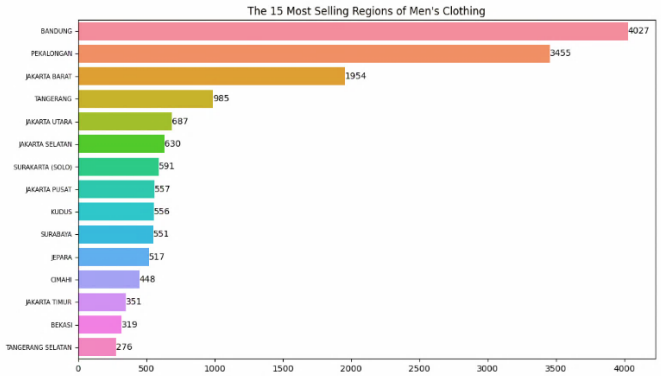


Figure 11. The 15 most selling regions of men's clothing

3. Muslim Fashion Product Category

Based on the scatter graph in Figure 5(b), the information obtained for the 3 clusters are:

Table 5. Muslim fashion product cluster information

Cluster	Product	Average Price	Store Rating
Most in demand	Outerwear	Rp.107,084	4.8
	Women's prayer clothes	Rp.125,080	
	Muslim men's bottoms	Rp.86,563	
	Muslim dress	Rp.107,710	
	Muslim women's bottoms	Rp.60,628	
In demand	Veil	Rp.14,714	4.8
	Muslim accessories	Rp.5,317	
	Muslim men's top	Rp.103,962	
	Prayer equipment	Rp.37,203	
Less in demand	Kids muslim clothes	Rp.20,460	4.8
	Muslim women's top	Rp.22,219	

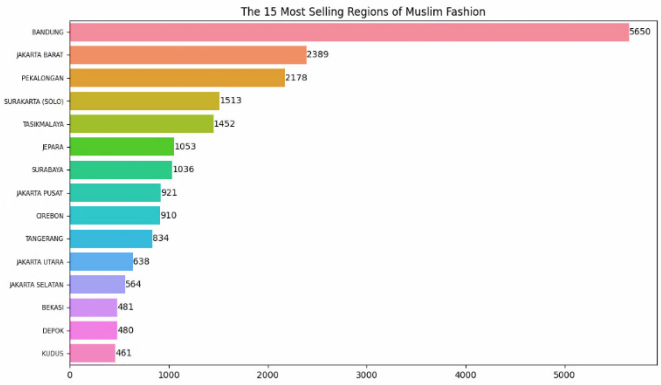


Figure 12. The 15 most selling regions of Muslim fashion

4. Electronics Product Category

Based on the scatter graph in Figure 6(b), the information obtained for the 4 clusters are:

Table 6. Electronics product cluster information

Cluster	Product	Average Price	Store Rating
Most in demand	Kitchen utensil	Rp.317,322	4.8
	TV	Rp.372,048	
	Clothes care	Rp.226,133	
	Refrigeration	Rp.143,611	
Quite in demand	Cables and spare parts	Rp.3,807	4.9
	Lighting	Rp.32,581	
	Electricity	Rp.29,025	
	Speakers and microphone	Rp.63,862	
In demand	Video games	Rp.198,218	4.8
	Surveillance camera	Rp.150,000	
	Phone	Rp.90,588	
	Media player	Rp.75,000	
Less in demand	Vaporizer	Rp.10,412	4.8
	Outdoor device	Rp.20,089	
	Vacuum cleaner	Rp.103,775	

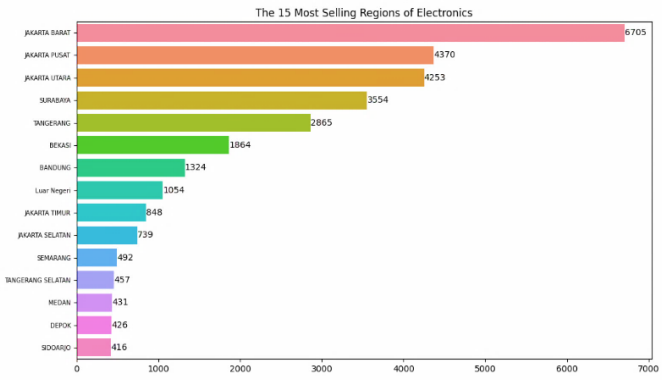


Figure 13. The 15 most selling regions of electronics

5. Care and Beauty Product Category

Based on the scatter graph in Figure 7(b), the information obtained for the 3 clusters are:

Table 7. Care and beauty product cluster information

Cluster	Product	Average Price	Store Rating
Most in demand	Face treatment	Rp.65,686	4.9
	Body care	Rp.49,028	
	Other beauty treatments	Rp.58,444	
	Lip cosmetics	Rp.72,839	
	Hair care	Rp.46,466	
In demand	Perfume	Rp.31,058	4.9
	Nail care	Rp.11,302	
	Beauty tools	Rp.6,696	
	Eye cosmetics	Rp.5,523	
	Beauty pack	Rp.167,608	
Less in demand	Men's care	Rp.74,945	4.8
	Facial cosmetics	Rp.30,555	
	Hair tools	Rp.10,616	

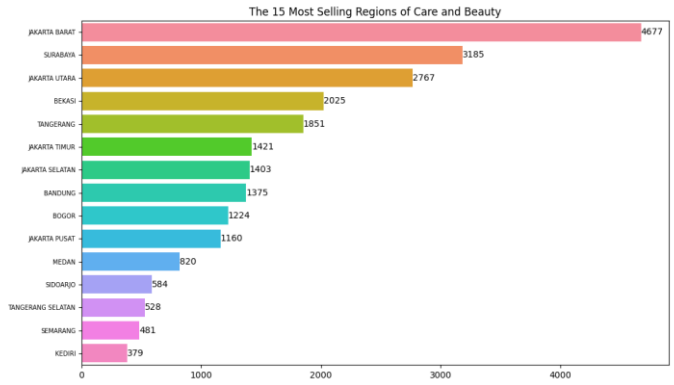


Figure 14. The 15 most selling regions of care and beauty

6. Household Appliances Product Category

Based on the scatter graph in Figure 8(b), the information obtained for the 3 clusters are:

Table 8. Household appliances product cluster information

Cluster	Product	Average Price	Store Rating
Most in demand	Religious	Rp.54,540	4.8
	Umbrella	Rp.67,838	
	Cooking ware	Rp.31,031	
	Bathroom equipment	Rp.23,755	
	Home care tools	Rp.18,648	
In demand	Decoration	Rp.24,749	4.8
	Storage	Rp.4,941	
	Dining room equipment	Rp.5,537	
	Garden	Rp.9,828	
	Kitchen utensils and accessories	Rp.11,559	
Less in demand	Carpentry tools	Rp.6,374	4.9
	Cleaning and laundry	Rp.40,363	
	Furniture	Rp.46,576	
	Bedroom equipment	Rp.14,486	

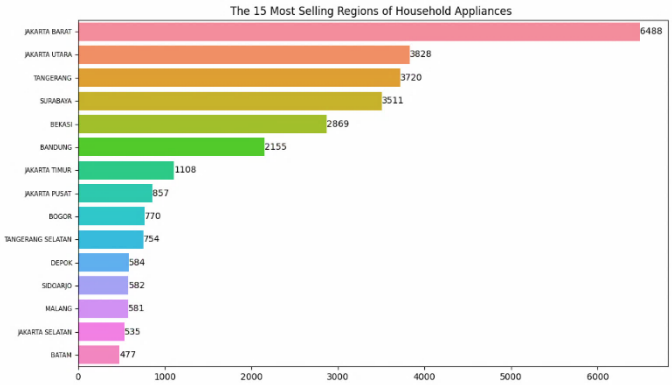


Figure 15. The 15 most selling regions of household appliances

4. Conclusion

Product clustering using the K-Means approach with the amount of data processed as much as 184,120. It can show that product clusters are of interest to less desirable by the public based on the similarity in each data. The optimal number of K clusters is different in each category. Women's clothing, men's clothing and electronics are at K=4 then Muslim fashion, care and beauty, and household appliance are at K=3.

The information generated from each cluster, such as the average price of products and the region with the most sellers, can help sellers, especially new sellers, start engaging in digital business activities as a basic reference in making decisions when managing online sales. Product grouping using the K-Means method by determining the number of K and the optimal center point has provided cluster structure results with a fairly good level of information.

It should be noted that the quality of clustering results is very dependent on the method used, where each method has its approach in its execution to allow differences in the final results of clustering. In addition, this study found that determining the optimal center point with the N iteration technique takes a long time, so it is more suitable when used on small datasets.

References

We Are Social And Hootsuite, accessed December 27, 2020, *Digital 2020 Indonesia : All The Data, Trends, And Insights You Need To Help You Understand How People Use The Internet, Mobile, Social Media, And Ecommerce*, https://datareportal.com/reports/digital-2020-indonesia?rq=indonesia.

Google, Temasek and Bain Company, accessed December 29, 2020, *E-conomy SEA 2019 Report*, https://www.temasek.com.sg/en/news-and-views/subscribe/google-temasek-e-conomy-sea-2019

Yustiani, R. dan Yunanto, R. 2017. *Peran Marketplace Sebagai Alternatif Bisnis di Era Teknologi Informasi*, Jurnal Ilmiah Komputer dan Informatika Vol. 6 No.2 : 2089-9033.

iPrice, accessed December 30, 2020, *Peta E-Commerce Indonesia*,https://iprice.co.id/insights/mapofecommerce/

Budiharto, W., 2018, *Pemrograman Python untuk Ilmu Komputer dan Teknik,* Andi Offset : Yogyakarta.

Thomas, D.M. and Mathur, S., 2019, *Data Analysis by Web Scraping Using Python*, Proceedings of the Third International conference on Electronics, Communication and Aerospace Technology (ICECA) : 10.1109/ICECA.2019.8822022.

Vulandari, R.T., 2017, *Data Mining : Teori dan Aplikasi Rapid Miner*, Yogyakarta : Gava Media.

Kantardzic, M., 2019, *Data Mining Concepts, Models, Methods and Algorithms (Third Edition),* IEEE Press : United States of America.

Primandana, A., Adinugroho, S. dan Dewi, C., 2019, *Optimasi Penentuan Centroid pada Algoritme K-Means Menggunakan Algoritme Pillar (Studi Kasus: Penyandang Masalah Kesejahteraan Sosial di Provinsi Jawa Timur),* Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol.3 No.11 : 2548-964X.

Maheswari, K., 2019, *Finding Best Possible Number Of Clusters using K-Means Algorithm*, International Journal of Engineering and Advanced Technology (IJEAT) : 10.35940.A1119.1291S419

Nainggolan, R. and Lumbantoruan, G., 2018, *Optimasi Performa Cluster K-Means Menggunakan Sum of Squared Error (SSE)*, Jurnal Manajemen Informatika dan Komputerisasi Akuntansi Vo.2 No.2.

Singh., A.K., Mittal, S., Srivastava, Y.V. and Malhotra, P., 2020, *Clustering Evaluation by Davies Bouldin Index (DBI) In Cereal Data Using K-Means Clustering*, Proceedings of the Fourth International conference on Computing Methodologies and Communication (ICCMC) : 10.1109/iccmc48092.2020.iccmc-00057.

Xiao, J., Lu, J. and Li, X., 2017, *Davies Bouldin Index based Hierarchical Initialization K-means*. Intelligent Data Analysis : 10.3233.