Indonesian Journal of Data and Science

Volume 6 Issue 3
ISSN 2715-9930
https://doi.org/10.56705/ijodas.v6i3.316

Research Article

Application Of K-Means Clustering Algorithm to Identify the Best-Selling Digital Printing Services

Ana Fatahali Ramadhan [1,*]; Sudin Saepudin [2]; Carti Irawan [3]; Mupaat [4]
[1] Universitas Nusa Putra, Sukabumi 43152, Jawa Barat, Indonesia, anafatahali@gmail.com
[2] Universitas Nusa Putra, Sukabumi 43152, Jawa Barat, Indonesia, sudin.saepudin@nusaputra.ac.id
[3] Universitas Nusa Putra, Sukabumi 43152, Jawa Barat, Indonesia, carti@nusaputra.ac.id
[4] Universitas Nusa Putra, Sukabumi 43152, Jawa Barat, Indonesia, mupaat@nusaputra.ac.id

Correspondence should be addressed to Ana Fatahali Ramadhan; anafatahali@gmail.com

Received 30 September 2025; Accepted 12 November 2025; Published 31 December 2025

© Authors 2025. CC BY-NC 4.0 (non-commercial use with attribution, indicate changes).
License: https://creativecommons.org/licenses/by-nc/4.0/ — Published by Indonesian Journal of Data and Science.

Abstract:
The digital printing industry in Indonesia is experiencing rapid growth thanks to the increasing demand from companies for printing services such as banners, stickers, brochures, and business cards. CV. Copy Paste is one of the companies operating in the digital printing industry that fulfills various printing orders every month. However, the company has difficulty identifying the most popular printing services, which makes it difficult to develop a targeted promotional strategy. In view of this problem, the aim of this study is to group digital printing services according to their popularity using the K-Means Clustering method. This study uses a quantitative approach, collecting sales data from the last 12 months, covering 160 types of services. The steps taken include preliminary data processing, namely attribute selection, data cleaning, and data transformation so that it can be effectively processed using the K-Means algorithm, implemented in the Python programming language. The test results show that digital printing services can be divided into three clusters: 115 less popular services (C1), 31 fairly popular services (C2), and 14 very popular services (C3). The results of this study provide information that can be used as a basis for strategic decisions regarding promotion and service management. In this way, the K-Means Clustering algorithm has proven effective in helping companies group products in a more objective and measurable way based on historical data.

Keywords: K-Means Clustering, Data Mining, Digital Printing, Python, Best-selling Services.

1. Introduction

Digital printing is one of the modern printing technologies that is currently undergoing rapid development. This method involves processing designs in the form of images, text, illustrations, and colors using a computer, and then printing them directly onto the medium using a digital printing machine [1]. In recent years, Indonesia's digital printing sector has experienced notable expansion, recording a growth rate of 14.9%. In line with the increase in the global printing industry which reached USD 47 billion. In addition, supporting sectors such as the packaging and advertising industries also recorded growth of 13.2% and 12.1% respectively, which also drove the demand for digital printing services [2].

Globally, the commercial printing market was valued at USD $339.57 billion in 2024 and is expected to grow to USD $357.8 billion by 2025, at a compound annual growth rate of 4.3% (CAGR), with the market projected to reach $423.75 billion by 2029. As shown in the following graph: [3]

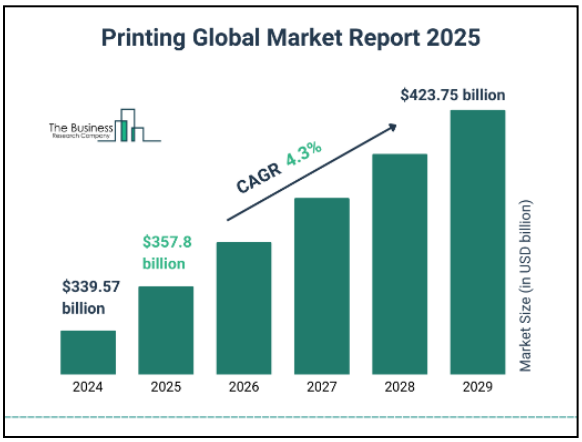


Figure 1. Global Printing Industry Market Growth Chart 2025 [3]

Data mining can be utilized by large companies to gather valuable information that helps improve and optimize business processes [4], [5]. By exploring databases, Data Mining can identify hidden patterns that may provide insights and predictions that business practitioners might not be aware of, as these patterns fall outside their expectations [6], [7]. CV. Copy Paste is a company engaged in digital printing and serves various types of services such as printing banners, stickers, brochures, business cards, and other custom products. The demand for services in this company is relatively high every month, with a variety of products and varying order volumes from consumers. However, in facing market dynamics and increasingly competitive business competition, the company faces obstacles in identifying which services are the most popular and have the potential to be developed as superior services. The absence of data-based analysis causes the decision-making process related to product promotion and management to not run optimally.

To overcome these problems, this study proposes the application of the K-Means Clustering algorithm as an unsupervised learning method that can group data based on characteristic similarities [8], [9], [10], [11]. This method divides the data into different groups so that data with similar characteristics are placed in the same group, while data with different characteristics are placed in another group [12], [13]. Clustering algorithms, as one of the most commonly applied methods in data mining, assist businesses in uncovering user characteristics and behavioral trends within large datasets, allowing for more targeted and accurate marketing strategies [14], [15]. Although K-Means Clustering was chosen for its ability to efficiently group data based on characteristic similarities, this algorithm has important limitations. One of these is its sensitivity to outliers, as very extreme values can pull the centroid and affect the overall cluster formation [16]. This condition requires careful data preprocessing to ensure stable and representative clustering results. However, studies focusing specifically on clustering for digital printing services remain limited. Prior research often lacks transparent analytical workflows, clear feature definitions, cluster validation metrics, and discussion on managerial implications. This algorithm is used to group digital printing services based on sales data into three main clusters, namely less popular, quite popular, and very popular services. The implementation is carried out using the Python 3.11 programming language with NumPy 2.0, Pandas 2.2, and Scikit-Learn 1.6, thus enabling efficiency in the data analysis and visualization process [8]. Python [17] is a general high-level programming language. Furthermore, Python is simpler to use compared to compiled languages because numerous complex aspects, like memory management, are managed automatically. The quick processing speed of Python is a significant benefit in product development, which necessitates swift iterations [18],[19].

The scope of this study is limited to sales data collected over the last twelve months, covering a total of 160 different services. The analysis focuses on two relevant numerical features Initial Stock and Total Sold which represent service availability and sales outcomes. Clustering is conducted using three popularity levels to classify the services into less popular, fairly popular, and very popular categories. However, the study has several limitations, particularly the reliance on the completeness and accuracy of historical sales data. Additionally, this research does not take into account external factors such as seasonal fluctuations, consumer preferences, or the influence of marketing campaigns, all of which may significantly affect service performance. This research has several advantages that provide added

value, both from an academic and practical perspective. One of the main advantages is that the use of the K-Means Clustering algorithm allows the efficient grouping of digital printing service sales data based on the level of sales [20].

A related study applied the K-Means Clustering algorithm to identify the best-selling Muslim fashion products in a retail store. Using Orange Data Mining, the researchers analyzed 282 products and produced three clusters: 228 items were classified as low-selling, 52 as moderately selling, and 2 as highly selling. The findings support inventory optimization and more effective sales promotion strategies [21].

Another study titled "Application of K-Means Clustering for Customer Segmentation in Grocery Stores in Kenya" identified homogeneous customer groups based on transaction data and demographic factors such as age, income, and spending scores. Using K-Means and the Elbow method to determine the optimal number of clusters, the research found that customers could be grouped into 4 to 6 clusters depending on the variable combinations. Customers aged 20-40 showed higher spending tendencies, making them a strategic target for promotions. High-income customers with low spending scores also exhibited significant potential for increased purchase value. These insights help grocery store managers design targeted marketing strategies, improve inventory management, and enhance customer experience through personalized approache [22].

The contribution of this study lies not only in the technical application of the K-Means algorithm, but also in demonstrating how the resulting clusters can serve as a foundation for more measurable and data-driven business strategies. Companies can focus promotions on services in the best-selling cluster, evaluate services in the less popular cluster, and allocate resources more precisely. In addition, the study offers a transparent analytical workflow supported by reproducible Python code, visualizations, and a clear cluster profile table that managers can directly use for decision-making. Beyond its relevance for CV. Copy Paste, this research also provides practical value as a reference for other digital printing businesses seeking to manage their service offerings based on actual sales data.

It is hoped that this research will serve as an initial step toward data-driven digital transformation in printing businesses like CV. Copy Paste through the implementation of the K-Means Clustering method. The results of the study are expected to be used to formulate promotional strategies, compile service catalogs, and more efficient stock planning. In the future, this research can also be developed through integration with other algorithms, such as machine learning-based sales predictions or the development of service recommendation systems, as well as considering external variables such as seasonal trends and consumer preferences.

2. Method

This study employs a quantitative method, which is designed to objectively describe or clarify an issue and allows for generalization. This approach relies on actual data that can be measured and statistically analyzed [23]. In this study, Knowledge Discovery in Database (KDD) will be used as part of the data analysis. The KDD process involves collecting and using data to identify specific rules or patterns in very large data sets [24]. The initial dataset consisted of four attributes Category, Service Name, Initial Stock, and Total Sold. Because Category and Service Name are categorical, they were converted into numerical form using frequency encoding, where each unique value is replaced with its relative frequency in the dataset. This approach preserves distributional patterns and provides a compact numerical representation suitable for distance-based clustering.

Although all four attributes were successfully prepared, exploratory experiments showed that the most stable clustering structure was achieved when only the two numerical attributes Initial Stock and Total Sold were used. Therefore, these two features were selected as the final input for K-Means, ensuring consistency across preprocessing steps, Python implementation, and visualization outputs.

In general, KDD involves the following steps: selection, preprocessing/cleaning, transformation, data mining, and evaluation. A more detailed explanation follows below:

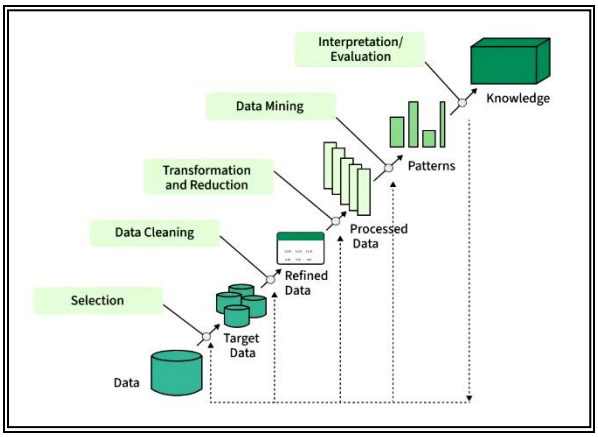


Figure 2. Knowledge Discovery in Database (KDD) [25]

a. Selection

Before the data is processed, the first step is to select the data. The data is retrieved from the inventory of grocery stores. This information includes the characteristics and quantities of the available goods. Data selection is based on the needs of Knowledge Discovery in Database (KDD). Therefore, not all characteristics collected during data collection will be used.

b. Pre-processing

At this stage, special attention is paid to the data that will be processed. This process includes removing duplicate elements, checking for data inconsistencies, and correcting errors such as spelling mistakes.

c. Transformation

This transformation data distribution variables into unique initial codes for use in calculations using the k-means method. Since the k-means algorithm can only handle numerical data, these variables are converted into numerical form when using the utility to apply them.

d. Data Mining

The data mining process applies the K-Means Clustering algorithm to categorize services based on specific predefined attributes or features. Then, the most appropriate number of groups or clusters to be studied is determined.

e. Interpretation/Evaluation

Interpretation and evaluation is the phase in which research results are evaluated after data mining. This stage is carried out after the data is processed using the k-means clustering algorithm with the Python programming language, and the aim is to evaluate the results of the data mining process.

The research method applied in this study is the K-Means algorithm, which is an algorithm that functions to perform grouping (clustering) on the data being analyzed. In this section, we describe the method for determining initial centroids in our proposed algorithm, which aims to improve clustering accuracy by considering the values of each column in the dataset [26]. Selecting initial centroids is a crucial step in clustering algorithms because it significantly affects the quality of the resulting clusters [27]. The stages carried out in this research process can be seen in the following picture:

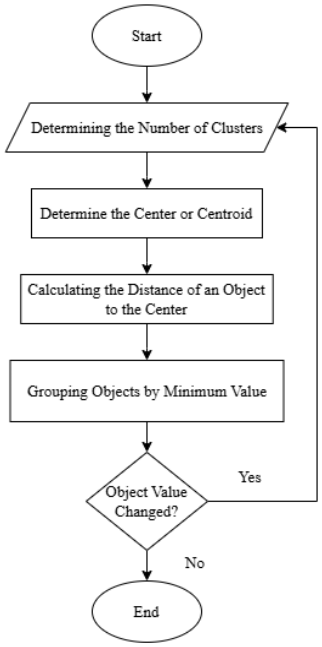


Figure 3. Flowchart of the K-Means Clustering Algorithm

The grouping stage using the K-Means algorithm is as follows: [28],[21]

1) The number of k (Clusters) and the initial centroid are determined randomly.
2) The distance between each data point and the cluster center is measured using the Euclidean Distance, calculated with the following formula:

$$D(i,j) = \sqrt{(X1i - X1j)^2} + (X2i - X2j)^2 + \cdots + (Xki - Xkj)^2 \quad (1)$$

Description:

$D(i,j)$= Distance from data i to cluster center j

Xki = Data i on data attribute k

Xkj = Center point j on attribute k

3) Grouping data into clusters with the closest distance.
4) Calculate the new cluster center by finding the average of the data groups.
5) The center of the cluster is determined when all data are assigned to the nearest cluster.
6) The process of determining the cluster midpoint and assigning the data to the clusters is repeated until the midpoint value no longer changes.

The Within-Cluster Sum of Squares (WCSS) is used to assess how densely clustered members are by calculating the sum of the squares of the distances of each data point from its cluster centroid [29]. This measure reflects the degree of cluster compactness: the smaller the WCSS value, the closer the data points are to their respective centroids. The optimization process minimizes the Within-Cluster Sum of Squares (WCSS):

$$WCSS = \sum_{i=1}^{K} \sum_{x_j \in S_i} \|(x_j, c_i)\|^2 \quad (2)$$

The experiments were conducted using Python 3.11, NumPy 1.26, Pandas 2.0, and Scikit-Learn 1.5.0. The dataset

contains no personal information, was obtained with company permission, and may be shared in anonymized form upon request.

3. Result and Discussion

Results

The sample used is sales data for twelve (12) months with a total of 160 data. Two numeric attributes Initial Stock and Total Sold were used for clustering because categorical attributes can distort Euclidean distance when encoded. This decision ensures consistency, avoids bias from arbitrary encoding, and aligns the feature set with clustering requirements. From the existing data, calculations or tests are carried out on the data using the Python 3.11 programming language.

a. Pre-processing

Once the sales data are obtained, whose conditions need to be selected again, a data preprocessing stage needs to be performed depending on the results of the literature review before moving to the system implementation stage. The data pre-processing performed is described below:

b. Attribute Selection

Of the various attributes in the original data, after selecting the attributes, there are 2 (two) attributes used in this study, namely: Total Sold and Initial Stock.

c. Data Cleaning

After the 2 (two) attributes were determined, data that did not fit before further data processing was also cleaned, such as data that was incorrectly entered by mistake or even data that had no meaning at all.

d. Data Transformation

Datasets are derived from the results of prior attribute selection and data cleansing, and then mathematically transformed from raw data type data into processable data. This transformation is performed from alphanumeric data or textual data to numeric data or data in the form of numbers.

Table 1. Dataset

No	Category	Name of Goods/Services Purchased	Total Sold	Initial Stock
1	Stiker	Stiker Vinyl A3+	60	250
2	Stiker	Stiker Kromo A3+	19	120
3	Dokumentasi	Sampul Ijazah	208	600
…		…	…	
159	Percetakan	Baliho / Spanduk 2x1,5	6	70
160	Stiker	Stiker Motor 110	5	50

Implementation of K-Means Clustering Using Python

Based on the pre-processed data described earlier, the system test is executed using the Python 3.11 programming language and applying the K-Means clustering method as outlined below:

At the initial stage, after the data set is prepared for processing in the system, which is the result of previous processing, the data set loading process is performed as shown in the following figure:

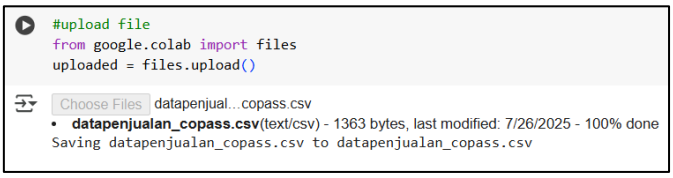


Figure 4. Data File Upload Process

Next, import libraries as needed for system testing. The libraries used include Numpy, which serves to convert data into arrays; Matplotlib, which is used to visualize data to make it easier to understand; Pandas, for numerical table manipulation.

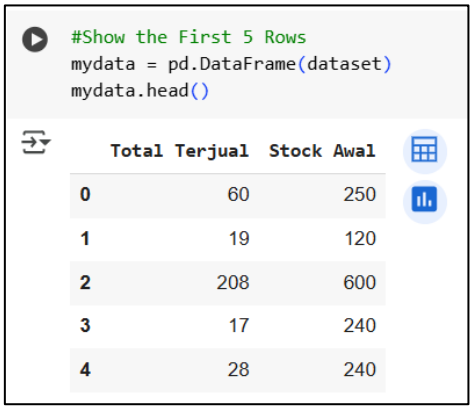
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

Dataset import process Once the dataset file is successfully loaded, the data will be stored in a variable. The variables used here are Category and Total number. The dataset stored in the variable is named as shown below:



Figure 5. Dataset Storage and Retrieval Process

In the next stage, the first five rows are displayed first to ensure that the dataset used is as required.



	Total Terjual	Stock Awal
0	60	250
1	19	120
2	208	600
3	17	240
4	28	240

Figure 6. The First Five Rows of the Dataset

Data conversion After the dataset is suitable and can be applied in the system, the data is converted into an array using the Numpy library so that it can be processed according to the existing system requirements. As seen in the picture below:

```
#Convert Data to Array
X = np.asarray(dataset)
print (X)
```

```
[[  60  250]
 [  19  120]
 [ 208  600]
 [  17  240]
 [  28  240]
 [  30  180]
 [  77 1200]
 [  35  240]
 [  61  400]
```

Figure 7. Data Conversion Process

In its implementation, this study uses K-Means++ with a random seed value = 45. This seed setting ensures that the initialization process can be replicated and the cluster results obtained are consistent.

Apply the Elbow Method to determine the optimal number of clusters by calculating the Within-Cluster Sum of Squares (WCSS) value for various numbers of clusters. The elbow point on the graph indicates the best number of clusters. This process is assisted by Scikit-Learn, a Python library that provides essential functions in implementing the K-Means Clustering algorithm, so that clustering can be done efficiently and accurately.

```
from sklearn.cluster import KMeans

# Drop rows with NaN values in the relevant columns
X = dataset.dropna(subset=['Stock Awal', 'Total Terjual']).iloc[:, [0,1]].values

wcss = []
for i in range (1,11):
    kmeans = KMeans(n_clusters=i, init = 'k-means++', random_state= 45)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
plt.plot(range(1,11), wcss)
plt.title('Metode Elbow untuk Menentukan Jumlah Cluster Optimal')
plt.xlabel('Jumlah cluster')
plt.ylabel('WCSS')
plt.show()
```

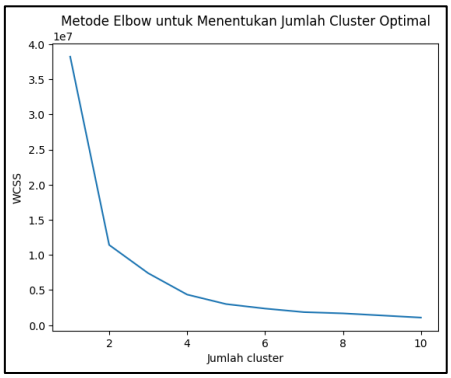


Figure 8. Elbow Method Results

Figure 9 of the Elbow method shows that the WCSS value experiences a very sharp decrease from $k = 1$ to $k = 2$ and decreases again quite significantly at $k = 3$. After that point, the decrease in WCSS becomes much more gradual until $k = 10$, as also shown by the summary table of WCSS values. This pattern shows a clear elbow point at the number of clusters of three, so $k = 3$ is determined as the optimal number of clusters used in this study.

To facilitate understanding of the data to be analyzed, a visualization process was performed using a scatterplot graph. This visualization was created using the Matplotlib library, as shown in the following image:

```
plt.scatter(X[:, 0], X[:, 1], s = 30, label='Data Awal', c='black')
plt.xlabel('Stock Awal')
plt.ylabel('Total Terjual')
plt.title('Data Awal')
plt.legend()
plt.grid()
```

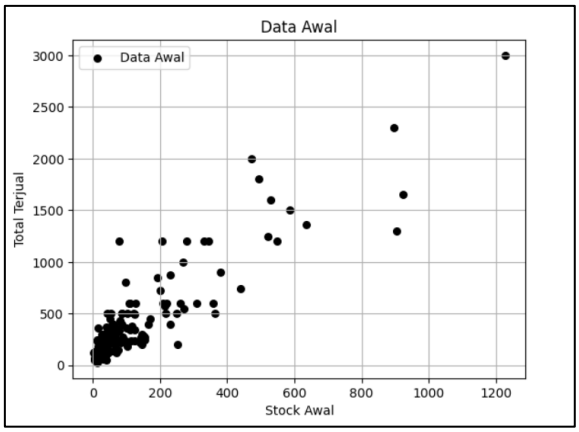


Figure 9. Data Visualization in Scatterplot Form

The next step is to activate the K-Means Clustering algorithm using the Scikit-Learn library, while determining the desired number of clusters.

```
#Activate K-Means with the Number k=3
kmeans = KMeans(n_clusters=3, init = 'k-means++', random_state= 45)
y_kmeans = kmeans.fit_predict(X)
```

In the next stage, the results of the centroid value of each cluster obtained previously will be reviewed.

```
#Show Centroid Values
print(kmeans.cluster_centers_)
```

```
[[  49.7826087   209.26086957]
 [ 197.61290323  646.58064516]
 [ 620.64285714 1611.42857143]]
```

Figure 10. Shows Centroid Values

In **Figure 10**, the centroid value for cluster 1 is 209.26, for cluster 2 is 646.58, and for cluster 3 is 1611.42.

Table 2. Centroid Distance

Cluster	Centroid Value (Initial Stock, Total Sold)	Number of Members	Top Items Based on Total Sold
0	(≈ 49.78 , 209.26)	115	- Kalender Duduk A5 AC 210 (145) - Banner FL 280 3x2 (61) - Stiker Vinyl A3+ (60)
1	(≈ 197.61 , 646.58)	31	- Kaos (439) - Sampul Raport (364) - Gantungan Kunci Pin 2S (380)
2	(≈ 620.64 , 1611.42)	14	- Art Paper 260gr Satu Sisi (1.227) - Notebook A5 (924) - Kertas Undangan Jasmine (903)

In addition to evaluating the centroid value, it is also important to check the results of the preprocessed data labeling. The process and results can be seen in the following figure:

```
#show data points labels
print(y_kmeans)
```

```
[0 0 1 0 0 0 1 0 0 0 0 0 0 2 1 2 1 1 1 2 0 2 1 1 0 0 0 1 1 1 2 2 2 0 0 0 0
 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 2 2 0 0 0 0 0 2 2 2 0 0 0 0 0 1 0
 0 1 2 1 1 1 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2 0 0 0 0 0 0 0 0 0 0
 1 0 0 1 0 0 1 0 0 0 0 0]
```

Figure 11. Display of Point Data Labels

Based on the results of the previous stage, visualization is carried out using scatterplot to show the results of data clusterization processed using the Python programming language. Each cluster is given a different color. The process and results of scatterplot visualization can be seen in the image displayed below.

```
plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s = 30, c = 'red',
label = 'Cluster 1')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s = 30, c = 'blue',
label = 'Cluster 2')
plt.scatter(X[y_kmeans == 2, 0], X[y_kmeans == 2, 1], s = 30, c = 'green',
label = 'Cluster 3')
plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:,1], s =
100, c = 'yellow', label = 'Centroid')
plt.xlabel('Stock Awal')
plt.ylabel('Total Terjual')
plt.title('Hasil Klasterisasi')
plt.legend()
plt.grid()
```

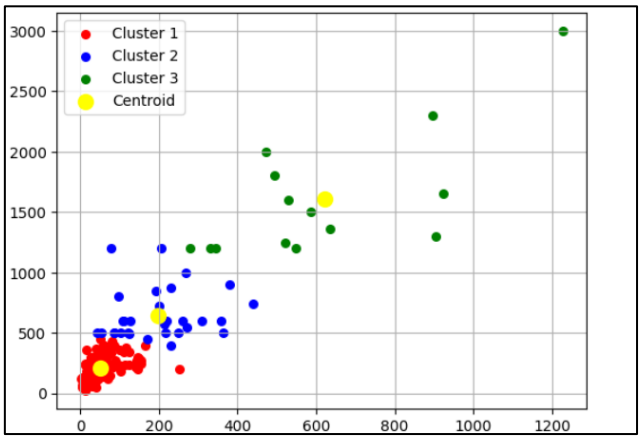


Figure 12. Display of Clustering Results and Centroid Data

The next stage is to add data from the clusters obtained with the initial dataset, so that clear information can be obtained to represent the cluster results that have been generated previously. The following is the process and results of adding clusters to the dataset.

```
dataset['cluster'] = kmeans.labels_
dataset
```

At the end, the process of downloading data obtained from the unification of the initial dataset with the clusters obtained from the results of data processing carried out through the application of the Python programming language is carried out. Thus, the data obtained will provide more comprehensive and valuable information. The following are the steps taken when downloading the data.

```
dataset.to_csv('result.csv')
files.download('result.csv')
```

Discussion

This study successfully demonstrated that the K-Means Clustering algorithm is effective in grouping digital printing services at CV. Copy Paste into three categories based on their level of popularity: less popular, moderately popular, and very popular. With this grouping, the company can more easily understand the performance of each service offered and use it as a basis for service evaluation.

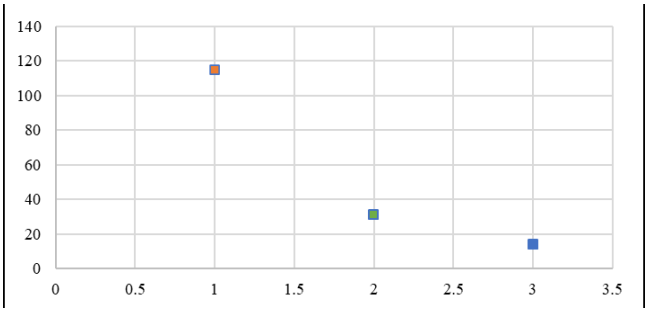


Figure 13. Clustering results based on Python processing

The application of the K-Means Clustering algorithm to digital printing service sales data at CV. Copy Paste is carried out to group services based on their level of popularity or best-selling. The three main categories used are Less Popular (C1), Quite Popular (C2), and Very Popular (C3). Based on these findings, the data distribution shows that out of a total of 160 types of digital printing services, 115 are categorized as less popular (C1), 31 are categorized as

quite popular (C2), and 14 services are categorized as very popular (C3). The following services are included in the best-selling group:

Table 3. Best-selling Group Services

No	Name of Goods/Services Purchased	Total Sold	Cluster
1	Lanyard 2cm 2S	587	2
2	Lanyard 1,5cm 2S	344	2
3	*Gantungan Kunci Plastik*	529	2
4	Art Paper 230gr Dua Sisi	549	2
5	Goodie Bag 20x26	330	2
6	Goodie Bag 30x40	895	2
7	Goodie Bag 25x35	495	2
8	Art Paper 260gr Satu Sisi	1227	2
9	Art Paper 260gr Dua Sisi	471	2
10	Notebook A5	924	2
11	Notebook A6	280	2
12	*Kertas Undangan A3+ Jasmine*	903	2
13	*Gantungan UV*	520	2
14	*Pulpen Custom*	635	2

The membership rules for services classified in the top cluster (Cluster 2) are determined based on the proximity of each service's feature values to the cluster centroid, using Euclidean distance calculations. Considering only two key features: Total Sold and Initial Stock, a service is classified into Cluster 2 if the values of both features are closest to the cluster centroid compared to the centroids of other clusters. Empirically, services in Cluster 2 are characterized by very high sales volumes and relatively large initial stocks, reflecting a strong and consistent demand pattern. These services are positioned in the cluster because they numerically resemble the centroid profile of the "high-demand, high-volume" group. Thus, cluster membership reflects mathematical proximity to the cluster center and illustrates the service categories that dominate overall sales contributions.

Through the results of these clustering, the company obtains relevant information to develop a more targeted marketing strategy. By identifying services with the highest sales performance, the company can prioritize them in promotional initiatives. Understanding which service groups generate stronger engagement or demand also enables managers to direct improvement efforts more effectively and design targeted interventions that align with customer behavior patterns [30]. Thus, companies can optimize resource allocation and increase competitiveness in the digital printing industry through data-based decision making.

4. Conclusion

This study successfully demonstrated the effectiveness of the K-Means Clustering algorithm in grouping digital printing services at CV. Copy Paste into three categories based on their level of demand: less popular, moderately popular, and very popular. This grouping allows the company to more easily understand the performance of each service offered and use it as a basis for service evaluation.

The clustering process used the Python 3.11 programming language, which is capable of efficiently processing sales data and producing informative cluster divisions. Of the 160 services analyzed, 115 fell into the less popular (C1) category, 31 into moderately popular (C2), and 14 into the very popular (C3) category, namely Art Paper 260gr One Side with a total of 1227 sold, A5 Notebook with a total of 924 sold, A3+ Jasmine Paper Invitation Paper with a total of 903 sold, and Goodie Bag 30x40 with a total of 895 sold.

With this clustering result, CV. Copy Paste obtained information that can be used to formulate a more targeted marketing strategy. Services in the high-selling cluster can be prioritized for intensive promotion, while services in the less popular cluster can be further evaluated to determine whether they need improvement, innovation, or even

discontinuation. This supports data-driven decision-making for efficient resource allocation and increased company competitiveness in the digital printing market.

This study has limitations, particularly the absence of external factors such as seasonal demand patterns and promotional campaigns that may affect sales performance. Future work could incorporate temporal dynamics, such as monthly sales trends, and evaluate alternative clustering methods to test the stability and consistency of the resulting service groups.

References:

[1] K. Saharja and R. Gobal, "*Pengaruh Waktu Proses Produksi Digital Printing Terhadap Kepuasan Konsumen Pengguna Produk Cetak*," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 1, pp. 458–469, 2021, doi: 10.30645/j-sakti.v5i1.339.

[2] Bhinneka, "*Masa Depan Bisnis Digital Printing di Indonesia*," *Bhinneka Update*, May 14, 2024.

[3] *Printing Global Market Report 2025*. 2025.

[4] M. F. Haryanti *et al.*, "*Pengaruh Data Mining, Strategi Perusahaan Terhadap Laporan Kinerja Perusahaan*," *J. Manaj. dan Bisnis*, vol. 3, no. 1, pp. 71–90, 2024, doi: 10.70704/jpjmb.v3i1.285.

[5] H. S. Nugraha, H. Mutaqin, A. Fathah, and C. Juliane, "*Mengidentifikasi Strategi Promosi pada Jasa Penjualan Saldo Digital menggunakan Pendekatan Clustering," Edumatic J. Pendidik. Inform., vol. 7, no. 1, pp. 11–19, 2023, doi: 10.29408/edumatic.v7i1.7385*.

[6] Z. I. Alfianti, M. A. Azis, and A. Fauzi, "Grouping of Covid-19 Affected Areas in Bogor City Using The K-Means Algorithm," *J. Mantik*, vol. 4, no. 4, pp. 2336–2341, 2021, doi: https://doi.org/10.35335/mantik.Vol4.2021.1142.pp2336-2341.

[7] V. S. Moertini, "Data Mining Sebagai Solusi Bisnis," *Integral*, vol. 7, no. 1, 2002.

[8] A. F. AlShammari, "Implementation of Clustering using K-Means in Python," *Int. J. Comput. Appl.*, vol. 186, no. 40, pp. 12–17, Sep. 2024, doi: 10.5120/ijca2024923990.

[9] G. Gunadi, "*Penerapan Algoritma K-Means Clustering Untuk Menganalisa Transaksi Penjualan Jasa Cetak Pada Unit Print on Demand (Pod) Percetakan Gramedia*," *Infotech J. Technol. Inf.*, vol. 8, no. 2, pp. 117–126, 2022, doi: 10.37365/jti.v8i2.148.

[10] Wanto, Anjar, M. N. H. Siregar, and and A. P. Windarto, *Data Mining : Algoritma Dan Implementasi*, 1st ed. Yayasan Kita Menulis., 2020.

[11] I. Taufik, N. Sa'adah, N. Suparna, C. Alam, and P. Dauni, "Scoring System and K-Means Algorithm for Mutaba'ah Yaumiyah Activity," 2020, doi: 10.4108/eai.11-7-2019.2297566.

[12] N. K. Zuhal, "Study Comparison K-Means Clustering dengan Algoritma Hierarchical Clustering," *Pros. Semin. Nas. Teknol. Dan Sains*, vol. 1, pp. 200–205, 2022, doi: https://doi.org/10.29407/stains.v1i1.1495.

[13] A. Yani, Z. Azmi, and D. Suherdi, "Implementasi Data Mining Menganalisa Data Penjualan Menggunakan Algoritma K-Means Clustering," *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, vol. 2, no. 2, p. 315, 2023, doi: 10.53513/jursi.v2i2.6357.

[14] M. Rochmawati, G. W. C. Bagaskara, I. A. Adha, Y. Umaidah, and A. Voutama, "Implementation of the K-Means Algorithm in Sales Clustering at a Company using the KDD Methodology," *SISTEMASI*, vol. 13, no. 1, p. 54, Jan. 2024, doi: 10.32520/stmsi.v13i1.3074.

[15] D. Wu and L. Xin, "HC-means clustering algorithm for precision marketing on e-commerce platforms," *Syst. Soft Comput.*, vol. 7, 2025, doi: https://doi.org/10.1016/j.sasc.2025.200236.

[16] P. Kirst, T. Bajbar, and M. Merkel, "A bisection method for solving distance-based clustering problems globally," *TOP*, vol. 33, no. 3, pp. 437–469, 2025, doi: 10.1007/s11750-024-00684-w.

[17] R. S. Chauhan, A. Munshi, and A. Pradhan, "The Role of Python in Enhancing Radiotherapy Department Workflow Efficiency and Promoting Open-source Software Utilisation," *Clin. Oncol.*, vol. 45, p. 103897, Sep. 2025, doi: 10.1016/J.CLON.2025.103897.

[18] L. Simorangkir, E. Sany, and M. F. N, "*Penerapan Metode K-Means Untuk Pengelompokan Data Kunjungan Wisata Pada Dinas Kebudayaan Dan Pariwisata Provinsi Jambi*," *J. Akad.*, vol. 17, no. 2, pp. 35–40, Jun. 2025, doi: 10.53564/akademika.v17i2.1496.

[19] K. Rakesh *et al.*, *Python for Beginners : A Comprehensive Guide to Learning Python Programming*, First Edit., no. October. Warta Saya, 2024.

[20] M. Suyal and S. Sharma, "A Review on Analysis of K-Means Clustering Machine Learning Algorithm based on Unsupervised Learning," *J. Artif. Intell. Syst.*, vol. 6, no. 1, pp. 85–95, 2024, doi: 10.33969/ais.2024060106.

[21] M. Syukron Nawawi, F. Sembiring, and A. Erfina, "Implementasi Algoritma K-Means Clustering Menggunakan Orange Untuk Penentuan Produk Busana Muslim Terlaris," *Progr. Stud. Tek. Inform. Pgri Madiun*, pp. 789–797, 2021.

[22] E. Omol, D. Onyangor, L. Mburu, and P. Abuonji, "Application Of K-Means Clustering For Customer Segmentation In Grocery Stores In Kenya," *Int. J. Sci. Technol. Manag.*, vol. 5, no. 1, pp. 192–200, 2024, doi: 10.46729/ijstm.v5i1.1024.

[23] A. J. Alifah, S. Saepudin, and C. Irawan, "Implementation of the K-Means Clustering Algorithm in Analyzing Public Satisfaction Regarding Public Services (Studi Case : Balai Pengujian Standar Instrumen Tanaman Industri Dan Penyegar) Implementasi Algoritma K-Means Clustering Dalam Menganalisis Kep," vol. 5, no. 4, pp. 487–496, 2024, doi: 10.52436/1.jutif.2024.5.4.2125.

[24] F. Zafira, B. Irawan, and A. Bahtiar, "Penerapan Data Mining Untuk Estimasi Stok Barang Dengan Metode K-Means Clustering," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 156–161, 2024, doi: 10.36040/jati.v8i1.8319.

[25] GeeksforGeeks, "KDD Process in Databases." Accessed: Jun. 06, 2025.

[26] C. W. Id, "The impact of neglecting feature scaling in k-means clustering," pp. 1–19, 2024, doi: 10.1371/journal.pone.0310839.

[27] A. A. Khan, M. S. Bashir, A. Batool, M. S. Raza, and M. A. Bashir, "K-Means Centroids Initialization Based on Differentiation Between Instances Attributes," *Int. J. Intell. Syst.*, vol. 2024, no. 1, 2024, doi: 10.1155/2024/7086878.

[28] R. Afifa, M. I. Mazdadi, T. H. Saragih, F. Indriani, and M. Muliadi, "Implementasi Principal Component Analysis (PCA) dan Gap Statistic untuk Clustering Kanker Payudara pada Algoritma K-Means," *SISTEMASI*, vol. 13, no. 5, p. 1852, Sep. 2024, doi: 10.32520/stmsi.v13i5.4015.

[29] J. Meng *et al.*, "Heliyon Nano-integrating green and low-carbon concepts into ideological and political education in higher education institutions through K-means clustering," *Heliyon*, vol. 10, no. 10, p. e31244, 2024, doi: 10.1016/j.heliyon.2024.e31244.

[30] T. P. Scholdra, J. R. K. Wichmann, and W. J. Reinartz, "Reimagining personalization in the physical store," *J. Retail.*, vol. 99, no. 4, pp. 563–579, 2024, doi: 10.1016/j.jretai.2023.11.001.