Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

PENERAPAN ALGORITMA *K-MEANS* UNTUK PENGELOMPOKAN PADA DATA SUPERMARKET MENGUNAKAN PEMOGRAMAN *PYTHON*

Rizky Prastya Wibowo[1]
Erlin Elisa[2]

[1]Program Studi Sistem Infromasi Universitas Putera Batam
[2]Program Studi Sistem Informasi, Universitas Putera Batam
email: pb211510008@upbatam.ac.id

ABSTRACT

The rapid growth of transactional data in the retail sector, particularly in supermarkets, has resulted in large and complex sales datasets that require effective analytical methods to identify meaningful sales patterns. This study applies data mining through clustering techniques by implementing the K-Means algorithm using Python to classify supermarket sales patterns and group transactions with similar characteristics. The research methodology includes data collection, preprocessing, normalization using StandardScaler, determination of the optimal number of clusters through the Elbow Method, clustering with the K-Means algorithm, and evaluation of clustering quality using the Silhouette Score, Davies–Bouldin Index, and Inertia (Within-Cluster Sum of Squares). The results indicate that from 1,000 sales transactions, three clusters were formed, consisting of 515 transactions in the low sales cluster, 314 in the medium sales cluster, and 171 in the high sales cluster. The clustering evaluation produced a Silhouette Score of 0.6055, a Davies–Bouldin Index of 0.502, and an Inertia value of 122.01, indicating that the resulting clusters are compact and well separated. These findings confirm that the K-Means algorithm is effective for grouping supermarket sales patterns and can support sales management and strategic planning.

Keywords: : *K-Means;Clustering;Sales Patterns;Supermarket Data;Python*

PENDAHULUAN

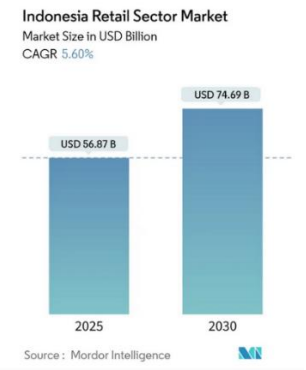
Industri ritel di Indonesia memiliki potensi pertumbuhan yang sangat besar dan menjadi salah satu sektor usaha yang paling diminati. Peningkatan nilai pasar ritel yang signifikan, didukung oleh perubahan gaya hidup masyarakat, pertumbuhan ekonomi nasional, serta pembangunan infrastruktur distribusi, menunjukkan bahwa sektor ini masih akan terus berkembang dalam beberapa tahun ke depan. Kondisi tersebut menjadikan industri ritel tidak hanya sebagai tulang punggung ekonomi domestik, tetapi juga sebagai peluang strategis bagi pelaku usaha dan investor dalam mengembangkan bisnis yang berorientasi pada efisiensi, kenyamanan, dan pemanfaatan teknologi. Namun demikian, sejumlah penelitian menunjukkan bahwa manajemen usaha ritel atau supermarket masih menghadapi kendala utama dalam mengidentifikasi dan memahami pola penjualan produk

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

secara menyeluruh (Fajar Maulana Adji & Dwilestari, 2024). Data penjualan yang besar dan bervariasi sering kali belum diolah secara optimal untuk mengungkap tren permintaan, pengaruh musiman, maupun perubahan perilaku pelanggan, sehingga menyulitkan manajemen dalam membedakan produk dengan penjualan stabil, musiman, atau fluktuatif (Fadel et al., 2025). Kondisi ini berdampak pada pengambilan keputusan yang kurang efektif, terutama dalam perencanaan stok, strategi promosi, dan evaluasi kinerja produk, serta meningkatkan risiko kesalahan dalam pengadaan dan penjualan (Fahrizal et al., 2024).



Gambar 1. Perkembangan Sektor Ritel Di Indonesia
Sumber: Data Penelitian 2026

Oleh karena itu, diperlukan metode analisis berbasis data yang mampu mengelompokkan produk berdasarkan kesamaan pola penjualan untuk mendukung pengambilan keputusan yang lebih adaptif dan efisien. Penelitian ini bertujuan menerapkan algoritma K-Means dalam proses pengelompokan pola penjualan pada data supermarket, sehingga dapat diidentifikasi kelompok produk dengan karakteristik penjualan yang serupa. Berbagai penelitian sebelumnya menunjukkan bahwa algoritma K-Means efektif dalam pengelompokan data penjualan ritel dan segmentasi pelanggan, baik pada tingkat toko maupun cabang (Hidayat et al., 2025; Alvianatinova et al., 2024), serta dalam pemodelan perilaku pembelian pelanggan menggunakan pendekatan RFM (Anitha & Patil, 2022). Meskipun demikian, sebagian besar penelitian terdahulu masih berfokus pada penerapan algoritma menggunakan perangkat lunak siap pakai dan belum mengintegrasikan eksplorasi parameter serta visualisasi data secara mendalam. Keterbaharuan penelitian ini terletak pada penerapan algoritma K-Means secara langsung menggunakan bahasa pemrograman Python, yang memungkinkan fleksibilitas eksplorasi model, evaluasi kinerja secara kuantitatif, serta penyajian hasil yang lebih aplikatif bagi pengelola supermarket dalam manajemen persediaan dan strategi pemasaran berbasis data.

KAJIAN TEORI

1. *Knowledge Discovery In Database*

Knowledge Discovery in Database (KDD) merupakan proses sistematis untuk mengekstraksi informasi yang bernilai dari kumpulan data dalam basis data berskala besar. KDD didefinisikan sebagai proses untuk memperoleh informasi penting dari data yang tersimpan dalam database besar (Maulana et al., 2024). Selain itu, KDD juga mencakup serangkaian kegiatan pengolahan data yang bertujuan mengidentifikasi kecenderungan dan pola, kemudian mengubah hasilnya menjadi informasi yang mudah dipahami (Utomo & Kurniasari, 2023). KDD

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

dipahami sebagai proses untuk memperoleh informasi baru, bernilai, dan bermakna dari kumpulan data yang besar dan kompleks (Amaliah et al., 2024).

Gambar 2. *Knowledge Discovery In Database*
Sumber:Data Penelitian, 2026

2. Data Mining

Data mining merupakan aktivitas pengolahan data yang memanfaatkan data historis untuk menemukan keteraturan serta pola hubungan dalam kumpulan data berukuran besar. Data mining juga dikenal sebagai Knowledge Discovery in Database (KDD), yaitu proses penemuan pengetahuan dari basis data (Siahaan, 2022). Selain itu, data mining didefinisikan sebagai proses mengekstraksi data yang bermanfaat dari gudang basis data yang sangat besar guna mendukung pengambilan keputusan (Sulistio et al., 2023). Proses ini melibatkan penelusuran data untuk membangun suatu model yang selanjutnya digunakan dalam menemukan pola tambahan yang sebelumnya tidak terlihat, termasuk untuk memenuhi kebutuhan prediksi. Data mining juga mencakup teknik pengelompokan data sebagai salah satu metode utama dalam analisis data. Menurut Sulastri dan Gufroni (2017), data mining merupakan proses penggalian dan penyaringan data dalam jumlah besar melalui tahapan tertentu untuk menghasilkan informasi yang bernilai dari kumpulan data berskala besar.

3. Clustering

Clustering merupakan teknik analisis data yang digunakan untuk membagi sekumpulan data ke dalam beberapa kelompok berdasarkan tingkat kesamaan karakteristik yang dimiliki. Setiap cluster terdiri dari objek-objek data yang memiliki tingkat kemiripan. tinggi dalam kelompok yang sama dan perbedaan yang signifikan dengan objek pada kelompok lain (Darmi & Setiawan, 2017). Prinsip utama clustering adalah memaksimalkan kesamaan antar objek dalam satu cluster dan meminimalkan kesamaan antar cluster, yang umumnya direpresentasikan sebagai titik-titik dalam ruang multidimensi berdasarkan atribut data. Clustering menjadi salah satu teknik penting dalam data mining untuk mengelompokkan data ke dalam subset yang serupa berdasarkan pola atau karakteristik tertentu (Hendrastuty, 2024). Tujuan utama dari clustering adalah mengidentifikasi struktur tersembunyi dalam data guna mendukung pemahaman yang lebih mendalam terhadap kelompok atau kategori yang terbentuk. Selain itu, teknik clustering banyak dimanfaatkan dalam berbagai bidang, seperti klasifikasi, pengolahan citra, dan pengenalan pola, karena kemampuannya dalam mengungkap struktur data yang kompleks (Rofiqo et al., 2018).

4. *K-Means*

Algoritma K-Means merupakan metode *clustering* berbasis partisi yang digunakan untuk mengelompokkan data ke dalam sejumlah kelompok berdasarkan kesamaan atribut. Tujuan

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

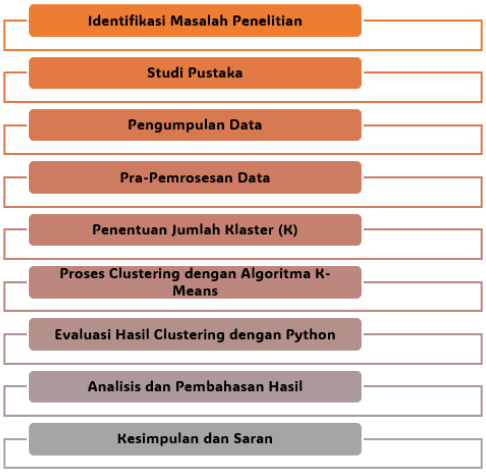
utama K-Means adalah meminimalkan perbedaan dalam satu kelompok dan memaksimalkan perbedaan antar kelompok dengan mengelompokkan data yang memiliki karakteristik sebanding (Sippan & Setiyawati, 2025). K-Means termasuk algoritma *unsupervised learning* yang melakukan pengelompokan data tanpa pelabelan awal, di mana setiap cluster memiliki karakteristik internal yang serupa namun berbeda dengan cluster lainnya (Harsono et al., 2023). Proses pengelompokan ini bertujuan untuk menghasilkan struktur data yang jelas dan mudah diinterpretasikan sebagai dasar analisis lebih lanjut (Sulastri & Gufroni, 2017).

5. Python

Python merupakan bahasa pemrograman tingkat tinggi yang banyak digunakan oleh pemula maupun profesional karena sintaksisnya yang sederhana dan mudah dipahami (Apriyanto & Sitio, 2025). Python bersifat serbaguna dan mendukung berbagai paradigma pemrograman, seperti pemrograman berorientasi objek, fungsional, dan prosedural. Selain itu, Python didukung oleh ekosistem pustaka yang kaya, antara lain NumPy untuk komputasi numerik, Pandas untuk analisis data, Matplotlib untuk visualisasi data, serta scikit-learn untuk pembelajaran mesin, sehingga sangat mendukung pengolahan dan analisis data dalam penelitian ilmiah (Apriyanto & Sitio, 2025).

METODE PENELITIAN

metode *clustering* K-Means sebagai pendekatan utama dalam desain penelitiannya. Metode ini dipilih karena kemampuannya dalam mengelompokkan data ke dalam beberapa klaster berdasarkan kemiripan karakteristik, tanpa memerlukan label sebelumnya. berikut adalah alur metode penelitian ini.



Gambar 3. Metode Penelitian
Sumber: Data Penelitian 2026

Tahap awal penelitian dimulai dengan identifikasi masalah penelitian yang dilanjutkan dengan studi pustaka untuk memperoleh landasan teoritis yang relevan. Selanjutnya, data penjualan dikumpulkan dan dilakukan pra-pemrosesan data yang mencakup pembersihan, penanganan data hilang, serta normalisasi data agar siap dianalisis. Setelah itu, jumlah klaster optimal ditentukan menggunakan metode tertentu, seperti Elbow Method. Proses clustering kemudian dilakukan dengan menerapkan algoritma K-Means menggunakan bahasa pemrograman Python. Hasil pengelompokan selanjutnya dievaluasi menggunakan metrik evaluasi clustering untuk menilai kualitas klaster yang terbentuk. Tahap akhir penelitian meliputi analisis dan pembahasan hasil clustering, serta penarikan kesimpulan dan pemberian saran berdasarkan temuan penelitian.

HASIL DAN PEMBAHASAN

1. Deskripsi Data

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

Data yang digunakan dalam penelitian ini merupakan data sekunder yang bersumber dari dataset publik penjualan supermarket yang tersedia pada platform Kaggle dengan nama *SuperMarket Analysis*. Dataset tersebut berisi 1.001 data transaksi penjualan yang berasal dari beberapa cabang supermarket yang berlokasi di berbagai kota.

2. Impor Libray

Tahap awal penelitian ini diawali dengan proses impor library yang diperlukan untuk mendukung analisis data dan penerapan algoritma K-Means menggunakan bahasa pemrograman Python. Untuk lebih jelasnya dapat dilihat bada gambar 4.

```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
```

Gambar 4. Impor Libray
Sumber: Data Penelitian 2026

3. Load Data

Setelah library berhasil diimpor, dataset penjualan supermarket dimuat menggunakan fungsi read_csv() dari library *pandas*.

```
import pandas as pd

df = pd.read_csv('/content/SuperMarket Analysis 1.csv')
display(df.head())
```

Gambar 5. Load Data
Sumber: Data Penelitian 2026

Selanjutnya, beberapa baris awal data ditampilkan sebagai langkah verifikasi untuk memastikan data telah terbaca dengan benar dan memiliki struktur yang sesuai untuk tahap analisis berikutnya. Seperti gambar 6 berikut.

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Sales	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	Alex	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	1:08:00 PM	Ewallet	522.83	4.761905	26.1415	9.1
1	226-31-3081	Giza	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29:00 AM	Cash	76.40	4.761905	3.8200	9.6
2	631-41-3108	Alex	Yangon	Normal	Female	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	1:23:00 PM	Credit card	324.31	4.761905	16.2155	7.4
3	123-19-1176	Alex	Yangon	Member	Female	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	8:33:00 PM	Ewallet	465.76	4.761905	23.2880	8.4
4	373-73-7910	Alex	Yangon	Member	Female	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37:00 AM	Ewallet	604.17	4.761905	30.2085	5.3

Gambar 6. Tampilan Dataset Supermarket Analysis

4. Memilih Variabel Penjualan

Pemilihan variabel yang akan digunakan dalam proses analisis, yaitu variabel Sales, yang merepresentasikan nilai penjualan pada setiap transaksi. Variabel ini dipilih karena menjadi indikator utama dalam pengelompokan pola penjualan menggunakan algoritma K-Means.

```
sales = df[['Sales']]
```

Gambar 7. Memilih Variabel Penjualan
Sumber: Data Penelitian 2026

Selanjutnya, dilakukan pemeriksaan struktur dan karakteristik dataset menggunakan fungsi info(), yang bertujuan untuk mengetahui jumlah data, jumlah atribut, tipe data masing-masing variabel, serta keberadaan nilai kosong (missing values). Hasil pemeriksaan menunjukkan bahwa dataset terdiri dari 1.000 data transaksi dengan 17 atribut, dan seluruh variabel memiliki nilai lengkap tanpa adanya data kosong. Kondisi ini menunjukkan bahwa dataset telah memenuhi syarat untuk dilanjutkan ke tahap analisis dan proses klasterisasi. Untuk lebih jelas dapat dilihat pada gambar 8.

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 17 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Invoice ID              1000 non-null   object
 1   Branch                  1000 non-null   object
 2   City                    1000 non-null   object
 3   Customer type           1000 non-null   object
 4   Gender                  1000 non-null   object
 5   Product line            1000 non-null   object
 6   Unit price              1000 non-null   float64
 7   Quantity                1000 non-null   int64
 8   Tax 5%                  1000 non-null   float64
 9   Sales                   1000 non-null   float64
 10  Date                    1000 non-null   object
 11  Time                    1000 non-null   object
 12  Payment                 1000 non-null   object
 13  cogs                    1000 non-null   float64
 14  gross margin percentage 1000 non-null   float64
 15  gross income            1000 non-null   float64
 16  Rating                  1000 non-null   float64
dtypes: float64(7), int64(1), object(9)
memory usage: 132.9+ KB
```

Gambar 8. Hasil Pemeriksaan Variabel
Sumber: Data Penelitian 2026

Berikut statistik deskriptif dari dataset penjualan supermarket pada gambar 9.

```
display(df.describe())
```

	Unit price	Quantity	Tax 5%	Sales	cogs	gross margin percentage	gross income	Rating
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000	1.000000e+03	1000.000000	1000.00000
mean	55.672130	5.510000	15.379369	322.966749	307.58738	4.761905e+00	15.379369	6.97270
std	26.494628	2.923431	11.708825	245.885335	234.17651	6.131498e-14	11.708825	1.71858
min	10.080000	1.000000	0.508500	10.678500	10.17000	4.761905e+00	0.508500	4.00000
25%	32.875000	3.000000	5.924875	124.422375	118.49750	4.761905e+00	5.924875	5.50000
50%	55.230000	5.000000	12.088000	253.848000	241.76000	4.761905e+00	12.088000	7.00000
75%	77.935000	8.000000	22.445250	471.350250	448.90500	4.761905e+00	22.445250	8.50000
max	99.960000	10.000000	49.650000	1042.650000	993.00000	4.761905e+00	49.650000	10.00000

Gambar 9. Statistik Penjualan
Sumber: Data Penelitian 2026

Pada gambar 9 Hasil analisis menunjukkan bahwa seluruh variabel numerik memiliki jumlah data sebanyak 1.000 transaksi, yang menandakan tidak adanya nilai kosong pada data. Rata-rata nilai penjualan (*Sales*) sebesar 322,97, dengan nilai minimum 10,68 dan maksimum 1.042,65, yang menunjukkan variasi nilai transaksi yang cukup tinggi. Variabel *Unit price* memiliki nilai rata-rata 55,67, sedangkan *Quantity* menunjukkan rata-rata pembelian sebanyak 5,51 unit per transaksi. Nilai *Tax 5%* dan *gross income* memiliki karakteristik sebaran yang serupa, sejalan dengan perhitungan pajak dan pendapatan kotor dari transaksi penjualan. Selain itu, variabel *Rating* memiliki nilai rata-rata 6,97, yang menunjukkan tingkat kepuasan pelanggan berada pada kategori cukup baik. Statistik deskriptif ini memberikan gambaran awal mengenai sebaran dan karakteristik data penjualan yang menjadi dasar dalam proses pengelompokan pola penjualan menggunakan algoritma K-Means.

5. Normalisasi Data

Proses normalisasi data penjualan menggunakan metode StandardScaler. Normalisasi bertujuan untuk mengubah skala data sehingga memiliki nilai rata-rata mendekati nol dan simpangan baku satu, sehingga perbedaan skala nilai tidak memengaruhi perhitungan jarak pada algoritma K-Means.

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
sales_scaled = scaler.fit_transform(sales)
```

Gambar 10. Normalisasi Data
Sumber: Data Penelitian 2026

Hasil normalisasi ditampilkan dalam bentuk nilai numerik terstandarisasi, di mana nilai positif menunjukkan penjualan di atas rata-rata, sedangkan nilai negatif menunjukkan penjualan di bawah rata-rata. sebagai berikut.

```
display(sales_scaled[:10])
```

```
array([[ 0.91960685],
       [-0.98772956],
       [ 0.07144605],
       [ 0.67577985],
       [ 1.26712548],
       [ 1.2396111 ],
       [ 0.45053787],
       [ 1.82864963],
       [-1.00430655],
       [-0.61124392]])
```

Gambar 11. Hasil Normalisasi
Sumber: Data Penelitian 2026

6. Visualisasi Distribusi Data Penjualan Setelah Normalisasi

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

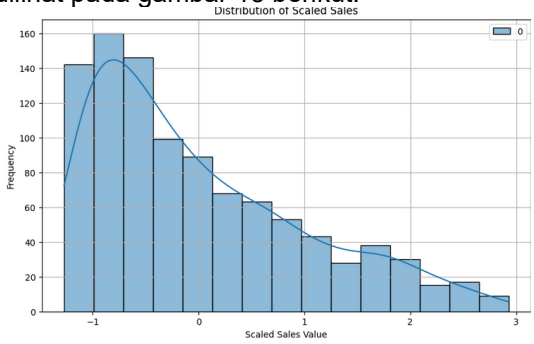
ISSN (Online) 2715-6265

Visualisasi disajikan dalam bentuk histogram yang dilengkapi dengan kurva kepadatan (*kernel density estimation*), dengan tujuan untuk mengamati pola sebaran nilai penjualan setelah proses standarisasi.

```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.histplot(sales_scaled, kde=True)
plt.title('Distribution of Scaled Sales')
plt.xlabel('Scaled Sales Value')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```

Gambar 12. Code Visualisasi Distribusi Data Penjualan
Sumber: Data Penelitian 2026

Berdasarkan grafik yang dihasilkan, data penjualan terdistribusi di sekitar nilai nol, yang menunjukkan bahwa proses normalisasi telah berjalan dengan baik. Distribusi data memperlihatkan variasi nilai penjualan dari yang berada di bawah rata-rata hingga di atas rata-rata, sehingga data siap digunakan dalam proses klasterisasi menggunakan algoritma K-Means tanpa adanya bias akibat perbedaan skala nilai. Lebih jelah dilihat pada gambar 13 berikut.



Gambar 13. Grafik Distribusi Data Penjualan
Sumber: Data Penelitian 2026

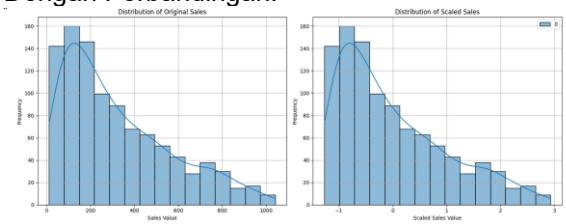
7. Visualisasi Perbandingan Distribusi Data Penjualan

Pada tahap ini dilakukan visualisasi perbandingan distribusi data penjualan sebelum dan sesudah dilakukan normalisasi.

```
import matplotlib.pyplot as plt
import seaborn as sns

fig, axes = plt.subplots(1, 2, figsize=(15, 6))

# Plot for original 'Sales' data
sns.histplot(df['Sales'], kde=True, ax=axes[0])
axes[0].set_title('Distribution of Original Sales')
axes[0].set_xlabel('Sales Value')
axes[0].set_ylabel('Frequency')
axes[0].grid(True)

# Plot for 'sales_scaled' data
sns.histplot(sales_scaled, kde=True, ax=axes[1])
axes[1].set_title('Distribution of Scaled Sales')
axes[1].set_xlabel('Scaled Sales Value')
axes[1].set_ylabel('Frequency')
axes[1].grid(True)

plt.tight_layout()
plt.show()
```

Gambar 14. Code Perbandingan Distribusi Data Penjualan
Sumber: Data Penjualan 2026

Dengan Perbandingan:



Gambar 15. Grafik Perbandingan Distribusi Data Penjualan Sebelum Dan Sesudah
Sumber: Data Penelitian 2026

Distribusi penjualan asli menunjukkan rentang nilai yang cukup lebar dengan kecenderungan sebaran tidak simetris, di mana sebagian besar transaksi berada pada nilai penjualan rendah hingga menengah dan hanya sebagian kecil yang memiliki nilai penjualan tinggi. Setelah dilakukan normalisasi, pola distribusi data tetap terjaga, namun skala nilai telah diubah sehingga data terpusat

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

di sekitar nilai nol dengan sebaran yang lebih proporsional. Kemudian dihitung kolerasi Penjualan dan Kuantitas sebagai berikut.

```
correlation = df['Sales'].corr(df['Quantity'])
print(f"Korelasi antara 'Sales' dan 'Quantity': {correlation}")
```

```
Korelasi antara 'Sales' dan 'Quantity': 0.7055101859433065
```

Gambar 16. Korelasi Penjualan Dan Kuantitas
Sumber: Data Penelitian 2026

Hasil perhitungan korelasi menunjukkan bahwa nilai koefisien korelasi antara variabel *Sales* dan *Quantity* sebesar 0,7055, yang mengindikasikan adanya hubungan positif yang kuat antara kedua variabel tersebut.

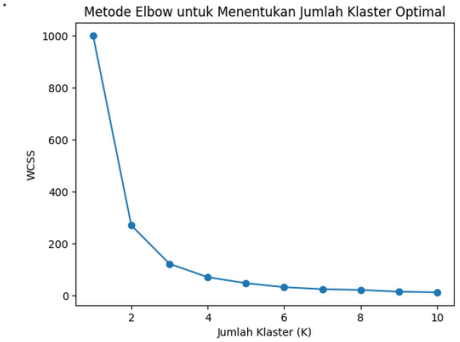
8. Penentuan Jumlah Kluster (Elbow Method)

penentuan jumlah klaster optimal menggunakan Metode Elbow. Proses ini dilakukan dengan menjalankan algoritma K-Means pada rentang jumlah klaster dari K = 1 hingga K = 10, kemudian menghitung nilai *Within Cluster Sum of Squares* (WCSS) pada setiap nilai K. WCSS digunakan untuk mengukur tingkat kekompakan data dalam setiap klaster, di mana nilai yang lebih kecil menunjukkan klaster yang lebih homogen.

```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

wcss = []

# Menentukan range jumlah klaster
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(sales_scaled)
    wcss.append(kmeans.inertia_)

# Visualisasi Elbow Method
plt.figure()
plt.plot(range(1, 11), wcss, marker='o')
plt.xlabel('Jumlah Klaster (K)')
plt.ylabel('WCSS')
plt.title('Metode Elbow untuk Menentukan Jumlah Klaster Optimal')
plt.show()
```

Gambar 17. Penentuan Jumlah Cluster
Sumber: Data Penelitian 2026

Hasil perhitungan WCSS kemudian divisualisasikan dalam bentuk grafik Elbow. Berdasarkan grafik yang dihasilkan, terlihat adanya titik siku (*elbow point*) pada nilai K = 3, yang menunjukkan bahwa penambahan jumlah klaster setelah titik tersebut tidak memberikan penurunan WCSS yang signifikan. Oleh karena itu, jumlah klaster optimal yang digunakan dalam penelitian ini ditetapkan sebanyak tiga klaster. Untuk lebih jelas dilihat pada gambar 18.



Gambar 18. Grafik Klaster
Sumber: Data Penelitian 2026

9. Algoritma K-Means

Pada tahap ini dilakukan proses klasterisasi data penjualan menggunakan algoritma K-Means dengan jumlah klaster sebanyak tiga klaster (K = 3), sesuai dengan hasil penentuan jumlah klaster optimal menggunakan Metode Elbow.

```
cluster_counts = df['Cluster'].value_counts()
print("Jumlah catatan di setiap cluster:")
print(cluster_counts)
```

```
Jumlah catatan di setiap cluster:
Cluster
2    515
0    314
1    171
Name: count, dtype: int64
```

Gambar 19. Proses Klasterisasi Algoritma K-Means

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

10. Hasil Akhir Clustering

Hasil akhir pengelompokan penjualan pada dataset supermarket dapat dilihat dibawah ini.

```
df[['Sales', 'Cluster', 'Kategori Penjualan']].head()
```

	Sales	Cluster	Kategori Penjualan
0	548.9715	0	Rata-rata
1	80.2200	2	Rendah
2	340.5255	0	Rata-rata
3	489.0480	0	Rata-rata
4	634.3785	1	Tinggi

```
display(cluster_analysis.describe())
```

	Unit price	Quantity	Tax 5%	Sales	cogs	gross margin percentage	gross income	Rating
count	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000
mean	62.953232	6.346552	20.465484	429.775173	409.309689	4.761905	20.465484	6.918535
std	21.164098	2.421155	14.953959	314.033144	299.079185	0.000000	14.953959	0.191292
min	41.777845	3.831068	6.266224	131.590710	125.324485	4.761905	6.266224	6.697661
25%	52.376836	5.189419	12.661064	265.882352	253.221287	4.761905	12.661064	6.862365
50%	62.975828	6.547771	19.055904	400.173994	381.118089	4.761905	19.055904	7.027070
75%	73.540926	7.604295	27.565115	578.867405	551.302290	4.761905	27.565115	7.028972
max	84.106023	8.660819	36.074325	757.568816	721.406491	4.761905	36.074325	7.030874

Gambar 20. Hasil Akhir Clustering
Sumber: Data Penelitian 2026

Pada tabel pertama ditampilkan beberapa contoh data yang memuat nilai Sales, label Cluster, serta Kategori Penjualan yang telah ditetapkan, yaitu rendah, rata-rata, dan tinggi. Hal ini menegaskan bahwa setiap transaksi berhasil dipetakan ke dalam klaster tertentu sesuai dengan karakteristik nilai penjualannya. Selanjutnya, statistik deskriptif pada tabel kedua menyajikan ringkasan nilai rata-rata, sebaran, serta rentang setiap variabel numerik untuk masing-masing klaster. Hasil ini memperlihatkan perbedaan yang jelas antar klaster, khususnya pada variabel penjualan, harga satuan, dan jumlah barang, sehingga dapat disimpulkan bahwa algoritma K-Means mampu mengelompokkan data penjualan supermarket secara efektif dan bermakna.

SIMPULAN

Dari hasil analisi Algoritma K-Means yang dilakukan terhadap dataset supermarket maka dapat disimpulkan:

1. Algoritma K-Means berhasil mengelompokkan data penjualan supermarket ke dalam tiga klaster utama, yaitu klaster penjualan tinggi, penjualan rata-rata, dan penjualan rendah. Pengelompokan ini didasarkan pada variabel penjualan (Sales) yang telah dinormalisasi, sehingga hasil klasterisasi mencerminkan perbedaan karakteristik transaksi secara nyata dan konsisten. Temuan ini menjawab rumusan masalah pertama bahwa K-Means dapat digunakan untuk mengidentifikasi kelompok produk atau transaksi dengan pola penjualan yang serupa, serta mendukung tujuan penelitian dalam menyediakan dasar analisis berbasis data untuk pengambilan keputusan manajerial di sektor ritel.
2. Implementasi K-Means dengan bahasa pemrograman Python, menggunakan pustaka seperti *pandas*, *scikit-learn*, dan *matplotlib*, mampu menghasilkan visualisasi yang jelas serta evaluasi klasterisasi yang baik. Hal ini dibuktikan melalui nilai Silhouette Score sebesar 0,6055, Davies–Bouldin Index sebesar 0,502, dan Inertia sebesar 122,01, yang secara keseluruhan menunjukkan kualitas klasterisasi yang cukup optimal, dengan klaster yang kompak dan terpisah dengan baik.

DAFTAR PUSTAKA

Akbar, A. A., Izzulhaq, A. B., Nursabila, N., & Hananto, V. R. (2023). Analisis Data Penjualan Pada Supermarket Xyz Menggunakan Metode Market Basket. *Jurnal Sistem Informasi Dan Informatika (Simika)*, *6*(2), 142–152.

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

https://doi.org/10.47080/simika.v6i2.2711

Alvianatinova, V., Ali, I., Rahaningsih, N., & Bahtiar, A. (2024). Penerapan Algoritma K-Means Clustering Dalam Pengelompokan Data Penjualan Supermarket Berdasarkan Cabang (Branch). *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(2), 1529–1535. https://doi.org/10.36040/jati.v8i2.8993

Amaliah, R., Tohidi, E., Wahyudin, E., Rizki Rinaldi, A., & Iin, I. (2024). Pengelompokkan Data Bencana Alam Berdasarkan Wilayah Menggunakan Algoritma K-Means. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(6), 3572–3579. https://doi.org/10.36040/jati.v7i6.8253

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(5), 1785–1792. https://doi.org/10.1016/j.jksuci.2019.12.011

Apriyanto, B., & Sitio, S. L. M. (2025). Penerapan K-Means dalam Menganalisis Pola Pembelian Pelanggan Pada Data Transaksi E-Commerce. *Bit-Tech*, *7*(3), 790–797. https://doi.org/10.32877/bt.v7i3.2195

Darmi, Y. D., & Setiawan, A. (2017). Penerapan Metode Clustering K-Means Dalam Pengelompokan Penjualan Produk. *Jurnal Media Infotama*, *12*(2), 148–157. https://doi.org/10.37676/jmi.v12i2.418

Fadel, D., Maulana, F., Harahap, A., Fauzan, I., Azriel, M., Fadlan, M., & Fansyuri, M. (2025). Implementasi Algoritma K-Means Clustering Data Penjualan Pada Warung Sembako Isan Menggunakan. *Journal of Information Technology and Informatics Engineering*, *1*(1), 7–11.

Fahrizal, F., Irawan, B., & Bahtiar, A. (2024). Analisis Produk Terlaris Dan Pengujian K-Means Untuk "Umkm Cetom." *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(3), 3055–3061. https://doi.org/10.36040/jati.v8i3.8379

Fajar Maulana Adji, M., & Dwilestari, G. (2024). Analisis Data Transaksi Penjualan Barang Menggunakan Teknik K-Means Clustering. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*(1), 619–625. https://doi.org/10.36040/jati.v9i1.12433

Harsono, S., Prihatin, T. D., Sadad, A., Kusrini, K., & Maulina, D. (2023). Penerapan Algoritma K-Means Untuk Pemetaan Biodiversity Kayu Bulat Di Indonesia. *CogITo Smart Journal*, *9*(1), 1–14. https://doi.org/10.31154/cogito.v9i1.402.1-14

Hendrastuty, N. (2024). Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa. *Jurnal Ilmiah Informatika Dan Ilmu Komputer (Jima-Ilkom)*, *3*(1), 46–56. https://doi.org/10.58602/jima-ilkom.v3i1.26

Hidayat, K., Adytama, Muhammad Rezky Darmawan, H. A., Arnando, Y., & Mukarim, A. (2025). Analisis Data Penjualan Menggunakan Algoritma K-Means Clustering Pada Toko Superindo. *Jurnal of Data Science*

Terbit *online* pada laman web jurnal : http://ejournal.upbatam.ac.id/index.php/comasiejournal

Jurnal Comasie

ISSN (Online) 2715-6265

Methods and Applications, *01*(01), 1–6. https://doi.org/10.30873/jodmapps.v1i1.pp1-6

Maulana, A., Danar Dana, R., & Dienwati Nuris, N. (2024). Implementasi Algoritma K-Means Clustering Dalam Pengelompokan Data Kerusakan Rumah Akibat Bencana Alam Di Kabupaten Cirebon. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(2), 1417–1424. https://doi.org/10.36040/jati.v8i2.9024

Muzakki, F., Ubaydillah, I., Assyiami, N. R., & Soleha, S. (2024). Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Jantung Menggunakan Rapidminer. *Jurnal Komputer Antartika*, *2*(2), 71–79. https://doi.org/10.70052/jka.v2i2.304

Rofiqo, N., Windarto, A. P., & Hartama, D. (2018). Penerapan Clustering Pada Penduduk Yang Mempunyai Keluhan Kesehatan Dengan Datamining K-Means. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *2*(1), 216–223. https://doi.org/10.30865/komik.v2i1.929

Siahaan, M. (2022). Data Mining Strategi Pembangunan Infrastruktur Menggunakan Algoritma K-Means. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *11*(3), 316–324. https://doi.org/10.32736/sisfokom.v11i3.1453

Sippan, R. B., & Setiyawati, N. (2025). *Pemetaan dan klasterisasi daerah rawan bencana alam di provinsi sulawesi tengah menggunakan k-means*. *10*(2), 1031–1045.

Sulastri, H., & Gufroni, A. I. (2017). Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *3*(2), 299–305. https://doi.org/10.25077/teknosi.v3i2.2017.299-305

Sulistio, M. R., Suarna, N., & Nurdiawan, O. (2023). Analisa Penerapan Metode Clustering X-Means Dalam Pengelompokan Penjualan Barang. *Jurnal Teknologi Ilmu Komputer*, *1*(2), 37–42. https://doi.org/10.56854/jtik.v1i2.49

Yudo Bismo Utomo, Iin Kurniasari, I. Y. (2023). Penerapan Knowledge Discovery in Database. *Jurnal Teknik Informatika Kaputama (JTIK)*, *7*(1).

	Rizky Prastya Wibowo, merupakan mahasiswa Prodi Sistem Informasi Universitas Putera Batam
	Erlin Elisa, S.Kom., M.Kom. Penulis kedua, Erlin Elisa, S.Kom., M.Kom, merupakan Dosen Prodi Sistem Informasi Universitas Putera Batam. Penulis banyak berkecimpung di bidang data mining dan data scientist di bidang sistem informasi.