INSTINK: (Jurnal Inovasi Pendidikan, Teknologi Informasi & Komputer)
Vol.4 No.1, April 2025
ISSN 2964-688X

Analisis Segmentasi Anggaran Pemasaran dan Penjualan Produk di Industri Retail Menggunakan K-Means Clustering Berbasis R Shiny

Anggelia Deli*[1], Petronela Kurniati Kondang[2], Wilnotus Daniel Awil[3], Astina Ranti[4]
[1,2,3,4]Program Studi Teknologi Informasi
[1,2,3,4]Institut Shanti Bhuana
e-mail: *[1]anggelia@shanti.bhuana.ac.id, [2]petronela@shantibhuana.ac.id,
[3]willnotusdanielawil@gmail.com, [4]astina@shantibhuana.ac.id

Abstrak

Penelitian ini bertujuan untuk menganalisis pola hubungan antara anggaran pemasaran dan penjualan produk di industri retail melalui segmentasi data menggunakan metode K-Means Clustering, baik secara manual menggunakan Excel maupun otomatis melalui aplikasi R Shiny. Topik ini diangkat karena pentingnya efektivitas alokasi anggaran pemasaran dalam meningkatkan kinerja penjualan di tengah persaingan sektor retail yang semakin ketat. Data yang digunakan sebanyak 68.619 titik data yang kemudian diolah untuk menentukan jumlah klaster dan centroid secara acak. Hasil penelitian menunjukkan terbentuknya tiga klaster dengan distribusi yang signifikan, di mana klaster ketiga memiliki jumlah data terbanyak dan rata-rata harga tertinggi setelah proses scaling. Evaluasi hasil clustering melalui nilai Within Cluster Sum of Squares (WCSS) dan rasio Between_SS/Total_SS sebesar 55,7% mengindikasikan kualitas segmentasi yang cukup baik. Hasil dari perhitungan manual dan otomatis menunjukkan kesamaan yang signifikan, sehingga dapat disimpulkan bahwa metode K-Means efektif dan efisien dalam mengelola data berskala besar serta konsisten dalam mendukung analisis segmentasi anggaran pemasaran terhadap penjualan produk.

Kata kunci— K-Means, Shiny R, Anggaran Pemasaran, Penjualan Produk, Clustering

Abstract

This study aims to analyze the relationship pattern between marketing budget and product sales in the retail industry through data segmentation using the K-Means Clustering method, both manually using Excel and automatically via the R Shiny application. This topic was chosen due to the critical importance of effective marketing budget allocation in enhancing sales performance amid the increasingly competitive retail sector. The dataset comprises 68,619 data points, which were processed to determine the number of clusters and randomly assigned centroids. The results revealed the formation of three significantly distributed clusters, with the third cluster containing the largest number of data points and the highest average price after scaling. The evaluation of clustering performance through the Within Cluster Sum of Squares (WCSS) and a Between_SS/Total_SS ratio of 55.7% indicates a reasonably good segmentation quality. The comparison between manual and automated computations showed a significant similarity, leading to the conclusion that the K-Means method is both effective and efficient in handling large-scale data and consistent in supporting marketing budget segmentation analysis related to product sales.

Keywords— K-Means, Shiny R, Marketing Budget, Product Sales, Clustering

1. PENDAHULUAN

Industri retail merupakan salah satu sektor bisnis yang berperan penting dalam perekonomian, terutama karena langsung berinteraksi dengan konsumen akhir tanpa melalui perantara. Produk-produk yang diperdagangkan meliputi barang kebutuhan pokok hingga gaya hidup seperti pakaian, elektronik, dan perlengkapan rumah tangga. Seiring perkembangan zaman, pertumbuhan ritel modern di Indonesia semakin pesat akibat perubahan gaya hidup, peningkatan pendapatan, serta pergeseran preferensi konsumen terhadap kenyamanan dan kemudahan berbelanja [1].

Dalam menjalankan bisnis retail, salah satu aspek krusial yang memengaruhi performa penjualan adalah strategi pemasaran. Pemasaran mencakup upaya mengenalkan produk kepada konsumen melalui berbagai saluran, baik offline maupun digital. Namun, upaya pemasaran ini membutuhkan biaya atau anggaran yang tidak sedikit, mencakup promosi, iklan, diskon, hingga biaya tenaga pemasaran. Oleh karena itu, perusahaan perlu mengelola anggaran pemasaran secara efektif agar dapat mendorong peningkatan penjualan tanpa membuang sumber daya [2].

Meskipun demikian, belum semua pelaku bisnis memahami bagaimana pola anggaran pemasaran berkaitan dengan hasil penjualan secara keseluruhan. Oleh karena itu, diperlukan pendekatan analitik yang mampu mengelompokkan atau melakukan segmentasi data berdasarkan karakteristik tertentu. Salah satu pendekatan yang dapat digunakan adalah data mining, khususnya metode K-Means Clustering. Metode ini efektif untuk mengelompokkan data dalam jumlah besar berdasarkan kemiripan karakteristik tertentu, dalam hal ini antara nilai anggaran pemasaran dan penjualan produk [3].

Beberapa penelitian sebelumnya telah membahas penggunaan metode K-Means Clustering dalam analisis data penjualan dan preferensi konsumen. Misalnya, Prasetiyo, Lestari, dan Atima (2024) menerapkan metode K-Means untuk mengelompokkan menu favorit pelanggan di Tetra Coffeeshop, yang menunjukkan efektivitas metode ini dalam mengidentifikasi pola tersembunyi dalam data preferensi konsumen [4]. Penelitian lain oleh Muhaimin et al. (2024) mengombinasikan metode Rank Order Centroid (ROC) dan Simple Additive Weighting (SAW) dalam sistem pendukung keputusan untuk pemilihan kafe terbaik, menunjukkan efektivitas kombinasi metode ini dalam memproses data berskala besar secara efisien [5].

Namun, sebagian besar studi tersebut belum mengintegrasikan analisis anggaran pemasaran secara langsung, khususnya dalam konteks industri retail. Oleh karena itu, penelitian ini menawarkan kontribusi baru dengan mengaitkan segmentasi anggaran pemasaran dan penjualan produk menggunakan K-Means berbasis aplikasi R Shiny sebagai alat bantu visualisasi dan analisis yang lebih interaktif.

Penelitian ini menggunakan dataset retail berskala besar yang berisi informasi penjualan produk seperti pakaian, sepatu, dan buku, dengan total 68.619 titik data. Untuk pengolahan dan analisis data, digunakan algoritma K-Means Clustering yang diimplementasikan secara manual menggunakan Microsoft Excel dan secara otomatis menggunakan R Shiny, sebuah aplikasi berbasis web yang memungkinkan visualisasi data interaktif secara real time. Penggunaan R Shiny memberikan keunggulan dalam hal efisiensi, otomatisasi, dan kemudahan akses lintas platform tanpa perlu instalasi perangkat lunak tambahan [6].

Tujuan utama dari penelitian ini adalah untuk melakukan analisis segmentasi antara anggaran pemasaran dan penjualan produk dengan menggunakan metode K-Means Clustering. Melalui segmentasi ini, diharapkan pelaku usaha dapat memperoleh gambaran yang lebih jelas tentang pola distribusi anggaran dan hasil penjualan dalam kelompok-kelompok tertentu. Hasil ini diharapkan dapat menjadi dasar dalam pengambilan keputusan strategis, khususnya dalam mengalokasikan anggaran pemasaran secara lebih efektif di masa mendatang.

2. METODE PENELITIAN

Algoritma K-Means Clustering merupakan salah satu metode dalam analisis data dan data mining yang digunakan untuk mengelompokkan data berdasarkan tingkat kemiripannya tanpa memerlukan label atau panduan tertentu (unsupervised learning). Algoritma ini berfungsi untuk membagi data ke dalam beberapa kelompok yang memiliki kesamaan karakteristik. Proses pengelompokan dilakukan dengan menghitung jarak antara setiap data dengan titik pusat cluster (centroid), yang digunakan sebagai acuan untuk menentukan tingkat kemiripan antar data. Tujuan utama dari metode ini adalah mengurangi nilai fungsi objektif, yakni meminimalkan variasi data di dalam cluster yang sama, sekaligus memaksimalkan perbedaan data antar cluster[7].

Tahapan dalam perhitungan Algoritma K-Means Clustering Analysis adalah:

1. Menentukan jumlah cluster (k).
2. Memilih centroid awal secara acak.
3. Mengelompokkan data berdasarkan jarak terdekat ke centroid.
4. Menghitung ulang centroid berdasarkan rata-rata dalam cluster.
5. Mengulangi langkah 3–4 hingga centroid tidak berubah lagi (konvergen).

Rumus K-Means yaitu:

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

Keterangan:

$D(x,y)$: jarak antara data pada titik x dan y

x: titik data objek

y: titik data centroid

i: jumlah atribut data

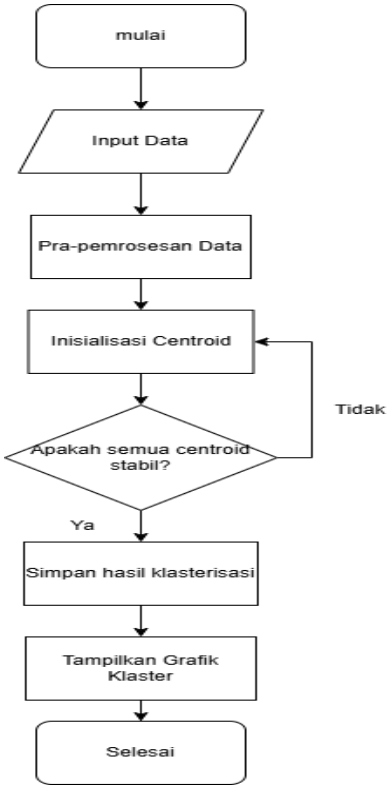
2.1 Pengumpulan Data

Penelitian ini menggunakan pendekatan kuantitatif untuk menganalisis hubungan antara anggaran pemasaran dan penjualan produk di industri retail. Data diolah menggunakan metode data mining dengan algoritma K-Means Clustering dan divisualisasikan dengan menggunakan aplikasi Shiny R. Data penjualan produk diambil dari Kaggle dataset yang berisi informasi penjualan berbagai jenis produk retail.

2.2 Flowchart

Algoritma flowchart adalah sebagai berikut:

1. Start (mulai): Proses dimulai dengan menjalankan algoritma K-Means untuk mengelompokkan data.
2. Input Data: Data dimasukkan dalam bentuk format CSV, atau dataset mentah.
3. Pra-pemrosesan Data: Membersihkan dan menyiapkan data, termasuk pemilihan kolom relevan dan normalisasi agar berada dalam rentang nilai yang sama. Hasil normalisasi ditampilkan melalui boxplot.
4. Inisialisasi Centroid: Pemilihan centroid awal dilakukan secara acak atau berdasarkan kebutuhan analisis.
5. Assign Data to Cluster: Menghitung jarak antara data dan centroid, lalu menetapkan data ke cluster terdekat.
6. Keputusan (diamond): Memeriksa apakah posisi centroid berubah. Jika berubah, ulangi proses clustering; jika tidak, lanjut ke tahap berikutnya.
7. Simpan Hasil Clustering: Menyimpan hasil klasterisasi setelah mencapai konvergensi untuk dianalisis lebih lanjut.
8. Visualisasi: Menampilkan hasil visualisasi dalam bentuk grafik atau diagram.
9. End: Proses algoritma K-Means selesai dan hasil akhir siap digunakan.

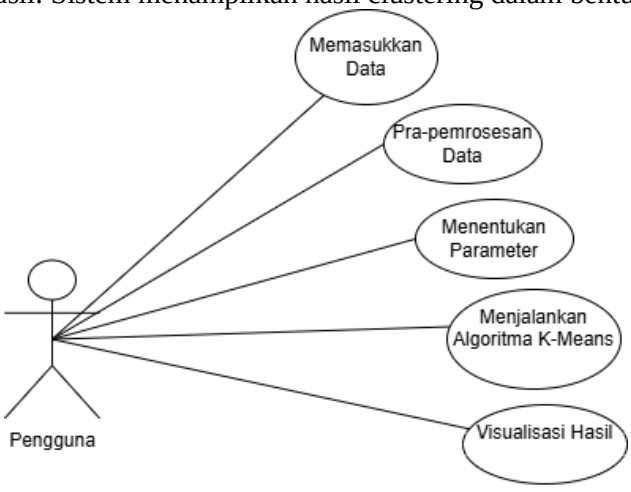


Gambar 1. Flowchart

2.3 Use Case

Use case di atas melibatkan:

1. Pengguna: Aktor utama sebagai peneliti yang menganalisis data menggunakan K-Means.
2. Memasukkan Data: Pengguna mengunggah dataset untuk diproses.
3. Pra-pemrosesan Data: Sistem melakukan normalisasi dan pembersihan data.
4. Menentukan Parameter: Pengguna memilih jumlah cluster (k).
5. Menjalankan Algoritma K-Means: Sistem menjalankan proses clustering.
6. Visualisasi Hasil: Sistem menampilkan hasil clustering dalam bentuk visual.



Gambar 2. Use Case Diagram

2.4 K-Means

K-Means merupakan algoritma unsupervised learning yang berfungsi mengelompokkan data ke dalam beberapa cluster. Algoritma ini bekerja tanpa label kategori (non-hierarki) dan mengelompokkan data yang memiliki kemiripan karakteristik. Tujuan dari clustering adalah

mengelompokkan data dalam satu kelompok yang seragam serta berbeda dengan kelompok lainnya [8].

Clustering juga berkaitan dengan konsep cluster sampling, yaitu teknik pengambilan sampel secara acak dari kelompok yang telah ada (cluster), yang diterapkan dalam konteks data mining untuk menganalisis data besar tanpa pengawasan langsung.

2.5 R Shiny

R Shiny merupakan framework yang dikembangkan oleh RStudio untuk membangun aplikasi web interaktif langsung dari bahasa pemrograman R. Dengan R Shiny, pengguna dapat membuat dashboard visualisasi data secara real-time, menjadikan eksplorasi data lebih efisien dan mudah. Keunggulannya terletak pada kemampuan lintas platform tanpa memerlukan instalasi perangkat lunak tambahan, sehingga sangat sesuai untuk kebutuhan analisis data yang bersifat interaktif dan dinamis [9].

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Data penelitian ini diperoleh dari *Customer Shopping Dataset* yang tersedia di platform Kaggle (https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset). Dataset tersebut dipercaya valid dan dapat diandalkan, dengan total jumlah data sebanyak 99.458 entri yang merepresentasikan transaksi penjualan di sektor ritel.

Dataset ini terdiri dari 11 atribut, yaitu:

1. No: Nomor urut transaksi.
2. Invoice_no: Nomor faktur atau nota pembelian.
3. Customer_id: ID unik pelanggan.
4. Gender: Jenis kelamin pelanggan (Female/Male).
5. Age: Usia pelanggan.
6. Category: Kategori produk yang dibeli (seperti *Clothing*, *Shoes*, *Books*, *Cosmetics*, *Souvenir*, dan lainnya).
7. Quantity: Jumlah barang yang dibeli.
8. Price: Total harga dalam transaksi.
9. Payment_method: Metode pembayaran yang digunakan (seperti *Credit Card*, *Debit Card*, atau *Cash*).
10. Invoice_date: Tanggal pembelian dilakukan.
11. Shopping_mall: Nama pusat perbelanjaan tempat transaksi terjadi.

Dataset ini digunakan sebagai basis untuk melakukan analisis segmentasi dengan algoritma K-Means Clustering, yang kemudian divisualisasikan melalui aplikasi R Shiny.

no	invoice_no	customer_id	gender	age	category	quantity	price	payment_method	invoice_date	shopping_mall
1	I138884	C241288	Female	28	Clothing	5	1500.4	Credit Card	5/8/2022	Kanyon
2	I317333	C111565	Male	21	Shoes	3	1800.51	Debit Card	12/12/2021	Forum Istanbul
3	I127801	C266599	Male	20	Clothing	1	300.08	Cash	9/11/2021	Metrocity
4	I173702	C988172	Female	66	Shoes	5	3000.85	Credit Card	16/05/2021	Metropol AVM
5	I337046	C189076	Female	53	Books	4	60.6	Cash	24/10/2021	Kanyon
6	I227836	C657758	Female	28	Clothing	5	1500.4	Credit Card	24/05/2022	Forum Istanbul
7	I121056	C151197	Female	49	Cosmetics	1	40.66	Cash	13/03/2022	Istinye Park
8	I293112	C176086	Female	32	Clothing	2	600.16	Credit Card	13/01/2021	Mall of Istanbul
...	I714965	C317224	Female	58	Clothing	3	900.24	Cash	7/10/2021	Mall of Istanbul
99458	I232867	C273973	Female	36	Souvenir	3	35.19	Credit Card	15/10/2022	Mall of Istanbul

Gambar 3. Dataset

3.2 Preprocessing Data

Tahap preprocessing data dilakukan untuk memastikan bahwa data yang digunakan memenuhi asumsi distribusi yang diperlukan untuk proses klasterisasi. Proses ini dilakukan menggunakan aplikasi RStudio, dengan bantuan framework Shiny R. Dari 11 atribut yang tersedia dalam dataset, penelitian ini hanya menggunakan 3 atribut, yaitu: age, category, dan price, sesuai dengan tujuan segmentasi yang difokuskan pada karakteristik usia pelanggan, jenis produk, dan nilai transaksi.

Langkah pertama dalam preprocessing adalah memuat dataset. Proses ini dapat dilakukan secara manual atau melalui kode program. Berikut adalah sintaks untuk memuat data secara langsung:

```
data <- read.csv("C:/Users/NETY/Documents/data1.csv")
```

Letak file harus disesuaikan dengan direktori masing-masing pengguna. Untuk menampilkan data yang telah dimuat ke dalam tampilan tabular di RStudio, digunakan perintah:

```
View(data1)
```

Selanjutnya dilakukan eksplorasi awal terhadap data untuk memperoleh ringkasan statistik dari setiap kolom. Perintah yang digunakan adalah:

```
summary(data1)
```

Output dari perintah tersebut menunjukkan bahwa dataset memiliki 99.457 entri untuk seluruh kolom. Kolom invoice_no, customer_id, gender, category, payment_method, invoice_date, dan shopping_mall memiliki tipe data karakter. Sementara kolom numerik menunjukkan bahwa usia pelanggan berkisar antara 18 hingga 69 tahun, dengan rata-rata sebesar 43,43 tahun. Jumlah barang yang dibeli per transaksi berkisar antara 1 hingga 5 unit, dengan rata-rata 3 unit. Nilai transaksi berada dalam rentang Rp5,23 hingga Rp5.250,00, dengan median sebesar Rp203,30 dan rata-rata Rp689,26.

```
> summary(data)
  invoice_no        customer_id           gender              age        
 Length:99457       Length:99457       Length:99457       Min.   :18.00  
 Class :character   Class :character   Class :character   1st Qu.:30.00  
 Mode  :character   Mode  :character   Mode  :character   Median :43.00  
                                                          Mean   :43.43  
                                                          3rd Qu.:56.00  
                                                          Max.   :69.00  
   category           quantity         price          payment_method    
 Length:99457       Min.   :1.000   Min.   :   5.23   Length:99457      
 Class :character   1st Qu.:2.000   1st Qu.:  45.45   Class :character  
 Mode  :character   Median :3.000   Median : 203.30   Mode  :character  
                    Mean   :3.003   Mean   : 689.26                     
                    3rd Qu.:4.000   3rd Qu.:1200.32                     
                    Max.   :5.000   Max.   :5250.00                     
 invoice_date       shopping_mall     
 Length:99457       Length:99457      
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

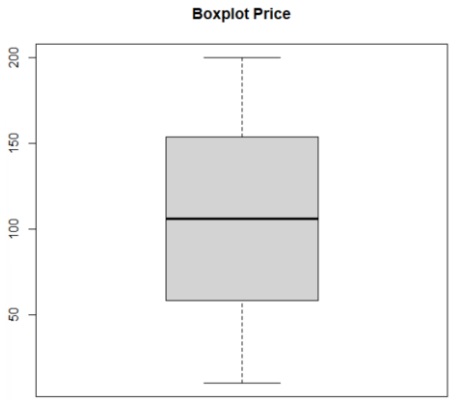
Gambar 4. Output dari perintah

Langka selanjutnya yaitu membuat *boxplot*. Yang akan di *boxplot* hanya dua kolom dari ketiga kolom yang di butuhkan, karena satu kolom yang di butuhkan akan di clusteringkan dalam bentuk teks. Yang pertama kolom *price*, berikut adalah *code r* nya,

```
boxplot(data1$price, main="Boxplot price")
```

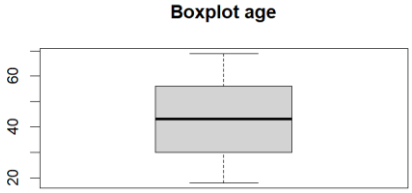
Karena, hasil nya belum terdistribusi normal dan hasilnya lebih turun, maka akan dikuadratkan, code r untuk kuadratnya adalah berikut;

```
data1$price_sqrt <- data1$price^0.1
boxplot(data1$price_sqrt, main="Boxplot price")
```



Gambar 5. Boxplot Price

```
boxplot(data1$age, main="Boxplot age")
str(data$age)
```



Gambar 6. Boxplot age

Boxplot adalah representasi grafis dari distribusi data numerik yang merangkum lima ukuran statistik utama yaitu, minimum, kuartil pertama (Q1), median (Q2), kuartil ketiga (Q3), dan maksimum. *Boxplot* digunakan untuk menunjukkan persebaran, kecendrungan sentral, dan potensi pencilan *(outliers)* dalam suatu *dataset.*
Gambar 5 dan gambar 6 di atas menjelaskan bahwa data *price* dan data *age* sudah terdistribusi normal. *Boxplot* tidak memiliki pencilan atau *outliers*, yang di tunjukan dengan tidak adanya titik atau lingkaran di luar *whiskers.* Dengan demikian tidak ada nilai yang jauh di luar rentang yang dianggap normal.

Setelah semua data yang akan di gunakan untuk di *cluster* sudah terdistribusi normal, dari banyaknya data perlu juga mengambil data sampel yang akan di clusteringkan, data sampelnya yaitu sebagai berikut.

no	age	category	price_sqrt	price_category	age_category
1	28	Clothing	2.077881	Tinggi	Muda
2	62	Shoes	2.032031	Tinggi	Tua
3	68	Clothing	2.077881	Tinggi	Tua
4	44	Cosmetics	1.448494	Rendah	Dewasa
5	43	Clothing	2.077881	Tinggi	Dewasa
6	24	Clothing	2.077881	Tinggi	Muda
7	58	Clothing	2.077881	Tinggi	Tua
8	40	Food & Beverage	1.179913	Rendah	Dewasa
...	59	Books	1.507465	Rendah	Tua
69620	58	Clothing	1.974403	Tinggi	Tua

Gambar 7. Data Sampel yang akan di Clusteringkan

3.3 Clustering K-Means dengan Perhitungan Manual

Perhitungan manual dilakukan menggunakan Microsoft Excel. Tahapan perhitungan dimulai dengan menentukan jumlah cluster yang diinginkan, kemudian memilih nilai centroid

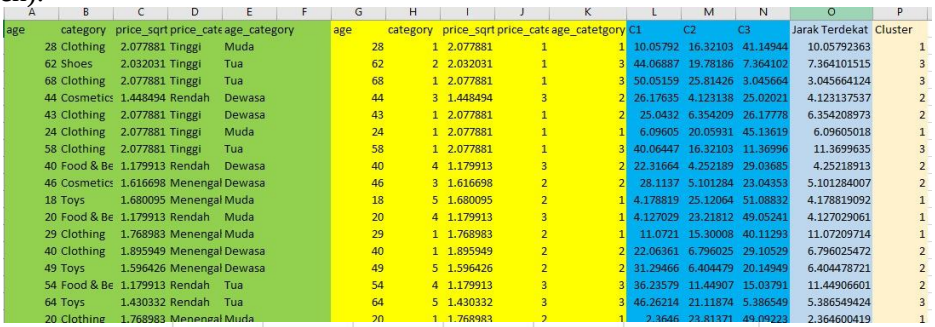
secara acak. Nilai centroid awal yang digunakan dapat dilihat pada Gambar X (centroid), sedangkan hasil akhir dari proses clustering manual ditunjukkan pada Gambar Y (hasil clustering).

Sebelum proses perhitungan dilakukan, data kategorikal terlebih dahulu diubah ke dalam bentuk numerik untuk memudahkan proses perhitungan jarak. Setelah itu, perhitungan dilakukan secara berurutan (baris demi baris) menggunakan rumus Euclidean distance dalam Microsoft Excel.

Perhitungan ini mencakup:

1. Penentuan jarak setiap data terhadap masing-masing centroid (C1, C2, C3),
2. Pemilihan jarak terdekat dari setiap data ke centroid tertentu,
3. Penetapan cluster berdasarkan jarak terdekat tersebut.

Langkah-langkah ini diulangi hingga centroid tidak mengalami perubahan nilai (konvergen).



age	category	price_sqrt	price_cate	age_category	age	category	price_sqrt	price_cate	age_categgory	C1	C2	C3	Jarak Terdekat	Cluster
28	Clothing	2.077881	Tinggi	Muda	28	1	2.077881	1	1	10.05792	16.32103	41.14944	10.05792363	1
62	Shoes	2.032031	Tinggi	Tua	62	2	2.032031	1	3	44.06887	19.78186	7.364102	7.364101515	3
68	Clothing	2.077881	Tinggi	Tua	68	1	2.077881	1	3	50.05159	25.81426	3.045664	3.045664124	3
44	Cosmetics	1.448494	Rendah	Dewasa	44	3	1.448494	3	2	26.17635	4.123138	25.02021	4.123137537	2
43	Clothing	2.077881	Tinggi	Dewasa	43	1	2.077881	1	2	25.0432	6.354209	26.17778	6.354208973	2
24	Clothing	2.077881	Tinggi	Muda	24	1	2.077881	1	1	6.09605	20.05931	45.13619	6.09605018	1
58	Clothing	2.077881	Tinggi	Tua	58	1	2.077881	1	3	40.06447	16.32103	11.36996	11.3699635	3
40	Food & Be	1.179913	Rendah	Dewasa	40	4	1.179913	3	2	22.31664	4.252189	29.03685	4.25218913	2
46	Cosmetics	1.616698	Menengal	Dewasa	46	3	1.616698	2	2	28.1137	5.101284	23.04353	5.101284007	2
18	Toys	1.680095	Menengal	Muda	18	5	1.680095	2	1	4.178819	25.12064	51.08832	4.178819092	1
20	Food & Be	1.179913	Rendah	Muda	20	4	1.179913	3	1	4.127029	23.21812	49.05241	4.127029061	1
29	Clothing	1.768983	Menengal	Muda	29	1	1.768983	2	1	11.0721	15.30008	40.11293	11.07209714	1
40	Clothing	1.895949	Menengal	Dewasa	40	1	1.895949	2	2	22.06361	6.796025	29.10529	6.796025472	2
49	Toys	1.596426	Menengal	Dewasa	49	5	1.596426	2	2	31.29466	6.404479	20.14949	6.404478721	2
54	Food & Be	1.179913	Rendah	Tua	54	4	1.179913	3	3	36.23579	11.44907	15.03791	11.44906601	2
64	Toys	1.430332	Rendah	Tua	64	5	1.430332	3	3	46.26214	21.11874	5.386549	5.386549424	3
20	Clothing	1.768983	Menengal	Muda	20	1	1.768983	2	1	2.3646	23.81371	49.09223	2.364600419	1

Centroid	age	category	price_sqrt	price_cate	age_category
C1	18	1	1.974403	1	1
C2	43	7	1.464716	3	2
C3	69	3	1.552458	3	3

Gambar 8. Perhitungan Manual Clustering dengan Excel

3.4 Clustering K-Means Menggunakan R Shiny

Dalam proses clustering menggunakan metode K-Means, data yang telah diambil sebagai sampel terlebih dahulu diubah ke dalam format numerik. Tiga kolom yang akan digunakan dalam proses klasterisasi dipastikan seluruhnya berupa data numerik. Proses ini bertujuan agar algoritma K-Means dapat mengukur jarak antar data secara valid.

Berikut adalah cuplikan kode R yang digunakan untuk memilih kolom numerik dan melakukan penskalaan data sebelum dilakukan proses clustering:

Memilih hanya kolom numerik

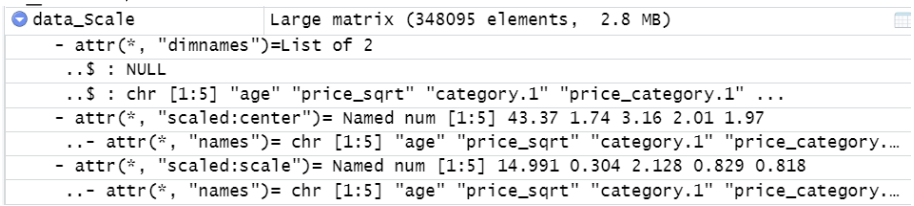
```
data_numeric <- data[sapply(data, is.numeric)]
```

Menskalakan data numerik

```
data_Scale <- scale(data_numeric)
```

Menampilkan data hasil penskalaan

```
head(data_Scale)
```



```
data_Scale          Large matrix (348095 elements,  2.8 MB)
  - attr(*, "dimnames")=List of 2
   ..$ : NULL
   ..$ : chr [1:5] "age" "price_sqrt" "category.1" "price_category.1" ...
  - attr(*, "scaled:center")= Named num [1:5] 43.37 1.74 3.16 2.01 1.97
   ..- attr(*, "names")= chr [1:5] "age" "price_sqrt" "category.1" "price_category.…
  - attr(*, "scaled:scale")= Named num [1:5] 14.991 0.304 2.128 0.829 0.818
   ..- attr(*, "names")= chr [1:5] "age" "price_sqrt" "category.1" "price_category.…
```

Gambar 9. Isi Data_Scale

Selanjutnya, dilakukan pemilihan terhadap kolom-kolom yang akan digunakan dalam proses klasterisasi. Dalam hal ini, dipilih tiga variabel yaitu price_sqrt, age_category, dan category.1 yang telah dikonversi menjadi bentuk numerik.

Proses seleksi kolom dilakukan dengan mengambil subset dari data utama menggunakan perintah berikut:

Memilih kolom yang digunakan untuk clustering

```
data_3cols <- data[, c("price_sqrt", "age_category", "category.1")]
```

Setelah kolom dipilih, dilakukan verifikasi bahwa seluruh kolom yang dipilih adalah bertipe numerik:

Memastikan hanya kolom numerik

```
data_3cols <- data_3cols[sapply(data_3cols, is.numeric)]
```

Tahap berikutnya adalah proses penskalaan (normalisasi) data numerik menggunakan fungsi scale(), agar setiap variabel memiliki kontribusi yang seimbang terhadap proses clustering:

Menskalakan data numerik

```
data_3cols_scaled <- scale(data_3cols)
```

Penskalaan ini bertujuan untuk menghindari bias akibat perbedaan satuan antar variabel. Misalnya, variabel usia dapat memiliki rentang puluhan tahun, sedangkan kategori hanya memiliki nilai diskrit dari 1 hingga 5. Dengan normalisasi, seluruh variabel memiliki rata-rata 0 dan standar deviasi 1, sehingga metode K-Means dapat mengelompokkan data secara lebih akurat.

```
data_3cols                69619 obs. of 3 variables
    $ price_sqrt    : num  2.08 2.03 2.08 1.45 2.08 ...
    $ age_catetgory: int  1 3 3 2 2 1 3 2 2 1 ...
    $ category.1    : int  1 2 1 3 1 1 1 4 3 5 ...
```

Gambar 10. Data_3cols

Memeriksa hasil scaling

```
head(data_3cols_scaled)
```

```
> head(data_3cols_scaled)
     price_sqrt age_catetgory  category.1
[1,]  1.1207979   -1.19159966 -1.01516175
[2,]  0.9701392    1.25326652 -0.54530119
[3,]  1.1207979    1.25326652 -1.01516175
[4,] -0.9473199    0.03083343 -0.07544063
[5,]  1.1207979    0.03083343 -1.01516175
[6,]  1.1207979   -1.19159966 -1.01516175
>
```

Gambar 11. Output Dari Perintah R

Mengatur seed untuk hasil yang konsisten

```
set.seed(123)
```

Melakukan clustering menggunakan k-means

```
km.out <- kmeans(data_3cols_scaled, centers = 3, nstart = 20)
```

Menampilkan hasil clustering

```
print(km.out)
```

```
> print(km.out)
K-means clustering with 3 clusters of sizes 14943, 25862, 28814

Cluster means:
  price_sqrt age_catetgory category.1
1  0.6360136   -1.1915996600 -0.5667457
2  0.7255476    0.6875686137 -0.6183910
3 -0.9810531    0.0008389753  0.8489522

Clustering vector:
    [1] 1 2 2 3 2 1 2 3 3 3 3 1 2 3 3 3 1 2 1 2 2 3 3 2 1 1 2 2 3 3 3 1 3 3 2
   [36] 3 1 2 2 2 2 3 1 1 1 3 3 3 1 2 2 3 2 1 3 2 2 3 2 1 2 2 2 3 3 3 3 3 3 2
   [71] 2 2 2 2 1 3 1 2 1 3 1 3 2 2 3 3 3 2 3 3 3 2 1 3 3 2 3 1 3 2 2 1 2 1 3 2 3
  [106] 3 3 2 2 3 2 2 3 2 2 1 3 3 2 2 3 3 1 2 2 3 3 1 1 2 2 3 3 2 2 3 3 1 2 3 3
  [141] 3 2 3 3 1 2 2 2 3 1 1 3 1 1 3 3 2 2 3 3 2 1 3 3 2 2 3 3 2 2 2 3 3 1 2 3
  [176] 2 2 1 1 2 2 3 2 2 3 3 3 3 2 2 3 2 3 3 1 2 1 2 2 3 2 2 2 2 2 2 3 2 2 3 1
  [211] 3 3 3 1 2 3 3 1 2 2 2 1 1 2 2 2 1 1 1 3 3 1 1 3 2 3 3 3 1 2 3 3 3 2 3 3
  [246] 3 2 1 3 2 2 2 3 2 1 1 2 2 2 2 2 1 2 3 3 1 2 2 1 3 2 2 2 3 2 2 2 2 2 3 2 2 3 3 3
  [281] 1 1 2 2 3 2 2 3 2 2 2 3 3 3 2 2 2 3 3 3 2 2 2 1 3 3 2 2 2 2 2 3 3 1 1 1
  [316] 3 1 1 3 1 3 2 3 3 2 2 3 1 2 2 1 3 3 2 1 3 3 1 3 3 2 2 2 3 3 1 1 2 2 2 2
  [351] 2 1 3 3 2 2 2 2 3 2 2 3 3 2 2 2 3 3 2 1 2 2 1 3 2 3 3 2 3 3 1 3 3 3 3 2 1 1 3 3
  [386] 2 2 2 2 1 2 3 2 2 3 2 2 2 3 3 3 2 2 3 3 2 3 3 1 2 2 1 1 2 2 3 3 1 2 2 3 3
  [421] 1 2 3 3 1 2 2 3 3 2 2 2 1 1 3 3 3 3 1 1 2 2 2 2 3 1 1 3 2 3 3 3 2 2 3 1 3 2 2
  [456] 3 2 1 1 2 2 2 3 1 1 3 3 3 3 3 1 3 2 2 2 1 2 2 2 2 2 1 3 1 1 2 2 3 3 3 3 2 2 2 3 1 2
  [491] 3 3 2 2 1 3 3 3 2 3 2 2 1 1 1 1 2 2 2 3 2 2 2 1 2 2 1 3 2 2 3 3 3 3 3 1 2 3
  [526] 3 3 2 1 3 3 3 2 2 2 3 2 2 2 2 3 1 1 1 3 3 3 1 1 3 3 2 2 2 3 3 3 3 2 3 1 1 2
  [561] 3 2 1 3 3 3 3 2 2 2 3 2 2 2 2 1 3 3 2 2 1 3 3 1 2 2 1 3 2 2 1 1 2 2 3 3 3 1 2 1 3
  [596] 3 2 2 1 2 3 3 3 3 1 1 3 2 2 2 1 1 1 2 2 3 3 1 1 3 3 2 2 3 3 3 2 3 3 2 3
  [631] 2 2 2 2 1 3 2 3 3 2 2 3 1 1 3 2 2 3 2 2 2 3 3 3 2 2 1 1 3 3 1 2 2 1 1 2 3 3 3 3
  [666] 1 1 1 3 3 3 3 2 3 3 1 3 1 1 1 2 2 2 3 2 2 1 3 3 2 3 3 2 2 1 3 2 2 1 1 3 3 3 3 2
  [701] 2 1 2 2 3 1 1 3 3 1 3 2 2 1 3 2 2 2 2 2 3 3 3 3 1 2 3 3 3 3 3 3 3 3 1 3
  [736] 2 3 2 2 2 3 1 1 3 3 3 2 3 1 1 3 3 1 3 1 2 2 3 3 3 3 3 1 3 2 3 3 1 3 2
  [771] 2 3 3 2 2 3 2 1 1 3 2 2 1 1 2 1 3 2 2 3 3 1 1 2 2 1 3 3 3 3 3 1 1 2
  [806] 3 3 3 3 3 3 2 3 3 3 3 3 3 1 1 3 3 2 1 1 3 3 1 2 1 3 3 1 2 1 1 2 2 1 1 3 3 3 3 1 2
  [841] 3 2 3 3 2 3 3 2 3 2 2 1 3 3 3 3 1 2 3 3 1 3 1 1 3 2 3 3 3 3 3 3 3 1 3 2 2 1
  [876] 2 2 1 2 2 1 2 1 2 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
  [911] 3 1 1 2 2 2 1 2 2 1 3 2 3 2 2 3 1 3 3 3 3 2 2 2 3 2 2 2 3 2 3 1 1 1 2
  [946] 2 3 3 3 3 3 2 1 1 2 2 2 2 2 3 2 2 3 3 1 3 1 2 2 3 3 3 1 3 2 3 2 3 3 3 1
  [981] 1 1 3 3 3 2 3 3 3 3 2 1 1 2 2 1 3 3 3 3
  [ reached getOption("max.print") -- omitted 68619 entries ]

Within cluster sum of squares by cluster:
[1] 12879.26 30235.84 49447.18
 (between_SS / total_SS =  55.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Gambar 11. Output Clustering Dengan K-Means di R

Output hasil clustering menunjukkan bahwa terbentuk tiga klaster dengan jumlah anggota sebagai berikut: Cluster 1 terdiri dari 14.943 data poin, Cluster 2 sebanyak 25.862 data poin, dan Cluster 3 sebanyak 28.814 data poin. Distribusi ini menggambarkan sebaran jumlah data pada masing-masing klaster, di mana Cluster 3 merupakan klaster dengan jumlah data terbanyak, sedangkan Cluster 1 memiliki jumlah paling sedikit.

Rata-rata dari masing-masing variabel yang telah distandarkan (scaled) pada setiap klaster ditunjukkan oleh nilai cluster means. Untuk variabel price_sqrt, nilai rata-rata pada:

1. Cluster 1: 0.6360
2. Cluster 2: 0.7255
3. Cluster 3: -0.9810

Hal ini menunjukkan bahwa harga rata-rata (setelah diskalakan) pada Cluster 1 dan 2 lebih tinggi dibandingkan Cluster 3.

Untuk variabel age_category, nilai rata-rata pada:

1. Cluster 1: -1.1916
2. Cluster 2: 0.6876
3. Cluster 3: 0.0008

Ini mengindikasikan bahwa rata-rata usia pada Cluster 1 lebih rendah dibandingkan Cluster 2 dan 3.

Sedangkan pada variabel category.1, diperoleh rata-rata:

1. Cluster 1: -0.5667
2. Cluster 2: -0.6184
3. Cluster 3: 0.8489

Dengan demikian, Cluster 3 memiliki nilai kategori tertinggi, yang bisa jadi mengindikasikan kelompok kategori tertentu yang lebih dominan dibandingkan dua klaster lainnya.

Hasil clustering juga menunjukkan nilai Within-Cluster Sum of Squares (WCSS) sebagai berikut:

1. Cluster 1: 12.879,26
2. Cluster 2: 30.235,84
3. Cluster 3: 49.447,18

WCSS mengukur sebaran data dalam masing-masing klaster. Nilai yang lebih kecil mengindikasikan bahwa data dalam klaster tersebut lebih dekat ke centroid-nya. Dengan demikian, Cluster 1 memiliki penyebaran paling kecil, sedangkan Cluster 3 paling besar.

Selain itu, rasio Between_SS / Total_SS sebesar 55,7% menunjukkan bahwa sebesar 55,7% variasi total dalam data dapat dijelaskan oleh pemisahan klaster, sedangkan sisanya (44,3%) merupakan variasi dalam klaster. Semakin tinggi rasio ini, semakin baik kualitas pemisahan klaster, dan nilai 55,7% dapat dikatakan menunjukkan kualitas clustering yang cukup baik.

Komponen-komponen yang tersedia dalam objek hasil k-means (km.out) meliputi:

a) cluster: Penugasan klaster untuk masing-masing data
b) centers: Koordinat centroid dari masing-masing klaster
c) totss: Total variasi dalam dataset
d) withinss: WCSS untuk masing-masing klaster
e) tot.withinss: Total WCSS keseluruhan
f) betweenss: Variasi antar klaster
g) size: Jumlah anggota tiap klaster
h) iter: Jumlah iterasi hingga algoritma mencapai konvergensi
i) ifault: Status akhir algoritma (misalnya apakah berhasil atau terdapat kesalahan)

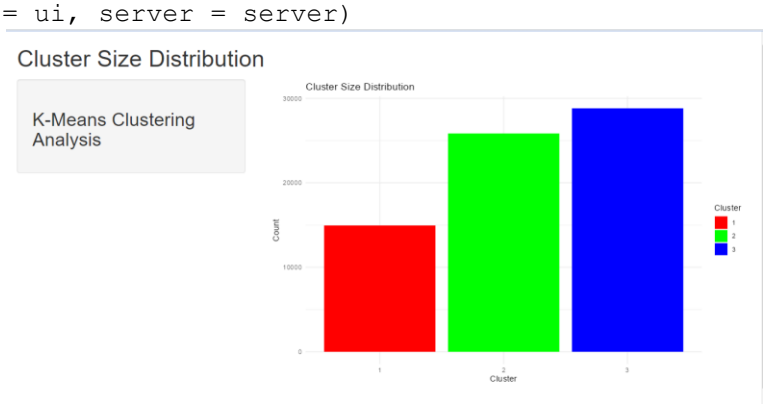
3.5 Visualisasi Hasil Clustering K-Means Shiny R

3.5.1 Visualisasi hasil clustering K-Means dalam bentuk Bar Plot

Visualisasi hasil clustering K-Means dalam bentuk Bar Plot, untuk menampilkan distribusi jumlah anggota per klaster. Berikut adalah code R-nya:

```
# UI Shiny
ui <- fluidPage(
  titlePanel("Cluster Size Distribution"),
  sidebarLayout(
    sidebarPanel(
      h3("K-Means Clustering Analysis")
    ),
    mainPanel(
      plotOutput("barPlot"),
      plotOutput("clusterPlot")
    )
  )
)
# Server Shiny
server <- function(input, output) {
  # Visualisasi bar plot distribusi cluster
  output$barPlot <- renderPlot({
    ggplot(cluster_counts, aes(x = Cluster, y = Count, fill = Cluster))
+
      geom_bar(stat = "identity") +
      labs(title = "Cluster Size Distribution", x = "Cluster", y =
"Count") +
      scale_fill_manual(values = c("red", "green", "blue")) +
      theme_minimal()
  })

  # Visualisasi clustering menggunakan fviz_cluster
  output$clusterPlot <- renderPlot({
    fviz_cluster(km.out, data = data_3cols_scaled, geom = "point",
stand = FALSE) +
      theme_minimal()
  })
}
# Menjalankan aplikasi Shiny
shinyApp(ui = ui, server = server)
```



Gambar 12. Visualisasi Bar Plot

Gambar di atas menunjukkan visualisasi distribusi ukuran klaster menggunakan diagram batang. Sumbu horizontal (X) merepresentasikan label klaster (1, 2, dan 3), sedangkan sumbu vertikal (Y) menunjukkan jumlah data dalam setiap klaster. Warna merah, hijau, dan biru digunakan untuk mengidentifikasi masing-masing klaster. Dari grafik ini, terlihat bahwa klaster 1 (merah) memiliki jumlah data paling kecil, yaitu sekitar 14.943 atau mendekati 15.000.

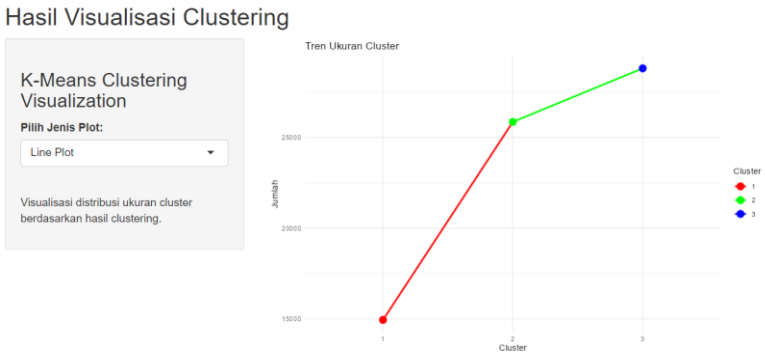
Selanjutnya, klaster 2 (hijau) memiliki jumlah data yang lebih besar, yaitu sekitar 25.862 mendekati 26.000. Sementara itu, klaster 3 (biru) memiliki jumlah data terbesar, mencapai sekitar 28.814 mendekati 29.000. Visualisasi ini membantu memahami distribusi jumlah data antar klaster dengan membandingkan tinggi setiap batang.

3.5.2 Visualisasi Clustering K-Means dalam Bentuk Line Plot

Gambar 13 di bawah ini menampilkan tren ukuran klaster menggunakan diagram garis. Sumbu horizontal (X) menunjukkan label klaster (1, 2, dan 3), sedangkan sumbu vertikal (Y) merepresentasikan jumlah data dalam setiap klaster. Setiap titik pada grafik menunjukkan ukuran klaster, yang dihubungkan oleh garis untuk memperlihatkan perubahan jumlah data antar klaster.

Klaster 1, yang direpresentasikan oleh titik merah, memiliki jumlah data terkecil. Klaster 2, ditunjukkan dengan titik hijau, memiliki jumlah data yang lebih besar. Sedangkan Klaster 3, direpresentasikan oleh titik biru, memiliki jumlah data terbesar.

Garis yang cenderung naik dari klaster 1 ke klaster 3 mengindikasikan adanya tren peningkatan jumlah data pada masing-masing klaster. Visualisasi ini memberikan gambaran yang lebih intuitif terhadap perbandingan ukuran antar klaster dalam bentuk tren.



Gambar 13. Line Plot

4. KESIMPULAN

Berdasarkan hasil clustering, pengaruh anggaran pemasaran terhadap penjualan produk di industri retail dapat dianalisis secara lebih mendalam melalui interpretasi klaster yang terbentuk. Beberapa poin utama yang mendukung analisis ini adalah sebagai berikut:

1. Distribusi Data dalam Cluster
   Cluster 3, yang memiliki jumlah data terbesar (28.814 data poin), menunjukkan rata-rata harga (price_sqrt) yang lebih rendah dibandingkan Cluster 1 dan Cluster 2. Hal ini mengindikasikan bahwa kelompok ini kemungkinan mencakup produk atau segmen pelanggan dengan harga rendah, yang mungkin memerlukan pendekatan pemasaran yang lebih intensif untuk mendorong penjualan.
2. Rata-rata Harga dalam Cluster
   Cluster 1 dan Cluster 2 memiliki rata-rata harga yang lebih tinggi, mengarah pada segmen produk atau pelanggan premium. Oleh karena itu, strategi pemasaran pada kelompok ini dapat difokuskan pada pendekatan yang menekankan nilai atau kualitas produk.
3. Rata-rata Usia Pelanggan
   Variabel age_category menunjukkan bahwa Cluster 1 memiliki rata-rata usia yang lebih rendah dibandingkan klaster lainnya, yang mengindikasikan segmen pelanggan yang lebih muda. Segmentasi ini dapat ditargetkan dengan strategi pemasaran digital yang lebih dinamis, seperti media sosial dan konten interaktif.
4. Kualitas Pemisahan Klaster
   Nilai Within Cluster Sum of Squares (WCSS) terbesar pada Cluster 3 menunjukkan

tingkat heterogenitas yang tinggi. Hal ini menyiratkan bahwa strategi pemasaran yang ditujukan pada klaster ini perlu disesuaikan dengan keberagaman preferensi pelanggan. Rasio Between_SS / Total_SS sebesar 55,7% menunjukkan bahwa model clustering memiliki performa yang cukup baik dalam memisahkan data ke dalam kelompok yang bermakna.

Selain itu, hasil clustering yang dihasilkan dari Excel dan R Shiny menunjukkan konsistensi yang tinggi, baik dari segi jumlah cluster, nilai centroid, maupun distribusi data antar cluster. Nilai-nilai centroid untuk variabel seperti age, price_sqrt, price_cate, dan age_category menunjukkan pola yang serupa.

Pendekatan manual melalui Excel memungkinkan pemahaman langkah per langkah dalam proses clustering, namun menjadi kurang efisien untuk dataset besar seperti yang digunakan (68.619 data). Sebaliknya, penggunaan R Shiny memberikan efisiensi tinggi dalam pemrosesan data besar dan otomatisasi visualisasi. Dengan demikian, metode K-Means terbukti andal dan konsisten, terlepas dari alat yang digunakan, dan dapat dijadikan dasar untuk strategi pemasaran berbasis data yang lebih terarah.

5. SARAN

Untuk penelitian selanjutnya, disarankan untuk melakukan eksplorasi terhadap variabel-variabel tambahan yang berpotensi memengaruhi hasil clustering, seperti faktor demografis (jenis kelamin, tingkat pendapatan, wilayah geografis), tren pasar terkini, maupun efektivitas berbagai saluran pemasaran (seperti iklan digital, promosi langsung, atau media sosial). Dengan menambahkan variabel-variabel ini, model clustering yang dikembangkan diharapkan mampu memberikan segmentasi pasar yang lebih akurat dan mendalam.

Selain itu, penggunaan metode clustering lain seperti DBSCAN atau hierarchical clustering dapat dijadikan perbandingan untuk mengevaluasi keandalan dan kualitas hasil klaster yang diperoleh dari metode K-Means. Integrasi dengan pendekatan prediktif berbasis machine learning juga dapat dipertimbangkan untuk memproyeksikan pola pembelian di masa depan berdasarkan hasil segmentasi saat ini.

Pendekatan ini tidak hanya dapat meningkatkan ketajaman analisis pemasaran, tetapi juga membantu pengambilan keputusan yang lebih tepat dalam pengalokasian anggaran dan strategi promosi pada masing-masing segmen pelanggan dalam industri retail.

DAFTAR PUSTAKA

[1] Kementerian Koordinator Bidang Perekonomian Republik Indonesia. (2021, November 11). *Peran penting kontribusi perdagangan ritel dalam mendukung pertumbuhan ekonomi nasional*. https://ekon.go.id/publikasi/detail/3442/peran-penting-kontribusi-perdagangan-ritel-dalam-mendukung-pertumbuhan-ekonomi-nasional

[2] Sutrismi, S., & Anggraeni, N. (2023). *Pengaruh Biaya Pemasaran terhadap Penjualan (Studi Kasus pada PT. Suling Mas Tritunggal Abadi Tulungagung)*. JAT: Journal Of Accounting and Tax, 2(1), 23–31. https://doi.org/10.36563/jat.v2i1.785

[3] Saputra, R., Karenina, A., Parinduri, Z. A., Dellyco, A. V., & Qois, A. F. (2025). Penggunaan Metode K-Means Clustering Untuk Segmentasi Pasar Konsumen. *JRIIN: Jurnal Riset Informatika dan Inovasi*, 2(10), 1895–1901. https://jurnalmahasiswa.com/index.php/jriin/article/view/2216

[4] Prasetiyo, D., Lestari, W., & Atima, V. (2024). Penerapan clustering dengan K-means untuk pemilihan menu favorit di Tetra Coffeeshop. *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, *11*(3). https://doi.org/10.35957/jatisi.v11i3.8347

[5] Muhaimin, M. A., Niswatin, R. K., Wulanningrum, R., & Muttaqien, H. (2024).

Penerapan metode Rank Order Centroid (ROC) dan Simple Additive Weighting (SAW) dalam sebuah sistem pendukung keputusan pemilihan cafe terbaik. *Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi)*, 8(2), 739–748.

[6] Algoritma Data Science School. (2022, Maret 29). *Mengenal R-Shiny, Aplikasi Web Interaktif Modern*. Algoritma. https://algorit.ma/blog/r-shiny-adalah-2022/

[7] Eko, Y., Rema, Y. O. L., Ullu, H. H., & Baso, B. (2023). Implementasi metode *K-Means Clustering* untuk menentukan kondisi gizi balita (Studi kasus: Puskesmas Mamsena). *Jurnal TEKNO KOMPAK*, 19(1), 163–177.

[8] Wahyudi, M., Masitha, M., Saragih, R., & Solikhun. (2020). *Data mining: Penerapan algoritma K-Means clustering dan K-Medoids clustering*. Kita Menulis.

[9] Beeley, C., & Sukhdeve, S. (2013). *Web Application Development with R Using Shiny*. Packt Publishing.