LAPORAN PROJECT PENAMBANGAN DATA
SEMESTER GENAP 2024/2025

IDENTITAS PROYEK	
Judul	Segmentasi Tipe Konsumen Menggunakan Algoritma *K-Means* dan *Agglomerative Clustering* untuk Optimalisasi Strategi Promosi Penjualan pada Pusat Perbelanjaan.
Topik	Segmentasi Konsumen untuk Optimasi Strategi Promosi Penjualan pada Pusat Perbelanjaan.
Identitas Penyusun	1. Fadhil Muhamad (23031554090) 2. Sintiya Risla Miftaqul Nikmah (23031554204) 3. Dimas Fatkhul Rahman (23031554211)
Kelas	2023B

1. PENDAHULUAN

Pendahuluan penelitian tidak lebih dari 1000 kata yang terdiri dari:
A. Latar belakang dan rumusan permasalahan yang akan diteliti
B. Pendekatan pemecahan masalah

1.1. Latar Belakang (min. 250 kata)

Dalam dunia bisnis modern, khususnya di industri ritel dan pusat perbelanjaan, memahami karakteristik perilaku dan preferensi konsumen menjadi kunci utama dalam menyusun strategi pemasaran yang efektif, meningkatkan kepuasan pelanggan, dan meningkatkan profit [1]. Perubahan perilaku konsumen yang dinamis menuntut perusahaan untuk tidak lagi menggunakan pendekatan promosi yang bersifat umum, melainkan mengarah pada pendekatan yang lebih personal dan spesifik. Oleh karena itu, segmentasi konsumen menjadi langkah strategis untuk mengelompokkan pelanggan ke dalam kelompok-kelompok yang memiliki kesamaan karakteristik, kebutuhan, atau perilaku sehingga memfasilitasi pendekatan yang ditargetkan dan dipersonalisasi [2]. Segmentasi konsumen didefinisikan sebagai pembagian konsumen berdasarkan aspek demografi (usia, jenis kelamin, status pernikahan) dan perilaku [3].

Pada penelitian ini dua algoritma *clustering* berbeda diterapkan untuk melakukan segmentasi pelanggan dan akhirnya membandingkan hasil cluster yang diperoleh dari algoritma [4]. Teknik segmentasi konsumen berbasis *data mining*, seperti *K-Means Clustering* dan *Agglomerative Clustering* merupakan metode yang efisien untuk menemukan pola tersembunyi dalam data konsumen. *K-Means Clustering* mampu mengelompokkan data ke dalam jumlah *cluster* yang optimal, di mana data yang serupa ditempatkan di cluster yang sama [5] secara cepat dan

meningkatkan kualitas pengelompokan informasi data pelanggan [6], sedangkan *Agglomerative Clustering* membangun hierarki kelompok yang memberikan gambaran hubungan antar segmen konsumen. Dengan penerapan kedua algoritma ini, pusat perbelanjaan dapat mengidentifikasi tipe-tipe konsumen yang berbeda, memahami kebutuhan spesifik mereka, serta menyusun strategi promosi yang lebih terarah dan efektif.

Melalui segmentasi yang akurat, pusat perbelanjaan tidak hanya dapat meningkatkan efektivitas promosi dan pemasaran, tetapi juga meningkatkan pengalaman belanja pelanggan secara keseluruhan. Penyesuaian strategi promosi sesuai dengan tipe konsumen yang ada diharapkan mampu meningkatkan loyalitas pelanggan, memperbesar tingkat konversi penjualan, serta memperkuat posisi pusat perbelanjaan dalam menghadapi persaingan pasar yang semakin ketat. Oleh karena itu, proyek ini bertujuan untuk menerapkan algoritma *K-Means* dan *Agglomerative Clustering* dalam melakukan segmentasi tipe konsumen pada pusat perbelanjaan, guna mendapatkan wawasan tentang perilaku dan preferensi konsumen, serta menggunakan wawasan tersebut untuk merancang dan menerapkan pemasaran bertarget dan rekomendasi produk yang dipersonalisasi [7].

1.2. Rumusan Masalah dan Tujuan

1.2.1 Rumusan Masalah

Rumusan masalah dari penelitian ini adalah:

1. Bagaimana penerapan algoritma K-Means dan Agglomerative Clustering dalam segmentasi tipe konsumen di pusat perbelanjaan?
2. Bagaimana kualitas hasil segmentasi yang diperoleh dari kedua algoritma tersebut?
3. Bagaimana interpretasi hasil segmentasi yang telah dilakukan?

1.2.2 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Menerapkan algoritma K-Means dan Agglomerative Clustering untuk melakukan segmentasi tipe konsumen pada data konsumen pusat perbelanjaan.
2. Mengevaluasi serta membandingkan performa dan kualitas hasil segmentasi dari masing-masing algoritma.

3. Menghasilkan profil segmen konsumen yang dapat digunakan sebagai dasar dalam merancang strategi promosi penjualan yang lebih efektif dan tepat sasaran.

1.2.3 Pendekatan Penyelesaian Masalah

Pendekatan penyelesaian masalah dalam segmentasi tipe konsumen untuk optimalisasi strategi promosi penjualan di pusat perbelanjaan dimulai dengan melakukan studi literatur dari penelitian sebelumnya untuk memahami teori-teori terkait segmentasi konsumen, teknik clustering, serta penerapan algoritma *K-Means* dan *Agglomerative Clustering* dalam konteks *data mining* dan pemasaran. Selanjutnya, data konsumen yang berisi atribut-atribut relevan dikumpulkan dari dataset yang ada di *kaggle*. Proses berikutnya adalah pembersihan data (data cleaning) untuk mengatasi masalah *missing value*, duplikasi, dan *outlier*, serta melakukan pre-processing seperti normalisasi data untuk memastikan kualitas data yang siap dianalisis. Selanjutnya, teknik reduksi dimensi *Principal Component Analysis* (PCA) diterapkan untuk menyederhanakan kompleksitas data dan mempercepat proses clustering. Setelah itu, algoritma *clustering* seperti *K-Means* dan *Agglomerative Clustering* diterapkan untuk membentuk cluster dan hierarki segmen pelanggan berdasarkan jarak antar data. Evaluasi model dilakukan dengan melakukan *Exploratory Data Analysis* untuk mempelajari pola dalam *cluster* yang terbentuk dan menentukan sifat pola cluster, serta menarik kesimpulan setiap *cluster*. Setelah terbentuk, karakteristik setiap *cluster* dianalisis untuk menghasilkan deskripsi profil konsumen, yang mencakup kebutuhan, kebiasaan, dan perilaku belanja dari masing-masing segmen. Berdasarkan hasil ini, strategi promosi yang lebih efektif, efisien, dan tepat sasaran dapat disusun untuk meningkatkan efektivitas promosi di pusat perbelanjaan.

2. Metodologi

Metodologi atau cara untuk mencapai tujuan yang telah ditetapkan ditulis tidak melebihi 1000 kata. Bagian ini berisi metode pre-processing dan/atau metode post processing yang dilengkapi dengan diagram alir penelitian yang menggambarkan apa yang sudah dilaksanakan dan yang akan dikerjakan selama waktu yang diusulkan. Format diagram alir dapat berupa file JPG/PNG. Metode penelitian harus dibuat secara utuh dengan penahapan yang jelas.

2.1. Eksplorasi Dataset

Pemahaman dataset yang dimiliki

Dataset yang digunakan adalah **Customer Personality Analysis** yang diambil dari situs kaggle. Dataset ini berisi informasi mengenai pelanggan dari

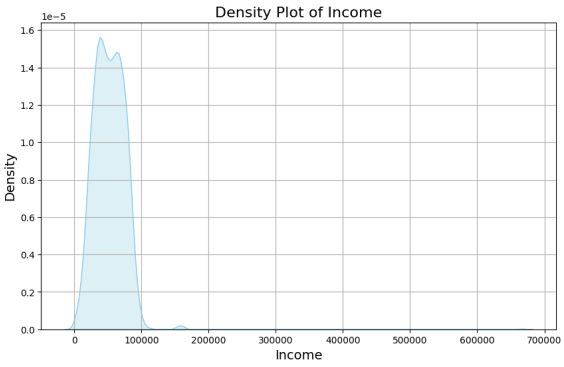
sebuah perusahaan ritel, termasuk data demografis, kebiasaan pembelian, dan respon terhadap kampanye pemasaran. Tujuan dari dataset ini adalah untuk memahami perilaku pelanggan dan meningkatkan strategi pemasaran.

Dataset ini mengandung kolom yang berisi informasi pelanggan, informasi pengeluaran terhadap produk tertentu selama 2 tahun terakhir, jenis promosi, serta media dimana pembelian/interaksi terjadi, terdiri dari fitur antara lain:

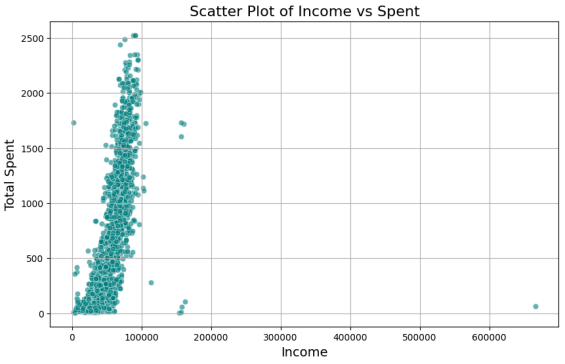
- ID : id untuk setiap pelanggan
- Year_Birth : Tahun kelahiran pelanggan
- Education : Tingkat pendidikan pelanggan
- Marital_Status : Status pernikahan pelanggan
- Income : Pendapatan tahunan pelanggan
- Kidhome : Jumlah anak di rumah
- Teenhome : Jumlah remaja di rumah
- Dt_Customer : Tanggal pelanggan menjadi pelanggan perusahaan
- Recency : Jumlah hari sejak pelanggan terakhir melakukan pembelian
- MntWines : Jumlah uang yang dihabiskan untuk wine dalam 2 tahun terakhir
- MntFruits : Jumlah uang yang dihabiskan untuk buah dalam 2 tahun terakhir
- MntMeatProducts : Jumlah uang yang dihabiskan untuk daging dalam 2 tahun terakhir
- MntFishProducts : Jumlah uang yang dihabiskan untuk ikan dalam 2 tahun terakhir
- MntSweetProducts : Jumlah uang yang dihabiskan untuk makanan manis dalam 2 tahun terakhir
- MntGoldProds : Jumlah uang yang dihabiskan untuk emas dalam 2 tahun terakhir
- NumDealsPurchases : Jumlah pembelian dengan diskon
- NumWebPurchases : Jumlah pembelian melalui web
- NumCatalogPurchases : Jumlah pembelian melalui katalog
- NumStorePurchases : Jumlah pembelian melalui toko
- NumWebVisitsMonth : Jumlah kunjungan ke situs web dalam sebulan terakhir

Target dari dataset ini dapat berupa pengelompokkan natural terbaik yang paling bermakna (artinya bagaimana caranya menentukan pelanggan menjadi kategori tertentu dengan berbagai informasi yang tersedia secara tidak langsung) untuk membagi pelanggan ke dalam beberapa kelompok yang memiliki keunikan tersendiri

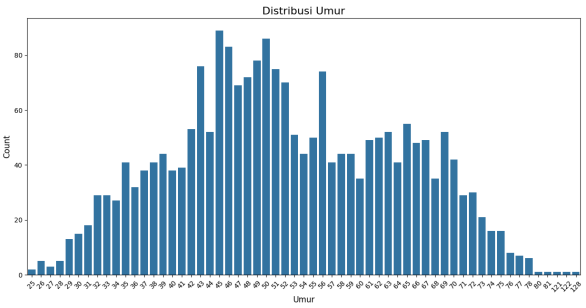
antara satu sama lain, sehingga kita dapat menentukan keputusan berdasarkan interpretasi hasil segmentasi dari pengolahan distribusi data itu.



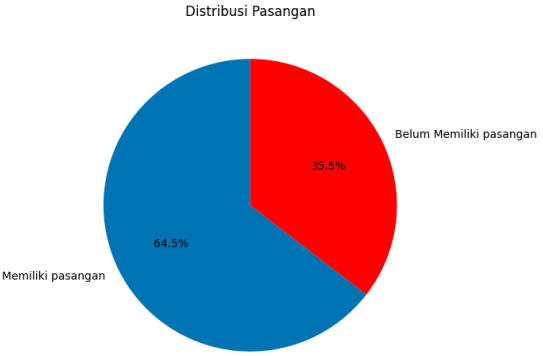
Gambar 1. Distribusi data Income



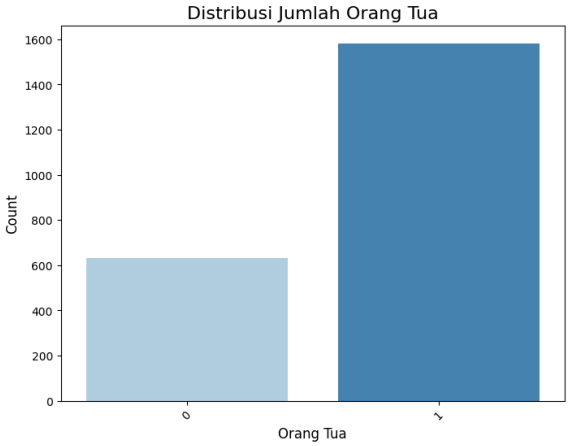
Gambar 2. Scatter Plot Income vs Spent



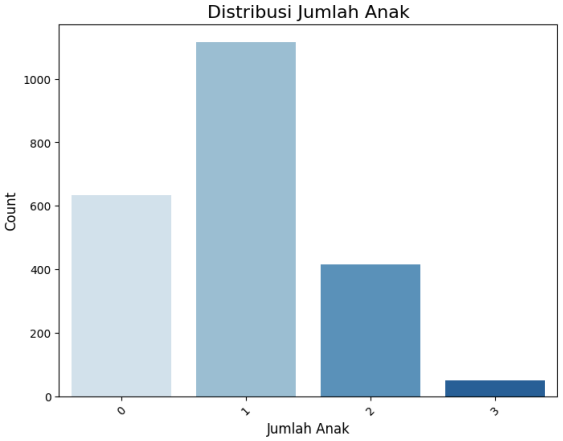
Gambar 3. Distribbusi Umur Konsumen



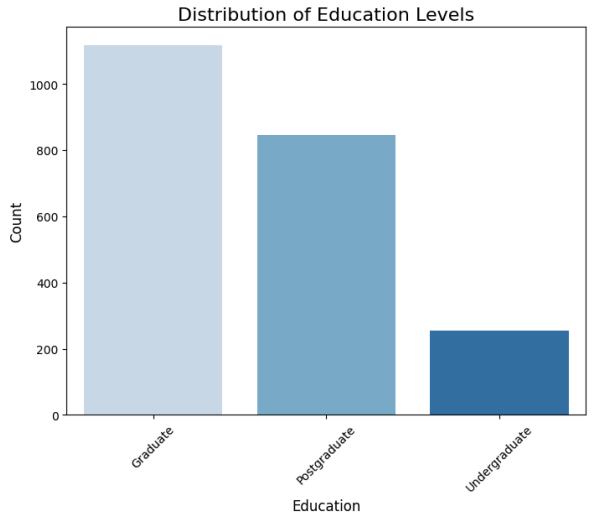
Gambar 4. Distribusi Status Pernikahan



Gambar 5. Distribusi Jumlah Orang Tua



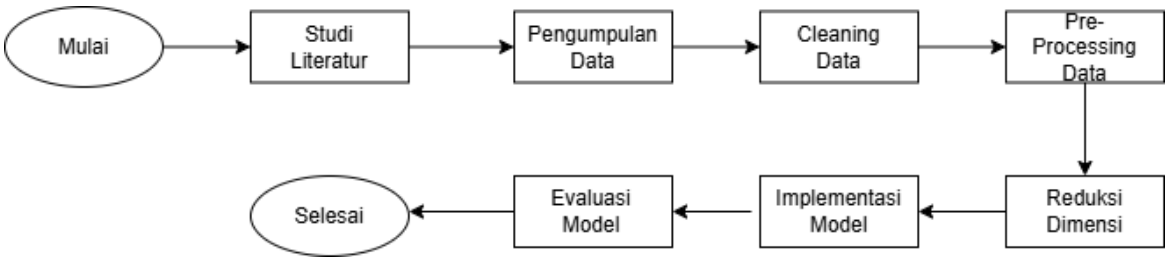
Gambar 6. Distribusi Jumlah Anak



Gambar 7. Distribusi Level Pendidikan Konsumen

2.2. Langkah Penelitian

Berikut adalah diagram alir pada penelitian ini:



Gambar 8. Diagram Alir

1. **Studi Literatur**

Segmentasi konsumen merupakan aspek penting dari analisis modern, yang memungkinkan bisnis untuk memahami dan memenuhi berbagai kebutuhan pelanggan mereka [8]. Penelitian ini diawali dengan menganalisis berbagai penelitian sebelumnya mengenai Segmentasi Konsumen. Berbagai penelitian dengan menggunakan metode *K-Means Clustering* [4],[5],[6],[7],[8],[9] dipelajari untuk memahami keunggulan serta keterbatasan masing-masing. Selain itu, penelitian dengan menggunakan *Agglomerative Clustering* juga dipelajari guna meningkatkan wawasan lebih lanjut mengenai model [9],[10].

2. **Pengumpulan Data**

Dataset diperoleh dari situs kaggle **Customer Personality Analysis**. Dataset ini berisi informasi yang relevan terkait dengan profil konsumen, yang mencakup atribut-atribut seperti id, tahun lahir, pendidikan, status pernikahan, dll. Data ini digunakan untuk menganalisis pola-pola konsumen dalam upaya mengidentifikasi segmen-segmen pasar yang berbeda dan menyusun strategi promosi yang lebih terarah dan efisien. Dengan

menggunakan data ini, analisis segmentasi konsumen dapat dilakukan untuk memahami lebih mendalam kebutuhan dan perilaku konsumen di pusat perbelanjaan.

3. ***Cleaning Data***

Pada tahap ini, data yang telah dikumpulkan dibersihkan *(data cleaning)* untuk memastikan bahwa data tersebut layak dan siap digunakan dalam analisis lanjutan, serta dilakukan *feature engineering* untuk memperbaiki, menyesuaikan, atau menambahkan fitur-fitur baru yang relevan. Proses ini mencakup identifikasi dan penghapusan data duplikat yang dapat menyebabkan bias dalam hasil analisis. Selain itu, dilakukan penanganan terhadap *missing value* melalui teknik seperti imputasi atau penghapusan data tergantung pada tingkat keparahan dan proporsinya. *Outlier* yang dapat mengganggu akurasi model juga dideteksi menggunakan metode statistik seperti IQR *(Interquartile Range)*, dan kemudian ditangani sesuai dengan proporsinya. Di samping itu, ketidakkonsistenan dalam format data, seperti perbedaan penulisan kategori atau kesalahan pengetikan, juga diperbaiki untuk memastikan keseragaman.

Pembersihan data merupakan tahap penting karena kualitas data sangat mempengaruhi ketepatan hasil *clustering* dan interpretasi segmentasi konsumen. Data yang bersih dan konsisten akan meningkatkan akurasi model analisis, mengurangi kemungkinan kesalahan, serta memastikan bahwa keputusan strategis yang diambil berdasarkan hasil analisis dapat lebih akurat dan efektif.

4. ***Pre-Processing***

Pre-processing mencakup berbagai tahapan transformasi data untuk mempersiapkan data sebelum analisis *clustering* dilakukan. Salah satu langkah yang dilakukan adalah *label encoding* untuk mengkodekan fitur kategoris ke dalam format numerik, sehingga dapat dikenali dan diproses oleh model. Selanjutnya, semua fitur diskalakan menggunakan *standard scaler* untuk memastikan bahwa setiap fitur memiliki skala yang seragam, dengan rata-rata nol dan standar deviasi satu. Selain itu, dibuat kerangka data subset khusus yang dirancang untuk pengurangan dimensi, yang kemudian akan digunakan sebagai input dalam proses clustering. Secara keseluruhan, tahap ini bertujuan untuk memastikan bahwa semua fitur yang relevan dapat diolah secara optimal.

5. **Reduksi Dimensi**

Pada dataset di atas, ada banyak fitur yang menjadi dasar klasifikasi. Semakin banyak jumlah fitur, semakin sulit pula dalam mengolahnya. Sehingga reduksi dimensi dilakukan untuk menyederhanakan kompleksitas data, mempercepat proses komputasi, dan membantu visualisasi hasil clustering. *Principal Component Analysis* (PCA) dilakukan mengurangi dimensionalitas kumpulan data tersebut, meningkatkan interpretabilitas tetapi pada saat yang sama meminimalkan hilangnya informasi.

6. **Implementasi Model**

   Pada tahap ini, dilakukan penentuan jumlah *cluster* optimal menggunakan ***Elbow Method***. Setelah jumlah *cluster* ditentukan, proses *clustering* dilakukan menggunakan dua pendekatan, yaitu *K-Means Clustering* dan *Agglomerative Clustering*. Kedua algoritma ini digunakan untuk membentuk segmen konsumen berdasarkan pola kesamaan dalam data. Selanjutnya, hasil pembentukan *cluster* divisualisasikan melalui *scatter plot* untuk mengevaluasi penyebaran dan pemisahan antar *cluster*, serta membandingkan hasil *clustering* yang diperoleh dari algoritma *K-Means* dan *Agglomerative Clustering* guna melihat perbedaan struktur segmen yang terbentuk.

7. **Evaluasi Model**

   Karena proyek ini menggunakan pendekatan *unsupervised learning*, tidak ada fitur target yang memungkinkan evaluasi menggunakan metrik *supervised* seperti akurasi atau *F1-score*. Oleh karena itu, evaluasi model dilakukan melalui pengukuran nilai *silhouette score* dan *Exploratory Data Analysis* (EDA) dengan cara menganalisis distribusi data dalam setiap cluster, memvisualisasikan hasil *clustering* melalui *scatter plot*, serta mengamati karakteristik utama dari masing-masing kelompok. Cara ini memungkinkan penilaian kualitas *cluster* berdasarkan pola perilaku konsumen yang terbentuk, serta relevansi hasil segmentasi terhadap kebutuhan bisnis, tanpa bergantung pada skor evaluasi berbasis label.

Rencana output dari penelitian ini mencakup terbentuknya model segmentasi konsumen berdasarkan hasil *clustering* menggunakan algoritma *K-Means* dan *Agglomerative Clustering*. Selain itu, penelitian ini juga menghasilkan penjelasan untuk setiap *cluster*, yang menggambarkan karakteristik utama dari masing-masing segmen konsumen. Hasil lainnya adalah visualisasi distribusi *cluster* melalui *scatter plot*, serta rekomendasi strategi promosi yang disesuaikan dengan profil dan

kebutuhan dari setiap kelompok konsumen. Dengan output ini, diharapkan pusat perbelanjaan dapat mengoptimalkan strategi pemasaran secara lebih terarah dan efektif.

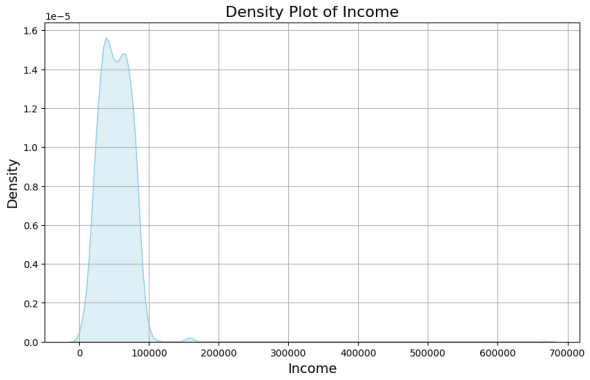
3. Hasil dan Analisis

3.1 Exploratory Data Analysis

1. Cleaning Data

- **Penanganan *Missing Value***

Pada dataset yang telah dikumpulkan, pengecekan nilai *missing value* dilakukan untuk menghindari distorsi dalam analisis dan hasil clustering. Hasil pengukuran menunjukkan bahwa hanya kolom *Income* yang memiliki missing value, sebanyak 24 nilai kosong. Karena jumlahnya yang relatif sedikit, penanganan dilakukan dengan menghapus baris yang mengandung missing value tersebut guna menjaga kualitas data. Berikut adalah visualisasi distribusi kolom *income* setelah dilakukan penanganan *missing value*:



Gambar 9. Distribusi Income Setelah Penanganan Missing Value

- **Penanganan Ketidakkonsistenan Data**

Pada kolom *Dt_Customer* ditemukan ketidakkonsistenan format tanggal yang dapat mempengaruhi proses analisis. Beberapa baris memiliki format lengkap berupa tahun-bulan-tanggal-jam (YYYY-MM-DD HH:MM:SS), sementara yang lain hanya berisi format tahun-bulan-tanggal (YYYY-MM-DD). Berikut contoh ketidakkonsistenan datanya:

Dt_Customer
15-11-2013
20-02-2013
2013-05-11 00:00:00
2014-01-01 00:00:00

Tabel 1. Ketidakkonsistenan Data

Untuk mengatasi hal ini, seluruh data pada kolom tersebut dikonversi ke format *datetime* yang seragam agar dapat diproses secara konsisten dalam analisis selanjutnya. Hasilnya adalah sebagai berikut:

```
5   Kidhome              2216 non-null   int64
6   Teenhome             2216 non-null   int64
7   Dt_Customer          2216 non-null   datetime64[ns]
8   Recency              2216 non-null   int64
```

Gambar 10. Hasil Penanganan Ketidakkonsistenan Data

- ***Feature Engineering***

  Pada bagian ini terdapat beberapa fitur baru untuk mempermudah dalam analisis lanjutan nantinya. Feature Engineering yang kami lakukan diantaranya:

  1) Menghitung jumlah hari pelanggan

     Membuat fitur baru bernama "*Days_As_Customer*" yang berisi jumlah hari pelanggan mulai berbelanja. Jumlah hari pelanggan dihitung berdasarkan selisih hari antara tanggal terbaru pelanggan dengan setiap tanggal pelanggan lainnya.

  2) Menghitung usia pelanggan

     Membuat fitur baru bernama "*Age*" yang berisi usia setiap pelanggan. Disini kami menggunakan tahun 2021 dan tahun kelahiran setiap pelanggan untuk menghitung usia pelanggan.

  3) Menghitung total pengeluaran pelanggan

     Membuat fitur baru bernama "*Spent*" yang berisi total pengeluaran setiap pelanggan. Total pengeluaran pelanggan dihitung dari penjumlahan total yang dibelanjakan oleh pelanggan dalam berbagai kategori, seperti wine, buah-buahan, daging, ikan, makanan manis, dan emas.

  4) Menyederhanakan status pernikahan menjadi dua kategori

     Membuat fitur baru bernama "*Marital_Status_Simplified*" yang merupakan versi terbaru dari fitur "*Marital_Status*". Setiap value pada fitur *"Marital_Status*" diganti menjadi dua kategori saja, diantaranya:

Value *"Marital_Status*"	Value "*Marital_Status_Simplified*"
Married	Memiliki pasangan
Together	Memiliki pasangan
Single	Belum Memiliki pasangan
Divorced	Belum Memiliki pasangan

Widow	Belum Memiliki pasangan
YOLO	Belum Memiliki pasangan
Absurd	Belum Memiliki pasangan

Tabel 2. Penyederhanaan Status Pernikahan

5) Menghitung jumlah anak dalam rumah tangga

   Membuat fitur bernama "*Children*" yang berisi jumlah anak dalam rumah tangga. Jumlah anak ini terdiri dari anak-anak dan remaja yang berasal dari kolom "*Kidhome*" dan "*Teenhome*".

6) Menghitung jumlah anggota rumah tangga

   Membuat fitur bernama "*Number_of_Family_Members*" yang berisi banyaknya anggota dalam rumah tangga. Banyaknya anggota dalam rumah tangga dihitung dengan memanfaatkan kolom "*Marital_Status_Simplified*" dan "*Children*", penjelasannya sebagai berikut:

   a. Pada kolom "*Marital_Status_Simplified*", untuk kategori "Memiliki pasangan" diubah menjadi angka 2 dalam bentuk integer, sedangkan kategori "Belum memiliki pasangan diubah menjadi angka 1 dalam bentuk integer juga.
   b. Menjumlahkan kolom "*Marital_Status_Simplified*" yang sudah diubah tadi dengan kolom "*Children*" yang berisi jumlah anak dalam rumah tangga

7) Menentukan apakah seseorang dikatakan sebagai orang tua

   Membuat fitur baru bernama "*Is_Parent*" yang berisi data kategorik biner 1 jika seseorang punya anak/dikatakan sebagai orang tua, dan 0 jika seseorang belum punya anak/belum bisa dikatakan sebagai orang tua. Cara menentukannya adalah dengan memanfaatkan kolom "*Children*", dengan ketentuan:

   a. Jika valuenya lebih dari 0, maka value kolom "*Is_Parent*" akan diisi angka 1
   b. Jika valuenya kurang atau sama dengan 0, maka value kolom "*Is_Parent*" akan diisi angka 0

8) Menyederhanakan level pendidikan

   Setiap value pada fitur *"Education*" diganti menjadi tiga kategori saja, diantaranya:

Sebelum	Sesudah

Basic	Undergraduate
2n Cycle	Undergraduate
Graduation	Graduate
Master	Postgraduate
PhD	Postgraduate

Tabel 3. Penyederhanaan Level Pendidikan

9) Mengganti nama kolom pengeluaran

Setiap kolom pengeluaran diganti namanya menjadi lebih singkat dan mudah dipahami, diantaranya:

Sebelum	Sesudah
MntWines	Wine
MntFruits	Buah-buahan
MntMeatProducts	Daging
MntFishProducts	Ikan
MntSweetProducts	Manisan
MntGoldProds	Perhiasan

Tabel 4. Penyederhanaan Kolom

10) Menghapus fitur yang tidak relevan atau sudah digantikan

Beberapa fitur seperti "*Marital_Status*", "*Dt_Customer*", "*Year_Birth*", dan "*ID*" sudah tidak lagi relevan untuk analisis selanjutnya sehingga fitur-fitur ini diputuskan untuk dihapus.

- **Penanganan Outlier**

  Outlier adalah sebuah istilah yang berkaitan dengan nilai atau data yang terjadi penyimpangan yang signifikan atau sangat berbeda dari data yang lain. Outlier tentunya mengganggu performa model, hal ini karena outlier bisa mengganggu analisis data dan menyebabkan hasil yang tidak akurat. Outlier perlu ditangani agar analisis data lebih akurat, reliable, dan tidak terpengaruh oleh nilai ekstrim yang tidak representatif.

  Pada penelitian ini value pada kolom "*Age*" yang kurang dari 90 dan value pada kolom "*Income*" yang kurang dari 600.000 dianggap sebagai outlier. Penanganan yang dilakukan adalah dengan menghapus baris yang terdapat outlier tersebut.

- **Korelasi Fitur (tidak termasuk data kategorik)**

Untuk memahami bagaimana variabel-variabel numerik dalam dataset saling berkaitan, dilakukan analisis korelasi yang divisualisasikan dalam bentuk heatmap. Korelasi sendiri adalah ukuran yang menunjukkan seberapa kuat hubungan linier antara dua variabel. Dalam analisis ini digunakan koefisien korelasi Pearson, yang nilainya berkisar dari −1 hingga +1. Nilai mendekati +1 menunjukkan hubungan positif yang kuat—artinya saat satu variabel naik, variabel lain cenderung ikut naik. Sebaliknya, nilai mendekati −1 menunjukkan hubungan negatif yang kuat—saat satu naik, yang lain justru turun. Jika nilainya mendekati 0, berarti hubungan liniernya lemah atau tidak ada. Warna terang seperti merah muda atau coklat menandakan korelasi positif yang tinggi, sementara biru gelap menunjukkan korelasi negatif yang kuat.

Dari hasil visualisasi dalam kode, terlihat bahwa variabel Spent (jumlah pengeluaran) punya korelasi sangat tinggi dengan pembelian Wine (0.89) dan Daging (0.84), yang berarti dua produk ini jadi penyumbang utama pengeluaran konsumen. Selain itu, ada hubungan positif kuat juga antara NumCatalogPurchases dan Spent (0.75), yang menunjukkan bahwa belanja lewat katalog cenderung meningkatkan total pengeluaran. Di sisi lain, variabel Recency punya korelasi negatif cukup besar dengan Spent (−0.66), yang artinya pelanggan yang sudah lama tidak berbelanja cenderung mengeluarkan uang lebih sedikit.

Di sisi lain, variabel seperti Days_As_Customer tidak menunjukkan korelasi kuat dengan variabel-variabel lain, yang bisa diartikan bahwa lamanya seseorang menjadi pelanggan belum tentu menentukan perilaku belanjanya. Hal ini relevan untuk strategi retensi, karena loyalitas jangka panjang belum tentu berbanding lurus dengan intensitas atau volume pembelian.

2. *Pre-Processing Data*

- ***Label Encoding***

Label Encoding merupakan teknik pra-pemprosesan data untuk mengubah nilai kategorik menjadi format numerik agar dapat digunakan dalam algoritma *machine learning*. Pada dataset ini, terdapat dua kolom bertipe object yang perlu dikonversi, yaitu kolom *Education* dan *Marital_Status_Simplified*.

```
Kolom dengan tipe data object:
Index(['Education', 'Marital_Status_Simplified'],
dtype='object')
```

Oleh karena itu, dilakukan proses *label encoding* untuk menggantikan masing-masing kategori dengan nilai numerik yang sesuai, sehingga data dapat diproses lebih lanjut dalam tahap clustering. Hasilnya adalah sebagai berikut:

```
print(df_encoded['Education'].unique())
```

```
[0 1 2]
```

Pada kolom *Education* dengan hasil di atas, telah berhasil dilakukan *label encoding* dengan interpretasi nilai 0 mewakili *graduate*, 1 mewakili *postgraduate*, 2 mewakili *undergraduate.*

```
print(df_encoded['Marital_Status_Simplified'].unique())
```

```
[0 1]
```

Sementara pada kolom *Marital_Status_Simplified* dengan hasil di atas, hasil encodingnya dapat diinterpretasikan menjadi nilai 0 mewakili individu yang belum memiliki pasangan, dan nilai 1 mewakili individu yang sudah memiliki pasangan.

- **Normalisasi Data**

Dalam penelitian ini, normalisasi dilakukan menggunakan metode *Standard Scaler*, yang mengubah nilai setiap fitur agar memiliki mean 0 dan standar deviasi 1. Langkah ini penting untuk memastikan hasil clustering yang lebih akurat dan seimbang. Berikut adalah hasil normalisasi dari setiap fitur pada dataset:

	Education	Income	Kidhome	Teenhome	Recency	Wine	Buah-buahan	Daging	Ikan	Manisan
0	-0.893586	0.287105	-0.822754	-0.929699	0.310353	0.977660	1.552041	1.690293	2.453472	1.483713
1	-0.893586	-0.260882	1.040021	0.908097	-0.380813	-0.872618	-0.637461	-0.718230	-0.651004	-0.634019
2	-0.893586	0.913196	-0.822754	-0.929699	-0.795514	0.357935	0.570540	-0.178542	1.339513	-0.147184
3	-0.893586	-1.176114	1.040021	-0.929699	-0.795514	-0.872618	-0.561961	-0.655787	-0.504911	-0.585335
4	0.571657	0.294307	1.040021	-0.929699	1.554453	-0.392257	0.419540	-0.218684	0.152508	-0.001133

	Perhiasan	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth
0	0.852576	0.351030	1.426865	2.503607	-0.555814	0.692181
1	-0.733642	-0.168701	-1.126420	-0.571340	-1.171160	-0.132545
2	-0.037254	-0.688432	1.426865	-0.229679	1.290224	-0.544908
3	-0.752987	-0.168701	-0.761665	-0.913000	-0.555814	0.279818
4	-0.559545	1.390492	0.332600	0.111982	0.059532	-0.132545

Days_As_Customer	Age	Spent	Marital_Status_Simplified	Children	Number_of_Family_Members	Is_Parent
1.973583	1.018352	1.676245	-1.349603	-1.264598	-1.758359	-1.581139
-1.665144	1.274785	-0.963297	-1.349603	1.404572	0.449070	0.632456
-0.172664	0.334530	0.280110	0.740959	-1.264598	-0.654644	-1.581139
-1.923210	-1.289547	-0.920135	0.740959	0.069987	0.449070	0.632456
-0.822130	-1.033114	-0.307562	0.740959	0.069987	0.449070	0.632456

Gambar 11. Hasil Normalisasi Data

Dari hasil di atas, terlihat bahwa setiap fitur sudah berhasil dilakukan normalisasi, di mana nilai-nilai pada setiap kolom telah disesuaikan ke dalam skala yang sebanding dengan mean mendekati 0 dan standar deviasi mendekati 1.

3. Reduksi Dimensi

Metode reduksi dimensi dilakukan dengan PCA dimana PCA bekerja dengan mengubah fitur asli menjadi sejumlah *principal components* yang tidak saling berkorelasi, di mana komponen-komponen tersebut merupakan kombinasi linear dari fitur awal. Langkah-langkah pengerjaan PCA yang dilakukan adalah sebagai berikut:

- Menghitung Nilai Varians dan Kovarians Matriks

  Langkah pertama dalam PCA adalah membentuk matriks kovarians dari dataset yang telah dinormalisasi. Kovarians mengukur sejauh mana dua variabel berubah bersama-sama. Berikut hasil dari beberapa fitur:

	Education	Income	Kidhome	Teenhome	Recency	Wine	Buah-buahan	Daging	Ikan
Education	1.000452	-0.086292	0.022432	-0.032187	-0.026182	-0.027711	-0.087370	-0.087295	-0.065825
Income	-0.086292	1.000452	-0.514755	0.034580	0.007969	0.688521	0.507583	0.692592	0.520275
Kidhome	0.022432	-0.514755	1.000452	-0.039083	0.010627	-0.497428	-0.373427	-0.439230	-0.388819
Teenhome	-0.032187	0.034580	-0.039083	1.000452	0.014398	0.003947	-0.175984	-0.261252	-0.205328
Recency	-0.026182	0.007969	0.010627	0.014398	1.000452	0.015988	-0.005259	0.022924	0.000788
Wine	-0.027711	0.688521	-0.497428	0.003947	0.015988	1.000452	0.386019	0.568338	0.397095
Buah-buahan	-0.087370	0.507583	-0.373427	-0.175984	-0.005259	0.386019	1.000452	0.546987	0.593307
Daging	-0.087295	0.692592	-0.439230	-0.261252	0.022924	0.568338	0.546987	1.000452	0.573245
Ikan	-0.065825	0.520275	-0.388819	-0.205328	0.000788	0.397095	0.593307	0.573245	1.000452

Gambar 12. Hasil Varians dan Kovarians

- Menghitung Nilai Eigen dan Vektor Eigen

  Eigenvalue menunjukkan seberapa besar variansi data yang bisa dijelaskan oleh masing-masing komponen utama. *Eigenvector* menentukan arah dari masing-masing komponen utama tersebut. Nilai-nilai ini merupakan inti dari PCA karena membantu kita menemukan dimensi baru (principal components) yang paling merepresentasikan data asli. Berikut adalah hasil dari nilai eigennya:

  Nilai Eigen (λ): [8.28505283e+00 2.91329944e+00 1.49317885e+00 1.31724994e+00 1.11063290e+00 1.00837773e+00 9.47934549e-01 8.44510443e-01 7.63484983e-01 6.48332958e-01 6.11921771e-01 5.55103088e-01 1.73596231e-01 1.99477131e-01 2.41092058e-01 2.77519493e-01 3.51324474e-01 3.91898395e-01 4.50048396e-01 4.26366869e-01 -6.06381892e-17 1.88851380e-15 1.90782563e-15]

Berdasarkan *Kaiser Rule*, jumlah *Principal Components* yang dipilih adalah sejumlah nilai eigen yang lebih besar dari 1. Sehingga, jika mengikuti aturan tersebut jumlah PC yang diambil adalah 6 PC.

- Menghitung Proporsi Variansi dan Persentase Kumulatif

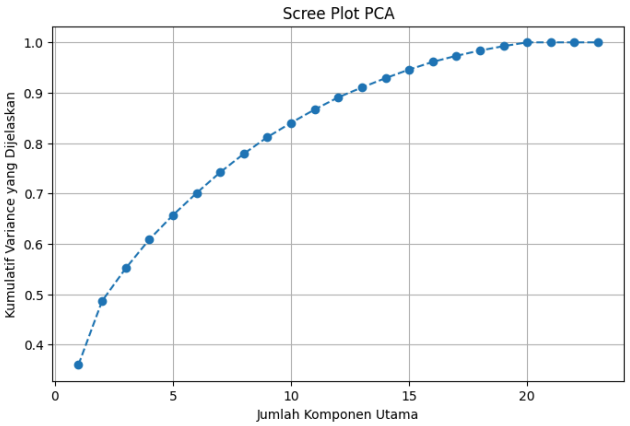
  Untuk meyakinkan jumlah pengambilan PC, dilakukan perhitungan persentase kumulatif dari variansi yang dijelaskan oleh setiap komponen utama. Dengan cara ini, kita dapat melihat seberapa banyak informasi yang dapat dijelaskan oleh setiap komponen utama yang dipilih. Berikut adalah hasil nya:

PC	Eigenvalue	Persentase (%)	Kumulatif Persentase (%)
PC1	8.285053	36.005684	36.005684
PC2	2.913299	12.660793	48.666477
PC3	1.493179	6.489147	55.155624
PC4	1.317250	5.724585	60.880209
PC5	1.110633	4.826656	65.706864
PC6	1.008378	4.382269	70.089133
PC7	0.947935	4.119591	74.208725
PC8	0.844510	3.670125	77.878849
…….	……	…….	…….
PC23	0.000000	0.000000	100.000000

  Dari hasil di atas, dapat dilihat bahwa dengan menggunakan 6 PC, sudah dapat menjelaskan lebih dari 70 persen variansi dalam data. Hal ini sesuai dengan aturan *Cumulative Variance Criterion*, yang menyarankan pemilihan PC yang dapat menjelaskan sebagian besar variansi dalam data, biasanya lebih dari 70% hingga 90%. Aturan ini juga mendukung hasil pemilihan jumlah PC menggunakan *Kaiser Rule.*

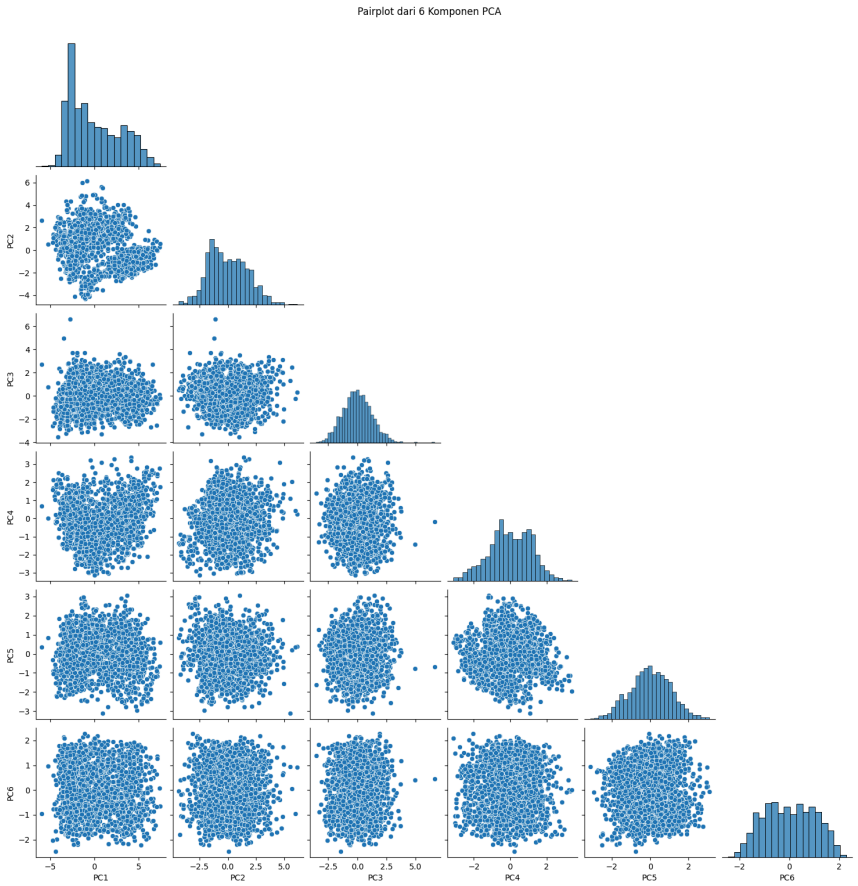
- Mendapatkan Hasil *Principal Component*

  Dari hasil perhitungan, didapatkan bahwa jumlah PC yang diambil adalah 6 PC, yang mana sesuai dengan *Kaiser Rule* dan *Cumulative Variance Criterion.* Berikut adalah hasil *scree plot* yang menunjukkan titik elbow mendukung pemilihan 6 komponen tersebut:



Gambar 13. Hasil PCA

Dengan demikian reduksi dimensi yang telah dilakukan menghasilkan sejumlah 6 *Principal Component* dengan visualisasi pairplot sebagai berikut:



Gambar 14. Pairplot Hasil PCA

Pairplot di atas menunjukkan bahwa komponen yang diambil sudah cukup baik merepresentasikan data, dengan sebaran yang merata dan distribusi yang mendekati normal. Setiap pasangan komponen tampak tidak saling berkorelasi.

4. Implementasi Model

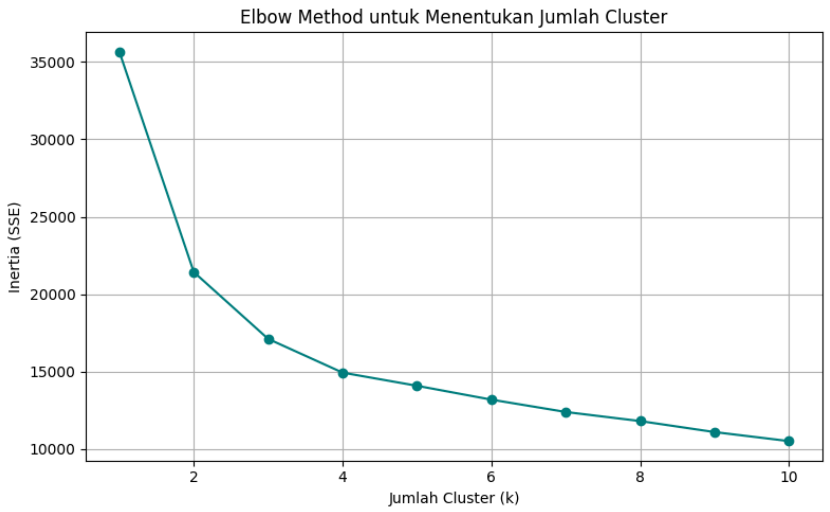
Setelah melakukan reduksi dimensi, maka bagian selanjutnya adalah implementasi model. Pada projek ini, algoritma clustering yang digunakan

adalah *Agglomerative Clustering* dan *K-Means Clustering*. Tahapan implementasi model adalah sebagai berikut:

- ***Elbow Method***

  Elbow method adalah salah satu teknik dalam *unsupervised learning* yang digunakan untuk menentukan jumlah *cluster* optimal dalam suatu dataset. *Elbow method* bukan bagian dari algoritma *K-Means* itu sendiri, melainkan metode bantu (*tool*) yang digunakan bersama *K-Means* untuk menentukan jumlah cluster (k) yang paling optimal sebelum benar-benar menjalankan algoritma *K-Means Clustering* atau algoritma *clustering* yang lainnya. Cara kerja *Elbow method* adalah sebagai berikut:

  1) Jalankan K-Means untuk berbagai jumlah *cluster*, misalnya dari k = 1 sampai k = 10.
  2) Inisialisasi list kosong "*inertia*" untuk menyimpan nilai *inertia* (SSE atau WCSS) dari masing-masing jumlah *cluster*
  3) Lakukan perulangan untuk setiap nilai k dalam rentang jumlah *cluster*:
     a. Buat objek KMeans dengan *cluster* sebanyak k dan random_state sebesar 42 agar hasil konsisten.
     b. Latih model K Means menggunakan data hasil PCA sebelumnya.
     c. Simpan nilai *inertia* (SSE atau WCSS) hasil pelatihan ke dalam list *inertia*
  4) Buat plot *Elbow method* dengan *matplotlib*
  5) Titik "*elbow*" (siku) pada grafik adalah titik di mana penurunan WCSS mulai melambat secara signifikan. Titik ini dianggap sebagai jumlah *cluster* optimal.



Gambar 15. Visualisasi Elbow Method

Plot atau hasil di atas menunjukkan bahwa jumlah *cluster* optimal untuk data ini adalah 3.

- ***Agglomerative Clustering***

  Agglomerative clustering adalah salah satu metode dari *Hierarchical Clustering* yang bekerja dengan pendekatan *bottom-up*. Singkatnya, dimulai dengan setiap data sebagai *cluster* tersendiri, lalu secara bertahap dua *cluster* yang paling mirip atau paling dekat akan digabung, dan proses ini diulang terus sampai semua data tergabung menjadi satu *cluster* besar atau tercapai jumlah *cluster* yang diinginkan.

  Namun, pada bagian sebelumnya sudah ditentukan jumlah *cluster* dengan menggunakan *elbow method* sebanyak 3 *cluster*. Oleh karena itu pada projek ini, *agglomerative clustering* akan menggunakan *cluster* sebanyak 3 dan sebagai awalan menggunakan *principal component* sebanyak 6. Jika hasil *silhouette score*-nya terlalu rendah, maka *principal component* akan dikurangi. Lebih lengkapnya seperti ini:

  1) *Clustering* untuk : *Cluster* = 3 dan PC1-PC6

     Didapat hasilnya:

     ```
     Silhouette Score: 0.267180484582771
     Calinski-Harabasz Index: 1052.7597370940828
     Davies-Bouldin Index: 1.355670987691319
     ```

     Hasil ini menunjukkan bahwa *clustering* sudah terbentuk, tetapi masih belum optimal. *Silhouette Score* yang rendah dan *Davies-Bouldin Index* yang tinggi menyiratkan adanya tumpang tindih atau kurangnya kejelasan antar cluster meskipun *Calinski-Harabasz Index* yang tinggi sedikit memberikan sinyal positif.

  2) *Clustering* untuk : *Cluster* = 3 dan PC1-PC3

     Didapat hasilnya:

     ```
     Silhouette Score: 0.39939088578789816
     Calinski-Harabasz Index: 2011.886443382473
     Davies-Bouldin Index: 0.958167109720749
     ```

     Hasil ini menunjukkan bahwa *clustering* sudah mengalami perbaikan yang signifikan. *Cluster* sekarang lebih terpisah, lebih konsisten secara internal, dan kurang tumpang tindih dibandingkan sebelumnya.

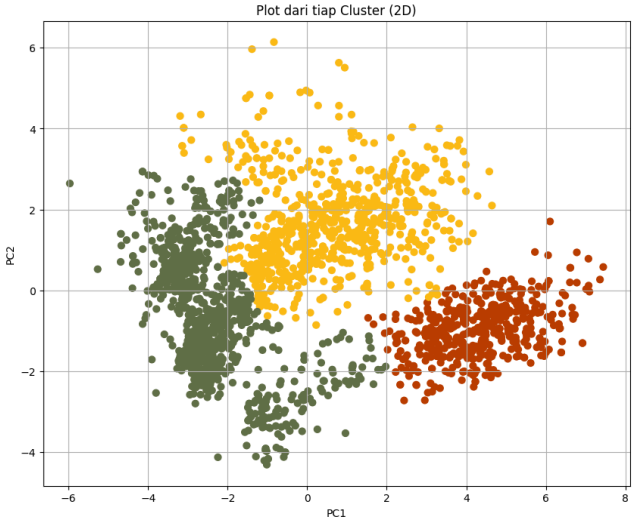
  3) *Clustering* untuk : *Cluster* = 3 dan PC1-PC2

     Didapat hasilnya:

     ```
     Silhouette Score: 0.465428915251249
     Calinski-Harabasz Index: 2827.2563909536175
     Davies-Bouldin Index: 0.7755982297672879
     ```

Hasil ini menunjukkan bahwa model *clustering* ini adalah yang terbaik dari sebelumnya dengan *cluster* lebih jelas dan terpisah, struktur internal lebih baik (kompak), dan tumpang tindih antar cluster minim.

4) Hasil Visualisasi antar *Cluster*



Gambar 16. Visualisasi hasil clustering

Dari gambar diatas dapat dilihat bahwa dengan menggunakan 3 cluster saja untuk membentuk algoritma *agglomerative clustering* dapat secara baik memisahkan antar cluster tanpa ada yang tumpang tindih, sehingga evaluasi lanjutan bisa dilakukan.

- **Evaluasi *Agglomerative Clustering***

Algoritma clustering seperti *Agglomerative Clustering* adalah algoritma *unsupervised* sehingga untuk mengevaluasinya tidak dengan akurasi. Sebagai gantinya adalah dengan melihat indikator seperti *Silhouette Score*, *Calinski-Harabasz Index*, dan *Davies-Bouldin Index*.

Silhouette Score berada pada kisaran -1 hingga 1, dengan nilai yang lebih tinggi menunjukkan hasil *clustering* yang lebih baik. *Calinski-Harabasz Index* menunjukkan rasio antara variansi antar *cluster* dengan variansi dalam *cluster*. Semakin tinggi nilai indeks ini, semakin baik pemisahan antar *cluster*. *Davies-Bouldin Index* mengukur sejauh mana *cluster* berlapis atau tumpang tindih. Nilai indeks ini semakin rendah, semakin baik.

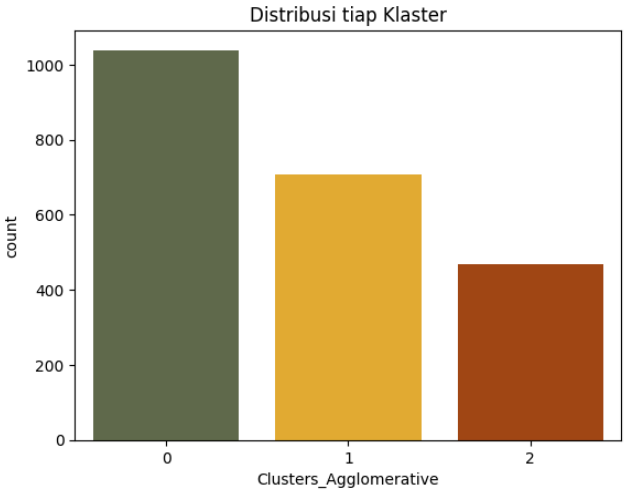
Sebelumnya telah didapat bahwa:

```
Silhouette Score: 0.465428915251249
Calinski-Harabasz Index: 2827.2563909536175
Davies-Bouldin Index: 0.7755982297672879
```

Hal ini menunjukkan bahwa PC1 dan PC2 ternyata sudah cukup untuk menangkap struktur penting dalam data, menambah komponen justru memasukkan *noise* atau variansi kecil yang mengaburkan *cluster*.

Beberapa evaluasi sekaligus EDA yang lain adalah sebagai berikut:

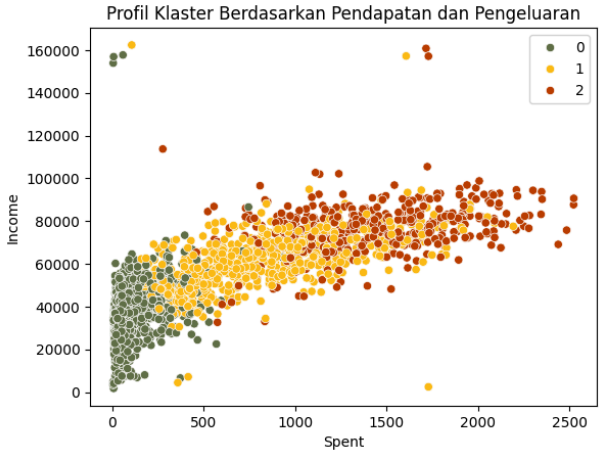
1) Melihat distribusi tiap *cluster.*



Gambar 16. Visualisasi Distribusi Tiap Cluster

Terlihat bahwa *cluster* 0 memiliki jumlah anggota terbanyak, diikuti oleh *cluster* 1, dan yang paling sedikit adalah *cluster* 3. Hal ini mengindikasikan bahwa pembagian *cluster* tidak sepenuhnya seimbang, namun tetap mencerminkan adanya variasi karakteristik antar kelompok dalam data.

2) Melihat *Cluster* berdasarkan Pendapatan dan Pengeluaran.

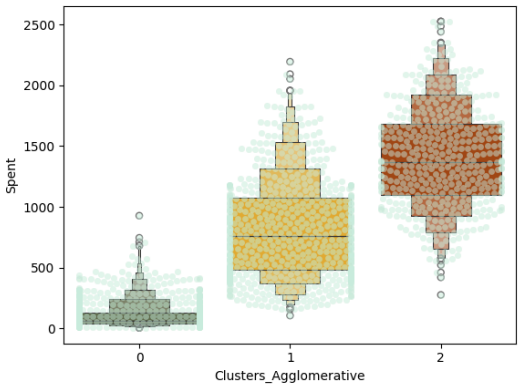


Gambar 17. Visualisasi Profil Cluster berdasarkan Pendapatan dan Pengeluaran

Terdapat perbedaan karakteristik yang cukup jelas antar cluster. Cluster 2 terdiri dari individu dengan pendapatan dan pengeluaran tinggi, mencerminkan daya beli besar. Cluster 0 berada di sisi sebaliknya, yaitu kelompok dengan pendapatan dan pengeluaran rendah, menunjukkan daya beli terbatas. Sementara itu, Cluster 1 cenderung memiliki

pendapatan rendah hingga sedang, namun dengan pengeluaran yang lebih rendah dari cluster lain, menggambarkan kelompok yang lebih hemat atau terbatas dalam konsumsi.

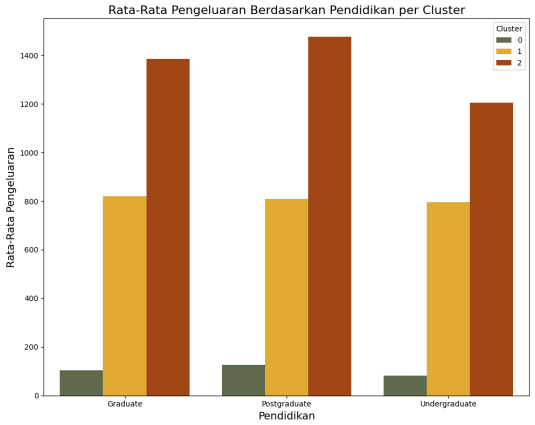
3) Melihat distribusi pengeluaran pada masing-masing *cluster.*



Gambar 18. Visualisasi distribusi pengeluaran masing-masing cluster

Visualisasi ini menunjukkan sebaran data, pusat nilai, dan outlier dalam tiap cluster. Cluster 2 memiliki pengeluaran tertinggi secara konsisten dengan distribusi yang lebar dan median tinggi, mencerminkan daya beli paling besar. Cluster 1 berada di posisi menengah dengan rentang pengeluaran yang cukup luas, menunjukkan variasi perilaku belanja yang tinggi dalam kelompok berpengeluaran sedang. Cluster 0 memiliki pengeluaran rendah dengan rentang dan median yang kecil.

4) Melihat rata-rata pengeluaran setiap cluster berdasarkan tingkat pendidikan.

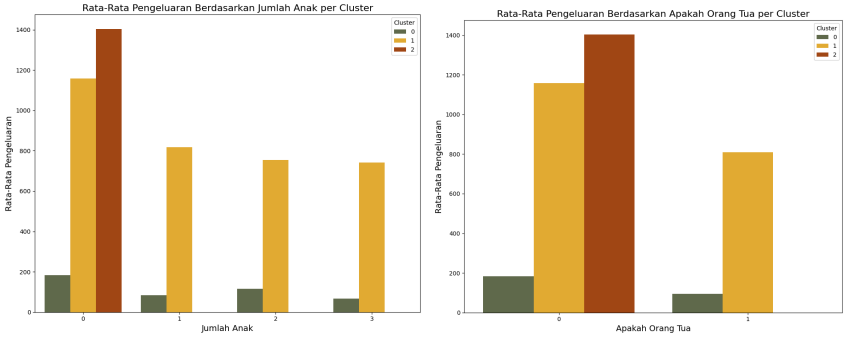


Gambar 19. Visualisasi rata-rata pengeluaran berdasar pendidikan

Dari visualisasi di atas, dapat dilihat bahwa untuk semua tingkat pendidikan, Cluster 2 memiliki rata-rata pengeluaran tertinggi, menunjukkan bahwa individu dalam cluster ini cenderung berbelanja lebih banyak terlepas dari latar belakang pendidikannya. Cluster 0 menunjukkan rata-rata pengeluaran yang rendah di setiap jenjang pendidikan, mencerminkan keterbatasan daya beli. Sementara itu,

Cluster 1 berada di posisi menengah dengan pengeluaran yang stabil untuk semua tingkatan pendidikan.

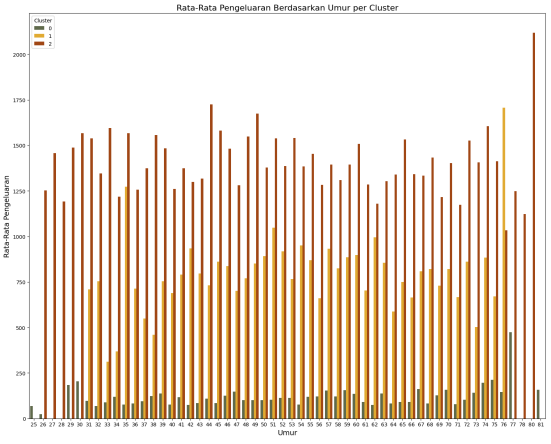
5) Melihat rata-rata pengeluaran berdasarkan status sebagai orang tua dan jumlah anak setiap cluster.



Gambar 19. Visualisasi rata-rata pengeluaran berdasar jumlah anak dan status orang tua

Visualisasi di atas menampilkan rata-rata pengeluaran setiap cluster berdasarkan jumlah anak dan status sebagai orang tua. Cluster 1 didominasi oleh individu yang merupakan orang tua dengan jumlah anak 0–3, dan memiliki pengeluaran menengah yang cukup stabil. Cluster 2 justru terdiri dari individu yang bukan orang tua dan tidak memiliki anak, namun menunjukkan pengeluaran tertinggi di antara semua cluster. Sementara itu, Cluster 0 berisi individu yang merupakan orang tua dengan jumlah anak 0–3 dan memiliki pengeluaran yang tergolong rendah.

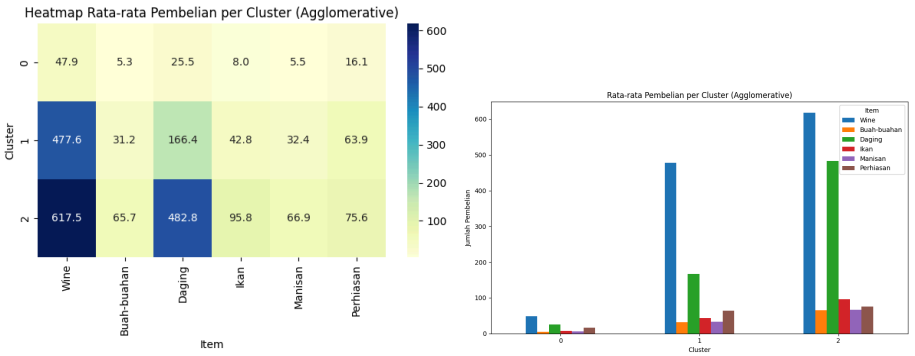
6) Visualisasi pengeluaran berdasarkan umur setiap cluster.



Gambar 20. Visualisasi rata-rata pengeluaran berdasarkan usia

Cluster 1 merupakan kelompok dengan rentang usia 31 hingga 76 tahun, menunjukkan konsumen dewasa hingga lansia dengan karakteristik pengeluaran yang bervariasi. Cluster 0 memiliki sebaran usia yang luas dan tidak terfokus pada rentang tertentu. Cluster 2 juga memiliki distribusi usia yang tersebar, namun menunjukkan pengeluaran yang sangat tinggi bahkan di usia yang masih muda.

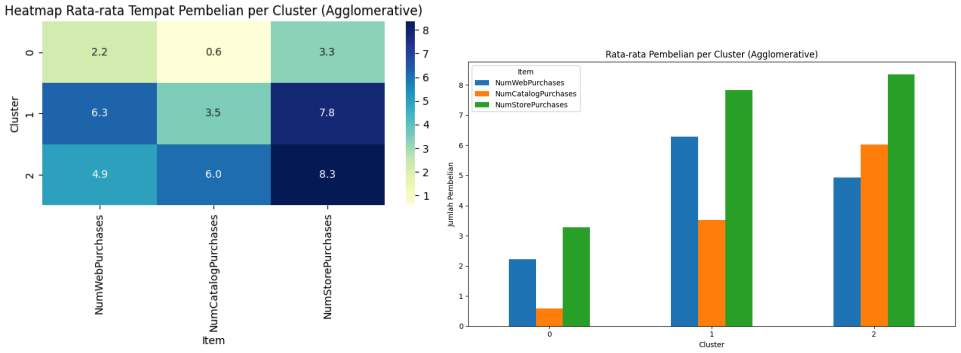
7) Visualisasi jumlah pembelian dan rata-rata pembelian per cluster.



Gambar 20. Visualisasi jumlah pembelian dan rata-rata pengeluaran barang-barang

Berdasarkan visualisasi dan data rata-rata pembelian di atas, Cluster 2 merupakan kelompok dengan pengeluaran tertinggi di semua kategori produk, terutama pada wine dan daging. Cluster 1 berada di posisi menengah dengan fokus pada wine dan daging serta pembelian cukup stabil pada kategori lain seperti ikan, manisan, dan perhiasan. Sebaliknya, Cluster 0 menunjukkan pengeluaran yang jauh lebih rendah di semua kategori, dengan wine sebagai item tertinggi mereka meskipun jumlahnya tetap kecil, mencerminkan karakteristik konsumen yang hemat atau dengan daya beli terbatas.

8) Visualisasi rata-rata tempat pembelian dan jumlah pembelian.



Gambar 21. Visualisasi jumlah pembelian dan rata-rata tempat pembelian barang

Berdasarkan visualisasi dan data rata-rata frekuensi pembelian di atas, Cluster 2 merupakan cluster dengan aktivitas belanja tertinggi di semua kategori, terutama melalui katalog (≈6.0) dan toko fisik (≈8.3). Cluster 1 juga cukup aktif, terutama dalam pembelian di toko fisik (≈7.8) dan web (≈6.3), namun lebih sedikit melalui katalog. Sementara itu, Cluster 0 menunjukkan frekuensi pembelian yang rendah di semua kategori, dengan toko fisik sebagai saluran paling umum, mencerminkan konsumen yang cenderung pasif atau hemat dalam berbelanja.

- ***Agglomerative Clustering* Tanpa PCA**

  Untuk melihat perbandingan hasil segmentasi dari algoritma *agglomerative* maka dilakukan pula, *clustering* dengan algoritma ini tanpa menggunakan PCA. Tujuannya adalah melihat apakah penggunaan PCA memiliki pengaruh yang signifikan terhadap hasil segmentasi. Tanpa PCA, semua fitur akan dipakai untuk membentuk *cluster.* Hasilnya nanti akan dibandingkan dengan yang menggunakan PCA dan akan dipilih yang terbaik. Berikut adalah hasil yang telah didapatkan:

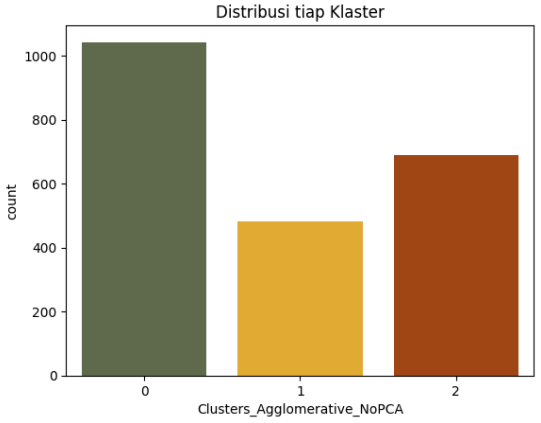
```
Silhouette Score: 0.17121259805155878
Calinski-Harabasz Index: 554.8762963365112
Davies-Bouldin Index: 1.9613773139843829
```

  Dari hasil di atas dapat dilihat bahwa nilai *silhouette score* dan *calinski-harabasz index* yang didapat masih sangat kecil jika dibandingkan dengan hasil *agglomerative clustering* yang menggunakan PCA. Selain itu, nilai *Davies-Bouldin Index* yang didapat pun masih relatif lebih besar jika dibandingkan dengan yang menggunakan PCA. Jika dilihat dari sini, dapat disimpulkan bahwa dengan penggunaan PCA memberikan kualitas *clustering* yang lebih bagus.

- **Evaluasi *Agglomerative Clustering* Tanpa PCA**

  Selain menggunakan pendekatan matrik evaluasi seperti yang dilakukan di atas, evaluasi dilakukan pula dengan melihat visualisasi sekaligus EDA dari hasil *clustering* dengan *agglomerative* tanpa PCA. Adapun visualisasi yang dilakukan:

  1) Melihat distribusi setiap *cluster*

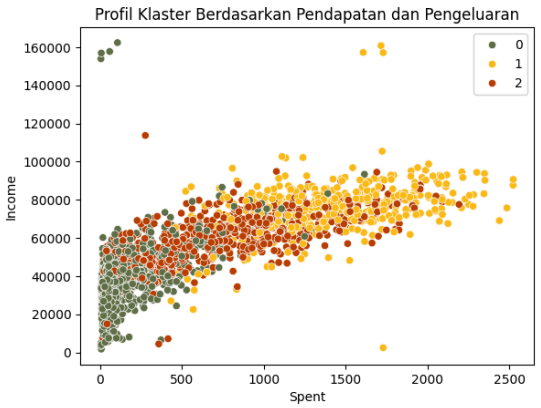


Gambar 28. Visualisasi Distribusi Tiap Cluster

  Dari hasil diatas terlihat bahwa *cluster* 0 memiliki jumlah anggota terbanyak sama seperti yang menggunakan PCA dan diikuti oleh *cluster 2,*

dan yang paling sedikit adalah *cluster 1.* Hal ini menunjukkan bahwa distribusi setiap *cluster* tidak seimbang namun bervariasi.

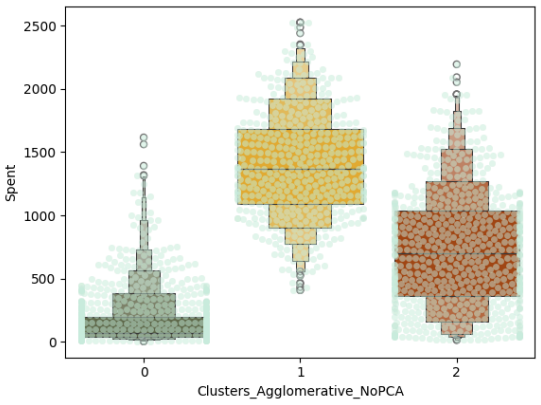
2) Melihat *Cluster* berdasarkan Pendapatan dan Pengeluaran.



Gambar 29. Visualisasi Profil Cluster berdasarkan Pendapatan dan Pengeluaran

Terdapat perbedaan karakteristik antar *cluster. Cluster* 1 merupakan cluster dengan individu yang memiliki pengeluaran dan pendapatan tertinggi dari semua cluster yang ada, dan disusul cluster 2 dengan individu yang memiliki pendapatan dan pengeluaran yang menengah. Sementara itu, untuk cluster 0 memiliki pengeluaran dan pendapatan yang relatif rendah jika dibandingkan dengan kelompok cluster yang lain.

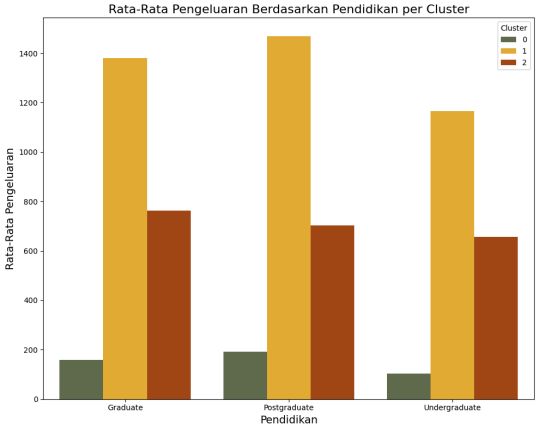
3) Melihat distribusi pengeluaran pada masing-masing *cluster.*



Gambar 30. Visualisasi distribusi pengeluaran masing-masing cluster

Visualisasi ini menunjukkan sebaran data, pusat nilai, dan outlier dalam tiap cluster. Cluster 1 memiliki pengeluaran tertinggi secara konsisten dengan nilai median yang cukup tinggi, dimana hal ini mencerminkan daya beli yang sangat tinggi. Berbanding terbalik dengan cluster 1, cluster 0 justru memiliki pengeluaran yang sangat rendah dan menunjukkan individu yang paling hemat dari semua cluster yang ada. Dan untuk cluster 2 merupakan cluster dengan pengeluaran individunya menengah.

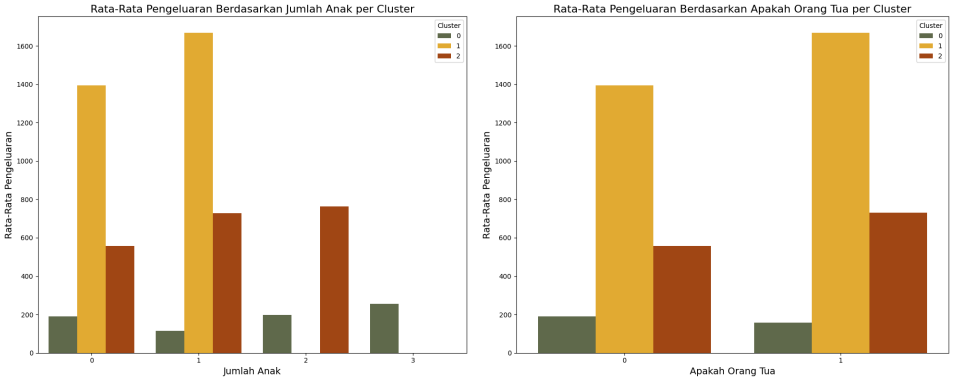
4) Melihat rata-rata pengeluaran setiap cluster berdasarkan tingkat pendidikan.



Gambar 31. Visualisasi rata-rata pengeluaran berdasar pendidikan

Dari visualisasi di atas, dapat dilihat bahwa untuk semua tingkat pendidikan, Cluster 1 memiliki rata-rata pengeluaran tertinggi, menunjukkan bahwa individu dalam cluster ini cenderung berbelanja lebih banyak terlepas dari latar belakang pendidikannya. Cluster 0 menunjukkan rata-rata pengeluaran yang rendah di setiap jenjang pendidikan, mencerminkan keterbatasan daya beli. Sementara itu, Cluster 2 berada di posisi menengah dengan pengeluaran yang stabil untuk semua tingkatan pendidikan.

5) Melihat rata-rata pengeluaran berdasarkan status sebagai orang tua dan jumlah anak setiap cluster.

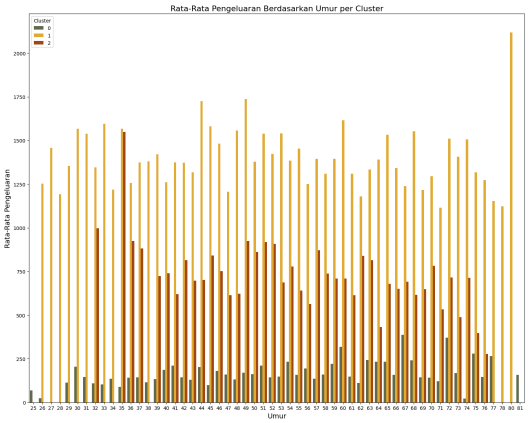


Gambar 32. Visualisasi rata-rata pengeluaran berdasar jumlah anak dan status orang tua

Visualisasi di atas menampilkan rata-rata pengeluaran setiap cluster berdasarkan jumlah anak dan status sebagai orang tua. Mulai dari cluster 0, 1 dan 2 semuanya merupakan orang tua dengan jumlah anak yang beragam. Cluster 0 berisi individu dengan jumlah anak 0-3 dengan pengeluaran yang sangatr rendah. Sementara itu, cluster 1 terdiri dari individu yang memiliki anak 0-1 dengan pengeluaran yang tertinggi

diantara yang lainnya. Untuk cluster 2 merupakan individu dengan jumlah anak 0-2 dan memiliki pengeluaran yang menengah.

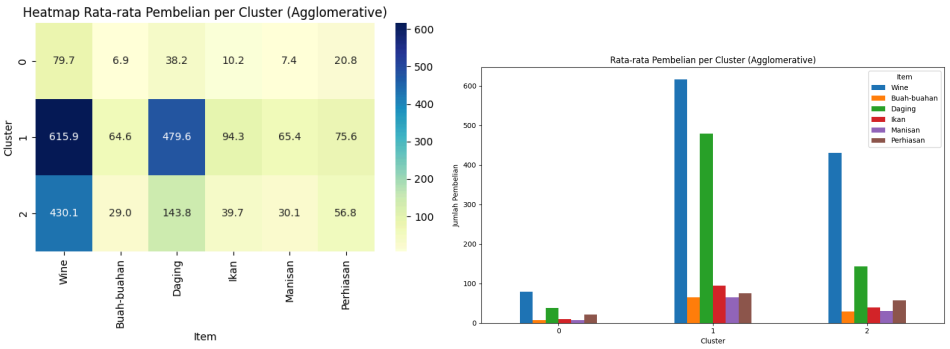
6) Visualisasi pengeluaran berdasarkan umur setiap cluster.



Gambar 32. Visualisasi rata-rata pengeluaran berdasarkan usia

Cluster 2 merupakan kelompok dengan rentang usia 32 hingga 76 tahun, menunjukkan konsumen dewasa hingga lansia dengan karakteristik pengeluaran yang bervariasi. Cluster 0 memiliki sebaran usia yang luas dan tidak terfokus pada rentang tertentu. Cluster 1 juga memiliki distribusi usia yang tersebar, namun menunjukkan pengeluaran yang sangat tinggi bahkan di usia yang masih muda.

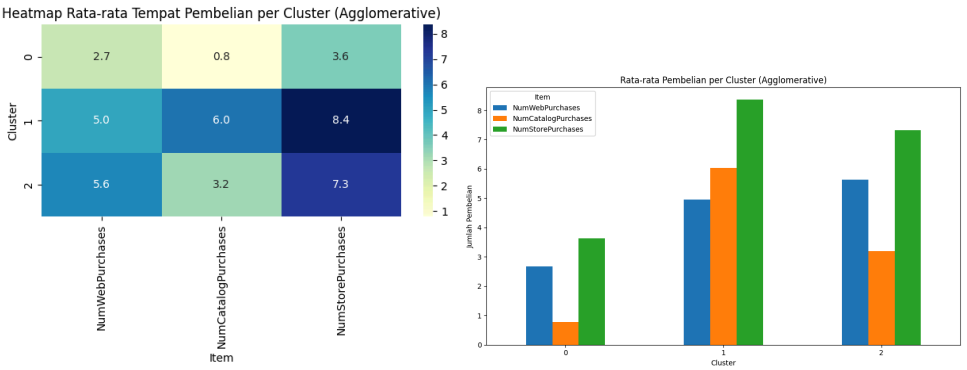
7) Visualisasi jumlah pembelian dan rata-rata pembelian per cluster.



Gambar 33. Visualisasi jumlah pembelian dan rata-rata pengeluaran barang-barang

Berdasarkan visualisasi dan data rata-rata pembelian di atas, Cluster 1 merupakan kelompok dengan pengeluaran tertinggi di semua kategori produk, terutama pada wine dan daging. Cluster 2 berada di posisi menengah dengan fokus pada wine dan daging serta pembelian cukup stabil pada kategori lain seperti ikan, manisan, dan perhiasan. Sebaliknya, Cluster 0 menunjukkan pengeluaran yang jauh lebih rendah di semua kategori, dengan wine sebagai item tertinggi mereka meskipun jumlahnya tetap kecil, mencerminkan karakteristik konsumen yang hemat atau dengan daya beli terbatas.

8) Visualisasi rata-rata tempat pembelian dan jumlah pembelian.



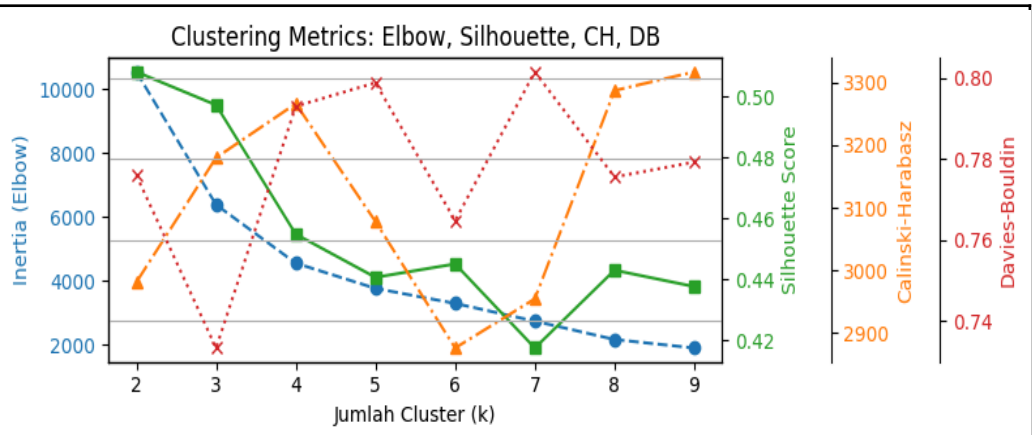
Gambar 34. Visualisasi jumlah pembelian dan rata-rata tempat pembelian barang

Berdasarkan visualisasi dan data rata-rata frekuensi pembelian di atas, Cluster 1 merupakan cluster dengan aktivitas belanja tertinggi di semua kategori, terutama melalui katalog (≈6.0) dan toko fisik (≈8.4). Cluster 2 juga cukup aktif, terutama dalam pembelian di toko fisik (≈7.3) dan web (≈5.6), namun lebih sedikit melalui katalog. Sementara itu, Cluster 0 menunjukkan frekuensi pembelian yang rendah di semua kategori, dengan toko fisik sebagai saluran paling umum, mencerminkan konsumen yang cenderung pasif atau hemat dalam berbelanja.

- ***K-Means Clustering***

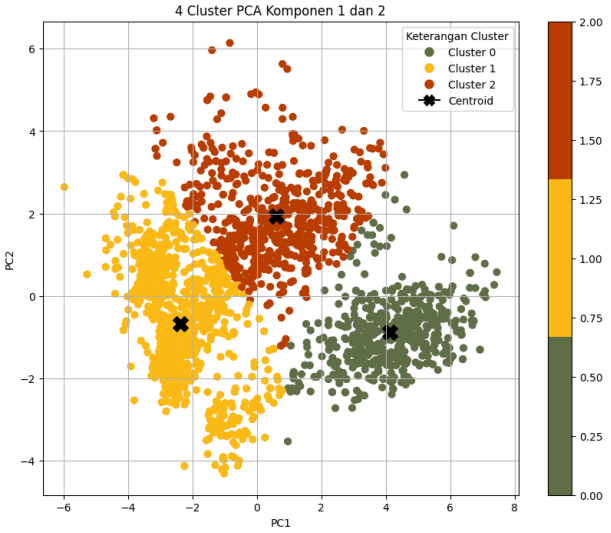
  K Means Clustering merupakan salah satu metode *unsupervised learning* yang termasuk dalam kategori partition-based clustering. Algoritma ini mengelompokkan data ke dalam sejumlah cluster yang telah ditentukan sebelumnya dengan cara meminimalkan jarak antar data dan pusat cluster (centroid), umumnya menggunakan metrik jarak Euclidean. Pendekatan ini bekerja secara iteratif untuk mengoptimalkan posisi centroid sehingga tiap data berada dalam cluster dengan kemiripan tertinggi.

  Pada bagian sebelumnya, jumlah cluster telah ditentukan menggunakan metode elbow dan evaluasi metrik *Agglomerative Clustering* lainnya, yaitu sebanyak 3 cluster. Berdasarkan grafik skor & indeks parameter penilaian kualitas cluster lebih lanjut, berikut tiap pola proyeksi pada *K Means*.



Gambar 35. Visualisasi Penentuan Jumlah Cluster

Agar hasil pembagian kelompok data kebiasaan konsumen bisa diinterpretasikan & diterapkan perbandingan antar performa model, maka jumlah cluster serta *PCA* yang dilibatkan berdasarkan aturan yang sudah terpilih, akan sama. Selanjutnya, untuk mengartikan setiap pemisahan cluster, tidak akan ada perbedaan signifikan dengan yang sudah dinyatakan pada metode *clustering* sebelumnya.



Gambar 36. Visualisasi Hasil Clustering

Dari hasil visualisasi clustering di atas dapat disimpulkan bahwa dengan menggunakan 2 PC mampu membentuk hasil clustering yang baik tanpa adanya tumpang tindih antar cluster.

- **Evaluasi Algoritma *K-Means Clustering***

  K-Means merupakan algoritma *unsupervised* yang dievaluasi menggunakan metrik seperti *Silhouette Score, Calinski-Harabasz Index,* dan *Davies-Bouldin Index.* Nilai *Silhouette* dan *Calinski-Harabasz* yang tinggi serta *Davies-Bouldin* yang rendah menunjukkan kualitas clustering yang

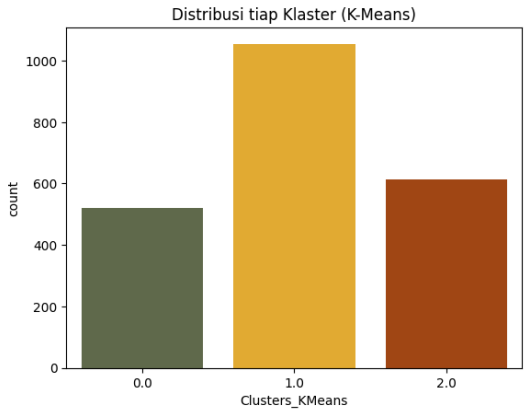
baik, dengan pemisahan antar cluster yang jelas dan struktur yang solid. Berikut adalah hasilnya:

```
Silhouette Score: 0.49717267008432164
Calinski-Harabasz Index: 3179.8098781803737
Davies-Bouldin Index: 0.7333450545790788
```

Sama seperti algoritma *Agglomerative* di atas, hasil ini menunjukkan bahwa PC1 dan PC2 ternyata sudah cukup untuk menangkap struktur penting dalam data, menambah komponen justru memasukkan *noise* atau variansi kecil yang mengaburkan *cluster*.

Beberapa evaluasi sekaligus EDA yang lain adalah sebagai berikut:

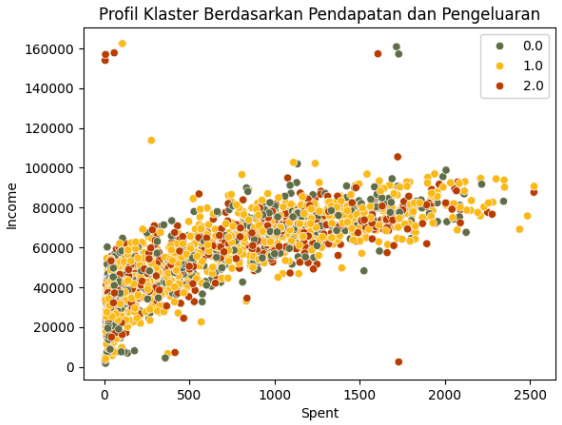
1) Melihat distribusi setiap cluster.



Gambar 37. Distribusi tiap Cluster

Dari hasil distribusi di atas, terlihat bahwa cluster 1 memiliki jumlah anggota terbanyak, diikuti cluster 2, dan yang paling sedikit adalah cluster 0. Hal ini menunjukkan bahwa data tersebut dan ada variasi karakteristik antar kelompok dalam data.

2) Melihat *Cluster* berdasarkan Pendapatan dan Pengeluaran.

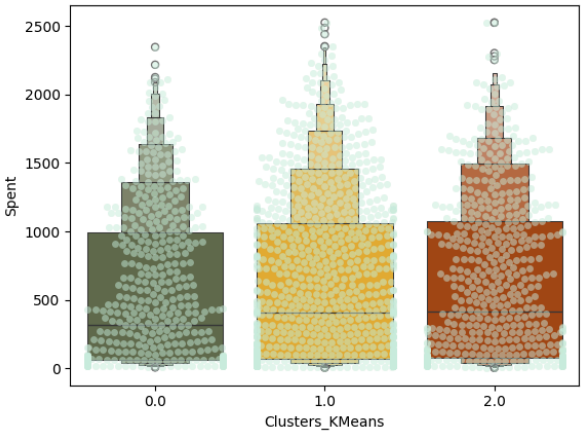


Gambar 38. Visualisasi Profil Cluster berdasarkan Pendapatan dan Pengeluaran

Dari hasil di atas terlihat bahwa secara umum terdapat pola positif, di mana semakin tinggi pendapatan, pengeluaran juga cenderung meningkat. Namun, ketiga cluster (0–2) terlihat bertumpang tindih,

yang artinya pemisahan antar cluster tidak terlalu tegas. Setiap cluster mencakup individu dengan rentang pendapatan dan pengeluaran yang cukup luas.

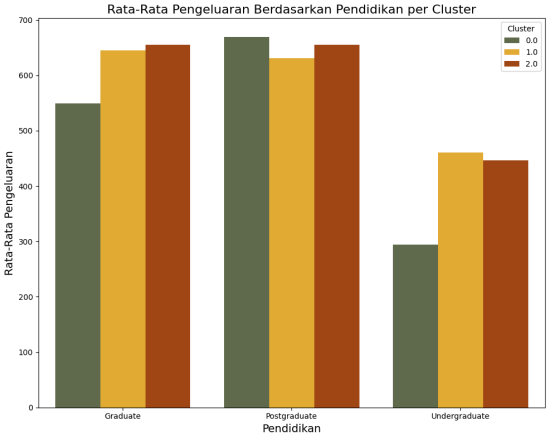
3) Melihat distribusi pengeluaran pada masing-masing *cluster.*



Gambar 39. Visualisasi distribusi pengeluaran masing-masing cluster

Terlihat bahwa semua cluster memiliki rentang pengeluaran yang hampir serupa, dengan distribusi yang cukup lebar dan adanya beberapa outlier (titik ekstrem di bagian atas). Cluster 1 tampak memiliki sedikit kecenderungan ke pengeluaran yang lebih tinggi dibanding cluster lainnya, sementara cluster 0 menunjukkan sebaran yang sedikit lebih sempit, mengindikasikan kelompok dengan perilaku belanja yang lebih seragam.

4) Melihat rata-rata pengeluaran setiap cluster berdasarkan tingkat pendidikan.

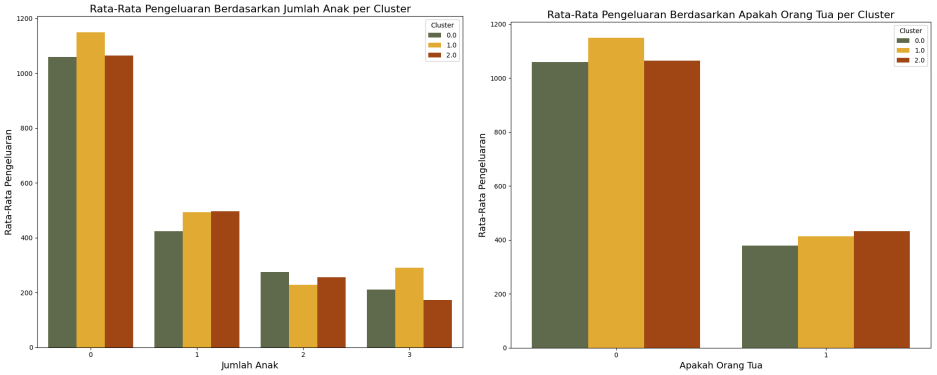


Gambar 40. Visualisasi rata-rata pengeluaran berdasar pendidikan

Hasil di atas menunjukkan rata-rata pengeluaran berdasarkan tingkat pendidikan di setiap cluster hasil *K-Means Clustering*. Terlihat bahwa semua cluster memiliki pola pengeluaran yang cukup merata di semua jenjang pendidikan, tanpa perbedaan yang terlalu mencolok. Untuk jenjang pendidikan *graduate* pengeluaran tertinggi ada pada cluster 2.

Sementara itu, untuk jenjang *postgraduate,* cluster 0 merupakan cluster dengan pengeluaran paling tinggi. Dan, untuk jenjang *undergraduate*, cluster 1 lah yang memiliki pengeluaran paling tinggi.

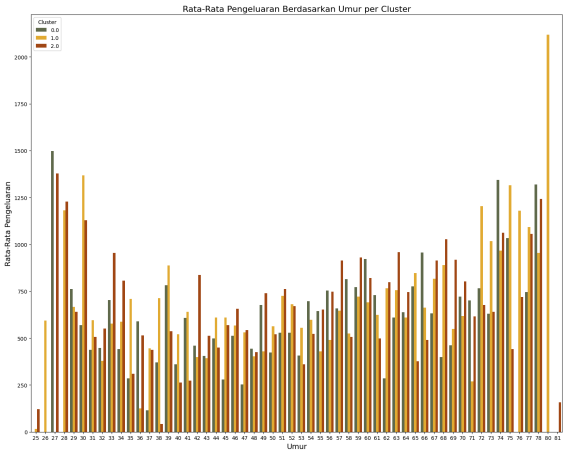
5) Melihat rata-rata pengeluaran berdasarkan status sebagai orang tua dan jumlah anak setiap cluster.



Gambar 41. Visualisasi rata-rata pengeluaran berdasar jumlah anak dan status orang tua

Dari hasil visualisasi di atas, seluruh pengeluaran individu yang bukan orang tua lebih tinggi dibandingkan dengan mereka yang sudah menjadi orang tua. Cluster 0 dan 2 menunjukkan penurunan pengeluaran yang tajam seiring bertambahnya jumlah anak. Sedangkan Cluster 3, meskipun juga terdiri dari individu dengan anak 0–3, menunjukkan bahwa mereka yang tidak memiliki anak memiliki daya beli paling tinggi dalam cluster tersebut. Pola ini menegaskan bahwa status sebagai orang tua dan jumlah anak berpengaruh signifikan terhadap perilaku pengeluaran di semua cluster.

6) Visualisasi pengeluaran berdasarkan umur setiap cluster.

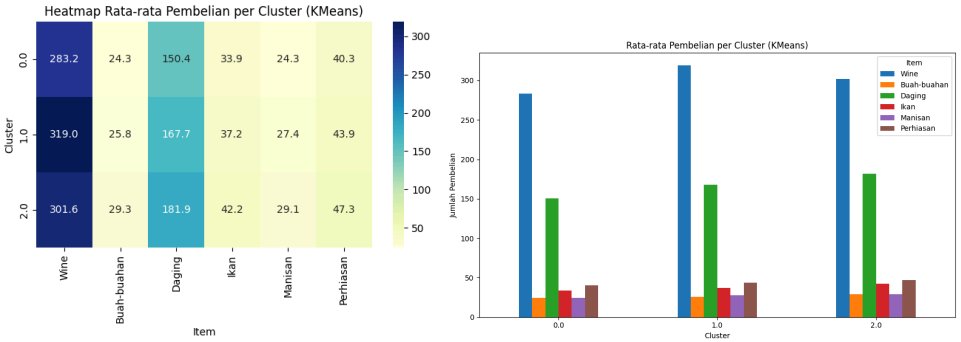


Gambar 34. Visualisasi rata-rata pengeluaran berdasarkan usia

Visualisasi berdasarkan umur menunjukkan bahwa semua cluster (0–2) memiliki sebaran yang luas dan mencakup hampir seluruh rentang usia, dari usia muda hingga lanjut. Tidak ada cluster yang secara dominan

terfokus pada kelompok usia tertentu, sehingga tidak ditemukan kecenderungan khusus antara umur dengan segmentasi cluster.

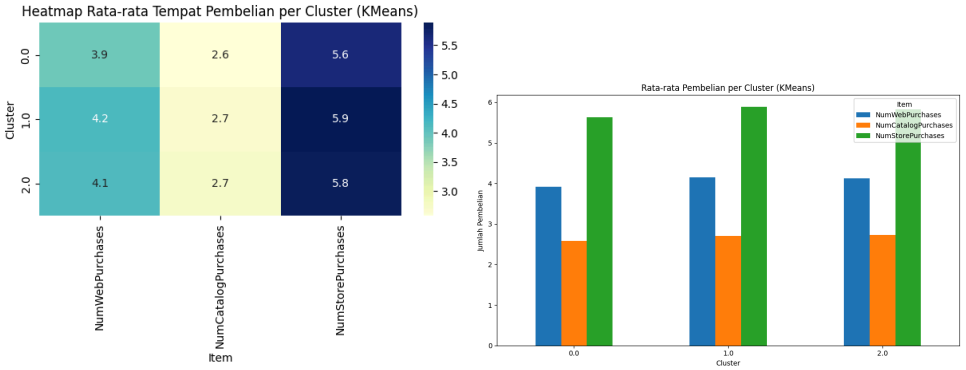
7) Visualisasi jumlah pembelian dan rata-rata pembelian per cluster.



Gambar 35. Visualisasi jumlah pembelian dan rata-rata pengeluaran barang-barang

Hasil visualisasi menunjukkan bahwa di semua cluster, wine dan daging adalah dua kategori dengan rata-rata pembelian tertinggi. Cluster 2 mencatat pembelian tertinggi untuk hampir semua item, terutama wine (≈301.6) dan daging (≈181.9), diikuti oleh Cluster 1. Sebaliknya, Cluster 0 memiliki rata-rata pembelian terendah secara umum, meskipun tetap cukup tinggi untuk wine dan daging. Perhiasan, ikan, manisan, dan buah-buahan menunjukkan pola peningkatan bertahap dari Cluster 0 ke Cluster 2.

8) Visualisasi rata-rata tempat pembelian dan jumlah pembelian.



Gambar 36. Visualisasi jumlah pembelian dan rata-rata tempat pembelian barang

Visualisasi di atas menunjukkan bahwa di semua cluster hasil K-Means, pembelian di toko fisik memiliki frekuensi tertinggi dibandingkan kanal lainnya. Cluster 1 memiliki rata-rata pembelian tertinggi di ketiga saluran, terutama di toko fisik (≈5.9) dan web (≈4.2), menandakan konsumen yang paling aktif. Cluster 0 memiliki frekuensi pembelian paling rendah di web dan katalog, meskipun masih cukup tinggi di toko. Secara umum, seluruh cluster menunjukkan preferensi terbesar terhadap belanja di toko fisik, diikuti oleh web, dan paling sedikit melalui katalog.

- ***K-Means Clustering* Tanpa PCA**

  Sama halnya dengan *agglomerative* yang diuji dengan tanpa PCA, *K-Means* juga diuji dengan tanpa PCA. Tujuannya tentu juga untuk melihat apakah penggunaan PCA memiliki pengaruh yang signifikan terhadap hasil segmentasi. Tanpa PCA, semua fitur akan dipakai untuk membentuk *cluster.* Hasilnya nanti akan dibandingkan dengan yang menggunakan PCA dan akan dipilih yang terbaik. Berikut adalah hasil yang telah didapatkan:

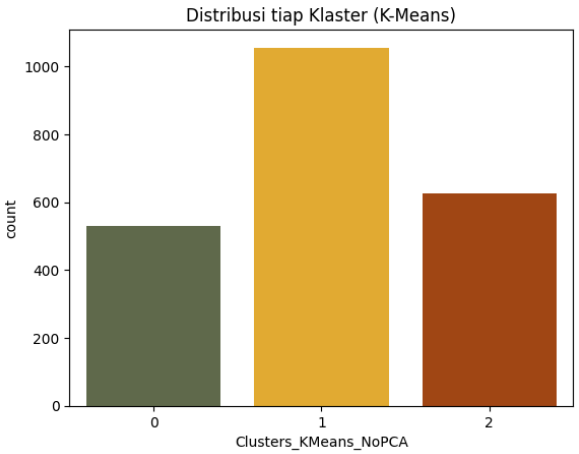
```
Silhouette Score: 0.20932021366754708
Calinski-Harabasz Index: 644.5026600135105
Davies-Bouldin Index: 1.7693273274270638
```

  Dari hasil di atas dapat dilihat bahwa hasil yang didapat sangat kecil jika dibandingkan dengan hasil *K-Means clustering* yang menggunakan PCA. Jika dilihat dari sini, dapat disimpulkan bahwa dengan penggunaan PCA memberikan kualitas *clustering* yang lebih bagus.

- **Evaluasi *K-Means Clustering* Tanpa PCA**

  Evaluasi juga dilakukan dengan melihat visualisasi sebagai berikut:

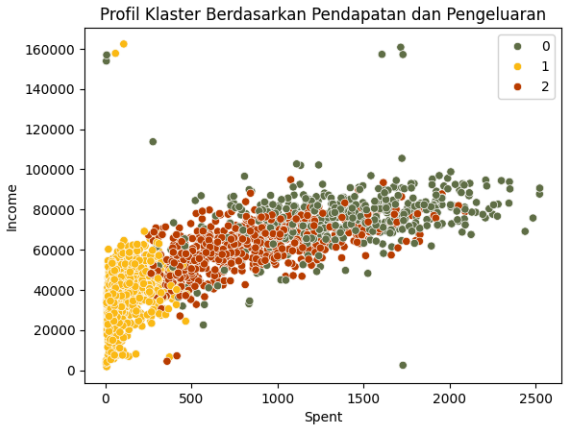
  1) Melihat distribusi setiap *cluster*



Gambar 37. Visualisasi Distribusi Tiap Cluster

  Dari hasil diatas terlihat bahwa hasil visualisasi tidak terlihat berbeda secara signifikan jika dibandingkan dengan *K-Means* yang menggunakan PCA.

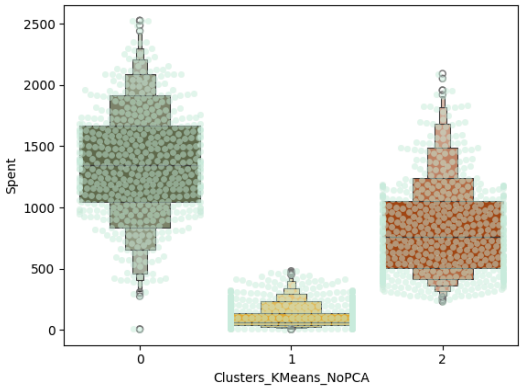
  2) Melihat *Cluster* berdasarkan Pendapatan dan Pengeluaran.



Gambar 38. Visualisasi Profil Cluster berdasarkan Pendapatan dan Pengeluaran

Terdapat perbedaan karakteristik antar *cluster. Cluster* 0 merupakan cluster dengan individu yang memiliki pengeluaran dan pendapatan tertinggi dari semua cluster yang ada, dan disusul cluster 2 dengan individu yang memiliki pendapatan dan pengeluaran yang menengah. Sementara itu, untuk cluster 1 memiliki pengeluaran dan pendapatan yang relatif rendah jika dibandingkan dengan kelompok cluster yang lain.

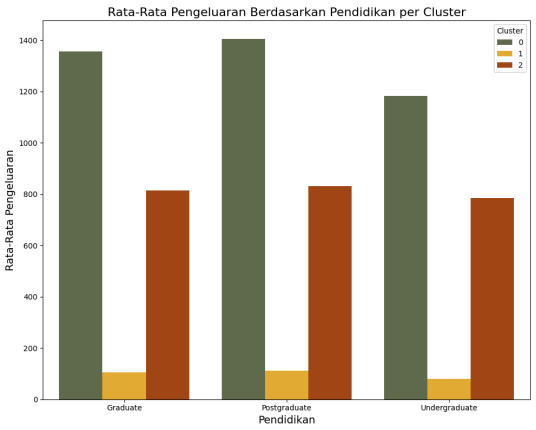
3) Melihat distribusi pengeluaran pada masing-masing *cluster.*



Gambar 39. Visualisasi distribusi pengeluaran masing-masing cluster

Visualisasi ini menunjukkan sebaran data, pusat nilai, dan outlier dalam tiap cluster. Cluster 0 memiliki pengeluaran tertinggi secara konsisten dengan nilai median yang cukup tinggi, dimana hal ini mencerminkan daya beli yang sangat tinggi. Berbanding terbalik dengan cluster 0, cluster 1 justru memiliki pengeluaran yang sangat rendah dan menunjukkan individu yang paling hemat dari semua cluster yang ada. Dan untuk cluster 2 merupakan cluster dengan pengeluaran individunya menengah.

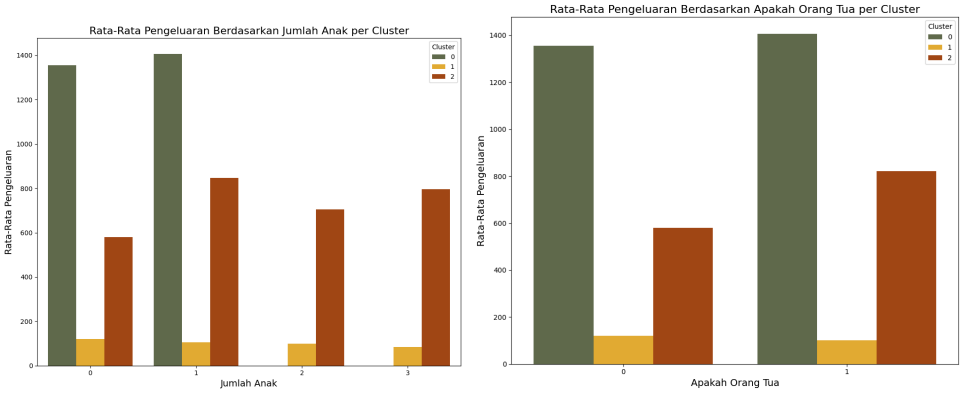
4) Melihat rata-rata pengeluaran setiap cluster berdasarkan tingkat pendidikan.



Gambar 40. Visualisasi rata-rata pengeluaran berdasar pendidikan

Dari visualisasi di atas, dapat dilihat bahwa untuk semua tingkat pendidikan, Cluster 0 memiliki rata-rata pengeluaran tertinggi, menunjukkan bahwa individu dalam cluster ini cenderung berbelanja lebih banyak terlepas dari latar belakang pendidikannya. Cluster 1 menunjukkan rata-rata pengeluaran yang rendah di setiap jenjang pendidikan, mencerminkan keterbatasan daya beli. Sementara itu, Cluster 2 berada di posisi menengah dengan pengeluaran yang stabil untuk semua tingkatan pendidikan.

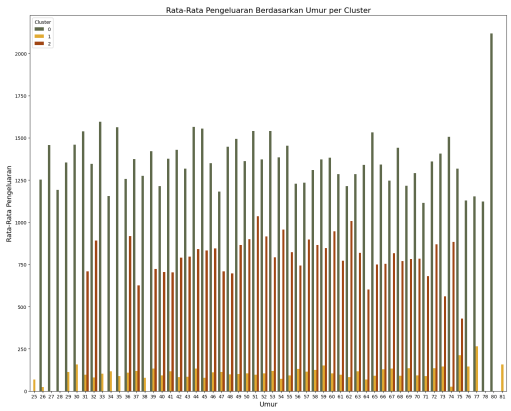
5) Melihat rata-rata pengeluaran berdasarkan status sebagai orang tua dan jumlah anak setiap cluster.



Gambar 41. Visualisasi rata-rata pengeluaran berdasar jumlah anak dan status orang tua

Visualisasi di atas menampilkan rata-rata pengeluaran setiap cluster berdasarkan jumlah anak dan status sebagai orang tua. Mulai dari cluster 0, 1 dan 2 semuanya merupakan orang tua dengan jumlah anak yang beragam. Cluster 1 berisi individu dengan jumlah anak 0-3 dengan pengeluaran yang sangat rendah. Sementara itu, cluster 0 terdiri dari individu yang memiliki anak 0-1 dengan pengeluaran yang tertinggi diantara yang lainnya. Untuk cluster 2 merupakan individu dengan jumlah anak 0-3 dan memiliki pengeluaran yang menengah.

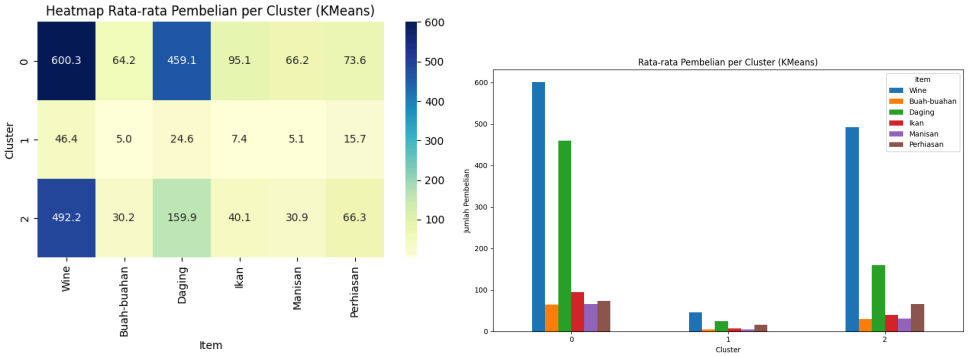
6) Visualisasi pengeluaran berdasarkan umur setiap cluster.



Gambar 42. Visualisasi rata-rata pengeluaran berdasarkan usia

Cluster 2 merupakan kelompok dengan rentang usia 31 hingga 75 tahun, menunjukkan konsumen dewasa hingga lansia dengan karakteristik pengeluaran yang bervariasi. Cluster 1 memiliki sebaran usia yang luas dan tidak terfokus pada rentang tertentu. Cluster 0 juga memiliki distribusi usia yang tersebar, namun menunjukkan pengeluaran yang sangat tinggi bahkan di usia yang masih muda.

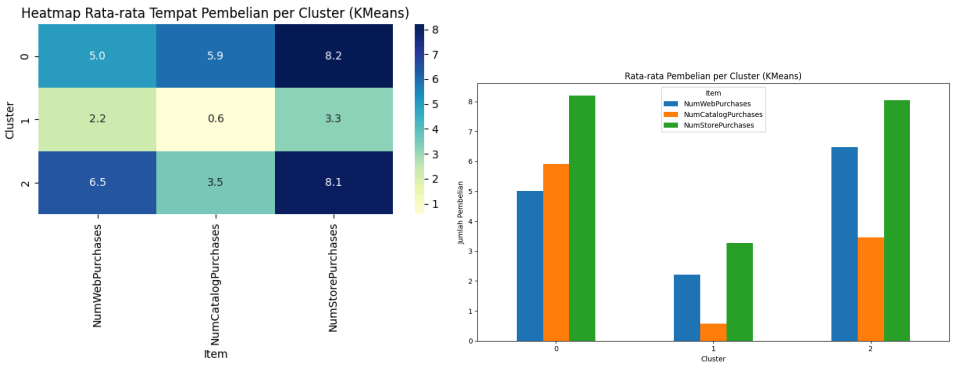
7) Visualisasi jumlah pembelian dan rata-rata pembelian per cluster.



Gambar 43. Visualisasi jumlah pembelian dan rata-rata pengeluaran barang-barang

Berdasarkan visualisasi dan data rata-rata pembelian di atas, Cluster 0 merupakan kelompok dengan pengeluaran tertinggi di semua kategori produk, terutama pada wine dan daging. Cluster 2 berada di posisi menengah dengan fokus pada wine dan daging serta pembelian cukup stabil pada kategori lain seperti ikan, manisan, dan perhiasan. Sebaliknya, Cluster 1 menunjukkan pengeluaran yang jauh lebih rendah di semua kategori, dengan wine sebagai item tertinggi mereka meskipun jumlahnya tetap kecil, mencerminkan karakteristik konsumen yang hemat atau dengan daya beli terbatas.

8) Visualisasi rata-rata tempat pembelian dan jumlah pembelian.



Gambar 44. Visualisasi jumlah pembelian dan rata-rata tempat pembelian barang

Berdasarkan visualisasi dan data rata-rata frekuensi pembelian di atas, Cluster 0 merupakan cluster dengan aktivitas belanja tertinggi di semua kategori, terutama melalui katalog (≈5.9) dan toko fisik (≈8.2). Cluster 2 juga cukup aktif, terutama dalam pembelian di toko fisik (≈8.1) dan web (≈6.5), namun lebih sedikit melalui katalog. Sementara itu, Cluster 1 menunjukkan frekuensi pembelian yang rendah di semua kategori, dengan toko fisik sebagai saluran paling umum, mencerminkan konsumen yang cenderung pasif atau hemat dalam berbelanja.

3.2 Insight dan Hasil Mining dari Project

Berdasarkan hasil clustering menggunakan algoritma *Agglomerative Clustering* dan *K-Means Clustering* dengan jumlah *cluster* optimal sebanyak 4 (berdasarkan *Elbow Method*), diperoleh segmentasi konsumen yang memiliki karakteristik berbeda. Dari proses ini, diperoleh beberapa insight penting yang dapat dimanfaatkan dalam menyusun strategi promosi di pusat perbelanjaan.

- ***Agglomerative Clustering***

  Berdasarkan algoritma *Agglomerative Clustering* dengan jumlah optimal 3 cluster, diperoleh pembagian pelanggan yang memiliki berbagai karakteristik unik. Berikut adalah ringkasan insight per cluster berdasarkan hasil yang telah didapatkan.

Cluster 0	Cluster 1	Cluster 2
Cluster dengan jumlah terbanyak sekitar 1.000+ orang.	Cluster dengan jumlah terbanyak kedua sekitar 700 orang.	Cluster dengan jumlah sekitar 460 orang
Rata-rata pendapatan dan daya beli nya rendah.	Rata-rata pendapatan dan pengeluarannya menengah.	Rata-rata tingkat pengeluaran dan pendapatan tertinggi.

Latar belakang pendidikan yang menyebar di semua bagian namun semuanya berdaya beli rendah.	Latar pendidikan beragam, namun didominasi oleh *graduate dan postgraduate* dengan pengeluaran stabil.	Latar belakang pendidikan Anggotanya umumnya berpendidikan tinggi dengan pengeluaran tinggi di setiap bagian.
Pengeluaran seseorang yang bukan orang tua lebih tinggi dari pada yang sudah menjadi orang tua.	Pengeluaran seseorang yang bukan orang tua lebih tinggi dari pada yang sudah menjadi orang tua.	Bukan merupakan orang tua sehingga pengeluarannya sangat tinggi.
Jumlah anak pada cluster ini ada dalam rentang 0-3 dengan pengeluaran kurang dari $200.	Pengeluarannya menengah dan semakin menurun seiring bertambahnya jumlah anak.	Tidak memiliki satupun anak dan memiliki pengeluaran mendekati $1.400
Usia tersebar di berbagai rentang.	Merupakan kelompok dengan rentang usia 31-76 Tahun.	Usia tersebar di berbagai rentang, dan pengeluarannya sangat tinggi walaupun di usia yang muda.
Cluster 0 menghabiskan paling banyak untuk wine, meskipun jumlahnya tetap rendah dibanding kelompok lain.	Cluster 1 paling banyak menghabiskan uang untuk wine (≈478) dan daging (≈166).	Cluster 2 adalah pengeluar terbesar di semua kategori, terutama pada wine (≈617), daging (≈483), dan ikan (≈96).
Cluster 0 memiliki frekuensi belanja terendah di semua kelompok. Mereka jarang berbelanja, baik online, katalog, maupun di toko fisik.	Cluster 1 paling sering melakukan pembelian di toko fisik (≈7.8) dan web (≈6.3). Mereka jarang belanja lewat katalog.	Cluster 2 paling menonjol dalam pembelian melalui katalog (≈6.0) dan juga cukup tinggi di toko fisik (≈8.3) serta web (≈4.9).
Karena karakteristiknya yang *price-sensitive,* strategi promosi yang sesuai untuk cluster ini adalah diskon besar, cashback, serta kampanye promosi yang masif melalui media digital.	Cluster ini cocok untuk diberikan promosi tentang produk kebutuhan keluarga, bundling, dan program loyalitas jangka panjang.	Cluster ini sangat potensial untuk ditargetkan dengan produk eksklusif, layanan premium, serta program membership VIP atau personalisasi pengalaman berbelanja.

Segmentasi yang dihasilkan melalui algoritma *Agglomerative Clustering* membantu memahami perilaku konsumen dari sisi ekonomi dan demografi. Setiap cluster memiliki karakteristik berbeda yang memerlukan pendekatan pemasaran yang spesifik. Cluster 2, sebagai kelompok dengan daya beli tertinggi, sebaiknya difokuskan pada strategi yang menekankan keuntungan tinggi dan loyalitas pelanggan, seperti layanan eksklusif atau program VIP.

Sementara itu, Cluster 0 yang memiliki daya beli lebih rendah cocok disasar dengan strategi berbasis kuantitas dan frekuensi pembelian, seperti diskon besar atau paket hemat. Adapun Cluster 1, yang cenderung stabil, lebih sesuai diberi pendekatan yang menyeimbangkan variasi produk dengan kenyamanan berbelanja, misalnya program bundling dan diskon keluarga. Pendekatan yang berbeda ini akan meningkatkan efektivitas promosi dan retensi pelanggan di setiap segmen.

- ***Agglomerative Clustering tanpa PCA***

  Meskipun hasil dari *agglomerative clustering* tanpa PCA maupun dengan PCA mirip, namun insight yang didapat bisa berbeda. Berikut adalah insight yang didapat dari hasil segmentasi ini.

Cluster 0	Cluster 1	Cluster 2
Cluster dengan jumlah anggota terbanyak, yakni di atas 1.000 orang.	Cluster dengan jumlah anggota paling sedikit, sekitar 550 orang.	Anggota kelompoknya mendekati 650 orang.
Rata-rata pengeluaran dan pendapatannya rendah dari semua cluster.	Rata-rata pengeluaran dan pendapatannya paling tinggi.	Pengeluaran dan pendapatannya menengah.
Latar belakang pendidikannya cenderung menyebar, namun semua berdaya beli rendah.	Latar belakang pendidikannya menyebar di semua tingkatan dengan daya beli paling tinggi.	Latar belakang pendidikannya menyebar di semua tingkatan dengan daya beli menengah.
Pengeluaran seseorang yang merupakan orang tua ataupun bukan sama-sama di bawah $200.	Pengeluaran seseorang yang merupakan orang tua lebih tinggi dibandingkan yang bukan orang tua.	Pengeluaran seseorang yang bukan orang tua lebih rendah dari pada yang sudah menjadi orang tua.
Jumlah anak pada cluster ini ada dalam rentang 0-3 dengan pengeluaran kurang dari $200.	Pengeluaran semakin meningkat seiring bertambahnya jumlah anak.	Pengeluaran semakin meningkat seiring bertambahnya jumlah anak.
Usia menyebar di semua rentang.	Usia menyebar di semua rentang namun berdaya beli tinggi.	Di dominasi kelompok usia produktif dengan pengeluaran menengah.
Cluster 0 menghabiskan paling banyak untuk wine, meskipun jumlahnya tetap rendah dibanding kelompok lain.	Cluster 1 memiliki pengeluaran terbanyak untuk semua kategori pembelian kebutuhan	Cluster 2 paling banyak menghabiskan uang untuk *wine* (≈430.1) dan paling sedikit untuk

	pokok.	buah-buahan (≈29.0)
Cluster 0 memiliki frekuensi belanja terendah di semua kelompok. Mereka jarang berbelanja, baik online, katalog, maupun di toko fisik.	Cluster 1 paling sering melakukan pembelian di toko fisik (≈8.4) dan katalog (≈6.0). Mereka jarang belanja lewat web.	Cluster 2 paling menonjol dalam pembelian melalui web (≈5.6) dan juga cukup tinggi di toko fisik (≈7.3).
Strategi promosi yang cocok adalah dengan menawarkan diskon, paket hemat, atau promo beli 1 gratis 1.	Karena masyarakatnya *fancy*, strategi yang efektif adalah dengan menciptakan paket VIP yang unik, mewah, dan eksklusif.	Untuk promosi ke kelompok menengah, fokus pada nilai dan kualitas produk serta penggunaan media yang mereka sukai termasuk promosi yang relevan dengan gaya hidup mereka.

Hasil segmentasi dengan menggunakan *agglomerative clustering* tanpa PCA menghasilkan insight yang sangat jelas pembagian antar clusternya. Setiap cluster memiliki karakteristik tersendiri sehingga membutuhkan pendekatan promosi yang berbeda. Cluster 0 didominasi oleh kelompok masyarakat yang cenderung hemat dan berdaya beli rendah, sehingga strategi promosi yang cocok adalah dengan memberikan diskon untuk setiap barang, adanya paket hemat, dan promo beli 1 gratis 1, sehingga meningkatkan daya beli pada masyarakat ini. Cluster 1 merupakan kalangan masyarakat dengan daya beli tinggi, sehingga promosi yang cocok adalah dengan memberikan paket eksklusif, barang-barang *limited,* dan paket VIP. Sementara itu, untuk cluster 2 strategi yang cocok adalah dengan memfokuskan pada hal-hal yang mereka sukai, termasuk meningkatkan kualitas produk, dan kegunaan produk.

- ***KMeans Clustering***

Sama seperti *Agglomerative Clustering***,** jumlah cluster yang dihasilkan *K-Means Clustering* juga sejumlah 3 cluster yang masing-masing clusternya memiliki karakteristik yang beragam. Berikut adalah insight yang bisa di dapat dari hasil segmentasi ini.

Cluster 0	**Cluster 1**	**Cluster 2**
Cluster dengan jumlah paling sedikit dengan jumlah dibawah 600 orang.	Cluster dengan jumlah terbanyak yakni lebih dari 1.000 orang.	Cluster dengan jumlah sekitar 600 orang.

Rata-rata pendapatan dan pengeluaran menengah.	Rata-rata pendapatan dan pengeluaran menengah.	Rata-rata pendapatan dan pengeluaran menengah.
Latar pendidikan tersebar di semua bagian dimana seseorang dengan tingkat pendidikan *undergraduate* memiliki rata-rata pengeluaran paling kecil dari semua cluster.	Latar pendidikan tersebar di semua bagian dimana seseorang dengan tingkat pendidikan lebih tinggi memiliki rata-rata pengeluaran paling banyak.	Latar pendidikan tersebar di semua bagian dengan seseorang dengan pendidikan *postgraduate* memiliki rata-rata pengeluaran paling banyak.
Pengeluaran seseorang yang bukan orang tua lebih tinggi dari pada yang sudah menjadi orang tua.	Pengeluaran seseorang yang bukan orang tua lebih tinggi dari pada yang sudah menjadi orang tua.	Pengeluaran seseorang yang bukan orang tua lebih tinggi dari pada yang sudah menjadi orang tua.
Pengeluaran semakin menurun tajam seiring bertambahnya jumlah anak.	Pengeluaran cenderung naik turun seiring bertambahnya jumlah anak.	Memiliki jumlah anak sebanyak 0-3 dengan seseorang dengan anak 3 memiliki daya beli paling yang rendah.
Usia tersebar di berbagai rentang.	Usia tersebar di berbagai rentang.	Usia tersebar di berbagai rentang.
Paling banyak menghabiskan uang untuk daging (≈151.4) dan wine (≈283), dengan pengeluaran terendah untuk buah-buahan dan manisan dibanding cluster lain.	Pengeluarannya mirip dengan Cluster 0, tapi sedikit lebih tinggi di hampir semua kategori, terutama wine (≈319) dan yang paling sedikit adalah buah-buahan.	Pengeluaran tertinggi ada pada daging (≈182) dan wine (≈302), serta nilai perhiasan paling tinggi (≈47.3).
Lebih sering membeli barang pada toko fisik dan web dari pada katalog.	Lebih sering membeli barang pada toko fisik dan web dari pada katalog.	Lebih sering membeli barang pada toko fisik dan web dari pada katalog.
Cluster ini cocok diberi promosi penawaran produk dengan kualitas menengah dan promosi berbasis kebutuhan rutin, khususnya untuk kelompok dewasa produktif.	Promosi yang ideal untuk cluster ini adalah paket diskon untuk keluarga atau promosi edukatif tentang efisiensi belanja.	Cluster ini cocok diberi promosi dengan program diskon konstan, bundling hemat, atau program keanggotaan berbasis kebutuhan pokok.

Segmentasi yang dilakukan menggunakan algoritma *K-Means Clustering* menghasilkan empat cluster dengan jumlah anggota yang relatif seimbang. Cluster 2 merupakan kelompok dengan tingkat pengeluaran tertinggi dibandingkan kelompok lainnya, sehingga cocok dengan layanan premium, dan produk eksklusif. Sementara, Cluster 0 menunjukkan perilaku belanja

yang lebih hemat, sehingga lebih cocok diberikan promosi diskon, ataupun paket bundling. Sementara itu, Cluster 1 memiliki karakteristik tingkat pendapatan dan pengeluaran menengah, sehingga promosi yang paling sesuai adalah penawaran produk dengan kualitas baik namun tetap terjangkau, serta produk yang seimbang antara nilai dan fungsionalitas. Dengan demikian, hasil segmentasi ini dapat menjadi acuan awal dalam merancang strategi pemasaran yang lebih tepat sasaran berdasarkan perilaku belanja konsumen.

- ***K-Means Clustering* tanpa PCA**

  Sama seperti *agglomerative clustering,* segmentasi dengan K-Means tanpa PCA pun juga dilakukan. Hal ini digunakan untuk melihat apakah penggunaan PCA mempengaruhi hasil segmentasi secara signifikan atau tidak.

Cluster 0	Cluster 1	Cluster 2
Cluster dengan jumlah paling sedikit dengan jumlah dibawah 600 orang.	Cluster dengan jumlah terbanyak yakni lebih dari 1.000 orang.	Cluster dengan jumlah sekitar 600 orang.
Rata-rata pengeluaran dan pendapatannya paling tinggi.	Rata-rata pengeluaran dan pendapatannya paling rendah diantara semua cluster.	Rata-rata pengeluaran dan pendapatannya menengah.
Latar pendidikan tersebar di semua bagian dimana seseorang dengan tingkat pendidikan lebih tinggi memiliki rata-rata pengeluaran paling banyak.	Di semua tingkat pendidikan memiliki daya beli yang paling rendah dibandingkan dengan cluster lain.	Pengeluaran cenderung menengah dengan pengeluaran yang stabil untuk semua tingkatan pendidikan.
Baik bukan orang tua atau merupakan orang tua memiliki pengeluaran yang tinggi.	Pengeluaran yang bukan orang tua dan orang tua sama-sama di bawah $200.	Pengeluaran seseorang yang bukan orang tua lebih rendah dari pada yang sudah menjadi orang tua.
Cluster ini memiliki anak 0-1 dan sama-sama memiliki pengeluaran tinggi.	Cluster ini memiliki jumlah anak 0-3 dengan pengeluaran yang sangat rendah.	Kelompok yang memiliki anak 1 memiliki daya beli paling tinggi di antara jumlah anak yang lain.
Cluster 1 memiliki sebaran usia yang luas dan tidak terfokus pada rentang tertentu namun dengan pengeluaran	Cluster 1 memiliki sebaran usia yang luas dan tidak terfokus pada rentang tertentu.	Rentang usia pada cluster ini ada di 31-75 Tahun.

tinggi.		
Paling banyak menghabiskan uang untuk daging (≈459.1) dan wine (≈600.3).	Rata-rata pembelian bahan-bahan pokok sangat rendah dan yang tertinggi adalah wine (≈46.4).	Paling banyak menghabiskan uang untuk wine (≈492.2), dengan pengeluaran terendah untuk buah-buahan dan manisan.
Lebih sering membeli barang pada toko fisik dan katalog dari pada web.	Pembelian di semua tempat pembelian sangat rendah.	Lebih sering membeli barang pada toko fisik dan web dari pada katalog.
Strategi promosi untuk masyarakat kelas atas berfokus pada kualitas, eksklusivitas, dan pengalaman.	Strategi promosi yang cocok adalah dengan menawarkan paket hemat, atau promo beli 1 gratis 1, serta diskon untuk semua bahan-bahan pokok.	Strategi promosi untuk cluster ini memberikan nilai yang seimbang antara kualitas dan harga.

Hasil segmentasi *K-Means Clustering* tanpa PCA menghasilkan insight yang lebih tertata jika dibandingkan dengan penggunaan PCA. Dari hasil ini dapat ditarik kesimpulan bahwa penggunaan PCA memberikan pengaruh terhadap hasil segmentasi dengan *K-Means Clustering.* Cluster 0 merupakan masyarakat kelas atas yang cocok diberikan barang-barang bermerek dengan kualitas terbaik, dan paket VIP. Sementara itu, cluster 1 merupakan cluster dengan masyarakat yang hemat dan cocok diberi promosi paket hemat, dan diskon untuk semua bahan-bahan yang diperlukan. Sedangkan, untuk cluster 2 cocok diberikan promosi barang-barang yang sesuai dengan kebutuhan mereka, dan seimbang antara kualitas dengan harga.

- **Perbandingan Hasil dari *Agglomerative Clustering* dan *K-Means Clustering***

Aspek	***Agglomerative Clustering***	***K-Means Clustering***
Jumlah cluster	Sebanyak 3 cluster	Sebanyak 3 cluster
Banyak PC	PC1 sampai PC2	PC1 sampai PC2
Inisialisasi	Tidak membutuhkan inisialisasi	Membutuhkan inisialisasi awal centroid
Durasi	0.4405 detik	0.4299 detik

runtime		
Silhouette Score	0.465428915251249	0.49717267008432164
Calinski-Harabasz Index	2827.2563909536175	3179.8098781803737
Davies-Bouldin Index	0.7755982297672879	0.7333450545790788
Posisi *cluster* 0	Tengah-atas (PC1 sedang, PC2 tinggi)	Kanan bawah (PC1 tinggi, PC2 ≈ 0)
Posisi *cluster* 1	Bawah tengah (PC1 -3 s.d. 0, PC2 negatif)	Kiri atas (PC1 negatif, PC2 positif)
Posisi *cluster* 2	Kanan bawah (PC1 > 2, PC2 < 0)	Tengah atas (PC1 0–3, PC2 tinggi)

Kesimpulannya:

Perbandingan antara algoritma *Agglomerative Clustering* dengan algoritma *K-Means Clustering* di atas menunjukkan tidak ada perbedaan yang signifikan. Dengan jumlah *cluster* dan PC yang sama, didapatkan:

- *Agglomerative Clustering* memiliki durasi *runtime* sedikit lebih cepat dari *K-Means Clustering*
- *Silhouette Score* dari *K-Means Clustering* sedikit lebih besar yang berarti sedikit lebih baik *Agglomerative Clustering*
- *Calinski-Harabasz Index* dari *K-Means Clustering* jauh lebih besar yang berarti jauh lebih baik *Agglomerative Clustering*
- *Davies-Bouldin Index* dari *Agglomerative Clustering* lebih rendah yang berarti lebih baik *K-Means Clustering*

- **Perbandingan Hasil dari *Agglomerative Clustering* PCA dan Non-PCA**

Aspek	PCA	Non-PCA
Jumlah Cluster	Sebanyak 3 cluster	Sebanyak 3 cluster
Silhouette Score	0.465428915251249	0.17121259805155878
Calinski-Harabasz Index	2827.2563909536175	554.8762963365112
Davies-Bouldin Index	0.7755982297672879	1.9613773139843829

Kesimpulannya:

Dari hasil di atas dapat dilihat bahwa dengan menggunakan PCA maupun non-PCA, algoritma *Agglomerative Clustering* mampu memisahkan 3 cluster dengan baik. Jika dilihat dari hasil visualisasinya, penggunaan PCA jauh lebih

mudah untuk diinterpretasikan karena lebih jelas dan terstruktur. Dan jika dilihat dari hasil metrik evaluasinya, penggunaan PCA juga jauh lebih unggul. Nilai *Silhouette Score* dan *Calinski-Harabasz Index* yang lebih tinggi serta *Davies-Bouldin Index* yang lebih rendah pada clustering dengan PCA menunjukkan bahwa hasil segmentasi lebih bagus dan terpisah dengan baik antar cluster.

- **Perbandingan Hasil *K-Means Clustering* dengan PCA dan Non-PCA**

Aspek	PCA	Non-PCA
Jumlah Cluster	Sebanyak 3 cluster	Sebanyak 3 cluster
Durasi runtime	0.4149 detik	0.1231 detik
Silhouette Score	0.49717267008432164	0.20932021366754708
Calinski-Harabasz Index	3179.8098781803737	644.5026600135105
Davies-Bouldin Index	0.7333450545790788	1.7693273274270638

Kesimpulannya:

Dari hasil di atas dapat dilihat bahwa dengan menggunakan PCA maupun non-PCA, algoritma *K-Means* mampu memisahkan 3 cluster dengan baik. Jika dilihat dari hasil metrik evaluasi di atas, penggunaan PCA lebih baik dari pada yang Non-PCA, dimana nilai *Silhouette Score* dan *Calinski-Harabasz Index* lebih tinggi, serta *Davies-Bouldin Index* lebih rendah. Namun, jika dilihat dari hasil visualisasinya, segmentasi tanpa PCA menghasilkan visualisasi yang lebih baik dan lebih mudah untuk diinterpretasikan. Dimana antar cluster sudah terpisah dan tidak ada yang tumpang tindih.

- **Dokumentasi Percobaan Algoritma Lainnya**

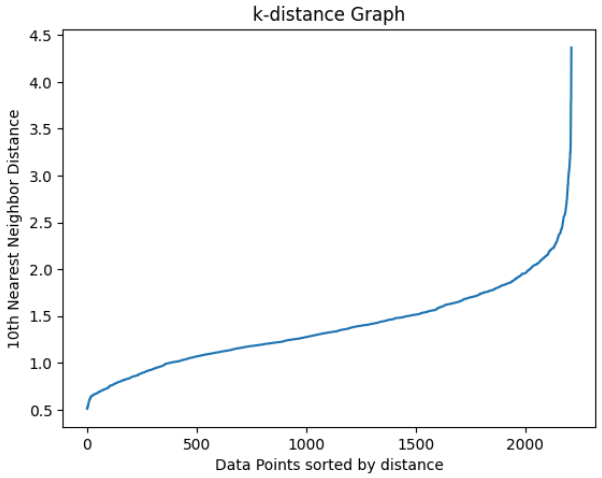
1) **DBSCAN (dengan PCA)**

Algoritma DBSCAN (Density-Based Spatial Clustering of Applications with Noise) adalah algoritma klasterisasi yang digunakan untuk mengelompokkan data berdasarkan kepadatan (density) titik-titik data. Kelebihan DBSCAN adalah bisa menemukan klaster dengan bentuk yang tidak beraturan, tidak perlu menentukan jumlah klaster di awal seperti halnya K-Means, dan juga mampu menangani outlier/titik noise. DBSCAN menggunakan dua parameter yaitu:

a. ε (epsilon), yaitu jarak maksimum antara dua titik untuk dianggap sebagai "tetangga".

b. MinPts (Minimum Points) atau *min_samples* pada *scikit-learn*, yaitu jumlah minimum tetangga (termasuk titik itu sendiri) agar titik dianggap sebagai bagian dari *cluster*.

Langkah pertama adalah menentukan nilai epsilon, umumnya adalah dengan menggunakan *k-distance graph* seperti ini:



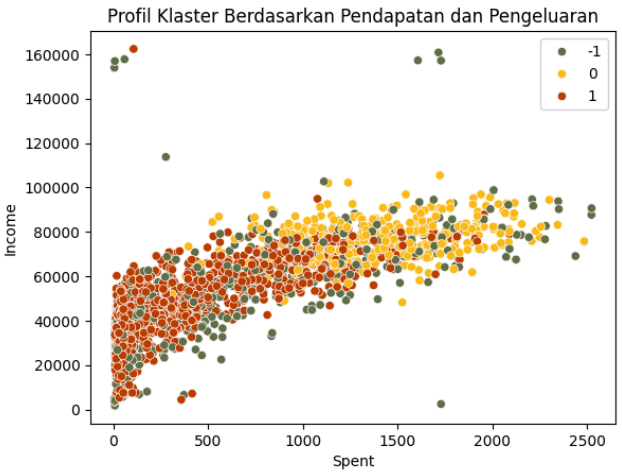
Gambar 45. Visualisasi k-distance Graph

k-distance graph di atas menggunakan jumlah tetangga sebanyak 10 dan terlihat bahwa ada *knee point* (titik siku) di sekitar nilai $\varepsilon \approx 2.0$ sampai 2.5 (di sumbu Y).

Setelah dilakukan beberapa kali percobaan akhirnya percobaan dihentikan pada nilai epsilon sebesar 1.76 dan minimum points sebesar 30 dengan hasil sebagai berikut:

```
Silhouette Score          : 0.4077693077669633
Davies-Bouldin Index      : 0.9027217605362177
Calinski-Harabasz Index   : 1228.5742369429756
```

Dari hasil di atas, terlihat bahwa DBSCAN dengan epsilon sebesar 1.76 dan minimum points sebesar 30 berhasil membentuk klaster yang baik. Karena ini hanyalah algoritma tambahan, maka untuk evaluasi diambil evaluasi *cluster* berdasarkan pendapatan dan pengeluaran saja seperti ini:

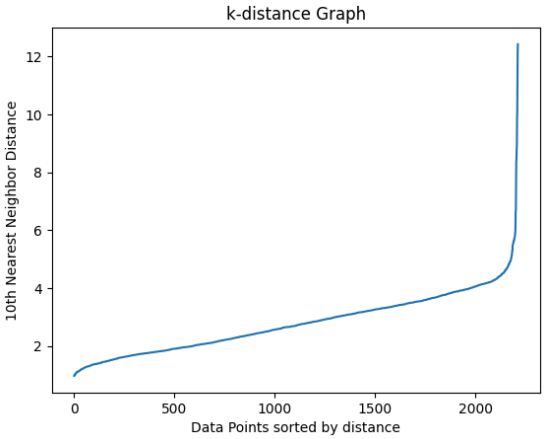


Gambar 46. Visualisasi Profil Cluster berdasarkan Pendapatan dan Pengeluaran

Dari visualisasi di atas, didapat bahwa DBSCAN berhasil mengelompokkan konsumen ke dalam 2 klaster utama berdasarkan perilaku pengeluaran dan pendapatan. Selain itu, algoritma juga mengidentifikasi sejumlah outlier (label -1) yang memiliki perilaku ekstrem (baik sangat rendah maupun sangat tinggi).

2) DBSCAN (tanpa PCA)

Sama seperti sebelumnya, langkah pertama adalah menentukan nilai epsilon, umumnya dengan menggunakan *k-distance graph*, dan hasilnya seperti ini:



Gambar 47. Visualisasi k-distance Graph

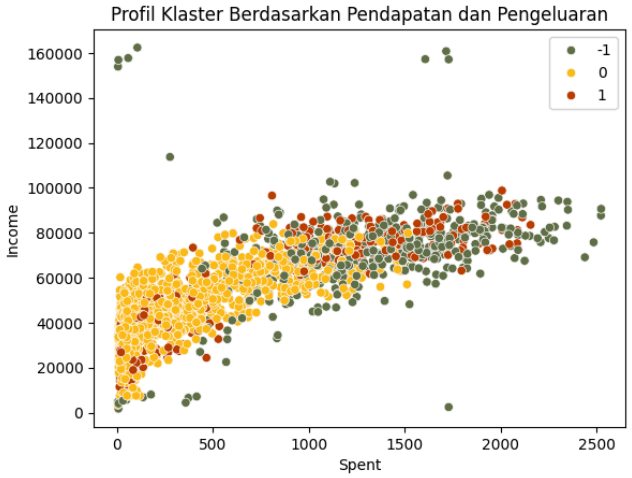
k-distance graph di atas menggunakan jumlah tetangga yang sama dengan sebelumnya yaitu sebanyak 10 dan terlihat bahwa ada *knee point* (titik siku) di sekitar nilai $\varepsilon \approx 4.0$ sampai 5.0 (di sumbu Y).

Setelah dilakukan beberapa kali percobaan akhirnya percobaan dihentikan pada nilai epsilon sebesar 3 dan minimum points sebesar 6 dengan hasil sebagai berikut:

```
Silhouette Score           : 0.2475118570185295
Davies-Bouldin Index       : 1.7410709669620565
Calinski-Harabasz Index    : 345.74148729936775
```

Hasil di atas menunjukkan bahwa DBSCAN bekerja lebih baik ketika menggunakan PCA. Data berdimensi tinggi cenderung membuat jarak antar titik jadi seragam, sehingga algoritma berbasis jarak seperti DBSCAN menjadi kurang efektif.

Kemudian karena ini hanyalah algoritma tambahan, maka untuk evaluasi diambil evaluasi *cluster* berdasarkan pendapatan dan pengeluaran juga seperti ini:



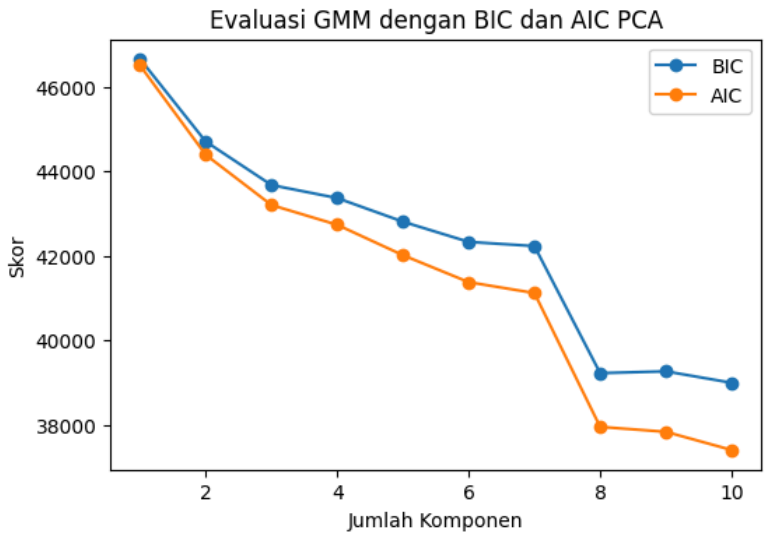
Gambar 48. Visualisasi Profil Cluster berdasarkan Pendapatan dan Pengeluaran

Dari visualisasi di atas, didapat bahwa klasterisasi kurang optimal jika dibandingkan dengan DBSCAN (dengan PCA) sebelumnya. Banyak data yang diklasifikasikan sebagai outlier (label -1) dan antar *cluster* yang saling tumpang tindih.

3) Gaussian Mixture Model (dengan PCA)

Gaussian Mixture Model (GMM) adalah algoritma klasterisasi berbasis probabilistik yang mengasumsikan data berasal dari gabungan beberapa distribusi Gaussian. Berbeda dengan K-Means yang membagi data secara tegas, GMM memberikan probabilitas keanggotaan setiap titik data ke masing-masing klaster.

Karena GMM membutuhkan jumlah klaster yang ditentukan di awal, maka dalam percobaan ini digunakan metode evaluasi model dengan Akaike Information Criterion (AIC) dan Bayesian Information Criterion (BIC) untuk memilih jumlah klaster terbaik secara otomatis. Namun karena sudah ditentukan untuk semua metode digunakan 3 klaster, maka hasil ini hanya sebagai evaluasi.

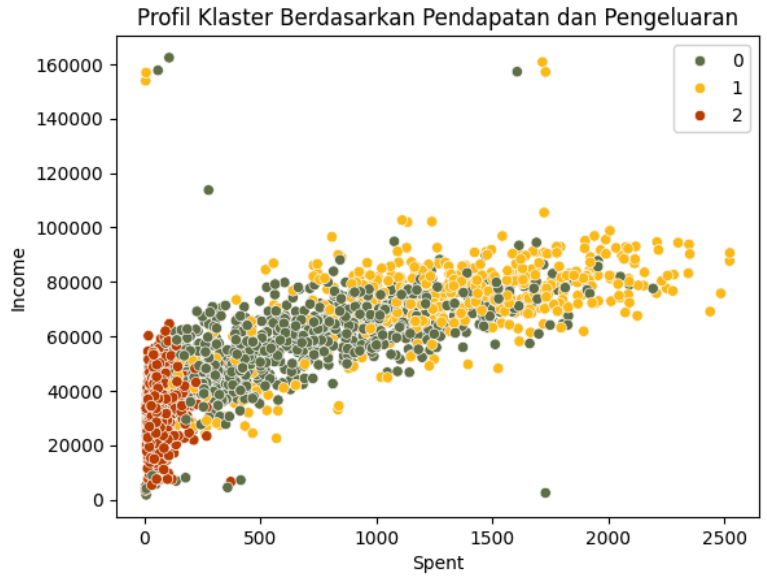


Gambar 49. Penentuan klaster berdasarkan BIC AIC

Berikut penilaian kualitas 3 klaster yang dihasilkan GMM:

```
Jumlah Klaster: 3
Silhouette Score        : 0.2616
Davies-Bouldin Index    : 1.3781
Calinski-Harabasz Score: 1055.18
```

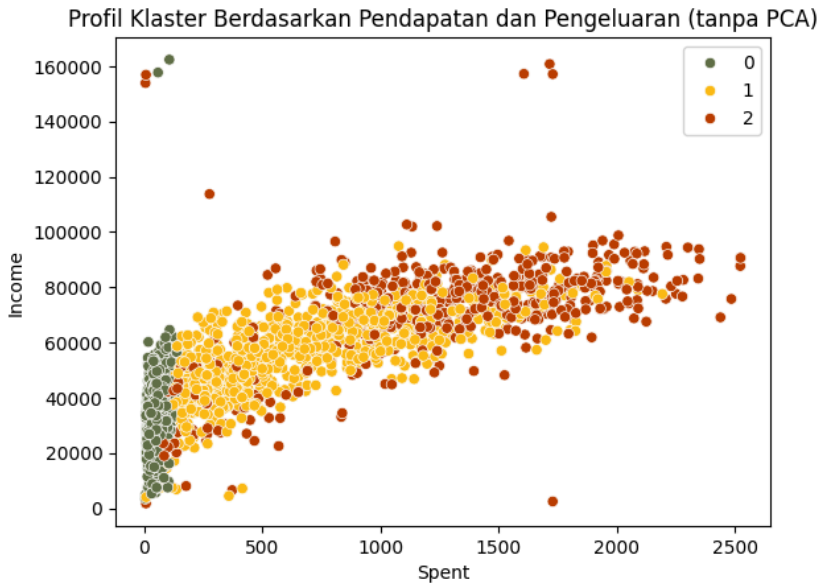
Grafik *profiling* pendapatan dan pengeluarannya sebagai berikut:



Gambar 50. Visualisasi Profil Cluster berdasarkan Pendapatan dan Pengeluaran

4) Gaussian Mixture Model (tanpa PCA)

```
Jumlah Klaster: 3
Silhouette Score        : 0.2616
Davies-Bouldin Index    : 1.3781
Calinski-Harabasz Score: 1055.18
```



Gambar 51. Visualisasi Profil Cluster berdasarkan Pendapatan dan Pengeluaran

Kesimpulan

Berdasarkan hasil segmentasi dan evaluasi, *Agglomerative Clustering* maupun *K-Means Clustering* mampu membentuk cluster konsumen yang dapat dijadikan dasar strategi pemasaran yang tepat sasaran. Masing-masing cluster menunjukkan karakteristik ekonomi dan demografi yang berbeda, sehingga memerlukan pendekatan promosi yang disesuaikan, mulai dari layanan eksklusif untuk konsumen dengan daya beli tinggi hingga penawaran diskon untuk kelompok yang lebih hemat. Meskipun hasil segmentasi keduanya mirip, evaluasi metrik menunjukkan keunggulan masing-masing dimana *K-Means* unggul pada *Silhouette Score* dan *Calinski-Harabasz Index,* sedangkan *Agglomerative Clustering* lebih baik dalam kecepatan runtime dan *Davies-Bouldin Index.*

Untuk algoritma DBSCAN dengan dan tanpa PCA, dapat disimpulkan bahwa DBSCAN dengan PCA menghasilkan klasterisasi yang jauh lebih baik dan terstruktur. Hal ini dibuktikan dengan matrik evaluasi yang lebih baik pada DBSCAN dengan PCA daripada metrik evaluasi pada DBSCAN tanpa PCA. Visualisasi hasil klasterisasi juga memperkuat hal ini, di mana hasil dengan PCA menunjukkan distribusi klaster yang lebih bersih dan outlier yang lebih terkendali, sedangkan tanpa PCA, klaster tampak tumpang tindih dan banyak data menjadi outlier. Oleh karena itu, penggunaan PCA sebelum DBSCAN sangat direkomendasikan terutama pada data berdimensi tinggi.

GMM menunjukkan evaluasi metrik yang identik baik dengan maupun tanpa PCA, menunjukkan kualitas segmentasi yang cukup baik namun tidak signifikan meningkat

oleh PCA. Visualisasi menunjukkan distribusi klaster yang serupa, meski tanpa PCA klaster tampak lebih tersebar. GMM mampu menangkap variasi data dengan pendekatan probabilistik, namun manfaat PCA dalam konteks ini relatif minimal.

3. Daftar Pustaka

Sitasi disusun dan ditulis berdasarkan sistem nomor sesuai dengan urutan pengutipan, mengikuti format APA. Hanya pustaka yang disitasi pada usulan penelitian yang dicantumkan dalam Daftar Pustaka. Pustaka yang disitasi maksimal 8 tahun terakhir sebanyak minimal 10 pustaka.

[1] Isabel, M. (2020). *A Clustering Approach to Market Segmentation Using Integrated Business Data*. http://erepository.uonbi.ac.ke/handle/11295/155810

[2] Asha Panyako Makana, B., & Evans K Miriti, S. A. (2020). *Customer Segmentation On Mobile Money Users In Kenya*. http://erepository.uonbi.ac.ke/handle/11295/152965

[3] Patankar, N., Dixit, S., Bhamare, A., Darpel, A., & Raina, R. (2021). Customer Segmentation Using Machine Learning. *Advances in Parallel Computing*, *39*, 239–244. https://doi.org/10.3233/APC210200

[4] Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer Segmentation using K-means Clustering. *Proceedings of the International Conference on Computational Techniques, Electronics and Mechanical Systems, CTEMS 2018*, 135–139. https://doi.org/10.1109/CTEMS.2018.8769171

[5] Yazar, S., & Author, C. (2023). Customer Segmentation Using K-Means Clustering Algorithm and RFM Model. *DEÜ FMD*, *25*(74), 491–503. https://doi.org/10.21205/deufmd.2023257418

[6] Deng, Y., & Gao, Q. (2020). A study on e-commerce customer segmentation management based on improved K-means algorithm. *Information Systems and E-Business Management*, *18*(4), 497–510. https://doi.org/10.1007/S10257-018-0381-3

[7] Shiradwade, P., Munnole, S., & Birje, M. N. (2023). CUSTOMER SEGMENTATION WITH K-MEANS CLUSTERING. *Www.Irjmets.Com @International Research Journal of Modernization in Engineering*, *661*. https://doi.org/10.56726/IRJMETS44609

[8] Omol, E., Onyangor, D., Mburu, L., & Abuonji, P. (2024). Application Of K-Means Clustering For Customer Segmentation In Grocery Stores In Kenya. *International*

Journal of Science, Technology & Management, *5*(1), 192–200. https://doi.org/10.46729/IJSTM.V5I1.1024

[9] Shihab, S. H., Afroge, S., & Mishu, S. Z. (2019). RFM Based Market Segmentation Approach Using Advanced K-means and Agglomerative Clustering: A Comparative Study. *European Conference on Cognitive Ergonomics*. https://doi.org/10.1109/ECACE.2019.8679376

[10] Hung, P. D., Thuy Lien, N. T., & Ngoc, N. D. (2019). Customer Segmentation Using Hierarchical Agglomerative Clustering. *Proceedings of the 2nd International Conference on Information Science and Systems*, *Part F148384*, 33–37. https://doi.org/10.1145/3322645.3322677

4. Kontribusi

Nama	Kontribusi
Fadhil Muhamad (23031554090)	• Mencari Dataset • Eksplorasi Dataset • Interpretasi Korelasi Fitur • Implementasi Model • Evaluasi Model • *K-Means Clustering* • GMM *Clustering* PCA dan Non-PCA • Poster
Sintiya Risla Miftaqul N. (23031554204)	• Mencari Dataset • Mengerjakan Latar Belakang • Tujuan Penelitian • Rumusan Masalah • Pendekatan Penyelesaian Masalah • Langkah Penelitian • Mencari Jurnal • Sitasi • Membuat PPT • Penanganan Missing Value • Penanganan Ketidakkonsistenan Data • Code Python • *Pre-processing* Data • Reduksi Dimensi • Evaluasi Model • Insight dan Hasil Mining dari Project *Agglomerative Clustering* dan *K-Means Clustering*

	• Perbandingan *Agglomerative Clustering* PCA dan Non-PCA • Perbandingan *K-Means* PCA dan Non-PCA • Poster
Dimas Fatkhul Rahman (23031554211)	• Mengerjakan Eksplorasi Dataset • Code Python • Membuat PPT • Merapikan PPT • Feature Engineering • Penanganan Outlier • Implementasi Model • *Agglomerative Clustering* PCA dan Non-PCA • Perbandingan Hasil dari *Agglomerative Clustering* dan *K-Means Clustering* • *DBSCAN clustering* PCA dan Non-PCA • Poster

5. Kendala

1. Kendala yang dialami adalah ketidakkonsistenan data pada kolom *Dt_Customer* yang awalnya tidak disadari sehingga saat mengerjakan bagian selanjutnya mengalami error yang membuat sedikit kebingungan.
2. Kesulitan dalam memahami pairplot dari hasil PCA sehingga membutuhkan waktu tambahan untuk menginterpretasikannya.
3. Kesulitan dalam menginterpretasikan hasil mining dari algoritma *K-Means Clustering*
4. Menentukan nilai parameter terbaik untuk DBSCAN