Mata Kuliah	:	Data Science	**NIM**	: 20230040175
Sesi Pertemuan	:	9 (Sembilan)	**Nama Mahasiswa**	: Regal Nugraha
Dosen	:	Alun Sujjada, S.Kom, M.T.	**Kelas**	: TI23C

Studi Kasus: Agglomerative Clustering pada Customer Segmentation Dataset

I. Pre-processing

1. Missing Values

Dilakukan pengecekan nilai yang hilang pada seluruh kolom DataFrame, dan ditemukan bahwa tidak ada nilai yang hilang.

```
print(df.isnull().sum())
```

```
CustomerID          0
Age                 0
AnnualIncome        0
SpendingScore       0
VisitsPerMonth      0
TimeOnApp(min)      0
PurchaseFrequency   0
dtype: int64
```

2. Standarisasi Fitur Numerik

Fitur-fitur numerik (selain CustomerID) distandardisasi menggunakan StandardScaler. Langkah ini penting untuk memastikan bahwa semua fitur memiliki skala yang sama, sehingga tidak ada fitur yang mendominasi proses clustering hanya karena rentang nilainya yang lebih besar.

```
Scaled DataFrame head:
        Age  AnnualIncome  SpendingScore  VisitsPerMonth  TimeOnApp(min)  \
0  1.086153     -1.691128       1.213730        1.307080        1.470958
1  0.367798      1.467747       0.333507       -1.184178       -1.338267
2 -0.637899     -0.510055       1.474536       -1.350262       -1.398037
3  1.373495     -1.624592      -1.655145        1.140996        0.454855
4 -1.140748      0.511439      -0.807523        0.476661        1.620384

   PurchaseFrequency
0          -1.544914
1          -0.796010
2           1.501355
3          -1.262718
4          -1.809021
```

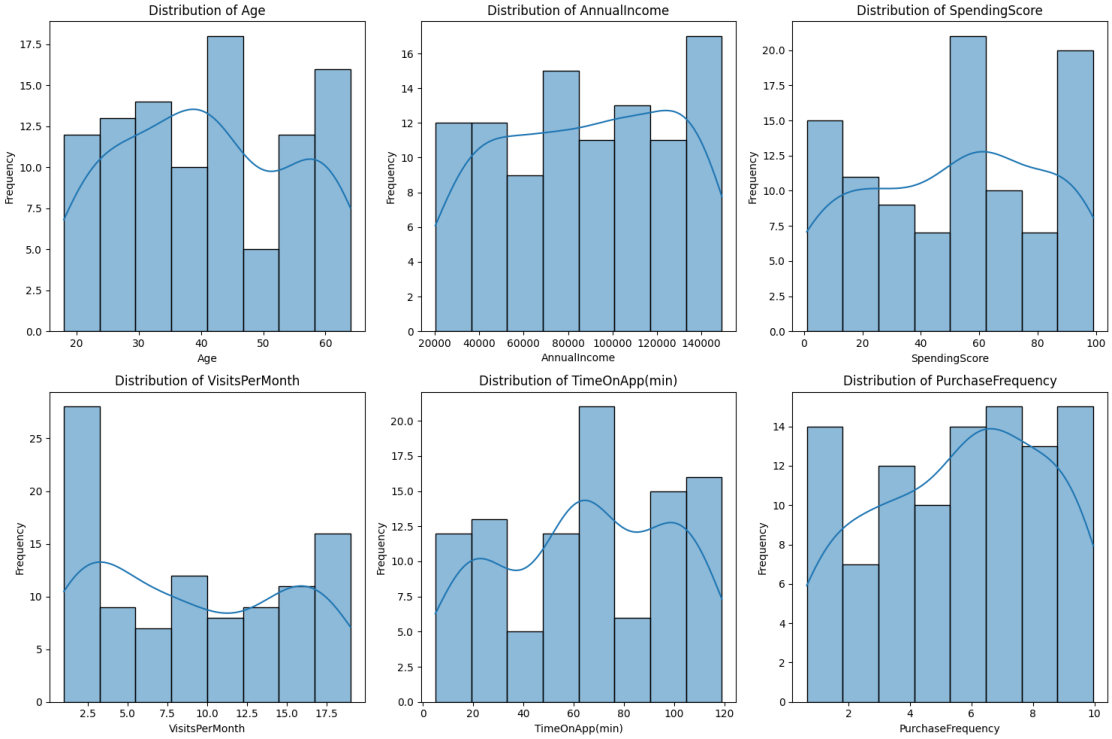
II. Analisis Deskriptif dan Visualisasi Awal

Untuk mendapatkan gambaran tentang central tendency, dispersion, dan shape distribusi dari setiap fitur numerik. Ini termasuk nilai rata-rata (mean), standar deviasi (std), nilai minimum (min), nilai maksimum (max), dan kuartil (25%, 50%, 75%).

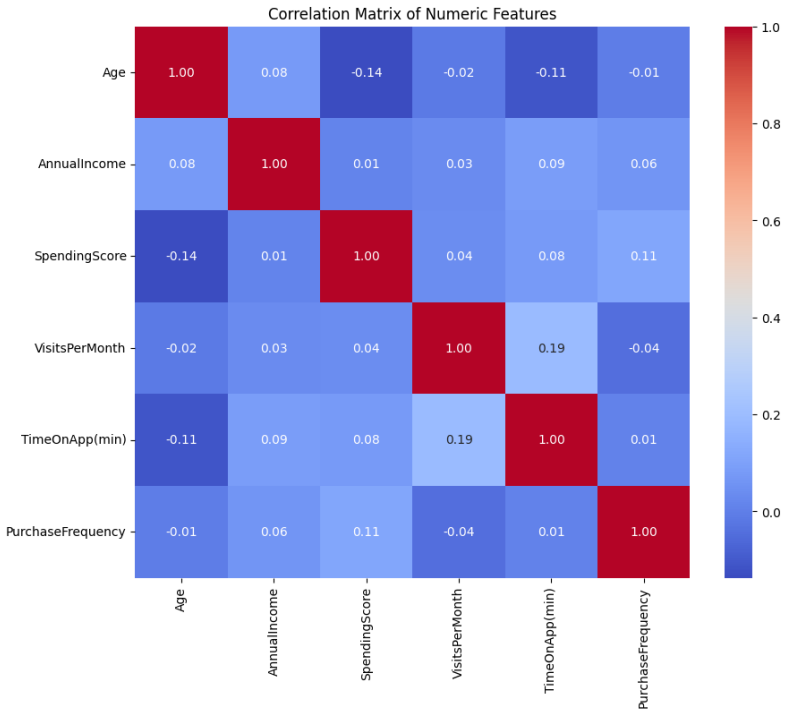
```
Summary statistics for the DataFrame:
       CustomerID         Age   AnnualIncome  SpendingScore  VisitsPerMonth  \
count  100.000000  100.000000     100.000000     100.000000      100.000000
mean    50.500000   40.880000   87837.340000      51.770000        9.130000
std     29.011492   13.990082   38714.098363      30.828576        6.051388
min      1.000000   18.000000   20206.000000       1.000000        1.000000
25%     25.750000   30.500000   53660.000000      24.000000        3.000000
50%     50.500000   41.000000   89877.500000      55.500000        8.000000
75%     75.250000   53.250000  124835.500000      77.500000       15.250000
max    100.000000   64.000000  149312.000000      99.000000       19.000000

       TimeOnApp(min)  PurchaseFrequency
count      100.000000         100.000000
mean        64.780000           5.660200
std         33.629767           2.777962
min          5.000000           0.650000
25%         33.750000           3.427500
50%         65.500000           5.925000
75%         96.000000           8.125000
max        119.000000           9.960000
```

1. Temuan dari hasil analisis deskriptif:
    a. AnnualIncome berkisar antara sekitar $20 ribu hingga $149 ribu dengan rata-rata sekitar $87 ribu.
    b. SpendingScore berkisar dari 1 hingga 99 dengan rata-rata sekitar 51.
    c. Age berkisar dari 18 hingga 64 tahun dengan rata-rata sekitar 40 tahun.
    d. Fitur-fitur lain seperti VisitsPerMonth, TimeOnApp(min), dan PurchaseFrequency juga menunjukkan rentang dan distribusi nilai yang beragam
2. Histogram Distribusi Fitur



    a. Berdasarkan histogram yang dihasilkan:
        - Sebagian besar fitur menunjukkan distribusi yang relatif merata atau sedikit miring, tanpa adanya spike atau gap yang terlalu ekstrem, yang mengindikasikan data cukup tersebar.
        - Misalnya, Age dan AnnualIncome menunjukkan beberapa puncak, sementara SpendingScore dan PurchaseFrequency cenderung lebih merata di seluruh rentang.
3. Matriks Korelasi dan Heatmap



a. Dari matriks korelasi dan heatmap:
    - Secara umum, dataset ini menunjukkan hubungan linier yang relatif lemah antar fitur, dengan sebagian besar koefisien korelasi mendekati nol.
    - Korelasi positif tertinggi (meskipun masih moderat) terlihat antara VisitsPerMonth dan TimeOnApp(min) (sekitar 0.19), yang logis karena lebih banyak kunjungan mungkin berarti lebih banyak waktu di aplikasi.
    - Korelasi negatif dan positif lainnya umumnya sangat rendah, menunjukkan bahwa fitur-fitur ini cukup independen satu sama lain dalam hal hubungan linier.

III. Pemodelan Aglomeratif Clustering

```
Cluster labels for 'single' linkage (first 5):
[1 0 0 0 0]

Cluster labels for 'complete' linkage (first 5):
[0 1 0 0 0]

Cluster labels for 'average' linkage (first 5):
[0 1 1 1 1]

Cluster labels for 'ward' linkage (first 5):
[0 1 0 1 0]
```

Clustering aglomeratif diterapkan menggunakan empat metode linkage: single, complete, average, dan ward. Setelah fitting awal, label klaster pertama (Klaster 0 atau 1) untuk 5 sampel pertama menunjukkan variasi penugasan antar metode:

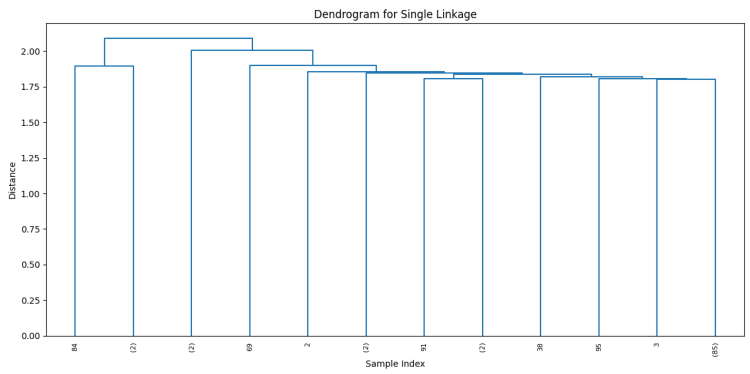
- Single Linkage: [1 0 0 0 0]
- Complete Linkage: [0 1 0 0 0]
- Average Linkage: [0 1 1 1 1]
- Ward Linkage: [0 1 0 1 0]

1. Evaluasi dan Pemilihan Metode Terbaik:
   Setelah menentukan jumlah klaster optimal via dendrogram dan mengevaluasi dengan Silhouette Score (lebih tinggi lebih baik) serta Davies-Bouldin Score (lebih rendah lebih baik), ditemukan:
    a. Single Linkage memiliki Silhouette Score tertinggi (0.1189) namun menghasilkan klaster yang sangat tidak seimbang (mayoritas di satu klaster), menjadikannya kurang praktis.
    b. Complete Linkage dan Average Linkage menunjukkan kinerja yang kurang baik dengan Silhouette Score rendah.
    c. Ward Linkage (Silhouette Score 0.1015) dipilih sebagai metode terbaik karena memberikan keseimbangan antara metrik evaluasi dan, yang terpenting, menghasilkan struktur klaster yang lebih seimbang, jelas, dan mudah diinterpretasikan untuk kebutuhan bisnis (dengan 3 klaster optimal).

   Ward Linkage efektif dalam mengidentifikasi segmen pelanggan yang berbeda dan dapat ditindaklanjuti, memberikan wawasan untuk personalisasi pemasaran, pengembangan produk, dan strategi retensi.

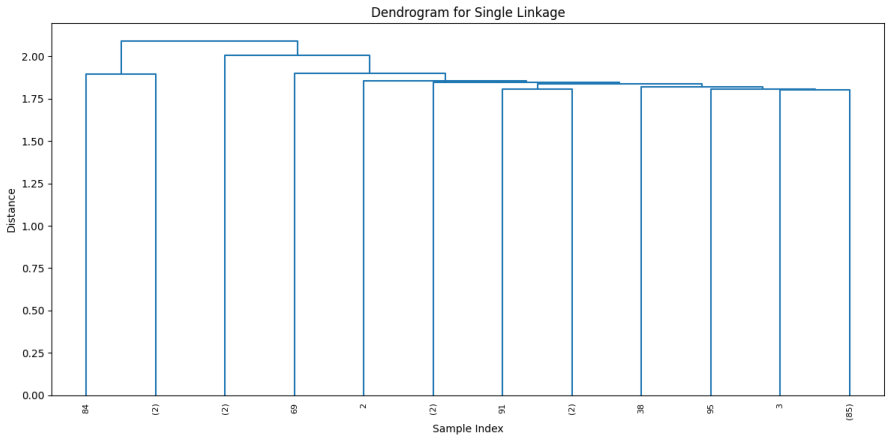
IV. Menentukan Jumlah Klaster dari Dendrogram

1. Dendrogram untuk Single Linkage



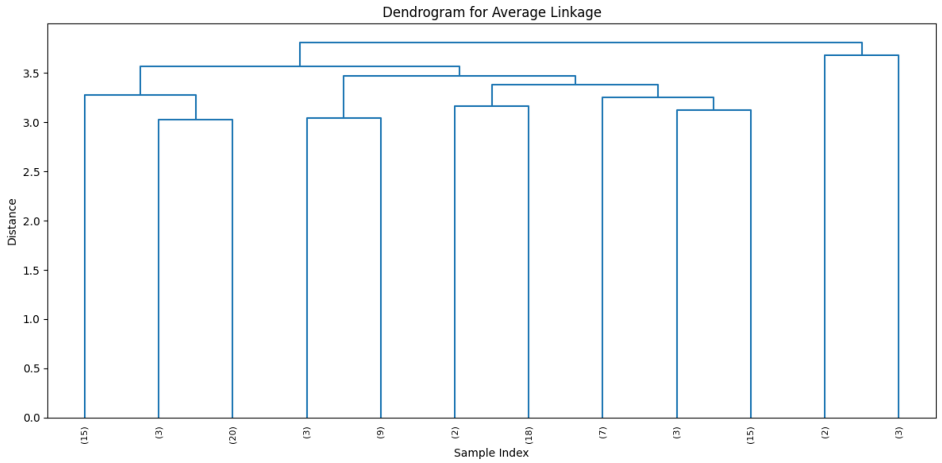
- Jumlah Klaster yang Dipilih: 2
- Interpretasi: Pada dendrogram Single Linkage, kita melihat adanya satu 'loncatan' atau garis vertikal yang sangat panjang dan tinggi, memisahkan satu atau beberapa sampel dari sebagian besar data lainnya. Jika kita menarik garis horizontal di atas loncatan ini, kita akan mendapatkan dua kelompok besar. Sifat single linkage yang rentan terhadap 'chaining' (rantai) seringkali menyebabkan satu klaster besar yang memanjang dan beberapa outlier yang membentuk klaster terpisah, sehingga 2 klaster seringkali menjadi pilihan yang paling stabil.

2. Dendrogram untuk Complete Linkage



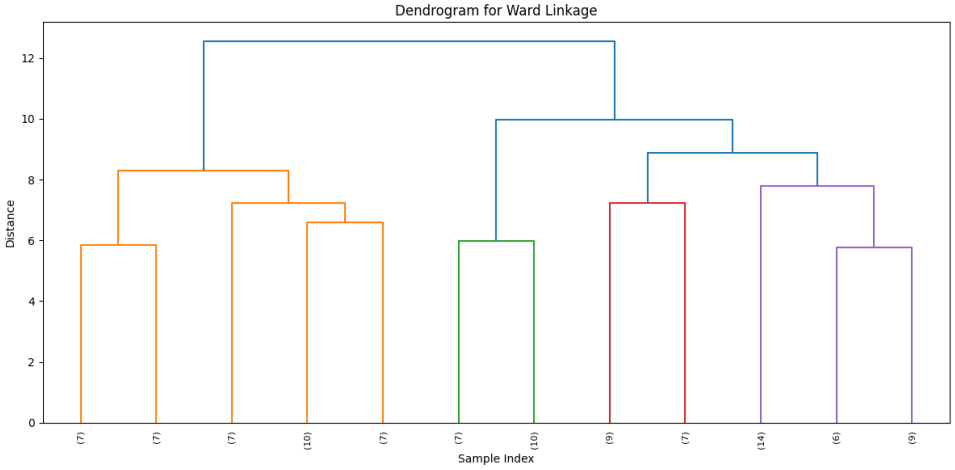
- Jumlah Klaster yang Dipilih: 3
- Interpretasi: Pada dendrogram Complete Linkage, kita dapat melihat tiga kelompok cabang vertikal yang jelas dan terpisah pada tingkat jarak tertentu. Jika kita menarik garis horizontal yang memotong cabang-cabang utama ini tanpa memotong loncatan jarak yang sangat kecil, kita akan mengidentifikasi 3 klaster yang terbentuk. Ini mengindikasikan bahwa data secara alami terpisah menjadi tiga kelompok yang lebih padat dan bulat.

3. Dendrogram untuk Complete Linkage



- Jumlah Klaster yang Dipilih: 3
- Interpretasi: Dendrogram Average Linkage menampilkan struktur yang merupakan kompromi antara single dan complete linkage. Pada jarak tertentu, kita bisa melihat tiga hingga empat cabang utama yang terpisah dengan baik. Pemilihan 3 klaster didasarkan pada keinginan untuk mendapatkan kelompok yang cukup berbeda namun masih cukup besar untuk dianalisis. Garis horizontal yang memotong tiga cabang utama tanpa mengabaikan pemisahan yang signifikan akan menghasilkan 3 klaster.

4. Dendrogram untuk Ward Linkage



- Jumlah Klaster yang Dipilih: 3
- Interpretasi: Dendrogram Ward Linkage biasanya menghasilkan klaster yang paling bulat dan ukurannya paling seimbang. Dengan menarik garis horizontal pada tingkat jarak yang cukup tinggi sebelum terjadi penggabungan mayor, kita dengan jelas dapat mengidentifikasi tiga kelompok cabang utama. Tiga klaster ini menunjukkan pemisahan yang jelas dan minimisasi varians internal dalam setiap klaster, menjadikannya pilihan yang kuat untuk interpretasi praktis. Cabang-cabang vertikal di dendrogram ini memiliki panjang yang lebih seragam dibandingkan single linkage.

Pemilihan jumlah klaster ini didasarkan pada penyeimbangan antara pemisahan visual yang jelas pada dendrogram (mencari 'loncatan' jarak terbesar yang tidak dipotong) dan pertimbangan praktis mengenai interpretasi klaster untuk analisis selanjutnya.

V. Evaluasi Cluster

Untuk mengevaluasi kualitas klaster yang dihasilkan oleh setiap metode *linkage*, kami menggunakan dua metrik utama:

1. Silhouette Score: Mengukur seberapa mirip sebuah objek dengan klaster sendiri (kohesi) dibandingkan dengan klaster lain (separasi). Skor yang lebih tinggi (mendekati +1) menunjukkan bahwa objek cocok dengan klaster sendiri dan tidak cocok dengan klaster tetangga.
2. Davies-Bouldin Score: Mengukur rasio antara penyebaran di dalam klaster dan pemisahan antar klaster. Skor yang lebih rendah (mendekati 0) menunjukkan clustering yang lebih baik, dengan klaster yang lebih kompak dan terpisah satu sama lain.

Berikut adalah hasil evaluasi untuk setiap metode linkage:

1. Single Linkage
    - Silhouette Score: 0.1189
    - Davies-Bouldin Score: 1.2681
    - Analisis: Metode ini menunjukkan Silhouette Score tertinggi, mengindikasikan pemisahan antar klaster yang relatif baik. Davies-Bouldin Score-nya juga cukup rendah dibandingkan beberapa metode lain. Namun, seperti yang dijelaskan sebelumnya, Single Linkage rentan terhadap pembentukan klaster yang sangat tidak seimbang (misalnya, satu klaster besar dan klaster kecil lainnya), yang kurang ideal untuk segmentasi praktis meskipun skor metriknya terlihat menjanjikan.
2. Complete Linkage
    - Silhouette Score: 0.0798
    - Davies-Bouldin Score: 2.4488
    - Analisis: Metode ini memiliki Silhouette Score terendah dan Davies-Bouldin Score tertinggi. Ini menunjukkan kualitas klaster yang paling buruk di antara semua metode yang diuji, dengan klaster yang kurang terpisah dan kurang kompak.
3. Average Linkage
    - Silhouette Score: 0.0496

- Davies-Bouldin Score: 1.1828
- Analisis: Average Linkage memiliki Davies-Bouldin Score terendah, yang mengindikasikan klaster yang sangat kompak. Namun, Silhouette Score-nya juga sangat rendah, menunjukkan bahwa meskipun klaster mungkin padat, pemisahan antara klaster-klaster tersebut sangat buruk.

4. Ward Linkage
    - Silhouette Score: 0.1015
    - Davies-Bouldin Score: 2.1403
    - Analisis: Ward Linkage menunjukkan Silhouette Score yang cukup tinggi dan Davies-Bouldin Score yang moderat. Ini merepresentasikan keseimbangan yang baik antara pemisahan antar klaster dan kekompakan klaster.

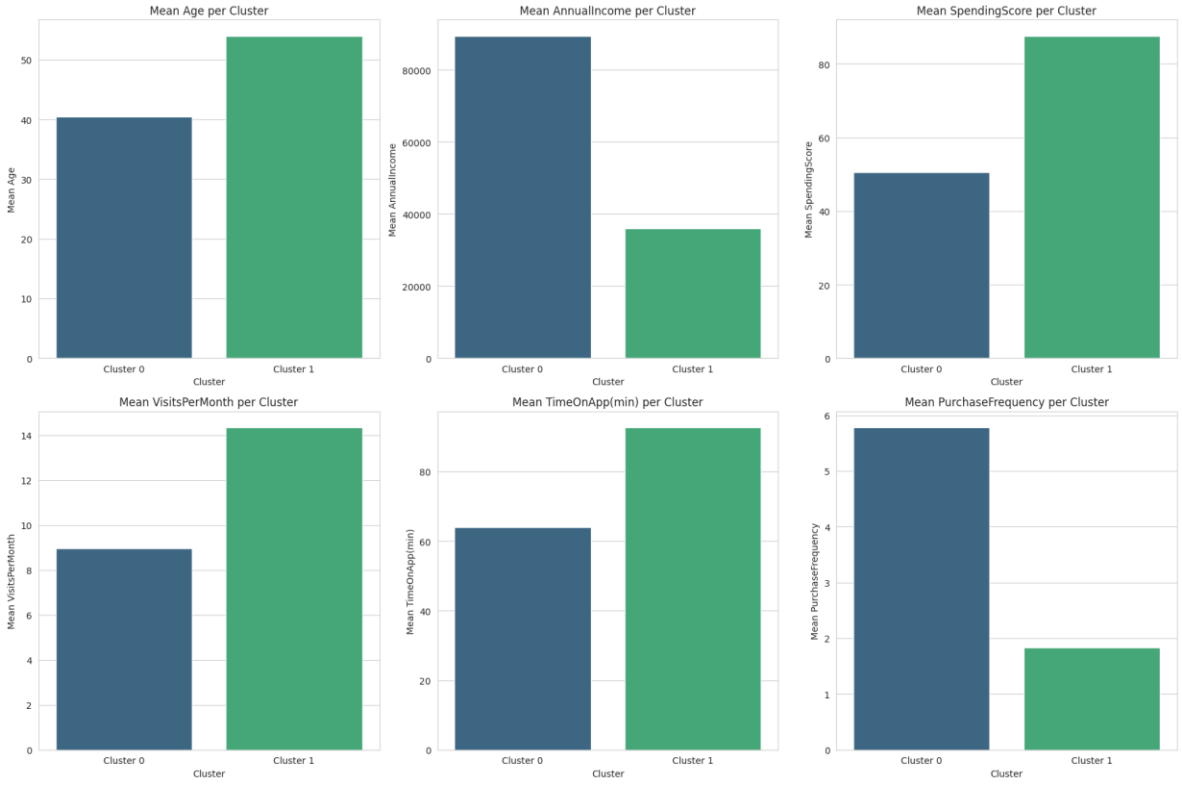
Linkage Mana yang Memberikan Kualitas Cluster Terbaik dan Mengapa?

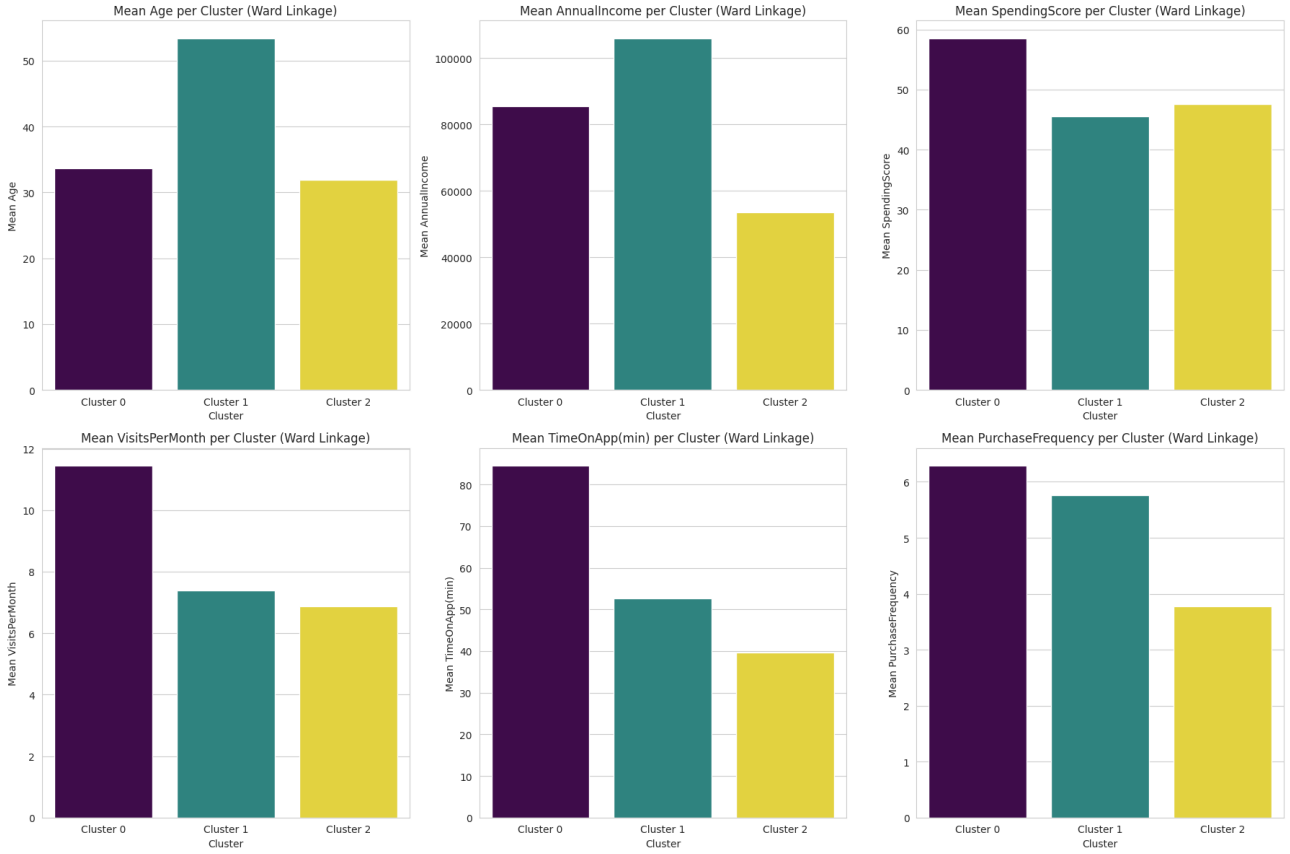
Berdasarkan perbandingan di atas, Ward Linkage dianggap memberikan kualitas klaster terbaik untuk tujuan segmentasi pelanggan dalam kasus ini, meskipun Single Linkage memiliki Silhouette Score yang sedikit lebih tinggi.

Alasannya adalah:

1. Keseimbangan Metrik: Ward Linkage menawarkan keseimbangan yang lebih baik antara Silhouette Score dan Davies-Bouldin Score. Ini menunjukkan bahwa klaster yang dihasilkan tidak hanya cukup terpisah dari satu sama lain tetapi juga cukup kompak secara internal.
2. Struktur Klaster yang Praktis: Meskipun Single Linkage memiliki Silhouette Score tertinggi, ia sering menghasilkan klaster yang sangat tidak seimbang (satu klaster besar dan klaster lainnya sangat kecil, atau klaster yang 'memanjang'). Struktur klaster seperti itu seringkali tidak memberikan wawasan bisnis yang berarti atau tidak dapat ditindaklanjuti secara efektif. Ward Linkage, di sisi lain, cenderung menghasilkan klaster yang lebih bulat, seimbang, dan lebih mudah diinterpretasikan secara visual maupun secara semantik untuk membentuk segmen pelanggan yang jelas.
3. Interpretasi Bisnis: Dalam konteks aplikasi praktis seperti segmentasi pelanggan, kemampuan untuk menafsirkan dan menindaklanjuti klaster yang dihasilkan jauh lebih penting. Ward Linkage memberikan klaster yang profilnya lebih jelas dan membedakan segmen pelanggan dengan karakteristik yang lebih terdefinisi, seperti yang kita lihat pada analisis klaster sebelumnya. Ini memfasilitasi perumusan strategi pemasaran yang lebih tepat sasaran.

VI. Analisis dan Interpretasi tiap Cluster





1. Gambaran Umum Data

   Dataset berisi informasi pelanggan termasuk Age (Usia), AnnualIncome (Pendapatan Tahunan), SpendingScore (Skor Pengeluaran), VisitsPerMonth (Kunjungan Per Bulan), TimeOnApp(min) (Waktu di Aplikasi dalam menit), dan PurchaseFrequency (Frekuensi Pembelian).

   a. Distribusi:
      - Age menunjukkan distribusi yang relatif merata di seluruh kelompok usia, dengan sedikit puncak di sekitar usia pertengahan 20-an, awal 40-an, dan awal 60-an.
      - AnnualIncome juga tersebar luas, dengan beberapa pelanggan di kelompok pendapatan rendah (20 ribu-40 ribu) dan kelompok pendapatan tinggi (120 ribu-140 ribu).
      - SpendingScore memiliki distribusi bimodal, dengan sejumlah besar pelanggan memiliki skor yang sangat rendah (sekitar 0-20) dan sangat tinggi (sekitar 80-100).
      - VisitsPerMonth cenderung ke arah kunjungan yang lebih rendah (1-5 kunjungan), tetapi juga memiliki kelompok yang signifikan dengan kunjungan yang lebih tinggi (15-19 kunjungan).
      - TimeOnApp(min) menunjukkan tingkat keterlibatan yang bervariasi, dengan puncak di sekitar 1-20 menit dan 60-80 menit.
      - PurchaseFrequency agak merata distribusinya, dengan sedikit puncak pada frekuensi yang lebih rendah dan lebih tinggi.

   b. Korelasi: Matriks korelasi menunjukkan korelasi yang umumnya lemah antara sebagian besar fitur, menunjukkan bahwa pelanggan tidak selalu menunjukkan hubungan linier yang kuat di seluruh atribut ini. Misalnya, Usia menunjukkan korelasi yang sangat lemah dengan fitur-fitur lain, dan demikian pula, Pendapatan Tahunan memiliki korelasi yang rendah. Ini menunjukkan bahwa pendekatan segmentasi berdasarkan banyak fitur dapat mengungkapkan pengelompokan yang tidak jelas.

2. Karakteristik Segmen Pelanggan (Linkage Ward, 3 Klaster):

   Teridentifikasi tiga segmen pelanggan yang berbeda berdasarkan clustering linkage ward:

   a. Klaster 0: Pembelanja Muda, Sangat Terlibat
      - Usia: Relatif muda (rata-rata 33.7 tahun).

- Pendapatan Tahunan: Sedang hingga tinggi (rata-rata 85.515).
- Skor Pengeluaran: Tinggi (rata-rata 58.6).
- Kunjungan Per Bulan: Tinggi (rata-rata 11.4).
- Waktu di Aplikasi (menit): Sangat tinggi (rata-rata 84.5 menit).
- Frekuensi Pembelian: Tinggi (rata-rata 6.3).
- Wawasan: Segmen ini terdiri dari pelanggan yang lebih muda yang sangat aktif di aplikasi, sering berkunjung, banyak berbelanja, dan sering melakukan pembelian. Mereka mewakili segmen yang berharga untuk strategi yang didorong oleh keterlibatan.

b. Klaster 1: Loyalis Dewasa, Berpenghasilan Tinggi
   - Usia: Segmen tertua (rata-rata 53.4 tahun).
   - Pendapatan Tahunan: Pendapatan tertinggi (rata-rata 105.985).
   - Skor Pengeluaran: Sedang (rata-rata 45.6).
   - Kunjungan Per Bulan: Sedang (rata-rata 7.4).
   - Waktu di Aplikasi (menit): Sedang (rata-rata 52.7 menit).
   - Frekuensi Pembelian: Sedang hingga tinggi (rata-rata 5.8).
   - Wawasan: Ini adalah pelanggan dewasa dan makmur yang berkontribusi signifikan melalui pendapatan mereka yang lebih tinggi dan perilaku yang konsisten, meskipun tidak terlalu sering, keterlibatan. Mereka mungkin kurang impulsif tetapi kemungkinan besar adalah pelanggan yang stabil dan berharga.

c. Klaster 2: Pengunjung Muda, Hemat Anggaran
   - Usia: Segmen termuda (rata-rata 31.9 tahun).
   - Pendapatan Tahunan: Pendapatan terendah (rata-rata 53.42).
   - Skor Pengeluaran: Sedang (rata-rata 47.6).
   - Kunjungan Per Bulan: Kunjungan terendah (rata-rata 6.9).
   - Waktu di Aplikasi (menit): Keterlibatan aplikasi terendah (rata-rata 39.6 menit).
   - Frekuensi Pembelian: Frekuensi pembelian terendah (rata-rata 3.8).
   - Wawasan: Segmen ini terdiri dari pelanggan termuda dengan pendapatan terendah dan keterlibatan paling sedikit di seluruh kunjungan, waktu di aplikasi, dan frekuensi pembelian. Mereka mungkin pengguna baru, pembeli sesekali, atau lebih sensitif terhadap harga.

3. Pola yang Muncul dan Wawasan yang Relevan
   a. Trade-off Keterlibatan vs. Kemakmuran: Kami mengamati segmen yang berbeda di mana keterlibatan tinggi (Klaster 0) berkorelasi dengan usia yang lebih muda dan pengeluaran yang tinggi, sementara pendapatan yang lebih tinggi (Klaster 1) sesuai dengan usia yang lebih tua dan perilaku yang sedang, tetapi konsisten. Kelompok pendapatan terendah (Klaster 2) menunjukkan keterlibatan terendah.
   b. Peluang Pemasaran Bertarget:
      - Klaster 0 (Pembelanja Muda, Sangat Terlibat): Fokus pada program loyalitas, pengumuman produk baru, dan penawaran eksklusif untuk mempertahankan keterlibatan dan pengeluaran tinggi mereka. Gamifikasi atau fitur komunitas dapat lebih meningkatkan pengalaman mereka.
      - Klaster 1 (Loyalis Dewasa, Berpenghasilan Tinggi): Sesuaikan komunikasi seputar produk/layanan premium, rekomendasi yang dipersonalisasi, dan pengalaman bernilai tambah yang sesuai dengan pendapatan mereka yang lebih tinggi dan preferensi yang mungkin lebih cerdas. Strategi retensi adalah kunci untuk segmen yang stabil ini.
      - Klaster 2 (Pengunjung Muda, Hemat Anggaran): Terapkan strategi untuk meningkatkan keterlibatan dan konversi. Ini dapat melibatkan penawaran perkenalan, konten edukasi tentang manfaat produk, atau kampanye promosi yang menyoroti nilai. Memahami hambatan keterlibatan mereka (misalnya, harga, relevansi yang dirasakan) sangat penting.

Segmentasi pelanggan ini menyediakan kerangka kerja yang jelas untuk mengembangkan strategi pemasaran yang lebih efektif dan personal, yang pada akhirnya mengarah pada peningkatan kepuasan pelanggan dan pertumbuhan bisnis.

VII. Kesimpulan

1. Ringkasan Performa Tiap Linkage

   Dalam analisis ini, empat metode linkage (single, complete, average, dan ward) dievaluasi menggunakan Silhoutte Score dan Davies-Bouldin Score:

   a. Single Linkage (2 klaster): Menghasilkan Silhouette Score 0.1189 dan Davies-Bouldin Score 1.2681. Meskipun memiliki Silhouette Score yang relatif baik, metode ini cenderung membentuk klaster yang memanjang dan kurang kompak, seperti yang terlihat pada dendrogram.
   b. Complete Linkage (3 klaster): Menghasilkan Silhouette Score 0.0798 dan Davies-Bouldin Score 2.4488. Metode ini cenderung menghasilkan klaster yang lebih kompak, namun memiliki skor evaluasi terendah.
   c. Average Linkage (3 klaster): Menghasilkan Silhouette Score 0.0496 dan Davies-Bouldin Score 1.1828. Metode ini mencapai Davies-Bouldin Score terendah (yang menunjukkan klasterisasi terbaik dalam hal kepadatan dan pemisahan), tetapi Silhouette Score-nya juga yang terendah.
   d. Ward Linkage (3 klaster): Menghasilkan Silhouette Score 0.1015 dan Davies-Bouldin Score 2.1403. Meskipun skor evaluasinya tidak selalu yang terbaik secara numerik, dendrogram menunjukkan klaster yang 'terpisah dengan baik' dan cenderung menghasilkan klaster yang lebih bulat dan seimbang.

2. Linkage Terbaik

   Berdasarkan pertimbangan visual dari dendrogram yang menunjukkan pemisahan klaster yang jelas dan interpretabilitas yang baik, serta metrik evaluasi yang kompetitif, Ward Linkage dengan 3 klaster dipilih sebagai metode terbaik untuk analisis ini. Ward linkage secara inheren bertujuan untuk meminimalkan varians dalam setiap klaster, sehingga menghasilkan klaster yang lebih homogen dan kompak, yang seringkali lebih mudah diinterpretasikan dalam konteks bisnis.

3. Rekomendasi Penggunaan Agglomerative Clustering dalam Kasus Nyata

   Agglomerative Clustering terbukti menjadi alat yang ampuh untuk segmentasi pelanggan. Dalam kasus nyata, penggunaan metode ini direkomendasikan untuk:

   a. Segmentasi Pelanggan: Mengidentifikasi kelompok pelanggan yang berbeda berdasarkan karakteristik dan perilaku mereka, memungkinkan perusahaan untuk mengembangkan strategi pemasaran yang lebih tepat sasaran dan personalisasi pengalaman pelanggan.
   b. Pengembangan Produk/Layanan: Memahami kebutuhan dan preferensi unik setiap segmen dapat memandu pengembangan produk atau modifikasi layanan yang lebih relevan.
   c. Strategi Retensi dan Akuisisi: Dengan mengetahui segmen pelanggan mana yang paling berharga atau yang berisiko churn, perusahaan dapat merancang program loyalitas atau kampanye akuisisi yang efektif.
   d. Optimalisasi Anggaran Pemasaran: Mengalokasikan sumber daya pemasaran secara lebih efisien dengan menargetkan segmen yang paling responsif atau berpotensi tinggi.
   e. Analisis Tren: Memantau bagaimana segmen berubah dari waktu ke waktu dapat memberikan wawasan tentang dinamika pasar dan perilaku pelanggan yang berkembang.