PROYEK UAS STATISTIKA MULTIVARIAT
CUSTOMER SEGMENTATION

Disusun oleh:
Kelompok C

Leonardo Alindra Mauliate	6162001111
Jesslyn Angelica	6162001007
Jessica Caroline Joshua	6162001026
Cristiana Devina Wigono	6162001119
Elysia Vanessa Prasetyo	6162001153

PROGRAM STUDI MATEMATIKA
FAKULTAS TEKNOLOGI INFORMATIKA DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN

2023

BAB I

PENDAHULUAN

Deskripsi Masalah

Dalam berbisnis, menentukan target pasar yang baik adalah salah satu hal yang perlu dipertimbangkan agar bisnis dapat berkembang. Penentuan target pasar dimulai dengan menentukan segmentasi pasar, yaitu dengan membentuk pemetaan atau pengelompokkan target pelanggan berdasarkan karakteristik, kebutuhan, serta perilaku pelanggan. Dengan mengkategorikan pelanggan berdasarkan hal-hal tersebut, segmentasi pasar dapat meningkatkan pelayanan serta meningkatkan efektifitas strategi pemasaran produk. Melalui segmentasi pasar juga, kebutuhan pelanggan dapat terpenuhi dan daya tarik pelanggan dapat meningkat. Suatu perusahaan juga dapat memfokuskan kelompok target tertentu agar pemasaran menjadi lebih terarah.

Pada laporan ini, akan dilakukan *customer segmentation* dengan *agglomerative clustering* untuk dataset yang berisikan data pelanggan dari toko bahan makanan. Selain itu, akan dilakukan pengembangan dengan melakukan analisis *K-Means clustering* untuk mengetahui jumlah klaster atau pengelompokkan yang optimal agar toko dapat memodifikasi produk sesuai dengan kebutuhan dan perilaku pelanggan, dan membantu toko memenuhi kebutuhan berbagai jenis pelanggan. Analisis ini pun diharapkan dapat meningkatkan penjualan dengan kegiatan promosi yang ditargetkan pada pelanggan tertentu.

BAB II

PEMBAHASAN

Deskripsi Data

Data yang digunakan berasal dari website *Kaggle*. Dataset ini terdiri atas informasi pribadi *customer*, jumlah pengeluaran berdasarkan produk, jumlah pembelian yang dilakukan dengan diskon, serta tempat pembelian produk. Data terdiri dari 2.240 observasi dan 27 variabel. Adapun rincian dari variabel-variabel untuk informasi pribadi pelanggan yang digunakan adalah sebagai berikut:

People	
ID	Nomor identitas pelanggan
Year_Birth	Tahun lahir pelanggan
Education	Pendidikan terakhir pelanggan
Martial_Status	Status perkawinan pelanggan
Income	Pendapatan per tahun pelanggan
Kidhome	Jumlah anak dalam rumah tangga pelanggan
Teenhome	Jumlah remaja dalam rumah tangga pelanggan
Dt_Customer	Tanggal pendaftaran pelanggan dengan perusahaan
Recency	Jumlah hari sejak pembelian terakhir pelanggan
Complain	1 jika pelanggan melakukan komplain dalam 2 tahun terakhir, 0 sebaliknya

Kemudian, rincian dari jumlah pengeluaran produk adalah sebagai berikut:

Produk	
MntWines	Jumlah pengeluaran yang dihabiskan pada *wine* selama 2 tahun terakhir
MntFruits	Jumlah pengeluaran yang dihabiskan pada buah selama 2 tahun terakhir
MntMeatProducts	Jumlah pengeluaran yang dihabiskan pada daging selama 2 tahun terakhir
MntFishProducts	Jumlah pengeluaran yang dihabiskan pada ikan selama 2 tahun terakhir
MntSweetProducts	Jumlah pengeluaran yang dihabiskan pada manisan selama 2 tahun terakhir
MntGoldProds	Jumlah pengeluaran yang dihabiskan untuk emas selama 2 tahun terakhir

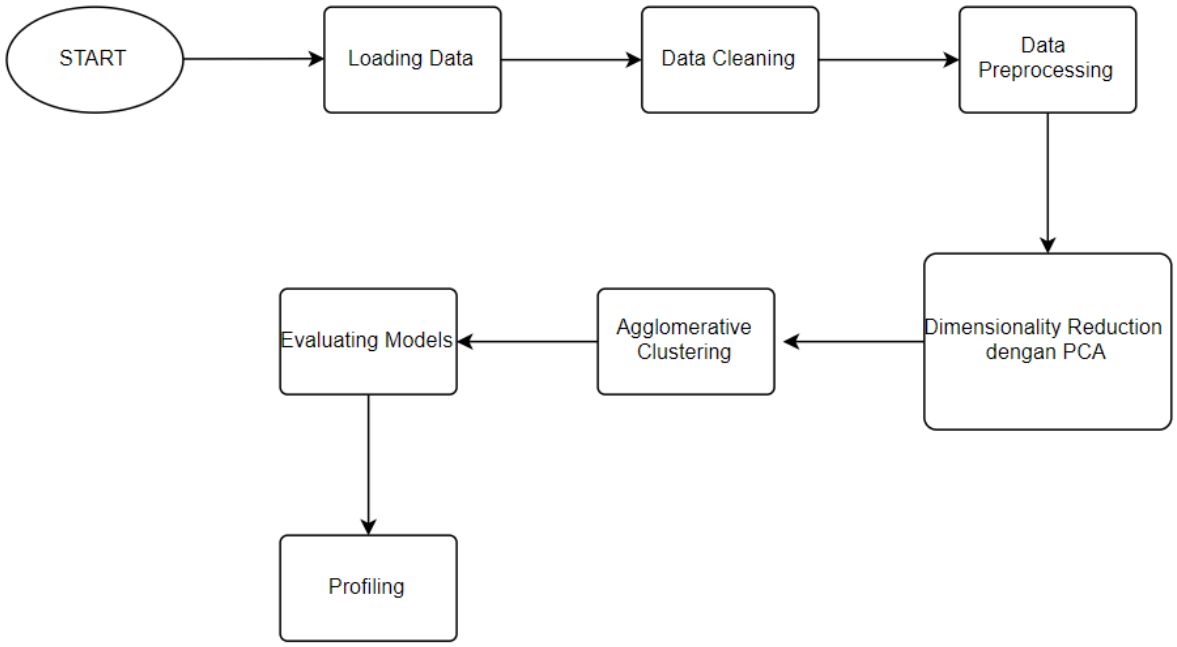
Berikut merupakan rincian variabel dari jumlah pembelian yang dilakukan dengan promosi atau diskon:

Promosi	
NumDealsPurchase	Jumlah pembelian yang dilakukan dengan diskon
AcceptedCmp1	1 jika pelanggan menerima tawaran pada *campaign 1,* 0 sebaliknya
AcceptedCmp2	1 jika pelanggan menerima tawaran pada *campaign* 2, 0 sebaliknya
AcceptedCmp3	1 jika pelanggan menerima tawaran pada *campaign 3*, 0 sebaliknya
AcceptedCmp4	1 jika pelanggan menerima tawaran pada *campaign 4,* 0 sebaliknya
AcceptedCmp5	1 jika pelanggan menerima tawaran pada *campaign 5*, 0 sebaliknya
Response	1 jika pelanggan menerima tawaran pada *campaign* terakhir, 0 sebaliknya

Berikut merupakan rincian variabel dari tempat pembelian produk yang dilakukan pelanggan:

Tempat	
NumWebPurchases	Jumlah pembelian melalui situs perusahaan
NumCatalogPurchases	Jumlah pembelian dengan menggunakan katalog
NumStorePurchases	Jumlah pembelian melalui toko
NumWebVisitsMonth	Jumlah kunjungan pada situs perusahaan selama 1 bulan terakhir

Alur penyelesaian



1. *Loading Data*

   Pertama-tama, data akan dimuat dengan bantuan Python agar data dapat digunakan untuk pemrosesan serta analisis. Proses *loading data* ini antara lain adalah mengimport *library* untuk membuka sumber data, membaca data, dan menyimpannya. Data kemudian akan diolah sesuai dengan kebutuhan.

2. *Data Cleaning*

   Sebelum data diproses lebih lanjut, perlu adanya *data cleaning* yaitu proses membersihkan dan mempersiapkan data agar data menjadi lebih terstruktur. Selain itu, *data cleaning* dilakukan untuk mengidentifikasi dan memperbaiki apabila terdapat ketidaksesuaian di data mentah, seperti data hilang atau data yang terulang.

3. *Data Preprocessing*

   Data kemudian diproses dengan normalisasi atau standarisasi data untuk meningkatkan kualitas data, menghilangkan kecacatan, dan mempersiapkan data agar sesuai dengan persyaratan analisis *clustering* atau segmentasi yang akan digunakan.

4. *Dimensionality Reduction* dengan *Principal Component Analysis* (PCA)

   Reduksi dimensi atau jumlah variabel pada dataset yang kompleks digunakan untuk mengurangi kompleksitas data. Pada laporan ini, PCA digunakan agar dapat mengidentifikasi pola tersembunyi dalam data dan diubah ke dalam ruang fitur yang lebih rendah disebut komponen utama. PCA dapat menghilangkan jumlah komponen sambil mempertahankan informasi yang sesuai.

5. *Clustering*

   Metode yang digunakan adalah *agglomerative clustering*. Metode ini termasuk dalam analisis klaster hierarki. Pertama, setiap data diasumsikan sebagai satu klaster. Kemudian, kita hitung jarak *Euclid* dari setiap pasang data untuk mendapatkan ukuran kesamaan dari setiap klaster. Untuk mengelompokkan klaster yang mirip, kita gunakan metode *Ward*. Metode *Ward* memanfaatkan informasi jarak *Euclid* untuk menggabungkan pasangan klaster yang memiliki variansi minimum.

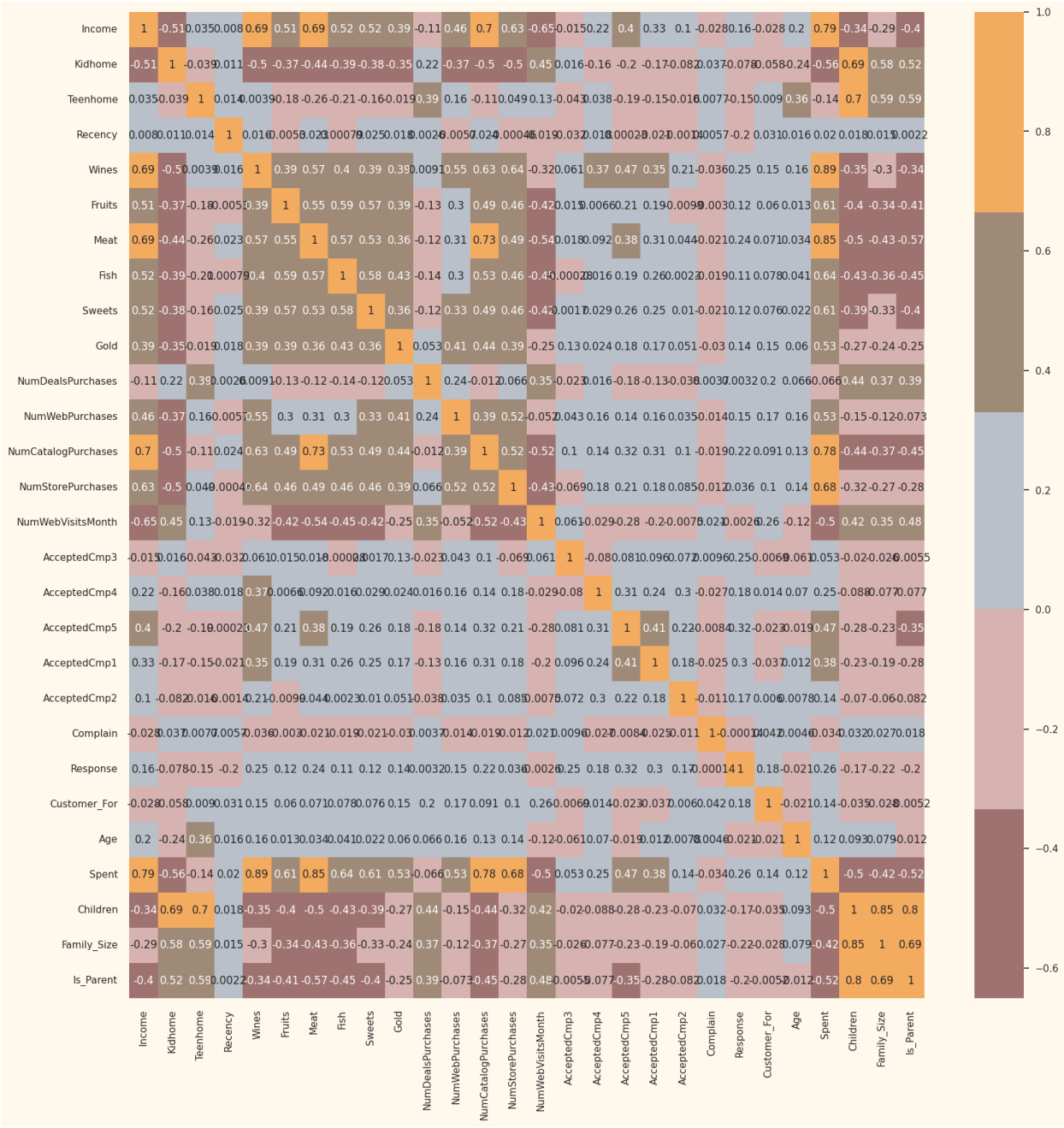
6. *Evaluating models*

   Pada tahap ini, akan dipelajari pola dalam setiap klaster yang terbentuk dan menentukan sifat dari pola tersebut. Hasilnya akan digunakan untuk menilai apakah model atau *clustering* yang dibuat sudah baik.

7. *Profiling*

   Data yang sudah diklaster akan dianalisis. Dengan mengidentifikasikan persamaan dan perbedaan dari masing-masing klaster melalui visualisasi data, akan dilakukan *profiling* untuk setiap klaster yang terbentuk dan menyimpulkan siapa pelanggan utama dan pelanggan yang membutuhkan perhatian lebih dari tim pemasaran.

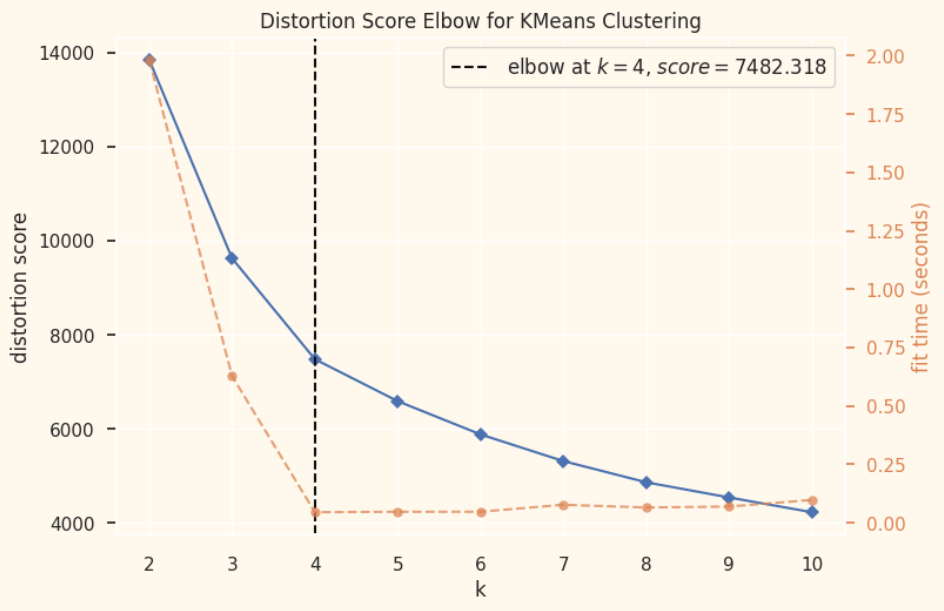
Analisis Hasil



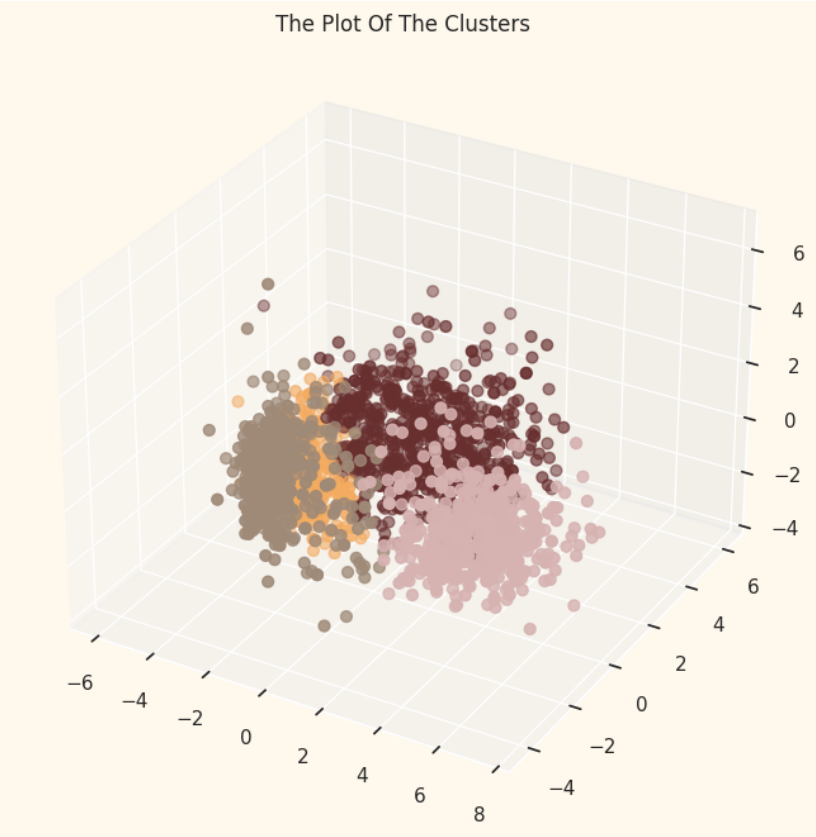
Gambar di atas merupakan matriks korelasi dari setiap variabel yang digunakan. Matriks korelasi berfungsi untuk melihat korelasi atau hubungan antar dua variabel. Jika nilai dari koefisien korelasi mendekati 1, artinya kedua variabel memiliki korelasi positif yang kuat. Tetapi jika koefisien korelasi mendekati -1, artinya kedua variabel memiliki korelasi negatif yang kuat. Jika

koefisien korelasi mendekati 0, artinya kedua variabel tidak memiliki korelasi. Koefisien dari matriks korelasi berdasarkan gambar di atas berkisar dari 1 hingga -0,6.

Korelasi paling kuat adalah antara variabel ‘Spent’ dan ‘Wines’ dengan koefisien sebesar 0,89. Artinya pelanggan lebih banyak menghabiskan uangnya untuk membeli *wine*. Hal tersebut dapat dilihat juga dari koefisien korelasi antara variabel “Income” dan “Wines” bernilai 0,69, artinya semakin tinggi pendapatan pelanggan, makan semakin banyak *wine* yang dibeli oleh pelanggan tersebut. Selain itu, variabel ‘Spent’ juga memiliki korelasi yang kuat dengan variabel ‘Meat’. Artinya selain membeli *wine*, pelanggan juga banyak membeli daging.

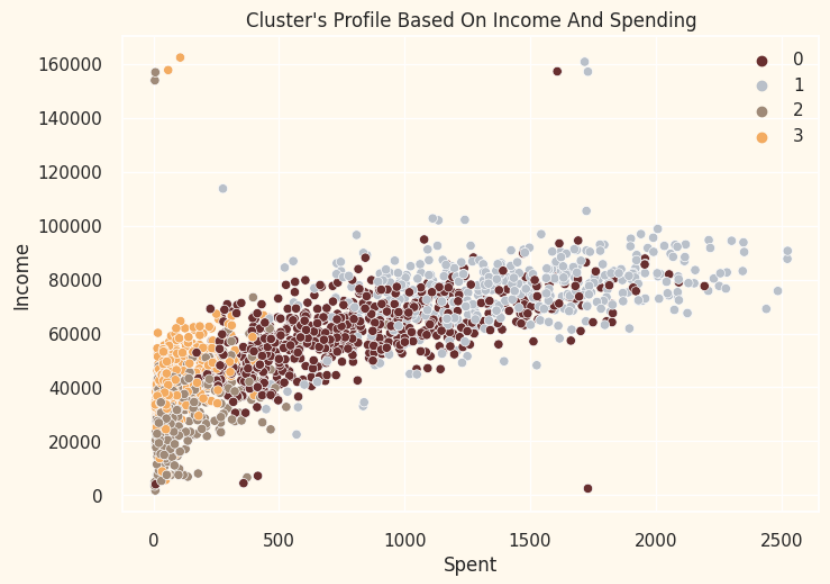


Berdasarkan metode *Elbow*, dapat dilihat bahwa pada jumlah klaster 4, garis menyerupai bentuk siku atau *Elbow*. Artinya, 4 merupakan jumlah klaster yang optimal. Hal tersebut dikarenakan penambahan klaster lebih lanjut tidak memberikan penurunan *Sum of Square Error* (SSE) yang signifikan.

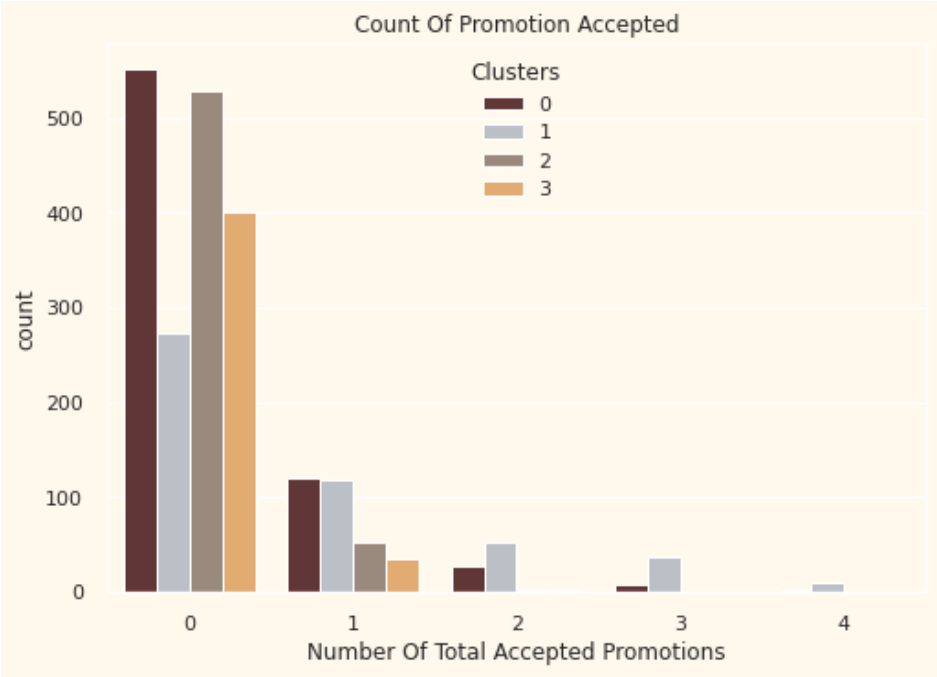


Gambar di atas merupakan visualisasi hasil *clustering* pelanggan menggunakan metode *agglomerative clustering*. Pelanggan dibagi menjadi 4 klaster yang dibedakan berdasarkan warnanya. Klaster 0 berwarna merah tua, klaster 1 berwarna abu-abu, klaster 2 berwarna coklat muda, dan klaster 3 berwarna oranye.

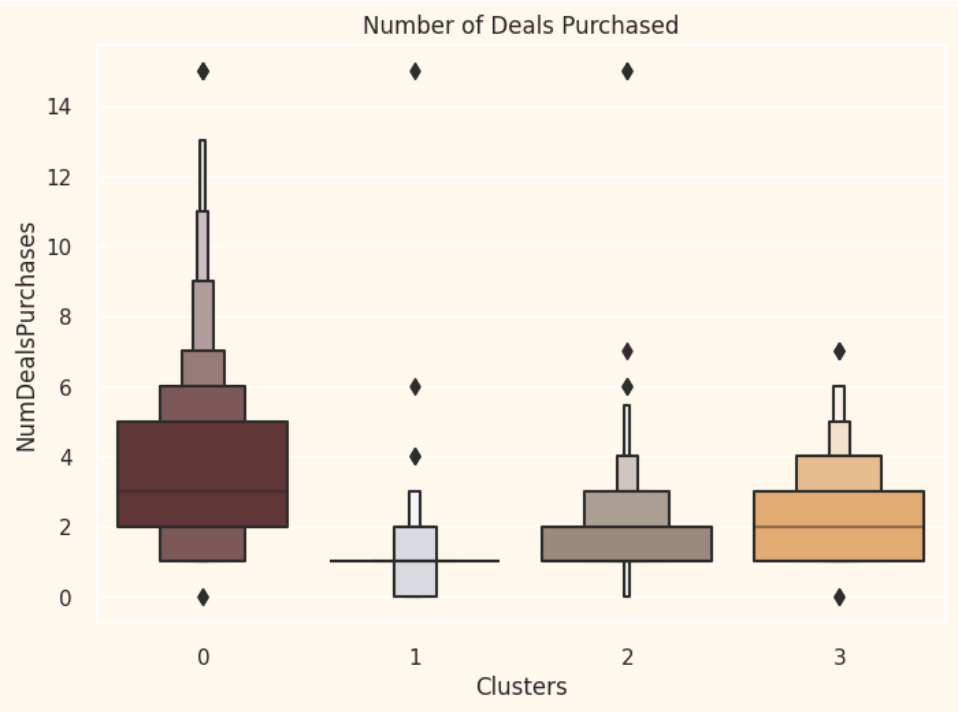
Gambar di atas adalah *bar plot* untuk menunjukkan jumlah pelanggan dari setiap klaster yang telah terbentuk, yaitu klaster 0, 1, 2, dan 3. Berdasarkan *bar plot* tersebut, dapat dilihat bahwa jumlah observasi atau pelanggan pada tiap klaster terbagi secara cukup merata. Klaster dengan jumlah pelanggan terbanyak ada pada klaster 0 yaitu sekitar 700 orang, sedangkan klaster dengan jumlah pelanggan paling sedikit adalah klaster 3 yaitu sekitar 400 orang.



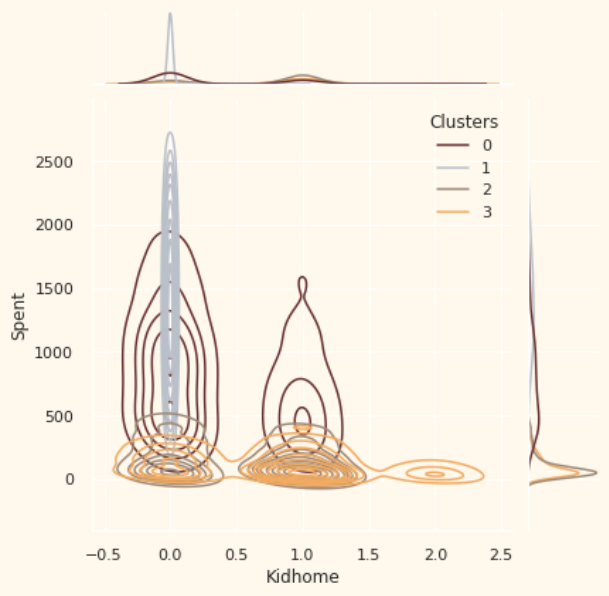
Berdasarkan gambar di atas, dapat disimpulkan beberapa ciri dari setiap klaster. Klaster 0 memiliki pengeluaran rata-rata dan pendapatan rata-rata yang besar, di mana pendapatan dan pengeluaran pelanggan pada klaster 0 berada di antara pendapatan dan pengeluaran pelanggan pada klaster 1 sampai klaster 3. Klaster 1 ditandai dengan titik-titik berwarna abu. Pada klaster 1, pelanggan memiliki pengeluaran dan juga pendapatan yang besar. Klaster 2 ditandai dengan titik berwarna ungu muda, di mana klaster 2 memiliki pengeluaran dan pendapatan yang kecil. Lalu, klaster 3 ditandai dengan titik-titik berwarna oranye. Pada klaster 3, pelanggan memiliki pendapatan rata-rata namun pengeluaran yang kecil.

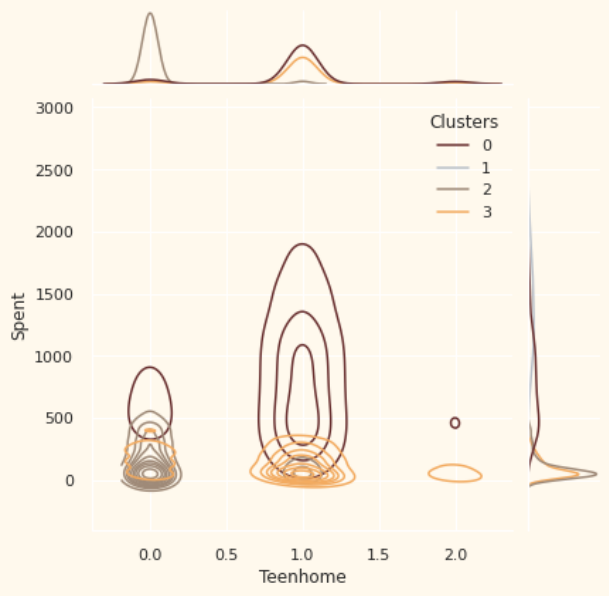


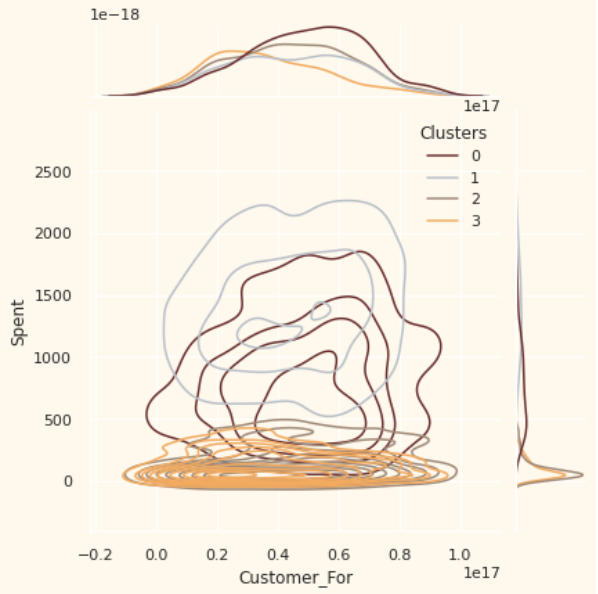
Berdasarkan gambar di atas, mayoritas pelanggan tidak tertarik untuk mengikuti program promosi yang diselenggarakan. Hanya beberapa pelanggan pada klaster 1 yang melakukan 4 kali pembelian pada saat berlangsungnya program promosi. Dari seluruh pelanggan, tidak ada satupun pelanggan yang mengikuti program promosi sebanyak 5 kali. Artinya, program promosi yang dilakukan masih kurang menarik untuk membuat pelanggan membeli produk yang dipromosikan.

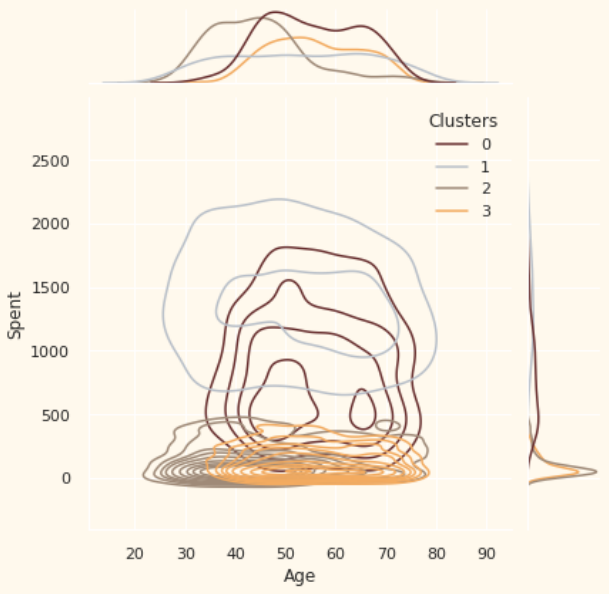


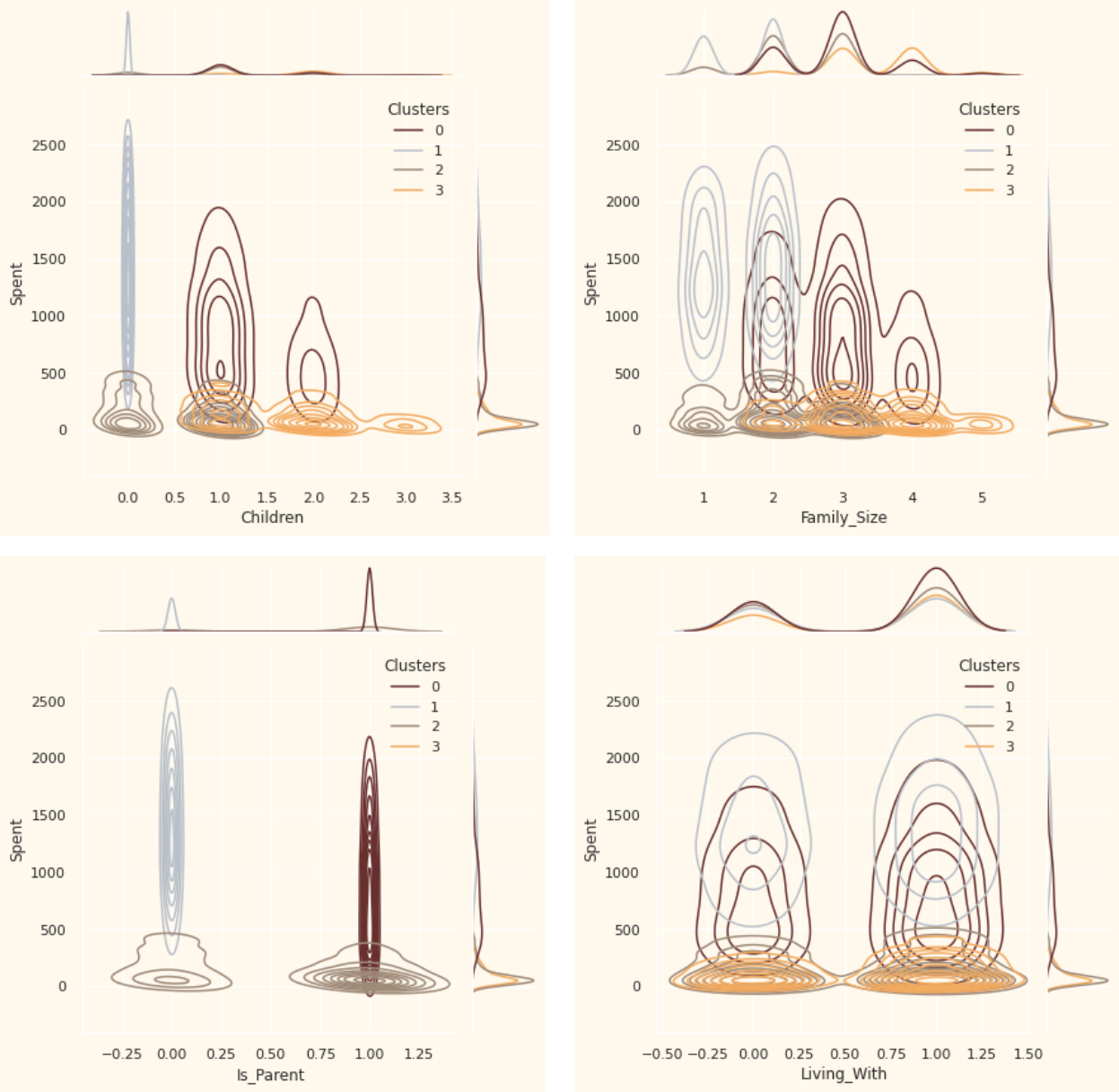
Berdasarkan gambar di atas, cukup banyak pelanggan yang membeli produk dengan harga diskon. Mayoritas pelanggan yang melakukan pembelian produk dengan harga diskon berada di klaster 0, sedangkan pelanggan pada klaster 1 dan 2 tampak tidak terpengaruh dengan potongan harga yang diberikan pada produk.


Clusters
0
1
2
3
Spent
Kidhome


Clusters
0
1
2
3
Spent
Teenhome


1e−18
1e17
Clusters
0
1
2
3
Spent
Customer_For
1e17


Clusters
0
1
2
3
Spent
Age



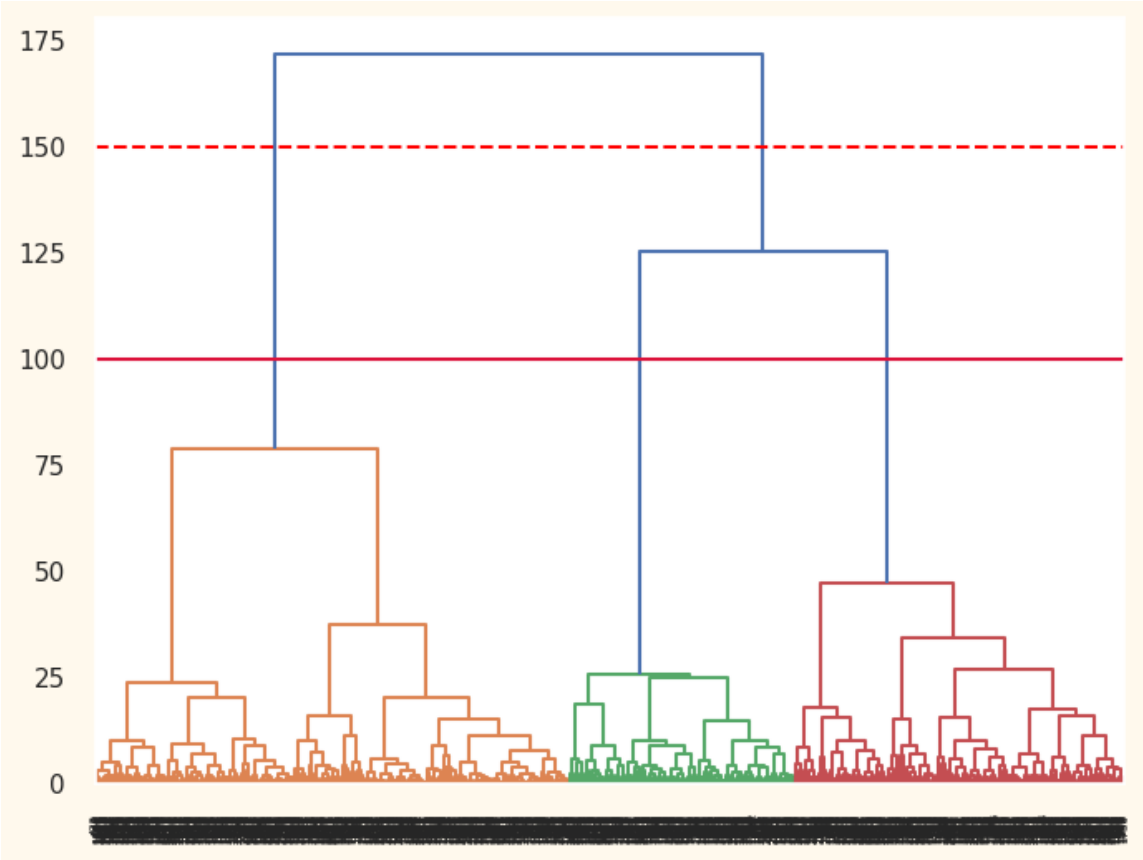
Berdasarkan 8 gambar di atas, profil dari masing-masing klaster dapat ditentukan. Pelanggan pada klaster 0 merupakan orang tua dengan anggota keluarga minimal berjumlah 2 dan maksimal 4 orang. Pelanggan pada klaster 0 berusia lebih tua dibandingkan pelanggan lain dan mayoritas memiliki anak remaja di rumah. Pelanggan pada klaster 1 bukan orang tua yang memiliki anak, melainkan seorang lajang atau memiliki pasangan. Mereka memiliki pendapatan yang besar dan tersebar di segala usia. Pelanggan pada klaster 2 mayoritas merupakan orang tua berusia muda yang memiliki 1 anak kecil. Mereka memiliki maksimal 3 anggota keluarga. Pelanggan pada

klaster 3 merupakan orang tua yang memiliki anak remaja dan minimal 2 anggota keluarga dan maksimal 5 anggota keluarga. Pelanggan pada klaster 3 berusia lebih tua dibandingkan pelanggan lain dan memiliki pendapatan yang kecil.

BAB III

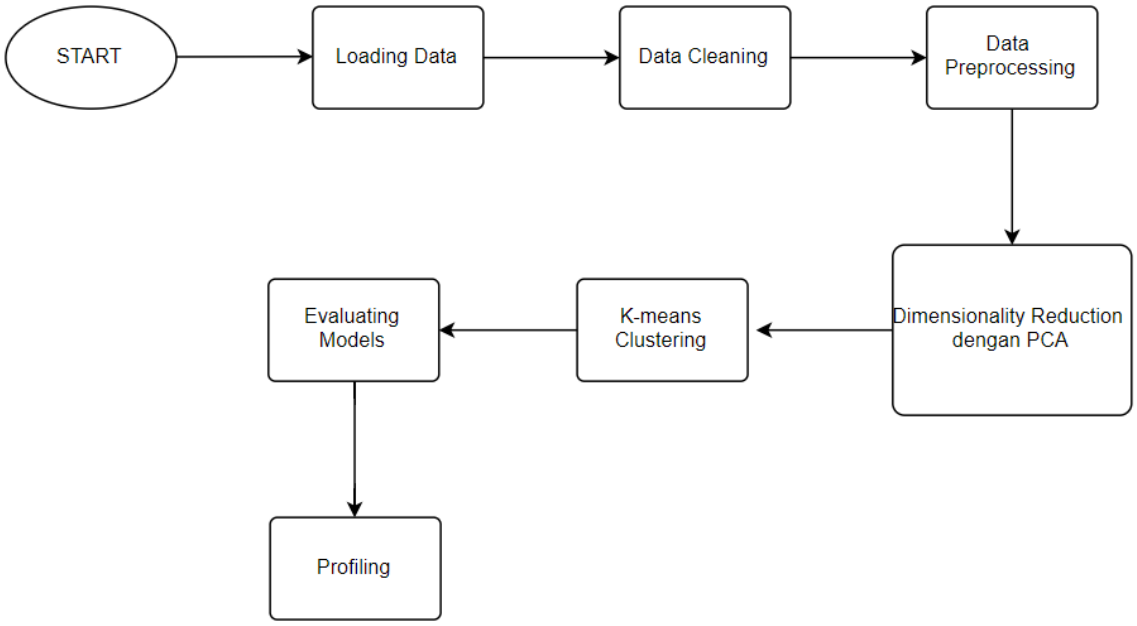
PENGEMBANGAN

Salah satu kelemahan dari *agglomerative clustering* yaitu sangat bergantung pada inisialisasi yang tepat sehingga pemilihan titik awal dapat mempengaruhi hasil akhir. Apabila inisialisasi tidak optimal, maka klaster hierarki yang terbentuk tidak akan sesuai. Selain itu, *agglomerative clustering* yang termasuk ke dalam *hierarchy clustering* ini tidak memerlukan jumlah klaster yang sudah ditentukan sebelumnya, di mana jumlah klaster yang digunakan belum tentu merupakan jumlah klaster optimal. Pada sumber *Kaggle*, nilai klasternya sudah ditentukan terlebih dahulu untuk *hierarchy clustering,* padahal seharusnya jumlah clusternya tidak boleh ditentukan. Ketika dilakukan *clustering* secara hierarki tanpa menentukan jumlah clusternya, diperoleh bahwa jumlah klasternya sebanyak 3 dan dapat dilihat di gambar berikut.



Oleh karena itu, pengembangan akan digunakan menggunakan metode yang lebih baik yaitu metode *K-Means clustering* dengan penentuan jumlah klaster menggunakan metode *Elbow*. Metode *K-Means clustering* adalah salah satu metode dalam analisis klaster yang digunakan untuk mengelompokkan data menjadi beberapa klaster berdasarkan kesamaan atribut atau ciri-ciri yang dimiliki. Setiap klaster kelompok akan diwakili oleh sebuah *centroid* atau pusat klaster. Tujuan utama dari metode ini adalah meminimalkan variansi di dalam setiap klaster dan memaksimalkan perbedaan antara klaster-klasternya.

Alur Penyelesaian



Langkah-langkah dalam metode ini sama dengan sebelumnya. Perbedaannya terdapat pada metode clustering yang digunakan, yaitu *K-Means clustering*. Langkah-langkah dalam metode *K-Means clustering* adalah sebagai berikut.

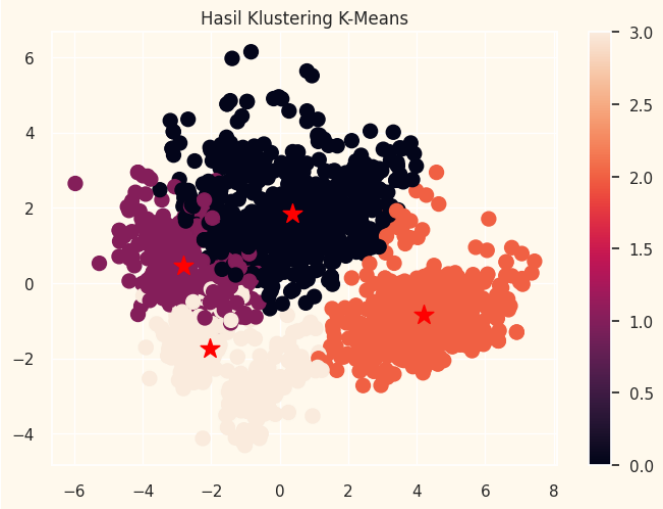
1. Mencari klaster optimal dengan menggunakan *Elbow method*.
2. Pilih k buah titik awal secara acak sebagai pusat klaster.
3. Hitung jarak antara setiap titik data dengan pusat klaster yang terdekat.
4. Setiap data akan dikelompokkan ke dalam klaster berdasarkan pusat klaster terdekat.

5. Tentukan pusat klaster yang baru untuk setiap klaster berdasarkan data yang sudah dikelompokkan dengan mengambil rata-rata dari titik data dalam klaster tersebut.
6. Jika tidak ada lagi perubahan pusat klaster maupun penempatan data, maka proses *K-Means* selesai.

Untuk menentukan metode mana yang lebih baik antara *hierarchy clustering* dan *K-Means clustering*, akan dibandingkan *silhouette score* dari kedua metode tersebut. *Silhouette score* merupakan suatu alat ukur yang digunakan untuk mengukur kebaikan dari hasil *clustering* yang nilainya berkisar antara -1 dan 1. Nilai 1 menandakan bahwa klaster terpisah satu sama lain dan dapat dibedakan dengan jelas, nilai 0 menandakan bahwa klaster tidak berbeda atau saling tumpang tindih, dan nilai -1 menandakan bahwa data dikelompokkan di klaster yang salah. Setelah dicari *silhouette score* untuk masing-masing metode, didapat hasil sebagai berikut:

Metode	*Hierarchy Clustering*	*K-Means Clustering*
Silhouette Score	0,4707980141662763	0,4788958290381151

Dari hasil di atas, *silhouette score* untuk metode *K-Means clustering* lebih besar dibandingkan dengan klaster hierarki, sehingga dapat disimpulkan bahwa metode *K-Means clustering* lebih baik dalam melakukan segmentasi *customer* dan diperoleh hasil klusteringnya seperti gambar berikut.



Berdasarkan gambar di atas, dapat dilihat bahwa terdapat 4 klaster yang dibedakan oleh warna dengan setiap warna memiliki nilai yang berbeda. Klaster 0 berwarna hitam, klaster 1 berwarna ungu, klaster 2 berwarna orange, dan klaster 3 berwarna krem. Berdasarkan hasil yang sudah dibahas pada bab sebelumnya, diperoleh klaster 0 terdiri dari orang tua, keluarga yang terdiri dari 2-4 anggota, bisa jadi *single parent*, mayoritas memiliki anak remaja, dan relatif lebih tua. Klaster 1 terdiri dari bukan orang tua, keluarga yang terdiri dari maksimal 2 anggota, mayoritas orang berpasangan dibandingkan yang *single*, semua usia, dan memiliki pendapat yang tinggi. Klaster 2 terdiri dari mayoritas orang tua, keluarga yang terdiri dari maksimal 3 anggota, mayoritas memiliki 1 anak kecil, dan relatif lebih muda. Klaster 3 terdiri dari orang tua, keluarga yang terdiri dari 2-5 anggota, mayoritas keluarga memiliki anak remaja, relatif lebih tua, dan memiliki pendapatan yang lebih rendah.

BAB 4
KESIMPULAN

Berdasarkan analisis dan pengembangan mengenai *customer segmentation* dengan metode *agglomerative clustering* dan *K-Means clustering* di atas, diperoleh beberapa kesimpulan sebagai berikut:

1. Berdasarkan *hierarchy clustering*, diperoleh klaster sebanyak 3 klaster.
2. Berdasarkan *silhouette score*, metode *K-Means clustering* lebih baik dalam melakukan *customer segmentation* dibandingkan dengan *hierarchy clustering,* dengan perbedaan nilai *silhouette* sebesar 0,008.
3. Berdasarkan *K-Means clustering*, diperoleh klaster optimal sebanyak 4 klaster. Karena metode *K-Means clustering* lebih unggul, dipilih jumlah klaster optimal sebanyak 4 klaster.
4. Klaster terbagi berdasarkan status kekeluargaan, jumlah anggota keluarga, usia anak, serta pendapatan.

LAMPIRAN

https://youtu.be/jOR5OuAIveo