Link Notebook :
https://colab.research.google.com/drive/1AtgpTPeY_01Eil9XL-aqKZKoJlGoBn-g?usp=sharing

A. PENDEFINISIAN MASALAH

Sebuah supermarket ABC sudah berdiri selama kurang lebih 10 tahun, namun beberapa tahun belakangan telah terjadi penurunan profit atau keuntungan. Terdapat beberapa aksi yang sudah dilakukan oleh pihak supermarket, seperti memanfaatkan jasa periklanan, namun masih belum mendapatkan hasil yang diharapkan dan secara signifikan.

B. PENGENALAN, EKSPLORASI DATASET, DAN USULAN TASK YANG DILAKUKAN

1. 50_SupermarketBranches.csv

a. Informasi Dataset

Dataset ini berisi data biaya iklan, biaya promosi dan profit yang diperoleh berdasarkan iklan dan promosi yang dilakukan pada masing-masing lokasi.

b. EDA (Exploratory Data Analysis)

Berikut adalah wawasan tentang data yang didapatkan setelah dilakukannya proses EDA:

1. Dataset ini memiliki ukuran data sebesar 50 baris dengan 5 kolom, yaitu:
   - AdvertisementSpend = Total pengeluaran untuk biaya iklan
   - PromotionSpend = Total pengeluaran untuk biaya promosi
   - AdministrationSpend = Total Pengeluaran untuk biaya administrasi
   - State = Nama negara bagian
   - Profit = Total keuntungan yang melebihi pengeluaran, biaya, dan pajak yang terlibat dalam mempertahankan aktivitas yang bersangkutan.

   Kemudian, dilakukan penambahan kolom baru yaitu TotalSpending yang nilainya merupakan gabungan seluruh spend yaitu AdvertisementSpend, PromotionSpend, dan AdministrationSpend. Sehingga, ukuran data yang akan dilanjutkan pada tahapan selanjutnya berjumlah 50 baris dan 6 kolom.
2. Tidak terdapat nilai null.
3. Terdapat total jumlah cabang untuk setiap negara bagian yaitu, 17 New York (NW), 17 California (CA), dan 16 Florida (FL).

4. Tabel rangkuman keuangan dari setiap cabang dengan urutan dari terbesar adalah sebagai berikut:

Urutan	Total Pendapatan	Total Profit	Total Pengeluaran	Margin Bersih
1.	**Florida (9097447.83)**	**New York (1933859.6)**	**Florida (7197063.44)**	**California (22,02%)**
2.	New York (8786296.9)	Florida (1900384.4)	New York (6852437.3)	New York (22.00%)
3.	California (8021454.9)	California (1766387.98)	California (6255066.88)	Florida (20.88%)

dimana:

- Total Pengeluaran adalah gabungan seluruh pengeluaran yang telah dikeluarkan untuk setiap cabang, dengan rumus:
  Total Pengeluaran = AdvertisementSpend + PromotionSpend + AdministrationSpend

  Dari tabel di atas, dapat diketahui bahwa **CA memiliki pengeluaran paling sedikit** dibandingkan dengan yang lainnya. Artinya, CA dapat melakukan manajemen keuangan dari sisi pengeluaran dengan baik.

- Total profit adalah total sisa uang dari pendapatan dikurangi biaya.
  Profit = Total Pendapatan – Total Pengeluaran

  Dari tabel diatas, dapat diketahui bahwa **NY memiliki profit paling tinggi** dibandingkan dengan yang lainnya. Artinya, NY memiliki keuntungan yang paling tinggi dibandingkan dengan cabang yang. Bisa jadi, hal ini dikarenakan NY mendapatkan pendapatan yang baik dengan pengeluaran yang masih terawasi.

- Total pendapatan adalah total seluruh pendapatan yang diperoleh ditambah dengan total pengeluaran.
  Total Pendapatan = Profit + Total Pengeluaran

  Dari tabel diatas, dapat diketahui bahwa **FA memiliki total pendapatan paling tinggi** dibandingkan dengan yang lainnya dan memiliki total pengeluaran lebih tinggi, sedangkan profitnya rendah. Artinya, total pendapatan dari FA didominasi oleh pengeluaran dibandingkan dengan profitnya.

- Margin laba bersih merupakan rasio yang digunakan untuk mengukur besarnya persentase laba bersih atas penjualan bersih. Rasio ini dihitung dengan membagi laba bersih terhadap penjualan bersih.

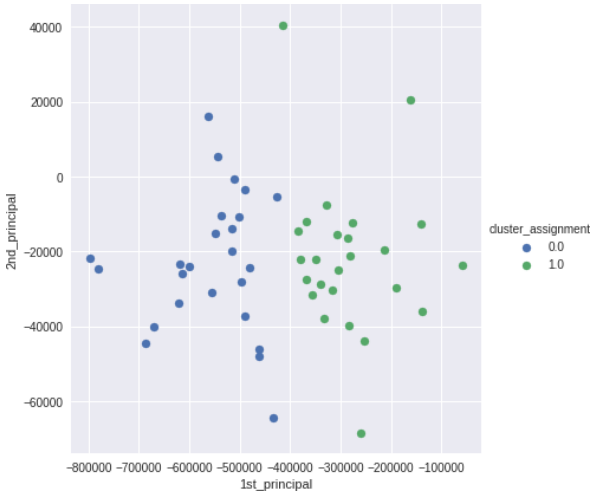
  Margin Laba Bersih $= \frac{Profit}{Total\ Pendapatan}$ **x 100%**

Dari tabel diatas, dapat diketahui bahwa **CA memiliki margin laba bersih paling tinggi** dibandingkan dengan yang lainnya. Artinya, CA memiliki keuntungan yang lebih tinggi karena, dapat menyimpan 2.202 untuk setiap 10.000 pendapatan penjualan. Rasio ini digunakan untuk memberi analis gambaran tentang stabilitas keuangan setiap cabangnya. Cabang yang menghasilkan keuntungan lebih besar per nilai dari penjualan berarti lebih efisien. Efisiensi itu membuat cabang lebih mungkin bertahan ketika lini produk tidak memenuhi harapan, atau ketika periode kontraksi ekonomi menghantam perekonomian yang lebih luas. Sebaliknya, rasio ini juga menunjukkan jumlah pendapatan yang hilang melalui biaya dan pengeluaran yang terkait dengan bisnis setiap cabangnya. Ini dapat membantu analis untuk mengetahui apakah sebuah bisnis harus fokus pada pengurangan pengeluaran atau tidak.

Kesimpulannya adalah, dari keempat metrics yang digunakan, CA adalah state cabang yang menghasilkan keuntungan yang lebih besar dan efisien. Hal ini dikarenakan **CA memiliki karakter state paling HEMAT** karena mampu mengatur keuangan untuk pengeluaran, meskipun memiliki profit paling kecil. Sedangkan, FA adalah state cabang yang menghasilkan keuntungan paling kecil dan tidak efisien. Hal ini dikarenakan **FA memiliki pengeluaran paling banyak, sedangkan profitnya tidak tergolong tinggi dan terindikasi berkarakter BOROS.** Oleh karena itu, diperlukan evaluasi yang lebih lanjut terhadap cabang-cabang yang berada di daerah negara bagian FA, mulai dari pengaturan dan penggantian strategi terhadap pengeluaran dan lebih fokus untuk menekankan pendapatan terlebih dahulu.

c. Task yang dikerjakan

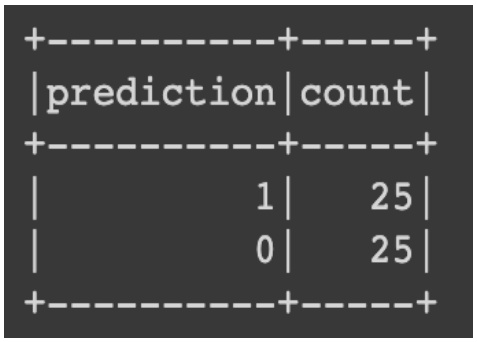
Dari hasil EDA yang sebelumnya telah dilakukan, maka penulis mencoba untuk melakukan pembagian kelompok menggunakan metode Clustering untuk mengetahui jenis-jenis kelompok pada setiap state berdasarkan perbandingan Profit dan Total Pengeluaran sebagai knowledge dan insight untuk mengambil Wisdom (keputusan). Pada eksperimen ini, penulis menggunakan 2 metode clustering yaitu berdasarkan jumlah *cluster* berdasarkan metode optimasi menggunakan Method Elbow dan saran penulis.



Secara garis besar, berikut adalah tahapan yang dilakukan dan *insight* yang didapatkan untuk membangun model c*lustering*:

1. Memilih data inputan kolom yaitu ‘Profit’ dan ‘TotalSpending’ yang akan dilakukan tahapan Vector Assembler. VectorAssember dari salah satu library dari Spark ML adalah modul yang memungkinkan konversi fitur numerik menjadi satu vektor yang digunakan oleh model ML.

2. Pada skema 1, dilakukan pemilihan jumlah cluster menggunakan Method Elbow (ME), dimana ME dapat menghasilkan jumlah cluster optimal berdasarkan nilai evaluasi Silhouette Score (SS) paling tinggi. **Pada skema 1 ini, didapatkan jumlah cluster terbaik adalah 2 dengan nilai SS 0.735.** Sehingga, pada tahapan visualisasi, akan muncul 4 warna yang berbeda. Berikut adalah visualisasi hasil pembangunan model **clustering untuk Skema 1** yang dilakukan.



```
+----------+-----+
|prediction|count|
+----------+-----+
|         1|   25|
|         0|   25|
+----------+-----+
```

Pada skema 1 ini, dapat dilihat bahwa penyebaran cluster terpisah dengan baik dan sempurna. Artinya, tidak ada himpunan anggota dari cluster lain yang melenceng dan memasuki area cluster lainnya. Dari hasil jumlah data setiap cluster, didapatkan bahwa cluster 0 dan 1 memiliki jumlah data yang sama banyak yaitu 25 data untuk masing-masing cluster.

Berikut adalah rangkuman karakteristik untuk setiap cluster yang didapatkan:

No. Cluster	Profit (Mean)	Profit (%)	Pengeluaran (Mean)	Pengeluaran (%)
0	142042.027	20.84%	482320.167	79.15%
1	81983.251	23.10%	272863.209	76.89%

Dari tabel di atas, dapat diambil kesimpulan bahwa:

- Cluster 0 memiliki nilai Profit > Pengeluaran, dengan rasio 20.84% profit dan 79.15% pengeluaran.
- Cluster 1 memiliki nilai Profit > Pengeluaran, dengan rasion 23.10% profit dan 76.89% pengeluaran.

Hal ini berarti, Cluster 0 memiliki pengeluaran yang lebih besar dan profit yang lebih kecil daripada Cluster 1. Artinya, Cluster 0 masih mengeluarkan biaya yang lebih besar dengan sisa keuangan yang lebih sedikit dibandingkan Cluster 1. Solusi yang diberikan oleh penulis adalah, Cluster 0 dapat melakukan tinjauan kembali terhadap distribusi pengeluaran yang dilakukan untuk masing-masing cabang pada setiap negara bagian yang termasuk dalam cluster ini. Dengan menggunakan konsep ekonomi yang paling sering disebut-sebutkan yaitu, mendapatkan keuntungan sebesar-besarnya dengan pengeluaran sekecil-kecilnya.

Berikut adalah rangkuman jumlah negara bagian untuk masing-masing cluster:

No. Cluster	FA	NW	CA
0	**9**	8	8
1	9	9	7
Total	18	17	15

Dari tabel di atas, dapat diambil kesimpulan bahwa:

1. Pada cluster 0 yang memiliki pengeluaran yang lebih besar sedangkan profit paling sedikit, cabang pada state FA (9 cabang) lebih mendominasi pada cluster 0.
2. Cabang pada state NW lebih mendominasi pada cluster 1 dengan karakter profit tinggi dan pengeluaran yang sedikit. Artinya, cabang ini masih tergolong sehat dan berjalan efektif.
3. Cabang pada state CA lebih mendominasi pada cluster 0 dengan karakter yang sama dengan FA.

Dari kesimpulan EDA dan Clustering yang dilakukan, maka solusi yang dapat dilakukan adalah:

1. Untuk cabang di FA, dibutuhkan evaluasi kinerja keuangan yang lebih mendetail lagi, mempertimbangkan kembali rasio pembagian untuk pengeluaran, dan memperbaiki strategi penjualan agar didapatkan profit yang lebih tinggi dibandingkan hanya memiliki nilai yang besar pada pengeluaran.
2. Untuk cabang di NW, disarankan untuk mempertahankan kinerja yang sekarang karena masih tergolong perusahaan dengan cabang yang sehat. Hal ini dibuktikan profit yang tinggi walaupun pengeluaran juga tergolong tinggi. Artinya, kedua parameter ini masih berjalan secara tegak lurus.
3. Untuk cabang di CA, disarankan juga untuk mempertahankan kinerja yang sekarang karena memiliki nilai laba bersih paling tinggi diantara yang lainnya. Selain itu, cabang ini juga memiliki pengeluaran yang kecil walaupun profit

juga tergolong masih sedikit. Artinya, cabang pada CA memiliki mawas diri yang baik terhadap manajemen *cash flow* cabang masing-masing.

2. Ads_CTR_Optimisation.csv

a. Informasi Dataset

Dataset ini berisi data optimasi iklan berdasarkan click through rate pada iklan digital yang dipasang di berbagai web dari 10.000 user pada 10 iklan yang berbeda.

b. EDA (Exploratory Data Analysis)

Berikut adalah wawasan tentang data yang didapatkan setelah dilakukannya proses EDA:

1. Dataset ini memiliki ukuran data sebesar 10000 baris yang merepresentasikan user dengan 10 kolom yang merepresentasikan jenis iklan yang berbeda pada platform website.
2. Tidak terdapat nilai null.
3. Value yang ditampung oleh dataset ini hanya terdiri dari 2 jenis representasi, yaitu 0 yang artinya ikan tersebut tidak dipilih atau diklik oleh user yang bersangkutan, sedangkan 1 yang artinya dipilih atau diklik.
4. Berikut adalah rangkuman jumlah *click rate* untuk masing-masing ad:

No. Ad	Not Click (0)	Click (1)
1	8297	1703
2	8705	1295
3	9272	728
4	8804	1196
5	**7305**	**2695**
6	**9874**	**126**
7	8888	1112
8	7909	2091
9	9048	952
10	9511	489
Total	**87613**	**12387**
Rasio	**7,0729797**	**0,1413831**

Pada tabel di atas, dapat diketahui bahwa:

- **Ad6 paling jarang dikunjungi diantara ad yang lain.** Hal ini dibuktikan dengan jumlah Not Click yang paling besar yaitu 9874 user yang tidak tertarik. Sedangkan, yang paling sedikit untuk jumlah Not Click didapatkan oleh ad5 dengan total 7305 user yang tidak tertarik.
- **Ad5 paling sering dikunjungi diantara ad yang lain.** Hal ini dibuktikan dengan jumlah click sebesar 2695. Sedangkan, jumlah click paling sedikit didapatkan oleh ad6.
- **Total Not Click > Click dengan rasio perbandingan adalah 7.07:0.14** Artinya, ketika kondisi tidak ada 1 orang pun yang mengunjungi iklan, malah terdapat 7 orang yang tidak mengunjungi iklan. Yang artinya, ketertarikan akan iklan yang dipampang di web memang masih tergolong sangat kecil. Hal ini dibuktikan dengan jumlah not click yang lebih besar daripada click.
- Jumlah Click masih jauh dari nilai yang seharusnya diharapkan yaitu 10 (iklan) x 10000 (user) = (100.000) click. Sedangkan, total click yang sekarang baru dicapai hanya berada pada nilai 12387.
- Masih dibutuhkan beberapa perbaikan yang harus dilakukan untuk menaikkan click rate yang memiliki selisih yang sangat besar yaitu 87613.
- **Ad terbaik didapatkan oleh Ad5 dan ad terburuk didapatkan oleh Ad6.**

Dari kesimpulan EDA yang dilakukan, maka solusi yang dapat dilakukan adalah:

1. Disarankan agar Ad5 untuk mempertahankan metode periklanan yang sedang dilakukan agar jumlah click rate yang didapatkan dapat bertahan, bahkan diharapkan untuk cenderung naik.
2. Disarankan agar Ad6 untuk mengganti metode periklanan yang dilakukan, seperti:
    - Beralih dari platform web lama dan mengganti platform web baru untuk bekerjasama dengan iklan yang akan dipasang.
    - Melakukan pembaharuan ide platform iklan seperti melalui sosial media yang sering digunakan oleh masyarakat, dibandingkan dengan hanya mengandalkan iklan via website.
    - Mengganti metode iklan yang sedang dilakukan, seperti desain iklan, user experience dari iklan yang dipasang, maupun kalimat atau kata yang persuasif dan menarik user.
    - Memasarkan Ad kepada kategori target yang sesuai dengan objek yang diluar dan latar belakang dari target.
    - Mengganti metode periklanan, misalnya dengan memasang baliho atau spanduk di tempat-tempat umum yang ramai dikunjungi oleh masyarakat dibandingkan dilakukan secara online melalui website.

3. **Market_Basket_optimization.csv**

a. **Informasi Dataset**

Dataset ini berisi data 7500 transaksi dalam waktu 1 minggu. Value yang ada pada dataset ini berisikan item yang dijual pada supermarket, dan setiap barisnya merepresentasikan 1 transaksi.

b. **EDA (Exploratory Data Analysis)**

- Terdapat banyak nilai Nan/null yang artinya item tersebut tidak dibeli pada transaksi yang sedang berjalan.
- Pada tahapan ini, dilakukan penjumlahan total setiap item untuk setiap transaksi. Sehingga, dapat diketahui bahwa terdapat 119 item yang dijual di setiap cabang Supermarket ABC untuk setiap Statenya.
- Selama seminggu waktu untuk transaksi jual beli, didapatkan 5 item yang paling sering dibeli dan 5 item yang paling jarang dibeli.
- Berikut adalah 5 data dengan item yang paling sering dibeli dalam kurus 1 minggu pada data:

No.	Nama Item	Total Item
1.	Mineral Water	1788.0
2.	Eggs	1348.0
3.	Spaghetti	1306.0
4.	French Fries	1282.0
5.	Chocolate	1230.0

Dari tabel di atas, dapat diketahui bahwa 5 item ini merupakan minuman, makanan, maupun bahan makanan sehari-hari dan berkemungkinan besar dibeli oleh pembeli dengan rasio sesering mungkin.

- Berikut adalah 5 data dengan item yang paling jarang dibeli dalam kurus 1 minggu pada data:

No.	Nama Item	Total Item
1.	Tea	29.0
2.	Bramble	14.0.0
3.	Cream	7.0
4.	Napkins	5.0
5.	Water Spray	3.0

Dari tabel di atas, dapat diketahui bahwa 5 item ini bukan merupakan minuman, makanan, maupun bahan makanan sehari-hari dan berkemungkinan besar tidak dibeli oleh pembeli dengan rasio sesering mungkin. Misalnya, seperti Bramble dan Napskins yang hanya beberapa kali dibeli oleh pembeli apabila jika membutuhkan saja.

Dari kesimpulan EDA yang dilakukan, maka solusi yang dapat dilakukan adalah:

- Tetap mempertahankan ketersediaan item-item yang paling sering dibeli oleh pembeli.
- Mencari ide alternatif yang lain untuk lebih menggiatkan promosi item-item yang jarang dibeli. Selain itu, solusi yang lain adalah dengan melakukan gaya dan pendekatan promosi yang berbeda dari sebelumnya. Hal ini diharapkan perputaran stok dalam supermarket dapat berjalan dengan teratur dan berkesinambungan.
- Mengatur tata letak *display* setiap itemnya sesuai dengan kategorinya masing-masing. Seperti, meletakkan bahan baku makanan menjadi 1 area yang sama, baik itu untuk item yang sering maupun jarang dibeli. Hal ini juga akan membantu item yang jarang dibeli untuk lebih dilirik oleh pembeli atau menjadi alternatif item belanja yang lainnya.
- Melakukan evaluasi keterbaharuan jenis produk, harga, maupun kualitas dari item-item yang dijual. Ada baiknya disarankan untuk menjual item-item dengan nama brand yang sudah terkenal dan sering diminati oleh pembeli.

4. Supermarket_CustomerMember.csv

a. Informasi Dataset

Dataset ini berisi data yang mencatat deskripsi dari pelanggan di supermarket ABC dengan ukuran 5 kolom dan 200 baris. Kolom tersebut terdiri dari:

- CustomerID = ID unik dari Customer
- Genre = Jenis Kelamin
- Age = Umur
- Annual Income (k$) = Pemasukan
- Spending Score (1-100 = Pengeluaran

b. EDA (Exploratory Data Analysis)

- Tidak ada nilai nan/null pada dataset.
- Berikut adalah informasi deskripsi terhadap customer;
    - Berdasarkan gendernya, terdiri atas 112 perempuan dan 88 laki-laki. Hal ini mengartikan bahwa mayoritas customer adalah perempuan.
    - Rata-rata umum customer yang berbelanja di Supermarket ABC adalah 39 tahun, dengan nilai terkecil adalah 18 tahun dan nilai terbesar adalah 70 tahun.

- Rata-rata pendapatan adalah 60.560000 dan pengeluaran adalah 50.200000. Artinya, mayoritas dari customer memiliki pendapatan yang lebih besar daripada pengeluarannya.

c. Data Pre-Processing

- One Hot Encoder

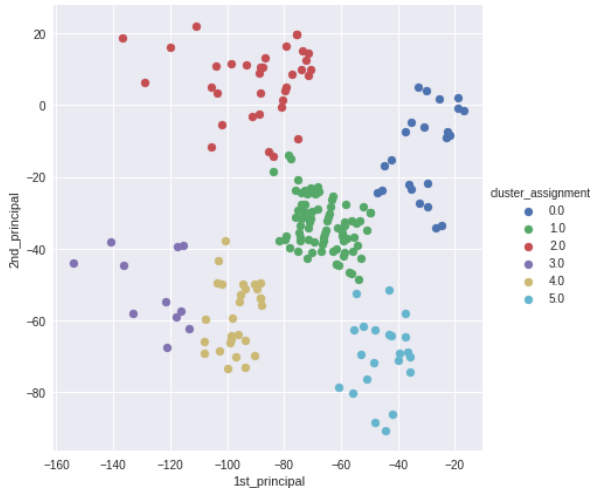
Pada tahapan ini, dilakukan perubahan konversi representasi data terhadap kolom umur (Age) yang awalnya memiliki nilai kategorikal menjadi nilai numerikal dengan *rules* sebagai berikut:

- Umur < 25 = Gen Z
- Umur 25 - 40 = Gen Y
- Umur 41 - 56 = Gen X
- Umur > 56 = Baby Boomers
- Missing Value = No Definition

Untuk setiap nilainya, akan dienkripsi dengan nilai (0-4).

- Vector Assembler for Clustering

Memilih data inputan kolom yaitu ‘Annual Income (k$)’, ‘Spending Score (1-100)’, dan ‘Genre’ yang akan dilakukan tahapan Vector Assembler. VectorAssember dari salah satu library dari Spark ML adalah modul yang memungkinkan konversi fitur numerik menjadi satu vektor yang digunakan oleh model ML. Hal ini berarti, akan dibangun sebuah pemodelan clustering dengan 3 data inputan tersebut.



d. Clustering

Dari hasil EDA yang sebelumnya telah dilakukan, maka penulis mencoba untuk melakukan pembagian kelompok menggunakan metode Clustering untuk mengetahui jenis-jenis kelompok pada setiap state berdasarkan perbandingan ‘Annual Income (k$)’, ‘Spending Score (1-100)’, ‘Age’, dan ‘Gender’ sebagai knowledge dan insight untuk mengambil Wisdom (keputusan). Pada eksperimen ini, penulis menggunakan metode clustering berdasarkan jumlah cluster berdasarkan metode optimasi menggunakan Method Elbow.

Secara garis besar, berikut adalah tahapan yang dilakukan dan *insight* yang didapatkan untuk membangun model c*lustering*:

1. Memilih data inputan kolom yaitu ‘Annual Income (k$)’, ‘Spending Score (1-100)’, ‘Age’, dan ‘Genre’ yang akan dilakukan tahapan Vector Assembler.
2. Pada skema ini, dilakukan pemilihan jumlah cluster menggunakan Method Elbow (ME), dimana ME dapat menghasilkan jumlah cluster optimal berdasarkan nilai evaluasi Silhouette Score (SS) paling tinggi.

   Pada skema ini ini, didapatkan jumlah cluster terbaik adalah 6 dengan nilai SS 0.7273. Sehingga, pada tahapan visualisasi, akan muncul 6 warna yang berbeda.

   Pada skema ini, dapat dilihat bahwa penyebaran cluster terpisah dengan baik dan sempurna. Artinya, tidak ada himpunan anggota dari cluster lain yang melenceng dan memasuki area cluster lainnya. Dari hasil jumlah data setiap cluster, didapatkan total data per-cluster sebagai berikut dengan nilai terbanyak didapatkan oleh cluster 1 (total: 81 data):

```
+----------+-----+
|prediction|count|
+----------+-----+
|         1|   81|
|         3|   11|
|         5|   22|
|         4|   28|
|         2|   35|
|         0|   23|
+----------+-----+
```

Berikut adalah rangkuman karakteristik untuk setiap cluster yang didapatkan:

No. Cluster	Annual Income	Spending Score	Age	Gender
0	26.304	20.913	45	0 (Perempuan)
1	55.296	49.518	43	0 (Perempuan)
2	88.2	17.114	41	1 (Laki-laki)
3	108.18	82.727	32	0 (Perempuan)

4	78.035	81.892	33	0 (Perempuan)
5	25.727	79.363	25	0 (Perempuan)

Berikut adalah rangkuman jumlah negara bagian untuk masing-masing cluster:

No. Cluster	Laki-laki	Perempuan
0	**9**	14
1	33	48
2	16	19
3	5	6
4	13	15
5	9	13

Dari tabel di atas, dapat diambil kesimpulan bahwa:

- Cluster 0 memiliki pemasukan > pengeluaran, gender perempuan, dengan umur diantara 45 tahun, selisih tabungan sekitar 6000, dan terindikasi kelompok ibu rumah tangga yang memiliki pemasukan tambahan dari bekerja ataupun berwirausaha.
- Cluster 1 memiliki pemasukan > pengeluaran, gender perempuan, dengan umur diantara 43 tahun, selisih tabuhan sekitar 6000, dan terindikasi kelompok Ibu rumah tangga dengan pemasukan yang besar namun disertai dengan pengeluaran yang besar juga.
- Cluster 2 memiliki pemasukan > pengeluaran, gender laki-laki, dengan umur diantara 41 tahun, namun sisa tabungan sangat banyak. Cluster ini terindikasi termasuk golongan Bapak-bapak umur 40 tahun-an yang hemat.
- Cluster 3 memiliki pemasukan > pengeluaran, gender perempuan, dengan umur 32 tahun (termasuk gen Y). Hal ini terindikasi bahwa kelompok ini adalah Gen Y yang berprinsip dan mampu mengelola keuangannya dengan baik.
- Cluster 4 memiliki pemasukan < pengeluaran, gender perempuan, dengan umur 33 tahun, dan terindikasi boros karena pengeluaran lebih besar daripada pemasukannya walaupun dengan selisih yang tidak jauh berbeda.
- Cluster 5 memiliki pemasukan < pengeluaran, gender perempuan, dengan umur 25 tahun (Gen Z), dan terindikasi boros karena pengeluaran sangat besar sedangkan penghasilan sangat kecil.

Dari kesimpulan pemodelan Clustering yang sudah dibangun, maka solusi yang dapat dilakukan adalah:

- Mempertimbangkan metode pemasaran yang sesuai dengan target pasar berdasarkan kategori setiap clusternya, baik itu dari sisi pemasukan, pengeluaran, umur, maupun gender. Misalnya, apabila target pasarnya adalah Cluster 2 (Bapak-bapak), maka metode iklan yang dapat digunakan adalah dengan metode pendekatan berbasis konten yang sedang viral dengan menawarkan kebutuhan bapak-bapak sehari-hari.
- Memberikan saran kepada cluster yang pengeluarannya lebih besar daripada pemasukannya untuk lebih giat dan telaten dalam melakukan manajemen keuangan dan pengeluaran agar mendapatkan keadaan finansial yang stabil.

e. Regresi

Secara garis besar, berikut adalah tahapan yang dilakukan dan *insight* yang didapatkan untuk membangun model regresi:

- Regresi adalah suatu metode analisis yang biasa digunakan untuk melihat pengaruh antara dua atau banyak variabel. Umumnya, analisis regresi digunakan untuk melakukan prediksi atau ramalan.
- Memilih data inputan kolom yaitu ‘Annual Income (k$)', 'Spending Score (1-100)', dan 'Genre_ohe' dengan target outputnya adalah ‘Age’.
- Terdapat 3 algoritma regresi yang akan digunakan, yaitu Linear Regression, DecisIon Tree, dan Gradient-boosted Tree Regression untuk meninjau pengaruh fitur inputan terhadap prediksi.
- Dilakukan pembagian dataset dengan rasio 70 untuk data latih dan 30 untuk data uji.
- Kemudian, dilakukan pengujian untuk setiap algoritma yang sudah disebutkan sebelumnya yang selanjutnya akan dilakukan evaluasi terhadap performa algoritma masing-masing.
- Metric evaluasi yang digunakan adalah Root Mean Squared Error (RMSE). RMSE merupakan salah satu cara untuk mengevaluasi model regresi linear dengan mengukur tingkat akurasi hasil perkiraan suatu model. RMSE dihitung dengan mengkuadratkan error dibagi dengan jumlah data, lalu diakarkan. Metode estimasi yang mempunyai RMSE lebih kecil dikatakan lebih akurat daripada metode estimasi yang mempunyai RMSE lebih besar. Dengan rumus sebagai berikut:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}}$$

- Berikut adalah hasil evaluasi RMSE untuk setiap algoritma yang digunakan sebagai uji coba:

No.	Algoritma	Nilai RMSE
1.	Linear Regression	1.0047258773637804
2.	DecisIon Tree	1.2338008234635234
3.	Gradient-boosted Tree Regression	1.1285804415116063

Dari ketiga algoritma yang digunakan, dapat diketahui bahwa algoritma Linear Regression mendapatkan nilai RMSE paling baik diantara algoritma yang lainnya. Sebab nilainya semakin dekat dengan 0. Ketika kita tahu hubungan antara variabel independen dan dependen memiliki hubungan linier, algoritma ini adalah yang terbaik untuk digunakan karena ini adalah yang paling kompleks dibandingkan dengan algoritma lain yang juga mencoba menemukan hubungan antara variabel independen dan dependen.

Dari kesimpulan pemodelan Regresi yang sudah dibangun, maka solusi yang dapat dilakukan adalah:

- Model yang dibangun dapat digunakan untuk memprediksi umur seorang customer berdasarkan pemasukan, pengeluaran, dan umur. Sehingga, target pemasaran yang akan dilakukan, beserta dengan implementasi dapat tepat sasaran dan memberikan probabilitas kenaikan penjualan dan keuntungan masing-masing cabang di setiap state/negara bagian.
- Dengan melakukan metode periklanan yang tepat, maka diharapkan dapat muncul ide-ide baru terkait pembaharuan metode pendekatan kepada customer. Seperti, memberikan hak, bonus, diskon, perlakuan khusus, dan lain-lain berdasarkan dari deskripsi masing-masing customer.