Ananda Fitri Karimah
2301212013
Big Data Analysis

Assesment 3
Big Data Analysis

Link Code Google Colab :
https://colab.research.google.com/drive/1EB5x4mhX3OaaYNQmidg44E-FPdCJIqQI?usp=sharing

A. Masalah

Permasalahan yang akan dicoba selesaikan pada studi kasus ini adalah permasalahan yang dihadapi oleh Supermarket ABC dimana supermarket ini sudah beroperasi selama 10 tahun, namun beberapa tahun belakangan ini mengalami penurunan keuntungan.

Dikarenakan permasalahan ini Supermarket ABC ingin meningkatkan keuntungannya lagi seperti tahun-tahun sebelumnyanya dengan memanfaatkan beberapa aspek data yang dimiliki Supermarket ABC seperti jasa periklanan, item transaksi perminggu, dan lain-lain. Namun, hasil yang diperoleh belum signifikan sehingga akan diberikan saran aksi dan tindakan melalui tulisan ini.

B. Dataset dan Task

1. 50_SupermarketBranches.csv

a. Pengenalan Dataset

Dataset ini merupakan dataset yang berisikan lima features/kolom yaitu biaya iklan, biaya promosi, biaya administrasi, state, dan profit. Dimana dataini adalah data penggunaan biaya iklan, biaya promosi, biaya administrasi di cabang yang terdaftar pada dataset beserta profit yang didapatkan dalam kurun waktu tertentu.

b. Task

i. Data Preprocessing

- Perhitungan jumlah kolom dan jumlah data dan didapatkan dataset ini memiliki jumlah data sebanyak 50 row dan kolom/features sebanyak 5.
- Pengecekan nilai null dengan hasil bahwa tidak ada nilai null terdapat didalam dataset ini.

ii. EDA (Exploratory Data Analysis)

Pada dataset ini dilakukan EDA untuk mendapatkan insight pada dataset. Beberapa diantaranya adalah sebagai berikut.

- Dilakukan perhitungan jumlah masing-masing value unique yang terdapat didalam kolom/features state. Pada langkah ini didapatkan bahwa tercatat ada 17 cabang supermarket kota New york dan California, dan 16 cabang supermarket kota Florida tercatat dalam data penggunaan biaya dan profit ini.
- Dilakukan perhitungan kolom/feature total pengeluaran (nama variabel tot_spending pada code) dengan menambahkan nilai pada kolom/feature biaya iklan, biaya promosi, biaya administrasi disetiap barisnya.
- Dilakukan perhitungan dan penambahan feature/kolom total pendapatan (nama variabel adalah omzet pada code) pada setiap cabang dengan menggunakan perhitungan dibawah ini.

Total Pendapatan = Profit + Total Pengeluaran

Ananda Fitri Karimah
2301212013
Big Data Analysis

Dengan asumsi bahwa total pengeluaran yang dilakukan hanya mengacu pada dataset dan profit yang tertera pada dataset adalah profit bersih. Hasil yang didapatkan pada perhitungan ini bahwa total pendapatan paling besar didapatkan oleh cabang Florida dengan total pendapatan sebesar 9097447.83, disusul oleh cabang New York sebesar 8786296.919, dan terakhir adalah cabang California sebesar 8021454.859.

- Selanjutnya dilakukan penjumlahan untuk total profit untuk semua data pada setiap cabang. Didapatkan hasil bahwa profit paling besar didapatkan oleh cabang New York dengan nilai sebesar 1933859.589, diikuti oleh cabang Florida yaitu sebesar 1900384.39, dan yang terakhir adalah cabang California sebesar 1766387.98.
- Kolom/feature pengeluaran juga dilakukan penjumlahan untuk melihat besar total pengeluaran pada setiap cabang yang tercatat, dan dihasilkan bahwa total pengeluaran terbesar ada pada cabang Florida dengan total pengeluaran sebesar 7197063.44, kemudian total pengeluaran terbesar kedua adalah cabang New York dengan total sebesar 6852437.329, dan yang terakhir adalh cabang California dengan total pengeluaran sebesar 6255066.88.

```
Total Profit California :  1766387.98
Total Spending California :  6255066.88
Total Pendapatan Cabang California :  8021454.859999999
==========================================
Total Profit Florida :  1900384.39
Total Spending Florida :  7197063.44
Total Pendapatan Cabang Florida :  9097447.83
==========================================
Total Profit New York :  1933859.5899999996
Total Spending New York :  6852437.329999999
Total Pendapatan Cabang New York:  8786296.919999998
```

- Perhitungan lainnya yang dilakukan adalah perhitungan margin profit bersih dimana margin profit bersih adalah sebuah pengukuran untuk menghitung berapa banyak profit yang dihasilkan dari pendapatan. Perhitungan ini dilakukan untuk setiap cabang pada dataset yang dilakukan dengan menggunakan rumus dibawah ini.

  Margin Profit Bersih = (Profit/Total Pendapatan)*100

  Sehingga dari perhitungan ini didapatkan bahwa margin profit bersih paling besar didapatkan oleh cabang California dengan margin sebesar 22.02%, lalu kedua didapatkan oleh Cabang New York dengan nilai margin sebesar 22.00%, dan terakhir adalah Cabang Florida dengan nilai margin sebesar 20.88%.

```
Margin Profit yang didapatkan Cabang California :  22.020793120813 %
Margin Profit yang didapatkan Cabang Florida :  20.889203494338695 %
Margin Profit yang didapatkan Cabang New York :  22.00995035346472 %
```

Ananda Fitri Karimah
2301212013
Big Data Analysis

iii. Insight EDA

Didapatkan kumpulan insight pada dataset ini adalah sebagai berikut.

Cabang	Total Pendapatan	Total Profit	Total Pengeluaran	Margin Profit/Laba Bersih
California	8021454.859	1766387.98	6255066.88	**22.02%**
Florida	**9097447.83**	1900384.39	**7197063.44**	20.88%
New York	8786296.919	**1933859.589**	6852437.329	22.00%

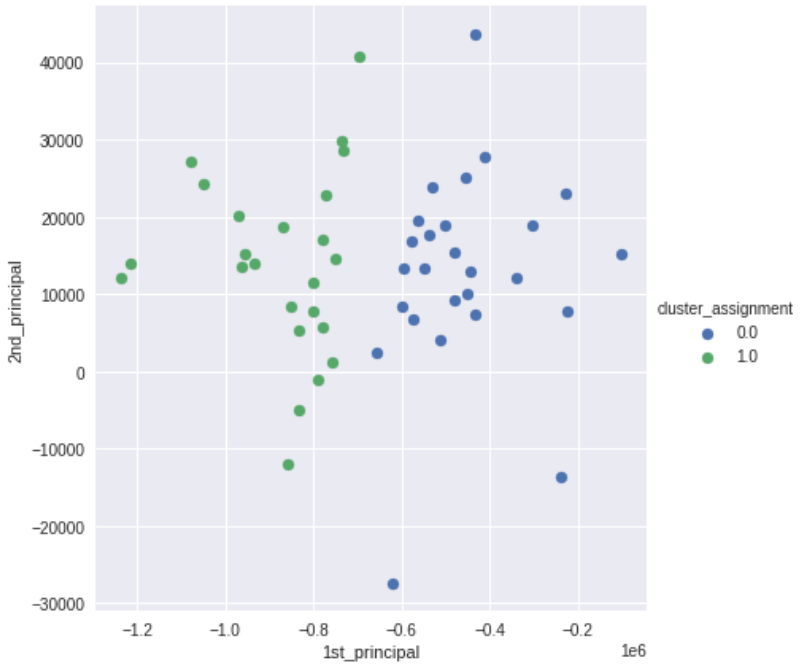
Sehingga dapat diambil kesimpulan bahwa:

- Cabang California adalah cabang yang paling bijaksana, dan efisien dalam memanfaatkan pengeluaran yang dapat dilihat pada total pengeluaran yang paling sedikit namun memiliki margin profit yang paling tinggi.
- Sedangkan Cabang Florida adalah cabang yang paling kurang bijaksana dan efisien dalam memanfaatkan pengeluaran yang juga dapat dilihat pada total pendapatan yang paling besar diantara ketiga cabang namun hanya memiliki margin profit paling kecil dibandingkan ketiga cabang lainnya.
- Terakhir Cabang New York adalah cabang yang seimbang dalam memanfaatkan biaya pengeluaran yang dapat dilihat pada total profit dan yang paling tinggi dan memiliki margin profit terbesar kedua setelah Cabang California.

iv. Clustering

Pada dataset ini juga dilakukan tahapan clustering dengan menggunakan algoritma KMeans untuk mengetahui karakteristik kelompok data yang terbentuk dengan menggunakan beberapa feature/kolom yang pada studi kasus ini digunakan feature/kolom total pengeluaran dan total pendapatan. Nilai K yang direkomendasikan oleh silhouette coefficient adalah K=2 (nilai silhouette coefficient = 0.77) dan K=4 (nilai sihouette coefficient = 0.77), dipilih untuk digunakan nilai K=2 (nilai K=2 dipilih berdasarkan eksperimen yang dilakukan pada kedua nilai K, yang hasilnya adalah keduanya memiliki pengelompokan yang sama-sama terpisah antara kedua feature/kolom yang terpilih tersebut. Sehingga dipilihlah nilai K=2 dengan alasan bahwa nilai K=2 adalah nilai K yang direkomendasikan terlebih dahulu oleh silhouette coefficient) dan didapatkan 2 buah clustering yaitu Cluster 0 dan Cluster 1 yang terbentuk yang dapat dilihat pada berikut ini.

Ananda Fitri Karimah
2301212013
Big Data Analysis



Jumlah masing-masing kelompok clustering yang terbentuk memiliki jumlah 26 cabang dimana didalamnya terdapat 10 cabang California, 9 cabang New York, dan cabang Florida untuk Cluster 0 dan 24 buah dimana ada 9 cabang Florida, 8 cabang New York, dan 7 Cabang California untuk cluster 1

```
+----------+-----+
|prediction|count|
+----------+-----+
|         1|   24|
|         0|   26|
+----------+-----+
```

```
df_super_clust_pandas_0['State'].value_counts()
```

```
California    10
New York       9
Florida        7
Name: State, dtype: int64
```

```
df_super_clust_pandas_1['State'].value_counts()
```

```
Florida       9
New York      8
California    7
Name: State, dtype: int64
```

Ananda Fitri Karimah
2301212013
Big Data Analysis

Karakteristik yang terbentuk dimasing-masing Cluster 0 yaitu didapatkan rata-rata total pengeluaran sebesar 278326.841 dan rata-rata omzet yaitu sebesar 360876.614 , dan pada Cluster 1 didapatkan rata-rata total pengeluaran sebesar 360876.614 dan rata-rata omzet pada Cluster 1 sebesar 688433.651

```
Cluster ke-0
Rata-rata Total Spending  :  278326.84192307695
Rata-rata Omzet :  360876.61423076916
Cluster ke-1
Rata-rata Total Spending  :  544502.9066666666
Rata-rata Omzet :  688433.6516666667
```

v. Insight Clustering

Dari proses clustering yang dilakukan didapatkan beberapa insight yang dapat dilihat sebagai berikut.

- Setelah dilakukan Clustering dilakukan dengan menggunakan nilai K = 2 didapatkan dua cluster yaitu Cluster 0 dan Cluster 1 dimana jumlah masing masing kelas terdapat 26 cabang dan 24 cabang secara berturut-turut.
- Didalam Cluster 0 terdapat 10 cabang California, 9 cabang New York, dan cabang Florida, sedangkan didalam Cluster 1 terdapat 9 cabang Florida, 8 cabang New York, dan 7 Cabang California.
- Hasil cluster yang terbentuk dengan nilai K=2 mendapatkan pengelompokan yang terpisah yang dapat dilihat dari karakteristik masing-masing feature/kolom yang terpilih.

Cluster	Rata-rata Total Pengeluaran	Rata-rata Total Pendapatan
Cluster 0	278326.841	360876.614
Cluster 1	544502.906	688433.651

Dari hasil pada tabel diatas, didapatkan karakteristik bahwa.

- Cluster 0 adalah Cluster yang memiliki rata-rata total pengeluaran dan rata-rata total pendapatan paling kecil
- Cluster 1 adalah cluster yang memiliki rata-rata total pengeluaran dan rata-rata total pendapatan paling besar.

Yang mana jika dikaitkan dengan proses EDA sebelumnya dapat dicocokkan bahwa.

- Cabang Florida memiliki total pendapatan terbesar yang dapat dibuktikan dengan mendominasinya Cabang Florida yaitu sebanyak 9 cabang pada Cluster 1.
- Sedangkan Cabang California terbukti memang cabang yang memiliki total pendapatan paling kecil yang dibuktikan dengan mendominasinya Cabang California yaitu sebanyak 10 cabang.

Ananda Fitri Karimah
2301212013
Big Data Analysis

2. Ads_CTR_Optimisation.csv

a. Pengenalan Dataset

Dataset ini merupakan dataset yang memiliki value click through rate pada iklan digital dengan jumlah iklan sebanyak 10 iklan yang dipasang oleh supermarket pada platform web dari 10000 user yang telah mengklik iklan tersebut.

b. Task

i. Data Pre Processing

- Pengecekan pada dataset apakah terdapat nilai null/nan pada dataset, dan hasilnya adalah pada dataset ini tidak ada nilai yang null/nan
- Penghitungan jumlah row dan kolom/feature yang didapatkan terdapat sebanyak 10000 row yang merepresentasikan jumlah user dan 10 kolom/feature yang merepresentasikan ada 10 jenis iklan yang dipasang pada web.

ii. EDA (Exploratory Data Analysis)

Pada proses ini diketahui bahwa value yang terkandung pada dataset tersebut adalah nilai 0 dan 1. Nilai 0 merepresentasikan bahwa iklan tersebut tidak di klik dan 1 adalah iklan tersebut diklik 1 kali oleh user. Sehingga berikut adalah EDA yang dilakukan pada dataset ini.

- Mencari jumlah value 0 dan 1 di setiap kolom/feature ads. Pada tahapan ini seluruh kolom dilakukan pencarian untuk mengetahui pasti berapa kali sebuah ads diklik oleh user.

```
==============================
Ad 1
0    8297
1    1703
Name: Ad 1, dtype: int64
==============================
Ad 2
0    8705
1    1295
Name: Ad 2, dtype: int64
==============================
Ad 3
0    9272
1     728
Name: Ad 3, dtype: int64
==============================
Ad 4
0    8804
1    1196
Name: Ad 4, dtype: int64
==============================
Ad 5
0    7305
1    2695
Name: Ad 5, dtype: int64
```

```
==============================
Ad 6
0    9874
1     126
Name: Ad 6, dtype: int64
==============================
Ad 7
0    8888
1    1112
Name: Ad 7, dtype: int64
==============================
Ad 8
0    7909
1    2091
Name: Ad 8, dtype: int64
==============================
Ad 9
0    9048
1     952
Name: Ad 9, dtype: int64
==============================
Ad 10
0    9511
1     489
Name: Ad 10, dtype: int64
```

- Mendapatkan kolom/feature ads yang rata-rata click rate, paling sering diklik dan paling sedikit diklik, yang mana didapatkan bahwa iklan yang paling sedikit di klik pada web adalah kolom/feature Ads 6 dengan jumlah click sebesar 126 dan iklan yang paling banyak diklik adalah kolom/feature Ads 5 dengan jumlah click sebesar 2695, dan dataset ini memiliki rata-rata click rate sebesar 1238.7.

Ananda Fitri Karimah
2301212013
Big Data Analysis

```
print('Rata-rata Click rate : ', np.mean(arr_click_rate))
print('Max Click rate : ', np.max(arr_click_rate))
print('Min Click rate : ', np.min(arr_click_rate))
```

```
Rata-rata Click rate :  1238.7
Max Click rate :  2695
Min Click rate :  126
```

iii. Insight EDA

Pada proses EDA yang telah dilakukan diatas didapatkan bahwa.

- Minimal, Maksimal, dan Rata-rata Click Rate

Minimal Click Rate	Maksimal Click Rate	Rata-rata Click Rate
126 (Ads 6)	2695 (Ads 5)	1238.7

Sehingga didapatkan insight bahwa Ads6 dapat menjadi acuan/contoh bagi Ads lainnya agar bisa membuat iklan yang sama atau lebih menarik dari pada Ads6, sehingga iklan yang dipasang minimal bisa mendapatkan jumlah click rate yang sama dengan Ads5. Sedangkan Ads6 adalah iklan yang patut menjadi evaluasi karena nilai click ratenya yang rendah.

- Dapat diketahui bahwa nilai rata-rata click rate yang masih jauh dari nilai 10000 dimana artinya bahkan ada user yang bahkan tidak mengklik satu iklanpun yang dipasang di web. Ini menandakan bahwa kualitas iklan supermarket yang dipasang di web masih jauh dari kata menarik dan harus dievaluasi lebih dalam lagi agar iklan yang dipasang diweb bisa menjadi semenarik mungkin sehingga membuat user ingin mengunjungi iklan tersebut, atau bisa jadi iklan yang dipasang diweb tidak tepat sasaran sehingga iklan tersebut tidak diklik oleh user. Dapat juga mengganti platform web menjadi platform media sosial lainnya yang lebih ramai untuk menjangkau pelanggan lebih dekat lagi.

3. Market_Basket_optimization.csv

a. Pengenalan Dataset

Data Market Basket adalah dataset yang berisikan data transaksi sebanyak 7500 transaksi dalam seminggu. Dimana value dari dataset ini adalah sebanyak 7500 transaksi/row yang berisikan barang-barang supermarket yang dibeli oleh pelanggan.

b. Task

i. Data Preprocessing

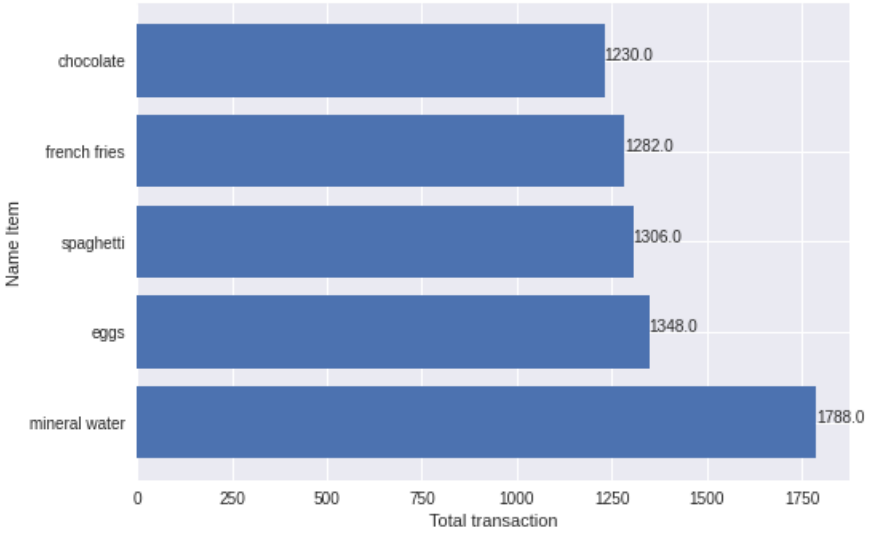
- Dilakukan penambahan header pada saat pembacaan data csv dengan nama col1 sampai dengan col20 dikarenakan tidak ada header pada dataset.
- Penghitungan jumlah row dan kolom/feature dengan jumlah sebanyak 7501(1 adalah header) dan jumlah/panjang item barang yang dibeli mencapai 20 kolom.
- Mengecek nan/null, yang ternyata ada didalam dataset yang semudian di replace sebagai -

Ananda Fitri Karimah
2301212013
Big Data Analysis

- Menghitung jumlah masing-masing item yang muncul di setiap kolom, yang kemudian dimasukkan pada dataframe lain agar lebih ringkas.

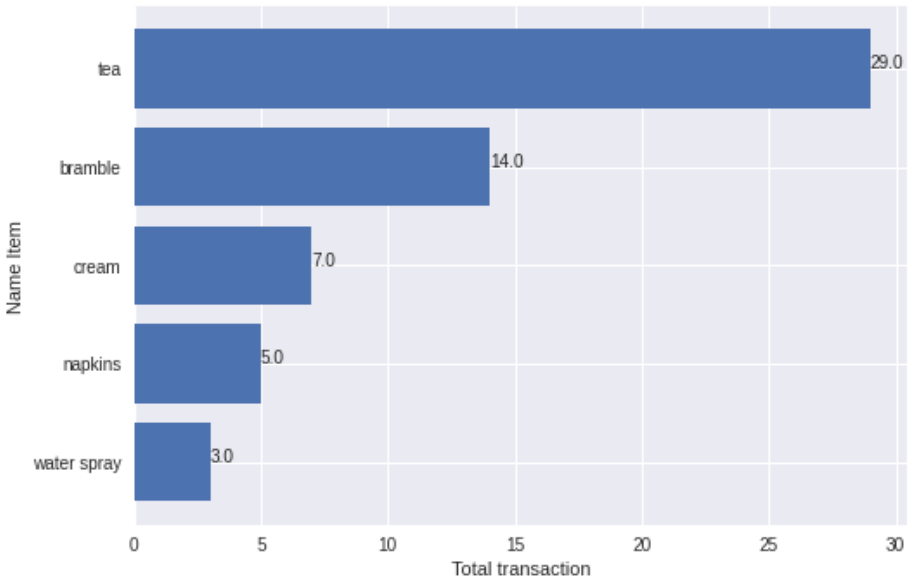
ii. EDA (Exploratory Data Analysis)

- Dilakukan pengecekan pada dataset untuk mengetahui jumlah item yang terbeli dalam satu dataset transaksi ini. Hasil yang didapatkan adalah sebanyak 119 item dengan total sebanyak 29363.0 terjual dalam satu minggu.
- Perhitungan total per-item yang terjual dalam seminggu, kemudian dilakukan pengurutan berdasarkan yang terbesar dan yang terkecil. Didapatkan hasil sebagai berikut.
  - Pengurutan 5 item dengan total item terjual tertinggi yang didapatkan oleh item mineral water, eggs, spagetti, french fries, dan chocolate yang terurut secara membesar ke terkecil



  - Pengurutan 5 item dengan total item terjual terendah yang didapatkan oleh item water spray, napkins, cream, bramble, dan tea yang terurut secara mengecil ke terbesar

Ananda Fitri Karimah
2301212013
Big Data Analysis



iii. Insight EDA

Hasil insight yang didapatkan dati proses EDA yang telah dilakukan adalah.

- Berdasarkan hasil EDA yang telah dilakukan didapatkan 5 item yang memiliki penjualan tertinggi dan terendah yang dapat dilihat pada tabel berikut.

Peringkat	5 item Penjualan Tertinggi (Top-5)	5 item Penjualan Terendah (Last-5)
1	Mineral Water (Total = 1788)	Water Spray (Total =3)
2	Eggs (Total = 1348)	Napkins (Total =5)
3	Spagetti (Total = 1308)	Cream (Total =7)
4	French Fries (Total = 1282)	Bramble (Total =14)
5	Chocolate (Total = 1230)	Tea (Total = 29)

- Dapat diketahui dari ringkasan tabel yang tertera diatas bahwa antara top-5 penjualan item dan last-5 memiliki gap atau selelisih yang cukup besar, sehingga bisa diambil tindakan dengan mengecek status dan kondisi dari ke-5 barang yang memiliki penjualan paling sedikit.
  - Apakah barang dengan penjualan paling sedikit ini memang paling sedikit penjualannya dikarenakan stock barang yang habis ?
  - Apakah dikarenakan kelangkaan pada barang tersebut ?
  - Apakah barang yang disediakan pada supermarket ABC ini kurang baik kualitasnya sehingga pelanggan enggan untuk membeli item tersebut ?

Ananda Fitri Karimah
2301212013
Big Data Analysis

- Apakah harga barang tersebut terlalu mahal sehingga pembeli enggan membeli barang tersebut ?
- Apakah tata letak barang tersebut terlalu tersembunyi/kurang strategis sehingga pelanggan supermarket abc tidak dapat menemukan barang tersebut ?

4. Supermarket_CustomerMember.csv

a. Pengenalan Dataset

Dataset ini merupakan dataset yang berisikan profil atau deskripsi dari pelanggan Supermarket ABC, dimana didalam dataset tersebut terdapat 5 kolom dengan jumlah data sebanyak 200 baris/row.

b. Task

i. Data Preprocessing

Beberapa tahapan prerocessing yang dilakukan pada dataset adallah sebagai berikut.

- Pengecekan data terhadap null/nan, yang didapatkan hasil bahwa didalam dataset ini tidak ada nilai nan/null sehingga tidak ada data yang didrop
- Pengecekan untuk melihat jumlah data dan baris daya yang didapatkan ada 5 kolom dengan jumlah data/kolom sebnayak 200

ii. EDA (Exploratory Data Analysis)

- Dilakukan pengecekan rasio pelanggan wanita dan laki laki dan didapatkan total pengunjung wanita sebanyak 112 dan laki-laki sebanyak 88 orang.
- Pengecekan pada distribusi dataset yang didapatkan hasil bahwa pada pelanggan Supermarket ABC memiliki rentang umur 18-70 tahun, dan rata-rata umur pelanggan yang berbelanja pada supermarket ABC ini adalah berumur 38 tahun, kemudian pada kolom/feature annual income berada pada rentang 15 sampai dengan 137 dan dengan nilai rata-rata 60.56, dan terakhir nilai spending score dengan rentang nilai 1 sampai dengan 99 dengan rata-rata sebesar 50.20. Berikut adalah visualisasi dari pengecekan distribusi data yang dilakukan.

```
df_member_pandas.describe()
```

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

Ananda Fitri Karimah
2301212013
Big Data Analysis

iii. Insight EDA

- Diketahui bahwa pelanggan Supermarket ABC didominasi oleh wanita
- Dapat diketahui juga bahwa rata-rata umur pelanggan Supermarket ABC adalah 38 tahun, sedangkan rata-rata annual income pelanggan Supermarket ABC adalah 137, dan rata-rata spending scorenya adalah 50.20.

iv. Clustering

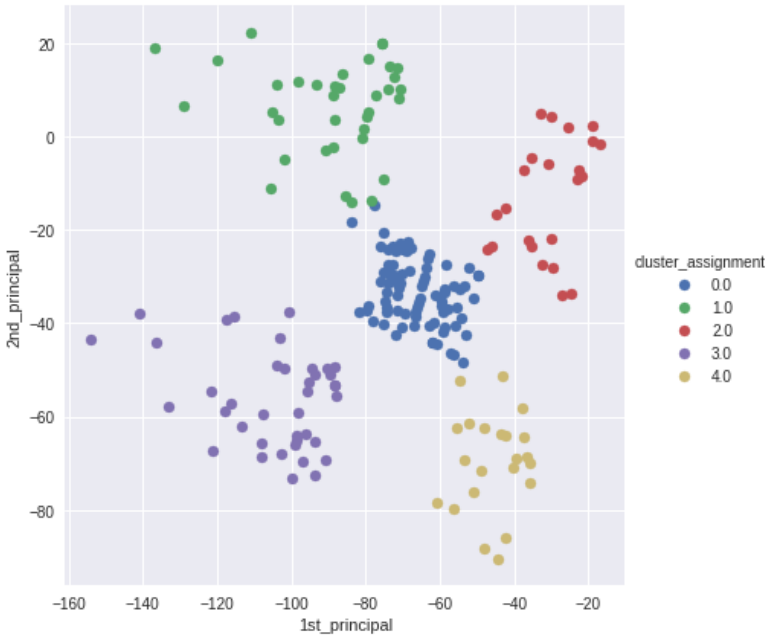
1. Data Preprocessing pada Clustering
   - Dilakukan diskritisasi pada kolom/feature age berikut adalah rule diskritisasi yang dibangun.
     - Umur < 25 termasuk kedalam kelompok “usia muda”
     - Umur >= 25 dan Umur <=34 termasuk kedalam kelompok “usia pekerja awal”
     - Umur >= 35 dan Umur <=44 termasuk kedalam kelompok “usia paruh baya”
     - Umur >= 45 dan Umur <=54 termasuk kedalam kelompok “usia pra-pensiun”
     - Umur >= 55 termasuk kedalam kelompok “usia pensiun”
     - Selain itu tidak terdefinisikan

   - Kemudian dilakukan onehotencoder untuk mengubah rule sebelumnya menjadi angka sehingga didapatkan
     - usia pekerja awal berganti menjadi nilai 0
     - usia paruh baya berganti menjadi nilai 1
     - usia pra-pensiun berganti menjadi nilai 2
     - usia muda berganti menjadi nilai 3
     - usia pensiun berganti menjadi nilai 4

   - Dilakukan onehotencoder juga untuk kolom/fitur genre yang mana value female pada dataset berubah menjadi 0 dan value male menjadi 1

2. Clustering
   - Akan dibangun Clustering pada dataset untuk mengetahui karakteristik pengelompokan yang terbentuk berdasarkan kolom atau fitur yang terpilih. Pada dataset dilakukan clustering untuk semua kolom/fitur yaitu annual income, spending score, age (yang sudah di onehotencoder), dan genre (yang sudah di onehotencoder)
   - Clustering yang dilakukan pada studi kasus ini menggunakan nilai K=5 sesuai dengan nilai silhouette score tertinggi pada nilai K=5 yaitu sebesar 0.74.
   - Didapatkan hasil 5 buah cluster yang cukup terpisah dengan nama Cluster 0, Cluster 1, Cluster 2, Cluster 3, dan Cluster 4 yang dapat dilihat dibawah ini

Ananda Fitri Karimah
2301212013
Big Data Analysis



- Masing-masing cluster memiliki jumlah Cluster 0 sebanyak 80 orang, Cluster 1 sebanyak 36 orang, Cluster 2 sebanyak 23 orang, Cluster 3 sebanyak 39 orang, dan Cluster 4 sebanyak 22 orang.

prediction	count
1	36
3	39
4	22
2	23
0	80

- Didapatkan karakteristik masing masing pengelompokan Cluster yaitu
  - Cluster 1 didominasi oleh laki-laki dan dengan rata-rata umur 40 tahunan memiliki annual income terbesar pertama (1), namun memiliki spending score paling kecil (5).
  - Cluster 3 didominasi oleh perempuan dengan rata-rata umur 32 tahunan memiliki annual income terbesar kedua (2), dan memiliki spending score terbesar pertama(1)
  - Cluster 0 didominasi oleh perempuan dengan rata-rata umur 42 tahun memiliki annual income terbesar ketiga (3), dengan spending score terbesar ketiga (3).
  - Cluster 2 didominasi oleh perempuan dengan rata-rata umur 45 tahun yang memiliki annual income dengan urutan ke empat(4), yang memiliki spending score terkecil kedua (4)
  - Cluster 3 didominasi oleh perempuan dengan rata-rata umur 25 tahun yang memiliki anniaan income terkecil (5), namun memiliki spending score terbesar kedua (2)

Ananda Fitri Karimah
2301212013
Big Data Analysis

```
Cluster ke-0
Rata-rata Annual Income (k$)  :  55.0875
Rata-rata Spending Score (1-100) :  49.7125
Rata-rata Age  :  42.9375
Modus (nilai yang paling banyak keluar) Gender :  0.0
Cluster ke-1
Rata-rata Annual Income (k$)  :  87.75
Rata-rata Spending Score (1-100) :  17.583333333333332
Rata-rata Age  :  40.666666666666664
Modus (nilai yang paling banyak keluar) Gender :  1.0
Cluster ke-2
Rata-rata Annual Income (k$)  :  26.304347826086957
Rata-rata Spending Score (1-100) :  20.91304347826087
Rata-rata Age  :  45.21739130434783
Modus (nilai yang paling banyak keluar) Gender :  0.0
Cluster ke-3
Rata-rata Annual Income (k$)  :  86.53846153846153
Rata-rata Spending Score (1-100) :  82.12820512820512
Rata-rata Age  :  32.69230769230769
Modus (nilai yang paling banyak keluar) Gender :  0.0
Cluster ke-4
Rata-rata Annual Income (k$)  :  25.727272727272727
Rata-rata Spending Score (1-100) :  79.36363636363636
Rata-rata Age  :  25.272727272727273
Modus (nilai yang paling banyak keluar) Gender :  0.0
```

v. Insight Clustering

Berikut adalah kumpulan insight berdasarkan tahapan clustering yang dilakukan.

- Telah dilakukan Clustering dilakukan dengan menggunakan nilai K = 5 yang mana masing-masing cluster mendapatkan jumlah Cluster 0 sebanyak 80 orang dengan rasio 47 perempuan: 33 laki-laki, Cluster 1 sebanyak 36 orang dengan rasio 19 perempuan: 17 laki-laki, Cluster 2 sebanyak 23 orang dengan rasio 14 perempuan: 7 laki-laki, Cluster 3 sebanyak 39 orang dengan rasio 21 perempuan: 18 laki-laki, dan Cluster 4 sebanyak 22 orang dengan rasio 13 perempuan: 9 laki-laki.

Ananda Fitri Karimah
2301212013
Big Data Analysis

```
df_member_clust_pandas_1['Genre'].value_counts()
```

```
Male      19
Female    17
Name: Genre, dtype: int64
```

```
df_member_clust_pandas_2['Genre'].value_counts()
```

```
Female    14
Male       9
Name: Genre, dtype: int64
```

```
df_member_clust_pandas_3['Genre'].value_counts()
```

```
Female    21
Male      18
Name: Genre, dtype: int64
```

```
df_member_clust_pandas_4['Genre'].value_counts()
```

```
Female    13
Male       9
Name: Genre, dtype: int64
```

- Hasil cluster yang terbentuk dengan nilai K=5 mendapatkan pengelompokan yang terpisah (yang dapat dibuktikan pada gambar clustering diatas), dan memiliki masing-masing karakter yang dapat diidentifikasi. Berikut adalah karakteristik masing-masing cluster.

Cluster	Rata-Rata Umur	Modus Gender	Rata-Rata Annual Income	Rata-Rata Spending Score
Cluster 0	42.93	Perempuan	55.08 (3)	49.71 (3)
Cluster 1	40.64	Laki-Laki	87.75 (1)	17.58 (5)
Cluster 2	45.21	Perempuan	26.30 (4)	20.91 (4)
Cluster 3	32.69	Perempuan	86.53 (2)	82.12 (1)
Cluster 4	25.2	Perempuan	25.72 (5)	79.36 (2)

Ananda Fitri Karimah
2301212013
Big Data Analysis

Dari hasil pada tabel diatas, dapat diidentifikasi dan diasumsikan bahwa.

- Cluster 0 merupakan kelas bagi perempuan yang sudah berumah tangga sekaligus sedang bekerja, hal ini dapat dilihat dari annual incomenya cukup banyak dan spending scorenya yang juga cukup banyak
- Cluster 1 adalah kelas bagi laki-laki paruh baya yang sudah berumur yang pelit, yang dapat dibuktikan dengan annual incomenya yang paling besar pertama, namun spending skor nya paling sedikit.
- Cluster 2 didominasi oleh perempuan yang dapat diidentifikasikan sebagai perempuan yang berumah tangga tanpa bekerja, yang dapat dibuktikan dengan annual incomenya yang termasuk terbesar keempat dan spending scorenya yang juga berbanding lurus.
- Cluster 3 adalah perempuan yang memiliki umur yang cukup matang namun belum menikah, yang dapat dibuktikan dengan tingginya annual income dan tingginya juga spending scorenya.
- Cluster 4 adalah perempuan yang masih berada di kepala dua, sedang merintis karirnya yang dapat dibuktikan dengan rendahnya annual income namun memiliki spending score yang lumayan berbanding terbalik.

vi. Regresi

- Fitur yang dipilih dalam pembangunan regresi ini adalah fitur spending score, asnnual income, dan genre ohe. Untuk target model regresi yang dibangun pada studi kasus ini adalah fitur Age.
- Algoritma yang dipilih untuk menjadi algoritma yang membangun model ini adalah algorima linear regresi dengan alasan untuk melihat pengaruh dua variable terhadap prediktor.
- Pada pembangunan model ini pembagian dataset dibagi menjadi 75% data train : 25% data test dengan alasan bahwa sudah dilakukan eksperimen dengan menggunakan pembagian data lainnya namun tidak memiliki hasil yang lebih baik dari pada pembagian dataset 75:25
- Didapatkan hasil prediksi regresi dengan menggunakan algoritma linear regresi.

feat_reg	Age_ohe	prediction
[3.0,19.0,1.0]	4.0	3.2538515310527742
[5.0,73.0,1.0]	3.0	2.316420059558435
[5.0,81.0,1.0]	3.0	2.1819132437294018
[6.0,16.0,0.0]	3.0	3.107199580380984
[9.0,71.0,1.0]	2.0	2.291025834218966
[13.0,87.0,1.0]	1.0	1.9629912732641726
[14.0,33.0,0.0]	2.0	2.7033307381508345
[16.0,120.0,0.0]	2.0	1.211058651361732
[17.0,34.0,0.0]	1.0	2.64225168919966
[17.0,103.0,0.0]	1.0	1.4821304026742466
[20.0,78.0,0.0]	1.0	1.8581985051674308
[20.0,86.0,1.0]	1.0	1.8765179989735294
[28.0,39.0,0.0]	2.0	2.3958773737405146
[29.0,70.0,0.0]	3.0	1.8599082300788286
[31.0,29.0,0.0]	1.0	2.519745196554261
[32.0,28.0,0.0]	2.0	2.521803316208709
[34.0,72.0,0.0]	0.0	1.7525053645006616
[35.0,24.0,1.0]	1.0	2.6976173367858127
[36.0,39.0,1.0]	2.0	2.430661824782193
[40.0,76.0,0.0]	0.0	1.5967205626410543

only showing top 20 rows

Ananda Fitri Karimah
2301212013
Big Data Analysis

- Untuk mengukur hasil evaluasi prediksi yang diberikan digunakan beberapa metrik pengukuran yaitu MAE yang mencari rata-rata selisih mutlak nilai sebenarnya (aktual) dengan nilai prediksi (prediksi) dengan rumus sebagai berikut.

$$MAE = \frac{1}{n}\sum_{1}^{n}\left|d_i - \hat{d}_i\right|$$

Metrik lainnya yang digunakan adalah MSE yang digunakan untuk mengecek estimasi berapa nilai kesalahan pada peramalan. Berikut adalah rumus MSE.

$$MSE = \frac{\sum_{t=1}^{n}(At - Ft)^2}{n}$$

Berikut adalah hasil metrik pada model regresi yang dibangun

```
print('MAE for train set:', pred_train.meanAbsoluteError)
print('MAE for test set:', pred.meanAbsoluteError)
```

```
MAE for train set: 1.1179612350515
MAE for test set: 0.9538799777316394
```

```
print('MSE for train set:', pred_train.meanSquaredError)
print('MSE for test set:', pred.meanSquaredError)
```

```
MSE for train set: 1.7591334088790367
MSE for test set: 1.237183540566773
```

Didapatkan hasil bahwa model regresi ini mendapatkan nilai 1.11 untuk data train dan 0.95 untuk data test pada matrik evaluasi MAE. Sedangkan pada matrik evaluasi MSE didapatkan nilai 1.75 pada data train dan 1.23 untuk data test.
MSE dan MAE memiliki rentang yang semakin mendekati nilai 0 maka semakin baik model tersebut dan apabila semakin mendekati nilai 1 maka model yang dibangun maka model tersebut semakin buruk. Pada model regresi yang dibangun didapatkan nilai metric yang masih belum optimal, sehingga kedepannya model ini harus terus diperbaiki untuk mendapatkan nilai metrik dan hasil prediksi yang lebih baik lagi.

vii. Insight Regresi

Pembangunan model regresi yang dibangun ini bisa digunakan untuk memprediksi umur pelanggan Supermarket ABC, namun model yang dibangun ini masih belum optimal sehingga di masa yang akan datang model ini masih harus terus di development untuk mendapatkan nilai terbaik.

Ananda Fitri Karimah
2301212013
Big Data Analysis

5. Saran dan Aksi yang Diusulkan Secara Keseluruhan

Beberapa Saran aksi yang diusulkan untuk supermarket ABC adalah.

a. Perbaikan pada manajemen keuangan dan biaya pengeluaran pada Cabang Florida agar biaya pengeluaran dan total pendapatan dapat berbanding lurus dengan margin profit dan profit yang dihasilkan.
b. Melakukan perbaikan pada Cabang California untuk menaikkan total pendapatan dan total profit yang berbanding lurus dengan margin profit bersih.
c. Melakukan reformasi besar-besaran terhadap iklan supermarket yang ditampilkan pada web dengan berbagai cara beberapa diantaranya adalah:
    i. Iklan ditampilkan di platform lain selain web seperti pada social media instagram, tiktok, dan lain-lain yang memiliki penonton/pengguna social media yang banyak
    ii. Melakukan riset dan evaluasi pada konten iklan yang sudah dibangun agar dapat menarik pelanggan lebih banyak lagi minimal dapat mencontoh seperti iklan yang dibangun pada Ads5 agar bisa mendapatkan click rate dengan minimal sama dengan Ads5.
    iii. Evaluasi lainnya yang dapat lakukan adalah dengan melakukan riset pasar/target customer kemudian menyesuaikan konten iklan dengan penempatan iklan pada social media (contoh iklan pada facebook disesuaikan dengan pengguna social media facebook)
d. Melakukan inspeksi pada item-item yang sedikit terjual untuk mengetahui titik permasalahannya. Apakah item tersebut tidak terjual karena kualitasnya yang buruk sehingga pelanggan enggan membeli barang tersebut, atau apakah barang tersebut sedikit terbeli dikarenakan kelangkaan barang tersebut, atau barang tersebut memang memiliki stok yang sedikit sehingga barang yang terjual memang hanya sisa barang yang ada, atau item tersebut terjual sedikit dikarenakan harganya yang terlalu tinggi atau tata letak barang tersebut yang terlalu tersembunyi sehingga pelanggan tidak dapat menemukan barang tersebut.
e. Melakukan inspeksi juga pada item-item yang paling banyak terjual untuk melihat berapa banyak barang yang habis terjual dikarenakan permintaan yang tinggi.
f. Dengan melakukan inspeksi pada kedua item yang paling sedikit terjual (d) dan item paling banyak terjual (e) maka supermarket ABC dapat
    i. Mengatur berapa banyak item yang harus di stok di gudang berdasarkan total penjualan
    ii. Mengatur ulang tata letak barang-barang pada supermarket ABC sehingga barang yang terjual terlalu sedikit ini bisa dijangkau oleh pelanggan.
    iii. Menginspeksi harga barang pada barang-barang yang terjual terlalu sedikit, jika terlalu tinggi harga dapat disesuaikan.
    iv. Memutuskan apakah item yang memiliki penjualan paling sedikit dihilangkan saja dari supermarket ABC atau tidak dengan manfaat memangkas biaya operasional, jika diperlukan.
g. Melakukan kegiatan promosi sesuai dengan penggolongan umur, gender, annual income, dan spending score yang sesuai dengan kelompok penggolongan

pelanggan yang sudah dibangun menggunakan clustering. Contoh promosi yang bisa dilakukan diantaranya adalah.

i. Memberikan promosi berupa potongan harga pada produk yang dibeli oleh pelanggan perempuan dengan umur sekitar 25 tahunan agar pelanggan ini semakin banyak membeli produk yang ada pada supermarket ABC sehingga meningkatkan spending score.
ii. Memberikan benefit kepada pelanggan yang memiliki spending score paling tinggi seperti memasukkan pelanggan ini kedalam keanggotaan pelanggan VIP.