**PROJECT AKHIR**

**Metode Density Based Scan dalam “Menganalisis Data Customer Segmentation di dalam Mall”**

Dibuat untuk Memenuhi Tugas Akhir Pertemuan 14

Mata Kuliah IS388 – Data Analyst



**Disusun oleh Kelompok 3:**

Christopher Darren 00000054804

Christian Alexander 00000054292

Michelle Stephanie 00000055477

Meisha Geovanni Mulin 00000055487

Azzahra Shaffira Wijaya 00000055376

**PROGRAM STUDI: SISTEM INFORMASI**

**FAKULTAS TEKNIK DAN INFORMATIKA**

**UNIVERSITAS MULTIMEDIA NUSANTARA**

**TANGERANG**

**2022**

# **Kata Pengantar**

Puji dan syukur kami panjatkan kepada Tuhan Yang Maha Esa karena berkat dan rahmat-Nya kami dapat menyelesaikan pembuatan laporan ini sebagai tugas akhir dan menjadi acuan untuk memperluas wawasan, pengetahuan, dan mengembangkan kemampuan kami dalam menganalisis sebuah data yang berguna bagi kami serta orang banyak.

Tujuan pembuatan laporan ini adalah untuk membantu memberikan informasi dan pengetahuan mengenai penerapan *Machine Learning* khususnya pada metode DBScan (metode untuk clustering). Proyek yang kami buat juga bertujuan untuk memberikan wawasan sederhana kepada para pembaca agar mampu menganalisis hal yang sederhana dalam kehidupan sehari hari.

Dalam menyelesaikan laporan ini penulis banyak mendapatkan dukungan, bimbingan, pengarahan, dan bantuan dari berbagai pihak. Pada kesempatan ini penulis ingin menyampaikan terima kasih kepada:

1. Ibu Ririn Ikana Desanti, S.Kom., M.Kom. selaku Ketua Program Studi Universitas Multimedia Nusantara.
2. Ahmad Faza, S.Kom. selaku Dosen Pengampu Mata Kuliah Data Analyst and sebagai pembimbing yang memberikan arahan dan bimbingan agar laporan ini bisa selesai sebagaimana mestinya.
3. Teman Teman jurusan Sistem Informasi yang memberikan dukungan dan semangat kepada kami sehingga laporan ini bisa selesai dengan baik.

Semoga laporan ini dapat bermanfaat untuk keperluan proyek penulis, pembaca, dan calon peneliti selanjutnya yang akan melakukan penelitian dalam bidang yang serupa. Kami juga meminta maaf sebesar besarnya apabila ada kesalahan dalam penulisan, pengejaan, sistemasi EYD dalam pembuatan laporan ini. Kami juga terbuka terhadap saran dan kritik sehingga bisa menjadi masukkan bagi kami dalam meningkatkan penulisan projek ini kedepannya agar lebih baik.

Tangerang, 20 Desember 2022

Penulis

# **DAFTAR ISI**

[Kata Pengantar](#_t5iko24ulcdn)…………………………………………………………………………………..i

[DAFTAR ISI](#_t87muc878e0h)…………………………………………………………………………………….ii

[BAB I](#_ywuw0679y90j)……………………………………………………………………………………………1

[1.1 Latar Belakang](#_y575hop851lq)…………………………………………………………………………...1

[1.3 Tujuan](#_89quuufp1nsu)……………………………………………………………………………………1

[BAB II](#_xa0m2lcq1p5)…………………………………………………………………………………………………….2

[2.1 Definisi Machine Learning](#_olm19ycaqazz)……………………………………………………………2-3

[2.2 Definisi DBScan](#_h1anrumx4f6a)………………………………………………………………………...3

[2.3 Metode Algoritma](#_ea7mgyb3wy3f)…………………………………………………………………….3-4

[BAB III](#_7t1tkpwidt6y)…………………………………………………………………………………………………5-7

[BAB IV](#_l2szuhz5h5wy)…………………………………………………………………………………………8

[4.1 Visualisasi(Distribusi, dkk, boxplot)](#_384bgrkfyba7)..............................................................................8-12

[4.1.1 Bivariate Analysis](#_jp4y0pqlkcag)……………………………………………………………………….12-13

[4.2 Pembahasan (DBSCAN)](#_hrbb2s6pfjw4)..............................................................................................14-17

[4.3 Kesimpulan](#_oo2gs3uf06kz)…………………………………………………………………………...17-18

[4.4 Saran](#_jm6w52m958d2)…………………………………………………………………………………….18

[DAFTAR PUSTAKA](#_oxlhneidkssk)……………………………………………………………………………19

# **BAB I**

**PENDAHULUAN**

## **1.1 Latar Belakang**

Manajemen dan pemeliharaan hubungan pelanggan selalu memainkan peran penting dalam memberikan intelijen bisnis kepada organisasi untuk membangun, mengelola, dan mengembangkan hubungan pelanggan jangka panjang yang berharga. Pentingnya memperlakukan pelanggan sebagai aset utama organisasi semakin meningkat nilainya di era sekarang ini. Organisasi - organisasi tertarik untuk berinvestasi dalam pengembangan strategi akuisisi, pemeliharaan, dan pengembangan pelanggan. Kecerdasan bisnis memiliki peran penting dalam memungkinkan perusahaan menggunakan keahlian teknis untuk lebih memahami pelanggan yang lebih baik dan Program untuk penjangkauan. Segmentasi pelanggan dilakukan menggunakan teknik *Clustering* seperti DBscan orang - orang yang memiliki minat yang sama dapat dikelompokkan. Dengan adanya segmentasi pelanggan, dapat membantu tim pemasaran untuk mengenali dan memaparkan berbagai segmen pelanggan yang memiliki pemikiran yang berbeda serta mengikuti strategi pembelian yang berbeda - beda.

**1.2 Rumusan Masalah**

* Bagaimanakah penerapan DBscan pada dataset mall customer ?

## **1.3 Tujuan**

* Tujuan utama melakukan segmentasi pelanggan adalah untuk melakukan pengelompokan terhadap orang-orang yang memiliki minat yang sama sehingga tim pemasaran dapat membentuk perencanaan pemasaran yang efektif.

# 

# **BAB II**

**LANDASAN TEORI**

## **2.1 Definisi Machine Learning**

*Machine Learning* adalah sebuah mesin pembelajaran yang dirancang untuk belajar secara otomatis tanpa instruksi pengguna. Pembelajaran mesin ini didasarkan pada matematika, statistik, penambangan data(mining), dan ilmu lainnya. Bergantung pada kebutuhan keahlian si data analyst sendiri.

Contoh penggunaan Machine Learning adalah:

1. Rekomendasi lagu di spotify berdasarkan kesukaan si pengguna
2. Sugesti untuk menambah teman yang sesuai dengan jaringan si pengguna(Instagram , Facebook, Twitter, dll)
3. Chat Bot yang sering ditemukan di dalam sebuah aplikasi forum kesehatan, toko jual beli online dan lain lain.
4. Security checkup pada sebuah aplikasi ataupun web based untuk segi keamanan

Metode Machine Learning dibagi menjadi 2 teknik yaitu:

* **Supervised Learning**

Supervised learning merupakan pembelajaran yang diawasi dilakukan dengan memberi nama kumpulan data yang digunakan oleh pembelajaran mesin dan mengklasifikasikannya oleh pengembang untuk memungkinkan algoritme mengenali tingkat akurasi kinerja. Dalam metode ini, pembelajaran mesin dipantau pada data berlabel, yang kemudian dipelajari oleh pembelajaran mesin tentang hubungan dan ketergantungan antar data.

* **Unsupervised Learning**

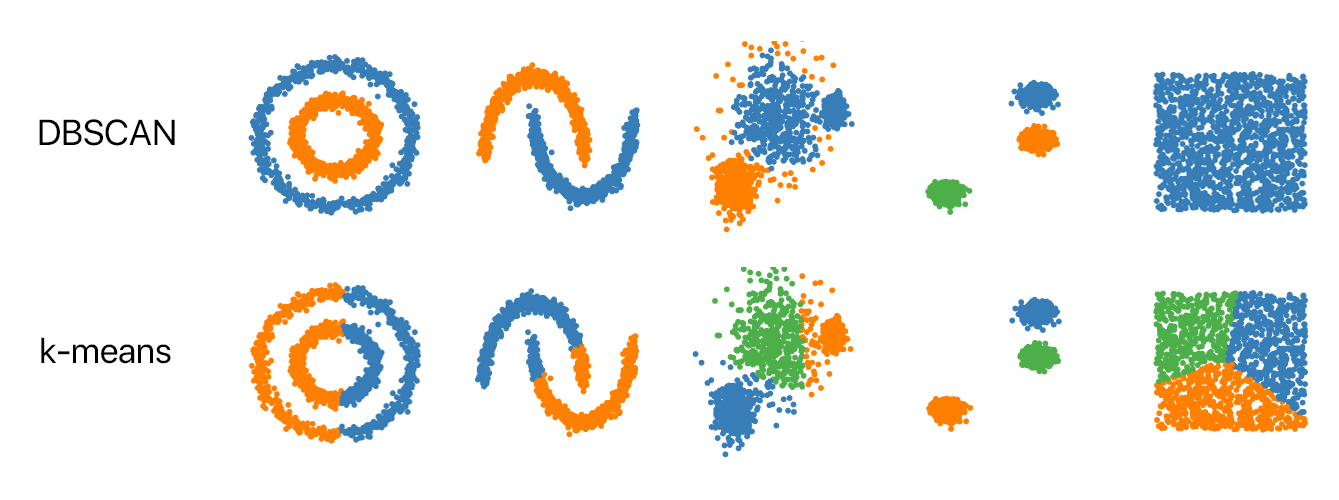
Unsupervised adalah teknik pembelajaran mesin tanpa pengawasan(security). Oleh karena itu, proses dilakukan pada kumpulan data mentah yang tidak berlabel, dan algoritma pembelajaran mesin mencoba mengidentifikasi pola dan hubungan antar data tanpa bantuan pemrogram.

Secara umum, unsupervised learning tidak memerlukan bantuan manusia, sehingga komputer benar-benar mempelajari data dan hubungannya sendiri. Dalam hal ini, kumpulan data tidak diberi nama dan mesin secara komputasi mengidentifikasi pola dalam data. Pembelajaran tanpa pengawasan digunakan untuk memudahkan pengembang membuat keputusan.

## **2.2 Definisi DBScan**

Density Based Scan adalah sebuah metode pengolahan data dengan cara *clustering* berbasis density atau *density based*. Metode clustering merupakan suatu proses yang bertujuan untuk membagi data menjadi kelompok-kelompok tertentu berdasarkan kategori atau tingkat kesamaan data. DBSCAN juga merupakan teknik ML dengan unsupervised machine learning.

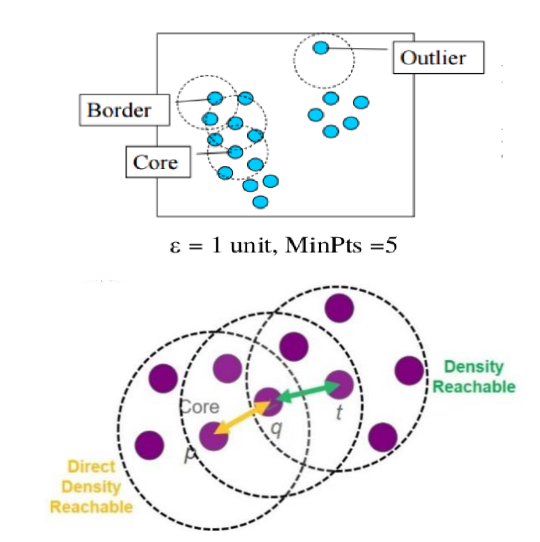
Metode ini sering digunakan untuk menghitung atau mengolah data yang membutuhkan pengelompokan data secara detail. Beberapa contoh data yang dikelola dengan data ini antara lain analisis *data sumber api* dan *analisis titik bencana pada suatu wilayah*.



## **2.3 Metode Algoritma**

Seperti yang sudah diketahui pada penjelasan di atas. Metode DBscan ini menggunakan algoritma density based. Dan ada beberapa parameter yang patut dipertimbangkan yaitu *epsilon* (eps) dan *minimum points*( minPts). Epsilon adalah jarak maksimal antara dua data dalam satu cluster. Sedangkan minimum points adalah banyak data minimal pada jarak epsilon agar terbentuk menjadi cluster. Berikut parameter lain yang harus diperhatikan dalam menggunakan metode Dbscan:

1. Directly density-reachable
2. Density -reachable
3. Core point
4. Border point
5. Outlier

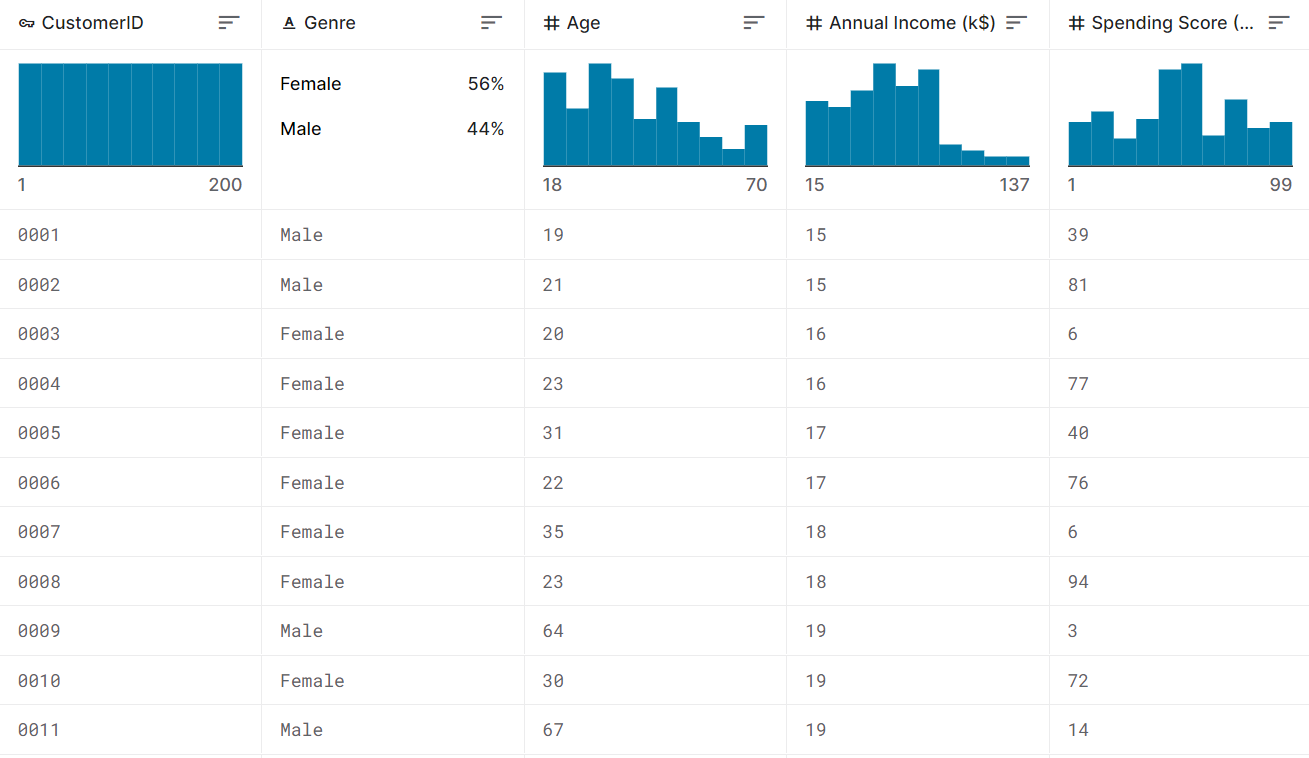


Cara kerja DBScan:

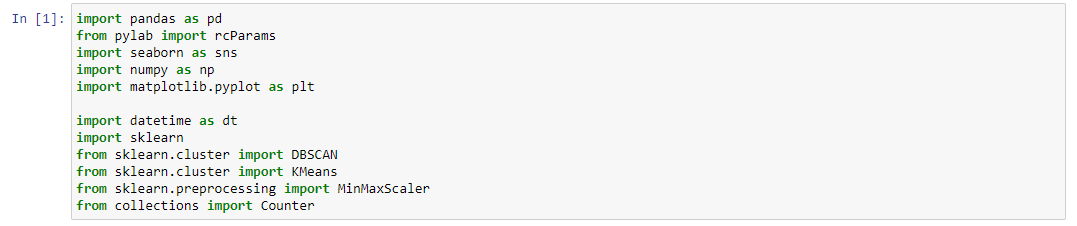
1. Tentukan nilai minimal points serta epsilon yang akan digunakan
2. Pilih data awal “p” secara random
3. Hitung jarak antara data “p” terhadap semua data menggunakan distance Euclidean.
4. Ambil semua pengamatan dengan density-reachable dengan amaatan “p”
5. Apabila pengamatan memenuhi nilai epsilon lebih dari jumlah minimal pengamatan dalam 1 group maka pengamatan “p” dikategorikan sebagai *core points*  dan group terbentuk
6. Apabila pengamatan “p” adalah border points dan tidak ada amatan yang densitasnya reachable dengan pengamatan “p”, maka bisa dilanjutkan ke pengamatan lainnya
7. Algoritma dbscan akan terus berulang sampai semua titik sudah dikunjungi,(titik q,t ,p)

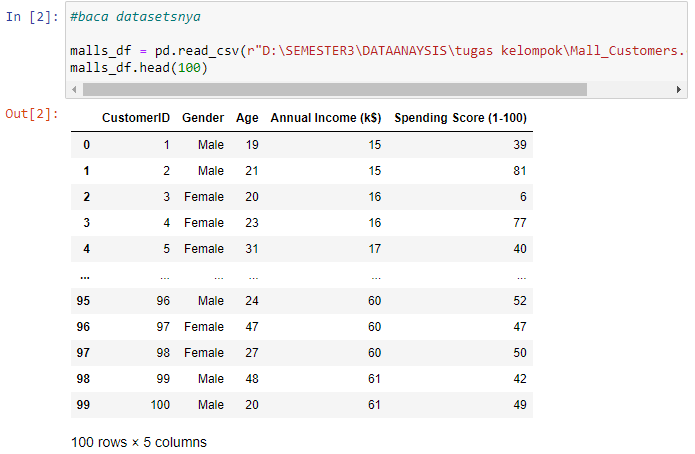
# **BAB III**

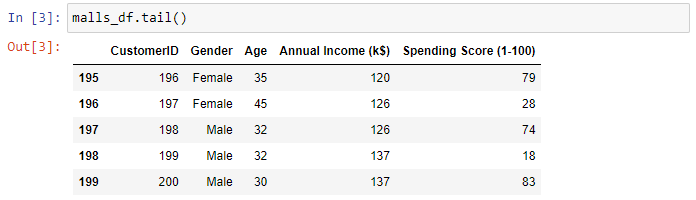
**PENGUMPULAN & PENGOLAHAN DATA**



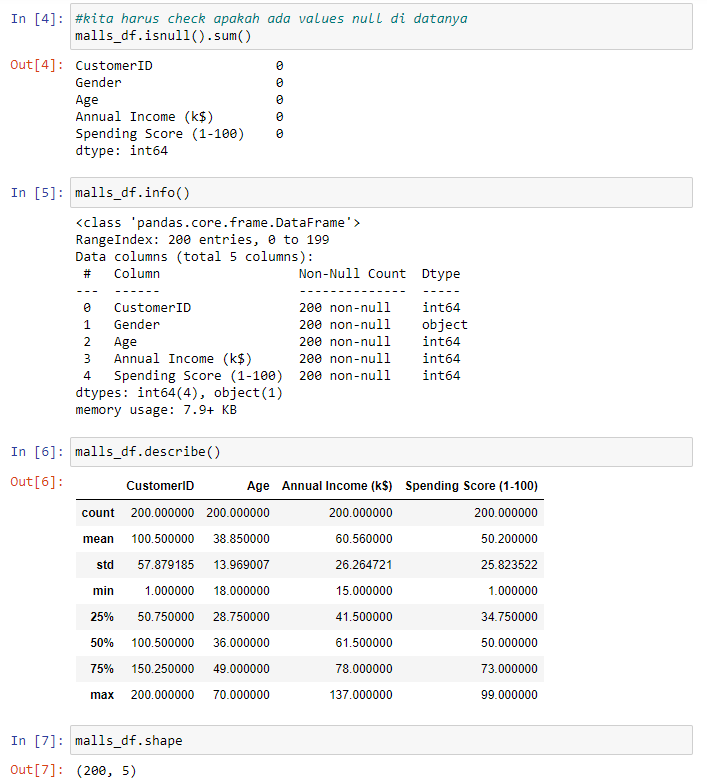
Data yang kami gunakan berasal dari website **kaggle.com** . Data mall customers yang terdiri dari 5 kolom dan 200 baris yang dibagi menjadi id customer (CustomerID) dari 0001 sampai 0200, jenis kelamin (Genre) Male dan Female, umur (Age) dengan range usia 18-70 tahun, pendapatan tahunan (Annual Income) dalam range 15-137$, dan skor pengeluaran (Spending score) dari 1 sampai 100.







Gambar diatas merupakan Code untuk Membaca dataset



Gambar diatas merupakan code yang digunakan untuk mencari data yang memiliki values yang null

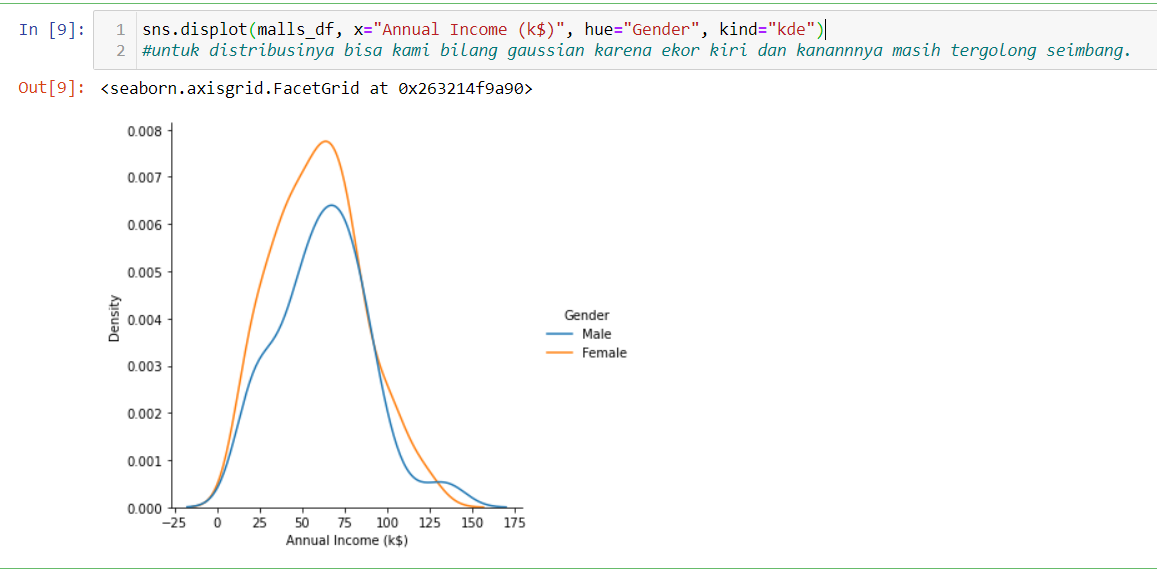
# **BAB IV**

**HASIL & PEMBAHASAN**

## **4.1 Visualisasi(Distribusi, dkk, boxplot)**

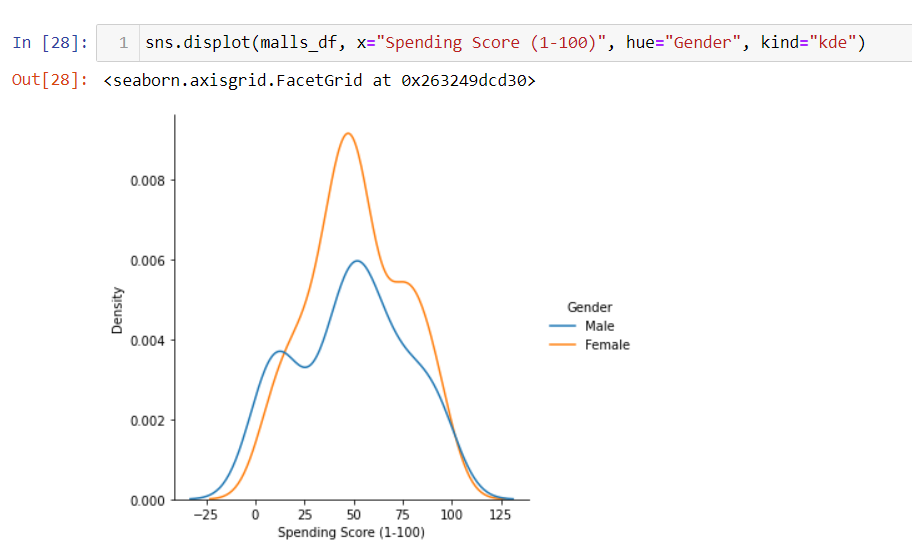
Gambar 4.1 Pairplot Mall Customer Data

Gambar diatas merupakan sns seaborn untuk merangkap semua data didalam 1 gambar penuh dimulai dari scatter plot hingga histogram.

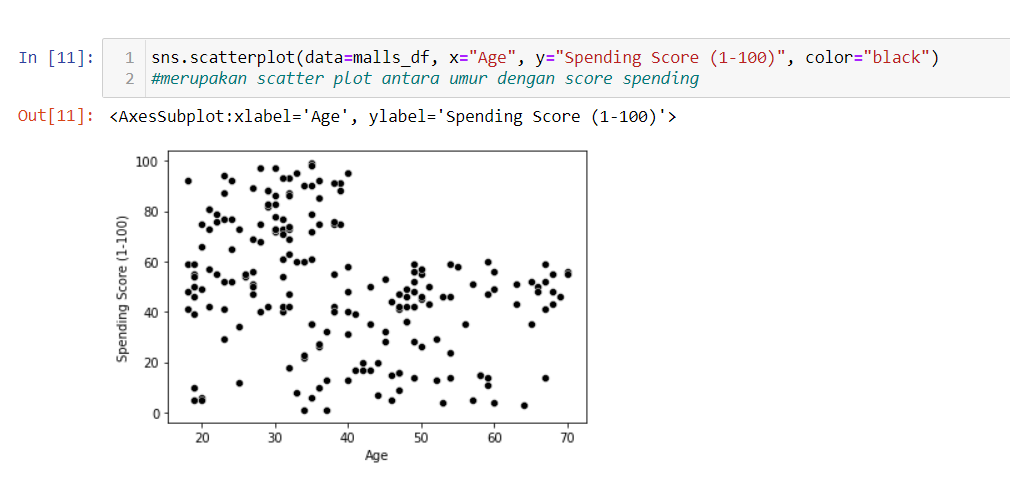


Gambar 4.2 Displot antara Annual Income dengan Gender

Distribusinya untuk annual income dengan gender bisa kita bilang agak gaussian(datanya normal) karena ekor berbentuk oval.

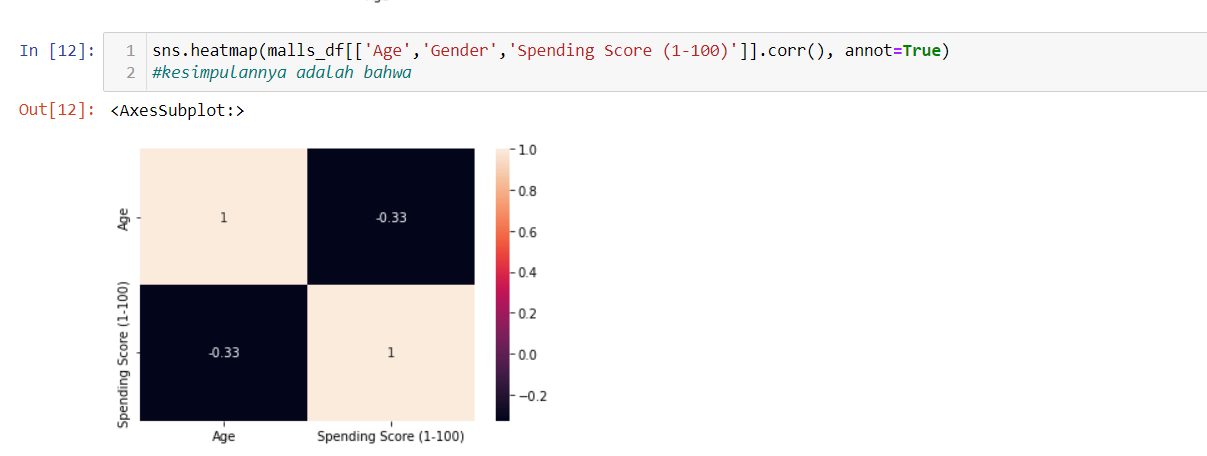


Gambar 4.3 Displot Annual Spending Score dengan Gender

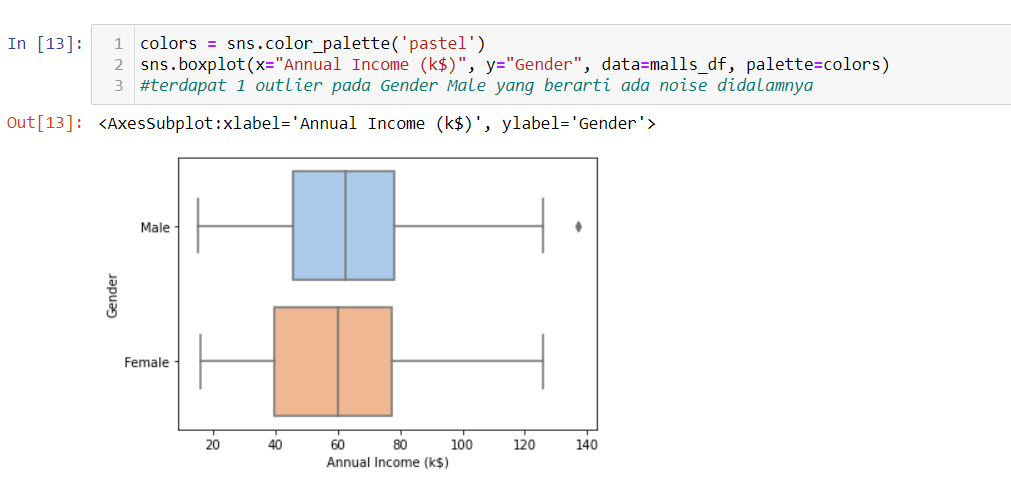


Gambar 4.4 Scatter plot antara *Age* dengan *Spending Score*

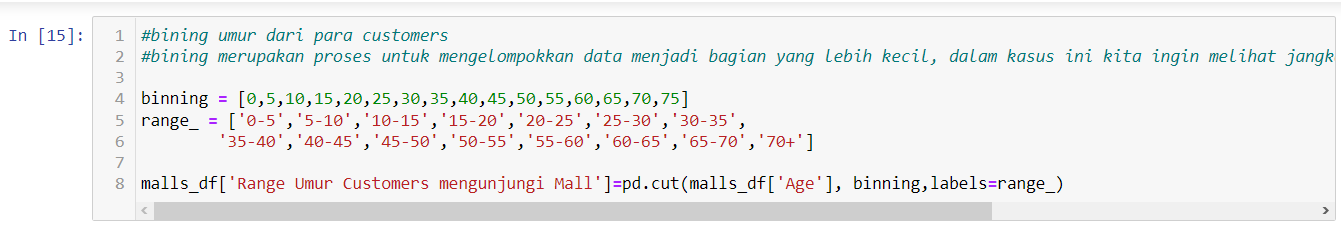
Kesimpulannya adalah data tersebut tingkat kerapatannya rendah, beserta tidak ada relationship karena menyebar kemana mana, dan arahnya juga tidak jelas antara positif maupun negatif.



Gambar 4.5 Heatmap antara korelasi Age, Gender, dengan Spending Score.

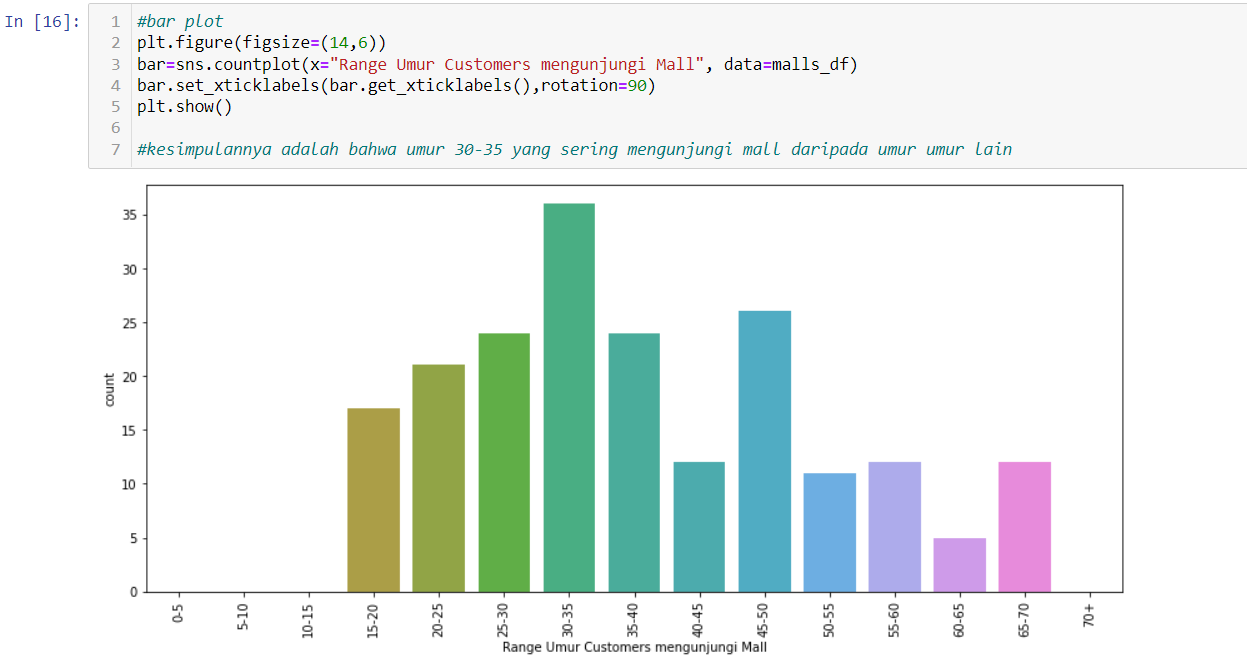


Gambar 4.6 Boxplot antara Annual Income dengan Gender



Binning merupakan proses untuk mengelompokkan data ke dalam bagian-bagian yang lebih kecil yang disebut *bin* berdasarkan kriteria tertentu. Juga merupakan salah satu teknik pra proses data yang akan digunakan untuk meminimalisir kesalahan dalam pengamatan serta terkadang dapat meningkatkan akurasi dari model prediktif.

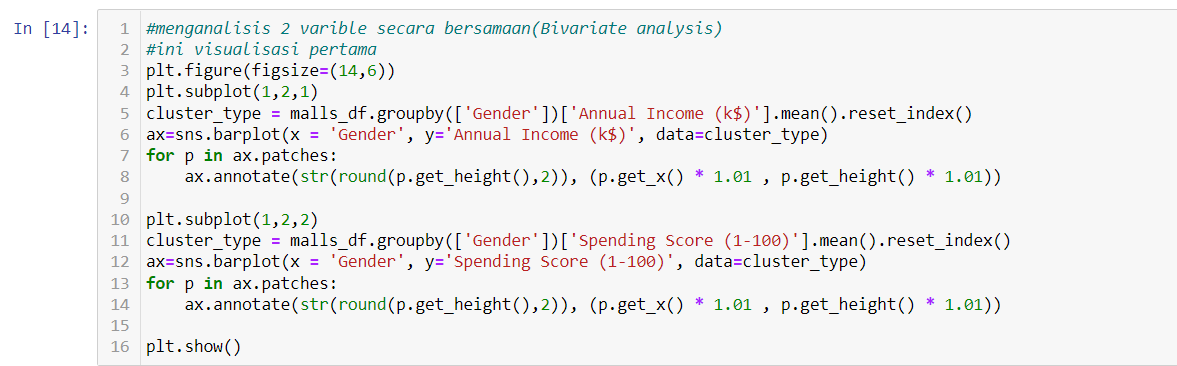
Binning biasanya digunakan untuk mengelompokkan ***data numerik*** menjadi beberapa bin agar sebaran data lebih mudah dipahami. Contohnya mengelompokkan usia.

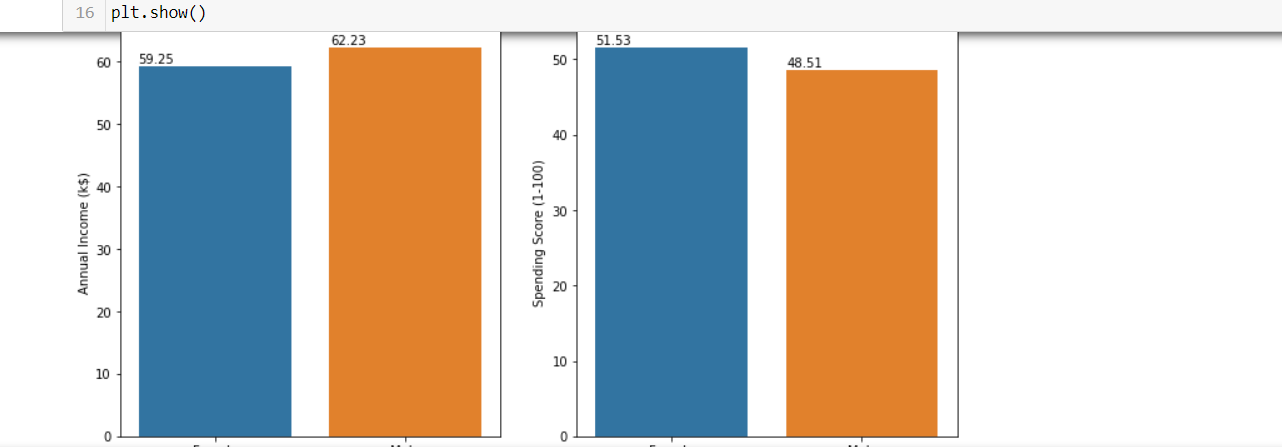


Gambar 4.7 Histogram range umur customer di Mall

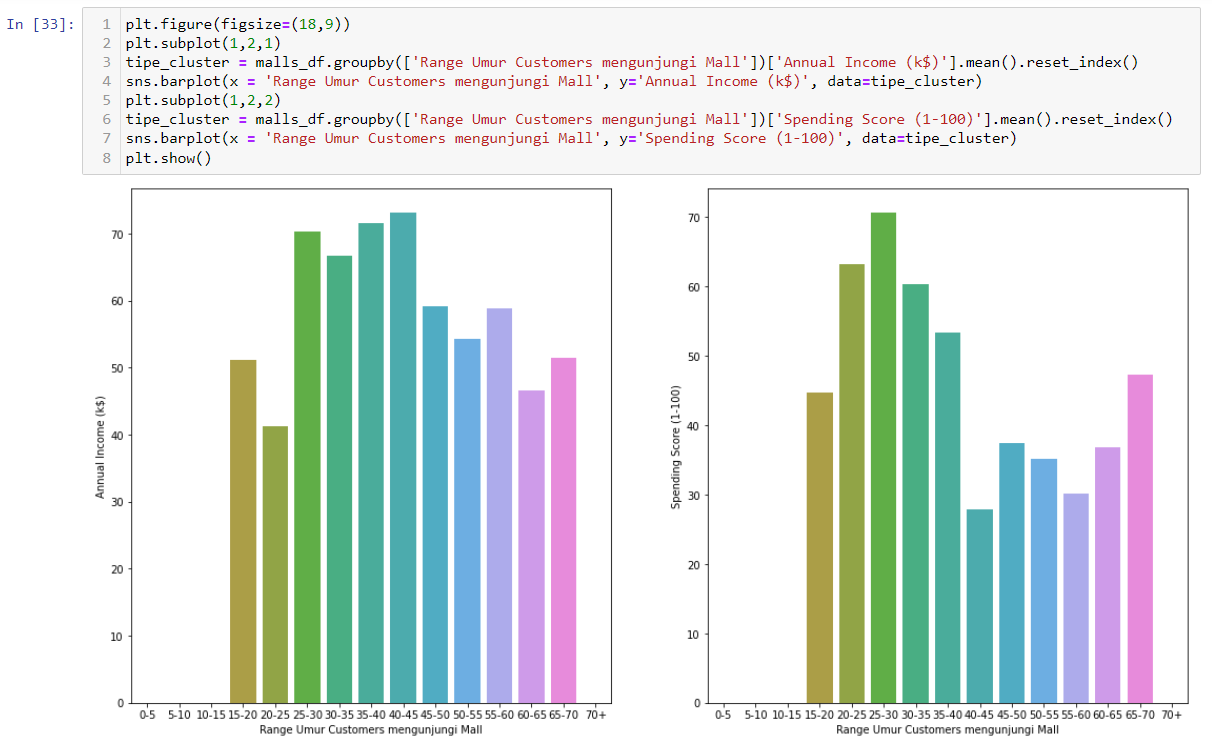
Dari data diatas range umur 30-35 adalah pengunjung yang paling sering mengunjungi mall.

### **4.1.1 Bivariate Analysis**



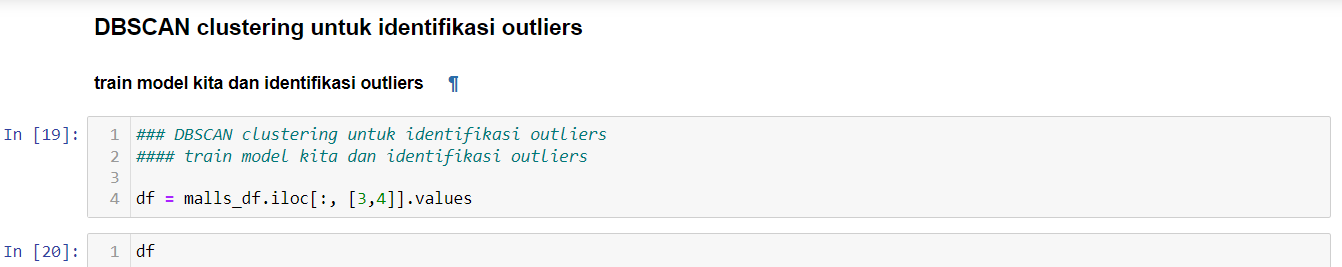


Gambar 4.8 Diagram Batang



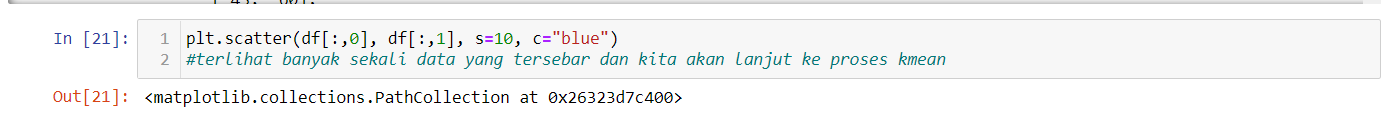
Gambar 4.9 Perbandingan histogram Annual Income dengan Spending Score

## **4.2 Pembahasan (DBSCAN)**

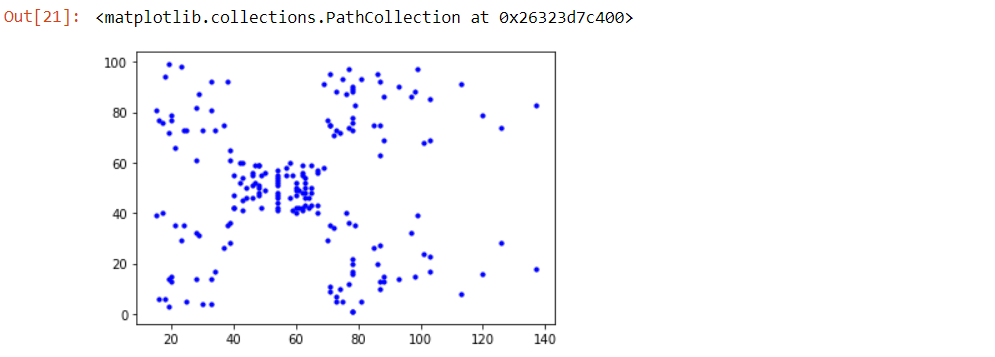
****

****

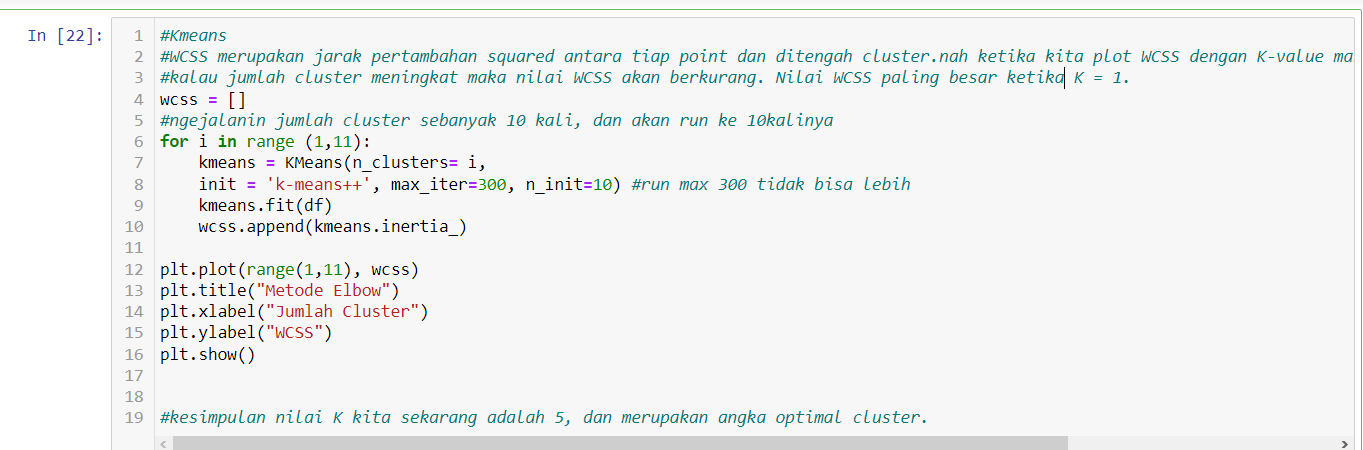
Inisialisi DBscan pertama tama kita membuat variable df , lalu memasukkan data frame sebelumnya yakni **malls\_df**. Setelah itu tambahkan code .iloc[:, [3,4] yang artinya semua data array akan di list pada code dibawah ini. Dimulai dari column yang pertama.

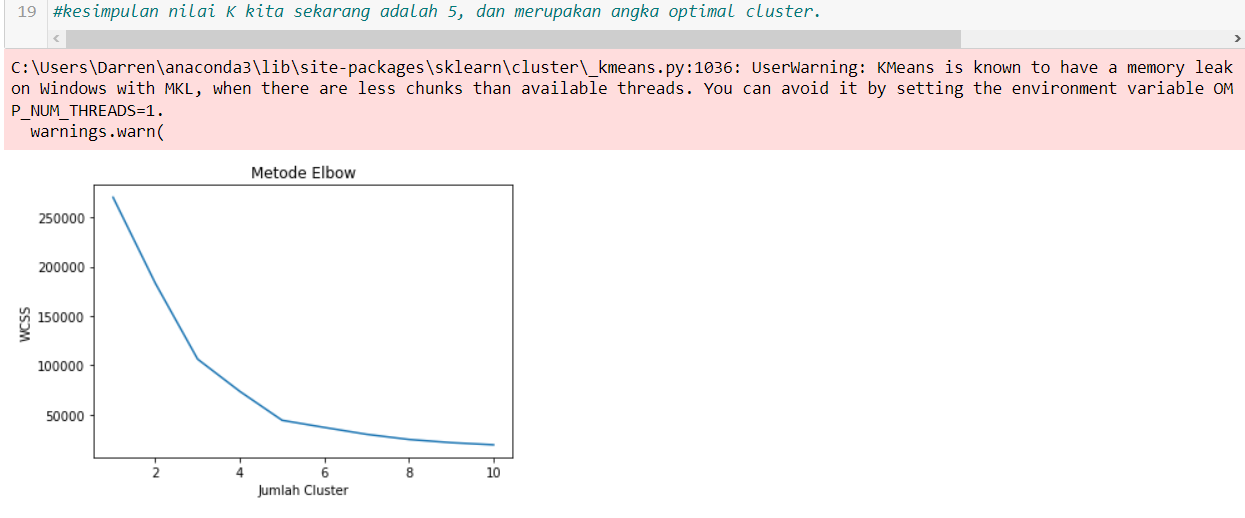


Kemudian kita tambahkan visualisasi yakni scatter plot untuk melihat seberapa besar cluster yang ada pada mall customer segmentation data. Kita menggunakan warna biru, dan bisa diganti dengan warna bebas lainnya.

****

Sebenarnya K Means tidak diperlukan dalam metode dbscan ,namun kita tetap memasukan karena ada WCSS. pada code dibawah ini menggunakan for untuk pengulangan agar di run sebanyak 10 kali. Dan menerapkan code KMeans lalu kita juga membuat sebuah plot agar bisa divisualisasikan dengan benar. Nah pada code ini nantinya akan menggambarkan elbow method yang merupakan sebuah menentukan jumlah cluster yang akan digunakan pada k-means clustering. Hasilnya akan ada dibawah ini.

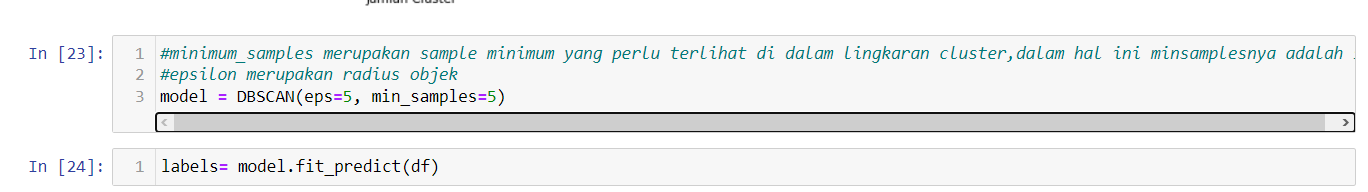
****

****

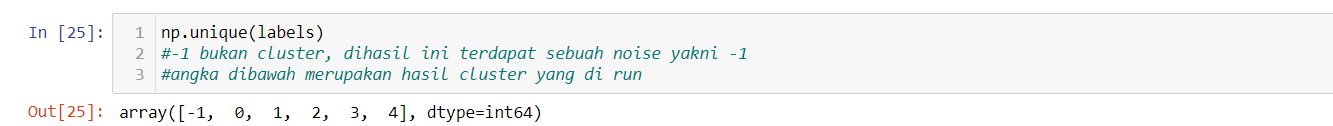
Elbow method sendiri merupakan metode untuk menentukan jumlah cluster yang tepat melalui persentase hasil perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik.

**Tujuan K-Means pada DBScan:**

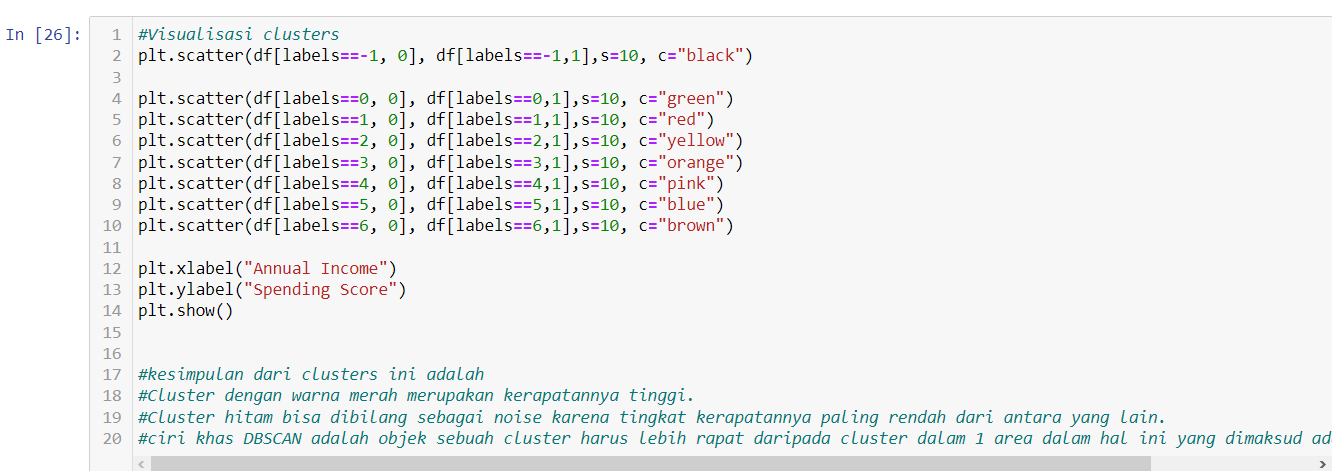
Membantu dalam pembuatan elbow method untuk DBSCAN sehingga bisa menentukan siku pada 1 titik yang palis pas(dlm hal ini cluster optimal).

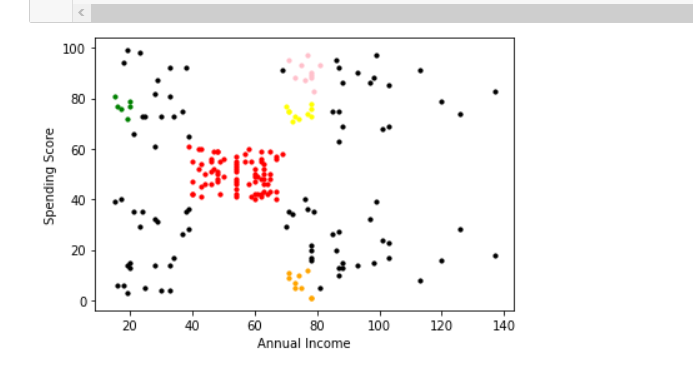
****

Selanjutnya kita masuk ke proses DBScan untuk menemukan samples pada cluster yang sudah dijalankan pada code sebelumnya. Kita sekarang berada di numpy np.unique labels, untuk melihat array yang ada noise atau tidak karena di dataset mall customers segmentation ada noise maka kita mengambil angka -1 sebagai noisenya.



Langkah terakhir kita memodifikasi cluster yang sudah dibuat sebelumnya pada bagian atas dengan visualisasi warna agar kelihatan jelas yang mana density tinggi maupun rendah.



****

Terlihat bahwa cluster merah merupakan cluster yang paling rapat densitasnya. Ciri khas dbscan yakni sebuah objek cluster harus lebih rapat daripada cluster dalam 1 area. Dan noise paling keliatan di data ini adalah cluster berwarna hitam(yang mempunyai kerapatan data yang rendah).

## **4.3 Kesimpulan**

Kesimpulan dari projek ini dengan menerapkan algoritma DBscan adalah

1. DBscan bisa membantu para data analyst untuk mendeteksi adanya noise atau outlier dalam mengelompokkan sebuah data berdasarkan tingkat kepadatan data sehingga memunculkan cluster dengan tingkat kepadatan yang tinggi maupun yang rendah. Kemudian DBScan membutuhkan 2 buah parameter untuk melakukan clustering yakni *eps* dan MinPts(Minimum points). Dalam mall customer kasus ini si data analyst bisa tahu cluster cluster mana saja pengunjung dengan umur, serta spending score serta annual spending yang digambarkan pada cluster cluster di atas.

## **4.4 Saran**

Semoga dengan adanya proyek ini bisa membuka wawasan bagi para calon data analyst dalam melakukan clustering data menggunakan dbscan method. Penulis berharap juga untuk para pembaca yang membaca laporan ini bisa menambah wawasan dalam pengenalan DBscan dalam clustering sebuah data serta dapat mengimplementasikan terhadap data lain menggunakan DBscan.

# **DAFTAR PUSTAKA**

Raharja, A. D. B. (n.d.). *Machine learning: Pengertian, Cara Kerja, Dan 3 Metodenya!* TALENTS. Retrieved December 8, 2022, from <https://www.ekrut.com/media/apa-itu-machine-learning>

ArcGIS, K. D. P. &. (2022, February 8). *Mengenal algoritma DBSCAN Dan Manfaatnya*. Konsultan Analis Data Penelitian dan Peta Digital. Retrieved December 8, 2022, from <https://patrastatistika.com/mengenal-algoritma-dbscan-dan-manfaatnya/>

Awan, A. A. (2022, August 17). *Implementing DBSCAN in python*. KDnuggets. Retrieved December 8, 2022, from <https://www.kdnuggets.com/2022/08/implementing-dbscan-python.html>

Jagadish, K. (2019, April 29). *Mall customers*. Kaggle. Retrieved December 8, 2022, from <https://www.kaggle.com/datasets/kandij/mall-customers>

Link gambar online

[https://miro.medium.com/max/1339/0\*xu3GYMsWu9QiKNOo.png](https://miro.medium.com/max/1339/0*xu3GYMsWu9QiKNOo.png)