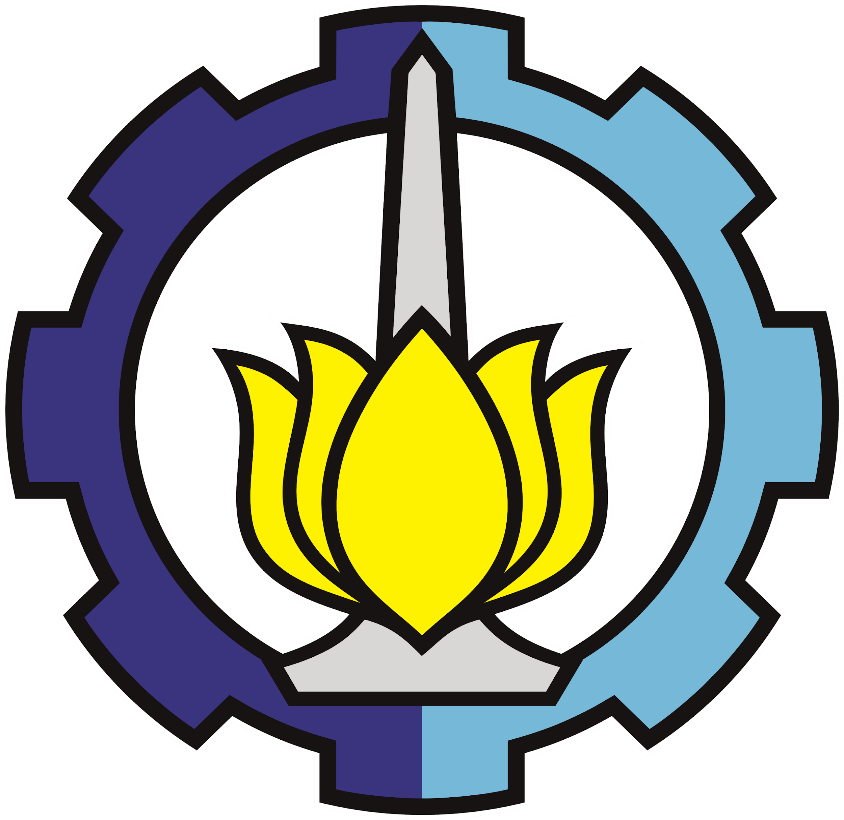
**TUGAS 2**

**KECERDASAN KOMPUTASIONAL**

****

**Kelas : B**

**Gema Adi Perwira 05111840000138**

**Vieri Fath Ayuba 05111840000153**

**M. Samsu Dhuha 05111840000155**

**Dosen : Dr. Diana Purwitasari, S.Kom., M.Sc.**

**Departemen Informatika**

**Fakultas Teknologi Elektro Dan Informatika Cerdas**

**Institut Teknologi Sepuluh Nopember**

**Surabaya**

**2020**

**Rumusan Masalah**

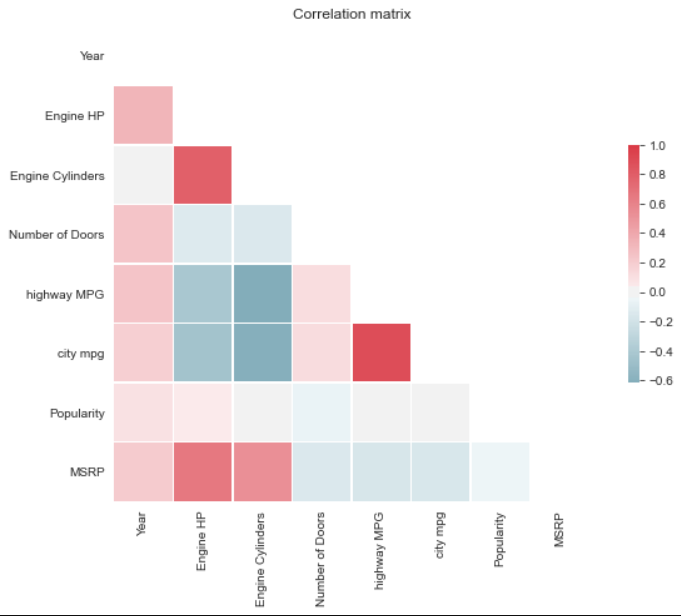
1. Apa kombinasi fitur yang memberikan cluster terbaik menurut indikator Sillhoutte score !
2. Bagaimana hubungan korelasi antar fitur2 yang terpilih?
3. Apakah ada proses cleaning data (impute, dll) !
4. Bagaimana hasil cluster jika centroid terbentuk secara random atau dipiihkan yang terbaik kmeans ++ !
5. Bandingkan KMeans dgn Hiearchical (divisive, agglomerative) !
6. Bandingkan dengan algoritma jika menggunakan SkLearn !

Tujuan: utk Data Cars mana yang paling sesuai dalam melakukan pengelompokkan? partitional atau hierarchical

**Penjelasan**

1. Apa kombinasi fitur yang memberikan cluster terbaik menurut indikator Sillhoutte score!

Pendekatan pertama kita harus mencari korelasi yang terbaik dari kombinasi fitur. Disini kita menggunakan heat map untuk mencari korelasi yang terbaik. Heat map itu sendiri merupakan metode grafis untuk memvisualisasikan nilai atribut menurut kelas dalam matrix 2 arah, serta nilainya direpresentasikan dengan warna. Dibawah ini merupakan observasi kami untuk mencari korelasi yang terbaik dengan metode heatmap :



Gambar 1. Heatmap

Dilihat dari heat map terlihat bahwa fitur yang paling berkorelasi adalah city mpg dan highway mpg. Kita mengambil 3 fitur yang paling berkorelasi untuk dianalisis berikutnya menggunakan sillhoute score, sillhoute plot, serta scatter plot. Tiga kombinasi fitur yang paling berkorelasi berdasarkan heat map sebagai berikut :

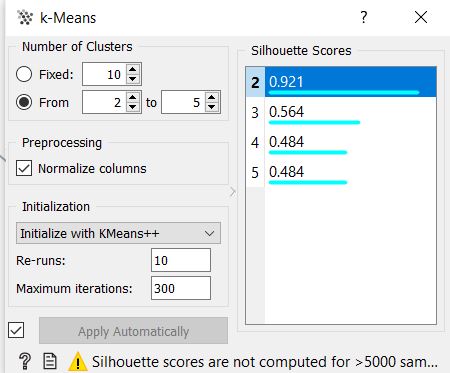
1. City MPG dengan highway mpg
2. Engine cylinder dan engine HP
3. MSRP dan Engine HP

Pendekatan selanjutnya kita mencari nilai sillhoute score dari tiga fitur tersebut dengan dengan cara mencari nilai silhoute score yang mendekati 1, karena nilai yang mendekati satu menunjukkan bahwa dekat dengan pusat kluster dan jauh dari cluster lain. Dengan menggunakan pendekatan sillhoute score, sillhoute plot, serta scatter plot, berikut ini merupakan ekspolarasi kami terhadap 3 kombinasi fitur :

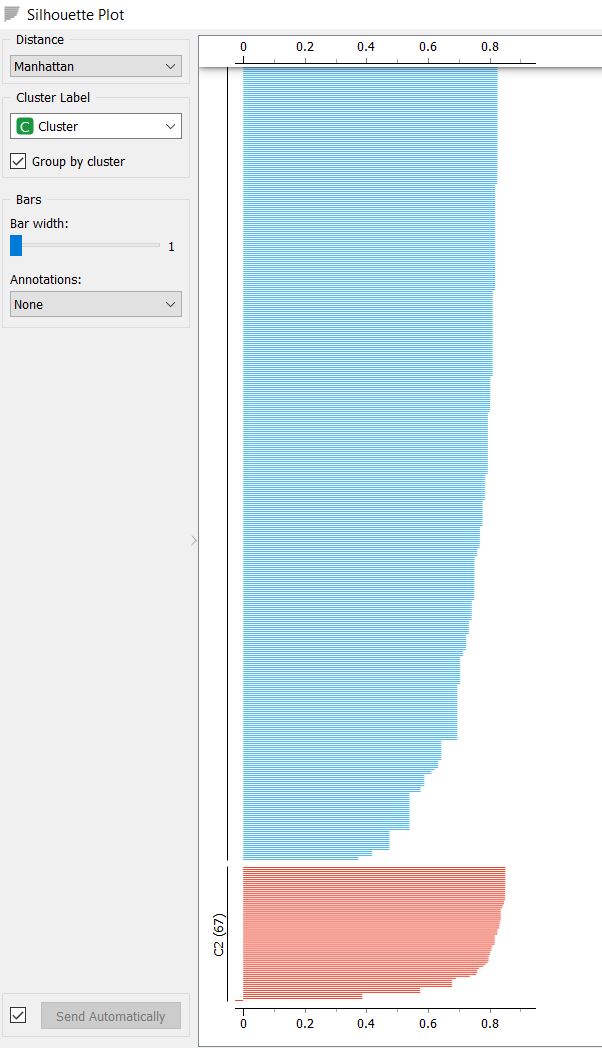
1. City mpg dengan Highway mpg

Sillhoute score terbesar yang didapatkan dari city mpg dengan highway mpg yaitu 0.921. Kemudian, observasi tersebut dilanjutkan menggunakan widget sillhoute plot. Dapat dilihat dari silhoute plot, tidak ada nilai sillhoute plot yang bernilai negatif, hal itu menunjukkan bahwa cluster sudah terkelompok dengan baik serta tidak ada yang tumpang tindih.

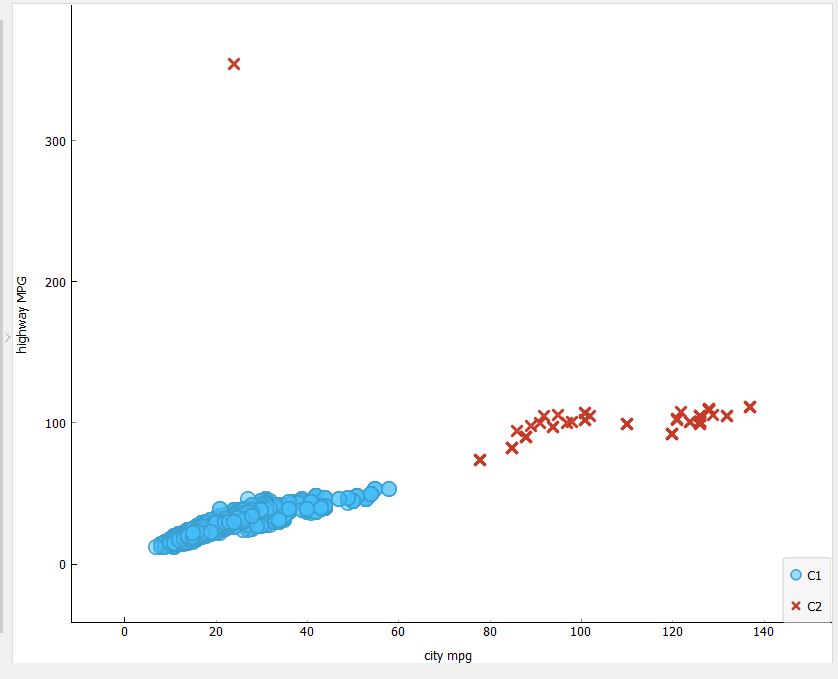
Observasi terakhir yaitu menggunakan scatter plot. Pada scatter plot terlihat bahwa cluster terkelompok dengan baik, tidak ada yang tumpang tindih antar cluster.



Gambar 2. K-Means - Highway MPG & City MPG



Gambar 3. Sillhouate Plot -H-MPG & C-MPG

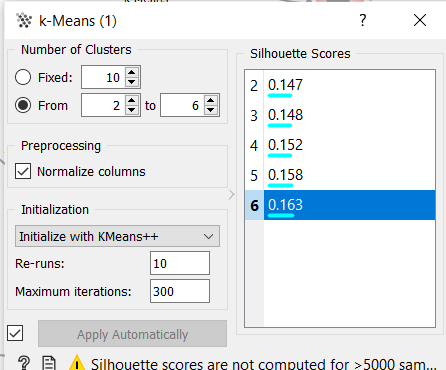


Gambar 4. Scatter Plot H-MPG & C-MPG

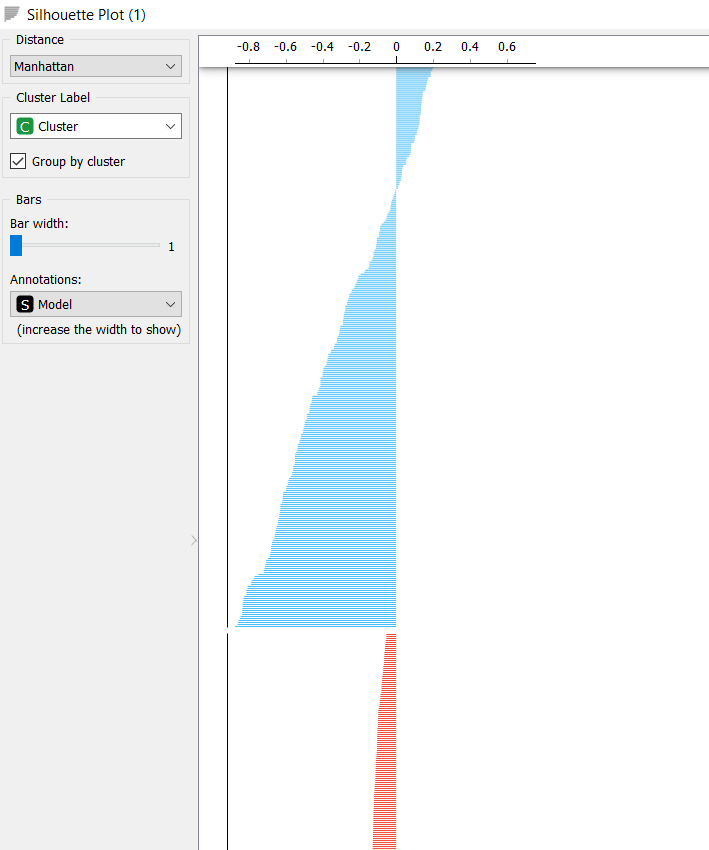
1. Engine cylinder dan engine HP

Sillhoute score terbesar yang didapatkan dari kombinasi engine cylinder dan engine hp dengan highway mpg yaitu 0.163. Setelah itu, dilanjut observasi sillhoute plot. Dari informasi yang didapat dengan di sillhoute plot, banyak cluster yang tumpang tindih, itu dapat dilihat banyak silhoute score yang bernilai negatif.

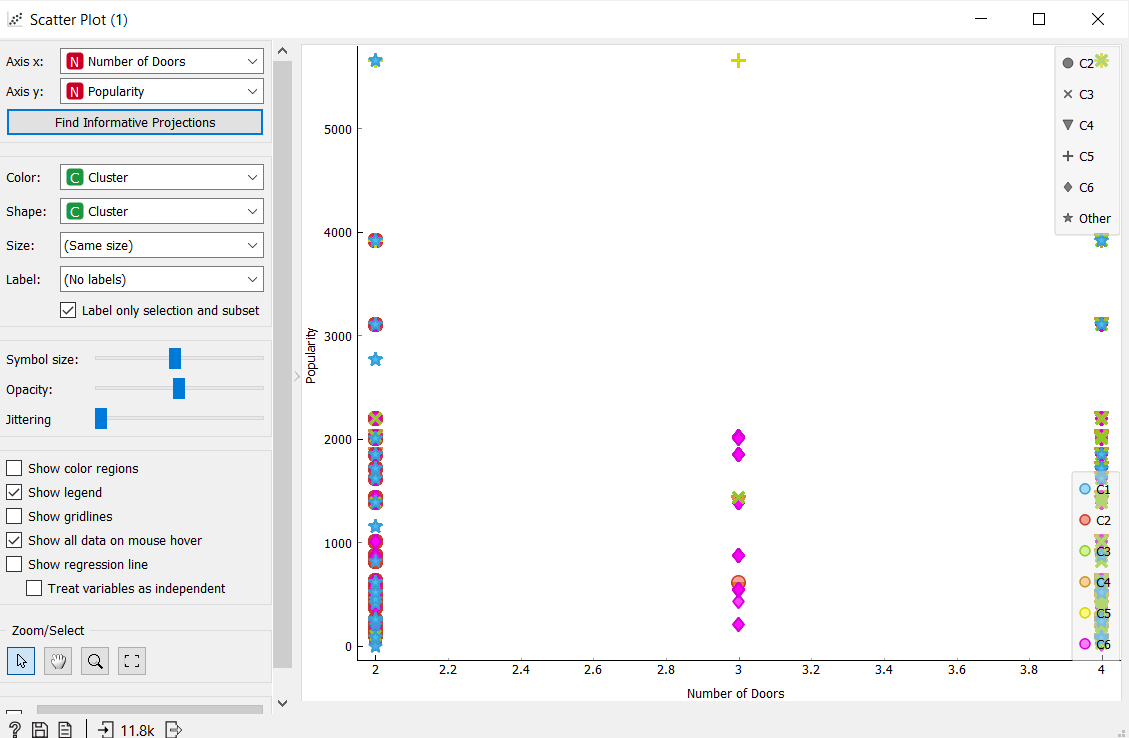
Observasi yang terakhir menggunakan scatter plot. Pada scatter plot terlihat bahwa cluster tidak terkelompok dengan baik, terdapat anggota cluster yang tumpang tindih.



Gambar 5 - K-Means Engine Cylinder dan Engine HP



Gambar 6 – Sillhoute Plot - Engine Cylinder dan Engine HP

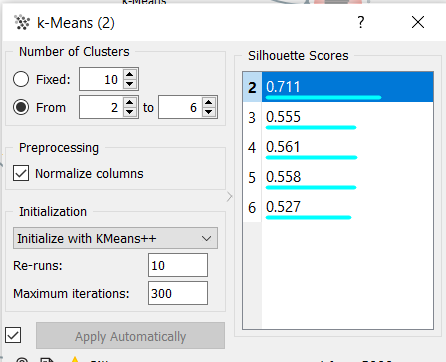


Gambar 7 – Scatter Plot - Engine Cylinder dan Engine HP

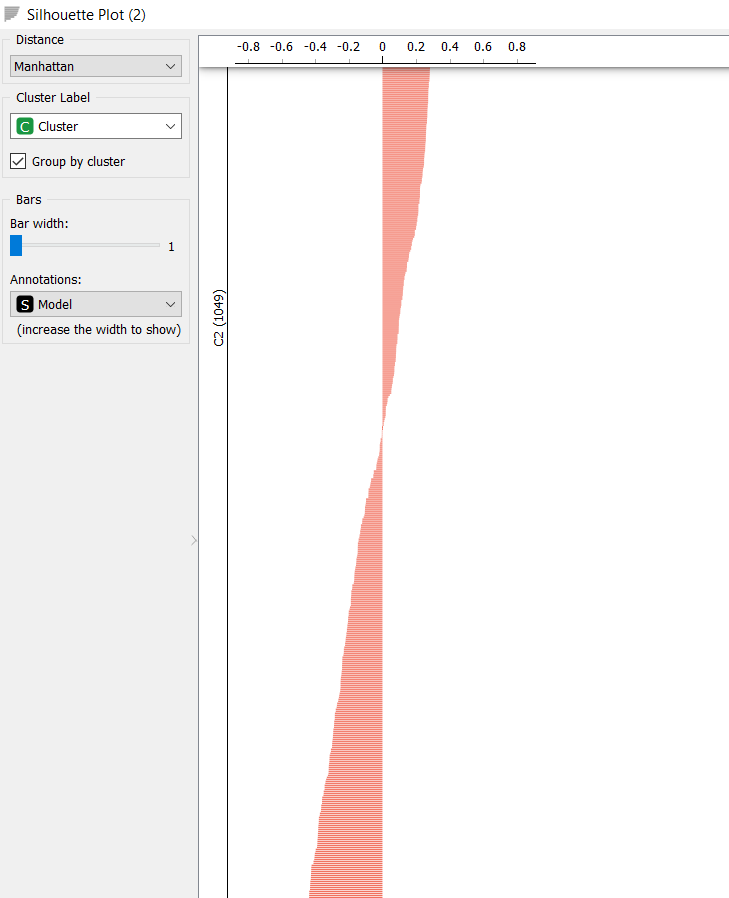
1. MSRP dan Engine HP

Sillhoute score terbesar yang didapatkan dari kombinasi engine cylinder dan engine hp dengan highway mpg yaitu 0.711. Setelah itu, dilanjut observasi sillhoute plot. Dari informasi yang didapat dengan di sillhoute plot, banyak cluster yang tumpang tindih, itu dapat dilihat banyak silhoute score yang bernilai negatif.

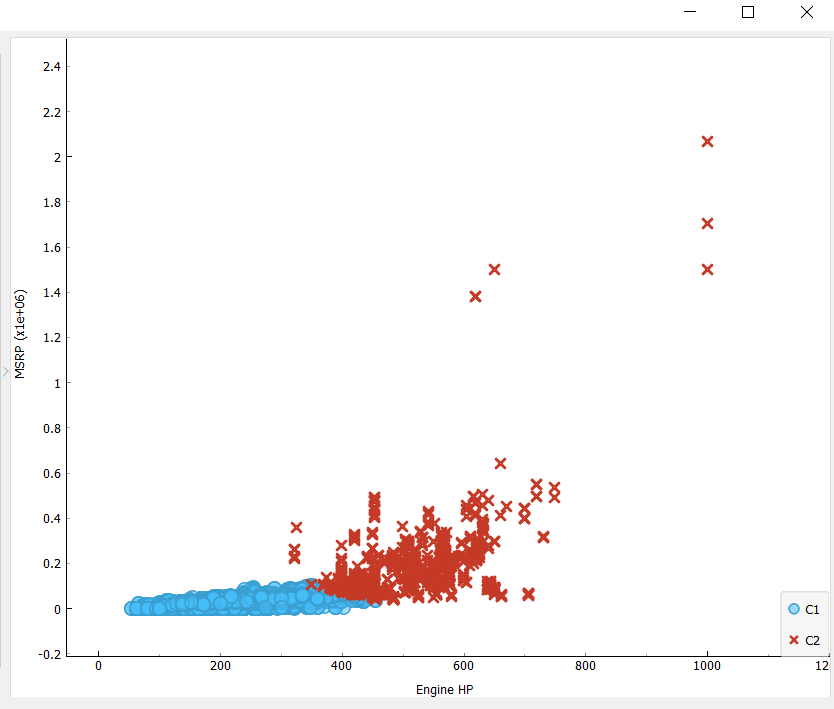
Observasi yang terakhir menggunakan scatter plot. Pada scatter plot terlihat bahwa cluster tidak terkelompok dengan baik, terdapat anggota cluster yang tumpang tindih.



Gambar 8 – K-Means - MSRP dan Engine HP



Gambar 9 – Sillhouete Plot - MSRP dan Engine HP

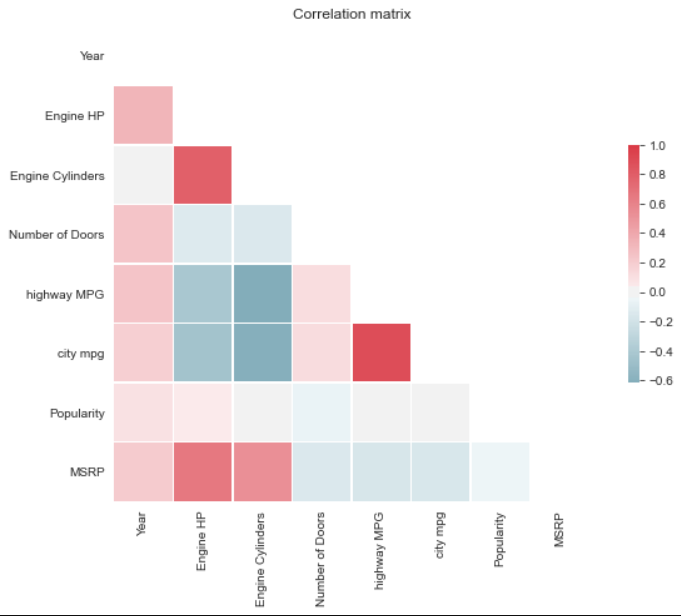


Gambar 10 – Scatter Plot - MSRP dan Engine HP

Kesimpulan :

Kombinasi fitur yang memberikan cluster terbaik menurut indikator Sillhoutte score adalah kombinasi fitur City MPG dan Highway MPG dengan sillhoute score terbesar yang didapatkan yaitu 0.921. Kemudian, observasi menggunakan widget sillhoute plot. Dapat dilihat dari silhoute plot, tidak ada nilai sillhoute plot yang bernilai negatif, hal itu menunjukkan bahwa cluster sudah terkelompok dengan baik serta tidak ada yang tumpang tindih. Observasi terakhir yaitu menggunakan scatter plot. Pada scatter plot terlihat bahwa cluster terkelompok dengan baik, tidak ada yang tumpang tindih antar cluster.

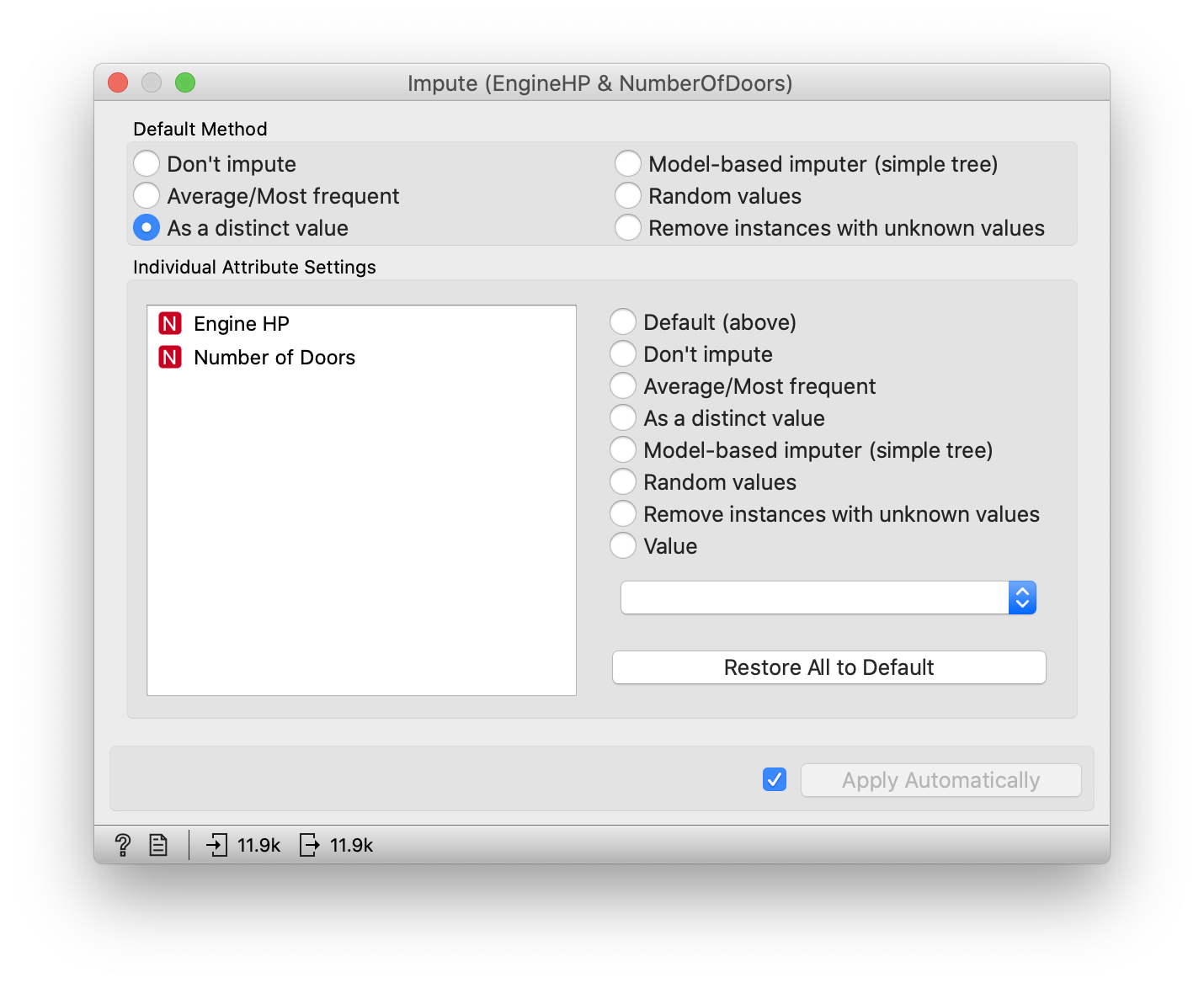
2. Bagaimana hubungan korelasi antar fitur2 yang terpilih?



Gambar 11 – Heat Map

* City MPG adalah fitur rata - rata pemakaian bahan bakar mobil di kota.
* Highway MPG adalah fitur rata - rata pemakaian bahan bakar di jalan raya.

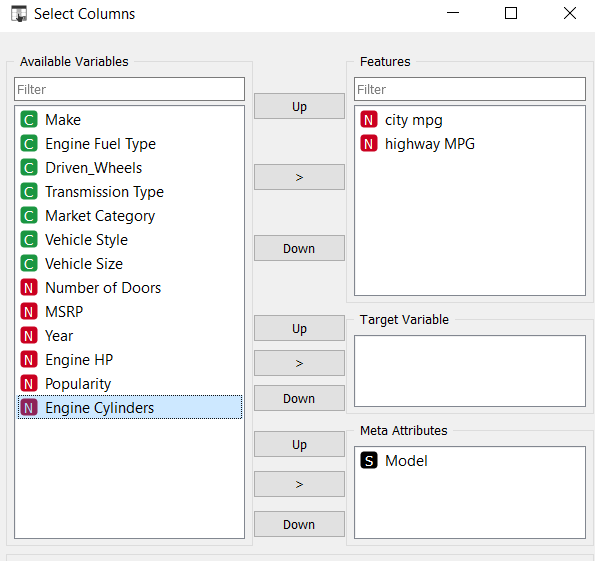
Hubungan korelasi dari kombinasi fitur City MPG dan Highway MPG adalah jika nilai city MPG dan Highway MPG semakin tinggi, maka kualitas dari suatu mobil akan semakin baik, karena pemakaian bahan bakar baik pemakaian kota dan jalan raya semakin hemat.

3. Apakah ada proses cleaning data (impute, dll)

Gambar 12 – Impute Data

Proses cleaning yang dilakukan yaitu dengan menggunakan widget “Impute”. Di widget “Impute” dilakukan metode yaitu menghapus data dengan nilai yang tidak diketahui (NULL Values).

Kami juga melakukan cleaning data untuk kolom yang tidak kami butuhkan dengan menggunakan select column.



Gambar 13 – Select Column

4. Bagaimana hasil cluster jika centroid terbentuk secara random atau dipilihkan yang terbaik kmeans ++ !

Pengelompokkan k - means bertujuan untuk menyatu pada kumpulan optimal pusat cluster yang disebut centroid, dan anggota cluster nya berdasarkan jarak dari centroid tersebut melalui iterasi yang berurutan. Jika semakin optimal posisi awal centroid, maka semakin sedikit iterasinya.

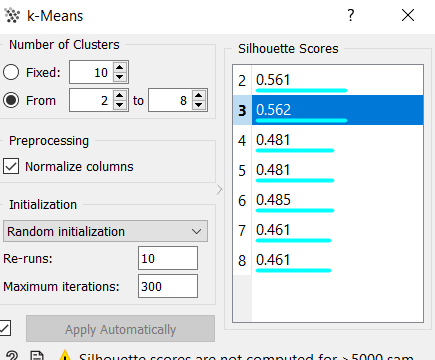
* Menggunakan random data centroid

Pada formula ini, k random data centroid dipilih dari dataset dan digunakan sebagai centroid awal, formula ini jelas sangat mudah berubah, serta membuat skenario dimana centroid yang dipilih tidak pada posisi yang tepat.

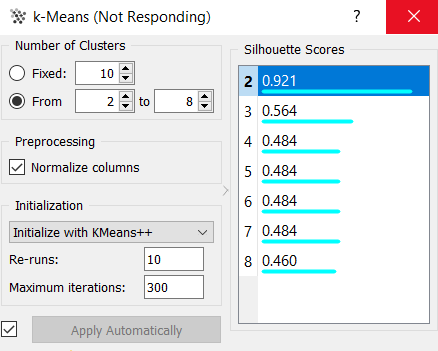
* Menggunakan K - means++

Pada formula ini, K-means++ menentukan centroid awal dengan menetapkan centroid awal ke lokasi titik data yang dipilih secara acak. Kemudian memilih centroid berikutnya dari titik data yang tersisa berdasarkan probabilitas yang sebanding dengan jarak kuadrat dari titik pusat terdekat yang ada pada titik tertentu. Sehingga, formula ini mendorong centroid sejauh mungkin dari satu sama lain mencangkup sebanyak mungkin ruang data yang ditempati dari inisialisasi.

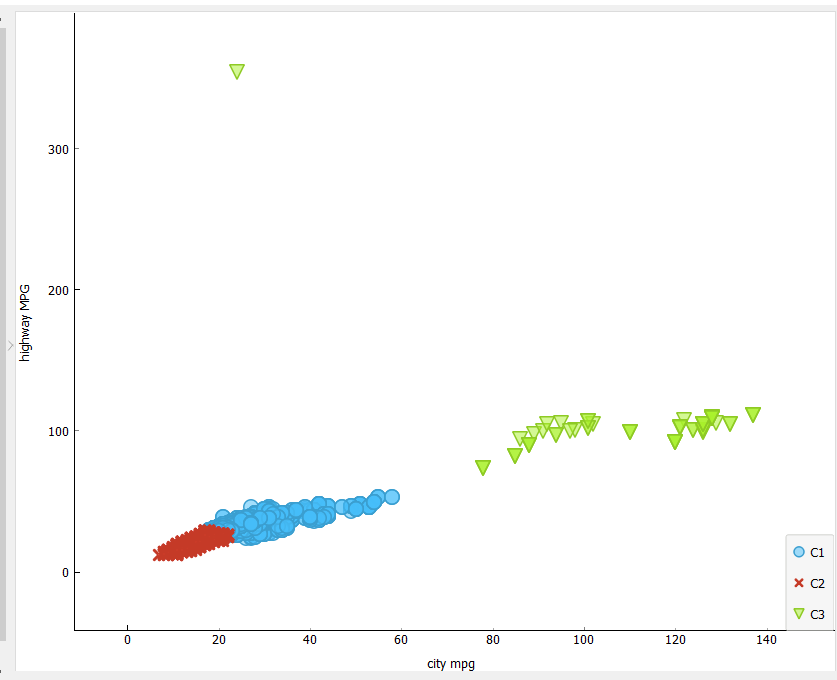
Berikut ini adalah observasi dari kelompok kami terhadap hasil cluster dari centroid yang terbentuk secara random dan centroid yang dipilih oleh kmeans++. Menggunakan kombinasi fitur city MPG dan highway MPG.



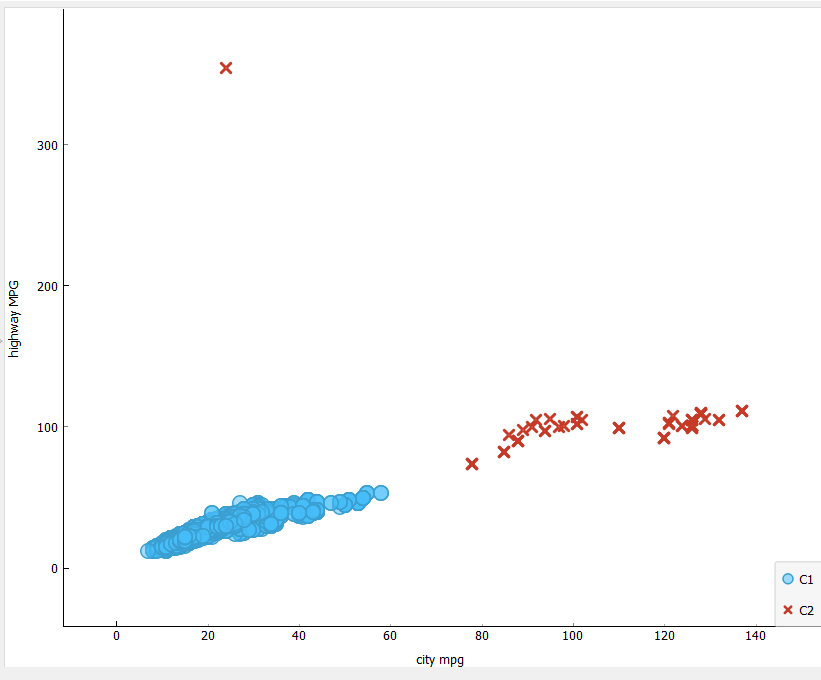
Gambar 14 – K – Means menggunakan random initialization



Gambar 15 – K – Means menggunakan Kmeans++



Gambar 16 – Scatter Plot menggunakan Random Intialization



Gambar 17 – Scatter Plot menggunakan Kmeans++

Dari observasi yang kami lakukan, centroid yang ditentukan dengan Kmeans++ memiliki sillhoute score lebih tinggi dengan centroid yang ditentukan dengan random initialization. Begitu pula observasi menggunakan scatter plot yang terbentuk dengan centroid yang ditentukan oleh means++ terbentuk lebih baik tidak ada tumpang tindih antar cluster.

5. Bandingkan KMeans dgn Hiearchical (divisive, agglomerative)

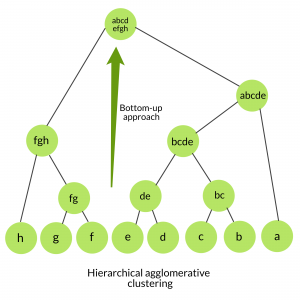
K-means adalah metode kluster analisis yang menggunakan jumlah cluster yang ditentukan sebelumnya. metode ini memberikan record ke setiap kluster untuk menemukan kluster berdasarkan jarak k.

K-means klustering membutuhkan pengetahuan awal tentang k yaitu banyaknya kluster yang membagi data.

Hiearchical merupakan metode analisis kelompok yang berusaha untuk membangun sebuahi hirarki kelompok data. strategi pengelompokannya memiliki dua jenis, yaitu

* Agglomerative

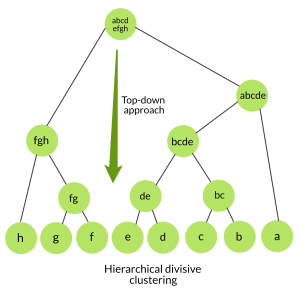
Agglomerative juga dikenal sebagai pendekatan bottom up. Algoritma bottom up memperlakukan setiap data sebagai cluster tunggal diawal dan kemudian secara rekursif menggabungkan pasangan cluster sampai semua cluster digabungkan menjadi satu cluster yang berisi semua data. Ilustrasi singkat sebagai berikut



Gambar 18 – Hierarchical Agglomerative Clustering

* Divisive

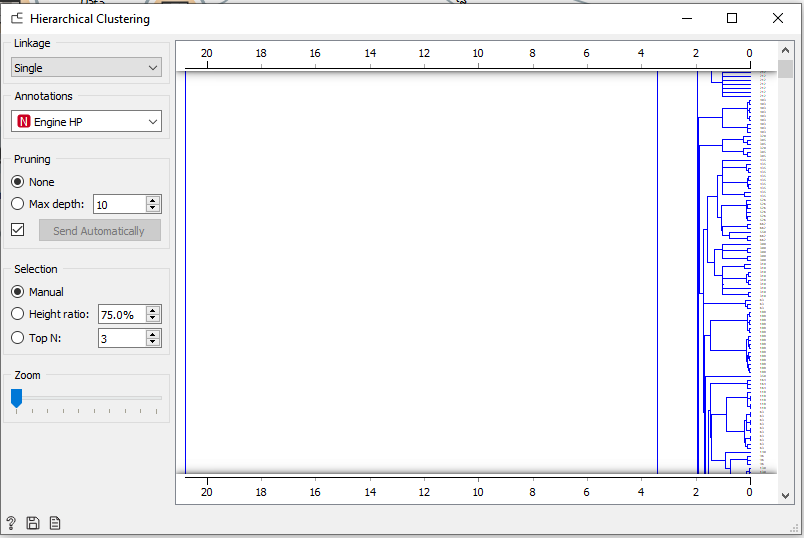
Divisive juga dikenal sebagai pendekatan top down. Algoritma ini tidak perlu menentukan jumlah cluster sebelumnya. Top down clustering memerlukan metode untuk memisahkan cluster yang berisi seluruh data dan melanjutkan dengan memisahkan cluster secara rekursif hingga data individu dipecah menjadi cluster tunggal.



Gambar 19 – Hierarchical Divisive Clustering

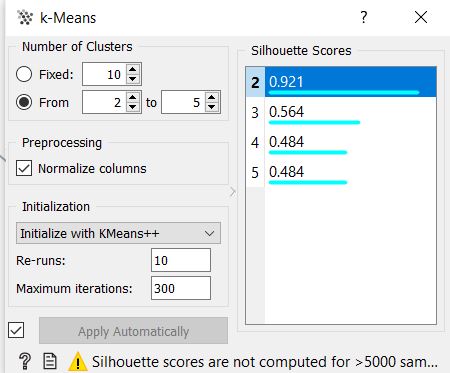
Perbandingan di Orange :

1. Hierarchical Clustering



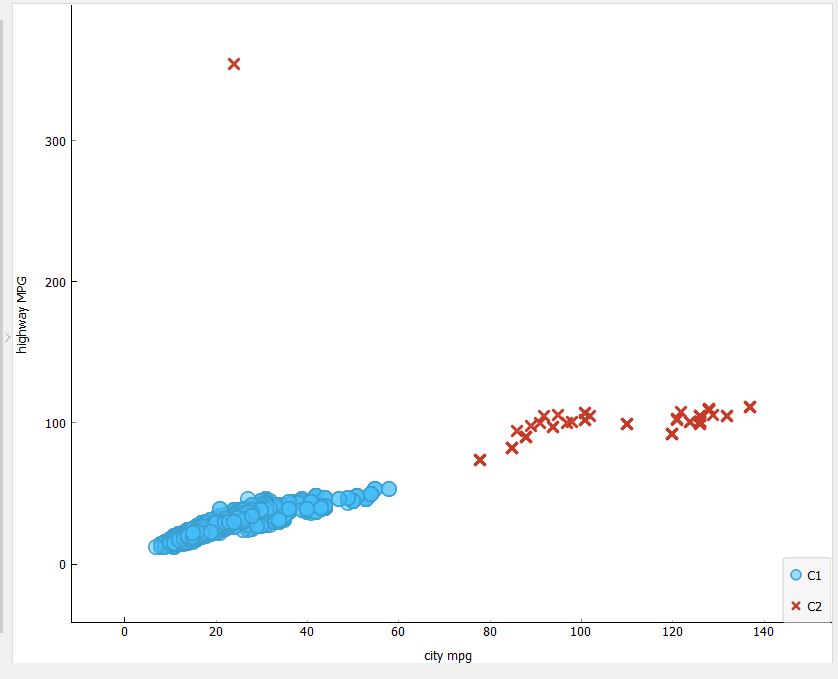
Gambar 20 – Hierarchical Clustering

1. KMeans Clustering



Gambar 21 – KMeans Clustering

Gambar 23 – Scatter Plot – Kmeans Clustering



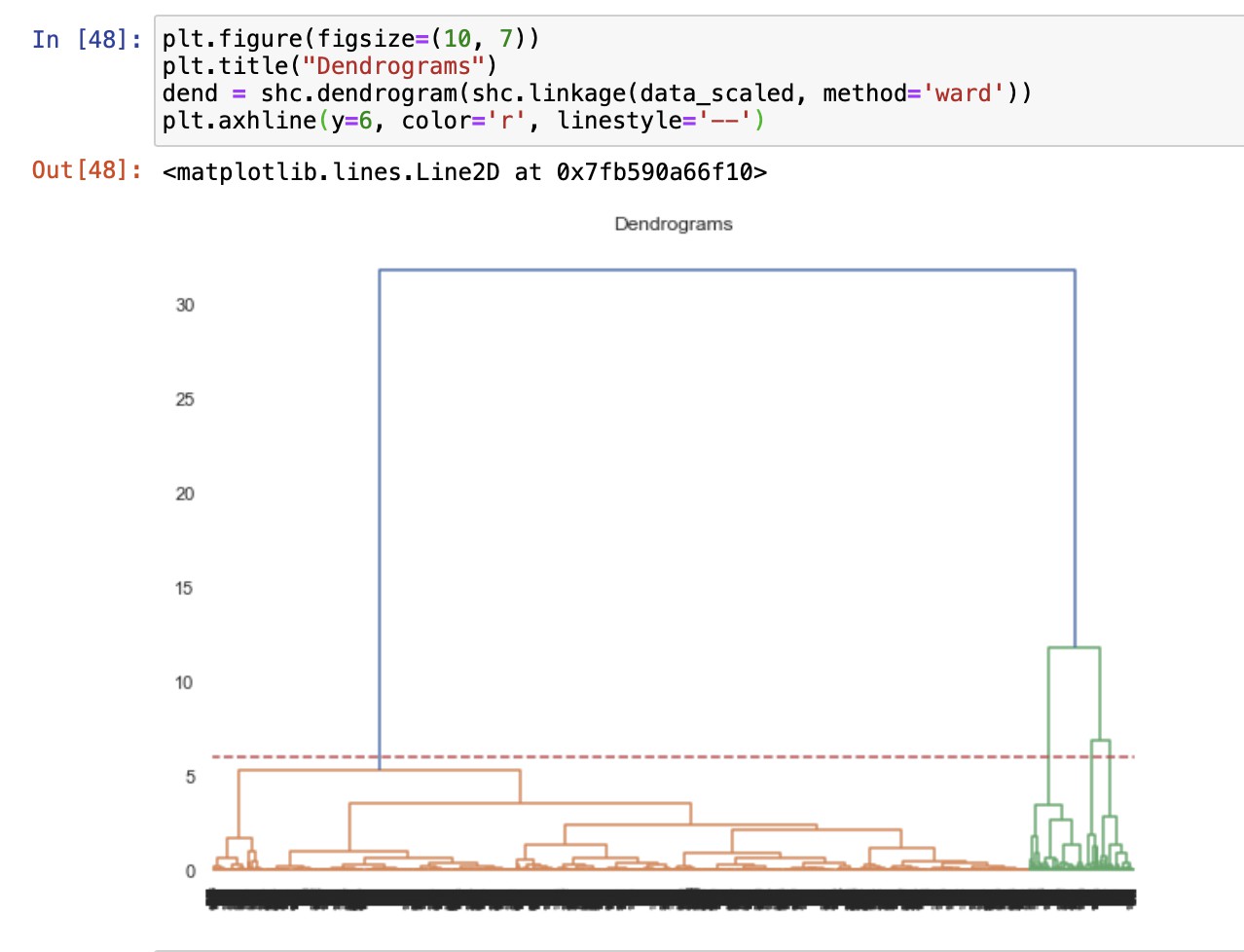
Kesimpulan : untuk dataset yang kami gunakan akan lebih mudah pengelompokkannya dengan algoritma K-means karena data yang kami gunakan besar jumlahnya. Dan akan menjadi lebih efisien menggunakan K-means dibandingkan hierarchical. Selain itu, k-means akan terjamin akurasinya.

6. Bandingkan dengan algoritma jika menggunakan SkLearn

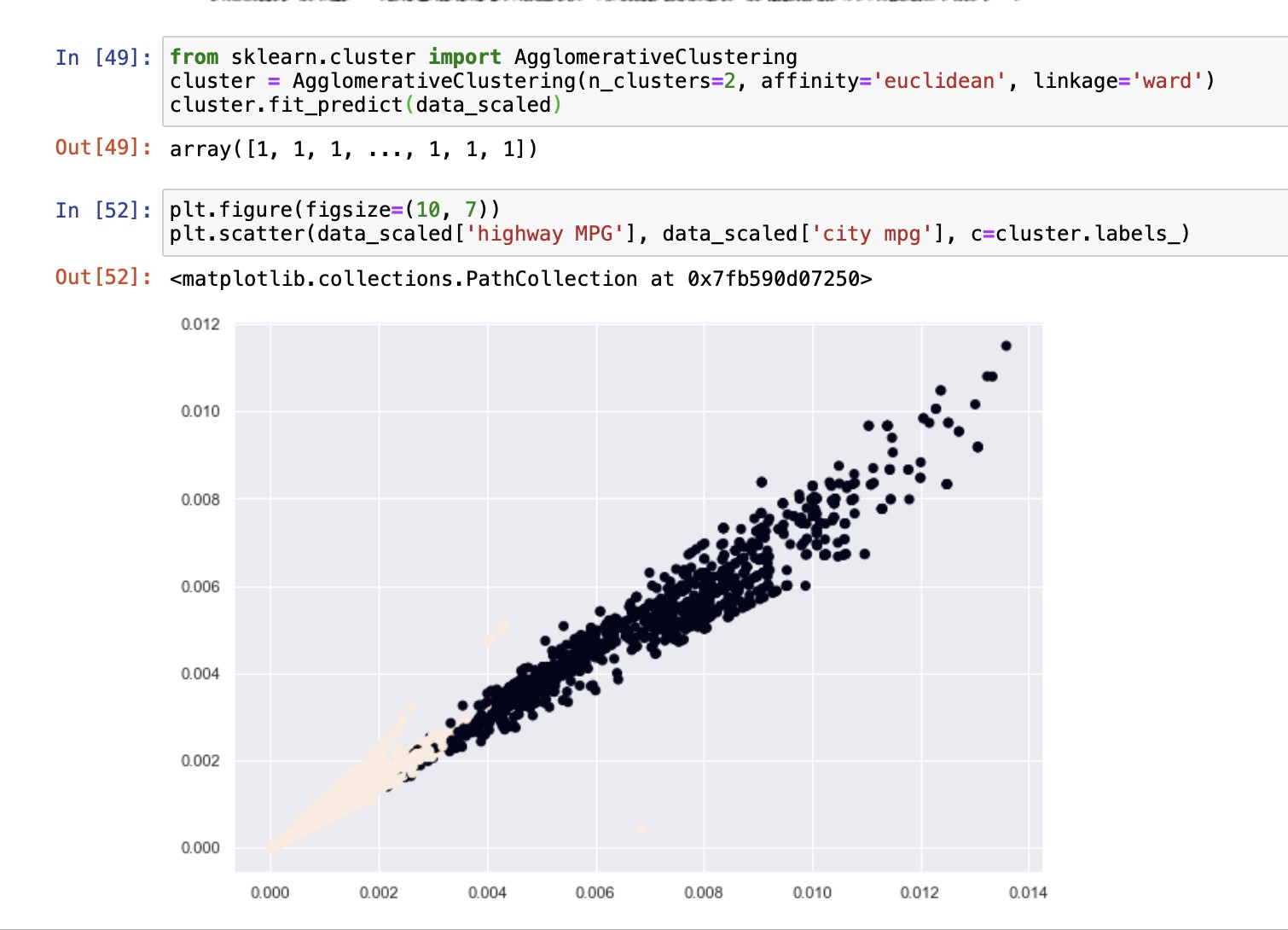
Tujuan: utk Data Cars mana yang paling sesuai dalam melakukan pengelompokkan? partitional atau hierarchical

[Scikit-learn](https://scikit-learn.org/) atau sklearn adalah modul untuk bahasa pemograman python yang dibangun diatas NumPy, SciPy, dan matplotlib, fungsinya dapat membantu melakukan processing data ataupun melakukan training data untuk kebutuhan machine-learning.

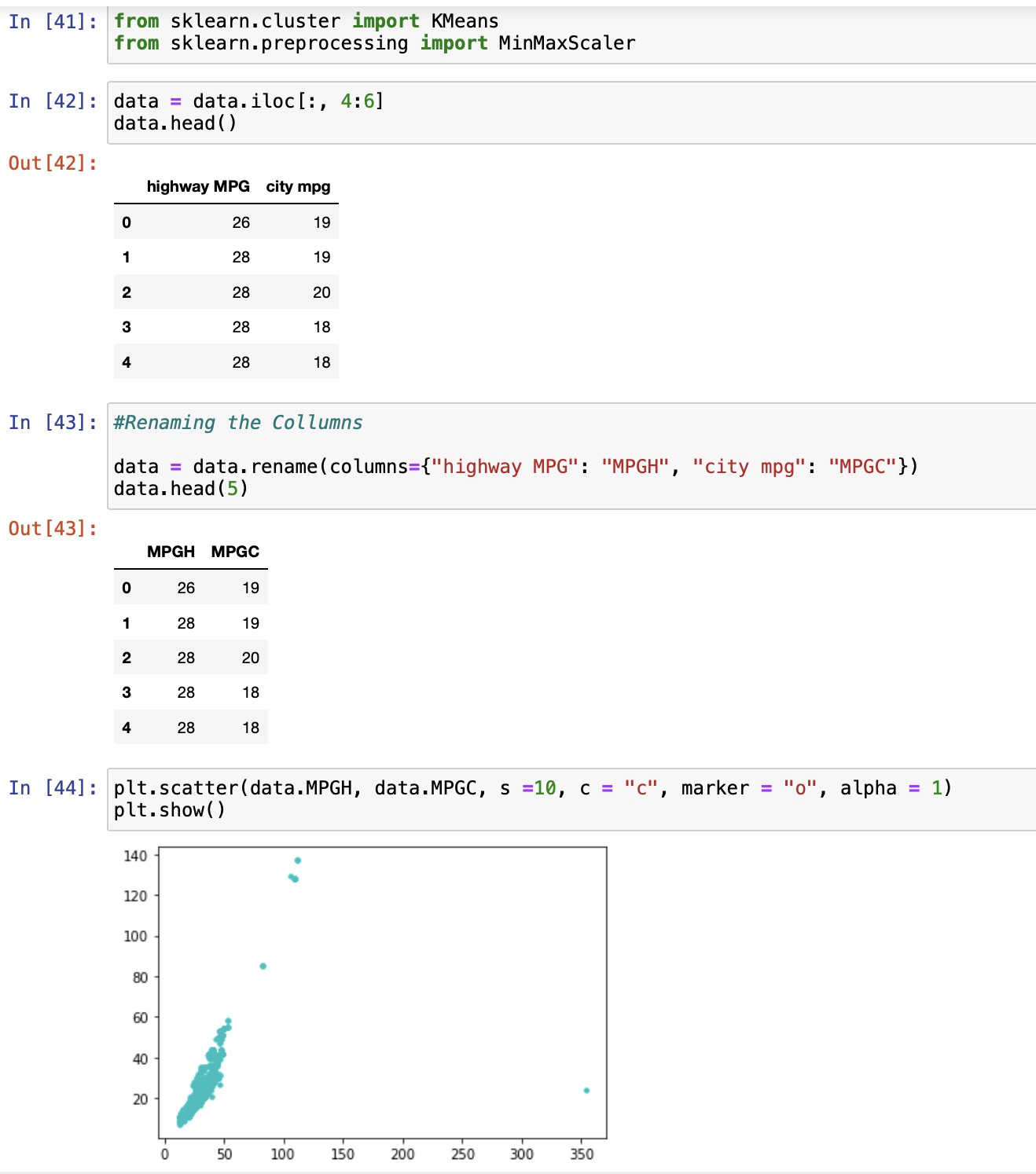
1). Hierarchical Clustering



Gambar 24 – Dendogram – Hierarchical Clustering



Gambar 25 – Agglomerative - Hierarchical Clustering dengan SKLearn



Gambar 26 – K-Means Clustering dengan SKLearn

Tujuan: utk Data Cars mana yang paling sesuai dalam melakukan pengelompokkan? partitional atau hierarchical.

Yang paling sesuai adalah dengan partitional clustering lebih tepatnya menggunakan algoritma K-means. Karena K-means bisa mencakup data yang banyak, sedangkan hierarchical tidak bisa. Data Cars sendiri mencakup 11.914 data. Kesimpulan : untuk dataset yang kami gunakan akan lebih mudah pengelompokkannya dengan algoritma K-means karena data yang kami gunakan besar jumlahnya. Dan akan menjadi lebih efisien menggunakan K-means dibandingkan hierarchical. Selain itu, k-means akan terjamin akurasinya.

**REFERENSI**

* Cooper Union. (2016). Car Features and MSRP Dataset. *<https://www.kaggle.com/CooperUnion/cardataset>* . Diakses pada 15 Oktober 2020 pukul 18.30.
* Enes Gokce. (2020). Predicting Used Car Prices With Machine Learning. [*https://towardsdatascience.com/predicting-used-car-prices-with-machine-learning-techniques-8a9d8313952*](https://towardsdatascience.com/predicting-used-car-prices-with-machine-learning-techniques-8a9d8313952) *.* Diakses 15 Oktober 2020 pukul 19.35.
* Master\_abhig. (2020). Different Between K-Mean and Hierarchical. [*https://www.geeksforgeeks.org/difference-between-k-means-and-hierarchical-clustering/#:~:text=A%20hierarchical%20clustering%20is%20a,the%20clusters%20is%20hyper%20spherical*](https://www.geeksforgeeks.org/difference-between-k-means-and-hierarchical-clustering/#:~:text=A%20hierarchical%20clustering%20is%20a,the%20clusters%20is%20hyper%20spherical) . Diakses pada 16 Oktober 2020 pukul 15.00.