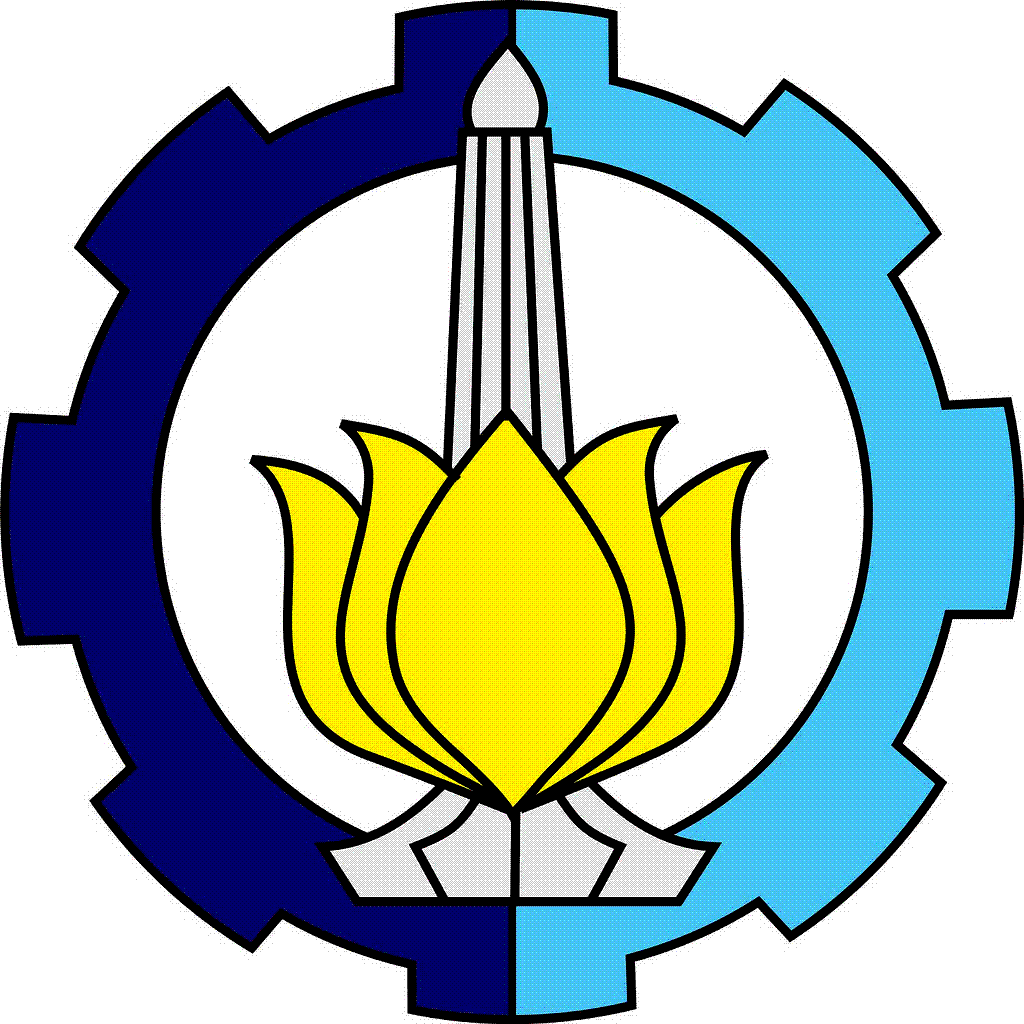
**Laporan Kecerdasan Komputasional**

**Tugas 3**

**Kelompok 7**



**Nama Kelompok :**

**Muhamad Haris Wicaksono (5111840000029)**

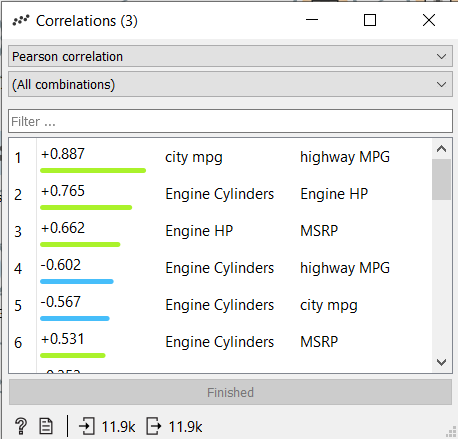
**M Ridho Daffa Ardista (5111840000065)**

**Kenji Hikmatullah (5111840000074)**

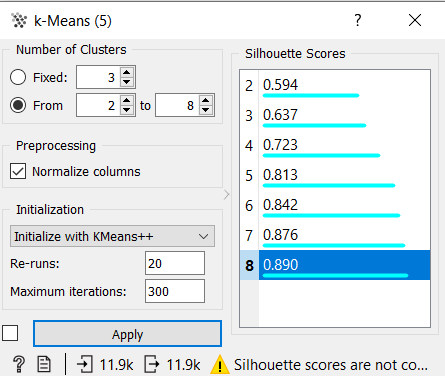
**1. Apa kombinasi fitur yang memberikan cluster terbaik menurut indikator *Silhouette score*?**

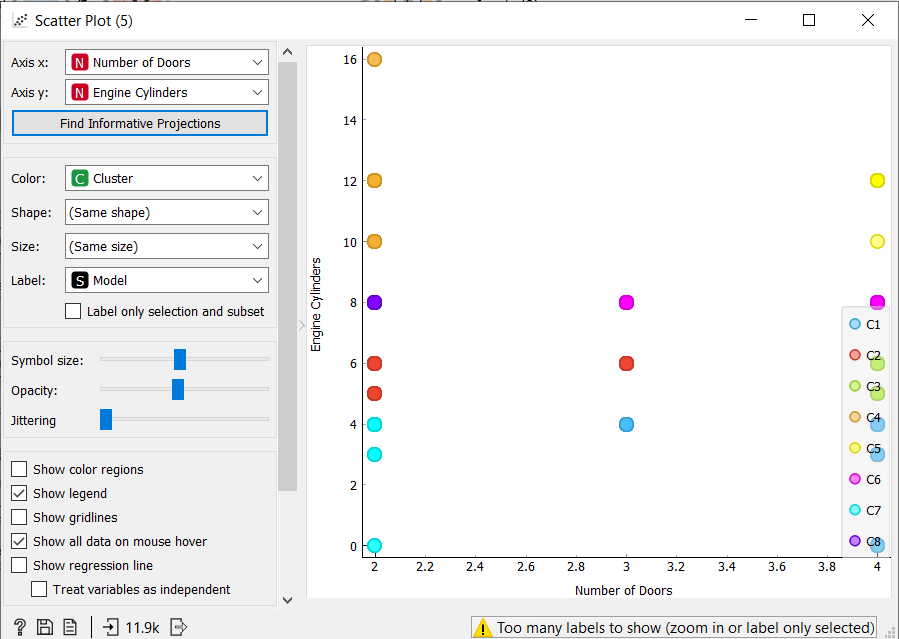
Dengan menggunakan aplikasi orange, disini akan membandingkan kombinasi fitur yang memberikan cluster terbaik menurut indikator Silhouette score. Silhouette score yang baik adalah yang nilainya mendekati 1.

Pertama kami melihat pada fungsi correlation pada orange sehingga dapat melihat korelasi terbesar antara fitur-fitur yang yang ada pada data cars. Nilai correlations tertinggi ada pada kombinasi fitur city mpg dan highway MPG

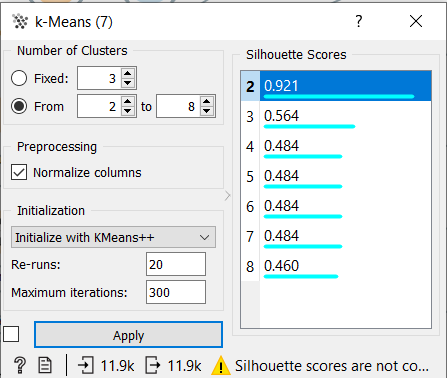
**

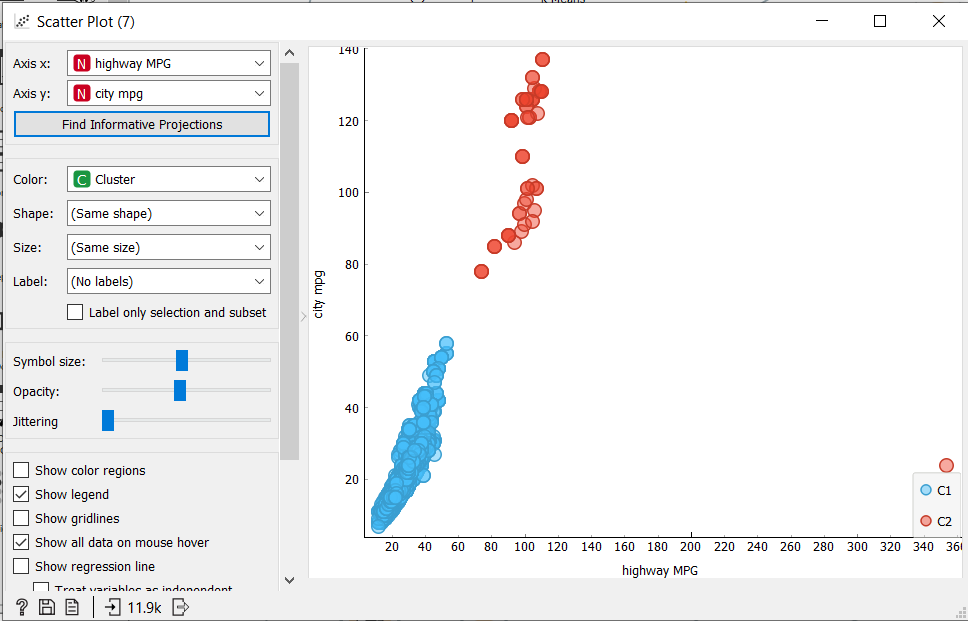
Pada fitur *Engine Cylinder* dan *Number of Doors*



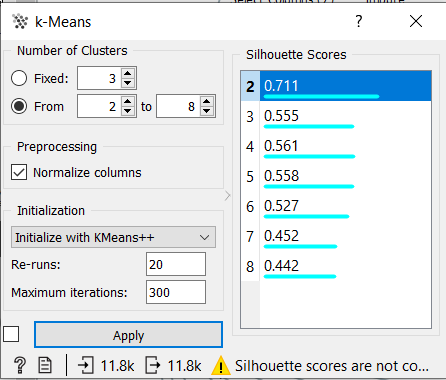


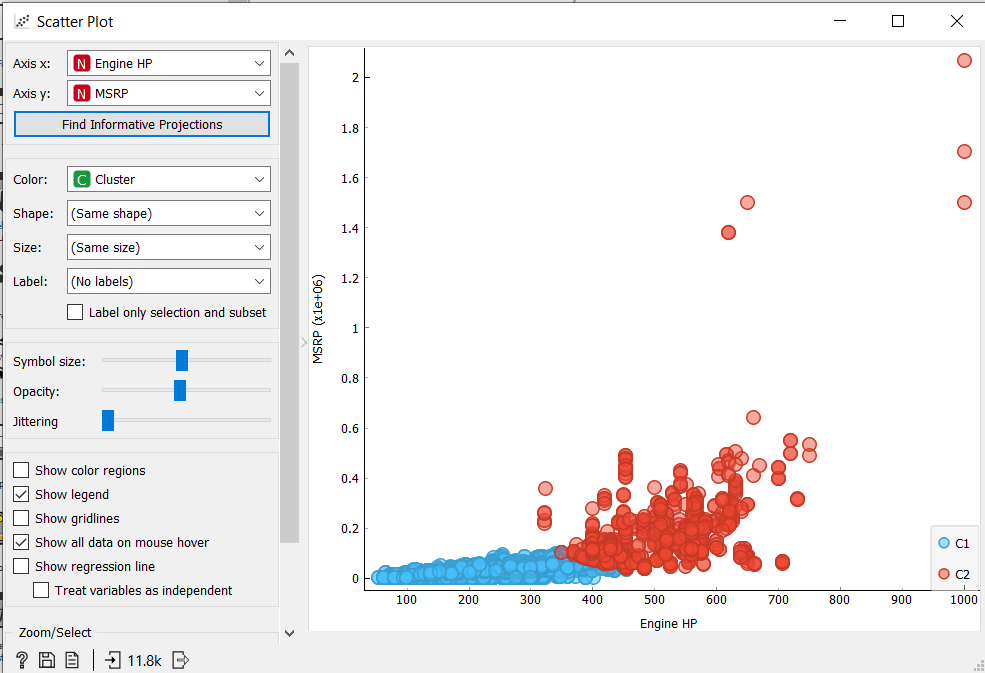
Pada fitur *city mpg* dan *highway MPG*





Pada fitur *Engine HP* dan *MSRP*

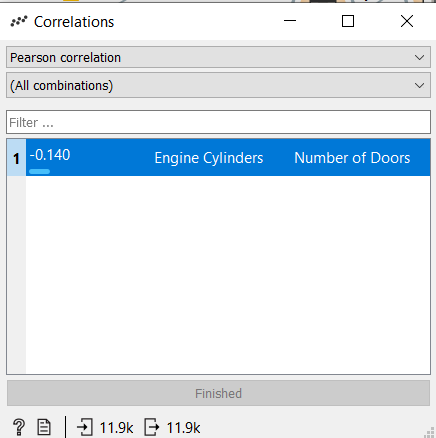


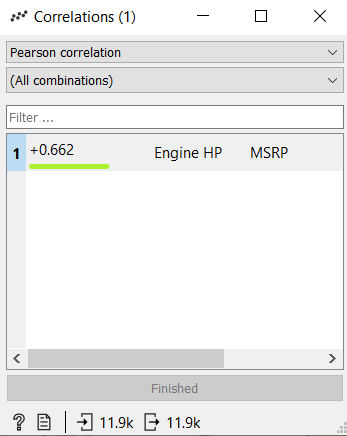
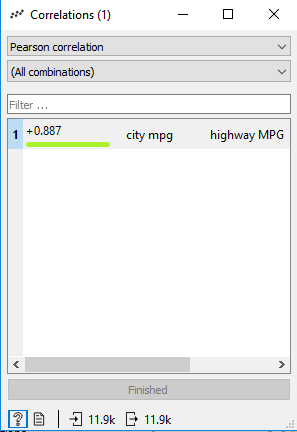


Setelah membandingkan dari berbagai fitur dalam beberapa percobaan yang ditampilkan pada laporan ini adalah nilai dari *silhouette score* yang paling baik dari percobaan diatas adalah antara *city mpg* dan *highway MPG*

**2. Bagaimana korelasi antara fitur-fitur yang dipilih?**

Dibawah ini adalah nilai-nilai correlations antara fitur-fitur yang kami pilih pada no 1. Kalo dilihat dari nilai-nilai di bawah hubungan korelasi antara fitur-fitur tidak ada, misalnya Engine Cylinders dan Number of Doors tidak ada hubungan satu sama lain sebenarnya





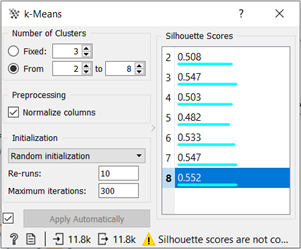
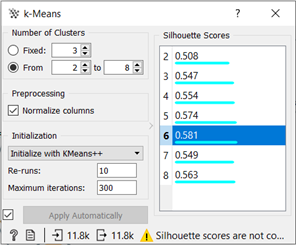
**3. Apakah ada proses *cleaning data* (impute, dll)?**

Ada sebagian yang tabel nya mengandung null value dan di situlah akan terjadi cleaning data, misalnya pada pengecekan fitur Engine cylinder dan Engine HP.. Untuk percobaan fitur yang lainnya tidak menggunakan cleaning data. Sebenarnya sendiri untuk di orange tidak diperlukan adanya cleaning data, karena tidak terlalu berpengaruh kecuali data yang kosong sangatlah banyak

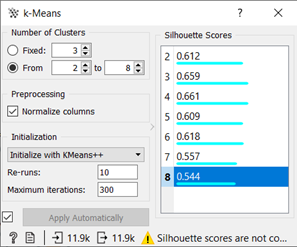
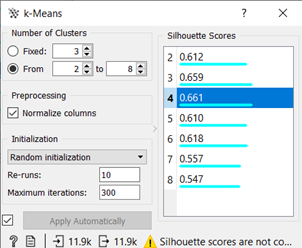
**4. Bagaimana hasil cluster jika centroid terbentuk secara random atau dipilihkan yang terbaik k means ++?**

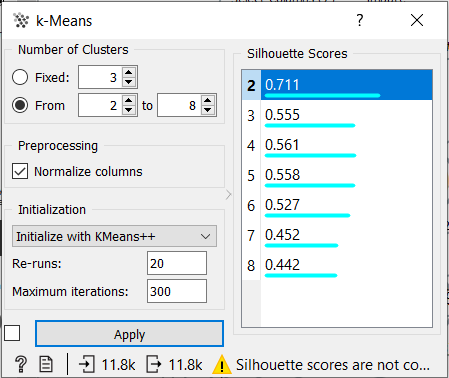
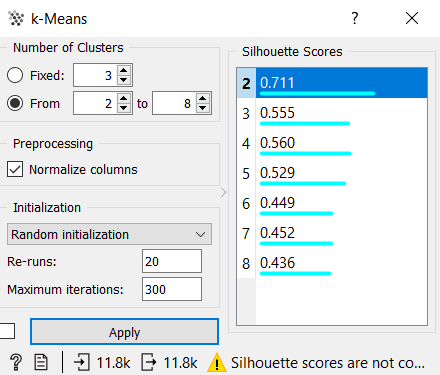
Pada poin kali ini akan melakukan percobaan untuk mengetahui hasil cluster jika centroid yang terbentuk secara random atau centroid yang terbentuk secara dipilihkan k means ++ . Dengan melakukan percobaan kolaborasi antara 2 fitur yang berbeda yang dimana untuk preprocessing menggunakan Normalize columns.

Pada fitur *Engine Cylinder* dan *Engine HP*



Pada fitur *MSRP* dan *Year*





Pada umumnya pada centroid yang terbentuk dari yang dipilihkan secara random dan menggunakan k-means ++ lebih baik yang menggunakan k-means ++.

**5. Bandingkan K Means dengan Hierarchical (divisive, agglomerative)!**

K Means clustering

Salah satu metode yang banyak digunakan dalam melakukan clustering dengan partisi ini adalah metode k-means. Secara umum metode k-means ini melakukan proses pengelompokan dengan prosedur sebagai berikut:

1. Tentukan jumlah cluster
2. Alokasikan data secara random ke cluster yang ada
3. Hitung rata-rata setiap cluster dari data yang tergabung di dalamnya
4. Alokasikan kembali semua data ke cluster terdekat
5. Ulang proses nomor 3, sampai tidak ada perubahan atau perubahan yang terjadi masih sudah dibawah treshold

Prosedur dasar ini bisa berubah mengikuti pendekatan pengalokasian data yang diterapkan, apakah *crisp* atau *fuzzy*. Setelah meneliti clustering dari sudut yang lain, saya menemukan bahwa *k-means clustering* mempunyai beberapa kelemahan.

Hierarchical clustering

Clustering dengan pendekatan hirarki mengelompokkan data yang mirip dalam hirarki yang sama dan yang tidak mirip di hirarki yang agak jauh. Ada dua metode yang sering diterapkan yaitu *agglomerative hierarchical clustering* dan *divisive hierarchical clustering*. Agglomerative melakukan proses clustering dari *N* cluster menjadi satu kesatuan cluster, dimana *N* adalah jumlah data, sedangkan divisive melakukan proses clustering yang sebaliknya yaitu dari satu cluster menjadi *N* cluster.

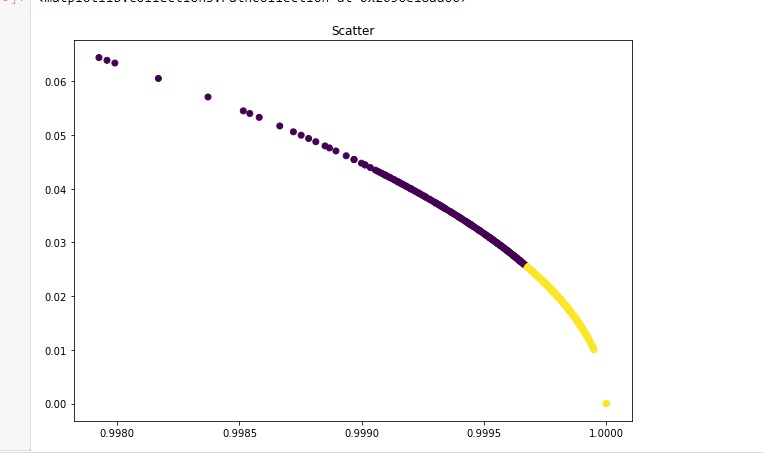
Beberapa metode *hierarchical clustering* yang sering digunakan dibedakan menurut cara mereka untuk menghitung tingkat kemiripan. Ada yang menggunakan *Single Linkage*, *Complete Linkage*, *Average Linkage*, *Average Group Linkage* dan lain-lainnya. Seperti juga halnya dengan *partition-based clustering*, kita juga bisa memilih jenis jarak yang digunakan untuk menghitung tingkat kemiripan antar data.

Salah satu cara untuk mempermudah pengembangan dendogram untuk hierarchical clustering ini adalah dengan membuat similarity matrix yang memuat tingkat kemiripan antar data yang dikelompokkan. Tingkat kemiripan bisa dihitung dengan berbagai macam cara seperti dengan Euclidean Distance Space. Berangkat dari similarity matrix ini, kita bisa memilih linkage jenis mana yang akan digunakan untuk mengelompokkan data yang dianalisa.

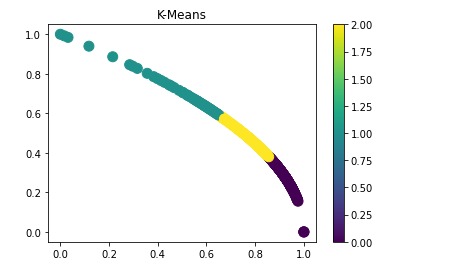
Waktu kompleksitas K means linier O(n) sedangkan Hierarchical O(n).

**6. Bandingkan dengan algoritma jika menggunakan SkLearn**

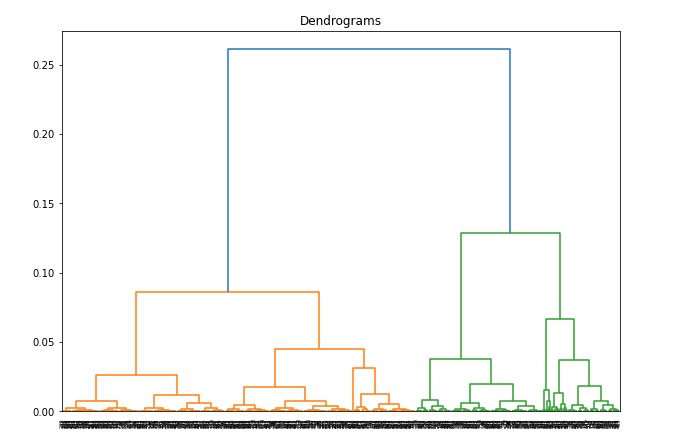
**Kelompok kami melakukan pengecekan pada fitur Engine HP dan Engine Cylinder**



Agglomerative Hierarchical Clustering



Scatter plot K Means



Dendogram untuk hierarchical clustering Agglomerative

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

Saat melakukan percobaan fitur fitur diatas kelompok kami menggunakan sklearn via orange dengan python script, dan hasilnya lebih cepat menggunakan k-means daripada yang hierarchical