Nama : Rizky Fauzi Ramadhani

Kelas : IF-42-10

NIM : 1301184144

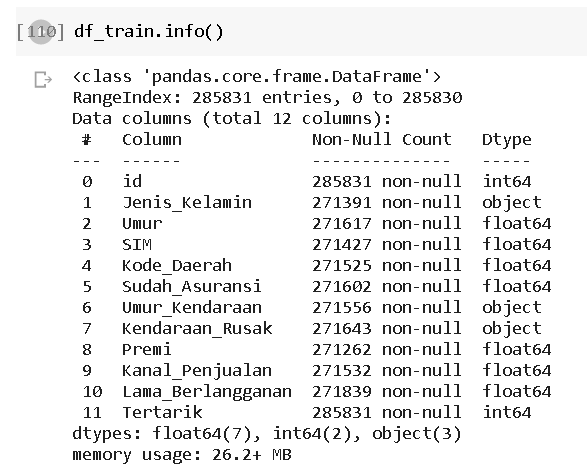
**Formulasi Masalah**

Terdapat 2 dataset yang diberikan yaitu, salju dan kendaraan. Dataset salju diminta untuk mencari tahu apakah besok akan turun salju atau tidak, sedangkan dataset kendaraan diminta untuk mencari tahu apakah pembeli tertarik untuk membeli kendaraan baru atau tidak. Dataset tersebut digunakan dengan menggunakan clustering K-Mean. Dataset yang akan digunakan yaitu dataset kendaraan untuk mencari tahu apakah pembeli tertarik untuk membeli kendaraan baru atau tidak.

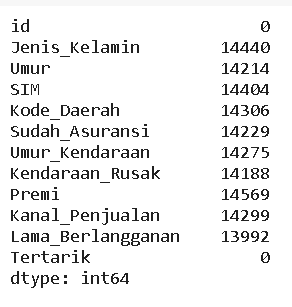
**Eksplorasi dan Persiapan Data**

Saya menggunakan dataset Kendaraan, kemudian eksplorasi data yang saya lakukan adalah dengan melihat info data, penguraian data agar dapat melihat count, mean, std, min, max, dll. Kemudian saya mengecek missing values pada dataset, kemudian saya melakukan preprocessing pada data dengan mendrop/menghapus baris yang memiliki isi yang kosong, agar nanti dataset dapat diproses dengan sangat baik, kemudian saya melakukan encode data untuk mengubah dataset yang bersifat objek agar menjadi numerik agar lebih mudah diproses dengan menggunakan library label encoder, sehingga mengubah data yang awalnya pria, wanita menjadi 0 dan 1 kemudian encode umur\_kendaraan dan juga encode kendaraan\_rusak. Kemudian saya mengecek korelasi heatmap dataset dengan tujuan untuk mengambil 2 kolom yang akan dicluster, dengan melihat warna yang lebih terang di luar dari yang bernilai 1 itu merupakan korelasi heatmap yang baik untuk di cluster, dan saya mendapatkan korelasi yang bagus yaitu kendaraan\_rusak dan sudah\_asuransi dengan hasil 0.83. Selanjutnya sebelum ke pemodelan saya melakukan terlebih dahulu normalisasi pada dataset dengan tujuan mengubah nilai dari range 0 – 1 agar tidak terjadinya timpang tindih data.

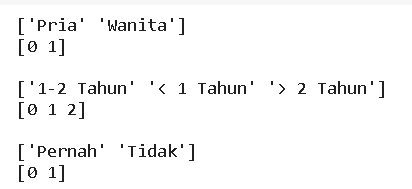
Info dataset



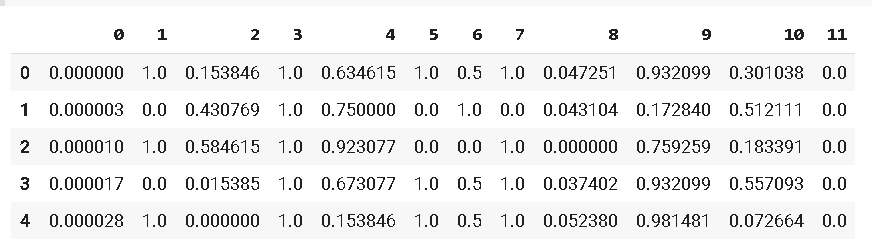
Sum missing value



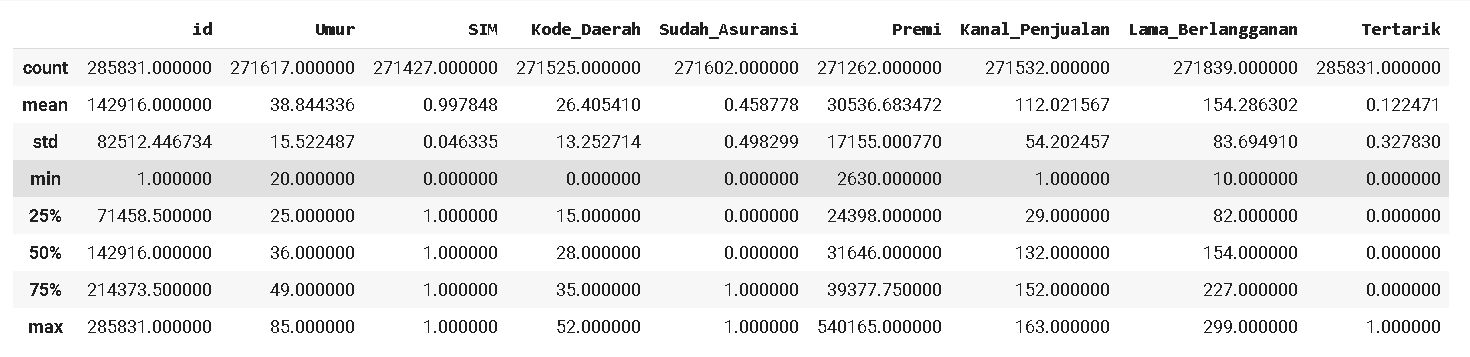
encode



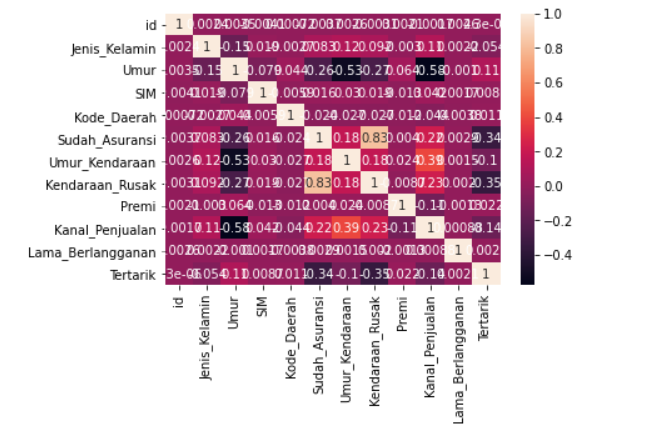
normalisasi



Dataset describe



Correlation

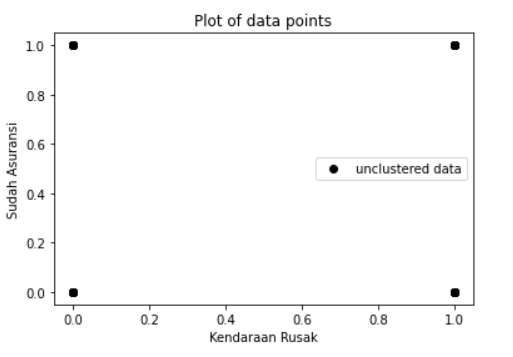


**Pemodelan**

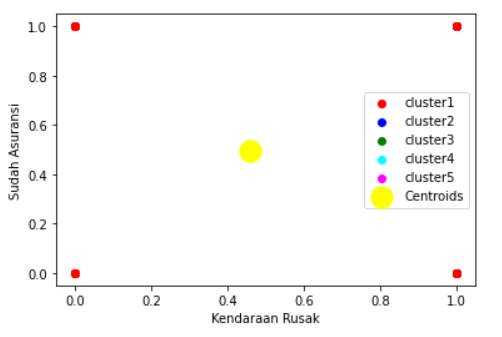
Pertama, dilakukan terlebih dahulu encode untuk mengubah nilai pada kolom yang non-numerik menjadi numerik. Kedua, dilakukan normalisasi untuk menghindari terjadinya tumpang tindih data dengan mengubah nilai dan range data dari kisaran 0 – 1. Ketiga, melakukan clustering menggunakan algoritma K-mean dengan memetakan data dengan 2 kolom berdasarkan korelasi yang didapat (kendaraan\_rusak dan sudah\_asuransi) sehingga membentuk node-node. Keempat, inisialisasi secara acak pusat dari cluster. Dengan ansumsi awal nilai K = 5 yang saya ambil, jadi kita memilih 5 titik data secara acak sebagai centroid. Kemudian untuk setiap titik data dihitung jarak Euclid dari semua centroid dengan rumus sebagai berikut :

Kemudian menetapkan cluster berdasarkan jarak minimal ke semua centroid. Kemudian mengambil setiap titik hitam, lalu menghitung jarak eucliddiannya dari semua centroid. Dan kemudian mewarnai titik hitam dengan warna titik terdekatnya. Kelima, menyesuaikan pusat massa setiap cluster dengan mengambil rata-rata dari semua titik data yang termasuk dalam cluster tersebut, setelah itu menghitung mean dari semua cluster individu untuk menetapkan semua titik data ke salah satu cluster. Sehingga didapatkan hasil cluster dari dataset kendaraan berdasarkan algoritma K-mean.

Sebelum clustering :

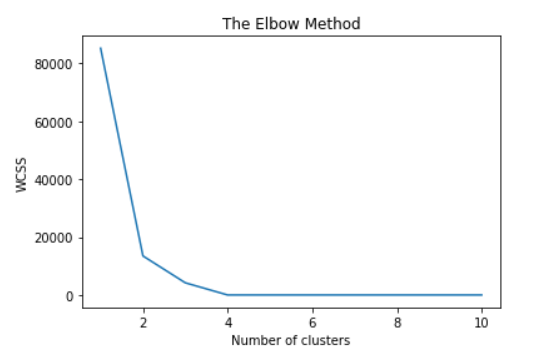


Sesudah clustering :



**Evaluasi**

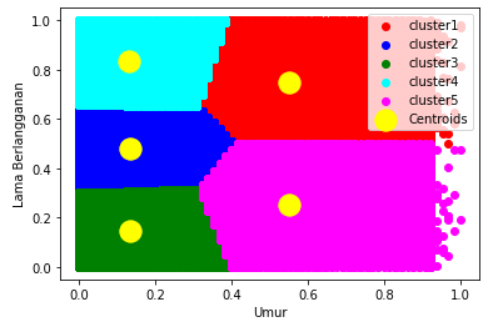
Evaluasi yang saya lakukan adalah dengan Teknik siku (elbow method). Langkah-langkah yang dilakukan oleh elbow method yaitu pengelompokan pada nilai K yang berbeda mulai dari range 1- 10. Untuk setiap K, dihitung WCSSnya. Kemudian memplot nilai WCSS sesuai dengan jumlah kluster. Gambar pada plot yang terdapat siku/elbow merupakan cluster yang paling sesuai. Yang dimana saya mengambil nilai K = 5 pada elbow method menunjukkan bahwa nilai K = 3 lebih optimal.



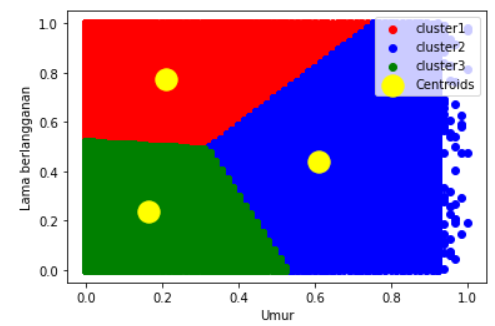
**Eksperimen**

Eksperimen yang saya lakukan mengubah nilai K dan mengubah kolom yang akan dicluster, dan mendapatkan hasil sebagai berikut :

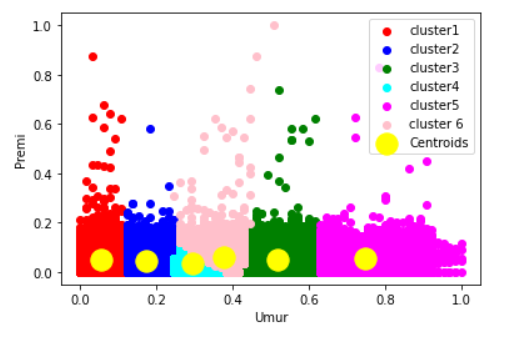
Nilai K = 5 , Kolom Umur dan Lama Berlangganan

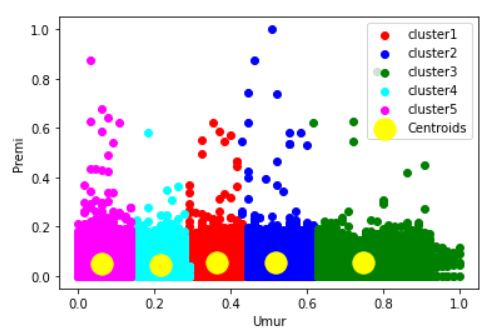


Nilai k = 3, kolom umur dan lama berlangganan



Nilai k = 6, kolom umur dan premi



Nilai k = 5, kolom umur dan premi

**Kesimpulan**

Kesimpulan yang saya dapatkan berdasarkan eksperimen clustering yang saya lakukan, disimpulkan bahwa kolom dengan nilai yang unik akan lebih optimum yaitu adalah kolom umur dan lama berlangganan.

**Link Video Presentasi**

<https://youtu.be/cH7H6UgHdik>

**Link Colab**

<https://colab.research.google.com/drive/18VEihICP2PE_BreZjiGnIoCBo96sVv3L?usp=sharing>