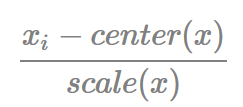
**Eksplorasi Data**

**Penghapusan Fitur**

Dataset ini memiliki 2 fitur yang seluruh nilai elemennya adalah “ - ”, yakni Jahe dan Kapulaga. Karena isi dari fitur tersebut adalah “ - ” dan diasumsikan tidak dilakukan pengukuran pada data tersebut, maka fitur tersebut dihapus dari dataset.

**Pra-Proses**

Dataset tersebut memiliki banyak fitur sehingga hanya beberapa fitur saja yang akan digunakan dalam proses clustering. Supaya dataset tersebut dapat digunakan secara baik, dilakukan proses *scaling* dengan rumus seperti berikut:



Keterangan:

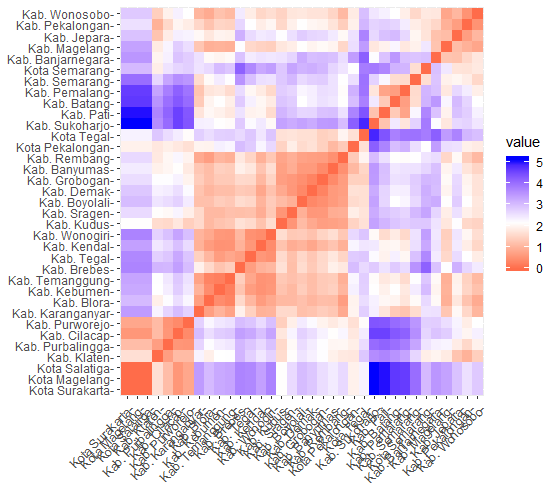
Xi = data dari fitur X

Center(x) = nilai tengah dari fitur x

Scale(x) = nilai ragam dari fitur x (standar deviasi)

**Menghitung Dissimilarity Matrix**

Untuk mengetahui seberapa besar ketidakmiripan antara satu objek dengan objek lainnya, digunakan *Dissimilarity Matrix*.

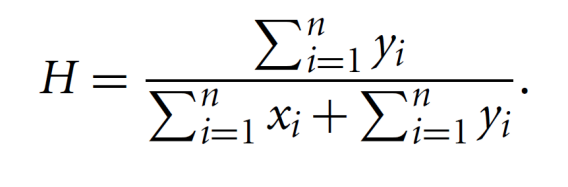


Jika kotak tersebut bewarna biru maka ketidakmiripan antar objek tersebut sangat besar dan dipastikan tidak berada dalam satu *cluster* yang sama. Sebaliknya, jika kotak bewarna merah maka ketidakmiripan antar kedua objek dalam kotak tersebut sangat kecil sehingga dapat dikatakan bahwa kedua objek tersebut berada di dalam satu *cluster.*

**Hopkins Statistic Test**

*Hopkins Statistics* adalah salah satu metode yang digunakan untuk menentukan apakah *dataset* yang di-*cluster* tersebut memiliki *cluster* yang lebih bermakna dengan cara menguji keacakan dari data tersebut.

Untuk rumus dari Hopkins Statistic (H) adalah:



Keterangan:

Xi = Jarak minimum antara objek pi dengan objek p yang lain, dengan p adalah himpunan objek yang memiliki sebaran normal di dataset

Yi = Jarak minimum antara objek qi dengan objek q yang lain, dengan q adalah himpunan objek yang memiliki sebaran normal di dataset

H = Nilai *Hopkins Statistic*

Dalam pengujian tersebut, terdapat hipotesis null yang menyatakan jika dataset tersebut menyebar normal dan tidak memiliki *cluster* yang bermakna. Sedangkan hipotesis alternatif nya adalah dataset tersebut tidak terdistribusi secara normal dan memiliki *cluster* yang bermakna. Jika nilai H > 0.5, maka hipotesis null dapat ditolak sehingga dataset tersebut bisa dikatakan memiliki *cluster* yang bermakna.

Untuk dataset ini, nilai dari *Hopkins Statistic* nya adalah:



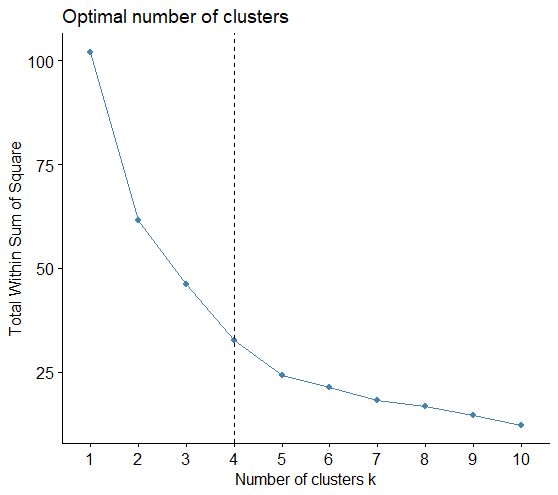
Karena nilai dari H > 0.5, maka hipotesis null dapat ditolak sehingga *cluster* tersebut memiliki makna.

**Penentuan Jumlah Cluster**

Setelah fitur tersebut dihapus dari dataset, maka dilakukan penentuan jumlah optimal banyaknya *cluster* (k) terhadap dataset tersebut. Metode yang digunakan adalah *elbow method.*

*(*yang mau disitasi*)*

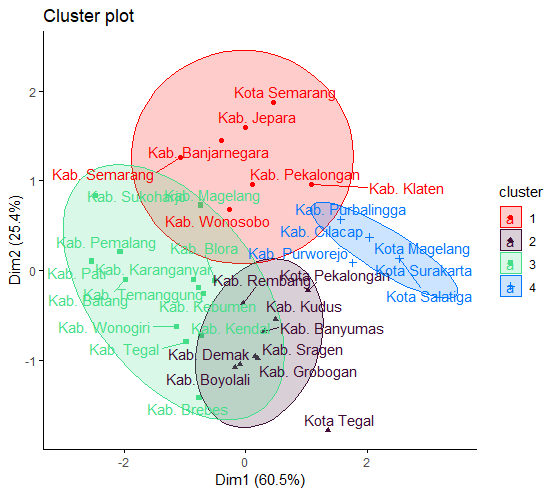
***Elbow Method* adalah suatu metode untuk menentukan jumlah cluster optimal dengan melihat persentase ragam terhadap banyaknya jumlah cluster.**



Sesuai dari grafik tersebut, k yang diambil untuk *clustering* adalah 4 karena dilihat dari kemiringan, k yang lebih dari 4 tidak memiliki perubahan yang signifikan sehingga k tersebut tidak diambil.

**Pembentukan Cluster**

Setelah nilai k telah ditemukan, proses clustering dilakukan dengam algoritma *partitioning around medoids* (PAM).



Dari grafik tersebut, didapatkan 4 *cluster* yang mengelompokkan seluruh kabupaten dan kota di Jawa Tengah. Dengan keterangan cluster 1 sangat produktif, cluster 2 cukup produktif, cluster 3 kurang produktif, dan cluster 4 tidak produktif.