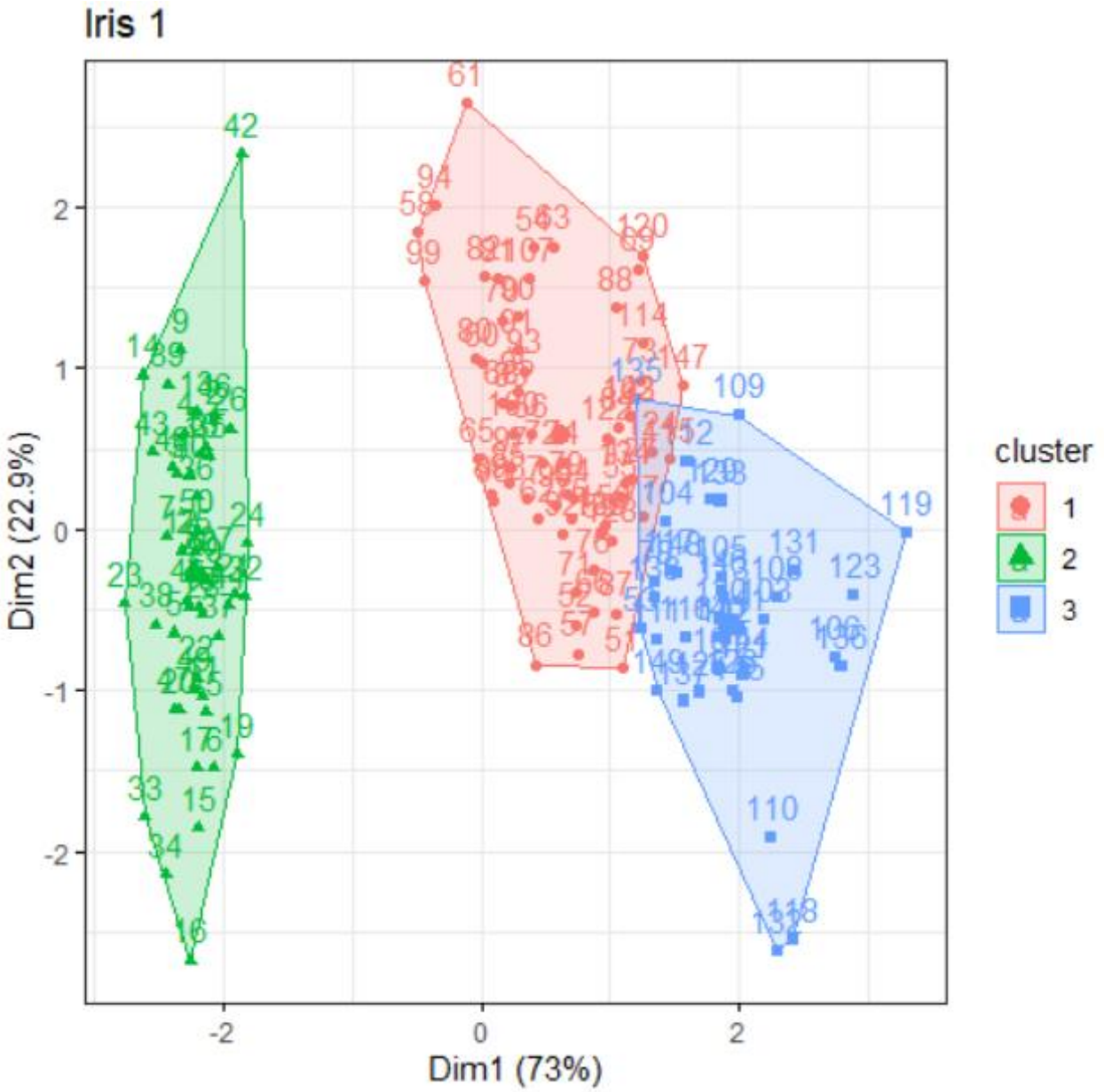
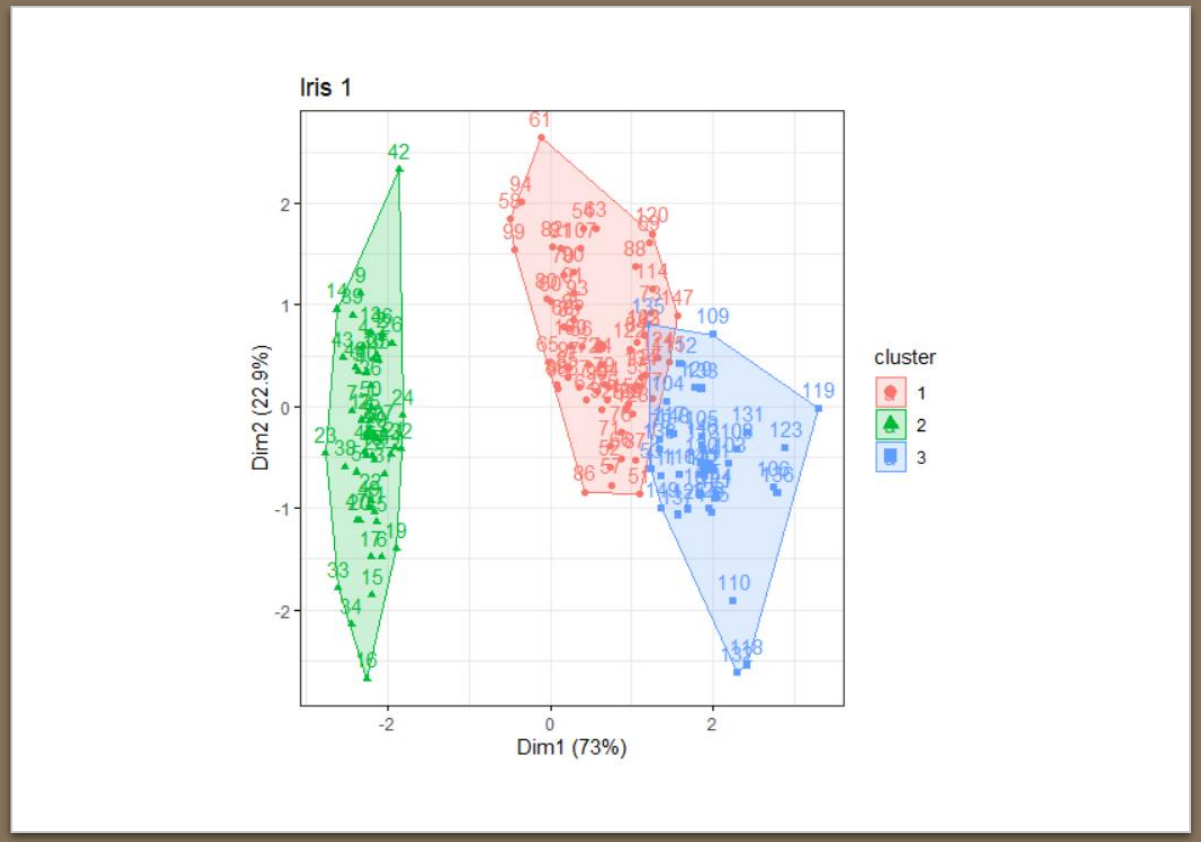
Agrupamentos

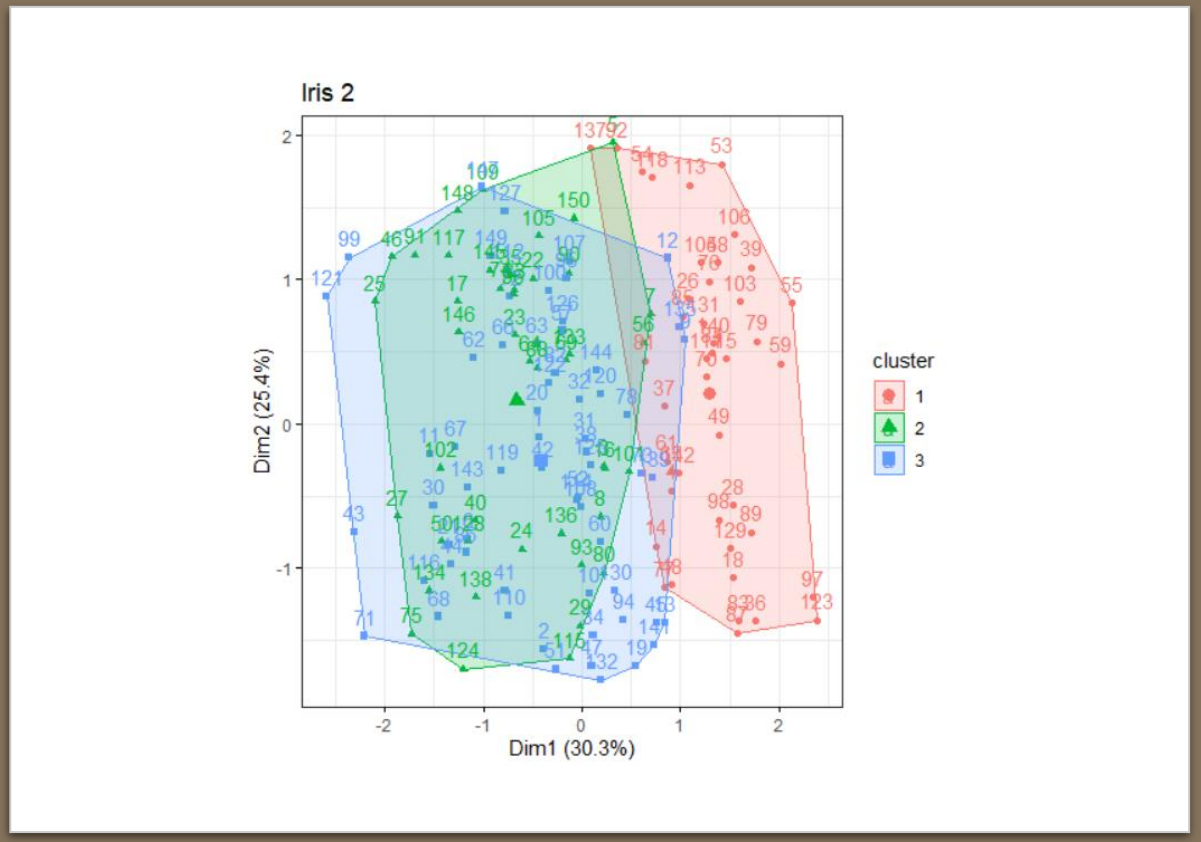


Perguntas:

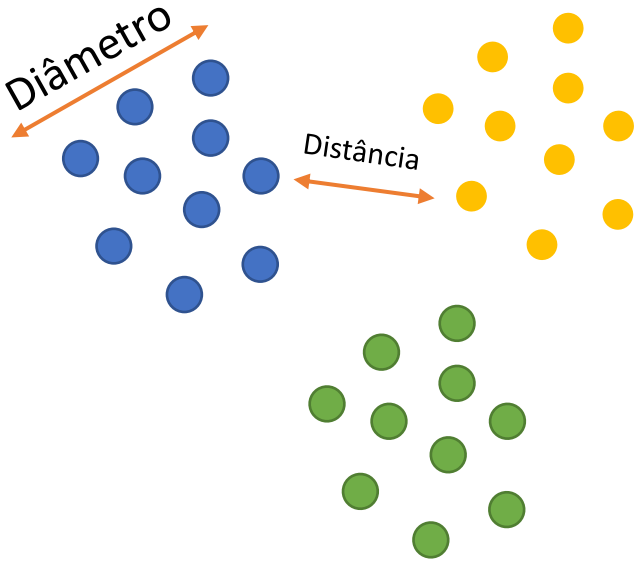
1. De fato existem Clusters nos dados?
   - Sabemos que K-means irá agrupar sempre os dados
2. O número ideal de clusters é de fato 3?
   - Usamos 3 porque conhecemos os labels!
3. Foram produzidos bons clusters?
4. Usamos o melhor agrupador?

Produzindo um cluster

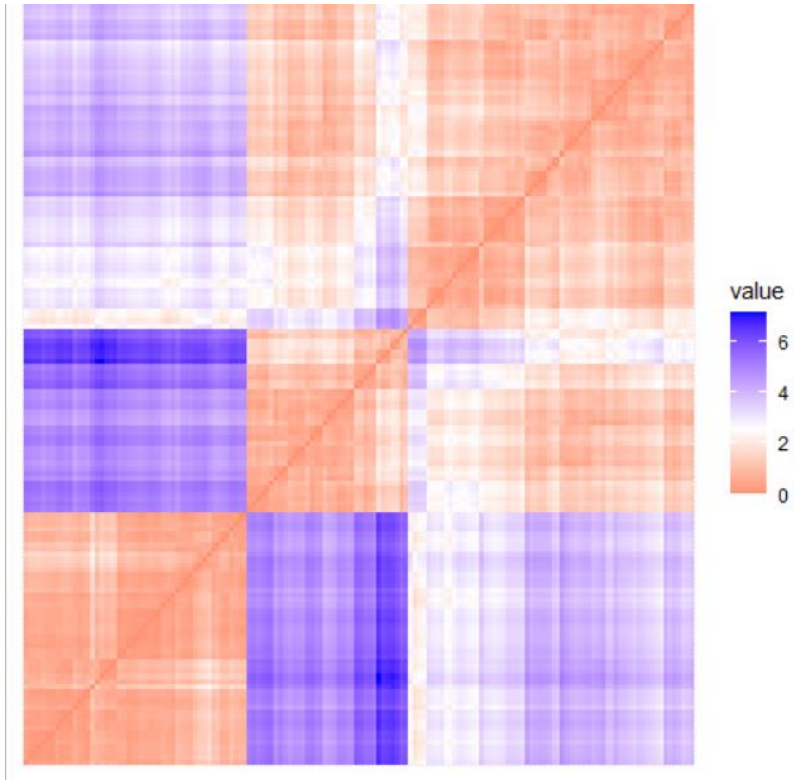




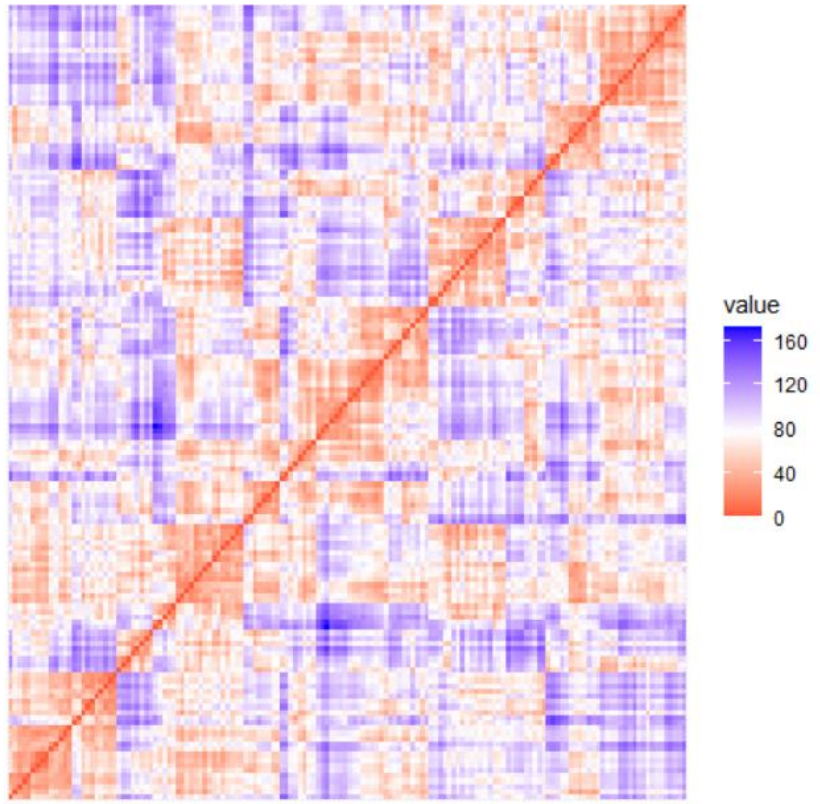
1.De fato existem Clusters nos dados?



1.De fato existem Clusters nos dados?

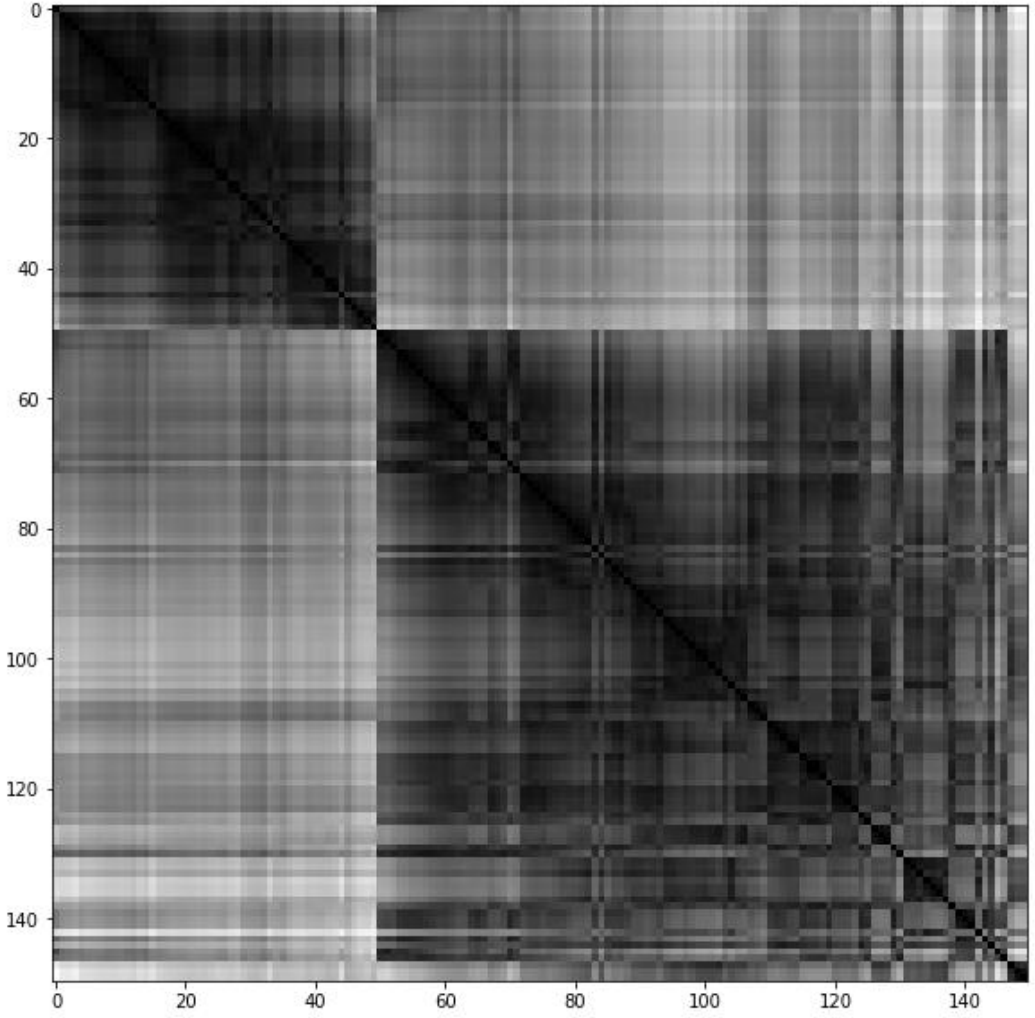


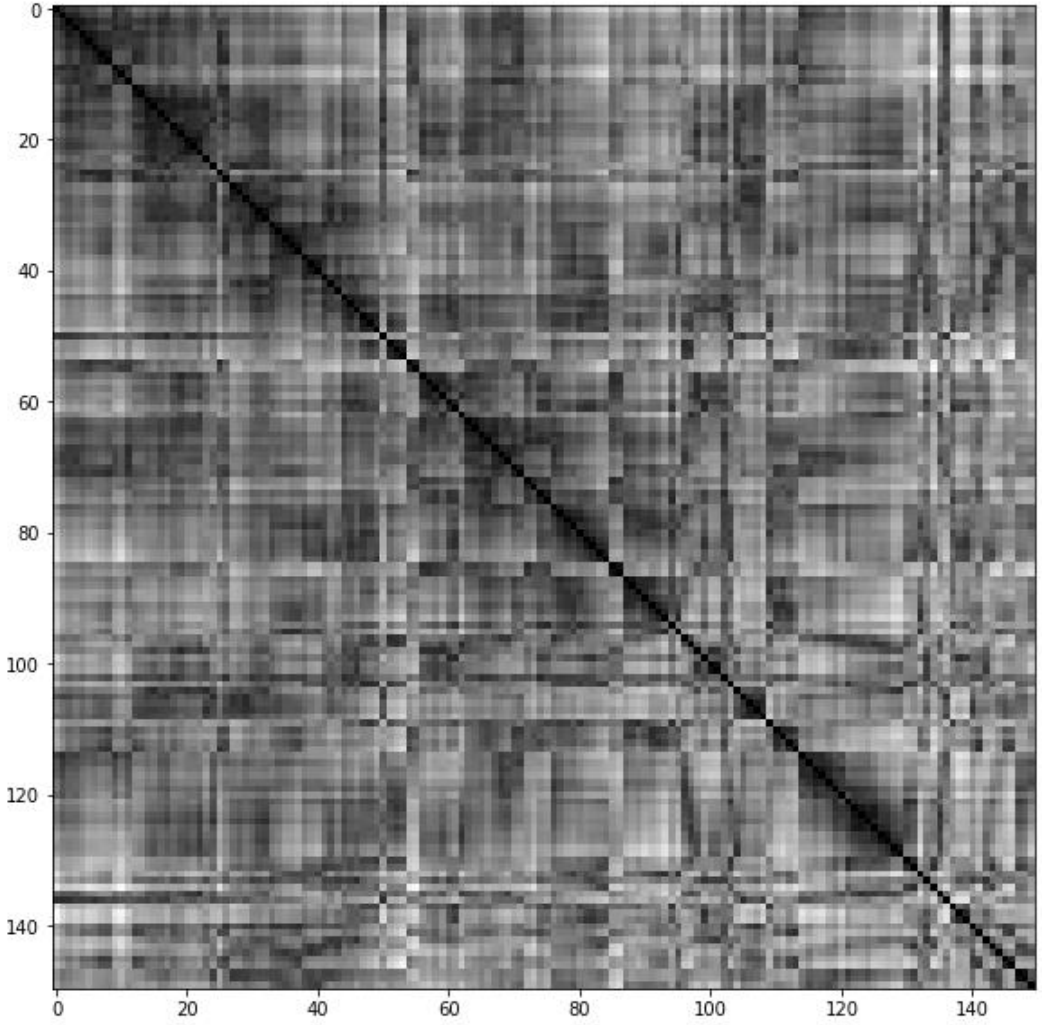
H 0.8191482



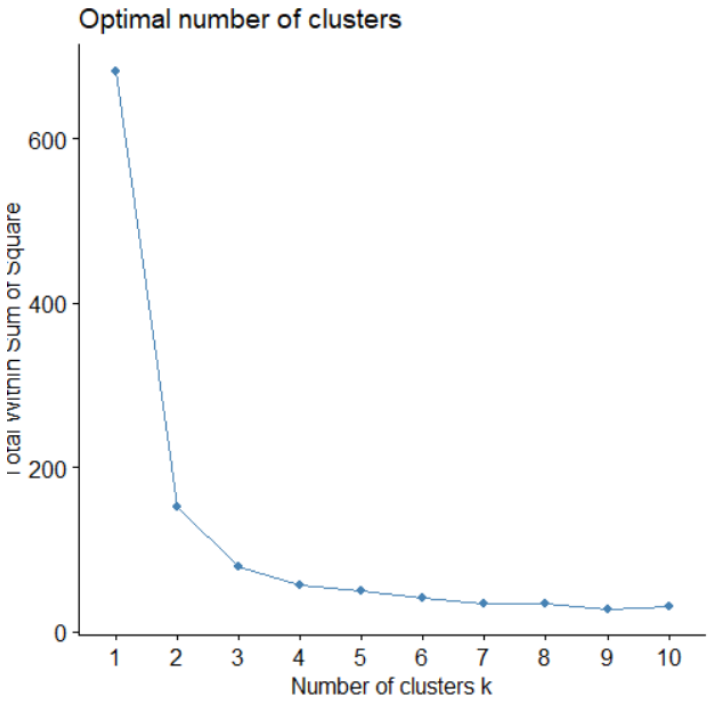
H 0.4814079

1.De fato existem Clusters nos dados?

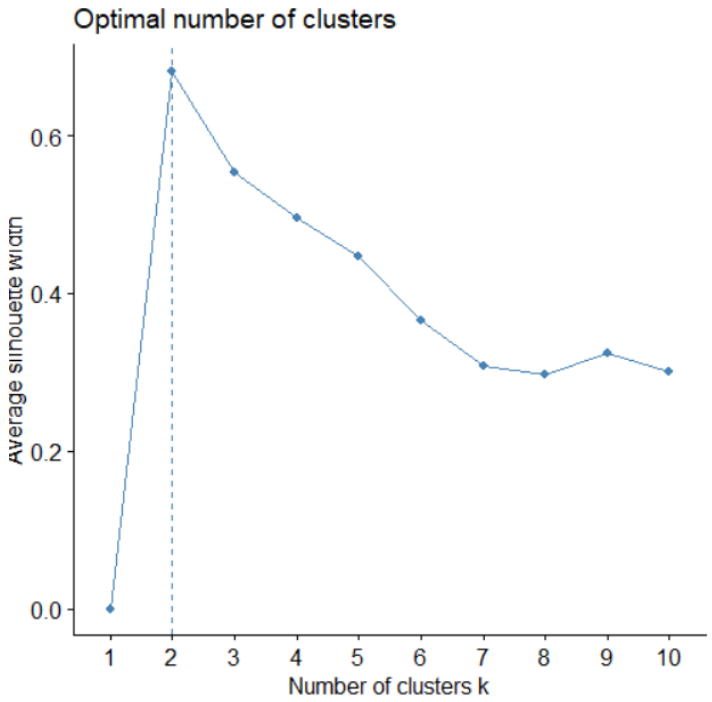




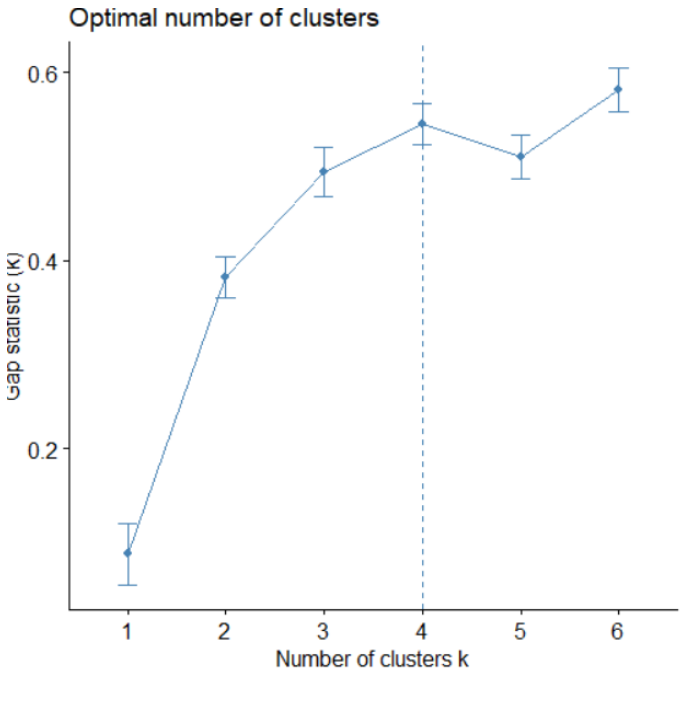
2. Qual o número ideal de clusters?



Elbow

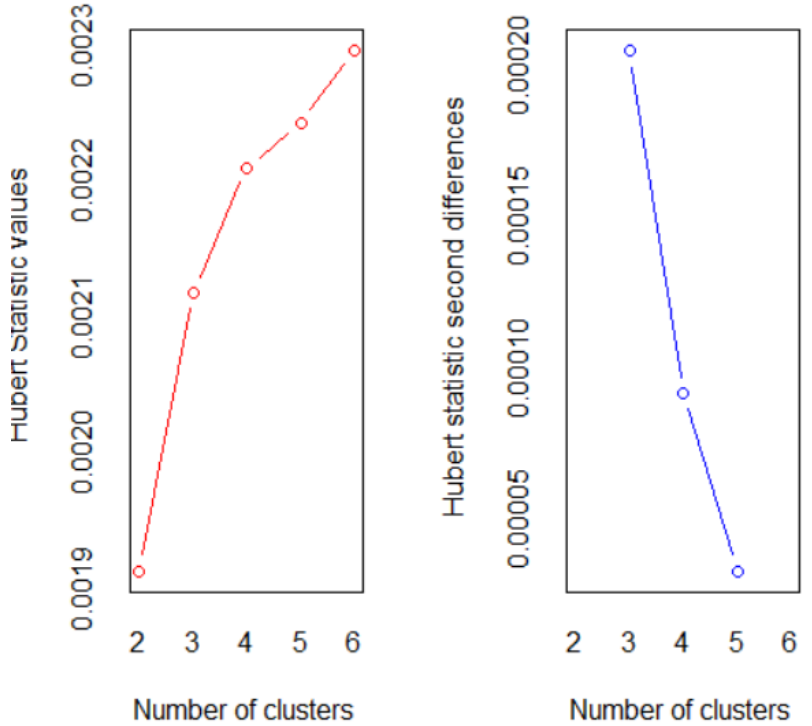


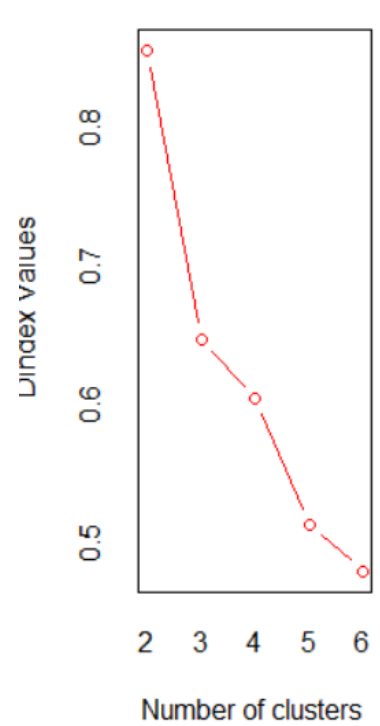
Average silhouette

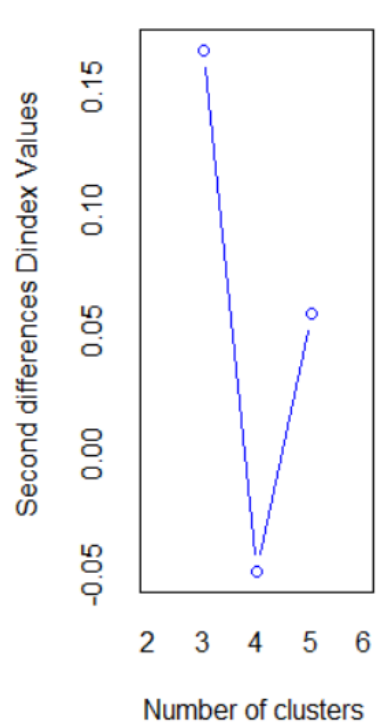


Gap

Hubert index e D index







3. Foram produzidos bons clusters?

- Em um bom cluster:
  - Diâmetro do cluster pequeno
  - Distância entre os clusters deve ser grande
- Índice "Dunn" mede a qualidade do cluster
- Busca-se maximizar este índice

4. Usamos o melhor agrupador?

```
Validation Measures:
                                 2        3        4        5        6

hierarchical Connectivity   0.0000   4.4770   8.9929 15.4893 18.4183
             Dunn           0.3389   0.1378   0.1540   0.1540   0.1668
             Silhouette     0.6867   0.5542   0.4720   0.4307   0.3420
kmeans       Connectivity   6.1536 10.0917 17.5194 27.9373 36.4873
             Dunn           0.0765   0.0988   0.1365   0.0823   0.0853
             Silhouette     0.6810   0.5528   0.4981   0.4887   0.3648
pam          Connectivity   3.9623 10.0917 24.1675 28.5984 43.8631
             Dunn           0.0811   0.0988   0.1008   0.1235   0.0921
             Silhouette     0.6858   0.5528   0.4897   0.4867   0.4704

Optimal Scores:

             Score  Method       Clusters
Connectivity 0.0000 hierarchical 2
Dunn         0.3389 hierarchical 2
Silhouette   0.6867 hierarchical 2
```