Opdracht 6 - K-Means

Aymane Machrouki en Berke Ozmuk

1. Call tree

main()

-----importData()

-----createLabels()

-----normalizeData()

-----kMeans()

----------centroidAvg()

----------getCentroids()

-----assignLabels()

-----countLabels()

-----createSkreePlot()

2. Totaal-uitleg

Als eerste wordt de data uit dataset1.csv geimporteerd door **importData()** , waarna het direct wordt genormaliseerd met **normalizeData()** Hierna wordt de **kMeans()** functie aangeroepen met de gekozen k. In de functie worden random centroids gekozen waarvan het aantal besloten wordt door de meegegeven k door de **getCentroids()** functie aan te roepen. Na de centroids te hebben gekozen wordt er door alle datapunten gelooped en wordt er gechecked welke centroid het dichtste bij is. Als alle datapunten zijn gechecked, dan worden ze in clusters opgedeeld. Aan het eind van de functie wordt het gemiddelde van elk cluster berekend met behulp van de **centroidAvg()** functie, waarna deze gemiddelden gebruikt worden als de nieuwe centroids. Na dit te hebben gedaan start de funtie opnieuw op, met de nieuwe centroids in de plaats van de oude. De functie blijft dit herhalen totdat er geen vernadering meer is. Nadat de functie is afgerond wordt de **assignLabels()** functie aangeroepen met de resulaten van de **kMeans()** functie. De **assignLabels()** functie voegt het bijbehorende label toe aan elk datapunt. Hierna wordt er geteld welke seizoenen er in elk cluster voorkomen met de **countLabels()** functie. Als laatste wordt de **createSkreePlot()** functie aangeroepen om een screeplot te maken van de meegeleverde k.

3. Resulaten

Gegeven dat k 4 is, geven de resultaten aan dat...

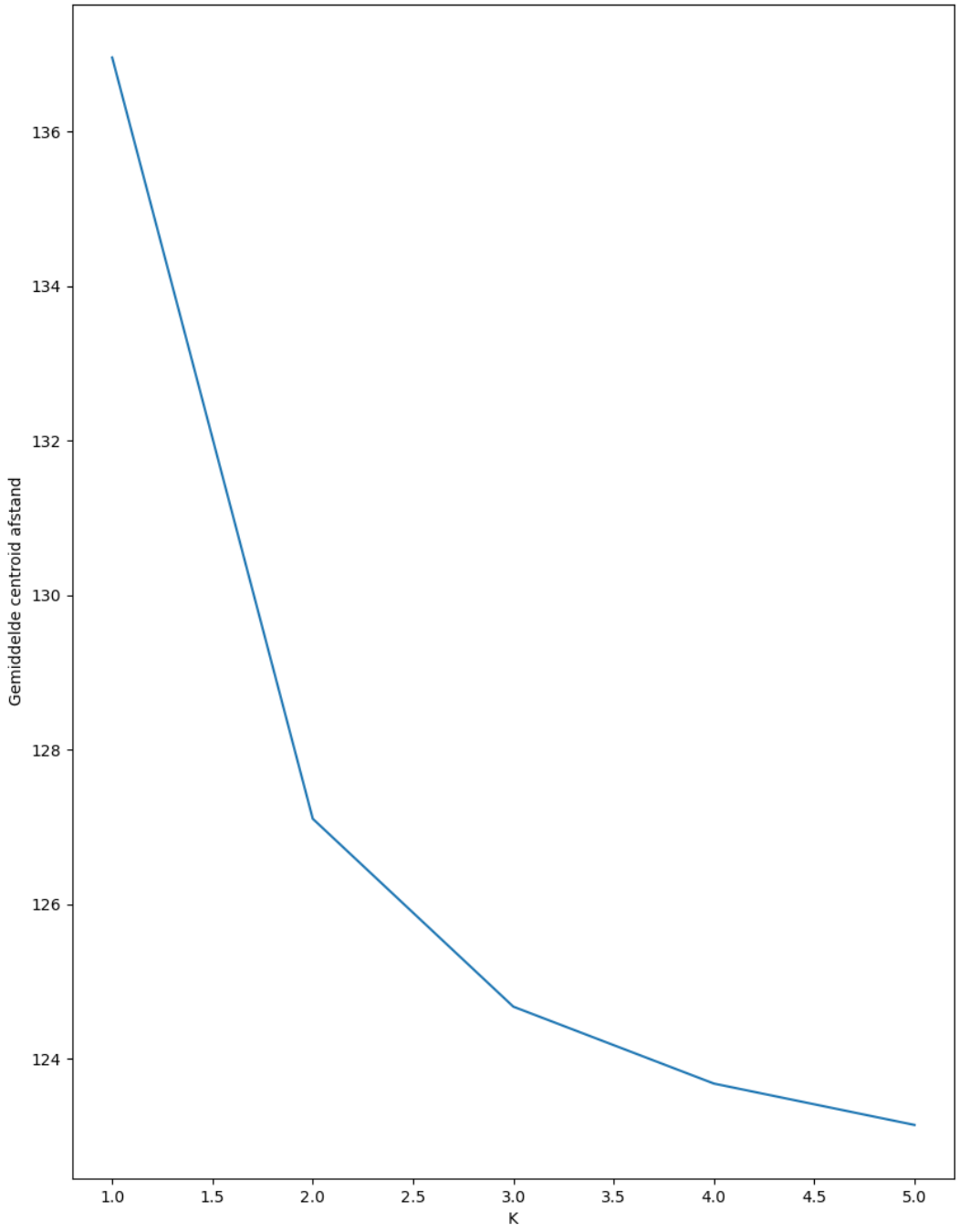
...cluster 1 53 winters, 21 lentes, 0 zomers en 3 herfsten bevat

...cluster 2 0 winters, 14 lentes, 38 zomers en 8 herfsten bevat

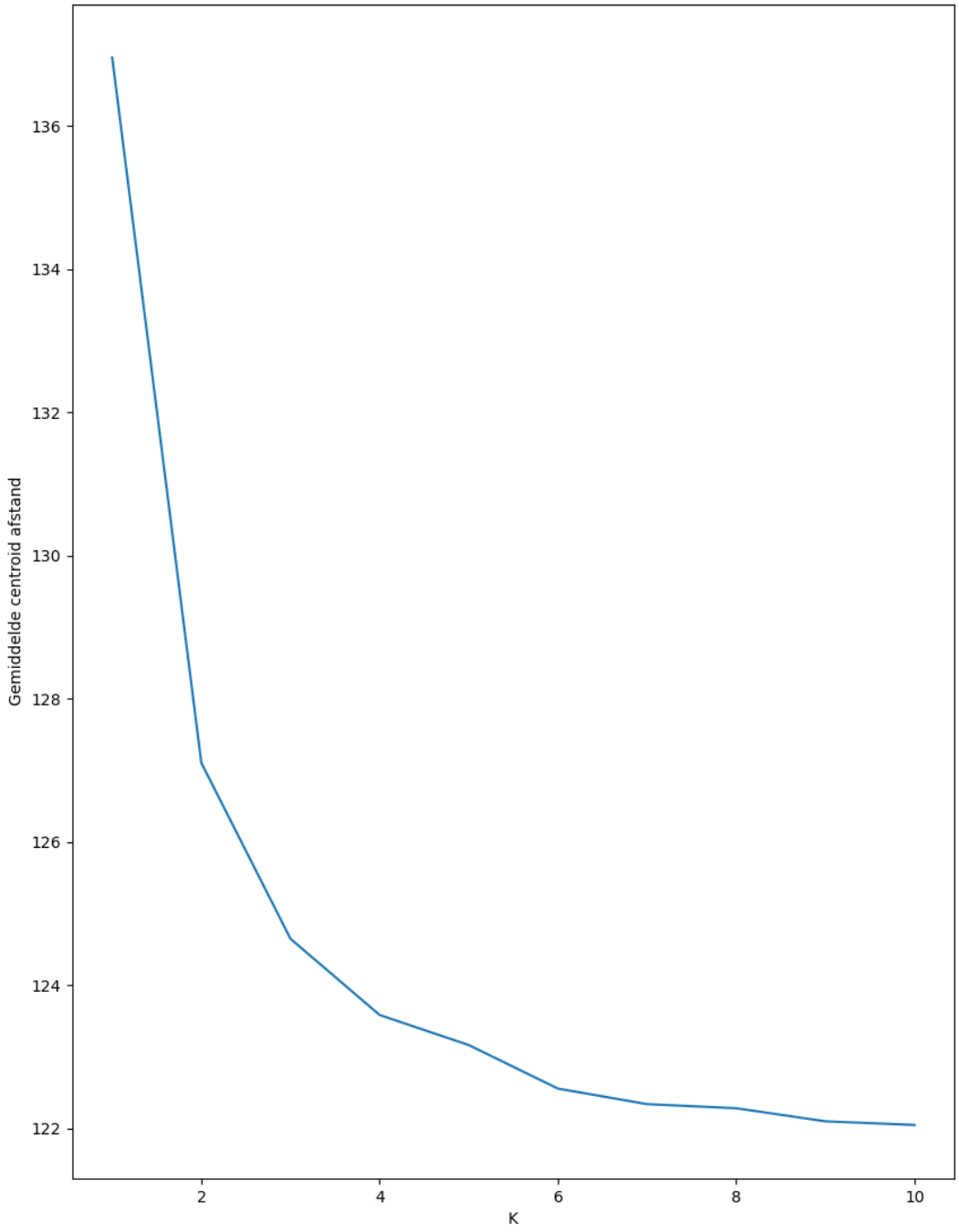
...cluster 3 30 winters, 36 lentes, 7 zomers en 39 herfsten bevat

...cluster 4 8 winters, 21 lentes, 47 zomers en 41 herfsten bevat

Op de volgende paginas zijn 2 skree plots te zien.



Bij deze skree plot is er een k van 5 gekozen.



Bij deze skree plot is er een k van 10 gekozen.

4. Antwoorden

Hoeveel clusters kan je (betrouwelijk) detecteren?

Door de twee skree plots met elkaar te vergelijken is te zien dat na een k van 6 en 7 de plot al begint af te vlakken. Daarom zou een k van 6 optimaal moeten zijn.