Eindopdracht

Mohammed Al Hor

2022-12-31

Opdracht 1: World Values Survey

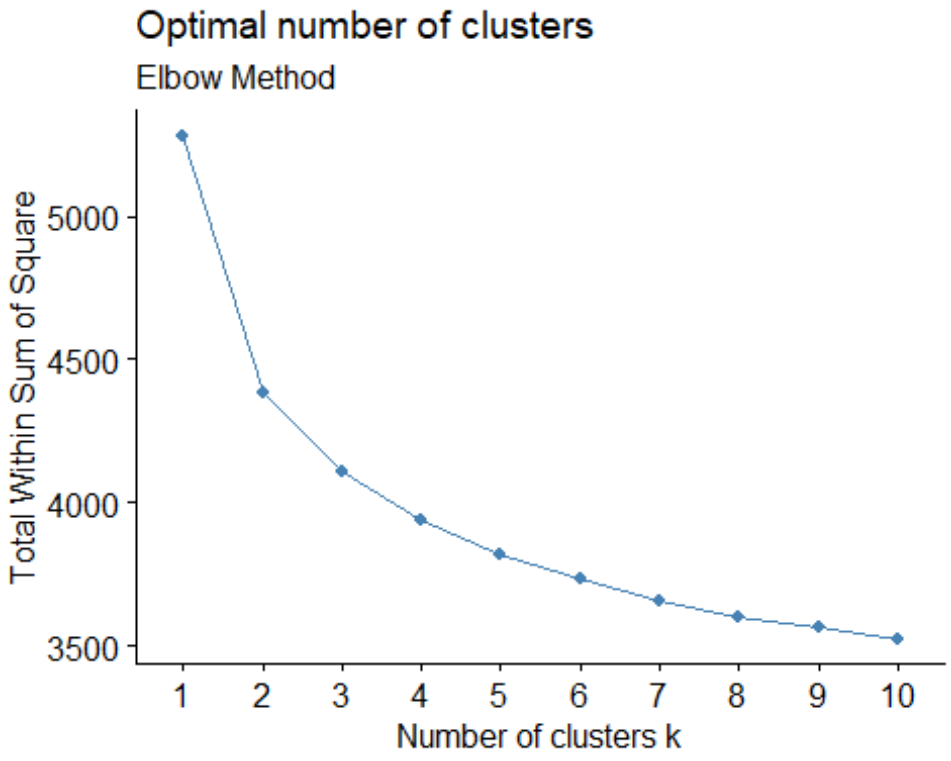
Probeer door middel van cluster analyse groepen van mensen te vinden die vergelijkbare antwoorden hebben gegeven op de 45 vragen. Normaliseer de data. (Dus: Herschaal alle antwoorden zodat deze van 0 -1 lopen).

a. Normaliseer de data. (Dus: Herschaal alle antwoorden zodat deze van 0 -1 lopen).

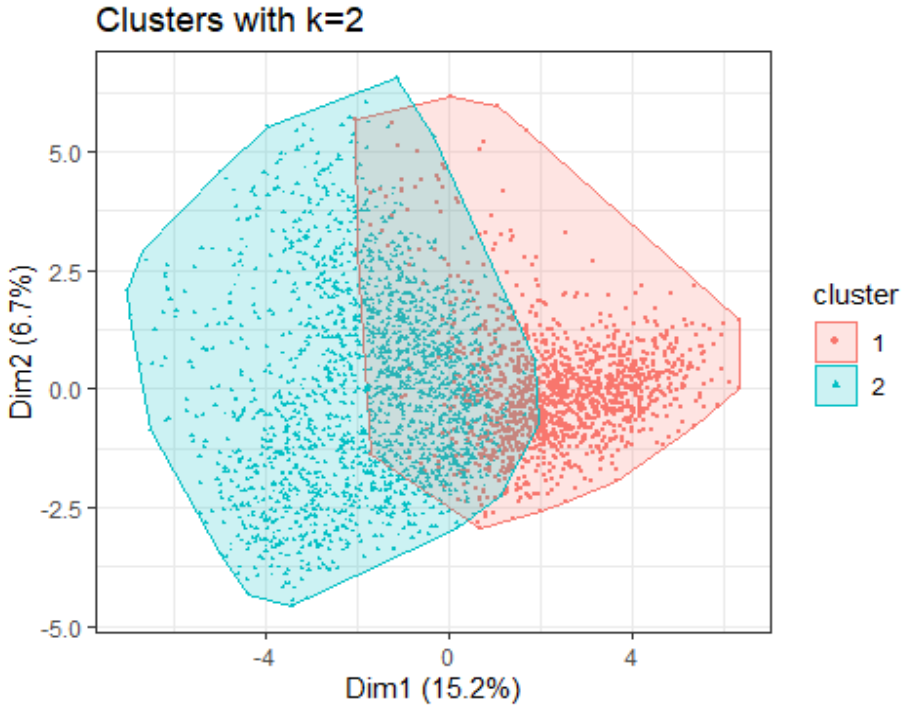
In the following part of code we first load the data, we normalize all values in the dataframe between 0 and 1, except the first column. Furthermore, we use column bind to get this column back in the dataframe.

b. Gebruik kmeans om deze data te clusteren. Leg uit hoe u tot een uiteindelijke keuze voor K komt.

To figure out what the optimal K is to use in kmeans clustering we use the 'Elbow' method from the factoextra package. This technique or method runs k-means clustering on the dataset for a range of k-values and computes an average score for these clusters. These average scores are plotted, making it possible to visually determine the best value for K.



In an ideal world this plot above would look like an arm with a clear elbow at K, however the graph shown above does not seem to have a clear or sharp elbow. It looks like K sits at 2 or 3. The 'total within sum of squares' decreases significantly after two cluster and to a lesser extent when adding another cluster. Therefore, we try two clusters to begin with.



c. Beschrijf uw bevindingen; interpreteer uw resultaten

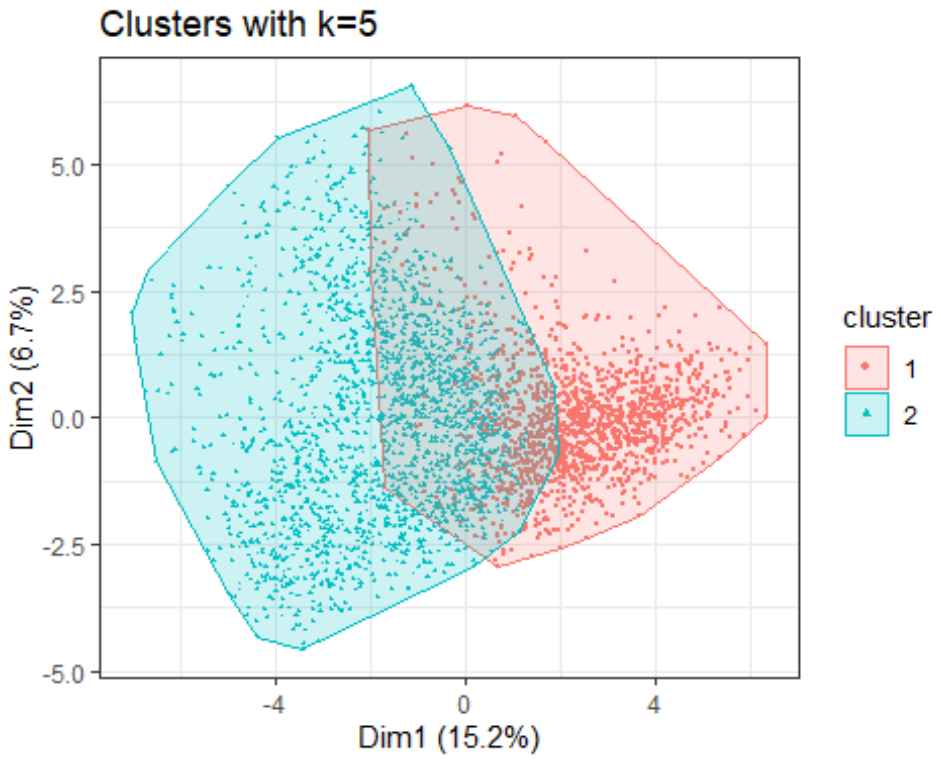
The first thing we notice when looking at the plot above is that there is significant overlap between the clusters. This makes sense, the respondents from the different countries are going to have at least some overlap in their answers. In other words, some overlap in their 'social values, norms and stereotypes'. Let's delve into this further by looking at the specific countries that were assigned to each cluster. We do this using the table function:

```
##
##      AUS CHN DEU NGA PAK PER USA
##    1 422  77 262  11  37  89 547
##    2  44 748  54 321 427 232 128
```

We see that cluster 1, plotted in red in the figure above, is dominated by responses from the USA, Germany and Australia. This data would suggest that western countries have more in common with each other and are assigned more to the same cluster. In cluster 2 we can observe an over representation of China, Pakistan, Peru and Nigeria. This would suggest that these countries, on average, have more in common with each other than western countries. However, as mention before there is overlap between the clusters. Further analysis is needed.

d. Ongeacht wat u bij b heeft gedaan, doe kmeans met 5 clusters op deze data en interpreteer de clusters

In the next section we do kmeans clustering with k=5 and plot this in the figure shown below.



This figure, showing the 5 clusters, is a bit more difficult to interpret. We observe much more overlap between the cluster than before. Let's look at this by country:

```
##
##     AUS CHN DEU NGA PAK PER USA
##   1   5  40   4 250 147  27  19
##   2  17 711  28   6   5  36  31
##   3   2  15   8  16 253  11  10
##   4 320  27 182   0   9  12 320
##   5 122  32  94  60  50 235 295
```

The Usa, Germany and Australia are again over represented in the same cluster (4), but also have a lot of responses in cluster 2. We could speculate that these 'free' societies are not homogeneous in their 'social values, norms and stereotypes', which results in different responses from the same countries. We see that Nigerian responses are concentrated in the first cluster(almost 80%), suggesting a homogeneity of 'social values, norms and stereotypes'. China in the same way is also over represented in cluster 5(85% of its responses are clustered in this cluster). Furthermore, we can see that China also makes up 85% of responses in this specific cluster. So not only would this suggest that China has homogeneity of its 'social values, norms and stereotypes', but also that these are their own and differ from the rest of the countries.

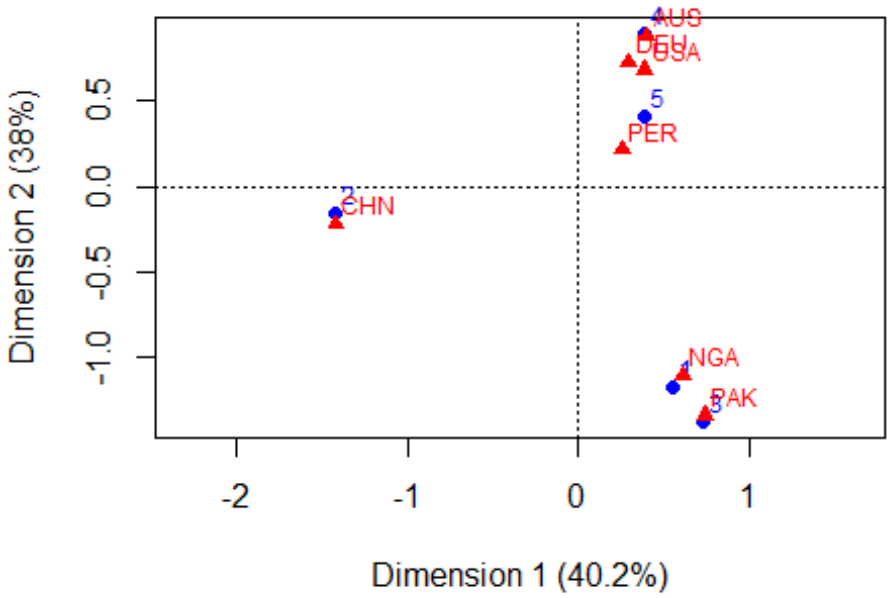
e. Doe correspondentie analyse op de kruistabel die u verkrijgt wanneer u kijkt naar de verdelingen over de landen voor de 5 gevonden clusters

Let's build on the previous question and do some correspondence analysis. We use the 'ca' package for this.

```
##
##  Principal inertias (eigenvalues):
##             1        2        3        4
## Value      0.663938 0.627559 0.230525 0.129112
## Percentage 40.21%   38.01%   13.96%   7.82%
##
##
##  Rows:
##                 1         2         3        4        5
## Mass     0.144748  0.245366  0.092674 0.255958 0.261253
## ChiDist  1.537997  1.427234  1.945902 1.070299 0.794248
## Inertia  0.342393  0.499810  0.350914 0.293210 0.164806
## Dim. 1   0.678987 -1.739929  0.888579 0.479163 0.473271
## Dim. 2  -1.476580 -0.200721 -1.733915 1.129633 0.514956
##
##
##  Columns:
##              AUS       CHN       DEU       NGA       PAK       PER       USA
## Mass    0.137099  0.242718 0.092969  0.097676  0.136511 0.094440 0.198588
## ChiDist 1.054073  1.434147 0.821408  1.751781  1.716938 1.081541 0.790946
## Inertia 0.152327  0.499218 0.062727  0.299741  0.402417 0.110469 0.124236
## Dim. 1  0.491600 -1.738274 0.360417  0.746413  0.909595 0.315181 0.474169
## Dim. 2  1.110753 -0.276650 0.913196 -1.396146 -1.688983 0.268996 0.863579
```

f. Beschrijf uw bevindingen; interpreteer het plaatje.

First, we take a look at the dimensions and the amount of variance explained by each dimension. The first dimension accounts for about 40 percent of the variance in the data, the second dimension accounts for 38 percent of the variance. These dimensions combined explain 78% of the variance. The third dimension accounts for only 13.96% of the variance. Because the first two dimension explain away a lot of the variance in this data, dimensionality reduction is appropriate in this case and may provide us with (additional) insights.



The plot provided above presents a visual summary of the data in two dimensions. In the previous question we observed that the responses from the USA, Australia and Germany were related, they were concentrated in the same cluster. The fact that these countries are very close to each other in the plot presented above, would suggest a strong relationship between the answers the respondents from these countries gave. Thus, 'social values, norms and stereotypes' of these countries are strongly related. Furthermore, in the previous question we saw 85% of the responses coming out of China being isolated in cluster 5. We speculated that not only does China have a homogeneous responses, but that these responses were different from other countries. In other words, China has its own homogeneous 'social values, norms and stereotypes'. This plot also confirms this. Lastly, we see that Nigeria and Pakistan are very close to each other in the plot, suggesting that the responses from these countries are closely related to each other. Both Nigeria and Pakistan have majority Muslim populations. Religion plays an important role in the 'social values, norms and stereotypes' of a population. This is what we could be seeing in the plot.

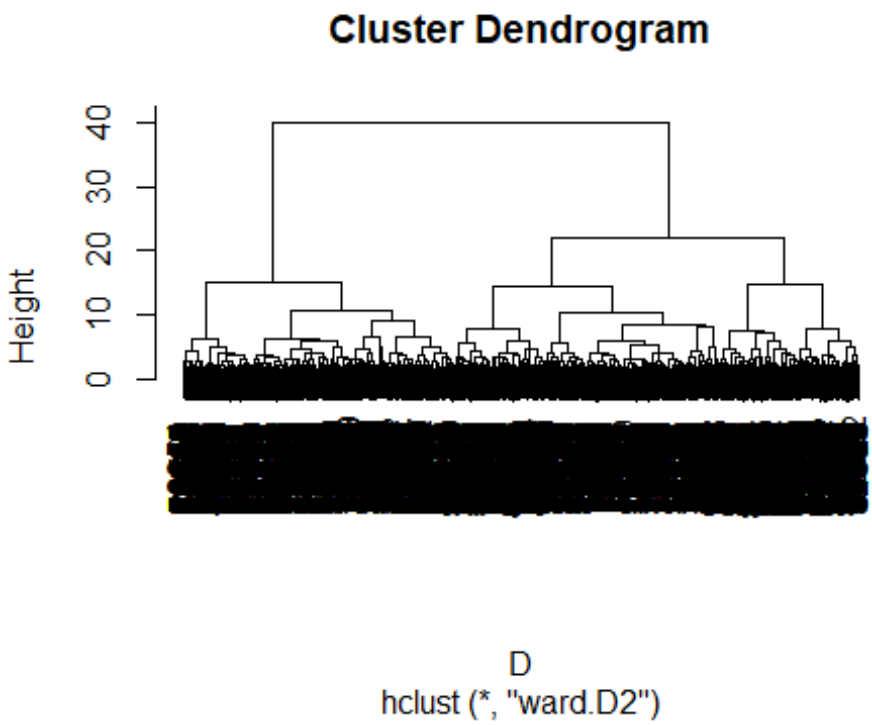
2. Gebruik nu hierarchische cluster analyse om clusters van waarnemingen te vinden.

a. Kies een afstandsmaat en een linkage method, en motiveer uw keuzes.

In this section we delve in to hierarchical clustering. In hierarchical clustering there are two approaches. The agglomerative approach, where each observation is considered a cluster, the distance between these clusters is calculated and the clusters closest to each other are combined. This process is repeated until there's only 1 cluster left. The divisive approach does the opposite, it starts off with 1 cluster and splits the clusters until every observation is its own cluster. In this case we will be using the agglomorative approach, simply for the fact that the divisive approach is more computationally intensive. Crucial to cluster analysis is the definition of 'distance' or simmilarity. Observations that are similar to one another must be clustered into the same cluster. Let's look at how we define this distance or similarity? We consider two types, 'Euclidian' distance and 'Manhattan' distance. Euclidian distance captures the distance between two points by aggregating the squared difference in each variable . 'Manhattan' distance captures the distance by aggregating the pairwise difference between each variable. Manhattan distance is appropriate when you have high dimensionality in the data and when you're dealing with categorical data. Given the relatively low dimensionality of our data, Euclidian distance is more appropriate and will be used in the next section. Lastly, we need to consider the linkage method we will use in our clustering. From what I've read there's no real method to predetermine which linkage method is best, single linkage is fast, but performs poorly when the data is noisy. Average and complete linkage perform well on cleanly separated clusters, but have mixed results otherwise. The ward method is most effective on noisy data. We use the 'agnes' function to determine the 'agglomerative coefficient'. This coefficient measures the amount of clustering structure found.

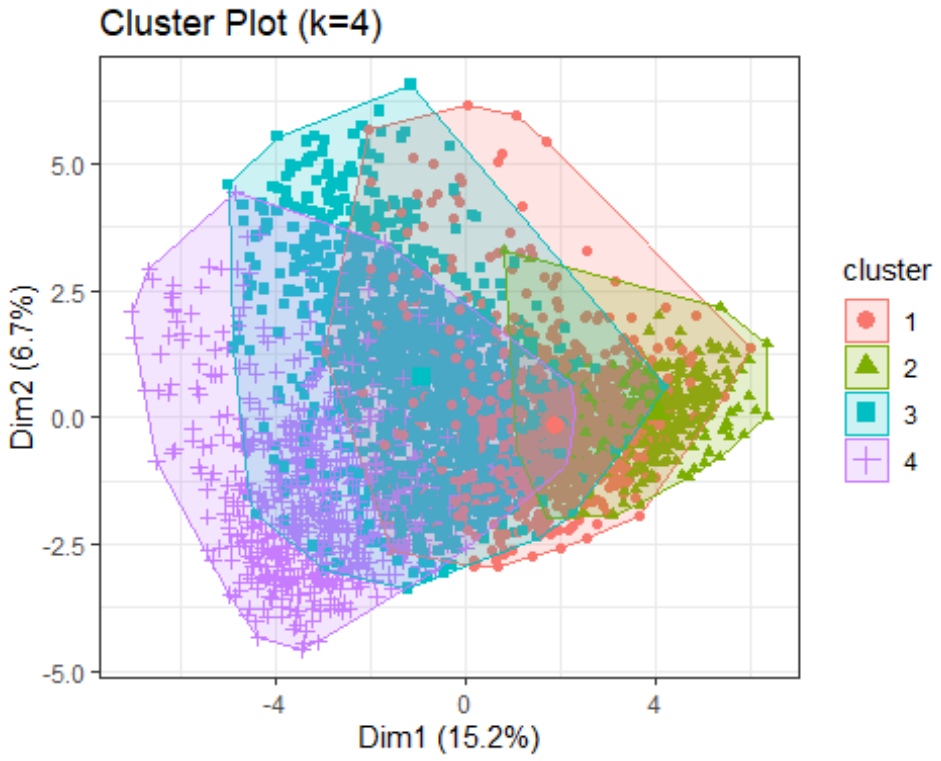
```
##   average    single  complete      ward
## 0.6297230 0.4743831 0.7315128 0.9755946
```

Again, the AC describes the strength of the clustering structure. The coefficient takes values from 0 to 1. The AC for the 'Ward' method would suggest this is the best linkage method and this will be used in the next section.



The figure above presents a dendogram. A dendogram is a branching diagram that represents the relationship of similarity among groups. The way to interpret this dendogram is by looking at the largest vertical difference between nodes, draw a horizontal line and count the number of vertical lines that intersect with this line. In this case it would be 2, suggesting that the number of clusters is 2. Now we can move on to the next question, where we use the 'cutree' function.

b. Kies 4 clusters en interpreteer deze clusters.



We see significant overlap between the clusters. Furthermore, dim1 and dim2 explain only roughly 22% of the variance. Let's look at this by country.

```
table(clus.ident, normalized_data[,1])

##
## clus.ident AUS CHN DEU NGA PAK PER USA
##          1 273  62 158  12  33  65 419
##          2 135   0  91   0   0   2 118
##          3  51 742  56  97  37 234 116
##          4   7  21  11 223 394  20  22
```

We observe that Australia, Germany and the USA are over represendted in cluster 1. Confirming our previous results that these countries are similar when it comes to 'social values, norms and stereotypes'. Pakistan and Nigeria are concentrated in cluster 4, which could be explained by religious similarities between these countries. China and Peru seem to be over represented in cluster 3, which would suggest there being similarities between these countries.