ADS2 Problem set: Clustering

Rob Young (based on Dmytro Shytikov and Melanie Stefan)

Semester 2, 2023/24

We expect this problem set to take around one hour to complete. But professors are sometimes wrong![citation missing].

More about Dr. Fischer's seating plan

We are going to work with the same dataset as for the practical: The guest list for Dr. Fischer's conference dinner, which is provided in file `guests.csv`.

Finding the best 4-means cluster

In this week's practical, you did a k-means clustering with k=4 and for one specific (randomly chosen) set of initial centroids. Now, let's improve on that a bit.

As you know from lecture, different initial centroids can give you different final results. The solution is to do the clustering several times and pick the best result.

But to know what's "best", we need some quality criterion.

- Assume Dr. Fischer has an additional constraint: She does not want a table with just two people, and another table with 9 people! Instead, the number of people per table should be quite similar between clusters (though they don't have to be exactly the same).
- This is the first task: Design a quantitative measure of "similar numbers of people per table". (To test your intuition, check what your measure does if there are exactly 5 people on each table. Check what it does if there are 20 people at one table and nobody at the other tables. Does your measure behave as expected?)
- Now that we have a measure of "quality" (a good cluster is one where number of peoples at each table are similar), we can run the 4-means clustering algorithm several time, compute our quality measure, and determine the cluster that does best with respect to this quality measure. (In terms of code, this is not much harder than what you did for Practical. All you need is an extra loop to run the clustering several times, and some way to keep track of the quality criterion.)

Is 4-means the answer?

In practical, we assumed that 4-means clustering would be best, but is that true?

- Run the same analysis as before using other numbers of clusters. For each, keep track of the quality measure (as defined above) for the "best" cluster. Plot those against cluster number. This should give you an elbow plot. Does it? What do you conclude about optimal cluster size?