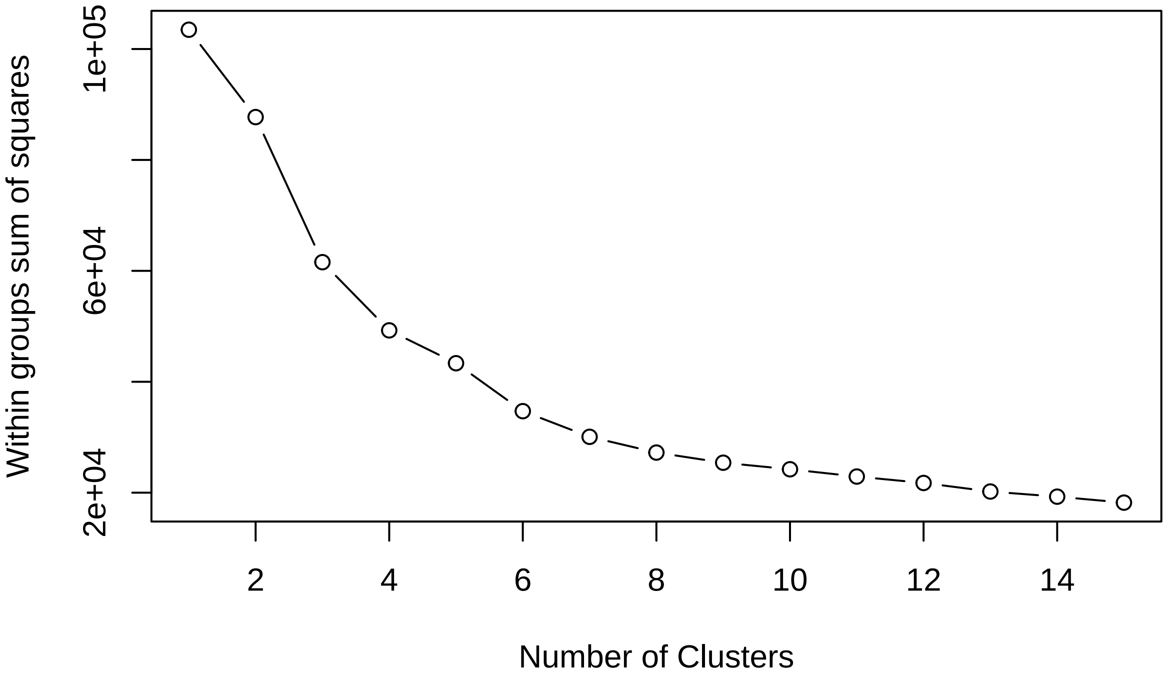
Clustering LENA segments

George

3/30/2022

Our basic approach is to 1) select the number of clusters (k) for k-means that looks justified for the segments, based only on LENA features, and then 2) to look at the proportion of sleep, tCDS, ODS in each of these clusters. The key questions are: 1) Are there cluster(s) that correspond primarily to particular activity types (e.g., sleep)? and if so: 2) What do these clusters looks like, in terms of their average LENA features?

(Note: we used this within-groups sum of squares plot to pick 7 clusters, although here it changes every time as k-means is run again with a random seed. The 'elbow' often looks to be at $k = 7$.)



Load clustered data.

Merge demographic data, mostly to check whether clusters differ much by language.

Do our clusters all capture several children? (We wouldn't want clusters that only capture a few children, as these may be idiosyncratic and unlikely to generalize well to future datasets.) Below, we see that there are segments from 47-147 children in each cluster, with most clusters having >100 children represented.

What are the means of the LENA variables for each cluster?

Table 1: Children represented per cluster.

cluster	Nclust
1	147
2	129
3	103
4	118
5	47
6	135
7	139

Table 2: Means of LENA variables by cluster.

cluster	N	sleep	tCDS	ODS	AWC	CTC	CVC	noise	silence	distant	TV	meaningful
4	2041	**0.64**	0.22	0.14	3.01	0.07	0.49	0.01	0.85	0.08	0.02	0.03
5	142	**0.53**	0.30	0.17	3.44	0.11	0.86	0.63	0.12	0.16	0.04	0.04
6	1256	0.00	**0.73**	0.27	54.55	3.78	9.51	0.02	0.27	0.25	0.01	0.45
1	3450	0.01	**0.60**	0.39	21.97	1.14	4.76	0.03	0.37	0.33	0.03	0.25
7	1485	0.01	0.33	**0.66**	76.10	1.40	2.61	0.01	0.21	0.33	0.03	0.42
2	3475	0.04	0.45	**0.51**	13.63	0.37	1.80	0.03	0.17	0.66	0.02	0.12
3	1087	0.27	0.28	**0.45**	7.33	0.16	0.73	0.02	0.15	0.07	0.69	0.06