Homework 4

CAP 4453

Spring 2023

Assume you have the following data:

Data:

Name	Gender	Age	Height
Carlos	Male	24	6.2
Jhon	Male	2	2.4
Jessica	Female	41	5.1
Keneddy	Female	12	4.4
Peter	Male	19	6.0
Lathia	Female	17	5.3
Nancy	Female	13	4.8
Ana	Female	22	5.8
Jared	Male	27	6.1
Lebron	Male	35	6.6
David	Male	16	5.6
Henry	Male	8	4.6
Claire	Female	8	4.5
Jude	Female	10	4.9
Grace	Female	12	5.0
Mason	Male	14	5.5
Blake	Male	5	4.3

You are going to cluster the data base in two features simultaneously. They are: Age and Height.

Given that the units are not comparable, you are going to use a value to scale one of the axis when you are computing distances.

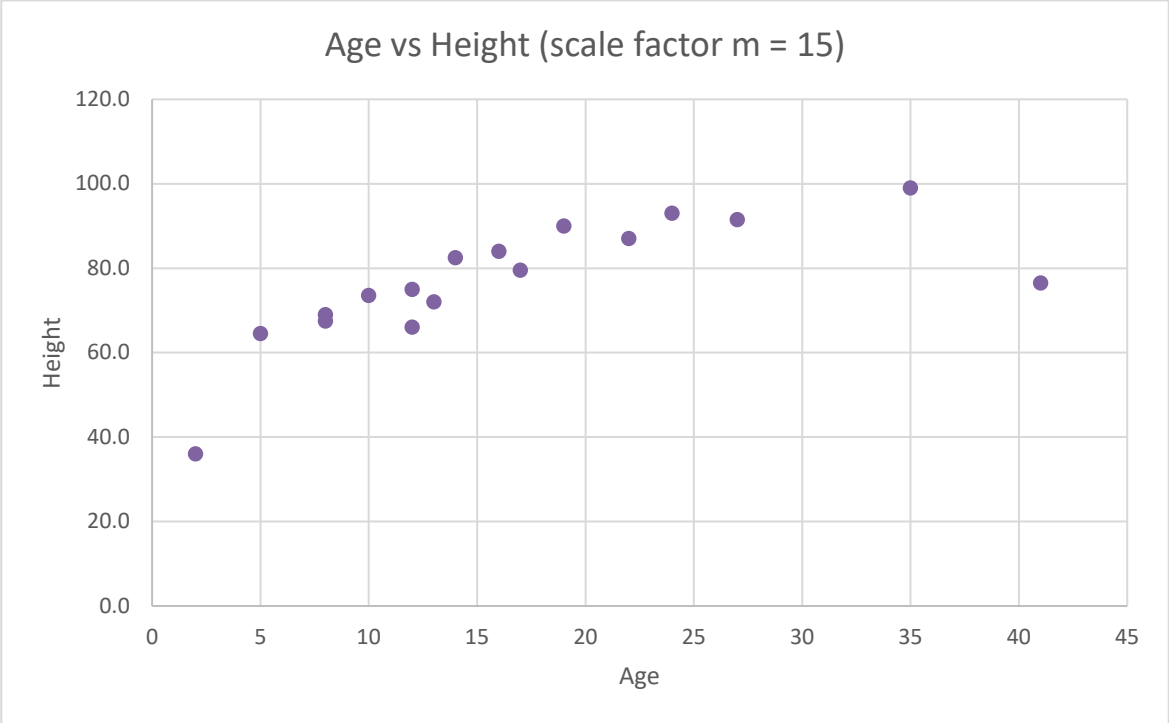
Define the distance between any two persons i and j as:

$$d_{ij} = \left|Age_i - Age_j\right| + m\left|Height_i - Height_j\right|$$

Note that m weights which component (age or height) is more important when you are computing distances.

Assuming a m=15,

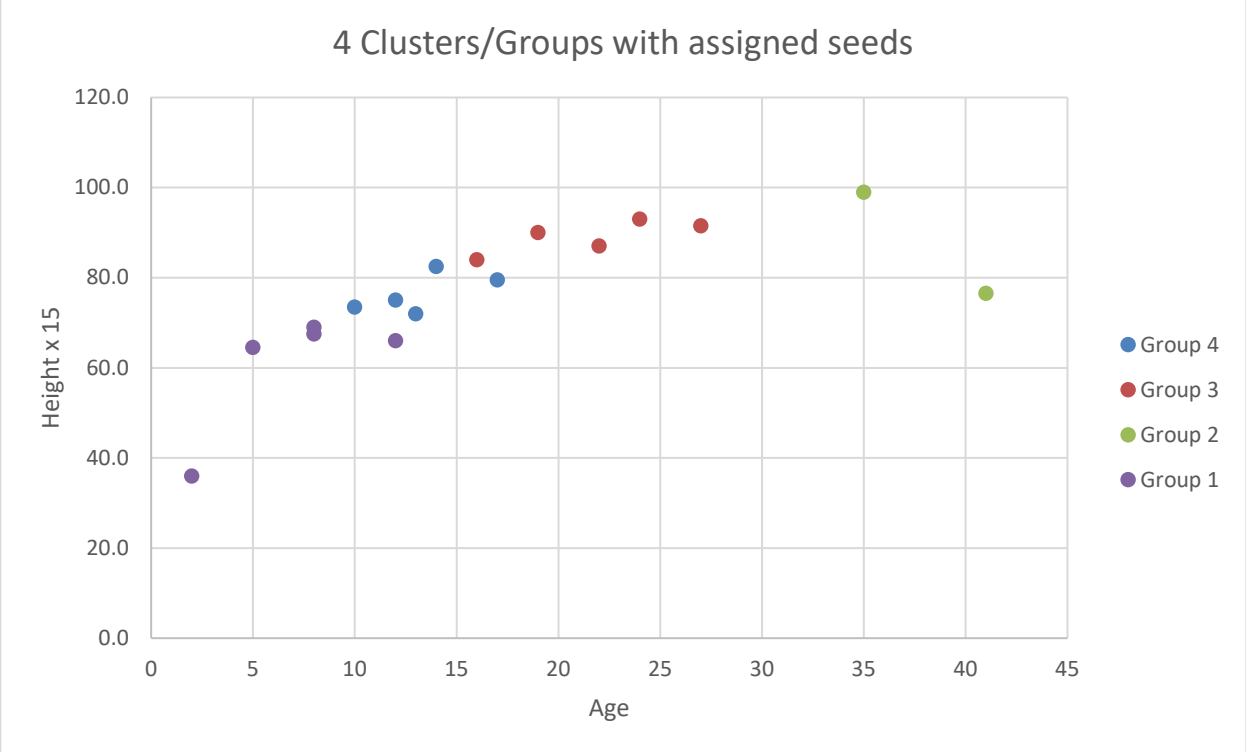
1. [20%] Plot your data in a cartesian plane.
    a. Put Age in the x-axis, Height in the Y-axis (you can use the factor of m = 15 in this axis) to visualize how close they are using the defined distance



2. Define the number of clusters K=4, with seeds on Blake, LeBron, Peter and Grace, and compute:
    a. [15%] Assign a group for each of the persons of the table
    b. [15%] Compute new center centers

Shows your results in tables

Name	Gender	Age	Height	Seed 1 Blake	Seed 2 LeBron	Seed 3 Peter	Seed 4 Grace	Min dist. per each case	Group
Jhon	Male	2	2.4	31.5	96	71	49	31.5	**1**
Keneddy	Female	12	4.4	8.5	56	31	9	8.5	**1**
Henry	Male	8	4.6	7.5	57	32	10	7.5	**1**
Claire	Female	8	4.5	6	58.5	33.5	11.5	6	**1**
Blake	Male	5	4.3	0	64.5	39.5	17.5	0	**1**
New Centers Group 1		**7**	**4.0**						
Jessica	Female	41	5.1	48	28.5	35.5	30.5	28.5	**2**
Lebron	Male	35	6.6	64.5	0	25	47	0	**2**
New Centers Group 2		**38**	**5.9**						
Carlos	Male	24	6.2	47.5	17	8	30	8	**3**
Peter	Male	19	6.0	39.5	25	0	22	0	**3**
Ana	Female	22	5.8	39.5	25	6	22	6	**3**
Jared	Male	27	6.1	49	15.5	9.5	31.5	9.5	**3**
David	Male	16	5.6	30.5	34	9	13	9	**3**
New Centers Group 3		**22**	**5.9**						
Lathia	Female	17	5.3	27	37.5	12.5	9.5	9.5	**4**
Nancy	Female	13	4.8	15.5	49	24	4	4	**4**
Jude	Female	10	4.9	14	50.5	25.5	3.5	3.5	**4**
Grace	Female	12	5.0	17.5	47	22	0	0	**4**
Mason	Male	14	5.5	27	37.5	12.5	9.5	9.5	**4**
New Centers Group 4		**13**	**5.1**						

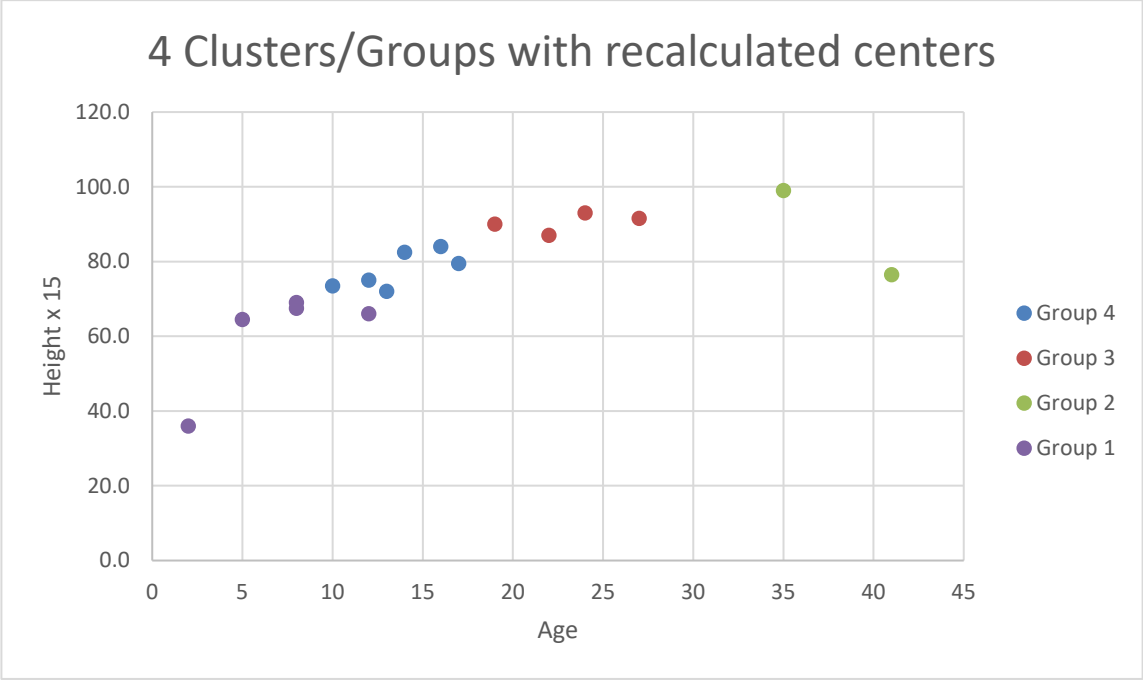


3. Compute and show results in tables:

   a) [15%] Assign a group for each of the persons of the table

   b) [15%] Compute a new center

Name	Gender	Age	Height	New Center 1	New Center 2	New Center 3	New Center 4	Min dist. per each case	Group
Jhon	Male	2	2.4	29.6	87.75	72.7	51.7	29.6	**1**
Keneddy	Female	12	4.4	10.4	47.75	32.7	11.7	10.4	**1**
Henry	Male	8	4.6	9.4	48.75	33.7	12.7	9.4	**1**
Claire	Female	8	4.5	7.9	50.25	35.2	14.2	7.9	**1**
Blake	Male	5	4.3	5.9	56.25	41.2	20.2	5.9	**1**
New Centers Group 1		**7**	**4.0**						
Jessica	Female	41	5.1	49.9	14.25	32	27.8	14.25	**2**
Lebron	Male	35	6.6	66.4	14.25	23.3	44.3	14.25	**2**
New Centers Group 2		**38**	**5.9**						
Carlos	Male	24	6.2	49.4	19.25	6.3	27.3	6.3	**3**
Peter	Male	19	6.0	41.4	21.25	3.5	19.3	3.5	**3**
Ana	Female	22	5.8	41.4	16.75	2.5	19.3	2.5	**3**
Jared	Male	27	6.1	50.9	14.75	7.8	28.8	7.8	**3**
New Centers Group 3		**23**	**6.0**						
David	Male	16	5.6	32.4	25.75	10.7	10.3	10.3	**4**
Lathia	Female	17	5.3	28.9	29.25	14.2	6.8	6.8	**4**
Nancy	Female	13	4.8	17.4	40.75	25.7	4.7	4.7	**4**
Jude	Female	10	4.9	15.9	42.25	27.2	6.2	6.2	**4**
Grace	Female	12	5.0	19.4	38.75	23.7	2.7	2.7	**4**
Mason	Male	14	5.5	28.9	29.25	14.2	6.8	6.8	**4**
New Centers Group 4		**14**	**5.2**						



4. [10%] Do you see any change in the assignations of the clusters? Would you recommend an extra round of computations?

Yes, David was in group 3 when the initial cluster centers were assigned. After recalculating the centers as the mean of the points in a cluster and the distance from each data point to the revised center of a cluster, David is now in group 4. After the second iteration, the cluster are stabilized so an extra round of computation is not required. However, to realize that there is no change in the assignation of the clusters, the extra round is needed to compare the results.

5. [10%] Could you tell some characteristics about the clusters created? Check the age, and/or height of the elements of each cluster. Do the cluster corresponds makes sense in the real world?

The 4 clusters created have a correlation between the average age and height of a human as shown in the table below. For example, elements of cluster 1 (children) have the lowest height, as they get older (teenagers) the height increases until they become adults and the height stabilizes. Hence, we can say that the cluster assignations make sense in the real world.

Cluster centers	Age	Height
Average cluster 1	**7**	**4.0**
Average cluster 4	**14**	**5.2**
Average cluster 3	**23**	**6.0**
Average cluster 2	**38**	**5.9**