John David Conley

Machine Learning assignment 7 (in-class assignment 6)

1: (Provide only mathematical solutions for this question) Six points with the following attributes are given, calculate and find out clustering representations and dendrogram using Single, complete, and average link proximity function in hierarchical clustering technique.

X-Y coordinates of six points:

|  |  |  |
| --- | --- | --- |
| point | x coordinate | y coordinate |
| p1 | 0.4005 | 0.5306 |
| p2 | 0.2148 | 0.3854 |
| p3 | 0.3457 | 0.3156 |
| p4 | 0.2652 | 0.1875 |
| p5 | 0.0789 | 0.4139 |
| p6 | 0.4548 | 0.3022 |

Distance Matrix for six points:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | p1 | p2 | p3 | p4 | p5 | p6 |
| p1 | 0.0000 | 0.2357 | 0.2218 | 0.3688 | 0.3421 | 0.2347 |
| p2 | 0.2357 | 0.0000 | 0.1483 | 0.2042 | 0.1388 | 0.2540 |
| p3 | 0.2218 | 0.1483 | 0.0000 | 0.1513 | 0.2843 | 0.1100 |
| p4 | 0.3688 | 0.2042 | 0.1513 | 0.0000 | 0.2932 | 0.2216 |
| p5 | 0.3421 | 0.1388 | 0.2843 | 0.2932 | 0.0000 | 0.3921 |
| p6 | 0.2347 | 0.2540 | 0.1100 | 0.2216 | 0.3921 | 0.0000 |

Least-to-greatest:

p3-p6 (0.1100), p2-p5 (0.1388), p2-p3 (0.1483), p3-p4 (0.1513), p2-p4 (0.2042)

p4-p6 (0.2216), p1-p3 (0.2218), p1-p6 (0.2347), p1-p2 (0.2357), p2-p5 (0.2540)

p3-p5 (0.2843), p4-p5 (0.2932), p1-p5 (0.3421), p1-p4 (0.3421), p5-p6 (0.3921)

Single:

1: p3 (0.3457, 0.3156) - p6 (0.4548, 0.3022), distance 0.1100

2: p2 (0.2148, 0.3854) - p5 (0.0789, 0.4139), distance 0.1388

3: 2=p2 (0.2148, 0.3854) - 1=p3 (0.3457, 0.3156), distance 0.1483

4: 3=p3 (0.3457, 0.3156) - p4 (0.2652, 0.1875), distance 0.1513

5: p1 (0.4005, 0.5306) - 4=p3 (0.3457, 0.3156), distance 0.2218

Complete:

1: p3 (0.3457, 0.3156) - p6 (0.4548, 0.3022), distance 0.1100

2: p2 (0.2148, 0.3854) - p5 (0.0789, 0.4139), distance 0.1388

3: p4 (0.2652, 0.1875) - 1=p6 (0.4548, 0.3022), distance 0.2216

4: p1 (0.4005, 0.5306) - 2=p5 (0.0789, 0.4139), distance 0.3421

5: 4=p1 (0.4005, 0.5306) - 3=p4 (0.2652, 0.1875), distance 0.3688

Average:

1: p3 (0.3457, 0.3156) - p6 (0.4548, 0.3022), distance 0.1100, midpoint (0.40025, 0.3089)

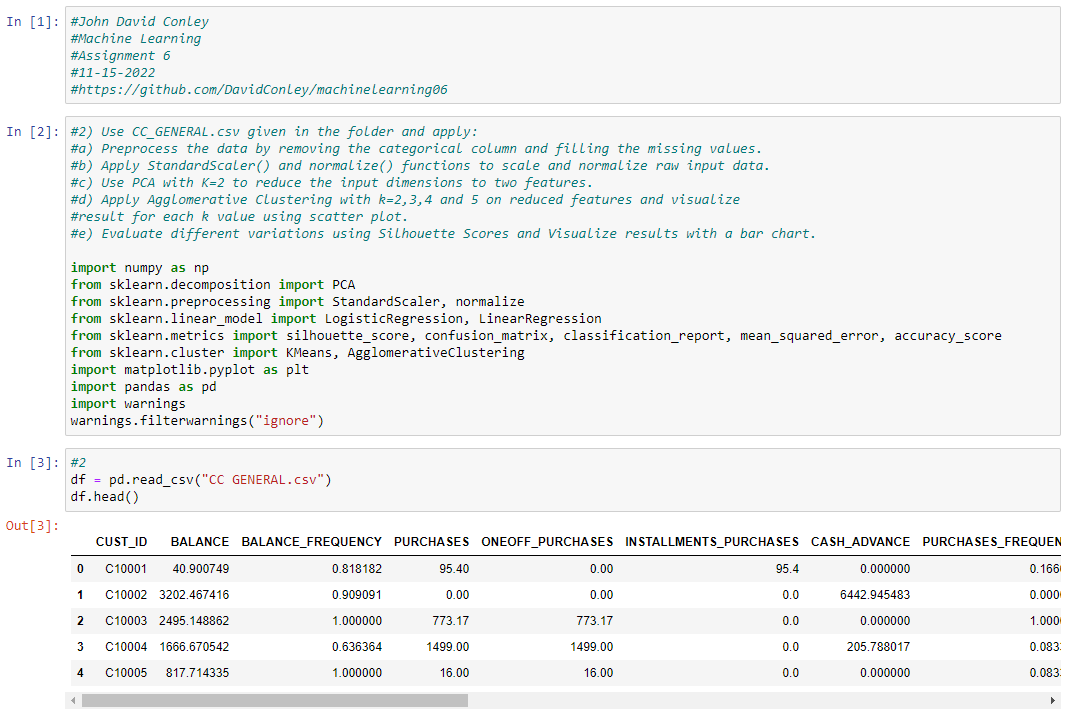
2: p2 (0.2148, 0.3854) - p5 (0.0789, 0.4139), distance 0.1388, midpoint (0.14685, 0.39965)

3: 1 (0.40025, 0.3089) - p4 (0.2652, 0.1875), distance 0.1816, midpoint (0.332725, 0.2482)

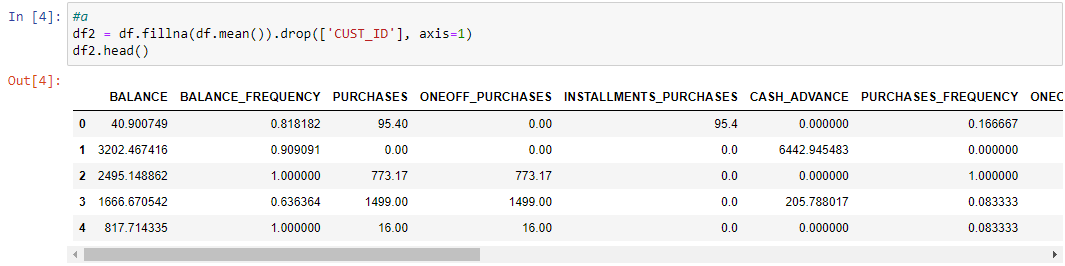
4: 2 (0.14685, 0.39965) - 3 (0.332725, 0.2482), distance 0.2398, midpoint (0.2397875, 0.323925)

5: p1 (0.4005, 0.5306) - 4 (0.2397875, 0.323925), distance 0.2618, midpoint (0.32014375, 0.4272625)

2 (report)

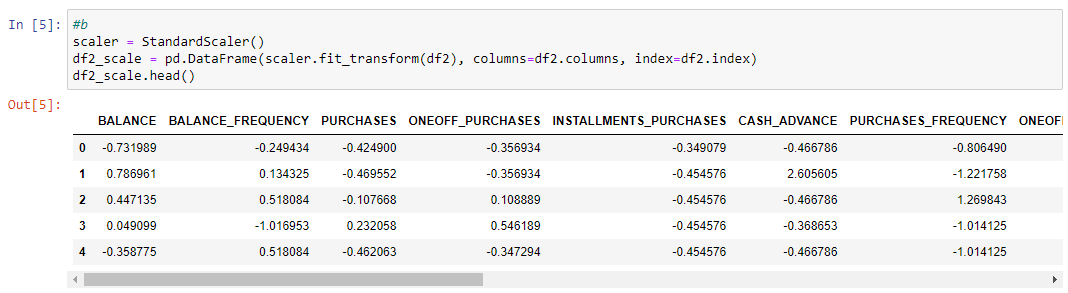
Here we begin.

a:



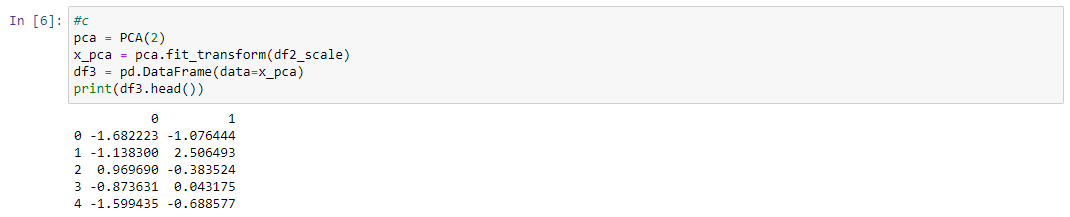
Removing non-numeric column and filling nulls with means

b:



Standard scaler and normalization

c:



PCA

d:



Various agglomerates

e:



**MemoryError**: Unable to allocate 611. MiB for an array with shape (8950, 8950) and data type float64

It always returns that no matter what I do, so I can’t test the code.

Video link: <https://youtu.be/RqxoMNb0U08>