C:\Users\Nelson\SG-DAT-NL\Project>python 05\_NL\_Project\_kMeansA.py

loading c:\Users\Nelson\SG-DAT-NL\Project\data\nifty\_one\_minute\_checkpoint.pkl

---We try to find the best clustering based on all available indicators

copying dataframe

analyzing

Trying 2 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

Trying 3 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

Trying 4 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

Trying 5 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

Trying 6 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

Trying 7 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

Trying 8 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

clusters mean std

2 0.20368 0.00696

3 0.17468 0.00553

4 0.15010 0.00344

5 0.15174 0.00504

6 0.11614 0.00503

7 0.11463 0.00538

8 0.10771 0.00474

This seems to recommend a 2-cluster solution. But, we are looking at Silhouette score, perhaps we should look at elbows of Scree plot instead?

Now we try DBSCAN

C:\Users\Nelson\SG-DAT-NL\Project>python 05\_NL\_Project\_kMeansB.py

loading c:\Users\Nelson\SG-DAT-NL\Project\data\nifty\_one\_minute\_checkpoint.pkl

---We try to find the best clustering based on all available indicators

copying dataframe

analyzing

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10 [-0.28580133776713273, -0.32525475720276426, -0.29724534542523773, -0.30758284147390019, -0.32026554568670473, -0.28047395251891599, -0.17682837123199396, -0.27954898235716835, -0.25759592754955135, -0.11417698036592802]

-0.264477404158

DBSCAN with all the data doesn’t seem to produce particularly good clusters

We reduce the number of predictors:

C:\Users\Nelson\SG-DAT-NL\Project>python 05\_NL\_Project\_kMeansB2.py

loading c:\Users\Nelson\SG-DAT-NL\Project\data\nifty\_one\_minute\_checkpoint.pkl

---We try to find the best clustering based on a reduced set of indicators selected based on prior knowledge

copying dataframe

analyzing

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10 [-0.25054021336101023, -0.21526459917219171, -0.23692857300666179, -0.21920891386248859, -0.22273788179455634, -0.21047613226871226, -0.22579936176503937, -0.27114061026889363, -0.17698205384246504, -0.2171418320360351]

-0.224622017138

But it still isn’t too good.

Let’s try k-means again, with the reduced set:

C:\Users\Nelson\SG-DAT-NL\Project>python 05\_NL\_Project\_kMeansA2.py

loading c:\Users\Nelson\SG-DAT-NL\Project\data\nifty\_one\_minute\_checkpoint.pkl

---We try to find the best clustering based on a reduced set of indicators

copying dataframe

analyzing

Trying 2 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

Trying 3 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

Trying 4 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

Trying 5 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

Trying 6 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

Trying 7 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

Trying 8 cluster solution

silhouette calculation (total 10):

1 2 3 4 5 6 7 8 9 10

clusters mean std

2 0.36327 0.02139

3 0.31181 0.01435

4 0.32074 0.01349

5 0.22591 0.01166

6 0.12810 0.01222

7 0.15633 0.00642

8 0.12374 0.00699

Maybe this is better, but we need to output the elbows to assess, probably.