Joe Griffin
Assignment 8
May 9, 2021

1. K-means clustering aims to place objects in groups so that the objects are as similar as possible to the others in their group, and as different as possible from those objects not in the same group. First, one must select the variables to be used in the clustering. Only variables that characterize the objects and relate specifically to the objectives of the cluster analysis should be used. Once the variables have been selected, one should consider several factors before continuing. The sample size should be sufficient, in certain cases the data should be standardized, and a check for outliers should be done. If outliers exist, it should be decided whether they should be deleted or not.

   Next, it needs to be decided how similarity will be measured between the objects. There are three main methods: correlational measures, distance measures, and association. Within each of these methods there are specific techniques that can be used. A hierarchical or non-hierarchical partitioning procedure must be decided on, and lastly the number of clusters must be determined. This can be difficult as there is no direct formula to determine the optimal number of clusters, but certain observations about the data can help determine the optimal number.

   There are many different examples of how this can be applied. It can be used by a pizza chain to optimize how many stores they need in an area and where exactly the stores should be based on where the orders come from. In a similar fashion, it could be used to optimize hospital or school locations as well.

   There are many advantages of this method -- the examples above support that. By minimizing the distance of the pizza places to the homes where they will deliver, they will make their business run more efficiently. One main disadvantage that was mentioned above is that there is no one answer as to the exact optimal number of clusters. This is certainly one limitation. Other limitations include the fact that the technique can only utilize numerical data, and that spherical clusters are assumed.

2. 3 clusters
   - Cluster 1: 24
   - Cluster 2: 12
   - Cluster 3: 64

4 clusters
   - Cluster 1: 11
   - Cluster 2: 52
   - Cluster 3: 12
   - Cluster 4: 25

5 clusters

- Cluster 1: 9
- Cluster 2: 34
- Cluster 3: 28
- Cluster 4: 22
- Cluster 5: 7

I think all of these clusters seem to give similar results, but if I had to pick one I think I would choose 5 clusters. This has the highest Pseudo F statistic, which suggests the optimal number of clusters. Furthermore, the distances between cluster centroids remain high despite adding more clusters. The standard deviations for each cluster are also low.

Replace=FULL Drift Radius=0 Maxclusters=3 Maxiter=1

Initial Seeds					
Cluster	**x6**	**x8**	**x12**	**x15**	**x18**
1	1.281978445	0.937628564	2.496456809	0.904190706	0.563696948
2	-0.938207688	-1.479950311	-1.327029525	2.913503384	-0.797889883
3	-1.654396764	-0.434510798	-0.207960354	-1.774892866	0.563696948

Replace=FULL Drift Radius=0 Maxclusters=3 Maxiter=1

Cluster Listing		
Obs	Cluster	Distance from Seed
1	3	2.4177
2	3	2.2946
3	1	1.2763
4	3	1.8793
5	3	1.0192
6	2	2.0787
7	3	2.9912
8	3	1.6399
9	1	2.0758
10	1	1.6341
11	2	1.4218
12	3	2.2748
13	1	1.1808
14	3	1.6503
15	1	1.9234
16	3	1.8448
17	3	2.0101
18	1	2.1039
19	1	2.1854
20	1	1.7291
21	3	2.1995
22	1	1.7784
23	3	1.5776
24	3	2.5049
25	3	1.5388
26	3	1.4725
27	3	2.1850
28	3	1.1141
29	3	1.6469
30	2	0.5528
31	3	2.2946
32	3	1.9568
33	3	1.1544
34	3	0.9323
35	1	1.7687

Replace=FULL Drift Radius=0 Maxclusters=3 Maxiter=1

Cluster Listing		
Obs	Cluster	Distance from Seed
36	3	2.8112
37	3	1.0259
38	1	1.8219
39	2	1.2778
40	2	1.8024
41	3	2.6053
42	3	1.6930
43	1	1.4796
44	1	2.5129
45	3	2.2481
46	1	2.0493
47	3	2.0224
48	1	2.3095
49	1	2.2495
50	3	1.6731
51	3	1.5851
52	3	2.5533
53	2	2.5150
54	3	0.9328
55	3	1.3919
56	3	1.9139
57	1	2.0394
58	3	1.3355
59	3	1.6593
60	3	2.6678
61	3	2.3545
62	3	1.4029
63	2	1.6238
64	3	2.0833
65	3	1.6696
66	1	2.5648
67	3	2.6063
68	3	2.4237
69	3	1.3701
70	2	1.4648

Replace=FULL Drift Radius=0 Maxclusters=3 Maxiter=1

Cluster Listing		
Obs	Cluster	Distance from Seed
71	1	0.8322
72	3	2.5885
73	2	1.7780
74	1	2.0045
75	3	1.4857
76	3	2.5544
77	3	1.8194
78	3	2.1911
79	1	2.0038
80	3	2.0920
81	3	2.3880
82	3	0.9263
83	3	1.4784
84	3	3.4069
85	3	1.1844
86	3	1.7439
87	2	2.2810
88	3	2.6568
89	3	1.0587
90	1	2.8086
91	3	1.5577
92	3	2.3344
93	2	1.7936
94	1	1.1970
95	3	1.7328
96	3	1.3104
97	2	0.7520
98	3	2.9702
99	1	0.9342
100	3	1.1332

Criterion Based on Final Seeds =	0.8653

Replace=FULL Drift Radius=0 Maxclusters=3 Maxiter=1

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	24	0.8472	2.8086		3	2.1083
2	12	0.7922	2.5150		3	2.3499
3	64	0.8882	3.4069		1	2.1083

Statistics for Variables				
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
x6	1.00000	1.00281	0.014693	0.014912
x8	1.00000	0.87865	0.243570	0.321999
x12	1.00000	0.73000	0.477863	0.915207
x15	1.00000	0.84078	0.307378	0.443789
x18	1.00000	0.86654	0.264280	0.359212
OVER-ALL	1.00000	0.86814	0.261557	0.354200

Pseudo F Statistic =	17.18

Approximate Expected Over-All R-Squared =	0.29462

Cubic Clustering Criterion =	-1.977

WARNING: The two values above are invalid for correlated variables.

Cluster Means					
Cluster	x6	x8	x12	x15	x18
1	-0.114590252	0.088208959	1.214190050	0.142326315	0.904093656
2	-0.233955098	-1.322045385	-0.153561158	1.400937451	-0.128443024
3	0.086837925	0.214805150	-0.426528552	-0.316048140	-0.314952054

Cluster Standard Deviations					
Cluster	x6	x8	x12	x15	x18
1	1.274001090	0.736722094	0.743126157	0.675226959	0.643669578
2	1.002587193	0.672403577	0.649096092	0.766033476	0.819955136
3	0.883334424	0.954768423	0.738461756	0.905260584	0.942064441

Replace=FULL Drift Radius=0 Maxclusters=3 Maxiter=1

Distance Between Cluster Centroids			
Nearest Cluster	**1**	**2**	**3**
1	.	2.554220554	2.108254496
2	2.554220554	.	2.349926396
3	2.108254496	2.349926396	.

Replace=FULL Drift Radius=0 Maxclusters=4 Maxiter=1

Initial Seeds					
Cluster	**x6**	**x8**	**x12**	**x15**	**x18**
1	1.281978445	0.937628564	2.496456809	0.904190706	0.563696948
2	1.067121722	-1.806650159	-0.580983411	-0.636282348	0.836014314
3	-0.938207688	-1.479950311	-1.327029525	2.913503384	-0.797889883
4	-1.869253486	1.133648473	-0.114704590	-0.368373991	-0.797889883

Replace=FULL Drift Radius=0 Maxclusters=4 Maxiter=1

Cluster Listing		
Obs	Cluster	Distance from Seed
1	2	2.3040
2	2	2.0266
3	2	1.2711
4	4	1.1257
5	2	1.2982
6	3	2.0211
7	4	2.1794
8	2	1.7708
9	2	1.8977
10	2	1.4294
11	3	1.5173
12	4	2.0665
13	1	1.3766
14	2	1.5409
15	2	2.1080
16	2	1.5036
17	2	1.9499
18	2	2.0305
19	2	2.0585
20	1	1.6071
21	4	1.7788
22	1	1.4589
23	2	1.6232
24	2	2.1477
25	2	1.6538
26	4	1.6375
27	2	2.0921
28	2	1.3895
29	2	1.4368
30	3	0.5128
31	4	1.8219
32	4	1.6180
33	4	0.8247
34	2	0.8932
35	1	1.5979

Replace=FULL Drift Radius=0 Maxclusters=4 Maxiter=1

Cluster Listing		
Obs	Cluster	Distance from Seed
36	4	2.4559
37	4	1.3469
38	2	1.6896
39	3	1.4055
40	3	1.7333
41	2	2.5826
42	2	1.4069
43	2	1.7505
44	1	2.4019
45	2	2.0555
46	2	1.8123
47	4	2.1379
48	2	2.5376
49	1	2.1196
50	2	1.3885
51	4	0.8163
52	2	2.2170
53	3	2.6106
54	4	0.6296
55	2	1.1371
56	4	2.0098
57	1	2.0478
58	2	1.1950
59	2	1.7892
60	2	2.6640
61	2	2.3373
62	4	1.5775
63	3	1.6071
64	4	1.3872
65	4	0.8037
66	2	2.6541
67	2	2.5880
68	2	2.4469
69	3	1.6010
70	3	1.4569

Replace=FULL Drift Radius=0 Maxclusters=4 Maxiter=1

Cluster Listing		
Obs	Cluster	Distance from Seed
71	1	0.7628
72	4	2.5215
73	3	1.8248
74	1	2.0785
75	2	1.5477
76	2	2.5210
77	2	1.9341
78	2	2.2191
79	2	1.9416
80	4	1.5854
81	2	2.3998
82	2	0.9464
83	4	1.1573
84	4	2.4868
85	2	1.1935
86	4	1.8508
87	3	2.1711
88	2	2.6291
89	2	1.0787
90	1	2.7129
91	4	1.5139
92	4	1.7369
93	2	1.5182
94	2	1.4033
95	2	1.7876
96	2	1.4812
97	3	0.7434
98	4	2.7770
99	1	0.6582
100	2	1.3234

Criterion Based on Final Seeds =	0.8161

Replace=FULL Drift Radius=0 Maxclusters=4 Maxiter=1

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	11	0.8522	2.7129		2	2.1939
2	52	0.8489	2.6640		4	1.8278
3	12	0.7853	2.6106		2	2.2357
4	25	0.8040	2.7770		2	1.8278

Statistics for Variables				
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
x6	1.00000	0.99322	0.043404	0.045373
x8	1.00000	0.86385	0.276370	0.381923
x12	1.00000	0.74610	0.460209	0.852570
x15	1.00000	0.83180	0.329081	0.490493
x18	1.00000	0.68733	0.541892	1.182892
OVER-ALL	1.00000	0.83111	0.330191	0.492964

Pseudo F Statistic =	15.77

Approximate Expected Over-All R-Squared =	0.39795

Cubic Clustering Criterion =	-3.935

WARNING: The two values above are invalid for correlated variables.

Cluster Means					
Cluster	x6	x8	x12	x15	x18
1	0.240248881	0.201068906	1.903011036	0.520594649	0.600831134
2	0.137453211	0.074387042	-0.175679513	-0.154562514	0.500854479
3	-0.335415217	-1.365605365	-0.176875100	1.322797514	-0.219215480
4	-0.230612882	0.412295208	-0.387011422	-0.542514423	-1.200919585

Cluster Standard Deviations					
Cluster	x6	x8	x12	x15	x18
1	1.260315192	0.926395845	0.497452378	0.619739633	0.743508472
2	1.006318275	0.896395030	0.776923168	0.888159252	0.628000121
3	0.925499262	0.629884824	0.659817359	0.856357217	0.813074753
4	0.860003627	0.859098441	0.800823977	0.771448926	0.719926295

Replace=FULL Drift Radius=0 Maxclusters=4 Maxiter=1

Distance Between Cluster Centroids				
Nearest Cluster	**1**	**2**	**3**	**4**
1	.	2.193946711	2.903065560	3.144365875
2	2.193946711	.	2.235682783	1.827801247
3	2.903065560	2.235682783	.	2.767526481
4	3.144365875	1.827801247	2.767526481	.

Replace=FULL Drift Radius=0 Maxclusters=5 Maxiter=1

Initial Seeds					
Cluster	**x6**	**x8**	**x12**	**x15**	**x18**
1	-1.654396764	1.525688290	2.869479866	-0.167442723	1.380649046
2	0.637408277	1.983068077	-1.233773761	-0.435351080	0.836014314
3	1.496835167	-1.087910494	1.470643402	1.373030331	0.019062216
4	-0.651732058	-0.238490889	0.258318467	-1.774892866	-2.567952762
5	-2.012491301	-2.656069764	-0.207960354	1.640938688	-1.070207249

Replace=FULL Drift Radius=0 Maxclusters=5 Maxiter=1

Cluster Listing		
Obs	Cluster	Distance from Seed
1	3	1.8855
2	2	2.0889
3	3	1.2120
4	4	1.1475
5	2	1.1326
6	5	1.9627
7	4	1.8229
8	4	1.6388
9	1	1.6096
10	3	1.1006
11	2	1.7228
12	3	2.1865
13	3	1.3565
14	2	1.7463
15	3	1.7309
16	2	1.6453
17	1	1.3224
18	1	1.5519
19	1	1.6556
20	3	1.9211
21	4	1.8323
22	3	2.6110
23	2	0.9881
24	3	2.4026
25	4	1.6065
26	4	1.5902
27	2	1.3804
28	2	1.4893
29	2	0.9532
30	5	0.6706
31	4	2.0864
32	2	1.7692
33	4	0.9528
34	2	1.3088
35	3	1.6147

Replace=FULL Drift Radius=0 Maxclusters=5 Maxiter=1

Cluster Listing		
Obs	Cluster	Distance from Seed
36	2	2.5673
37	2	1.0615
38	2	1.8455
39	3	1.7434
40	5	1.5137
41	4	2.5393
42	2	1.2571
43	3	1.3295
44	1	1.3895
45	2	2.1689
46	1	1.3259
47	2	1.7280
48	1	0.8881
49	3	2.2869
50	3	1.1555
51	4	1.1080
52	2	2.3630
53	2	2.8238
54	4	0.5617
55	3	1.5385
56	2	1.6343
57	3	2.4498
58	3	1.0466
59	2	1.3813
60	2	1.6794
61	2	1.8192
62	3	1.6593
63	5	1.6910
64	4	1.6400
65	4	0.5134
66	3	2.2233
67	1	1.5540
68	4	2.2723
69	4	1.5301
70	5	1.4101

Replace=FULL Drift Radius=0 Maxclusters=5 Maxiter=1

Cluster Listing		
Obs	Cluster	Distance from Seed
71	3	1.6455
72	4	1.8289
73	3	1.9422
74	3	2.4296
75	3	1.4042
76	2	1.7800
77	2	1.6439
78	2	2.2162
79	3	2.0403
80	4	1.2997
81	2	1.5903
82	2	0.3834
83	4	1.3420
84	4	2.8272
85	2	1.3168
86	4	1.2701
87	5	1.6962
88	2	2.1460
89	2	1.0411
90	1	1.7250
91	2	1.6116
92	4	1.2475
93	2	1.3843
94	3	1.2181
95	3	1.7238
96	3	1.5696
97	5	0.8147
98	4	2.5969
99	3	1.3381
100	2	1.5586

Criterion Based on Final Seeds =	0.7573

Replace=FULL Drift Radius=0 Maxclusters=5 Maxiter=1

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	9	0.6275	1.7250		3	2.2054
2	34	0.7692	2.8238		3	1.8877
3	28	0.8049	2.6110		2	1.8877
4	22	0.7797	2.8272		2	2.1745
5	7	0.7067	1.9627		3	2.6856

Statistics for Variables				
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
x6	1.00000	0.68979	0.543411	1.190151
x8	1.00000	0.82843	0.341433	0.518449
x12	1.00000	0.70688	0.520511	1.085553
x15	1.00000	0.85032	0.306174	0.441283
x18	1.00000	0.74767	0.463581	0.864216
OVER-ALL	1.00000	0.76731	0.435022	0.769980

Pseudo F Statistic =	18.29

Approximate Expected Over-All R-Squared =	0.48419

Cubic Clustering Criterion =	-3.067

WARNING: The two values above are invalid for correlated variables.

Cluster Means					
Cluster	x6	x8	x12	x15	x18
1	-1.542989574	0.639968702	0.880023562	0.167442723	1.108331680
2	0.702707869	0.461031138	-0.712638608	0.196991439	0.155220899
3	0.325354466	-0.327166562	0.924431069	0.035880584	0.461577936
4	-0.645221249	-0.021680990	-0.377516289	-0.849391269	-1.057829187
5	-0.702888421	-1.685304502	-0.181315850	1.353894019	-0.700633681

Replace=FULL Drift Radius=0 Maxclusters=5 Maxiter=1

Cluster Standard Deviations					
Cluster	**x6**	**x8**	**x12**	**x15**	**x18**
1	0.175835398	0.688485427	1.061232569	0.475961647	0.333519298
2	0.553739022	0.804042840	0.647772249	1.013103399	0.748028322
3	0.896474608	0.852339858	0.704163605	0.803411240	0.753671075
4	0.637486478	0.950591793	0.608896833	0.739392272	0.901128426
5	0.875894931	0.496796610	0.757066594	0.826395494	0.476323772

Distance Between Cluster Centroids					
Nearest Cluster	**1**	**2**	**3**	**4**	**5**
1	.	2.919082344	2.205364165	2.924274103	3.452409176
2	2.919082344	.	1.887712815	2.174543206	2.989259461
3	2.205364165	1.887712815	.	2.413029181	2.685589992
4	2.924274103	2.174543206	2.413029181	.	2.791329073
5	3.452409176	2.989259461	2.685589992	2.791329073	.