Florida Atlantic University
COLLEGE OF BUSINESS

ISM 6136 – Datamining/Predictive Analytics
Class Assignment 9
10 points

TASK: Performing Clustering – Data Mining Task using XLMiner or RapidMiner

1. For the Public Utilities Dataset: 22 US Utility firms and 8 variables. Using the **Hierarchical clustering algorithm** (Try out 3 different number of clusters values and determine the following:

    a) The cluster with maximum no. of utilities that are operating in a similar manner based on the 8 variables

Cluster Labels

Record ID	Cluster	Sub-Cluster
Record 1	1	1
Record 2	1	2
Record 3	1	3
Record 4	1	2
Record 6	1	5
Record 9	1	8
Record 10	1	2
Record 13	1	2
Record 14	1	1
Record 18	1	1
Record 19	1	1
Record 20	1	2
Record 22	1	2
Record 5	2	4
Record 7	3	6
Record 12	3	6
Record 15	3	6
Record 17	3	10
Record 21	3	6
Record 8	4	7
Record 11	4	9
Record 16	4	7

Based on these results, cluster number 1 has the most similar acting utilities which are: Arizona, Boston, Central, Common, Florida, Kentucky, Madison, Northern, Oklahoma, Southern, Texas, Wisconsin, Virginia.

b) Any outlier utilities that are not combined with another one to form a cluster?

There is an outlier utility in this case (screenshot above), it is utility Consolid.

c) Any other cluster# identification and talk about the number and which utilities are under it?

*Screenshots provided below

Interestingly, the cluster 4 in the screenshot in part A is the same as cluster 2 below which both consist of Idaho, Nevada, Puget, but Nevada becomes an outlier when the max number of clusters is increased (shown in the second screenshot). From the screenshot above cluster 3 seems to be comprised of coastal areas or areas near water as the utilities are Hawaiian, New England, Pacific, San Diego, United.

Cluster Labels

Record ID	Cluster	Sub-Cluster
Record 1	1	1
Record 2	1	2
Record 3	1	3
Record 4	1	2
Record 5	1	4
Record 6	1	5
Record 7	1	6
Record 9	1	8
Record 10	1	2
Record 12	1	6
Record 13	1	2
Record 14	1	1
Record 15	1	6
Record 17	1	10
Record 18	1	1
Record 19	1	1
Record 20	1	2
Record 21	1	6
Record 22	1	2
Record 8	2	7
Record 11	2	9
Record 16	2	7

Cluster Labels

Record ID	Cluster	Sub-Cluster
Record 1	1	1
Record 3	1	3
Record 6	1	5
Record 9	1	8
Record 14	1	1
Record 18	1	1
Record 19	1	1
Record 2	2	2
Record 4	2	2
Record 10	2	2
Record 13	2	2
Record 20	2	2
Record 22	2	2
Record 5	3	4
Record 7	4	6
Record 12	4	6
Record 15	4	6
Record 21	4	6
Record 8	5	7
Record 16	5	7
Record 11	6	9
Record 17	7	10

Note: This is an example where clustering would be useful as a study to predict the cost impact of deregulation. To perform the requisite analysis, economists would be required to build a detailed cost model of the various utilities. It would save a considerable amount of time and effort by clustering similar types of utilities, building a detailed cost model for just one typical utility in each cluster, then scaling up from these models to estimate results for all utilities.

2. Using **K-means clustering algorithm** on the cereal dataset - find out the following and and explain along with screen shots for each of the answer. Try out at least 3 different number of clusters to determine the following:

   Explain how many clusters you had to create to get these answers.

   a) Cluster of 'healthy cereals (low fat, low salt etc)'. Which cereals are part of that cluster?
   Cluster 2 on tab KMC Clusters which is 4 max clusters had the most healthy cereals out of the clusters. The cereals included Puffed rice, Puffed wheat, Shredded wheat, shredded wheat n bran, shredded wheat spoon size, raisin squares, strawberry fruit wheels, cream of wheat (these are some of the top healthy cereals when the original data is sorted based on low fat and low sodium and low sugars which is what I sorted based on what a healthy cereal would have).

Record 9	2	3.0385605	2.2996022	5.8199688	3.1771196
Record 16	2	3.5636389	2.6768845	8.5257763	4.5797256
Record 17	2	3.7183798	2.4074918	7.9448223	4.5413589
Record 21	2	4.7606238	2.1120079	7.4754855	4.643692
Record 22	2	3.621033	2.0171666	7.6248623	3.6446442
Record 24	2	3.0271961	1.6485895	6.866466	2.9935063
Record 27	2	3.8185121	2.4059474	5.9654802	3.3360213
Record 33	2	3.352354	1.8006009	5.6964864	2.0045159
Record 34	2	3.7913358	1.6490887	6.0028642	2.6808866
Record 41	2	3.0606281	2.2529993	8.0589837	3.5439796
Record 44	2	4.2295556	2.4992929	6.8394858	3.3995029
Record 51	2	4.3253877	1.6130594	5.6894923	3.3539705
Record 54	2	5.2296956	4.2820456	8.1131476	4.7624728
Record 55	2	5.559922	4.0855637	7.5314074	6.0465231
Record 56	2	5.6371596	4.0455754	6.5631178	5.6948889
Record 61	2	3.948675	2.510771	6.1831779	3.5465414
Record 62	2	3.7075033	2.8220461	8.7754636	4.9019014
Record 63	2	3.5842135	2.7378711	8.4947589	4.5668223
Record 64	2	5.1262638	3.0183316	6.4938717	5.2206198
Record 65	2	5.758313	3.3185861	6.1694833	5.2135419
Record 66	2	5.6072451	3.1135434	6.6001785	5.2127504
Record 68	2	5.0947469	3.662398	7.3059858	4.4723276
Record 69	2	3.804744	2.0639083	6.0659925	3.7247127
Record 70	2	4.6938438	4.0616046	8.6344585	4.3809868
Record 73	2	3.3993476	2.4617554	7.8248902	3.2787874
Record 75	2	3.4632056	1.9844889	6.196254	3.101911
Record 76	2	3.4489253	1.8002646	6.1677831	3.0688069
Record 1	3	7.1818087	6.1425306	1.2131571	5.3647422
Record 3	3	7.023131	6.0430571	1.7575334	5.1638969

   b) Cluster with lowest consumer ratings?
   Cluster 10, which had the most cereals with the lowest ratings. On tab KMC Clusters3

Record 6	10	4.7522864	6.0133418	7.5441525	4.128078	5.7605015	3.1113376	5.3069463	3.2634515	3.2075902	1.645437894
Record 7	10	4.6076797	5.3633285	7.7190024	3.4721275	5.4922522	3.716401	5.177009	3.1182086	4.0731835	1.473702102
Record 11	10	4.7671453	6.6130802	8.7665668	5.0962793	5.7493057	4.0438572	6.5041078	3.6192888	4.0729217	1.545974609
Record 13	10	4.8334961	6.6507529	8.7660588	5.1982603	5.896164	4.0881846	6.4535018	3.7278824	3.7518301	2.33276289
Record 15	10	4.5931008	5.8705005	8.2722851	4.3331187	5.5340749	3.7715063	5.872723	3.2700144	4.1222748	0.775619342
Record 18	10	4.6492664	4.9474807	7.9893191	3.2537226	5.5594155	3.7190215	4.8157671	3.3009937	4.4004244	1.630020104
Record 19	10	4.5784246	5.8914588	8.2217854	4.3287379	5.5454313	3.7770567	5.8751214	3.2380158	4.0895402	0.83002129
Record 25	10	4.4307831	5.354607	7.6590897	3.4571643	5.4932167	3.5495253	5.1756729	2.866	3.3952259	0.979236162
Record 26	10	4.8718107	5.7986843	8.2651013	4.1534862	5.3375269	2.9299507	5.1067287	3.6426804	4.8981177	1.570505844
Record 30	10	4.5255065	5.4075916	8.3121717	3.8494279	5.4883799	3.6157365	5.3859263	3.2173025	4.0430636	0.765493699
Record 31	10	5.4389251	5.3996385	8.164029	3.5455425	6.3817078	4.1590443	5.0079583	4.1975534	4.7069614	2.33949478
Record 32	10	4.4037511	6.0597836	8.349916	4.7145449	4.7410819	3.0033689	5.8205276	3.134907	4.4867233	1.721039466
Record 36	10	4.5604778	6.400953	8.3175375	4.8101601	5.5822715	3.7605097	6.1797061	3.2674287	3.7456693	1.450711081
Record 37	10	4.5762675	6.2433214	7.1732416	4.1657153	5.1398282	2.4649236	5.2016829	2.9415512	3.5559817	2.134260802
Record 38	10	4.9125391	5.779374	8.4850862	4.1443715	5.4457498	3.0766504	5.1707583	3.7371368	4.9219604	1.490546389
Record 43	10	4.2650185	5.6295809	7.8342091	3.8275081	5.1460495	3.2272762	5.4132241	2.7058655	3.5159192	0.664264126
Record 49	10	3.9386123	5.6510274	8.0251608	3.6766509	4.6945819	2.5739183	5.0029616	2.392554	3.5268384	1.191348718
Record 67	10	4.8895832	5.5332465	7.7827647	3.6525643	6.1746482	4.3423263	5.52434	3.4551163	3.58062	1.721353921
Record 74	10	4.5207864	5.4342954	8.316452	3.8844209	5.4681101	3.6012468	5.4121821	3.2126553	4.0536107	0.736157118

c) Any other cluster identification?

61	Record 1	3	7.1818087	6.1425306	1.2131571	5.3647422
62	Record 3	3	7.023131	6.0430571	1.7575334	5.1638969
63	Record 4	3	9.7759581	7.9953642	2.3537191	8.0487129
64	Record 2	4	5.3829165	6.1643132	7.8958115	4.206245
65	Record 5	4	2.6630251	3.2822743	6.7186511	1.6919678
66	Record 8	4	3.1534515	3.2818714	7.2422239	1.6226959
67	Record 10	4	4.4058608	2.9708393	4.1190298	2.9478692
68	Record 12	4	5.4969865	4.2751608	7.075608	4.2358928
69	Record 14	4	2.8633963	2.8635614	6.299525	1.0869574
70	Record 20	4	3.8942884	4.1206808	5.6681038	1.7379623

On tab KMC Clusters

The highlighted cluster 3 includes cereals 100% bran, all bran, and all bran extra fiber, while cluster 4 the first record is 100% natural bran which by name would look to be included in cluster 3 but is more similar to higher calorie cereals as what the records after it look to be.

*For parts A and B, different KMC were used as the number of total clusters would have roughly the same number of records in multiple clusters so I chose to use two outputs that held more records of what was being asked.