Assignment 4

Garcia, Jorge A.
Department of Physics
New Mexico State University
(Dated: November 11 2019)

Course: C S 508

Instructor: Dr. Tuan Le

I. DATA PREPROCESSING

The script "prob1_preprocess.py" first removes any instances with NaN values present. Afterwards, points with outliers present were removed by converting the data to z-scores and eliminating instances with a z-score $z > 3$ in any of its dimensions. The new, clean data was saved in the "data.csv" file.

II. K-MEANS

Clustering using the K-Means algorithm is done in the "prob2_kmeans.py" script. The resulting average silhouette score for each number of clusters is seen in the top plot of Figure 1. To take into account the statistical behavior of the algorithm due to initialization, the best number of clusters is that with the most potential instead of just the highest average score, with the potential being defined as the sum of the average score and its standard deviation. Given enough random initializations, this score ceiling for a specific number of clusters can be achieved. Using K-Means, the maximum silhouette score was found using 20 clusters. The centroids of these clusters are shown in Table I.

The bottom plot in Figure 1 shows the data projected into two dimensions using PCA, with the points colored according to their corresponding cluster ID. The clusters are not well separated in the visualization; arguably, cluster #13 is only one that is well separated from the other groups. With an average silhouette score of 0.183, it is a fairly low score and thus separation of the clusters is not the best.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
1.29	4.89	4.24	4.03	4.54	3.70	3.02	2.53	2.66	3.04	1.94	1.80	1.67	1.67	1.64	1.13	0.36	0.24	0.29	0.61	0.72	0.92	1.06	1.22
1.06	4.98	2.86	2.70	2.69	3.17	4.10	2.76	4.10	2.81	2.40	2.32	2.01	3.63	4.87	0.70	0.69	0.69	0.59	0.71	0.78	0.83	0.94	1.08
1.22	2.09	2.82	3.62	4.35	3.56	2.34	2.21	1.97	1.90	1.71	2.00	1.75	1.41	1.05	0.86	0.68	0.70	0.66	0.67	0.68	0.73	0.79	0.96
1.05	1.46	2.01	2.41	3.08	4.72	4.76	3.24	4.56	2.69	1.58	1.57	3.80	4.82	4.10	0.61	0.59	0.60	0.69	0.72	0.94	0.99	0.98	1.01
1.97	2.20	2.80	4.22	4.06	2.93	3.28	2.96	3.58	3.44	4.18	1.66	5.00	2.79	1.75	1.02	0.90	0.78	0.72	0.71	0.72	4.23	3.04	1.74
1.28	1.40	1.72	2.43	2.60	3.49	4.32	4.28	4.88	4.20	1.41	1.36	1.28	1.32	0.94	0.65	0.41	0.26	0.35	0.74	0.82	1.12	1.26	1.27
1.52	2.03	3.12	4.64	4.36	3.59	2.39	1.94	2.01	2.31	2.45	1.73	1.30	1.12	1.20	1.21	0.87	0.61	0.53	0.52	0.59	4.82	1.01	1.65
1.28	1.43	1.62	2.16	2.29	2.35	3.44	3.62	4.39	4.40	4.98	2.15	1.78	1.32	1.22	0.92	0.66	0.32	0.27	0.26	0.49	0.91	1.03	1.23
0.37	0.99	1.51	1.55	1.69	1.72	2.98	2.31	3.07	3.22	4.16	4.80	4.83	4.12	0.92	0.88	0.85	0.84	0.89	0.34	0.22	0.23	0.29	0.43
0.92	4.86	1.87	1.91	2.24	3.30	4.82	3.64	4.23	3.40	2.58	2.19	1.67	2.96	1.10	0.73	0.71	0.72	0.70	0.82	0.81	0.79	0.81	0.99
0.65	0.75	4.47	1.41	1.37	2.19	4.54	3.02	4.92	4.12	2.77	1.78	1.64	3.70	2.73	0.78	0.77	0.78	0.85	0.80	0.76	0.46	0.46	0.52
1.18	1.43	2.12	2.06	2.07	2.03	3.90	2.20	2.90	2.79	2.82	2.77	3.30	4.80	4.78	0.63	0.58	0.59	0.81	0.79	0.93	1.06	1.04	1.19
0.54	0.65	1.50	1.54	1.53	1.54	2.55	1.71	2.59	2.93	3.07	4.89	4.80	5.00	5.00	0.74	0.72	0.73	0.88	0.79	0.54	0.51	0.49	0.51
1.14	1.71	4.51	4.35	4.48	4.29	4.72	2.24	2.59	2.59	2.80	1.89	1.73	1.54	1.53	1.18	0.75	0.25	0.24	0.23	0.43	0.73	0.95	1.00
1.66	2.59	2.43	3.59	4.45	3.38	3.64	2.99	2.78	2.70	2.53	3.38	3.11	1.92	1.08	0.78	0.89	0.73	0.81	0.76	0.80	0.91	2.95	3.98
2.48	2.76	2.52	2.24	2.02	1.83	1.82	1.56	1.59	1.55	1.52	1.49	1.55	1.79	1.80	1.22	1.18	1.35	1.61	1.64	1.37	2.62	2.51	3.19
1.02	2.53	1.63	1.91	2.13	2.90	4.72	3.89	4.99	4.60	2.69	1.84	1.44	1.51	4.98	0.66	0.57	0.56	0.64	0.79	0.87	1.02	1.00	1.01
1.13	1.31	1.85	2.66	3.75	4.37	4.95	2.00	3.20	1.88	1.66	1.65	1.72	1.76	1.96	0.75	0.62	0.57	0.67	0.65	0.75	0.91	1.12	1.12
1.60	2.37	3.65	4.07	3.95	3.15	2.98	2.20	2.17	2.21	2.58	1.33	1.02	0.92	1.14	0.94	0.95	0.75	0.70	0.70	0.73	4.03	4.92	2.04
1.56	1.77	1.83	2.53	2.70	2.55	3.11	3.20	4.26	4.00	4.97	2.55	2.66	1.19	1.28	1.03	0.72	0.46	0.41	0.41	0.48	4.65	1.63	1.46

TABLE I. Cluster centroids using K-Means, showing their values across the attributes.

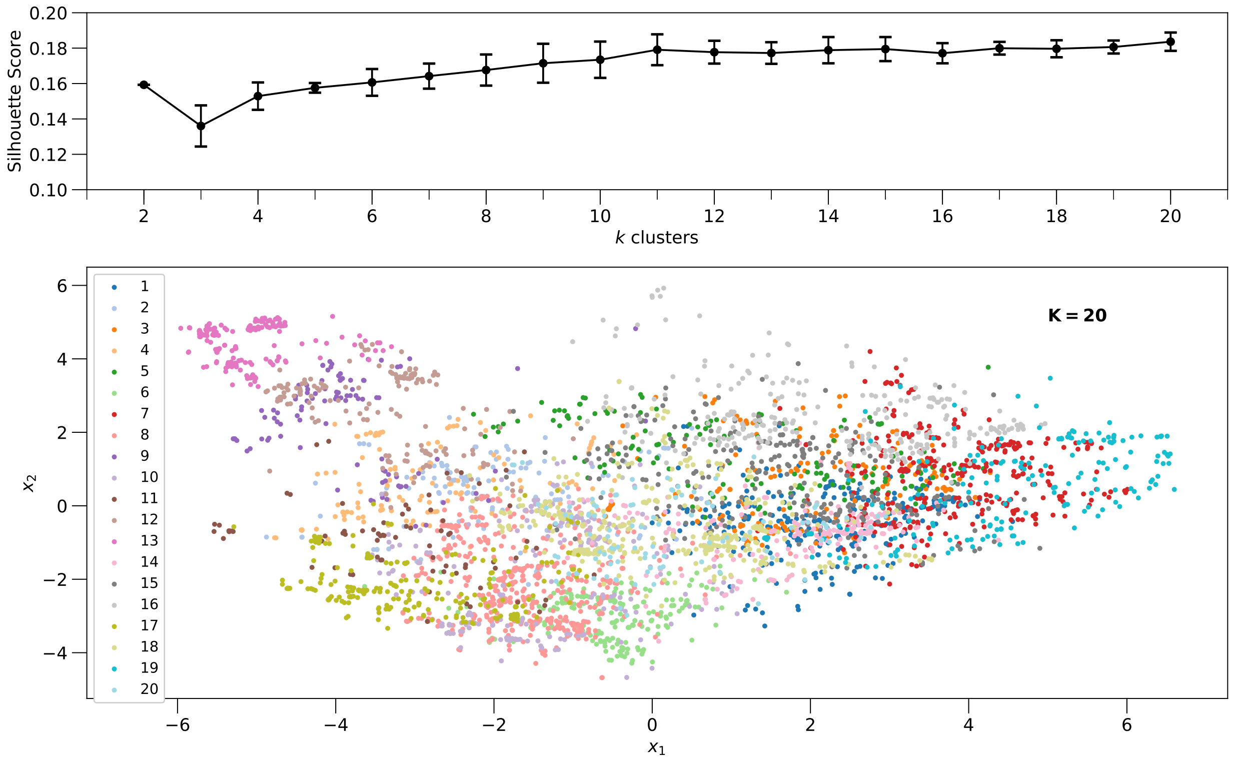

FIG. 1. Silhouette score as a function of number of clusters (top) and distribution of clusters (bottom) using K-Means.

III. K-MEANS++

Clustering using the K-Means++ algorithm is done in the "prob3_kmeans++.py" script. The resulting average silhouette score for each number of clusters is seen in the top plot of Figure 2. The best number of clusters is again defined as that with the highest silhouette score potential. This number of clusters was found to be 14 clusters. The centroids of these clusters are shown in Table II.

The bottom plot in Figure 2 shows 2-D projection using PCA. Again, the clusters aren't very well separated. Cluster 7 is the one that is best separated, having some overlap with cluster 14. The average silhouette score of 0.181 indicates that the cluster separation will be about the same as before, using K-Means.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
1.05	1.30	2.21	2.25	2.58	3.63	4.76	3.21	4.60	3.34	2.03	1.73	2.16	2.90	4.87	0.67	0.62	0.62	0.67	0.77	0.86	0.95	0.96	1.03
1.01	3.19	2.40	2.26	2.80	3.53	4.53	3.29	3.64	2.86	2.55	2.28	2.41	4.72	0.89	0.74	0.77	0.70	0.74	0.88	0.77	0.78	0.84	1.51
1.54	2.05	3.10	4.58	4.35	3.58	2.44	1.99	2.05	2.36	2.54	1.82	1.30	1.12	1.19	1.21	0.86	0.61	0.54	0.53	0.59	4.82	1.03	1.66
1.13	1.57	2.75	3.38	4.24	4.29	4.31	2.15	2.62	2.06	1.96	1.77	1.69	1.50	1.35	0.90	0.66	0.48	0.53	0.51	0.62	0.80	1.04	1.06
2.48	2.76	2.52	2.24	2.02	1.83	1.82	1.56	1.59	1.55	1.52	1.49	1.55	1.79	1.80	1.22	1.18	1.35	1.61	1.64	1.37	2.62	2.51	3.19
1.24	1.50	1.65	2.20	2.28	2.39	3.53	3.65	4.45	4.50	4.96	2.16	1.43	1.35	1.41	0.89	0.62	0.32	0.26	0.26	0.49	1.30	0.98	1.20
0.79	0.95	1.69	1.64	1.64	1.65	2.95	1.82	2.66	2.87	2.96	4.07	4.33	5.00	4.96	0.68	0.65	0.66	0.87	0.80	0.74	0.74	0.71	0.76
1.74	2.67	2.43	3.79	4.60	3.37	3.51	2.88	2.75	2.70	2.54	3.62	2.96	1.43	1.14	0.79	0.87	0.79	0.80	0.76	0.79	0.93	3.32	3.73
0.98	4.98	2.48	2.23	2.30	2.83	4.41	3.04	4.43	3.56	2.60	2.20	1.89	2.87	4.81	0.69	0.65	0.66	0.64	0.73	0.80	0.84	0.86	0.98
1.63	2.34	3.69	4.06	3.90	3.07	2.91	2.20	2.21	2.31	2.66	1.41	1.02	0.92	1.16	0.95	0.96	0.75	0.70	0.70	0.73	4.20	4.94	2.05
1.89	2.12	2.60	3.83	3.77	2.75	3.28	2.98	3.74	3.45	4.32	1.72	4.97	2.42	1.63	1.04	0.90	0.75	0.70	0.68	0.70	4.17	2.85	1.72
1.29	4.67	4.24	4.04	4.47	3.58	2.97	2.45	2.65	2.94	1.95	1.74	1.63	1.63	1.61	1.11	0.43	0.29	0.33	0.60	0.72	0.90	1.03	1.20
1.18	2.31	1.78	2.18	2.35	3.49	4.50	4.25	4.86	4.14	1.60	1.49	1.38	1.28	1.06	0.65	0.48	0.37	0.46	0.73	0.85	1.05	1.14	1.18
0.65	1.19	1.56	1.61	1.76	1.72	3.00	2.55	3.48	3.37	4.40	4.04	4.91	3.22	1.05	0.97	0.87	0.76	0.78	0.38	0.31	0.45	0.54	0.69

TABLE II. Cluster centroids using K-Means++, showing their values across the attributes.

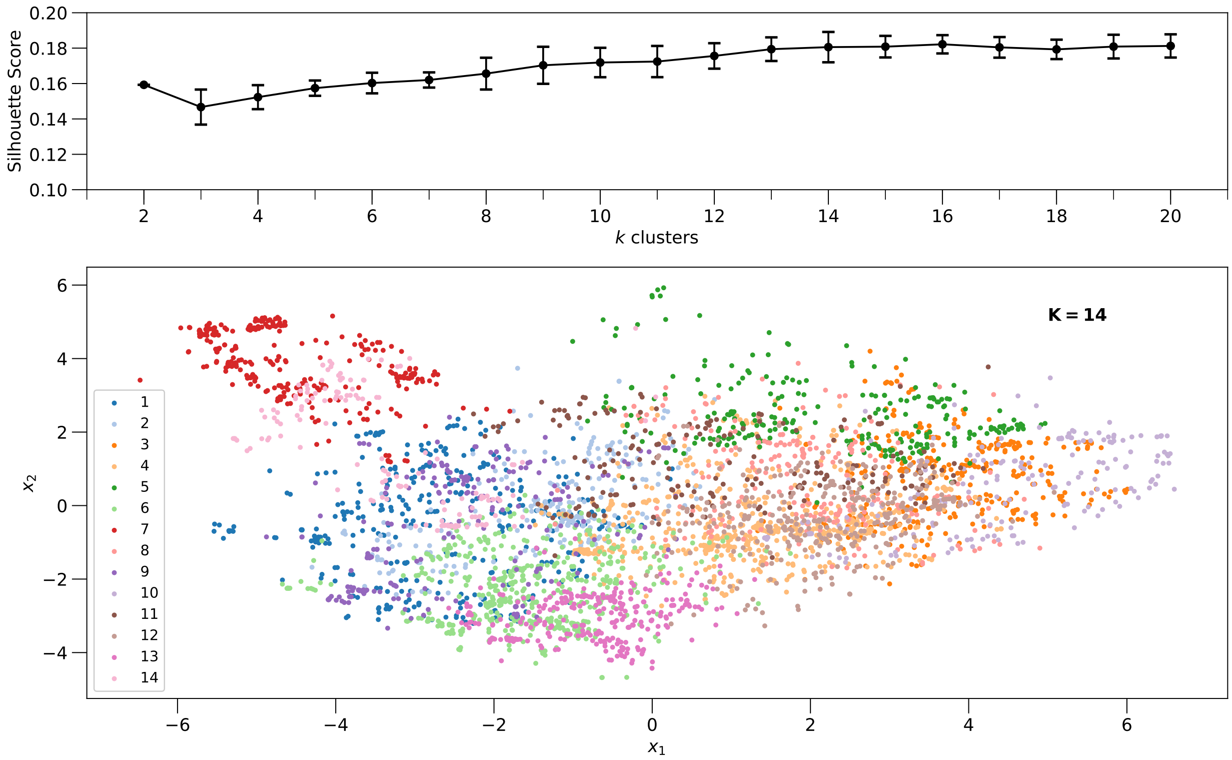

FIG. 2. Silhouette score as a function of number of clusters (top) and distribution of clusters (bottom) using K-Means++.

IV. AGGLOMERATIVE CLUSTERING

Clustering using the Agglomerative Clustering algorithm is done in the "prob4_aggclust.py" script. The resulting average silhouette score for each number of clusters is seen in the top plot of Figure 3. Since this is a deterministic method, the resulting silhouette score is constant and thus only the maximum score is considered. The best number of clusters was found to be 20 clusters. The centroids of these clusters are shown in Table III.

The bottom plot in Figure 3 shows 2-D projection using PCA. With a lower silhouette score of 0.170, the clustering looks messier in the projection. Cluster 14, a group of points that consistently appears, is the one that is best separated.

V. GAUSSIAN MIXTURE MODEL

Clustering using the Gaussian Mixture Model algorithm is done in the "prob5_gmm.py" script. The resulting average silhouette score for each number of clusters is seen in the top plot of Figure 4. For this method, the maximum silhouette potential is again used. The optimal number of clusters was found to be 19 clusters. The centroids of these clusters are shown in Table IV.

The bottom plot in Figure 4 shows 2-D projection using PCA. This method has the lowest of the silhouette scores, with the highest average score being 0.135. This score is translated in a much messier visualization, and the clusters not being easily separable.

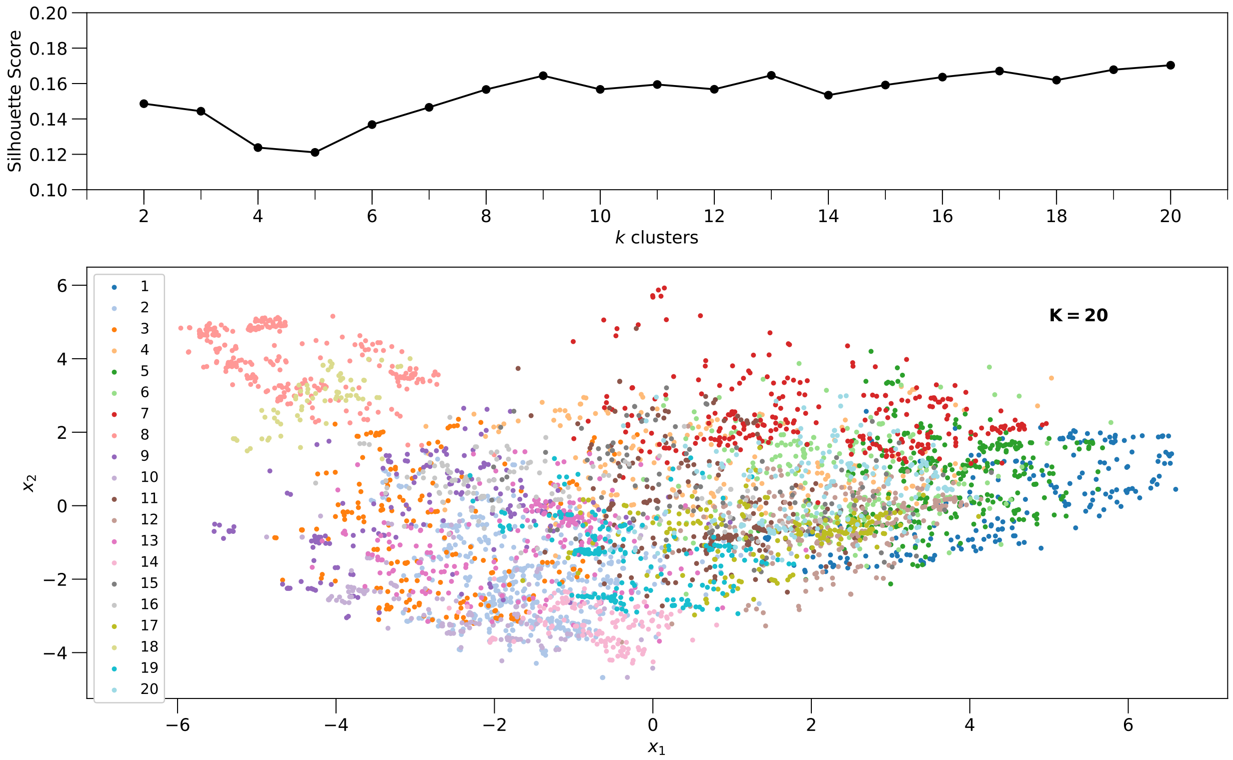

FIG. 3. Silhouette score as a function of number of clusters (top) and distribution of clusters (bottom) using Agglomerative Clustering.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
1.48	2.10	3.53	4.08	4.04	3.51	3.19	2.23	2.14	2.09	2.59	0.95	0.90	0.88	1.09	0.94	0.94	0.73	0.67	0.65	0.70	3.46	4.69	2.36
1.25	1.43	1.55	2.13	2.24	2.35	3.31	3.60	4.41	4.36	5.00	2.27	1.87	1.29	1.22	0.92	0.66	0.33	0.25	0.23	0.45	1.31	0.97	1.21
1.12	1.60	1.73	2.20	2.77	4.27	4.81	3.90	4.96	3.42	1.47	1.38	2.67	3.35	4.65	0.60	0.54	0.54	0.64	0.77	0.99	1.10	1.07	1.11
1.86	2.11	2.60	3.81	3.76	2.75	3.24	2.95	3.70	3.43	4.32	1.56	5.00	2.47	1.67	1.06	0.91	0.75	0.69	0.68	0.69	4.09	2.80	1.69
1.53	2.05	3.11	4.62	4.31	3.53	2.37	1.91	2.01	2.35	2.41	1.64	1.29	1.15	1.19	1.23	0.87	0.60	0.52	0.50	0.57	4.79	1.01	1.50
1.93	2.85	2.51	3.91	4.57	3.21	3.41	2.92	2.82	2.93	2.71	4.27	2.69	1.26	1.19	0.81	0.81	0.82	0.80	0.80	0.82	2.07	3.93	2.97
2.40	2.82	2.52	2.29	2.05	1.86	1.86	1.56	1.60	1.54	1.52	1.49	1.54	1.79	1.96	1.21	1.17	1.33	1.58	1.61	1.35	2.56	2.48	3.07
0.78	0.94	1.64	1.61	1.60	1.63	2.91	1.81	2.65	2.85	2.91	4.13	4.39	4.99	4.99	0.67	0.64	0.66	0.88	0.79	0.75	0.73	0.70	0.75
0.97	1.13	2.60	2.39	2.42	2.58	4.55	2.73	4.07	3.84	3.38	2.26	1.98	2.82	4.54	0.77	0.67	0.66	0.65	0.75	0.67	0.76	0.85	0.92
0.88	4.99	1.78	1.56	1.71	2.66	4.88	3.77	4.98	4.65	2.81	1.96	1.57	1.47	3.38	0.67	0.61	0.61	0.72	0.76	0.87	0.88	0.81	0.85
1.06	1.29	1.80	2.92	4.14	4.49	4.45	2.42	2.14	2.12	1.62	1.76	1.93	2.17	1.29	0.80	0.57	0.45	0.55	0.58	0.57	0.86	0.89	0.92
1.27	4.84	4.32	4.06	4.66	3.95	2.95	2.64	2.48	3.25	1.80	1.74	1.73	1.70	1.54	1.32	0.34	0.17	0.23	0.58	0.68	0.90	1.02	1.23
0.93	3.08	2.76	2.07	2.33	3.14	4.61	3.26	4.07	3.34	2.83	2.29	1.70	4.06	0.93	0.79	0.74	0.72	0.66	0.82	0.69	0.72	0.78	0.83
1.32	1.42	1.58	2.45	2.52	3.35	4.26	4.16	4.95	4.96	1.46	1.40	1.29	1.21	0.89	0.65	0.37	0.20	0.30	0.74	0.79	1.13	1.30	1.30
1.34	2.51	2.44	2.96	4.07	3.67	3.90	3.08	2.75	2.57	2.40	2.36	2.89	2.85	1.00	0.77	0.98	0.65	0.86	0.78	0.77	0.82	0.93	4.90
1.15	4.80	2.88	3.04	3.06	3.17	3.79	2.56	4.19	2.66	2.44	2.44	2.02	3.59	4.56	0.61	0.59	0.60	0.46	0.60	0.81	0.86	0.98	1.11
1.04	1.90	4.79	4.15	4.34	3.96	4.51	2.42	2.69	2.72	2.76	1.67	1.75	1.58	1.73	1.10	0.71	0.15	0.18	0.24	0.42	0.69	0.89	0.99
0.25	0.80	1.46	1.45	1.46	1.49	2.99	2.13	2.83	3.13	3.95	4.76	4.82	5.00	0.86	0.85	0.81	0.85	1.05	0.35	0.19	0.18	0.18	0.39
1.08	1.79	1.81	2.33	3.02	4.13	4.92	2.46	4.74	1.31	1.28	1.33	1.40	1.44	2.25	0.64	0.61	0.60	0.76	0.73	0.99	1.02	1.00	1.02
1.53	2.92	3.01	3.77	4.45	3.62	2.94	1.97	2.10	1.88	2.09	2.66	1.98	0.98	0.92	0.85	0.75	0.92	0.85	0.77	0.79	0.80	0.85	0.90

TABLE III. Cluster centroids using Agglomerative Clustering, showing their values across the attributes.

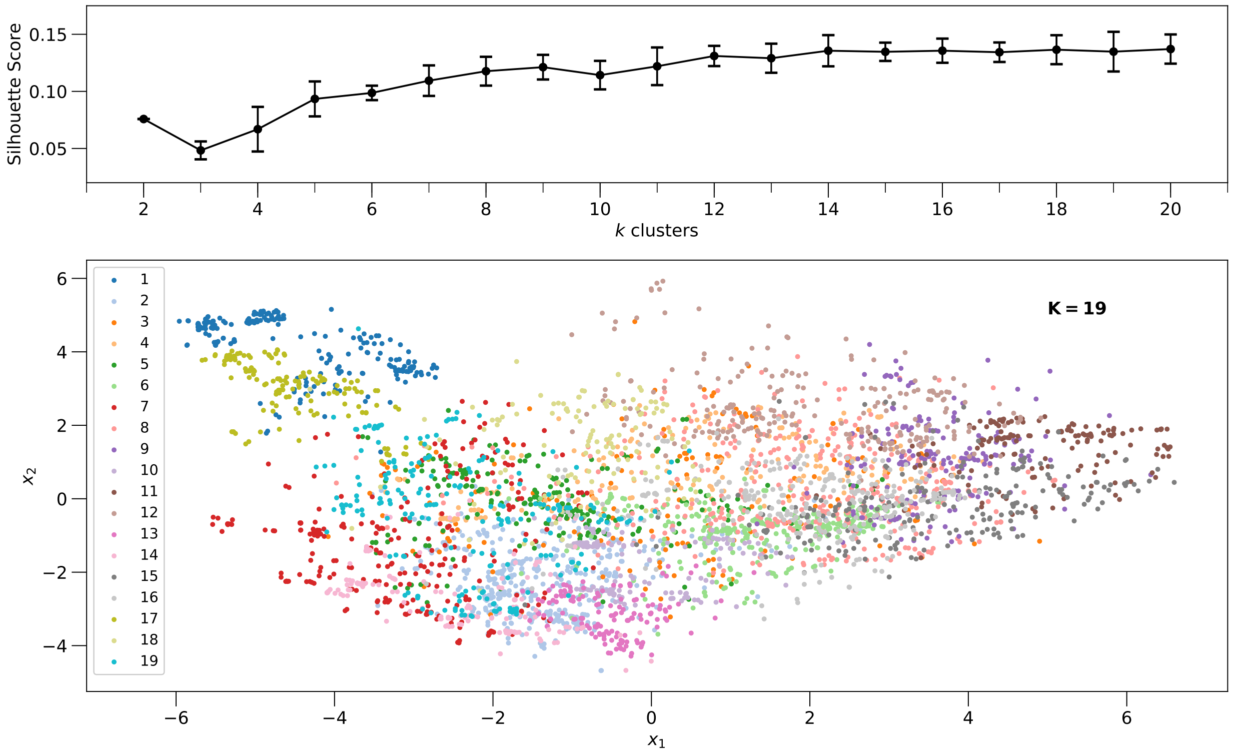

FIG. 4. Silhouette score as a function of number of clusters (top) and distribution of clusters (bottom) using Gaussian Mixture Model.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24
0.55	0.82	1.58	1.54	1.54	1.55	1.68	1.75	2.71	2.93	3.20	4.34	4.40	5.00	4.17	0.73	0.69	0.72	0.90	0.64	0.53	0.51	0.48	0.57
1.32	1.44	1.57	2.29	2.42	2.46	3.32	3.62	4.39	4.45	5.00	2.16	1.49	1.30	1.34	0.91	0.61	0.26	0.20	0.19	0.44	1.46	1.02	1.28
1.06	1.88	2.06	2.53	3.78	3.75	3.85	2.75	2.39	2.22	2.39	2.53	2.08	2.16	0.93	0.80	0.85	0.90	1.03	0.67	0.57	0.64	0.81	1.04
1.77	2.06	2.42	3.49	3.49	2.59	3.24	3.02	3.83	3.42	4.30	2.09	5.00	1.86	1.15	1.06	0.90	0.73	0.69	0.67	0.68	3.28	2.84	1.62
1.03	5.00	2.56	2.72	2.87	3.35	4.29	2.99	3.80	2.81	2.47	2.41	1.92	3.74	3.04	0.71	0.70	0.71	0.57	0.83	0.78	0.79	0.89	1.10
1.09	2.04	3.84	3.93	4.40	4.28	4.49	2.36	2.68	2.83	2.48	1.69	1.69	1.64	1.75	1.05	0.61	0.08	0.08	0.20	0.38	0.75	0.90	0.95
0.82	1.42	2.74	1.64	1.67	2.80	4.74	3.20	4.67	4.07	3.13	2.06	1.77	2.95	3.34	0.72	0.69	0.70	0.75	0.79	0.74	0.75	0.72	0.76
1.56	2.47	2.40	3.67	4.61	3.61	3.54	2.59	2.55	2.43	2.44	2.97	2.37	1.33	1.09	0.84	0.82	0.84	0.77	0.75	0.77	0.82	2.43	2.68
1.45	1.66	2.39	4.69	4.57	3.94	2.36	1.89	1.88	2.33	2.37	2.10	1.70	1.25	1.27	1.29	0.95	0.55	0.45	0.40	0.46	4.67	1.34	1.43
1.06	1.69	1.76	2.38	2.95	4.10	4.91	2.83	5.00	1.30	1.28	1.28	1.40	1.46	1.35	0.62	0.58	0.57	0.72	0.75	0.99	1.04	1.02	1.04
1.67	2.75	4.02	4.23	3.63	2.03	1.96	1.64	1.68	1.91	1.39	0.88	0.87	0.87	0.91	0.96	0.91	0.70	0.69	0.70	0.73	4.45	3.27	2.12
2.46	2.75	2.47	2.25	2.02	1.84	1.90	1.55	1.58	1.55	1.55	1.54	1.61	1.87	1.91	1.26	1.21	1.39	1.67	1.72	1.42	2.46	2.45	3.07
1.37	1.44	1.75	2.51	2.62	3.26	4.13	4.16	4.84	5.00	1.36	1.32	1.26	1.25	0.88	0.65	0.34	0.14	0.23	0.75	0.77	1.15	1.35	1.36
0.87	4.99	1.78	1.53	1.67	2.60	4.88	3.76	5.00	4.70	2.80	1.97	1.57	1.36	3.27	0.67	0.60	0.61	0.71	0.75	0.88	0.88	0.81	0.84
1.73	2.35	3.61	3.82	3.98	3.72	3.28	2.58	2.58	2.61	3.47	1.98	0.99	0.95	1.24	0.99	0.81	0.74	0.68	0.71	0.77	4.17	3.00	2.25
1.28	3.95	3.79	3.76	4.14	3.59	2.97	2.36	2.59	2.68	1.59	1.58	1.61	1.62	1.83	1.15	0.34	0.30	0.37	0.63	0.81	0.96	1.02	1.22
0.88	1.07	1.62	1.70	1.70	1.73	5.00	2.10	2.64	2.88	3.08	4.05	4.46	5.00	4.14	0.68	0.65	0.67	0.90	0.81	0.75	0.81	0.80	0.86
1.33	2.23	2.57	3.08	3.74	3.80	3.92	3.14	3.36	2.83	2.67	1.93	4.18	4.89	2.04	0.78	0.90	0.58	0.78	0.77	0.85	1.59	1.00	2.55
1.13	1.45	1.86	2.44	2.96	4.16	4.86	3.33	4.65	2.95	1.45	1.48	2.42	3.06	4.94	0.61	0.58	0.58	0.65	0.77	0.96	1.07	1.05	1.10

TABLE IV. Cluster centroids using Gaussian Mixture Model, showing their values across the attributes.

VI. COMPARISON

Figure 5 plots the silhouette coefficient across the various cluster sizes for all 4 methods used. K-Means and K-Means++ perform the same within their uncertainties. The highest silhouette potential is seen with K-Means++ when considering 14 clusters, and thus this is likely the best model and parameter to use for clustering this data set.

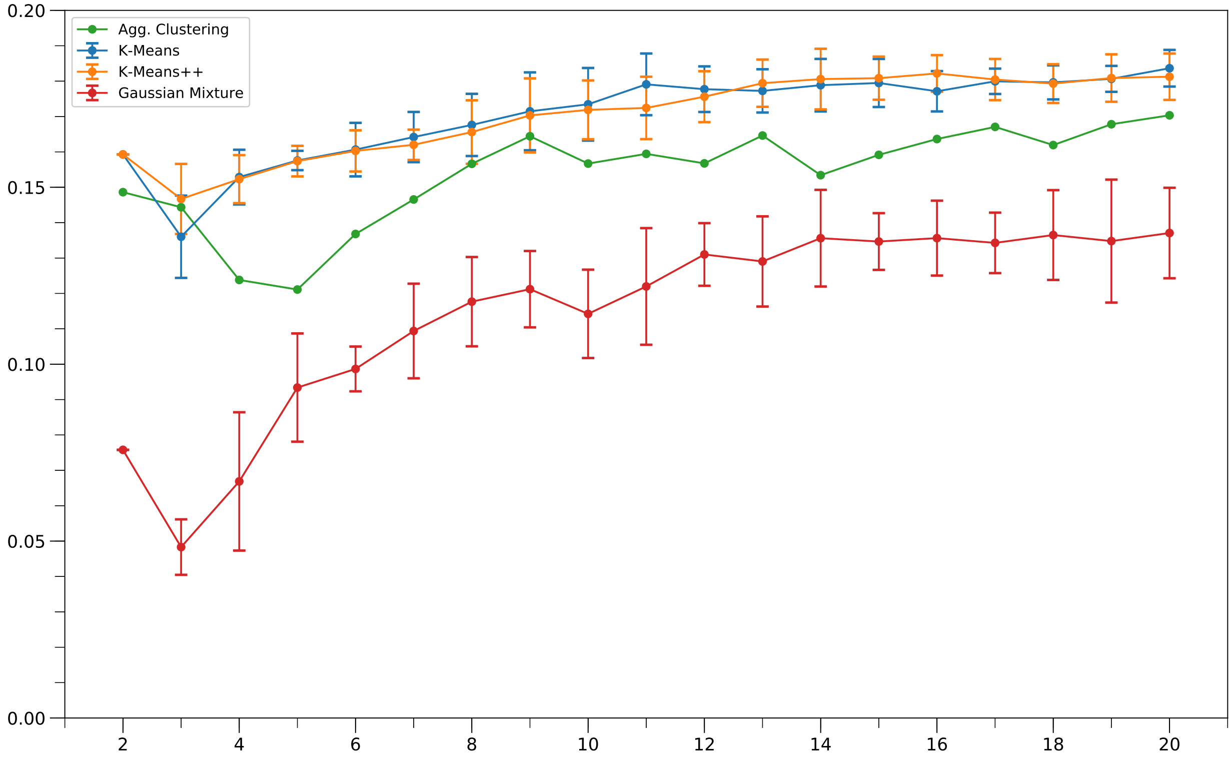

FIG. 5. Comparison of average silhouette score per cluster for the different algorithms.