William Tran

CS 373

PUID: 0029823770

NOTE: USING 1 DAYS OF EXTENSION

Homework 4

Part 2.1: Theory

1. **Weaknesses/Limitations:**
   • Terminates at local optimum (sensitive to initial seeds)
   • Applicable only when mean is defined
   • Need to specify k
   • Susceptible to outliers/noise

2. **Can you always discover structure in data using k-means algorithm? Does the algorithm always generate meaningful clusters (e.g., in uniformly distributed data, in binary data)?**
   No, sometimes k-means algorithm creates meaningless structure or is unable to create a structure. When the data is categorical, supervised learning algorithms is a more appropriate choice because k-means uses a distance measure, which does not apply to categorical data. Non-continuous data includes binary data and calculating distance measures for binary variables for k-means algorithm is meaningless.

3. **Theoretical time complexity:** $O(n^2)$
   Parallel computing is when multiple calculations/executions of processes are run simultaneously. Instead of using 1 process to match each data entry to a cluster, we can divide the data into $k > 1$ distinct groups, then match the entries in each group into a cluster simultaneously using k processes.

4. The k-means algorithm creates clusters with similar values, so the new feature will have boundaries that are appropriate to create a new discrete variable to predict the outcome. Therefore, the loss functions will be more accurate and have less errors for B than for A.

5. The goal here is to minimize within-cluster sum of squared errors and maximize between-cluster sum of squared errors so that the clusters are compact and far away from other clusters. Therefore, dividing within-cluster sum of squared errors (WC-SSE) by between-cluster sum of squared errors (BC-SSE) would give a better estimation of errors for the score function:

$$\text{Score}(C, D) = \frac{\text{WC\_SSE}}{\text{BC\_SSE}} \qquad (*)$$

This is better than only WC-SSE because the larger the between-cluster distance, the smaller the errors (we are trying to minimize the score function). Therefore, the score function reflects how good the clusters are more accurately when both WC-SSE and BC-SSE are in consideration.

For example, collection of clusters A has WC-SSE = 5 and BC-SSE = 5, while collection of clusters B has WC-SSE = 2 and BC-SSE = 2. Here, if we only look at WC-SSE, clusters in B seems to have much better scoring functions. However, using scoring function in (*) above, both clusters have the same scores, 1, which means clusters in A is just as good as clusters in B and it is a more accurate judgement.

Part 2.2: Implementation

Option 1:

a. K = 3:

```
WC-SSE= 1315554775.5261614
Centroid 1 = [38.705543, -91.556022, 14.720177, 28.64294]
Centroid 2 = [35.462838, -108.800144, 231.066336, 903.973495]
Centroid 3 = [35.663642, -105.28736, 57.785861, 191.196791]
```

b. K = 6:

```
WC-SSE= 859035528.0912684
Centroid 1 = [35.120619, -109.439936, 138.542597, 523.555394]
Centroid 2 = [35.506838, -106.10757, 64.329312, 212.593648]
Centroid 3 = [35.128428, -112.84924, 454.060531, 1835.29158]
Centroid 4 = [34.758894, -112.082319, 12.353014, 24.414274]
Centroid 5 = [36.242507, -102.177398, 32.767682, 88.907382]
Centroid 6 = [43.849767, -65.483032, 12.266061, 18.022799]
```

c. K = 9:

```
WC-SSE= 691042800.8486245
Centroid 1 = [36.310899, -101.720653, 59.452081, 131.621332]
Centroid 2 = [40.512933, -77.140662, 26.328372, 53.679209]
Centroid 3 = [35.107157, -110.893377, 201.29254, 769.647844]
Centroid 4 = [35.278002, -107.361034, 79.371714, 284.563873]
Centroid 5 = [35.406076, -109.99401, 9.202798, 13.195888]
Centroid 6 = [35.15677, -113.722217, 640.171029, 2632.356011]
Centroid 7 = [34.655959, -112.647071, 18.762424, 43.759036]
Centroid 8 = [44.636175, -61.398339, 10.622767, 14.130778]
Centroid 9 = [34.962129, -109.684466, 19.191051, 100.43287]
```

d. K = 12:

```
WC-SSE= 638421899.3013129
Centroid 1 = [35.100243, -112.040057, 277.294998, 1028.17184]
Centroid 2 = [55.362249, -2.21546, 17.814362, 21.254273]
Centroid 3 = [39.855276, -80.970821, 25.180188, 53.749474]
Centroid 4 = [34.380299, -113.074175, 10.603676, 18.836]
Centroid 5 = [35.051857, -110.1695, 133.967698, 567.541471]
Centroid 6 = [37.392525, -96.8362, 71.827895, 107.075759]
Centroid 7 = [35.280342, -107.463749, 119.980146, 370.853987]
Centroid 8 = [35.124029, -113.547353, 603.563145, 2533.354812]
Centroid 9 = [34.602607, -113.051079, 21.086208, 66.353839]
Centroid 10 = [35.324502, -106.46977, 34.834645, 151.179159]
Centroid 11 = [41.798165, -78.840019, 9.510036, 13.487966]
Centroid 12 = [35.24758, -107.39277, 68.53469, 249.470759]
```

e. K = 24:

```
WC-SSE= 551257114.6180156
Centroid 1 = [34.234372, -112.878255, 23.752927, 21.841893]
Centroid 2 = [39.127093, -83.013302, 8.119191, 10.265832]
Centroid 3 = [35.123102, -112.714144, 318.993275, 1256.424716]
Centroid 4 = [34.535621, -113.265285, 11.320863, 45.255082]
Centroid 5 = [41.922687, -82.100899, 27.933921, 20.098013]
Centroid 6 = [39.387381, -89.511799, 58.920894, 66.937668]
Centroid 7 = [38.945107, -80.761733, 15.887568, 59.789239]
Centroid 8 = [36.640637, -102.109092, 91.937463, 132.771067]
Centroid 9 = [35.310263, -106.817031, 44.864067, 188.979155]
Centroid 10 = [34.651752, -113.149714, 17.202029, 77.699247]
Centroid 11 = [44.842531, -74.486925, 7.182062, 8.433236]
Centroid 12 = [34.36953, -113.069493, 6.91629, 9.314585]
Centroid 13 = [40.367976, -79.182498, 10.092791, 28.398096]
Centroid 14 = [55.851791, -3.024435, 43.530131, 74.776208]
Centroid 15 = [40.979658, -80.24317, 27.586357, 40.745611]
Centroid 16 = [34.419811, -113.125672, 37.447428, 55.254125]
Centroid 17 = [34.472588, -113.189724, 7.249278, 25.754308]
Centroid 18 = [38.44784, -81.222496, 35.686564, 105.176301]
Centroid 19 = [34.644437, -112.395942, 26.522329, 121.331369]
Centroid 20 = [55.297084, -2.10789, 14.933732, 15.51913]
Centroid 21 = [35.168588, -107.502273, 83.230409, 281.54494]
Centroid 22 = [35.115304, -113.62347, 757.005988, 3072.884009]
Centroid 23 = [35.19688, -108.720701, 116.56784, 431.24938]
Centroid 24 = [35.027725, -110.85206, 176.696419, 691.161801]
```

Option 2:

a. K = 3:

```
WC-SSE= 224771.93831493132
Centroid 1 = [55.38289, -2.249759, 1.140508, 1.191398]
Centroid 2 = [40.715545, -79.856363, 1.17833, 1.449523]
Centroid 3 = [34.529685, -113.257446, 1.383364, 1.882434]
```

b. K = 6:

```
WC-SSE= 43880.8662268917
Centroid 1 = [54.852897, -6.080046, 1.134103, 1.190841]
Centroid 2 = [42.453666, -89.156822, 1.308855, 1.477742]
Centroid 3 = [36.121643, -115.174989, 1.442403, 1.974758]
Centroid 4 = [33.480627, -111.993835, 1.344457, 1.821605]
Centroid 5 = [37.408424, -80.498788, 1.231183, 1.600058]
Centroid 6 = [45.372796, -73.901559, 1.020774, 1.181079]
```

c. K = 9:

```
WC-SSE= 53298.43235615244
Centroid 1 = [54.945723, -5.382152, 1.135083, 1.190701]
Centroid 2 = [33.477101, -111.998622, 0.966533, 1.360197]
Centroid 3 = [36.128112, -115.167188, 1.017339, 1.488523]
Centroid 4 = [43.427774, -76.410583, 1.106245, 1.306806]
Centroid 5 = [35.18511, -80.824039, 0.892446, 1.273904]
Centroid 6 = [35.223512, -80.826309, 1.64322, 2.163592]
Centroid 7 = [33.484823, -111.988143, 1.794145, 2.370635]
Centroid 8 = [36.114002, -115.184196, 1.944316, 2.548874]
Centroid 9 = [42.453673, -89.156823, 1.308856, 1.477744]
```

d. K = 12:

```
WC-SSE= 63515.868315214895
Centroid 1 = [33.488435, -111.993676, 1.766076, 2.367742]
Centroid 2 = [33.490062, -112.008252, 1.000848, 1.663535]
Centroid 3 = [54.215113, -9.514884, 1.137205, 1.202772]
Centroid 4 = [37.239046, -80.474017, 1.247551, 1.6266]
Centroid 5 = [40.11182, -88.230842, 1.33882, 1.452782]
Centroid 6 = [36.121643, -115.174989, 1.442409, 1.974748]
Centroid 7 = [33.462727, -111.965556, 1.443104, 1.79807]
Centroid 8 = [33.496491, -112.058846, 0.663068, 1.350865]
Centroid 9 = [33.484349, -111.974596, 2.310988, 2.916453]
Centroid 10 = [43.079382, -89.404233, 1.300827, 1.484367]
Centroid 11 = [33.469025, -111.982229, 0.854007, 0.860695]
Centroid 12 = [45.073784, -74.356889, 1.008721, 1.168521]
```

e. K = 24:

```
WC-SSE= 3843.279694207217
Centroid 1 = [33.461431, -111.975721, 0.712114, 1.396522]
Centroid 2 = [36.129946, -115.162546, 0.882021, 1.081584]
Centroid 3 = [33.475341, -111.978071, 1.062974, 0.949738]
Centroid 4 = [40.135431, -88.239943, 1.328755, 1.440021]
Centroid 5 = [43.017561, -89.382062, 1.772837, 2.041254]
Centroid 6 = [49.331336, 7.860879, 1.30432, 0.82404]
Centroid 7 = [33.493314, -111.985586, 2.175547, 2.776731]
Centroid 8 = [36.113175, -115.187688, 2.633187, 3.153724]
Centroid 9 = [36.127287, -115.165836, 0.87022, 1.732697]
Centroid 10 = [45.509713, -73.599064, 1.020397, 1.181502]
Centroid 11 = [45.211004, -74.606986, 1.129661, 1.085591]
Centroid 12 = [40.755913, -80.037685, 1.215622, 1.469014]
Centroid 13 = [33.427145, -111.904069, 1.294001, 1.491623]
Centroid 14 = [55.944446, -3.187982, 1.125298, 1.225501]
Centroid 15 = [36.11092, -115.179594, 1.725807, 2.334338]
Centroid 16 = [37.176327, -80.506276, 1.478643, 1.892848]
Centroid 17 = [33.466117, -111.960483, 1.673621, 2.295364]
Centroid 18 = [35.191039, -80.826986, 1.228758, 1.67287]
Centroid 19 = [36.111679, -115.190491, 2.112616, 2.682637]
Centroid 20 = [33.503004, -112.040572, 0.963235, 1.907221]
Centroid 21 = [36.126377, -115.177145, 1.390133, 1.940666]
Centroid 22 = [33.535312, -112.081615, 1.558043, 1.838839]
Centroid 23 = [43.079599, -89.403159, 1.044575, 1.181792]
Centroid 24 = [33.497561, -112.072195, 0.636971, 0.840571]
```

Option 3:

a. K = 3:

```
WC-SSE= 94509.75815364164
Centroid 1 = [0.801539, 1.388425, 3.02894, 3.530731]
Centroid 2 = [0.804195, 1.354337, -1.391944, -1.329315]
Centroid 3 = [0.883788, 1.211319, -2.317968, -2.340439]
```

b. K = 6:

```
WC-SSE= 62222.400740409634
Centroid 1 = [0.787047, 1.389761, -2.140611, -1.717044]
Centroid 2 = [0.794566, 1.4334, 10.792755, 11.53715]
Centroid 3 = [0.889459, 1.201768, -2.352935, -2.388071]
Centroid 4 = [0.860888, 1.23945, -1.488794, -1.97058]
Centroid 5 = [0.791522, 1.403978, 1.396146, 1.940971]
Centroid 6 = [0.793981, 1.37518, -0.822062, -0.674998]
```

c. K = 9:

```
WC-SSE= 73633.34091597241
Centroid 1 = [0.798328, 1.365997, -0.833677, -0.965776]
Centroid 2 = [0.770299, 1.434638, -2.411863, -2.375484]
Centroid 3 = [1.324843, 0.301182, -2.211415, -2.387125]
Centroid 4 = [0.793931, 1.418931, 6.170486, 6.430799]
Centroid 5 = [0.960284, 1.080753, -2.379023, -2.41382]
Centroid 6 = [0.788467, 1.400159, 0.22959, 1.29596]
Centroid 7 = [0.80209, 1.36027, -2.017381, -2.215705]
Centroid 8 = [0.843027, 1.277817, -1.466119, -1.940022]
Centroid 9 = [0.7905, 1.371086, -2.180738, -1.587463]
```

d. K = 12:

```
WC-SSE= 58366.43283710291
Centroid 1 = [0.791072, 1.405009, 1.48867, 2.119763]
Centroid 2 = [0.77268, 1.435752, -2.362195, -2.282687]
Centroid 3 = [0.796054, 1.434152, 12.70391, 13.115012]
Centroid 4 = [0.785825, 1.399429, -2.063697, -2.312073]
Centroid 5 = [0.783279, 1.392732, -1.740275, -0.685392]
Centroid 6 = [0.785444, 1.393408, -2.26913, -1.844868]
Centroid 7 = [0.853273, 1.108223, -2.34395, -2.351551]
Centroid 8 = [1.328153, 0.29314, -2.207236, -2.384756]
Centroid 9 = [0.769182, 1.434145, -2.454277, -2.444225]
Centroid 10 = [0.832767, 1.295626, -1.617364, -1.972214]
Centroid 11 = [1.034277, 1.064014, -2.388922, -2.447024]
Centroid 12 = [0.813327, 1.341674, -0.419008, -1.080495]
```

e. K = 24:

```
WC-SSE= 65603.03617528682
Centroid 1 = [0.963412, 1.103371, -2.200211, -2.379149]
Centroid 2 = [0.900013, 1.109338, -1.916101, -2.063728]
Centroid 3 = [1.32716, 0.295139, -2.242927, -2.398702]
Centroid 4 = [0.8803, 1.116795, -2.437807, -2.452252]
Centroid 5 = [0.80011, 1.361735, -1.172439, -1.770732]
Centroid 6 = [0.76998, 1.434582, -2.488609, -2.449372]
Centroid 7 = [0.78587, 1.406364, 0.513978, 1.539131]
Centroid 8 = [0.773871, 1.434971, -1.907684, -1.929533]
Centroid 9 = [0.767042, 1.43281, -1.74769, -2.234609]
Centroid 10 = [0.783204, 1.388589, -1.474132, -0.964375]
Centroid 11 = [0.991148, 1.08545, -1.924522, -2.323277]
Centroid 12 = [0.775301, 1.436928, -2.439566, -2.193871]
Centroid 13 = [0.773519, 1.436118, -2.183574, -2.188374]
Centroid 14 = [0.786643, 1.389648, -2.185954, -0.981051]
Centroid 15 = [1.063078, 1.027523, -2.430122, -2.460574]
Centroid 16 = [0.763395, 1.431268, -2.156642, -2.395573]
Centroid 17 = [0.774932, 1.436307, -2.290777, -1.837955]
Centroid 18 = [0.975095, 1.024482, -1.470107, -2.043863]
Centroid 19 = [0.841436, 1.101119, -2.23398, -1.963329]
Centroid 20 = [0.772217, 1.435555, -2.386484, -2.344509]
Centroid 21 = [0.813261, 1.34377, 0.079779, -0.899295]
Centroid 22 = [0.767426, 1.433263, -2.38364, -2.464056]
Centroid 23 = [0.868728, 1.104437, -2.384408, -2.244133]
Centroid 24 = [0.794201, 1.428886, 8.247669, 8.880431]
```

Option 4:

a. K = 3:

```
WC-SSE= 3433299269.1288433
Centroid 1 = [37.523021, -97.168493, 49.414506, 165.784028]
Centroid 2 = [0, 0, 0, 0]
Centroid 3 = [0, 0, 0, 0]
```

b. K = 6:

```
WC-SSE= 3433299269.1274714
Centroid 1 = [0, 0, 0, 0]
Centroid 2 = [0, 0, 0, 0]
Centroid 3 = [0, 0, 0, 0]
Centroid 4 = [37.522996, -97.168531, 49.41452, 165.783989]
Centroid 5 = [0, 0, 0, 0]
Centroid 6 = [0, 0, 0, 0]
```

c. K = 9:

```
WC-SSE= 3433299269.1235433
Centroid 1 = [37.523018, -97.168473, 49.414168, 165.783523]
Centroid 2 = [0, 0, 0, 0]
Centroid 3 = [0, 0, 0, 0]
Centroid 4 = [0, 0, 0, 0]
Centroid 5 = [0, 0, 0, 0]
Centroid 6 = [0, 0, 0, 0]
Centroid 7 = [0, 0, 0, 0]
Centroid 8 = [0, 0, 0, 0]
Centroid 9 = [0, 0, 0, 0]
```

Option 5: (2/5 samples each for illustration)

a. K = 3:

```
WC-SSE= 16207252.589770151
Centroid 1 = [34.901213, -111.529034, 260.861072, 1125.5622]
Centroid 2 = [38.149158, -93.885202, 19.000526, 40.784063]
Centroid 3 = [35.133142, -106.883347, 86.701261, 300.237093]

WC-SSE= 17103713.246180777
Centroid 1 = [38.237285, -93.606669, 18.387666, 38.369012]
Centroid 2 = [34.916676, -110.974043, 252.385173, 1078.09307]
Centroid 3 = [35.1667, -106.492426, 80.076048, 275.295205]
```

b. K = 6:

```
WC-SSE= 15897418.727264037
Centroid 1 = [35.040518, -107.809166, 91.381781, 317.33706]
Centroid 2 = [34.568834, -112.392739, 10.681282, 21.273523]
Centroid 3 = [44.739258, -63.052983, 10.151398, 13.592347]
Centroid 4 = [40.467391, -81.493692, 38.671794, 49.624644]
Centroid 5 = [34.964787, -110.474283, 241.259667, 1022.646355]
Centroid 6 = [34.804062, -107.65762, 32.56477, 104.675355]

WC-SSE= 11160926.832041817
Centroid 1 = [34.967395, -112.430168, 294.92914, 1277.061635]
Centroid 2 = [44.256588, -64.833023, 12.475167, 17.154549]
Centroid 3 = [34.718694, -111.921517, 10.970528, 22.10736]
Centroid 4 = [35.005565, -105.780573, 53.139826, 181.726674]
Centroid 5 = [35.067641, -107.879491, 122.411566, 448.912716]
Centroid 6 = [36.171606, -102.166423, 34.542344, 75.351606]
```

c. K = 9:

```
WC-SSE= 7365554.718881389
Centroid 1 = [35.00013, -110.531598, 177.995141, 669.866274]
Centroid 2 = [53.553245, -16.798857, 89.12912, 60.813186]
Centroid 3 = [34.31537, -112.995574, 8.937896, 14.402337]
Centroid 4 = [34.96966, -107.641837, 88.365336, 306.660727]
Centroid 5 = [42.090732, -78.180233, 11.578028, 17.38273]
Centroid 6 = [54.975399, -1.585544, 16.427532, 16.750363]
Centroid 7 = [35.120688, -112.209858, 347.945868, 1605.136752]
Centroid 8 = [35.242853, -104.703826, 41.23836, 129.742004]
Centroid 9 = [35.652665, -105.89645, 24.281141, 54.860386]

WC-SSE= 8875303.10726991
Centroid 1 = [35.101665, -108.552113, 16.404486, 52.1634]
Centroid 2 = [54.987273, -1.604025, 18.801293, 18.125]
Centroid 3 = [34.944253, -109.214713, 145.583998, 540.155224]
Centroid 4 = [34.326442, -113.008063, 9.18654, 14.024616]
Centroid 5 = [42.124478, -78.205153, 10.001378, 15.724315]
Centroid 6 = [34.980048, -105.69759, 33.883911, 108.072663]
Centroid 7 = [35.047115, -106.632587, 70.965525, 234.204506]
Centroid 8 = [35.115541, -112.534843, 310.71256, 1411.864734]
Centroid 9 = [40.495182, -84.18655, 56.512009, 53.998479]
```

d. K = 12:

```
WC-SSE= 11118685.448775945
Centroid 1 = [34.245691, -112.915385, 14.535292, 35.222947]
Centroid 2 = [34.371785, -113.061828, 8.307798, 10.760524]
Centroid 3 = [34.911726, -104.893896, 41.05783, 162.272713]
Centroid 4 = [54.744198, -3.007323, 18.821078, 18.256535]
Centroid 5 = [40.587957, -80.630217, 25.360642, 47.659378]
Centroid 6 = [41.16471, -78.998305, 12.384854, 21.918942]
Centroid 7 = [35.011733, -108.640795, 118.939369, 456.335155]
Centroid 8 = [43.042515, -77.420617, 6.625149, 7.467192]
Centroid 9 = [37.782905, -96.642333, 62.371798, 74.953093]
Centroid 10 = [34.944758, -112.079099, 277.139659, 1192.761979]
Centroid 11 = [34.458958, -110.790988, 20.01689, 78.326223]
Centroid 12 = [35.506628, -105.373285, 98.531689, 253.223974]
```

```
WC-SSE= 10968605.521326972
Centroid 1 = [35.002205, -106.646402, 62.238459, 209.255969]
Centroid 2 = [42.075221, -78.2202, 10.286431, 16.307105]
Centroid 3 = [35.17087, -104.8925, 38.799743, 111.072004]
Centroid 4 = [34.955045, -112.231829, 285.715728, 1231.836504]
Centroid 5 = [34.378198, -112.973455, 7.597215, 9.123239]
Centroid 6 = [35.06534, -108.483694, 15.929173, 65.477908]
Centroid 7 = [55.31236, -7.495209, 100.039301, 74.441048]
Centroid 8 = [54.952461, -1.546752, 17.146898, 17.037717]
Centroid 9 = [0, 0, 0, 0]
Centroid 10 = [35.077467, -107.864393, 124.068465, 455.351886]
Centroid 11 = [38.748233, -91.989585, 46.97014, 49.635371]
Centroid 12 = [34.277422, -112.948539, 12.431563, 28.702002]
```

e. K = 24:

```
WC-SSE= 2737773.6160663157
Centroid 1 = [34.836988, -113.592593, 13.177717, 6.831168]
Centroid 2 = [48.559704, -49.321159, 89.091247, 65.564245]
Centroid 3 = [34.765893, -109.599521, 74.129845, 239.295889]
Centroid 4 = [35.452453, -114.304743, 408.242307, 667.784615]
Centroid 5 = [35.089469, -106.657162, 86.816461, 346.371722]
Centroid 6 = [35.10841, -103.221324, 46.437736, 157.608133]
Centroid 7 = [35.095356, -111.03078, 87.809024, 487.265724]
Centroid 8 = [34.416865, -113.112315, 9.505162, 18.194334]
Centroid 9 = [39.220699, -80.432468, 26.662162, 68.990347]
Centroid 10 = [42.82923, -77.740985, 5.723309, 6.502031]
Centroid 11 = [34.223527, -112.916257, 7.747413, 33.15862]
Centroid 12 = [40.441388, -80.793296, 20.823815, 36.505467]
Centroid 13 = [34.463084, -113.163065, 17.371877, 58.901211]
Centroid 14 = [54.842676, -2.497189, 18.757204, 18.170749]
Centroid 15 = [42.436008, -77.737642, 11.321229, 16.940169]
Centroid 16 = [47.244385, -59.626504, 6.442307, 7.75]
Centroid 17 = [35.3003, -113.832099, 366.233038, 1912.651917]
Centroid 18 = [34.543471, -108.090691, 129.747191, 678.798314]
Centroid 19 = [35.998567, -100.707128, 254.5685, 418.26645]
Centroid 20 = [34.765176, -111.837344, 258.959046, 1052.134474]
Centroid 21 = [34.15237, -112.776579, 28.441545, 26.20572]
Centroid 22 = [34.482645, -111.478732, 26.939996, 96.219004]
Centroid 23 = [36.165616, -104.570796, 62.022231, 64.867435]
Centroid 24 = [34.114098, -112.758344, 4.557209, 7.569039]
```

```
WC-SSE= 2382645.301634759
Centroid 1 = [34.557832, -109.16862, 75.783974, 363.109294]
Centroid 2 = [34.85964, -104.408704, 29.562357, 140.272623]
Centroid 3 = [35.121593, -113.944433, 406.5, 2062.173913]
Centroid 4 = [35.314874, -114.24785, 201.941634, 1290.951361]
Centroid 5 = [34.988677, -105.926211, 17.624772, 94.068306]
Centroid 6 = [40.682957, -78.501751, 7.298507, 11.484396]
Centroid 7 = [34.34525, -113.032846, 7.440281, 8.796029]
Centroid 8 = [40.965313, -79.436072, 14.666441, 23.974367]
Centroid 9 = [45.475874, -73.622632, 6.369961, 6.123574]
Centroid 10 = [35.181343, -107.282955, 46.488968, 69.672652]
Centroid 11 = [41.87822, -78.891753, 45.833542, 45.182788]
Centroid 12 = [45.570377, -60.078155, 111.884514, 55.561679]
Centroid 13 = [34.476009, -113.186117, 20.495456, 44.612801]
Centroid 14 = [36.213705, -102.075754, 146.261709, 266.880562]
Centroid 15 = [54.597755, -3.955851, 18.358019, 18.058155]
Centroid 16 = [39.707583, -81.925617, 17.540045, 49.138215]
Centroid 17 = [35.208099, -109.515149, 433.89981, 1072.011342]
Centroid 18 = [42.114508, -88.995848, 6.343939, 8.524242]
Centroid 19 = [34.443506, -113.129851, 14.124733, 65.697583]
Centroid 20 = [35.263904, -110.147474, 159.935912, 553.430068]
Centroid 21 = [34.366815, -110.993305, 166.104721, 825.181974]
Centroid 22 = [35.569179, -104.074074, 58.045186, 107.628192]
Centroid 23 = [34.817541, -107.229913, 53.094879, 200.214511]
Centroid 24 = [34.211827, -112.87282, 11.273592, 25.797392]
```

Option 6:

a. K = 3:

```
WC-SSE= 1384773692.4405353
Improved f(C, D)= 1234.2835436788935
Centroid 1 = [35.230997, -109.689195, 207.940635, 810.558636]
Centroid 2 = [38.836949, -90.995909, 13.279535, 25.051053]
Centroid 3 = [36.033998, -103.442457, 51.07781, 160.998129]
```

b. K = 6:

```
WC-SSE= 1251323532.9598818
Improved f(C, D)= 282.8933932053334
Centroid 1 = [54.541847, -7.764771, 17.684291, 21.450791]
Centroid 2 = [35.733286, -105.521663, 28.519058, 74.066066]
Centroid 3 = [35.453894, -106.428952, 70.384669, 246.053414]
Centroid 4 = [35.196109, -110.375018, 245.54621, 957.438859]
Centroid 5 = [34.497604, -112.678415, 9.983304, 17.702022]
Centroid 6 = [41.473692, -78.99474, 11.683284, 18.783635]
```

c. K = 9:

```
WC-SSE= 918525101.0652432
Improved f(C, D)= 47.66582373019229
Centroid 1 = [36.254324, -102.914703, 33.028858, 76.309536]
Centroid 2 = [54.067406, -9.698335, 20.368234, 24.226582]
Centroid 3 = [39.896798, -80.905193, 18.788711, 40.024127]
Centroid 4 = [35.249504, -107.425965, 71.837398, 260.561263]
Centroid 5 = [35.206667, -109.273152, 148.480523, 537.209418]
Centroid 6 = [35.09273, -112.408127, 384.405974, 1555.642714]
Centroid 7 = [34.412416, -113.115519, 11.631215, 23.987095]
Centroid 8 = [42.397385, -76.725045, 8.749844, 10.938089]
Centroid 9 = [35.538378, -105.673496, 42.938616, 138.101931]
```

d. K = 12:

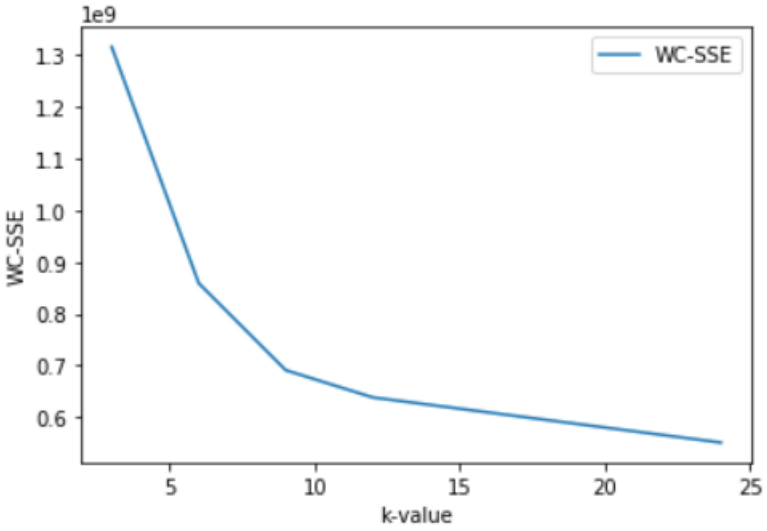
```
WC-SSE= 463266734.06119776
Improved f(C, D)= 1.3308812898929607
Centroid 1 = [35.17745, -108.462221, 110.080923, 405.004649]
Centroid 2 = [34.557629, -113.151447, 17.842955, 54.131788]
Centroid 3 = [34.355894, -112.946496, 4.652931, 5.459746]
Centroid 4 = [35.705499, -105.036903, 46.465477, 142.179333]
Centroid 5 = [34.438631, -113.148353, 8.529232, 22.959901]
Centroid 6 = [41.930131, -77.963211, 9.333552, 13.502804]
Centroid 7 = [35.564347, -114.511162, 1570.404921, 5423.494407]
Centroid 8 = [35.072219, -112.681776, 335.096449, 1410.713393]
Centroid 9 = [40.520536, -77.83, 26.107497, 54.009397]
Centroid 10 = [34.36036, -113.049509, 9.347624, 10.004126]
Centroid 11 = [53.216387, -14.543686, 15.358412, 16.719159]
Centroid 12 = [35.294899, -109.426807, 38.724473, 26.505363]
```

e. K = 24:

```
WC-SSE= 259337466.7522631
Improved f(C, D)= 0.13812997979512226
Centroid 1 = [55.097845, -1.775282, 13.093533, 10.02452]
Centroid 2 = [35.036056, -108.5308, 84.967574, 350.979806]
Centroid 3 = [39.213667, -80.25979, 19.050544, 65.587294]
Centroid 4 = [35.698398, -104.271827, 74.1358, 163.484114]
Centroid 5 = [41.624712, -79.869718, 13.766338, 17.413175]
Centroid 6 = [34.380862, -113.084537, 6.025817, 9.371856]
Centroid 7 = [42.276074, -78.110958, 6.464637, 7.597526]
Centroid 8 = [35.170598, -107.762432, 29.915144, 96.208521]
Centroid 9 = [55.871301, -3.065127, 21.971028, 34.467757]
Centroid 10 = [38.390334, -95.637175, 140.561244, 162.40622]
Centroid 11 = [35.8357, -114.855728, 2454.992424, 8628.136363]
Centroid 12 = [35.161846, -109.410503, 144.416666, 537.848351]
Centroid 13 = [35.059898, -111.802686, 252.397907, 956.058845]
Centroid 14 = [35.229919, -107.224543, 28.177147, 163.600178]
Centroid 15 = [40.653458, -79.858195, 17.696984, 31.065851]
Centroid 16 = [55.830938, -2.988873, 49.291614, 86.210262]
Centroid 17 = [34.335387, -113.021004, 32.028902, 47.54928]
Centroid 18 = [34.43884, -113.147629, 8.831476, 27.272146]
Centroid 19 = [35.132643, -113.641621, 503.207272, 2160.185454]
Centroid 20 = [35.140374, -107.874396, 73.331103, 245.291416]
Centroid 21 = [34.682398, -113.439699, 10.760196, 53.47051]
Centroid 22 = [34.23812, -112.890254, 22.316824, 15.128911]
Centroid 23 = [38.79358, -81.559215, 11.921036, 43.399284]
Centroid 24 = [40.181118, -85.934721, 54.362186, 61.68936]
```

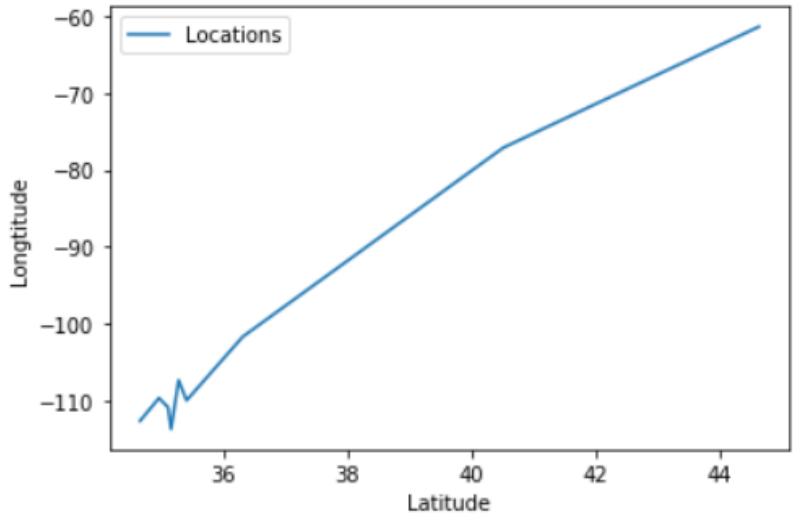
Part 3: Analysis

1. Plot of WC-SSE vs. K:

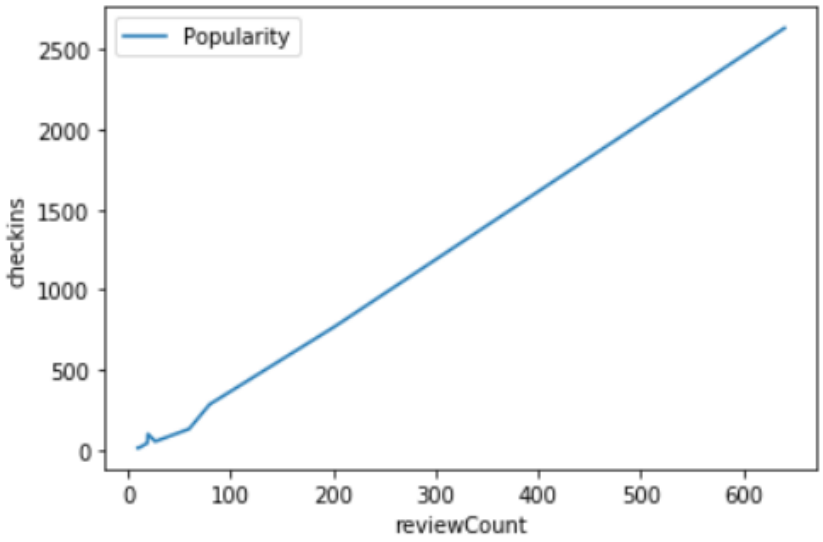


According to this plot, k = 9 would be a good balance between score function and the number of clusters. Both errors and resources would be optimized to form k = 9 clusters.

Plot of latitude vs longitude:



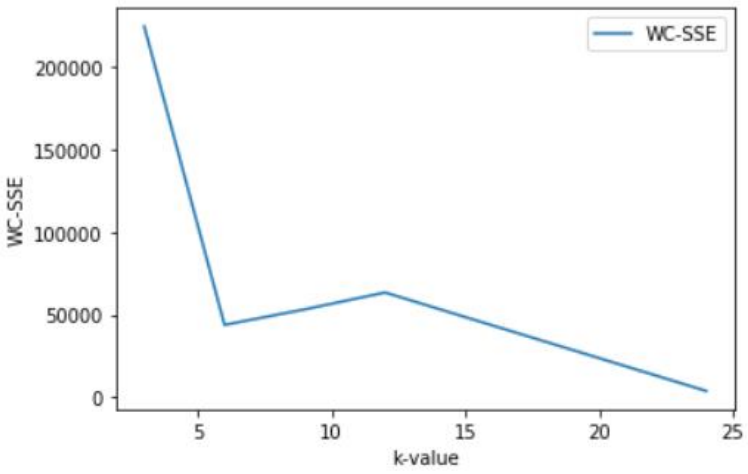
Plot of reviewCount vs. checkins:



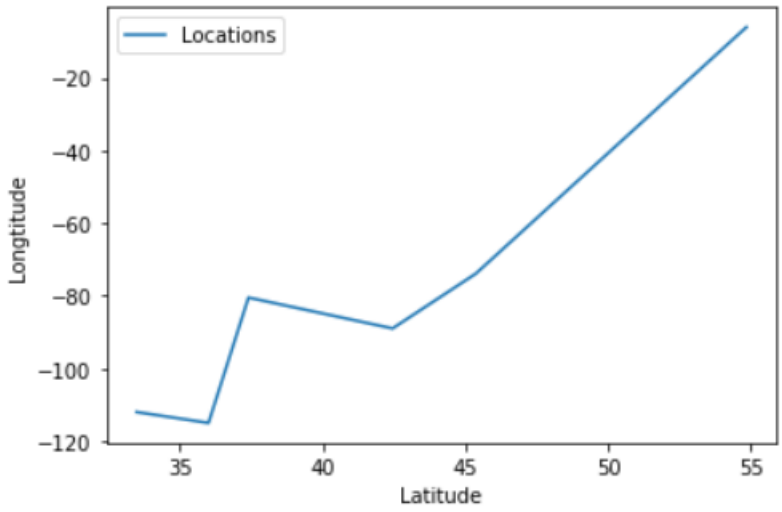
In both graph above, the slope is positive and y increases as x increases. In both, some clusters are very close to each other on the lower end while the rest are far away.

2. Expectation: reviewCount and checkins would have much smaller standard deviations, and the WC-SSE would be much smaller than in (1).

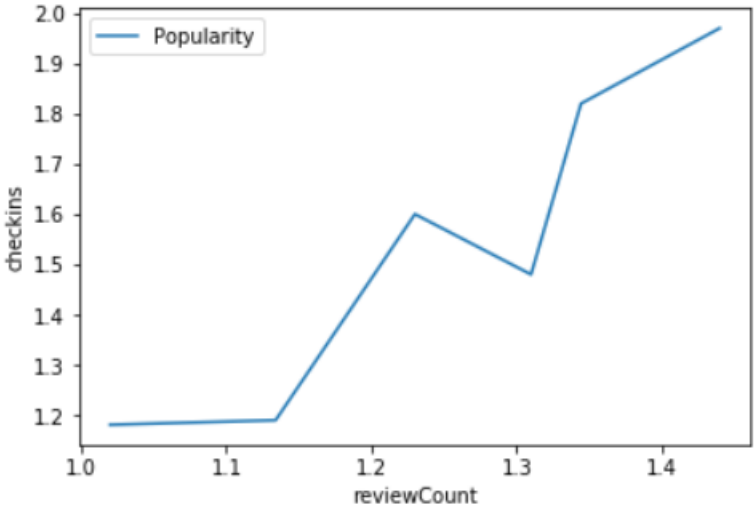
   Plot of WC-SSE vs. K:



According to this plot, k = 6 would be a good balance between score function and the number of clusters. Both errors and resources would be optimized to form k = 6 clusters.
Plot of latitude vs longitude:



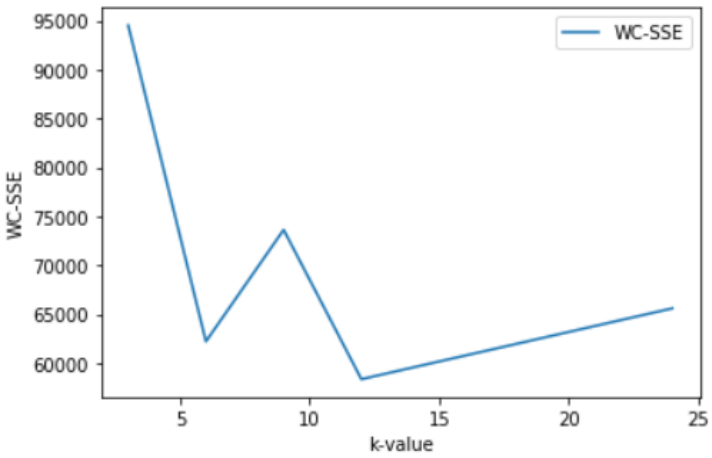
Plot of reviewCount vs. checkins:



Different from the centroids in (1), the centroids here in part 2 are well-distributed along the line, indicating that the clusters are more separated, and BC-SSE is larger. As expected, the WC-SSE is much smaller than that in (1), indicating a better set of clusters.

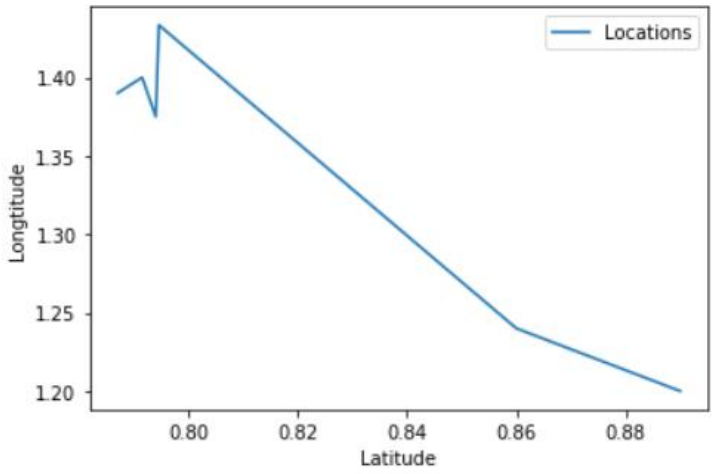
3. Expectation: Since all 4 attributes are standardized here, these clusters are expected to have even less errors than (2).

   Plot of WC-SSE vs. K:

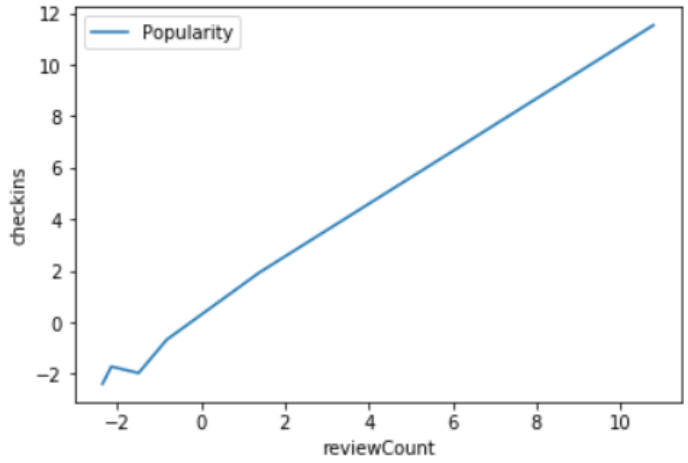


According to this plot, k = 6 would be a good balance between score function and the number of clusters. Both errors and resources would be optimized to form k = 6 clusters.

Plot of latitude vs longitude:

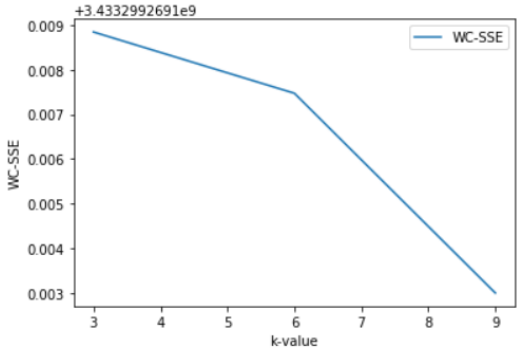


Plot of reviewCount vs. checkins:



On the contrary to what I expected, the centroids here are very close together. This makes sense because they are all normalized, making the cluster not as well-separated as in (2).

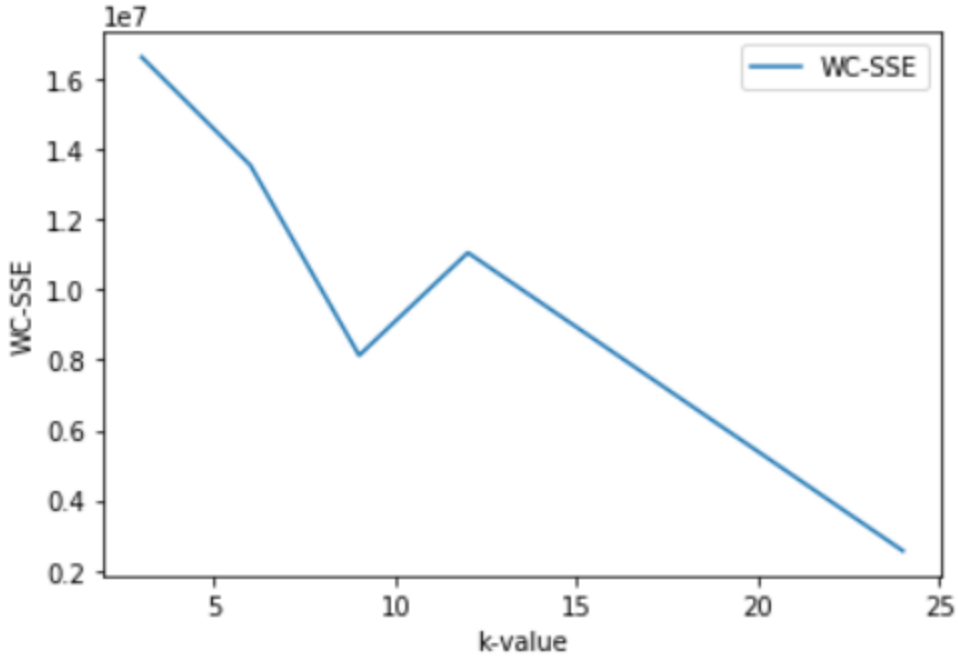
4. Expectation: Less accurate distance, so WC-SSE is larger.
   Plot of WC-SSE vs. K:



   According to this plot, k = 6 would be a good balance between score function and the number of clusters. Both errors and resources would be optimized to form k = 6 clusters.
   Plot of latitude vs longitude:

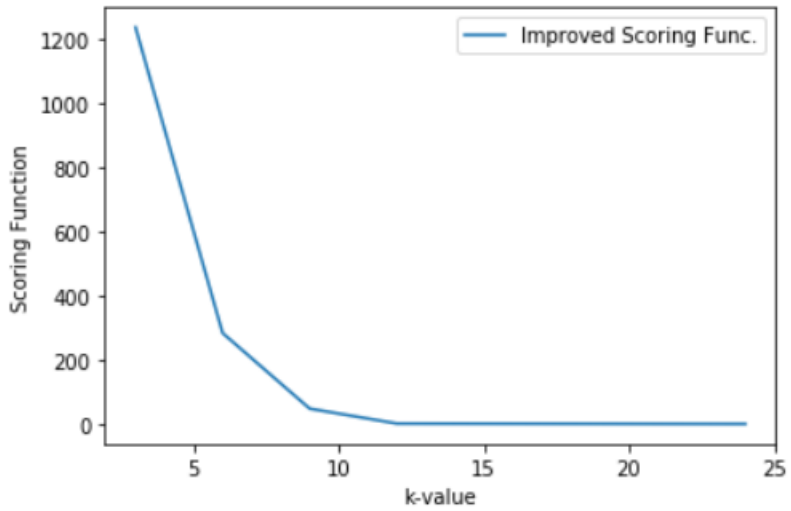
5. Expectation: High variance because the sample size is smaller. Therefore, different trials of the same k-value will yield values with big difference. WC-SSE will also be lower due to having less data.

   Average of WC-SSE:
   K = 3: 16607252.59
   K = 6: 13529172.78
   K = 9: 8120428.91
   K = 12: 11043645.4851
   K = 24: 2560209.45882
   Plot of WC-SSE vs. K:



Just like in part (3.1), as k increases, the sum of squared errors decreases. However, the graph is not as clear and stable as the one in 3.1 because the dataset here is smaller and has more variability, so it is more sensitive to noise.

6. Plot of f(C, D) and WC-SSE vs. K:



This plot is very similar to the plot of WC-SSE vs K in 3.1. That is because both graphs represent the scoring function of the same dataset. However,

f(C, D) in (6) will produce more consistently accurate result because it also takes in the between-cluster SSE.