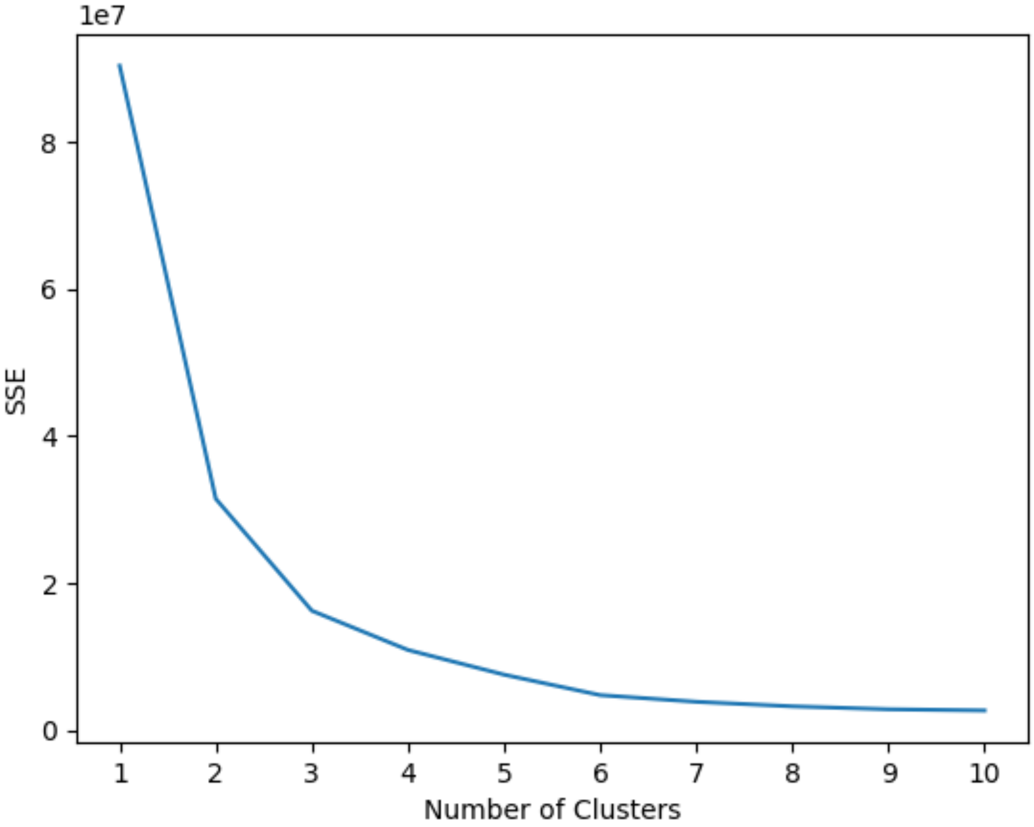
HW3 Report
Gina Roh

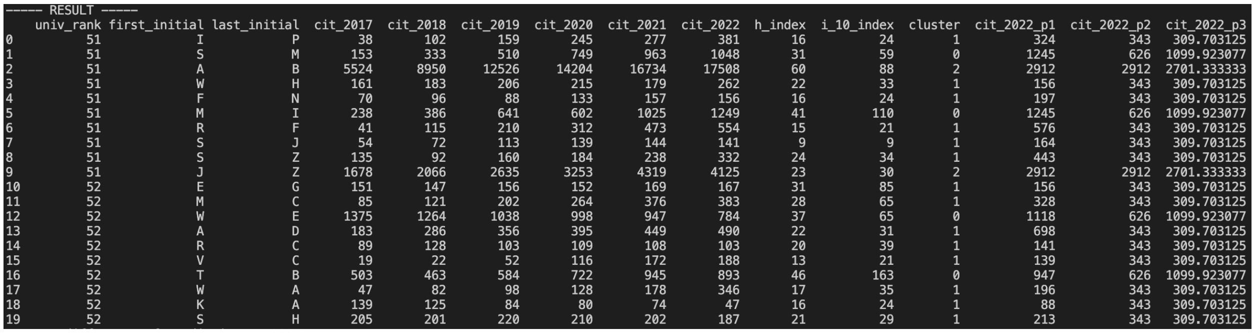
1. What is the right number of clusters for this problem? Why?

I used elbow method. Based on the graph below, after k=3, there is not many changes in SSE. So, I decided k = 3.



2. For each of the test data, find the nearest cluster centroid and place the test data into that cluster.

Please see the ‘cluster’ column in the following chart.



```
----- RESULT -----
    univ_rank first_initial last_initial  cit_2017  cit_2018  cit_2019  cit_2020  cit_2021  cit_2022  h_index  i_10_index  cluster  cit_2022_p1  cit_2022_p2  cit_2022_p3
0          51             I            P        38       102       159       245       277       381       16          24        1          324          343   309.703125
1          51             S            M       153       333       510       749       963      1048       31          59        0         1245          626  1099.923077
2          51             A            B      5524      8950     12526     14204     16734     17508       60          88        2         2912         2912  2701.333333
3          51             W            H       161       183       206       215       179       262       22          33        1          156          343   309.703125
4          51             F            N        70        96        88       133       157       156       16          24        1          197          343   309.703125
5          51             M            I       238       386       641       602      1025      1249       41         110        0         1245          626  1099.923077
6          51             R            F        41       115       210       312       473       554       15          21        1          576          343   309.703125
7          51             S            J        54        72       113       139       144       141        9           9        1          164          343   309.703125
8          51             S            Z       135        92       160       184       238       332       24          34        1          443          343   309.703125
9          51             J            Z      1678      2066      2635      3253      4319      4125       23          30        2         2912         2912  2701.333333
10         52             E            G       151       147       156       152       169       167       31          85        1          156          343   309.703125
11         52             M            C        85       121       202       264       376       383       28          65        1          328          343   309.703125
12         52             W            E      1375      1264      1038       998       947       784       37          65        0         1118          626  1099.923077
13         52             A            D       183       286       356       395       449       490       22          31        1          698          343   309.703125
14         52             R            C        89       128       103       109       108       103       20          39        1          141          343   309.703125
15         52             V            C        19        22        52       116       172       188       13          21        1          139          343   309.703125
16         52             T            B       503       463       584       722       945       893       46         163        0          947          626  1099.923077
17         52             W            A        47        82        98       128       178       346       17          35        1          196          343   309.703125
18         52             K            A       139       125        84        80        74        47       16          24        1           88          343   309.703125
19         52             S            H       205       201       220       210       202       187       21          29        1          213          343   309.703125
```

3. Tabulate the following predictions for the 2022 citation numbers for the test set, using the average difference magnitude to evaluate them:

(1) same as the 2022 citation number of the nearest neighbor from the training set;

Refer to the ‘cit_2022_p1’ column.

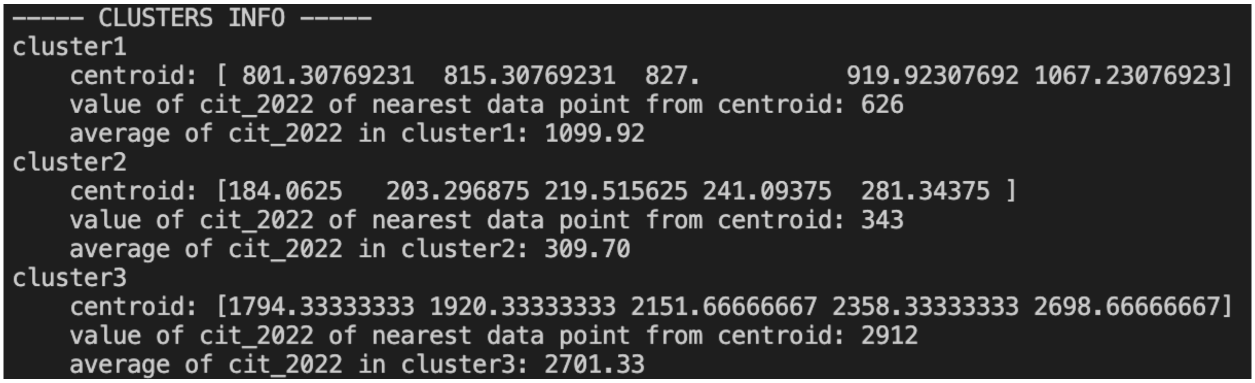
(2) same as the point nearest the cluster centroid;

Refer to the ‘cit_2022_p2’ column.

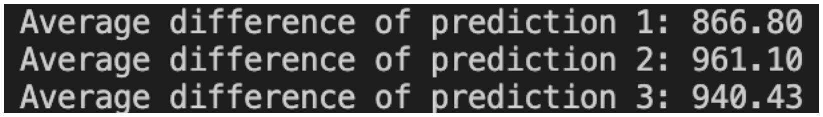
(3) average of all others from the training set in the same cluster.

Refer to the ‘cit_2022_p3’ column.

4. Draw conclusions from the comparison.



```
----- CLUSTERS INFO -----
cluster1
    centroid: [ 801.30769231  815.30769231  827.          919.92307692 1067.23076923]
    value of cit_2022 of nearest data point from centroid: 626
    average of cit_2022 in cluster1: 1099.92
cluster2
    centroid: [184.0625   203.296875 219.515625 241.09375  281.34375 ]
    value of cit_2022 of nearest data point from centroid: 343
    average of cit_2022 in cluster2: 309.70
cluster3
    centroid: [1794.33333333 1920.33333333 2151.66666667 2358.33333333 2698.66666667]
    value of cit_2022 of nearest data point from centroid: 2912
    average of cit_2022 in cluster3: 2701.33
```



```
Average difference of prediction 1: 866.80
Average difference of prediction 2: 961.10
Average difference of prediction 3: 940.43
```

The prediction method (1) using values same as the 2022 citation number of the nearest neighbor from the training set have the lowest difference with the actual value of cit_2022 in test datasets.