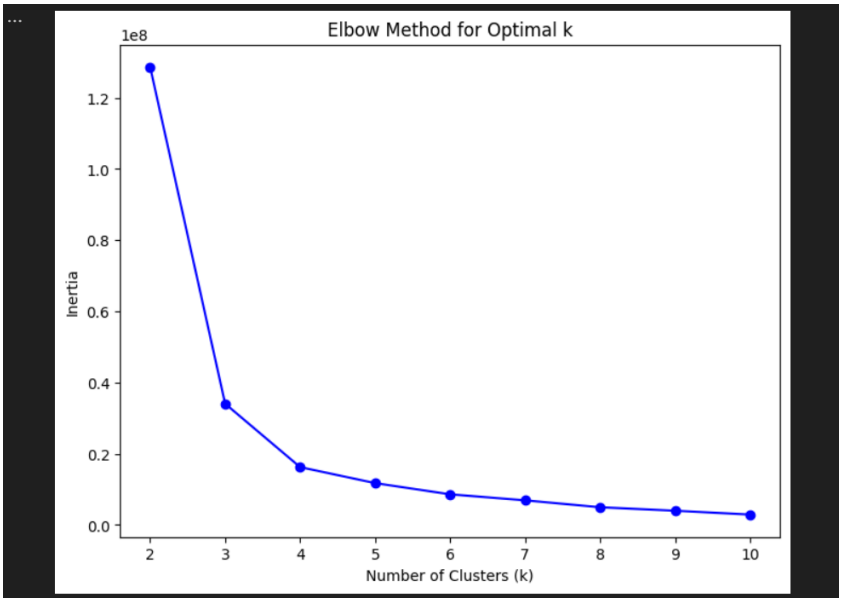
HW3

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

The right number of clusters for this problem, based on the Elbow Method, seems to be 4. In this case, we've calculated the inertia for k values in the range from 2 to 10 and plotted the results. The "elbow" point on the plot is the point where the inertia starts to decrease at a slower rate. This is often considered a good choice for the number of clusters because it represents a trade-off between maximizing data points within clusters while minimizing the number of clusters. As we can see in the graph below, the graph after point 4 is almost at the same level. So, considering 4 as elbow point and the number of clusters in this case.



The Tabulated predictions for the 2022 citations are as follows:

	Cit_2022	Pred_NN	Pred_NC	Pred_AC
83	406	374	827.882353	915.000000
53	679	687	827.882353	915.000000
70	406	370	249.743590	284.714286
45	399	255	249.743590	284.714286
44	488	370	249.743590	284.714286
39	233	252	249.743590	284.714286
22	737	356	249.743590	284.714286
80	424	411	249.743590	284.714286
10	310	255	249.743590	284.714286
0	473	356	249.743590	284.714286
18	61	62	249.743590	284.714286
30	102	62	249.743590	284.714286
73	237	276	249.743590	284.714286
33	232	190	249.743590	284.714286
90	605	668	827.882353	915.000000
4	50	37	249.743590	284.714286
76	58	62	249.743590	284.714286
77	184	140	249.743590	284.714286
12	663	789	827.882353	915.000000
31	760	789	827.882353	915.000000

HW3

```
Average Difference Magnitude for Nearest Neighbour method: 66.2
Average Difference Magnitude for Nearest Centroid method: 167.8077677224736
Average Difference Magnitude for Average Cluster method: 191.3357142857143
```

1.Same as the 2022 citation number of the nearest neighbor from the training set(Pred_NN):

For each data point in the test set, this strategy selected the nearest neighbor from the training set based on the features used for clustering (in this case, citation numbers for previous years). The predicted value for Cit_2022 is set to be the same as the Cit_2022 value of the nearest neighbor from the training set.
For example, we will consider our 1st row in data. If a test data point has the Cit_2022 value of 406 and its nearest neighbor from the training set has a Cit_2022 value of 374, the prediction for that test data point using this strategy is 374. This strategy essentially assigns each test data point the Cit_2022 value of its nearest neighbor from training set.

2. Same as the point nearest the cluster centroid (Pred_NC):

For each data point in the test set, this strategy assigns the same Cit_2022 value as the cluster centroid that the data point is assigned to. In other words, it takes the Cit_2022 value of the centroid of the cluster to which the data point belongs.
For example, if a test data point belongs to Cluster 1, and the centroid of Cluster 1 has a Cit_2022 value of 827.882353, then the prediction for that test data point using this strategy is 827.882353.

3. Average of all others from the training set in the same cluster (Pred_AC):

For each data point in the test set, this strategy calculates the average Cit_2022 value of all the data points in the same cluster as the test data point (excluding the test data point itself).
For example, if a test data point belongs to Cluster 2, this strategy computes the average Cit_2022 value of all the other training data points in Cluster 2 and uses that average as the prediction for the Cit_2022 value of the test data point.

In terms of average difference magnitude, the "Nearest Neighbor" method performs the best, as it has the lowest average difference magnitude. The "Nearest Centroid" method has a higher average difference magnitude, indicating less accurate predictions, and the "Average Cluster" method has the highest average difference magnitude among the three, suggesting that it's the least accurate. Therefore, based on the provided average difference magnitudes, the "Nearest Neighbor" method is the better choice for predicting Cit_2022 values in this specific context.

References

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html