HW1 Report

Patrick Gardocki

2023-09-05

1 Concept Questions

1. What's the main difference between supervised and unsupervised learning? Give one benefit and drawback for supervised and unsupervised learning, respectively.

The main difference between supervised and unsupervised learning is that supervised learning requires a labeled dataset. The goal is to map inputs to outputs and each datapoint has a target. Supervised learning is capable of producing accurate prediciton models given labeled dataset, but its drawback is the need for labeled data. Labeled data is not always accessible. On the other hand, unsupervised learnign does not require labeled data. These algoritms try to find patterns and groupings within the given data. The lack of labeled data allows for the ability to find hidden patterns in data but the results are then difficult to interpert and connect to real world applications.

2. Will different initializations for k-means lead to different results?

Yes, k-means minimizes the objective function which can contain local minima. Also the initial centrodis are randomly determined which can lead to the obejctive function converging to different local minima.

3. Give a short proof (can be in words but using correct logic) of why k-means algorithm will converge in a finite number of iterations.

The k-means algorithm will converge in a finite number of iterations because the algorithm minimizes the WCSS and will only iterate of it deceases. Given a finite set of data points, there is a finite number of ways to assign the data to clusters. Each iteration, the WCSS will decrease or not change. If the WCSS can not increase, and there are finite cluster groups, the algorithm will converge in a finite number of iterations.

4. What is the main difference between k-means and generalized k-means algorithm? Explain how the choice of similarity/dissimilarity/distance will impact the result.

K-Means uses the Euclidean distance to determine the similarity and distance of the data and centers. The Euclidean distance is a good measure when the clusters are spherical and generally equally sized. Generalized K-Means allows for various distance metrics to be used. The choice of distance metric will impact the results. Based on the characteristics of the data, other distance metrics will be better. The Manhattan distance is another metric that is useful when the data has different scales.

5. Write down the graph Laplacian matrix and find the eigenvectors associated with the zero eigenval- ues. Explain how you find out the number of disconnected clusters in the graph and identify these disconnected clusters using these eigenvectors.

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$D = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$L = D - A = \begin{bmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix} \text{ So that } L * v = 0 \; v_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \; v_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

2 Math of K-Means Clustering

1.

Given: $J = \sum_i \sum_j r_{ij} \cdot \|x_i - \mu_j\|^2$

$\frac{\partial J}{\partial \mu_j} = \frac{\partial}{\partial \mu_j} \left(\sum_i \sum_j r_{ij} \cdot \|x_i - \mu_j\|^2 \right) = 2 \sum_i r_{ij} \cdot (x_i - \mu_j)$

$0 = 2 \sum_i r_{ij} \cdot (x_i - \mu_j)$

Solve for μ_j:

$\mu_j = \frac{1}{\sum_i r_{ij}} \sum_i r_{ij} \cdot x_i$

2.

Given: $J = \sum_i \sum_j r_{ij} \cdot \|x_i - \mu_j\|^2$

$\frac{\partial J}{\partial r_{ij}} = \frac{\partial}{\partial r_{ij}} \left(\sum_i \sum_j r_{ij} \cdot \|x_i - \mu_j\|^2 \right) = 2\|x_i - \mu_j\|^2$

$0 = 2\|x_i - \mu_j\|^2$

Solve for r_{ij}:

$\|x_i - \mu_j\|^2 = 0$

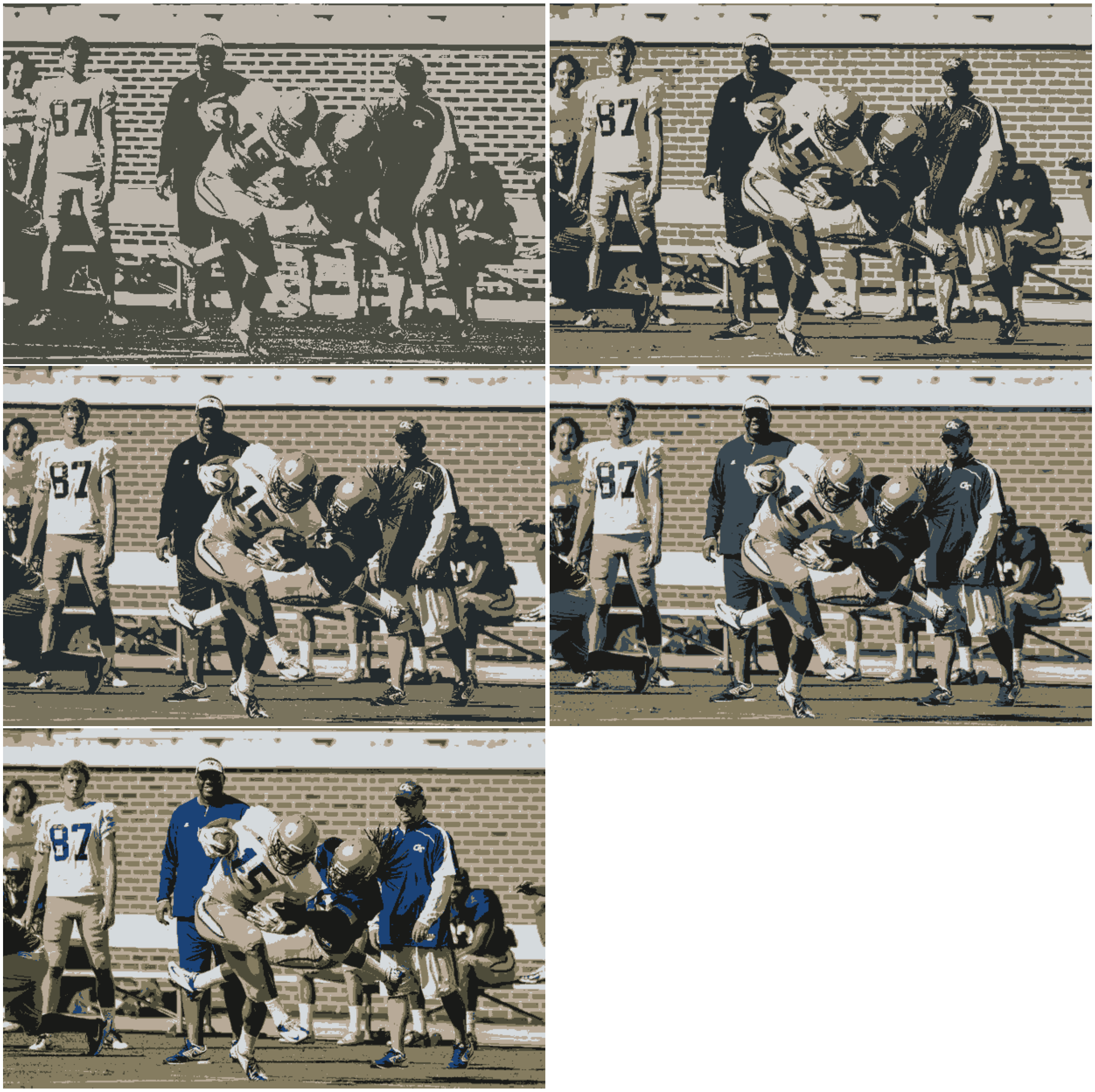
$r_{ij} = \begin{cases} 1, & \text{if } j = \arg\min_k \|x_i - \mu_k\|^2 \\ 0, & \text{otherwise} \end{cases}$

When the centroids μ_j are fixed, r_{ij} should be set to 1 for the cluster j that minimizes the squared Euclidean distance and set to 0 for all other clusters.

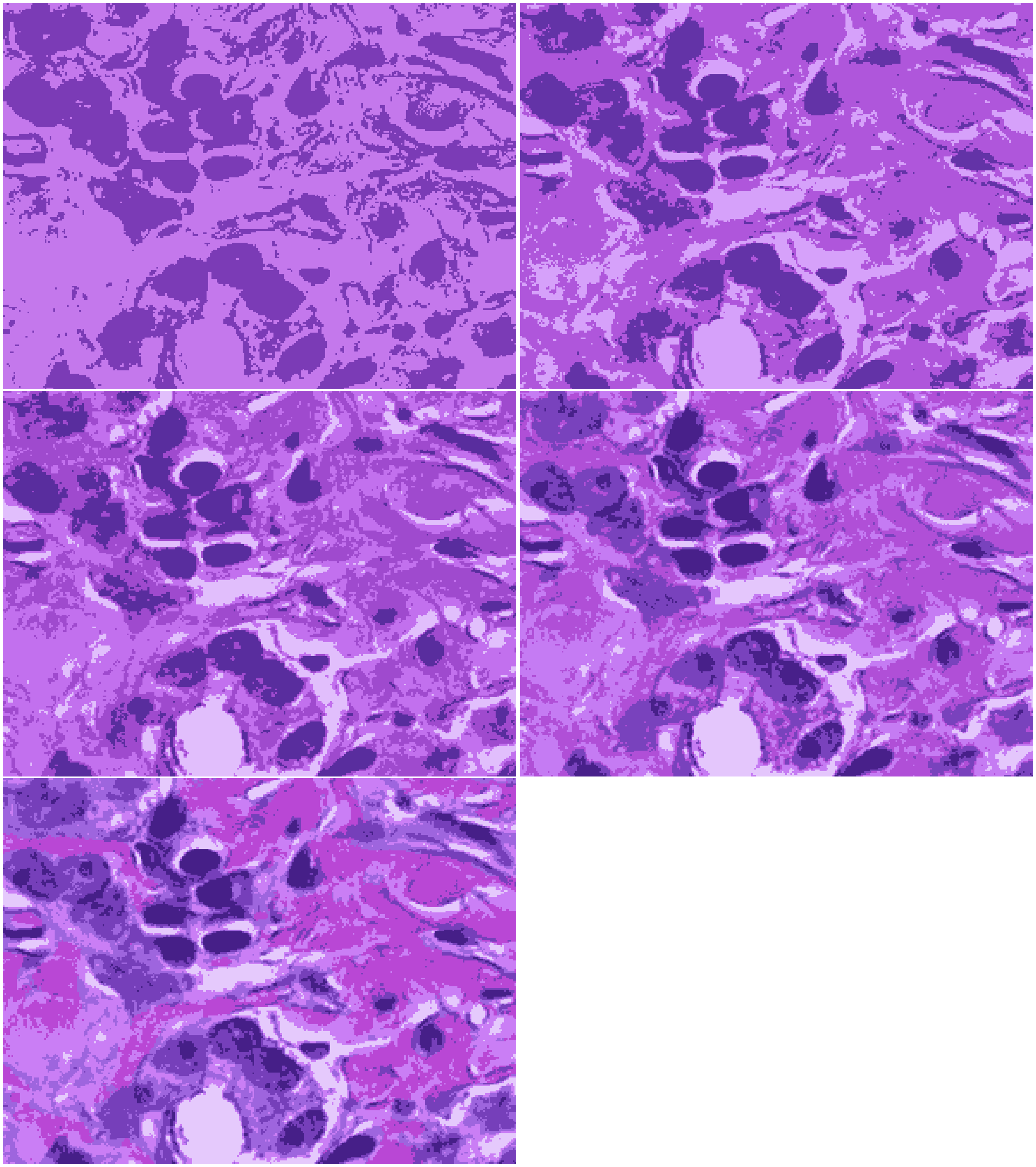
3 Image Compression using Clustering

1.

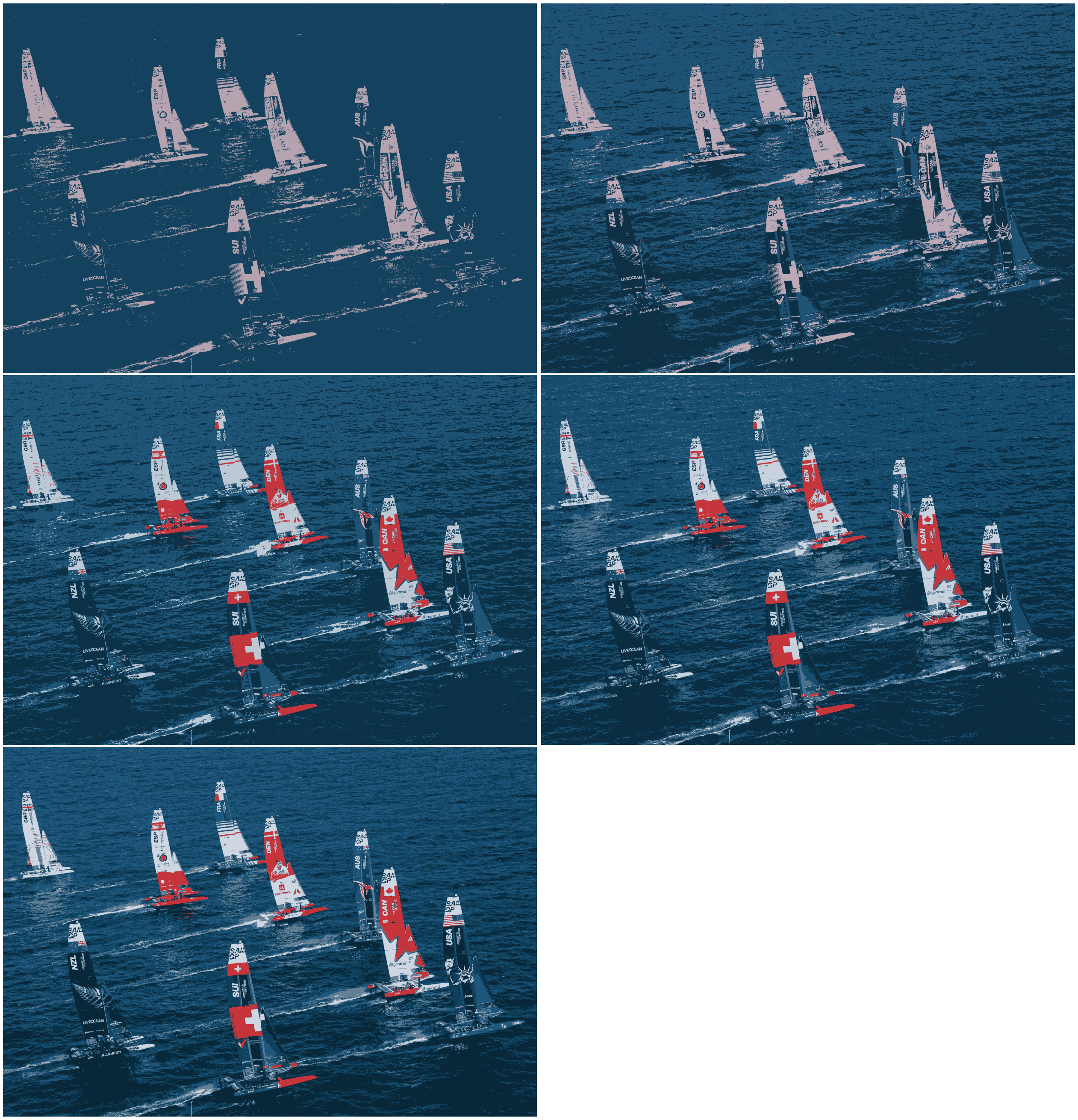
Football Image compression for k={2,6}



Hestain Image compression for k={2,6}



Sailing Image compression for k={2,6}



2.

Table 1: Football Image Results

Clusters	Iterations	RunTime
2	18	0.94
3	15	1.11
4	26	2.63
5	31	2.87
6	32	3.52

Table 2: Hestain Image Results

Clusters	Iterations	RunTime
2	7	0.104
3	24	0.412
4	32	0.637
5	61	1.480
6	57	1.670

Table 3: Sailing Image Results

Clusters	Iterations	RunTime
2	20	15.77
3	20	18.77
4	37	50.77
5	67	94.63
6	48	88.72

3.

One method of finding the best k for k-means is utilizing the elbow plot. The elbow plot has varying k values on the x-axis and the Within-Cluster Sum of Squares (WCSS) value for each k on the y-axis. The WCSS calculates the variance within each cluster. The aim of the plot is to see when the addition of additional clusters has diminishing improvements in the WCSS. Usually, an elbow, is observed, and that is where the rate of decrease in WCSS diminishes. The elbow plot method is a subjective method and is a helpful heuristic. Depending on the dataset, there can be no distinct 'elbow' in the plot.

The example below is for the Football image. With additional clusters, the WCSS decreases, but after about 5 clusters, the rate of decrease is much less apparent than with the addition of the first 4 clusters. Based on this plot, the optimal k value would be 5.

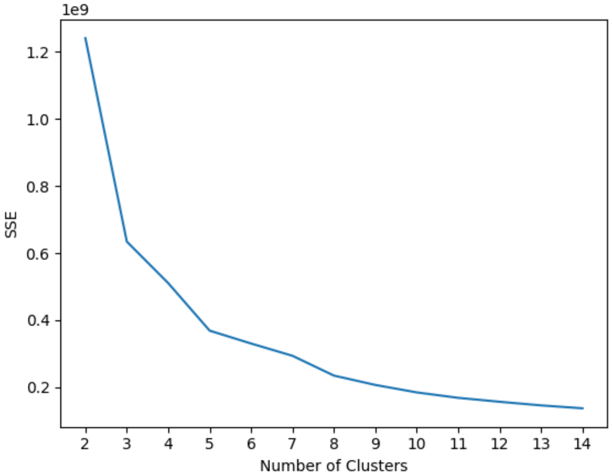


Figure 1: Elbow Plot of Football image Compression for k={2,14}

4 MNIST Dataset Clustering

Table 4: Purity Scores for Euclidean Distance Metric

Cluster	Purity
1	0.789
2	0.550
3	0.703
4	0.634
5	0.374
6	0.337
7	0.775
8	0.450
9	0.530
10	0.415

It seems that the best purity scores came from using the Manhattan distance. The data processed was image data represented in pixel form. Manhattan distance restricts the metric to vertical and horizontal directions. This may have aided in getting better results in its image analysis.

Table 5: Purity Scores fot Manhattan Distance Metric

Cluster	Purity
1	0.418
2	0.997
3	0.441
4	0.432
5	0.436
6	0.352
7	0.601
8	0.683
9	0.394
10	0.518

5 Political Blog Dataset

1.

Table 6: Purity Scores for Euclidean Distance Metric

Cluster	MajorityLabel	MismatchRate
1	1	0.481
2	1	0.000

Table 7: Purity Scores fot Manhattan Distance Metric

Cluster	MajorityLabel	MismatchRate
1	1	0.067
2	0	0.000
3	0	0.021
4	1	0.000
5	1	0.000

Table 8: Purity Scores fot Manhattan Distance Metric

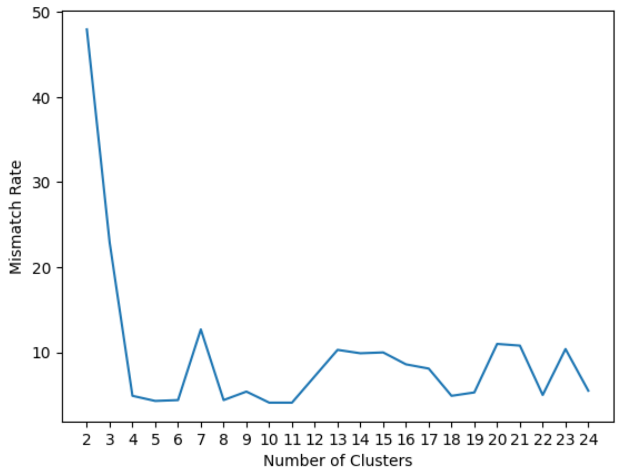
Cluster	MajorityLabel	MismatchRate
1	0	0.000
2	0	0.022
3	1	0.000
4	0	0.000
5	0	0.000
6	0	0.000
7	1	0.000
8	1	0.000
9	1	0.070
10	0	0.000

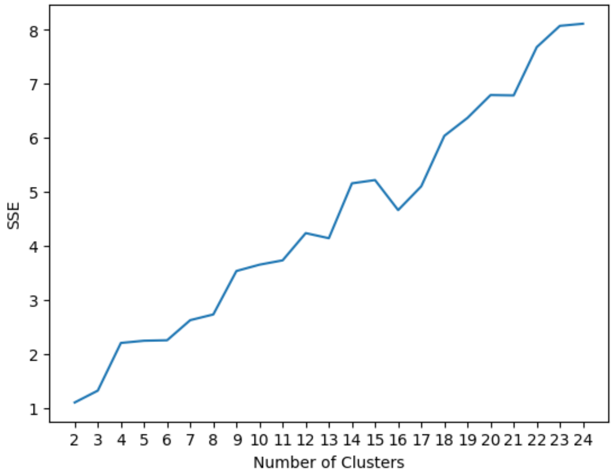
Table 9: Purity Scores fot Manhattan Distance Metric

Cluster	MajorityLabel	MismatchRate
1	1	0.028
2	0	0.000
3	0	0.477
4	0	0.000
5	1	0.000
6	0	0.000
7	0	0.000
8	1	0.000
9	1	0.000
10	0	0.000
11	0	0.000
12	0	0.250
13	0	0.000
14	0	0.166
15	0	0.000
16	0	0.000
17	0	0.000
18	0	0.000
19	1	0.000
20	1	0.034
21	0	0.000
22	0	0.066
23	0	0.000
24	0	0.409
25	0	0.111

2.

Looking at overall mismatch rate, it seems that additional clusters past 4 clusters do not help minimize it. The rate hovers between 4 and 10 %. The mismatch rate was high for 2 and 3 clusters but dropped signifigantly at 4 clusters. Further, utilizing an elbow plot, it is observed that the WCSS increases consistently past 2 clusters. This implies that there are two to four distinct political communities in the dataset.


Mismatch Rate
Number of Clusters


SSE
Number of Clusters