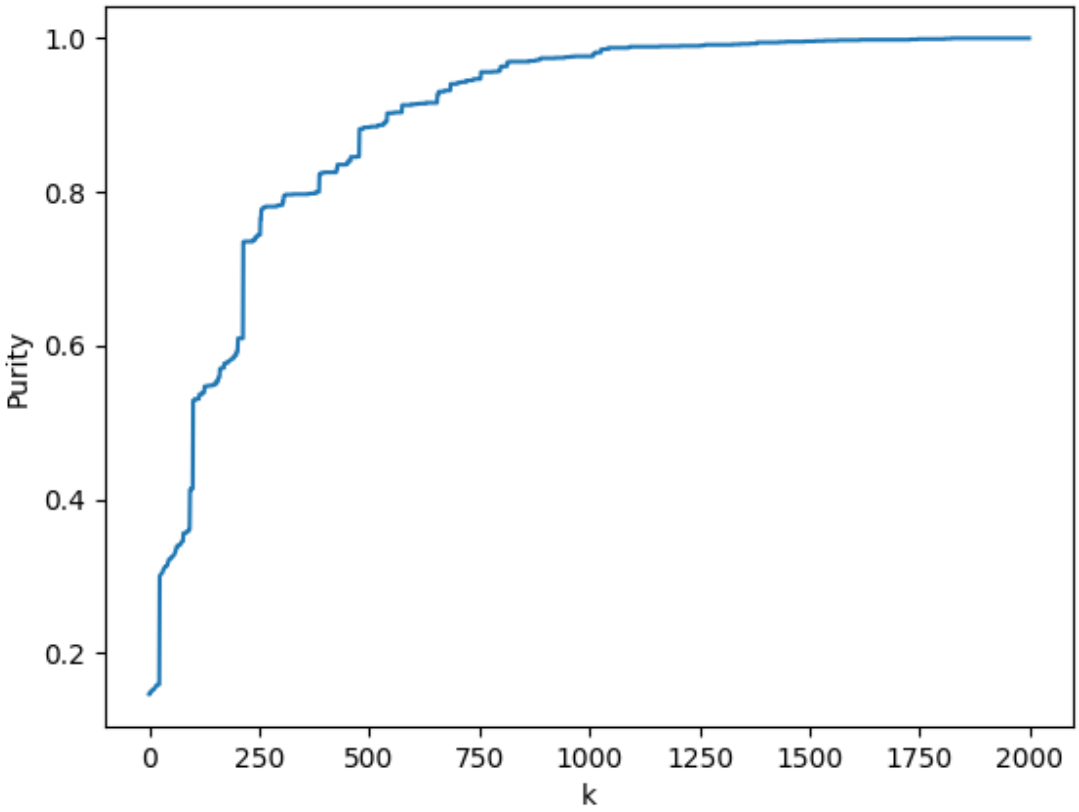
1).



2).

Sample K values and their corresponding Purities

k	Purity
10	0.1525
20	0.1585
30	0.304
40	0.3135
50	0.323

* Note - Taking into account the running time of the code for a data of 2000 instances, only purities for sample k values are in this table.

Prevalent Labels in the first 25 clusters

Most prevalent label for cluster 1 is: b'path'

Most prevalent label for cluster 2 is: b'foliage'

Most prevalent label for cluster 3 is: b'sky'

Most prevalent label for cluster 4 is: b'path'

Most prevalent label for cluster 5 is: b'grass'

Most prevalent label for cluster 6 is: b'foliage'

Most prevalent label for cluster 7 is: b'brickface'

Most prevalent label for cluster 8 is: b'brickface'

Most prevalent label for cluster 9 is: b'cement'

Most prevalent label for cluster 10 is: b'cement'

Most prevalent label for cluster 11 is: b'sky'

Most prevalent label for cluster 12 is: b'foliage'

Most prevalent label for cluster 13 is: b'cement'

Most prevalent label for cluster 14 is: b'grass'

Most prevalent label for cluster 15 is: b'grass'

Most prevalent label for cluster 16 is: b'window'

Most prevalent label for cluster 17 is: b'window'

Most prevalent label for cluster 18 is: b'window'

Most prevalent label for cluster 19 is: b'sky'

Most prevalent label for cluster 20 is: b'cement'

Most prevalent label for cluster 21 is: b'cement'

Most prevalent label for cluster 22 is: b'cement'

Most prevalent label for cluster 23 is: b'cement'

Most prevalent label for cluster 24 is: b'cement'

Most prevalent label for cluster 25 is: b'sky'

Purity is: 0.7785

Formula for calculating purity:

1/2000*(217+236+109+6+266+2+37+58+2+4+48+3+6+1+2+5+8+23+1+2+1+10+4+1+16+1+2+1+1+10+2+8+1+7+2+2+1+13+1+2+1+4+4+1+4+3+3+3+3+5+5+2+1+14+8+11+1+20+3+1+3+3+1+3+1+2+1+2+5+1+2+12+2+1+1+5+1+6+1+1+1+1+5+1+8+14+1+2+6+1+1+1+1+1+1+6+1+1+3+1+1+1+1+4+2+6+1+2+2+1+1+3+1+2+1+3+2+1+1+3+2+3+1+1+1+2+3+1+1+2+2+1+1+1+2+1+3+2+1+1+3+2+1+3+1+1+2+2+1+3+2+1+2+1+2+2+1+1+1+1+2+1+1+1+1+4+1+1+2+2+1+3+2+1+1+1+1+3+1+1+1+2+1+1+1+1+1+1+1+1+1+1+1+1+1+2+1+1+1+3+1+1+1+1+1+1+1+2+1+1+1+5+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+2+1+1+3+1+1+1+1+1+1+1+1+1+1+1+1+2+1+2+1+1+1+1+1+1+1+1+1+1+1+1+1+)

Optimal K value : 258

3). Clustering is a wonderful technique widely used to find groups of observations called clusters that share similar characteristics. The observations in the same cluster have greater similarity between them than other data points in a different group. The goal of the clustering algorithm is to obtain all similar data points in the same cluster and so data points in different clusters are dissimilar. K- means clustering is greatly useful for getting to know our data and to provide insights on almost all datatypes. There are different methods used to evaluate the performance of a clustering model or clustering quality and one such metric is purity. Each cluster is assigned with the most frequent label and purity is calculated by adding the number of most frequent labels in each cluster and finally dividing by the total number of data points. In general purity increases as the number of clusters increases. So if we have a clustering model that groups each observation as a separate cluster, then the purity is 1 and high purity can be easily obtained when the number of clusters is large. Poor clusterings have a purity value close to 0 and good clusterings have a purity of 1. Thus when we use purity as an evaluation metric for clustering quality, it does provide an understanding of how good a cluster is. However, purity alone cannot be the basis for choosing the optimal k. This is because the higher the k, the smaller the size of the clusters which makes it very difficult to obtain any useful information from the data. Hence it is important to note that even though a higher k results in more number of clusters, it affects the clustering quality which is undesirable. This shows that it is important to use different metrics to evaluate the performance of our clustering model and to not base the evaluation on purity alone.

4). Given the cons specified in question 3, I decided to choose the optimal k value by not only taking into consideration the purity generated by that k, but also penalizing k values that are very big. The reasoning is because the clustering algorithm offers no important information if the k is

very big, even if the purity is near 100%. So, my method is as following: choose the k value that minimizes purity * (# of vertices - k) / (# of vertices). This method factors in the purity but also penalizes large k values.