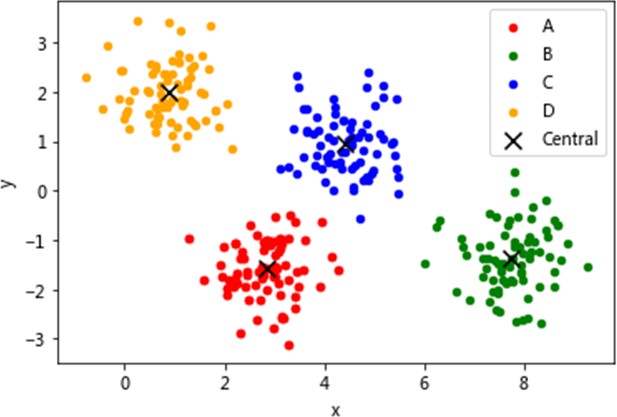
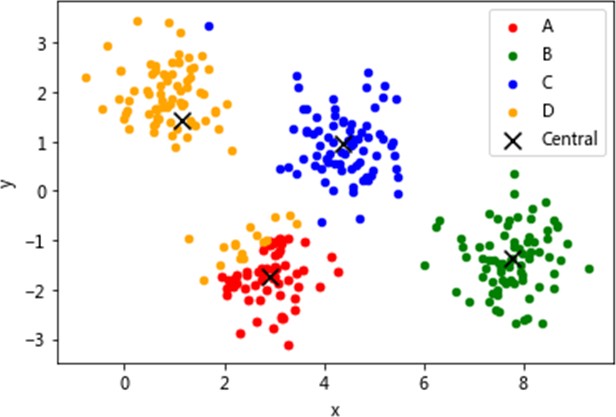
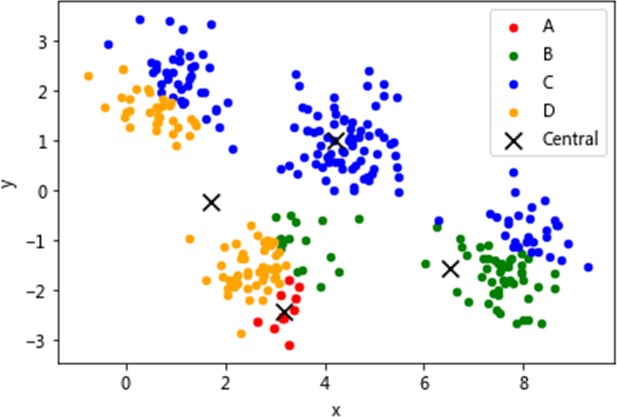
**K‐means Clustering Implementation**

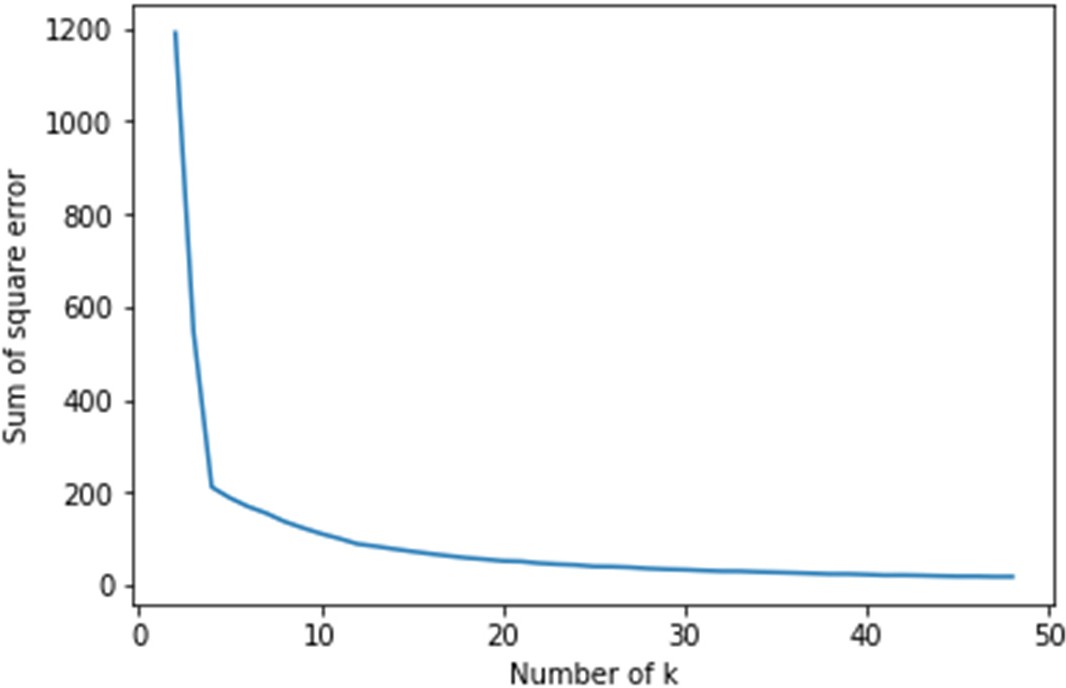
Your task is to use Python (along with numpy and Pandas) to implement the well‐known clustering algorithm, K‐means, based on a synthetic dataset **cdata.csv**. This dataset contains two data columns, “X” and “Y”, and one “cluster” column (1, 2, 3, and 4). In implementing K‐means, you need to use “X” and “Y” as **features** for clustering while the “cluster” column is for your validation. Note that it is not necessary to perfectly clustering all of the data points into clusters. Also note that the “cluster” column cannot be used in clustering.

1. Randomly select data points as the initialized centroids. By default, please set K=4. Report and plot the process until convergence. The centroids also need to be plotted. An example is shown below. Note that it may not have 3 rounds (it can be 4 or 5 rounds, depend on initialized centroids). 

Round 1 Round 2 Round 3

1. Re‐execute your K‐means clustering algorithm by changing K from 2 to 50 (from 2 to 10 is also

okay). Plot the K value (x‐axis) vs. the value of Sum of Squared Error (SSE) (y‐axis) as below. Note that it is reasonable and acceptable if the curve is 凹凸不平. 



1. Try 10 times of randomly initialized centroids, and plot their SSE values (y‐axis)

