Methodology

1. To find the “natural clusters”, I first plotted the given 2D-data. From the plot it is very clear that the data has 3 natural clusters.
2. On the given data, I have performed k-means with 3 clusters. For k-means, I used ‘sklearn’ package. In the code, ‘kMeans\_clustering\_main’ method has most of the k-means logic.
3. To find the outliers, code takes an input ‘threshold\_percentile’ from user, I followed below approach: Found the distance of each data point from its corresponding cluster centroid, this information is stored in ‘distance\_from\_assigned\_cluster\_centriod’ column. Found the 95th percentile (assuming ‘threshold\_percentile’ input value is 95) value (say threshold\_distance) on the ‘distance\_from\_assigned\_cluster\_centriod’ column. Any data point having distance from its cluster centroid more than the threshold\_distance value is labelled as Outlier. I have added a new column ‘is\_outlier’ which indicates if the data point is Outlier or not, depending on the input threshold\_percentile.

Inputs to the code:

1. **input\_file** : Location of the input csv data file
2. **threshold\_percentile** : threshold percentile distance after which data is labelled as Outlier
3. **output\_folder** : Location to save output files

Output Files:

1. **given\_data.png** : Basic plot of the given data
2. **cluster\_centers.csv** : This file has k-means centriods co-ordinates.
3. **data\_with\_cluster\_assignment.csv** : This file has given input data, with an extra column ‘assigned\_cluster’. This column indicates the cluster\_id that data point belongs to
4. **plot\_with\_cluster\_assignment.png** : This image shows how clusters are assigned to the data. Each cluster is given a different color.
5. **final\_data\_with\_outlier\_info.csv** : This file has complete output details including is\_outlier, distance of the data point from its cluster centroid

Sample command from CLI:

python kmeans\_cluster.py --input\_file /Users/sandy/Documents/data\_2d.csv --threshold\_percentile 95 --output\_folder /Users/sandy/Documents/