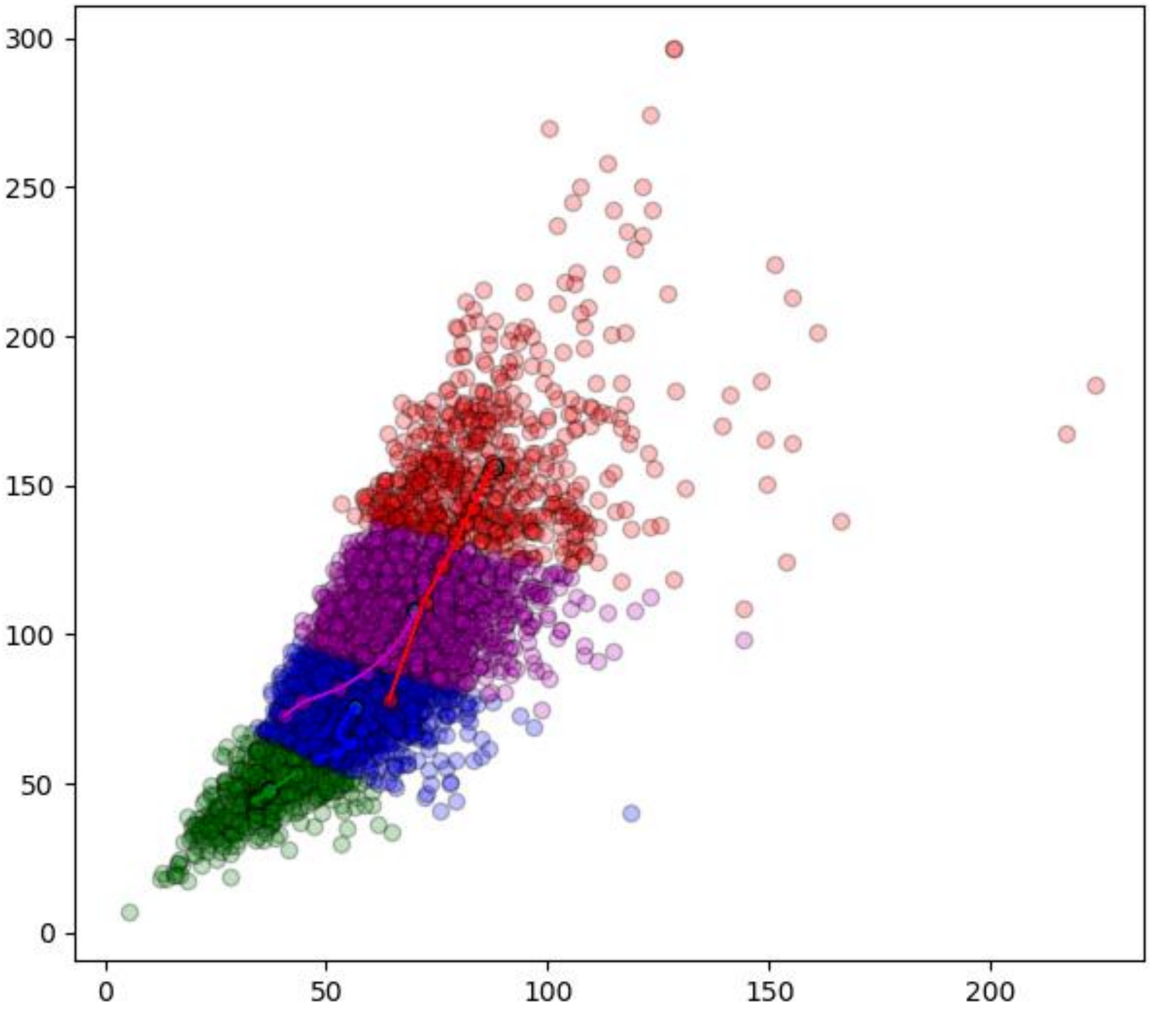
K-Means Clustering

1 Find the Cluster

All the raw program output can be found in the *Resut-KMeans.txt* file.

1.1 Plot

Here is the plot for **Training Set** and the movements of the cluster centroids:

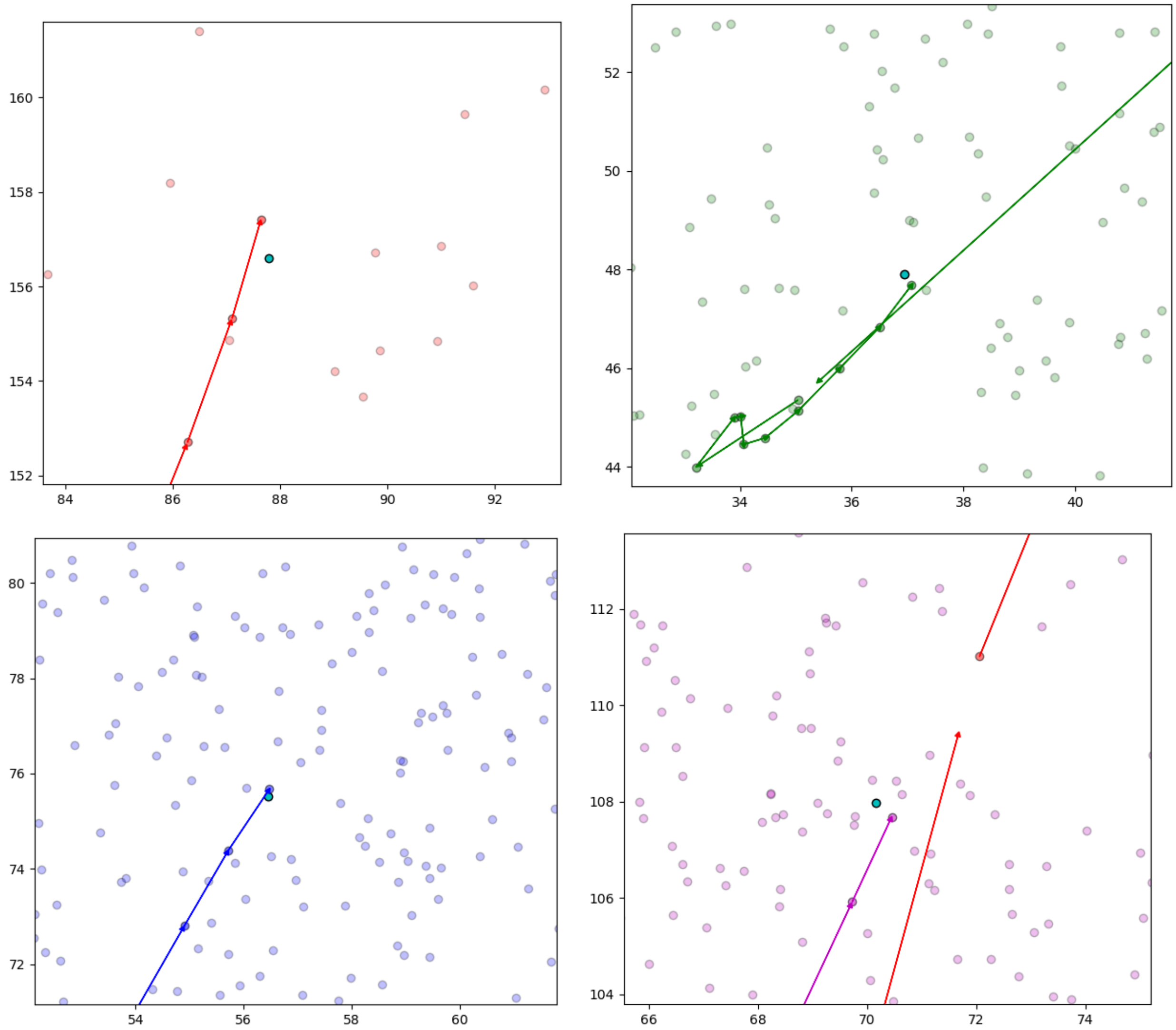


1.2 Closest Training Samples

From the program output, we can know that these samples are the closest one to their cluster center:

• For centroid [87.648034, 157.417381] (red)	[87.784722, 156.593750]
• For centroid [37.069545, 47.683058] (green)	[36.935764, 47.911458]
• For centroid [56.475492, 75.671486] (blue)	[56.446181, 75.526042]
• For centroid [70.449837, 107.687520] (magenta)	[70.149306, 107.979167]

And the in the plot image, those 4 samples are marked as color cyan. And here is the zoom-in image for them:



1.3 Clustering

1.3.1 Process

For the initialize of the cluster's centroids, just random choose 4 samples as the centroids.

Then with each iteration, we will get a new centroid for each cluster that have a less cost. So it will become better and better to converge to the optimal.

1.3.2 Mathematically Prove

First, there are at most k^N ways to partition N data points into k clusters. This is a large but finite number.

For each iteration of the algorithm, we produce a new clustering based only on the old clustering. And it has only 2 ways:

- if the old clustering is the same as the new, then the next clustering will again be the same.
- If the new clustering is different from the old one, then the newer one has a lower cost

Since the algorithm iterates a function whose domain is a finite set, the iteration must eventually enter a cycle. The cycle cannot have length greater than 1, because otherwise, there will have some clustering which has a lower cost than itself which is impossible.

Hence the cycle must have length exactly 1. That's also means *K-Means Clustering* converges in a finite number of iterations.

1.3.3 Closest Sample

The closest samples are almost the center of the Cluster, so it make sense.