Home Work Assignment-6

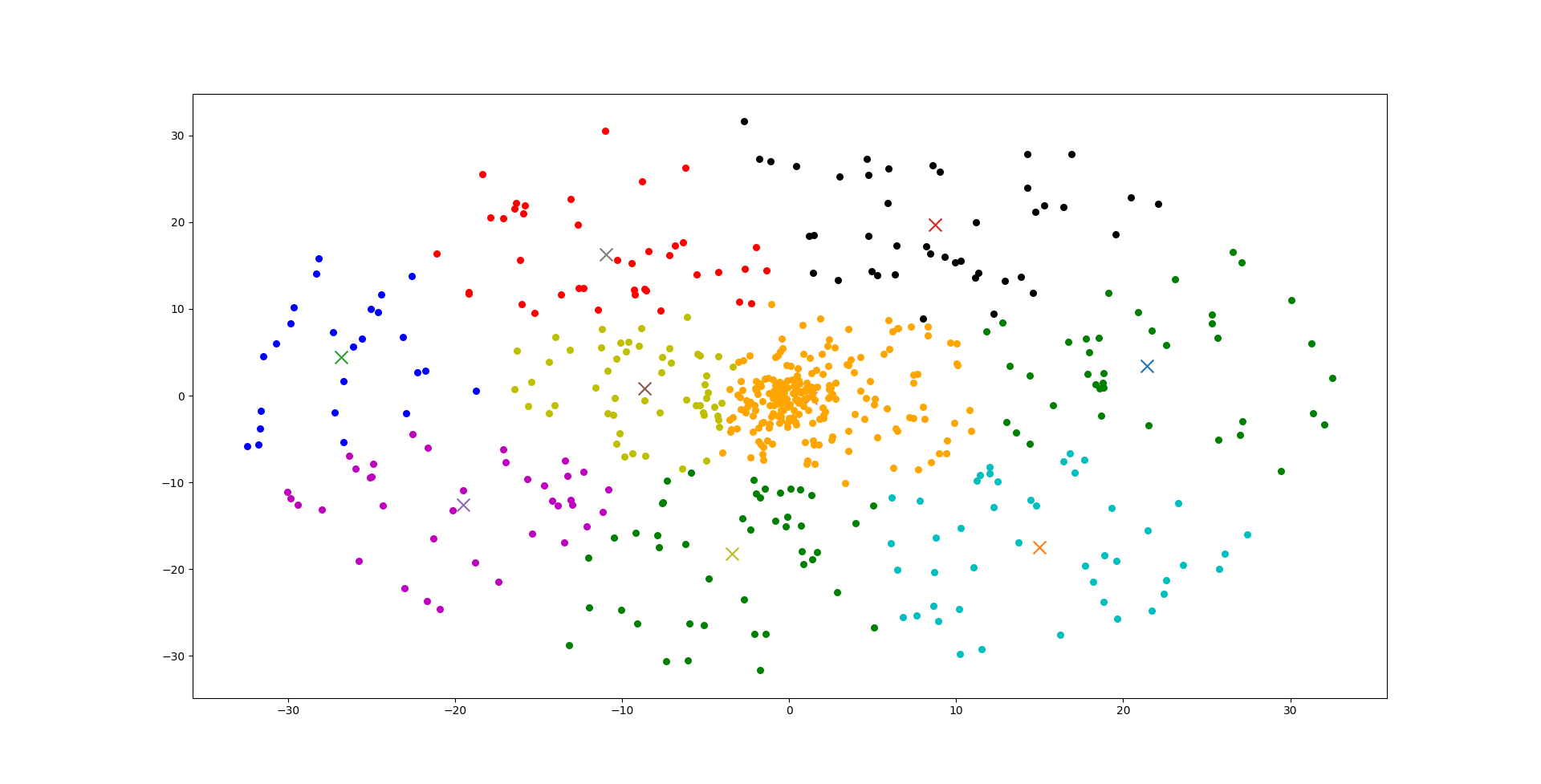
In the homework assignment 6, I used the K-means clustering strategy to do clustering of the data without using libraries and clustering the data from scartch. I initialized the cluster centroids trivially to the intial data points.I have done clustering of the data points by calculating the distances from the cluster and assigning the data points to clusters closest to them.

After the assignment, we recalculate the mean data point of each cluster assigned previously and assign mean as centroids of the current clustering . Then we calculate the distance between the previously calculated centroids and the current centroid which have calculated based on the clustering. If the difference is greater than the tolerance value, then the loop continues with the current centroids as the cluster centroids. The loop continues until the distances between the current centroid and the previous centroid is less than the the tolerance value. When the loop reaches this stage, we know that the algorithm has reached the optimal stage and loop stops. Also we set the maximum number of iterations so that the loop does not turn into an infinite loop.

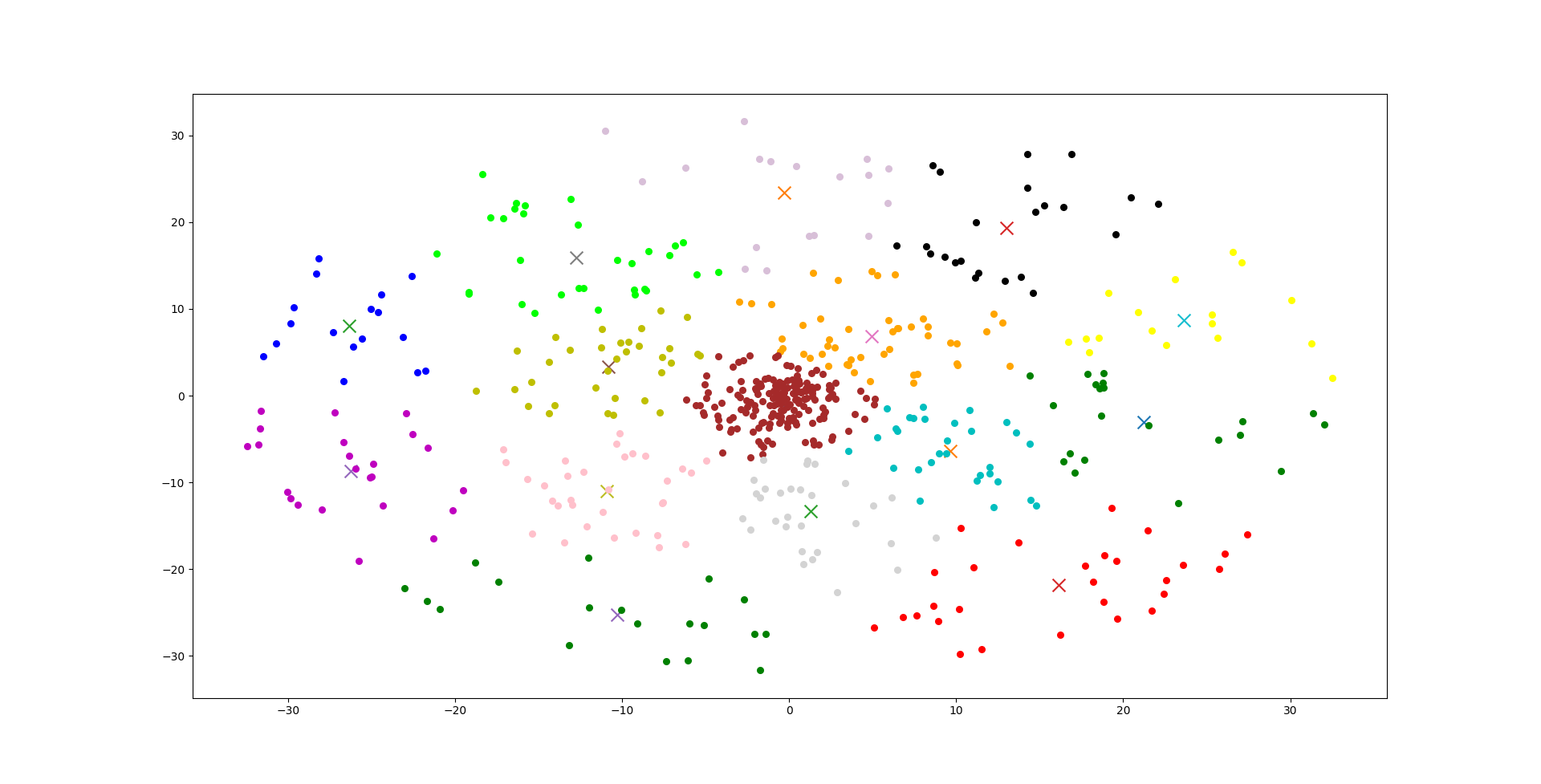
The number of clusters(k) is also to be considered. Various k values can be used to get the output, I plotted the points of the data and based on the data plot that I got, I understood that k value has to be something above k>5 because the data points expand across a vast region has a circular kind of shape. I came to a conclusion that k=15 would be an optimal value because of the size, shape of the plot of the data and the distance between the centroids and the data points assigned in each cluster was optimal and the variation existing in the data.

Further, I plot the datapoints that are assigned to each cluster.

When k=10,



At k=15,



So, I used the k=15, for the output predictions data that I have submitted as it was optimal for the data.