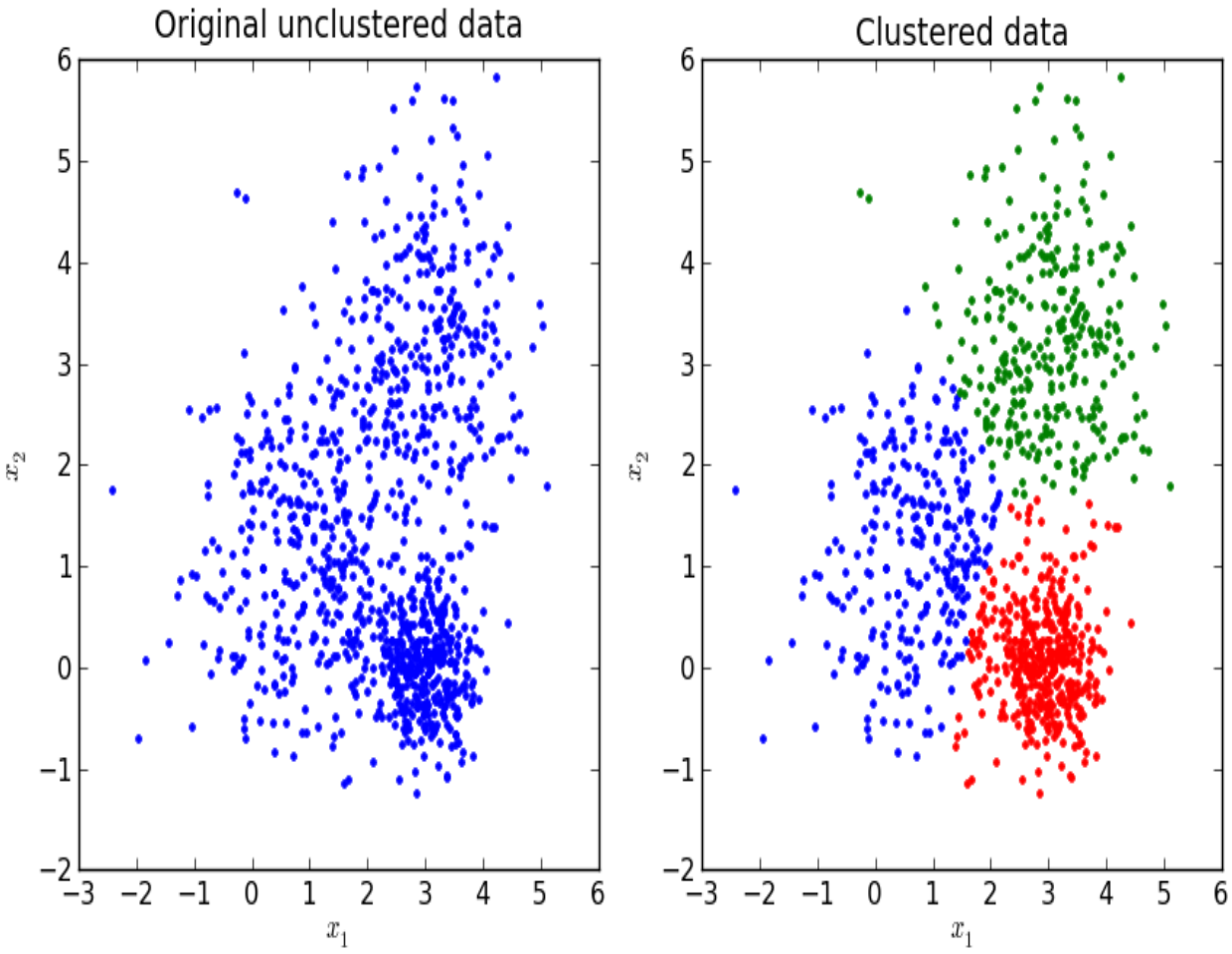


Artificial Intelligence

K-Means Clustring

Name : Ziad Ashraf Shreet

ID : 3947

K-means clustering → is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity.

Objective:

Machine learning is to derive techniques for unsupervised learning on data. This kind of data analysis is very helpful in many applications that require classification of data, such as identifying cancerous cells within a large sample, clustering words with similar definitions for better search engine accuracy,identifying outliers in student's academic performance for better refinement of habits, or even for detecting landmines in a battlefield.

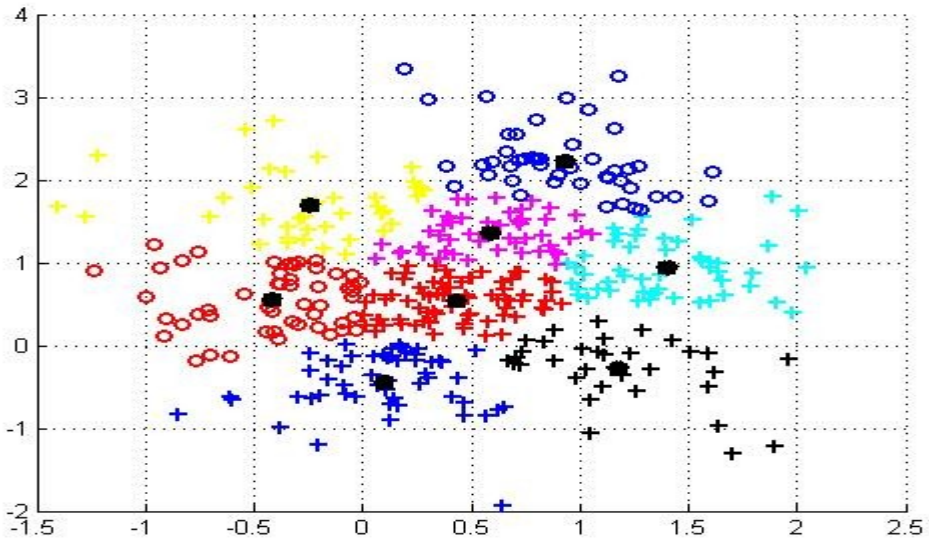
The results of the K-means clustering algorithm are:

1. The centroids of the *K* clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have

formed organically. The "Choosing K" section below describes how the number of groups can be determined.

Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.



K-means algorithm iteratively minimizes the distances between every data point and its centroid in order to find the most optimal solution for all the data points.
1. Random points of the data-set are chosen to be the initial centroids.

2. Distances between every data point and the centroids are calculated and stored.

3. Based on distance calculates, each point is assigned to the nearest cluster

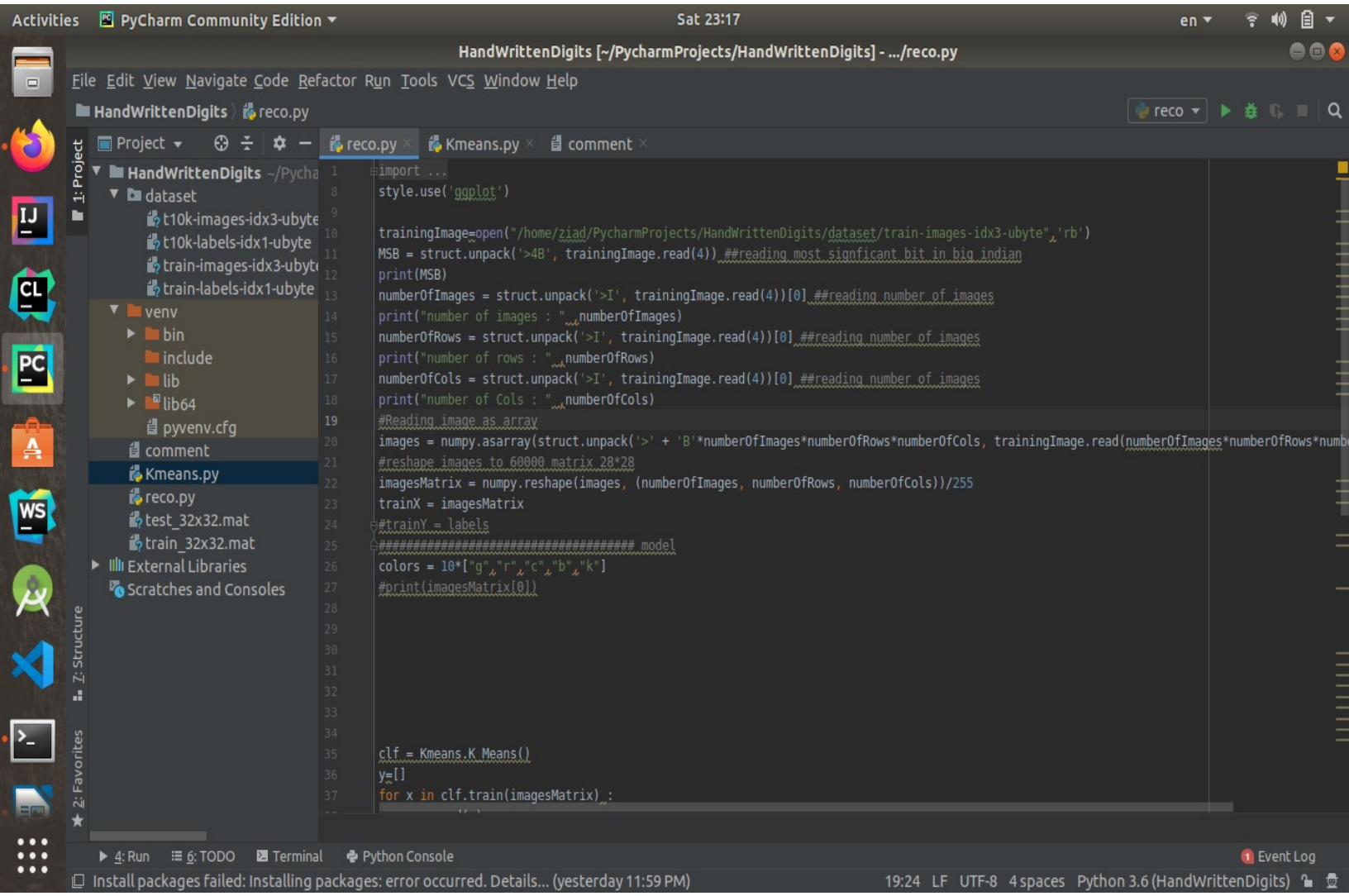
4. New cluster centroid positions are updated: similar to finding the mean

5. If the centroid locations changed, the process repeats from step 2, until the calculated new center stays the same, which signals that the clusters' members and centroids are now set and will be constant movment.

I choosed the MINIST Handwritten Digits Data-Set

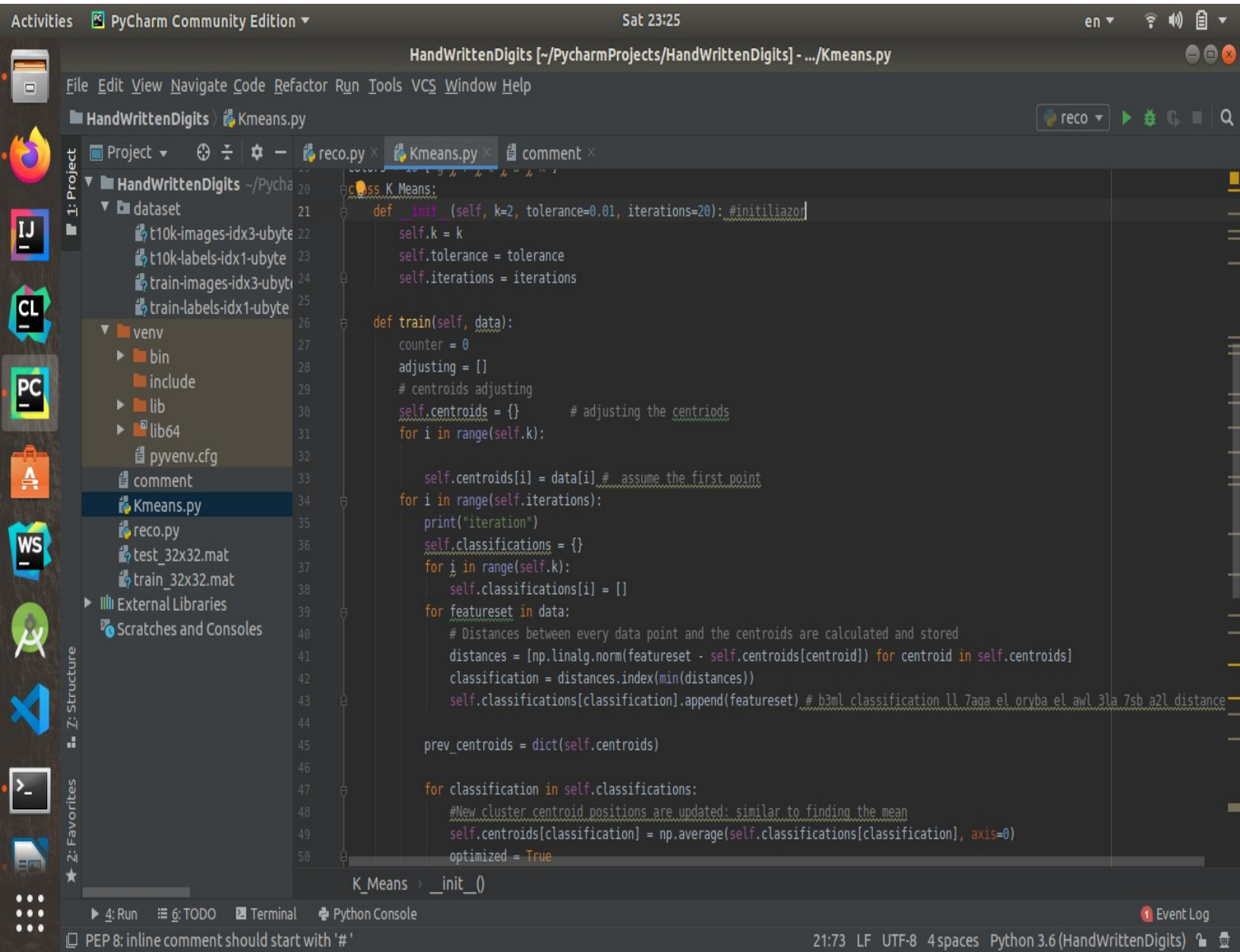
The Data-Set is a batch of 60,000 images

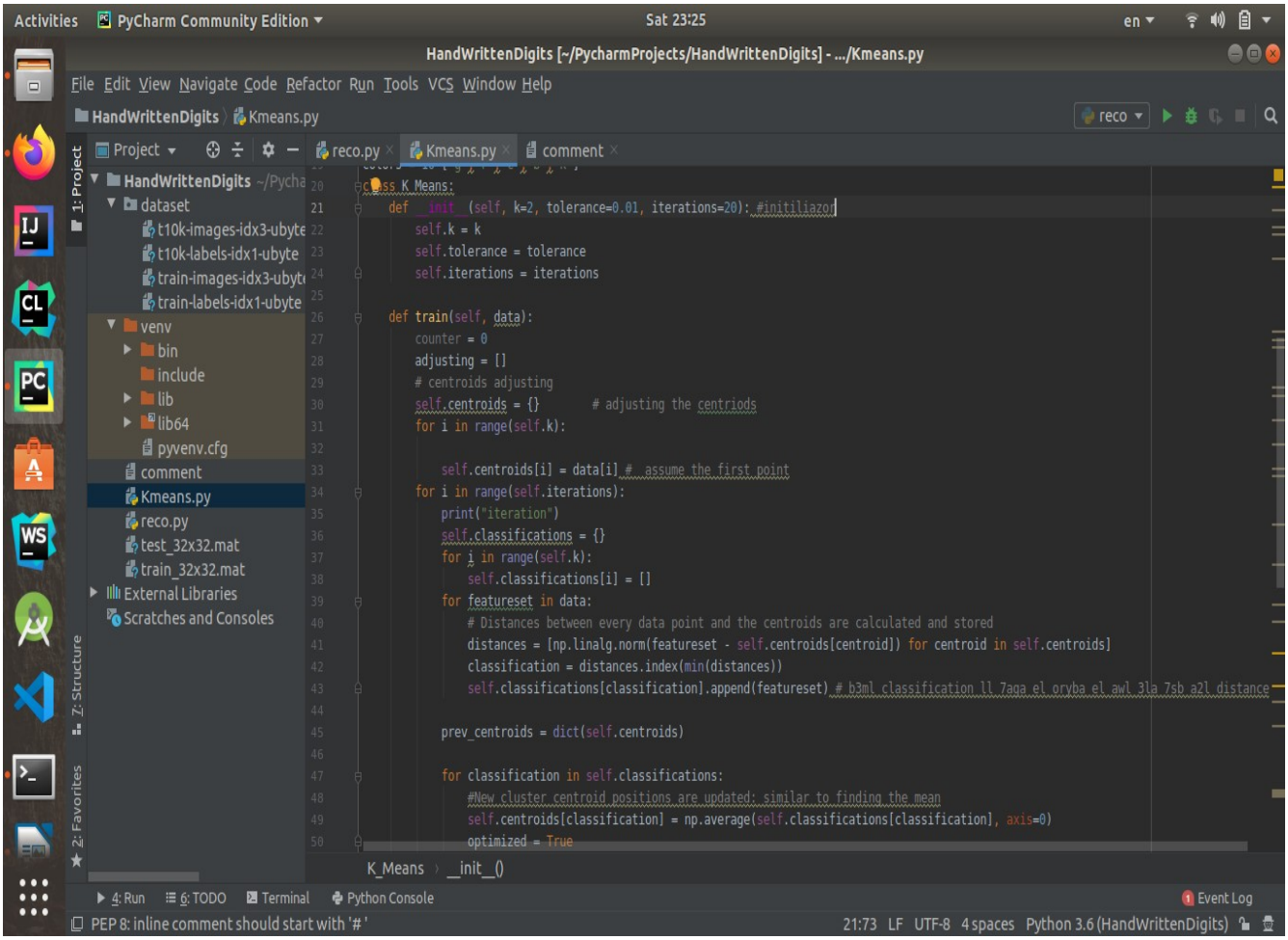
First I began with extracting and loading the Data-set file to nd-array and reshaped it into Matrix then adjusting the attributes to be in the interval [0,1].



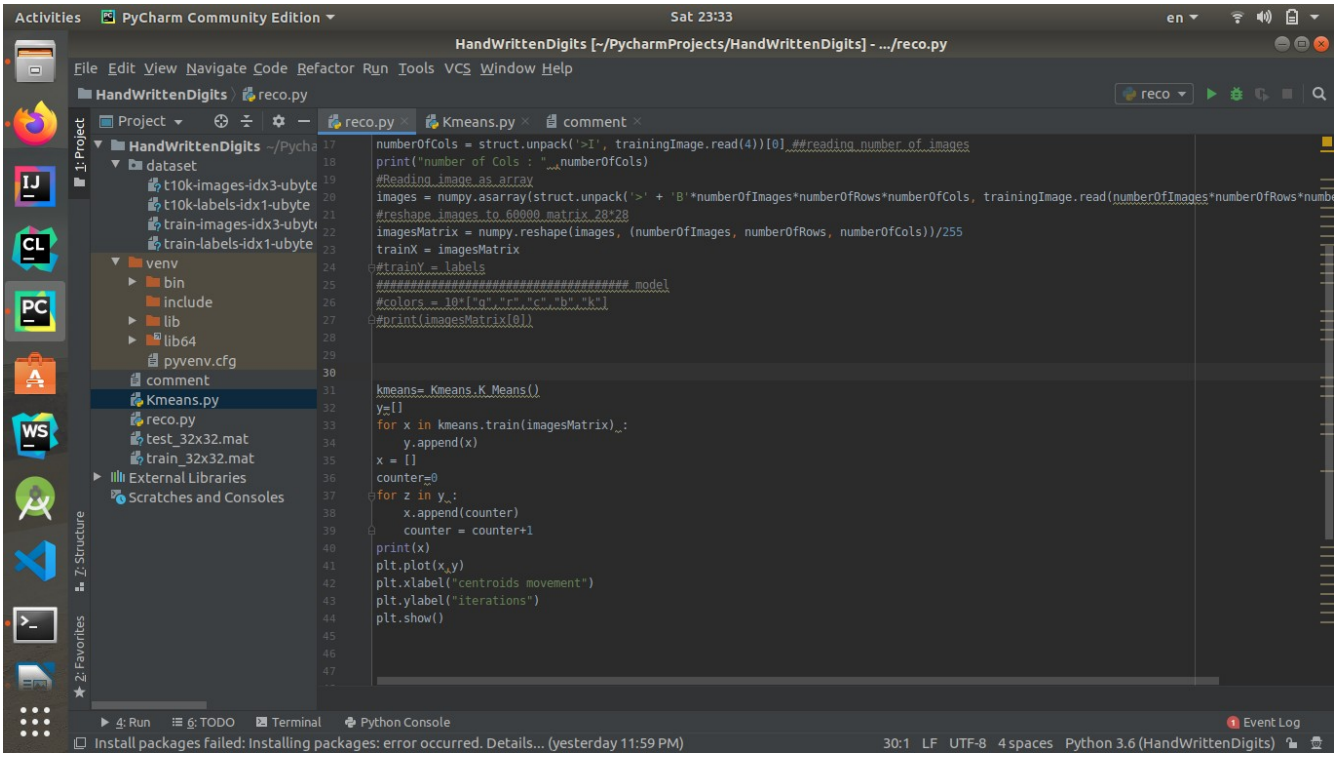
I used struct.unpack() to *unpack the* data from strings using format specifiers made up of characters representing the type of the data and optional count and endianness indicators.

K-Means Implementation with comments and the steps are the same as I explained in the introduction :



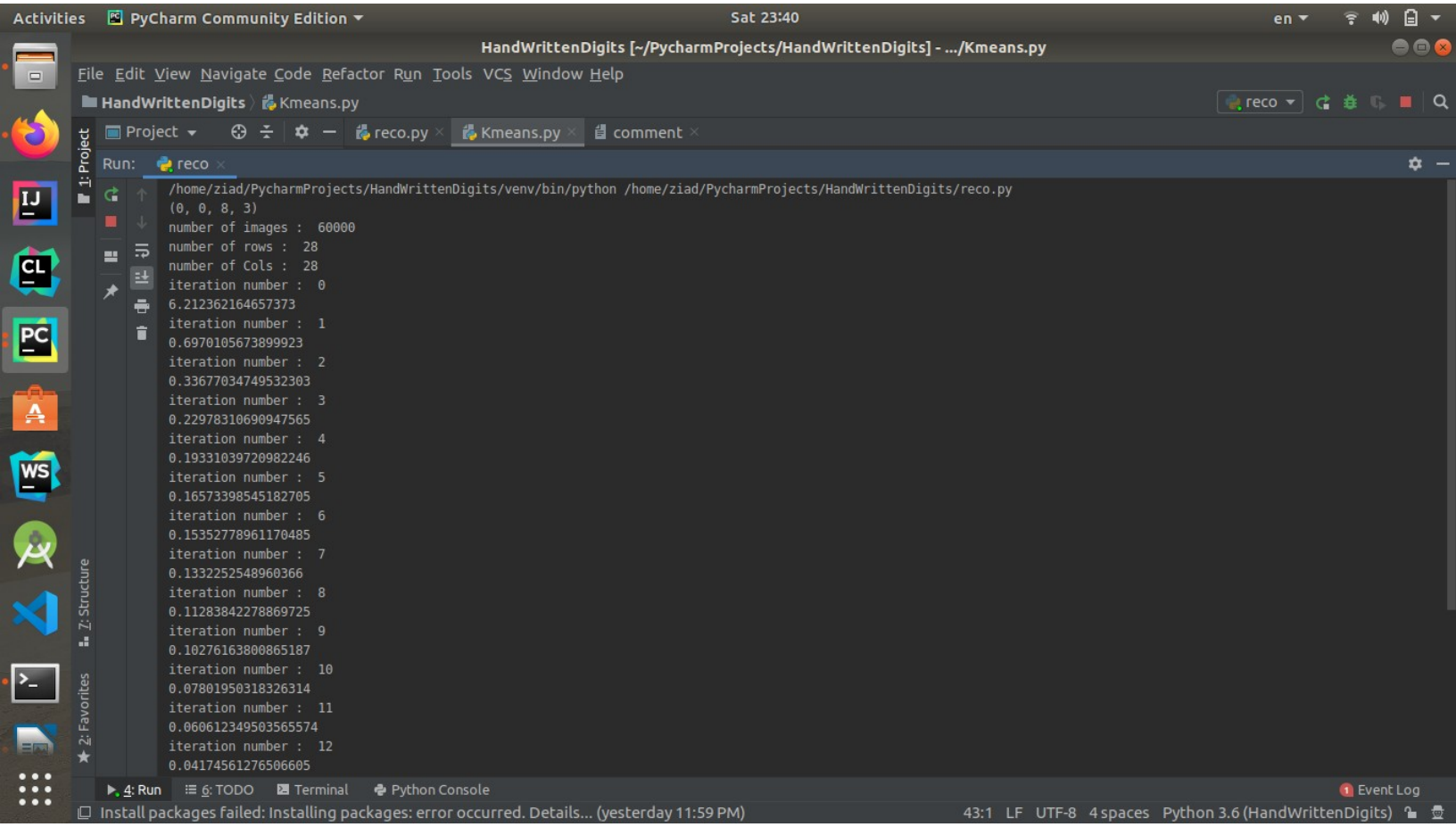


First I tested the algorithm for small data-set then I tested it for the Big Data-set:

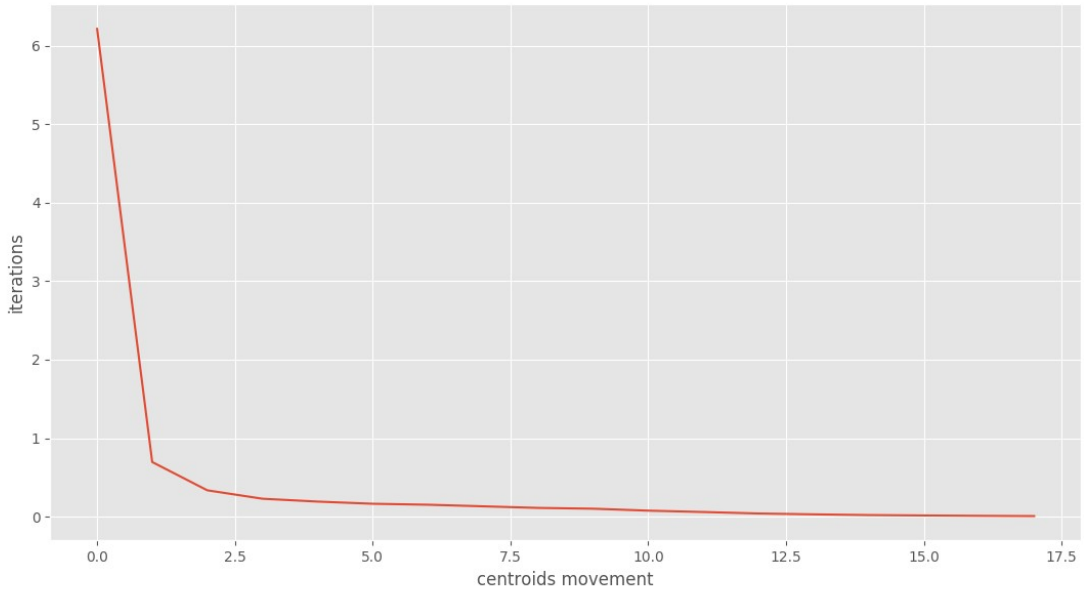


for k = 2 :

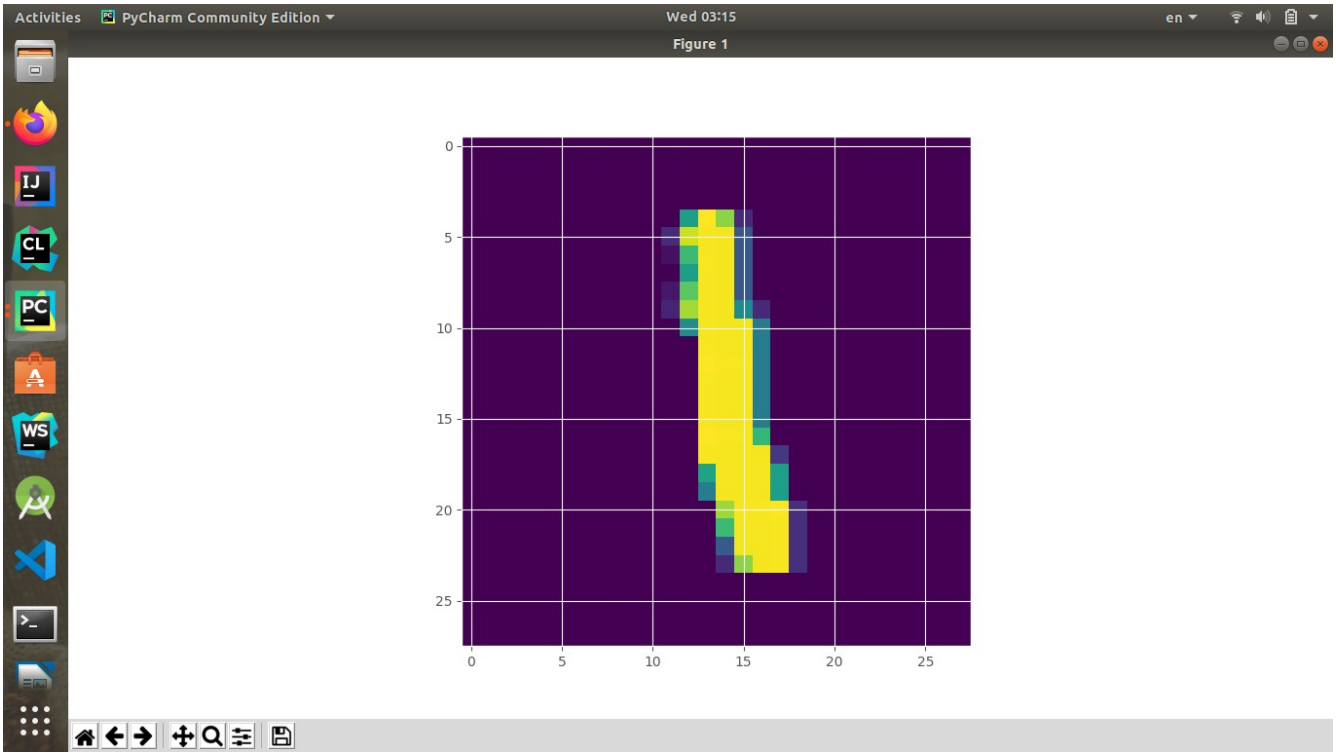
choose random 2 images to be the initial centriods then calculate the distances from the images to the centriods the nearest centriod will be the cluster that the image will belong to if it exceeds the tolerance the algorithm will break .



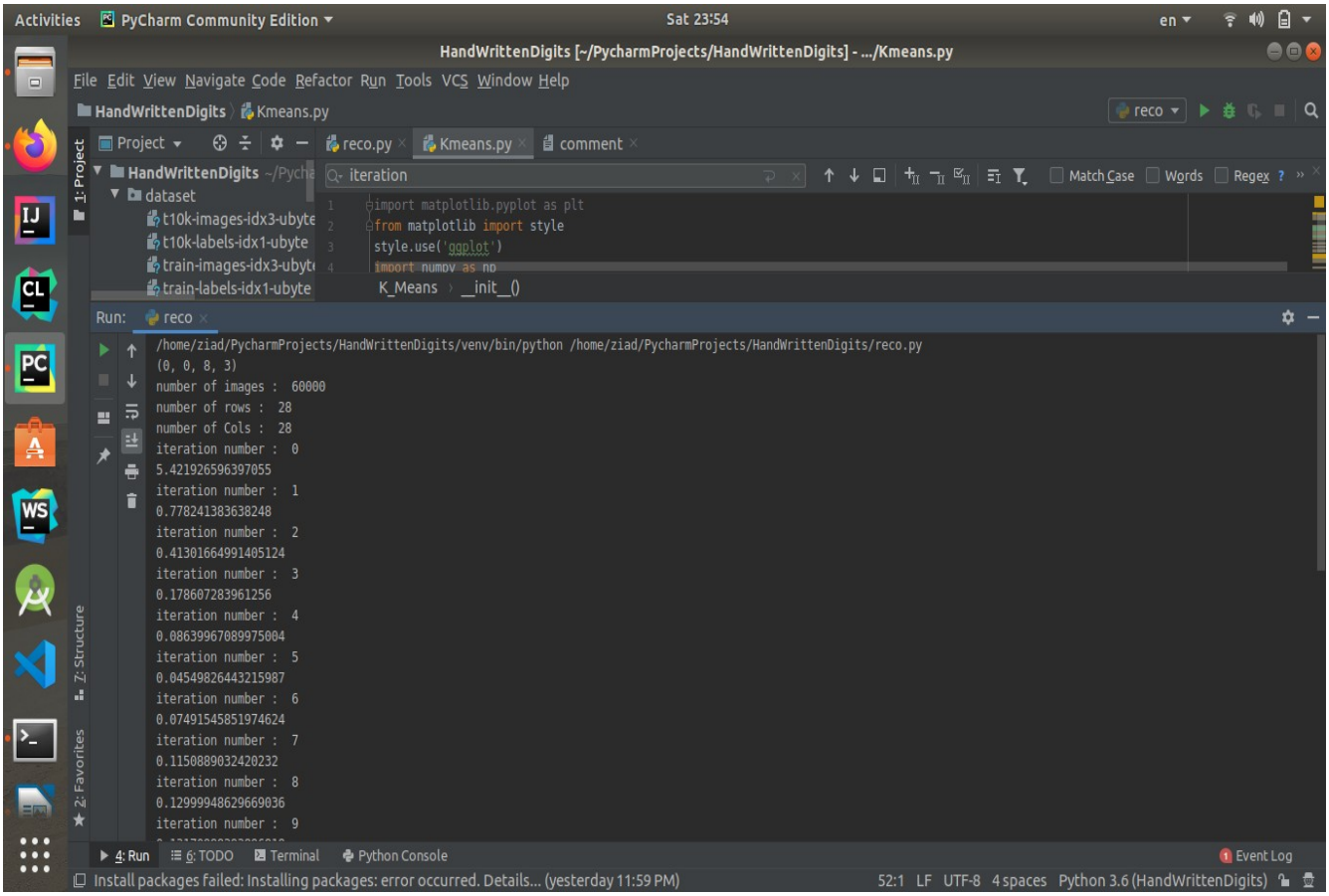
Plot the K-Means objective function (distortion measure) as a function of iteration and verify that it never increases.
The objective function :



some representative images :

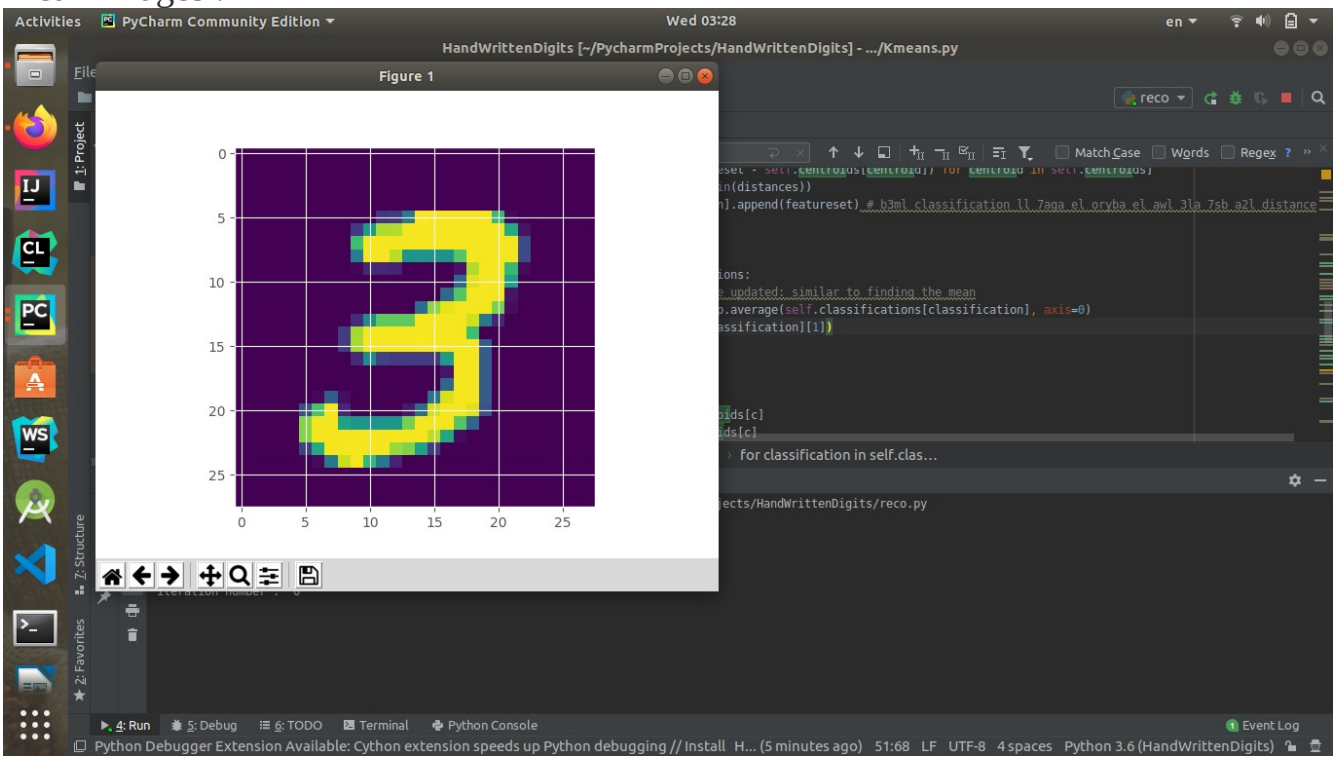


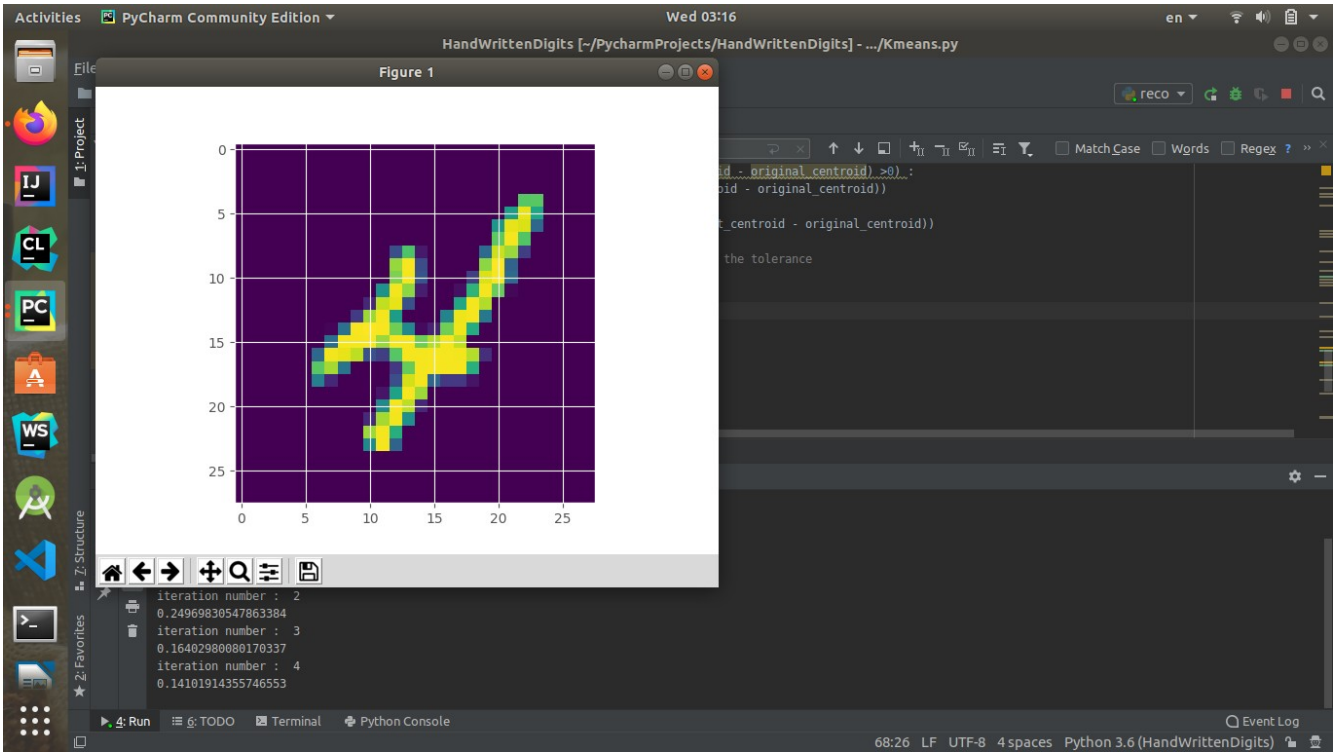
for k = 4:

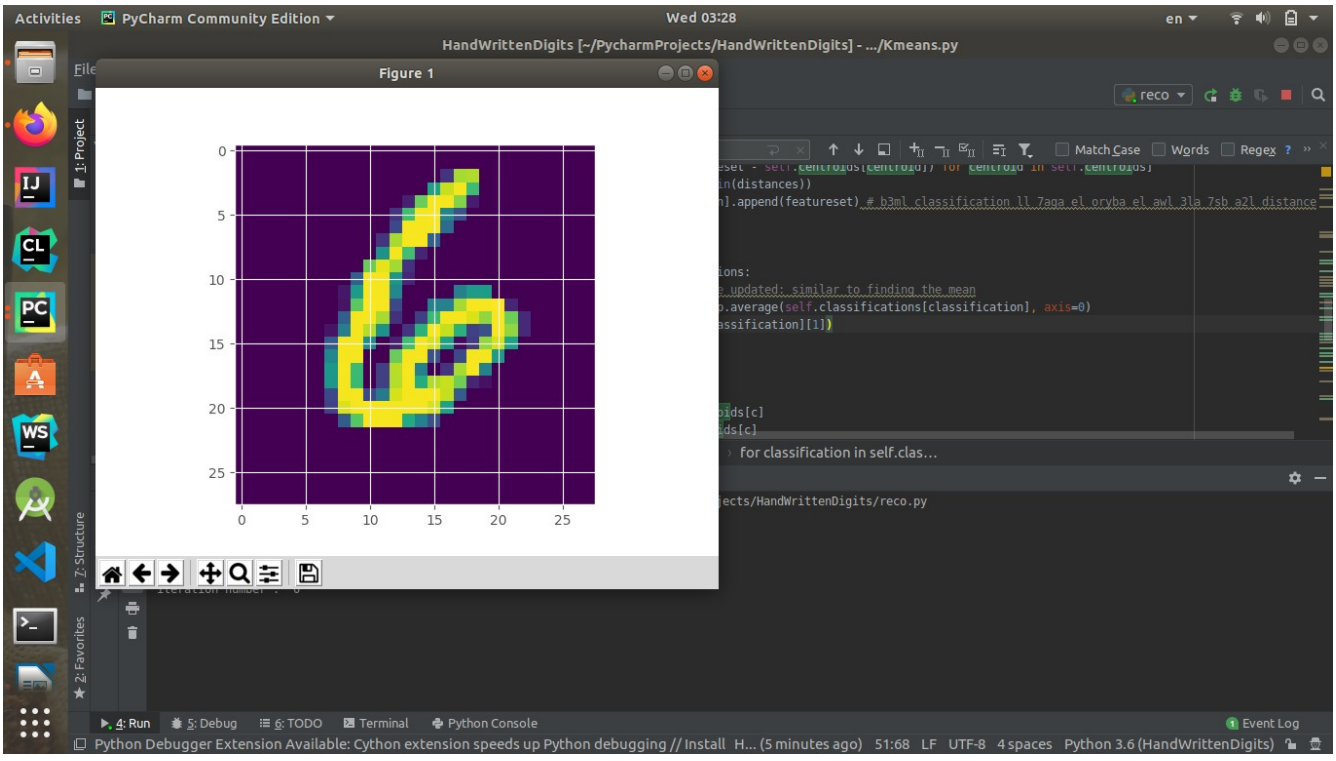


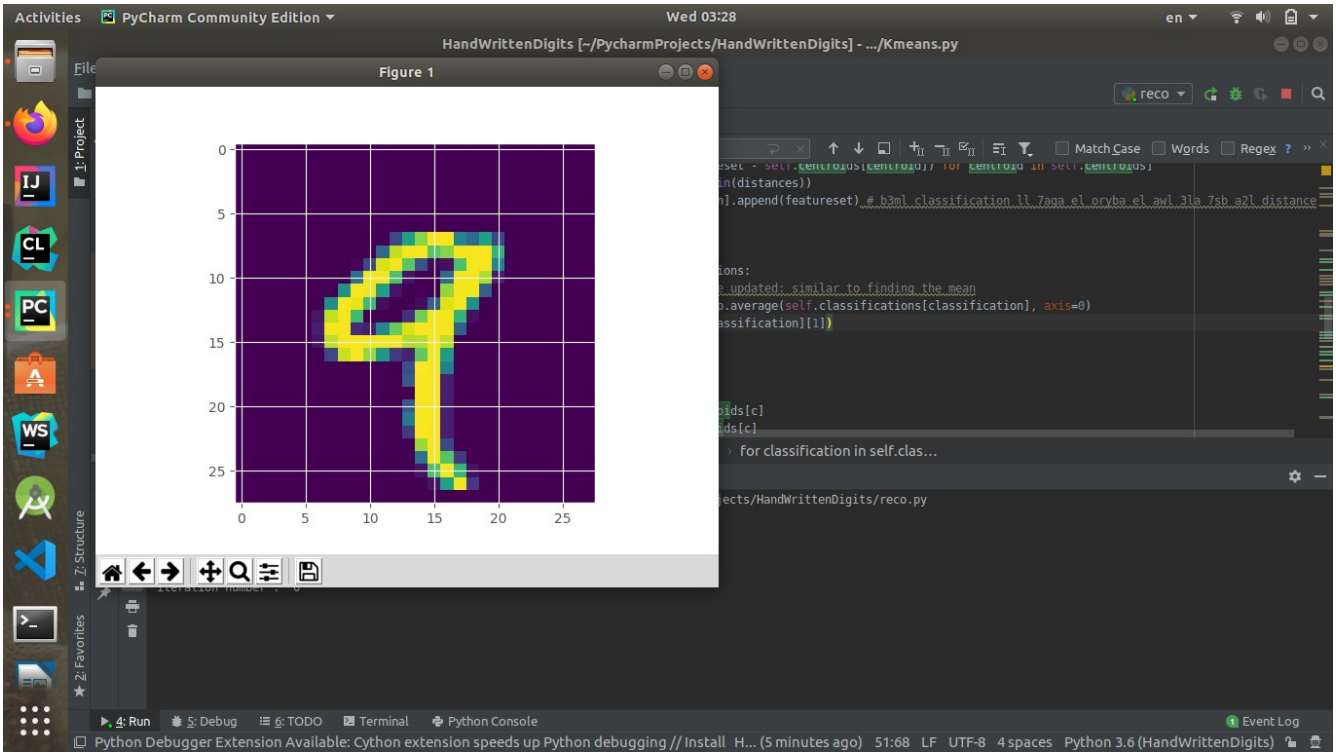
Train Accuracy = 0.4719, Test Accuracy = 0.4632

Mean Images :

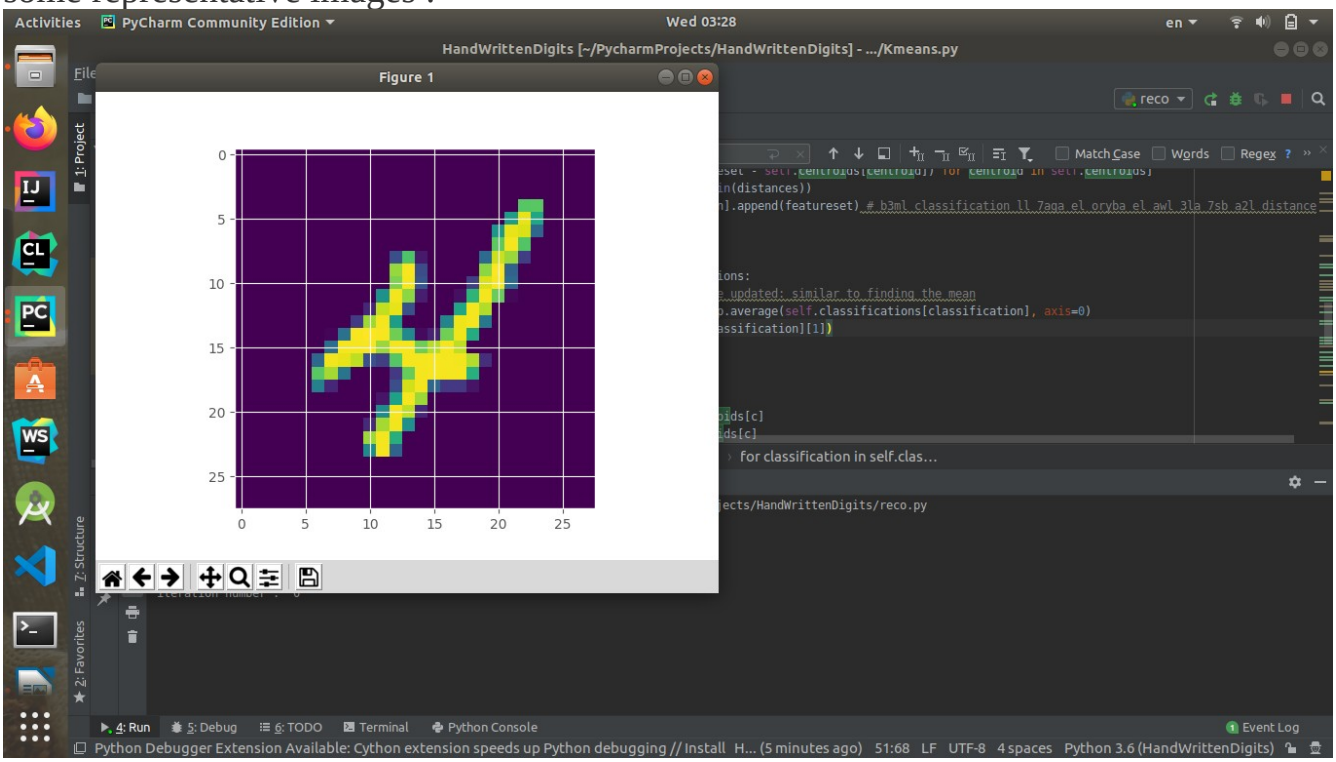


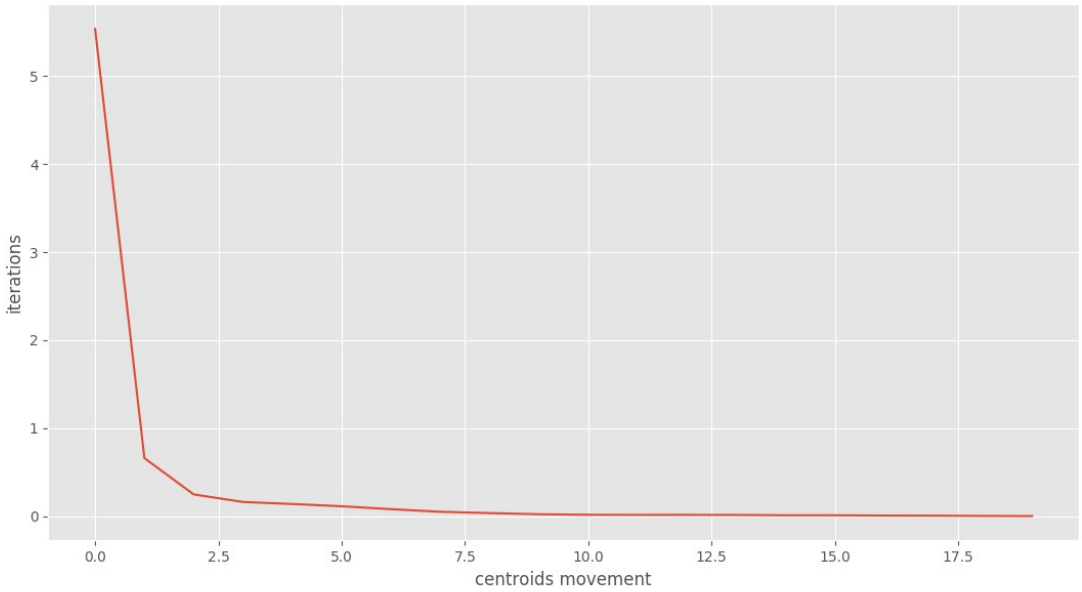

Activities
PyCharm Community Edition
Wed 03:16
en
HandWrittenDigits [~/PycharmProjects/HandWrittenDigits] - .../Kmeans.py
Figure 1
reco
Match Case
Words
Regex
iteration number : 2
0.24969830547863384
iteration number : 3
0.16402980080170337
iteration number : 4
0.14101914355746553
4: Run
6: TODO
Terminal
Python Console
Event Log
68:26 LF UTF-8 4 spaces Python 3.6 (HandWrittenDigits)


Activities
PyCharm Community Edition
Wed 03:28
en
HandWrittenDigits [~/PycharmProjects/HandWrittenDigits] - .../Kmeans.py
Figure 1
reco
Match Case
Words
Regex
for classification in self.clas...
4: Run
5: Debug
6: TODO
Terminal
Python Console
Event Log
Python Debugger Extension Available: Cython extension speeds up Python debugging // Install H... (5 minutes ago) 51:68 LF UTF-8 4 spaces Python 3.6 (HandWrittenDigits)

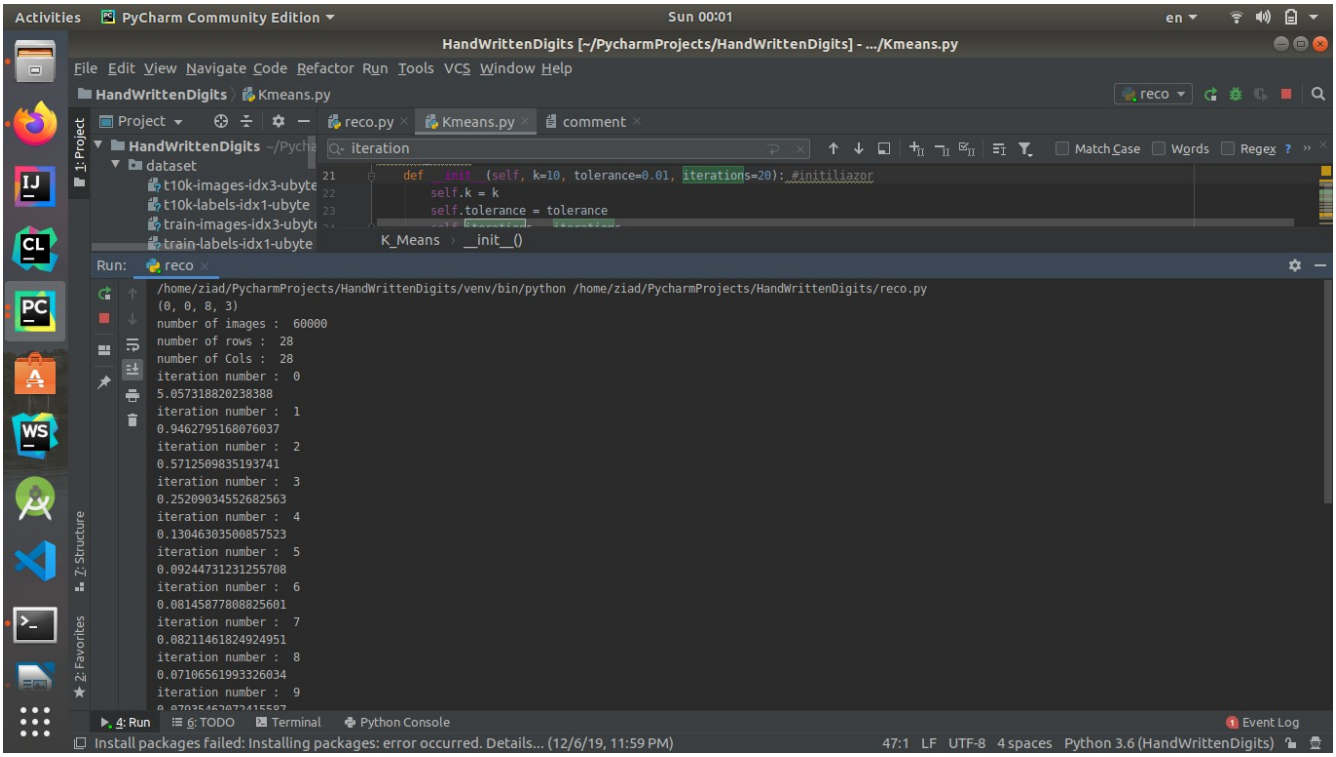


some representative images :




iterations
centroids movement
0
1
2
3
4
5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

for K = 10 :



Train Accuracy = 0.5193, Test Accuracy = 0.5092

The objective function :