**Unsupervised Clustering for images-**

**Algorithms or packages used-**

* OpenCV- Image reading and resizing
* Scikit-Learning- Data Shuffling, Kmeans and Guassian Mixture clustering, Principal Component Analysis, and performance metrics.
* Keras with tensorflow backend- VGG16 Convolution Network models with ImageNet weights.

Principle​ ​component​ ​analysis​ ​is​ ​a​ ​technique​ ​that​ ​brings​ ​out​ ​the​ ​variations​ ​in​ ​the​ ​dataset.​ ​It reduces​ ​the​ ​number​ ​of​ ​features​ ​of​ ​a​ ​data​ ​set​ ​such​ ​that​ ​the​ ​features​ ​it​ ​provides​ ​produce maximum​ ​variance​ ​between​ ​the​ ​points.​ ​This​ ​allows​ ​us​ ​to​ ​represent​ ​the​ ​differences​ ​between​ ​the different​ ​points​ ​in​ ​less​ ​features,​ ​and​ ​makes​ ​it​ ​easier​ ​for​ ​our​ ​clustering​ ​algorithms​ ​to​ ​operate.

We​ ​therefore​ ​use​ ​PCA​ ​to​ ​further​ ​reduce​ ​the​ ​output​ ​of​ ​the​ ​pooling​ ​layer​ ​of​ ​the​ ​convnet​ ​before passing​ ​on​ ​the​ ​data​ ​to​ ​our​ ​clustering​ ​algorithm.

KMeans​ ​clustering​ ​is​ ​a​ ​method​ ​that​ ​clusters​ ​data​ ​points​ ​together​ ​based​ ​on​ ​feature​ ​similarity.​ ​It starts​ ​by​ ​randomly​ ​assigning​ ​data​ ​points​ ​to​ ​clusters​ ​and​ ​then​ ​works​ ​on​ ​minimising​ ​the​ ​eucledian distances​ ​of​ ​points​ ​from​ ​their​ ​cluster​ ​centroids.

In​ ​Gaussian​ ​Mixture​ ​Model​ ​(GMM)​ ​clustering,​ ​the​ ​algorithm​ ​assumes​ ​that​ ​each​ ​cluster​ ​is​ ​in​ ​a gaussian​ ​distribution​ ​and​ ​assigns​ ​it​ ​to​ ​the​ ​cluster​ ​it​ ​has​ ​the​ ​largest​ ​probability​ ​of​ ​belonging​ ​to. GMM​ ​clustering​ ​has​ ​an​ ​advantage​ ​of​ ​not​ ​requiring​ ​clusters​ ​to​ ​be​ ​spherical,​ ​i.e.​ ​the​ ​cluster boundary​ ​does​ ​not​ ​have​ ​to​ ​be​ ​equidistant​ ​from​ ​its​ ​centroid.

**Data-** Park image data is taken and placed and 4 different folders. Let say the four different folder names as- aa,bb,cc,dd. Each of these folders have 5-6 images of bench, bridge, frisbee and kite keywords. These folders are used for training and other two folders say t10, t12 are used for testing.   
**Clustering**- 4 clusters are specified since there are 4 type of images. All the images in t10, t12 may fall under these 4 clusters say (0,1,2,3)

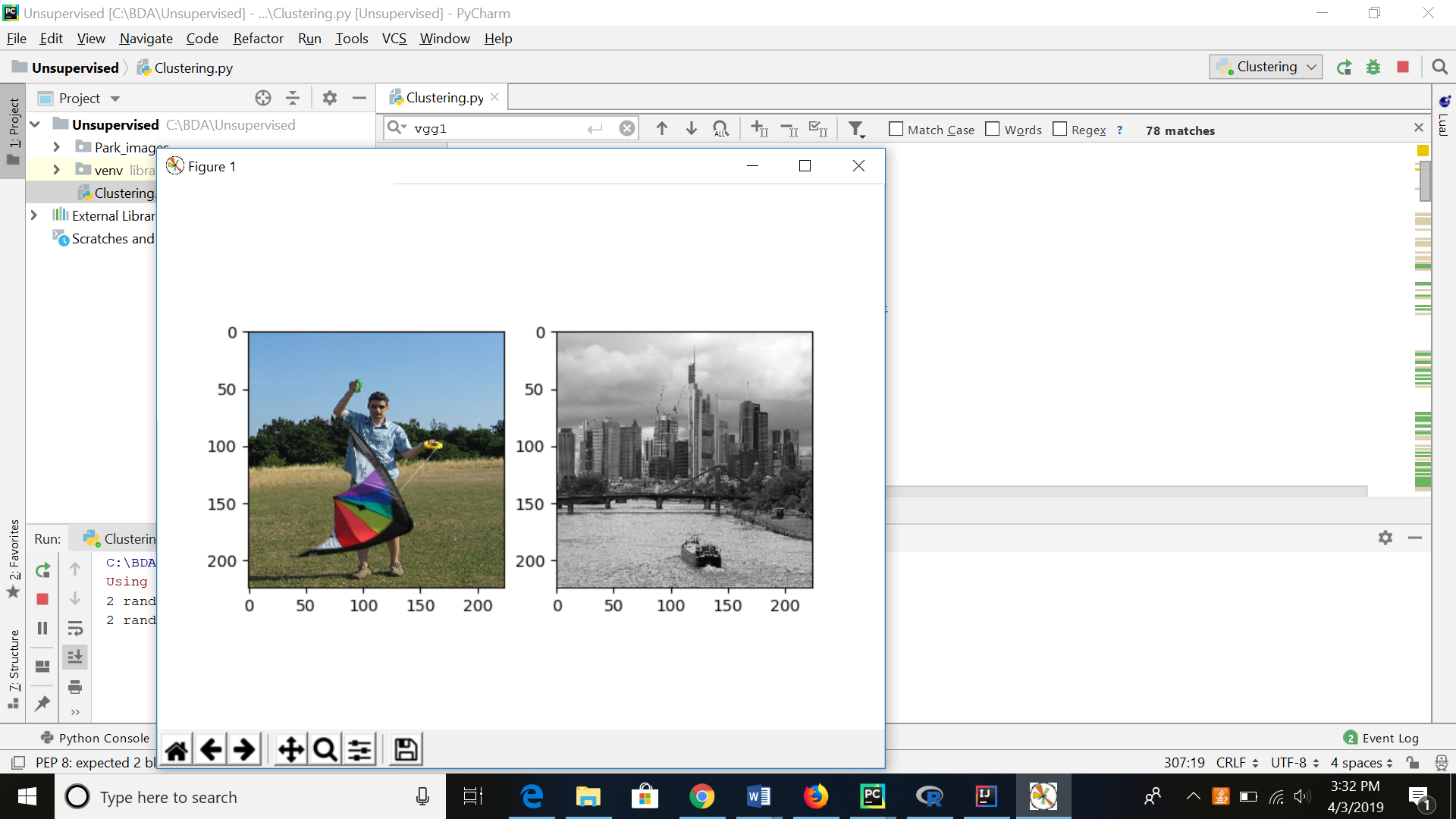
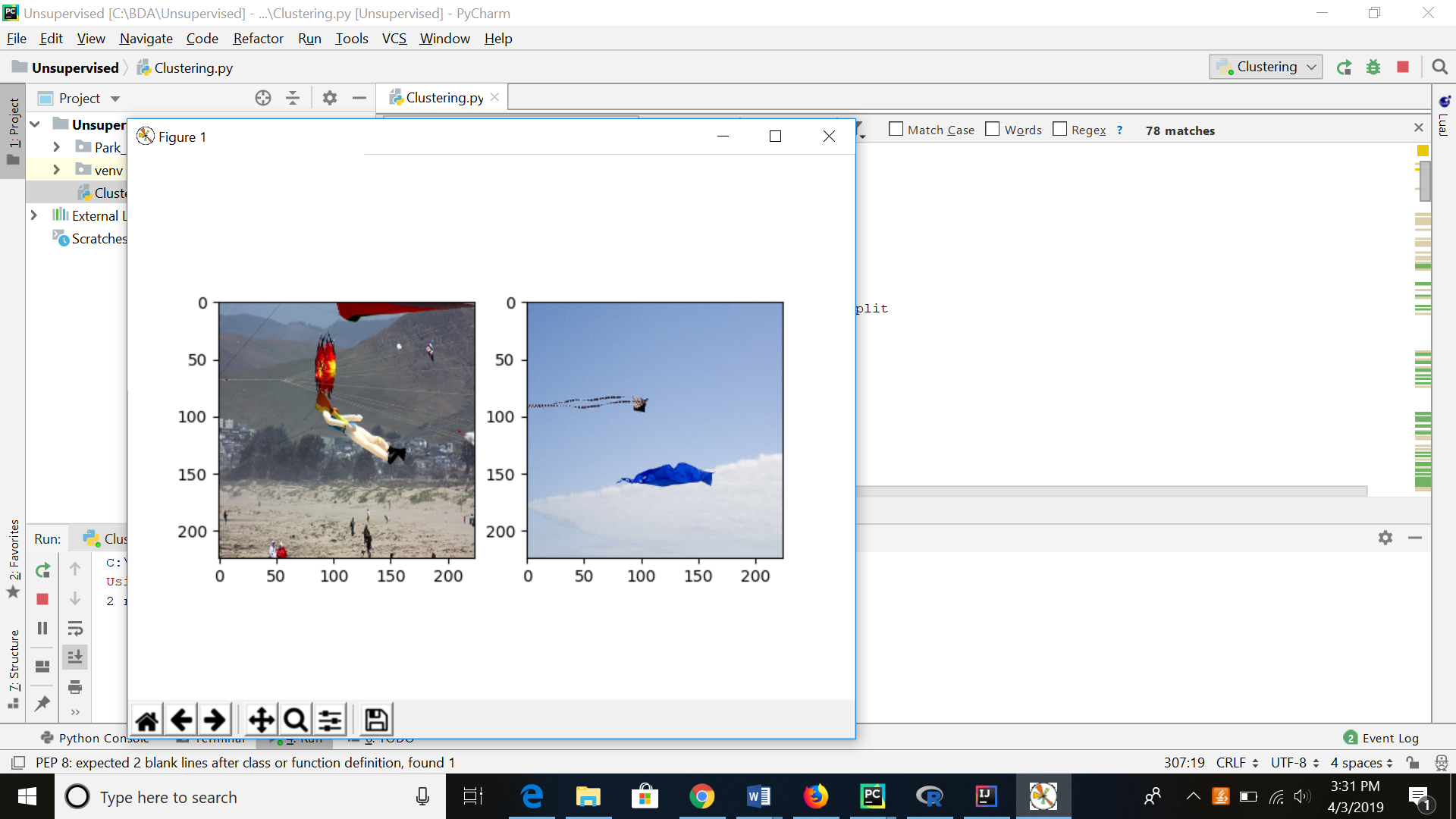
**Unsupervised Learning processing steps-**

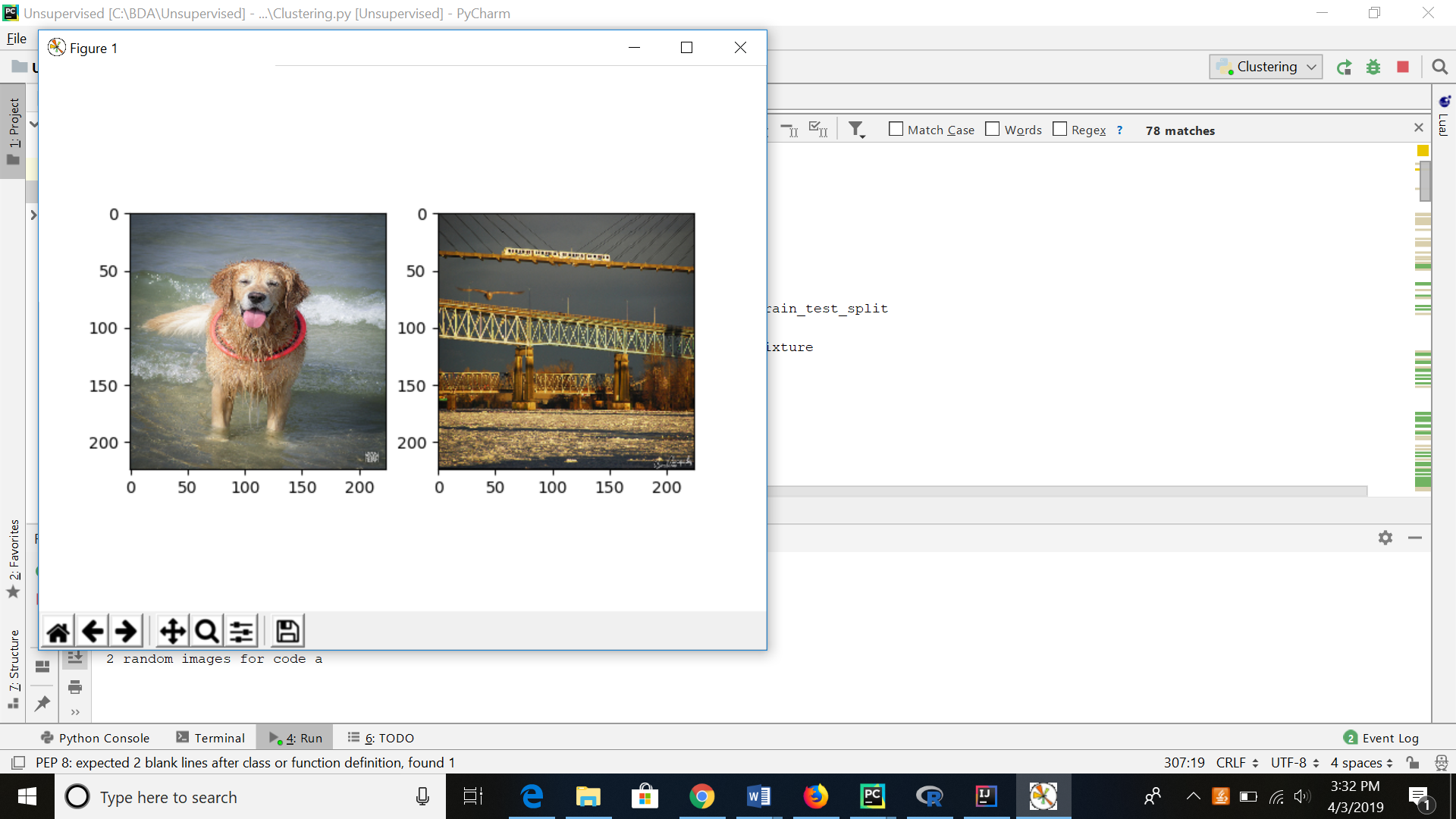
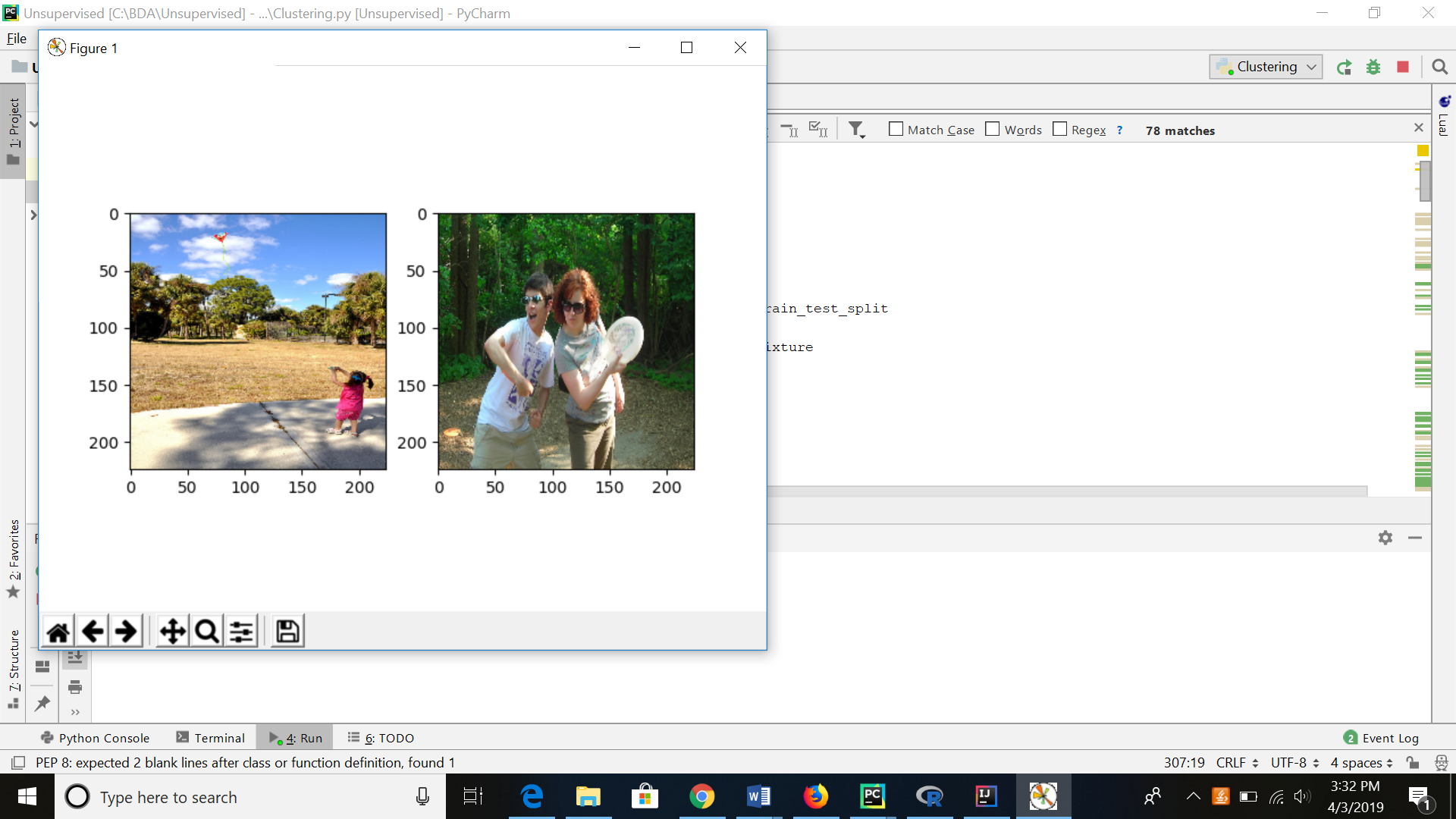
**Step1-** Load images that are in folders and get codes for these set of images, let code be a,b,c,d

**Step2-** Generating a few random images from all the 4 set of folders

**Step3-** Normalizing image pixels by 255

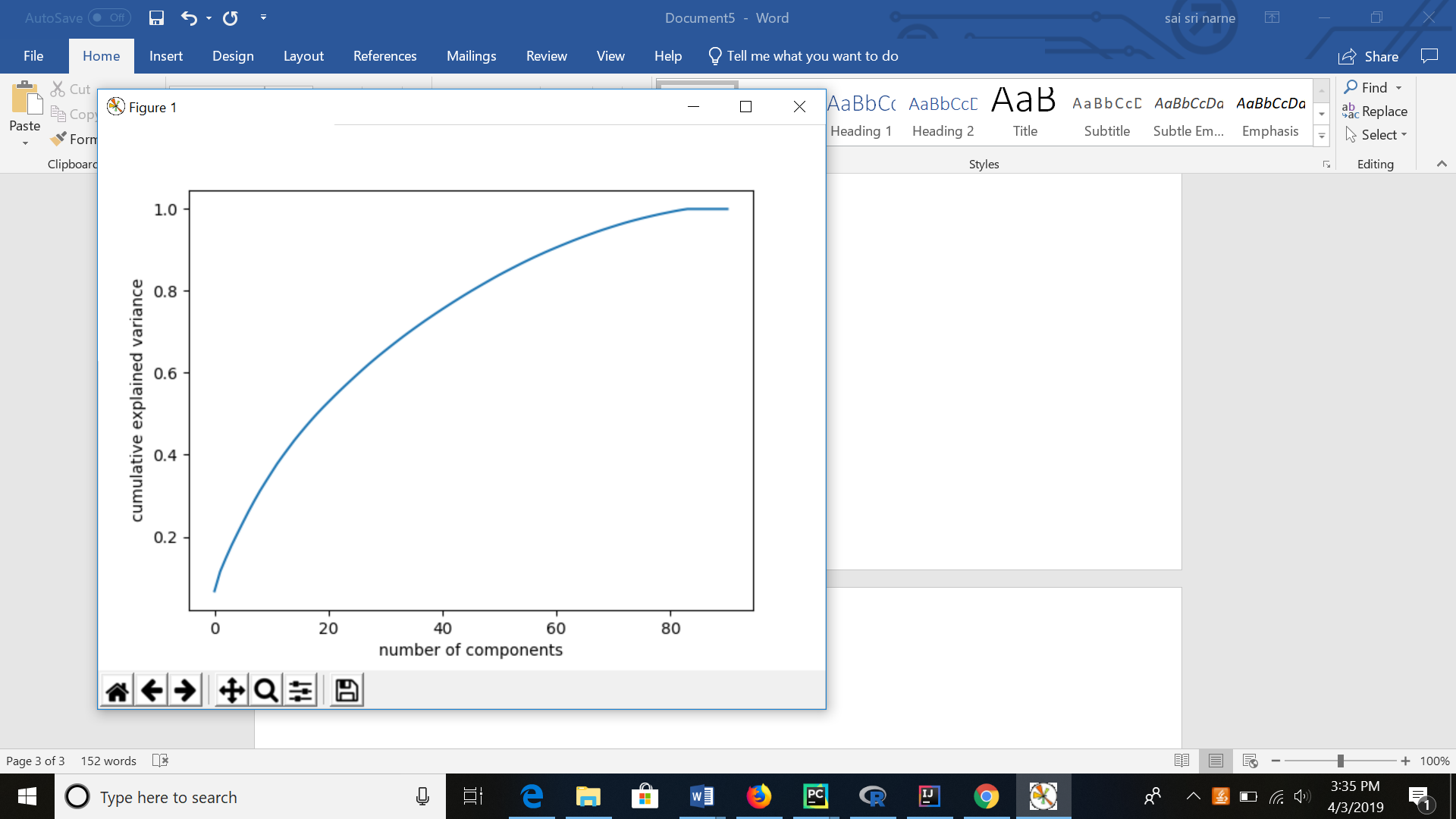
**Step4-** Shuffle data take only training images to make other images to fall under a cluster. Random images from three folders are-



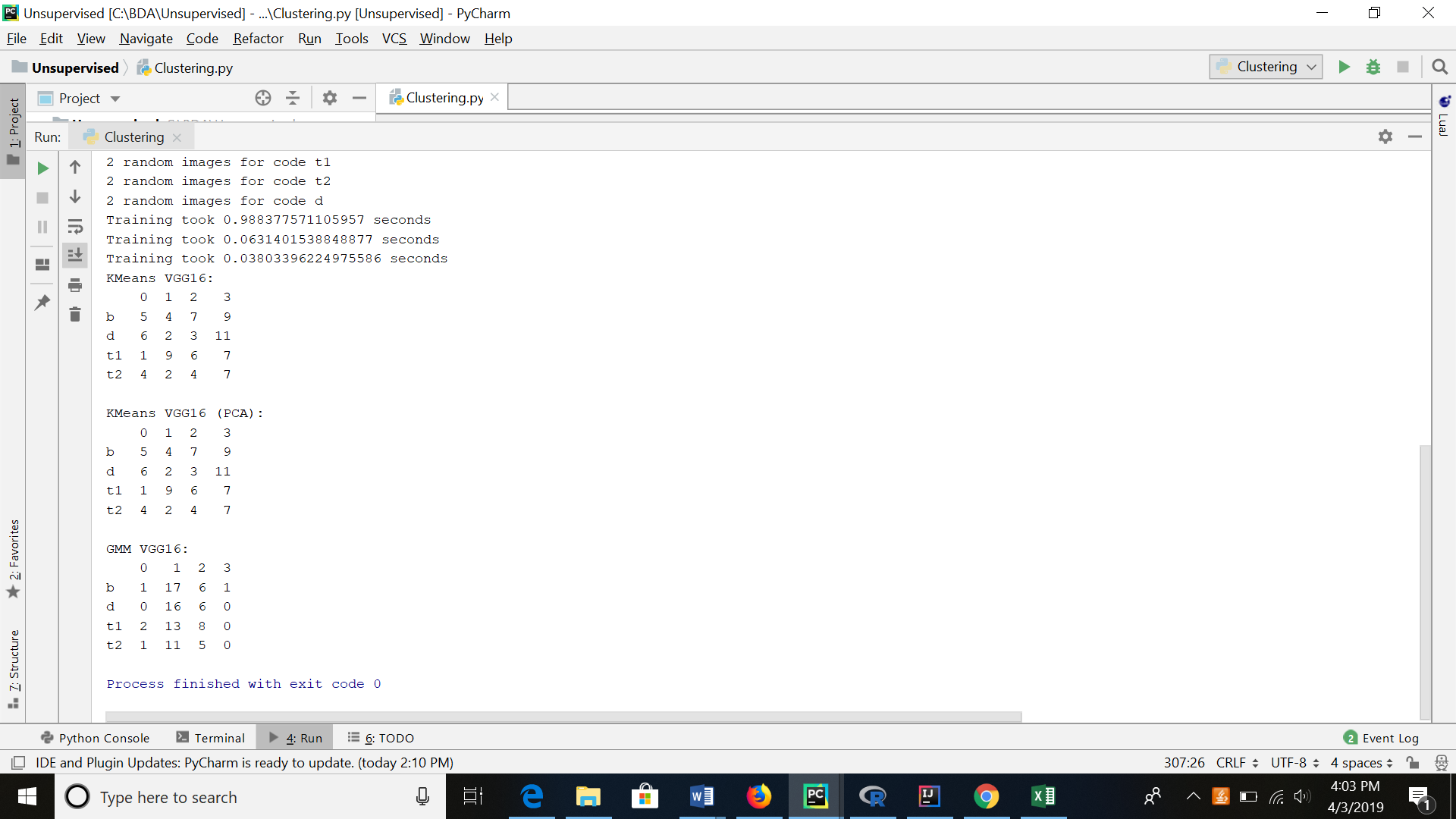


Step5- Pass the data to vgg16 pretrained model inherited from keras. Get the flatted image features.

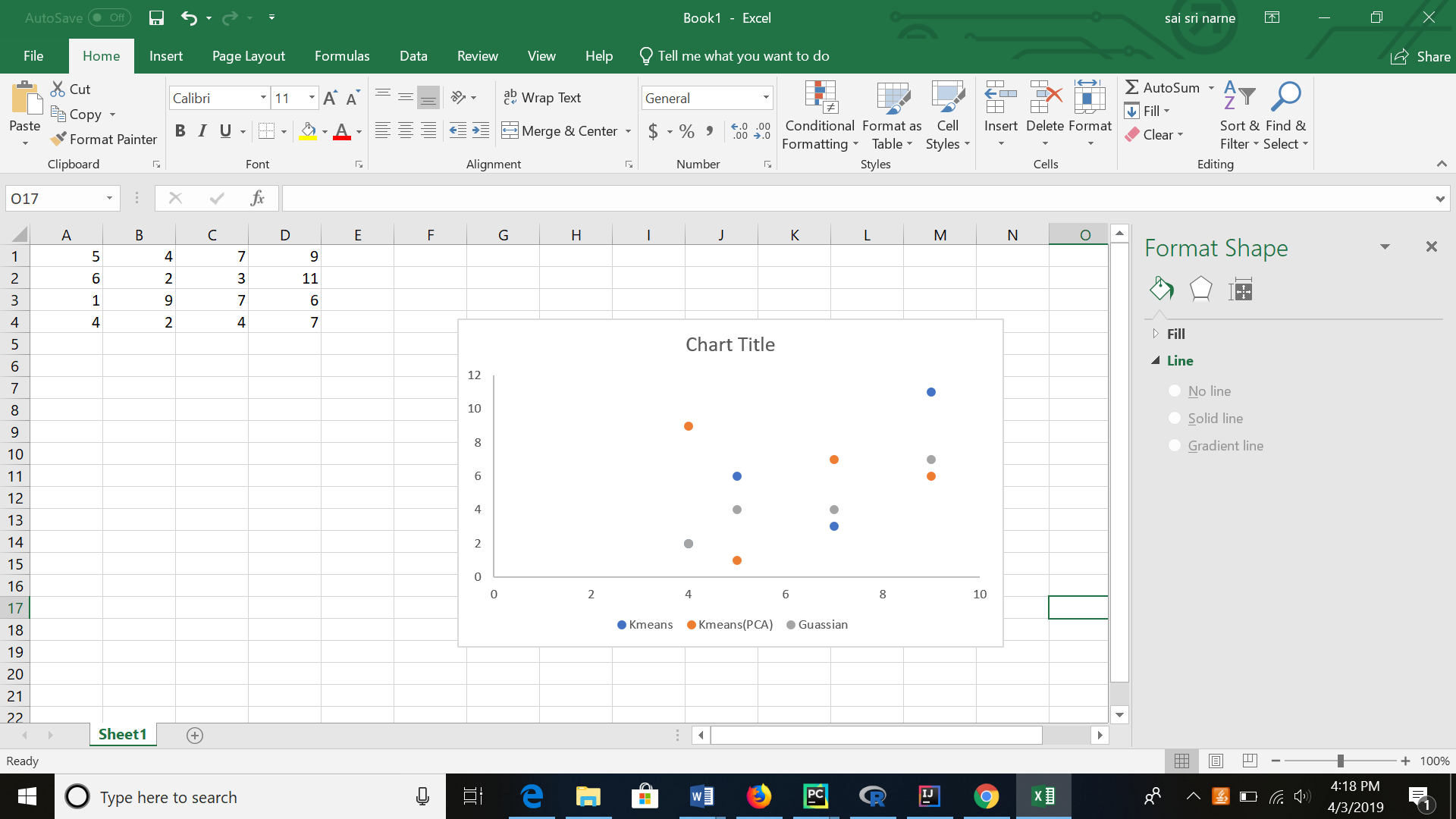
Step6- This input features are passed as input to perform PCA of Kmeans and PCA of Guassian (GMM)



No. of images from testing data set under cluster is shown in matrix format-



Visualization-



Link to code-

<https://github.com/SaiSriNarne/BigDataAnalytics/tree/master/Unsupervised>