K-Mean Clustering

- **Objective**

  To implement k-means clustering and analyse the clusters formed for various values of k. Display the centroids of the clusters.

- **Procedure**

  1. Choose a known dataset to analyse the results obtained from the implementation.
  2. Initialize the k number of centroids randomly and well spread in the given space.
  3. Assign all the examples an appropriate cluster i.e according to the minimum distance between points and the centroids.
  4. Re-calculate the centroids according to the distribution of all the points in space to the respective cluster(By taking the mean of all points in a cluster).
  5. Repeat step 3 and 4 repetitively till the maximum number of iterations or till the position of the centroids doesn't change in the subsequent iteration.
  6. Observe the final position of the centroids and the overall distortion involved.
  7. Repeat all the steps for different values of k.

- **Observations**

  ➢ **Dataset Parameters**

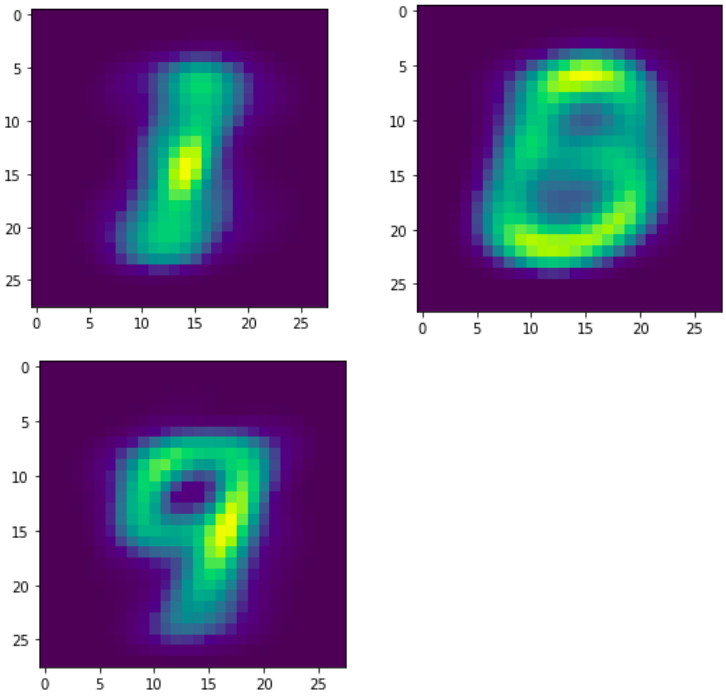
  - Number of classes = 10, the dataset contains the images of handwritten digits from 0 to 9.
  - There are a total 60,000 images in the dataset.
  - The distribution of images for each digit is as follows:

Digit	Number Of Images
0	5923
1	6742
2	5958
3	6131

4	5842
5	5421
6	5918
7	6265
8	5851
9	5949

➢ **K = 3**

- The Images Obtained for the centroids are as shown below:



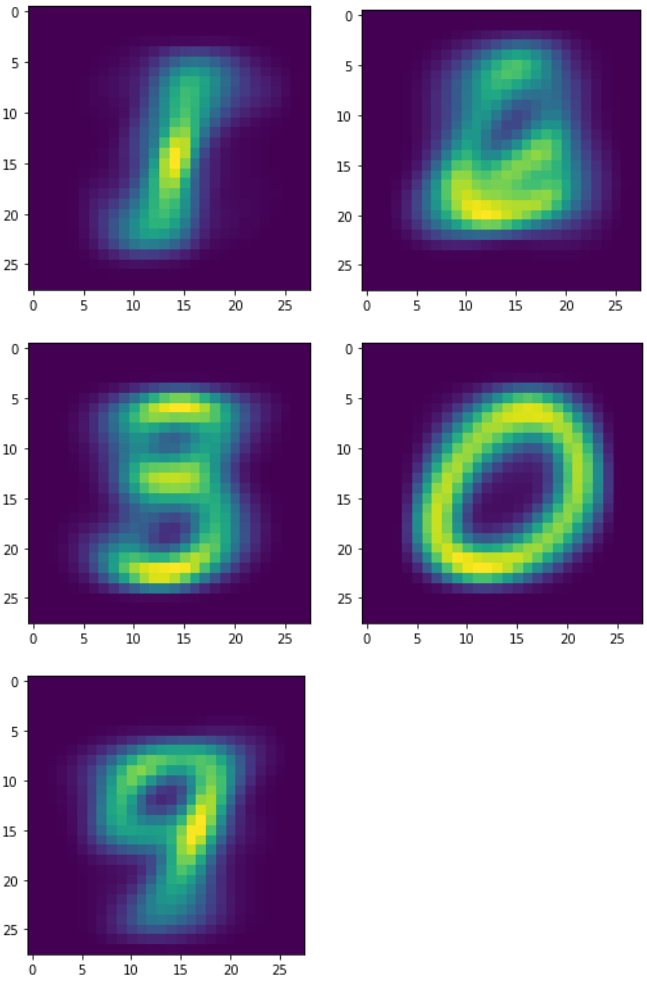
- The distribution of the examples in each cluster is as follows:

```
The number of images in cluster 0 is:  21180
The number of images in cluster 1 is:  19412
The number of images in cluster 2 is:  19408
```

- The number of iterations required to converge the model is `39`.

➢ **K = 5**

- The Images Obtained for the centroids are as shown below:



- The distribution of the examples in each cluster is as follows:

```
The number of images in cluster 0 is:  13902
The number of images in cluster 1 is:  10799
The number of images in cluster 2 is:  12573
The number of images in cluster 3 is:  5417
The number of images in cluster 4 is:  17309
```

- The number of iterations required to converge the model is `44`.

➢ **K = 7**

- The distribution of the examples in each cluster is as follows:

```
The number of images in cluster 0 is:  9166
The number of images in cluster 1 is:  10958
The number of images in cluster 2 is:  9810
The number of images in cluster 3 is:  8428
The number of images in cluster 4 is:  5068
The number of images in cluster 5 is:  7599
The number of images in cluster 6 is:  8971
```

- The number of iterations required to converge the model is `30`.

➢ **K = 10**

- The distribution of the examples in each cluster is as follows:

```
The number of images in cluster 0 is:  5446
The number of images in cluster 1 is:  7205
The number of images in cluster 2 is:  4495
The number of images in cluster 3 is:  10057
The number of images in cluster 4 is:  7444
The number of images in cluster 5 is:  5321
The number of images in cluster 6 is:  4514
The number of images in cluster 7 is:  6482
The number of images in cluster 8 is:  4072
The number of images in cluster 9 is:  4964
```

- The number of iterations required to converge the model is equal to the maximum iteration allowed = 100.

➢ **K= 12**

- The distribution of the examples in each cluster is as follows:

```
The number of images in cluster 0 is:  5433
The number of images in cluster 1 is:  5291
The number of images in cluster 2 is:  6601
The number of images in cluster 3 is:  2954
The number of images in cluster 4 is:  4732
The number of images in cluster 5 is:  5282
The number of images in cluster 6 is:  5198
```

```
The number of images in cluster 7 is:  6952
The number of images in cluster 8 is:  5548
The number of images in cluster 9 is:  2769
The number of images in cluster 10 is:  4504
The number of images in cluster 11 is:  4736
```

- The number of iterations required to converge the model is equal to the maximum iteration allowed = 100.

❖ Not showing the images of centroids for k = 7, 10, 12 as value is large

➢ **Variations Of Distortion With k**

- The Overall distortion can be calculated by measuring the distance between the points and centroid within a single cluster and then adding it for all clusters.

- The Distortion versus number of cluster is shown below:

- **Conclusion**
    - The convergence of the model depends upon the initialization of the centroids which is a random process.There is no relationship between the iterations required and the number of clusters to be formed.
    - Since the dataset is known, It can be observed that the formation of the clusters is around the classes of the examples.For e.g the most recurring centroid is the image of digit 9.