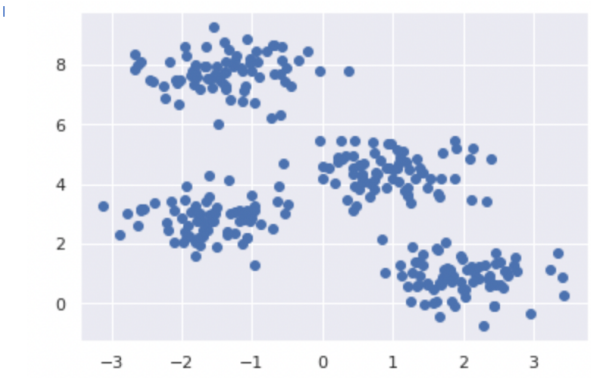
Mini Project – K-means Clustering and Principal Component Analysis(PCA)

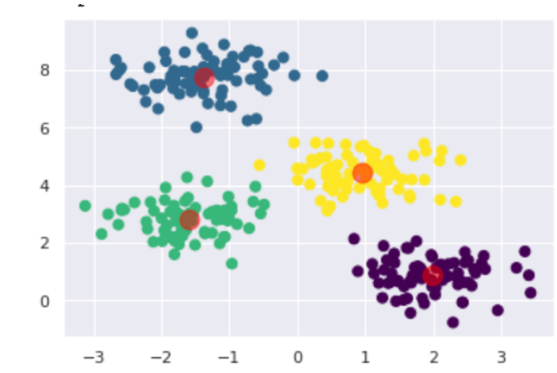
Part I – K-means clustering

Instruction on how to run my code:
Please run this project on Google Collab and use the file named "**clustering_ukm202.ipynb**". Go to Runtime and click on Run all and you will see the result.

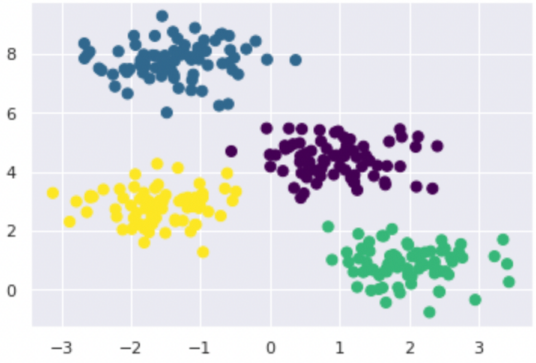
Result:



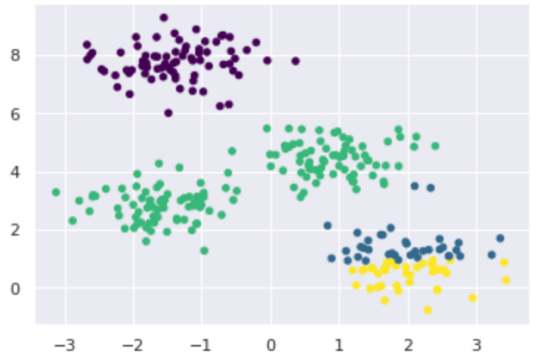
1.Manually generated data – Original dataset Plotted in four clusters



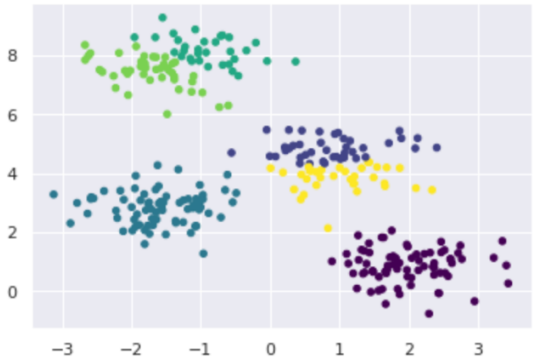
2.Visualizing the clusters with its cluster centers which was determined by k-means estimator



3. Fitted the model on the dataset and plotted the figure with default seed which is 2



4. Fitted the model on the dataset and plotted the figure with seed=0



5. Fitted the model on the dataset and plotted the figure with seed=2 and clusters=6

6. Implemented the K-means++ function, fitted the model on the dataset and plotted the figure with seed= 0

7. Implemented the K-means++ function, fitted the model on the dataset and plotted the figure with the default seed=2

Result Analysis:

First of all, answering the requirements 6 and 10. Yes, there is difference in the result when different seed values are passed. This is because k-means is sensitive to the starting conditions and since the seed generator generates the completely different random values as the start point of the algorithm, the clusters centers generally go far off from the points and in such case, there is chances that more than one cluster could be linked to a single centroid or more than one center can fall into a single cluster.

Now, answering the requirements 11. When clustered were made using k means and k mean++, we can observe that the quality of clustered is seemed to be improved and initialization of centroids looks better than what was observed in k means.

Part II – Principal Component Analysis

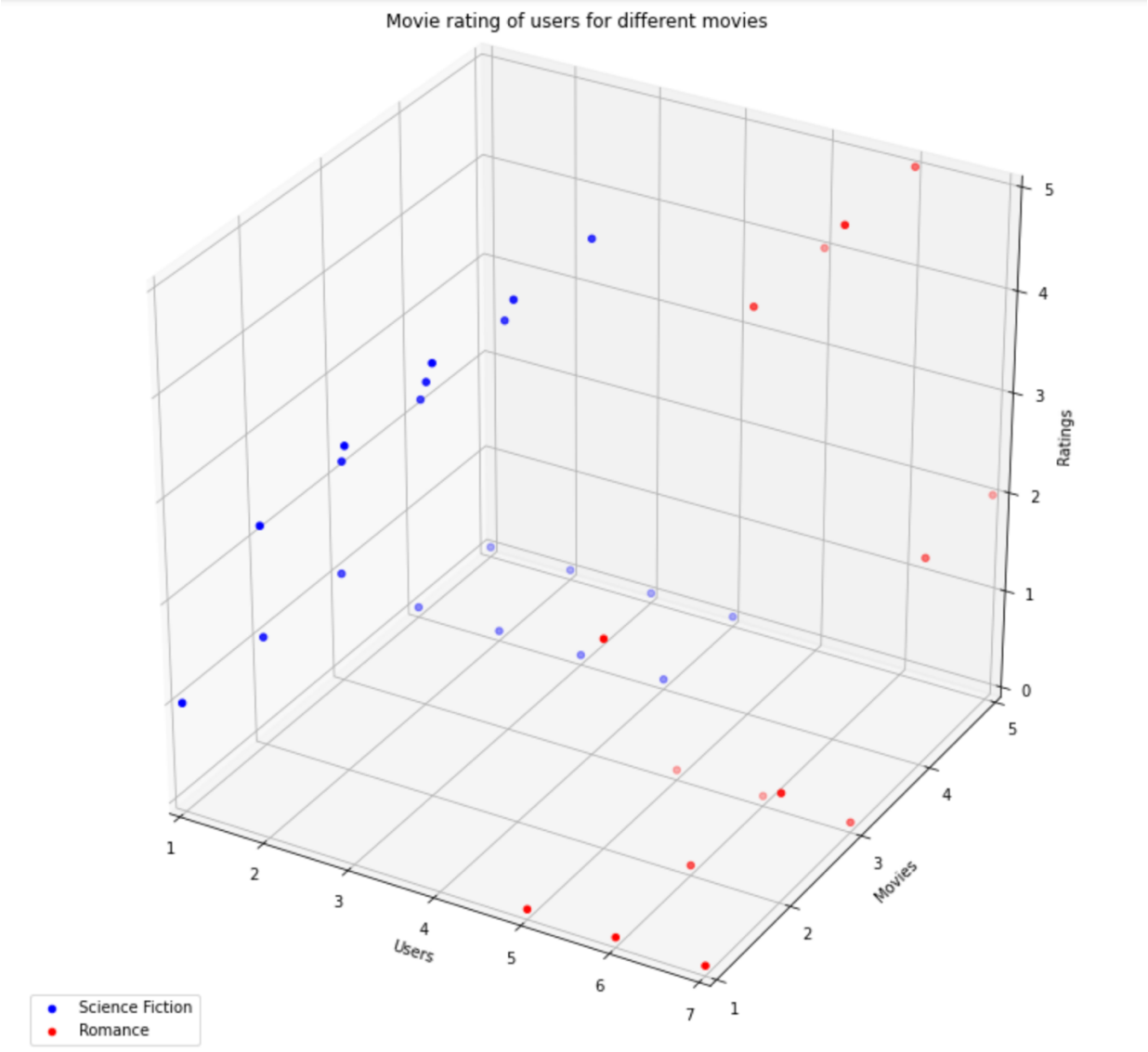
Instruction on how to run my code:

Please run this project on Google Collab and use the file named "**PCA_ukm202.ipynb**". Go to Runtime and click on Run all and you will see the result.

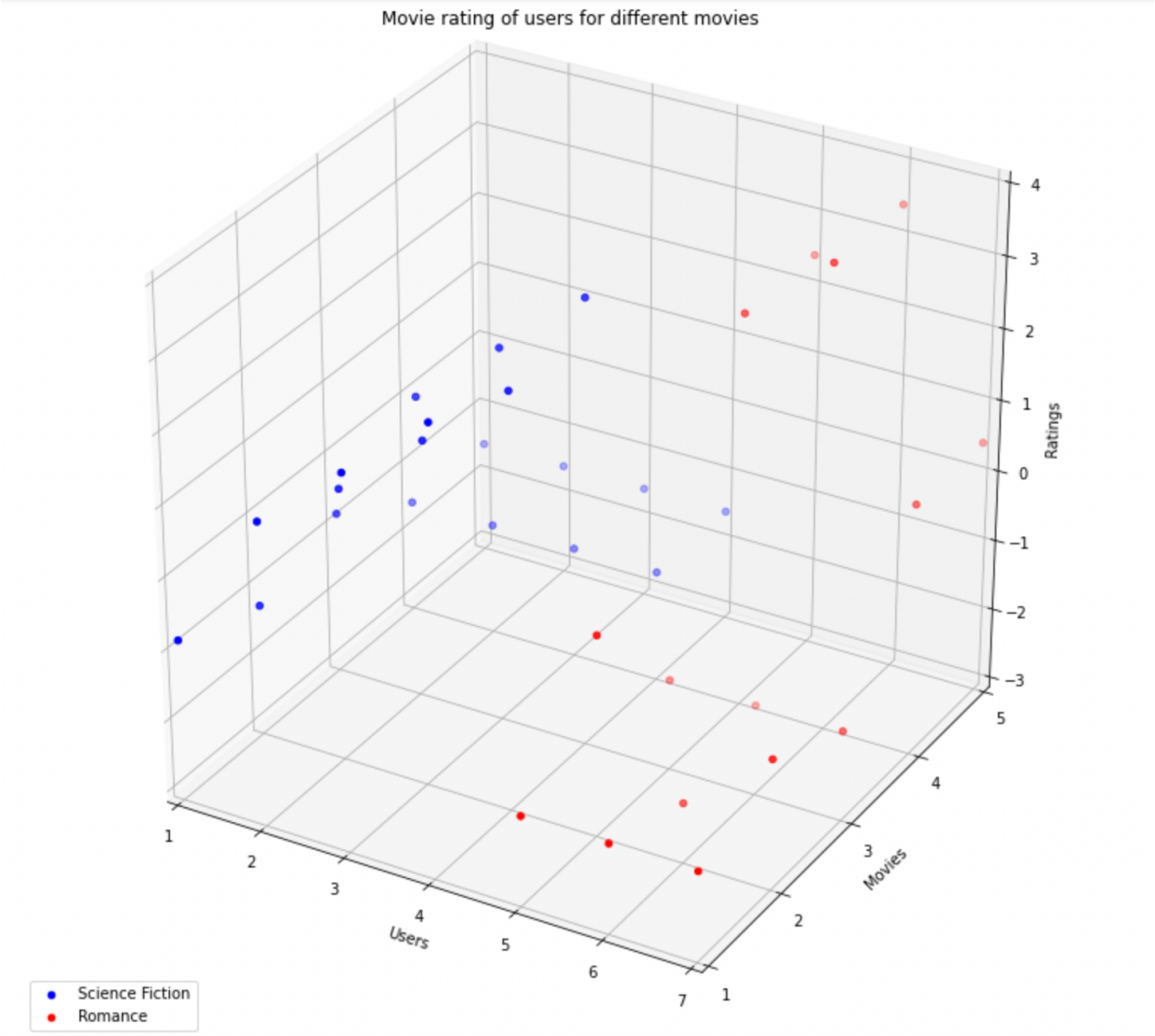
Result:

This part has 2 tasks.

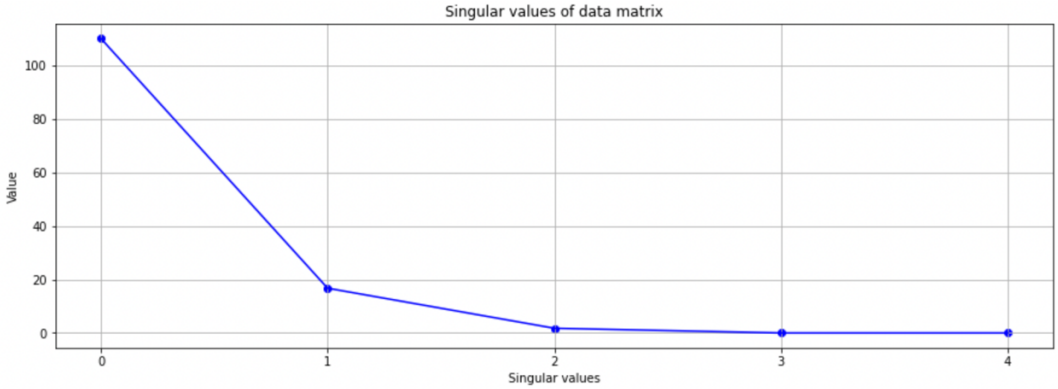
Task-1 Users to Movies



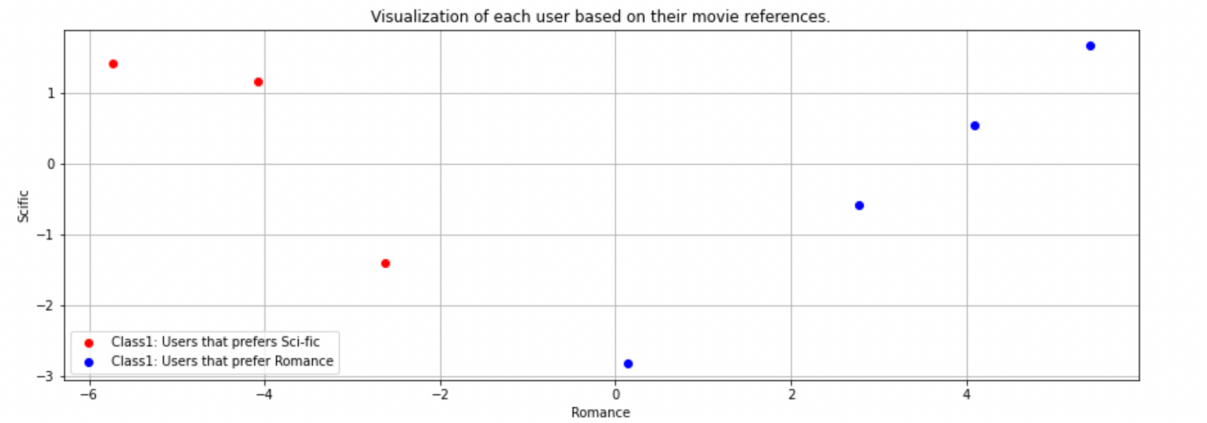
1.The Original Dataset



2.The Dataset Centered at 0



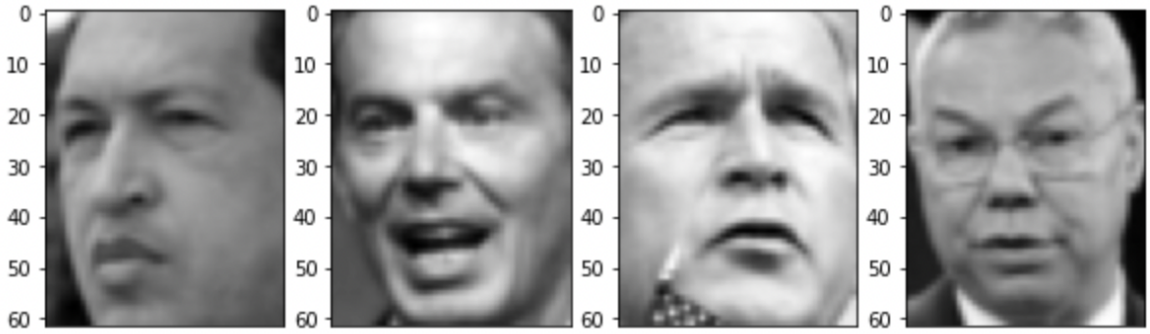
3.Singular values of data matrix



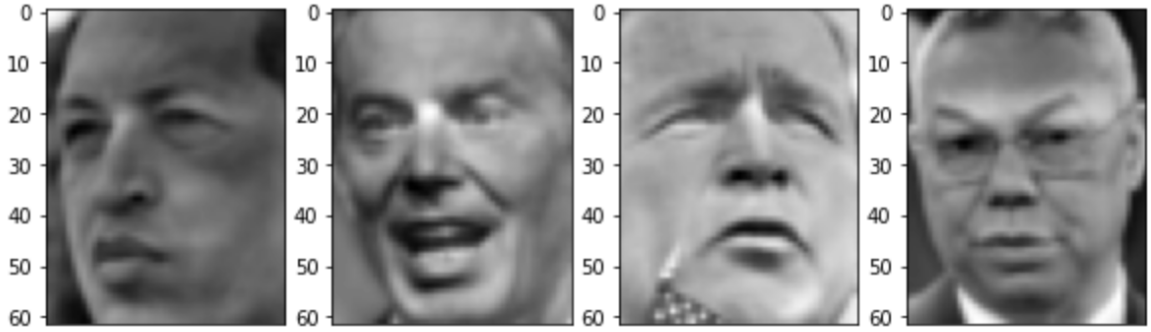
4. Visualization of each user based on their movie references

Task-1 Result Analysis:

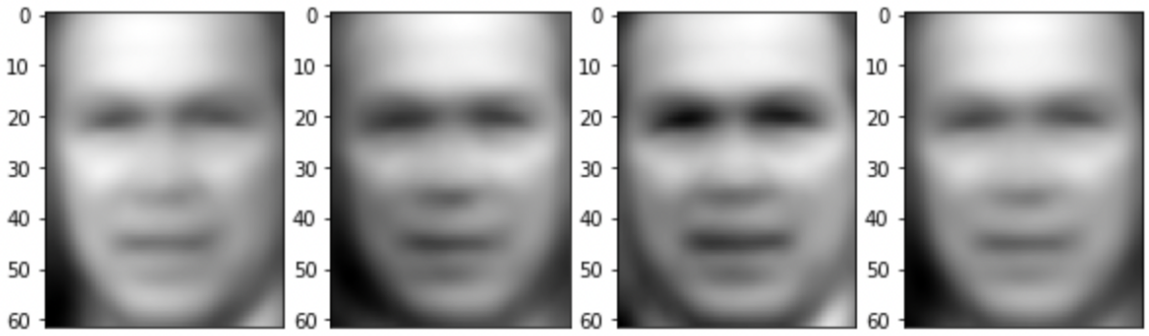
Task-2 Human Faces



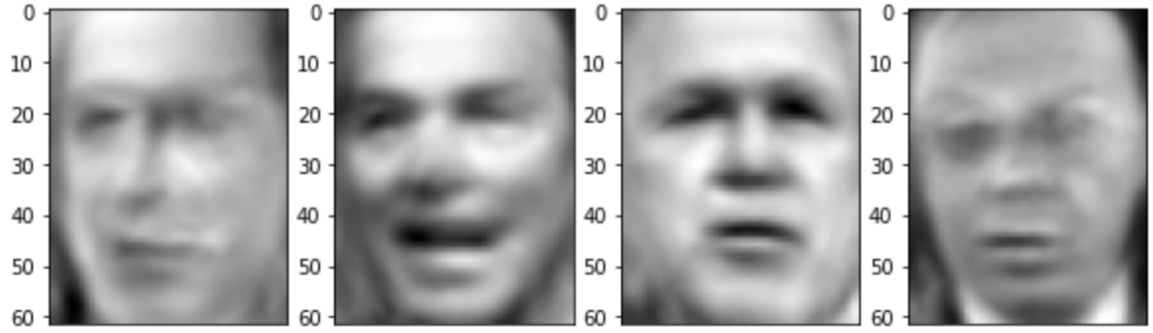
1.The Original Images



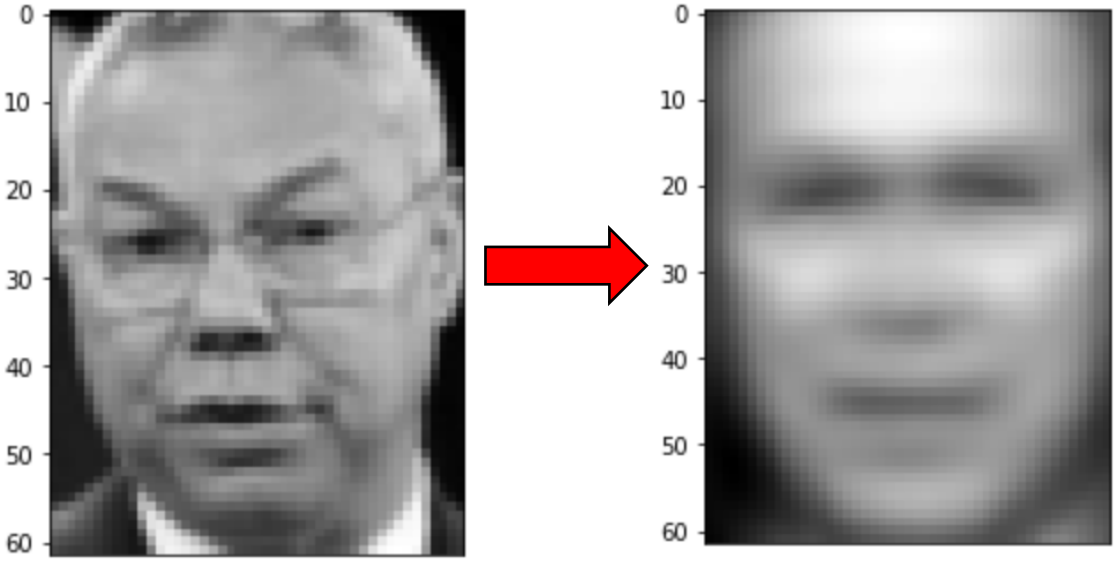
2.Images with mean removed from all data



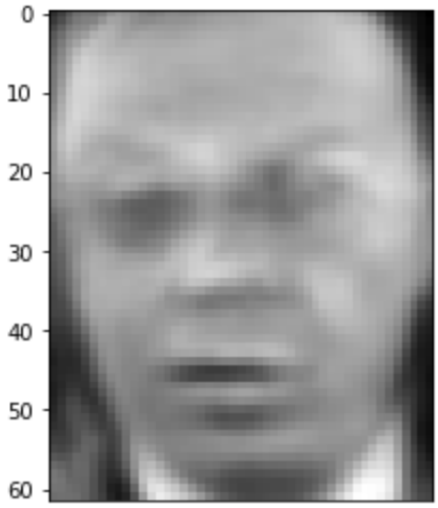
3.Images after using top 5 PCA



4.Images after using top 50 PCA



Original 4th Image, 4th Images after using top 5 PCA



4th Image after using top 50 PCA