Unsupervised Learning

1- Comparison of the error with different Q matrix:

- With num_compenents=40 , the average reconstruction error is = 8.
- With num_compenents=30 , the average reconstruction error is = 10.
- With num_compenents=20 , the average reconstruction error is = 13.
- With num_compenents=9 , the average reconstruction error is = 17.
- With num_compenents=3 , the average reconstruction error is = 18.

Significant Dimensionality Reduction: Reducing from 64 to 3 features is a drastic reduction in dimensionality. This can be very advantageous in aspects like: **simplifying the model**, **reducing computational costs**, and **enabling visualization in 3D space.**

Average Error of 18%: An average error of 18% means that the three principal components retain 82% of the variance in the original dataset. This is a substantial amount of information being captured, considering the level of dimensionality reduction.

The best num_components is equal to 3 as we reduced the number of features from 64 to 3, and the error is acceptable (=18).

2- *Comparison of DB scan Before PCA and after PCA:*

I. Before PCA:

1-Computational Complexity: The computational complexity increases with the number of dimensions. DBSCAN's performance can degrade in high-dimensional spaces, both in terms of speed and memory usage.

2-Noise Sensitivity: In its original form, DBSCAN is sensitive to noise, and high-dimensional data often contains more noise and sparsity, potentially leading to many points being labeled as noise. **As the resulted noise is = 301**

II. After PCA:

1- Improved Performance: Lower dimensions can significantly improve the computational efficiency of DBSCAN, both in terms of speed and memory requirements.

2- Noise Reduction: PCA can help in filtering out noise and irrelevant features, as it focuses on the directions with the most variance.

As the resulted noise is = 0 which is less than before the PCA

Clustering Efficiency: Utilizing dimensionality reduction, particularly PCA, prior to applying a clustering algorithm like DBSCAN, can be highly advantageous. This approach effectively mitigates the challenges posed by the curse of dimensionality “a common issue in high-dimensional data” that often degrades the performance of clustering algorithms.