**Analysis Report: (Question 1) PCA on MNIST Dataset**

**Author: Ujjwal Poudel**

**1. Objective & Data Preparation**

The objective was to apply Principal Component Analysis (PCA) to the mnist\_748 dataset to visualize its structure and perform efficient dimensionality reduction. The dataset, containing 70,000 images, was loaded into a Pandas DataFrame named df\_ujjwal with a final shape of (70000, 785). A 2x5 grid of sample digits was plotted to visually inspect the data.

A number with text and numbers

AI-generated content may be incorrect.

**2. PCA for 2D Visualization (2 Components)**

The 784 pixel features were reduced to 2 principal components for visualization.

Key Findings:

**Total Explained Variance**: 9.68%

The two components captured less than a majority of the variance, which is expected for such a drastic simplification of complex image data.

The resulting 2D scatter plot showed some distinct clusters (e.g., '1') but significant overlap for digits, confirming that 2 components are not sufficient to fully separate the classes.

A graph showing a number of colored dots

AI-generated content may be incorrect.

**3. Dimensionality Reduction (154 Components)**

IncrementalPCA was used to efficiently reduce the dataset to 154 components.

**Key Findings:**

Justification: 154 components were chosen to retain **80% of the original data's variance**, providing a strong balance between reducing complexity and preserving information.

Shapes:

**Original: (70000, 784)**

**Reduced: (70000, 154)**

**Reconstructed: (70000, 784)**

**Reconstruction Analysis**: When the reduced data was reconstructed back to 784 dimensions, the resulting images were recognizable but blurrier than the originals. This blurriness is the visual representation of the 5% of information that was discarded.

A collage of numbers

AI-generated content may be incorrect.

**4. Challenges & Conclusion**

**Challenges**: Initial challenges included loading the data into a DataFrame (solved using the .frame attribute) and understanding the batch-processing logic of IncrementalPCA (solved using np.array\_split and .partial\_fit).

**Conclusion**: PCA is a powerful tool for both visualization and dimensionality reduction. This analysis demonstrates the direct trade-off between the number of dimensions and the amount of information retained, where reducing dimensions for efficiency results in a measurable and visible loss of detail.