PCA

**Task 1: Exploratory Data Analysis (EDA):**

1. Load the dataset and perform basic data exploration.
2. Examine the distribution of features using histograms, box plots, or density plots.
3. Investigate correlations between features to understand relationships within the data.

**Task 2: Dimensionality Reduction with PCA:**

1. Standardize the features to ensure they have a mean of 0 and a standard deviation of Implement PCA to reduce the dimensionality of the dataset.
2. Determine the optimal number of principal components using techniques like scree plot or cumulative explained variance.
3. Transform the original dataset into the principal components.

**Task 3: Clustering with Original Data:**

1. Apply a clustering algorithm (e.g., K-means) to the original dataset.
2. Visualize the clustering results using appropriate plots.
3. Evaluate the clustering performance using metrics such as silhouette score or Davies–Bouldin index.

**Task 4: Clustering with PCA Data:**

1. Apply the same clustering algorithm to the PCA-transformed dataset.
2. Visualize the clustering results obtained from PCA-transformed data.
3. Compare the clustering results from PCA-transformed data with those from the original dataset.

**Task 5: Comparison and Analysis:**

1. Compare the clustering results obtained from the original dataset and PCA-transformed data.
2. Discuss any similarities or differences observed in the clustering results.
3. Reflect on the impact of dimensionality reduction on clustering performance.
4. Analyze the trade-offs between using PCA and clustering directly on the original dataset.

**Task 6: Conclusion and Insights**

1. Summarize the key findings and insights from the assignment.
2. Discuss the practical implications of using PCA and clustering in data analysis.
3. Provide recommendations for when to use each technique based on the analysis conducted.

ANSWER:

**Task 6**)

1. The study reveals that clustering results from both original and PCA-transformed data show similar shapes and distributions, but slight differences in visualization and evaluation metrics.
2. PCA reduces the dimensionality of high-dimensional datasets while preserving variance, making it easier to visualize and interpret complex data. Clustering directly on the original dataset provides cluster labels that are interpretable in the original feature space. Both techniques are useful in data analysis, with PCA for reducing computational complexity and improving visualization, while clustering helps identify inherent structures and patterns.
3. The choice of technique depends on the specific goals of the analysis and the dataset's characteristics.

**Collab file:**

<https://colab.research.google.com/drive/1tdEqO4bORcK0NuEXDR2TnkMuwloitluC?usp=sharing>