Here is a summary of the analysis:

Objective:

The primary objective was to **segment students into distinct groups or "clusters"** based on various academic and engagement features (study_hours, assignments_completed, attendance_percent, previous_score, and final_score). The analysis used **K-Means clustering** and its results were evaluated using the **Elbow Method** and **Silhouette Score** to determine the optimal number of clusters.

Dataset Source:

The analysis was performed on a **synthetic demo dataset** that was procedurally generated within the notebook itself, as the primary source file (student_data.csv) was not found.

The synthetic data generation created three distinct groups:

- **High-performing (35%):** High scores, study hours, assignments, and attendance.
- **Medium-performing (45%):** Medium values across all metrics.
- **Low-performing (20%):** Low values across all metrics.

Steps Followed:

1. **Setup and Data Loading:** Imported necessary libraries (NumPy, Pandas, Matplotlib, Seaborn, scikit-learn modules) and set a random state for reproducibility. A synthetic dataset was generated with a shape of (350, 5).
2. **Data Preprocessing:**
    - Identified numeric columns (study_hours, assignments_completed, attendance_percent, previous_score, final_score).
    - Handled missing values by filling them with the **median** of the respective column (though the synthetic data was complete).
    - **Scaled** the data using StandardScaler to ensure all features contribute equally to the clustering process.
3. **Dimensionality Reduction:** Applied **Principal Component Analysis (PCA)** to reduce the feature space to 2 principal components for visualization purposes.

    - The first two principal components (**PCA 1 and PCA 2**) explained approximately **83.5% and 4.96%** of the variance, respectively (Total $\approx 88.5\%$).

4. **Optimal K Determination (K-Means):**
    - Used the **Elbow Method** (plotting Inertia vs. *k*) and the **Silhouette Score** to find the best number of clusters (*k*).
    - The Silhouette Score method indicated that the **best *k* is 3**.
5. **K-Means Clustering:** Performed K-Means clustering with the optimal $k=3$ on the scaled data and assigned the resulting cluster labels (kmeans_label) back to the DataFrame.
6. **DBSCAN Experimentation:** Briefly explored **DBSCAN** clustering with various eps values, which resulted in 5, 4, and 2 clusters, or 0 clusters (all noise), depending on the eps value. The DBSCAN results were not used for the main analysis.
7. **Cluster Profiling and Visualization:**
    - Calculated the **mean** of each numeric feature for every cluster to create a cluster summary/profile.
    - Visualized the clusters in the 2D PCA space using a scatter plot.
    - Visualized the distribution of final_score across the clusters using a box plot.
    - Visualized the size of each cluster using a count plot.

Tools Used:

Type	Tool	Purpose
Programming/Data	Python (Pandas, NumPy)	Data manipulation, data generation, and numerical operations.
Clustering	KMeans (sklearn.cluster)	Performed the primary clustering analysis.
Scaling	StandardScaler (sklearn.preprocessing)	Standardized features for clustering.

Type	Tool	Purpose
Dimensionality Reduction	PCA (sklearn.decomposition)	Reduced data to 2 dimensions for visualization.
Evaluation	silhouette_score (sklearn.metrics)	Determined the optimal number of clusters (*k*).
Visualization	matplotlib.pyplot, seaborn	Created scatter, bar, box, and count plots for cluster analysis.
Exploratory Clustering	DBSCAN (sklearn.cluster), NearestNeighbors (sklearn.neighbors)	Explored an alternative clustering algorithm.

Key Insights and Graphs

The K-Means clustering identified **3 distinct student segments**, which strongly correspond to the synthetic **low, medium, and high-performing** groups.

Cluster Summary:

The table below shows the average values for each feature within the three identified clusters, providing clear profiles:

kmeans_label	study_hours (Avg.)	assignments_completed (Avg.)	attendance_percent (Avg.)	previous_score (Avg.)	final_score (Avg.)	count (Size)
0 (Low)	**1.62**	**2.70**	**63.54**	**41.40**	**46.19**	70

kmeans_label	study_hours (Avg.)	assignments_completed (Avg.)	attendance_percent (Avg.)	previous_score (Avg.)	final_score (Avg.)	count (Size)
1 (High)	**6.44**	**9.08**	**92.11**	**77.92**	**84.69**	122
2 (Medium)	**3.47**	**6.25**	**80.36**	**60.99**	**64.90**	157

Insights

- **Cluster 0 ("Low Performers"):** This is the smallest group (70 students) and is characterized by the lowest mean values across **all metrics**, including the poorest **attendance** and **final scores** (average 46.19).
- **Cluster 1 ("High Performers"):** This group (122 students) exhibits the highest engagement and performance, with averages well above 84 for **final score** and **attendance**, and over 6 hours of **study hours**.
- **Cluster 2 ("Medium Performers"):** This is the largest group (157 students) with intermediate scores across all features, showing moderate **study habits** (average 3.47 hours) and **final scores** (average 64.90).

GRAPHS:

K-Means Clusters (PCA view)

This scatter plot shows the distinct separation of the three clusters when the data is projected onto the first two principal components. The separation is clean, suggesting a good clustering result.

Cluster Feature Means (K-Means)

This bar chart visually confirms the three distinct profiles, showing a clear hierarchy for all features from Cluster 0 (lowest) to Cluster 1 (highest).

Final Score Distribution by Cluster

The box plot illustrates the final score distribution for each cluster. The boxes are well-separated with minimal overlap, confirming that the clustering successfully segmented students based on their final academic performance.

Cluster Size Distribution:

This count plot shows the number of students assigned to each cluster. Cluster 2 (Medium) is the largest, followed by Cluster 1 (High), and Cluster 0 (Low).