Unsupervised Learning Summary

This analysis aimed to explore patterns in college basketball team performance using unsupervised learning techniques. The dataset contained metrics related to offensive rating (OR), defensive rating (DR), and adjusted tempo (AT) for both home and away teams, along with their deltas. The study applied dimensionality reduction (DR) and clustering techniques to uncover latent structures within the data.

**Methods**

**Dimensionality Reduction (DR) Techniques:**

Five DR methods were tested:

1. **Principal Component Analysis (PCA)** – Linearly reduces dimensions while preserving variance.
2. **t-Distributed Stochastic Neighbor Embedding (t-SNE)** – Captures nonlinear relationships.
3. **Uniform Manifold Approximation and Projection (UMAP)** – Preserves local and global structures.
4. **Locally Linear Embedding (LLE)** – Captures local neighborhood structure.
5. **Independent Component Analysis (ICA)** – Extracts statistically independent features.

**Clustering Techniques:**

Three clustering methods were applied:

1. **K-Means** – Partitions data into k clusters based on centroids.
2. **Agglomerative Clustering** – Hierarchical clustering based on merging points iteratively.
3. **DBSCAN** – Density-based clustering to detect arbitrary-shaped clusters and noise.

**Results and Interpretations**

**PCA + K-Means**

* PCA reduced the dataset to two principal components, capturing most variance.
* K-Means assigned teams into four clusters, suggesting distinct styles of play based on team performance.
* The clustering boundary shows clear differentiation, with some overlap, indicating some teams have hybrid playing styles.

**t-SNE + Agglomerative Clustering**

* t-SNE effectively spread data points, revealing nonlinear relationships.
* Agglomerative clustering identified four clusters, potentially reflecting different playstyles (e.g., fast-paced vs. defensive-oriented teams).
* Clusters are more evenly distributed compared to PCA, suggesting better separation.

**UMAP + DBSCAN**

* UMAP retained both local and global structure, showing more organic groupings.
* DBSCAN primarily grouped most teams into a single cluster (label 0) with a few noise points (-1), indicating that many teams share similar performance metrics, with only a few outliers.
* This result suggests DBSCAN may not be optimal for this dataset, as basketball team performance data may not have distinct density-based clusters.

**Cluster Distributions**

* **DBSCAN:** Most teams fell into one cluster, with only a few outliers.
* **K-Means:** Produced four well-balanced clusters, suggesting meaningful segmentation.
* **Agglomerative:** Showed slightly uneven distributions but still reasonable separation.

**Conclusion**

* **Best DR-Clustering Pairing:** t-SNE with Agglomerative Clustering provided the most distinct and interpretable clusters.
* **PCA & K-Means** performed well but had overlapping regions.
* **UMAP & DBSCAN** struggled to separate teams, suggesting DBSCAN may not be ideal for this dataset.
* The clusters likely correspond to different team strategies, such as offensive vs. defensive dominance, high vs. low tempo, or balanced vs. extreme playstyles.