FIFA 2018 Player Clustering Report

1. Data Preparation

- **Dataset:** FIFA 18 sample data with player attributes.
- **Preprocessing steps:**
    1. Dropped irrelevant columns: IDs, names, photos, clubs, flags, birth dates, body type, real face, work rates, preferred foot, and special traits.
    2. Removed goalkeeping stats (gk_*) and all position preference columns.
    3. Removed constant columns and columns with very high multicollinearity.
    4. Calculated **BMI** from height and weight, then dropped height and weight.
- **Scaling:** StandardScaler applied to all numerical features to normalize ranges.

---

2. Dimensionality Reduction

- **Method:** Principal Component Analysis (PCA).
- **Components:** First 3 principal components chosen for visualization and clustering.
- **Explained variance:**
    - First 3 components explain **62.48%** of the variance in the dataset.
- **Insight:** PCA effectively reduced dimensionality while retaining most of the information, making clustering more reliable.

---

3. Clustering

- **Algorithm:** K-Means.
- **Number of clusters:** 4 (determined via silhouette score).
- **Silhouette Score: 0.4155**
    - This score indicates **moderate clustering quality**. Clusters are reasonably distinct, though some overlap exists.

---

4. Cluster Interpretation

- After clustering, clusters were **mapped to positions** based on the centroids and player stats:

Cluster	Assigned Role
Cluster 0	CM (Midfielder)
Cluster 1	DEF (Defender)
Cluster 2	AT (Attacker)
Cluster 3	GK (Goalkeeper)

- **Cluster sizes:**
  - CM: 125
  - DEF: 333
  - AT: 429
  - GK: 113
- **Comparison to actual assigned positions:**

Assigned_Position	Count	Cluster Count
CM	450	125
DEF	302	333
AT	135	429
GK	113	113

Insight:

- Defenders and goalkeepers are clustered reasonably well.
- Attackers are overrepresented in clusters, while midfielders are underrepresented. This may be due to overlapping characteristics between CM and AT in PCA space.

---

5. Visualizations

- **3D Scatter plot:** Using the first 3 principal components, clusters are clearly separable in PCA space, providing a visual validation of the clustering.
- PCA axes help reduce complexity while keeping players with similar overall profiles close together.

---

6. Key Insights

1. **Cluster quality:** Moderate silhouette score indicates clusters capture structure but are not perfectly separated.
2. **Dimensionality reduction:** PCA retained 62% of variance in 3 dimensions—good tradeoff between simplicity and information retention.
3. **Position mapping:** Useful for interpreting player types but some overlap between attacking and midfield roles remains.
4. **Scaling & cleaning:** Removing irrelevant and multicollinear columns improved clustering performance.
5. **Next steps:**
    - Consider using **more PCA components** for better variance capture.
    - Try **other clustering algorithms** (e.g., Gaussian Mixture Models, Agglomerative Clustering) for improved silhouette score.
    - Include boolean prefers_* columns as 0/1 to potentially enhance clustering.