FIFA Player Segmentation: Identifying Distinct Player Profiles

1. Project Objectives

1.1. PO1

Segmentation of FIFA Player Data using K-means Clustering to identify distinct groups based on player attributes

1.2. PO2

Identification of the optimal number of clusters (k) in the FIFA Player data to ensure meaningful and interpretable groupings.

1.3. PO3

Determination of the characteristics of each cluster to understand the unique features and trends within each identified player cluster by analyzing how various player positions, attributes and skills differ between clusters.

2. Description of Data

2.1. Index Variables

The dataset did not have a natural identifier like an "ID" column, which could be set as the index for easier retrieval by specific identifiers.

Hence, an index variable was created bu using `player_id` without modifying the actual data itself to significantly improve the organization and interpretability of the data for further analysis and exploration.

> Refer Code Block (Cell)

2.2. **Categorical Variables (CV)**

2.2.1. Categorical Variables - Nominal Type

- `fifa_version`: The version of the FIFA game the data is associated with.
- `league_id`: A numerical identifier for the league in which the player is currently active.
- `club_team_id`: A numerical identifier for the club the player is currently playing for.
- `club_position`: The primary position the player occupies within their club.
- `nationality_id`: A numerical identifier for the player's nationality.
- `nation_team_id`: A numerical identifier for the national team the player represents (if applicable).
- `preferred_foot`: The player's preferred foot for kicking the ball (Left or Right).
- `body_type`: The general body shape or build of the player (e.g., Normal, Lean, Stocky).

2.2.2. **Categorical Variables - Ordinal Type**

- `league_level`: The hierarchical level of the league the player is in, implying a ranking or order of leagues.
- `weak_foot`: A rating (from 1 to 5) indicating the skill level of the player's weaker foot.
- `skill_moves`: A rating (from 1 to 5) representing the player's ability to perform technical moves or tricks.
- `international_reputation`: A rating (from 1 to 5) reflecting the player's renown and recognition on the international stage.
- `work_rate`: A categorical variable describing the player's work ethic and stamina both in attack and defense (e.g., High/Low, Medium/Medium).

2.3. **Non-Categorical Variables (NCV)**

- `fifa_update`: An integer likely representing an update or patch number within a FIFA version
- `overall`, `potential`: Overall and potential ratings of the player, indicating their current and future ability
- `value_eur`, `wage_eur`: The player's estimated market value and weekly wage in Euros
- `age`: The player's age in years.
- `height_cm`: The player's height in centimeters.
- `weight_kg`: The player's weight in kilograms
- `club_jersey_number`: The jersey number the player wears for their club
- `pace`, `shooting`, `passing`, `dribbling`, `defending`, `physic`: Core attributes representing the player's abilities in different aspects of the game
- `attacking_crossing`, `attacking_finishing`, `attacking_heading_accuracy`, `attacking_short_passing`, `attacking_volleys`: Specific attacking attributes
- `skill_dribbling`, `skill_curve`, `skill_fk_accuracy`, `skill_long_passing`, `skill_ball_control`: Skill-related attributes
- `movement_acceleration`, `movement_sprint_speed`, `movement_agility`, `movement_reactions`, `movement_balance`: Movement-related attributes
- `power_shot_power`, `power_jumping`, `power_stamina`, `power_strength`, `power_long_shots`: Power-related attributes
- `mentality_aggression`, `mentality_interceptions`, `mentality_positioning`, `mentality_vision`, `mentality_penalties`, `mentality_composure`: Mentality or psychological attributes.
- `defending_marking_awareness`, `defending_standing_tackle`, `defending_sliding_tackle`: Defensive attributes.
- `goalkeeping_diving`, `goalkeeping_handling`, `goalkeeping_kicking`, `goalkeeping_positioning`, `goalkeeping_reflexes`, `goalkeeping_speed`: Goalkeeping-specific attributes

Refer Code Block (Cell)

NOTE: Dropping irrelevant columns from the DataFrame 'df' to simplify the dataset not needed for analysis.

Refer Code Block (Cell)

3. Analysis of Data

3.1. Data Pre-Processing

3.1.1. Missing Data Statistics and Treatment

3.1.1.1. Missing Data Statistics: Records

The dataset is divided into two subsets based on categorical and non-categorical variables. The code calculates and prints the number of missing values for each variable and record.

Refer Code Block (Cell)

3.1.1.2. Missing Data Treatment: Records

Refer Code Block (Cell)

3.1.1.2.1. Imputation of Missing Data

Missing categorical data is imputed using the most frequent value, while missing non-categorical data is also imputed using the most frequent value.

Refer Code Block (Cell)

3.1.1.2.2. Removal of Records with More Than 50% Missing Data

Empty records and variables are excluded from the dataset.

Refer Code Block (Cell)

3.1.2. Numerical Encoding of Categorical Variables

Categorical data is encoded numerically using the Ordinal Encoder from scikit-learn.

Refer Code Block (Cell)

3.1.3. Outlier Statistics and Treatment

3.1.3.1. Outlier Treatment: Non-Categorical Variables

Non-categorical variables are normalized using Min-Max Scaler to handle outliers.

Refer Code Block (Cell)

3.2. Data Analysis

3.2.1. Assessment Criteria

1. **Silhouette Score (SS)**:
   - The Silhouette Score is a measure of how similar an object is to its own cluster compared to other clusters.

  - It quantifies the separation between clusters. A high Silhouette Score indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.
  - The Silhouette Score ranges from -1 to 1, where a high value indicates that the object is well-clustered, a value near 0 indicates overlapping clusters, and negative values suggest that the object may have been assigned to the wrong cluster.
  - In the context of K-means clustering, the average Silhouette Score across all data points can be used to evaluate the quality of clustering. Higher average Silhouette Scores indicate better-defined clusters.
2. **Davies-Bouldin Index (DBI)**:
   - The Davies-Bouldin Index is a measure of cluster compactness and separation.
   - It evaluates the average similarity between each cluster and its most similar cluster, weighted by the cluster sizes.
   - A lower DBI indicates better clustering, with clusters that are well-separated from each other and internally compact.
   - The DBI considers both intra-cluster and inter-cluster distances, aiming to minimize intra-cluster distance while maximizing inter-cluster distance.
   - Like the Silhouette Score, the Davies-Bouldin Index is used to assess the quality of clustering algorithms, with lower values indicating better-defined clusters.

NOTE:

- DBI: Lower values are better. A DBI close to 0 indicates well-separated clusters.
- Silhouette Score: Higher values are better. A score close to 1 indicates dense, well-separated clusters, while negative values suggest overlapping clusters.

These metrics provide insights into the quality and interpretability of the clustering results, helping to guide the selection of the optimal number of clusters for a given dataset.

3.2.2. K-Means Clustering | Metrics Used - Euclidean Distance

1. **Simple and Fast**: K-means is computationally efficient and relatively easy to understand and implement. It works well with large datasets, making it suitable for analysis even when dealing with a significant amount of data.

2. **Scalability**: K-means clustering is scalable to a large number of samples and has been used in many large-scale data processing scenarios.

3. **Interpretability**: K-means produces clusters that are easy to interpret. Each cluster is represented by its centroid, which is the mean of all the data points assigned to that cluster. This centroid can provide insight into the characteristics of the cluster.

4. **Versatility**: K-means can be applied to various types of data and can handle both numerical and categorical variables (after appropriate preprocessing). This versatility makes it applicable to a wide range of datasets.

5. **Well-suited for Convex Clusters**: K-means performs well when clusters are spherical or close to spherical in shape. It tries to minimize the within-cluster variance, which makes it suitable for convex clusters.

6. **Initial Centroid Selection**: While the performance of K-means can be sensitive to the initial choice of centroids, there are strategies to mitigate this issue, such as multiple initializations with different seeds and more advanced methods like k-means++.

However, it's essential to consider potential limitations as well:

1. **Sensitive to Initial Centroid Selection**: The results of K-means clustering can be sensitive to the initial placement of centroids. Different initializations may lead to different results.

2. **Assumes Spherical Clusters**: K-means assumes that clusters are spherical and isotropic, which may not always hold true for complex datasets with irregularly shaped clusters.

3. **Number of Clusters (K) Selection**: Determining the appropriate number of clusters (K) can be challenging and may require domain knowledge or additional validation techniques, such as the elbow method or silhouette analysis.

4. **Sensitive to Outliers**: K-means is sensitive to outliers, as it tries to minimize the within-cluster variance. Outliers can significantly impact the positions of cluster centroids.

5. **Equal Variance Among Clusters**: K-means assumes that clusters have equal variance, which may not always be the case in practice.

Overall, while K-means clustering has its limitations, it can still be a valuable tool for exploratory analysis and pattern discovery in your dataset, especially if the assumptions of the algorithm are met and appropriate preprocessing steps are taken.

3.2.2.1. Determining Value of 'k' | Elbow Curve & K-means Inertia

The elbow curve is used to determine the optimal number of clusters (k) for the K-means clustering algorithm. It plots the Within Cluster Sum of Squared Distances (WCSS) on the y-axis and the number of clusters (k) on the x-axis.

The elbow curve appears to have a distinct bend and decreases steadily around k=3. This suggests that optimal number of clusters for this dataset will be in the range of 2 to 4.

Refer Code Block (Cell)

3.2.2.2. K-means 4 Clustering Analysis

Refer Code Block (Cell)

3.2.2.2.1. Model Performance Evaluation

- **Davies-Bouldin Index (DBI): 0.058**

A very low DBI value indicates excellent clustering performance. It suggests that the clusters are well-separated and compact, with minimal overlap between them.

- **Silhouette Score: 0.976**

  A Silhouette Score close to 1 signifies outstanding clustering quality. It implies that the data points within each cluster are very similar to each other and dissimilar to points in other clusters.

- **Overall Assessment:**

  Based on both the DBI and Silhouette Score, the K-means clustering model with 4 clusters exhibits exceptional performance on the given dataset. The clusters are highly distinct and internally cohesive, suggesting that the model has effectively captured the underlying structure of the data. This strong performance increases confidence in the meaningfulness and interpretability of the identified clusters, providing a solid foundation for further analysis and insights.

Refer Code Block (Cell)

3.2.2.2.2. Cluster 4 Profile Analysis

ANOVA Results

- **Significant Differences:**
  - Most non-categorical variables, including `overall_mmnorm`, `potential_mmnorm`, `value_eur_mmnorm`, `wage_eur_mmnorm`, and various skill attributes, showed **extremely small p-values (close to 0)**, indicating highly significant differences between clusters. This suggests that these attributes play a crucial role in distinguishing player clusters.
  - `age`, `height_cm_mmnorm`, and `weight_kg_mmnorm` also exhibited statistically significant differences between clusters.
  - Interestingly, `movement_balance_mmnorm` did not show a significant difference (p-value = 0.077), suggesting that balance might not be a key factor in differentiating these player groups.
  - Goalkeeping attributes, while showing some significant differences, had relatively higher p-values, likely due to the smaller sample size of goalkeepers in the dataset.
  - The warning about constant input arrays for `fifa_update` implies that this variable has the same value across all clusters and thus doesn't contribute to cluster differentiation.

Chi-Square Test Results

- **Strong Associations:**
  - All categorical variables, except for `preferred_foot_oe`, demonstrated **very small p-values (close to 0)**, indicating a strong association between these variables and the cluster assignments. This highlights their importance in defining the player clusters.

    - `league_id`, `club_team_id`, `nationality_id`, and `nation_team_id` showed particularly strong associations, suggesting that these factors heavily influence player grouping.
    - `preferred_foot_oe` showed a p-value of 0.0199, which is still statistically significant, although the association with cluster labels is weaker compared to other categorical variables.

Cluster Profile Analysis

1. Centricity Analysis

Centricity analysis involves examining the cluster centers (centroids) to understand the typical characteristics of players within each cluster. Below are key observations from the provided centroids:

- **Cluster 0 (Goalkeepers):**
    - Characterized by significantly higher values in goalkeeping attributes (diving, handling, kicking, positioning, reflexes, speed).
    - Lower values in outfield attributes like pace, shooting, passing, dribbling, defending, and physic.
    - Club position is predominantly 'GK.'
- **Cluster 1 (High-Rated Players):**
    - Possesses the highest overall and potential ratings among all clusters.
    - High values in most key skill attributes (shooting, passing, dribbling, etc.), indicating well-rounded players.
    - Higher market value (`value_eur`) and wage (`wage_eur`).
    - Includes a mix of positions, but likely skewed towards attacking roles given the higher emphasis on offensive skills.
- **Cluster 2 (Mid-Tier Players):**
    - Exhibits mid-range overall and potential ratings, falling between the high-rated and lower-rated clusters.
    - Skill attributes are generally balanced, suggesting a mix of players with diverse skill sets.
    - Market value and wages are lower than Cluster 1 but higher than Cluster 3.
    - Likely encompasses a wider range of positions, including midfielders and defenders.
- **Cluster 3 (Lower-Rated Players):**
    - Displays the lowest overall and potential ratings.
    - Skill attributes are generally lower across the board, implying less developed or specialized players.
    - Market value and wages are the lowest among all clusters.
    - May include a mix of young players with high potential and older players in the twilight of their careers.

2. Cluster Sizes

The distribution of players across clusters provides additional context for understanding the relative prevalence of each player archetype:

- **Cluster 0 (Goalkeepers):** 132 players (relatively small cluster, as expected for goalkeepers)
- **Cluster 1 (High-Rated Players):** 10,122 players (a substantial group representing the elite players)
- **Cluster 2 (Mid-Tier Players):** 14,612 players (the largest cluster, indicating a majority of players fall in this category)
- **Cluster 3 (Lower-Rated Players):** 987 players (a smaller group, likely consisting of young talents and less prominent players)

3. Conclusion

The centricity and cluster size analysis reveals meaningful distinctions between the identified player groups. Cluster 0 represents specialized goalkeepers, Cluster 1 comprises the top-tier, well-rounded players, Cluster 2 encompasses a large group of balanced, mid-tier players, and Cluster 3 includes less developed or specialized players. These insights offer a valuable framework for understanding the diversity of player skills and potential within the FIFA dataset.

Refer Code Block (Cell)

Refer Code Block (Cell)

Refer Code Block (Cell)

4. Results | Observations

Clustering Performance Summary

Clusters	Silhouette Score	Davies-Bouldin Index	Memory Usage (MiB)
k=2	0.908	0.165	1048.59
k=3	0.968	0.201	1078.74
k=4	0.976	0.058	1094.34

5. Managerial Insights

Player Archetype Identification

The segmentation provides valuable insights into distinct player archetypes, allowing managers to identify players with specific skill sets and potential.

Team Building Strategies

Managers can utilize these insights to build balanced teams by strategically selecting players from different clusters based on their roles and attributes.

Player Valuation and Transfer Decisions

The clustering analysis can inform player valuation and transfer decisions by providing a framework for understanding the relative value of players within different clusters.