**Data Analysis Report on**

**PRCP-1004-Fifa20**

**By: Sameera Tanveer,**

**Arunmozhi Muthamizhan**

**Date: 28-11-2024**

**Table of Contents:**

1. Introduction
2. Data Preprocessing & Feature Engineering
3. Task2 & Task 3
4. Conclusion

**Introduction:**

Purpose of the report:

The report aims to give you the complete data analysis report on the provided data.

**Overview of the dataset:**

* The dataset is provided with attributes to explore the football skills and cluster football players.
* The domain of the dataset is “Sports”.
* The dataset contains a total of 102 features. It has 16 floats, 45 integers, and 41 object features.
* The data set has a lot of null values in multiple columns.

**Objectives:**

The objectives of the dataset are to explore the skills, cluster football players based on the given attributes.

To list the top 10 countries based on the players’ rank and the countries which have the players to play this level.

To determine the plot distribution of the overall ratings age of the players and the age at which the players stop improving.

To find the remuneration of the players based on their playing position.

**Data Preprocessing & Feature Engineering.**

**- Data Cleaning:**

Firstly, the unwanted columns are removed from the data set since there won’t be any impact on the features.

**-Feature Extraction:**

The unwanted features are extracted before dropping the null values. The birth month and date are not required for the analysis so they are not extracted.

**- Handling Missing Values:**

The data set has a lot of null values so it is difficult to proceed further. Since the number of null values in the data set is high. So, the null values with the percentage are determined.

Now the columns with greater than 50% null values are dropped.

**- Check for duplicates:**

There are no duplicate values present in the data set.

**- Handling Outliers:**

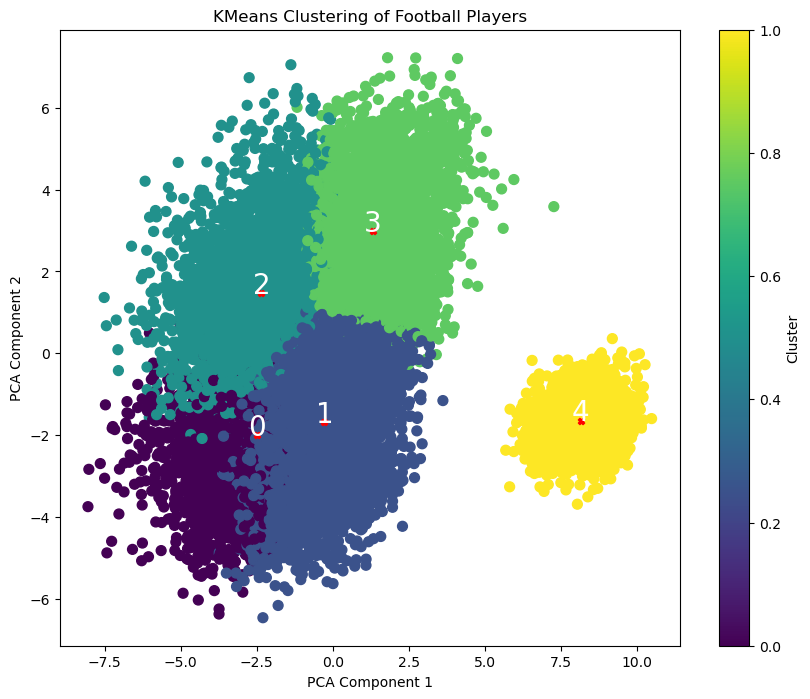
* After the completion of feature extraction and missing values, our next step is to handle the outliers present in this data set.
* We found the outliers for various features such as age, height\_cm, weight\_kg, overall, potential, value\_eur, wage\_eur, etc.
* At last, we found the number of outliers in this data set is very high and planned to implement the DBSCAN clustering algorithm which can identify and handle noise points as outliers.

**Task-2:**

1. **To Explore football skills and cluster football players based on their attributes:**

Firstly, we have to handle the categorical columns. Followed by, we applied scaling and clustering algorithms.

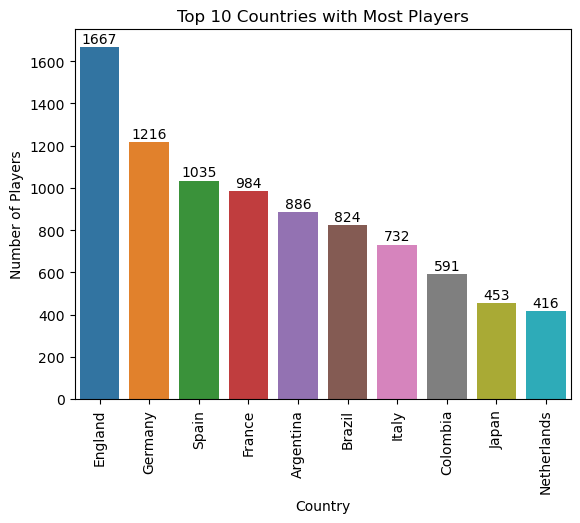
* The features are scaled by the standard scaler.
* Apply the Kmeans cluster and check the cluster counts.
* Visualizing the clusters using PCA to reduce the dimensionality.
* The clustering of the football players is mentioned in the form of a scatter plot.



**Task-3:**

**1. To prepare a list of the top 10 countries with the most players:**

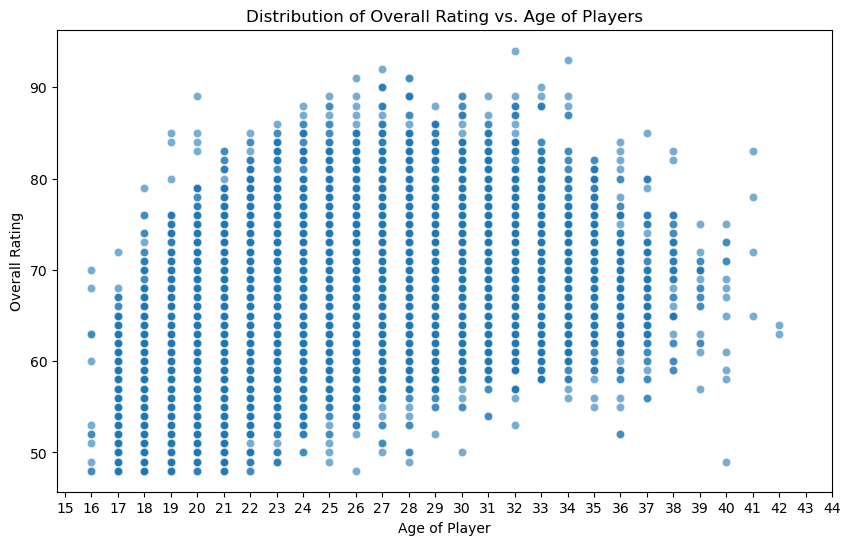
The 10 countries with the greatest number of players have been determined with a bar plot. The below graph portrays the number of players that each countries have:



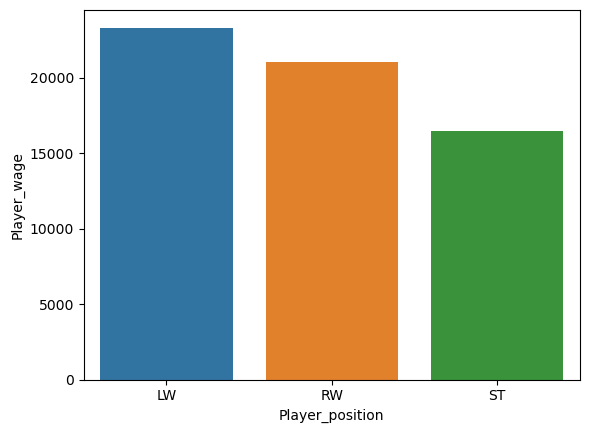
The graph clearly says that England has the highest number of players with 1667 players and at the last we have the Netherlands with just 416 players. The rest of the countries with the number of players are arranged based on the number of players.

**2. To plot the distribution of overall rating vs age of players:**

The overall rating of a player against their age has been determined by a scatterplot. The graph depicts that the players start to improve their rating from the age of 16 and gradually improve as they age. After the age of 32, the overall rating starts to decline. The graph is attached below for better understanding.



**3. To determine the highest-paid player based on the position:**

We have three positions such as striker, right-winger, left-winger. By checking the salary by position of the players of each team, the left-winger holds the first position, the right-winger is in second and the third is the striker. The wage has been determined using a bar plot and mentioned below.

**Additional Analysis:**

**1. Comparison between Messi and Ronaldo:**

The stats of both players are listed and compared, we can highlight some of the facts:

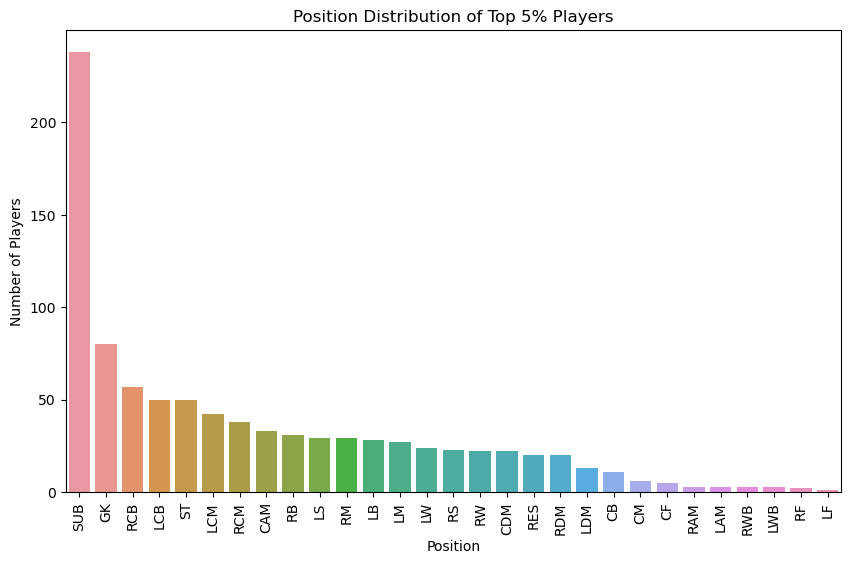
* Messi is the youngest player to start playing from the year 2004.
* Ronaldo first featured in the game in 2018 only.
* Both the players are well-established and known to all but, Ronaldo tops Messi in terms of strength, power and stamina.
* Messi earns a higher wage than Ronaldo, possibly because of his long-term significance to Barcelona, his global appeal, and his historically high earnings. In contrast, Ronaldo’s move to Juventus might have involved a different salary structure and financial considerations specific to the club.
* Even though Messi has been in form for many years, based on their achievements Ronaldo outperforms Messi.

**2. Budget of a Team:** The budget to form a competitive team with the top 11 players is 3410000 Euro.

**3. Average statistics of top players:** overall- 81.040526, potential- 82.778751, wage\_eur- 67474.260679, age- 27.889376, dribbling- 74.296755, shooting- 64.698046, passing- 70.646415, defending- 62.233520.

**4. Average statistics of other players:** overall- 65.467089, potential- 70.956349, wage\_eur- 6406.564929, age- 25.146271, dribbling- 61.846055, shooting- 50.877525, passing- 56.578777, defending- 52.015552.

**5. The position of the top 5 players:** The top players are usually kept as a substitute. Secondly, the goalkeeper position has the higher number of players. The position of the top 5 players is placed in the below graph:

****

**6. Additional information:** The average age of a player is 25, the highest wage of a player is 105500000 Euro, and the weight of the heaviest player is 110 kgs. The height of the tallest player is 205 cm.

**4. Conclusion:**

The mentioned tasks are completed using the best-fitting models and the data set has been analyzed properly using the implementation of the libraries that are required.

1. The football skills and cluster football players have been explored using Kmeans Clustering.

2. Top countries based on the number of players are listed. Germany holds the number 1 position and the Netherlands holds the last.

3. The distribution of overall rating vs. age of players is plotted. 32 is the age when the player stops improving.

4. The left-winger position is highly paid.

**Model Comparison Report:**

For calculating the outliers, we used a box plot manually but this data set has a high number of outliers and we planned to implement the DBSCAN to handle the outliers. The DBSCAN has the advantage of handling the outliers on its own. Since this data set is a cluster classification, we planned to use the DBSCAN.

Boxplot has been done for almost all the features which you can see in our notebook file.

**Challenges faced:**

1. Handling the null values in this data set is a great deal. Then the columns with the null values are dropped based on the missing values percentage.

2. Feature extraction is quite hard in this data because it is difficult to find the features which are not required for this analysis.

3. To bring the sum of null values to 0 in this project is quite challenging.

4. Outlier detection consumed more time in this project. Since we got to know the number of outliers is high and varying, we implemented the DBSCAN as it is the best suited for this data set.