NBA Player Analysis

Morau Horia-Andrei - NIE: 1720314
Balagiu Darian - NIE: 1719581
Valentin Micu Hontan - NIE: 1718971

May 27, 2024

1 Introduction

This project provides a comprehensive data analysis of NBA player data spanning 27 seasons, focusing on various statistical aspects and correlations. The analysis includes an in-depth examination of player performance metrics, correlation between several features, and how the game changed over the seasons. Key statistical categories such as weight, height, points, rebounds, assists are analyzed to identify patterns and significant changes over time.

2 Data Analysis

2.1 Data Types and Missing Records

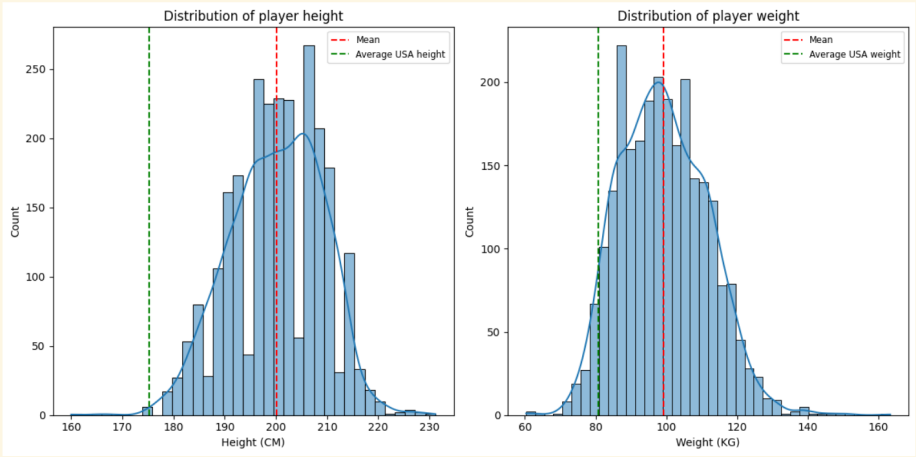
We begin by reading the necessary data and checking the data types and identifying any missing records in the dataset.

2.2 Statistical Summary

We provide information about the mean and standard deviation, minimum and maximum values, and the quartiles of the dataset which is contained of 2551 unique players over the course of 27 seasons.

2.3 Height and Weight Distribution

Both height and weight are distributed normally in the NBA. However, NBA athletes stand out in terms of their height and weight when compared to regular adult males. The tallest player ever to step on the NBA floor was Gheorghe Muresan with a height of 231cm, and he was Romanian! The shortest player is Muggsy Bogues, standing at 160cm.

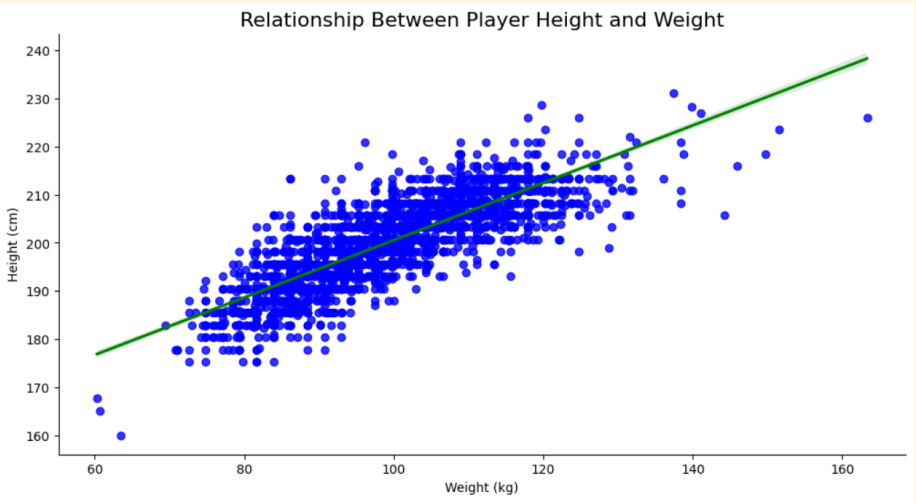


3 Correlation Analysis

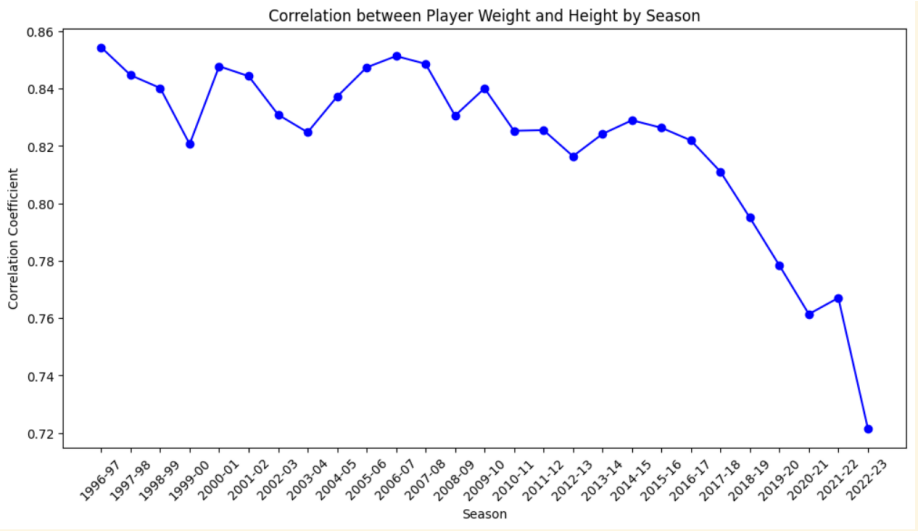
3.1 Height and Weight Correlation

To calculate the similarity between weight and height, we have used the Pearson correlation formula:

$$\text{sim}(a,b) = \frac{\sum_{p \in P}(ra,p - \overline{ra})(rb,p - \overline{rb})}{\sqrt{\sum_{p \in P}(ra,p - \overline{ra})^2}\sqrt{\sum_{p \in P}(rb,p - \overline{rb})^2}}$$



Not surprisingly, height and weight are closely related variables. The correlation between height and weight of players is shown for each season.



It is interesting to note that the correlation between height and weight has decreased over the years, indicating that NBA players' body types are changing.

4 Geographical Distribution

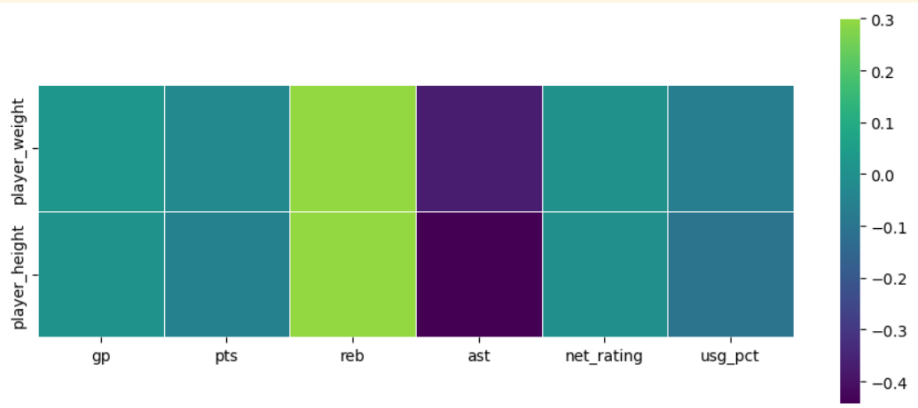
This section presents a detailed analysis of the geographical distribution of NBA players, focusing on the variation in height and weight across different countries. By visualizing these physical attributes on a global scale, we can gain insights into the diversity and physical characteristics of NBA players from various regions.

The analysis includes data representation through maps and charts, illustrating how player physiques differ by country. This geographical perspective helps highlight trends and anomalies in the physical profiles of players from different parts of the world. For example, we can notice that the highest average height of players comes from China!

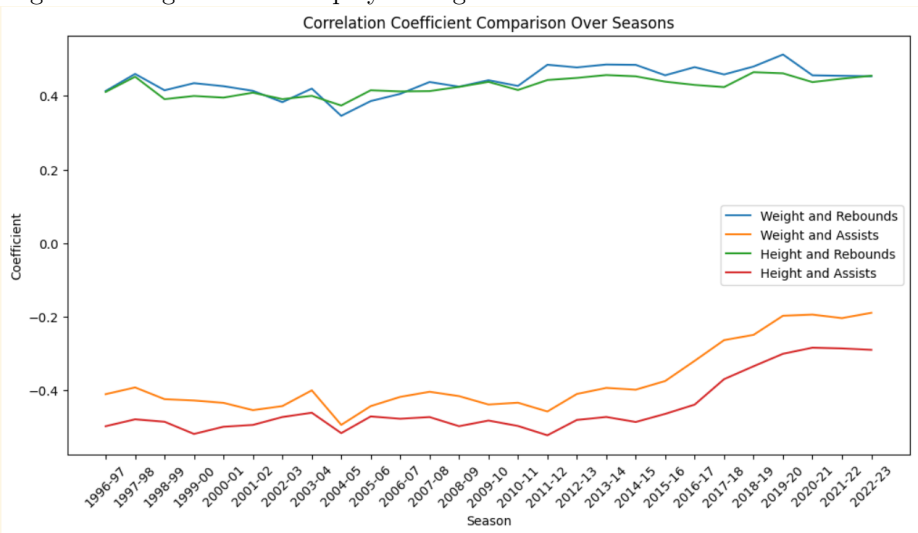
5 Performance Analysis

5.1 Height and Weight Impact on Performance

Neither height nor weight seem to have a significant impact on the total games played and average points scored. However, as expected, height and weight do impact average assist and rebound statistics.



Rebound coefficients are stable. However, the negative correlation for assists (the taller or heavier the player, the fewer assists he makes) has been reducing since 2011. This illustrates how the game is changing, with taller players becoming more integral to overall playmaking.



6 Principal Component Analysis and Standardization

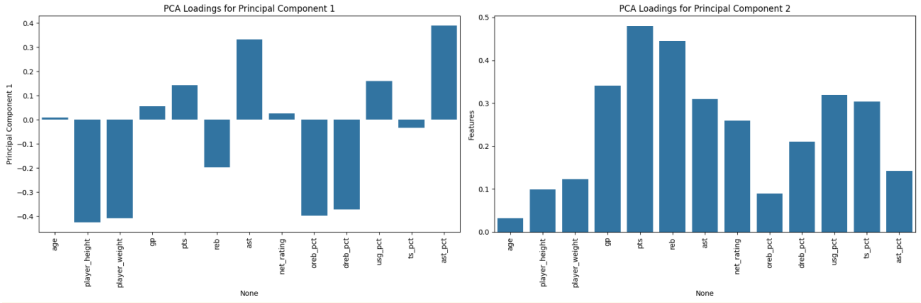
6.1 Standardization

We first apply standardization to all numerical data, removing all non-numerical values from the dataframe. The purpose of standardization is to transform the data so that it has a mean of 0 and a standard deviation of 1. By standardizing our dataset, we enhance the reliability and interpretability of our results, ensuring that each variable contributes equally to the analysis, and

preparing the data for PCA.

6.2 Principal Component Analysis (PCA)

Using PCA for reducing the dimensionality of the data optimizes clustering results, simplifies visualization and greatly improves clustering computation times. We can visualize how each feature influences each principal component.



Principal Component 1 is heavily influenced by physical attributes (player height, player weight), offensive rebounds (oreb pct), and defensive rebounds (dreb pct). It is negatively influenced by assist-related metrics (ast, ast pct). Thus, PC1 might be interpreted as a contrast between physical size/rebounding ability and playmaking ability.

Principal Component 2 is strongly influenced by scoring (pts), rebounding (reb), and games played (gp). These metrics have strong negative contributions, suggesting that PC2 may represent overall player activity and effectiveness in games.

7 K-Means Clustering

7.1 Elbow Method for Optimal Number of Clusters

Using the Elbow Method, we determine the optimal number of clusters. The optimal number of clusters in this case is 10 from reading the plot elbow point.

7.2 Clustering

After transforming the dataset into these two principal components, we applied the K-Means clustering algorithm with a predetermined number of clusters (K=10). This clustering approach helped us categorize the players into 10 distinct groups based on their playmaking skills and physical attributes. Each cluster represents a unique combination of these skills, allowing for meaningful analysis and comparison of player types across the NBA.

7.3 Mean Clustering

We noticed that players in the dataset may appear more than once due to players playing in multiple seasons. So we thought that for a much more accurate clustering, it would be a good idea to get the mean stats from all the player and essentially combine them into one player. This reduced computation time for clustering

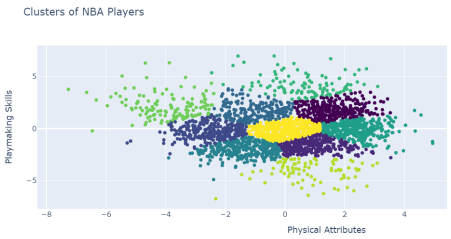


Figure 1: Cluster with Mean

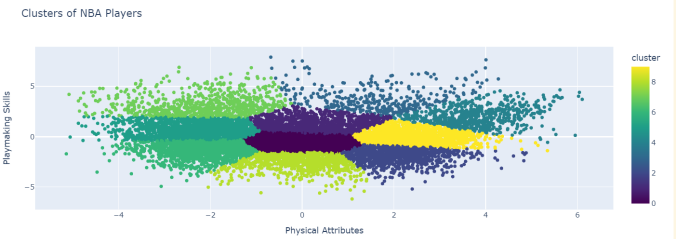


Figure 2: Clustering without Mean

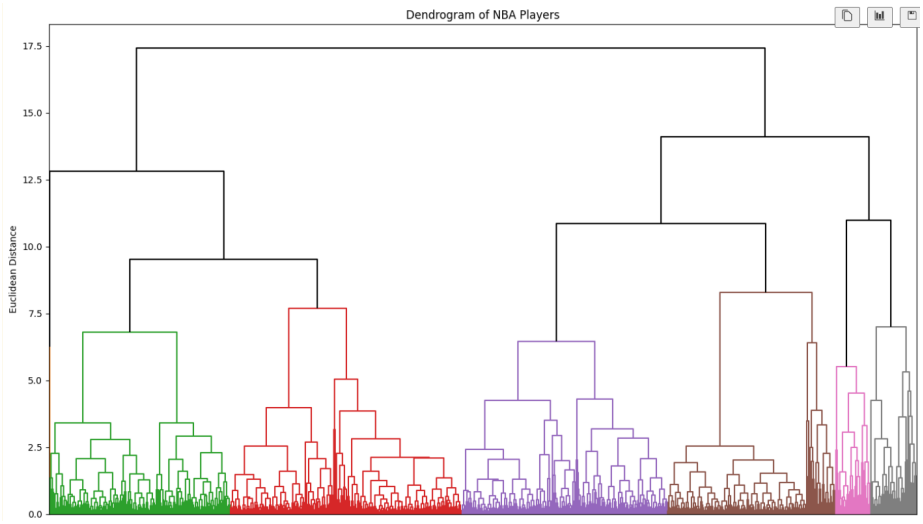
Despite some players previously belonging to multiple clusters, the clusters remain widely spread, indicating consistency for most players across seasons.

7.4 PCA vs Non-PCA Clustering

Without PCA, players are clustered primarily based on height and weight, showing a strong correlation between these features. With PCA, clustering also considers performance and playmaking abilities.

8 Hierarchical Clustering

Hierarchical clustering highlights outliers more clearly. For example, Cluster 6 contains only two players who are far more scattered than the rest, a distinction not as evident in K-Means clustering.



9 Recommendation System using K-Means

This recommendation system identifies and suggests similar players based on their statistical features and K-Means Clustering.

9.1 Recommendation Process:

Input Player: When a user inputs a player's name, the system checks if the player exists in the dataset.

Player's Cluster Identification: The system identifies the cluster to which the player belongs.

Distance Calculation: It calculates the Euclidean distances between the input player and other players in the same cluster, using their statistical vectors.

Sorting and Selection: The players are sorted based on their distance from the input player. The closest players are considered the most similar.

Output Recommendations: The system outputs the top N recommended players, excluding the input player.

10 Feature Prediction using k-NN

The target feature for prediction is 'age'.

The ages are binned into three categories: 'Young', 'Prime', and 'Old'. The bins are defined as follows: Young: 18-27 years. Prime: 28-35 years. Old: 36-44 years. The reason for binning the ages is to achieve a more accurate prediction.

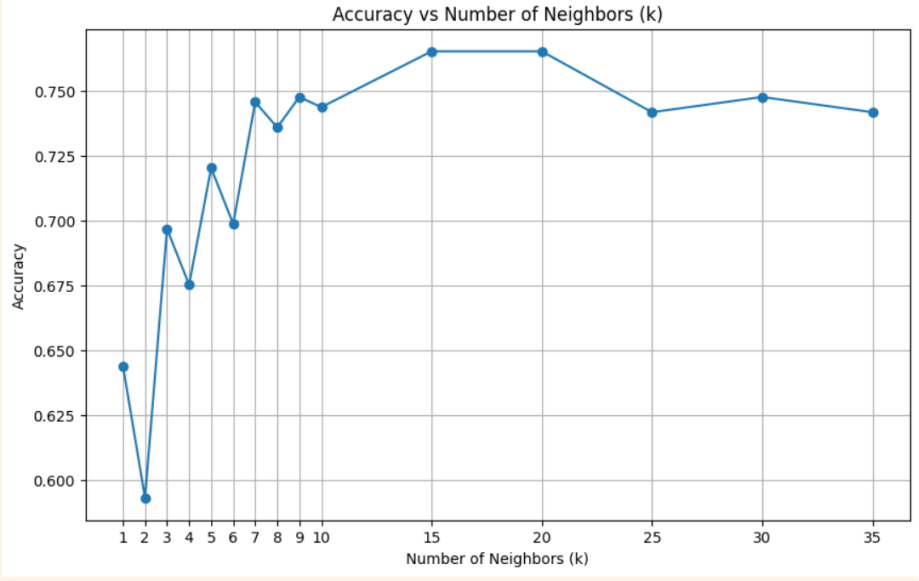
The features for the prediction model exclude the original 'age' and 'age binned' columns. The dataset is split into features (X) and the target vari-

able (y), where X contains the player statistics and y contains the binned age categories.

The dataset is split into training (80%) and testing (20%) sets .

A range of k values (number of neighbors) is tested to find the optimal value that yields the highest accuracy. The classifier is trained on the training data and evaluated on the test data for each k value, storing the accuracy results.

The accuracy of the k-NN classifier is plotted against different k values to visualize the performance. The optimal k value (in this case, 15) is selected based on the highest accuracy observed in the plot.



The system achieves an accuracy of 76%.

11 Conclusion

The analysis reveals several insights into the physical and performance attributes of NBA players, their correlations, and how these attributes have evolved over time. Clustering methods provide different perspectives on player categorization based on various features. The k-NN helps us predict player features such as their age.