Why do I believe that an analysis of NBA statistic will be helpful to my understanding of EDA?

I am the definition of a nerd. I was not aware about the statistics in the world of basketball. However, I sought to look at a dataset I received from Kaggle to see if I could reveal the predictors of the more successful players. During my preliminary analysis I found that some metrics were better predictors of success on the basketball court. My research revealed that the top positively correlated metrics of NBA teams were 2-pointer percentage, field goal percentage, 3-pointer percentage, assistant to turn over, free throw attempts over field gold attempts and percentage points from the free throw line.

The author collected his data from basketball reference website. I wanted to do that as well, but I felt that I was already invested in the data I had from Kaggle. Unfortunately, the Kaggle data did not have all the features mentioned in the authors analysis. So, my decision is to analyze and come up with the best fit features. The features I chose were net\_rating (of each player), oreb\_pct (Offensive rebound percentage of each player), dreb\_pct (defensive rebound percentage of each player), and ts\_pct (true shooting percentage).

What I intend to do is use these features to predict which player is likely to be the rookie of the year as I am aware of when each player starts his NBA career but not when the player has stopped or is injured. Also, I had made the effort to download from the basketball reference site which of the players were selected MVP and Rookie of year from 1946-47 to 2020-21.

While exploring the data I realize that the age for the NBA players range from age 18 to age 44. However, there is a disproportionate number of players who are at or around 24 years of age. On the feature of player height, I found that the minimum height is 160.02 centimeters, and the maximum is 231.14 centimeters. However, there are precious few players below the height of 180 cm or 5ft 10 inches while majority of the players are at a height close to the mode 205.74 which represent 12.3% of the 11 thousand records in the data although there are only 2,235 players in this database. I thought that this number would more align with what is seen on the television when we watch basketball. The top 5 players that are repeated in this database represent are either close to or taller than the mean height of 200.81 cm or 6ft 7in. Those top repeated players are more likely to be top players and more likely to be seen on television than a shorter player who only played 1 season.

I would like to have had more practice with regression analysis and time to demonstrate more of the skills I learned in this course.

Github link: <https://github.com/theojam2/NBA_EDA_Class_Submit_21>

References

<https://www.kaggle.com/justinas/nba-players-data>

<https://www.basketball-reference.com/leagues/>

<https://towardsdatascience.com/how-different-metrics-correlate-with-winning-in-the-nba-over-30-years-57219d3d1c8>

