210 Project

Austin Brown and Ryan Yu

Introduction

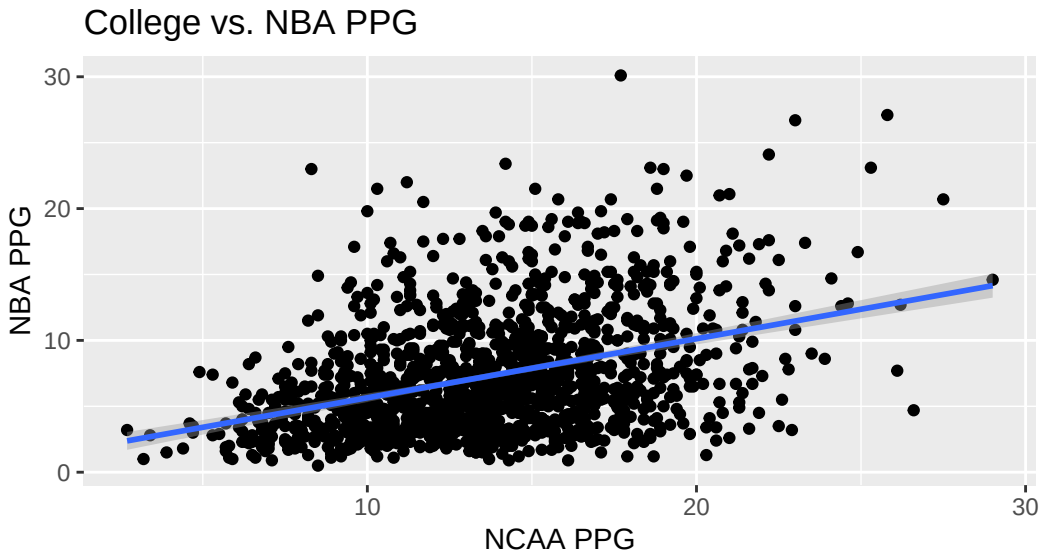
For NBA team owners and general managers, the college basketball level represents a key scouting opportunity as they search for their next potential draft prospects. In fact, for the 2020-2021 NBA season, 84.5% of players on NBA rosters played division I college basketball (DNA of Sports). The college level presents players an opportunity to showcase their abilities on both ends of the floor, and prove to scouts that they are ready for the next level. However, it's also important to note that the college game is vastly different from the NBA game in certain ways. In addition to the rules being slightly altered, the NBA game is faster-paced, more spread out, and has a higher concentration of talent on the floor (Brokke). Because of this, a player who fills the stats sheet and dominates at the college level may not be a guaranteed top pick and success in the NBA draft. The goal of this research report is to examine the relationship between college basketball statistics and NBA statistics, specifically with the focus of scoring ability, and ultimately derive an optimal model to predict the NBA's future prolific scorers based on college performance.

Data

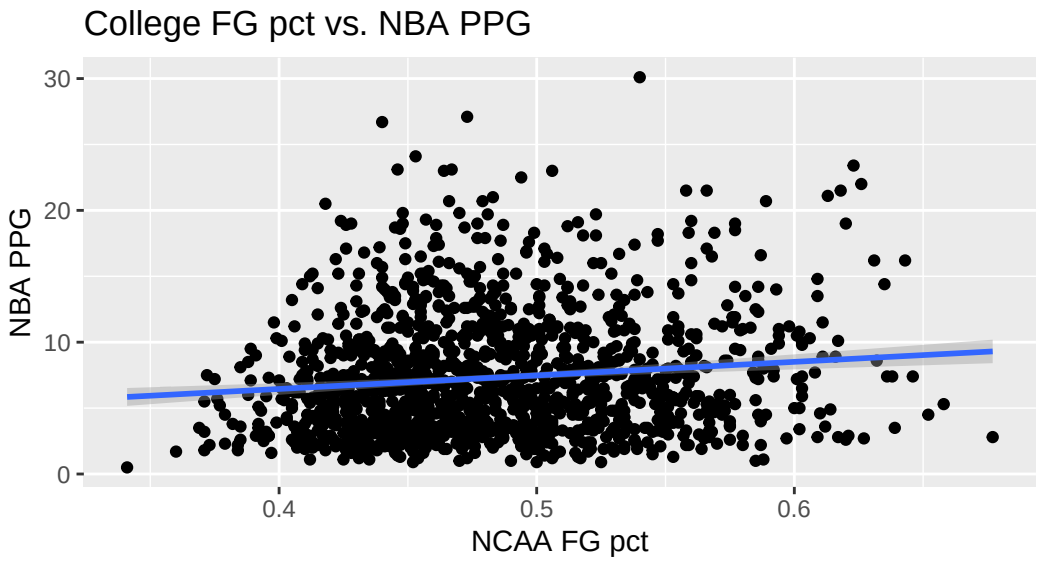
The dataset we are using for this project comes from Data World in a table that originally contained 34 columns and 4,576 rows. A link to the original data dictionary can be found here. This dataset contains observations of people who have had both NBA and NCAA basketball careers, with data from the beginning of the NBA until 2020. Key identifiers in the dataset include name, date of birth, and college. For the purpose of this project, however, we are focused on performance statistics. Specifically, we are looking at NBA PPG (total points/total games played) as a measure of offensive proficiency in the NBA. As our main predictors of interest, we are looking at college performance statistics. This includes PPG, total games played, and also includes measures of efficiency such as field goal percentage (includes both two and three point shots), three point field goal percentage, and free throw percentage. All shooting percentages are measured as total makes divided by total attempts. For the purpose of our research, we eliminated url, birth date, X (number identifier) from our dataset. We also removed NCAA effective field goal percentage, as this is not a commonly measured statistic at

the college level, and thus had many missing observations. After we removed these columns, we used the na.omit function to remove all observations with NAs in the columns. In addition to this, we trimmed the data by filtering for observations with a minimum of 6 college games and 21 NBA games, to avoid samples without a reasonable sample size. For the purpose of our investigation, we also altered the position column to only represent a player's primary position in the case that they were labeled as more than one. For example, if a player was labeled as "G-F" (meaning Guard/Forward), we changed their label to just "G". We also created a career length variable by subtracting the player's starting NBA year from their final NBA year. Lastly, we created a new categorical variable, "NCAA_length", to categorize a player's college career into one of four categories: "short", "mid-short", "mid-long", and "long". We did this by splitting the data into quarters given the range of NCAA games in the dataset from 16 to 152.

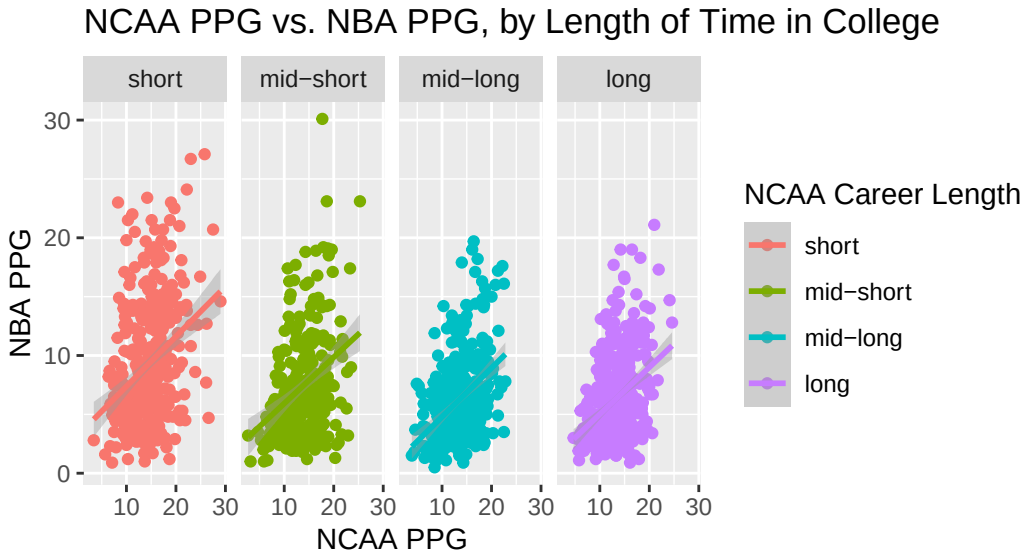
EDA

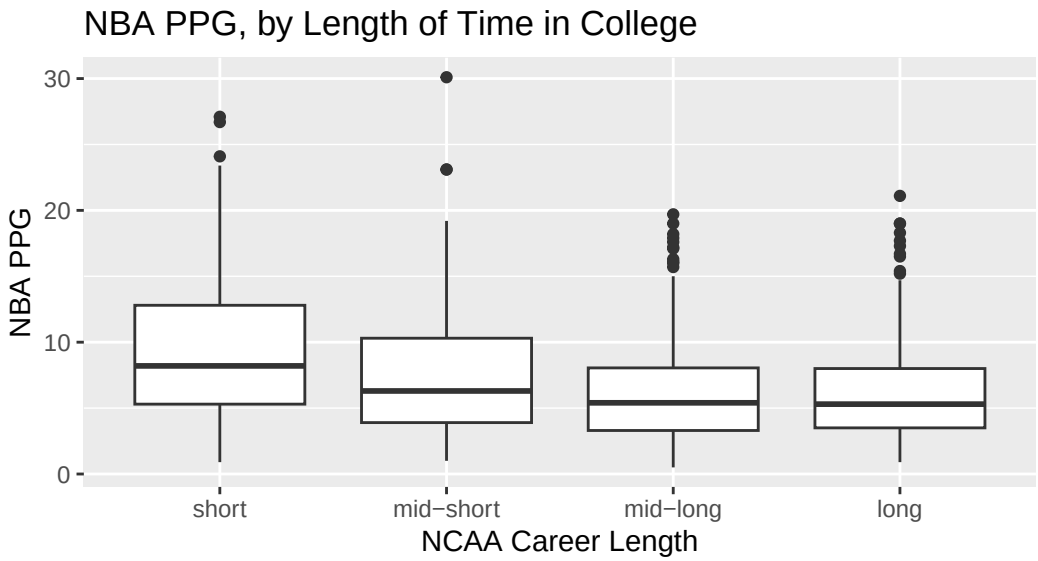


Our first initiative was to compare the direct analog to NBA ppg, which is NCAA ppg. As can be seen in the scatter plot above, there is a direct positive relationship between NCAA PPG and NBA PPG. One interesting thing to note, however, is that there are cases of 'high' NBA scoring averages (> 15) even for players who had NCAA scoring averages of less than 15 PPG.



The second relationship we wanted to explore is between scoring efficiency in college (in this case NCAA FG pct) and NBA PPG. We also see a generally positive relationship between NCAA FG pct and NBA PPG, however, the fitted line in this plot has a distinctly less extreme slope than in the previous plot. Both of these plots suggest a relationship between scoring ability and efficiency in college and NBA, however, it is likely that there are more factors involved in predicting, especially given that at the college level, most players may not be fully developed from a skill, physical, and mental standpoint. We explore this further in the next two graphs.





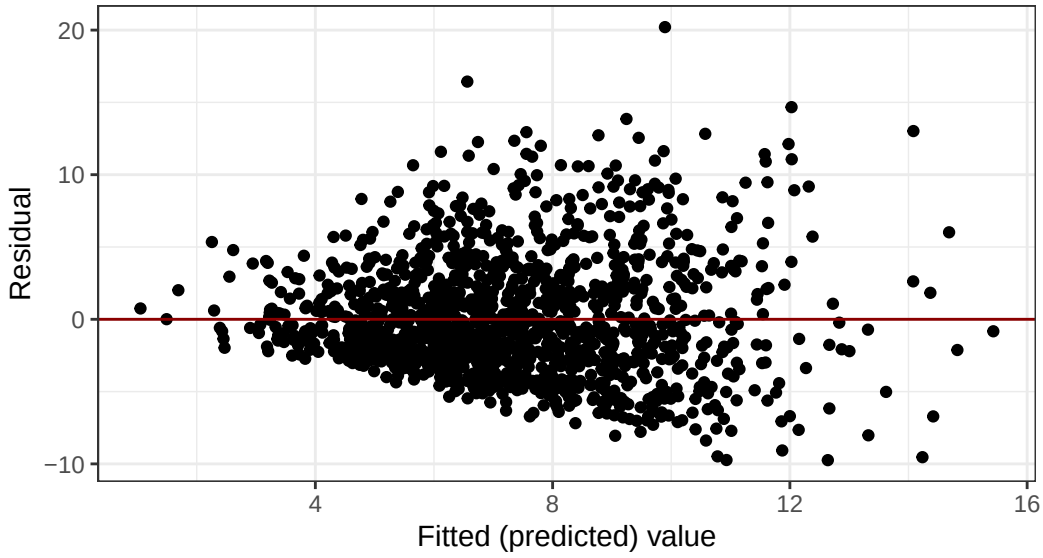
In the two graphs above, we wanted to get a sense of NBA scoring on a college career length basis. What we can see from both plots is that the two shortest categories for college career (likely players that left early for the NBA) have observations that spread higher in the NBA ppg category. All four categories maintain the upward sloping fitted line in the scatterplot, however, there is a clear difference amongst the four categories in the NBA scoring averge distributions. This is likely due to the fact that players who have 'lottery pick' potential for the NBA draft will often leave college early. These players are more often the players who develop into the most dangerous scoreres in the NBA. As we continue to explore these relationships and begin to craft a model, college career length will certainly be taken into consideration.

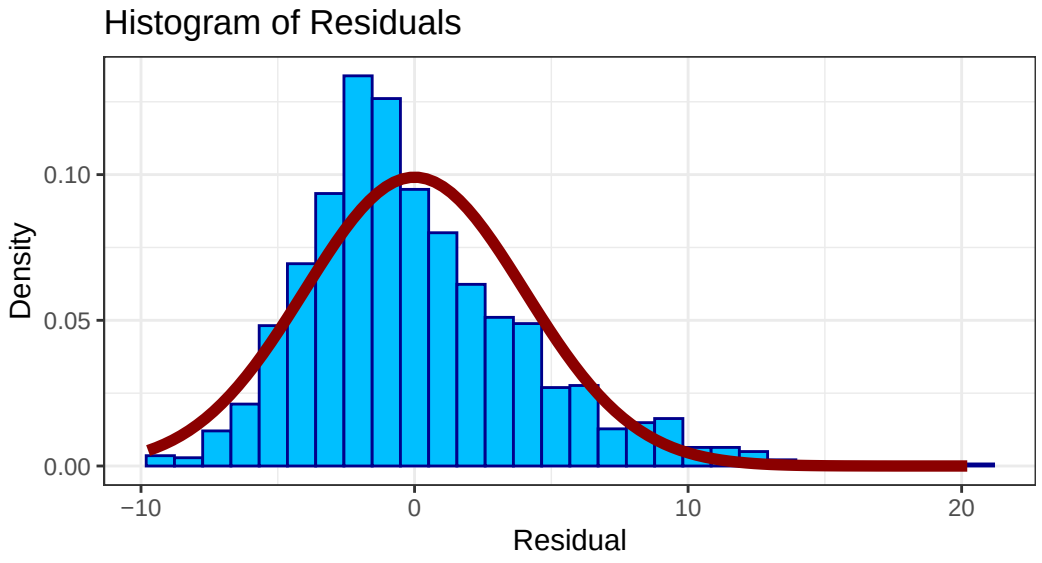
Methodology

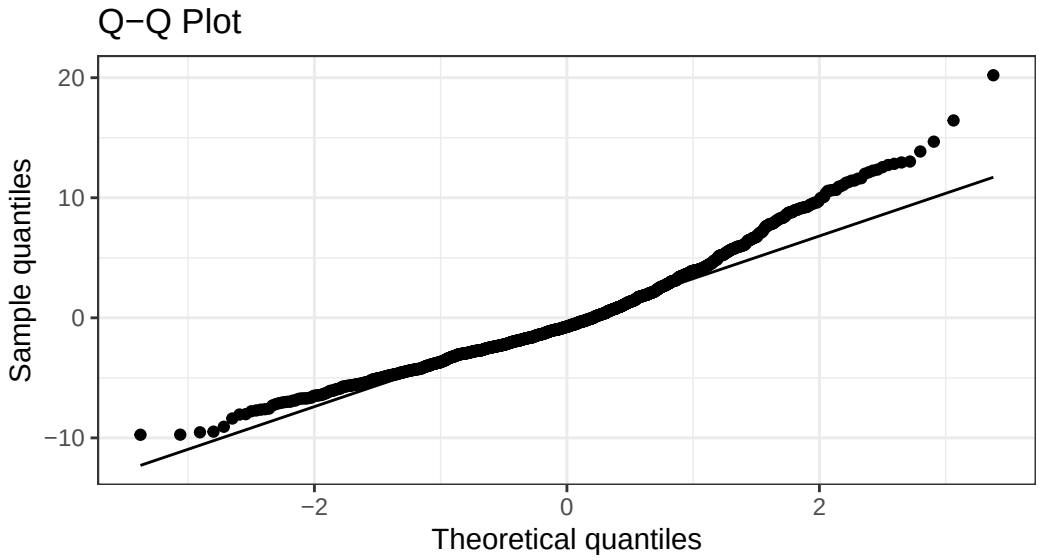
```
15 x 1 sparse Matrix of class "dgCMatrix"
                                      s0
(Intercept)                    .
NCAA_ppg                       0.389509975
NCAA_ft                        1.665822912
NCAA__3ptpct                   1.512781668
NCAA_fgpct                    15.292720454
NCAA_lengthmid-short          -1.226583839
NCAA_lengthmid-long           -2.481722769
NCAA_lengthlong               -2.032078068
positionF                     -0.038877682
positionG                      0.670013209
```

```
NCAA_ft:positionF                .
NCAA_ft:positionG                .
NCAA_ppg:NCAA_lengthmid-short -0.021265024
NCAA_ppg:NCAA_lengthmid-long  -0.006795547
NCAA_ppg:NCAA_lengthlong      -0.037416291
```

Assumptions Evaluation







Results

```
# A tibble: 11 x 5
   term                           estimate std.error statistic  p.value
   <chr>                             <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                     -5.06       1.86     -2.72   6.58e- 3
 2 NCAA_ppg                         0.398      0.0534    7.46   1.59e-13
 3 NCAA_ft                          2.63       1.68      1.56   1.18e- 1
 4 NCAA__3ptpct                     1.95       0.935     2.09   3.72e- 2
 5 NCAA_fgpct                      12.6        2.19      5.74   1.15e- 8
 6 NCAA_lengthmid-short            -1.15       1.18     -0.980  3.27e- 1
 7 NCAA_lengthmid-long             -2.57       1.14     -2.26   2.42e- 2
 8 NCAA_lengthlong                 -2.00       1.12     -1.80   7.25e- 2
 9 NCAA_ppg:NCAA_lengthmid-short -0.0314     0.0798   -0.394  6.94e- 1
10 NCAA_ppg:NCAA_lengthmid-long  -0.00686    0.0787   -0.0871 9.31e- 1
11 NCAA_ppg:NCAA_lengthlong      -0.0462     0.0784   -0.589  5.56e- 1
```

From this model, based off of p-values and a significance level of .05, we can identify the following terms as significant: NCAA_ppg, NCAA_3ptpct, NCAA_fgpct, and NCAA_lengthmid-long, as all of these terms have p values of less than 0.05. As for interpretations, if NCAA ppg increases by 1 point, we can expect that on average, NBA ppg will increase by 0.398 points while holding all other predictors constant. For NCAA_3ptpct, if that increases by 1, or 100%, we would expect NBA ppg to increase on average by 1.951 points, while holding all other predictors constant. To interpret the percentage terms on a different scale, for NCAA_fgpct, for every 1% increase, we would expect on average NBA ppg to increase by 0.126 points while

holding all other predictors constant. Lastly, if a player has a 'mid-long' college career, we would on average expect their NBA ppg to be 2.566 lower in comparison to a player with a 'short' college career, while holding all other predictors constant.

In terms of answering our research questions, the interpretation of the first three terms was logical and in line with our expectations. In general, players who score more in college, as well as are more efficient both from the field overall but also from three point range are likely to be better scorers in the NBA. The last term we interpret is interesting, as it is suggesting that players who stay longer in college are expected on average to score less than the shortest college career length. This can be explained by the 'one and done' phenomenon we briefly mentioned earlier, in that the players with the most potential NBA upside usually leave college early, even if they haven't fully developed yet as a player. In comparison, players who stay in college longer may have less natural scoring ability, and thus on average may score less in their NBA careers.

To check the predictive power of our model, we are going to compare our model to a baseline model, that uses college points per game as the sole predictor as one of our goals was to evaluate whether NBA General Managers should evaluate player scoring ability with more than just their college points per game.

Baseline Model Results:

```
     RMSE  Rsquared      MAE
1 4.21319 0.1506385 3.256958
```

Experimental Model Results:

```
      RMSE  Rsquared      MAE
1 4.046686 0.2173287 3.136414
```

Discussion and Conclusion

From the research we did, we can conclude that the most important college statistics to consider when predicting NBA scoring average are NCAA ppg, NCAA 3 point percentage, NCAA field goal percentage, and the length of the college career for the player. What our research shows is that it is important for general managers and NBA scouts to look beyond the basic scoring numbers a college player shows. How efficiently a player can score can be an important factor, especially because college players transitioning to the NBA may get a lower volume of shots at the next level, which would affect their total scoring volume. In addition, it is also important to look at the age of a player and how long they've been in college. Along with consideration of other important factors, a young, first year college player with a lot of raw talent may have more upside than one of the best college players in the nation that is a 3rd or 4th year player.

Predicting NBA points per game from college basketball statistics is a challenging task due to several potential limitations. First and foremost, it is notoriously difficult to predict individual player performance with a high degree of accuracy. If it were easy, there would be no such thing as "draft busts," where players selected with high expectations fail to live up to their potential. Moreover, there are numerous external factors that can significantly affect a player's career trajectory. For instance, injuries, coaching changes, changes in team roles, and off-court distractions can all impact a player's ability to perform at their best. Additionally, the transition from college basketball to the NBA is itself a challenging process that can take time to adjust to, which can further complicate efforts to predict future performance based on college statistics alone. Therefore, while college basketball statistics can be useful in predicting NBA success to some extent, it is essential to recognize the limitations of this approach and to consider other factors that may impact player performance.

In terms of ideas for future work, we can look to incorporate other types of statistics such as defensive metrics, as well as more general basketball metrics such as on/off ratings. This would allow us to assess college players from a more holistic perspective, as scoring is obviously only one part of a total player's success and impact on a team. In addition, we could look to categorize the type of school that players come from, and investigate the impact of that on their future NBA success as well as on how reliable college statistics are for predicting NBA success. High-major players from schools like Duke and Kentucky, for example, may be easier to predict than players from less-known, smaller schools.

Sources

"What Percent of NBA Players Played in College?" DNA of SPORTS, 24 Feb. 2022, https://www.dnaofsports.com/basketball/what-percent-of-NBA-players-played-in- college/.

Brokke, Nathan. "NCAA Mania: Why NBA Is Much Better Basketball than College Hoops." Bleacher Report, Bleacher Report, 3 Oct. 2017, https://bleacherreport.com/articles/636365-ncaa-mania-why-nba-is-much-better- basketball-than-college-hoops.

"NBA NCAA Comparisons - Project by BGP12." Data.world, 6 Feb. 2020,

https://data.world/bgp12/nbancaacomparisons/workspace/data-dictionary.