Key Statistical Metrics: Analyzing the NBA with Advanced Analytics

Tahsin Rahman tr174, Henry Sun hs325, Rohan Bose rb404, William Zhou wwz4, Alp Altınbaşak aa638

I. Introduction and Research Questions

The National Basketball Association (NBA) has generated vast amounts of data for the past 70 years, making it a prime candidate for data analysis. Data ranges from broad topics such as field goal percentage to very niche topics such as medical history. Our research aims to utilize statistical analysis in order to identify strengths and weaknesses in player and team performance. Specifically, the following questions will be investigated in this report:

1. **Question 1:** The frequency of three-pointers in the NBA has increased substantially over the last 10 years. Has this had an effect on the average efficiency of NBA players?

This question identifies trends in the way the game is played and allows teams to improve their offensive strategy. It is relevant because NBA teams rely heavily on analytics, and identifying this trend affects the way the game is played and a team's offensive strategy. Our second question addresses how experience affects salary by asking:

2. **Question 2:** The average age of an NBA player in the 2021 season was 26.2 years old. Is there a relationship between age and salary in the NBA?

This question shows how value is determined in the NBA which is relevant as it directly affects free agency for players, contract negotiations, and more. Our third question looks at the past of NBA players by asking:

3. **Question 3:** On average, which colleges have produced the most successful NBA players over the last 20 years?

This question addresses even further what NBA teams are specifically looking for within players and allows us to identify certain programs which historically create successful players, which is relevant because NBA teams may choose to pick more players from a program with a track record of success.

NBA analytics are relevant to the broader scientific community and society as a whole, as they can provide valuable insights into player evaluation and team-building strategies. These methods can be applied to other professional sports and sports betting markets, which constitutes a trillion-dollar industry. Additionally, this project is worthwhile to our time in this course as it allows us to apply the knowledge and skills gained throughout the course to a real-world problem in a popular and exciting sport.

II. Data Sources

All of the datasets we used can be found in our Google Drive folder here. All of our data sets are originally sourced from basketball-reference.com. We downloaded the datasets from Kaggle or other websites in .csv format and then wrangled it for our own purposes. For our first research question on the effect of three-pointers on scoring efficiency, we will use the NBA Player Stats dataset, which includes the player-level statistics for the past 40 years, including field goal percentage, three-point percentage, and true shooting percentage. This dataset is appropriate for addressing our research question since it includes the necessary variables to calculate offensive efficiency metrics. For our second research question on the relationship between age and salary in the NBA, we used the NBA Player Salaries dataset and combined it with the players dataset, which includes player-level salary data for the past 40 years. This dataset is appropriate since it contains salary and age information of players, which we used to identify the relationship between player age and player salary. For our third research question on the most successful NBA players produced by colleges over the last 20 years, we determined the sum of total career points in the NBA per college, the number of top 5 draft pick players per college, and number of drafted players per college. This was done using the data in draft-data-20-years dataset, which had information about the players in the draft for the last 20 years, their college, and their stats in the NBA. We did not use much from the Draft_Combine dataset, but it holds information about physical

characteristics (height, weight, how much they bench press, etc.), and it could be relevant for future analysis regarding any correlation between draft combine statistics and career performance.

III. Modules Used

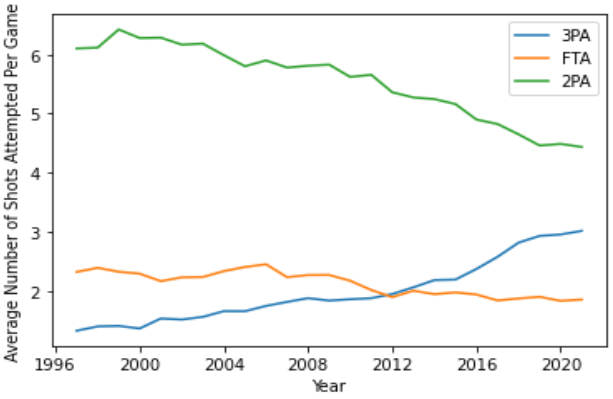
Module 4 and Module 6: Data Wrangling and Data Combining were **justifiably** used to clean and manipulate the NBA Player Stats dataset to calculate true shooting percentage and find player ages from birth date information. We also used wrangling **concepts** and techniques to combine data from salary and age datasets into one data frame for analysis. We used concepts such as split(), merge(), and join() to perform these tasks and combine the data. These modules were mainly used during the data cleaning and data investigation **stages** of our project.

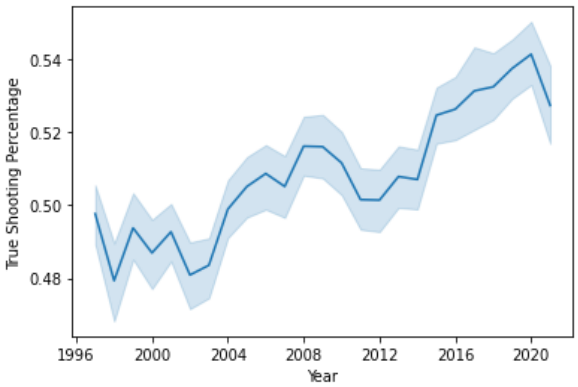
Module 5: Hypothesis Testing was **justifiably** used to test for the relationship between age and salary in the NBA, and to compare the difference in true shooting percentage between 1997-1998 and 2021-2022. We used **concepts** such as t-tests and p-values to formulate an answer as to whether there is a significant difference in the true shooting percentage between the 1997-1998 and 2021-2022 seasons. We mainly used this module during the data analysis **stage** of our project.

Module 8: Visualization was **justifiably** used to visualize the NBA data and make interpretations before our statistical inference stage. We used **concepts** such as Seaborn's barplot and line plot functions to visualize the distribution of variables and to visualize the relationship between our variables. This module was mainly used during the data investigation and analysis **stages** of our project.

IV. Results and Methods: Code Available Here

Question 1: The first question answered was "has the increase in 3-point frequency increased offensive efficiency?" To measure offensive efficiency, we used true shooting, a measure of how efficient a player is at scoring the basketball factoring in 3-pointers, 2-pointers, and free throws. Looking at field goal percentage independently can be skewed as frontcourt players who play closer to the basket tend to have higher field goal percentages, but this may not indicate them being more efficient scorers as they may not shoot as well from behind the 3-point line or the free throw line. A similar metric is effective field goal percentage, but this excludes free throws, making it a good determinant of a player's capability as a shooter, but not as a scorer - getting to the foul line and knocking down free throws at a high clip can very much be an efficient offense. After wrangling the data into a usable format and converting the year variable to datetime format, we created a column for true-shooting percentage (TS%), using its formula found online and combining the field goal percentage, three point percentage, and free throw percentage columns. Then, the two line-plots seen below were generated using Seaborn's line plot function. The visualizations generated below indicate that the average number of 3-pointers attempted per player per game has nearly tripled since 1997-1998, while the average number of 2-pointers and free throws attempted actually decreased. Similarly, the mean true shooting percentage has increased from 0.505 to 0.545, a significant jump.





Figures 1 and 2: Change in 3-point attempts and 2-point/free throw attempts over time, change in true shooting over time

To confirm the visualized increase in true shooting percentage, we used a t-test to see if the difference in TS was statistically significant from 1997-1998 vs. 2021-2022. Our null hypothesis for this question was H_o: There is no difference in TS% between players from the years 97-98. On the contrary, our alternative hypothesis was H_a: There is a difference in TS% between basketball players from the years 97-98. The p-value for this two-sided t-test was $2.22*10^{-16}$, which is far lower than the significance level of 0.05. Thus, we can reject the null hypothesis and conclude there is a significant difference in TS% between basketball players from the years 97-98, indicating NBA players and offenses are indeed becoming more efficient.

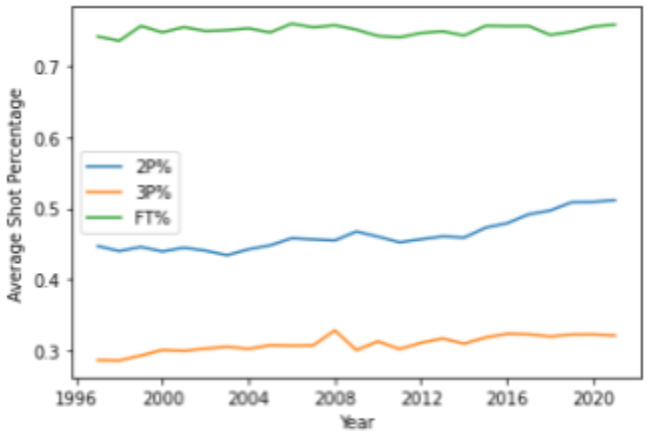


Figure 3: Change in field goal percentages and free throw percentage over time

Now, building from our work in the prototype, the main question was whether we could isolate the effect of increased true shooting and explain it via the increase in three point frequency. Because the average number of free throws and two pointers attempted per game has declined, the only other explanation for an increase in true shooting percentage would be if players were now shooting a higher percentage from inside the three point line and at the foul line. This graph above generated using Seaborn indicated that *all* methods of scoring (twos, threes, and free throws) saw their percentages increase, which is a testament to the increasing skill of NBA players (or weakened defense, depending on who you ask) and complicates our analysis using this particular dataset.

Question 2: The second question that was addressed is, "is there a relationship between age and salary in the NBA?" This question is fundamental to NBA teams, as they have to decide between paying their new up and coming players, or pooling their resources to play more experienced, older players. To answer this question, we first had to merge the player and salary datasets together based on player ID to be able to generate visualizations using one dataframe. Another challenge with our data was that the data spanned several decades which generated significant noise during analysis. To combat this, we created several bins for the data based on the decade in the variable year_interval while generating our visualizations.

Then, we created two graphs using Seaborn, one histogram showing the number of NBA players based on their age, and another barplot showing salaries of different ages. For the age data in figure 4, we found that the vast majority of players across all decades were in the age range of 22-30 years old, especially in the decade 2010-2017. Another interesting result was that there were fewer players aged 30 and above in the most recent time period than in past decades, indicating that the NBA as a whole is becoming younger and prioritizing younger, more athletic players over older, more experienced veterans.

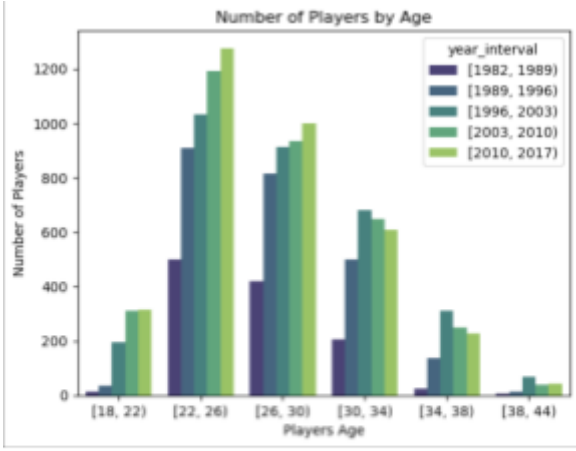


Figure 4: distribution of NBA player ages over time

Subsequently, we plotted the salary and age of NBA players over time. Building from our work in the prototype, we switched from a dot plot to using a similar bar chart that was more readable. One challenge with interpreting the results of the data is the fact that our data was not adjusted for inflation.

Hence, even the average rookie player during 2010-2017 made more money than an average 30 year old player during the 1980s, based on our data. Despite this, we believe the underlying trend in the data has not changed significantly over time, based on our analysis. We found that the distribution of salary vs age of NBA players across multiple decades suggests that as player age increases, salary tends to also increase, reaching a peak at around age 30-34, while slowly declining afterwards. The reason for this could be that players in this 'prime' age have had enough time to prove their abilities and earn a higher contract, while still performing at a high level. Younger players may lack the experience or resumé to justify a larger deal, whereas older players may be tapering off in terms of ability and are not worthy of a maximum contract either. An interesting result emerging within the last decade is that players past the age of 38 are now making significantly more than they were in the past. One reason could be that advances in treatment and sports medicine have allowed players to remain healthy and play at a high level longer than they were able to in the past, resulting in increased performance and salary.

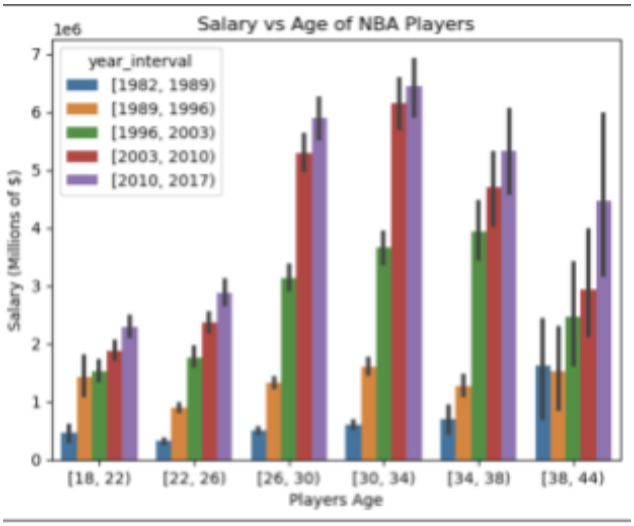


Figure 5: Player age vs salary in each decade

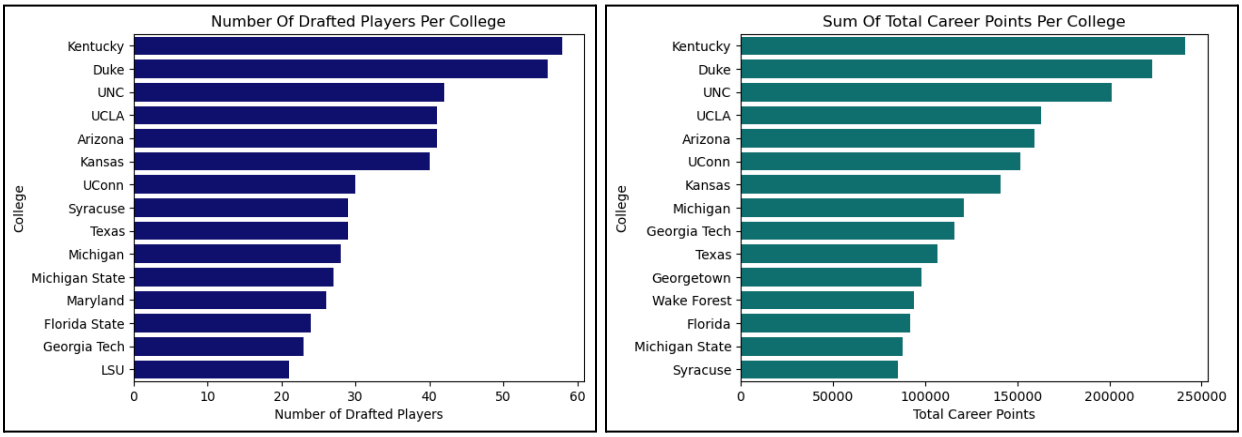
Question 3: The third question that was addressed is, "On average, which colleges have produced the most successful NBA players over the last 30 years?" Answering this question is more difficult than it seems, as there is not one clear cut definition of success. To address this problem, we selected certain metrics that would help us quantify the success of college basketball programs in producing NBA players, which are:

1. Number of players that got drafted to the NBA
2. Number of players that got drafted from the first five picks to the NBA
3. Total points scored in a player's career, grouped by college
4. Total points/assists/rebounds combined per player, grouped by respective college
5. Total salaries accumulated in the last 30 years for each player, grouped by respective college

We started our analysis by importing the datasets we needed based on the metrics we chose. We used the "draft-data-20-years", "players" and "salaries_1985to2018" datasets for our evaluations. In order to use the data effectively, we merged the "players" and "salaries_1985to2018" datasets using their common "player_id" values. Once we had the two main datasets we were going to use, we cleaned the datasets by dropping the columns with duplicate values or irrelevant criteria (ex: height, weight) using the np.drop and np.dropna commands. We then filtered our data to only get the players drafted in the past 30 years, beginning with 1990 using dataframe slicing. These steps allowed us to have the draft data, career salary information, and career stats of all players that have been drafted to the NBA in the last 30 years without any duplicate, missing or irrelevant values in two separate dataframes.

Finally we added some new columns to the datasets to make our analysis easier and used the "groupby" function of pandas to group the datasets by colleges and sum the values that we wanted to use for our

metrics. We used this process to find the total number of players drafted from each program, the total number of top-5 draft picks, the total career salaries of players combined for each school and more. As our last step, we ordered the grouped data by the criteria we were interested in and presented the top 15 schools for each metric in bar graphs. Since we wanted to represent each school individually and the schools did not have a pattern connecting them to each other, we found it logical to use scatter plots or bar graphs to visualize the data. We chose bar graphs over scatter plots because we decided it was easier to see the difference between the programs and understand the magnitude of the result in each metric using bar plots (as the length of the bars gives more visual information than dots at the same locations).



Figures 6 and 7: The number of drafted players per college and sum of total career points per college

The graphs show how the schools compare with each other on the categories of "Number of Drafted Players" and "Sum of Total Career Points". Below we provided a summary table that reflects our findings.

Ranking	Result-1: Number of Drafted Players	Result-2: Number of Top-5 Picks	Result-3: Total Career Points Per College	Result-4.1: Total Career Stats Per College	Result-4.2: Total Career Per Game Stats Per College	Result-5: Total Salaries Per Player Per College 1990-2017
1	Kentucky	Duke	Kentucky	Kentucky	Kentucky	Duke
2	Duke	Kentucky	Duke	Duke	Duke	UNC
3	UNC	Kansas	UNC	UNC	UNC	Kentucky
4	UCLA	UNC	UCLA	UCLA	UCLA	Arizona
5	Arizona	LSU	Arizona	Arizona	Arizona	UConn
6	Kansas	Georgia Tech	UConn	UConn	Kansas	UCLA
7	UConn	Georgetown	Kansas	Kansas	UConn	Kansas
8	Syracuse	Syracuse	Michigan	Michigan	Michigan	Georgetown
9	Texas	UConn	Georgia Tech	Georgia Tech	Texas	Georgia Tech
10	Michigan	Ohio State	Texas	Texas	Georgia Tech	Michigan

Figure 8 : Summary of findings for each metric

The table shows the results of each metric and how each school performed in the given criteria. The reason why we did not come up with a unified score for each school by combining their performances in each metric is because the weights of these categories are subjective. However, given the selected metrics, it is still abundantly clear that Kentucky, Duke and UNC (University of North Carolina) have produced more successful NBA players than other schools over the last 20 years.

V. Limitations and Future Work

Question 1: While we were able to demonstrate the correlation between 3-point percentage and true shooting, there are ways we could have done more to show why 3-point percentage was directly responsible for the increase. First, using the given dataset, we could have plotted each player's percentage

of points scored from behind the 3-point line over time to see if 3-pointers accounted for a significantly larger share of scoring. Even then, this would not be wholly conclusive. We were limited by the dataset used, which only gave us box score statistics for us to wrangle a metric out of (that being true shooting). A better method would have been to find team-specific data that showed the changes in net offensive rating (a team-wide measure of offensive efficiency), and whether that was due to the increase in 3-pointers taken by the team. This would be a better indicator since this trend, if present, would most likely represent a conscious decision by a team to shoot more threes. Another way we could answer this question with more time would be to find data showing efficiency by shot location on the court. This could help us prove whether NBA players were becoming more efficient at two-pointers overall and whether that was responsible for the increase in true shooting. Finally, a last step could be to join this dataset with another dataset containing information regarding player position or height to see if taller players or big men are shooting more threes, as they traditionally play in the paint. If this was the case, it provides more evidence for the three-point revolution heralding in a new age of NBA offense.

Question 2: While it was possible to visualize how different ages earned different salaries throughout the years, attempting to construct a linear regression and a neural network to model the parameters was unsuccessful. For instance, the fully connected neural network had a loss of around 90 percent between epochs. Even other regression patterns failed to sufficiently model the relationship between the two variables, indicating a more complex relationship between the two variables was present. Due to this, we decided to exclude these regression models from our final analysis and focused on the information we could find from visualizations. Another challenge with the salaries dataset was that the values were not inflation-adjusted. Today, even the lowest-paid NBA players make far more than the highest-paid players decades ago, meaning we had to split our bar graphs into multiple bars for each decade to adjust for this, adding to the complexity of the graphs. With more time, it would be worth attempting to see if there is a model that can fit the data, or if salary and age are normally distributed as appears to be the case. We could also expand on our research questions by investigating the impact of other factors such as player performance, position played, and market size on player salaries. For instance, the relationship between position played and player salary over time could also be indicative of a changing trend in NBA strategy, if we found that certain player archetypes became more sought-after over time.

Question 3: Although we used a very extensive dataset that included information on all players that have been drafted since the year 1990, for the purposes of answering our selected research question, we had to exclude a lot of data and be very specific about which criteria we were using. The aim of this question is to determine the successfulness of college basketball programs based on the players that they sent to the NBA, rather than to determine the overall success of college basketball programs. Two, the metrics used for this evaluation were selected by our team, and another party might choose different metrics in their evaluation. We can only make specific claims about the program's success based on specific metrics, and those claims could not be generalized beyond the criteria they were based on. Since it is almost impossible to conduct an analysis including every metric available, we had to choose the statistics that we found most important, like points, assists, rebounds and career salaries. In addition to all basic statistics, player performance also includes intangibles like leadership, work ethic and energy and advanced statistics which go beyond basic statistics. The rankings we found for each metric in our table are facts that were drawn directly from data, but once someone decides to combine those results in a way of their choosing and make claims about relative success, then these results inevitably become disputable and subjective. Future work in this area could involve somehow measuring other advanced metrics (such as leadership) and seeing if any colleges produce NBA players that stand out in this regard. Another interesting aspect would be to see if there was a correlation between the best-performing college teams (based on winrate, conference championships, and national titles) and the subsequent performance of their NBA players.

VI. Conclusion

Our analysis focused on three separate dimensions of the NBA landscape: player performance, player salary, and correlation between college and NBA success. To conduct analysis, we relied on a combination of hypothesis testing, data wrangling, and data visualization. We combined traditional statistics over a 25-year time period to generate measures of offensive efficiency and measured the change in those metrics (namely true shooting) over time. We found a statistically significant increase in both true shooting percentage and three-point frequency, but were unable to prove a causal link between the two. We also found that field goal percentages across the board were increasing over time in NBA offenses. As for salary, we used player and salary data throughout NBA history to examine the correlation between age and salary. We discovered that salary and age appear to be approximately normally distributed, revealing the players with the highest contracts were neither young, inexperienced players nor older veterans. This trend is not completely uniform, as plenty of outliers were present, but the highest-paid players tended to be within 27 to 34 years of age. Finally, we generated differing metrics of measuring a player's success in the NBA and tracked which colleges were most successful under each one. We find strong evidence indicating Duke, Kentucky, and North Carolina are a cut above other universities at producing NBA players with successful careers, regardless of the metric chosen, although we cannot discern the specific reasons with the data used. All these results obtained would serve as an effective framework for further questioning and research.