NBA Draft Dataset Analysis

Introduction

The NBA is the sports league where the highest level of basketball is played. In order for players to make it to this league, however, the most common route is the NBA draft. More specifically, the traditional route for players to make it to the NBA. The goal of this study is to determine how draft position and the college attended impact the success of NBA players. For this project, we define a successful NBA player by the number of years they played in the NBA.

Research Questions

Does draft pick position predict the length of a player's career?

Do players from specific colleges have longer careers on average?

How does draft position impact years played?

These questions are relevant in the context of social data analytics because it covers topics such as opportunity and career outcomes. Having a fair opportunity is not only essential in professional sports but in broader social contexts as well. For this study, I hypothesize that factors such as draft position and college attended will be key predictors in the length of a player's career. Analysis like this offer insight into how early opportunities can impact future career success.

Dataset

The dataset selected is from Kaggle. The data was collected from Basketball Reference, which is a site that posts NBA player information. It contains NBA player information relating to the NBA draft with players drafted from the year 1989 until 2021.

Source: Kaggle, Basketball Reference

Link: https://www.kaggle.com/datasets/mattop/nba-draft-basketball-player-data-19892021

Key Variables:

- Player Name
- College player Attended

- Year drafted
- Pick drafted
- Team drafted
- Years active in the NBA
- Games played in the NBA
- Minutes played in the NBA

File Format/Size: The file is in a CSV format and is 212.4kb. There are 1922 players recorded in this dataset.

	id	year	rank	overall_pick	team	player	college	years_active	games	minutes_played	...	3_point_percentage	free_throw_percentage	average_minutes_played	po:
0	1	1989	1	1	SAC	Pervis Ellison	Louisville	11.0	474.0	11593.0	...	0.050	0.689	24.5	
1	2	1989	2	2	LAC	Danny Ferry	Duke	13.0	917.0	18133.0	...	0.393	0.840	19.8	
2	3	1989	3	3	SAS	Sean Elliott	Arizona	12.0	742.0	24502.0	...	0.375	0.799	33.0	
3	4	1989	4	4	MIA	Glen Rice	Michigan	15.0	1000.0	34985.0	...	0.400	0.846	35.0	
4	5	1989	5	5	CHH	J.R. Reid	UNC	11.0	672.0	15370.0	...	0.135	0.716	22.9	

Justification

This dataset is a good fit for answering the previously indicated research questions. It includes key variables such as the college attended, year, pick, and team drafted. These are all important variables in predicting future success in the NBA. Additionally, to measure their success, the dataset contains key variables such as years active, games played, and minutes played in the NBA. Moreover, the dataset contains a large sample size, with information from 1989 until 2021. Considering these factors, this dataset is a great choice to find answers to the selected research questions.

Limitations

One limitation is that the dataset only contains data until the 2021 NBA draft, meaning we are missing out on 4 years of data. There could be more recent trends that cannot be identified because of this, potentially biasing the older generation of basketball. Furthermore, young players who only started their careers recently may skew the data. Moreover, there are beneficial stats in measuring career success such as points, rebounds, and assists per game. Awards would also be beneficial. Additionally, players stats in college would also be beneficial in predicting future success. With this information, deeper insight could be provided. Lastly, there may be

other factors at play such as injuries or coaching decisions that can impact the amount of time a player plays in the NBA that are not captured by this dataset.

Methodology

Data Cleaning:

Since there were null values in 'years_active,' I removed players who did not play in the NBA as they do not have measurable careers. Additionally, I created a separate dataframe with only college players who have played in the NBA. I also created another dataframe only including valid colleges.

The following types of analysis were conducted:

Descriptive Statistics:

- Grouped players based on draft position into ranges: top 5, lottery (top 14), first round (top 30), and second round (30-60) and determined the mean career length by draft position
- Categorized draft picks and determined the average years active using mean
- Displayed the top 10 colleges with the highest mean career lengths

Regression Analysis:

- Conducted linear regression analysis displaying the correlation between the number draft pick and the years played in the NBA
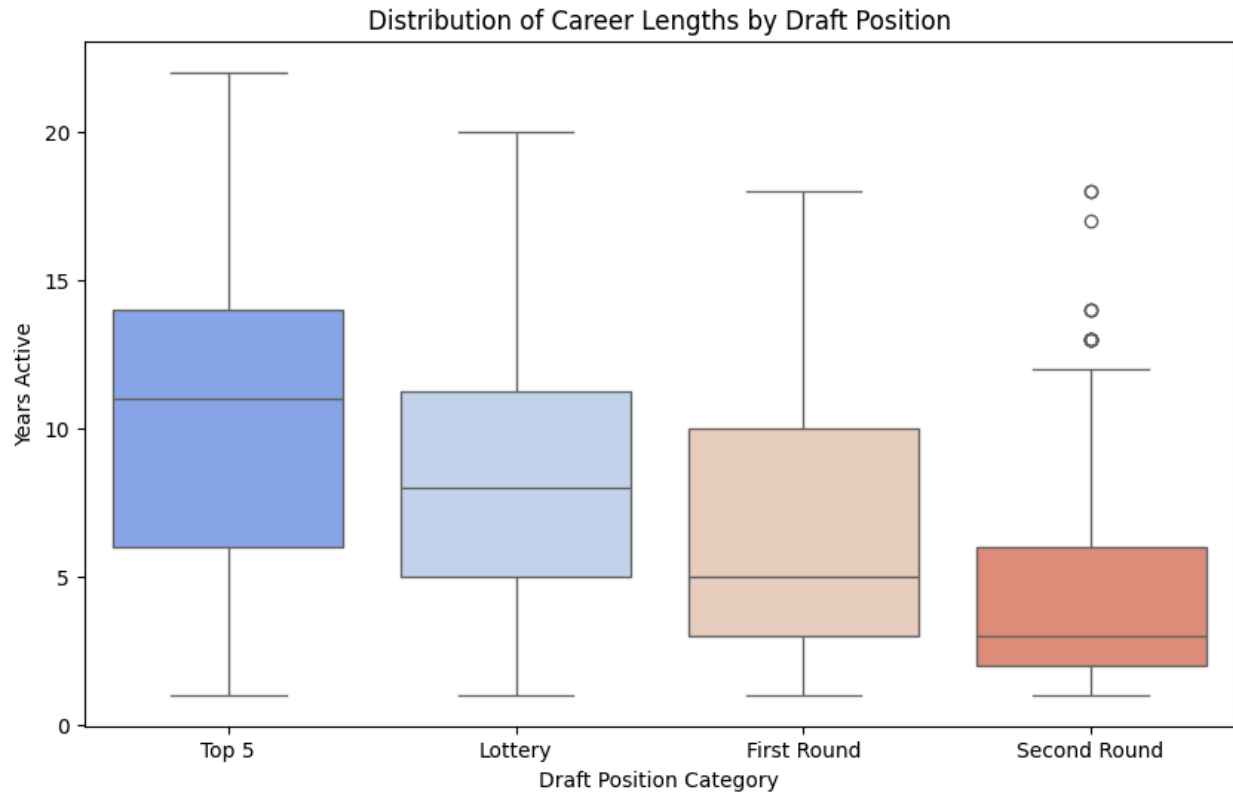
Results

Figure 1: Bar plot highlighting the career lengths of players by draft position

Top 10 Colleges with Highest Average Career Lengths

College	Years Active
Wake Forest	10.181818
Georgetown	9.133333
UConn	8.481481
UNC	8.425000
Texas-El Paso	8.400000
Temple	8.250000
Georgia Tech	8.086957
Cincinnati	8.083333
Oregon State	8.000000

Utah   7.888889

Figure 2: Table displaying the top 10 colleges and the average career length of players drafted from them
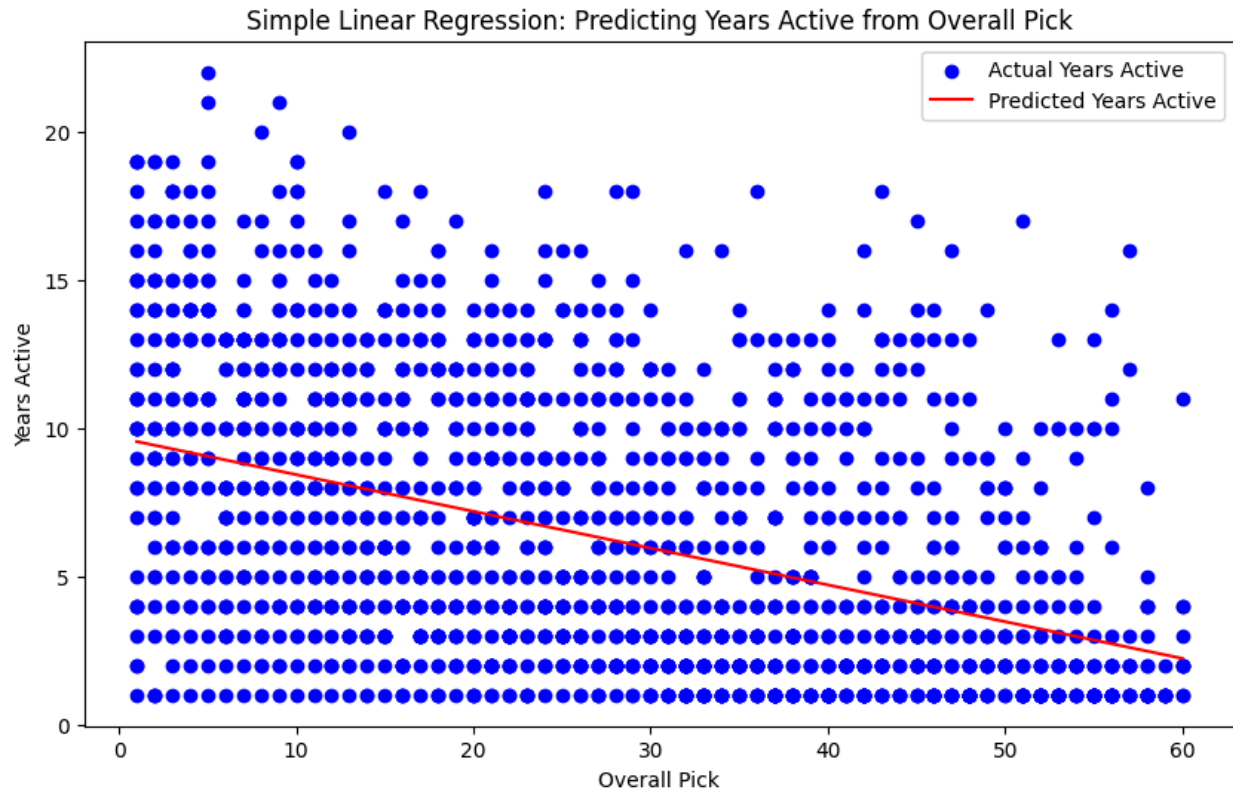

Figure 3: Linear Regression Analysis: Correlation between draft pick and years active

Regression Coefficient (Slope): -0.12395238104327806

Regression Intercept: 9.666846075086287

Discussion

Through this analysis of NBA draft data, we can gain insight into the impact of draft position and college on the success of players careers. Figure 1 reveals that players selected in the top 5 are more likely to have longer careers than those selected in later draft ranges. In Figure 3 we can see a weak negative correlation between the number draft pick a player is selected, and the amount of years they spend active in the NBA. This means that the later a player is drafted the less likely they are to remain in the NBA for many years. This is expected as players selected high are who teams believe to be the best player available. This correlation is weak, however, as indicated by the slope value because NBA players' success prediction is not linear. There are

many factors that can impact a player's chances of success such as injuries, inconsistent play, and off-court issues including a bad relationship with the team.

Conclusion

Overall, these results speak to a broader issue that is greater than sports. The fact that NBA players' success can be somewhat predetermined not only is related to talent but opportunity. Players selected low do not receive the equal chance that players selected higher get. This relates back to the idea that opportunity is necessary for success. As displayed in figure 2, by simply selecting a college, a player's chances of performing well in the NBA can shift dramatically.