Deep Dive Into NBA Draft Data

Dank Stan

Kyleena Xin (kx63)

Jacky Xu (hx289)

Sean Dreifuss (sjd262)

Eric Verdes (edv23)

Introduction

The National Basketball League is one of if not *the most* competitive basketball leagues in the world. With only 30 teams, and a whole world of talented players, only the best talent makes their way into the league. Since 1985, the NBA has hosted a lottery-style draft where the next generation of players are selected onto each team. The order of the draft sets expectations on incoming basketballers before they even set foot on the court. In this project, we seek to explore if the draft changes the trajectory of player's professional careers.

The motivation behind our project stems from a desire to better understand factors influencing NBA player performance relative to their draft position. Specifically, we sought to examine how the order in which players are drafted impacts their subsequent playing opportunities and overall effectiveness in the league. Using a dataset that we compiled through scraping data from NBA draft data spanning 2004 to 2024 – which includes metrics such as draft order, player minutes per game, and Value Over Replacement Player (VORP) – we formulated two key hypotheses.

Our first research hypothesis investigates whether players selected earlier in the NBA draft receive more playing time per game than those drafted later. Through linear regression analysis, our results confirmed that earlier draft picks indeed tend to play more minutes per game, supporting our initial hypothesis. We found this to be an interesting and important question to analyze because it tells us whether the draft is an accurate predictor for whether the player may have success (in terms of playing time) down the line when they are playing professionally.

Our second research hypothesis explores whether players drafted later (with higher numerical draft positions) have lower VORP values, indicating that they are more easily replaceable. This hypothesis was also supported by our analysis, as results indicated a clear trend where later draft picks correlate with lower VORP values. We found this question to also be interesting to look into because it tells us more about the nature of the draft, and whether it can accurately predict the value of a player in the NBA, whether they will be of high or low value.

Overall, our findings highlight significant trends in the NBA draft system, demonstrating that earlier draft positions not only influence immediate opportunities (e.g., playing time) but also longer-term effectiveness and value within teams. These insights contribute valuable information for understanding draft strategies and player development within the NBA.

Data description

What is the Data?

"NBA-drafts-all" is the data set that we created and used for the entirety of our project. The data set is web-scraped from the SportsReference.com site, with each row representing a different player. We scraped all drafts from 2004 to 2025, with each observation in our data set being representative of a different draft pick for the NBA. The top observations of the data set being from 2004, then 2005, 2006, and so on.

The data set contains all career-related statistics such as total points or total assists, percentages such as free throw and 3-point percentages, per game stats such as minutes and points per game, as well as some other stats such as a player's Value Over Replacement Player (VORP). These stats gave us an extensive overview of these players' career and impact in the NBA. To help with clarification we added a year column that specifies which year the player was picked in. The other columns are all player info, and correspond to their stats during their time **in the NBA** aside from what college they went to, if applicable.

In terms of funding, there was no external funding for the creation of this dataset; it was independently compiled by students at Cornell University for academic and exploratory purposes for the course INFO 2951 .

Data Source

The basketball draft statistics were collected from Sports-Reference.com. They were first collected in 2004 as a part of Basketball-Reference.com, curated by Justin Kubatko who ran the website until 2013 when it merged in Sports-Reference.com and now includes data outside of NBA Basketball as well in various other sports and leagues. The database has data on NCAA basketball stats which we will use, dating back to 1947 until the current date and continues to update regularly.

Key Variables

The following variables are the key variables that we selected from the data set to utilize in our research project.

- **Draft Pick**: The player's overall draft pick number.
- **Minutes Played Per Game**: The average minutes the player played per game.
- **Value Over Replacement Player (VORP)**: A statistic that measures how much a player contributes to a team compared to a replacement-level player, essentially quantifying the difference in production.

Data Purpose

This dataset was created to analyze relationships between NBA draft positions and player performance, initially focusing on pre-NBA factors but later shifted towards NBA career statistics due to consistent data availability.

Outside Influences

- Data availability from Basketball Reference heavily influenced observed variables. Data not uniformly recorded or easily accessible—such as pre-NBA international performance or detailed injury histories—was omitted.
- Additionally, instances of forfeited draft picks may have influenced the completeness of draft-position data.

Data Pre-Processing/Cleaning

- Scraped HTML pages from the Basketball Reference site using R (rvest).
- Looped through all the draft years from 2004 to 2024, scraping all the data into a single unified dataset.
- Cleaned and tidied the dataset, reformatting the column names and cell values.
- Created dedicated column for draft year clarity.
- Managed missing values (e.g., college attended marked as NA).
- Saved the cleaned and structured dataset as a CSV file called "nba-drafts-all.csv".

Data Collection Awareness

- The data involves publicly available historical NBA player performance statistics, meaning individuals (NBA players) were not directly involved or notified explicitly about this particular data collection. Players generally expect their performance data to be publicly analyzed and used for sports reporting, historical records, and statistical analyses.

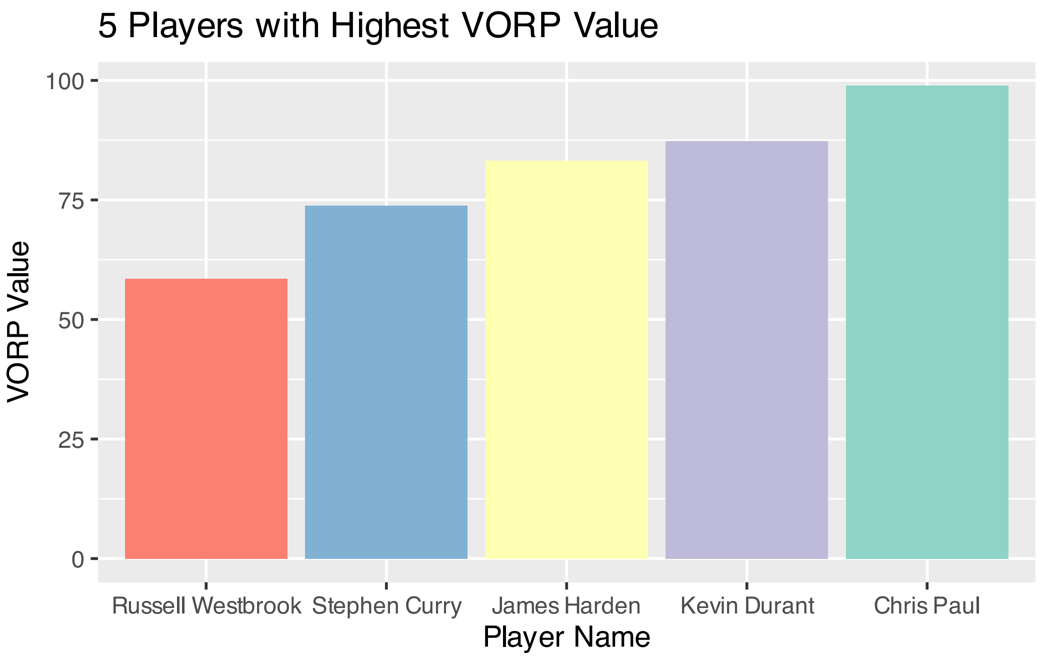
Data analysis

- As part of the data analysis, we examined two key indicators of player performance: minutes per game and Value Over Replacement Player (VORP). VORP is an advanced box-score metric that estimates how many points per 100 team possessions a player contributes above what a "replacement-level" player would provide. It is calculated using the formula: VORP = (BPM−(−2.0)) × % of minutes played × (team games)/82 (BPM = Box Plus/Minus (a box-score estimate of "points per 100 possessions" above average).

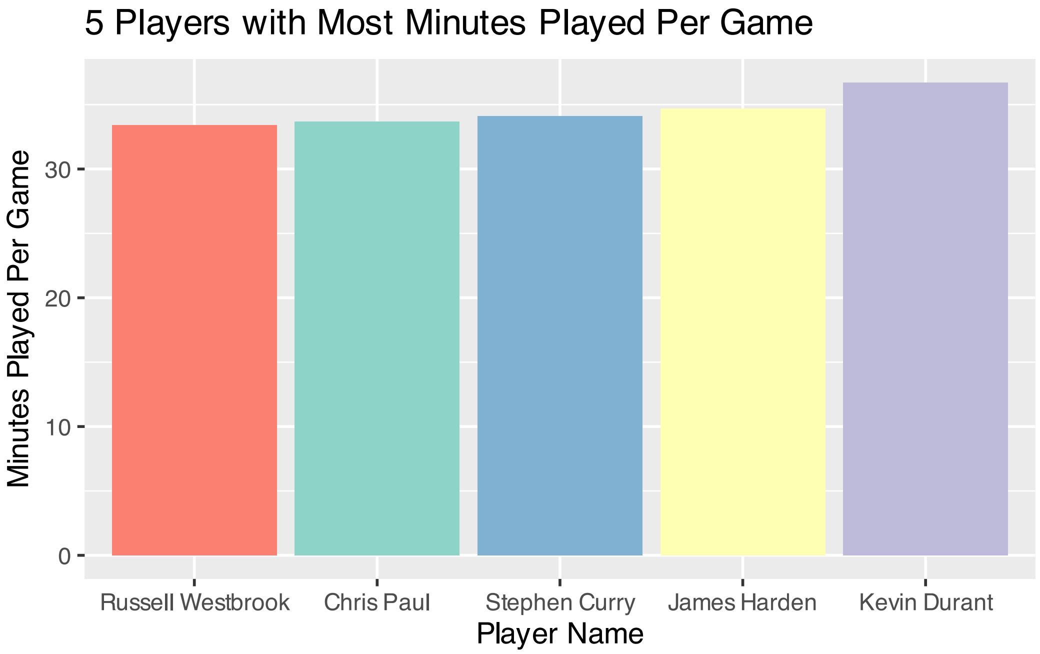
  We began by calculating basic summary statistics, including the mean and standard deviation, to capture the central tendencies and variability within the dataset. To visualize the distribution of minutes played and the VORP values overall, we constructed two histograms.

Summary statistics

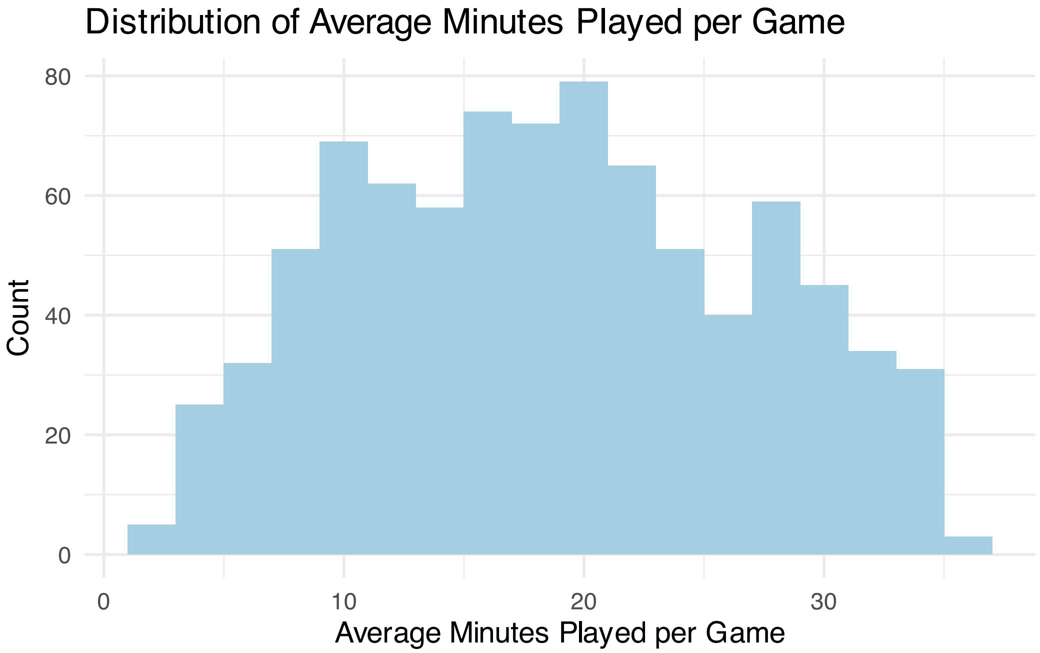
- The average minutes played per game across all players in our dataset is 18.71 minutes. This is 38.9% of the full game!
- The average VORP value across all players in our dataset is 3.91. Steph Curry has a VORP of 73.8 in comparison.



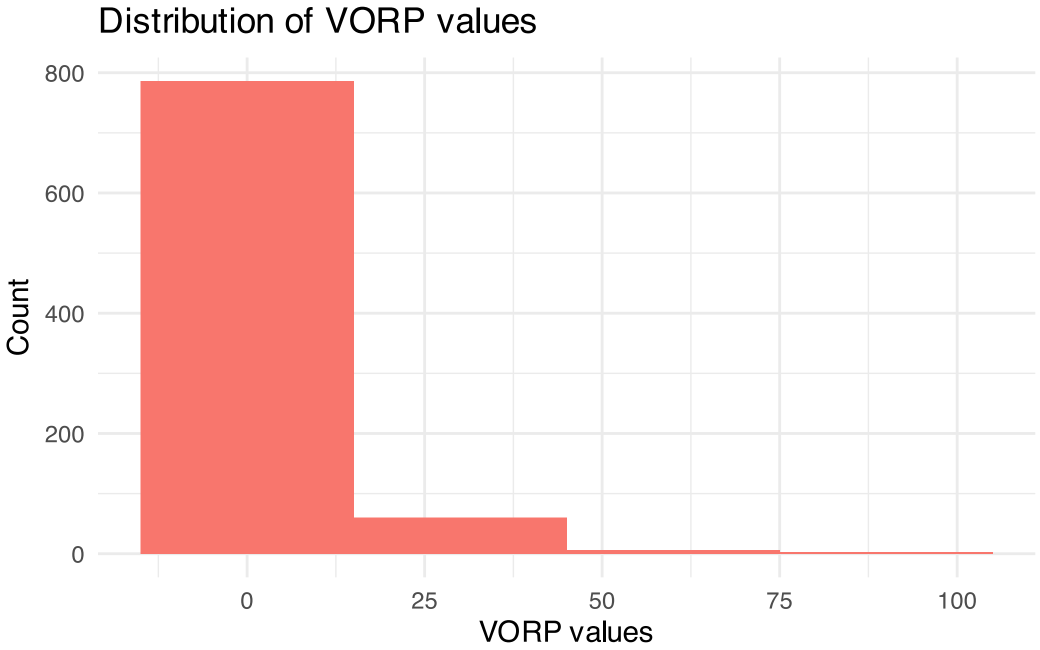
Here are the 5 players in our data set with the highest VORP values. These may be familiar names, with starts like Chris Paul and Steph Curry. These high VORP values indicate that they are much more valuable than the next replacement player.



Here are the 5 players in our data set with the most minutes played per game. It is interesting and logical that these are the same 5 players with the highest VORP value in the data set as well. Here, it is interesting that even through Chris Paul has the highest VORP value out of the 5, he does not play the most minutes per game. This may show that although some players may be more valuable on a team, they do not play as much.



As seen, the distribution of minutes per game for all players is a roughly normal distribution with most players averaging around 20 minutes per game.



As seen, VORP values cluster around −5 to 10, with occasionally some players having much higher values. Minutes per game display a roughly unimodal distribution, with most players having around 20 minutes per game and no significant outliers.

Hypothesis 1

Earlier draft picks play more minutes per game.

Analysis: For data where each row represents a different player, we run a linear regression where we input their draft number (the lower it is, the earlier the player is picked in the draft), and output the minutes played per game. The draft number will be the reference variable, so we will test whether $\beta_{draft} > 0$. This topic interests us because examining whether a player's draft pick number correlates with the number of minutes they play is highly relevant in the basketball world, and we believe that by uncovering this, we add value to our project.

Hypothesis 2

Lower draft picks have a smaller VORP Value.

Analysis: We will run a linear regression where the input variable is pick_num (draft position) and the output variable is vorp (Value Over Replacement Player). Since a higher pick_num corresponds to a later pick (i.e., a "lower" draft status), a positive coefficient for pick_num would suggest that VORP decreases as players are picked later in the draft. In other words, later picks tend to be more easily replaceable. We will specifically test whether the slope coefficient $\beta_{\text{vorp}} > 0$, indicating that later picks are associated with lower VORP values. This is interesting to us as we are trying to understand trends that are happening in the NBA draft, and testing to see the correlation between the draft pick and the player's VORP value would add value to our project.

Evaluation of significance

Hypothesis 1

Null hypothesis (H_0): There is no linear relationship between minutes per game and draft position

$$H_0 : \beta_{\text{pick\_num}} = 0$$

Alternative hypothesis (H_A): There is a negative linear relationship between minutes per game and draft position

$$H_A : \beta_{\text{pick\_num}} < 0$$

Since the p-value is less than 0.05 (less than 0.0005), we reject the null hypothesis, suggesting that there is a negative linear relationship between minutes per game and draft position. However, this result may be sensitive to outliers and assumes linearity. (see "Limitations" below for more details).

Evaluating Standard Deviation of Targeted Variable (Minutes Per Game)

- Mean Minutes Per Game: 18.71 (SD = 8.13)

  - On average, players play about 19 minutes per game, which is roughly half of a 48-minute NBA game.
  - The SD of 8.13 shows considerable variation—some players are role players, others starters.

Hypothesis 2

Null hypothesis (H_0): There is no positive effect of draft position on VORP

$$H_0 : \beta_{\text{pick\_num}} \leq 0$$

Alternative hypothesis (H_A): There is a positive effect of draft position on VORP

$$H_A : \beta_{\text{pick\_num}} > 0$$

Since the p-value is less than 0.05, we reject the null hypothesis, suggesting that there is in fact a **positive** relationship between VORP and pick number. However, this result may be sensitive to outliers and assumes linearity, therefore requiring further analysis (see "Limitations" below for more details).

Evaluating Standard Deviation of Targeted Variable (VORP)

- Mean VORP: 3.912 (SD = 9.902)
  - The average Value Over Replacement Player (VORP) across players is about 3.9, but the large SD (9.9) means there's wide variability—some players significantly outperform replacement level, while others may have little to no impact.
  - This spread suggests that draft position alone likely doesn't explain all the variation in value.

Interpretation and conclusions

Descriptive Analysis

The descriptive statistics tell us:

Mean draft position (pick_num): 26.995 (SD = 16.276)

Mean VORP: 3.912 (SD = 9.902)

Mean minutes per game: 18.71 (SD = 8.13)

The statistics for mean draft position us that, on average, players are picked near the late-second round (pick ≈ 27) but with a wide spread (SD ≈ 16 picks). The career VORP of players also varies substantially (SD ≈ 9.9, versus a mean of ≈ 3.9). Players in the dataset averaged 18.71 minutes per game, with a standard deviation of 8.13. This means that while the typical player saw modest playing time, the spread in minutes was wide. Some players see significant court time, while others barely play, which reinforces the need to test whether draft position correlates with minutes played.

Hypothesis 1

We started our analysis on hypothesis 1 by calculating the correlation value between minutes played per game and draft number. Our result was an R value of −0.5738117, showing that there is a negative correlation. As drift pick number increases and players are picked later in the draft, their minutes played per game decreases. This was an early trend that we saw in the data without linear regression.

We then proceeded to fit a linear regression model to further test our hypothesis. Some interesting results from our linear regression model was its coefficients. The results show that for each increase in draft pick number by 1, a player is expected to play ~0.296 fewer minutes per game. This supports our hypothesis that earlier picks/players with smaller pick number are given more playing time per game, most likely due to higher expectations/perceived talent level.

Furthermore, the r value sheds light on the R^2 value, which would be roughly $(-0.5738117)^2 = 0.329264$, which is around 0.33. In the context of our hypothesis and research question, this tells us that about 33% of the variation in minutes played per game is explained by the draft pick number. The other 67% of the variation however is due to other factors, such as injuries, player development, team fit, etc.

To statistically test the significance of this relationship, we used a hypothesis test where the null hypothesis was that there is no linear relationship between draft position and minutes played per game. The p-value from our regression output was 8.32e-101, an extremely small value; far less than both 0.05 and even 0.001. This indicates that the observed slope is highly unlikely under the assumption that the null hypothesis is true.

In essence, the slope we observed which represents the relationship between draft position and minutes per game, falls far in the tail of the null distribution. This tells us that later draft picks consistently receive fewer minutes, while earlier picks generally play more. We can therefore reject the null hypothesis and conclude that there is a statistically significant negative relationship between draft position and playing time in the NBA. This likely reflects teams' greater investment and confidence in early picks, who may be expected to contribute immediately to team performance.

Earlier draft picks consistently earn more minutes because teams have already "put their money on them" and want a quick return on that investment—an expectation bias in action. But since ability isn't determined by draft slot, this pattern shows how first impressions can gatekeep early opportunities, underscoring the need to ensure that later picks get fair chances to prove themselves.

However, the moderate R^2 value of around 0.33 also reminds us that draft position is far from the only factor influencing playing time. A substantial portion of the variation (about 67%) is likely due to other real-life considerations. These might include injuries, team needs on different positions, player development progress, coaching preferences, system fit and other off-court factors such as work ethic, attitude, and team chemistry.

Understanding this dynamic is useful for analysts and decision-makers when evaluating player development, and for players and agents when considering long-term growth beyond draft position.

Hypothesis 2

Over the course of our analysis, we've found clear, statistically robust evidence that later NBA draft picks—those with higher pick_num values—tend to generate lower lifetime VORP. The moderate negative correlation of −0.3197 between pick_num and vorp_val tells us that roughly 10% (r^2 ~ 0.10) of the variation in VORP across players is linearly tied to where they were selected in the draft. Furthermore, looking at the slope and intercept, we have: Intercept: If the pick_num is 0, we expect the vorp_value to be 8.35 points, on average. Slope: For every one point increase in the pick_num (i.e. draft pick number), we expect the VORP value to decrease by 0.168 points, on average.

The p-value (<10^-21) and the analysis we have conducted provide overwhelming evidence that the negative relationship is not due to chance. Our sample spans 2004–2024, covering hundreds of players, so we're highly confident in the direction and significance of this effect—even as we acknowledge that draft position alone explains only about 10 percent of VORP variance.

Now, focusing on interpreting the results in the wider context of real-life application, we find:

Teams invest heavily in lottery and top-10 selections because those slots deliver, on average, the highest VORP returns. Our model implies that the difference between the 1st and 30th pick is roughly 30 × 0.168 ≈ 5 VORP units, which can translate to multiple wins above replacement over a player's career.

Our finding—each draft slot costs ~0.17 VORP—means that moving up 10 picks can "buy" about 1.7 extra wins in a full season. In today's NBA, a single win can translate to $500K–$1 M in incremental revenue via ticket prices, local TV contracts, and merchandise sales. Front offices routinely monetize these win estimates when trading picks: acquiring a pick that reliably delivers +2 VORP over a later slot justifies surrendering assets valued at roughly $2 M. Moreover, expected VORP guides rookie scale extensions and informs draft-night decisions under salary-cap constraints.

That said, draft position explains only ~10 % (R^2) of VORP variance—teams still lean on in-person scouting, medical evaluations, and character assessments to manage the 90 % of performance driven by non-box-score factors. Recognizing both the power and limits of VORP helps franchises blend quantitative projections with qualitative insights for better draft outcomes.

In sum, our findings validate long-standing draft theory—earlier picks yield more value—but also highlight the substantial "noise" around individual outcomes. As a result, teams that can consistently identify and cultivate talent beyond pure pick order stand to gain a competitive edge in NBA.

Limitations

A significant limitation of the statistical methods used in this analysis is the reliance on linear regression models. Linear models assume a constant, linear relationship between variables, which may not fully capture the complexities of player development and performance in the NBA. Linear regression does not account for potential interactions between draft position and other factors, such as team needs, player development opportunities, or injury history, which could influence a player's minutes played and overall performance. For example, it can be seen through the work we completed in hypothesis 1 when calculating the R^2 value. Our R^2 for hypothesis 1 was 0.33, meaning that draft pick number could be used to explain about 33% of the variation in minutes played per game. However this is a great limitation because it leaves around 67% of the variation unexplained, or unable to be explained by draft pick alone.

The linear regression model may also be sensitive to outliers. Extreme values in the data can disproportionately affect the slope. For instance, if a high draft pick plays very few minutes due to injury or other circumstances, or if a late pick becomes a star and plays many minutes, these cases could distort the overall trend.

Overall, this approach may result in oversimplifying the true dynamics between draft position and player success. More sophisticated models, such as nonlinear regression or machine learning algorithms, could offer a better fit and provide a better understanding of how draft position interacts with various other factors in NBA.

Acknowledgments

We thank our TA for her continuous feedback at each project milestone, which helped us structure our final report effectively.

We also want to acknowledge that our NBA draft dataset (nba_draft_data) was scraped from SportsReference.com.