# Optimal Roster Construction

Alex Gaynor

June 2024

## 1 Introduction

Every team spends time searching for the best draft pick, the best player to sign in free agency, or the best player to trade for. However, it is just as important to take a step back and look at the team as a whole. Every team is built in a certain way, such as the Heatles with 3 veteran stars surrounded by cheaper role players, or the 2-headed monster of the 2024 Boston Celtics with more premiere role players. In this report, I will be presenting the best way to construct a roster, based on data from the 2017-18 season through the 2023-24 season.

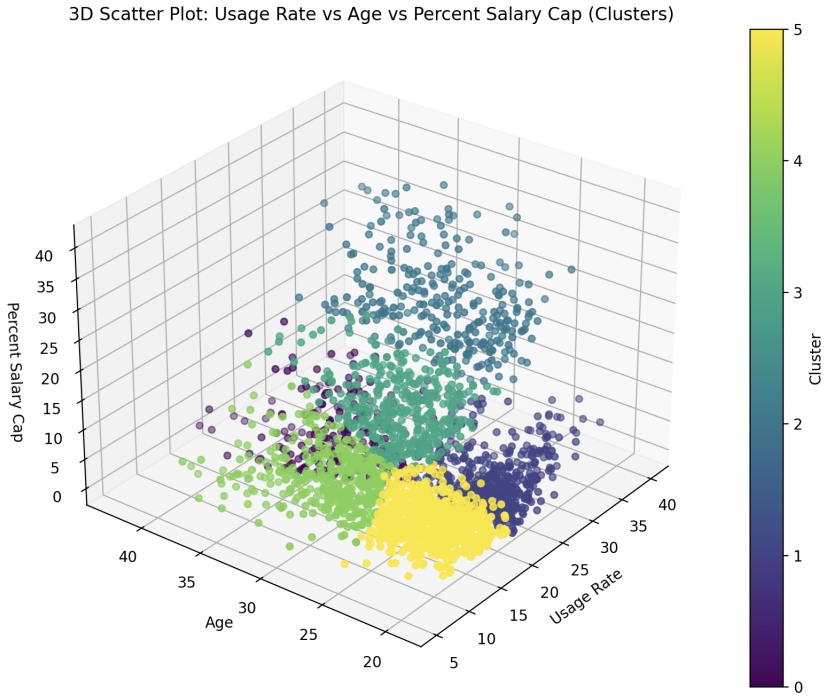
## 2 Data

To get the data necessary for this project, I scraped spotrac.com for all salary data. Then, I used basketball-reference.com for usage rates, ages, and team stats. I only used data beginning in the 2017-18 season because of the 2017 Collective Bargaining Agreement (CBA), which redefined the NBA salary cap. After I scraped the necessary data, I combined them into one file that featured each player's age, usage rate, salary, and team in each season from 2017-18 to 2023-24. Then, to get the final data, I only took the top 12 salaries from each team, as this is the typical rotation seen throughout an NBA regular season.

## 3 Clustering Players

I chose to cluster each player's season based on three factors: **age**, **salary**, and **usage rate**. First, I chose age as this is an important determining factor in the player's timeline. It is important to differentiate between two similar contract values, such as a year of a rookie's contract or a veteran minimum. It is a common theme for teams to be called too young (2024 OKC Thunder), or too old (2022 LA Lakers). Next, salary was used as it is the biggest restriction on roster construction. A team can only spend a certain amount of money during a season. To account for the increasing salary cap over the last few years, I used the percentage of salary cap to measure the player salary. Finally, usage rate is the final variable because it is important to see which players are actually

making an impact on the floor. However, I did not want the bias of players being on winning teams versus losing teams. One must differentiate between a highly paid player traded to take up salary and a highly paid player signed to be an all-star. I implemented K-Means Clustering to group every player from 2017-18 to 2023-24 into six clusters based on these three factors and got the following result:




The algorithm split up the players into distinct groups, which can be defined based on their cluster averages:

| Cluster | Age | % of Salary Cap | Usage Rate |
|---|---|---|---|
| 0 | 32.94 | 6.06 | 20.53 |
| 1 | 22.34 | 4.50 | 24.07 |
| 2 | 28.47 | 27.00 | 27.96 |
| 3 | 27.63 | 15.15 | 19.80 |
| 4 | 30.21 | 5.98 | 13.43 |
| 5 | 23.01 | 2.95 | 15.85 |

Using these figures, I set names for each of the clusters: "Cheap, Old, Medium Usage", "Cheap, Young, High Usage", "Expensive, Prime, High Usage", "Highly Paid, Prime, Medium Usage", "Cheap, Old, Low Usage", "Cheap, Young, Low Usage"

# 4 Model

After clustering each of the players, every team was defined based on how many players they had from each cluster. This was then combined with their win percentage and playoff wins in that season. I used a linear regression model to fit the number of players in each cluster to the success of that team. I changed the constraint of the percent of the total salary of the team as the standard salary cap (100%), the luxury tax threshold (121.5%), and the highest of any team in the dataset (2023-24 GS Warriors, 151.38%) for the 2023-24 season. Also, I bounded the model to only allow a maximum of 5 players for each cluster when determining the optimal distribution; it usually isn't feasible to have more than this amount for any given cluster. Lastly, the model is limited to 12 players on the team as this was the number that each team was counted for.

# 5 Results

When the amount of players in each cluster was fit with the regular season win percentage of the team, the model had a Training $R^2$ of 0.349 and Test $R^2$ of 0.330. These are solid numbers for predicting the win percentage of an NBA team. However, when looking at playoff wins, the Training $R^2$ and Test $R^2$ fell to 0.060 and 0.042, respectively. This is because of the high variance of the NBA playoffs and the increased importance of coaching and matchups, among other factors.

The optimal regular season team under the salary cap had 2 "Expensive, Prime, High Usage" players, 5 "Cheap, Old, Low Usage" players, and 5 "Cheap, Young, Low Usage" players. This would be a 2-star team surrounded by cheap role players split by age. The model predicted this team to have a 0.600 winning percentage or a 49-33 season. The results changed when the analysis was run for the playoffs, swapping the 5 "Cheap, Old, Low Usage" for 5 "Cheap, Young, High Usage" players. It created a much younger team, as many times older players wear down and do not perform as well in the playoffs. Also, younger players tend to be more versatile. This team was predicted to have 6.29 playoff wins.

Next, when looking at a team in the luxury tax, it is able to add another star. This optimal team has 3 of the "Expensive, Prime, High Usage" players, 5 "Cheap, Old, Low Usage" players, and 4 "Cheap, Young, Low Usage" players. When going into the luxury tax, the team was able to replace a rookie contract player with a star in their prime. This led to an increase in win percentage to 0.668 for a record of 55-27. The same thing happened when looking at playoffs, where the vet min players were flipped for rookie contract players. The added star added a playoff win for a total of 7.28 wins.

When given the leeway of going above the 2nd apron, the optimal team again traded a "Cheap, Young, Low Usage" player for an "Expensive, Prime, High Usage" player. This yielded a team of 4 stars, 5 older role players, and 3 younger role players. This team has a predicted 0.736 win percentage for a great 60-22 record. When looking at the respective playoff team, it added yet another star for 5 stars, 2 high usage rookie contracts, and 5 low usage rookie contracts. This team has a predicted 9.27 playoff wins.

# 6 Discussion

It is important to see examples of optimally constructed rosters in order to conceptualize. The most optimally built team when compared to the model's output from staying under the salary cap was the 2020-21 Phoenix Suns. They had 2 "Expensive, Prime, High Usage" players (Chris Paul and Devin Booker), 6 "Cheap, Young, Low Usage" players (Deandre Ayton, Mikal Bridges, Cam Johnson, etc.), and 3 "Cheap, Old, Low Usage" players (Jae Crowder, E'Twaun Moore, and Langston Galloway). This team was very well constructed as it featured a great core of star-power surrounded by solid cheap role players. This team ended 2nd in the West at 51-21 and lost in The Finals to the Milwaukee Bucks.

A great example of an optimal roster that reaches the luxury tax is the 2017-18 New Orleans Pelicans. The Pelicans had a big 3 of Jrue Holiday, Anthony Davis, and DeMarcus Cousins. They had many "Cheap, Old, Low Usage" players, such as Rajon Rondo, Solomon Hill, and E'Twaun Moore. The team also had a couple "Cheap, Young, Low Usage" players in Ian Clark in Cheick Diallo. The Pelicans made the playoffs at 48-34, but lost in the Second Round to the eventual champion, the Golden State Warriors.

The best example of a roster with the most amount of salary available was the 2022-23 Milwaukee Bucks. The all-stars, Giannis Antetokounmpo, Khris Middleton, and Jrue Holiday, were supported by a "Cheap, Old, Medium Usage" player (Brook Lopez) and "Highly Paid, Prime, Medium Usage" (Bobby Portis). They were surrounded by a bunch of "Cheap, Old, Low Usage" players (Jae Crowder, Grayson Allen, Pat Connaughton, Wesley Matthews, and Joe Ingles) and had a couple "Cheap, Old, Low Usage" players (MarJon Beauchamp and Jevon Carter). This team was well constructed, as they supported their core of stars with cheap role players that fit in well. The Bucks finished 1st in the East at 58-24, but lost in the first round to the Miami Heat.

In order to build an optimal roster, it is imperative to lock up 2-3 stars and surround them with cheap role players, both young and old. The data suggests moving away from "Cheap, Old, Medium Usage", such as 2022 Mike Conley and 2021 Kemba Walker, and "Highly Paid, Prime, Medium Usage" players, like 2023 Rudy Gobert and 2023 Gordon Hayward. It is important to maximize

salary allocation by giving more expensive contracts to the best players and cheap contracts to role players, eliminating the middle ground.

# 7 Conclusion

When looking at the best method of constructing a roster, data suggests that the most important part is maximizing the number of all-stars to give expensive contracts. Then, fill the rest of the roster with cheap role players: to focus on regular season success, find a combination of veterans and young players with lower usages to support the all-stars. Alternatively, if the team's goal is to succeed in the playoffs, pair low-usage young players with higher-usage young players to support the team's all-stars. While this is just the foundation of a team, a building is only as strong as its foundation.