NBA Data Exploration

Authors: Melanie Cooray, Vaibhav Gattani, Michelle Lin

Abstract

The access to NBA player and team tracking data has enhanced the dimensionality of basketball. With a simple goal of putting the ball in the basket, analyzing how well a player is playing has now been transformed into something much more complicated. Luckily, Principal Component Analysis (PCA) can help reduce the dimensionality of this data to make it much more understandable for those of us who are not basketball fans. Thus, we explore the data set through the technique of PCA to understand a player's performance and the correspondence to winning. We also transformed this to calculate a team's PC score and evaluate a team's performance across the NBA. After exploring the dataset, our goal was to accurately predict the relationship between a player's statistics and the position he plays in. There are five possible positions a player can play in: point guard, shooting guard, small forward, power forward, or center. To understand which statistics correlate to position, we used feature engineering and visualizations to depict what features do help and do not help in predicting one's position. We then created logistic regression models and random forest models to help classify the players based on these features.

Introduction

The National Basketball Association (NBA) has always valued keeping track of player statistics whether it be for the purpose of viewers or recruiters to evaluate player performances. On top of the team box score that tracks how many points each team has scored throughout the game, the NBA also keeps track of more granular player statistics such as players' points per game, steals per game, free throw percentage, and more. With these statistics, we want to see if we are able to deduce anything about a player's play style, such as their position. Hence, in this exploration, we will be delving deeper into the 2012-2018 seasons player box scores by (1) conducting general exploratory data analysis to understand what we have to work with, (2) data cleaning, (3) principal component analysis, (4) building a model to predict a player's position -- point guard, shooting guard, small forward, power forward, or center, and (5) testing our model.

Description of Data

We are conducting our data exploration on the box_player_scores.csv dataset. In this dataset, each row represents a player's statistics for a game that they played during the regular season from 2012-2018. Each column represents some type of statistic whether it be rebounds, blocks, assists, game result, and more. To work with each player's statistics, we group by each player's name, height, weight to account for players who may have the same name. In fact, we found 5 players in our dataset who had the same names but weren't the same players: Chris Johnson, Chris Wright, Elfrid Payton, Mike James, and Tony Mitchell. When looking at Elfrid Payton's data, we found that there seemed to be only one player, but his height and weight changed throughout his career. Thus, we changed his height and weight so it would remain consistent.

We also cleaned our data to only include players that have played at least 2 full games (96 minutes) and players with at least 10 points scored, to avoid players that can skew our classification because of a lack of game time.

We decided to only keep the following features for the rest of our exploration: points, assists, turnovers, steals, blocks, personal fouls, field goals attempted/missed/%, 2 pointers attempted/missed/%, 3 pointers

attempted/missed/%, free throws attempted/missed/%, offensive/defensive rebounds, team name and minutes played.

Exploratory Data Analysis

Principal Component Analysis

First, we conducted principal component analysis in an effort to reduce dimensionality and gain a better understanding of our data. In a player box score, we are given dozens of different statistics, many of which may be derived from each other or have some correlation. Through principal component analysis, we hope to better understand player archetypes and their correspondence to winning through top principal components that capture the most variance. Additionally, we hope to make sense and provide some interpretations of these top principal components. We will use the top principal components to help build our player position prediction model.

Our first principal component analysis looks at all games for all players. From the scree plot in Figure 1, the first principal component explains about 35% of the variance and the second principal component explains about 15%. When we plot PC1 against PC2 with jitter in Figure 2, we have way too many data points for us to deduce much from the visualization. However, games of players with large negative PC1 scores are all players that we recognize, such as Russell Westbrook, Kobe Bryant, and Anthony Davis, since they are star players of their teams. One outlier in the graph is Devin Booker on the far left, this must be due to his outstanding game against the Celtics on March 24, 2017 where he scored 70 points (Figure 6). We tried to color code each point by the game result to visualize the player's impact with the correspondence to winning; however, it doesn't seem to share any information.

In our second principal component analysis, we group the dataset by players and position to get their average PC score (not just for one game) when playing a certain position, so when we look at our scatterplot of PC1 vs PC2, we will have less clutter. In the scree plot of Figure 3, the first principal component explains about 70% of the variance and the second principal component explains about 15% of the variance. When we plot PC1 against PC2 with jitter in Figure 4, we have a much more digestible visualization. By color coding each point by the player's position instead of win/loss, we see that players with high PC1 values are renowned star players who start and play long minutes in games. Players with low PC2 values are point guards or shooting guards, and players with high PC2 values are centers or power forwards.In Figure 7, we can see that Andre Drummond, Carmelo Anthony, and Kyrie Irving all have PC1 values of 10 or more with the exception of Carmelo Anthony's PC1 as a shooting forward. On the other hand, we have players like Shavlik Randolph who have a small and negative PC1 value which reflects on his minimal contributions to his team. Hence, players with low PC1 values are benchwarmers who don't have as much contribution to their teams. We also see that Kyrie Irving, a point guard, has a small, negative PC2 value; Carmelo Anthony, a power forward, has a PC2 value, close to 0; and Andre Drummond, a center, has a large, positive PC2 value. Thus, PC1 seems to represent how much the player contributes to the team and PC2 represents a player's position/archetype. Something interesting we see with Carmelo Anthony since he plays both the power forward and shooting forward position is that he has a much lower PC1 value as a power forward which means that as a power forward, he makes less contributions. This makes sense because he generally plays as a shooting forward which is where most of his player statistics come from.

Transformations

After looking at the principal components for each player, we've understood a lot more about individual players and what impact they make/what position they play in. From this, we wanted to visualize the correspondence to winning, but as shown by Figure 1, there are too many players for a clean visualization and while players have an impact on winning, basketball is a team game. Therefore, the team wins and loses together, it is important to visualize our principal component in a team context. So, we utilized the formula to calculate the team PC score from a research report (Bruce), "The team PC score can be found by taking an average of the kth PC scores across the n players weighted by the minutes played throughout the season." Doing this transformation and mapping it to the win rate of each team (%) between 2012-2018 produced Figure 5.

From Figure 5, we see that the teams with a larger positive team PC score also tend to be the teams that have a higher win rate. We can see a likely positive linear relationship between the win rate and a team's PC1 score. However, there are outliers to this relationship for teams with PC scores between 2.5 and 3.5. One notable outlier is San Antonio Spurs, who won the championship in 2014, have a very high win rate but an average PC score. This could be down to the fact that the Spurs won games as a team rather than individual players scoring the majority of the points. This is also highlighted through Tony Parker, who led San Antonio in scoring in 2014, had the fewest points-per-game of any leading scorer on a championship team.

Description of Methods

Feature Engineering

Feature engineering became the crux of creating a good model. We standardized all our features that are dependent on games (points, assists, rebounds, turnovers, steals etc.) by dividing by the number of minutes played of each player. This is because the sum of assists made by one player can only be compared to other players if it was standardized. We also utilized shot percentages as features because they depicted how accurate players shoot from different areas of the court, even though these were provided, we did compute our own percentages to double check the values.

We also built a new feature that combines many player statistics into one overall feature called "efficiency." The idea is that it sums the positive features to create a standardized formula for player calculations. There are many formulas that are used to different sports analytics companies, but for the purpose of our exploration, we will be using the following formula to calculate efficiency:

$$Efficiency = \frac{points + rebounds + assists + steals + blocks + field\ goals\ made + free\ throws\ made}{number\ of\ minutes\ played}$$

Lastly, since the 'Team' column is currently categorical, we one-hot-encoded the values to be able to use them as features. Looking at the team distribution (Figure 10), the type of players are not evenly distributed among the teams, ensuring that the feature captures some sort of variance and would be useful in our model.

Building the Model

Before selecting our features, we split the data set into our x/y values and ensured the data in each of them was valid. When checking the possible positions in our dataset post-filtering players that made the cutoff, we found that there are 7 players who were marked as a G and 7 players who are marked as F. We removed these 14 rows because 14 players out of 1827 players shouldn't have too much of an impact on our model.

With the given dataset, we have a lot of features to select from. However, not all features will be useful, so we created a heatmap and pairplot to help us intuitively select features that we think will be predictive. As we can see from Figure 8, features such as blocks/assists, 3 pointers made/blocks, and assists/rebounds all have low correlation. This shows that we should think about using those features in our model. Features with correlation coefficients of 1 or -1 imply that there will be redundancy in using both features. However, features with some, but not perfect, correlation can still contain predictive power when looking at players who may be categorized at the border of two positions. We will further look into how player positions are distributed across features by creating a pairplot. This will tell us if there is any clear clustering or separation between the positions to see if they could be capturing any important differences. In Figure 9, we can see that there is clean separation in player positions for assists/offensive rebounds. Thus, these are features we definitely want to include in our models.

We repeatedly analyzed the features through these two visualizations to improve and create our models and normalized our x values before training/testing the model. In the end, we concluded Rebounds, Steals, and the 2P/FG/FT attempted/made features were not helpful in determining positions, so we dropped these features. The rebounds feature was not helpful because of overlap with the two other features offensive rebounds and defensive rebounds. Steals told us nothing about a players position type. The attempted/made features were not helpful in determining positions except for 3P because only certain positions shoot 3-pointers.

Since we wanted to classify players to a position based on their statistics, we first built a logistic regression model. It didn't perform as well as we had hoped, so we tried building a random forest model as we have seen it perform better as it benefits from ensemble learning and can weigh features differently. After playing around with hyperparameters, we decided to set the max depth of our random forest model to 9 to minimize overfitting but also ensure that our model trains correctly and the n_estimators to 100. We then ran further iterations of these model types including and not including our one-hot-encoded feature of teams and our principal component 2.

Summary of Results

In our first iteration, our logistic regression model had 75.9% training accuracy and 72.5% testing accuracy. Our random forest model performed at a 99.4% training accuracy and 74.4% testing accuracy. The random forest model's testing accuracy is significantly lower than its training accuracy which is a sign of overfitting, and it performs a little better than our logistic regression model.

In our next iteration of modelling, we included PC2 which was the principal component that we concluded to capture player position. The logistic regression model that includes PC2 does not perform any better than the logistic regression model without PC2 with a 76.0% training accuracy and a 71.8% testing accuracy. However, the random forest model performs better with PC2 with a 99.5% training accuracy and 75.1% testing accuracy. It seems like PC2 adds some value on top of the previous random forest model.

In our third iteration of modelling, we include teams one-hot-encoded along with the quantitative features. The logistic regression model performs with a 76.4% training accuracy and 70.9% testing accuracy. This is not better than either of our previous logistic regression models. Including teams in our random forest model also does not improve its performance (98.7% training accuracy, 73.8% testing accuracy).

It seems like Principal Component 2 helped with the random forest model, and one-hot-encoding did not. Nonetheless, in our last iteration of modelling, we created a logistic regression model that includes all the original features, Principal Component 2, and all the teams (one-hot-encoded). This logistic regression model still does not perform better than our first logistic regression model and has a training accuracy of 76.9% and a testing accuracy of 72.2%. Our random forest model with all original features, Principal Component 2, and teams (one-hot-encoded) also doesn't perform better with a training accuracy of 99.0% and a testing accuracy of 72.2%.

Figure 11. Table Depicting Model Performances

Model	**Training Accuracy (%)**	**Testing Accuracy (%)**
Logistic Regression	75.9	72.5
Random Forest	99.4	74.4
Logistic Regression w/ PC2	76.0	71.8
Random Forest w/ PC2	**99.5**	**75.1**
Logistic Regression w/ Team	76.4	70.9
Random Forest w/ Team	98.7	73.8
Logistic Regression w/ PC2 and Team	76.9	72.2
Random Forest w/ PC2 and Team	99.0	72.2

Overall, our second random forest model with PC2 had the best testing accuracy. However, all the random forest models seemed to overfit with high training accuracies and lower testing accuracies. Unfortunately, including teams didn't seem to improve our model. Even with our best performing random forest model, we are still given an accuracy that is relatively low compared to implemented models. This is because the problem of classifying player positions is very hard to solve as players don't tend to play basketball according to a fixed role but like to be versatile and play their best basketball. We will continue this conversation in the next section.

Discussion

Interesting Features

From our pairplot of player statistics to player positions, we found interesting correlations between positions and assists, rebounds and 3-point %. This makes sense because each position has a role and therefore, we are able to differentiate the positions by their different attributes e.g. centers are generally the tallest on the team and will normally be the players with many offensive and defensive rebounds. Additionally, our second principal component proved to be very interesting as it seemed to represent player archetype through player position.

Ineffective Feature

We expected that free throws should have been a useful feature because certain positions are fouled more often, so the number of attempts or the number of free throws made should be included in our model. Alternatively, we thought free throw % could help define which players are good shooters. Contrary to our belief, including any of the free throw statistics actually made our model less accurate so we did not end up using the feature and this is likely because being good at free throws is not relative to position.

Challenges and Obstacles

Given the limitations of this data, it was hard to come up with questions that would create meaningful and accurate models because the college data provided was not holistic of player performance. It only included

points, but had nothing about the rebounds, blocks, or assists. We originally wanted to predict how a player would perform in the NBA given their statistics in the NCAA; however it seemed as though the statistics we had for college basketball wouldn't be sufficient, so we had to pivot our exploration. We first came up with a predictor for the number of points a team scores in a game given the rest of the statistics but this just turned out to be a linear regression problem and we trained a model with 99%.

Limitations and Assumptions

In our analysis, we assumed that we are only predicting the position for players that have played at least 96 minutes (two full games) and have scored at least 10 points. We also assumed that there exists only one Elfrid Peyton and his height/weight did not change when he switched teams. Two limitations we faced: firstly, how accurate we could get our model because players don't always play like other players in their position, and secondly, having more complex data to work with like play-by-play data to make even better predictions.

Ethical Dilemmas

When we standardized the statistics played by each player, we chose to group by sum of games and divide by the minutes played. This makes some players who play less often appear much better than they are, while also making other players who play full games look mediocre overall. Alternatively, we could have grouped by sum and not standardized by minutes but this would lead to unethically considering players with lower minutes played for justified reasons like recently joined the NBA or have been injured.

Additional Data

If we had access to more specific statistics such as location on the court where a field goal was scored or how many touches per game each player gets, we would be able to improve our model. Centers are much more likely to be scoring closer to the basket while point guards and shooting guards are more likely to be shooting further from the basket around the 3 point range.

Ethical Concerns

In this analysis, we are predicting player positions based on their performances which may be inaccurate. The idea that we are conforming players to follow a certain expectation or a certain role in a team is not always justified. When we look at the bigger picture and ignore the statistics, the end goal in a basketball game is to outscore your opponent by putting the ball in the basket. The whole concept of positions and player roles is artificially created to group players of similar play styles and play to their advantages. But in the end of the day, every player is different and human, and thus, not every Center will conform to the way our model will predict the Center position.

Future Research

To further improve our model, we hope to incorporate stronger features in the future. From our research, we found a lot of past research papers that created attributes that could predict player archetypes such as FiveThirtyEight's CARMELO. Looking forward, we hope to include such features and build our models on more complex models such as the XGBoost. We would like to build future models on ones that rely on decision trees, such as an XGBoost model, given the successful performance of the random forest model.

References

https://arxiv.org/pdf/1511.04351.pdf
https://en.wikipedia.org/wiki/Player_tracking_(National_Basketball_Association)
https://projects.fivethirtyeight.com/carmelo/
https://nycdatascience.com/blog/student-works/predicting-nba-player-positions/
http://cs229.stanford.edu/proj2014/Luke%20Lefebure,%20Understanding%20Player%20Positions%20in%20the%20NBA.pdf

Appendix

Figure 1.

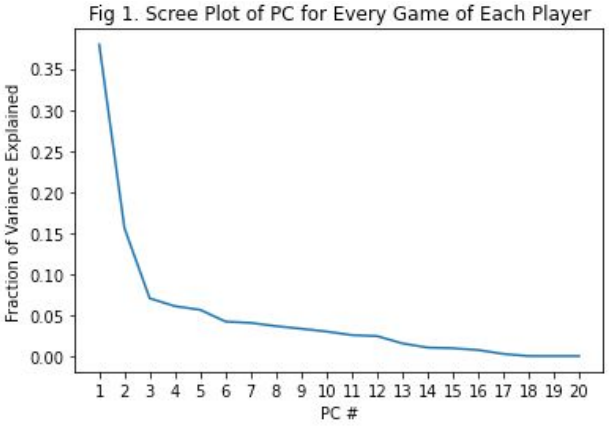

Figure 2.

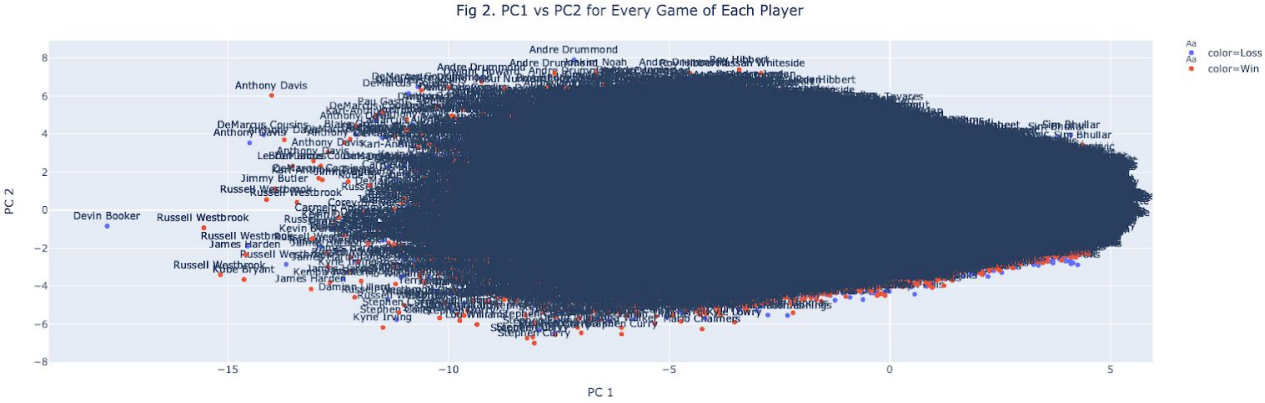

Figure 3.

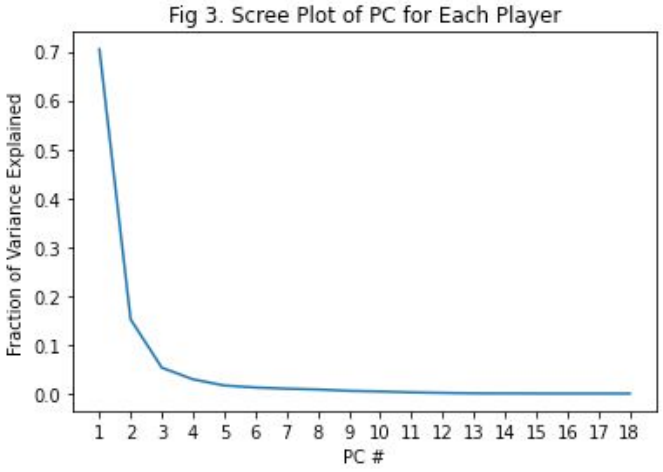

Figure 4.

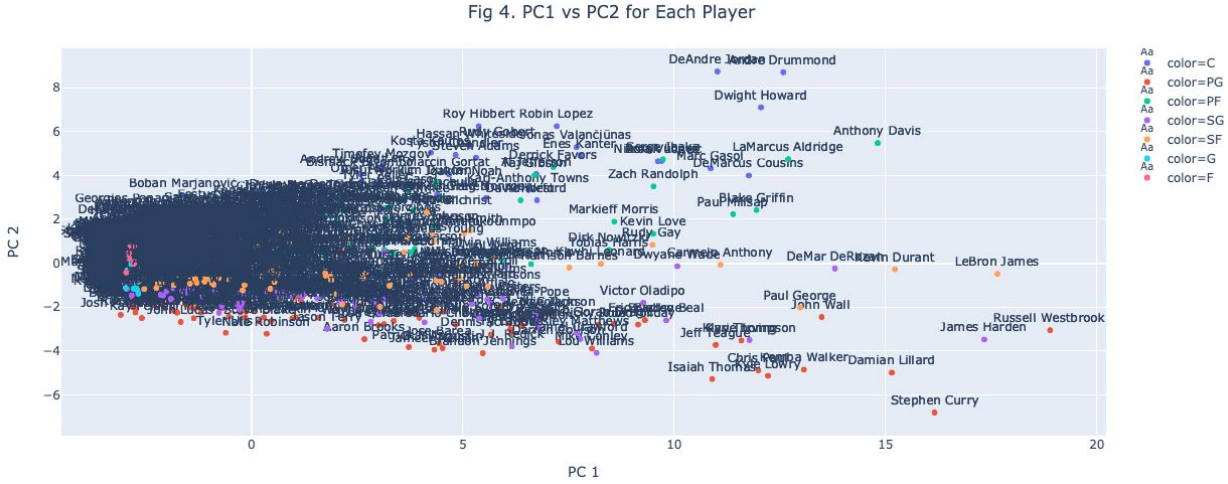

Figure 5.

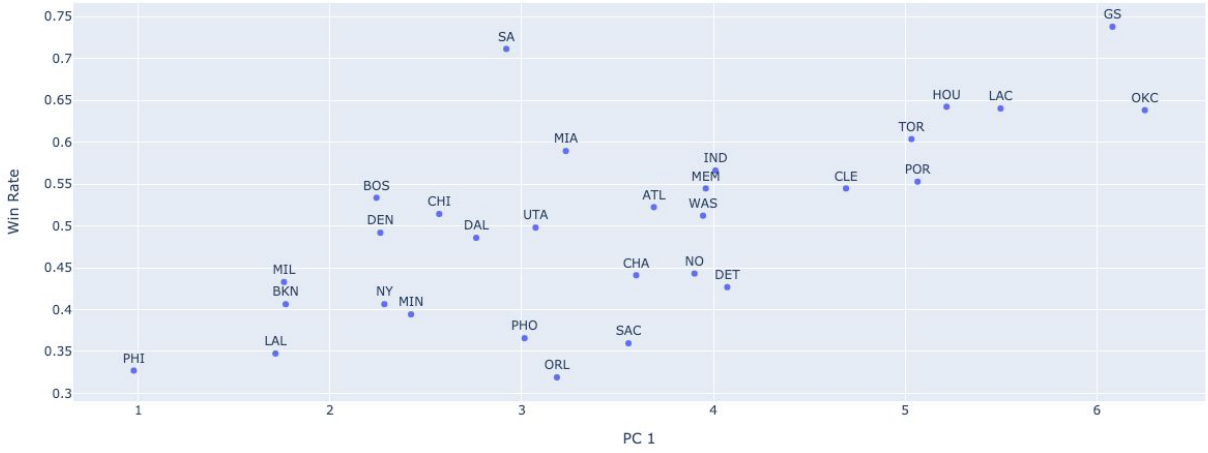

Figure 6.

Fig 6. Devin Booker Game	gmDate	playDispNm	teamAbbr	playPTS	playAST	playTO	playSTL	playFGA	playFGM	play2PA	play3PA	playFTA	playORB	playDRB	playMin
126352	2017-03-24	Devin Booker	PHO	70	6	5	3	40	21	29	11	26	2	6	45

Figure 7.

Fig 7. PC Players	playDispNm	playPos	playPTS	playAST	playTO	playSTL	playBLK	playFGA	play2PA	play3PA	playFTA	playORB	playDRB	pc1	pc2
0	Andre Drummond	C	6287	513	795	590	708	4922	4892	30	2188	2189	4007	12.493047	8.593446
0	Carmelo Anthony	PF	1109	92	87	40	42	1049	624	425	167	67	356	-1.027777	0.119985
3	Carmelo Anthony	SF	8360	1047	801	317	174	6740	4957	1783	2080	509	1852	10.989638	-0.086654
0	Kyrie Irving	PG	8735	2143	1006	513	123	6910	4804	2106	1767	269	1054	11.549991	-3.532042
0	Shavlik Randolph	PF	111	9	18	16	11	97	92	5	50	51	82	-2.640559	0.884852

Figure 8.

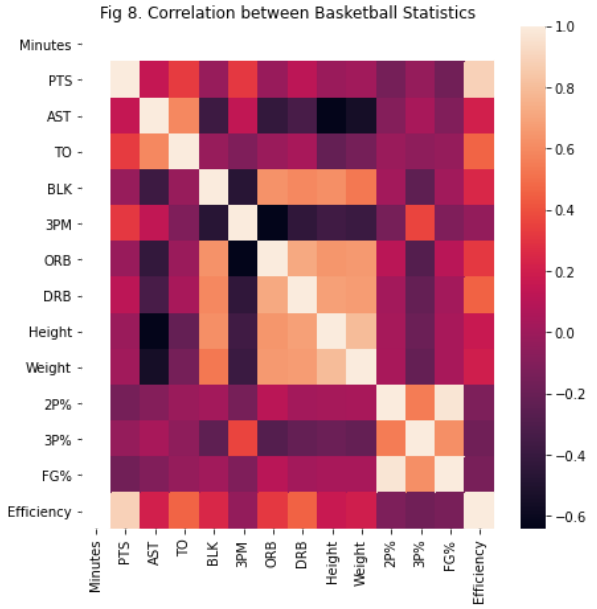

Figure 9.

Fig 9. Relationship Between Basketball Statistics for each Position

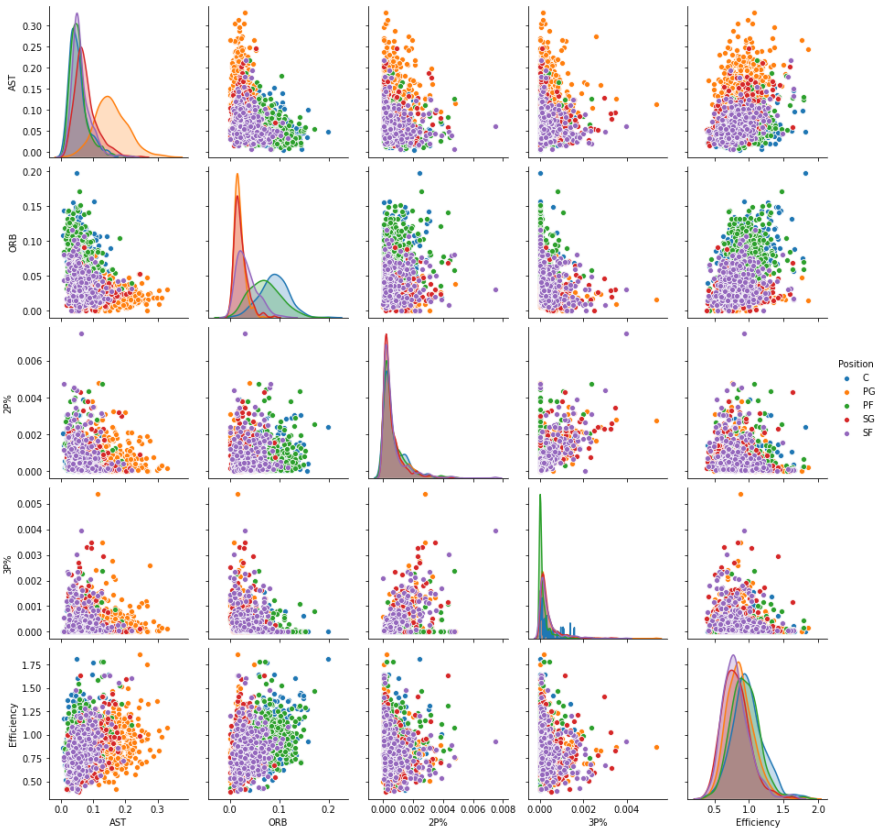

Fig 10.

Fig 10. Team Distribution