NBA Automated General Manager

DAEN 690 Capstone Project Report

3 May 2018

Adam Cook
Colin Bowers
Shruti Patil
Vaibhav Trivedi

Abstract

The National Basketball Association (NBA) is the premier basketball league in the world and the competition for the best talent is fierce. NBA teams are run by general managers, who attempt to win games, and ultimately championships, by signing and developing the best combinations of players. Based on over fifty years if evidence, simply signing the best players is not enough. Due to overlapping skill sets, some combinations of players fit together better than other. Additionally, the NBA has a salary cap system that prevents one or two team hoarding all the talent. The rules of the NBA salary system are complicated and governed by several factors. The goal of an NBA general manager is build the best combination of players subject the salary cap and other league rules.

This system seeks to identify what makes an NBA lineup successful. Most NBA team-building analysis is player and team-centric. These sorts of analyses ignore that the lineup is the basic building block of an NBA team. This approach uses NBA lineup data and player data to identify the individual player attributes and statistics that contribute to overall lineup success. Since NBA teams are ultimately a collection of lineups made up of players, this information can be used to guide personnel transactions. The final step of this approach is to apply the findings to a real NBA team and advise the Washington Wizards what personnel moves they should make during the upcoming 2018 NBA offseason, subject to the Wizards salary cap situation and what we estimate the available players will cost on the free agent market.

Table of Contents

1 Project Definition

1.1 Problem Space

Competition among teams in the National Basketball Association (NBA) is fierce and every team is on the lookout for the next competitive advantage. Teams are increasingly turning to mathematics and analytics to provide that edge. NBA teams are run by general managers, who must win games and compete for championships to keep their jobs. However, the league salary cap prevents a team from just signing the best players, even if a team has a clear way to determine who those players are. It is necessary for league decision makers to determine which basketball skills are worth paying for and which may be overvalued. Each NBA team has up to fifteen active roster spots and those sports are typically all filled. At any one time in a basketball game, there are five players on the court. This collection of five players is known as a lineup.

1.2 Analytics in Basketball

As with other sports, statistics have been around in basketball since the beginning of the sport. The progression of analytics in basketball can be divided into three eras that show how the sport has evolved.

1) Box Score Era – From the beginning of basketball until the 1990s, the primary statistics in the sport were those found in a common box score: points, rebounds, assists, blocks, steals, etc. These statistics still have real value, but with the rise of the Internet and increases in computing, they were supplemented with newer statistics.
2) Aggregation Era – Throughout the 1990's and 2000's, the Internet allowed a community of basketball fans to collaborate and combine box score statistics in a way to provide more useful data. These advancements also informed NBA teams and in fact several leading aggregation era analysts (Dean Oliver, John Hollinger) ended up working in the league. Important statistics from the aggregation era include player effectiveness rating, true shooting percentage, and net rating.
3) Tracking Era – Starting with the 2010-11 season, NBA teams begin using cameras hung in the rafters to collect data multiple times every second on the location of all 10 players on the court as well as the ball. This data is available to the league, the teams, and some media companies, but most of it is not publicly available. The ongoing race in basketball analytics is how to best leverage this tracking data. Important statistics from the tracking era include shots chart, quantified shot quality, and specific shooting breakdowns by shot type and situation.

This project uses attributes from all three eras, a practice which is becoming increasingly common in basketball analytics. Many of the player and lineup features we leverage are from the box score era, but we also leverage net rating (along with offensive and defensive ratings) and shooting data from the aggregation and tracking eras, respectively.

1.3 Solution Space

Our system seeks to identify what makes an NBA team successful. Most NBA team-building analysis is player and team-centric. These sorts of analyses ignore that the lineup is the basic building block of an NBA team. Basketball is played in units of five players from each team on the court at any one time and our approach is to analyze the game at that level. Our project uses NBA lineup data and player data from the 2007-08 season through the 2016-17 season to identify the individual player attributes and statistics

that contribute to overall lineup success. Lineup success is a nebulous concept, but we are using the offensive and defensive components of the statistic net rating to describe lineup quality. Net rating describes a team's point differential per 100 possessions. It can be divided into offensive and defensive components, describing how many points a team scores and allows the other team to score per 100 possessions. For this analysis, net rating has two advantages over win percentage, even though our goal is to increase a team's winning percentage:

1) Winning percentage is a binary statistic whereas net rating is a continuous one. Win percentage treats a 20-point win and a 1-point win the same whereas they are quite different in net rating. This lack of nuance makes win percentage a much noisier statistic than net rating.
2) Net rating can be applied at the lineup level in addition to the team level, unlike win percentage which describes a team outcome. Therefore, net rating is much more suited to this analysis.

Still the statistics are highly correlated. For the 2017-18 NBA season, over 91% of the variation in team winning percentage can be explained by the team's net rating.

Our project identifies what basketball statistics and attributes are most correlated with lineup success. To account for injury and minute disparities, all statistics are analyzed per 100 possessions. We then apply those factors to individual players to identify which players should be contribute to a successful lineup. This information can be used by an NBA decision maker to guide decision making. The final step of our project is to demonstrate the applicability of our approach by applying the knowledge gleaned from our lineup analyses to a real NBA team by advising the Washington Wizards what personnel moves they should make during the upcoming 2018 NBA offseason, subject to the Wizards salary cap situation and what we estimate the available players will cost on the free agent market. The Washington Wizards are already over the salary cap heading into the 2018-19 season, so they will only be able to add one free agent via the taxpayer mid-level exception (MLE). Our project will advise them on the best players they can realistically add using the taxpayer MLE. This approach is visualized in Figure 1.

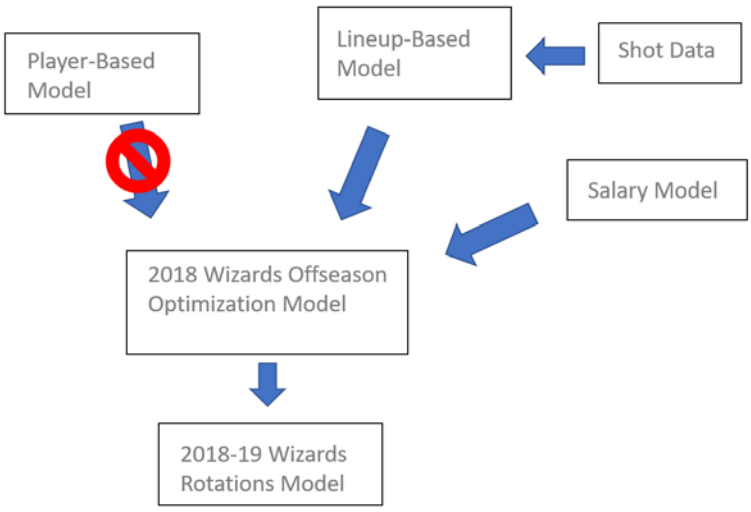


Figure 1. Visualization of Project Approach

On important consideration is that NBA general managers are at the pinnacle of their professional careers and are disinclined to take advise from any methodology/algorithm that they can't completely understand. Therefore, every step of our project must be transparent and explain why a recommended decision is being made in addition to what that decision is. Different NBA teams also have different comfort levels with statistical and analytical reasoning. However, analytical approaches do seem to work. The Golden State Warriors, who have won two of the past three NBA titles and set the record for most NBA wins in a season, are famously analytically driven. As are the Houston Rockets, who are led by former statistical consultant Daryl Morrey. These two teams currently have the best records in the NBA for the 2017-18 season and are the favorites to win the NBA title.

2 Datasets

2.1 Overview

Our approach takes advantage of the thoroughness and robustness of sports data to collect player and lineup data going back to the 2007-08 season and player salary data going back to the 2012-13. The player and lineup data come from stats.nba.com, which is the public facing API of nba.com, the league's website. This data was pulled using Python for the player and lineup data and R for the shot chart data. The salary data was built manually using data from ESPN.com and Basketball-Reference because no complete resource of historic salary values is readily available in collected form. This data does not include minimum salaries, which is a modeling feature because they would not factor into the model we're building.

Our other data sets provide information on the Washington Wizards current and future roster commitments and a list of the 2018 free agents. Both datasets come from spotrac.com, an online sports contracts resource.

2.2 Field Descriptions

The relevant fields from the data sets employed during this project are described below. A full data dictionary, describing all the fields, can be found in Appendix C.

Lineup Data – This data comes from stats.nba.com and was extracted using Python. It goes back to the 2007-08 season. Figure 2 shows the distribution of net rating, offensive rating, and defensive rating for lineups that have played more than 100 minutes together. Offensive rating describes how good a team is at scoring points, so a good offensive rating is above 100. Conversely, defensive rating describes how good a team is at stopping their opponent from scoring, so a good defensive rating is below 100. A good net rating is positive since quality teams score more points than they allow their opponents.

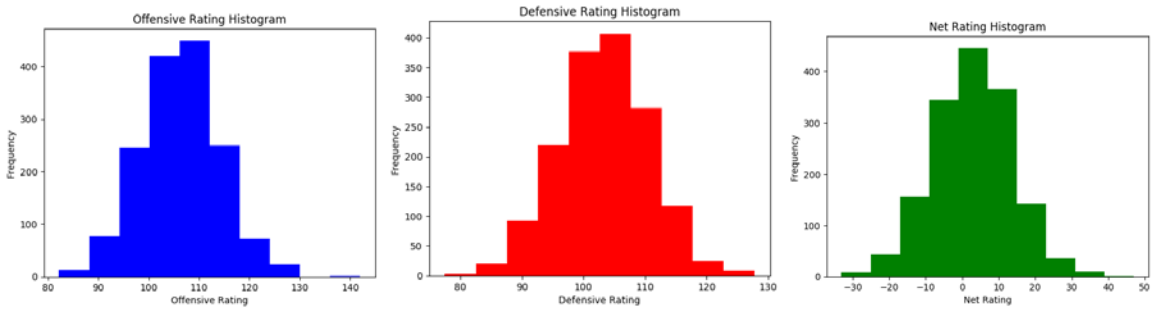


Figure 2. Net, Offensive, and Defensive Rating for Lineups that have played >100 minutes together

NBA teams typically play their best lineups, so the 100-minute filter should bias this sample towards above average lineups. It is interesting that both offensive and defensive ratings have an average higher than 100, suggesting that NBA teams value offensive performance over defensive performance.

- Group ID – The 5 players present in the lineup are presented as a list of their player IDs. This field is used as the key to pair this data with individual player data.

- Group Name – Similar to the Group ID, this field is a list of the 5 player names in the lineup. It serves no analytical purpose but can be used to identify the lineup during exploratory data analysis
- Minutes – This field contains the number of minutes the lineup has played on the season, rounded to the nearest whole minute. To restrict the noise-ness of our data, we only examined lineups with more than 100 minutes played together.
- Offensive Rating – This field contains the offensive rating for the lineup in question. Offensive rating is defined in the lexicon of this report and served as one of our two dependent variables.
- Defensive Rating – This field contains the offensive rating for the lineup in question. Offensive rating is defined in the lexicon of this report and served as one of our two dependent variables.
- Net Rating - This field contains the net rating for the lineup in question. Net rating is offensive rating minus defensive rating and is a measure of overall lineup value.

Player Data – This data comes from stats.nba.com and was extracted using Python. It goes back to the 2007-08 season. Unless otherwise note, the data is on the average for the season on a per game basis.

- Player ID - This field contains the Player ID of the player in question. It is the key for the database.
- Player Name - This field identifies the name of the player in question.
- Season - This field identifies the season in which the player's statistics took place.
- Steals - This field identifies the number of times per game during the season in question that the player in question stole the ball from the other team.
- Blocks - This field identifies the number of times per game during the season in question that the player in question blocked a shot by the other team.
- Personal Fouls Allowed - This field identifies the number of times per game during the season in question that the player in question was fouled by a player on the other team.
- Assist Ratio - This field identifies the percentage of a player's possessions during the season in question that end in an assist.
- Defensive Rebound % - This field identifies the percentage of defensive rebounds that the player in question grabbed out of all the available defensive rebounds while they were on the court.
- Rebound % - This field identifies the percentage of rebounds that the player in question grabbed out of all the available rebounds while they were on the court.
- True Shooting % - This field identifies the player in questions Field Goal Percentage, accounting for the increased value of 3-pointers as well as Free Throws. It is calculated via: Points/ [2*(Field Goals Attempted+0.44*Free Throws Attempted)].

- Defensive Win Shares - This field identifies the overall amount that a player's defense contributes to his team winning.
- % of Team Points - This field identifies the percentage of the team's points that the player in question had while they were on the court.
- Player Height – This field identifies how tall the player is, in inches.

Shot Chart Data – This data comes from NBA.com and was extracted using R. It was collected by the NBA using SportsVU, a camera system that is hung from the rafters in NBA arenas and collects data at 25 frames per second. This dataset is the largest single file we have in our database at nearly 500 MB. To ease analysis, we've summarized the data by player and season for 6 zones of the court: Above the Break 3, Right Corner 3, Left Corner 3, Mid-Range, In the Paint, and Restricted Area.

- Game ID – This field contains the game ID in question.
- Player ID – This field contains the nba.com Player ID of the player in question.
- Team ID –This field the team ID for the player in question.
- Period – This field contains which quarter of the game the shot took place.
- Time Remaining – This field contains the amount of time left in the quarter is a combination of Minutes Remaining and Seconds Remaining.
- Event Type – This field contains whether the shot in question was Made or Missed.
- Action Type – This field contains whether the shot was a Jumper or Layup.
- Shot Type - This field identifies whether the shot was a 2 or 3-pointer.
- Shot Zone – This field identifies which zone of the court the shot took place. The six zones of the court are: Above the Break 3, Right Corner 3, Left Corner 3, Mid-Range, In the Paint, and Restricted Area.
- Shot Distance – This field identifies how far from the goal the shot took place from. It can be one of several categorical values describing distance.
- Shot Location x – This field identifies the location of the x-coordinate of a shot on an xy-plane.
- Shot Location y – This field identifies the location of the y-coordinate of a shot on an xy-plane.

League Salary Data – This data was extracted manually from NBA.com and basketball-reference.com to provide information on NBA salaries over time. It goes back to the 2012-13 season and does not include minimum contracts.

- Player ID – This field contains the nba.com Player ID of the player in question. It was matched to the contract list using R and is the key for the database.
- Player Name –This field identifies the name of the player in question.

- Salary Value – This field contains the one-season monetary value of the contract in the year it was signed. Contract values change slightly over time (typically increasing) but this value is adequate for the scope of modeling in this analysis.
- Rookie Scale - This field identifies whether the contract in question was a part of the rookie scale. The rookie scale structures the contract values for younger NBA players and must be analyzed separately from non-rookie scale contracts.
- Percent of Maximum - This field contains the percentage of the maximum salary that a player could sign compared to what they did sign. It was derived using R based on the contract year, salary value, rookie scale, and experience and incorporated the NBA's salary cap rules.

Wizards Salary Data – This data was extracted manually from sportrac.com and contains information on the Washington Wizards current and future salary situation

- Player ID – This field contains the nba.com Player ID of the player in question. It was matched to the contract list using R and is the key for the database.
- Player Name – This field identifies the name of the player in question.
- 18-19 Salary – This field contains the amount of salary the player is paid during the 2018-19 season.

2018 Free Agents Data – This data was extracted manually from sportrac.com and contains information on which players will be free agents during the 2018-19 off-season.

- Player ID – This field contains the nba.com Player ID of the player in question. It was matched to the contract list using R and is the key for the database.
- Player Name –This field identifies the name of the player in question.
- Experience – This field contains the years of NBA experience of the player in question.
- Free Agent Type– This field can be either Restricted or Unrestricted and impacts the type of salary that the player in question can sign for.

2.3 Data Context

The data we've collected should paint as complete a picture as possible of how players interact while on the court together, how that interaction and resulting performance impacts team quality, and allow for those insights to be used to provide player acquisition recommendations to a specific team. It is important to clarify that the picture is not totally complete because individual teams and the NBA collect numerous measurement and develop analytics that are not publicly available. Moreover, the number of basketball statistics is seemingly endless, and our project scope required that decisions be made about which ones to use. Despite this, we feel that the data we've collected should provide the opportunity to develop thorough insights about the context of performance in basketball and how that context should impact team building.

2.4 Data Conditioning and Fusion

The NBA player and lineup data are readily accessible using standard RESTful endpoints with very little data conditioning required. Teams, Players and Lineups all have unique identifiers making fusion amongst sources directly from the NBA straightforward. Additional data for salary and contract information was manually downloaded and also did not require data conditioning. Free agent data was manually extracted but did not contain a common player ID to fuse with the NBA stats data. To approach this, we choose to use the name and team as a unique key to join the two datasets.

2.5 Data Quality Assessment

Below is a summary of the overall quality of our data sets. Table 1 describes each individual dataset, with green meaning no issues and yellow meaning potential issues. There are no large issues with our data quality.

- Completeness: Our datasets are complete in the information they provide. There is a nearly endless universe of basketball statistics, even publicly available ones, and we had to make some choices about what data to collect. This decision was made with capturing the totality of player performance and interactions through our data and if we have any issues, the stats.nba.com API presents an easy way to fill in any gaps. As noted below, it is possible the 2018 free agents list could change slightly due to changing conditions within the league.

- Uniqueness: Due to high quality of our data sets and based on exploratory data analysis, all our data sets are unique. We do not have repeating entries for the same player in the same game or season (depending on the data set).

- Accuracy: All our datasets come from enterprise websites that are in the business of posting accurate and complete data. This is particularly true of the stats.nba.com data since the NBA is the system we're modeling. Our 2018 free agent list may not be accurate because the list of impending free agents can change. Players are eligible to negotiate extensions, which would remove them from the list. To account for this, we checked the list before finalizing the results.

- Atomicity: Our data is mostly atomized. The exception is some of our player, shot chart, and lineup statistics are divisible and colinear. However, the goal of the data exploration and dimensionality reduction phases of the analysis is to reduce the collinearity of input variables of the model. Any final inputs to the lineup model that are not atomized will be that way by design. The other salary and optimization model will use entirely atomized values unless otherwise noted.

- Conformity: All our data scores very highly with regards to conformity. Most of our data is in the same format because it came from the same source (stats.nba.com). The data that did not come from this source was easily parred with the stats.nba.com dating using player names and team as a unique identifier. All our data uses standardized formats.

- Overall Quality: One of the benefits of using sports data is the high quality and completeness of the data. As noted above, all of our datasets are complete, and they are all accurate because

they all come from enterprise websites that are in the business of posting accurate and complete data. All the data used in this analysis is of a very high quality.

	Player Statistics	Shot Chart Statistics	Lineup Statistics	League Salary Statistics	Wizards Salary Statistics	2018 Free Agent List
Completeness						
Uniqueness						
Accuracy						
Atomicity						
Conformity						
Overall Quality						

Table 1. Data Quality Assessment by Data Set

2.6 Other Data Sources

There is additional information on players and lineups on stats.nba.com that we did not utilize for this analysis. This decision was partly a scoping one and partly since this unused data is similar and, in most cases, colinear to data we chose to use. Competitors to nba.com, such as ESPN, also have their own proprietary data we could have used. There are also new data sources that the NBA teams are using that are not publicly available. Finally, our analysis focused on lineup data rather than team data (except in the context of building the Wizards). This distinction is a specific hypothesis of our approach and is based on the theory that basketball should be analyzed primarily at the player and lineup level rather than the player and team level.

3 Analytics and Algorithms

3.1 Overview

Our overall modeling and analysis approach is shown below in Figure 3. This approach was refined over the course of the project. It involves multiple rounds testing and removing insignificant factors and testing for collinearity. Each step also features a sanity check to ensure that our findings are consistent with expectations or if the findings are surprising, they are defensible. For all our modeling, we used the 2007-08 season through 2015-16 season data as training data and the 2016-17 data as test data.

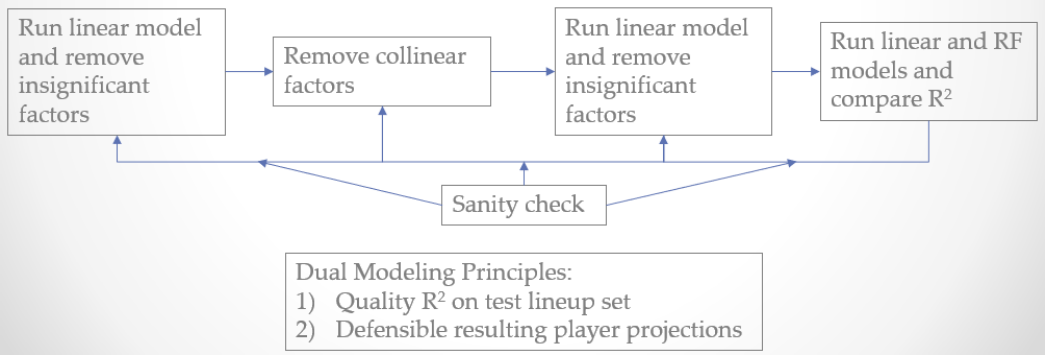


Figure 3. Overall Modeling Approach

3.2 Player-Based Model

As described in the Data section of this report, the datasets employed in this analysis contain many features. Specifically, we've collected many player statistics and attributes that can be modeled against offensive and defensive rating to identify the key characteristics when building a lineup. However, the number of possible features far exceeds the number that can be employed in a well-constructed model. Therefore, feature reduction is an important part of the lineup model portion of our approach. We approached feature reduction in two ways: an expertise-driven approach and principal component analysis (PCA).

3.2.1 Expertise Driven Feature Reduction

The expertise-driven approach used both statistical and basketball knowledge to identify the most analytically valuable set of attributes. The first step was to divide the attributes into like clusters, as shown below in Figure 4.

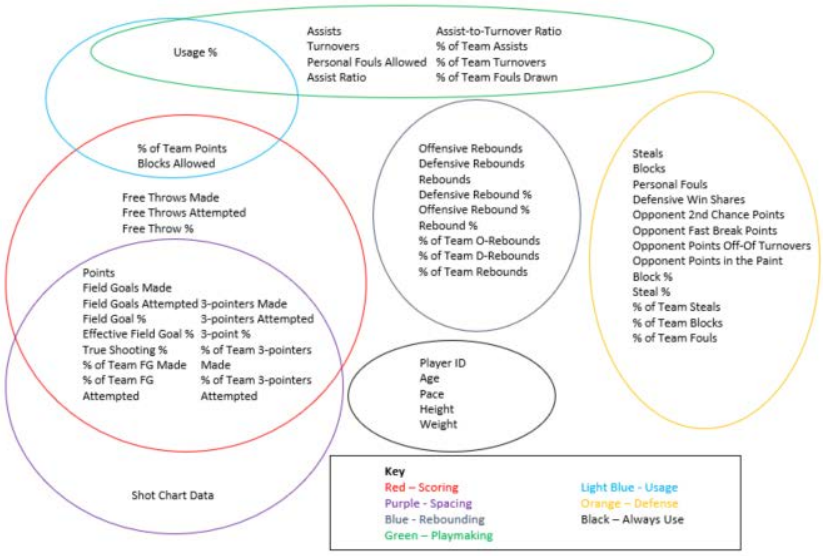


Figure 4. Feature Clustering for Expertise-Driven Feature Reduction

The scoring, spacing, playmaking, and usage clusters were analyzed against offensive rating since they are offensive components of basketball. The defense cluster was compared against defensive rating. Rebounding was compared against both offensive and defensive rating since it is both an offensive and defensive component of the game. Each feature was compared using the mean, the maximum, the minimum, and the standard deviation of the five players in the lineup. For each statistical cluster, the most significant one or two attributes for each cluster were identified through trial-and-error testing. In certain cases, one attribute was chosen over another because the correlation was more likely be causation based on basketball knowledge. These attributes were combined to make the lineup models for offensive and defensive rating. An example of the correlation matrices employed for this feature reduction is shown in Figure 5. This correlation matrix is for the Scoring attribute. In this case, none of the remaining attributes is very heavily correlated with one another.

	0	1	2	3
0	1	0.239513	-0.0966284	-0.00896038
1	0.239513	1	-0.346288	0.00789329
2	-0.0966284	-0.346288	1	-0.075617
3	-0.00896038	0.00789329	-0.075617	1

Figure 5. Correlation matrix for Scoring attribute

The expertise-driven player-based model ultimately lead to the attributes shown in Table 2. As stated elsewhere, the modeling for offense and defense was done separately.

Offensive Feature	Defensive Feature
True Shooting %	Defensive Win Shares
Points %	Steals + Blocks
Field Goals Allowed	Rebound %
Personal Fouls Drawn	Height
Assist Ratio	
Defensive Rebounds %	

Table 2. Features used in player-based modeling

3.2.2 PCA Based Feature Reduction

In addition to expertise-driven feature reduction, we also explored how principal components analysis could be used to model the data. For this analysis, the offensive data set had 56 variables and the defensive data set had 31 variables. These variables are fully described in Appendix C. PCA was able to reduce the number of variables significantly to 8 for both offense and defense. These variables explain 80% of the variance in data. Figures 6 and 7 show the importance of the individual components for offensive and defensive data.

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8
Standard deviation     3.680439 3.3026061 1.72812777 1.46785095 1.17679791 1.07144097 0.96829837 0.92784797
Proportion of Variance 0.356464 0.2870318 0.07859015 0.05669964 0.03644351 0.03021015 0.02467373 0.02265531
Cumulative Proportion  0.356464 0.6434958 0.72208590 0.77878555 0.81522906 0.84543921 0.87011294 0.89276825
```

Figure 6. Offensive Data PCA

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8
Standard deviation     3.3324630 1.7033520 1.44823116 1.21492214 1.04533624 1.01008576 0.92824879 0.86653553
Proportion of Variance 0.4627212 0.1208920 0.08739056 0.06150149 0.04553033 0.04251139 0.03590191 0.03128683
Cumulative Proportion  0.4627212 0.5836132 0.67100380 0.73250529 0.77803562 0.82054701 0.85644891 0.88773574
```

Figure 7. Defensive Data PCA

The number of components to ultimately use can be determined using a scree plot, as shown in Figures 8 and 9. In this case, the offensive data is best explained through three components and the defensive data is best explained through 2 components

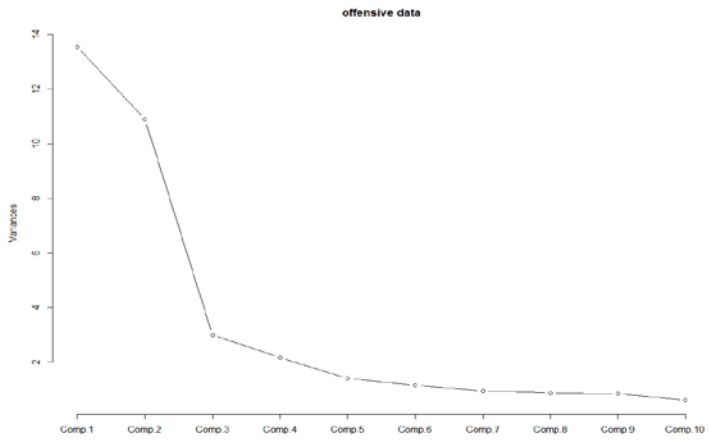


Figure 8. Offensive Data Scree Plot

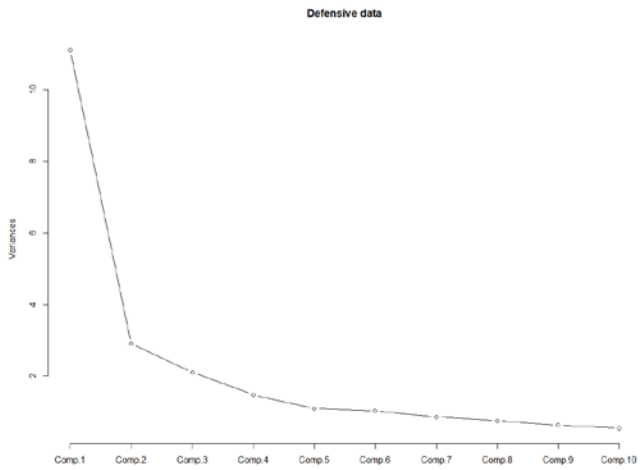


Figure 9. Defensive Data Scree Plot

We made a modeling decision to use the expertise-driven approach rather than the PCA-based approach due to the ability to better explain the expertise-driven approaches results and the possibility of applying the insights to other applications within basketball.

3.2.3 Player-Based Model Analysis

For both the offensive and defensive player-based models, our next step was to build a random forest and a linear model and test which one was more predictive. In all cases, we used the 2007-08 through 2015-16 season's data as training data and the 2017-17 season's data as test data. Since the next steps of our analysis make use of 2017-18 data, we did not employ that in any part of training or validating any of the models.

For both the offensive and defensive player-based models, the random forest approach was more predictive than the linear model. The offensive model had an r^2 value of 0.542 and the defensive model had an r^2 value of 0.48. Figures 10 and 11 show the predicted results on the test data, where the X-axis is the true rating and the y-axis is the predicted rating.

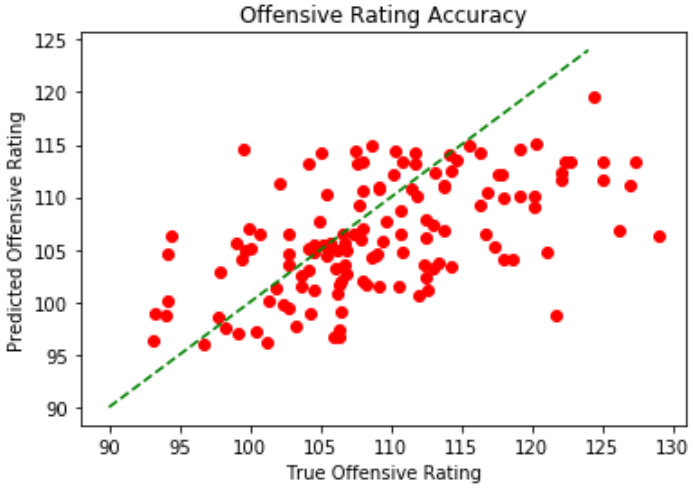


Figure 10. Player-Based Offensive Model Results

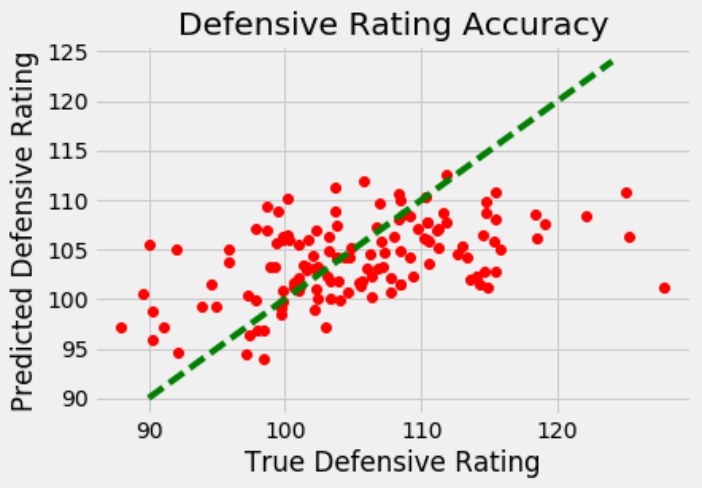


Figure 11. Player-Based Defensive Model Results

Both models have a slight trend line in the positive direction, but do not appear to be particularly predictive. This finding is confirmed by their middling r^2 values. After receiving these results, our team adjusted its approach and decided to build the model based off lineup statistics rather than aggregated player statistics. This set of models is explored in Section 3.5.

While we did not use the player aggregated models in our final results, it is still illuminating to look at how the random forest models ranked the individual attributes in importance. The important figures for both models is shown in Figures 12 and 13.

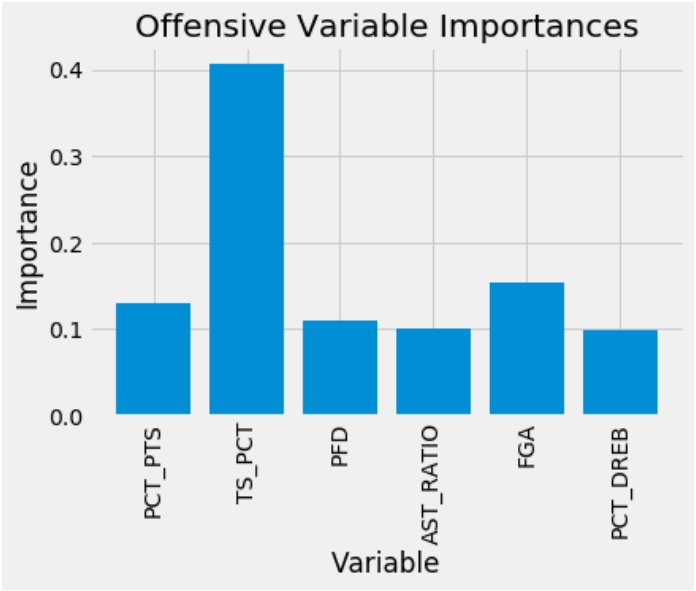


Figure 12. Player-Based Offensive Model Feature Importance

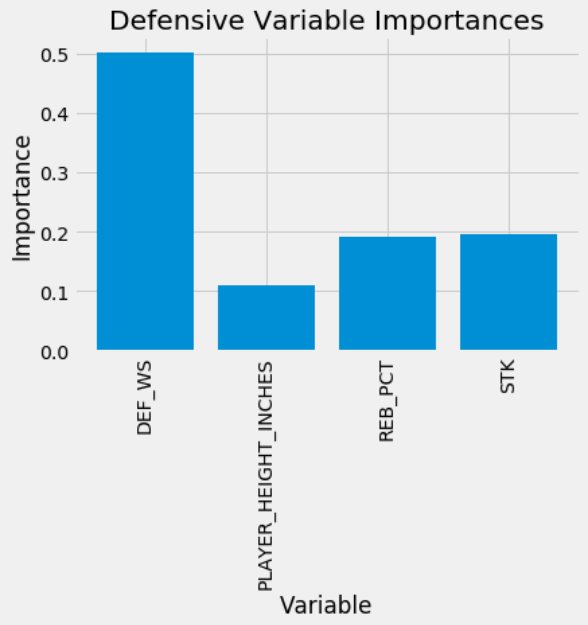


Figure 13. Player-Based Defensive Model Feature Importance

Both models featured one variable that was significantly more important than the others (true shooting percentage) for offensive and defensive win shares for defense. Due to the interconnectedness of basketball statistics, this finding foreshadowed some issues our team had when building the lineup driven models.

3.3 Lineup-Based Model

Since the player-based models were less predictive than expected, our team decided to change our approach slightly and use the statistics lineups earned while they were on the court together to identify the attributes that make lineups successful. Using this information, the Wizards can then acquire players who do those things.

3.3.1 Modeling with True Shooting Percentage

The feature reduction portion of the lineup-based modeling used the same approach as described in the player-based modeling, except with a different set of data. One difference was that we did not attempt to subdivide the statistics into categories beyond offensive and defensive. As before, rebounds were examined as both an offensive and a defensive statistic.

Using this approach, the first iteration of our offensive model found a fit that was shockingly accurate at predicting team success, with an r^2 of 0.996. The true vs. predicted offensive ratings of this model is shown in Figure 14 and the coefficients of the features is shown in Figure 15.

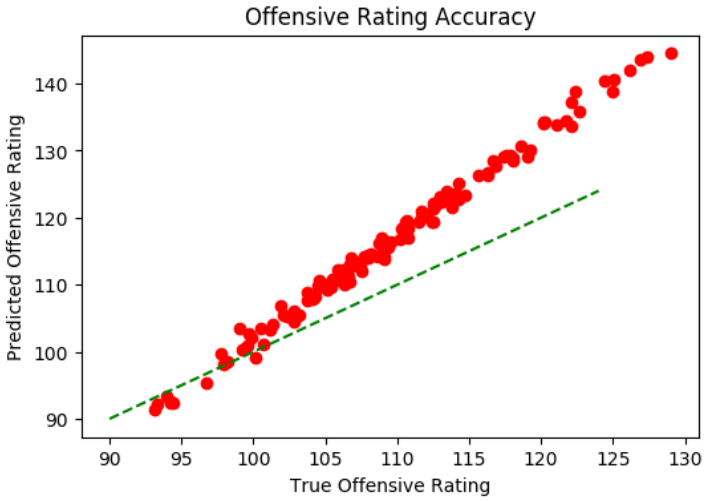


Figure 14. First Lineup-Based Offensive Model Results

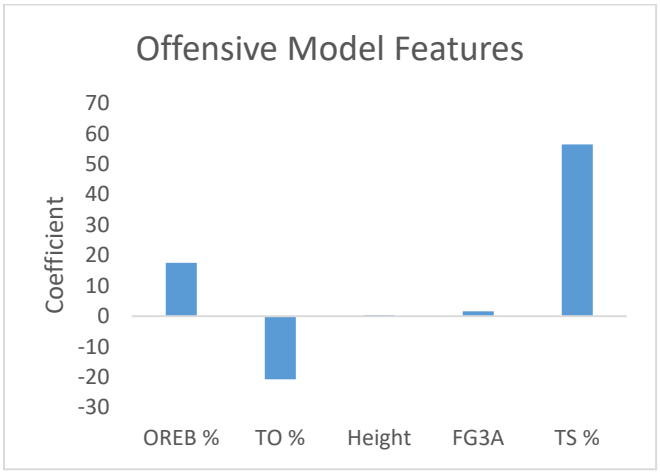


Figure 15. First Lineup-Based Offensive Model Feature Coefficients

The most important factor in this model is true shooting percentage, which is consistent with our early findings on the value of shooting and shot-making. However, upon using the model to predict individual player values for 2016-17, we discovered that the model was systematically overrating big men who play close to the basket, and therefore make more accurate shots, at the expense of all other players. Table 3 shows the top 5 players from 2016-17per the model as well as 5 prominent players who were not top 5. Russell Westbrook, who ranked 273rd, won the MVP of the league for the 2016-17 season.

Rank	Player	Off. Rating
1	Clint Capela	138.5
2	Enes Kanter	138.3
3	Steven Adams	137.8
4	DeAndre Jordan	137.6
5	Montrezl Harrell	137.3
19	Stephen Curry	133.2
60	Kevin Durant	128.0
83	LeBron James	126.4
98	James Harden	125.6
273	Russell Westbrook	116.9

Table 3. First Lineup-Based Model Player Predictions

This finding caused our team to re-examine our approach again and ultimately resulted in our removing true shooting percentage as a feature and including the second constraint that, in addition to a good r^2 value, our model must produce individual player rankings that are defensible. It makes sense to remove true shooting percentage because ultimately it is an aggregated statistic that combines several different sets of data on shooting and scoring and it rewards two types of players: players who take low-risk shots close to the basketball and players who take and make a lot of 3-pointers. These two player types are different but the statistic doesn't distinguish between them, causing big men such as Clint Capela to be over-presented at the expense of an elite 3-point shooter like Steph Curry. Furthermore, it is somewhat tautological to say that a team making shots leads to scoring more points. Scoring and shot-making area important skills in basketball, but removing true shooting percentage as a feature leads to a more predictive and defensible modeling effort overall.

3.3.2 Modeling without True Shooting Percentage

After removing true shooting percentage from the feature set, we re-ran the feature reduction and modeling process for both the offensive and defensive models using the lineup-based data. Figures 16 and 17 show the true versus predicted offensive and defensive rating the models that were developed. In both cases a linear model was more predictive than a random forest model. The offensive model had an r^2 of 0.861 and the defensive model had an r^2 of 0.578. The lower r^2 for the defensive model is to be expected because defensive statistics in basketball are significantly less mature than offensive statistics.

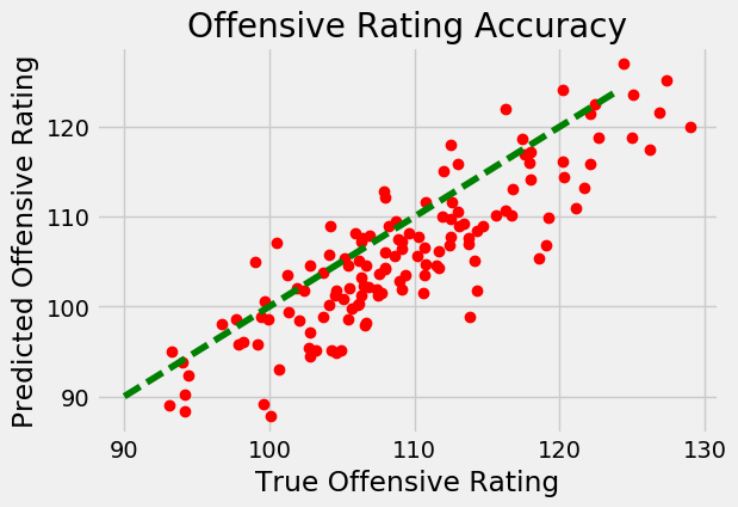


Figure 16. Lineup-Based Offensive Model Results

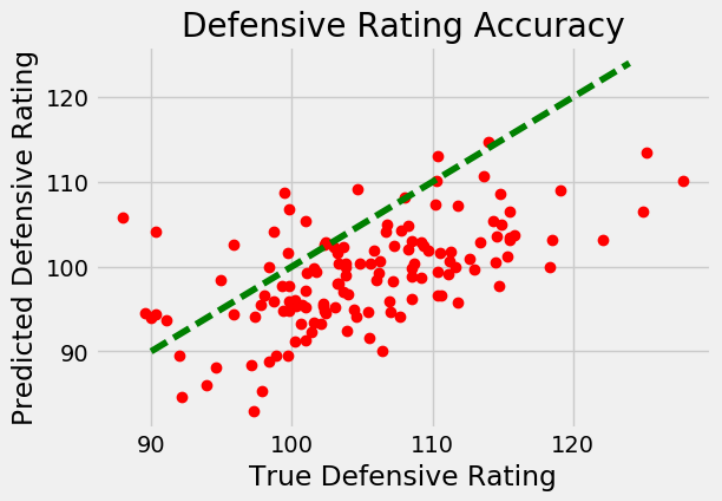


Figure 17. Lineup-Based Defensive Model Results

The values for the coefficients of the offensive model are shown in Figure 18. These values show that shot taking and shot making are the most important skills. It's notable that players provide positive value just by taking 3-pointers and free throws. These findings hint at further information in the shot chart data that we couldn't fully explore during this project. Interestingly, rebounding and assists both have negative coefficients, suggesting that players who provide those statistics without also making shots are overvalued.

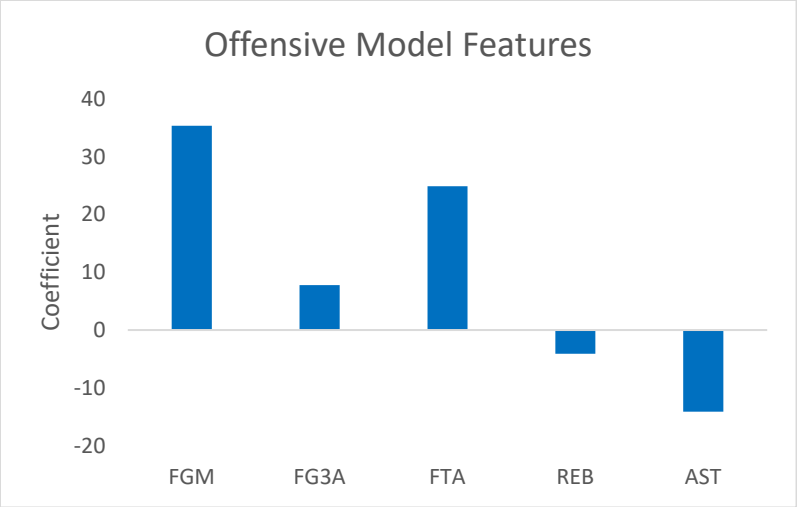


Figure 18. Lineup-Based Offensive Model Feature Coefficients

Figure 19 shows the coefficients for the defensive model. Since good defensive teams have a lower rating, a negative coefficient in this model means that the attribute is positively correlated with good defense. This model's coefficients are more intuitive, with rebounds, blocks, and steals all positively correlated with good defensive and fouling negatively correlated with good defense. As noted before, the set of data we used in our modeling does not include many features that can be expected to make up a great defensive model, since most of that data is still proprietary.

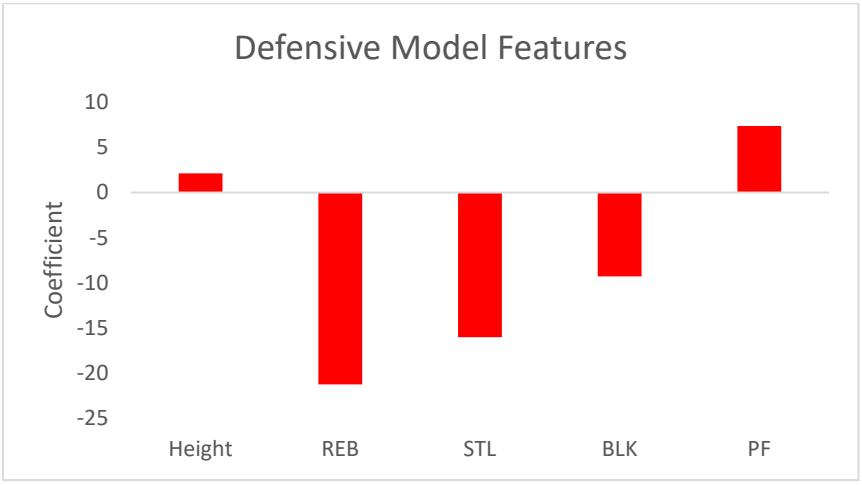


Figure 19. Lineup-Based Defensive Model Feature Coefficients

As an input to the optimization model, the coefficients from the lineup-based models were using to predict offensive, defensive, and net ratings on a player basis from the 2017-18 data. This data was combined with salary model data to inform the Wizards' 2018 offseason.

3.4 Salary Model

In addition to player values, this projects approach also requires a way to predict how much a player will earn on the open market of free agency. To do that, we gathered a data set of players who can be

acquired during the 2018 NBA offseason and built a model that predicts how much each player could expect to earn on the open market of free agency.

To simplify the salary model and to simulate the wide range of analytical competency among NBA teams, our approach began with a more limited set of data than our other models. Specifically, we only used the traditional NBA box score statistics of points, rebounds, assists, steals, and blocks, along with the age and years of experience of the player.

This set of inputs was used to predict percentage of maximum salary. We derived percentage of maximum salary because actual NBA salaries differ depending on the experience of the player and the total amount of the NBA salary cap for that season. Percentage of maximum salary is an attempt to remove these complicating factors and allow the model to predict a single value per player.

As with our models, we used the 2007-08 through 2015-16 data as training data and the 2016-17 season as test data and experimented with both random forest and linear models. For predicting NBA salaries, a random forest was found to be more predictive, with an r^2 on the test data of 0.84. Figure 20 shows the actual percentage of maximum salary for the test data plotted against the model's predicted result. Due to specific NBA rules intended to assist franchises with retaining their own star players, it is possible for a player to earn more than 100% of the maximum salary. As noted in the optimization model, these edge cases were not relevant to the Wizards' offseason so we did not explore them further.

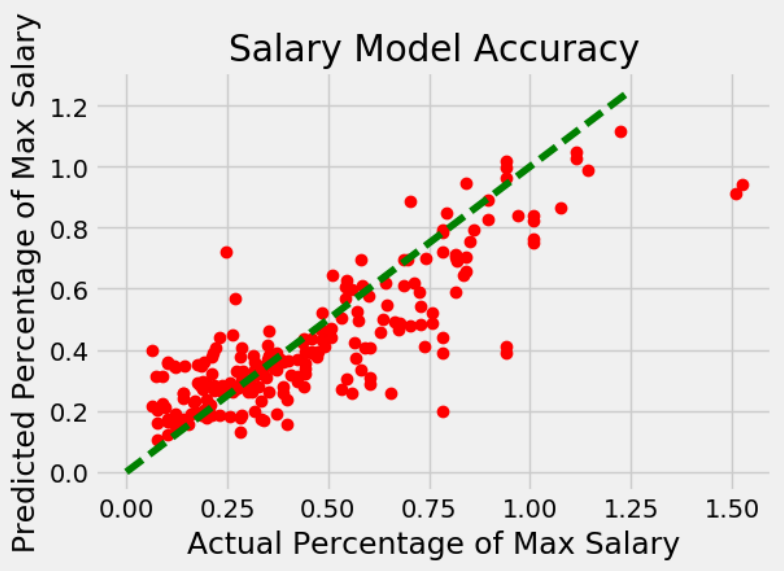


Figure 20. Salary Model Results

Figure 21 shows the relative importance of the different factors to the random forest model. Not unsurprisingly, NBA teams value points the most when negotiating a new contract. This finding is consistent with our findings elsewhere that scoring, especially efficient scoring, is the single most valuable NBA skill.

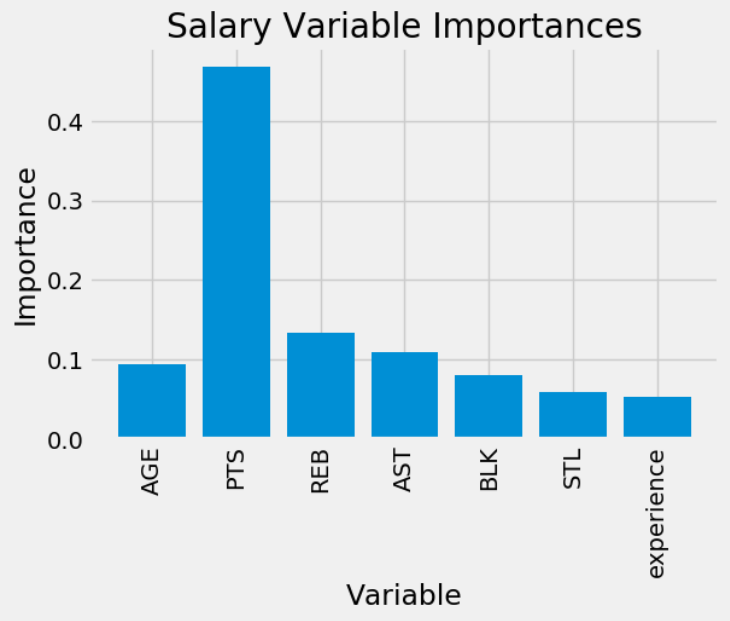


Figure 21. Salary Model Feature Importance

As an input to the optimization model, the salary model was used to predict the percentage of maximum salary that every 2018 free agent would earn. This value was then converted back into dollars using the same NBA rules as were used to convert the value into a percentage. The resulting salary value provides the optimization model a framework to determine which players the Wizards can afford to sign.

3.5 Optimization Model

While the lineup-based models and the salary model reveal useful information about player values in the NBA, the goal of this project is to inform the Wizards' 2018 offseason. Doing that requires combining the predictions of both models along with information on the status of the Wizards team and salary cap situation. This approach is described further in Figure 22.

Figure 22. Optimization Model Approach

The 2018 Wizards are already over the projected salary cap of $108 million and, for the purposes of this project, can afford to add one player at a specially designated salary slot known as the taxpayer's mid-level exception (MLE). The taxpayer's MLE for the 2018-19 season is $5,451,600. Therefore, our model suggests that the Wizards should add the best player per net rating (offensive rating minus defensive

rating) that the salary model predicts will cost less than $5,451,600. The results of this analysis and our ultimate player recommendation are described in section 4.2.

After identifying the best player that the Wizards can afford to add, we took our project one step further and optimized the Wizards rotations for next season using mixed-integer linear programming. This optimization used Gurobi solver and sought to optimize the net rating of the Wizards rotations, subject to the players who are on the team and the fact that no player can play more than 33 minutes per game. The results of this optimization are shown in section 4.3 and suggest that there is significant value to be derived by better optimizing a team's rotations.

3.6 Shot Chart Features

An aspect of our data that we were not able to fully incorporate into our analysis was shot chart data. While much more can be done in the future, we began the process by targeting a specific feature that was easy to obtain within our timeframe. To increase the performance of our model we tried to added the concept of spacing provided by players to their teammates, enabling them to score more points. Specifically, we extracted the average shot distances from the hoop for each player for each season. These results were partitioned by shots attempted and shots made for 2- and 3-point field goals. When looking at the comparison of distributions between attempted and made 3-point shots we noticed a variation that may have been useful to describe player spacing as shown in Figure 23 and 24.

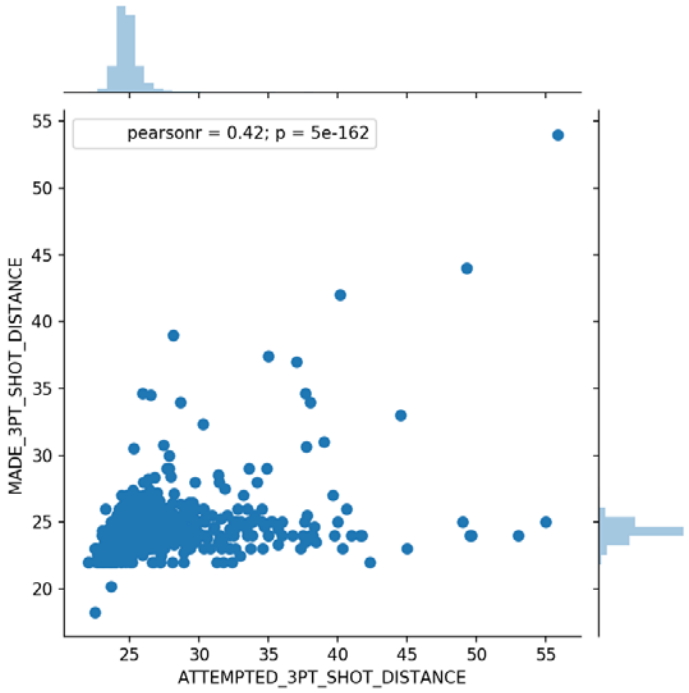


Figure 23. Shot Distance Comparison

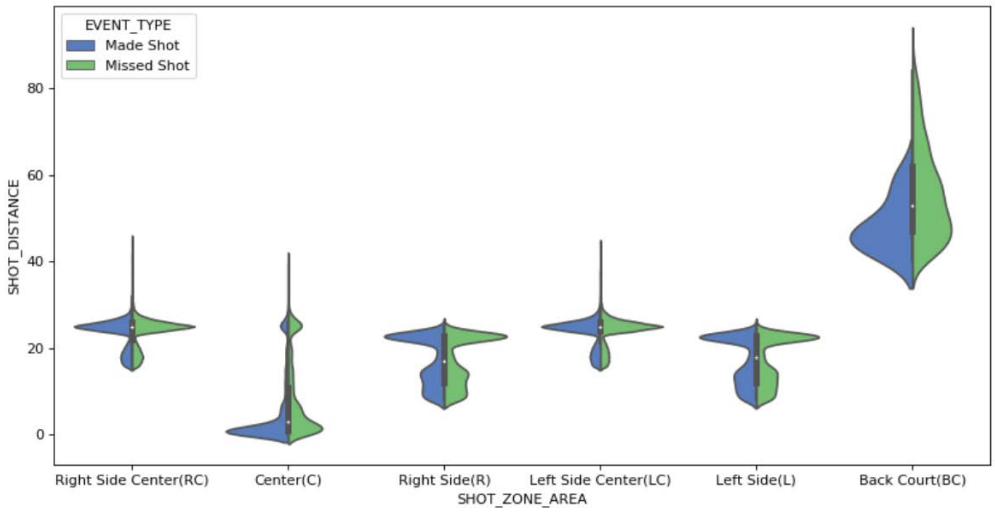


Figure 24. Shot Distance by Court Area

Unfortunately, this approach did not produce the predictive features that we expected and did not improve the performance of our model. More work can be done in the area as described in section 8 Future Work.

4 Findings and Results

4.1 Player Ratings

The offensive and defensive lineup-based models can be used to describe overall player value. Figure 23 shows the distribution of offensive, defensive, and net rating when the model is applied to the 2017-18 season's players. These distributions are like those seen on the lineup offensive and defensive rating, which is to be expected.

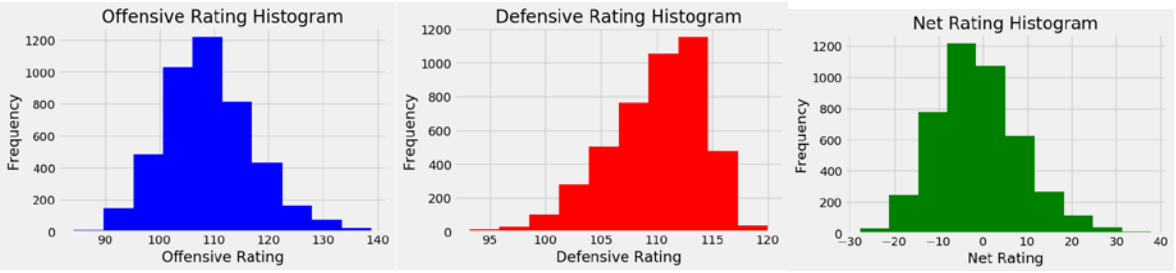


Figure 23. Offensive, Defensive, and Net Rating for 2017-18 Players

The same methodology can be used to identify the top players in the league according to this model. These predictions should be defensible according to our current understanding of NBA statistics or the discrepancy should be identifiable and defensible. Table 4 shows the top 10 players by net rating for the 2017-18 season. This list includes many of the consensus best players in the league and is much more defensible than the early list that include true shooting percentage as a feature.

Player	Net Rating
Anthony Davis	33.0
Kawhi Leonard	31.2
Joel Embiid	31.0
Giannis Antetokounmpo	28.7
DeMarcus Cousins	27.2
James Harden	26.8
Boban Marjanovic	26.5
Hassan Whiteside	24.5
LaMarcus Aldridge	24.4
Kristaps Porzingis	24.0

Table 4. Top 2017-18 Net Ratings as Predicted by Model

While this list does contain many top players, the lack of players like Steph Curry and Lebron James (which the model ranks as great but not elite) does suggest that something is missing. Our team has identified three likely candidates for this missing information, all of which are worth further investigation.

1) Spacing is key - Our model does allow for the value of 3-point shots but it may not fully be accounting for the value a good shooter brings his team mates by the attention the defense must give him. Our team began examining the impact of this spacing by looking at shot chart data but were ultimately not able to incorporate it into the model.

2) Not all passing is equal – The primary NBA statistics to measure passing is the assist and our model gives assists a negative coefficient. What his means in practice is that a player who generates a lot of assists but doesn't score is not generating the value their statistics may suggest. This insight is most likely true but crucially not all assists are created equal, even though the statistic doesn't allow for a distinction. A pass to a wide-open player laying up the ball does not generate the same added value as a pinpoint pass that creates a shot that otherwise would not exist. Hopefully tracking data will allow for distinctions between types of assists in the next few years. Adding that level of granularity to our model would most likely find that exceptional assists are extremely valuable while routine assists have the negative value our model currently shows.
3) Most defense isn't numeric – The nature of basketball is that offense has many more statistics than defense. This occurs because offense is fundamentally the act of doing something (scoring) whereas defense is the act of not allowing something (also scoring). An elite defender may never accumulate statistics such as blocks or steals because the player he's guarding never touches the ball. In practice, most elite defenders do accumulate statistics, but our model may be underrating elite on-ball defenders at the expense of statistics accumulators like Hassan Whiteside.

4.2 Wizards' 2018 Offseason

The 20117-18 Washington Wizards have 13 players under contract who played at least 200 minutes this season. Those players are shown in Table 5, along with their offensive, defensive, and net rating, as given by our model. Tim Frazier and Mike Scott's names are italicized because they are free agents and will not be back on the 2018-19 team unless they are signed to a new contract.

Player	Offensive Rating	Defensive Rating	Net Rating
Bradley Beal	122.7	111.8	10.9
John Wall	113.6	110.7	2.9
Otto Porter Jr.	112.0	106.4	5.6
Kelly Oubre Jr.	112.5	111.0	1.5
Markieff Morris	109.7	110.0	-0.3
Marcin Gortat	104.0	105.8	-1.7
Tomas Satoransky	99.2	112.4	-13.2
Mike Scott	*114.0*	*113.8*	*0.2*
Ian Mahinmi	105.9	108.0	-2.0
Tim Frazier	*90.1*	*109.9*	*-19.8*
Jodie Meeks	112.0	113.3	-1.2
Jason Smith	110.1	113.3	-3.3

Table 5. 2017-18 Wizards' Roster

Given their current salary cap situation, Washington can afford to sign one player this offseason for the taxpayer's MLE, meaning the contract must cost $5,451,600 or less. Based on the premise that the best player is the one with the highest net rating, the best player Washington can sign is point guard Devin Harris, currently with the Denver Nuggets. Harris's net rating per the model is -3.4, so he may not be a valuable contributor to the 2018-19 team. However, Harris can still add value by allowing the Wizards to

play negative contributors like Tomas Satoransky for less minutes, even in the case of an injury to John Wall. For $5,650,000, the Wizards can sign power forward Jeff Green, currently with the Cleveland Cavaliers. His net rating per the model is 1.4, so the Wizards will get a positive contributor if they can convince him to sign for the taxpayer's MLE. Details about both Devin Harris and Jeff Green are provided below in Table 6.

	Devin Harris	Jeff Green
Position	PG	PF
Age	35	31
Experience	14	11
PTS	21.5	22.4
AST	5.5	2.6
REB	4.6	6.5
BLK	0.4	0.9
STL	1.7	1.1
Salary	$4,100,000	$5,650,000
Off. Rating	111.5	114.7
Def. Rating	114.9	113.3
Net Rating	-3.4	1.4

Table 6. Summary Details on Devin Harris and Jeff Green

Figures 24 and 25 show Harris and Green's shot charts, respectively. In both cases, the left chart shows the attempts and the right chart shows the made shots. Harris is a better 3-point shooter than Green, but both players struggle in the mid-range area between the rim and the 3-point arc. The Wizards should push any players they acquire, along with their current roster, to take as many shots at the rim and behind the 3-point arc as they can.

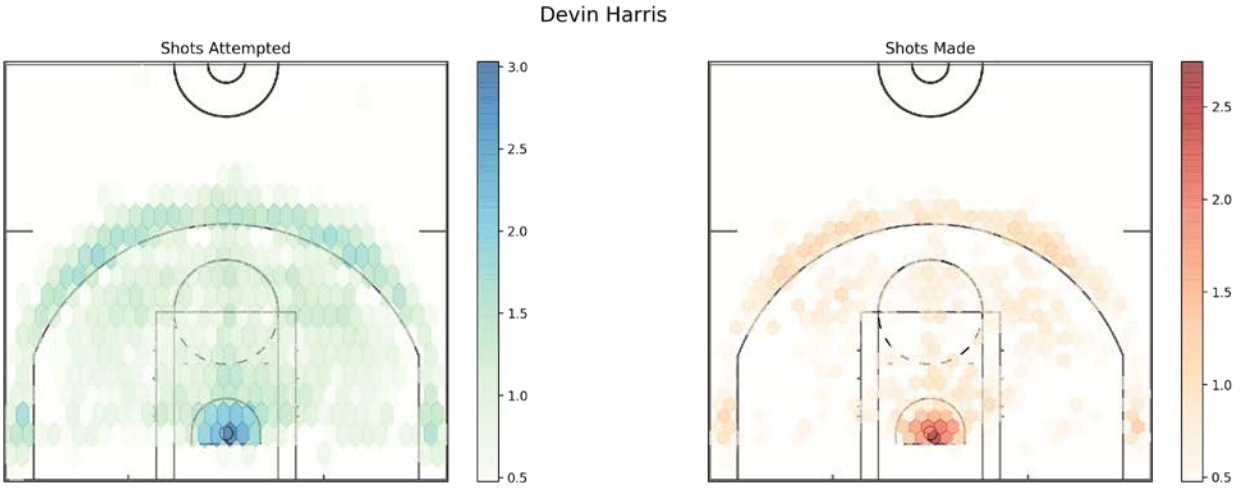


Figure 24. Devin Harris' 2017-18 Shot Chart

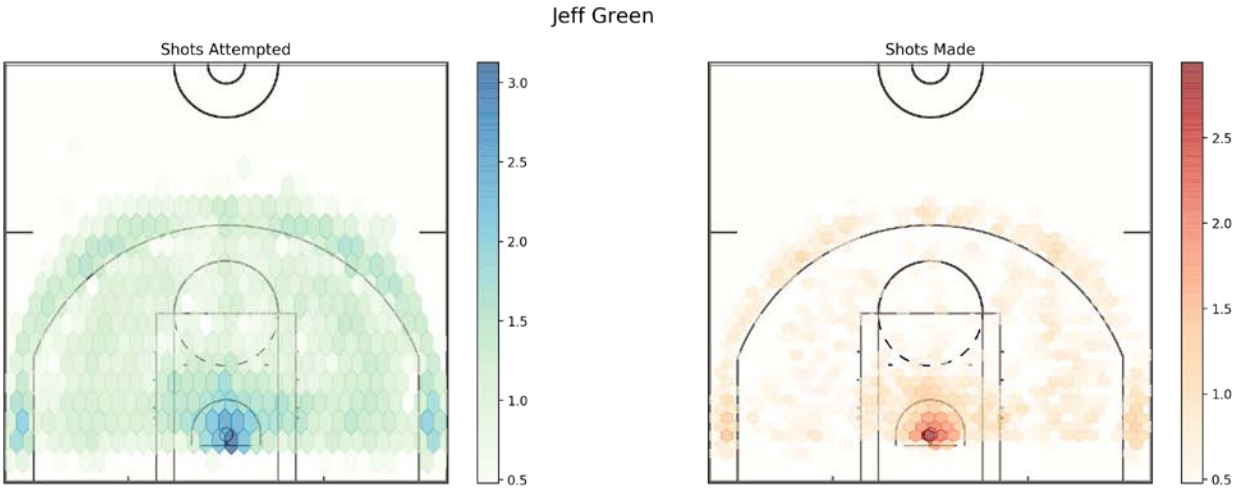


Figure 25. Jeff Green's 2017-18 Shot Chart

4.3 Wizards' Rotations

In addition to advising the Wizards on their offseason choices, it is possible to use a player's net rating to optimize the Wizards' rotation. The goal of this exercise is to inform the Wizards with how to best deploy their new team. Tables 7-9 shows the results of this analysis, with Table 7 showing the 2017-18 rotation, Table 8 showing the proposed 2018-19 rotation with Devin Harris, and Table 9 showing the proposed 2018-19 rotation with Jeff Green. It is notable that Devin Harris does not actually appear in Table 7 because the Wizards have enough players better than him to fill out their rotation. However, he still provides much needed depth in the case of an injury to John Wall, Bradley Beal, or Jodie Meeks.

Player	MPG
Bradley Beal	33
John Wall	31
Otto Porter Jr.	29
Kelly Oubre Jr.	26
Markieff Morris	25
Marcin Gortat	23
Tomas Satoransky	20
Mike Scott	17
Ian Mahinmi	12
Jodie Meeks	12
Tim Frazier	12

Table 7. 2017-18 Wizards' Rotation

Player	MPG
John Wall	33
Bradley Beal	33
Otto Porter Jr.	33
Kelly Oubre Jr.	33
Jodie Meeks	33
Markieff Morris	30
Ian Mahinmi	15
Marcin Gortat	15
Jason Smith	15

Table 8. 2018-19 Wizards' Rotation with Devin Harris

Player	MPG
John Wall	33
Bradley Beal	33
Otto Porter Jr.	33
Kelly Oubre Jr.	33
Markieff Morris	33
Jeff Green	30
Ian Mahinmi	15
Marcin Gortat	15
Jodie Meeks	15

Table 9. 2018-19 Wizards' Rotation with Jeff Green

Based on the 2017-18 season's minute distribution (which is derived from their actual rotations for the season), the model predicts that the 2017-18 Wizards should have a net rating of 1.2, which is close to their actual net rating of 0.6, especially when accounting for injuries that occurred during the season. The same methodology predicts that the rotation in Table 7 would have a net rating of 11.0 and the rotation in Table 8 would have a net rating of 13.3. Those values are certainly high, and may overestimate the true change, but they reveal that there is great value to be achieved by NBA teams better optimizing their rotations. Doing so could extract significantly more value from their current rosters or from a roster with minimal other changes.

5 Assumptions and Limitations

The full-space of NBA performance and contract analysis is well outside the scope of this project. Below is the set of assumptions that were used to constrain the solution space. One prime area of future work for this approach would be to relax these assumptions and incorporate that scope into a more powerful future approach.

1) The model ignored the impact of NBA coaching on performance and assumed that the two drivers of a player's performance are his own ability and his team context. NBA coaches do have an impact on player performance but quantifying this value is extremely difficult.
2) The model attempts to consider the on-court fit between players but does not attempt to model the personality fit of players. NBA teams throughout history have had issues due to non-basketball reasons (i.e., 2004 Lakers) but the analysis of this subject is outside the scope of this project. Further exploration could also better incorporate elements of on-court forward, specifically related to playmaking and spacing.

3) As stated throughout this report, the analysis relies heavily on the statistics Net Rating and its two components, Offensive Rating and Defensive Rating. Net Rating is based off plus/minus statistics, which are among the noisiest in basketball, requiring a lot of data before it becomes representative. Our model attempts to account for this quality of Net Rating by only analyzing lineups with more than 100 minutes played together, but Net Rating's flaws do represent a limitation on our approach.

4) Our overall project goal is to automate the roster building and maintenance decisions of an NBA general manager. However, there are several types of roster decisions that we have ignored for scoping purposes.

    a. A key roster building resource for an NBA general manager is the NBA draft. A good draft pick represents several years of quality performance at a discount price and as such, draft picks themselves are valuable assets in trades between players. However, modeling draft picks in transaction and especially predicting the performance of college players once they get to the NBA is a project of similar scope to this one unto itself. If this model were being deployed by a team, creating a draft module to plug into the overall model would be a top priority.

    b. Our model is built only from the perspective of the Washington Wizards (although it could be generalized to look at any team). Actual NBA player transactions involve many other independent actors, including opposing team, the players themselves, and their sport agents. Our model only recommends what the Wizards should do and does not take those other actors into account.

5) Our analysis makes two key assumptions about the roster building goals of team. Both assumptions serve the purpose of scoping the project. The first assumption is that an NBA team wants to win as many games as possible. Due to the incentive structure of the NBA draft lottery, some number of teams every year are not trying to maximize their chances of winning as many games as possible that season. They do this because they perceive their odds of winning enough to be successful to be extremely low and the best way to increase their odds in future seasons is

to be bad now to get a better draft pick. By selecting the Washington Wizards, we minimized the impact of this assumption because they are a team that is trying to maximize their roster

6) The second assumption is related to the first and is that different times value present roster talent versus future roster talent differently. Team A may have 45-win talent but believe that playing their younger stars now could make them a 60-win team in 2 years, even if it's not the optimal strategy now. Team B may have 55-win talent but older stars and so they may be trying to win now. By selecting the Washington Wizards, we minimized the impact of this limitation because they are trying to win now. Selecting the Wizards allows us to focus on the 2018 offseason and 2018-19 team because the Wizards value near-term wins highly compared to more distant wins. A more complete model would need to weigh these competing goals based on individual team situation.

7) Our model, especially the roster optimization component, assumes that all the Wizards' players will remain healthy for the full season. This assumption is obviously incorrect, but it does serve to simplify the problem significantly. A more complete analysis could develop an injury model that assigned each player an injury probability based on past injuries and body type.

6 Lessons Learned

Many lessons were learned while performing this project. The most significant ones are summarized below.

1) Prepare multiple approaches to allow for failures – As detailed throughout the report, the team encountered several issues that caused plans and approaches to be changed. Due to planning we completed early in the project, we had contingencies in place. A good example of this concept is the change in approach we performed from player-based to lineup-based modeling. Having a contingency plan allowed the team to keep the project on schedule despite encountering issues.
2) No need to re-invent the wheel – Basketball analytics are a rich environment and there are numerous high-quality statistics to be built from. This project sought to add to that environment by developing our own understanding of the relationship between lineup structure and lineup performance and translating that structure into a statistic that can be used to inform decisions. However, the overall success of this project would have been improved by relying more on some player value statistics that have been developed by other sources, such as Win Shares or Real Plus-Minus. This lesson is an important reminder that a good analysis doesn't have to build everything from scratch. Rather the goal is to add value to whatever subject is being explore.
3) Surprising solutions must be defensible – Our original lineup model had an extremely high R^2 but produced player values that were far from expected. The model overvalued shooting accuracy at the expense of all else, which caused players who only took low-risk shots at the rim to be ranked highest. It also ranked last year's MVP, Russell Westbrook, as the 273rd best offensive player. This solution was not defensible so it caused us to reconsider our approach. We found that true shooting percentage was causing us to overfit our model and produce indefensible results, so we removed it as a feature.
4) Interaction effects are an essential part of multi-actor analysis – Everything that happens on a basketball court depends on the other nine people on the court. Our model was not able to fully take advantage of these interaction effects and that shortcoming is the part of the most significant area of improvement for the model. Our underrating of playmaking as an attribute is one example of this shortcoming. This lesson regarding interaction effects is applicable to many problems outside the realm of basketball, or even sports analytics.
5) Model transparency is as important as accuracy – Throughout our modeling process, we sought to use simple statistics and models wherever possible to enhance the transparency of our approach and solution. This guiding principle led to some shortcomings, as described elsewhere, but a transparent solution is one that is more likely to be adopted, especially by an expert in the field like an NBA general manager.

7 Summary

There are four main takeaways from this project.

1) There is no elite offseason move for the Wizards to make in free agency – The best player the Wizards can afford with the taxpayer's MLE is Devin Harris, who has a -3.4 net rating. An optimized rotation with Harris on the team doesn't even feature him with regular minutes. The team can add Jeff Green if they can convince him to take $200k less than our model predicts he should earn, but even his net rating is only 1.4. Ultimately, the Wizards already have a very high payroll and their biggest avenue for improvement is not through free agency.
2) Optimizing rotation could be a source of significant value – Luckily for the Wizards, there may be a way to significantly improve their performance even without adding another elite player. Our model suggests that the team could add another 10 point of net rating value, going from an average team to an elite team, by better optimizing the minutes distribution of the team. This approach is likely overestimating the value of this optimization, but it does seem to be a source of significant value for the Wizards, or any other team.
3) Shot making, and shot taking, is king – We tried multiple modeling approaches throughout this analysis and every one of them revealed the supremacy of taking and making shots. This finding is somewhat tautological when you consider that the goal of basketball is to score points (and prevent them team from doing the same) and you can't score if you don't take and make shots. Our final model found that other statistics that are generally considered to be valuable, such as assists and rebounds, have a negative coefficient in a player value model. What this finding means in practice is that players who produce a lot of rebounds and assists without also scoring are not producing as much value as commonly believed.
4) New statistics may add granularity to spacing, playmaking, and defense – Our model as currently constructed underrates these three skills and thus underrates the players who excel at them. These three skills all rely heavily on interaction effects, which we were unable to fully incorporate.
    a. Our model does allow for the value of 3-point shots but it may not fully be accounting for the value a good shooter brings his team mates by the attention the defense must give him. Our team began examining the impact of this spacing by looking at shot chart data but were ultimately not able to incorporate it into the model.
    b. There is not currently a passing statistic that considers that not all passes are created equal. A pass to a wide-open player laying up the ball does not generate the same added value as a pinpoint pass that creates a shot that otherwise would not exist. Hopefully tracking data will allow for distinctions between types of assists in the next few years. Adding that level of granularity to our model would most likely find that exceptional assists are extremely valuable while routine assists have the negative value our model currently shows.
    c. The nature of basketball is that offense has many more statistics than defense. An elite defender may never accumulate statistics such as blocks or steals because the player he's guarding never touches the ball. In practice, most elite defenders do accumulate statistics, but our model may be underrating elite on-ball defenders at the expense of statistics accumulators like Hassan Whiteside.

8 Future Work

To scope our project to an achievable workload for a semester of development our team made many simplifying assumptions. These assumptions are discussed in greater detail in section 5. A more advanced model could relax any of these assumptions to produce a more powerful model. Some noteworthy examples include accounting for changes in coaching style to include how coaching impacts player performance and finding optimal coaching styles for a given roaster or attempting to incorporate player injuries into the overall modeling approach.

Another area for future work would be to extend our optimization model to include recommendations for draft picks or trades in addition to offseason free agents. Each of these presents their own set of complexities such as predicting future player performance with limited data and optimizing trade transactions when the set of possible moves if much broader.

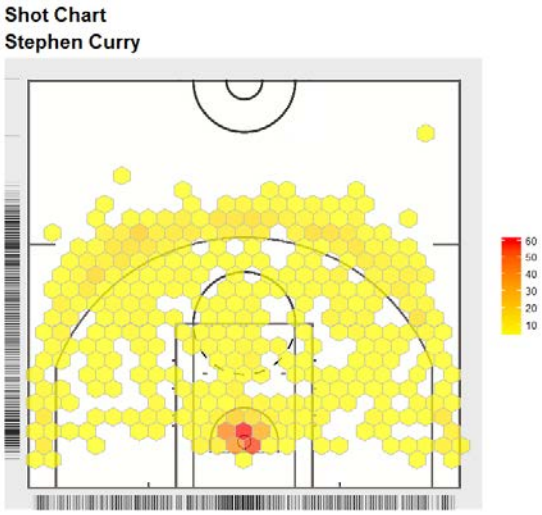
More work could also be done in exploring and extracting features that make lineups most successful. The core of our approach involves finding the most important lineup level features to apply to players but there is some evidence that more information is needs to adequately describe player contributions to lineups, such as spacing and athleticism. A more targeted analysis of shot chart data, along with other tracking data if it becomes available, could be the key to this sort of analysis.

Finally, our model does not account for temporal factors, such as a lineup playing together for several seasons. A future iteration could do several things with time such as accounting for changes in play style over time, player performance with age and how performance for a lineup changes over time.

Appendix A – Definition of Terms

Below is a lexicon of some terms used during this analysis. Additional terms and definitions can be found as part of the Dataset section, in the Field Descriptions.

- Net rating: A stat to measure a team's point differential per 100 possessions. Can be divided into offensive and defensive components and broken out into individual player contributions.
- Lineup: Collection of five players on the basketball court for one team. Statistics can be evaluated at the lineup level in addition to the player and team level.
- Shot chart: A visual representation that graphically shows where a specific player or group of players attempts and completes shots. An example of a shot chart produced using R is shown below.



- Salary cap: An agreement between the NBA and the NBA player's union that governors the amount of money that teams can spend on salaries and the types of salaries that can be signed for specific types of players.
- Maximum contract: The maximum amount that a player can sign for given the NBA salary rules. This value is different for different players given their accomplishments and years played in the league.
- Fast Break: A basketball strategy to move the ball up the court as quickly as possible following a defensive rebound or turnover. The goal of a fast break is to give the offense a chance to score without the defensive having an opportunity to get ready.
- Paint: Also called the key, the paint is the area on a basketball court inside the free throw lane. Descriptions using "in the paint" are meant to convey proximity to the basket.
- Restricted Area: The area of the court closest to the basket where secondary defenders cannot gain position to take a charge. Shots in the restricted area are typically dunks or layups.
- Mid-Level Exception: The NBA collective bargaining agreement provides several different ways that an NBA team can acquire a player. The mid-level exception or MLE, is one of those ways are is so named because it is an exception to the normal cap rules and it allows a team to sign one player to a contract at a middle amount of money. There are two MLEs, one for teams not in the luxury tax and one for teams in the luxury tax. The non-luxury tax MLE is higher than the luxury tax MLE.
- Per-100 Possessions: Basketball statistics can be expressed in many different ways. The most common of these is on a per-game basis (I.e., points per game). However, the per-game framing

misses out on differences in minutes played and team pace. A different way to express basketball statistics is per-100 possessions, which means a player's total statistics are divided by the total number of possessions they played and then multiplied by 100. It better accounts for disparities in pace and minutes, which is useful for this application because we are trying to apply players to new contexts rather than just understand where they came from.

- 3-pointer: A field goal that takes place behind the 3-point line and is worth 3-points rather than the typical 2 points.
- Free throw: A shot that is awarded upon being fouled. Depending on when the foul occurred in the quarter and whether the fouled played was in the act of shooting, fouls can be worth 1, 2, or 3, points.
- Tracking data: Data collected using overheard cameras that tracks the location of every player on the court as well as the ball. Tracking data is revolutionizing the world of basketball analytics.
- Rotation: The set of players a team will deploy in a given game as well as the specific combinations that will be deployed. Teams typically play 9-11 players in a given game and those players must play together in groups of five.
- Free agent: An NBA player who is not currently under contract with an NBA team and may sign with any team in the league. Free agency is not truly free and player/team transactions are restricted by numerous rules related to the salary cap.
- Rebound: The act of collecting the basketball after a missed shot. Rebounds can either be offensive, if the player's own team is shooting, or defensive, if the other team is shooting.
- Assist: The act of passing the ball to a scoring player in a way that leads to his basket. Assists are judgement calls and ultimately up to the game's official scorer.
- Steal: The act of stealing the ball from the opposing team without going out of bounds or fouling in the process. Steals result in a change of possession and often lead to quick points.
- Block: The act of adjusting the arc of a ball after it's been shot. Blocks can stay in-bounds or go out-of-bounds but the best blocks stay in-bounds and result in a change of possession. A block must occur while the ball is still rising in its arc. Otherwise, it is considered a goaltend and the shooting team gets credit for the points from the shot.
- Turnover: The act of losing possession of the ball to the other team without shooting. Turnovers most often occur via steals but can also occur when the ball is knocked or thrown out of bounds.
- Effective Field Goal Percentage: A statistics that quantifies the field goals made divided by the field goals attempted by a player, adjusted for the increased value that 3-pointers have compared to 2-pointers. It takes Field Goal Percentage and weights 3-pointers 1.5x compared to 2-pointers.
- True Shooting Percentage: A statistics that attempts to improve on field goal percentage by accounting for the increased value of 3-pointers as well as free throws. It is calculated via: Points/ [2*(Field Goals Attempted+0.44*Free Throws Attempted)].
- Usage Percentage: A statistic that quantifies the percentage of team's plays that the player in question finishes when he is on the court. It is calculated via: (Field Goals Attempted + (0.44 * Free Throws Attempted) + Turnovers) / Possessions.
- Defensive Win Shares: An advanced statistic that quantifies that defensive contribution to winning that a player makes to his team.

Appendix B – Data and Code Repository

The full selection of data and code that was used for this analysis can be found on the projects GitLab site at https://gitlab.com/auto-nba/auto-nba.

Appendix C – Full Data Dictionary

Lineup Data – This data comes from stats.nba.com and was extracted using Python. It goes back to the 2007-08 season. Figure 2 shows the distribution of net rating, offensive rating, and defensive rating for lineups that have played more than 100 minutes together.

- Group ID (Type: list of strings) – The 5 players present in the lineup are presented as a list of their player IDs. This field is used as the key to pair this data with individual player data.
- Group Name (Type: list of string) – Similar to the Group ID, this field is a list of the 5 player names in the lineup. It serves no analytical purpose but can be used to identify the lineup during exploratory data analysis
- Team Abbreviation (Type: string) –This field identify the NBA team that the lineup plays for. It must be one of the 30 NBA teams.
- Minutes (Type: integer) – This field contains the number of minutes the lineup has played on the season, rounded to the nearest whole minute. In order to restrict the noise-ness of our data, we only examined lineups with more than 100 minutes played together.
- Offensive Rating (Type: double) – This field contains the offensive rating for the lineup in question. Offensive rating is defined in the lexicon of this report and served as one of our two dependent variables.
- Defensive Rating (Type: double) – This field contains the offensive rating for the lineup in question. Offensive rating is defined in the lexicon of this report and served as one of our two dependent variables.
- Net Rating (Type: double) - This field contains the net rating for the lineup in question. Net rating is offensive rating minus defensive rating and is a measure of overall lineup value.

Player Data – This data comes from stats.nba.com and was extracted using Python. It goes back to the 2007-08 season. Unless otherwise note, the data is on the average for the season on a per game basis.

- Player ID (type: string) - This field contains the Player ID of the player in question. It is the key for the database.
- Player Name (type: string) - This field identifies the name of the player in question.
- Season (type: string) - This field identifies the season in which the player's statistics took place.
- Age (Type: integer) - This field identifies the age of the player at the end of the season in which the statistics took place.
- Points (Type: double) – This field identifies the number of points per game the player in question scored during the season in question.
- Field Goals Made (Type: double) - This field identifies the number of fields goals the player made per game during the season in question.

- Field Goals Missed (Type: double) - This field identifies the number of fields goals the player missed per game during the season in question.
- Field Goal % (Type: double) - This field identifies the percentage of field goals made during the season in question. It is Fields Goal Made divided by Fields Goals Made plus Field Goals Missed.
- 3-pointers Made (Type: double) - This field identifies the number of 3-pointers made by the player per game during the season in question.
- 3-pointers Attempted (Type: double) - This field identifies the number of fields goals the player attempted per game during the season in question.
- 3-point % (Type: double) - This field identifies the percentage of 3-pointers made by the player during the season in question. It is 3-pointers Made divided by 3-pointers Attempted.
- Free Throws Made (Type: double) - This field identifies the number of free throws made by the player per game during the season in question.
- Free Throws Attempted (Type: double) - This field identifies the number of free throws attempted by the player during the season in question.
- Free Throw % (Type: double) - This field identifies the percentage of free throws made by the player during the season in question. It is Free Throws Made divided by Free Throws Attempted.
- Offensive Rebounds (Type: double) - This field identifies the number of offensive rebounds per game made by a player during the season in question. Offensive rebounds are those that occur when the team is attempting to score on their opponent's basket.
- Defensive Rebounds (Type: double) - This field identifies the number of defensive rebounds per game made by a player during the season in question. Defensive rebounds are those that occur when the team is attempting to prevent the opponent from scoring on their basket.
- Rebounds (Type: double) - This field identifies the total number of rebounds made per game by a player during the season in question. It is Offensive Rebounds plus Defensive Rebounds.
- Assists (Type: double) - This field identifies the number of assists per game by the player in question during the season in question.
- Turnovers (Type: double) - This field identifies the number of times per game during the season in question that the player in question turned the ball over to the other team.
- Steals (Type: double) - This field identifies the number of times per game during the season in question that the player in question stole the ball from the other team.
- Blocks (Type: double) - This field identifies the number of times per game during the season in question that the player in question blocked a shot by the other team.
- Blocks Allowed (Type: double) - This field identifies the number of times per game during the season in question that the player in question had his own shot blocked by a player on the other team.

- Personal Fouls (Type: double) - This field identifies the number of times per game during the season in question that the player in question fouled a player on the other team.
- Personal Fouls Allowed (Type: double) - This field identifies the number of times per game during the season in question that the player in question was fouled by a player on the other team.
- +/- (Type: double) - This field identifies on a per game by basis how many more (or less) points a player's team scored when he was on the court as compared to off the court.
- Fantasy Points (Type: double) - This field identifies how many fantasy basketball points a player averaged per game during the season in question. The scoring is calculated using nba.com fantasy rules.
- Double-Doubles (Type: double) - This field identifies how many double-doubles the player in question had during the season in question. A double-double is when a player has 10 or more of any two of the five primary statistics: points, rebounds, assists, blocks, and steals.
- Triple-Doubles (Type: double) - This field identifies how many triple-doubles the player in question had during the season in question. A triple-double is when a player has 10 or more of any three of the five primary statistics: points, rebounds, assists, blocks, and steals.
- Assist Ratio (Type: double) - This field identifies the percentage of a player's possessions during the season in question that end in an assist.
- Assist-to-Turnover Ratio (Type: double) - This field identifies the ratio during the season in question between the number of assists and turnovers a player produces.
- Defensive Rating (Type: double) - This field identifies, per 100 possessions, how much a player's team performs on defense when he is on the court as compared to off the court. A defensive rating of 100 is average and below 100 is better than average.
- Defensive Rebound % (Type: double) - This field identifies the percentage of defensive rebounds that the player in question grabbed out of all the available defensive rebounds while they were on the court.
- Effective Field Goal % (Type: double) - This field identifies the field goals made divided by the field goals attempted by a player during the season in question, adjusted for the increased value that 3-pointers have compared to 2-pointers. It takes Field Goal Percentage and weights 3-pointers 1.5x compared to 2-pointers.
- Offensive Rating (Type: double) - This field identifies, per 100 possessions, how much a player's team performs on offense when he is on the court as compared to off the court. An offensive rating of 100 is average and above 100 is better than average.
- Net Rating (Type: double) - This field identifies, per 100 possessions, how much better a player's team performs overall when he is on the court as compared to off the court. It is Offensive Rating minus Defensive Rating.

- Offensive Rebound % (Type: double) - This field identifies the percentage of offensive rebounds that the player in question grabbed out of all the available offensive rebounds while they were on the court.
- Pace (Type: double) - This field identifies how fast the player's team plays while he is on the court. It is normalized to 100 being average.
- Player Impact Estimate (Type: double) - This field is a nba.com attempt at producing a single statistic quantify a player's contribution while on the court. It uses other field described in this report.
- Rebound % (Type: double) - This field identifies the percentage of rebounds that the player in question grabbed out of all the available rebounds while they were on the court.
- True Shooting % (Type: double) - This field identifies the player in questions Field Goal Percentage, accounting for the increased value of 3-pointers as well as Free Throws. It is calculated via: Points/ [2*(Field Goals Attempted+0.44*Free Throws Attempted)].
- Usage Percentage (Type: double) - This field identifies the percentage of team's plays that the player in question finishes when he is on the court. It is calculated via: (Field Goals Attempted + (0.44 * Free Throws Attempted) + Turnovers) / Possessions
- Defensive Win Shares (Type: double) - This field identifies the overall amount that a player's defense contributes to his team winning.
- Opponent 2nd Chance Points (Type: double) - This field identifies number of points per game during the season in question that the opposing team scored after already attempting a shot on offense, while the player in question was on the court.
- Opponent Fast Break Points (Type: double) - This field identifies number of fast break points per game during the season in question that the opposing team scored while the player in question was on the court.
- Opponent Points Off-Of Turnovers (Type: double) - This field identifies number of points after a turnover per game during the season in question that the opposing team scored while the player in question was on the court.
- Opponent Points in the Paint (Type: double) - This field identifies number of points in the paint per game during the season in question that the opposing team scored while the player in question was on the court.
- Block % (Type: double) - This field identifies the percentage of the team's blocks that the player in question had while they were on the court.
- Steal % (Type: double) - This field identifies the percentage of the team's steals that the player in question had while they were on the court.
- Blocks Allowed % (Type: double) - This field identifies the percentage of the team's blocked shots that the player in question had while they were on the court.

- % of Team FG Made (Type: double) - This field identifies the percentage of the team's made field goals that the player in question had while they were on the court.
- % of Team FG Attempted (Type: double) - This field identifies the percentage of the team's attempted field goals that the player in question had while they were on the court.
- % of Team 3-pointers Made (Type: double) - This field identifies the percentage of the team's made 3-pointers that the player in question had while they were on the court.
- % of Team 3-pointers Attempted (Type: double) - This field identifies the percentage of the team's attempted 3-pointers that the player in question had while they were on the court.
- % of Team O-Rebounds (Type: double) - This field identifies the percentage of the team's offensive rebounds that the player in question had while they were on the court.
- % of Team D-Rebounds (Type: double) - This field identifies the percentage of the team's defensive rebounds that the player in question had while they were on the court.
- % of Team Rebounds (Type: double) -This field identifies the percentage of the team's rebounds that the player in question had while they were on the court.
- % of Team Assists (Type: double) -This field identifies the percentage of the team's assists that the player in question had while they were on the court.
- % of Team Turnovers (Type: double) - This field identifies the percentage of the team's turnovers that the player in question had while they were on the court.
- % of Team Steals (Type: double) - This field identifies the percentage of the team's made 3-pointers that the player in question had while they were on the court.
- % of Team Fouls (Type: double) - This field identifies the percentage of the team's fouls that the player in question had while they were on the court.
- % of Team Fouls Drawn (Type: double) - This field identifies the percentage of the team's drawn fouls that the player in question had while they were on the court.
- % of Team Points (Type: double) - This field identifies the percentage of the team's points that the player in question had while they were on the court.

Shot Chart Data – This data comes from NBA.com and was extracted using R. It was collected by the NBA using SportsVU, a camera system that is hung from the rafters in NBA arenas and collects data at 25 frames per second. This dataset is the largest single file we have in our database at nearly 500 MB. To ease analysis, we've summarized the data by player and season for 6 zones of the court: Above the Break 3, Right Corner 3, Left Corner 3, Mid-Range, In the Paint, and Restricted Area.

- Game ID (type: string) – This field contains the game ID in question.
- Player ID (Type: string) – This field contains the nba.com Player ID of the player in question.
- Team ID (Type: string) –This field the team ID for the player in question.

- Period (Type: integer) – This field contains which quarter of the game the shot took place.
- Time Remaining (Type: datetime) – This field contains the amount of time left in the quarter is a combination of Minutes Remaining and Seconds Remaining.
- Event Type (Type: string) – This field contains whether the shot in question was Made or Missed.
- Action Type (Type: string) – This field contains whether the shot was a Jumper or Layup.
- Shot Type (Type: string) - This field identifies whether the shot was a 2 or 3-pointer.
- Shot Zone (Type: string) – This field identifies which zone of the court the shot took place. The six zones of the court are: Above the Break 3, Right Corner 3, Left Corner 3, Mid-Range, In the Paint, and Restricted Area.
- Shot Distance (Type: string) – This field identifies how far from the goal the shot took place from. It can be one of several categorical values describing distance.
- Shot Location x (Type: integer) – This field identifies the location of the x-coordinate of a shot on an xy-plane.
- Shot Location y (Type: integer) – This field identifies the location of the y-coordinate of a shot on an xy-plane.

League Salary Data – This data was extracted manually from NBA.com and basketball-reference.com to provide information on NBA salaries over time. It goes back to the 2012-13 season and does not include minimum contracts.

- Player ID (Type: string) – This field contains the nba.com Player ID of the player in question. It was matched to the contract list using R and is the key for the database.
- Player Name (Type: string) –This field identifies the name of the player in question.
- Team Name (Type: string) – This field contains the team that signed the contract in question.
- Contract Year (Type: integer) – This field contains the season the contract in question was signed.
- Contract Length (Type: double) – This field contains the length of the contract in question. A model that analyzed contracts and the NBA salary structure temporally would use this data. However, that is outside the scope of this analysis.
- Salary Value (Type: double) – This field contains the one-season monetary value of the contract in the year it was signed. Contract values change slightly over time (typically increasing) but this value is adequate for the scope of modeling in this analysis.
- Rookie Scale (Type: boolean) - This field identifies whether the contract in question was a part of the rookie scale. The rookie scale structures the contract values for younger NBA players and must be analyzed separately from non-rookie scale contracts.

- Experience (Type: integer) - This field contains the number of years of NBA experience the player in question had when they began their contract. Years of experience are a major factor in NBA contract rules.
- Percent of Maximum (type: double) - This field contains the percentage of the maximum salary that a player could sign that they actually did sign. It was derived using R based on the contract year, salary value, rookie scale, and experience and incorporated the NBA's salary cap rules.

Wizards Salary Data – This data was extracted manually from sportrac.com and contains information on the Washington Wizards current and future salary situation

- Player ID (Type: string) – This field contains the nba.com Player ID of the player in question. It was matched to the contract list using R and is the key for the database.
- Player Name (Type: string) – This field identifies the name of the player in question.
- Position (Type: string) – This field contains the position of the player in question. It can be PG (point guard), SG (shooting guard), SF (small forward), PF (power forward), or C (center).
- 17-18 Salary (Type: integer) – This field contains the amount of salary the player is paid during the 2017-18 season. This field will be used to identify the Wizards current team structure, salary situation, and team needs.
- 18-19 Salary (Type: integer) – This field contains the amount of salary the player is paid during the 2018-19 season.
- 19-20 Salary (Type: integer) – This field contains the amount of salary the player is paid during the 2019-20 season.

2018 Free Agents Data – This data was extracted manually from sportrac.com and contains information on which players will be free agents during the 2018-19 off-season.

- Player ID (Type: string) – This field contains the nba.com Player ID of the player in question. It was matched to the contract list using R and is the key for the database.
- Player Name (Type: string) –This field identifies the name of the player in question.
- Position (Type: string) – This field contains the position of the player in question. It can be PG (point guard), SG (shooting guard), SF (small forward), PF (power forward), or C (center).
- Age (Type: integer) – This field contains the current age of the player in question.
- Experience (Type: integer) – This field contains the years of NBA experience of the player in question.
- Free Agent Type (Type: string) – This field can be either Restricted or Unrestricted and impacts the type of salary that the player in question can sign for.
- Rights Type (Type: string) – This field can be either None, Bird, or Early Bird and impacts the type of salary that the player in question can sign for.

Appendix D – References

1) Basketball Statistics and History. (n.d.). Retrieved April 23, 2018, from https://www.basketball-reference.com/
2) Bibey, C. (n.d.). An Introduction to Advanced Basketball Statistics. Retrieved April 23, 2018, from https://www.sportingcharts.com/articles/nba/an-introduction-to-advanced-basketball-statistics.aspx
3) Maia, E. (2017, March 26). How to create NBA shot charts in R. Retrieved April 23, 2018, from https://thedatagame.com.au/2015/09/27/how-to-create-nba-shot-charts-in-r/
4) NBA 2017-2018 Cap Tracker. (n.d.). Retrieved April 23, 2018, from http://www.spotrac.com/nba/cap/
5) NBA salary cap. (2018, April 23). Retrieved April 23, 2018, from https://en.wikipedia.org/wiki/NBA_salary_cap
6) NBA Salary Cap FAQ. (n.d.). Retrieved April 23, 2018, from http://www.cbafaq.com/salarycap.htm
7) NBA Statistics. (n.d.). Retrieved April 23, 2018, from http://www.espn.com/nba/statistics
8) NBA Stats. (n.d.). Retrieved April 23, 2018, from http://stats.nba.com/
9) Oliver, D. (2013, November 15). How numbers have changed the NBA. Retrieved April 23, 2018, from http://www.espn.com/nba/story/_/id/9980160/nba-how-analytics-movement-evolved-nba
10) Polacek, S. (2017, September 29). NBA Salary Cap Reportedly Projected to Be $101M in 2018, $108M in 2019. Retrieved April 23, 2018, from http://bleacherreport.com/articles/2735791-nba-salary-cap-reportedly-projected-to-be-101m-in-2018-108m-in-2019
11) S. (n.d.). Savvastj/nbashots. Retrieved April 23, 2018, from https://github.com/savvastj/nbashots
12) Schneider, T. (n.d.). BallR: Interactive NBA Shot Charts with R and Shiny. Retrieved April 23, 2018, from http://toddwschneider.com/posts/ballr-interactive-nba-shot-charts-with-r-and-shiny/
13) Second Spectrum Data. (n.d.). Retrieved April 24, 2018, from https://www.nbastuffer.com/analytics101/second-spectrum/
14) Tjortjoglou, S. (2015, July 28). How to Create NBA Shot Charts in Python. Retrieved April 23, 2018, from http://savvastjortjoglou.com/nba-shot-sharts.html
15) Values Of 2017/18 Mid-Level, Bi-Annual Exceptions. (n.d.). Retrieved April 23, 2018, from https://www.hoopsrumors.com/2017/06/values-of-201718-mid-level-bi-annual-exceptions.html