ABSTRACT

Title of Document: DATA-DRIVEN APPROACHES TO NBA TEAM EVALUATION AND BUILDING

Hyunsoo Chun, David Creegan, Olivia Majedi, Daniel Smolyak, Brian Valcarcel

Directed by: Dr. Margret Bjarnadottir

Department of Decisions, Operations, and Information Technologies, R.H. Smith School of Business, University of Maryland

In the National Basketball Association (NBA), it has historically been difficult to build and sustain a team that can consistently compete for championships. Given this challenge, we have developed a series of analyses to support NBA teams in making data-driven decisions. Relying on a variety of datasets, we examined several facets related to the construction of NBA rosters and their performance. In our analysis of on-court performance, we have used clustering algorithms to classify teams in terms of play style, and determined which play styles tend to lead to success. In our analysis of roster construction and transactions, we have investigated the relative value of draft picks and the impact of trades involving draft picks, as well as the effect of roster continuity (i.e. maintaining the same players across seasons) on team success. Additionally, we have developed a model for predicting player contract values and performance versus contract

value, which will help teams in identifying the most cost-effective players to acquire. Ultimately, this assembly of analyses, in conjunction, can be used to inform any NBA team's decisions in its pursuit of success.

DATA-DRIVEN APPROACHES TO NBA TEAM EVALUATION AND BUILDING

By

Team PROCESS

Hyunsoo Chun

David Creegan

Olivia Majedi

Daniel Smolyak

Brian Valcarcel

Thesis submitted in partial fulfillment of the requirements of the Gemstone Honors Program

University of Maryland, College Park

2020

Advisory Committee:

Dr. Margret Bjarnadottir, Mentor

Dr. Sean Barnes, Discussant

Dr. David Klossner, Discussant

Dr. Nicholas Montgomery, Discussant

Dr. David Waguespack, Discussant

ACKNOWLEDGEMENTS

We would like to thank our mentor, Dr. Margret Bjarnadottir. It cannot be overstated how much her guidance, expertise, and support helped our team find success.

We would like to thank Dr. Sean Barnes, who mentored our team along with Dr. Bjarnadottir our first two years as a team. His positivity and knowledge of both data analysis and sport made him an excellent mentor.

We would like to thank Dr. Susan Weisner, our librarian, who often guided us to use all available resources to the best of our ability.

We would like to thank Matthew Moore and Will Disher, who started on this team but were unfortunately unable to stay with us for the completion of this thesis. Their contribution to this team was significant and we would be remiss not granting them credit for their work.

We would like to thank the Gemstone staff: Dr. Kristan Skendall, Dr. Frank Coale, Dr. Vickie Hill, Leah Kreimer Tobin, and Jessica Lee. The whole Gemstone staff made this research possible, and helped us every step of the way.

Finally, we would like to thank our thesis discussants Dr. David Klossner, Dr. Nicholas Montgomery, and Dr. David Waguespack for their contributions.

TABLE OF CONTENTS

1. INTRODUCTION AND BACKGROUND

Data analytics provide important insights to inform decision making through the discovery of patterns and trends in datasets, and relating these observations to outcomes. With the increasing collection and availability of large datasets, data analytics has the potential to revolutionize decision making in a multitude of industries, from health to finance; the sports industry is no exception. The sports analytics market had an estimated value of $125 million in 2014, and Marketwatch (2015) projects that spending will grow to $4.7 billion by 2021. Today, every major sports team has an analytics department or an expert (Neese, 2015). Using analytics enables teams to increase the understanding of past performance and factors that drove success (or lack thereof). Teams can also leverage analytics to make predictions of future success, and to inform decisions that lead to success.

The roots of American sports analytics can be found in baseball. In 1876, an English statistician, Henry Chadwick, created the modern baseball box score to summarize the events of the game (Neese, 2015). Chadwick began by recording hits, home runs, and total bases. Soon, statisticians started aggregating these statistics to create new metrics such as batting average. The next big step for sports analytics, however, was not until the 1980's when Bill James created a yearly publication on baseball analytics. James's most influential publication, *The Bill James Historical Baseball Abstract* (1985), exposed the masses to data analytics research in baseball. James coined the term "sabermetrics" to describe the new baseball analytics, because the work was driven by the Society of American Baseball Research (SABR). With the arrival of sabermetrics, the public's interest in sports analytics began to grow (Birnbaum, 2017).

An early example of the impact sports analytics on team performance came between 1969 and 1971, when Baltimore Orioles manager Earl Weaver began utilizing pitching and batting splits, which capture how players perform against right- and left-handed opponents (Neese, 2015). Subsequently, the Orioles won three straight pennants in the American League. Later, in 2002, the Oakland Athletics would leverage sports analytics to build a team that won 20 consecutive games (Neese, 2015). The role of sports analytics in their success was chronicled in the book *Moneyball*. The success of the book in popular culture acted as a catalyst for more data driven decision making in sports. The phenomenon soon spread to other major sports, and the use of data analytics in all sports has been growing at an exponential rate since then. In this thesis, we focus on team and player success in the National Basketball Association (NBA).

1.1 Success in the NBA

The NBA is split into two conferences of 15 teams, the Eastern Conference and the Western Conference. During the regular season the teams play 82 games, and the top 8 teams from each conference advance to the playoffs. The playoffs conclude with the Eastern Conference champion and Western Conference Champion competing in the NBA Finals.

Winning a championship is each team's priority. In a league with limited talent, 30 teams, and only one champion, this is a difficult task. Even making consistent playoff appearances is no small feat. According to Joel Litvin, NBA's president of basketball operations in 2015, the way to build a successful team is to have younger players, a low payroll, draft well, and be smart about free agency signings (Litvin qtd. in Abbott, 2012), but this doesn't always work (Abbott, 2012). To gain the upper hand, NBA teams utilize the draft, free agency, and trade to build their vision of a successful team (see 1.3).

A team's financial situation influences their capability of attracting talent. While there are league rules that restrict a team's spending on player salaries (see 1.3.2), large-market teams still hold some financial advantage. However, there is not a direct link between the financial worth of a team and their success. In 2017, the two most valuable teams in the NBA, the Knicks and the Lakers, were both worth over $3 billion dollars, and yet both failed to make the playoffs in that same year (Badenhausen, 2017). The Golden State Warriors have resources comparable to the Knicks and Lakers, being valued at $2.6 billion, the third highest in the league, and with an annual revenue of over $300 million (Badenhausen, 2017). In contrast to the Knicks and Lakers, the 2017 Warriors were highly successful in the regular season and won the NBA championship, in large part due to a star studded roster. In contrast, the San Antonio Spurs have remained a playoff quality team every year for two decades because of great front office decision making and great coaching, despite having fewer resources than other top teams. Forbes ranks the Spurs at $1.175 billion with a yearly revenue of $187 million, 12th in the NBA (Badenhausen, 2017). While money affects decision making, it is by no means a predictor of success (Mandle, 1998).

The presence of the salary cap, a league-defined maximum threshold for player salaries (with exceptions -see 1.3.2), limits teams' resources for gathering talent. The inability to spend freely means that teams must be prudent about its allocation of resources to construct a winning team. In section 1.3, we discuss the various means of constructing a roster. Namely, we discuss the NBA draft in section 1.3.1, free agency in Section 1.3.2, and trading in Section 1.3.3.

1.1.1 Success cycles

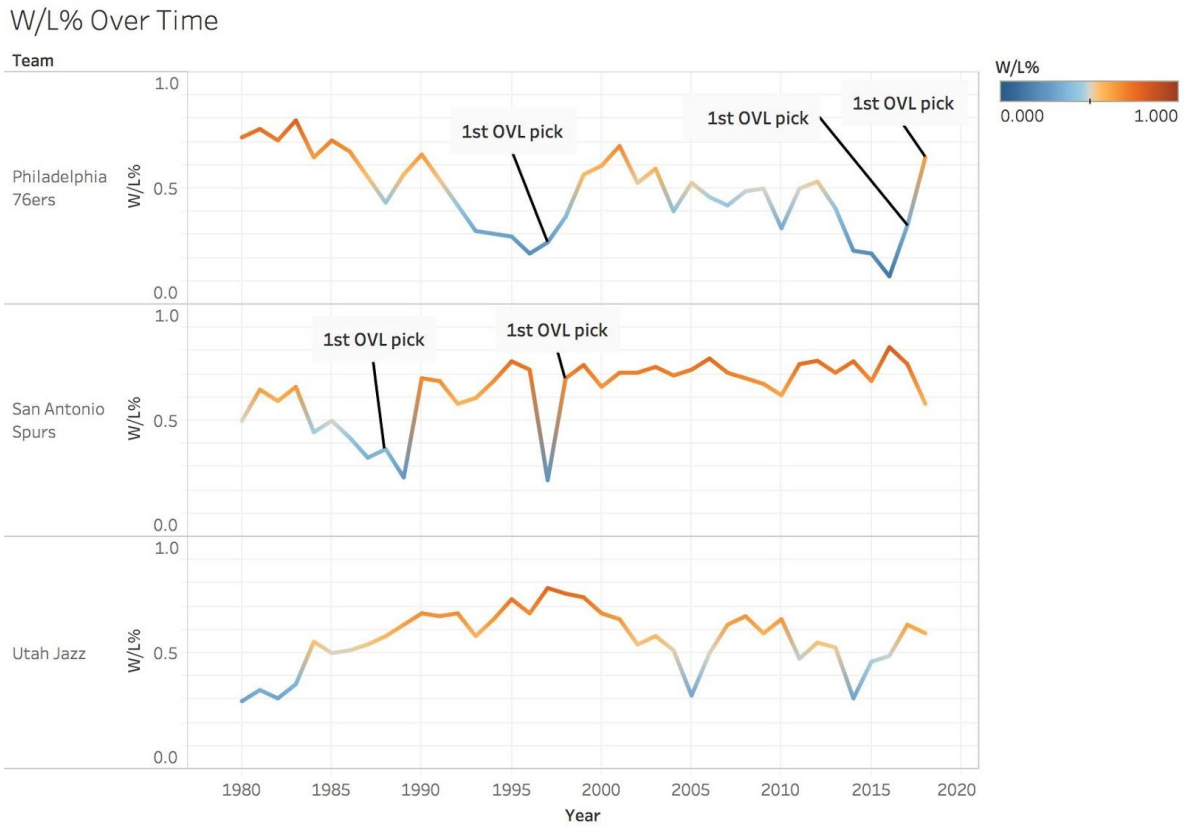


Fig. 1.1 Success Cycles: Plot shows winning percentage (the number of wins divided by the total number of games) each season from 1980 to present for three teams

The term success cycle refers to the potential existence of a pattern in the ups and downs of team performance across many seasons. There is reason to believe both that team standing is relatively stable and that it is prone to fluctuation. In the lottery (see 1.3.1) era (1985-present), 150 teams had 55 wins or more in a single season. Of those teams, only 23% had fewer than 30 wins in any of the four previous seasons (Berri qtd. in, Abbott, 2012). Meanwhile, teams with 55 wins or more in a given season stay in playoff contention for a decade on average (Oliver qtd. in Abbott, 2012). However, our analysis in Chapter 8.2 shows that team standing is fluid from year

to year for middle-performing teams and becomes fluid within 5 years for teams at the extreme ends of performance.

If teams typically maintained their performance year over year, that would challenge the notion that tanking is a good strategy. Tanking refers to the action of teams intentionally fielding mediocre or poor teams in hopes of acquiring future talent via early draft picks[1]. Many believe this harms competition and the overall NBA product and have proposed solutions to prevent tanking. Adam Gold, a presenter at the 2012 Sloan Sports Analytics Conference suggested that the first team that is eliminated from the playoffs should be given the first draft pick, and as teams are eliminated they get the following picks. His reasoning: "We should never have to consider that a loss can be more helpful than a win."(Gold qtd. in Wade 2013). In 2019, the league adopted new policies to discourage tanking by decreasing the odds for the teams with the worst records to get the top pick. We discuss the draft in more detail in 1.3.1.

1.2 Basketball Data and Statistics

Traditional NBA statistics describe the basic events in the game, such as 2-point attempts, 3-point percentage, and assists. These traditional statistics can highlight team differences, for example, a team with a higher 3-point percentage can be assumed to have better 3-point shooters. Statistics such as pace or offensive/defensive rating are a formulaic combination of those in the former category.

Simple statistics from box-scores do not always tell the whole story. A comparison based on simple statistics does not allow for robust comparisons between different styles of playing

[1] The Philadelphia 76ers, under general manager Sam Hinkie in the early 2010s, famously employed a particularly extreme tanking strategy. During the years of losing seasons, 76ers fans were told to "trust the process". The phrase became wildly popular and is our team's namesake.

basketball. Take two figurative point guards: a guard who scores 30 points a game, but plays poor defense and only scores so much because he takes many attempts, is likely not as valuable as a high percentage shooter who is also effective on defense. To better describe and capture the value of a player, analysts created new metrics to capture the efficiency of a player. These metrics take into account their total production normalized in one way or another by attempts (e.g., three-point shooting percentage, rebound percentage, and more). These stats enable more relevant comparisons between players; if the two figurative point guards shoot field goals at 20% and 40% success rates, respectively, the second guard is clearly a more effective shooter. More advanced analyses and metrics have been developed by both researchers and NBA teams to better quantitatively describe performance. One of those statistics is Value Over Replacement Player (VORP).

1.2.2 VORP. In this thesis, VORP will be used as an overall measure of a player's 'on-court value'. This metric is calculated by estimating a player's contribution to a team relative to a theoretical "replacement player" that represents the approximate value of a G League talent (the G league is the NBA's equivalent of a minor league). VORP is useful because it is a relatively simple but accurate measure of player performance, which proves useful in analysis. VORP is calculated for each season, and a player's career VORP is simply the sum of the VORP for each season. Per FiveThirtyEight, NBA players tend to improve through the age of 27, and then begin to decline, which is reflected in their VORP (Silver, 2015). This means a crude estimate of a player's remaining career value in terms of VORP is possible, which can be useful for example when analyzing trades.

However these advanced metrics are not perfect; players that are known to be good sometimes have lower values than would be expected. Brian Skinner's 2009 article "The Price of Anarchy in Basketball" attempts to explain these anomalies by discussing the relationship between an individual's success in a given game and how the opposite team responds to that player´s performance. For example, Russell Westbrook, a high-volume scorer, draws more defensive attention than an average player. This leads to what Skinner coins as "Braess's Paradox", which states that the highest percentage play is not always the most efficient. Westbrook has shot at a field-goal percentage or 43.4% over his career, a relatively good shooting percentage, but he is often marked by the best defender, or double teamed. This can make passing to another player more valuable even if Westbrook is a better shooter. The defensive pressure on Westbrook will also be lighter in the following possessions. In a nutshell, it is harder for Westbrook to for example score points because of the efforts of the opposing teams to stop him; a fact not captured by VORP and other statistics. A version of Braess's paradox is seen in other sports as well. In football, for example, passing plays produce more yards than run plays on average, yet most teams still run on nearly half their plays. This keeps the defense from focusing their strategy solely on the pass, and allows future passes to be more effective.

1.3 Team Construction

Arguably the most crucial aspect of a team's success is putting a group of players on the court that are both individually talented, and cohesive as a squad. There are three avenues to add new players to a roster: the draft, free agency, and trading. All NBA franchises employ some combination of the three, but teams in different stages of the success cycle and with different

resources, will use these strategies differently. Successful teams that wish to prolong their current success might trade future draft picks for veterans that will have an immediate impact on the team's performance. On the contrary, a team near the bottom of the standings would benefit more long term from trading away contracted players for earlier picks. We discuss each of these options below.

1.3.1. The Draft. The draft is the primary way for young players to enter the league. As of 2018, for players to be eligible for the NBA draft, the player must be at least 19 years old and declare his eligibility 60 days before the draft (Berri et. al, 2011). In the draft, which occurs in the off-season, teams select two players from the pool of rookies, one in each of two rounds, who they believe will benefit their team the most. As the primary method of gathering young talent, the draft is arguably the most important aspect in building a successful team. The top prospect in a given draft is usually selected first, and the projected production of a given player descends with draft order. The value of the top pick was studied by Christopher Williams and Tyler Walters, who examined teams that had the first pick in the draft and measured two variables over the next 5 years: game attendance and win percentage. The results showed that for a team with the first draft pick, fan attendance increased by 5 to 6 percent every subsequent year, and win percentage increased by between 8 and 9 percent by the 4^{th} year after the draft (Walters and Williams, 2012). This highlights the importance of the top pick in the NBA draft. We will study the value of draft picks in chapter 6.

Most major American sports leagues assign draft picks in inverse order of the final standings- the last-place team gets the first draft pick, the next-worst team gets the second draft pick, and so on. In contrast, the NBA draft order is determined partially by weighted

randomization; the 14 teams that missed the playoffs the season leading up to the draft are each given a weighted chance at one of the top three picks (Patt, 2015). Currently, the three teams with the worst record have the highest chance of the first pick. The subsequent teams have decreasing chances at those picks according to the standings (the fourth-worst team has a slightly higher chance at the top three picks than the fifth-worst team, and so on. This ensures that the teams who performed poorly in a given season get the best chance at signing future stars. Once the draft lottery is over and the top four picks have been distributed, the rest of the picks are given in inverse order of last season's standings. There has been some controversy in the past with the lottery, given that teams who barely missed the playoffs could acquire one of the top few picks (Patt, 2015). This is a product of the lottery system's attempt to discourage teams from intentionally losing in order to get a high pick, by not guaranteeing the highest pick to the worst team. Prior to 2019, the worst team in the league had more favorable odds than the rest of the league. The changes in 2019 to give the worst three teams equal odds was aimed to reduce tanking, the practice of hoarding draft capital and fielding poor teams to get high draft picks.

1.3.2 Free Agency Each off-season, NBA franchises are allowed to adjust their rosters during a period referred to as free agency. During this time, free agents, players without NBA contracts, are allowed to market their skills to teams in an attempt to get the best contract. Players become free agents either by going undrafted out of college or by allowing their current contracts to expire. There are two categories of free agents: unrestricted free agents (UFA), who are allowed to sign with any team, and restricted free agents (RFA), who are allowed to sign offer sheets with any team, which can then be matched by their previous team in order to retain

the player. Restricted free agency occurs when a player finishes his rookie contract[2], and the team wants to exercise the option to keep the player for a fourth year (Rosen, 2016).

The salary cap is the most important regulation related to free agency, and works by financially punishing teams if the aggregate payroll of their players exceeds a preset amount. Teams that go over this cap pay a dollar for dollar tax on every penny their payroll exceeds the cap, called the "luxury tax." There are some exceptions to the tax rules that have been implemented for various reasons. The 2017 Collective Bargaining Agreement (CBA) defines exactly what the rules and exceptions are for the salary cap (NBA, 2017). Large market teams, such as the Golden State Warriors, are more likely to exceed the cap to keep their rosters star-studded, paying huge penalties as a result. Teams can not, however, sign players freely if the contract would put them over the salary cap. Instead, to exceed the salary cap, a signing must fulfill an 'exception,' the details of which are outlined in the CBA. The most used exceptions are the mid-level and the bi-annual exceptions (refer to Appendix A for details), which allow teams over the cap to sign free agents to contracts of a predetermined value (NBA, 2017). Derrick Rose signed with the Pistons as a mid-level exception in 2019. There is a second "cap" called the apron, over which teams forfeit their right to the exceptions (Wade, 2013). As of 2018, this was set at $6 million above the cap. There are certain benefits for staying under the salary cap. Mainly, these include avoiding the luxury tax and having the ability to acquire players in free agency and via trade without restriction. (Please refer to Appendix A for additional details).

[2] Rookie Contract - Contract given to first year players drafted by the NBA teams which are guaranteed for first two years with teams having the option to extend the contracts in third and fourth year with the player's salary increasing exponentially each year.

1.3.3 Trade. Given that the draft occurs in the summer, and there are typically only a few available quality free agents, NBA managers that are dissatisfied with the structure of their team will often make trades. If a team decides to make a roster change mid-season, or in the offseason, they can swap future draft picks and/or contracted players with other teams. Cash is also occasionally included in these trades. Player value in a trade is influenced by several factors: experience, potential, contract terms, performance and even popularity (Borgehsi, 2009). Draft pick value is affected by the year of the draft, the pick number and the prospect pool available for that draft year (Borgehsi, 2009). Teams in different scenarios will benefit differently from different trades. A perennial contender and a team in the process of rebuilding would both be happy with a trade in which future picks and prospects from the contender are swapped for veteran talents. This can mean there is often no outright "winner" in trades; one team may benefit immediately from a trade, while the other benefits in 5 years, but both teams may simultaneously assess the trade as beneficial. We study trades in Chapter 7.

1.4 Important Rule/Regulation Notes.

NBA rules have evolved over the years, as our discussion of the salary cap above indicated. Some of the rules that have been updated relate to the game itself, while other rules relate to the manner in which the front offices conduct their business. In particular, we have chosen to perform our analysis only on the 1980 NBA season and beyond, because we believe the addition of the three-point shot drastically changed the sport. There are numerous other examples of rule changes, such as the change in the rules that prohibit hand checking, which was the act of using one's hands to physically limit an offensive player's movement (this change occurred in 2004). Changes in rules affecting the front office during this period include changes

in the trade deadline, and a rise in the salary cap (Conway, 2017). Details on rule changes can be found in Appendix A.

The rest of this thesis is organized as follows. In Chapter 2 we conduct a literature review, describing scholarly works that form the context for our own work. In Chapter 3 we describe the data we have collected, its source, and its organization. In Chapters 4 through 8 we provide in depth analyses of important aspects of NBA team building and evaluation. In Chapter 9 we provide concluding remarks.

2. LITERATURE REVIEW

The literature review is organized as follows. First we summarize the research that has been done on NBA team performance and a few leading publications of sport team performance in different sports. Followed by an overview of studies evaluating player performance in the NBA. After discussing studies of the impact of coaches, team cohesion and continuity, we summarize the research of strategic team decision making in the NBA.

2.1 Team Performance

There are certain statistics factors that are especially important to look at when quantifying a team's success. These are effective field goal percentage, turnovers, free throw percentage, and offensive rebounds. Together, these statistics are called *The Four Factors* and they have been shown to be the most important factors to incorporate when analysing a team's potential playoff success (Baghal, 2012). The paper developed measures of "offensive and defensive quality" of teams to assess a team's general play style and based on structural equation modeling concluded that offensive quality affects winning more than defensive quality (Baghal,

2012). The conclusion of another study that analyzed team's playoff success, including teams' offense, defense, their bench, coaching, and other intangible factors concluded that offensive ability is often the factor that is most important in playoff predictions (Jenkins, 2019).

A different study analyzed the relationship between teamwork (as measured by assists) and the teams' win/loss percentage (W/L%) and found a statistically significant correlation between the two factors. Higher correlations were measured when the entire team's assists were factored in, not just those of the five starters. In addition, assisted points were positively correlated with winning, meaning higher assisted points correlated to a higher win loss percentage. Unassisted points showed a negative correlation, meaning the more unassisted points a team scores, the lower their W/L% (Melnick 2001).

In 2013, a paper investigated the Spanish professional basketball league to find indicators of team success. In the regular season, winning teams dominated several categories as expected: assists, defensive rebounds, and both 2 and 3 point field goals . In the playoffs, however, winning teams were superior only in defensive rebounding . Despite these results the best defensive rebounders were not always the highest paid players, illuminating a mismatch between players' on-court value and their compensation (Garcia et al., 2013).

A study (Hofler and Payne, 1997) looked to measure efficiency among NBA teams in the 1992-1993 season and determine if teams were living up to their "potential", where potential is defined as the maximum attainable wins given a team's assets (players, coach, etc.). Their models included as independent variables a variety of statistics including wins, field goal percentage, free throw percentage, steals assists, offensive rebounds, defensive rebounds, turnovers, and differences in blocked shots. They concluded defensive rebounds contribute to

winning, however their model had a lot of multicollinearity. Offensive rebounds, defensive rebounds, field goal percentage, steals, turnovers were all collinear, which challenges their interpretation and warrants further evaluation of the stats individual importance. The teams were estimated to be 89% efficient on average, where efficiency is what proportion of potential is achieved. No team was 100% efficient (Hofler and Payne 1997).

Another study conducted by Young Hoon Lee and David Berri additionally used position types to measure efficiency in the same as the aforementioned Hofler and Payne study. They divided players into guards, small forwards, and big men (power forwards and centers). The study found that big men are more valuable to teams (in regards to efficiency) and as a result, teams tend to draft big men over small forwards or guards (Lee and Berri 2008). This study was the first to quantify the "big men" phenomenon.

Finally, there is not much other literature available on team play style and success. A single paper published in 2019 explored the relationship between a team's three-point attempts and teams' revenue and found that three-point attempt rate had a negative effect on revenues (Harrison, 2019). We study the relationship between W/L% and teams' play style in Chapter 4.

2.2 Player Performance

NBA teams want to make sure that they are spending their money efficiently. A major decision by NBA teams is deciding which players to acquire via trade, draft, or free agency. In these decisions they want to ensure the acquired player will help them win games. Players are usually evaluated by how much they contribute to their team compared to their salary. To that end, factors such as rebounds and field goal percentage as well as minutes played are often used

to determine player efficiency and performance. Multiple academic papers focus on evaluating player efficiencies, which we summarize below.

2.2.1 Player Evaluation in the NBA. A 2012 study suggests that the best way for a team to determine if a certain player is worth bringing to the team is not by using past statistics, but rather to try to predict their future. The author used a simple model using only points to predict a player's future and normalized based on the number of games played. The model was validated on players who had already finished playing in the NBA, so their future was known. In addition, their model factored in time in the NBA and a player's age. Players often need a bit of adjustment time when they first enter the league, and players past their prime will likely decline. These two factors should be taken into consideration (Hwang 2012).

A second study aimed at valuing player efficiency used a slacks-based[3] efficiency measure to study the price of performance, focusing on field goals, free throws, offensive rebounds, defensive rebounds, steals, blocks, and points, while also taking into account minutes played. The model evaluated the efficiency (which the authors named Player Impact Estimate and captures the amount spent for statistical performance) of 95 players from the 2008-09 season through the 2015-16 season. Four players, one of which was Kevin Durant, had a efficiency ranking of 1, the highest possible score. Most other players were lower and closer to each other, indicating most players are not performing with 100% efficiency (Asghar et. al 2018).

In a third efficiency study, a DEA model was used to evaluate efficiency in a similar way as the slacks-based model mentioned in the previous paragraph. The DEA model used salary and minutes played as inputs. The outputs were slightly different as this model used points, assists,

[3] Slacks-based models must meet two conditions: "measure should be invariant concerning units of data" and "the measure should be monotone decreasing in each slack input and output" (Asghar et. al 2018).

rebounds, steals, turnovers, and blocked shots. Their measure of efficiency again is based on statistical output as a function of salary and minutes played. They applied their model to data from just the 2011-12 season and only on 26 players, all guards. Using a DEA model, 10 players were efficient, including Derrick Rose and Russell Westbrook. Of the 26 guards measured, Kobe Bryant ranked 25 out of 26 in terms of efficiency (at the time Kobe Bryant was towards the end of his career and was being paid very well; only Joe Johnson was less efficient (Radovanović et. al 2013).

Player impact was studied in a 2016 paper. A player who scores when their team is in a close game is much more impactful than a player who scores when their team is up by a large number of points. This paper accounted for the situation the player's team was in during their play using a Bayesian linear regression model. Controlling for other players on the court, the model identifies highly paid players who are not impactful to their team and also identifies players who are making an impact that is not being measured by current statistics. They used data from the 2006-2007 season through the 2012-2013 season and determined how likely a team was to win at each point in the game. Players were given an impact rating based on how impactful they were to their team. Since these rankings are team-based, rankings are determined only within their team and not with respect to the entire NBA (Deshpande and Jensen 2016).

A player's performance is influenced by his teammates. A recent study analyzed how a player who plays with other players who complement his play style will impact his performance, highlighting that the style of play can impact a player's performance. A tree model was used to capture the various different outcomes during a team's possession. An individual player model was then used to evaluate a player's performance with respect to his other teammates to estimate

the probability an event in the tree model will occur (ie. 2-point basket, 3-point basket, turnover, etc.). Players were then given ratings based on how likely they were to perform certain actions (shooting, scoring, turnovers, etc.) and then players were evaluated based on the other players they were on the court with to determine how they impacted their team. According to the model that analyzed the 2014-15 season, Anthony Davis had the largest expected increase in team points of any player in the league. This model also looked at how other players impacted their teammates performance (based on the increase of their teammate's expected points per possession) to determine who was a good teammate. Wesley Johnson was considering a good teammate as he increased his other teammates' expected points per possession the most, and Russell Westbrook was the worst teammate as he decreased his teammates' expected points per possession the most (Kuehn 2016).

Another study took a similar approach of examining synergies of skills within certain team line-ups. The authors found the probabilities of various game events occurring given a current game state, and were thus able to simulate whole basketball games. With these simulations, they found that players' skill sets could produce either positive or negative synergies amongst their line-up. With these predicted synergies, mutually-beneficial trades were then suggested for pairs of NBA teams (Maymin 2013).

Another major decision teams make is whether or not to invest in a "superstar" player. They must determine if that player is worth the expenditure and if that investment in a single player is the best way to help the team win. A study published in 2012 sought to determine the effect of a "superstar" player on a team. Based on data from 2006-07 through the 2010-11 season the authors constructed models to predict win percentage taking into account the players´salary

differences. The authors found that payroll inequality had a negative impact on overall team performance, but a player´s individual salary was positively correlated with his individual performance. The researchers concluded that teams should try to sign a big name free agent and pay him well, but keep other player's salaries on the team somewhat even or closer together in value (Lundgren 2012).

A study assessed changes in player performance in "contract years" - the year before a player enters free agency or signs an exception with his current team. The researcher utilized a regression model with crafted variables accounting for contract length and position as well as traditional statistics. Ultimately the analysis shows that players perform better in contract years by 3-5 percent in terms of win shares and PER (player efficiency rating). This shows that players may be additionally incentivized to play harder in seasons leading up to new contracts (Ryan 2015).

2.2.2 Player Evaluation in Other Sports. There is a major gap in the literature in terms of evaluating NBA player's salaries versus their on court performance and using on-court performance to predict salaries. We have found no published works on the subject for the NBA, and we seek to address that gap in literature in this thesis. While there is a lack of these studies in the NBA, similar studies have been done in other sports. "Great Expectations", published in 2016 is one such study. The researchers looked at player contracts in Major League Baseball from 1998 to 2014, and used advanced metrics to project both player contract value and player performance throughout the life of the contract. The paper draws valuable conclusions about the relative valuing of players by teams and inefficiencies in the baseball free agency market (Barnes, Bjarnadottir, 2016). We model our free agency analysis on the work done in this study.

In a 2012 study on player performance in the National Hockey league, players were divided by position, and further divided into position subgroups (Chan et al., 2012). The researchers employed neural networks that considered several factors such as individual statistics, team statistics, and players' contribution to the salary cap. They then classified players into different 'types' and analyzed which type of player is most effective at generating team success (Chan et al., 2012). While the conclusion that goalies generate the most value per player of any position does not apply to our research, the methods contained within the paper are relevant.

In summary, there are multiple ways to evaluate a player and many factors can impact the performance of a player. Most commonly players are evaluated using measures such as free throws, rebounds, and field goal percentages which is then compared to their salary to determine if they're playing efficiently (Radovanović et. al 2013). There are also studies from other sports that compare the predicted performance of players to their expected salary (Barnes, Bjarnadottir 2016). The literature also highlights the importance of considering the impact players have on each other since basketball is a team sport; teams should consider if having a player on their team, positively or negatively impacts the team beyond what is captured by commonly used player performance statistics (Deshpande and Jensen 2016). Player evaluation can be done in a variety of ways, using a variety of models, and often, these models must be used together to get the most accurate description of a player's performance and contribution.

2.3 Impact of Indirect Factors

There are other factors besides individual player performance that impact sport outcomes. How well players work together, how long they've been together, and coaching are other factors that impact a team's success.

2.3.1 Cohesion and Roster Continuity. Cohesion refers to how well a team perceives themselves working together as a unit. A meta-analysis of 46 different sports cohesion studies found a strong relationship between cohesion and performance across teams in different competitive levels and different sports, albeit less pronounced at a professional level (Carron et. al, 2002a). Specifically, in basketball a study of elite university basketball teams in Canada, it was found that teams that had a stronger sense of cohesion were expected to have a higher winning percentage in their league (Carron et. al 2002b). There are a myriad of other studies, (see Marcos 2010, Heuze 2006, Farneti 2008, and more) studying cohesion in basketball teams but not in the NBA. This is due to the fact that many of these studies are quantifying cohesion through directly surveying the players, something which may not be possible in the NBA. These studies found that generally, cohesion has a positive contribution to team success,

One of the factors that may play into cohesion is roster continuity. According to Basketball Reference, roster continuity is defined as the percentage of players from a previous year's roster that is on the roster for the following year. The effect of roster cohesion is understudied in the literature, but its impact has been discussed by both pundits and hobbyists, but not in a rigorous fashion (Adams, 2019) (Brooke, 2019). The intuition is that greater roster continuity would contribute to greater cohesion. Roster turnover in the NBA, when defined as the percentage of players that played in more than 60% of games in a season, has been found to show a negligible effect on game attendance in the NBA (Morse, et. al 2007).

2.3.2 Coaching. In the NBA, coaches play a large role in how individual games are played. However, their contributions are more difficult to quantify and aggregate because they do not have their own box scores like players do. The majority of the literature on coaching deals primarily with non-performance analysis of coaching styles (see Colquitt 2007), play by analysis (see McIntyre 2016), and qualitative analysis of coaching (see Mielke 2007 and Kihl and Richardson 2009). A recent study developed a framework for examining leadership effectiveness, called RIFLE, and applied it to basketball. They found that coaches explain about 20-30% of the variation in points scored and allowed in the NBA and Division I basketball (Berry and Fowler, 2019). To the best of our knowledge, only a single additional study aims at quantifying coaching impact. The study found that 61% of NBA teams improve their performance in the year after a coach is fired (Martinez and Caudill, 2013). In our study of team performance in Chapter 5 we will include key coaching variables.

2.4 Salary Drivers

Determining the appropriate pay for any player, including free agents is a challenging multifaceted problem. Players with a significant improvement in performance during the stress of the playoffs are sometimes referred to as “clutch” players. A 2005 study (Berri and Eschker, 2005) showed that almost none of the big stars in NBA history actually played consistently better in the playoffs. Even Michael Jordan, often touted as the greatest of all time, was not ‘clutch’ by the definition applied in the study. One paper found that teams in the NBA had been compensating players for a single statistic, points scored, more than any other (Mandle, 1998). Team and defensive stats, such as assists and rebounds, had no statistically significant correlation with salary. Mandle argued that players responded to this trend by shooting more, even if passing

might be the better option. These studies highlight the fact that NBA teams may have been overpaying scorers and "clutch" players while neglecting cheaper players who could potentially impact the team just as much. There is a significant gap in research in this area, as the few studies available are dated. We contribute to this literature with our study in Chapter 8 where we study, using recent data, players' salaries and further discuss the inefficiencies of player market value and how to exploit them.

2.5 Evaluating Draft Picks

When building a roster, determining the value of draft picks and the emphasis teams should place on acquiring them is a similarly difficult task, full of uncertainties and risks. In one paper (Motomura et al, 2016), the authors sought to determine the effectiveness of the strategy of "tanking" in order to gain higher draft picks in the future. Running several regression models on the number of recently acquired draft picks, in addition to other factors, showed that these draftees did not on expectation lead to higher rates of success. The authors indicated that their analysis instead supported the build-up of a roster through excellence in general management and other front-office decisions. Another study looked at a subsection of draft picks, specifically those of international players (Motomura, 2016). Motomura found that the valuation of these international players in the draft, measured by the order in which they were picked, relative to their subsequent performance in the NBA fluctuated. Teams first undervalued international players, with those picked in 1999-2001 outperforming expectations at their draft position. However, teams soon overreacted to these promising outcomes, and in the following years consistently overvalued international players in the draft. These papers both highlight the fact that teams are not perfectly rational in their reliance on and valuations of draft picks as a source

of team success. We further analyze actual draft pick performance in Chapter 6 and perceived valuations of draft picks in trades in Chapter 7.

3. DATA

3.1 Data Collection

Our analysis relies on a wide range of publicly available NBA data. Using web-scraping libraries in Python, such as Pandas and Beautiful Soup, we collected data from the 1979-1980 to 2017-2018 NBA seasons from three online sources: Basketball Reference, Pro-Sports Transactions, and Spotrac. For a sense of scale, this included approximately 1000 seasons of team performance data and 18,000 seasons of player performance data for over 3000 players. We also scraped over 35,000 transactions, including free agent signings, trades, and drafts, as well as 26,000 injury related events. After initially storing the data in comma separated value files, we organized the data into a SQLite database. Most fields are linked to either a team in a given season, or a player in a given season. This allowed for easier access to relevant data, especially when attempting to conduct an analysis across multiple data sources or categories. An overview of the data included in the database is displayed below.

Table 3.1: Representative fields in each of the database tables.

Table Name	Highlighted Fields
Franchise	Name, Abbreviation
Team	Franchise, Name, Year, W/L%
Team Totals	Field Goals, Field Goals Attempted, Assists, Steals, Offensive Rebounds
Team Advanced	Pace, Offensive Rating, Effective Field Goal Percentage
Player	Name, Height, Birth Year

Player Totals	Field Goals, Field Goals Attempted, Assists, Steals, Offensive Rebounds
Player Advanced	Player Efficiency Rating (PER), Win Shares, Value Over Replacement Player (VORP)
Salary	Salary (Annualized)
Injury	Relinquished/Acquired, Injury Description
Trade	Teams, Player Traded, Draft Pick Traded
Draft	Round, Pick

Throughout our study, we use a common transformation to translate salaries in dollars to salaries as a percentage of the salary cap. For example, in 2018, Giannis Antetokounmpo made $22,471,910, which accounted for 22.67% of the salary cap. This addresses two issues: inflation - one dollar in 1980 has a different value than one dollar in 2018, and the change in salary cap - one million dollars was a much greater portion of the salary cap in 1980 than it is today, even after accounting for inflation.

3.2 Important Events

There are two types of major events that have impacted the league in a large way during our study period: lockouts and league expansions.

Every five to six years, the National Basketball Players' Association, a union for players in the NBA, negotiates with the NBA itself in order to negotiate a new Collective Bargaining Agreement (CBA). The CBA is used to determine revenue share between the owners and players, the salary cap, and all other terms and conditions of players' employment in the NBA. If a CBA agreement is not reached before the previous agreement expires, a lockout (the owners hold out) or a strike (the players hold out) occurs, meaning all league activities are suspended

until a new CBA is agreed upon. Most recently, the NBA had the fourth lockout in their history delaying the start of the 2011-2012 regular season from November 1 to December 25. In our analysis, we have made adjustments to account for the impact of lockout shortened seasons and the effect they have on player and team statistics by favoring per-game stats and winning percentages over cumulative player stats and total wins.

League expansions change the size of the league by adding additional teams, which causes changes in the playoff system so more teams can be accommodated. In the 1979 season, the league had 22 teams, and it now has 30. The Dallas Mavericks joined in 1980, the Miami Heat and Charlotte Hornets in 1988, the Minnesota Timberwolves and Orlando Magic in 1989, the Memphis (then Vancouver) Grizzlies and Toronto Raptors in 1995, the New Orleans Pelicans (then, the New Orleans Hornets) in 2002, and in 2004 a new team, the Charlotte Bobcats was founded and joined the NBA (when the New Orleans Hornets changed their name to the Pelicans in 2014, and the Bobcats subsequently changed their name to the Hornets, the league decided that the current Charlotte Hornets should assume the history of the original Charlotte Hornets before they moved to New Orleans). Expansions also include a redistribution of players in the 'expansion draft'. In the expansion draft, the new franchise is allowed to acquire certain 'unprotected' players from other teams. The specifics of forming expansion teams are not relevant to our project, but the change in number of teams is obviously relevant.

4. TEAM PLAY STYLE AND ITS IMPACT ON SUCCESS

It has long been thought that a team's play style affects any team's chance of winning. However limited formal studies have been conducted. We hypothesised that the way a team plays has a

direct effect on the team's chance of winning a championship or games and therefore certain play styles should have a better chance of winning the league. In this chapter we applied formal analysis to define play styles and classify teams by their play style. In particular we applied clustering to understand play style, which enabled us to identify the distinct play styles played throughout the decades. Using these distinct play styles, we then studied the teams' performance as a function of play styles.

4.1 Play Style Clustering

In order to study how play styles affect success it was important that the statistics not be directly linked to how good a team was. Rather, we wanted to characterize different aspects of a team's style. Therefore we selected, for example, to include three point attempts, but to leave out statistics related to the efficiency or percentages of the shots. The goal was to group teams in how they play but not necessarily how successful they are in their play styles. The statistics included and the justification for each is included in list below:

3PA – Three Points Attempts – Included due to its significant impact on the way a team plays. 3PA is a significant indicator of how a team breaks aparts a defense.

2PA - Two Points Attempts – Included due to it being the main way a team scores most of the time. Also can be indicative of a team play style

FTA – Free Throw Attempts – shows how much the other team is fouling which reflects your play style. Sign of aggressiveness/faster pace/screen plays/etc

AST – Assist – an indication of how much the team moves the ball around. Less means that the team could be relying on one individual's skill and isolation plays.

ORB – Offensive Rebounds – an indication of the team's offensive ability as well as your rebounding ability – indication of what types of shots the team shoots

DRB – Defensive Rebounds – an indication of the team's defensive abilities as well as your rebounding – indication of how well the team defends

Opp_FTA – opponent free throws – indication of how aggressive the team is on defense

Pace – pace – indication of how fast the team plays

Opp_3PA - opponent 3 point attempts - indication of type of defense the team plays

Opp_2PA - opponent 2 point attempts - indication of type of defense the team plays

In this analysis, we have included all the teams in the NBA since and including the 1980-1981 season, a total of 38 NBA seasons consisting of 1073 team seasons.

We applied k-mean clustering, a commonly applied clustering algorithm (implemented with Pythons' sklearn package). To account for the improved athleticism over time, we first normalized the stats by year. This enabled us to study the relative characteristics of a team across our study period.

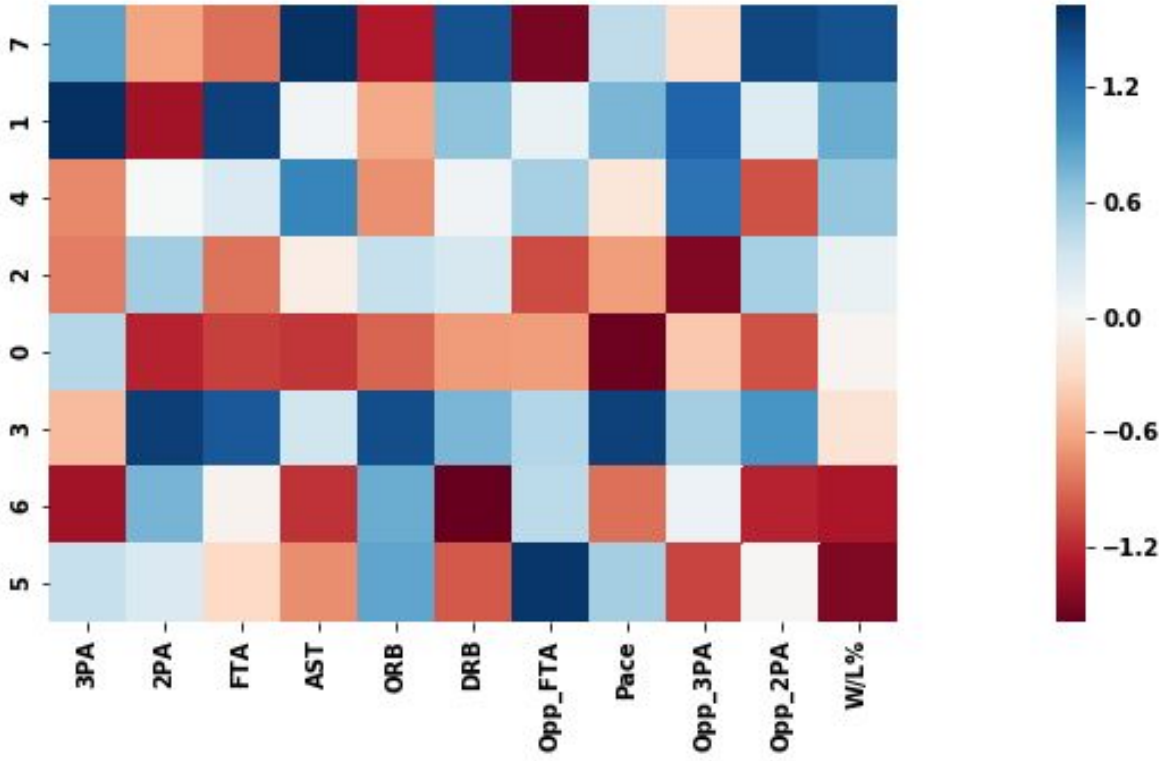


Figure 4.1. Heat Map of the Cluster vs the win percentage

Figure 4.1 is a heat map that highlights team play styles as a function of their clusters (numbered from 0 through 7). The heat map can be interpreted in following ways. The red indicates for the team-seasons in the cluster have on average lower than the average statistics compared to the other teams during a season for that particular stat measured by standard deviation. The blue indicates the opposite of red, the teams' stat is typically higher than average, compared to the other teams. For example, teams in cluster 0 have typically higher than average 3PA compared to the other teams but lower than average for 2PA. The heatmap shows that the teams in the cluster are characterized by a play style that has high FTA, low AST, higher than average ORB, lower than average DRB, and lower than average opponent FTA, opponent 3

point attempts, opponent 2 point attempts and Pace. This indicates that the teams in cluster 0 mainly like to like to slow down the plays and get behind the arc to score with isolation plays. Furthermore, the teams like to attack the basket when they are on offense but during the defense, the team is less aggressive. The last column represents the average win loss percentage for that certain cluster with blue representing higher than average W/L% and red representing lower than average W/L%.

Cluster 0 is characterized by high 3 point attempts, and low 2 points attempts, free throw attempts, assists, offensive rebounds, defensive rebounds, pace, and opponent free throw attempts. Generally, teams in cluster 0 like to slow down the plays and get behind the arc to score with isolation plays.

Cluster 1 is characterized by high 3 points attempts, free throw attempts, opponent free throws, pace, opponent 3 point attempts and low 2 point attempts, offensive rebounds. This play style likes to score behind the paint with fast playstyle in both offense and defense leading to high free throw attempts on both ends of the court.

Cluster 2 is characterized by high 2 point attempts, offensive rebounds, opponent free throw attempts, and low 3 point attempts. This play style likes to play fast and get inside the paint to score leading to higher free throw attempts than other play styles. High offensive rebounds could mean this play style is not very efficient in shooting or shooting contested shots. This play style utilizes aggressive defense leading to high opponent free throw attempts.

Cluster 3 is characterized by high 2 point attempts, free throw attempts, offensive rebounds, pace, opponent 2 point attempts, and low 3 point attempts. This play style utilizes fast paced ball

movements and shoots inside the paint leading to higher offensive rebounds. Their defense likes to keep the other teams from shooting inside the paint than outside.

Cluster 4 is characterized by high assist, opponent free throw, opponent 3 point attempts, and low 3 point attempts, offensive rebounds and opponent 2 point attempts. This play style relies on getting inside the paint to score efficiently with good ball movements while playing a fast paced game. This play style utilizes aggressive defense with focus on trying to stop players from getting inside the paint.

Cluster 5 is characterized by high opponent free throw attempts, offensive rebounds, and low assists, defensive rebounds, and opponent 3 point attempts. The play style likes to utilize isolation plays with focus on getting inside the paint leading to high offensive rebounds. In defense, the playstyle's focus is to stop players from shooting outside the paint.

Cluster 6 is characterized by high 2 point attempts, offensive rebounds and low 3 point attempts, assists, defensive rebounds, pace, opponent 2 point attempts. This play style gets inside the paint with emphasis on stopping the other team from getting inside the paint.

Cluster 7 is characterized by high 3 point attempts, assists, defensive rebounds, opponent 2 point attempts, and low 2 points attempts, free throw attempts, offensive rebounds, opponent free throw attempts. This play style likes to shoot from behind the arc with good ball movement with good defence with emphasis on stopping the other teams from shooting threes.

This analysis highlights that a high win percentage is correlated with multiple play styles. For example, play styles represented by clusters 7 and 4 are two different ways a team can play and still achieve a high win percentage and to certain extent, play style 1 and 2 also demonstrate

a way that team can play to achieve good results. In summary, the heatmap in Figure 4.1 highlighted that certain play styles (represented by clusters) had better win percentages than the others. This can be better seen in Figure 4.2 which is a boxplot of each play style and the spread of win percentage for each cluster. Figure 4.2 again demonstrates that a certain team's play style has a higher win percentage than the others - but also highlighting a wide variability in W/L%, and in most cases it is possible to perform badly independent of play style.

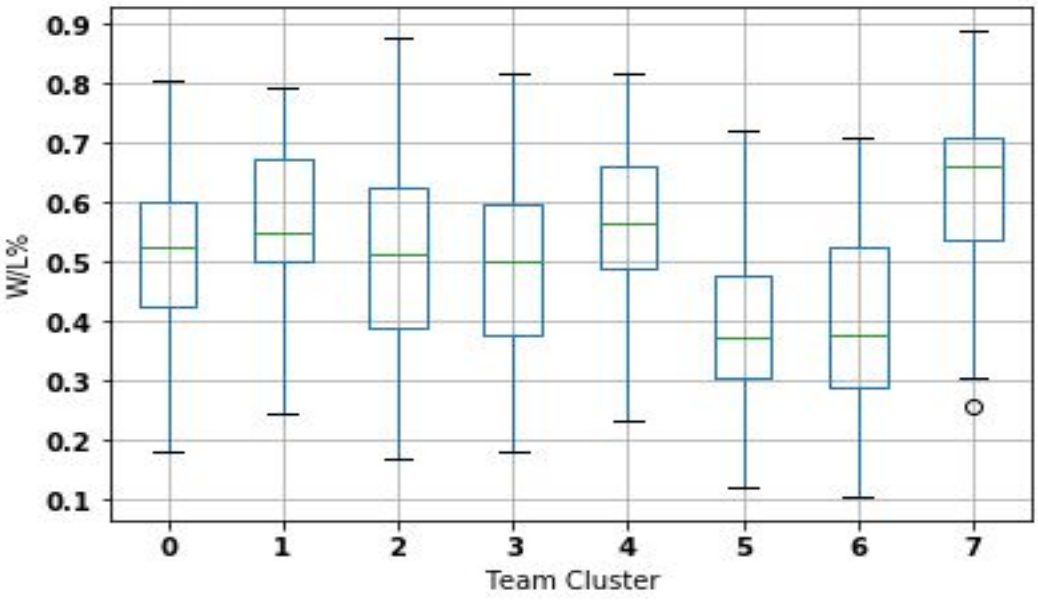


Figure 4.2. Team Cluster vs Team Win Percentage

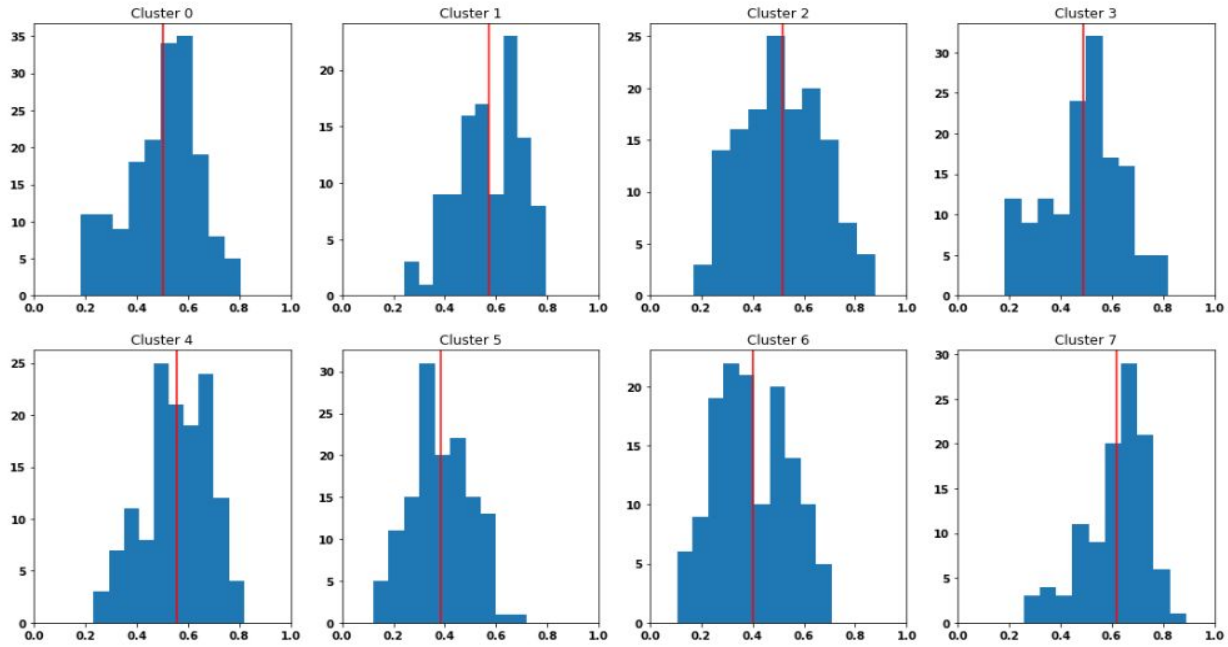


Figure 4.3. W/L Percentage Distribution per Cluster for Normalized Data with red line as the mean.

Figure 4.3 dives even further into the W/L% of each cluster and shows the W/L% distribution for each cluster. For example, cluster 0 almost resembles a normal distribution with 0.5 as the peak which is a play style characterized by slow pace with offensive tactics that gets inside the paint to score with average defense. On the other hand, cluster 7's distribution is shifted to the right highlighting the relative success of this type of play style which is characterized by good ball movements and 3 point shootings. Interestingly, most cluster's W/L% resembles a normal distribution.

Figure 4.4 represents the year distribution for each cluster. For example, cluster 0 consists of many teams from 1980 to 2018 with a slight dip during the late 1990s. Significantly, cluster 2 and 6 playing style has had a steady rise in usage through recent years which is interesting

because this play style involves shooting inside the arc contrary to the rise of 3 point shooting. Also, cluster 1 has seen a steady rise in usage which makes sense because this play style revolves around shooting behind the arc to score which has become a popular tactic the past decade. On the whole all of the clusters have been prevalent in the NBA throughout the 1980 to 2018 period. The clusters are capturing the play styles based on the league's average for that season and not play styles compared to teams across 1980 to 2018. By capturing the playing styles compared to the playing styles of that era (through normalization of the data), we saw which playing styles were strongly correlated with win percentage and which was not. We contrast our results with not normalizing the data in the sensitivity analysis below.

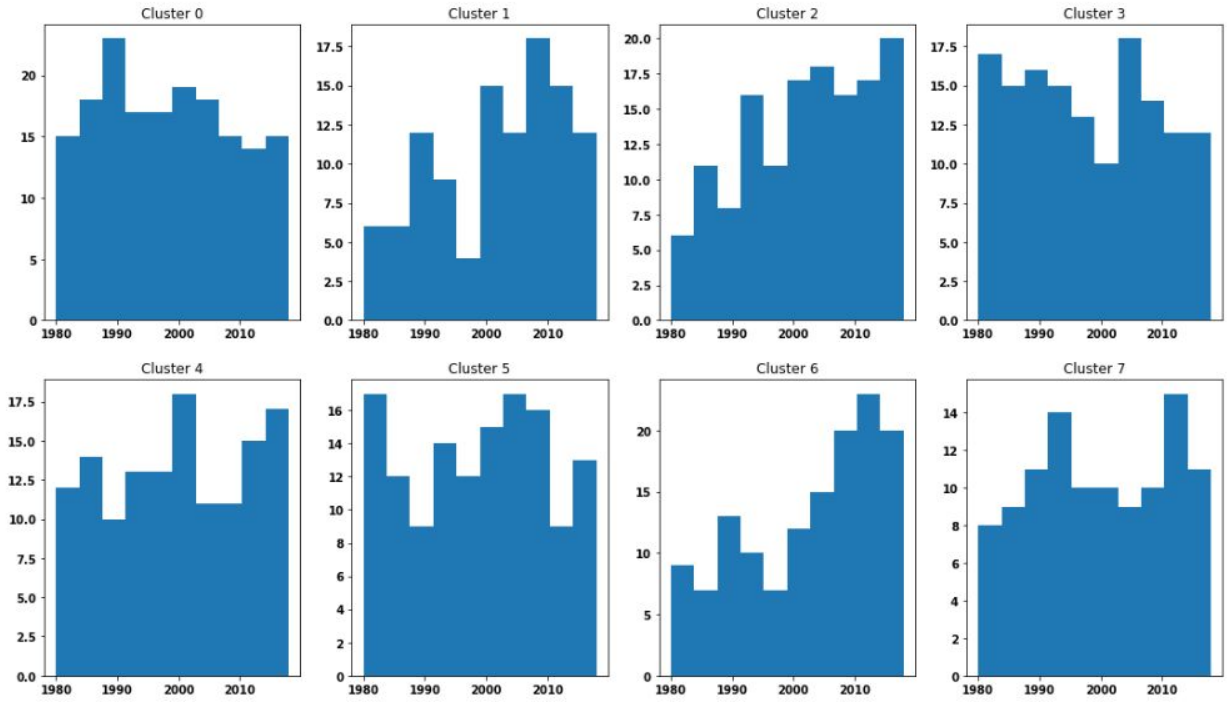


Figure 4.4. Year Distribution per Cluster for Normalized Data

4.2 Cluster Movements

Our analysis above showed that the team's play style is correlated with team's performance. Therefore, we investigated the impact of team movement from one cluster to another; each cluster movement represents a change in a team's play style. We define cluster movements as the change in the team's cluster during a 2 season period. We refer to the original cluster as the initial cluster and the ending cluster as the final cluster. For example, if a team was at cluster 0 then in the next season moved to cluster 3, then the initial cluster would be cluster 0 and final cluster would be cluster 3. There are a total of 56 possible cluster movements (there are 8 clusters, and from any of these clusters a team can move any of the other 7 clusters). As a result, 56 play styles were analyzed focusing on the change in W/L%.

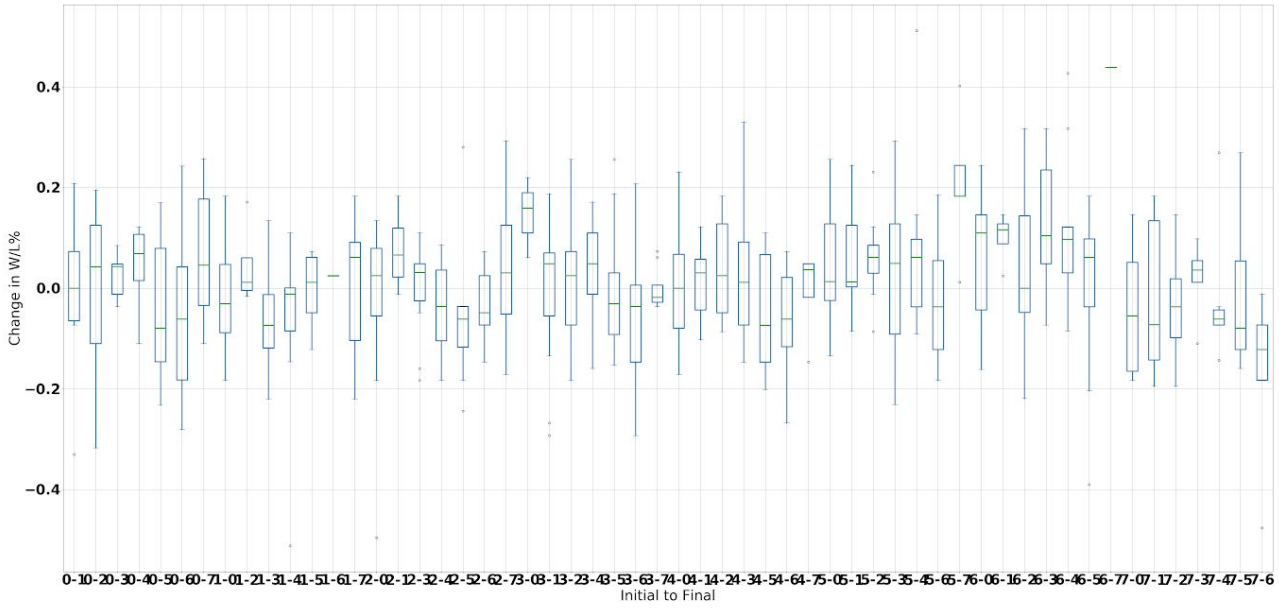


Figure 4.5. Plot of change in W/L% vs the change in team cluster in 2 season

Figure 4.5 indicates that certain movement from one play style to another has more impact than others. For example, the change from cluster 6 to cluster 3 generally has a large positive change in win percentage while movement from cluster 7 to 4 is correlated with a slight reduction in win percentage. This highlights that historically, certain cluster movements have had a greater impact on a team's W/L% than others which can be used to inform a team's decision when shifting from one playing style to another. However, sample size is worth considering, as for many of the cluster movements the number of teams making the chance is often small.

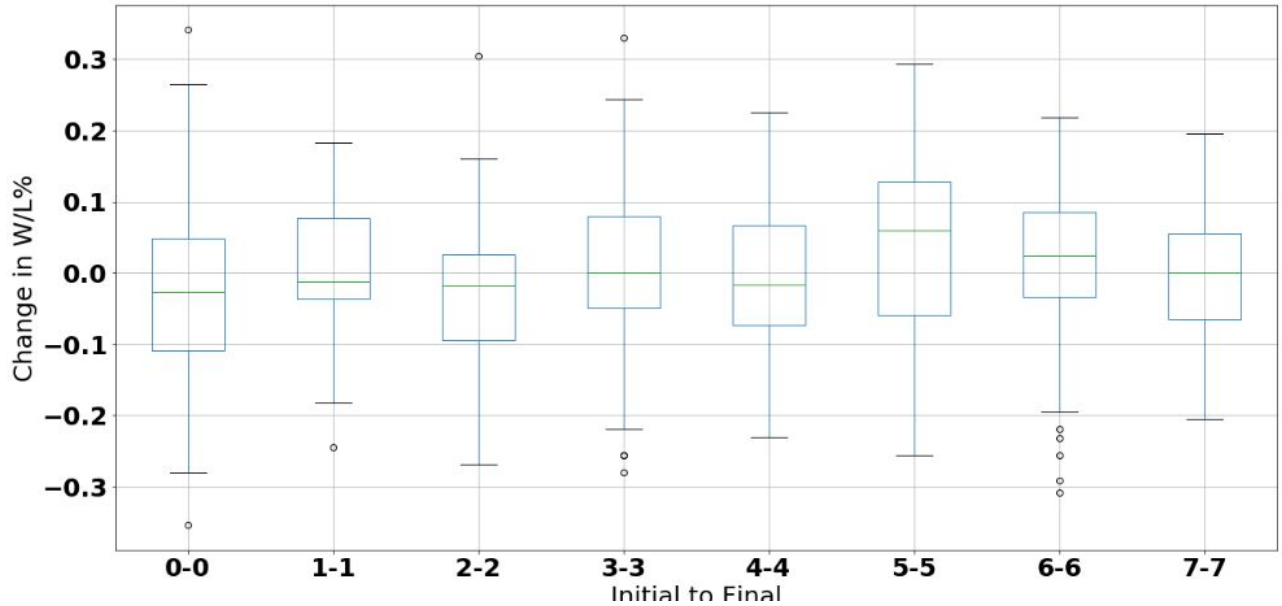


Figure 4.6. Change in W/L% vs Staying in same Cluster/Play style

Figure 4.6 shows the change in win percentages when the team stays in the same cluster in two consecutive years. The figure indicates that on average , there is limited impact on W/L% when a team stays in the same cluster. We however note on average positive impact on staying in

clusters 5 and 6, while the on average negative impact when staying in cluster 1. We also note the large variability in the impact of staying within the same play style cluster.

4.3 Sensitivity Analysis - Results Without Time-Normalization

In the NBA there are certain team play styles that are more popular than the others each year, and the game has evolved through the decades. For example, the current trend is to shoot threes often and play small ball - a play style defined by floor spacing and playing without traditional “big men” on the floor. In this sensitivity analysis we do not account for trends in play styles and do not normalize the team statistics by season. In this case the results will highlight how play styles and team statistics have evolved as opposed to comparing how a certain play style will do across our stuy horizon.

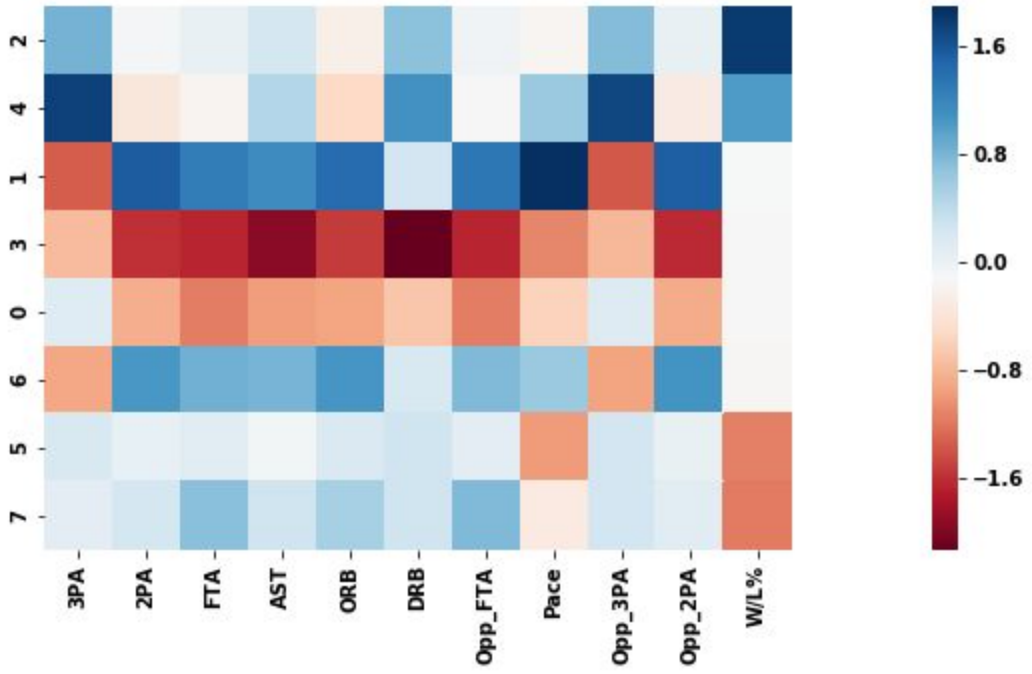


Figure 4.7. Heatmap of the Cluster vs the win percentage with the non normalized approach

In figure 4.7, we can first notice that except for cluster 0 and 3, most squares are blue. This is because of the lockouts in the 1998-1999 and 2010-2011 seasons. Because there were fewer games during those seasons, the numbers for those years are low which makes the rest of the years significantly positive compared to those years. This is highlighted in Figure 4.8 that shows that Cluster 3 contains the 1998-1999 season teams and cluster 0 contains the 2010-2011 seasons teams.

Without normalizing the data the clustering shows the transitions of play style throughout the years, which is highlighted in Figure 4.8. In contrast to the normalized results, where a play style typically was represented over a decade or two, the non-normalized clusters typically apply to a short period of time, typically about five years. As stated above, cluster 0 and 3 represents the lockout year which is why it has been separated into its own category by the clustering algorithm. More interesting things to note would be clusters 1,2, 4 and 6. We can see that for cluster 1 and 6, year distribution ranges from 1980 to 2000 while cluster 2 and 4 ranged from mostly mid 2010s to current. This shows the development of playing style throughout the NBA. The most significant change from early days of basketball and now would be the change from big man focused tactics to small ball. Big man tactics tended to concentrate their offensive capability to the biggest man on the team to score inside the paint and focused on physical defending. This can be reflected on cluster 1 and 6. Cluster 1 and 6's year distribution was from the 1980s to 2000s and both clusters liked to play big man focused tactics. Both cluster 1 and 6 have less 3PA, high 2PA, with high pace. In the modern NBA, many teams are focusing on the movement of balls using screens and each player's shooting ability from behind the arc leading

to an increase in 3 point attempts. This trend can be seen in Cluster 2 and 4. Cluster 2 and 4 have very high 3PA and less 2PA which indicates the departure from scoring from the paint to behind the arc. Furthermore, the increase in AST can be contributed toward the developments of ball movements and screen tactics.

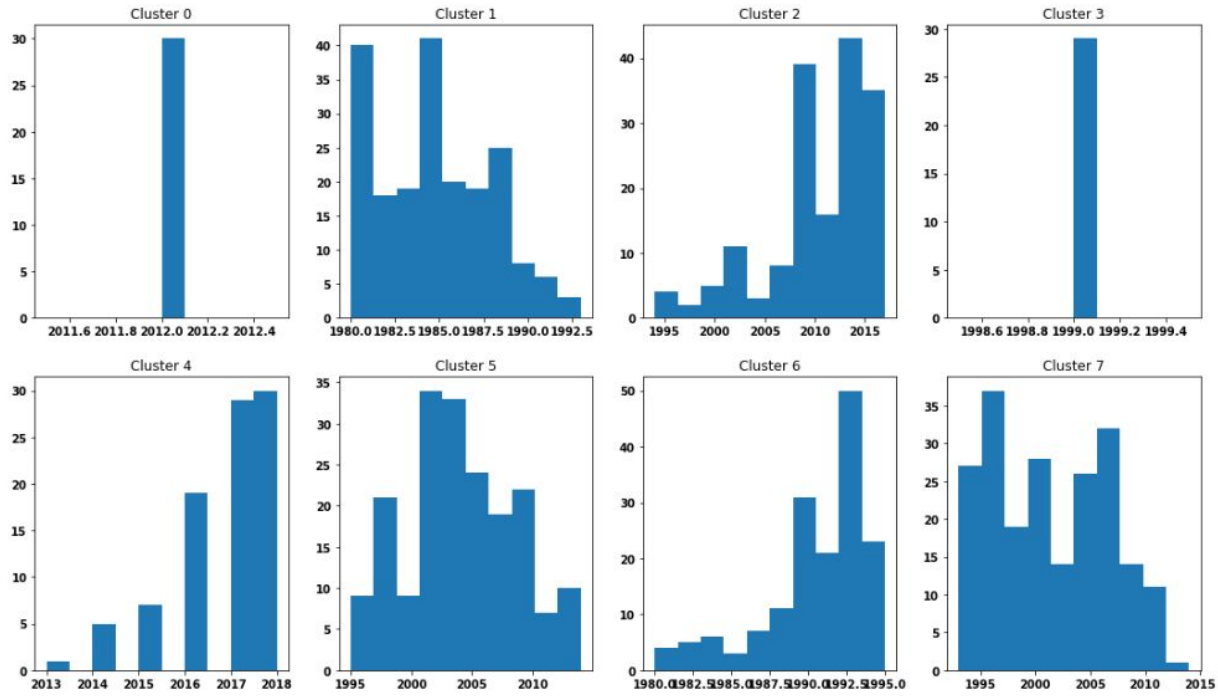


Figure 4.8. Year Distribution per Cluster for Not Normalized Data

Figure 4.9 shows the distribution of the W/L% distribution per cluster. The Figure highlights the fact that the non-normalized analysis, since they play style clusters capture overarcing play styles at a period in time, results in clusters with a W/L% median of close .5 while the normalized analysis is able to better relate play styles to performance.

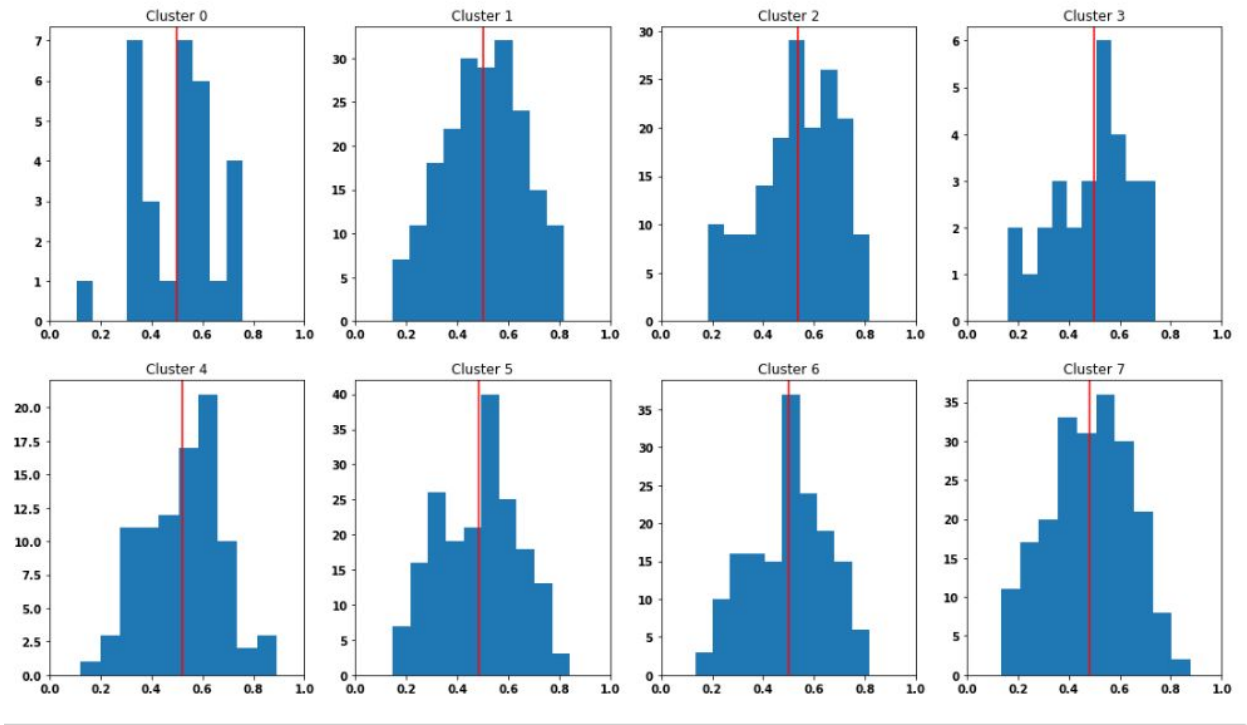


Figure 4.9. W/L Percentage Distribution per Cluster for Normalized Data with red line representing mean.

4.4 Conclusion

Our analysis showed that while there is no single “best” play style but some play styles have stronger correlation with win percentage than others. From figure 4.2, it can be seen that clusters 3 and 5 have on average better winning percentages compared to other team play styles. Therefore, it might be better for teams to try to emulate team play style similar to the cluster 7 and 1 by either playing fast ball movement oriented team with clean defense or shooting from behind the arc with clean defense.. However, teams should be careful on how to change their team play style as certain movement from each play style has differing effects. For example moving from cluster 6 to 3 tends to have a significant increase in win percentage but moving

from cluster 7 to 5 has a negative change in win percentage. However our cluster change analysis is limited by small sample sizes. However, the clustering showed that there are few clusters that are correlated with higher win percentage. Interestingly, the sensitivity analysis indicates that play style has evolved throughout the decades which seems to back up what many of the data analysts has been saying for the past few years.

5. TEAM PERFORMANCE: IMPACT OF COACHING AND CONTINUITY

In the NBA, cohesion is linked with success. Cohesion refers to how well a team works together as a single entity. As discussed in the literature review above, while the literature on team cohesion and team dynamic is vast, to the best of our knowledge there are no previous studies on team cohesion in the NBA. Typically, studies quantify cohesion through surveys given to participants. In contrast, in this chapter, we measure cohesion using roster continuity as a proxy.

Teams may either choose to maintain most or all of their contributing players and compete with what they have, or conversely, they may choose to offload their team's core for different players and/or future assets. This offloading could occur in the form of trading players and not signing players to new contracts so that they could go into free agency. More often than not, teams' off-seasons reflect something in between these two extremes, as teams will often retain their core players and seek to build a winning roster around that core.

Coaches are the off the court leaders of the team. In basketball, coaches are the driving force behind the plays that are made on the court and which players play at different points of the game. In addition, coaches work with the franchise organization as a whole to help build the team. As a result coaches determine or influence both in-game and longitudinal factors. Because

coaches are off the floor, they have a more nuanced impact on games, typically not measured by statistics. In this chapter, we combine factors from roster continuity, and coaching in order to estimate the impact on team W/L%, while controlling for team quality.

We hypothesize that both coaching and cohesion are important factors for success and therefore, at the beginning of a season we aim to predict the upcoming seasons W/L% using previous performance of the team, statistics measuring prior coaching characteristics and success, cohesion measured by how much the players have played together prior to the season with our novel metrics, and with variables denoting how much of the assets the team has retained. We will investigate the impact of these features on two dependent variables, W/L% and the regular season conference standing (the ranking of the coming seasons).

Finally in Chapter 5.4, in addition to studying the value of coaching and cohesion on future performance, we adapt our variable definitions to a explanatory model to explore the role of these factors in current season performance.

5.1 Data

In this study we used team-level performance statistics, coaching statistics, and roster data. In addition, we also used our transaction data to create features related to continuity. We examined the seasons from the 1992-1993 season to the 2017-2018 season, as the coaching data was first recorded in the 1992-1993 season.

To measure roster continuity, we aggregated recent transactions. Specifically we created the following variables: the sum of the number of minutes played by players that were retained, number of transactions in the previous season, number of transactions in the previous three seasons, number of players retained since last season, sum of VORP of retained players

(motivated by the fact that retaining players in the core is more important than retaining bench players), and change in team VORP calculated as VORP retained + VORP gained - VORP lost. In addition we created a variable we named *togetherness score* based on the time each pair of players have played together on a team. Specifically for each pair, we count the number of seasons those players have played together (at any point), and sum these pairwise counts for a given team. We define a *weighted togetherness* score, that weighs each pairwise count by the minutes played by the pair. If m_i is the minutes played by player *i* and m_j are the minutes played by player *j*, then the weighted togetherness score for a team is defined as $\frac{1}{10^6}\sum_i \sum_{j \neq i} \frac{(m_i + m_i)^2}{10^6}$. The intuition behind weighted togetherness score was the recognition that the cohesion of players that play more minutes is more important than those of players that sit on the bench. We also employ a "normalized" togetherness score that is computed in the same way, with a different weighting. If If m_i is the minutes played by player *i* and m_j are the minutes played by player *j*, we create a new variable by weighting togetherness score by a factor intended to normalize the togetherness score: $\sum_i \sum_{j \neq i} \sqrt{m_i^2 + m_j^2}$." The purpose of this weighting was to penalize pairs that included players that did not see a great deal of playing time.

For team-level statistics, we decided to use performance-based metrics including Offensive Rating, Defensive Rating, Pace, and Previous Year's Win Loss Percentage. These metrics allowed us to control for the team's ability.

After considering different coaching related metrics, we found that a coach's career win loss percentage helped predict future win loss percentage, although this may be very closely correlated with a team's historical performance. In addition, the number of games a coach has

coached is correlated with current win loss percentage. The complete list of variables considered in this study are summarized in Table 5.1.

Table 5.1: Variable Summary

Variable	Meaning
prevwl	The win loss percentage of a team in the previous season.
isEastern	Indicator Variable, 1if the team is in the Eastern Conference, 0 Otherwise
Performance Variables:	
Ly PS/G	Last Year's Points Scored Per Game
Ly Rel Pace	The team's last year's relative pace
Ly Rel DRtg	The team's last year's relative defensive rating
Ly Rel ORtg	The team's last year's offensive rating
Ly Top WS Amount	The Number of Win Shares of the player with the most win shares last year on a given team
Coaching Variables	
Seasons Overall	Number of Seasons a coach has coached in the NBA
CoachingW/L%	Coaching Win Loss Percentage over their career.
Continuity Variables	
Minutes Retained	Sum of the minutes played of retained players on a team from the previous season
VORP Retained	Sum of the VORP of players retained
toget_score	Togetherness score (explained above)
norm_toget	Normalized togetherness (explained above)
weighted_toget	Weighted togetherness (explained above)

Transactions	Transactions (Free Agent Signings + Trades) in the previous season
Y3Transactions	Transactions (Free Agent Signings + Trades) in the past three seasons
Starters Retained	Number of Players Retained That Started More than 40 Games
VORP_AR_DELT	Total VORP of current roster minus last year's roster

5.2 Methodology

We first built two different models to highlight the importance of team cohesion and coaching on performance in the upcoming season. We first built a model to estimate the impact of continuity and coaching on a future season W/L% (CCNext). In other words, we sought to understand whether current performance, cohesion and coaching is predictive of next year's performance. Second we studied the impact of continuity and coaching in a season on a team's resulting regular season conference ranking, that we will refer to as seed (CCSeed). In addition we studied an in-season model, please refer to Section 5.4 for setup and results.

During the model development we used a number of different machine learning techniques including regularized linear regression (with LASSO) and a decision tree regressor, but we found that none of them provided a great improvement over ordinary-least-squares regression. Therefore, below we specify the regression models that we ran, their results and conduct outliers analysis.

5.3 Results and Discussion

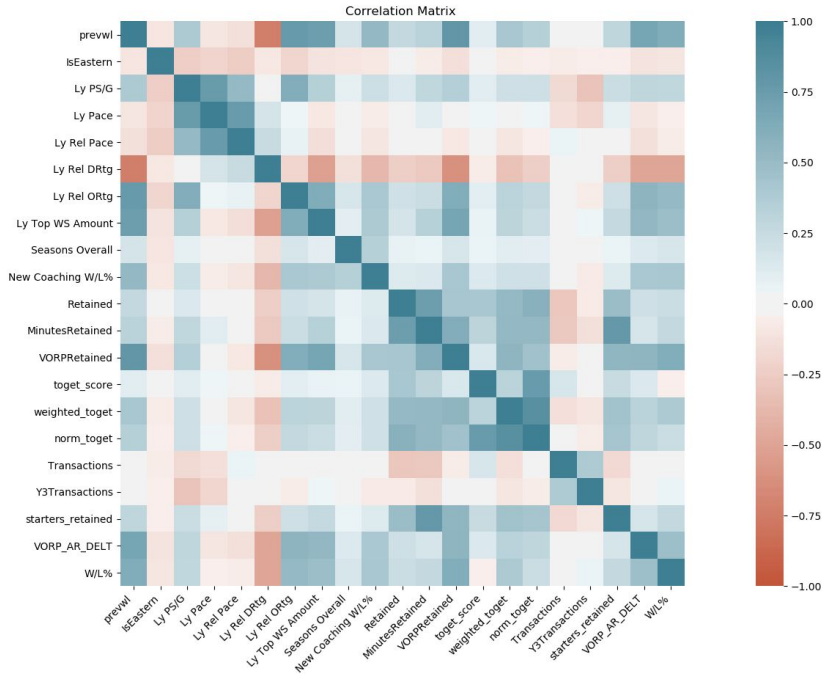


Figure 5.1: Variable Correlation Matrix

5.3.1 Variable Analysis. Figure 5.1 depicts a correlation matrix of the variables considered in the study. The Figure highlights that many of the variables that correspond to a team's performance from the previous year are correlated with each other. Looking at the last row, the win loss percentage, most of the ability controls have a relatively strong correlation with win loss percentage, with the exception of Pace and Relative Pace. Points Allowed and Defensive Rating are better if they are lower, which is why they have a negative correlation with win loss percentage. The remainder of the variables have a small positive correlation with the Win Loss percentage, with the exception of our togetherness score metric, and isEastern. Recently, the eastern conference teams in the NBA have been considered to be weaker, and this data supports this observation. Another row of interest is the one corresponding to "prevwl", or the win loss percentage in the year before. Expectedly, it has stronger correlations with the "Ly",

or last year variables, and exhibits about the same correlations with the other variables, but stronger.

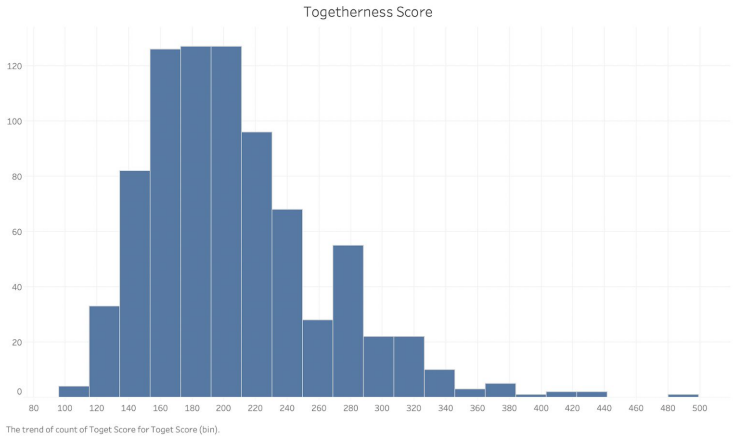


Figure 5.2a: Histogram of Togetherness Score

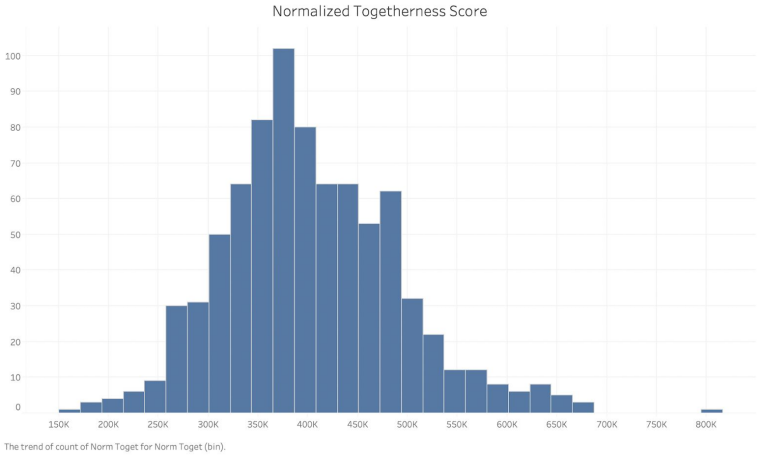


Figure 5.2b: Histogram of Normalized Togetherness Score

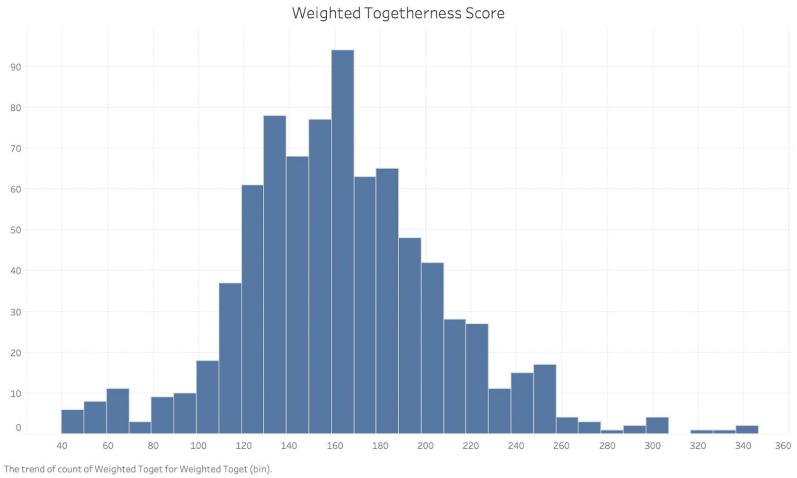


Figure 5.2c: Histogram of Weighted Togetherness Score

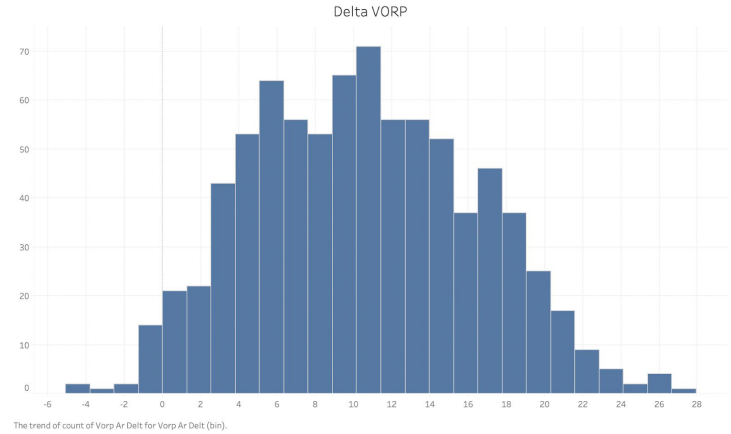


Figure 5.2d: Histogram of Delta VORP

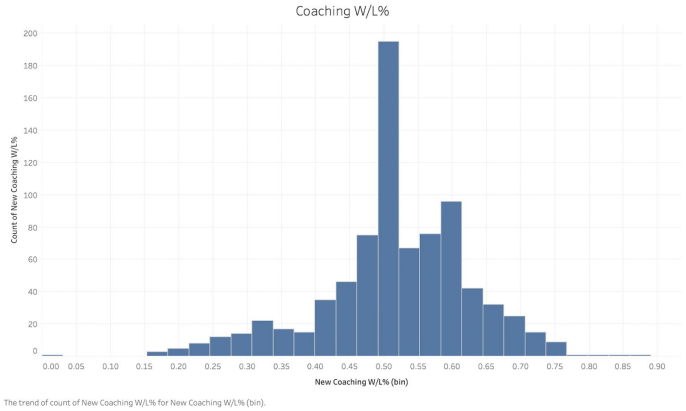


Figure 5.2e: Histogram of Coaching W/L%

Figure 5.2 shows histograms of selected variables of interest.The togetherness metrics exhibit a slight skew in the original and weighted forms, while for the coaching metric, the coaches' W/L%, there is a very large amount of values at 0.5, in part because we assign 0.5 to any coaches without historical data.

5.3.2 Regression Results. Tables 5.2a and 5.2b summarize the regression results of CCNext and CCSeed. Table 5.2a contains the information for the full regression models. As expected from the correlation analysis, many of the coefficients are not statistically significant and therefore we also provide Table 5.3 for models obtained by choosing only statistically significant variables, by first looking at the full model, and then removing the variables with p-values that lay above the significance level threshold. Note that the performance variables and the win loss percentage are strongly correlated, so for CCNext and CCSeed, we chose to preserve previous win loss percentage as a variable, and taking out the other performance variables, regardless of statistical significance in the full model. In all instances W/L%, was statistically significant in the truncated models. In order to construct the second set of predictive models, we used backwards elimination of features until we got a feature set of all statistically significant variables.

*Table 5.2a: Variable Coefficients (with R^2 , * denotes significance at α=0.05)*

Variable	**CCNext (R^2 =0.51)**	**CCSeed (R^2 =0.46)**
const	0.1292	21.4234
prevwl	-0.0174	-2.5015
IsEastern	-0.0038	-0.4701
Performance Vars:		

Ly PS/G	0.0021	-0.0752
Ly Rel Pace	-0.0005	0.0154
Ly Rel DRtg	0.0119	-0.1756
Ly Rel ORtg	-0.0127	0.2895
Ly Top WS Amount	-0.0035	0.1091
Coaching Vars:		
Seasons Overall	$-2.088 * 10^{-5}$	0.0031
Coaching W/L%	0.2062* (0.000)	-6.0495
Continuity Vars:		
Minutes Retained	$-2.573 * 10^{-6}$	$-5.576 * 10^{-5}$
VORP Retained	0.0054* (0.000)	-0.1099
toget_score	0.0001	-0.0015
norm_toget	0.0016	$1.932 * 10^{-5}$
weighted_toget	$-7.44 * 10^{-7}$ * (0.001)	-0.0432
Transactions	-0.0003	-0.0244
Y3Transactions	0.0023* (0.0015)	-0.0391
Starters Retained	0.0098	-0.3095
VORP_AR_DELT	0.0013	-0.0292

Table 5.2b: Models obtained with statistically significant variables (p-value). Variables that are not retained in any of the models are excluded from the table.

Variable	**CCNext (R^2 =0.46)**	**CCSeed (R^2 =0.41)**
const	0.0641 (p=0.0032)	18.6798 (p<0.001)
prevwl	0.3541 (p<0.001)	-7.061 (p<0.001)

IsEastern	-	-0.4632 (p=0.045)
Performance Vars:		
Coaching Vars:		
Coaching W/L%	0.1780 (p<0.001)	-7.117 (p<0.001)
Continuity Vars:		
VORP Retained	0.0069 (p<0.001)	-0.1597 (p<0.001)
weighted_toget	0.0004 (p=0.001)	-0.017 (p<0.001)
Transactions	-	-
Y3Transactions	0.0015 (p=0.003)	-
Starters Retained	-	-
VORP_AR_DELT	-	-

5.3.3 CCNext Results and Discussion. The full model has $R^2 = 0.51$, and the statistically significant variables were: last year's win loss percentage, a coach's previous win loss percentage, VORP Retained, weighted togetherness score, and transactions over the previous three years. In contrast the truncated model has a $R^2 = 0.46$. It is interesting to note that after controlling for the current ability of the team through VORP retained and the current W/L%, then we find that coaching W/L%, togetherness score and past transaction all contribute to future performance, beyond the basic ability of the team.

Table 5.3 lists the largest outliers for this model, those with a magnitude above 0.35. These outliers can be attributed to historical events in NBA history that may not be easily explained with any general model.

Table 5.3: CCNext Outliers

Team	Season	Actual W/L%	Predicted W/L%	Residual
Boston	2007-2008	0.805	0.420	0.384
Chicago	1995-1996	0.878	0.492	0.385
Chicago	1998-1999	0.260	0.635	-0.375
San Antonio	1997-1998	0.683	0.327	0.355

The two Chicago rows, corresponds to Michael Jordan's rejoining to the Chicago Bulls for the 1995-1996 season and departing before the 1998-1999 season. The 2008 Boston row corresponds to when Kevin Garnett and Ray Allen first joined the Boston Celtics to form an outstanding team, and the 1998 San Antonio Spurs corresponds to the Spurs coming off of a year where one of their star players, David Robinson, was injured, to having David Robinson and another star player, Tim Duncan.

When we considered a baseline result derived from just using the previous win loss percentage to get the next year's win loss percentage, that model achieves an $R^2 = 0.408$, which means our full model increases R^2 by more than 0.1, and our truncated model improves it by 0.05. We therefore note that continuity and coaching significantly improve our ability of predicting future win loss percentages.

5.3.4 CCSeed Results and Discussion. The model only has $R^2 = 0.46$. A point to note about this model is that it attempts to predict discrete values, but instead makes its predictions over a continuous domain. For this model, the set of statistically significant variables is: Last Year's Relative Offensive Rating, Coach's W/L%, VORP Retained, Weighted Togetherness

Score, and Is Eastern Conference. Below is a table of selected outliers. We considered data points with residuals over 8 for the purpose of this table.

Table 5.4: CCSeed Outliers

Team	Season	Seed	Predicted Seed	Residual
Brooklyn (NJN)	2001-2002	1	11.8	-10.8
Boston	2007-2008	1	11.3	-10.3
Charlotte	2016-2017	11	2.7	-8.3
Chicago	1998-1999	15	5.4	9.6
Miami	2004-2005	1	9.3	-8.3
Orlando	2003-2004	15	6.8	8.2
Phoenix	2004-2005	1	10.5	-9.5
San Antonio	1997-1998	4	13	-9

The outliers for the CCSeed model are similar to those of the CCNext model, with notable additions, reflecting the fact that predicting the seed is a more difficult prediction problem, as seed is dependent also on how other teams do, and the fact that often closely performing teams will have different seeds. In fact, if the conference is underperforming, a team with a lower win percentage may be able to land a top seed. Some of the outliers here were explained above, but there are others, such as the Phoenix Suns row, which can be explained because they had an extremely strong roster at the time, named the "7 Second or Less Suns," because of their offense that ran plays in 7 seconds or less. . We compared this model to a baseline predictive model that only uses previous year's seed as a predictor, that model achieves $R^2 = 0.358$, so this model significantly improves upon that R^2 by 0.08.

5.4 Explanatory Model

Next we estimate the impact of continuity and coaching in a current season (CCCurrent). In this model we estimate teams' current year's win loss percentage, using cohesion variables, coaching variables and team's change in VORP, in order to investigate the correlation between success and roster moves, as opposed to trying to predict success. We prepared these variables by taking the variables from our predictive models and making some adjustments. We removed the performance variable because they were too strongly correlated with performance. Also, the coaching variables were adjusted so that the characteristics for the coach for the current season instead of the upcoming season were taken into account.

5.4.1 CCCurrent Results and Discussion. The resulting $R^2 = 0.738$, with significant variables being VORP Retained, Transactions, Past 3 Years Transactions, Starters Retained, and the net change in VORP of the team. The coefficients of the variables in a model with all of the variables, along with the variables in a truncated model with only hand-pick statistically significant variables is presented below in table 5.5.

Table 5.5 CCCurrent Results, the third column is statistically significant variables

Variable	**redCCCurrent** (R^2 **=0.76)**
const	0.024
IsEastern	0.0069
Coaching Vars:	
Seasons Overall	$-5.612 * 10^{-6}$
Coaching W/L%	0.029
Continuity Vars:	
Minutes Retained	$-6.115 * 10^{-6}$

VORP Retained	0.0210 (0.000)
toget_score	-0.0002
norm_toget	$2.693 * 10^{-7}$
weighted_toget	0.0004
Transactions	0.0012
Y3Transactions	-0.0012
Starters Retained	-0.0067
VORP_AR_DELT	0.0074

That is after accounting for the ability of the team (through VORP retained and VORP change), we see transactions impacting the expected performance. Interestingly the different continuity variables were not retained by the models. The high explanatory power is unsurprising as the variables are intrinsically related to performance, such as starters retained and net change in VORP. It is worth noting that in the statistically significant variables for the CCCurrent models, transactions in the current year have a positive coefficient, while transactions over the past 3 years and starters retained have a negative coefficient. One possible explanation for this is that teams that are making changes to their starting lineup in the short term are more likely to be teams with a higher win percentage, as having more transactions and less starters retained give a higher win percentage according to the model, but teams that have been continually making transactions may expect on average a lower performance.

The outliers associated with the CCCurrent are listed below. Because of the relative good fit, we list outliers greater than 0.2.

Table 5.6: CCCurrent Outliers

Team	Season	W/L%	Fit Value	Residual
Charlotte	2011-2012	0.106	0.355	-.249
New Orleans	1998-1999	0.622	0.416	0.205
Chicago	1996-1997	0.878	0.648	0.229
Chicago	1997-1998	0.841	0.602	0.238
Chicago	2000-2001	0.183	0.407	-0.224
Chicago	2011-2012	0.758	0.510	0.248

Interestingly, the majority of these outliers are related to the Chicago Bulls. In the 95-96 and 96-97 seasons, the high positive residuals point to Michael Jordan's historical dominance during that period. In the 00-01 season, the Chicago Bulls went through three coaches in the same season in the Post-Jordan Era, so that may have affected these results. In the 2011-2012 season, Derrick Rose put up a career performance that led the Chicago Bulls to the first in the East with a 50-16 record. In the same lockout 2011-2012 season, the Charlotte Hornets performed well under even a pessimistic expectation by going 7-59, which holds the record for the worst in NBA history. These two outliers may point to the fact that extraordinary circumstances, such as a lockout may cause some abnormalities, regardless if the data is aggregated to account for them.

5.5 Conclusion

We have presented novel variables with the goal of capturing team's cohesion. We demonstrated that cohesion measures (Transactions, Y3 Transactions, Delta VORP) and one of our coaching variables (Coaching W/L%) improve prediction of a team's next year's performance. We also looked at a correlation with the current year using continuity variables to

see how the variables affect how a team is doing currently. These improvements are significant compared to strong baselines built on previous performance. However, as with any set of models, they do not account for special cases such as great teams that have one unlucky year as the model only goes back one year. A direction for future improvement is a longitudinal model that builds predictions on more than one year of data. For example, to include the past three years' data as inputs to the model. For coaching variables, it may be worth looking into additional ways to quantify a coach's impact that is more independent of team player quality.

6. DRAFT ANALYSIS

Teams that want to be successful - particularly small market teams with limited financial capabilities - must take advantage of undervalued assets that are overlooked by other teams. Thaley and Masser, in a 2005 study of NFL teams concluded that picks closer to the front of the order draft picks are consistently overvalued by NFL teams, and that by 'trading back' in the draft a team could receive more value than what they give away. Trading back refers to giving up an early draft pick for multiple later picks, contracted players, or a combination of the two. We hypothesize that NBA teams often undervalue draft picks that occur later in the draft. Below we conducted statistical analysis of the NBA draft to determine if an earlier pick in the draft is better, and, if it is, exactly how much better.

6.1 Methodology

The study was based on NBA draft data from 1980 through 2014. The data consisted of the picked players and the order in which they were picked in each draft and their career VORP. As stated in earlier chapters, VORP is designed to measure overall performance, and was used as a measure of a player's value in order to quantitatively compare players. We also considered the

VORP over the first three years of a player's career, to account for the fact that players that were drafted in the 2000s may still be playing. In addition, we conduct the analysis using the first three years of a player's career to reflect the fact that a player may choose to leave their team after their rookie contract is completed. One consideration we had to account for was that some of the players drafted never played in the NBA, and therefore have no value for VORP with respect to the NBA, as they did not produce for their team. These players were given a -2, which is the minimum possible VORP, and is indicative of a player who could be replaced by a player from the NBA's G-League (development league).

6.1.1 Pick VORP Analysis. The data was organized by pick, and the average and standard deviation of players selected at each pick was calculated. We aggregated both the career VORP of players as well as their VORP in the first three years. Figure 1 shows the correlation between the career VORP and the three year VORP, colored by draft pick, with picks closer to the front of the order being lighter, while later picks being darker.

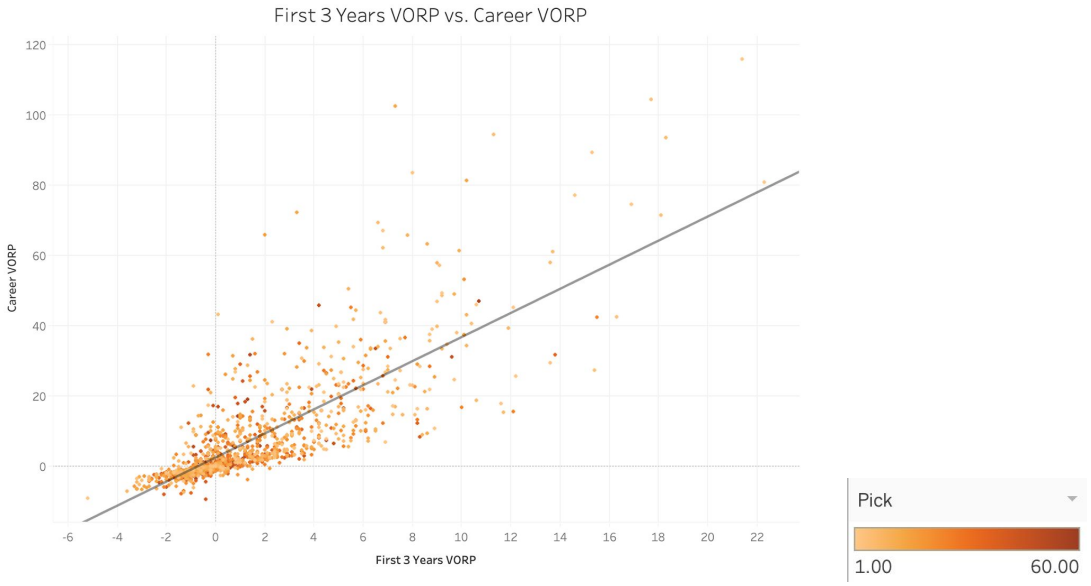


Figure 6.1: Correlation Between First 3 Years VORP and Career VORP

The correlation between the VORP for a player's first three years and their career VORP is statistically significant with a p-value of less than 0.0001 and an $R^2 = 0.67$. However as Figure 6.1 highlights it is not a guarantee that players that are good early will be good later, and vice-versa. Also, there are picks later in the order that can be seen in the graph to have low VORP, although as higher VORP is reached in the figure, the points tend to become lighter in color, signifying a pick earlier in the order.

We chose to conduct a two-sided Wald-type t-test on each adjacent pair of picks (for example, the 1st and 2nd, 2nd and 3rd, etc) for both career VORP and the 3 year VORP. The goal of these tests were to indicate whether there is a statistically significant difference between the expected values among consecutive picks. After conducting the t-tests on the adjacent pairs of picks, we decided to do pairwise t-tests between all of the picks, not just the adjacent ones, We express these in a visualization below.

6.1.2 Draft Pick Trading Analysis. After performing our statistical analyses, we sought to find real examples of how teams are over or undervaluing draft picks. In order to do this, we assigned expected values to the respective picks; the VORP assigned to each pick is the mean career VORP of the pick. Then, we analyzed a year of transactions to see if the picks that were being traded were being correctly valued by teams in terms of the resources that they were trading for them. We considered the trade to be a 'correct' decision for a given team if it results in the team gaining a net positive VORP in the trade. For the purposes of our analysis, we only considered the VORP of the players and draft picks.

6.2 Pick VORP Analysis Results

Figures 6.2a and 6.2b summarize the distribution of career and first 3 year VORP respectively. The Figures highlight that the expected VORP of the first pick is significantly higher than any other pick. After the first pick, however, the relative value of the picks becomes unclear, although it is clearly decreasing both when the career VORP and the 3 year VORP is analyzed. Interestingly the variance in the value of the pick is in both cases decreasing the pick order. In addition, this visualization suggests that the third pick is slightly more valuable that the second pick, which is illogical based on the wisdom that the best players are selected first. This quirk is partly due to the fact that Michael Jordan, who is widely considered by many to be one of the best players in NBA history, was picked 3rd.

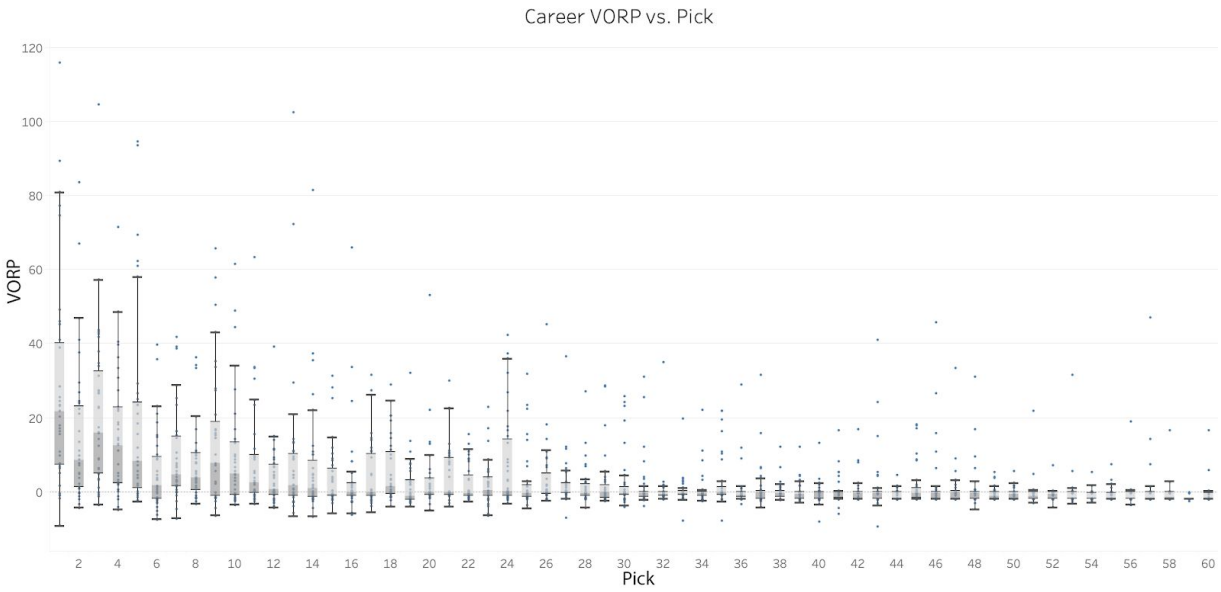


Figure 6.2a: Career VORP Plotted along each pick

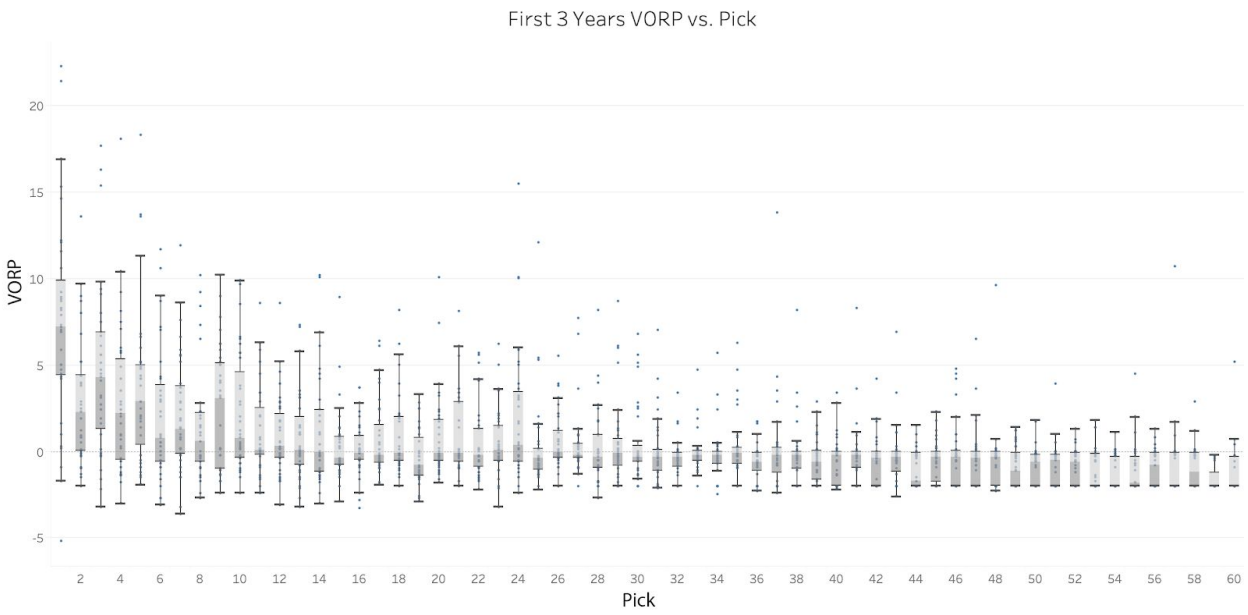


Figure 6.2b: 3 Year VORP plotted along each pick

To test for statistical differences in the average values of different picks, a t-test was performed on each pair of consecutive picks. The results for career VORP are visualized in Figure 6.3a. Each point in the figure represents the lifetime VORP of a player. The color changes for from one pick to the next when the adjacent t-test result was significant at the 0.05 significance level. The results of the same analysis for the first 3 year VORP are displayed in Figure 6.3b. The resulting groups are summarized in Tables 6.1a and 6.1b.

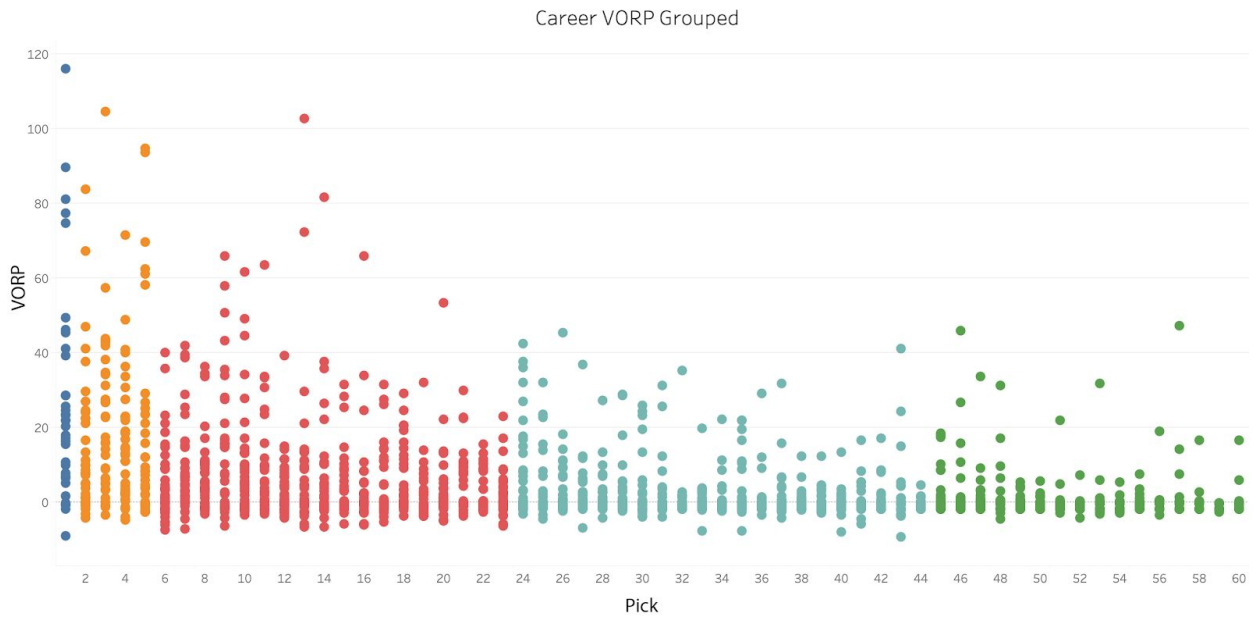


Figure 6.3a: *Results of Pairwise Wald t-tests Groups for Career VORP*

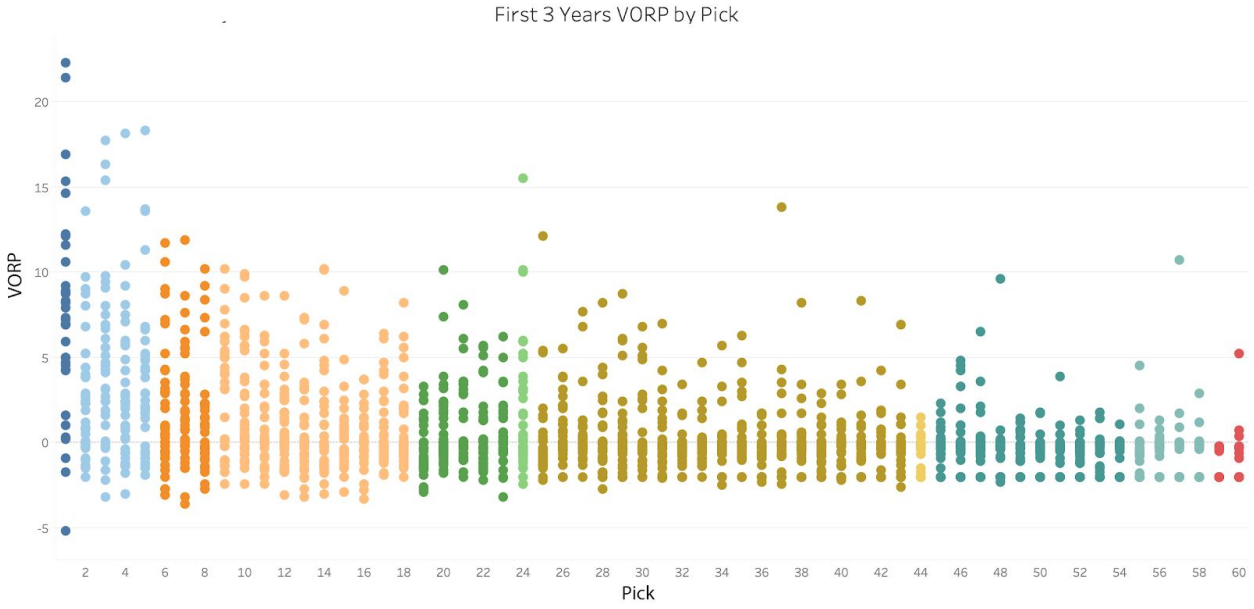


Figure 6.3b: *Results of Pairwise Wald t-tests Groups for 3 Year VORP*

Table 6.1a : Average VORP by adjacency groups(Career VORP Groups)

Group	Mean Career VORP
1	27.9
2-5	18.0
6-22	5.9
23-44	1.5
45-60	0.0

Table 6.1b: Average VORP by adjacency groups (3 Year VORP Groups)

Group	Mean 3 Year VORP
1	7.5
2-5	3.5
6-8	1.8
9-18	1.1
19-23	0.5
24	2.0
25-43	0.3
44	-0.1
45-54	0.1
55-58	0.3
59-60	0.1

The t-tests analyses of the career VORP suggest 5 ‘groups’ of picks with similar value. The first pick is in its own group, followed by pick groups: 2-5, 6-23, 23-44, 45-60 (reported in

Table 6.1a). The results highlight that in general, picks closer to the front of the order are more valuable than picks closer to the back. Between the groups, the mean VORP is strictly decreasing as we approach the end of the draft order. Table 6.1b summarizes the adjacency groups for the three year VORP. Here, the average VORP does not strictly decrease as we approach the end of the draft order. This is partly due to the existence of some outliers. One particular outlier of interest is the 24th pick, which gets its own bucket because the average is raised by Andrei Kirlenko with 15.5 VORP in his first 3 years, and Tim Hardaway with 10.1. Kirlenko's and Hardway's career VORP were 42.4 and and 37.4 respectively, so they had long careers of value ahead of them at the time.

For the career VORP, our methodology suggests that the 6th pick is essentially as good as the 22nd pick. This conclusion is an artifact of the approach; the pairwise tests only test consecutive picks, and therefore while there may not be statistical significant differences between picks 20th and 21st, and again between the 21st and the 22nd pick (etc) there may be statistical differences between the 6th pick and the 22nd pick. As a result of this limitation, we ran a full set of statistical significance tests below, between groups of players at any two picks. The results are displayed in Figure 6.4a for the career VORP and for the 3 year VORP in Figure 6.4b.

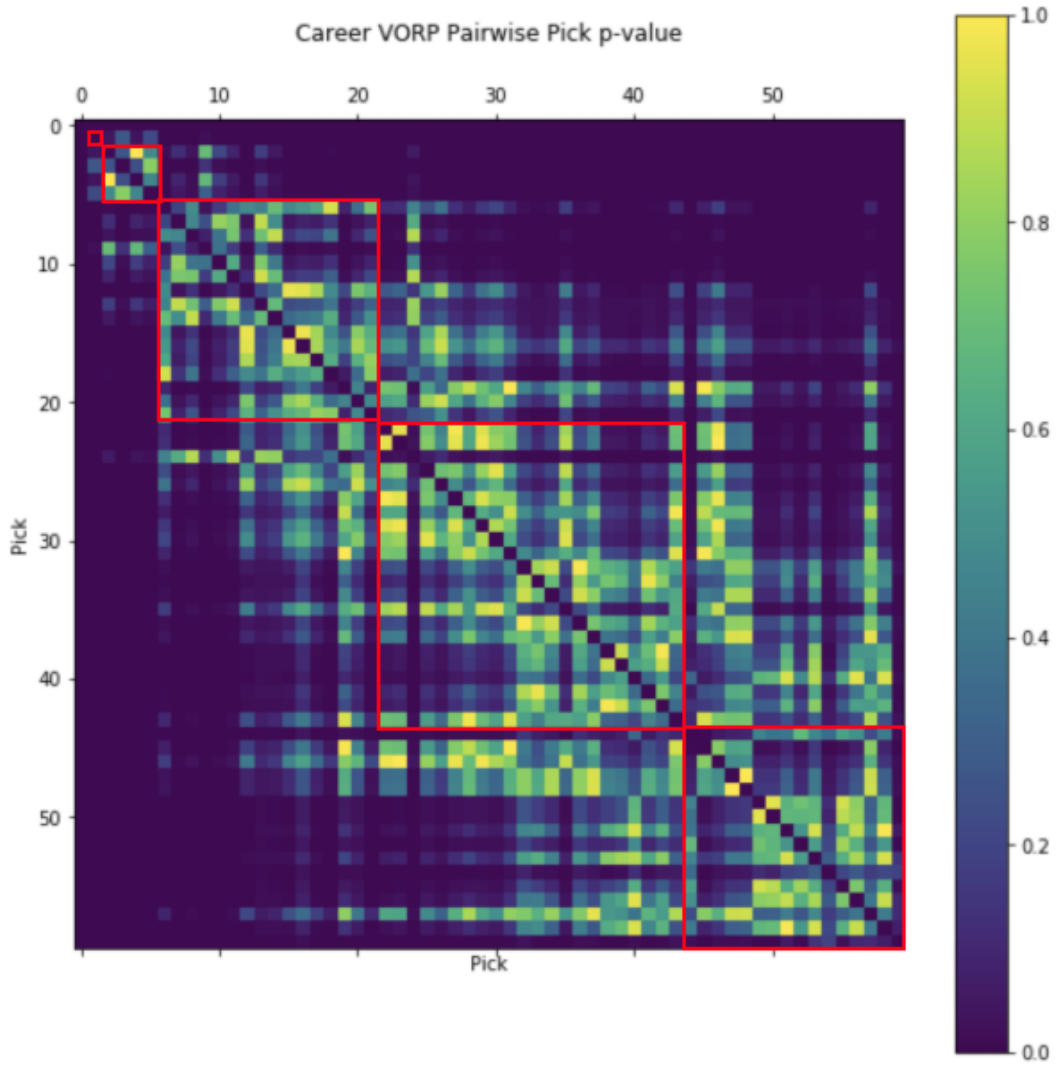


Figure 6.5a: Career VORP Groups visualized on t-test matrix, where each square summarizes the results of a t-test for the differences in means between two picks.

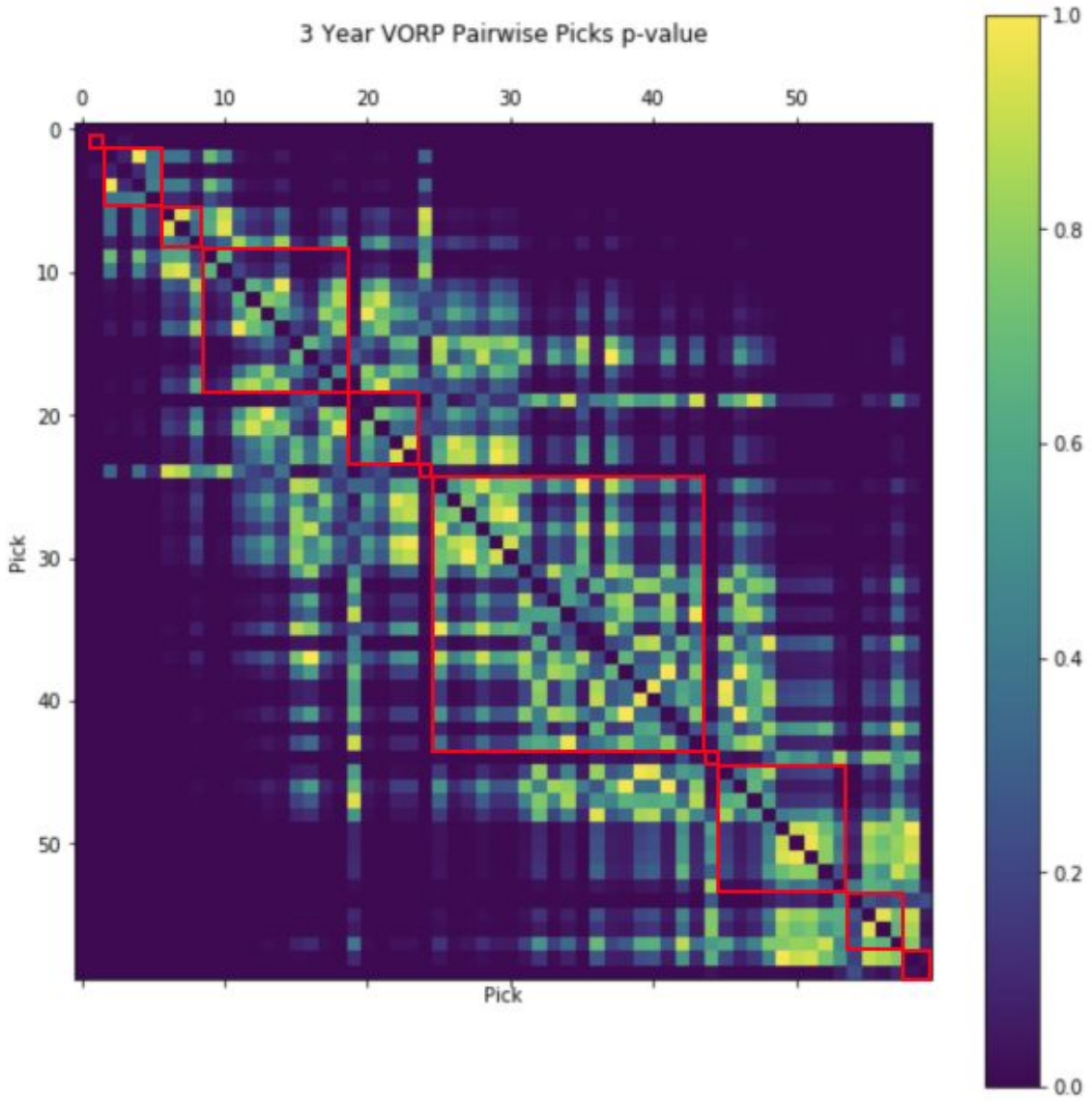


Figure 6.5b: 3 Year VORP Groups visualized on t-test matrix, where each square summarizes the results of a t-test for the differences in means between two picks.

In Figures 4a and 4b, we have superimposed the groups that were derived in the previous step with red squares on the graph of the full pairwise t-tests between picks. In the graphs, the cell corresponding to *i,j* denotes the result of the t-test of the differences in mean VORP of pick *i*

and pick *j*. On the diagonal, in cells *i,i* for any given *i,* we did not conduct a t-test, so we assigned the value 0. In the above figures, approximate groups can be seen in clusters of cells closer to 1.0 in p-value. The Figures highlight the higher variability in the analysis of the first 3 year VORP compared to career VORP; the groups are harder to identify in the 3 year VORP figure. The way that we can identify statistically similar clusters of picks visually is by looking at this matrix, and finding regions of a lighter color, because these regions signify that the picks within them are statistically similar, we can see that our previous method, with the adjacent pairwise t-tests was able to get partially capture clusters of picks with similar value, but it fails in some cases. For example, the 6th pick is grouped with the 22nd pick, but according to the matrix, the 6th pick and the 22nd pick are not similar. Also, these groups are susceptible to error if one of the picks is an outlier, such as at pick 24, where the distribution at pick 24 is higher than the one at pick 23, when looking at Figure 6.2a. The effect of having an outlier that is higher than its predecessors is that the groups would split at these outliers. In this case, the 23rd pick is statistically similar to the 26th pick, but not according to the groups.

Figure 6.6 highlights an alternate way of extracting these groups of picks, for which each groups consists of picks where none are statistically significantly different from one another (with a few notable outlier exceptions, like the 24th pick and the 44th pick). The way that the new red squares were created were by drawing regions around the visually "lighter" regions, because they indicate possible groups of picks that would still not be statistically significantly different from one another.

.

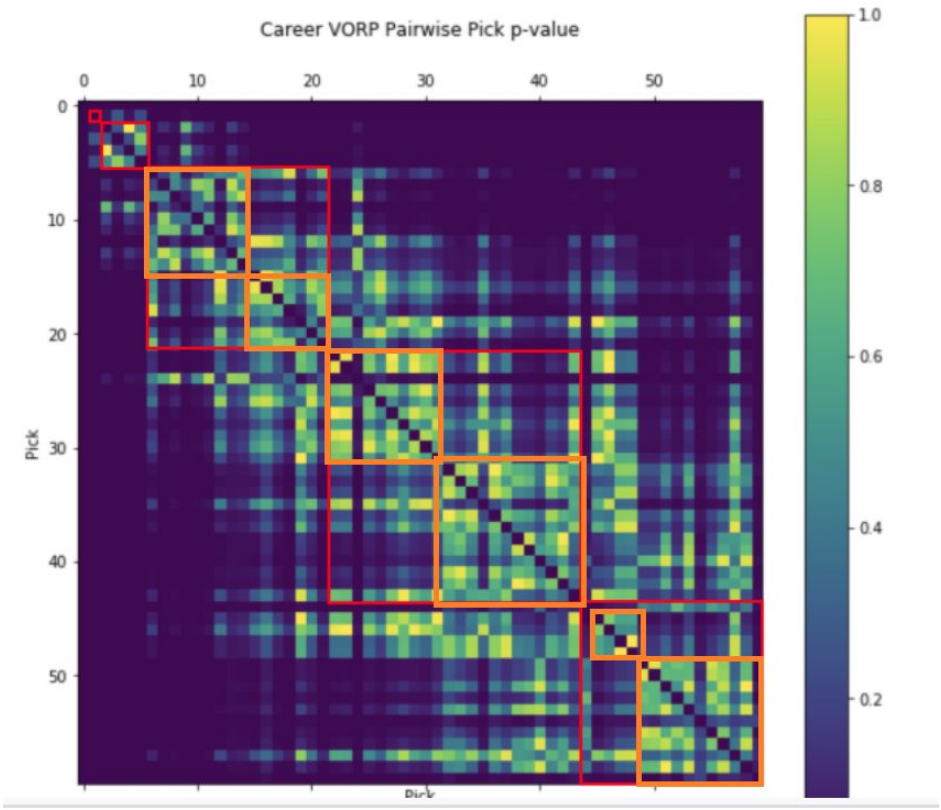


Figure 6.6: An alternative way to find groups, from drawing squares some light colored regions

A way to make this useful in practice would be, for each pick give a range of picks that are similar, and that way a team can see if it is worth it to "trade up," that is get a pick higher in the order. Figure 6.7 presents an alternate way of visualizing the pairwise interactions, coloring the cells yellow if there is a statistically significant difference. In row *i,* column *j,* the cell is colored yellow if the p value is less than 0.05, signifying those picks are significantly different. For all *i,* row *i,* column *i* is just colored purple.

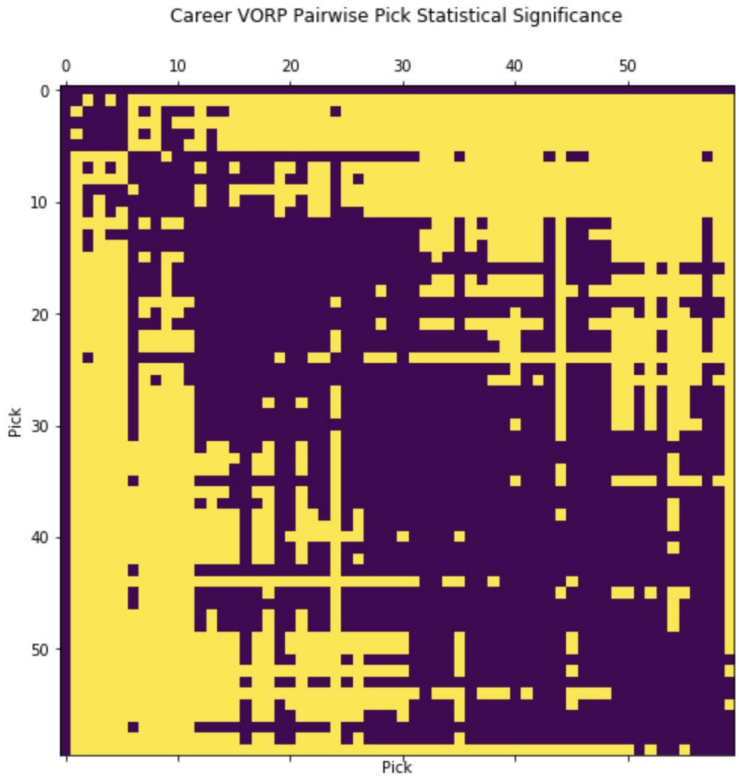


Figure 6.7: An alternate visualization of the pick pairs.

6.2.1 Analysis using Medians. One of the issues with the use of means as a metric was that means are not resistant to outliers unlike medians. In order to address this, we repeated the same analyses mentioned in the previous section using medians. To test for the difference in population means, we swapped out the Wald t-test for Mood's Median Test. At the 0.05 significance level, we found that there were no instances where adjacent pairs had a test statistic with a p-value less than 0.05. This is more reflective of the conventional wisdom that picks at around the same point in the draft are similar to each other, rather than some picks being especially bad or good. This means that all of the adjacent pairs had similar medians under the statistical test, further motivating the need to compare non-adjacent pairs. The visualization of the p-values of the full pairwise comparisons picks, like in Figure 6.7, but instead done with the

median test, can be found below. The purple regions are clearer, as there are less yellow cells, indicating more pairs of picks with statistically similar medians.

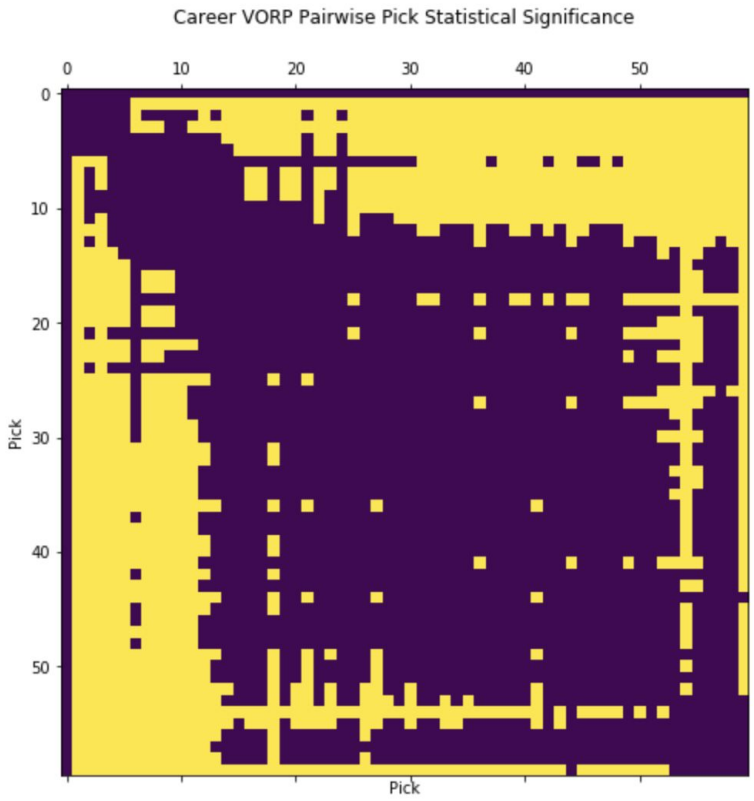


Figure 6.8: The same figure as 6.7 done with Mood's Median Test instead of Wald t-test

The regions in the figure are similar to those in Figure 6.7, but it seems to create three clear tiers of picks, which are overlapping: the first group is around 1-5, the next group is around 4-15, and then 10-60. This suggests that the picks that are very early in the order are strong, the subsequent 10 picks are similar but weaker, and the picks after those are more or less the same in terms of value.

6.3 Conclusions

In this study, we established that the expected values of consecutive (adjacent) draft picks are often not statistically different, nor are the expected values strictly decreasing. As a result,

there are some clusters of picks that can be treated as similar, or at the very least, there may exist an interval around every given pick of other picks of similar values. This is an important result for NBA teams looking to take advantage of over/under valued assets because throughout the NBA, and other sports leagues with drafts, it is believed that having an earlier draft pick is always better. This analysis shows that there may not be as much value added as previously thought, and teams that are willing to take advantage of that could benefit. NBA teams often hesitate to trade down because of the belief that players taken with the first few picks are much better, especially when the top few prospects were extraordinary college players and fans want to see big, young, names. The reality is that they may be able to trade down a few picks and acquire assets, coming out with a net positive expected value, but if they have the first pick, they should not trade down.

With all of our results, there are other factors to consider when discussing the draft. First, some draft classes are stronger than others, like the 1984 NBA draft, which included players like Michael Jordan and Hakeem Olajuwon. Also, if a team is ill-equipped to develop their young players, even with the first pick in the draft, the team will not derive as much value from drafting, because the player will not grow. Finally, pundits and front offices may have a general idea of which players will be drafted when in the draft, there may be situations where trading down is not optimal because the other teams would want to draft a desired player.

7. TRANSACTION ANALYSIS

A key to understanding team success is in analyzing how teams recruit players. One of the most important methods for acquiring talent is through trading with other teams, whether for

draft picks or players. Building on our previous analyses of the draft and of how players are overvalued or undervalued in free agency, we use transaction data to determine the comparative value placed on players and draft picks.

Using the available information from the Pro Sports Transactions website, we built a comprehensive database of team transactions, taking place from 1980 to 2018, as described in Chapter 3. This came to include just over 1500 trades between two teams. We combined this dataset with statistics on player performance, in particular VORP, and data on team standings and salaries. More background on the composition of trades can be found in Chapter 1.3.3.

7.1 Expected Draft Pick VORP

Our initial goal in this analysis was to use this trade data to study the value teams are placing on draft picks. This would allow us to confirm whether or not teams trade according to an empirical understanding of draft pick value. Our approach started by assuming each trade was balanced for both teams, where the VORP acquired/relinquished by each team is equal. This allowed us to view these trades as equations: the equation has known values, the upcoming year VORP of current players, and we would solve for the unknown values, the the VORP equivalence of future draft picks. To best isolate the value of draft picks, we examined the approximately 900 trades where a single draft pick was exchanged. This approach is formalized below:

$$VORPEQ(Draft\ pick) = \sum VORP(Players\ Acquired) - \sum VORP(Players\ Relinquished)$$

From this equation framework we conducted several analyses on the resulting estimated draft pick values. We first examined the relative value of first round and second round picks, as teams specify within trades the round of the exchanged picks. After separating out the two

rounds and running a t-test, we found that first round picks were indeed valued more highly (p = .001). The respective averages for first and second round pick values were 0.40 VORP and 0.15 VORP, and are distributed as shown in the box plot below.

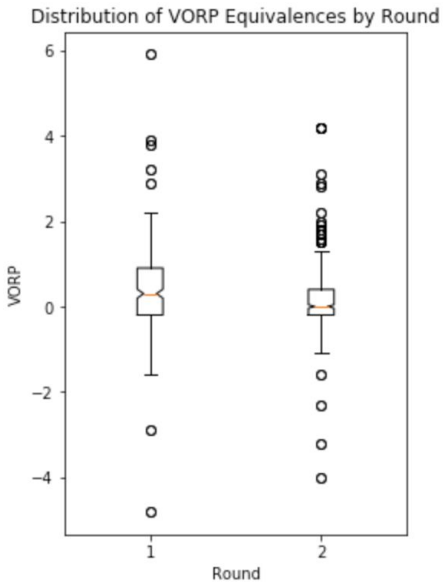


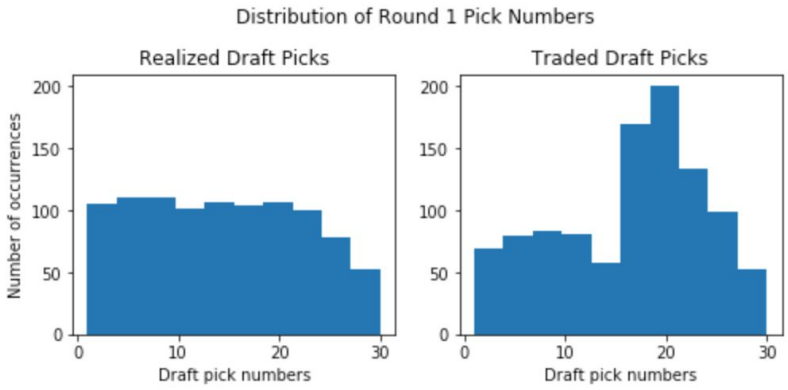
Figure 7.1: The distribution of VORP equivalence of traded draft picks in each round

We then compared these VORP equivalence values to the realized VORPs of all past draft picks, including those not used in trades as well, using their average VORP per year over their whole career. We additionally compared this to the average over all years of performance of first round picks, without any intermediate averaging. One would hypothesize that the trade VORP equivalence value should be equivalent. As Table 7.1 shows that is not the case.

Table 7.1: Draft Pick VORPs, comparing trade expectations to realized values

	Trade VORP equivalence	Realized Average VORP (By Player)	Realized Average VORP (By Season)
Round 1	0.36	0.56	0.83
Round 2	0.13	0.05	0.27

Treating the VORP equivalence as actual VORP, we ran a t-test for first round draft pick values. Comparing the VORP equivalence expectations to the realized VORPs of all draft picks, averaging by player, we found there was a significant difference between the two groups ($p<0.05$, $n_{tradeVORP}=270$, $n_{realizedVORP}=956$), even while the latter was higher. When comparing VORP equivalence to realized values, averaging by season, we also found there indeed was a significant difference ($p<0.01$, $n_{tradeVORP}=270$, $n_{realizedVORP}=8736$). This analysis indicates that, under the assumptions that teams value the expected VORP of a draft pick at the same level as the VORP of players being traded, teams undervalued draft picks in trades. This undervaluing may have several explanations. This includes the fact that the distribution of traded draft pick numbers is different from the distribution of realized draft pick numbers. As shown below, many more low value first round draft picks are traded (15-25) than higher values, especially due to "protections" placed on some traded draft picks.



Figures 7.2a (left) and 7.2b (right): The frequency of round 1 draft picks at each pick number realized (7.2a) vs. traded (7.2b). Drop off at tail in 7.2a is due to the expansion of the number of teams, and thus expansion of the draft.

Another explanation is that teams are risk averse, and prefer to value draft picks at below their expectation. This may especially be due to the lack of exact knowledge of which number pick the team will receive, as well as the variability in the observed value of any given pick number (as detailed Chapter 6). One final explanation is that the value of a traded draft pick may decrease as the number of years until realization increases, as future gains are inherently less valuable than present gains. We further consider these hypotheses as we continue to the next stage of analysis.

7.2 Team Standing and Effects on Draft Pick Number

Because traded draft picks only specify their round number, this often leaves vast uncertainty as to which exact pick number a team will end up receiving. Referring back briefly to the workings of the draft system, while the typically the top 3 picks are distributed by lottery (this number recently has been raised to 4), the remaining picks are deterministically assigned based on team standing from the previous season. And as seen from previous analysis, there is a significant difference between getting a 5^{th} pick and a 25^{th} pick. Given that trades generally occur at least one year out, it is important to understand the stability of a team's standing over time. The following series of plots show just how quickly a team's standing can change in the span of a few years.

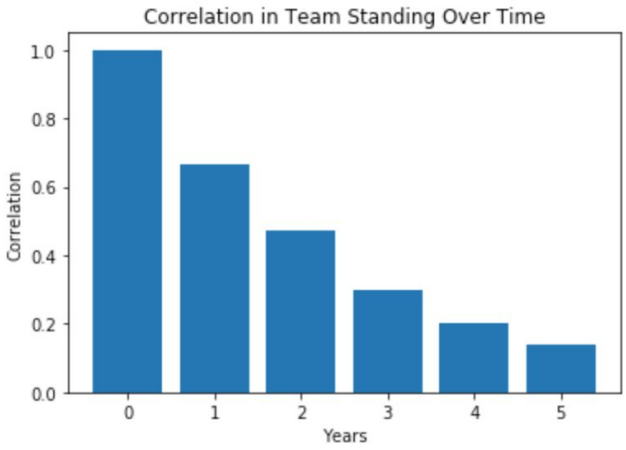


Figure 7.3: Correlation in team standing after varying years later.

This first plot shows that the correlation between a team's original standing and future standing degrades rapidly. For instance, after 2 years, this correlation falls to near one half, and in 3 years it is closer to one third. Furthermore, inspired by related analyses of social mobility, we also examined team mobility in terms of change in quartile over time.

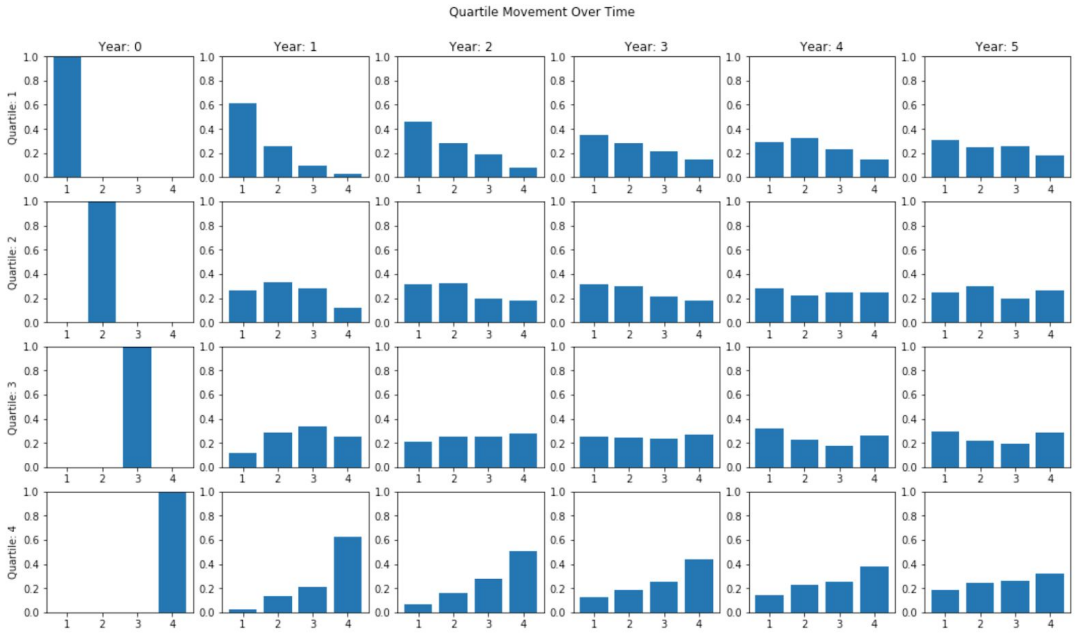


Figure 7.4: Probabilities of remaining in original quartile after varying years later. (Quartile 4 is the highest performing, while Quartile 1 is the lowest performing)

The plot above also illustrates the speed with which teams move between quartiles for their standing. Teams in the middle quartiles have practically even likelihoods for ending up in any of the four quartiles after just 2 years. The outer quartiles are close to even likelihoods by 5 years, with still greatly increased mobility by years 3 and 4. Teams trading draft picks more than one or two years out, especially with middling teams, may have little to no guidance on which number pick they will ultimately receive, within the designated round. This is especially relevant considering how many picks are traded years before their realization, as shown in Figure 7.4. Further adding complication, teams also oftentimes place "protection" conditions on their traded draft picks, preventing the trade from occurring if the draft pick number realizes within a certain range (i.e top 5).

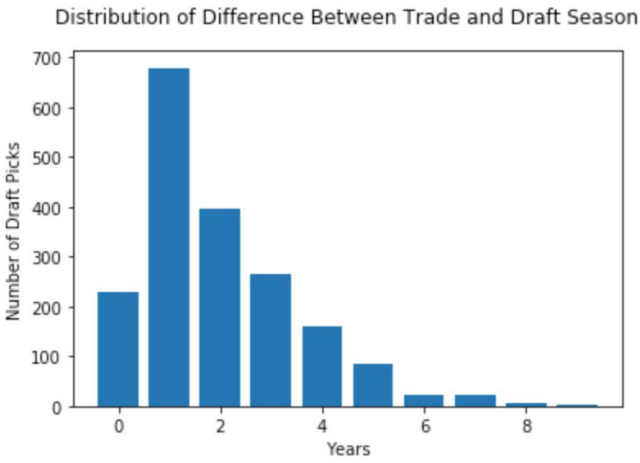


Figure 7.5: Number of draft picks traded at each difference between year of trade and year of draft pick realization. Zero years refers to trades taking place in the postseason months directly prior to the draft, at which point standing has already been determined.

To summarize the difficulty of predicting the eventual draft pick number, we then estimated a regression model, predicting the draft pick number of first round picks based on the

team standing of the team relinquishing it, and whether the traded draft pick included protections. The model and results are provided below.

$$pickNumber = \beta_0 + \beta_1 teamStanding + \beta_2 protected$$

*Table 7.2: Results of regression estimating draft pick number. Team standing and pick number were both normalized to 0-1. Significance at p<0.05 level indicated by *.*

	Model Results (R^2 = .132)
constant	0.7682*
teamStanding	-0.2768*
protected	0.0718*

Interpreting the above results, it is promising that *teamStanding* and *protected* both have significant impacts on the outcome. However, we would expect a coefficient of -1.0 (as opposed to -0.28) on *teamStanding*, had the picks been set deterministically with full information. This difference, together with the low R^2, is indicative of just how unpredictable the pick number remains. This is further emphasized by running a comparable regression on draft picks that hadn't been traded, which comprise approximately 75% of the 991 recorded first round draft picks in our data. Out of approximately 1000 first round draft picks 25% have been traded. To briefly summarize, that model had an R^2 = 0.767, and coefficient of -0.8754 on *teamStanding*, also statistically significant and much closer to the expected -1.0 coefficient. The two regression models show that the uncertainty added by the gap between trade year and draft year drastically decreases the predictability of draft pick numbers.

We next continued to a more complex analysis of trades, and in particular draft picks, accounting for team standing and years until realization in addition to salary cap considerations.

7.3 Regression Analysis of Transactions

The final component of this analysis is multilinear regression building off the formulaic approach initially taken. Based on the intuition from previous analyses, we decided to construct a model that would take as independent variables various team, player, and draft pick characteristics as related to those components of the trade. As a dependent variable, we predicted the differential between the VORP of players acquired by the designated "team 1", and the VORP of the players that team relinquishes, quite similar to the formula approach (PlayerVORP) above. For the independent variables, we took differentials on some of the variables, and split apart others into acquired and relinquished variables.

In our first regression model, we analyzed just trades with a single, first round draft pick, of which there are 96. To account for the effect of team standing and the time difference between trade and draft, we created a standing/year multiplier to weigh each draft pick accordingly. The standing was normalized from 0 to 1, where 0 indicates a high performing team and 1 indicates a low performing team. This weight was then calculated as follows:

$$Weight = (normalized\ standing) \times 0.9^{(years\ in\ the\ future)}$$

This weight then gives the most value to a draft pick that will occur soon and whose original team was a low performer, as we would expect. Additionally, the exponential discounting of value over time matches with general economics thought on valuation of future gains and losses. We further decided to test whether team standing would affect the balance of a trade, adding in a difference term that subtracts the second team's standing from the first (higher standing indicates worse performance). Also included are two binary terms for overall team salary, where the term for each team would be equal to 1 if the salary exceeded 1.25 times the

salary cap, and 0 otherwise. We expected to observe whether teams greatly exceeding the salary cap are more willing to offload expensive, higher VORP players in order to decrease their financial burden. Lastly, we account for whether or not cash was involved in the trade: 1 for acquired, -1 for relinquished, and 0 for not involved. Thus, we regressed the following model:

$$playerVORP = \beta_1 draftWeight + \beta_2 standingDiff + \beta_3 salaryAcq + \beta_4 salaryRel + \beta_5 cash$$

For the second model that we estimated, we analyzed trades that also included 2nd round draft picks and any number of draft picks, totalling to 1789 trades. We introduced additional independent variables: two variables (*draft1* and *draft2*) account for the number of draft picks from each round on either side of the trade, once again as a differential between the two sides. We additionally find a corresponding weight for 2nd round draft picks. This model is below.

$$playerVORP = \beta_1 draft1 + \beta_2 draft2 + \beta_3 draftWeight1 + \beta_4 draftWeight2 +$$

$$\beta_5 standingDiff + \beta_6 salaryAcq + \beta_7 salaryRel + \beta_8 cash$$

The results from estimating these two regression models are in the following table.

Table 7.3: Regression results for both models. (indicates significance at p<0.1 level)*

Variable	**Model 1** (R^2 = .239)	**Model 2** (R^2 = .136)
draft1	-	-0.7975*
draft2	-	-0.1411*
draftWeight1	-1.0923*	0.4639*
draftWeight2	-	-0.0215
standingDiff	-0.0242*	-0.0189*
salaryAcq	0.3946	-0.0378
salaryRel	-0.0528	0.0393
cash	-0.3934	0.1374*

As a general remark when it comes to interpreting these results, negative coefficients generally indicate variables that relate to higher draft value. When the dependent variable, player VORP acquired differential, decreases, that indicates the team is willing to give up more player VORP in order to acquire this draft pick. Along those lines, we found two significant variables from our first model, *draftWeight1* and *standingDiff*. The coefficient for *draftWeight1* is negative, as we would expect, as a higher weight indicates a stronger draft pick candidate (lower team standing, and fewer years). The coefficient for *standingDiff* is also negative, indicating that teams with higher standing, or worse teams, generally get less *playerVORP* in their trades, especially when trading with much better teams. This unbalance may be explained by the fact that worse teams likely became that way due to poor front office decisions. As a result, a worse team is more likely to continue to make poor decisions, such as accepting an unbalanced trade.

Examining model 2's significant variables, we also get intuitively sound results when examining the coefficients for *draft1* and *draft2*. As they are both negative, this correctly implies that a team has to give away more VORP to gain additional draft picks. Additionally, we see that the coefficient for *draft1* is much larger than that for *draft2*, confirming that first round picks are worth more in terms of player VORP. The coefficient for *cash* is also significant and negative, indicating that providing cash does indeed serve to "sweeten" a trade. The other two significant variables in model 2 are the same ones as in model 1. However, while *standingDiff*'s coefficient is approximately equivalent in both models, *draftWeight1* takes on a positive coefficient in model 2, opposite of model 1. This may be due to a difference in calculation, as the weight

variable accounts for several draft picks on both sides of the trade, but it is otherwise unclear what this result indicates.

7.4 Conclusions

This analysis of transactions has yielded several compelling results, especially as relates to draft picks. We first determined that valuations of draft picks as determined by player VORP in trades are sufficiently similar to the actual resulting value of drafted players. We then further expanded our analysis to include variables that may cause differences in valuations of draft picks, including team standing and years until draft pick realization. However, the low R^2 from each model also indicates there is still much to be explored when it comes to analyzing transactions. For instance, one might come to the conclusion that some trades will naturally be uneven, where one team will gain more from a trade than another team. The potential existence of such trades certainly warrants further exploration.

8. PLAYER MARKET VALUE IN FREE AGENCY

Among the four major American professional sports leagues, free agency plays a particularly strong role in the NBA. It is rare in other sports to see the very best and most impactful players hit the open market, but this is a fairly common occurrence in the NBA, with all-stars, surefire hall-of-famers, and even MVPs such as LeBron James and Kevin Durant changing teams in free agency in recent memory. The number of free agents in any given season can vary, but there are generally 100-165 free agents in a given year[4]. While it is obvious that players of that calibre should receive the highest possible salaries, for many players in free agency, it is not so clear. As a result, many players get overpaid or underpaid relative to their

[4] In 2011 there were 66 free agents, and in 2013 there were 204, but the rest of the years in this study fall into that range.

performance. We study the relative efficiency of the free agency market in terms of player's pay relative to their performance. To achieve this, we construct linear regression models of both *players' salaries* and their performance in terms of *Value Over Replacement Player (VORP)*.

8.1 Methodology

8.1.1 Data Selection and Cleaning We gathered salary data from Spotrac's NBA contracts dataset and used the full range of data available that could be integrated with our database, corresponding to salary information from 2011 through 2018. From this dataset we extracted information about players' salaries in their first year of a new contract, a total of 625 entries. This includes free agents, restricted free agents, and cases where a player or team option was picked up on a contract that would otherwise expire. We integrated this with our existing database for analysis. Further cleaning these data included collapsing instances where players played for multiple teams in a season to single entries in the dataframe, correcting rounding errors, and manually calculating some statistics such as effective field goal percentage. The salary values were adjusted to reflect percentages of the salary cap that season. Keeping salaries in terms of dollars would be misleading for two reasons. The first is inflation: the value of 1 dollar in 2011 is different from the value in 2018. The second is the value of a salary relative to the salary cap has changed. For example, a 10 million dollar salary in 2011 is a greater commitment for a team than the same salary in 2018, since teams have more money available to spend.

Roughly 200 data points were lost in merging the contracts and advanced statistics dataset. This is because the advanced statistics dataset was incomplete. The lost data points tend towards less impactful players both in terms of the salary they command and their VORP. We

were ultimately left with 625 entries in the dataset for predicting salaries and 409 entries in the dataset for predicting VORP. The discrepancy is due to a lack of calculated advanced statistics for players with few minutes played.

8.1.2 Salary and VORP Modeling We originally considered 22 variables in constructing the predictive models. This included career averages of box-score statistics, the previous season's stats, the player's age, and the player's previous season salary (or VORP). Preliminary testing showed that the previous season statistics alone were stronger predictors than the career average statistics. This may be explained in part by younger players whose career averages are not reflective of the player's abilities since they are likely improving each season. For older players, the previous season statistics are often similar to their career averages, so there is not much to gain by including them.

We chose to partition the data into three age groups, one for players under 25 years old, one for players 25 years old to 32 years old, and one for players over 32 years old. Each of these age groups represents a different stage in a player's career. Players under 25 who sign new contracts are usually signing their first contracts after their rookie deals. This is crucial as rookie deals are defined in the CBA, so the contracts we study in this age group are often the first market-defined contracts the players sign. These players are usually still developing their skills, and have not reached their peak performance. The middle age group represents players both in their primes on the court and in earning potential, as they are more or less known entities and still have many productive years of basketball left. Finally, the oldest age group represents players past their primes on the court and often in earning potential.

For our predictive models, we implemented a backwards elimination for variable selection. For the under 25 salary model, only the previous season's points and rebounds were retained, and this yielded R^2 of 0.665. For the between 25 and 32 salary model, the previous season's points, rebounds, assists, steals, and salary were retained. This yielded R^2 of 0.672. For the over 32 salary model, only the previous season's points and salary were retained. This yielded R^2 of 0.421. For the VORP models, the previous season's steals, blocks and VORP were retained for the under 25 model, the previous season's rebounds, steals, and VORP were retained for the between 25 and 32 model, and only previous season's points, rebounds and VORP were retained. The models initially yielded R^2 of 0.729, 0.757, and 0.496, respectively.

We opted to maintain the same variables across all age groups in the salary models and to do the same in the VORP models, since there was significant overlap in the retained variables. This simplifies our modeling results, and as detailed below only marginally affected our R^2 values. This left us with three variables for each model. The previous season's points per game, rebounds per game, and salary for the salary prediction. The previous season's rebounds per game, steals per game, and VORP were used for the VORP prediction.

We implement the linear regression model using the python package scikit learn. Variable selection was performed on a random 80-20 train-test split. The final model coefficients are reported on the entire data. We measure model performance with R^2; the models' R^2 are reported as the average over a 5-fold cross-validation sample.

We chose our baselines for the salary and VORP models to be the previous season's salary and the previous season's VORP, respectively. Both of these were the strongest single

predictors of the dependent variables. Additionally, one would expect in most cases a player's current salary and performance to be indicative of their future salary and future performance.

8.2 Results

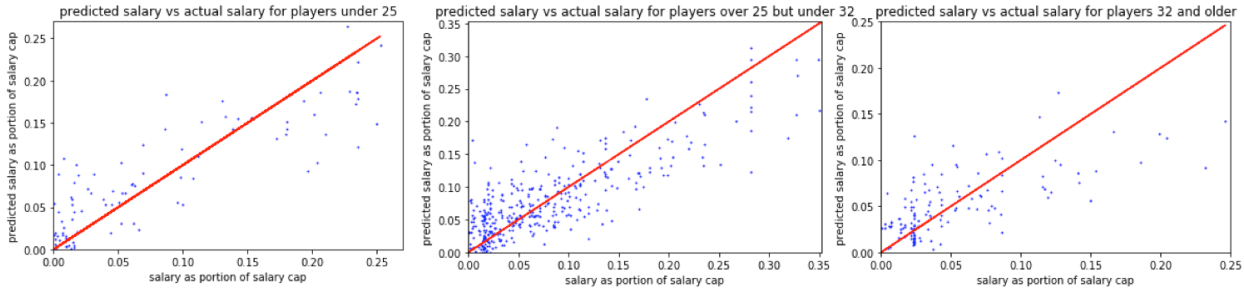
8.2.1 Salary Model Salaries in the NBA range wildly. The lowest salaries in the league in a given season take up only a fraction of a percent of the salary cap. In 2018, Antonius Cleveland, who played 6 games for the Hawks, made $225,792. The salary cap in 2018 was $99.093 million. In contrast, Russell Westbrook made $28,530,608 as one of the highest paid players in the league. The two salaries accounted for 0.23% of the cap, and 28.79% of the cap, respectively.

Each of the linear regression models showed significant improvement over the best baseline predictor, which was the previous season's salary. Using only the previous season's salary as a predictor, linear regression models yielded R^2 scores of 0.353, 0.391, and 0.304 for the under 25, 25-32, and over 32 age groups, respectively. The final models yielded k-fold cross-validated (n=5) R^2 scores of 0.721, 0.669, and 0.427, respectively. The oldest age group was the most difficult to predict. The baseline for all ages in a single dataset scored R^2= 0.276, demonstrating that the age split fostered better models.

*Table 8.1: Coefficients of Linear Regression Salary Models. Shows the contributions of each parameter to the predictions of salary l. For example, in the over 32 age group, for each point per game the player had in the previous season, on average that player's salary in the current season is 0.0058 * the salary cap greater, everything else held constant.*

	Age<25 n=106	**25≤Age≤32** n=385	**Age>32** n=134
Last Season PTS/G	0.0110 (*p<0.001*)	0.0073 (*p<0.001*)	0.0058 (*p<0.001*)

Last Season TRB/G	0.0051 (*p=0.011*)	0.0129 (*p<0.001*)	0.0016 (*p=0.389*)
Last Season Salary (as portion of the salary cap)	0.0498 (*p=0.478*)	0.1715 (*p<0.001*)	0.1216 (*p=0.038*)



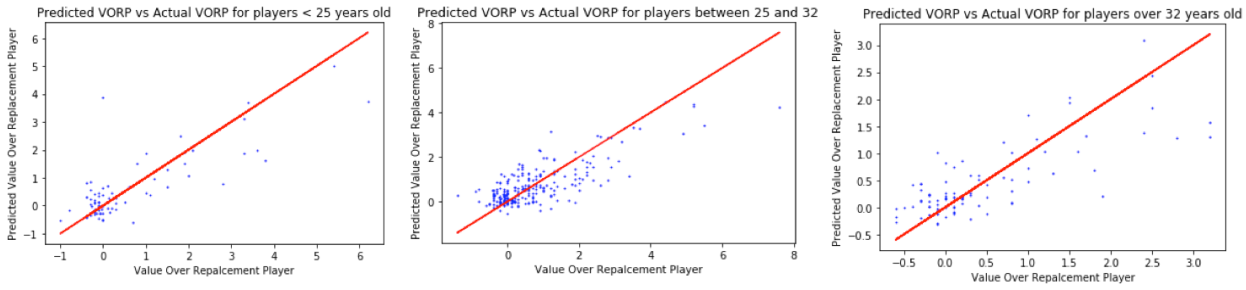
Figures 8.1a (left) 8.1b (center) 8.1c (right): Plots of predicted salary (y-axis) vs. actual salary (x-axis) for each age group, as labeled. The red line represents a slope of 1- points above the line are instances where the predicted salary is greater than the actual, and points below the line are instances where the predicted salary is less than the actual.

8.2.2 VORP Model VORP is defined such that the worst players in the league in a given year will have a negative VORP. In 2018, 19-year old Frank Ntilikina of the New York Knicks had a VORP of -1.3. The league leader that year, LeBron James, had a VORP of 8.2. The highest recorded VORP in a single season belongs to Michael Jordan, who in 1988 had a VORP of 12.47. A player has had VORP over 10.0 in a season only 10 times in recorded history. Similarly to the salary models, the VORP models performed better than the baseline of solely the previous season's VORP. These baseline models scored $R^2 = 0.529$ for players under 25, $R^2 = 0.484$ for players between 25 and 32, and $R^2 = 0.478$ for players older than 32. For these same age groups, the final models yielded R^2 scores of 0.757, 0.759, and 0.531, respectively. Again,

the oldest age group was the most difficult to predict. The baseline for all age groups scored R^2= 0.449.

Table 8.2: Coefficients of Linear Regression VORP Models. Shows the contributions of each parameter to the predictions of VORP. For example, in the under 25 age group, for each steal per game the player had in the previous season, on average that player's VORP in the current season to be 0.5169 higher, everything else held constant.

	Age<25 n=72	**25≤Age≤32** n=245	**Age>32** n=92
Last Season TRB/G	0.1068 (*p=0.310*)	0.0309 (*p=0.088*)	0.0615 (*p=0.017*)
Last Season STL/G	0.5169 (*p=0.258*)	0.3755 (*p=0.020*)	0.0358 (*p=0.800*)
Last Season VORP	0.5284 (*p<0.001*)	0.6076 (*p<0.001*)	0.6244 (*p<0.001*)



Figures 8.2A(left) 8.2B(center) 8.2C(right): Plots of predicted VORP (y-axis) vs. actual VORP (x-axis) for each age group, as labeled. The red line represents a slope of 1- points above the line are instances where the predicted VORP is greater than the actual, and points below the line are instances where the predicted VORP is less than the actual.

8.3 Discussion of Results

The predictive models show great improvement over a single baseline predictor and demonstrate satisfactory accuracy in the two younger age groups. For players over 32, we were able to produce models that performed better than baselines but not at the same level as the models for the younger group. In both the salary (fig. 8.1c) and the VORP (fig. 8.2c), it can be observed that there are several points near the high-end of the x-axis (the actual values) where the salary and VORP are under predicted. In the case of the salary, this group of players includes great players well beyond their primes such as Kevin Garnett and Steve Nash in 2013. Those players commanded large salaries because of the players they once were and the salaries they once earned. In these cases, the on-court performance dips but earning potential does not. In the case of VORP, under-predictions occur because it is difficult to predict when older players will decline, or at what rates they will decline. Some players in this under-predicted group include Tim Duncan in 2013 and Manu Ginobili in 2014, who were on Spurs teams that played extremely well due to excellent coaching and unexpected longevity of the players. Conversely, we predicted Tim Duncan in 2016 would have VORP of 3.1, but his actual VORP was 2.4. Duncan retired after the 2016 season and had by far his least productive season, signifying a rather rapid decline. Players declining at varied rates, and, in some cases players having late-career renaissances like JJ Reddick, makes predicting both salary and VORP a challenge for older players.

In the other two age groups, we see most of the predictions are fairly close to the actual values, with some exceptions. Injuries and other extreme circumstances explain some of the most egregious outliers such as the case we predicted VORP of about 3.89 versus an actual VORP 0 (fig. 8.2a). This instance is Paul George in 2017. Paul George famously broke his leg training for

the Olympics in 2014, and only played six games in the 2015 season because of the injury. In the salary model, an example is Greg Monroe in 2015, whose market value plummeted after he was arrested for a DWI in 2014. He was a promising young player, and we predicted he would make roughly 18.1% of the salary cap, when in reality he made 8.7% (fig. 8.1b). For under predictions in the salary model, we found that most instances were cases where the players were widely considered overpaid. For example, we predicted Wesley Matthews would make about 13.1% of the salary cap in 2016, when the actual number was about 23.5% (fig. 8.1b). His closest peers in salary were Kevin Love and Draymond Green, both perennial all stars. In contrast his closest peers in VORP in 2016 included Blake Griffin, Manu Ginobili, and Thabo Sefolosha. In that season, Griffin only played 38 games and Ginobili was 38 years old, well past his prime. Matthews' next contract, which he signed in 2019, paid him roughly $2.5 million per season. The average annual value of the contract he signed in 2016 was around $17.5 million. For the VORP model, under predictions often occurred when players took large leaps in production. The model predicted Kawhi Leonard to have a VORP of 3.7 in 2016 after he had a similar VORP in 2017. However, the actual value was 6.2. (fig 8.2a) "Board Man"'s breakout season was 2016, in which he secured his first Finals MVP, first all-star appearance and began playing at an elite level.

Despite the above described difficulties and irregularities present in these prediction problems, the models on the whole perform well and we are able to draw interesting conclusions about market inefficiencies from them, as discussed below.

8.4 Conclusions

Tables 8.1 and 8.2 highlight that points per game was not a strong enough predictor of VORP to be included in the best-performing model, but was strong enough to be included in the best-performing salary model. Mindful of the fact that the magnitude of regression coefficients is not directly related to their impact, we find that the defensive statistics (blocks and steals) have a higher relative impact than is obvious. For example, in the 25-32 salary model, the coefficient for points is 0.0073, meaning that, holding all other stats the same, a player who scored one more point per game would expect a change in salary of 0.73% of the salary cap. The coefficient for rebounds is 0.0129, meaning that we expect a change in salary of about 1.29% of the salary cap per block per game. In 2005, the league leader in points per game was Allen Iverson with 30.7. The league leader in rebounds was Kevin Garnett with 13.5. If both players signed a new contract the next season (both fall into the middle age group), we would expect Iverson's points to earn him 22.4% of the salary cap, and Garnett's blocks to earn him 17.4% of the salary cap. Points mattered more than rebounds in this salary model, and especially mattered more in the model for players over 32, where rebounds have such small coefficients they are essentially negligible. In particular, we see that points scored is the strongest predictor of salary (and the only significant predictor for the <25 age group), but is not a significant predictor of VORP, while rebounds and steals per game, both considered hustle stats, are.

We note in the literature that researchers have found a slight statistical boost in performance in the year before a contract is signed[5]. In the context of this study, this phenomenon in conjunction with the fact that contract year statistics are stronger predictors of

[5] See section 2.2.1 for details

salary reveal the driving principle in NBA contract negotiations. New contract values seem to rely on “what have you done for me lately?”, and players know this.

The plots below show an interesting trend of the relative undervaluing of defense. In this first set of plots (fig. 8.3a, 8.3b, 8.3c), the red highlighted points are for players who scored more than 15 points per game in the previous season.

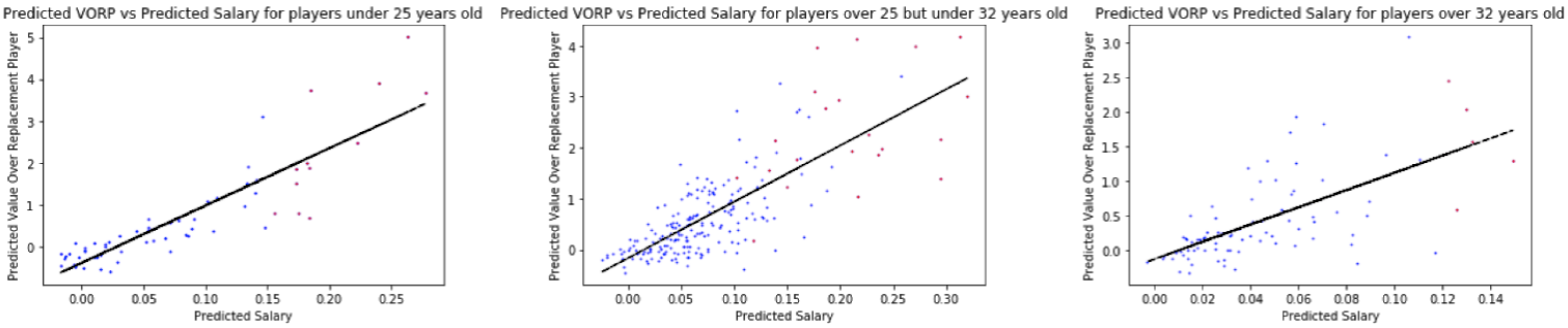


Fig. 8.3a(left),8.3b(center),8.3c(right). Plots show predicted VORP (y-axis) versus predicted salary (x-axis). In each case, the red points represent players who averaged over 15 points per game in the previous season. The black line represents the trend line.

In the next set of plots (fig. 8.4a, 8.4b, 8.4c), the red highlighted points are for players who averaged more than 1.0 steals per game.

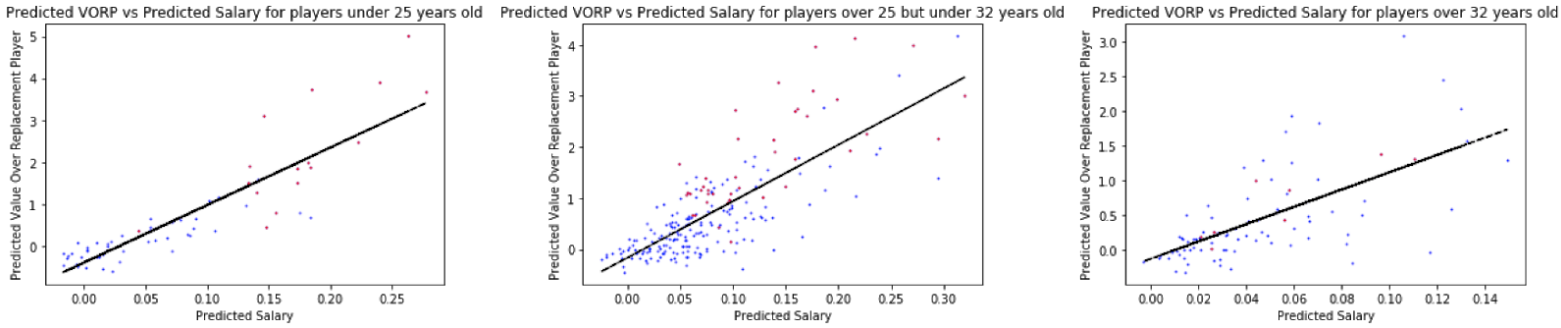


Fig. 8.4a(left),8.4b(center),8.4c(right). Plots show predicted VORP (y-axis) versus predicted salary (x-axis). In each case, the red points represent players who averaged over 1.0 steals per game in the previous season. The black line represents the trend line.

These plots, in conjunction with the coefficients in the models, illustrate the tendency of teams to overvalue high scorers and undervalue players who are effective on defense. Note in the plots with high scorers highlighted (fig. 8.3a, 8.3b, 8.3c) the high concentration of red points on the far right side but below the trendline. This group represents the players who are expected to be paid the most but perform the worst relative to their pay. Conversely, in the plots with high steals (fig. 8.4a,8.4b,8.4c) we see a high concentration of red points above the trendline, representing players who outperform their salaries. In particular, there are many players who are projected to make a high, but not top, salary but are projected to perform at the highest level. This represents an inefficiency that is exploitable.

We have created six models that accurately predict both player's salaries and VORP with good accuracy. From these models, we have demonstrated inefficiency in the market where many players are paid more or less than their expected performance dictates. In particular, we note the trend of teams to overvalue players who are high scorers and generally perform well on offense, and undervalue players who may be not as strong on offense but strong defensively and in rebounding.

9. CONCLUSIONS

In this thesis we present a series of analyses that contribute to the NBA analytic literature. We quantified team play styles, and study the effect that team play styles have on success.

We concluded that while there is no one always superior playing style, there are playing styles that are more indicative of success and we provide preliminary analysis on transitions between play styles and how conducive they are to success. For example, we would encourage teams in our defined cluster 5 to transition to our defined cluster 7, by emphasizing ball movement (to increase assists), spacing (by increasing three-point attempts), and perimeter defense (limiting opponent three point attempts).

We studied the effect that team continuity, in terms of both roster construction and coaching, has on team success. In doing this, we effectively developed novel statistics that enable more accurate prediction of team winning percentage. From this analysis, we would encourage teams to lean towards maintaining the cores of their rosters and retooling as opposed to shipping off assets and entirely rebuilding.

We studied the NBA draft and the value of different draft picks as they translate to on-court value. We presented observations about the similarity in expected value of various groupings of consecutive draft picks. In particular, we suggest teams to re-evaluate how they value draft picks as picks beyond the early first round have similar expected value.

We studied NBA transactions between two teams, particularly analyzing how teams value draft picks, comparing to both actual draft pick value and the value of current players. We observed that team standing and years until draft pick realization play a role in how teams value traded draft picks. However, due to the large variability in draft pick value we found, even after accounting for several related team and trade factors, we suggest teams make a more careful evaluation of draft picks in trades.

We studied the free agency market and the difference in how players are valued in terms of salary and their performance on the court. We successfully modeled both players salaries and their performance, and found exploitable discrepancies in driving factors of salary and performance. In particular, we would suggest teams seek to sign defensively oriented players as they often outperform their market value and high scoring players who are not strong defensively are often overpaid.

Building a winning NBA roster is a complex, multifaceted problem. There are multiple avenues for gathering talent and 30 organizations attempting to utilize those avenues to their gain. As data analytics continue to grow in influence in the sports world, and particularly the NBA, teams are seeking ways to best leverage available data to their advantage. With so many factors into building a team, there is no one analytic tool that can hold all of the answers. The series of analyses we present address gaps in available literature and provide additional pieces to the analytic puzzle of constructing a winning team.

Works Cited

Abbott, H. (2012, April 10). Does tanking even work? TrueHoop. http://www.espn.com/blog/truehoop/post/_/id/40055/does-tanking-even-work

Adams, L. (2019). NBA Teams With Most, Least Roster Continuity. Hoops Rumors. https://www.hoopsrumors.com/2019/10/nba-teams-with-most-least-roster-continuity-3.html

Asghar, F., Asif, M., Nadeem, M. A., Nawaz, M. A., & Idrees, M. (2018). A NOVEL APPROACH TO RANKING NATIONAL BASKETBALL ASSOCIATION PLAYERS. Journal of Global Economics, Management and Business Research, 176–183.

Badenhausen, K. (2017, February 15). The Knicks And Lakers Top The NBA's Most Valuable Teams 2017. Forbes.

Baghal, T. (2012). Are the "Four Factors" Indicators of One Factor? An Application of Structural Equation Modeling Methodology to NBA Data in Prediction of Winning Percentage. Journal of Quantitative Analysis in Sports, 1355.

Barnes, S. L., & Bjarnadóttir, M. V. (2016). Great expectations: An analysis of major league baseball free agent performance. Statistical Analysis & Data Mining, 9(5), 295–309. https://doi.org/10.1002/sam.11311

Berri, D. J., Brook, S. L., & Fenn, A. J. (2011). From college to the pros: Predicting the NBA amateur player draft. Journal of Predictive Analysis, 35(1), 25–33.

Berri, D. J., & Eschker, E. (2005). Performance When It Counts? The Myth of the Prime Time Performer in Professional Basketball. Journal of Economic Issues, 39, 798–807. https://doi.org/10.1080/00213624.2005.11506847

Berry, C. R., & Fowler, A. (2019). How Much Do Coaches Matter? 25.

Birnbaum, P. (2017). A Guide to Sabermetric Research. Society for American Baseball Research. http://sabr.org/sabermetrics/single-page

Borghesi, R. (2009). An Examination of Prediction Market Efficiency: NBA Contracts on Tradesports. The Journal of Prediction Market, 3(2), 65–77.

Brooke, B. (2019, July 2). NBA Roster Continuity vs Winning. https://observablehq.com/@benjaminadk/nba-roster-continuity-vs-winning

Carron, A., Bray, S., & Eys, M. (2002). Team cohesion and team success in sport. Journal of Sports Sciences, 20, 119–126. https://doi.org/10.1080/026404102317200828

Carron, A. V., Colman, M. M., Wheeler, J., & Stevens, D. (2002). Cohesion and performance in sport: A meta analysis. Journal of Sport & Exercise Psychology, 24(2), 168–188.

Chan, T. C. Y., Justin A. Cho, & David C. Novati. (2012). Quantifying the Contribution of NHL Player Types to Team Performance. Interfaces, 42, 131–145. https://doi.org/10.1287/inte.1110.0612

Colquitt, L. L., Godwin, N. H., & Shortridge, R. T. (2007). The Effects of Uncertainty on Market Prices: Evidence from Coaching Changes in the NBA. Journal of Business Finance & Accounting, 34(5–6), 861–871.

Conway, T. (2017, October 17). Explaining the NBA Rule Changes for 2017-18 Season. Bleacher Report.

Deshpande, S. K., & Jensen, S. T. (2016). Estimating an NBA player's impact on his team's chances of winning. Journal of Quantitative Analysis of Sports, 12(2), 51–72.

Farneti, C. (2008). Exploring Leadership Behaviors and Cohesion in NCAA Division III Basketball Programs [The Ohio State University]. https://etd.ohiolink.edu/pg_10?0::NO:10:P10_ACCESSION_NUM:osu1211907565

García, J., Ibáñez, S. J., Martinez De Santos, R., Leite, N., & Sampaio, J. (2013). Identifying Basketball Performance Indicators in Regular Season and Playoff Games. Journal of Human Kinetics, 36(1), 161–168.

Harrison, J. (2019). The Effect of Play Style on NBA Revenues [University of California, Berkeley]. https://pdfs.semanticscholar.org/5fd0/c23f31e1adebe8b1dc60111514fd70c0d1fb.pdf

Heuzé, J.-P., Raimbault, N., & Fontayne, P. (2006). Relationships between cohesion, collective efficacy, and performance in professional basketball teams: Investigating mediating effects. Journal of Sports Sciences, 24, 59–68. https://doi.org/10.1080/02640410500127736

Hofler, R., & Payne, J. (1997). Measuring efficiency in the National Basketball Association. Economics Letter, 55, 293–299.

Hwang, D. (2012, March 2). Forecasting NBA Player Performance using a Weibull-Gamma Statistical Timing Model. MIT Sloan Sports Analytics Conference.

James, B. (1985). The Bill James Historical Baseball Abstract. Villard Books.

Jenkins, B. (2019, May 28). NBA Finals, Warriors vs. Raptors: Matchup analysis and predictions. San Francisco Chronicle.

Kihl, L., & Richardson, T. (2009). “Fixing the Mess”: A Grounded Theory of a Men’s Basketball Coaching Staff’s Suffering as a Result of Academic Corruption. Journal of Sport Management, 23(3), 278–304. https://doi.org/10.1123/jsm.23.3.278

Kuehn, J. (2016, March 11). Accounting for Complementary Skill Sets When Evaluating NBA Players’ Values to a Specific Team. MIT Sloan Sports Analytics Conference.

Lee, Y. H., & Berri, D. (2008). A RE‐EXAMINATION OF PRODUCTION FUNCTIONS AND EFFICIENCY ESTIMATES FOR THE NATIONAL BASKETBALL ASSOCIATION. Scottish Journal of Political Economy, 55.

Lundgren, Z. (2012). The Superstar Effect in the National Basketball Association: An Evaluation of Salary Distribution and Team Performance. The St. Olaf College Economic’s Department Omicron Delta Epsilon Journal of Economic Research, 34–52.

Mandle, J. R. (1998). Team Play and the Compensation System in Professional Basketball. 9, 13.

Marcos, F. M. L., Miguel, P. A. S., Oliva, D. S., & Calvo, T. G. (2010). Interactive Effects of Team Cohesion on Perceived Efficacy in Semi-Professional Sport. Journal of Sports Science & Medicine, 9(2), 320–325.

MarketWatch. (2015). Sports Analytics Market Worth $4.7 Billion by 2021. MarketWatch. http://www.marketwatch.com/story/sports-analytics-market-worth-47-billion-by-2021-2015-06-25-142032056

Martinez, J. A., & Caudill, S. B. (2013). Does Midseason Change of Coach Improve Team Performance? Evidence From the NBA. Journal of Sport Management, 27, 108–113.

Maymin, A., Maymin, P., & Shen, E. (2013). NBA chemistry: Positive and negative synergies in basketball. International Journal of Computer Science in Sport, December.

McIntyre, A., Brooks, J., Guttag, J., & Wiens, J. (2016). Recognizing and Analyzing Ball Screen Defense in the NBA. 10.

Melnick, M. J. (2001). Relationship between team assists and win-loss record in The National Basketball Association. Perceptual and Motor Skills, 92(2), 595–602. https://doi.org/10.2466/pms.2001.92.2.595

Mielke, D. (2007). Coaching Experience, Playing Experience and Coaching Tenure. International Journal of Sports Science & Coaching, 2(2), 105–108. https://doi.org/10.1260/174795407781394293

Morse, A. L., Shapiro, S. L., McEvoy, C. D., & Rascher, D. A. (2007). The Effects of Roster Turnover on Demand in the National Basketball Association (SSRN Scholarly Paper ID 1690885). Social Science Research Network. https://doi.org/10.2139/ssrn.1690885

Motomura, A. (2016). MoneyRoundball? The Drafting of International Players by National Basketball Association Teams. Journal of Sports Economics, 17(2), 175–206. https://doi.org/10.1177/1527002514526411

Motomura, A., Roberts, K. V., Leeds, D. M., & Leeds, M. A. (2016). Does it pay to build through the draft in the National Basketball Association? Journal of Sports Economics, 17(5), 501–516.

NBA. (2017). 2017 NBA COLLECTIVE BARGAINING AGREEMENT – PRINCIPAL DEAL POINTS. https://official.nba.com/wp-content/uploads/sites/4/2017/01/NBA-2017-CBA-Principal-Deal-Points.pdf

Neese, B. (2015). Inside the Numbers: An Introduction to Sports Analytics.

Patt, J. (2015, May 19). Explaining how the NBA Draft Lottery works. SBNation. https://www.sbnation.com/nba/2015/5/19/8625903/nba-draft-lottery-2015-rules-how-works

Radovanović, S., Radojičić, M., Jeremić, V., & Savić, G. (2013). A Novel Approach in Evaluating Efficiency of Basketball Players. Journal for Theory and Practice Management, 37–45. https://doi.org/10.7595/management.fon.2013.0012

Rosen, J. (2016). Determining NBA Free Agent Salary from Player Performance.

Ryan, J. (2015). Examining the Validity of the Contract Year Phenomenon in the NBA. 87.

Silver, N. (2015, October 9). We're Predicting The Career Of Every NBA Player. Here's How. FiveThirtyEight. https://fivethirtyeight.com/features/how-were-predicting-nba-player-career/

Skinner, B. (2009). The price of anarchy in basketball. Journal of Quantitative Analysis in Sports, 6(1). https://doi.org/10.2202/1559-0410.1217

sklearn.cluster.KMeans—Scikit-learn 0.22.2 documentation. (n.d.). Retrieved April 5, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Wade, J. (2013, May 30). Idiots' Guide to the New NBA Free Agency Rules. Bleacher Report.

https://bleacherreport.com/articles/1656251-idiots-guide-to-the-new-nba-free-agency-rules

Walters, C., & Williams, T. (2012). To Tank or Not to Tank? Evidence from the NBA. MIT Sloan, March 2-3, 2012.

APPENDIX A: Rules and Regulations

A.1: The Salary Cap

The salary cap, a soft upper limit to how much teams can pay all their players, is the primary legislation concerning team payrolls. Before 1985, there was no salary cap, and teams were free to pay players as much as they wanted. From 1985 to 2002, there were just two 'levels' a team could be in: under or over the cap. Teams over the cap need to get permission from the league in order to increase their payroll any further, while teams under the cap are free to sign any players as long as it doesn't put them over. Following 2002, the team introduced a third level, tax-paying teams. This second line is called the luxury tax and is higher than the cap. The only changes this line made was that teams over this second line, the third level, had to pay a dollar for dollar tax on every cent they went over the luxury tax. These teams are still over the cap and need exceptions to increase their payroll. Teams under the luxury tax are in what is called the tax apron, and they, as well as under-cap teams, are treated the same as over-cap teams from before 2002. There are exceptions that exist such that teams can go over the salary cap and avoid paying the luxury tax. These are summarized in the below table.

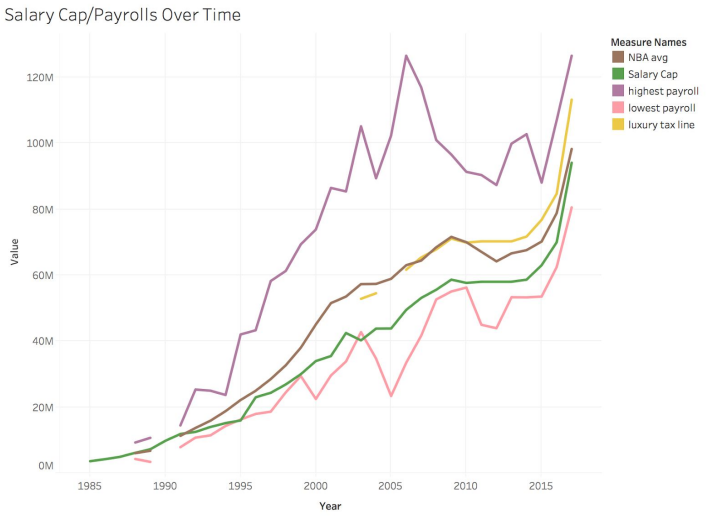


Figure A.1: Salary cap/ payrolls over time. Plot displays the salary cap, the highest NBA payroll, the lowest NBA payroll, and the luxury tax line each year from 1980 to 2018.

Table A.1: Exceptions to the salary cap.

	Larry Bird	**Early Bird**	**Non-Bird**	**Mid-Level**	**$1 Million**	**Rookie**	**Min.**	**Disabled**	**Qualifying Offer**
Who Qualifies	Own free agent, 3 seasons without changing teams as a free agent	Own free agent, 2 seasons without changing teams as a free agent	Own free agent, if not Larry Bird or Early Bird	Any	Any	Team's first round draft pick(s)	Any	Any	Own free agent, first three years in league, entered NBA in 98-99 or later
Minimum Years of Contract	1	2	1	1	1	3	1	1	1
Maximum Years of Contract	7	6	6	6	2	3 + team option	2	6	1
Maximum Salary	Max.	Greater of 175% of previous salary or average salary	Greater of 120% of previous salary or 120% of minimum salary	Average salary	$1 mil in 98-99, increases $100,000 per year thereafter	120% of scale amount	Min.	Lesser of 50% of injured player's salary or average salary	Greater of 125% of previous salary or minimum salary plus $150,000
Maximum Raises	12.5%	12.5%	10%	10%	10%	10%	10%	10%	N/A
Can be split?	No	No	No	Yes	Yes	No	No	No	No
Other				Expanded later to include 3 varieties	Cannot be used in consecutive seasons	Restricted free agency following option year		Approval from league, time limit.	Makes player a restricted free agent

A.2: Rules Regarding the NBA Draft, Free Agency, and Trades

A.2.1 NBA Draft

- Every summer, prospective players (typically from American universities) enter the draft

- Teams are granted one pick each round (picks can be traded), amount of rounds has changed:
  - 10, 7, 3, 2 : 1980-84, 1985-87, 1988, 1989-2018
- Picks were in reverse order of the standings until 1985
  - 1985-89, lottery introduced, all non-playoff teams had an equal chance
  - 1990-present, worse teams have better odds of first pick, specific odds have changed
- Beginning in 2005, a player had to be 19 to enter the draft (encouraged players to take at least one year of college)

A.2.2 Free Agency

- From our first year of study, 1980, to 1988, there was only restricted free agency
  - Teams had 'right of first refusal' (ROFR) : a player could work out a contract with another team at the end of current contract, but original team can match the offer to keep the player
- Unrestricted free agency introduced in 1988
  - Initially had to be in league 7+ years and played through 2 contracts
    - This has since changed
  - Could sign anywhere (no ROFR)
- Cap comes into play, teams over the cap must use an exception to sign free agents

A.2.3 Trades

- Teams can trade assets, in a two way trade both teams must send and receive: contracted player, future draft pick, draft rights to an “NBA prospect”, right to swap unencumbered picks in future draft, 75k or more
- No limit on how many teams can be involved, but all teams in a multi way trade must ‘touch’ another team, i.e. sending or receiving: active player contract, 750k cash, future pick(s) that will become a pick (not a protected pick that may become cash), draft rights to an “NBA prospect”