The NBA in the 2010s

Matthew Zimolzak, Li Zhou
University of Michigan

Author Note

This project was developed in fulfillment of an educational requirement as specified by the SIADS 591/592 milestone course offered through the Master of Applied Data Science program by the School of Information at the University of Michigan. The authors Matthew Zimolzak and Li Zhou may be contacted with inquiries via email at zimolzak@umich.edu and lizhoula@umich.edu respectively.

Table of Contents

Motivation

For years, the use of analytics in sports has been an effective way to diagnose what is happening on the field of play. This has become increasingly true with recent developments in data collection and management technologies that facilitate the application of diverse numerical approaches to sports data. In this project, we will explore NBA data specific to the 2010s starting with the 2010-2011 season and ending with the 2019-2020 season. As the sport of basketball continues to evolve, so do the data and trends within. We aim to address specific questions that include the following:

- Are there any trends that indicate a change in playing style?
- Who are the stars of the NBA past, present, and future?
- Which players are the most overvalued/undervalued?
- How has the league-wide distribution of star players changed over the last decade?

Data Sources

The first data set was compiled from Basketball Reference which offers one of the most complete and in-depth catalogues of basketball statistics on the web. Utilizing HTML web scraping techniques to create a CSV file, we gathered regular season per-game box score statistics for every player from each league year starting with the 2010-2011 season and ending with the 2019-2020 season for a total of 6,208 records. Variables of interest include, but are not limited to, points, assists, rebounds, steals, blocks, turnovers, and shooting percentage. Additionally, we included a new field to denote which year ('Year') is associated with each record. For example, the 2010-2011 season is labeled simply as 2011. (https://www.basketball-reference.com/leagues/NBA_2020_per_game.html)

The remaining datasets were gathered from sportsdata.io, an API that returns JSON formatted data. We selected unique data outside the realm of standard basketball statistics such as height, weight, salary, and years of experience for *active* players as well as team data consisting of city, division, conference, and colors for a total of 576 and 30 records respectively. (Player data: https://api.sportsdata.io/v3/nba/scores/json/Players?key=a3824595f2f740dbb21dd847e49ba332
Team data: https://api.sportsdata.io/v3/nba/scores/json/Teams?key=a3824595f2f740dbb21dd847e49ba332)

Data Manipulation Methods

Our box score dataset first required some basic cleaning. Cleaning steps include removing intermediate headers, relabeling hyphenated player positions to just a single position (e.g., SF-SG became SG), removing non-alphanumeric characters (asterisks which denote Hall of Fame status), replacing non-English characters with their English equivalent (to allow for the joining of the datasets), and consolidating organizations that underwent name changes into a single name. Additionally, some players had multiple records per season which indicates playing for multiple teams in a season. In league wide analysis, we removed the records with the team name 'TOT', the aggregate record for the player during that season, instead electing to use the team splits so as to avoid double counting. Then, when doing individual player analysis, we

kept the records with the team name 'TOT' in order to accurately reflect that individual's performance. Other manipulations unique to this dataset (outside of cleaning) involve converting string values to numeric values, aggregating relevant statistics by both season and position in terms of both absolute values and proportions, and sorting/ordering/aggregating by a variety of fields. Finally, we made use of some advanced metrics, the explanations and formulas for which are given below:

Approximate Value (AV)
Approximate value is the metric which is an estimate of a player's value, making no fine distinctions, but, rather, distinguishing easily between very good seasons, average seasons, and poor seasons. (Approximate Value (AV) Explained)

$$\begin{aligned} Credits \;=\; & (Points) + (Rebounds) + (Assists) + (Steals) + (Blocks) \\ & - (Field\ Goals\ Missed) - (Free\ Throws\ Missed) - (Turnovers) \end{aligned}$$

$$Approximate\ Value\ (AV) \;=\; (Credits^{(3/4)})/21$$

Trade Value (TV)
Trade value is the estimate using a player's age and his approximate value to determine how much value a player has left in his career. Invented by Bill James. (Trade Value Explained)

$$Trade\ Value\ (TV) \;=\; [(AV - 27 - 0.75Age)^2(27 - 0.75Age + 1)AV]/190 \;+ 2AV/13$$

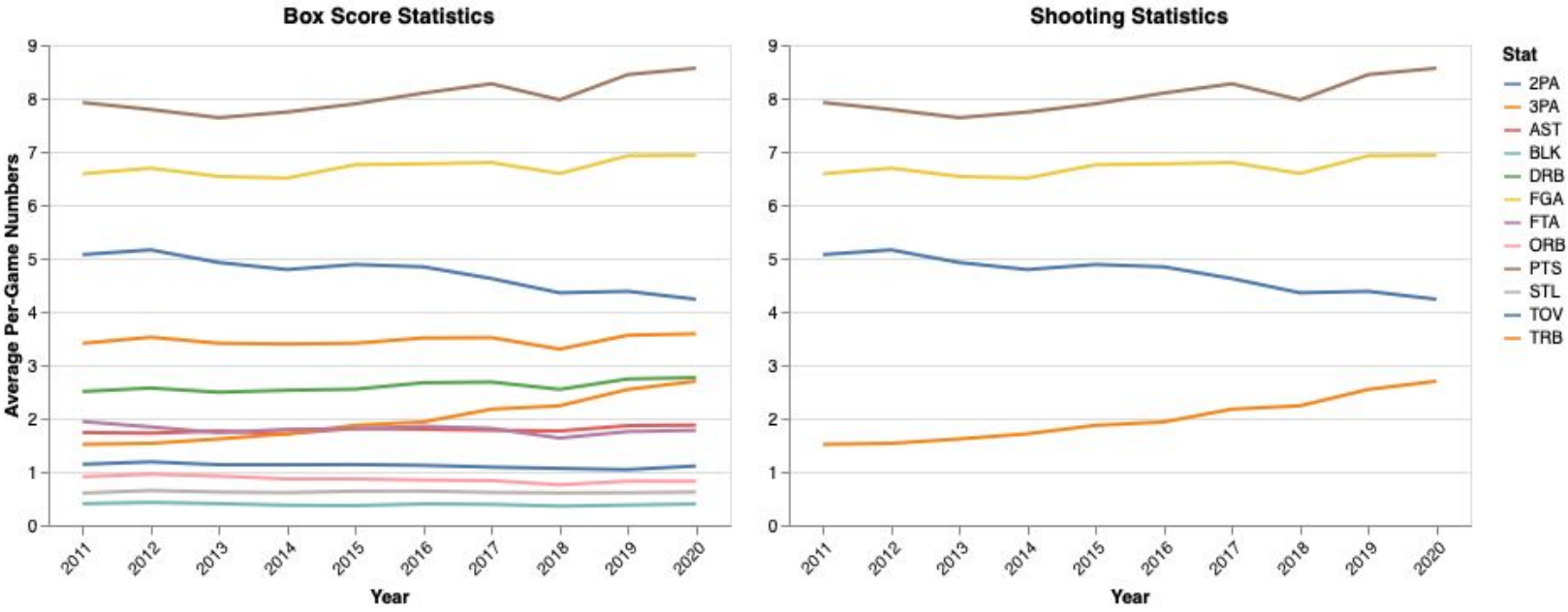
The datasets from the sportsdata.io API did not require nearly as much cleaning or manipulation. Aside from reconciling team names as was necessary with the first dataset, we only needed to add a pound sign (#) to the color fields so that they could be used in creating our visualizations. These two datasets alone did not prove useful in our analysis and only after joining with the per-game statistics was their value realized.

When necessary, we joined (pd.merge) the per-game statistics dataset with either the API player data using player name as the key or the API team data using team name as the key. The purpose of the former was to connect salary and experience with each player while the purpose of the latter was to connect team name with team colors to allow for coloring based on team. In order to allow for the comparison between Approximate Value and salary, we performed a z-transformation so that the values for each would be on a comparable scale. These values were then filtered by experience for one specific situation as we discuss below. Additionally we fit a linear regression line involving salary and Approximate Value which was then plotted on top of a scatter plot.

Analysis and Visualization

Question 1: Are there any trends that indicate a change in playing style?

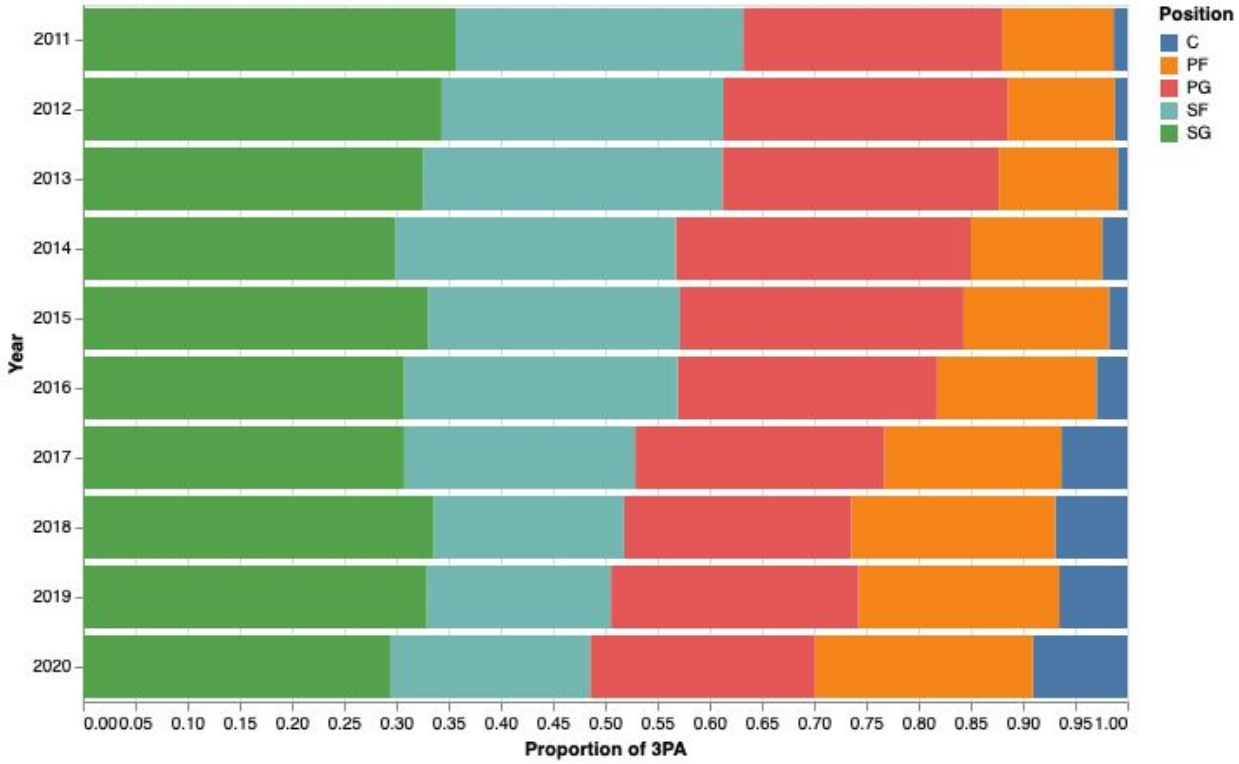
We begin by examining the mean per-game box score statistics across all players over the course of the last decade. The justification for using the mean, as opposed to the total counts, is to mitigate the effect of outlier seasons on the underlying trends. The two offenders here are the 2011-2012 strike-shortened season in which only 66 games were scheduled for each team (as opposed to the standard 82 games) and the 2019-2020 COVID-19-shortened season where varying impacts resulted in an average of 69.2 games per team. Opting for total counts would produce considerable declines for the corresponding seasons in the charts pictured below.



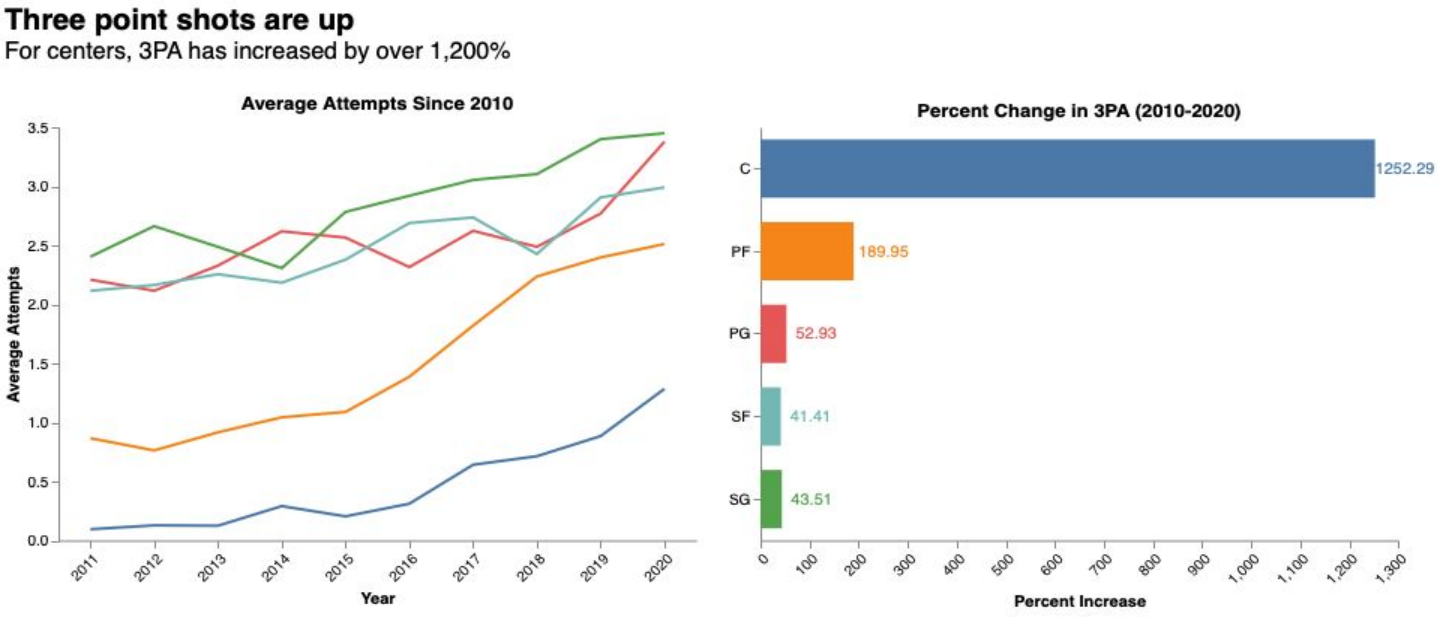
As you can see on the left, most of the statistics have remained relatively stable with the exception of two-point shots attempted (2PA) and three-point shots attempted (3PA). In fact, the number of two-point shots attempted per player per game has gone down from 5.07 to 4.23, a decrease of approximately 16.6% . Over the same time period, the number of three-point shots attempted per player per game rose from 1.51 to 2.70, an increase of approximately 78.8%. Alternatively, one might note that the proportion of three-point field goals attempted relative to all field goals attempted (FGA) rose from 0.23 during the 2010-2011 season to 0.39 by the end of the decade which likely attributed to the modest increase in points (PTS).

As the game of basketball continues to evolve, it certainly seems that the three point shot has become increasingly important over the years. A natural follow-up to this observation is to explore the positional distribution of three-point shots attempted. To do so, we approximated the total number of three-point shots attempted for each player by computing the product of games played and three-point shots attempted per game. Subsequently, we grouped these totals by position and year, then computed the proportion of all three-point field goals attempted by each position for each year. A keen eye might notice the trend taking place on the right side of the

chart where the proportion of all three-point field goals attempted attributed to the position of center (C) and power forward (PF) increase rather substantially from 0.01 to 0.09 and 0.11 to 0.21 respectively.



Considering only the above visualization, one might think to conclude that centers and power forwards are stealing three-point shot opportunities from point guards (PG) and small forwards (SF), whose proportions actually did decrease, but that is not what's unfolding. Rather, what we're seeing is a notable increase in three-point shot attempts across the board with the most pronounced effect occurring amongst big men. These findings suggest that not only is the NBA transitioning towards a style of play that places more of an emphasis on the three-point shot but also that both centers and power forwards, positions that traditionally play inside, are becoming more versatile with an increased ability to stretch the floor.



Question 2: Who are the stars of the NBA past, present, and future?

We first begin by computing the Approximate Value (AV) and Trade Value (TV) metrics that we have previously discussed. As AV can be used as a guide to evaluate player performance, some natural follow up questions include: who was the best player over the course of last decade? Which players produced the best individual season performances? For the purpose of *these rankings only*, we implement one additional condition by requiring that players have participated in *at least half* of the games for that season. Simply put, we feel that a player's durability should factor into what should be considered a *great* season. Among a few others, this disqualifies Anderson Varejão's 2012-2013 season where he had the 3rd highest AV while playing in only 25 games and Karl Anthony-Towns' 2019-2020 season where he had the 4th highest AV while playing in only 35 games. In determining the best players of the decade, we consider the top 10 for each year and assign points for their placement where 1st nets 10 points, 2nd nets 9 points, 3rd nets 8 points, and so on. Here are our findings:

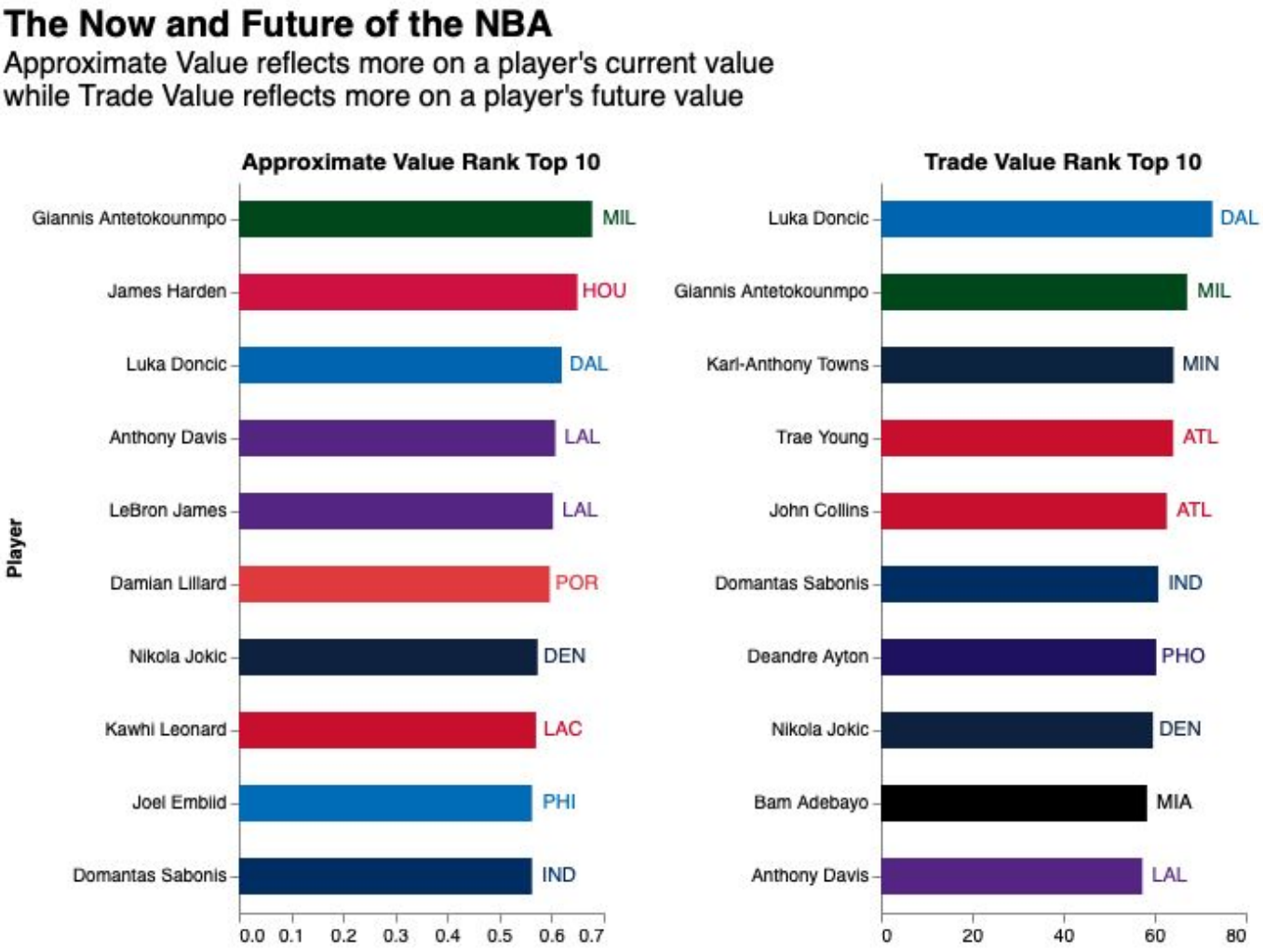
Year	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
2011	LeBron James	Dwight Howard	Kevin Love	Blake Griffin	Pau Gasol	Dwyane Wade	Kevin Durant	Amar'e Stoudemire	Zach Randolph	Deron Williams
2012	LeBron James	Kevin Love	Kevin Durant	Dwight Howard	Chris Paul	Andrew Bynum	Blake Griffin	Pau Gasol	Al Jefferson	LaMarcus Aldridge
2013	LeBron James	Kevin Durant	Kobe Bryant	Tim Duncan	James Harden	Chris Paul	David Lee	Al Horford	Carmelo Anthony	LaMarcus Aldridge
2014	Kevin Durant	Kevin Love	LeBron James	Anthony Davis	Blake Griffin	Chris Paul	DeMarcus Cousins	Carmelo Anthony	Stephen Curry	LaMarcus Aldridge
2015	Anthony Davis	Russell Westbrook	DeMarcus Cousins	James Harden	Stephen Curry	Chris Paul	LeBron James	Pau Gasol	LaMarcus Aldridge	DeAndre Jordan
2016	Stephen Curry	Kevin Durant	Russell Westbrook	James Harden	LeBron James	Anthony Davis	DeMarcus Cousins	Chris Paul	Draymond Green	Kawhi Leonard
2017	Russell Westbrook	James Harden	Anthony Davis	LeBron James	Karl-Anthony Towns	Kevin Durant	DeMarcus Cousins	DeMarcus Cousins	Giannis Antetokounmpo	Jimmy Butler
2018	Anthony Davis	LeBron James	Giannis Antetokounmpo	James Harden	DeMarcus Cousins	Russell Westbrook	Karl-Anthony Towns	Kevin Durant	Stephen Curry	Andre Drummond
2019	Giannis Antetokounmpo	Anthony Davis	James Harden	Joel Embiid	Karl-Anthony Towns	LeBron James	Nikola Jokic	Russell Westbrook	Nikola Vucevic	Kevin Durant
2020	Giannis Antetokounmpo	James Harden	Luka Doncic	Anthony Davis	LeBron James	Damian Lillard	Nikola Jokic	Kawhi Leonard	Joel Embiid	Domantas Sabonis

Player	Points	Appearances
LeBron James	75	10
Anthony Davis	56	7
James Harden	53	7
Kevin Durant	49	8
Russell Westbrook	35	5
Giannis Antetokounmpo	30	4
DeMarcus Cousins	29	6
Kevin Love	26	3
Chris Paul	24	5
Stephen Curry	20	4

Player	Year	AV
Giannis Antetokounmpo	2019	0.689623
Giannis Antetokounmpo	2020	0.677868
Russell Westbrook	2017	0.669007
Anthony Davis	2019	0.661592
James Harden	2019	0.654150
Anthony Davis	2018	0.654150
LeBron James	2018	0.651166
James Harden	2020	0.649672
Joel Embiid	2019	0.646680
James Harden	2017	0.646680

Rather unsurprisingly, we find that LeBron James was the best player over the last decade and by a fairly sizable margin. It is worth mentioning, however, that the two seasons in which Kevin Durant did not land amongst top 10 were obstructed by injury where he played in only 27 games during the 2014-2015 season and missed the 2019-2020 season entirely. Another noteworthy observation is how the AV metric correctly indicates the league MVP in 7 out of 10 years with the most egregious omission being Derrick Rose who fell outside of the top 10 despite being named MVP. As far as the best individual performance is concerned, we see that Giannis Antetokounmpo's two most recent seasons were atop the list and that each of the top 10 occurred in the latter half of the decade. So how does this relate to TV?

Recall that TV takes AV into account while also considering a player's age to estimate how much a player has left in his career. As such, we can use this metric to gain some insight regarding potential future stars of the NBA. It is important to mention that in calculating TV, we have not restricted AV here with a minimum participation qualifier, as we have above, because one shortened season is not necessarily indicative of a player's career. Here we have performed a join with pd.merge on the key of 'Tm' (team) in order to combine team colors with not only TV but also AV to allow for comparison between metrics. There is certainly some overlap between the two metrics as 5 players appear in the top 10 for each but a regular observer of the NBA might notice that TV is heavily influenced by age (with an average age of 22.8 compared to 26.5 for AV). Notable snubs from the TV top 10 that can be considered "names to watch" include Ben Simmons and Ja Morant who, in combination with Luka Doncic, account for the last 3 rookie of the year awards, along with Devin Booker, Jayson Tatum, and Zion Williamson.



Question 3: Which players are the most overvalued/undervalued?

An alternative way to think of this question is which players are underperforming or over performing compared to their current salary. In order to evaluate this, we will consider AV as a gauge for performance and salary as the *true* value of a player (remember that neither AV nor TV account for salary). More specifically, we calculate z-scores for AV and z-scores for salary, then identify the largest differences between the two. A high, positive z-score for AV means that an individual player is a top tier player while a low, negative z-score indicates that a player has performed poorly (of course, this is relative as simply playing in the NBA is an incredible accomplishment). Similarly for salary, a high, positive z-score for salary means that a player's salary is among the highest in the league while a low, negative z-score indicates that a player is not paid much (again, relative – the league minimum salary has been around $900,000 per season in recent years). We then subtract the salary z-score from the AV z-score. Consequently, the lowest, most negative differences indicate the most overvalued/overpaid players while the highest, most positive differences indicate the most undervalued/underpaid players. We begin with those that are overvalued:

Player	Pos	Tm	AV	Salary	Experience	AV_z	Salary_z	z_difference
Blake Griffin	PF	DET	0.305099	36810996.000000	10.000000	0.420069	2.937060	-2.516992
Mike Conley	PG	UTA	0.327895	34502130.000000	13.000000	0.597698	2.694662	-2.096965
Stephen Curry	PG	GSW	0.467138	43006362.000000	11.000000	1.682725	3.587485	-1.904760
Chris Paul	PG	OKC	0.482074	41358814.000000	15.000000	1.799112	3.414516	-1.615404
Kemba Walker	PG	BOS	0.428215	34379100.000000	9.000000	1.379426	2.681746	-1.302320
Paul George	SF	LAC	0.460449	35450412.000000	10.000000	1.630599	2.794219	-1.163619
Isaiah Roby	PF	OKC	0.008468	1517981.000000	1.000000	-1.891372	-0.768201	-1.123172
James Johnson	PF	MIA	0.204929	16047100.000000	11.000000	-0.360485	0.757149	-1.117633
Russell Westbrook	PG	HOU	0.549871	41358814.000000	12.000000	2.327409	3.414516	-1.087107
Tobias Harris	PF	PHI	0.457092	34358850.000000	9.000000	1.604442	2.679620	-1.075178

Blake Griffin at the top is an interesting discovery. His AV for the 2019-2020 season was a career low by a pretty sizable difference and could qualify as an outlier season. With that being said, performing at his career mean AV, about 0.488, would still keep him in or near the top 10. Coupled with the facts that his pay is set to increase per his contract and that he's played in 75% of games only once over the last 5 seasons it's safe to say that the return on investment is not ideal. Mike Conley checks out – his AV was on the lower end for this season but his career mean AV is not much higher and he is only getting older which tends to take a toll on performance. However, the remaining names on this list, specifically the high salaried players, suggest that an underlying trend is at play. In general, it seems that when a player is paid a salary among the highest in the league, it is fairly uncertain whether they perform at an equivalent level. In fact, the likelihood of a top 50 salaried player also being in the top 50 AV is not

Top N salaried players	# in top N salary and AV	Proportion
10	3	0.3
20	7	0.35
30	14	0.47
40	23	0.57
50	31	0.62

much greater than the likelihood of a coin toss turning up heads.

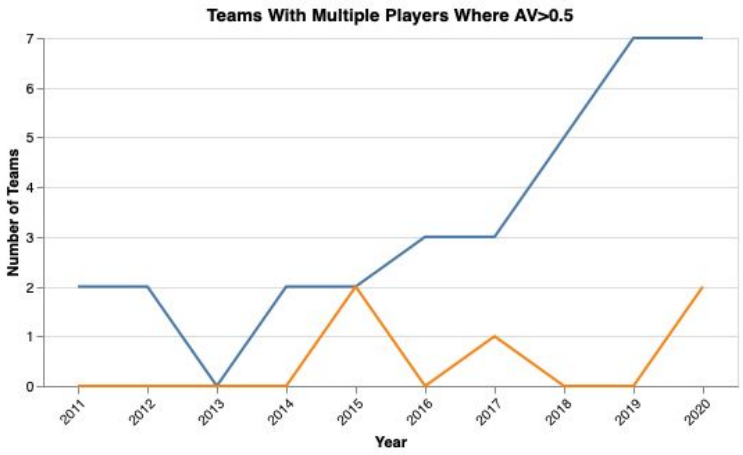
We now transition toward those players that are undervalued. Here we have restricted player experience to greater than or equal to four years. Without this restriction, we would exhibit a list similar to the top 10 TV players that we discussed in the previous section. This makes sense as rookie contracts, which can last as long as four years, tend to be relatively low which would give any exceptional young talent a fairly decent chance to land on this list. Effectively, we have limited eligibility for this list to those players who have signed at least a second contract.

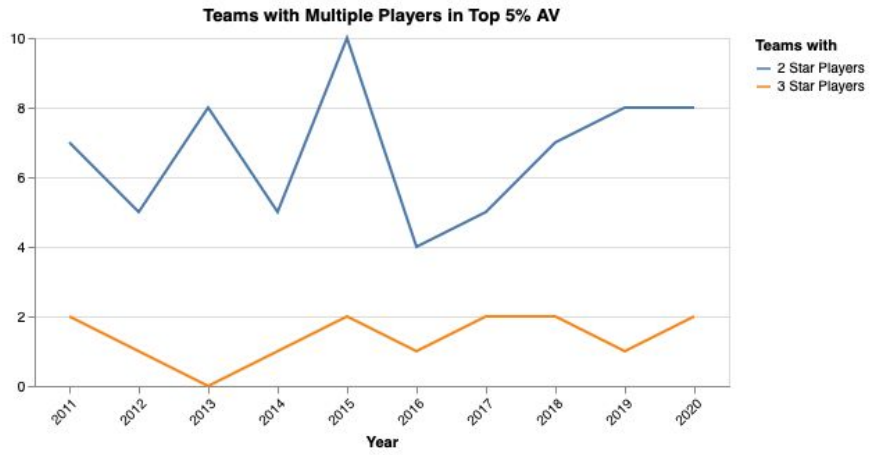
Player	Pos	Tm	AV	Salary	Experience	AV_z	Salary_z	z_difference
Hassan Whiteside	C	POR	0.556180	1620564.000000	8.000000	2.376566	-0.757431	3.133997
Jusuf Nurkic	C	POR	0.541952	12888889.000000	6.000000	2.265698	0.425582	1.840116
Richaun Holmes	C	SAC	0.435067	5005350.000000	5.000000	1.432815	-0.402077	1.834892
Jeff Teague	PG	MIN	0.364765	1620564.000000	11.000000	0.885005	-0.757431	1.642436
Carmelo Anthony	PF	POR	0.355668	1620564.000000	17.000000	0.814121	-0.757431	1.571552
Paul Millsap	PF	DEN	0.344649	1000000.000000	14.000000	0.728252	-0.822581	1.550833
Marquese Chriss	PF	GSW	0.352008	1824003.000000	4.000000	0.785597	-0.736073	1.521670
Montrezl Harrell	C	LAC	0.452041	9258000.000000	5.000000	1.565085	0.044390	1.520694
Glenn Robinson III	SF	GSW	0.342801	1620564.000000	6.000000	0.713852	-0.757431	1.471283
Clint Capela	C	HOU	0.532397	16000000.000000	6.000000	2.191245	0.752204	1.439041

Hassan Whiteside seems to have outperformed his salary by quite a large margin while Carmelo Anthony continues to prove his worth in the association even at the tail end of his career. Moreover, it appears that Neil Olshey, the general manager for the Portland Trailblazers (POR), has a knack for cost efficiency as evidenced by having three rostered players on this list. Also of note is that 8 of the top 10 here are either power forwards or centers which seems to suggest that the league does not prioritize these positions as much with the increasing prevalence of the three point shot. (See Appendix for a scatter plot containing AV and salary data)

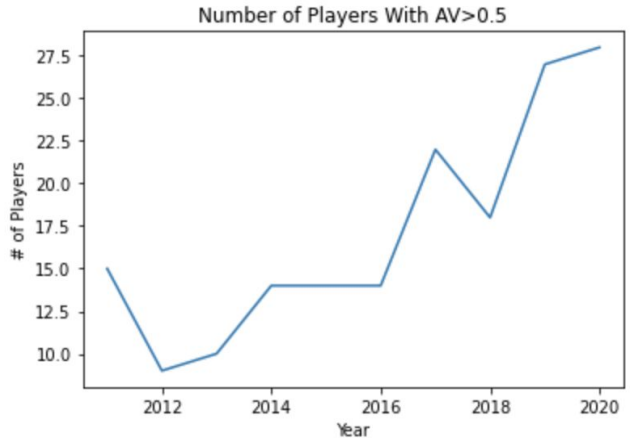
Question 4: How has the league-wide distribution of star players changed over the last decade?

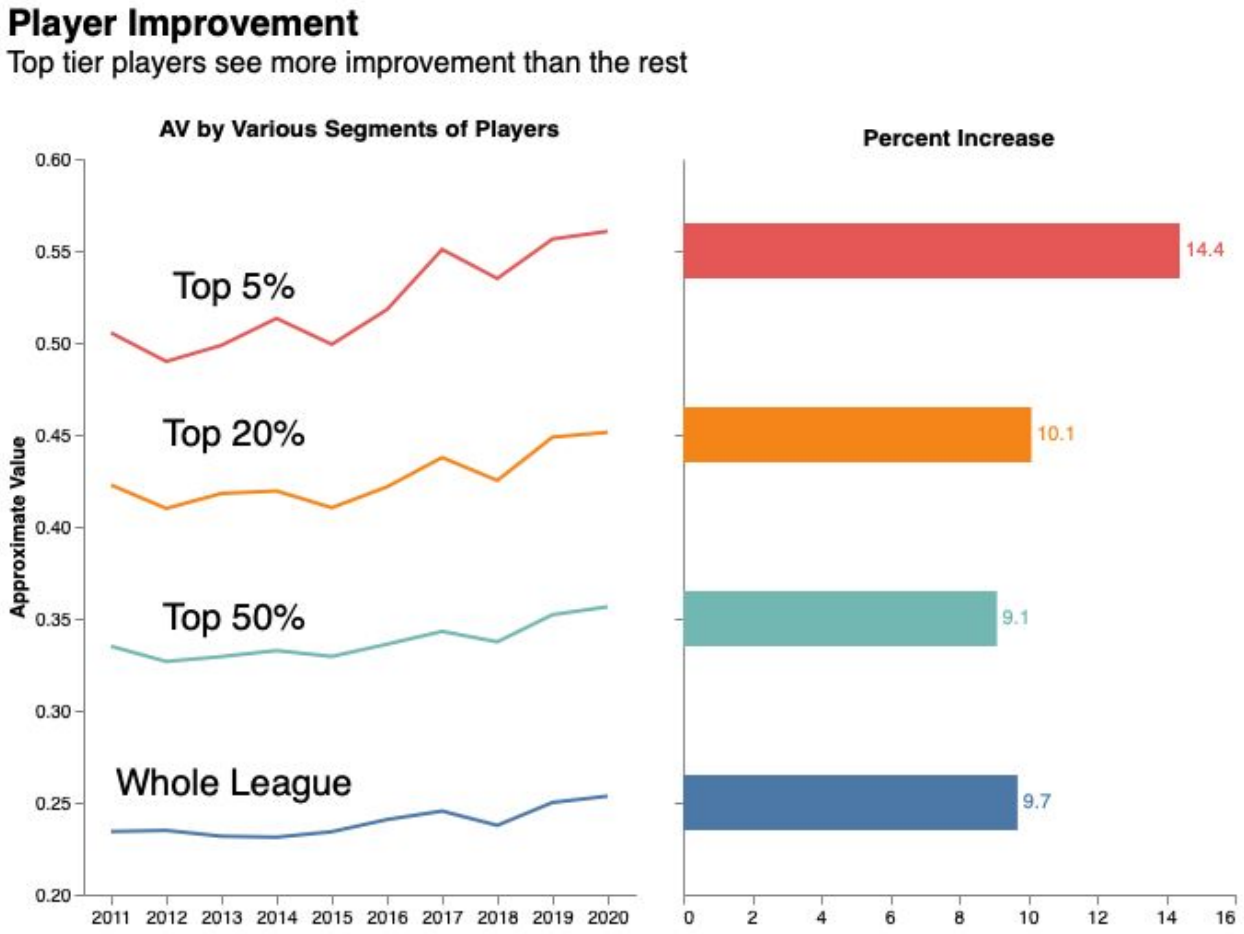
This question was directly inspired by formation of teams like the 2010-2014 Miami Heat with Lebron James, Dwayne Wade, and Chris Bosh, the 2014-2018 Cleveland Cavaliers with Lebron James, Kyrie Irving, and Kevin Love, and the 2016-2019 Golden State Warriors with Steph Curry, Kevin Durant, and Klay Thompson. We operate under the general assumption that an increase in teams with more than 1 star player indicates that the distribution of star players has become more concentrated, e.g. that they are more likely to play together. We tried to assess this question first by defining a star player as having an AV greater than 0.5 and then by defining a star player as having an AV in the top 5% of all players.





Given these two graphs it looks like our answer depends on the definition of a star player. If we consider a star player to be those with an AV greater than 0.5, the answer to our question seems to be that the distribution of star players *is* more concentrated. However, if we consider a star player to be those with an AV in the top 5% of all players, the answer to our question seems to be that the distribution of star players *is not* more concentrated. So which is right? Consider the total number of players with an AV greater than 0.5 – we see that it has nearly doubled by the end of the decade. This suggests that defining a star player as those having an AV greater than 0.5 is not an appropriate way to evaluate whether the distribution of star players has changed because we have more players that meet the criteria which consequently drives the number of teams with more than one star player upwards. Furthermore, we actually see an increase in AV across the league as a whole with the most pronounced increase being amongst players that are in the top 5%.





We can now provide a more definitive answer to our question by concluding that the distribution of star players has *not* become more concentrated – rather, it has remained relatively the same. What we *are* seeing is an overall increase in the quality of players throughout the league, particularly amongst the top tier players. Ultimately, this trend may give the appearance that the distribution of star players is more concentrated but there is simply not enough evidence here to directly support any such claim.

Statement of Work

Matthew was responsible for collecting the per-game statistics through the utilization of HTML web scraping from Basketball Reference. His role entailed a greater emphasis on writing and elaboration. Li was responsible for gathering the data from sportsdata.io. His role entailed a greater emphasis on coding and creating visualizations. There was, however, considerable overlap between roles as we were both responsible for, to some extent, the necessary exploration, visualization, and analysis.

References

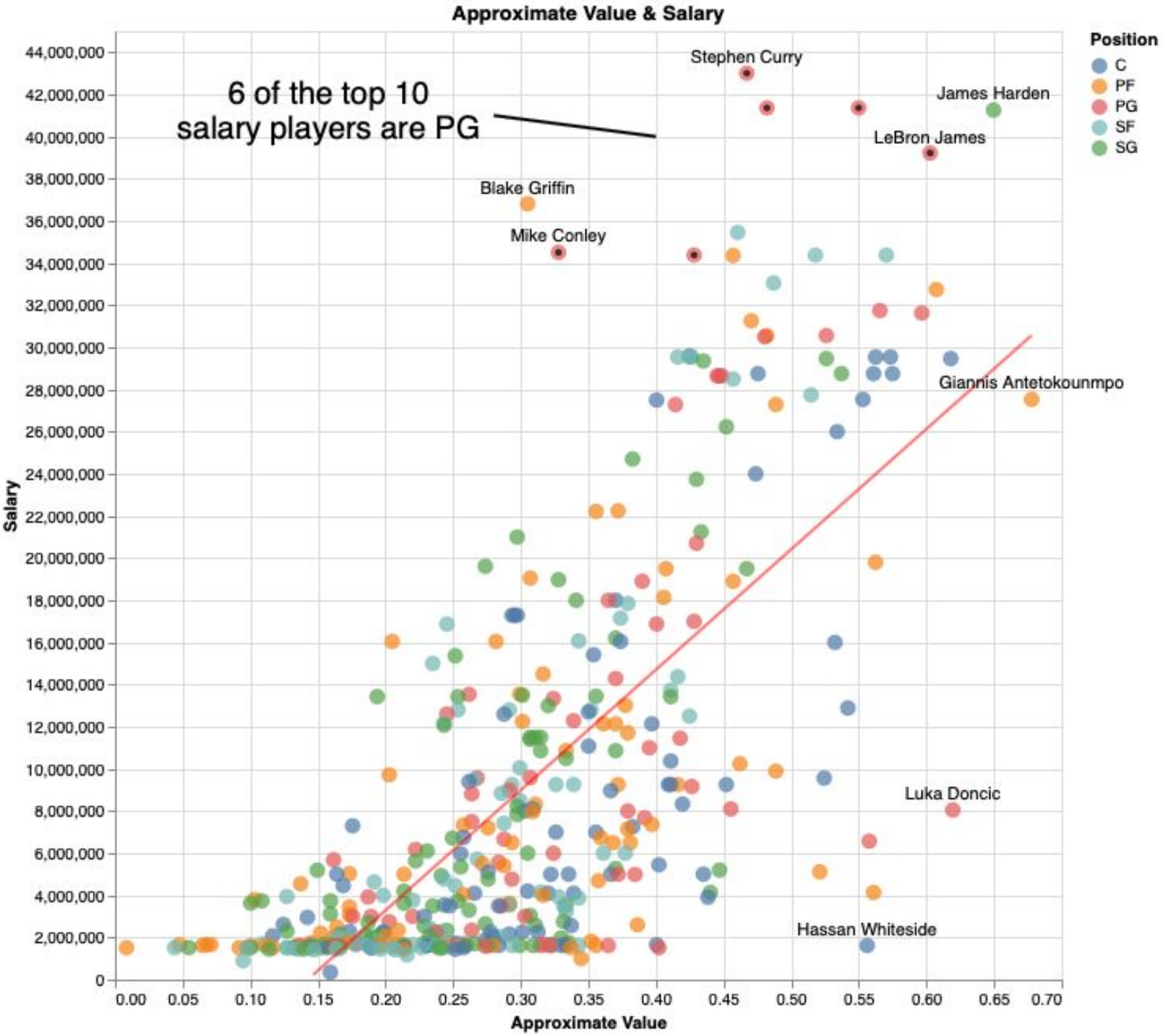
Basketball Statistics and History. Retrieved February, 2021, from https://www.basketball-reference.com/

NBA Database: NBA API: Sports Data API. Retrieved February, 2021, from https://sportsdata.io/nba-api

Approximate Value (AV) Explained. (2020, June 19). Retrieved from https://www.nbastuffer.com/analytics101/approximate-value/

Trade Value Explained. (2020, June 23). Retrieved from https://www.nbastuffer.com/analytics101/trade-value/

Appendix



If every point on this scatter plot is to be classified as either overvalued or undervalued, those that fall above the line would be classified as overvalued while those that fall under the line would be classified as undervalued