Team 055 Final Report

Introduction/Motivation

With an interest in sports, our team focused on analytics for the NBA (National Basketball Association). Currently, players are commonly evaluated based on a metric called "real plus/minus" (RPM). RPM is a measure of a player's impact on the court, but it doesn't account for individual defensive assignments. A player's performance can change based on defensive assignments, and thus RPM can be inconsistent based on the individual defensive matchup. Accounting for individual defensive assignments is the motivation for this project, as it will impact how teams strategize their defensive (and offensive) matchups, and inform coaches on the best player matchups, resulting in better overall team performance.

Problem Definition

Our goals were to assess which methods are optimal for analyzing player versus player matchup data and to create and visualize a new performance metric that accounts for defensive assignments on the court.

The questions we are looking to answer are listed below:

- How can we incorporate defensive matchup results as a contributing factor of a player's defensive-performance metric?
- Based on defensive matchup metrics, can we identify groups (or 'positions') of players who defend each other outside of the typical Point Guard, Shooting Guard, Small Forward, Power Forward, and Center positions?
- Are there other quantifiable factors which affect these metrics?
- Will this metric be more meaningful than the current performance metric, RPM and/or its defensive sub-metric: DRPM (Defensive Real Plus-Minus)?
- Can we use this metric to help coaches choose defensive assignments that will increase their likelihood of winning games?
- Can we use this metric to support strategic NBA analysis (e.g., NBA Presidents approaching free agency or the trade market)?
- Can this metric be used as leverage within NBA gambling (e.g., Daily Fantasy Sports)?

Literature Survey

In our literature review, we found 4 different types of articles: team performance, individual performance, patterns of play, and general analytics. The team performance, patterns of play and general analytics articles were useful in understanding how the game is played, and insight into the generally important player factors and metrics. We have also used them to gain knowledge on the general state of NBA analytics.

The individual performance articles have helped us to understand the key metrics currently in place for players, particularly the plus-minus metric, which is an easy-to-understand baseline number to evaluate player value and performance. This overall metric encompasses both offensive and defensive team output, in the context of when each player is in the game. While other individual metrics, such as points scored per game, shooting percentage, and rebounds/blocks/steals per game are available, the plus-minus metric is the main one cited within the literature as most used to directly evaluate and compare player performance and value.

Our analysis is focused on individual defense, but there are relatively few articles in the literature that specifically address this game factor. Most that address defense do so with relatively high-level metrics, which collectively create an overall Defensive Player Rating. While this methodology provides some good insight into factors that may play key roles in our analysis, it does not address specific player interactions, and how that rating may change when players are assigned to guard different player archetypes. The literature review confirms that our analysis is unique and valuable, as our data began being publicly tracked during the 2017-18 NBA season. It will occupy an unfilled niche in this analytical world, as it seems that until now, specific defensive assignments have not been analyzed, and as such, our analysis is both useful and effective.

Proposed Method

Intuition & Innovation

Our approach drills defensive impact down to individual defensive assignments by player and position, which has not been publicly analyzed in the past. This method provides coaches with information on which defensive assignments allow players to be most effective on defense, thereby improving upon the team's overall performance on the court. For example, this approach is innovative in producing optimal defensive assignments for a coach rather than using intuition to determine which player should guard each opposing offensive player. As such, this approach should increase the likelihood of better team performance and hence, winning games.

This analysis provides coaches with direct information on the player archetypes which individual players are most (and least) effective against, both on offense and defense. For example, our analysis shows that Stephen Curry is most effective on offense when guarded by an opposing PF, and he is most effective on defense when matched up against a fellow PG. Or more specifically, he performs best against shorter-than-average and lighter-than-average PFs and is at his best defensively against taller-than-average and lighter-than-average PGs. With this information, his coach can design schemes on both sides of the ball to take advantage of these matchup benefits.

At a more technical level, the impact metrics are derived by using a percent different relative to a player or team's average output. Specifically, we evaluate three domains: player efficiency, player points and team points. By using a percent difference from average, we introduce a relative statistic that will considers player variability and thus mitigates risk of our metrics being impacted by absolute differences. Per our literature survey, most NBA statistics are either impacted by a player's aggregate numbers and/or fail to account for variability from average in their statistics, and thus we believe this is a key innovation. In addition, many statistics (e.g., RPM) attempt to encompass a player's impact into one number. While this is convenient in nature, we believe it lacks differentiation and thus our three metrics can better tell the story of a player's impact.

Approaches

We focused on 3 metrics for this project: true shooting percentage (TS%) player points per minute, and team points per minute. TS% is a commonly used measure of a player's offensive efficiency—a player can score many points, but if they take too many shots to do so, their team's offense will suffer due to inefficient scoring. Player points per minute is used to account for how many points a given player scores for their team per minute in a game. Team points per minute is used to account for the player's effect on the team's overall performance on the court, as a player can have a significant impact even if they aren't directly scoring points. For example, using the Stephen Curry example above, our analysis shows that, when he is assigned to guard a fellow PG, that player's shooting performance is almost 11% worse than their "normal" performance. Additionally, he holds the opposing PG to 53% less than their scoring average, and the opposing team only scores about 45% of their normal offensive output when Curry is matched up as such. So not only is he effective at limiting the opposing PG, but his matchup enhances his team's overall defensive performance further than expected.

Data Collection and Processing

We used the NBA_API to scrape the previous four seasons worth of NBA player's matchup results (e.g., the result of one player guarding another), each player's individual statistics, and each team's statistics from NBA.com. In addition, we scraped each player's listed position from https://www.basketball-reference.com/. After performing extensive cleaning on the initial datasets, we merged them into an aggregated dataset based on either the player ID (per NBA.com) or player names (per basketball-reference.com). This scraping resulted in a matchup data file of about 1.3 million rows and 36 columns, with the other two files being reference data sets to add to the matchup file.

We then split this foundational dataset into two datasets for further analysis. First, we augmented the base dataset, which maintains each season as a unique entity for each player, with statistical calculations such as player-points/minute, team-points/minute, and TS%. Concurrently, we developed a second dataset by grouping each players statistics across the various seasons. For example, Stephen Curry's individual datapoints across the seasons were aggregated into a single datapoint. Similarly, we then calculated all per-minute averages across the seasons, as well as each player's TS% across the accumulated seasons. Finally, we calculated the percent difference of an offensive player's performance when guarded by a given defender versus how they do on average within each dataset to develop our impact metrics. Please see the following Figure 1 for a flowchart of this process along with related files.

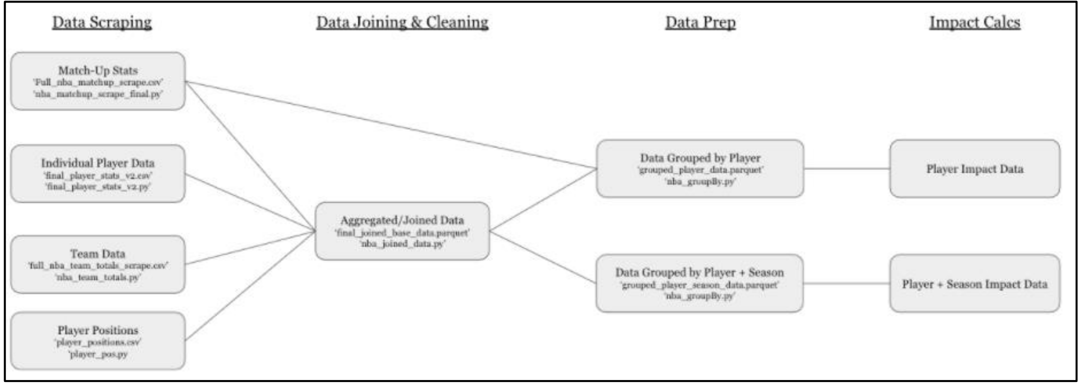

Data Scraping
Data Joining & Cleaning
Data Prep
Impact Calcs
Match-Up Stats
'Full_nba_matchup_scrape.csv'
'nba_matchup_scrape_final.py'
Individual Player Data
'final_player_stats_v2.csv'
'final_player_stats_v2.py'
Team Data
'full_nba_team_totals_scrape.csv'
'nba_team_totals.py'
Player Positions
'player_positions.csv'
'player_pos.py'
Aggregated/Joined Data
'final_joined_base_data.parquet'
'nba_joined_data.py'
Data Grouped by Player
'grouped_player_data.parquet'
'nba_groupBy.py'
Data Grouped by Player + Season
'grouped_player_season_data.parquet'
'nba_groupBy.py'
Player Impact Data
Player + Season Impact Data

Fig. 1 – Data cleaning and processing flow chart

Exploratory Data Analysis

Our initial approach was a player network, which is a directed graph comprised of nodes and weighted edges. The nodes in our graph represent players in the NBA, and the edges represent matchups between offensive and defensive players. The nodes in our network were built using the Player IDs of the offensive players in each matchup. For a player to be represented as a node, they had to have been the offensive player in a matchup at least once, and the defensive player in a matchup at least once. Player names identify each node, with the weight of the edges representing the amount of time that the matchup endured. The graph is directed, because we are analyzing the one-way relationship of a defensive player guarding an offensive player.

Subsequently, we explored the network graph for potential groupings. Initially, we hoped to identify clusters of players that could represent more nuanced position groups. Currently, NBA players are typically categorized into 5 common positions (PG, SG, SF, PF, C) which can be limiting in nature. We used the NetworkX and scikit-learn Python libraries in various manners, including clustering functions, greedy_modularity_communities, cliques, connected components, Louvian Method, and Affinity Propagation. While Affinity Propagation provided us the most usable clusters based on defensive player matchups, it still provided us at minimum 217 groupings of players. After further review, we were not able to associate these clusters with meaningful groupings (based on human judgement) and thus tabled this piece of exploratory data analysis for future endeavors.

Continuing with our exploratory analysis, we used principal components analysis and linear regression in Python and R to determine which factors to use for our metrics comparisons. These methods showed us that the two factors most affecting our metrics are the offensive and defensive player position matchup and the height difference between the two players. While we determined that complex combinations of these variables provided excellent computational results, we the importance of distilling these variables into factors that are easily understandable and interpretable by NBA coaches (and any other end-user) in their subsequent application. We first created 20 categories of position matchups, one for each combination of the 5 positions (e.g., PG-PG, PG-SG, etc.). Then we broke down each position into 'short' and 'tall' players, with the division being the median height of each position. While the individual height differences provided a more granular view into defensive player performance, this short/tall paradigm provides coaches and analysts a useful way of judging how the height difference factors into the metrics. We used the same median-based paradigm for player weight, with the categories being ‘light’ and ‘heavy’.

Visualization

Using Tableau as our front-end visualization tool, we first integrated our network graph results into an interactive program to visualize our graph of player matchups. This allows for exploratory research based on our network results with filterable data, using quick slicing of the data to see matchups based on position groupings and total matchup minutes.

Secondly, we created 6 interactive dashboards to present the data in various digestible formats. These dashboards are the primary means for coaches, analysts, NBA gambler, or any other end-user to analyze individual player matchups. The dashboards use tables to present player data, along with high-level graphs, charts, and flash facts for summary information. They incorporate filters and search functionalities for interactive insight into specific teams, players, and matchups to provide an end-user insight into the detailed data from various perspectives. For example, coaches can use certain views on their own players, or upcoming opponent’s players, during game plan preparation. Alternatively, they provide management an opportunity to search for underrated defensive players when approaching the trade market or free agency or an NBA gambler insight into tonight’s most efficient matchups. Examples of the dashboard analyses are in our poster presentation. A brief summary of each dashboard is as follows:

1. Player Summary: A dashboard that provides a list of each player along with his basic information such as name, position, height, weight, and current team.
2. Matchup Network: A dashboard that presents a matchup network. Specifically, a network graph of NBA player matchups.
3. Summary: Position-Position Impact: A dashboard that shows the position vs position matchup impact data.
4. Detailed: Position-Position Impact: A dashboard that shows the position vs position impact data with additional groupings for height and weight.
5. Defensive Player Impact: A dashboard that shows each player’s defensive impact metrics based on the offensive player’s position, height, and weight.
6. Offensive Player Impact: A dashboard that shows each player’s offensive impact metrics based on the defensive player’s position, height, and weight.
7. Detailed Matchup Data: A dashboard that displays the entirety of the matchup data.

Evaluation

Our testbed is a series of Python and R scripts that perform our exploration and analysis and generate the final data files for our Tableau visualizations.

The questions we are answering are as follows:

- Can we establish an innovative method of grouping players based on defensive matchups?
- Are we able to develop a new set of insightful and useful metrics to analyze a player’s defensive impact?
- Can we effectively visualize the matchup network to provide opportunity for any end-user to explore player groupings?
- Are there optimal defensive matchups, based on various stratifications such as positional and height groupings, to provide recommendations for coaching staffs to consider when assigning defensive matchups?
- Can we provide an interactive and easy-to-use dashboard tool that presents the matchup and impact data, and allows coaching staff to game plan for the best matchups in their upcoming games?
- Do these dashboards also provide deep insight into defensive matchup tendencies and performance for interested NBA fans, such as fantasy players and bettors?

Experiments & Observations

As noted above, we used PCA and linear regression to identify the factors that most affect our three metrics. Overall, we discovered that the combination of position matchups and height and weight differences best accounted for the matchup effectiveness (or not) of players and their archetypes.

Using the last four seasons of matchup data, we ran linear regression experiments on the full dataset, to determine which matchups and height/weight combinations were the best and the worst. We experimented with using the individual positions and height/weight categories as factors, along with different combinations of them. We also experimented with splitting and subsetting the data by the available factors, and by only using a smaller sample of full dataset.

We found that categories which combined the player positions and height/weight categories into a single category performed best in predicting player performance. For example, a single category has four pieces of data: The offensive and defensive player positions, and each player's relative height and weight within their position category. One such category would be PG-PG-Short-Short-Light-Light (a short and lightweight point guard matched against another short and lightweight point guard).

Our second key observation is that we must weigh the individual metrics (player/team points and TS%) by the number of matchup minutes recorded. We found that when we used the absolute metric values, matchups that only had a few minutes across the four seasons had an outsize effect on the models, much more so than their effect in the actual game. We then used the number of minutes for each matchup as a weighting factor, and this gave us a much better model, with highly significant p-values and low residual standard errors. We found that some matchups are significantly effective (on both offense and defense), and those are the ones that we would, in general, highlight to coaches as ones they should seek out in their game planning. For example, a matchup with a tall and heavy center on any Guard or Forward position shows significant increases across all three metrics, with p-values in the range of 1.00e-05 and smaller.

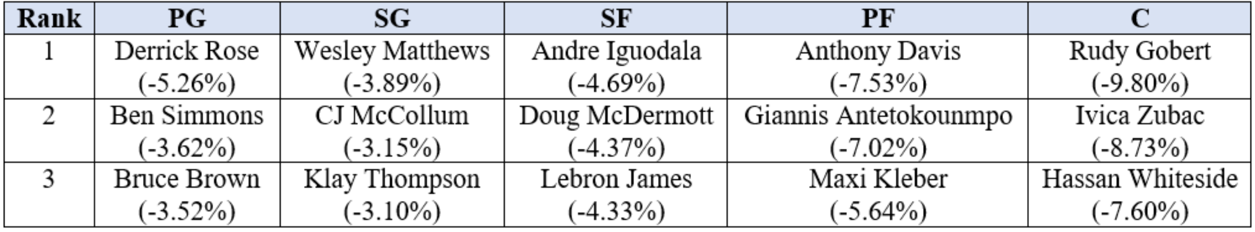
Finally, when aggregating across position groups in addition to weighting by the number of minutes, we discovered that using only matchups with the number of minutes greater than the median across the four seasons was most effective. The median number of matchup minutes at the matchup level is 38, so subsetting the data to only include those rows provided the best model. This granularity is player versus player, aggregated across the four seasons.

We further evaluated our metrics' effectiveness by comparing what it determines are the best defensive players, with those whom the NBA itself deems the best. This NBA standard is the NBA All-Defensive First and Second teams, which are the top 10 players (5 per team) whom the NBA deems as the best on defense. Across the four seasons, our metrics align completely with the NBA's evaluation. We removed seldom-occurring matchups from the evaluation, to prevent skewing the data (those that occurred for less than 150 minutes across the four seasons, or about 5% of the total time possible).

Our metrics show that these players perform, at a minimum, in the top 32% of the total NBA player population, as their metrics are generally at least one standard deviation to the left of the norm. In about half of the cases, the cumulative player metrics are 1.5-2 standard deviations from the norm, putting them in the top 5-20% of the NBA, for defensive performance.

Filtering by the most-often occurring matchups for each player, they are typically holding their opponent to 30-60% fewer points scored individually, and 40-70% lower team scoring. So, their effectiveness is not only in shutting down their individual matchup opponent, but also in contributing to the overall team defense. The mean for each of these measures is 0%, so it is easy to see how good these defenders are. Our metrics clearly align with the NBA's subjective player evaluation, providing quantifiable data points for coaches, players, and management.

Finally, we evaluated our results by using the ‘Defensive Player Impact’ dashboard to identify the top three ranked defenders by position, using the TS% Differential Metric and a minutes minimum of 1500 as shown in the following table:



Rank	PG	SG	SF	PF	C
1	Derrick Rose (-5.26%)	Wesley Matthews (-3.89%)	Andre Iguodala (-4.69%)	Anthony Davis (-7.53%)	Rudy Gobert (-9.80%)
2	Ben Simmons (-3.62%)	CJ McCollum (-3.15%)	Doug McDermott (-4.37%)	Giannis Antetokounmpo (-7.02%)	Ivica Zubac (-8.73%)
3	Bruce Brown (-3.52%)	Klay Thompson (-3.10%)	Lebron James (-4.33%)	Maxi Kleber (-5.64%)	Hassan Whiteside (-7.60%)

Based on these rankings, we identified the following:

- Among these 15 players, six of them have finished within the top-5 voting for defensive player of the year at least once with finish position in parenthesis: Simmons (2nd, 4th), Iguodala (5th), James (2nd x2, 4th x2), Davis (2nd, 3rd, 4th), Antetokounmpo (1st, 2nd, 5th), Gobert (1st x3, 2nd, 3rd, 5th)
- Of the remaining 9, one has received All-Defensive Honors (top-10 defenders per season): Klay Thompson
- Of the remaining 8, per a qualitative assessment there are four players that are either historically strong defenders or currently upcoming and likely underrated defenders (best DRPM ranking during 2017-21 period): Bruce Brown (8), Wesley Matthews (2), Maxi Kleber (7), Ivica Zubac (6)
- The final four players are surprising in nature. These players are either known for their offense creation (Rose & McCollum), shooting ability (McDermott) or are typically backups due to generally poor performance (Whiteside). In all four cases, DRPM is low on these players.

In summary, we use these results from two perspectives. First, to evaluate our own methodology, they provide additional confidence that our results are suggesting players that are agreed-upon by all parties involved as some of the best defensive players in the NBA. Secondly, it provides results worth further investigation. As a coach, I would consider playing Brown or Zubac more minutes as they are likely underrated defenders. As a team president, I might investigate if signing or trading for Whiteside or McDermott is an efficient and strategic move, as they might not be fully valued by the rest of the NBA.

Conclusion & Discussion

Our analysis provides an optimal and easy-to-use method for coaches to game plan for defensive and offensive matchups. It is based on a rigorous statistical analysis that provides confidence in the results and recommendations, but it is also simple to use and understand.

We foresee coaches using our visualization tool and metrics in their game planning to identify the optimal matchups on both offense and defense, team presidents and general managers using it to identify underrated players, and bettors to find inefficiencies in the NBA gambling market. Altogether, it provides a new and optimal methodology to improve performance and results across various domains.

There are many opportunities for additional efforts based on our completed analysis. For example, a lineup optimization program would greatly enhance our interactive dashboards. Such a program would take a team and its opponent’s players as input and would output the optimal player matchups. Coaches could use this program to optimize their player rotations and defensive assignments, to maximize the likelihood of winning the game. Conversely, a coach could utilize this function to determine which matchups they should try to avoid when developing their offense strategy. Our statistical model could also be altered for prediction purposes. If an organization is considering signing or drafting a new player, the player’s characteristics (position, height, etc.) could be used to predict how likely it is for that player to perform well against another type of player. This would be extremely useful since this player would not have any previous statistics against other NBA players.

Distribution of Team Member Effort: All team members have contributed a similar amount of effort.

References

1. Govan, A. Y., Langville, A. N., & Meyer, C. D. (2009). Offense-Defense Approach to Ranking Team Sports. Journal of Quantitative Analysis in Sports, 5(1). https://doi.org/10.2202/1559-0410.1151. Uses a ranking model to compute team ratings and the likelihood of winning games. Useful in that it goes to winning and losing, but it may be too high-level for our analysis (does not drill into individual player performance).
2. Piette, J., Anand, S., & Zhang, K. (2010). Scoring and Shooting Abilities of NBA Players. Journal of Quantitative Analysis in Sports, 6(1). https://doi.org/10.2202/1559-0410.1194. Uses offensive shooting statistics to provide two new metrics for player evaluation. Useful in providing ideas about how to apply a baseline and how to reduce sampling bias and outliers. It is focused on offensive performance, not defense, and the data is old (from 2004-2007).
3. Piette, J., Pham, L., & Anand, A. (2011, March). Evaluating Basketball Player Performance via Statistical Network Modeling. 1–11. https://www.scribd.com/document/320915007/Evaluating-Basketball-Player-Performance-via-Statistical-Network-Modeling-pdf. Examines interactions between players to better understand player performance. Focus is on the plus/minus metric. Is a bit higher-level and does not specifically focus on defensive performance.
4. Ahmadalinezhad, Mahboubeh, Masoud Makrehchi, and Neil Seward. "Basketball lineup performance prediction using network analysis." Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining. 2019. Builds ML / Network Analysis model to predict line up performance. Useful as a methodology framework for one analysis. Does not focus specifically on defensive performance.
5. Fewell, Jennifer H., et al. "Basketball teams as strategic networks." PloS one 7.11 (2012): e47445. Considers players as nodes and ball movements as edges – does various network analysis. Useful as methodology framework for analysis. Like above, does not specifically focus on defensive performance.
6. Koster, Jeremy, and Brandy Aven. "The effects of individual status and group performance on network ties among teammates in the National Basketball Association." PloS one 13.4 (2018): e0196013. Considers a network among NBA players but edges are not related to game time but related to social media. Finds that high status players on bad teams less likely to follow teammates than high status players on good teams. Focus on social media versus on court performance is a limitation.
7. Bhandari, I., Colet, E., Parker, J. et al. Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. Data Mining and Knowledge Discovery 1, 121–125 (1997). https://doi.org/10.1023/A:1009782106822. Discusses a data-mining application used by NBA teams for data analysis. Older article, but very good for understanding how to pre-process our data.
8. Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2014, February). Pointwise: Predicting points and valuing decisions in real time with NBA optical tracking data. In *Proceedings of the 8th MIT Sloan Sports Analytics Conference, Boston, MA, USA* (Vol. 28, p. 3). Uses player position tracking data to formulate Expected Possession Value metric. Very detailed analysis but provides a good reference framework for our defensive analysis.
9. Franks, A., Miller, A., Bornn, L., & Goldsberry, K. (2015). Characterizing the spatial structure of defensive skill in professional basketball. *The Annals of Applied Statistics*, *9*(1), 94-121. Excellent defensive-focused analysis. Good idea generation and follow on for our project.
10. Franks, A., Miller, A., Bornn, L., & Goldsberry, K. (2015, February). Counterpoints: Advanced defensive metrics for NBA basketball. In *9th Annual MIT Sloan Sports Analytics Conference, Boston, MA*. Another excellent defensive article from the same team as #9. Provides who guarding whom matrix, like our methodology.

11. Wang, J., Fox, I., Skaza, J., Linck, N., Singh, S., & Wiens, J. (2018). The advantage of doubling: a deep reinforcement learning approach to studying the double team in the NBA. *arXiv preprint arXiv:1803.02940*. Good article focused on defensive performance, but we are not planning to use double-teaming specifically in our analysis.
12. Terner, Z., & Franks, A. (2021). Modeling player and team performance in basketball. Annual Review of Statistics and Its Application, 8, 1-23. Good overall article on the current state of NBA analytics. A bit broad for our specific use but does provide some insight into current defensive analytics.
13. Nagarajan, R., & Li, L. (2017, November). Optimizing NBA player selection strategies based on salary and statistics analysis. In 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech) (pp. 1076-1083). IEEE. Provides two metrics to evaluate player performance, selection, and team win rate. Formulation of the Player Defense Rating provides some insights for the project, but not to the level of detail we are planning.
14. Lamas, L., Santana, F., Heiner, M., Ugrinowitsch, C., & Fellingham, G. (2015). Modeling the offensive-defensive interaction and resulting outcomes in basketball. PloS one, 10(12), e0144435. Good article for understanding NBA play patterns and outcomes. A bit general for our use, and not specific to matchups and individual players.
15. D'Amour, A., Cervone, D., Bornn, L., & Goldsberry, K. (2015). Move or die: How ball movement creates open shots in the NBA. In Sloan Sports Analytics Conference. Detailed analysis of how ball movement positively affects possession outcome and points scored. Good for understanding play patterns and overall game play, but not specific to players themselves.
16. Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA game result prediction using feature analysis and machine learning. Annals of Data Science, 6(1), 103-116. Good article for identifying important features to help in prediction game outcomes. A bit high-level for our use, and the features identified are not necessarily useful in our analysis.
17. Mikołajec, K., Maszczyk, A., & Zając, T. (2013). Game indicators determining sports performance in the NBA. Journal of human kinetics, 37, 145. Good overall analysis of NBA game play. Not sufficiently detailed at the player level for us to directly use.
18. Fearnhead, P., & Taylor, B. M. (2011). On estimating the ability of nba players. Journal of Quantitative analysis in sports, 7(3). Uses game statistics to evaluate player performance and team value. Excellent article for our use, although the data is older (2008-2009 season).
19. Luke (no last name). "What is the Average Career Length of an NBA Player?". https://dunkorthree.com/nba-player-career-length/, October 11, 2021.