ATTITUDE OF THE NBA COMMUNITY TOWARDS ANALYTICS:

REVOLUTION OR RUIN?

by

JACK WENTZIEN

A THESIS

Presented to the School of Computer and Data Science
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Science

May 2024

An Abstract of the Thesis of

Jack Wentzien for the degree of Bachelor of Science
in the School of Computer and Data Sciences to be taken June 2024

Title: Attitude of the NBA Community Towards Analytics: Revolution or Ruin?

Approved: *Dr. Eren Çil*
Primary Thesis Advisor

Since the turn of the century, the use of analytics has become more prominent in the National Basketball Association (NBA). Teams are using data-driven approaches to construct rosters and develop strategies, which have proven useful in turning previously underwhelming teams into championship contenders. Today, all NBA teams have designated analytics departments that use advanced metrics to evaluate the success of their players and teams. Media coverage of the NBA reflects this change; more than ever player evaluations in the news are bolstered with evidence such as shooting percentage, adjusted plus minus, or other advanced metrics. This analytics overhaul in the news has made it easier than ever for NBA fans to interact with analytics and has caused many to take a "Moneyball" approach to consuming NBA basketball, which relies on using numerical data to evaluate players and teams.

However, NBA affiliates are resistant to the league's rapid integration of analytics into team operations, claiming that analytics are destroying the integrity of the game by reducing players to numbers. This could be problematic for players, as it creates a larger gap between how players evaluate their own performances, versus how organizations evaluate them, which makes it more difficult for players to meet team expectations. This, coupled with other ethical concerns, is why some NBA affiliates argue that the reliance on analytics furthers the divide between players and management, which in turn is harmful to the players and threatens their livelihoods.

This thesis seeks to evaluate the discourse surrounding analytics in the NBA, specifically looking at reactions by NBA fans to media posts about basketball analytics. Gauging the reaction of an entire community to one specific issue is difficult when relying solely on literature review, so this project will make use of Twitter scraping and sentiment analysis to quantify the nature of the discourse surrounding NBA analytics. Twitter is a representative sample of the NBA community because fans, announcers, and other NBA affiliates can post thoughts, as well as interact with other posts. The use of sentiment analysis to quantify tweets has been performed before and is useful in getting the tone of a discourse efficiently. Once this evaluation is complete, this project will conclude with re-examining the current implementations of analytics through the lens of the NBA community's discourse and provide suggestions on how the NBA could change their current implementations of analytics to better suit players, fans, and league operations personnel alike.

Acknowledgements

First, I would like to thank the University of Oregon for an unforgettable 4 years. I have had the most incredible learning opportunities, and every day I have felt welcomed, accepted, and appreciated as a student. I am proud to forever call myself a Duck. Next, a huge thanks to my primary advisor, Dr. Eren Çil, who has supported me throughout the course of this project. Despite me never having taken one of his classes and being from a different major entirely, Dr. Çil graciously agreed to advise me on this project, and has been extremely helpful, responsive, and encouraging every step of the way. I would also like to thank my CHC advisor, Dr. Lindsay Hinkle, who has shown me nothing but support and kindness over the course of this project.

Finally, I would like to thank my family. My parents are the light of my life and have pushed me to keep moving forward and apply my best self to everything I do. My family has supported me so much these last four years, and I feel blessed every day to have them at my side.

Table of Contents

List of Figures

1 Literature Review

Background of Analytics in the NBA

Over the last decade, analytics have taken a more prominent role in the National Basketball Association (NBA). Operations personnel within the association are using data-driven decision making to construct rosters, develop strategy, and monitor or maintain player health. This integration has not come without controversy, as many individuals within and affiliated with the association argue that the integrity of the game of basketball is being ruined because of analytics. Despite this pushback, analytics have proven useful in turning previously underwhelming teams into championship contenders (Abbas 2019). Analytics are giving NBA teams tremendous insight into how to generate a winning team; however, this is not without its drawbacks. This review will begin by examining specific data collection techniques implemented by the NBA, as well as the advantages teams league wide are gaining because of data analytics and data-driven decision making, and it will conclude with the ethical concerns arising from this rapid integration of analytics into standard team operations.

Massive amounts of data are collected every game, and it fits into two categories: continuous data and discrete data (D'Amour 2016). The most useful example of continuous data is positional player tracking. In all NBA arenas, there are several cameras mounted in fixed positions above and beside the court. These cameras record a video feed of the game, and then locations of the players and the ball are extracted from this feed at a rate of 25 times per second (Macdonald 2020). Through a process called computer vision, these positions are converted into XYZ coordinates, which are readily available for analysis. Machine learning has made it easy to decipher explicit positions and tactics from these coordinates (Min-Chun 2011) such as screen and rolls, offensive patterns like triangles, and drives. Discrete data is far more accessible, and it is the data that most fans of basketball are familiar with such as shot attempts, steals, rebounds, etc.

These two types of data can be cross referenced to draw out an unbelievable amount of insight. The Toronto Raptors have developed a concept called a “ghost team”, which is essentially the optimal defensive positioning of all players based on a metric called the Expected Possession Value (EPV), and the type of defensive scheme they are running (Macdonald 2020). The EPV is an average of every possible decision the ball handler can make weighted with the probability that they will make that decision. The ghost team is used by the Raptors to develop extremely refined defensive schemes. Often it is difficult for players on the floor to make long rotations or switch who they are guarding on the fly, so letting the ghost team make these adjustments and then retroactively coaching the players allows for a predictive defense that is much more effective.

Some teams even collect biometric data. The players wear monitoring devices that measure heart rate, start-stop velocity, and other physiological responses. These devices collect over 1000 points of data per second, which are directly transferred to the fitness trainers working with the teams (Berger 2015). Pre-set danger thresholds are determined before training, and if the players enter these thresholds, then it is likely that they will be recommended to sit out the next game (Berger, 2015). As of now, this wearable technology is only used during practices, but some advocates of biometric monitoring propose collecting data from athletes off the court, for instance sleeping patterns or dietary choices.

This extensive collection of data by NBA teams does not come without its ethical concerns. Biometric data collection poses several privacy issues, especially when used while athletes are not in practice or during games. Even if this data was collected during games, there are questions as to how much a team must adhere to the conclusions gathered from it. For example, imagine the Lakers were playing in the tiebreaker of the NBA finals and their readings indicate that LeBron James has entered the danger threshold in the last few minutes of the game. Despite this increased risk for injury, would the coaches be receptive to taking him out? Would LeBron himself even accept that? (Berger 2015) There is a question of how much to balance risk of injury with a player's ability and desire to push past adversity in close games, and if strict regulations are imposed regarding player safety based on this biometric data, then we could see pushback from coaches and players alike. Another issue that arises from this data collection and interpretation is resting star players during games. More than ever before, NBA teams are resting stars during games they deem less important to their overall record and playoff standings. Teams are giving players time off and even paying them less when they are deemed "high risk" for injury (Kopf 2017). Fans are becoming frustrated when they pay for expensive tickets just to find out moments before tip-off that their favorite players are resting. Despite this, the league benefits from star players having prolonged careers, and bench players are becoming more skilled because of this extra playing time.

Even standard positional player tracking has ethical quandaries and is often subject to criticism from those who have been affiliated with the NBA for a long time. Some commentators argue that analytics are ushering in a "repetitive" form of basketball (Abbas 2019). Analytics directly contributes to the rise of the three-point shot, a shot that was historically frowned upon. However, finding efficient three-point shooters has proven extremely effective in transforming a mediocre team into a championship contender. Coaches who use this strategy may fall into the repetitive form of offense mentioned above, but there is no denying its effectiveness. Critics of this strategy argue that the integrity of the game itself is being destroyed and suggest that the three-point line be moved back to discourage it (Abbas 2019). This phenomenon also bleeds into the question of how much influence the front office should have in determining the style in which a team plays. Coaches and front offices alike are much more inclined to allow players to find their own strategies for success if they remain mathematically sound (Lowe 2013). Finding this synergy between coaches, analysts, and players has proven challenging for many teams when it comes to being successful.

Another problem that arises from this is the ownership of the data. The NBA has the rights to all the data it collects, but what about the players? It is them who are generating this data, after all. Players are usually on the outside when it comes to analyzing their own data, and often the front office will make decisions based off a player's data without them being involved. This proves troublesome for players being scouted or drafted, as they often have no idea what conclusions are being drawn from their game data, and seldom have a chance to make a case for themselves.

An unforeseen ethical challenge that has come from recording games and practices is media leaks. During the 2022 NBA season, TMZ leaked a video of two Golden State Warriors getting into an altercation during practice (Thomas 2022). This caused a storm of drama and bad press to befall the team in the middle of their season, and the two players involved were interviewed relentlessly for details about their "beef". While it is frowned upon for media to pay operations personnel within the NBA for leaked footage, it is not an uncommon occurrence. It is also difficult to hold individuals within the league accountable, and these leaks often come at the expense of the players (Thomas 2022).

As analytics become rapidly integrated into the NBA, and subsequently the game of basketball in its entirety, there are just as many ethical concerns as tactical benefits. What is the next step? Even though all teams are on board with fully using analytics, it is still a polar topic for some in the the NBA community, and many retired players see the integrity of the game being corrupted. This thesis serves to determine whether discourse within the NBA community is being positively or negatively affected by discussions concerning analytics.

Twitter Discourse

To get an accurate assessment of the NBA community's attitude towards analytics, there are two reasonable approaches. First, extensive literature review: NBA related media is constantly uploaded digitally, and is up to date with the latest opinions, trends, rumors, and other NBA-related topics. This media is easily accessible, and comes from a variety of authors including journalists, former players, coaches, or analysts, which makes it useful for gauging opinions of individuals close to the NBA. However, this approach is time consuming, and would exclude the largest portion of the NBA community: the fans. Fans are important to the success of the league, and their opinions are influential in the league's decision making.

So how can the opinions of the entire community, including the fans, be sampled accurately? Fans don't write professional NBA related journalism, so how can we include their voice? This is where the second approach comes in: Twitter scraping. Twitter is a social media platform where fans, former players, coaches, and other NBA related personnel post their voices, which makes it an excellent source for assessing the discourse within the NBA community about a variety of issues. Other studies have used data from Twitter to draw conclusions about real-world issues. For example, Kusumasari & Prabowo used Twitter data to identify different ways that it is used in disaster response, claiming that Twitter had "considerable potential as a communication channel" (Kusumasari & Prabowo, 2020).

2 Data Collection

Scraping is a data collection method in which software takes data directly from online platforms. The scraper will receive an input URL, from which it will load all the website's HTML code and return data based on the user's given parameters. Users can hyper-parameterize the scraping bots to return specific data from a website's source code, allowing for the efficient collection of online data. The NBA has a strong online media presence, especially on Twitter. League affiliates and fans alike tweet news, opinions, and reactions to a variety of NBA media sources, making Twitter a suitable population for this project.

This data was collected with a scraper, designed in Python, specifically built to gather tweets and replies when given a Twitter username. All data collection as well as statistical analyses were conducted in Jupyter Labs, a web-based Python environment. For this project, a list of 30 NBA basketball focused Twitter accounts were collected to input into the scraper. From each account, 100 tweets along with the first 100 replies to each Tweet were collected and processed using NLTK's sentiment analysis model, which will be covered in detail later. Sentiment analysis is a technique in which a program can quantify the emotional tone of a piece of text, which is very useful for short blocks of text such as tweets. The use of sentiment analysis allows tones and trends in this Twitter discourse to be quantified and analyzed efficiently, smoothing the transition from qualitative data to quantitative data.

2.1 Tweet and Reply Collection

The tweet and reply collection algorithm uses a series of functions to collect, filter, and store tweets and their associated replies. In Python, functions are blocks of code that perform a specified function when called upon, and take inputs known as parameters. *Figure 1* illustrates the series of functions that are used to collect tweets for a given user. The output is a Python dataframe, which is a table-like data structure. These dataframes are easily exportable into .csv (comma separated value) files, which can then be accessed and combined for later analysis.

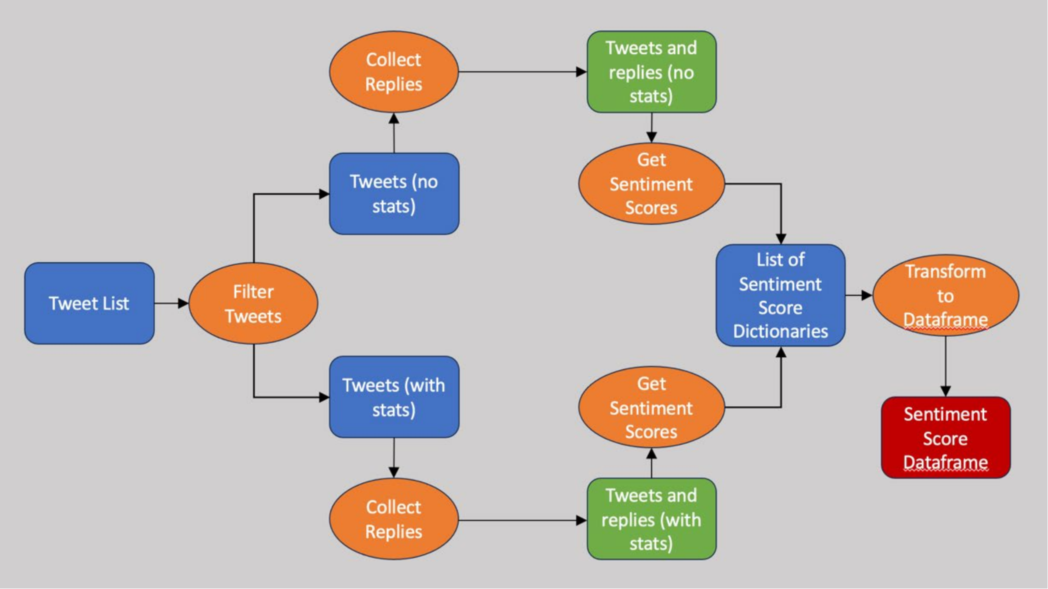


Figure 1: **Tweet and Reply Collection into Dataframe Algorithm**

This figure represents the series of functions used to collect, filter, and store tweets from a given user. Blue cells represent lists, orange cells are functions, green cells are dictionaries, and red cells are dataframes. The arrows are the chain of inputs, so for all data structures (lists, dictionaries, and dataframes), an arrow coming out of them means they are the input for the next cell. When arrows are coming out of a function cell, it means that the function produced the following data structure at the end of the arrow.

Tweet Collection

The scraper is designed to collect tweets when given a username. Each input username from the list of 30 was hand selected due to their affiliation with the NBA, high follower count, and tendency to post frequently about basketball analytics. However, there are two factors to consider that may lead to bias in the data due to the manual selection of these users:

- Interactions with followers:

    - Some accounts interact more with their followers, replying to their comments or retweeting their posts. This may incentivize regular followers to reply more provocatively to get a reaction from the account, thus promoting more false negative replies in the data.
- League affiliation status:
    - While some of the users from the input list are directly affiliated with the NBA, others are basketball buffs with high follower counts that post regularly about basketball analytics. League affiliated personnel may generate more replies than non-league personnel.

Each username in the list was inputted into the scraping function 'fetch_tweets', which returned a list of the most recent 100 tweets for each user since the beginning of the most recent NBA season (October 24, 2023). This was accomplished using the Tweepy package, which is an open-source library that interacts with the latest Twitter Application Program Interface (API). An API is essentially the portal that programmers need to use to have their own program interact with a given application, in this case Twitter. However, despite Tweepy being free and open source, the Twitter API is not. It limits the number of tweets that can be scraped every month to 10,000 (including replies), which severely restricts the amount of data that can be collected for this project. This should not be an issue because, despite this limitation, the dataset is still sufficiently large. Each tweet still retains all its metadata, most importantly the tweet ID numbers. Each tweet has a unique ID number associated with it, which will be used later in gathering the replies for a specific tweet.

Tweet Sorting

To perform analysis, the collected tweets need to be flagged into two categories: tweets about analytics, and tweets *not* about analytics. This was accomplished processing the text with a function to check if the text of each tweet contained any of the following characters:

- %: Percentage, commonly used for field goal percentage or true shooting percentages.
- +/-: Plus-minus, a basketball statistic that calculates a team's net points while a specific player is on the floor.
- '3PM', '3PA': Three pointers made, and three pointers attempted.
- '2PM', '2PA': Two pointers made, and two pointers attempted.
- 'pts', 'reb', 'ast', 'stl': Points, rebounds, assists, and steals.
- 'Efficiency': A composite statistic that is a measure of the four statistics above ('pts', 'reb', 'ast', 'stl').
- r'\d+-for-\d+': This is a regex pattern (a pattern that helps isolate certain words with changing numbers) that looks for "number-for-number". Shooting performance for players is often denoted as "7-for-15", but with varying numbers, so a regex pattern was needed to account for those variations.

If these patterns were detected in a tweet, that tweet was then stored in an array tweet with analytics mentioned. If not, the tweet was stored in a separate array for tweets with no analytics mentioned.

Reply Collection

Tweets often generate more replies based on how popular or engaging the original tweet is. Thus, for each tweet collected, another scraping function collects the first 100 replies. If a tweet has fewer than 5 replies, that tweet is deleted from the dataset. One caveat of this is that neutral tweets may be underrepresented due to garnering less replies.

Replies were collected for all tweets. These replies are then stored in a Python dictionary data structure, which uses key and value pairs. In this case the key is the text from the tweet, and the value is the text of all replies to that tweet. The text from the original tweet is stored as a string (collection of characters), and the texts from all the replies are stored in a list of strings. This is accomplished using two functions: ‘fetch_replies’ and ‘collect_replies’. The first function accepts a tweet’s id as an input, then returns a list of all the texts of the replies to that tweet. The second function loops through a list of tweets for a given user, then makes use of the first function to collect the replies for each tweet. Once the replies are collected, the function then assigns them to the dictionary, then once all the tweets have been looped through, it returns the dictionary.

2.2 Sentiment Analysis

For this project, sentiment analysis was conducted using the Natural Language Toolkit (NLTK) sentiment analysis model. NLTK is an open-source Python library that specializes in working with human language data. For each tweet, the sentiment analysis performed uses NLTK’s algorithm to classify samples of words into positive, negative, and neutral categories. This is accomplished using the VADER lexicon, which can score each word based on the words and punctuation in the rest of the sentence. In the context of this project, the specifics of this algorithm are opaque and will be used as a ‘black box’. NLTK is a flexible, powerful, and reputable language library, so despite the complexity of its model, it is a suitable choice for sentiment analysis in this context.

Text Pre-Processing

Before sentiment analysis can be performed, the text from each tweet and reply needs to be cleaned, which means removing two patterns: mentions and links. Mentions are when users tag other users using the '@' symbol and are often use to alert friends to posts that the user thinks they will enjoy (or sometimes despise). Links, or URLs, are present in lots of tweets and serve to direct people to websites relevant to the tweet. All links begin with the same characters: 'http://'. Since both mentions and links have unique starting characters, a regex pattern can be used again to isolate all links and mentions in a text, then replace them with other characters.

The reason that mentions and links need to be filtered is that they can interfere with the accuracy of sentiment analysis. Sometimes, users respond to tweets with a series of mentions followed by no text, which means they are simply alerting other users to the tweet and not actually expressing a sentiment. This should be reflected in the data, however replacing all mentions and links with nothing (empty strings) in a tweet such as that will result in sentiment scores of 0 in all categories, which will pollute the dataset. This is why replacement words need to be inserted instead of empty strings: mentions will be replaced with the word 'user', and links will be replaced with the word 'link'. These words return neutral sentiment scores on their own, so they are suitable replacement words that will not skew the sentiment analysis in a positive or negative direction. Some might argue that this skews the data towards neutrality, however if a post is consisted of exclusively mentions and links, this should be classified as a fully neutral tweet because there is no way to extract any sort of sentiment or discourse from text like that.

Polarity Scores

After the text for each tweet and its associated list of replies is processed, sentiment scores are calculated for each piece of text using NLTK's 'polarity_scores' function. This function returns a dictionary with 4 scores: positive, neutral, negative, and compound score. The first three metrics are proportions of the total inputted text that are classified as positive, neutral, and negative, in other words the sum of these three scores should add up to one. For example, the phrase “The Portland Trail Blazers are the best team in the NBA!” returns the following output when inputted into the ‘polarity_scores’ function: {'neg': 0.0, 'neu': 0.69, 'pos': 0.31, 'compound': 0.6696}. The first three scores all sum up to 1, however the compound score is not included. The compound score is essentially a sum of the sentiment scores of each individual word in the text which is then normalized to be between -1 (most negative) and 1 (most positive). This is accomplished using the VADER lexicon, which can calculate the sentiment score of a word based on the words and punctuation that come before and after it. NLTK uses the VADER lexicon in its built-in functions (such as ‘polarity_scores’) to calculate sentiment scores.

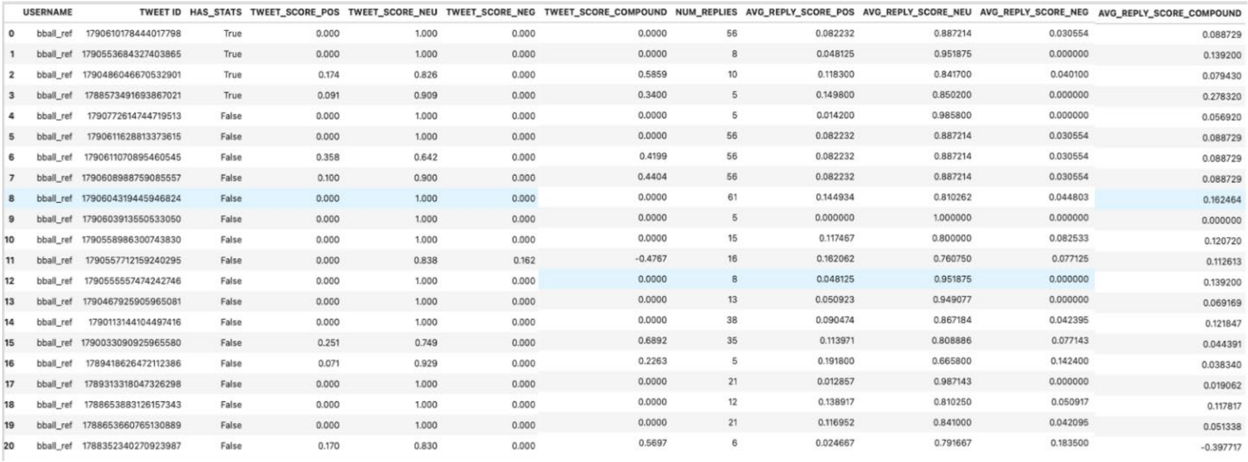
2.3 Data Storage

The data for each twitter user from the original list of NBA affiliate users is ultimately stored in a Python dataframe with the following columns:

- 'USERNAME': username of the account that posted the tweet.
- 'TWEET ID': the specific tweet’s id number.
- 'HAS_STATS': True or False on whether the tweet is about NBA stats.
- 'TWEET_SCORE_POS': Proportion of tweet text with positive score.
- 'TWEET_SCORE_NEU': Proportion of tweet text with neutral score.
- 'TWEET_SCORE_NEG': Proportion of tweet text with negative score.
- 'TWEET_SCORE_COMPOUND': Compound tweet score.
- 'NUM_REPLIES': Number of replies.

- 'AVG_REPLY_SCORE_POS': Average proportion of each reply with positive score.
- 'AVG_REPLY_SCORE_NEU': Average proportion of each reply with neutral score.
- 'AVG_REPLY_SCORE_NEG': Average proportion of each reply with negative score.
- 'AVG_REPLY_SCORE_COMPOUND': Average compound score.

Python dictionaries are easily convertible to dataframes, which is accomplished using the 'Transform to Dataframe' function in *Figure 1*. This function takes input as a list of dictionaries, and for each one the keys are the column names listed above, and the values are the data corresponding to each column for a specific tweet. The output is a python dataframe, which can easily be exported to a .csv file. *Figure 2* is an example of the structure of such a dataframe for one user, in this case '@bball_ref'.



	USERNAME	TWEET ID	HAS_STATS	TWEET_SCORE_POS	TWEET_SCORE_NEU	TWEET_SCORE_NEG	TWEET_SCORE_COMPOUND	NUM_REPLIES	AVG_REPLY_SCORE_POS	AVG_REPLY_SCORE_NEU	AVG_REPLY_SCORE_NEG	AVG_REPLY_SCORE_COMPOUND
0	bball_ref	1790610178444017798	True	0.000	1.000	0.000	0.0000	56	0.082232	0.887214	0.030554	0.088729
1	bball_ref	1790553684327403865	True	0.000	1.000	0.000	0.0000	8	0.048125	0.951875	0.000000	0.139200
2	bball_ref	1790486046670532901	True	0.174	0.826	0.000	0.5859	10	0.118300	0.841700	0.040100	0.079430
3	bball_ref	1788573491693867021	True	0.091	0.909	0.000	0.3400	5	0.149800	0.850200	0.000000	0.278320
4	bball_ref	1790772614744719513	False	0.000	1.000	0.000	0.0000	5	0.014200	0.985800	0.000000	0.056920
5	bball_ref	1790611628813373615	False	0.000	1.000	0.000	0.0000	56	0.082232	0.887214	0.030554	0.088729
6	bball_ref	1790611070895460545	False	0.358	0.642	0.000	0.4199	56	0.082232	0.887214	0.030554	0.088729
7	bball_ref	1790608988759085557	False	0.100	0.900	0.000	0.4404	56	0.082232	0.887214	0.030554	0.088729
8	bball_ref	1790604319445946824	False	0.000	1.000	0.000	0.0000	61	0.144934	0.810262	0.044803	0.162464
9	bball_ref	1790603913550533050	False	0.000	1.000	0.000	0.0000	5	0.000000	1.000000	0.000000	0.000000
10	bball_ref	1790558986300743830	False	0.000	1.000	0.000	0.0000	15	0.117467	0.800000	0.082533	0.120720
11	bball_ref	1790557712159240295	False	0.000	0.838	0.162	-0.4767	16	0.162062	0.760750	0.077125	0.112613
12	bball_ref	1790555557474242746	False	0.000	1.000	0.000	0.0000	8	0.048125	0.951875	0.000000	0.139200
13	bball_ref	1790467925905965081	False	0.000	1.000	0.000	0.0000	13	0.050923	0.949077	0.000000	0.069169
14	bball_ref	1790113144104497416	False	0.000	1.000	0.000	0.0000	38	0.090474	0.867184	0.042395	0.121847
15	bball_ref	1790033090925965580	False	0.251	0.749	0.000	0.6892	35	0.113971	0.808886	0.077143	0.044391
16	bball_ref	1789418626472112386	False	0.071	0.929	0.000	0.2263	5	0.191800	0.665800	0.142400	0.038340
17	bball_ref	1789313318047326298	False	0.000	1.000	0.000	0.0000	21	0.012857	0.987143	0.000000	0.019062
18	bball_ref	1788653883126157343	False	0.000	1.000	0.000	0.0000	12	0.138917	0.810250	0.050917	0.117817
19	bball_ref	1788653660765130889	False	0.000	1.000	0.000	0.0000	21	0.116952	0.841000	0.042095	0.051338
20	bball_ref	1788352340270923987	False	0.170	0.830	0.000	0.5697	6	0.024667	0.791667	0.183500	-0.397717

Figure 2: **Example Dataframe for User 'bball_ref"**

This figure is a subset of the final dataframe, but exclusively shows tweets for the user 'bball_ref'.

Once a dataframe has been created for each user on the list, it is then exported into a .csv file. Another script then takes those files and combines them into a cumulative dataframe that contains all the users. This is the dataframe that will be used to conduct the analysis on this data.

3 Analysis

The analysis of this final dataset will begin with some preliminary summary statistics about the dataset. Then, two scatterplots showing the relationship between tweet scores and reply sentiment scores for tweets with analytics mentioned, and tweets without analytics mentioned. Linear regression will be performed on the data from the scatterplots, and the subsequently calculated best fit lines will be plotted over them. This will be followed by histograms exploring the distribution of tweet vs. reply scores with Kernel Density Estimation (KDE) on the same plots.

To assess whether the mean scores of tweets with stats are significantly different from tweets without stats, a two-sample t-test will be conducted on the tweet compound scores (stats vs. no stats) and the average compound reply scores (stats vs. no stats). Finally, 3 linear regression models will be tested to see which features most impact the average sentiment scores of the replies to a tweet. These features include the tweet itself having stats, the sentiment score of the tweet, and the number of replies. The information from the tests and visualizations should be enough to determine if having stats is a significant factor in determining how a tweet will be received by users.

3.1 Data Summary

The final dataset consists of n = 589 total tweets scraped, with 22086 total replies. Out of those 589, 22.07% of them contained information about analytics and 77.93% of them were not. There is an imbalance in the two classes of tweets, which will be relevant in later analysis. For tweets with stats, the average sentiment score was minimally negative (μ = -0.01803, σ^2 = 0.22043, n = 130), but with a high variance. Tweets without analytics have a mean sentiment score that is positive (μ = 0.14421, σ^2 = 0.1453, n = 459) with a lower variance relative to the tweets with analytics present. The large spread of the scores of tweets with stats can be visualized in *Figure 3A*.

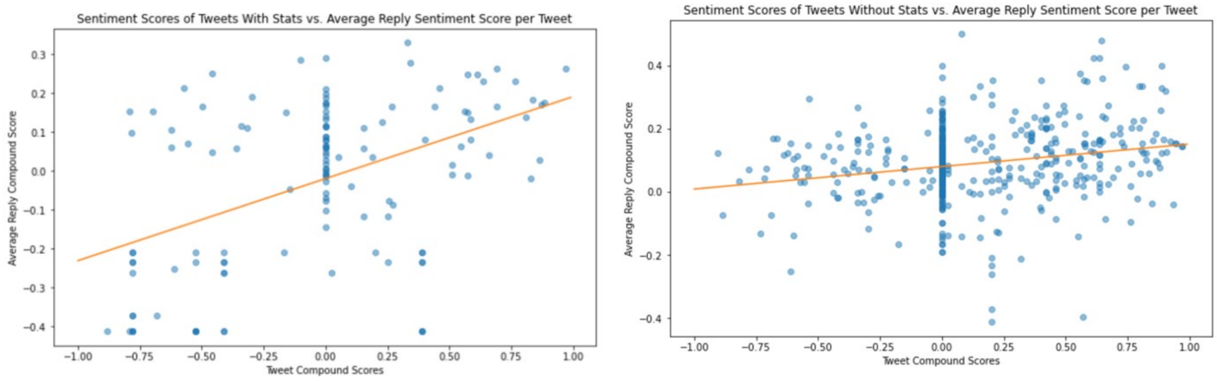


Figure 3: **Scatterplots of Tweet Compound Scores vs. Average Reply Compound Scores with Best Fit Lines**

A: Compound scores of tweets with stats were plotted against their average reply scores. A best fit line was calculated using the 'linregress' function from Scipy's open-source library and yielded values of (β = 0.2109, α = -0.01997, r^2 = 0.22435). The line was plotted using the equation $y = \beta x + \alpha$.

B: Compound scores of tweets without stats were plotted against their average reply scores. Another best fit line was calculated using the same method as *3A*, and yielded values of (β = 0.07122, α = 0.079527, r^2 = 0.050106).

Notice from the scatterplots for the tweets with stats that the data is far more spread, explaining the high variance. However, another important observation from both plots is the cluster of points present in both when the tweet compound score is 0. Many of the tweets were classified by the 'polarity_scores' function as true neutral, meaning they had positive and negative proportions of 0, and neutral proportions of 1, which leads to a compound score of 0. This is a result of the text pre-processing where mentions and links were replaced with words that would lead to a true neutral compound score. While the best fit lines indicate minor correlation, the wide clusters of points in addition to the large clusters when the tweet compound score is 0 indicate that linear regression is not the best technique for drawing conclusions for this data. Despite this, it still revealed important features about the data. For tweets with stats, the correlation coefficient (r^2 = 0.22435) was larger than the correlation coefficient for tweets without stats (r^2 = 0.050106). This means that for tweets with analytics mentioned that have more positive scores tend to generate even more positive responses, and similarly more negative tweets with analytics mentioned tend to generate even more negative responses, more so than tweets without stats.

Figure 4 further supports this point by visualizing the distribution of sentiment scores for tweets and replies separately.

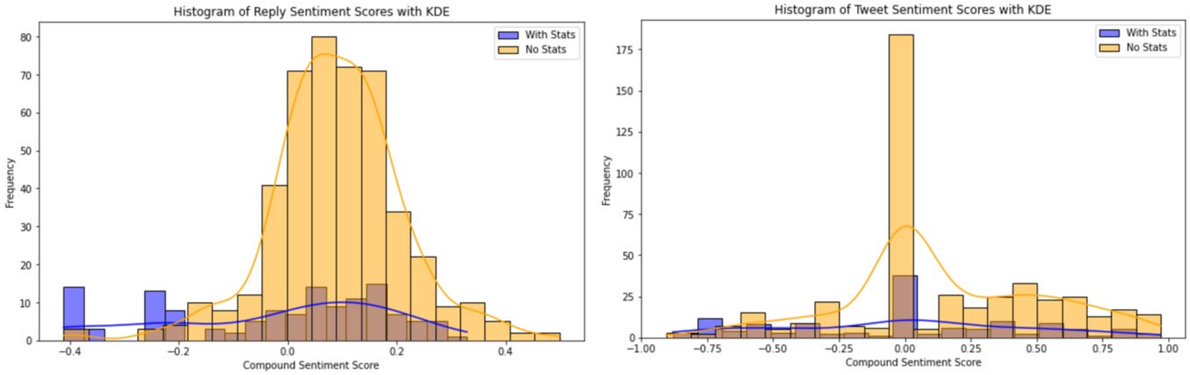


Figure 4: **Distribution of Sentiment Scores for Replies (with/without stats) and Tweets (with/without stats)**

A: Distribution of average reply compound sentiment scores for tweets with stats (blue) and tweets without stats (yellow). This plot was calculated using the function ‘sns.histplot’ from Seaborn’s open source library. The ‘kde’ parameter was toggled to true to calculate and plot the KDE for each distribution.

B: Distribution of tweet compound sentiment scores for tweets with stats (blue) and tweets without stats (yellow). Calculation of KDE was the same as 4A.

There is a noticeable difference in sample sizes between the tweets with stats (blue) and tweets without stats (yellow) in *Figure 4*, however the differences in the shapes of the distributions and KDE curves that highlight differences in the distributions regardless. In *Figure*

4A, there is a large accumulation of negative scores for replies to tweets with stats that makes the distribution more right skewed, indicating a larger than usual presence of negative responses to tweets with stats. In the same figure, the reply scores for tweets without stats are centered around 0, whereas the scores for the tweets with stats are centered slightly above 0 (0-0.2). This could lead one to believe that the responses for replies to tweets with stats are more polar, which is true on this scale, but notice the difference in axes between *Figure 4A* and *4B*. In figure *4B,* the scores range from roughly -0.9 to 0.9, whereas figure *4A* has less than half of that total spread. Despite the distribution in *4B* also being centered at 0, the spread of the scores is larger, meaning the scores are more volatile (or polarizing) than the replies.

3.2 Two Sample T-Test for Differences in Sample Means

A two-sample t-test was conducted using the "ttest_ind" function from scipy's open-source library. This function accepts two arrays as input, then conducts a two-sample t-test and returns two values: the t-statistic and the p-value. In this context, the p-value will be used to assess significance at the α = 0.05 significance level, meaning if p < α, then the results are statistically significant. First, a test was conducted to compare the compound scores of tweets containing stats and tweets without stats. This returned significant results at the standard α = 0.05 significance level (t = –4.0562, p = $5.661 * 10^{-5}$). This means that having stats did cause significant differences to appear in the sentiment scores of the tweets themselves vs. not, however it does not say in which direction. In other words, while this test concludes significant differences in the scores, it does not clarify whether having stats makes a score more positive or negative.

A second t-test was performed on the average reply compound scores of tweets with stats vs. tweets without stats, which also yielded statistically significant results at the $\alpha = 0.05$ significance level (t = –7.857, p = $1.886 * 10^{-14}$). Again, this only reveals significant differences in the sample means, but fails to elaborate on which direction. However, for both tests, the p-value was extremely small compared to the significance level, meaning the magnitude of the difference was of a large magnitude, regardless of the direction.

3.3 Ordinary Least Squares Linear Regression

Ordinary least squares (OLS) is a linear regression technique that aims to minimize the values of the squared differences (also called residuals) between observed values and values predicted by the model. For this project, 3 OLS models were created, each with one additional coefficient corresponding to a feature of the data, to attempt to predict the average reply score for a tweet.

Model 1 was the first OLS model and began with two coefficients to represent a tweet having stats present vs. not having stats, and a tweet's compound sentiment score. It can be represented with the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

β_0 is the intercept term, which helps fit the data better. Without it, the model would be forced through the origin (0,0), which may not necessarily be the best fit for the data. β_1 is the coefficient for the array of compound tweet scores (x_1: TWEET_SCORE_COMPOUND), and β_2 is the coefficient for the binary array indicating whether a tweet has stats or not (x_2: HAS_STATS_01), with a 1 indicating true, and a 0 indicating false. This array is special because it is a dummy array created specifically for this model, and it represents a column of the data which is a Boolean column (a column of trues and falses). It was necessary to change this Boolean column into a binary column for the OLS function to work properly.

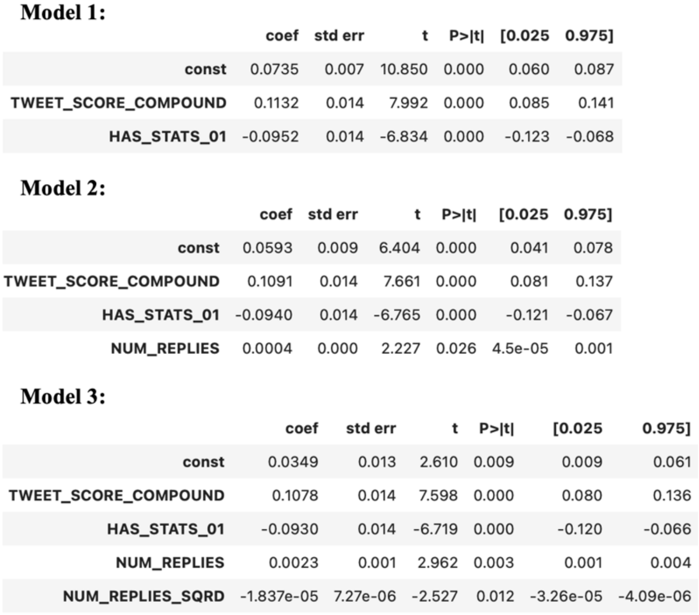
The equation derived from Model 2 has a similar form to Model 1, but with one additional parameter:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

β_3 is the coefficient for the array of the number of replies (x_3: NUM_REPLIES). This seems like a thorough model, that contains all the useful features from the data. However, the number of replies may not be normally distributed. In theory, more polar tweets (tweets with a very high or very low sentiment score) will generate more replies, whereas more neutral tweets (tweets with a score closer to 0) will have less replies. To account for this, Model 3 contains an additional parameter: the number of replies squared. It can be represented with this equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

β_4 is the coefficient for a new dummy column (x_4: NUM_REPLIES_SQRD) which contains the number of replies squared. Theoretically, this will be able to account for non-linearity in the distribution of the replies. *Figure 5* displays the results of the three models.



Model 1:

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0735	0.007	10.850	0.000	0.060	0.087
TWEET_SCORE_COMPOUND	0.1132	0.014	7.992	0.000	0.085	0.141
HAS_STATS_01	-0.0952	0.014	-6.834	0.000	-0.123	-0.068

Model 2:

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0593	0.009	6.404	0.000	0.041	0.078
TWEET_SCORE_COMPOUND	0.1091	0.014	7.661	0.000	0.081	0.137
HAS_STATS_01	-0.0940	0.014	-6.765	0.000	-0.121	-0.067
NUM_REPLIES	0.0004	0.000	2.227	0.026	4.5e-05	0.001

Model 3:

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0349	0.013	2.610	0.009	0.009	0.061
TWEET_SCORE_COMPOUND	0.1078	0.014	7.598	0.000	0.080	0.136
HAS_STATS_01	-0.0930	0.014	-6.719	0.000	-0.120	-0.066
NUM_REPLIES	0.0023	0.001	2.962	0.003	0.001	0.004
NUM_REPLIES_SQRD	-1.837e-05	7.27e-06	-2.527	0.012	-3.26e-05	-4.09e-06

Figure 5: **Results of OLS Regression for 3 Models**

A: Model 1 results with only two parameters: tweet compound scores and whether the tweet has stats.

B: Model 2 results with same parameters as model 1, but with one additional parameter: number of replies.

C: Model 3 results with same parameters as model 2, but with number of replies squared parameter to account for non-linearity.

At the the α = 0.05 significance level, we can see that the p-values for all the coefficients were statistically significant. This means that all coefficients that have p < α are significantly different from 0, meaning they have a profound effect on the the model. If a p-value in *Figure 5* is listed as 0.000, it means that the value is smaller than 0.001.

The t-statistics from each model can be used to evaluate the impact of each feature on the model. A t-statistic is equivalent to the number of standard errors away from the mean of the distribution. Generally, if a t-statistic is greater than 2 or less than -2, it is said to be statistically significant. From Model 1, the compound tweet score has a t-statistic of $t = 7.992$, meaning that the tweet score has a profound impact on the average reply score. This conclusion is supported by the t-statistics from Model 2 ($t = 7.661$) and Model 3 ($t = 7.598$) for the same parameter. To infer the direction of the impact, we can look at the coefficient, which is positive for all 3 models. This means that a tweet with a positive sentiment score will generally have more positive replies, and a tweet with a negative sentiment score will have more negative replies.

Another extremely significant parameter was the binary parameter “HAS_STATS”, however the coefficients for each model with respect to this parameter are negative. Models 1, 2, and 3 all had statistically significant t-statistics, and negative coefficients. This means that having stats decreases the average reply score by a statistically significant magnitude.

The parameters involving the number of replies were just barely statistically significant but had a far smaller magnitude than the previous two parameters mentioned. In Model 2, the "NUM_REPLIES" parameter had a very minutely positive coefficient, whereas in Model 3, the coefficient was larger. However, in Model 3, there was the additional parameter of "NUM_REPLIES_SQRD", which had a negative coefficient. If the relationship between the number of replies and the average reply sentiment score is non-linear as hypothesized before, then these coefficients are unhelpful. Despite this, the coefficients for the last two parameters in Model 3 can be used to calculate the threshold number of replies where the average reply sentiment score will start to lower. For polynomials of degree 2 with the form $y = ax^2 + bx + c$, the vertex can be calculated using the formula: $\frac{-b}{2a}$. In this case, a = β_4 and b = β_3. Using the values of these two coefficients from *Figure 5C*, the threshold is 62.6 replies. In the context of the model, this means that after ~62 replies, the average sentiment score of replies will start going down as opposed to before 62 replies, where the average sentiment score of replies was directly correlated with the number of replies.

4 Conclusion

Based on the results from Models 1, 2, and 3, there is a clear relationship between a tweet containing information about stats and how that tweet is received by users on Twitter. Tweets that contain stats generally lead to lower sentiment scores in the replies, implying that they cause more controversy. This was consistent with the literature review, highlighting how statements with analytics cause controversy or heavier discourse in the basketball community.

There are several factors that influence how a tweet is received, and this project covers a select few of them. There are two very important features that this project was unable to incorporate that future studies should consider. First is NBA player interaction with tweets. Tweets that players respond to directly are far more likely to get replies and media attention, and consequently cause controversy. Another Boolean column in the dataset showing if a player had interacted with a tweet would have been helpful in eliminating this factor. Next is some sort of correlative variable with the NBA season itself, specifically which games are happening on the day the tweet is posted. This feature is vaguer, but it is important not to ignore that people tweet based on what is happening in the world. For example, if the Denver Nuggets blew a big game against the Portland Trail Blazers, there would be many angry Denver fans tweeting at that specific time, just as there would be many happy Blazer fans gloating to those angry Denver fans. Tweets are not independent of real-world events, so establishing a link between what is happening in the timeline of the season with the timing of a tweet will help provide some insight into if analytics make these reactions more intense. There are a few lesser features that were also ignored or standardized for simplicity in this project, such as the number of followers each account has or how active accounts are on Twitter. The amount of data that can be incorporated and studied is limitless, and future researchers with more time, resources, and computing power should incorporate them to paint a more complete picture of the role analytics plays in NBA twitter discourse.

This project shows that technology is an effective way to learn insights about people, and conversely people are an effective way to learn about technology. Not necessarily how technology works, but what role it plays in society. The rapid integration of analytics in the NBA comes with great promise and is advancing the way basketball is played and studied. However, with the integration of any new technology with widespread impact, it is important to gage the effect it has on communities. As this project has demonstrated, NBA analytics on Twitter can be a polarizing topic, as people will often use numbers to defend their own views about players, coaching styles, and strategies. The league is not slowing down with the integration of analytics, with new basketball technologies are being invented every day. This is why it is even more important that the NBA organization remains conscious of the logistical and ethical impacts of these technologies on the basketball community. Constant interaction with and research on the NBA community, players, and organizations ensures that the league remains a space of healthy competition, growth, and joy for basketball lovers around the world.

Bibliography

Abbas, Nabil M. “NBA Data Analytics: Changing the Game.” *Medium*, 21 Aug. 2019, towardsdatascience.com/nba-data-analytics-changing-the-game-a9ad59d1f116.

Berger, Ken. “Warriors ‘wearable’ Weapon? Devices to Monitor Players While on the Court.” *CBSSports.Com*, 5 June 2015, www.cbssports.com/nba/news/warriors-wearable-weapon-devices-to-monitor-players-while-on-the-court/.

Burns, Ed. “Sports Data Analytics Isn’t Always a Slam Dunk: TechTarget.” *Business Analytics*, 6 Oct. 2016, www.techtarget.com/searchbusinessanalytics/opinion/Sports-data-analytics-isnt-always-a-slam-dunk.

Chin, Barry. “New Age of NBA Analytics: Advantage or Overload? - The Boston Globe.” *BostonGlobe.Com*, 30 Mar. 2014, www3.bostonglobe.com/sports/2014/03/29/new-age-nba-analytics-advantage-overload/1gAim4yKYXGUQ2CTAe7iCO/story.html?arc404=true.

Cohen, Ben, director. *SSAC23 - Leveling Up: How Basketball Teams Deploy Analytics to Develop Players*. *YouTube*, YouTube, 14 Mar. 2023, https://www.youtube.com/watch?v=knZ8ImMzBX4&list=PLhbPeSFiFnAaCOQ8GBqw6ES5pfSmlul0V&index=13. Accessed 19 Apr. 2023.

Chotiner, Isaac. “Bomani Jones on the N.B.A., Analytics, and Race.” *The New Yorker*, 17 June 2019, www.newyorker.com/news/q-and-a/bomani-jones-on-the-nba-analytics-and-race.

D'Amour, Alex, et al. “A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes.” *Taylor & Francis*, 2016, https://www.tandfonline.com/doi/abs/10.1080/01621459.2016.1141685?journalCode=uasa20.

Daskal, Ouriel. “Data Tracking of Athletes Tests Ethical Boundaries.” *CTECH - Www.Calcalistech.Com*, 31 Jan. 2021, www.calcalistech.com/ctech/articles/0,7340,L-3890711,00.html.

Facchinetti, Tullio, et al. “Filtering Active Moments in Basketball Games Using Data from Players Tracking Systems - Annals of Operations Research.” *SpringerLink*, Springer US, 16 Nov. 2021, https://link.springer.com/article/10.1007/s10479-021-04391-8.

Flannery, Paul. “How and Why NBA Coaches Communicate Advanced Metrics to Players.” *SBNation.Com*, 23 July 2013, www.sbnation.com/nba/2013/7/23/4547374/nba-coaches-analytics-players-celtics-brad-stevens.

Franks, Alexander, et al. “Counterpoints: Advanced Defensive Metrics for NBA Basketball.” *MIT Sloan Sports Analytics Conference*, 2015, http://www.lukebornn.com/papers/franks_ssac_2015.pdf.

Kopf, Dan. “Data Analytics Have Made the NBA Unrecognizable.” *Quartz*, 18 Oct. 2017, qz.com/1104922/data-analytics-have-revolutionized-the-nba.

Kusumasari, B., & Prabowo, N. P. A. (2020, June 30). *Scraping social media data for disaster communication: How the pattern of Twitter users affects disasters in Asia and the Pacific - Natural Hazards*. SpringerLink. https://link.springer.com/article/10.1007/s11069-020-04136-z

Lowe, Zach. “Lights, Cameras, Revolution.” *" Lights, Cameras, Revolution*, 19 Mar. 2013, http://grantland.com/features/the-toronto-raptors-sportvu-cameras-nba-analytical-revolution/.

Macdonald, B. (2020). Recreating the Game: Using Player Tracking Data to Analyze Dynamics in Basketball and Football . Harvard Data Science Review, 2(4). https://doi.org/10.1162/99608f92.6e25c7ee

Min-Chun, Hu, et al. “Robust Camera Calibration and Player Tracking in Broadcast Basketball ...” *IEEE Explore*, Apr. 2011, https://ieeexplore.ieee.org/abstract/document/5671490.

Sampaio, Jaime, et al. “Exploring Game Performance in the National Basketball Association Using Player Tracking Data.” *PLOS ONE*, Public Library of Science, 15 July 2015, https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0132894.

Skinner, Brian, and Stephen J. Guy. “A Method for Using Player Tracking Data in Basketball to Learn Player Skills and Predict Team Performance.” *PLOS ONE*, Public Library of Science, 9 Sept. 2015, https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0136393.

Thomas, Etan. “Is It Ethical for the Media to Pay NBA Employees for Leaked Footage?” *NBA News and Rumors, Advanced NBA Stats, NBA Odds, Basketball News*, 10 Oct. 2022, www.basketballnews.com/stories/nba-media-golden-state-warriors-draymond-green-jordan-poole-tmz-tabloid-reputable-journalism-espn-the-athletic-etan-thomas-david-aldridge-howard-bryant-michael-lee.

Tian, Changjia, et al. “Use of Machine Learning to Automate the Identification of Basketball Strategies Using Whole Team Player Tracking Data.” *MDPI*, Multidisciplinary Digital Publishing Institute, 18 Dec. 2019, https://www.mdpi.com/2076-3417/10/1/24.

Vangelis, Sarlis, and Tjortjis Christos. “Sports Analytics - Evaluation of Basketball Players and Team Performance.” *Information Systems*, Pergamon, 23 May 2020, https://www.sciencedirect.com/science/article/abs/pii/S0306437920300557.

Yu, Andrew, and Sunnie Chung. “Framework for Analysis and Prediction of NBA Basketball Plays: On-Ball Screens” *IEEE Explore*, 2019, https://ieeexplore.ieee.org/abstract/document/9060200/.