Datasheet for NBA player stats*

Dataset retrieved from Basketball Reference

Yihang Cai

April 19, 2024

This datasheet details the NBA player stats dataset, which includes performance metrics for NBA players during the 2023-2024 season. Designed to support a variety of analytical tasks, the dataset contains data on points, assists, rebounds, and other performance indicators from official NBA games. It describes the dataset's creation, structure, distribution, and maintenance, providing essential information for users such as researchers, analysts, and sports enthusiasts. The document ensures users understand how to use the dataset effectively for sports analytics and related projects.

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset was created to enable detailed analysis and predictive modeling of NBA players' performance for the 2023-2024 season. It aims to predict the number of points a player scores based on various performance metrics.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was created by Basketball Reference, and I downloaded following the instruction given. Utilizing publicly available data from Basketball Reference, which sources its current-season data from SportRadar, the official statistics provider of the NBA.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

[*] Code and data are available at: https://github.com/peachvegetable/NBA-player-points

   - Unsure. Often such datasets are funded by the hosting organizations themselves or through partnerships with sports leagues and media outlets.

4. *Any other comments?*

   - None

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

   - Each instance represents statistical data for an NBA player for a particular season, including metrics like points scored, assists, rebounds, etc.

2. *How many instances are there in total (of each type, if appropriate)?*

   - The dataset contains 718 observations, each corresponding to a unique player or a player's stint with a particular team.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

   - The dataset contains a comprehensive set of NBA player statistics for the specified season, rather than a sample. It includes all players who participated in the NBA during the 2023-2024 season, accounting for any who played sufficient minutes to be included in standard statistical aggregations.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

   - Personal and Demographic Information: Player's name, position, age, and team.

- Game Participation: Games played and started, minutes played.
- Scoring Metrics: Field goals, three-point field goals, and free throws (made, attempted, and percentages).
- Rebounding: Offensive, defensive, and total rebounds.
- Playmaking and Defense: Assists, steals, blocks.
- Game Conduct: Turnovers, personal fouls.
- Overall Performance: Total points scored.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - The dataset does not inherently contain a designated label or target for predictive modeling purposes. However, any of the numerical statistics, such as total points scored (PTS), could serve as a target variable for predictive analyses.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - None

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - The dataset does not explicitly encode relationships between instances (players). However, relationships or comparisons could be inferred or constructed based on team affiliations, positions, or other shared attributes.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

   - For predictive modeling, such as predicting a player's points based on their performance, a common approach is to split the data into training, validation, and testing sets. A typical split ratio could be 70% for training, 15% for validation, and 15% for testing. The training set is used to train the model, the validation set is for tuning model parameters and preventing overfitting, and the testing set is for evaluating the final model's performance. This split ensures that the model is evaluated on unseen data, providing an unbiased estimate of its performance.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

   - The dataset might contain redundancies, especially if it includes separate entries for players who have been traded or played for multiple teams within the season. These instances might need to be consolidated or treated separately, depending on the analysis.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

- The dataset is self-contained in terms of providing a comprehensive set of season-long player statistics. It does not rely on external resources for its primary content, but additional context or data (e.g., team performance, player salaries) might require external sources.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - The dataset contains publicly available NBA player statistics and does not include any personal or confidential information. Data like player performance metrics are standard in sports analytics and are widely disseminated without privacy concerns.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - The dataset is purely statistical and related to professional basketball performance. It does not contain any content that could be considered offensive, insulting, threatening, or anxiety-inducing.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - The dataset may allow for the identification of sub-populations based on age (as age is a recorded attribute) and position played, which might correlate with certain physical attributes or skill sets. However, it does not explicitly categorize players by other demographic variables such as gender, as the dataset pertains solely to an all-male professional sports league.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - Yes, individuals can be directly identified from the dataset as it contains the names of NBA players. The dataset is intended to provide statistics at the individual player level, making identification not only possible but a fundamental feature of the data.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - The dataset does not contain sensitive personal information. It is focused on professional performance metrics in the context of NBA basketball games, such as points scored, assists, rebounds, etc. It lacks personal data such as race, ethnic origins, sexual orientations, religious beliefs, political opinions, financial or health data, or any forms of government identification.

16. *Any other comments?*

    - This dataset is valuable for sports analytics purposes, offering insights into player performance and league dynamics. Its content is ethically unproblematic, being limited to professional achievements in a public sporting context.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was directly observable and publicly available, compiled from official NBA games and player performance statistics. These statistics are typically recorded by official scorers and statisticians during games and are not reported by the subjects (players) themselves. The data is validated through the official NBA data validation processes, which involve real-time and post-game reviews to ensure accuracy.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The collection involved both hardware apparatuses (like tracking devices and cameras in arenas) and software programs (for data aggregation and presentation). NBA and SportRadar, as the official statistics provider, likely use a combination of these tools to compile and validate game statistics. The validation processes are rigorous, involving multiple checks to ensure data accuracy and reliability.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - This dataset is not a sample but rather a comprehensive compilation of all player statistics for the 2023-2024 NBA season, representing a complete set of data points for that timeframe.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

- The data was collected by official NBA statisticians and game officials, possibly with support from SportRadar personnel. These individuals are professionals employed or contracted by the NBA and SportRadar, compensated according to their professional agreements, not crowdworkers.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

   - The data was collected over the course of the 2023-2024 NBA season. The timeframe of data collection matches the creation timeframe of the data, as it is recorded live during games and updated in near real-time.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - Given the public and professional nature of the data, formal ethical review processes like those conducted by institutional review boards are generally not applicable. The data collection is part of standard operations within professional sports leagues.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - The data was obtained via a third party, the Basketball Reference website, which aggregates official NBA statistics.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

   - As professional athletes, NBA players are aware that their performance metrics are recorded and publicly disseminated as part of their participation in the league. Formal individual notifications for each data collection instance are not standard practice.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

   - Players consent to the recording and public sharing of their game performance as part of their professional contracts with the NBA and their respective teams. Explicit consent for each instance of data collection in the form of game statistics is inherent to their professional engagement.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - The public sharing of game performance data is a standard part of professional sports contracts. Individual players typically do not have the option to revoke consent for the sharing of standard game performance data.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - An analysis of this nature is generally not conducted for publicly available professional sports performance data, as it is considered non-sensitive and is a standard aspect of professional sports analytics.

12. *Any other comments?*

    - The collection and use of this data are standard practices within the domain of professional sports analytics and are not subject to the typical privacy concerns associated with personal or sensitive data.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Yes, preprocessing typically includes correcting typographical errors, standardizing team names and player identifiers, and handling missing or incomplete data entries. This ensures consistency across the dataset for analysis purposes.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - Unsure. It is common for organizations like Basketball Reference to maintain both "raw" and processed data for internal use, but publicly available data might only include processed data. For access to the "raw" data, direct contact with the data provider would be required.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - Unsure. The specific tools used for data preprocessing by Basketball Reference or similar organizations are not typically disclosed publicly. However, common data manipulation packages in languages like Python (pandas) or R (dplyr) are likely used.

4. *Any other comments?*

    - The preprocessing steps are crucial for ensuring the accuracy and usability of the data, especially in a dynamic and detail-oriented field like sports analytics. It is recommended for users to conduct their own checks and cleaning routines suited to their specific analytical needs to ensure data integrity.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - Yes, the dataset has been extensively used for various analytical tasks such as performance analysis, predictive modeling, player valuation, and strategy development. Additionally, sports journalists, fantasy sports enthusiasts, and researchers use this data for articles, reports, and academic studies on sports science and analytics.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

    - While there is no specific repository that links to all papers or systems that use this dataset, research articles, and analyses citing such datasets can often be found through academic databases like Google Scholar, or directly on sports analytics blogs and websites like Basketball Reference itself.

3. *What (other) tasks could the dataset be used for?*

    - Beyond traditional sports analysis, the dataset could also be used for machine learning projects, such as developing algorithms to predict future player performance or career trajectory. It can also be employed in economic studies assessing the financial impact of player performances on team revenues and sports marketing.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

    - The dataset's utility might be limited by its focus on quantitative metrics without contextual qualitative data such as player injuries, psychological factors, or team dynamics. These omissions can lead to misinterpretations if users are not cautious. To mitigate these risks, users should consider integrating this dataset with qualitative analyses and not rely solely on the data for making broad assessments about player capabilities or team performance.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

   - This dataset should not be used for making personal judgments about athletes or for decision-making that could affect the personal lives or careers of the players without corroborative evidence from more comprehensive sources. Decisions based solely on statistical data without context can lead to ethical and fairness issues.

6. *Any other comments?*

   - Users of this dataset should remain aware of the temporal nature of sports data; what may be true for one season might not hold in another due to changes in team compositions, player conditions, and other external factors such as rule changes in the sport.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

   - Yes, the dataset is intended to be accessible to researchers, analysts, and sports enthusiasts, which means it will be distributed beyond the original data curating entity.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

   - The dataset is distributed via an API and a downloadable format on sports analytics websites like Basketball Reference. It does not usually have a digital object identifier (DOI) but is referenced by its source and season.

3. *When will the dataset be distributed?*

   - The dataset is updated and distributed annually, following the end of each NBA season to include the latest statistical data.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

   - The dataset is distributed under terms of use that restrict the data for non-commercial, research, or educational purposes only. Users are typically required to acknowledge the source in any publications or presentations. Detailed terms can be found on the provider's website.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - No additional third-party IP restrictions are typically imposed beyond the terms of use specified by the data provider.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - No export controls or regulatory restrictions apply to this dataset, as it contains publicly available sports performance data.

7. *Any other comments?*

    - Users should ensure they remain compliant with the terms of use when using the dataset for any purpose.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The dataset is maintained by the data team at Basketball Reference or the respective sports analytics firm responsible for its compilation.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - The dataset curator can be contacted via email provided on the Basketball Reference contact page or the dataset's main webpage.

3. *Is there an erratum? If so, please provide a link or other access point.*

    - Yes, users can report errors or issues with the dataset through a dedicated section on the website where the dataset is hosted, and errata are published in subsequent updates.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - The dataset is updated annually with new data from the latest NBA season. Updates are communicated to users via newsletters and update logs on the website. This dataset for the 2023-2024 season is not likely to be updated, since the season is already ended and player statistics are fixed.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

   - The data, being historical sports performance records, is retained indefinitely as part of the public record. No limits are imposed on data retention.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

   - Older versions are archived and remain accessible for historical comparison and research purposes.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

   - Contributions or suggestions for dataset extensions can be submitted via the platform's contribution section, but all contributions are reviewed and validated by the data team to ensure accuracy and consistency.

8. *Any other comments?*

   - Users are encouraged to check for updates regularly and use the latest version of the dataset for the most accurate analysis.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92.