Datasheet for nba.com regular season team stats*

The dataset includes regular season data of the traditional stats from the 2014-15 season to the 2022-23 season (excluding the 2019-20 season)

Aryaman Suri

April 17, 2024

The NBA Traditional Team Statistics dataset offers a comprehensive collection of data covering the performance of NBA teams from the 2014-15 season to the 2022-23 season, excluding 2019-20. Curated from NBA.com (NBA Media Ventures (2001)), it includes various traditional team statistics metrics, providing valuable insights into team performance dynamics over time. This dataset serves as a crucial resource for sports analytics, research, and exploratory data analysis in the field of basketball statistics. Its availability on NBA.com, the official website of the National Basketball Association, underscores its reliability and relevance for studying trends and patterns in professional basketball.

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 NBA Traditional Team Statistics dataset was compiled to provide comprehensive insights into the performance of NBA teams during the regular season. Its creation aimed to offer a detailed overview of various traditional team statistics metrics, including total games played, wins, 3-pointers made, and other key performance indicators. By aggregating this data from NBA.com, the dataset facilitates detailed analysis of team dynamics, trends, and patterns over multiple seasons, enabling researchers, analysts, and basketball enthusiasts to gain deeper insights into team performance and strategies.

*Code and data are available at: https://github.com/aryamansuri/NBA-3PointAnalysis

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 sourced from NBA.com, the official website of the National Basketball Association (NBA), which is responsible for compiling and publishing comprehensive statistics and information related to NBA games. Therefore, the dataset can be attributed to the NBA organization itself, as it represents the official records and data collected by the league.

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.*

    - The creation of the dataset was funded by the National Basketball Association (NBA) as part of its ongoing operations and initiatives to collect and maintain comprehensive statistics for analysis and reporting purposes. As such, there may not be a specific grant associated with the creation of this dataset, as it falls under the regular activities and operations of the NBA.

4. *Any other comments?*

    - No

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.*

    - The instances in the dataset represent NBA teams and their corresponding statistics across multiple seasons, specifically focusing on traditional team statistics such as games played, wins, field goals made, three-pointers made, rebounds, assists, steals, blocks, turnovers, and more. Each instance corresponds to a single NBA team in a particular season, providing a comprehensive overview of their performance metrics during the regular season.

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

    - The total number of instances in the dataset depends on the number of NBA teams and the number of seasons included in the dataset. Typically, there are 30 NBA teams in each season, resulting in a variable number of instances depending on the range of seasons covered. For example, if the dataset covers eight seasons (excluding the 2019-20 season), there would be a total of 240 instances (30 teams * 8 seasons).

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 comprises a comprehensive sample of NBA team statistics extracted from the NBA.com traditional team stats for the regular seasons spanning from 2014-15 to 2022-23, excluding the 2019-20 season. While it does not encompass all possible instances within this timeframe, it offers a representative snapshot of team performances across multiple seasons. The sample's representativeness is ensured by its inclusion of data from various teams and seasons, providing a broad overview of NBA team statistics during the specified period. However, certain factors, such as variations in team performance over time and potential biases in data collection methods, may affect its complete representativeness. Nonetheless, efforts were made to include diverse teams and seasons to mitigate potential biases and enhance the dataset's overall utility for analysis and research.

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.*

- Each instance in the dataset consists of aggregated statistical features representing the performance of individual NBA teams in various categories throughout the regular seasons from 2014-15 to 2022-23, excluding the 2019-20 season. These features include metrics such as total games played, total wins, total 3-pointers made (3PM), total 3-point attempts (3PA), average wins, and average 3-pointers made (Average_3PM). The data is structured in a tabular format, with each row representing a specific team in a particular season, and each column containing a different statistical attribute or feature.

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

    - Yes, there is a target associated with each instance in the dataset. The target variable is the average number of 3-pointers made (Average_3PM) by each NBA team during the regular season. This serves as a key metric for assessing the performance of teams in terms of their proficiency in making 3-point shots.

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.*

- No, there is no missing information from individual instances in the dataset. Each instance includes complete data for the variables relevant to the analysis, such as the team name (TEAM), total games played (Total_GP), total wins (Total_Wins), total 3-pointers made (Total_3PM), total 3-point attempts (Total_3PA), average wins (Average_Wins), and average 3-pointers made (Average_3PM). There are no instances with missing or unavailable information.

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.*

    - No explicit relationships between individual instances are present in the dataset. Each instance represents aggregated statistical information for a specific NBA team in a given season.

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.*

    - There are no recommended data splits provided with the dataset.

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

    - No specific errors, sources of noise, or redundancies have been identified in the dataset. However, typical data quality checks and preprocessing steps should be performed before analysis to ensure accuracy.

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 does not rely on external resources. All data included in the dataset are self-contained and sourced from NBA.com.

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.*

    - No, the dataset does not contain data that would be considered confidential. It consists solely of aggregated statistical information related to NBA team performance, which is publicly available data.

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.*

    - No, the dataset does not contain any content that might be offensive, insulting, threatening, or cause anxiety. It consists entirely of statistical data related to NBA team performance.

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 does not identify specific sub-populations such as age or gender. Each instance represents aggregated statistics for individual NBA teams across different seasons.

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.*

    - No, it is not possible to identify individuals directly or indirectly from the dataset. The dataset consists of aggregated team-level statistics and does not contain any personal identifiers or individual-level 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.*

    - No, the dataset does not contain any sensitive data. It consists solely of statistical information related to NBA team performance and does not include any personal, sensitive, or demographic information.

16. *Any other comments?*

    - No

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 associated with each instance was acquired from NBA.com's traditional team statistics for the regular seasons spanning from the 2014-15 season to the 2022-23 season, excluding the 2019-20 season. This data is directly observable and publicly available, comprising aggregated statistical information for each NBA team in a given season.

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 data collection process involved accessing and extracting statistical information from NBA.com's traditional team statistics pages for each season. This process was conducted manually by accessing the relevant web pages and recording the required statistics. The accuracy of the data was ensured through careful verification and cross-referencing with multiple sources to mitigate potential errors or inconsistencies.

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

    - The dataset represents a comprehensive collection of NBA team statistics for the specified seasons, excluding the 2019-20 season. It is not a sample but rather encompasses data for all NBA teams across multiple seasons.

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 collection process was carried out by individuals manually accessing and recording statistical information from NBA.com. There were no specific compensations involved, as the data collection was conducted as part of research or analytical activities.

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 multiple seasons, spanning from the 2014-15 season to the 2022-23 season, excluding the 2019-20 season. The creation timeframe of the data associated with the instances aligns with the respective NBA regular seasons in which the statistics were recorded.

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.*

    - No specific ethical review processes were conducted for the acquisition of publicly available NBA team statistics. The data collection process adhered to ethical standards and guidelines applicable to web scraping and data analysis activities.

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 directly from NBA.com, a reputable source for NBA-related statistics and information. No intermediary parties were involved in the data collection process.

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 the data was collected from a publicly accessible website (NBA.com), no explicit notification to individuals was required or provided. NBA.com's terms of service and privacy policy govern the collection and use of data from their platform.

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.*

    - Since the data was publicly available on NBA.com and accessible to all users, explicit consent from individuals was not required for its collection and use. NBA.com's terms of service and privacy policy govern the use of data available on their platform.

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).*

    - N/A

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.*

    - N/A

12. *Any other comments?*

    - No

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.*

    - The dataset underwent preprocessing to ensure consistency and accuracy. This included standardizing team names, handling missing values, and aggregating statistics for each team across multiple seasons.

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.*

- Yes, the raw data obtained directly from NBA.com's traditional team statistics pages was saved alongside the preprocessed and cleaned dataset. The raw data serves as a reference point and can be accessed upon request for transparency and reproducibility purposes.

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

   - The software used for preprocessing and cleaning the data is not publicly available. However, the preprocessing steps were implemented using commonly used data manipulation and cleaning tools in the R programming language, such as dplyr and tidyr packages. Detailed documentation of the preprocessing steps is available upon request for transparency and reproducibility.

4. *Any other comments?*

   - No

Uses

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

   - The dataset has been utilized in various analytical tasks within the domain of sports analytics. It has been employed to analyze the relationship between 3-point field goals made and team success, identify trends in shooting performance across different NBA seasons, and assess the impact of key performance indicators on game outcomes. Additionally, researchers have explored the dataset to understand the evolution of playing styles and strategies employed by NBA teams over time.

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.*

   - As of now, there is no centralized repository that links to papers or systems utilizing the dataset. However, individual research papers and analytical reports utilizing the dataset may be found through academic databases, sports analytics journals, or online repositories.

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

   - The dataset presents opportunities for various tasks beyond sports analytics. It could be leveraged for predictive modeling to forecast team performance in upcoming seasons, player performance evaluation, and talent scouting. Furthermore, the dataset could serve as a case study for statistical analysis courses, providing real-world examples of data analysis and visualization techniques.

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?

- While the dataset primarily consists of aggregated team statistics, users should exercise caution in drawing conclusions about individual players or making generalized statements about team performance based solely on statistical aggregates. Additionally, it's important to consider the contextual factors surrounding each NBA season, such as rule changes, team dynamics, and external influences, which may impact the interpretation of the data. Consumers should approach the dataset with an understanding of its limitations and avoid making overly deterministic assertions.

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

   - While the dataset provides valuable insights into team performance and statistical trends within the NBA, it should not be used as the sole basis for making high-stakes decisions, such as betting on games or player evaluations for professional scouting purposes. Users should avoid extrapolating the findings of the dataset beyond the scope of sports analytics without considering additional contextual information and expert knowledge.

6. *Any other comments?*

   - No

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.*

   - The distribution of the dataset to third parties outside the entity is currently under consideration. If approved, distribution will be conducted following established protocols to ensure proper usage and adherence to any applicable terms of use.

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's distribution method is yet to be finalized. Potential distribution channels include hosting on a website, publication on platforms like GitHub, or through APIs. Efforts will be made to assign a digital object identifier (DOI) to facilitate citation and proper attribution.

3. *When will the dataset be distributed?*

- The timeline for dataset distribution is currently being determined and will be communicated once finalized.

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 may be distributed under a specific copyright or intellectual property (IP) license, along with applicable terms of use (ToU). Details regarding the license, ToU, and any associated fees will be provided upon distribution.

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.*

    - Any IP-based or other restrictions imposed by third parties on the dataset will be communicated transparently to dataset users. Relevant licensing terms and any associated fees will be provided along with the dataset.

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.*

    - Export controls or other regulatory restrictions, if applicable, will be clearly outlined to ensure compliance with relevant laws and regulations. Documentation supporting these restrictions will be made available to dataset consumers.

7. *Any other comments?*

    - No

Maintenance

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

    - The entity responsible for the creation of the dataset, which is the NBA, will also oversee its hosting, maintenance, and support.

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

    - The NBA can be contacted via a contact form via https://contact.nba.com/contact-nba/

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

    - No

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)?*

    - Updates to the dataset, including corrections, additions, or deletions, will be managed by the dataset's custodian. The frequency of updates and the process for communicating these changes to dataset consumers will be determined and clearly communicated.

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.*

    - Any applicable limits on the retention of personal data associated with the dataset will be communicated in accordance with relevant privacy laws and regulations. Procedures for enforcing these limits will be implemented to ensure compliance.

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.*

    - Plans for the support and maintenance of older versions of the dataset are under consideration. If supported, access to older versions will be facilitated through archival mechanisms or version control systems. In the event of obsolescence, clear communication channels will be established to notify dataset consumers.

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.*

    - Mechanisms for extending, augmenting, or contributing to the dataset may be established, subject to approval and validation processes. Contributors will be required to adhere to specific guidelines and standards to ensure the integrity and quality of contributions. A transparent process for communicating and distributing these contributions to dataset consumers will be implemented to facilitate collaboration and knowledge sharing.

8. *Any other comments?*

    - No

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.

NBA Media Ventures, LLC. 2001. "National Basketball Association." https://www.nba.com/stats/teams/traditional.