Datasheet for the Lahman Baseball Database*

Sean Eugene Chua

30 November 2024

This is the datasheet for the dataset used in the analysis in `paper/paper/pdf`. The extract of the questions below is sourced 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 Lahman database was created to provide a free, comprehensive, and accessible resource for historical baseball statistics to address the lack of a unified source for baseball data. The database was designed to support advanced statistical analysis, allowing users to explore trends and conduct in-depth research.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - Developed by Sean Lahman in the early 2000s, it serves researchers, statisticians, journalists, and baseball enthusiasts by offering extensive data on players, teams, and seasons from 1871-2023. For the longest time, the database was not tied to a specific company, or organization, but it was recently donated to the Society for American Baseball Research (SABR) in October 2024 (Lahman 2024).
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 database was not funded by any external organization, grant, or institution.
4. *Any other comments?*
    - None

Composition

*Code and data are available at: https://github.com/jfhasj/mlb_lahman_analysis.

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 Lahman database primarily represent baseball-related entities and events. These include multiple types of instances that describe various aspects and are related through tables. These tables can be subsumed under one of five categories: players (where each instance represents an individual baseball player, with detailed statistics such as batting, pitching, and fielding records, career data, and biographical information), teams (where each instance represents individual baseball teams, including team-level performance data and roster information), games or seasons (where each instance captures the performance and results of games and entire seasons), managers and coaches (where each instance contains information on team managers and coaches, including their career records and performance metrics (if applicable), and awards and honors (where each instance describes individual player achievements, awards, and recognition).
    - Naturally, these instances are linked relationally such as players being associated with teams, teams being associated with seasons, and players' performance being tied to games and specific seasons.

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

    - In the database, there are 21,010 players, 3,045 different teams throughout 153 seasons, 3,749 instances of managers (with repetition), and 6,990 instances of awards for both managers and players (with repetition).

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 database contains a near-complete set of data for Major League Baseball (MLB) history; it does not yet contain data for the 2024 season, but it is not a random sample from a larger dataset. Instead, it is a comprehensive collection of instances that represent all possible MLB players, teams, games, and seasons until the 2023 MLB season.

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.*

    - Players: Each instance represents a baseball player and includes features such as player name, a unique player ID, birth information, career statistics, and career timeline (start and end years of MLB career, and teams played for).

- Teams: Each instance represents an MLB team and includes features such as team name, unique team ID, franchise history (team's years active and location changes), and season statistics (wins, losses, batting and pitching statistics for the season).
- Seasons: Each instance represents a specific MLB season and includes features such as season year, team ID, and team statistics for that season.
- Games: Each instance represents a specific game and includes features such as game ID, date, teams involved, game result, and individual player statistics for that game.
- Managers: Each instance represents an MLB manager and includes features such as manager name, unique manager ID, and career statistics (win-loss record, winning percentage, games managed).
- Awards and Honors: Each instance represents an award or honor and includes features such as award name, player or manager ID (recipient), year the award was given, and the type of award.

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

   - There is no label or target associated with each instance.

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.*

   - The `Salaries` table in the dataset contains information on player salaries starting from 1985 up until 2016. This is because "there is no real coverage of player's salaries until 1985."

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.*

   - Yes, information on explicit relationships can be found in the "Description" file of the `Lahman` R package.

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.*

   - The dataset does not inherently come with predefined data splits.

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

   - None

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.

- While the Lahman database is self-contained in terms of its data structure, the dataset itself was created by compiling data from a variety of publicly available external sources such as Baseball Reference and Retrosheet (but does not dynamically link to these resources when the dataset is downloaded). The majority of the data in the Lahman database comes from Baseball Reference, which is a comprehensive resource for MLB statistics. This includes player and team statistics, season records, and historical performance data. Retrosheet provides detailed game-level data, including play-by-play information for MLB games going back to 1871.
- The Lahman database itself is static and and does not require external resources once downloaded. While there are no official guarantees that Baseball Reference and Retrosheet will remain constant, they are well-established, and it's unlikely that they will be taken offline or changed significantly without notice. The Lahman database also maintains its own internal consistency, and new versions of the database are released periodically, but it does not automatically fetch data from external sources once downloaded.
- The Lahman database itself is available as a downloadable file (CSV format or other structured formats), and versions of the database are archived. Users can access these versions through the Lahman Database GitHub Repository. The dataset is regularly updated and archived in the GitHub repository for public access. Official versions of the dataset from previous years can be downloaded as well by following the instructions in the `README` of the said GitHub repository.
- The Lahman database is freely available for use and download under an open-source (MIT) license, which allows anyone to use, modify, and distribute the data freely, as long as they provide proper attribution. With regards to the external resources, Baseball Reference and Retrosheet are both free and made publicly available for use. However, while Baseball Reference provides a lot of data for free, it has restrictions on how its data can be scraped or used for commercial purposes. They offer a Pro membership for enhanced features and deeper datasets. Retrosheet data is typically used under specific terms of use, especially for large-scale usage or redistribution.

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 does not contain confidential data; all data therein is public.

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 data does not contain any data that might be offensive, insulting, threatening, or might otherwise cause anxiety.

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 Lahman database does not explicitly identify sub-populations in the sense of sensitive categories. However, it does include some data (through its various tables) that can be used to analyze certain sub-populations based on age, player position, team, country of origin, and other characteristics that might create natural groupings of players within the database. Most players are in their mid- to late 20s, with fewer rookies and veterans. A large proportion of players are from the United States, with notable representation from Latin American countries like the Dominican Republic, Venezuela, and Cuba, as well as smaller numbers from Japan and Canada. Players are spread across the 30 MLB teams, with historically larger franchises showing higher turnover. Most players have shorter careers (under 5 years), with fewer playing 5-15 years and even fewer having careers lasting 15+ years.

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, individual players and managers or coaches can be directly identified through their unique ID or name (as columns of different tables in the database).

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 information. It only contains birth information, personal identifiers, and physical characteristics such as height and weight.

16. *Any other comments?*

    - None

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 in the Lahman database was primarily acquired through direct observation from publicly available records of Major League Baseball (MLB) players. This includes official MLB statistics, player rosters, and other publicly accessible sources like team and league records, historical databases, and official player biographies. Most data such as player and team information, career statistics, and the like are directly observable from these. For example, career statistics like batting averages, home runs, and pitching data are derived from official game records. Information about players' birthplaces, team affiliations, and career timelines is obtained from historical rosters and player data maintained by the MLB and made available on the Lahman database's aforementioned external sources such as Retrosheet and Baseball Reference. The data has been validated and verified through multiple sources over time. This includes cross-referencing with other databases, official team rosters, and MLB records, as well as regular updates and revisions to ensure accuracy. Discrepancies or errors are typically corrected through verification by users, as well as the MLB and the Chadwick Baseball Bureau.

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 Lahman database uses a combination of manual human curation with the help of a team of researchers, software programs, and external APIs to collect and compile data. The accuracy of the dataset was ensured through regular cross-referencing with official MLB records, continual updates, and a structured process of validation and revisions.

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 Lahman database is a comprehensive dataset that aims to cover all MLB players and their statistics over time, so this is not applicable.

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)?*

    - Sean Lahman created the database out of personal interest and did not receive compensation. Contributions from researchers and the community were voluntary, driven by shared interest in baseball and data analysis. No formal monetary compensation was involved in the data collection or contribution process.

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 initial creation of the Lahman database began in 1996. Sean Lahman started compiling the data from various encyclopedias and other publicly available sources. The database is updated annually to include new player and team data, as well as seasonal statistics. The most recent updates are typically made after each MLB season, with Lahman or contributors correcting or adding new information.

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.*

    - As the dataset consists of publicly accessible information, an ethical review process was not needed.

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 collected via a publicly available R package, and Lahman has made the database available on his website as well.

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.*

    - The data used in the Lahman database is not private and so does not require the creators to be notified.

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.*

    - Sean Lahman publicly announced the availability of the database on his X account in this post.

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).*

    - No revocation mechanism exists because there was no consent process in the first place (unless Lahman or SABR suddenly deletes all existing and pre-existing versions of the database).

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.*

    - No such analysis has been made.

12. *Any other comments?*

    - None

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 Lahman database underwent preprocessing to ensure its accuracy, consistency, and usability. Tasks like cleaning, standardization (of dates and places for example), verification of statistics, and data integration were performed to ensure that the data compiled from multiple public sources could be used effectively.

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.*

    - No

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

    - No

4. *Any other comments?*

    - None

Uses

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

    - The dataset has already been used my millions of baseball fans to conduct analyses and further their own understanding of baseball statistics. Numerous books have also been written such as "Analyzing Baseball Data with R" by Max Marchi and Jim Albert, where the data used was taken from the Lahman database.

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.*

    - No such repository currently exists.

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

    - The database can be used for statistical analysis, predictive modeling, historical research

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?*

    - No such concerns arise from with regards to the Lahman database.

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

    - The database should not be used tasks involving real-time data, injury tracking, or analyses related to financials beyond basic player salaries.

6. *Any other comments?*

    - None

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 dataset was donated to the Society of American Baseball Research in October 2024 (Lahman 2024).

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 can be downloaded either on Sean Lahman's website or the database's GitHub repository (where users will be able to access instructions to download the database in R).

3. *When will the dataset be distributed?*

    - The dataset is already publicly available and has been distributed for several years. The database is typically distributed annually to include new seasons and data as they become available.

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 data in the Lahman Baseball Database is made freely available to the public to use for any purpose (commercial or non-commercial). However, Sean Lahman holds the copyright for the dataset and must be given credit when it is used, most especially in academic or professional contexts.

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

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

7. *Any other comments?*

    - None

Maintenance

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

    - Since Lahman donated the database to the Society of American Baseball Research in October 2024, they will be maintaining the dataset in collaboration with other researchers. In addition, a small team is also maintaining the database on its GitHub repository, namely: Chris Dalzell as the maintainer, Michael Friendly as the author, Denis Murphy, Martin Monkman, and Sean Lahman

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

    - Sean Lahman can be reached on his X account (with username: seanlahman), and Chris Dalzell can be reached at his email address (cdalzell@gmail.com).

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

    - There is no formal document for errata, but the community has provided potential issues with the dataset through GitHub Issues on the database's GitHub repository.

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 typically updated after the conclusion every season; updates can be found on Sean Lahman's X account or on the aforementioned GitHub repository.

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.*

    - No, player and team data will always be available on the dataset as long as such data exists and remains public.

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.*

    - Yes, official versions of older versions of the dataset can be downloaded by following the instructions in the `README` of the database's GitHub repository.

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.*

    - Suggestions can be proposed through GitHub Issues on the database's GitHub repository or that of the Chadwick Baseball Bureau's, where those maintaining or managing the dataset will respond to queries.

8. *Any other comments?*

    - None

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.

Lahman, Sean. 2024. "Sean Lahman Donates Lahman Baseball Database to SABR." Society for American Baseball Research (SABR). https://sabr.org/latest/sean-lahman-donates-lahman-baseball-database-to-sabr/.