Data in Brief
Open access
Article template

ARTICLE INFORMATION

Article title

NBA odds dataset linking sportsbook implied probabilities to game outcomes (2020–2024)

Authors

Hieu Vu

Affiliations

University of Virginia

Corresponding author's email address and Twitter handle

brr2tu@virginia.edu

Keywords

sports wagering; basketball; implied probability; calibration; relational database; API integration; time-series snapshots

Abstract

This dataset was created by collecting and integrating National Basketball Association (NBA) game results with sportsbook betting odds in order to support reproducible analyses of implied probabilities, calibration, and betting-return metrics across seasons and bookmakers. Data collection was performed programmatically in Python using two public web APIs. NBA team metadata and game-level results were retrieved from the balldontlie API, including game identifiers, UTC start timestamps, season labels, home/away teams, and final scores. Betting odds were retrieved from The Odds API using its historical odds endpoint to capture daily snapshots over a multi-year window. Each snapshot returns event metadata (including commence time and listed teams), bookmaker metadata, and market outcomes for head-to-head (moneyline) markets with American odds.

During transformation, team names from both sources were standardized via lightweight string normalization to reduce naming mismatches. American odds were converted into implied probabilities for each outcome, and the odds data were stored in a relational schema (SQLite) separating events, bookmakers, and quotes. A linking table was produced to connect odds events to completed NBA games by matching normalized home/away team names and selecting the closest event by absolute time difference within a configurable threshold, with an additional pass that considers swapped home/away orientation. The final dataset supports reuse for evaluating sportsbook probability calibration, favorite/underdog performance, bookmaker comparisons, and strategy backtesting over multiple seasons, and it can be extended with additional markets (spreads/totals) and more robust entity-resolution rules.

35 # SPECIFICATIONS TABLE

Subject	Social Sciences
Specific subject area	Sports analytics and betting markets analysis using implied probabilities and game outcomes
Type of data	Type of data - Table (relational database tables) - Chart - Graph - Figure Data format - Raw - Processed - Analyzed
Data collection	NBA game results and team metadata were collected via the balldontlie API using scripted HTTP requests in Python. Historical sportsbook odds were collected daily using The Odds API historical endpoint, capturing head-to-head markets with American odds across multiple bookmakers. Data were normalized by standardizing team names, converting timestamps to UTC, and transforming American odds into implied probabilities. Odds events were linked to completed games by matching normalized team names and minimizing time differences within a fixed threshold. Data ingestion, cleaning, and storage were implemented using Python, pandas, SQLAlchemy, and SQLite.
Data source location	Data was collected from publicly available web APIs and stored locally and version-controlled at the University of Virginia, Charlottesville, Virginia, USA.
Data accessibility	Repository name: NBA_SPORTS_BETTING_DB Data identification number: N/A Direct URL to data: **https://github.com/CSHieuV/NBA_SPORTS_BETTING_DB** Instructions for accessing these data: Clone the repository using Git, then open the SQLite database (sports_pipeline.db) locally. Data tables can be queried using standard SQL tools or accessed programmatically via Python using pandas and SQLAlchemy. The repository also includes scripts for reproducing data extraction,transformation, and visualization.

Related research article	*None*

VALUE OF THE DATA

Why are these data valuable?

- This data provides a link between historical NBA game outcomes and sportsbook betting odds, allowing researchers to work with both realized results and implied probabilities in a single, normalized dataset. The inclusion of standardized team names, bookmaker identifiers, odds, and relational joins reduces the technical overhead typically required to combine sports results with betting-market data from multiple sources. As such, a typical user will not have to do any of the data cleaning required before using this data.

How can these data be reused by other researchers?

- The dataset can be reused to study probability calibration, forecasting accuracy, and uncertainty in prediction markets by comparing implied probabilities against observed outcomes. Because both raw American odds and derived implied probabilities are included, researchers may apply alternative probability conversions or statistical models without needing to recollect the underlying data.

What types of analyses do these data support?

- This data supports descriptive, comparative, and methodological analyses, including bookmaker-level comparisons, season-by-season trends, favorite versus underdog performance, and return-on-investment simulations. The relational structure also enables time-based filtering and aggregation for other relevant studies.

How can the dataset be extended or combined with other data sources?

- The data can be extended by incorporating additional betting markets, other sports leagues, or external covariates such as team statistics or player-level metrics. The existing extraction and transformation scripts provide a reusable framework for integrating new sources while preserving the original schema.

BACKGROUND

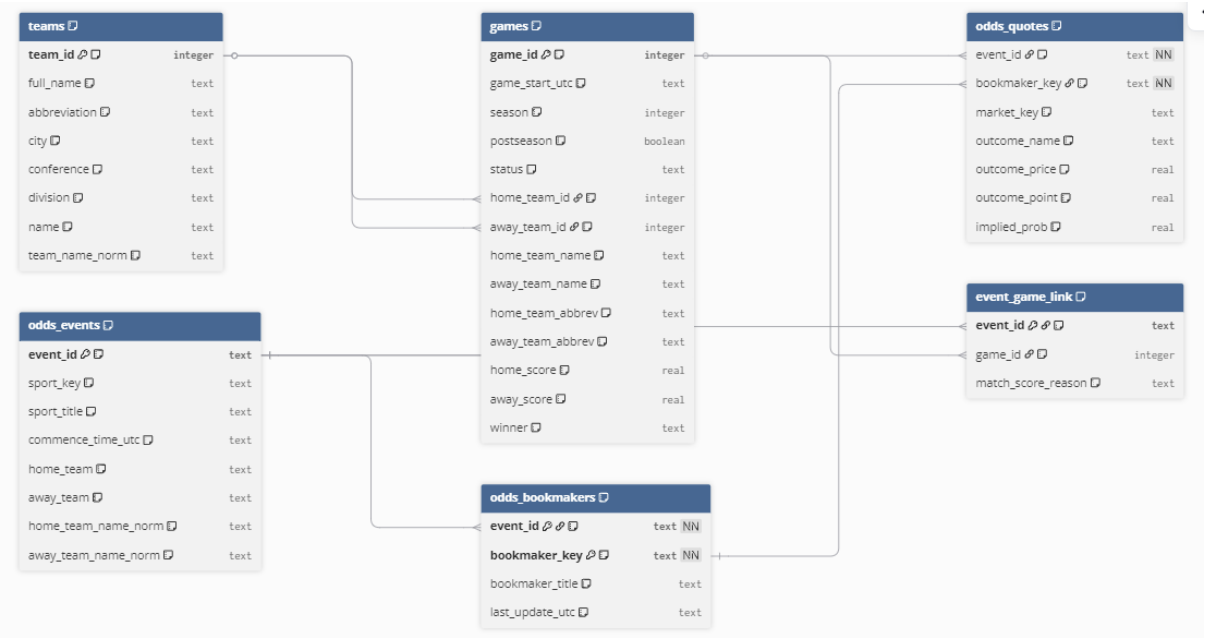
This dataset was compiled to support structured analysis of sports betting markets in the context of professional basketball, with a focus on linking sportsbook odds to realized game outcomes in a reproducible manner. In sports analytics, betting odds are commonly interpreted as expressions of implied probabilities, yet these data are often dispersed across proprietary platforms, inconsistent in format, and difficult to align with official game results over extended time periods.

Methodologically, the dataset was generated within a data engineering framework that emphasizes reproducible extraction, transformation, and loading (ETL) pipelines. Public web APIs were used to

collect NBA game metadata and results alongside historical sportsbook odds. The data was organized into a relational schema to preserve the hierarchical structure of events, bookmakers, and market outcomes, while minimizing redundancy.

The dataset was developed in the context of coursework and independent research in data science and computer science, where there is a need for realistic, multi-source datasets that demonstrate challenges such as entity resolution, time alignment, and probability transformation. By providing both raw odds and derived implied probabilities, the dataset enables reuse for methodological demonstrations without requiring direct access to the original APIs.

DATA DESCRIPTION



Link to Database Documentation: https://dbdocs.io/brr2tu/DS-6600-Data-Science-Final-Project

The dataset is organized as a single SQLite relational database (sports_pipeline.db) that integrates NBA game results with historical sportsbook betting odds. The database schema is designed to preserve source provenance, support reproducible joins across APIs, and enable querying at the team, game, event, bookmaker, and market-outcome levels. The following tables comprise the dataset.

Teams table:

The teams table is a reference (dimension) table containing National Basketball Association team metadata retrieved from the BallDontLie API. Each row corresponds to one NBA team and includes a stable team identifier, full team name, abbreviation, city, conference, and division. A normalized

team name field is included to support joins with external data sources that use different naming conventions.

Games table:

The games table contains NBA game-level data collected from the BallDontLie API, with one row per game. Fields include the game identifier, season label, postseason indicator, game status, start time in UTC, home and away team identifiers, team names and abbreviations at ingestion time, final scores, and a computed winner label. Foreign keys link home and away teams to the teams table.

Odds events table:

The odds_events table stores event-level metadata retrieved from The Odds API historical snapshot endpoint. Each row represents a single betting event and includes the event identifier, sport metadata, commence time in UTC, and raw home and away team names as provided by the odds source. Normalized team name fields are included to facilitate matching with NBA games.

Odds bookmakers table:

The odds_bookmakers table contains bookmaker-specific metadata for each odds event. Rows are uniquely identified by a composite primary key consisting of the event identifier and bookmaker key. This table records the bookmaker's stable identifier, display name, and last update timestamp for the event.

Odds quotes table:

The odds_quotes table stores individual market outcome quotes associated with each event and bookmaker. Fields include the market type (e.g., head-to-head), outcome label, American odds price, optional point values, and computed implied probabilities. This table may contain multiple rows per event and bookmaker due to repeated historical snapshots.

Event game link table:

The event_game_link table connects betting events from The Odds API to completed NBA games from BallDontLie. Each row maps a single odds event to at most one NBA game using normalized team names and time proximity heuristics. A descriptive field records the matching rationale.

EXPERIMENTAL DESIGN, MATERIALS AND METHODS

The dataset was generated using a scripted data engineering pipeline implemented in Python to reproducibly collect, normalize, and store NBA game data and sportsbook betting odds from two independent public web APIs. NBA team metadata and game-level results were retrieved from the BallDontLie API via authenticated HTTP requests, with team data collected as a reference table and game data collected for specified NBA seasons. Retrieved fields included stable team and game identifiers, season labels, postseason indicators, game status, team names and abbreviations, game start timestamps, and final scores when available. All timestamps were converted to UTC and stored as ISO 8601 strings, and a winner label was computed from final scores when both team scores were

present. Historical sportsbook odds were collected from The Odds API using its historical snapshot endpoint, restricted to the basketball_nba sport key and head-to-head (moneyline) markets, with data queried on a per-date basis to capture repeated snapshots. Each API response included event-level metadata, bookmaker metadata, and outcome-level pricing data expressed as American odds, which were stored in separate relational tables to preserve source hierarchy.

Data normalization and transformation were applied uniformly across sources. Team names from both APIs were normalized, with both raw and normalized values retained. American odds were converted to implied probabilities using standard formulas without vigorish removal, and boolean values were stored using SQLite compatible integer representations. Odds events were linked to NBA games using a heuristic matching procedure based on normalized team names and time proximity, considering both original and swapped home-away orientations and selecting the closest match within a predefined time threshold; each event was linked to at most one game, with a descriptive matching rationale recorded. All data was stored in a single SQLite database (sports_pipeline.db) using a relational schema comprising team, game, event, bookmaker, quote, and linkage tables. Database creation and population were managed using SQLAlchemy, with pandas used for intermediate data manipulation and validation. API credentials were handled via environment variables stored in .env files, and all scripts used for data acquisition, transformation, and storage are included in the associated public GitHub repository.

LIMITATIONS

One limitation of the dataset is that the volume and temporal coverage of sportsbook odds data were constrained by the number of API requests permitted due to budget constraints on the creator's end for The Odds API. As a result, historical odds were collected at discrete snapshot intervals rather than continuously, and some games may have fewer recorded odds updates than others. This limitation may affect the density of temporal snapshots available for certain events or seasons but does not impact the structure or integrity of the stored data. No other significant limitations were encountered during data collection or curation.

ETHICS STATEMENT

The authors have read and complied with the ethical requirements for publication in Data in Brief. This work does not involve human subjects, animal experiments, or data collected from social media platforms. All data were obtained from publicly accessible web APIs and do not contain personal, sensitive, or identifiable information.

CRediT AUTHOR STATEMENT

Hieu Vu: Conceptualization; Data curation; Software; Methodology; Validation; Formal analysis; Investigation; Resources; Writing - original draft; Writing - review & editing; Visualization; Supervision; Project administration.

ACKNOWLEDGEMENTS

There are no contributors to acknowledge who do not meet the criteria for authorship. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

REFERENCES

[1] S. D. Levitt, “Why are gambling markets organised so differently from financial markets?” The Economic Journal, vol. 114, no. 495, pp. 223–246, 2004. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0297.2004.00207.x

[2] ——, “Why are gambling markets organised so differently from financial markets?” University of Chicago Price Theory, Tech. Rep., 2004, working paper. [Online]. Available: https://pricetheory.uchicago.edu/levitt/Papers/LevittWhyAreGamblingMarkets2004.pdf

[3] E. Snowberg and J. Wolfers, “Explaining the favorite-longshot bias: Is it risk-love or misperceptions?” IZA Institute of Labor Economics, Tech. Rep. DP4884, 2010. [Online]. Available: https://www.iza.org/publications/dp/4884/explaining-the-favorite-longshot-bias-is-it-risk-love-or-misperceptions

[4] M. Ottaviani and P. Sørensen, “The favorite-longshot bias: An overview of the main explanations,” NBER, Tech. Rep. w15923, 2010. [Online]. Available: https://www.nber.org/system/files/working papers/w15923/w15923.pdf

[5] FiveThirtyEight, “Fivethirtyeight data repository,” https://data.fivethirtyeight.com/, 2017, cC-BY-4.0. Accessed Oct. 23, 2025.

[6] ——, “Nba elo dataset,” https://www.kaggle.com/datasets/

fivethirtyeight/fivethirtyeight-nba-elo-dataset, 2019, accessed Oct. 23, 2025.

[7] ——, “Nfl historical scores and elo (forecasting game data),” https://github.com/fivethirtyeight/nfl-elo-game, 2018, accessed Oct. 23, 2025.

[8] The Odds API, “Sports odds api,” https://the-odds-api.com/, 2025, jSON feed for bookmaker odds. Accessed Oct. 23, 2025.

[9] ——, “Odds api documentation (v4),” https://the-odds-api.com/liveapi/guides/v4/, 2025, endpoints for moneyline, spreads, totals. Accessed Oct. 23, 2025.

[10] Odds API IO, “Odds api: Real-time sports betting odds (250+ bookmakers),” https://odds-api.io/, 2025, accessed Oct. 23, 2025.

[11] football-data.org, “Api reference,” https://www.football-data.org/documentation/api, 2025, rEST endpoints for fixtures and results. Accessed Oct. 23, 2025.

[12] ——, “Api quickstart (v4),” https://www.football-data.org/documentation/quickstart, 2022, accessed Oct. 23, 2025.

[13] BALLDONTLIE, “Balldontlie sports apis,” https://balldontlie.io/, 2025, multi-league coverage. Accessed Oct. 23, 2025.

[14] ——, “Nba api (documentation),” https://nba.balldontlie.io/, 2025, historical and current NBA data. Accessed Oct. 23, 2025.

[15] CollegeFootballData.com, “College football data api,” https://api.collegefootballdata.com/, 2025, games, play-by-play, advanced stats. Accessed Oct. 23, 2025.

[16] ——, “College football data api: Documentation,” https://www.postman.com/api-evangelist/college-football-data/documentation/el6ukhf/college-football-data-api, 2025, postman docs. Accessed Oct. 23, 2025.

[17] S. Gilani and SportsDataverse, “cfbfastr: College football data in r (sportsdataverse),” https://cfbfastr.sportsdataverse.org/reference/index.

html, 2025, open-source wrappers around CollegeFootballData API.
Accessed Oct. 23, 2025.
[18] H. Vu, “NBA sports betting database,” GitHub repository, 2025.
[Online]. Available: https://github.com/CSHieuV/NBA_SPORTS_BETTING_DB.
Accessed Oct. 23, 2025.