Malcolm Darukhanawalla

Dataset Selection Report

For this project, I selected the NBA Betting Data (2007-2024). This dataset provides detailed information on NBA betting odds, including point spreads and lines, from October 2007 to June 2024. It captures data across multiple seasons, making it ideal for historical analysis and trend identification in betting markets. The source of this dataset is Kaggle, where it can be downloaded at:

<https://www.kaggle.com/datasets/cviaxmiwnptr/nba-betting-data-october-2007-to-june-2024>

The dataset is comprised of a single comprehensive file containing multiple data points such as team names, dates, game results, and associated odds. Although the data is consolidated in one file, it can be normalized into several tables for the database. Based on a preliminary review, I plan for the database to have at least six tables, some of which would probably be the following:

1. Teams – Stores team identifiers, names, and abbreviations (if provided).
2. Games – Contains the date, home/away teams, location, and final scores for each game.
3. Betting Odds – Includes opening and closing odds, point spreads, and moneylines for each game.
4. Over/Under Data – Tracks predicted total points and whether the game result went over or under.
5. Seasons – Segments games by season and tracks the timeline of each NBA season.
6. Player/Team Performance Metrics – Stores cumulative team or player statistics (e.g., points per game, win streaks).

This dataset aligns with my personal interest in sports and betting, offering opportunities to explore questions related to betting markets, team performances, and market efficiencies. Some potential business questions could possibly include:

* How often do teams cover the spread compared to bookmaker expectations?
* Are there trends in betting odds that indicate market inefficiencies?
* Do specific teams consistently outperform betting lines over multiple stretches of games or seasons?