Project Title

Moneyball NBA: A 20-Year Statistical Analysis of Evolution, Efficiency, and MVPs

Difficulty

Intermediate

This project moved beyond basic queries into complex data cleaning, window functions, self-joins, and feature engineering.

Project Description

Project Overview

This project analyzes over 12,000 rows of NBA player data (1996–2023) to solve longstanding fan debates using data science. Moving beyond simple averages, this analysis uses SQL to uncover deep trends in player efficiency, physical evolution of the league, and statistical dominance. The goal was to replace subjective opinions with objective, data driven metrics to crown the true "statistical GOATs" and build an analytically perfect starting lineup.

Workflow & Methodology

The project was executed in DBeaver (SQLite) following a standard Data Science lifecycle:

1. Data Cleaning & Preprocessing

- **Structural Repair:** Renamed unlabeled index columns to create a proper Primary Key (row_id).
- **Data Type Standardization:** Solved the "Undrafted" issue by converting mixed-text columns into pure numeric types (converting 'Undrafted' text to NULL) to enable aggregation.
- **Categorical Normalization:** Merged inconsistent country names (e.g., "Bosnia" vs. "Bosnia & Herzegovina") and harmonized team abbreviations.
- **Quality Assurance:** Verified zero duplicates and trimmed whitespace to ensure data integrity.

2. Exploratory Data Analysis (EDA)

- **Era Comparison:** Used CASE statements to segment data into decades (90s, 00s, 10s, 20s), proving the "Small Ball" theory (average player weight dropped ~3kg while scoring increased).
- **Rookie vs. Veteran Impact:** quantified the "Rookie Wall," showing that rookies average a -5.5 Net Rating compared to the -1.9 of veterans.
- **Most Improved Players:** Implemented **Self-Joins** to compare a player's current season against their previous one, successfully identifying historical breakout seasons like CJ McCollum (+14.0 PPG).

3. Feature Engineering & Modeling

- Custom MVP Index: Developed a weighted formula to calculate value beyond raw points:

  $$MVP Score = (Points \times 0.4) + ((Rebounds + Assists) \times 0.3) + (True Shooting \% \times 30)$$

- **Positional Classification:** Created dynamic logic to classify players as Guards, Forwards, or Centers based on height, rather than archaic position labels.

Key Outcomes & Insights

- **The Statistical GOAT Season:** Nikola Jokić's 2022-23 season achieved the highest "MVP Score" in the dataset, driven by near-triple-double averages and elite efficiency (70% TS).
- **The Dream Team:** Built a statistically optimal starting 5 that maximizes efficiency:
  - **G:** James Harden (2019)
  - **G:** Damian Lillard (2023)
  - **F:** Luka Dončić (2023)
  - **F:** Kevin Durant (2014)
  - **C:** Nikola Jokić (2023)
- **Efficiency vs. Volume:** Disproved the myth that high scorers are inefficient; identified "unicorns" like Steph Curry who maintain >65% True Shooting even with high usage rates.

Tools Used

- **SQL (SQLite/DBeaver):** For all data cleaning, transformation, and analysis.
- **Techniques:** Window Functions (RANK, ROW_NUMBER), Self-Joins, CTEs, Case Logic, Aggregations.