NBA Draft and Player Data (1989-2021)

By: Ciara Fasullo, Lesly Veizaga, Tiffanie Kwakye

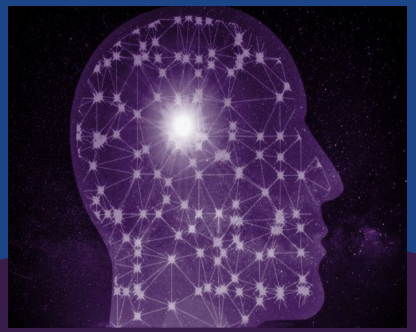
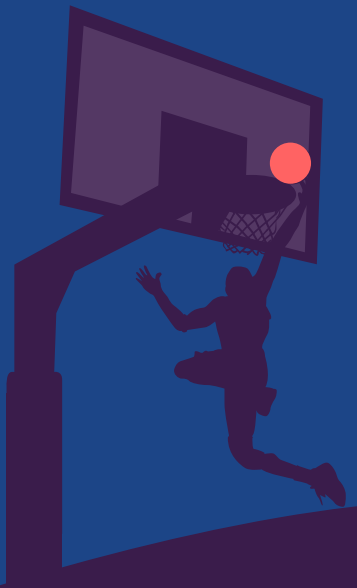


TABLE OF CONTENTS



01 Datasets

Data Selection Overview: NBA Draft Data

- **NBA Draft Data (1989-2021):**
  - This dataset is from kaggle and contains all NBA Draft picks from 1989-2021.
  - **Relevance to Research Question:** This dataset is essential as it provides detailed information about players entering the NBA through the draft, their draft positions (overall pick and round), and their universities. These factors are central to exploring the relationship between the university attended, draft round/pick, and subsequent NBA performance.
  - **Key Data Points:**
    - Draft year and pick order give context to a player's entry into the league.
    - The university attended enables analysis of the impact of collegiate background on NBA performance.
    - The drafting team provides additional context for player development and opportunities.
  - **Expected Insights:**
    - Trends in drafting players from specific universities.
    - Correlation between draft position and perceived player potential versus actual performance.

Data Selection Overview: NBA Player Stats

- **NBA Player Stats (1950-2022):**
  - This dataset is from kaggle and consists of NBA player stats from 1950-2022.
  - **Relevance to Research Question:** This dataset contains long-term performance metrics, including games played, games started, and minutes played—key indicators of a player's contribution and success in the NBA.
  - **Key Data Points:**
    - Career statistics offer a comprehensive view of player performance post-draft.
    - Historical depth (1950-2022) provides a robust sample for longitudinal analysis, even though the focus will be on players drafted between 1989 and 2021.
  - **Expected Insights:**
    - How draft position correlates with career metrics such as games played, games started, and minutes played.
    - Differences in performance among players from various universities, revealing which schools consistently produce high-performing NBA talent.

Why These Datasets are Ideal

- **Complementary Scope:** The draft dataset focuses on entry into the league, while the player stats dataset tracks long-term performance. Together, they provide a complete picture of a player's career trajectory.

- **Historical Depth:** With over three decades of data, the draft dataset and the extended timeline of the player stats dataset allow for robust trend analysis and control for generational shifts in the game.

Expected Insights from Combined Analysis

- **University-NBA Performance Association:**
    - We expect colleges with strong basketball programs (e.g., Duke, Kentucky, North Carolina, Kansas, UCLA) to have players who perform better in the NBA. These schools often attract highly talented recruits and provide strong training and development. We will base this on metrics like total minutes played, games started, and games played.
    - Historically, lower-round picks include undervalued players who outperform expectations, and we would expect them to come from lesser-known programs or international leagues.
- **Draft Position-NBA Performance Association:**
    - Correlation between draft round/pick and long-term performance indicators, addressing whether earlier picks generally perform better or if later picks exceed expectations.
    - Insights into the value of specific draft positions or rounds.

2 ETL Process

ETL Process

Extract:

- Retrieved data by loading data from the CSV files (NBA Player Stats and NBA Players Draft) into Pandas Dataframe

Transform:

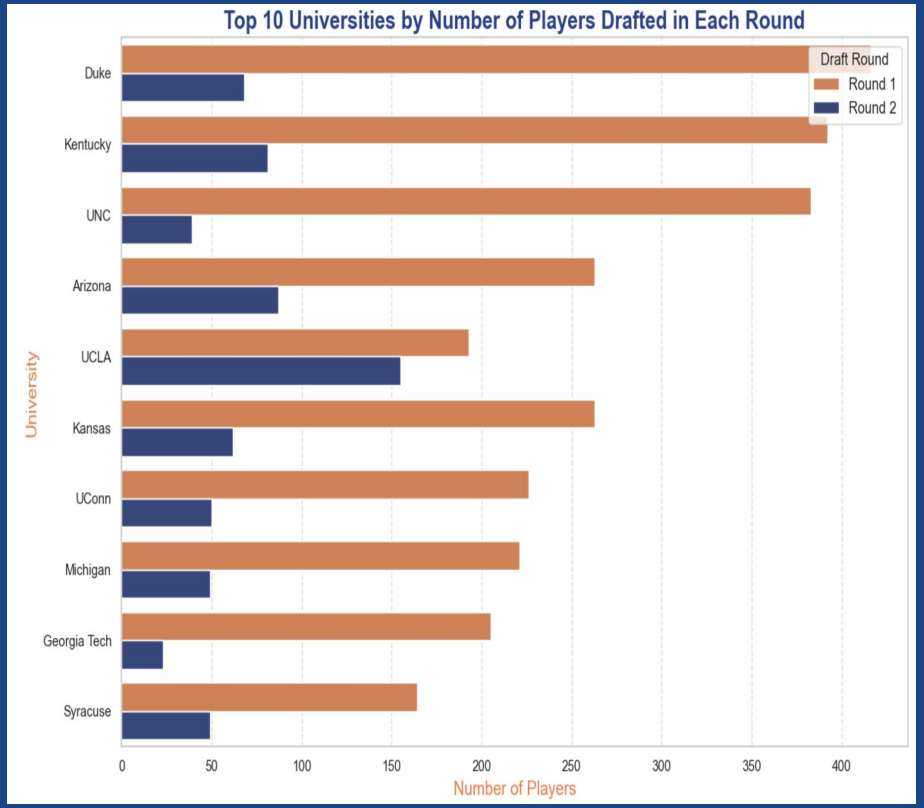
- Filtered the draft dataset to include only players drafted after 1989
- Cleaned and merged the player stats dataset to include only players who appear in the cleaned draft data
- Renamed a column to ensure that the names match between datasets
- Transformed the data by filtering it to ensuring consistency between datasets.

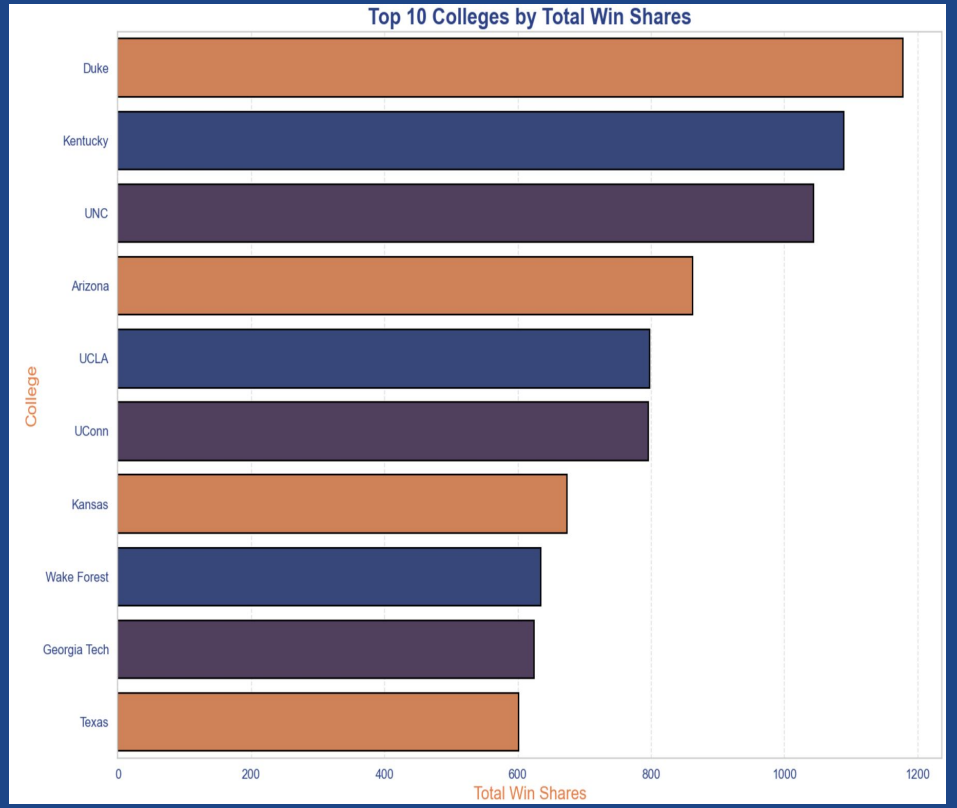
Load:

- Loaded to MySQL database using SQLAlchemy
- Created a command to load the cleaned dataset
- Set up Google cloud storage setting up buckets for data

3 Data Analysis

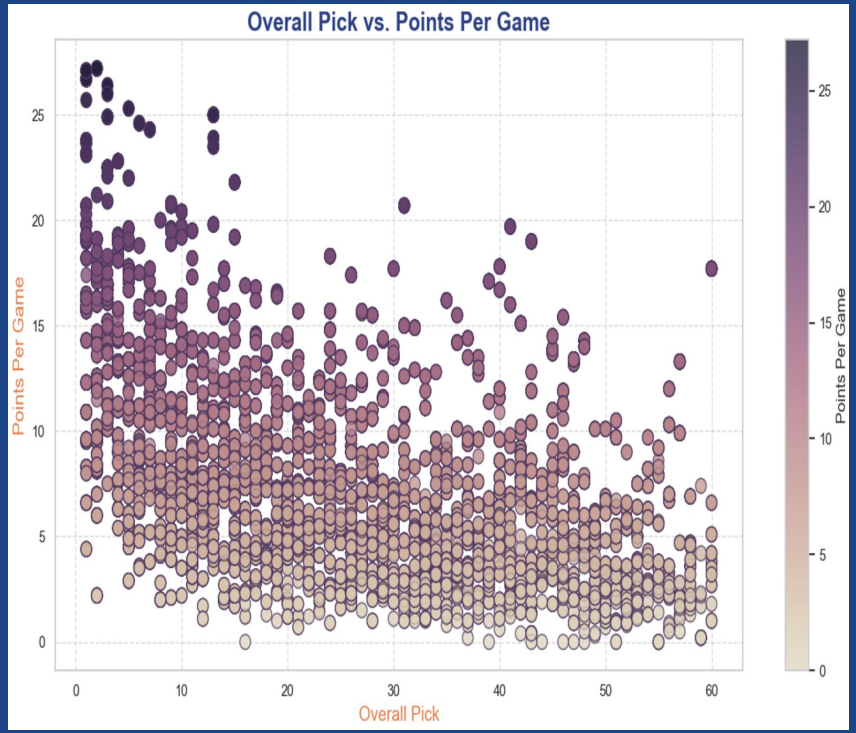
Drafted Players and Universities

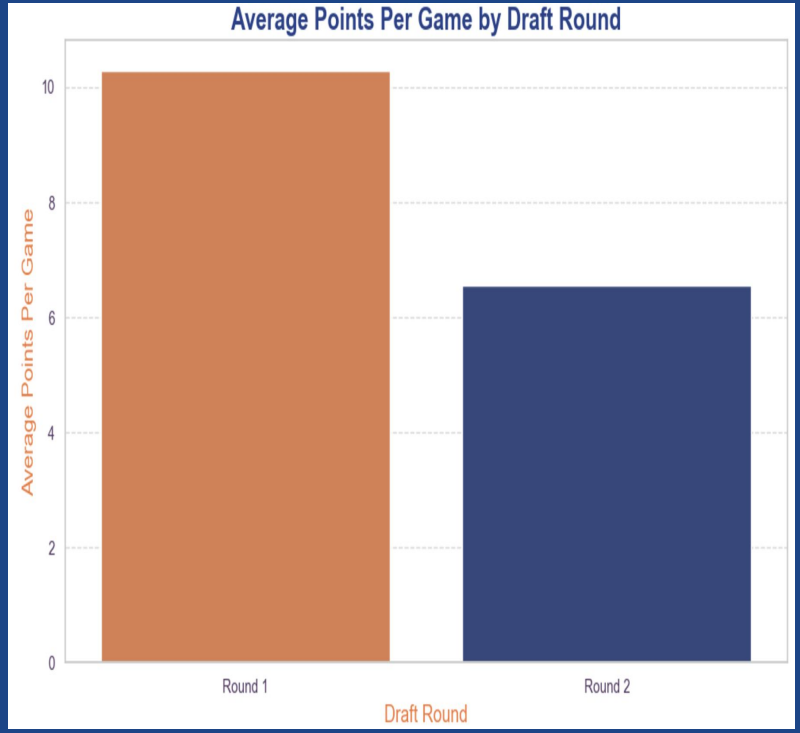




The top 10 universities dominate NBA talent pipelines, with Duke and Kentucky leading in first-round picks and total win shares, highlighting their elite player development. UCLA stands out for producing more second-round picks than other top schools, showcasing its ability to develop impactful players across draft levels.

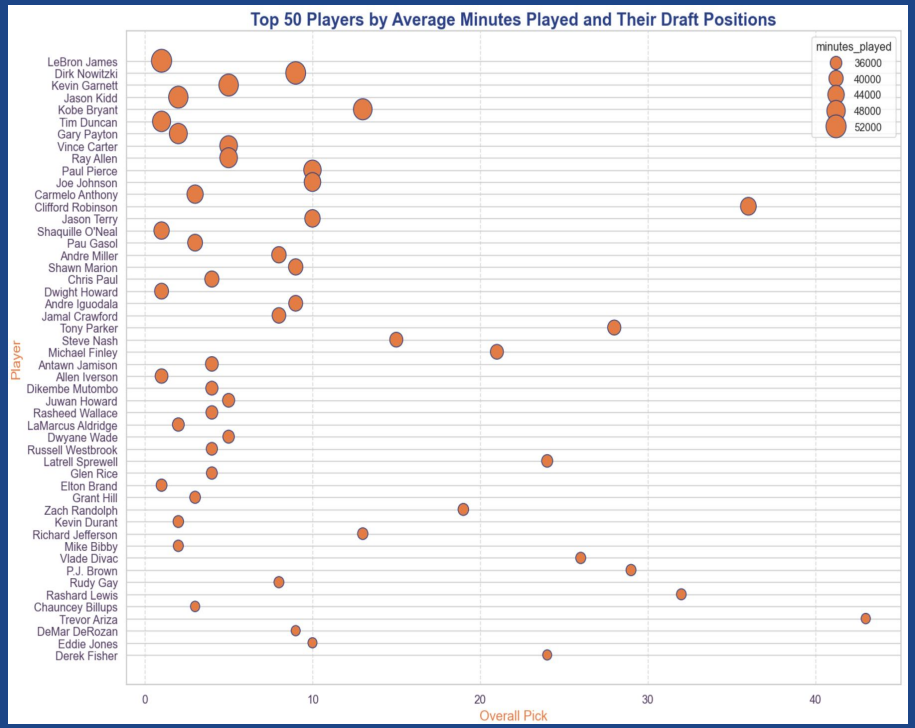
Points Per Game

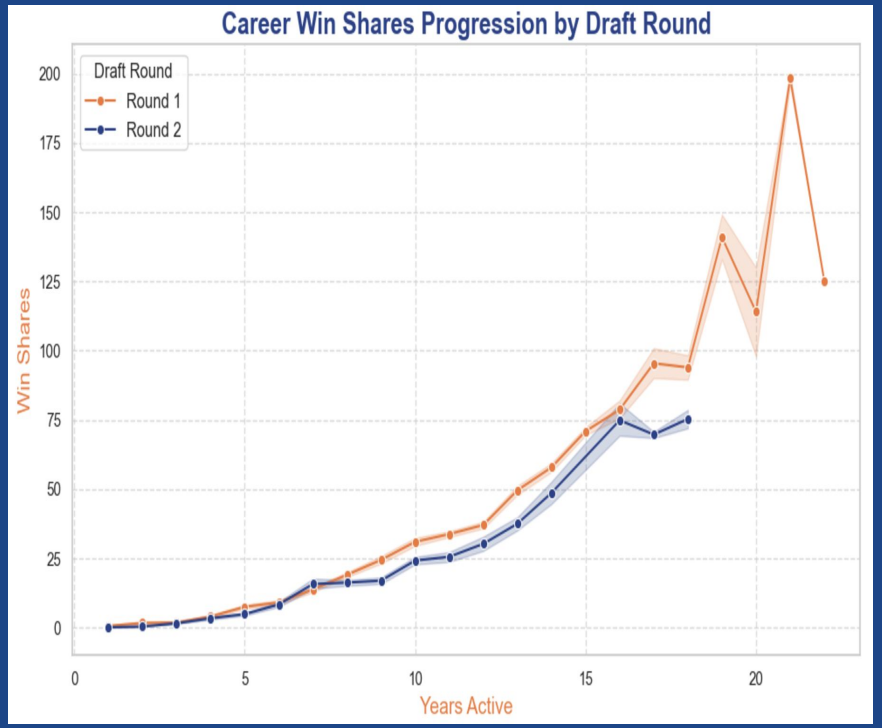




Players selected earlier in the draft (lower overall pick numbers) tend to have higher points per game, with a noticeable decline as the pick number increases. Additionally, first-round draft picks generally have a higher average points per game compared to second-round picks, highlighting the emphasis on early selections for impactful players

Players and Career Progression





Many of the top 50 players by average minutes played are within the first 10 overall draft picks but notable exceptions like Tony Parker and Rashard Lewis show that later picks can still achieve high career minutes. First-round draft picks generally accumulate more win shares over their careers compared to second-round picks, with the gap widening as players stay active longer.

Analysis

University-NBA Performance:

- The top universities, such as Duke, Kentucky, and UCLA, dominate NBA pipelines, producing both high draft picks and impactful players.
    - Duke and Kentucky lead in first-round picks and total win shares, reflecting their ability to develop elite players who succeed in the NBA.
    - UCLA stands out for its significant contribution of second-round picks, demonstrating depth beyond the top tier.
- Additionally, these schools consistently produce players with high career minutes and win shares, showcasing their ability to prepare players for longevity and success in the league.

Draft-NBA Performance:

- Early draft picks generally perform better, with players selected in the lottery and first round showing higher points per game, career minutes, and win shares compared to later picks.
    - However, there are notable exceptions, as some second-round picks and later selections (e.g., Tony Parker and Rashard Lewis) outperform expectations.
- First-round picks consistently accumulate more win shares over time, with the gap widening as players stay active longer, further emphasizing the predictive value of draft position.
    - While later picks show occasional standout performances, the data confirms the importance of early selection for sustained NBA success.

4 Challenges and Solutions

Challenges

PROBLEMS:

- Format Data Conversion to SQL
  - incorrect column types or structures.
- ETL Pipeline Errors
  - Errors with datasets
- Data Alignment
  - Columns were imported as strings instead of integers or floats,

SOLUTION

validated the data before saving, ensuring that all columns adhered to the expected format.

Debugged using small testing and verifying each output

Used pd.to_numeric() to make sure numbers were compatible

Takeaways

Lessons Learned:

- Technical skills
  - Python libraries
  - Data manipulation
- Data Preprocessing
  - Importance in data cleaning to avoid errors
  - ETL design

Takeaways:

- Data-Driven Hypotheses
  - Trends and patterns in the data don't always align with assumptions. (However, our expected insights were close)
- Google Cloud
  - We gained hands-on experience with Google Cloud storage, including setting up buckets and managing access permissions