Data Mining Patterns in WNBA Career Performance and Draft Effectiveness

Elizabeth Serrano
eserran3@students.kennesaw.edu
College of Computing and Software Engineering
Kennesaw State University
Kennesaw, Georgia, USA

Course Information
CS 4412: Data Mining
Spring Semester 2026

Instructor
Professor Christopher Regan
cregan1@kennesaw.edu

Figure 1: WNBA All-Star MVP Performance Highlight.

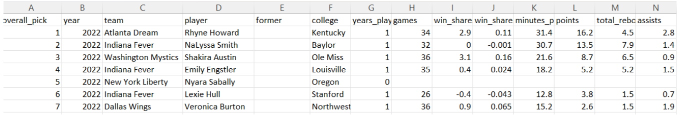
1. Dataset Description

The dataset used for this project is titled *WNBA Draft Player Data Analysis (1997–2022)* and is publicly available on Kaggle at https://www.kaggle.com/code/mattop/wnba-draft-player-data-analysis-1997-2022. The dataset contains 1,065 rows and 14 columns. Each row represents a WNBA draft pick and includes information about the player's draft position (`overall_pick`), draft year (`year`), drafting team (`team`), player name (`player`), former team or league (`former`), college (`college`), years played in the WNBA (`years_play`), total career games played (`games`), career win share (`win_share`), win share per 40 minutes (`win_share_40`), average minutes per game (`minutes_p`), average points per game (`points`), average total rebounds per game (`total_rebc`), and average assists per game (`assists`).

This dataset captures player performance statistics as well as draft and collegiate background information, making it suitable for examining patterns such as career productivity, team drafting success, and player overperformance or underperformance relative to expectations. Known data quality issues include missing values for career statistics for players who did not play, and occasional missing information for college or former teams.

2. Discovery Questions

This project is guided by three primary discovery questions designed to uncover statistical patterns from the data set.



A	B	C	D	E	F	G	H	I	J	K	L	M	N
overall_pick	year	team	player	former	college	years_play	games	win_share	win_share	minutes_p	points	total_rebc	assists
1	2022	Atlanta Dream	Rhyne Howard		Kentucky	1	34	2.9	0.11	31.4	16.2	4.5	2.8
2	2022	Indiana Fever	NaLyssa Smith		Baylor	1	32	0	-0.001	30.7	13.5	7.9	1.4
3	2022	Washington Mystics	Shakira Austin		Ole Miss	1	36	3.1	0.16	21.6	8.7	6.5	0.9
4	2022	Indiana Fever	Emily Engstler		Louisville	1	35	0.4	0.024	18.2	5.2	5.2	1.5
5	2022	New York Liberty	Nyara Sabally		Oregon	0							
6	2022	Indiana Fever	Lexie Hull		Stanford	1	26	-0.4	-0.043	12.8	3.8	1.5	0.7
7	2022	Dallas Wings	Veronica Burton		Northwest	1	36	0.9	0.065	15.2	2.6	1.5	1.9

Figure 2: Sample Dataset

- Which WNBA teams most consistently draft players who achieve successful long-term careers based on career performance metrics (win share, games played, years in league)?
- Which drafted players most significantly overperformed or underperformed career expectations relative to their draft position?
- Which players delivered the greatest career impact per minute played, and what patterns distinguish high-efficiency players from high-volume players?

These questions are valuable because they provide insight into organizational drafting effectiveness, player development, and talent evaluation within the WNBA. By discovering which teams have historically been more successful at identifying players who achieve long-term success can reveal strategic advantages in scouting and player selection. Similarly, identifying players who significantly exceeded or fell short of expectations can highlight trends in draft valuation and the reliability of draft position as an indicator of future performance. Analyzing players based on their efficiency

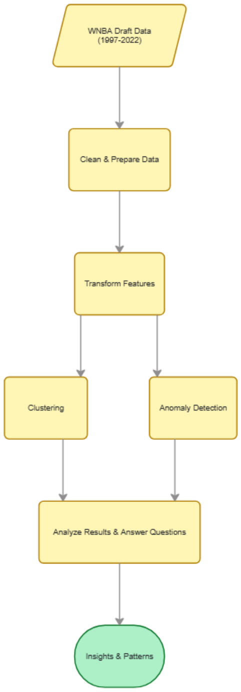


Figure 3: Analysis Pipeline: flowchart

(impact per minute) is important because some players produce strong results even when they are given limited opportunities.

Importantly, these are discovery questions rather than prediction questions because they focus on analyzing and interpreting existing data rather than attempting to predict the future outcomes. This project is focused on discovering insights from existing career performance data by identifying trends and anomalies in past WNBA drafts, rather than developing prediction models.

3. Planned Techniques

For this project, I plan to apply two primary data mining techniques: clustering and anomaly detection. These techniques are well suited to exploring patterns in career performance and identifying meaningful insights within historical WNBA draft data.

Clustering will be used to group players based on similar performance characteristics such as points, assists, rebounds, minutes played, and win share. Techniques such as K-Means or Hierarchical Clustering will help identify natural groupings of players, such as high-impact contributors, role players, and low-impact or limited-opportunity players. This directly relates to my first discovery question by helping determine whether certain teams consistently draft players that fall into higher-performing clusters.

Anomaly Detection will be used to identify players who significantly overperformed or underperformed relative to expectations based on their draft position and career trajectory. This technique supports my second and third discovery questions by highlighting outliers such as late-round picks who became strong contributors or high draft picks who did not meet expectations.

If time allows, I may also explore additional techniques such as dimensionality reduction (PCA) to help visualize player groupings more clearly, but the primary focus will remain on clustering and anomaly detection.

4. Preliminary Timeline

- Week 1 (Jan 12 - Feb 8) - M1: Project Proposal
- Weeks 2-4 (Feb 9 - Mar 8) - M2: Initial Implementation
- Weeks 5-7 (Mar 9 - Apr 5) - M3: Complete Implementation
- Weeks 8-9 (Apr 30 - May 4) - M4: Final Deliverable

Anticipated Challenges

- Cleaning and preprocessing raw data
- Handling missing or incomplete information
- Normalizing and standardizing features
- Choosing appropriate clustering parameters
- Distinguishing true anomalies from normal variation