Chain Reaction: Modeling Basketball Player and Lineup Effectiveness Using Markov Chains

Maria Vazhaeparambil

1. Introduction

This goal of this project is to investigate how individual basketball players and lineup combinations influence the probability of desirable or undesirable possession outcomes, such as scoring, turnovers, fouls, and defensive stops. The central research question that this investigation aims to answer is ”How can Markov chain modeling be used in college basketball in order to quantify and interpret the impact of players and lineups on offensive and defensive possession outcomes?”

This problem is important for several reasons. In the era of advanced basketball analytics, traditional statistics often fail to capture the dynamic and interconnected nature of possessions. They are biased towards what is easy to measure, like points, rebounds, and assists, while smart cuts, effective screens, or lockdown defense rarely translate to the stat sheet. Box score statistics tend to overlook the sequential structure of play and the contextual contributions of each player in a lineup. In addition, chemistry, off-ball effort, and other subtle contributions play a huge role in a team's success, but remain undervalued when teams look at numbers to make decisions about trades, drafts, or game strategies.

A Markov chain framework allows us to model basketball as a sequence of probabilistic events, revealing nuanced effects of players and their combinations on team performance. This has implications for coaching decisions, player development, and lineup optimization. The core theoretical framework assumes that basketball possessions can be modeled as Markov processes, where each action, such as a pass, screen or shot corresponds, to a state, and transitions are influenced by the involved player. This framework hypothesizes the following:

1. Players can be more meaningfully evaluated based on their Markov Chain transition profiles than simple box-score or usage statistics.
2. Markov Chain-based lineup evaluations will correlate more strongly with team success metrics.
3. Decisions made based on intermediate events (plays like Pick and Roll, Drive, Cut) will correlate more strongly with team success.

Ultimately, the analytics in this research should move beyond existing Markov Chain-based metrics by modeling not just simply starts and outcomes of possessions, but more granular play-by-play data that drives transitions between in-possession states, revealing even more accessible insights into player effectiveness, lineup effectiveness, and team strategies. Specifically, this model should give insights from a player perspective on which players are most effective. Additionally, from a lineup perspective this model should answer questions such as which lineup is the most effective in general, which lineup is the most effective when the team needs a 2 Ptr/3 Ptr, and which lineup is the most effective when the team needs a stop. More detailed insights that could be analyzed might include what play should be run when the team needs a 2 Ptr/3 Ptr.

2. Concept Overview

A Markov Chain is a mathematical model that describes a sequence of events in which the outcome of each event depends only on the current state, and not on the sequence of events that preceded it.

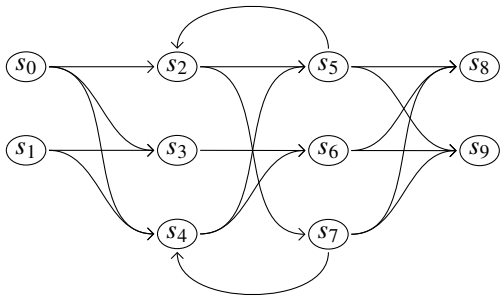

Figure 1. Look into a simplified sample Markov Chain that could be encountered. Each arrow corresponds to a probability of how likely it is to go from the previous state to the next one.

This property is known as the Markov property or memoryless property, and it forms the foundation of the theory behind Markov processes. In formal terms, a Markov Chain is defined by a set of states $S = \{s_1, s_2, \ldots, s_n\}$, a transition probability matrix P, and a starting state distribution.

2.1. States and Transitions

In a Markov Chain, the system is assumed to be in one of a finite set of states at any given time. The transition from one state to another is governed by probabilities that are contained in the transition matrix P, where each element p_{ij} represents the probability of moving from state s_i to state s_j. These probabilities must satisfy the condition that the sum of probabilities from any given state equals 1, i.e.,

$$\sum_j p_{ij} = 1, \quad \forall i.$$

2.2. The Transition Matrix

The **transition matrix** P encapsulates the dynamics of the Markov process. It is a square matrix where each entry p_{ij} gives the probability of transitioning from state s_i to state s_j. Over time, repeated application of the transition matrix allows us to predict the future states of the system. Mathematically, the probability distribution of states after n steps is obtained by multiplying the initial state distribution by the transition matrix raised to the n-th power:

$$\mathbf{P}^n \mathbf{x_0},$$

where $\mathbf{x_0}$ is the initial state distribution.

2.3. Application

Markov chains are well-suited to modeling basketball possessions because they capture the sequential nature of game events — such as passes, shots, fouls, and turnovers — using state transitions. Each state represents a moment in the game, for example a player possessing the ball, a shot attempt, or a change of possession. The Markov property assumes that the next event depends only on the current state.

This framework allows analysts to estimate the probability of outcomes such as scoring or turning the ball over from a given state and to simulate entire possessions or games. It also makes it possible to compute derived metrics like Expected Possession Value (EPV), transition probabilities, and state value functions, which provide insight into team strategies and player effectiveness.

3. Background

3.1. Existing Metrics

3.1.1. Traditional Stats (Box Score)

Traditional box score statistics have long served as the foundation for evaluating individual players and, to a lesser extent, lineups. Metrics such as points, rebounds, assists, steals, field goal percentage, free throw percentage, and three-point percentage offer a straightforward view of a player's contribution during games. While these statistics are easy to understand and widely accessible, they often fail to capture the full context of performance – such as defensive impact or off-ball movement—and can be misleading in isolation. In lineup evaluations, box score stats are typically aggregated or averaged across five-player units, but this method lacks nuance, especially regarding how players interact or influence each other's performance.

3.1.2. Advanced Metrics

Advanced metrics aim to provide a more comprehensive and efficient measure of player and lineup impact by combining and contextualizing traditional statistics. Metrics like Player Efficiency Rating (PER), True Shooting Percentage (TS%), and Win Shares attempt to account for efficiency, usage, and overall value. More sophisticated models like Box Plus/Minus (BPM) and Value Over Replacement Player (VORP) estimate a player's per-possession impact while adjusting for pace and team context. For lineup evaluations, advanced metrics such as Net Rating (offensive rating minus defensive rating) and Expected Possession Value (EPV) from player tracking data are commonly used to assess how effective certain combinations of players are on the court.

3.1.3. On/Off Court Impact

On/off court impact metrics provide insight into a player's value by measuring the difference in team performance when the player is on the floor versus when they are off. These include basic net rating splits as well as more refined methods like Adjusted Plus-Minus (APM), which control for the quality of teammates and opponents. Variants such as Regularized APM (RAPM) and player impact metrics like LEBRON and RAPTOR further improve stability and predictive power. In lineup analysis, similar adjusted plus-minus techniques can be applied to five-man units to isolate the synergistic or detrimental effects of specific lineup combinations, accounting for game context and substitution patterns.

Overall, these techniques provide valuable insights into player and lineup effectiveness but often ignore the structure of possession flow and magnify typical box score statistics.

3.2. Markov Chain

Some studies have started to use Markov chains to address possession flow, such as the following:

3.2.1. Paul Kvam & Joel Sokol – "A Logistic Regression/Markov Chain Model for NCAA Basketball" (2006)

Kvam and Sokol introduce the LRMC (Logistic Regression/Markov Chain) model to rank NCAA basketball teams and predict game outcomes. The model uses logistic regression to estimate transition probabilities between teams, forming a Markov chain that reflects the likelihood of one team defeating another. Applied to NCAA tournament data, the LRMC model outperforms traditional ranking methods, providing more accurate predictions of game results (Kvam & Sokol, 2006).

3.2.2. Kenny Shirley – "A Markov Model for Basketball" (2007)

Kenny Shirley also models basketball as a discrete-time Markov chain, defining states based on possession, method of gaining possession, and points scored in the previous possession. The model comprises up to 30 states, although simplifications reduce this number to 18 in practical applications. Using data from the 2003–2004 NBA season, Shirley estimates transition probabilities and simulates games to predict

outcomes. The model effectively captures team performance, with simulated win percentages that closely match actual results and offered insights into the probabilities of winning in the game and the impact of different types of possession on the scoring. Instead of modeling on team-level outcomes, this uses the Markov chain idea towards individual possessions (Shirley).

3.2.3. Igor Jelaska et al. – "Analysis of Basketball Game States and Transition Probabilities Using the Markov Chains" (2012)

This study expands on the previous analysis, presenting a Markov chain model that segments basketball gameplay into four phases: positional play, transition play, and others. By discretizing the game's continuous flow into distinct states, the model calculates transition probabilities between these phases. The approach helps predict future game states and provides a framework for analyzing team strategies and performance. It moves from overall possession outcomes and matchup outcomes to analyzing the flow and phases, identifying more about the effectiveness of different play styles (Jelaska et al., 2012).

3.2.4. Daniel Cervone et al. – "A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes" (2016)

This study introduced the concept of possession outcome probabilities using Bayesian models, shifting from treating possessions as discrete static states to modeling continuous player movement and event-based transitions. Cervone et al. develop a multiresolution stochastic process model that uses optical tracking data to estimate the Expected Possession Value (EPV) in real time. The model differentiates between continuous player movements and discrete events (e.g. shots, turnovers), using hierarchical spatiotemporal models to estimate transition probabilities. This approach provides detailed insights into player decision-making and offensive strategies, enabling a deeper understanding of possession dynamics and their impact on scoring (Cervone et al., 2016).

3.3. Lineup Evaluation

Although typically most lineups are evaluated based on the previously mentioned traditional statistics, there have been some strides taken towards more complex data-driven methods of approaching lineup optimization.

3.3.1. Winston, Nestler, Pelechrinis - "NBA Lineup Analysis" (2019)

Chapter 32 of Mathletics delves into the quantitative evaluation of basketball lineups using advanced statistical methods. It builds on Adjusted Plus-Minus (APM) and ESPN's Real Plus-Minus (RPM), by focusing specifically on the performance of five-player units rather than individual players. The chapter emphasizes the importance of analyzing entire lineups to understand team dynamics better. By evaluating the combined performance of five-player units, the authors provide insights into how different player combinations contribute to a team's success. It introduces methods to assess lineup effectiveness, considering factors such as offensive and defensive efficiencies, point differentials, and overall impact on game outcomes (Winston et al., 2019).

3.3.2. Lian - "Modeling Offensive Efficiency of NBA Lineups with Bayesian Methods" (2022)

The article explores efficiency of NBA lineups, aiming to determine how different combinations of player roles impact a lineup's scoring performance. The author utilized Principal Component Analysis (PCA) and Spectral Clustering on play-type data from Synergy Sports to categorize NBA players (2017–2022) into seven distinct offensive role clusters. Clusters included roles such as Ballhandlers, Bigs, Wings, and Off-Screen Specialists, each characterized by specific offensive actions and average Offensive Box Plus-Minus (OBPM) scores. They then analyzed lineups based on the composition of these player role clusters, focusing on lineups with at least 100 possessions to ensure statistical significance. Finally, they applied Bayesian inference to compare the offensive efficiency of different lineup configurations, specifically testing hypotheses such as whether lineups with multiple ballhandlers outperform those with "Off-Screen Specialists" or lineups lacking wings (Lian, 2022).

3.3.3. Yichen and Yamashita - "Line-up Optimization Model of Basketball Players and the Simulation Evaluation" (2022)

This paper presents a lineup optimization model for basketball that aims to maximize team performance based on player statistics. It also evaluates different lineup combinations through simulation to validate their effectiveness. The authors use real-world player stats from the NBA, including offensive and defensive performance metrics (e.g., points per game, rebounds, assists, defensive ratings). A RNN model is constructed to identify optimal lineups from a set of available players. After selecting optimal lineups using the model, the authors run Monte Carlo simulations to play out virtual games between teams. They assess which lineups consistently yield higher win probabilities and performance scores.

(Wang & Yamashita, 2022).

4. Data

In order to gain more detailed information within each possession, the constructed Markov Chain will utilize Synergy data for college basketball players. Synergy data includes many standard columns such as Title, Game, Team, Result, Duration, Date, Period, Clock, and Play Type. The most detailed information that can be accumulated from this lies in the "Synergy String," which contains each of the events that happens in that possession. Below are 5 examples of Synergy Strings from both the UConn Women's Basketball team during their 2024 - 2025 season.

Offensive Synergy String
2 KK Arnold > Hand Off > From Dribble > Right > Dribble Jumper > Short to < 17' > Make 2 Pts Off
21 Sarah Strong > Transition > Ballhandler > To Basket > Make 2 Pts Off
2 KK Arnold > Spot-Up > Drives Right > To Basket > Make 2 Pts Off
8 Jana El Alfy > Cut > Flash > Make 2 Pts Off
35 Azzi Fudd > P&R Ball Handler > Left P&R > Side > Dribble Off Pick > Dribble Jumper > Short to < 17' > Miss 2 Pts Off

Defensive Synergy Strings
11 Sonia Citron > Spot-Up > Drives Left > Turnover Def
13 Kate Koval > Cut > Basket > Make 2 Pts Def
3 Hannah Hidalgo > P&R Ball Handler > High P&R > Dribble Off Pick > To Basket > Make 2 Pts Def
3 Hannah Hidalgo > Spot-Up > Drives Left > Foul Def
5 Olivia Miles > P&R Ball Handler > High P&R > Dribble Off Pick > Defense Commits > Turnover Def

Each of the events within the possession can be considered a state in the model and be used to probabilistically infer how likely each final state would be of the following outcomes for both offense and defense: Made 2 Pointer, Missed 2 Pointer, Made 3 Pointer, Missed 3 Pointer, Made Free Throw, Missed Free Throw, Turnover, and Foul.

5. Methodology

To quantify the effectiveness of individual players within possession sequences, offensive and defensive play progressions are modeled using a Markov Chain framework. Each possession is represented as a series of discrete events, captured from play-by-play data in the form of Synergy Strings transitions. These sequences are parsed into state transitions, where each state is defined as a player-specific action, and terminal states are labeled as final outcomes. These terminal outcomes are treated as absorbing states in the Markov model.

5.1. Transition Matrix

We construct a weighted transition matrix where each edge between states is incremented based on a possession-specific weight. This weight is computed using a decaying exponential function of the game's point differential, giving higher significance to actions taken in closer games. Transition frequencies are aggregated across all observed possessions, normalized to yield probabilities, and used to form the transition matrix T. From this, we isolate transient and absorbing states to extract the submatrices Q and R. The fundamental matrix $N = (I - Q)^{-1}$ is computed to model the expected number of visits to

transient states before absorption. The matrix of absorption probabilities $B = N\dot{R}$ is then used to calculate the probability of ending a possession in each absorbing state given a starting transient state.

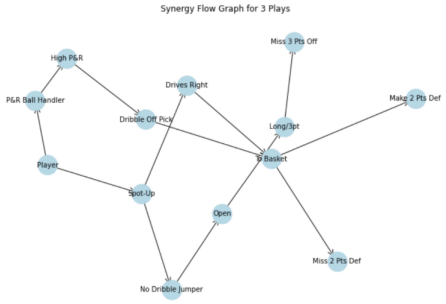

Figure 2. Example of 3 plays with edges drawn between states.

5.2. Metric Design

In Dean Oliver's book *Basketball on Paper*, he discusses the statistics behid how both the offensive and defensive sides of the game contribute to winning and breaks down the performance of a team over time based on empirical analysis. His book aims to create a holistic view of what drives success in basketball.

To evaluate offensive effectiveness, this study applies *Dean Oliver's Four Factors*, a well-established framework that decomposes offensive success into four key components, each assigned a relative weight based on empirical importance:

- **Shooting (40%)**: Represented by the effective field goal percentage (eFG%), which accounts for the added value of three-point shots:
$$\text{eFG\%} = \frac{\text{FG} + 0.5 \times \text{3P}}{\text{FGA}}$$
- **Turnovers (25%)**: Measured by turnover rate (TOV%), which captures the frequency of turnovers relative to overall possession-ending events:
$$\text{TOV\%} = \frac{\text{TOV}}{\text{FGA} + 0.44 \times \text{FTA} + \text{TOV}}$$
- **Offensive Rebounding (20%)**: Quantified by offensive rebound percentage (ORB%), which estimates a team's ability to retain possession after missed shots:
$$\text{ORB\%} = \frac{\text{ORB}}{\text{ORB} + \text{Opponent DRB}}$$
- **Free Throws (15%)**: Evaluated using free throw rate, measuring points generated from the free throw line per field goal attempt:
$$\text{FT\%} = \frac{\text{FT}}{\text{FGA}}$$

To assess defensive contribution, this analysis incorporates Oliver's concept of *Defensive Rating*, which centers on the estimation of "stops"—instances where a defender successfully ends the opponent's possession. These include forced turnovers, contested missed shots, and other defensive actions not fully captured by traditional box score metrics such as steals or blocks. The total number of stops is modeled as:

$$\text{Stops} = \text{Stops}_1 + \text{Stops}_2$$

where:

$$\text{Stops}_1 = \text{STL} + \text{BLK} \times \text{FM}_{\text{wt}} \times (1 - 1.07 \times \text{DOR\%}) + \text{DRB} \times (1 - \text{FM}_{\text{wt}})$$

$$\text{Stops}_2 = \left(\frac{\text{Opponent FGA} - \text{Opponent FGM} - \text{Team BLK}}{\text{Team MP}} \times \text{FM}_{\text{wt}} \times (1 - 1.07 \times \text{DOR\%}) \right.$$

$$\text{Stops}_2 = \left(\frac{\text{Opponent FGA} - \text{Opponent FGM} - \text{Team BLK}}{\text{Team MP}} \times \text{FM}_{\text{wt}} \times (1 - 1.07 \times \text{DOR\%}) \right.$$
$$\left. + \frac{\text{Opponent TOV} - \text{Team STL}}{\text{Team MP}} \right) \times \text{MP} + \left(\frac{\text{PF}}{\text{Team PF}} \right) \times 0.4 \times \text{Opponent FTA} \times \left(1 - \frac{\text{Opponent FTM}}{\text{Opponent FTA}} \right)^2 \tag{1}$$

with the following intermediate terms defined as:

$$\text{FM}_{\text{wt}} = \frac{\text{DFG\%} \times (1 - \text{DOR\%})}{\text{DFG\%} \times (1 - \text{DOR\%}) + (1 - \text{DFG\%}) \times \text{DOR\%}}$$

$$\text{DOR\%} = \frac{\text{Opponent ORB}}{\text{Opponent ORB} + \text{Team DRB}}, \quad \text{DFG\%} = \frac{\text{Opponent FGM}}{\text{Opponent FGA}}$$

This formula attempts to quantify not only a player's observable defensive actions (e.g., steals, blocks, defensive rebounds) but also their indirect impact on missed shots and forced turnovers, incorporating contextual team data for accurate scaling.

This formulation captures both observable and inferred defensive impacts. Oliver also posits that basketball success is approximately 60% offense and 40% defense, a principle that guides the balanced weighting of offensive and defensive contributions in this model (Oliver, 2004).

5.3. Aggregation

After determining probabilities for expected outcomes by mapping them using the absorption probability matrix, we are left with an expected impact estimate for each transient state. These values are then averaged across all tracked possessions in which a player appears, producing a per-player expected impact contribution metric.

In addition to analyzing expected points contributed by individual players, we extended our Markov chain framework to evaluate player duos and entire lineups. Because lineup is was not explicitly provided, we inferred pseudo-lineups by maintaining a rolling window of the five most recent unique players involved in each possession. Each possession was then labeled with a "lineup key" representing this inferred group. Using the same event-based transition modeling approach, we constructed transition counts and built a transition probability matrix for each duo within these lineups and the lineups themselves. Absorption probabilities into final scoring outcomes were calculated from this matrix, and expected points were computed based on those same probabilities and assigned point values for each terminal state. We then aggregated expected points across all possessions associated with each lineup, providing a ranking of the most effective duos and lineups observed in the dataset.

5.4. Regression Modeling

To complement the Markov chain analysis of possession-level outcomes, we implemented a regression-based approach to evaluate and predict the offensive performance of both observed and unobserved lineups. This

Algorithm 1 Estimate Player/Lineup Effectiveness via Inferred Transitions

```
Initialize transition counts
for each possession in play-by-play data do
    Get player/duo/lineup involved
    Assign player/duo/lineup state labels for each event in possession
    Increment transition counts between states, weighted by point differential
end for
Construct transition matrix T and extract submatrices Q and R
Compute fundamental matrix N = (I − Q)^{-1}
Calculate absorption probabilities B = N · R
Multiply B with point value vector to get expected points per transient state
Compute rating based on probabilities of outcomes by player/duo/lineup
```

approach aimed to quantify how much individual player skills and pairwise synergies contribute to overall lineup effectiveness.

For each lineup, we constructed a 15-dimensional feature vector composed of two components: (1) the five individual offensive ratings (`OffRtg`) of the players in the lineup, and (2) ten pairwise offensive synergy ratings derived from a duo-level rating dataset. All individual and duo ratings were sorted prior to modeling to ensure consistency across equivalent lineups regardless of player order.

Let $x \in \mathbb{R}^{15}$ represent the input feature vector for a given lineup, and $y \in \mathbb{R}$ represent the observed offensive rating of that lineup. We fit a Ridge regression model of the form:

$$\hat{y} = \beta_0 + \sum_{i=1}^{15} \beta_i x_i$$

where β_0 is the intercept, β_i are the learned coefficients, and an L_2 regularization penalty is applied to mitigate overfitting. Specifically, the model minimizes the following objective:

$$\min_{\beta} \left\{ \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 + \alpha \sum_{i=1}^{15} \beta_i^2 \right\}$$

where α controls the strength of regularization.

The model was trained using an 80/20 train-test split of the lineup dataset, and performance was evaluated using mean squared error (MSE) on the test set. The trained model not only provided insight into the marginal contributions of individual players and pairwise combinations but also enabled generalization to unobserved lineups not present in the original data. This generalizability is particularly useful for lineup optimization and strategic decision-making in settings where not all combinations are empirically tested.

5.5. Assumptions of the Markov Chain Model

Applying a Markov chain framework to model basketball possessions relies on several critical assumptions that must be acknowledged for the integrity of the analysis.

First and foremost, the model assumes the Markov property, meaning the next state of a possession depends only on the current state and not on the sequence of events that preceded it. In practice, this implies that player decisions, defensive adjustments, or momentum effects from previous plays are not directly incorporated into the transition probabilities.

Second, the chain is assumed to be time-homogeneous, with transition probabilities remaining fixed over time. This simplifies modeling but may not reflect the dynamic nature of possessions that change based

on game context (e.g., travel, fatigue, hot hands, or lineup substitutions).

Third, the model presumes a well-defined state space with clear classifications of transient states (e.g., ball-handler actions, screens, cuts) and absorbing states (e.g., made shot, turnover, foul). This discretization is necessary for computation but inherently abstracts away continuous or nuanced in-game decision-making.

While these assumptions provide mathematical tractability and enable insightful estimation of expected values, they also limit the expressive power of the model. Future work may explore semi-Markov or context-dependent extensions to more accurately reflect the complexity of in-game basketball strategy.

6. Results

For all evaluative methods, we analyzed the regular results from both the Duke Men's Basketball team and the UConn Women's Basketball team from the 2024 - 2025 season.

6.1. Player Effectiveness

For the player analysis, the result is a ranking of players based not solely on raw statistics, but on their modeled contribution to scoring outcomes through possession flow.

	eFG	TOV	FT	ORB	OffRtg	Stops1	Stops2	DefRtg	Rtg
Paige Bueckers	0.6711	0.9309	0.9792	0.7087	0.7898	0.1470	0.2646	0.2717	0.5825
Sarah Strong	0.6725	0.8955	0.1845	0.7244	0.6654	0.1750	0.2764	0.3687	0.5467
KK Arnold	0.7953	0.8521	0.3462	0.4567	0.6744	0.2788	0.2723	0.2795	0.5164
Ashlynn Shade	0.5848	0.9154	1.0000	0.4724	0.7073	0.1367	0.2491	0.1754	0.4945
Azzi Fudd	0.5782	0.9617	1.0000	0.5276	0.7272	0.0963	0.2364	0.1304	0.4885
Ice Brady	0.7140	0.8672	1.0000	0.3071	0.7138	0.1379	0.2895	0.1479	0.4875
Morgan Cheli	0.7049	0.8901	1.0000	0.1417	0.6828	0.2270	0.2886	0.1379	0.4649
Kaitlyn Chen	0.7402	0.8654	0.1967	0.5039	0.6427	0.1044	0.2653	0.1788	0.4571
Jana El Alfy	0.5616	0.8833	0.3373	0.4724	0.5905	0.1707	0.2595	0.2441	0.4520
Aubrey Griffin	0.7127	0.8413	1.0000	0.1339	0.6722	0.3320	0.2195	0.0773	0.4342
Qadence Samuels	0.4432	0.5418	1.0000	0.0866	0.4800	0.5700	0.3216	0.1122	0.3329
Allie Ziebell	0.3230	0.9225	0.0000	0.1654	0.3929	0.2473	0.2507	0.0590	0.2593
Caroline Ducharme	0.4797	0.7137	0.0000	0.0157	0.3735	0.3790	0.0000	0.0051	0.2261

Table 1. Player Ratings

Top Performers. **Paige Bueckers** leads all players with the highest composite rating (`Rtg` = 0.5825), driven by her elite shooting efficiency (eFG% = 0.6711), extremely low turnover rate (TOV% = 0.9309), and strong contributions in all Four Factors. Defensively, she posts solid numbers in both Stops_1 and Stops_2, suggesting consistent impact beyond traditional steals and blocks. As the number 1 pick of the draft this year, this is what we would expect from a player of her caliber.

Sarah Strong and **KK Arnold** follow closely behind, as would expect based on their performances this year. Sarah Strong pairs excellent shooting and rebounding with low turnovers, while KK Arnold boasts the highest eFG% in the dataset (0.7953), indicating exceptional shot selection and finishing. Both also contributes heavily on the defensive end, with strong Stops_1 and Stops_2 values, reflecting her activity forcing misses and turnovers.

Specialists and Balance. **Azzi Fudd** and **Ashlynn Shade** show strong offensive profiles, particularly in free throw efficiency. However, their lower defensive ratings (especially individually being able to get blocks and steals) slightly dampen their overall value, which aligns with their minutes this past season.

Lower-Rated Players. **Qadence Samuels**, **Allie Ziebell**, and **Caroline Ducharme** appear at the lower end of the ratings distribution. All have gotten limited minutes this past season, so this is expected.

Overall Patterns. As expected, players with balanced contributions across both ends of the floor achieve the highest composite ratings. The data reinforces that shooting efficiency (eFG%) and turnovers (TOV%) are the most influential offensive factors, consistent with Dean Oliver's weighting. On defense, Stops_2 often distinguishes players who consistently disrupt possessions beyond traditional box score stats.

These results support the validity of the player evaluation framework and provide actionable insights for lineup construction, substitution decisions, and talent development focus areas.

6.2. Player Clustering

KMeans Clustering is applied to players' absorption profiles to look into possible archetypes that emerge and test hypotheses about diversity and synergy. Principal Component Analysis (PCA) was conducted to

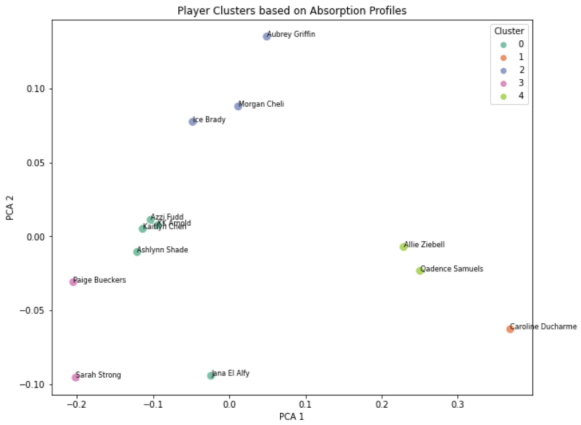

Figure 3. PCA analysis of UConn women's players transitions.

reduce the dimensionality of the player rating metrics and reveal underlying patterns in player contributions. The resulting projection revealed distinct clusters within the player pool. Notably, Paige Bueckers and Sarah Strong formed a unique cluster, indicating their dominant and well-rounded impact across offensive and defensive metrics. A second, tightly grouped cluster included high-impact rotational players such as Kaitlyn Chen, Azzi Fudd, Ashlynn Shade, and Jana El Alfy, who each contribute meaningfully in specific dimensions of play. A third cluster captured players with more limited but consistent minutes, reflecting moderate contributions across fewer areas. Finally, bench players who received minimal playing time formed a separate cluster, largely driven by low statistical engagement across the measured features. These results support the role-based stratification of the roster and validate the broader rating framework.

6.3. Duo Effectiveness

Using the duo analysis, we can rank lineups based on their modeled impact on possession outcomes, capturing how effective each pair of players is at driving sequences toward both scoring and defending.

The highest-rated two-player combination is **Paige Bueckers and Sarah Strong** (`Rtg` = 0.3381), combining elite offensive execution (`OffRtg` = 0.4382) with strong defensive synergy (`DefRtg` = 0.1881). Given that both players individually rank among the top three in overall player rating, it is unsurprising that their pairing also yields the most impactful duo score. Their ability to complement each other—Bueckers with high-efficiency shot creation and Strong with rebounding and defensive versatility—makes this combination highly effective across both ends of the floor.

The duo of **KK Arnold and Sarah Strong** follows closely with a composite rating of 0.2868. KK Arnold brings elite finishing ability (highest individual eFG%) and aggressive on-ball defense, while Sarah Strong continues to contribute in nearly every offensive metric. Together, they form a dynamic and balanced backcourt-frontcourt pair.

Pair	OffRtg	DefRtg	Rtg
Paige Bueckers, Sarah Strong	0.4382	0.1881	0.3381
KK Arnold, Sarah Strong	0.3784	0.1493	0.2868
Kaitlyn Chen, Sarah Strong	0.3876	0.1178	0.2797
KK Arnold, Paige Bueckers	0.3879	0.1094	0.2765
Kaitlyn Chen, Paige Bueckers	0.3959	0.0826	0.2706
Azzi Fudd, Sarah Strong	0.3797	0.0895	0.2636
Azzi Fudd, Paige Bueckers	0.3908	0.0663	0.2610
Jana El Alfy, Sarah Strong	0.3351	0.1372	0.2559
Jana El Alfy, KK Arnold	0.3746	0.0712	0.2533
Ashlynn Shade, KK Arnold	0.3672	0.0811	0.2527

Table 2. Duo or Pair Ratings

Kaitlyn Chen and Sarah Strong and **KK Arnold and Paige Bueckers** also emerge as top-tier combinations. Chen's strong individual shooting efficiency and low turnover rate pair well with Strong's interior presence. Meanwhile, the combination of Arnold and Bueckers excels on offense (`OffRtg` = 0.3879) and is especially potent in transition actions, both key areas in modern offensive schemes.

Several other duos, including **Azzi Fudd paired with either Strong or Bueckers**, demonstrate high offensive synergy, reflecting Fudd's value as a perimeter shooter and off-ball mover. However, slightly lower defensive ratings compared to other top duos limit their composite scores.

Lastly, the presence of **Jana El Alfy** and **Ashlynn Shade** in the top 10 highlights their adaptability when paired with primary ballhandlers like Strong and Arnold. El Alfy's off-ball cutting and Shade's spot-up shooting complement more dominant creators, enabling efficient possessions even without commanding high usage themselves.

These results validate the lineup construction logic and suggest that optimal lineup strategies should emphasize pairing complementary skill sets across positions and play styles—such as combining high-efficiency guards with rebounding forwards or pairing shooters with slashers.

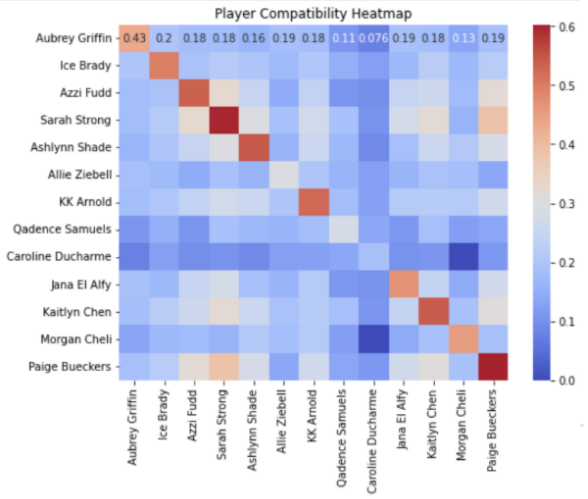

Figure 4. Compatability matrix UConn women's players pairs.

6.4. Lineup Effectiveness

Using the lineup analysis, we can rank lineups based on their modeled impact on possession outcomes, capturing how effective each group of players is at driving sequences toward both scoring and defending.

Lineup	OffRtg	DefRtg	Rtg
Azzi Fudd, Jana El Alfy, Kaitlyn Chen, Paige Bueckers, Sarah Strong	0.4248	0.0039	0.2564
Ashlynn Shade, Ice Brady, Kaitlyn Chen, Paige Bueckers, Sarah Strong	0.3322	0.0000	0.1993
Ashlynn Shade, Azzi Fudd, Kaitlyn Chen, Paige Bueckers, Sarah Strong	0.3077	0.0006	0.1849
Ashlynn Shade, KK Arnold, Kaitlyn Chen, Paige Bueckers, Sarah Strong	0.3026	0.0070	0.1844
Ashlynn Shade, Azzi Fudd, KK Arnold, Paige Bueckers, Sarah Strong	0.3042	0.0042	0.1842
Azzi Fudd, KK Arnold, Kaitlyn Chen, Paige Bueckers, Sarah Strong	0.3015	0.0017	0.1816
Azzi Fudd, Ice Brady, Jana El Alfy, KK Arnold, Sarah Strong	0.2948	0.0019	0.1777
Jana El Alfy, KK Arnold, Morgan Cheli, Paige Bueckers, Sarah Strong	0.2834	0.0014	0.1706
Jana El Alfy, KK Arnold, Kaitlyn Chen, Paige Bueckers, Sarah Strong	0.2763	0.0116	0.1704
Ashlynn Shade, Azzi Fudd, Ice Brady, KK Arnold, Sarah Strong	0.2685	0.0000	0.1611

Table 3. Lineup Ratings

The highest-rated lineup in the dataset—**Azzi Fudd, Jana El Alfy, Kaitlyn Chen, Paige Bueckers, and Sarah Strong**—achieves a composite rating (`Rtg`) of 0.2564. This group is anchored by elite offensive execution (`OffRtg` = 0.4248) and near-flawless defensive results (`DefRtg` = 0.0039). The blend of strong floor-spacing (Fudd), interior activity (El Alfy), facilitating (Chen), and all-around elite performance (Bueckers and Strong) allows for dynamic versatility across all phases of play. Unsurprisingly, this is also Coach Geno Auriemma's choice for the starting lineup.

Multiple other top lineups revolve around a similar core, and the presence of **KK Arnold** in multiple high-performing lineups—particularly in combinations with Bueckers and Strong—confirms her ability to amplify transition scoring and on-ball defense. Overall, these results reinforce that optimal lineup construction benefits from a combination of elite individual talents and facilitators (Bueckers, Strong, Chen), efficient complementary scorers (Fudd, Arnold), and defensively sound supporting players (Brady, El Alfy, Shade). The best units leverage floor balance, decision-making, and two-way consistency, rather than relying solely on individual scoring.

6.5. Situation Specific Lineups

When optimizing lineups based on outcome-weighted preferences—specifically favoring possessions that end in either a made 2-point or 3-point field goal—the model identifies two distinct optimal configurations.

For maximizing **2-point scoring outcomes**, the most effective lineup comprises **Azzi Fudd, Ice Brady, Jana El Alfy, Kaitlyn Chen, and Paige Bueckers**. This group features a higher concentration of frontcourt players, with **Brady** and **El Alfy** providing interior presence and finishing ability, and **Fudd** and **Chen** capable of penetrating or facilitating inside opportunities. The inclusion of **Bueckers**, a versatile scorer with strong midrange and rim-finishing efficiency, further enhances this group's ability to generate high-quality looks at the basket. The lineup's composition reflects a deliberate emphasis on interior spacing, offensive rebounding, and rim gravity, ideal for maximizing shot attempts close to the hoop.

In contrast, the lineup optimized for **3-point shot success** consists of **Azzi Fudd, Jana El Alfy, Kaitlyn Chen, Paige Bueckers, and Sarah Strong**. This unit reduces frontcourt presence slightly by removing a traditional post like Ice Brady in favor of a more additional perimeter-oriented post in **Sarah Strong**. The result is a lineup with elite spacing, multiple long-range threats, and excellent ball movement. **Fudd**, **Bueckers**, and **Strong** are all capable of high-efficiency spot-up shooting, while **Chen** facilitates off the dribble and creates open looks for o

6.6. Game Situation Insights

As anticipated, the two-point field goal attempt with the highest probability of success was the shot taken at the basket. This outcome aligns with conventional basketball analytics, which consistently highlight the high efficiency of close-range attempts.

Analysis of the play types most frequently leading to successful two-point field goals revealed that actions such as *Dribble Off Pick*, *Face-Up*, *Cut*, and *Pick-and-Pop* were among the most effective. These plays are typically designed to create space or mismatch opportunities, facilitating cleaner paths to the basket.

Conversely, the two-point shot type most likely to result in a miss was the jump shot, particularly from the midrange or "long two" distance (approximately 17 feet). This finding is consistent with existing literature that considers midrange jumpers to be among the least efficient shot types due to their relatively low expected value.

For three-point field goals, *Spot-Up* attempts exhibited the highest success rate. Regarding play types, the most common actions leading to three-point shot attempts included *Pick-and-Pop*, *High Pick-and-Roll*, and *Off-Screen* movement. These plays typically generate open looks for perimeter shooters by leveraging screens and spacing principles.

Future Work

While this study presents a data-driven approach to evaluating player and lineup effectiveness through Markov models and regression-based insights, several opportunities exist for future exploration and refinement.

Validation and Practical Application. A critical next step is the validation of these models through real-world outcomes. While the current analysis identifies high-performing players, duos, and lineups based on estimated scoring transitions and ratings, applying these insights to actual game decisions requires further testing. Future work should explore how these analytics translate to in-game success, particularly during high-leverage postseason scenarios where sample sizes become large enough to offer reliable inference. Some initial work I have done to look into this involves separating the regular season and post season results for numerous teams. Then, I do Markov chain lineup analysis on just the regular season. Finally, I attempt to evaluate a team's post-season based on the rating that each lineup used had during the regular season. Here are some preliminary results: Markov Ratings, which account for the probabilistic flow of

NCAA March Madness	Team	Transformed Markov_Rtg_Avg	PlusMinus_Avg
Champions	Connecticut Huskies	0.21840	0.2132
Runner-Up	South Carolina Gamecocks	0.13319	0.1222
Elite Eight	LSU Tigers	0.133005	0.0576
Final Four	Texas-(Austin) Longhorns	0.11714	0.0934
Sweet Sixteen	North Carolina State Wolfpack	0.11351	-0.0137
Sweet Sixteen	North Carolina Tar Heels	0.109325	0.1034
Sweet Sixteen	Oklahoma Sooners	0.096315	0.1242
Sweet Sixteen	Mississippi Rebels	0.074570	0.0867
Final Four	UCLA Bruins	0.073105	0.0673
Elite Eight	USC Trojans	0.071815	0.0853
Sweet Sixteen	Tennessee Lady Volunteers	0.066520	0.0617
Elite Eight	Duke Blue Devils	0.063935	0.0935
Sweet Sixteen	Notre Dame Fighting Irish	0.057495	0.1129
Sweet Sixteen	Kansas State Wildcats	0.01242	0.1357
Sweet Sixteen	Maryland Terrapins	0.007085	0.0263
Elite Eight	TCU Horned Frogs	0.005925	0.1069

Table 4. Team Ratings with Transformed Markov Rating

possessions and scoring transitions, aligned more closely with actual NCAA tournament outcomes than raw plus-minus metrics. Teams with higher Markov Ratings—like Connecticut and South Carolina—tended to advance further or perform above seed expectations, indicating that possession-level efficiency may be a stronger predictor of postseason success than average point differentials alone. However, these are still preliminary results and require more extensive exploration.

Leveraging Historical Data. Due to the limited availability of possession-level data within a single season, especially for college teams, in-season application of these methods may be constrained. As such, aggregating and integrating multi-year data sets could strengthen model robustness and allow for the generalization of insights across player cohorts and game contexts. Techniques for adjusting historical data to align with current rosters and play styles would enhance predictive power.

Extending Regression Approaches. The ridge regression model used to estimate unseen lineup performance presents a promising foundation for evaluating lineup synergies. Further exploration into advanced regularization techniques (e.g., Lasso, Elastic Net) or non-linear models (e.g., gradient boosting, random forests) could improve predictive accuracy. Additionally, incorporating interaction terms or latent synergy variables may help uncover non-obvious lineup strengths.

Situational and Contextual Analytics. Current models evaluate possessions agnostic to game context. Future work could investigate how lineup performance varies with situational factors such as score differential, time remaining, or opponent scheme. While individual-level data for these micro-contexts remains sparse, approximation through clustering or simulation could provide directional insights. Ultimately, this would make analytics more actionable by linking lineup decisions to specific game states.

Collectively, these directions aim to bridge the gap between theoretical evaluation and real-time decision-making, advancing the use of analytics in strategic planning and roster optimization.

References

Cervone, D., D'Amour, A., Bornn, L., and Goldsberry, K. A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association*, 111(514):585–599, April 2016. ISSN 1537-274X. doi: 10.1080/01621459.2016.1141685. URL http://dx.doi.org/10.1080/01621459.2016.1141685.

Jelaska, I., Trninić, S., and Perica, A. Analysis of basketball game states and transition probabilities using the markov chains. *Analysis of Basketball Game States and Transition Probabilities Using the Markov Chains*, 66:15–24, 01 2012. doi: 10.5937/fizkul1201015J.

Kvam, P. and Sokol, J. S. A logistic regression/markov chain model for NCAA basketball. *Nav. Res. Logist.*, 53(8):788–803, December 2006.

Lian, X. Modeling offensive efficiency of nba lineups with bayesian methods. https://medium.com/@xulianrenzoku/modeling-offensive-efficiency-of-nba-lineups-with-bayesian-methods-55c695e95e1d, August 2022. Accessed: 2025-05-14.

Oliver, D. *Basketball on Paper: Rules and Tools for Performance Analysis*. Potomac Books, Inc., Washington, D.C., 2004. ISBN 9781597979514.

Shirley, K. A markov model for basketball.

Wang, Y. and Yamashita, H. Line-up optimization model of basketball players and the simulation evaluation. In *Proceedings of the 11th International Conference on Data Science, Technology and Applications (DATA 2022)*, pp. 492–499. SCITEPRESS – Science and Technology Publications, Lda., 2022. ISBN 978-989-758-583-8. doi: 10.5220/0011307900003269.

Winston, W. L., Nestler, S., and Pelechrinis, K. Nba lineup analysis. In *Mathletics*, chapter 32, pp. 289–295. Princeton University Press, Princeton, NJ, 2nd edition, 2019.

Appendix: Terminology

- 3P: Number of 3-Pointers Made
- BLK: Number of Blocks
- DRB: Number of Defensive Rebounds
- FG/FGM: Number of Field Goals Made
- FGA: Number of Field Goals Attempted
- FT/FTM: Number of Free Throws Made
- FTA: Number of Free Throws Attempted
- MP: Minutes Played
- ORB: Number of Offensive Rebounds
- PF: Number of Personal Fouls
- STL: Number of Steals
- TOV: Number of Turn Overs