SpatialSportsR Analysis — Kaleb Coleman

Kaleb Coleman

Abstract

This paper presents a reproducible, end-to-end examination of NBA offensive strategy and player performance. We detail a comprehensive shot dataset compiled via a custom data pipeline spanning the 2014-15 through 2025-26 seasons. Using this data, we develop an expected field goal (xFG) model via logistic regression, achieving 63% accuracy with 0.653 AUC-ROC. Building on this foundation, we deliver three advanced analytics products: (1) Residual Analysis to identify players who over/underperform expectations, (2) a Shot Difficulty Index (SDI) quantifying shot complexity, and (3) K-Means clustering to identify player shot archetypes. We then integrate salary data to contextualize performance via POE per $1M. Results highlight Nikola Jokić as a top overperformer (+12.9% residual) and surface distinct shot-profile archetypes (e.g., rim-heavy and high-3PT profiles).

Introduction

The National Basketball Association (NBA) has undergone a profound data revolution, fundamentally altering strategies and analysis. This paper addresses the challenge of translating shot data into actionable insights by leveraging a comprehensive, custom-built NBA shot dataset spanning the 2014-15 through 2025-26 seasons.

Contributions

- A reproducible data pipeline consolidating NBA shot data into an analysis-ready SQLite database.
- An expected field goal (xFG) model using logistic regression with 63% accuracy.
- Three advanced analytics products: Residual Analysis, Shot Difficulty Index, and Player Archetypes.
- A salary-adjusted value layer (POE per $1M) for performance context.
- League-wide shot selection summaries including efficiency metrics and attempt distribution.

Paper Roadmap

Section 2 reviews relevant background. Section 3 details methods. Section 4 reports results including the xFG model and advanced analytics. Section 5 discusses findings and limitations.

Related Work

Prior research on basketball analytics has emphasized the relationship between shot location and scoring efficiency. This study builds on that work by providing a fully reproducible xFG model and extending it with residual analysis, shot difficulty quantification, and player clustering.

Methods

Data Sources and Pipeline

The dataset was compiled through a custom R package interfacing with the NBA Stats API and ESPN NBA API. Raw JSON data was parsed, cleaned, and stored in a SQLite database (`nba.sqlite`) containing 2.5M+ shots across 12 seasons. (*NBA Stats API*, n.d.; *ESPN NBA API*, n.d.)

Pipeline overview

1. **Collect**: `collect_raw()` downloads game-level JSON from ESPN and NBA Stats with caching and rate-limit handling.
2. **Parse**: `parse_raw()` normalizes raw JSON into standardized tables (games, events, shots, box scores, play-by-play).
3. **Validate**: `validate_tables()` enforces schema and key constraints to ensure referential integrity.
4. **Write**: `write_tables()` and `write_sqlite_from_rds()` persist data to `data/parsed/` and a consolidated SQLite file.

Analysis scripts

- `expected_points_analysis.py`: trains the xFG model and writes `shots_with_xp_*.parquet`
- `advanced_analytics.py`: residuals, SDI, and player archetype clustering
- `player_performance_analysis.py`: POE leaderboards and player shot charts
- `shot_density.py`: league-wide shot density heatmaps
- `salary_collector.py`: player salary data from Basketball-Reference (*Basketball-Reference Player Contracts*, n.d.)
- `value_analysis.py`: POE per $1M salary rankings

xFG Model

The expected field goal model uses logistic regression with the following features:

- **Spatial**: LOC_X, LOC_Y, shot distance, shot angle
- **Shot Type**: layup, dunk, jump shot, hook, floater indicators
- **Context**: period, seconds remaining, clutch indicator
- **Zones**: SHOT_ZONE_BASIC, SHOT_ZONE_AREA (one-hot encoded)

The model was trained on 80% of 2025-26 regular season shots and evaluated on the held-out 20%.

Shot Difficulty Index (SDI)

SDI quantifies shot difficulty using a weighted combination:

$$SDI = 0.30 \times distance + 0.20 \times clock + 0.20 \times type + 0.15 \times zone + 0.15 \times angle$$

Higher SDI indicates a more difficult shot.

Player Clustering

K-Means clustering was applied to player-level features:

- Zone percentages (6 zones)
- Average shot distance
- Pull-up rate (jump shot percentage)
- Average xFG and SDI
- Usage% (minutes-weighted from NBA Stats usage table)
- Attempts per game (usage proxy / fallback)

Clusters were reduced to 2D via PCA for visualization.

Archetype names are assigned *post-hoc* using simple heuristics on cluster averages (e.g., 3PT share, SDI, and usage level). These labels describe **shot profiles** only and do not incorporate defensive role or on-ball context.

Analysis Dataset

```
shots <- arrow::read_parquet("../data/shots_with_xp_2025-26.parquet")
residuals <- read.csv("../data/player_residuals.csv")
clusters <- read.csv("../data/player_clusters.csv")
```

Results

xFG Model Performance

```
model_metrics <- data.frame(
  Metric = c("Accuracy", "AUC-ROC", "Log Loss"),
  Value = c("62.9%", "0.653", "0.645")
)
kable(model_metrics, caption = "xFG Model Performance (Logistic Regression)")
```

Table 1: xFG Model Performance (Logistic Regression)

Metric	Value
Accuracy	62.9%
AUC-ROC	0.653
Log Loss	0.645

Summary Metrics

```
summary_metrics <- shots |>
  mutate(
    shot_value = if_else(SHOT_TYPE == "3PT Field Goal", 3L, 2L),
    made = SHOT_MADE_FLAG == 1L
  ) |>
  summarise(
    attempts = n(),
    fg_pct = mean(made),
    fg2_pct = mean(made[SHOT_TYPE == "2PT Field Goal"]),
    fg3_pct = mean(made[SHOT_TYPE == "3PT Field Goal"]),
    avg_distance = mean(shot_distance_feet, na.rm = TRUE)
  )

kable(
  summary_metrics,
  digits = 3,
  caption = "League-wide shooting summary for the 2025-26 regular season."
)
```

Table 2: League-wide shooting summary for the 2025-26 regular season.

attempts	fg_pct	fg2_pct	fg3_pct	avg_distance
124518	0.469	0.547	0.359	13.832

Shot Distribution

The shot chart and distance distribution contextualize league-wide shot selection. As expected in the modern NBA, attempts are concentrated at the rim and beyond the three-point line, while mid-range volume is comparatively low.

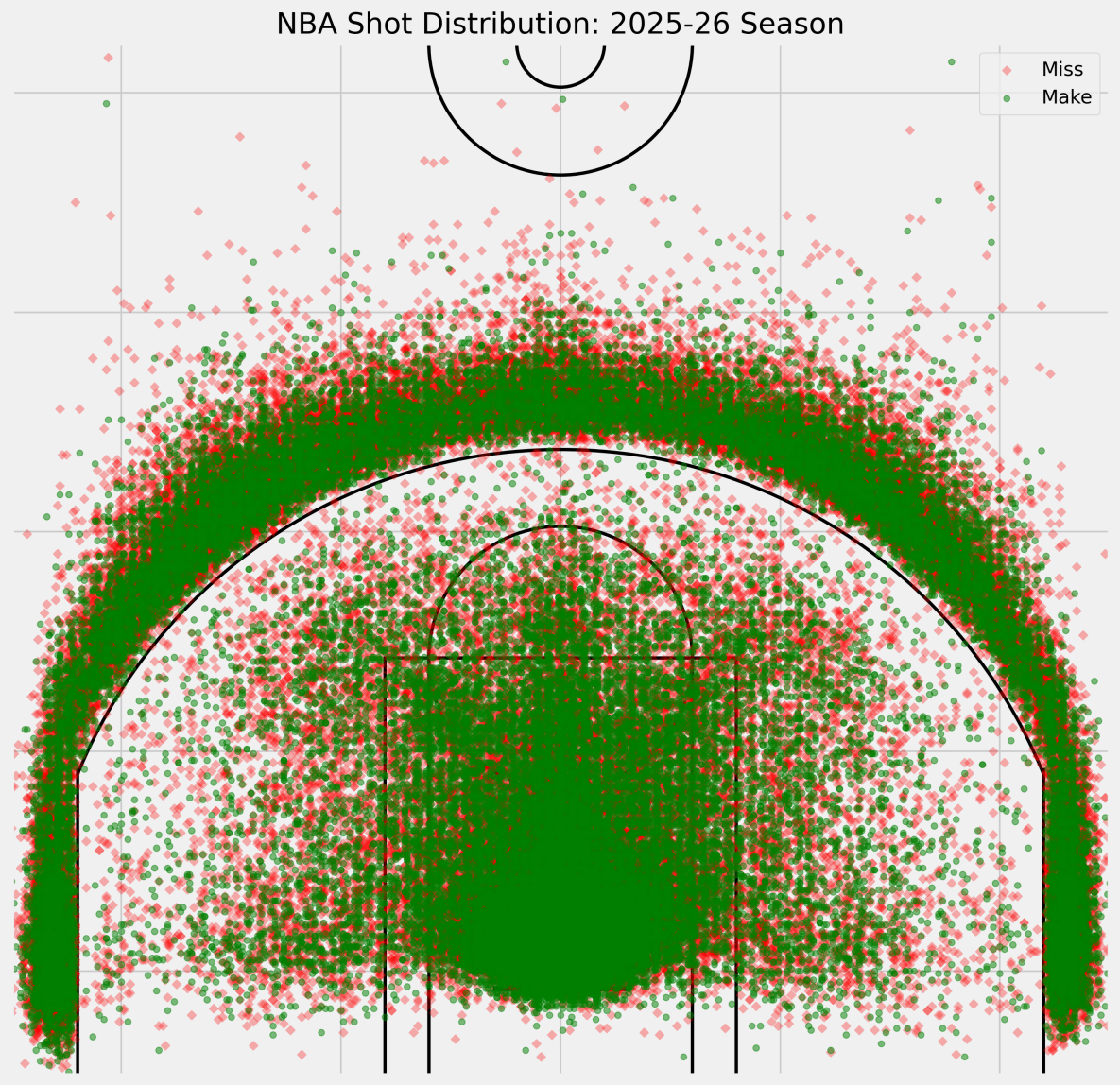


Figure 1: Shot distribution for the 2025-26 regular season.

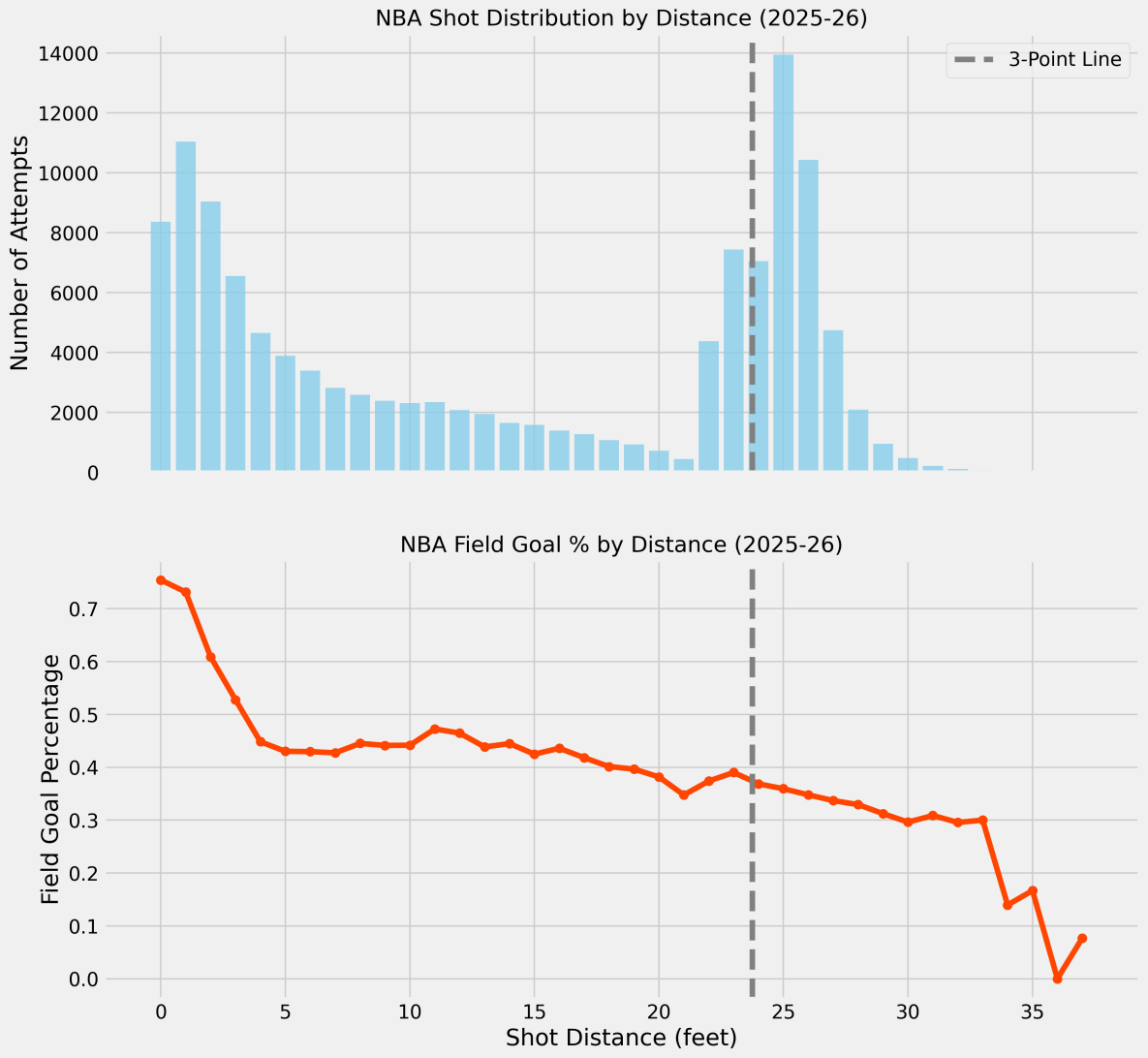


Figure 2: Shot distance distribution for the 2025-26 regular season.

Residual Analysis

```
top_over <- head(residuals, 10)
kable(
  top_over[, c("player", "avg_xFG", "actual_fg_pct", "residual_fg_pct", "attempts")],
  digits = 3,
  caption = "Top 10 Overperformers (Positive Residuals)"
)
```

Table 3: Top 10 Overperformers (Positive Residuals)

player	avg_xFG	actual_fg_pct	residual_fg_pct	attempts
Nikola Jokić	0.476	0.605	0.129	560
Luke Kennard	0.411	0.537	0.126	218
Shai Gilgeous-Alexander	0.464	0.554	0.090	908

player	avg_xFG	actual_fg_pct	residual_fg_pct	attempts
Sam Merrill	0.385	0.470	0.085	230
Cam Spencer	0.397	0.480	0.083	344
Kevin Durant	0.437	0.514	0.077	759
Deandre Ayton	0.584	0.660	0.076	373
Naji Marshall	0.470	0.545	0.075	453
AJ Green	0.373	0.446	0.073	314
T.J. McConnell	0.462	0.533	0.071	289

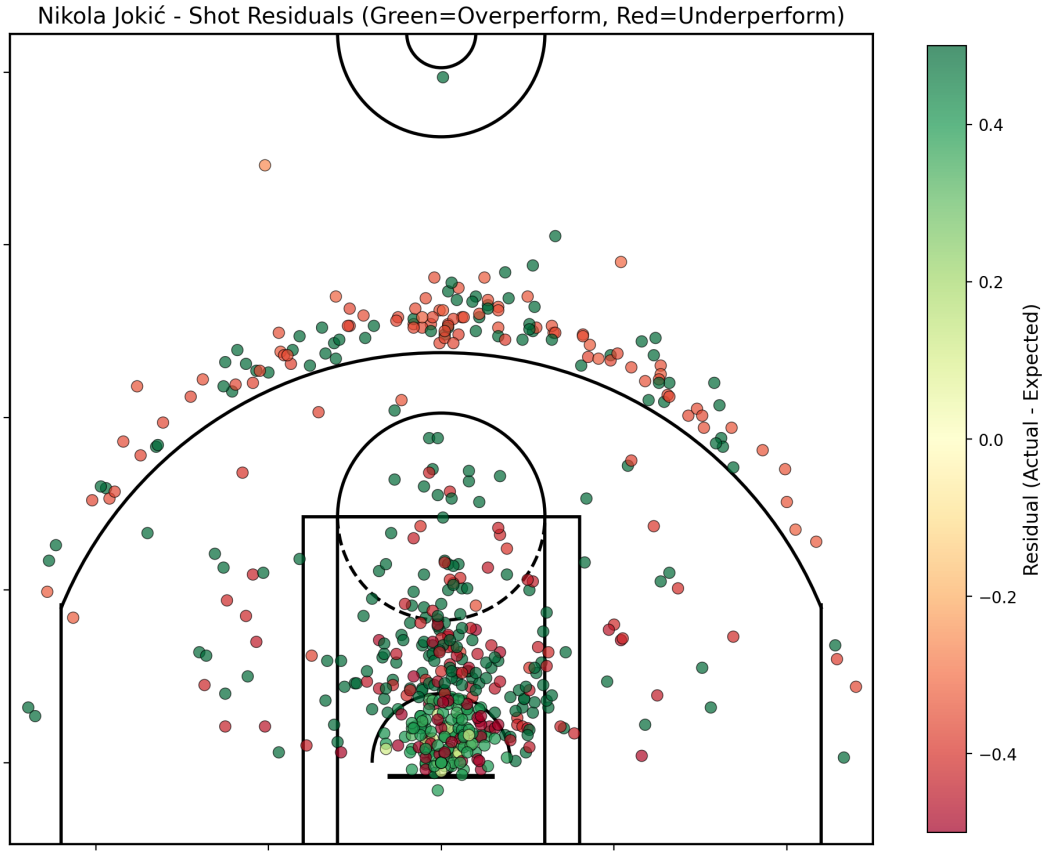


Figure 3: Residual heatmap for top overperformer (Jokić).

Shot Difficulty Index

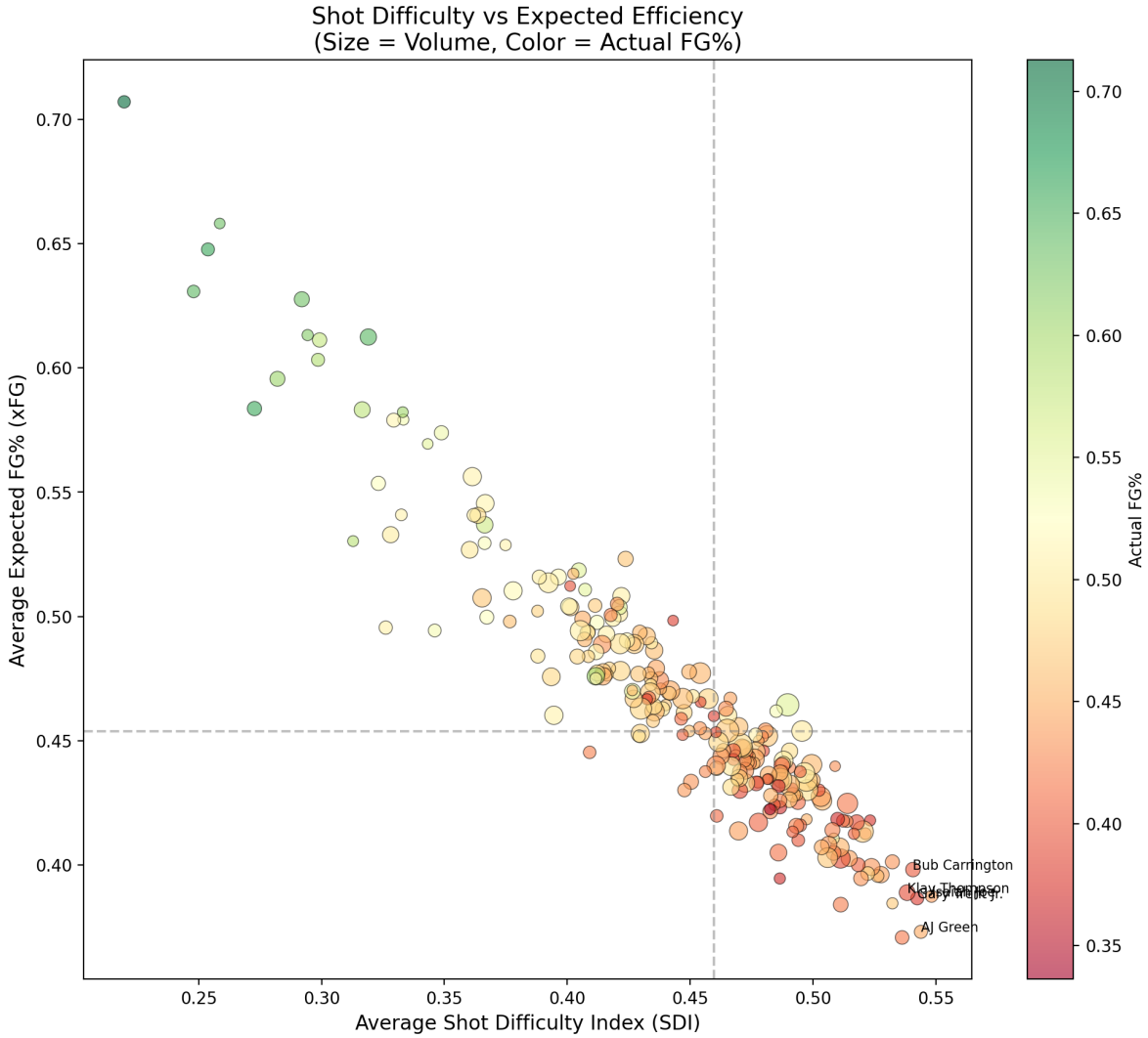


Figure 4: SDI vs xFG scatter plot. Size indicates volume, color indicates actual FG%.

Player Archetypes

```
cluster_summary <- clusters |>
  group_by(archetype) |>
  summarise(
    count = n(),
    avg_distance = mean(avg_distance),
    avg_xFG = mean(avg_xFG),
    avg_sdi = mean(avg_sdi),
    avg_usage = mean(usage_pct, na.rm = TRUE),
    avg_apg = mean(attempts_per_game, na.rm = TRUE),
    .groups = "drop"
  )
kable(cluster_summary, digits = 3, caption = "Player Archetype Summary")
```

Table 4: Player Archetype Summary

archetype	count	avg_distance	avg_xFG	avg_sdi	avg_usage	avg_apg
Balanced	75	11.091	0.497	0.403	0.209	11.526
High 3PT / High SDI	72	17.179	0.435	0.478	0.169	8.327
High SDI / Non-3	84	15.714	0.438	0.482	0.237	13.232
Rim Heavy / Low Distance	16	4.398	0.609	0.291	0.190	8.530

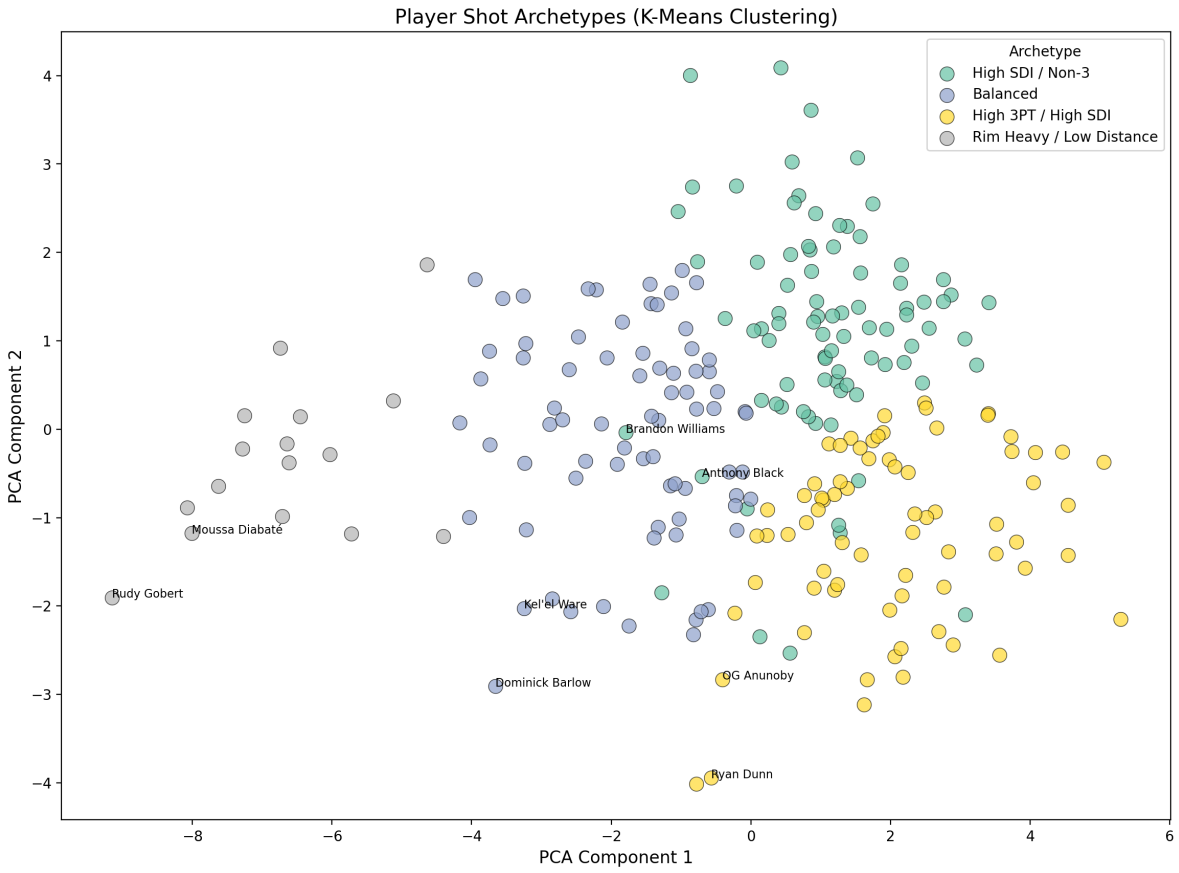


Figure 5: PCA visualization of player shot archetypes.

Shot Density

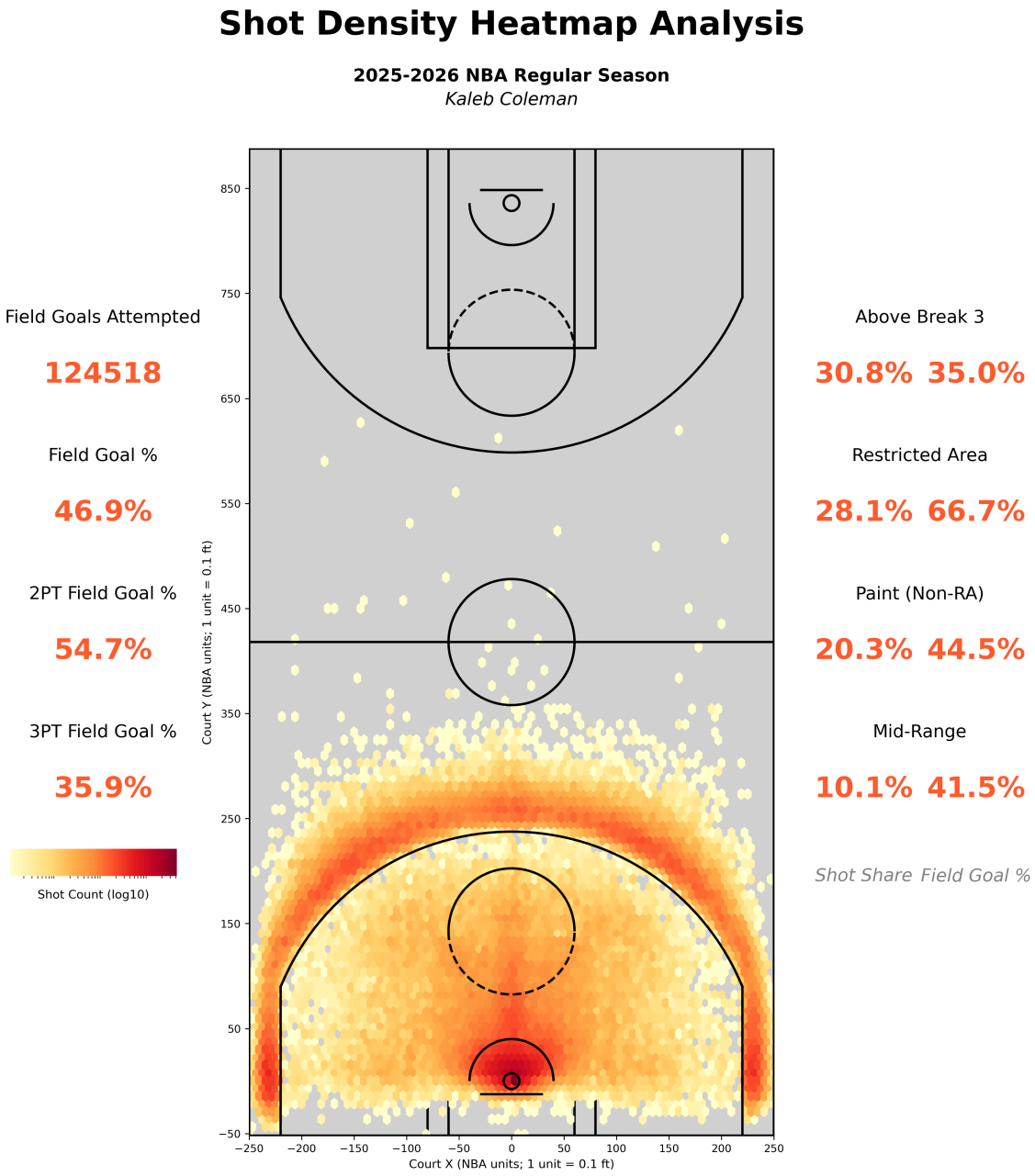


Figure 6: League-wide shot density heatmap for the 2025-26 regular season.

Player Performance (POE Shot Charts)

Player-level POE shot charts show where individual shot-making exceeds or falls short of model expectations.

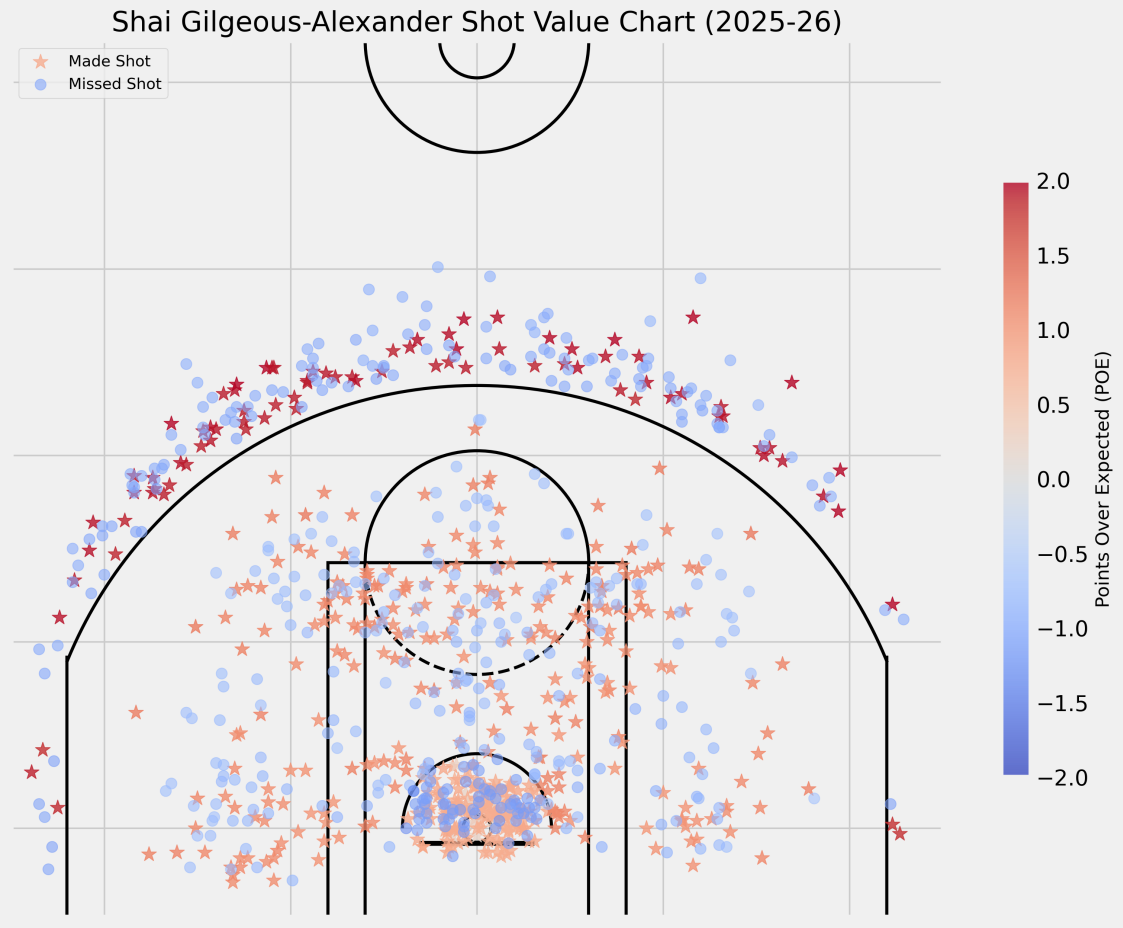
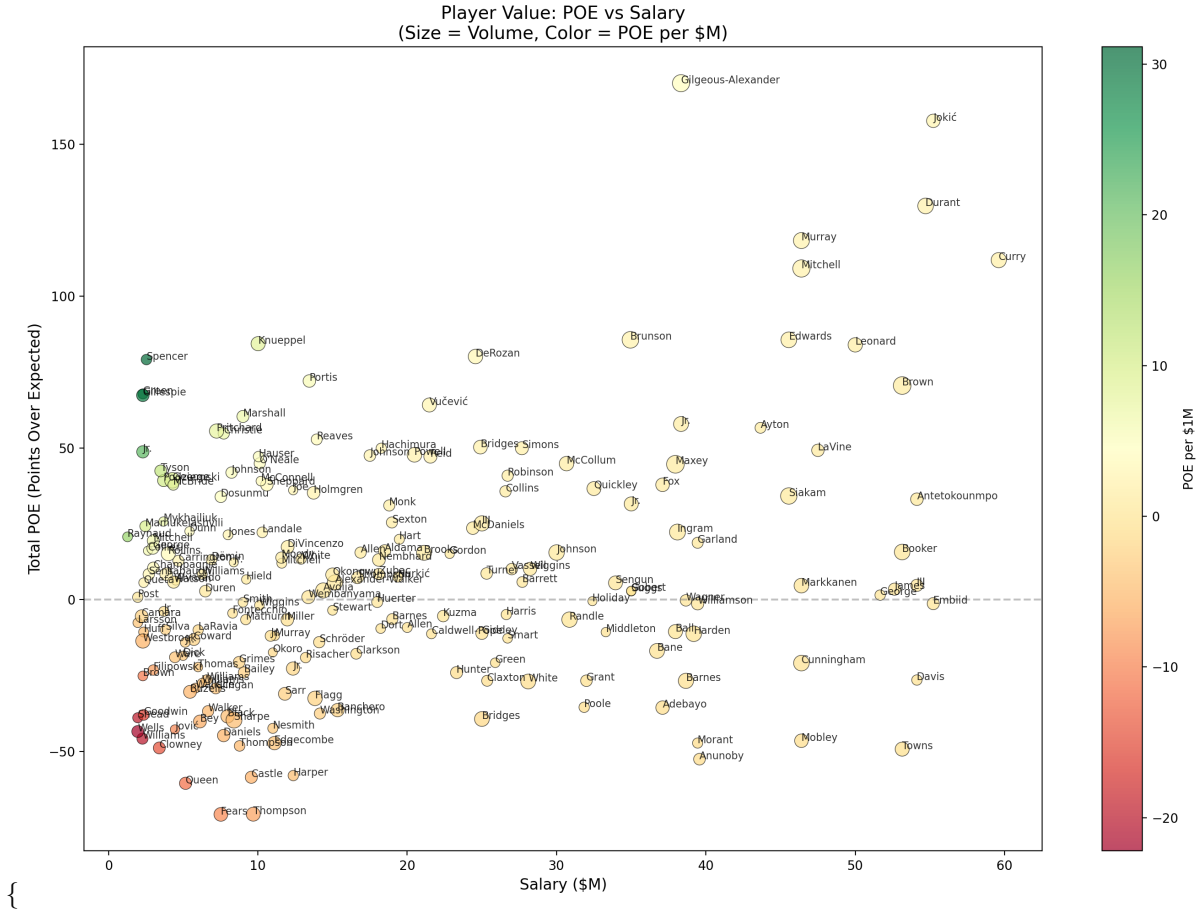


Figure 7: Shai Gilgeous-Alexander POE shot chart (2025-26).

Player Value (POE vs Salary)

We combine POE with salary estimates to contextualize performance efficiency. This is not a definitive value model (it omits defense and usage context), but it provides a first-order lens on shot-making return per $1M.

\begin{figure}[H]



{

}

\caption{Player value scatter: POE vs Salary with color encoding POE per \$1M.} \end{figure}

Discussion

Key Findings

1. **Model Performance**: The logistic regression xFG model achieves 63% accuracy, near the theoretical ceiling without defender tracking data.
2. **Overperformers**: Nikola Jokić leads with +12.9% residual, indicating elite shot-making ability beyond what location and context predict.
3. **Shot Selection**: League-wide charts confirm the rim/three prioritization with suppressed mid-range volume.
4. **Shot Difficulty**: Players like Isaiah Joe and AJ Green take the most difficult shots (high SDI), while rim attackers like Gobert have low SDI.
5. **Archetypes**: Distinct *shot-profile* types emerged based on zone mix, SDI, and usage. Labels are heuristic and role-agnostic:
   - **Rim Heavy / Low Distance**: Low distance, low SDI, high efficiency
   - **High 3PT / High SDI**: High distance, high SDI, moderate efficiency
   - **Balanced**: Moderate distance, moderate SDI
6. **Value Context**: POE per $1M adds a salary lens for identifying efficient shot makers relative to contract cost.

Implications

- **Player Evaluation**: Residual analysis identifies players who consistently beat expectations, valuable for player development and trade evaluation.
- **Shot Selection**: SDI helps distinguish players who take difficult shots by choice vs. those forced into them.
- **Team Building**: Archetype clustering can inform roster construction and lineup optimization.
- **Contract Efficiency**: Salary overlays provide a first-order signal for value-based decision making.

Limitations

This study uses league-wide averages and does not account for defensive pressure, which is the primary driver of shot difficulty. The xFG model is limited by the absence of tracking data. Salary data is scraped from a public source and may include reporting lags or contract nuances.

Conclusion

This paper presented an xFG model and three advanced analytics products for NBA shot analysis. The residual analysis, SDI, and player clustering provide actionable insights for player evaluation and team strategy. Future work should incorporate tracking data for defender distance and extend the analysis to team-level and lineup-level aggregations.

References

Basketball-Reference Player Contracts. n.d. Online. https://www.basketball-reference.com/contracts/players.html.

ESPN NBA API. n.d. Online. https://site.api.espn.com/apis/site/v2/sports/basketball/nba/scoreboard.

NBA Stats API. n.d. Online. https://stats.nba.com/stats/.