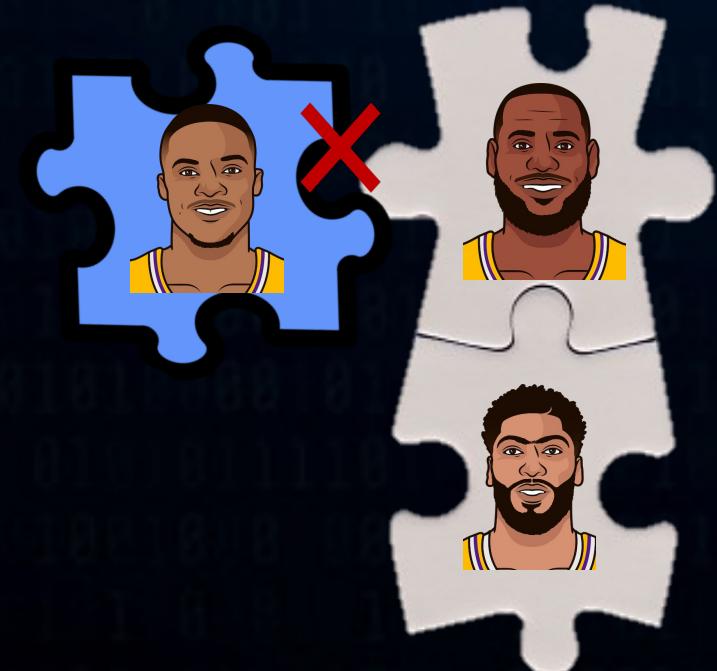


Constructing the Ideal NBA Roster



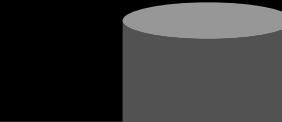
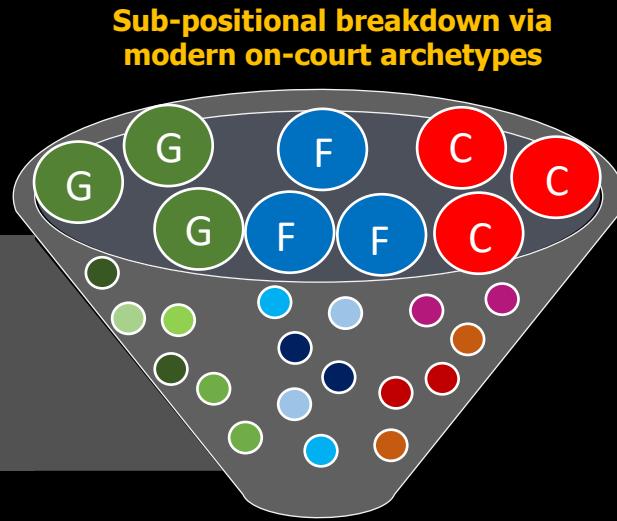
Romith Challa

Project Pipeline:



Data Preparation

Unsupervised
Clustering
Algorithms



2021 - 2022
NBA Lineup Data

Feature
Engineering

Interactive
Web-App

Regression &
Classification
Modeling

**"How will a [specific lineup]
fit and perform?"**

Use-Cases:



Data Engineering



Data Retrieval

40+ NBA Stats-API Endpoints

500+ Total Attributes



< 80 Total Features

Scoring Stats

- 2-PT vs. 3-PT vs. FT
- Catch-And-Shoot vs. Pull-Up
- RA vs. Non-RA-Paint vs. Mid-Range
- Corner3 vs. Above-the-Break3

Offensive Style

- Transition vs. PnR vs. Isolation
- Cuts vs. Drives vs. Putbacks
- Assisted vs. Unassisted Attempts
- Elbow vs. Paint touches
- Touch time, # of dribbles, dist. run

Playmaking Stats

- Primary vs. Secondary Ast
- Screen Assists
- Turnovers
- % Ast on Drives, etc.

Defensive Stats

- Deflections, Steals, Blocks
- Loose Balls, Charges, Fouls
- Opponent volume & efficiency by zones, shot type, offensive style type



Data Preparation

Goal: Balance between O vs. D and style vs. production

- Standardize player & attribute names between player stats vs. lineup data (account for trades)
- Truncate redundant features & filter low-PT players
- Transform non-attempt scoring stats into per-basis (by possession or minute); percentile-based conversions for efficiency stats
- Custom aggregations for grouped players (lineups)
- ML Pre-Requisites: Missing Data, Scaling, Encoding
- PCA for dimensionality reduction

Exploratory Analysis

Goals for EDA:

- Assess the level of separability among guards, wings, forwards, bigs
- Feature distributions & correlations to aid feature engineering / selection process
- Determine characteristics of players belonging to “Net-Positive” vs. “Net-Negative” lineups

High-Level Insights:

- High correlations among attributes --> truncate feature set
- Individual base-level stats not correlated with lineup ORTG or DRTG
- Specialized offensive play-types for position – has to be taken into account while handling missing data; reinforced need to cluster by position

Methods Implemented:

PANDAS PROFILING | AUTOVIZ | SWARM-VIOLINS

PAIR-WISE FT. INTERACTIONS | SHOT-CHARTS (O & D)

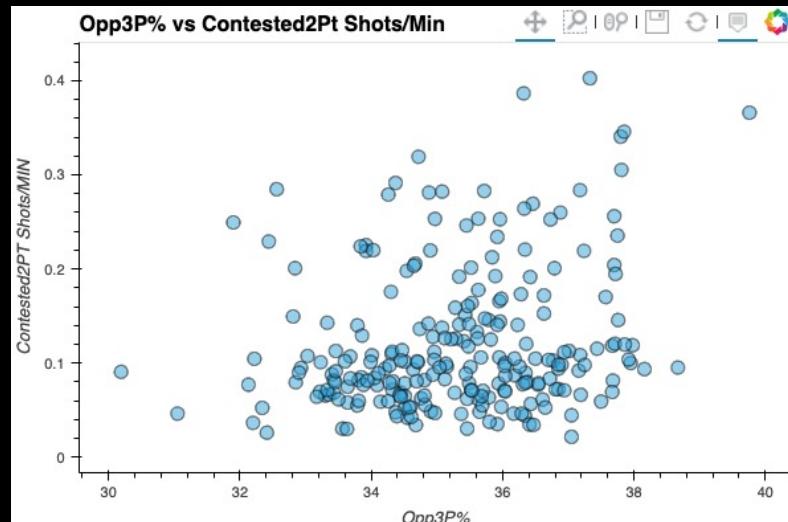
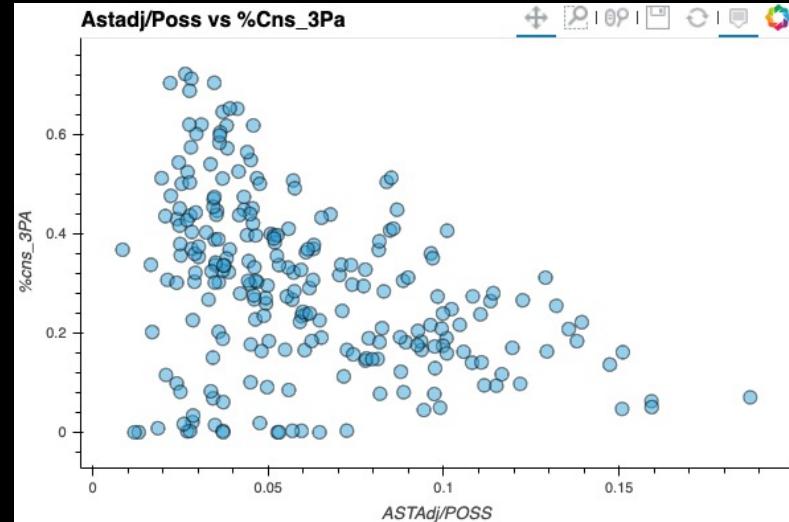
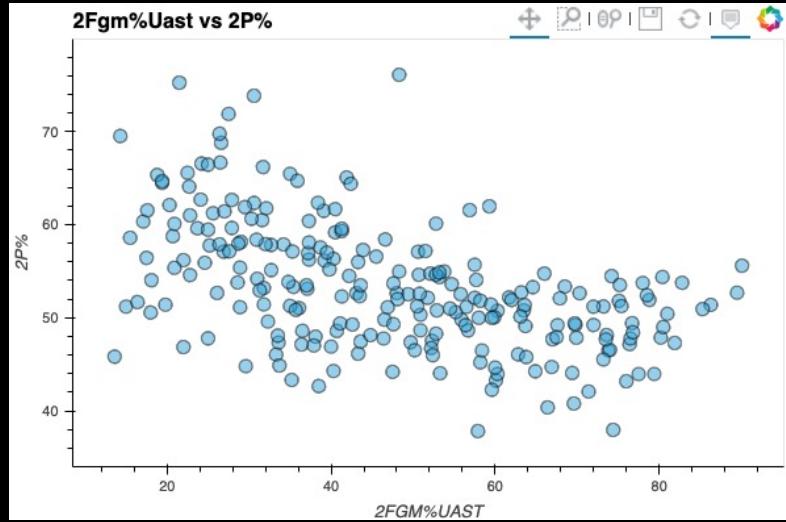


Replies to @SplashMyers and @HPbasketball

Who the f... wants to look at graphs while having a hoop convo?

11:15 AM · Oct 15, 2019 · Twitter for iPhone

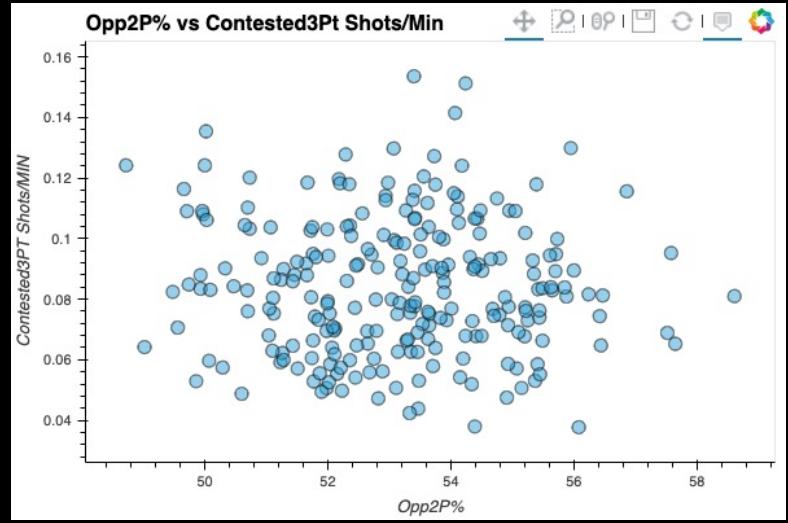
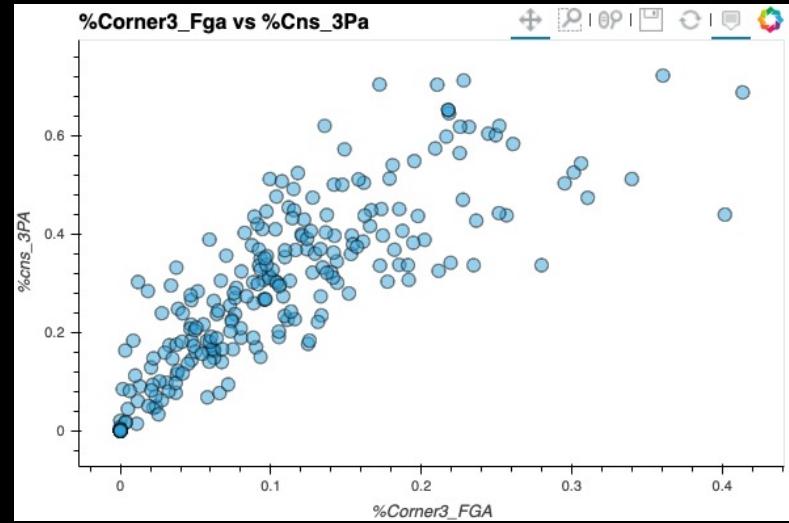
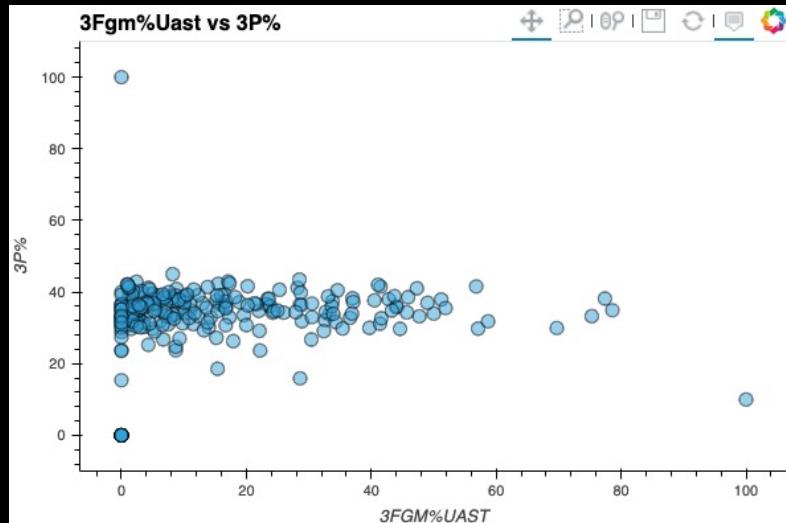
...

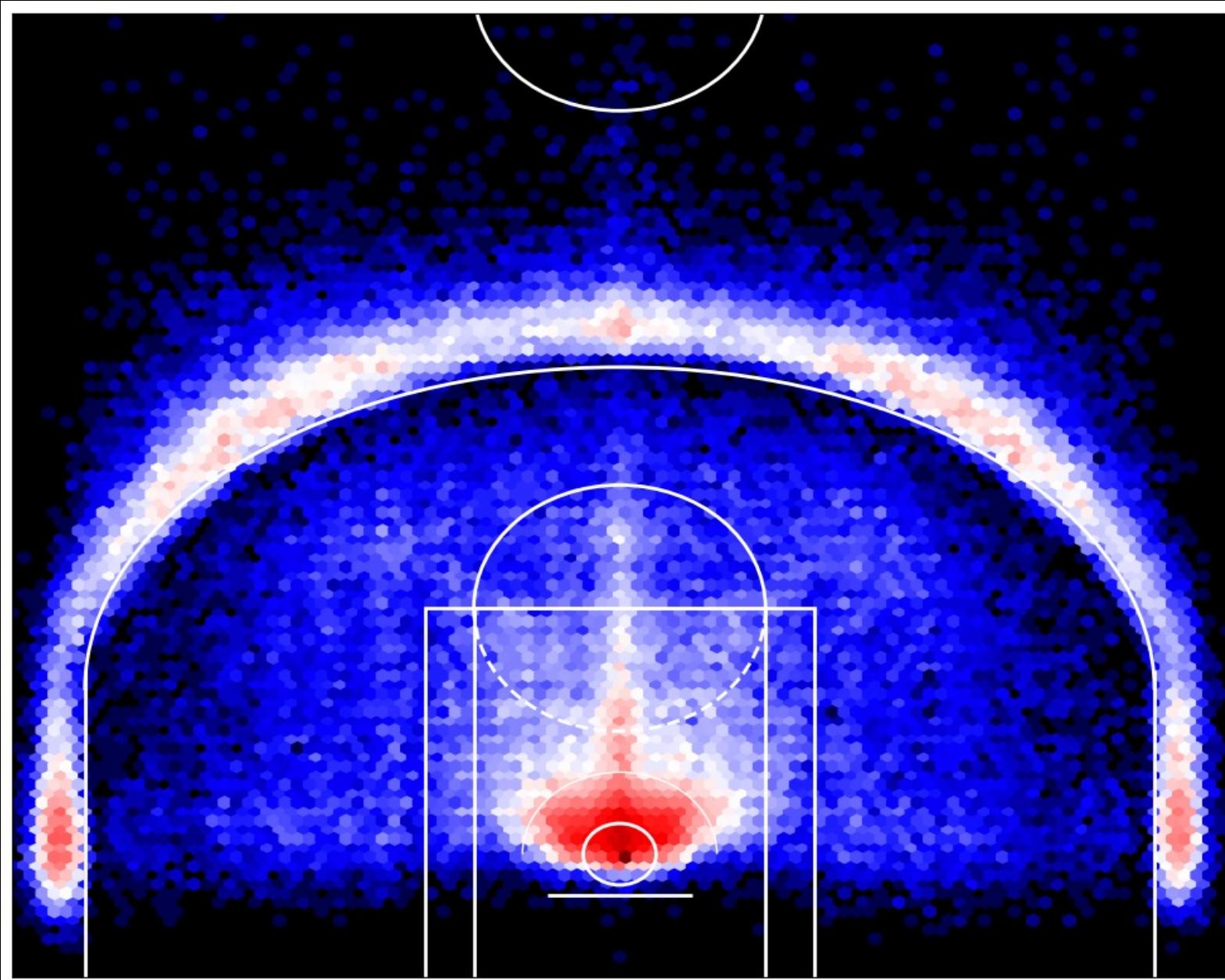


Proportion of unassisted vs. assisted FG% is impacted for 2-pointers but no such relationship exists for 3s

Players less likely to pass if their shot profile consists of mainly C&S shots. "Hidden redundant" attribute as C&S and Corner3s have strong linear relationship.

Perimeter defenders have better success against 2-pt shots compared to wing/paint defenders against 3-pt shots.





Below League Average

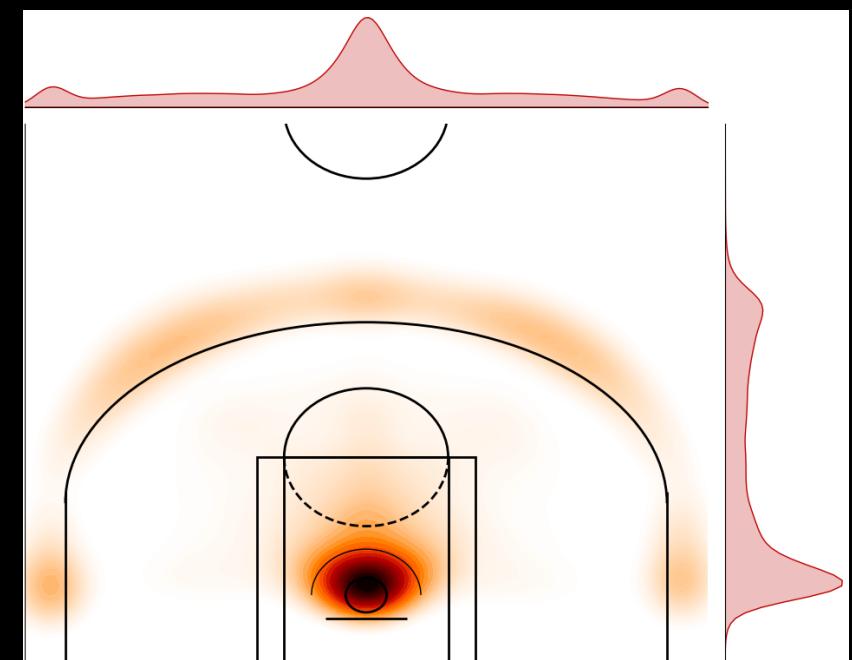
Above League Average

SHOT-CHARTS:

$\geq 5+$ NetRtg

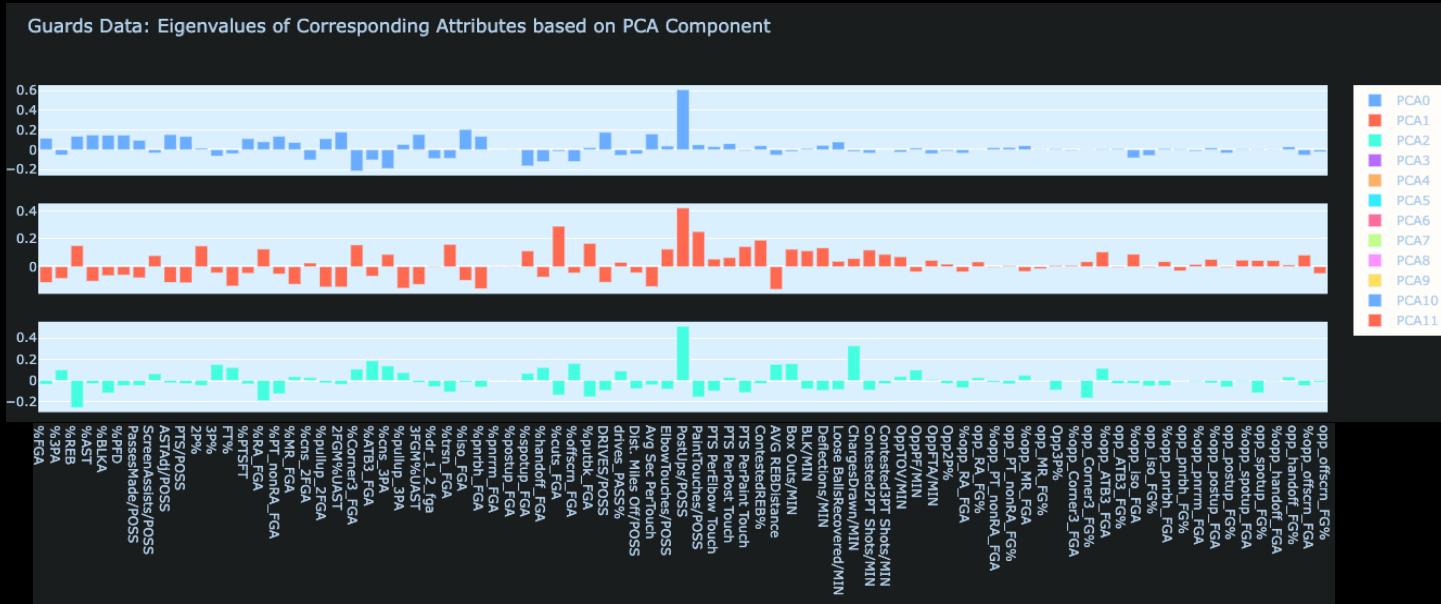
≥ 40 min.

Efficiency in RA-Paint & Corners3 are most indicative of a successful lineup.

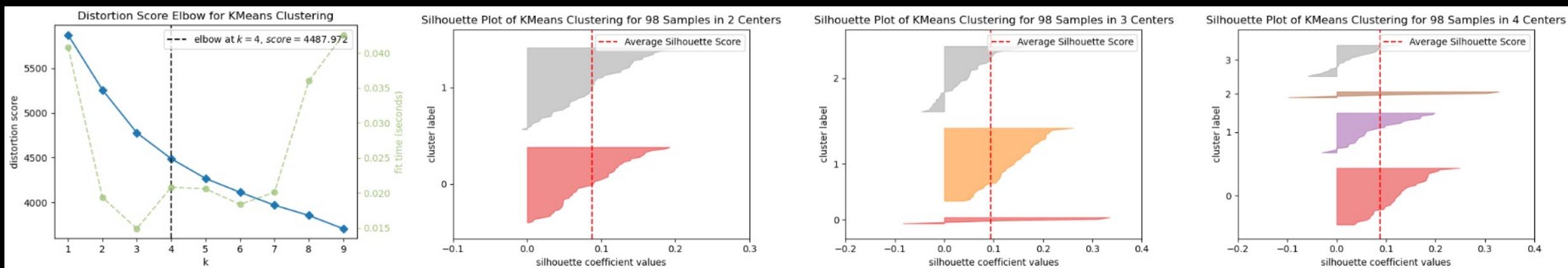


Clustering Pre-Requisites: Guards

(PROCESS REPEATED FOR WINGS, FORWARDS, BIGS)



- Plotting explained variance for PCA components
- Charting eigenvalues of each attribute based on PCA component to examine the information captured by each engineered component
- Elbow & Silhouette plots to inform hyperparameter choice
- Baseline clustering & evaluation among relevant metrics



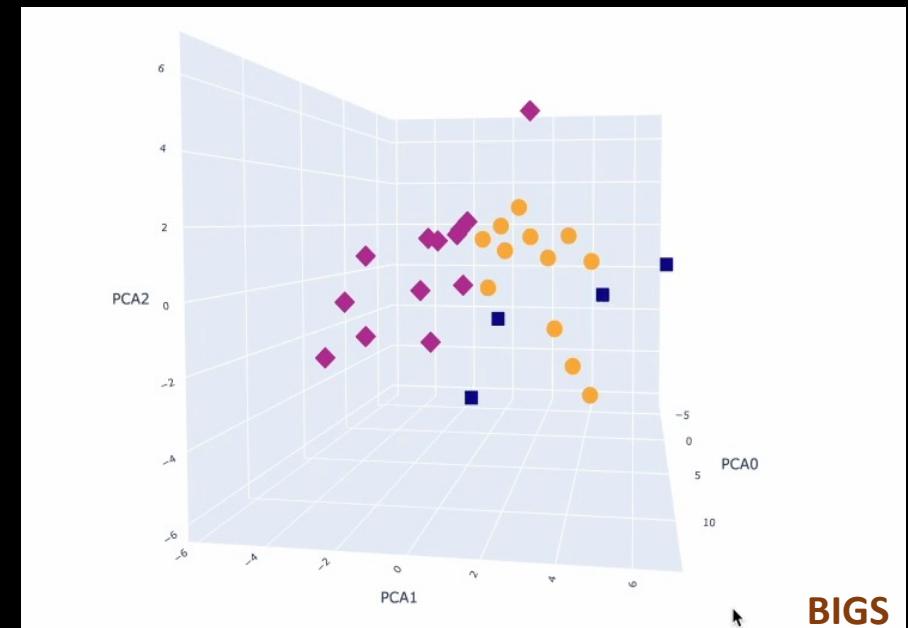
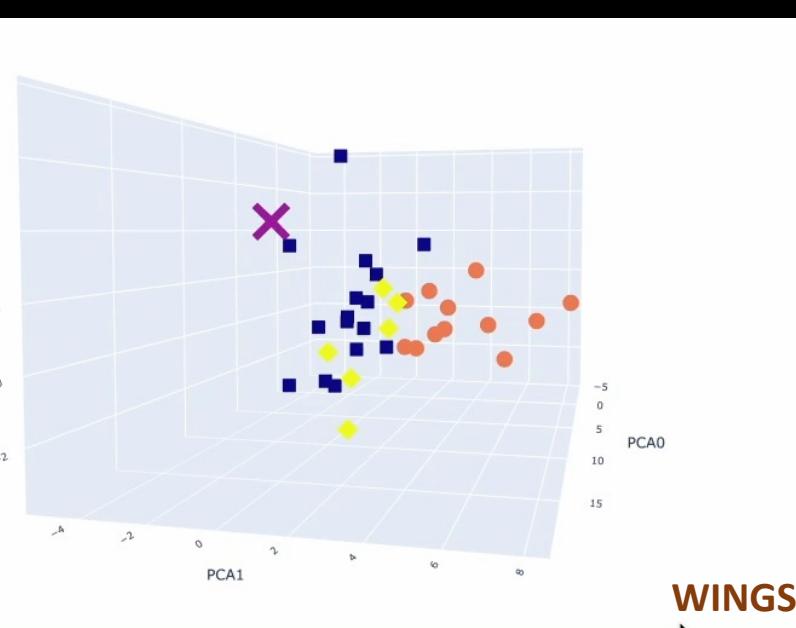
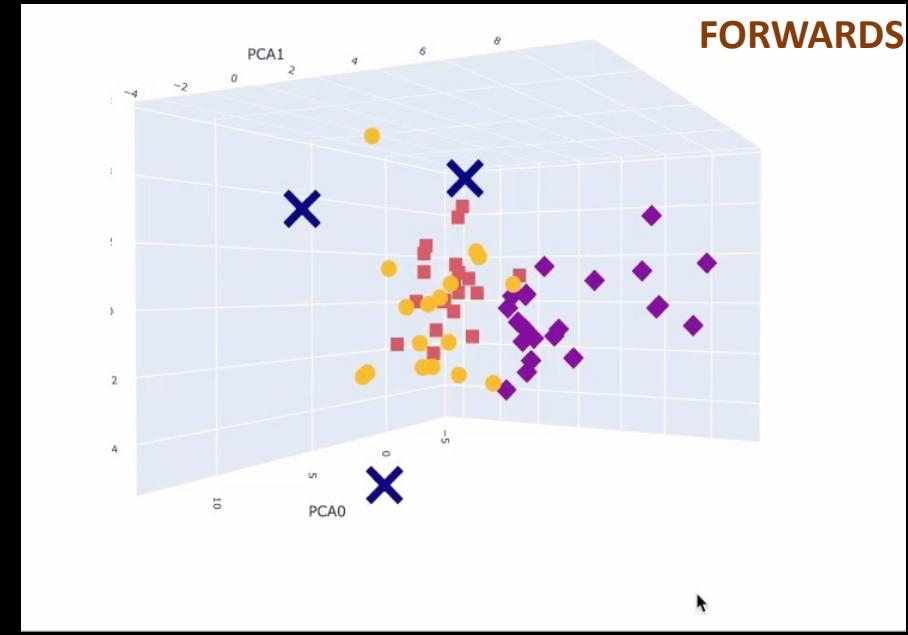
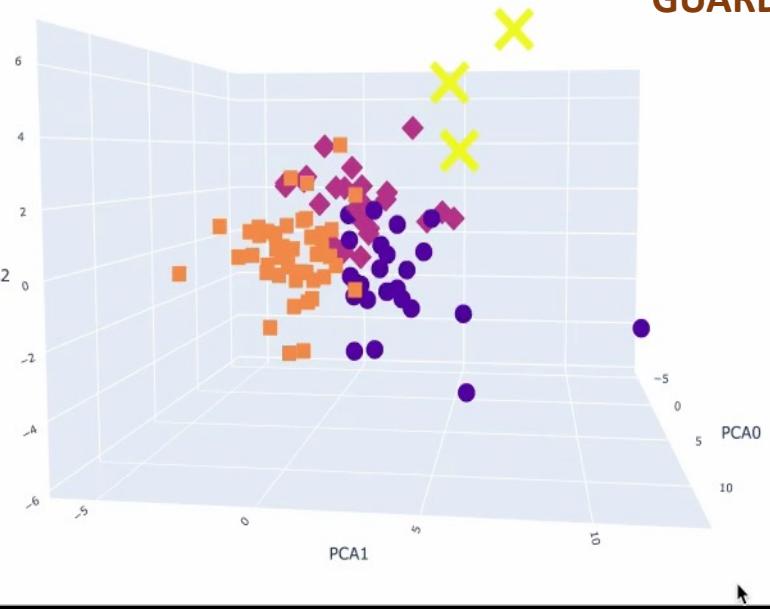
Distortion score: avg sum of squared dist. from each point to its assigned center. Silhouette Coefficient: diff. between avg intra-cluster dist. & mean nearest-cluster dist. for each sample, normalized by max value.

Cluster Analysis

Compared cluster models with appropriate metrics when there is no “ground-truth”:

- Silhouette Coefficient
- Calinski-Haribasz Score: ratio of dispersion between and within clusters, averaged out by k
- Davies-Bouldin Index: avg. similarity measure of each cluster with its most similar cluster
- K-Means, Spectral, Birch, Gaussian-Mixture

K-Means performed most consistently and had more evenly dense cluster sizes



BIGS

Regression

- BASELINE MODELS: Linear, Lasso, ElasticNet, Ridge, Decision-Tree, Random-Forest
- EVALUATION METRICS: R², MSE, RMSE, MAE, MedAE, MAPE
- HYPERPARAMETER TUNING: Random-Forest (tree depth & qt)
- TARGETS: O.RTG & D.RTG (separate models; NET RATING estimated after)

OPTIMIZED MODEL PERFORMANCE:

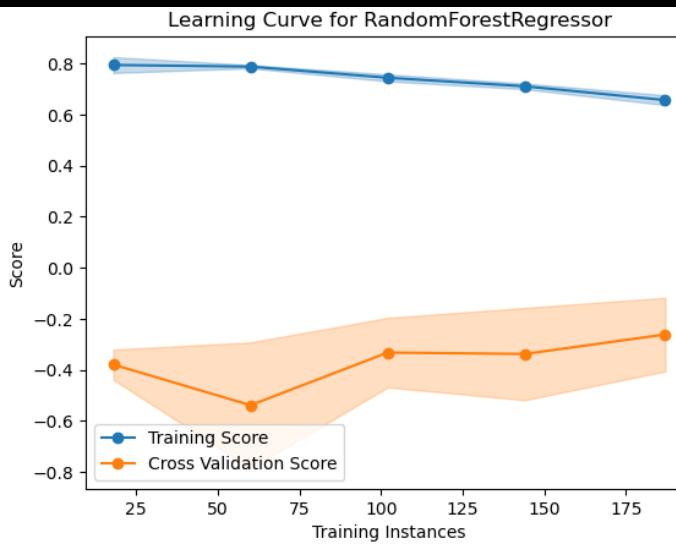
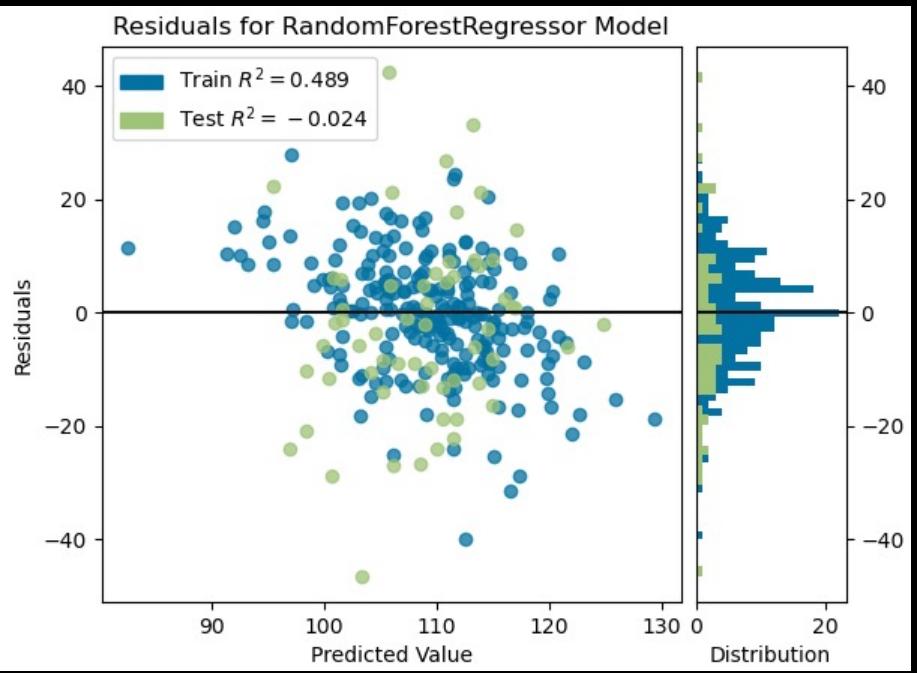
- NET RTG | TRAIN | RMSE (computed): 13.03
- NET RTG | TEST | RMSE (computed): 23.47

OPTIMIZED MODEL PERFORMANCE:

- TRAIN | ACCURACY: **89%**
- TEST | ACCURACY: **69%**
- “Good Lineup” predictions have less inaccuracies than “Bad Lineup” predictions
 - Test Recall: 77%
 - Test Precision: 77%

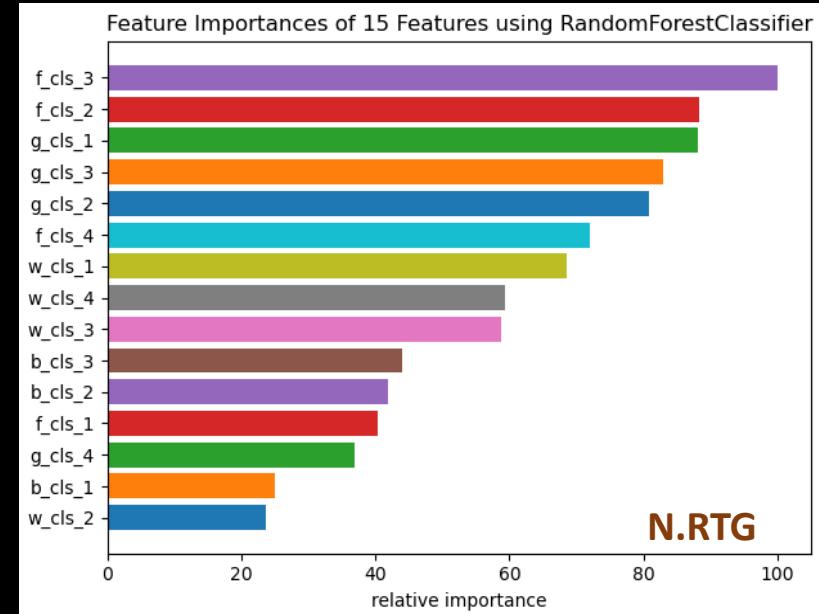
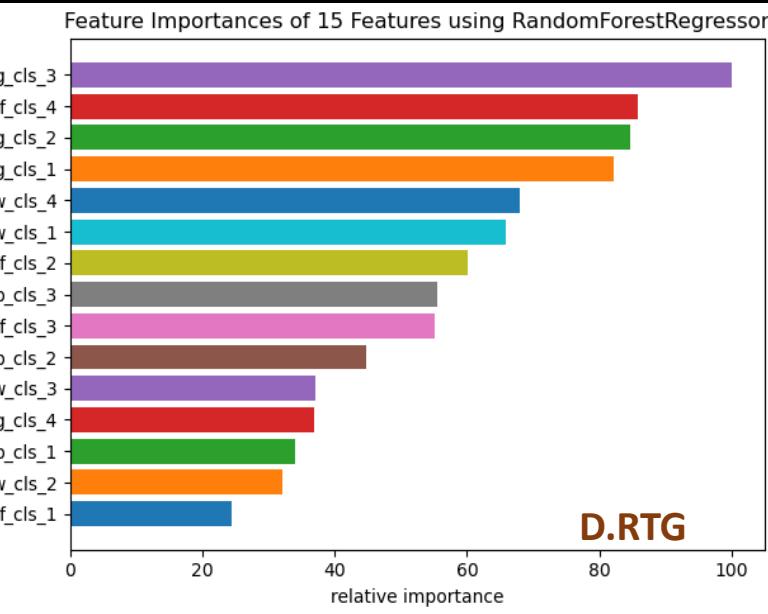
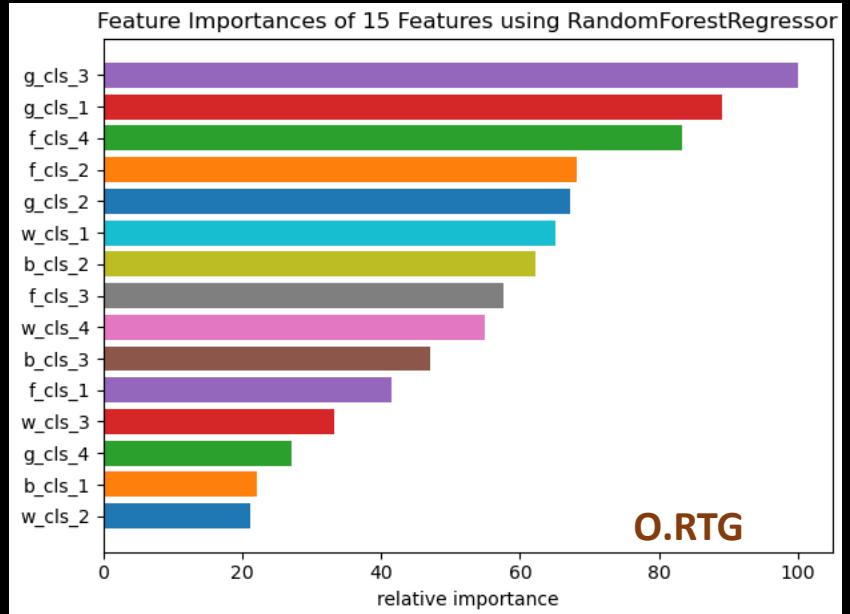
Classification

- BASELINE MODELS: Logistic, Support-Vector Machine, K-Nearest Neighbor, Random-Forest
- EVALUATION METRICS: Accuracy, Precision, Recall, F1-Score
- HYPERPARAMETER TUNING: Random-Forest (tree depth & qt)
- TARGETS: Positive or Negative Net Rating (Binary-class / Boolean)



TAKEAWAYS:

- Models valued cluster groups slightly different for offensive vs. defensive ratings
- Classification version saw completely different cluster group importances
- Overall, GUARDS & FORWARDS dictate how well a lineup performs over WINGS & BIGS
- Further tuning of models saw no improvement



Conclusions & Recommendations

ALTERNATIVE STRATEGIES:

- Using aggregated player stats in lieu of clustering-created feature set
- Re-Calibrating the balance between play-style vs. production
- Tune learners to account for lineup playing-time by using “weights” on minutes attribute

Aggregated combinations of individual player statistics are not conducive to consistently forecasting a lineup's performance.

ALTERNATIVE USE-CASE:

~~Comprehensive tool to drive lineup / roster decisions~~

Complementary tool to assess positional diversity on roster



Lineup Evaluator

This app is designed to aid NBA coaching staff in lineup selections and the front-office in trade/free-agency decisions.

Test out a lineup:

Point-Guard:

LaMelo Ball



Shooting-Guard:

Nicolas Batum



Small-Forward:

Kevin Durant



Power-Forward:

Giannis Antetokounmpo

Center:

Bam Adebayo

Projected Offensive Rating

113.2

Projected Defensive Rating

103.7

Estimated Net Rating

+9.5

