Baskets and Ballers

Using Machine Learning to Profile Pro Players

The Team

Sneha

Ryan

Aaron

Shikha

Yun

★★★ 2023 ★★★

Using Machine Learning to Profile Pro Players

API - Pandas - Machine learning - Javascript - Flask API - Web App

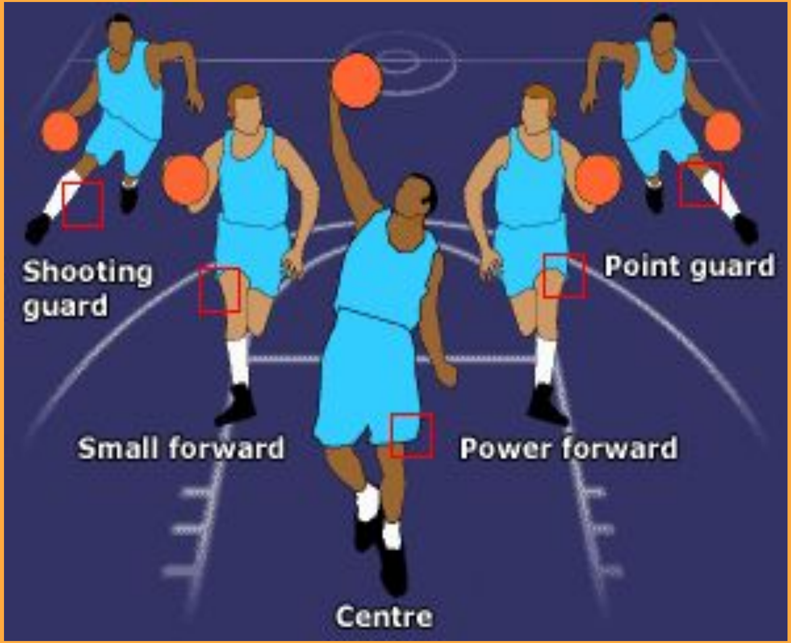
Purpose

Traditional groups for NBA players are based on their position in a game and the skills that position requires :

1. **Point guard** - Assists, Steals, Three Point Shots
2. **Shooting guard** - Points, Three Point Shots, Free Throw%
3. **Small forward** - Points, Defense Rebound, Field Goal %
4. **Power forward** - Total Rebound, Field Goal %, Blocks
5. **Center** - Total Rebounds, Blocks, Field Goal %

In reality, no player has the perfect combination of skills for a specific position. So the game play is governed by how well the skills of each player compliments the others' on the team.

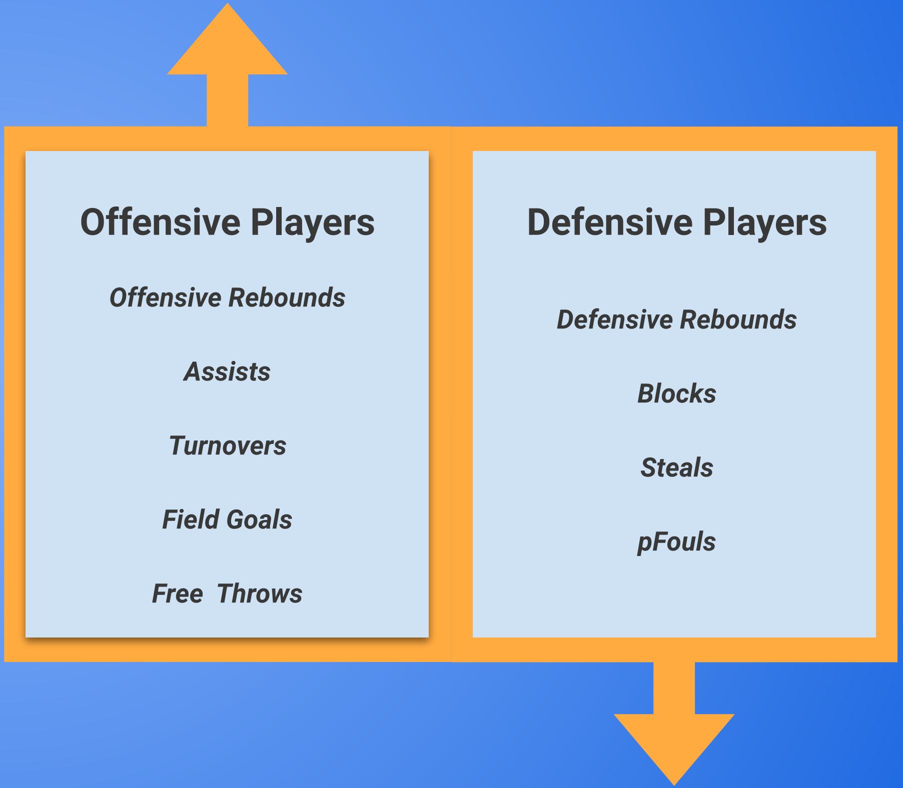
Clustering allows us to group players based on observed unique set of skills represented through their performance statistics. These groups provide insights for talent evaluation, team composition, and strategic decision-making in the context of basketball management and coaching.



Player Stats

19 metrics from **season 2021**, were used for each player who had at least **500 minutes** of playtime:

- **Minutes:** Total #minutes spent on court
- **PTS**: Total points scored through the season
- **FGM, FGA, FG%:** field goals made, attempted & %. All points made outside of free throws.
- **FTM, FTA, FTP%:** free throws made, attempted & %. Awarded after fouls on the player from the opposing team.
- TPM, TPA, TPM%: three-point field goals made, attempted & %.
- **totReb, offReb, defReb:** Possession after a missed shot by either offensive or defensive
- **Assists:** passes the ball to a teammate or defensive goaltending, that leads directly to a score
- **Steals:** Positive aggression that leads to turnover
- **Blocks:** legally deflects a field goal attempt from an offensive player to prevent a score
- **Turnovers:** ball stolen, out of bounds, pass intercepted, committing a violation/foul that leads to loss of team possession
- **pFouls:** rule breach that concerns illegal personal contact with an opponent



Offensive Players

Offensive Rebounds

Assists

Turnovers

Field Goals

Free Throws

Defensive Players

Defensive Rebounds

Blocks

Steals

pFouls

Data Sources and Cleaning

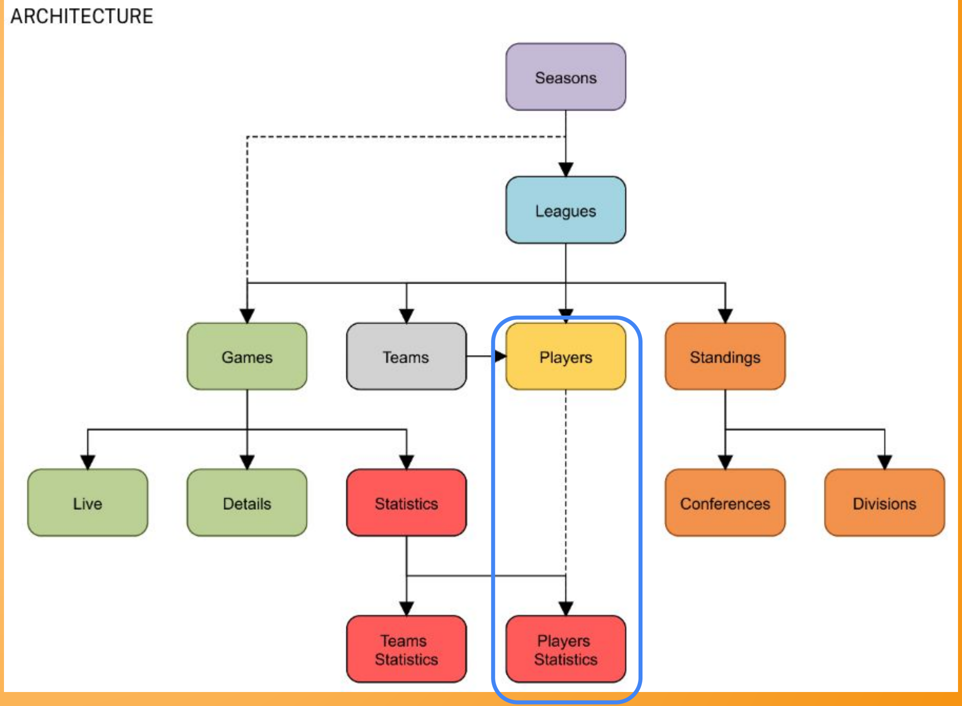
Source - https://rapidapi.com/api-sports/api/api-nba

Endpoints:

- **Players** for demographic data
- **Player Statistics** for game metrics

Steps:

1. Player demographics from Players Endpoint
2. Player stats for above Player IDs
3. Missing value treatment and format changes
4. Aggregation at player level and recalculation of derived metrics
5. Filter players with at least 500 minutes of playtime



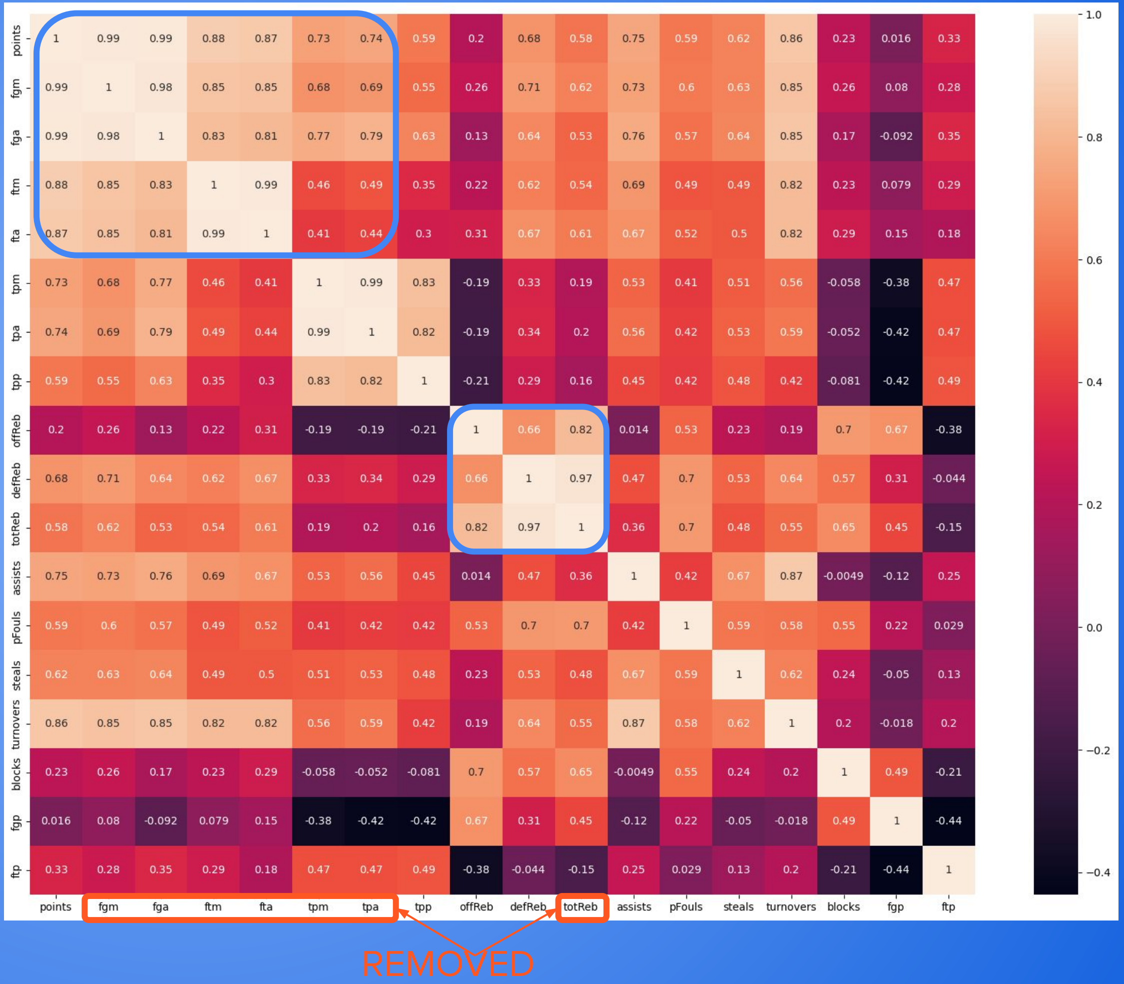
Machine Learning

Correlation Matrix

- Field Goals , 3-Point goals , Free Throws were highly correlated - we hence retained only the derived % metrics FGP, FTP, TPP & points
- Offensive and Defensive Rebounds replaced the sum ie. Total Rebounds

Data Normalization

- Features were measured in different units.
- Used ‘StandardScaler’ to prevent any single feature from dominating the clustering process.



	points	fgm	fga	ftm	fta	tpm	tpa	tpp	offReb	defReb	totReb	assists	pFouls	steals	turnovers	blocks	fgp	ftp
points	1	0.99	0.99	0.88	0.87	0.73	0.74	0.59	0.2	0.68	0.58	0.75	0.59	0.62	0.86	0.23	0.016	0.33
fgm	0.99	1	0.98	0.85	0.85	0.68	0.69	0.55	0.26	0.71	0.62	0.73	0.6	0.63	0.85	0.26	0.08	0.28
fga	0.99	0.98	1	0.83	0.81	0.77	0.79	0.63	0.13	0.64	0.53	0.76	0.57	0.64	0.85	0.17	-0.092	0.35
ftm	0.88	0.85	0.83	1	0.99	0.46	0.49	0.35	0.22	0.62	0.54	0.69	0.49	0.49	0.82	0.23	0.079	0.29
fta	0.87	0.85	0.81	0.99	1	0.41	0.44	0.3	0.31	0.67	0.61	0.67	0.52	0.5	0.82	0.29	0.15	0.18
tpm	0.73	0.68	0.77	0.46	0.41	1	0.99	0.83	-0.19	0.33	0.19	0.53	0.41	0.51	0.56	-0.058	-0.38	0.47
tpa	0.74	0.69	0.79	0.49	0.44	0.99	1	0.82	-0.19	0.34	0.2	0.56	0.42	0.53	0.59	-0.052	-0.42	0.47
tpp	0.59	0.55	0.63	0.35	0.3	0.83	0.82	1	-0.21	0.29	0.16	0.45	0.42	0.48	0.42	-0.081	-0.42	0.49
offReb	0.2	0.26	0.13	0.22	0.31	-0.19	-0.19	-0.21	1	0.66	0.82	0.014	0.53	0.23	0.19	0.7	0.67	-0.38
defReb	0.68	0.71	0.64	0.62	0.67	0.33	0.34	0.29	0.66	1	0.97	0.47	0.7	0.53	0.64	0.57	0.31	-0.044
totReb	0.58	0.62	0.53	0.54	0.61	0.19	0.2	0.16	0.82	0.97	1	0.36	0.7	0.48	0.55	0.65	0.45	-0.15
assists	0.75	0.73	0.76	0.69	0.67	0.53	0.56	0.45	0.014	0.47	0.36	1	0.42	0.67	0.87	-0.0049	-0.12	0.25
pFouls	0.59	0.6	0.57	0.49	0.52	0.41	0.42	0.42	0.53	0.7	0.7	0.42	1	0.59	0.58	0.55	0.22	0.029
steals	0.62	0.63	0.64	0.49	0.5	0.51	0.53	0.48	0.23	0.53	0.48	0.67	0.59	1	0.62	0.24	-0.05	0.13
turnovers	0.86	0.85	0.85	0.82	0.82	0.56	0.59	0.42	0.19	0.64	0.55	0.87	0.58	0.62	1	0.2	-0.018	0.2
blocks	0.23	0.26	0.17	0.23	0.29	-0.058	-0.052	-0.081	0.7	0.57	0.65	-0.0049	0.55	0.24	0.2	1	0.49	-0.21
fgp	0.016	0.08	-0.092	0.079	0.15	-0.38	-0.42	-0.42	0.67	0.31	0.45	-0.12	0.22	-0.05	-0.018	0.49	1	-0.44
ftp	0.33	0.28	0.35	0.29	0.18	0.47	0.47	0.49	-0.38	-0.044	-0.15	0.25	0.029	0.13	0.2	-0.21	-0.44	1

REMOVED

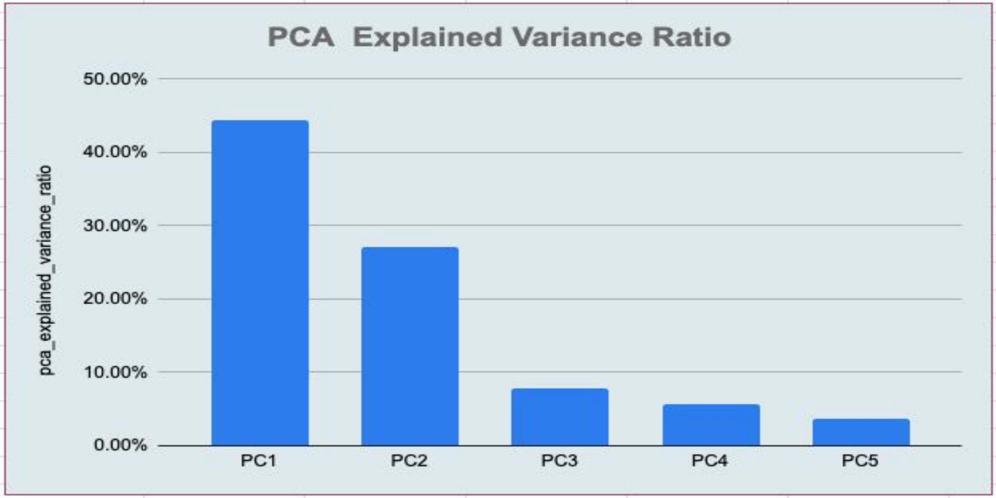
Machine Learning

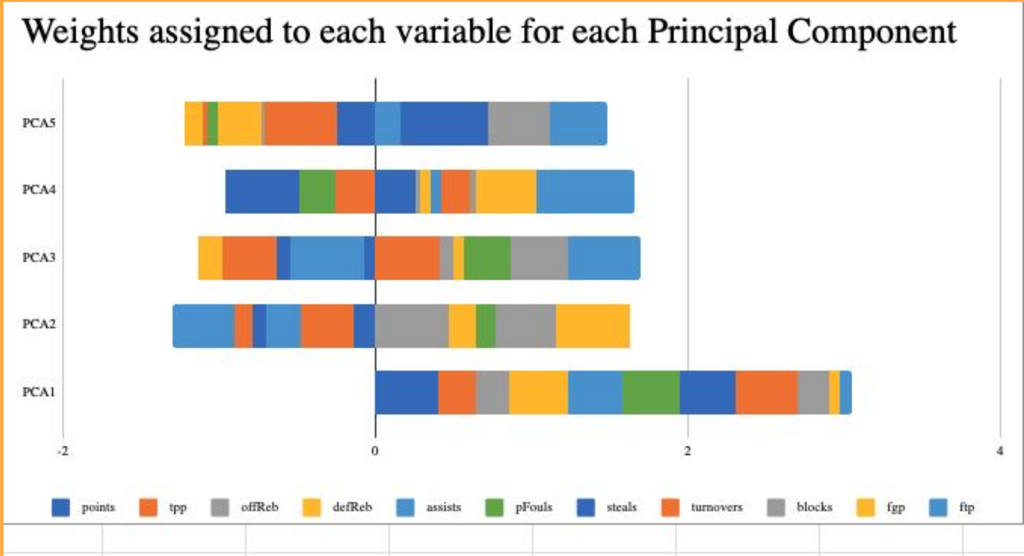
Principal Component Analysis

- 11 metrics was the 'curse of dimensionality' in this dataset
- PCA was performed to reduce the dimensionality to 5 features hence improve the model speed and efficiency

Principal Components Results

- 5 principal components explained 88.6% of the variance
- PC1 and PC2 contributed to 71% of this variance
- PC1 and PC2 have very different distribution of weights across the 11 features as shown here





Machine Learning

K-Means Elbow Curve

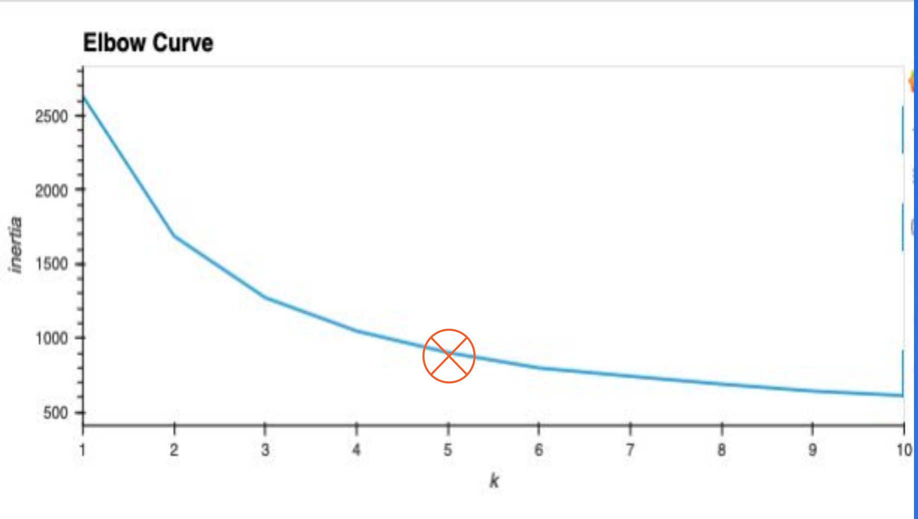
- Using 5 PC features - the k means elbow curve determined 5 optimal clusters before drop in inertia was negligible

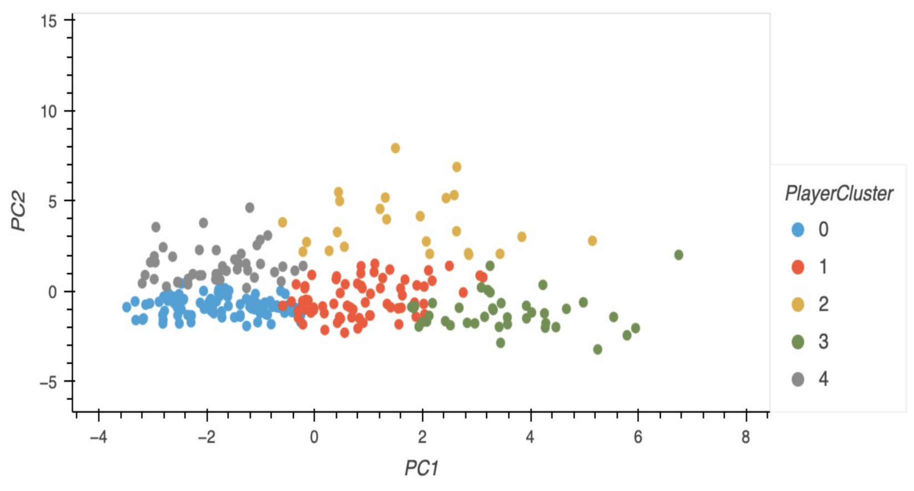
K Means Clustering

- Using PC1 and PC2 that explained 71% of the variance - the distinction in clusters was clear from the scatter plot

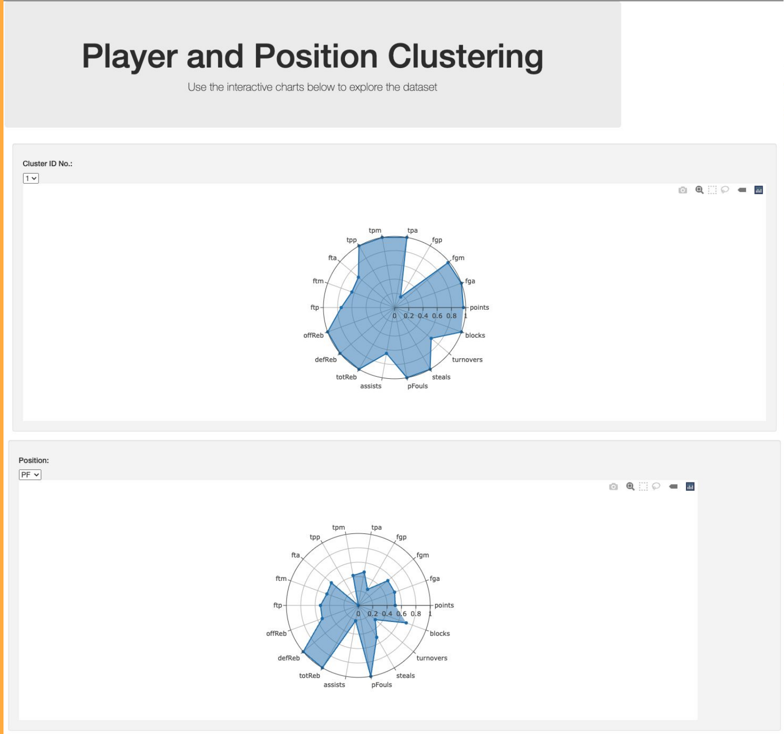
Data for App Visualizations:

- Data was aggregated at the Cluster, Position and Player level
- MinMaxScaler was used to ensure they could be rendered on a radar chart that requires only positive normalized values





Visualizations - Demo

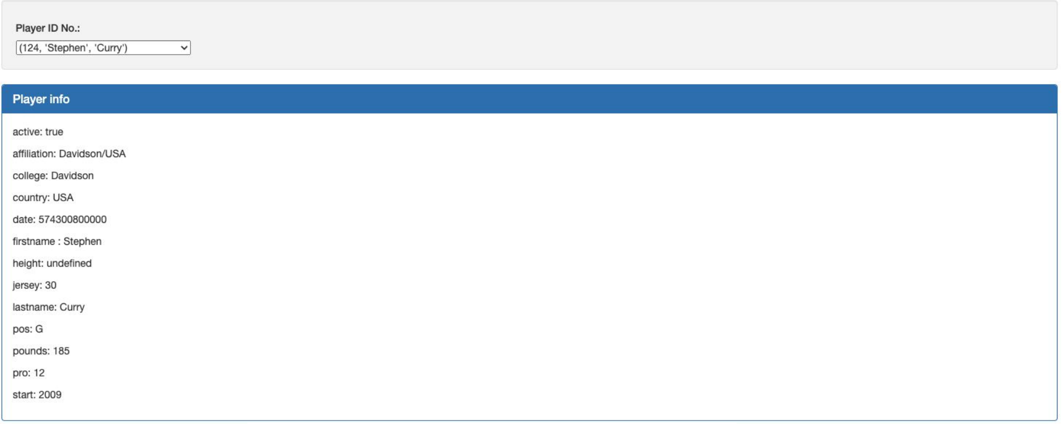


Player and Position Clustering

Use the interactive charts below to explore the dataset

Cluster ID No.: 1

Position: PF



Player ID No.: (124, 'Stephen', 'Curry')

Player info

active: true

affiliation: Davidson/USA

college: Davidson

country: USA

date: 574300800000

firstname : Stephen

height: undefined

jersey: 30

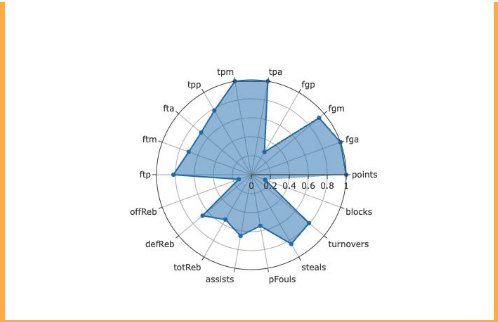
lastname: Curry

pos: G

pounds: 185

pro: 12

start: 2009



Insights

- Skills that players presented did not necessarily match what is expected from their given position (e.g. Point Guard: there were higher free throw values than three point shots made)
- Clusters showed players had opposing metrics (e.g. Cluster 2 high offReb and blocks)
- PG, Cluster 1, & Cluster 3 groups with the highest assists also have the highest turnover rates

SASYR Team Neural Network Results:

✓ Open to ideas and discussion early in the process

✓ Success is not the only path to learning - as seen from the effort we put into creating spider charts in Tableau

✓ Positive attitude towards roadblocks

✓ Moral support is valuable support

✓ Recognize strengths and weaknesses are like sticks that band together to form a log

★ If it is convoluted- there’s a better tool out there, use it!

★ Real world problem solving - require many iterations based on data understanding!

Thank You