Part A - Tableau Visualizations

1. Data and Methods

For Part A, I used the “NBA, ABA and BAA Stats” Kaggle dataset. I focused on the regular-season player-level statistics by season. I imported the CSV files into Tableau and created an extract to aggregate statistics at both the career level and the season level.

Before building the visuals, I cleaned the dataset by:

- Filtering for valid league (NBA/ABA/BAA) and team values
- Excluding rows with missing games or missing points
- Converting Season into a numeric variable and Age into a continuous dimension

2. Data Dictionary (Key Variables Used)

Field	Meaning
Player	Player full name
Season	Season year (e.g., 2020 = 2019–20 season)
Age	Player age during that season
G	Games played
MP	Minutes played
FGM	Field goals made (2PT or 3PT combined)
3PM	Three-pointers made
FTM	Free throws made
PTS	Total points scored that season
FG%	Field goal percentage
3P%	Three-point percentage
FT%	Free throw percentage

3. Derived Measures Created in Tableau

- **Career Points** = SUM(PTS) per player
- **Cumulative Career Points by Age** = RUNNING_SUM(SUM(PTS)) partitioned by Player
- **Shot Type Shares per Season**
  - FT Share = FT Points / Total Points
  - 2PT Share = Two-point Points / Total Points
  - 3PT Share = Three-point Points / Total Points

After preparing the data, I built three dashboards representing three NYT-inspired visuals.

Part A-1. Top 250 NBA Scorers (Career Total Points)

Caption (for Tableau dashboard)

This bar chart shows the top 250 all-time regular-season scorers. LeBron James and Kareem Abdul-Jabbar are highlighted to emphasize their position relative to all other scorers.

Explanation

This visualization recreates the NYT “Top 250 Scorers” chart. Using the per-season dataset, I aggregated total points for each player to compute career totals and filtered for the top 250 players. Players are sorted in descending order, with LeBron and Kareem highlighted in distinct colors.

The visual shows that LeBron currently holds a narrow lead over Kareem in total career points, while both are significantly ahead of the rest of the top scorers. The steep drop-off after the top few players shows how rare it is to surpass 30,000–35,000 career points.

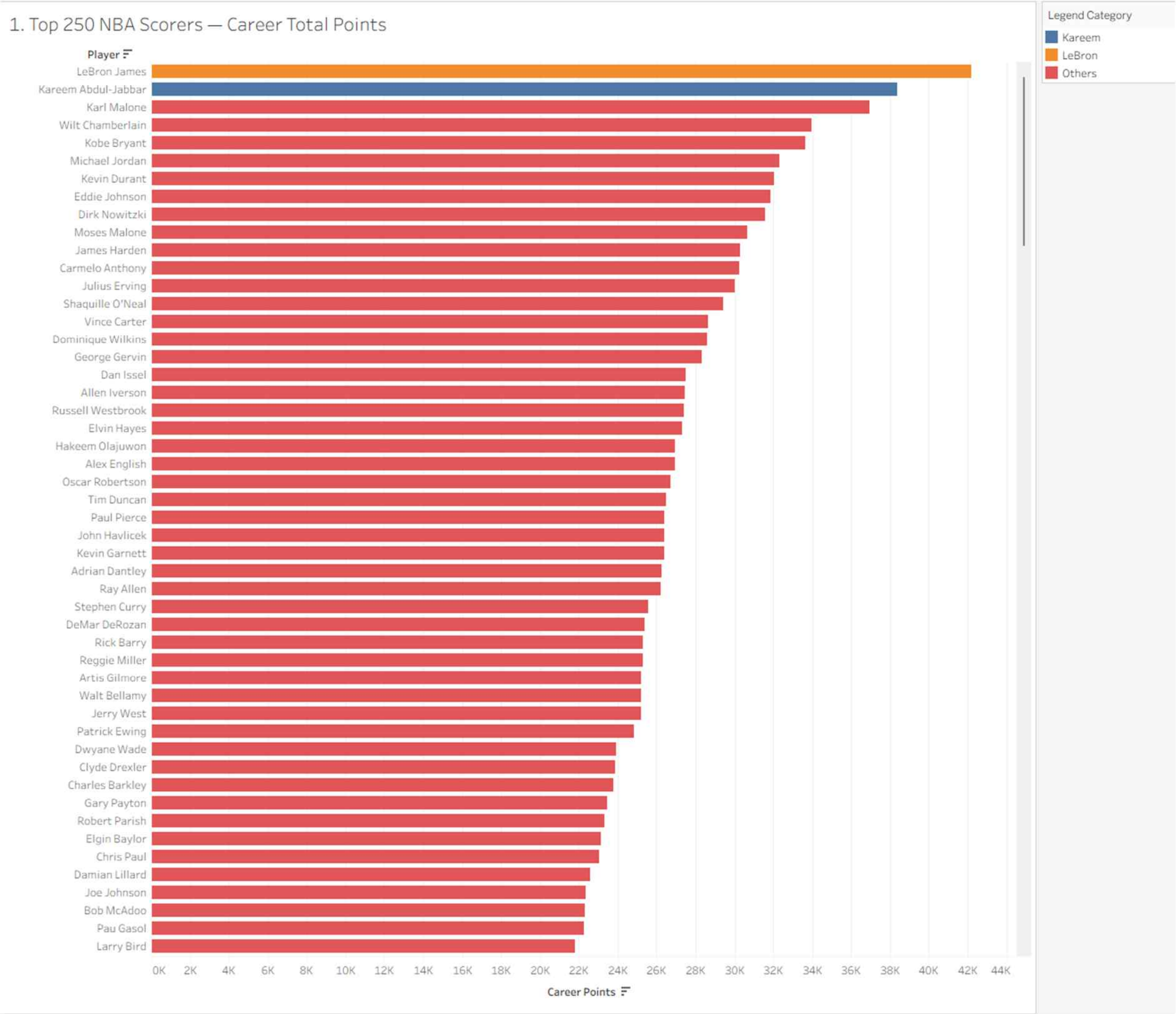


Figure A-1. Top 250 NBA Scorers — Career Total Points (Tableau)

Part A-2. History of NBA Shot Selection by Year (Field Goal Type Share)

Caption (for Tableau dashboard)

This line chart shows how the league’s scoring distribution shifted among free throws, two-pointers, and three-pointers across NBA history.

Explanation

This visualization shows the evolution of shot selection in the NBA. For each season, I aggregated league-wide made free throws, two-point shots, and three-point shots, and calculated their shares of total scoring.

- In the early decades, nearly all scoring came from two-pointers and free throws.
- After the introduction of the three-point line in 1979, 3PT share slowly increased.
- Beginning around 2010, three-point usage rose sharply, while two-point share declined.
- Free throw share has gradually decreased over time.

This matches the modern trend toward perimeter-oriented offenses and confirms the league-wide shift toward 3-point heavy play.

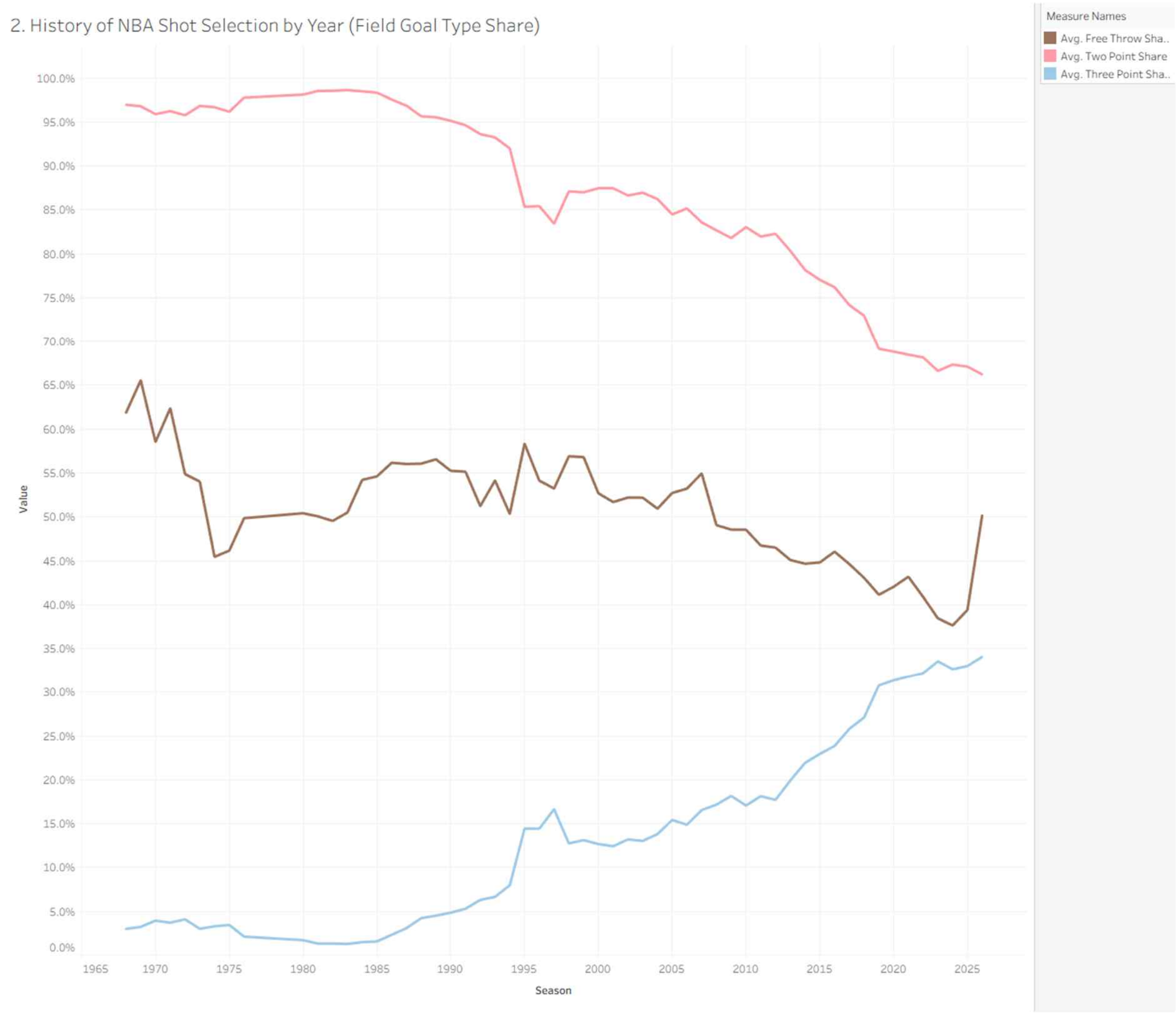


Figure A-2. History of NBA Shot Selection by Year (Tableau)

Part A-3. LeBron vs Kareem: Career Points by Age

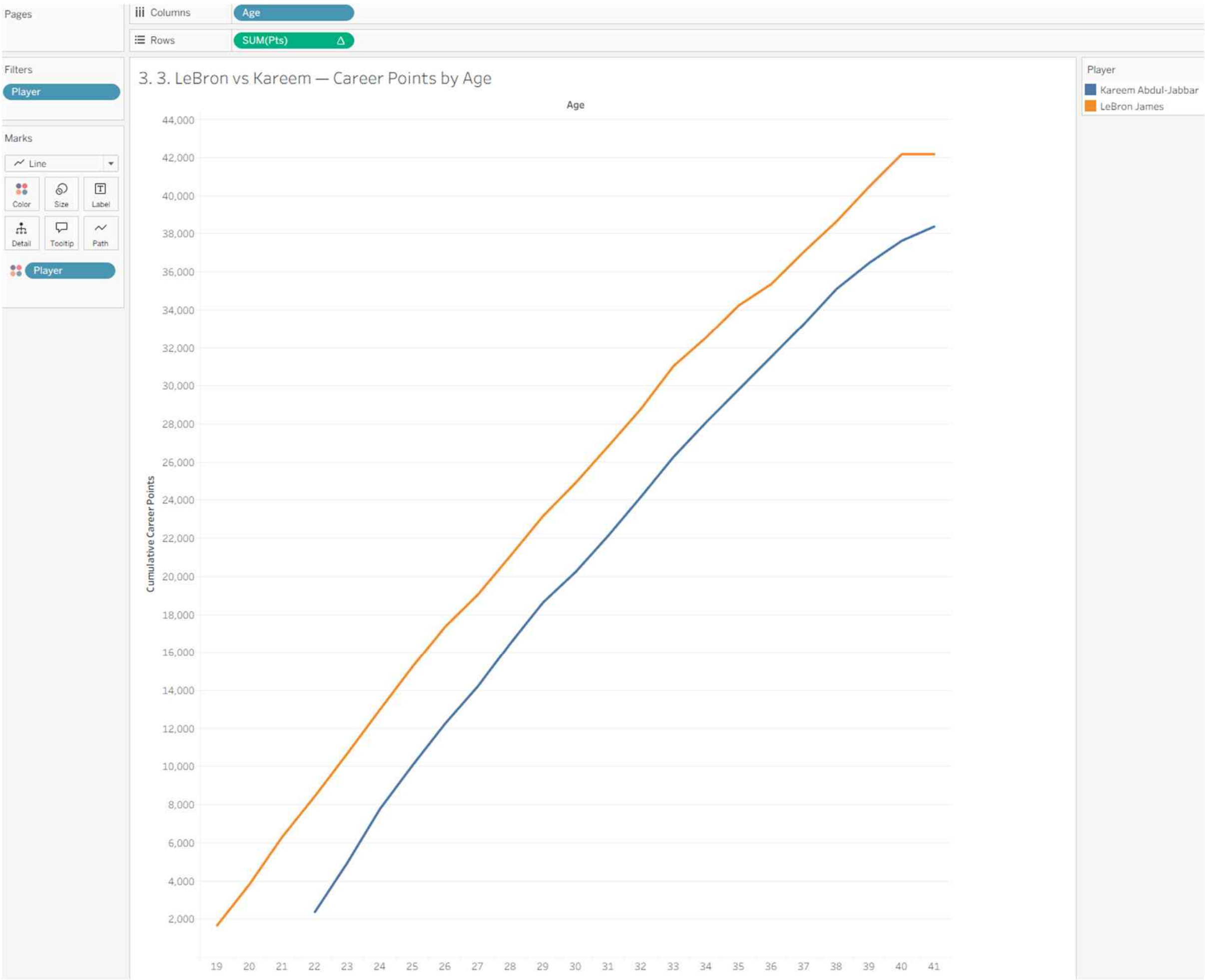
Caption

This line chart compares LeBron James and Kareem Abdul-Jabbar’s cumulative career points by age.

Explanation

I filtered for LeBron and Kareem and grouped the data by Age. Using RUNNING_SUM(SUM(PTS)) in Tableau, I calculated cumulative career points for each age. The visualization shows that:

- LeBron’s curve is consistently above Kareem’s from age 19–34.
- Kareem’s curve extends to older ages, showing his exceptional longevity.
- LeBron’s early scoring volume helps explain why he eventually passed Kareem despite starting his career later in the modern 3-point era.



Columns: Age

Rows: SUM(Pts) Δ

Filters: Player

Marks: Line

Color: Player

Figure A-3. LeBron vs Kareem : Career Points by Age (Tableau)

4. Overall Findings and Reflections

Across the three visuals, I explored both NBA scoring history and the specific career comparison between LeBron and Kareem.

- Visualization 1 shows the historical ranking and confirms LeBron and Kareem as outliers in career scoring.
- Visualization 2 explains the league context and how shot selection evolved over time.
- Visualization 3 illustrates two superstars’ different but converging paths to the all-time scoring record.

Together, the visuals provide a complete picture: the environment of scoring, where LeBron and Kareem stand historically, and how their careers developed year by year.

Part B - PCA, MDS, and LOF Analysis (Orange & Python)

1. Data & Preprocessing

For Part B, I used the Per 36 Minutes dataset from the Kaggle "NBA, ABA, BAA Stats" project. This dataset normalizes all box-score statistics to a per-36-minute basis, enabling fair comparisons across players with different playing times.

Steps Taken

1. Variable selection
   - I selected all numeric per-36 variables: FG, 3P, FT rates, rebounds, assists, steals, turnovers, points, etc.
   - ID columns retained: *Player Name*, *Season*.
2. Cleaning
   1. Replaced *NaN*, *inf*, *–inf* with np.nan.
   2. Dropped rows with missing numeric values.
   3. Final clean sample ≈ 20,040 players.
3. Scaling
   - Applied StandardScaler() to all numeric columns.
   - Necessary because PCA, MDS, and LOF are distance-based

2. PCA Analysis

2.1 Scree Plot (Explained Variance)

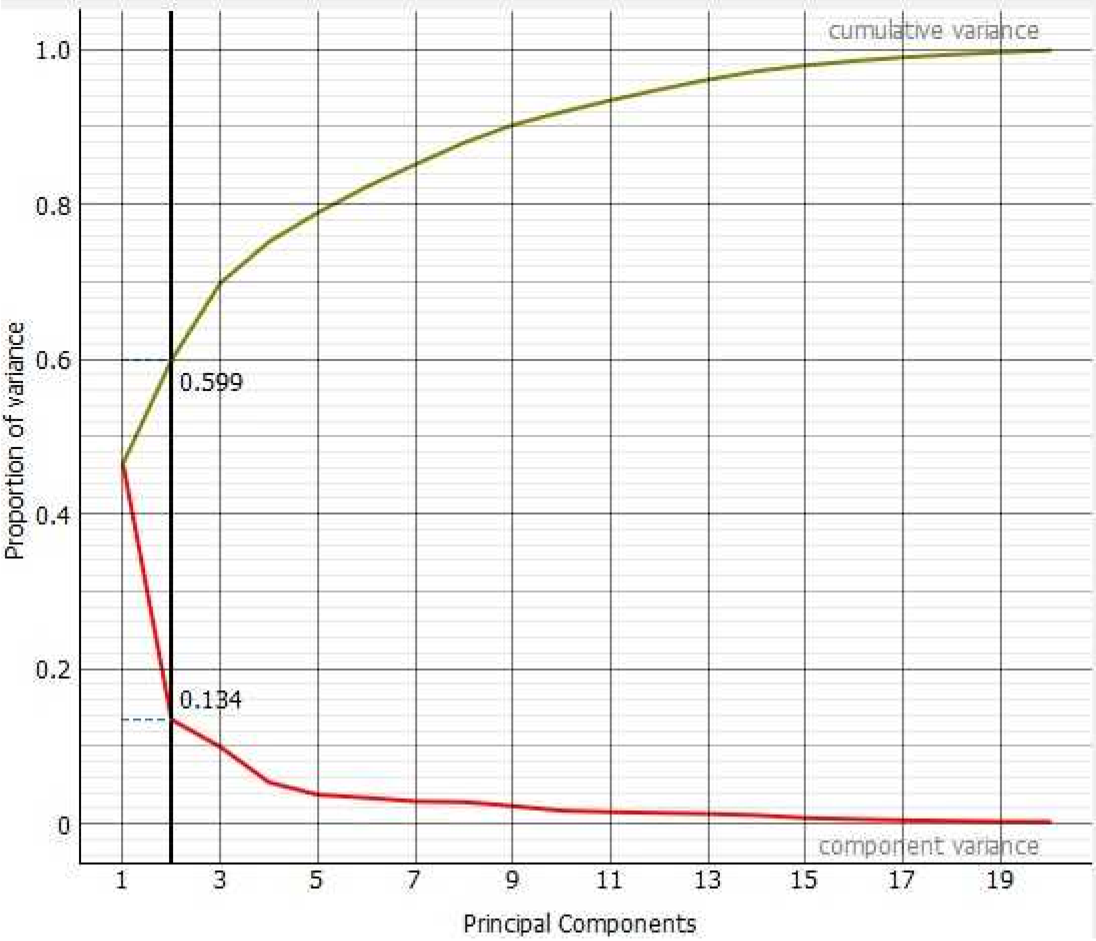


Figure B-1. Scree Plot for PCA (Explained Variance)

The scree plot shows:

- PC1 explains ~47–50% of variance
- PC2 explains ~13–14%
- Cumulative variance surpasses ~60% by PC2 and ~75% by PC3

Thus, the PCA1–PCA2 plane captures the main structure of skill variation among players.

2.2 PCA Visualization (Colored by Position)

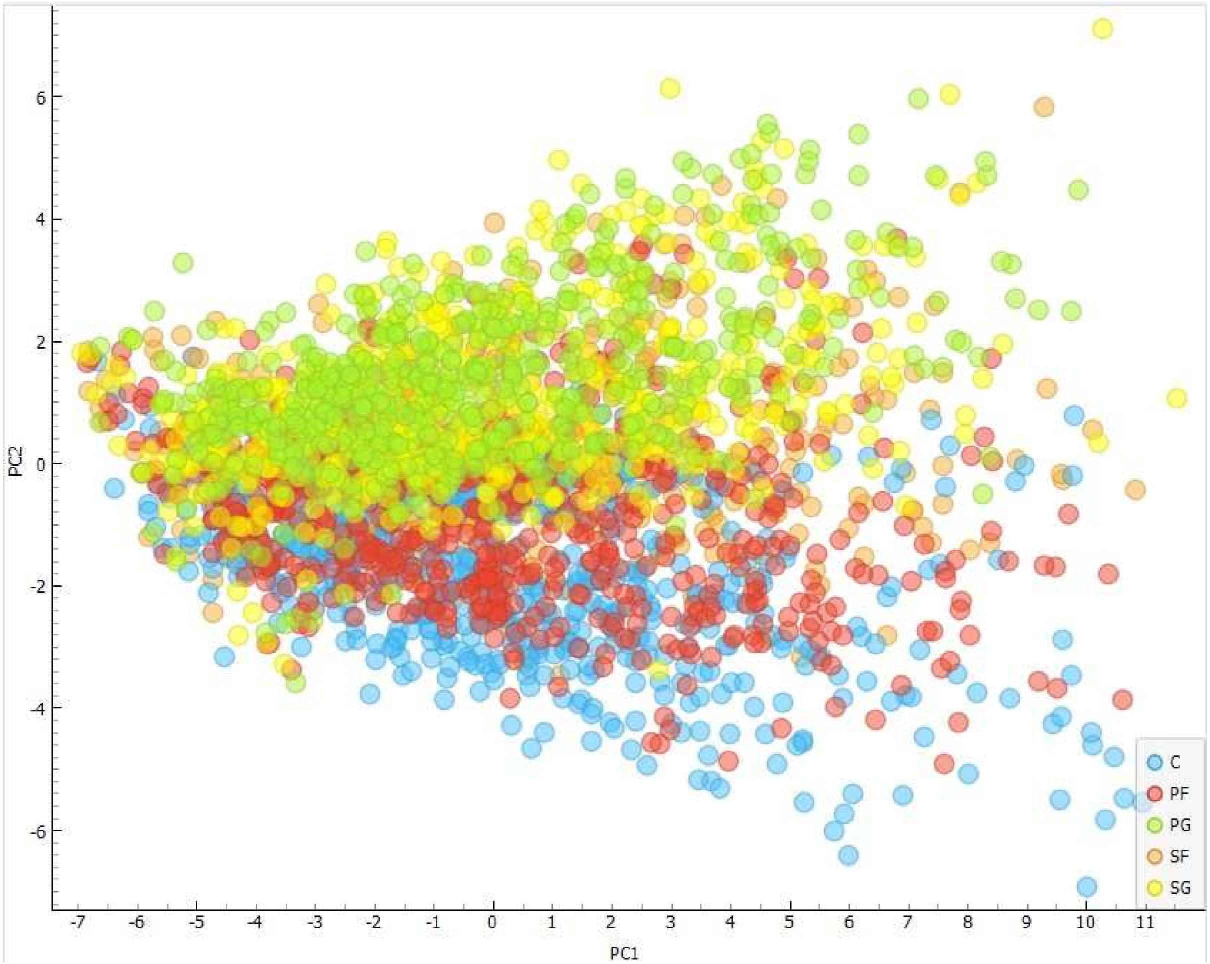


Figure B-2. PCA Scatter (PC1 vs PC2)

Interpretation:

- The PCA shape is an elongated oval, showing continuous skill variation rather than hard clusters.
- Rough positional pattern:
  - Centers (C) cluster lower (left/bottom), associated with rebounds & blocks.
  - Guards (PG/SG) appear on the upper/right area, corresponding to assists, steals, 3PT shooting.
  - Forwards (SF/PF) spread across the dense middle area.

This matches our basketball intuition: PCA is grouping players by broad role and statistical tendencies.

3. MDS Projection

3.1 MDS 2D Layout (Sample n = 3000)

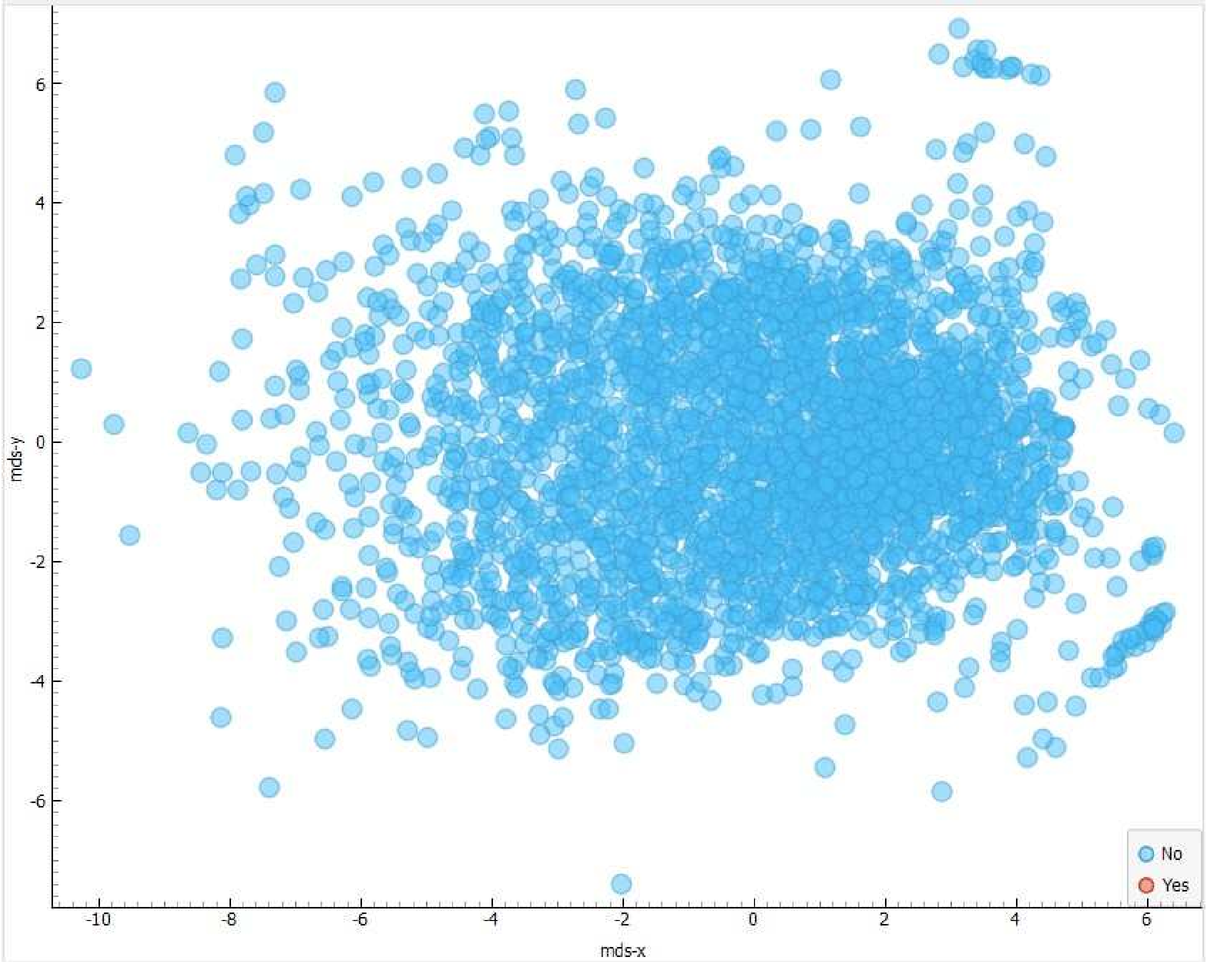


Figure B-3. Scatter Plot (MDS)

Because MDS is O(n²) in time and memory, I used a random sample of 3000 players. Observations:

- The MDS shape is also "blob-like," but more circular/compact than PCA.
- Distances emphasize local similarities, so extreme players appear slightly more separated.
- Compared to PCA, MDS loses the axes' "interpretability," but preserves pairwise distances better.

3.2 PCA vs MDS Comparison

Aspect	PCA	MDS
Goal	Maximize variance	Preserve pairwise distances
Axes meaning	Interpretable linear combinations	No interpretable axes
Shape	Elliptical	Rounder, dense center
Clusters	Better separation by position	More local similarity-based

4. Outlier Detection via LOF (Local Outlier Factor)

4.1 Method

- Used LocalOutlierFactor(n_neighbors=20, contamination='auto').
- LOF > 1 indicates "more outlier-like."
- Added LOF score to dataframe

4.2 LOF Scatter in PCA Space

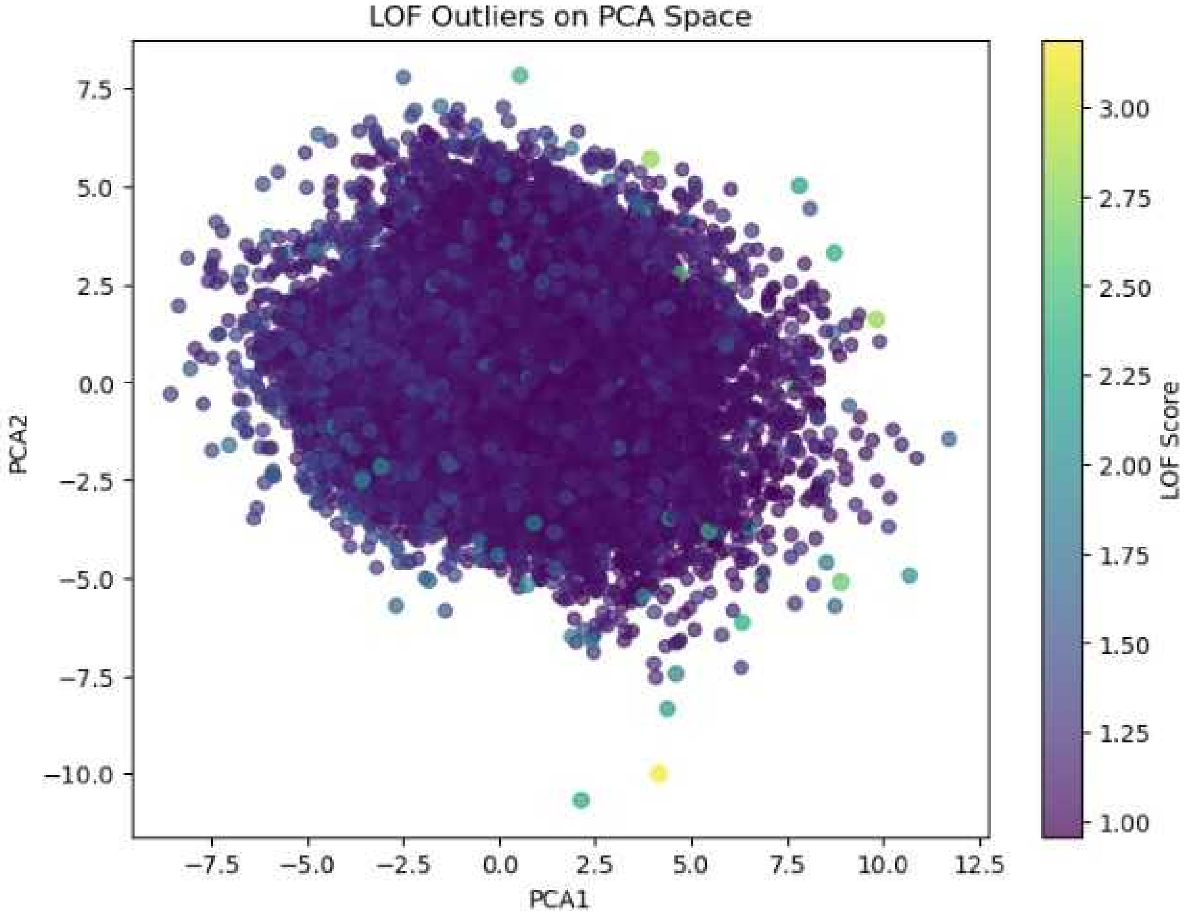


Figure B-4. LOF Outliers

Lighter dots represent higher LOF scores, indicating more unusual per-36 statistical profiles.

Color = LOF score (yellow = strong outlier).

Most players fall near LOF ≈ 1.0, with a few noticeable outliers.

4.3 Top 5 Outliers (with context)

The LOF method identifies players whose per-36 statistical profiles differ strongly from the league-wide patterns.

Below are the top 5 outliers:

Player	Season	Position	LOF	Notes
Darrell Armstrong	1995	PG	3.19	Very high steals & assists per-36 with low minutes; highly active defensive profile.
Paris Bass	2022	SF	2.82	Played only 2 games; inflated per-36 scoring/rebounding, small-sample anomaly.
Andre Barrett	2005	PG	2.79	Low efficiency + relatively high assists; unusual guard profile.
Nate Robinson	2010	PG	2.64	Extremely high usage bursts; high 3PT attempts & steals for his position/size.
Ryan Anderson	2017	PF	2.55	Pure stretch-four: extreme 3PT volume with low rebounding.

5. Findings & Insights

- PCA reveals that most NBA players fall on a continuous spectrum of scoring, shooting, rebounding, and playmaking abilities—rather than strict clusters.
- Player positions loosely group in PCA space, validating the idea that Per-36 statistics reflect on-court roles.
- MDS highlights local relationships, showing clusters of similar players but with less linear separation.
- LOF successfully identifies rare statistical profiles, often belonging to:
  - Low-minute specialists
  - Defensive or rebounding-only players
  - Players with extreme efficiency values

Together, these methods offer a strong picture of NBA skill similarity and structural patterns in player performance.