Clustering of NBA Players to Investigate Changes in Player Specialization from 1981 to 2021

Joseph Mintz

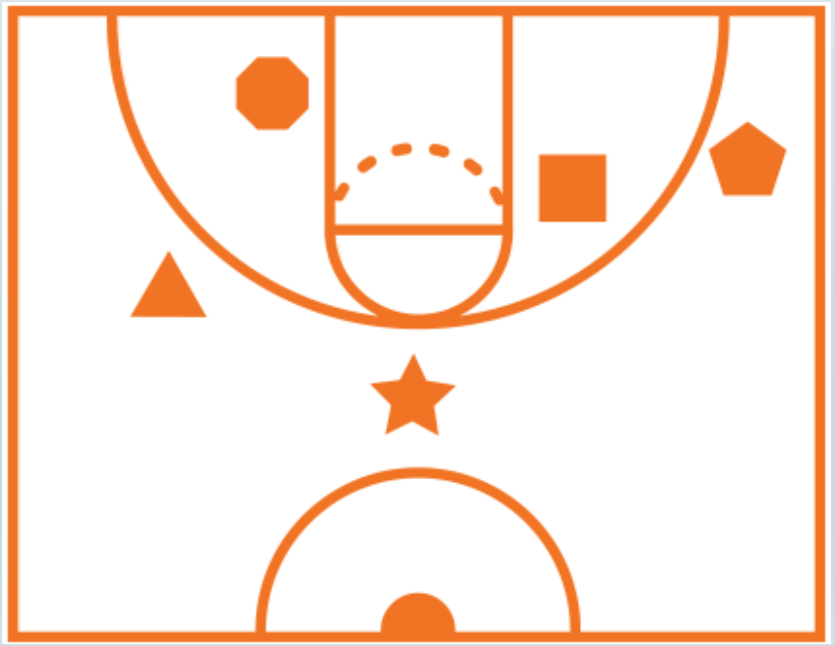
March 12, 2022

Overview and Problem Statement

- In the NBA, the specialization of players based on position has given way over the years to a demand for more multidimensional players, e.g. centers that can shoot 3-pointers.

- How does grouping of NBA player performance stats align with player position (PF, SG, SF, PG, C)?

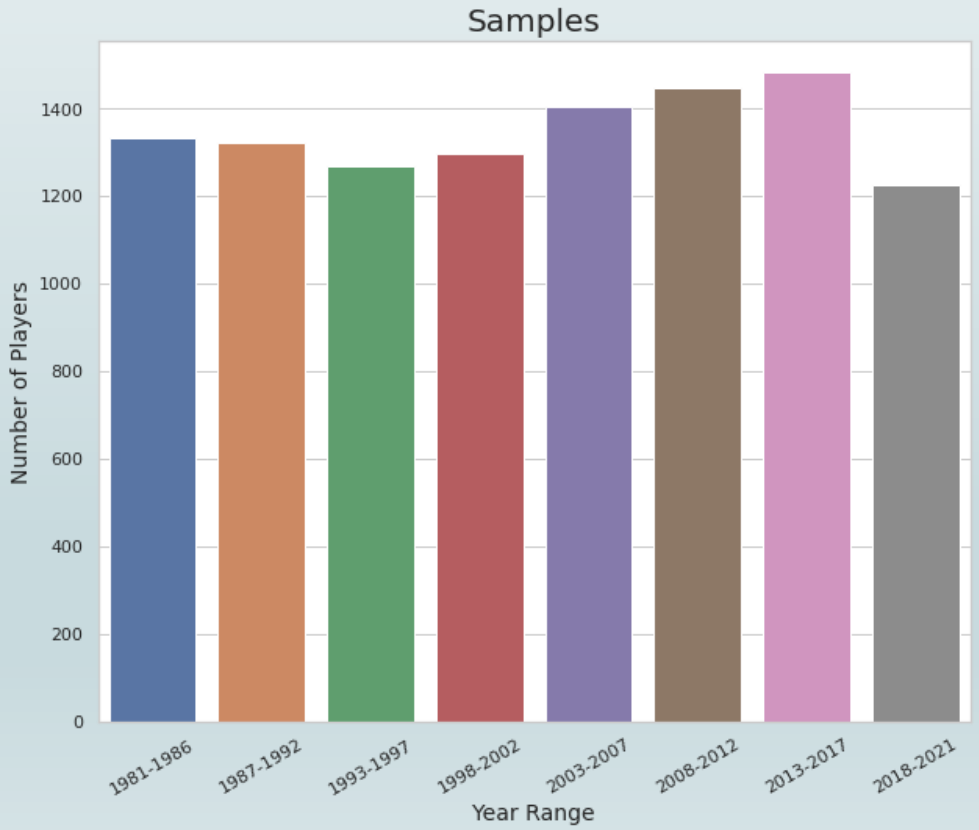
- Can clustering techniques clearly identify distinct groups of players, and does that clarity diminish over the decades as players have become less specialized?



Data

- NBA player stats were scraped from www.basketball-reference.com for the years 1981-2021.
- Only individual player performance stats were kept as features (22 total features):

FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS

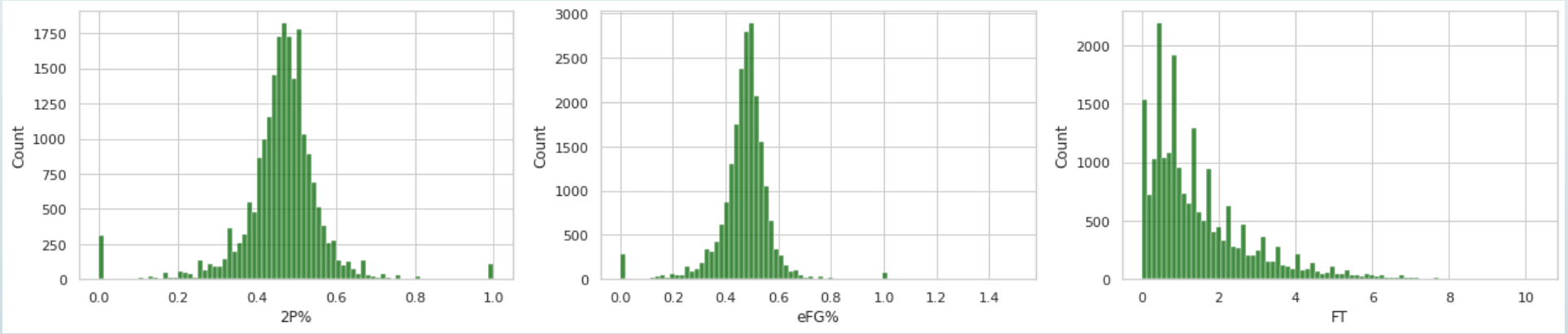


- The data was split into eight different year ranges to analyze changes in clustering over time.
- The eight samples are well-balanced.
- Total number of player entries = 10782

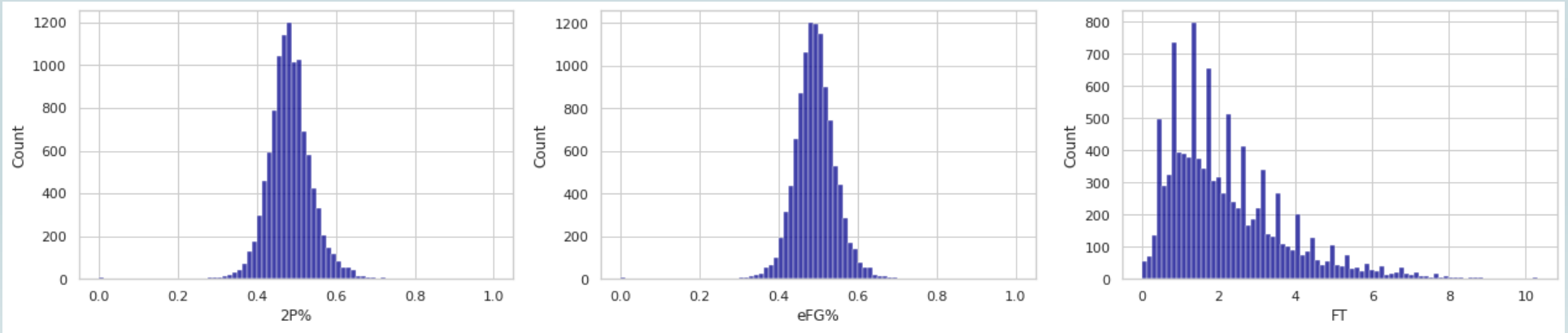
Data

- The data was subset by “Games Started” (GS), keeping only players that started in 8 or more games in the year.
- This helped to eliminate outliers and noise in the distributions.
- Distributions of 3 of the 22 features:

Before Subset

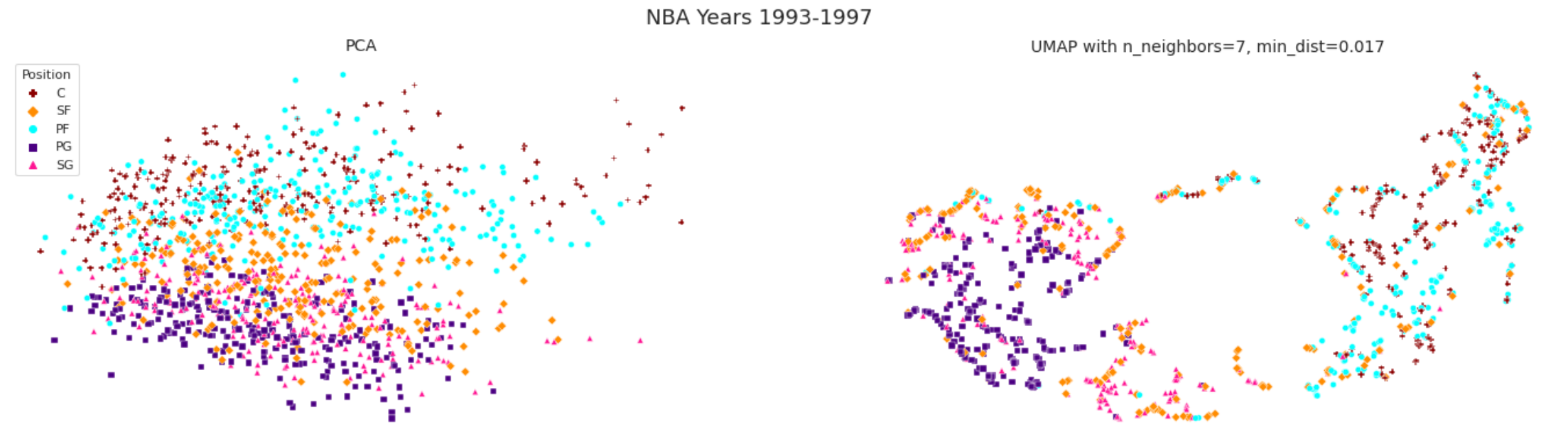


After Subset



Dimensionality Reduction

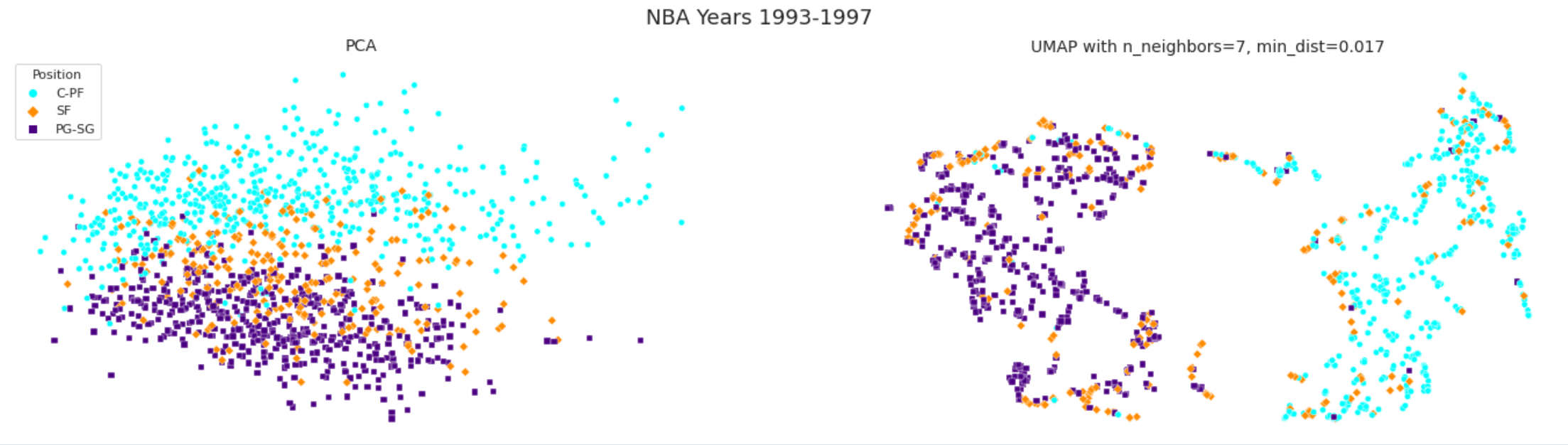
- Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) are used for visualization.
- Player positions (PF, SG, SF, PG, C) are set as the ground truths.
- PCA and UMAP colored by ground truths:



- The PCA and UMAP point to a blurring of player positions in the early years, such as 1993-1997.
- The greatest similarities appear to be between PG and SG, and between PF and C.
- The greatest dissimilarity appears to be between PG and C, which is expected.

Dimensionality Reduction

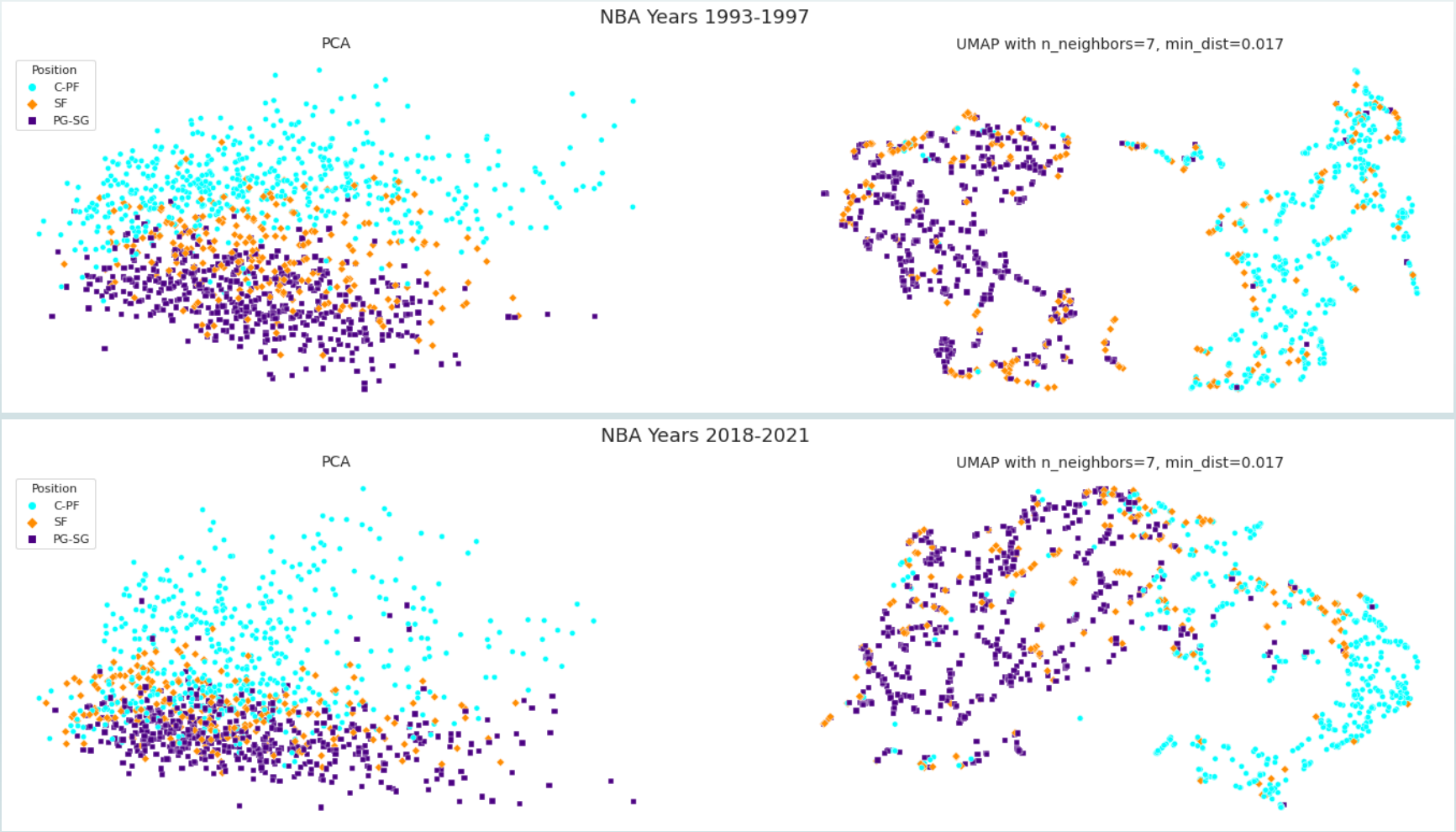
- Due to local similarity observed with 5 ground truths, the study was refocused with an aim to get a 3-cluster solution.
- Ground truths were reduced to 3:
  - Small Forward (SF)
  - Center (C) combined with Power Forward (PF). Note, existing C-PF players were added back into the dataset.
  - Point Guard (PG) combined with Shooting Guard (SG). Note, existing PG-SG players were added back into the dataset.
- This study focuses on two main changes over time:
  - Change in dissimilarity, e.g. the C position changing relative to PG
  - Change in the 3-cluster solution relative to the ground truths
- Updated PCA and UMAP colored by ground truths:



Dimensionality Reduction: Comparison of Years

The PCA and UMAP show that, by the years 2018-2021, C-PF is more similar to PG-SG, and SF is more intertwined in the middle ground, supporting the notion that NBA player roles have become increasingly blurred over the years.

Colored by Ground Truths



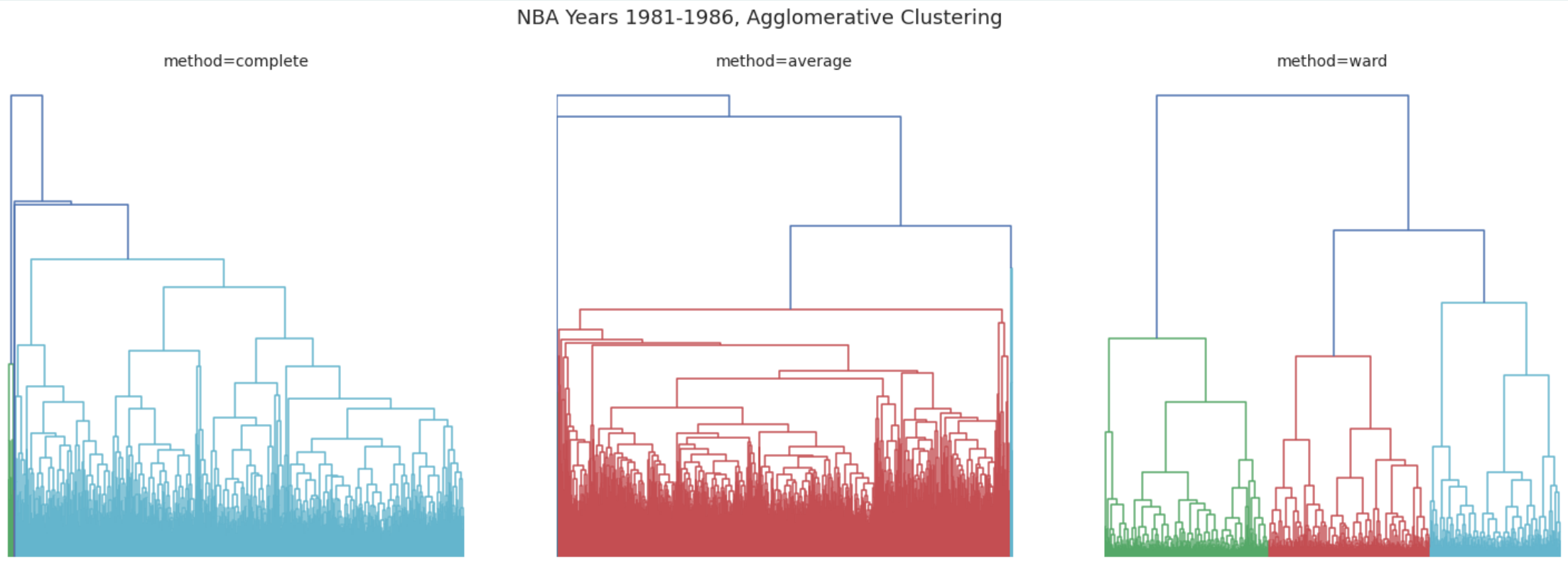
Modeling

- K-means, Agglomerative Clustering, and Gaussian Mixture Models (GMM) were tested for varying parameters.
  - 25 total models tested per year range
- The models were tuned to the 1981-1986 data.
- Adjusted Rand Index (ARI) was used as the primary evaluation metric, since player positions are considered as ground truths.
- Even though the ground truths are 3 classes, 4 and 5-cluster models were evaluated as well.
- Based on ARI, the best performing 3, 4, and 5-cluster models are highlighted:

	years	method	n_clusters	ARI	silhouette_score	params
0	1981-1986	GMM	5	0.2902	0.1256	{'covariance_type': 'tied', 'n_components': 5,...
1	1981-1986	k-means	5	0.2593	0.1795	{'n_clusters': 5, 'random_state': 123}
2	1981-1986	k-means	4	0.2550	0.1922	{'n_clusters': 4, 'random_state': 123}
3	1981-1986	GMM	4	0.2451	0.1838	{'covariance_type': 'spherical', 'n_components...
4	1981-1986	GMM	5	0.2405	0.1546	{'covariance_type': 'spherical', 'n_components...
5	1981-1986	agglomerative_clustering	5	0.2397	0.1102	{'linkage': 'complete', 'affinity': 'cosine', ...
6	1981-1986	agglomerative_clustering	3	0.2187	0.1762	{'linkage': 'ward', 'affinity': 'euclidean', '...

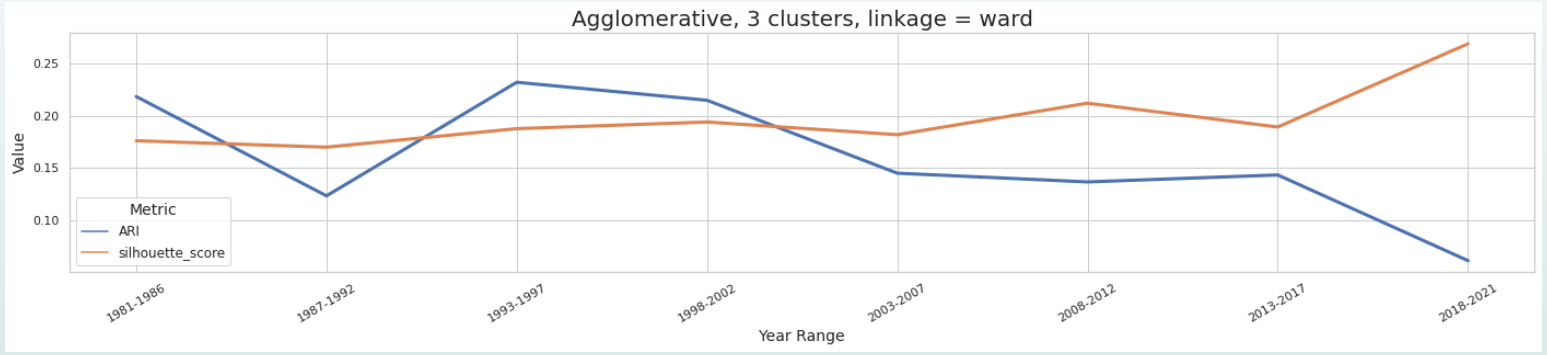
Modeling

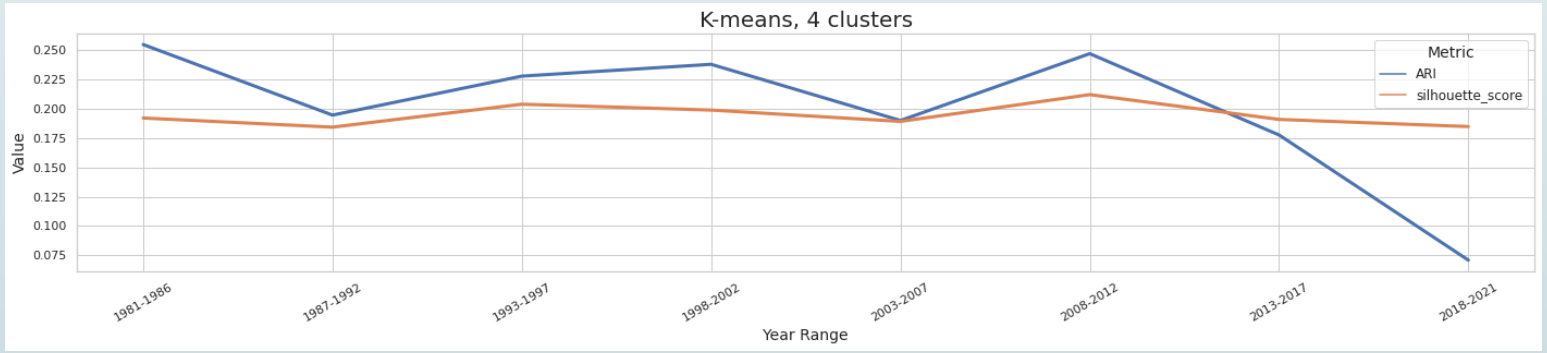
The dendrogram illustrates the clarity of the 3-cluster solution obtained by agglomerative clustering with ward linkage.

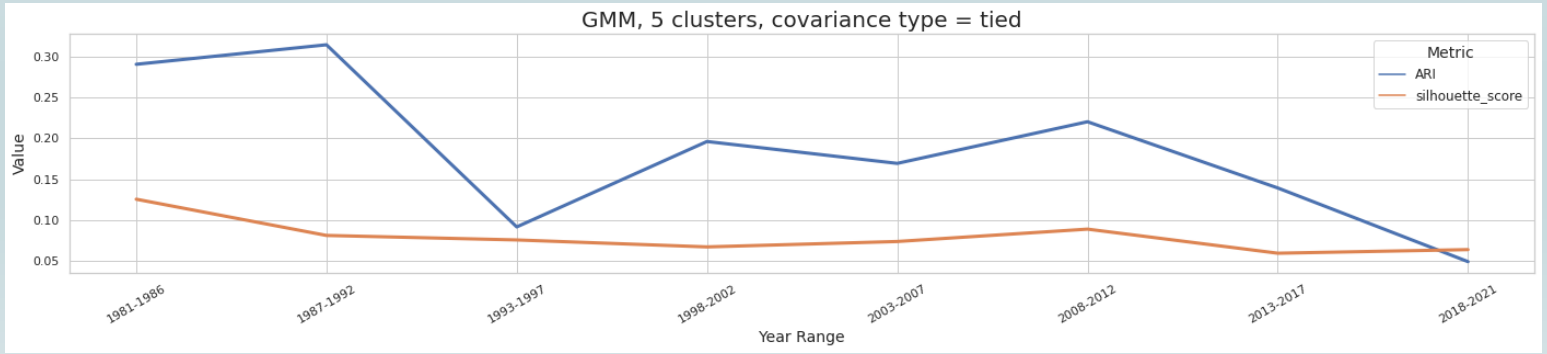


Modeling: Comparison of Years

- Each top model shows the ARI decreasing over the years, while the silhouette score does not change as much.
- Decreasing ARI supports the notion that player roles have blurred or become less specialized relative to position.
- The drop in ARI is most evident between the last two year ranges.
- Consistent silhouette score points to the ability of the models to find player groupings with about the same level of clarity, regardless of year range, even though groupings change relative to ground truths.

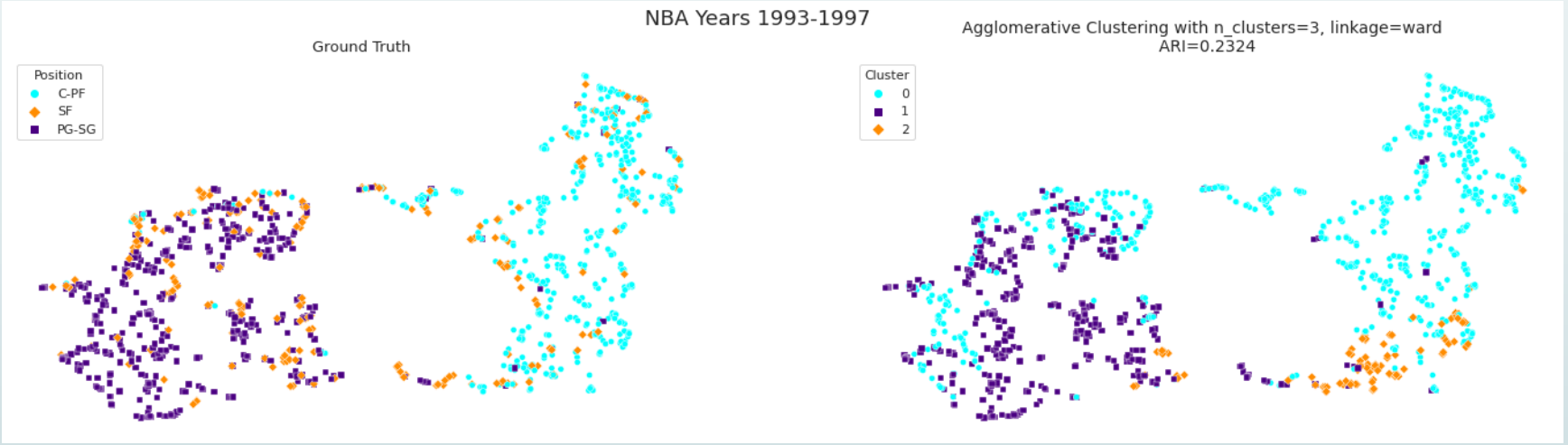


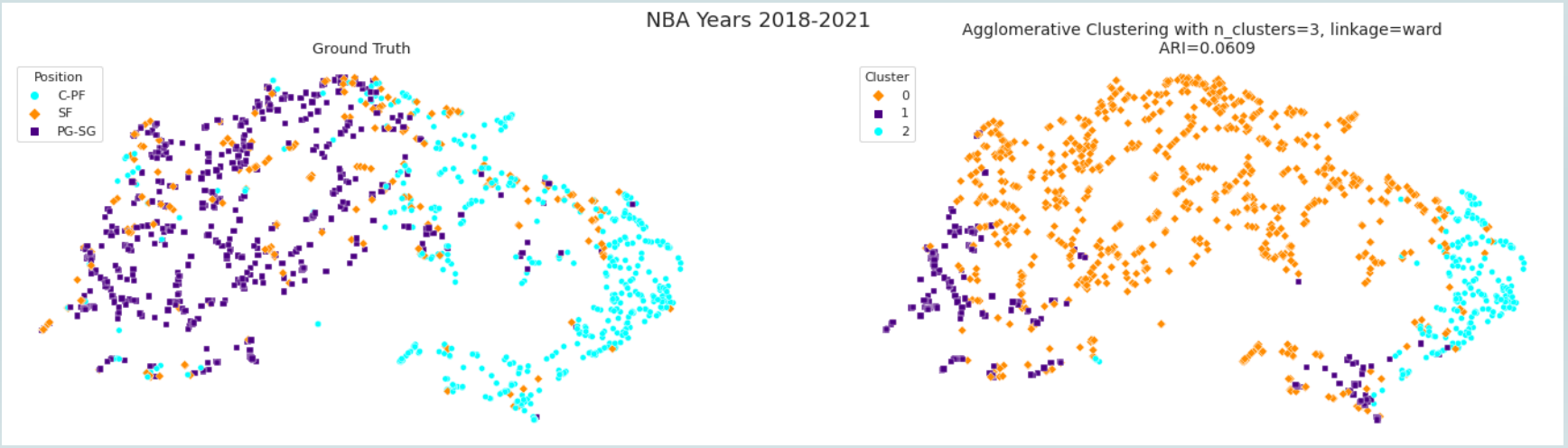




Change in 3-cluster Solution Relative to Ground Truth

The UMAP demonstrates the degradation of ARI between year ranges as the clusters identified by the model have shifted relative to the ground truths.





Conclusions and Next Steps

- This study has provided evidence supporting the notion that from the 1980s to the present, NBA player roles have blurred or become less specialized relative to position.
- Visual evidence is seen with PCA and UMAP, which illustrate that, by the years 2018-2021, C-PF is more similar to PG-SG, and SF is more intertwined in the middle ground.
- Numerically, further evidence is provided with the ARI metric, as it clearly decreases over time, indicating player groupings changing relative to position.
  - The drop in ARI is most apparent between the last two year ranges, 2013-2017 and 2018-2021.
- Based on ARI, agglomerative clustering yielded the best performing 3-cluster solution; thus, it was used to investigate the cluster changes over time, given the 3 ground truths.
- A practical use case for this study is with sports analysts or historians researching the evolution of professional basketball.

- Recommended next steps:
  - Focus on a more granular view of the years, especially in the later years where the blurring of player roles is most apparent.
  - Explore individual clusters to determine what features factor most into the groupings, and how the characteristics of those groups change between the year ranges.

Questions?