Clustering NBA Players Offensively

Nehal Chigurupati

November/December 2023

1 Objective

The objective of this project is to cluster together NBA players based on their offensive game, as determined by their performance on the following types of plays:

- Isolation
- Cut
- Off screen
- Pick and roll ball handler
- Pick and roll roller
- Post up
- Spot up

The clusters could be used by a team in a number of ways. If a front office is looking for a hidden gem in free agency to replace a player, they could look at everyone within the cluster associated with the departing player. Strategy-oriented coaches could use the model to craft lineups and plays, and development coaches could use clusters to find appropriate film in the off-season.

2 Methodology

For each play type, the following statistics over the 2023-24 season were collected on a per player basis (averaging over games):

- number of possessions
- frequency
- average points per possession
- total points

- field goals made
- field goals attempted
- field goal percent
- effective field goal percent
- free throw frequency
- turnover frequency
- shooting foul frequency
- and one frequency
- score frequency

For each play type, the corresponding player data was scaled and normalized. Three different types of models were tested: KMeans, spectral clustering with affinity matrix calculated via RBF, and spectral clustering with affinity matrix calculated via nearest neighbors. For each model type, seven different models were trained, one for each play type.

For KMeans, the number of clusters for each play type model was computed via the maximum of silhouette scores over a range from 2 to 20 clusters. The same technique was used for spectral clustering.

Next, all the models were trained. The average of all Davies-Bouldin scores for the KMeans models was computed, and compared to the same averages for the Spectral RBF and nearest neighbor models. The average score was lowest on the KMeans models, leading to it being chosen.

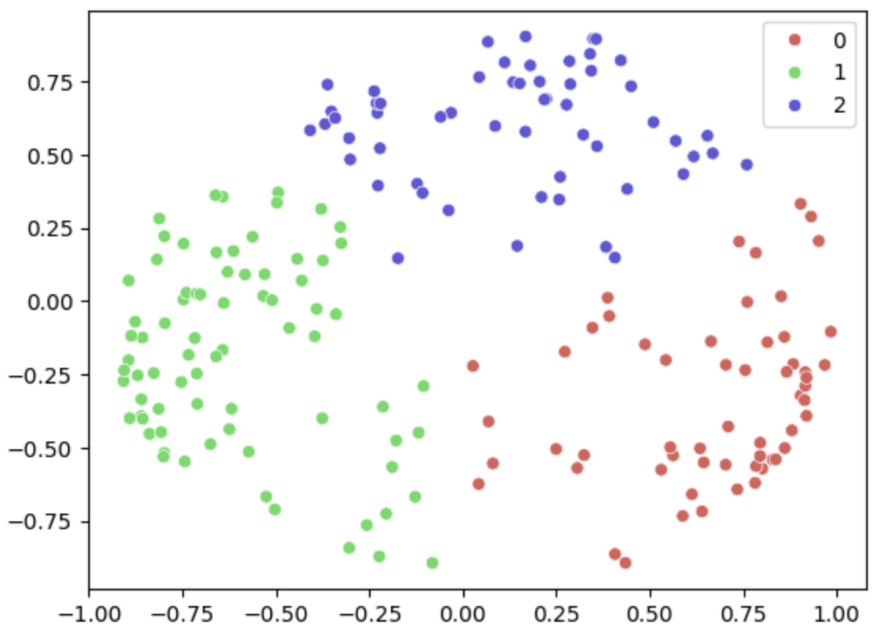
3 Results

The KMeans models were trained on the following numbers of clusters:

Iso: 3 clusters
Cut: 3 clusters
Off screen: 3 clusters
PnR Handler: 3 clusters
PnR Roller: 3 clusters
Post up: 3 clusters
Spot up: 3 clusters

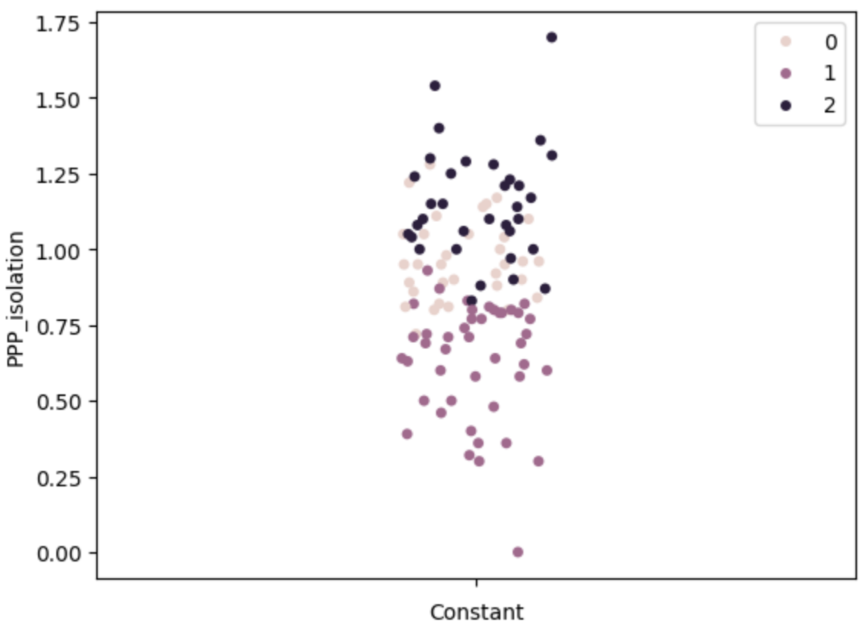
The average Davies-Bouldin score over all KMeans models was 1.39, indicating a moderately strong clustering. To visualize the clusters (a difficult task, because the feature space is greater than three dimensions), two different techniques were used.

First, PCA on two components was run on the data to project into 2-space, then displayed on a scatterplot with points colored in accordance with their label. The clusters were well visible. Below is the scatterplot for pick-and-roll handler plays. The axes are the two PCA components, and the colors are the labels from clustering.

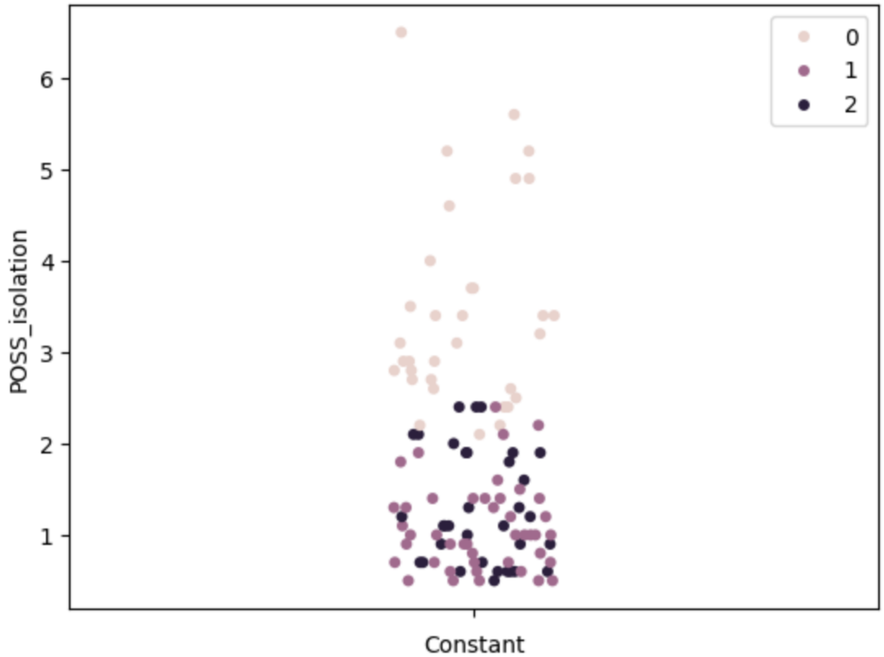


All of the other play type models were similarly well clustered in PCA visualization.

Another visualization technique employed was to plot how clusters vary over a specific stat. For example, let's view how the isolation model clusters vary with points per possession.



So, broadly, we can see that cluster 0 has a PPP on isolation attempts between .75 and 1.25, cluster 1 between 0 and .75, and cluster 2 between .8 and 1.75. We could also compute this same plot for a different statistic. Let's see how the same clusters vary over the player's average number of iso possessions per game.



Cluster 0 players attempt a lot of isolation plays (2+ per game), while cluster 1 and 2 players attempt fewer than 2 per game. Broadly speaking, cluster 0 players attempt a lot of isolation plays, with middle-of-the-road points per possession. Cluster 1 players attempt very few isolation plays, and generally don't generate many points off such plays. Cluster 2 players don't attempt many isolation plays, but efficiently generate points out of such possessions.

Examples of players in cluster 0 include Darius Garland, Collin Sexton, Jalen Brunson and Dennis Schroder. A gut check confirms the clustering: all of these players attempt a lot of isolation plays, and all of them are shifty enough to score at a relatively good efficiency on an island. However, they don't have the size of a player like Jokic (he belongs to cluster 2), who can easily generate points in isolation.

4 Further work

Next, I want to create some sort of holistic visualization that takes players and shows the proficiencies/deficiencies associated with their clusters.