Executive Summary: Clustering NBA Regular Season and Postseason Play Styles

Malcolm Gaynor (gaynor1@kenyon.edu) and Parker Gibbons (gibbons1@kenyon.edu)
Kenyon College, STAT 306, Prof. Brad Hartlaub

Objectives:

In our analysis, we investigate methods for clustering team styles of play. However, we specifically want to dive into the differences in team styles of play between the regular season and postseason. There are various questions related to this that our clustering will seek to answer. For example, do teams change their style of play in the postseason as compared to the regular season? What aspects of their style do they change? Does either changing style of play or not changing style of play have adverse effects in the context of postseason success?

Data and variables:

Data is from the 5 most recent NBA regular seasons and postseason (2018-2019 to 2022-2023). The variables that were considered (all on a per-game basis) were: field goal, three point, two point, and free throw attempts, offensive, defensive, and total rebounds, assists, steals, blocks, turnovers (offensive), personal fouls, average age, pace, free throw attempt rate (number of free throw attempts per field goal attempt), average shot distance, percent of shots that are two pointers, percent of shots that are three pointers, percent of shots that are attempted between 0 and 3 feet, 3 and 10 feet, 10 and 16 feet, and over 16 feet from the hoop, percent of three point and percent of two point shots that are assisted, percent of shots that are dunks, percent of three pointers that are corner threes, bench minutes, and bench field goals attempted. All of the statistics were taken from basketball-reference.com except for bench minutes and bench field goals attempted, which were taken from nba.com.

We chose stats such as three pointers attempted instead of three pointers made or three point percentage in order to pick up on the team's play style choices, not their success or talent. In other words, we care about how often a team chooses to take threes, not how often they make them. Also, because the goal of our analysis focuses on the difference between postseason and regular season play styles, we will only consider data from teams that went to the postseason during each respective season. Thus, for each season, there are 16 teams considered, and all 16 teams have two sets of data, one from the regular season and one from the postseason.

Preliminary comparisons:

Before we begin clustering, we will use some basic statistical techniques to analyze the differences in regular season and postseason play styles, just to get a general idea of what to

expect in our clusters. The following analysis does not directly seek to answer the questions posed above, rather to inform our expectations for the following clustering.

First, we will complete a MANOVA test to compare the mean statistics between the regular season and postseason statistics, in order to demonstrate that there is a general difference in playstyle in the regular season and postseason.

With a p-value approximately zero, we are able to reject the null hypothesis that the true means of every statistic are the same in the regular season and the postseason. This is fully in line with our expectations, especially considering the nature of MANOVA tests, where the groups are considered different if just one of the parameter mean's is statistically significantly different. Also, we will not dive deeply into the conditions of this test, because the findings do not directly relate to the conclusions of our analysis, but it is important to keep in mind that MANOVA testing requires strict conditions.

Next, we will complete logistic regressions to explore specifics about the differences between regular season and postseason playstyles. Basically, the response variable is whether or not the given statistics correspond to the regular season or the postseason. We will separately consider each input variable that we are considering in its own simple logistic regression model. Then, using k-Fold Cross-Validation, we measure the accuracy of each model to determine which variables are independently better at predicting if the stats came from the regular season or the postseason.

Using k-Fold Cross-Validation with 10 folds, the accuracy of nine variables are above 65%: Bench field goals attempted (76%), Bench minutes (75%), assists (71%), total rebounds (70%), percent field goals attempted from between 0 and 3 feet from the hoop (68%), field goals attempted (67%), personal fouls (67%), defensive rebounds (66%), and percent field goals attempted that are dunks (66%). Also the accuracy of four variables are below 55%: Three point field goals attempted (49%), percent of field goals attempted from between 3 and 10 feet from the hoop (50%), age (54%), and percent of field goals attempted from 16 feet or further from the hoop (54%). The rest of the variables are between 65 and 55% accuracy.

We are not sure that these variables will be especially important in the clustering, they are just variables we will pay special attention to. Again, this is all just a preliminary investigation that will be used to inform the clustering analysis, which will dive deeper into our questions about clustering play style in the postseason compared to the regular season, and whether or not changing this play style in the postseason is a good idea.

Clusters:

To cluster the different types of playstyles, we chose to use K-means clustering in R, which is a popular unsupervised machine learning algorithm used for partitioning data points into distinct groups called clusters. This is done by iteratively assigning each data point to the nearest cluster centroid, typically based on Euclidean distance, and recalculating the centroids until they converge. This results in a set of clusters where data points within each cluster are more similar to each other than to those in other clusters. To determine the number of clusters, we used the elbow method, which is conducted by plotting the explained variation as a function of the number of clusters and using the elbow of the curve to determine the number of clusters to use. Below are summaries of the clustering of all regular seasons, postseason, and both combined along with qualitative descriptions of each cluster.

All Regular Seasons from 2018-2019 to 2022-2023 Clustered		
Cluster	# teams	Description
1	33 (1 champion, 5 runner ups)	Lowest pace, least aggressive, least mistakes, least bench minutes, least bench production
2	16 (2 champions)	Oldest, highest pace, most 3s attempted, most aggressive, most mistakes
3	31 (2 champions)	Youngest, moderate pace, least 3s attempted, moderately aggressive, most bench minutes, most bench production

All postseason from 2018-2019 to 2022-2023 Clustered		
Cluster	# teams	Description
1	20 (1 champion, 2 runner ups)	Moderate pace, most possessions, shortest distance shots
2	19 (2 champions, 2 runner ups)	Oldest, higher pace, most physical
3	15	Older, slower pace, least physical, high number of 3s attempted
4	10 (1 champion)	Least shots, least rebounds, slower pace, farther shots,

5	16 (1 champion, 1 runner up)	Youngest, highest pace, most shots, most 3s attempted, most rebounds, most dist

All Regular Seasons and postseason from 2018-2019 to 2022-2023 Clustered				
Cluster	# teams	Regular Season	Postseason	Description
1	69	4 champions 2 runner up	1 champion 1 runner up	Slowest pace, least shots attempted, most possessions, least aggressive, most turnovers, least bench minutes and production, most physical
2	39	0 champions 0 runner up	3 champions 3 runner up	Highest pace, most shots attempted, least 3s attempted, least possessions, shortest distance shots, most bench minutes and production
3	52	1 champion 3 runner up	1 champion 1 runner up	High pace, most rebounds, most 3s attempted, least physical, most turnovers, high bench minutes and production

Changing play styles:

Along with the above clusters, we also clustered each NBA season separately to investigate which teams changed their play style in the postseason. First, for each season, we clustered all 16 regular season and postseason teams. Then, each team was divided into one of two groups, depending on if they were in the same cluster in the postseason and the regular season, or if they changed clusters. Functionally, this represents whether or not the teams significantly modified their play styles or strategies in the postseason as compared to the regular season.

Of the 80 teams considered (16 postseason teams during 5 seasons), 47 teams ended up in different clusters in the postseason, while 33 stayed in the same. The teams that were in different clusters (i.e. changed their play style in the postseason) had a mean of 5.7 postseason wins, with a standard deviation of 5.2 and a median of 3. The teams in the same cluster in both the regular season and the postseason had a mean of 4.7 postseason wins, with a standard deviation of 4.2 and a median of 4.

Also, interestingly enough, the last four NBA finals winners all were in different clusters in the regular season compared to the postseason. However, in the 2018-2019 season, the champion Toronto Raptors were in the same cluster in both the regular season and the postseason. Also,

over the five years in the analysis, three of the teams that lost in the NBA finals were in different clusters in the postseason as compared to the regular season.

For example, the 2023 NBA Finals Champion Denver Nuggets were in cluster 2 during the regular season, and cluster 3 during the postseason. The biggest difference between these two clusters is that cluster 3 shoots more three pointers, is more aggressive on rebounds, and plays at a higher pace than cluster 2.

Already, it does not appear that there is a statistically significant difference between the mean number of postseason wins in each group. This suspicion is confirmed when conducting a Two-Sample t-Test to test for a difference in means between the two groups. As expected, with a large p-value of about 0.36 (greater than our significance level of 0.05), we do not have significant evidence that there is a difference in mean postseason wins in the two groups. Therefore, it appears that there is no evidence that changing play styles has a negative or a positive effect on postseason wins.

Conclusion and areas for future research:

Our use of k-means clustering revealed three to four distinct play styles during the regular season and postseason among the sixteen postseason teams each year. Out of the 80 teams considered, 47 teams were placed in a different cluster in the postseason as compared to the regular season. While these teams averaged slightly higher success in the postseason, the difference was not significant, and thus it could not be concluded that teams that switched play styles were more successful in the postseason than the teams that did not change play styles.

There are a few simple ways to extend this study, such as to consider more variables (number of substitutions, percent of shots taken by the number one offensive player, etc.) and to expand the time frame to take into account more years. However, outside of this, one flaw in our analysis is that it doesn't take into account changes in play style due to opponent quality. To get around this, we could eliminate statistics taken from regular season games against non-postseason opponents.

In terms of the statistical methods used, it would be interesting to also take into account different clustering methods, such as hierarchical clustering. This probably would not have led to an enormous difference in our findings, but could have offered some more flexibility in terms of algorithm or parameter adjustments that potentially could have given us another perspective.

Finally, because we are taking on a more specific question than just clustering teams by play style, it might make sense to do some sort of supervised learning method to examine what aspects of a team's play style should and shouldn't be changed when moving on to the postseason.