CMPT 318 Final Project Report – Olympic Qualifying Men's Basketball Tournaments

Background

To qualify for the 2016 Men's Olympic Basketball Tournament in Rio there were multiple paths a country's team could take. In fact, there were 12 different tournaments that teams competed in for a berth to the Olympics. Those tournaments took place between 2014 and 2016 across Africa, the Americas, Asia, Europe and Oceania. Some were regional tournaments with teams participating based on location and others were open to all teams that not yet qualified for the Olympics. This create variability in tournament strength and gives us the opportunity to test the applicability of long-standing basketball sayings to different levels of competition.

The teams that qualified for the Olympics were the USA, Brazil, Australia, Nigeria, Venezuela, Argentina, Spain, Lithuania, China, Serbia, Croatia and France. By following the graphic below, you can see the path that each team took to secure their spot. Serbia, Croatia and France and not qualified yet when the graphic was made and fill the 3 remaining spots.

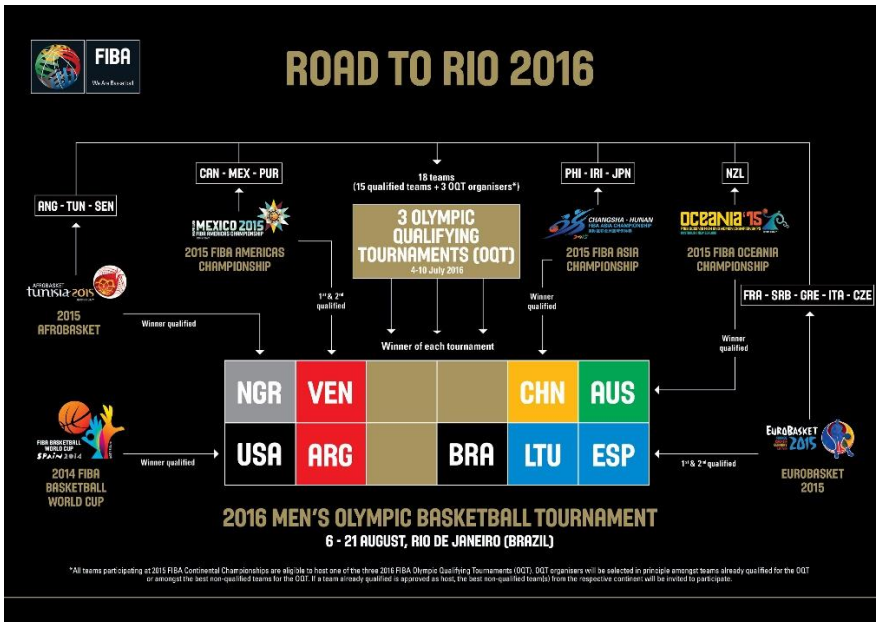


Photo Courtesy of FIBA Basketball: http://www.fiba.basketball/olympic/qualifying

Data

The Data for this project was provided by the Men's Canada Basketball program. It was scraped from the FIBA archives, (https://archive.fiba.com/) the official online provider of statistics for FIBA tournaments. There are 4 data files provided, cleanedboxscorestats.csv, gameinformation.csv, PlayerIDs.csv and tournamentlist.csv. All files can be linked together through various primary keys.

Box Score Stats

The box score file has information on both player performance and team performance for every game played throughout the qualification process. It describes shot accuracy and shot volume for different point valued shots (1, 2, 3), rebounding (offensive, defensive, total), assists, fouls, turnovers, steals,

blocks and points. There are 11395 rows of data for player performance and 956 rows detailing team performance. Unfortunately, the player data does not contain information about the player's position.

Game Information

The game information file contains information about all 478 games that were completed during the qualification process. It notes when and where the game was played, the playoff round of the game, and the home and away team. In most sporting events there is a phenomenon called home court advantage, and the home/away team variables are important. In this case the home/away team variables should not matter as they do not describe where the game is being played. Instead it indicates which team should wear dark and light-colored jerseys. Since the tournaments were often played in only one country, a home team advantage effect could be looked for by examining the location variable and whether the city where the game was being played was in the country of one of the teams playing. Often this is not the case.

Player IDs and Tournament List

Both the player ID file and tournament file are quite small. The player ID file contains a unique player ID to distinguish between players with the same name. Additionally, it contains the birthday of the given player. The tournament list simply contains the name of each tournament.

Problems

We will try to answer 2 main questions with this data.

1. ### Clustering for Player Positions

Due to the missing information about player position in the box score data set we can ask an interesting question about player positions. Instead of classifying players based on their height, and our opinion of the style of play they should have we will cluster players with similar of styles of play together. From these clusters we can give players a position based on their on-court performances. This let's the data speak for itself. We can then compare players more clearly.

Most of the data we have describes offensive style of play; therefore, we will be clustering players on their offensive performances. We will define offensive playing style by the following characteristics: 2 point shots attempted, 3 point shots attempted, free throws attempted, offensive rebounds, assists, and turnovers.

2. ### The game slows down in the playoffs. Or does it?

Common points among basketball tv personalities is that the "game slows down" and that defense gets better in the playoffs. We would like to verify these claims with data. To do this we will first classify playoff games and non-playoff games. Non-playoff games are any games in the group phase, group round, preliminary round, or classification games there are 287 of these games. Playoff games are all other games and are games where the losing team loses out on qualification contention for the Olympics from that tournament. There are 191 of these games.

To verify the first claim, we will compare the total number of offensive possessions for both teams in playoff games and non-playoff games. We will calculate the number of offensive possessions for one team using this common formula (https://www.nbastuffer.com/analytics101/possession/):

0.5*((Field Goal Attempts + 0.4*Free Throw Attempts – 1.07*(Offensive Rebounds/(Offensive Rebounds + Opponent Defensive Rebounds))*(Field Goal Attempts – FG) + Turnovers) + (Opponent Field Goal Attempts + 0.4*(Opponent Free Throw Attempts) – 1.07*(Opponent Offensive

Rebounds)/(Opponent Offensive Rebounds + Defensive Rebounds))*(Opponent Field Goal Attempts – Opponent FG) + Opponent Turnovers))

Each team should have roughly the same number of possessions. The number of possessions is often used to describe the pace of play in a game, so we will compare the number of possessions in playoff games and non-playoff games. Next, we will assess the claim that defense is better in the playoffs. To do this we will not look simply at the number of points team scored in playoff and non-playoff games as the score does not account for pace of play. Instead we will look at the points per possession metric. Using our possession calculation from before we will compare whether offense efficiency is worse during playoff games.

Clustering for Player Positions

Methodology

As mentioned above, most of the data we will be clustering on a player's offensive attributes. As many players played multiple games in multiple tournaments there are multiple rows for each player corresponding to his performance in an individual game. We will cluster in two ways. The first will by on a per game basis. This will benefit us in two ways. The first is that we will allow players "different positions" depending on the game. This makes basketball sense. Depending on the opponent, tournament, and performance, a player will sometimes play differently. The second is that it will allow us to validate our clustering method. We hope that a player plays most of his games the same way, and that over all his games he is clustered into 3 different player types at most. After, assigning a cluster to each player's game we will assign a player a position based on their most common cluster.

The drawback to this method is that we don't use all the data we have about the player in the first clustering step. The second method we will use is to cluster on a player's aggregated per game statistics. This will allow us to account for consistency in a player's play when we first cluster the data.

In both methods we will follow the same steps.

1. Extract offensive style descriptors: 2 point shots attempted, 3 point shots attempted, free throws attempted, offensive rebounds, assists, and turnovers.
2. Scale the data. This will make sure that 2 point field goal attempts are not overly powerful in distinguishing clusters.
3. Perform Principle Component Analysis to reduce our features space from 6 variables to 2 variables. This makes sure that the Euclidean distance formula we use for clustering works well. In addition, it makes the plotting of the clusters feasible
4. Use KMeans clustering with 6 clusters. We chose 6 clusters for basketball reasons. There are many players in our dataset who barely play; additionally, there are 5 players on the basketball court. Therefore, our hypothesis is that there should be one cluster for the players who don't play and 5 clusters for each role of player on the basketball court.
5. Summarize the data and interpret the clusters.

Results and Discussion

Individual Game Clusters

For the individual games, principle component analysis captured roughly 70% of the data's variance in 2 components. The clustering did not do what we expected but still gave interpretable and relevant results. Typically in basketball, we have seen 5 positions, the point guard, shooting guard, small forward, power forward, and center. However, the game is changing, and coach Brad Stevens of the

Boston Celtics is quote as saying “it may be as simple as three positions now, where you’re either a ball-handler, a wing or a big” (http://bleacherreport.com/articles/2720250-brad-stevens-says-celtics-have-3-not-5-positions-now). This is what our clustering revealed.

1. Bench Players
    - Bench players are characterized by their lack of production. They are the largest cluster as most teams have approximately 15 players per team but play only 9 of them in a game.
2. High Usage Ball Handlers
    - Ball Handlers typically attempt the most three-point attempts per game, and assist on the most baskets. They do not rebound the ball very often. They have a relatively high number of turnovers due to the high number of times they touch the ball.
3. High Usage Bigs
    - Bigs play close to the hoop and so they shoot the ball often from two-point range. They get fouled the most, so they shoot the most free throws. As they are close to the basket on offense, and bigger they grab the most offensive rebounds. They do not pass as often as they are often the finishers of plays. This leads to less assists. Finally, they are not good ball handlers and still turnover the ball at a high rate.
4. High Usage Wings
    - Wings do a combination of shooting threes, setting up teammates, and attacking the hoop. This leads to moderate volume across all statistics.
5. Low Usage Ball Handlers/Wings
    - These low usage players follow the same trends mentioned above but at a lesser degree.
6. Low Usage Wings/Bigs
    - These low usage players follow the same trends mentioned above but at a lesser degree.

Cluster #	Name	Size (n)	FGA2	FGA3	FTA	OReb	Ast	TO
6	Bench	4021	0.7	0.4	0.26	0.2	0.2	0.3
1	High Usage Ball Handler	834	4.5	6.9	2.4	0.6	4.0	2.3
0	High Usage Big	734	10.4	0.9	5.5	3.2	1.5	2.3
2	High Usage Wing	1230	6.6	3.5	3.7	1.2	2.8	2.4
5	Low Usage, BH/Wing	2414	2.4	3.3	1.1	0.4	1.8	1.1
4	Low Usage Wing/Big	2162	4.7	0.7	2.1	1.4	0.9	1.2

*Note cluster # has no relevant interpretation, it is labeled for the purpose of reporting consistent results

These results were based on the variables we clustered on. We will now look at summaries of features that we did not cluster on and see if the results match what we would expect for the labels of our clusters.

The first feature we will look at is the average minutes played for each cluster. We would expect the Bench cluster to have low minutes played. High Usage players to have high minutes played, and low Usage Players to have moderate minutes played. Note that players can be low usage for 2 reasons. First, they played a lot but did not contribute, or second, they contributed a lot but did not play enough minutes to be considered high usage.

	Bench	High BH	High Big	High Wing	Low BH/Wing	Low Wing/Big
Minutes	6.1	29.0	27.8	27.8	20.0	19.0

These results match our expectations. Next, we will look at defensive statistics such as defensive rebounds, steals and blocks. We expect that bigs have high defensive rebounds and blocks and that ball handlers have high steals. Wings should have moderate defensive rebounds, comparable steals to ball handlers and moderate blocks. Low Usage players should follow the same trends but at smaller quantities.

Name	DReb	Steals	Blocks
Bench	0.6	0.2	0.1
High Usage Ball Handler	3.0	1.1	0.2
High Usage Big	4.8	1.0	0.7
High Usage Wing	3.4	1.1	0.3
Low Usage, BH/Wing	2.0	0.7	0.1
Low Usage Wing/Big	2.6	0.6	0.4

Again, these results give us confidence that our clusters make sense. If theses statistics were more evenly spread across clusters there would be a problem. Since we have diversity of statistics within the clusters and they match our intuitive ideas we have faith in our clusters.

Next, we will look at shooting percentages by cluster. We expect ball handlers to shoot a high percentage from the three-point line and the free-throw line. They may shoot a lower percentage from two-point range. Bigs should shoot a high percentage for two-point field goals and very low percentages from the three-point line. Again, wings should be a moderation of the ball handlers and bigs. Our low usage players should be closely aligned with their high usage groups. However, we expect their percentages to be a little worse as if they were just as good we believe they would play more often.

Name	2FG%	3FG%	FT%
Bench	41.1	23.1	63.2
High Usage Ball Handler	48.8	37.0	78.8
High Usage Big	52.6	26.2	67.0
High Usage Wing	50.0	32.4	73.8
Low Usage, BH/Wing	47.9	33.0	72.5
Low Usage Wing/Big	50.1	26.9	66.5

Again, these results match what we would expect from the labels of the clusters. Finally, we will look at 2 advanced stats, Usage Percentage and Assist Percentage (https://www.basketball-reference.com/about/glossary.html). Usage percentage estimates the percentage of team plays used by a player while on the floor, and assist percentage estimates the percentage of teammate field goals a player assisted on while he was on the floor. They are both smarter ways of analysing the usual counting statistics.

	Bench	High BH	High Big	High Wing	Low BH/Wing	Low Wing/Big
USG Percentage	9.4	23.7	26.3	23.6	16.2	17.2
AST Percentage	1.8	11.4	4.7	8.3	5.9	3.2

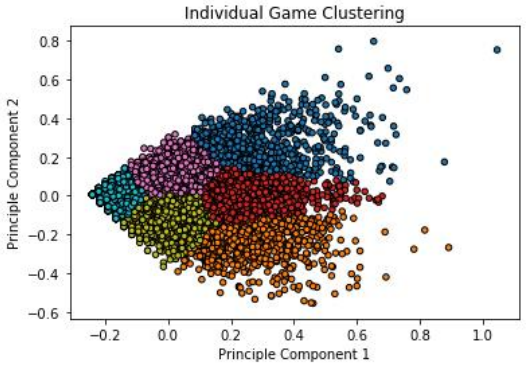
Each element of a cluster was a game that a player played. To get a player's overall position, we took their most common position across all games (ties were broken by the following priority list [HU BH, HU B, HU W, LU W/B, LU BH/W, Bench]). After assigning each player a position we had: 76 Bigs, 89 Ball Handlers, 134 Wings, 227 low usage bigs. 260 low usage ball handlers, and 548 bench players. On averages players played as their classified position in two thirds of their games. Almost every player played a game where they were classified as a different position than their final classification. We will investigate these classifications to make sure player classifications make sense. High Usage Bigs often had games as low usage bigs and sometimes as high usage wings. They rarely had games as High Usage

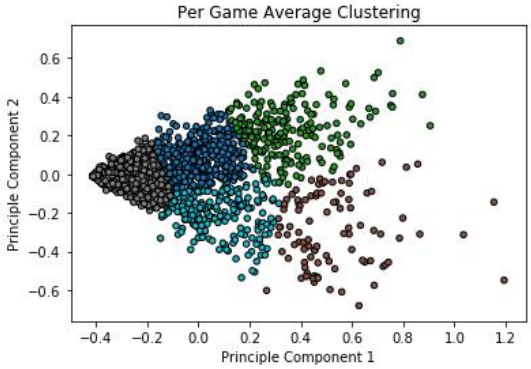
Ball Handlers or Low Usage Ball Handlers. Similarly, High Usage Ball Handlers had games as Low Usage Ball Handlers and High Usage Wings. Lastly, Bench players most often had games as Low Usage players and rarely had games as High Usage players. This gives us confidence that our clusters do in fact make sense.

Our last validation method was to scan the names of players and their assigned position to see if it made sense. Kevin Durant was a High Usage Wing, Pau Gasol a High Usage Big, and J.J. Barea a High Usage Ball handler, these are all good classifications. One interesting assignment was that of Mirza Teletovic. Teletovic is a NBA player often classified as a Power Forward (Big or Wing in our positions). He was classified as a Ball Handler in our analysis. This is because of his style. He is a bigger player who shoots threes and makes good passes; therefore, despite disagreeing with conventional knowledge we believe his classification makes sense.

Per Game Averages Clusters

We will only present surface level results of the per game average clusters. We found that after averaging a player's performance over multiple games it made more sense to use 5 clusters. The principle component analysis accounts for 85% of the variation. We are left with clusters for Bench players, Low Usage Bigs, Low Usage Ball Handlers, High Usage Bigs and High Usage Ball Handlers. Their characteristics align with those of our previous clusters and will not be presented. What is interesting is that the Wing position does not appear in the clustering when we used 6 clusters (got 2 bench type clusters). We theorize that this is because it is harder to sustain constant production in all categories across multiple games. These wing players are still there on the boundary of the High Usage Bigs and Ball Handlers cluster. However, there are not enough of them to create their own cluster.

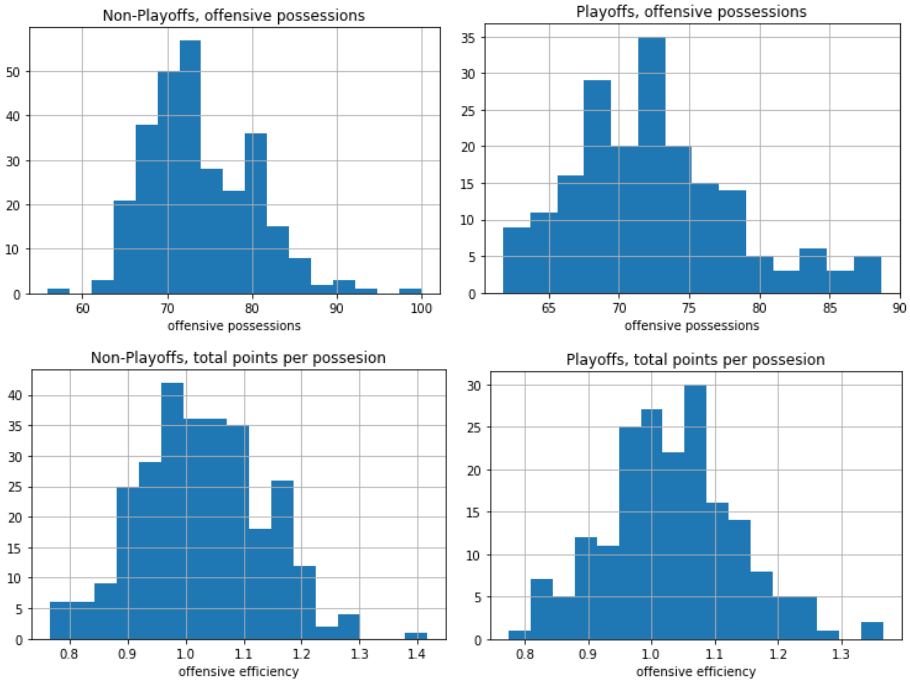




The game slows down in the playoffs. Or does it?

Methodology

First, we classified games as playoff or non-playoff, calculated the estimated possessions per game for each game and the offensive efficiency for each game. With these variables we plotted histograms to make sure the data looked normal. We had enough observations to assume the data was normal enough for our t-test purposes. Then we tested each hypothesis with a t-test.



Results and Discussion

Offensive Possessions Per Game

The results of the t-test measuring the pace of the playoff games vs non-playoff games were significant and indicated the rejection of the null hypothesis that the pace was the same. However, in basketball sense there is not much different between an average of 72 possessions to 74 possessions. The game is slightly slower on average. From the histograms we can see that non-playoff games have more extremes for number of offensive possessions.

Offensive Efficiency

We fail to reject the null hypothesis of the t-test that offensive efficiency is different in playoff games vs non-playoff games. This is an interesting contradiction against classical basketball knowledge. This may be because as we saw above playoff games are slower on average. This leads to less chances for teams to score. In this next test we see that efficiencies are indistinguishable. Therefore, the total scores in playoffs games will be lower than non-playoff games. This may be what leads to the conclusion that defense is better in the playoffs; however, we disagree with that conclusion based on our results.

Limitations

For the clustering project the best validation metric for our clusters would be to check the positions against a players commonly assigned position. This data is available but would have to be scraped from the web. This process would involve distinguishing players with the same name and those who have different assigned positions depending on the team and website. Other issues are that we don't get great separation in our clusters after performing principle component analysis. While the results make great basketball sense we look at a mostly means. This means that the edges of clusters are almost indistinguishable. It does however, give us the notion that the first component gives the notion of frequency while the second dictates the area of the court a player occupies on offense. This is a nice thought but one that we did not explore further. We could improve clusters by adding attributes that define players such as shot location, passes per game or dribbles per game.

For the second problem the biggest limitation is the lack of normality. However, due to the large sample size we can still perform our analysis. The other limitation would be that our significant result indicates a difference of means of 2 possessions. I do not think a coach of basketball team would care if he learned that the number of possessions is only 2 less on average in playoff games.

Project Experience Summaries

Wilson Adhikari

Obtained statistics for number of offensive possessions for each team, derived from an equation model from an unofficial NBA stats website using various statistics implemented in pandas DataFrames. Applied one-sided t-test, comparing independent pace data of playoff vs non-playoff games to determine if the expected pace for playoff games is lower than the expected pace for non-playoff games. Evaluated for total points per possession for each game to further distinguish playoff vs non-playoff offenses. Provided analysis on findings using prior basketball knowledge. Communicated effectively with group, creating a Discord group to centralize discussion. Organized workflow through use of Feature Branch Workflow, by creating branches for new problems followed by a merge after completion, to better optimize development process and future history readability.

Dani Chu

Clustered, using K-Means clustering after performing PCA for dimension reduction, basketball players who played in FIBA tournaments to qualify for the 2016 Men's Olympic Basketball Tournament based on their offensive impacts in games. These clusters were created to better understand positions in basketball, and because the data given was missing player position data. Analysed, and discussed the resulting summaries of the clusters using my basketball knowledge. Demonstrated communication and interpretation skills by detailing our groups work in a report explaining the background of the problem, the data, problems, methodology and results. Designed, the analysis of playoff and non-playoff games to better understand common basketball sayings.

Viresh Soedhwa

Implemented commonly used basketballs formulas to create advanced attributes for each player. These attributes are used to feed the PCA analysis in an attempt to get a better classification of player styles. Part of the challenge was understanding basketball terms and how these relate to the functions that are used.

Merging, joining and using lambda functions to shape the data in preparation for the function parameters. The basic parameters are observations while the advanced parameters are a combinations of several observations including attributes from the opposing team that affect the player. Adding the advanced parameters did not affect the shape of the classification by much. The players are already clearly classified from the available data. An improvement here would be including attributes that could describe player behaviour more clearly such as shot location, passes per game or dribbles per game.