NBA Players Segment Analysis

August 1, 2022

A Comparison of Unsupervised Machine Learning Algorithms, KMeans Clustering and Gaussian Mixture Model

Clustering Analysis of NBA Players Statistics from the 2021-22 Season

by John K. Hancock jkhancock@gmail.com

1 Introduction

Clustering is the use of unsupervised machine learning algorithms to identify how different data points are related to one another. The practical use of cluserting is to identify those similar characteristics that define segments. Marketers use segments to differentiate and target consumers who are most likely to buy their particular products or services.

For a professional sports league, clustering players based on their performance statistics enable the players, General Managers, and team owners to segment players along the dimensions of their overall value to the team. Players in one cluster/segment may provide more scoring value than players in another segment.

In this project, my goal is to solve a problem for a brand new owner of a NBA team. The owner wants to have a segmented ordering of players in the NBA in order to have some guidance when negotiating contracts or trades.

To solve this problem, I used two unsupervised clustering algorithms, KMeans and Gaussian Mixture Model to create player segments based on their 2021-22 performances.

2 Problem Statement

James Duarte is the new owner of an NBA team, the San Antonio Spurs. He is getting bombarded with trade requests and several of his players are up for new contracts. In the NBA, teams are limited by how much they can spend on players. If owners go over that limit, they have to pay a luxury tax to the other teams in the league.

Given these circumstances, Duarte needs more information about all of the players in the NBA. He wants to differentiate players based on their performance. Doing so, gives him insights on how to build his roster, renew contracts, and make trades with other teams. Most importantly, he wants to avoid the NBA's luxury tax. He doesn't want to over spend on new contracts.

3 Project Plan

To solve James Duarte's problem, this project clusters NBA players using two unsupervised machine learning algorithms, KMeans Clustering and Gaussian Mixture Models.

The following is an outline for how I devised a solution:

I. I identified and collected NBA player statistics from the 2021-22 NBA season from the website, www.nbastuffer.com. More information can be found in the "About the Data" section.

II. I imported and cleaned the data.

III. I explored the data by looking at some of the distributions and the correlations among and between the features.

IV. I then processed and modeled the data

V. I summarized the findings.

VI. I provided a conclusion.

4 I. About the Data

For this project, I used data from the NBA 2021-22 regular season. I collected the player statistics from the website, www.nbastuffer.com.

From the site:

NBAstuffer, started out as a hobby-site by Serhat Ugur in 2007, has grown into a reputable stats-reference that delivers unique metrics and NBA analytics content some of which can't be found anywhere else.

The site also maintains data for the post-season as well. In keeping within the scope of this report, I used data from the regular season.

5 II. Data Import and Cleaning

Reading the data from NBAStuffer into a pandas data file shows that there are 26 features over 716 observations. Each observation represents a player in the NBA.

After an inspection, the following fixes will be made to the data:

1. Preserve a copy of the downloaded data and re-import the local copy just in case the link to the original data is lost.
2. Remove the "RANK" column which is nothing more than another index.
3. The column names contain information about each feature. These names will be preserved into a dictionary.
4. The column names will then be re-named with shorter names.

[ ]:
```
(716, 28)
```

Some of the column names contain additional information about each statistic.

[ ]:
```
Empty DataFrame
Columns: [FULL NAME, TEAM, POS, AGE, GP, MPG, MIN%Minutes PercentagePercentage
of team minutes used by a player while he was on the floor, USG%Usage RateUsage
rate, a.k.a., usage percentage is an estimate of the percentage of team plays
used by a player while he was on the floor]
Index: []
```

[ ]:
```
Index(['FULL NAME', 'TEAM', 'POS', 'AGE', 'GP', 'MPG', 'Minutes_pct',
       'Usage_pct', 'Turnover_Rate', 'FTA', 'FT%', '2PA', '2P%', '3PA', '3P%',
       'Effective_Shooting_pct', 'True_Shooting_pct', 'Points_per_game',
       'Rebounds_per_game', 'Total_Rebound_pct', 'Assists_per_game',
```

```
 'Assist_pct', 'Steals_per_game', 'Blocks_per_game',
 'Turnovers_per_game', 'Versatility Index',
 'Offensive_Rating_Individual', 'Defensive Rating'],
dtype='object')
```

Cleaning Duplicate Records Data is duplicated for players that were traded during the season. Below there are two entries for Seth Curry, James Harden, and Kristaps Porzingas. Statistics are split for each team that the player played for.

These records will have to grouped and aggregated accordingly so that each player is represented once in the dataset.

[ ]:

```
             FULL NAME TEAM
140         Seth Curry  Phi
141         Seth Curry  Bro
251       James Harden  Bro
252       James Harden  Phi
533  Kristaps Porzingis  Dal
534  Kristaps Porzingis  Was
```

[ ]:

```
          FULL NAME POS    AGE  GP   MPG  Minutes_pct  Usage_pct  Turnover_Rate
284  Aaron Holiday   G  25.53  41  16.2         33.7       18.0           14.5
285  Aaron Holiday   G  25.53  22  16.3         33.9       19.9           16.3
330  Alize Johnson   F  25.97   4   7.1         14.8       19.2            7.8
329  Alize Johnson   F  25.97   3   6.1         12.6       23.1           31.8
328  Alize Johnson   F  25.97  16   7.5         15.7       12.5           23.5
```

Group the duplicate records by Full Name and aggregate their statistical measures accordingly. For example, to get the traded players full number of games played, you have to sum their games played across two teams.

[ ]:

```
                       FULL NAME POS    AGE  GP    MPG  Minutes_pct  Usage_pct  \
FULL NAME
Aaron Holiday      Aaron Holiday   G  25.53  63  16.25    33.800000  18.950000
Alize Johnson      Alize Johnson   F  25.97  23   6.90    14.366667  18.266667
Andre Drummond    Andre Drummond   C  28.67  73  20.35    42.400000  19.550000
Armoni Brooks      Armoni Brooks   G  23.85  54  14.30    29.850000  14.650000
Brad Wanamaker    Brad Wanamaker   G  32.71  23  20.20    42.100000  12.850000

                Turnover_Rate
FULL NAME
Aaron Holiday       15.400000
Alize Johnson       21.033333
Andre Drummond      17.950000
Armoni Brooks        8.700000
Brad Wanamaker      11.000000
```

```
[ ]: (507, 28)
```

```
[ ]:           FULL NAME TEAM POS    AGE  GP   MPG  Minutes_pct  Usage_pct
     0  Precious Achiuwa  Tor   F  22.56  73  23.6         49.2       18.5
     1      Steven Adams  Mem   C  28.73  76  26.3         54.8       12.0
```

```
[ ]: (98, 27)
```

```
[ ]:           FULL NAME POS    AGE  GP   MPG  Minutes_pct  Usage_pct
     0  Precious Achiuwa   F  22.56  73  23.6         49.2       18.5
     1      Steven Adams   C  28.73  76  26.3         54.8       12.0
```

```
[ ]: (605, 27)
```

There are now 605 observations over 27 features. Below is a check on how the duplicates were aggregated into one record each.

```
[ ]:               FULL NAME  POS    AGE  GP    MPG  Minutes_pct  Usage_pct  \
     546       James Harden    G  32.63  65  37.35        77.85       26.6
     564  Kristaps Porzingis  F-C  26.69  51  28.85        60.05       30.0
     589         Seth Curry    G  31.63  64  32.35        67.40       18.5

          Turnover_Rate
     546           18.1
     564            8.3
     589           11.6
```

A check of the fields shows that Offensive_Rating_Individual and Defensive Rating have 34 observations that are missing. These are players that play very limited minutes. The average Minutes per Game ("MPG") is only 2.54 minutes.To address the missing data for this feature, I set the missing values to 0.

```
[ ]: FULL NAME                       0
     POS                             0
     AGE                             0
     GP                              0
     MPG                             0
     Minutes_pct                     0
     Usage_pct                       0
     Turnover_Rate                   8
     FTA                             0
     FT%                             0
     2PA                             0
     2P%                             0
     3PA                             0
     3P%                             0
     Effective_Shooting_pct          9
     True_Shooting_pct               8
```

```
Points_per_game                0
Rebounds_per_game              0
Total_Rebound_pct              0
Assists_per_game               0
Assist_pct                     0
Steals_per_game                0
Blocks_per_game                0
Turnovers_per_game             0
Versatility Index              0
Offensive_Rating_Individual   34
Defensive Rating              34
dtype: int64
```

[ ]:
```
2.54
```

[ ]:
```
           FULL NAME  GP  MPG  Offensive_Rating_Individual  Defensive Rating
17        Joel Ayayi   7  2.9                          NaN               NaN
31        Paris Bass   2  3.7                          NaN               NaN
38       Jordan Bell   1  2.0                          NaN               NaN
85       Ahmad Caver   1  0.9                          NaN               NaN
102   Sharife Cooper  13  3.0                          NaN               NaN
103    Petr Cornelie  13  2.9                          NaN               NaN
114       Sam Dekker   1  0.9                          NaN               NaN
115  Javin DeLaurier   1  2.8                          NaN               NaN
132  Jaime Echenique   1  3.1                          NaN               NaN
166   Jordan Goodwin   2  3.0                          NaN               NaN
```

[ ]:
```
FULL NAME                      0
POS                            0
AGE                            0
GP                             0
MPG                            0
Minutes_pct                    0
Usage_pct                      0
Turnover_Rate                  0
FTA                            0
FT%                            0
2PA                            0
2P%                            0
3PA                            0
3P%                            0
Effective_Shooting_pct         0
True_Shooting_pct              0
Points_per_game                0
Rebounds_per_game              0
Total_Rebound_pct              0
Assists_per_game               0
```

```
Assist_pct                    0
Steals_per_game               0
Blocks_per_game               0
Turnovers_per_game            0
Versatility Index             0
Offensive_Rating_Individual   0
Defensive Rating              0
dtype: int64
```

```
[ ]: Index(['FULL NAME', 'POS', 'AGE', 'GP', 'MPG', 'Minutes_pct', 'Usage_pct',
            'Turnover_Rate', 'FTA', 'FT%', '2PA', '2P%', '3PA', '3P%',
            'Effective_Shooting_pct', 'True_Shooting_pct', 'Points_per_game',
            'Rebounds_per_game', 'Total_Rebound_pct', 'Assists_per_game',
            'Assist_pct', 'Steals_per_game', 'Blocks_per_game',
            'Turnovers_per_game', 'Versatility Index',
            'Offensive_Rating_Individual', 'Defensive Rating'],
           dtype='object')
```

6 III. Data Exploration

Exploration of Key Features Age

Well over 25% of NBA players are between 22-25 years old.

Games Played

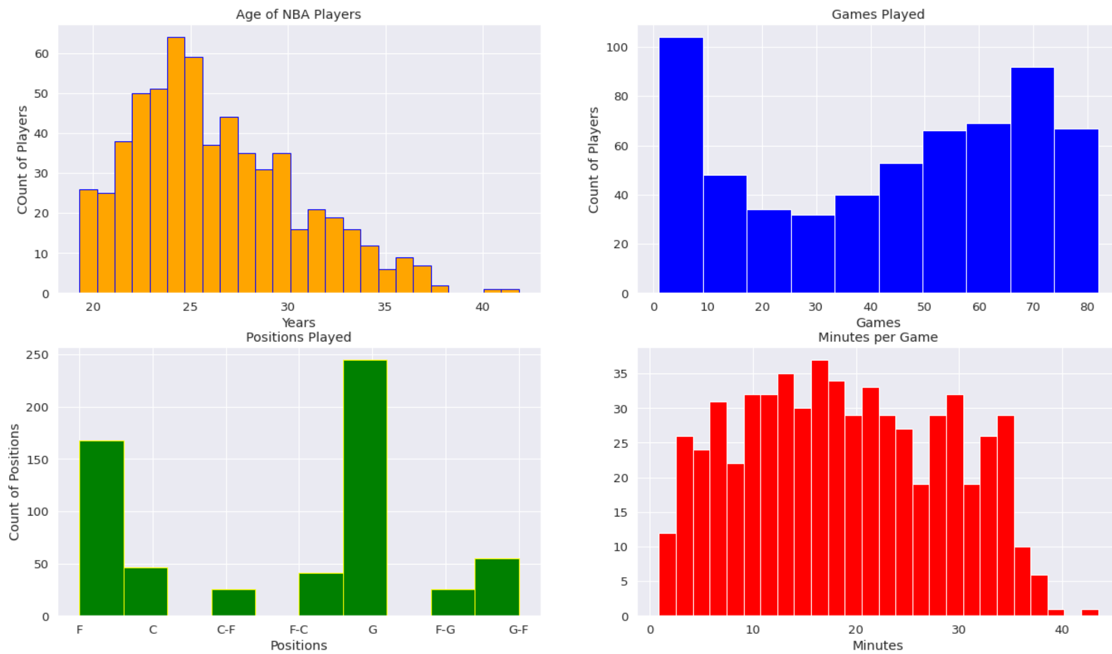
The distribution of the number of games played is bi-modal with the largest number of players playing 10 games or less followed by players plaing 68 to 75 games. The average number of Games Played was 43 and the median was 48.

Positions

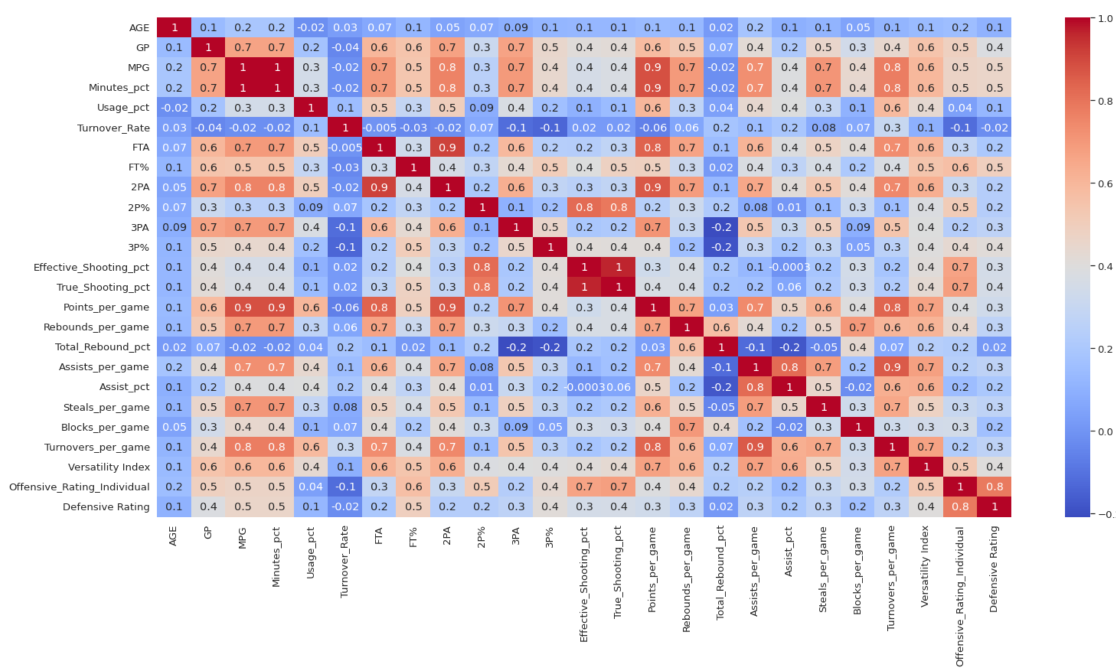
Guards account for the most positions played followed by Forwards and Centers. Other players play a combination of positions.

Minuters per Game

The average number of minutes per game for all players is approx. 19 minutes. The standard deviation is 10 minutes per game.



Multicollinearity In the HeatMap below, you can see that there are a few features that are highly correlated. This will be handled using Principal Component Analysis (PCA) in the Data PreProcessing section.



Explanation of Some Key Features **'Minutes__pct'** is the percentage of team minutes used by a player while he was on the floor.

'Usage__pct' is the usage rate, a.k.a., usage percentage is an estimate of the percentage of team plays used by a player while he was on the floor

'Turnover__Rate' a metric that estimates the number of turnovers a player commits per 100 possessions

'Effective__Shooting__pct' Effective Shooting Percentage With eFG%, three-point shots made are worth 50% more than two-point shots made. eFG% Formula=(FGM+ (0.5 x 3PM))/FGA

'True__Shooting__pct' True Shooting PercentageTrue shooting percentage is a measure of shooting efficiency that takes into account field goals, 3-point field goals, and free throws.

'Total__Rebound__pct' Total Rebound PercentageTotal rebound percentage is estimated percentage of available rebounds grabbed by the player while the player is on the court.

'Assist__pct' Assist PercentageAssist percentage is an estimated percentage of teammate field goals a player assisted while the player is on the court

'Versatility Index' Versatility index is a metric that measures a player's ability to produce in points, assists, and rebounds. The average player will score around a five on the index, while top players score above 10

'Offensive__Rating__Individual' Offensive RatingIndividual offensive rating is the number of points produced by a player per 100 total individual possessions

'Defensive Rating' Defensive rating estimates how many points the player allowed per 100 possessions he individually faced while staying on the court.

7 IV. Data Processing and Modeling

For this section, I took the following steps:

1. Drop the Full Name and POS from the dataset
2. Standardized the features
3. Conduct PCA Analysis of the features
4. Use KMeans Clustering to segment the data
5. Concatenate the datasets into one dataframe
6. Analyze the KMeans Segmentation results
7. Use Gaussian Mixture Model to segment the data
8. Analyze the GMM Segmentation results

1. Drop the FULL NAME feautre

2. Scale the Features using sklearn Standard Scaler Scaling features is essential for machine learning algorithms. Many algorithms assume that all of the features are centered around zero. Features with large variances may inhibit the algorithm's ability to learn.

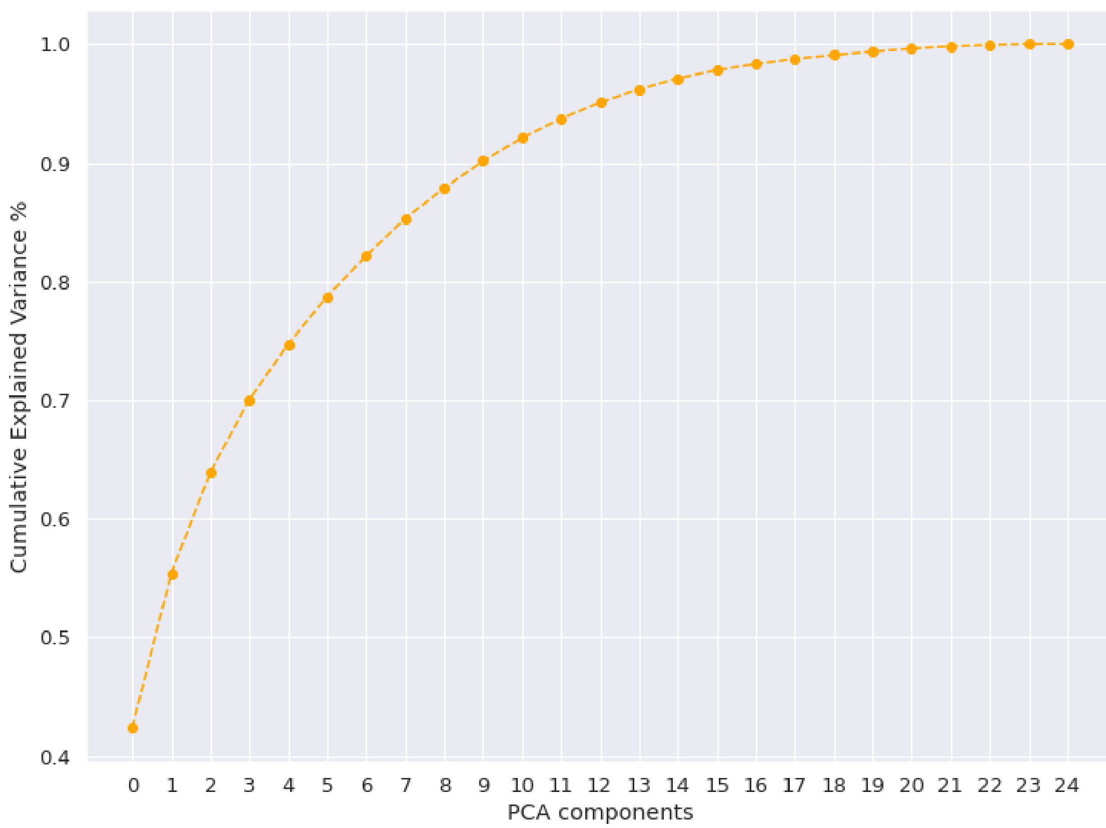
3. Reduce the Dimensionality of the Features using PCA As discussed earlier, several features were correlated with each other. Thus, "Principal Component Analysis ("PCA") is used below to reduce the overall number of features. PCA projects each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. The first principal component can equivalently be defined as a direction that maximizes the variance of the projected data." Wikipedia: Principal component analysis

A good resource for KMeans clustering and PCA Analysis is an article by Chris Ding and Xiaofeng He.

Chris Ding and Xiaofeng He, *K-means Clustering via Principal Component Analysis*, K-means Clustering via Principal Component Analysis, Proc. of Int'l Conf. Machine Learning, (2004)

```
[ ]: array([4.23925803e-01, 1.29960208e-01, 8.50489016e-02, 6.12148626e-02,
            4.69838399e-02, 4.01799430e-02, 3.40532636e-02, 3.15050856e-02,
            2.56038368e-02, 2.29202161e-02, 1.94738429e-02, 1.61496619e-02,
            1.37301560e-02, 1.10248568e-02, 9.30962164e-03, 6.93846463e-03,
            5.27303488e-03, 4.22443574e-03, 3.16330956e-03, 3.07934189e-03,
            2.37269522e-03, 2.01029237e-03, 1.06604018e-03, 7.88103373e-04,
            1.82276673e-07])
```

Using 80% of the explained variance, there are 5 pca components to use for the model,



- Build the final model using n_components = 5
- Fit the model to the scaled data
- Get the scores for each component by transforming the data

```
[ ]: PCA(n_components=5)
```

The data has been reduced from a 25 features set down to 5 principal components.

```
[ ]: (605, 5)
```

4. Unsupervised Learning Algorithm, KMeans Clustering The first unsupervised learning algorithm I used for this project was KMeans clustering. The term ”unsupervised learning algorithm” simply means that the algorithm looks for patterns in the data using features without labelled outcomes. The algorithm works by grouping data together into a fixed number of clusters. This number of clusters is the ”K” in KMeans clustering.

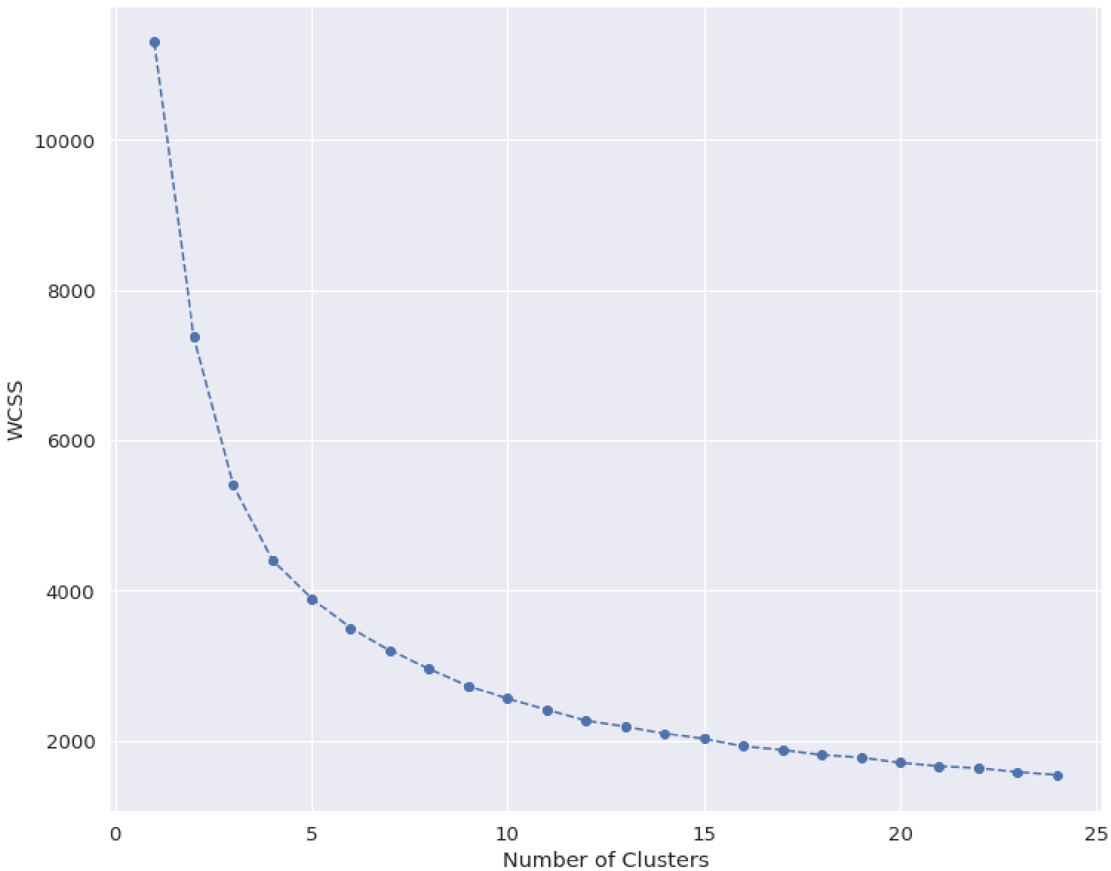
Data points are assigned to clusters based on their distance from the centroid of the cluster. It then calculates the means of each cluster. It iterates this process by taking the variation of the cluster.

”The quality of the cluster assignments is determined by computing the sum of the squared error (SSE) after the centroids converge, or match the previous iteration’s assignment. The SSE is defined as the sum of the squared Euclidean distances of each point to its closest centroid. Since this is a measure of error, the objective of k-means is to try to minimize this value.”

Kevin Arvai, *K-Means Clustering in Python: A Practical Guide*, RealPython.com(2021)

The number of clusters is determined by plotting each cluster against the within cluster sum of squares (WCSS) sum of the squared deviations from each observation and the cluster centroid.

When the graph levels out after steep declines shows us the number of clusters.



As shown above, the number of clusters is 6 which is the number of clusters used to create and fit the model to the reduced PCA component dataset.

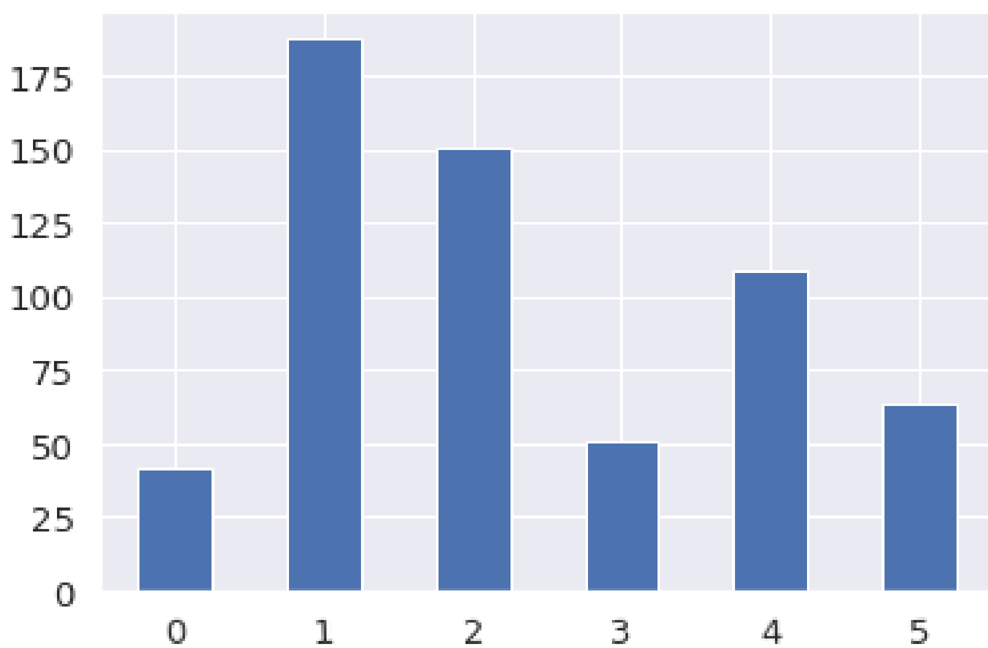
[ ]:
```
KMeans(n_clusters=6, random_state=42)
```

5. Concatenate the datasets into one dataframe

[ ]:
```
          FULL NAME  POS    AGE  GP   MPG  Minutes_pct  Usage_pct  \
0  Precious Achiuwa    F  22.56  73  23.6         49.2       18.5
1      Steven Adams    C  28.73  76  26.3         54.8       12.0
2       Bam Adebayo  C-F  24.73  56  32.6         67.9       25.0

   Turnover_Rate    FTA    FT%  …  Turnovers_per_game  Versatility Index  \
0           11.3  131.0  0.595  …                1.15                6.8
1           19.6  199.0  0.543  …                1.51                9.4
2           14.4  340.0  0.753  …                2.64               10.7

   Offensive_Rating_Individual  Defensive Rating  PCA Component 1  \
0                        105.4             104.0         1.207780
1                        124.7             103.9         2.424570
```

```
2                          117.2              98.2         5.021161

   PCA Component 2  PCA Component 3  PCA Component 4  PCA Component 5  \
0        -0.318339         1.135486        -1.523582         0.156698
1        -0.640418         3.551679         0.096325        -1.835523
2         0.868942         3.833111        -0.094303        -0.009745

   Segment KMeans PCA
0                   5
1                   5
2                   3

[3 rows x 33 columns]
```

6. Analyze the KMeans Segmentation results The KMeans segmentation algorithm shows six clusters of NBA players with segments 1, 2, and 4 with the most number of players.



Segments 0,1, and 2

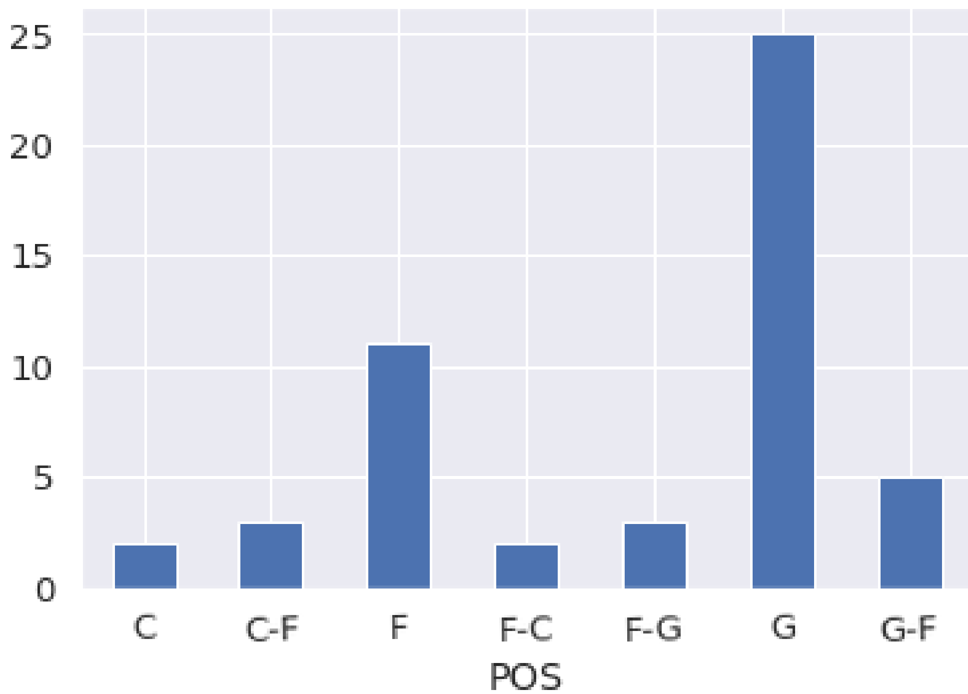
Players in segments, 0,1,and 2, average 13.24 minutes per game and play about 32 games during the season. These are very low impact players. Segment 1 players did play in more games than players in Segments 0 or 2, but they played in substantially less games than players in Segments 3,4, and 5. The Offensive and Defensive ratings for Segment 1 players are comparable to Segments 3,4, and 5, but given their low playing time, we cannot put this segment on their level.

```
[ ]: MPG      13.24
     GP       31.75
     dtype: float64
```

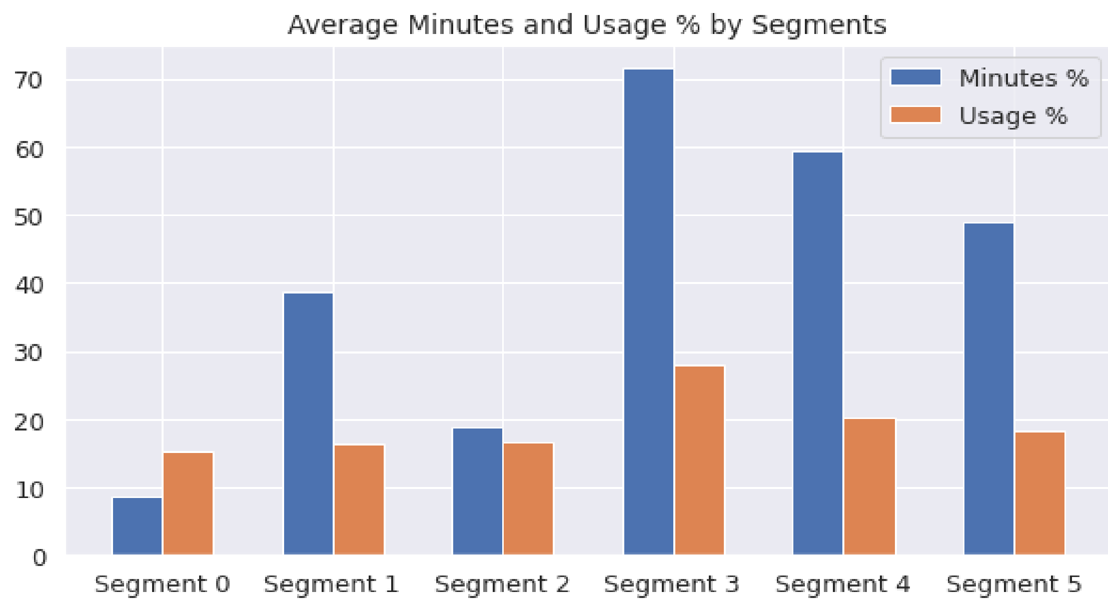
```
[ ]:                  FULL NAME  GP   MPG  Segment KMeans PCA
     3             Santi Aldama  32  11.2                   2
     5            Grayson Allen  66  27.3                   1
     7            Jose Alvarado  54  15.4                   1
     10  Thanasis Antetokounmpo  48   9.9                   2
     14        Ryan Arcidiacono  10   7.6                   2
```

Segment 3

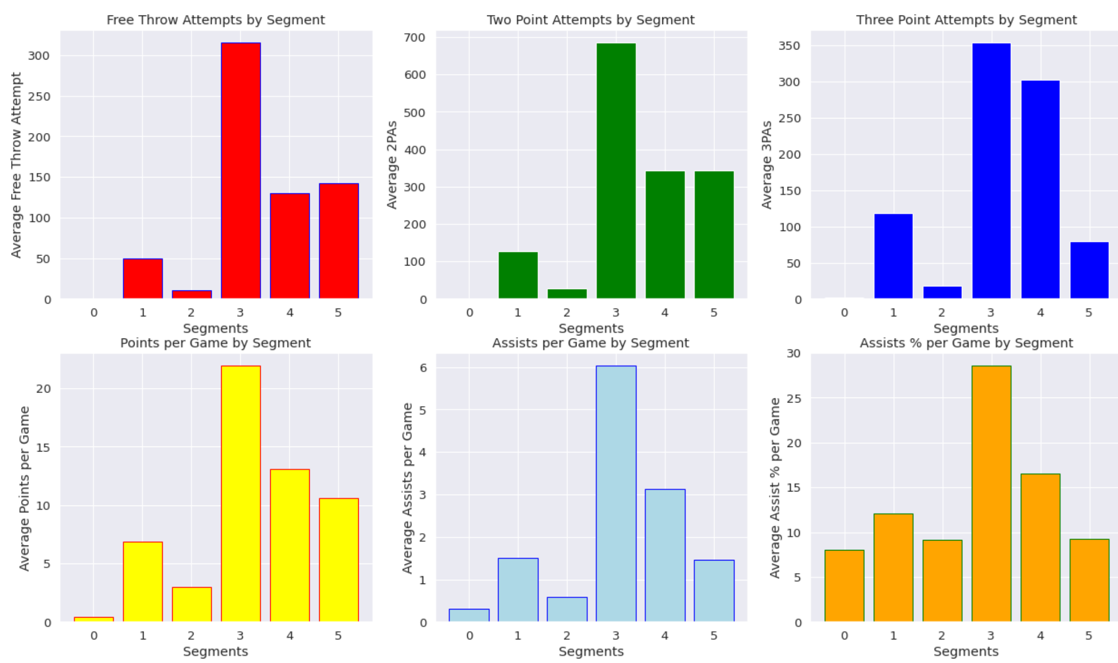
Segment 3 has 51 players with 44 of those players being either Guards, Forwards, or a combination of the two.



This segment also had the highest Minutes_pct and Usage_pct, which are the percentage of team minutes and the number of plays used by a player while he was on the floor.



Segment 3 players also lead in Free Throw Attempts, Two- and Three-Point Attempts, Points Per Game, Assists per game, and Assist percentage.



Interestingly enough, Segment 3 players have a lower True Shooting percentage than do Segment 5 players. True Shooting percentage is a ratio of a player's total points against 2 * their field goals attempted plus 44% of their free throws attempted.

$1/2 * PointsScored/(FGA + .44(FTA))$

This stat rewards players for the additional point of three pointers as well as shooting well from the free throw line. If a player checks into a game and hits a three pointer and then leaves the game, his True Shooting percentage would be 150, a maximum value.

[ ]:

	True_Shooting_pct
Segment KMeans PCA	
3	0.571833
5	0.612578

This segment has the high impact players. In particular, their impact comes from their presence on the court and their contributions offensively.This segment has some of the most notable players in the game, e.g. current and former Most Valuable Player award recipients, LeBron James, Stephen Curry, Giannis Antetokounmpo, and Nikola Jokic.

[ ]:

```
2                Bam Adebayo
9      Giannis Antetokounmpo
12              Cole Anthony
20               LaMelo Ball
28                RJ Barrett
36              Bradley Beal
52              Devin Booker
59             Miles Bridges
67              Jaylen Brown
76              Jimmy Butler
107          Cade Cunningham
108            Stephen Curry
116            DeMar DeRozan
121              Luka Doncic
131             Kevin Durant
133          Anthony Edwards
139              Joel Embiid
147             De'Aaron Fox
153            Darius Garland
158              Paul George
Name: FULL NAME, dtype: object
```

Segment 4

Segment 4 is the third largest segment of NBA players, but unlike segments 1 and 2, which are larger, this segment does get a lot of playing time and usage. This segment ranks as the highest in Defensive Rating. For other measures, this segment has the second highest games played, minutes, and usage rate. They have the second highest points and assists per game as well.

60% of the players in this segment are guards.

[ ]:

	GP	MPG	Usage_pct
Segment KMeans PCA			
4	62.899083	28.466055	20.350459

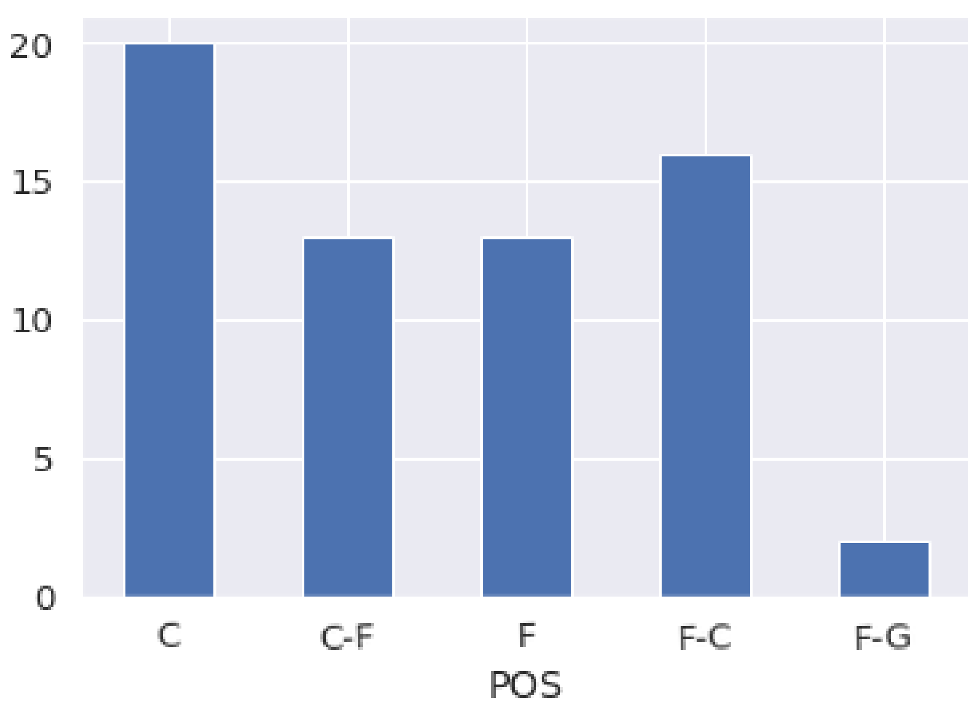
```
[ ]: G      66
     F      27
     G-F    10
     F-G     4
     F-C     2
     Name: POS, dtype: int64
```

```
[ ]: 489     Brandon Williams
     515          Buddy Hield
     516         CJ McCollum
     518         Caris LeVert
     531      Dennis Schroder
     533        Derrick White
     551            Josh Hart
     579        Norman Powell
     589           Seth Curry
     590    Spencer Dinwiddie
     Name: FULL NAME, dtype: object
```

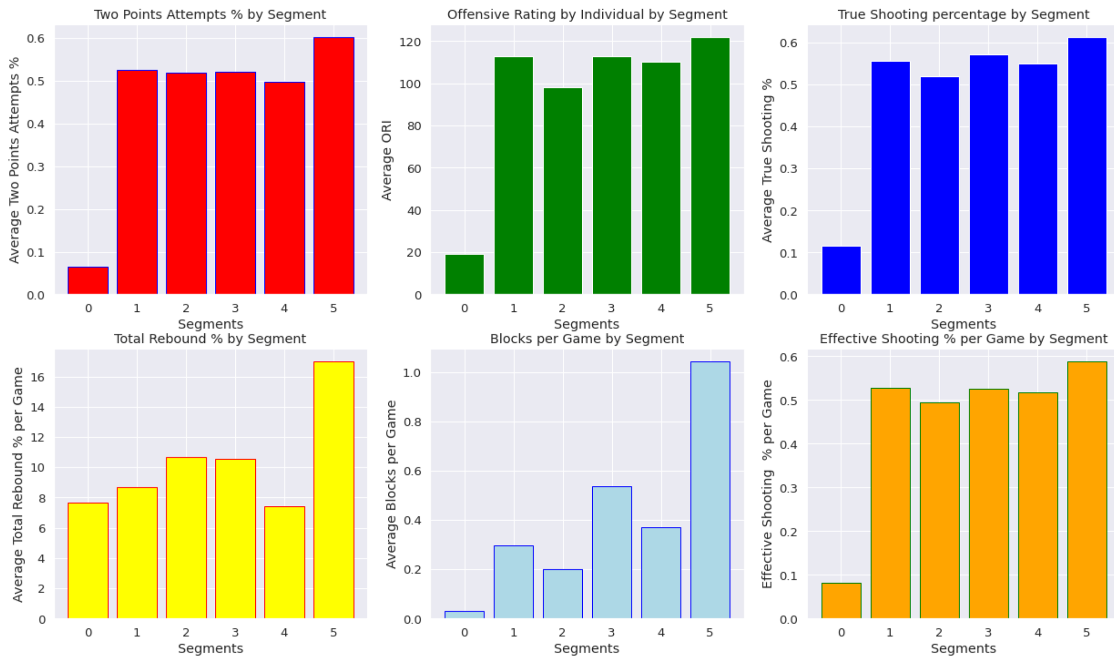
This segment has medium impact players. The only statistical category that they lead in is Defensive Rating, and even then, it's marginal. As stated earlier, they do play regularly and get the second most minutes, but they are largely complementary players.

Segment 5

Segment 5 has the second smallest number of players at 64 total players. It's comprised almost entirely of centers and forwards or a combination of the two.



Players in this segment lead in two point scoring percentage, true shooting percentage, total rebounds per game percentage, offensive rating, and blocks.




```
[ ]: 0          Precious Achiuwa
     1              Steven Adams
     4         LaMarcus Aldridge
     6             Jarrett Allen
     18            Deandre Ayton
     22                 Mo Bamba
     35            Darius Bazley
     43          Bismack Biyombo
     54            Chris Boucher
     82             Clint Capela
     83       Wendell Carter Jr.
     91           Brandon Clarke
     94              Nic Claxton
     96             John Collins
     97             Zach Collins
     109           Anthony Davis
     113          Dewayne Dedmon
     151          Daniel Gafford
     164             Rudy Gobert
     192      Isaiah Hartenstein
     Name: FULL NAME, dtype: object
```

Segment 5 contains the second highest impact players. They are typically the biggest players on the court given the number of Centers and Forwards in the segment. They play close to the basket which is why they lead in blocks per game.

These are players whose position calls form to play close to the hoop. They score off of rebounds or passes close to the basket.

Segment Analysis

The KMeans algorithm identified 6 segments. After analyzing the segments, I can rename them accordingly:

Segment 3 - "Elite Player" Segment 5 - "High Impact Player" Segment 4 - "Moderate Impact Player" Segment 1 - "Low Impact Player" Segment 2 - "Extremely Low Impact Player" Segment 0 - "Marginal Player"

```
[ ]: Segment KMeans PCA                Elite  Extremely Low Impact  \
AGE                             27.188431             25.151126
GP                              63.313725             19.198675
MPG                             34.281373              9.099779
Minutes_pct                     71.422549             18.949779
Usage_pct                       27.885294             16.551987
Turnover_Rate                   13.861765             13.006954
FTA                            315.215686             10.777594
FT%                              0.809902              0.502072
2PA                            685.176471             27.596578
2P%                              0.521206              0.518639
3PA                            353.764706             17.664459
3P%                              0.339039              0.206817
Effective_Shooting_pct           0.526314              0.495092
True_Shooting_pct                0.571833              0.520253
Points_per_game                 21.946078              2.949172
Rebounds_per_game                6.609804              1.771965
Total_Rebound_pct               10.528431             10.650497
Assists_per_game                 6.038235              0.582726
Assist_pct                      28.595098              9.103918
Steals_per_game                  1.166961              0.279823
Blocks_per_game                  0.536961              0.200193
Turnovers_per_game               2.977647              0.431187
Versatility Index               10.784314              4.964183
Offensive_Rating_Individual    112.661765             98.168322
Defensive Rating               107.148039             96.446799

Segment KMeans PCA           High Impact Player  Low Impact   Marginal  \
AGE                                   26.451250   27.117500  24.418810
GP                                    60.312500   48.037234   3.952381
MPG                                   23.435156   18.597872   4.129762
Minutes_pct                           48.806250   38.746543   8.594048
Usage_pct                             18.171094   16.386348  15.145238
```

Turnover_Rate	12.753125	11.041223	11.207143
FTA	141.945312	49.664007	0.833333
FT%	0.691992	0.773758	0.130405
2PA	342.195312	126.191489	2.119048
2P%	0.601070	0.524603	0.064905
3PA	79.742188	117.895390	2.178571
3P%	0.243109	0.336718	0.049345
Effective_Shooting_pct	0.588555	0.528079	0.081393
True_Shooting_pct	0.612578	0.556962	0.116738
Points_per_game	10.553906	6.861613	0.392857
Rebounds_per_game	7.287500	2.932004	0.690476
Total_Rebound_pct	17.007812	8.685372	7.642857
Assists_per_game	1.464063	1.509840	0.307143
Assist_pct	9.237500	12.115160	8.069048
Steals_per_game	0.644531	0.589344	0.226071
Blocks_per_game	1.044531	0.296738	0.029762
Turnovers_per_game	1.202734	0.764911	0.231667
Versatility Index	7.917969	6.401596	1.090476
Offensive_Rating_Individual	121.841406	112.795745	19.307143
Defensive Rating	102.113281	107.092465	44.445238

Segment KMeans PCA	Medium Impact Player
AGE	26.839450
GP	62.899083
MPG	28.466055
Minutes_pct	59.306422
Usage_pct	20.350459
Turnover_Rate	11.275688
FTA	130.389908
FT%	0.808023
2PA	343.082569
2P%	0.496335
3PA	302.036697
3P%	0.352913
Effective_Shooting_pct	0.517679
True_Shooting_pct	0.550408
Points_per_game	13.098624
Rebounds_per_game	3.894954
Total_Rebound_pct	7.444037
Assists_per_game	3.128899
Assist_pct	16.584862
Steals_per_game	0.942615
Blocks_per_game	0.370000
Turnovers_per_game	1.485367
Versatility Index	7.374771
Offensive_Rating_Individual	110.328440
Defensive Rating	108.554128

7. Unsupervised Learning Algorithm, Gaussian Mixture Model ("GMM") In this section, I used a different unsupervised machine learning algorithm,Gaussian Mixture Models ("GMM") to segment the data.

The previous algorithm, KMeans clustering, uses a simple distance from cluster center to assign segments. If a data point overlaps two segments, KMeans forcibly assigns them to a cluster.

Gaussian Mixture Models look for a mixture of multi-dimensional Gaussian probability distributions to fit the data. It's more of a probability distribution or a mixture of different distributions.

"GMMs can be used to find clusters in data sets where the clusters may not be clearly defined. Additionally, GMMs can be used to estimate the probability that a new data point belongs to each cluster. Gaussian mixture models are also relatively robust to outliers, meaning that they can still yield accurate results even if there are some data points that do not fit neatly into any of the clusters. This makes GMMs a flexible and powerful tool for clustering data. It can be understood as a probabilistic model where Gaussian distributions are assumed for each group and they have means and covariances which define their parameters."

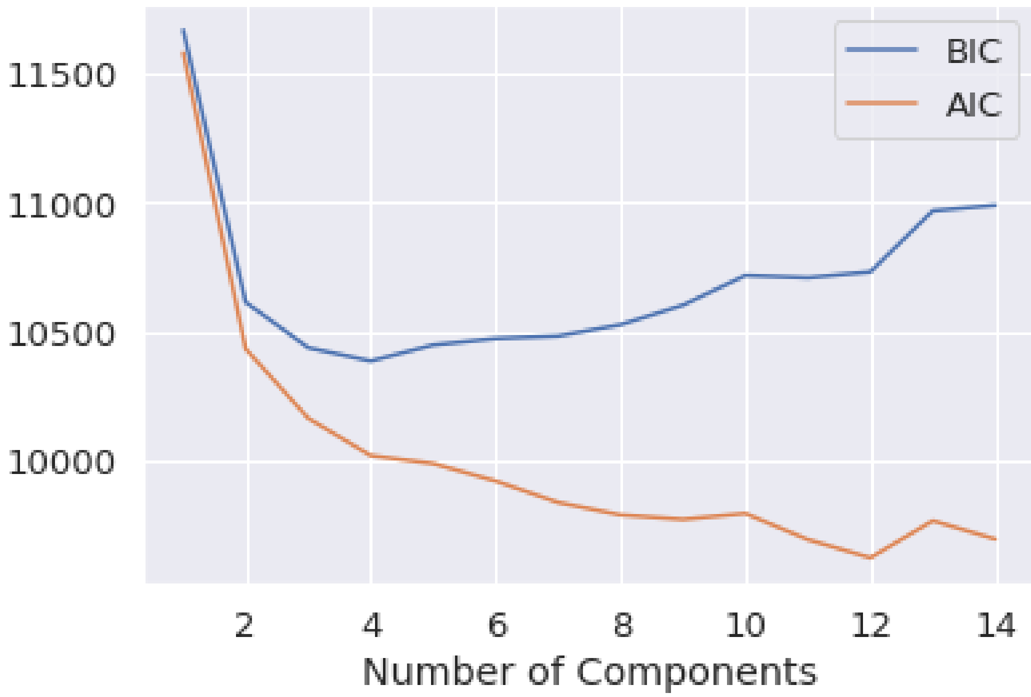
Ajitesh Kumar, *Gaussian Mixture Models: What are they & when to use?*, Data Analytics (April 14, 2022)

Created a DataFrame using the 5 PCA components

```
[ ]:    PCA Component 1  PCA Component 2  PCA Component 3  PCA Component 4  \
     0         1.207780        -0.318339         1.135486        -1.523582
     1         2.424570        -0.640418         3.551679         0.096325
     2         5.021161         0.868942         3.833111        -0.094303

        PCA Component 5
     0         0.156698
     1        -1.835523
     2        -0.009745
```

Finding the Optimal number of components I created a list comprehension which applied the Gaussian Mixture algorithm for 15 components and plotted the results from the Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC). The AIC measures the goodness of fit for the model while the BIC penalizes additional parameters.

In the model below, the best GMM model for the data is with 4 components.



Building the Model and Assigning Segments

The GMM model is fitted against the data dataframe which creates 4 segments. The GMM dataframe is created from a copy of the players dataframe along with a new feature called "segment" which is the segment that each player belongs to.

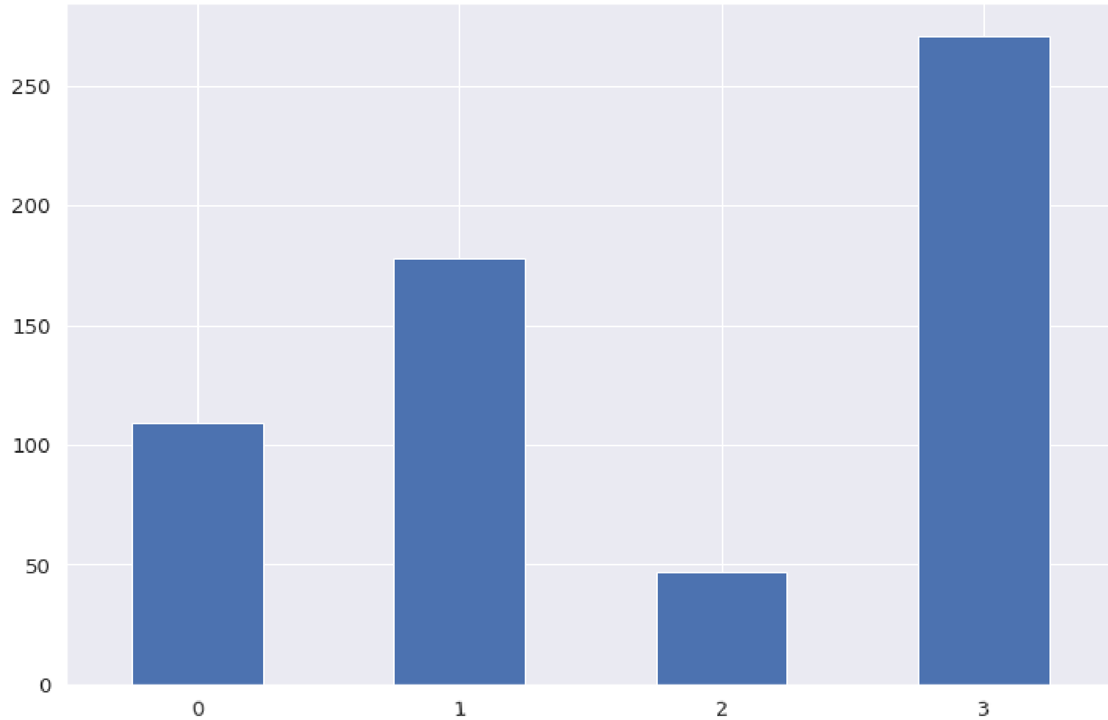
[ ]:

```
          FULL NAME  POS    AGE  GP   MPG  Minutes_pct  Usage_pct  \
0  Precious Achiuwa    F  22.56  73  23.6         49.2       18.5
1      Steven Adams    C  28.73  76  26.3         54.8       12.0
2       Bam Adebayo  C-F  24.73  56  32.6         67.9       25.0

   Turnover_Rate    FTA    FT%  …  Total_Rebound_pct  Assists_per_game  \
0           11.3  131.0  0.595  …               14.9               1.1
1           19.6  199.0  0.543  …               19.9               3.4
2           14.4  340.0  0.753  …               17.5               3.4

   Assist_pct  Steals_per_game  Blocks_per_game  Turnovers_per_game  \
0         6.9             0.51             0.56                1.15
1        16.1             0.87             0.79                1.51
2        17.5             1.43             0.79                2.64

   Versatility Index  Offensive_Rating_Individual  Defensive Rating  segment
0                6.8                        105.4             104.0        3
1                9.4                        124.7             103.9        0
2               10.7                        117.2              98.2        0
```

[3 rows x 28 columns]

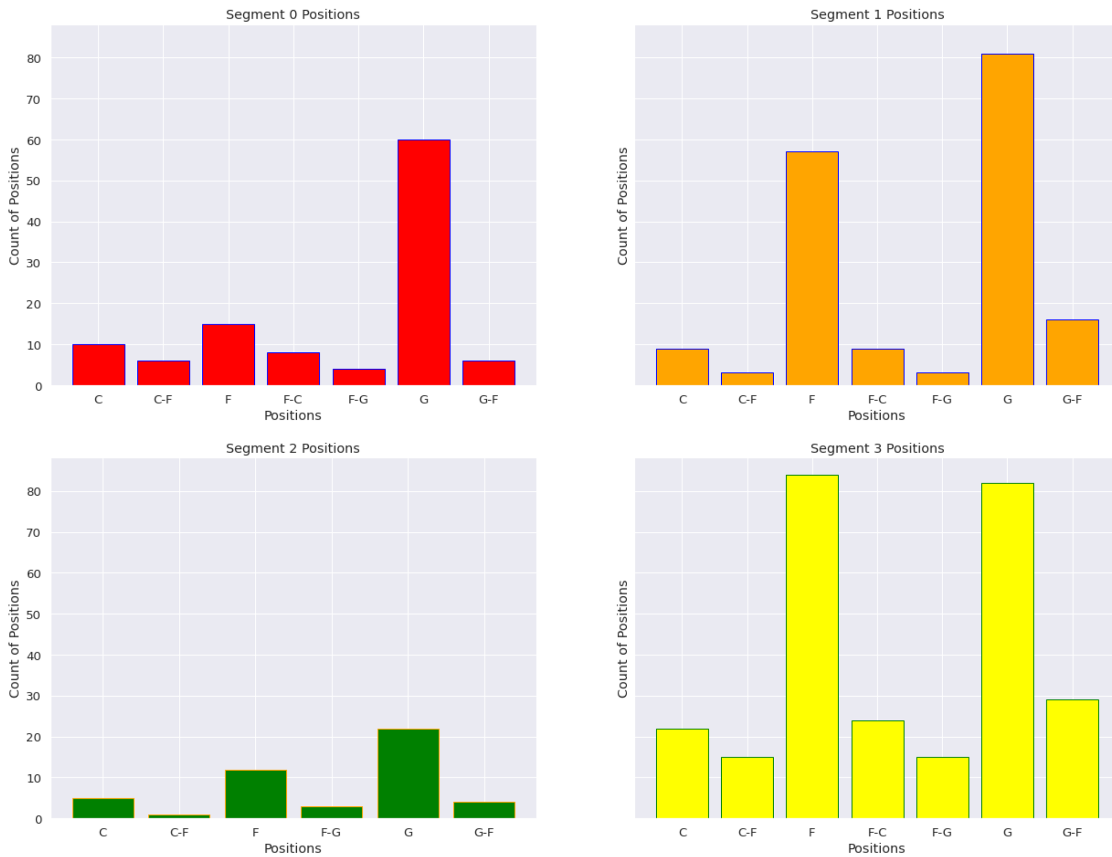
8. Analyze the GMM Segmentation results

Number of Players The model labeled four segments of players which I rename 0 through 3. Segments 1 and 3 contain the most players while Segment 2 has the least.



By Position For all of the segments, the position of Guard ("G") is the most common except for Segment 3 where the forward position edges it out.

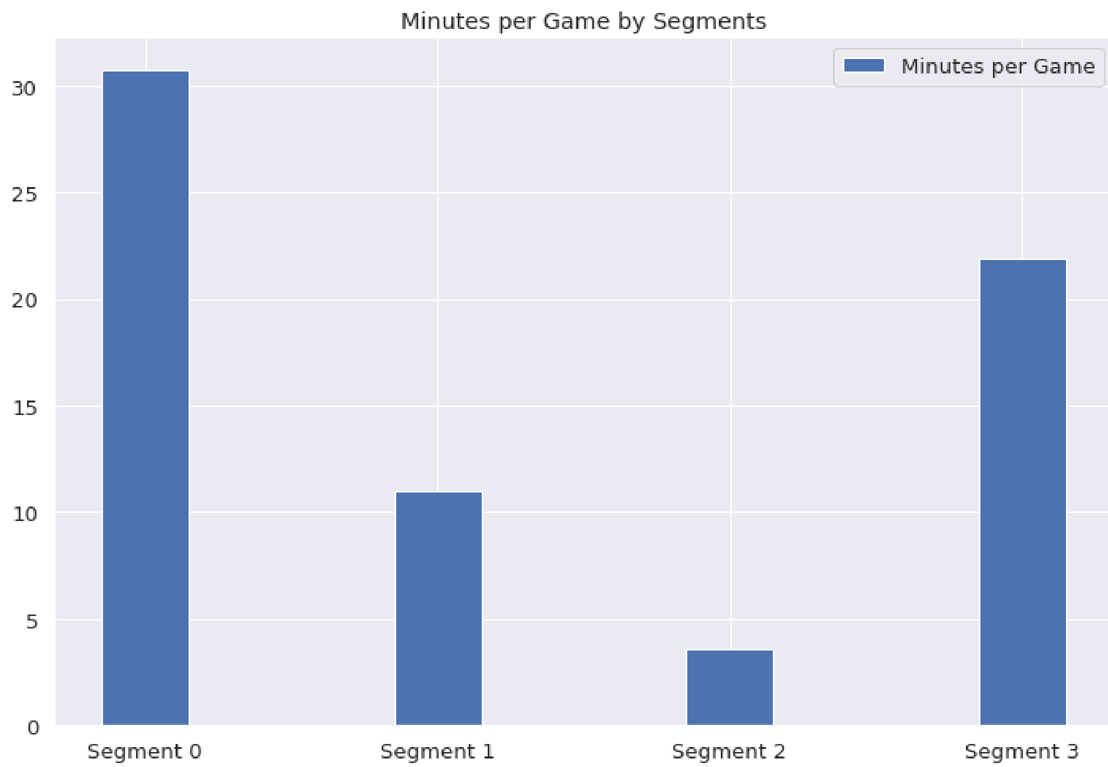
```
[ ]: Text(0, 0.5, 'Count of Positions')
```

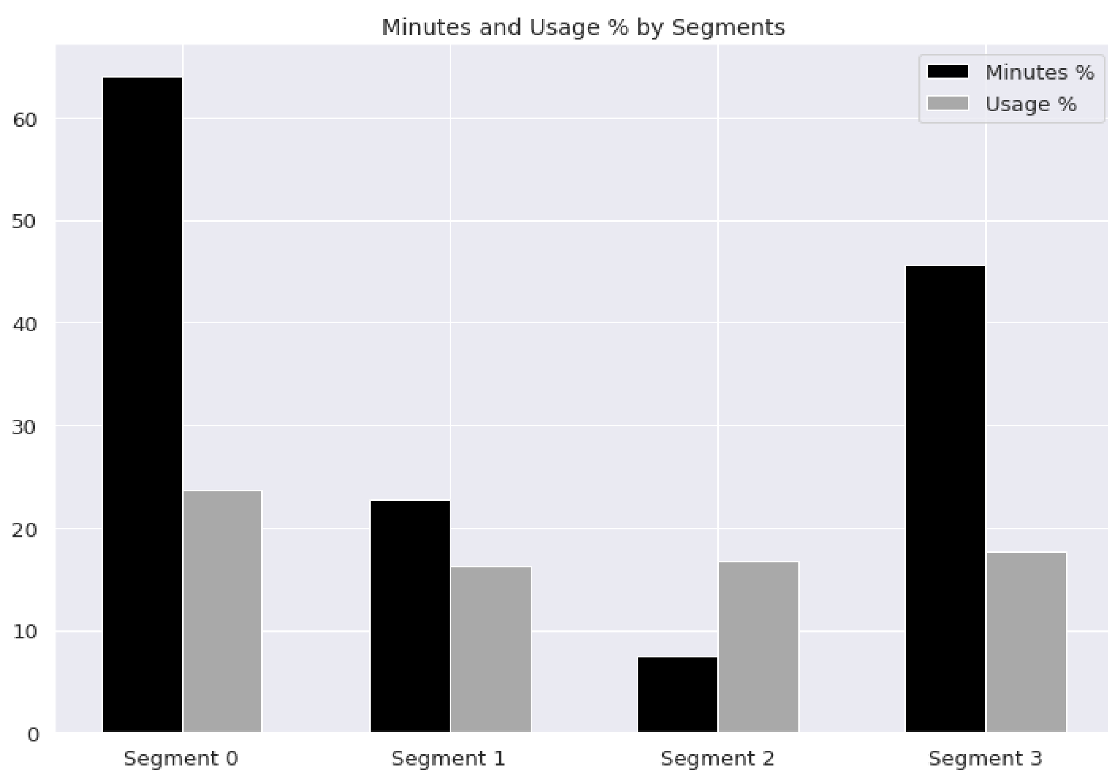


Discussion of Segments 1 and 2 Segment 2, the segment with the least number of players, also has the lowest average of number of Games Played, Minutes per Game, and Minutes Percentage. These are players that don't get a lot playing time.

Much like Segment 2, Segment 1 players don't get much playing time. These players have the second lowest Games Played, Minutes per Game, and the lowest Usage percentage. However, their Minutes per Game are more than double than that of Segment 2. These are bench players that do see significant time.

The players in these two segments are more than likely not every day players. These are very low impact players

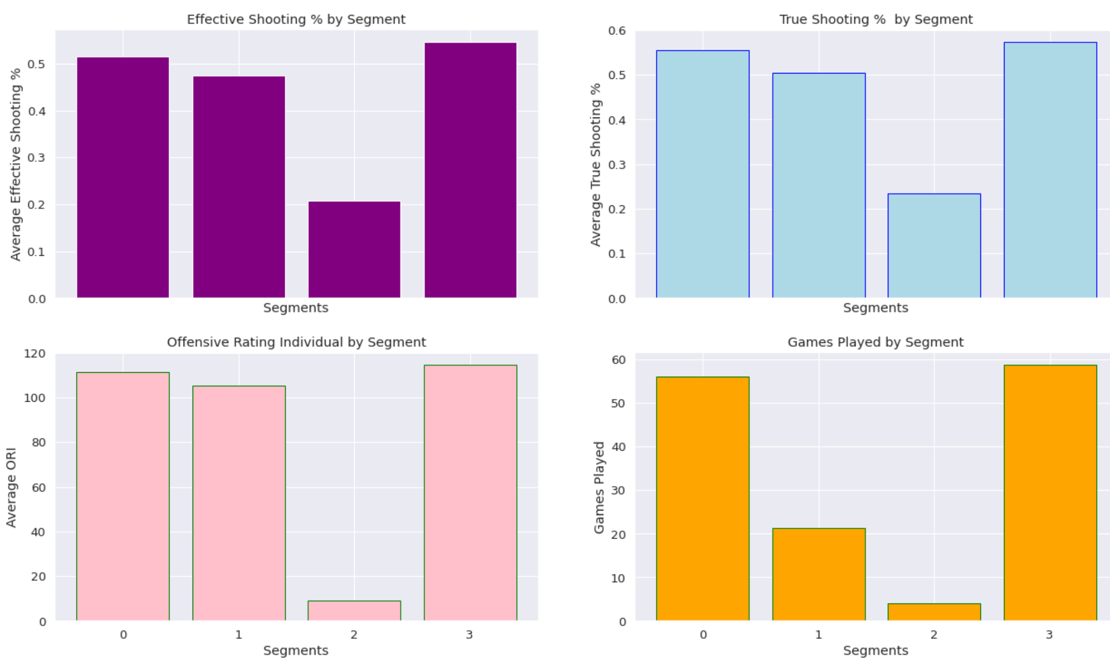

Minutes per Game by Segments
Minutes per Game
0
5
10
15
20
25
30
Segment 0
Segment 1
Segment 2
Segment 3


Minutes and Usage % by Segments
Minutes %
Usage %
0
10
20
30
40
50
60
Segment 0
Segment 1
Segment 2
Segment 3

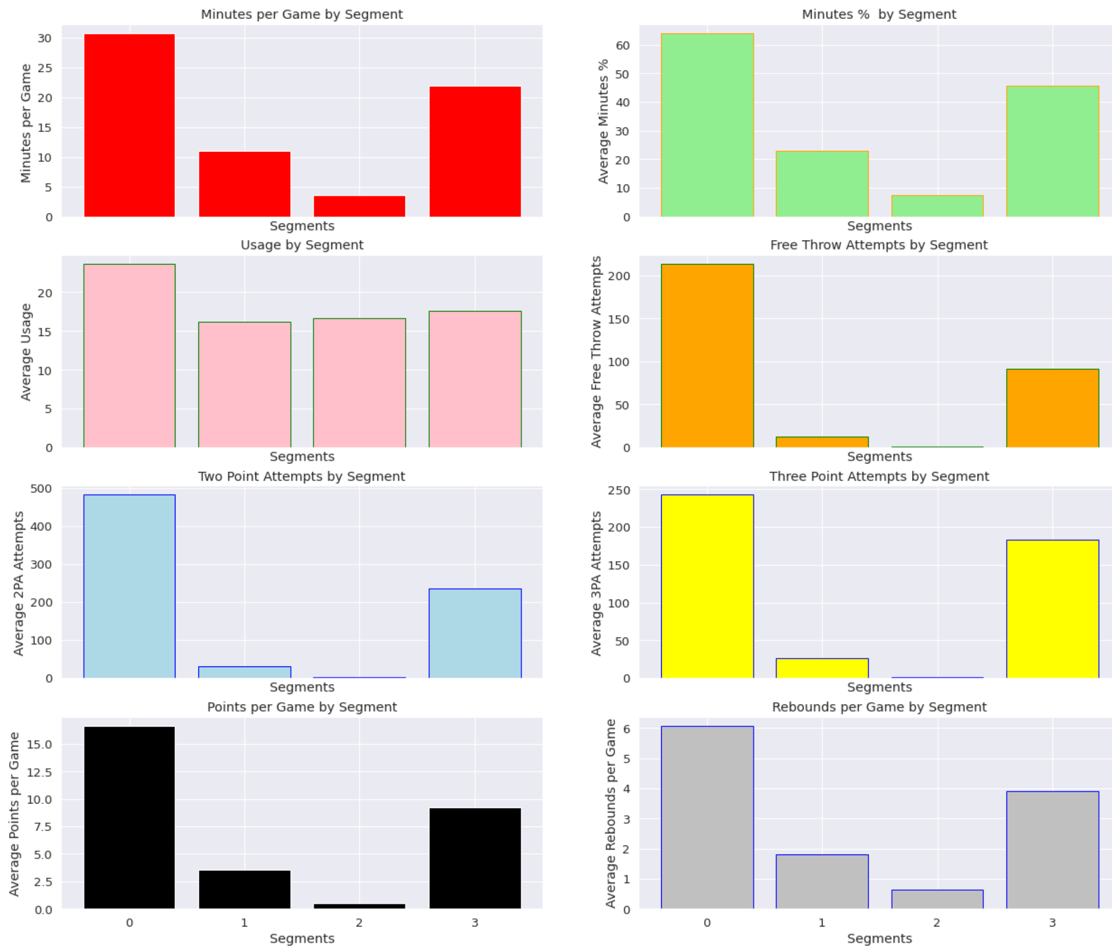
[ ]:

	FULL NAME	POS	GP	MPG	Points_per_game	segment
3	Santi Aldama	F-C	32	11.2	4.1	1
10	Thanasis Antetokounmpo	F	48	9.9	3.6	1
14	Ryan Arcidiacono	G	10	7.6	1.6	1
15	Trevor Ariza	F	24	19.3	4.0	1
17	Joel Ayayi	G	7	2.9	0.3	2
19	Udoka Azubuike	C-F	17	11.5	4.7	1
24	Dalano Banton	F	64	10.9	3.2	1
25	Cat Barber	G	3	4.3	0.0	2
30	Charles Bassey	C-F	23	7.3	3.0	1
31	Paris Bass	F	2	3.7	3.0	2
34	Kent Bazemore	G-F	39	14.0	3.4	1
38	Jordan Bell	F	1	2.0	0.0	2
46	Keljin Blevins	G	31	11.3	3.1	1
49	Bol Bol	C-F	13	6.2	2.5	1
50	Leandro Bolmaro	F	35	6.9	1.4	1
51	Isaac Bonga	G	15	4.6	0.8	1
55	James Bouknight	G	31	9.8	4.6	1
68	Sterling Brown	G-F	49	12.8	3.3	1
71	Shaq Buchanan	G	2	4.9	1.0	2
73	Trey Burke	G	42	10.5	5.1	1

Segment 3 Segment 3, the largest segment, has the highest number of Games Played, Effective Shooting percentage, True Shooting percentage, and Offensive Rating Individual. This segment makes major contributions, but they're not the elite players in the game. These are medium impact players



Segment 0 Segment 0, the second smallest segment, represents the elite players in the NBA. This segment leads in Minutes per Game, Minutes percentage, Free Throw Attempts, Two-and-Three point attempts, Points, and Rebounds per game.



The last 10 NBA MVPs are in this segment.

[ ]:

```
                FULL NAME POS  segment
9    Giannis Antetokounmpo   F        0
108          Stephen Curry   G        0
131          Kevin Durant    F        0
228          LeBron James    F        0
241          Nikola Jokic    C        0
481      Russell Westbrook   G        0
546          James Harden    G        0
```

GMM Analysis

The Gaussian Mixture Model algorithm identified 4 segments, and after the above analysis, I can rename them accordingly:

Segment 0 - "Elite Player" Segment 1 - "Low Impact Player" Segment 2 - "Extremely Low Impact Player" Segment 3 - "Medium Impact Player"

[ ]:

```
segment                             Elite  Low Impact   Marginal  Medium Impact
AGE                             27.486789   25.531348  24.372766      26.717085
GP                              56.174312   21.280899   4.021277      58.822878
MPG                             30.715138   10.961236   3.574468      21.874047
Minutes_pct                     63.987156   22.829307   7.440426      45.569557
Usage_pct                       23.651835   16.240730  16.687234      17.587392
Turnover_Rate                   14.178440   11.794757  13.553191      10.999200
FTA                            213.509174   11.948502   0.936170      91.988622
FT%                              0.780711    0.571949   0.135149       0.764453
2PA                            482.160550   31.959738   2.191489     235.694649
2P%                              0.513096    0.477739   0.230234       0.542547
3PA                            242.605505   26.059925   1.723404     183.863469
3P%                              0.303028    0.246070   0.054106       0.326748
Effective_Shooting_pct           0.514963    0.475292   0.207723       0.545744
True_Shooting_pct                0.554018    0.504710   0.234213       0.573058
Points_per_game                 16.659174    3.600749   0.529787       9.197694
Rebounds_per_game                6.075229    1.811049   0.629787       3.917589
Total_Rebound_pct               10.738991    9.436049   8.140426      10.207226
Assists_per_game                 4.790367    0.823596   0.248936       1.649416
Assist_pct                      24.303211   10.660487   7.614894      10.863561
Steals_per_game                  1.080550    0.327631   0.200000       0.666261
Blocks_per_game                  0.616330    0.182097   0.048511       0.442648
Turnovers_per_game               2.330046    0.476348   0.252553       0.954788
Versatility Index                9.501376    5.117322   1.572340       6.786193
Offensive_Rating_Individual    111.501376  105.339700   9.195745     114.660517
Defensive Rating               106.759633  104.967509  27.476596     106.211562
```

[ ]:

```
     FULL NAME  POS    AGE  GP   MPG  Minutes_pct  Usage_pct  Turnover_Rate  \
28  RJ Barrett  F-G  21.82  70  34.5         71.9       27.6            9.9

      FTA    FT%  …  Total_Rebound_pct  Assists_per_game  Assist_pct  \
28  406.0  0.714  …                9.0               3.0        14.9

    Steals_per_game  Blocks_per_game  Turnovers_per_game  Versatility Index  \
28             0.61             0.23                2.16                8.2

    Offensive_Rating_Individual  Defensive Rating  segment
28                        103.4             108.3    Elite

[1 rows x 28 columns]
```

8 V. Summary of Findings

The project was a success. Both unsupervised learning cluster algorithms identified patterns in the data that reflected reality. Both were able to discern low, high, and elite impact players. The end deliverable are two data tables where the owner, James Duarte, can look up a player and get a clear idea of his value in relation to other players in the league. The project will continue with PowerBI where it will be more graphically interactive. For future projects, we can segments within segments analysis as well as supervised machine learning projects using the segment labels.

9 VI. Conclusions

Both algorithms were sucessful in their objectives to cluster the players based on their performances. The KMeans algorithm created two additional clusters that the GMM algorithm did not which was a surprising finding since KMeans is more inflexible than GMM.

Additionally, both algorithms showed how important Games Played and Minutes are to creating clusters. Originally, I omitted players with low playing time, but these algorithms were able to differentiate these players from the rest.

10 VII. References

1. K-Means Clustering in Python: A Practical Guide 2021 English Real Python https://realpython.com/k-means-clustering-python/
2. Wikipedia: Principal component analysis
3. Chris Ding and Xiaofeng He, *K-means Clustering via Principal Component Analysis*, K-means Clustering via Principal Component Analysis, Proc. of Int'l Conf. Machine Learning, (2004)
4. Ajitesh Kumar, *Gaussian Mixture Models: What are they & when to use?*, Data Analytics (April 14, 2022)

Write the output to csv files

```
[NbConvertApp] Converting notebook /content/drive/MyDrive/NBA Stats/NBA Players
2021-22 Segment Analysis/NBA Players Segment Analysis.ipynb to pdf
[NbConvertApp] Support files will be in NBA Players Segment Analysis_files/
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
```

```
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Making directory ./NBA Players Segment Analysis_files
[NbConvertApp] Writing 188970 bytes to ./notebook.tex
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: ['xelatex', './notebook.tex', '-quiet']
[NbConvertApp] Running bibtex 1 time: ['bibtex', './notebook']
[NbConvertApp] WARNING | bibtex had problems, most likely because there were no
citations
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 995017 bytes to /content/drive/MyDrive/NBA Stats/NBA
Players 2021-22 Segment Analysis/NBA Players Segment Analysis.pdf
```