Final Project

Alex Wako

2023-12-13

Problem

The NBA or the National Basketball Association is the premier basketball league in the United States and is also one of the most difficult leagues for players to play in. Only 450 players in the world can actively participate in a NBA roster making it the most prestigious job for basketball players. The problem for the NBA is that in their 30 team league, each team can only roster 15 players. This means that in a game with 5 players on the court, only 10 players are available for substitution. If a player were to get injured or feel fatigued from the constant game schedule of the NBA, a team would have one less player to either start or have as a potential substitute. Additionally, the skill levels of the very best to the bench varies (as in any sport), so a loss in a key player can potentially set back the team a large amount. One solution to deal with this problem is to utilize a player's versatility to their fullest.

Each player usually has one preferred position that they are effective in. Within these positions, the common names are:

- **PG:** Point Guard
- **SG:** Shooting Guard
- **PF:** Power Forward
- **SF:** Small Forward
- **C:** Center

These positions are usually assigned by a player's physical capabilities including their speed, size, and their skills, but the game is constantly evolving and a player's style in their main position can potentially translate over to another position. With that in mind, creating an environment where players play different positions will increase the adaptability of the team and increase the overall strength of the 5 players on the court. Understanding play styles of players can also help in developing teams around certain ways of playing basketball.

Having a versatile player in the NBA leads to a stronger and a more effective team, so for this project, the models used attempts to resolve the problem by understanding the different play styles the NBA has for each position and potentially identify a secondary position for the player. The official NBA website provides general positions that players had played including a secondary players, although for many players the minutes with the secondary position are close or equal to zero, and we will use that as reference to see if player stat lines can seperate these players in to a primary and secondary position.

The Models

The model used for this projects will be K-Means Clustering and Hierarchical clustering as the plan is to identify the many different play styles of each position and attempt to categorize each player to a second position through different clusters.

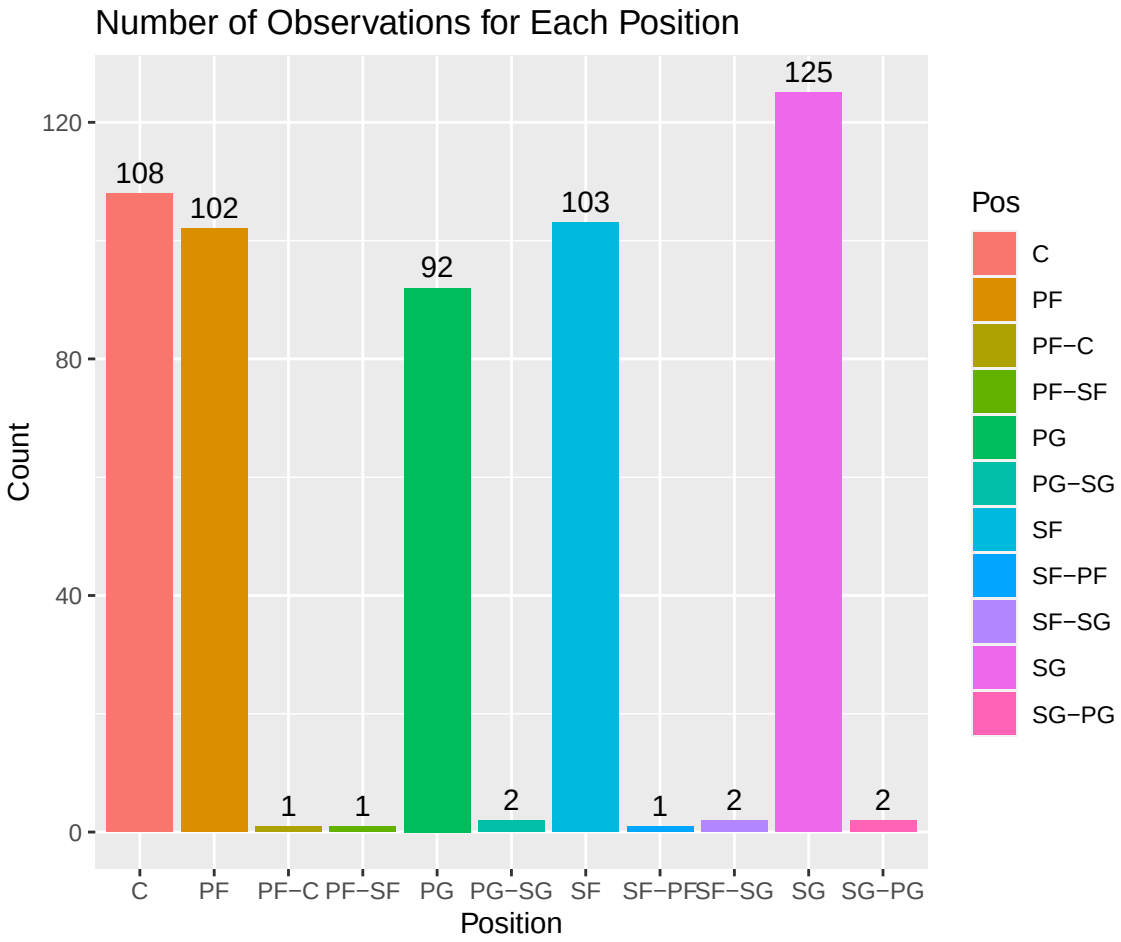
K-Means Clustering: This model is a form of an unsupervised machine learning model that is used to cluster data into K distinct groups or clusters. In this problem, because there are five positions in basketball, the number of distinct groups that will be considered is K=25; five groups for each primary and secondary position combination. The second group of clusters being considered is K=9. The two guard and forward

positions can hold similar responsibilities, so grouping them together can easily identify three group for each primary and secondary position combination. The general idea of K-means clustering is to visualize a plane, where players with similar statistical profiles are grouped together.

Hierarchical Clustering: This model is a form of an unsupervised machine learning model that is used to cluster data into a tree like figure called a dendrogram. The goal of the model is to group individual clusters of data together through a bottom up approach so in the end there will only be one group. The height represents the measure of dissimilarity between two clusters and does not represent the number of clusters grouped together.

The Data

The data used for the project is the per game basketball stats of each player for the 2022-2023 NBA season. All data is sourced by the 'Sports Reference' website, specifically the Basketball reference website.



The data contains a combination of 9 different observations of positions not in the standard positions. As these observations are not useful for the model and removing them will likely not affect the data too much, these 9 observations can be removed.

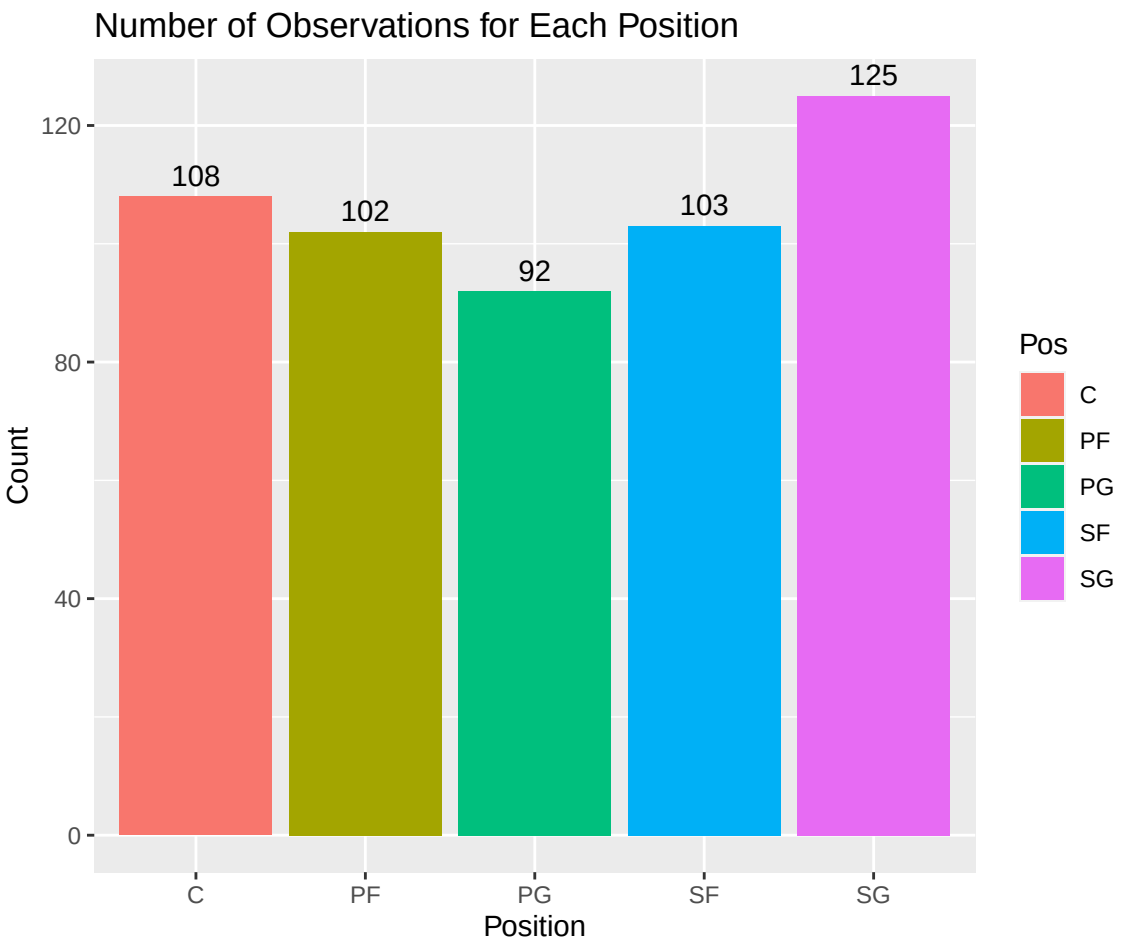
```
##             Player Pos  G  eFG.   FT. ORB DRB AST STL BLK TOV  PF  PTS
## 1 Precious Achiuwa   C 55 0.521 0.702 1.8 4.1 0.9 0.6 0.5 1.1 1.9  9.2
## 2     Steven Adams   C 42 0.597 0.364 5.1 6.5 2.3 0.9 1.1 1.9 2.3  8.6
## 3      Bam Adebayo   C 75 0.541 0.806 2.5 6.7 3.2 1.2 0.8 2.5 2.8 20.4
## 4     Ochai Agbaji  SG 59 0.532 0.812 0.7 1.3 1.1 0.3 0.3 0.7 1.7  7.9
## 5     Santi Aldama  PF 77 0.560 0.750 1.1 3.7 1.3 0.6 0.6 0.8 1.9  9.0
```

The data now includes the following columns:

- **Player:** Name of the player
- **Pos:** The position played
- **G:** Games Played
- **eFG.:** Effective field goal percentage
- **FT.:** Freethrow percentage
- **ORB:** Offensive rebound
- **DRB:** Defensive rebound
- **AST:** Assist

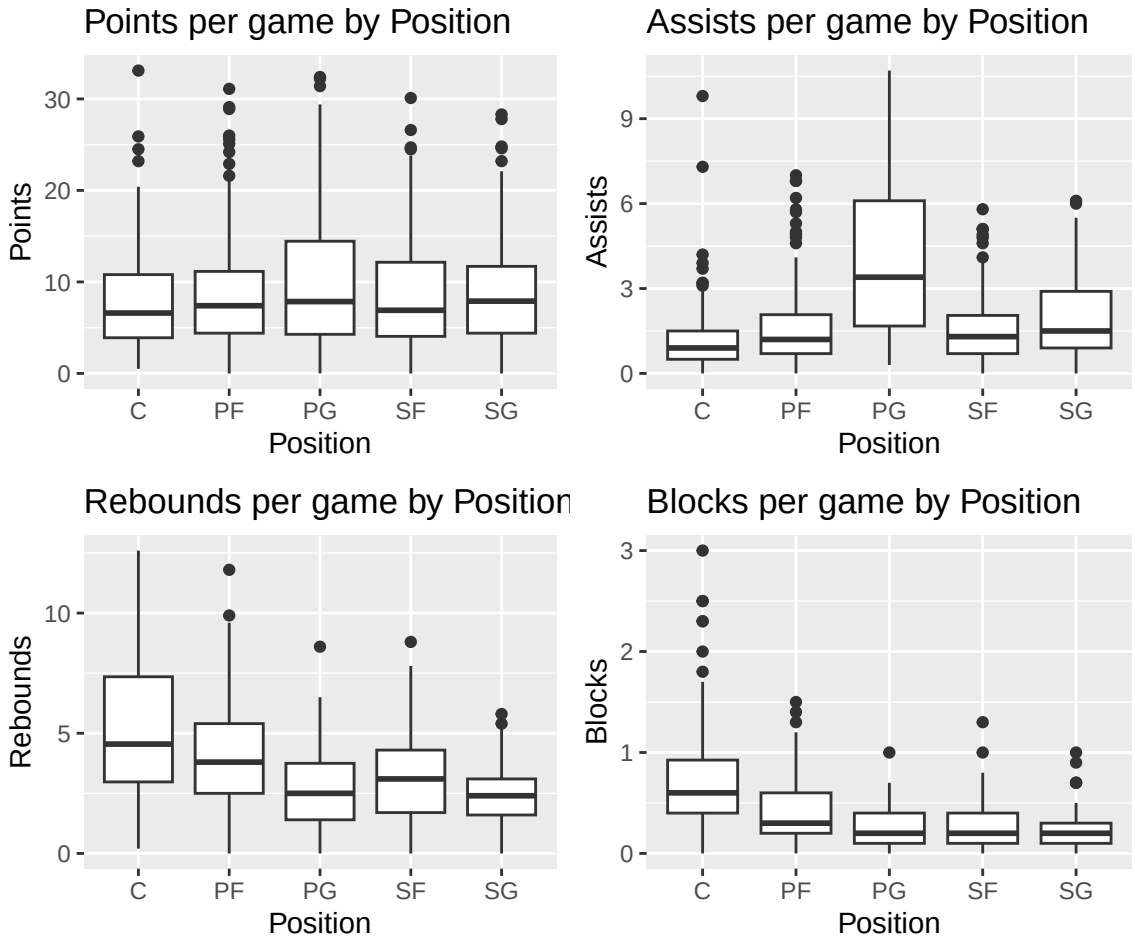
- **STL:** Steal
- **BLK:** Block
- **TOV:** Turnover
- **PF:** Personal Fouls
- **PTS:** Points

To get a more information of the data, split the data into each position. Information can be visualized to better understand the roles of each position, and what can be the expected secondary positions.



There aren't any positions with considerably more data, so no adjustments have to be made to account for bias.

The three core stats of basketball are points scored, rebounds, and assists per game. These are the easiest to visualize in a basketball and usually the easiest and most effective way in knowing a players per game output. Of course defensive responsibilities are hard to see through these three stats alone, but NBA rule changes have shown that NBA basketball prioritizes teams with attacking prowess, so taking that into consideration, it is simple and effective to compare players and positions on their attack. Additionally blocks are another easy way to know the position of a player. The center is a position primed for big and tall players which usually result in a higher block count, while the two guard position have smaller players that can easily facilitate the game through quicker movements resulting in a lower block count.

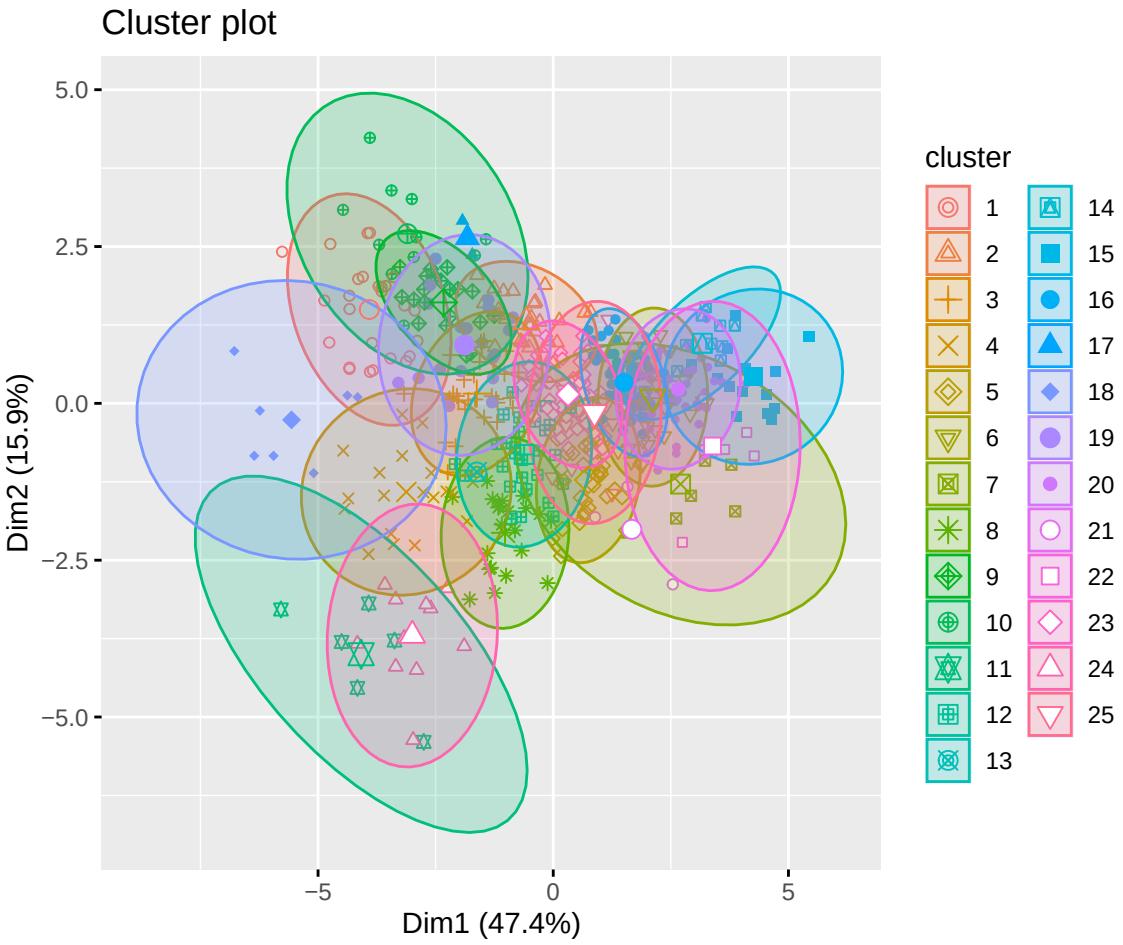


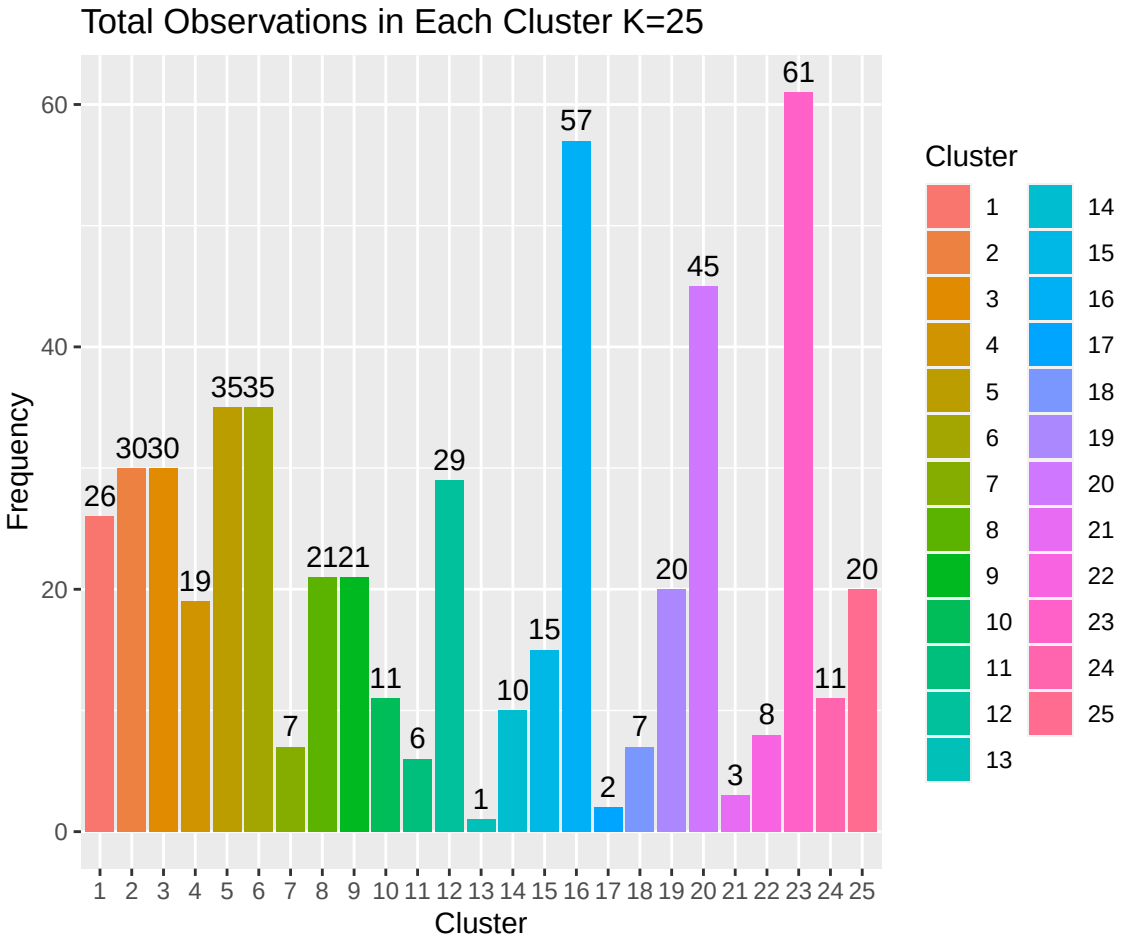
As the data shows, stats where physical attributes are taken to account ie. rebounds and blocks are more likely to be dominated by larger players, mainly centers then forwards. On the other hand, as previously mentioned, point guards whom facilitate the game have a much larger range of assists than any other positions. In the NBA, teams will have a main ball carrier - a person in the team that is most comfortable/skilled on the ball - which can skew the data a little, but overall the stats showcases the role of each position.

Modeling

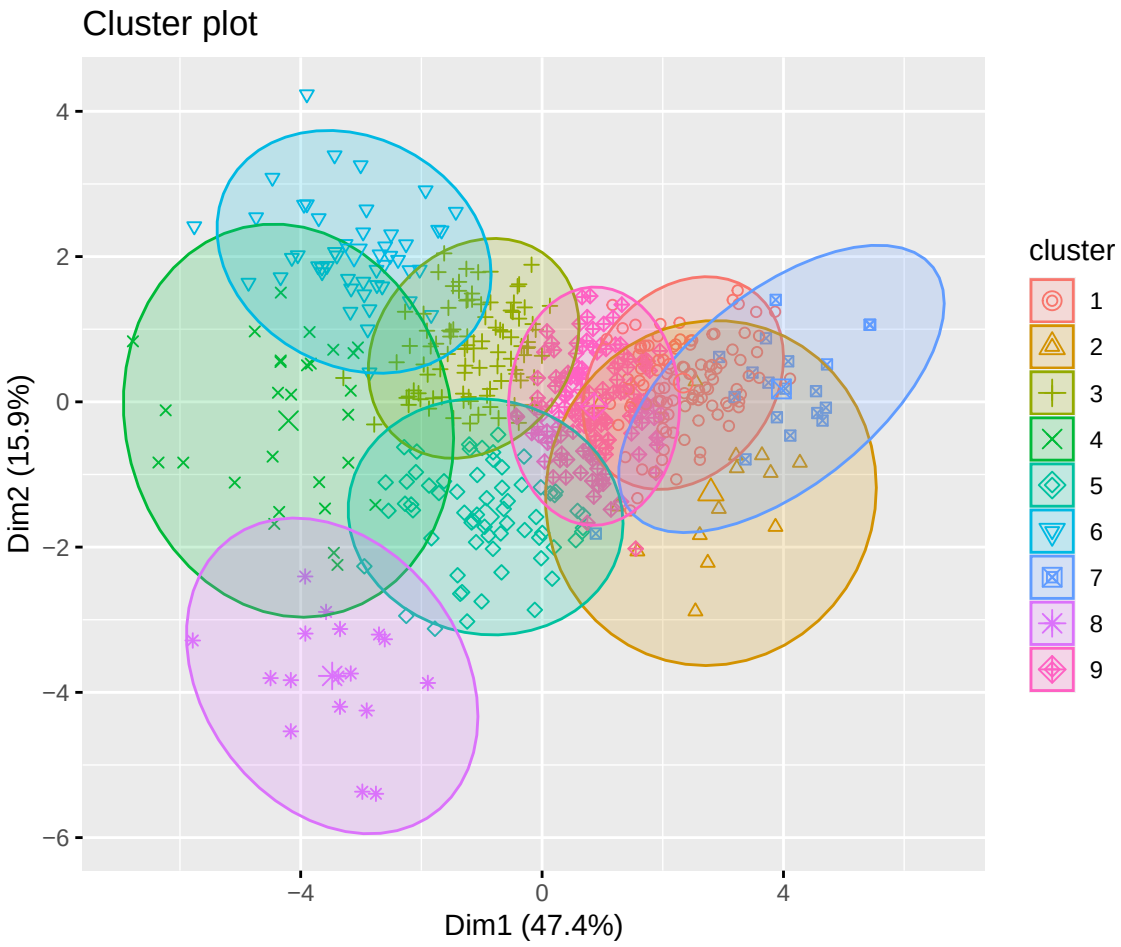
First approach the problem using K-means modeling.

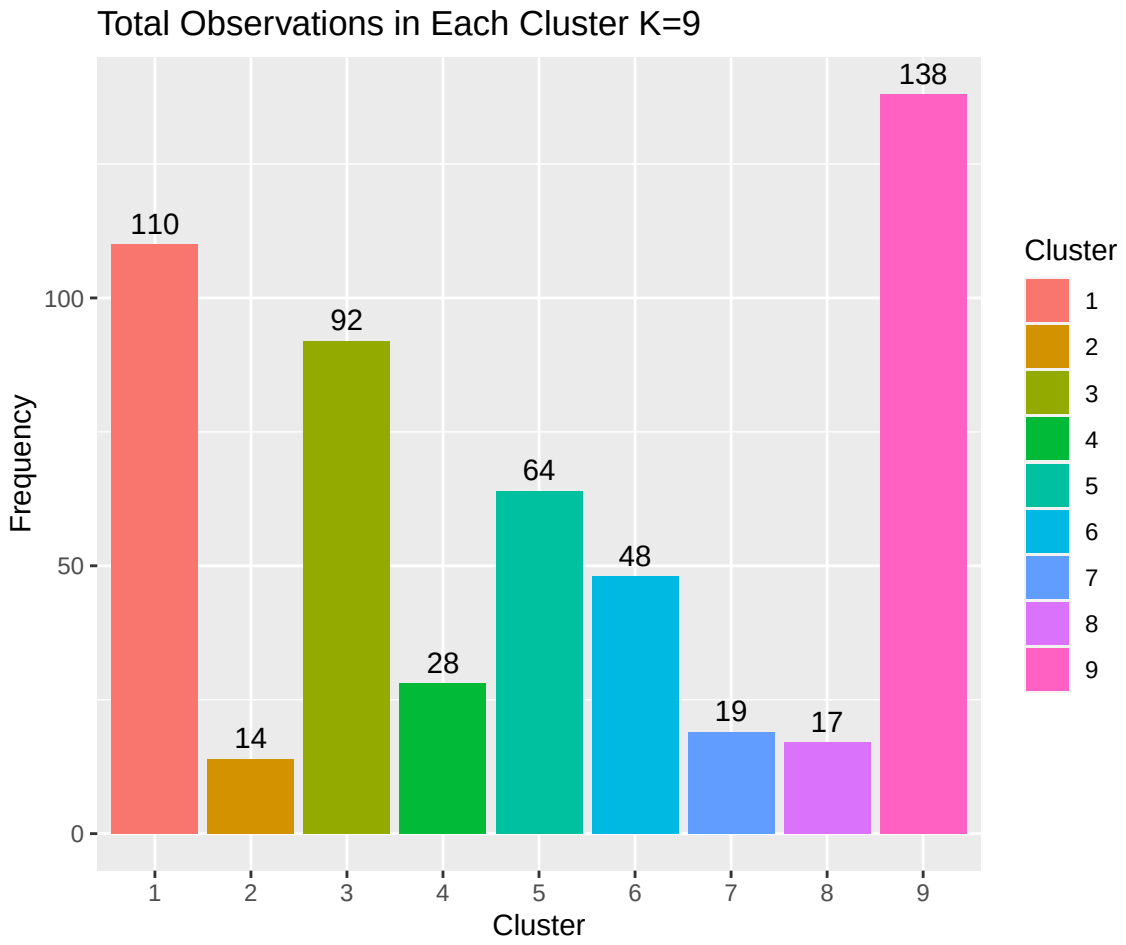
```
## Too few points to calculate an ellipse
## Too few points to calculate an ellipse
## Too few points to calculate an ellipse
```



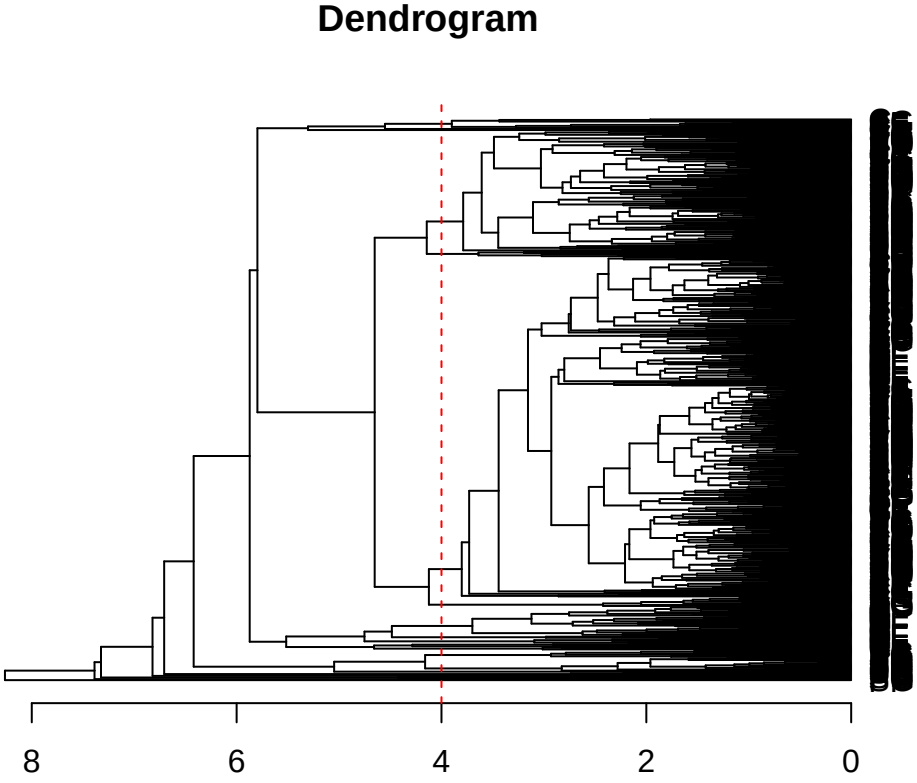


With 50 different sets of random centers and identifying the best model out of the 50 models using the minimum within sum of square, theoretically, if each player were to have one zero or one secondary postion, there would be 25 different combinations of primary and secondary positions. Therefore, a K=25 K-means clustering approach can allow for the model to assign primary and secondary positions based on statistical similarities. However, the Point Guard and Shooting Guard positions and the Power Forward and Small Forward positions can have similar responsibilities, so it can be useful to group the positions together for a better split of primary and secondary roles.

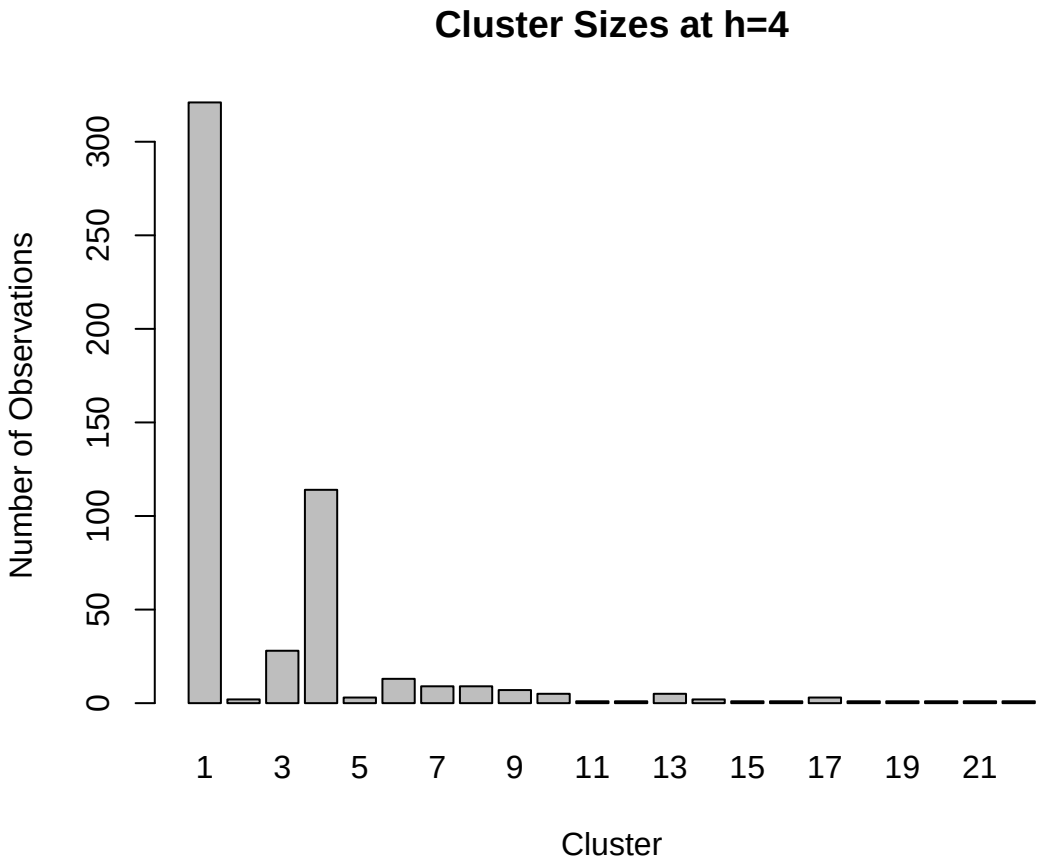

Cluster plot
Dim2 (15.9%)
Dim1 (47.4%)
4
2
0
−2
−4
−6
−4
0
4
cluster
1
2
3
4
5
6
7
8
9



As mentioned before, K-means clustering can be potentially limiting due to the nature of the model. An initial K value reduces the flexibility of the model, making it difficult to spot any statistical patterns that models without restrictions can otherwise spot. The next model used is a hierarchical clustering model.

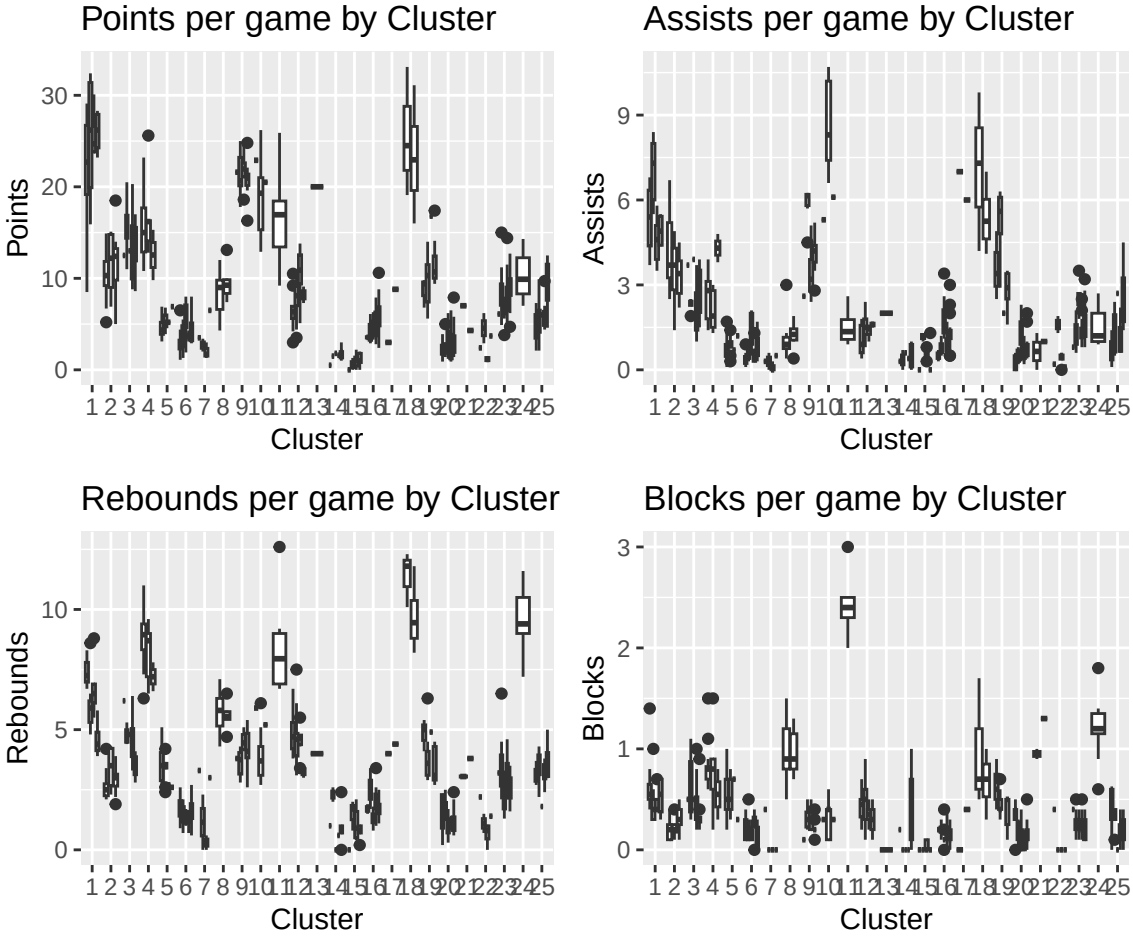


The hierarchical clustering model gives a much more different result to the K-means clustering model. With more clusters that spread individually, the visual can be harder to identify the positions. We can see the number of observations in the cluster through the bottom graphic at height equals to 4.

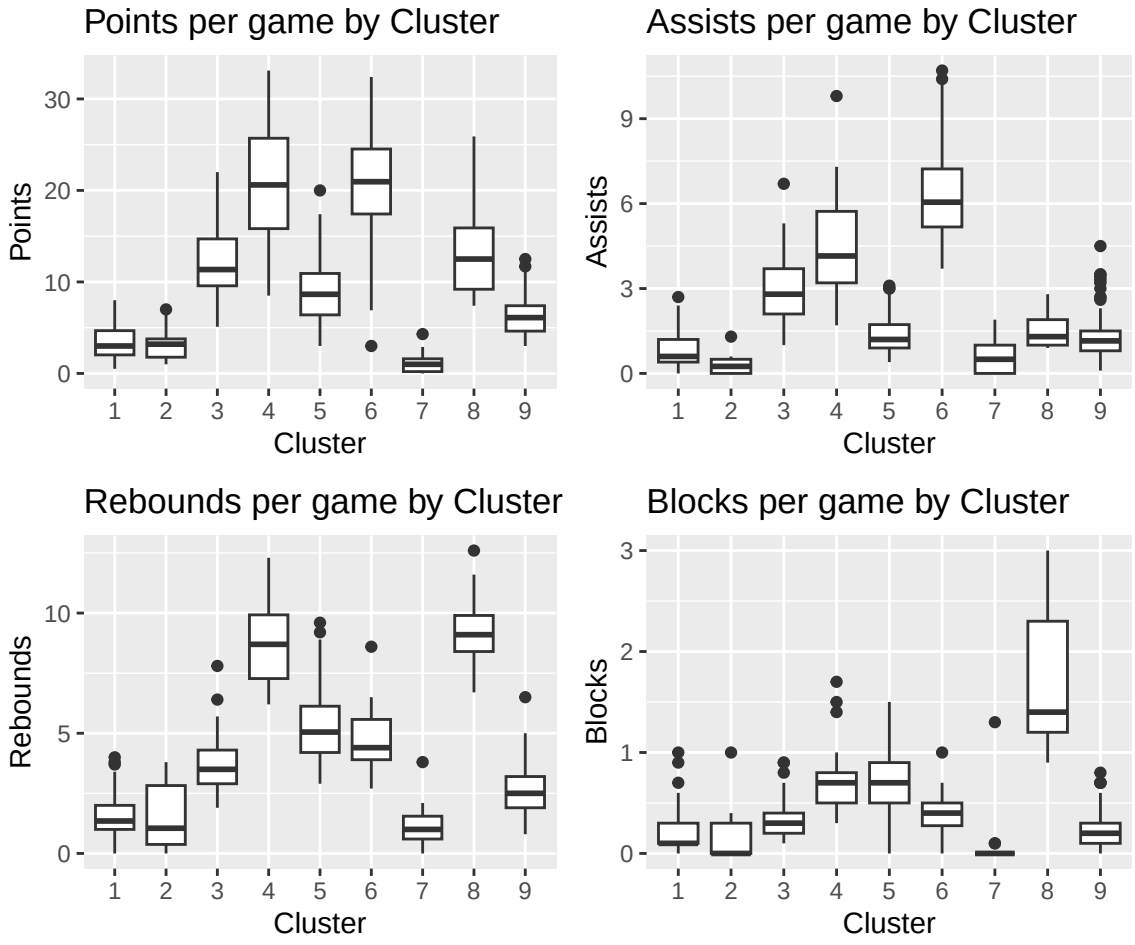

Cluster Sizes at h=4
Number of Observations
0
50
100
150
200
250
300
1
3
5
7
9
11
13
15
17
19
21
Cluster

Conclusion

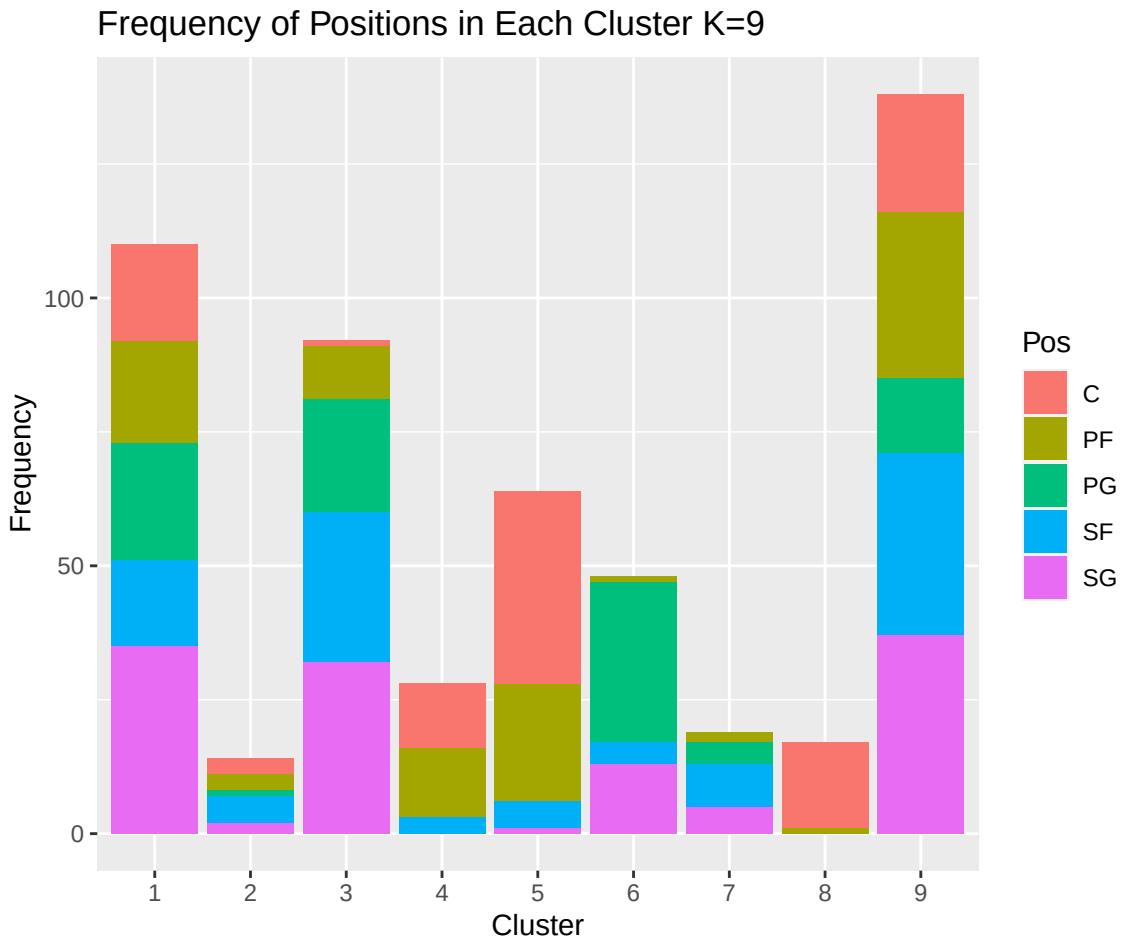
Looking at K-means clustering model, the data can be too small to understand a players primary and secondary position.



By looking back at the box plot created earlier when describing the data, it can be compared to the one above and can reach to the conclusion that information is lacking. Many of the box plots are just points on the plot, and as shown in the histogram earlier, a fraction of the clusters have less than 10 observations. Even when clustering, three error points are printed, showing that too little information is used to form three of the ellipses. In conclusion, a season's worth of data for all the players that played in the NBA that season lacks the data necessary to identify 25 different combinations of primary and secondary positons. On the other hand, when grouping similar positions together, a different result emerges.



Results can more easily be visualized with K=9 clusters. For example, cluster 8 is clearly a pure center with no secondary position. As mentioned earlier, centers are players with high rebounds and blocks and only a pure center will fit the description of cluster 8. Seperately, cluster 6 has high assists and scoring numbers, matching the discription of a guard. To better visualize the spread of postions by clusters, the below graph splits the observations of each cluster by their position from the data.



Both cluster 6 and 8 fit the exact description of the conclusion. Both either pure guards and centers, respectively, these players do not have a secondary position according to the model. On the other hand, both clusters 4 and 5 have an abundant amount of centers and forwards, meaning either the players in these clusters are primary forwards and secondary centers or primary centers and secondary forwards.

The guard position is much harder to distinguish then the rest of the positions. While cluster 6 shows a pure guard, the rest of the clusters have a mix of every position, making it harder to identify what other positions guards can play. This could mean that the guard position is usually exclusive to only guards and that main guards cannot branch out to other positions, or that in the NBA, guards and forward do similar things in terms of stat lines, making it harder to separate the two positions. Either way, it can be said that the center position is the easiest to identify, and likely forwards and centers are much more interchangeable than guards to any other position.

Alternatively, the hierarchical clustering model shows how the data is split differently. As shown in the vertical line at height equals 4 and the plot representing the numbers of observations split into clusters at height equals 4, the data is skewed to just one group. As the height goes up and the clusters become more generalized, there will be even more clusters of NBA players in just one position. Alternatively, if the height decreases, the positions will be more unique, but since there are already 22 different clusters in the model at height equals 4, these clusters will become unidentifiable to people as combinations of positions in a basketball court. Ultimately, the hierarchical clustering model fails to identify and represent the combination of two positions in a basketball court through NBA data.

Through both models I can conclude that some forwards and centers are interchangeable, while guards are likely exclusive to their position. In the clustering model, there are two different clusters representing a large group of power forwards and centers, meaning power forwards can play the center position, while small forwards, like guards, are more exclusive to their own position. Interestingly, in cluster 3, a large chunk of

guards and small forwards exist, which may make it so we can have two groups of combinations; those being small forwards and guards or power forwards and centers. By looking at NBA's official website, I can already see that players like Anthony Davis, who are listed as primary forward, also have a center tag with him. On the other hand, a smaller forward like Scottie Barns is listed as both a forward and guard. Size seems to be a big factor into what the players' positions are, which the data does not contain and should likely be used for future improvements. Overall, the clusters do notice a trend in positions based on players' play style and stat line and seems to fit the center and power forward role quite accurately. Guards seem harder to identify to a second position, which may just mean that guards are exclusive to their own position, but finding new roles is a way to improve the team, so these limitations in this project should be improved upon to help find new positions for players.

Future Improvements

There are several ways to enhance the future quality of the clustering models.

- Incorporating Body Information: Public NBA data is not limited to the thirteen variables used for this project, so adding new information can be done. Some data that can help identify player positions are their physical attributes. For example, centers are usually the largest players on the team, so information such as weight and height are important to the center role. Injury information can also help identify positions. Ball handlers such as the point guard position are more likely to get blocked when they attempt to score, and bad contact with an opposing player can lead to injuries. Therefore, players with more common injuries can indicate a certain position on the court.
- Including Contextual Data: Players already shown to have secondary positions can have a separate observation that includes their stats from that position. This can help identify the difference a players stats may have when playing in two different positions.
- Using Data from Previous Seasons: The game of basketball has evolved, but the rule and plays are still fundamentally the same, so using data from previous seasons will likely still help in identifying player positions. For example, players just recently began to take more three point shots due to the effectiveness of the Golden State Warriors and Steph Curry, who used his three point making skills to lead them to multiple championships, but the roles of each position is still the same. The center is still the big man and the point guard is still the game facilitator. Using previous season data will advance the models' intelligence and identify position playstyles influenced by older generations.

By addressing these potential points of improvement, the clustering models can develop into a much more sophisticated tool to identify and undestand player roles in basketball. Coaches and analysts can see the players on the court, but when it comes down to effectiveness and versatility, player statistics can analyze and group current and potential positions to better improve the player and the team.