An Analysis of Positions in the NBA

Julian, Kush, Trevor, Ethan

4/23/2020

Introduction

This project explores the NBA and looks at how the game of basketball has evolved over time.

Our first question was to determine what positions actually exist in the NBA today. As the game of basketball has evolved over time, the set positions we have defined over the course of history have altered. What the players do in the game and the positions they are assigned sometimes do not correlate with each other. Through k means analysis, we set out to determine what “positions” actually exist in the NBA today.

Our next question was to understand how positions in the NBA have changed over time. People often talk about the NBA evolving as seasons go by. For example, the game used to be centered around inside play, then it transitioned into “iso” basketball where each team would give the ball to their best wing player and get out of their way. Now, people talk about how teams shoot as many three point shots as possible, and that guards (the position that often shoots the most three point shots) are the most important.

Branching off that question, we also set out to discover how Steph Curry and the Warriors altered the way basketball in the NBA is played. After the warriors won a championship and then had a 73 win season by shooting (and making) tons of threes, many teams shift to adopt a similar style. Now there are teams (the Rockets) who have traded away their starting big-men in order to focus more on their star guards (James Harden and Russell Westbrook).

Our final question was to determine what factors affect a player’s shot. To do this we analyzed distance from defenders as well as position titles. This way we could get a comphrensive understanding of some factors that affect a person’s ability to make a shot.

Along with these questions, our group had a goal we set out to discover. This was to determine if changes in the way that players play impacts the way that NBA teams draft. Essentially our goal was asking do NBA teams try to draft to fit around the style of modern NBA play, or do they draft for their need with less regard for the style of play in the league as a whole?

We had another goal to predict the future of positions in the NBA. After recieving feedback emphasizing that kmeans is an exploratory model, not a predictive one, we chose to approach the goal in a different manner. We simply looked at younger players and their statistics in contrast to other players who had been in the league longer.

For our project we analzyed various datasets taken from the NBA [through SportRadar and Basketball Reference]. Therefore, we do not have any worries concerning the integrity and validity of our data. These datasets were curical for the understanding and comprehension of our questions. We had to analyze various datasets with lots of variables in order to answer the questions posed above.

At the conclusion of our project, some results were more surpising than others.

Through k-means, we created various plots looking at the clustering of different positions. We did this analysis to determine if changing the number of clusters gave insight. We initally chose five clusters because traditionally there are five NBA positions. What we saw from chaning the number of clisters is that there definately isnt more than five positons. Having six clusters does not appear to give any insight, there appears to be far too much

overlap between each cluster. Having four positons could be useful, but we see that one of the clusters (cluster 3) has far more players than any oyther cluster has, meaning that that cluster appears to be a combination of clusters 3 and 5 from the 5 cluster model. This is not a signifigant discovery though, because we could have already seen that those two clusters were very similar.

However, the current position titles do not fit what each player does in a game as players are more dynamic and are stepping outside the boundaries of their assigned roles.

From our K-clustering of the PCA's for the 2008-09 season, we found that 10 years ago, there used to be more of a presence of big men to have two clusters dedicated to them, whereas now we only have one clear cluster. Point Guard's ten years ago were more focused on assists, while Point Guards today are split evenly amongst assists and scoring three pointers.

Next, we plotted the distribution of players in each cluster for teams with over 50 wins in the 2019 NBA season. Eight teams accomplished this, including all four teams who appeared in the conference finals. What we found was that inside play was still important; traditional big-men are still utilized. This could mean that is not as useful as many think to have big-men who shoot threes and "stretch the floor" as some call it. These teams also have a fair amount of players in group five, possibly meaning that it is helpful to have bench players who shoot lots of threes.

We created a bar chart that compares the three pointers attempted during the 2008-09 season and 2018-19 season, and breaks them up by position. It is noteworthy to recognize that just over 50,000 three pointers were taken over this season. During the 2018-2019 season, over 30,000 more three pointers were attempted! This staggering number is a 60% increase from threes taken 10 years prior. Furthermore, more centers have been taking three, as well as power forwards. While the percentages are still low, it is a significant increase from what it was 10 years ago. It would be interesting to see 10 years from now if that percentage will increase even more.

These results showed us that although the game has changed in the favor of three point shooters, as seen in the bar graph above, some positions still hold certain value. In the graphs detailing distributions of positions amongst teams, we found that traditional big-men are being ulitized even though more Centers shoot threes.

From a summary table that detailed three pointers made between positions throughout the 2014-2015 season, we found that shooting guards and point guards do not need as much space to drain a three point shot, in comparison to power fowards and centers. The PG-SG position has a significantly higher distance than the rest; the three point line is roughly 23 feet from the basket, so stepping 2 feet further out is considered a 'deep shot'. This just reflects the capability of these "flex" guards and how dangerous they can be, even though a defender may think the shot is too far to make.

For recruitment of players, we found that as the seasons have gone by, there has been a shift of importance to how well the players shoot versus their ability to pass. Although we did see a weird discrepancy in 2019, there has been a strong upwards trend in True Shooting Percentage.

Finally, we wanted to look at if younger players are in different categories than the rest of the NBA players to see if we could predict what the future of the NBA would be. What we actually found is that there are more young players who fit into categories where they shoot less threes, instead of the beleif that the NBA is moving toward more threes. This probably means that younger players are worse at shooting them, and therefore shoot less threes than older players. This was our analysis of predicting future NBA positions.

Data Description and Exploration

The data was taken from NBA Stat data released by the league.(Provided by SportRadar). Values that had an N/A were changed to zero. There were very few of these(about 6-10). The indices that had a value of N/A were typically in the ThreePAr (3 Point Attempt Rate) colunm of the player. This makes sense since if a player was not a major 3 point shoooter, then he most likely did not make any 3s in game so assigning this variable a value of 0 did not mess up any clustering.

We noticed that there were many duplicate names of players but we decided to not delete them since a dupliate meant that a player had been traded or switched teams throughout the season. This means that that certain player was used for a different role or position in a new team so the stats and data for that player would be appropriate for his role in game and therefore, not make any cluster overlap.

Position Reference: PG - Point Guard - typically the team's best ball handler and passer SG - Shooting Guard - prolific from the three-point range and long mid range SF - Small Forward - versatile players who can dribble, shoot, and post up close to the basket PF - Power Forward - typically shoots close to the basket, taller than the first three positions C - Center - tallest players on the court, also shooting close to the basket and rebounding

Key Varaibles (The important ones to know): Pos - Position MP - Minutes Played BLK - Blocks Percentage ThreePAr - 3 Point Attempt Rate STL - Steal Percentage AST - Asist Percentage USG - Usage Percentage FTr- Free Throw Attempt Rate TRB - Total Rebound Percentage DRB - Defensive Rebound Percentage TOV - Turnover Percentage TRB - Total Rebound Percentage ORB - Offensive Rebound Percentage USG - Usage Percentage avg_shot_distance - average distance shot was made from avg_def_distance - average distance defender was from shooter FGM - Field Goal Made (0 means missed, 1 means made)

```
summary(nbaNow)
```

```
##      Name               Pos                 Age            Team
##  Length:304         Length:304         Min.   :19.00   Length:304
##  Class :character   Class :character   1st Qu.:23.00   Class :character
##  Mode  :character   Mode  :character   Median :26.00   Mode  :character
##                                        Mean   :26.44
##                                        3rd Qu.:29.00
##                                        Max.   :42.00
##     ThreePAr           FTr               ORB              DRB
##  Min.   :0.0000   Min.   :0.0590   Min.   : 0.700   Min.   : 5.90
##  1st Qu.:0.2617   1st Qu.:0.1678   1st Qu.: 2.000   1st Qu.:10.57
##  Median :0.3840   Median :0.2325   Median : 2.950   Median :13.55
##  Mean   :0.3726   Mean   :0.2490   Mean   : 4.437   Mean   :15.13
##  3rd Qu.:0.5162   3rd Qu.:0.3068   3rd Qu.: 5.325   3rd Qu.:18.15
##  Max.   :0.8300   Max.   :0.7330   Max.   :16.800   Max.   :35.90
##       TRB              AST              STL              BLK
##  Min.   : 3.700   Min.   : 4.000   Min.   :0.300   Min.   : 0.000
##  1st Qu.: 6.300   1st Qu.: 8.175   1st Qu.:1.100   1st Qu.: 0.700
##  Median : 8.400   Median :11.750   Median :1.400   Median : 1.250
##  Mean   : 9.787   Mean   :14.896   Mean   :1.512   Mean   : 1.687
##  3rd Qu.:12.025   3rd Qu.:19.025   3rd Qu.:1.800   3rd Qu.: 2.000
##  Max.   :25.900   Max.   :46.500   Max.   :3.100   Max.   :10.000
##       TOV              USG
##  Min.   : 3.70   Min.   : 9.50
##  1st Qu.: 9.70   1st Qu.:15.80
##  Median :11.40   Median :18.70
##  Mean   :11.79   Mean   :19.76
##  3rd Qu.:13.60   3rd Qu.:22.95
##  Max.   :26.80   Max.   :40.50
```

```
head(NBA_Player_Stats_18_to_19)
```

```
## # A tibble: 6 x 28
##      Rk First Last  Pos     Age Team  Games Minutes   PER    TS ThreePAr   FTr
##   <dbl> <chr> <chr> <chr> <dbl> <chr> <dbl>   <dbl> <dbl> <dbl>    <dbl> <dbl>
## 1     4 Stev… "Ada… C        25 OKC      80    2669  18.5 0.591    0.002 0.361
## 2     5 Bam   "Ade… C        21 MIA      82    1913  17.9 0.623    0.031 0.465
## 3     8 LaMa… "Ald… C        33 SAS      81    2687  22.9 0.576    0.032 0.312
## 4    11 Jarr… "All… C        20 BRK      80    2096  18.5 0.632    0.079 0.489
## 5    13 Al-F… "Ami… PF       28 POR      81    2292  13.2 0.568    0.472 0.292
## 6    15 Kyle  "And… SF       25 MEM      43    1281  12.8 0.569    0.123 0.232
## # … with 16 more variables: ORB <dbl>, DRB <dbl>, TRB <dbl>, AST <dbl>,
## #   STL <dbl>, BLK <dbl>, TOV <dbl>, USG <dbl>, OWS <dbl>, DWS <dbl>, WS <dbl>,
## #   `WS/48` <dbl>, OBPM <dbl>, DBPM <dbl>, BPM <dbl>, VORP <dbl>
```

A preview of the data we utilized in the project.

```
head(draft_by_year)
```

```
## # A tibble: 6 x 5
##   drafted count avg_shooting assists_pct dreb_pct
##     <dbl> <int>        <dbl>       <dbl>    <dbl>
## 1    1983    10        0.500       0.159    0.107
## 2    1985    21        0.492       0.118    0.150
## 3    1986    18        0.500       0.149    0.142
## 4    1987    22        0.454       0.139    0.146
## 5    1988    29        0.490       0.130    0.138
## 6    1989    23        0.498       0.158    0.131
```

Here, data to see how the percentage of Assists, shooting, and defensive Rebounds were changed throughout the years of the drafting. With this, we can see what the NBA recruiters wanted when scouting for players and how that has changed over the draft years.

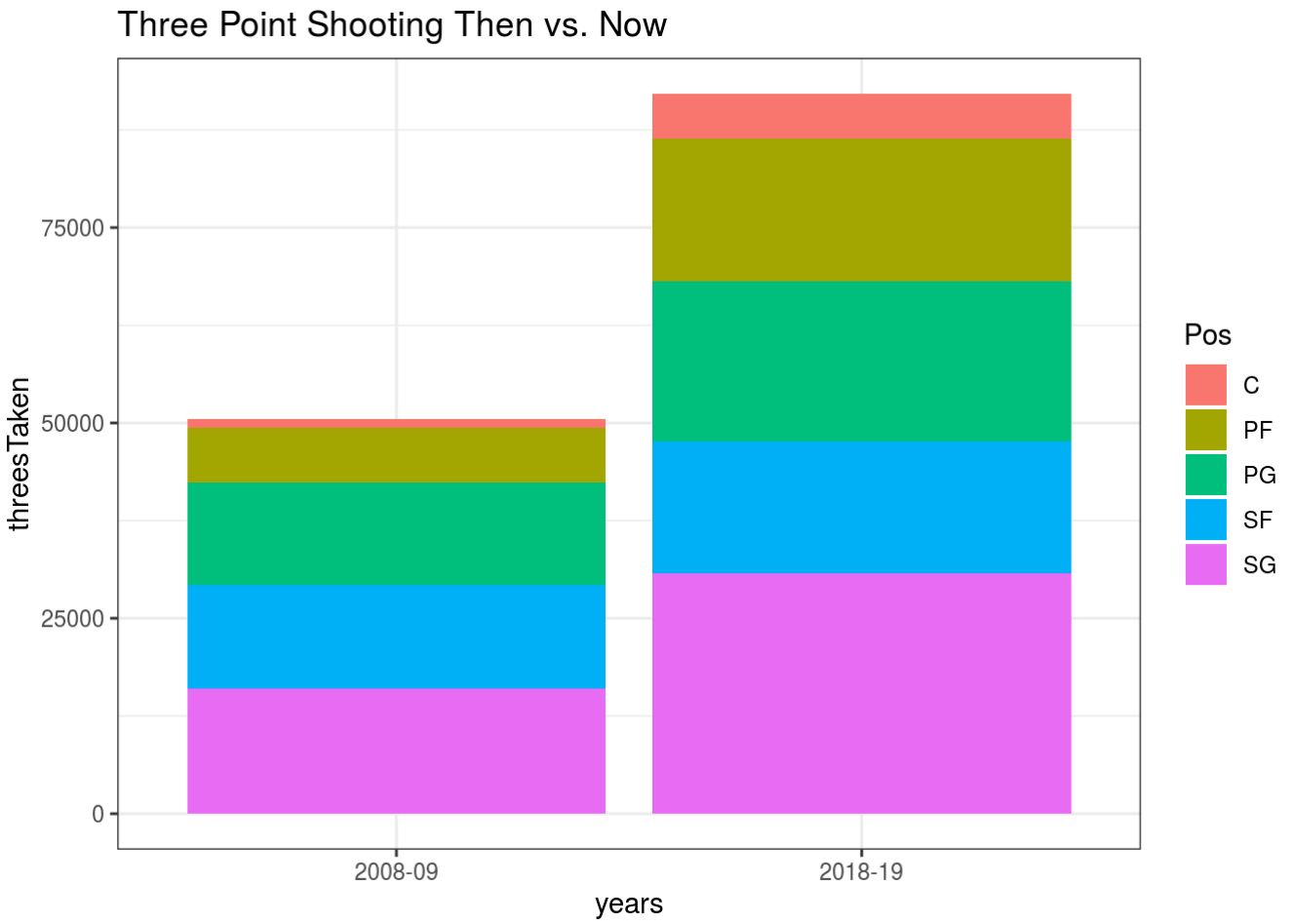
```
head(threesMade, 1)
```

```
##           Name Rk Pos Age Team Games Minutes  PER    TS ThreePAr   FTr ORB DRB
## 1 Aaron Brooks 63  PG  30  CHI    82    1885 14.4 0.534    0.383 0.213 1.9 7.5
##   TRB  AST STL BLK  TOV USG   IDK OWS DWS  WS WS/48   idk OBPM DBPM  BPM VORP
## 1 4.8 24.2 1.5 0.6 14.9  25 FALSE 1.7 1.5 3.3 0.083 FALSE  0.3 -1.1 -0.8  0.5
##    GAME_ID                  MATCHUP LOCATION W FINAL_MARGIN SHOT_NUMBER PERIOD
## 1 21400591 JAN 16, 2015 - CHI @ BOS        A W           16           6      2
##   GAME_CLOCK SHOT_CLOCK DRIBBLES TOUCH_TIME SHOT_DIST PTS_TYPE SHOT_RESULT
## 1   09:06:00        3.6        0          1      22.8        3        made
##   CLOSEST_DEFENDER CLOSEST_DEFENDER_PLAYER_ID CLOSE_DEF_DIST FGM PTS
## 1    Pressey, Phil                     203515            4.9   1   3
##    player_name player_id
## 1 aaron brooks    201166
```

When looking at the 2014-2015 shot data, two datasets were merged. The shot log data along with the seasons stats by the name of the player. This helps us then look at which positions made what type of shots (2s or 3s) the most and how far they were from a defender. We can also use this to see how certain positions shoot the ball depending on the distance they are from the defender.

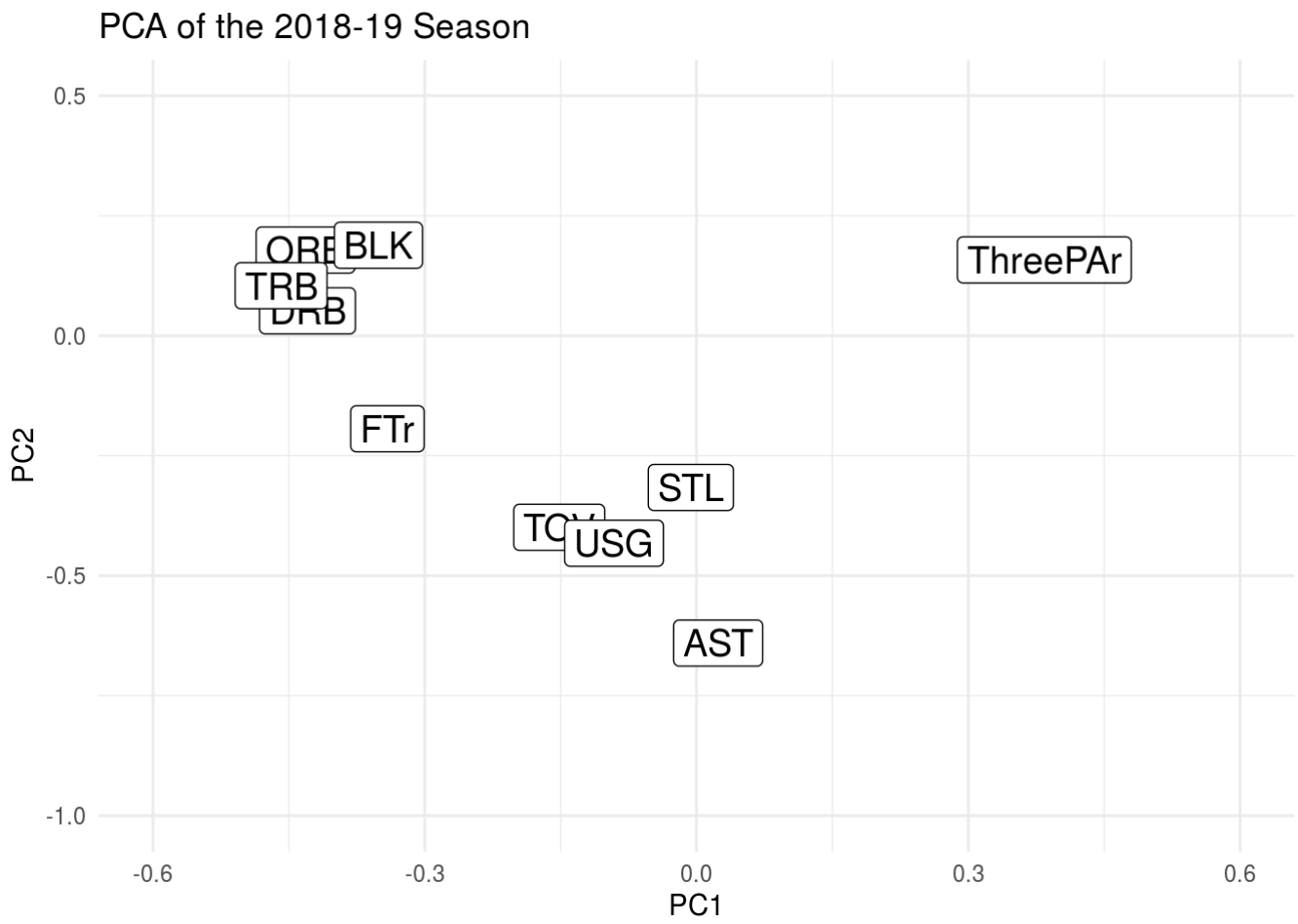
Results

```
ggplot(threesComparison, aes(x = years, y = threesTaken, fill = Pos)) +
  geom_bar(stat = 'identity') + theme_bw() + ggtitle("Three Point Shooting Then vs. Now"
)
```



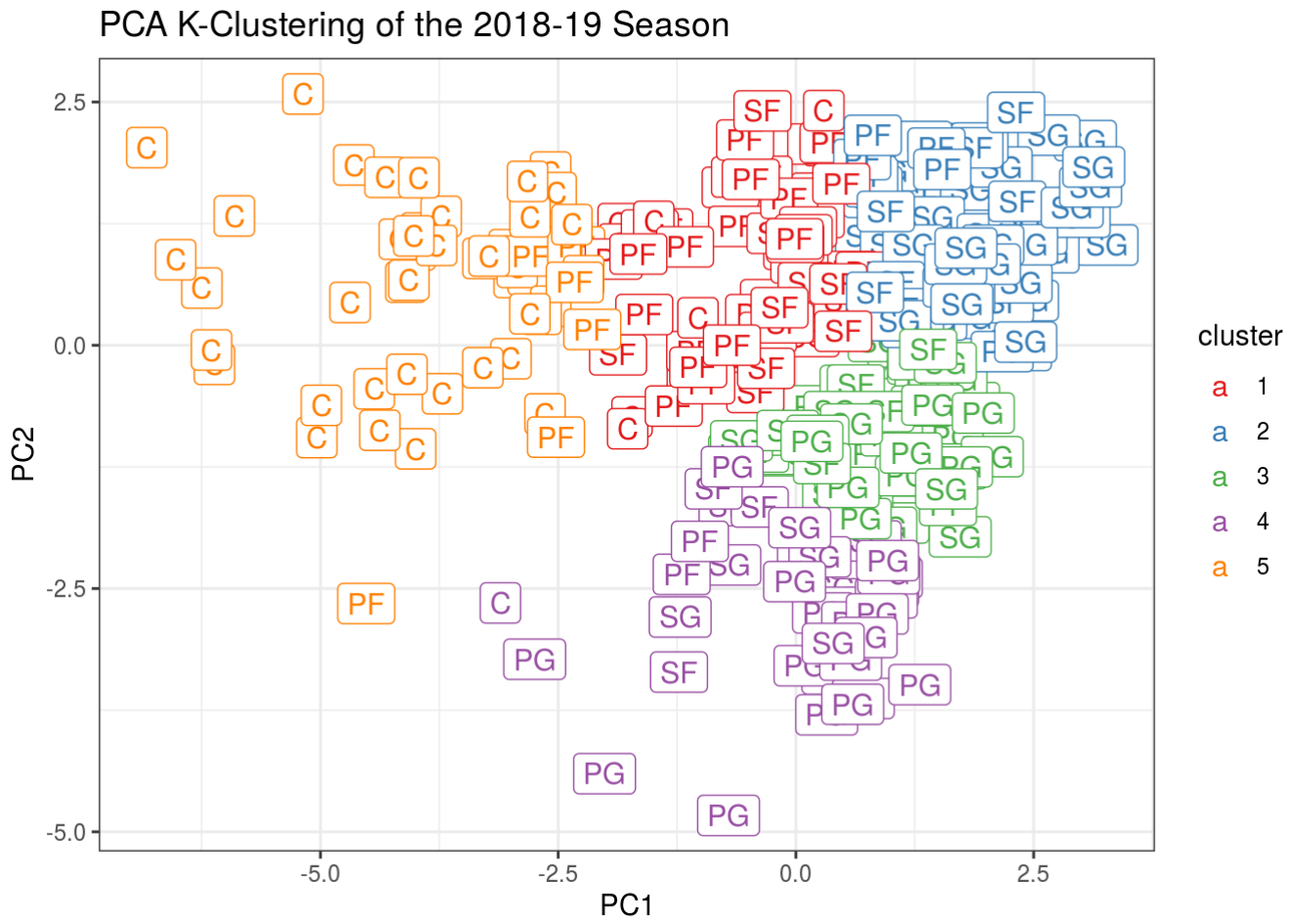
This stacked bar chart compares the three pointers attempted during the 2008-09 season and 2018-19 season, and breaks them up by position. The bar graph for 2008-2009 season illustrates an almost even split between shooting guards, point guards and small forwards. Centers has the smallest percentage of the three pointers attempted during this season. It is also noteworthy to recognize that just over 50,000 three pointers were taken over this season. During the 2018-2019 season, over 30,000 more three pointers were attempted! This staggering number is a 60% increase from threes taken 10 years prior. Furthermore, more centers have been taking three, as well as power forwards. While the percentages are still low, it is a significant increase from what it was 10 years ago. It would be interesting to see 10 years from now if that percentage will increase even more.

```
ggplot(data.frame(pcaNew$rotation), aes(x = PC1, y = PC2, label = rownames(pcaNew$rotati
on))) +
  geom_point(colour = "purple3") + geom_label(size = 5) + theme_minimal() +
  xlim(-.6, .6) + ylim(-1, .5) + ggtitle("PCA of the 2018-19 Season")
```



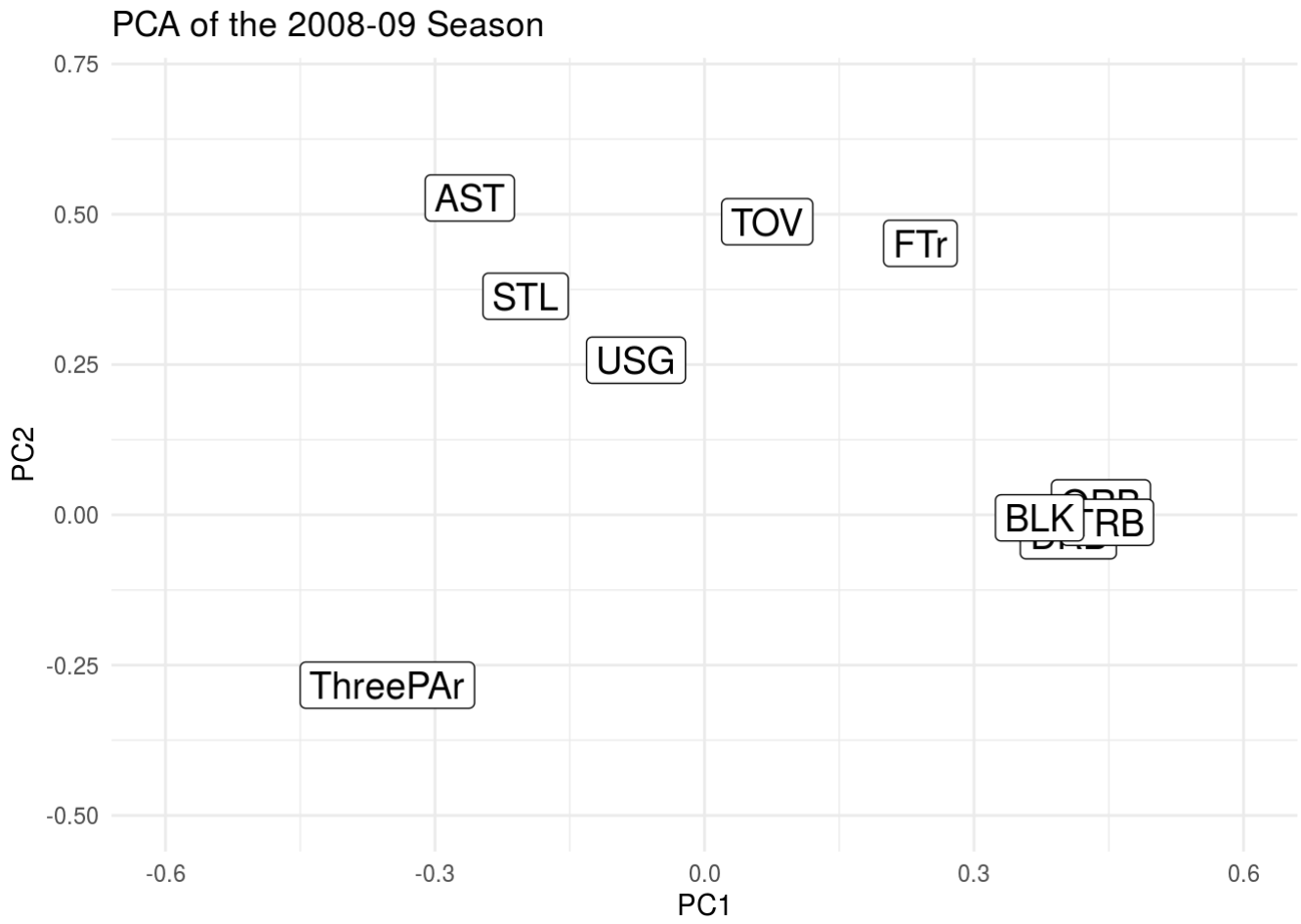
This is a PCA plot of the 2018-2019 season, showing the first and second principal components. We can see that ThreePAr had a significant holding value for the first principal compenent, showing how the most variation was connected to the amount of three's pointers players were taking. We can see a negative correlation with indicators of ORB, DRB, TRB, and BLK; it makes sense that they are clustered together as generally, players that can rebound effectively are those who are more aggressive on defense, making them effective at blocking shots. When looking at the second principal component, one can see the magnitude that assists percentage hold, nearly correlating with the usage percentage and steals.

```
pcaKmeansNew <- kmeans(pcaNew$x[, c(1,2)], 5)
as_tibble(pcaNew$x) %>% mutate(cluster = factor(pcaKmeansNew$cluster), Positions = posit
ions) %>%
  ggplot(aes(PC1, PC2, color = cluster, label = positions)) + geom_label(size = 4) +
  theme_bw() + scale_color_brewer(type = "qual", palette = "Set1") + ggtitle("PCA K-Clus
tering of the 2018-19 Season")
```



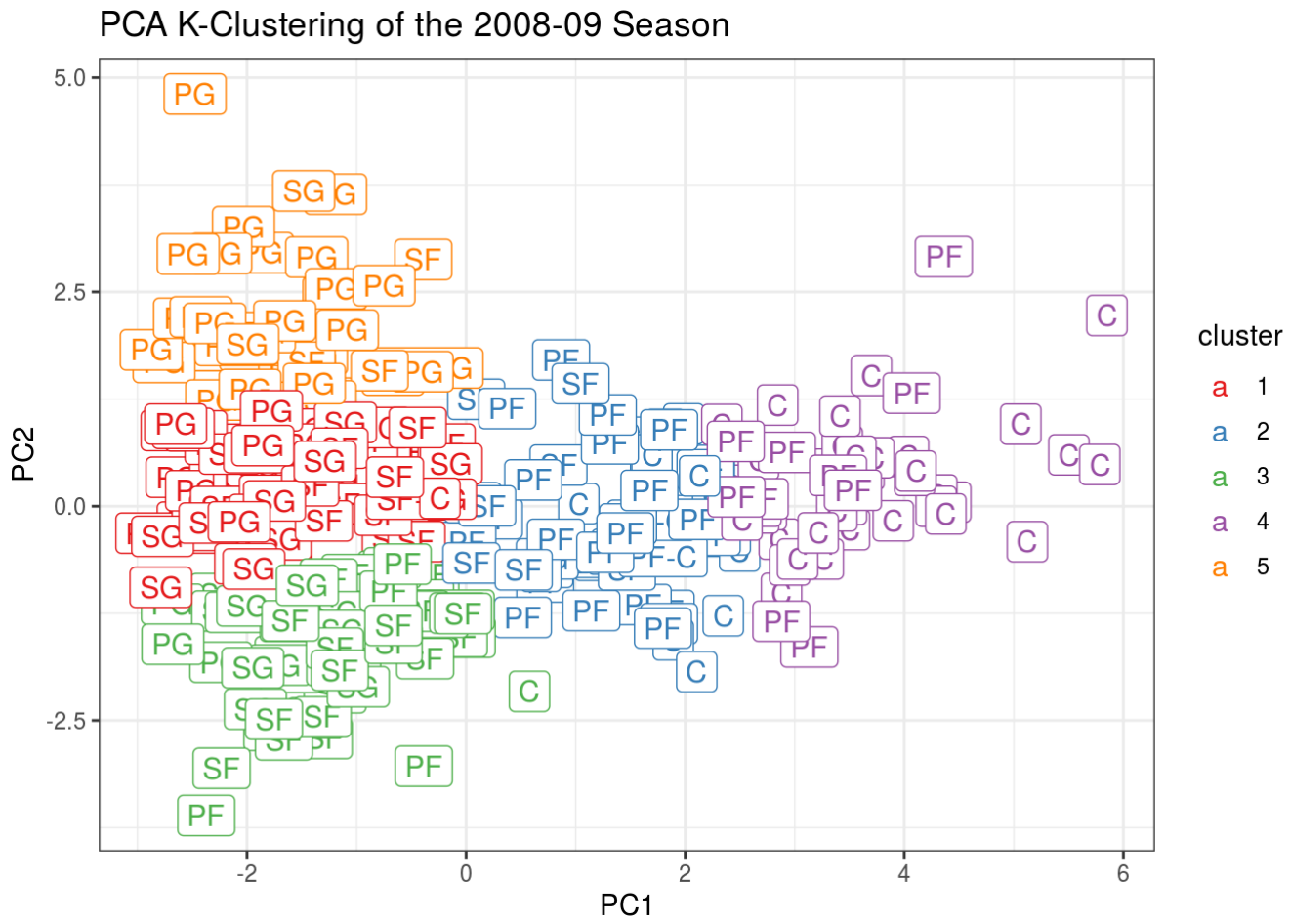
This plot depicts the K-clustering of the PCA's, this time showing where each player (the labels here show thier position) falls in terms of these principal components. When looking at this in conjunction with the previous plot, one can see that mainly centers (C's) and some power forwards (PF's) are in charge of the rebounding and blocks (top-left), and that we don't see many "C's" and "PF's" on the top right, which is surprising when one believes that these big men are taking more three pointers. When looking at cluster 5, which corresponds to the Assist percentage as it's near the bottom, we see many point guards but also glimpses of PF's and C's. The third and fourth clusters seem to be intermediary between those who can shoot and pass (mainly PG's, SG's, SF's), or those who can shoot and rebound (PF's and some C's). The 3rd cluster here is what would mainly define those big men who are taking more three point shots.

```
ggplot(data.frame(pcaOld$rotation), aes(x = PC1, y = PC2, label = rownames(pcaOld$rotati
on))) +
  geom_point(colour = "purple3") + geom_label(size = 5) + theme_minimal() +
  xlim(-.6, .6) + ylim(-.5, .7) + ggtitle("PCA of the 2008-09 Season")
```



This is a PCA plot of the 2008-2009 season, showing the first and second principal components. We can see that ThreePAr had a significant negative holding value for both the first and second principal compenents, showing how the most variation overall was connected to the amount of three's pointers players were taking. As did with the 2018-19 PCA, blocks and rebounds were clumped together, and this time we see it having a major positive effect on the first principal component, meaning these statistics are what varied the players the most. Assist percentage, steal percentage, and usage percentage were slightly moderate in the first PC, where as in the second PC, they hold a positive holding.

```
pcaKmeansOld <- kmeans(pcaOld$x[, c(1,2)], 5)
as_tibble(pcaOld$x) %>% mutate(cluster = factor(pcaKmeansOld$cluster), Positions = posit
ions2) %>%
  ggplot(aes(PC1, PC2, color = cluster, label = positions2)) + geom_label(size = 4) +
  theme_bw() + scale_color_brewer(type = "qual", palette = "Set1") + ggtitle("PCA K-Clus
tering of the 2008-09 Season")
```



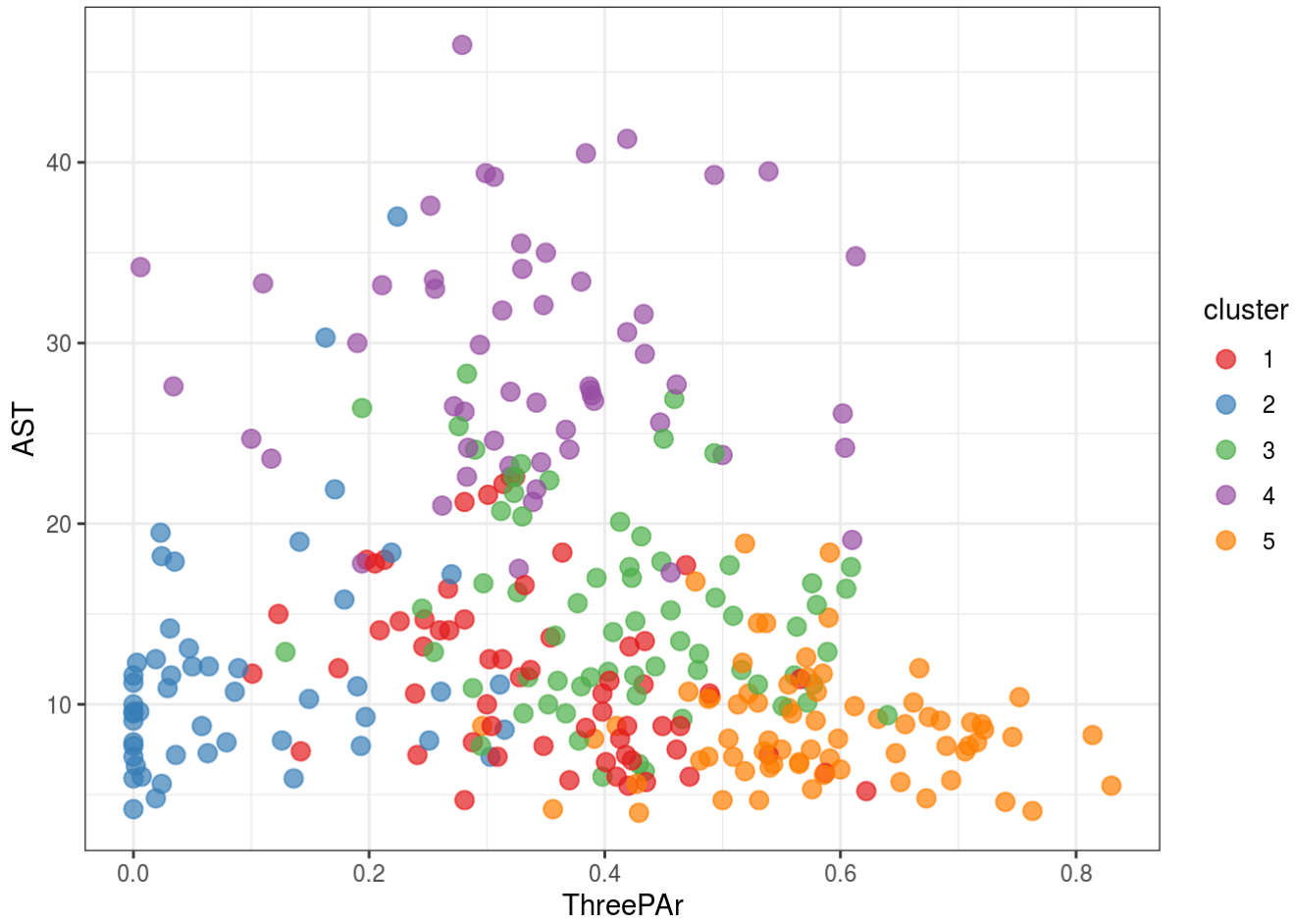
This plot depicts the K-clustering of the PCA's, this time for the 2008-09 season. Again, we see the centers and PF's are dominant in the side where rebounds and blocks were on the previous plot, this time on the right hand side. Not only do we see a cluster (blue) of some PF's that are on the 3 point shooting side of the graph, but we another PF/C dominant cluster (orange), where in the previous, we had only one such cluster. This shows how 10 years ago, there used to be more of a presence of big men to have two clusters dedicated to them, whereas now we only have one clear cluster, as shown in the previous plot. We see in the bottom left (where ThreePAr would be) a major presence of shooting guards and small forwards, where now we had PG's too in that area. PG's ten years ago were more focused on assists, which is why we see a good amount of them on the top left.

```
#trying to define 5 positions in 18-19 season with 5 k-clusters
K <- 5
set.seed(1305)
m <- kmeans(scale(data.matrix(nbaNow[,5:14])), centers = K)
df_clusters <- mutate(nbaNow, cluster = factor(m$cluster))

group_by(df_clusters, cluster) %>%
  summarise(n = n(),
            ThreePAr = mean(ThreePAr),
            FTr = mean(FTr),
            ORB = mean(ORB),
            DRB = mean(DRB),
            TRB = mean(TRB),
            AST = mean(AST),
            STL = mean(STL),
            BLK = mean(BLK),
            TOV = mean(TOV),
            USG = mean(USG))
```

```
## # A tibble: 5 x 12
##   cluster     n ThreePAr   FTr   ORB   DRB   TRB   AST   STL   BLK   TOV   USG
##   <fct>   <int>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 1          60   0.345  0.265  4.83  18.0 11.4  11.5   1.77 1.88   11.3  20.0
## 2 2          51   0.0907 0.376 11.1   24.4 17.8  11.6   1.33 3.80   12.8  20.4
## 3 3          67   0.416  0.207  2.11  10.5  6.27 15.2   1.33 0.761  10.6  20.7
## 4 4          55   0.340  0.278  2.61  13.1  7.9  29.1   1.85 1.08   15.2  24.3
## 5 5          71   0.583  0.161  2.90  12.0  7.48  8.78  1.33 1.34   10.0  14.7
```

After doing this K-means clustering, I get five clusters. Here are general descriptions of the types of players they describe: Cluster 1 - Three point shooting big men and forwards. Example: DeMarre Carroll Cluster 2 - Traditional big men who shoot very few threes. Example: Steven Adams Cluster 3 - Traditional Wings such as Andrew Wiggins. This category also has some backup point guards who play less with the ball in their hands because they are not starters, such as Quinn Cook. Cluster 4 - This cluster is generally players who shoot less threes than other guards or wings, but instead players who have very high usage rates and very high rates of assists. Example: LeBron James Cluster 5 - These players have very high rates of threes, but low usage rates and therefore lower rates of things like rebounds and assists. There are a lot of backup players in this category(although not all backup players), some might call it a “3 and D” category as these players come on to shoot threes and play defence, but not much else. Example: Danny Green

```
cent <- as_tibble(m$centers) %>% mutate(cluster = factor(1:K))
ggplot(df_clusters, aes(ThreePAr, AST, color = cluster)) + geom_point(size = 3, alpha =
.7) +
  theme_bw() + scale_color_brewer(type = "qual", palette = "Set1")
```



Two of the most imporant differentiaters appeared to be assists and threes, which as you can see allow for groups with some overlap on this graph.

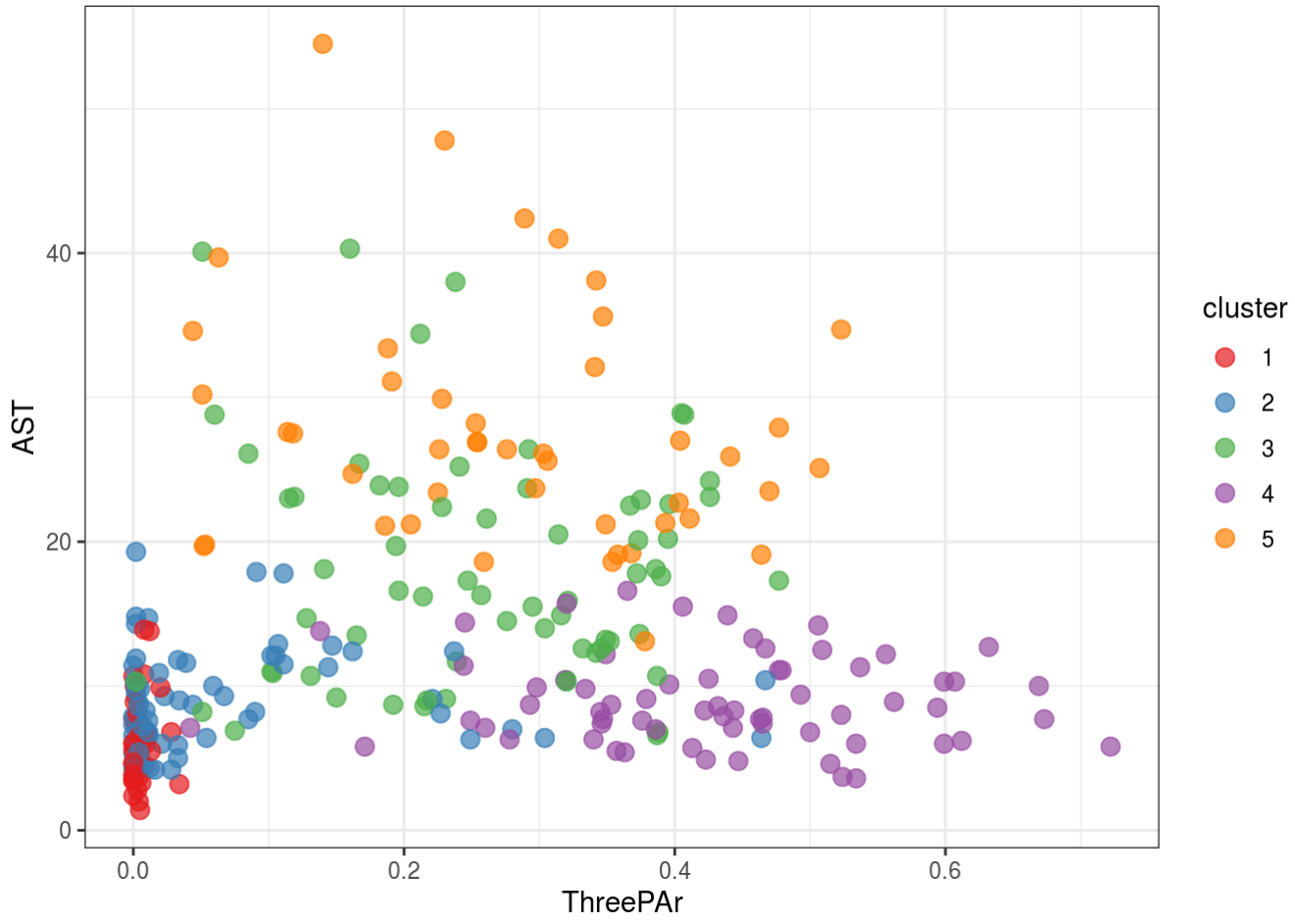
```
#trying to define 5 positions in 08-09 season with 5 k-clusters
K <- 5
set.seed(1305)
mOld <- kmeans(scale(data.matrix(nbaOld[,5:14])), centers = K)
df_clustersOld <- mutate(nbaOld, cluster = factor(mOld$cluster))

group_by(df_clustersOld, cluster) %>%
  summarise(n = n(),
            ThreePAr = mean(ThreePAr),
            FTr = mean(FTr),
            ORB = mean(ORB),
            DRB = mean(DRB),
            TRB = mean(TRB),
            AST = mean(AST),
            STL = mean(STL),
            BLK = mean(BLK),
            TOV = mean(TOV),
            USG = mean(USG))
```

```
## # A tibble: 5 x 12
##   cluster     n ThreePAr    FTr   ORB   DRB   TRB   AST   STL   BLK   TOV   USG
##   <fct>   <int>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 1          36  0.00542 0.484 11.8   21.6 16.7   6.4   1.24 4.15   16.1  16.3
## 2 2          63  0.0690  0.320  8.42  19.9 14.2   9.11  1.31 2.15   11.7  20.5
## 3 3          66  0.259   0.339  2.88  11.2  7.06 18.1   1.81 0.814  11.9  24.1
## 4 4          65  0.429   0.185  3.26  12.4  7.84  9.02  1.49 0.906  10.2  16.9
## 5 5          45  0.280   0.273  2.69  10.4  6.51 27.6   1.94 0.538  17.3  18.8
```

This k-means clustering shows much more distinct clusters than before, and the clisters correlate very well to the 5 traditional NBA positions. Cluster 1 - Centers Cluster 2 - Power Forwards Cluster 3 - Players who arent big men but who also do not shoot a lot of threes. Mostly shooting guards and small forwards, but even some point guards (Derrick Rose). Cluster 4 - Mostly Shooting Guards and some Small Forwards who shoot lots of threes; wings. Cluster 5 - Point Guards

This discrepancy in the clusters shows that, more than anything, the boundaries between positions are disappearing. They used to be much more distinct, but now many players are more hybrids than in past years.

```
cent <- as_tibble(mOld$centers) %>% mutate(cluster = factor(1:K))
ggplot(df_clustersOld, aes(ThreePAr, AST, color = cluster)) + geom_point(size = 3, alpha
= .7) +
  theme_bw() + scale_color_brewer(type = "qual", palette = "Set1")
```



The biggest conclution that can be drawn from comparing this graph to the one above is that many more players used to never shoot threes. In this graph, similar to the one above, groups can be seen with some overlap.

```
m$tot.withinss; mOld$tot.withinss
```

```
## [1] 1445.646
```

```
## [1] 1230.037
```

Comparing the 5 K-clustering for the 2018-19 season vs. the 2008-09 season, one can see that there was less variation in the 2008-09 clusters, meaning that the playstyles were more well defined than they are today. Looking at the lower variation in conjunction with the plot for the 2008-09 season, a big part of why the variation is lower is because the centers were sticking to their rebounding roles. They weren't involved with passing or three point shooting, but now, with players like Nikola Jokic, the positions are not as well defined as they used to be.

```
m$betweenss; mOld$betweenss
```

```
## [1] 1584.354
```

```
## [1] 1509.963
```

The betweenss variable represents the between cluster variation, which takes all the data points' sum of squares (totss) and subtracts the within cluster sum of squares (withinss), to see how much varied across clusters. The overall between cluster variation is slightly different, with that for the old season being slightly lower, meaning that in this case, the 2018-2019 season did have more variation across players to begin with, reflecting how different players' playstyles are in the modern NBA.

```
filter(df_clusters) %>%
  count(cluster)
```

```
## # A tibble: 5 x 2
##   cluster     n
##   <fct>   <int>
## 1 1          60
## 2 2          51
## 3 3          67
## 4 4          55
## 5 5          71
```

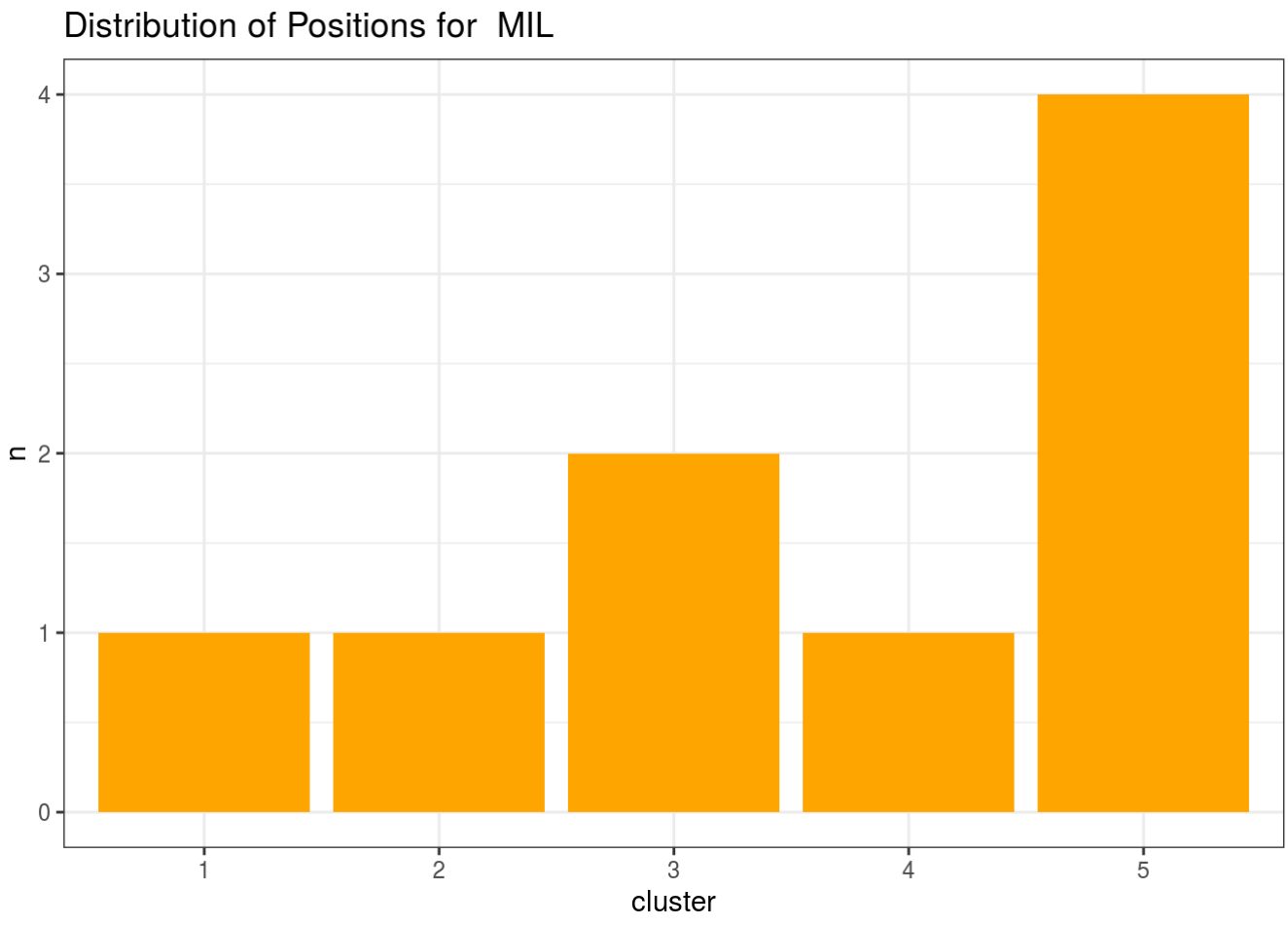
```
filter(df_clusters, Age <= 22) %>%
  count(cluster)
```

```
## # A tibble: 5 x 2
##   cluster     n
##   <fct>   <int>
## 1 1          10
## 2 2          14
## 3 3           8
## 4 4          12
## 5 5          10
```

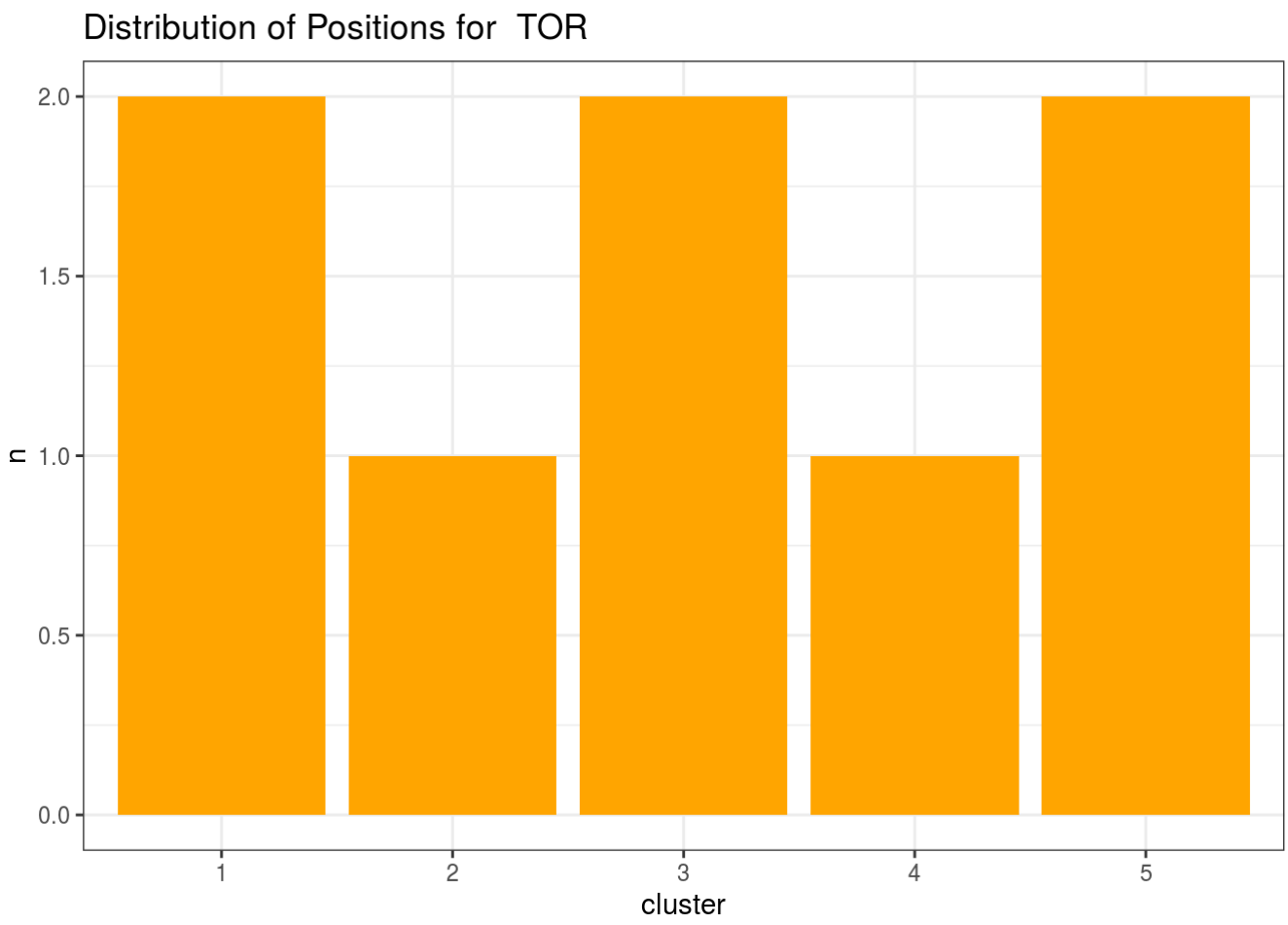
We wanted to look at if younger players are in different categories than the rest of the NBA players to see if we could predict what the future of the NBA would be. What we actually founnd is that there are more young players who fit into categories where they shoot less threes, instead of the beleif that the NBA is moving toward more threes. This probably means that younger players are worse at shooting them, and therefore shoot less threes than older players.

```
generatePlot <- function(team){
  d <- filter(df_clusters, Team == team) %>% count(cluster, Team)
  ggplot(d, aes(x = cluster, y = n)) +
  geom_bar(stat = 'identity', fill = 'orange') + theme_bw() +
    ggtitle(paste("Distribution of Positions for ", team))
}
```

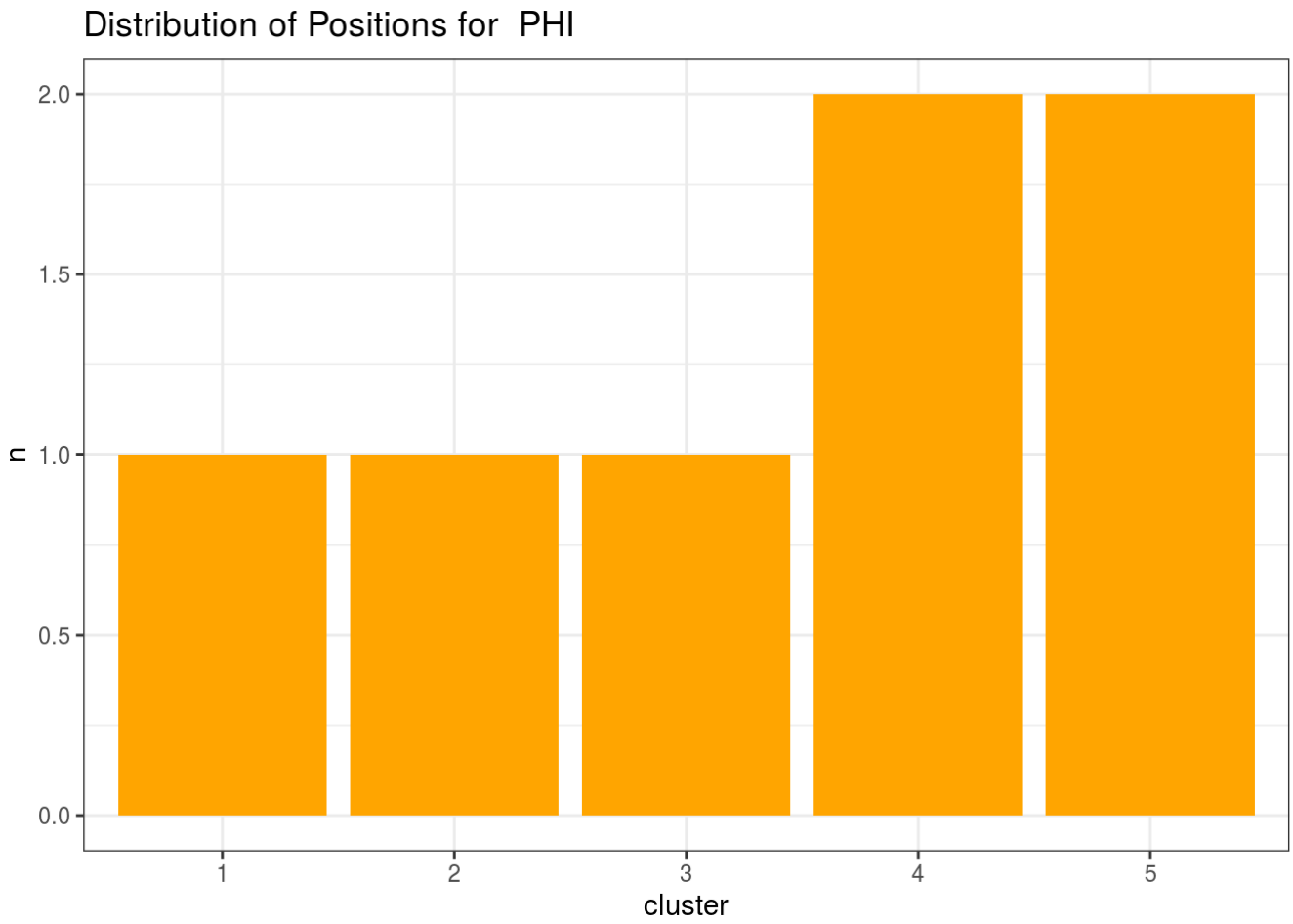
```
generatePlot("MIL")
```



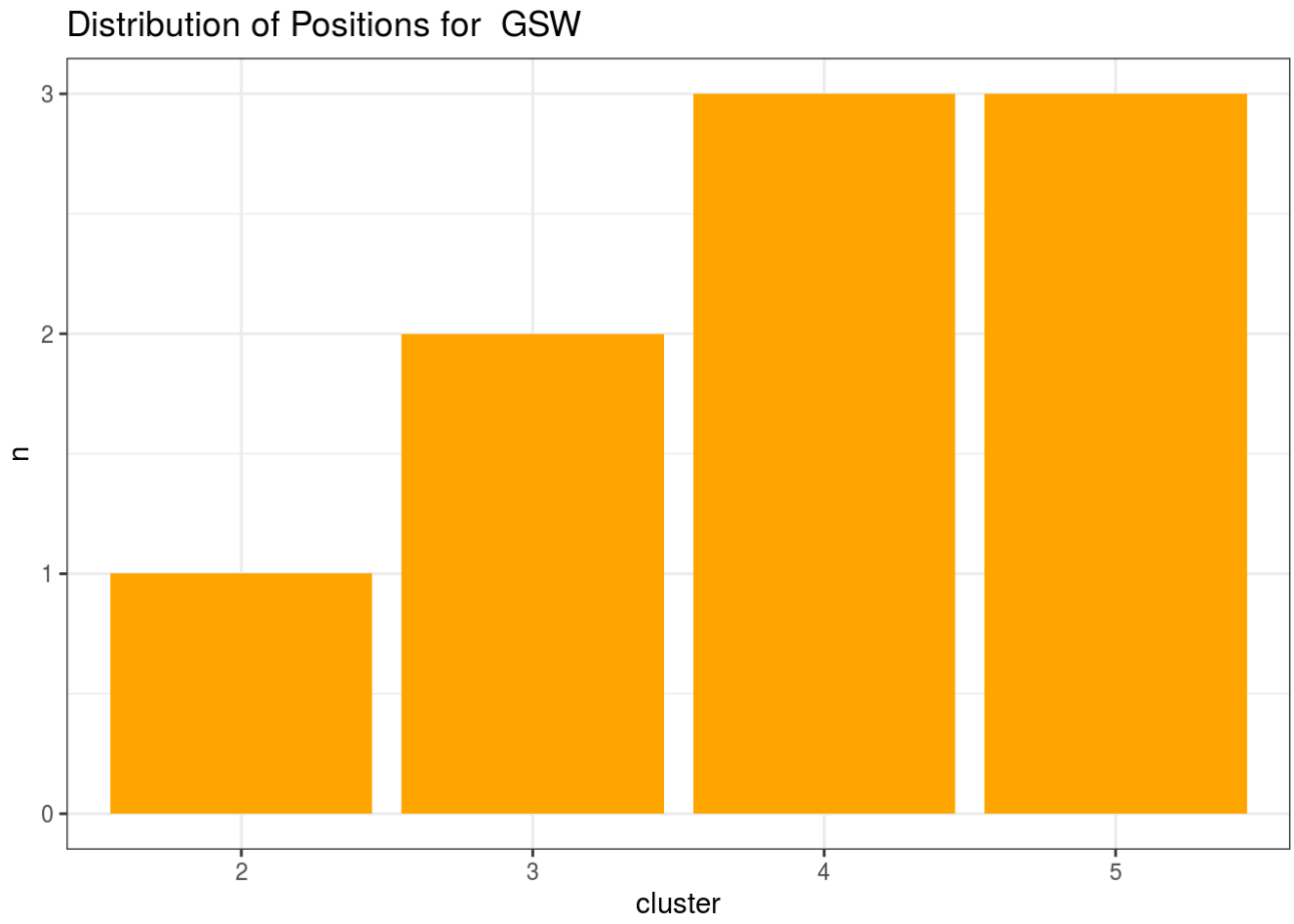
```
generatePlot("TOR")
```



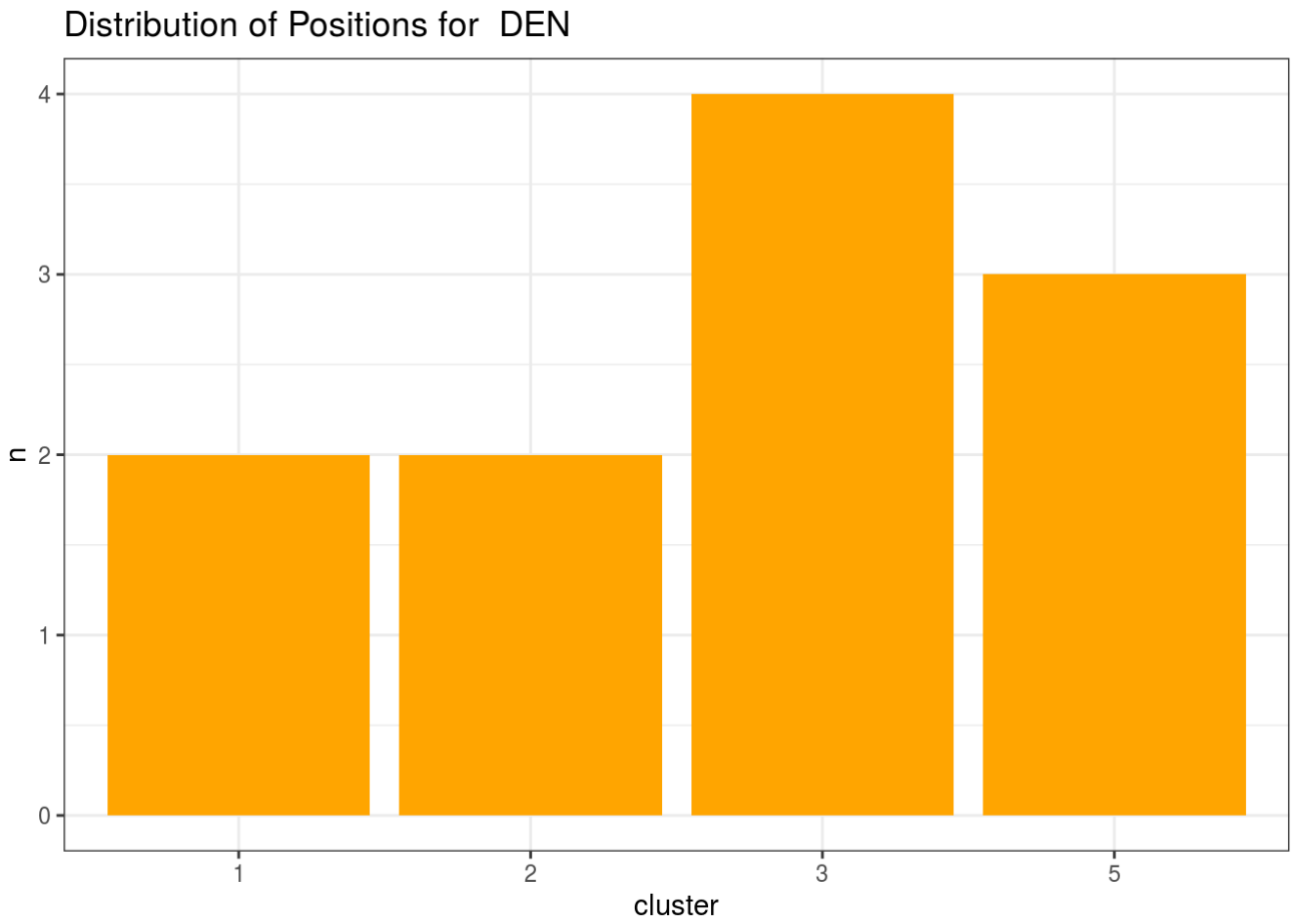

```
generatePlot("PHI")
```



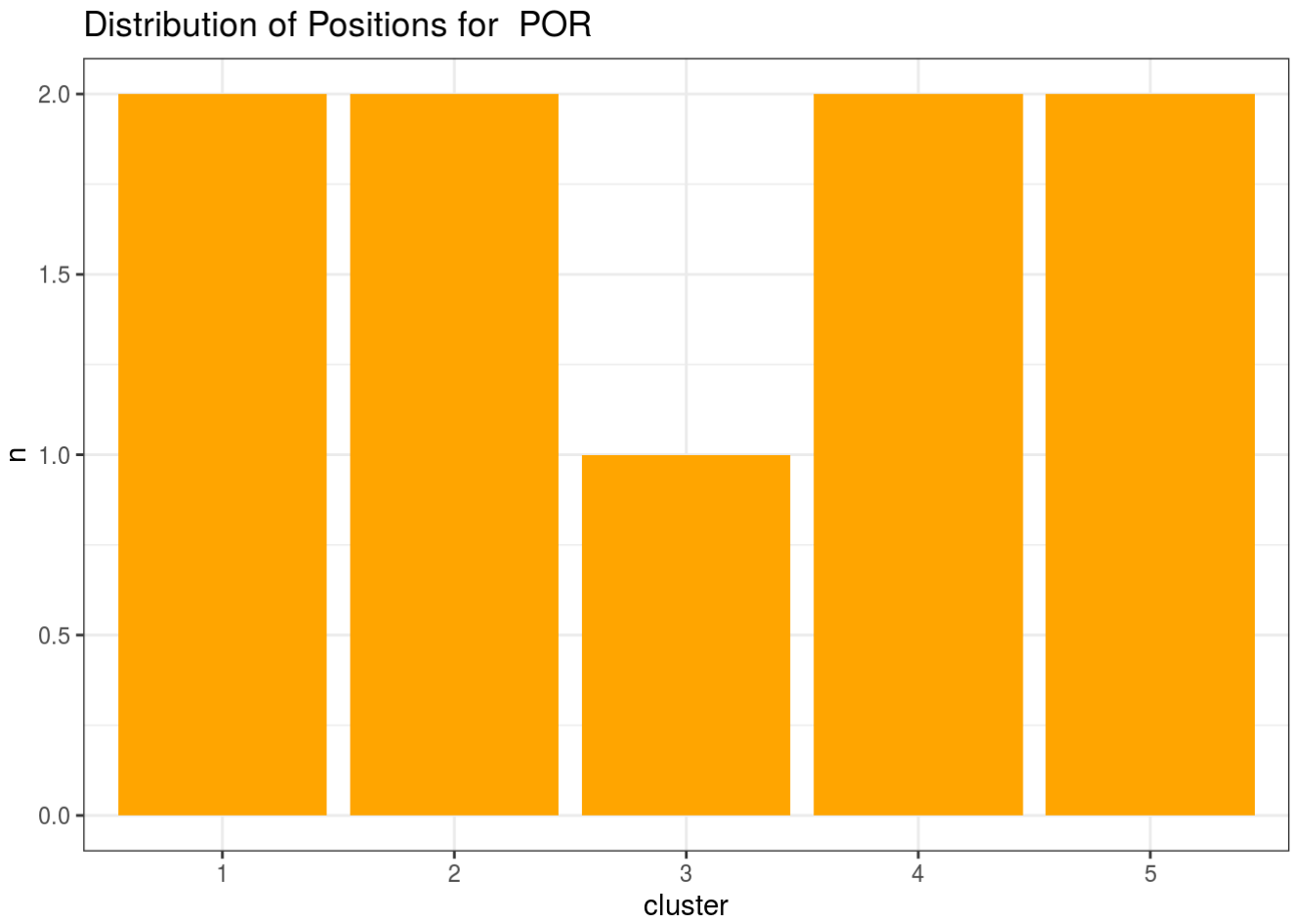

```
generatePlot("GSW")
```



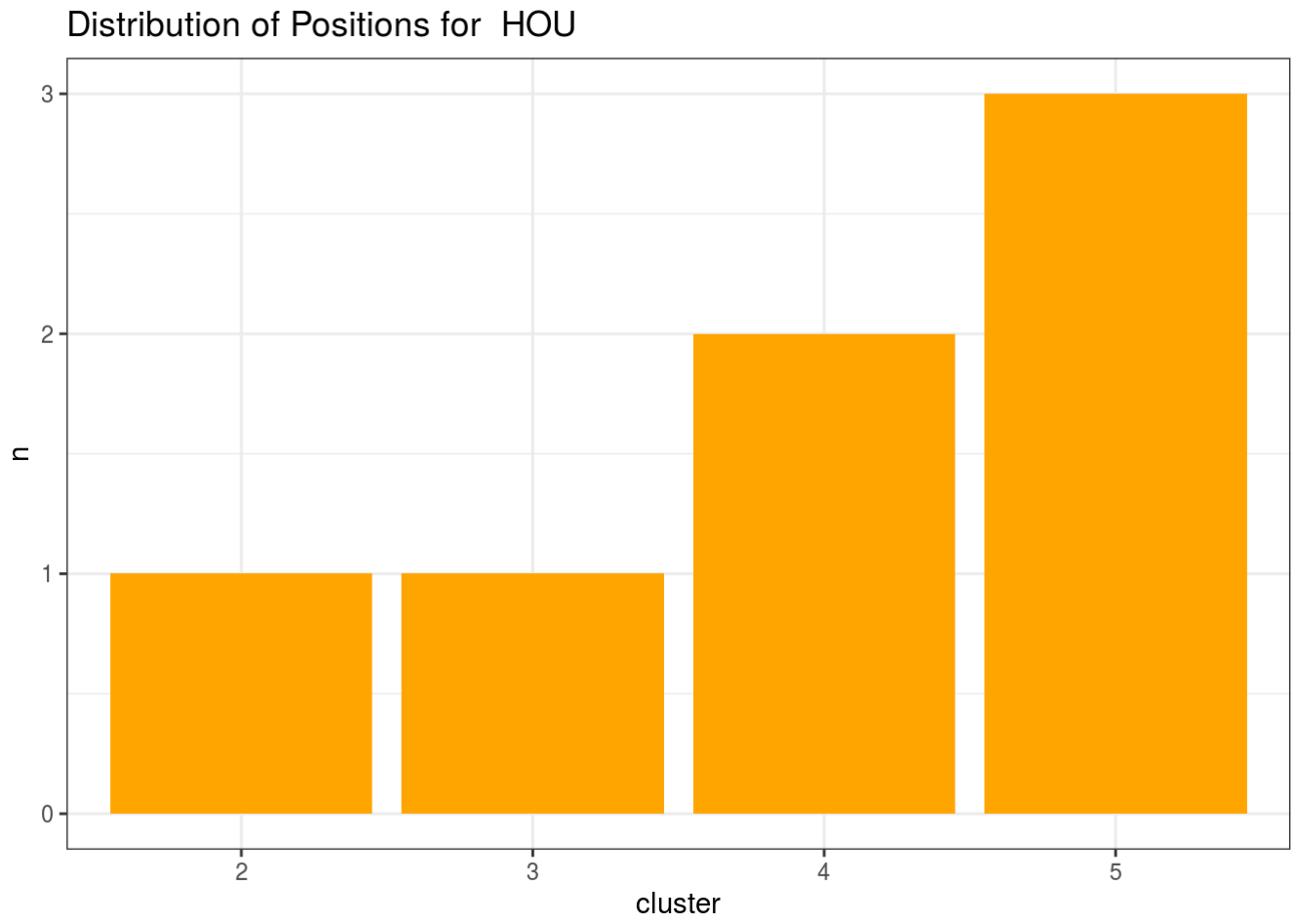

```
generatePlot("DEN")
```



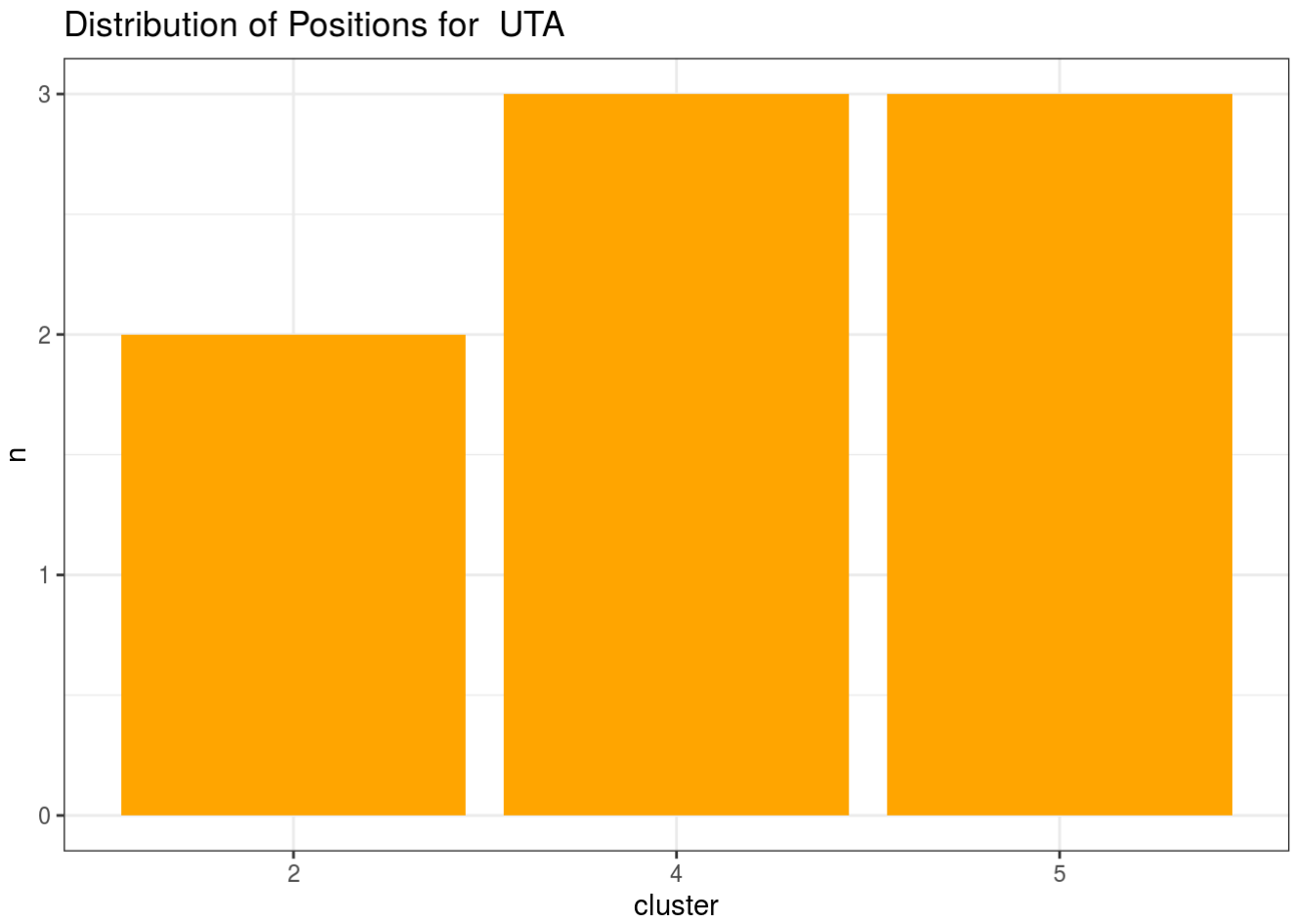

```
generatePlot("POR")
```




```
generatePlot("HOU")
```




```
generatePlot("UTA")
```

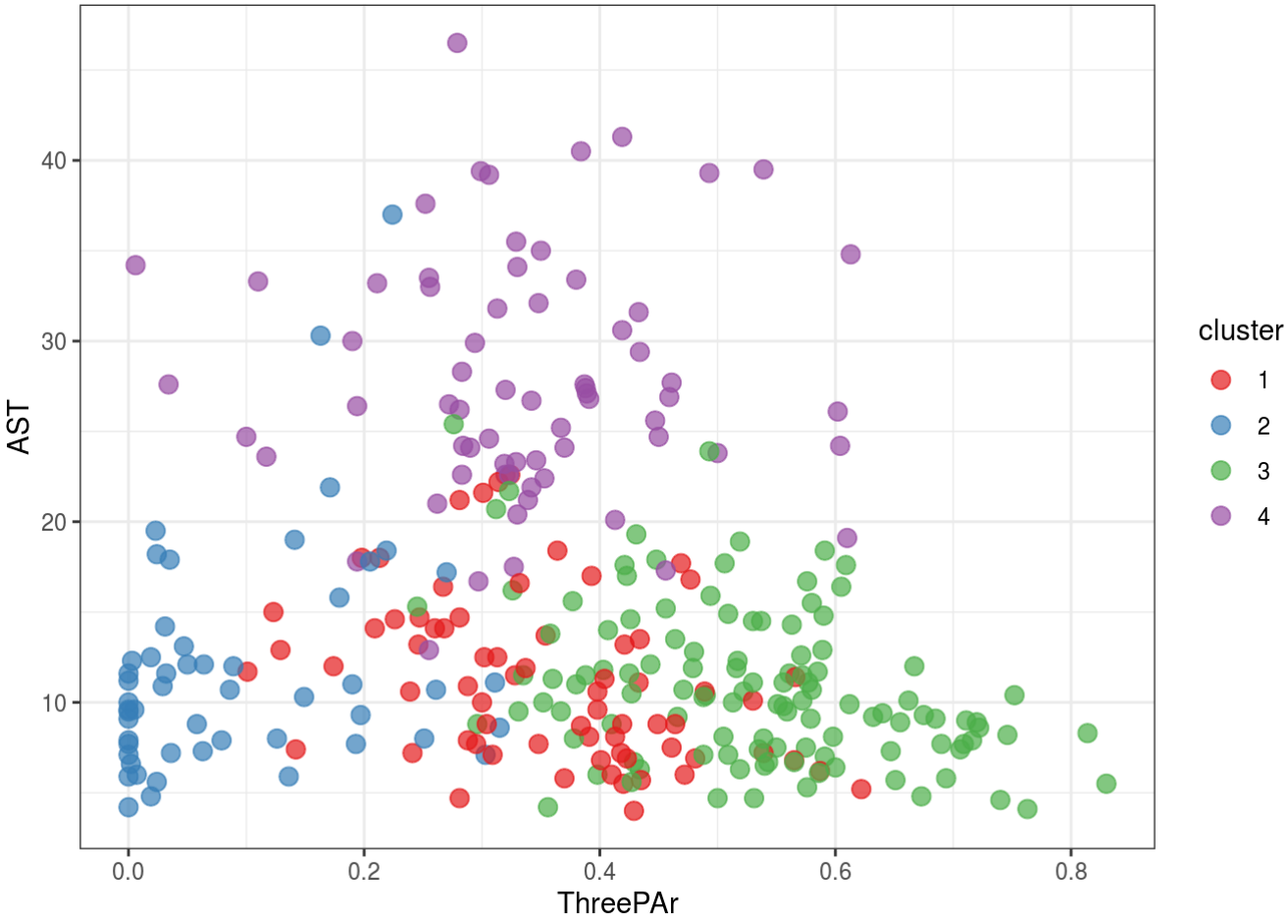


In this section, we wanted to see what the better NBA players were doing, so we decided to plot the distribution of players in each cluster for teams with over 50 wins in the 2019 NBA season. Eight teams accomplished this, including all four teams who appeared in the conference finals. What we found was that inside play was still important; traditional big-men are still utilized. This could mean that is isnt as useful as many think to have big-men who shoot threes and “stretch the floor” as some call it. These teams also have a fair amount of players in group five, possibly meaning that it is helpful to have bench players who shoot lots of threes.

```
#trying to define 4 positions with 4 k-clusters
K <- 4
set.seed(1305)
m4 <- kmeans(scale(data.matrix(nbaNow[,5:14])), centers = K)
df_clusters4 <- mutate(nbaNow, cluster = factor(m4$cluster))
group_by(df_clusters4, cluster) %>%
  summarise(n = n(),
            ThreePAr = mean(ThreePAr),
            FTr = mean(FTr),
            ORB = mean(ORB),
            DRB = mean(DRB),
            TRB = mean(TRB),
            AST = mean(AST),
            STL = mean(STL),
            BLK = mean(BLK),
            TOV = mean(TOV),
            USG = mean(USG))
```

```
## # A tibble: 4 x 12
##   cluster     n ThreePAr    FTr   ORB   DRB   TRB   AST   STL   BLK   TOV   USG
##   <fct>   <int>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 1          69   0.355  0.258  4.75  17.4 11.1   11.2  1.69 1.91   11.2  19.7
## 2 2          52   0.0929 0.374 11.1   24.4 17.8   11.7  1.36 3.77   12.8  20.3
## 3 3         115   0.530  0.174  2.37  11.2  6.77  10.9  1.32 1.04   10.0  17.2
## 4 4          68   0.338  0.270  2.51  12.5  7.51  27.8  1.77 0.969  14.6  23.7
```

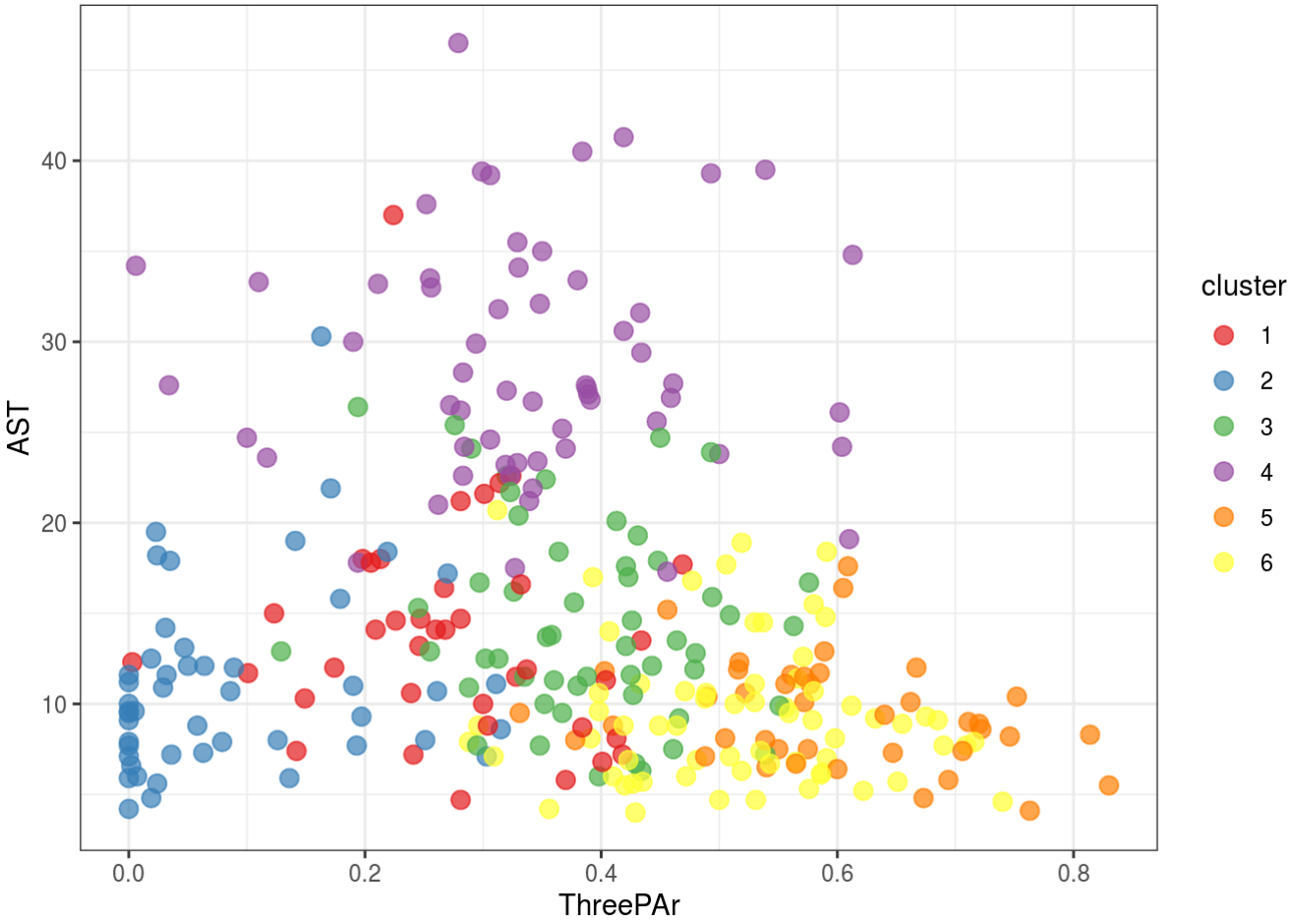
```
cent <- as_tibble(m4$centers) %>% mutate(cluster = factor(1:K))
ggplot(df_clusters4, aes(ThreePAr, AST, color = cluster)) + geom_point(size = 3, alpha =
.7) +
  theme_bw() + scale_color_brewer(type = "qual", palette = "Set1")
```



```
#trying to define 6 positions with 6 k-clusters
K <- 6
set.seed(1305)
m6 <- kmeans(scale(data.matrix(nbaNow[,5:14])), centers = K)
df_clusters6 <- mutate(nbaNow, cluster = factor(m6$cluster))
group_by(df_clusters6, cluster) %>%
  summarise(n = n(),
            ThreePAr = mean(ThreePAr),
            FTr = mean(FTr),
            ORB = mean(ORB),
            DRB = mean(DRB),
            TRB = mean(TRB),
            AST = mean(AST),
            STL = mean(STL),
            BLK = mean(BLK),
            TOV = mean(TOV),
            USG = mean(USG))
```

```
## # A tibble: 6 x 12
##   cluster     n ThreePAr   FTr   ORB   DRB   TRB   AST   STL   BLK   TOV   USG
##   <fct>   <int>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 1          38   0.274  0.278  5.52 19.0  12.2  13.8   1.94 2.14  11.9   21.2
## 2 2          48   0.0885 0.380 11.3  24.4  17.8  11.1   1.28 3.89  12.8   20.4
## 3 3          53   0.388  0.236  2.37 11.9   7.12 14.5   1.28 0.857 10.3   21.8
## 4 4          59   0.340  0.274  2.55 12.8   7.68 28.8   1.85 1.03  15     23.8
## 5 5          41   0.594  0.146  1.71  9.63  5.68  9.51  1.06 0.698  9.11  16.7
## 6 6          65   0.517  0.188  3.87 14.3   9.10  9.40  1.59 1.70  11.0   15.0
```

```
cent <- as_tibble(m6$centers) %>% mutate(cluster = factor(1:K))
ggplot(df_clusters6, aes(ThreePAr, AST, color = cluster)) + geom_point(size = 3, alpha =
.7) +
  theme_bw() + scale_color_brewer(type = "qual", palette = "Set1")
```



We did this analysis to determine if changing the number of clusters gave insight. We initally chose five clusters because traditionally there are five NBA positions. What we saw from changing the number of clusters is that there definitely is not more than five positons. Having six clusters does not appear to give any insight, there appears to be far too much overlap between each cluster. Having four positons could be useful, but we see that one of the clusters (cluster 3) has far more players than any oyther cluster has, meaning that that cluster appears to be a combination of clusters 3 and 5 from the 5 cluster model. This isnt a signifigant discover though, because we could have already seen that those two clusters were very similar.

```
m$tot.withinss; m4$tot.withinss; m6$tot.withinss
```

```
## [1] 1445.646
```

```
## [1] 1530.438
```

```
## [1] 1379.617
```

Comparing the variation of clusters between, 4, 5, and 6 clusters, we see how 6 clusters reduces the variation a lot compared to 5 clusters. This could mean that 6 positions might be a better way of looking at players, but when looking at certain variables, like in the previous plot, one can see lot of overlap. We can also see that the 6 cluster model only really split one existing cluster from the 5 cluster model, which doesn't say much in terms of 6 'new' positions. We also see more variation within the 4 cluster model, which makes sense as there is no way we

can go from defining 5 positions to 4 positions. One of the clusters has 120 players, which is huge compared to the even spread we saw with 5 clusters. In conclusion, more clusters will cause less variation, but it is a matter of how practical it is to have more positions when only 5 players from each team can play in the court at a time.

```
group_by(threesMade, Pos) %>% summarise(count = n(), avg_shot_distance = mean(SHOT_DIS
T), avg_def_distance = mean(CLOSE_DEF_DIST))
```

```
## # A tibble: 6 x 4
##   Pos   count avg_shot_distance avg_def_distance
##   <chr> <int>             <dbl>            <dbl>
## 1 C       178              23.9             7.16
## 2 PF     1491              24.5             7.26
## 3 PG     3612              24.6             5.82
## 4 PG-SG   119              25.2             6.32
## 5 SF     2736              24.0             6.60
## 6 SG     3567              24.2             5.93
```

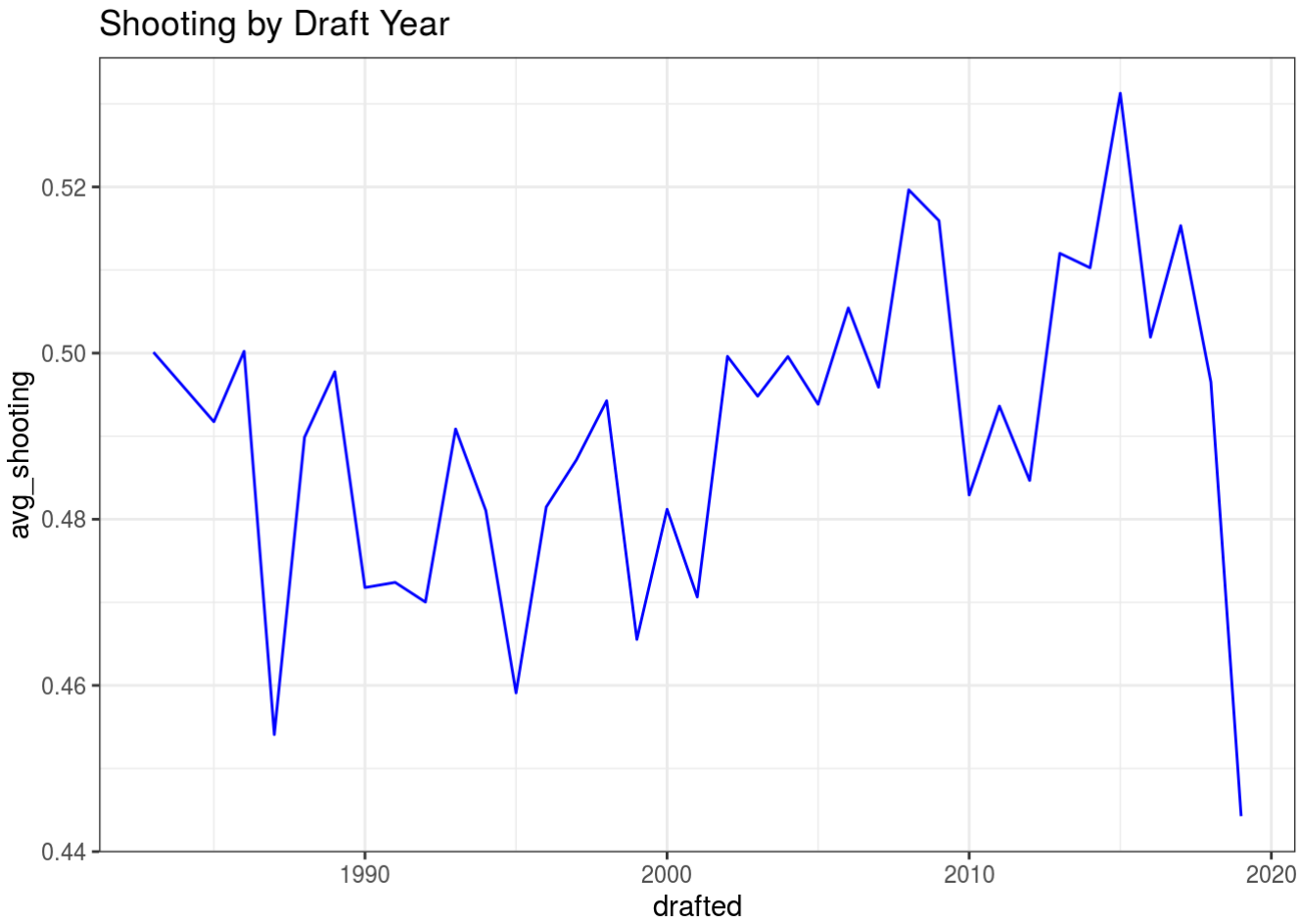
This summary shows the three pointers made differed between positions throughout the 2014-15 season, inlcuding statistics on how far the shot was from the basket and how far the nearest defender was when the player made that shot. It is interesting that those who are power forwards tend to make shots when defenders are slightly further away when compared to centers. This can be reflection of how centers only took shots they knew they could make, as 7 feet is a far distance, where as power forwards take more three's, and it just so happens that the ones they make are when they are left slightly more wide open. Point guards and shooting guards have the lowest average defender distance, a reflection of how well defenders guard that position as they generally make more threes. In terms of the distance of the shot itself, it is interesting that the PG-SG position has a significantly higher distance than the rest; the three point line is roughly 23 feet from the basket, so stepping 2 feet further out is considered a 'deep shot'. This just reflects the capability of these "flex" guards and how dangerous they can be, even though a defender may think the shot is too far to make.

```
group_by(twosMade, Pos) %>% summarise(sum = n(), avg_shot_distance = mean(SHOT_DIST), av
g_def_distance = mean(CLOSE_DEF_DIST))
```

```
## # A tibble: 6 x 4
##   Pos     sum avg_shot_distance avg_def_distance
##   <chr> <int>             <dbl>            <dbl>
## 1 C      8680              7.01             3.49
## 2 PF     9938              7.83             3.48
## 3 PG    10618              9.28             3.58
## 4 PG-SG   236             10.1              4.21
## 5 SF     6115              8.45             3.51
## 6 SG     7343              9.80             3.57
```

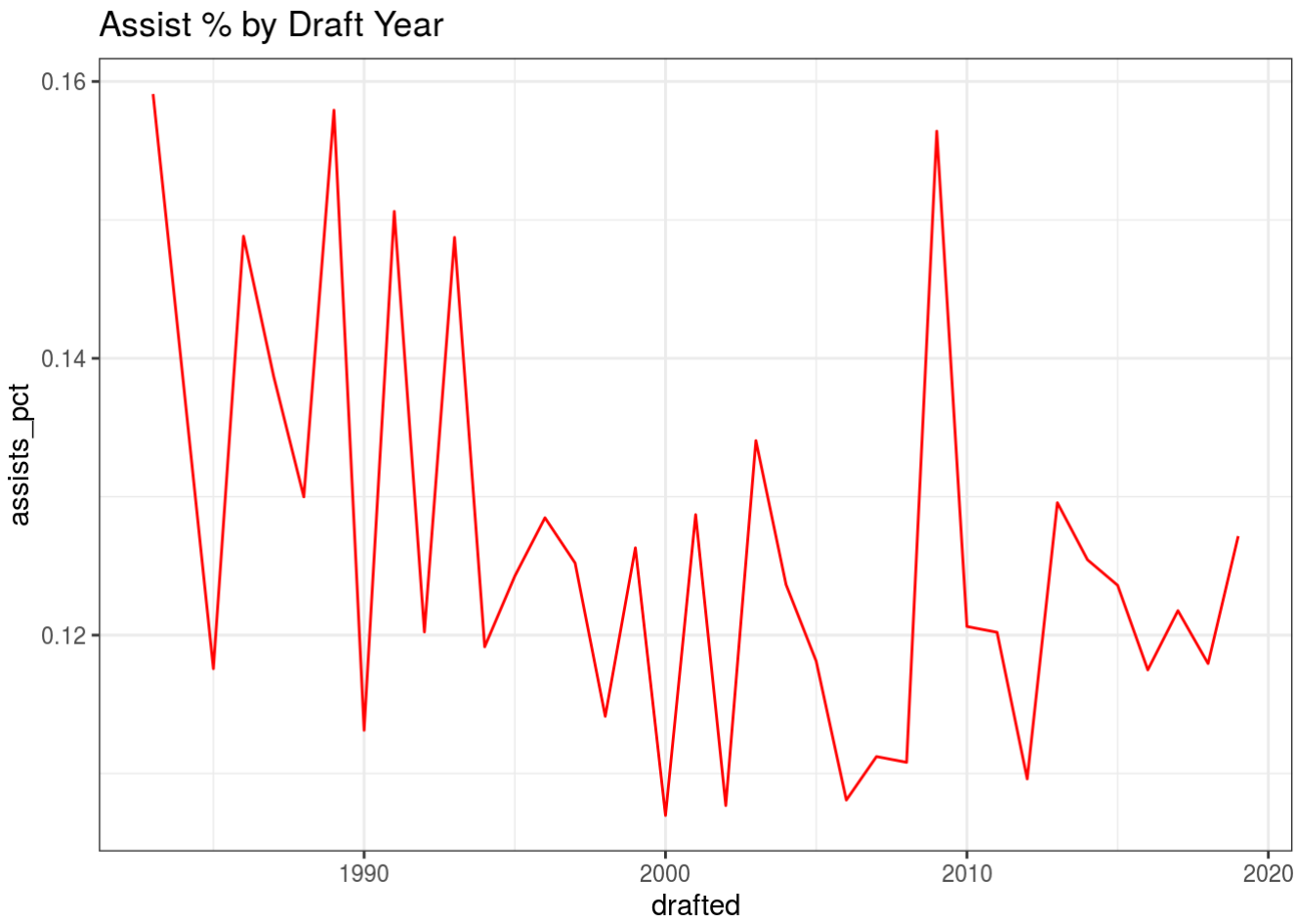
This summary shows the two point field goal makes and how they differ between positions throughout the 2014-15 season, also with statistics of the shot distance and nearest defender distance. There interestingly does not seem to be that big of a difference between the shot distance with which they make these shots. Centers stay close to the rim, but 7 feet away is definitely far from a simple layup. The point guards with the slight higher average shot distance reflects the mid-range shots that they go for, more often than centers. Regarding the defender distance, we really don't see that much variation, aside from those labelled as PG-SG.

```
ggplot(draft_by_year, aes(x = drafted, y = avg_shooting)) + geom_line(color = "blue") +
 theme_bw() +
  ggtitle("Shooting by Draft Year")
```



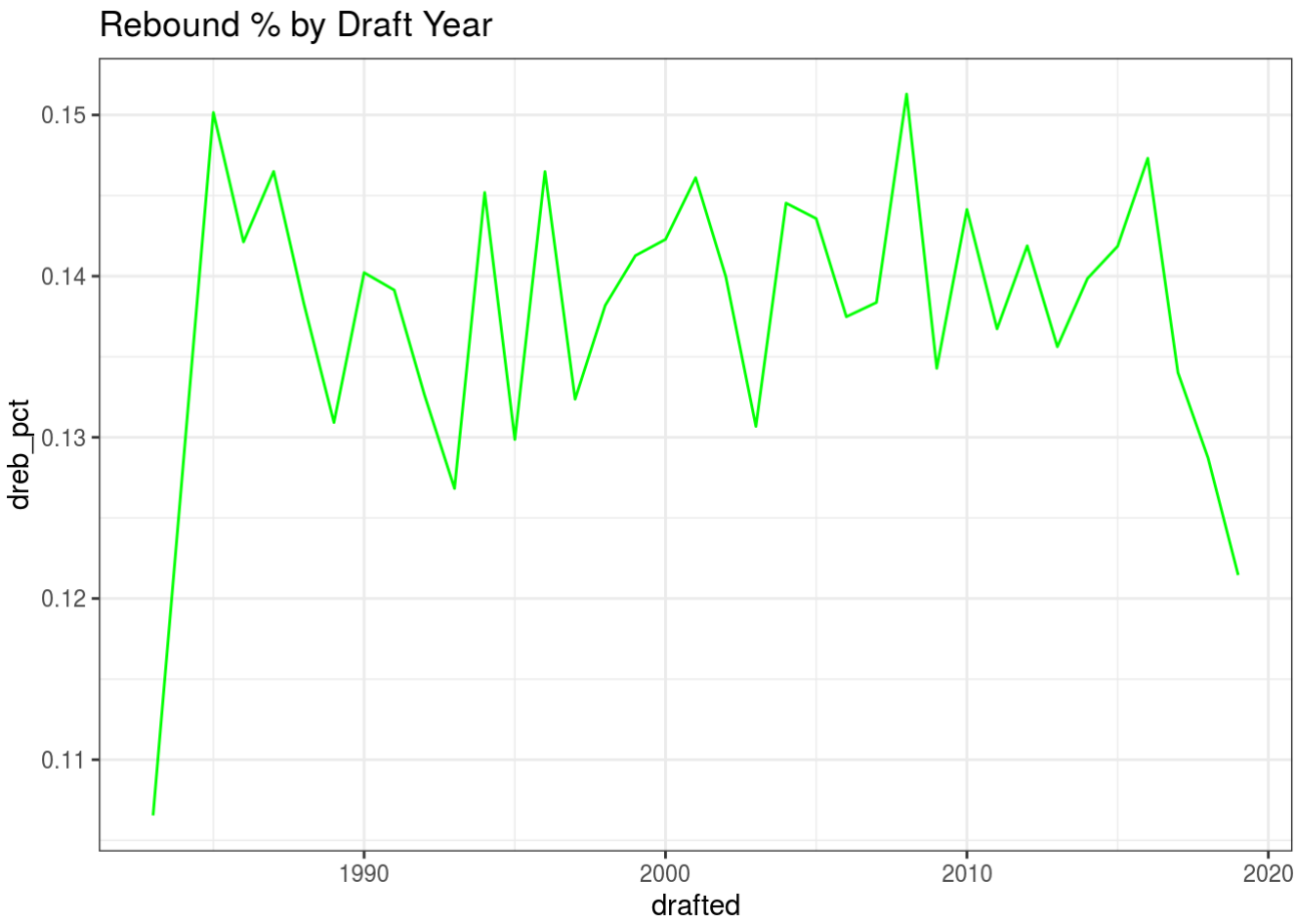
From this line graph, it is clearly shown that as the draft season increased, so did the players average shooting (an accumulation of free throws, 2 pointers and three pointers). This illustrates that recruiters were very focused on obtaining players that could play well in an NBA environment concerned largely with making high percentage shots. However, we are unsure what the steep drop in 2019 data is. It clearly does not follow the trend that began from 1986.

```
ggplot(draft_by_year, aes(x = drafted, y = assists_pct)) + geom_line(color = "red") + th
eme_bw() +
  ggtitle("Assist % by Draft Year")
```



This graph illustrates a decrease between the late 1900s to our present day regarding assists. This means that recruiters are not putting as much value onto this sort of playing style in our current NBA environment. This connects with the previous graph as recruiters are looking for players with higher average shooting, meaning they want their players to be able to make their own shots off the dribble.

```
ggplot(draft_by_year, aes(x = drafted, y = dreb_pct)) + geom_line(color = "green") + the
me_bw() +
  ggtitle("Rebound % by Draft Year")
```



Overall, the rebound percentage is a variable that stayed fairly constant during draft years. However in recent years it is starting to drop which is indicative of the NBA environment lessening the focus on rebounding, therefore there being less importance on the center and power forward position.