# Analysis of Player Performance in the NBA

*Gabriel Zanuttini-Frank*

*May 9, 2017*

## Introduction

General managers in the NBA have an extremely difficult job. With a limited amount of money and time, they are expected to assemble a group of basketball players that will be competitive, get along well with each other, and most importantly win championships. Making this job even tougher (or more interesting, depending on one's perspective) is the fact that there is no obvious way to determine which players are more valuable than others. Of course LeBron James is the best player in the world, but after him there is little consensus and constant debate about who is next. Traditionally, basketball players were evaluated based on four main factors: their size, effort, points per game, rebounds per game, assists per game, and the "eye test" (whether they have a good feel for the game and look comfortable and natural on the court). Recently, however, general managers have begun to realize that certain, more complicated, statistics are more important in identifying the superior players than the ones that appear in the standard box score. Furthermore, a general manager cannot simply identify the 12 best available players and expect them to become a competitive team. A team must consist of a number of players with different skill sets that complement each other, as well as address the many facets to the game of basketball (scoring, rebounding, defending, etc.). Similar to new statistics being used, teams have experimented using lineups that do not necessarily employ the typical five positions: point guard, shooting guard, small forward, power forward, and center.

### Design and Primary Questions

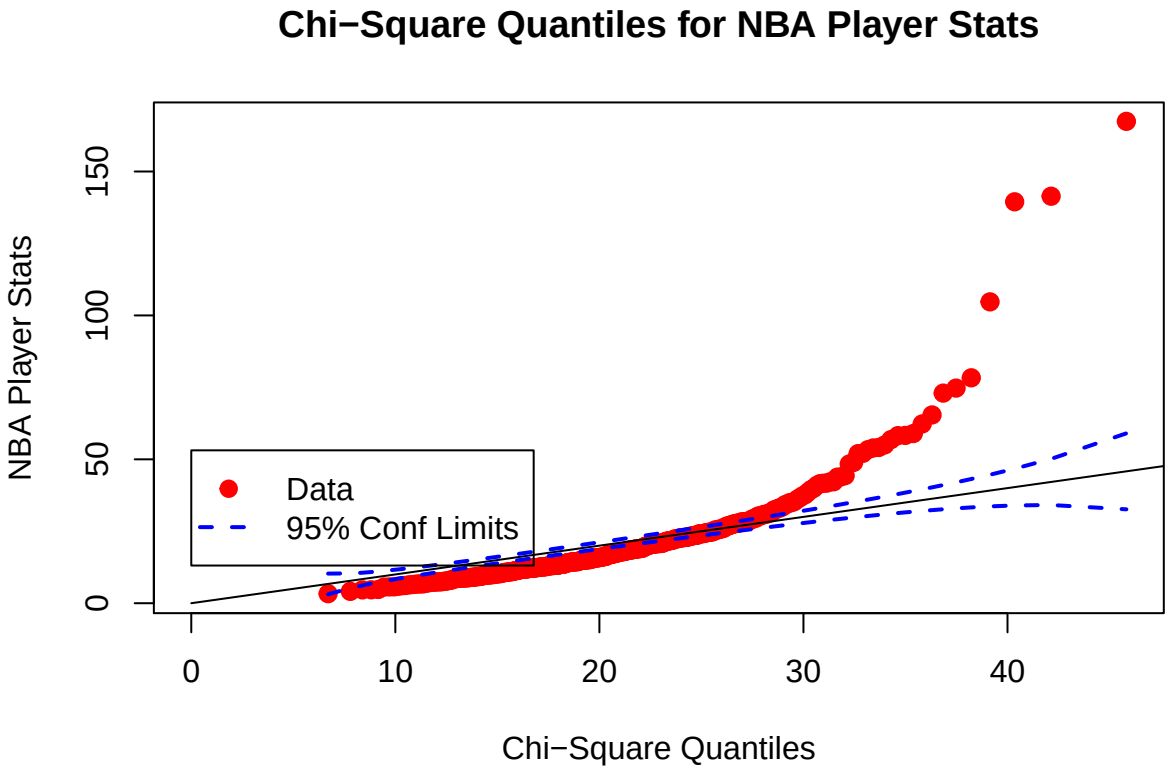
I will use three different multivariate techniques in an attempt to figure out ways to address the challenge of building a successful NBA team:

- **Principal components analysis** will allow me to reduce the dimensionality of my data in order to identify the components, or combinations of statistics, that account for the most variability among players. The results will be useful in showing the characteristics of a player that general managers should look into when comparing them to others.
- **Cluster analysis** will provide better insight into the different kinds of basketball players that exist. Rather than grouping players into categories based on the positions they are listed by, the clusters will group them based on their strengths and weaknesses as well as their roles on their respective teams. This will be useful in building a team because it will allow general managers to diversify the strengths of their players.
- I will use **MANOVA** to explore the variation among positions for players' average distance covered and average speed during a game. This will be a general way to determine the positions that play similar roles and in similar manners, and are therefore more interchangeable. In addition, I will explore how players' average distances and speeds vary with age. This will provide a baseline against which specific players can be compared, to recognize whether they may be physically slowing down more quickly than others of the same position.

## Data

I used two combined datasets for my analysis. For PCA and cluster analysis, I used data containing basic per-game statistics for each NBA player from the 2015-2016 season, and for MANOVA I used SportVU player tracking data, also from the 2015-2016 season.

- There are 57 variables in this dataset
  - Three are categorical: Team, Position, and Age Range
  - The remainder are continuous and fall into four categories:
    - *Shooting*: this includes the number of makes and attempts per game, as well as the percentage for field goals (all shots), three pointers, two pointers, and free throws. In addition, eFG. corresponds to effective field goal percentage, which uses this formula (FG + 0.5 * 3P) / FGA that adjusts for the fact that a 3-point field goal is worth one point more than a 2-pointer.
    - *Basic counting statistics*: these include statistics that show up commonly in box scores, namely: points, offensive and defensive rebounds, assists, steals, blocks, turnovers, fouls, minutes played, games played, and games started.
    - *Advanced statistics*: the statistics that I primarily used from this category are percentages for each basic statistic (for example, ORB. is offensive rebounding percentage, corresponding to the percentage of available offensive rebounds that a player got when in the game), X3PAr (three point attempt rate, or the percentage of one's shots that are 3-pointers), FTr (free throw rate), and TS. (true shooting percentage - a more advanced shooting metric that again takes into account the different values of shots).
    - *Player tracking*: this includes average distance traveled in feet per game, as well as average distance, average distance on offense, average distance on defense, average speed, average speed on offense, and average speed on defense, all measured in miles or miles per hour.
- The dataset contains 476 observations, or players. One challenged I faced was eliminating players that had multiple entries so that their numbers would not be counted twice. This occurred for players that were traded during the season, as their statistics for each team were split up into different rows. In addition, often in my analyses I found it useful to eliminate observations of players that played very few games or minutes because they might have very extreme values for certain statistics having not played enough for them to even out.
- The data was collected from two sources. The first three categories (shooting, basic and advanced statistics) were taken from basketball-reference.com, and the player tracking data was taken from NBA.com.
- The NBA measures and publishes data for hundreds of variables. Therefore, my data is by no means complete, and similar analyses conducted with different variables could produce different results. However, I tried to include the more general statistics that covered all of the major aspects of the game rather than going into great depth in one certain category, such as shooting or defense.
- I used chi-squared quantile plots to examine the multivariate normality of my continuous variables.




This data does not seem to quite have a multivariate normal distribution, as the data in the chi-squared quantile plot does not mostly lie within the 95% confidence boundaries. I tried to transform all of the variables by taking square roots of counts and computing logits of percentages, but found that it did not significantly improve the chi-squared plot, so I will proceed with the original data. - Because this dataset has lots of observations and variables, it sometimes was slightly harder to work with and understand everything that was going on.

## Principal Components Analysis

The first step toward performing principal components analysis is computing and observing correlations between the variables.

```
##         Age     G    GS    MP   FG.   X3P  X3PA  X3P.   X2P  X2PA  X2P.
## Age    1.00 -0.13 -0.03 -0.11  0.00  0.00 -0.01 -0.03 -0.14 -0.15  0.03
## G     -0.13  1.00  0.48  0.76  0.13  0.38  0.40  0.09  0.52  0.51  0.18
## GS    -0.03  0.48  1.00  0.79  0.18  0.39  0.40  0.06  0.70  0.70  0.14
## MP    -0.11  0.76  0.79  1.00  0.10  0.57  0.59  0.16  0.79  0.80  0.11
## FG.    0.00  0.13  0.18  0.10  1.00 -0.34 -0.40 -0.42  0.35  0.23  0.88
## X3P    0.00  0.38  0.39  0.57 -0.34  1.00  0.99  0.49  0.23  0.27 -0.21
## X3PA  -0.01  0.40  0.40  0.59 -0.40  0.99  1.00  0.48  0.24  0.29 -0.23
## X3P.  -0.03  0.09  0.06  0.16 -0.42  0.49  0.48  1.00 -0.03  0.01 -0.31
## X2P   -0.14  0.52  0.70  0.79  0.35  0.23  0.24 -0.03  1.00  0.99  0.26
## X2PA  -0.15  0.51  0.70  0.80  0.23  0.27  0.29  0.01  0.99  1.00  0.13
## X2P.   0.03  0.18  0.14  0.11  0.88 -0.21 -0.23 -0.31  0.26  0.13  1.00
## FT    -0.10  0.42  0.61  0.72  0.11  0.44  0.46  0.05  0.80  0.82  0.07
## FTA   -0.13  0.43  0.62  0.73  0.20  0.34  0.36 -0.02  0.82  0.82  0.14
## FT.    0.12  0.07  0.05  0.17 -0.41  0.44  0.45  0.36  0.09  0.14 -0.32
## ORB   -0.15  0.40  0.44  0.42  0.59 -0.25 -0.26 -0.40  0.57  0.50  0.45
```

```
## DRB  -0.11  0.55  0.65  0.71  0.41  0.13  0.13 -0.13  0.74  0.70  0.34
## AST   0.01  0.36  0.50  0.63 -0.11  0.47  0.49  0.24  0.54  0.59 -0.10
## STL  -0.09  0.54  0.63  0.78 -0.02  0.54  0.56  0.21  0.59  0.61  0.03
## BLK  -0.12  0.32  0.42  0.38  0.52 -0.16 -0.16 -0.32  0.51  0.44  0.44
## TOV  -0.12  0.51  0.66  0.79  0.04  0.47  0.50  0.13  0.78  0.81  0.02
## PF   -0.16  0.72  0.65  0.78  0.26  0.28  0.29 -0.04  0.63  0.62  0.23
##         FT   FTA   FT.   ORB   DRB   AST   STL   BLK   TOV    PF
## Age  -0.10 -0.13  0.12 -0.15 -0.11  0.01 -0.09 -0.12 -0.12 -0.16
## G     0.42  0.43  0.07  0.40  0.55  0.36  0.54  0.32  0.51  0.72
## GS    0.61  0.62  0.05  0.44  0.65  0.50  0.63  0.42  0.66  0.65
## MP    0.72  0.73  0.17  0.42  0.71  0.63  0.78  0.38  0.79  0.78
## FG.   0.11  0.20 -0.41  0.59  0.41 -0.11 -0.02  0.52  0.04  0.26
## X3P   0.44  0.34  0.44 -0.25  0.13  0.47  0.54 -0.16  0.47  0.28
## X3PA  0.46  0.36  0.45 -0.26  0.13  0.49  0.56 -0.16  0.50  0.29
## X3P.  0.05 -0.02  0.36 -0.40 -0.13  0.24  0.21 -0.32  0.13 -0.04
## X2P   0.80  0.82  0.09  0.57  0.74  0.54  0.59  0.51  0.78  0.63
## X2PA  0.82  0.82  0.14  0.50  0.70  0.59  0.61  0.44  0.81  0.62
## X2P.  0.07  0.14 -0.32  0.45  0.34 -0.10  0.03  0.44  0.02  0.23
## FT    1.00  0.97  0.26  0.32  0.59  0.61  0.62  0.33  0.80  0.52
## FTA   0.97  1.00  0.09  0.45  0.68  0.56  0.61  0.43  0.79  0.57
## FT.   0.26  0.09  1.00 -0.37 -0.15  0.25  0.11 -0.28  0.17 -0.07
## ORB   0.32  0.45 -0.37  1.00  0.79 -0.02  0.21  0.73  0.28  0.62
## DRB   0.59  0.68 -0.15  0.79  1.00  0.31  0.50  0.73  0.60  0.74
## AST   0.61  0.56  0.25 -0.02  0.31  1.00  0.71  0.00  0.85  0.38
## STL   0.62  0.61  0.11  0.21  0.50  0.71  1.00  0.18  0.75  0.63
## BLK   0.33  0.43 -0.28  0.73  0.73  0.00  0.18  1.00  0.29  0.56
## TOV   0.80  0.79  0.17  0.28  0.60  0.85  0.75  0.29  1.00  0.62
## PF    0.52  0.57 -0.07  0.62  0.74  0.38  0.63  0.56  0.62  1.00
```

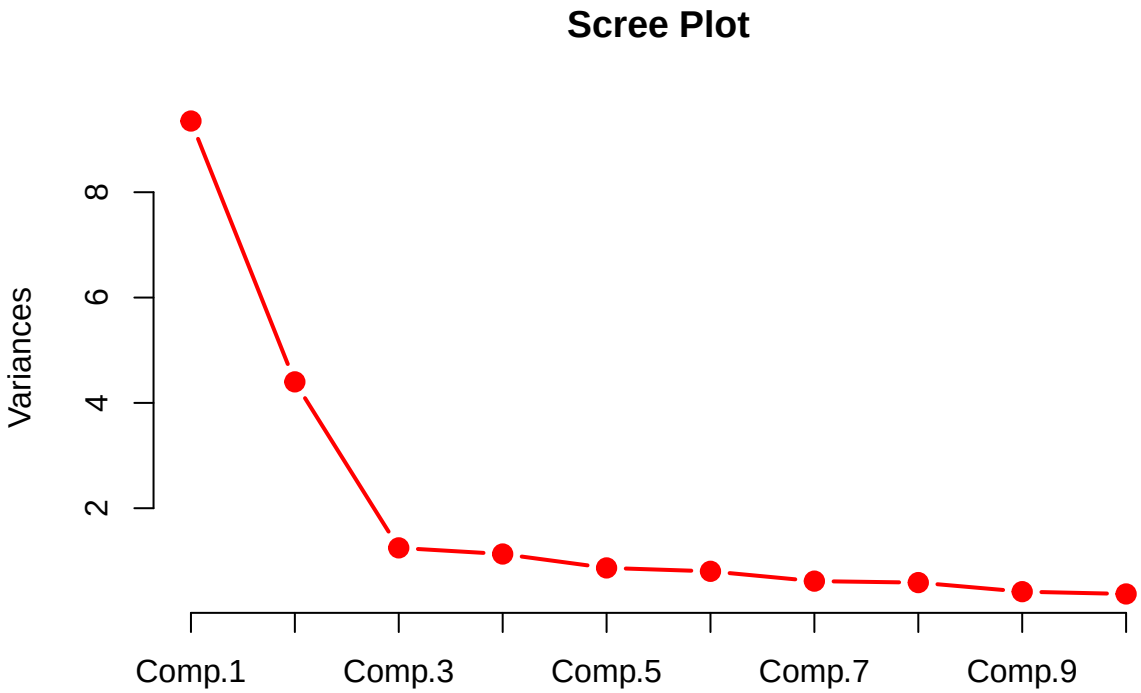
Many of the relationships I observed in this correlation matrix are intuitive. For example, minutes played is highly correlated with games played and games started, which makes sense because they all signify roughly the same characteristic about a player, namely that he plays a lot. Field goal % is strongly correlated with two point % and has a strong negative correlation with 3 point and free throw percentage. These factors point to a difference in the skills between players of different positions; guards tend to shoot better from the three point and free throw lines while big men shoot better inside (for two points).

The next step is to determine how many principal components to retain. There are four different methods or criteria that I will employ to guide me toward a decision in that regard: Eigenvalue>1 criterion, cumulative variance explained method, scree plot, parallel analysis.

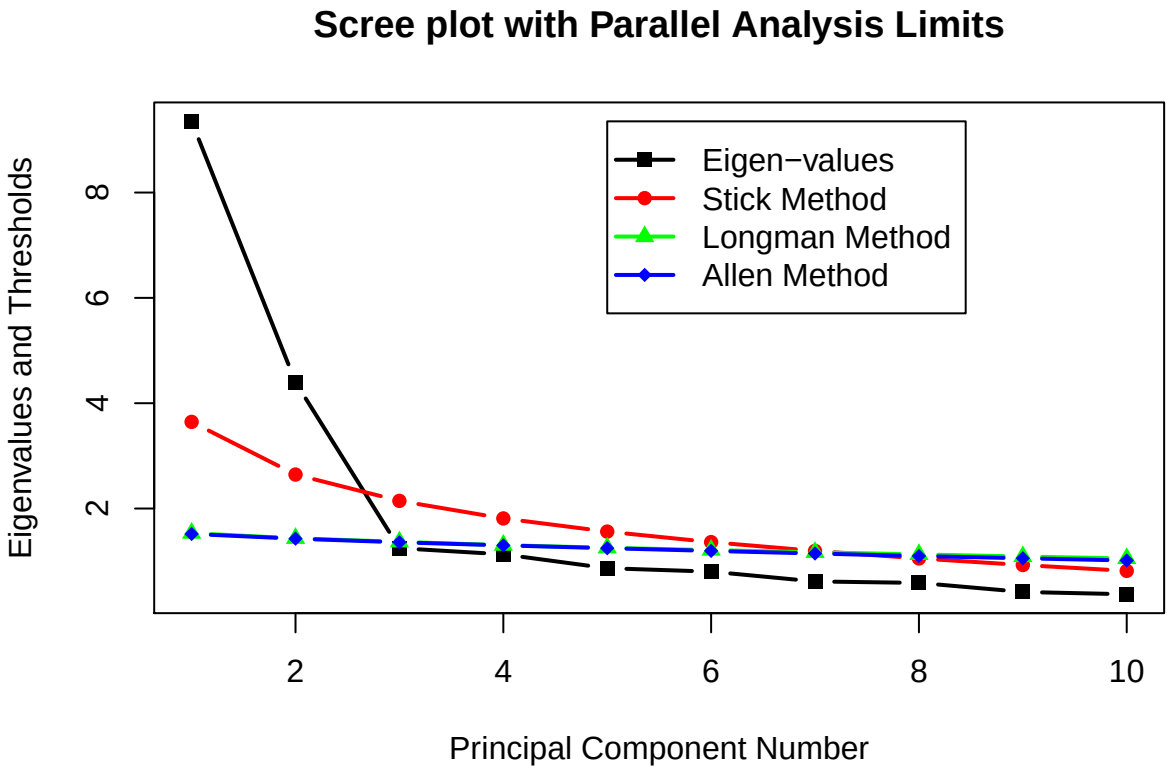
```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
## Standard deviation     3.0580997 2.0974373 1.11674566 1.06331645
## Proportion of Variance 0.4453321 0.2094878 0.05938671 0.05384009
## Cumulative Proportion  0.4453321 0.6548199 0.71420657 0.76804666
##                           Comp.5     Comp.6    Comp.7     Comp.8
## Standard deviation     0.93085113 0.89550662 0.7846355 0.76680771
## Proportion of Variance 0.04126113 0.03818724 0.0293168 0.02799972
## Cumulative Proportion  0.80930780 0.84749504 0.8768118 0.90481156
##                           Comp.9    Comp.10    Comp.11    Comp.12
## Standard deviation     0.6448060 0.61021035 0.52811545 0.47300999
## Proportion of Variance 0.0197988 0.01773127 0.01328123 0.01065421
## Cumulative Proportion  0.9246104 0.94234163 0.95562286 0.96627707
##                           Comp.13     Comp.14     Comp.15     Comp.16
## Standard deviation     0.454480811 0.431134151 0.321881089 0.285456295
```

```
## Proportion of Variance 0.009835848 0.008851269 0.004933687 0.003880252
## Cumulative Proportion  0.976112920 0.984964189 0.989897876 0.993778129
##                           Comp.17     Comp.18      Comp.19      Comp.20
## Standard deviation     0.255194787 0.212766060 0.1052382094 0.0756624862
## Proportion of Variance 0.003101161 0.002155686 0.0005273848 0.0002726101
## Cumulative Proportion  0.996879289 0.999034975 0.9995623598 0.9998349699
##                           Comp.21
## Standard deviation     0.0588696156
## Proportion of Variance 0.0001650301
## Cumulative Proportion  1.0000000000
```

- The Eigenvalue $> 1$ criteria would suggest keeping 4 components.
- The ~80% of total variance explained method would suggest keeping 5, which explain roughly 81% of total variance.




- Despite the previous indications that 4 or 5 was the correct number of components to retain, the scree plot above pretty convincingly suggests keeping only 2.




- Lastly, parallel analysis strongly supports keeping only two components as well.

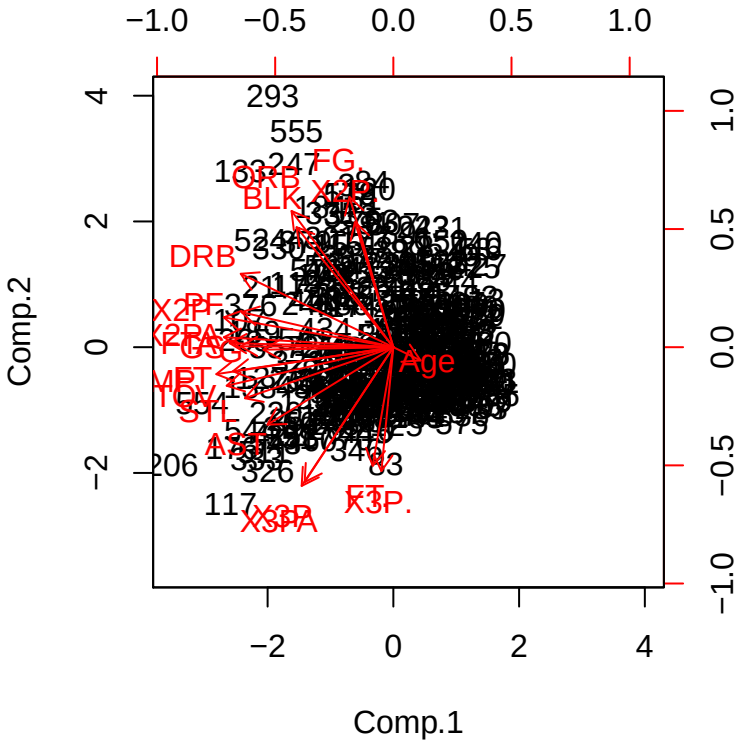
Looking more closely at the Eigenvalue and total explained variance methods that suggested four components, it becomes increasingly evident that 2 is the right choice. Putting aside the cutoff of Eigenvalue=1, there is a much more significant decrease from the second to the third Eigenvalue than from the third to fourth or fourth to fifth. Thus, keeping two components rather than four does not greatly decrease the total explained variance. As a result **I will choose to keep only the first two principal components.**

Next, I will examine the loadings of the two principal components I decided to retain.

```
##            Comp.1       Comp.2
## Age   0.04696254 -0.033446678
## G    -0.22469966 -0.006320433
## GS   -0.26415255  0.002258170
## MP   -0.30621378 -0.068007143
## FG.  -0.07585861  0.378878241
## X3P  -0.15341884 -0.340198969
## X3PA -0.15859919 -0.349848532
## X3P. -0.02070857 -0.312444250
## X2P  -0.29344557  0.075240825
## X2PA -0.29312424  0.025233944
## X2P. -0.06649758  0.315405040
## FT   -0.27745079 -0.060703104
## FTA  -0.28303823  0.010132901
## FT.  -0.03573448 -0.299167730
## ORB  -0.17608709  0.343568483
## DRB  -0.26400224  0.185182502
## AST  -0.21769978 -0.195161674
## STL  -0.25669177 -0.128732574
## BLK  -0.16740946  0.302646576
## TOV  -0.28830595 -0.097199332
```

```
## PF   -0.26254938  0.088374283
```

The first component can be interpreted as the overall skill of a player. The variables with the most weight in the first component are: minutes played, two-point attempts and conversions, free throw attempts and makes, defensive rebounds, and personal fouls. These are all statistics that can generally distinguish between players that play more and less often, which is usually a good measure of a player's skill. The second component represents the size or position of a player, specifically whether he is a guard or a big man. The most impactful variables in this component are field goal and two point % (positive), three point and free throw % (negative), three pointers attempted and made (negative), and offensive rebounds (positive). All of these variables highlight the biggest differences between guards and big men, and examining their the signs suggests that big men have high positive values while guards' values in this component are more negative.



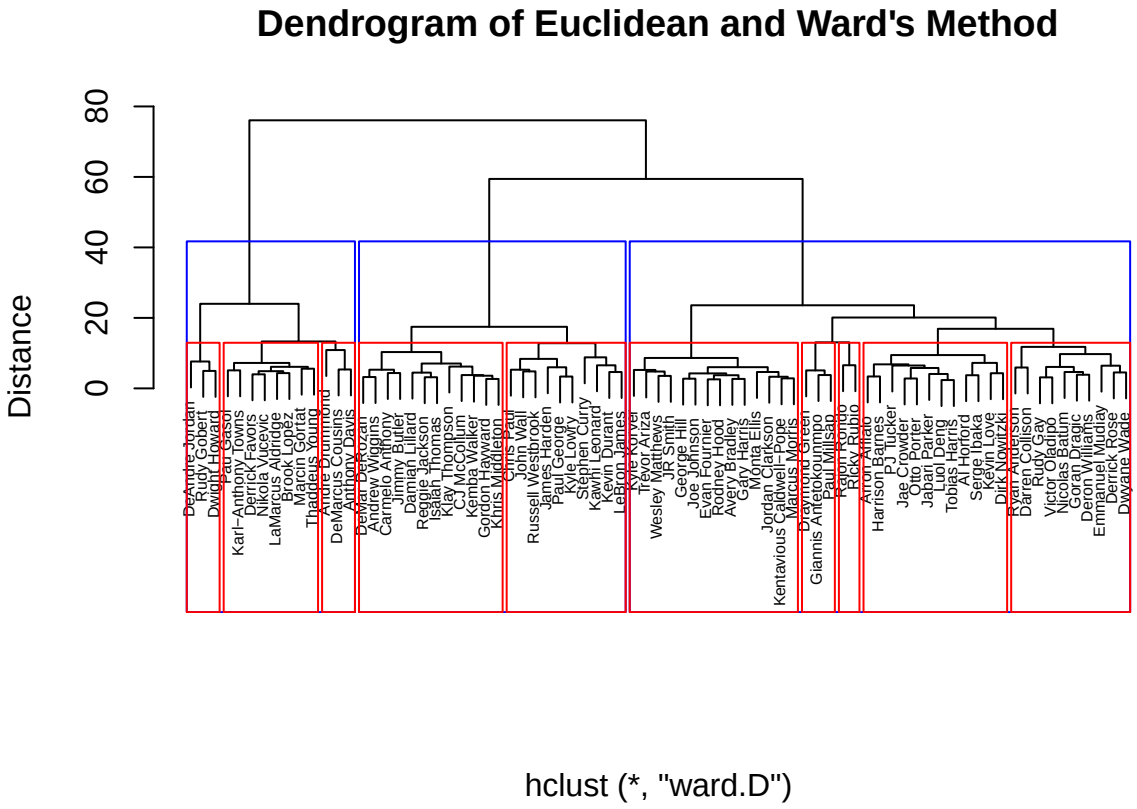

Though it is difficult to discern a pattern in the distribution of factors for component 1 in this biplot, component 2 is more clearly interpretable. In the negative direction are factors that would point towards smaller players, such as three point attempts and percentage, whereas the positive direction has factors that would point towards bigger players, such as blocks and field goal percentage. This provides further support for the interpretation of component 2 as position.

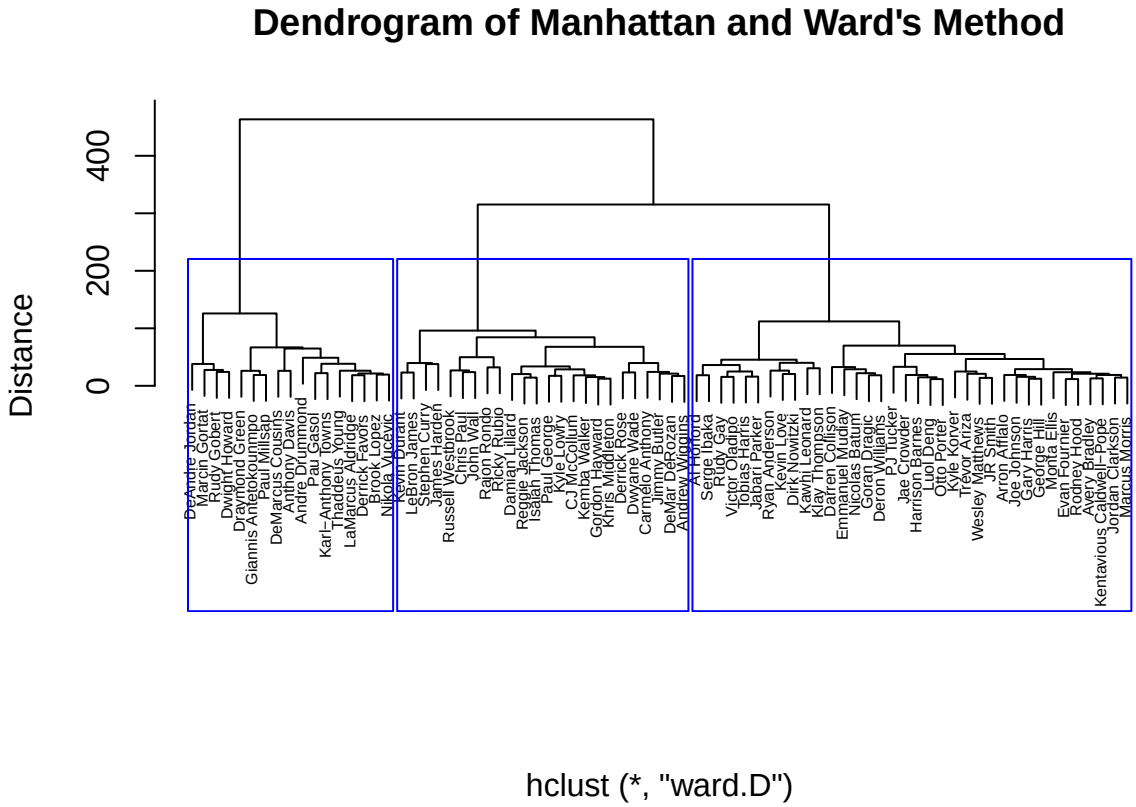
I think that PCA was effective in reducing the dimensionality of my data. The two components that I ended up keeping summarize the two characteristics that most describe NBA players: talent and position. This does not provide much new insight (perhaps using more nuanced data than general box score statistics could have accomplished this), but it definitely reinforces the general intuitions of which kinds of players accumulate which kinds of statistics, and which statistics characterize better players. From looking at the score plots, it is evident thatthe best players are outliers. This makes sense because they are stars due to the fact that their statistics differ most from the mean. Overall, I did not run into many problems with my dataset. I had to fill in some NA's and slightly alter certain statistics when doing transformations, but the sample size was plenty large enough relative to the number of variables even after filtering out less significant players.

## Cluster Analysis

For cluster analysis, I used the same statistics as I did for principal components analysis. Instead of trying log and square root transformations as I did for PCA, I standardized the variables. This is because there are a number of different scales and ranges on which variables are measured in this dataset. For example, certain variables are measured as percentages, others (such as blocks) have ranges of around 0-4, while others (such as points) range from 0 to 30. By standardizing all of them, each variable will have the same weight and effect for clustering the players.

The data I am using has all continuous variables, meaning that none of them are binary or categorical. This means that I should use one of the common continuous distance metrics, such as Euclidean or Manhattan. I generated dendrograms using both the Ward and average linkage agglomeration methods, each with both Euclidean and Manhattan distance. I found that Ward's Method produced the best looking dendrograms with better defined group boundaries. On the other hand, average linkage is more sensitive to outliers, so it creates many small groups, rather than a smaller number of more equally sized groups. Below are the two dendrograms produced using Ward's Method.



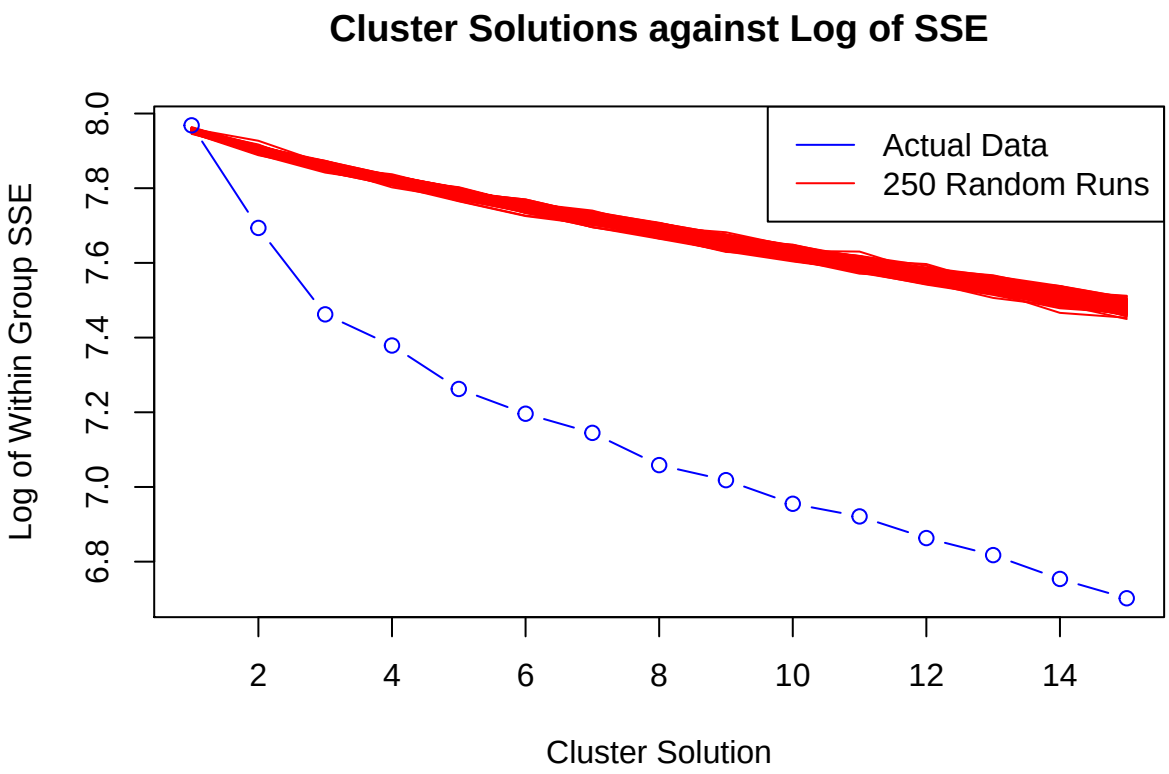



Simply from observing both dendrograms, it appears that there are three well-defined clusters, which contain many of the same players in both cases. If I try to divide the tree into four groups, rather than maintain equally sized groups, the fourth group consists of only three or four players from within one group. This leads me to believe that the fourth group defines only a very particular set of players, rather than helping to create a general classification of players. The three main groups that emerged from this clustering (outlined by the blue lines) can be explained (from left to right) as big men, high usage players (players with the ball in their hands most of the time), and more secondary/role players (essentially the players on teams of a high usage player who therefore play off of the ball more often). Some variables that might define each group could be high blocks, rebounds, and low three point shots for big men; high field goal attempts, usage rate, and assists for the high usage (star player) group; and high three point percentage, fewer minutes, field goal attempts, and points for the third, role player group.

If I had wanted to define more specific groups, I could have chosen 10 clusters, as shown by the red outline in the Euclidean and Ward's Method dendrogram. Since I am knowledgable about all of the players and their tendencies, I would be able to define each group's main characteristics. However, I believe that having such a large number of groups would be more worthwhile if I had included more players, so **I will proceed with the three cluster solution.**

### K-Means Clustering

I next performed k-means clustering on the data as another method of partitioning the players into different numbers of groups. Again, I used the standardized data.

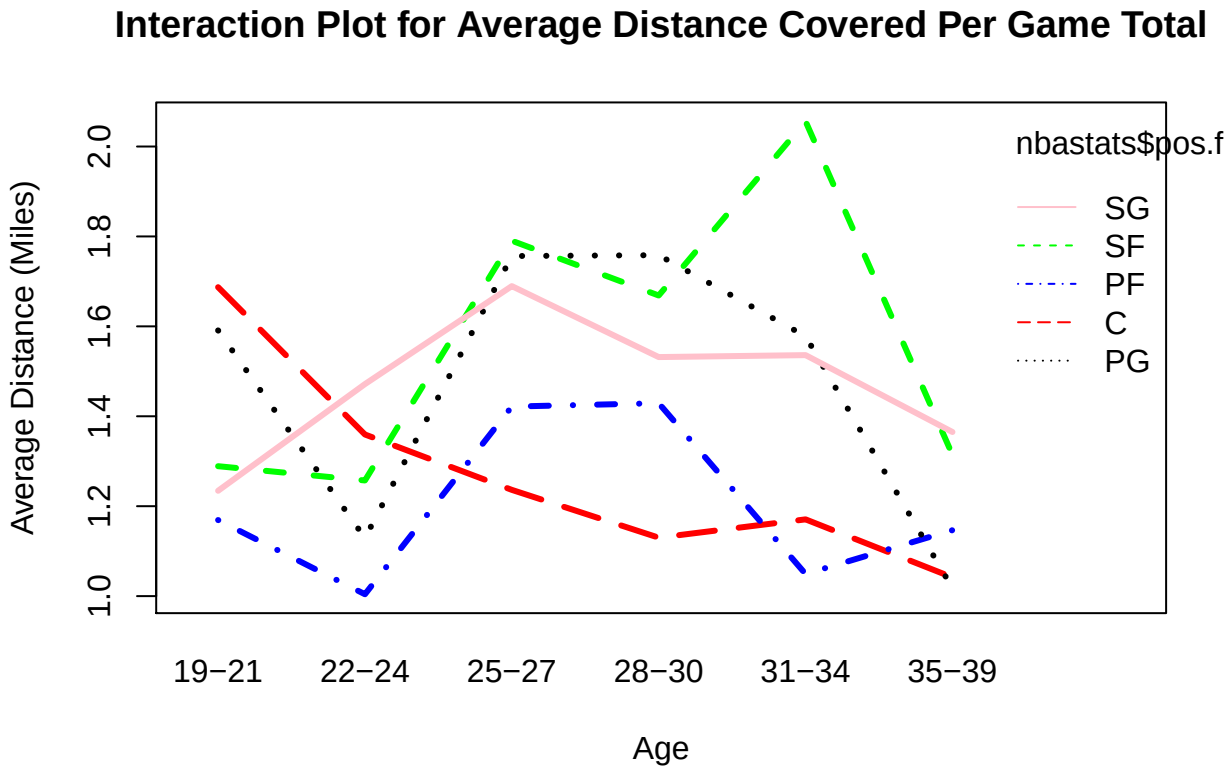



Above is the plot of within groups SSE against all tested cluster solutions for actual and randomized data. From looking at this plot, it is evident that **the correct number of groups is three.** After the third cluster, the within group SSE of the actual data begins to decrease at a rate very similar to that of the random runs. This yields the same result as Euclidean and Ward's Method did above, which suggests that the groups are fairly well defined. Further, I examined which players were placed in which cluster, and whether the clusters created using k-means and Ward's Method contained the same players. Over 88% of the players were clustered the same way using k-means and Ward's Method with Euclidean and Manhattan distances. This provides evidence that not only should there be three clusters, but that they are fairly well defined.

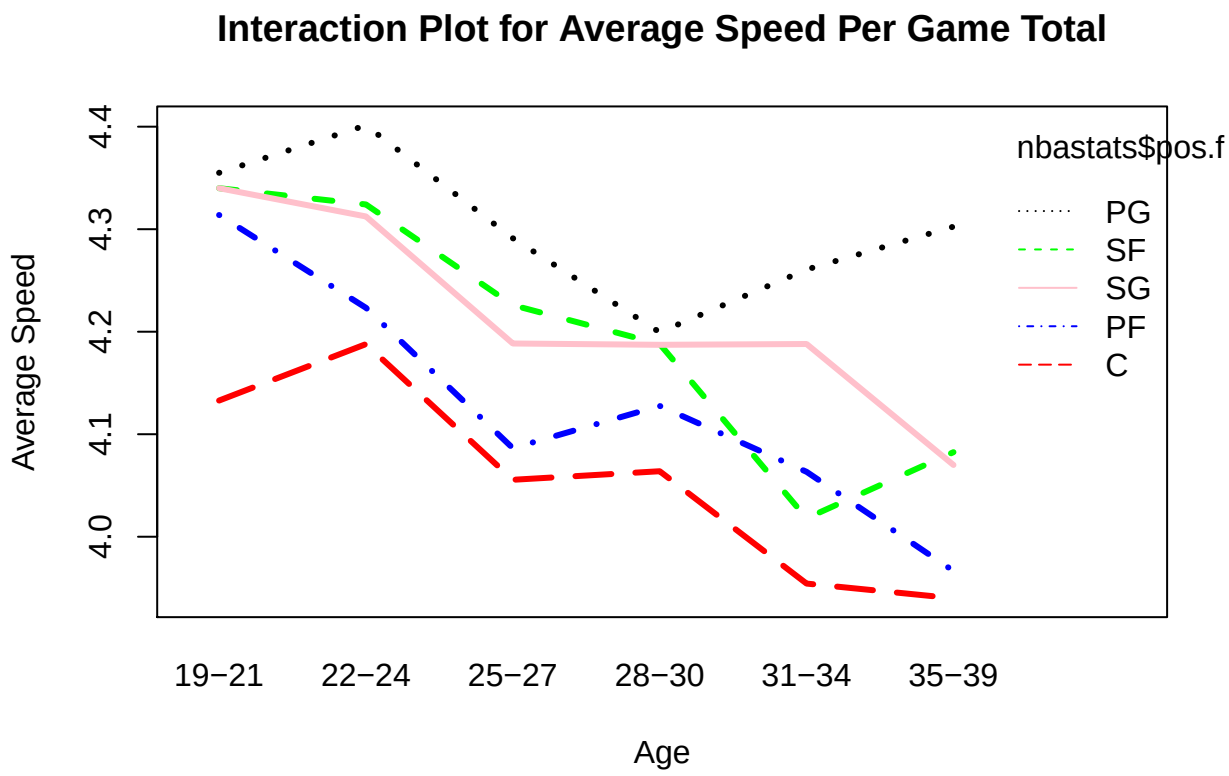
In terms of what effect this would have for a general manager, it would suggest that it is important for a team to have players that lie in each of these three clusters. In addition, a general manager should look to compare players within clusters rather than necessarily by their actual size, declared position, and overall statistics.

# MANOVA

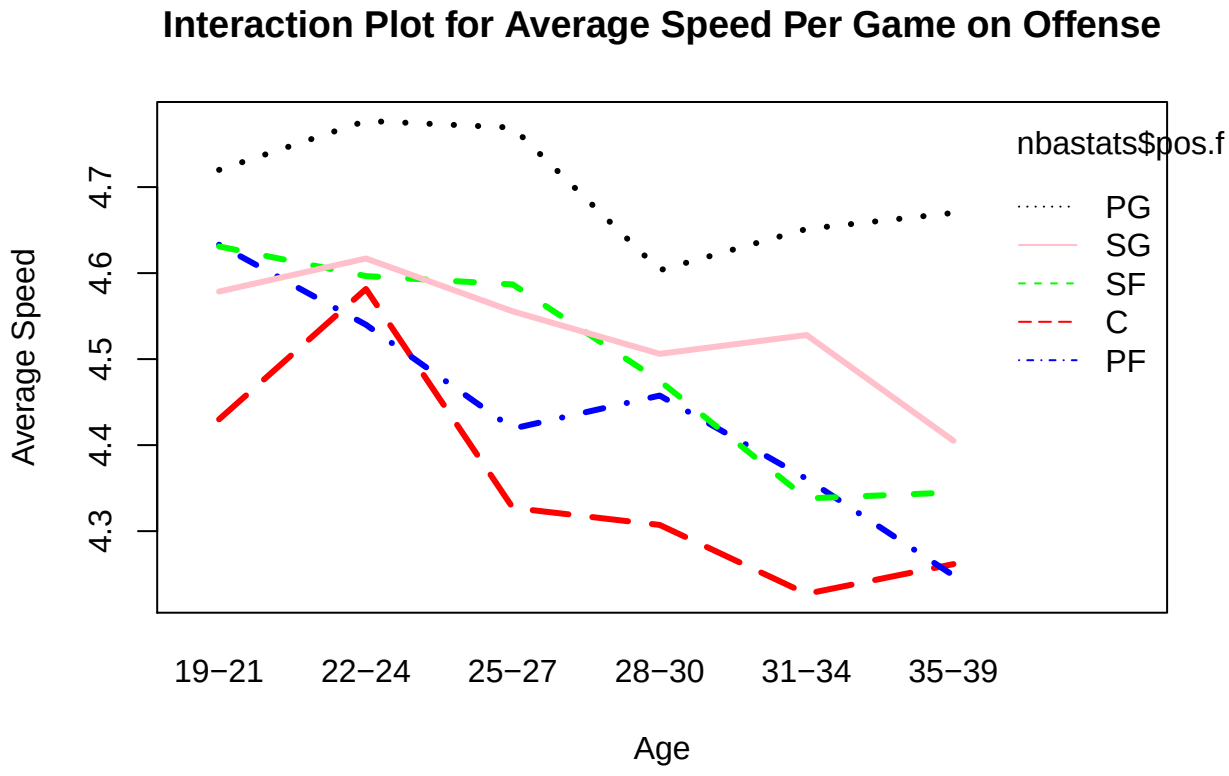
In this section I explore the effect that the categorical variables position and age-range have on players' on-court tendencies, specifically their average speed and distance traveled per game. Below are the interaction plots for each response variable.

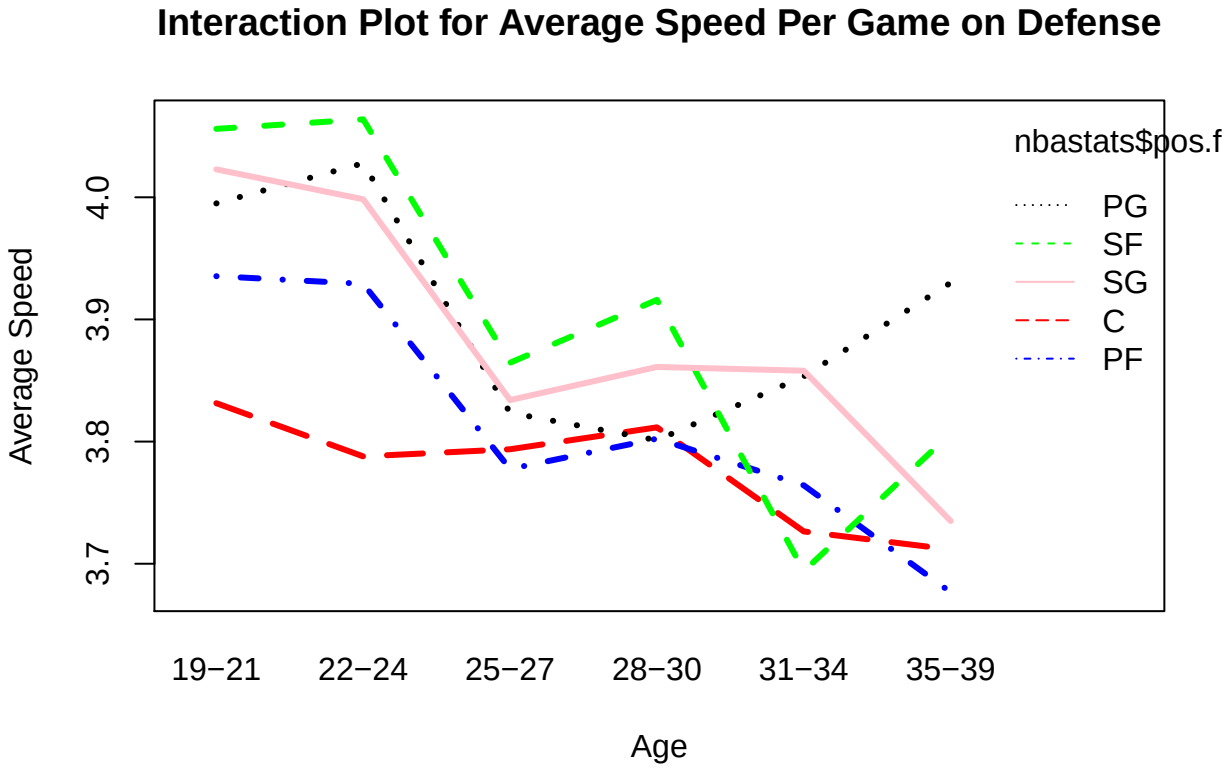



This interaction plot shows how the average distance covered per game varies with age for each position. Plots of the average distance on offense and defense are similar to this, meaning that these trends do not change much when looking at offense, defense, or the total. This plot reveals that big men (power forwards and centers) cover less distance in a game compared to smaller players. This makes sense because bigger players are generally more stationary and closer to the basket on both sides of the ball, so they naturally cover less distance over the course of a game. The other trend that becomes apparent in these plots is that most players reach their highest levels of distance covered per game between the ages of 25 and 30. This aligns well with the general intuition that players peak in those ages. At their peak, players should in theory be in their best physical shape, allowing them to move more during the game, both chasing players on defense and getting open on offense.




This interaction plot shows how the average speed per game of a player varies with age and position. This plot shows that there is a more consistent difference by position on average speed as compared to total distance traveled, and that age has a roughly linear (negative) correlation with average speed. In addition, the plots for average speed on offense and defense differ significantly:






On offense, there is a clear hierarchy as to which players move the fastest, with point guards being at the top followed by shooting guards, small forwards, power forwards and centers. Interestingly, bigger players appear to experience a larger dropoff in average speed over the course of their careers than guards do, perhaps because it is not as vital an aspect of their game and they therefore can focus on their strengths inside rather than their speed as they get older. On defense however, the average speeds of all five positions are quite similar. Power forwards and centers are still noticeably slower than smaller players, but the differences are much smaller than on offense. Finally, the plot for average speed total seems to even out the exaggerated differences on offense and the similar values on defense. Point guards clearly have the highest average speeds, followed by shooting guards and small forwards who are very similar, and then power forwards with centers coming in last. This third plot also provides more evidence for the fact that average speed of players of all positions decreases with age.

```
##  Response distMI :
##                                  Df  Sum Sq Mean Sq F value    Pr(>F)
## nbastats$pos.f                    4   8.639 2.15975  6.0501 9.522e-05 ***
## nbastats$agerange                 5   9.211 1.84217  5.1604 0.0001293 ***
## nbastats$pos.f:nbastats$agerange 20  11.186 0.55928  1.5667 0.0566947 .
## Residuals                       446 159.213 0.35698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response distMIO :
##                                  Df Sum Sq Mean Sq F value    Pr(>F)
## nbastats$pos.f                    4  2.732 0.68306  6.4933 4.376e-05 ***
## nbastats$agerange                 5  2.552 0.51050  4.8528 0.0002472 ***
## nbastats$pos.f:nbastats$agerange 20  3.445 0.17225  1.6374 0.0408821 *
## Residuals                       446 46.917 0.10520
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##  Response distMID :
##                                  Df Sum Sq Mean Sq F value    Pr(>F)
## nbastats$pos.f                    4  1.686 0.42140  5.5738 0.0002192 ***
## nbastats$agerange                 5  2.060 0.41203  5.4499 7.016e-05 ***
## nbastats$pos.f:nbastats$agerange 20  2.238 0.11189  1.4800 0.0832870 .
## Residuals                       446 33.719 0.07560
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response avgspeed :
##                                  Df  Sum Sq Mean Sq F value    Pr(>F)
## nbastats$pos.f                    4  3.1556 0.78891 23.5089 < 2.2e-16 ***
## nbastats$agerange                 5  2.8487 0.56973 16.9777 2.235e-15 ***
## nbastats$pos.f:nbastats$agerange 20  0.4902 0.02451  0.7305    0.7954
## Residuals                       446 14.9668 0.03356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response avgspeedO :
##                                  Df  Sum Sq Mean Sq F value    Pr(>F)
## nbastats$pos.f                    4  6.1375 1.53438 23.9322 < 2.2e-16 ***
## nbastats$agerange                 5  2.8776 0.57552  8.9766 3.857e-08 ***
## nbastats$pos.f:nbastats$agerange 20  1.1174 0.05587  0.8714    0.6243
## Residuals                       446 28.5946 0.06411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response avgspeedD :
##                                  Df  Sum Sq Mean Sq F value    Pr(>F)
## nbastats$pos.f                    4  1.5285 0.38214 10.7202 2.645e-08 ***
## nbastats$agerange                 5  2.7160 0.54319 15.2382 7.783e-14 ***
## nbastats$pos.f:nbastats$agerange 20  0.7877 0.03938  1.1049    0.3406
## Residuals                       446 15.8984 0.03565
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The **univariate results** of the two-way MANOVA show that there are significant differences between the univariate means of each position and age range with respect to each of the six variables. However, the interaction between position and age range only predicts a significant difference in the means of the variables relating to distance traveled per game (distMI p=.05, distMIO p=.04, distMID p=.08) but not those pertaining to average speed.

```
##                                  Df  Pillai approx F num Df den Df
## nbastats$pos.f                    4 0.34804   7.0523     24   1776
## nbastats$agerange                 5 0.26974   4.2293     30   2225
## nbastats$pos.f:nbastats$agerange 20 0.27190   1.0585    120   2676
## Residuals                       446
##                                     Pr(>F)
## nbastats$pos.f                   < 2.2e-16 ***
## nbastats$agerange                 2.03e-13 ***
## nbastats$pos.f:nbastats$agerange    0.3174
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                                   Df   Wilks approx F num Df den Df
## nbastats$pos.f                     4 0.67740   7.5776     24 1539.7
## nbastats$agerange                  5 0.74881   4.4147     30 1766.0
## nbastats$pos.f:nbastats$agerange  20 0.75480   1.0625    120 2556.2
## Residuals                        446
##                                     Pr(>F)
## nbastats$pos.f                   < 2.2e-16 ***
## nbastats$agerange                3.421e-14 ***
## nbastats$pos.f:nbastats$agerange    0.3076
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The **multivariate results** provide a similar result: there are significant differences between the multivariate means of different positions and age ranges, but not of the interaction of the two.

### Contrasts

This first contrast shows that there is a statistically significant difference between point guards and centers in their average distance on offense. This was expected and supports the observations discussed above about the interaction plots.

```
## lm model parameter contrast
##
##    Contrast       S.E.     Lower     Upper    t  df Pr(>|t|)
## 1 0.1675234 0.04916456 0.0709144 0.2641324 3.41 471    7e-04
##
## Contrast coefficients:
##   (Intercept) pos.fPF pos.fPG pos.fSF pos.fSG
## 1           0       0       1       0       0
```

The second contrast compares small forwards and power forwards, and reveals that there is a statistically significant difference between them, though it is not nearly as significant as the difference between point guards and centers, which is what one would expect.

```
## lm model parameter contrast
##
##    Contrast       S.E.      Lower     Upper    t  df Pr(>|t|)
## 1 0.1426924 0.04817003 0.04803768 0.2373472 2.96 471   0.0032
##
## Contrast coefficients:
##   (Intercept) pos.fPF pos.fPG pos.fSF pos.fSG
## 1           0      -1       0       1       0
```

There is not a statistically significant difference between power forwards and centers. Therefore, as was seen in the interaction plot, there are two distinct types of players on offense, at least with regard to the amount of distance they cover. Point guards, shooting guards, and small forwards are very similar to each other and cover more distance on average than the other group of power forwards and centers.

```
## lm model parameter contrast
##
##       Contrast       S.E.      Lower      Upper     t  df Pr(>|t|)
## 1 -0.006269176 0.04844483 -0.1014639 0.08892556 -0.13 471   0.8971
##
## Contrast coefficients:
##   (Intercept) pos.fPF pos.fPG pos.fSF pos.fSG
```

```
## 1           0       1       0       0       0
```

Second set of contrasts, this time with respect to the average speed of a player on offense:

- **Point guards vs. shooting guards**

```
## lm model parameter contrast
##
##    Contrast       S.E.      Lower     Upper    t  df Pr(>|t|)
## 1 0.1491842 0.03806696 0.07438213 0.2239863 3.92 471    1e-04
##
## Contrast coefficients:
##   (Intercept) pos.fPF pos.fPG pos.fSF pos.fSG
## 1           0       0       1       0      -1
```

- **Shooting guards vs. small forwards**

```
## lm model parameter contrast
##
##     Contrast       S.E.       Lower     Upper   t  df Pr(>|t|)
## 1 0.03056452 0.03827214 -0.04464075 0.1057698 0.8 471   0.4249
##
## Contrast coefficients:
##   (Intercept) pos.fPF pos.fPG pos.fSF pos.fSG
## 1           0       0       0      -1       1
```

- **Power forwards vs. small forwards**

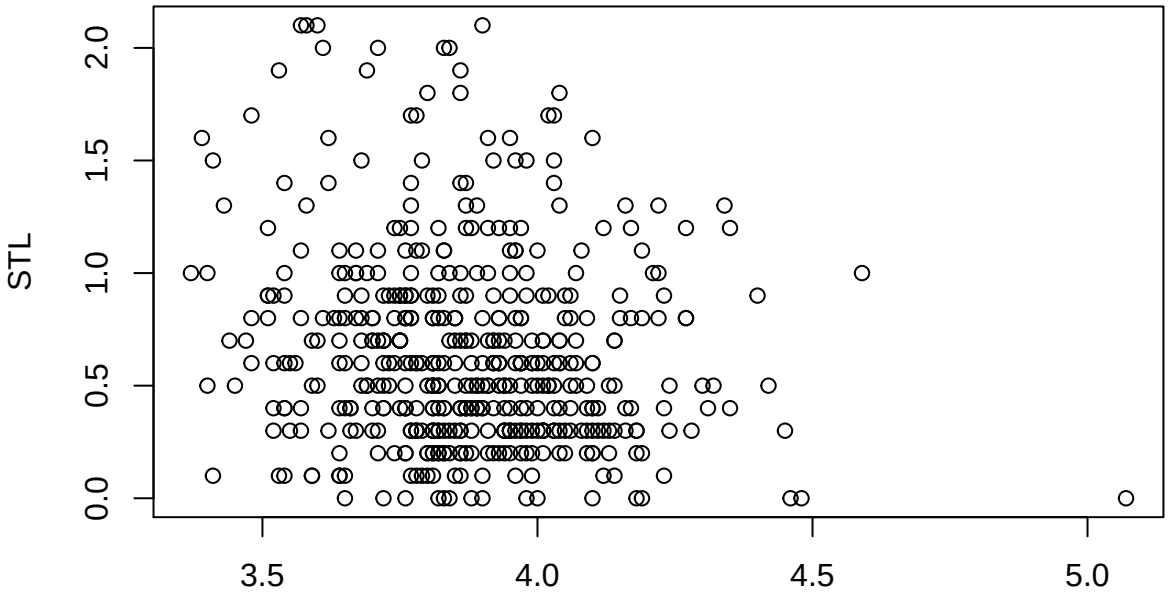
```
## lm model parameter contrast
##
##      Contrast       S.E.      Lower       Upper     t  df Pr(>|t|)
## 1 -0.06559885 0.03780312 -0.1398825 0.008684784 -1.74 471   0.0833
##
## Contrast coefficients:
##   (Intercept) pos.fPF pos.fPG pos.fSF pos.fSG
## 1           0       1       0      -1       0
```

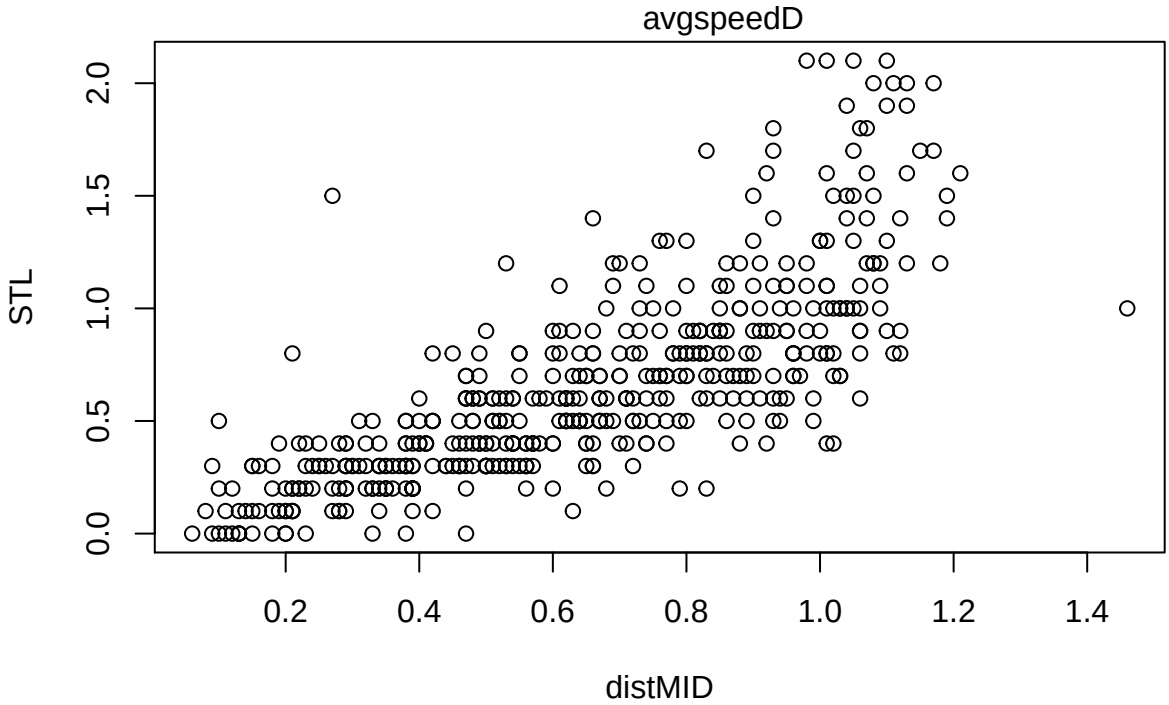
- **Power forwards vs. centers**

```
## lm model parameter contrast
##
##    Contrast       S.E.      Lower     Upper    t  df Pr(>|t|)
## 1 0.1018311 0.03801878 0.02712373 0.1765385 2.68 471   0.0077
##
## Contrast coefficients:
##   (Intercept) pos.fPF pos.fPG pos.fSF pos.fSG
## 1           0       1       0       0       0
```

From these contrasts, we can see that the positions are more clearly divided in average speed than in distance covered (previous set of contrasts). Point guards are in a class of their own, shooting guards and small forwards are indistinguishable, and power forwards and centers are each separate. An interesting conclusion that this finding leads to is that shooting guard and small forward are virtually the same position. From watching basketball games, one can see that their roles are very similar: neither of them are the primary ball handlers (other than certain exceptions like LeBron James) but both of them are usually stationed around the perimeter. Players are still often identified as one or the other because of long-standing convention, but once on the court it is essentially impossible to distinguish which player is occupying each of these positions. A player's average speed is by no means a perfect indicator of position, but it is interesting to see that similarities and differences observed by the naked eye are backed up by this statistic.

Now I will make plots to see whether there are linear relationships between my predictors (average speed and average distance on defense) and my response variable (steals).




Surprisingly, it is evident from these plots that there is a linear relationship between steals and distance, but not between steals and speed.

Next, I added assists as a covariate to the offensive model, and steals to the defensive one.

```
##  Response 1 :
##                    Df  Sum Sq Mean Sq F value    Pr(>F)
## nbastats$pos.f      4  2.7323  0.6831  12.298 1.631e-09 ***
## nbastats$agerange   5  2.5525  0.5105   9.191 2.356e-08 ***
## nbastats$AST        1 24.5346 24.5346 441.720 < 2.2e-16 ***
```

```
## Residuals         465 25.8277  0.0555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response 2 :
##                    Df  Sum Sq Mean Sq F value    Pr(>F)
## nbastats$pos.f      4  6.1375 1.53438 24.4763 < 2.2e-16 ***
## nbastats$agerange   5  2.8776 0.57552  9.1807 2.409e-08 ***
## nbastats$AST        1  0.5620 0.56199  8.9649  0.002899 **
## Residuals         465 29.1500 0.06269
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All three predictors show a statistically significant difference between the various univariate group means, both for distance traveled and average speed (both on offense).

```
##  Response 1 :
##                    Df  Sum Sq Mean Sq F value    Pr(>F)
## nbastats$pos.f      4  1.6856  0.4214  12.775 7.123e-10 ***
## nbastats$agerange   5  2.0602  0.4120  12.491 2.162e-11 ***
## nbastats$STL        1 20.6187 20.6187 625.069 < 2.2e-16 ***
## Residuals         465 15.3386  0.0330
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response 2 :
##                    Df  Sum Sq Mean Sq F value    Pr(>F)
## nbastats$pos.f      4  1.5285 0.38214  11.293 9.403e-09 ***
## nbastats$agerange   5  2.7160 0.54319  16.052 1.319e-14 ***
## nbastats$STL        1  0.9505 0.95045  28.087 1.795e-07 ***
## Residuals         465 15.7356 0.03384
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Similarly to the offensive MANOVA, this one (for defense) showed statistically significant differences between group means for each predictor.

```
##                    Df   Wilks approx F num Df den Df    Pr(>F)
## nbastats$pos.f      4 0.75864   17.180      8    928 < 2.2e-16 ***
## nbastats$agerange   5 0.82813    9.176     10    928 1.409e-14 ***
## nbastats$AST        1 0.50655  226.005      2    464 < 2.2e-16 ***
## Residuals         465
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The multivariate results of the offensive MANOVA reveal a difference in group means for different age ranges, positions, and number of steals, with respect to distance covered and average speed on offense.

```
##                    Df   Wilks approx F num Df den Df    Pr(>F)
## nbastats$pos.f      4 0.80087   13.622      8    928 < 2.2e-16 ***
## nbastats$agerange   5 0.76698   13.163     10    928 < 2.2e-16 ***
## nbastats$STL        1 0.42618  312.370      2    464 < 2.2e-16 ***
## Residuals         465
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Once again, multivariate group means appear to be different for the same predictors and response

variables as above, this time for defense.

# Discussion

In summary, these analyses provide a new categorization for players, as well as a way of distinguishing among player types.

- From PCA, we learned that variation among players is primarily divided into two components: overall skill (as measured by minutes played, two-point attempts and conversions, free throw attempts and makes, defensive rebounds, and personal fouls), and size (as measured by field goal and two point %, three point and free throw %, three pointers attempted and made, and offensive rebounds).
- Cluster analysis showed that the traditional division of players by position does not capture or describe their impact on a basketball game. Instead, players who are listed as playing different positions may indeed have similar impacts as measured by the similarity in their performance across the statistics that comprised my dataset. Speificially, cluster analysis revealed that the players can be best divided into three groups, that I interpret as big men, high usage players (players who tend to dominate the ball or be involved in a majority of the team's production), and secondary players (those who are not the dominant players on their respective teams).
- The MANOVA analysis ignored the individual player distinctions and instead explored differences between average speed and distance across positions and age. It was revealed that point guards and centers are at extremes with the other positions in the middle with respect to speed and distance, with points guards being the fastest and most traveled. However, the effect was less clear for distance. A clear effect was also detected for age, with players becoming slower as they aged. On the other hand, distance experienced a peak in the middle of a player's career (between the ages of 25 and 30). The effect of age was roughly uniform for speed across positions, but varied by position for distance traveled.

On the one hand, these results are not very surprising. Talent shows up as the first principal component, which means that those players who play the most during games produce the best results. However, the divisions yielded by the cluster analysis suggest that general managers might be better off acquiring the best players available within these clusters rather than necessarily filling their rosters with players of the five traditional positions. Finally, the MANOVA provides a way of assessing player performance as they age, and might provide guidance for deciding how much longer in their career a player might be able to make an impact on the court.

## Points for Further Analysis

- Based on the MANOVA analysis, it would be interesting to explore whether the change in a player's average speed and distance over time relative to the average of his position is predictive of his future contract length and salary or of other measured statistics, both during the observed time and in the future.
- Another path for further analysis might be to go into greater depth about a certain aspect of the game. For example, I could use more detailed data about a player's shooting tendencies in terms of location, time of game, remaining time on the shot clock, and how the shot was taken (with a teammate's assist or a play created alone).