CHICAGO BOOTH

The University of Chicago Booth School of Business

BUSN 41201 - Spring 2020

Big Data

Veronika Ročková

NBA Teams and Players

Exploring the characteristics of teams and players and predicting the possibility of wining and going to playoffs

Hanyang Jiang

Helen Jin

Hongji Liu

Michael Zhao

We pledge our honor that we have not violated the Booth Honor Code during this assignment

I. Executive Summary

In this report, we tried to use both dimension reduction and variable selection techniques to understand the characteristics of the National Basketball Association (NBA) teams and players, as well as to study the relationship between these characteristics and the qualification of the playoffs, i.e. being the top 8 teams in each conference (East and West), ranked in order by win-loss records. Since the NBA announced changes to the format of the NBA playoffs beginning with the 2017 NBA playoffs, we restricted ourselves to the 2017/18 season.

First, we used clustering to identify different groups of NBA teams with different characteristics. As an initial step, we used K-means approach to group NBA teams regarding their 3-pointer shooting capabilities. Although we found the Houston Rockets, the Golden State Warriors standing out from all the other teams, there seems not to be a strong correlation between the clusters and the probability of going to the playoffs. We then expanded the clustering using all meaningful variables. We still found the Houston Rockets being special but the component of most clusters changed quite a bit from the one only considering 3-pointer capabilities. In addition, through cluster regression (R-squared 52%), we also found the cluster, consisting of the Golden State Warriors, the Cleveland Cavaliers, the Houston Rockets and the Toranto Raptors, has a higher chance to go to the playoffs. At the same time, we used lasso to analyze what factors are more influential in determining whether a team will enter the playoffs or not. We gained a lasso model that have R-squared at around 0.5, and we found that steals, offensive rebounds, defensive efficiency are the most important factors in making a prediction.

Second, we attempted to investigate what factors are the most significant predictors for the outcome of each game. We split this into two questions, the final score of each game and whether a team wins each game, and performed lasso analysis. For the binary outcome (winning or losing) of the game, with the variables we have, we were able to create a lasso model that has an R-squared value over 0.5, meaning that these variables can explain considerably high amounts of information of the outcome. We then used different approaches such as AICc, BIC and CV to select our lambda value, and were able to figure out the coefficients of these variables, with number of steals and defensive rebounds being the two most significant predictors. We then split data into a training set and a test set and gained a prediction accuracy of 86.6%. We also performed lasso on the final scores of each game, and found that the number of 3 points and assists play the most important role for a team to gain a high final score.

Third, we used clustering methods to analyze the player data. We would like to see whether players can be divided into different groups and how their groups look like. It's interesting to see if the group has more to do with a player's position or simply is determined by his skills. It turns out that players' clusters have little to do with their position, and we can see the most excellent players tend to be in the same group, e.g. James Harden, LeBron James, Russel Westbrook. After that, we compared the result of k-means with hierarchical clustering and got the similar result. Besides, we are also interested in the effect of road games and home games. Is it possible that players perform differently in these different situations? Also, the result turns out that there exists some difference, but not huge. Then we also studied how a player's performance will affect his chance of getting into the playoff. Of course, the better he performs, the more chance he'll have to reach that. But the lasso shows the effect is really small, which means the final result still relies heavily on the whole team's performance. Also we checked the causal effect of single stats on the chance of getting into the playoff. One significant factor - "defensive rebound" turns out to have no causal effect, while the most significant factor - "score" still possesses a small effect on that.

Finally, we used principal component analysis to reduce the dimensions of our data, and did the analysis on the revised data. Our objective is still to look for the relationship between player stats and their chance of getting into the playoff. We took the first three principal components and did regression on them. The result shows that only the first principal component affects the outcome. So the whole data can be summarised into a single stats, which stands for the chance for each player. And we also took a look at the components of the first principal component, it appears that defensive rebounds and score are the most important factors. We also did lasso analysis on the player statistics. However, we realized that the R-squared value with numerical variables we had in the model is less than 0.1, meaning that these variables can barely be used to make a prediction. So based on the PCA analysis, if a player wants to add his chance of getting into the playoff, he needs to perform well both in scoring and getting defensive rebounds.

II. Introduction to Dataset and Analysis

The data for this analysis is sourced from BigDataBall (https://www.bigdataball.com/). The data consists of game level data (2,460 games) for both teams and players from NBA games for the 2017-2018 season. Regular season and playoff games are included in the dataset; All-Star games are excluded since they are not relevant for playoff qualification.

The dataset captures the some of the most important metrics evaluating both teams and players, including Final Scoring Results, Field Goal Made and Attempted, Three Point Field Goals Made and Attempted, Rebounds, Assists, Fouls, Steals, Turnovers, Blocks, Time Played, Efficiency, Rest Days, Referees, as well as odds from gambling websites. Based on the original dataset, we also constructed many variables to facilitate our analysis, including the Team Type and Player Type, which are based on our clustering analysis, as well as some factors based on the PCA analysis.

We restricted our analysis to the 2017/18 season only, for better comparability (before the 2017/18 season, the rules to qualify as playoff teams were different). Even so, our dataset is sufficient for the analysis as it contains the records of 30 teams and 541 players in 2,460 games. Our research questions follow:

A. What types of teams are more likely to go to the playoffs? For example, did teams with the most advanced "small ball" concept, i.e. being good at 3-pointer shooting, have a bigger chance to make it to the playoffs?

B. What are the most significant predictors on the outcome of each game?

C. Can we divide players into different clusters? Are those clusters capable of predicting whether a player can attend the playoff?

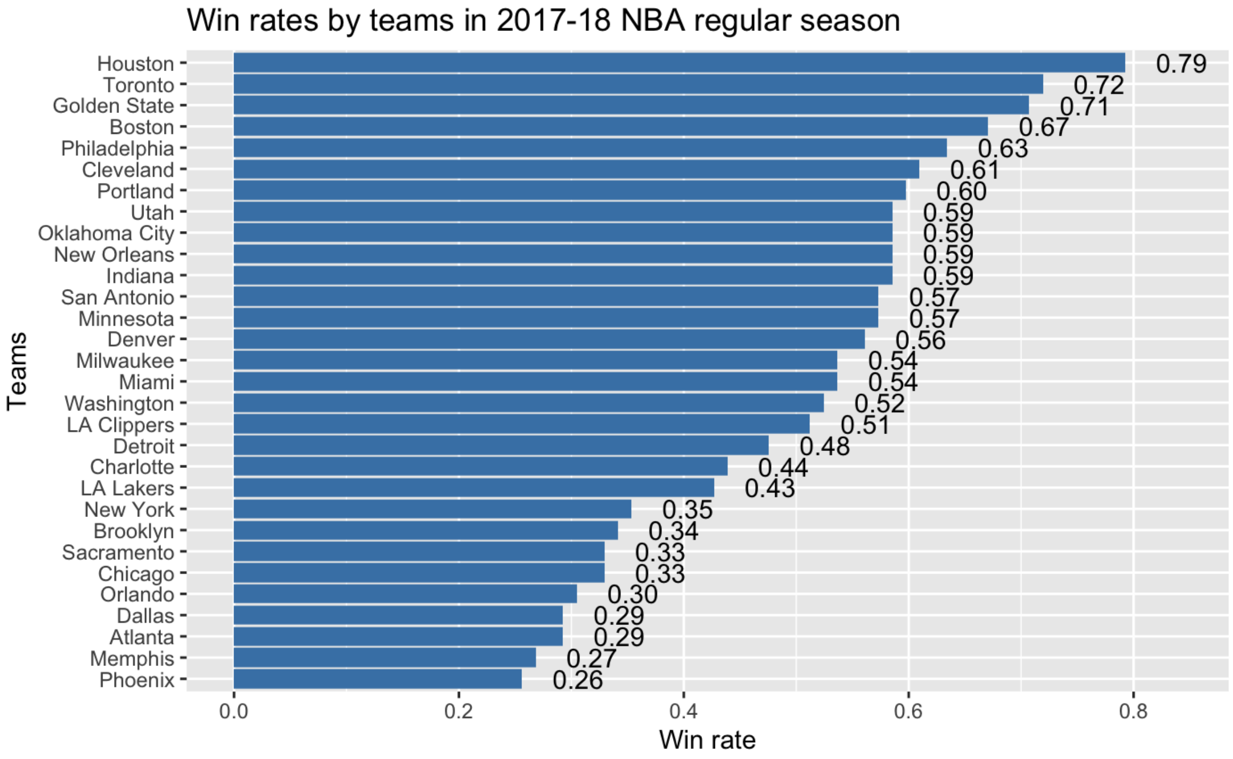
D. What skills should a player possess in order to get to the playoff?

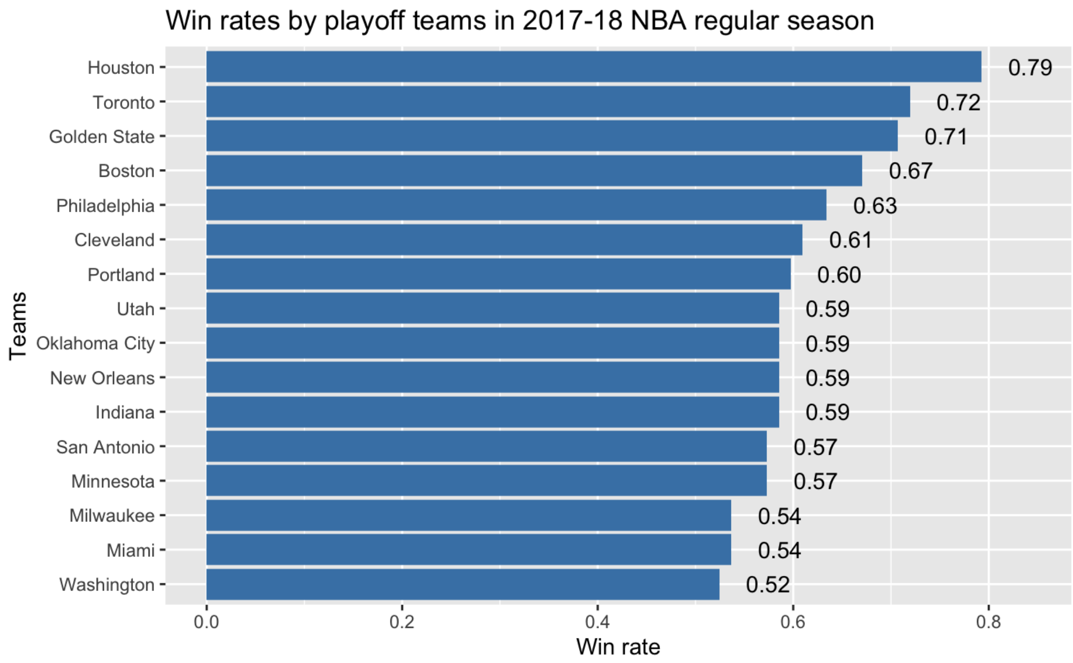
Below, we will conduct some exploratory analysis to better understand this dataset, and then address each of the research questions in order. We will end by sharing some preliminary conclusions from our study.

III. Data Summary

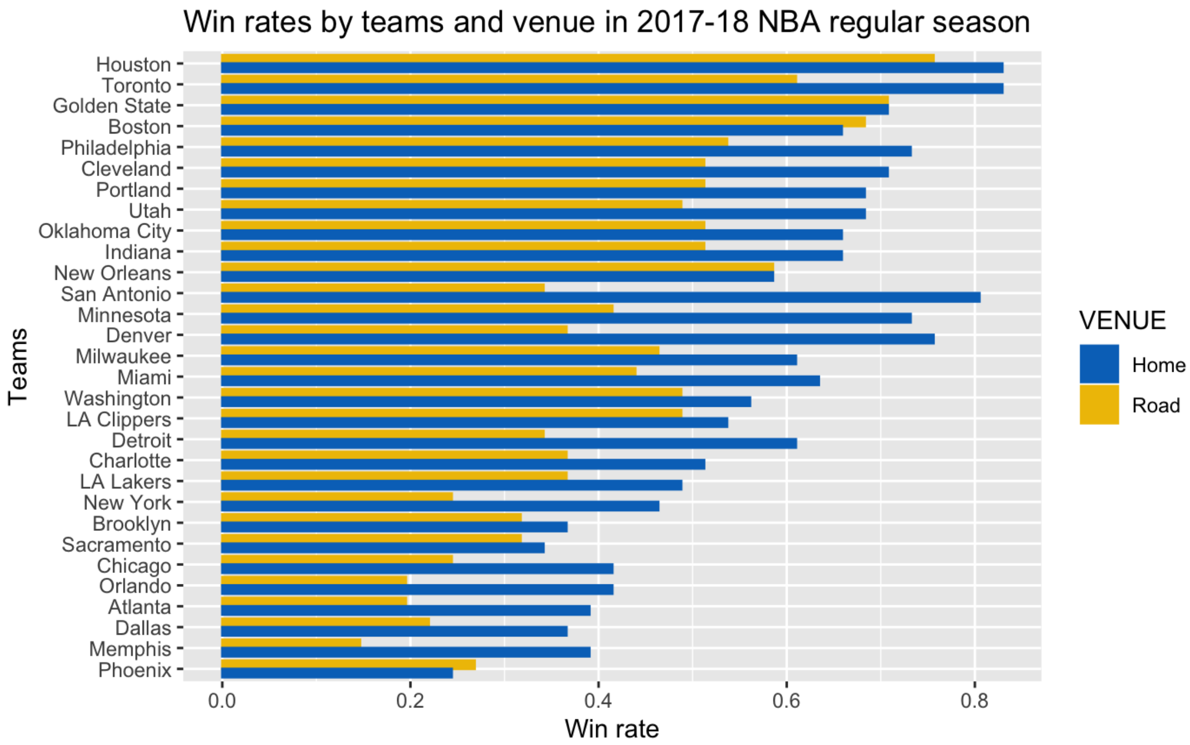
A. Win Rates

Within the team data, the win rates among teams are between 25.6% and 79.3%. The team with the highest win rate is Houston, and that with the lowest win rate is Phoenix; both belong to the Western Conference. If we look at the teams that made the playoffs, we can see that Denver, with its win rate in the top 16, did not make the playoffs. The reason is that Denver is in the Western Conference, and there are already 8 teams in the Western Conference that have better win rates. As a result, even though Denver's win rate beats Milwaukee, Miami and Washington, it being in the Western Conference unfortunately made it miss the playoffs, thereby showing that Western Conference are more competitive than the Eastern Conference, especially in the upper bracket.


Win rates by teams in 2017-18 NBA regular season

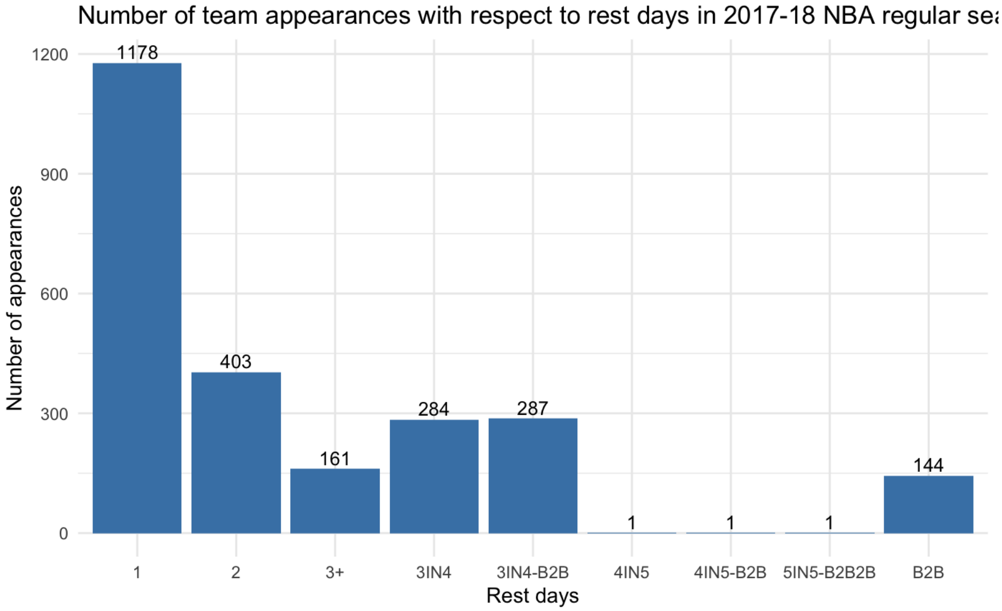


Looking at the win rates separated by the venue, most of the teams have better win rates when they are at home compared to when they are away, with exceptions of Golden States and New Orleans that share even win rates across venue, and Boston and Phoenix that even have higher win rates in away games than in home games.



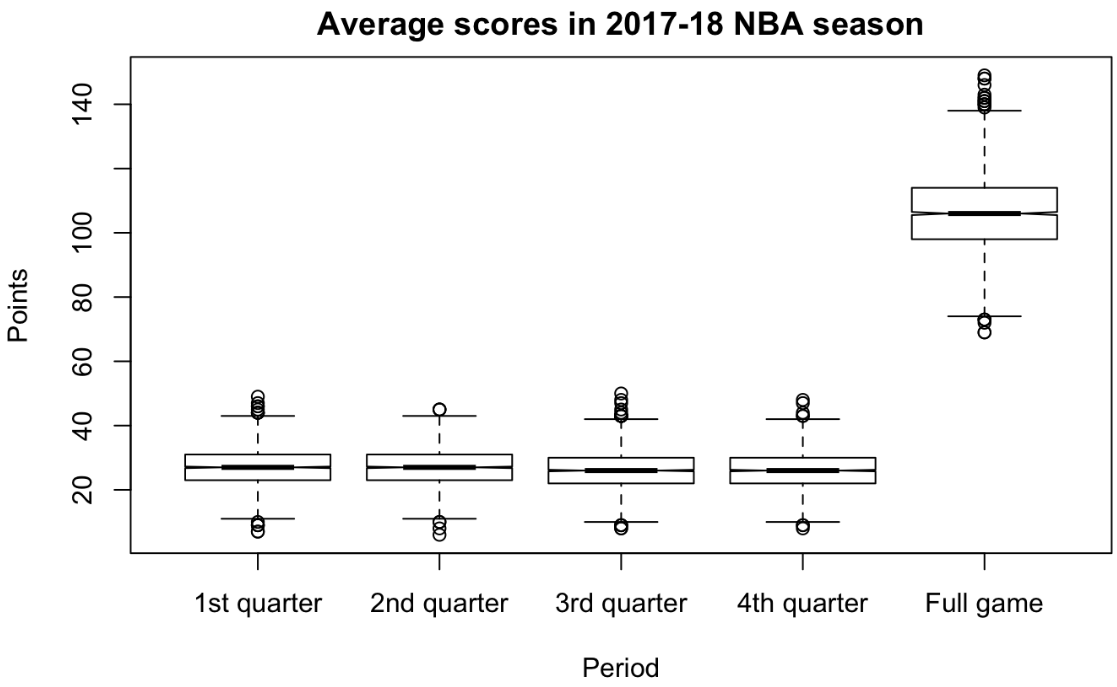
B. Rest Days

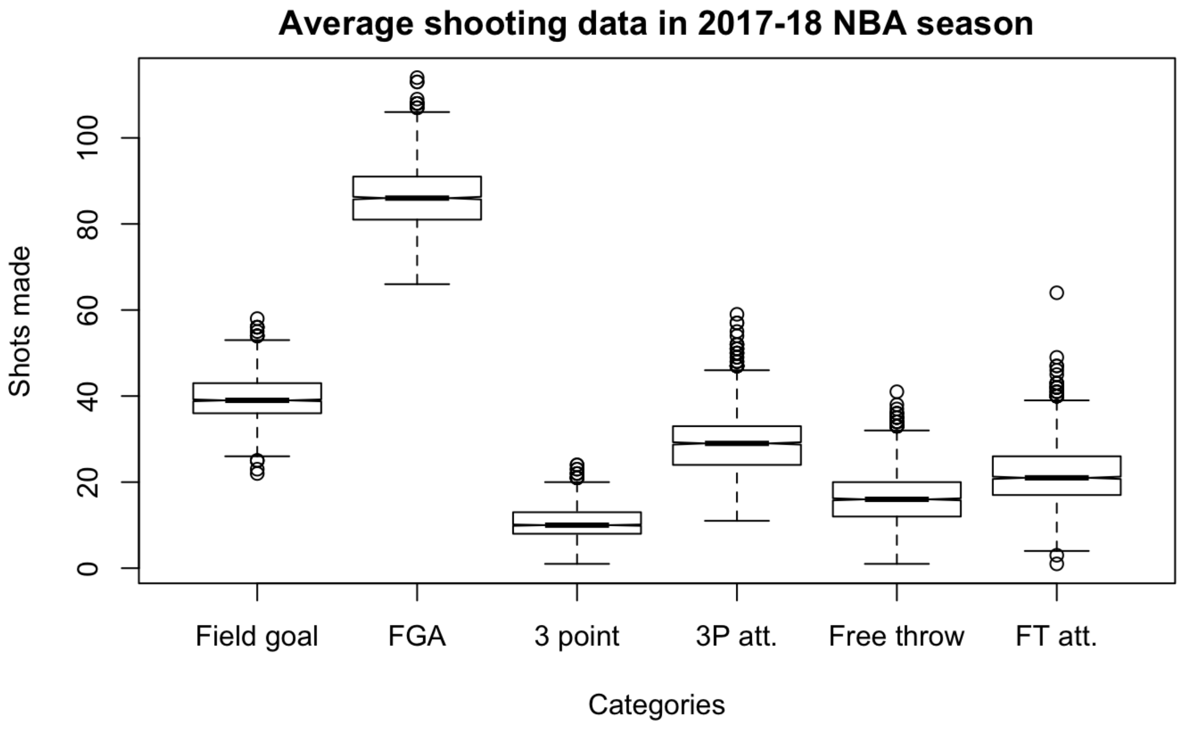
Among all the games that are played in the regular season, which is 1260 games in total and two teams per game give us 2520 matchups in total. Among these team appearances, the majority of them (1178) have 1 rest day before the game, while the rest are fairly scattered. Notice that there is 1 team that played a back-to-back-to-back game with 5 games in a row. This team, New Orleans, won this game against the LA Lakers, which had a two-day rest before. As a result, it spikes the interest of finding what covariates are significant in predicting the result of a game since New Orleans was a playoff team while Lakers failed to make the playoffs.

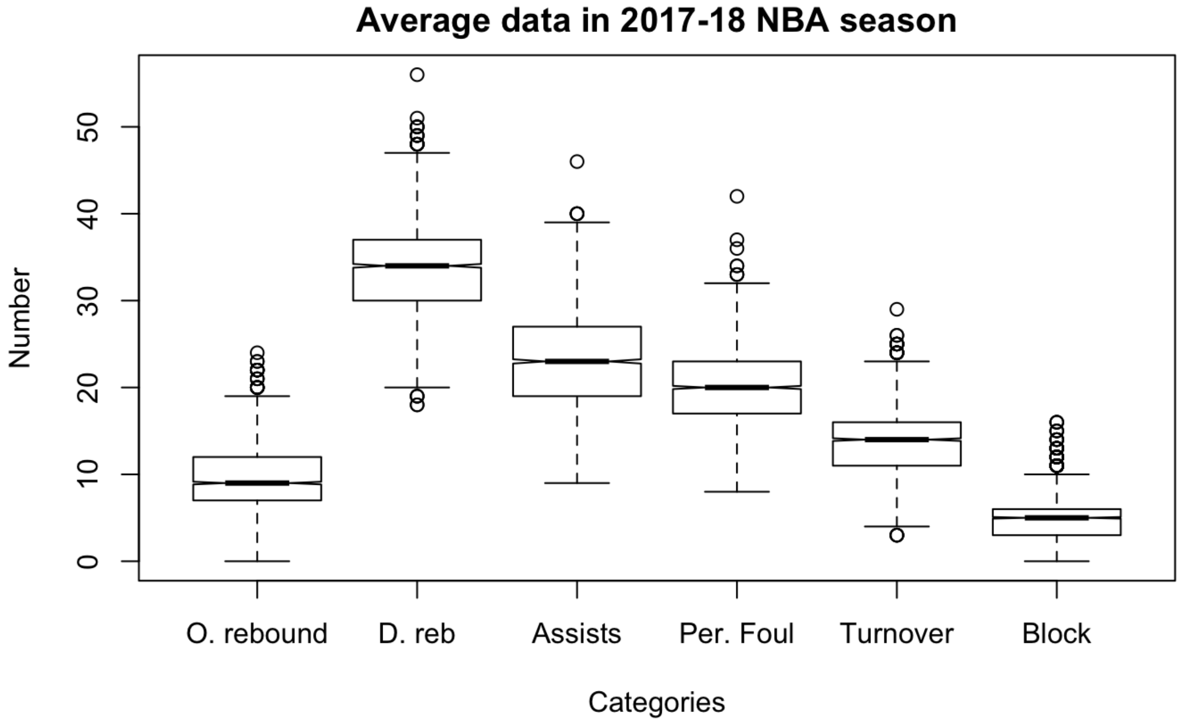


C. Team Stats

The score data suggests that each team scored at a median of 100 points, with each quarter at a median around 25 points. The shooting data suggests that although teams attempted far more 2-point field goals than 3-point field goals, in terms of percentiles, they are roughly the same at around 50%. The free throw percentage, on the other hand, is much higher.


Average scores in 2017-18 NBA season
Points
20
40
60
80
100
140
1st quarter
2nd quarter
3rd quarter
4th quarter
Full game
Period

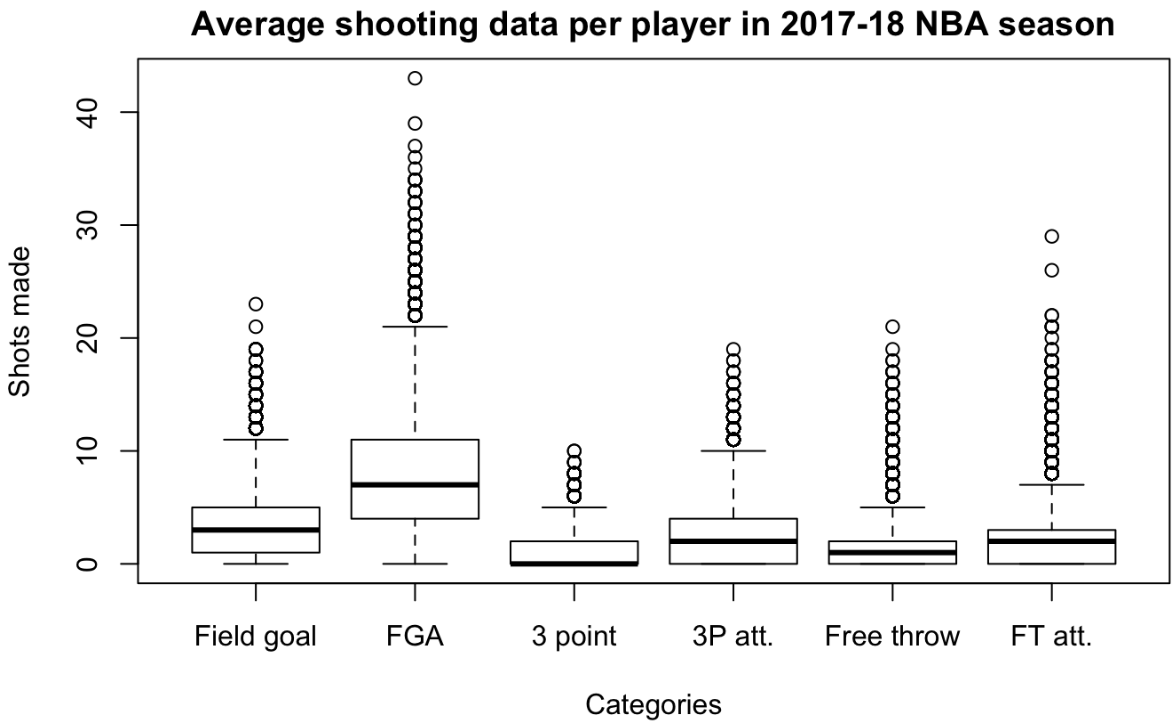

Average shooting data in 2017-18 NBA season
Shots made
0
20
40
60
80
100
Field goal
FGA
3 point
3P att.
Free throw
FT att.
Categories

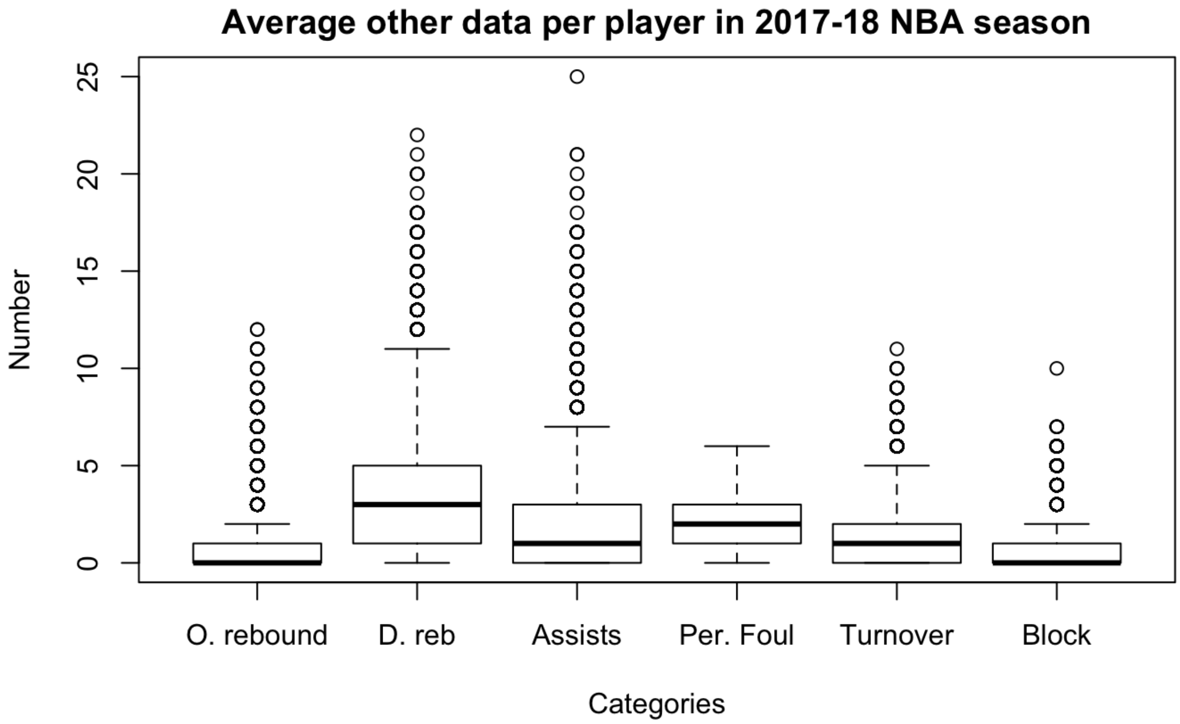


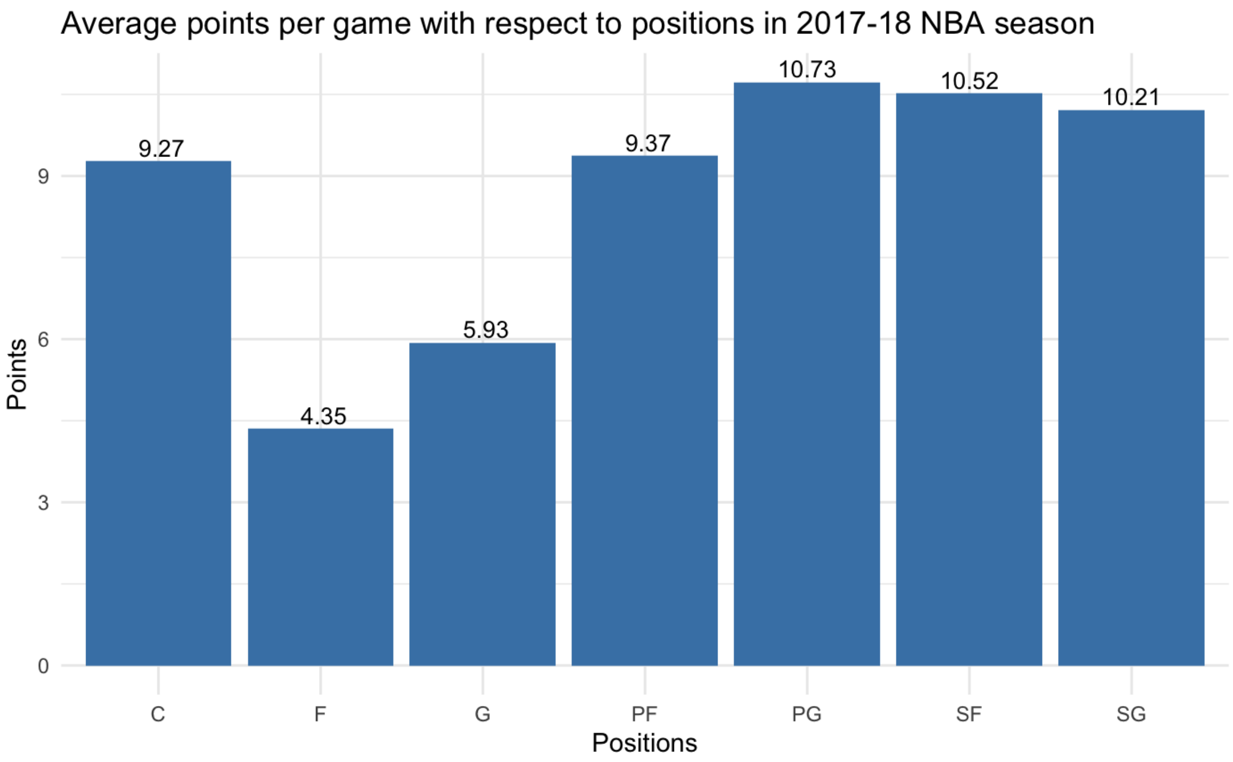
D. Player Stats

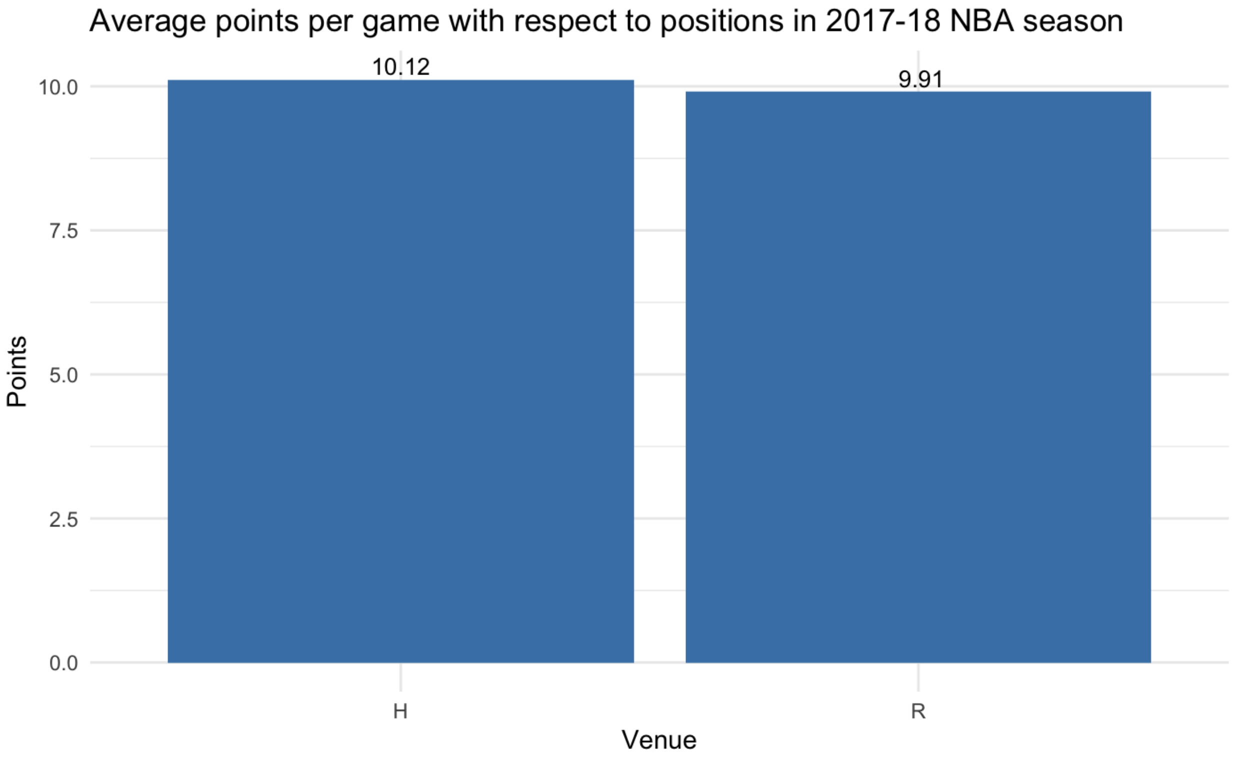
Comparing player data with the team data over the same categories, there are far more variations in the player data than in the team data since there are far more players than teams. There are a few shocking outliers in the data. Some players have nearly 30 free throw attempts in one game, which is 10 more than the median of the team data. In addition, some players had 25 assists in a game, which is higher than the team median. The only data that does not have outliers is the personal foul, since each player can only have at most 6 fouls per game and he is fouled out.

If we take a look at the points data, with the average points from all players around 10 points, we can see that when we group players by positions, there are slight variations between them (we want to not include F and G since there are not many data points of them, and they are not one of the traditionally defined positions). Centers the lower average points while point guards have the highest average scores. With PG, SF and SG's scores higher than that of C and PF, this might suggest that the play style of NBA in that season focuses more on smaller players in those positions. Grouping players by venues does not yield a significant advantage for home games, with a 0.2 difference.


Average shooting data per player in 2017-18 NBA season
Shots made
0
10
20
30
40
Field goal
FGA
3 point
3P att.
Free throw
FT att.
Categories


Average other data per player in 2017-18 NBA season
Number
0
5
10
15
20
25
O. rebound
D. reb
Assists
Per. Foul
Turnover
Block
Categories


Average points per game with respect to positions in 2017-18 NBA season
Points
0
3
6
9
9.27
4.35
5.93
9.37
10.73
10.52
10.21
C
F
G
PF
PG
SF
SG
Positions


Average points per game with respect to positions in 2017-18 NBA season
Points
0.0
2.5
5.0
7.5
10.0
10.12
9.91
H
R
Venue

IV. Research Question 1: What types of teams are more likely to go to the playoffs?

This analysis aims at figuring out the shared characteristics among NBA teams and how these characteristics correlate with the probability of going to the playoffs. We first built clusters and Hierarchical clusters based on intuition, for example, distinguishing teams with stronger 3-point field goals shooting capabilities to those with weaker 3-pointer shooters. And then we tried to build clusters based on all meaningful variables, without any pre-assumptions on which variables matter more. In the meanwhile, we used AICc and BIC tools to help us identify the optimal numbers of clusters. Finally, we ran cluster regressions and compared the results suggested by the two methods.

A. Data Preparation

First of all, we transform the data by separating home games and away games, as NBA teams usually play quite differently in these two scenarios. Then we averaged the regular season for each team over the following variables: 1Q,2Q,3Q,4Q,F, MIN, FG, FGA, 3P, 3PA, FT, FTA, OR,...Rest Days. Do this by home vs away (road) games, so that we have the indicators of each team's performance and strength.

Based on the original datasets, we also constructed several new variables to facilitate our analysis. First, we constructed variables TP_Hit_rate and FG_Hit_rate, indicating an NBA team's shooting capabilities. Second, constructed WIN to indicate the final result of each game for each team. Third, we created a new variable PLAYOFF, indicating whether a team eventually made it to the playoffs. But on the other hand, we removed some existing variables, which are defined and calculated by third parties, e.g. the odds of each game from different gambling websites. These variables are actually derived from the other basic variables and some of them are calculated using the formulas we don't know. Thus, to keep simplicity and reduce redundancy, we decided to get rid of these variables.

B. Clustering with intuitions – distinguishing teams based on 3-point goals

We want to understand in which way NBA teams are different with each other, but on the other hand, what characteristics do they share? A recent trend that is reshaping the NBA is the trend of "small ball". More and more teams tend to use height, physical strength and low post offense/defense in favor of a lineup of smaller players for speed, agility and increased scoring (often from the three-point line). Therefore, we first looked into clusters by 3-pointer shooting capabilities.

We plot each NBA's team average 3-point field goal hit rate against the average number of 3-point field goals attempted, in both home games and road games, in Figure 4.1 and Figure 4.2 respectively. From the chart of home games (Figure 4.1), we can clearly see some outstanding teams, e.g. Houston Rockets, which had a significantly higher number of 3-point field goals attempted; and Golden State Warriors, which was far better than all the other teams in terms of 3-point field goal hit rate. In addition, teams like the LA Lakers and Phoenix Suns were neither good at 3-pointers nor attempting to shoot 3-points.

However, figure 4.2 tells a different story. Golden State Warriors become more conservatinve in 3-pointers than they were in home games, with a lower average hit rate. The New York Knicks also did a much worse job in road games. Therefore, it would be necessary to separate teams in home games and road games to make it more comparable.

Figure 4.1: Avg. 3-point field goal attempted v.s. 3-point field goal hit rate (Home)

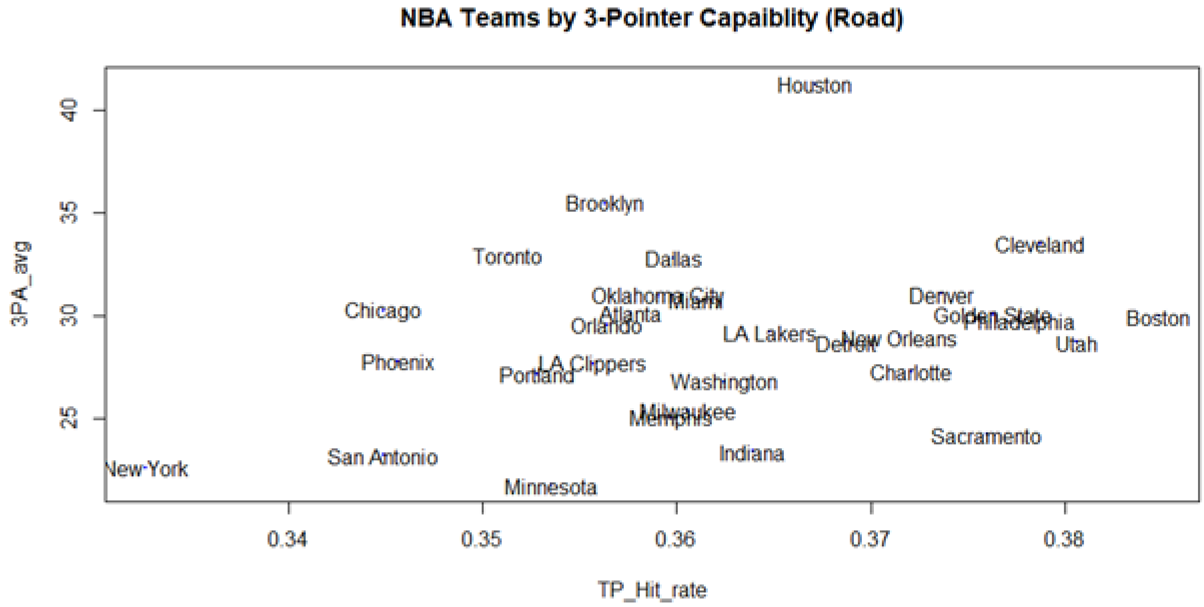
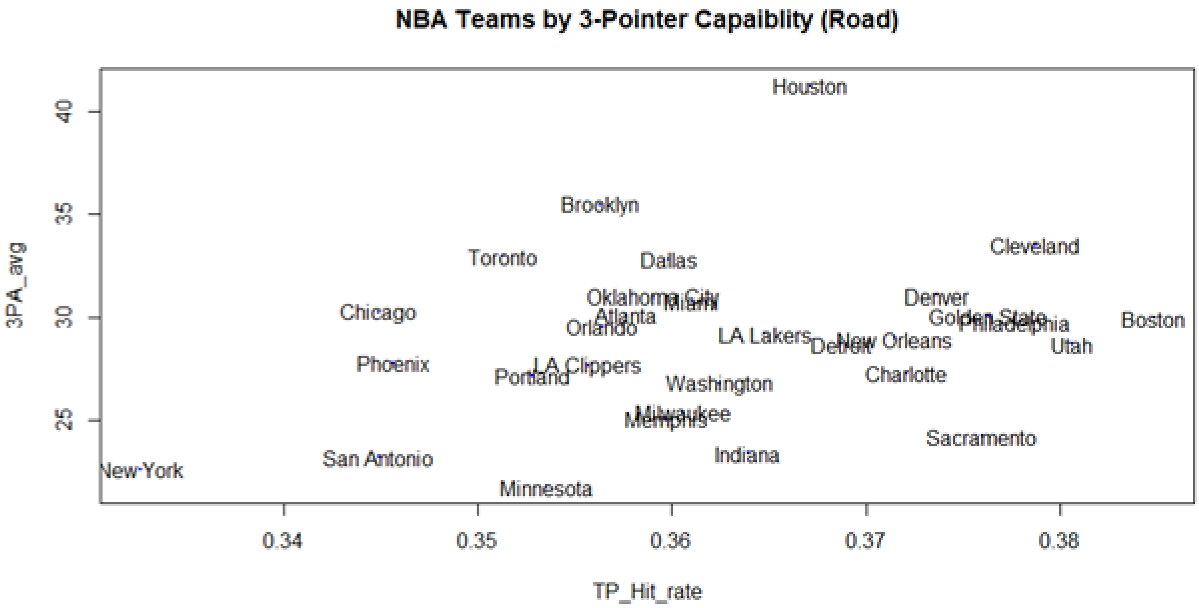


Figure 4.2: Avg. 3-point field goal attempted v.s. 3-point field goal hit rate (Road)



To further understand the clusters of NBA teams by 3-pointer capabilities, we first Now, applied the K-means clustering with 5 clusters and 1000 repeats, for both home games and road games respectively. The results are plotted in Figure 4.3 and Figure 4.4. In both home and road games, Houston Rockets stood out from all the others – there's only one team in its cluster.

Figure 4.3： Clustering by 3-Pointers (Home)

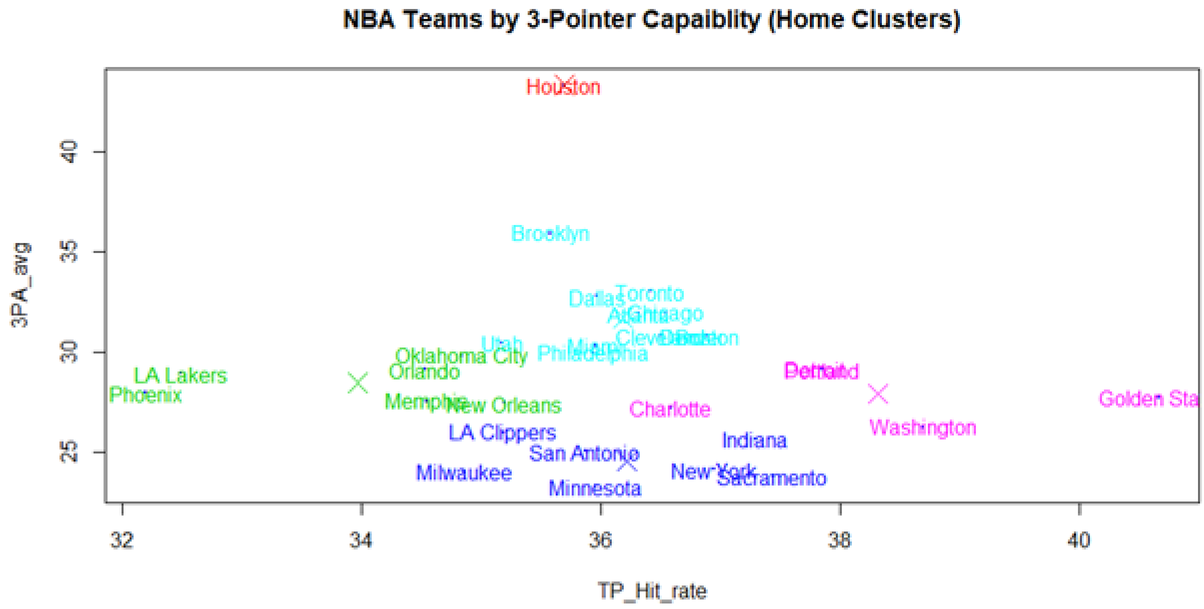
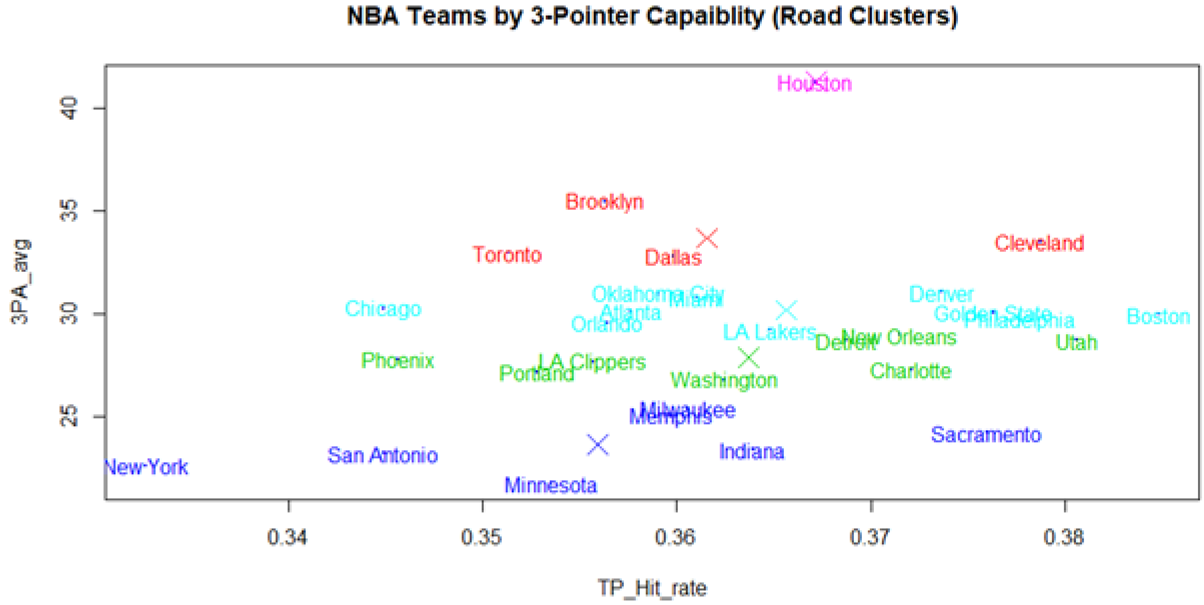


Figure 4.4: Clustering by 3-Pointers (Road)



We also did the Hierarchical clustering for both home and road games against 3-pointer capabilities. The hierarchical clustering suggests that, if we take only 5 clusters, the Brooklyn Nets, the Houston Rockets and the Golden State Warriors are so different that they should each occupy a cluster in home games. And in road games, the hierarchical clustering also suggests Houston Rockets really differentiates itself from all other teams.

Figure 4.5: Hierarchical clustering by 3-pointer (Home)

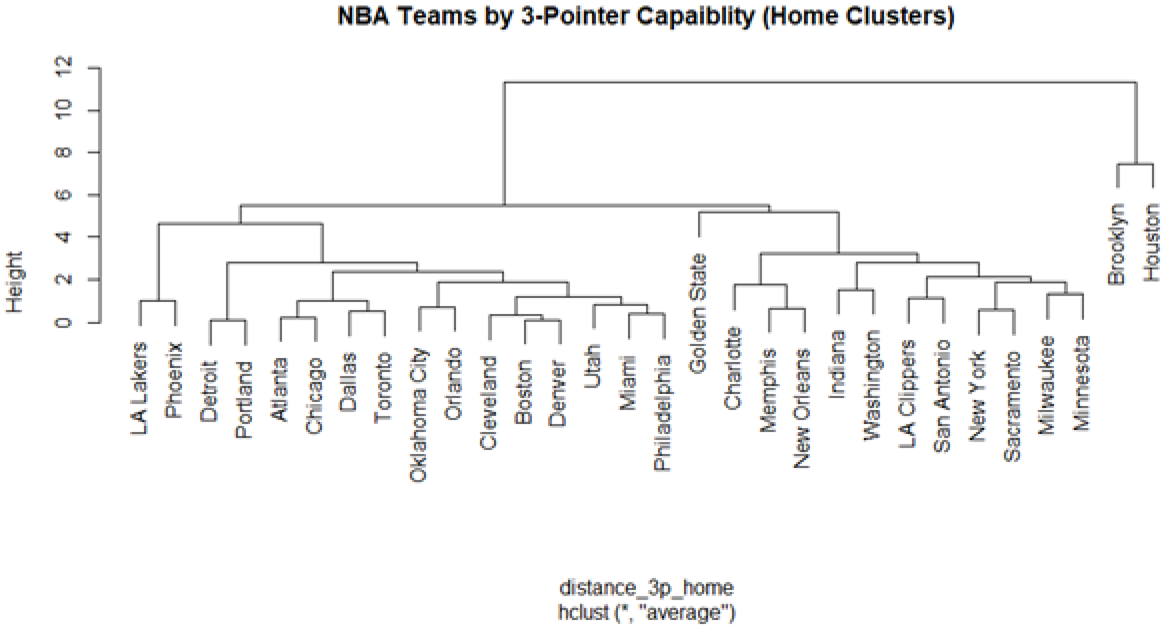
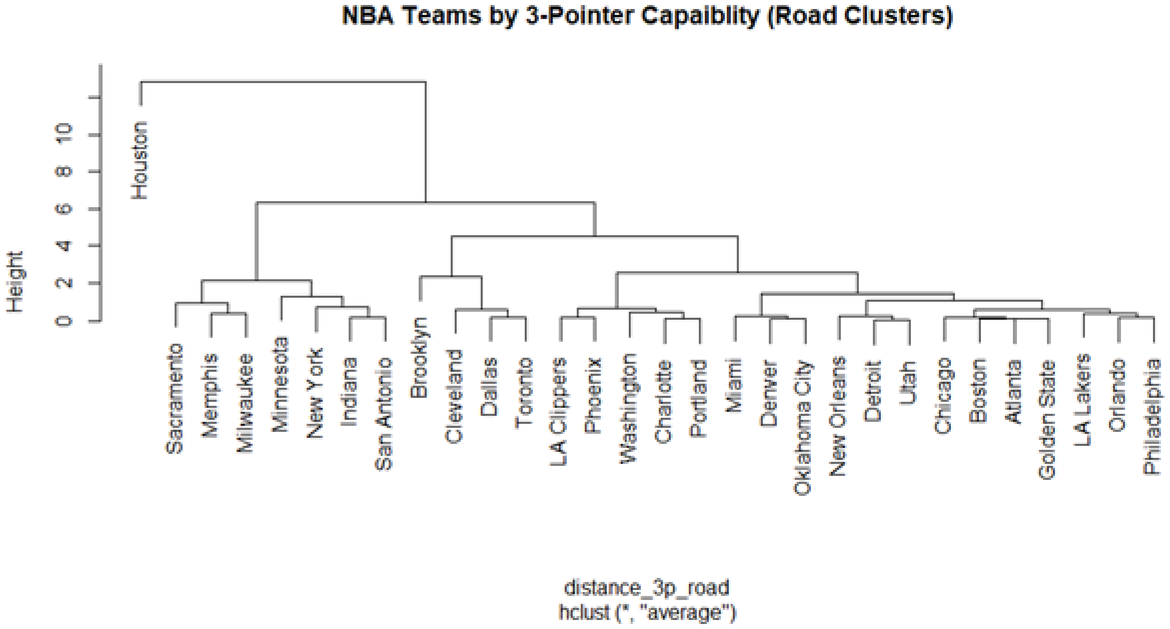


Figure 4.6: Hierarchical clustering by 3-pointer (Road)



However, we just randomly picked K=5. Is it a good guess? To answer this question, we used AICc and BIC approaches to come up with an arguably more accurate K. From the chart below (Figure 4.7), we can see the BIC approach suggested a K equals 7 rather than 5. But is K=7 better than K=5? We then re-calculated K-means clustering with K=7 and summarized the result in Figure 4.8. Since the chart on the RHS clearly shows a more accurate clustering than the chart on the LHS, we will keep K=7 going forward.

Figure 4.7: AICc and BIC analysis of clustering by 3-pointer capabilities (Home)

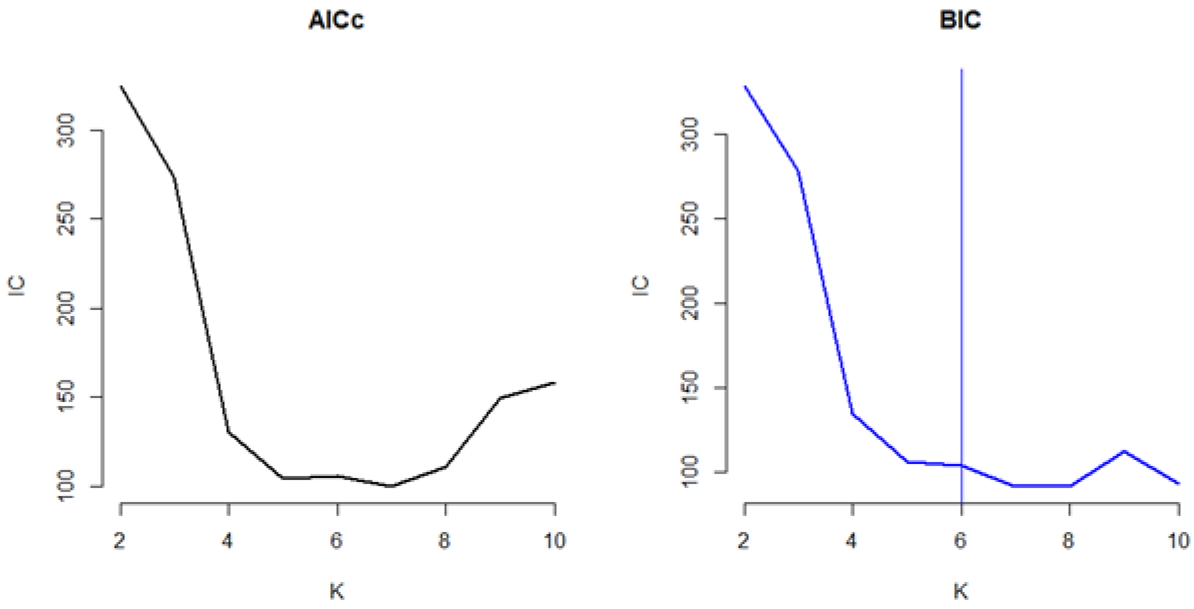
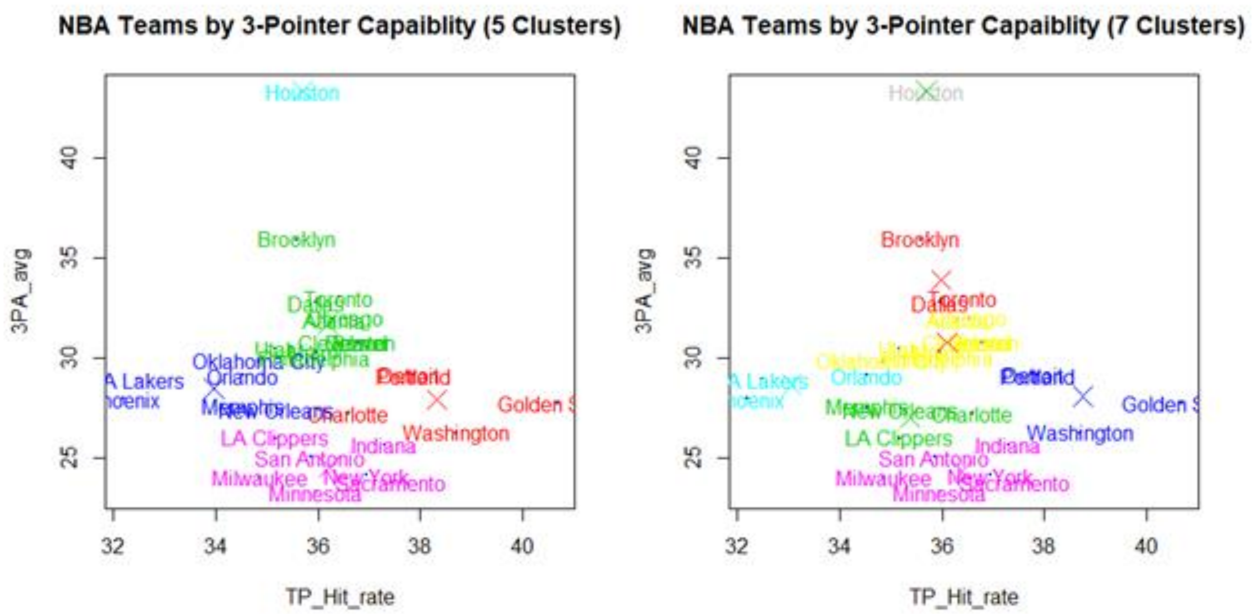


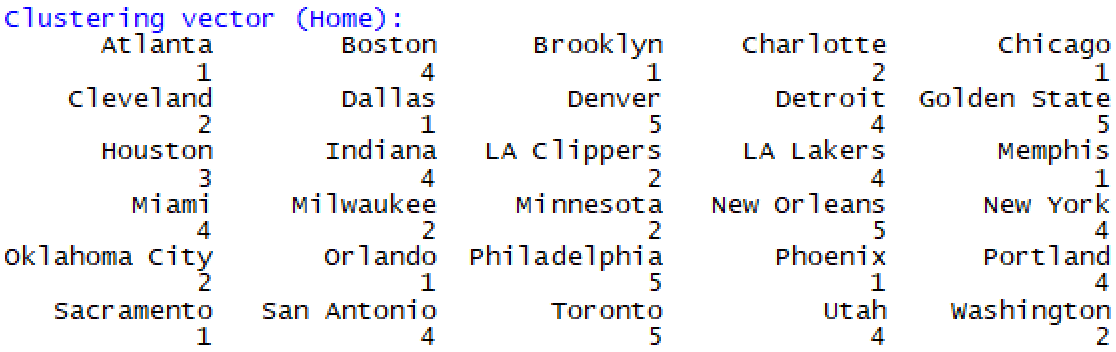
Figure 4.8: Comparison of clustering with different K values (Home):



C. Clustering with all meaningful variables

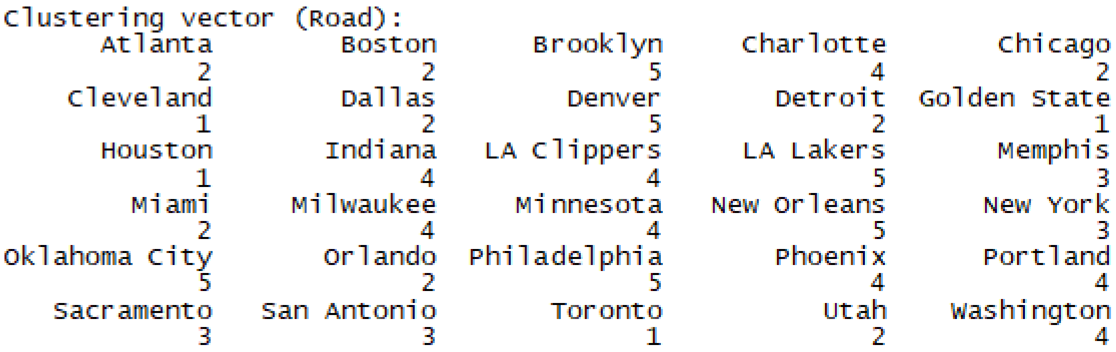
We then run the clustering based on all meaningful variables in the dataset, including scores made in each quarter, total points made, variables indicating field goals capabilities, depending and offending capabilities as well as team efficiencies, with a randomly picked K=5 and 1000 repeats. The results are shown in Figure 4.9 and Figure 4.10. Figure 4.9 tells us that Houston Rockets are still so special in home games even if we consider many other factors. While, the results here in road games are quite different from what we had in previous analysis when only 3-pointer capabilities are considered. Figure 4.10 shows Golden State Warriors, Cleveland Cavaliers, Houston Rockets and Toronto Raptors all belong to the same cluster. After looking into the stats of the above-mentioned teams, this cluster consists of the best teams in the league – good at both defending and offending, with a high team efficiency.

Figure 4.9: Clustering of NBA teams in home games:



```
Clustering vector (Home):
      Atlanta        Boston      Brooklyn     Charlotte       Chicago 
            1             4             1             2             1 
    Cleveland        Dallas        Denver       Detroit  Golden State 
            2             1             5             4             5 
      Houston       Indiana    LA Clippers     LA Lakers       Memphis 
            3             4             2             4             1 
        Miami     Milwaukee     Minnesota   New Orleans      New York 
            4             2             2             5             4 
Oklahoma City       Orlando  Philadelphia       Phoenix      Portland 
            2             1             5             1             4 
   Sacramento   San Antonio       Toronto          Utah    Washington 
            1             4             5             4             2 
```

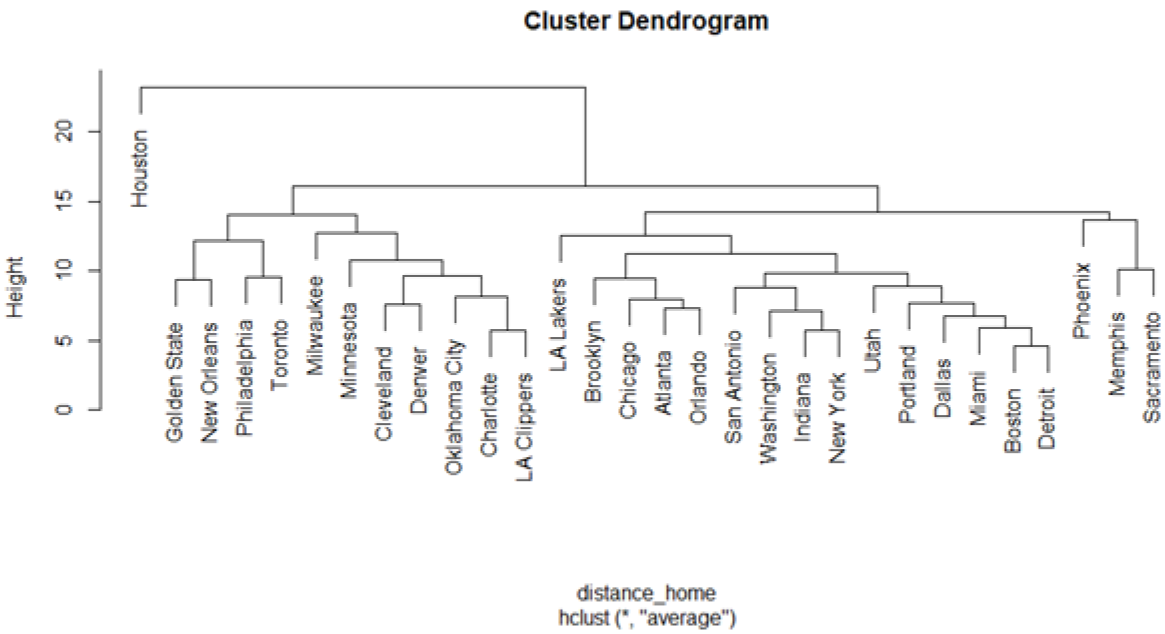
Figure 4.10: Clustering of NBA teams in road games



```
Clustering vector (Road):
      Atlanta        Boston      Brooklyn     Charlotte       Chicago 
            2             2             5             4             2 
    Cleveland        Dallas        Denver       Detroit  Golden State 
            1             2             5             2             1 
      Houston       Indiana    LA Clippers     LA Lakers       Memphis 
            1             4             4             5             3 
        Miami     Milwaukee     Minnesota   New Orleans      New York 
            2             4             4             5             3 
Oklahoma City       Orlando  Philadelphia       Phoenix      Portland 
            5             2             5             4             4 
   Sacramento   San Antonio       Toronto          Utah    Washington 
            3             3             1             2             4 
```

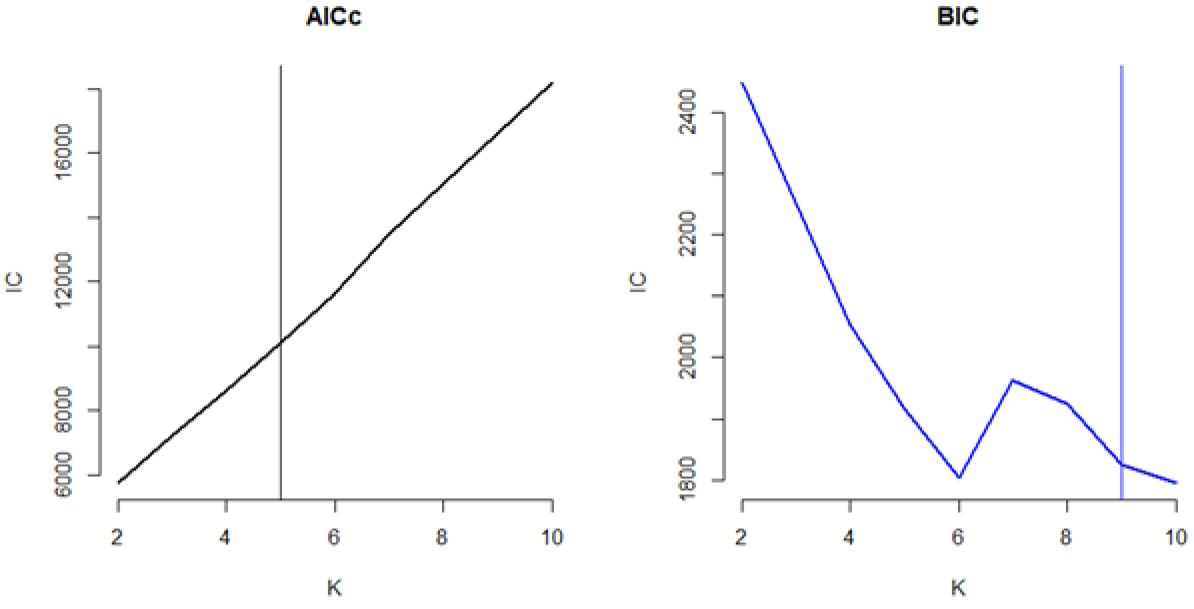
We then did a hierarchical clustering with all variables (Figure 4.11). Again, Houston Rockets are so outstanding. But we could also find some discrepancies, for example, the simple K-means clustering suggests Denver and Golden States belong to the same cluster, however the hierarchical clustering would suggest Golden States are closer to New Orleans.

Figure 4.11: Hierarchical clustering of NBA teams in home games (all variables)



As we mentioned above, K=5 is randomly picked. We then did AICc and BIC analysis to verify our hypothesis. Figure 4.12 shows AICc approach suggests a K=5 while BIC approach suggest a K=9. Since both AICc and BIC are only roughly right here, we will keep K=5 going forward.

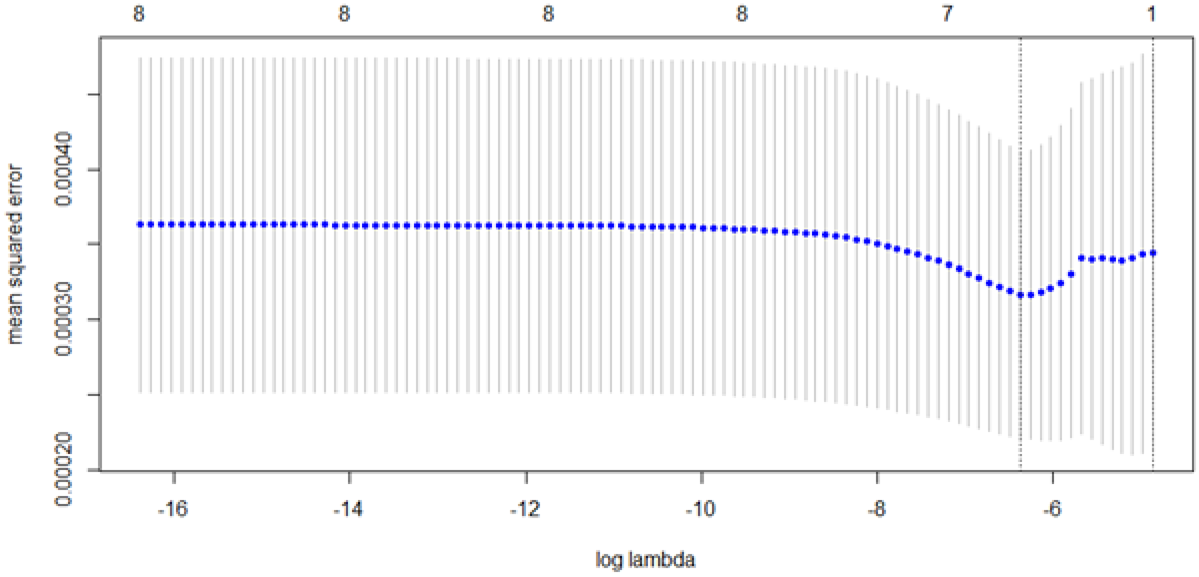
Figure 4.12: AICc and BIC – All variables (Home)



D. Cluster regression:

Finally, we run the cluster regression to understand the correlation between clusters and the probability of going to the playoffs. We first ran the cluster regression which only considered the difference in team's 3-pointer capabilities. However, the R-square of this regression is only 0.0807865, which means the probability of going to playoffs can't be well explained by the clusters only considering 3-pointer capabilities.

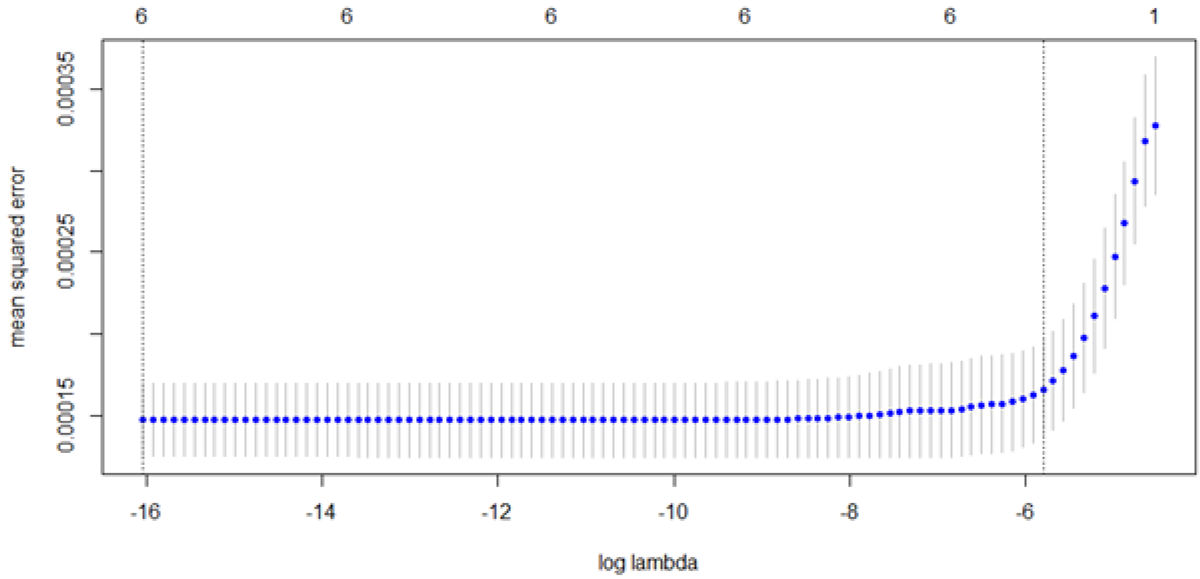
Figure 4.13: Deviance – Clustering by 3-pointer capabilities (Home)

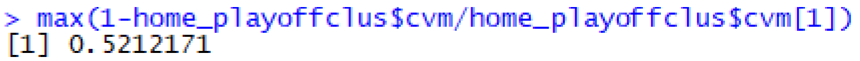


```
> max(1-home_3pt_playoffclus$cvm/home_3pt_playoffclus$cvm[1])
[1] 0.0807865
```

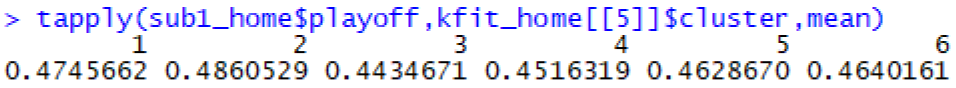
Then we also run the cluster regression considering all variables, with a much promising result. The R-square of the regression stood at 0.5212171. It also shows the 2nd Cluster, including Golden State Warriors, Cleveland Cavaliers, Houston Rockets and Toranto Raptors have a higher probability to go to the playoffs.

Figure 4.14: Deviance – Clustering by all variables (Home)





```
> max(1-home_playoffclus$cvm/home_playoffclus$cvm[1])
[1] 0.5212171
```



```
> tapply(sub1_home$playoff,kfit_home[[5]]$cluster,mean)
        1         2         3         4         5         6 
0.4745662 0.4860529 0.4434671 0.4516319 0.4628670 0.4640161 
```

E. Data Preparation for LASSO fit

Apart from clustering, we also used LASSO to select significant variables and see how much these variables can explain the result. We only used the team data for the regular season. We first group all the teams by team name and their venue, then for each team playing at home and road, we calculate the average of each numerical variable. Lastly, we performed LASSO to select variables with different criteria involving AICc, BIC and CV.

F. LASSO analysis

We have in total 26 numerical variables that represent the average of each teams' performance in factors such as defensive efficiency, turnovers, steals to fit the lasso model. With a minimum AICc, we obtained an R squared value of 0.57917, corresponding to a lambda value of 0.02721279, while minimizing BIC gives an R squared value of 0.4001179 with selected lambda 0.08706098. So these

26 variables give a pretty good explanation of the team's chance of getting into the playoffs. The six most significant predictors are given in the table below, and they are average steals, offensive rebounds, defensive efficiency, offensive efficiency, blocks, estimate of number of possessions and field goals attempted.

Figure 4.15: LASSO Plot

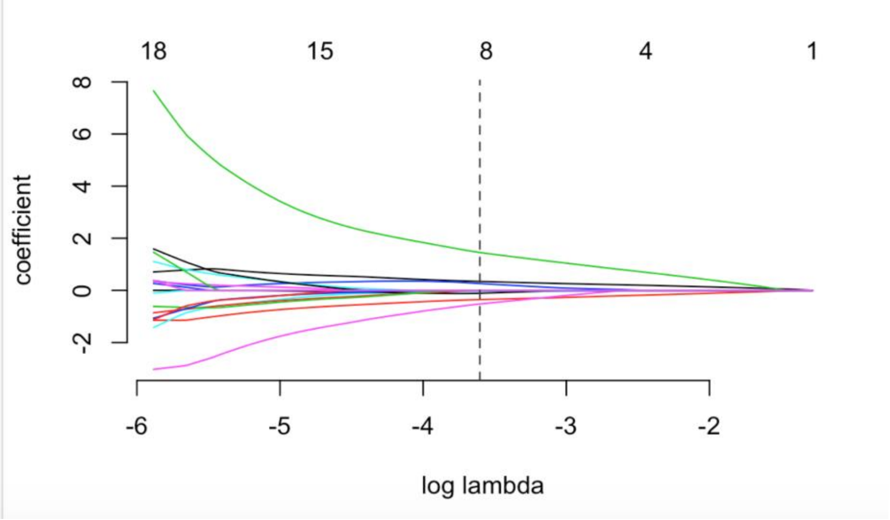
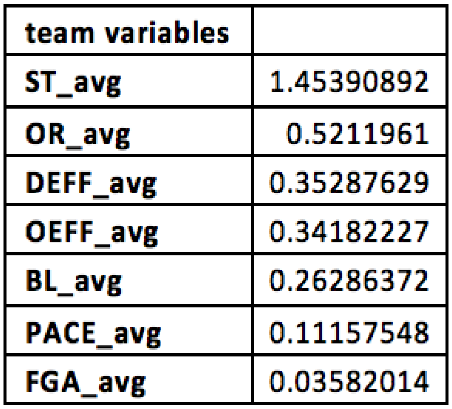


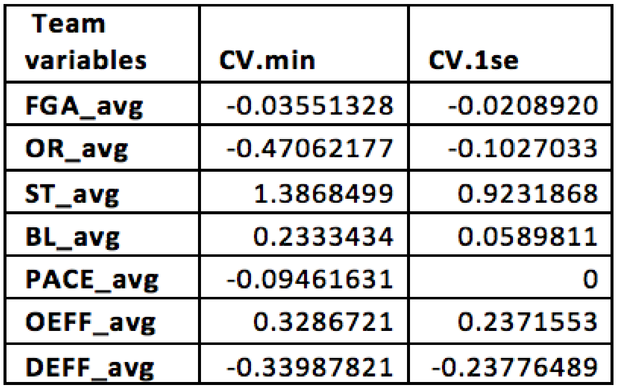
Table 4.1: Coefficient of LASSO fit



team variables	
ST_avg	1.45390892
OR_avg	0.5211961
DEFF_avg	0.35287629
OEFF_avg	0.34182227
BL_avg	0.26286372
PACE_avg	0.11157548
FGA_avg	0.03582014

We then perform the cross validation for gamma lasso penalty selection. The CV.min picks a lambda value of 0.029866, and CV.1se picks a lambda value of 0.06000775. The variable selected and their coefficients are presented below. CV.min picks one more variable than CV.1se, but both select similar variables as AICc above.

Table 4.1: Coefficient of CV fit



Team variables	CV.min	CV.1se
FGA_avg	-0.03551328	-0.0208920
OR_avg	-0.47062177	-0.1027033
ST_avg	1.3868499	0.9231868
BL_avg	0.2333434	0.0589811
PACE_avg	-0.09461631	0
OEFF_avg	0.3286721	0.2371553
DEFF_avg	-0.33987821	-0.23776489

V. Research Question 2: What are the most significant factors that determine the result of each game?

In this analysis, we will decompose the result of each game into two factors, the final score of each team and whether a team wins its opponent, and then investigate what variables would play the most important role. We first select the numerical variables that may be used to fit into a LASSO regression by running lasso regression on each individual factor and check R square value. After variables are selected, we perform LASSO and carry out the weight of each variable to see how they affect the result of the game. We then compare the different lambda and weight of factors based on different criteria such as AICc, BIC, CV.min, and CV.1se. Finally, we also carry out the accuracy of the LASSO model by splitting the data to test set and training set to see what percentage of the result are predicted correctly for whether a team wins its opponent.

A. Data Preparation

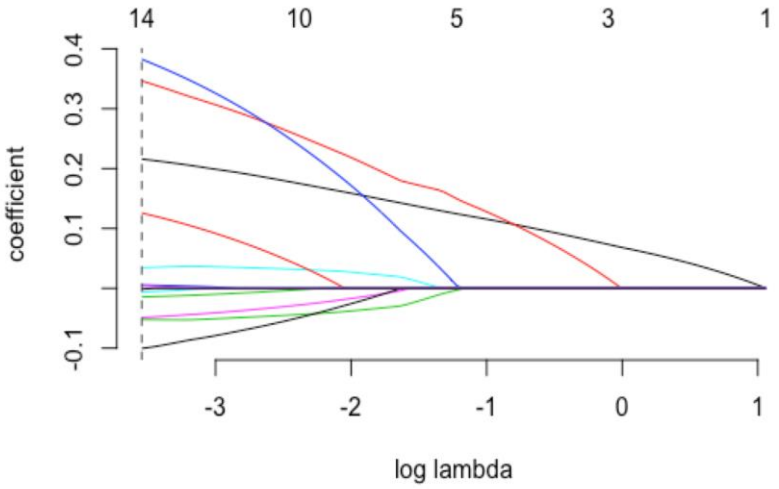
We want to perform LASSO analysis on whether the team will win each game in order to select the most influential factors that determine the result of each game. We used the team data including the ones in both regular season and playoffs. We first carry out the result of each game by comparing the final score of two teams in one game and create a column recording the result with 1 being winning and 0 being losing, so we will choose binary when we fit LASSO. Then we eliminate the factors that will directly influence the result of the game such as the final score and defensive efficiency. These factors are determined by performing LASSO with a single variable, and check the R squared value. If R squared values is higher than 0.9, we will eliminate this factor, since such a high R square value means that most of the information is obtainable for predicting the result using this single factor, and is not an ideal predictor we want to include in our LASSO model. Yet, 22 other variables are included as predictors for our model, including 3 points made, steals, offensive rebounds, quarter time scores and so on. The complete list of variables will be presented in the table below.

B. Binary Outcome (winning or losing)

B1. select variables with AICc and BIC

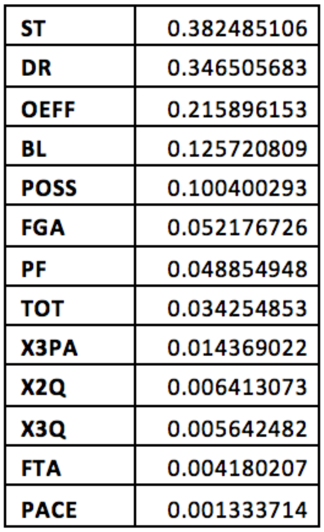
We will first study how factors determine whether a team wins the game. With these 22 inputs, we obtained a LASSO model that has R squared value 0.5504869, which is selected by minimizing AICc. The lambda selected is 0.02887238. The LASSO plot is shown below. We can see that from left to right, as the log lambda are getting larger and the penalty keeps increasing, less variables are kept in the model. If, alternatively we choose lambda by minimizing BIC, then we will have a lambda value of 0.03319622 that corresponds to a R squared value 0.5484366.

Figure 5.1: LASSO plot on whether a team wins a game



We rank the significance of each factor from high to low, and it is shown in the table below. We can see that the 9 factors are eliminated in the model, and number of steals and defensive rebounds being the two most important factors that determine the result of the game in this model. In this model, all the factors have a positive effect in assisting the team winning.

Table 5.1: Coefficient of LASSO fit



ST	0.382485106
DR	0.346505683
OEFF	0.215896153
BL	0.125720809
POSS	0.100400293
FGA	0.052176726
PF	0.048854948
TOT	0.034254853
X3PA	0.014369022
X2Q	0.006413073
X3Q	0.005642482
FTA	0.004180207
PACE	0.001333714

B2. Cross validation for gamma lasso penalty selection

We then perform the cross validation for gamma lasso penalty selection. The CV.min picks a lambda value of 0.0002581648, and CV.1se picks a lambda value of 0.009064661. So the lambda value of these three criteria follows that BIC > AICc > CV.min > CV.1se, where smaller and smaller penalties are followed. Below, the graphs of CV fit and coefficients select for CV.min and CV.1se are shown. CV.min eliminates 10 factors, while CV.1se only eliminates 2 variables. Different from the lasso model above, we see both positive and negative effects of each factor here, where defensive rebounds and steals still stand as the two most significant factors. For some factors such as X3Q, two selection methods have opposite conclusions on whether the factor has a positive or negative effect on the winning result.

Figure 5.2: binomial deviance VS log lambda for cross validation lasso

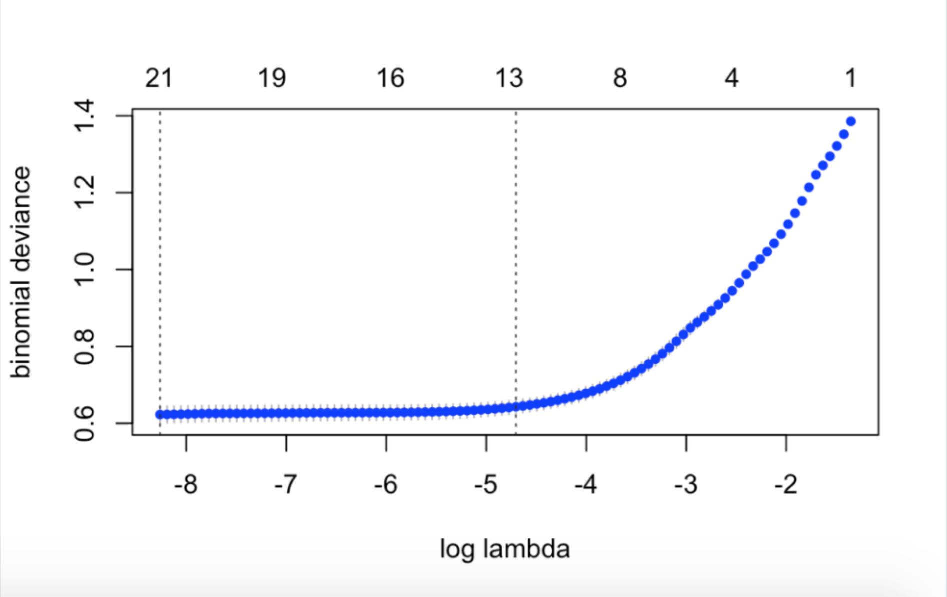
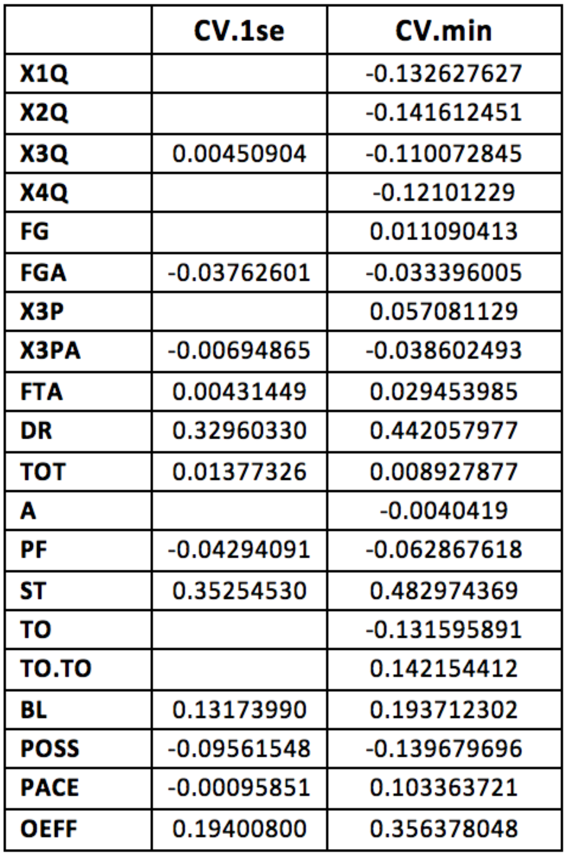


Table 5.2: Coefficient of CV fit



	CV.1se	CV.min
X1Q		-0.132627627
X2Q		-0.141612451
X3Q	0.00450904	-0.110072845
X4Q		-0.12101229
FG		0.011090413
FGA	-0.03762601	-0.033396005
X3P		0.057081129
X3PA	-0.00694865	-0.038602493
FTA	0.00431449	0.029453985
DR	0.32960330	0.442057977
TOT	0.01377326	0.008927877
A		-0.0040419
PF	-0.04294091	-0.062867618
ST	0.35254530	0.482974369
TO		-0.131595891
TO.TO		0.142154412
BL	0.13173990	0.193712302
POSS	-0.09561548	-0.139679696
PACE	-0.00095851	0.103363721
OEFF	0.19400800	0.356378048

B3. Model Prediction and Accuracy

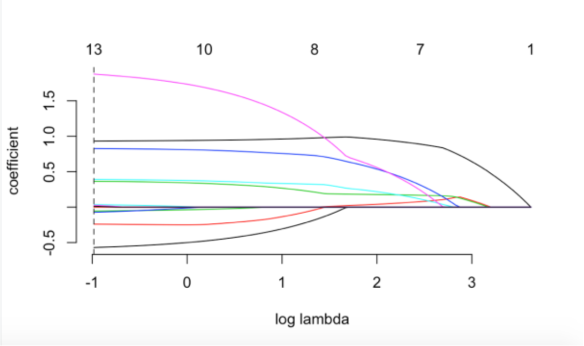
We then split the data set into training set and testing set with a proportion of 80% to 20%, set the seed as 120. We train our lasso model with the training set and make predictions on the test set, and compare the predicted result with the observed result. We get an accuracy of 86.6%, meaning this model is a pretty accurate fit for our data set.

C. Numerical Outcome (Final Score for the game)

We are also interested in investigating how each factor influences the final score. We fit LASSO with both Poisson and Gaussian. This time, we only keep 16 inputs, and the selection rule is the same as above. R squared for Gaussian is 0.7307123, and for Poisson it is 0.7229452. Poisson has a larger lambda compared to Gaussian. The plots show the same trend for Gaussian or Poisson in terms of how each variable’s weight changes as we tune lambda for penalty, yet Poisson returns fewer selected variables. However, the significance of each variable's influence on final score does not differ by much comparatively for two fits, and the number of 3 points and assist are the two main factors for high final score. In Gaussian, the number of 3 points is more highly valued. Below are the two plots for LASSO and the variable selected.

Figure 5.3: Lasso with Gaussian

Figure 5.4 Lasso with Poisson



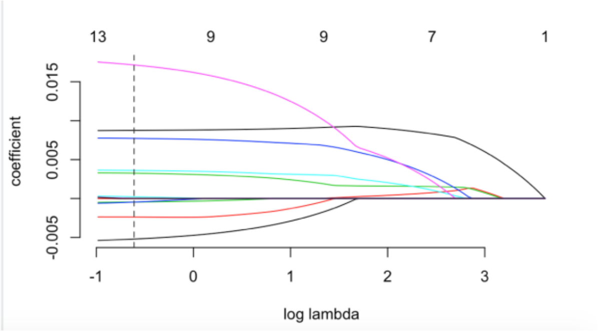
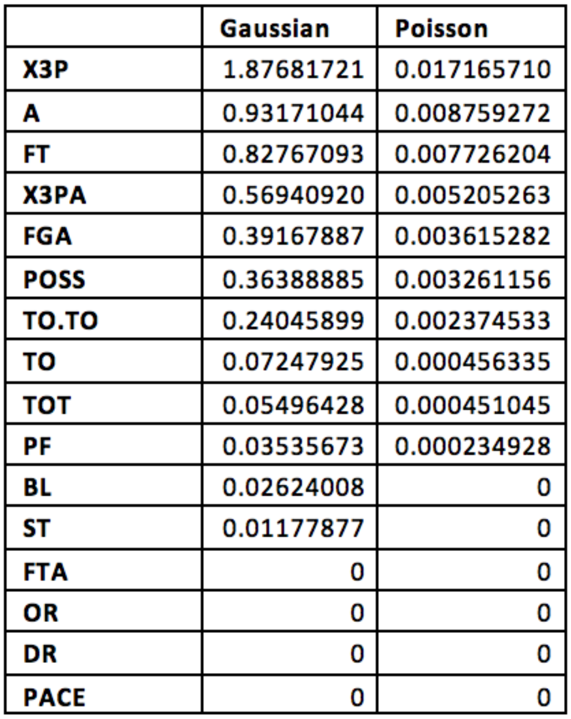


Table V.3 Coefficient of Lasso fit with Gaussian and Poisson



	Gaussian	Poisson
X3P	1.87681721	0.017165710
A	0.93171044	0.008759272
FT	0.82767093	0.007726204
X3PA	0.56940920	0.005205263
FGA	0.39167887	0.003615282
POSS	0.36388885	0.003261156
TO.TO	0.24045899	0.002374533
TO	0.07247925	0.000456335
TOT	0.05496428	0.000451045
PF	0.03535673	0.000234928
BL	0.02624008	0
ST	0.01177877	0
FTA	0	0
OR	0	0
DR	0	0
PACE	0	0

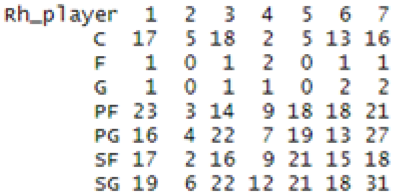
VI. Research Question 3: Can we divide players into different clusters? Are those clusters capable of predicting whether a player can attend the playoff?

In this part, we would like to do clustering to the player data so that we can divide players into different categories and explore the differences between them.

In the data, we have each player's performance for each game they play in the whole season (including regular season and playoff). We average the player's performance over the regular season (separately for road games and home games because it makes a difference) and use this to be a player's feature. Overall, a player will have two rows of data (one for road games and one for home games) in the regular season. If the player made it to the playoff, then he will have another two rows of data.

First, we'll do a simple k-means clustering on all of the players, which means we divide players into different categories with their performance. We would like to explore whether the clustering result has much to do with the position of each player. In the data set, we have 7 positions and they are C, F, G, PF, PG, SF, SG. Usually people use 5 of them. The other two, F and G, appear only several times since they are not widely used, but we still keep them here.

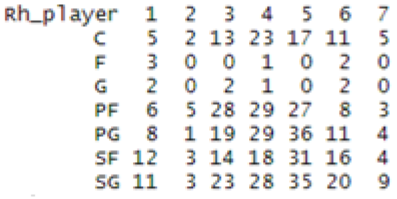
Figure 6.1: clustering result compared to positions



```
Rh_player  1  2  3  4  5  6  7
        C 17  5 18  2  5 13 16
        F  1  0  1  2  0  1  1
        G  1  0  1  1  0  2  2
       PF 23  3 14  9 18 18 21
       PG 16  4 22  7 19 13 27
       SF 17  2 16  9 21 15 18
       SG 19  6 22 12 21 18 31
```

This is the cluster and position comparison for players in home games of the regular season. As we can see, the table is dense and it appears that the positions don't have much to do with our clustering result. But that's the result when we use all the stats of the players (including minutes, points and so on). It's possible that some of them are not related to positions but some do. So we remove some stats and do the clustering again.

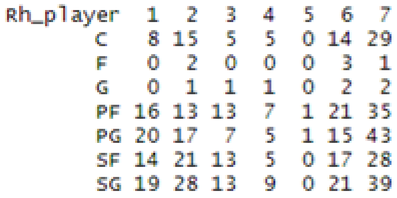
Figure 6.2: Clustering result compared to positions



```
Rh_player  1  2  3  4  5  6  7
       C   5  2 13 23 17 11  5
       F   3  0  0  1  0  2  0
       G   2  0  2  1  0  2  0
       PF  6  5 28 29 27  8  3
       PG  8  1 19 29 36 11  4
       SF 12  3 14 18 31 16  4
       SG 11  3 23 28 35 20  9
```

Obviously, the result is somewhat different but still disproving the relationship between clustering and positions. For the third time, we divide all the other stats of a player by his average time in games. Because in the analysis before, the stats have a lot to do with the time. Players with similar time tend to be in the same cluster.

Figure 6.3: Clustering result compared to positions



```
Rh_player  1  2  3  4  5  6  7
       C   8 15  5  5  0 14 29
       F   0  2  0  0  0  3  1
       G   0  1  1  1  0  2  2
       PF 16 13 13  7  1 21 35
       PG 20 17  7  5  1 15 43
       SF 14 21 13  5  0 17 28
       SG 19 28 13  9  0 21 39
```

For this result, we can see that the cluster 7 gathers a large number of players. Though now we believe the cluster has little relationship with positions, it probably reflects a player's efficiency instead.

We are also curious how different clusters look like.

Figure 6.4: Clustering plot

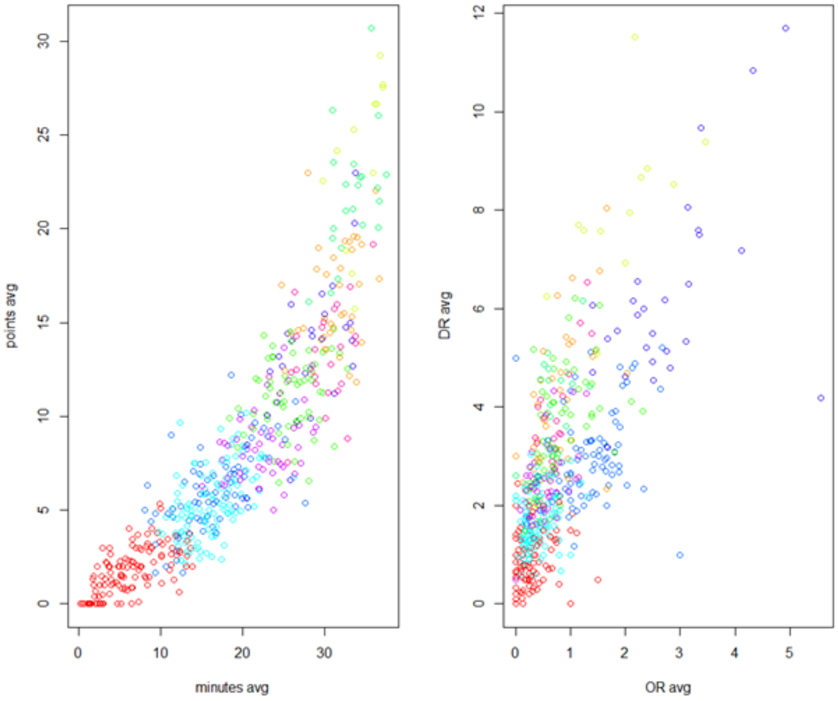
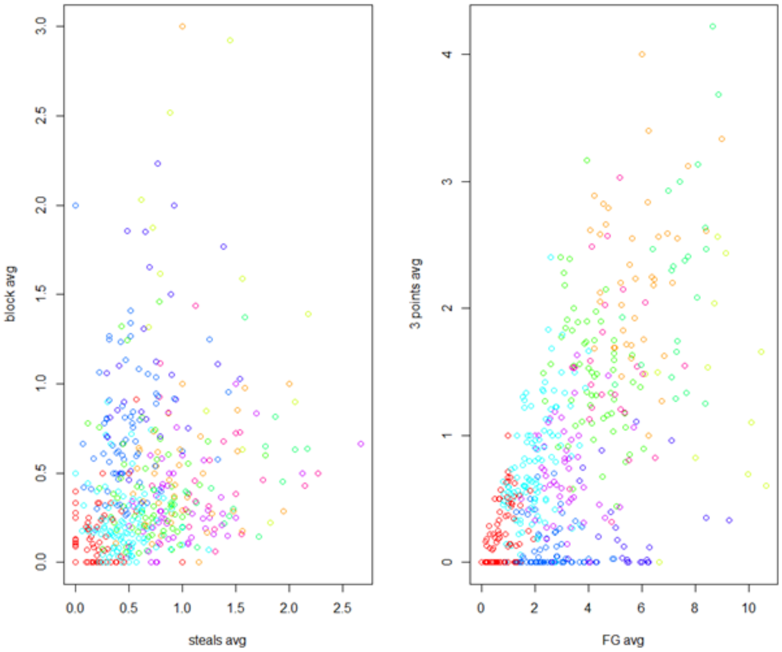


Figure 6.5: Clustering plot

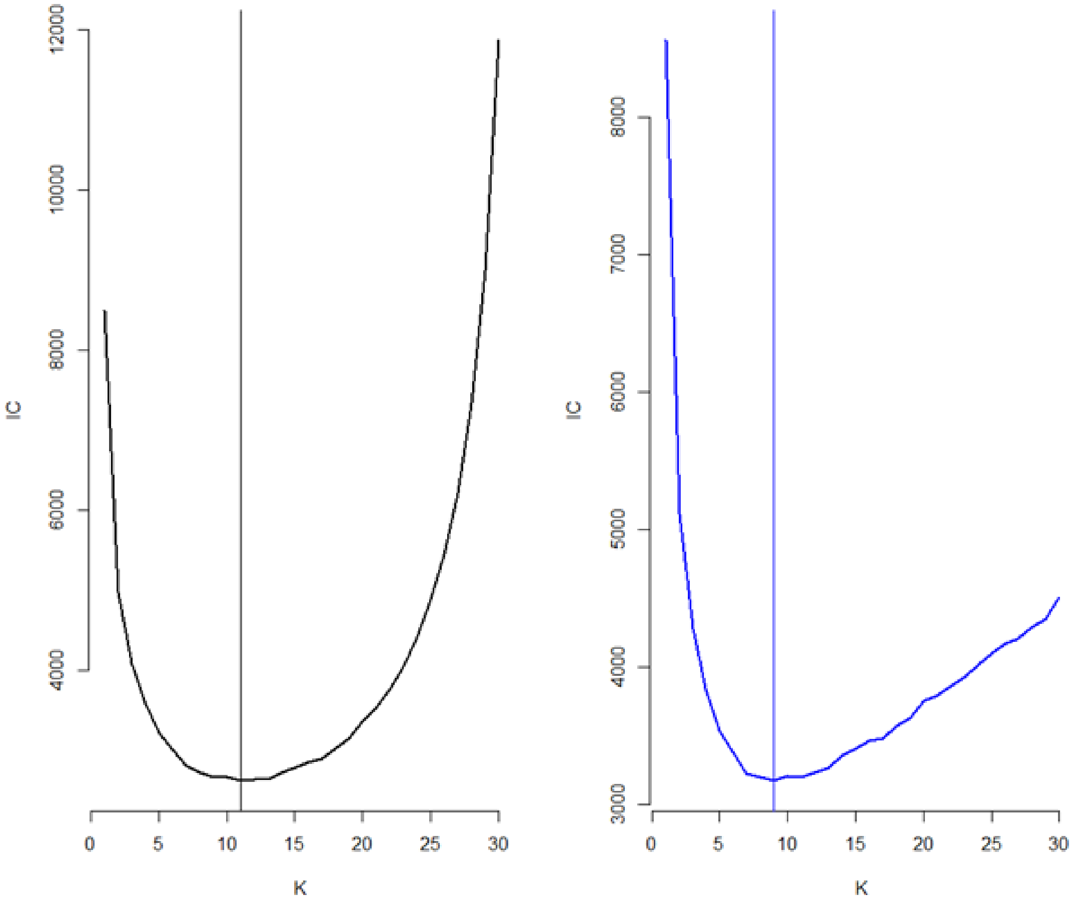


For the first plot, we can see a clear curve, and points in the same clusters are close to each other. In the second one, it appears that the players high in the top right are not in the same cluster as those high in the first plot, which means these represent two different types of players, one is good at scoring while the other is proficient in getting rebounds.

For the third and fourth plot, we cannot see a clear pattern, though players in the same cluster still tend to be close to each other. This is probably because the clusters are mainly decided by the previous two stats.

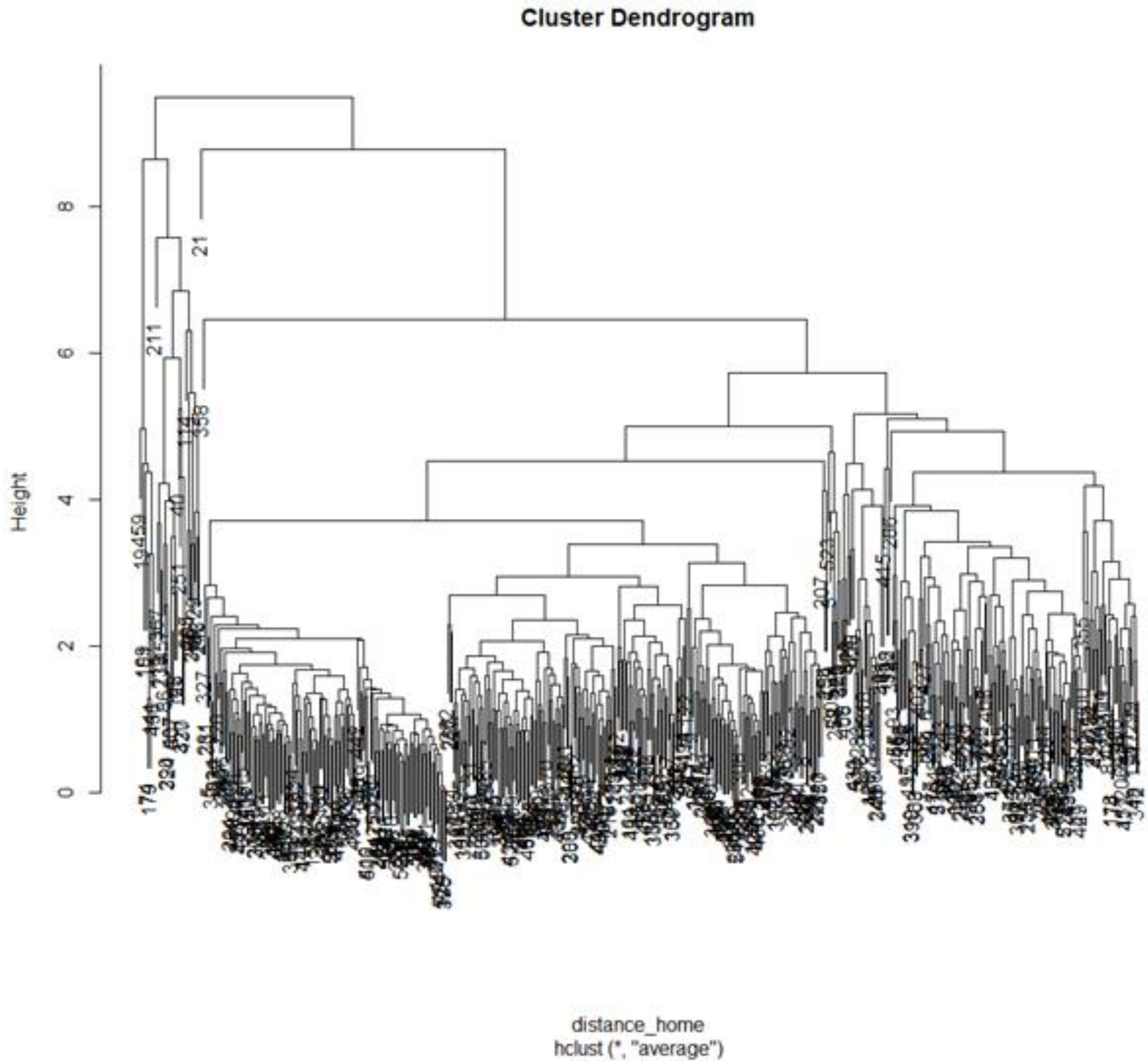
If we take a closer look at the players in a cluster, we can see players like James Harden, LeBron James, Russel Westbrook and so on in the same cluster, those turn out to be the best players in the NBA.

Figure 6.6: k-means for home games



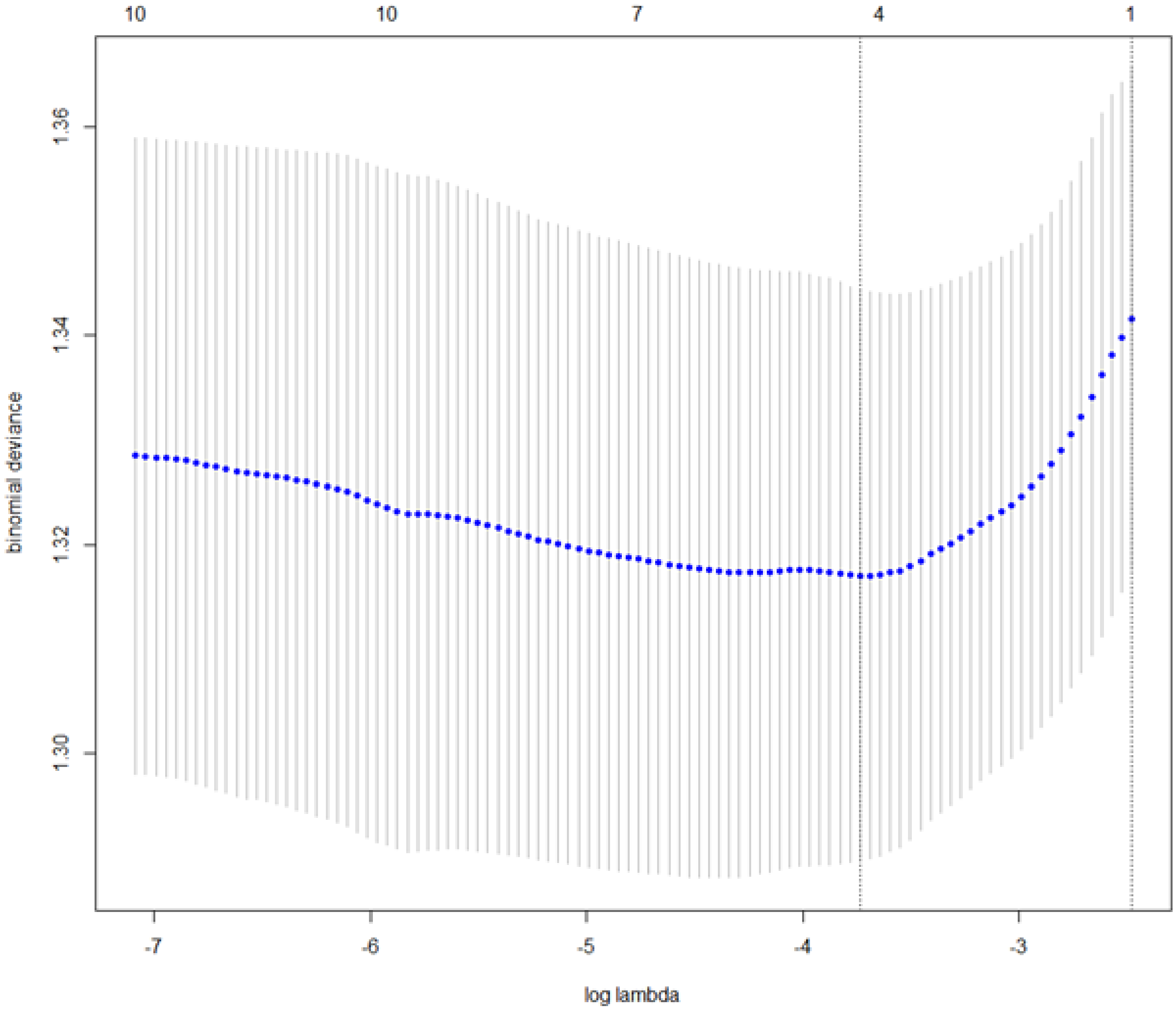
The left part of the plot shows the curve of the AICc criterion, and the right one is the BIC criterion. For AICc, the optimum k is 11. For BIC, the best one is 9. We'll be safe to use just =10 to be the parameter for our hierarchical clustering.

Figure 6.7: Hierarchical result for home games



The results of the two clustering are similar. Then we want to do cluster regression to see how well these clusters predict whether a player can enter the playoff.

Figure 6.8: Deviance plot



The R-squared of the lasso regression is about 0.03, which means the clusters only explain a fairly small part of the team's final performance. This makes sense because one player is only a part of a team and hardly decides the outcome of a match. Then we use the model to predict whether a player can enter the playoff. The accuracy turns out to be 0.625.

We would also like to see whether this cluster for players is similar in both road games and home games. Will players behave differently in home and road games?

In order to check this, we do a similar analysis for road games and compare their result.

Figure 6.9: k-means for road games

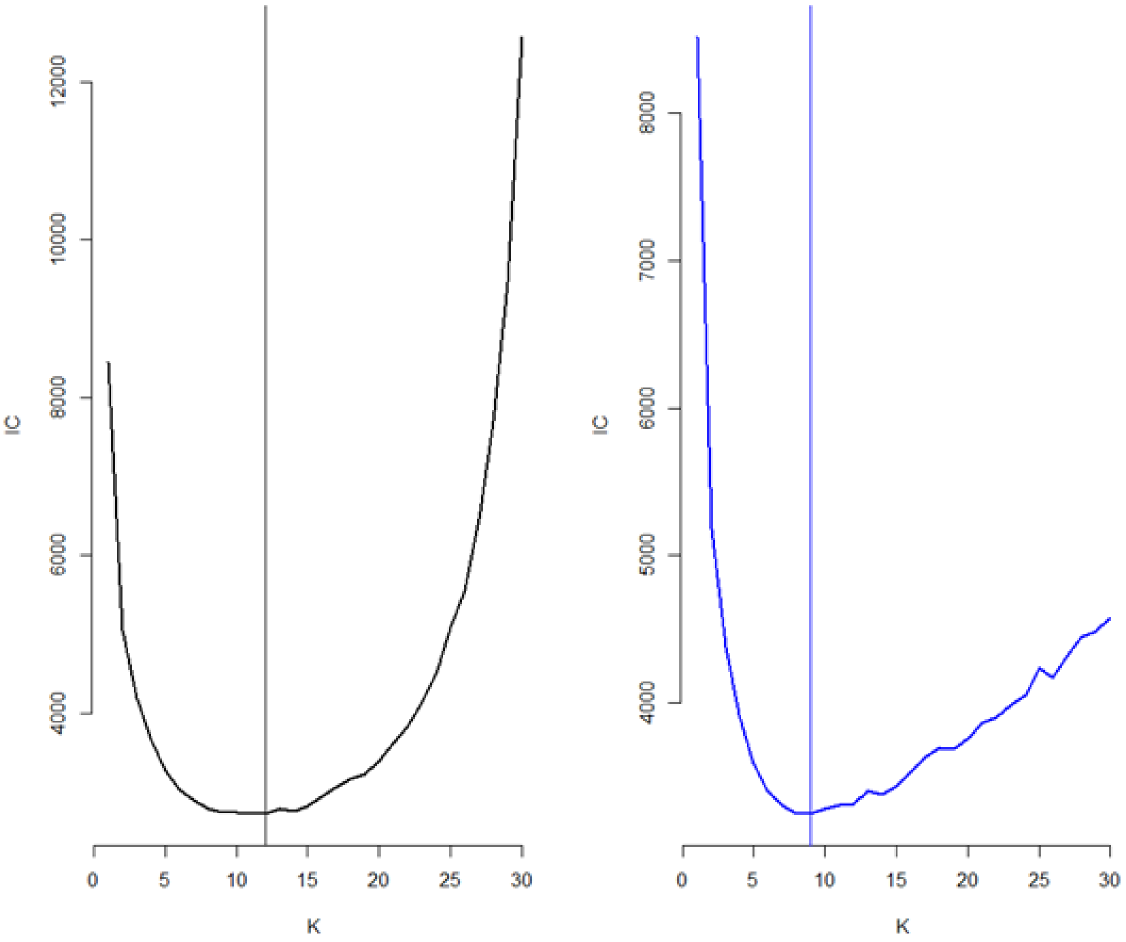
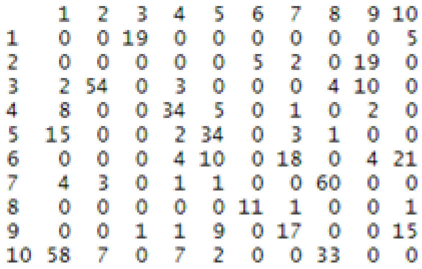
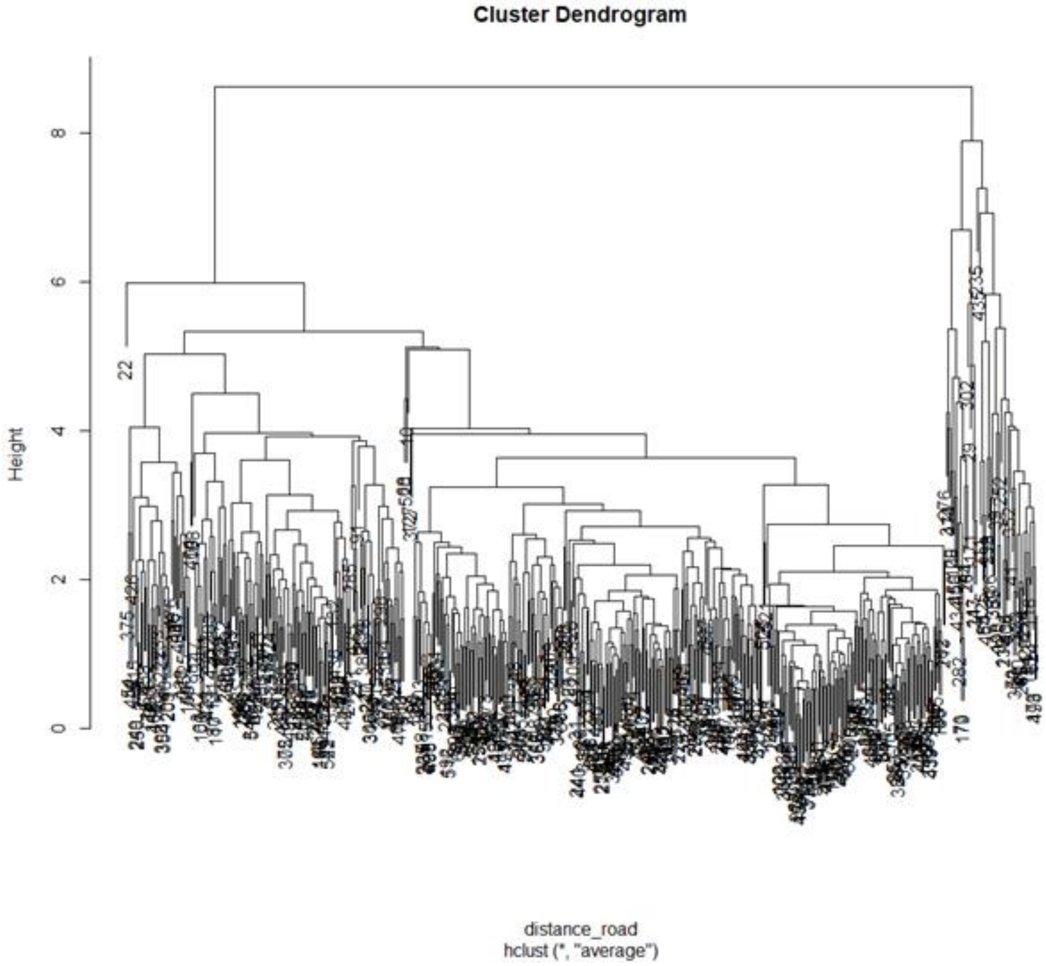


Figure 6.10: table for home games and road games



```
     1   2   3   4   5   6   7   8   9  10
1    0   0  19   0   0   0   0   0   0   5
2    0   0   0   0   0   5   2   0  19   0
3    2  54   0   3   0   0   0   4  10   0
4    8   0   0  34   5   0   1   0   2   0
5   15   0   0   2  34   0   3   1   0   0
6    0   0   0   4  10   0  18   0   4  21
7    4   3   0   1   1   0   0  60   0   0
8    0   0   0   0   0  11   1   0   0   1
9    0   0   1   1   9   0  17   0   0  15
10  58   7   0   7   2   0   0  33   0   0
```

Figure 6.11: Hierarchical result for road games



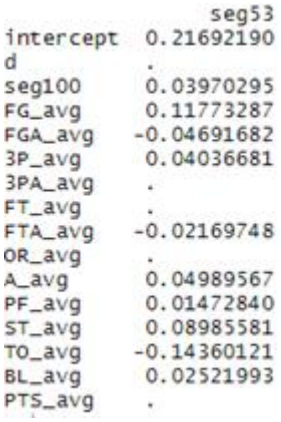
There are 517 players that appear in both road games and home games, so we only count for this part of players. For k value, we can see that it's similar for both games since they are all around 10. From the table of comparison, an obvious fact is that it's still sparse, which means most players' performance keeps similar in home games and road games. But we can also see for the 10th column (also for 7th and so on), the players are separated into several clusters, which means players in this cluster are not performing steadily.

Before we move on to the next part, we are also curious if some stats of a player have a causal effect on the success to the playoff. We use a causal lasso to analyze the effect.

The first column we would like to explore is the defensive rebounds. Through regression we've got that defensive rebound is one of the most important things in the games, but it's still unclear whether it has a causal effect.

We first regress it on other stats (except total rebounds and minutes) and get R square being around 0.77. Then we do a causal lasso and get the following table.

Figure 6.12: Lasso result for defensive rebounds

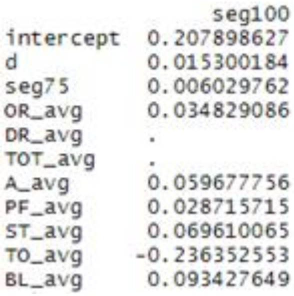


```
                 seg53
intercept  0.21692190
d          .
seg100     0.03970295
FG_avg     0.11773287
FGA_avg   -0.04691682
3P_avg     0.04036681
3PA_avg    .
FT_avg     .
FTA_avg   -0.02169748
OR_avg     .
A_avg      0.04989567
PF_avg     0.01472840
ST_avg     0.08985581
TO_avg    -0.14360121
BL_avg     0.02521993
PTS_avg    .
```

However, we can see the coefficient d (which represents DR) is 0 and that means it doesn't appear to have a causal effect.

Then we do it on the points that a player scores. The R square for the points regressing on other factors (exclude field goals, 3pts, foul shots and so on) is 0.73. Then we do the causal lasso.

Figure 6.13: Lasso result for scores



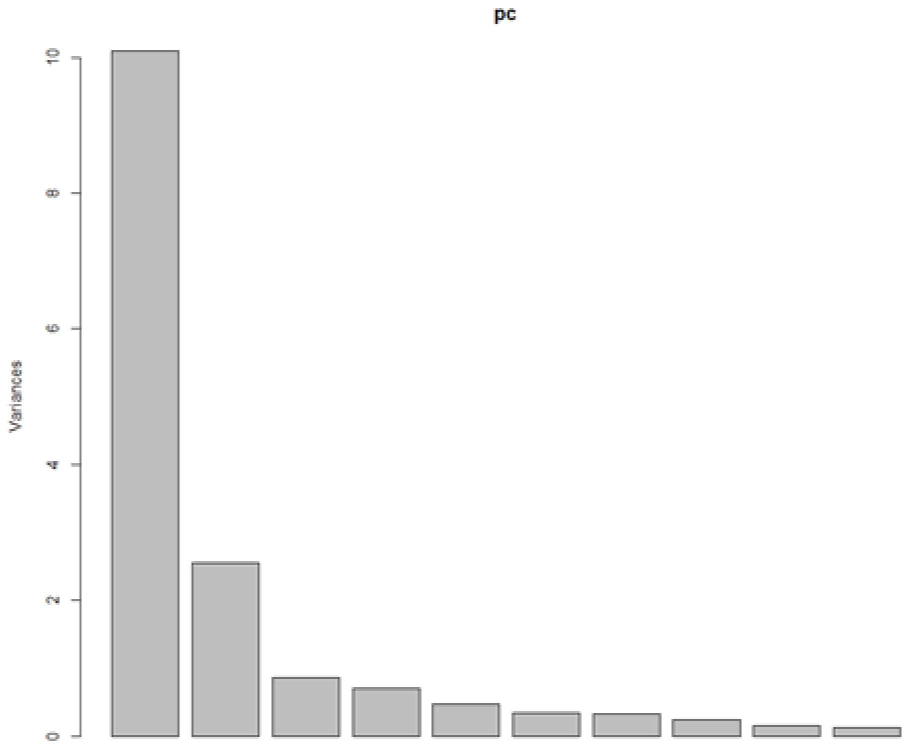
```
                 seg100
intercept  0.207898627
d          0.015300184
seg75      0.006029762
OR_avg     0.034829086
DR_avg     .
TOT_avg    .
A_avg      0.059677756
PF_avg     0.028715715
ST_avg     0.069610065
TO_avg    -0.236352553
BL_avg     0.093427649
```

Then we can see the score of a player still has a causal effect after excluding the part that's not independent from other factors.

VII. Research Question 4: What skills should a player possess in order to get to the playoff?

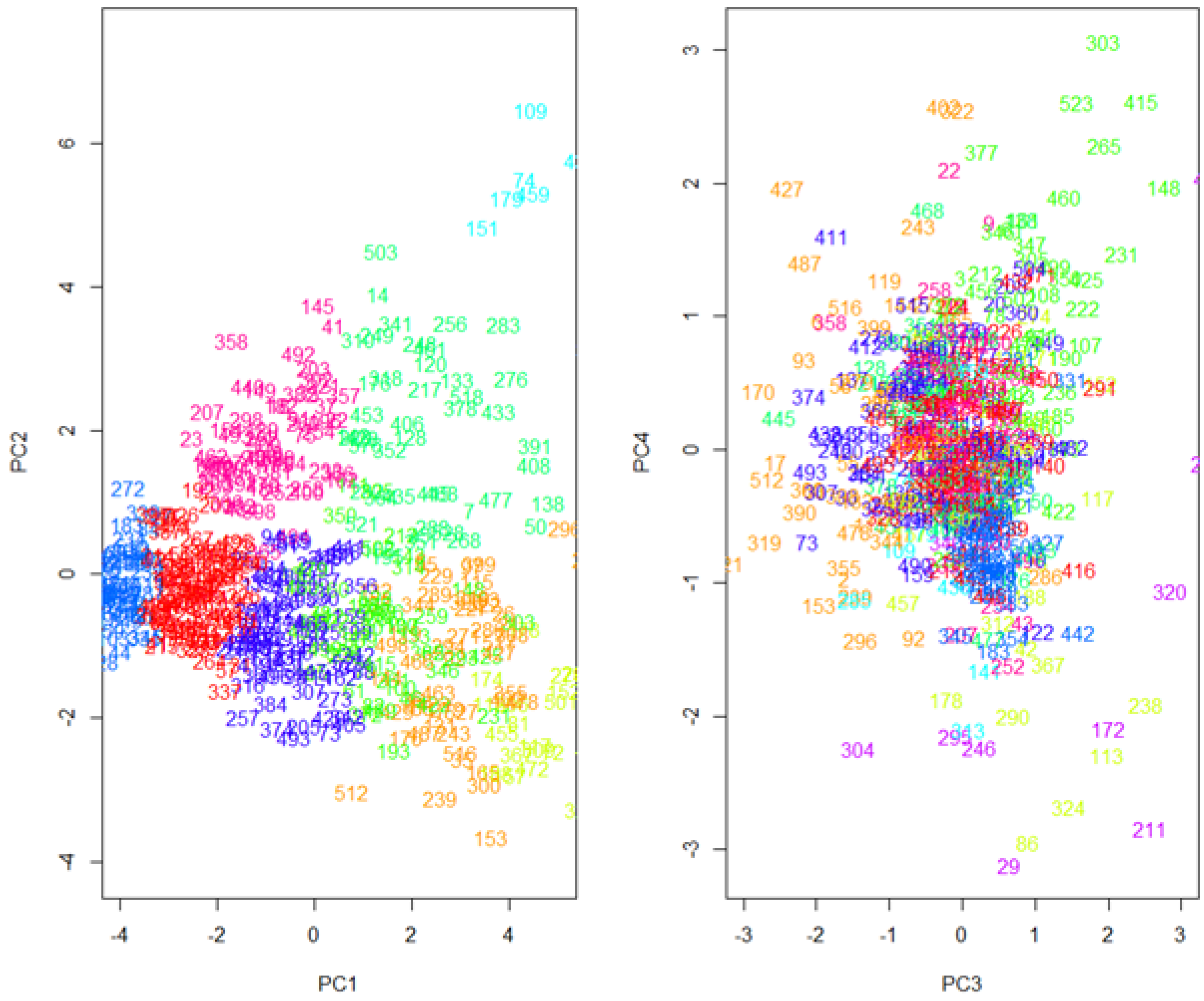
We would like to do dimensional reduction to the players' stats using the PCA method. In this way, we can acquire a simplified version of stats.

Figure 7.1: Variance of principal components



As we can see, the first principal component counts for the majority of the variance explained. It actually explains 63% of all. Then we plot the principal component according to the clustering results.

Figure 7.2: PCA plots



In the first plot, we can see that the players in the same cluster tend to be close to each other. Then we can get the plot of principal components and the playoff factor (1 represents getting to the playoff while 0 represents the opposite).

Figure 7.3: Histogram for first principal

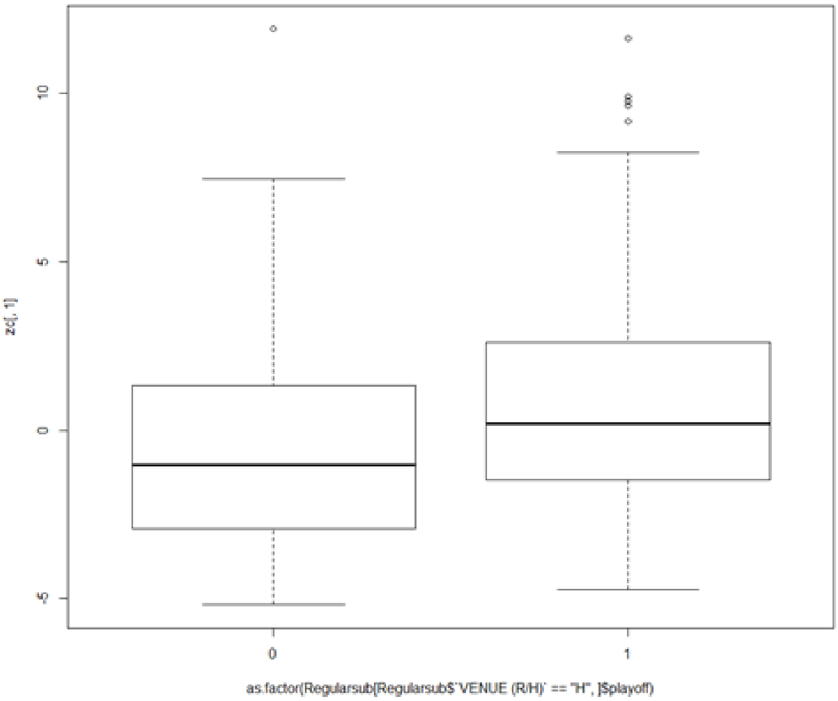
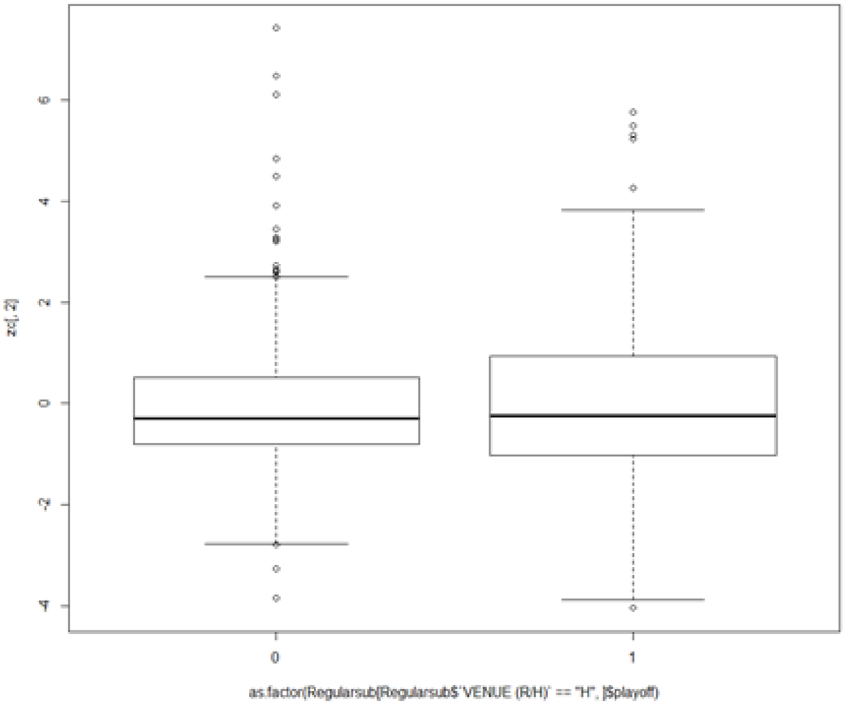


Figure 7.4: Histogram for second principal



The upper plot represents the first component while the lower one represents the second. Clearly the first component captures the relationship with the playoff factor when the second one doesn’t.

Figure 7.5: First principal component

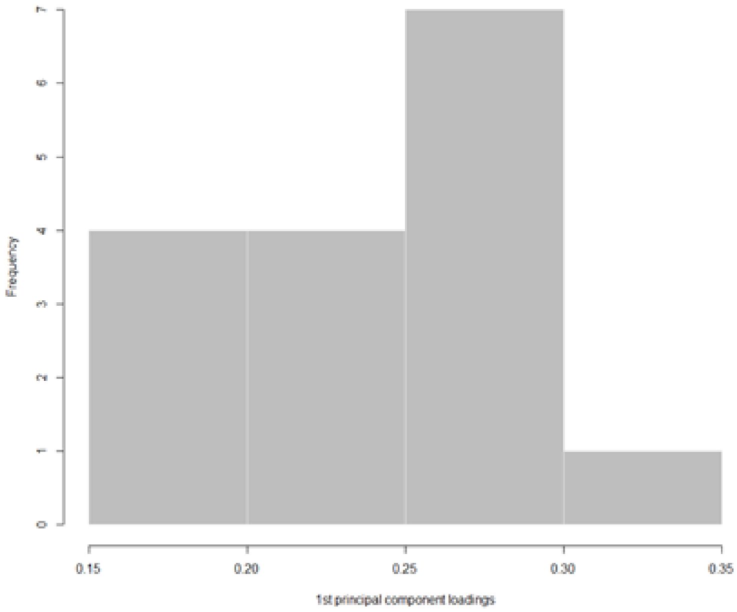
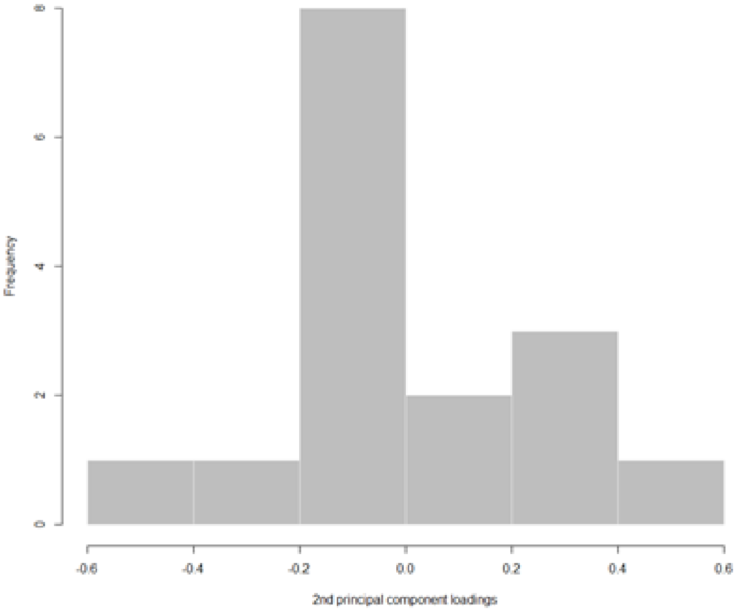


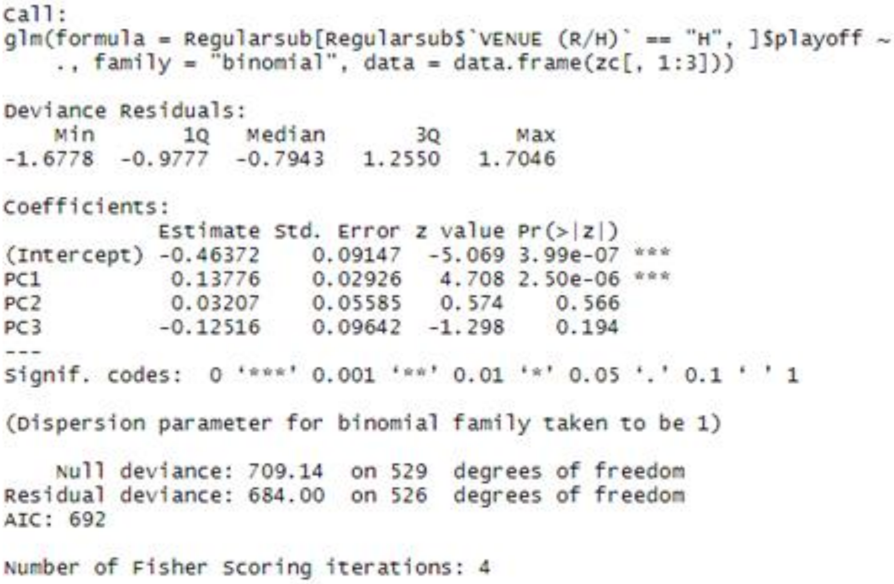
Figure 7.6: Second principal component



For the 1st principal component, it's positive for all factors. For the 2nd principal component, it divides the values into positive and negative parts. For the negative part, it gathers around zero.

Then we would like to do regression on the principal components to see how it works.

Figure 7.7: Regression on the first three principal components



```
Call:
glm(formula = Regularsub[Regularsub$`VENUE (R/H)` == "H", ]$playoff ~
    ., family = "binomial", data = data.frame(zc[, 1:3]))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6778  -0.9777  -0.7943   1.2550   1.7046

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.46372    0.09147  -5.069 3.99e-07 ***
PC1          0.13776    0.02926   4.708 2.50e-06 ***
PC2          0.03207    0.05585   0.574    0.566
PC3         -0.12516    0.09642  -1.298    0.194
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 709.14  on 529  degrees of freedom
Residual deviance: 684.00  on 526  degrees of freedom
AIC: 692

Number of Fisher Scoring iterations: 4
```

We use the first three principal components. The result shows that only the 1st principal component possesses a strong positive relationship with whether a player can attend the playoff, while the other two show no relationship. It also proves the observation we found before that the second principal component has nothing to do with the success. So, it means the first component explains almost all relationships between player stats and playoff. This new statistic is all one need to predict a player's success in the playoffs.

minutes	0.293
FG	0.299
3P	0.193
FT	0.276

OR	0.155
DR	0.255
ST	0.231
BL	0.173
PTS	0.302

From the table we can see that several most important stats are points, minutes, field goals and the several least important stats are offensive rebounds, blocks and 3 points. We can also ignore a gap between offensive rebounds and defensive rebounds. For players, it’s more important to get defensive rebound than get offensive rebound. But for the other part, we can see how many points a player get is of huge importance. So it seems to mean that if a player wants to get to the playoff, he needs to both get more points and do a good defense.

Apart from the PCA method, we also use LASSO to select the most influential factors. We have in total 30 variables, we obtained a LASSO model that has R squared value 0.07841415, which is selected by minimizing AICc. This means all these variables cannot be used to explain well on if a team will go into the playoffs. The lambda selected is 0.001087256. The LASSO plot is shown below. If, alternatively we choose lambda by minimizing BIC, then we will have a lambda value of 0.03398454 that corresponds to a R squared value 0.03636. This is also a very small R squared value, and meaning our variables contain very little information in predicting the outcome of a team going into the playoffs. Yet, we still select 12 most significant variables that have non-zero coefficients which is presented in the table below with the plot.

Figure 7.8: LASSO plot

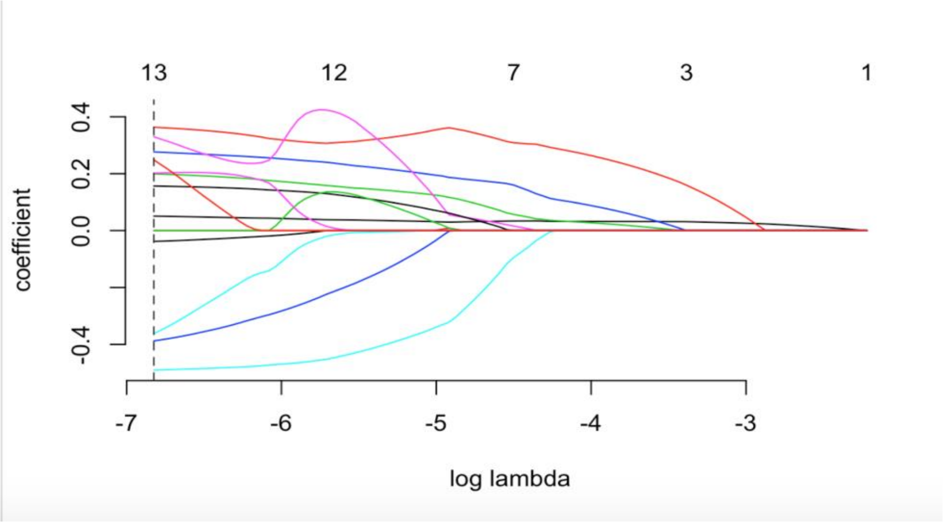
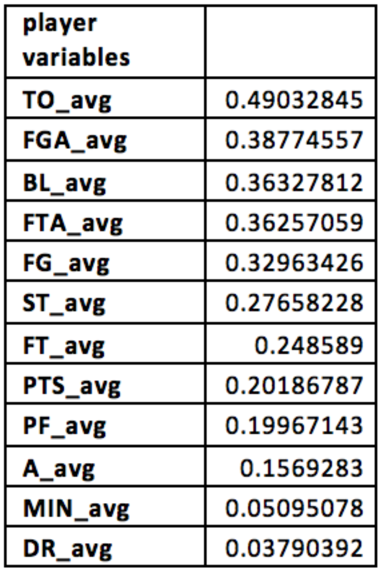


Table 7.1 Coefficients for lasso



player variables	
TO_avg	0.49032845
FGA_avg	0.38774557
BL_avg	0.36327812
FTA_avg	0.36257059
FG_avg	0.32963426
ST_avg	0.27658228
FT_avg	0.248589
PTS_avg	0.20186787
PF_avg	0.19967143
A_avg	0.1569283
MIN_avg	0.05095078
DR_avg	0.03790392

VIII. Conclusion

Cluster regression on teams overall did a decent job in predicting the possibility of going to the playoffs, though it could be further improved by introducing more instrumental variables. We also used k-means, hierarchical clustering, lasso, causal lasso, PCA, PCR and so on in analyzing data. Actually, some of them give similar results which proves their validity, and some of them give us a different aspect of looking at how things work. We can still improve our work through dealing with data elaborately and selecting useful components. Besides, the numerical variables we have are unable to make predictions on whether a team can go to the playoffs with a lasso model, other variables need to be investigated in the future in order to find good predictors for the model.

IX. Appendix

Variables Description	
1Q	1st quarter scoring result
2Q	2nd quarter scoring result
3Q	3rd quarter scoring result
4Q	4th quarter scoring result
OT1	1st over time period scoring result
OT2	2nd over time period scoring result
OT3	3rd over time period scoring result
OT4	4th over time period scoring result
F	Final scoring result
MIN	Total minutes that the team (sum all players' individual times) played.
FG	Field Goals Made
FGA	Field Goals Attempted
3P	Three Point Field Goals Made
3PA	Three Point Field Goals Attempted
FT	Free Throws Made
FTA	Free Throws Attempted
OR	Offensive Rebounds
DR	Defensive Rebounds
TOT	Total Rebounds
A	Assists
PF	Personal Fouls
ST	Steals
TO	Turnovers assigned to players
TO TO	(Turnovers assigned to players)+(Turnovers assigned to teams)
BL	Blocks
PTS	Points
POSS	Total possessions. It is assumed that both teams use same number of possessions in a game
PACE	Estimate of number of possessions per 48 minutes by a team
OEFF	Offensive Efficiency
DEFF	Defensive Efficiency
REST DAYS	Rest days