NBA 다변량 분석 보고서

201511508 통계학과 김민철

부산대학교 통계학과
PUSAN NATIONAL UNIVERSITY

1. Abstract

보통 남학생들은 축구를 가장 좋아한다. 남학생이라면 보통 점심시간에 학교 운동장에서 축구를 하는 추억이 흔히 있을 것이다. 하지만 축구만큼은 아닐지라도 남학생들이 좋아하고 많이 하는 스포츠가 바로 농구이다. 학창 시절의 추억 중에서 스포츠에 대한 로망은 보통은 축구가 아닌 농구였다. 필자 또한 대학교에 와서 농구를 접하게 된 이후로 꾸준히 관심을 가지며 지속적으로 해왔다. 최근 NBA에 대한 관심이 높아지고 있는 것을 인터넷의 기사와 많아진 댓글을 통해 확인할 수 있었다. 이에 대해 분석해 보니 NBA의 경기를 실시간으로 시청할 수 있다는 점이 인기 상승의 원인이었다. 상승하는 인기에 힘입어 많은 사람들이 NBA를 시청하고 수다를 떨 수 있도록 NBA 정규시즌 자료를 통해 다변량 분석을 실시해 보겠다.

2. Introduction

06월 14일 금요일 미국 NBA의 컨퍼런스 파이널이 막을 내렸다. NBA의 파이널을 실시간 스트리밍으로 시청하고 있던 시청자 수는 84만 명을 넘겼고 경기와 함께 엄청난 양의 기사가 쏟아져 나왔다. 농구는 한국에서 비인기 스포츠 종목이다. 하지만 전 세계가 기술로 연결되고 세계 최고 선수들의 플레이를 실시간으로 감상할 수 있게 되면서부터 우리나라에서도 NBA에 대한 관심이 늘기 시작했다. NBA의 인기 비결로는 역시 엄청난 퍼포먼스를 보여주는 NBA 선수들의 경기력과 세계 정상급 지도자들의 전술 덕분일 것이다. 이번 term project에서 NBA 공식자료들을 바탕으로 team 성적에 영향을 줄 것으로 예상이 되는 요인을 선별했다. 그 후 다변량 분석기법인 PCA, FA를 통해서 어떤 변수가 team 성적에 더 큰 영향을 끼치는지와 CA를 통해서 자료를 기준으로 그룹화시켜 그룹들의 특성을 알아보았다. 이를 통해 앞으로 NBA team 성적에 영향을 미치는 요인들을 알아보고 NBA를 즐겁게 시청하기 위한 시청 포인트를 제시하겠다.

3. Data Description

(1) 데이터

	WIN %	PTS	3PA	3P %	FT %	OREB	AST	TOV	STL	+/−
Milwaukee	.732	118.1	38.2	35.3	77.3	9.3	26.0	13.9	7.5	8.9
Toronto	.707	114.4	33.8	36.6	80.4	9.6	25.4	14.0	8.3	6.1
Golden	.695	117.7	34.4	38.5	80.1	9.7	29.4	14.3	7.6	6.5

Denver	.659	110.7	31.4	35.1	75.5	11.9	27.4	13.4	7.7	4.0
Houston	.646	113.9	45.4	35.6	79.1	10.2	21.2	13.3	8.5	4.8
Portland	.646	114.7	30.7	35.9	81.4	11.8	23.0	13.8	6.7	4.2
Philadelphia	.622	115.2	30.2	35.9	77.1	10.9	26.9	14.9	7.4	2.7
Utah	.610	111.7	34.0	35.6	73.6	10.0	26.0	15.1	8.1	5.3
Boston	.598	112.4	34.5	36.5	80.2	9.8	26.3	12.8	8.6	4.4
Oklahoma	.598	114.5	32.6	34.8	71.3	12.6	23.4	14.0	9.3	3.4
Indiana	.585	108.0	25.4	37.4	75.2	9.3	26.0	13.7	8.7	3.3
LA	.585	115.1	25.8	38.8	79.2	9.7	24.0	14.5	6.8	0.9
SanAntonio	.585	111.7	25.3	39.2	81.9	9.2	24.5	12.1	6.1	1.7
Brooklyn	.512	112.2	36.2	35.3	74.5	11.0	23.8	15.1	6.6	-0.1
Orlando	.512	107.3	32.1	35.6	78.2	10.0	25.5	13.2	6.6	0.7
Detroit	.500	107.0	34.8	34.8	74.7	11.4	22.5	13.8	6.9	-0.2
Charlotte	.476	110.7	33.9	35.1	79.7	9.9	23.2	12.2	7.2	-1.1
Miami	.476	105.7	32.4	34.9	69.5	11.2	24.3	14.7	7.6	-0.2
Sacramento	.476	114.2	29.9	37.8	72.6	11.0	25.4	13.4	8.3	-1.1
LAlakers	.451	111.8	31.0	33.3	69.9	10.2	25.6	15.7	7.5	-1.7
Minnesota	.439	112.5	28.7	35.1	78.7	11.3	24.6	13.1	8.3	-1.5
Dallas	.402	108.9	36.6	34.0	74.2	10.1	23.4	14.2	6.5	-1.3
Memphis	.402	103.5	28.9	34.2	77.2	8.8	23.9	14.0	8.3	-2.6
NewOrleans	.402	115.4	29.9	34.4	76.1	11.1	27.0	14.8	7.4	-1.3
Washington	.390	114.0	33.3	34.1	76.8	9.7	26.3	14.1	8.3	-2.9
Atlanta	.354	113.3	37.0	35.2	75.2	11.6	25.8	17.0	8.2	-6.0
Chicago	.268	104.9	25.9	35.1	78.3	8.8	21.9	14.1	7.4	-8.4
Cleveland	.232	104.5	29.1	35.5	79.2	10.7	20.7	13.5	6.5	-9.6
Phoenix	.232	107.5	29.3	32.9	77.9	9.1	23.9	15.6	9.0	-9.3
NewYork	.207	104.6	29.5	34.0	75.9	10.5	20.1	14.0	6.8	-9.2

(2) 변수 소개

● W - Win : 승리 횟수

이는 곧 팀의 성적이라는 결과로 해석할 수 있으며 각 요인이 경기 성적에 얼마만큼 영향을 끼치는지 확인할 수 있는 요인으로 볼 수 있다.

● PTS - Points : 경기당 평균 득점

최근 NBA는 업템포(경기의 흐름을 빠르게 하기 위해 경기속도를 올리는 것을 지칭하는 용어)와 3점슛의 영향을 많이 받아 경기당 평균 득점이 높아지는 경향이 있다. 특히 골든스테이트 워리어스와 휴스턴 로케츠처럼 3점슛 시도가 높은 팀이 고득점을 하는 경기가

많기에 변수로서 선택하게 되었다.

● 3PA – 3 Point Field Goals Attempted : 경기당 3점슛 시도 횟수

현대 농구에서 3점슛은 이제 승리를 위한 필수불가결한 요소가 되었다. 이를 얼마나 잘 받아들이고 시도하는 팀이 승리를 많이 확보하는지 분석을 통해 확인할 것이다.

● 3P% – 3 Point Field Goals Percentage : 3점슛 성공률

아무리 수준급 선수들이 즐비한 NBA라도 3점슛 시도에 비해 성공 확률이 높지 못하다면 3점슛은 팀 승리에 부정적인 영향을 끼치는 요인이 될 것이고 반대로 성공 확률이 높다면 팀 승리에 긍정적인 영향을 끼칠 것이다. 이를 분석을 통해 확인하도록 하겠다.

● FT% – Free Throw Percentage : 자유투 성공률

농구의 특성상 자유투는 한 번 성공할 때마다 1점씩 추가된다. 최근 트렌드인 3점슛에 비하면 3분의 1뿐인 점수이지만 경기의 승패를 가르는 접전 승부 구간에서는 수비 강도가 강해져 3점슛 보다는 수비자 파울로 인한 자유투로 인한 득점이 더 자주 발생하기에 선택하게 되었다. 자유투를 높은 확률로 성공시키는 팀이 많은 승리를 거뒀는지 확인하는 것 또한 이번 분석의 목적이다.

● OREB – Offensive Rebounds : 공격자가 슛 시도에 실패하고 그 공을 잡아낸 공격리바운드 횟수

앞서 말해왔듯이 현대 농구의 트렌드는 업템포와 3점슛이다. 업템포로 인한 빠른 공수전환(농구에서는 이를 트랜지션이라고 부른다.)과 슛 시도로 인해 리바운드 횟수가 증가하게 됐다. 공격에서 실패한 슛을 빠르게 잡아낸다면(공격 리바운드) 득점의 기회가 늘어나 득점 또한 높아지게 되어 승리에 기여하는 요인이 될 것으로 예상된다.

● AST – Assists : 득점으로 연결된 패스의 수

농구에서는 개인적으로 화려한 드리블과 움직임. 그리고 멋진 3점 슛을 떠올릴 것이다. 최근에는 개인적인 전술 중심에서 벗어나 팀원으로 움직이고 패스하는 팀 플레이가 떠오르고 있다. 경기에 막대한 영향을 끼치는 3점슛을 편하게 쏘기 위해 공간을 확보하는 움직임(농구에서는 이를 스페이싱이라 부른다.)을 경기에서 많이 볼 수 있다. 동시에 그 공간을 활용하여 패스를 통해 골대로 향하는 동료에게 손쉬운 득점을 만들어낼 수 있는 것이 어시스트이다. 어시스트 개수 또한 현대 농구 흐름의 중요한 요인이며 이는 경기 성적으로 직결된다고 예상된다.

● TOV – Turnovers : 경기에서 실책으로 인한 공격권 이전의 횟수를 나타내는 지표.(실책)

앞서 말한 문맥과 동일하듯이, 현대 농구는 업템포와 3점슛을 통한 빠른 농구를 하고 있다. 하지만 빠른 농구의 단점은 많은 실수를 저질러 상대에게 쉬운 득점기회를 내줄 수 있다는 것이다. 턴오버로 인한 어이없는 실점은 경기의 승패에 절대적인 영향을 끼친다. 이를 단속하기 위해 일부러 경기 템포를 낮추는 팀들이 있듯이(휴스턴 로케츠, 샌안토니오 스퍼스) 턴오버, 즉 실책을 관리하는 것이 승리에 얼마나 많은 영향을 끼치는지 확인할 것이다.

● STL – Steals : 공격자의 공을 뺏어내 횟수를 나타내는 지표.

위에서 말했듯이 가장 손쉬운 득점은 상대방의 실수를 통한 득점이다. 그렇다면 수비자들은 상대방의 실수를 기다릴 것이 아니라 실수를 유도하게끔 만드는 함정 수비와 협동 수비를 적극적으로 사용해야 한다. 대표적인 팀은 함정 수비와 투맨 게임을 활용하는 골든 스테이트, 공세적인 일선 압박 수비를 하는 오클라호마 시티이다. 이를 통해 스틸이 경기의 승패에 어떤 영향을 끼치는지 확인할 것이다.

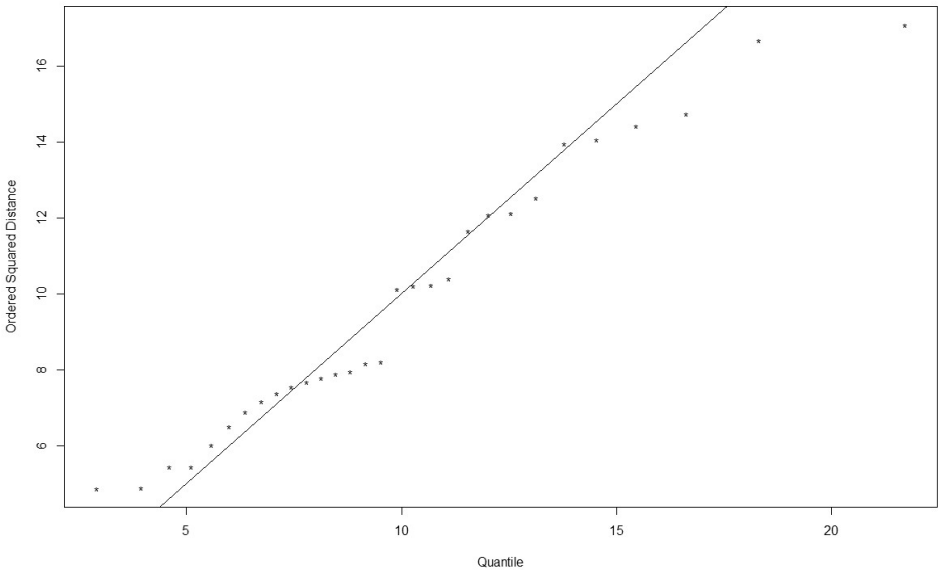
● +/- – Plus Minus : 경기당 평균 득실점 마진

엄밀히 얘기하자면 농구에서 1점 차로 이기는 것이든 20점 차로 이기는 것이든 다를 바 없

다. 하지만 정규 시즌은 82경기이고 북아메리카라는 거대한 땅에서 각 주를 돌아다니며 때로는 이틀 연속으로 경기를 치르다 보면 선수들은 많이 지칠 수밖에 없다. 이를 최대한 보완하기 위해서 경기의 가장 중요한 시간대인 3쿼터에 많은 득점을 넣고, 가비지 타임(이미 상당히 벌어진 점수 차이를 인정하고 경기를 포기함과 동시에 후보 선수들을 경기에 출전시키는 시간을 지칭함. 보통은 지고 있는 팀에서 먼저 4쿼터에 후보 선수들을 출전시킨다.)을 통한 주축 선수들의 휴식을 확보하는 전술이 많이 쓰이고 있다. 상위권인 골든 스테이트와 밀워키 벅스가 가장 대표적인 팀이라고 할 수 있다. 이를 통해 득실차가 정규 시즌 성적에 어떠한 영향을 끼치는지 확인할 수 있을 것이다.

3. 다변량 정규성 파악

```
> result
$multivariateNormality
             Test          Statistic           p value Result
1 Mardia Skewness  202.612743743536 0.793836080149637    YES
2 Mardia Kurtosis -1.25617610894169 0.209052126708971    YES
3             MVN               <NA>              <NA>    YES
```




```
> rq=cor(cbind(q, m))[1,2]
> rq
[1] 0.9833799
```

위는 R의 "MVN" 패키지 중 mvn 함수를 이용하여 수행한 왜도와 첨도를 이용한 다변량 정규성 결과 및 카이제곱 그림이다. 카이제곱 그림을 보았을 때 몇 개의 점을 제외한 점들이 직선위에 배열되어 있다. 그리고 분위수와 마할라노비스 거리의 상관계수(rq)가 0.9833799로 1에 가까워 상관성이 매우 인정된다. 또한 왜도와 첨도를 이용한 다변량 정규성 검정결과에서 왜도 검정 / 첨도 검정 모두 귀무가설을 기각하지 않으므로 위의 NBA data는 다변량 정규성을 만족하는 자료라고 볼 수 있다. 따라서 다변량 정규성을 바탕으로 여러 가지 다변량 분석기법을 사용해보도록 하겠다.

4. 주요인 분석(Principal Components Analysis)

주성분 분석은 자료에 대해 통계적 모형이나 어떤 특별한 가정을 필요로 하지 않으며 공분산 행렬의 고유값과 고유벡터에 의해 주성분이 결정된다. 주성분은 최대의 분산을 갖는 일차결합으로서 주성분들 간에는 서로 상관되지 않도록 결정된다. 서로 다른 두 개 주성분 간의 공분산은 0이어야 한다. 그리고 주성분 분산을 최대로 하는 벡터에 의한 일차 결합에 의해 주성분이 결정된다.

자료의 변수는 승수, 3점슛 시도, 공격리바운드 개수, 어시스트 개수, 턴오버 개수 등 단순히 개수를 단위로 하는 변수가 있고 그 와는 반대로 3점슛 성공률, 총득점 등 개수와는 다른 단위를 다르게 사용하는 변수가 있었다. 따라서 상관행렬을 이용한 주요인 분석을 실시했다.

```
> R
           W   PTS    3PA    3P%    FT%   OREB    AST    TOV    STL    +/-
W      1.000 0.661  0.313  0.542  0.161  0.080  0.498 -0.246  0.099  0.980
PTS    0.661 1.000  0.331  0.384  0.158  0.195  0.562  0.083  0.169  0.647
3PA    0.313 0.331  1.000 -0.225 -0.092  0.182  0.000  0.116  0.148  0.365
3P%    0.542 0.384 -0.225  1.000  0.419 -0.159  0.280 -0.370 -0.164  0.464
FT%    0.161 0.158 -0.092  0.419  1.000 -0.400 -0.054 -0.477 -0.212  0.104
OREB   0.080 0.195  0.182 -0.159 -0.400  1.000 -0.036  0.148  0.002  0.056
AST    0.498 0.562  0.000  0.280 -0.054 -0.036  1.000  0.190  0.261  0.520
TOV   -0.246 0.083  0.116 -0.370 -0.477  0.148  0.190  1.000  0.174 -0.252
STL    0.099 0.169  0.148 -0.164 -0.212  0.002  0.261  0.174  1.000  0.135
+/-    0.980 0.647  0.365  0.464  0.104  0.056  0.520 -0.252  0.135  1.000
```

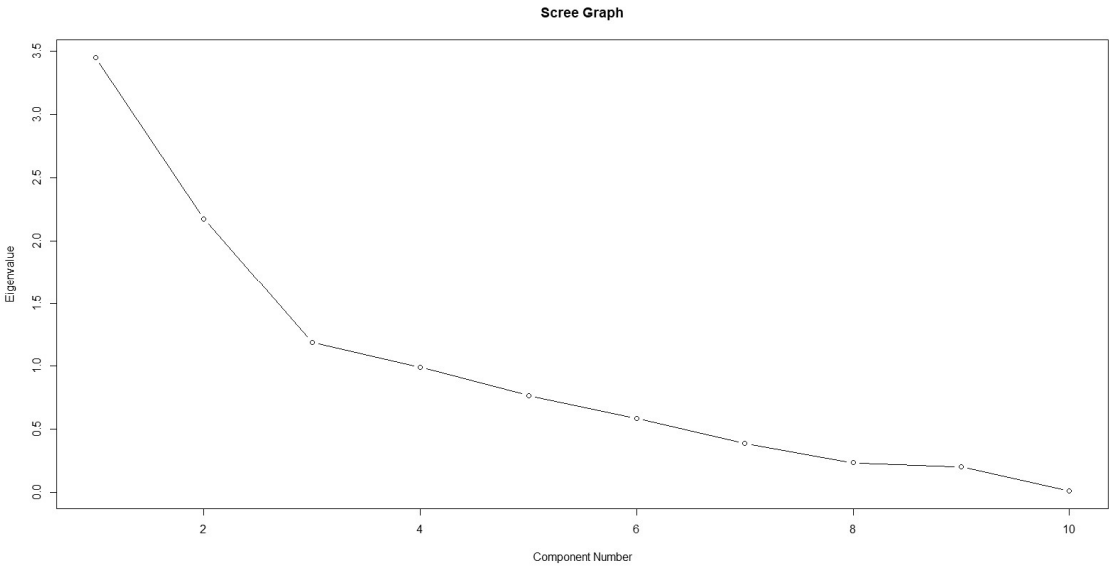
(1) 설명력 관찰

```
> round(eigen$values, 3) # Eigenvalues
 [1] 3.452 2.172 1.190 0.991 0.769 0.587 0.390 0.234 0.204 0.011
> round(gof, 3)
 [1] 34.513 21.714 11.905  9.909  7.689  5.870  3.903  2.349  2.036  0.111
```

위의 수치는 상관행렬 R을 spectral Decomposition 한 후에 얻은 고윳값들을 소수점 세 번째 자리까지 반올림한 수치이다. 밑의 수치는 고윳값의 총합에서 설명 비율을 수치화한 것이다.

이를 바탕으로 주요인의 개수를 결정할 수 있다. 최대 고윳값은 l_1 =3.452이며 전체 고윳값에서 약 34.513%를 설명하고, 그 다음으로 l_2 =2.172와 l_3 =1.190의 고윳값이 있으며 전체 고윳값에서 21.714%와 11.905%를 설명한다.

처음 세 개 주성분은 총 68.132%를 설명할 수 있기에 우리가 원하는 70% 설명력에 미치지 못하지만 대략적으로 70%를 달성할 수 있기에 세 개의 주성분을 고려해도 충분하다고 판단하며 이를 통해 주성분 분석을 시행하겠다. 그리고 본격적으로 분석에 앞서 scree plot을 통해 결과를 다시 확인하도록 하겠다.



위의 그림을 살펴보면 세 번째 고윳값 이후로 기울기가 완만해지므로 세 번째 고윳값이 팔꿈치가 되며 두 번째 고윳값까지를 주성분 요인으로 결정해야겠지만 네 번째 고윳값 이후로도 기울기가 완만하며 총 설명력을 고려하여 세 번째 고윳값까지 주요인으로 고려해도 된다고 판단했다.

(2) 주성분 해석

```
> V3
        [,1]  [,2]  [,3]
 [1,] -0.51 -0.01  0.12
 [2,] -0.43 -0.17 -0.02
 [3,] -0.16 -0.32  0.50
 [4,] -0.34  0.36 -0.13
 [5,] -0.14  0.50  0.02
 [6,] -0.02 -0.36  0.46
 [7,] -0.34 -0.18 -0.51
 [8,]  0.13 -0.47 -0.32
 [9,] -0.08 -0.32 -0.37
[10,] -0.50 -0.05  0.12
```

위의 행렬은 상관행렬 Spectral decomposition 한 후의 얻은 eigenvector 중 l_1, l_2, l_3의 eigenvector로서 주성분 PC1, PC2, PC3의 계수이다.

<제 1주성분>

$$p_1 = -0.51y_1 - 0.43y_2 - 0.16y_3 - 0.34y_4 - 0.14y_5 - 0.02y_6 - 0.34y_7 + 0.13y_8 - 0.08y_9 - 0.50y_{10}$$

<제 2주성분>

$$p_2 = -0.01y_1 - 0.17y_2 - 0.32y_3 + 0.36y_4 + 0.50y_5 - 0.36y_6 - 0.18y_7 - 0.47y_8 - 0.32y_9 - 0.05y_{10}$$

<제 3주성분>

$$p_3 = 0.12y_1 - 0.02y_2 + 0.50y_3 - 0.13y_4 + 0.02y_5 + 0.46y_6 - 0.51y_7 - 0.32y_8 - 0.37y_9 + 0.12y_{10}$$

제 1주성분에서 y8의 계수가 +0.13이고 나머지 변수들의 부호가 모두 음수인 것으로 보아 제 1주성분은 TOV(턴오버)과 TOV를 제외한 지표들의 대비를 나타내는 성분이라고 말할 수 있다.

여기서 y1과 y2, y10의 계수가 -0.51, -0.43, -0.50으로 높은 수치를 지니고 있으며 이는 승수와 총득점, 경기당 평균 득점마진이 제 1주성분 계수에 큰 영향을 끼친다고 말할 수 있다.

주요인 분석에 앞서 변수 설명과 함께 몇 가지 예상을 했었다. 제 1주성분의 계수를 통해 경기당 평균 득점마진 지표(+/-)가 시즌성적(=승수,W)에 얼마나 영향력을 끼치는지 확인할 수 있는데 -0.51과 -0.50은 서로간 강한 관계를 가지고 있음을 추측할 수 있게 해준다.

제 2주성분에서 변수의 부호를 살펴보았을 때, PTS(평균 득점), 3PA(3점슛 시도 횟수), OREB(공격 리바운드 개수), AST(어시스트 개수), TOV(턴오버 개수), STL(스틸 개수)은 모두 음수이고 3P%(3점슛 성공률), FT%(자유투 성공률)는 모두 양수이다. 따라서 제 2주성분은 (평균 득점, 3점슛 시도 횟수, 공격 리바운드 개수, 어시스트 개수, 턴오버 개수, 스틸 개수)과 (3점슛 성공률, 자유투 성공률)의 대비를 나타내는 주성분이라고 할 수 있다. 하지만 W의 계수가 0에 가까워 실질적으로 시즌 성적에는 영향을 미친다고 말할 수 없다.

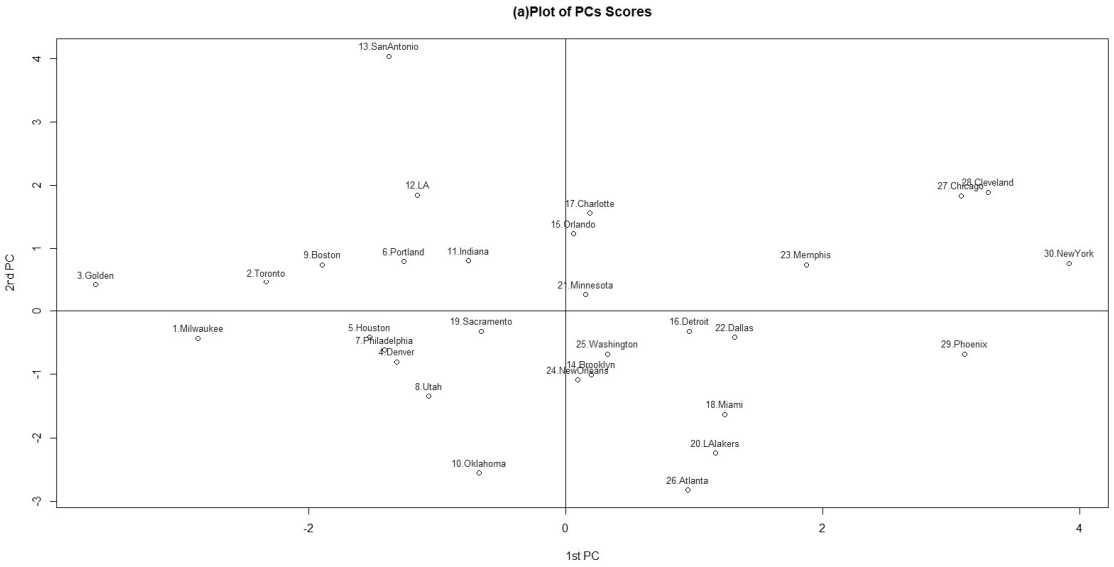
추가적으로 y5, y8은 각각 계수의 절댓값이 0.5에 달하며 자유투 성공률과 턴오버 지표로서 이는 앞서 변수 설명란에서 설명했듯이 경기의 승부를 결정짓는 승부처에 막대한 영향을 끼치는 두 가지 요인이다. 따라서 이 둘을 중심으로 제 2주성분을 설명하자면 제 2주성분은 승부를 결정짓는 승부처 요인으로 언급할 수 있을 것이다.

제 3주성분에서 y3, y6, y7, y8, y9의 계수가 +0.50, +0.46,-0.51, -0.32, -0.37이 나왔다. 이들은 3점슛 시도 횟수와 공격리바운드 개수(OREB), 어시스트 개수(AST), 턴오버 개수(TOV), 스틸(STL)의 개수를 나타내는 지표이다. 제 3주성분은 (3점슛 시도, 공격리바운드)와 (어시스트 개수, 턴오버 개수, 스틸의 개수)의 대비를 나타내는 주성분이라고 할 수 있다.

변수 중에서 3점슛 시도(3PA)와 공격리바운드 지표(OREB)의 계수가 양의 값인 것을 확인할 수 있으며 이는 3점슛 시도로 인해 생긴 롱리바운드 때문에 공격자가 리바운드를 따낼 기회가 많이 생기고 있음을 나타내는 것이라 말할 수 있다.

(2) 주성분 산점도

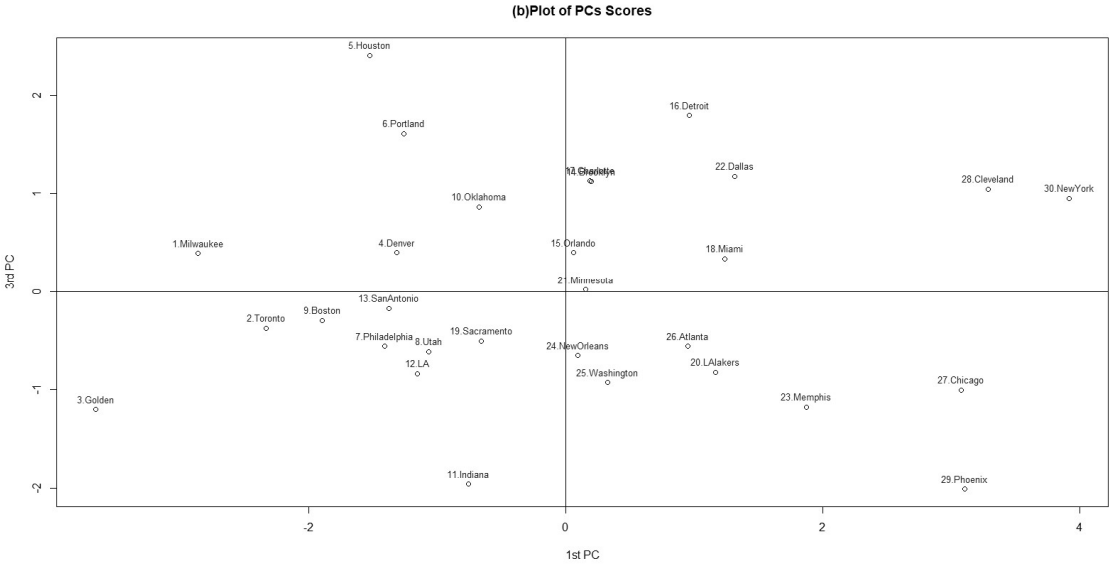
a. PC1 & PC2



1st PC축을 보면, 왼쪽에는 시즌 성적(W)이 높으면서 일반적인 성적이 높은 팀들이 위치해있다. 이름에 붙어있는 번호는 최종 순위를 나타내며 시즌 1, 2, 3위인 밀워키 벅스와 토론토 랩터스, 골든스테이트 워리어스가 위치했음을 볼 수 있다. 반대로 오른쪽에는 시즌 성적(승수)이 낮으면서 일반적인 성적이 낮은 팀들이 위치해 있다. 자세히 보면 27위부터 30위 팀들이 오른편에 위치해 있다.

2nd PC축을 보면, 위쪽에는 자유투 성공 확률이 높으며 턴오버 횟수가 낮은 팀들이 위치해있으며 밑에는 그 와는 반대 성향의 팀들이 위치해 있다. 덧붙이자면 가장 위쪽에 샌안토니오 스퍼스가 위치하며 그 다음으로는 LA clippers(LA를 연고지로 하는 두 팀이 있는데 바로 LA lakers와 LA clippers가 있다. 필자는 LA clippers를 LA를 대표하는 팀이라 생각하여 LA clippers를 LA로 기재하였다.)가 있다. 샌안토니오 스퍼스를 추가적으로 설명하자면 팀 농구를 중요시 여기는 포포비치 감독이 무리한 3점슛 시도(3PA: 음의 계수)는 자제하되 세트 플레이 상황에서 확실한 3점슛을 성공시키도록(3P%: 양의 계수) 선수들에게 요구하는 것을 알 수 있다.(포포비치 감독은 자유투 성공률이 낮은 것을 용서치 않아 샌안토니오 스퍼스 선수들은 대체적으로 평균 이상의 자유투 성공률을 가지고 있다.)

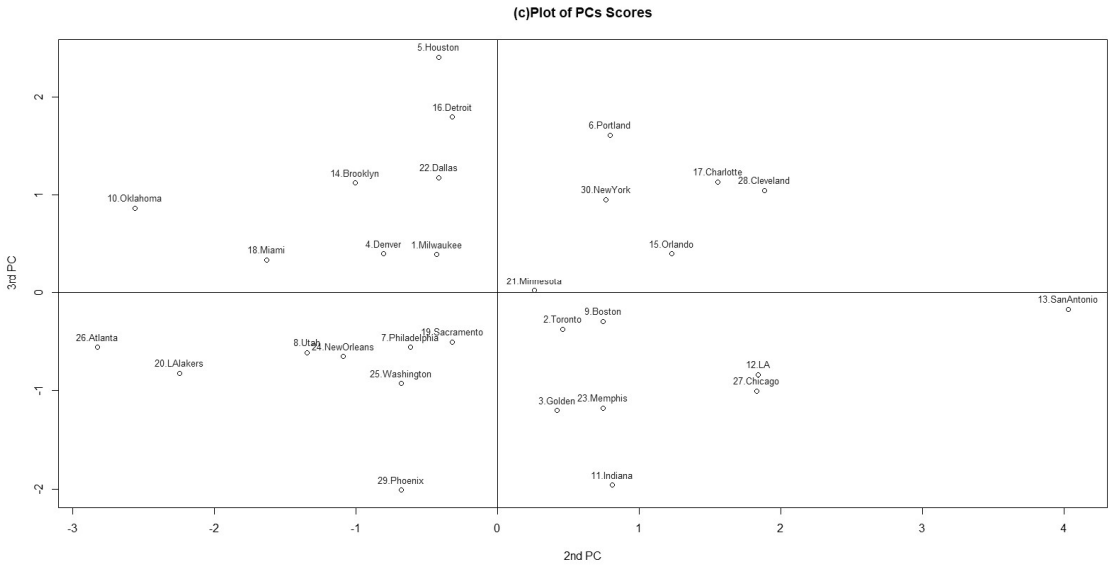
b. PC1 & PC3



1st PC축을 보면, 왼쪽에는 시즌 성적(W)이 높으면서 일반적인 성적이 높은 팀들이 위치해 있다. 반대로 오른쪽에는 시즌 성적(W)이 낮으면서 일반적인 성적이 낮은 팀들이 위치하고 있다.

3rd PC축을 보면 3점슛 시도(3PA), 공격리바운드 횟수(OREB)가 높은 팀들이 있다. 휴스턴이 맨 위에 있는 것을 주목하자. 휴스턴은 슛 시도의 30%에서 40% 정도를 3점슛으로 시도하기에 3점슛 광신도라고 불리는 팀이다(평균 3점슛 시도 45.4). 많은 3점슛 시도로 인해 정상적인 리바운드보다는 롱 리바운드 또한 많이 생기는 편이며 휴스턴의 모든 선수가 리바운드에 적극적으로 참여하여 공격 리바운드를 따내고 다시 3점슛을 시도한다. 아래편에는 그 반대의 경우가 위치해 있는데 여기서 골든 스테이트가 위치해 있다는 점은 예상외의 결과이다. 하지만 원인을 분석해 보자면 제 3주성분은 3점슛 시도와 공격리바운드에는 양의 계수를 갖지만 어시스트 턴오버, 스틸의 개수에는 음의 계수를 갖는다. 골든 스테이트는 스페이싱을 위한 스크린과 컷인을 주로 하기에 어시스트가 높으며 협력 수비로 인한 스틸의 수치 또한 높은 편이다. 반대로 화려한 경기력 뒤에 숨어있는 많은 턴오버로 인해 턴오버 수치가 높아 아래편에 위치하는 이유를 이해할 수 있다.

c. PC2 & PC3

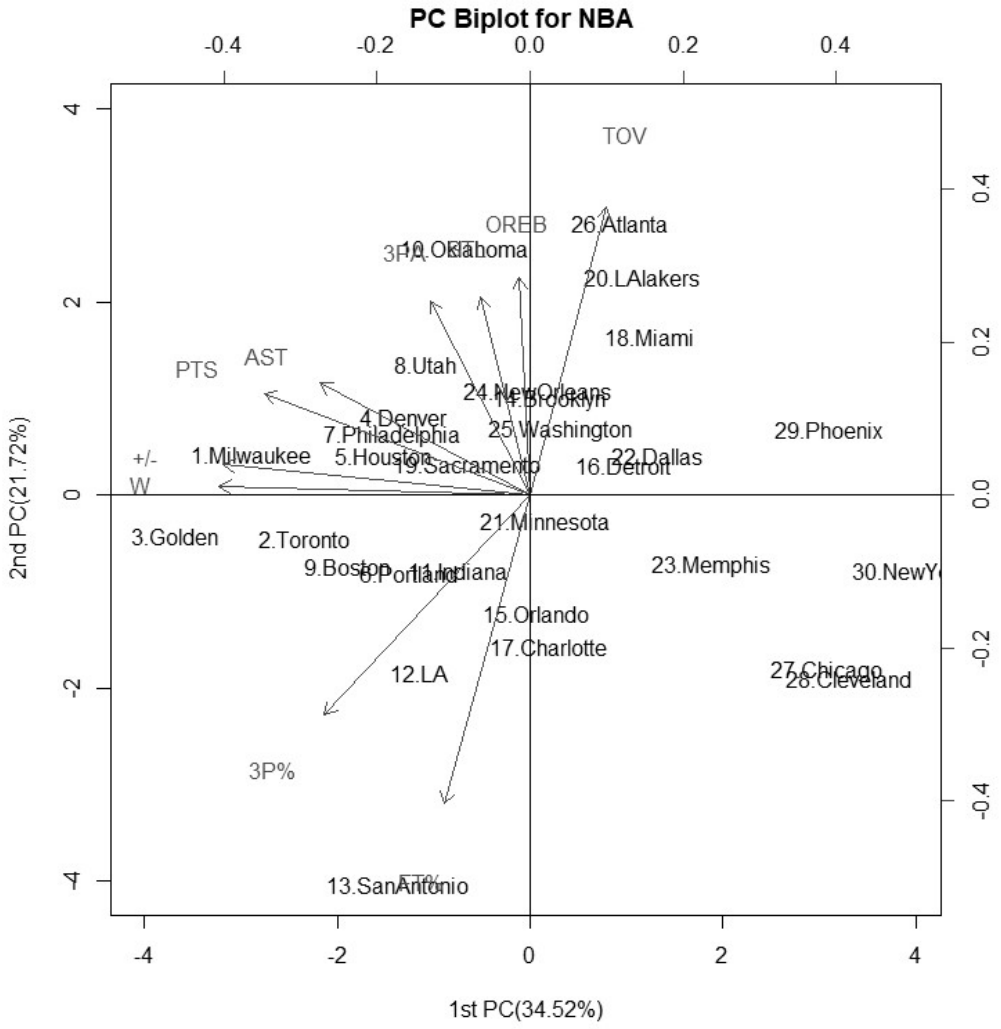


2nd PC축을 보면, 왼쪽에는 자유투 성공 확률(FT%)이 높으며 턴오버 횟수(TOV)가 낮은 팀들이 위치해 있으며 오른쪽에는 그 와는 반대 성향의 팀들이 위치해 있다.

3rd PC축을 보면 3점슛 시도(3PA), 공격리바운드 횟수(OREB)가 높은 팀들이 있다. 아래편에는 그 반대의 경우가 위치해 있다.

(3) 주성분 행렬도

a. PC1 & PC2



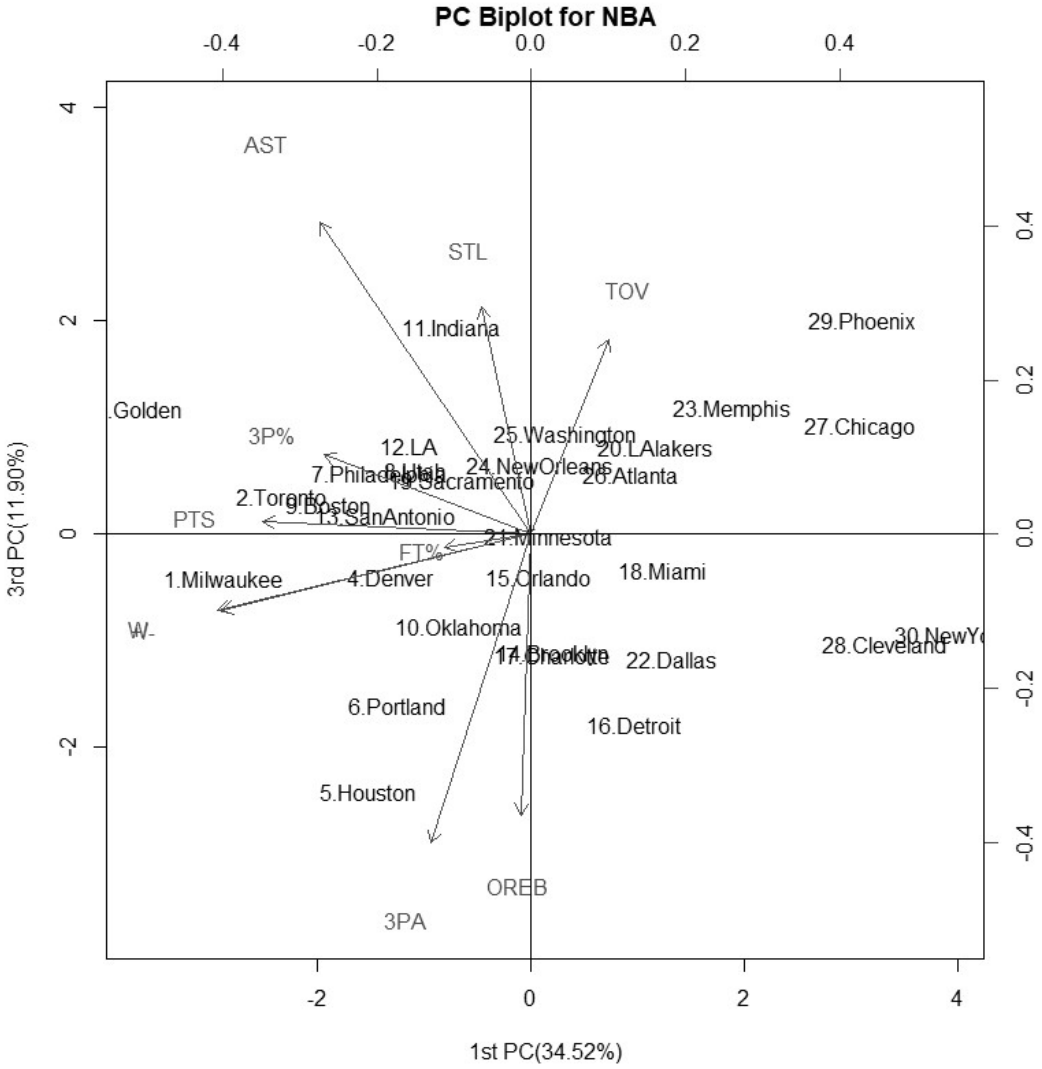
제 1st PC축(34.52%)과 2nd PC축(21.72%)에 의해서 행렬도의 설명력은 56.24%가 된다.

PC1 & PC2 주성분 행렬도에서 변수들의 화살표를 보자면, TOV, OREB, STL, 3PA 변수 화살표 사이의 각이 좁아 네 변수 사이의 상관관계가 어느 정도 높다고 할 수 있다. 또한 AST, PTS, +/-, W 변수 화살표 사이의 각이 좁아 상관관계가 어느 정도 높음을 알 수 있다. 반대로 TOV, FT%와 PTS, AST는 서로 직각을 이루고 있으므로 상관관계가 거의 없다고 할 수 있다.

제 1st PC축 왼편에 축에 가까운 1위부터 7위까지의 팀들은 AST, PTS, +/-, W 변수에 많은 영향을 받는 팀들이며 이는 시즌 성적(승수)이 높은 팀들이 평균 득점과 어시스트, 평균 득실점 마진이 높음을 알 수 있게 해준다.

제 1사분면에 있는 팀들은 TOV의 영향을 많이 받는 팀들이다. 턴오버의 영향을 많이 받은 팀들일수록 시즌 성적이 낮음을 알 수 있으며 제 1주성분에 대해 설명했을 때와 마찬가지로 TOV는 팀 승리와는 대비를 이루는 요인이라는 것을 다시 한번 확인할 수 있다.

b. PC1 & PC3

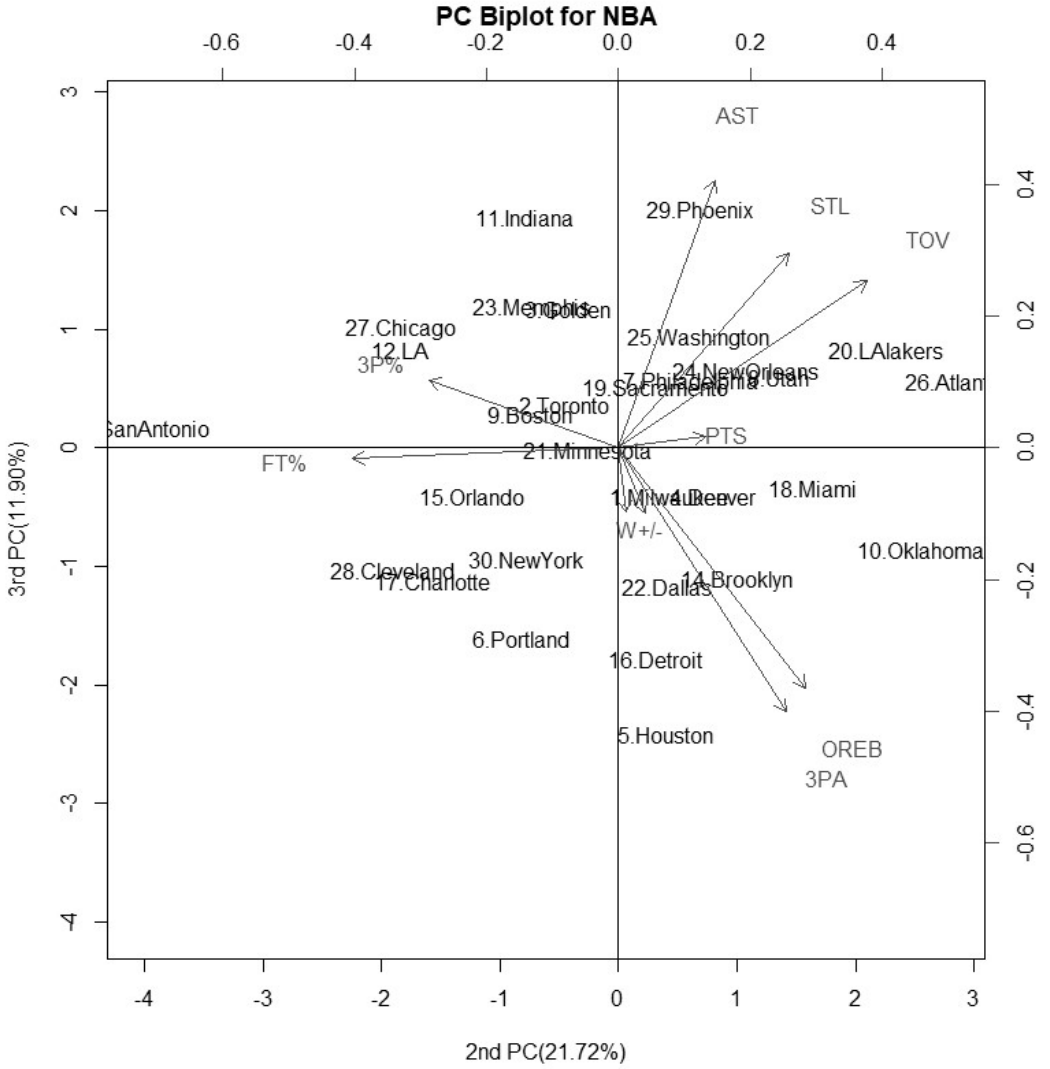


제 1st PC축(34.52%)과 3rd PC축(11.90%)에 의해서 행렬도의 설명력은 45.61%가 된다.

PC1 & PC3 주성분 행렬도에서 TOV, STL, AST변수 사이의 각이 좁아 이 세 변수 사이의 상관관계가 어느 정도 높은 것을 확인할 수 있다. 이와 마찬가지로 OREB, 3PA 두 변수. 그리고 3P%, PTS, FT%, W, +/- 네 변수의 각이 좁아 서로 상관관계가 어느 정도 높음을 알 수 있다. 반대로 TOV, 3PA와 3P%는 서로 직각을 이루고 있으므로 상관관계가 거의 없다고 할 수 있다.

제 3사분면을 보면 3PA의 변수의 화살표 근처에 5위 휴스턴과 6위 포틀랜드가 위치하고 있는데 이는 PC1 & PC3 주성분 산점도에서 봤듯이 휴스턴과 포틀랜드가 3점슛 시도와 공격리바운드 수치가 높음을 확인할 수 있다.

c. PC2 & PC3



제 2nd PC축(21.72%)과 3rd PC축(11.90%)에 의해서 행렬도의 설명력은 33.62%가 된다.

PC2 & PC3 주성분 행렬도에서 AST, STL, TOV 세 변수 사이의 각이 좁아 서로의 상관관계가 어느 정도 높은 것을 알 수 있다. 마찬가지로 3P%, FT% 두 변수. 그리고 OREB, 3PA, PTS, W, +/- 네 변수의 각이 좁아 서로 상관관계가 어느 정도 높음을 알 수 있다. 반대로 AST와 3P%, TOV와 (OREB, 3PA)는 서로 직각을 이루고 있으므로 상관관계가 거의 없다고 할 수 있다.

행렬도의 왼쪽을 보면 3P%와 FT% 변수의 화살표 방향 끝에 샌안토니오 스퍼스가 위치하고 있다. 이는 PC1 & PC2 주성분 산점도에서 앞서 설명했듯이, 포포비치 감독이 확실한 상황에서의 3점슛 시도와 가장 손쉬운 자유투 득점을 얼마나 중요시 여기고 선수들을 훈련시키는지 알 수 있게 해주는 자료이다(반대로 포포비치 감독은 무리한 3점슛은 끔찍하게 싫어해서 수비를 제대로 안 한다던가 3점 슛을 막 쐈다간 경기장에서 포포비치 감독의 고성을 심심치 않게 들을 수 있다).

5. 인자 분석 (FA)

인자 분석은 주성분 분석 PCA처럼 다변량 자료부터 차원축소 개념을 실현하는 분석기법이다. 공통인자분해를 통해 변수들에 대해 공통인자들의 선형결합 형태인 모형의 인자적재를 추정한다. 크게 주성분인자분석(PCFA)과 최대우도인자분석(MLFA)가 있다.

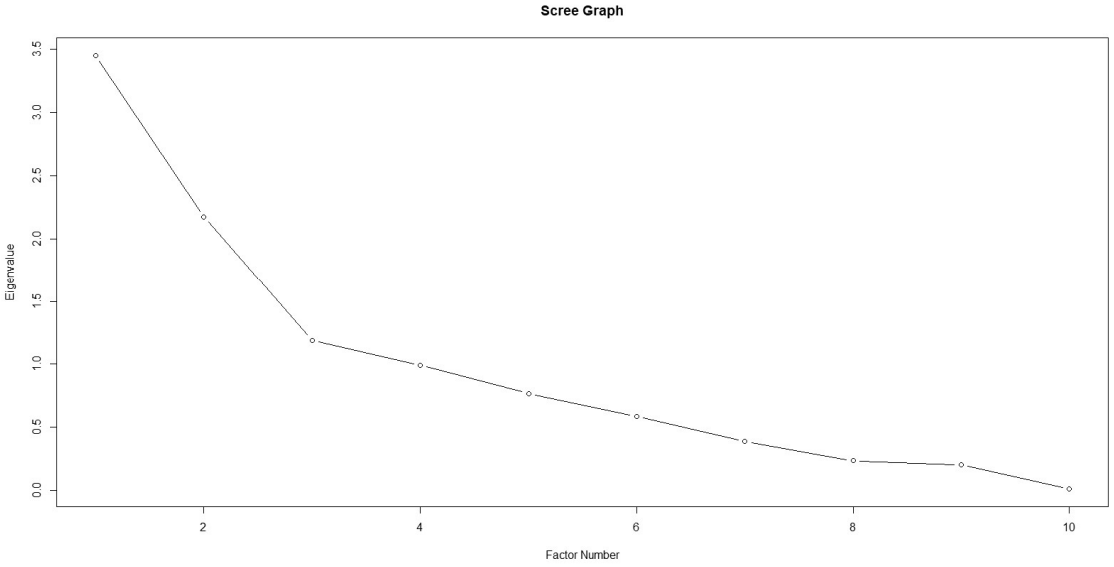
추가적으로 추정된 공통인자들을 구조적 공통인자분해가 유지되는 직교변환을 통해서 보다 쉽게 해석할 수 있다.

```
> round(eigen2$values ,3)
 [1] 3.452 2.172 1.190 0.991 0.769 0.587 0.390 0.234 0.204 0.011
> round(gof,3)
 [1] 34.513 21.714 11.905  9.909  7.689  5.870  3.903  2.349  2.036  0.111
```

주성분 분석에서와 마찬가지로 상관행렬 R의 Spectral Decomposition을 통해 구한 고윳값과 전체 고윳값 중에서 차지하는 총 설명력의 값이다.

이를 바탕으로 공통인자의 개수를 결정할 수 있다. 최대 고윳값은 l_1 =3.452이며 전체 고윳값에서 약 34.513%를 설명하고, 그 다음으로 l_2 =2.172와 l_3 =1.190의 고윳값이 있으며 전체 고윳값에서 21.714%와 11.905%를 설명한다.

이를 scree plot에서 한 번 더 확인할 수 있다.



위의 그림을 살펴보면 세 번째 고윳값 이후로 기울기가 완만해지므로 세 번째 고윳값이 팔꿈치가 되며 두 번째 고윳값까지를 주성분 요인으로 결정해야겠지만 네 번째 고윳값 이후로도 기울기가 완만하며 총 설명력을 고려하여 세 번째 고윳값까지 인자로 고려해도 된다 판단했다.

① PCFA의 인자 적재값 vs 회전 전의 MLFA 인자 적재값

```
> round(L,3)
        PC1    PC2    PC3
W     0.945  0.021 -0.134
PTS   0.808  0.244  0.023
3PA   0.302  0.467 -0.547
3P%   0.625 -0.530  0.141
FT%   0.262 -0.743 -0.023
OREB  0.032  0.524 -0.497
AST   0.638  0.269  0.553
TOV  -0.232  0.693  0.346
STL   0.151  0.477  0.403
+/-   0.933  0.073 -0.135
```

```
> round(Lm, 3)
      Factor1 Factor2 Factor3
W       0.875   0.453   0.078
PTS     0.465   0.528   0.169
3PA     0.213   0.354  -0.437
3P%     0.556  -0.010   0.762
FT%     0.320  -0.360   0.335
OREB   -0.027   0.180  -0.178
AST     0.311   0.533   0.150
TOV    -0.667   0.742   0.011
STL     0.014   0.251  -0.216
+/-     0.886   0.457  -0.032
```

PCFA(왼쪽)와 MFLFA(오른쪽)의 인자 적재 값이다. 우선 factor1은 TOV를 제외한 모든 변수의 부호가 동일하므로 턴오버(TOV)와 턴오버를 제외한 지표들의 대비를 나타내는 성분이라고 말할 수 있다. factor2를 보았을 때, 3PA(3점슛 성공률)와 STL(스틸)에서 인자 적재 값이 확연하게 커져 다른 변수들과 대비된다. W(승수)와 +/-(경기당 평균 득실점 마진)는 더 명확하게 작아졌음을 확인할 수 있다. 따라서 factor2 인자에서 대비되는 변수가 PCFA에서 더 확실하게 나타나므로 PCFA보다는 회전 전의 MLFA가 결과를 확인하는데 용이하다고 말할 수 있다.

② 회전 전의 MLFA 적재 값 / 그림 vs 회전 후의 MLFA 인자 적재 값 / 그림

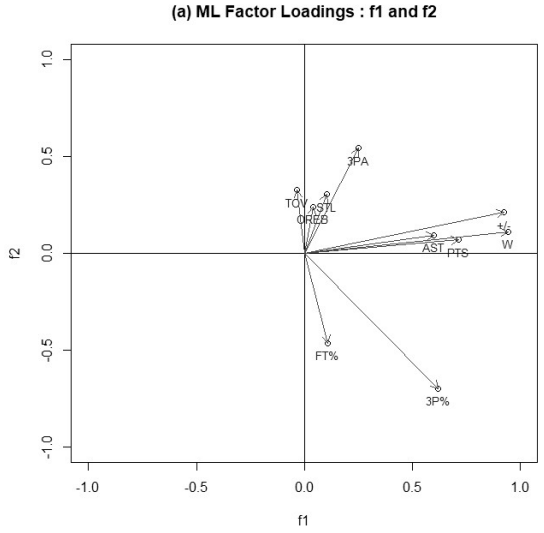
```
> round(Lm, 3)
      Factor1 Factor2 Factor3
W       0.875   0.453   0.078
PTS     0.465   0.528   0.169
3PA     0.213   0.354  -0.437
3P%     0.556  -0.010   0.762
FT%     0.320  -0.360   0.335
OREB   -0.027   0.180  -0.178
AST     0.311   0.533   0.150
TOV    -0.667   0.742   0.011
STL     0.014   0.251  -0.216
+/-     0.886   0.457  -0.032
```

```
> round(Lm.r, 3)
      Factor1 Factor2 Factor3
W       0.945   0.111  -0.267
PTS     0.715   0.070   0.089
3PA     0.250   0.543  -0.058
3P%     0.619  -0.700  -0.128
FT%     0.110  -0.464  -0.340
OREB    0.040   0.239   0.077
AST     0.599   0.093   0.190
TOV    -0.033   0.327   0.942
STL     0.103   0.304   0.083
+/-     0.925   0.212  -0.309
```

회전 전의 MLFA 적재 값(왼쪽)과 회전 후의 MLFA 인자 적재 값이다. Factor1을 비교했을 때 FT%(자유투 성공률)와 TOV(턴오버)를 제외한 모든 인자 적재 값이 커지고 이 둘의 인자 적재 값은 더 작아져 비교를 명확히 할 수 있다. 또한 회전을 통한 변수들의 화살표가 공통인자 축에 가까워지므로 공통인자인 변수들이 더욱 명확해졌다. 따라서 회전을 한 후의 MLFA를 통한 인자 분석을 실시 하겠다.

(1) 인자 적재 값 해석

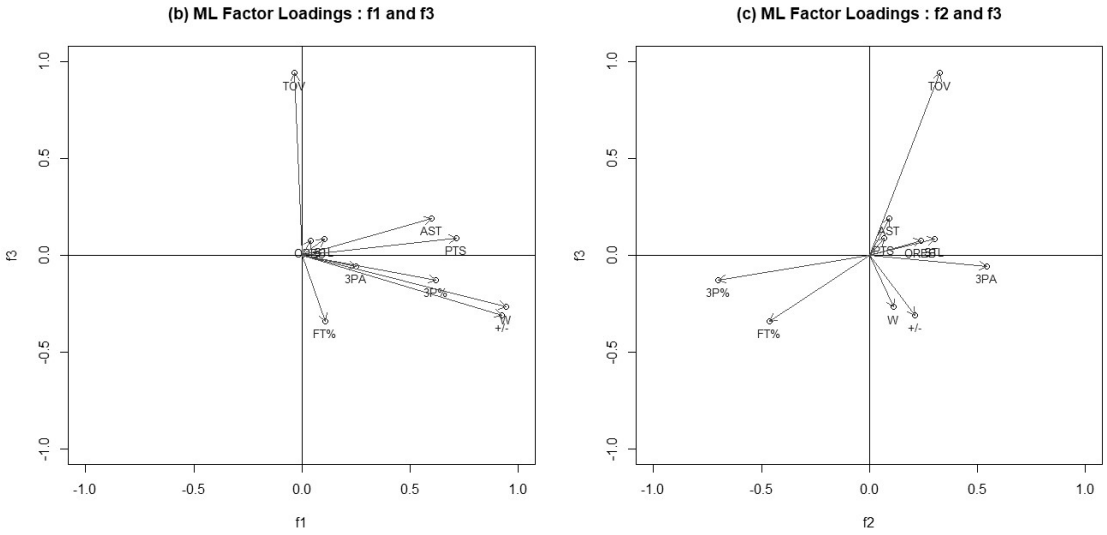
```
> round(Lm.r, 3)
     Factor1 Factor2 Factor3
W      0.945   0.111  -0.267
PTS    0.715   0.070   0.089
3PA    0.250   0.543  -0.058
3P%    0.619  -0.700  -0.128
FT%    0.110  -0.464  -0.340
OREB   0.040   0.239   0.077
AST    0.599   0.093   0.190
TOV   -0.033   0.327   0.942
STL    0.103   0.304   0.083
+/-    0.925   0.212  -0.309
```



factor1을 봤을 때, TOV를 제외한 모든 변수들의 부호가 양수이다. 그리고 OREB(공격 리바운드)와 STL 변수의 계수는 0.04와 0.103이므로 공통인자의 영향이 매우 적음을 알 수 있다. 반면 W와 +/-변수의 계수는 0.945와 0.925로서 매우 높은 수치임을 알 수 있다. 그리고 그에 못지 않게 PTS, 3P%, AST의 계수 또한 높은 인자 적재 값을 가지고 있다.

축을 기준으로 봤을 때, AST, PTS, +/-, W, 3P%는 f1축과 각이 좁은데 이는 factor1이 AST, PTS, +/-, W, 3P%에 가중을 둔 인자로 해석할 수 있다. f2축은 STL, OREB, 3PA변수와의 각이 좁은데 이는 factor2가 3P%, FT%, STL, OREB, 3PA에 가중을 둔 인자로 해석된다고 말할 수 있다.

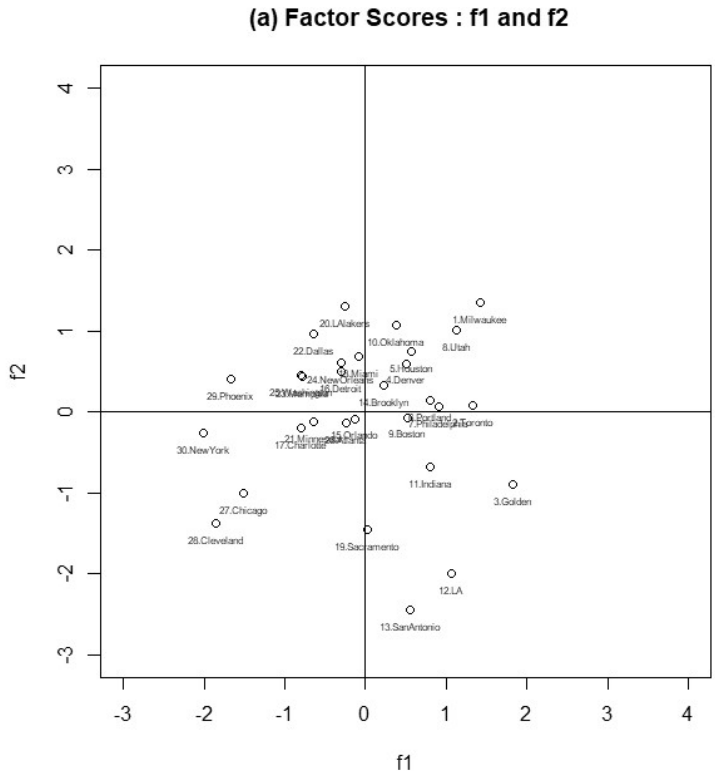
factor1에서 AST, PTS, +/-, W, 3P%는 서로 비례함을 알 수 있으며 factor2에서는 인자 적재 값과 화살표를 통해 3P%, FT%와 STL, OREB, 3PA가 반비례함을 알 수 있다.



f3축을 봤을 때, f3축은 TOV와 이루는 각이 좁은 것을 확인할 수 있다. 인자 적재 값을 확인해보면 TOV에서 0.942이다. 따라서 factor3는 TOV에 가중을 둔 인자로 해석된다고 말할 수 있다. 또한 인자 적재 값을 통해 TOV가 W, FT%, +/-와 반비례함을 알 수 있었다.

(2) 인자 점수그림

앞서 인자 적재 값을 해석했다. factor1은 AST, PTS, +/-, W, 3P%에 가중을 둔 인자로, factor2는 3P%, FT%, STL, OREB, 3PA에 가중을 둔 인자로, factor3는 TOV에 가중을 둔 인자로 해석했다. 이를 토대로 인자 점수그림을 해석하겠다.



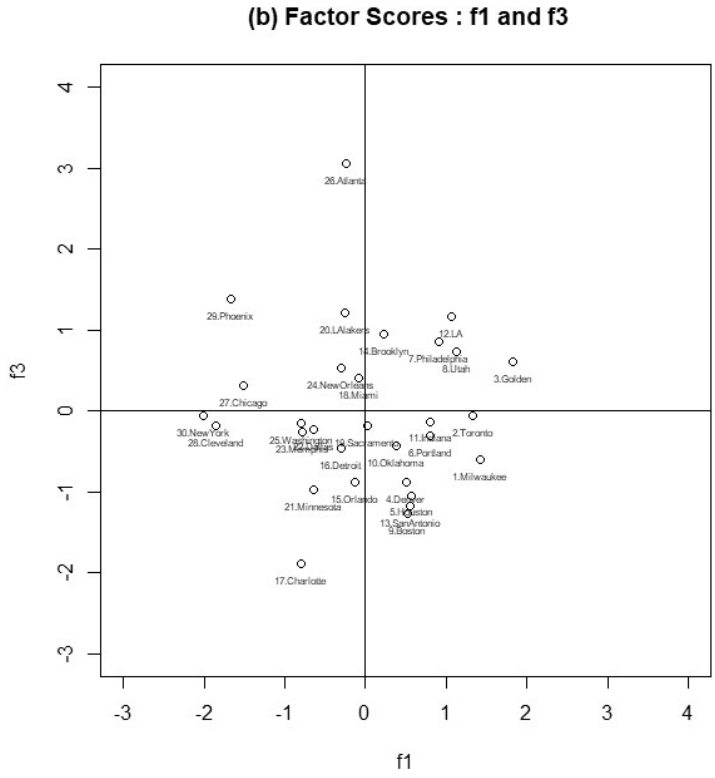
제 1축을 기준으로 봤을 때, factor1은 AST, PTS, +/-, W, 3P%에 가중을 둔 인자로 우측에 있는 팀들은 AST, PTS, +/-, W, 3P%에 있어서 높은 수치를 지니고 있는 팀으로 볼 수 있다. 반대로 좌측은 AST, PTS, +/-, W, 3P%에 있어서 낮은 수치를 지니고 있는 팀으로 볼 수 있다. 실제로 우측에는 1위부터 14위까지 플레이오프 시리즈에 진출한 팀들이 위치한 것을 볼 수 있으며 좌측은 플레이오프 시리즈에 탈락하거나 마지막까지 플레이오프 시리즈에 진출하기 위해 경쟁한 하위권 팀들이 위치한 것을 확인할 수 있다.(실제로 이번 시리즈 마지막 경기까지 동부 팀인 마이애미와 디트로이트, 샬롯, 올랜도가 경쟁을 했고 결국 디트로이트와 올랜도가 플레이오프에 진출하였다. 물론 두 팀 다 전반적인 전력 격차 때문에 1차전에서 탈락했다.)

제 2축을 기준으로 봤을 때, factor2는 3P%, FT%, STL, OREB, 3PA에 가중을 둔 인자로 상단에 있는 팀들은 STL, OREB, 3PA에 있어 높은 수치를, 3P%, FT%에 있어 낮은 수치를 팀으로 볼 수 있다. 반대로 하단에 있는 팀들은 STL, OREB, 3PA에 있어서 낮은 수치를, 3P%, FT%에 있어 높은 수치를 지니고 있는 팀으로 볼 수 있다.

1위부터 10위권에 위치한 팀들은 제 1사분면에 위치하고 있는데 유일하게 골든스테이트가 제 2사분면에 위치하고 있는데 이는 3점슛 팀으로 불리고 있는 골든스테이트가 높은 3점슛 성공률(음의 계수)과 자유투 성공률(음의 계수)을 보이며 많은 생각보다 많은 3점슛을 시도하지는 않는다는 것을 알 수 있게 해준다. 1사분면에 위치하고 있는 팀들은 factor2에 있어서도

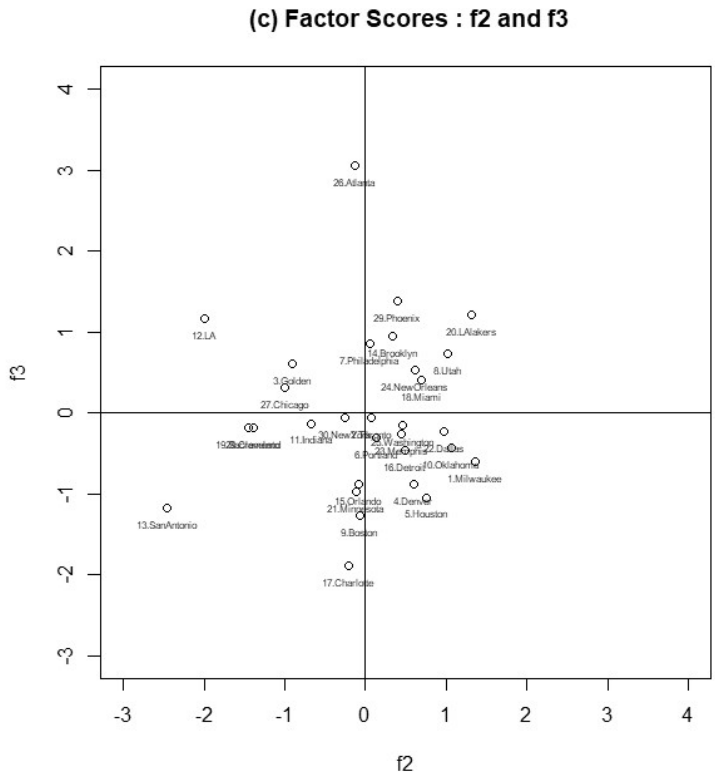
높은 수치를 지고 있다. 특히 밀워키 벅스와 오클라호마 시티가 factor2에 있어 높은 곳에 위치하고 있다. 이 두팀이 많은 3점슛을 시도하고 공격권을 얻기 위해 공격 리바운드에 적극적으로 참여하며 강력한 수비를 통해 스틸을 이끌어내는 것을 알 수 있게 해준다.

b. f1 & f3



아까와 같이 제 1축을 봤을 때 factor1은 AST, PTS, +/-, W, 3P%에 가중을 둔 인자로 우측에는 이들 수치가 높은 팀이, 좌측에는 낮은 팀이 위치하고 있다. 제 3을 기준으로 봤을 때, factor3는 TOV에 가중을 둔 인자로 상단에는 TOV의 수치가 높은 팀들이 하단에는 TOV의 수치가 낮은 팀들이 위치하고 있다. 특히 26위의 애틀란타가 앞도적으로 높은 곳에 위치하고 있다. 이는 애틀란타가 이번 시리즈 5순위로 지명된 트레이 영에게 주전 포인트 가드 자리와 경기 운영에 대한 전권을 부여해줬기 때문인데 특히나 트레이 영은 이번 시즌 데뷔한 신인이면서도 제 2의 스테판 커리로 불리는 만큼 많은 3점슛을 시도해서 턴오버 수치가 높을 수밖에 없다.(하지만 턴오버가 시즌 후반기가 되면서 절반에 가까이 떨어졌는데 이는 트레이 영의 성장 속도가 생각보다 빨리 진행되고 있는 것을 알려준다. 다음 시즌 애틀란타의 턴오버 개수가 궁금해지는 대목이기도 하다.)

C. f2 & f3



아까와 같이 제 2축을 기준으로 봤을 때, factor2는 3P%, FT%, STL, OREB, 3PA에 가중을 둔 인자로 우측에 있는 팀들은 STL, OREB, 3PA에 있어 높은 수치를, 3P%, FT%에 있어 낮은 수치를 지니고 있는 팀으로 볼 수 있으며 좌측은 그 반대의 팀들이 위치하고 있는 것으로 볼 수 있다. factor3는 TOV에 가중을 둔 인자로 상단에는 TOV의 수치가 높은 팀들이 하단에는 TOV의 수치가 낮은 팀들이 위치하고 있다.

여기서 주목해야할 점은 1, 4, 5위 팀들인 밀워키 벅스와 덴버 너게츠, 휴스턴 로케츠이다. 이들의 순위와 factor2가 높은 수치임에도 불구하고 TOV 수치에서 많이 낮은 것을 확인할 수 있다. 이는 변수 설명에서도 언급했듯이 턴오버가 경기의 승패에 더불어 시즌 성적에 얼마나 많은 영향을 끼치는지 알 수 있게 해주는 자료이다. 왜냐하면 TOV 수치가 낮은 팀들 속에서도 하위권 팀들이 산포해 있지만 상위권 팀으로만 분류했을 때 이들 중에서 턴오버 관리를 잘한 팀이 더 높은 순위를 기록하고 있기 때문이다.

샌안토니오 스퍼스는 제 3사분면에 위치하고 있는데 PCA 주성분 산점도에서 설명했을 때와 마찬가지로 무리한 3점슛은 자제하되 확실한 세트 플레이를 통한 완성도 높은 공격을 하므로 높은 3점슛 성공률과 자유투 성공률, 낮은 턴오버 수치를 기록한다고 말할 수 있다.
(샌안토니오 스퍼스의 시즌 3점슛 성공률이 39.2%이다! 3점슛 팀으로 불리는 골든스테이트와 휴스턴의 3점슛 성공률은 각각 38.5%, 35.6%이다.)

6. CA (Cluster Analysis)

군집 분석은 자료 간의 유사도를 정의하고 그 유사도에 가까운 것부터 순서대로 합쳐가는 방법이다. 군집화는 유사성 또는 거리를 근거로 이루어진다. 군집 분석에는 위계적 군집방법과 비위계적 군집방법이 있다. 위계적 군집방법은 단일연결, 완전연결, 평균연결, ward연결이 있다. 비위계적 군집방법에는 K-평균법이 있다.

(1) 위계적 군집방법

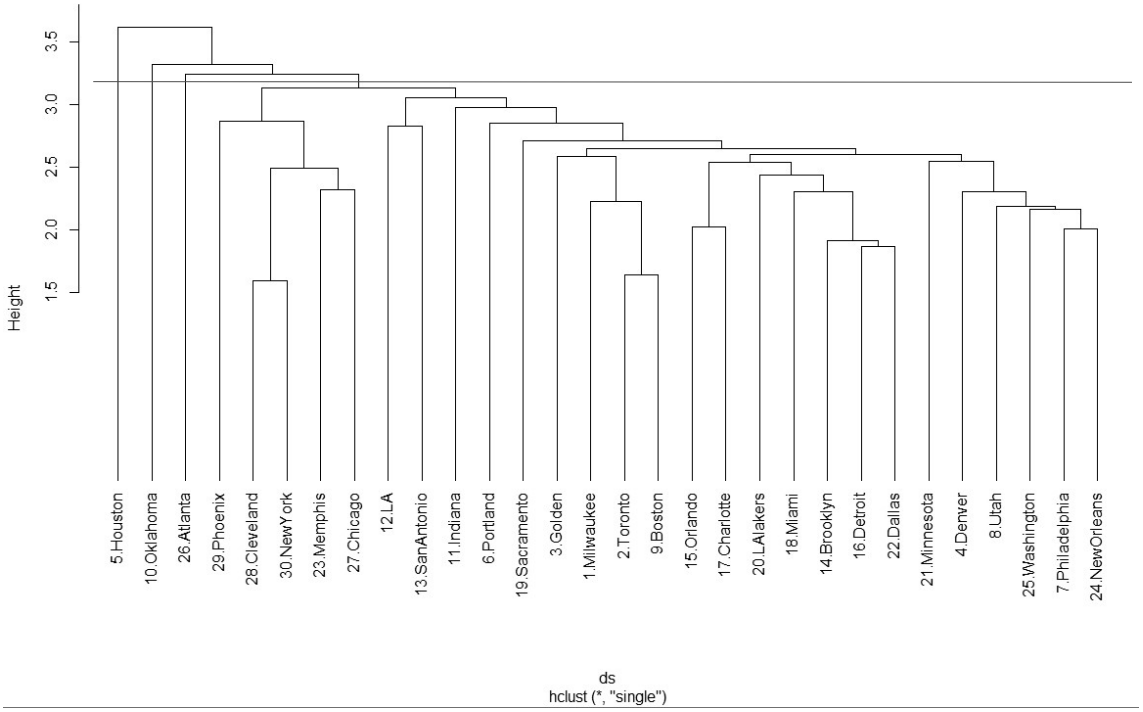
① Single Linkage(단일연결)

개체들 간의 거리 또는 유사성을 이용하여 최단거리를 갖는 쌍을 하나의 군집으로 묶는다. 최단거리는 거리가 가장 짧거나 가장 유사한 것을 의미한다. 덴드로그램을 보이기에 앞서, 군집의 수를 판단하기 위해 NbClust에서 allindex를 구해보았다.

```
> allindex <- NbClust(Z, distance= "euclidean", min.nc=2 , max.nc=8,
+                     method = "single" , index="all")
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 7 proposed 2 as the best number of clusters
* 2 proposed 3 as the best number of clusters
* 3 proposed 4 as the best number of clusters
* 6 proposed 5 as the best number of clusters
* 4 proposed 6 as the best number of clusters
* 2 proposed 8 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```

단일 연결법에서 가장 적절한 군집의 수는 두 개이며, 이를 바탕으로 군집을 나누겠다.

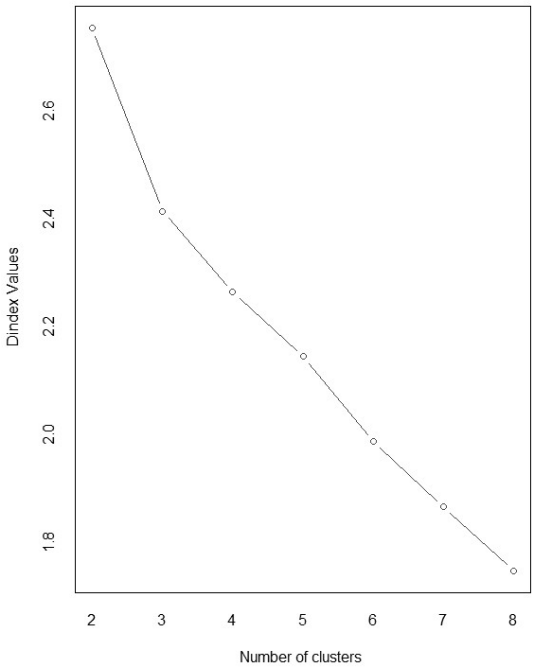
(a) Sinle Linkage

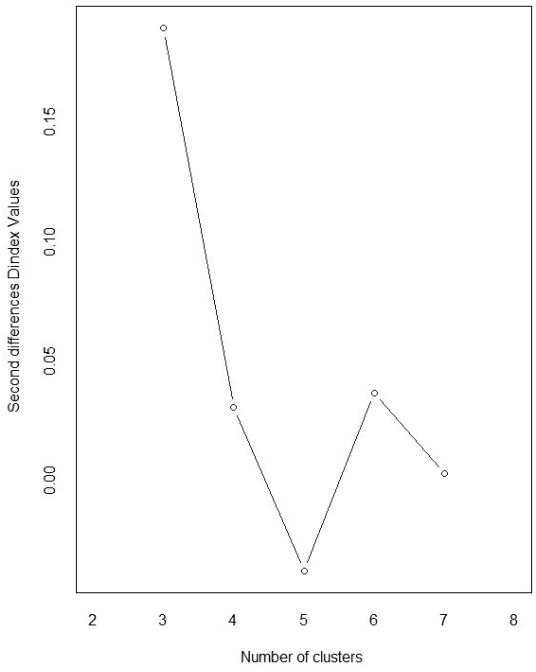


② 완전연결법

완전연결은 한 가지 사실만 빼고는 단일연결과 똑같은 방법으로 수행된다. 각 단계에서 군집 간의 거리는 두 군집에서 가장 멀리 떨어진 개체들의 거리에 의해 정의된다. 즉 하나로 병합된 군집의 모든 개체들은 병합 이전 두 군집간의 최대 거리 이내에 있다고 볼 수 있다.
완전 연결법에서도 마찬가지로 allindex를 이용하여 군집의 수를 결정했다.

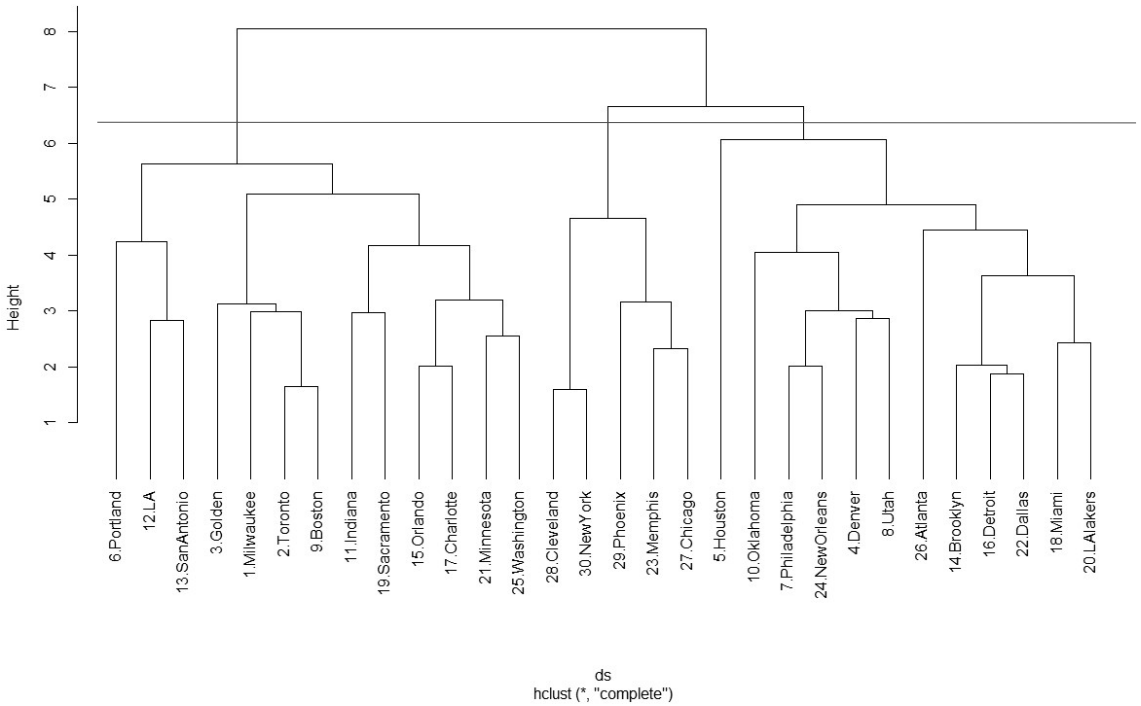
```
> allindex <- NbClust(Z, distance= "euclidean", min.nc=2 , max.nc=8,
+                     method = "complete" , index="all")
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 3 proposed 2 as the best number of clusters
* 13 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 4 proposed 8 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3
```





완전 연결법에서 가장 적절한 군집의 수는 세 개이며, 이를 바탕으로 군집을 나누겠다.

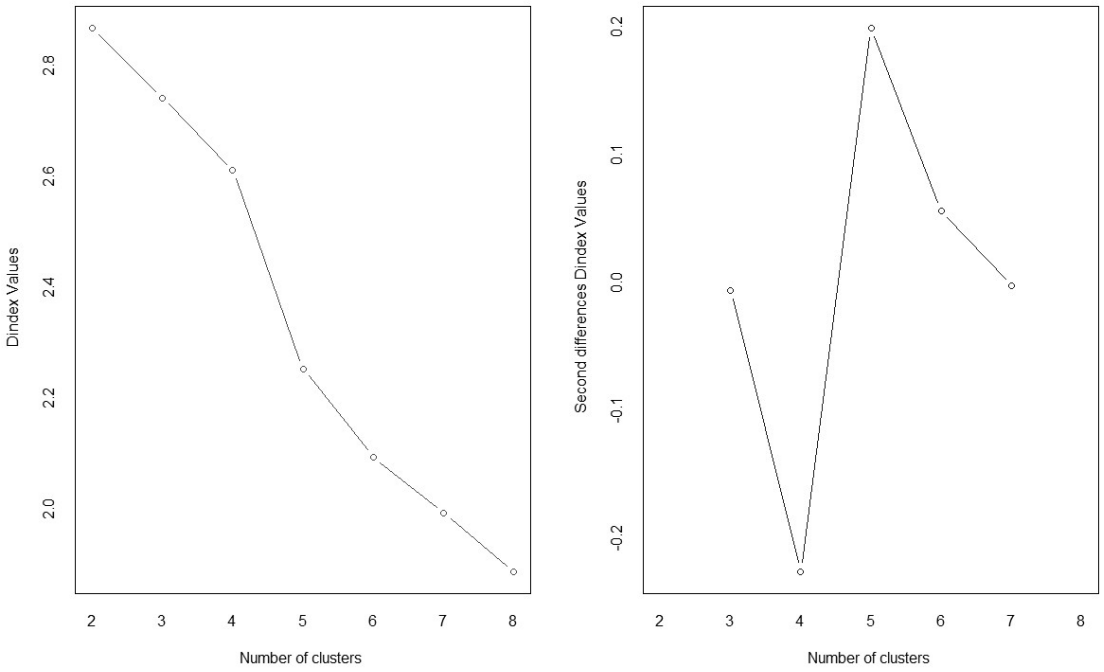
(b) Complete Linkage



③ 평균 연결법

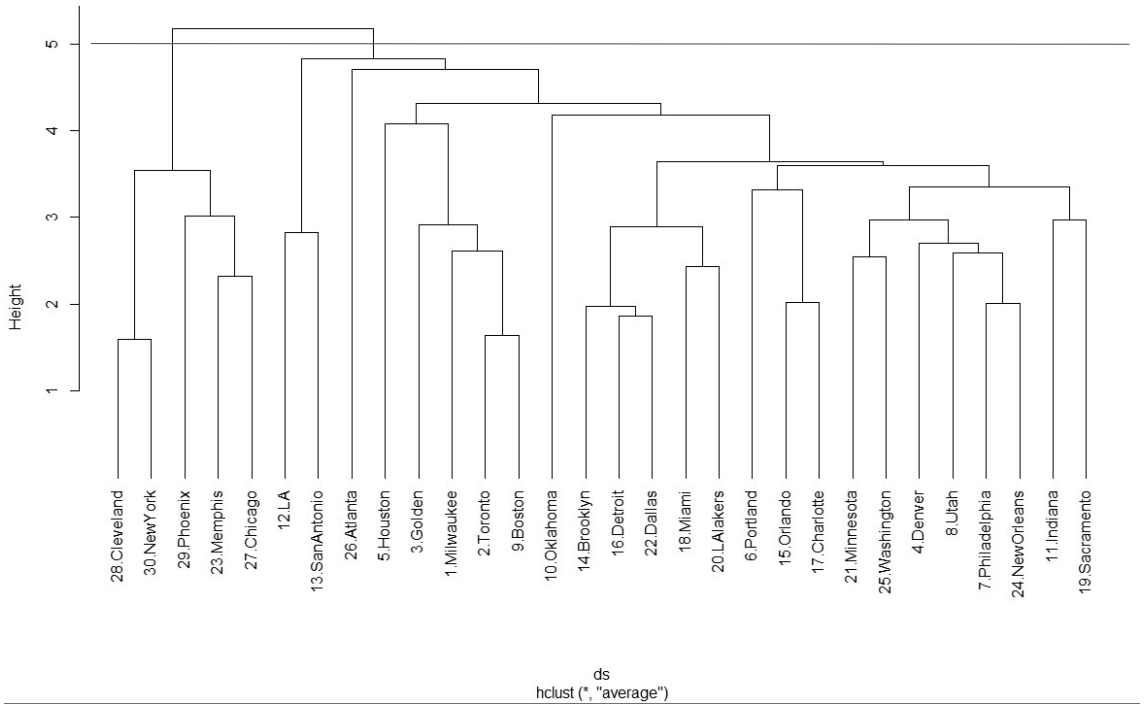
평균 연결은 서로 다른 두 군집에 속하는 개체들의 모든 쌍의 거리를 평균한 평균거리를 이용하여 군집화한다. 이 연결법에서도 거리 또는 유사성 측도를 입력 자료를 사용할 수 있으며 개체뿐만 아니라 변수들도 군집화에 사용할 수 있다.

```
> allindex <- NbClust(Z, distance= "euclidean", min.nc=2 , max.nc=8,
+                     method = "average" , index="all")
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 7 proposed 2 as the best number of clusters
* 3 proposed 3 as the best number of clusters
* 6 proposed 5 as the best number of clusters
* 3 proposed 7 as the best number of clusters
* 4 proposed 8 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```



평균 연결법에서 가장 적절한 군집의 수는 두 개이며, 이를 바탕으로 군집을 나누겠다.

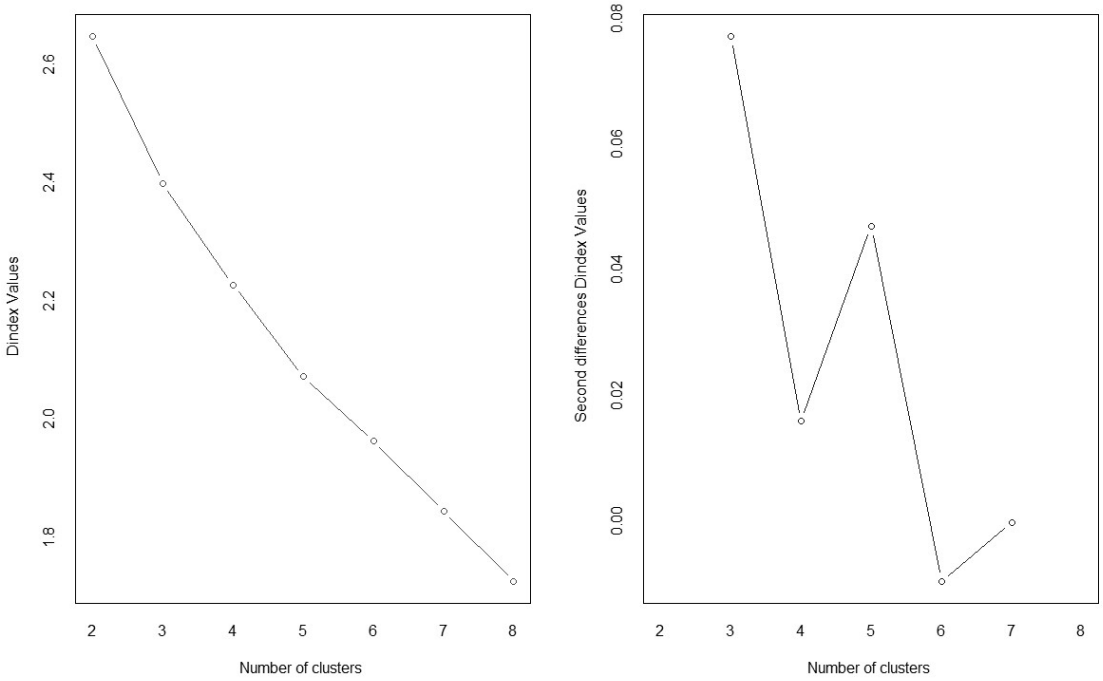
(c) Average Linkage



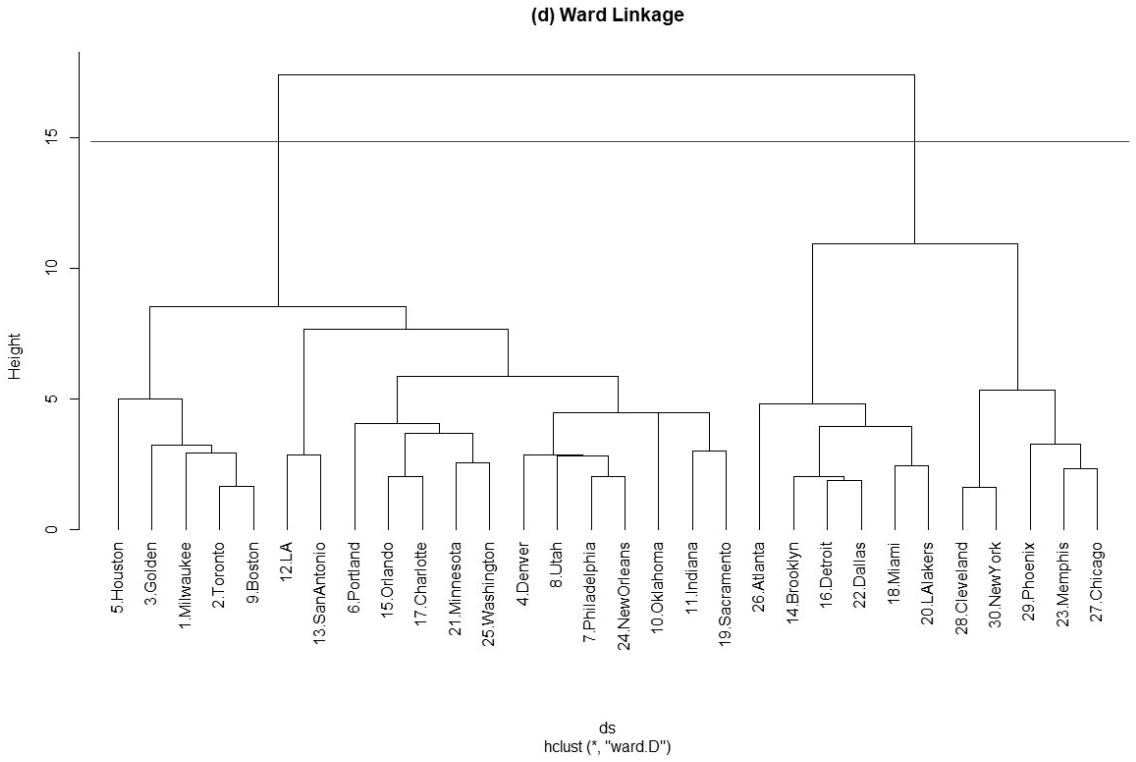
④ 와드 연결법

개체들을 병합하여 묶어 나가면 새로운 군집이 생길 때마다 정보의 손실이 초래된다. Ward는 군집 평균과 개체간의 편차제곱합에 의해 이러한 정보의 손실을 측정한 것이다.

```
> allindex <- NbClust(Z, distance= "euclidean", min.nc=2 , max.nc=8,
+                     method = "ward.D" , index="all")
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 6 proposed 2 as the best number of clusters
* 5 proposed 3 as the best number of clusters
* 4 proposed 4 as the best number of clusters
* 4 proposed 5 as the best number of clusters
* 4 proposed 8 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```



와드 연결법에서 가장 적절한 군집의 수는 두 개이며, 이를 바탕으로 군집을 나누겠다.




```
> colMeans(nba)
          W         PTS         3PA         3P%         FT%       OREB         AST
41.00000000 111.20333333  32.00666667  35.55000000  76.69666667  10.34666667  24.58000000
        TOV         STL         +/-
14.07666667   7.62333333   0.01333333
```

위계적 군집 방법으로 단일 연결법, 완전 연결법, 평균 연결법, 와드 연결법을 실행했다. 단일 연결법을 통해 2개의 군집을 얻었지만 군집 2에 너무 많은 관측치가 몰려 군집의 뚜렷한 특성을 파악하기 쉽지 않았다. 완전 연결법에서는 세 개의 군집을 얻었지만 최하위권 팀들이 모여 있는 군집 2를 제외하고 군집 1과 3의 특성을 파악하기 쉽지 않았다. 평균 연결법에서는 두 개의 군집을 얻었으며 최하위권 팀들이 모인 군집 1과 그 외 팀들이 모인 군집 2를 얻었다. 그러나 군집 2에 너무 많은 관측치들이 밀집하였다. 와드 연결법에서는 두 개의 군집을 얻었으며 각 군집은 특성이 확실하게 나타나 있다. 군집 1에는 21, 24, 25위를 제외한 상위권 팀들이 있으며 군집 2에는 14, 16위를 제외한 하위권 팀들이 있다. 각 군집에 관측치들이 고루 분포해 있는 것을 고려했을 때, 위계적 군집 방법에서는 와드 연결법이 군집을 나누는데 가장 효과적이라고 판단할 수 있다.

방법	군집1	군집2	군집3
단일 연결	5위, 10위, 26위	5, 10, 26위를 제외한 모든 팀	X
군집 특성	군집 2에 너무 많은 팀들이 밀집하여 뚜렷한 특징을 나타내기 쉽지 않다.		
완전 연결	1위, 2위, 3위, 6위, 9위, 11위, 12위, 13위, 15위, 17위, 19위, 21위, 25위	23위, 27위, 28위, 29위, 30위	4위, 5위, 7위, 8위, 10위, 14위, 16위, 18위, 20위, 22위, 24위, 26위,
군집 특성	군집 2에는 명확히 최하위권 팀들이 모인 군집이다. 군집 1과 군집 3은 플레이오프 시리즈에 진출한 상위권 팀들과 플레이오프 시리즈 진출에 실패한 하위권 팀들이 고루 포진해있으므로 특징을 뚜렷이 구분하기 힘들다고 할 수 있다.		
평균 연결	23위, 27위, 28위, 29위, 30위	그 외 나머지	X
군집 특성	군집1을 봤을 때, 우선 순위가 대체적으로 최하위인 팀들이 모여 있으며, 3점슛 시도가 30개가 되지 않고 어시스트 개수 또한 24개 밑으로 낮은 수치를 보인다. 이를 통해 군집 1의 팀들은 현대 농구 트렌드인 3점슛을 받아들이고 그에 맞는 경기 전략들을 사용하지 않음을 알 수 있다. 따라서 3점슛 중심전술의 파생상품인 어시스트 개수 또한 낮은 것이다.(물론 팀 로스터 특성상 3점슛을 시도하기 곤란한 경우도 있다.) 하지만 23위의 멤피스 그리즐리스 팀이 최하위권 팀과 같이 묶이는 것은 의문일 수 있다. 그 이유는 멤피스 그리즐리스 팀이 공격 전술 중심이 아닌 진흙탕 승부가 생길 수밖에 없는 수비를 중심으로 하는 다운 템포 전술을 주로 사용하기 때문이다. 앞서 강조해왔던 업템포와 반대되는 다운 템포 전술로 인해 낮은 3점슛 시도와 어시스트 개수를 기록하는 것이다.		

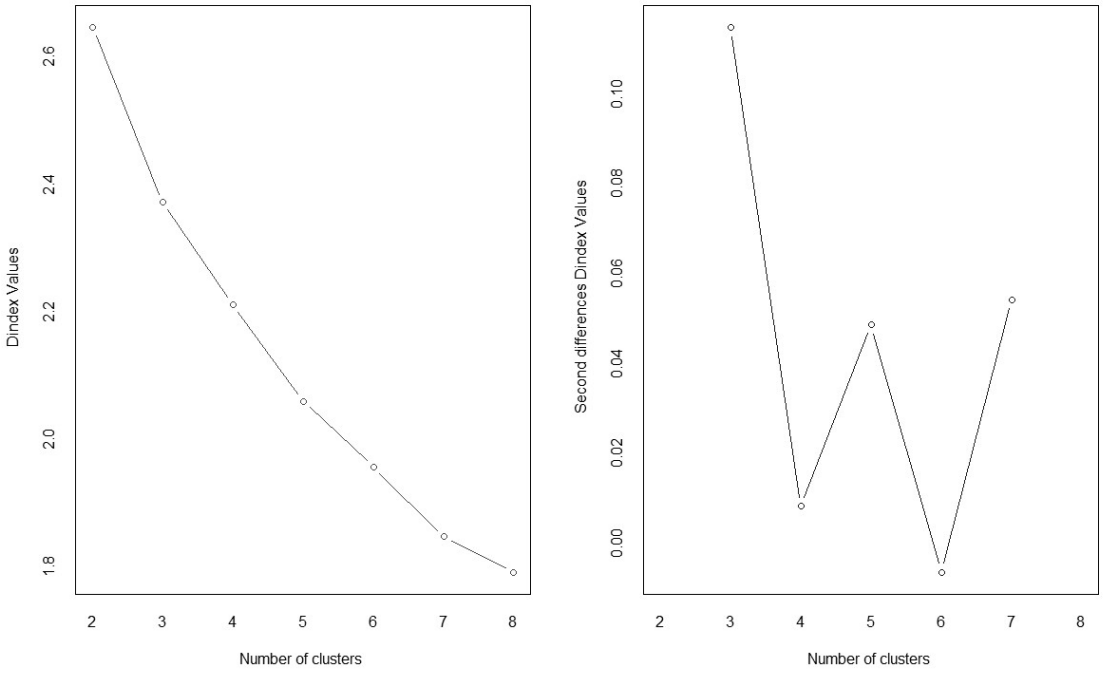
와드 연결	1위, 2위, 3위, 4위, 5위, 6위, 7위, 8위, 9위, 10위, 11위, 12위, 13위, 15위, 17위, 19위, 21위, 24위, 25위,	14위, 16위, 18위, 20위, 22위, 23위, 26위, 27위, 28위, 29위, 30위,	X
군집 특성	군집 1은 17, 19, 21, 24, 25위 팀을 제외하고 +/-에서 +마진을 남긴 팀이다. 반대로 군집 2에 모여 있는 팀들은 -마진을 기록하고 있다. 그 외에도 군집 1의 팀들은 몇 팀을 제외하고 대부분의 팀들이 상위권 팀들(W)이며 PTS와 AST 또한 몇 팀을 제외하면 평균을 넘고 있다. 그와는 반대로 군집 2의 팀들은 낮은 승수를 기록하고 있다.		

(2) 비위계적 군집 방법

① K-평균법

비위계적 군집분석은 변수들보다는 개체를 K개 군집으로 집단화하는데 이용된다. 군집 수 K는 미리 규정되거나 군집 절차의 한 부분으로 결정될 수 있다. 가장 가까운 중심점을 갖는 군집에 각 항목을 할당하는 알고리즘을 설명하기 때문에 K-평균법이라고 한다.

```
> allindex<-NbClust(Z, distance="euclidean", min.nc = 2, max.nc = 8,
+                   method = "kmeans", index = "all" )
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 6 proposed 2 as the best number of clusters
* 6 proposed 3 as the best number of clusters
* 3 proposed 4 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 4 proposed 7 as the best number of clusters
* 3 proposed 8 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```



K-평균법에서 가장 적절한 군집의 수는 두 개이며, 이를 바탕으로 군집을 나누고 각 군집의 특징을 파악해 보겠다.

```
> C1;C2
              rownames.nba. cluster
14.Brooklyn     14.Brooklyn       1
16.Detroit       16.Detroit       1
18.Miami           18.Miami       1
20.LAlakers     20.LAlakers       1
22.Dallas         22.Dallas       1
23.Memphis       23.Memphis       1
24.NewOrleans 24.NewOrleans       1
25.Washington 25.Washington       1
26.Atlanta       26.Atlanta       1
27.Chicago       27.Chicago       1
28.Cleveland   28.Cleveland       1
29.Phoenix       29.Phoenix       1
30.NewYork       30.NewYork       1
              rownames.nba. cluster
1.Milwaukee       1.Milwaukee       2
2.Toronto           2.Toronto       2
3.Golden             3.Golden       2
4.Denver             4.Denver       2
5.Houston           5.Houston       2
6.Portland         6.Portland       2
7.Philadelphia 7.Philadelphia       2
8.Utah                 8.Utah       2
9.Boston             9.Boston       2
10.Oklahoma       10.Oklahoma       2
11.Indiana         11.Indiana       2
12.LA                   12.LA       2
13.SanAntonio   13.SanAntonio       2
15.Orlando         15.Orlando       2
17.Charlotte     17.Charlotte       2
19.Sacramento   19.Sacramento       2
21.Minnesota     21.Minnesota       2
```

```
> aggregate(nba, by=list(kmeans$cluster),FUN=mean)
  Group.1        W      PTS      3PA      3P%      FT%     OREB      AST      TOV      STL       +/-
1       1 30.46154 108.7154 31.83846 34.43846 75.33846 10.32308 23.78462 14.66154 7.461538 -4.061538
2       2 49.05882 113.1059 32.13529 36.40000 77.73529 10.36471 25.18824 13.62941 7.747059  3.129412
```

k=5	Cluster1	Cluster2
	14, 16, 18, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 17, 19, 21
	군집 1과 군집 2를 비교했을 때 W, PTS, 3P%, FT%, AST, TOV, +/-에서 명확한 차이가 난다. 이는 군집 1이 하위권 팀이며, 군집 2가 상위권 팀인 것을 알 수 있게 해준다. 3PA, OREB, STL에선 군집 1이 군집 2에 비해 낮은 수치를 지니고 있지만 그리 명확한 차이가 아니므로 3PA, OREB, STL이 시즌 성적에 많은 영향을 끼친다고 볼 수 없다. 이를 통해 변수에 대한 추가적인 설명을 하자면 3PA 변수를 통해 이제는 모든 NBA 팀들이 3점슛의 중요성을 깨달아 이를 경기에 많이 반영하고 있는 것을 알 수 있게 해주며 OREB와 STL은 시즌 성적에 많은 영향을 끼치지 못한다고 결론을 내릴 수 있다.	

CA를 실행한 결과, 위계적 군집 방법에서는 군집 수가 2인 와드 연결법이 가장 군집의 특성이 잘 드러나며, 군집의 수도 적당한 방법이었고, 비위계적 방법에서는 K=2일 때의 K-평균법이 다양한 군집의 특성을 나타내며 군집의 수 또한 적당함을 알 수 있었다.

7. Conclusion

현대인의 욕구가 다양해지면서 사회는 그 욕구를 충족시키기 위해 점차 세부적이고 다양한 서비스를 제공하고 있다. 취미 또한 다르지 않다. 스포츠에서는 이제 자신이 직접 플레이를 하면서 자신이 좋아하는 선수들의 플레이를 실시간으로 감상할 수 있게 됐다. NBA를 실시간으로 시청할 수 있게 됨과 동시에 NBA에 대한 인기는 더욱더 증가하고 있음을 인터넷 검색창이나 기사로 확인할 수 있다. 이같이 인기가 늘어나는 NBA를 취미라고 남들에게 당당히 말하고 싶다면 이젠 NBA를 조금 더 전문적이고 심도 있게 시청하는 것이 필수불가결할 것이다. 그리고 이 또한 우리의 취미에 대한 욕구를 충족시켜주는 또 하나의 방법이라 생각한다.

이번 term project의 목표는 NBA team들의 시즌 성적에 영향을 미치는 요인을 분석함으로써 team들의 시즌 성적에 영향을 미치는 요인들을 알아보고 NBA를 즐겁게 시청하기 위한 시청 포인트를 제시하는 것이었다.

위의 분석결과들을 보면 주성분 분석의 결과 팀별 시즌 성적에 영향을 미치는 요인에 대해 알아봤으며 이들 간 어떤 연관이 있는지, 주성분에 따라 나누어진 팀들의 군집이 어떤 변수에 더 많은 영향을 받는지에 대해 알아보았다. 제 1주성분에서 턴오버와 턴오버를 제외한 지표들 간의 반비례함을 알 수 있었으며제 2주성분에서 (평균 득점, 3점슛 시도 횟수, 공격 리바운드 개수, 어시스트 개수, 턴오버 개수, 스틸 개수)과 (3점슛 성공률, 자유투 성공률)의 반비례함을 알 수 있었다. 그리고 제 2주성분을 승부처 요인으로 언급하기도 했다.
제 3주성분에서 또한 마찬가지로 (3점슛 시도, 공격리바운드)와 (어시스트, 턴오버, 스틸의 개수) 간의 반비례함을 알 수 있었으며 3점슛 시도의 증가가 공격리바운드의 증가로 이어짐을 알 수 있었다.

다음으로 인자 분석을 시행하기에 앞서 PCFA, 회전 전 MLFA, 회전 후 MLFA 중 어떤 것이 가장 분석을 하는 데에 용이한 지에 대해 알아보았다. 분석 결과, 회전을 한 MLFA가 인자 분석에 가장 용이했으며 이를 기반으로 인자 적재값 분석, 인자 점수그림을 그려봤다. 인자 분석 결과 선택한 인자를 기반으로 팀들이 어떻게 분포하는지에 대해 살펴보았다.
factor1을 통해 AST, PTS, +/-, W, 3P%는 서로 비례함으로 시즌 성적에 경기당 평균 득점과 3점슛 성공률, 어시스트, 득실점 마진이 영향을 끼침을 알 수 있었으며 factor2에서는 인자 적재 값과 화살표를 통해 3P%, FT%와 STL, OREB, 3PA가 반비례함을 알 수 있었다. 또한 factor3를 통해 TOV가 W, FT%, +/-와 반비례함을 알 수 있었다.

마지막으로 군집분석 결과, 위계적 군집방법에서는 군집수가 2인 와드 연결법이 가장 효과적으로 군집을 나눠줬고, 비위계적 방법에서는 k=2인 K-평균법을 통해 효과적으로 군집을 나눴는데 이는 W(승수)와 PTS, +/-에 따라 군집이 나누어진 것을 확인할 수 있었다.

결론적으로 턴오버를 제외한 모든 지표들이 좋으면 좋을수록 시즌 성적에 긍정적인 영향을 끼치는 것을 알 수 있었다. 이들 중에는 평균 득점과 평균 득실점 마진, 3점슛 성공률과 어시스트의 개수가 직접적인 영향을 끼쳤으며 스틸과 공격리바운드는 그리 많은 영향을 끼치지 못한 것을 알 수 있었다.

위의 결과들을 참고하여 우리가 봐왔던 NBA 경기들과 앞으로 있을 경기들을 조금이나마 더 즐겁게 시청할 수 있기를 희망한다.

Reference

● 자료 참조 : NBA 공식 홈페이지
https://stats.nba.com/teams/traditional/?sort=GP&dir=-1&Season=2018-19&SeasonType=Regular%20Season
● 기록 참조 : NBA.com, basketball-reference, ESPN.com, Elias Sports Bureau, spotrac.com
● 해석 참조 : 최용석.『R과 함께하는 다변량 자료분석 - 경문사』
● 코드 참조 : (이 후 페이지 참조)
● 칼럼 참조 : 열혈기자 염용근의 오늘의 NBA
https://sports.news.naver.com/column/columnList.nhn?expertId=724&page=4

----------감사합니다----------

R-code)

```
full<-read.table("nba full data.txt")
full1<-as.matrix(full,ncol=27,byrow=T)
fullnba<-matrix(full1,ncol=27,byrow=T)
colnames(fullnba)<-c(fullnba[1,])
rownames(fullnba)<-c(fullnba[,1])
fullnba<-fullnba[-1,]
fullnba<-fullnba[,-1]

nba<-fullnba[,cbind(2,6,11,12,15,16,19,20,21,26)]
nba<-as.numeric(nba)
nba<-matrix(nba,30,10)
colnames(nba)<-c(fullnba[1,c(3,7,12,13,16,17,20,21,22,
27)])
rownames(nba)<-rownames(fullnba)
dim(nba)
dim(fullnba)
View(nba)

###  다변량 정규성 검정
install.packages("MVN")
library(MVN)
result <- mvn(nba,
mvnTest="mardia",multivariatePlot="qq")
result

n=dim(nba)[1]
p=dim(nba)[2]
S=cov(nba)
xbar=colMeans(nba)
m=mahalanobis(nba, xbar, S)
m=sort(m)
id=seq(1, n)
pt=(id-0.5)/n
q=qchisq(pt, p)
plot(q, m, pch="*", xlab="Quantile", ylab="Ordered
Squared Distance")
abline(0, 1)

# Correlation Coefficient Test for Normailty
rq=cor(cbind(q, m))[1,2]
rq


### PCA 수행

R=round(cor(nba),3)
eigen=eigen(R)
round(eigen$values, 3) # Eigenvalues
V=round(eigen$vectors, 2) # Eigenvectors
gof=eigen$values/sum(eigen$values)*100 #
Goodness-of fit
round(gof, 3)
plot(eigen$values, type="b", main="Scree Graph",
xlab="Component Number", ylab="Eigenvalue")
#요인을 3개로 지정해야함을 알 수 있다.

# Plot of PCs Scores
V3=V[,1:3]
V3

Z=scale(nba, scale=T)
P=Z%*%V3
round(P, 3)

plot(P[,1], P[, 2], main="(a)Plot of PCs Scores",
xlab="1st PC", ylab="2nd PC")
text(P[,1], P[, 2], labels=rownames(P), cex=0.8,
col="blue", pos=3)
abline(v=0, h=0)

plot(P[,1], P[, 3], main="(b)Plot of PCs Scores",
xlab="1st PC", ylab="3rd PC")
text(P[,1], P[, 3], labels=rownames(P), cex=0.8,
col="blue", pos=3)
abline(v=0, h=0)

plot(P[,2], P[, 3], main="(c)Plot of PCs Scores",
xlab="2nd PC", ylab="3rd PC")
text(P[,2], P[, 3], labels=rownames(P), cex=0.8,
col="blue", pos=3)
abline(v=0, h=0)

## PC biplot
## 1 & 2
pcasvd.Z <- prcomp(nba, scale=T)
summary(pcasvd.Z)
round(pcasvd.Z$rotation,3) ## eigenvector의 값이
나옴
biplot(pcasvd.Z, scale=0 , xlab="1st PC(34.52%)" ,
ylab="2nd PC(21.72%)"
```

```
        , main="PC Biplot for NBA ", choices=c(1,2))
abline(v=0,h=0)
## 1 & 3
biplot(pcasvd.Z, scale=0 , xlab="1st PC(34.52%)" ,
ylab="3rd PC(11.90%)"
        , main="PC Biplot for NBA", choices=c(1,3))
abline(v=0,h=0)
## 2 & 3
biplot(pcasvd.Z, scale=0 , xlab="2nd PC(21.72%)" ,
ylab="3rd PC(11.90%)"
        , main="PC Biplot for NBA", choices=c(2,3))
abline(v=0,h=0)


### FA 수행
## 인자 갯수 파악
p <- ncol(nba)
R <- round(cor(nba),3)
eigen2<-eigen(R)
eigen2$values
round(eigen2$values ,3)
V<-round(eigen2$vectors, 2)
gof <- eigen2$values/p*100
round(gof,3)
plot(eigen2$values, type="b", main="Scree Graph",
xlab="Factor Number", ylab="Eigenvalue")

## 패키지를 통한 인자 갯수 파악
install.packages("psych")
library(psych)
pcfa <- principal(Z, nfactors=3, rotate="none")
gof <- pcfa$values/p*100 # Goodness-of fit
round(gof, 3)

### 인자 값 비교
pcfa <- principal(Z, nfactors=3, rotate="none")
L <- pcfa$loading[,1:3]
round(L,3)

mlfa <-factanal(covmat=R , factors = 3,
rotation="none")
mlfa.r <- factanal(Z, factors = 3,
rotation="varimax", score="regression")
Lm <- mlfa$loading[, 1:3]
Lm.r <- mlfa.r$loading[, 1:3]
round(Lm, 3)
round(Lm.r, 3)


### MLFA 인자 적재그림
## 1 & 2
win.graph()
lim<-range(pretty(Lm))
plot(Lm[,1], Lm[,2],main="(a) ML Factor Loadings :
f1 and f2",  xlab="f1", ylab="f2",
      xlim=lim, ylim=lim)
text(Lm[,1], Lm[, 2], labels=rownames(Lm), cex=0.8,
col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, Lm[,1], Lm[, 2], col=2, code=2,
length=0.1)


## 1 & 3

plot(Lm[,1], Lm[,3],main="(b) ML Factor Loadings :
f1 and f3",  xlab="f1", ylab="f3",
      xlim=lim, ylim=lim)
text(Lm[,1], Lm[, 3], labels=rownames(L), cex=0.8,
col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, Lm[,1], Lm[, 3], col=2, code=2,
length=0.1)

## 2 & 3

plot(Lm[,2], Lm[,3],main="(c) ML Factor Loadings :
f2 and f3",  xlab="f2", ylab="f3",
      xlim=lim, ylim=lim)
text(Lm[,2], Lm[, 3], labels=rownames(L), cex=0.8,
col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, Lm[,2], Lm[, 3], col=2, code=2,
length=0.1)

### 인자 점수 그림
mlfa<- factanal(Z, factors = 3, rotation="varimax",
score="regression")
fml <- mlfa$scores
round(fml, 3)
rownames(fml) <- rownames(nba)

## 1 & 2
lim<-range(pretty(fml))
par(pty="s")
plot(fml[,1], fml[,2],main=" (a) Factor Scores : f1
and f2",  xlab="f1", ylab="f2",
      xlim=lim, ylim=lim)
text(fml[,1], fml[,2], labels=rownames(fml), cex=0.5,
col="blue", pos=1)
abline(v=0, h=0)
```

```
## 1 & 3
par(pty="s")
plot(fml[,1], fml[,3],main=" (b) Factor Scores : f1
and f3",  xlab="f1", ylab="f3",
     xlim=lim, ylim=lim)
text(fml[,1], fml[,3], labels=rownames(fml), cex=0.5,
col="blue", pos=1)
abline(v=0, h=0)

## 2 & 3
par(pty="s")
plot(fml[,2], fml[,3],main=" (c) Factor Scores : f2
and f3",  xlab="f2", ylab="f3",
     xlim=lim, ylim=lim)
text(fml[,2], fml[,3], labels=rownames(fml), cex=0.5,
col="blue", pos=1)
abline(v=0, h=0)


### CA
Z <- scale(nba,scale=T)
ds <- dist(Z, method="euclidean")

## 단일연결법
install.packages("NbClust")
library(NbClust)
par(mfrow=c(1,1))
ccc <- NbClust(Z, distance="euclidean", min.nc = 2,
max.nc = 8,
                method = "single", index = "ccc")
plot(2:8, type="b", ccc$All.index, xlab="Number of
Clusters",
     ylab="CCC")
dindex <- NbClust(Z, distance= "euclidean",
min.nc=2 , max.nc=8,
                    method = "single" ,
index="dindex")
allindex <- NbClust(Z, distance= "euclidean",
min.nc=2 , max.nc=8,
                     method = "single" , index="all")
single <- hclust(ds, method="single")
plot(single, hang=-1, main="(a) Sinle Linkage")


## 완전연결법
complete <- hclust(ds, method="complete")
plot(complete,hang=-1, main="(b) Complete
Linkage")
ccc <- NbClust(Z, distance="euclidean", min.nc = 2,
max.nc = 8,
                method = "complete", index = "ccc")
plot(2:8, type="b", ccc$All.index, xlab="Number of
Clusters",
     ylab="CCC")
allindex <- NbClust(Z, distance= "euclidean",
min.nc=2 , max.nc=8,
                     method = "complete" ,
index="all")
dindex <- NbClust(Z, distance= "euclidean",
min.nc=2 , max.nc=8,
                   method = "single" ,
index="dindex")

#평균연결법
ccc <- NbClust(Z, distance="euclidean", min.nc = 2,
max.nc = 8,
               method = "average", index = "ccc")
plot(2:8, type="b", ccc$All.index, xlab="Number of
Clusters", ylab="CCC")

allindex <- NbClust(Z, distance= "euclidean",
min.nc=2 , max.nc=8,
                     method = "average" ,
index="all")
dindex <- NbClust(Z, distance= "euclidean",
min.nc=2 , max.nc=8,
                   method = "single" ,
index="dindex")
average=hclust(ds, method="average")
plot(average, hang=-1, main="(c) Average Linkage")

##와드연결법
ccc <- NbClust(Z, distance="euclidean", min.nc = 2,
max.nc = 8,
               method = "ward.D", index = "ccc")
plot(2:8, type="b", ccc$All.index, xlab="Number of
Clusters",
     ylab="CCC")
allindex <- NbClust(Z, distance= "euclidean",
min.nc=2 , max.nc=8,
                     method = "ward.D" ,
index="all")
dindex <- NbClust(Z, distance= "euclidean",
min.nc=2 , max.nc=8, method = "ward.D" ,
index="dindex")
ward=hclust(ds, method="ward")
plot(ward, hang=-1, main="(d) Ward Linkage")
colMeans(nba)

## k 평균법
kmeans <- kmeans(Z, 2)
cluster <- data.frame(rownames(nba) ,
```

```
cluster=kmeans$cluster)
C1 <- cluster[(cluster[,2]==1),]
C2 <- cluster[(cluster[,2]==2),]
C1;C2
ccc<-NbClust(Z, distance="euclidean", min.nc = 2, max.nc = 8,
                method = "kmeans", index = "ccc")
plot(2:8, type="b", ccc$All.index, xlab="Number of Clusters",
      ylab="CCC")
dindex<-NbClust(Z, distance="euclidean", min.nc = 2, max.nc = 8,
                    method = "kmeans", index = "dindex")
allindex<-NbClust(Z, distance="euclidean", min.nc = 2, max.nc = 8,
                      method = "kmeans", index = "all")
aggregate(nba, by=list(kmeans$cluster),FUN=mean)
```