REPORT

수강과목 :	다변량통계학(I)
담당교수 :	최용석 교수님
학 과 :	통계학과
학 번 :	201611523
이 름 :	안성빈
제출일자 :	2020.07.03

1. 소개

모든 구기종목에서처럼 농구에서도 포지션이 존재한다. 크게는 가드(G), 포워드(F), 센터(C)로 나뉘는데 이러한 포지션에는 각각의 특성이 있다. (예를 들어, 리바운드, 3점슛, 리딩 등) 하지만 축구, 야구와는 달리 선수 전원이 빠르게 공수전환을 하기에 경기 중 포지션의 역할이 희미해지는 경우가 많다. 또한 프로리그에서는 포지션에 가장 큰 영향을 주는 신장조차 큰 차이가 없다. 그럼에도 불구하고 포지션을 확실하게 구분하는 이유는 무엇이며, 과연 프로리그에서 선수들은 본인의 포지션에 맞는 경기를 하고 있는지 궁금해진다. 그렇기에 데이터(**"kbl 선수 득점순위 20"**)에 대한 다변량 자료분석을 통하여 이를 알아보고자한다.

2. 데이터

데이터는 "2019-2020 kbl 선수 득점순위 20"을 사용한다.(출처,네이버)
데이터는 센터 9명(캐디 라렌, 자밀 워니, 라건아, 브랜든 브라운, 리온 윌리엄스, 치나누 오노아쿠, 김종규, 머피 할로웨이, 김준일), 가드 5명(허훈, 이정현, 허웅, 김선형, 김낙현), 포워드 6명(닉 미네라스, 트로이 길렌워터, 송교창, 칼렙 그린, 양홍석, 최준용)으로 이루어져있다.
※이후 분석에서는 포지션을 간단히 C, G, F 로 나타낸다.

순위	선수	경기수	득점	AS	리바운드	스틸	블록	2점슛	3점슛	자유투	야투성공	3점슛성공	자유투성공
1	캐디 라렌(LG)	42	**21.40**	1.21	10.88	0.60	1.31	6.55	1.24	4.60	49.62	41.60	72.01
2	닉 미네라스(삼성)	43	**20.95**	0.77	5.86	0.88	0.19	6.00	1.56	4.28	48.58	27.92	83.26
3	자밀 워니(SK)	43	**20.42**	3.09	10.44	1.09	0.67	9.05	0.00	2.33	53.43	0.00	70.42
4	라건아(KCC)	41	**20.24**	1.98	12.49	0.66	0.98	8.51	0.05	3.07	56.34	20.00	70.79
5	브랜든 브라운(KGC)	42	**18.43**	2.76	8.88	1.57	0.24	6.10	0.93	3.45	50.86	29.32	63.88
6	트로이 길렌워터(전자랜드)	24	**16.63**	0.88	4.83	0.63	0.38	4.67	1.08	4.04	50.74	33.33	84.35
7	송교창(KCC)	42	**15.05**	3.19	5.64	1.00	0.60	4.43	1.36	2.12	45.94	38.00	67.94
8	허훈(KT)	35	**14.94**	7.23	2.63	1.20	0.06	3.49	2.00	1.97	42.38	35.18	77.53
9	리온 윌리엄스(현대모비스)	42	**14.64**	1.24	9.29	1.12	0.40	5.33	0.24	3.26	53.42	27.03	81.07
10	치나누 오누아쿠(DB)	40	**14.38**	2.45	10.30	1.35	1.53	5.65	0.20	2.48	53.79	27.59	66.89
11	칼렙 그란(DB)	42	**13.88**	2.38	5.76	0.79	0.24	3.26	1.40	3.14	47.46	36.42	80.49
12	이정현(KCC)	42	**13.74**	4.55	2.79	1.12	0.10	2.50	1.95	2.88	38.40	31.66	82.88
13	허웅(DB)	29	**13.69**	1.38	2.45	0.97	0.00	2.55	2.34	1.55	44.38	37.99	84.91
14	김종규(DB)	43	**13.28**	1.98	6.07	0.37	0.84	4.65	0.44	2.65	51.41	29.69	77.03
15	머피 할로웨이(전자랜드)	42	**13.12**	2.14	9.40	1.29	0.62	5.55	0.05	1.88	53.29	40.00	53.02
16	김선형(SK)	37	**12.59**	3.68	2.76	1.78	0.43	3.70	0.95	2.35	45.26	33.65	73.73
17	김낙현(전자랜드)	40	**12.18**	3.35	2.45	1.05	0.08	2.18	2.20	1.23	44.08	36.97	83.05
18	양홍석(KT)	43	**12.14**	1.79	5.74	0.47	0.53	3.35	1.14	2.02	43.57	28.32	77.68
19	최준용(SK)	38	**11.84**	3.37	5.97	0.89	0.79	2.08	1.95	1.84	39.84	35.41	67.31
20	김준일(삼성)	33	**11.82**	1.85	5.06	1.00	0.45	4.94	0.00	1.94	54.33	0.00	75.29

3. 변수소개

- **득점: 한 경기당 선수가 넣은 평균득점을 나타낸다.**
- **AS: 한 경기당 선수가 기록한 어시스트(골 도움)의 평균을 나타낸다.**
- **리바운드: 한 경기당 선수가 기록한 리바운드의 평균을 나타낸다**(주로 센터(C)의 기록이 높다)
- **스틸: 한 경기당 선수가 기록한 스틸의 평균을 나타낸다.** (주로 가드(G)의 기록이 높다)
- **블록: 한 경기당 선수가 기록한 블록의 평균을 나타낸다.** (주로 센터(C)의 기록이 높다)
- **2점슛: 한 경기당 선수가 성공시킨 2점슛의 평균을 나타낸다.**
- **3점슛: 한 경기당 선수가 성공시킨 3점슛의 평균을 나타낸다.**
- **자유투: 한 경기당 선수가 성공시킨 자유투(1점)의 평균을 나타낸다.**
- **야튜성공률: 한 경기당 선수가 성공시킨 2점슛의 확률의 평균을 나타낸다.**
- **3점슛성공률: 한 경기당 선수가 성공시킨 3점슛의 확률의 평균을 나타낸다.**
- **자유투성공률: 한 경기당 선수가 성공시킨 자유투(1점)의 확률의 평균을 나타낸다.**

(각 기록이 경기별 평균으로 제시되어 있으므로 경기수 변수는 사용하지 않는다.)

4. 다변량 자료분석

다중산점도, box-plot, 줄기-잎 그림 등 다양한 방법으로 다변량 자료에 대한 분석을 할 수 있다. 하지만 여기서는 선수별 특성을 가장 잘 보여주는 star그림과 변수끼리의 관계성을 잘 나타내는 상관계수 행렬을 통하여(변수별 단위가 다르기 때문에) 다변량 자료분석을 하고자 한다.

4-1. 상관계수 행렬

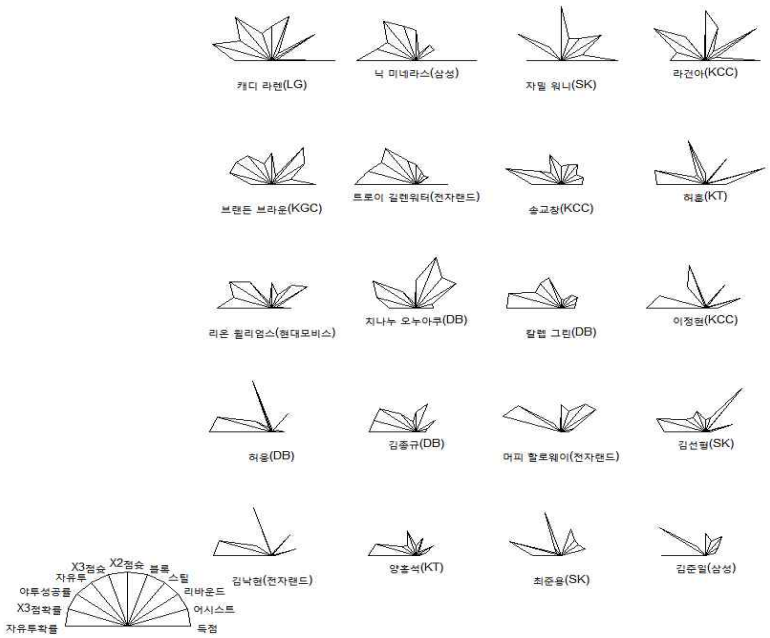
	득점	어시스트	리바운드	스틸	블록	X2점슛	X3점슛	자유투	야투성공률	X3점확률	자유투확률
득점	1.000	-0.254	0.587	-0.135	0.242	0.782	-0.191	0.719	0.413	-0.183	-0.061
어시스트	-0.254	1.000	-0.394	0.426	-0.268	-0.318	0.345	-0.443	-0.530	0.096	-0.090
리바운드	0.587	-0.394	1.000	-0.104	0.722	0.834	-0.715	0.402	0.741	-0.281	-0.576
스틸	-0.135	0.426	-0.104	1.000	-0.200	-0.026	-0.047	-0.237	-0.042	-0.005	-0.352
블록	0.242	-0.268	0.722	-0.200	1.000	0.489	-0.546	0.206	0.472	-0.100	-0.519
X2점슛	0.782	-0.318	0.834	-0.026	0.489	1.000	-0.739	0.451	0.810	-0.537	-0.391
X3점슛	-0.191	0.345	-0.715	-0.047	-0.546	-0.739	1.000	-0.144	-0.877	0.599	0.493
자유투	0.719	-0.443	0.402	-0.237	0.206	0.451	-0.144	1.000	0.307	0.068	0.154
야투성공률	0.413	-0.530	0.741	-0.042	0.472	0.810	-0.877	0.307	1.000	-0.496	-0.341
X3점확률	-0.183	0.096	-0.281	-0.005	-0.100	-0.537	0.599	0.068	-0.496	1.000	0.014
자유투확률	-0.061	-0.090	-0.576	-0.352	-0.519	-0.391	0.493	0.154	-0.341	0.014	1.000

> |

상관계수 행렬을 통하여 데이터를 분석해보니
1)득점은 2점슛과 자유투와 높은 연관성을 보인다.
2)리바운드는 블록, 2점슛, 야투 성공률과 높은 연관성을 보인다.
3)리바운드는 3점슛과는 음의 방향으로 높은 연관성을 보인다.
4)3점슛과 2점슛은 둘 다 야투성공률과 높은 연관성을 보인다.
5)2점슛은 2점슛과 음의 방향으로 높은 연광성을 보인다.

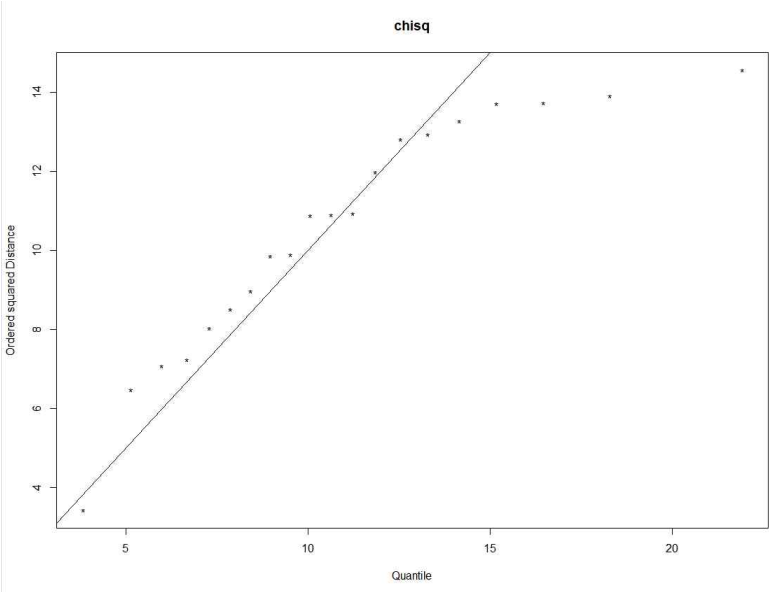
의외로 3점슛이 2점슛, 리바운드와 음의 관계를 보인다. 주로 센터(C)들이 리바운드와 근거리에서 슛을 하는 것을 고려한다면 충분히 가능한 결과이다.

4-2. STAR그림



같은 포지션임에도 다른 분포를 보이는 선수들이 있다.(캐디라렌과 자밀워니) 하지만 전체적으로 센터(C)들이 가드(G)나 포워드(F)보다는 왼쪽에서 큰 면적을 보인다. 또한, 득점의 순위가 높을수록 전체적인 면적이 큰 것을 알 수 있다.

4-3. 정규성 검정




```
> rq
[1] 0.9315532
```

※카이제곱 그래프와 rq를 통하여 데이터는 정규성을 따른다고 할 수 있다.

5. PCA(주성분분석)

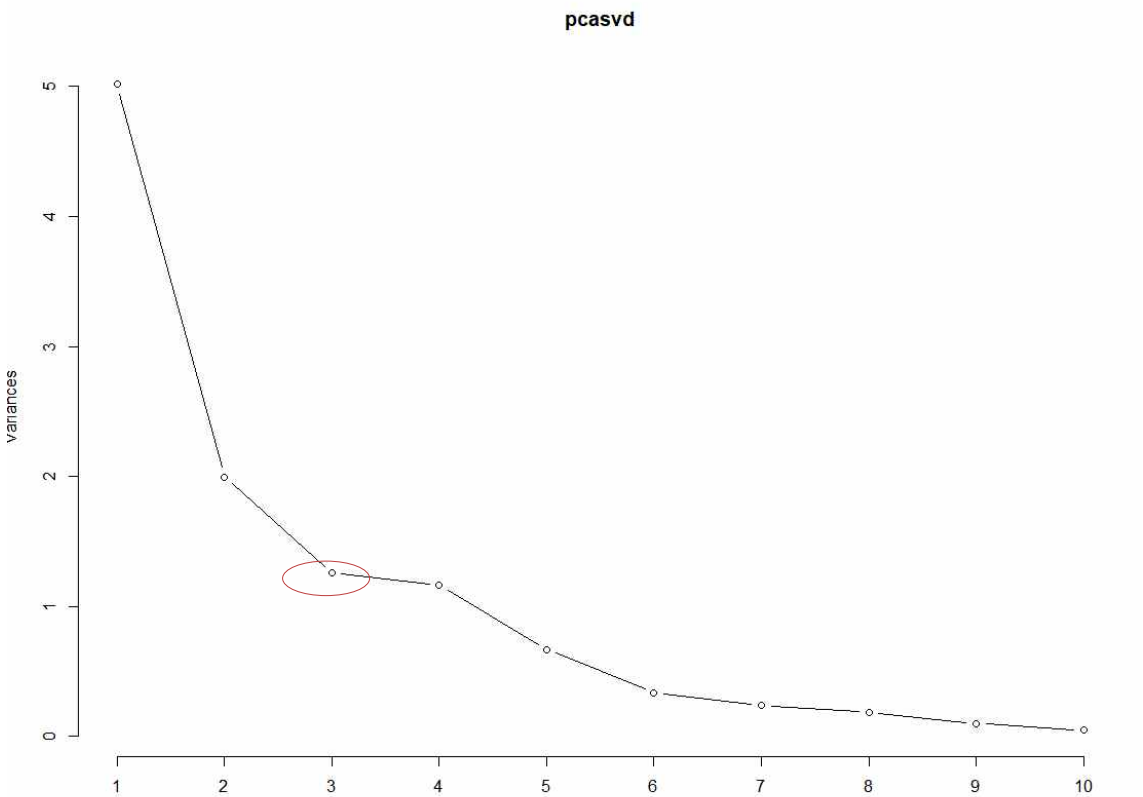
위의 다변량 자료분석을 통하여 우리는 변수끼리의 연관성이나 선수(개체)의 특성을 파악할 수 있다. 하지만 변수가 너무 많아 시각적으로 구성하기가 힘들어 차원축소를 해주고자 한다. 차원축소의 첫 방법으로 pca(주성분분석)를 사용한다. PCA는 상호 독립성을 보장할 수 없는 원 변수들의 수학적 선형결합을 통해 주성분(principal component)이라는 서로 독립인 새로운 변수들을 만들어 분석하는 기법이다.

5-1. 주성분의 선택

특이값 분해를 이용한 pca 함수인 prcomp를 사용하였다.

```
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     2.2398 1.4130 1.1204 1.0788 0.81788 0.57873 0.48757 0.42781 0.31222 0.21340
Proportion of Variance 0.4561 0.1815 0.1141 0.1058 0.06081 0.03045 0.02161 0.01664 0.00886 0.00414
Cumulative Proportion  0.4561 0.6376 0.7517 0.8575 0.91830 0.94875 0.97036 0.98700 0.99586 1.00000
                          PC11
Standard deviation     0.001372
Proportion of Variance 0.000000
Cumulative Proportion  1.000000
```

•pc3에서 누적 설명력이 0.7517이 되므로 3개의 pc를 선택한다. Screeplot을 통하여 한번 더 살펴보자.



pc3이 팔꿈치 부분이므로 pc는 3개를 선택한다.

5-2. 주성분의 성질

	PC1	PC2	PC3
득점	-0.285	-0.314	0.235
어시스트	0.226	0.355	0.178
리바운드	-0.411	0.050	0.207
스틸	0.056	0.462	0.163
블록	-0.300	0.102	0.287
X2점슛	-0.417	0.004	-0.017
X3점슛	0.374	-0.246	0.251
자유투	-0.215	-0.482	0.254
야투성공률	-0.397	0.059	-0.219
X3점확률	0.212	-0.154	0.651
자유투확률	0.205	-0.479	-0.402

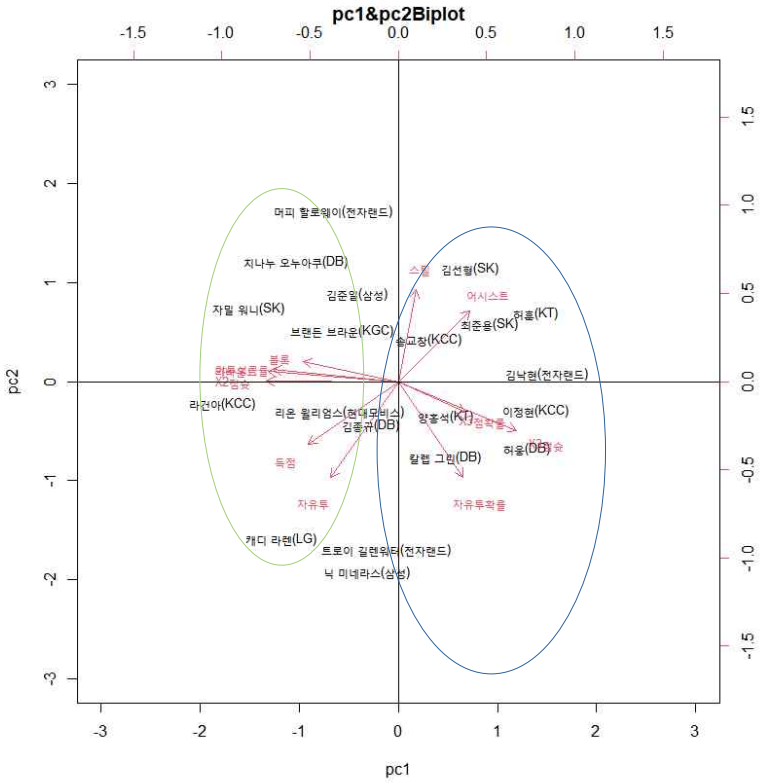
PC1: 리바운드, 2점슛, 야투성공률과는 음의 상관관계를 보이며, 3점슛과는 양의 상관관계를 보인다.

PC2: 자유투, 자유투 성공률과 음의 상관 관계를 보이며, 스틸과 양의 상관관계를 보이므로 스틸과 자유투를 나타내는 변수로 해석할 수 있다.

PC3: 3점슛 확률과 높은 양의 상관관계를 보이므로 3점슛 확률을 나타내는 변수로 볼 수 있다.

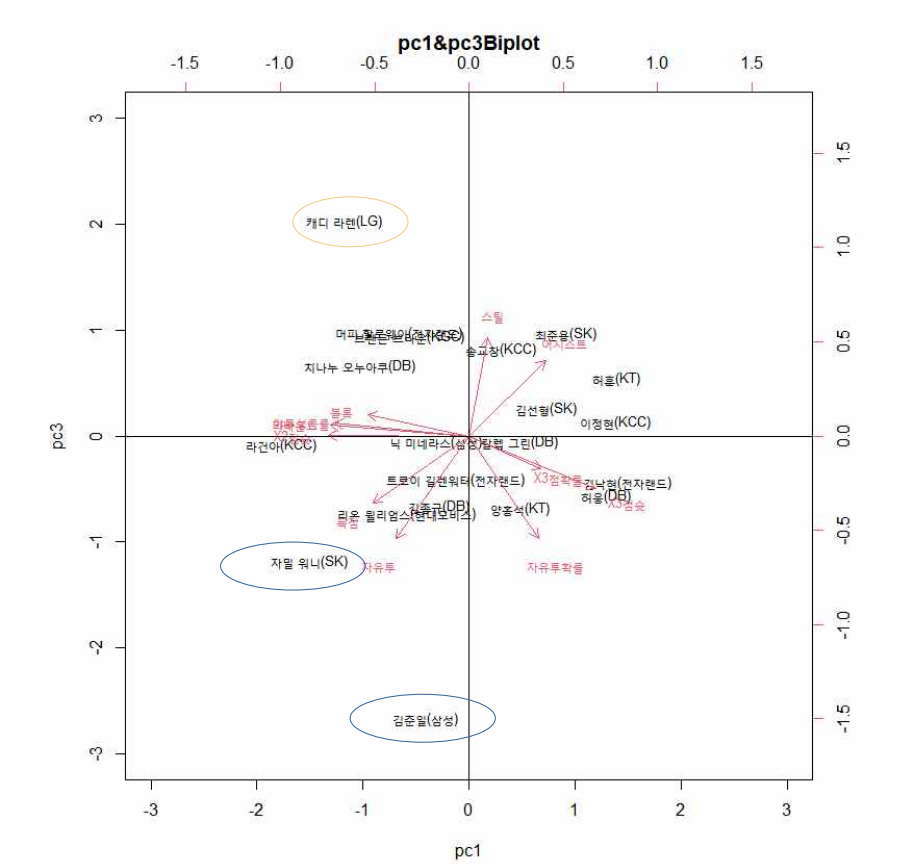
5-3. Biplot을 통하여 살펴보는 주성분의 성질과 선수들의 분포

1)pc1&pc2



우선, pc1을 기준으로 왼쪽은 2점슛, 블록, 리바운드 등 비교적 센터의 특성을 나타내는 변수가 큰 연관성을 가짐을 알 수 있다. 또한 pc2는 스틸과 높은 연관성을 보인다. 결과적으로 y축(pc2)를 기준으로 오른쪽에는 G,F의 선수들이 분포하고 왼쪽에는 C의 선수들이 분포함을 알 수 있다.

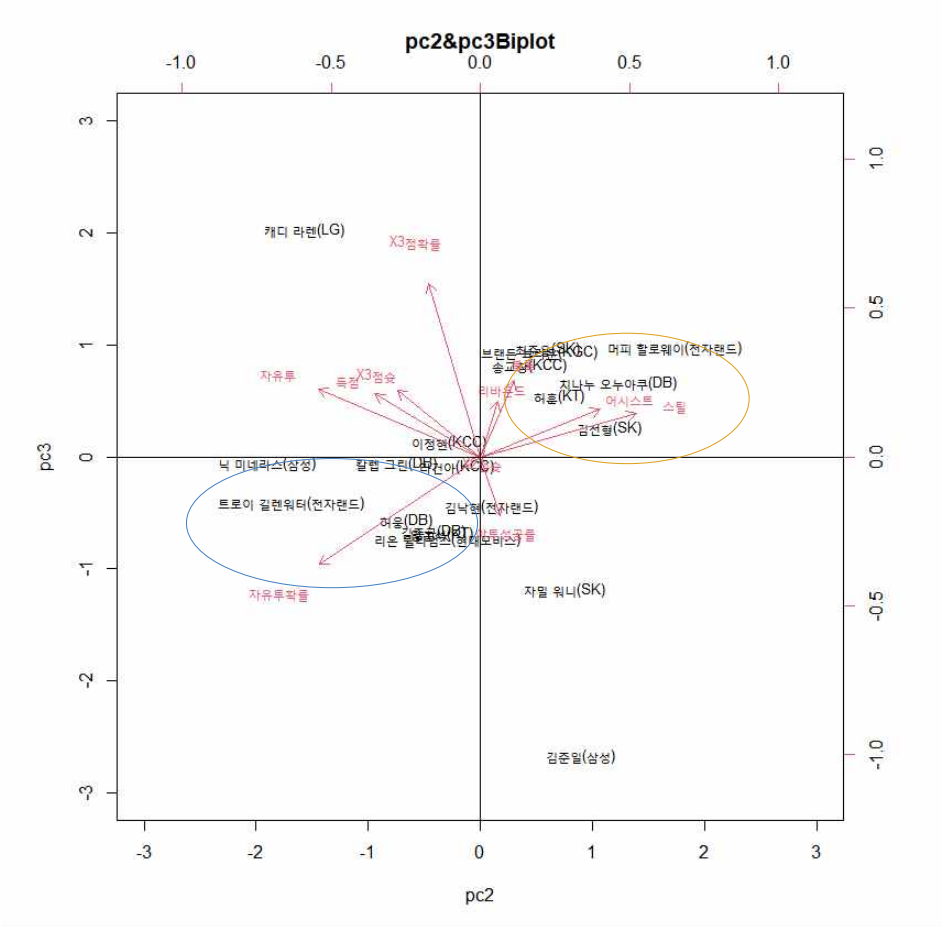
2)pc1&pc3



1) 이번 biplot역시 1변수에 의해 오른쪽에는 G와F, 왼쪽에는 C위주로 분포해 있다.

2) 캐디라렌은 센터 중 눈에 띄게 3점의 확률이 높은 선수이다. 그렇기에 다른 선수들과는 C들과는 다르게 PC3(3점 변수)에서 높은 지점에 존재함을 알 수 있다.
3) 반면 김준일과 자밀워니는 센터 중 3점슛이 하나도 없는 선수들로 PC3에서 낮은 지점에서 분포한다.

3)pc2&pc3



앞의 PC1 변수를 포함했던 Biplot들은 PC1에 의하여 오른쪽 왼쪽이 포지션 별로 나뉘었지만 PC2&PC3 Biplot은 3점슛이나 자유투등의 득점 방법과 플레이 스타일에 의하여 선수들의 분포가 이루어져있음을 볼 수 있다.

1)캐디 라렌, 김준일, 자밀 워니는 위의 PC1&PC3에서와 같이 3점슛에 의하여 y축(PC3)의 위치가 정해졌다.

2) 1사분면에 분포한 선수들은 자유투나 2점슛등 득점과 연관성이 높은 변수보다는 어시스트, 스틸, 리바운드에 더 치중한 선수들이므로 각자의 포지션에서 비교적 낮은 득점 순위를 보인다.

3) 3사분면에 분포한 선수들은 자유투나 2점슛 등의 득점과 연관성이 높은 변수에 치중되어 있으므로 비교적 높은 득점 순위를 보인다.

6. PCFA(주성분인자분석)

차원축소를 하는 또 다른 방법으로 FA(인자분석)가 있다. 변수들의 선형결합에 의한 주성분을 제공하는 주성분계수를 찾아내는 PCA와는 반대로 FA는 인자(factor)라는 잠재변수들의 선형결합으로 변수들을 나태내는 모형의 인자적재(factor loading)를 찾아내려는 것이다. 그리고 FA는 인자모형의 추정에 있어 PCA를 활용하는 PCFA와 MLE(최대우도법)을 활용하는 MLFA로 나뉜다. 그 중 PCFA를 먼저 이용하여 분석하여 본다.

6-1.인자선택

•인자선택은 앞서 시행했던 PCA에 그거를 두고 3개를 택하며 이번엔 함수를 통하여 설명력(gof)를 살펴본다.

```
> gof=pcfa$values/sum(pcfa$values)*100
> gof<-round(gof,3)
> gof
 [1] 45.606 18.152 11.412 10.579  6.081  3.045  2.161  1.664  0.886  0.414  0.000
>
```

•또한 잔차행렬의 비대각 원소들이 0에 가까울수록 좋으므로 잔차행렬도 한번 살펴본다.

```
> L=pcfa$loadings[,1:3]
> Psi=diag(pcfa$uniquenesses)
> Rm=R-(L%*%t(L)+Psi)
> round(Rm,2)
          득점 어시스트 리바운드  스틸  블록 X2점슛 X3점슛 자유투 야투성공률 X3점확률 자유투확률
득점      0.00     0.24    -0.03  0.19 -0.21   0.19   0.12   0.03     -0.05    -0.17       0.05
어시스트  0.24     0.00    -0.01  0.00 -0.06   0.16   0.04   0.09     -0.07    -0.18       0.11
리바운드 -0.03    -0.01     0.00 -0.08  0.02  -0.02   0.02  -0.06     -0.03     0.00       0.00
스틸      0.19     0.00    -0.08  0.00 -0.27   0.09   0.02   0.22      0.06    -0.06       0.11
블록     -0.21    -0.06     0.02 -0.27  0.00  -0.13  -0.02  -0.11     -0.06     0.02       0.03
X2점슛    0.19     0.16    -0.02  0.09 -0.13   0.00   0.05   0.01     -0.03    -0.08       0.03
X3점슛    0.12     0.04     0.02  0.02 -0.02   0.05   0.00  -0.06     -0.03    -0.08       0.00
자유투    0.03     0.09    -0.06  0.22 -0.11   0.01  -0.06   0.00      0.01    -0.06       0.04
야투성공률 -0.05   -0.07    -0.03  0.06 -0.06  -0.03  -0.03   0.01      0.00     0.12       0.01
X3점확률  -0.17    -0.18     0.00 -0.06  0.02  -0.08  -0.08  -0.06      0.12     0.00      -0.02
자유투확률 0.05     0.11     0.00  0.11  0.03   0.03   0.00   0.04      0.01    -0.02       0.00
```

Rm의 비대각 원소들도 0에 가까우므로 3개의 인자는 충분하다고 볼 수 있다.

6-2.인자의 성질

FA는 변수들의 관계를 유지한 채로 변환이 가능하다. 그리고 원래 인자가 인자회전한 인자보다 애매할 경우 인자회전을 한 인자를 사용할 수 있다. 그렇기에 먼저 "varimax"회전을 한 인자와 원래 인자의 비교를 통한 인자선정 후 인자의 해석을 실시한다.

rotate="varimax"

	RC1	RC2	RC3
득점	0.48	0.67	-0.03
어시스트	-0.09	-0.66	0.32
리바운드	0.84	0.31	-0.33
스틸	0.23	-0.65	0.07
블록	0.73	0.15	-0.14
X2점슛	0.68	0.37	-0.52
X3점슛	-0.57	-0.02	0.76
자유투	0.29	0.82	0.12
야투성공률	0.54	0.28	-0.70
X3점확률	0.01	0.02	0.90
자유투확률	-0.83	0.42	0.04

rotate="none"

	PC1	PC2	PC3
득점	0.64	0.44	0.26
어시스트	-0.51	-0.50	0.20
리바운드	0.92	-0.07	0.23
스틸	-0.13	-0.65	0.18
블록	0.67	-0.14	0.32
X2점슛	0.93	-0.01	-0.02
X3점슛	-0.84	0.35	0.28
자유투	0.48	0.68	0.28
야투성공률	0.89	-0.08	-0.25
X3점확률	-0.48	0.22	0.73
자유투확률	-0.46	0.68	-0.45

•회전후의 인자가 더 애매해지기에 회전하지 않은 인자를 사용한다.

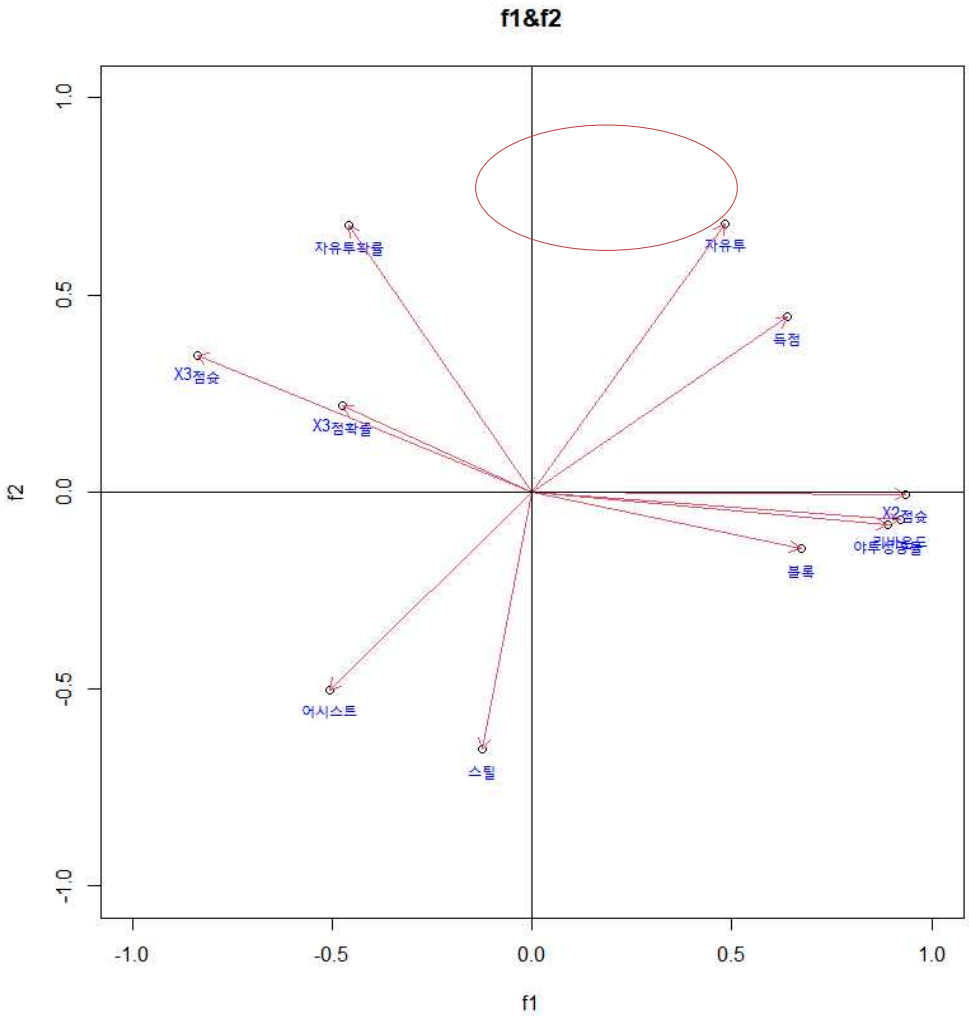
pcfa1: 리바운드, 2점슛, 야튜성공률과 높은 상관을 보이며, 3점슛과는 음의 높은 상관을 보인다. 이를 통해 pcfa를 2점과 리바운드 연관 변수이자 반 3점슛 변수라고 볼 수 있다.

pcfa2: 자유투, 자유투 확률과 높은 상관관계를 가지고 스틸과는 음의 높은 상관관계를 보인다. pcfa는 자유투 변수이자 반 스킬 변수라고 볼 수 있다.

pcfa3: 3점확률과 아주 높은 상관관계를 보인다. pcfa3은 2점 변수로 볼 수 있다.

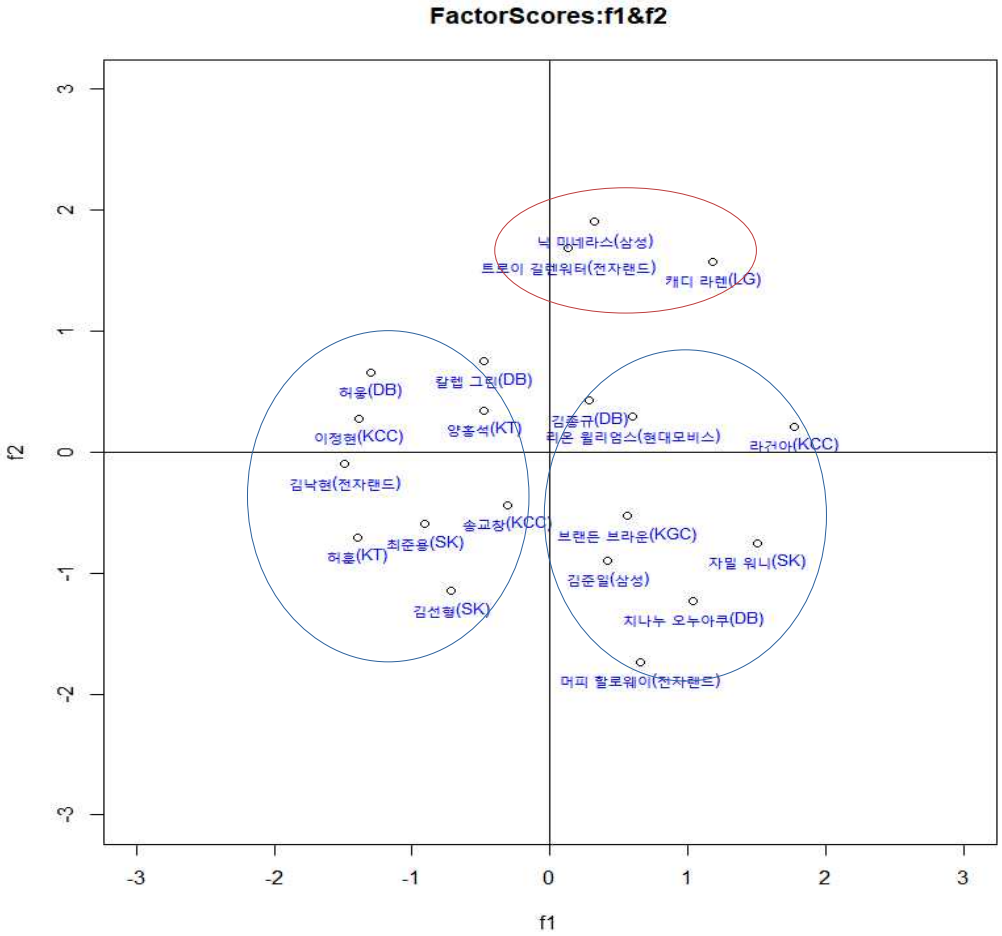
6-3. 인자적재그림과 인자점수그림을 통한 분석

1)f1&f2



---

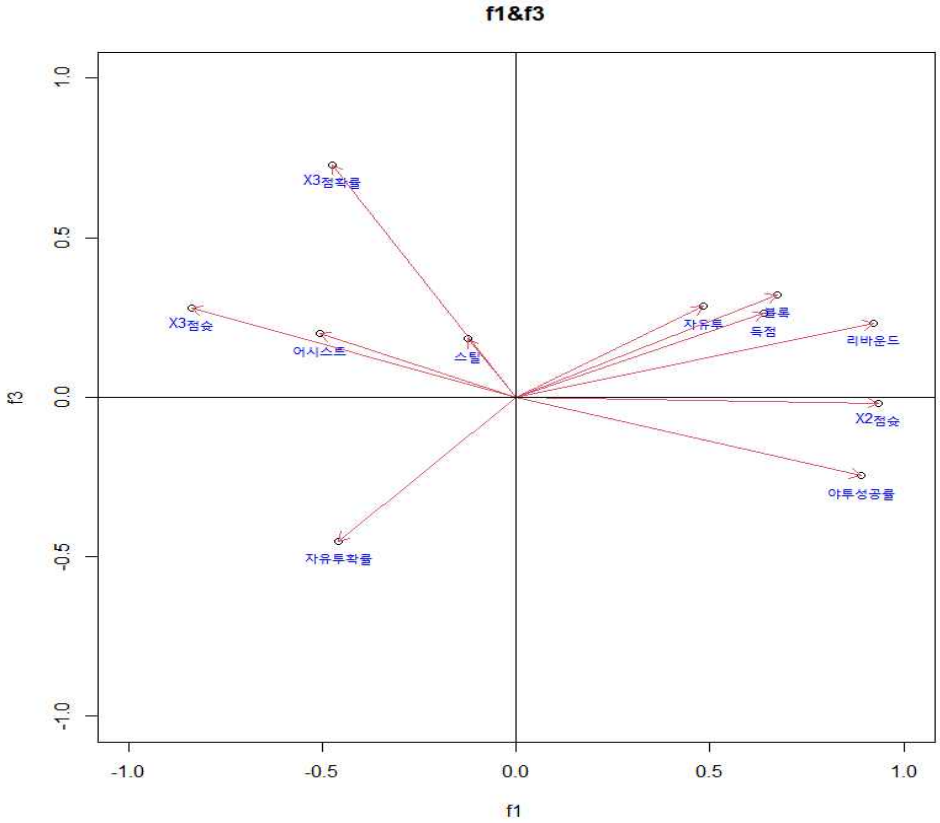
f2는 자유투, 스틸에 영향을 받으며, f1은 2점슛, 야투성공률, 리바운드, 블락 등에 영향을 많이 받음을 볼 수 있다.

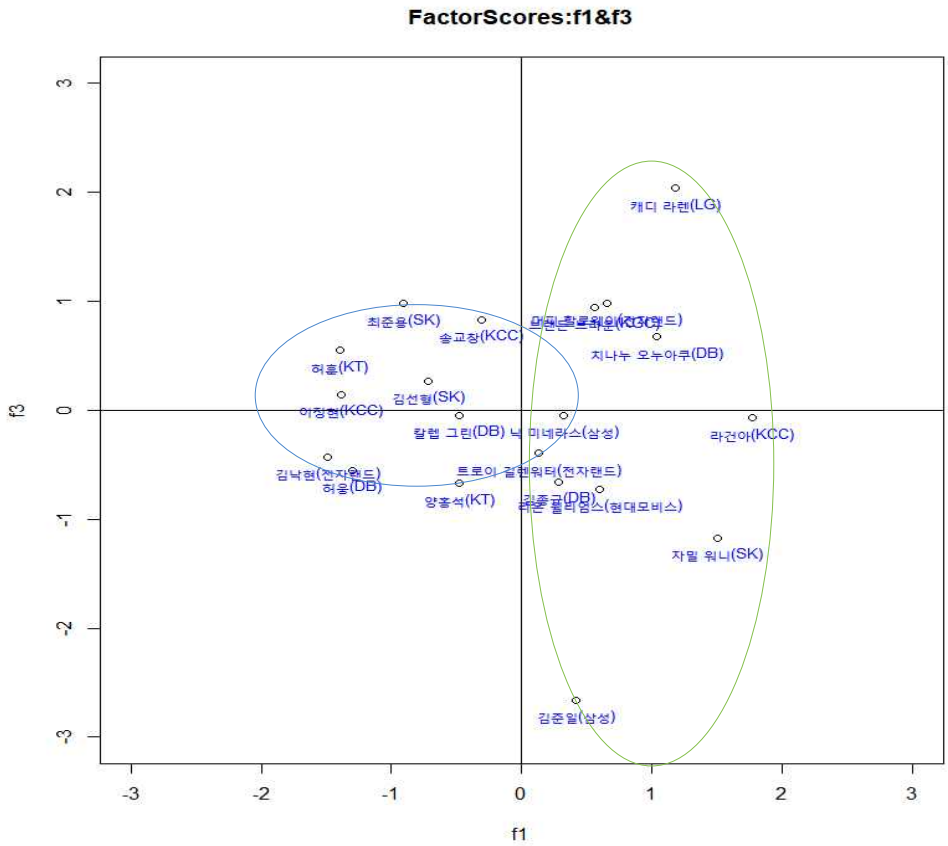


1) 1사분면 위쪽에 분포하는 선수들은 4점 이상의 3점슛을 기록하는 선수들이다.

2) f1을 기준으로 오른쪽은 C가 왼쪽은 G와F가 분포하고 있다.

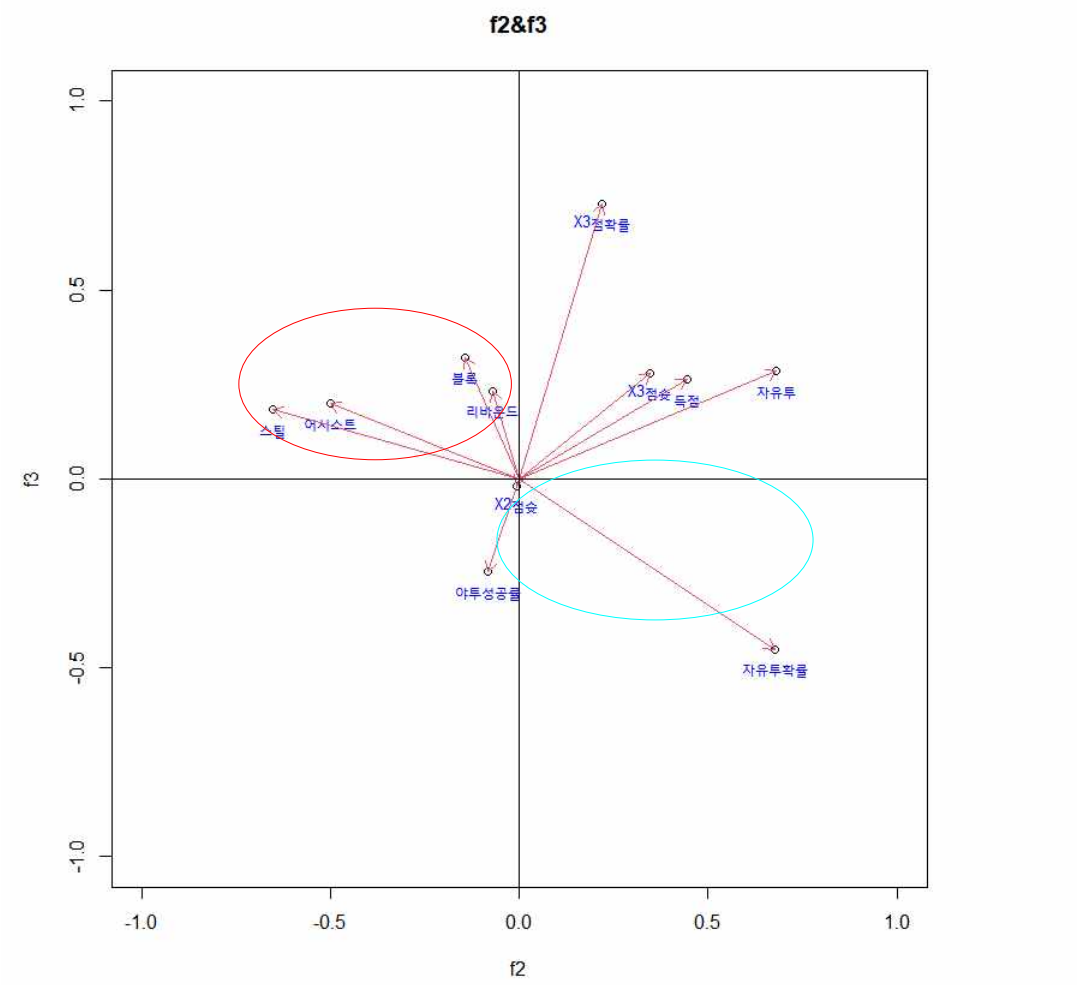
2)f1&f3

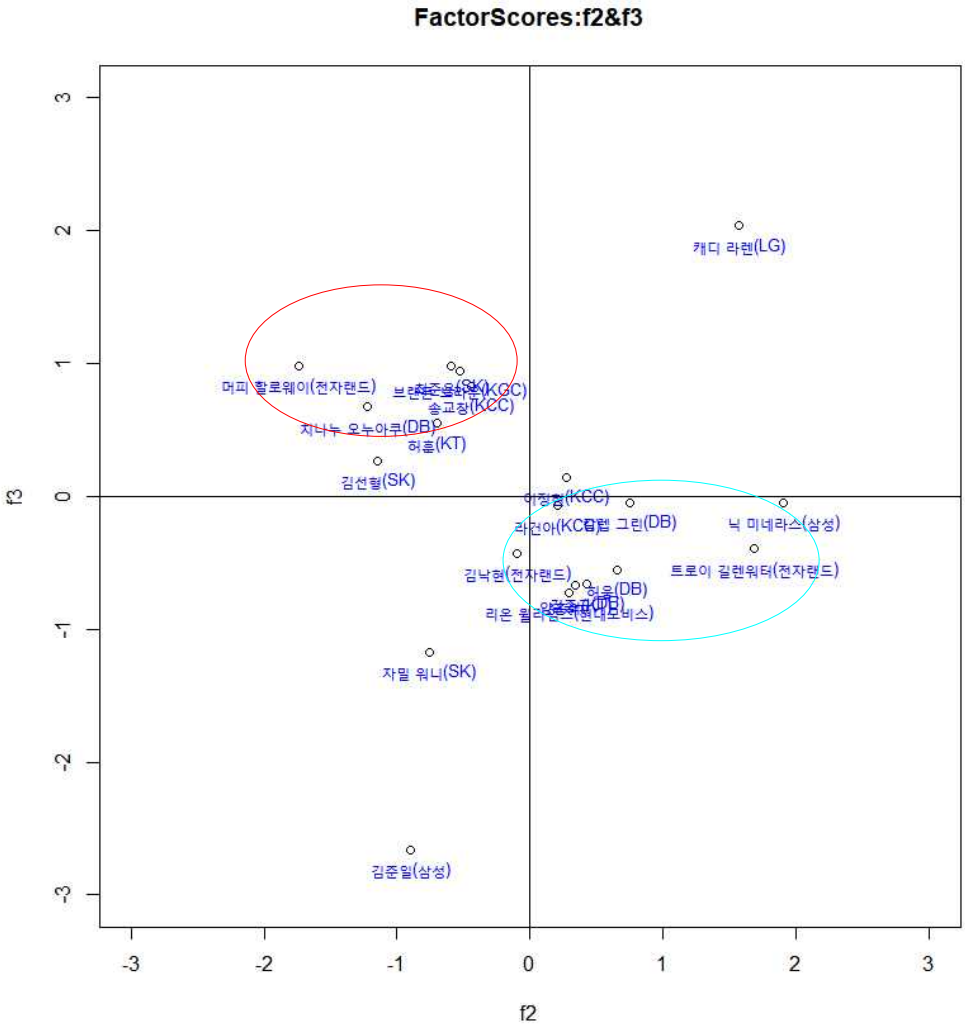

f1&f3
X3점확률
X3점슛
어시스트
스틸
자유투
블록
득점
리바운드
X2점슛
야투성공률
자유투확률
f1
f3


FactorScores:f1&f3
캐디 라렌(LG)
최준용(SK)
송교창(KCC)
치나누 오누아쿠(DB)
허훈(KT)
김선형(SK)
이정현(KCC)
라건아(KCC)
김낙현(전자랜드)
허웅(DB)
양홍석(KT)
자밀 워니(SK)
김준일(삼성)
f1
f3

1) f1에 의하여 오른쪽은 주로 C, 왼쪽은 주로 G와 F가 분포하고 있다. 예외로 닉 미네라스와 트로이 길렌워터가 F임에도 C의 영역에 분포한다. 이는 f1이 연관성 높은 변수중 하나가 리바운드임을 생각할 때 위의 둘은 F임에도 리바운드에 많이 참여함을 알 수 있다.
2) f3은 3점슛 변수로 선수들의 3점슛 기록으로 y값이 나뉘며 캐디라렌은 높은 3점기록으로 높은 y값을 기록하고, 자밀워니와 김준일은 3점 0개로 낮은 y값에 분포한다.

2)f2&f3





1) 2사분면에 분포하는 선수들은 4사분면에 분포하는 선수들보다 비교적 낮은 자유투확률을 가진다.
2) 유일하게 1사분면에 있는 캐디라렌은 자유투와 3점슛이 모두 좋다고 볼 수 있다.
3) 자밀워니와 김준일은 자유투와 3점슛 둘다 기록이 좋지 못하여 4영역에 분포한다.

7. MLFA(최대우도인자분석)

이번엔 FA중 MLE 방법을 활용한 MLFA를 이용한다. 앞서 데이터에 대한 정규성을 검토하였기에 MLFA의 적용에 문제는 없다.

7-1. 인자선택

```
> mlfa

Call:
factanal(factors = 4, covmat = R, rotation = "none")

Uniquenesses:
      득점     어시스트     리바운드        스틸        블록      X2점슛      X3점슛      자유투 야투성공률
     0.005       0.698       0.090       0.932       0.367       0.005       0.005       0.005       0.166
  X3점확률 자유투확률
     0.322       0.356

Loadings:
           Factor1 Factor2 Factor3 Factor4
득점         0.869   0.394           0.293
어시스트    -0.416                   0.358
리바운드     0.803  -0.297   0.421
스틸        -0.110  -0.194           0.129
블록         0.463  -0.316   0.553  -0.115
X2점슛       0.944  -0.255           0.205
X3점슛      -0.632   0.738           0.228
자유투       0.718   0.550          -0.420
야투성공률   0.747  -0.509          -0.124
X3점확률    -0.391   0.516   0.485  -0.155
자유투확률 -0.252   0.501  -0.556  -0.142

                Factor1 Factor2 Factor3 Factor4
SS loadings       4.380   2.070   1.032   0.574
Proportion Var    0.398   0.188   0.094   0.052
Cumulative Var    0.398   0.586   0.680   0.732

The degrees of freedom for the model is 17 and the fit was 8.1816
> screeplot(mlfa$)
```

일반적으로 설명력이 70%이상을 만들도록 인자를 선택한다. 그러므로 kbl데이터를 MLFA하기에 적합한 인자 수는 4이다. 하지만 편의를 위해 70%에 근사치인 68%를 나타내는 인자3까지를 인자로 채택한다.

6-2.인자의 성질

FA는 변수들의 관계를 유지한 채로 변환이 가능하다. 그리고 원래 인자가 인자회전한 인자보다 애매할 경우 인자회전을 한 인자를 사용할 수 있다. 그렇기에 먼저 "varimax"회전을 한 인자와 원래 인자의 비교를 통한 인자선정 후 인자의 해석을 실시한다.

rotate="varimax"

	Factor1	Factor2	Factor3
득점	0.12	0.14	0.41
어시스트	-0.18	-0.24	-0.46
리바운드	0.76	0.34	0.25
스틸	0.02	0.07	-0.24
블록	0.76	0.13	0.18
X2점슛	0.44	0.61	0.19
X3점슛	-0.55	-0.82	-0.12
자유투	0.01	0.05	0.92
야투성공률	0.46	0.73	0.21
X3점확률	0.03	-0.78	0.19
자유투확률	-0.77	-0.11	0.20

rotate="none"

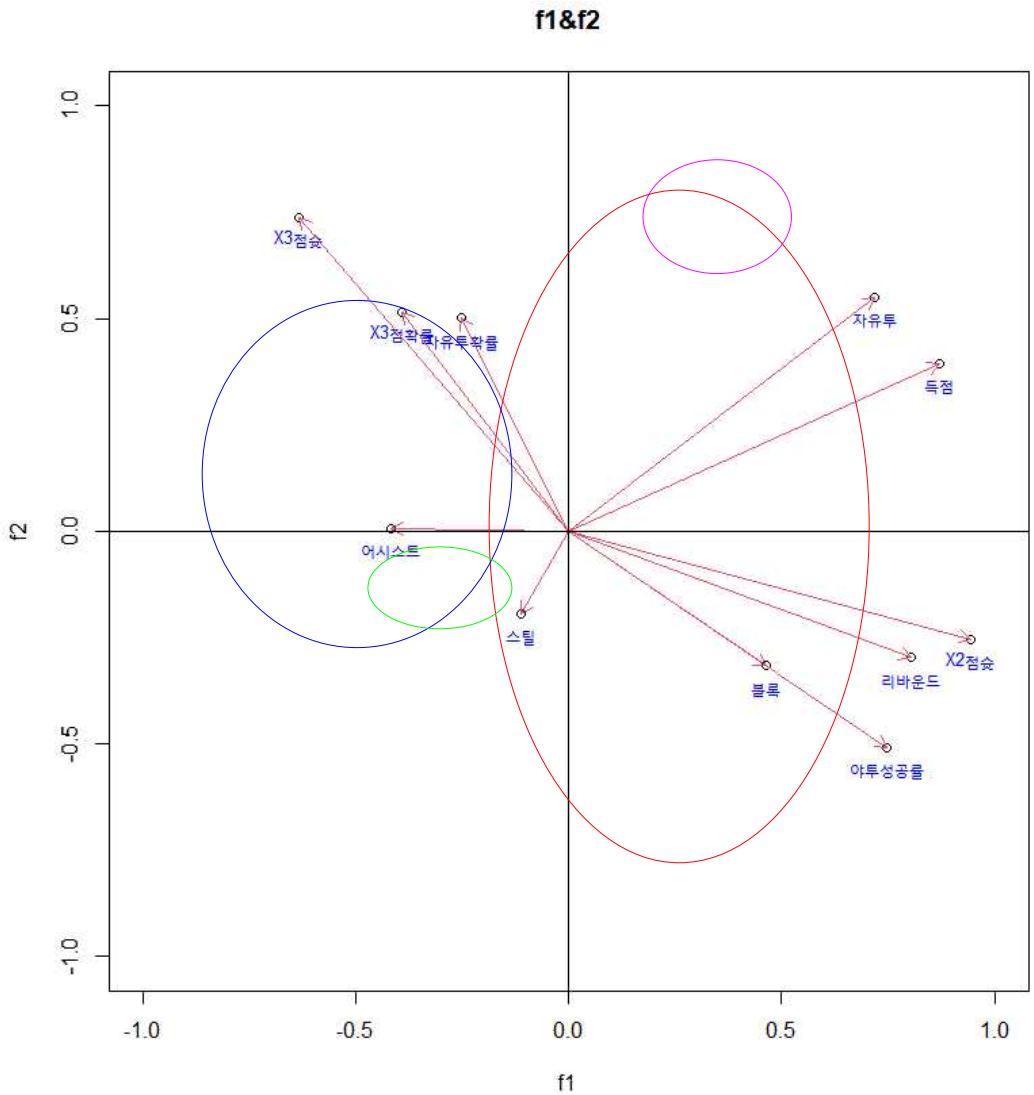
	Factor1	Factor2	Factor3
득점	0.87	0.39	0.00
어시스트	-0.42	0.01	-0.03
리바운드	0.80	-0.30	0.42
스틸	-0.11	-0.19	-0.03
블록	0.46	-0.32	0.55
X2점슛	0.94	-0.25	-0.01
X3점슛	-0.63	0.74	0.01
자유투	0.72	0.55	0.00
야투성공률	0.75	-0.51	-0.04
X3점확률	-0.39	0.52	0.49
자유투확률	-0.25	0.50	-0.56

회전을 한 후의 인자가 더 애매하기에 회전하지 않은 인자를 사용한다.

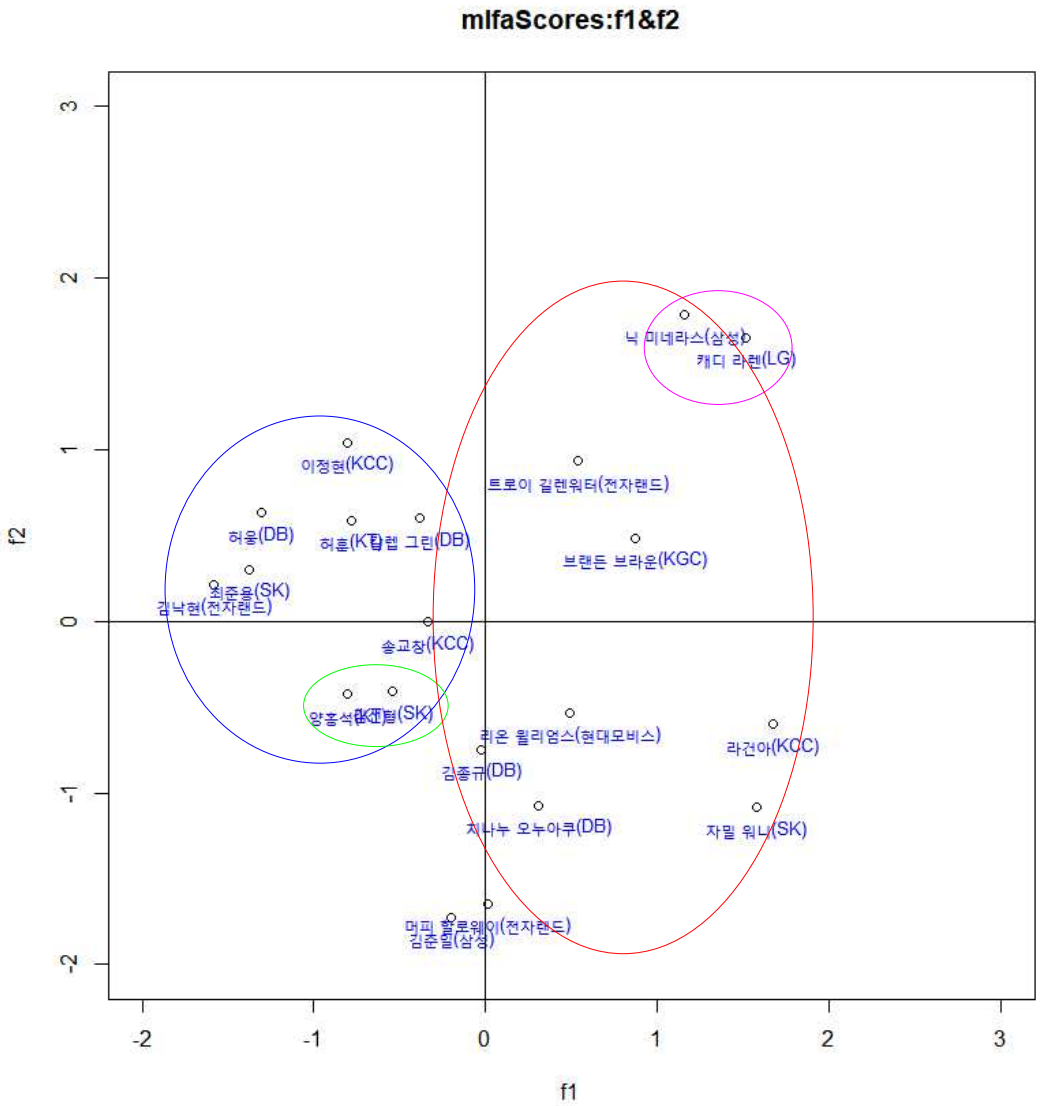
f1: 2점슛, 자유투, 야투성공률, 리바운드, 득점과 높은 상관관계를 가진다.
f2: 자유투와 3점과는 비교적 높은 수준의 양의 상관관계를 가지고 2점슛과는 음의 상관관계를 가진다.
f3: 블록과 3점슛확률 과는 비교적 높은 양의 상관관계를 가지고 자유투와는 음의 상관관계를 가진다.

7-3. 인자적재그림과 인자점수그림을 통한 분석

1)f1&f2

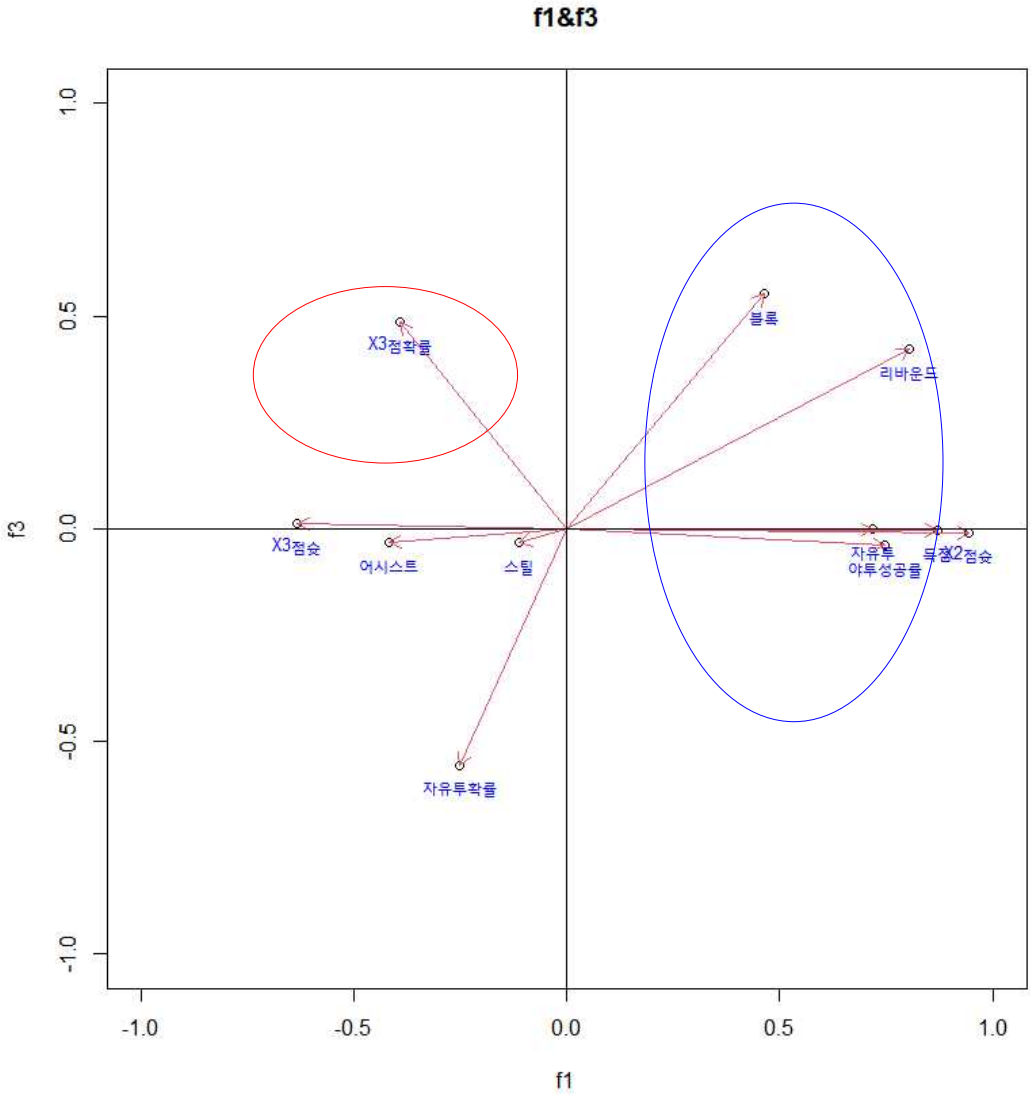


3점슛과 3점슛 성공률이 자유투 확률과 함께 fc2와 관련이 있어 보이며 fc1은 특히 어시스트와 큰 음의. 상관관계를 가진다

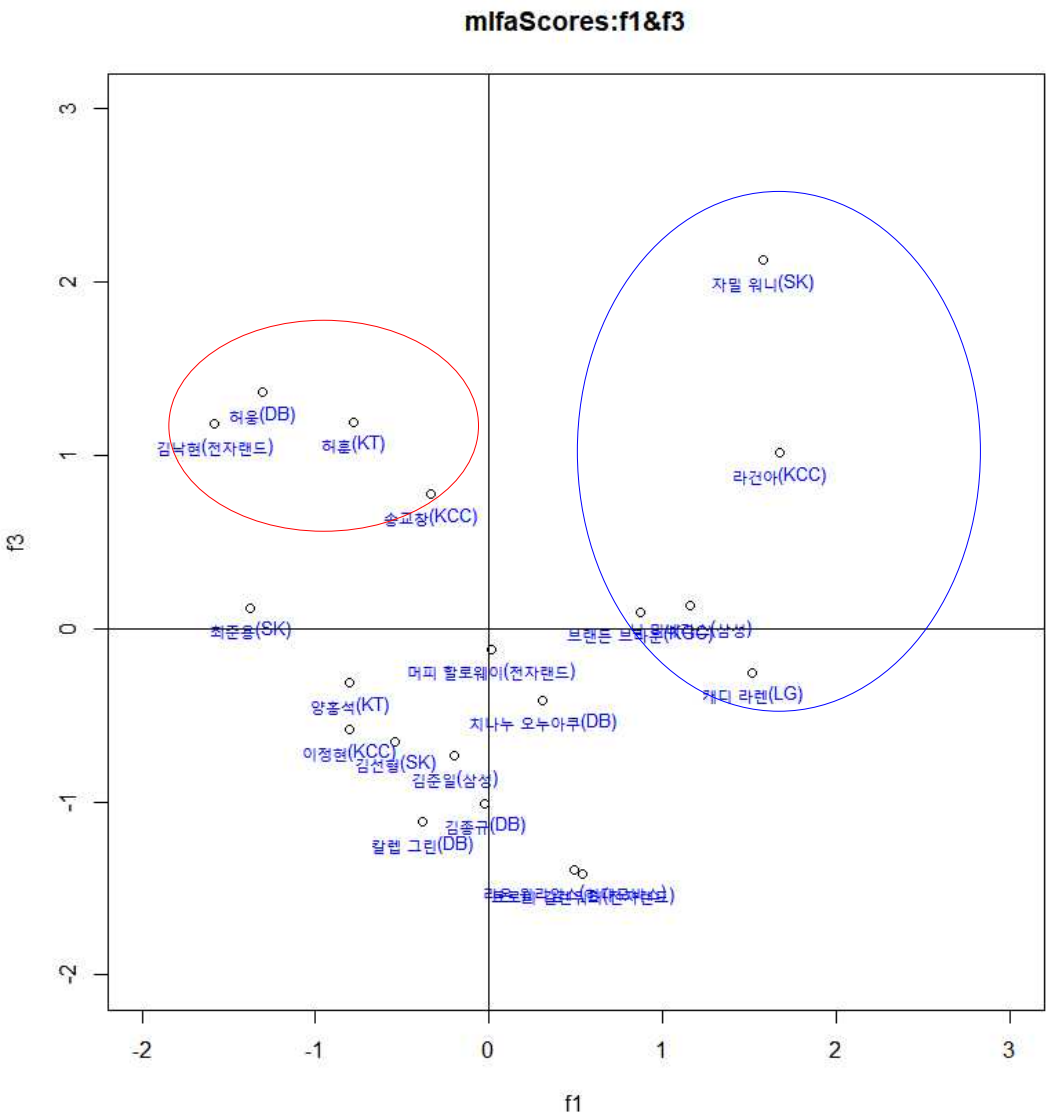


1) f1이 어시스트와 아주 음의 상관을 가지므로 왼쪽은 F,G위주의 선수들이 분포해 있고 오른쪽은 C위주의 선수들로 분포되어있다.
2) 특히 오른쪽 위에는 득점과, 자유투 둘 다 최상위권인 캐디 라렌과 닉 미네라스가 분포한다.
3) 왼쪽의 G와 F중에서도 밑에 있는 선수들은 비교적 저조한 3점 기록을 보인다.

2)f1&f3

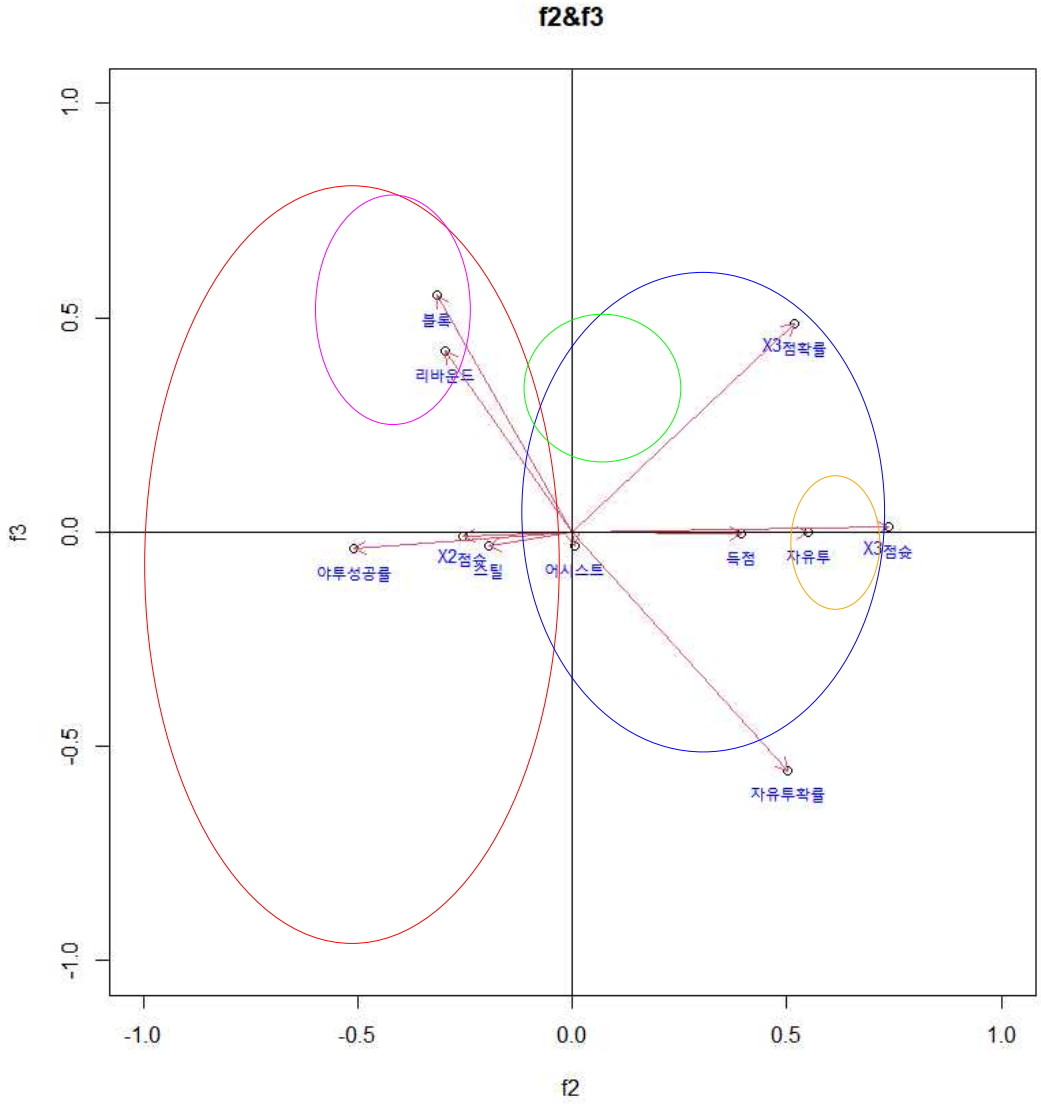


f1이 f2와의 plot에서 보다 더욱 2점슛, 야투성공률, 자유투, 득점과 강한 양의 상관관계를 가지며, 3점슛과 어시스트, 스틸과는 음의 상관관계를 가진다. 또한 f3점 확률과 양의 상관을 자유투확률과는 음의 상관관계를 가진다.

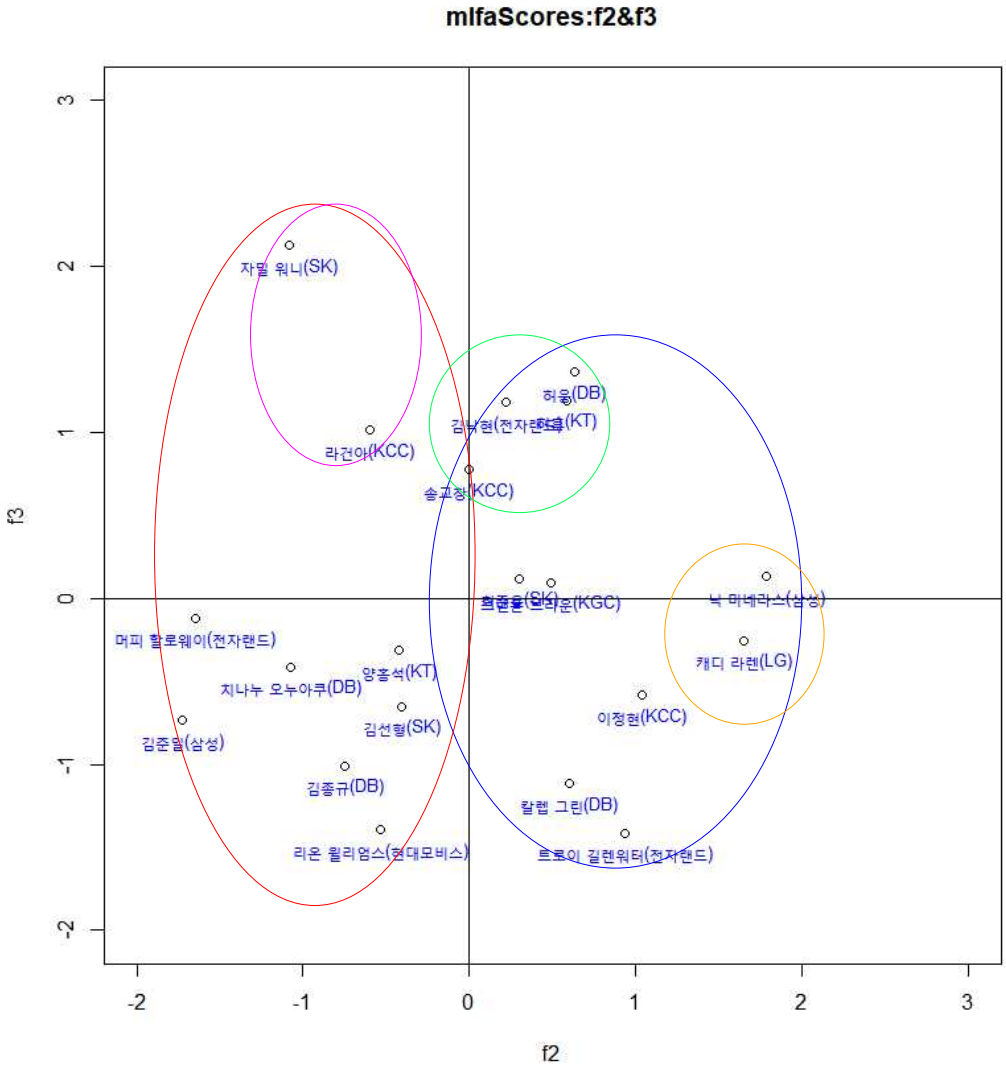


1) f1&f2에서와 동일하게 f1에 왼쪽은 F,G위주, 오른쪽은 C위주로 나뉜다.
2) F,G중에도 가장 3점 확률이 높은 4명이 왼쪽 상단에 분포한다.
3) 전체중에 가장 득점이 높은 5명이 비교적 왼쪽 위쪽에 분포한다.

3)f2&f3



f2는 3점슛, 자유투, 득점과 아주 강한 양의 상관관계를 가지고, 야투성공률, 2점슛, 스틸과는 아주 강한 음의 상관관계를 가진다. f3는 비교적 3점과 리바운드와 양의 상관관계를 가지며, 자유투확률과는 음의 상관관계를 가진다.



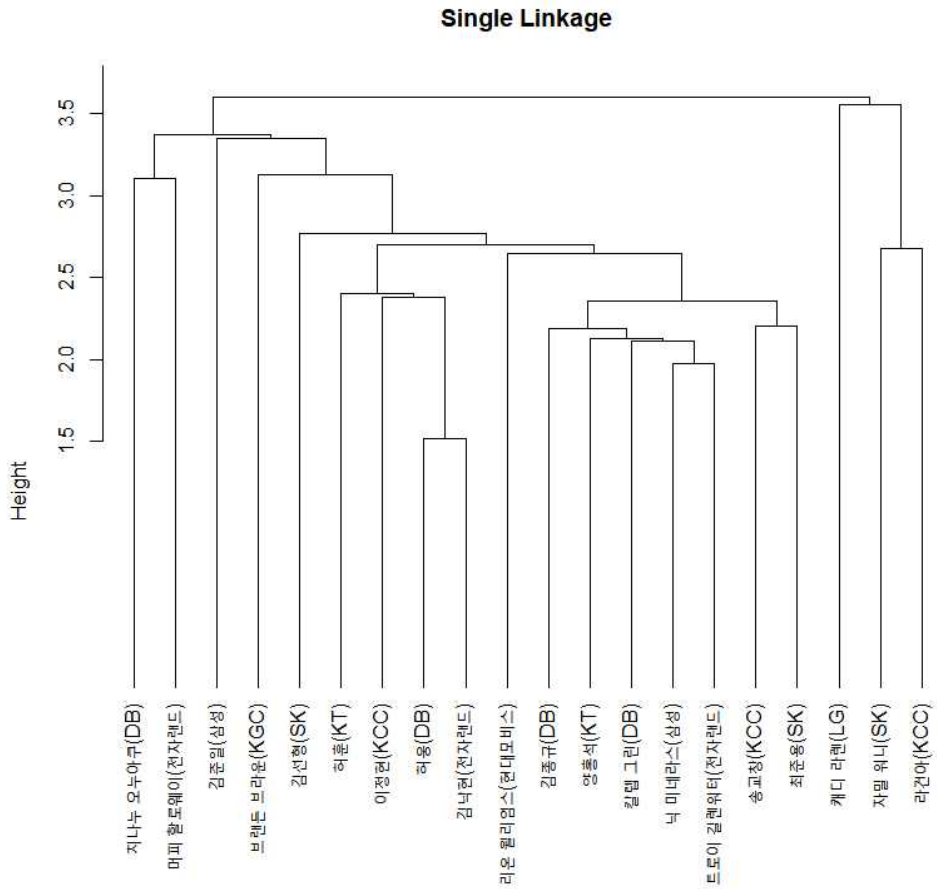
1) f2로 인해 오른쪽은 주로 G, F로 왼쪽은 C로 나눠져 분포된다.
2) G, F중 3점슛 확률이 가장 높은 3명은 f3 축에 의해 높게 분포한다.
3) 3점과 자유투위주의 득점이 높은 선수들은 오른쪽에 치우쳐있다.
4) 리바운드나 블록 등의 수비력이 높은 센터는 왼쪽 위에 존재한다.

7. CA(군집분석)

앞에서 실행한 PCA와 FA는 변수들 간의 관계를 나타내는데 초점을 맞춘 R-방법이었다. 이번엔 관측치 간의 관계를 나타내는 Q-방법 중 하나인 군집분석(CA)를 이용하여 분석을 해본다.

7.1계층 군집분석(표준화 유클리드 거리를 이용하여 실시)

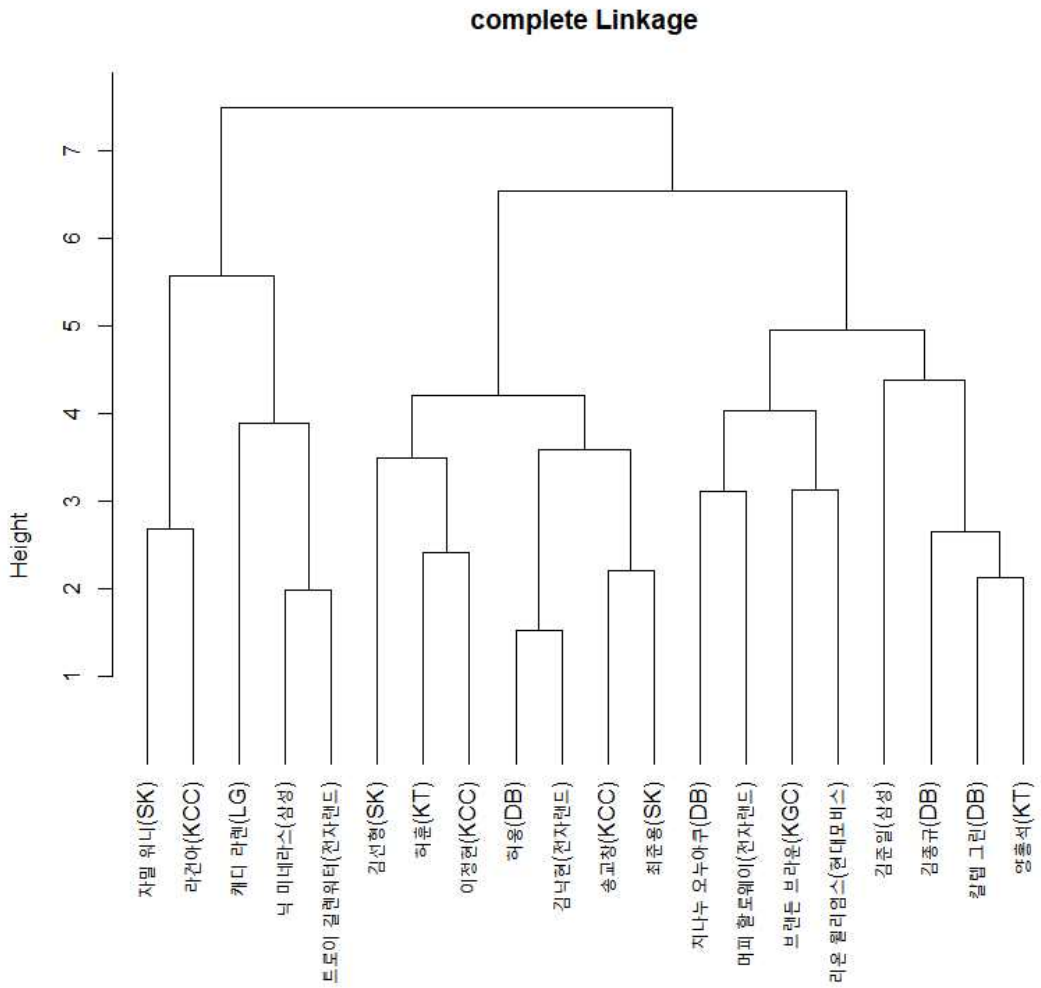
7.1-1. 단일 연결법



ds
hclust (*, "single")

C1	C2	C3	C4
오누아쿠, 할로웨이, 김준일, 브라운	김선형, 허훈, 이정현, 허웅, 김낙현	윌리엄스,김종규,양홍석,그린,미네라스, 송교창, 최준용	라렌, 자밀 워니, 라건아
C4의 득점 상위권 센터와 C3로 분류된 센터를 제외한 센터	포지션이 G인 선수들	포워드 위주의 군집으로 보인다.	센터 중 득점 최상위권 3선수이다.

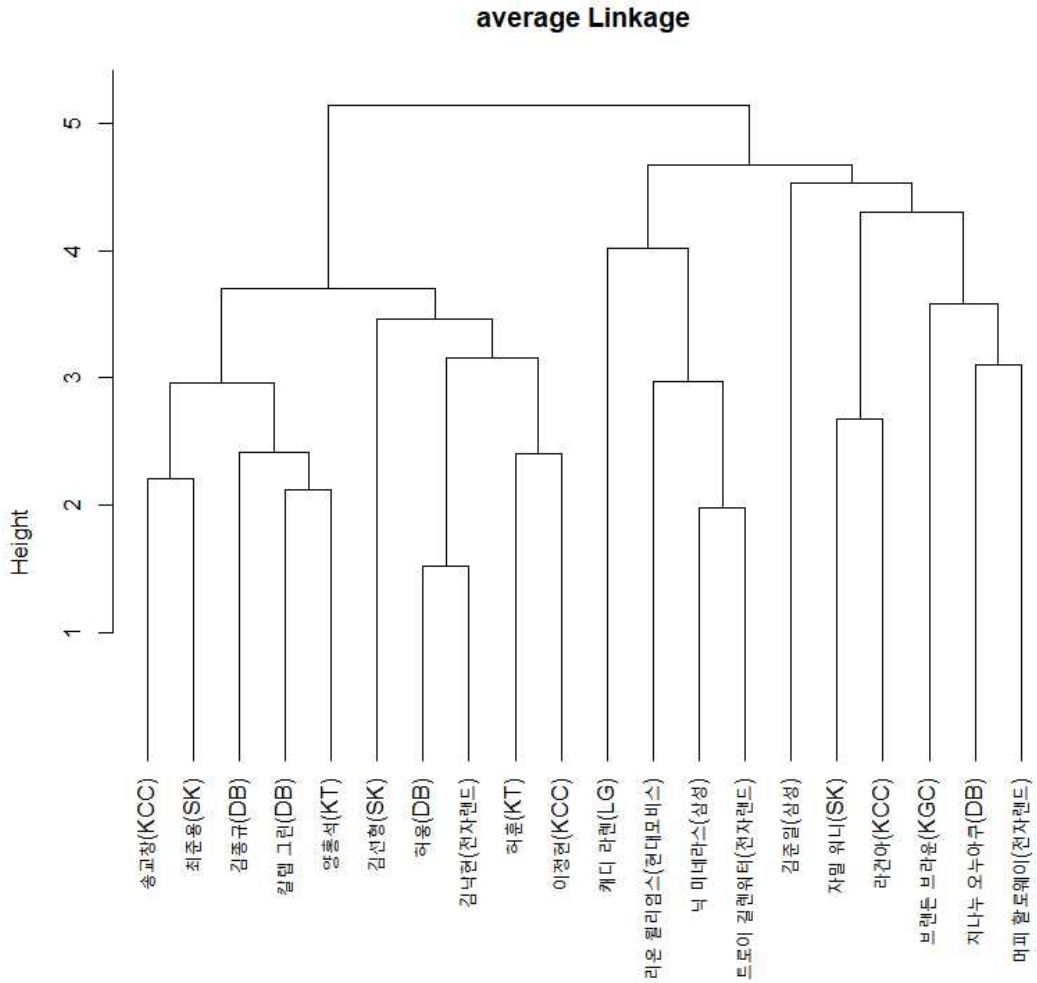
7.1-2. 완전연결법



ds
hclust (*, "complete")

C1	C2	C3
자밀 워니, 라건아, 라렌, 미네라스, 길렌워터	김선형, 허훈, 이정현, 허웅, 김낙현, 송교창, 최준용	나머지 선수들
포지션 관계없이 득점이 상위권인 5명의 선수	상위권을 제외한 G와 F 선수들 위주	상위권을 제외한 C 선수들 위주

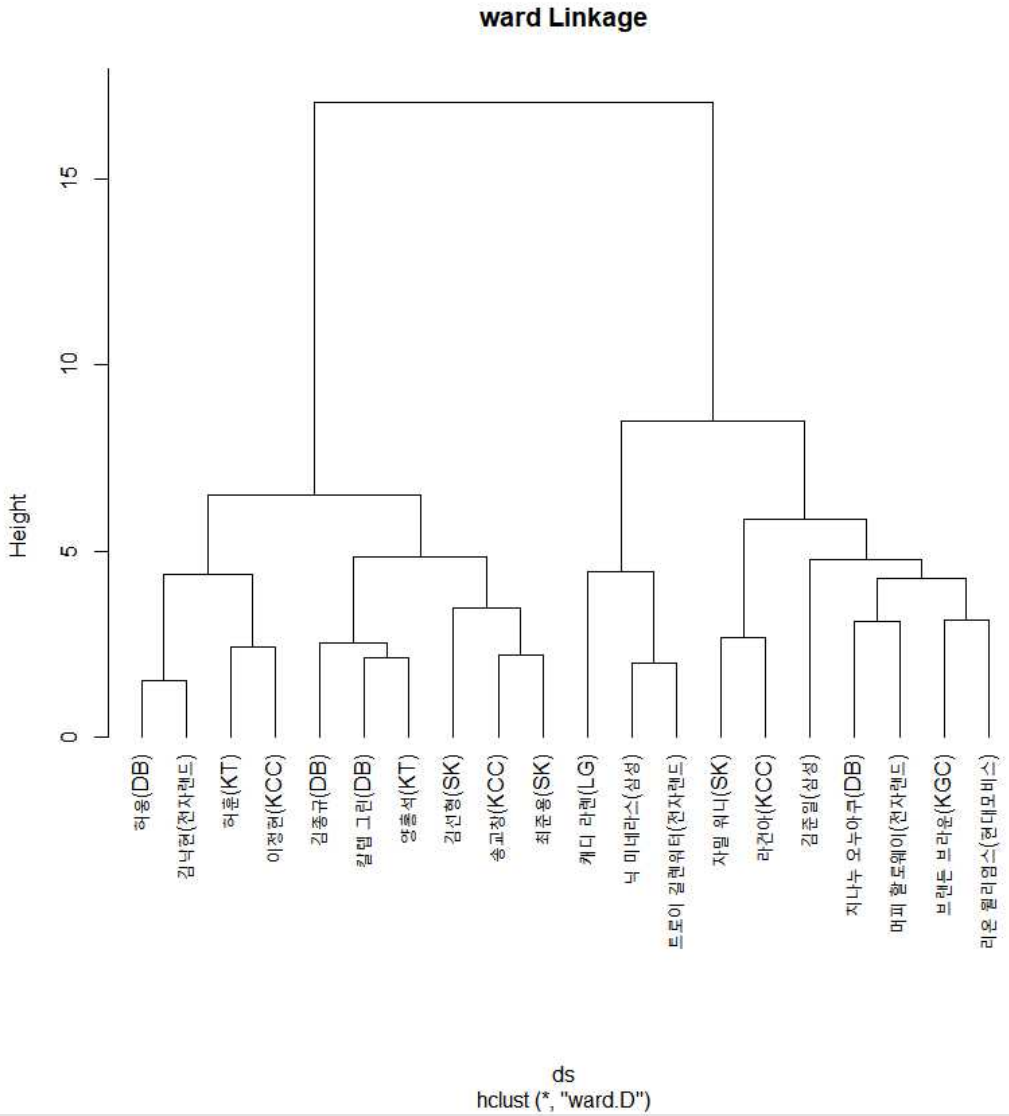
7.1-3. 평균연결법



ds
hclust (*, "average")

C1	C2	C3
송교창, 최준용, 김종규, 그린, 양홍석	김선형, 허웅, 김낙현, 허훈, 이정현	나머지 선수
F선수들 위주의 군집으로 보인다.	G선수들의 군집이다.	C선수들 위주의 군집이다

7.1-4. 와드연결법



C1	C2	C3
허웅, 김낙현, 허훈, 이정현, 김정규, 그린, 양홍석, 김선형, 송교창, 최준용	캐디라렌, 미네라스, 길렌워터, 워니, 라건아	김준일, 오누아쿠, 할로웨이, 브라운, 윌리엄스
G와 F선수들 위주의 군집	득점상위권C들 위주의 군집	상위권이 아닌 C들의 군집

•계층 군집분석은 전체적으로 포지션별로 군집을 이루며, 센터는 득점 순위에 따라 군집이 나누어지기도 했다. 약간의 이상점은 모든 군집분석에서 김종규(C)는 포워드로 주로 분류되었고, 미네라스(F)와 길렌워터(F)는 센터로 많이 분류되었다. 이러한 자료들은 꼭 선수의 포지션이 선수 각자의 특성과 맞지는 않다는 것을 보여준다고 생각한다.

7.2비계층 군집분석(allindex를 이용하여 군집의 수를 정한다.)

7.2-1. 군집의 수

```
> 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:
* 8 proposed 2 as the best number of clusters
* 7 proposed 3 as the best number of clusters
* 2 proposed 4 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 2 proposed 6 as the best number of clusters
* 1 proposed 7 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
```

allindex 결과 8개의 index가 2개의 군집을, 7개의 index가 3개의 군집을 나타낸다.
그러므로 군집을 2개로 나누었을 때의 결과를 kmeans 방법과 kmedoids 방법을 이용하여 도출해본다.

7.2-2. kmeans방법

```
> kmeans=kmeans(z,2)
> cluster=data.frame(player,cluster=kmeans$cluster)
> C1=cluster[(cluster[,2]==1),]
> C2=cluster[(cluster[,2]==2),]
> C1;C2
                                                player cluster
캐디 라렌(LG)                                  캐디 라렌(LG)       1
닉 미네라스(삼성)                            닉 미네라스(삼성)       1
자밀 워니(SK)                                  자밀 워니(SK)       1
라건아(KCC)                                    라건아(KCC)       1
브랜든 브라운(KGC)                          브랜든 브라운(KGC)       1
트로이 길렌워터(전자랜드) 트로이 길렌워터(전자랜드)       1
리온 윌리엄스(현대모비스) 리온 윌리엄스(현대모비스)       1
치나누 오누아쿠(DB)                        치나누 오누아쿠(DB)       1
김종규(DB)                                     김종규(DB)       1
머피 할로웨이(전자랜드)      머피 할로웨이(전자랜드)       1
김준일(삼성)                                   김준일(삼성)       1
                          player cluster
송교창(KCC)             송교창(KCC)       2
허훈(KT)                   허훈(KT)       2
칼렙 그린(DB)          칼렙 그린(DB)       2
이정현(KCC)             이정현(KCC)       2
허웅(DB)                   허웅(DB)       2
김선형(SK)               김선형(SK)       2
김낙현(전자랜드) 김낙현(전자랜드)       2
양홍석(KT)               양홍석(KT)       2
최준용(SK)               최준용(SK)       2
```

해석:

정확히 G와 F를 C로부터 분리한 군집으로 보이고 군집1 역시 C의 군집으로 보이나, 길렌워터와 미네라스는 F임에도 C로 분류되었다. 앞선 실행했던 분석들을 함께 고려해본 결과 길렌워터와 미네라스는 상당히 C의 성격을 가진다.

7.2-2. kmedoids방법

```
                                                player cluster
캐디 라렌(LG)                                 캐디 라렌(LG)       1
닉 미네라스(삼성)                           닉 미네라스(삼성)       1
자밀 워니(SK)                                 자밀 워니(SK)       1
라건아(KCC)                                   라건아(KCC)       1
브랜든 브라운(KGC)                         브랜든 브라운(KGC)       1
트로이 길렌워터(전자랜드) 트로이 길렌워터(전자랜드)       1
리온 윌리엄스(현대모비스) 리온 윌리엄스(현대모비스)       1
치나누 오누아쿠(DB)                       치나누 오누아쿠(DB)       1
칼렙 그린(DB)                                 칼렙 그린(DB)       1
김종규(DB)                                    김종규(DB)       1
머피 할로웨이(전자랜드)      머피 할로웨이(전자랜드)       1
김준일(삼성)                                  김준일(삼성)       1
                                  player cluster
송교창(KCC)                 송교창(KCC)       2
허훈(KT)                      허훈(KT)       2
이정현(KCC)                 이정현(KCC)       2
허웅(DB)                      허웅(DB)       2
김선형(SK)                   김선형(SK)       2
김낙현(전자랜드) 김낙현(전자랜드)       2
양홍석(KT)                   양홍석(KT)       2
최준용(SK)                   최준용(SK)       2
>
```

해석:

Kmean방법과는 약간 다르게 2군집은 국내 G와 F 만 분류가 되었고 C와 외국인 선수들은 군집 1로 분류되었다.

8.결론

데이터에 대한 다변량 자료분석 결과 프로리그에서도 득점과 관계없이 대부분의 선수들이 자신의 포지션에 맞는 역할을 하고 있으며, 포지션별 특성이 뚜렷하게 나타남을 알 수 있었다. 하지만 닉 미네라스, 트로이 길렌워터, 김종규 선수 등 현재 포지션과는 결과가 아예 다른 방향으로 나오는 선수도 있었다. 이를 이용하면 팀에서도 선수를 영입할 때 드러나는 포지션뿐만 아닌 정말로 팀에 필요한 선수인지를 조금 더 신중하게 판단할 수 있는 근거가 생길 것이다.

또한, 이는 스포츠팀에 데이터 분석이 필요한 이유를 설명해준다.

Reference
네이버 통계 자료

Rcode

```
setwd("C:/Users/stat/Desktop/새 폴더")
data<-read.csv("kbl.csv",header=T)
head(data)
player=data[,1]
kbl<-data[,c(-1,-2)]
kbl
summary(kbl)
round(cov(kbl),3)
round(cor(kbl),3)
Z<-scale(kbl)
rownames(Z)<-player
X<-t(Z)
barplot(X,legend=rownames(X),horiz=T)
player
stars(Z, key.loc=c(0, 1), full=FALSE)
#win.graph()
Z<-as.matrix(Z)
n=dim(Z)[1]
p=dim(Z)[2]
S<-cov(Z)
Zbar=colMeans(Z)
m=mahalanobis(Z,Zbar,S)
m=sort(m)
id=seq(1,n)
pt=(id-0.5)/n
q=qchisq(pt,p)
plot(q,m,pch="*",main="chisq",xlab="Quantile",ylab="Ordered squared Distance")
abline(0,1)
rq<-cor(cbind(q,m))[1,2]
rq
#mardiaTset
install.packages("MVN")
library(MVN)
result<-mvn(Z,mvnTest="mardia",multivariatePlot="qq")

#pca
pcasvd<-prcomp(Z,scale=T)
summary(pcasvd)
screeplot(pcasvd,type="lines")
round(pcasvd$rotation[,1:3],3)

#정확도를 위한 svdfmf 사용하여 pca를 한다
#오른쪽= 2점득점 ,왼쪽 3점슛 > 왼쪽에
#svd이용한 biplot 모든 pc에대한plot그릴려다가 말았음음
svd<-svd(Z)
n=nrow(Z)
joinnames=c(rownames(Z),colnames(Z))
U=svd$u
V=svd$v
D=diag(svd$d)
```

```
G=(sqrt(n-1)*U)
rownames(G)=rownames(Z)
colnames(G)<-colnames(Z)
G1=G[,1:2]
g=G[,-2]
G2=g[,1:2]   #13pc
G3=G[,2:3]
H=(sqrt(1/(n-1))*V%*%D)
rownames(H)=colnames(Z)
colnames(H)<-colnames(Z)
H1=H[,1:2]
h<-H[,-2]
H2=H[,1:2]
H3=H[,2:3]
C1=rbind(G1,H1)
C2=rbind(G2,H2)
C3=rbind(G3,H3)
C=rbind(G,H)
par(mfrow=c(1,1))
par(pty="s")
lim=range(pretty(C))
biplot(G1,H1,xlab="pc1",ylab="pc2",main="pc1&pc2Biplot",xlim=lim,ylim=lim,cex=0.8,pch=16)
abline(v=0,h=0)
biplot(G2,H2,xlab="pc1",ylab="pc3",main="pc1&pc3Biplot",xlim=lim,ylim=lim,cex=0.8,pch=16)
abline(v=0,h=0)
biplot(G3,H3,xlab="pc2",ylab="pc3",main="pc2&pc3Biplot",xlim=lim,ylim=lim,cex=0.8,pch=16)
abline(v=0,h=0)

R=cor(Z)
#FA
install.packages("psych")
library(psych)
pcfa<-principal(Z,nfactor=3,rotate="varimax")   #pca를 근거로 3개의factor를 이용한다다
pcfa
summary(pcfa)
round(pcfa$values,3)
gof=pcfa$values/sum(pcfa$values)*100
gof<-round(gof,3)
gof
plot(pcfa$values,main="screeplot",type = "b")
L=pcfa$loadings[,1:3]
round(L,2)
Psi=diag(pcfa$uniquenesses)
Rm=R-(L%*%t(L)+Psi)
round(Rm,2)
#plot factor
lim=range(pretty(L))
plot(L[,1],L[,2],main="f1&f2",xlab="f1",ylab="f2",xlim=lim,ylim=lim)
text(L[,1],L[,2],labels=rownames(L),cex=0.8,col="blue",pos=1)
arrows(0,0,L[,1],L[,2],col=2,length = 0.1)
abline(v=0,h=0)
```

```
plot(L[,1],L[,3],main="f1&f3",xlab="f1",ylab="f3",xlim=lim,ylim=lim)
text(L[,1],L[,3],labels=rownames(L),cex=0.8,col="blue",pos=1)
arrows(0,0,L[,1],L[,3],col=2,length = 0.1)
abline(v=0,h=0)

plot(L[,2],L[,3],main="f2&f3",xlab="f2",ylab="f3",xlim=lim,ylim=lim)
text(L[,2],L[,3],labels=rownames(L),cex=0.8,col="blue",pos=1)
arrows(0,0,L[,2],L[,3],col=2,length = 0.1)
abline(v=0,h=0)

#scores plot
fpc<-pcfa$scores
lim=range(pretty(fpc))
plot(fpc[,1],fpc[,2],main="FactorScores:f1&f2",xlab="f1",ylab="f2",xlim=lim,ylim=lim)
text(fpc[,1],fpc[,2],labels=player,cex=0.8,col="blue",pos=1)
abline(v=0,h=0)
plot(fpc[,1],fpc[,3],main="FactorScores:f1&f3",xlab="f1",ylab="f3",xlim=lim,ylim=lim)
text(fpc[,1],fpc[,3],labels=rownames(fpc),cex=0.8,col="blue",pos=1)
abline(v=0,h=0)
plot(fpc[,2],fpc[,3],main="FactorScores:f2&f3",xlab="f2",ylab="f3",xlim=lim,ylim=lim)
text(fpc[,2],fpc[,3],labels=rownames(fpc),cex=0.8,col="blue",pos=1)
abline(v=0,h=0)
Z<-scale(Z)

#mlfa
mlfa<-factanal(covmat=R, factors = 4, rotation="varimax" )
mlfa

#3fc는 설명력이 0.68로 목표치인 0.7에는 못미치지만 근차치로 3요소를 사용한다.
L=mlfa$loadings[,1:3]
round(L,2)
Psi=diag(mlfa$uniquenesses)
Rm=R-(L%*%t(L)+Psi)
#plotfc
lim=range(pretty(L))
plot(L[,1],L[,2],main="f1&f2",xlab="f1",ylab="f2",xlim=lim,ylim=lim)
text(L[,1],L[,2],labels=rownames(L),cex=0.8,col="blue",pos=1)
arrows(0,0,L[,1],L[,2],col=2,length = 0.1)
abline(v=0,h=0)

plot(L[,1],L[,3],main="f1&f3",xlab="f1",ylab="f3",xlim=lim,ylim=lim)
text(L[,1],L[,3],labels=rownames(L),cex=0.8,col="blue",pos=1)
arrows(0,0,L[,1],L[,3],col=2,length = 0.1)
abline(v=0,h=0)

plot(L[,2],L[,3],main="f2&f3",xlab="f2",ylab="f3",xlim=lim,ylim=lim)
text(L[,2],L[,3],labels=rownames(L),cex=0.8,col="blue",pos=1)
arrows(0,0,L[,2],L[,3],col=2,length = 0.1)
abline(v=0,h=0)
#scoreplot
Mlfa<-factanal(Z,factors=3,rotation="none",score="regression")
```

```
fml=Mlfa$scores
lim=range(pretty(fml))
plot(fml[,1],fml[,2],main="mlfaScores:f1&f2",xlab="f1",ylab="f2",xlim=lim,ylim=lim)
text(fml[,1],fml[,2],labels=rownames(fml),cex=0.8,col="blue",pos=1)
abline(v=0,h=0)
plot(fml[,1],fml[,3],main="mlfaScores:f1&f3",xlab="f1",ylab="f3",xlim=lim,ylim=lim)
text(fml[,1],fml[,3],labels=rownames(fml),cex=0.8,col="blue",pos=1)
abline(v=0,h=0)
plot(fml[,2],fml[,3],main="mlfaScores:f2&f3",xlab="f2",ylab="f3",xlim=lim,ylim=lim)
text(fml[,2],fml[,3],labels=rownames(fml),cex=0.8,col="blue",pos=1)
abline(v=0,h=0)

#CA
#HCM
ds=dist(Z,method="euclidean")

single=hclust(ds,method="single")
plot(single,labels=rownames(Z),hang=-1,main="Single  Linkage")
complete=hclust(ds,method="complete")
plot(complete,labels=rownames(Z),hang=-1,main="complete  Linkage")
average=hclust(ds,method="average")
plot(average,labels=rownames(Z),hang=-1,main="average  Linkage")
ward=hclust(ds,method="ward.D")
plot(ward,labels=rownames(Z),hang=-1,main="ward  Linkage")

#mahalanobis를 이용한 Hcm
install.packages("biotools")
library(biotools)
dm=D2.dist(Z,cov(Z))
wardm=hclust(dm,method="ward.D2")
plot(wardm,labels=player,main="mahalanobis")

#NHcM
#먼저 군집수를 NbClust의 all index를 이용하여 정하고 실시한다
install.packages("NbClust")
library(NbClust)
allindex=NbClust(Z,distance = "euclidean",min.nc=2,max.nc = 8, method = "kmeans",index="all")
#2가지 분류
kmeans=kmeans(Z,2)
cluster=data.frame(player,cluster=kmeans$cluster)
C1=cluster[(cluster[,2]==1),]
C2=cluster[(cluster[,2]==2),]
C1;C2

#2가지 분류
library(cluster)
kmedoids<-pam(Z,2,metric="euclidean")
cluster=data.frame(player,cluster=kmedoids$cluster)
C1=cluster[(cluster[,2]==1),]
C2=cluster[(cluster[,2]==2),]
C1;C2
```