EPL Player Analysis: PCA and Clustering

222DSN39 홍은정

1. Data explanation

1. 데이터 출처: Kaggle (English Premier League Players Dataset, 2017/18)

2. 데이터 설명: 17/18 시즌 EPL 선수들의 시장 가치

3. 변수 설명

- name: 선수 이름
- club: 선수가 속한 클럽
- age: 선수의 나이
- position: 선수 플레이 위치
- position_cat: 선수 위치 범주화
  (1 공격수, 2 중앙미드필더, 3 수비수, 4 골키퍼)
- market_value: 선수 이적 시장 가치
- page_views: 평균 일일 위키백과 페이지 조회수
- fpl_value: Fantasy 프리미어 리그 선수 가치
- fpl_sel: Fantasy 프리미어 리그 선수의 백분율
- fpl_points: 이전 시즌 동안 Fantasy 프리미어 리그 포인트
- region:선수의 지역 범주화(1 잉글랜드, 2는 EU, 3 아메리카, 4 기타)
- nationality: 선수의 국적
- new_foreign: 신규 선수 여부
- age_cat: 선수의 연령 범주화
- club_id: 클럽 식별용
- big_club: TOP 6 속하는지 여부
- new_signing: 신규 선수로 등록되어있는지 여부

1. Data explanation

4. 변수 유형

```
# 데이터 불러오기
epl_Data <- read.csv("/Users/enjunghong/Downloads/epldata_final.csv")
#View(epl_Data)
rownames(epl_Data) <- epl_Data$name
epl_Data <- epl_Data[, -1, drop = FALSE]

# 데이터 살펴보기
str(epl_Data)
```

```
## 'data.frame':    461 obs. of  16 variables:
##  $ club         : chr  "Arsenal" "Arsenal" "Arsenal" "Arsenal" ...
##  $ age          : int  28 28 35 28 31 22 30 31 25 21 ...
##  $ position     : chr  "LW" "AM" "GK" "RW" ...
##  $ position_cat: int  1 1 4 1 3 3 1 3 3 1 ...
##  $ market_value: num  65 50 7 20 22 30 22 13 30 10 ...
##  $ page_views   : int  4329 4395 1529 2393 912 1675 2230 555 1877 1812 ...
##  $ fpl_value    : num  12 9.5 5.5 7.5 6 6 8.5 5.5 5.5 5.5 ...
##  $ fpl_sel      : chr  "17.10%" "5.60%" "5.90%" "1.50%" ...
##  $ fpl_points   : int  264 167 134 122 121 119 116 115 90 89 ...
##  $ region       : int  3 2 2 1 2 2 2 2 2 4 ...
##  $ nationality  : chr  "Chile" "Germany" "Czech Republic" "England" ...
##  $ new_foreign  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age_cat      : int  4 4 6 4 4 2 4 4 3 1 ...
##  $ club_id      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ big_club     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ new_signing  : int  0 0 0 0 0 0 0 0 1 0 ...
```

2. Methods & Results

1. 데이터 전처리

1) 변수 지정

```
# 선택에 제외할 변수들 지정
no_sel_var <- c("club", "position", "fpl_sel", "nationality","age_cat","club_id","new_for
eign","page_views","name")

selected_data <- epl_Data[, !(names(epl_Data) %in% no_sel_var)]
names(selected_data)
```

```
## [1] "age"          "position_cat" "market_value" "fpl_value"    "fpl_points"
## [6] "region"       "big_club"     "new_signing"
```

2) 결측치 제거

```
# 결측치 제거
selected_data <- na.omit(selected_data)
```

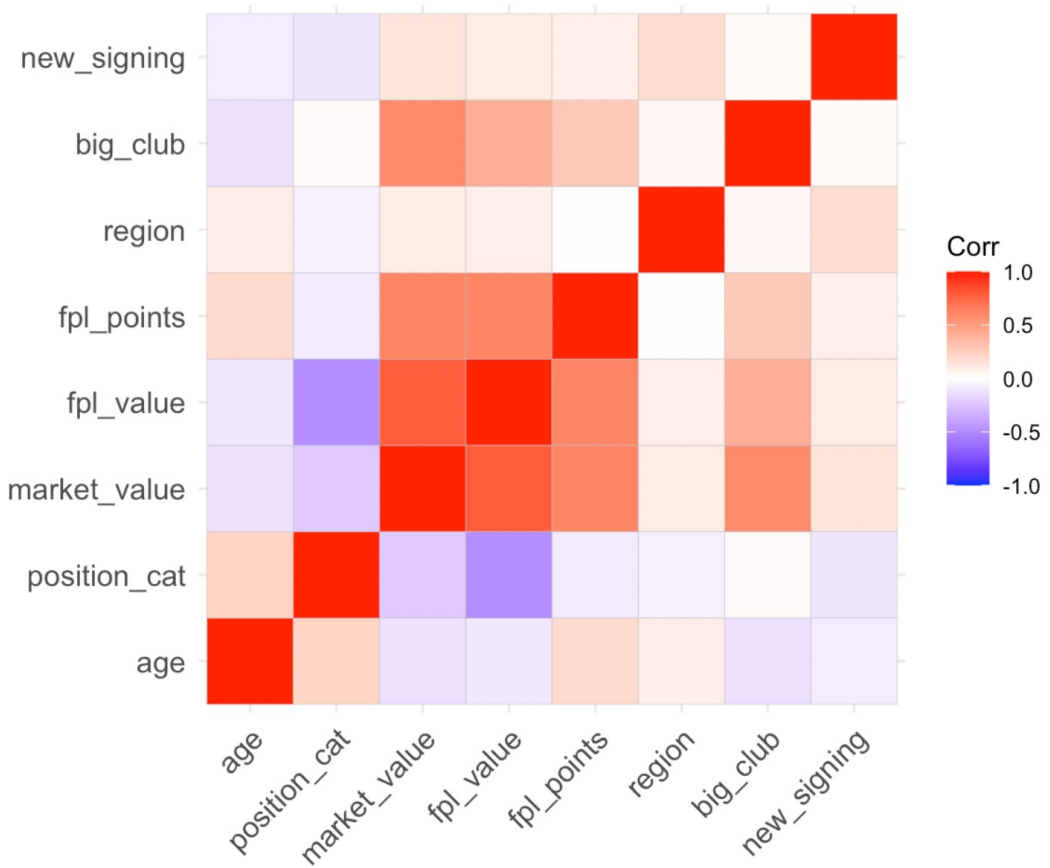
3) Scale 처리

```
standardized_data <- scale(selected_data)
```

2. Methods & Results

2. 상관관계 확인

```
cor_matrix <- cor(standardized_data)
ggcorrplot(cor_matrix)
```



2. Methods & Results

3. 데이터 모델링

1) PCA

* PCA(주성분 분석)이란?

- p차원의 다변량 데이터에 대해 분산-공분산 구조를 변수들의 선형결합식 m개를 설명하고자 하는 방법(m<<p)
- 주성분을 이용하면 정보의 손실을 최소화 하면서 저차원 공간상에서 데이터를 해석할 수 있게 됨
- 주성분은 서로 독립적인 새로운 변수로 또 다른 통계적 분석에 이용될 수 있음

* PCA의 목적

- 차원 축소, 변동이 큰 축 탐색, 주성분을 통한 데이터의 해석

```
pr.out = prcomp(standardized_data, scale=T)
names(pr.out)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
pr.out$scale
```

2. Methods & Results

* PCA 결과

```
pr.out$rotation
```

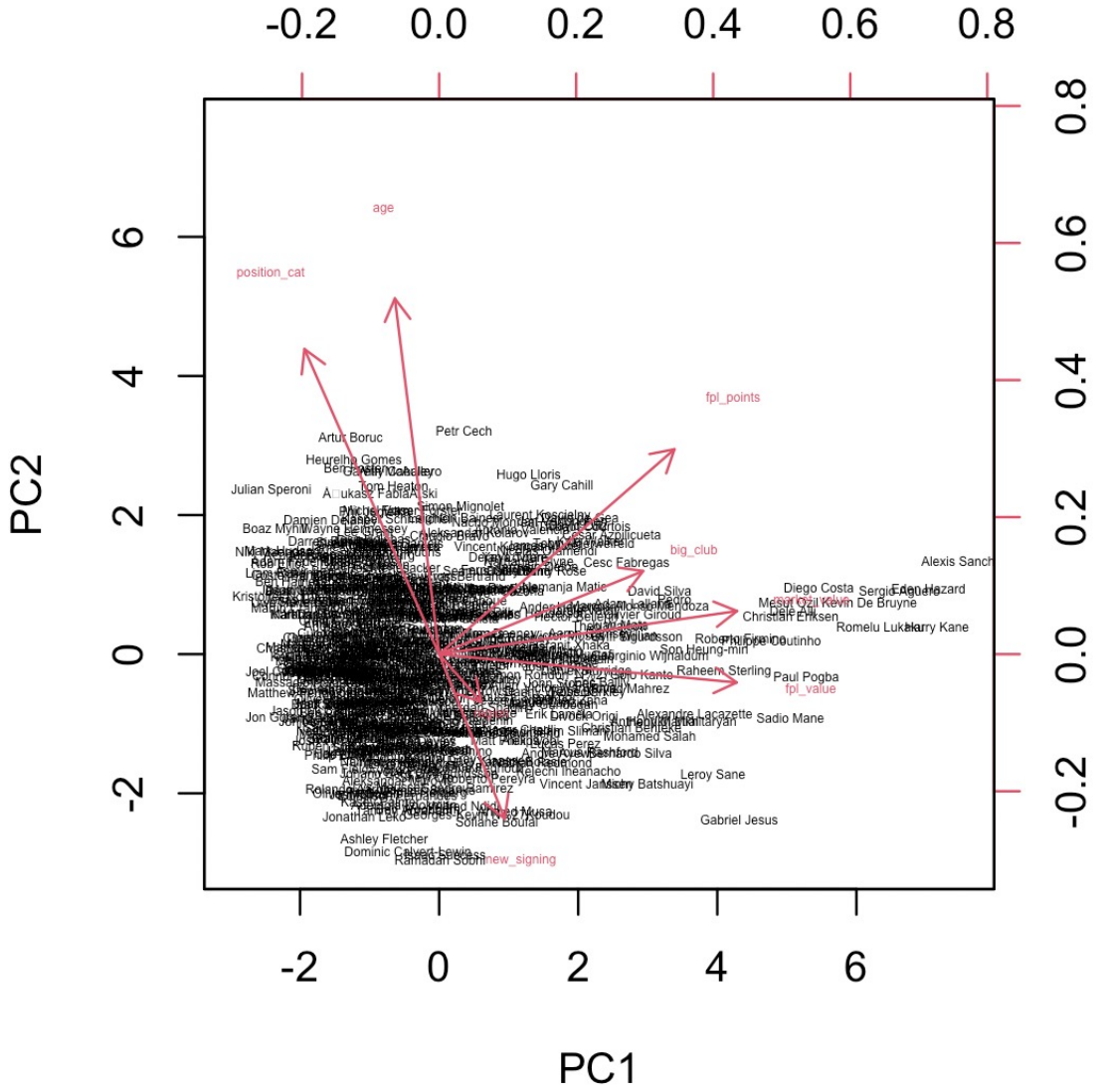
```
##                    PC1         PC2          PC3        PC4         PC5
## age          -0.08084312  0.64882372 -0.375175218 -0.3741868  0.02621856
## position_cat -0.24567470  0.55652575  0.009940599  0.5622762  0.13530351
## market_value  0.54348759  0.07873036  0.050753425  0.1326881 -0.02814093
## fpl_value     0.54280208 -0.05167855  0.032818151 -0.2247564 -0.05488560
## fpl_points    0.42937435  0.37384389 -0.052005202 -0.2304335  0.24792087
## region        0.07660308 -0.08718619 -0.721020545  0.1695439 -0.61696428
## big_club      0.37223816  0.15175345  0.186879615  0.5712766 -0.17138081
## new_signing   0.11959879 -0.30007987 -0.545887023  0.2598200  0.71114003
##                    PC6         PC7         PC8
## age          -0.51214381  0.16052355 -0.05452927
## position_cat  0.38788561  0.26802074  0.27031882
## market_value  0.11090382  0.53179493 -0.61833916
## fpl_value     0.01690694  0.34496316  0.72718975
## fpl_points    0.44604865 -0.59536606 -0.07203749
## region        0.21775916 -0.09789121  0.01637457
## big_club     -0.55321784 -0.36870920  0.07590004
## new_signing  -0.14912927  0.02453030  0.03984607
```

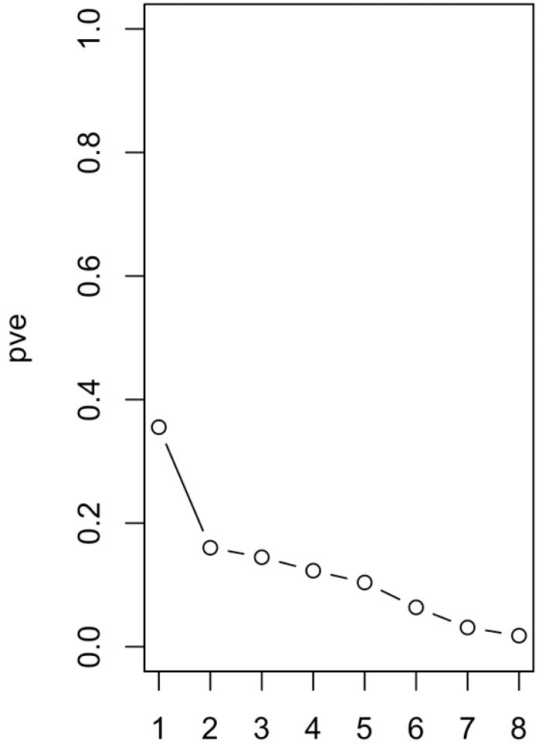
```
summary(pr.out)
```

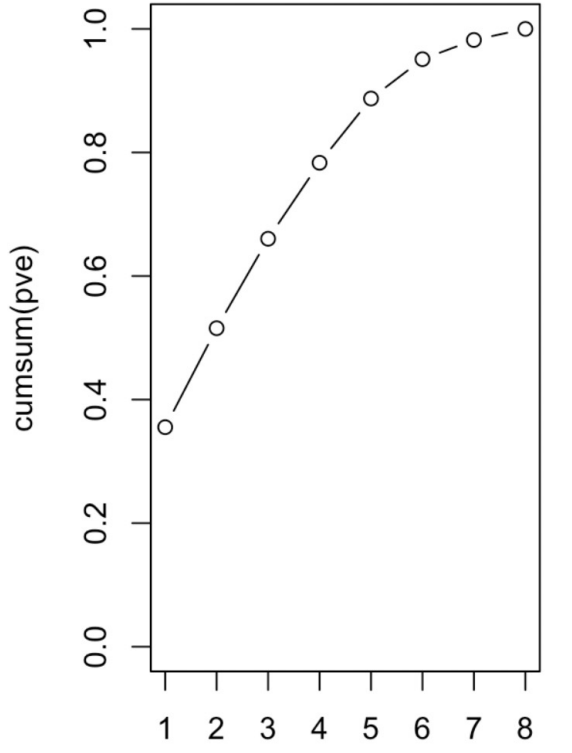
```
## Importance of components:
##                             PC1    PC2    PC3    PC4    PC5     PC6     PC7
## Standard deviation     1.6860 1.1321 1.0762 0.9920 0.9120 0.71382 0.49867
## Proportion of Variance 0.3553 0.1602 0.1448 0.1230 0.1040 0.06369 0.03108
## Cumulative Proportion  0.3553 0.5156 0.6603 0.7833 0.8873 0.95099 0.98208
##                             PC8
## Standard deviation     0.37866
## Proportion of Variance 0.01792
## Cumulative Proportion  1.00000
```

2. Methods & Results

* PCA 결과





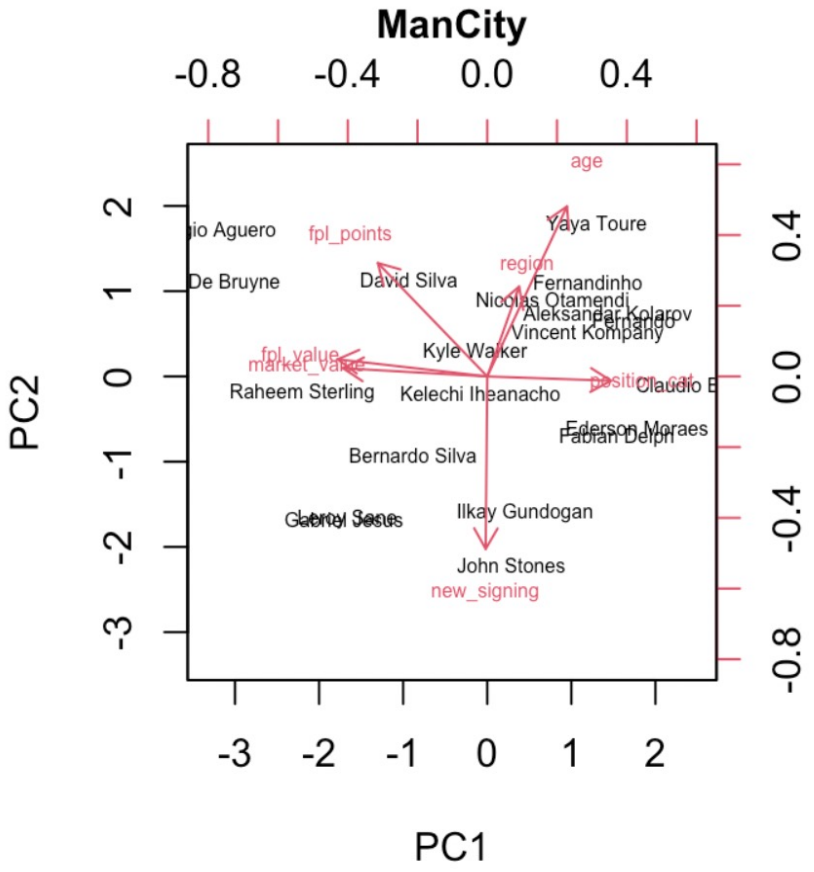


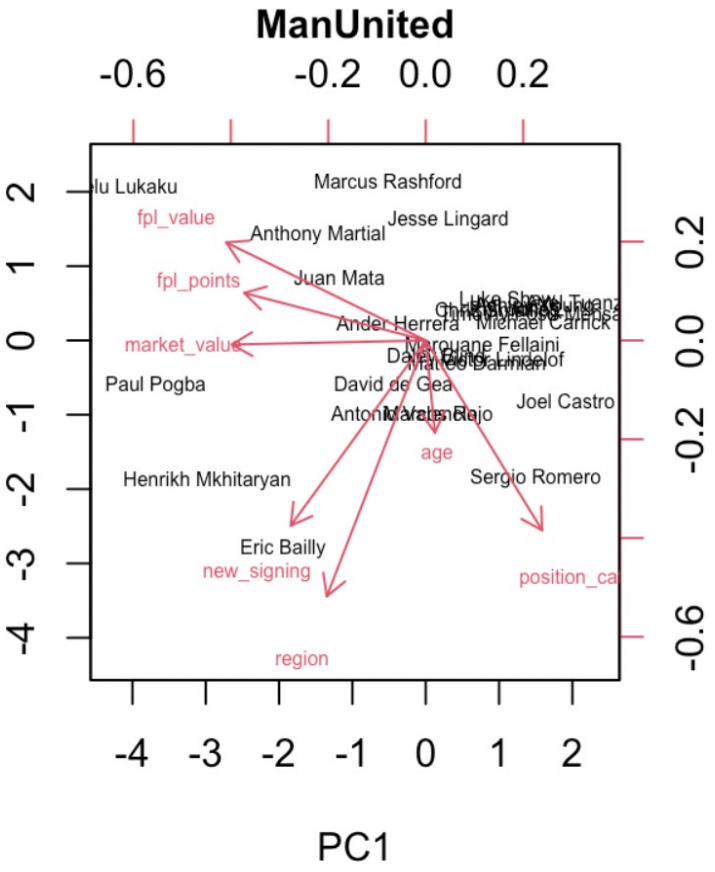
2. Methods & Results

* 17/18 시즌 상위 2팀, 하위 2팀 PCA 결과

- 맨시티, 맨유, 스토크시티, 웨스트브롬

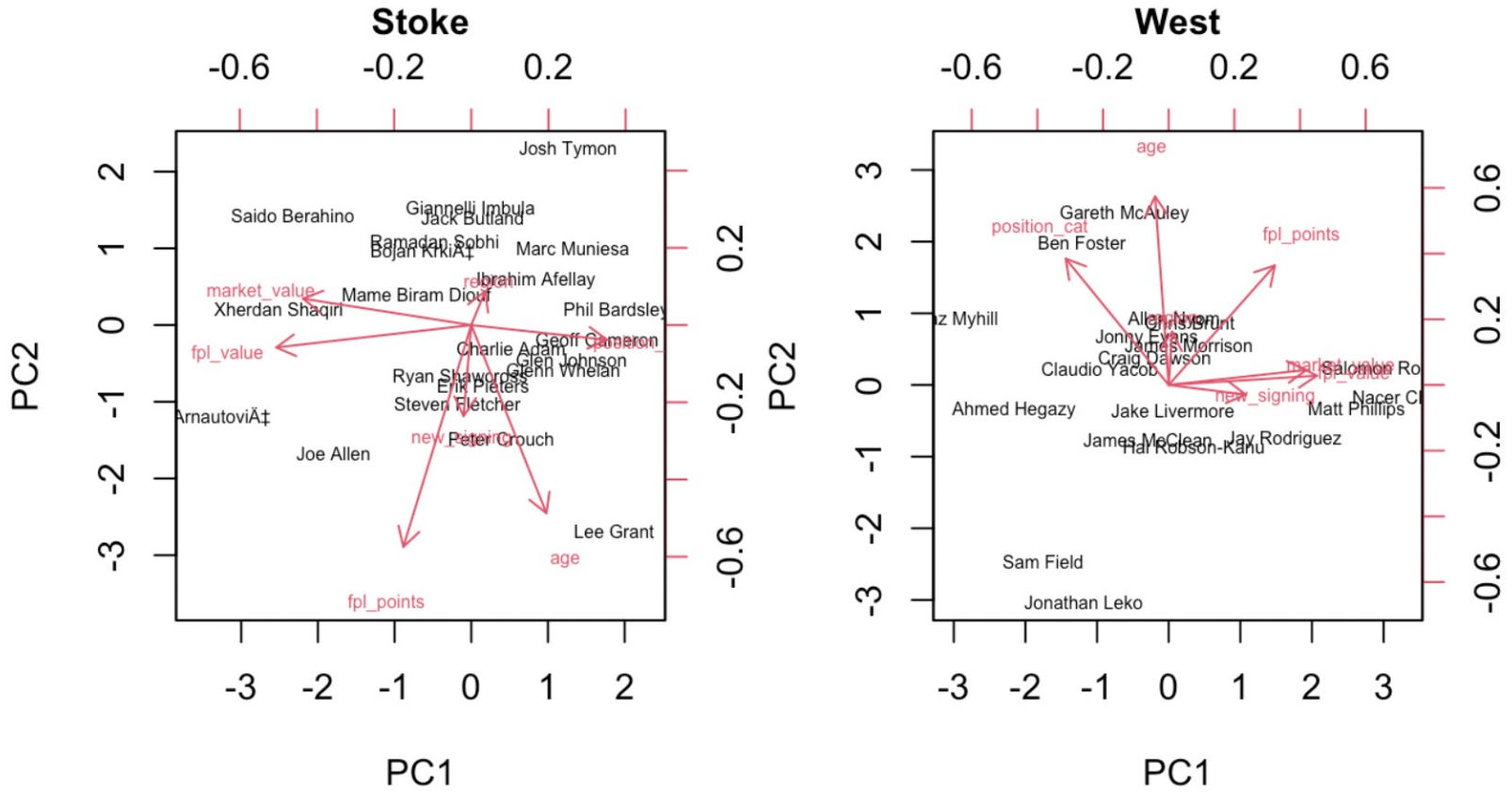
```
#unique(epl_Data$club)
selected_clubs <- c("Manchester+City", "Manchester+United", "Stoke+City","West+Brom")

ManCity <- epl_Data[epl_Data$club == selected_clubs[1],c("age","position_cat","market_value", "fpl_value", "fpl_points", "region","new_signing")]
ManUnited <- epl_Data[epl_Data$club == selected_clubs[2],c("age","position_cat","market_value", "fpl_value", "fpl_points", "region","new_signing")]
Stoke <- epl_Data[epl_Data$club == selected_clubs[3],c("age","position_cat","market_value", "fpl_value", "fpl_points", "region","new_signing")]
West <- epl_Data[epl_Data$club == selected_clubs[4],c("age","position_cat","market_value", "fpl_value", "fpl_points", "region","new_signing")]
```





2. Methods & Results

* 17/18 시즌 상위 2팀, 하위 2팀 PCA 결과



2. Methods & Results

2) K-Means Clustering

* K-means 클러스터링이란?

- 다변량 데이터를 k개의 클러스터로 그룹화하는 비지도 학습 알고리즘
- 각 클러스터는 중심(centroid)를 가지며, 각 데이터 포인트는 가장 가까운 클러스터 중심에 할당됨
- 클러스터 내의 데이터 포인트들은 서로 비슷한 특성을 공유하며, 클러스터 간의 유사성은 낮아야함

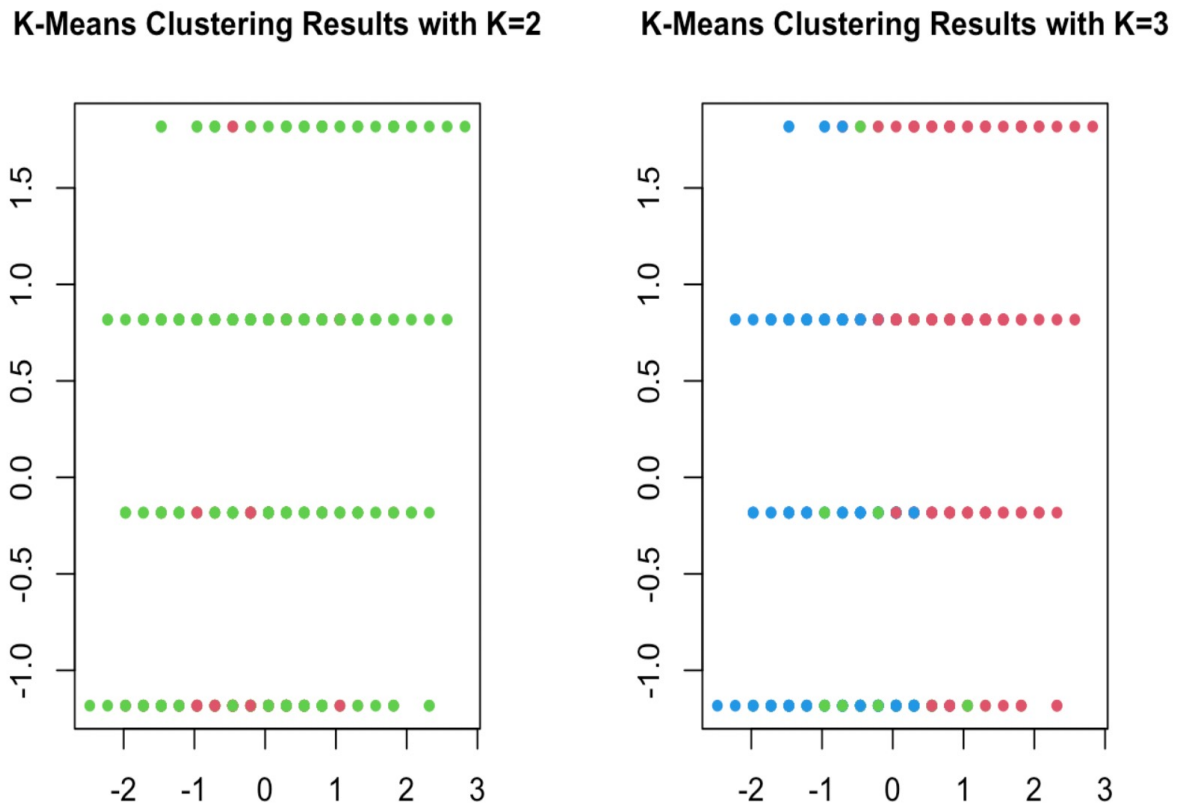
* K-means의 목적

- 데이터를 k개의 클러스터로 나눔으로 유사한 패턴이나 특징을 가진 그룹 형성
- 각 클러스터의 중심을 통해 데이터의 구조를 파악하고 각 데이터 포인트에 해당하는 클러스터로 할당

2. Methods & Results

* K-means 결과

```r
# K-means 클러스터링 수행
x=standardized_data
par(mfrow = c(1,2))
k <- 2  # 클러스터 수
kmeans_result <- kmeans(x, centers = k)
cluster_assignments <- kmeans_result$cluster

plot(x, col=(kmeans_result$cluster+1),
     main="K-Means Clustering Results with K=2",
     xlab="",ylab="", pch=20, cex=1, cex.main=1)

k <- 3  # 클러스터 수
kmeans_result <- kmeans(standardized_data, centers = k)
cluster_assignments <- kmeans_result$cluster

plot(x, col=(kmeans_result$cluster+1),
     main="K-Means Clustering Results with K=3",
     xlab="",ylab="", pch=20, cex=1, cex.main=1)
```



3. Conclusion

'EPL Player Analysis: PCA and Clustering'을 주제로하여 2017/2018 EPL 시즌의 선수들에 대한 종합적인 특성을 파악함

PCA를 활용하여 전체 선수의 다양한 특성을 차원 축소하고 이를 통해 선수의 위치, 연령, 시장 가치, 지역 등이 어떠한 영향을 미치는 지 확인 함

또한, 상위2개 클럽과 하위 2개 클럽을 대상으로 클럽 간의 선수 특성 차이를 살펴봄. 이를 통해 상위클럽과 하위 클럽 간의 어떠한 유의미한 차이가 있는 지 분석하고, 각 클럽 간의 전략적 특성을 도출함

K-means 클러스터링을 수행하여 선수들을 여러 그룹으로 나눔. 이를 통해 선수들의 군집을 확인하고 특성을 파악

이러한 다양한 분석을 통해 EPL 선수들의 특성과 클럽 간의 차이를 이해함