Practice 2

과제 2번

국민대학교 경영대학원 빅데이터경영 MBA
U2016054 이병준

Practice 2

Bulls.csv 는 경매시장에서 거래된 76마리의 어린(2살 이하) 황소의 특성과 거래가격(SalePr)에 관한 자료이다.변수 설명은 아래와 같다.

- Breed=1 if Angus, 5 if Hereford, 8 if Simental 소의 종류
- SalePr = Price of the bull
- FtFrBody=fat free body (pounds) 소의 무게
- Frame=Scale from 1(small) to 8 (large) 소의 크기별 등급 (1 ~ 8)
- SaleHt=Sale height at shoulder (inches)
- YrHgt=Yearling height at shoulder (inches)
- PrctFFB=Percent fat-free body
- BkFat=Back fat (inches)
- SaleWt=Sale weight (pounds) 판매 당시 무게

SalePr와 Breed 변수를 제외한 7개의 변수를 사용해 주성분분석을 시행하여 아래의 질문에 답하시오.
(공분산행렬 혹은 상관계수 행렬을 사용)

In [3]:
```
library(psych)
library(dplyr)
```

In [9]:
```
bulls <- read.csv("../0924_1001/bulls.csv", stringsAsFactors = F)
#bulls <- read.csv("../../KMU/Second Semester/Statistics/0924_1001/bulls.csv", stringsAsFactors = F)
#bulls$Breed <- as.factor(bulls$Breed)
#bulls$Frame <- as.factor(bulls$Frame)
str(bulls)
```

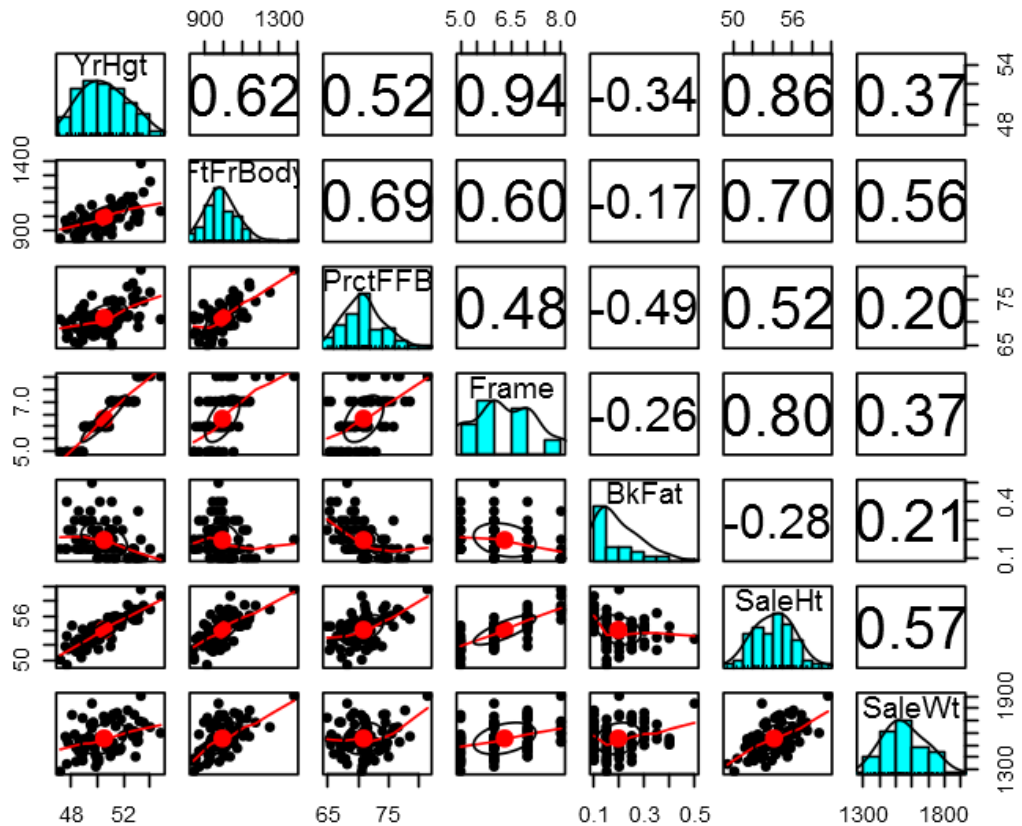
```
'data.frame':   76 obs. of  9 variables:
 $ Breed   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ SalePr  : int  2200 2250 1625 4600 2150 1225 2250 4000 1600 1525 ...
 $ YrHgt   : num  51 51.9 49.9 53.1 51.2 49.2 51 51.5 50.1 49.6 ...
 $ FtFrBody: int  1128 1108 1011 993 996 985 959 1060 979 1083 ...
 $ PrctFFB : num  70.9 72.1 71.6 68.9 68.6 71.4 72.1 69.3 71.2 75.8 ...
 $ Frame   : int  7 7 6 8 7 6 7 7 6 6 ...
 $ BkFat   : num  0.25 0.25 0.15 0.35 0.25 0.15 0.2 0.3 0.25 0.3 ...
 $ SaleHt  : num  54.8 55.3 53.1 56.4 55 51.4 54 55.6 51.5 54.6 ...
 $ SaleWt  : int  1720 1575 1410 1595 1488 1500 1522 1765 1365 1640 ...
```

- SalePr와 Breed 변수를 제외
  - breed 종 => 제외 필요.
  - Salespr 가격 => 제외 필요

In [9]:
```
head(bulls,3)
```

	Breed	SalePr	YrHgt	FtFrBody	PrctFFB	Frame	BkFat	SaleHt	SaleWt
1	1.00	2200.00	51.00	1128.00	70.90	7.00	0.25	54.80	1720.00
2	1.00	2250.00	51.90	1108.00	72.10	7.00	0.25	55.30	1575.00
3	1.00	1625.00	49.90	1011.00	71.60	6.00	0.15	53.10	1410.00

In [10]:
```
options(repr.plot.width = 6, repr.plot.height = 5)
pairs.panels(bulls[,-c(1,2)])
```



- BkFat 과 같은 경우 오른쪽 꼬리 그래프가 나온다. 한군대에 뭉쳐져 있다. => 분산시키기 위해 log 변환

In [5]:
```
shapiro.test(bulls$BkFat) # 엄청 작은 값이 나온다 P-value가 수치 변환이 필요하다. (귀무가설 기각)
```

```
        Shapiro-Wilk normality test

data:  bulls$BkFat
W = 0.87777, p-value = 2.616e-06
```

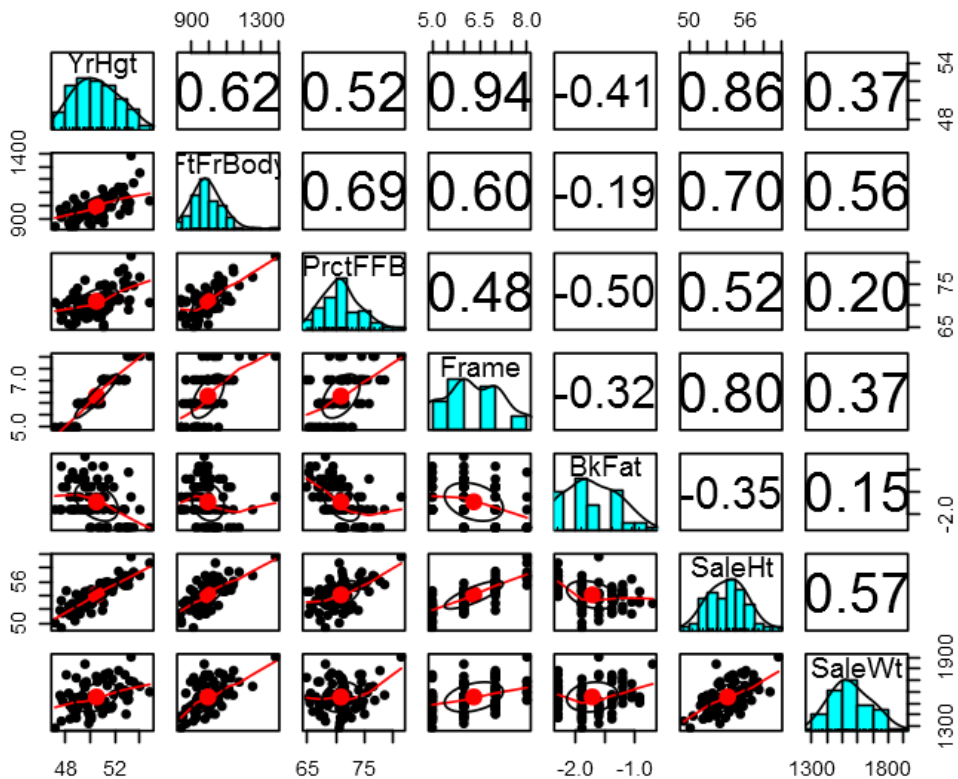
In [6]:
```
shapiro.test(log(bulls$BkFat)) # 그나마 값이 커진다.비록 0.05를 넘지 못하여 기각이지만 값들의 분포가 넓어졌다.
```

```
        Shapiro-Wilk normality test

data:  log(bulls$BkFat)
W = 0.92033, p-value = 0.0001472
```

In [11]:
```
bulls$BkFat <- log(bulls$BkFat)
```

In [12]:
```
pairs.panels(bulls[,-c(1,2)])
```



1. 주성분을 계산하는데 사용된 고유값과 고유벡터를 찾으시오.

- 각 항목당 나타내는 수치가 달라 표준화 작업이 필요할 것으로 판단 된다.

In [13]:
```
bulls_pca <- prcomp(bulls[,-c(1,2)], scale. = TRUE)
```

In [14]:
```
summary(bulls_pca)
```

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5    PC6     PC7
Standard deviation     2.0412 1.1336 0.8598 0.66207 0.42736 0.3760 0.21676
Proportion of Variance 0.5952 0.1836 0.1056 0.06262 0.02609 0.0202 0.00671
Cumulative Proportion  0.5952 0.7788 0.8844 0.94700 0.97309 0.9933 1.00000
```

1) 고유값

- 각각의 $y_1, \cdots, y_q$ 값의 분산이 고유값 이다.

In [16]:
```
sum(bulls_pca$sdev^2) # 총 분산의 합은 변수가 7개 이므로 7이 된다.
```

```
7
```

In [17]:
```
bulls_pca$sdev^2
```

```
4.16639401251579  1.28506492080151  0.739171981514117  0.438341375675231  0.182640831774896  0.141402023509068
0.0469848542093837
```

In [12]:
```
# 1. 방법
data.frame("고유값"=bulls_pca$sdev^2, row.names =c("PC1","PC2","PC3","PC4","PC5","PC6","PC7"))
```

	고유값
PC1	4.166394
PC2	1.285065
PC3	0.739172
PC4	0.4383414
PC5	0.1826408
PC6	0.141402
PC7	0.04698485

In [13]:
```
# 2. 고유 성질을 이용한 고유값 추출
apply(bulls_pca$x,2,var)
```

PC1	4.1663940125158
PC2	1.28506492080151
PC3	0.739171981514117
PC4	0.438341375675231
PC5	0.182640831774896
PC6	0.141402023509068
PC7	0.0469848542093837

2) 고유벡터

In [14]:
```
bulls_pca$rotation
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
YrHgt	-0.44820031	-0.05327032	-0.40235898	0.14962541	-0.07710727	-0.06107349	-0.77609454
FtFrBody	-0.40732427	0.16811879	0.45040383	0.24122005	0.69548050	-0.24680062	-0.01298538
PrctFFB	-0.3538218519	-0.2926851230	0.6107200980	0.2648702191	-0.5603882288	0.1788157016	0.0004718507
Frame	-0.4315783585	-0.0007095689	-0.4391195154	0.2849998420	-0.1982800989	-0.4004534412	0.5831044521
BkFat	0.21864508	0.69376718	-0.02332808	0.62123342	-0.15522909	0.24131212	-0.04559227
SaleHt	-0.45087936	0.09866641	-0.18622211	-0.20153998	0.18709720	0.78962316	0.23057670
SaleWt	-0.26581153	0.62624699	0.18497509	-0.58437544	-0.31290565	-0.24755848	-0.04746948

2. 적절한 주성분의 개수를 선택하고 근거를 설명하시오.

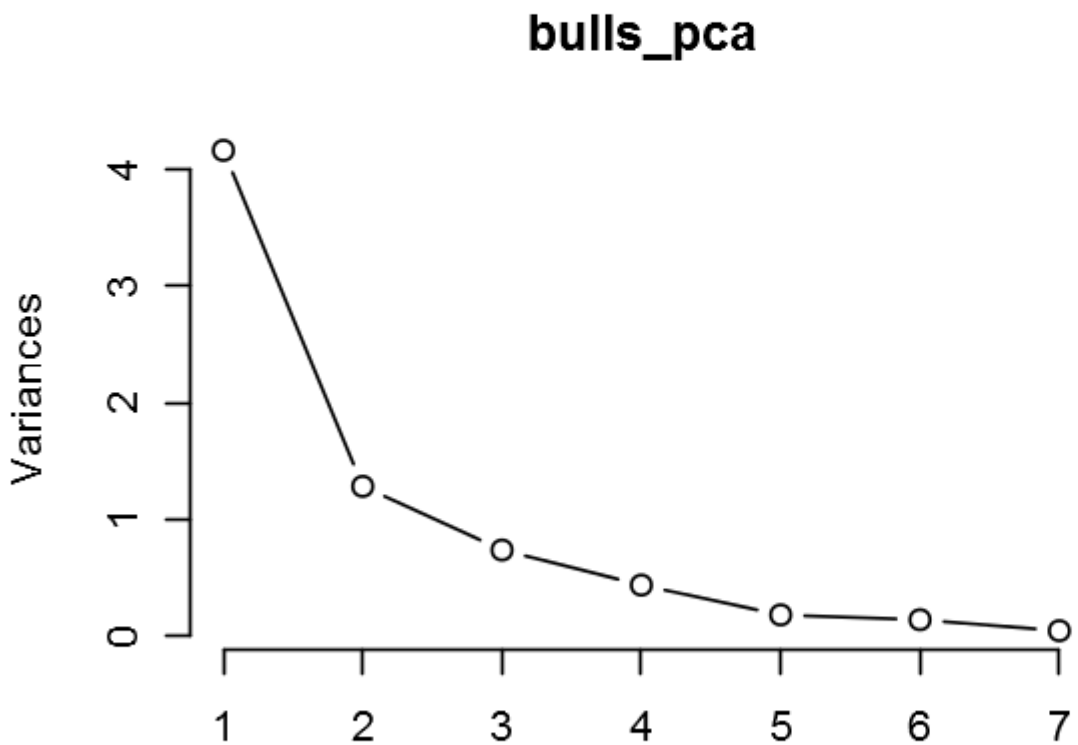
- 주성분 개수를 결정하는 것에 있어서 누적 분상의 합이 70~90%를 설명한다면 선택하는 것에 좋다고 판단됩니다.
- 그리고 Plot 의 Line차트를 이용하여 그래프의 기울기를 활용하여 설명도의 정도가 줄어드는 시점을 선택하면 될것으로 생각됩니다.

In [18]:
```
summary(bulls_pca) # 현재는 PC2~PC3에서 선택하면 될 것으로 판단됩니다.
```

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5    PC6     PC7
Standard deviation     2.0412 1.1336 0.8598 0.66207 0.42736 0.3760 0.21676
Proportion of Variance 0.5952 0.1836 0.1056 0.06262 0.02609 0.0202 0.00671
Cumulative Proportion  0.5952 0.7788 0.8844 0.94700 0.97309 0.9933 1.00000
```

- 그래프의 기울기로 판단하여 $PC2$ 이후로 기울기의 변화가 적은 것으로 생각되어 2개 즉, $PC2$까지 선택하는 것으로 판단.
- 그리고 $PC2$에서 $PC3$로 넘어갈때 10프로 정도 밖에 설명을 하지 못하므로 $PC2$까지 선택함.
- 대략 $PC1$ $PC2$ 합이 80% 됨.

In [19]:
```
options(repr.plot.width = 5, repr.plot.height=4)
plot(bulls_pca,type="l")
```



3.각 주성분의 rotation값을 표와 그래프를 사용해 비교하고 주성분의 의미를 해석하시오.

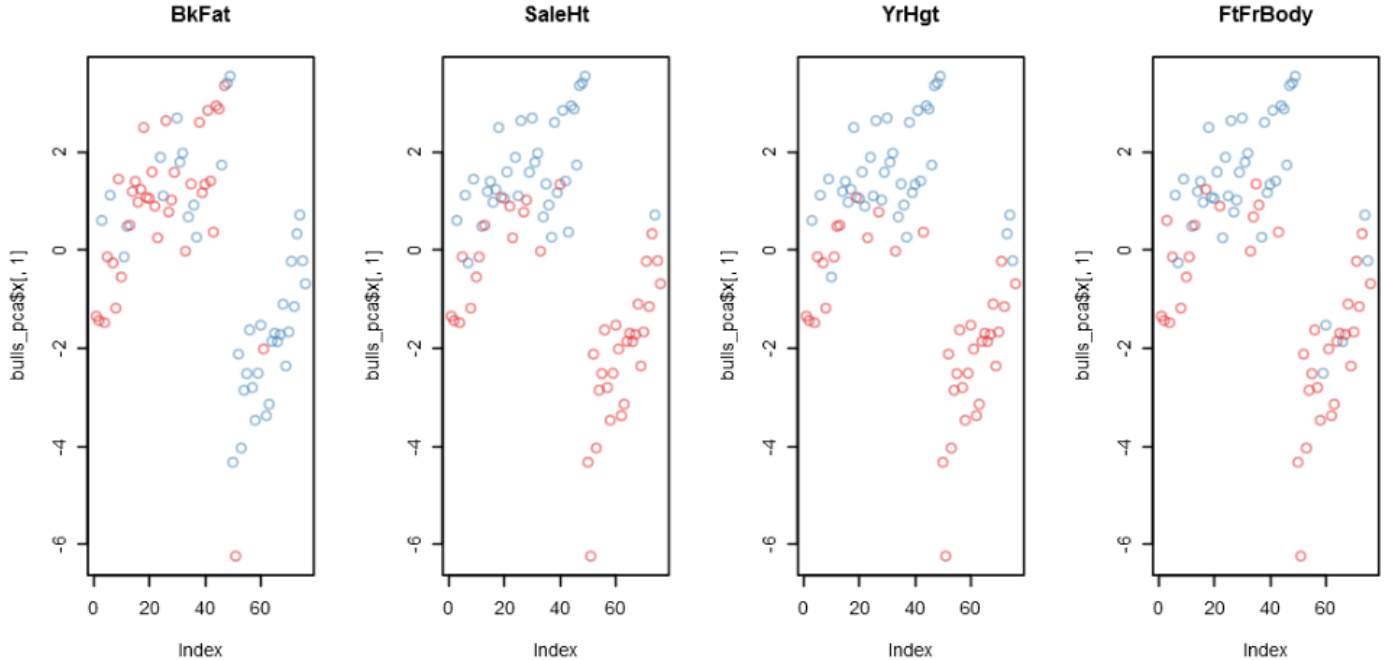
- 2개의 주성분을 이용하여 분석하려고 했으므로 PC1,PC2를 사용하겠습니다.

In [91]:
```
bulls_pca$rotation[,1:2]
```

	PC1	PC2
YrHgt	-0.44820031	-0.05327032
FtFrBody	-0.4073243	0.1681188
PrctFFB	-0.3538219	-0.2926851
Frame	-0.4315783585	-0.0007095689
BkFat	0.2186451	0.6937672
SaleHt	-0.45087936	0.09866641
SaleWt	-0.2658115	0.6262470

- $PC1$의 기준은 $BkFat$이 높을 수록 PC1의 값이 높고 나머지 6개 항목들이 높을수록 $PC1$의 $Y1$의 값이 낮게 나오는 모양을 취하고 있습니다.
- 즉, 축산업에 있어서 등의 지방 수치와 소의 등급에 미치는 영향을 절실하게 보여주는 데이터가 된다.
- $PC1$의 값이 낮을 수록 좋은 품질이라고 생각 됩니다.
- 축산업에서도 정육율 추정식에서 보정 등지방두께의 계수가 감산요인으로 정육등급의 하락에 기여도가 가장 크다. 라는 논문 내용도 발견하였다.

In [87]:
```
# 빨간색이 1 푸른색이 2
options(repr.plot.width=9, repr.plot.height=9)
par(mfrow=c(2,4))
tmp <- bulls
tmp$shape <- ifelse(tmp$BkFat > -1.8,1,2)
plot(bulls_pca$x[,1],col=tmp$shape, main="BkFat")
tmp$shape <- ifelse(tmp$SaleHt > 54,1,2)
plot(bulls_pca$x[,1],col=tmp$shape, main="SaleHt")
tmp$shape <- ifelse(tmp$YrHgt > 50.35,1,2)
plot(bulls_pca$x[,1],col=tmp$shape, main="YrHgt")
tmp$shape <- ifelse(tmp$FtFrBody > 990.5,1,2)
plot(bulls_pca$x[,1],col=tmp$shape, main="FtFrBody")
tmp$shape <- ifelse(tmp$PrctFFB > 70.85,1,2)
plot(bulls_pca$x[,1],col=tmp$shape, main="PrctFFB")
tmp$shape <- ifelse(tmp$Frame > 6,1,2)
plot(bulls_pca$x[,1],col=tmp$shape, main="Frame")
tmp$shape <- ifelse(tmp$SaleWt > 1538,1,2)
plot(bulls_pca$x[,1],col=tmp$shape, main="SaleWt")
```

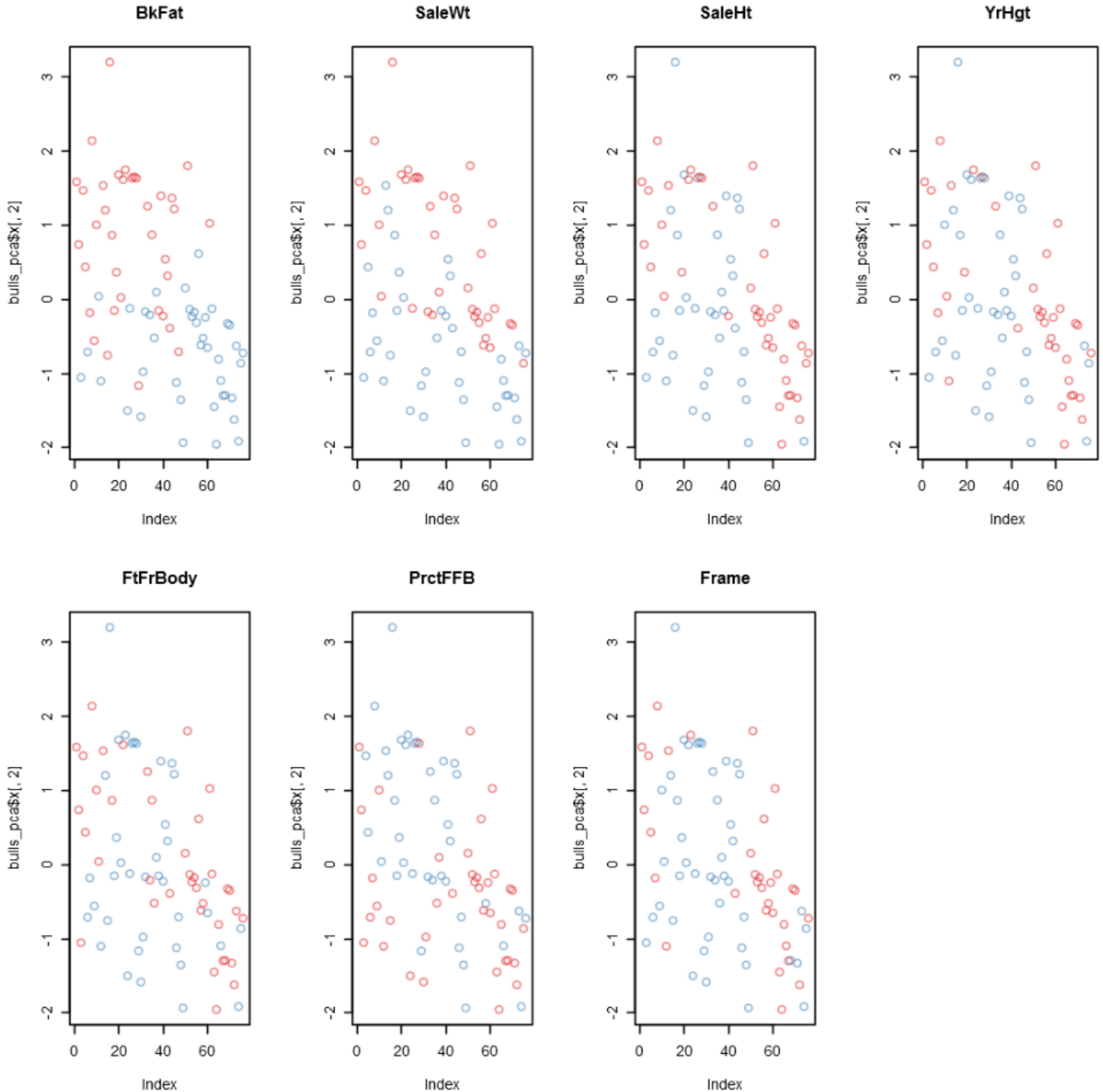

BkFat
SaleHt
YrHgt
FtFrBody
PrctFFB
Frame
SaleWt
bulls_pca$x[, 1]
Index

In [89]:
```
bulls_pca$rotation[,2]
```

YrHgt	-0.0532703215445766
FtFrBody	0.168118785977307
PrctFFB	-0.292685122976146
Frame	-0.000709568911840028
BkFat	0.69376718489631
SaleHt	0.0986664107513893
SaleWt	0.626246991605501

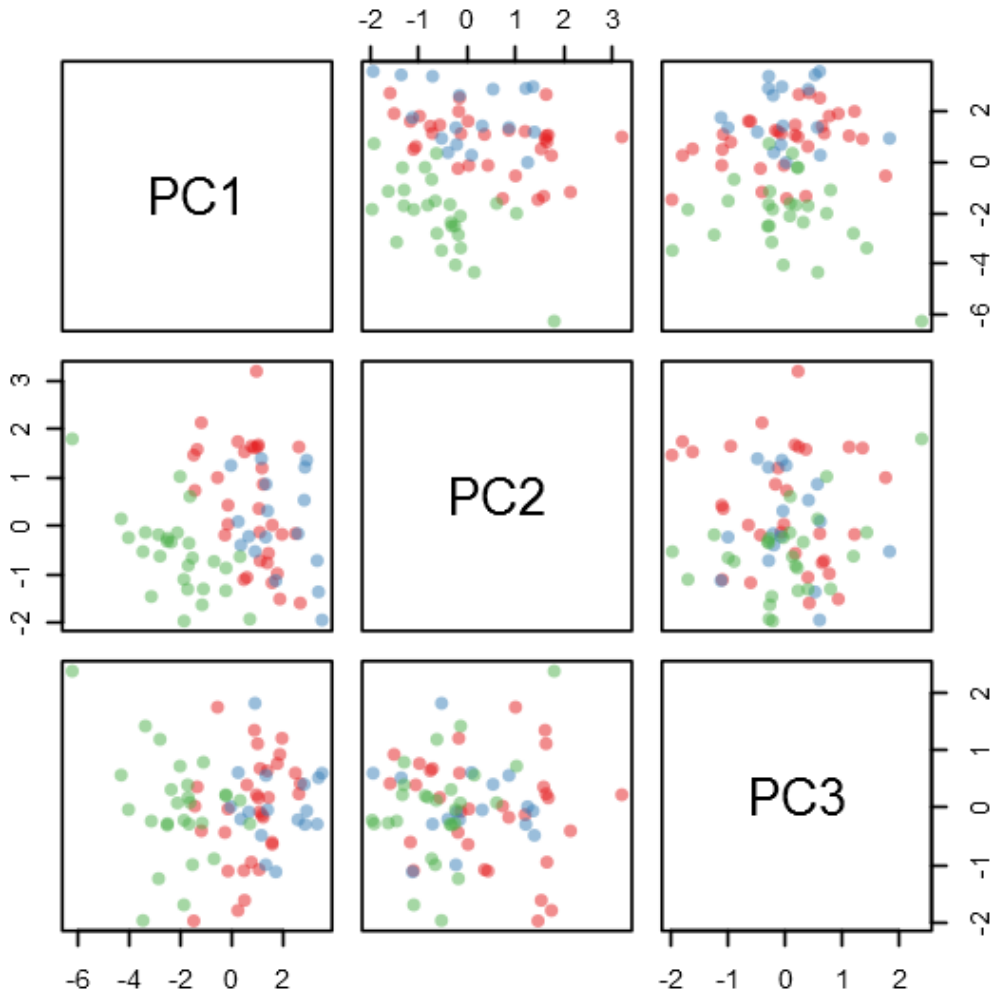
- $PC2$를 보면 알 수 있듯이 $BkFat$과 $SaleWt$을 제외한 나머지 부분에 있어서는 영향도가 적다. 그래프에서도 판단할 수 있는 부분이 크게 있지 않다고 판단됩니다.
- $PC2$에서 $BkFat$이 클수록 $PC2$의 값이 크고, $SaleWt$ 또한 값이 클수록 $PC2$의 값이 크게 보입니다.

In [90]:
```
par(mfrow=c(2,4))
tmp <- bulls
tmp$shape <- ifelse(tmp$BkFat > -1.8,1,2)
plot(bulls_pca$x[,2],col=tmp$shape, main="BkFat")
tmp$shape <- ifelse(tmp$SaleWt > 1538,1,2)
plot(bulls_pca$x[,2],col=tmp$shape, main="SaleWt")
tmp$shape <- ifelse(tmp$SaleHt > 54,1,2)
plot(bulls_pca$x[,2],col=tmp$shape, main="SaleHt")
tmp$shape <- ifelse(tmp$YrHgt > 50.35,1,2)
plot(bulls_pca$x[,2],col=tmp$shape, main="YrHgt")
tmp$shape <- ifelse(tmp$FtFrBody > 990.5,1,2)
plot(bulls_pca$x[,2],col=tmp$shape, main="FtFrBody")
tmp$shape <- ifelse(tmp$PrctFFB > 70.85,1,2)
plot(bulls_pca$x[,2],col=tmp$shape, main="PrctFFB")
tmp$shape <- ifelse(tmp$Frame > 6,1,2)
plot(bulls_pca$x[,2],col=tmp$shape, main="Frame")
```


BkFat
SaleWt
SaleHt
YrHgt
FtFrBody
PrctFFB
Frame
bulls_pca$x[, 2]
Index

In [45]:
```
comp <- data.frame(bulls_pca$x[,1:3])
```

In [52]:
```
palette(alpha(brewer.pal(9,'Set1'), 0.5))
DF <- bulls
DF$Breed <- as.factor(DF$Breed)
plot(comp, col=DF$Breed, pch=16)
```



4.행렬도를 사용해 원변수와 주성분의 관계, 원변수 간의 상관관계,특이한 관측치의 존재 유무 등을 파악하고 설명하시오.

※ 행렬도의 해석은 이미 3번에서 각 Rotation의 성분에 따라 무엇이 영향도가 있는지 되었는지 근거를 다시한번 생각하게 되는 부분으로 생각됩니다.

PC1

- $PC1$와 $YrHgt, Frame, FtFrBody, SaleHt$는 아주 높은 관계(음)를 가지고 있다.
- $PC1$과 $PrctFFB, SaleWt$는 위 보다는 적은 관계(음)를 가지고 있지만 여전히 관계를 가지고 있다.
- 유일하게 $PC1$와 $BkFat$는 양의 관계를 가지고 있지만 관계정도는 제일 적다.=> 가장 $PC1$과 수직방향이다.

PC2

- $PC2$와 $SaleWt, BkFat$은 높은 관계도를 가지고 있고, $PrctFFB$와는 약하지만 관계를 가지고 있다.
  하지만 나머지 부분에 있어서는 관계가 거의 없다. (수직방향)
- $BkFat$가 가장 높은 관계(양)을 가지고 있다. 가작 수평 방향이다.

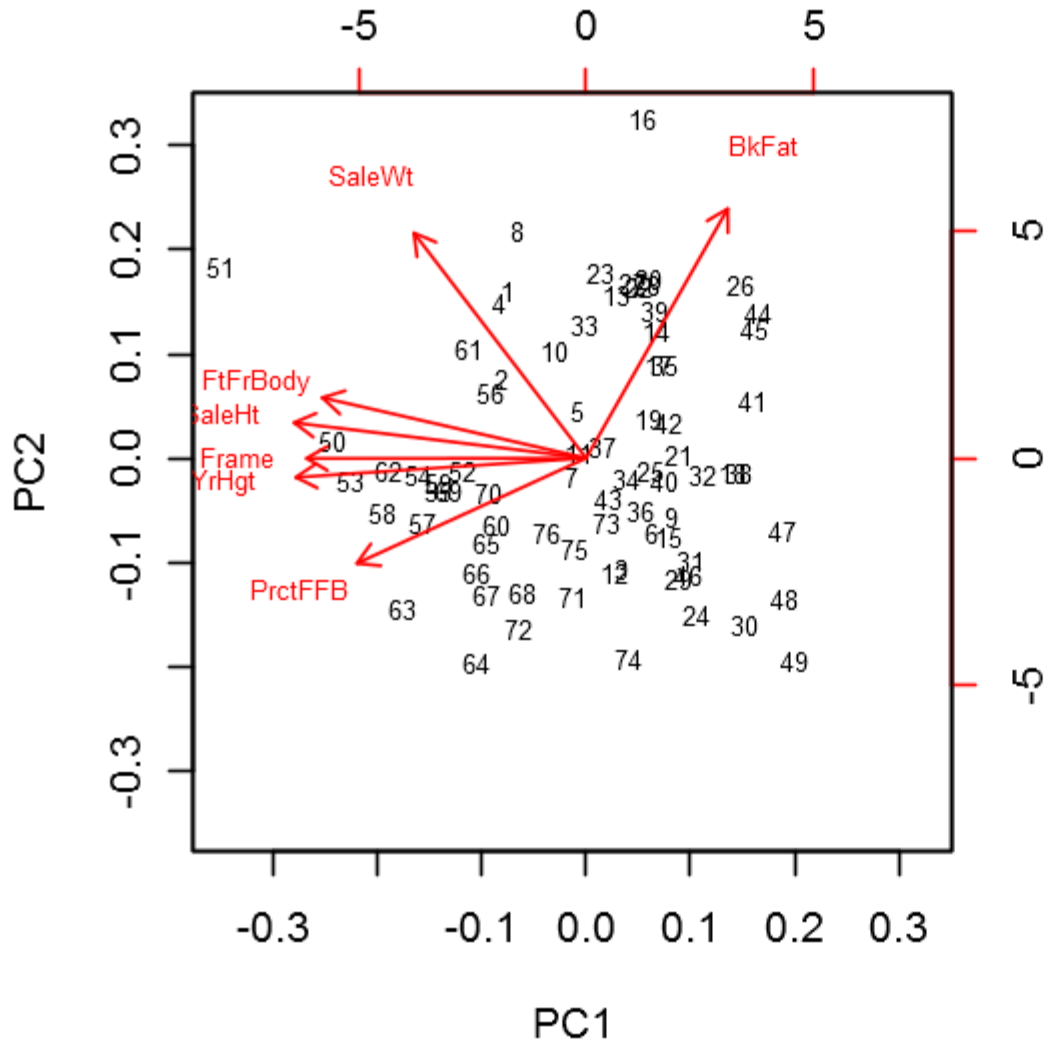
원변수간의 상관관계

- $YrHgt, Frame, FtFrBody, SaleHt$ 와 $SaleWt, BkFat$는 서로 무관한 수치라고 볼 수 있다. $PC1$와 $PC2$에 각각 서로 가장 큰 영향도를 주고 있는데, $PC1$과 $PC2$의 상관관계는 거의 0 이므로 두 그룹의 관계도는 아주 적다고 볼 수 있다.

이상치

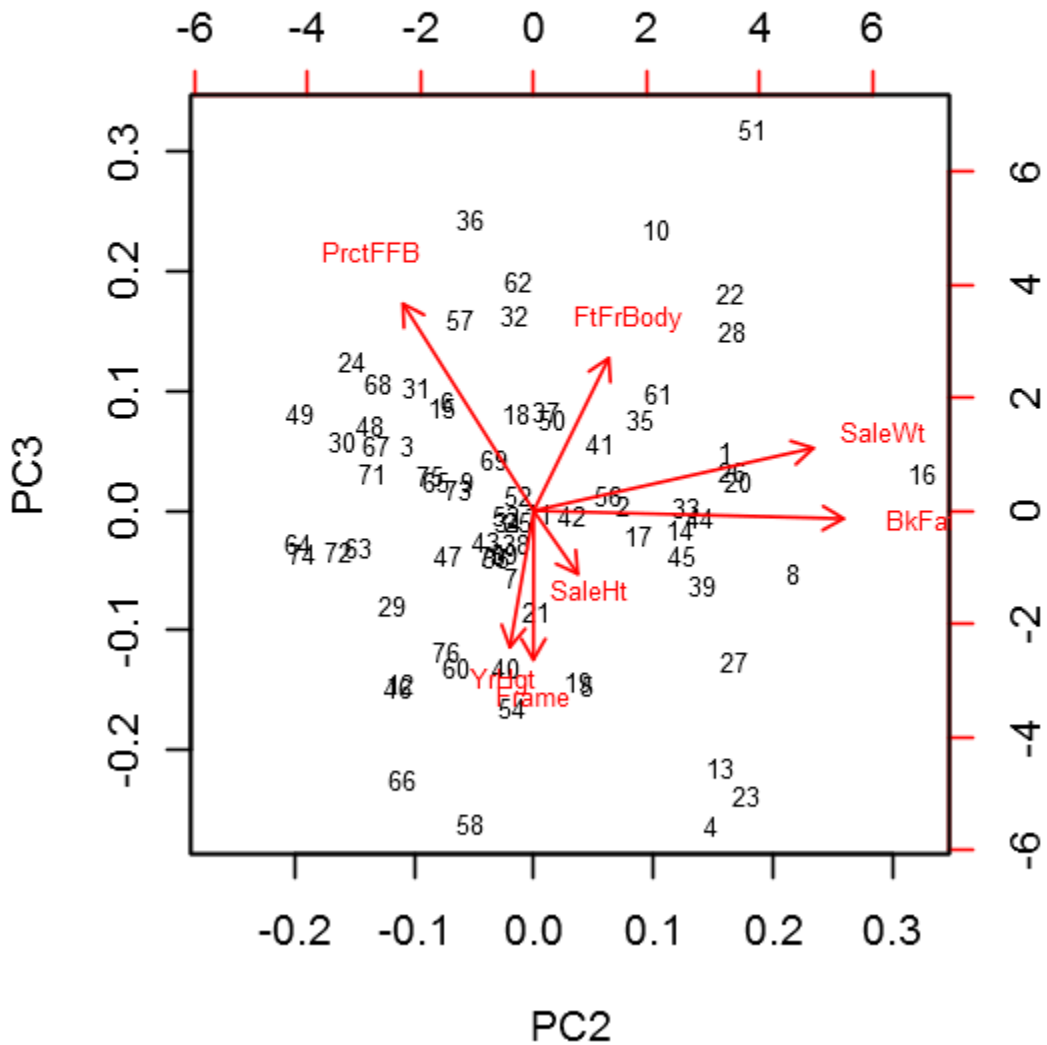
- 16, 51번 소가 가장 이상한 수치를 가지고 있다.
  - 51번은 $PC1$에서 가장 좋은 소로 판단된다. $PC1$이 낮을 수록 좋다.
  - 16번은 $PC2$에서 가장 좋지 못한 소로 판단된다. $PC2$은 높을 수록 좋지 않다.

In [18]:
```
options(repr.plot.width=5, repr.plot.height=5)
biplot(bulls_pca,cex=0.6)
```



PC3

In [19]:
```
biplot(bulls_pca,cex=0.6, c(2,3))
```



5. 첫 두개의 주성분을 사용해 산점도를 그리고 Breed를 서로 다른 색깔과 기호로 표시하시오. 주성분에 의해 다른 종의 황소를 구분할 수 있는가? 이상점이 있는가?있다면 어떤 특성을 가진 소인가?

In [14]:
```
library(RColorBrewer)
library(ggthemes)
library(ggplot2)
library(plotly)
```

In [15]:
```
DF <- data.frame(PC1 = bulls_pca$x[,1], PC2= bulls_pca$x[,2], Breed = as.numeric(bulls$Breed))
DF$Breed <- as.factor(DF$Breed)
head(DF)
```

	PC1	PC2	Breed
1	-1.34451290535535	1.58582898784658	1
2	-1.43500873306372	0.739540097487124	1
3	0.603381332348864	-1.05228877444158	1
4	-1.47819006545351	1.46961922474471	1
5	-0.137352258992633	0.436773771006885	1
6	1.11855495360577	-0.709489524084782	1

In [22]:
```
unique(DF$Breed)
```

```
1 5 8
```

- $PC1$ 값에 의하여 $Breed$ $1, 5$와 8이 대체적으로 구분이 갈 수 있으며 $PC2$에 의해서는 구분이 가기 힘든 것으로 판단됩니다.
- 이상점
  - 8번 종류의 소 중 $PC1$값이 -6 이하로 있는 소
  - 1번 종류의 소 중 $PC2$값이 4에 근접하는 소
  - 총 2개정도로 생각 할 수 있을 것 같습니다.

In [24]:
```
summary(DF)
```

```
      PC1                PC2               Breed
 Min.   :-6.2313   Min.   :-1.9535   1:32
 1st Qu.:-1.5529   1st Qu.:-0.7664   5:17
 Median : 0.4259   Median :-0.1769   8:27
 Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 1.3597   3rd Qu.: 0.9047
 Max.   : 3.5407   Max.   : 3.1998
```

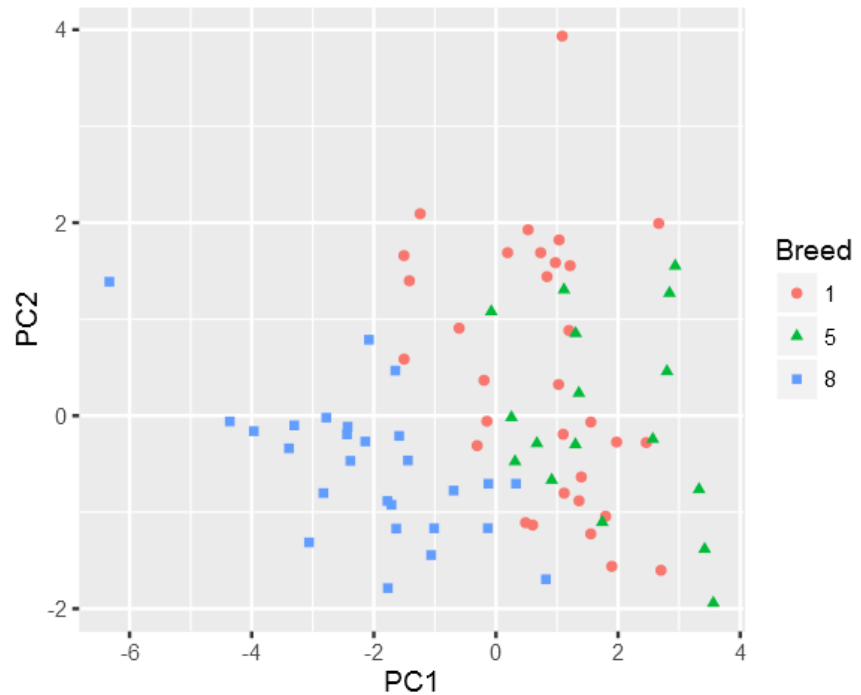
- SaleWt 가 큰 편에 속하고.

In [32]:
```
bulls[rownames(DF[DF$PC1 < -6,]),]
bulls[rownames(DF[DF$PC2 > 3.0,]),]
```

	Breed	SalePr	YrHgt	FtFrBody	PrctFFB	Frame	BkFat	SaleHt	SaleWt
51	8.000000	1450.000000	53.300000	1383.000000	81.400000	8.000000	-1.609438	59.600000	1904.000000

	Breed	SalePr	YrHgt	FtFrBody	PrctFFB	Frame	BkFat	SaleHt	SaleWt
16	1.0000000	2300.0000000	49.6000000	975.0000000	68.2000000	6.0000000	-0.6931472	52.9000000	1842.0000000

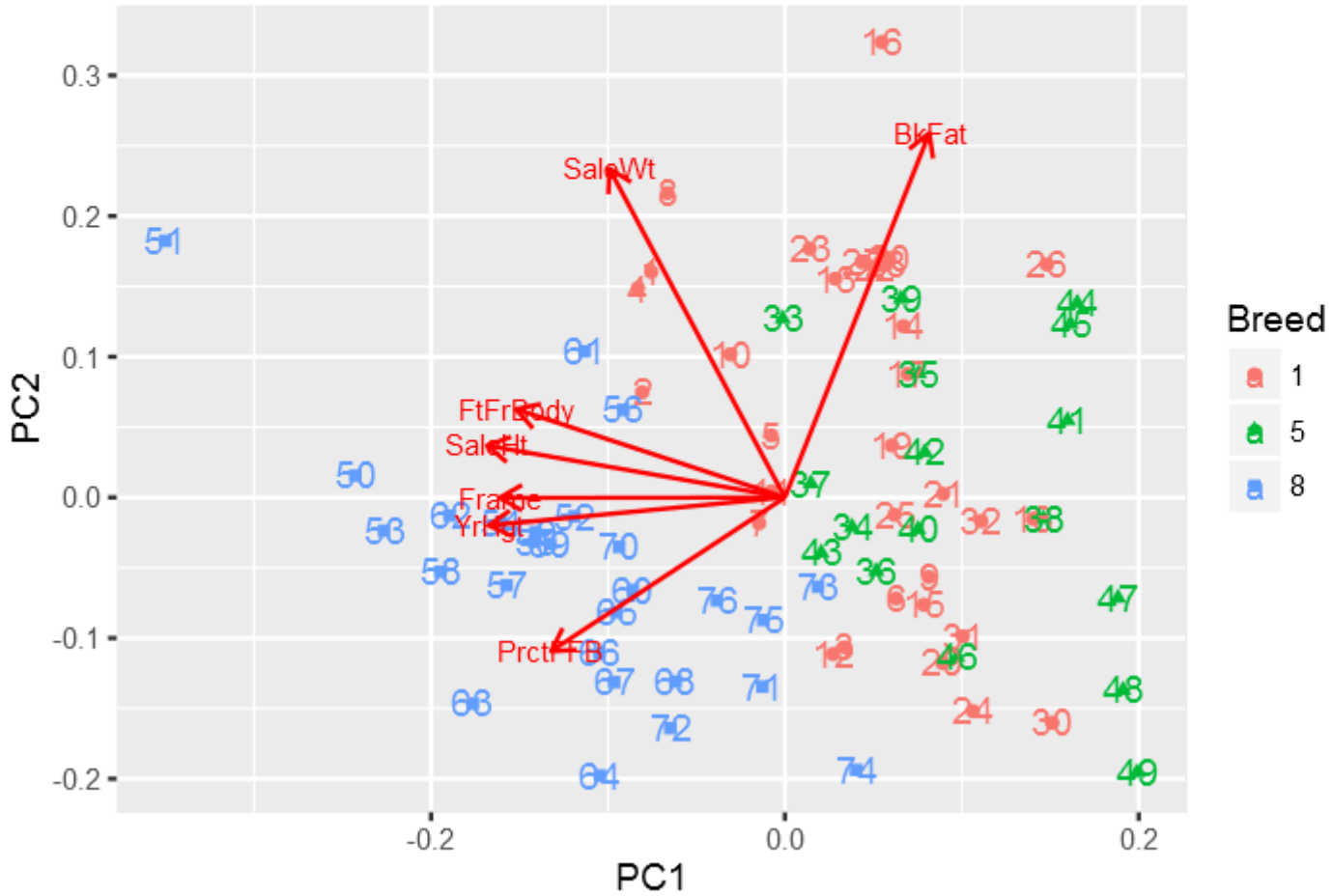
In [89]:
```
ggplot(DF, aes(x=PC1,y=PC2)) + geom_point(aes(colour=Breed,shape=Breed))
```



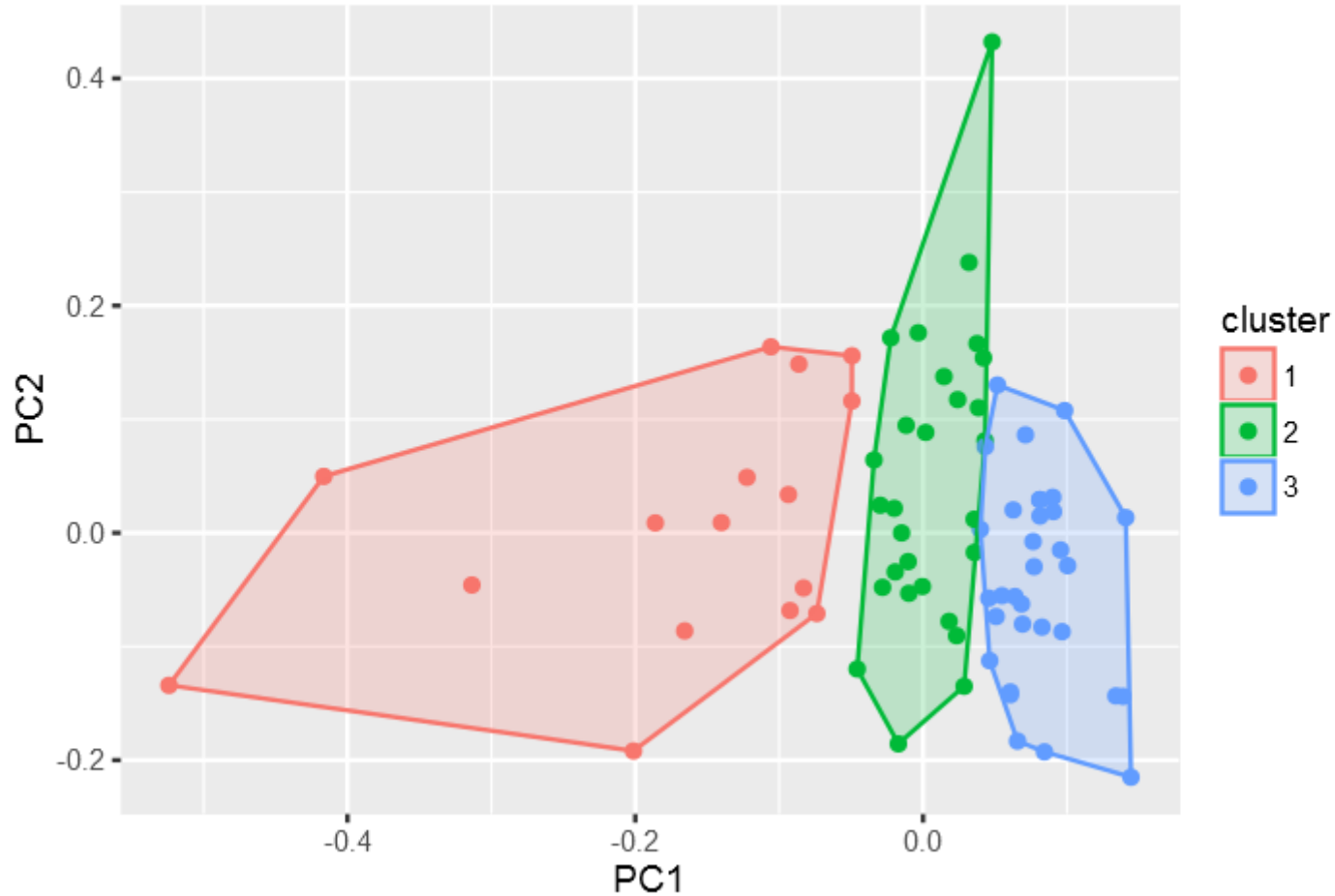
cluster, ggfortify 라이브러리를 활용한 Cluster 생성.

In [21]:
```
library(ggfortify)
library(cluster)
```

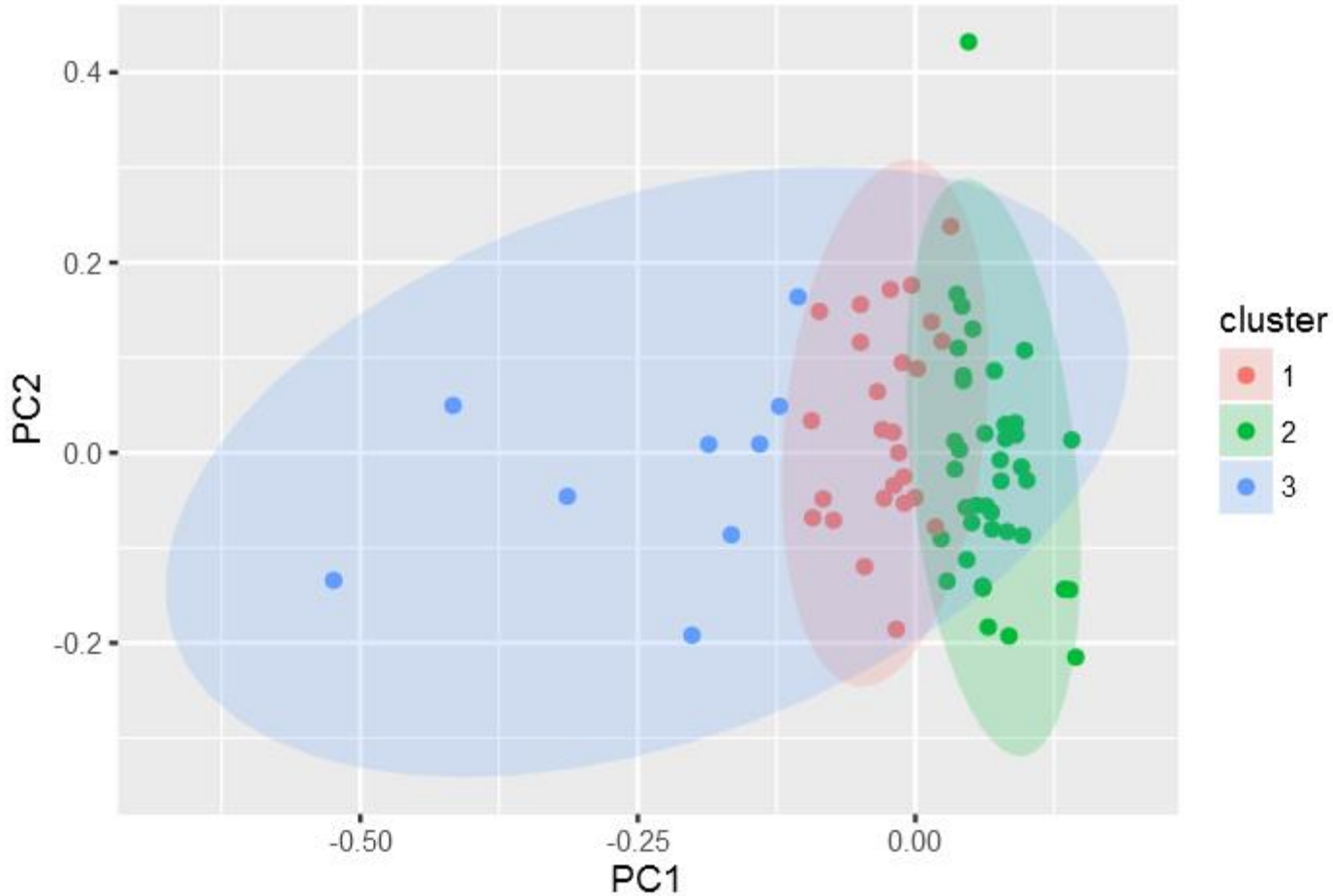
In [19]:
```
options(repr.plot.width = 6, repr.plot.height=4)
tmp <- bulls
tmp$Breed <- as.factor(tmp$Breed)
autoplot(bulls_pca,data=tmp,colour='Breed',label=T,shape='Breed', loadings=T,loadings.label = T,loadings.label.size = 3)
```



In [23]:
```
autoplot(fanny(tmp, 3), frame = TRUE)
```



In [24]:
```
autoplot(pam(tmp, 3), frame = TRUE, frame.type = 'norm')
```



6.첫 주성분을 사용해 Q-Q plot을 그리고 해석하시오.

- 정규성확인 결과 그래프의 변화가 일정한 기울기로 가고 있으므로 정규성을 잘 띄고 있다고 판단할 수 있을것이다.
- 주성분 분석 식 $PC1$ 이 정규성을 나타낸다는 것은 각 변수 즉, 항목들이 정규성을 나타낸다고 할 수 있다.
  하지만 개별 변수가 정규성을 띈다고 해도 주성분 식이 정규성을 띈다고 확실할 수 없는 부분이 있으므로 $PC1$ 의 정규성을 확인하는 것이 중요하다.
- 다시 정리하면, 다변량이 정규분포를 따른다면 각각의 단변량은 정규분포를 따르지만 각각의 단변량이 정규분포라고 하여 단변량을 합쳐
  다변량이 되었을때 정규분포를 따른다고 확신할 수 없다.

In [25]:
```
qqnorm(bulls_pca$x[,1])
qqline(bulls_pca$x[,1])
```