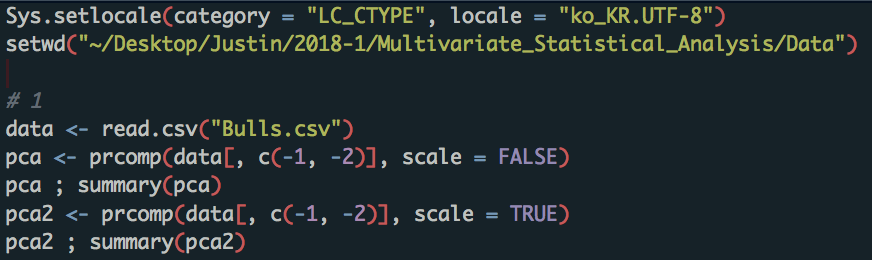
Bulls.csv 는 경매시장에서 거래된 76마리의 어린(2살 이하) 황소의 특성과 거래가격(SalePr)에 관한 자료이다. 변수 설명은 아래와 같다.

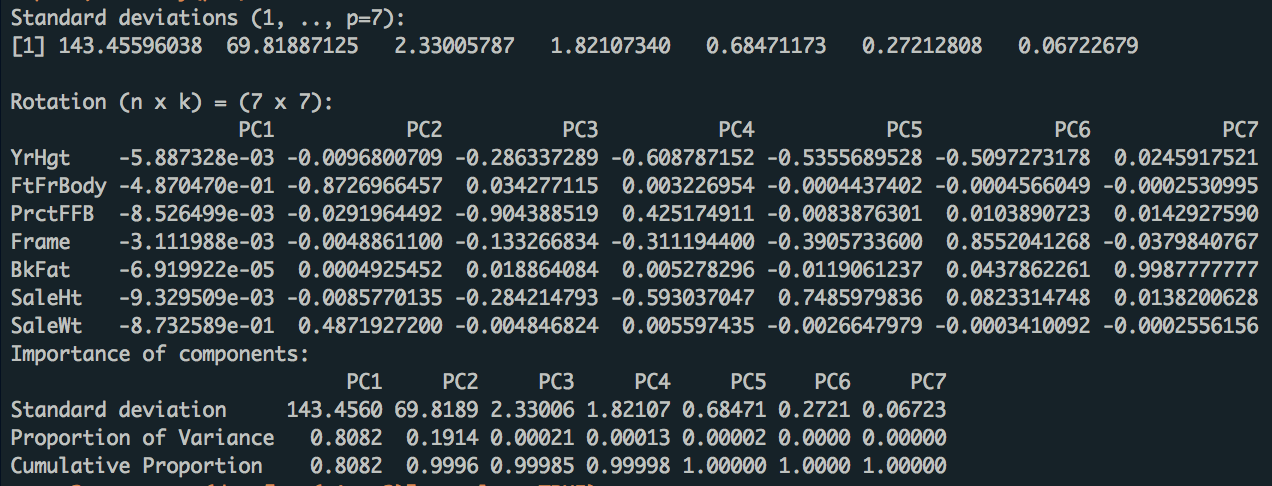
* Breed=1 if Angus, 5 if Hereford, 8 if Simental
* FtFrBody=fat free body (pounds)
* Frame=Scale from 1(small) to 8 (large)
* SaleHt=Sale height at shoulder (inches)
* YrHgt=Yearling height at shoulder (inches)
* PrctFFB=Percent fat-free body
* BkFat=Back fat (inches)
* SaleWt=Sale weight (pounds)

SalePr와 Breed 변수를 제외한 7개의 변수를 사용해 주성분분석을 시행하여 아래의 질문에 답하시오. 공분산 행렬과 상관계수 행렬을 사용하여 각각 분석하고 비교하시오.

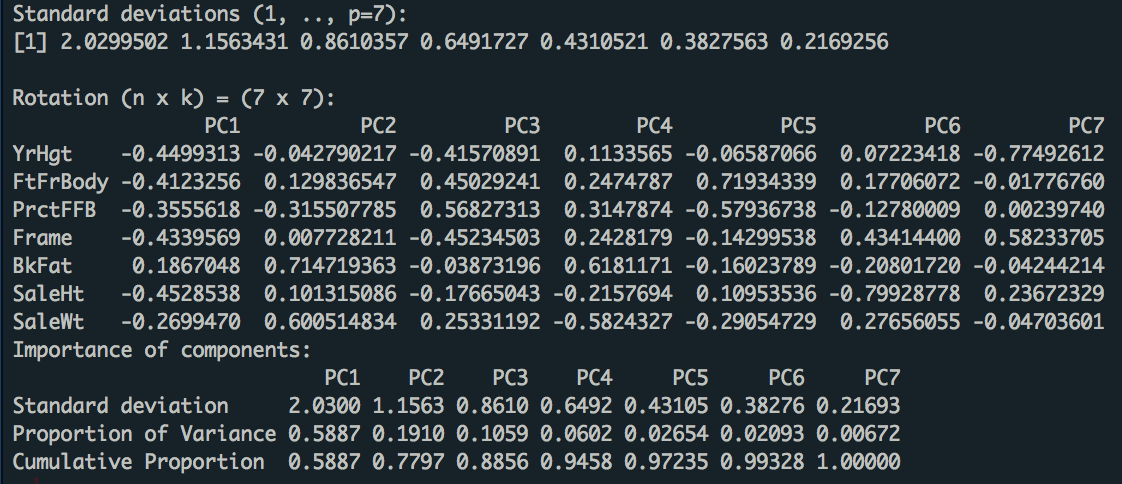
1. 각 주성분의 표준편차와 그 주성분을 계산하는데 사용된 rotation값을 찾으시오.



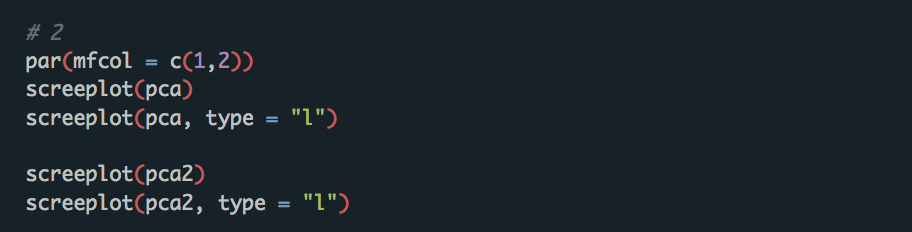
PCA : scale = FALSE (정규화X)

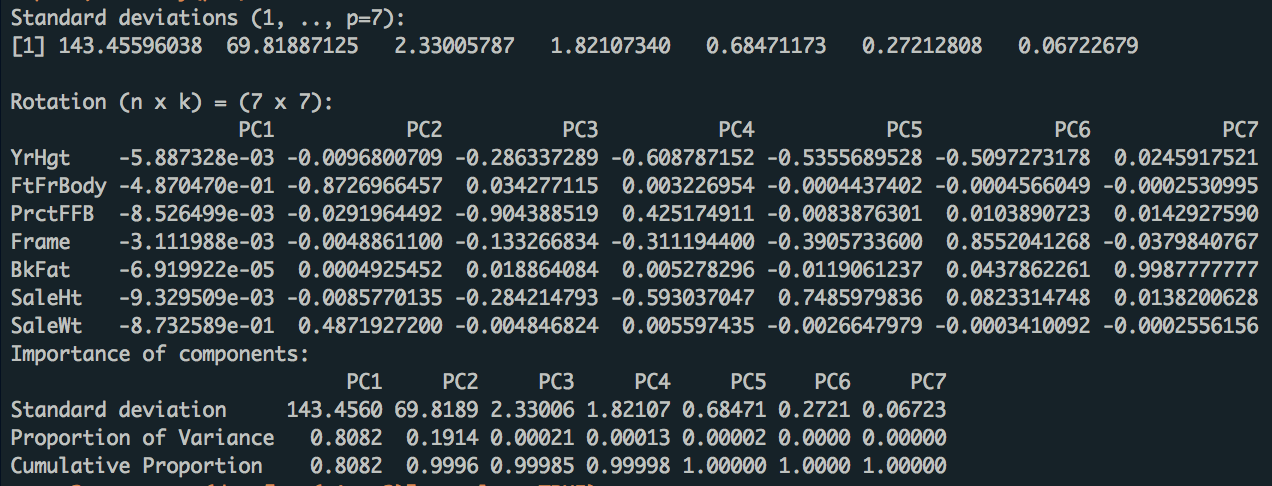
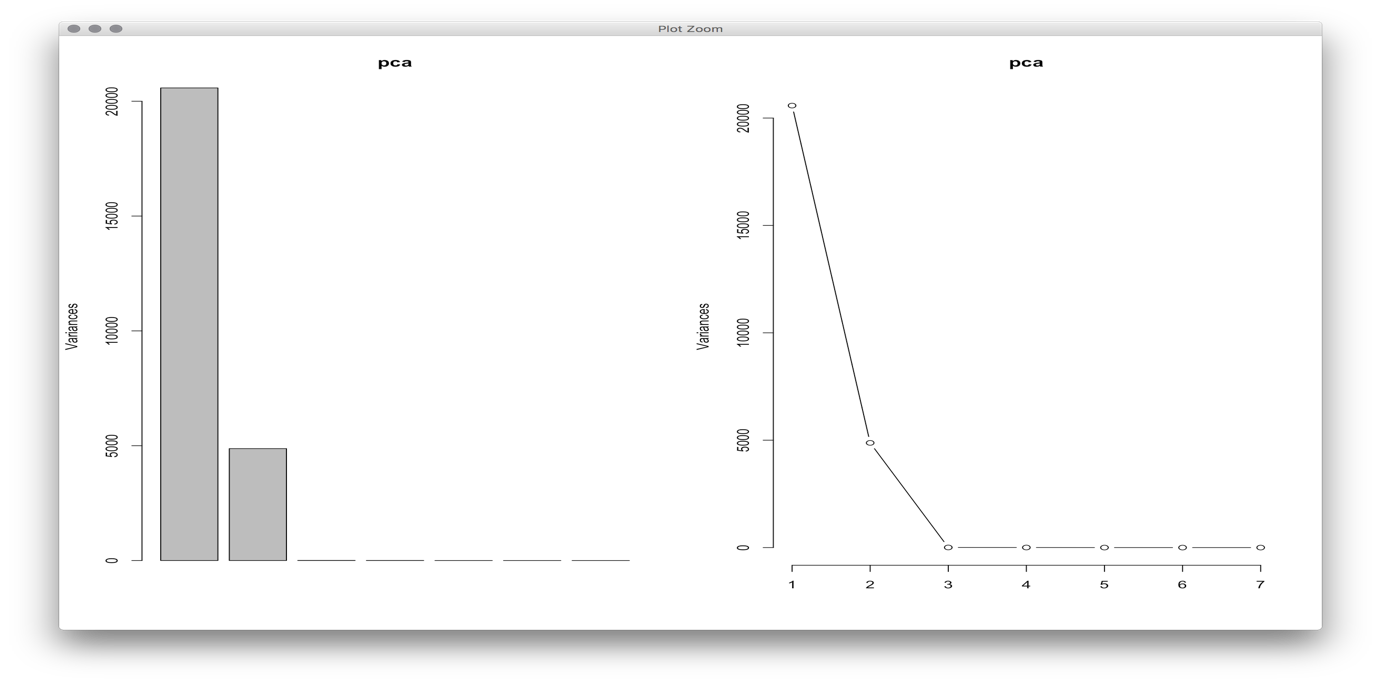


PCA2 : scale = TRUE (정규화O)



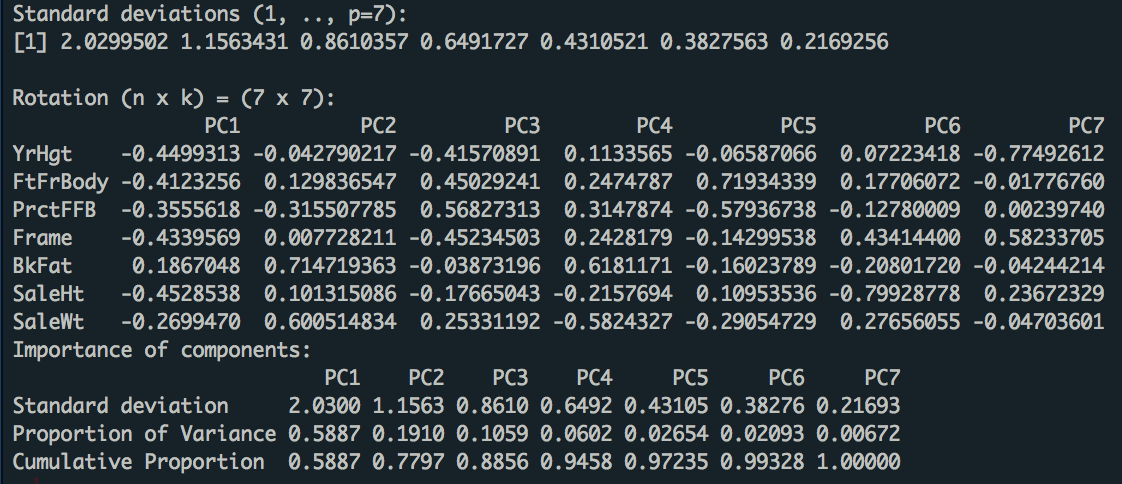
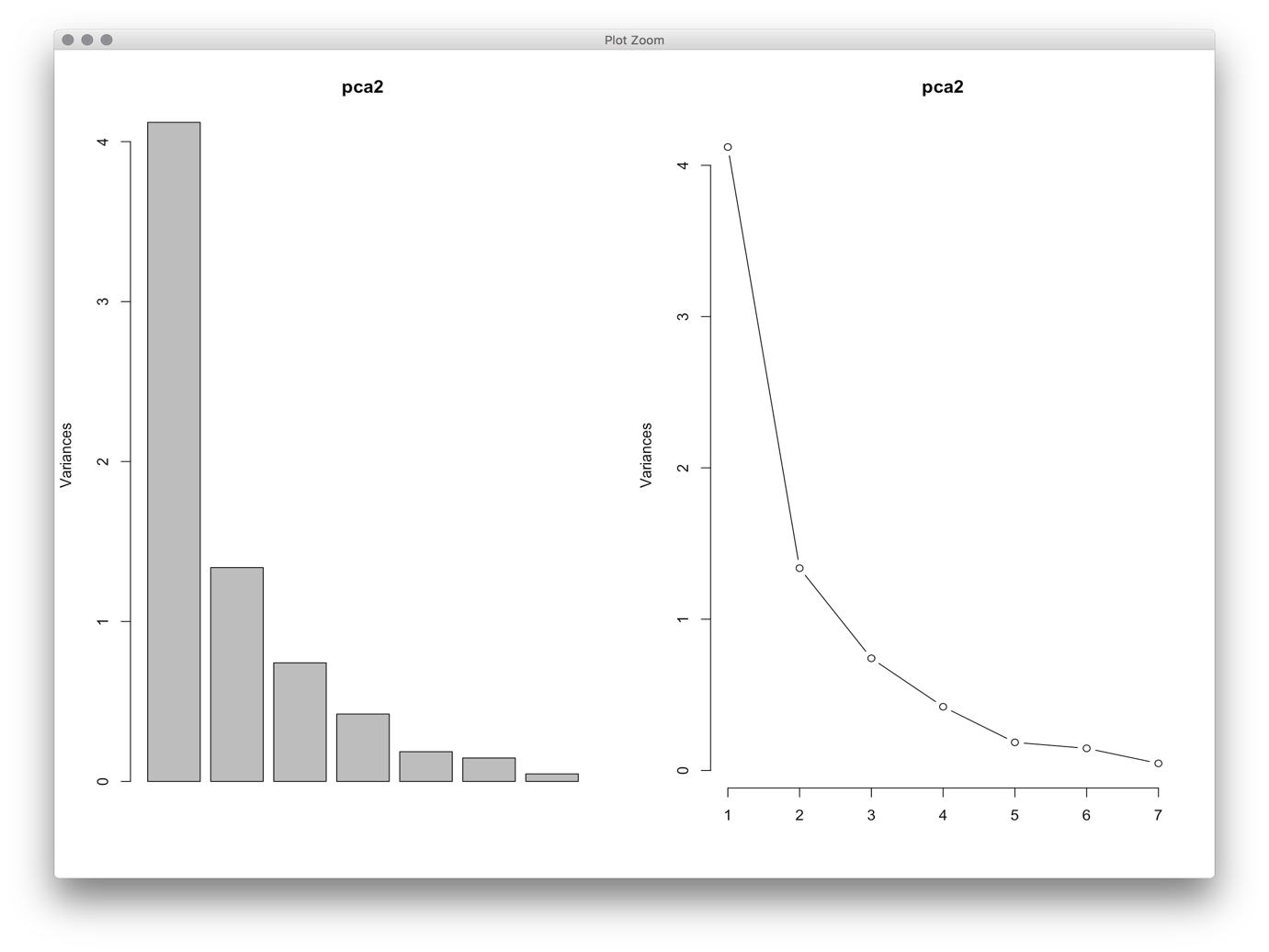
1. 적절한 주성분의 개수를 선택하고 근거를 설명하시오.





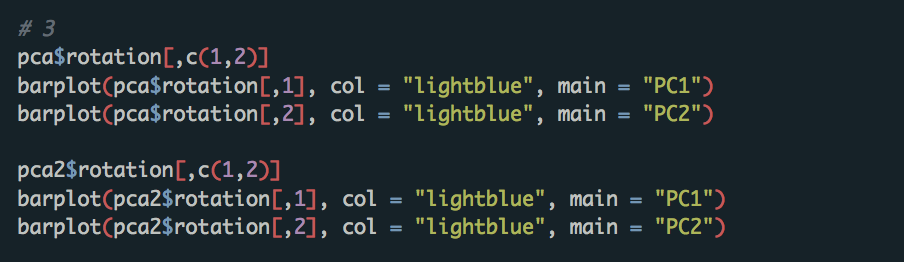
정규화를 하지 않은, 즉 PCA는 총분산을 설명하는 비중은, 주성분이 1개일 때, 약 80.82% 설명하고, 주성분이 2개일 때, 약 99.96% 설명한다. 주성분이 1개일 때도 높은 데이터 설명력을 가지지만, 약 99.96% 수준으로, 거의 모든 부분을 설명할 수 있는 주성분은 2개이므로 2개를 선택하는게 타당하다.

PCA2

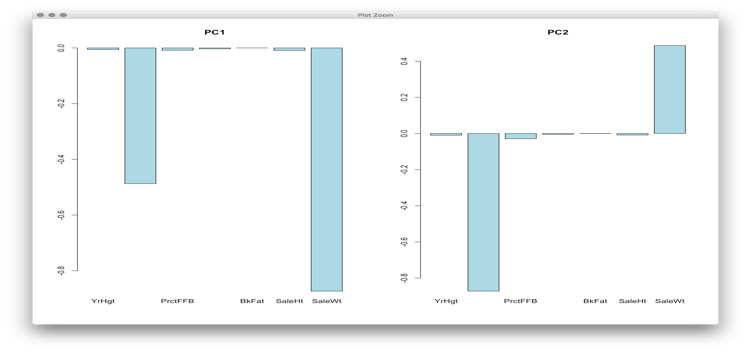
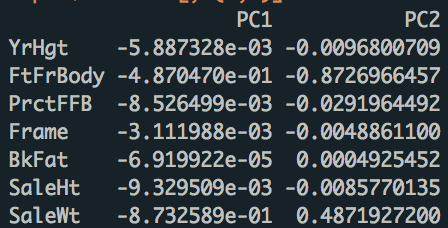


정규화를 거친, 즉 PCA2는 상관행렬로부터 추출되므로, 평균 분산이 1보다 작은 고유값을 제거하는게 이상적이다. 이에 따르면 PC2, 주성분은 2개까지 선택하는게 타당하다.

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



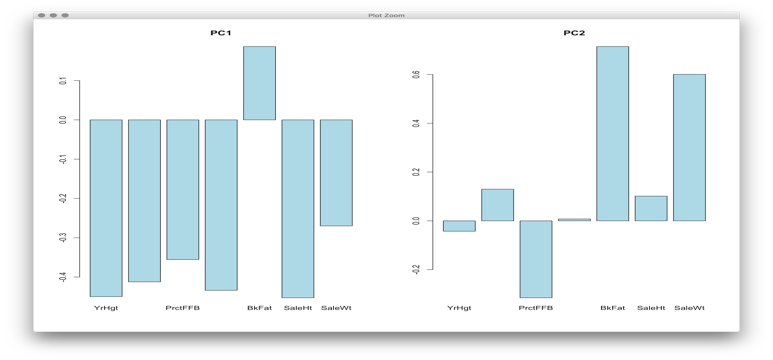
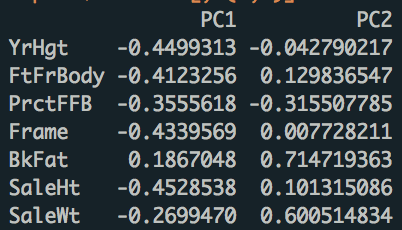
PCA



PC1은, FtFrBody값이 낮으며 SaleWt값이 낮은 황소를 나타낸다.

PC2는, FtFrBody값이 낮으며, SaleWt값이 높은 황소를 나타낸다.

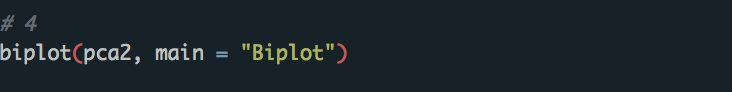
PCA2

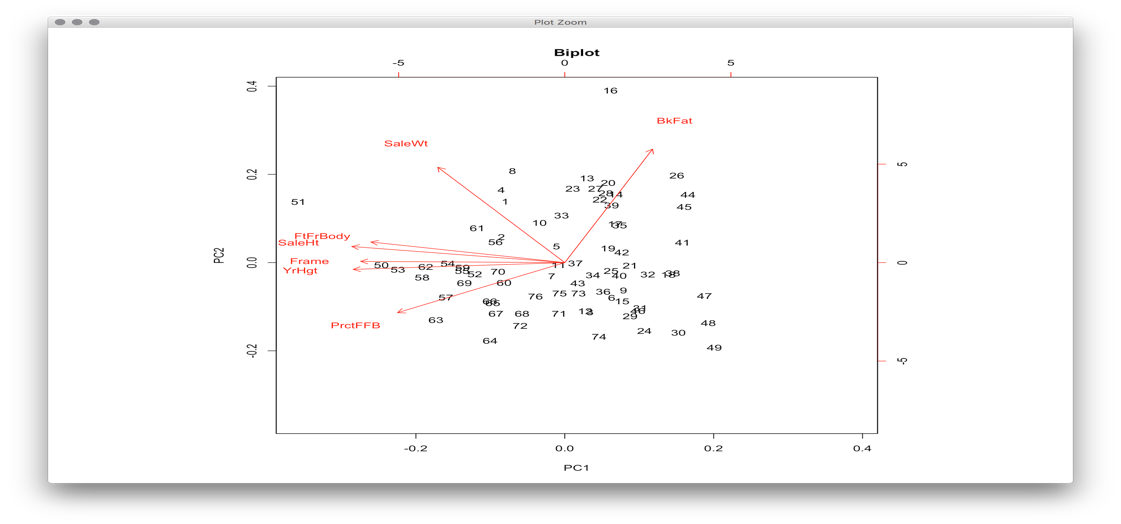


PC1은, BkFat값이 높고, 나머지 요소가 낮은 황소를 나타낸다.

PC2는, PrctFFB값이 낮으며, BkFat, SaleWt값이 높은 황소를 나타낸다.

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





FtFrBody, SaleHt, Frame, YrHgt 는 상관관계가 높다고 판단할 수 있다.

또한 이들은 PC1과 상관관계가 높은 것으로 나타난다.

BkFat은 다른 종목들과 비교적 상관관계가 적다고 판단할 수 있다.

반면 PC2는 BkFat, SaleWt와 비교적 상관관계가 높다고 판단할 수 있다.

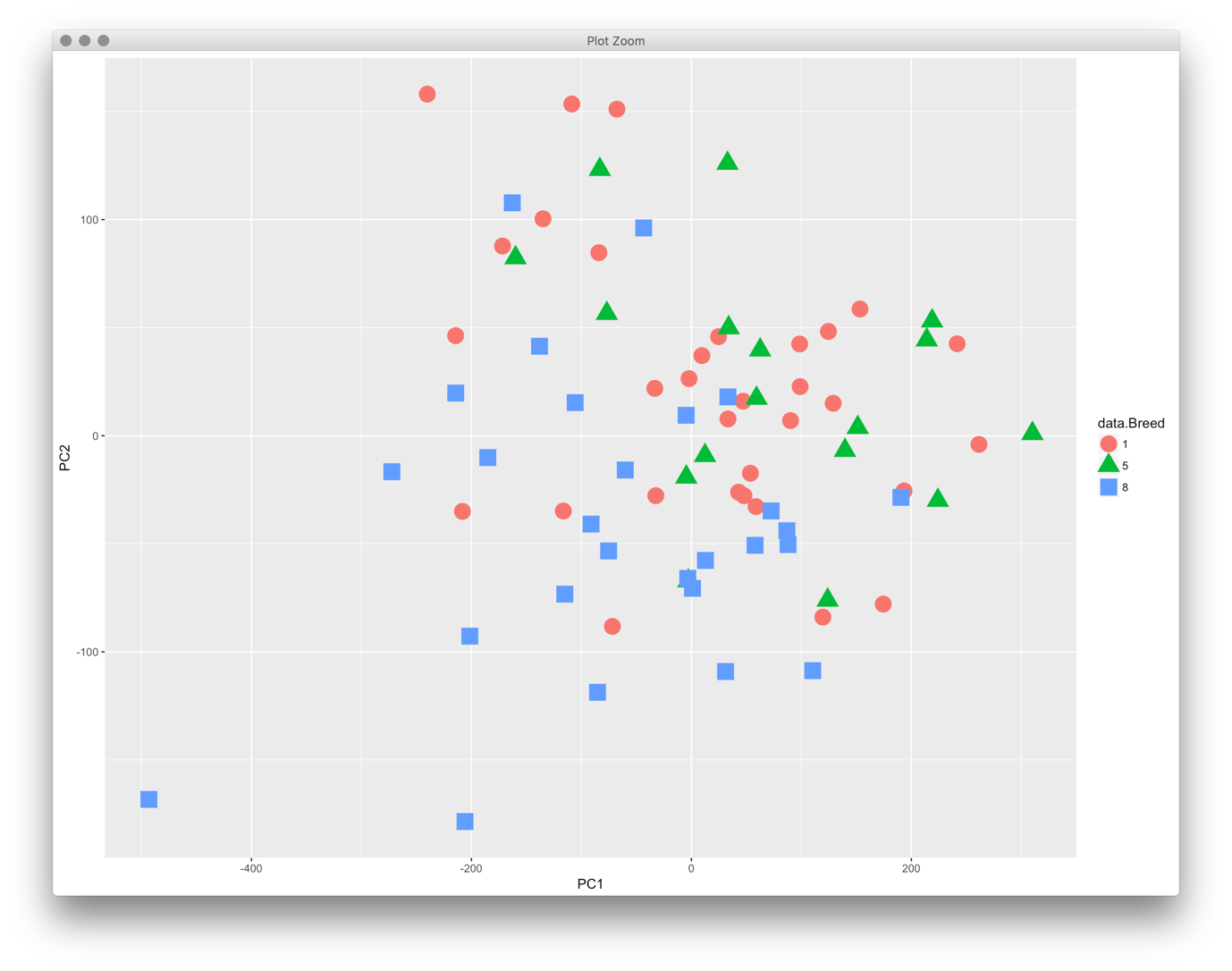
각 종목들의 화살표에 근접할수록, 해당 종목과 연관성이 높다고 판단할 수 있다.

중앙에서 멀어질수록 값이 높다고 판단할 수 있다.

16번과 51번 데이터는 타 데이터에 비해 동떨어져 있어 이상치로 판단할 수 있다.

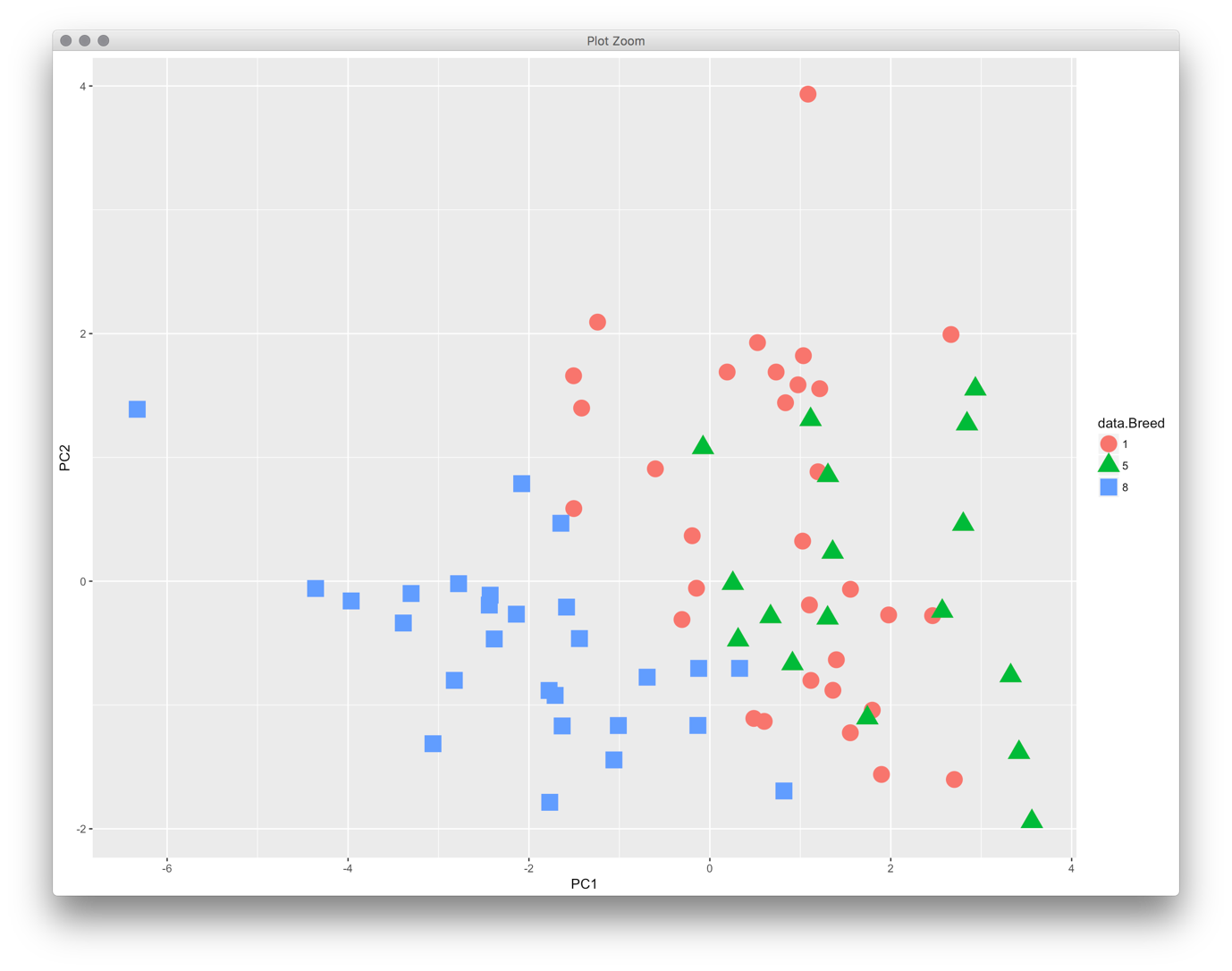
1. 첫 두개의 주성분을 사용해 산점도를 그리고 Breed를 서로 다른 색깔과 기호로 표시하시오. 주성분에 의해 다른 종의 황소를 구분할 수 있는가? 이상점이 있는가? 있다면 어떤 특성을 가진 소인가?

PCA



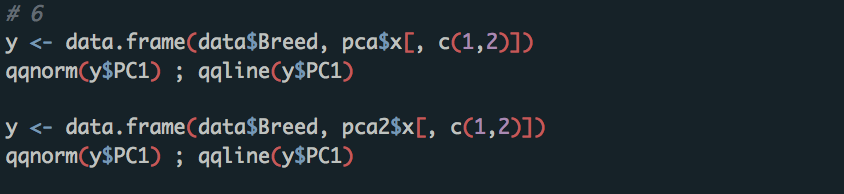
PCA경우, 주성분을 이용해서 황소를 구분하기 어려워 보이는 것을 직감적으로 확인할 수 있다. PC1은, FtFrBody값이 낮으며 SaleWt값이 낮은 황소를 나타내고, PC2는, FtFrBody값이 낮으며, SaleWt값이 높은 황소를 나타내며 PC1값과 PC2값이 낮은, 위의 빨간 원으로 표시된 데이터를 이상점으로 판단할 수 있다.

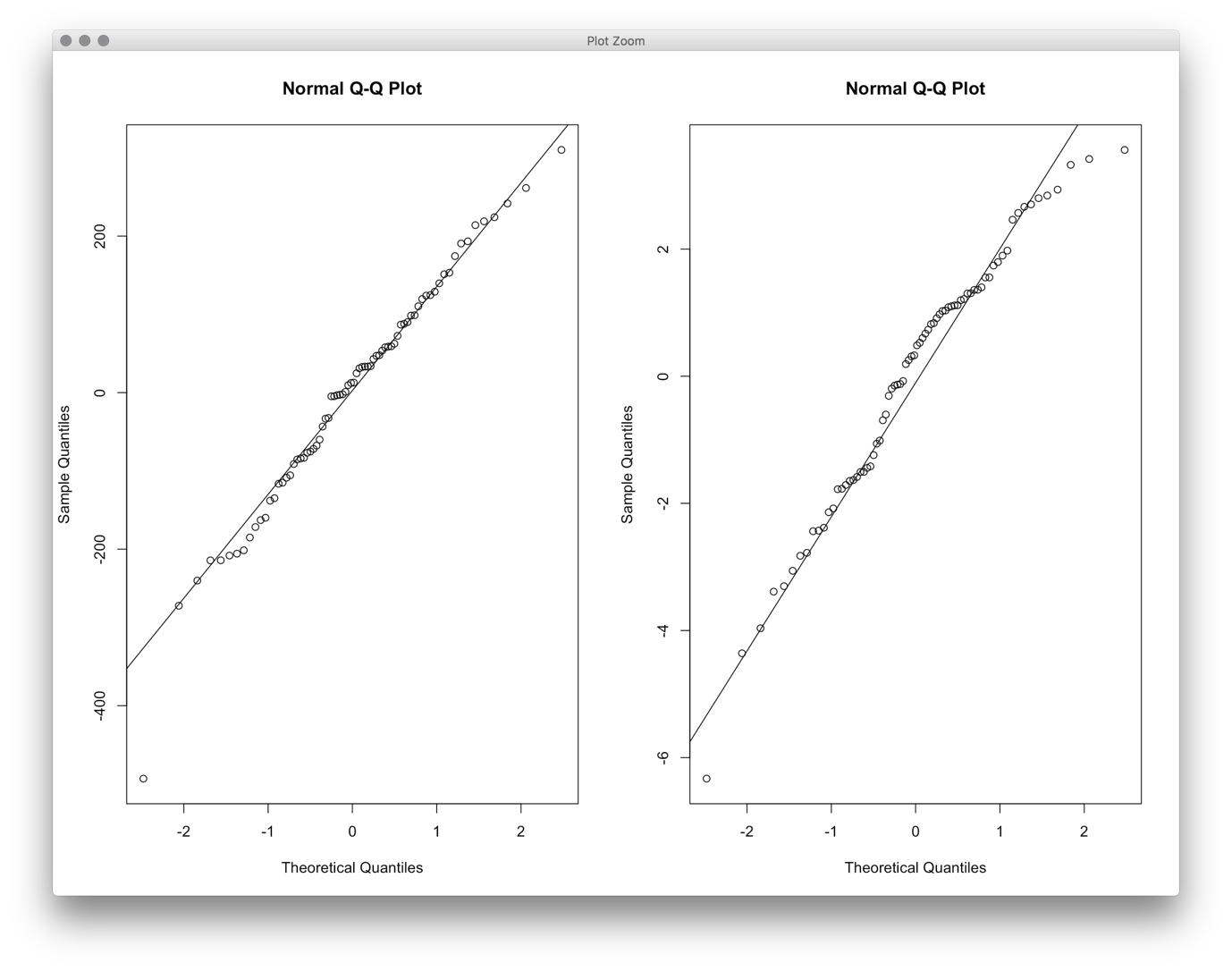
PCA2



PCA2경우에는, PCA1 경우보다 황소의 종을 구분하기에 적합하다고 판단할 수 있다. 좌측 상단의 푸른색 데이터와, 우측 상단의 붉은색 데이터를 이상치로 판단할 수 있다. 이 그래프의 경우, PC1은, BkFat값이 높고, 나머지 요소가 낮은 황소를 나타내며, PC2는, PrctFFB값이 낮으며, BkFat, SaleWt값이 높은 황소를 나타낸다. 푸른색 데이터인 경우, PC1이 매우 낮은 특징을 갖고 있으며, 붉은색 데이터인 경우, PC2가 매우 높은 특징을 갖고 있다.

1. 첫 주성분을 사용해 Q-Q plot을 그리고 해석하시오.





좌측이 PCA, 우측이 PCA2 의 Q-Q Plot 이다.

두 그래프 모두 어느정도 정규성을 따른다고 판단할 수 있지만, PCA 경우가 PCA2보다 좀 더 높은 수준으로 정규성을 따른다고 판단할 수 있다. 경우에 따라 그래프에서 상대적으로 멀리 떨어진 데이터들을 이상치로 판단할 수 있다.