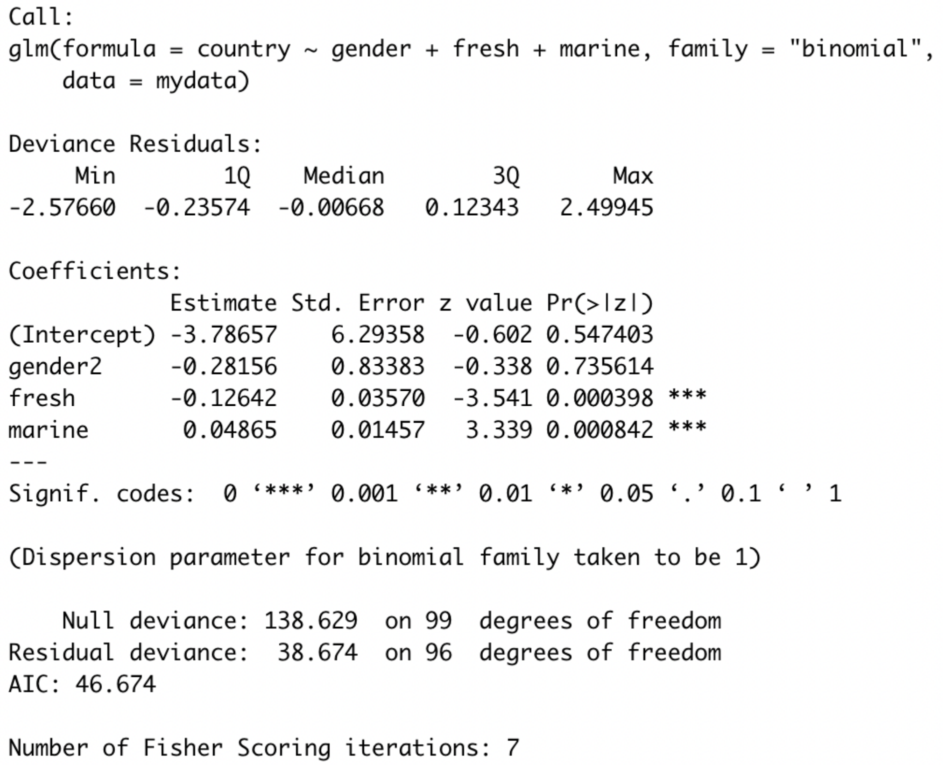
HW1

Han Ambrose

March 2021

Question 2

2a



```
Call:
glm(formula = country ~ gender + fresh + marine, family = "binomial",
    data = mydata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.57660  -0.23574  -0.00668   0.12343   2.49945

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.78657    6.29358  -0.602 0.547403
gender2     -0.28156    0.83383  -0.338 0.735614
fresh       -0.12642    0.03570  -3.541 0.000398 ***
marine       0.04865    0.01457   3.339 0.000842 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138.629  on 99  degrees of freedom
Residual deviance:  38.674  on 96  degrees of freedom
AIC: 46.674

Number of Fisher Scoring iterations: 7
```

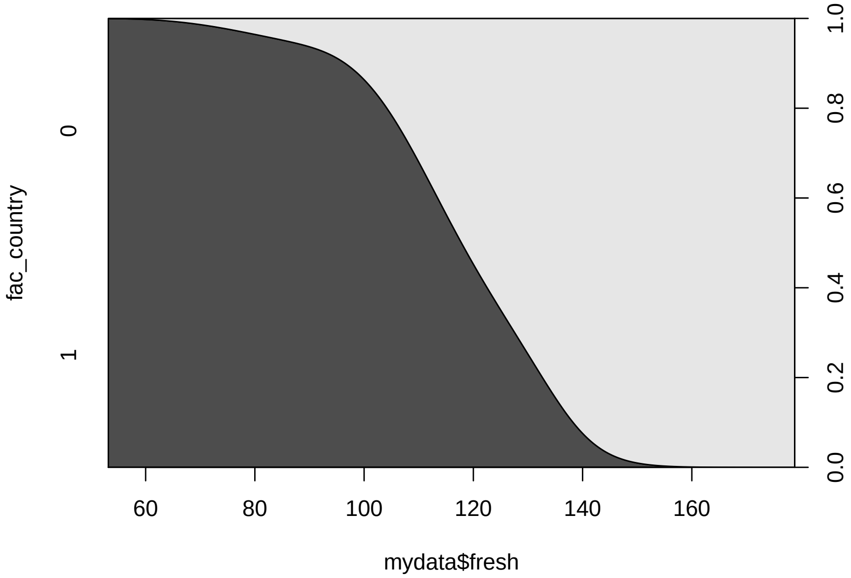
2b

Let 1 = Alaskan, 0 = Canadian.

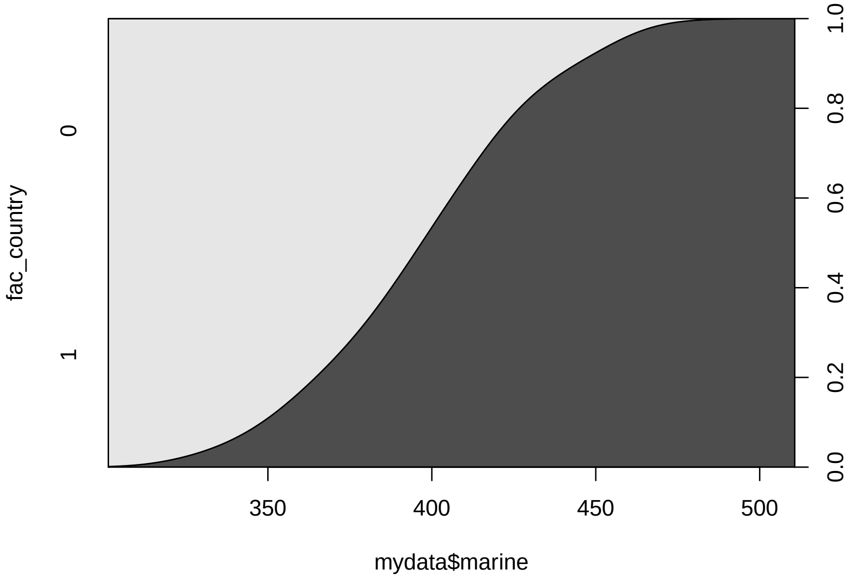
The dark shaded area corresponds to Result =1 (Alaskan) the light shaded area corresponds to Result =0 (Canadian).

The probability that Result =1 (Alaskan) when diameter of rings for first year freshwater fish =140 is approximately 5%. The probability that Result = 0 (Canadian) when diameter of rings for first year freshwater fish =140 is approximately 95%.

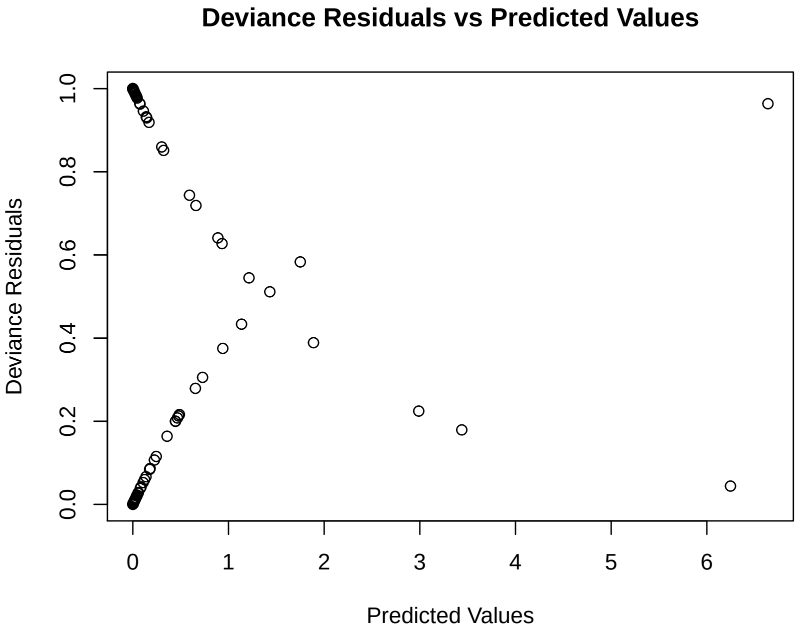
The rings associated with freshwater growth are smaller for the Alaskan-born than for the Canadian-born salmon



Similarly, The rings associated with marine growth are larger for the Alaskan-born than for the Canadian-born salmon


fac_country
0
1
350
400
450
500
mydata$marine
0.0
0.2
0.4
0.6
0.8
1.0

2d


Deviance Residuals vs Predicted Values
Deviance Residuals
0.0
0.2
0.4
0.6
0.8
1.0
Predicted Values
0
1
2
3
4
5
6

Question 6

6a

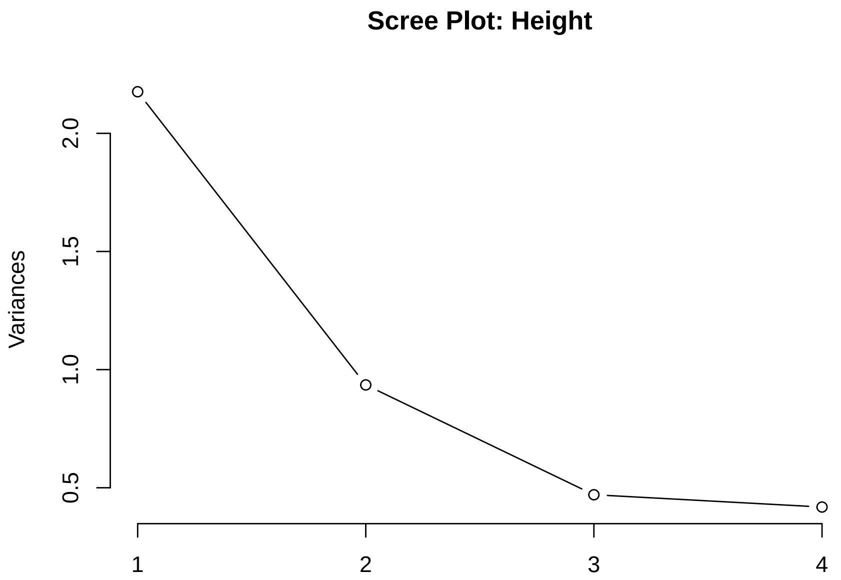
I used only those parent pairs who have 2 or more kids
I kept the heights of the first born male and female from each family
I also removed the N/A in the data, meaning family who only has daughters or sons are removed

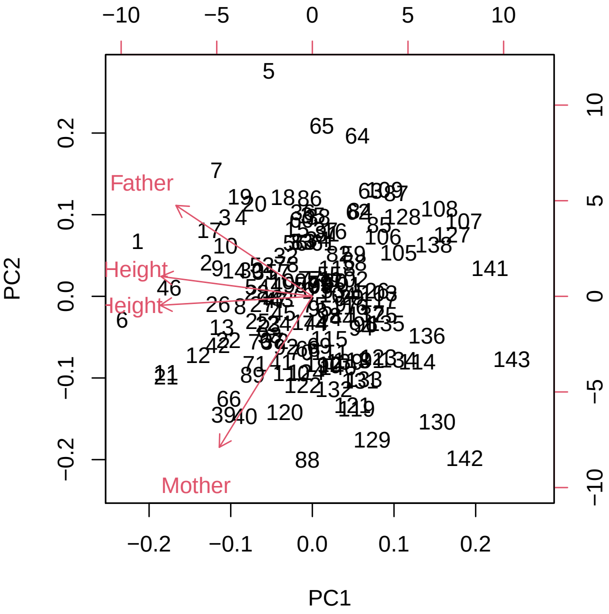
6b

Standardized

Percent Variation
0.5440050 0.2338080 0.1175965 0.1045905

Most of the variation is explained by the first 2 or three PCs. The first 2 PCs explain about 80%

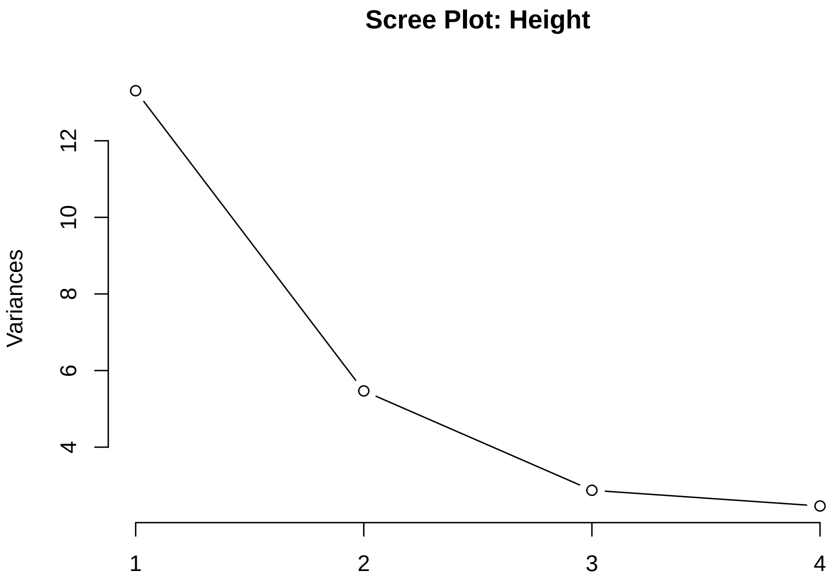


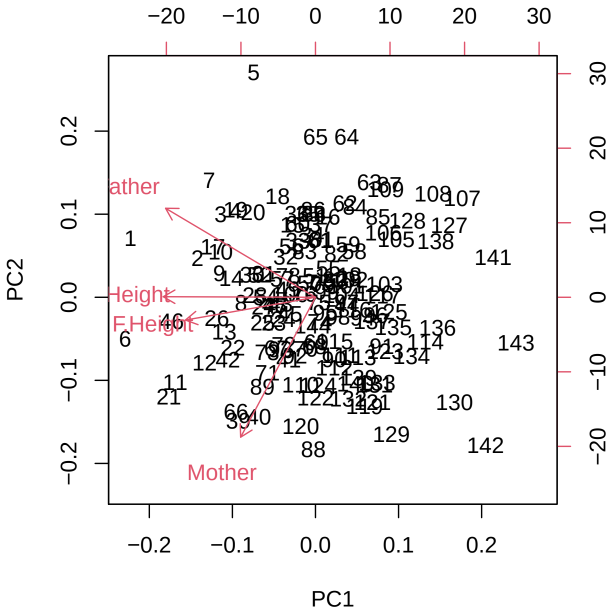


UnStandardized

Percent Variation
0.5519051 0.2267933 0.1191546 0.1021470

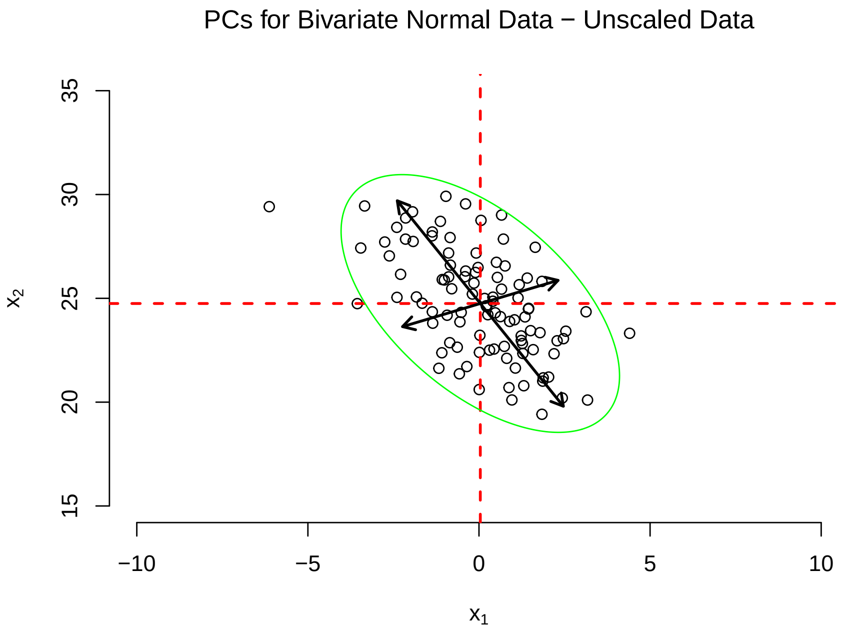
Most of the variation is explained by the first 2 or three PCs. The first 2 PCs also explain about 80%



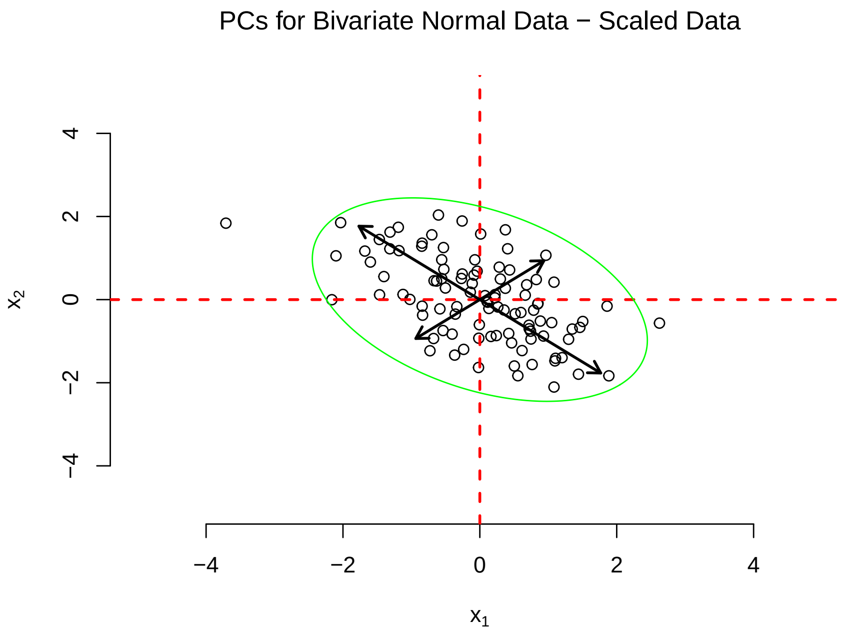


The results of standardized and unstandardized are about the same given they have the same unit of measurement

Question 7



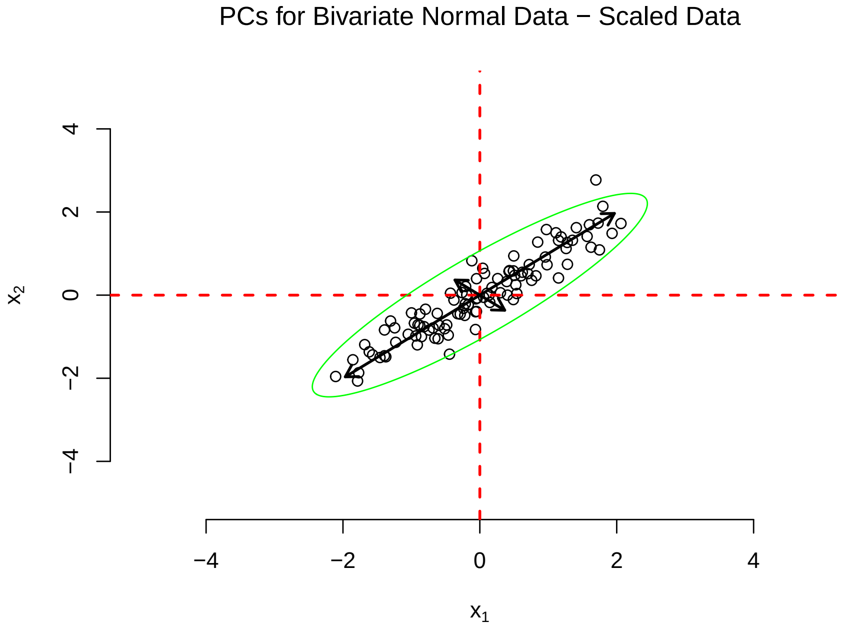
PCs for Bivariate Normal Data − Unscaled Data
35
30
25
20
15
x_2
−10
−5
0
5
10
x_1



PCs for Bivariate Normal Data − Scaled Data
4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1

Take any other arbitrary positive definite 2 x 2 correlation matrix and find its eigenvectors.

Pick $\rho_{12} = 0.9$, we observed that as it goes to 1, there is a perfect correlation between x_1 and x_2 with 45 degree eigenvectors.

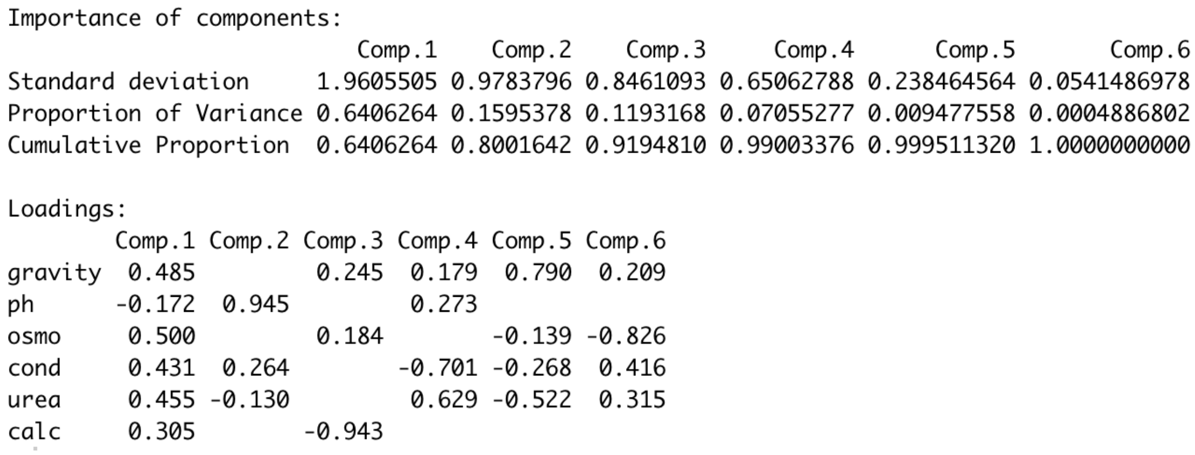


Question 8

8a

First we scaled and centered the data before PCA

CASE 1: calcium oxalate = 0: not present

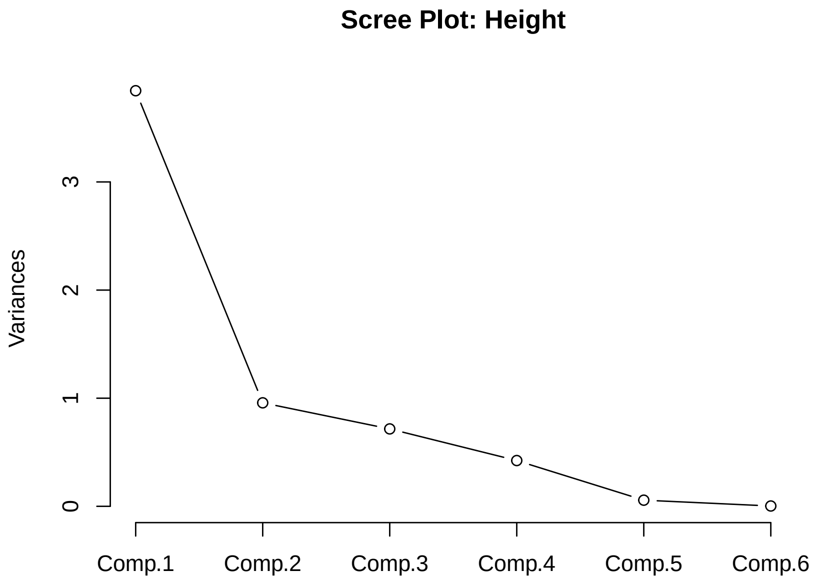


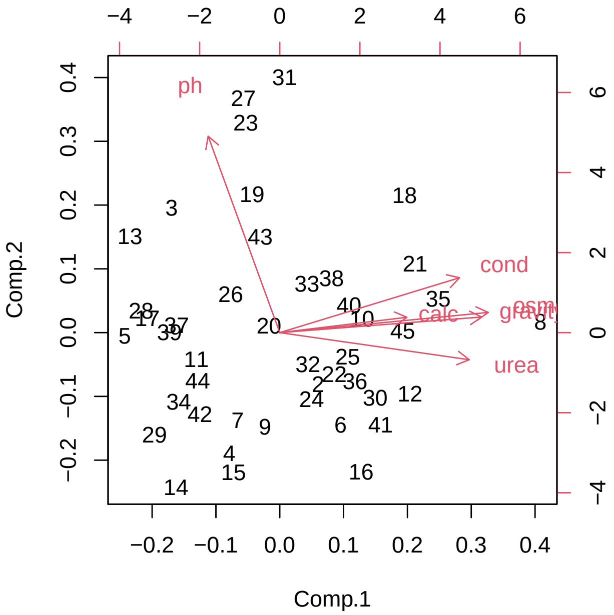
```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4      Comp.5       Comp.6
Standard deviation     1.9605505 0.9783796 0.8461093 0.65062788 0.238464564 0.0541486978
Proportion of Variance 0.6406264 0.1595378 0.1193168 0.07055277 0.009477558 0.0004886802
Cumulative Proportion  0.6406264 0.8001642 0.9194810 0.99003376 0.999511320 1.0000000000

Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
gravity  0.485         0.245  0.179  0.790  0.209
ph      -0.172  0.945         0.273              
osmo     0.500         0.184        -0.139 -0.826
cond     0.431  0.264        -0.701 -0.268  0.416
urea     0.455 -0.130         0.629 -0.522  0.315
calc     0.305        -0.943                     
```

The first two components from the correlation matrix accounts for more than 80% of the total variance of the observed variables.

We see that the first component might be regarded as concentration with high values of osmo, gravity and urea. The second component is largely concerned with ph level having high coefficients for ph and the third component is concerned with calcium concentration

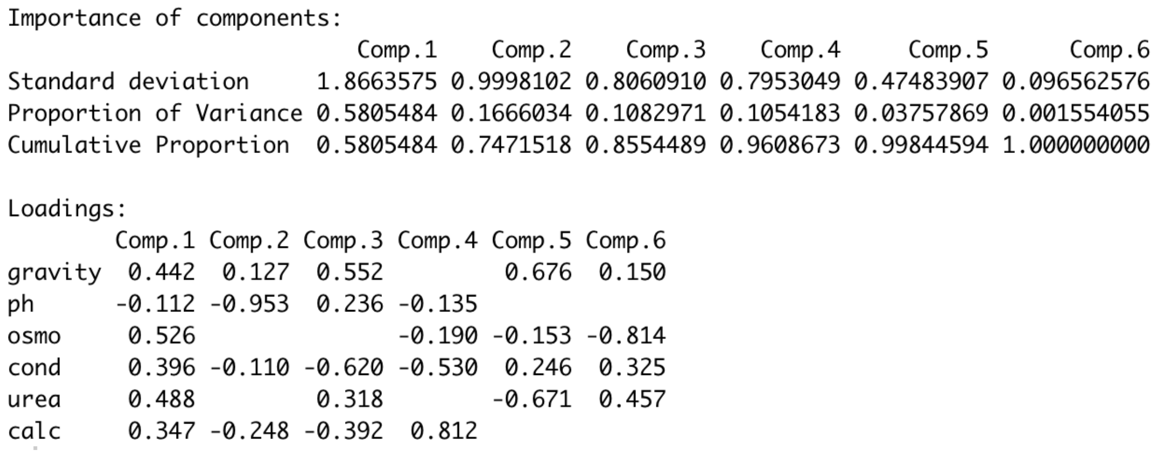




We can also see from the biplot that the results of the cond, osm, calc, gravity, urea are highly correlated.

On the other hand, the second component indicating ph uncorrelated with those with high coefficients from the first components.

CASE 2: calcium oxalate = 1: present

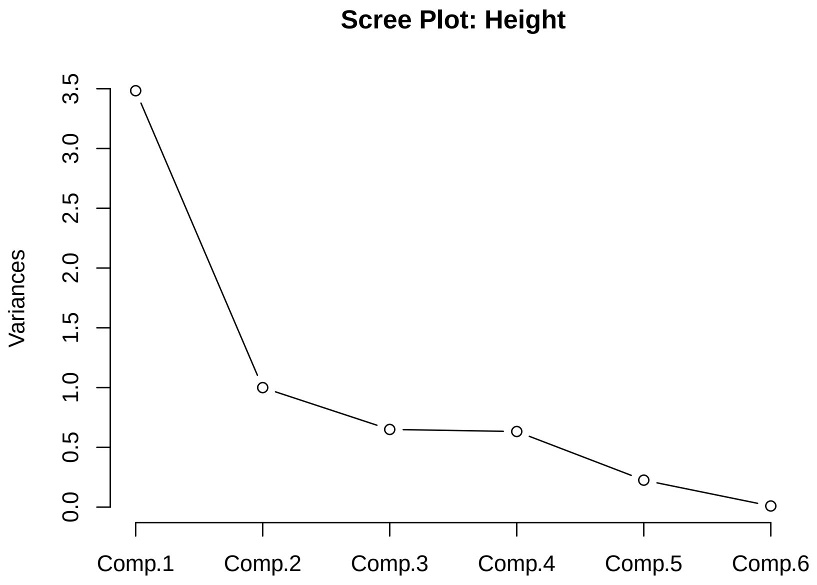


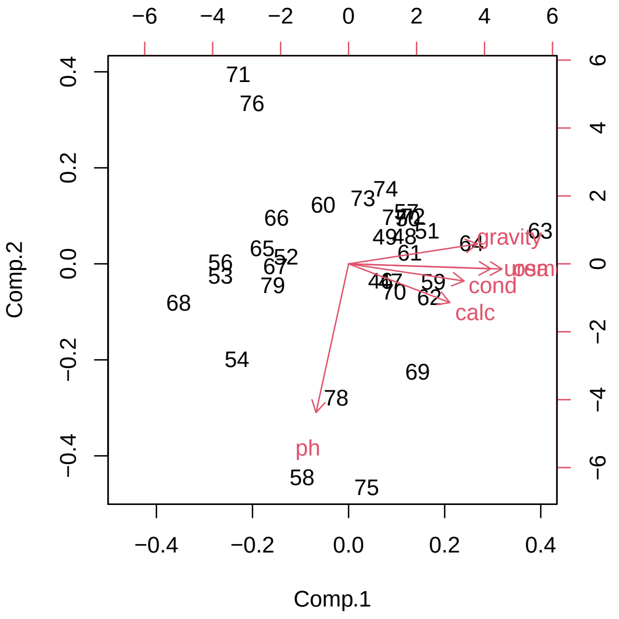
```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5      Comp.6
Standard deviation     1.8663575 0.9998102 0.8060910 0.7953049 0.47483907 0.096562576
Proportion of Variance 0.5805484 0.1666034 0.1082971 0.1054183 0.03757869 0.001554055
Cumulative Proportion  0.5805484 0.7471518 0.8554489 0.9608673 0.99844594 1.000000000

Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
gravity  0.442  0.127  0.552         0.676  0.150
ph      -0.112 -0.953  0.236 -0.135              
osmo     0.526               -0.190 -0.153 -0.814
cond     0.396 -0.110 -0.620 -0.530  0.246  0.325
urea     0.488         0.318        -0.671  0.457
calc     0.347 -0.248 -0.392  0.812              
```

The first two components from the correlation matrix accounts for more than 75% of the total variance of the observed variables.

We see that the first component might be regarded as concentration with high values of osmo, gravity and urea. The second component is largely concerned with ph level having high coefficients for ph and the third component is concerned with calcium concentration





Both subset of the PCs appear to sufficiently capture the variationin each dataset.

The only difference is the presence of ph level is negatively correlated with the presence of calcium oxalate for PC 2.

8b

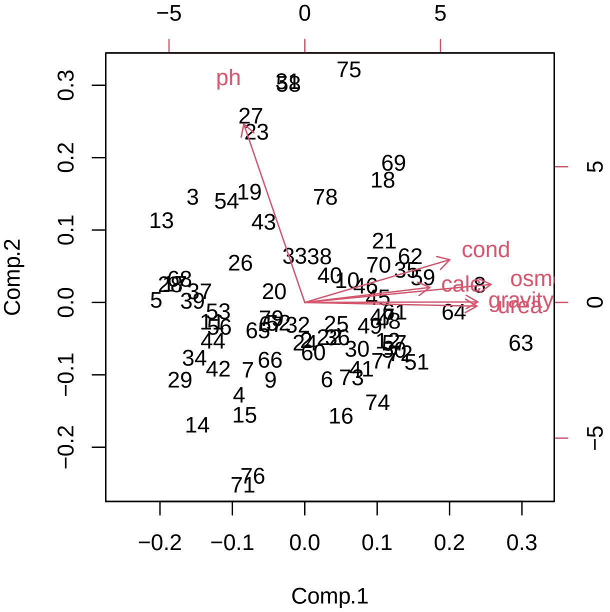


Figure 1: Combined data of when calcium oxalate is present and not present

We can see that the lower numbers (no presence of cal oxalate) are mostly on the left cluster and higher numbers (presence of cal oxalate) are mostly on the right side. However, it is not a very clear distinction.

Comparison of sub dataset and whole dataset

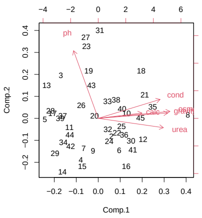


Figure 2: cal oxalate = 0

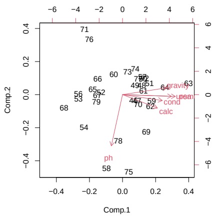


Figure 3: cal oxalate = 1

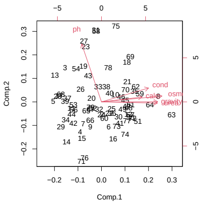


Figure 4: Combined