35459 Multivariate Statistics

Week 7 Seminar - Principal Components Analysis

The data in Notes.csv contain various characteristics of 100 genuine and 100 counterfeit Swiss bank notes. The characteristics include:

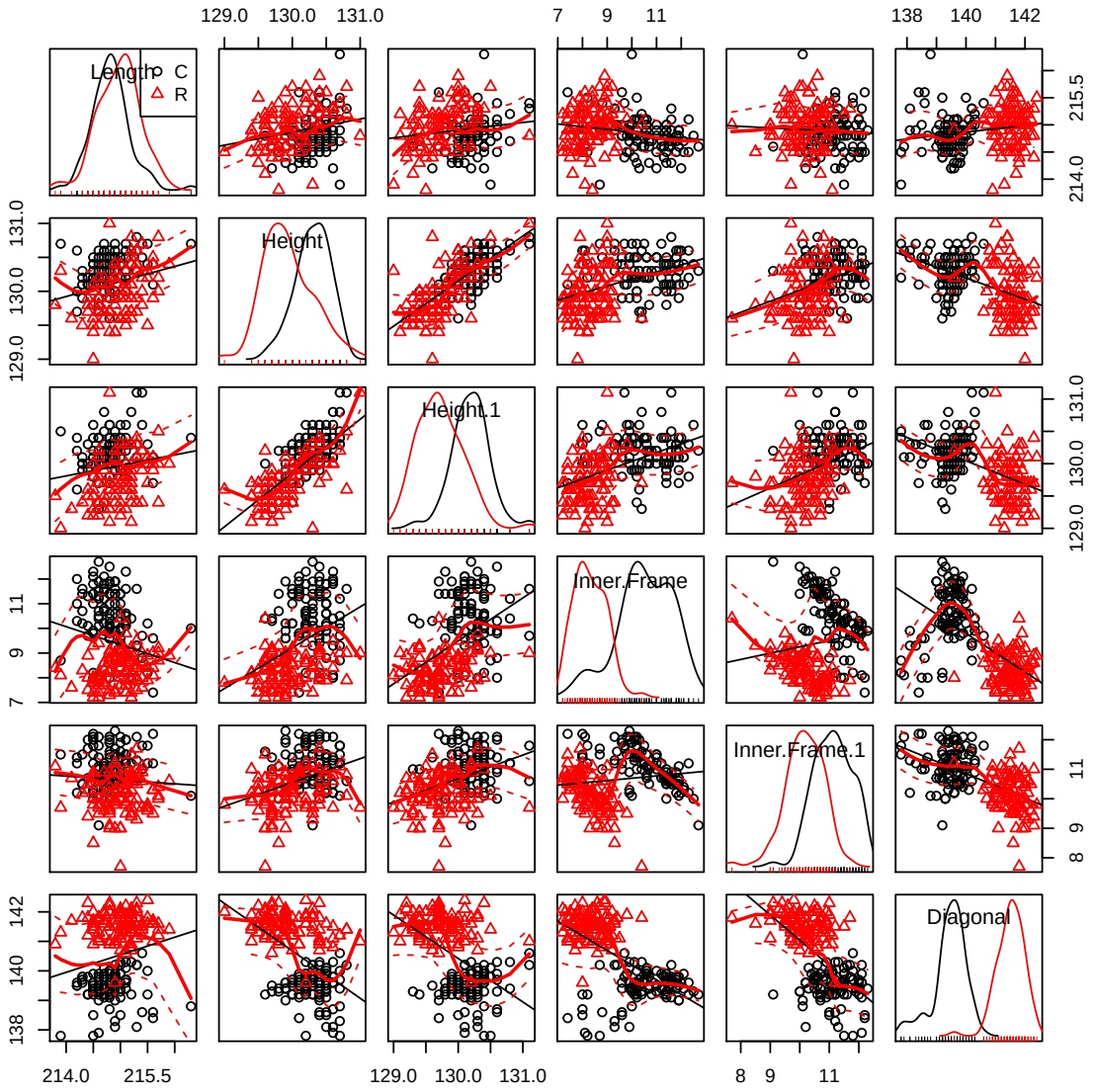
- Length of the bank note
- Height of the bank note, measured on the left
- Height of the bank note, measured on the right
- Distance of inner frame to the lower border
- Distance of inner frame to the upper border
- Length of the diagonal

Observations 1-100 are the genuine bank notes and the other 100 observations are the counterfeit bank notes. Determine whether these measures are different between the two types of note and produce confidence intervals for the difference between the notes.

We are going to perform PC analysis on all of the variables except for whether the notes are genuine or not.

```
library(car)
```

```
notes_data<-read.csv("C:/Documents/Notes.csv") notes_data$Group<-
substr(notes_data$Status,1,1) scatterplotMatrix(~Length+Height+Height.1
+Inner.Frame+Inner.Frame.1+Diagonal |
                    Group,data=notes_data)
```




```
cov(notes_data[,1:6])
```

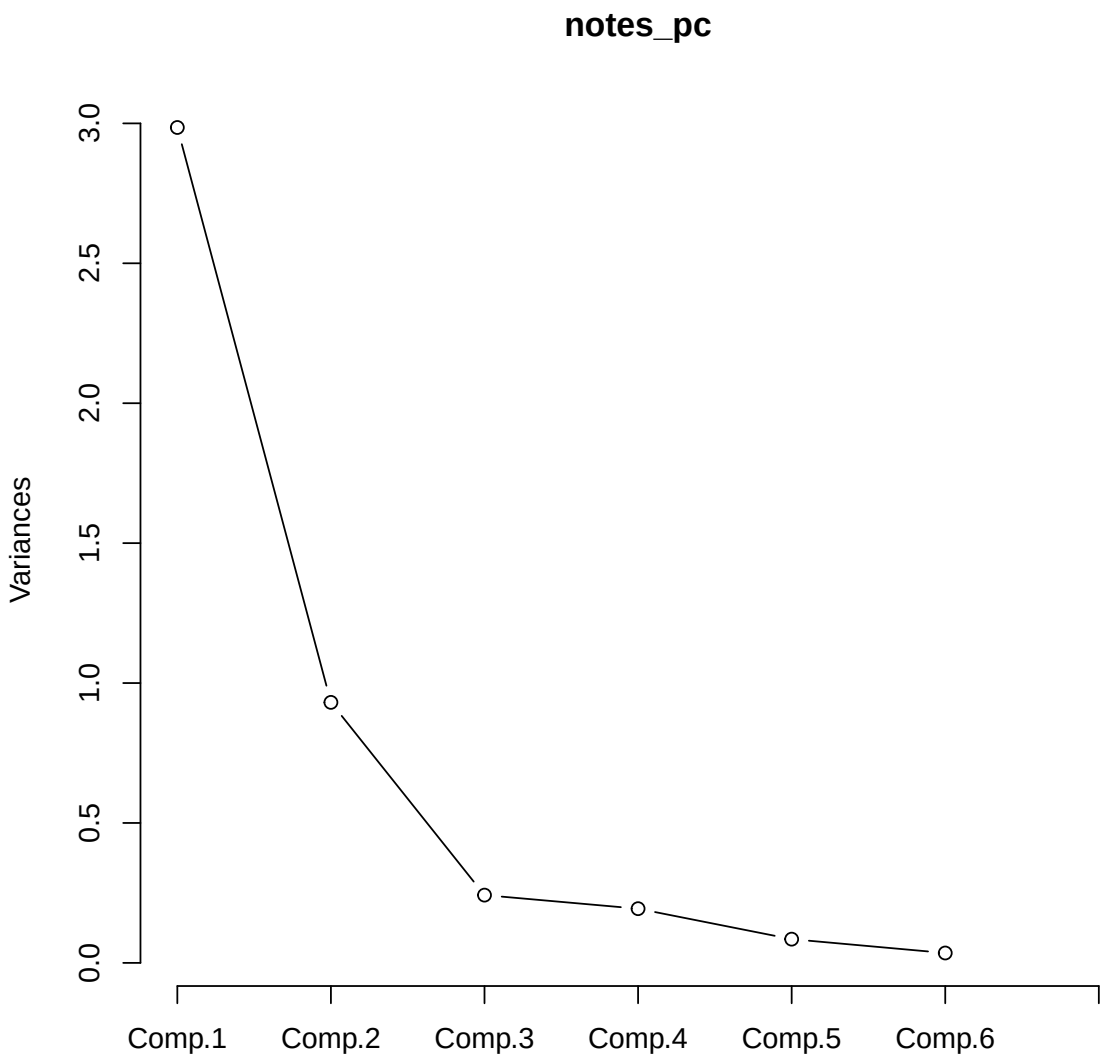
```
##                 Length   Height Height.1 Inner.Frame Inner.Frame.1 Diagonal
## Length         0.14179  0.03144  0.02309     -0.1032      -0.01854  0.08431
## Height         0.03144  0.13034  0.10843      0.2158       0.10504 -0.20934
## Height.1       0.02309  0.10843  0.16327      0.2841       0.13000 -0.24047
## Inner.Frame   -0.10325  0.21580  0.28413      2.0869       0.16454 -1.03700
## Inner.Frame.1 -0.01854  0.10504  0.13000      0.1645       0.64472 -0.54961
## Diagonal       0.08431 -0.20934 -0.24047     -1.0370      -0.54961  1.32772
```

```
notes_pc<-princomp(notes_data[,1:6])
summary(notes_pc, loadings = TRUE)
```

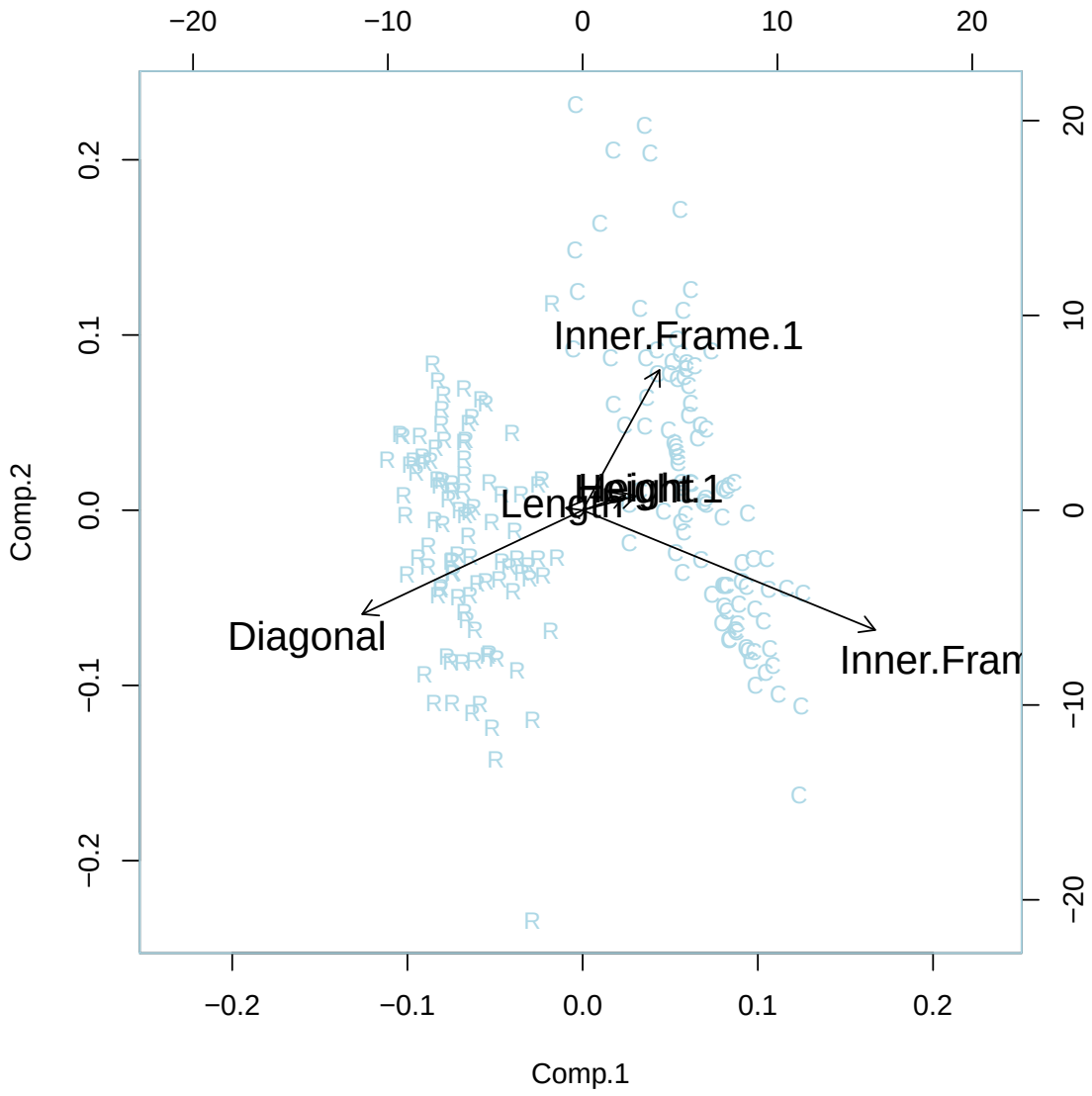
```
## Importance of components:
##                          Comp.1 Comp.2  Comp.3  Comp.4  Comp.5   Comp.6
## Standard deviation     1.7278 0.9649 0.49213 0.44010 0.29118 0.187982
## Proportion of Variance 0.6675 0.2082 0.05416 0.04331 0.01896 0.007901
## Cumulative Proportion  0.6675 0.8757 0.92983 0.97314 0.99210 1.000000
##
```

```
## Loadings:
##               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## Length                      -0.326  0.562  0.753
## Height         0.112        -0.259  0.455 -0.347 -0.767
## Height.1       0.139        -0.345  0.415 -0.535  0.632
## Inner.Frame    0.768 -0.563 -0.218 -0.186
## Inner.Frame.1  0.202  0.659 -0.557 -0.451  0.102
## Diagonal      -0.579 -0.489 -0.592 -0.258
```

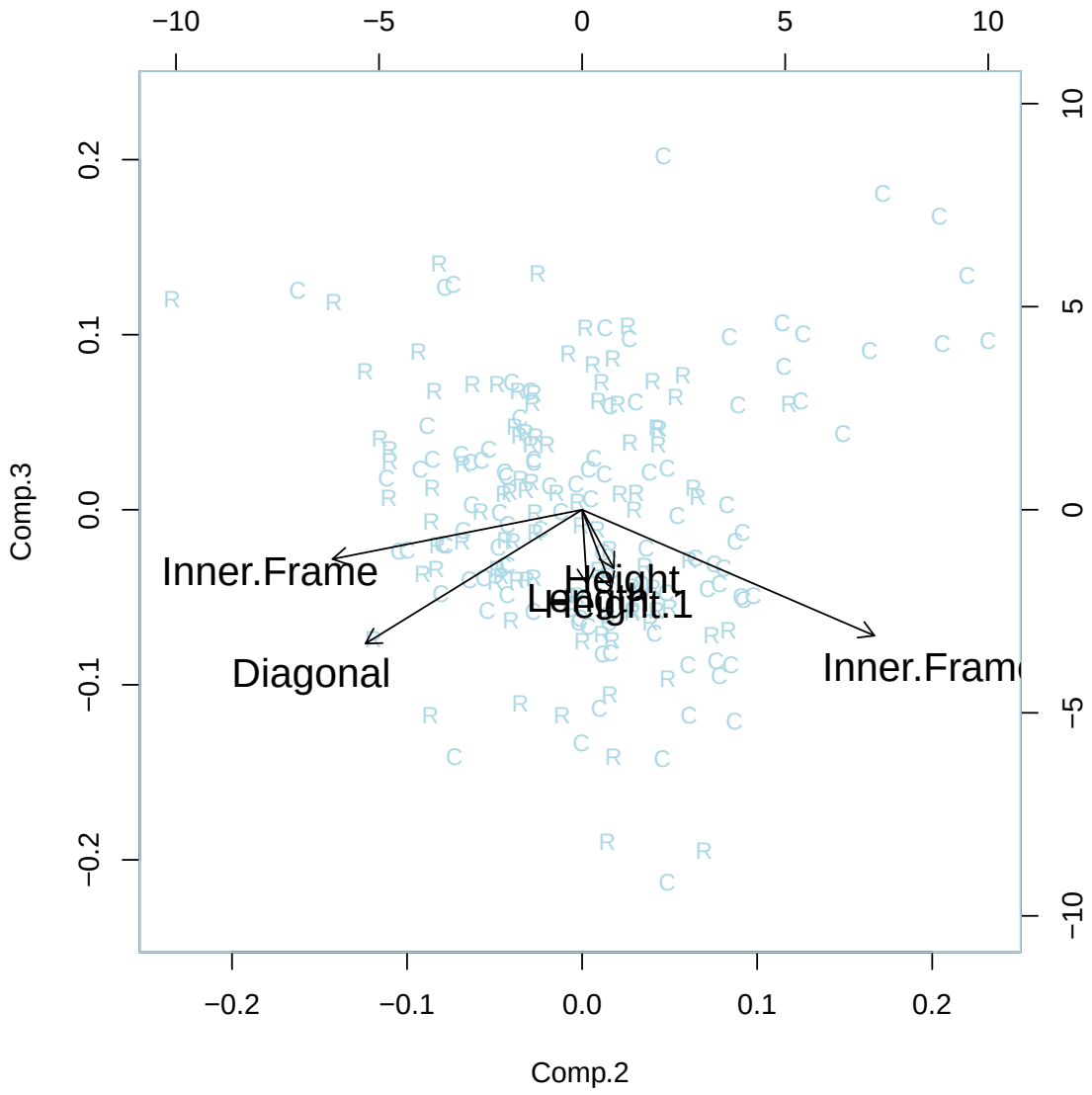
```r
screeplot(notes_pc, npcs = 7, type = "lines")
```



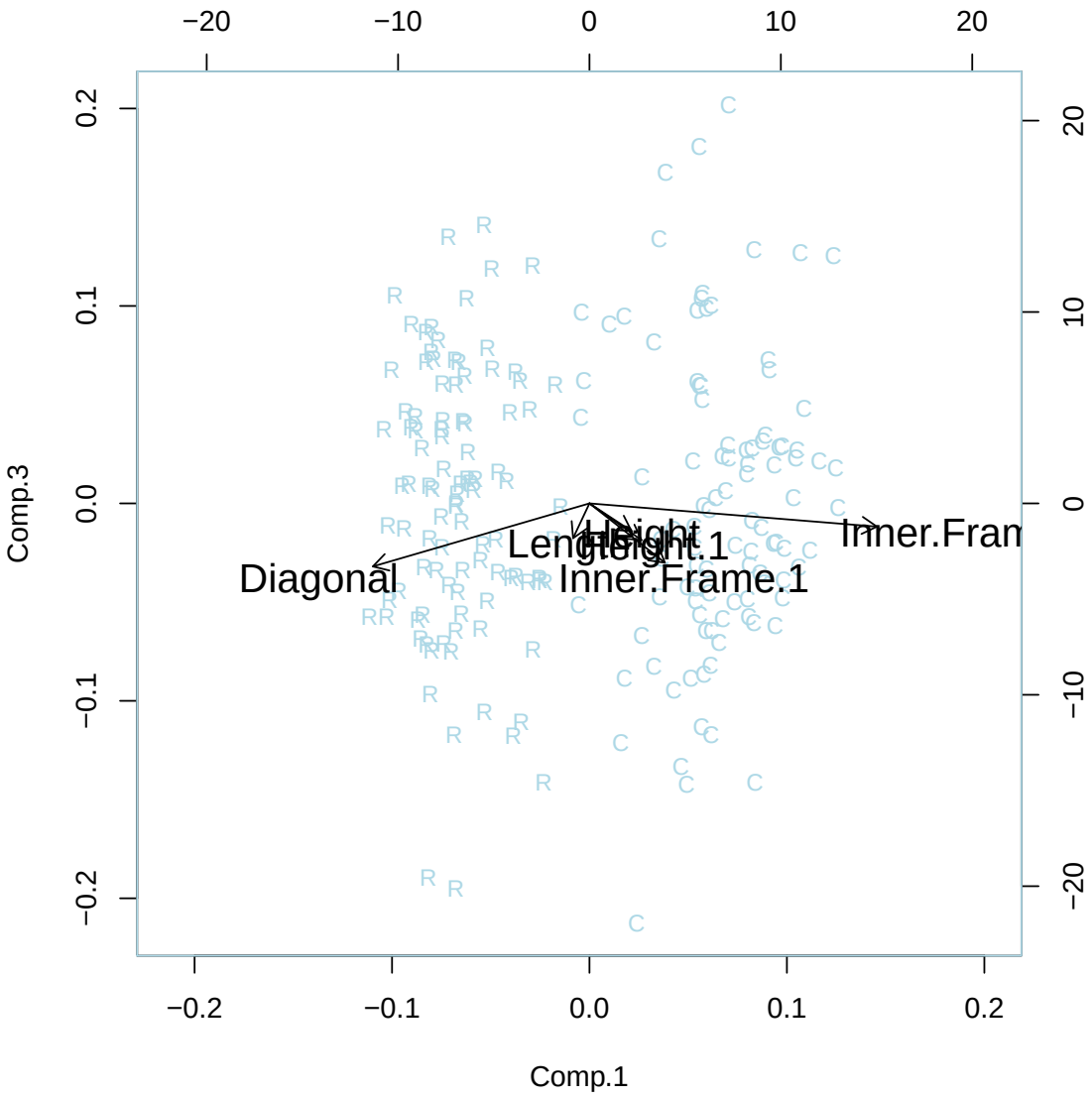
```
biplot(notes_pc,xlabs=notes_data[,8],cex=c(0.8,1.4),
       expand=0.9,col=c("lightblue","black"))
```



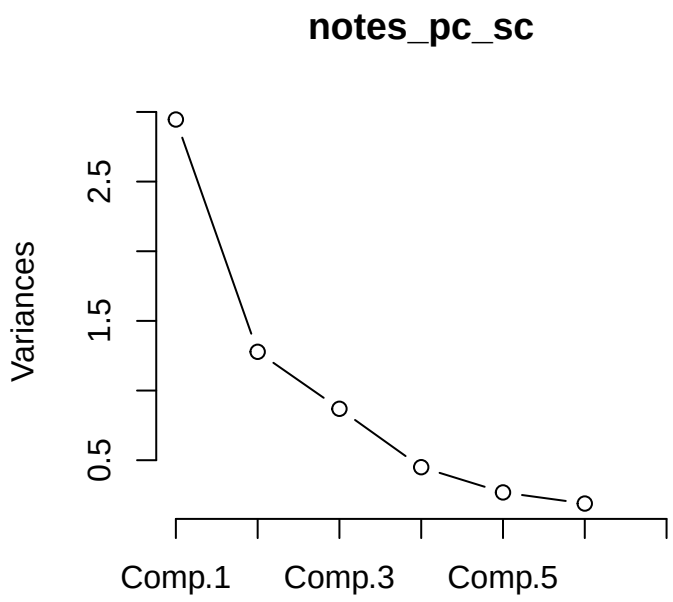
```
biplot(notes_pc,choices=2:3,xlabs=notes_data[,8],cex=c(0.8,1.4),
       expand=0.9,col=c("lightblue","black"))
```



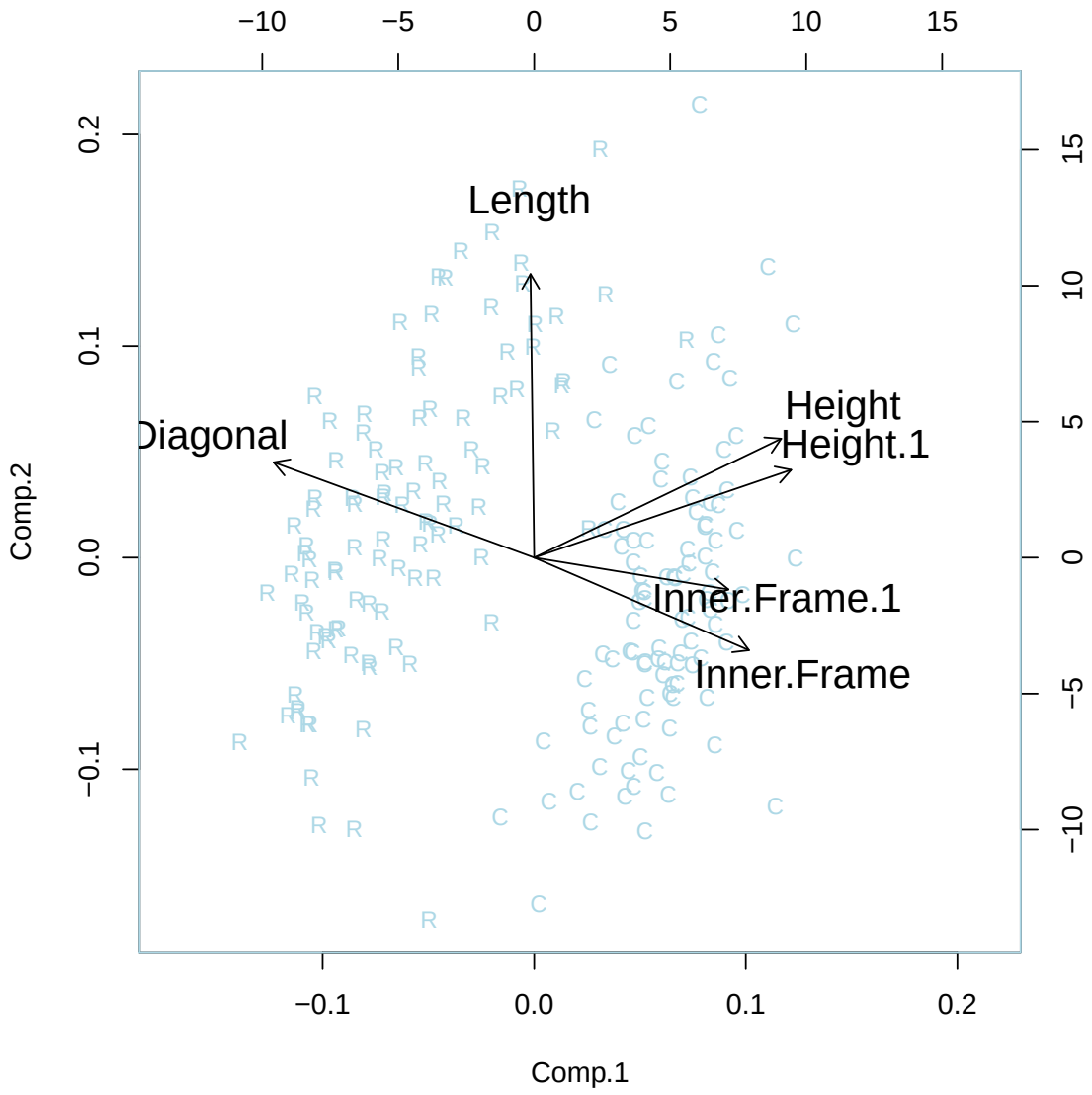
```
biplot(notes_pc,choices=c(1,3),xlabs=notes_data[,8],cex=c(0.8,1.4),
       expand=0.9,col=c("lightblue","black"))
```



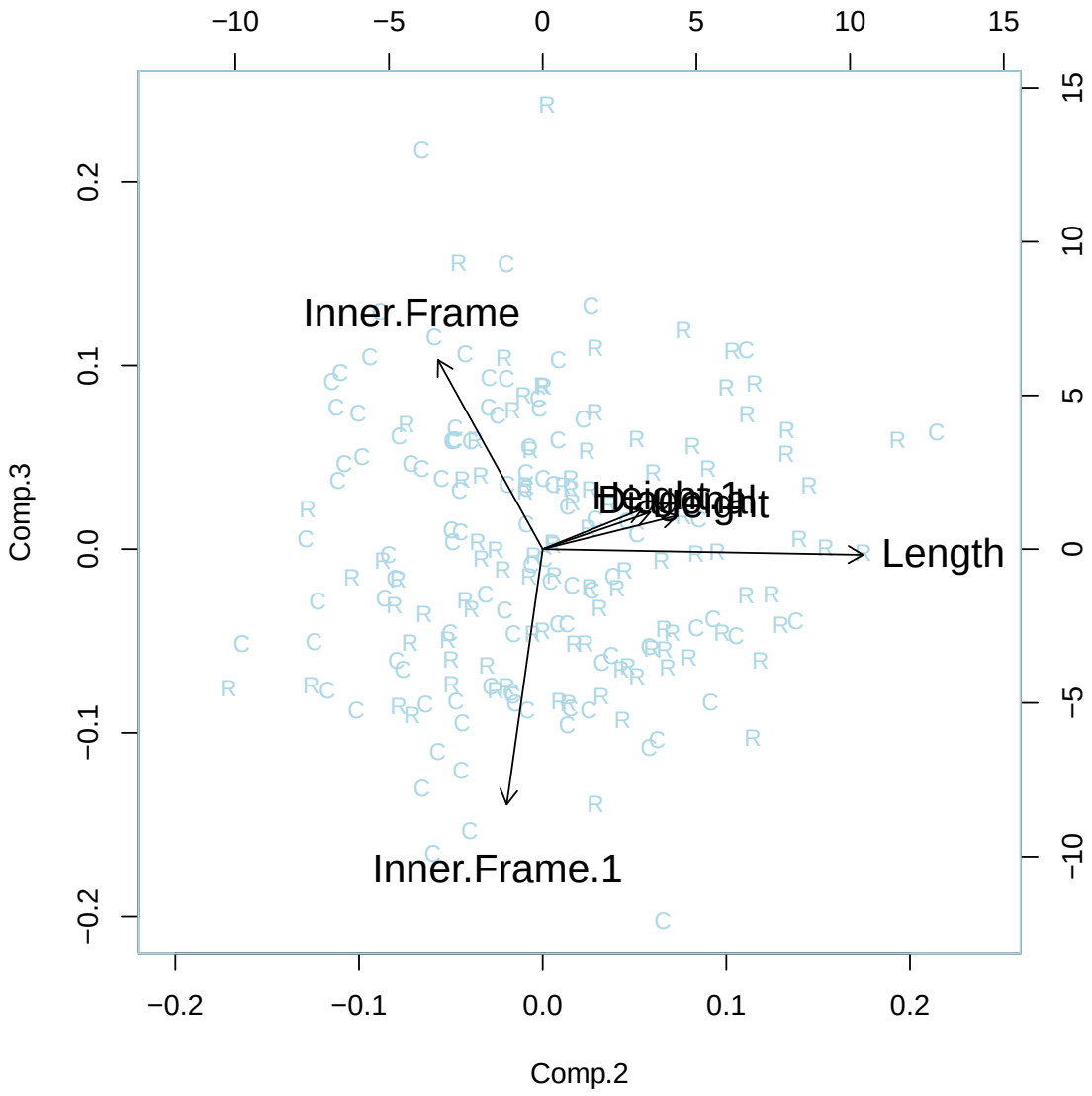
```r
notes_pc_sc<-princomp(notes_data[,1:6],cor=TRUE)
cor(notes_data[,1:6])

##                Length  Height Height.1 Inner.Frame Inner.Frame.1 Diagonal
## Length         1.00000  0.2313   0.1518     -0.1898      -0.06132   0.1943
## Height         0.23129  1.0000   0.7433      0.4138       0.36235  -0.5032
## Height.1       0.15176  0.7433   1.0000      0.4868       0.40067  -0.5165
## Inner.Frame   -0.18980  0.4138   0.4868      1.0000       0.14185  -0.6230
## Inner.Frame.1 -0.06132  0.3623   0.4007      0.1419       1.00000  -0.5940
## Diagonal       0.19430 -0.5032  -0.5165     -0.6230      -0.59404   1.0000

summary(notes_pc_sc, loadings = TRUE)

## Importance of components:
##                         Comp.1 Comp.2 Comp.3  Comp.4  Comp.5  Comp.6
## Standard deviation     1.7163 1.1305 0.9322 0.67065 0.51834 0.43460
## Proportion of Variance 0.4909 0.2130 0.1448 0.07496 0.04478 0.03148
## Cumulative Proportion  0.4909 0.7039 0.8488 0.92374 0.96852 1.00000
##
## Loadings:
##               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## Length                0.815         0.575
## Height         0.468  0.342  0.103 -0.395  0.639  0.298
## Height.1       0.487  0.252  0.123 -0.430 -0.614 -0.349
## Inner.Frame    0.407 -0.266  0.584  0.404 -0.215  0.462
## Inner.Frame.1  0.368        -0.788  0.110 -0.220  0.419
## Diagonal      -0.493  0.274  0.114 -0.392 -0.340  0.632

screeplot(notes_pc_sc, npcs = 7, type = "lines")
```



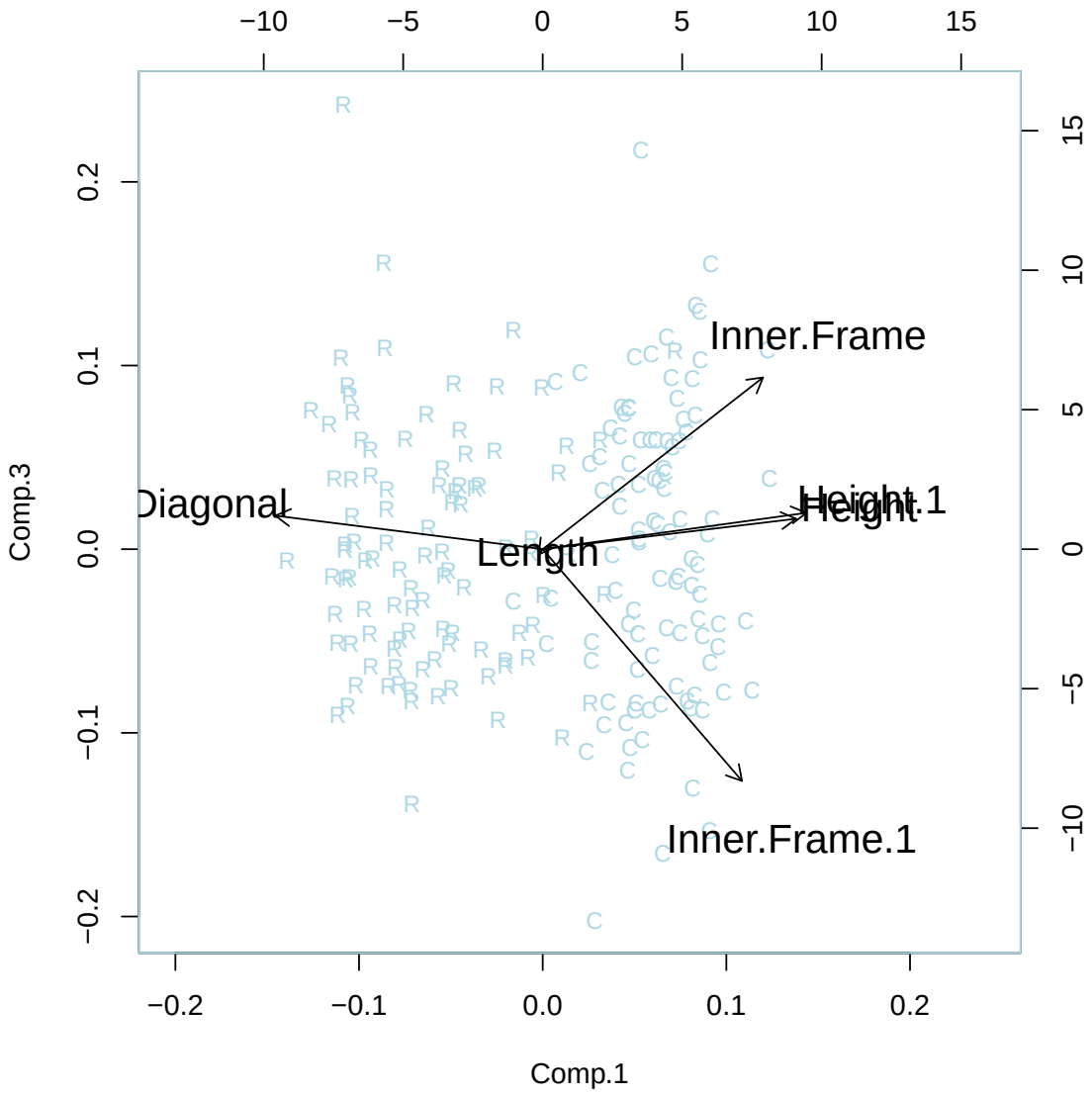
```
biplot(notes_pc_sc,xlabs=notes_data[,8],cex=c(0.8,1.4),
       expand=0.9,col=c("lightblue","black"))
```



```
biplot(notes_pc_sc,choices=2:3,xlabs=notes_data[,8],cex=c(0.8,1.4),
       expand=0.9,col=c("lightblue","black"))
```



```
biplot(notes_pc_sc,choices=c(1,3),xlabs=notes_data[,8],cex=c(0.8,1.4),
       expand=0.9,col=c("lightblue","black"))
```



Boston Data

Harrison and Rubenfeld (1978) collected data to determine whether clean air had any influence on house prices. The following variables were collected.

- CRIM: per capita crime rate by town
- ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: nitric oxides concentration (parts per 10 million)
- RM: average number of rooms per dwelling
- AGE: proportion of owner-occupied units built prior to 1940
- DIS: weighted distances to five Boston employment centres
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per $\$10,000$
- PTRATIO: pupil-teacher ratio by town
- B: $1000(Bk - 0.63)^2$ where Bk is the proportion of African Americans by town
- LSTAT: % lower status of the population
- MEDV: Median value of owner-occupied homes in \$1000's

We consider the variables after Box–Cox transfomations (see Seminar 1). We do not analyse the fourth variable (as it is binary).

```
boston <- read.csv("C:/Documents/boston.csv") boston<-
boston[,-4]
boston$CRIM<-log(boston$CRIM)
boston$ZN<-boston$ZN/10
boston$INDUS<-log(boston$INDUS)
boston$NOX<-log(boston$NOX)
boston$RM<-log(boston$RM)
boston$AGE<-(boston$AGE)^2.5/10000
boston$DIS<-log(boston$DIS)
boston$RAD<-log(boston$RAD)
boston$TAX<-log(boston$TAX)
boston$PTRATIO<-exp(0.4*boston$PTRATIO)/1000
```

```r
boston$B<-boston$B/1000
boston$LSTAT<-sqrt(boston$LSTAT)
boston$MEDV<-log(boston$MEDV)
```

```r
boston_pc<-princomp(boston,cor=TRUE)
cor(boston)
```

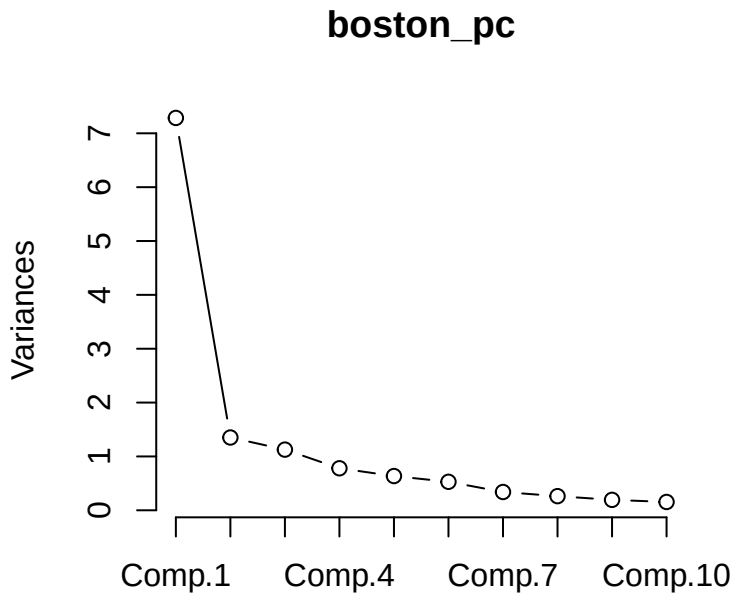
```
##            CRIM      ZN   INDUS     NOX      RM     AGE     DIS     RAD     TAX PTRATIO       B
## CRIM     1.0000 -0.5171  0.7396  0.8070 -0.3242  0.6968 -0.7439  0.8389  0.8100  0.4539 -0.4788
## ZN      -0.5171  1.0000 -0.6559 -0.5685  0.3094 -0.5263  0.5907 -0.3506 -0.3059 -0.3501  0.1755
## INDUS    0.7396 -0.6559  1.0000  0.7505 -0.4296  0.6581 -0.7303  0.5805  0.6593  0.4547 -0.3311
## NOX      0.8070 -0.5685  0.7505  1.0000 -0.3183  0.7831 -0.8600  0.6129  0.6683  0.3437 -0.3793
## RM      -0.3242  0.3094 -0.4296 -0.3183  1.0000 -0.2767  0.2807 -0.2134 -0.3064 -0.3208  0.1297
## AGE      0.6968 -0.5263  0.6581  0.7831 -0.2767  1.0000 -0.7960  0.4687  0.5409  0.3778 -0.2859
## DIS     -0.7439  0.5907 -0.7303 -0.8600  0.2807 -0.7960  1.0000 -0.5421 -0.5996 -0.3217  0.3248
## RAD      0.8389 -0.3506  0.5805  0.6129 -0.2134  0.4687 -0.5421  1.0000  0.8205  0.3982 -0.4113
## TAX      0.8100 -0.3059  0.6593  0.6683 -0.3064  0.5409 -0.5996  0.8205  1.0000  0.4763 -0.4279
## PTRATIO  0.4539 -0.3501  0.4547  0.3437 -0.3208  0.3778 -0.3217  0.3982  0.4763  1.0000 -0.2047
## B       -0.4788  0.1755 -0.3311 -0.3793  0.1297 -0.2859  0.3248 -0.4113 -0.4279 -0.2047  1.0000
## LSTAT    0.6223 -0.4522  0.6214  0.6094 -0.6394  0.6371 -0.5555  0.4612  0.5335  0.4338 -0.3610
## MEDV    -0.5672  0.3633 -0.5539 -0.5153  0.6104 -0.4821  0.4057 -0.4345 -0.5572 -0.5082  0.4024
##           LSTAT    MEDV
## CRIM     0.6223 -0.5672
## ZN      -0.4522  0.3633
## INDUS    0.6214 -0.5539
## NOX      0.6094 -0.5153
## RM      -0.6394  0.6104
## AGE      0.6371 -0.4821
## DIS     -0.5555  0.4057
## RAD      0.4612 -0.4345
## TAX      0.5335 -0.5572
## PTRATIO  0.4338 -0.5082
## B       -0.3610  0.4024
## LSTAT    1.0000 -0.8250
## MEDV    -0.8250  1.0000
```

```r
summary(boston_pc, loadings = TRUE)
```

```
## Importance of components:
##                          Comp.1 Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8 Comp.9 Comp.10
## Standard deviation     2.6991 1.1626 1.06143 0.88332 0.79742 0.72734 0.58282 0.51264 0.4400  0.3933
## Proportion of Variance 0.5604 0.1040 0.08666 0.06002 0.04891 0.04069 0.02613 0.02021 0.0149  0.0119
## Cumulative Proportion  0.5604 0.6644 0.75105 0.81107 0.85998 0.90067 0.92680 0.94702 0.9619  0.9738
##                          Comp.11 Comp.12  Comp.13
## Standard deviation     0.37480 0.33163 0.299950
## Proportion of Variance 0.01081 0.00846 0.006921
## Cumulative Proportion  0.98462 0.99308 1.000000
##
## Loadings:
##         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12
## CRIM     0.336  0.193  0.137         0.108                0.205        -0.210   0.326   0.239
## ZN      -0.237         0.476         0.441  0.519  0.320 -0.195        -0.285
## INDUS    0.318        -0.174               -0.248  0.164 -0.717  0.411 -0.253   0.113
## NOX      0.324  0.206 -0.168 -0.136  0.104  0.122  0.129        -0.529 -0.154   0.392  -0.525
## RM      -0.189  0.605               -0.288  0.321 -0.531 -0.290        -0.111  -0.115
## AGE      0.296  0.134 -0.278 -0.114         0.521         0.228  0.501  0.367   0.252
## DIS     -0.306 -0.250  0.281  0.121        -0.156 -0.299         0.298  0.130   0.567  -0.450
## RAD      0.279  0.246  0.358  0.223  0.249 -0.289 -0.217  0.331  0.127 -0.230
## TAX      0.301  0.142  0.346  0.196  0.268               -0.247         0.632  -0.302  -0.261
## PTRATIO  0.210 -0.229  0.141  0.684 -0.468  0.320  0.241  0.102        -0.136
## B       -0.182        -0.487  0.562  0.576        -0.228
## LSTAT    0.296 -0.366        -0.211         0.180 -0.361         0.209 -0.354  -0.440  -0.416
## MEDV    -0.273  0.444 -0.165  0.110        -0.146  0.427  0.275  0.361 -0.163  -0.194  -0.436
##         Comp.13
## CRIM     0.743
## ZN
```

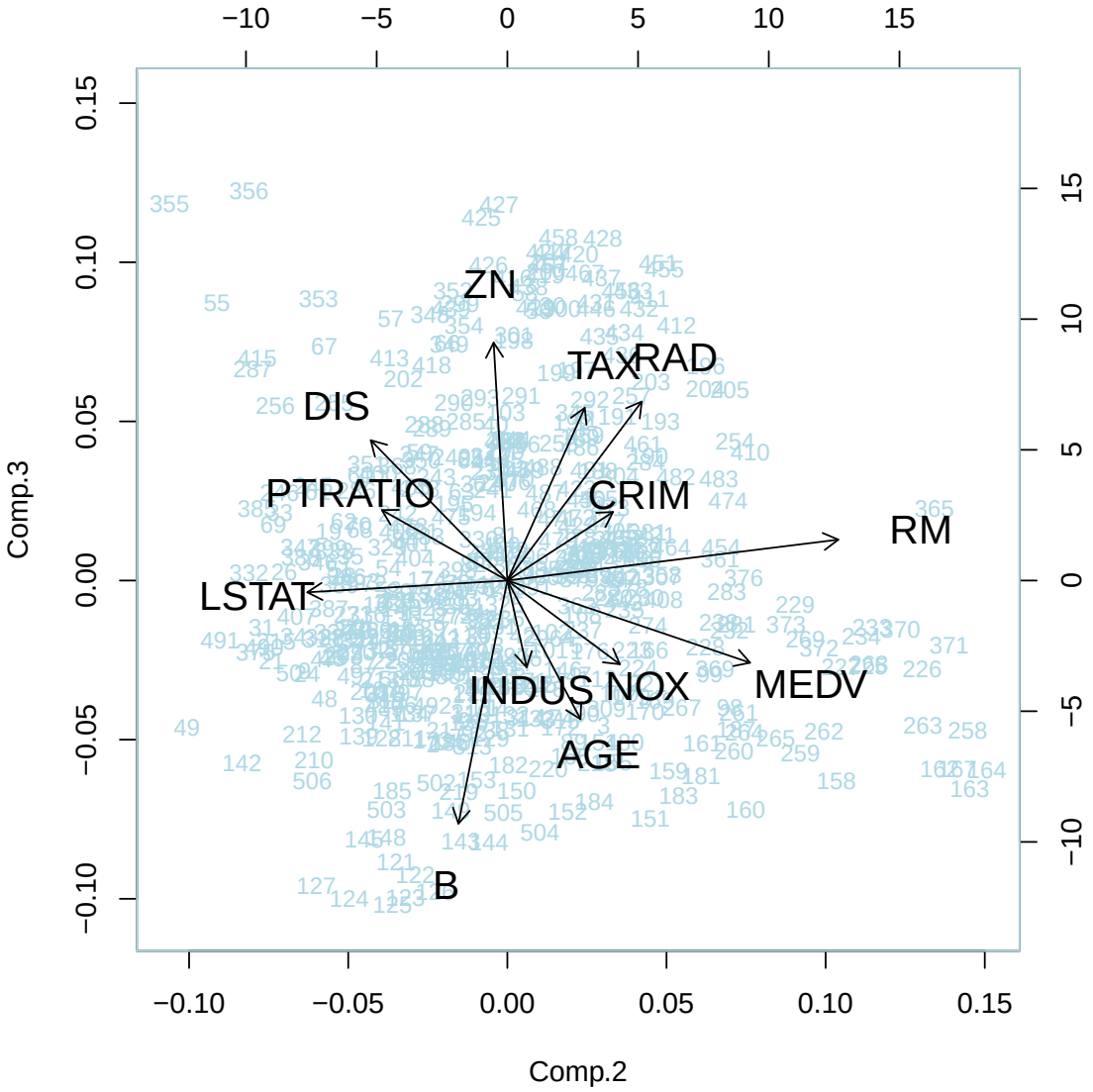
```
## INDUS
## NOX     -0.167
## RM
## AGE     -0.160
## DIS
## RAD     -0.551
## TAX      0.178
## PTRATIO
## B
## LSTAT    0.148
## MEDV     0.157
```

```
screeplot(boston_pc, type = "lines")
```

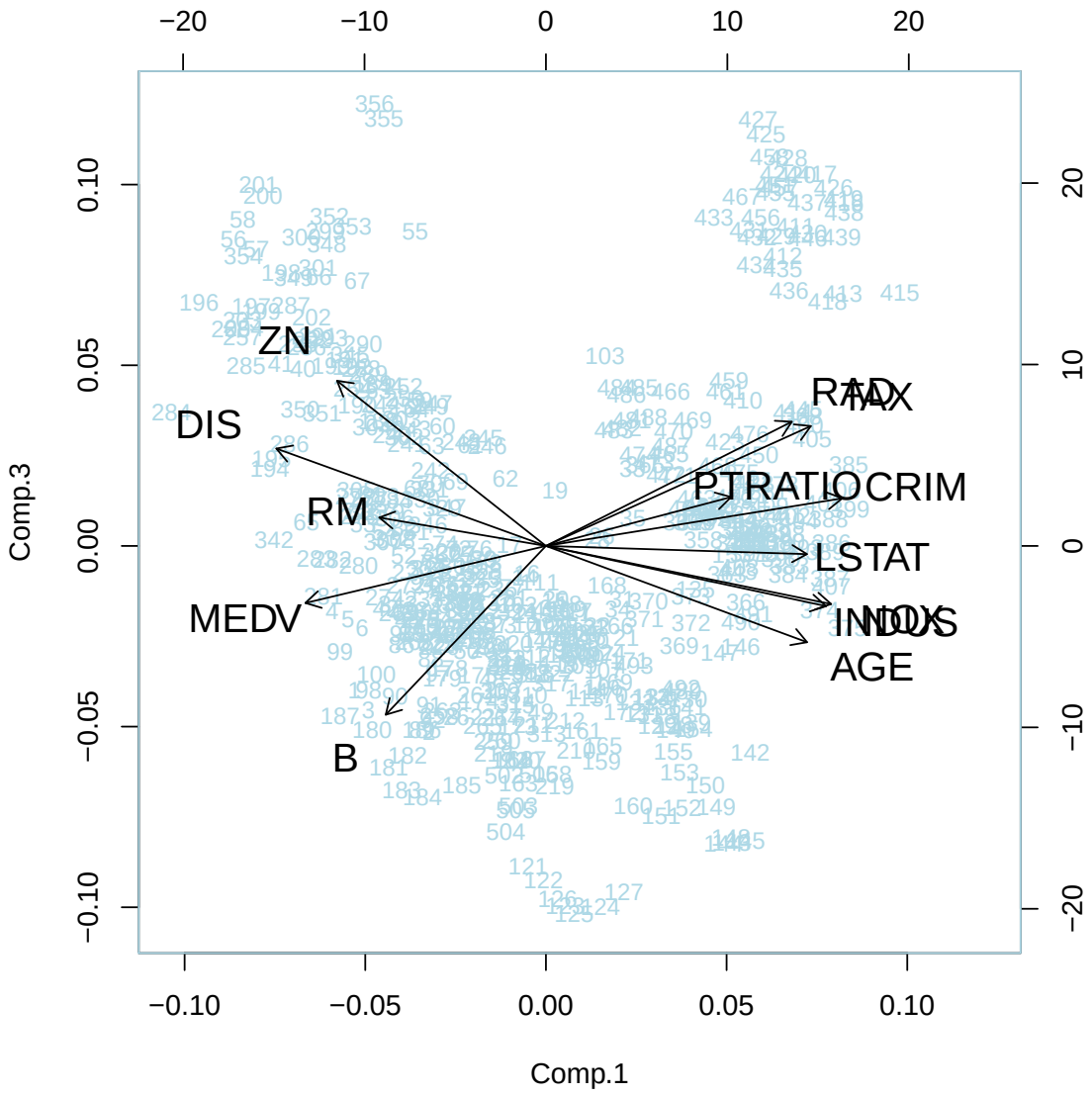


```
biplot(boston_pc,cex=c(0.8,1.4),
       expand=0.9,col=c("lightblue","black"))
```

```
biplot(boston_pc,choices=2:3,cex=c(0.8,1.4),
       expand=0.9,col=c("lightblue","black"))
```



```
biplot(boston_pc,choices=c(1,3),cex=c(0.8,1.4),
       expand=0.9,col=c("lightblue","black"))
```



In Class Exercise 1 - Turtles

Jolicoeur and Mosimann measures the carapace length (length of top of shell), width and height of painted turtles, which are native to north America. This data can be found in Turtles.csv. Perform a principal components analysis to determine whether the information in this data set can be summarised in fewer than 3 variables. Use the gender column to label the biplots - can you distinguish male and female turtles?

In Class Exercise 2 - Activities

The dataset Activities.csv contains data from 28 individuals, measuring the amount of time (in hours) spent on 10 activities over 100 days, as well as some demographic information. Perorm a principal components analysis on the 10 variables related to time spent on activities.

- prof: professional activity
- tran: transportation linked to professional activity
- hous: household occupation
- kids: occupation linked to children
- shop: shopping
- pers: time spent for personal care
- eat: eating
- slee: sleeping
- tele: watching television
- leis: other leisure activities

Exercises

Johnson and Wichern Exercises 8.18, 8.19, 8.20, 8.24, 8.28