```r
# pca.r
d1=USArrests
dim(d1)             # [1] 50  4
```

```
## [1] 50  4
```

```r
head(d1)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```r
#
# a) Original data -no scaling-
#
# covariance matrix
var(d1)
```

```
##              Murder   Assault   UrbanPop      Rape
## Murder    18.970465  291.0624   4.386204  22.99141
## Assault  291.062367 6945.1657 312.275102 519.26906
## UrbanPop   4.386204  312.2751 209.518776  55.76808
## Rape      22.991412  519.2691  55.768082  87.72916
```

```r
apply(d1,2,var)
```

```
##     Murder    Assault   UrbanPop       Rape
##   18.97047 6945.16571  209.51878   87.72916
```

```r
#
# column variances appear on main diagonal of cov matrix
#
# eigenvalues
#
eigen(var(d1))
```

```
## eigen() decomposition
## $values
## [1] 7011.114851  201.992366   42.112651    6.164246
##
## $vectors
##             [,1]        [,2]        [,3]        [,4]
## [1,] -0.04170432  0.04482166  0.07989066  0.99492173
## [2,] -0.99522128  0.05876003 -0.06756974 -0.03893830
## [3,] -0.04633575 -0.97685748 -0.20054629  0.05816914
## [4,] -0.07515550 -0.20071807  0.97408059 -0.07232502
```

```r
#
sum(eigen(var(d1))$values)
```

```
## [1] 7261.384
```

```r
sum(apply(d1,2,var))
```

```
## [1] 7261.384
```

```r
#
# sum of eigenvalues = sum variances
```

```
#
# b) SCALED DATA
#
m1=prcomp(d1, scale=T)
names(m1)
```

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

```
#
# "sdev": square-root of eigenvalues of columns of transformed data (of PCs)
# "rotation": matrix of eigenvectors
# "center"   "scale": # mean and sd of original data -unscaled-
# "x": transformed data set
#
# means -unscaled data
#
m1$center
```

```
##   Murder  Assault UrbanPop     Rape
##    7.788  170.760   65.540   21.232
```

```
#
# standard deviations -unscaled data
#
m1$scale
```

```
##    Murder   Assault  UrbanPop      Rape
##  4.355510 83.337661 14.474763  9.366385
```

```
apply(d1,2,sd)
```

```
##    Murder   Assault  UrbanPop      Rape
##  4.355510 83.337661 14.474763  9.366385
```

```
#
# rotation matrix has the eigenvectors
#
m1$rotation
```

```
##                 PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

```
#
eigen(var(scale(d1)))
```

```
## eigen() decomposition
## $values
## [1] 2.4802416 0.9897652 0.3565632 0.1734301
##
## $vectors
##            [,1]       [,2]       [,3]        [,4]
## [1,] -0.5358995  0.4181809 -0.3412327  0.64922780
## [2,] -0.5831836  0.1879856 -0.2681484 -0.74340748
## [3,] -0.2781909 -0.8728062 -0.3780158  0.13387773
## [4,] -0.5434321 -0.1673186  0.8177779  0.08902432
```

```
#
# squared-root of eigenvalues of scaled data
```

```
m1$sdev
```

```
## [1] 1.5748783 0.9948694 0.5971291 0.4164494
```

```
#
# transformed data on PC axes
d2 = m1$x
dim(d2)
```

```
## [1] 50  4
```

```
head(d2)
```

```
##                  PC1        PC2         PC3          PC4
## Alabama    -0.9756604  1.1220012 -0.43980366  0.154696581
## Alaska     -1.9305379  1.0624269  2.01950027 -0.434175454
## Arizona    -1.7454429 -0.7384595  0.05423025 -0.826264240
## Arkansas    0.1399989  1.1085423  0.11342217 -0.180973554
## California -2.4986128 -1.5274267  0.59254100 -0.338559240
## Colorado   -1.4993407 -0.9776297  1.08400162  0.001450164
```

```
apply(d2,2,mean)
```

```
##           PC1           PC2           PC3           PC4
##  3.695720e-17  3.619582e-16  2.375205e-16 -1.916184e-16
```

```
#
# PCs are centered (their means are zero)
#
# eigenvalues of scaled data
#
m1$sdev^2
```

```
## [1] 2.4802416 0.9897652 0.3565632 0.1734301
```

```
apply(d2,2,var)
```

```
##       PC1       PC2       PC3       PC4
## 2.4802416 0.9897652 0.3565632 0.1734301
```

```
#
# eigenvalues of cov matrix (of scaled data) = variances of Principal components
#
# eigen() function
#
cova=var(scale(d1))

m2 = eigen(cova)

# covariance matrix of transformed data
var(d2)
```

```
##               PC1           PC2           PC3           PC4
## PC1  2.480242e+00 -3.812359e-16  1.126674e-16  1.778258e-17
## PC2 -3.812359e-16  9.897652e-01 -2.024956e-16  9.907132e-17
## PC3  1.126674e-16 -2.024956e-16  3.565632e-01 -1.564569e-16
## PC4  1.778258e-17  9.907132e-17 -1.564569e-16  1.734301e-01
```

```
round(var(d2),5)
```

```
##         PC1     PC2     PC3     PC4
## PC1 2.48024 0.00000 0.00000 0.00000
```

```
## PC2 0.00000 0.98977 0.00000 0.00000
## PC3 0.00000 0.00000 0.35656 0.00000
## PC4 0.00000 0.00000 0.00000 0.17343
```

```
#
# This is Big lambda diagonal matrix (eigenvalues on main diagonal)
sum(diag(var(d2)))      # 4
```

```
## [1] 4
```

```
# covariances (off diagonal) all equal to 0  (PCs uncorrelated)
# PC1 with largest variance across states

# Use eigenvectors to define the PC variables.

m1$rotation
```

```
##                 PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

```
#
# Score vectors are PC1, PC2, defined as follows

# PC1 = 0.536 Murder + 0.58Assault + 0.28 UrbanPop + 0.543 Rape
# A weighted average of crime rates (almost exclude UrbanPop)

# PC2 = 0.4 Murder - 0.87 UrbanPop
# Weighted average of Urban Pop and Murder


# transformed variables in the principal components space.
#============================================================
# eigenvectors span a new p-dimensional space
# score vectors are the transformed data in this new space
d2 = m1$x
head(d2)
```

```
##                   PC1        PC2         PC3          PC4
## Alabama    -0.9756604  1.1220012 -0.43980366  0.154696581
## Alaska     -1.9305379  1.0624269  2.01950027 -0.434175454
## Arizona    -1.7454429 -0.7384595  0.05423025 -0.826264240
## Arkansas    0.1399989  1.1085423  0.11342217 -0.180973554
## California -2.4986128 -1.5274267  0.59254100 -0.338559240
## Colorado   -1.4993407 -0.9776297  1.08400162  0.001450164
```

```
tail(m1$x)
```

```
##                    PC1        PC2         PC3        PC4
## Vermont       2.7732561  1.3881944  0.83280797 -0.1434337
## Virginia      0.0953667  0.1977278  0.01159482  0.2092464
## Washington    0.2147234 -0.9603739  0.61859067 -0.2186282
## West Virginia 2.0873931  1.4105263  0.10372163  0.1305831
## Wisconsin     2.0588120 -0.6051251 -0.13746933  0.1822534
## Wyoming       0.6231006  0.3177866 -0.23824049 -0.1649769
```

```
#
# Variance of the PCs are the eigenvalues
```

```
#============================================================

apply(d2,2,var)
```

```
##       PC1       PC2       PC3       PC4
## 2.4802416 0.9897652 0.3565632 0.1734301
```

```
m2$values
```

```
## [1] 2.4802416 0.9897652 0.3565632 0.1734301
```

```
#
# proportion of variance explained (PVE) by each PC
#============================================================

# variance of PCs
aux=m1$sdev^2

sum(aux)   # 4
```

```
## [1] 4
```

```
pve=aux/sum(aux)
pve
```

```
## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

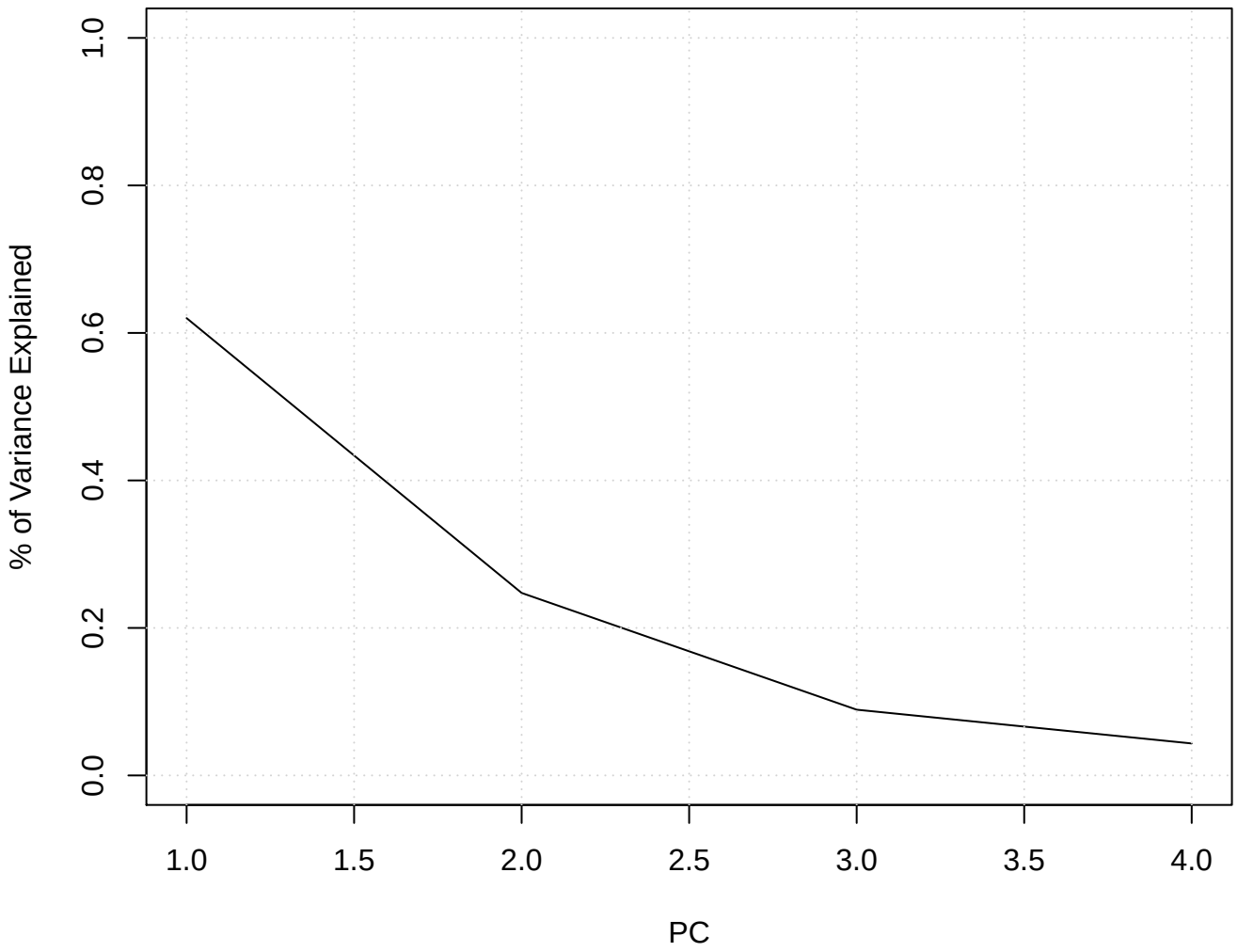
```
m2$values/4
```

```
## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

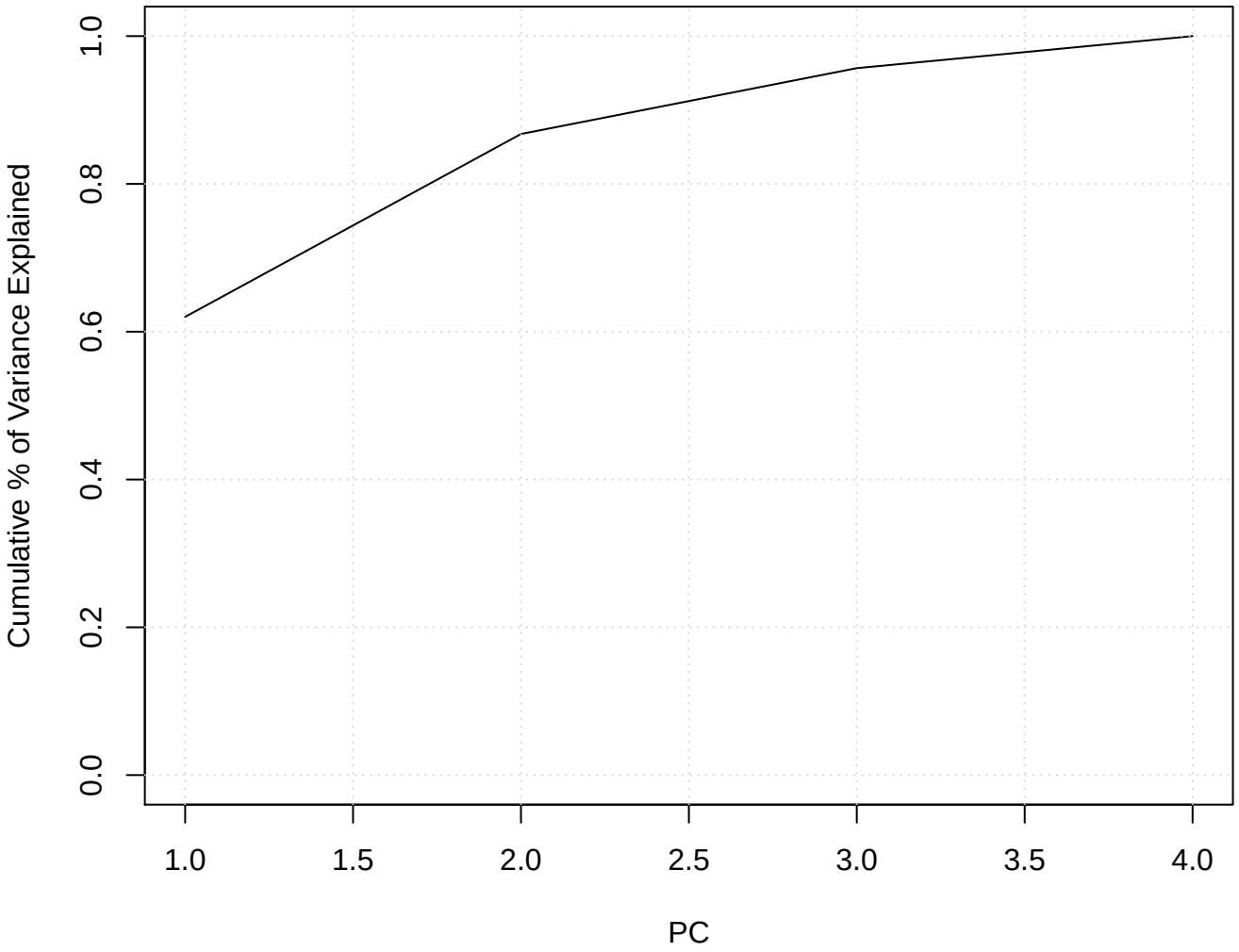
```
# each eigenvalue divided by 4
#
cumsum(pve)
```

```
## [1] 0.6200604 0.8675017 0.9566425 1.0000000
```

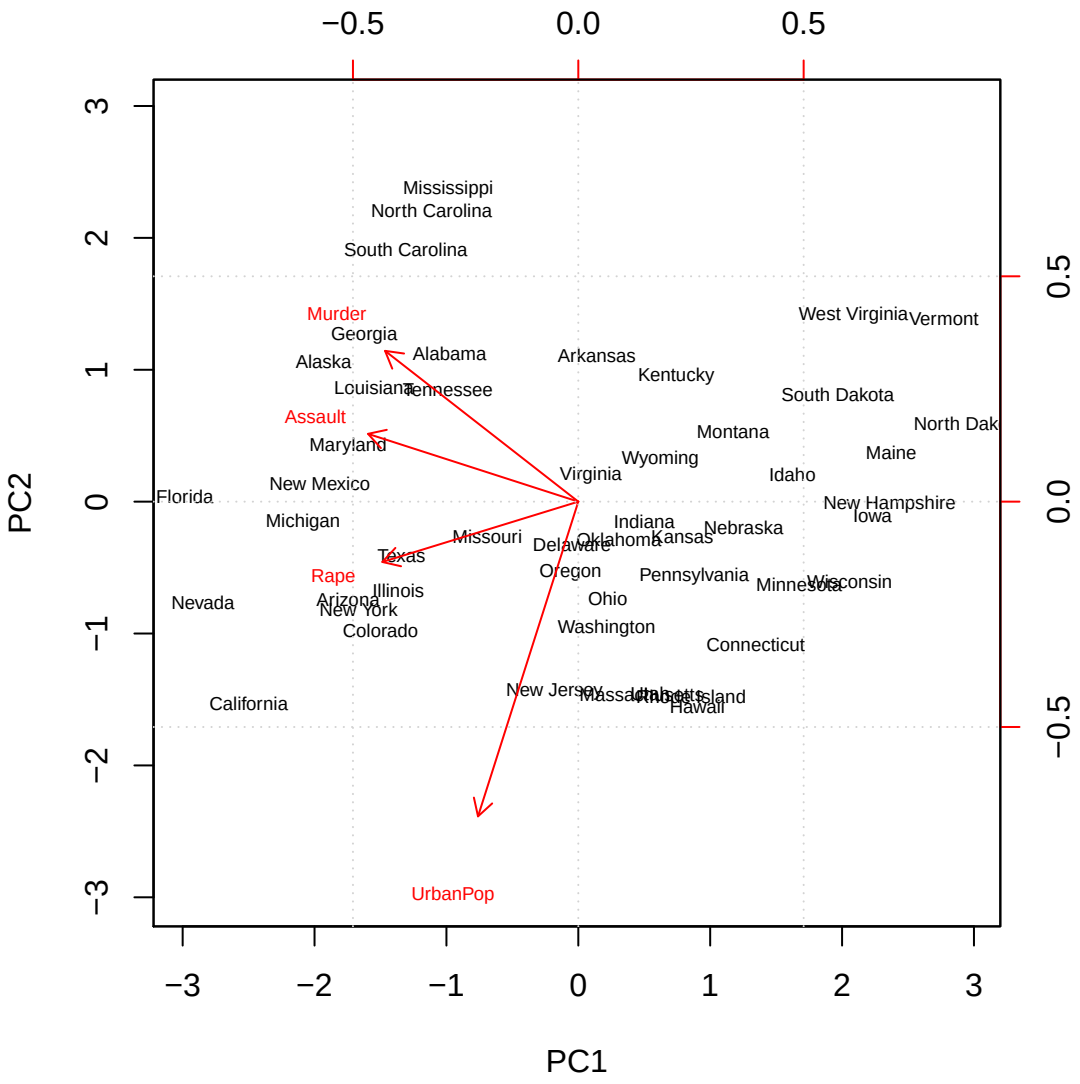
```
#
#  87% variability in the dataset explained by PC1 and PC2

# plots
plot(pve, xlab="PC", ylab="% of Variance Explained", ylim=c(0,1),type='l')
grid()
```




```
plot(cumsum(pve), xlab="PC", ylab="Cumulative % of Variance Explained", ylim=c(0,1),type='l')
grid()
```




```
#
# biplots
#
biplot(m1, scale=0,cex=0.6)
grid()
```




```
#
head(d2)
```

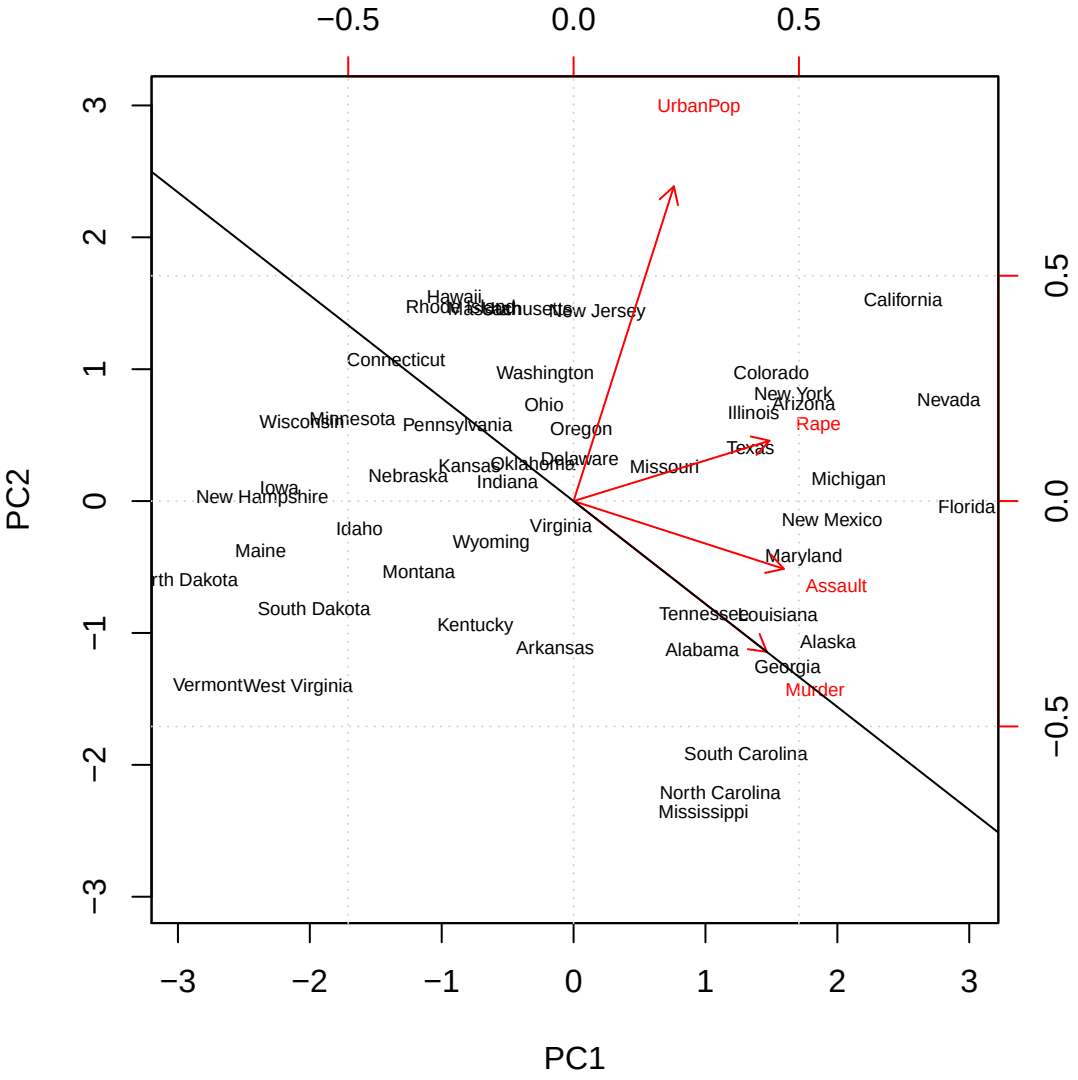
```
##                   PC1        PC2         PC3          PC4
## Alabama    -0.9756604  1.1220012 -0.43980366  0.154696581
## Alaska     -1.9305379  1.0624269  2.01950027 -0.434175454
## Arizona    -1.7454429 -0.7384595  0.05423025 -0.826264240
## Arkansas    0.1399989  1.1085423  0.11342217 -0.180973554
## California -2.4986128 -1.5274267  0.59254100 -0.338559240
## Colorado   -1.4993407 -0.9776297  1.08400162  0.001450164
```

```
#
# rowname is State name, located at (PC1,PC2) coordinates
#
# Rotate original axes (red colored)
#
# mirror image
#
m1$rotation=-m1$rotation
m1$x=-m1$x
biplot(m1, scale=0,cex=0.6)
grid()
#
```

```
rot=m1$rotation
#
# Murder axis identified
#
slope1=rot[1,2]/rot[1,1]
slope1   # -0.7803345
```

```
## [1] -0.7803345
```

```
abline(0,slope1)
```




```
# interpret the PCs
#============================================================
m1$rotation
```

```
##                PC1        PC2        PC3         PC4
## Murder   0.5358995 -0.4181809  0.3412327 -0.64922780
## Assault  0.5831836 -0.1879856  0.2681484  0.74340748
## UrbanPop 0.2781909  0.8728062  0.3780158 -0.13387773
## Rape     0.5434321  0.1673186 -0.8177779 -0.08902432
```

```
# states with large values in PC1 have high crime rates
#       (PC1 weights -col1- in rotation are 0.5359, 0.5831, 0.5434)
# California, Nevada, Florida   vs   North Dakota, Vermont
```

```r
# states with large values in PC2 have large urban areas
#       (PC2 largest weight -col2- in rotation is 0.8728)
# California  vs   Mississippi

# original vs transformed values
#============================================================
d3=data.frame(d1,d2)
head(d3)
```

```
##            Murder Assault UrbanPop Rape        PC1        PC2         PC3
## Alabama      13.2     236       58 21.2 -0.9756604  1.1220012 -0.43980366
## Alaska       10.0     263       48 44.5 -1.9305379  1.0624269  2.01950027
## Arizona       8.1     294       80 31.0 -1.7454429 -0.7384595  0.05423025
## Arkansas      8.8     190       50 19.5  0.1399989  1.1085423  0.11342217
## California    9.0     276       91 40.6 -2.4986128 -1.5274267  0.59254100
## Colorado      7.9     204       78 38.7 -1.4993407 -0.9776297  1.08400162
##                     PC4
## Alabama     0.154696581
## Alaska     -0.434175454
## Arizona    -0.826264240
## Arkansas   -0.180973554
## California -0.338559240
## Colorado    0.001450164
```

```r
tail(d3)
```

```
##               Murder Assault UrbanPop Rape       PC1        PC2         PC3
## Vermont          2.2      48       32 11.2 2.7732561  1.3881944  0.83280797
## Virginia         8.5     156       63 20.7 0.0953667  0.1977278  0.01159482
## Washington       4.0     145       73 26.2 0.2147234 -0.9603739  0.61859067
## West Virginia    5.7      81       39  9.3 2.0873931  1.4105263  0.10372163
## Wisconsin        2.6      53       66 10.8 2.0588120 -0.6051251 -0.13746933
## Wyoming          6.8     161       60 15.6 0.6231006  0.3177866 -0.23824049
##                      PC4
## Vermont       -0.1434337
## Virginia       0.2092464
## Washington    -0.2186282
## West Virginia  0.1305831
## Wisconsin      0.1822534
## Wyoming       -0.1649769
```