PCA

AUTHOR
DM/KA

PCA za pomocą prcomp

```
str(USArrests)
```

```
'data.frame':   50 obs. of  4 variables:
 $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
 $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
 $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
 $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

```
pca <- prcomp(USArrests, scale. = T)
pca
```

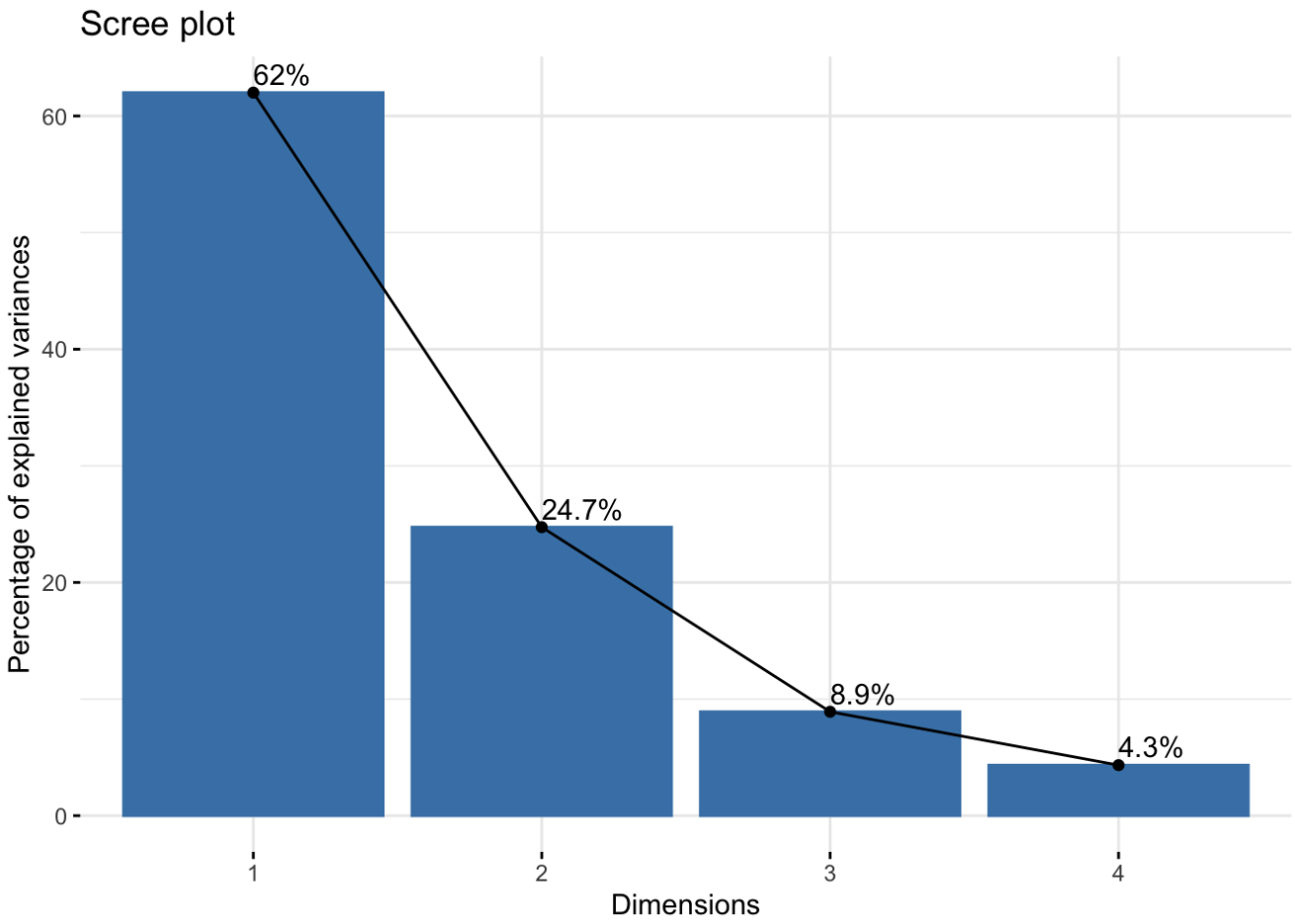
```
Standard deviations (1, .., p=4):
[1] 1.5748783 0.9948694 0.5971291 0.4164494

Rotation (n x k) = (4 x 4):
                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
```

$$PC1 = -0.54 * Murder - 0.58 * Assault - 0.28 * UrbanPop - 0.54 * Rape,$$

$$PC2 = 0.42 * Murder + 0.19 * Assault - 0.87 * UrbanPop - 0.17 * Rape,$$

Największy wkład w tworzenie się nowej zmiennej (PC1) ma zmienna Assault. Największy wkład w tworzenie PC2 ma zmienna UrbanPop.

```
library(factoextra)
fviz_screeplot(pca, addlabels=T)
```



Na podstawie kryterium procentu wyjaśnionej wariancji 2 składowe są wystarczające i wyjaśniają blisko 87% wariancji pierwotnej. Na podstawie kryterium Keisera należy wybrać tylko pierwszą składową ponieważ jedynie pierwsza wartość własna przekracza 1. Na podstawie kryterium osypiska trudno stwierdzić jaka liczba składowych będzie optymalna.

```
library(tidyverse)
summary(pca)
```

```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.5749 0.9949 0.59713 0.41645
Proportion of Variance 0.6201 0.2474 0.08914 0.04336
Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

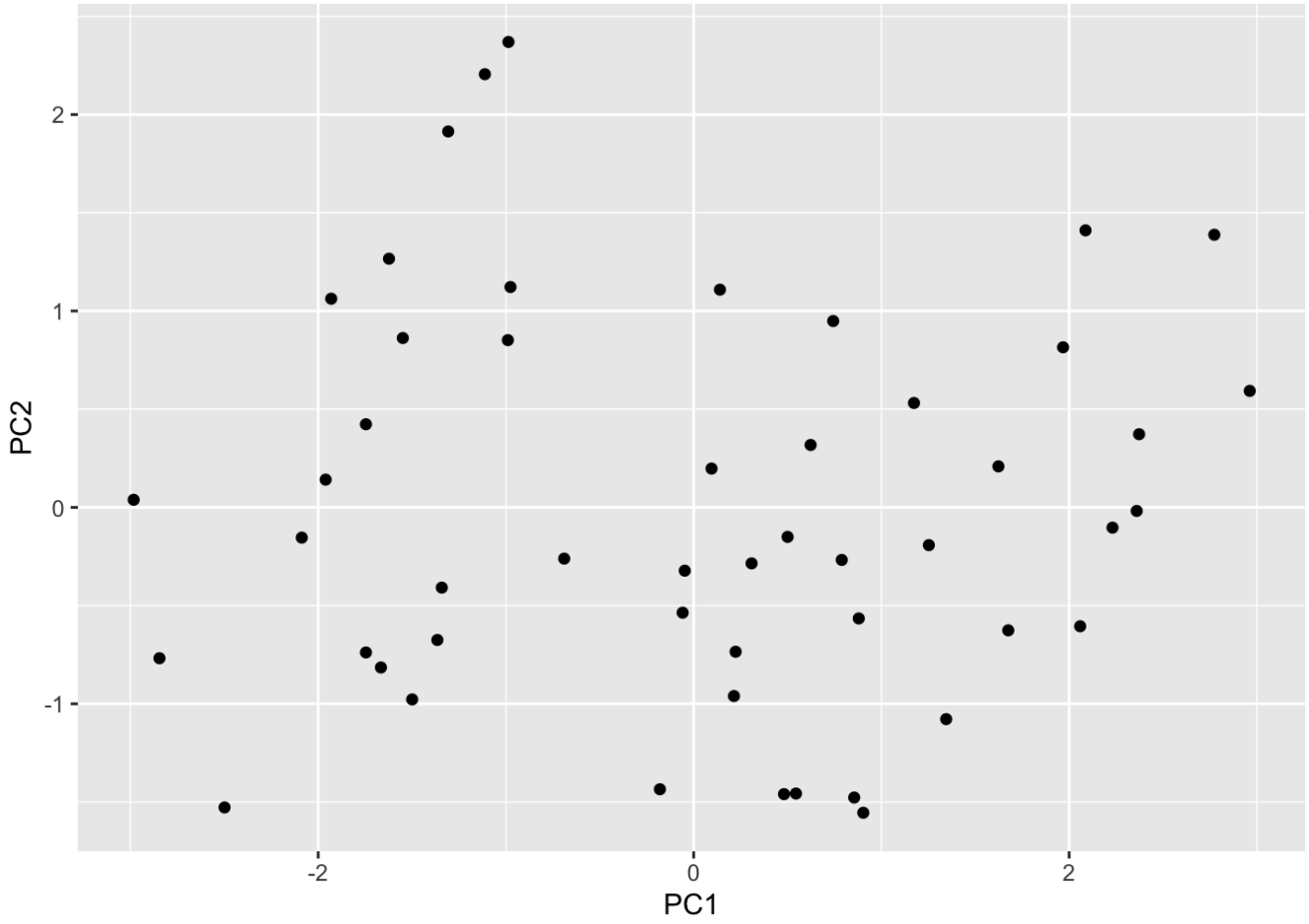
```
pca$x
```

```
                PC1         PC2         PC3          PC4
Alabama  -0.97566045  1.12200121 -0.43980366  0.154696581
Alaska   -1.93053788  1.06242692  2.01950027 -0.434175454
```

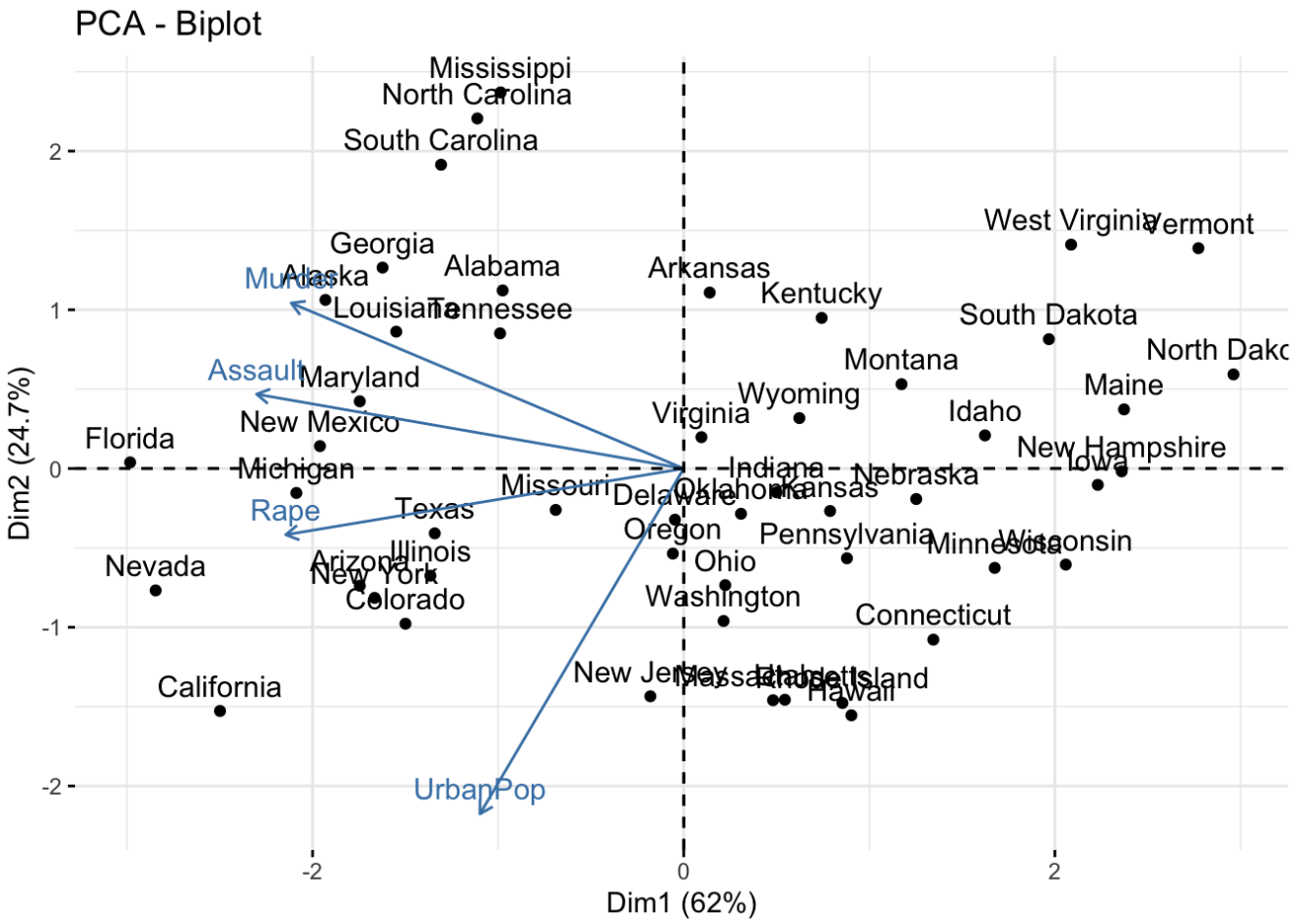
Arizona	-1.74544285	-0.73845954	0.05423025	-0.826264240
Arkansas	0.13999894	1.10854226	0.11342217	-0.180973554
California	-2.49861285	-1.52742672	0.59254100	-0.338559240
Colorado	-1.49934074	-0.97762966	1.08400162	0.001450164
Connecticut	1.34499236	-1.07798362	-0.63679250	-0.117278736
Delaware	-0.04722981	-0.32208890	-0.71141032	-0.873113315
Florida	-2.98275967	0.03883425	-0.57103206	-0.095317042
Georgia	-1.62280742	1.26608838	-0.33901818	1.065974459
Hawaii	0.90348448	-1.55467609	0.05027151	0.893733198
Idaho	1.62331903	0.20885253	0.25719021	-0.494087852
Illinois	-1.36505197	-0.67498834	-0.67068647	-0.120794916
Indiana	0.50038122	-0.15003926	0.22576277	0.420397595
Iowa	2.23099579	-0.10300828	0.16291036	0.017379470
Kansas	0.78887206	-0.26744941	0.02529648	0.204421034
Kentucky	0.74331256	0.94880748	-0.02808429	0.663817237
Louisiana	-1.54909076	0.86230011	-0.77560598	0.450157791
Maine	2.37274014	0.37260865	-0.06502225	-0.327138529
Maryland	-1.74564663	0.42335704	-0.15566968	-0.553450589
Massachusetts	0.48128007	-1.45967706	-0.60337172	-0.177793902
Michigan	-2.08725025	-0.15383500	0.38100046	0.101343128
Minnesota	1.67566951	-0.62590670	0.15153200	0.066640316
Mississippi	-0.98647919	2.36973712	-0.73336290	0.213342049
Missouri	-0.68978426	-0.26070794	0.37365033	0.223554811
Montana	1.17353751	0.53147851	0.24440796	0.122498555
Nebraska	1.25291625	-0.19200440	0.17380930	0.015733156
Nevada	-2.84550542	-0.76780502	1.15168793	0.311354436
New Hampshire	2.35995585	-0.01790055	0.03648498	-0.032804291
New Jersey	-0.17974128	-1.43493745	-0.75677041	0.240936580
New Mexico	-1.96012351	0.14141308	0.18184598	-0.336121113
New York	-1.66566662	-0.81491072	-0.63661186	-0.013348844
North Carolina	-1.11208808	2.20561081	-0.85489245	-0.944789648
North Dakota	2.96215223	0.59309738	0.29824930	-0.251434626
Ohio	0.22369436	-0.73477837	-0.03082616	0.469152817
Oklahoma	0.30864928	-0.28496113	-0.01515592	0.010228476
Oregon	-0.05852787	-0.53596999	0.93038718	-0.235390872
Pennsylvania	0.87948680	-0.56536050	-0.39660218	0.355452378
Rhode Island	0.85509072	-1.47698328	-1.35617705	-0.607402746
South Carolina	-1.30744986	1.91397297	-0.29751723	-0.130145378
South Dakota	1.96779669	0.81506822	0.38538073	-0.108470512
Tennessee	-0.98969377	0.85160534	0.18619262	0.646302674
Texas	-1.34151838	-0.40833518	-0.48712332	0.636731051
Utah	0.54503180	-1.45671524	0.29077592	-0.081486749

```
Vermont        2.77325613  1.38819435  0.83280797 -0.143433697
Virginia       0.09536670  0.19772785  0.01159482  0.209246429
Washington     0.21472339 -0.96037394  0.61859067 -0.218628161
West Virginia  2.08739306  1.41052627  0.10372163  0.130583080
Wisconsin      2.05881199 -0.60512507 -0.13746933  0.182253407
Wyoming        0.62310061  0.31778662 -0.23824049 -0.164976866
```

```r
pca$x |>
  as.data.frame() |>
  ggplot(aes(x = PC1, y = PC2)) +
  geom_point()
```

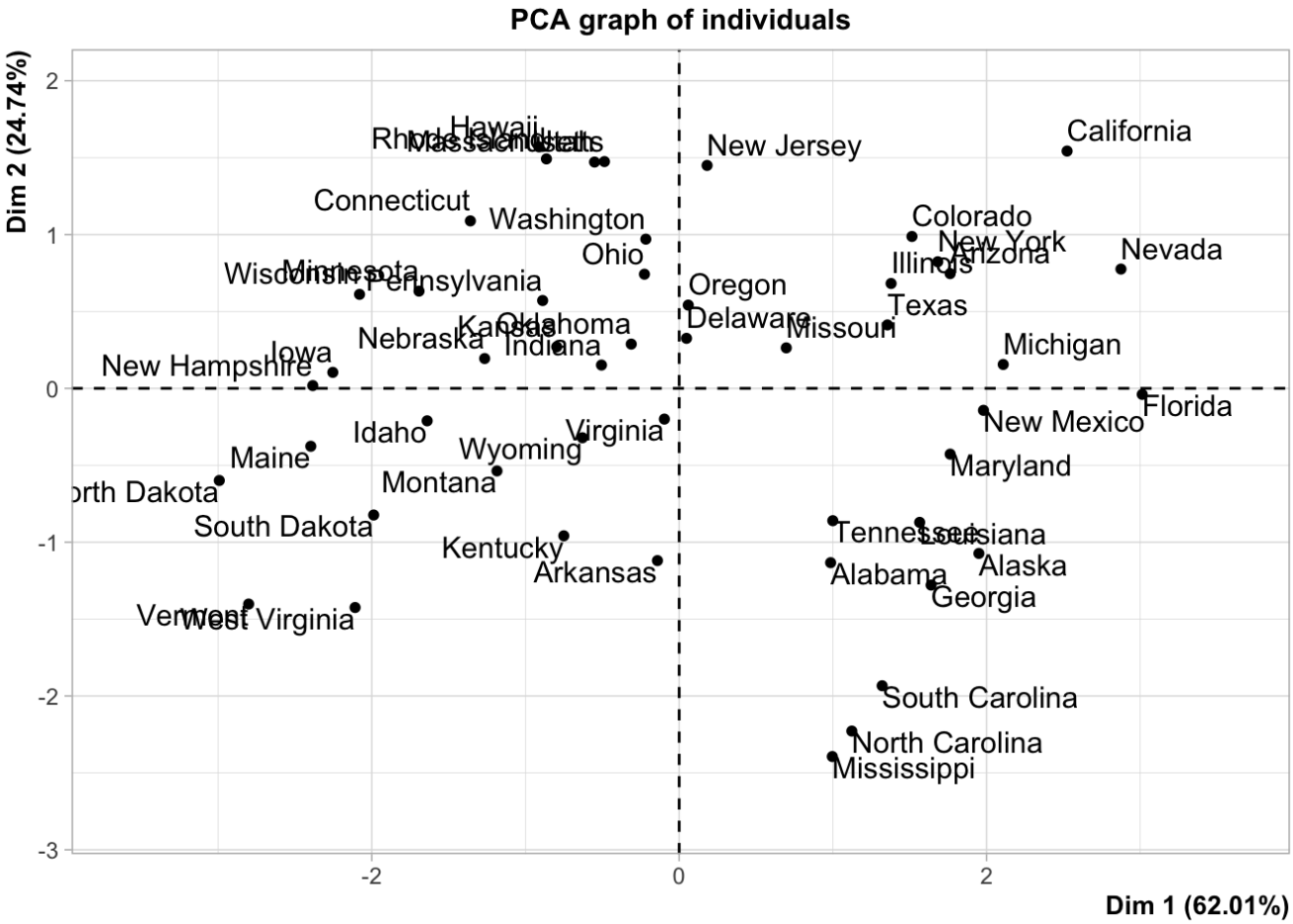


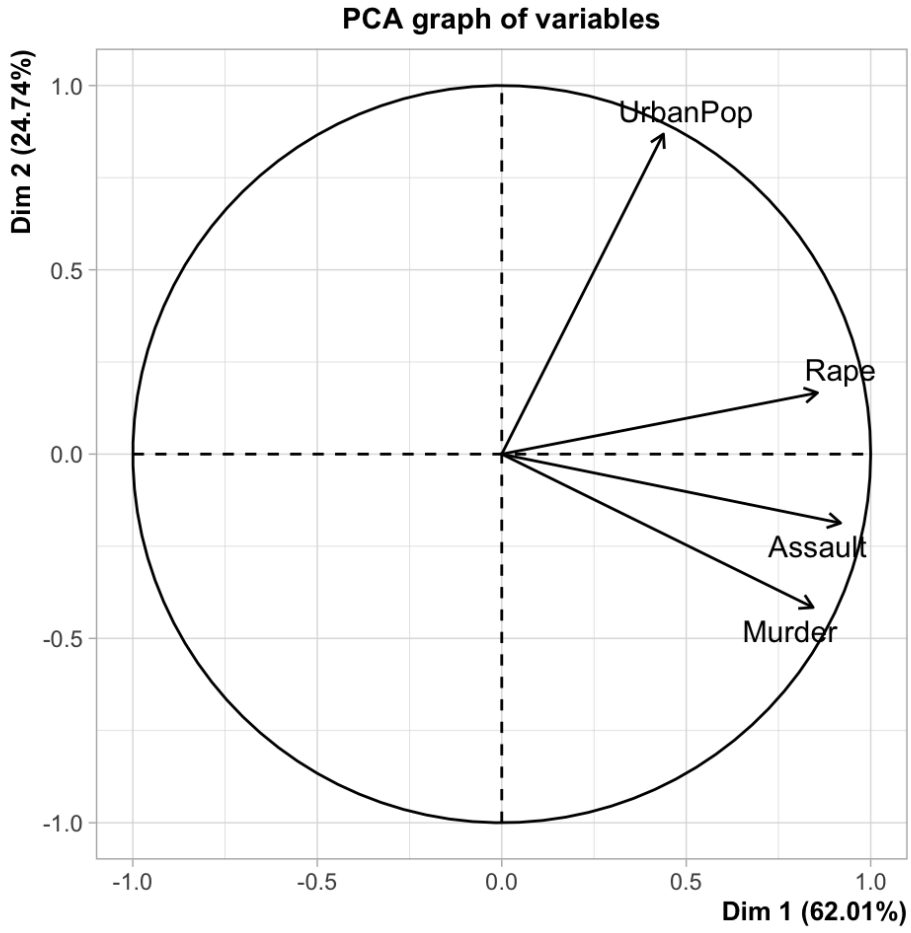

```r
# albo tak
fviz_pca(pca)
```



PCA z wykorzystaniem FactoMineR

```
library(FactoMineR)
pca <- PCA(USArrests, ncp = 4)
```


PCA graph of individuals
Dim 2 (24.74%)
Dim 1 (62.01%)
Hawaii
New Jersey
California
Connecticut
Washington
Colorado
New York
Ohio
Arizona
Nevada
Wisconsin
Minnesota
Pennsylvania
Oregon
Illinois
Texas
Delaware
Oklahoma
Kansas
Nebraska
Indiana
Missouri
Michigan
Iowa
New Hampshire
Florida
New Mexico
Idaho
Virginia
Wyoming
Maine
Maryland
Montana
South Dakota
Kentucky
Arkansas
Alabama
Alaska
Georgia
West Virginia
South Carolina
North Carolina
Mississippi

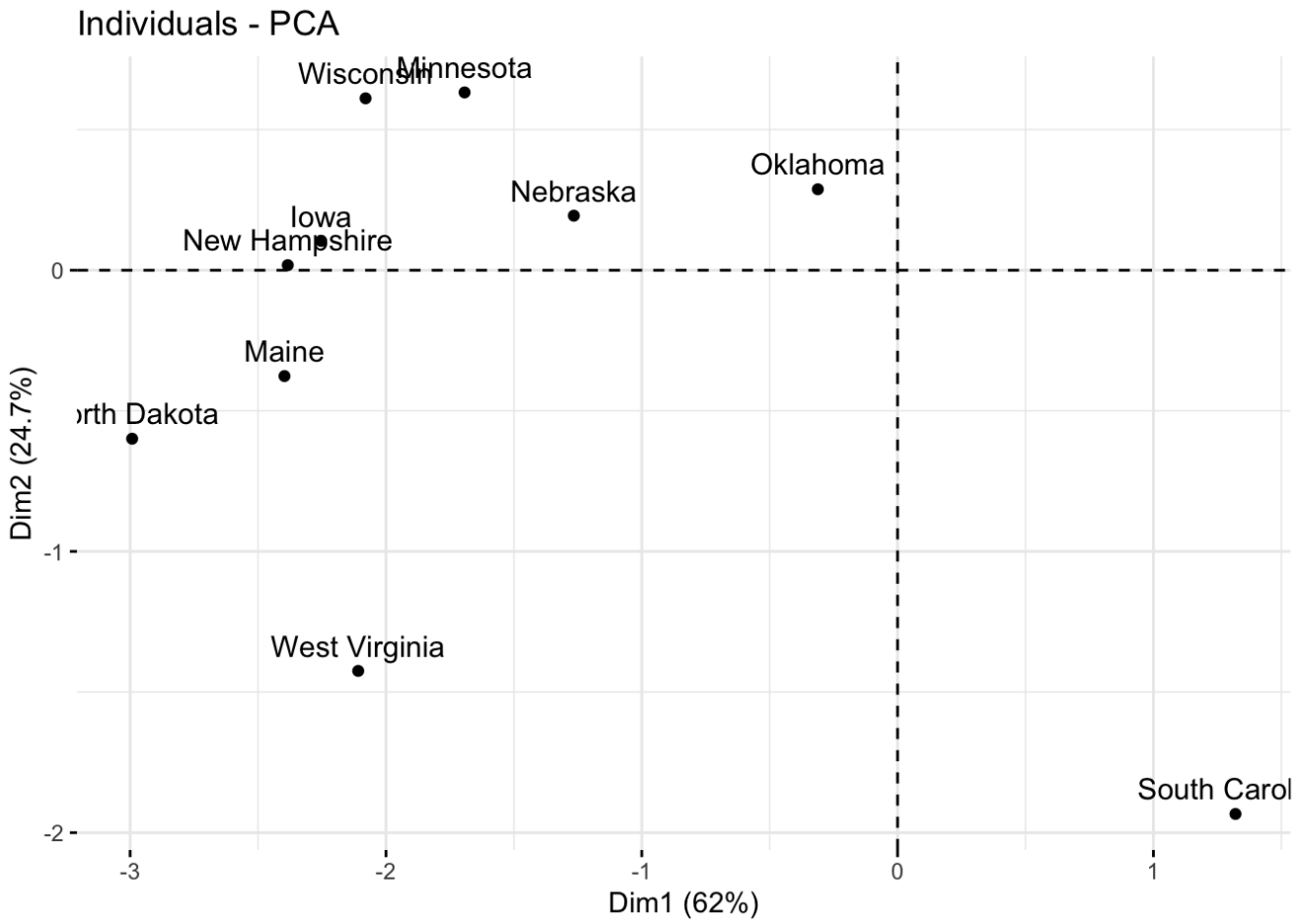

PCA graph of variables
Dim 2 (24.74%)
Dim 1 (62.01%)
UrbanPop
Rape
Assault
Murder

```
summary(pca)
```

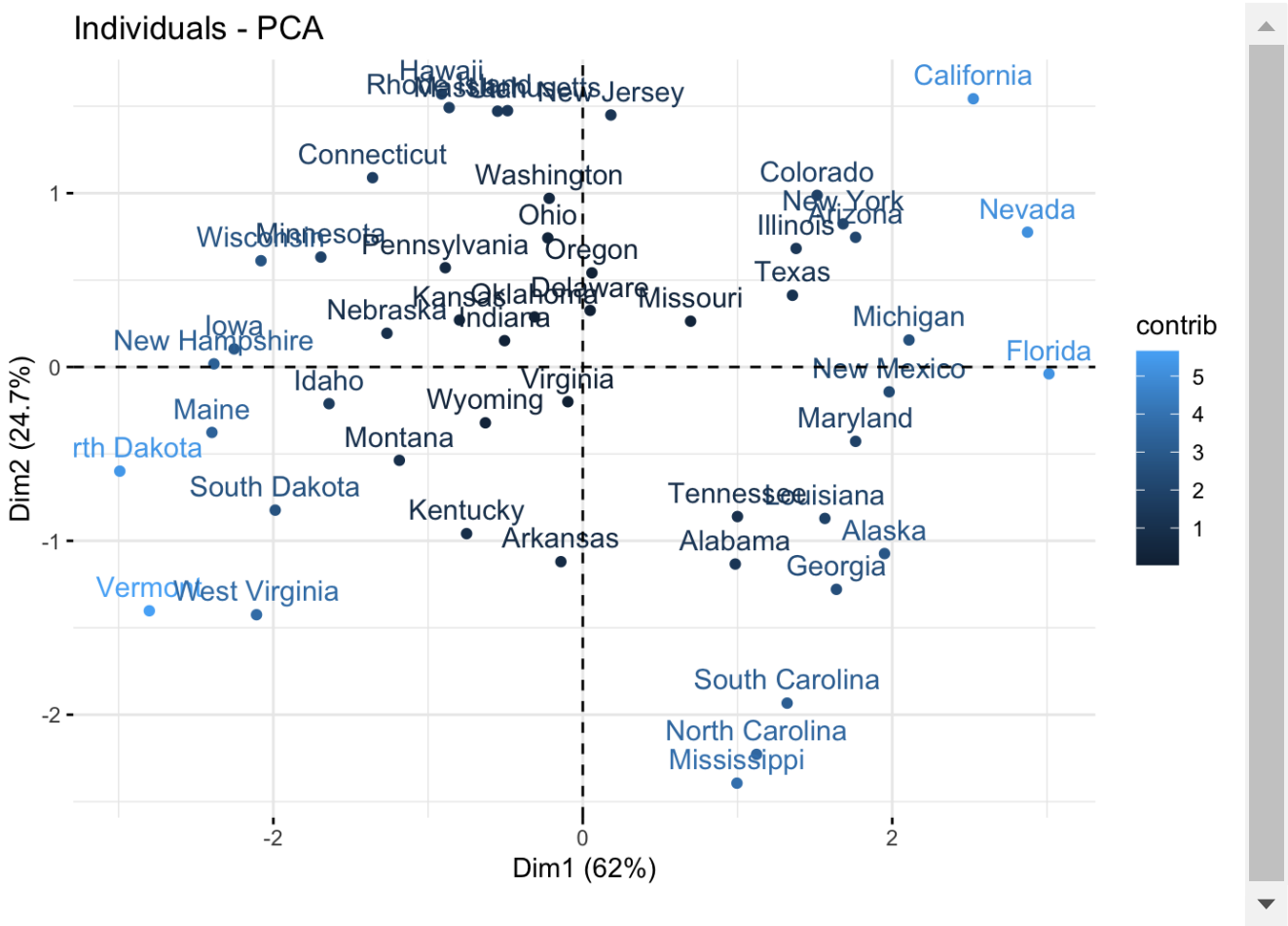
```
Call:
PCA(X = USArrests, ncp = 4)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4
Variance               2.480   0.990   0.357   0.173
% of var.             62.006  24.744   8.914   4.336
Cumulative % of var.  62.006  86.750  95.664 100.000

Individuals (the 10 first)
                Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2
Dim.3
Alabama     |  1.574 |  0.986  0.783  0.392 | -1.133  2.596  0.518 |
0.444
Alaska      |  3.051 |  1.950  3.067  0.409 | -1.073  2.327  0.124 |
-2.040
Arizona     |  2.089 |  1.763  2.507  0.712 |  0.746  1.124  0.127 |
-0.055
Arkansas    |  1.149 | -0.141  0.016  0.015 | -1.120  2.534  0.950 |
-0.115
California  |  3.037 |  2.524  5.137  0.690 |  1.543  4.811  0.258 |
-0.599
Colorado    |  2.114 |  1.515  1.850  0.513 |  0.988  1.971  0.218 |
-1.095
Connecticut |  1.860 | -1.359  1.489  0.534 |  1.089  2.396  0.343 |
0.643
Delaware    |  1.184 |  0.048  0.002  0.002 |  0.325  0.214  0.075 |
0.719
Florida     |  3.070 |  3.013  7.321  0.964 | -0.039  0.003  0.000 |
0.577
Georgia     |  2.366 |  1.639  2.167  0.480 | -1.279  3.305  0.292 |
0.342
               ctr   cos2
Alabama      1.107  0.080 |
Alaska      23.343  0.447 |
Arizona      0.017  0.001 |
Arkansas     0.074  0.010 |
California   2.010  0.039 |
Colorado     6.726  0.268 |
Connecticut  2.321  0.120 |
```

```
Delaware     2.897  0.368 |
Florida      1.866  0.035 |
Georgia      0.658  0.021 |

Variables
               Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
ctr
Murder       |  0.844 28.719  0.712 | -0.416 17.488  0.173 |  0.204
11.644
Assault      |  0.918 34.010  0.844 | -0.187  3.534  0.035 |  0.160
7.190
UrbanPop     |  0.438  7.739  0.192 |  0.868 76.179  0.754 |  0.226
14.290
Rape         |  0.856 29.532  0.732 |  0.166  2.800  0.028 | -0.488
66.876
              cos2
Murder       0.042 |
Assault      0.026 |
UrbanPop     0.051 |
Rape         0.238 |
```

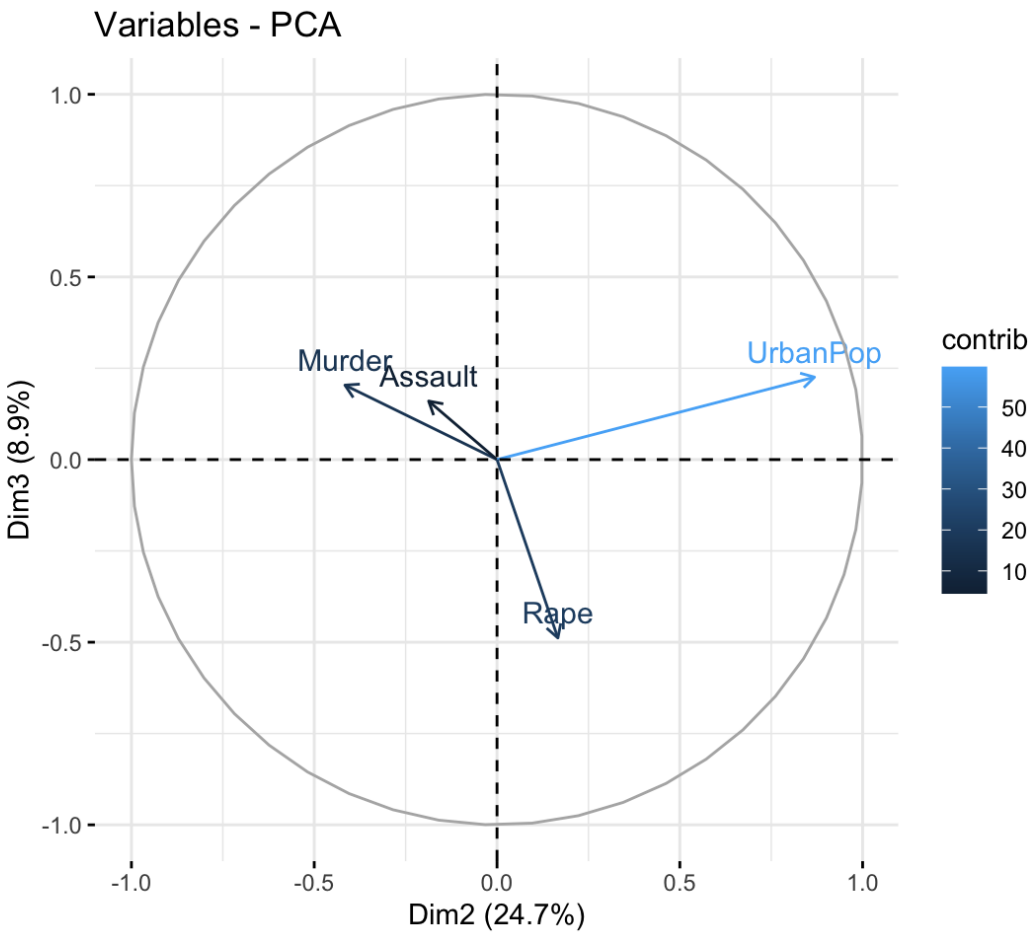
```r
fviz_pca_ind(pca, axes = 1:2, select.ind = list(cos2 = 10))
```



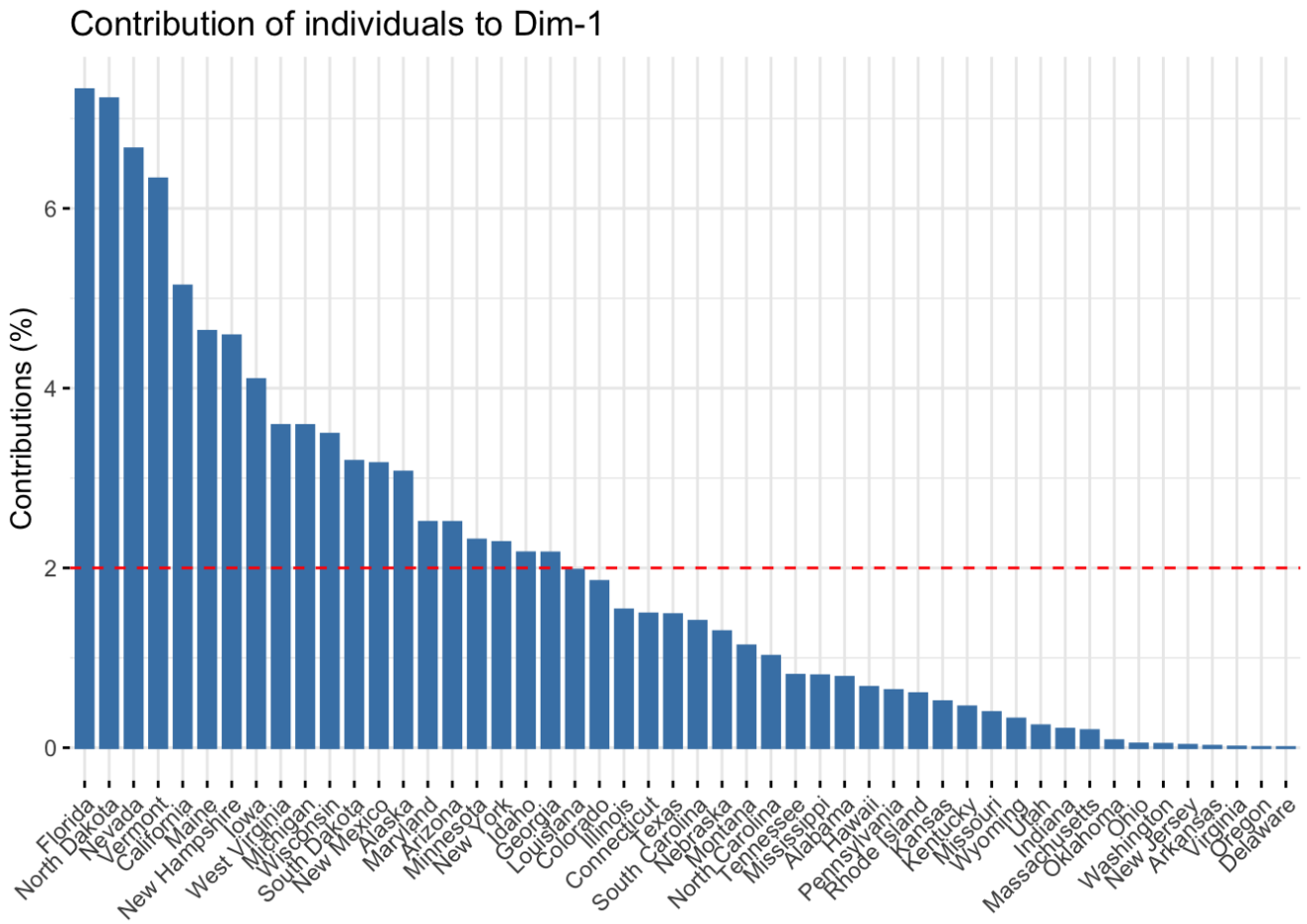

```
fviz_pca_ind(pca, axes = 1:2, col.ind = "contrib")
```



```
fviz_pca_var(pca, axes = 2:3, col.var = "contrib")
```




```
fviz_contrib(pca, choice = "ind", axes = 1)
```




```
fviz_cos2(pca, choice = "ind", axes = 1:2)
```