DA5

2025-06-24

```
states = row.names(USArrests)
states
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
##  [5] "California"     "Colorado"       "Connecticut"    "Delaware"
##  [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
## [13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
## [17] "Kentucky"       "Louisiana"      "Maine"          "Maryland"
## [21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
## [25] "Missouri"       "Montana"        "Nebraska"       "Nevada"
## [29] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
## [33] "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
## [37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
## [41] "South Dakota"   "Tennessee"      "Texas"          "Utah"
## [45] "Vermont"        "Virginia"       "Washington"     "West Virginia"
## [49] "Wisconsin"      "Wyoming"
```

we got the names of rows - states of the US

```
names(USArrests)
```

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

here are the names of colums

```
apply(USArrests, 2, mean)
```

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

apply works the same as in python 2 means that we want this function to be applied for columns

```
apply(USArrests, 2, var)
```

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

the biggest variance is among assault column

```r
arrestspca = prcomp(USArrests, scale = TRUE)
summary(arrestspca)
```

```
## 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
```

```r
names(arrestspca)
```

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

```r
arrestspca$scale
```

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

```r
arrestspca$center
```

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

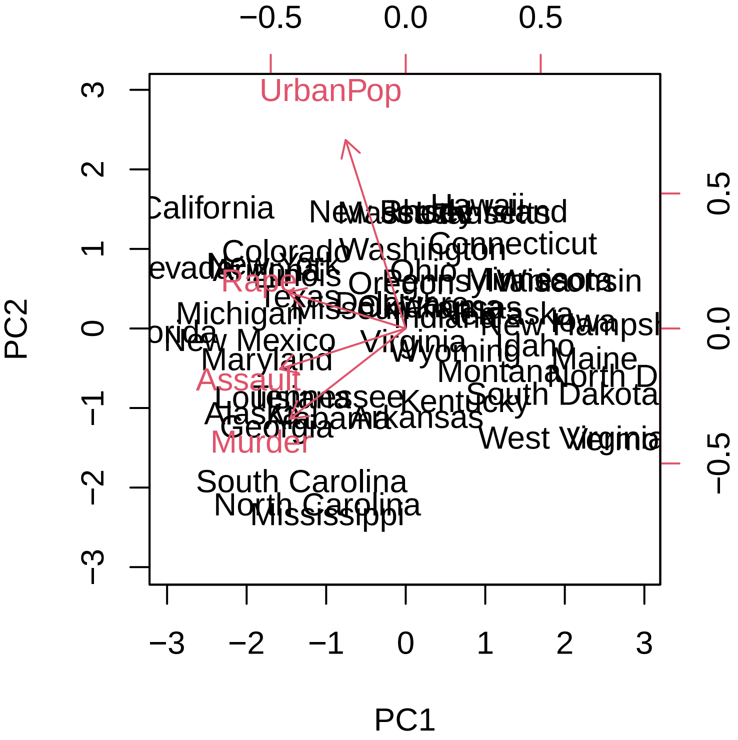
```r
arrestspca$rotation
```

```
##                 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
```

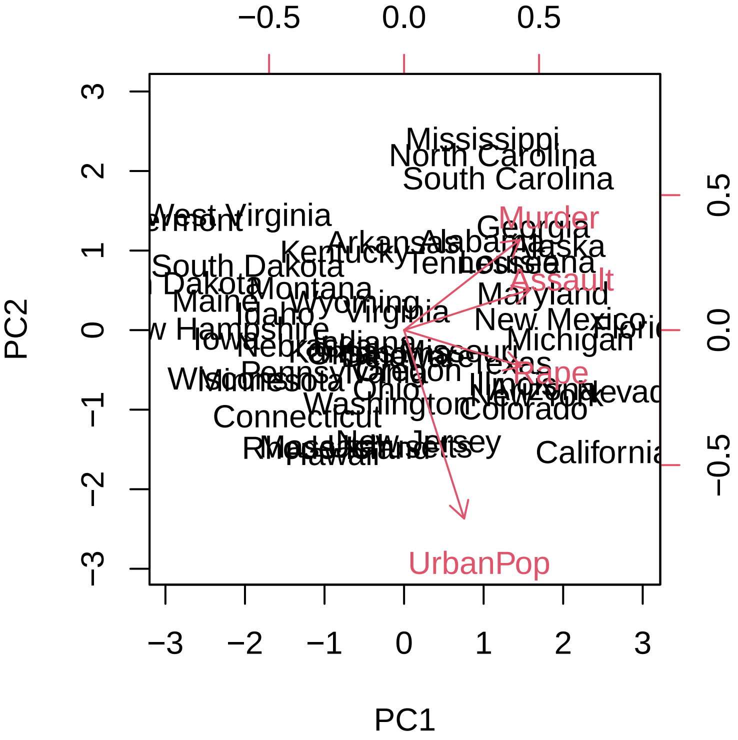
A large absolute loading (–0.583 for Assault on PC1) means that variable contributes heavily to that component. - sing tells the direction

We can see that PC1 is essentially an “overall crime level” axis (all four crimes load strongly and in the same direction). PC2 contrasts UrbanPop (–0.873) against the other three (positive but smaller): a “rural vs urban” dimension. PC3 is driven by Rape (0.818) versus the rest. PC4 pits Assault (–0.743) against Murder (0.649).

```r
biplot(arrestspca, scale=0)
```



```
arrestspca$rotation=-arrestspca$rotation
arrestspca$x=-arrestspca$x
biplot (arrestspca , scale =0)
```



```
vari = arrestspca$sdev^2
vari
```

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

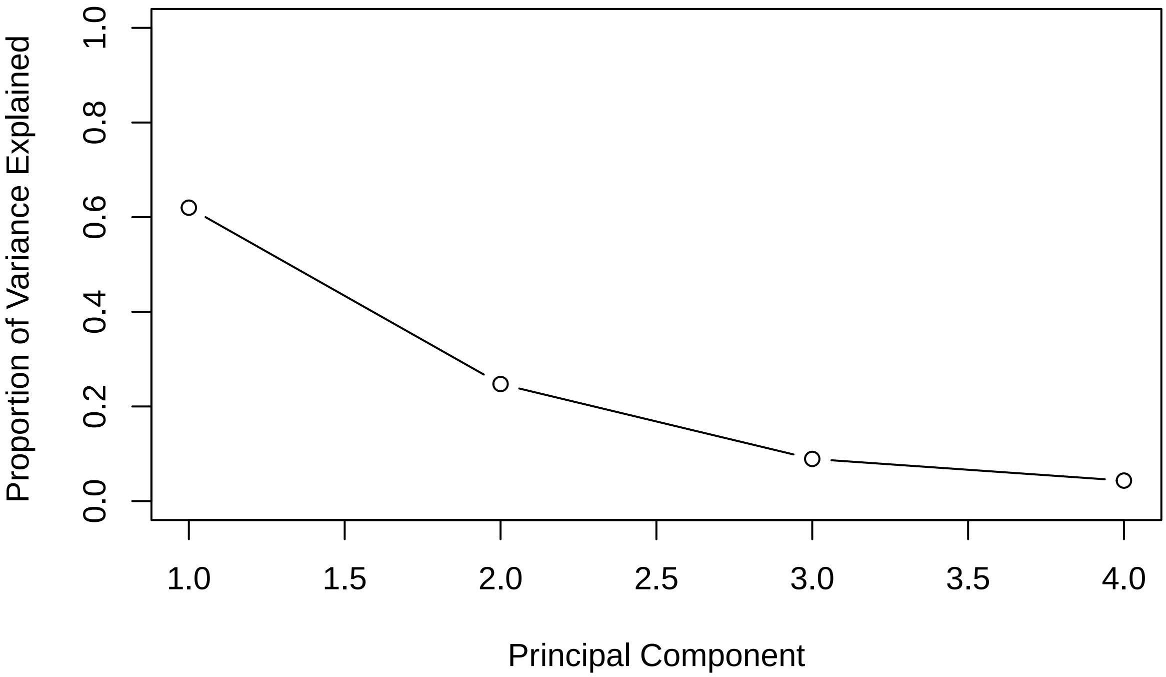
The variance

```
pve = vari / sum(vari)
pve
```

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

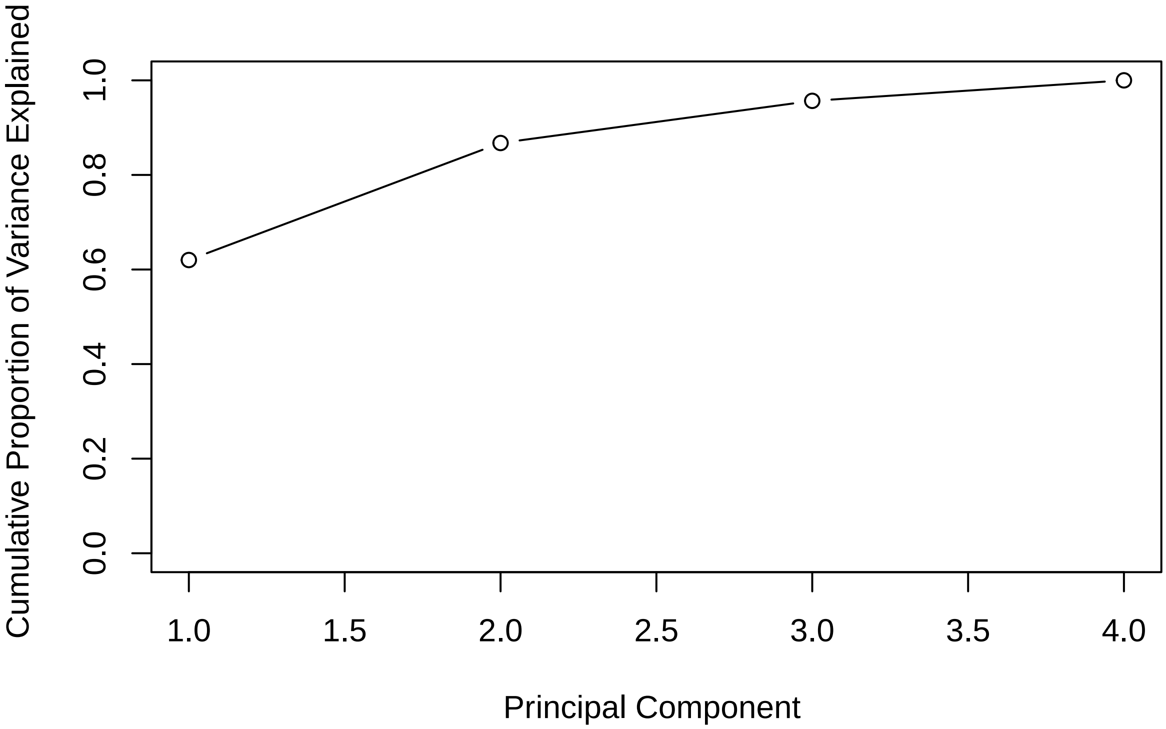
62% of variance is explained by pc1, 25 by pc2, 9 by pc3 and only 4 by pc4

```
plot(pve , xlab=" Principal Component ", ylab="Proportion of Variance Explained ", ylim=c(0,1), type='b
```



this is illustrated in the graph. We again see the hockey stick. The elbow is probably PC3 because after it it started to be more horizontal

```
plot(cumsum(pve), xlab="Principal Component ", ylab="Cumulative Proportion of Variance Explained ", ylin
```



this is for culminative one