PC example

Air pollution data from Sokal and Rohlf, *Biometry*, 1981.

```
City: City
SO2: Sulfur dioxide content of air in micrograms per cubic meter
Temp: Average annual temperature in degrees Fahrenheit
Man: Number of manufacturing enterprises employing 20 or more worl
Pop: Population size in thousands from the 1970 census
Wind: Average annual wind speed in miles per hour
Rain: Average annual precipitation in inches
RainDays: Average number of days with precipitation per year
```

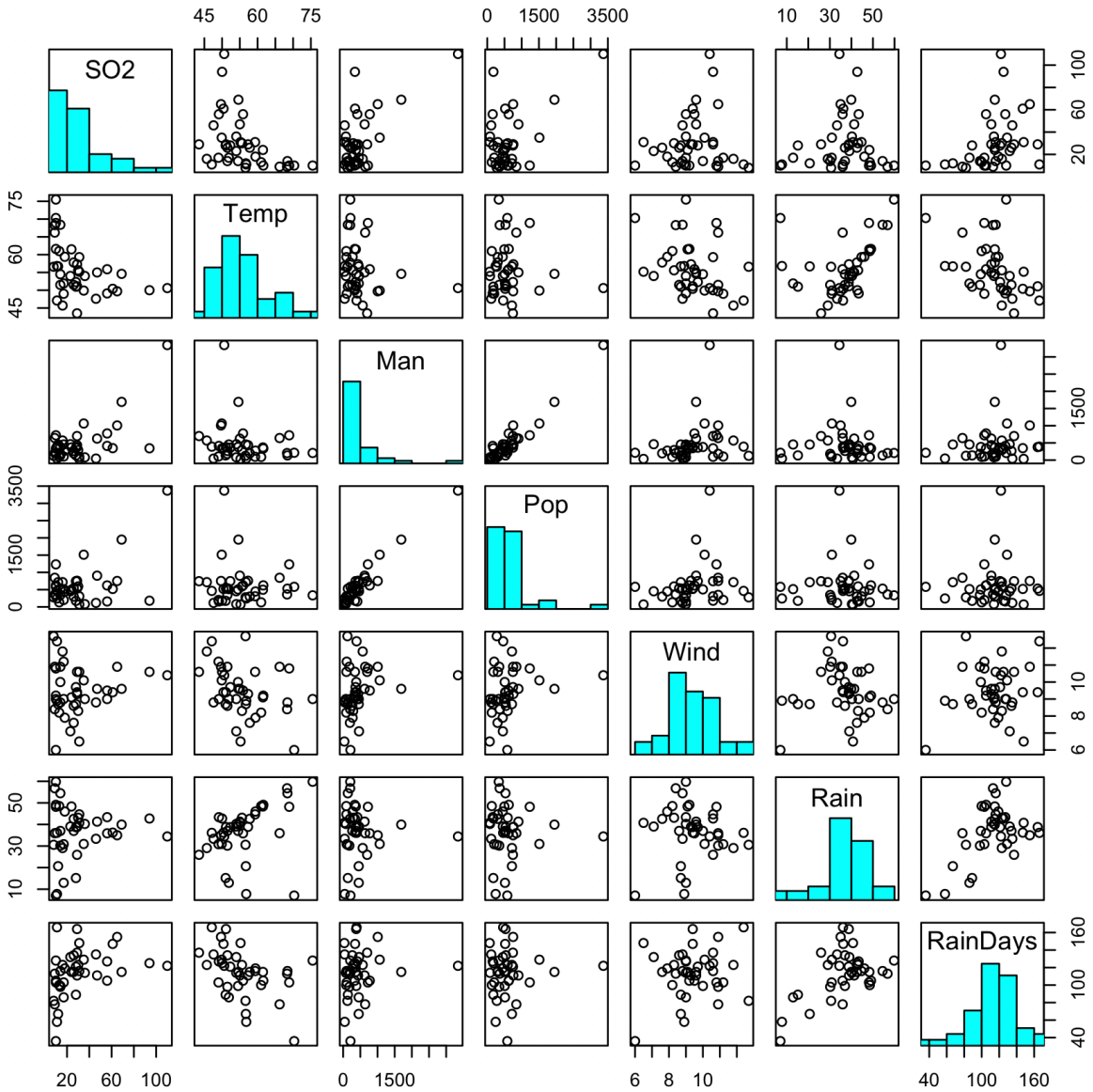
PC example

```
> head(x)
          City SO2 Temp Man Pop Wind  Rain RainDays
1      Phoenix  10 70.3 213 582  6.0  7.05       36
2   LittleRock  13 61.0  91 132  8.2 48.52      100
3 SanFrancisco  12 56.7 453 716  8.7 20.66       67
4       Denver  17 51.9 454 515  9.0 12.95       86
5     Hartford  56 49.1 412 158  9.0 43.37      127
6   Wilmington  36 54.0  80  80  9.0 40.25      114
> x[x$City=="Albuquerque",]
          City SO2 Temp Man Pop Wind Rain RainDays
23 Albuquerque  11 56.8  46 244  8.9 7.77       58
```

PC example

To visualize the data, note that all variables except `City` are quantitative, so I'll do a scatterplot matrix on all variables except `City`.

```
help(pairs)
     panel.hist <- function(x, ...)
      {
         usr <- par("usr"); on.exit(par(usr))
         par(usr = c(usr[1:2], 0, 1.5) )
         h <- hist(x, plot = FALSE)
         breaks <- h$breaks; nB <- length(breaks)
         y <- h$counts; y <- y/max(y)
         rect(breaks[-nB], 0, breaks[-1], y, col =
         "cyan", ...)
      }
pairs(x[,2:8],diag.panel=panel.hist)
```


SO2
Temp
Man
Pop
Wind
Rain
RainDays

PC example

Chicago is an outlier in some dimensions but not all dimensions.

```
> x$City[which(x$Pop==max(x$Pop))]
[1] Chicago
> options(digits=4)
> colMeans(x[,-1])
     SO2     Temp      Man      Pop     Wind     Rain RainDays
  30.049   55.763  463.098  608.610    9.444   36.769  113.902
> x[11,]
       City SO2 Temp  Man  Pop Wind  Rain RainDays
11 Chicago 110 50.6 3344 3369 10.4 34.44      122
```

We can do principal components on variables 2 through 8 using the scaled data using the cor=TRUE option.

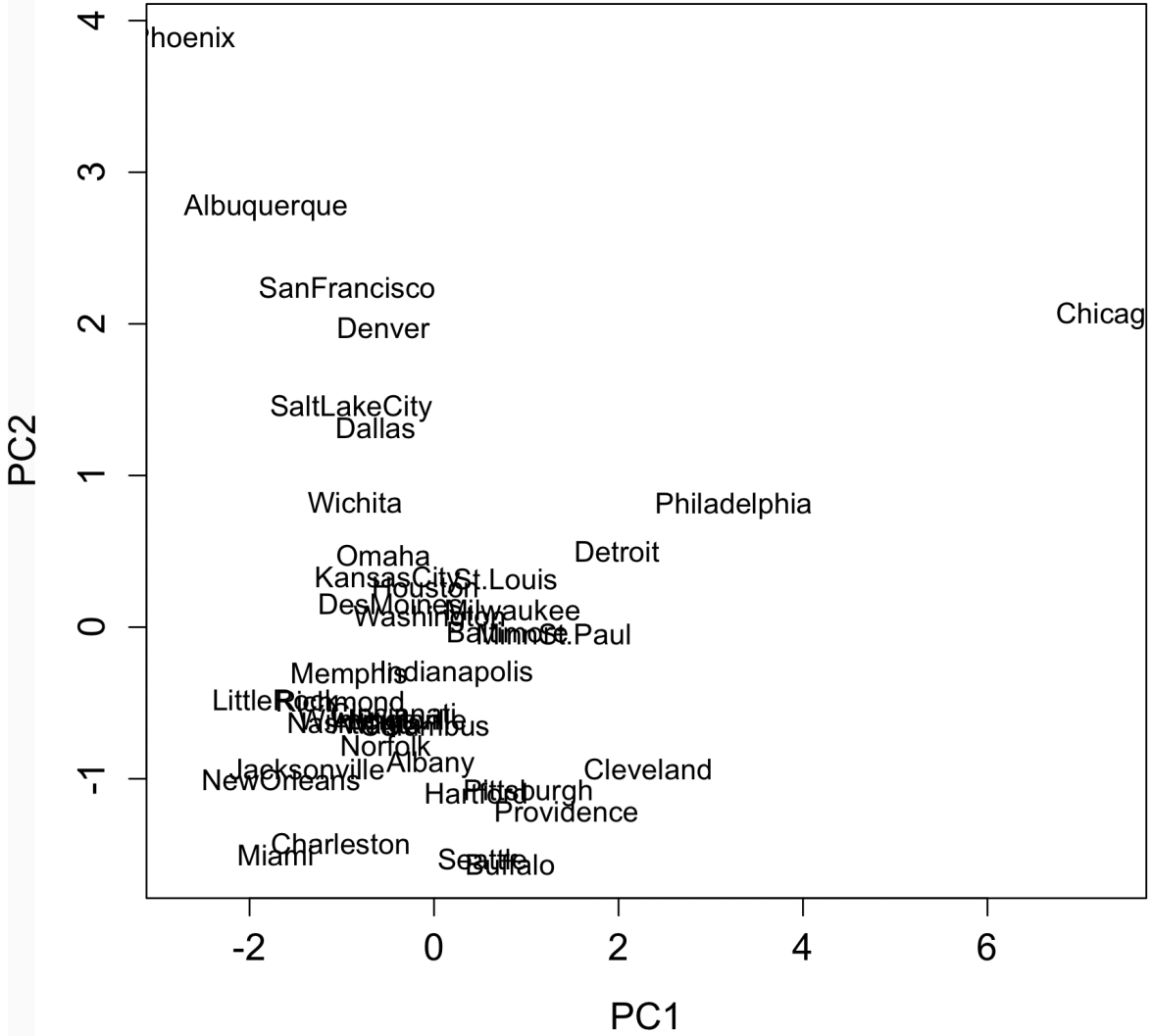
```
> pc <- princomp(x[,-1],cor=TRUE)
> options(digits=2)
> summary(pc,loadings=TRUE)
            Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
St. dev.      1.65   1.23   1.18   0.94   0.59  0.317 0.1597
Prop. Var.    0.39   0.22   0.20   0.13   0.05  0.014 0.0036
Cum. Prop.    0.39   0.61   0.81   0.93   0.98  0.996 1.0000
Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
SO2       0.490                0.404 -0.730 -0.183  0.150
Temp     -0.315        -0.677 -0.185 -0.162 -0.611
Man       0.541  0.226 -0.267         0.164        -0.745
Pop       0.488  0.282 -0.345 -0.113  0.349         0.649
Wind      0.250         0.311 -0.862 -0.268 -0.150
Rain             -0.626 -0.492 -0.184 -0.161  0.554
RainDays  0.260 -0.678  0.110  0.110  0.440 -0.505
```

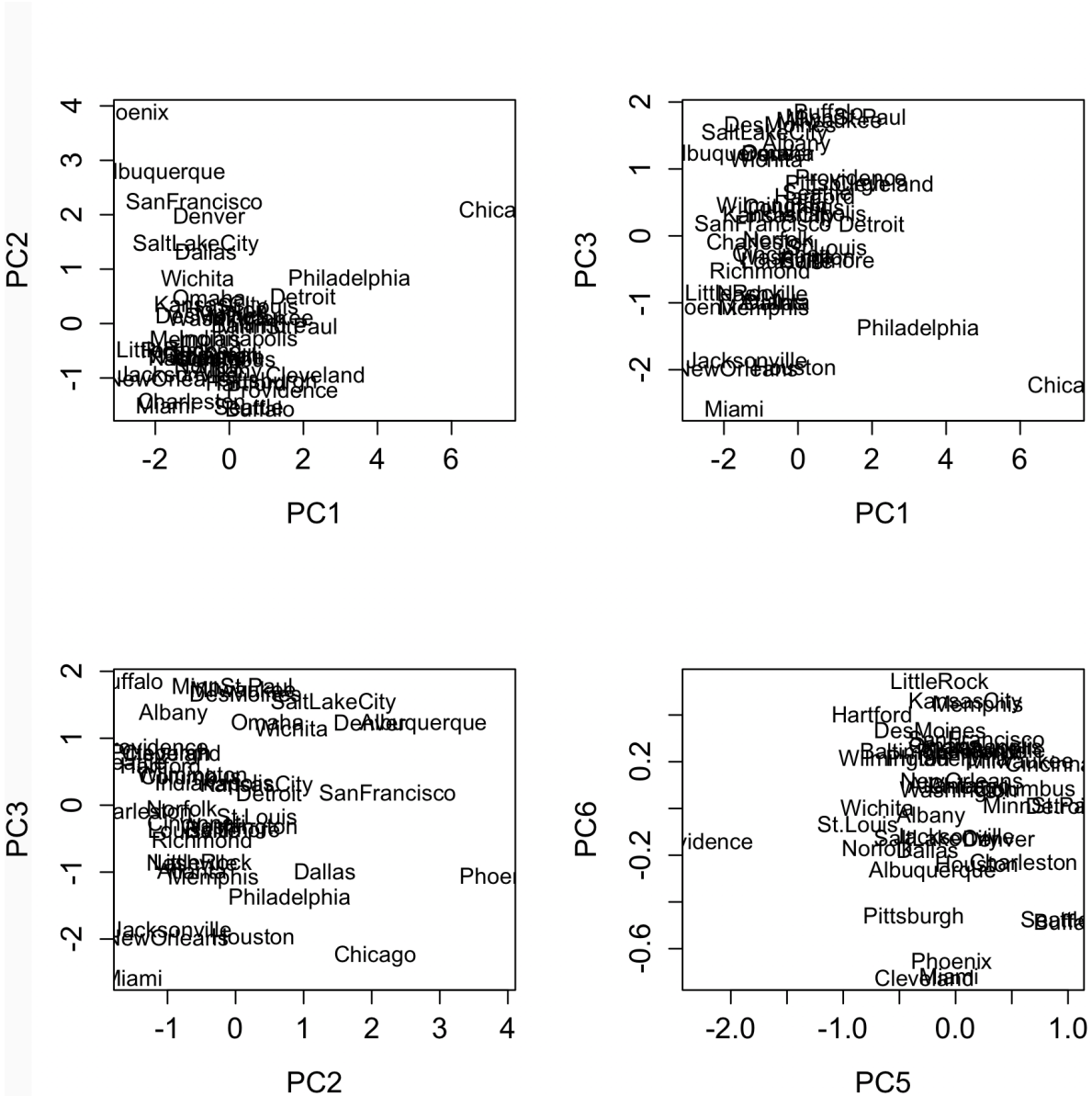
A rule of thumb often used is to consider components important if they have standard deviations (eigenvalues when scaled data is used) larger than one. For the pollution data, the first three principal components have standard deviations greater than 1, and this accounts for 81% fo the variation in the data. In practice, it is also hard to visualize more than 2 or 3 compontents.

To interpret the components, the first component is large for cities with high pollution, high manufacturing and population, high wind, many rainy days, and low temperatures. The second component is large for large cities that are drier. The third component is getting harder to interpret, but is large for smaller, colder cities with not much total rainfall (but maybe more rainy days...).

PC example

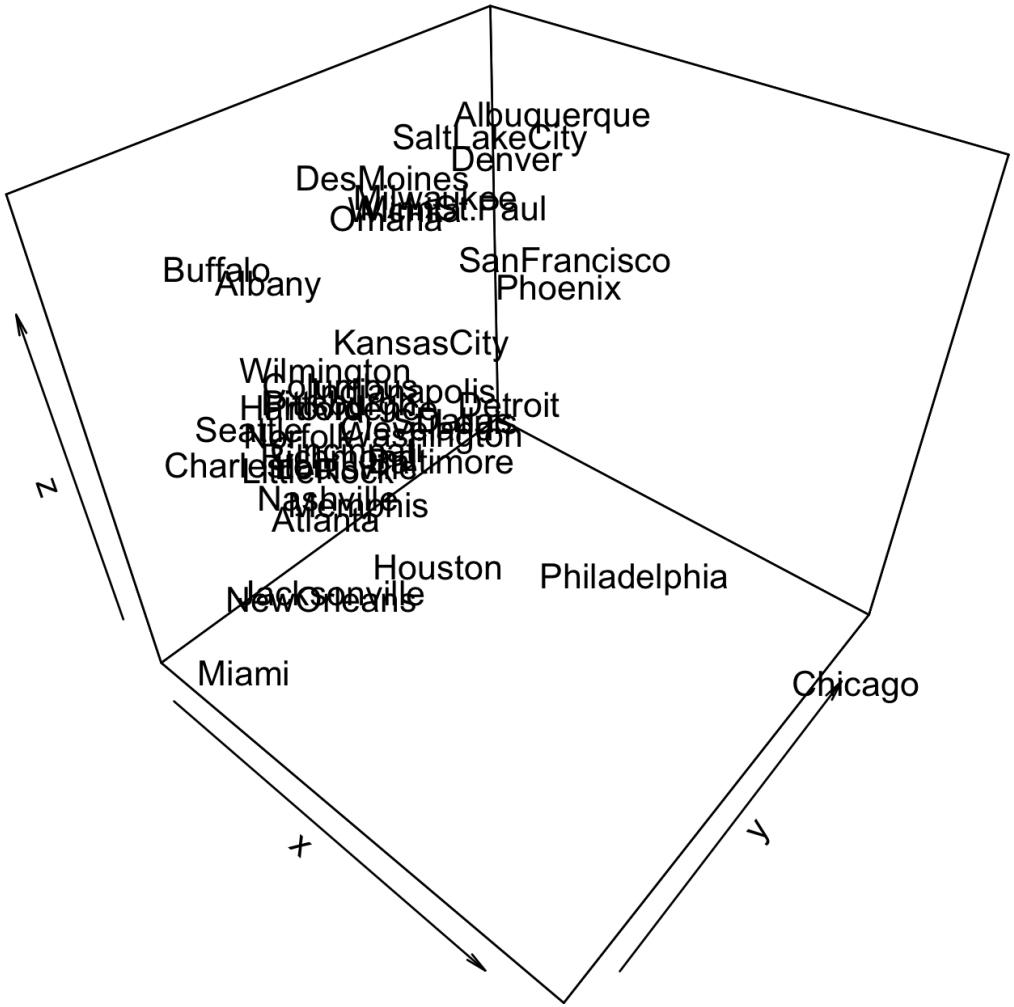
```
> plot(pc$scores[,1],pc$scores[,2],cex.lab=1.3,cex.axis=1.3,
xlab="PC1",ylab="PC2",type="n")
> text(pc$scores[,1],pc$scores[,2],x$City)
```


PC2
PC1
4
3
2
1
0
-1
-2
0
2
4
6
hoenix
Albuquerque
SanFrancisco
Denver
Chicag
SaltLakeCity
Dallas
Wichita
Philadelphia
Omaha
Detroit
KansasCity
St.Louis
Houston
DesMoines
Milwaukee
Washington
Baltimore
St.Paul
Memphis
Indianapolis
LittleRock
Richmond
Nashville
Columbus
Norfolk
Albany
Jacksonville
Cleveland
NewOrleans
Hartford
Pittsburgh
Providence
Charleston
Miami
Seattle
Buffalo


PC2
PC1
PC3
PC1
PC3
PC2
PC6
PC5

For. a three dimension plot of the first three PCs, I used the `plot3D` library:

```
> library(plot3D)
> scatter3D(pc$scores[,1],pc$scores[,2],pc$scores[,3])
> text3D(pc$scores[,1],pc$scores[,2],pc$scores[,3],x$City)
```


Albuquerque
SaltLakeCity
Denver
DesMoines
Omaha
SanFrancisco
Phoenix
Buffalo
Albany
KansasCity
Wilmington
Detroit
Houston
Philadelphia
Miami
Chicago
z
x
y