PCA: Houses

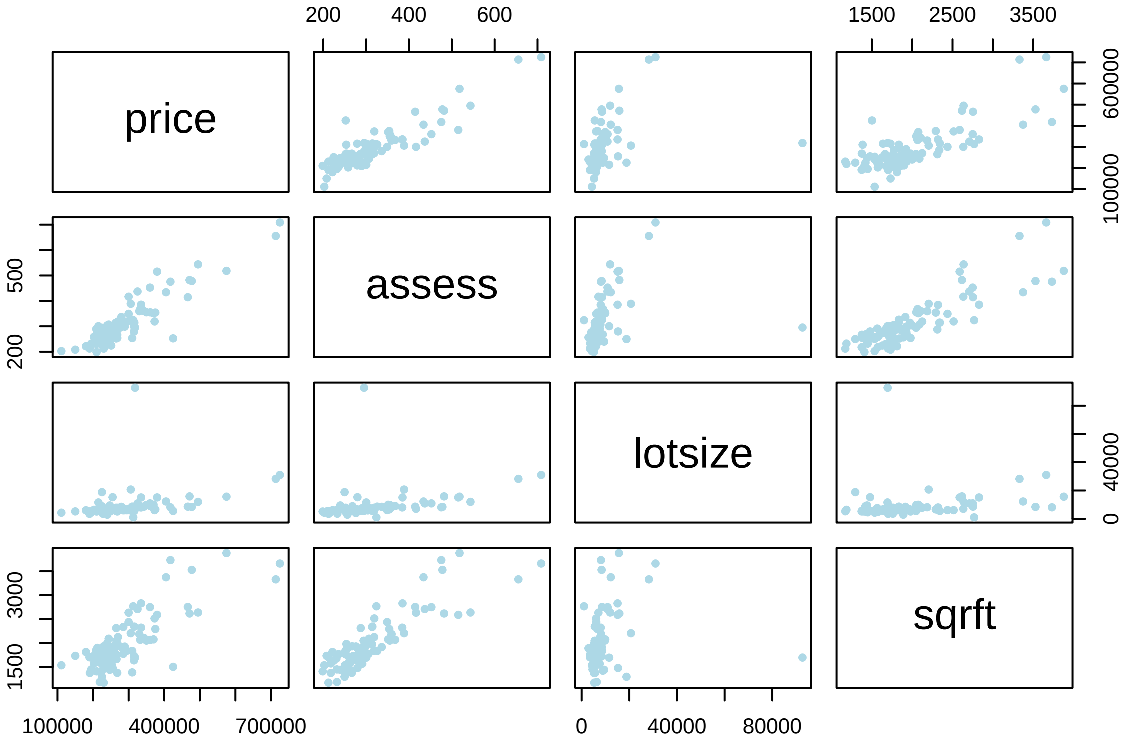
Gustavo Cepparo and Milica Cudina

First, we read in our house data.

```
data=read.csv("housepriceall.csv",header=TRUE)
attach(data)
#View(data)
```

Let me do a bit of exploratory data analysis.

```
plot(data, pch=20, col="lightblue")
```




```
cor(data)
```

```
##             price    assess   lotsize     sqrft
## price   1.0000000 0.9052794 0.3471245 0.7879065
## assess  0.9052794 1.0000000 0.3281463 0.8656345
## lotsize 0.3471245 0.3281463 1.0000000 0.1838422
## sqrft   0.7879065 0.8656345 0.1838422 1.0000000
```

Obviously, the scale of the price is different from the scale of the square footage. Also, the assessed price and the price are artificially on a different scale. Let's look at the means and variances.

```
apply(data, 2, mean)
```

```
##        price       assess      lotsize        sqrft
```

```
## 293546.0341    315.7364   9019.8636   2013.6932
```

```r
apply(data, 2, sd)
```

```
##         price        assess       lotsize         sqrft
## 102713.44517      95.31444  10174.15041     577.19158
```

Let's take a look at what the principal component analysis is telling us.

```r
pr.out<-prcomp(data, scale=TRUE)
```

Here are the outputs of `prcomp`.

```r
pr.out$center
```

```
##       price      assess     lotsize       sqrft
## 293546.0341    315.7364   9019.8636   2013.6932
```

```r
pr.out$scale
```

```
##         price        assess       lotsize         sqrft
## 102713.44517      95.31444  10174.15041     577.19158
```

```r
pr.out$rotation
```

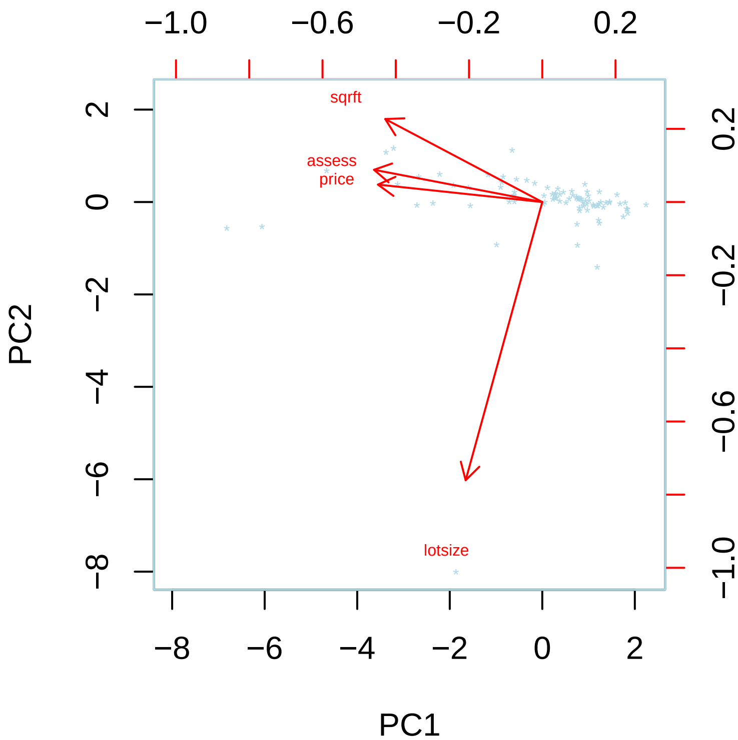
```
##                PC1         PC2        PC3         PC4
## price   -0.5608750  0.05950919 -0.6507820  0.50829197
## assess  -0.5742342  0.11020662 -0.1214913 -0.80209066
## lotsize -0.2615396 -0.95073388  0.1633602  0.03186797
## sqrft   -0.5359770  0.28358110  0.7314616  0.31188825
```

```r
dim(pr.out$x)
```

```
## [1] 88  4
```

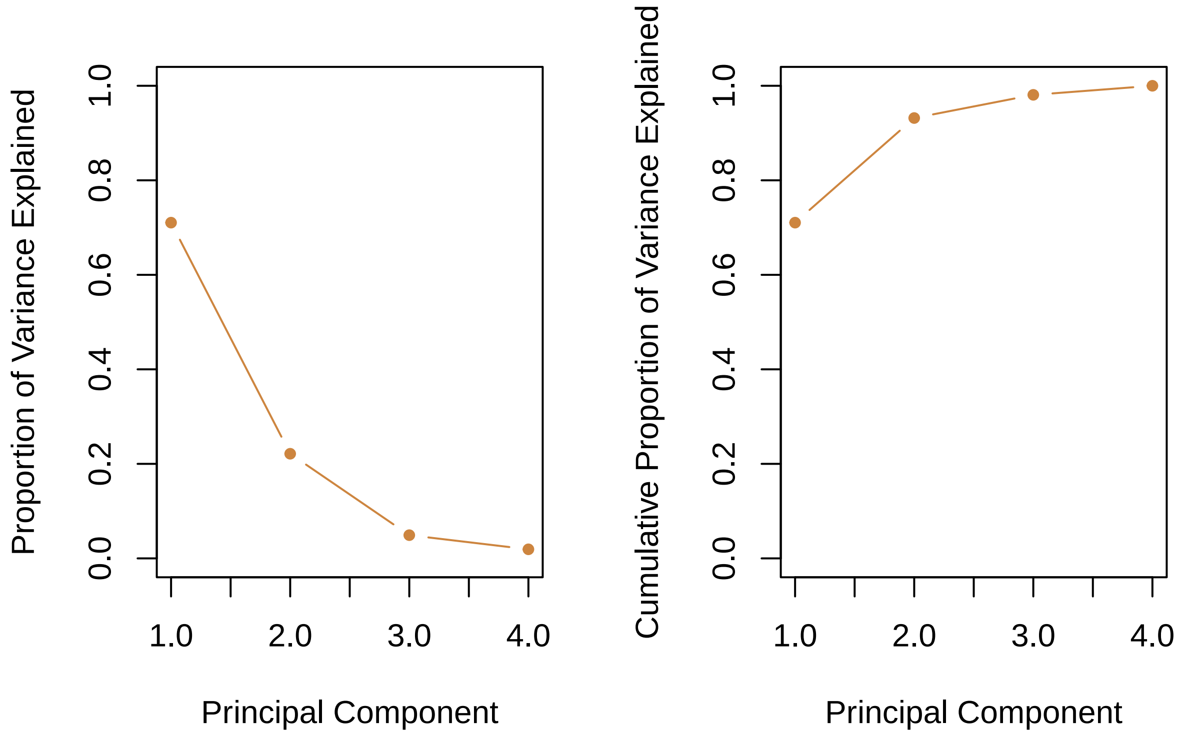
Of course, it's much better to look at the `biplot`.

```r
biplot(pr.out, scale=0, cex=0.5, xlabs=rep("*", length(price)),
       col=c("lightblue", "red"))
```



How many principal components are "sufficient"?

```
#transforming standard deviations to variances
pr.var=pr.out$sdev^2
#pr.var
#proportion of variance explained
pve=pr.var/sum(pr.var)
#pve
#plots
par(mfrow = c(1, 2))
plot(pve,xlab="Principal Component",
     ylab="Proportion of Variance Explained", col="peru",
     pch=20, ylim=c(0,1),type='b')
plot(cumsum(pve),xlab="Principal Component",
     ylab="Cumulative Proportion of Variance Explained",
     col="peru", pch=20, ylim=c(0,1),type='b')
```


Proportion of Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
1.0
2.0
3.0
4.0
Principal Component
Cumulative Proportion of Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
1.0
2.0
3.0
4.0
Principal Component