# Principal Component Analysis

*Rommel Bartolome*

*March 15, 2019*

## Data Loading

First, we set the city names as row names and remove it from the dataset that we will use in the analysis:

```
cities <- read.csv("cities.csv")
rownames(cities) <- cities[,1]
cities[,1] <- NULL
```

## 1) How many principal components can adequately explain the variation in the data? Justify your choice.

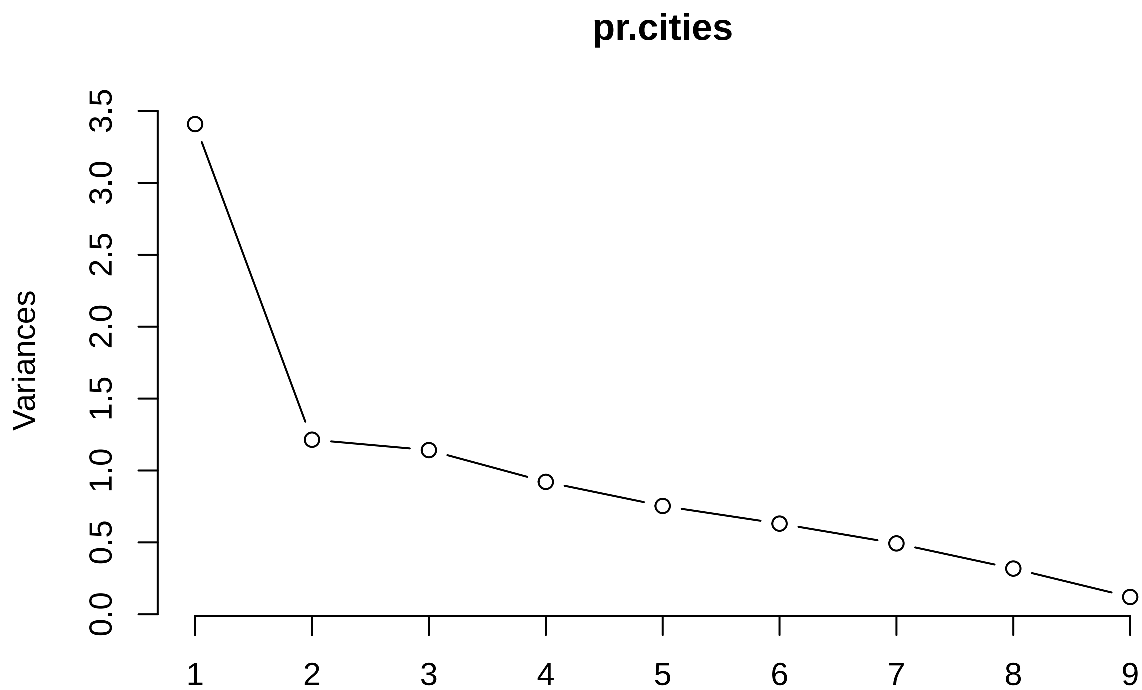
We first check principal components of each:

```
pr.cities <- prcomp(cities, scale = T)
pr.cities$rotation
```

```
##                   PC1        PC2          PC3         PC4        PC5
## Climate     0.2064140  0.2178353 -0.689955982  0.13732125 -0.3691499
## HousingCost 0.3565216  0.2506240 -0.208172230  0.51182871  0.2334878
## HlthCare    0.4602146 -0.2994653 -0.007324926  0.01470183 -0.1032405
## Crime       0.2812984  0.3553423  0.185104981 -0.53905047 -0.5239397
## Transp      0.3511508 -0.1796045  0.146376283 -0.30290371  0.4043485
## Educ        0.2752926 -0.4833821  0.229702548  0.33541103 -0.2088191
## Arts        0.4630545 -0.1947899 -0.026484298 -0.10108039 -0.1050976
## Recreat     0.3278879  0.3844746 -0.050852640 -0.18980082  0.5295406
## Econ        0.1354123  0.4712833  0.607314475  0.42176994 -0.1596201
##                     PC6         PC7         PC8           PC9
## Climate      0.37460469 -0.08470577 -0.36230833  0.0013913515
## HousingCost -0.14163983 -0.23063862  0.61385513  0.0136003402
## HlthCare    -0.37384804  0.01386761 -0.18567612 -0.7163548935
## Crime        0.08092329  0.01860646  0.43002477 -0.0586084614
## Transp       0.46759180 -0.58339097 -0.09359866  0.0036294527
## Educ         0.50216981  0.42618186  0.18866756  0.1108401911
## Arts        -0.46188072 -0.02152515 -0.20398969  0.6857582127
## Recreat      0.08991578  0.62787789 -0.15059597 -0.0255062915
## Econ         0.03260813 -0.14974066 -0.40480926  0.0004377942
```

We check the screeplot:

```
screeplot(pr.cities, type = 'l')
```




In the screeplot, we can see that 2 is an adequate number of principal components to explain the variation in the data (elbow).

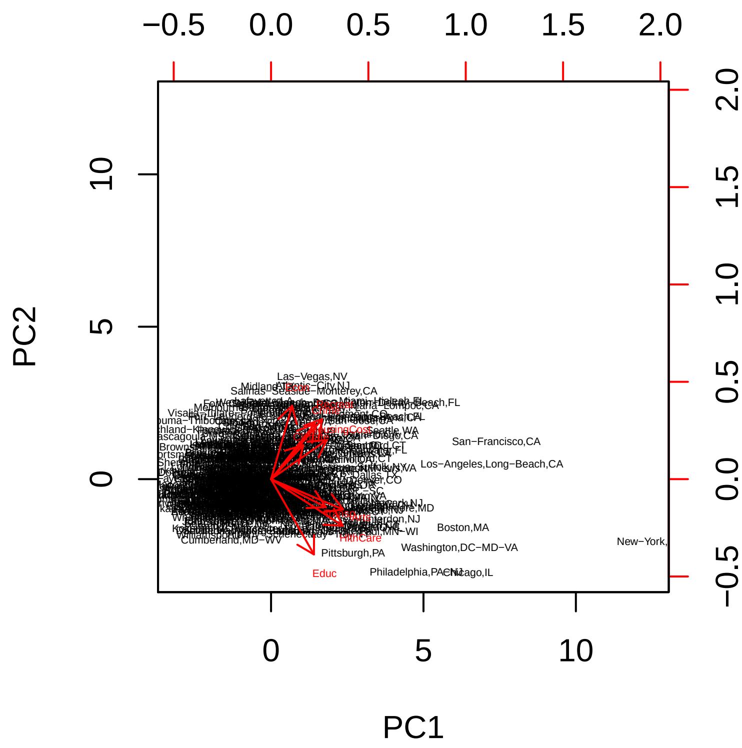
### 2) Interpret the principal components that you have chosen: basically, describe the characteristics contained in those PCs.

From `pr.cities$rotation`, `Educ` and `Econ` are the heaviest components of PC 2. `Education` or the pupil/teacher ratio in the public K-12 system, effort index in K-12 and academic options in higher education together with `Economics` or the average household income adjusted for taxes and living costs, income growth, job growth are the heaviest components of PC2. This is not surprising, as receiving a good education and having good economic status (better jobs because they have higher education) will likely explain a "better city". These components also affect other components. For example, if you have education, thus more money, you'll likely have better health care, no need for you to do crime and more time for recreation and the arts.

### 3) Construct a biplot containing the first two PCs. Provide a brief interpretation by identifying cities that tend to score high (or low) in the first and/or second PCs.

We construct the a biplot of the first PCs.

```
biplot(pr.cities, scale = 0, choices = 1:2, cex = 0.3)
```




Here, we can see that New York has a high PC1 score. On the other hand, Philadelphia and Chicago have low PC2 scores.