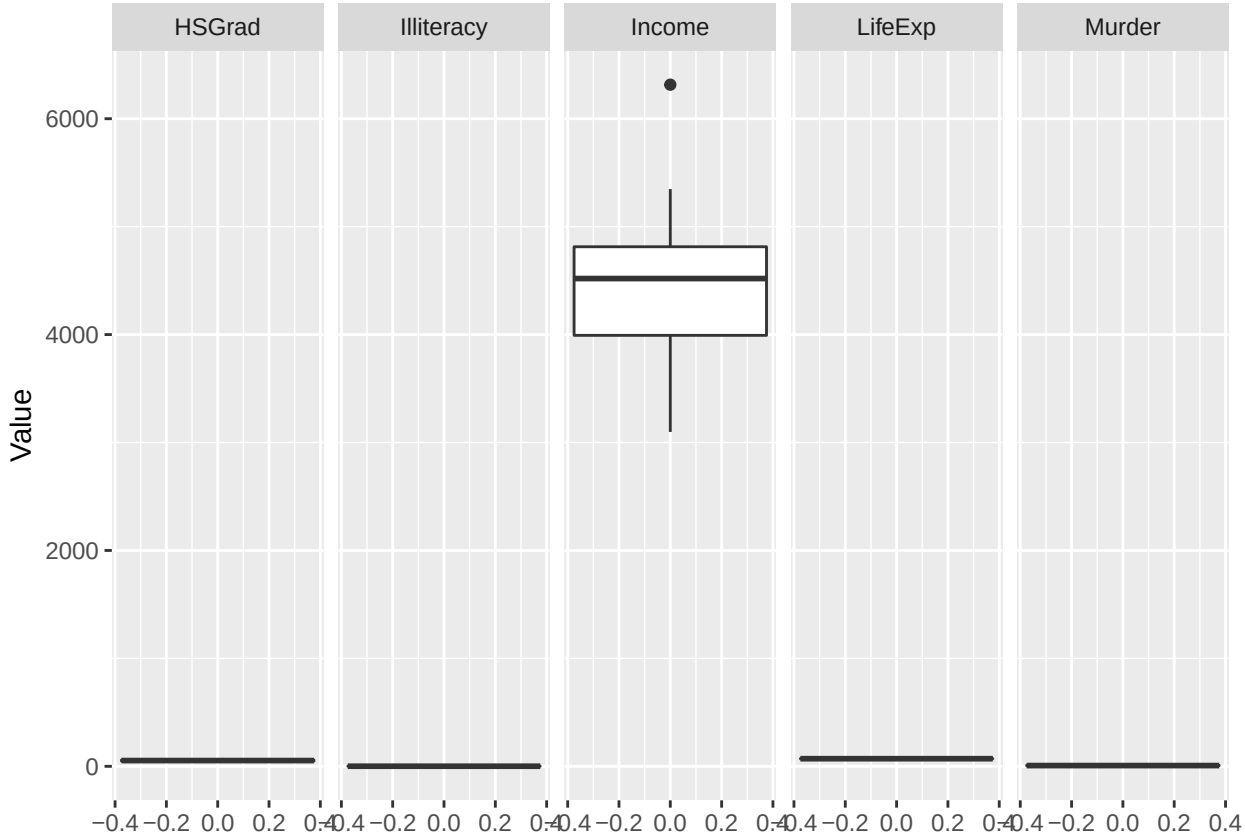
Assignment - A bad example

High Dimensional Data Analysis

9 August 2018

Preliminary Analysis

Before carrying out a Principal Components Analysis it is worth exploring the features of the original variables themselves. Box plots of each variable are provided below.



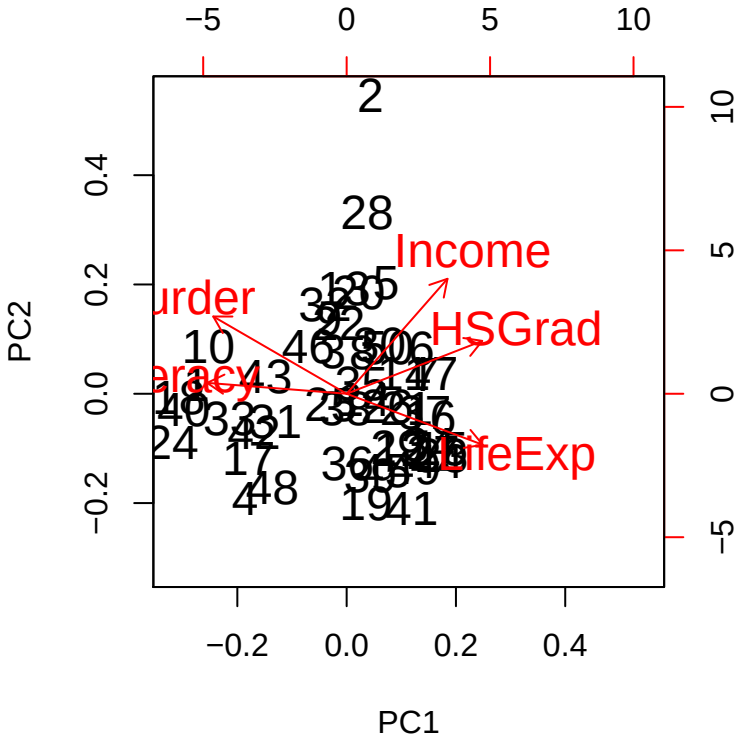
On a boxplot the line in the middle of the box is the median, while the lines at the top and bottom of the box are the third and first quartiles respectively. The median of income is about \$4500, while the first quartile is \$4000 and the third quartile is about \$4750.

Data description

Data are collected on five variables each representing different measures of welfare for the 50 states. These are Income, Illiteracy, LifeExp, Murder and the HSGrad. An analysis of these data, and in particualar a biplot based on Principal Components can help us to glean insight into the quality of life in different regions of the USA.

Principal Components Analysis and Biplot

Principal components analysis finds a small number of linear combinations of the orginal variables that explain a large proportion of overall variation in the data. The weights in principal components are normalised so that $\sum w^2 = 1$. The weights are found as solutions to the eigenvalue equation $Sw = \lambda w$ where S is the variance covariance matrix of the data. By selecting two principal components we are able visualise the data using a biplot which is included below

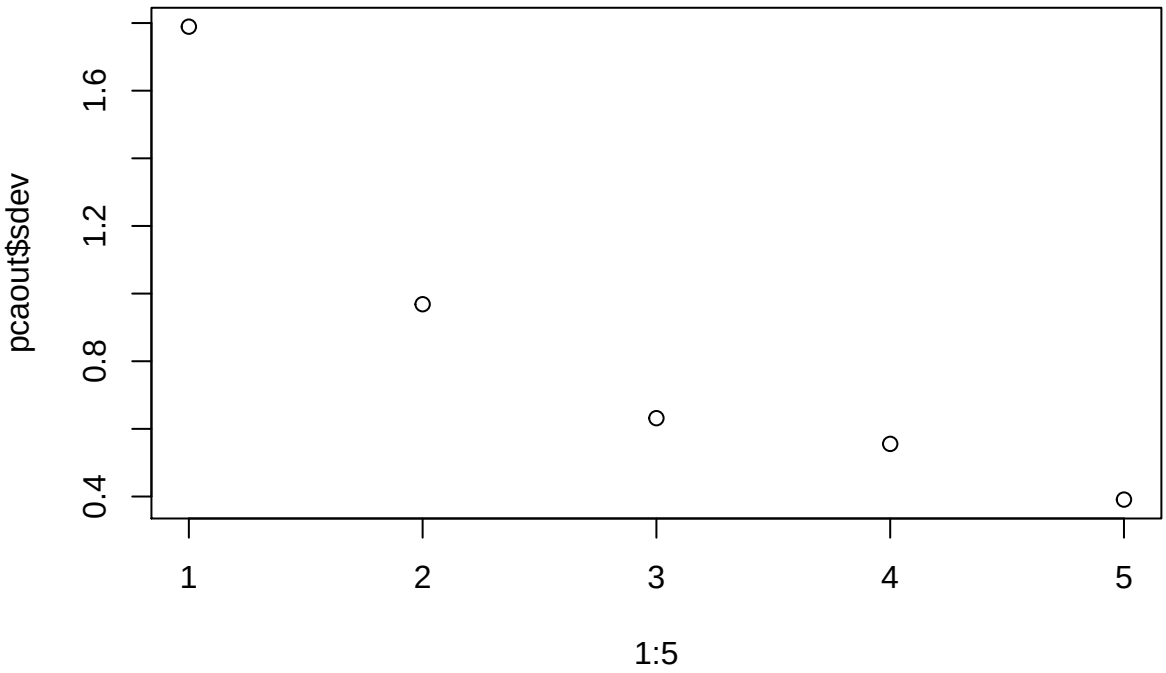


The biplot can be interpreted as follows. Points that are close to one another have similar characteristics, for example State 24 and State 40. If the angle between two arrows is small then they are likely to have a high correlation, if the angle between them is 90 degrees they are more likely to be uncorrelated and if the angle between them is 180 degrees they are likely to be negatively correlated.

The analysis was conducted using R. More specifically, first the data was loaded using the `load` function. Principal components analysis was conducted using the `prcomp` function. The output of this function is a prcomp object which was assigned the name `pcaout`. By running the command `summary(pcaout)` the following output was obtained

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.7892 0.9686 0.6317 0.55561 0.39093
## Proportion of Variance 0.6403 0.1876 0.0798 0.06174 0.03057
## Cumulative Proportion  0.6403 0.8279 0.9077 0.96943 1.00000
```

This table can be interpreted as follows. The standard deviations of the principal components are 1.7892, 0.9686, 0.6317, 0.55561 and 0.39093 respectively. The proportion of variances are 0.6403,0.1876,0.0798,0.06174,0.03057 respectively. The cumulative proportions are 0.6403,0.8279,0.9077,0.96943 and 1 respectively. The cumulative proportions are obtained by adding up the proportion of variance. Below is a plot of the standard deviation of each principal component.



Finally, here are the first two principal components for all states

```
##               PC1         PC2
##  [1,] -3.47364293  0.11782907
##  [2,]  0.55234580  3.74211720
##  [3,] -0.32181787  0.52727395
##  [4,] -2.35182398 -1.31101418
##  [5,]  0.91383190  1.38295436
##  [6,]  1.73193488  0.56709878
##  [7,]  1.82930700  0.24175232
##  [8,]  0.37084429  0.51666435
##  [9,] -0.40719735  0.94030180
## [10,] -3.20002323  0.59670753
## [11,]  1.32751386  0.26226425
## [12,]  1.24430964 -0.69209303
## [13,] -0.05866117  1.31483568
## [14,]  0.40598302 -0.07303673
## [15,]  2.19608922 -0.68458154
## [16,]  1.92568855 -0.29876790
## [17,] -2.26525696 -0.80388522
## [18,] -3.88265631 -0.04210000
## [19,]  0.45475710 -1.37518987
## [20,]  0.28444779  1.27187283
## [21,]  1.38689717 -0.17640257
## [22,] -0.17684646  0.88747403
## [23,]  2.20252811 -0.77221825
## [24,] -4.03622188 -0.60103001
## [25,] -0.36527017 -0.13636916
## [26,]  0.93592565 -0.15653928
## [27,]  2.00609608 -0.73333962
## [28,]  0.47198085  2.27210610
## [29,]  1.17273421 -0.67867616
## [30,]  0.76185895  0.61808081
```

```
## [31,] -1.64651957 -0.34537962
## [32,] -0.49376349  1.11381859
## [33,] -2.70360336 -0.30543307
## [34,]  1.90492368 -0.63934522
## [35,]  0.34446552  0.12790140
## [36,]  0.02272511 -0.87089641
## [37,]  1.80664829 -0.23652764
## [38,] -0.02423428 -0.18650858
## [39,]  0.55482029 -1.02885203
## [40,] -3.77227120 -0.19479975
## [41,]  1.51310495 -1.43468754
## [42,] -2.13795104 -0.47447133
## [43,] -1.87436137  0.22141008
## [44,]  2.09950898 -0.83819355
## [45,]  0.88055717 -0.96221085
## [46,] -0.88105357  0.59641273
## [47,]  1.96875345  0.26448348
## [48,] -1.71318052 -1.18019012
## [49,]  1.57284365 -0.92716268
## [50,]  0.94293158  0.57654255
```