# Lab 8: Principal Component Analysis

Final report due: Friday, October 15, 2021, by 11:59 pm.

Electronically submit one solution per group, including the Rmd file and the output file (either pdf or html). Please record the names of the group members who attended the lab.

# Purpose

• Learn how to perform and interpret principal component analysis (PCA). • Use principal componet scores to display and summarize differences in groups • Check the data prior to performing the PCA. • Diagnostic follow-up of the analysis.

# Data on Ames Livability

Ten rating variables of the livability in American small cities data. The data come from a book that was published in the late 90s, on livability of small cities in the USA. The Des Moines Register featured the book, because Ames was ranked the number 2 best small city in which to live! The study used 10 ratings variables -- Climate, Diversions, Economic, Education, Community, Health, House, Safety, Transportation, Urban -- with each city getting a rating between 0-100 on each of these varaibles. The scores were combined to give an overall rating for each city, Score.

towns <- read.csv("towns2.csv")  
dim(towns)

## [1] 193 16

head(towns)

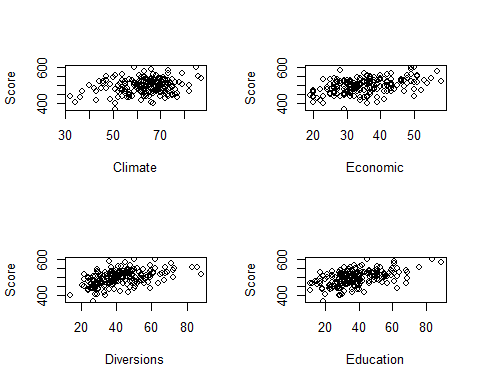
## City ID Population Climate Diversions Economic Education  
## 1 1.Albertville.AL 1 76802 68 44 43 12  
## 2 2.Auburn-Opelika.AL 2 91869 65 27 30 56  
## 3 3.Cullman.AL 3 71615 71 26 32 11  
## 4 4.Selma.AL 4 47991 64 28 26 28  
## 5 5.Talladega.AL 5 76034 62 26 20 15  
## 6 6.Fairbanks.AK 6 84711 32 59 50 63  
## Community Health House Safety Transportation Urban Score Latitude  
## 1 25 30 87 68 17 76 470 34.268  
## 2 31 55 77 60 36 72 509 32.610  
## 3 37 36 85 69 24 78 469 34.175  
## 4 49 35 83 32 36 61 442 32.407  
## 5 34 39 87 62 32 84 461 33.436  
## 6 45 37 45 55 56 0 442 50.000  
## Longitude  
## 1 -86.209  
## 2 -85.481  
## 3 -86.844  
## 4 -87.021  
## 5 -86.106  
## 6 -130.000

towns$ames <- 0  
towns$ames[52] <- 1

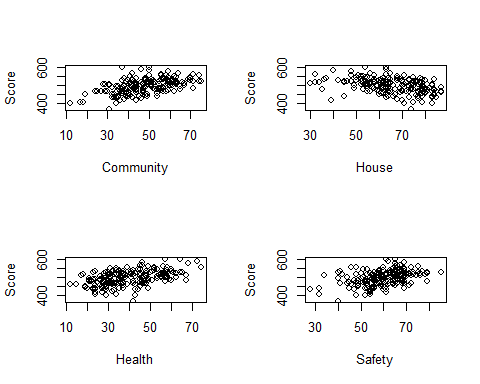
# Exercise 1

Examine how the the ten ratings variables are correlated with the livability scores (the Score variable) given to the cities. Write a short summary of your findings.

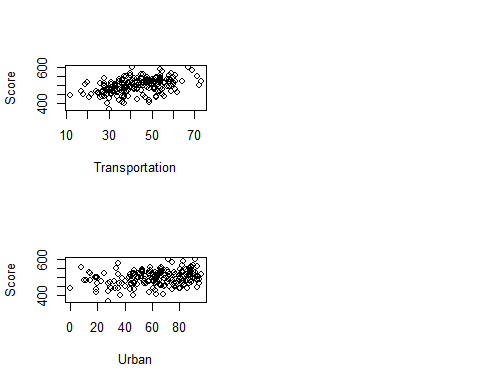
layout(matrix(1:4, nrow = 2, ncol = 2))  
plot(Score ~ Climate, data = towns, pch = 1)  
plot(Score ~ Diversions, data = towns, pch = 1)  
plot(Score ~ Economic, data = towns, pch = 1)  
plot(Score ~ Education, data = towns, pch = 1)



plot(Score ~ Community, data = towns, pch = 1)  
plot(Score ~ Health, data = towns, pch = 1)  
plot(Score ~ House, data = towns, pch = 1)  
plot(Score ~ Safety, data = towns, pch = 1)



plot(Score ~ Transportation, data = towns, pch = 1)  
plot(Score ~ Urban, data = towns, pch = 1)  
layout(matrix(1:1, nrow = 1, ncol = 1))



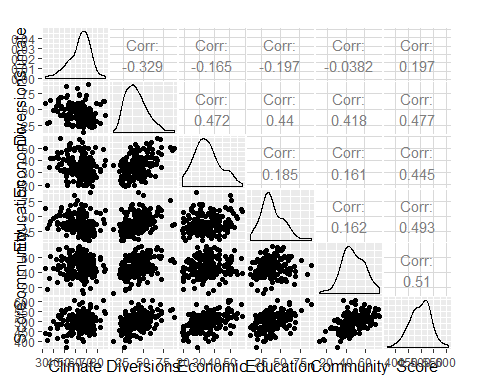
cor(towns$Score, towns[, 4:13])

## Climate Diversions Economic Education Community Health  
## [1,] 0.1974325 0.4768799 0.4449764 0.4930572 0.5098451 0.477609  
## House Safety Transportation Urban  
## [1,] -0.3868395 0.381705 0.436461 0.2447653

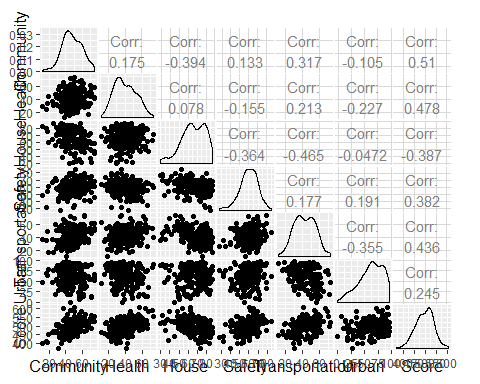
# Exercise 2

All the scores are measured on a scale of 0-100, so why is it still necessary to use the correlation matrix, or standardize the data, before doing a principla component analysis (PCA)? (Hint: Compute some summary statistics or make some plots.)

ggpairs(towns[, c(4:8, 14)])

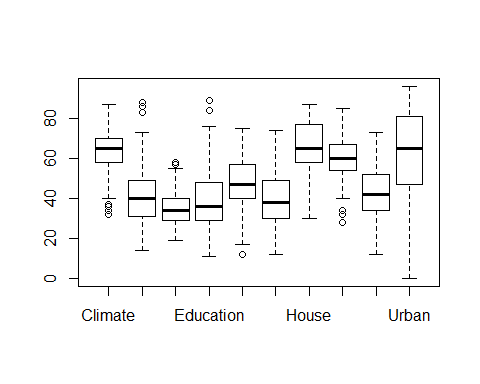


ggpairs(towns[, c(8:14)])



sapply(towns[, 3:13], sd)

boxplot(towns[, 4:13])



The best function to use in R is prcomp which handles a large number of variables better than other functions. The scale argument tells R to use the correlation matrix (TRUE) or covariance matrix (FALSE, default), and the retx argument tells R to compute the principal component scores and save them in the object that is created by the prcomp function.

The output has several main components: sdev contains the square root of the eigenvalues, rotation are the eigenvectors, and x contains the PC scores.

# Exercise 3

Present a summary of the PCA including the table of eigenvectors, a list of eigenvalues (variance), and cumulative percentage of total variance explained by the principal compontnes. (Be sure to make your output readable, eg rounding digits appropriately.) Make a scree plot.

1. How many PCs would you need to use to explain 80% of the total variation?
2. Explain how the Cumulative Proportion row of the summary of the PC was calculated.

towns.pca <- prcomp(towns[, 3:13], scale. = T) # include Population  
summary(towns.pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 1.8364 1.3937 1.0790 0.94910 0.90637 0.84211 0.8022  
## Proportion of Variance 0.3066 0.1766 0.1058 0.08189 0.07468 0.06447 0.0585  
## Cumulative Proportion 0.3066 0.4832 0.5890 0.67088 0.74556 0.81003 0.8685  
## PC8 PC9 PC10 PC11  
## Standard deviation 0.74464 0.61670 0.55731 0.44812  
## Proportion of Variance 0.05041 0.03457 0.02824 0.01826  
## Cumulative Proportion 0.91893 0.95351 0.98174 1.00000

round(towns.pca$rotation, 4)

## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Population -0.0673 0.3867 0.0285 -0.7216 0.4577 -0.0968 0.1446  
## Climate 0.2735 0.2297 0.5405 0.0745 -0.0480 -0.0566 -0.2517  
## Diversions -0.4440 -0.0790 0.0845 0.1483 0.0619 0.3743 0.0312  
## Economic -0.2668 -0.2633 0.4425 -0.2877 -0.1734 0.4455 0.3787  
## Education -0.3777 0.1921 -0.0721 -0.1111 -0.4316 -0.0357 -0.5235  
## Community -0.2789 0.0855 0.3598 0.4920 0.5843 -0.1595 -0.0167  
## Health -0.1788 -0.3706 0.4670 -0.2037 -0.1896 -0.5442 -0.1075  
## House 0.3842 -0.3853 0.0167 -0.0511 0.0120 -0.2543 0.2000  
## Safety -0.1158 0.4614 0.0322 0.2397 -0.4125 -0.2938 0.6531  
## Transportation -0.4230 0.0328 -0.1917 -0.0971 0.0741 -0.3876 -0.0097  
## Urban 0.2449 0.4220 0.3366 -0.0602 -0.1221 0.1573 -0.1369  
## PC8 PC9 PC10 PC11  
## Population 0.1413 0.1947 0.1067 -0.1243  
## Climate 0.6355 -0.2434 0.0348 0.2039  
## Diversions 0.1895 -0.1600 0.6712 -0.3325  
## Economic 0.0175 -0.0832 -0.4452 0.0591  
## Education 0.1673 0.2559 -0.2946 -0.4012  
## Community -0.1022 0.1907 -0.3093 -0.1839  
## Health -0.2541 0.1911 0.3313 0.1288  
## House 0.2031 -0.1499 -0.1016 -0.7226  
## Safety 0.0915 0.1258 0.0684 -0.0564  
## Transportation 0.0031 -0.7646 -0.1653 0.0778  
## Urban -0.6224 -0.3302 0.0641 -0.2935

# Exercise 4

1. Explain how the variables contribute to the first two principal components.
2. Using three pieces of information, where the elbow is in the scree plot, the proportion of total variation, and the interpretation of the PCs, make an argument for how many PCs would you recommend to summarise this data?

# Exercise 5

Compare the scores for the first principal component with the Score variable in the data (this is the rating the article gives for each city). Which city would be rated first using the Score variable? Which city would be rated first using the scores for the first principal component? (You could make a plot of the Score variable against PC1, and compute the correlation between the two variables.) Do the two approaches give cities similar ratings? (You may need to multiple the scores for the first principal component by -1).

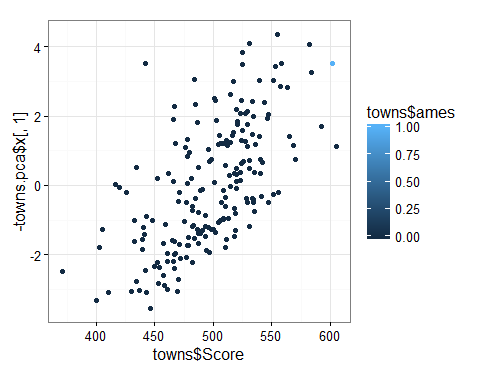
head(rownames(towns)[order(towns$Score, decreasing = T)])

## [1] "181" "52" "188" "108" "74" "182"

head(rownames(towns)[order(towns.pca$x[, 1], decreasing = T)])

## [1] "113" "26" "169" "9" "176" "3"

library(ggplot2)  
qplot(towns$Score, -towns.pca$x[, 1], col=as.factor(towns$ames)) + theme\_bw() + theme(aspect.ratio=1)



# Final Instructions

Put the names of all members of your group who participated in the Lab on Lab Report before you knit it.

Final report due: Friday, October 15, 2021, by 11:59 pm.