Q\_2.R

Nadim

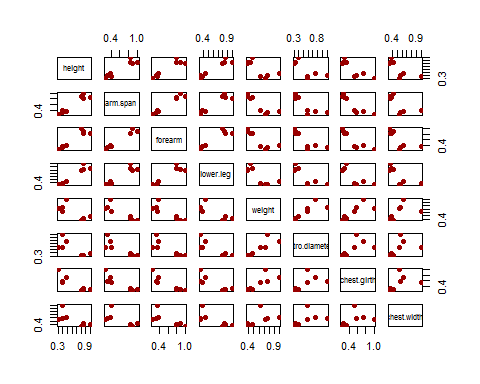
Thu Jan 31 02:58:08 2019

#######Question 2  
# 2Q) The dataset Harmon23.cor in the datasets package is a correlation matrix of eight   
# physical measurements made on 305 girls between the ages of 7 and 17.  
  
library('datasets')  
data\_<-Harman23.cor  
dat\_<-Harman23.cor$cov  
  
#####################################################################################  
# 2a) Provide a plot to visualize this matrix, and comment on any patterns you see.  
  
pairs(dat\_,pch=19,col="#990000")  
library(GGally)

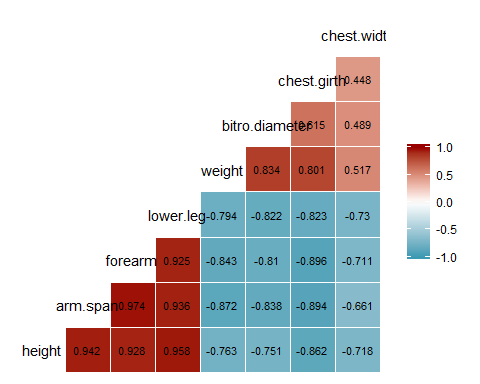
## Warning: package 'GGally' was built under R version 3.5.2

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.5.2



ggcorr(dat\_, low = "#3B9AB2", mid = "#FFFFFF", high = "#990000",label = T, label\_color = "black",label\_size = 3, label\_round = 3)



#####################################################################################  
# 2b) We want to perform PCA on this matrix (Harman23.cor$cov), and retain the first   
# two PCs. How can we do this? [Note: this is not the actual dataset, you only have   
# the correlation matrix of the variables.]  
# The function eigen() can compute the eigenvectors/eigenvalues  
eig.out<-eigen(dat\_)  
str(eig.out)

## List of 2  
## $ values : num [1:8] 4.673 1.771 0.481 0.421 0.233 ...  
## $ vectors: num [1:8, 1:8] -0.398 -0.389 -0.376 -0.388 -0.351 ...  
## - attr(\*, "class")= chr "eigen"

#Eigen Values  
lam <- eig.out$values  
tab <- rbind(lam,  
 lam/sum(lam), # proportion of variance explained  
 cumsum(lam)/sum(lam)) # cumulative proportion of var explained  
rownames(tab) <- c("Variance", "Proportion of variance", "Cumulative proportion")  
  
# Eigenvalues (variance of PCs)  
lambda <- tab[1,]  
# scree plots  
par(mfrow = c(1,2))  
plot(lambda, type="b", pch = 19, main = "Variance explained by each PC",  
 xlab = "PC number", ylab = "Variance explained")  
# 2c) How much of the total variation is captured by each of the first two PCs? How   
# much of the total variation is captured by the two PCs together?  
  
tab[3,1:2]

## [1] 0.5841099 0.8054828

# 2d) Interpret the first two PCs.  
###The frst PC Vector kind of represents overall body metrics  
###IN the Second PC vector, the first half represents body lengths and the second  
### half represents the body girth measurements()

