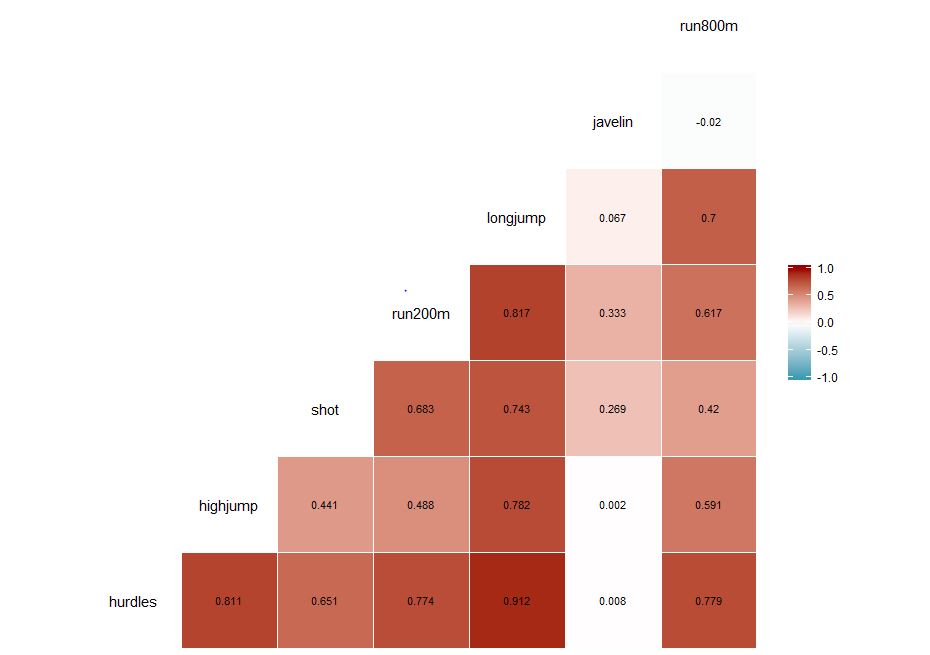
Q\_4.R

####Question 4  
  
# 4Q) Consider the heptathlon data in the HSAUR3 package. See ?heptathlon for details   
# about the dataset.  
############################################################################## 4a) Take only the first seven columns representing the seven events. Notice that in   
# the events high jump, long jump, shot and javelin, larger values indicate better   
# performance. But for the other three events (200m, 800m, and hurdles) smaller values   
# indicate better performance. To help with interpretation, transform the data of the   
# latter three events as "newx <- max(x) - x" so that for all the variables larger values   
# indicate better performance. Visualize the correlation matrix and comment on any pattern   
# you see  
  
data("heptathlon", package = "HSAUR3")  
a=heptathlon[,8]  
heptathlon[,8]=NULL  
heptathlon[,1]<-(max(heptathlon[,1])-heptathlon[,1])  
heptathlon[,4]<-(max(heptathlon[,4])-heptathlon[,4])  
heptathlon[,7]<-(max(heptathlon[,7])-heptathlon[,7])  
library(GGally)

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



*##we can observe obvious patterns or high corr scores for sports of similar nature  
##but also we can see the correlation between sports of different natures like shot and longjump  
##########  
#####################################################################################*# 4b) Perform PCA on the new dataset. Summarize and interpret the results, especially   
# the first two PCs. Note that you might need to standardize the data.  
  
std.data\_4<-scale(heptathlon,center=T,scale=T)  
dat\_4.pca<-prcomp(std.data\_4)  
summary(dat\_4.pca)

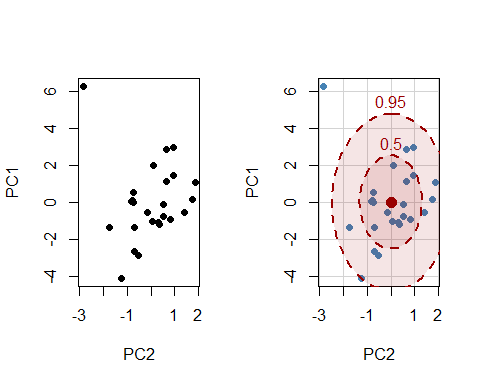
## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 2.1119 1.0928 0.72181 0.67614 0.49524 0.27010  
## Proportion of Variance 0.6372 0.1706 0.07443 0.06531 0.03504 0.01042  
## Cumulative Proportion 0.6372 0.8078 0.88223 0.94754 0.98258 0.99300  
## PC7  
## Standard deviation 0.2214  
## Proportion of Variance 0.0070  
## Cumulative Proportion 1.0000

round(dat\_4.pca$rotation[,1:2],3)

## PC1 PC2  
## hurdles -0.453 0.158  
## highjump -0.377 0.248  
## shot -0.363 -0.289  
## run200m -0.408 -0.260  
## longjump -0.456 0.056  
## javelin -0.075 -0.842  
## run800m -0.375 0.224

*#####We can observe in the first PC that there except for javelin, the rest of the variables  
#####have significant components.(which i think reflects the core and cardio strength)  
####the second PC has a huge javelin component.(which i think reflects the upper body strength)*  
  
*#####################################################################################*# 4c) Compute the PC scores for the first two PCs and create a scatterplot. Do you see   
# any pattern? If yes, investigate further and comment on your findings.  
par(mfrow=c(1,2))  
plot(dat\_4.pca$x[,2:1],pch=19)  
library(car)

dataEllipse(dat\_4.pca$x[,2:1], pch=19, col=c("steelblue","#990000"),lty=2,ellipse.label=c(0.5,0.95),levels=c(0.5,0.90),fill=TRUE,fill.alpha=0.1)



*###############  
##Observation: Except for an outlier the rest of the data cloud seems to be centered at around the mean  
###############*  
#4d) The last column of the heptathlon dataset provides the official scores given to   
# the athletes for the event. Note that your PCA did not involve the score information at   
# all. Consider PC1 (a summary of the performance). Plot the official scores versus your   
# PC1 scores in a scatterplot. Do you think your summary of performance, PC1, aligns with   
# the official scores? Comment on your findings.  
plot(a,dat\_4.pca$x[,1],xlab = 'Official\_scores',ylab = 'PCA\_1',pch=19)

