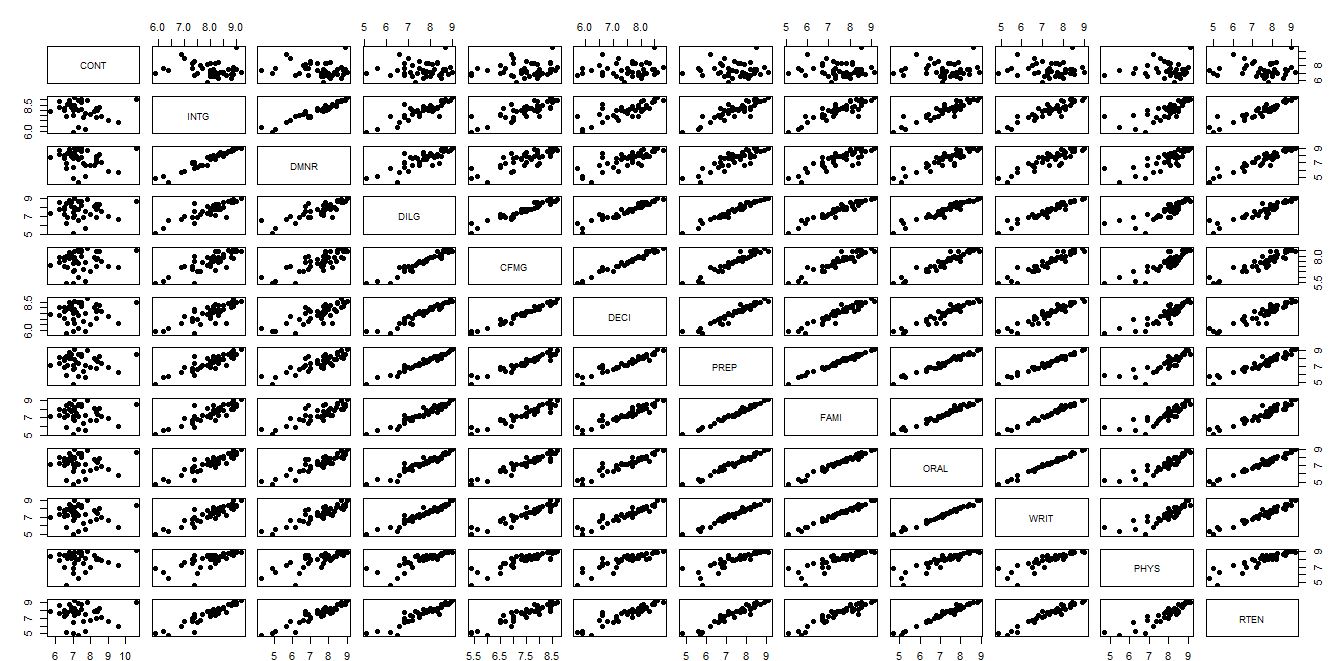
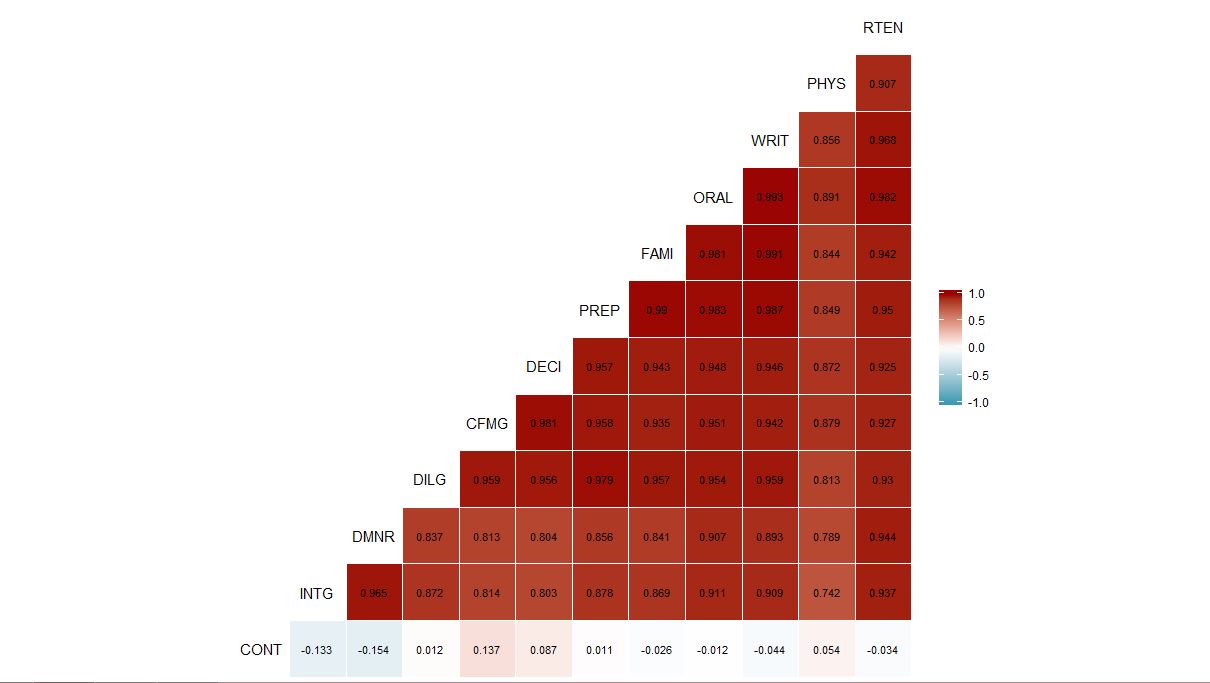
Question 1.R

# 1Q) Examine the USJudgeRatings data in the datasets library. This dataset contains   
# the ratings of 43 US Superior Court judges by attorneys. Each of the judges is evaluated   
# on each of 12 attributes such as demeanor, preparation for trial, sound rulings, and the   
# number of contacts each attorney had with that judge. See the R help file for more   
# information on this dataset.  
  
  
library('datasets')  
data\_<-USJudgeRatings  
#visulaising the datasetusing pairs plot  
pairs(data\_,pch=19,col="#990000")  
#visualising the dataset using correlation matrix  
library(GGally)



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



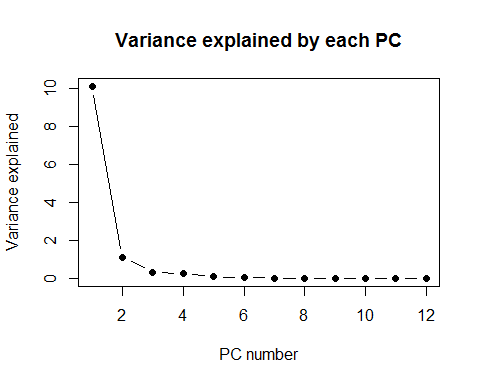
#to check if dataset has similar standard deviations, if not we need to   
#standardize them  
apply(data\_,2,sd)

## CONT INTG DMNR DILG CFMG DECI PREP   
## 0.9408768 0.7701447 1.1437054 0.9008978 0.8601102 0.8029362 0.9533702   
## FAMI ORAL WRIT PHYS RTEN   
## 0.9489868 1.0100437 0.9611328 0.9395753 1.1009711

##scaling  
std.data\_<-scale(data\_,center=T,scale=T)  
  
#Performing PCA  
  
data\_.pca<- prcomp(std.data\_)  
summary(data\_.pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 3.1833 1.05078 0.57698 0.50383 0.29061 0.19310  
## Proportion of Variance 0.8445 0.09201 0.02774 0.02115 0.00704 0.00311  
## Cumulative Proportion 0.8445 0.93647 0.96421 0.98537 0.99240 0.99551

lambda<- data\_.pca$sdev^2  
plot(lambda, type="b", pch = 19, main = "Variance explained by each PC",  
 xlab = "PC number", ylab = "Variance explained")



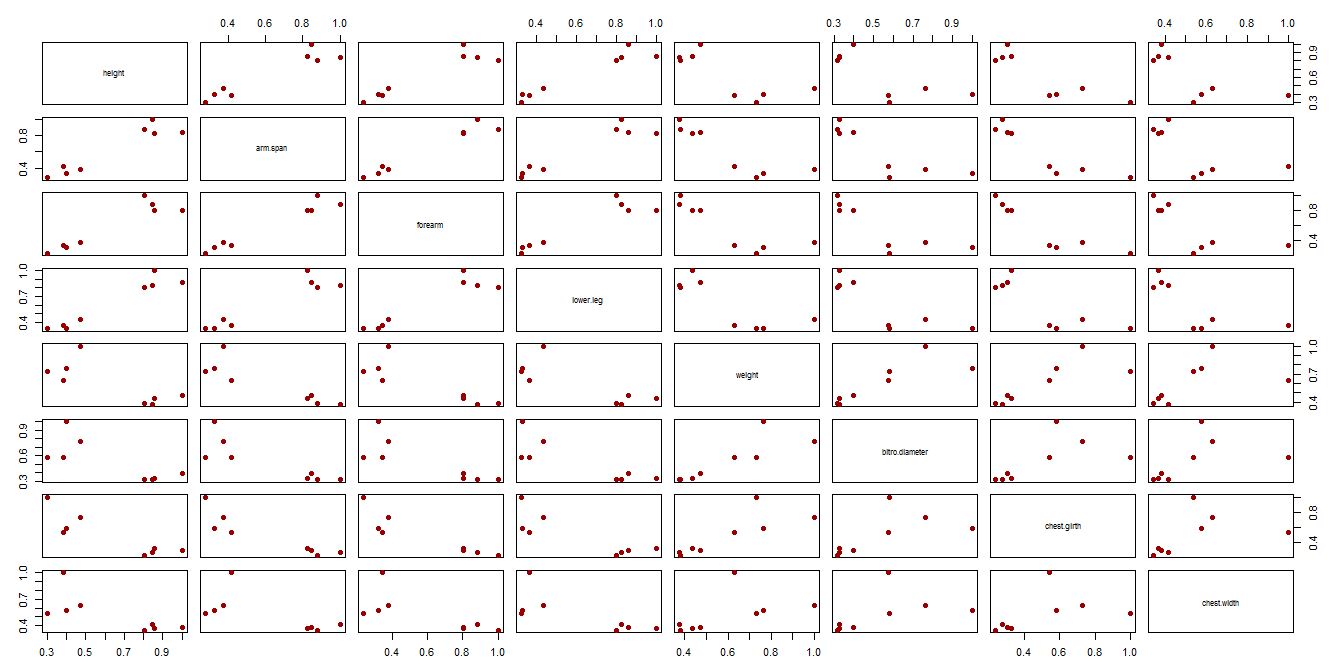
#first two components explain 93.6% of variability, so we can retain those two  
#since the loadings are the same as eigen vectors  
  
a<-round(data\_.pca$rotation,3)  
a[,1:2]

## PC1 PC2  
## CONT 0.003 -0.933  
## INTG -0.289 0.182  
## DMNR -0.287 0.198  
## DILG -0.304 -0.036  
## CFMG -0.303 -0.168  
## DECI -0.302 -0.128  
## PREP -0.309 -0.032  
## FAMI -0.307 0.001  
## ORAL -0.313 0.004  
## WRIT -0.311 0.031  
## PHYS -0.281 -0.089  
## RTEN -0.310 0.039

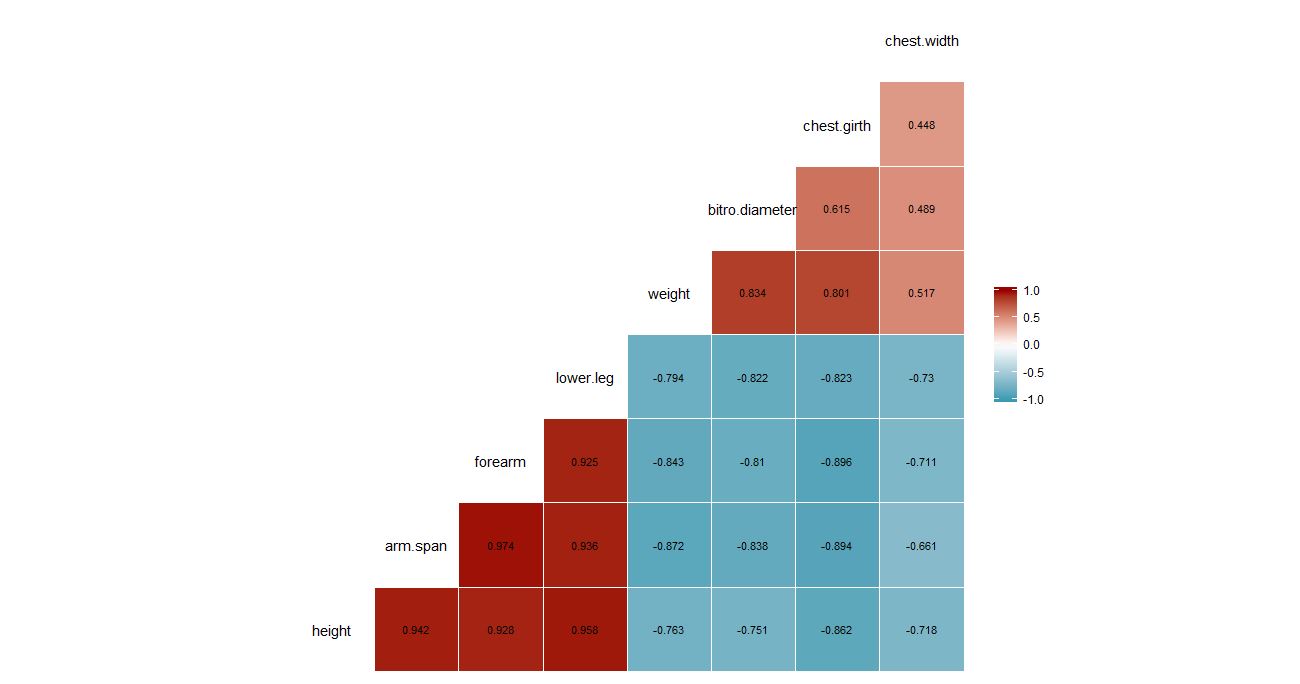
####Interpreting the loadings of the PCs  
####(i)First eigen shows component for every variable except CONT and second the PC   
####shows the opposite to the first  
####(ii)the same can be observed in the first plot where cont does not show any correlation  
  
###We can see from the matrix in part(a) that the covariance of cont with the rest\  
###of the variables is negligible. Hence the variance defined by CONT variable will have   
####have to be defined by a PC which has high weightage to the CONT variable. This   
###can be confirmed with the PC loadings

Question 2.R

#######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)



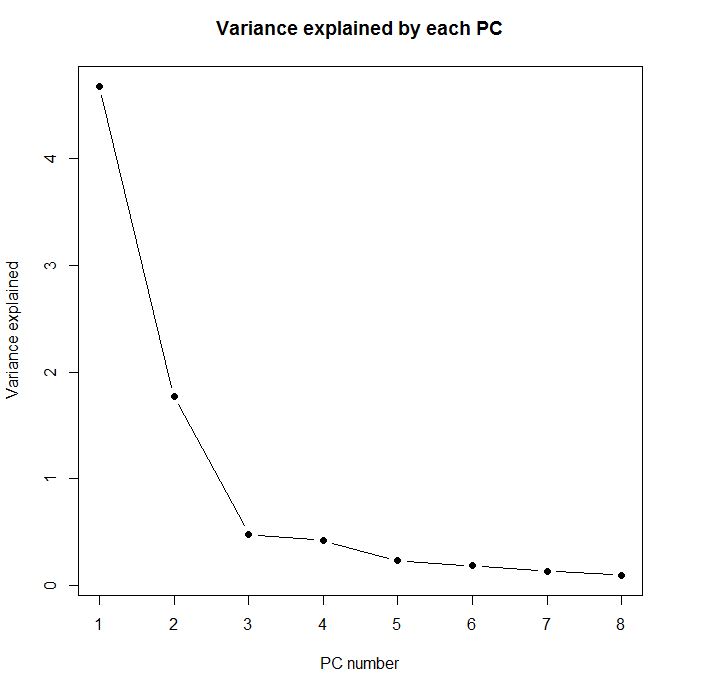
ggcorr(dat\_, low = "#3B9AB2", mid = "#FFFFFF", high = "#990000",label = T, label\_color = "black",label\_size = 3, label\_round = 3)



#####################################################################################  
# 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")

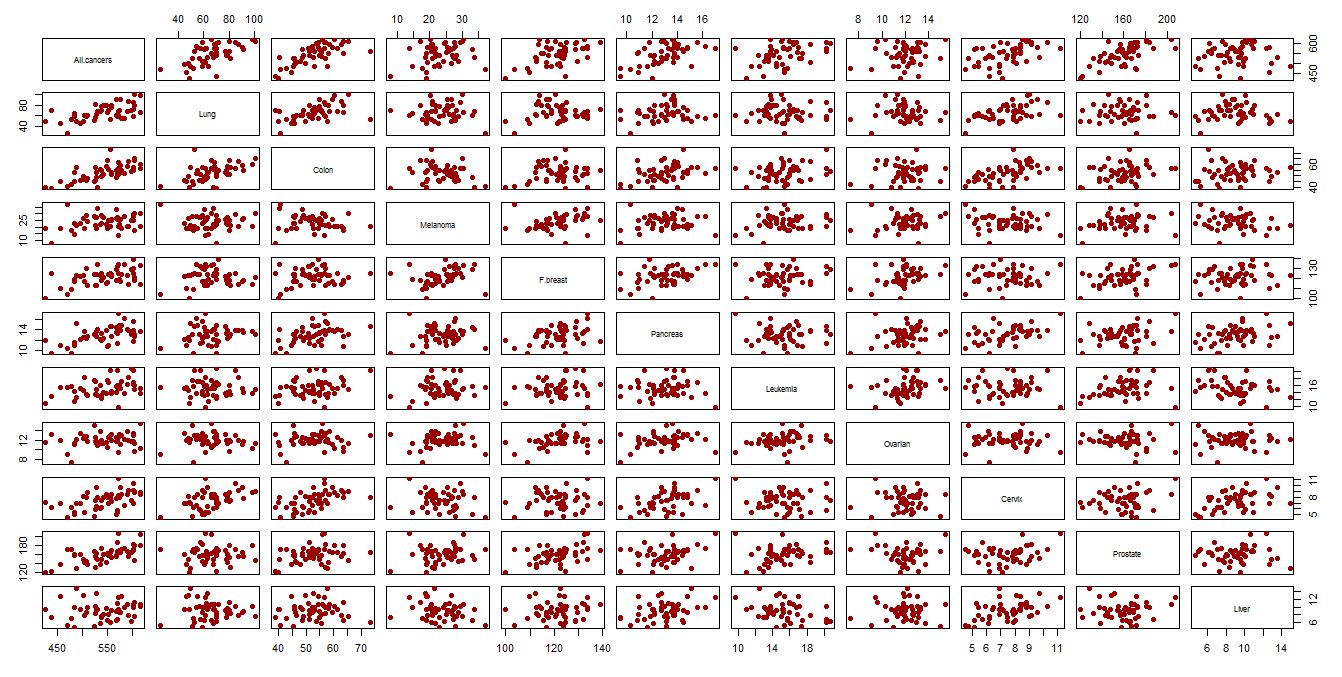
  
# 2c) How much of the total variation is captured by each of the first two PCs? How   
# much of the total variation is captured by the two PCs together?  
  
tab[3,1:2]

## [1] 0.5841099 0.8054828

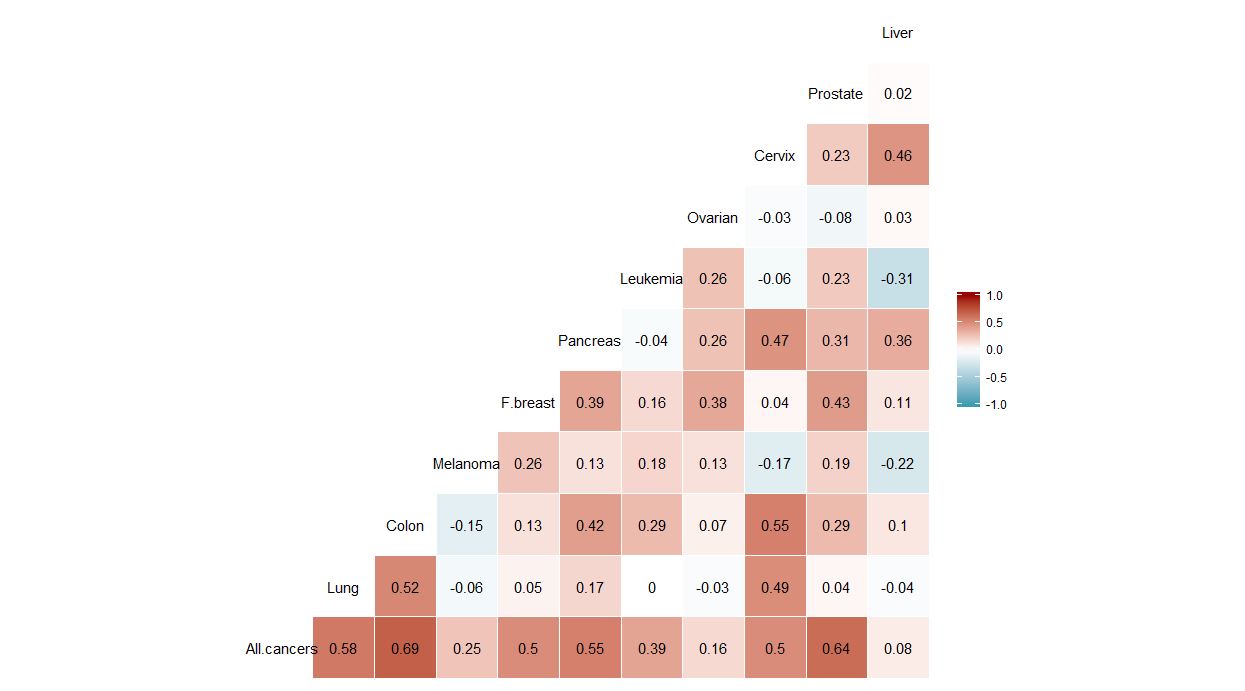
# 2d) Interpret the first two PCs.  
###The first PC Vector kind of represents overall body metrics  
###In the Second PC vector, the first half represents body lengths and the ###second half represents the body girth measurements

Question 3.R

#########Question 3  
# 3a) Examine the standard deviations of each cancer type, including the overall rates.  
data\_3=read.table("C:/Users/Nadim/Desktop/ST537/HW\_3/cancer\_dataset.txt",header = T,row.names = 1)  
pairs(data\_3, pch=19, col="#990000")  
library(GGally)



ggcorr(data\_3, low = "#3B9AB2", mid = "#FFFFFF", high = "#990000",label = T, label\_color = "black",label\_size = 3, label\_round = 3)



data\_3.sd<-apply(data\_3,2,sd)  
  
#####################################################################################  
# 3b) Do you think it is more appropriate to examine the covariance (i.e., use the   
# data matrix as it is) or the correlation (i.e., first standardize the data and then   
# compute covariance) in a principal components analysis of this data? Explain.  
###Covariance deals with similarity of variables of same kinds of fields. Correlation   
### provides a metric for variables from even different fields  
###Since the variables have diferent scales of std we standardise them  
std.data\_3<-scale(data\_3,center=T,scale=T)  
std.data\_3

## attr(,"scaled:center")  
## All.cancers Lung Colon Melanoma F.breast Pancreas   
## 544.120408 66.424490 52.777551 23.679592 121.510204 13.014286   
## Leukemia Ovarian Cervix Prostate Liver   
## 15.246939 11.959184 7.440816 159.953061 9.022449   
## attr(,"scaled:scale")  
## All.cancers Lung Colon Melanoma F.breast Pancreas   
## 47.536999 14.588088 7.210099 5.390848 8.008908 1.577709   
## Leukemia Ovarian Cervix Prostate Liver   
## 2.373649 1.395516 1.510094 18.048835 2.271588

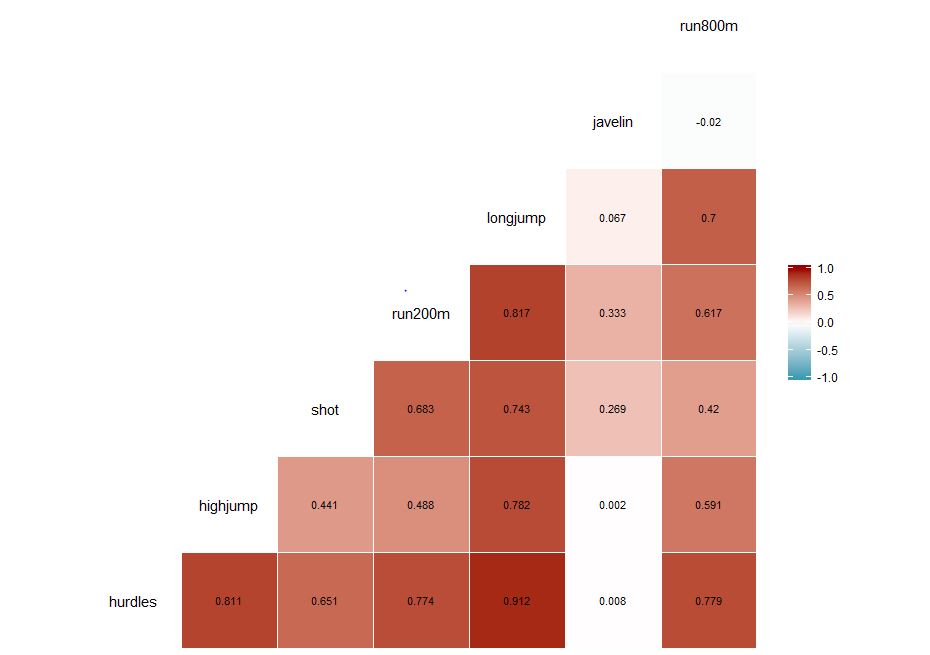
#####################################################################################  
# 3c) Perform a principal components analysis on this dataset and report the results.  
# Perform PCA  
dat\_3.pca<- prcomp(std.data\_3)  
# Extract the importance of each component  
summary(dat\_3.pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 1.8872 1.4110 1.2070 1.04549 0.92966 0.8028 0.69483  
## Proportion of Variance 0.3238 0.1810 0.1324 0.09937 0.07857 0.0586 0.04389  
## Cumulative Proportion 0.3238 0.5048 0.6372 0.73660 0.81517 0.8738 0.91765  
## PC8 PC9 PC10 PC11  
## Standard deviation 0.59711 0.52480 0.48278 0.20192  
## Proportion of Variance 0.03241 0.02504 0.02119 0.00371  
## Cumulative Proportion 0.95007 0.97511 0.99629 1.00000

Question 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)

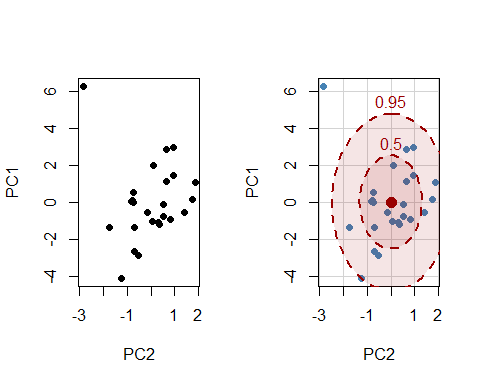
## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 2.1119 1.0928 0.72181 0.67614 0.49524 0.27010  
## Proportion of Variance 0.6372 0.1706 0.07443 0.06531 0.03504 0.01042  
## Cumulative Proportion 0.6372 0.8078 0.88223 0.94754 0.98258 0.99300  
## PC7  
## Standard deviation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

*# we can see that they are linearly related, this tells us that the first PCA contributes for the overall score as well.*

