Assignment:1

732A97 Multivariate Statistical Methods,Linkoping University

Group 18

Dimitra Muni - dimmu472
Karthikeyan Devarajan - karde799
Gowtham KM - gowku593
Biswas Kumar - bisku859

Question 1: Describing individual variables

a) *Describe the 7 variables with mean values, standard deviations e.t.c.*

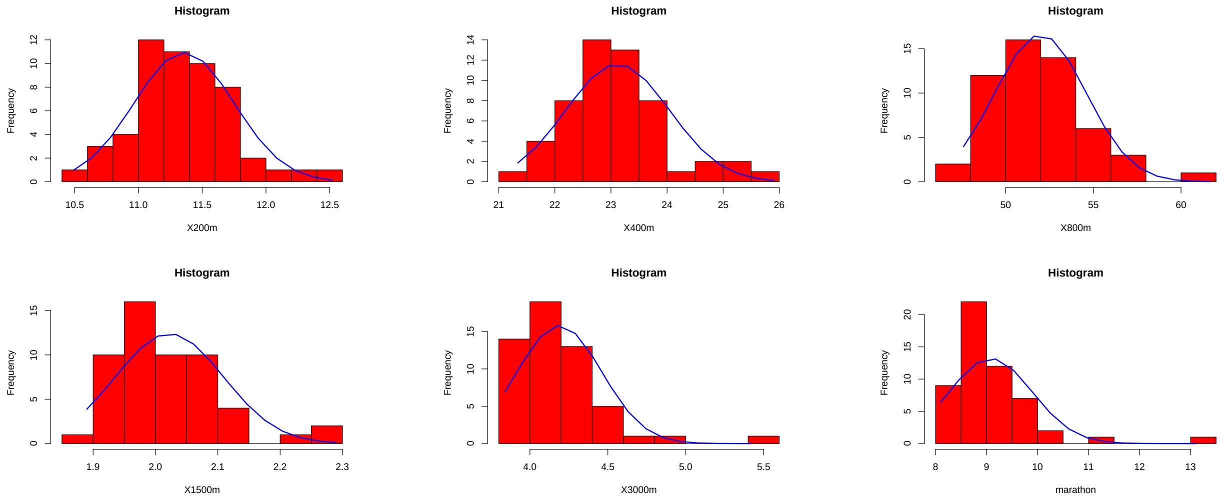
The mean value are

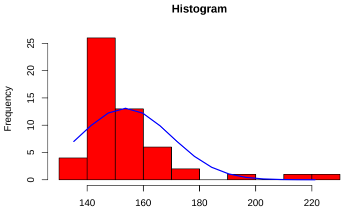
```
##       X100m      X200m      X400m      X800m     X1500m     X3000m
##   11.357778  23.118519  51.989074   2.022407   4.189444   9.080741
##    marathon
## 153.619259
```

The standard Deviations are

```
##        X100m       X200m       X400m       X800m      X1500m      X3000m
##   0.39410116  0.92902547  2.59720188  0.08687304  0.27236502  0.81532689
##     marathon
## 16.43989508
```

b) *Illustrate the variables with different graphs (explore what plotting possibilities R has). Make sure that the graphs look attractive (it is absolutely necessary to look at the labels, font sizes, point types). Are there any apparent extreme values? Do the variables seem normally distributed? Plot the best fitting (match the mean and standard deviation, i.e. method of moments) Gaussian density curve on the data's histogram. For the last part you may be interested in the hist() and density() functions*





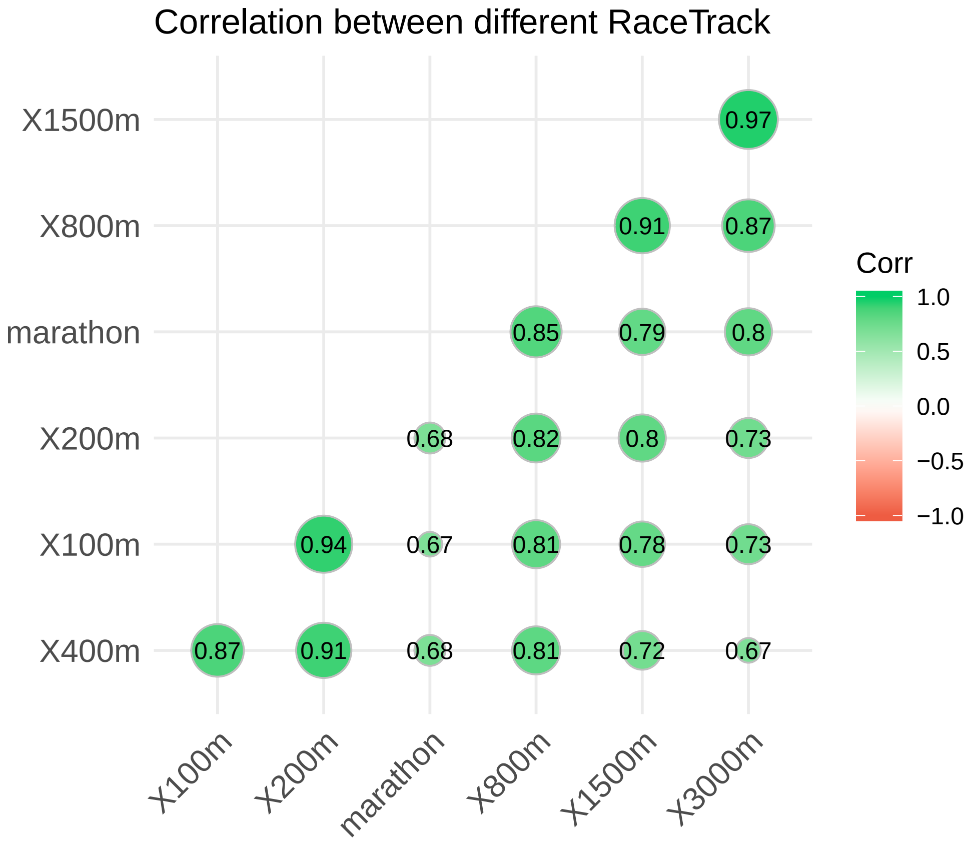
There are more extreme values in marathon histogram graph. All are normally distributed and with increase in race track distance the gaussian distribution fit curve are skewed towards negative. The deviation is more for marathon then 100m race. When a model is created, the accuracy will good for less skewed model. so, the accuracy will be better for 100m racetrack.

Question 2: Relantionship with the variables

a) *Compute the covariance and correlation matrices for the 7 variables. Is there any apparent structure in them? Save these matrices for future use*

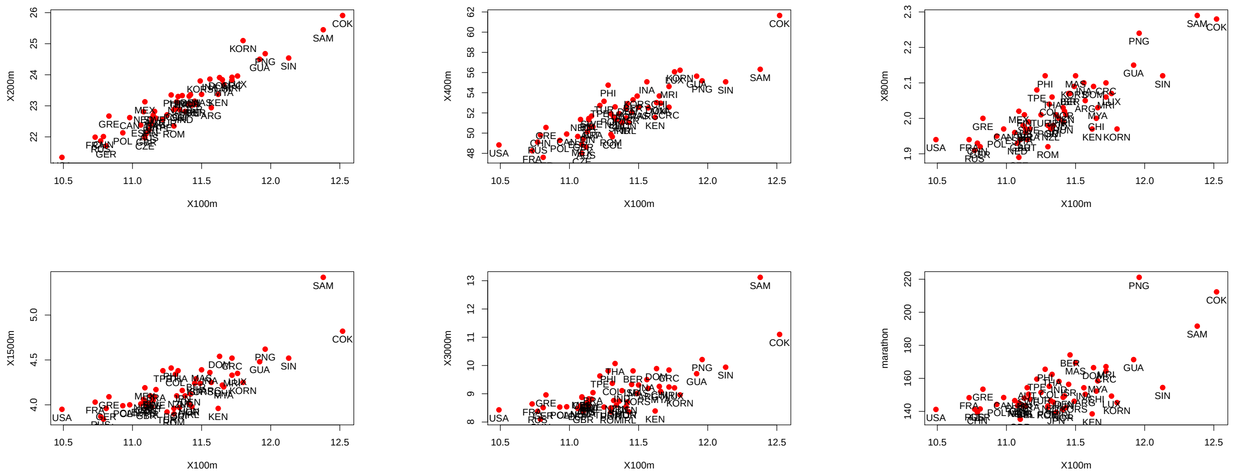
```
##                X100m      X200m      X400m       X800m     X1500m
## X100m    0.15531572  0.3445608  0.8912960 0.027703564 0.08389119
## X200m    0.34456080  0.8630883  2.1928363 0.066165898 0.20276331
## X400m    0.89129602  2.1928363  6.7454576 0.181807932 0.50917683
## X800m    0.02770356  0.0661659  0.1818079 0.007546925 0.02141457
## X1500m   0.08389119  0.2027633  0.5091768 0.021414570 0.07418270
## X3000m   0.23388281  0.5543502  1.4268158 0.061379315 0.21615514
## marathon 4.33417757 10.3849876 28.9037314 1.219654647 3.53983732
##               X3000m   marathon
## X100m     0.23388281   4.334178
## X200m     0.55435017  10.384988
## X400m     1.42681579  28.903731
## X800m     0.06137932   1.219655
## X1500m    0.21615514   3.539837
## X3000m    0.66475793  10.706091
## marathon 10.70609113 270.270150
```

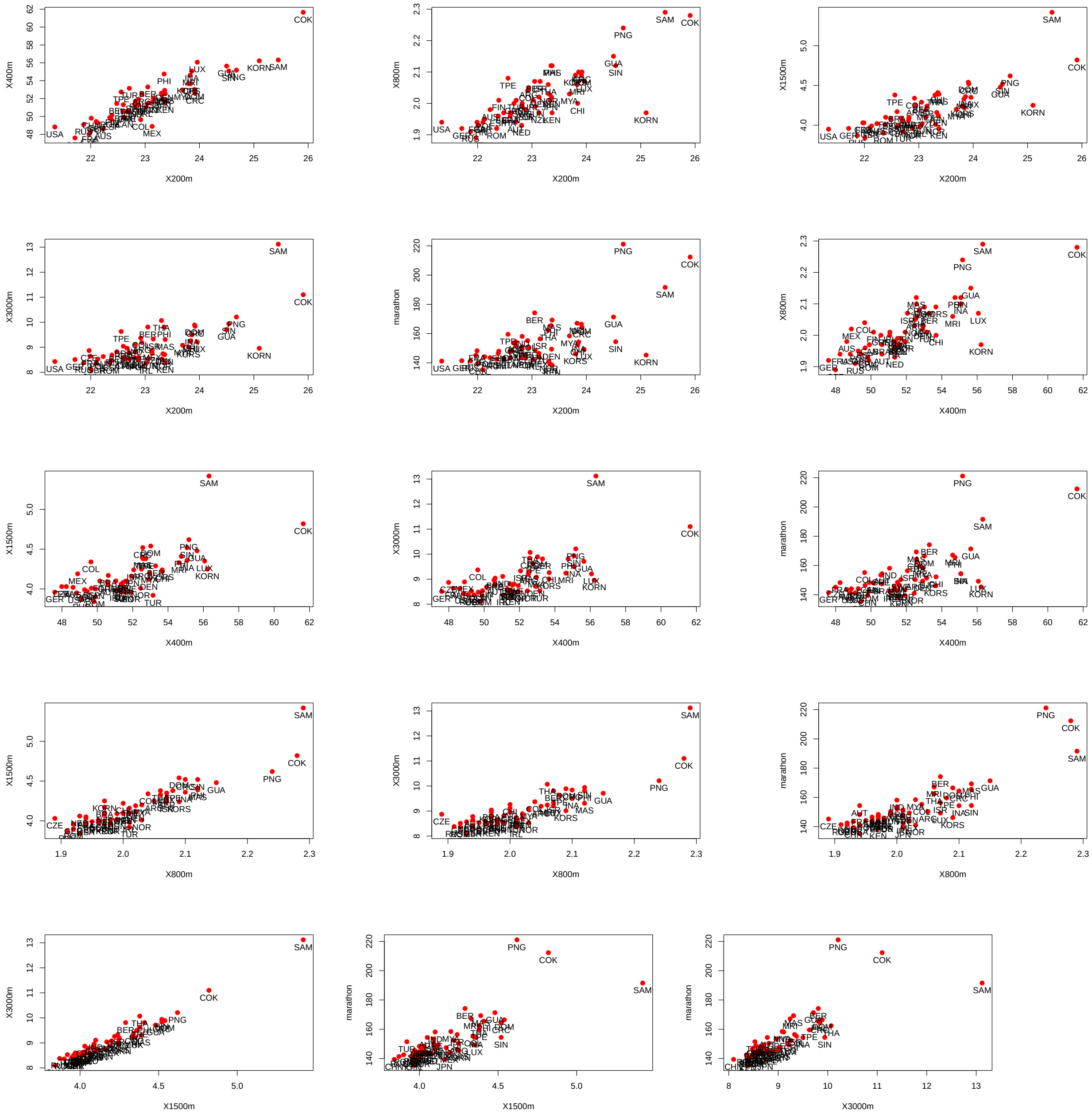
```
##              X100m     X200m     X400m     X800m    X1500m    X3000m
## X100m    1.0000000 0.9410886 0.8707802 0.8091758 0.7815510 0.7278784
## X200m    0.9410886 1.0000000 0.9088096 0.8198258 0.8013282 0.7318546
## X400m    0.8707802 0.9088096 1.0000000 0.8057904 0.7197996 0.6737991
## X800m    0.8091758 0.8198258 0.8057904 1.0000000 0.9050509 0.8665732
## X1500m   0.7815510 0.8013282 0.7197996 0.9050509 1.0000000 0.9733801
## X3000m   0.7278784 0.7318546 0.6737991 0.8665732 0.9733801 1.0000000
## marathon 0.6689597 0.6799537 0.6769384 0.8539900 0.7905565 0.7987302
##           marathon
## X100m    0.6689597
## X200m    0.6799537
## X400m    0.6769384
## X800m    0.8539900
## X1500m   0.7905565
## X3000m   0.7987302
## marathon 1.0000000
```



The correlation is decreasing for the racetrack with increase in distance of the racetrack. There is graph to represent the correlation matrix.

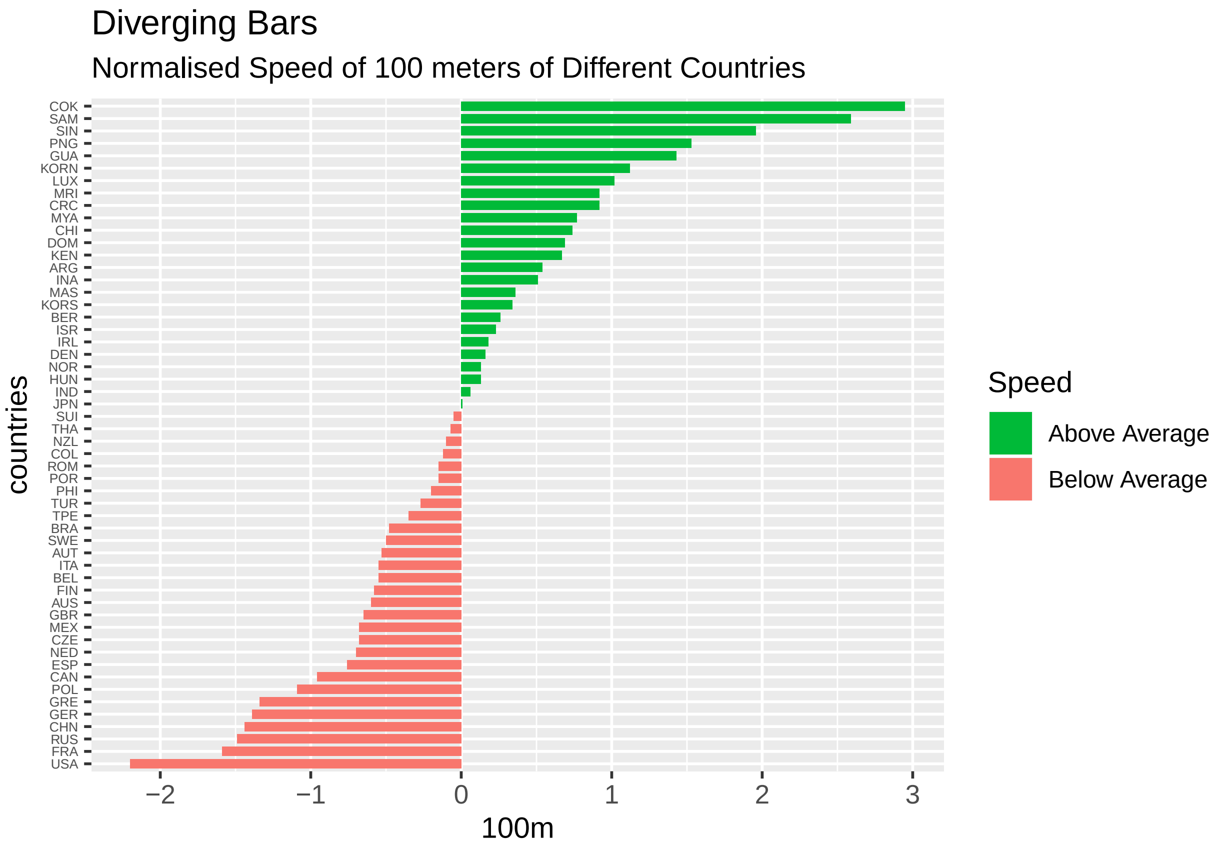
b) Generate and study the scatterplots between each pair of variables. Any extreme values?

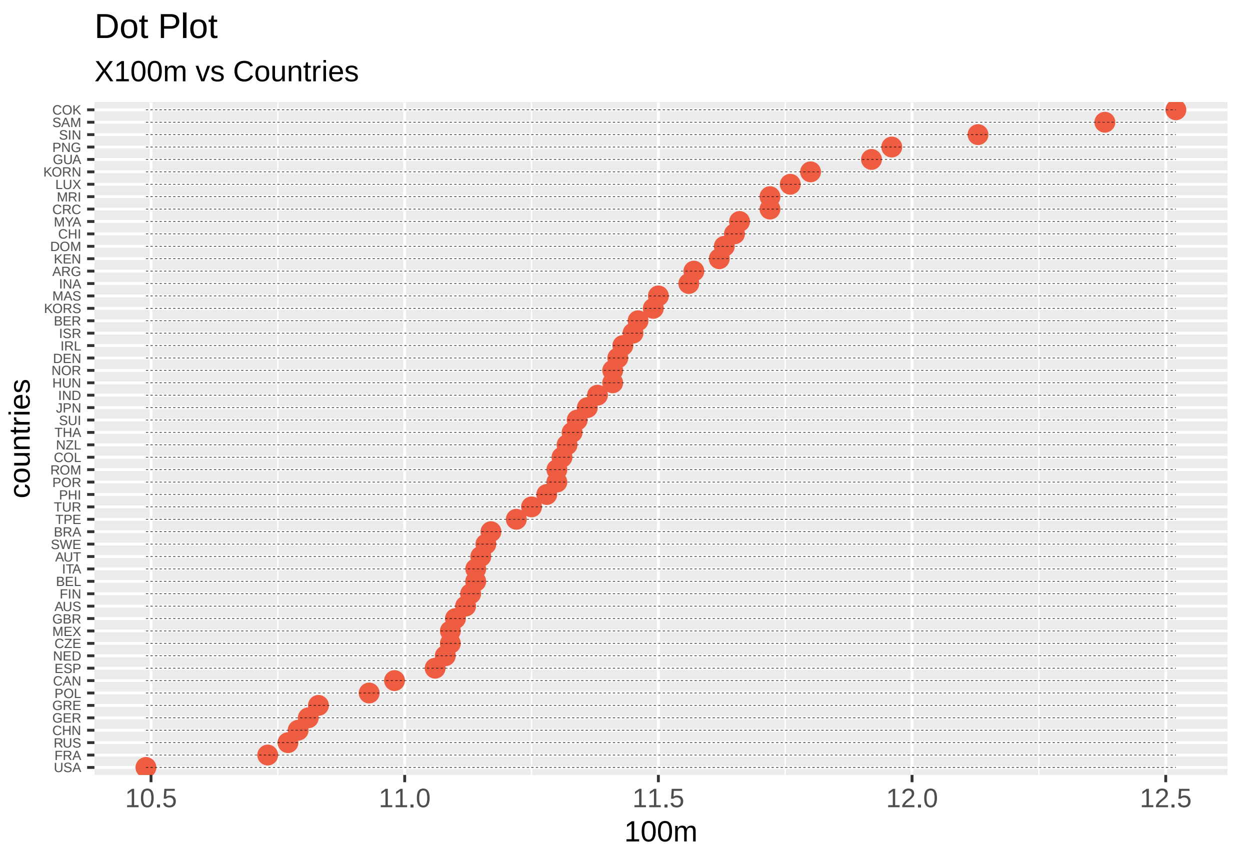




In most of the graphs between the variables, COK,SAM and PNG are the countries appears to extreme countries.

c) ***Explore what other plotting possibilities R offers for multivariate data. Present other (at least two) graphs that you find interesting with respect to this data set.***





Question 3: Examining for extreme values

a) ***Look at the plots (esp.scatterplots) generated in the previous question. Which 3-4 countries appear most extreme? Why do you consider them extreme?***

In previous plot of 100m,the countries peformed relatively better were Cook's Island(COK) and Samoa(SAM) while countries performing relatively poor were USA,France,Russia

Extreme countries lies between 2σ and 3σ away from the μ in each direction.

b) ***Finding five most extreme countries using Euclidean Distance***

```
## Countries which have the extreme values PNG COK SAM BER GBR
```

```
## Sweden's rank 48
```

c) Finding five most extreme countries using Normalized Euclidean Distance

```
## Countries which have the extreme values SAM COK PNG USA SIN
```

```
## Sweden's rank
```

d) Finding five most extreme countries using Mahalanobis distance

```
## Countries which have the extreme values SAM PNG KORN COK MEX
```

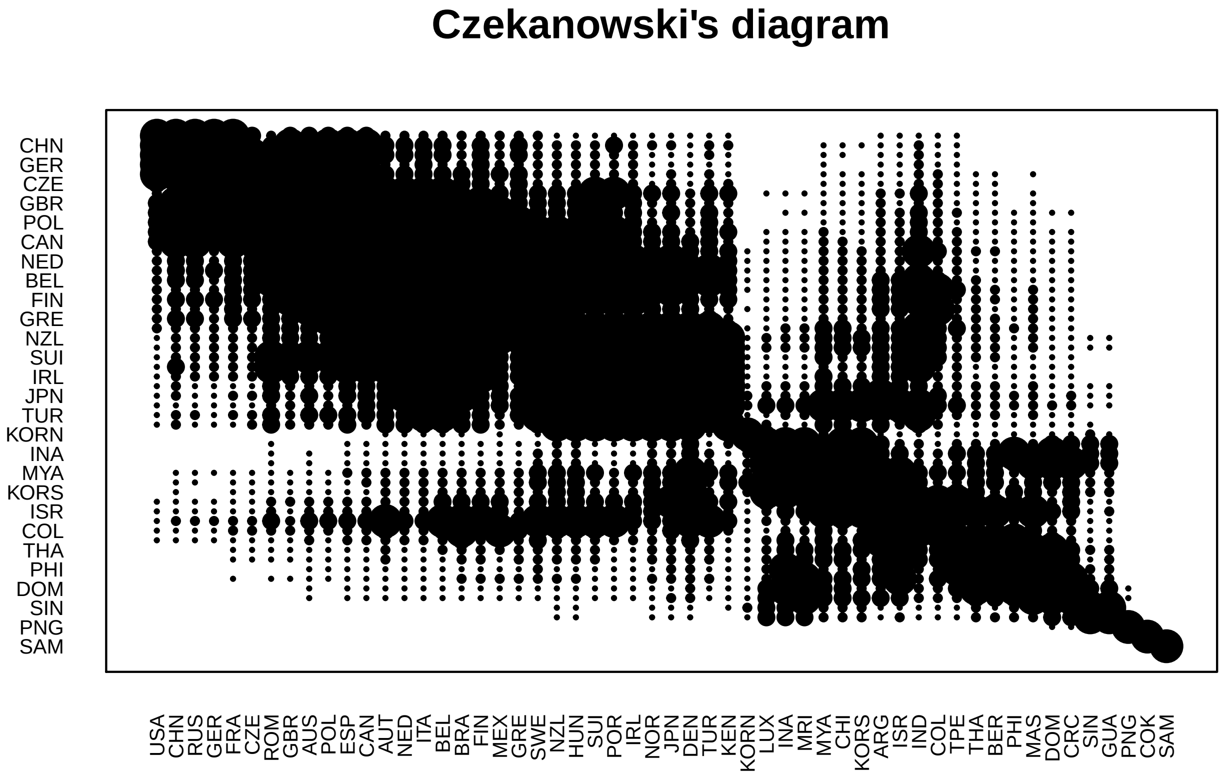
```
## Sweden's rank 54
```

e) ***Compare the results in b) and d),Some of the countries are in the upper end with all the measures and perhaps they can be classified as extreme.How does Sweden behave?.***

Countries performing extremly well or extremly poorly can be classified as extreme from the data.Samoa(SAM),Cook's Island(COK),Papua New Guinea(PNG) have relatively poor performance in the Olympics. Countries such as Russia and USA have performed well, still all these countries are extreme in terms of their distance from the mean.

Sweden is ranking 48,50 and 54 accordingly for **(b),(c) and (d)** which suggests that for Sweden the dispersion is less, the data is spread evenly from mean.Sweden has relatively consistent records in all different forms of competitions.

Produce Czekanowski's diagram using e.g. the RMaCzek package.



In case of problems please describe them

The input dataframe had to be conditioned in a such a way that for the data, rownames were column1.For the function **czek_matrix()** in input data.frame first column was excluded.

Code Appendix

```r
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(ggcorrplot)
library(RMaCzek)
sf <- read.csv('./T1-9.csv')
sf1 <- sf[,2:8]
mean = colMeans(sf[sapply(sf, is.numeric)])
mean
variance = var(sf[,-1])
deviation = apply(as.matrix(sf[,-1]),2, sd)
deviation
 x <- colnames(sf[,2:8])
 for(i in 2:8){
  h <-hist(sf[,i], breaks=10, col="red", xlab=x[i],
           main="Histogram")
  xfit <- seq(min(sf[,i]),max(sf[,i]),length=15)
  yfit <- dnorm(xfit,mean=mean(sf[,i]),sd=sd(sf[,i]))
  yfit <- yfit*diff(h$mids[1:2])*length(sf[,i])
  lines(xfit, yfit, col="blue", lwd=2)
}
sf_variance <- var(sf[,2:8])
sf_variance
sf_correlation <- cor(sf[,2:8])
sf_correlation
  ggcorrplot::ggcorrplot(sf_correlation, hc.order = TRUE,
           type = "lower",
           lab = TRUE,
           lab_size = 3,
           method="circle",
           colors = c("tomato2", "white", "springgreen3"),
           title="Correlation between different RaceTrack")

colna <- colnames(sf)
for(i in 2:8)
   for(j in 2:8)
       if(j!=i & j>i){
       plot(x = sf[,i], y = sf[,j],col = "red", pch = 19,cex = 1,lty = "solid",lwd = 2,xlab = colna[i],
       text(x =  sf[,i],y = sf[,j],labels = as.character(sf[,1]), pos = 1)
       }
sf$X100m_z <- round((sf$X100m - mean(sf$X100m))/sd(sf$X100m), 2)
sf$X100m_type <- ifelse(sf$X100m_z < 0, "below", "above")
sf <- sf[order(sf$X100m_z), ]
sf$countries <- factor(sf$countries, levels = sf$countries)
ggplot(sf, aes(x=countries, y=X100m_z, label=X100m_z)) +
  geom_bar(stat='identity', aes(fill=X100m_type),width = 0.75)  +
  scale_fill_manual(name="Speed",
                    labels = c("Above Average", "Below Average"),
                    values = c("above"="#00ba38", "below"="#f8766d")) +
  labs(subtitle="Normalised Speed of 100 meters of Different Countries",
       title= "Diverging Bars") + xlab("countries") + ylab("100m") +
  coord_flip() +  theme(axis.text.x = element_text(size=10),
                        axis.text.y = element_text(size=5))

ggplot(sf, aes(x=sf$countries, y=sf$X100m)) +
  geom_point(col="tomato2", size=3) +
  geom_segment(aes(x=sf$countries,
                   xend=sf$countries,
                   y=min(sf$X100m),
                   yend=max(sf$X100m)),
               linetype="dashed",
               size=0.1) +
  labs(title="Dot Plot",
       subtitle="X100m vs Countries") +  xlab("countries") + ylab("100m") +
  coord_flip() + theme(axis.text.x = element_text(size=10),
                    axis.text.y = element_text(size=5))
centered=scale(x = as.matrix(sf[,2:8]), center = TRUE, scale = FALSE)
squared=centered%*%t(centered)
rownames(squared)<-sf[,1]
colnames(squared)<-sf[,1]
diagonal<-diag(squared)
extrems=order(diagonal, decreasing=TRUE)[1:5]
cat("Countries which have the extreme values",as.character(sf[extrems,1]))
cat("Sweden's rank",which(names(diagonal[order(diagonal, decreasing=TRUE)])=="SWE"))
centered=scale(x = as.matrix(sf[,2:8]), center = TRUE, scale = FALSE)
variance<-var(sf[,2:8])
V<-variance*diag(7)
distance2<-centered%*%solve(V)%*%t(centered)
diagonal<-diag(distance2)
names(diagonal)<-sf[,2:8]
extrems=order(diagonal, decreasing=TRUE)[1:5]
cat("Countries which have the extreme values",as.character(sf[extrems,1]))
cat("Sweden's rank",which(names(diagonal[order(diagonal, decreasing=TRUE)])=="SWE"))
centered=scale(x = as.matrix(sf[,2:8]), center = TRUE, scale = FALSE)
covariance_mat=cov(sf[,2:8])
mahalnob<-centered%*%solve(covariance_mat)%*%t(centered)
diagonal<-diag(mahalnob)
names(diagonal)<-sf[,1]
extrems=order(diagonal, decreasing=TRUE)[1:5]
cat("Countries which have the extreme values",as.character(sf[extrems,1]))
cat("Sweden's rank",which(names(diagonal[order(diagonal, decreasing=TRUE)])=="SWE"))
suppressWarnings(suppressMessages(library(RMaCzek)))
#Czekanowski's diagram
rownames(sf)<-sf[,1]
x<-as.matrix(sf[,2:8])
mat<-czek_matrix(x)
plot(mat)
```