Assignment 1: Examining multivariate data

Laura Julià Melis

11/20/2019

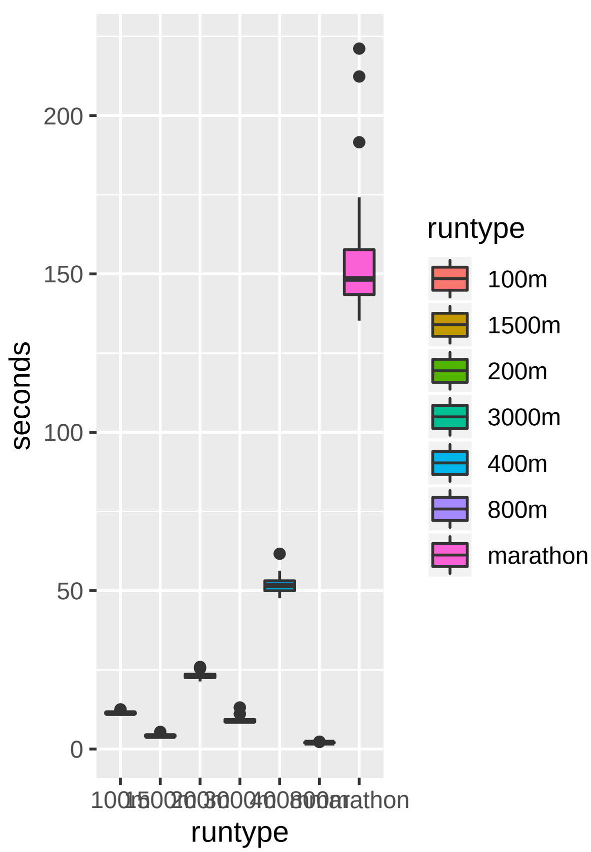
Question 1: Describing individual variables.

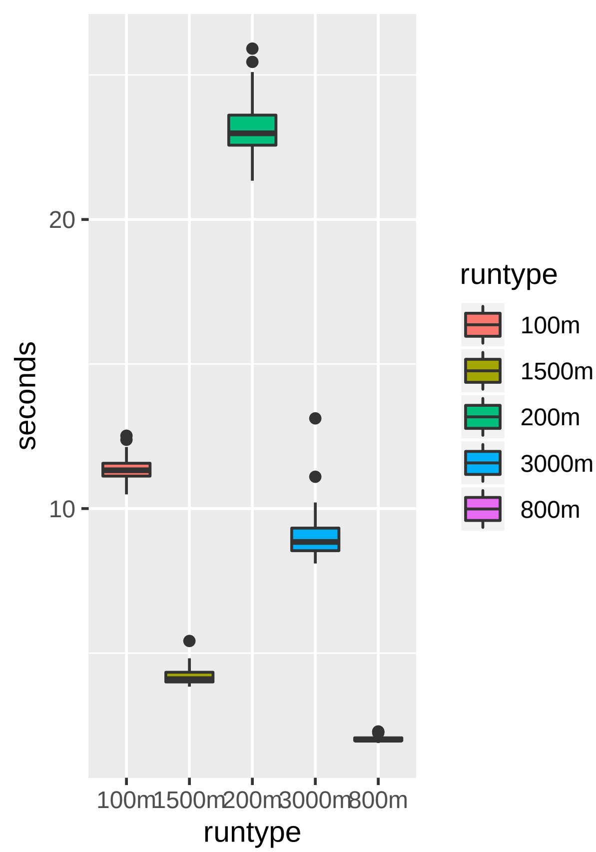
a) Describe the 7 variables with mean values, standard deviations e.t.c.

```
##              100m    200m    400m    800m   1500m   3000m marathon
## mean      11.3578 23.1185 51.9891 2.0224 4.1894  9.0807 153.6193
## median    11.3250 22.9800 51.6450 2.0050 4.1000  8.8450 148.4300
## mode      11.1400 22.6000 50.6200 1.9700 4.1000  8.5300 150.3200
## min       10.4900 21.3400 47.6000 1.8900 3.8400  8.1000 135.2500
## max       12.5200 25.9100 61.6500 2.2900 5.4200 13.1200 221.1400
## range      2.0300  4.5700 14.0500 0.4000 1.5800  5.0200  85.8900
## sd         0.3941  0.9290  2.5972 0.0869 0.2724  0.8153  16.4399
## skewness   0.5899  0.8087  0.9431 1.1820 1.9431  2.5373   2.3115
## kurtosis   0.7187  0.6608  1.7661 1.4178 5.7747  9.1981   6.0394
```

b) Illustrate the variables with different graphs. Are there any apparent extreme values? Do the variables seem normally distributed? Plot the best fitting Gaussian density curve on the data's histogram.

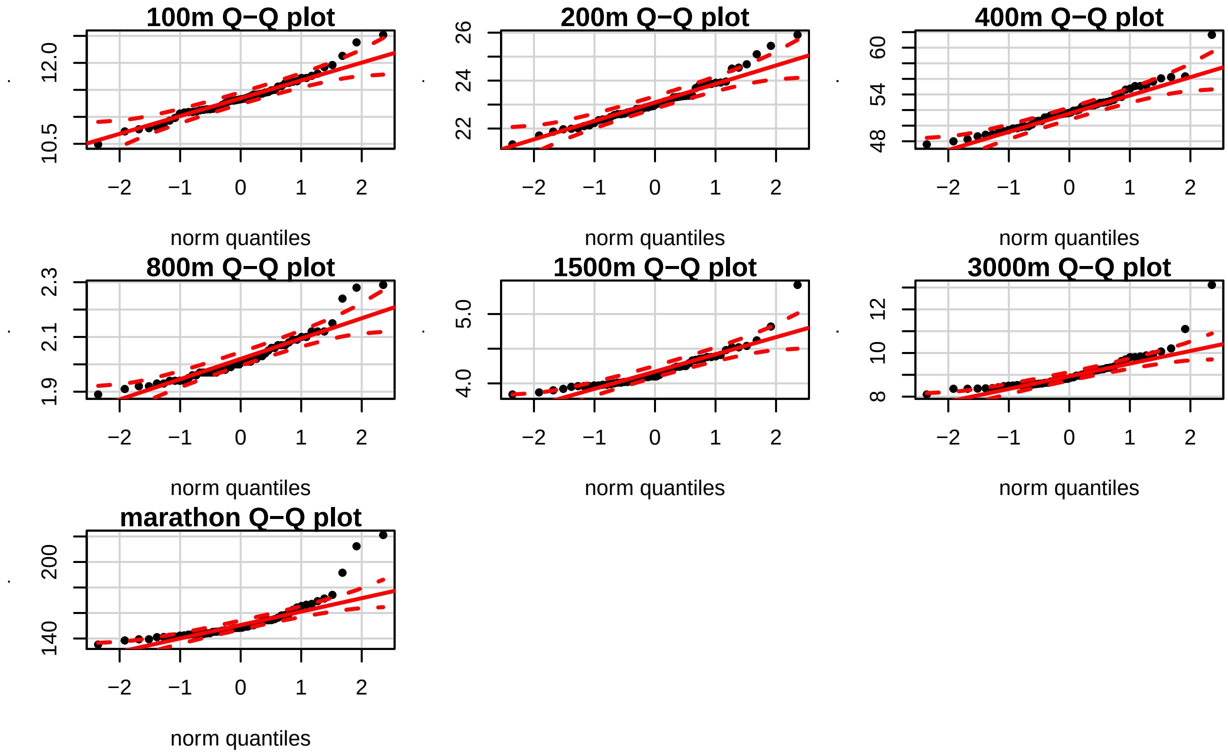
- Extreme values?





In terms of boxplots (i.e. fouth quatiles) marathon has 3 outliers and is the most spread distribution. Then, 400m has one outelier and is the second most spread.

- Normally distributed?

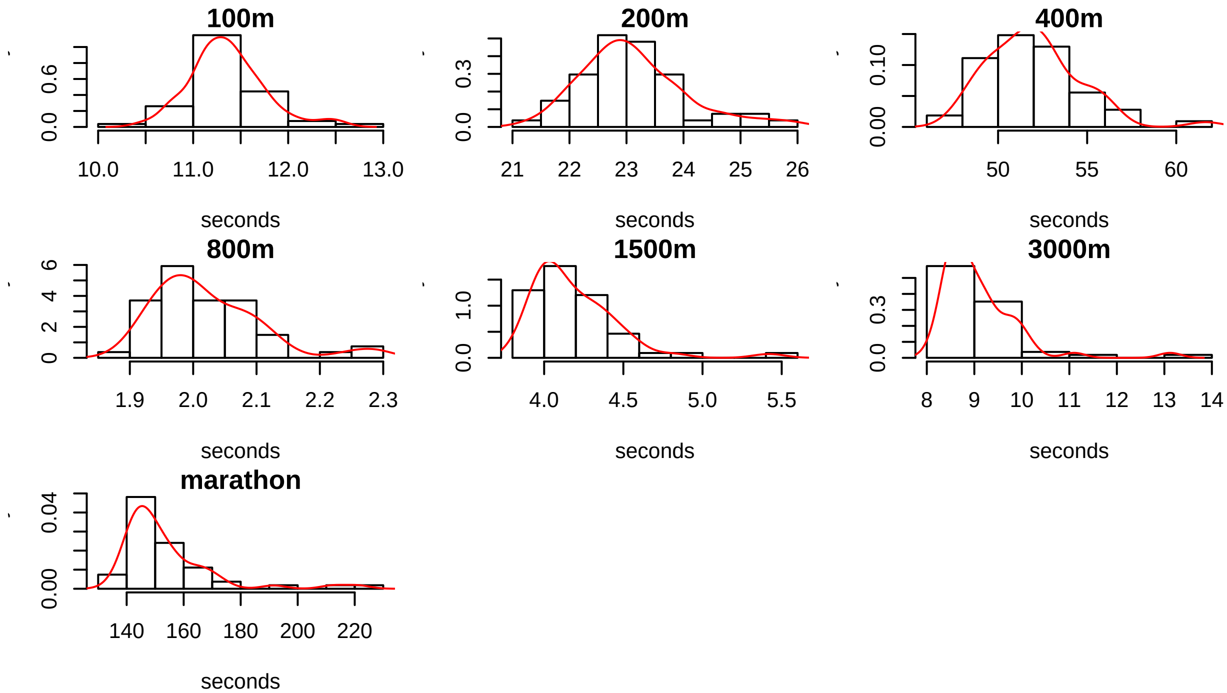


These qq-plots will help us assess if the data is plausibly coming from a Normal distribution. The closer all points lie to the line, the closer the distribution of the sample comes to the normal distribution. Furthermore, the red line should be almost y=x (i.e. 45 degrees angle from the horizontal line).

With 95% confidence level from Quantile-Comparison Plot, we can confirm the outliers as we said from box plots before (we could even specify which observations are outliers by putting the argument id=T in qqPlot). Moreover, exept from the ouliers all variables could be normal but a bit skewed. The variables that look more normal are 100m, 200m and 400m.

In all the variables, most of the points fall in the straight line except for the points in the tails (expecially in the right tail) meaning that the variables have more extreme values than expected if they came from a Normal distribution: the empirical (or sample) quantiles are less than the theoretical quantiles.

- Gaussian density curve on the data's histogram.



From histograms and densities we can confirm some conclusions from before. The variables '100m', '200m' and '400m' look normal. The variables are '1500m', '3000m' and 'marathon' look skewed and from the table of 1.a we can confirm that their skewed is high; 1.943142, 2.5372744, 2.311548 respectively.

Question 2: Relationships between the variables.

a) Compute the covariance and correlation matrices for the 7 variables.

- Covariance matrix:

```
##             100m    200m    400m   800m  1500m   3000m marathon
## 100m      0.1553  0.3446  0.8913 0.0277 0.0839  0.2339   4.3342
## 200m      0.3446  0.8631  2.1928 0.0662 0.2028  0.5544  10.3850
## 400m      0.8913  2.1928  6.7455 0.1818 0.5092  1.4268  28.9037
## 800m      0.0277  0.0662  0.1818 0.0075 0.0214  0.0614   1.2197
## 1500m     0.0839  0.2028  0.5092 0.0214 0.0742  0.2162   3.5398
## 3000m     0.2339  0.5544  1.4268 0.0614 0.2162  0.6648  10.7061
## marathon 4.3342 10.3850 28.9037 1.2197 3.5398 10.7061 270.2702
```

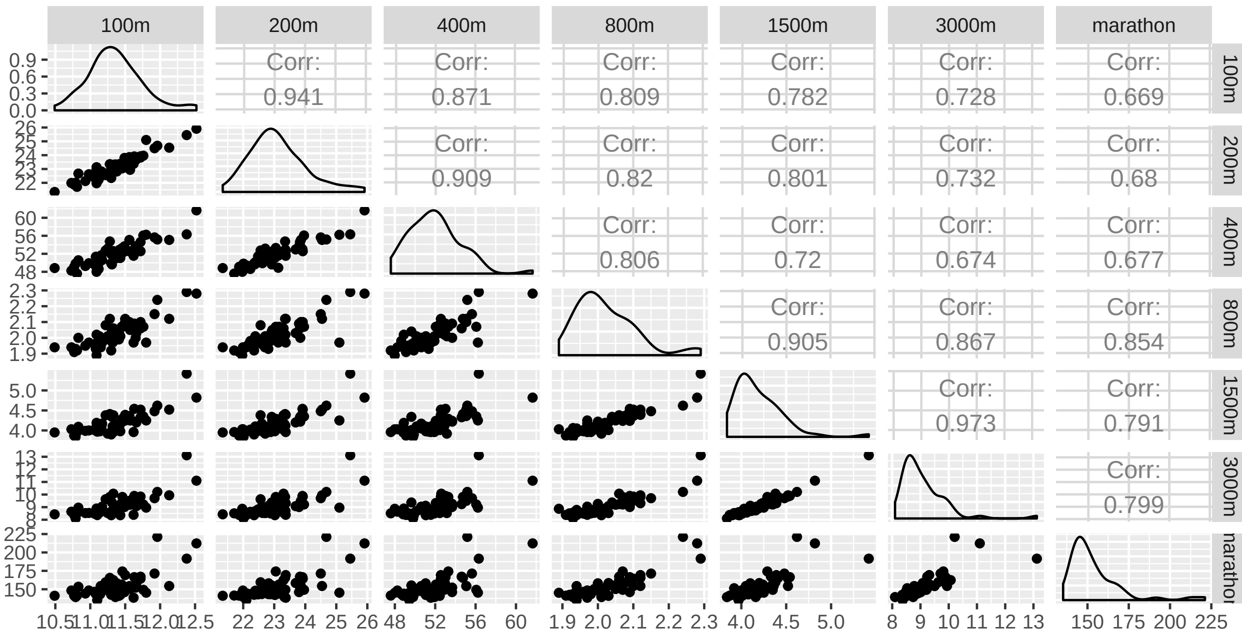
All the covariances between variables are positive, meaning that all pair of variables are positively related. Also, we can observe that the covariances are smaller between those variables representing short races and larger between variables that represent linger races (such as "Marathon"). The highest covariance coefficient is $\mathrm{Cov}(400\mathrm{m}, \mathrm{Marathon}) = 28.9$ and the lowest, $\mathrm{Cov}(800\mathrm{m}, 1500\mathrm{m}) = 0.0214$.

- Correlation matrix:

```
##             100m   200m   400m   800m  1500m  3000m marathon
## 100m      1.0000 0.9411 0.8708 0.8092 0.7816 0.7279   0.6690
## 200m      0.9411 1.0000 0.9088 0.8198 0.8013 0.7319   0.6800
## 400m      0.8708 0.9088 1.0000 0.8058 0.7198 0.6738   0.6769
## 800m      0.8092 0.8198 0.8058 1.0000 0.9051 0.8666   0.8540
## 1500m     0.7816 0.8013 0.7198 0.9051 1.0000 0.9734   0.7906
## 3000m     0.7279 0.7319 0.6738 0.8666 0.9734 1.0000   0.7987
## marathon 0.6690 0.6800 0.6769 0.8540 0.7906 0.7987   1.0000
```

The correlation coefficient is a measure that calculates the strength and direction of the linear relationship between two variables ($r \in \{-1, +1\}$). We observe that all the coefficients are positive (and greater than 0.65) so all pairs of variables have a positive linear relationship meaning that, as the value of one variable increases, the value of the other variable increases too. The highest coefficient is $\mathrm{Cor}(1500\mathrm{m}, 3000\mathrm{m}) = 0.9734$ and the lowest, $\mathrm{Cor}(100\mathrm{m}, \mathrm{Marathon}) = 0.6690$.

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



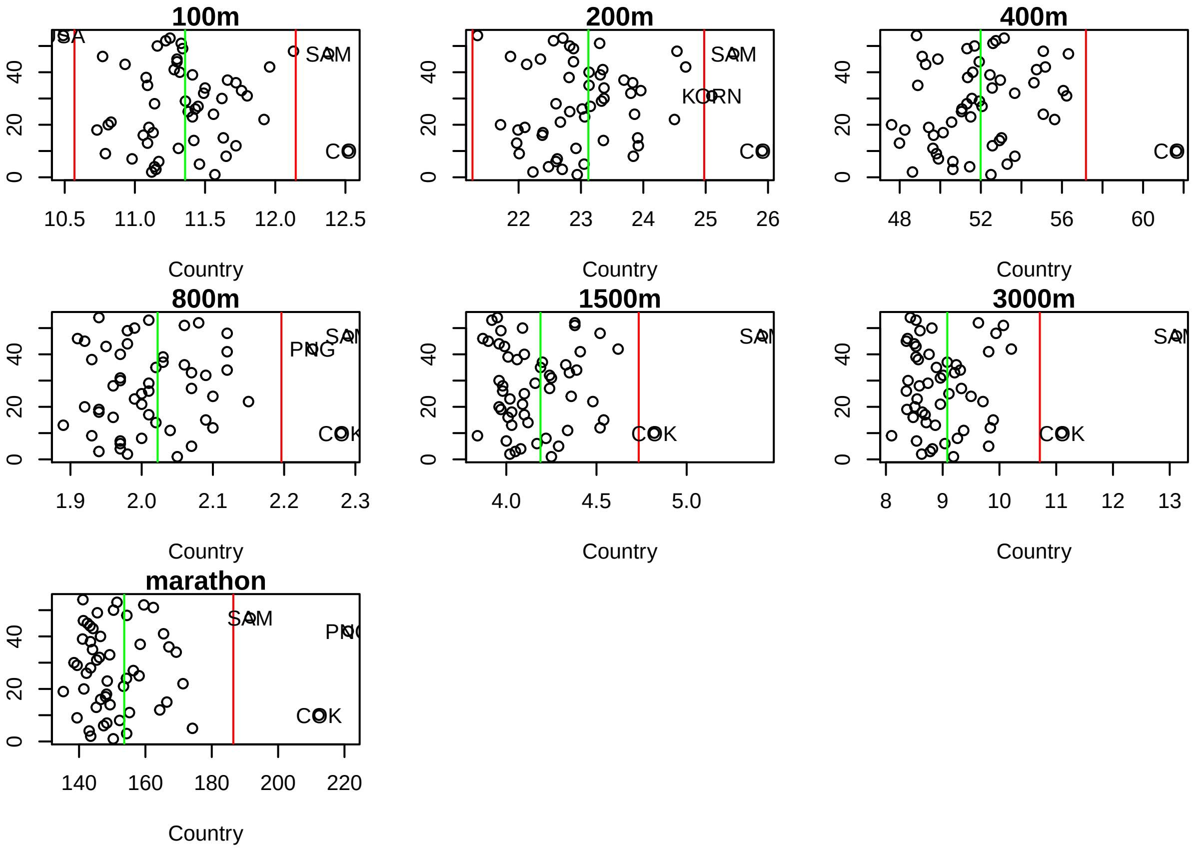
From pair scatterplot, we could see outliers at almost every pair (e.g. '200m' vs '800m','400m' vs '800m', '400m' vs '1500m'), but is is highly seen on every valiable vs 'marathon'.

c) Present other (at least two) graphs that you find interesting with respect to this data set.

- Scatterplots with labeled extreme points:

In the following plots, green lines represent the mean and red lines represent the lower and upper limit. These limits have been calculated as follows:

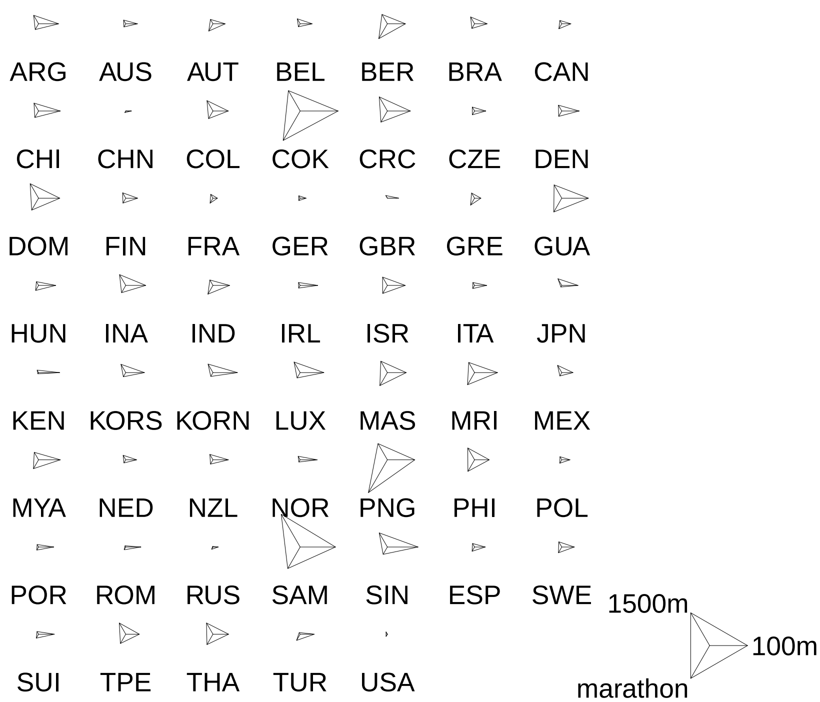
$$\bar{x} \pm 2\sigma \text{ because } \Pr\{\mu - 2\sigma \leq X \leq \mu + 2\sigma\} = 0.9545$$

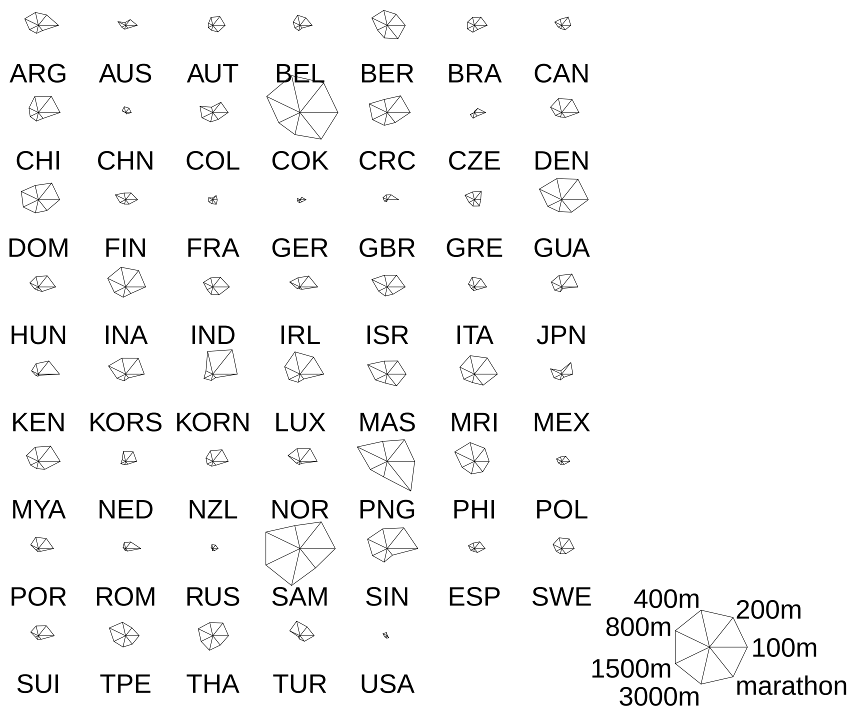


So, looking at these plots, we can observe that the most extreme countryes are COK, SAM and PNG.

- 3D Scatterplots:

We will take the 3 most uncorrelated variables, in order to have the most information, to plot in 3D scatterplot.





- Radar charts:

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?One approach to measuring "extremism" is to look at the distance (needs to be defined!) between an observation and the sample mean vector, i.e. we look how far one is from the average. Such a distance can be called an multivariate residual for the given observation.

As we can see from star plots above, the most extreme countries in terms of good performance are United States (USA), Russia (RU), China (CHN) and Germany (GER). In terms of bad performance the most extreme countries are Samoa (SAM), Cook Islands (COK), Papua New Guinea (PNG) and Guatemala (GUA).

We consider them extreme because the values of the times for the runtypes are either too small (good performance) or too big (bad performance).

b) The most common residual is the Euclidean distance between the observation and sample mean vector, i.e.

$$d(\vec{x}, \bar{x}) = \sqrt{(\vec{x} - \bar{x})^T \cdot (\vec{x} - \bar{x})}$$

This distance can be immediately generalized to the $L^r, r > 0$ distance as

$$d_{L^r}(\vec{x}, \bar{x}) = \left(\sum_{i=1}^{p} |\vec{x}_i - \bar{x}_i|^r \right)^{1/r}$$

where p is the dimension of the observation (here $p = 7$).

Compute the squared Euclidean distance (i.e. $r = 2$) of the observation from the sample mean for all 55 countries using R's matrix operations. First center the raw data by the means to get $\vec{x} - \bar{x}$ for each country. Then do a calculation with matrices that will result in a matrix that has on its diagonal the requested squared distance for each country. Copy this diagonal

to a vector and report on the five most extreme countries. In this questions you MAY NOT use any loops.

```
##      PNG      COK      SAM      BER      GBR
## 67.62796 59.61517 38.52476 20.61606 18.59146
```

The five most extreme countries are Papua New Guinea (PNG), Cook Islands (COK), Samoa (SAM), Bermuda (BER) and United Kingdom (GBR).

c) The different variables have different scales so it is possible that the distances can be dominated by some few variables. To avoid this we can use the squared distance

$$d^2_{\boldsymbol{V}}(\vec{x}, \bar{x}) = (\vec{x} - \bar{x})^T \boldsymbol{V}^{-1} (\vec{x} - \bar{x}),$$

where $\boldsymbol{V}$ is a diagonal matrix with variances of the appropriate variables on the diagonal. The effect, is that for each variable the squared distance is divided by its variance and we have a scaled independent distance.

It is simple to compute this measure by standardizing the raw data with both means (centring) and standard deviations (scaling), and then compute the Euclidean distance for the normalized data. Carry out these computations and conclude which countries are the most extreme ones. How do your conclusions compare with the unnormalized ones?

```
##      SAM      COK      PNG      USA      SIN
## 75.63643 64.65520 34.26764 12.88274 11.45269
```

Uding the Euclidean distance for thenormalized data, the five most extreme countries are Samoa (SAM), Cook Islands (COK), Papua New Guinea (PNG), United States (USA) and Singapore (SIN).

d) The most common statistical distance is the Mahalanobis distance

$$d^2_M(\vec{x}, \bar{x}) = (\vec{x} - \bar{x})^T \boldsymbol{C}^{-1} (\vec{x} - \bar{x}),$$

where $\boldsymbol{C}$ is the sample covariance matrix calculated from the data. With this measure we also use the relationships (covariances) between the variables (and not only the marginal variances as $d_{\boldsymbol{V}}(\mathring{u}, \mathring{u})$ does). Compute the Mahalanobis distance, which countries are most extreme now?

```
##      SAM      PNG     KORN      COK      MEX
## 35.02271 30.51926 26.75967 19.85290 14.35382
```

Using the Mahalanobis distance, the five most extreme countries are Samoa (SAM), Papua New Guinea (PNG), North Korea (KORN), Cook Islands (COK) and Mexico (MEX).

e) Compare the results in b)–d). Some of the countries are in the upper end with all the measures and perhaps they can be classified as extreme. Discuss this. But also notice the different measures give rather different results (how does Sweden behave?). Summarize this graphically. Produce Czekanowski's diagram using e.g. the `RMaCzek` package. In case of problems please describe them.

1. COMPARISON:

```
Results in b): PNG > COK > SAM > BER > GBR
Results in d): SAM > PNG > KORN > COK > MEX
```

In b, where the Eucliden distance (using the centered data by its mean) has been used, we haven't considered the variances and as a result, those variables with a diferent scale has dominated the distances: PNG, COK and SAM are three countries with extreme values in those variables where the times are scaled in minutes.

On the other hand, when we have calculated the Mahalanobis distance, we have divided each variable by its variance and also have considered the covariance between each pair of variables. As a result, we can see that for example the country KORN (that has not appeared in the plots analyzed in the questions above) is the third most extreme country.

Also, we can see that in b) SAM has obtained a distance of 38.52476 (third position), with a difference of 29 points with the first one while in d) has been the most extreme country. This might be explained because PNG and COK are quite more extreme in the marathon variable than SAM and this variable is dominating a lot the distance between the observations and the mean.

2. DISCUSSION:

In all the measures, SAM, PNG and COK have been the ones with biggest distances. This makes sense because in all the plots we have seen that these 3 countries have been always the ones with highest times for all the variables.

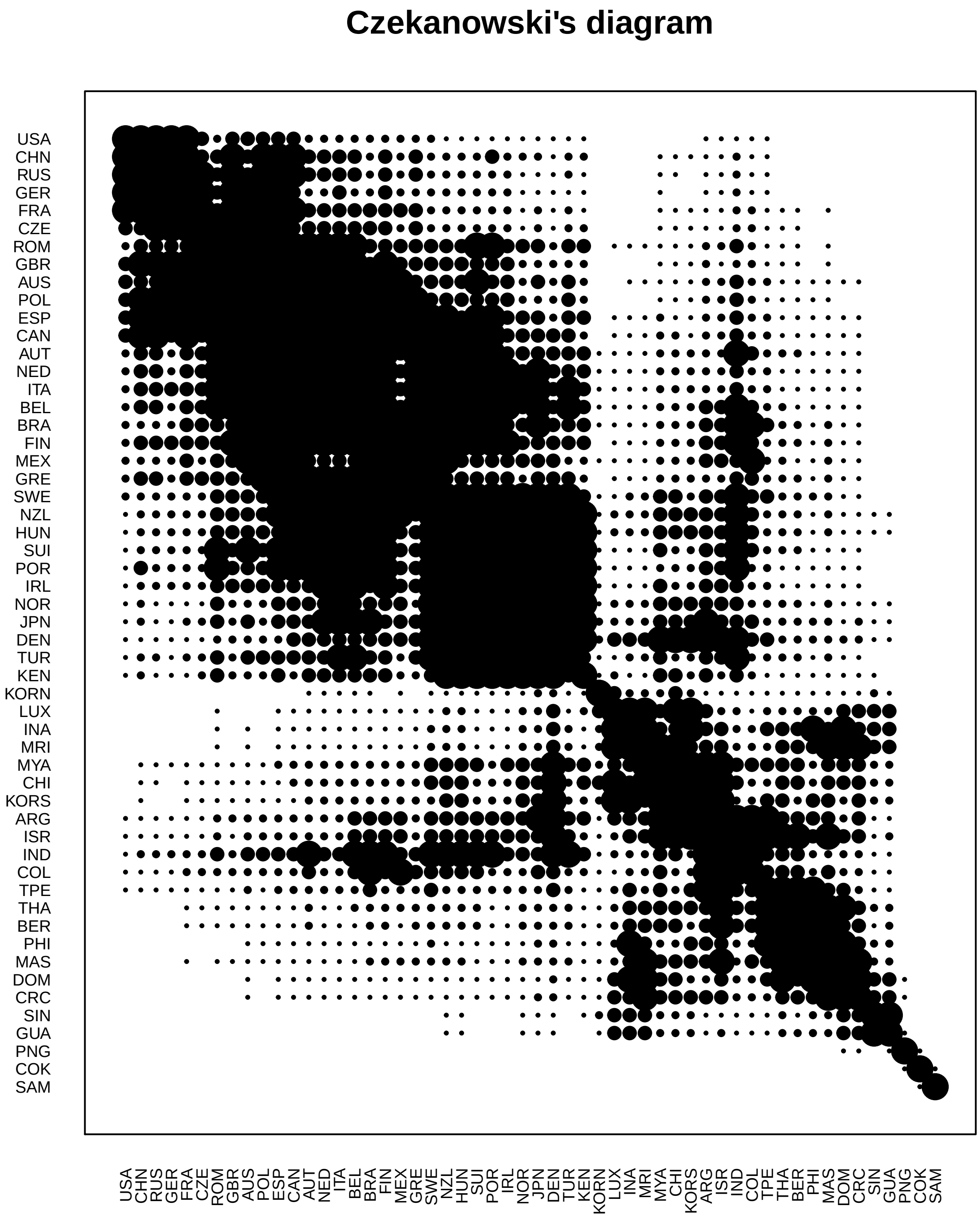
3. SWEEDEN:

In the next plot, each point represents a country (blue points represent Sweden), x-axis shows he three different distance measures that we have used and y-axis indicates the distance value.

(plot)

We observe that sweeden is in the three cases quite close to the mean (distance is close to 0)…

4. CZEKANOWSKI'S DIAGRAM: Czekanowski's function calculates the distance (by default, the Euclidean distance) between all possible pairs of objects (in our case, countries).



From the diagram we can conclude that USA, CHN and RUS (top left corner) are the three countries with the greatest lower distances (countries that performe better, they are far from the mean values in all variables) while again, in the bottom right corner, we see that PNG, COK and SAM are the three most extreme values, with the greatest upper distances (worst performances).

Appendix.

Question 1

```r
# NEEDED LIBRARIES
library(ggplot2)
library(tidyr)
library(gridExtra)
library(car)
library(fmsb)
library(GGally)
library(RMaCzek)

# Importing and modifying data file
df = read.table("T1-9.dat")
colnames(df) = c("country","100m", "200m", "400m", "800m", "1500m", "3000m" , "marathon")
df1 = df[,-1]  # without the column of the countries
```

Subquestion a.

```r
# Mode: the value that appears most often
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Table with some common statistics
statsdf = data.frame(cbind(mean=sapply(df1, mean),
                        median=sapply(df1, median),
                        mode=sapply(df1, getmode),
                        min=sapply(df1, min),
                        max=sapply(df1, max),
                        range=sapply(df1, function(x)max(x)-min(x)),
                        sd=sapply(df1, sd),
                        skewness=sapply(df1, timeDate::skewness),
                        kurtosis=sapply(df1, timeDate::kurtosis)
                        ))

as.data.frame(t(round(statsdf, 4)))
```

Subquestion b.

```r
# b1. EXTREME VALUES?
# tidyr::gather
# a table with columns the type of the run and the seconds
df2 = gather(df1, "runtype", "seconds")
df22 = gather(df1[,-c(3,7)], "runtype", "seconds")  #without: 400m, marathon
df222 = gather(df1[,-c(2,3,6 ,7)], "runtype", "seconds") #without: 200m, 400m, marathon

# https://www.r-graph-gallery.com/89-box-and-scatter-plot-with-ggplot2.html
#library(gridExtra)
par(mfrow=c(1,3))
ggplot(df2, aes(x=runtype, y=seconds, fill=runtype)) + geom_boxplot()
ggplot(df22, aes(x=runtype, y=seconds, fill=runtype)) + geom_boxplot()
ggplot(df222, aes(x=runtype, y=seconds, fill=runtype)) + geom_boxplot()
```

```r
# b2. NORMALLY DUSRIBUTED?
#library("car")
par(mfrow=c(3,3), oma = c(0,0,0.5,0) + 0.1, mar = c(4,3,1,1) + 0.1)
for(i in 1:7){
  qqPlot(df1[,i],ylab="sample",envelope=.95,col.lines="red",pch=19,id=F) + title(main=paste(names(df1)[
}

#pch = 19: solid circle
#envelope=.95: 95% confidence level
#id=TRUE: show the id of the outliers

# b3. GAUSSIAN DENSITY CURVE.
par(mfrow=c(3,3), oma = c(0,0,0.5,0) + 0.1, mar = c(4,3,1,1) + 0.1)
for(i in 1:7){
  dens = density(df1[,i])
  hist(df1[,i], freq=F, xlab="seconds",main=names(df1)[i]) #probability densities
  lines(dens,col="red")
}
```

Question 2

Subquestion a.

```r
(S <- round(cov(df1),4)) # Covariance matrix
(R <- round(cor(df1),4)) # Correlation matrix
```

Subquestion b.

```r
#library(GGally)
ggpairs(df1,progress=F)
```

Subquestion c.

```r
# C1. Scatterplots with labelled extreme points
par(mfrow=c(3,3), oma = c(0,0,0.5,0) + 0.1, mar = c(4,3,1,1) + 0.1)
for(i in 2:8){
 low_dist <- mean(df[,i])-2*sd(df[,i])  # 95.45% of the values are inside this interval:
 upp_dist <- mean(df[,i])+2*sd(df[,i])
 extreme <- which(df[,i] > upp_dist | df[,i] < low_dist)

plot(df[,i], df[,1], xlab="Country", ylab=names(df[i]), main=names(df)[i])+
  abline(v=c(low_dist,upp_dist, mean(df[,i])), col=c("red","red","green")) +
  text(df[extreme,i], df[extreme,1],as.vector(df[extreme,1]))
}

# C2. 3d scatterplots
df3 = df[,c(2,6,8)]
rownames(df3) = df$country
stars(df3, full = T, key.loc = c(20, 2))  #"100m", "1500m" and "marathon"

d = df[,-1]
rownames(d) = df$country
stars(d, full = T, key.loc = c(20, 2))

# C3. Radar chart
# #library(fmsb)
```

```
# #radar chart
# min_row <- sapply(df, min)
# max_row <- sapply(df, max)
# data_radar <- rbind(min_row, max_row, df)
# data_radar[1,1] <- "min"
# data_radar[2,1] <- "max"
# data_radar[,2:8] <- lapply(data_radar[,2:8], function(x) as.numeric(as.character(x)))
#
# # countries c(1,2, ..)
# radarchart(data_radar[1:5,2:8])
```

Question 3

Subquestion b.

```
resid = sapply(df[,-1],function(x)x-mean(x))  #center the raw data by the means
euclidist = sqrt(abs(resid%*%t(resid)))  #Euclidean distance between the observation and sample mean ve

dists = diag(euclidist)   #diag is squared distance for each country
names(dists) = df$country
head(sort(dists, decreasing=TRUE),5)  #sort with decreasing order and take the first five
```

Subquestion c.

```
V = diag(diag(S)) #first diag to take the diagonal
                  #then diag again to make the vector a diagonal matrix
sqdist = resid%*%solve(V)%*%t(resid)

sqdist = diag(sqdist)
names(sqdist) = df$country

head(sort(sqdist, decreasing=TRUE),5)  #sort with decreasing order and take the first five
```

Subquestion d.

```
mahdist = resid%*%solve(S)%*%t(resid)

mahdist = diag(mahdist)
names(mahdist) = df$country

head(sort(mahdist, decreasing=TRUE),5)  #sort with decreasing order and take the first five
```

Subquestion e.

```
#library(RMaCzek)
df3d = df1
rownames(df3d) = df$country
m = czek_matrix(df3d)
plot.czek_matrix(m)
```