assignment_1

Agustín Valencia

11/5/2019

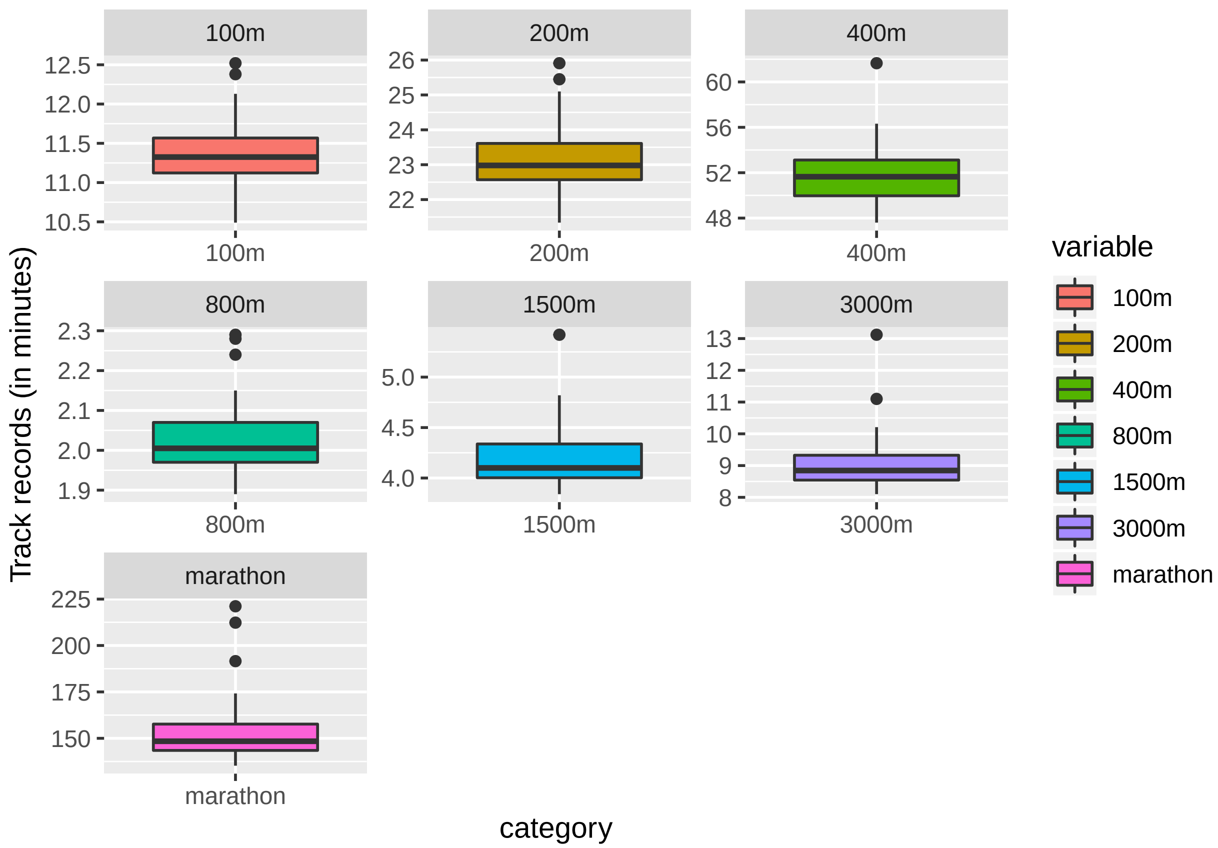
Question 1: Describing individual variables

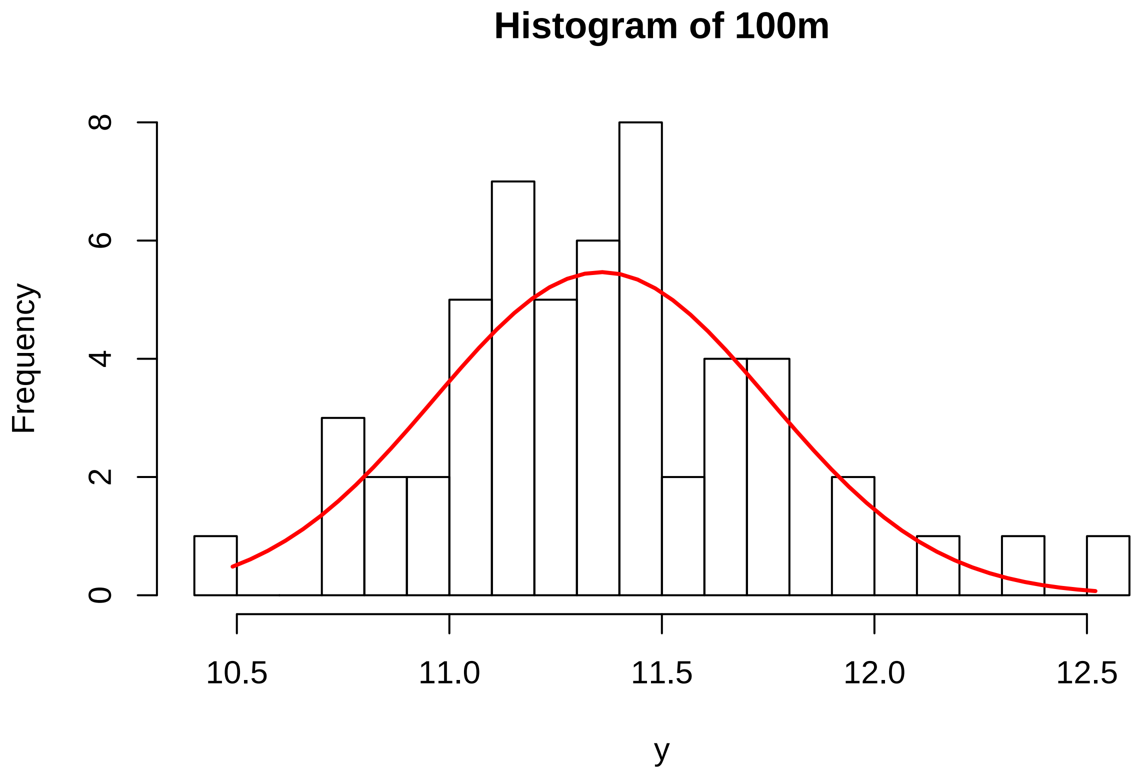
Consider the data set in the T1-9.dat file, National track records for women. For 55 different countries we have the national records for 7 variables (100, 200, 400, 800, 1500, 3000m and marathon). Use R to do the following analyses.

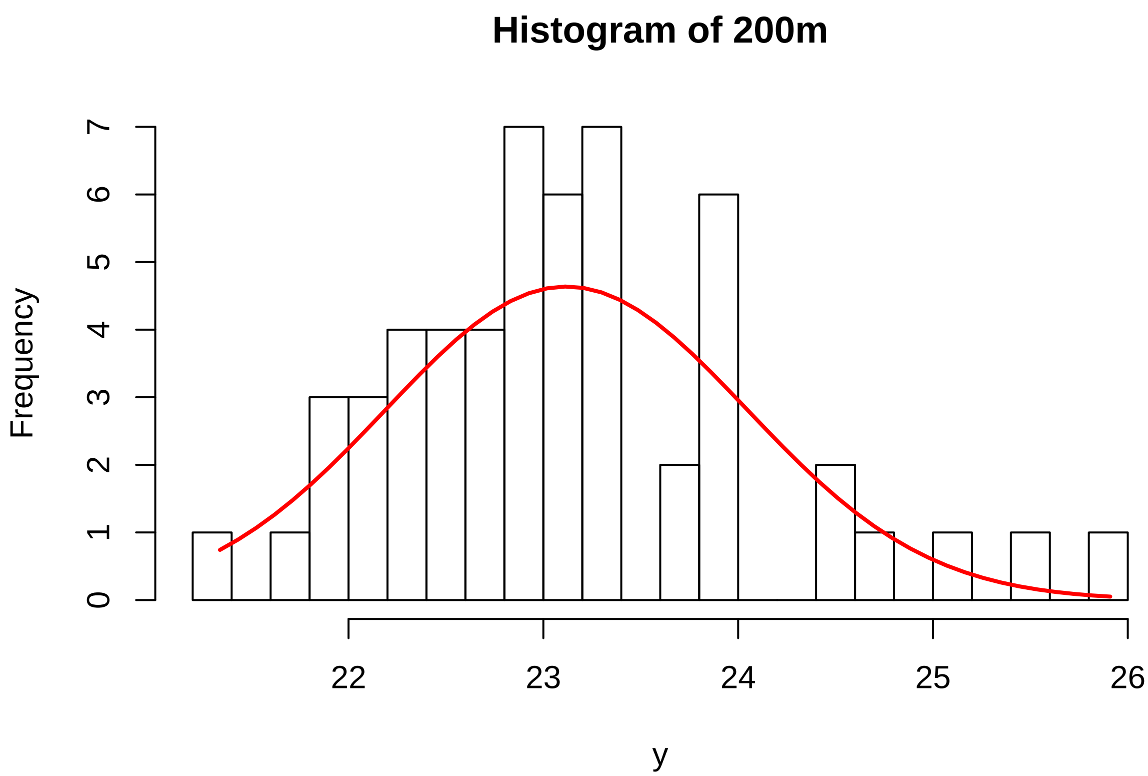
a) Describe the 7 variables with mean values, standard deviations e.t.c.

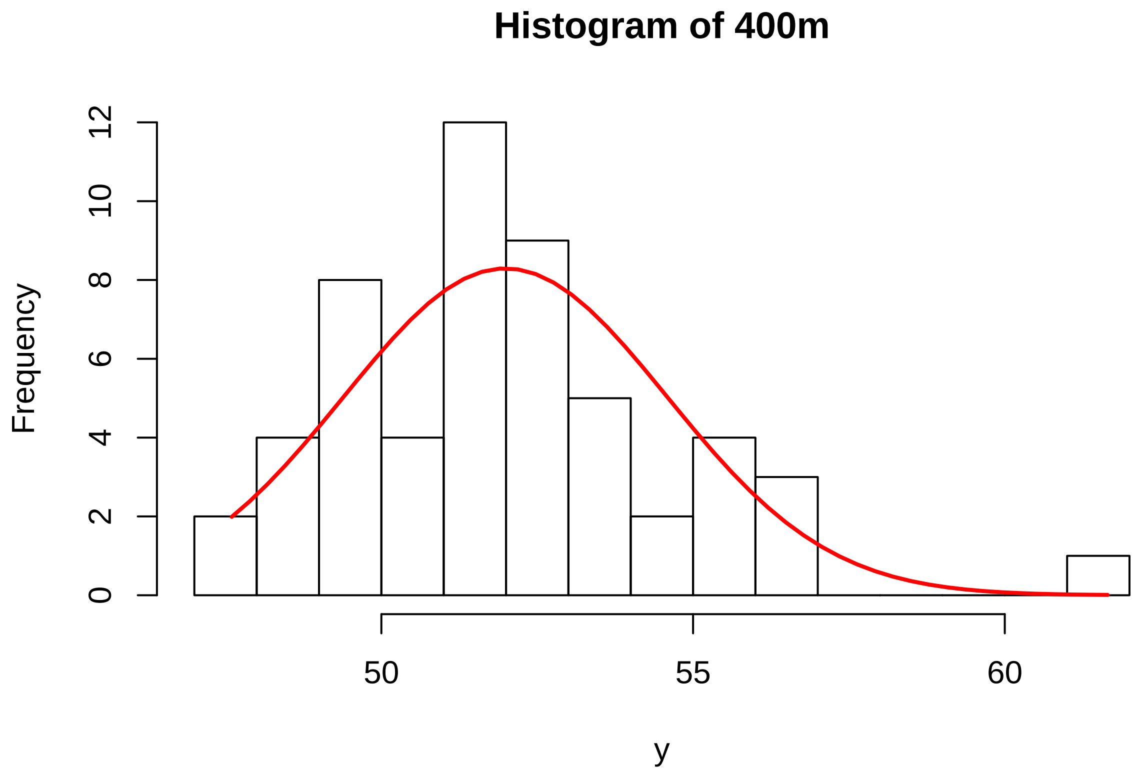
b) Illustrate the variables with different graphs (explore what plotting possibilities R has). Make sure that the graphs look attractive (it is absolutely necessary to look at the labels, font sizes, point types). Are there any apparent extreme values? Do the variables seem normally distributed? Plot 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.

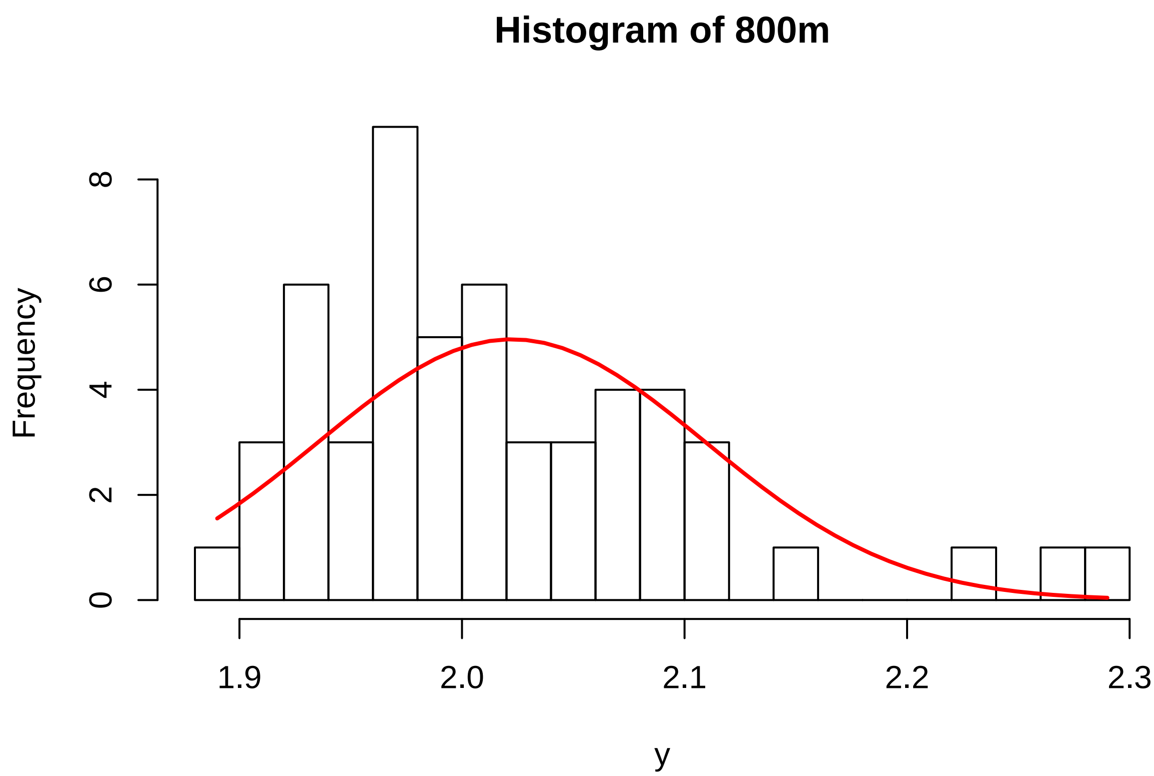
```
## Using country as id variables
```

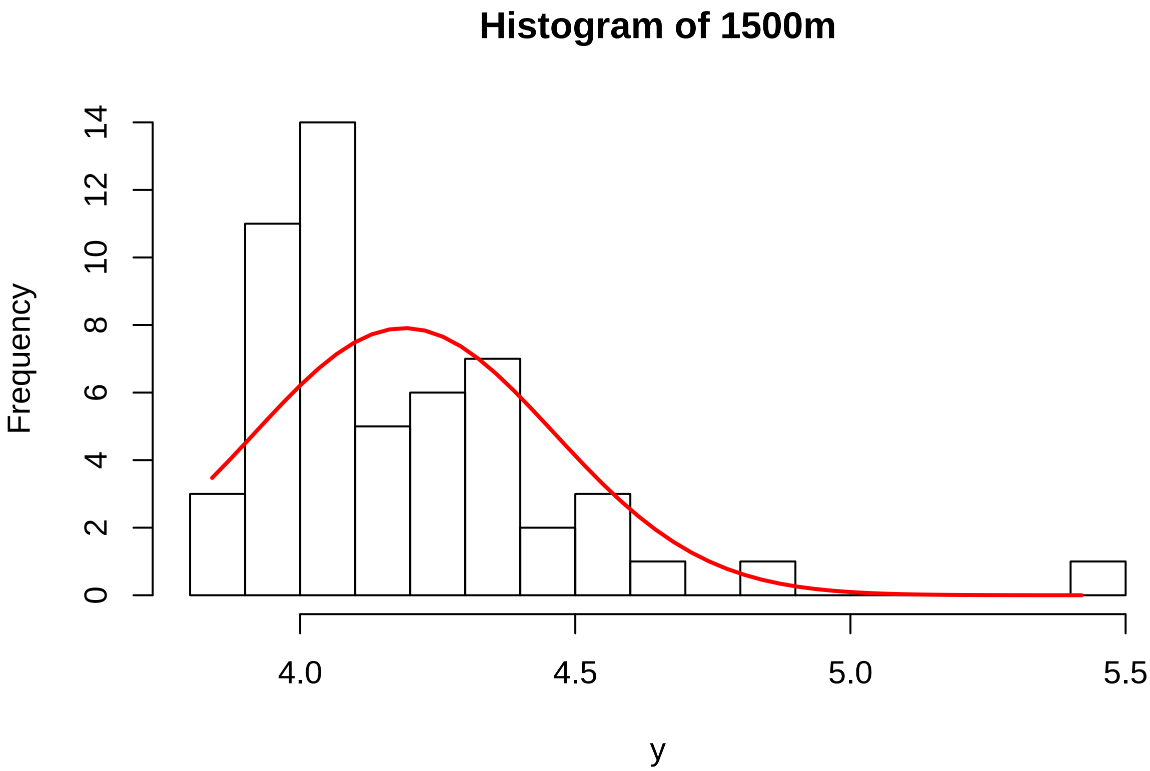


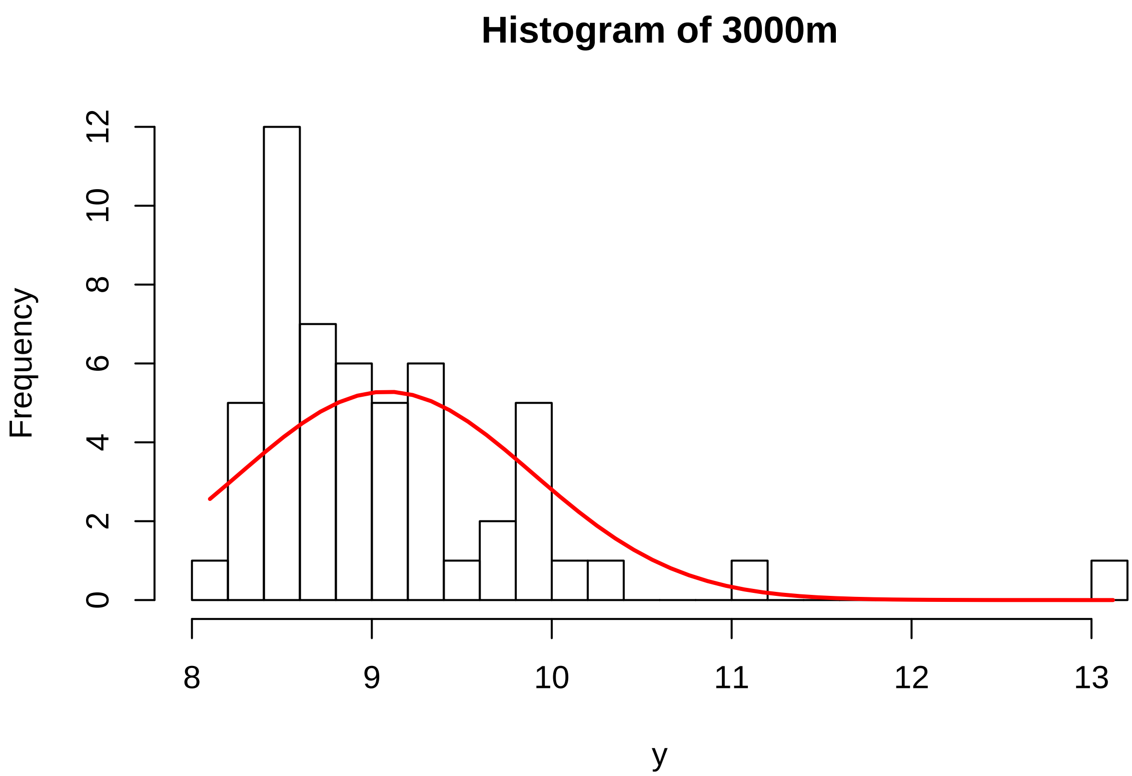

Histogram of 100m
Frequency
0
2
4
6
8
10.5
11.0
11.5
12.0
12.5
y

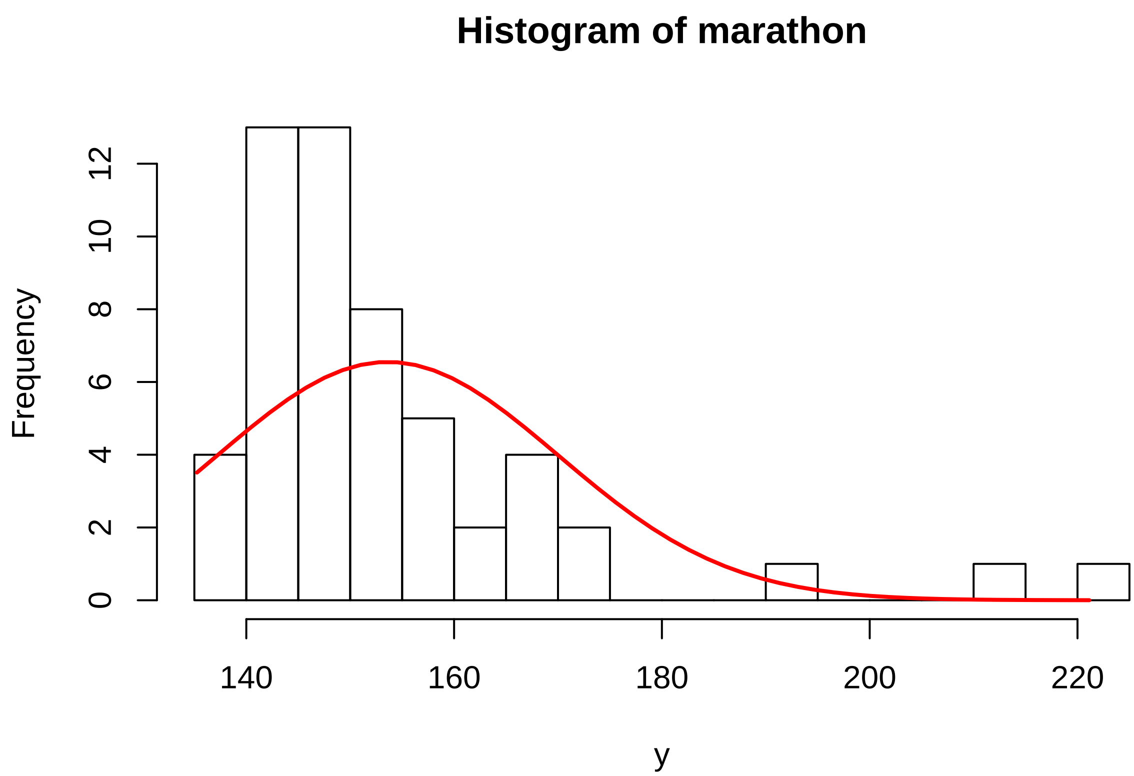

Histogram of 200m
Frequency
0
1
2
3
4
5
6
7
22
23
24
25
26
y


Histogram of 400m
Frequency
0
2
4
6
8
10
12
50
55
60
y


Histogram of 800m
Frequency
0
2
4
6
8
1.9
2.0
2.1
2.2
2.3
y


Histogram of 1500m
Frequency
0
2
4
6
8
10
12
14
4.0
4.5
5.0
5.5
y


Histogram of 3000m
Frequency
0
2
4
6
8
10
12
8
9
10
11
12
13
y



The first categories (100m, 200m and 400m) seem normally distributed by looking at the histograms. The longer races are have more skewed to the right histograms.

Question 2: Relationships between 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.

Covariance Matrix:

	100m	200m	400m	800m	1500m	3000m	marathon
100m	0.1553157	0.3445608	0.8912960	0.0277036	0.0838912	0.2338828	4.334178
200m	0.3445608	0.8630883	2.1928363	0.0661659	0.2027633	0.5543502	10.384988
400m	0.8912960	2.1928363	6.7454576	0.1818079	0.5091768	1.4268158	28.903731
800m	0.0277036	0.0661659	0.1818079	0.0075469	0.0214146	0.0613793	1.219655
1500m	0.0838912	0.2027633	0.5091768	0.0214146	0.0741827	0.2161551	3.539837
3000m	0.2338828	0.5543502	1.4268158	0.0613793	0.2161551	0.6647579	10.706091
marathon	4.3341776	10.3849876	28.9037314	1.2196546	3.5398373	10.7060911	270.270150

Correlation Matrix:

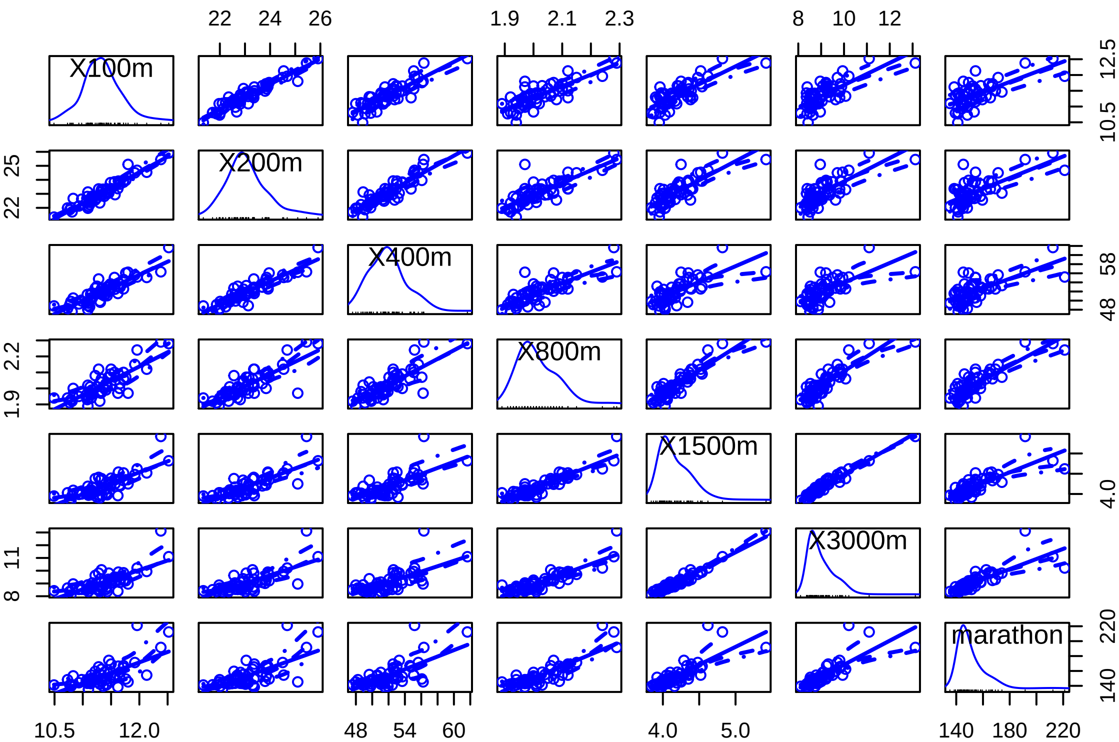
	100m	200m	400m	800m	1500m	3000m	marathon
100m	1.0000000	0.9410886	0.8707802	0.8091758	0.7815510	0.7278784	0.6689597
200m	0.9410886	1.0000000	0.9088096	0.8198258	0.8013282	0.7318546	0.6799537
400m	0.8707802	0.9088096	1.0000000	0.8057904	0.7197996	0.6737991	0.6769384
800m	0.8091758	0.8198258	0.8057904	1.0000000	0.9050509	0.8665732	0.8539900
1500m	0.7815510	0.8013282	0.7197996	0.9050509	1.0000000	0.9733801	0.7905565
3000m	0.7278784	0.7318546	0.6737991	0.8665732	0.9733801	1.0000000	0.7987302

	100m	200m	400m	800m	1500m	3000m	marathon
marathon	0.6689597	0.6799537	0.6769384	0.8539900	0.7905565	0.7987302	1.0000000

By analysing the variance covariance matrix, it can be concluded that countries that have a high performance in "shorter" races (100m, 200m and 400m) do not necessarily have high performance in "long distance" races (800m, 1500m, 3000m and marathon). This is coherent to the fact that short and long races require different training. Normally, the athletes are different altogether in these different categories.

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

```
scatterplotMatrix(data[,2:8])
```



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?

Visually inspecting all scatterplots, we can see that Papua New Guinea, are notable extremes.

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

$$d(\overrightarrow{x}, \overline{x}) = \sqrt{(\overrightarrow{x} - \overline{x})^T(\overrightarrow{x} - \overline{x})}$$

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

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

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 x - mean 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.

```
#euclidean distance
records <- as.matrix(data[,2:8])

#center the data
#deviation matrix
centered <- scale(x = records, center = TRUE, scale = FALSE)

euclid_dist <- sqrt((centered) %*% t(centered))
```

```
## Warning in sqrt((centered) %*% t(centered)): NaNs produced
```

```
#diagonal
euclid_diag <- sort(diag(euclid_dist), decreasing = TRUE, index.return = TRUE)
#extreme values (top 5)
euclid_extremes <- head(euclid_diag, n = 5)
#extract extreme countries
ind <- head(euclid_extremes$ix, 5)
euclid_extreme_countries <- data[ind,1]
#Sweden's index
SWE_ind <- which(data[,1] == "SWE")
#Sweden's Position
SWE_rank_euclid <- which(euclid_extremes$ix == SWE_ind)

print("Top 5 distance extreme countries:")
```

```
## [1] "Top 5 distance extreme countries:"
```

```
euclid_extreme_countries
```

```
## [1] PNG COK SAM BER GBR
## 54 Levels: ARG AUS AUT BEL BER BRA CAN CHI CHN COK COL CRC CZE DEN DOM ... USA
```

```
paste("Sweden's distance rank: ", SWE_rank_euclid)
```

```
## [1] "Sweden's distance rank:  48"
```

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, INSERT LATEX FORMULA HERE where 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 (centering) 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?

```
#diagonal of variance covariance matrix
covars <- cov(records)
V <- matrix(0, nrow = ncol(records), ncol = ncol(records))
diag(V) <- diag(covars)
#compute distance
dist3c <- (centered %*% solve(V) %*% t(centered))^(1/2)

#diagonal
diag3c <- sort(diag(dist3c), decreasing = TRUE, index.return = TRUE)
#extreme values (top 5)
extremes3c <- head(diag3c, n = 5)
#extract extreme countries
ind3c<- head(extremes3c$ix, 5)
extreme_countries3c <- data[ind3c,1]
#Sweden's Position
SWE_rank_3c <- which(extremes3c$ix == SWE_ind)

print("Top 5 distance extreme countries:")
```

```
## [1] "Top 5 distance extreme countries:"
```

```
extreme_countries3c
```

```
## [1] SAM COK PNG USA SIN
## 54 Levels: ARG AUS AUT BEL BER BRA CAN CHI CHN COK COL CRC CZE DEN DOM ... USA
```

```
paste("Sweden's distance rank: ", SWE_rank_3c)
```

```
## [1] "Sweden's distance rank:  50"
```

The euclidean distance for normalized data results in a different country list for most extreme ones, altough some countries are in both lists, but not necessarily in the same positions: Samoa (SAM), Cook Islands (COK) and Papua-New Guinea (PNG). This second set of extremes is more consistent because the for the first three race categories (100m, 200m and 400m), time is measured in seconds while the others (800m, 1500m, 3000m and marathon) are measured in minutes.

d) The most common statistical distance is the Mahalanobis distance, INSERT LATEX FORMULA HERE where 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 dV(· , ·) does). Compute the Mahalanobis distance, which countries are most extreme now?

```
dist3d <- (centered %*% solve(covars) %*% t(centered))^(1/2)
#diagonal
diag3d <- sort(diag(dist3d), decreasing = TRUE, index.return = TRUE)
#extreme values (top 5)
extremes3d <- head(diag3d, n = 5)
#extract extreme countries
ind3d<- head(extremes3d$ix, 5)
extreme_countries3d <- data[ind3d,1]
#Sweden's Position
SWE_rank_3d <- which(extremes3d$ix == SWE_ind)

print("Top 5 distance extreme countries:")
```

```
## [1] "Top 5 distance extreme countries:"
```

```
extreme_countries3d
```

```
## [1] SAM  PNG  KORN COK  MEX
## 54 Levels: ARG AUS AUT BEL BER BRA CAN CHI CHN COK COL CRC CZE DEN DOM ... USA
```

```
paste("Sweden's distance rank: ", SWE_rank_3d)
```

```
## [1] "Sweden's distance rank:  54"
```

The most extreme countries given by the Mahalanobis distance 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.

In this case, extreme countries normally have poor performance in races, but that is not always the case. For some distances definitions, high performance countries like USA and Great Britain appear as extremes as well. So "extremism" is not a measure of how "slow" a country is, but rather how far from the overall mean the country is. Graphically, this means that the region of equal euclidean distance (a hypersphere) gets distorted to regions of equal statistical distance (hyperellipsoids) with different statistical weights applied: the marginal variances in c) and variance-covariances in d).

By ranking the countries in decreasing order of distance, Sweden's lowers in position in the rank for the different distances definitions in questions b), c), and d) respectively (48th, 50th and 54th place). It is possible to conclude that Sweden's records are more consistent throught all race categories, since it is positioned in the bottom positions of all ranks.

The Czekanowski's diagram is showed below:

```
df = data.matrix(data)
rownames(df) = data[,1]
df = df[,2:8]
```

```r
x = czek_matrix(df)
plot(x)
```