Multivariate Statistical Methods

Assignment 1

Allan Gholmi, Emma Wallentinsson, Rasmus Holm

2017-11-24

Question 1

The data consists of national track times for women in 100m, 200m , 400m, 800m, 1500m, 3000m and marathon.

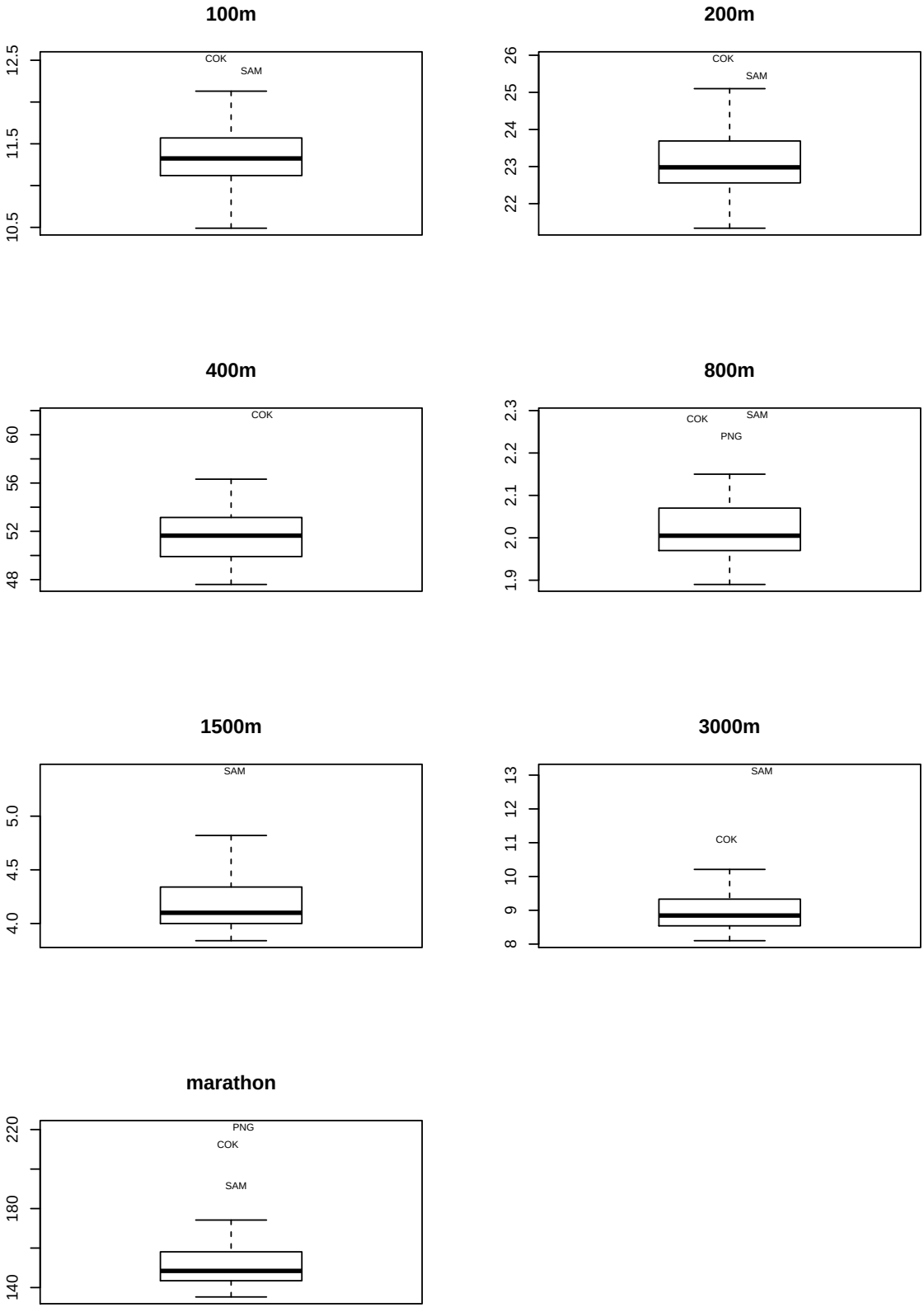
```
data <- read.table("../data/T1-9.dat")
names(data) <- c("country", "100m", "200m", "400m", "800m", "1500m", "3000m", "marathon")
numeric_data <- data[, -1]
```

a)

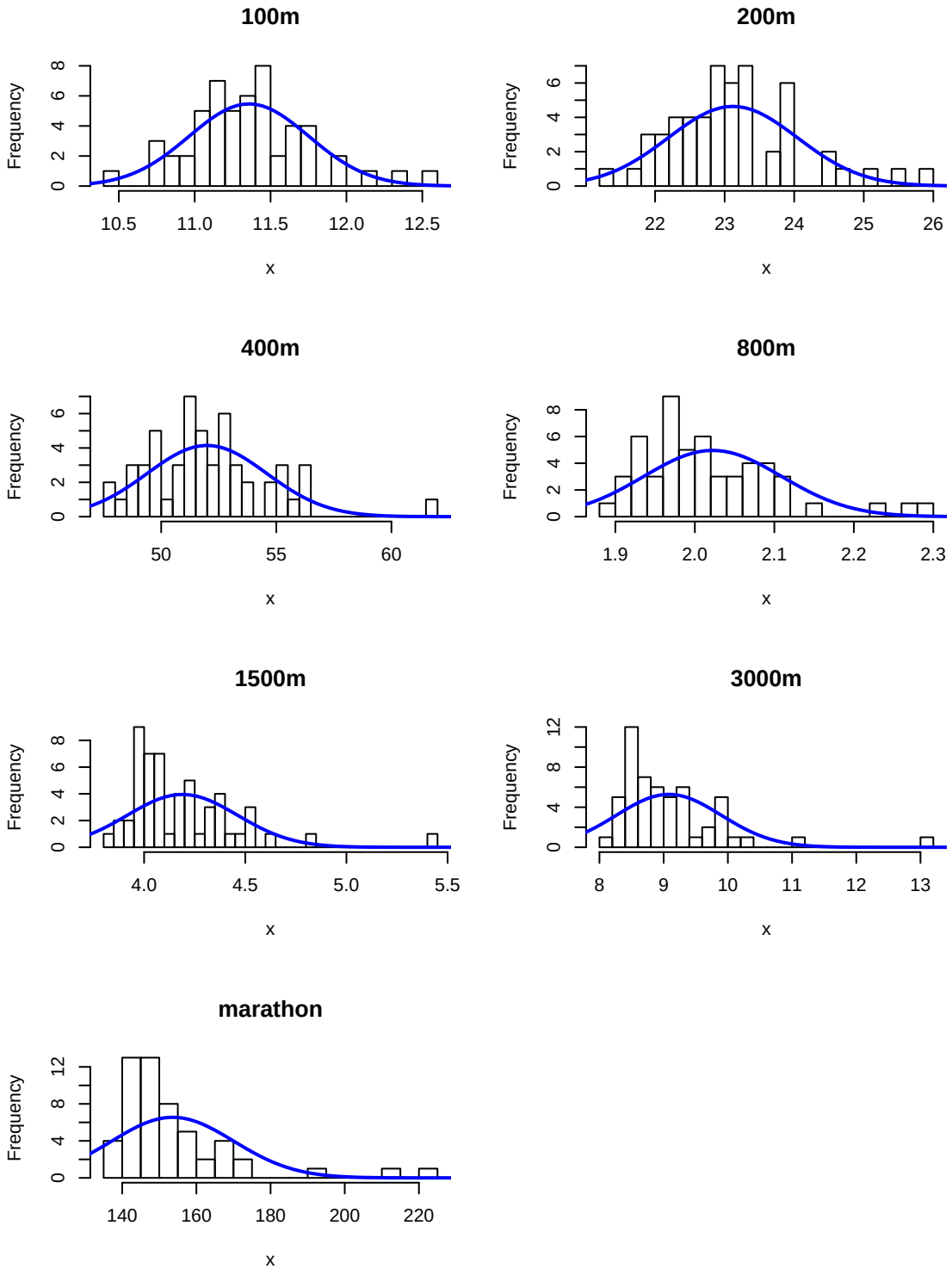
The variables are described with summary statistics in the table:

```
summary(numeric_data)
#>       100m            200m            400m            800m
#>  Min.   :10.49   Min.   :21.34   Min.   :47.60   Min.   :1.890
#>  1st Qu.:11.12   1st Qu.:22.57   1st Qu.:49.97   1st Qu.:1.970
#>  Median :11.32   Median :22.98   Median :51.65   Median :2.005
#>  Mean   :11.36   Mean   :23.12   Mean   :51.99   Mean   :2.022
#>  3rd Qu.:11.57   3rd Qu.:23.61   3rd Qu.:53.12   3rd Qu.:2.070
#>  Max.   :12.52   Max.   :25.91   Max.   :61.65   Max.   :2.290
#>      1500m           3000m           marathon
#>  Min.   :3.840   Min.   : 8.100   Min.   :135.2
#>  1st Qu.:4.003   1st Qu.: 8.543   1st Qu.:143.5
#>  Median :4.100   Median : 8.845   Median :148.4
#>  Mean   :4.189   Mean   : 9.081   Mean   :153.6
#>  3rd Qu.:4.338   3rd Qu.: 9.325   3rd Qu.:157.7
#>  Max.   :5.420   Max.   :13.120   Max.   :221.1
print("Standard Deviation")
#> [1] "Standard Deviation"
apply(numeric_data, 2, sd)
#>       100m       200m       400m       800m       1500m       3000m
#>  0.39410116 0.92902547 2.59720188 0.08687304 0.27236502 0.81532689
#>    marathon
#> 16.43989508
```

b)


100m
COK
SAM
10.5
11.5
12.5
200m
COK
SAM
22
23
24
25
26
400m
COK
48
52
56
60
800m
COK
SAM
PNG
1.9
2.0
2.1
2.2
2.3
1500m
SAM
4.0
4.5
5.0
3000m
SAM
COK
8
9
10
11
12
13
marathon
PNG
COK
SAM
140
180
220

The variables are displayed in box plots. We can notice from the scales of the variables that they are measured in different units. 100m, 200m, and 400m are probably measured in seconds while the others are measured in minutes. We can see that all outliers are above the 3rd quantile, i.e. have worse times than most countries. PNG, COK, and SAM are the only countries that deviate significantly.



A histogram for each variable is created to see the distributions. The first three variables, 100m, 200m and 300m do at least resemble the Gaussian (normal) distribution. The rest seems to resemble either the chisquare- or F distribution. The distances 300m, 800m, 1500m, 3000m and marathon has outliers on the far right of the histograms.

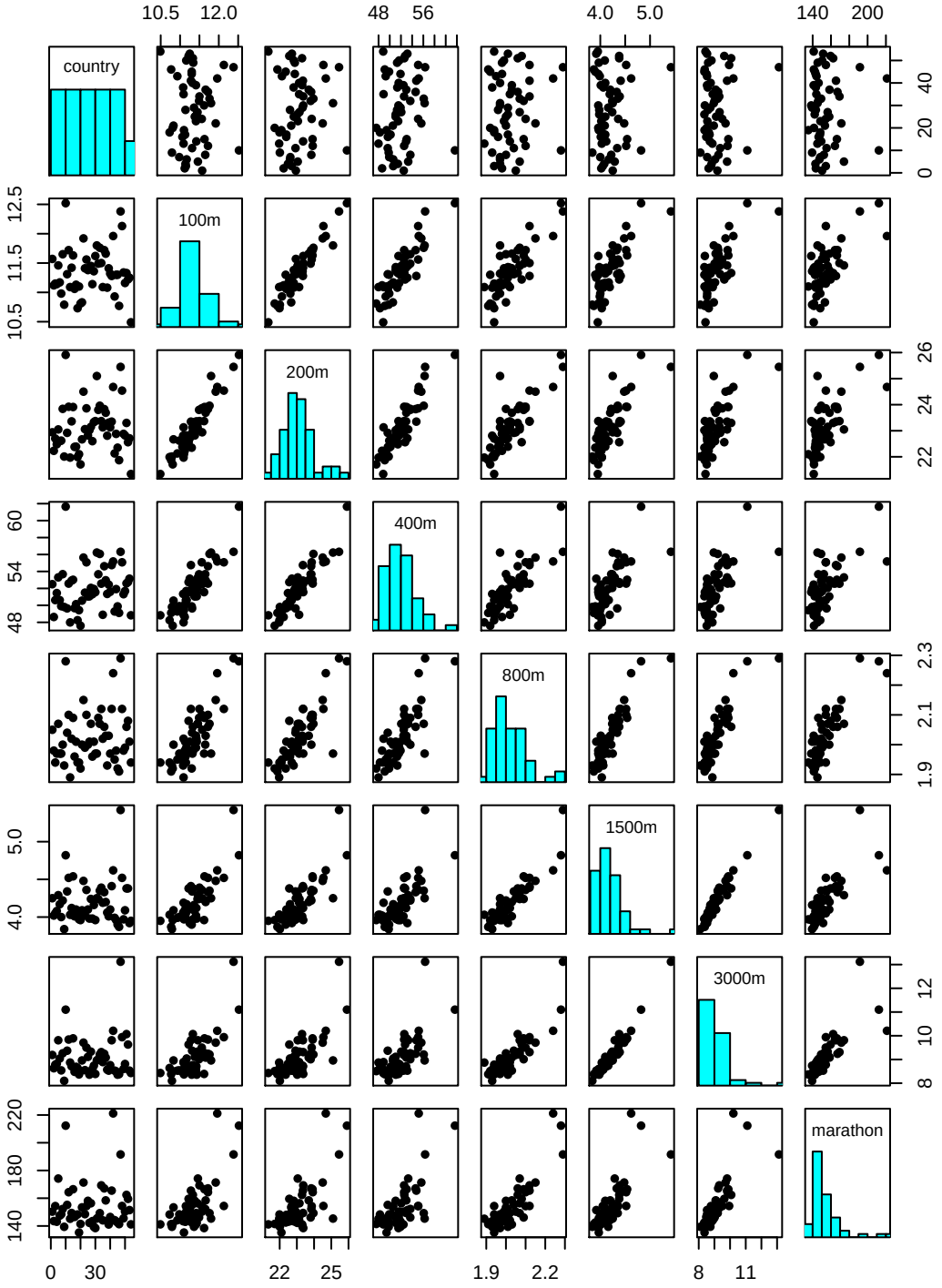
Question 2

a)

```
covariance_mat <- cov(numeric_data)
covariance_mat
#>                 100m       200m       400m        800m      1500m
#> 100m     0.15531572  0.3445608  0.8912960 0.027703564 0.08389119
#> 200m     0.34456080  0.8630883  2.1928363 0.066165898 0.20276331
#> 400m     0.89129602  2.1928363  6.7454576 0.181807932 0.50917683
#> 800m     0.02770356  0.0661659  0.1818079 0.007546925 0.02141457
#> 1500m    0.08389119  0.2027633  0.5091768 0.021414570 0.07418270
#> 3000m    0.23388281  0.5543502  1.4268158 0.061379315 0.21615514
#> marathon 4.33417757 10.3849876 28.9037314 1.219654647 3.53983732
#>                3000m   marathon
#> 100m      0.23388281   4.334178
#> 200m      0.55435017  10.384988
#> 400m      1.42681579  28.903731
#> 800m      0.06137932   1.219655
#> 1500m     0.21615514   3.539837
#> 3000m     0.66475793  10.706091
#> marathon 10.70609113 270.270150
correlation_mat <- cor(numeric_data)
correlation_mat
#>                100m      200m      400m      800m     1500m     3000m
#> 100m     1.0000000 0.9410886 0.8707802 0.8091758 0.7815510 0.7278784
#> 200m     0.9410886 1.0000000 0.9088096 0.8198258 0.8013282 0.7318546
#> 400m     0.8707802 0.9088096 1.0000000 0.8057904 0.7197996 0.6737991
#> 800m     0.8091758 0.8198258 0.8057904 1.0000000 0.9050509 0.8665732
#> 1500m    0.7815510 0.8013282 0.7197996 0.9050509 1.0000000 0.9733801
#> 3000m    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
#> 100m      0.6689597
#> 200m      0.6799537
#> 400m      0.6769384
#> 800m      0.8539900
#> 1500m     0.7905565
#> 3000m     0.7987302
#> marathon 1.0000000
```

In the covariance matrix, only positive values are present. So as one variable increase, every other variable also increases linearly. In the correlation matrix, we can see that there is strongest correlation between 1500m and 3000m at 0.97.

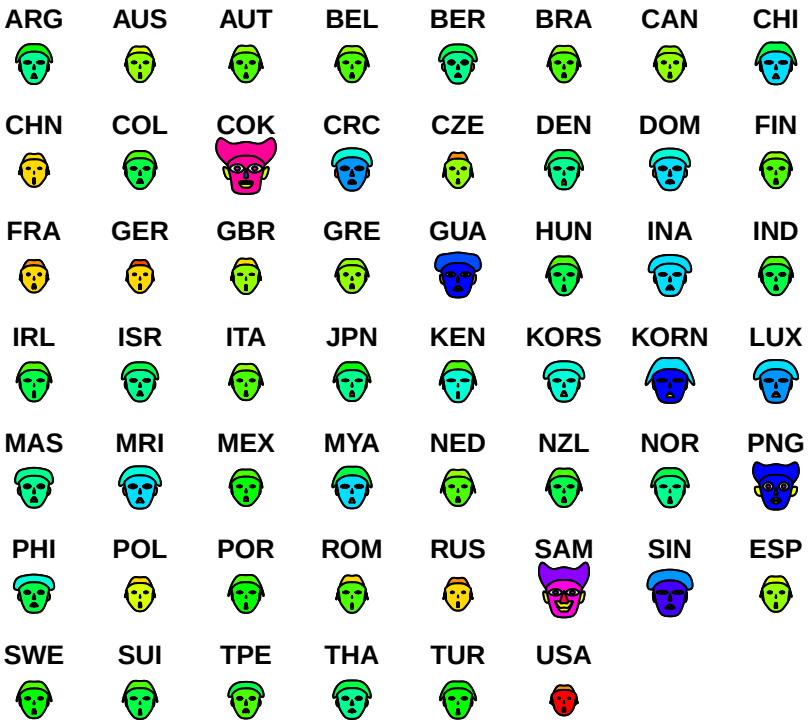
b)



Above we can see a scatterplot matrix of the different variables. The covariance and correlation matrices created before proves that we have a positive correlation between the variables. We can here see some outliers but comparing the plots we can see that every plot seems to have outliers where they appear mostly in the

topright corner of each plot. For example, between 1500m and 3000m we have an observation in the topright corner. Shortly speaking, there seems to be at least one country in each variable which runs the distance slower than the other women.

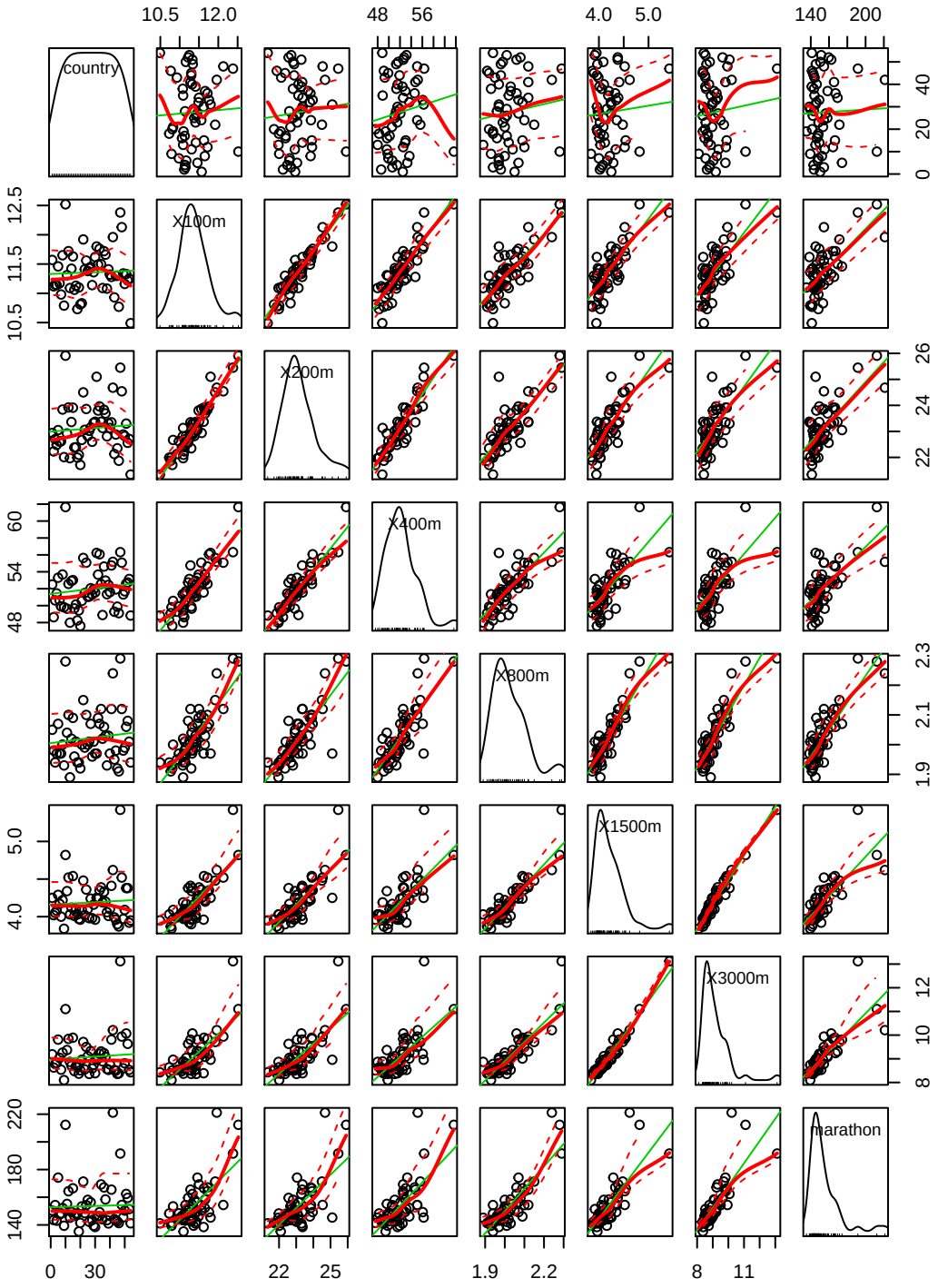
c)




```
#> effect of variables:
#>  modified item       Var
#>  "height of face   " "100m"
#>  "width of face    " "200m"
#>  "structure of face" "400m"
#>  "height of mouth  " "800m"
#>  "width of mouth   " "1500m"
#>  "smiling          " "3000m"
#>  "height of eyes   " "marathon"
#>  "width of eyes    " "100m"
```

```
#>  "height of hair   " "200m"
#>  "width of hair   "  "400m"
#>  "style of hair   "  "800m"
#>  "height of nose  "  "1500m"
#>  "width of nose   "  "3000m"
#>  "width of ear    "  "marathon"
#>  "height of ear   "  "100m"
```

As seen in the plot of Chernoff faces, outliers are COK, CRC, GUA, KORN, PNG, SAM, SIN, USA. For example, the big width of hair on the face for COK says that the country has a high time on the distance 400m.



The green line (straight) is the robust-regression line for each pair, and the thick red line (curvy) is a non-parametric regression smoother. The dotted red lines is a confidence band for the smoothed line.

Question 3

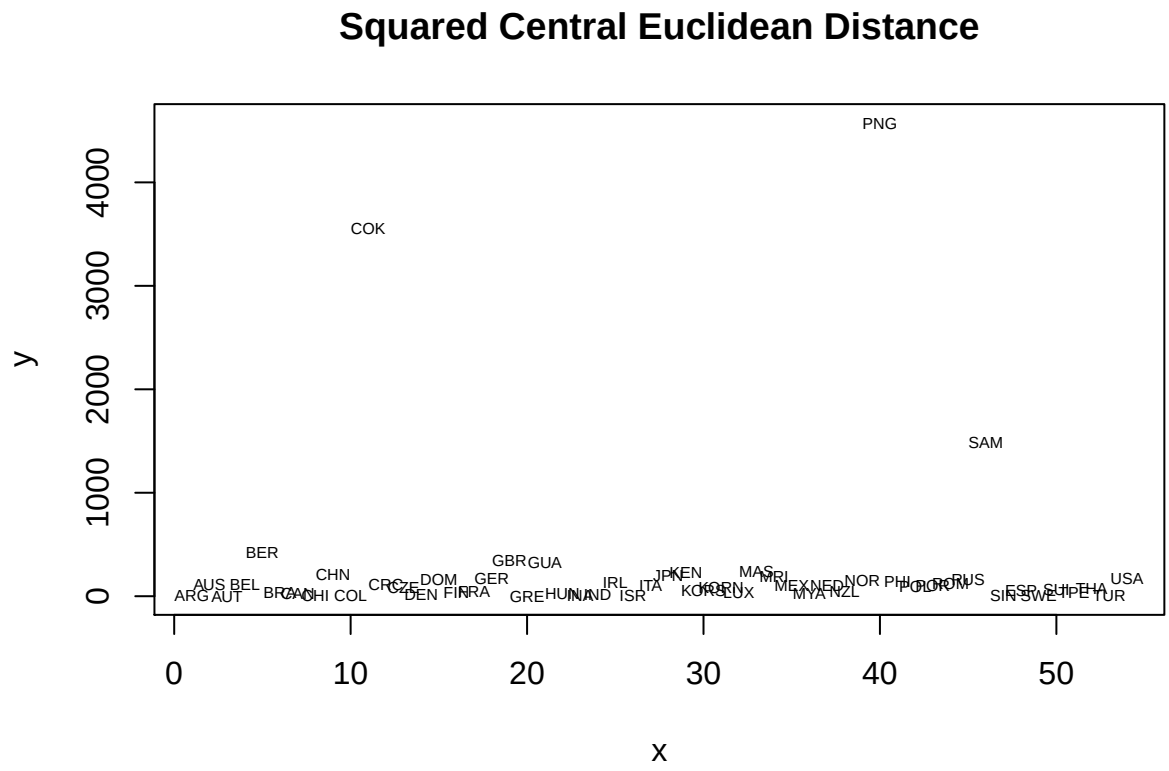
a)

The 5 countries that are most extreme are the dots that are far away from everybody else and we can see that Cook Islands, Samoa, Singapore, North Korea and Papa New guinea fits that description and the reason seems to be that they have higher records for each variables than the rest.

b)

```
edist_central_sq <- X_central %*% t(X_central)
country_central_edist <- diag(edist_central_sq)

central_edist_extreme_countries <- data[order(country_central_edist, decreasing=TRUE), 1][1:5]
as.character(central_edist_extreme_countries)
#> [1] "PNG" "COK" "SAM" "BER" "GBR"
```

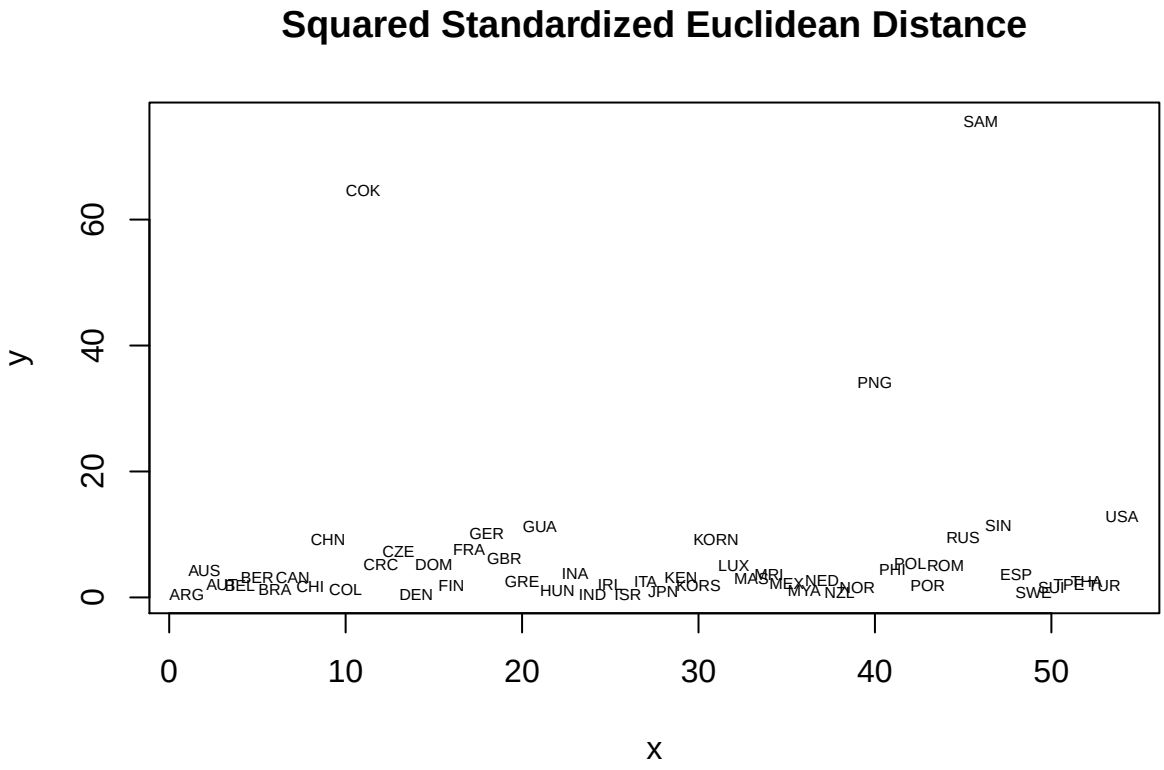


Here we can see the different Euclidean distances between the countries. The Euclidean distance that has the most extreme values are Papa New Guinea, Cooks Island, Samoa, Bermuda and Great Britain.

c)

```r
V_inv <- diag(1 / apply(X, 2, var))
edist_standard_sq <- X_central %*% V_inv %*% t(X_central)
country_standard_edist <- diag(edist_standard_sq)

standard_edist_extreme_countries <- data[order(country_standard_edist, decreasing=TRUE), 1][1:5]
as.character(standard_edist_extreme_countries)
#> [1] "SAM" "COK" "PNG" "USA" "SIN"
```

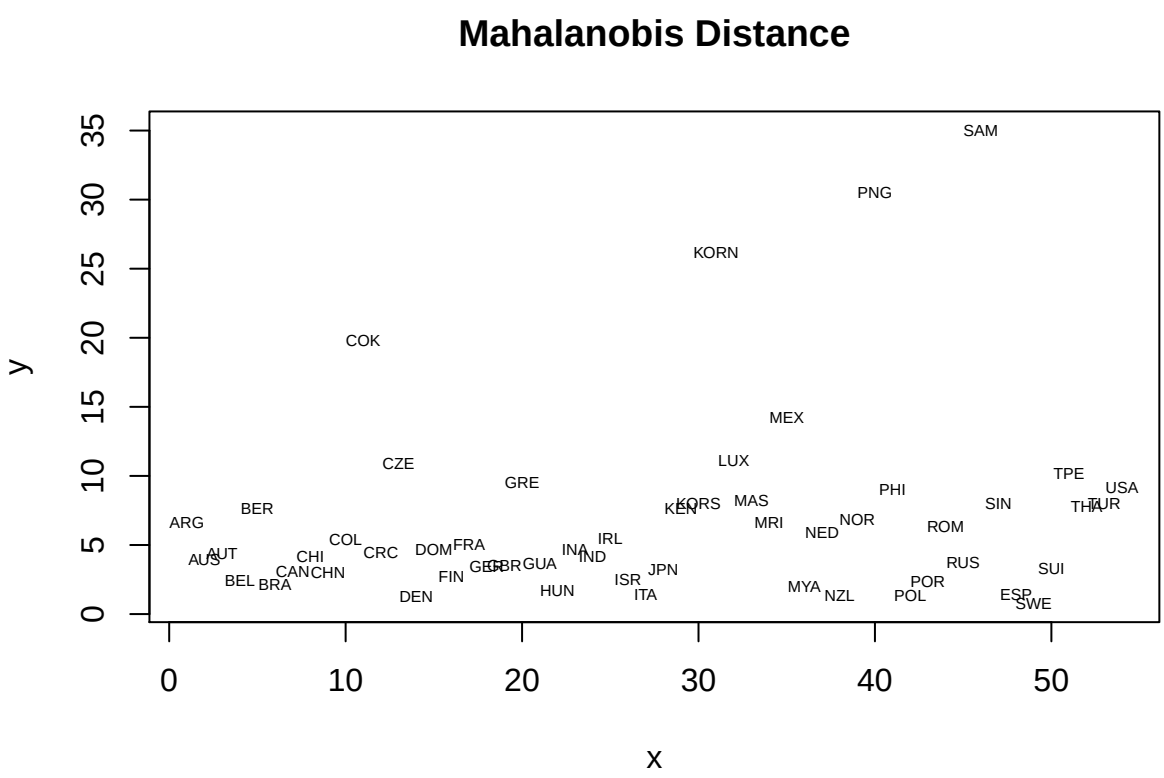


Here we have standardized with both mean and standard deviations and computed the Euclidean distance again. They are then as well and we can see that the most extreme values are Samo, Cooks Island, Papa New Guinea, USA and lastly Singapore.

d)

```r
mdist_sq <- X_central %*% solve(covariance_mat) %*% t(X_central)
country_mdist <- diag(mdist_sq)

mdist_extreme_countries <- data[order(country_mdist, decreasing=TRUE), 1][1:5]
as.character(mdist_extreme_countries)
#> [1] "SAM"  "PNG"  "KORN" "COK"  "MEX"
```



When using the Mahalanobis distance we can see that the most extreme values are Samoa, Papa New Guinea, North Korea, Cook's Island and Mexico.

e)

Samoa has been the most extreme in 2 cases, in assignment 3c and 3d. We also have Papua New Guinea and Cook's Island as extreme cases in all cases. Sweden appears in all cases very low, with low values in each assignment meaning that Sweden does not differ that much between the seven variables. Sweden is more balanced in the different seven variables compared to the more extreme cases mentioned before.

Appendix

Code

```
data <- read.table("../data/T1-9.dat")
names(data) <- c("country", "100m", "200m", "400m", "800m", "1500m", "3000m", "marathon")
numeric_data <- data[, -1]

summary(numeric_data)
print("Standard Deviation")
apply(numeric_data, 2, sd)

set.seed(123)

old <- par(mfrow=c(4, 2))

labels <- as.character(data$country)

for (col in 1:ncol(numeric_data)) {
    x <- numeric_data[, col]
    b <- boxplot(x, cex=0, main=names(data)[col + 1])
    idx <- which(x < b$stats[1] | x > b$stats[5])
    text(b$group + runif(length(b$out), -0.1, 0.1), b$out, labels[idx], cex=0.6)
}

par(old)

old <- par(mfrow=c(4, 2))

for (col in names(numeric_data)) {
    x <- numeric_data[, col]
    h <- hist(x, breaks=25, main=col)
    offset <- (max(x) - min(x)) / 2
    xfit <- seq(min(x) - offset, max(x) + offset, length = 100)
    yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))
    yfit <- yfit * diff(h$mids[1:2]) * length(x)
    lines(xfit, yfit, col="blue", lwd=2)
}

par(old)

covariance_mat <- cov(numeric_data)
covariance_mat
correlation_mat <- cor(numeric_data)
correlation_mat

panel.hist <- function(x, ...) {
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y / max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
```

```
}

pairs(data, pch=16, diag.panel=panel.hist)

library(aplpack)
ncolors=20
faces(numeric_data, labels=as.character(data$country),
      col.hair = rainbow(ncolors,  start = 0, end = 0.9),
      col.face = rainbow(ncolors,  start = 0, end = 0.9))

library(car)
scatterplotMatrix(data)

X <- as.matrix(numeric_data)
countries <- as.character(data$country)
x <- 1:length(countries)

means <- colMeans(X)
X_central <- X - rep(1, nrow(X)) %*% t(means)

edist_central_sq <- X_central %*% t(X_central)
country_central_edist <- diag(edist_central_sq)

central_edist_extreme_countries <- data[order(country_central_edist, decreasing=TRUE), 1][1:5]
as.character(central_edist_extreme_countries)

y <- country_central_edist
plot(x, y, main="Squared Central Euclidean Distance", type="n")
text(x=x, y=y, labels=countries, cex=0.5)

V_inv <- diag(1 / apply(X, 2, var))
edist_standard_sq <- X_central %*% V_inv %*% t(X_central)
country_standard_edist <- diag(edist_standard_sq)

standard_edist_extreme_countries <- data[order(country_standard_edist, decreasing=TRUE), 1][1:5]
as.character(standard_edist_extreme_countries)

y <- country_standard_edist

plot(x, y, main="Squared Standardized Euclidean Distance", type="n")
text(x=x, y=y, labels=countries, cex=0.5)

mdist_sq <- X_central %*% solve(covariance_mat) %*% t(X_central)
country_mdist <- diag(mdist_sq)

mdist_extreme_countries <- data[order(country_mdist, decreasing=TRUE), 1][1:5]
as.character(mdist_extreme_countries)

y <- country_mdist
plot(x, y, main="Mahalanobis Distance", type="n")
text(x=x, y=y, labels=countries, cex=0.5)
```