Multivariate Statistics Assignment 1

Aashana Nijhawan

07/11/2019

Question 1: Describing individual variables

a) Describe the 7 variables with mean values, standard deviations etc.

```
trackData = read.table("T1-9.dat")
colnames(trackData) = c("Countries", "100m", "200m", "400m", "800m", "1500m", "3000m", "Marathon")
samplesnames = c("100m", "200m", "400m", "800m", "1500m", "3000m", "Marathon")
trackData[,5:8] = trackData[,5:8]*60
trackData_Mean = colMeans(trackData[,2:8])
trackData_SD = apply(trackData[,2:8], 2, sd)
trackData_Median = apply(trackData[,2:8], 2, median)
```

b) Illustrate the variables with different graphs

** 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 **

```
plt100 = ggplot()+
  geom_histogram(aes(x=trackData$`100m`, y=..density..), color = "black", fill = "#343434",alpha=0.5,bi
  geom_vline(aes(xintercept = trackData_Mean[1],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[1],color="Median"))+
  geom_density(aes(x=trackData$`100m`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()
```

```
plt200 = ggplot()+
  geom_histogram(aes(x=trackData$`200m`, y= ..density..), color = "black", fill = "#343434",alpha=0.5,b
  geom_vline(aes(xintercept = trackData_Mean[2],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[2],color="Median"))+
  geom_density(aes(x=trackData$`200m`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()
```

```
plt400 = ggplot()+
  geom_histogram(aes(x=trackData$`400m`, y= ..density..), color = "black", fill = "#343434",alpha=0.5,b
  geom_vline(aes(xintercept = trackData_Mean[3],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[3],color="Median"))+
  geom_density(aes(x=trackData$`400m`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()
```

```
plt800 = ggplot()+
  geom_histogram(aes(x=trackData$`800m`, y= ..density..), color = "black", fill = "#343434",alpha=0.5,b
```

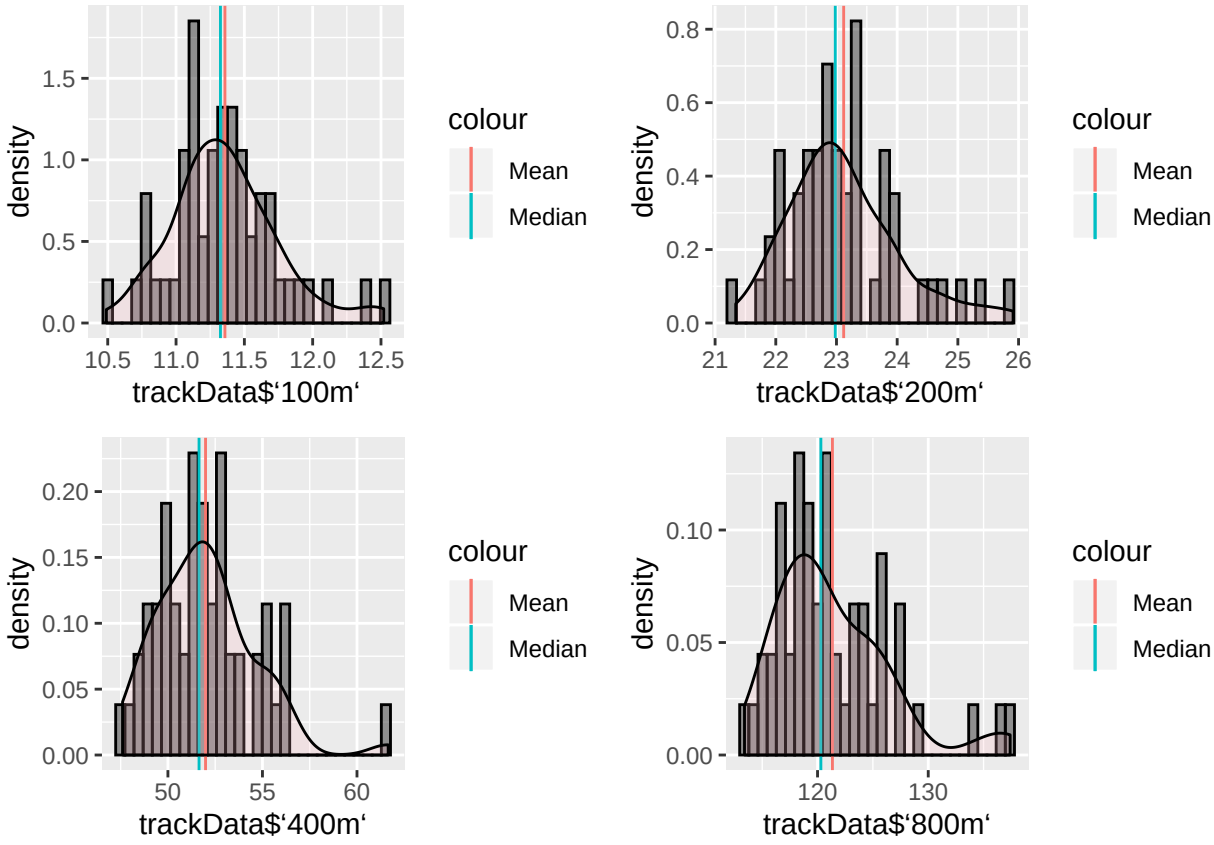
```
  geom_vline(aes(xintercept = trackData_Mean[4],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[4],color="Median"))+
  geom_density(aes(x=trackData$`800m`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()
```

```
plt1500 = ggplot()+
  geom_histogram(aes(x=trackData$`1500m`, y= ..density..), color = "black", fill = "#343434",alpha=0.5,
  geom_vline(aes(xintercept = trackData_Mean[5],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[5],color="Median"))+
  geom_density(aes(x=trackData$`1500m`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()
```

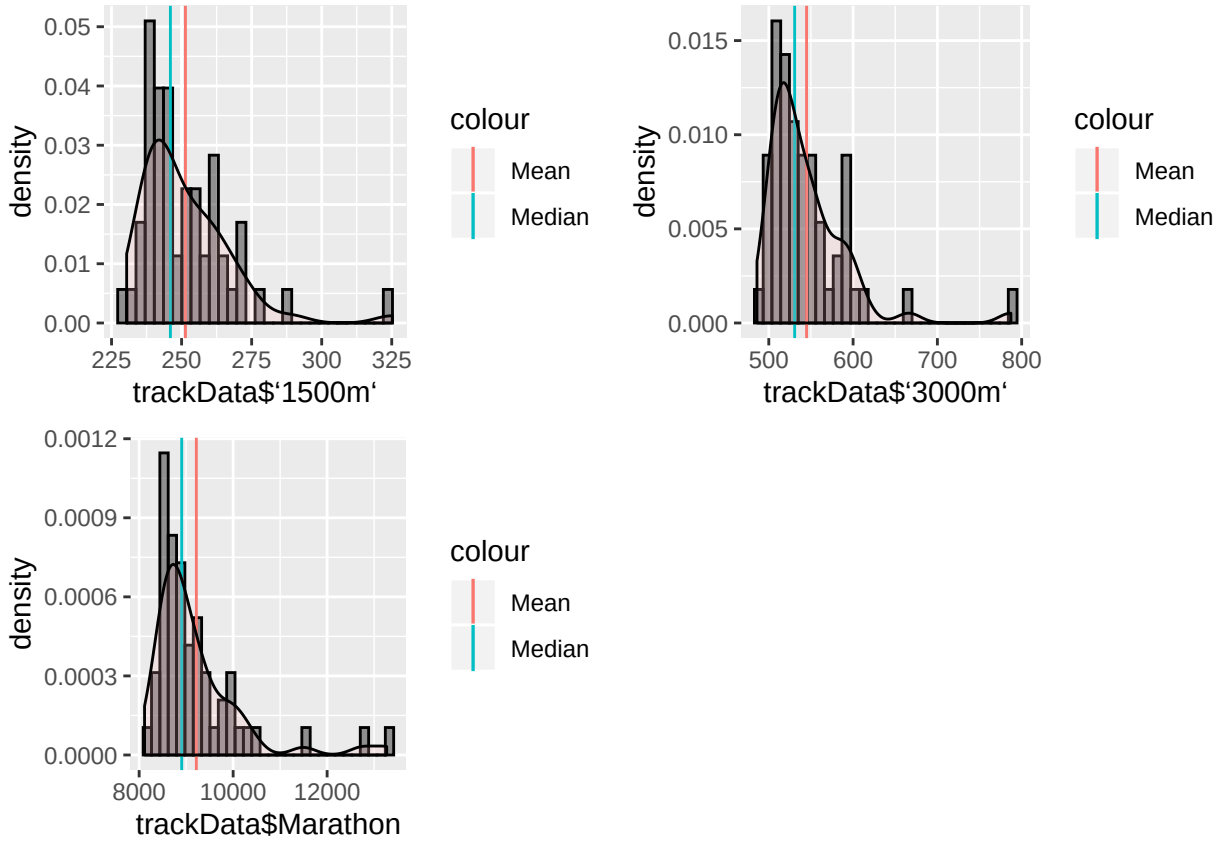
```
plt3000 = ggplot()+
  geom_histogram(aes(x=trackData$`3000m`, y= ..density..), color = "black", fill = "#343434",alpha=0.5,
  geom_vline(aes(xintercept = trackData_Mean[6],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[6],color="Median"))+
  geom_density(aes(x=trackData$`3000m`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()
```

```
pltM = ggplot()+
  geom_histogram(aes(x=trackData$`Marathon`, y= ..density..), color = "black", fill = "#343434",alpha=0
  geom_vline(aes(xintercept = trackData_Mean[7],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[7],color="Median"))+
  geom_density(aes(x=trackData$`Marathon`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()
```

```
grid.arrange(plt100,plt200,plt400,plt800, ncol=2)
```



```
grid.arrange(plt1500,plt3000,pltM, ncol=2)
```

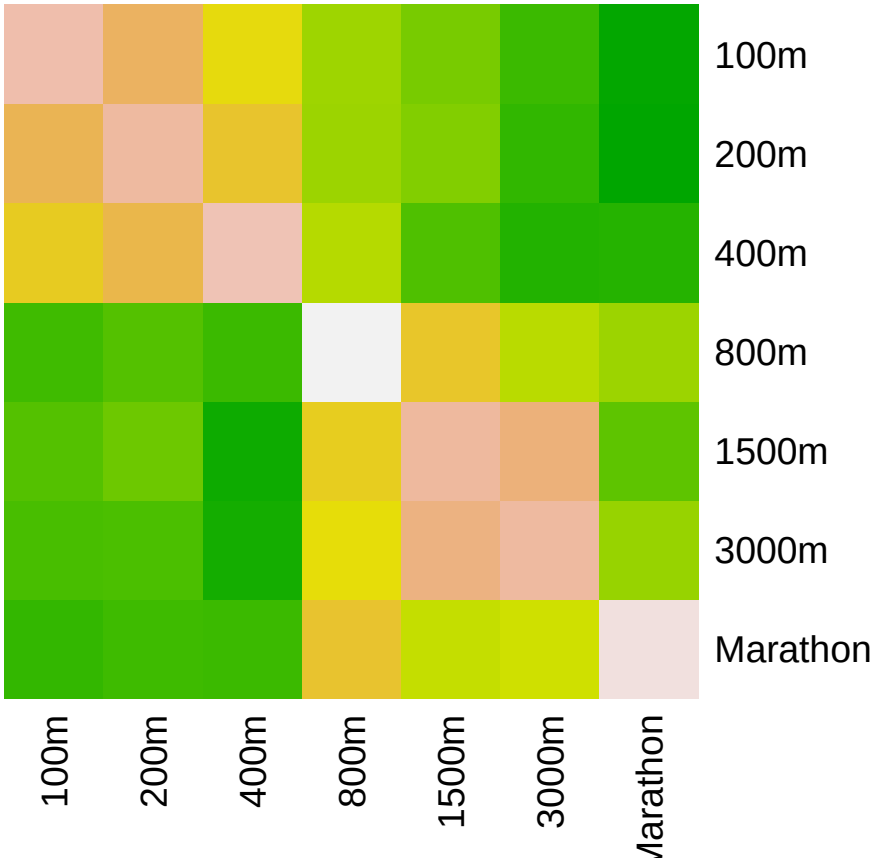


Question 2: Relationships between the variables

a) Compute the covariance and correlation matrices

```
trackData_Cov = cov(trackData[,2:8])
trackData_Cor = cor(trackData[,2:8])
```

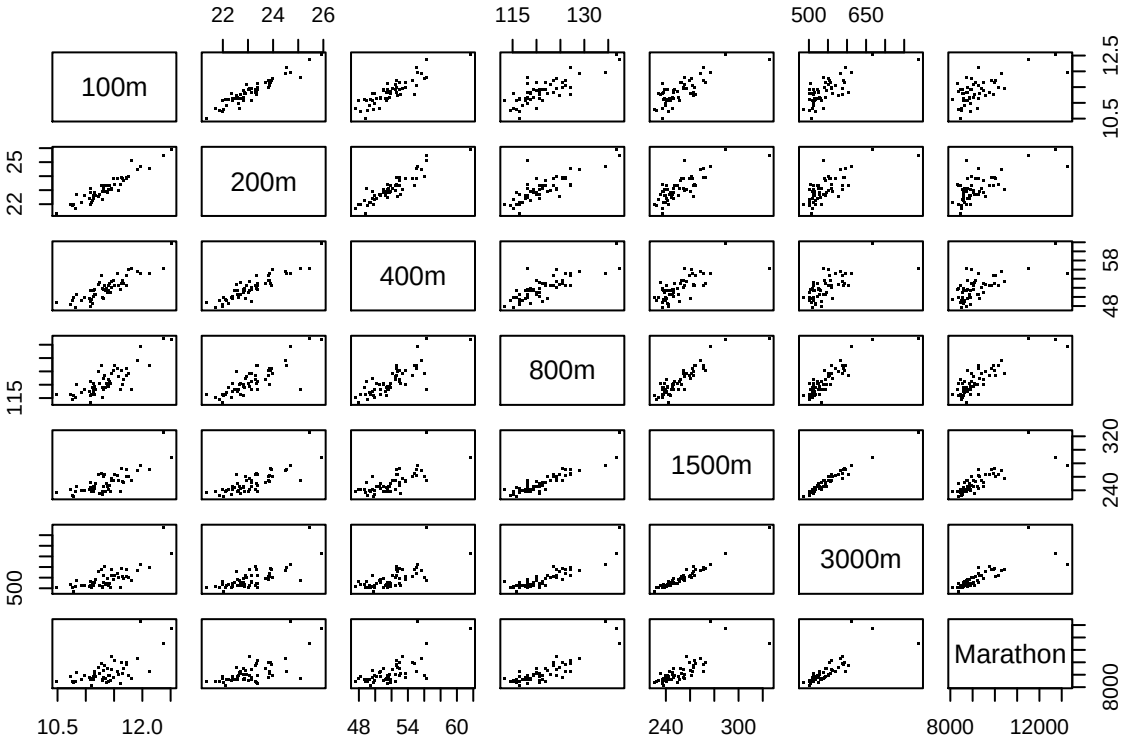
```
heatmap(trackData_Cor, col = terrain.colors(256), Rowv=NA,Colv=NA, revC=T)
```



Each square shows the correlation between the variables on each axis. Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables. The close to 1 the correlation is the more positively correlated they are; that is as one increases so does the other and the closer to 1 the stronger this relationship is. A correlation closer to -1 is similar, but instead of both increasing one variable will decrease as the other increases. The diagonals are all 1/dark green because those squares are correlating each variable to itself (so it's a perfect correlation). For the rest the larger the number and darker the color the higher the correlation between the two variables. The plot is also symmetrical about the diagonal since the same two variables are being paired together in those squares.

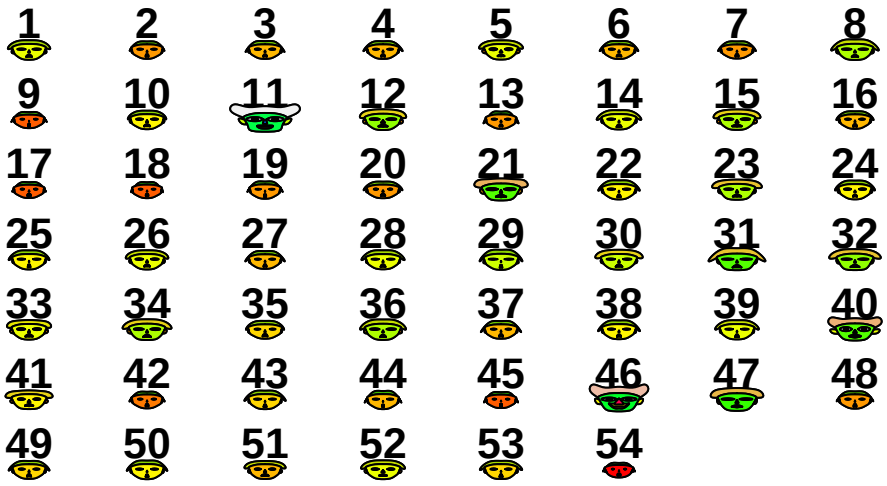
b) Scatterplots between each pair of variables

```
pairs(trackData[,2:8], pch = ".", cex = 1.5)
```



c)

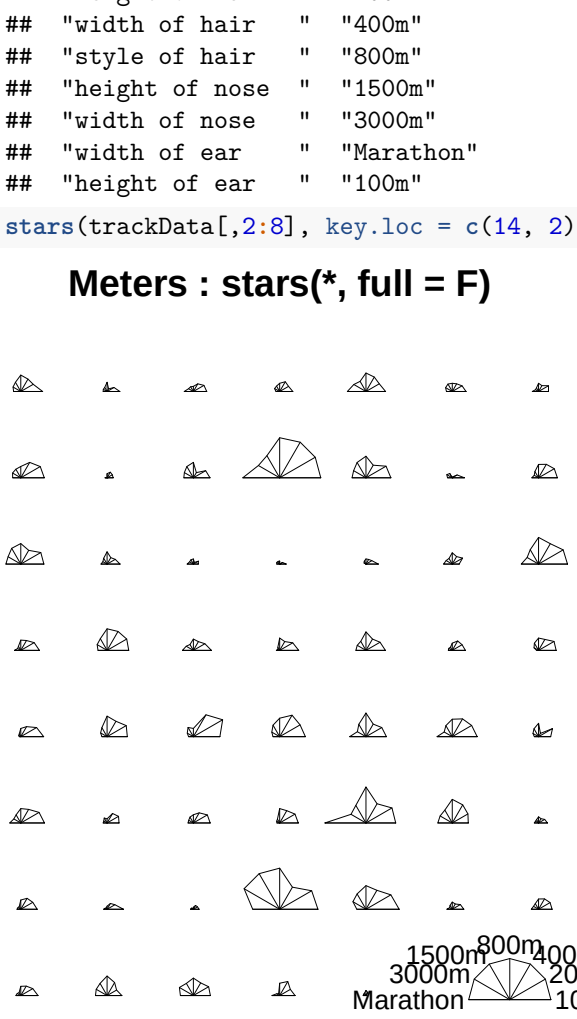
```
### Chernoff faces.
ncolors=c("pink","blue","red","yellow","green","purple","orange","magenta")
faces(trackData[,2:8],face.type=1, col.face =rainbow(50)) ## with colour
```



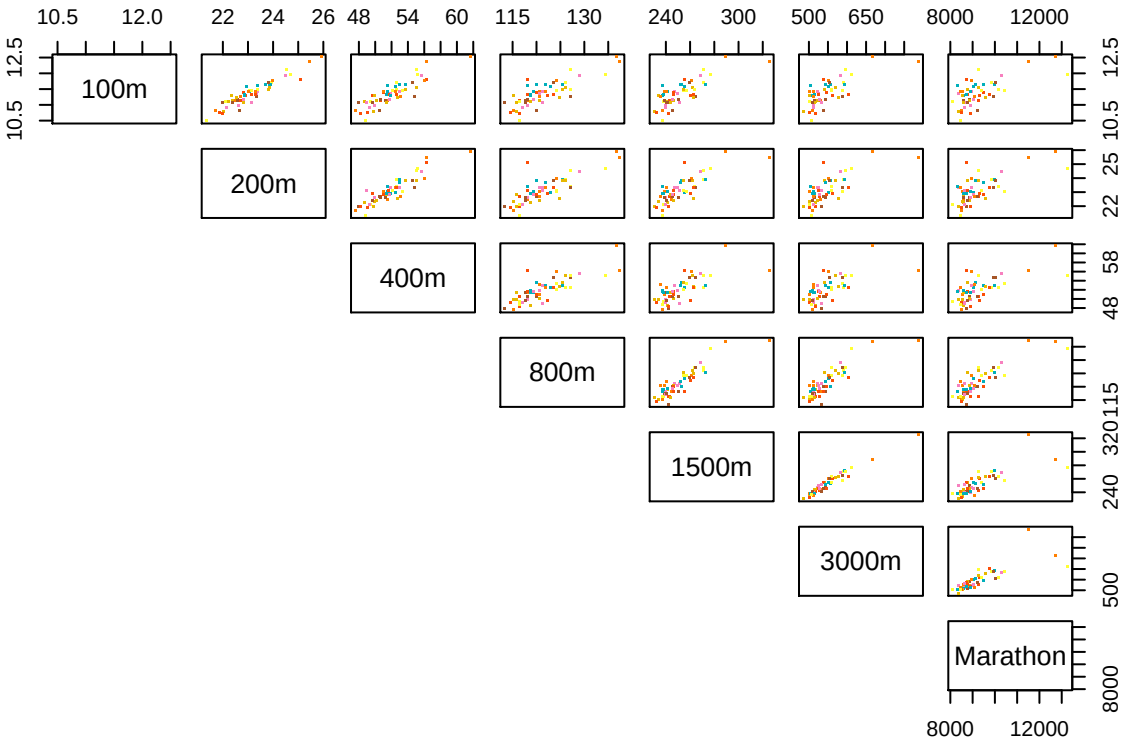

```
## effect of variables:
##  modified item       Var
##  "height of face   " "100m"
##  "width of face    " "200m"
##  "structure of face" "400m"
##  "height of mouth  " "800m"
##  "width of mouth   " "1500m"
```

```
##  "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"
```

```
stars(trackData[,2:8], key.loc = c(14, 2), main = "Meters : stars(*, full = F)", full = FALSE)
```




```
my_cols <- c("#00AFBB", "#E7B800", "#FC4E07" , "#ff7f00" , "#ffff33" , "#a65628" , "#f781bf")
pairs(trackData[,2:8], pch = ".",  cex = 1.5,
      col = my_cols,
      lower.panel=NULL)
```



Question 3: Examining for extreme values

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(aplpack) # for Chernoff faces
library(gridExtra)

trackData = read.table("T1-9.dat")
colnames(trackData) = c("Countries", "100m", "200m", "400m", "800m", "1500m", "3000m", "Marathon")
samplesnames = c("100m", "200m", "400m", "800m", "1500m", "3000m", "Marathon")
trackData[,5:8] = trackData[,5:8]*60
trackData_Mean = colMeans(trackData[,2:8])
trackData_SD = apply(trackData[,2:8], 2, sd)
trackData_Median = apply(trackData[,2:8], 2, median)




plt100 = ggplot()+
  geom_histogram(aes(x=trackData$`100m`, y=..density..), color = "black", fill = "#343434",alpha=0.5,bi
  geom_vline(aes(xintercept = trackData_Mean[1],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[1],color="Median"))+
```

```
  geom_density(aes(x=trackData$`100m`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()
plt200 = ggplot()+
  geom_histogram(aes(x=trackData$`200m`, y= ..density..), color = "black", fill = "#343434",alpha=0.5,bi
  geom_vline(aes(xintercept = trackData_Mean[2],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[2],color="Median"))+
  geom_density(aes(x=trackData$`200m`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()

plt400 = ggplot()+
  geom_histogram(aes(x=trackData$`400m`, y= ..density..), color = "black", fill = "#343434",alpha=0.5,bi
  geom_vline(aes(xintercept = trackData_Mean[3],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[3],color="Median"))+
  geom_density(aes(x=trackData$`400m`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()

plt800 = ggplot()+
  geom_histogram(aes(x=trackData$`800m`, y= ..density..), color = "black", fill = "#343434",alpha=0.5,bi
  geom_vline(aes(xintercept = trackData_Mean[4],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[4],color="Median"))+
  geom_density(aes(x=trackData$`800m`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()

plt1500 = ggplot()+
  geom_histogram(aes(x=trackData$`1500m`, y= ..density..), color = "black", fill = "#343434",alpha=0.5,b
  geom_vline(aes(xintercept = trackData_Mean[5],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[5],color="Median"))+
  geom_density(aes(x=trackData$`1500m`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()

plt3000 = ggplot()+
  geom_histogram(aes(x=trackData$`3000m`, y= ..density..), color = "black", fill = "#343434",alpha=0.5,b
  geom_vline(aes(xintercept = trackData_Mean[6],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[6],color="Median"))+
  geom_density(aes(x=trackData$`3000m`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()

pltM = ggplot()+
  geom_histogram(aes(x=trackData$`Marathon`, y= ..density..), color = "black", fill = "#343434",alpha=0
  geom_vline(aes(xintercept = trackData_Mean[7],color="Mean"))+
  geom_vline(aes(xintercept = trackData_Median[7],color="Median"))+
  geom_density(aes(x=trackData$`Marathon`, y=..density..), fill="lightpink", alpha=0.2)+
  theme()

grid.arrange(plt100,plt200,plt400,plt800, ncol=2)
grid.arrange(plt1500,plt3000,pltM, ncol=2)
trackData_Cov = cov(trackData[,2:8])
trackData_Cor = cor(trackData[,2:8])

heatmap(trackData_Cor, col = terrain.colors(256), Rowv=NA,Colv=NA, revC=T)
pairs(trackData[,2:8], pch = ".", cex = 1.5)
### Chernoff faces.
ncolors=c("pink","blue","red","yellow","green","purple","orange","magenta")
```

```r
faces(trackData[,2:8],face.type=1, col.face =rainbow(50)) ## with colour

stars(trackData[,2:8], key.loc = c(14, 2), main = "Meters : stars(*, full = F)", full = FALSE)
my_cols <- c("#00AFBB", "#E7B800", "#FC4E07" , "#ff7f00" , "#ffff33" , "#a65628" , "#f781bf")
pairs(trackData[,2:8], pch = ".",  cex = 1.5,
      col = my_cols,
      lower.panel=NULL)
```