Assignment 4 ST302

your name

2023-04-19

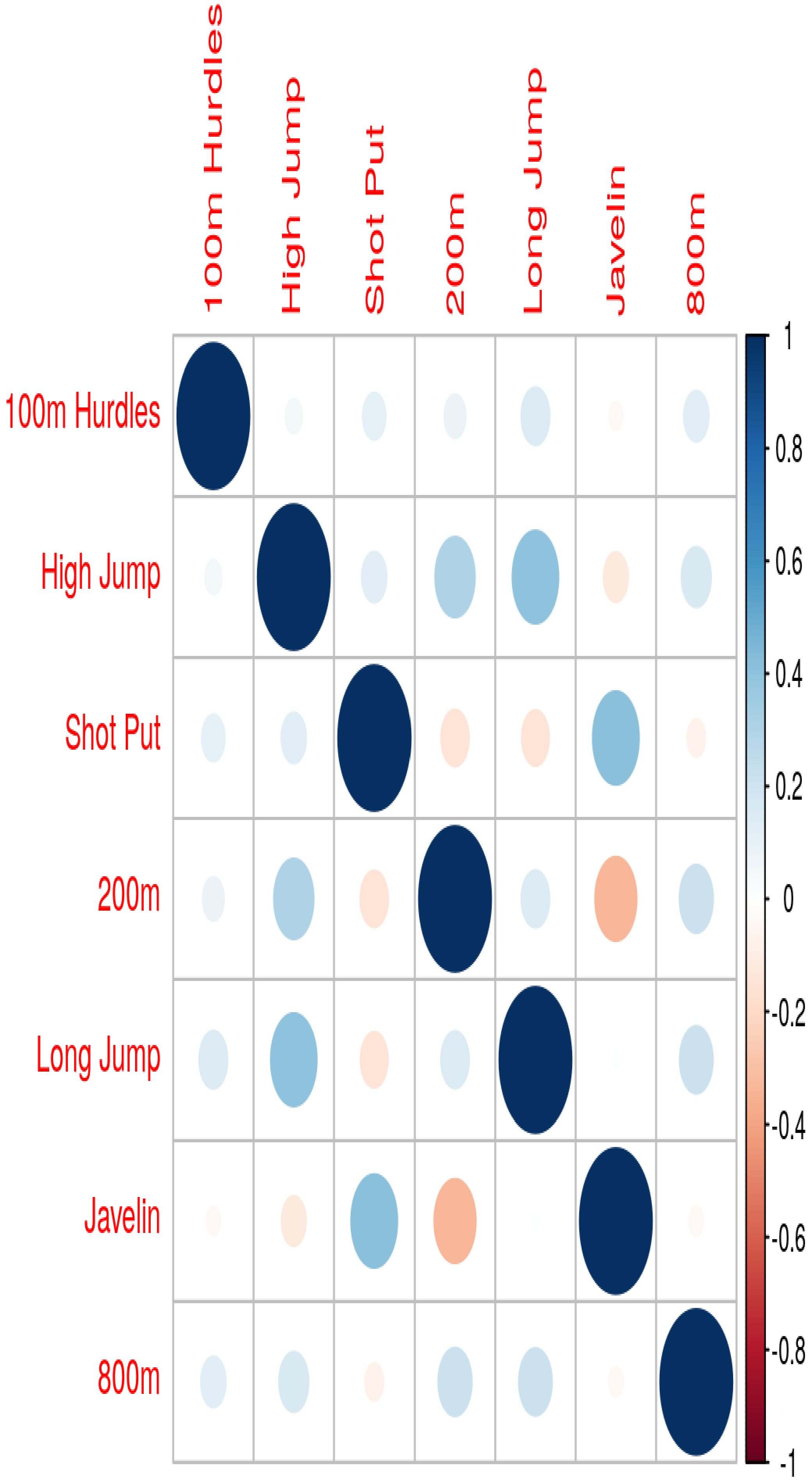
#Set up (Part 1)

```
#install.packages(c("GGally", "naniar", "plotly"))
library(tidyverse)
library(GGally)
library(naniar)
library(plotly)
library(corrplot)

hep <- read_csv(here::here("hwk", "Data", "Hep2012.csv"))
hep <- hep[,c("Athlete", "100m Hurdles","High Jump", "Shot Put","200m" , "Long Jump","Javelin", "800m"   )]
hep <- hep |> mutate(total = rowSums(across(`100m Hurdles`:`800m`))) |>arrange(desc(total))
```

Question 1.

Calculate the correlation of the event scores with option use="pairwise.complete.obs". Use corrplot with re-ordering to show the correlation matrix heatmap for the event scores. Identify pairs of variables with high positive correlation. Identify pairs of variables with high negative correlation.

```
Scores <- hep[, 2:8] #This filters out the Athletes.
Paired <- cor(Scores, use = "pairwise.complete.obs")
corrplot(Paired)
```


100m Hurdles
High Jump
Shot Put
200m
Long Jump
Javelin
800m
100m Hurdles
High Jump
Shot Put
200m
Long Jump
Javelin
800m
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

```
#The pair of variables with highest positive correlation are the dark blue  circles. The lighter the shade of blue, the less correlation it has. The variables with high negative correlation will be red in color.
#Of course, there is a high positive correlation with variables comparing to themselves (E.g 100m Hurdles has a very positive correlation with itself) Outside of those, (Long Jump & High Jump) & (Javelin & Shot Put) have the highest positve correlation with each other. The pair with the highest negative correlation are (Javelin & 200m), which is barely in the -0.4 mark as is. Thus I would say there is no pairs of variables with high negative correlation.
```

Question 2.

Make a scatterplot matrix of the event scores for the heptathlon data using ggpairs, using the order of variables identified by corrplot. Using ggplotly, identify the names of any outlying Athletes and explain what is unusual about them.

```
G <- ggpairs(hep, columns = c("Shot Put", "Javelin", "100m Hurdles", "800m", "200m", "High Jump", "Long Jump"), aes(text = Athlete))
ggplotly(G, tooltip = c("text", "x", "y")) |> highlight(on = "plotly_selected", off = "plotly_deselect")
```

4

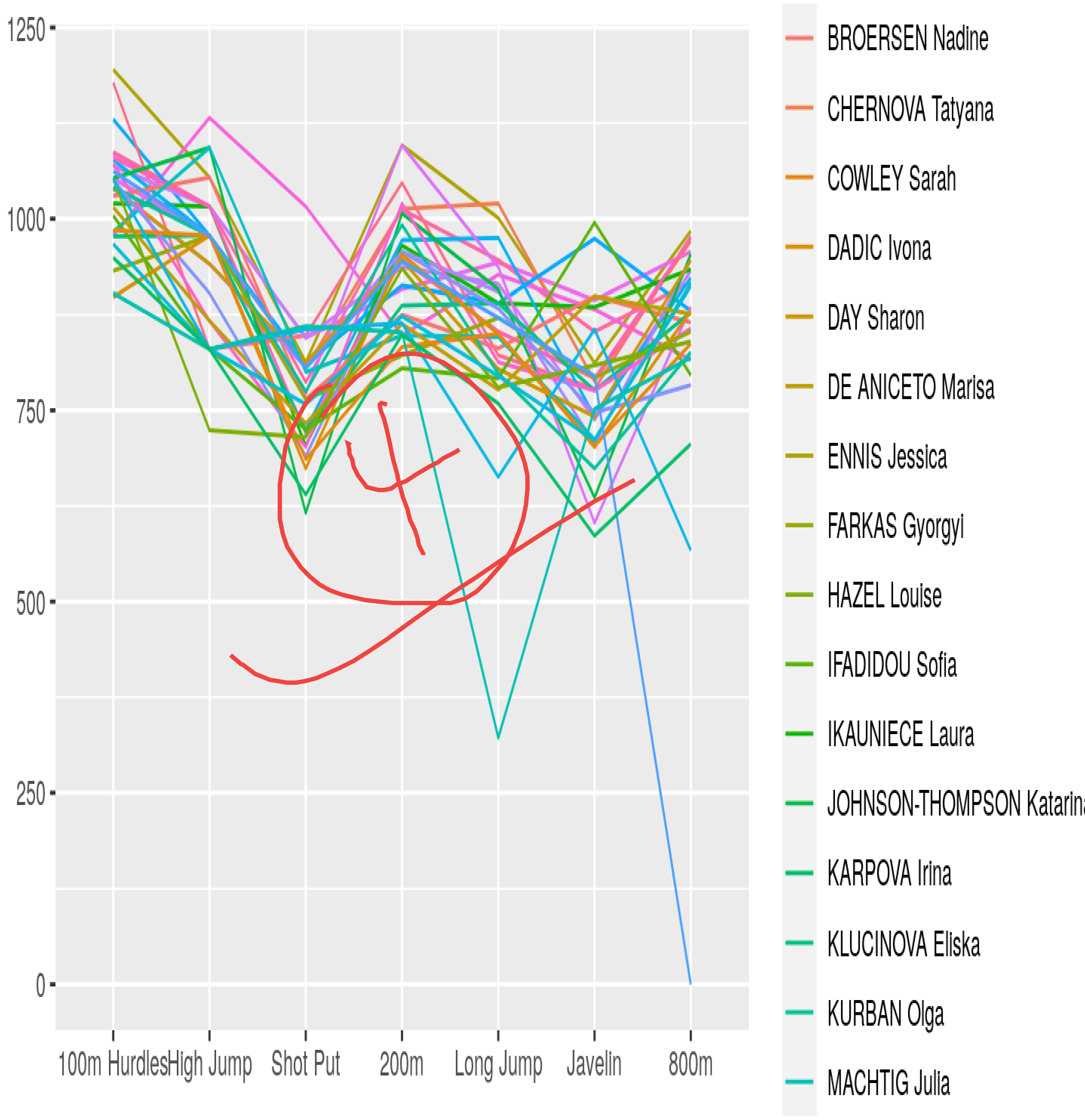
```
#AERTS Sara, DOBRYNSKA Nataliya, FOUNTAIN Hyleas, GRABUSTE Aiga, OSAZUWA Uhunoma & TYMINSKA Karolina all have events that haven't been recorded.
#AERTS Sara, TYMINSKA Karolina & OESER Jennifer partook in some events but got a score of 0.
#SKUJYTE Austra's Shot put is insanely high compared to her competitors.
#   FOUNTAIN Hyleas has a very low Javelin throw.
```

Question 3.

a. Draw a parallel coordinate plot of the heptathlon scores. Explain why scale=“globalminmax” is the appropriate choice. In this plot, can you see athletes with zero points in any event? Does this make sense vis-a-vis your findings from Question 2?

```
hepc <- hep
#for (i in 2:8) hepc[,i] <- rank(rowSums(hep[,2:i], na.rm=T))

ggparcoord(data = hepc, columns = 2:8, groupColumn = 1,scale = "globalminmax")+xlab("")+ylab("")
```

→ hepc is for question c



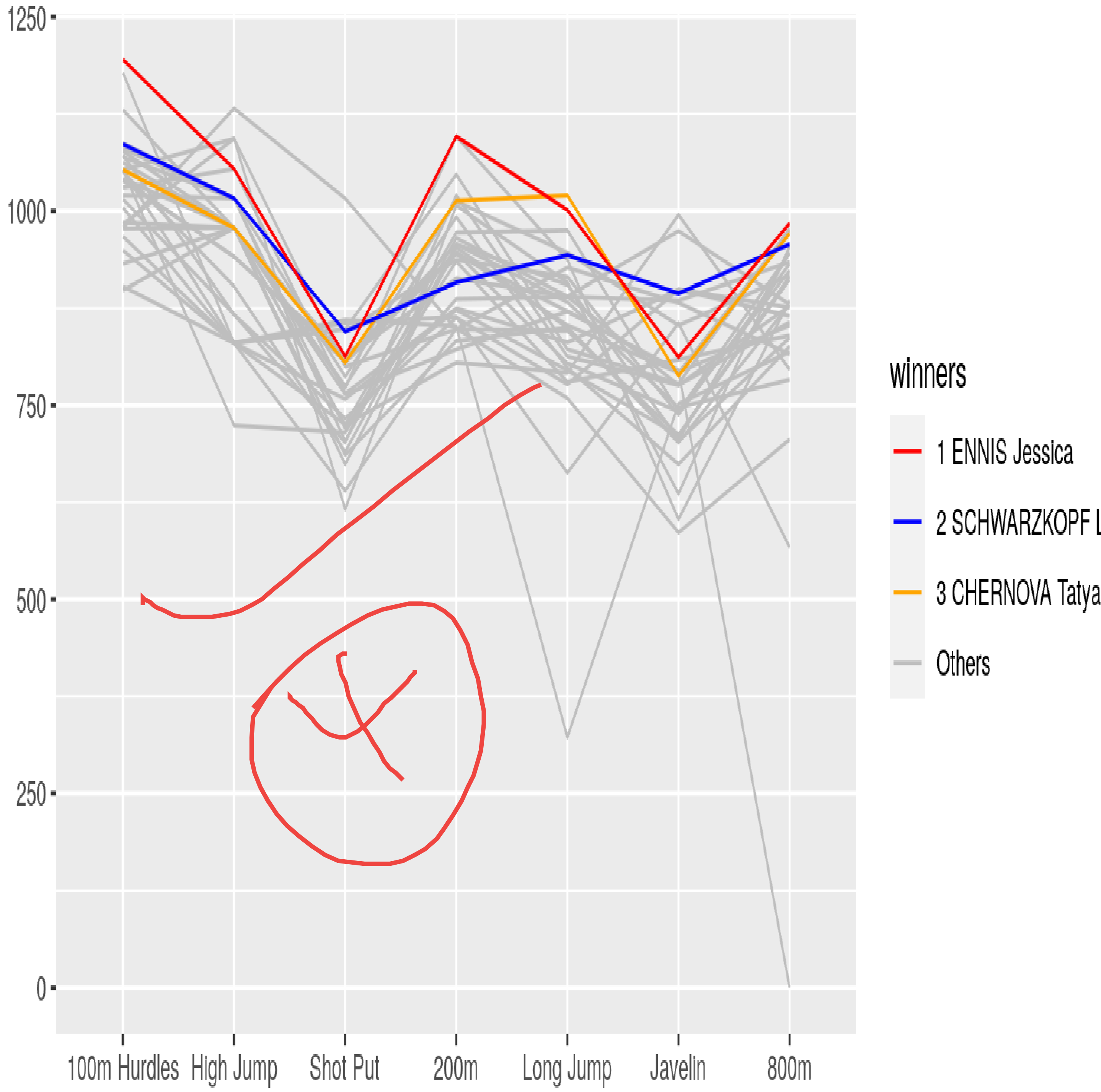
4

#Globalminmax scales each variable using the difference between its highest and lowest values across all observations. This is very useful at rooting the contestants with NA's. We can see that there is an athlete that has gotten zero points in an event. At first this doesn't make sense as there was multiple people that got a score of 0. However, a lot of these people didn't do all the events, thus for some events, their score was NA. Athletes with NA values are automatically removed.
#This is because all athletes will be considered, but once there's an NA, they are taken out of the system. Every point must have a value. Once it has NA, it doesn't automatically become 0. The following bit of code could be adding to allow the results of NA to be considered 0.

```
#for (i in 2:8) hepc[,i] <- rank(rowSums(hep[,2:i], na.rm=T))
```

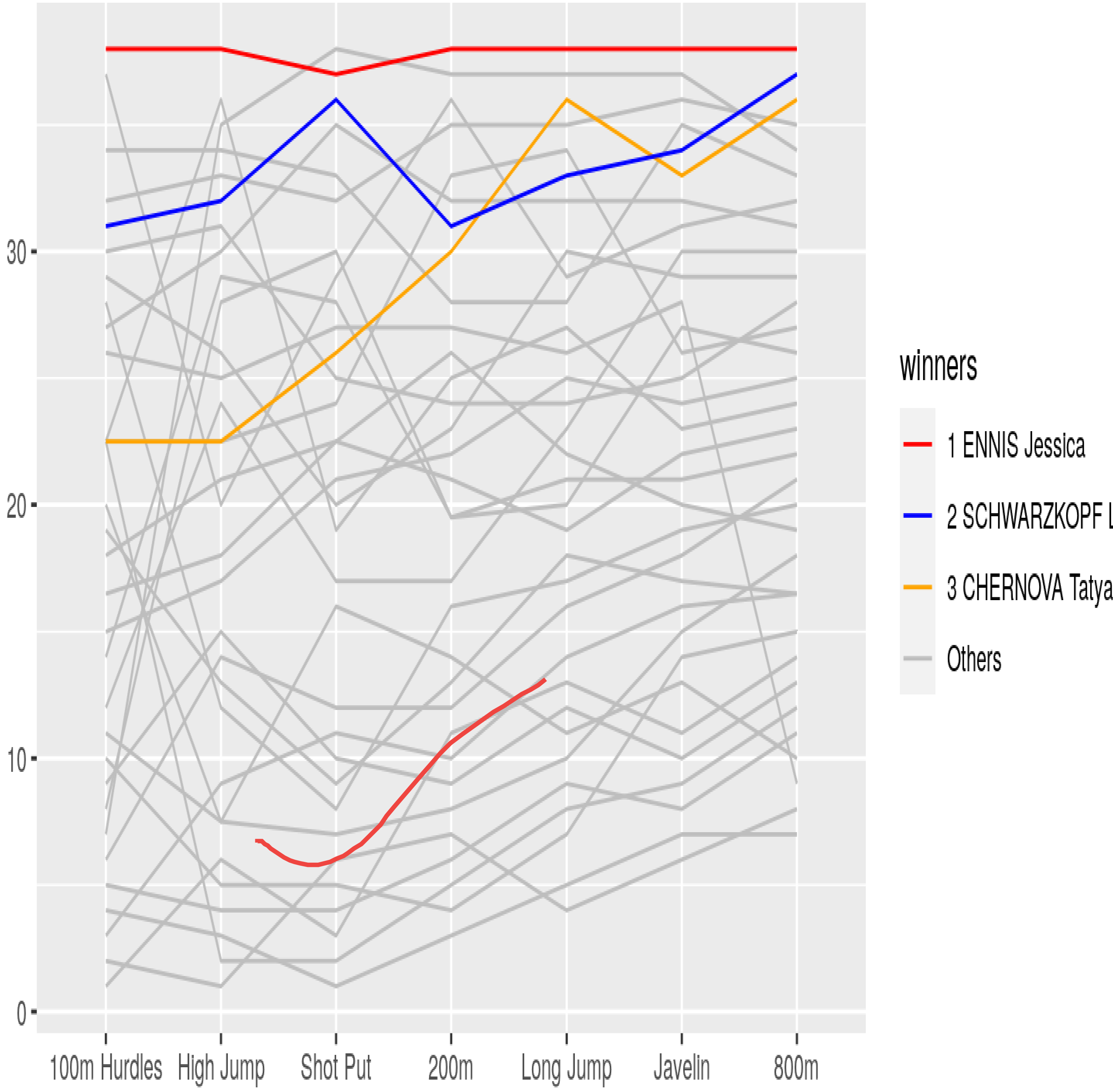
b. Make a group variable which has one level for each of the top three athletes, and a fourth level for the others. Use this to redraw the PCP with four different colours. Hint: use scale_color_manual to specify colours. Also reverse the order of the dataset rows so that the top performers are drawn last.

```
hep$winners <- paste(1:nrow(hep), hep$Athlete, sep=" ")
hep$winners[- (1:3)] <- "Others"

require(gridExtra)
plot1 <- hep %>% arrange(total) %>% na.omit() %>% ggparcoord( columns=2:8,
scale="globalminmax", groupColumn="winners" )+xlab("")+ylab("")+
  scale_color_manual(values=c("red","blue","orange","grey")) #First graph
plot2 <- hep %>% arrange(desc(total)) %>% na.omit() %>% ggparcoord( columns=2:8,
scale="globalminmax", groupColumn="winners" )+xlab("")+ylab("")+
  scale_color_manual(values=c("red","blue","orange","grey")) #Reversed Order
grid.arrange(plot1, plot2, ncol=2)
```



c. Calculate the ranks of the cumulative point totals using the given code. Make a plot which explores if Jessica Ennis and the other medal winners were ahead all of the way through the event.

```
hepc <- hep
for (i in 2:8) hepc[,i] <- rank(rowSums(hep[,2:i], na.rm=T))

hepc %>% arrange(total) %>% na.omit() %>% ggparcoord( columns=2:8, scale="globalminmax",
groupColumn = "winners") +xlab("")+ylab("")+
scale_color_manual(values=c("red","blue","orange","grey"))
```



#ENNIS was in the lead until the shotput, being overtaken by SKUJYTE. ENNIS got back to being first place during the 200m. The SCHWARZKOPF started in fifth, and went to third during the Shot Put. SCHWARZKOPF went to 7th place in the 200m but slowly and steadily rised to 2nd place in the 800m. CHERNOVA was tied 11th at the start of the 100m with 2 others that finished the Heptathlon. Chernova slowly rised to third place in th Long Jump, before going to sixth in Javelin and then back to third in the 800m.

#Question 4. Use vis_miss to explore the patterns of missings for the data. What are your findings? ggplotly will help find the rows with missing values.

```
#vis_miss(hep)
#v <- vis_miss(hep[,2:8])
#ggplotly(v)
```