Ljubljana Maraton 2014

A. Blejec

October 28, 2014

Contents


```
Warning: package 'Hmisc' was built under R version 3.0.3
Warning: package 'lattice' was built under R version 3.0.3
```

1 Introduction

At Ljubljana Marathon more than 15.000 runners compete on October 26, 2014. Results are available at http://vw-ljubljanskimaraton.si/en/result/16lm# and some preliminary checks showed nice statistical properties. So I decided to see how to get the data from the web page and get some insight into the results. Just to let you know - I am not a runner :)

2 Getting the results

First, we have to get the results form the web page. There are several categories and for now I will not go into the selection but rather use one of them. The names are nicely composed: `M42` is 42km marathon for all men (Moski in Slovenian), `Z42` is 42 marathon for all women (Zenske in Slovenian), `42MA` is Marathon - Men A group.

The URL for M42 is: http://www.pohod.si/lm/M42.asp. Let us read the web page.

```
> lfn <- "http://www.pohod.si/lm/Z42.asp"
> page <- readLines(lfn, enc = "UTF-8")
> length(page)
[1] 390
```

Total html page is now stored in page. First few lines

```
> head(page, 10)
```

```
 [1] "<html><head><META HTTP-EQUIV=\"Content-Type\" CONTENT=\"text/html; chars
 [2] "<title>Ljubljanski Maraton</title><style type=\"text/css\">"
 [3] "body      {font-family:\"Tahoma\"; font-face:\"Tahoma\"; font-size:9pt}"
 [4] "table     {font-family:\"Tahoma\"; font-face:\"Tahoma\"; font-size:9pt}"
 [5] "</style></head><body oncontextmenu=\"return false;\">"
 [6] "<p align=center style=\"font-size:18px; font-weight:bold; color:#73b508\
 [7] "<font style=\"font-size:12px; font-weight:bold; color:#73b508\"> Ljublja
 [8] "<p align=center><B>Z42 - Maraton - <U+008E>enske<br>Marathon - Women</b>
 [9] "<table cellpadding=1 align=center><tr><td style='background-color:white'
[10] "<TR class=r0><td><b>1</b></td><td>27</td><td>Janet Rono</td><td>1988</td
```

and the last few lines

```
> tail(page)
[1] "<TR class=r0><td><b></b></td><td>826</td><td>Ana Martinjak Ratej</td><td>
[2] "<TR class=r0><td><b></b></td><td>1147</td><td>Alenka Spori<U+009A>-loboda
[3] "<TR class=r0><td><b></b></td><td>1596</td><td>Renata Hojnik</td><td>1976<
[4] "<TR class=r0><td><b></b></td><td>932</td><td>Ula Kupec</td><td>1994</td><
[5] "<TR class=r0><td><b></b></td><td>110</td><td>Ana Marija Jeli<e8>i<e6></td
[6] "</table></td></tr></table><br><br></body></html>"
```

3 Using package XML

A better option is to use package **XML**.

```
> library(XML)
> doc <- xmlRoot(htmlTreeParse(lfn, enc = "UTF-8"))
> # doc
```

Node names

```
> table(names(doc))

body head
   1    1
```

All nodes have identical fields

```
> fields = xmlApply(doc, names)
> table(fields$body)

p
1
```

Extract table - this is a bit clumsy way!!

```
> tbl <- doc[["body"]]
> # names(xmlChildren(tbl))
> tbl <- doc[["body"]][["p"]][["font"]][["p"]][["font"]][["table"]][["tr"]][["
> xmlName(tbl)
[1] "table"
> xmlSize(tbl)
```

```
[1] 381
```

Get values and attributes

```
> tmp = xmlSApply(tbl, function(x) xmlSApply(x, xmlValue))
> tmp <- t(tmp)
> atr <- t(xmlSApply(tbl, function(x) xmlSApply(x, xmlAttrs)))
> head(atr)
   td   td   td   td   td   td   td   td   td   td   td   td   td
tr NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL
tr NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL
tr NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL
tr NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL
tr NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL
tr NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL
   td   td   td   td   td
tr NULL NULL NULL NULL NULL
tr NULL NULL NULL NULL NULL
tr NULL NULL NULL NULL NULL
tr NULL NULL NULL NULL NULL
tr NULL NULL NULL NULL NULL
tr NULL NULL NULL NULL NULL
> head(tmp)
   td  td       td                                td
tr "#" "St.Bib" "Ime in priimekName and Surname" "LRYoB"
tr "1" "27"     "Janet Rono"                      "1988"
tr "2" "30"     "Rehima Kedir Kedir"              "1985"
tr "3" "33"     "Agnes Mutane"                    "1986"
tr "4" "28"     "Caroline Chepkwony"              "1985"
tr "5" "31"     "Bornes Chepkirui"                "1988"
   td              td          td        td        td
tr "DravaCountry" "KlubClub"  "5km"     "10km"    "15km"
tr "KEN"            Character,0 "0:17:27" "0:34:47" "0:52:18"
tr "ETH"            Character,0 "0:17:28" "0:34:47" "0:52:15"
tr "KEN"            Character,0 "0:17:29" "0:34:54" "0:52:30"
tr "KEN"            Character,0 "0:17:28" "0:34:46" "0:52:14"
tr "KEN"            Character,0 "0:17:28" "0:34:47" "0:52:15"
   td        td        td        td        td        td
tr "20km"    "25km"    "30km"    "35km"    "40km"    "RezultatResult"
tr "1:09:33" "1:27:10" "1:44:44" "2:02:55" "2:21:09" "2:29:16"
tr "1:09:33" "1:27:11" "1:45:23" "2:04:25" "2:23:39" "2:32:29"
tr "1:10:02" "1:28:24" "1:46:55" "2:06:07" "2:24:55" "2:33:34"
tr "1:09:33" "1:27:11" "1:44:57" "2:04:26" "2:25:04" "2:35:00"
tr "1:09:33" "1:27:11" "1:44:57" "2:04:27" "2:25:46" "2:36:02"
   td     td  td
tr "Kat." "#" Character,0
tr "42ZA" "1" Character,0
tr "42ZA" "2" Character,0
tr "42ZA" "3" Character,0
tr "42ZA" "4" Character,0
tr "42ZA" "5" Character,0
```

Clean the table

```
> header <- tmp[1, ]
> data <- tmp[-1, ]
> dimnames(data) <- list(data[, 1], header)
> data <- data[dimnames(data)[[1]] != "", -1]
> data <- data.frame(data)
Warning: some row.names duplicated: 78,89,92,97,111,175,185,188,201,202,218,22
> head(data)
  St.Bib Ime.in.priimekName.and.Surname LRYoB DravaCountry KlubClub
1     27                     Janet Rono  1988            KEN
2     30             Rehima Kedir Kedir  1985            ETH
3     33                   Agnes Mutane  1986            KEN
4     28             Caroline Chepkwony  1985            KEN
5     31               Bornes Chepkirui  1988            KEN
6     29                 Edinah Kwambai  1980            KEN
     X5km    X10km    X15km    X20km    X25km    X30km    X35km    X40km
1 0:17:27 0:34:47 0:52:18 1:09:33 1:27:10 1:44:44 2:02:55 2:21:09
2 0:17:28 0:34:47 0:52:15 1:09:33 1:27:11 1:45:23 2:04:25 2:23:39
3 0:17:29 0:34:54 0:52:30 1:10:02 1:28:24 1:46:55 2:06:07 2:24:55
4 0:17:28 0:34:46 0:52:14 1:09:33 1:27:11 1:44:57 2:04:26 2:25:04
5 0:17:28 0:34:47 0:52:15 1:09:33 1:27:11 1:44:57 2:04:27 2:25:46
6 0:17:27 0:34:46 0:52:15 1:10:36 1:30:37 1:50:29 2:11:03 2:31:45
  RezultatResult Kat. X. character.0.
1        2:29:16 42ZA  1
2        2:32:29 42ZA  2
3        2:33:34 42ZA  3
4        2:35:00 42ZA  4
5        2:36:02 42ZA  5
6        2:40:42 42ZB  1
> data <- data[, -grep("character", names(data))]
> dim(data)
[1] 374  16

> result <- unlist(data[, grep("Result", names(data))])
> names(result) <- NULL
> head(result)
[1] "2:29:16" "2:32:29" "2:33:34" "2:35:00" "2:36:02" "2:40:42"
```

Function to convert strings to minutes

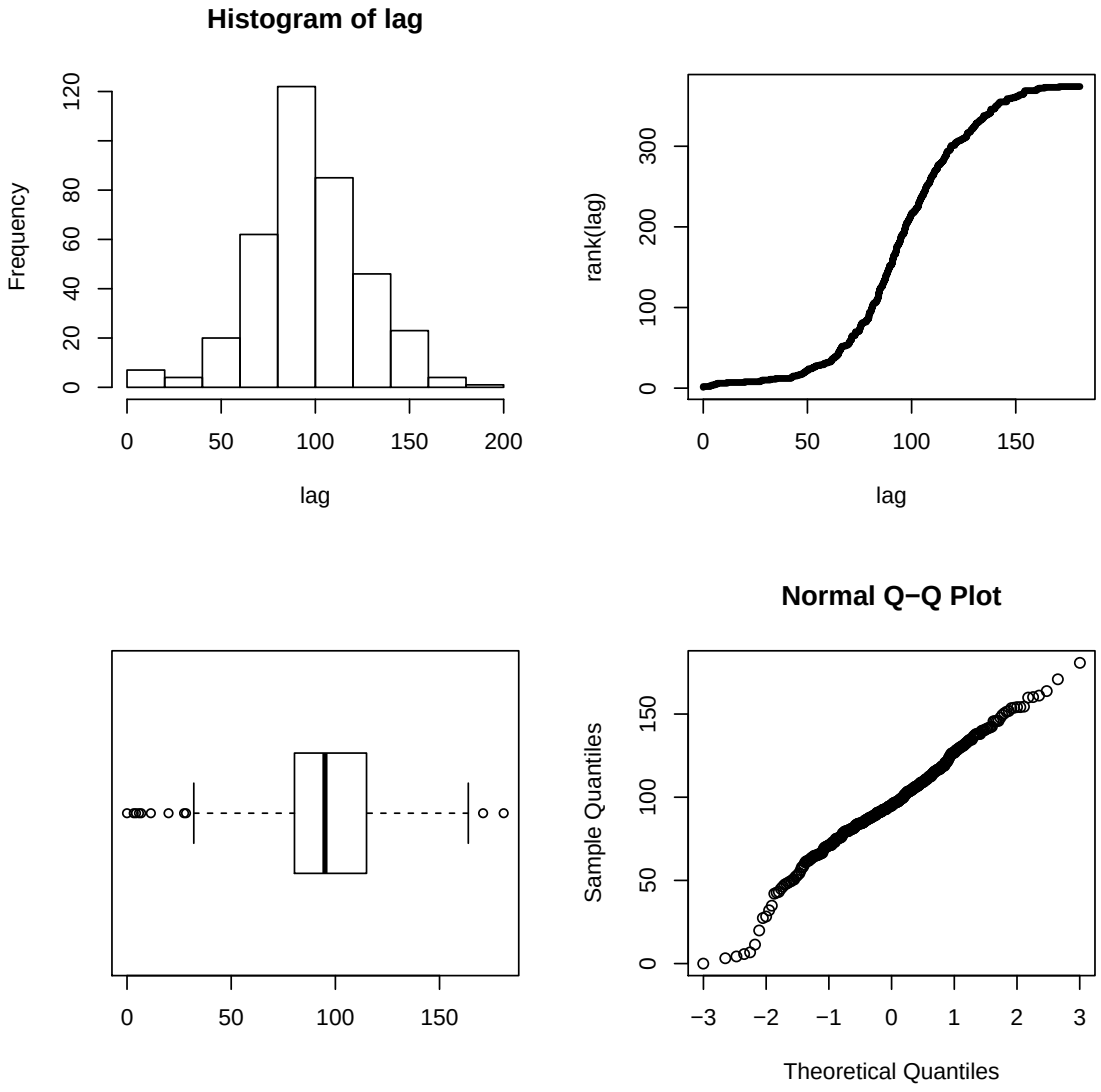
```
> as.mins <- function(x) {
+     pat <- "^([0-9]+):([0-9]+):([0-9]+)"
+     h <- as.numeric(gsub(pat, "\\1", x))
+     m <- as.numeric(gsub(pat, "\\2", x))
+     s <- as.numeric(gsub(pat, "\\3", x))
+     secs <- (h * 60 + m) * 60 + s
+     mins <- secs/60
+     return(mins)
+ }
```

Convert results

```
> result <- sapply(result, as.mins)
> lag <- result - result[1]
```

4 Distribution of results

```
> par(mfrow = c(2, 2))
>
> hist(lag)
> plot(lag, rank(lag), type = "S", lwd = 4)
> boxplot(lag, horizontal = TRUE)
> qqnorm(lag)
```



SessionInfo

Windows 7 x64 (build 7601) Service Pack 1

- R version 3.0.2 (2013-09-25), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=Slovenian_Slovenia.1250`, `LC_CTYPE=Slovenian_Slovenia.1250`, `LC_MONETARY=Slovenian_Slovenia.1250`, `LC_NUMERIC=C`, `LC_TIME=Slovenian_Slovenia.1250`
- Base packages: base, datasets, graphics, grDevices, grid, methods, splines, stats, utils
- Other packages: Formula 1.1-1, Hmisc 3.14-4, knitr 1.6, lattice 0.20-27, survival 2.37-7, XML 3.98-1.1
- Loaded via a namespace (and not attached): cluster 1.14.4, evaluate 0.5.5, formatR 0.10, latticeExtra 0.6-26, RColorBrewer 1.0-5, stringr 0.6.2, tools 3.0.2

Project path: `D:/_Y/R/CompStatistics`
Main file : `../doc/LjMarathon.Rnw`

View as vignette

Project files can be viewed by pasting this code to R console:

```
> projectName <-"CompStatistics";  mainFile <-"LjMarathon"


> commandArgs()
> library(tkWidgets)
> openPDF(file.path(dirname(getwd()),"doc",
> paste(mainFile,"PDF",sep=".")))
> viewVignette("viewVignette", projectName, #
> file.path("../doc",paste(mainFile,"Rnw",sep=".")))
> #
```