Elite Runners of the NYC Marathon

Valerie Molignano

NY Road Runners Club

- NYRR : Non-profit organization that helps and inspires people through running. Founded in 1958 by Fred Lebow. First marathon in 1970 with 127 entrants and 55 finishers. In 1976, takes race out of Central Park and into streets of NYC. First race to offer money.

- nyrr.org : Database of all runners from1970 to 2016. Men and Women run in separate races. 50,000 runners in 2016

- Entry is lottery unless you are an elite runner or raise $$$ for charity

- Personal marathon story

Marathon Data

- Format is different every few years. Early years - few variables. Later years lots of variables. 2001 - they started digitizing race times by ‘chip’ put on sneaker

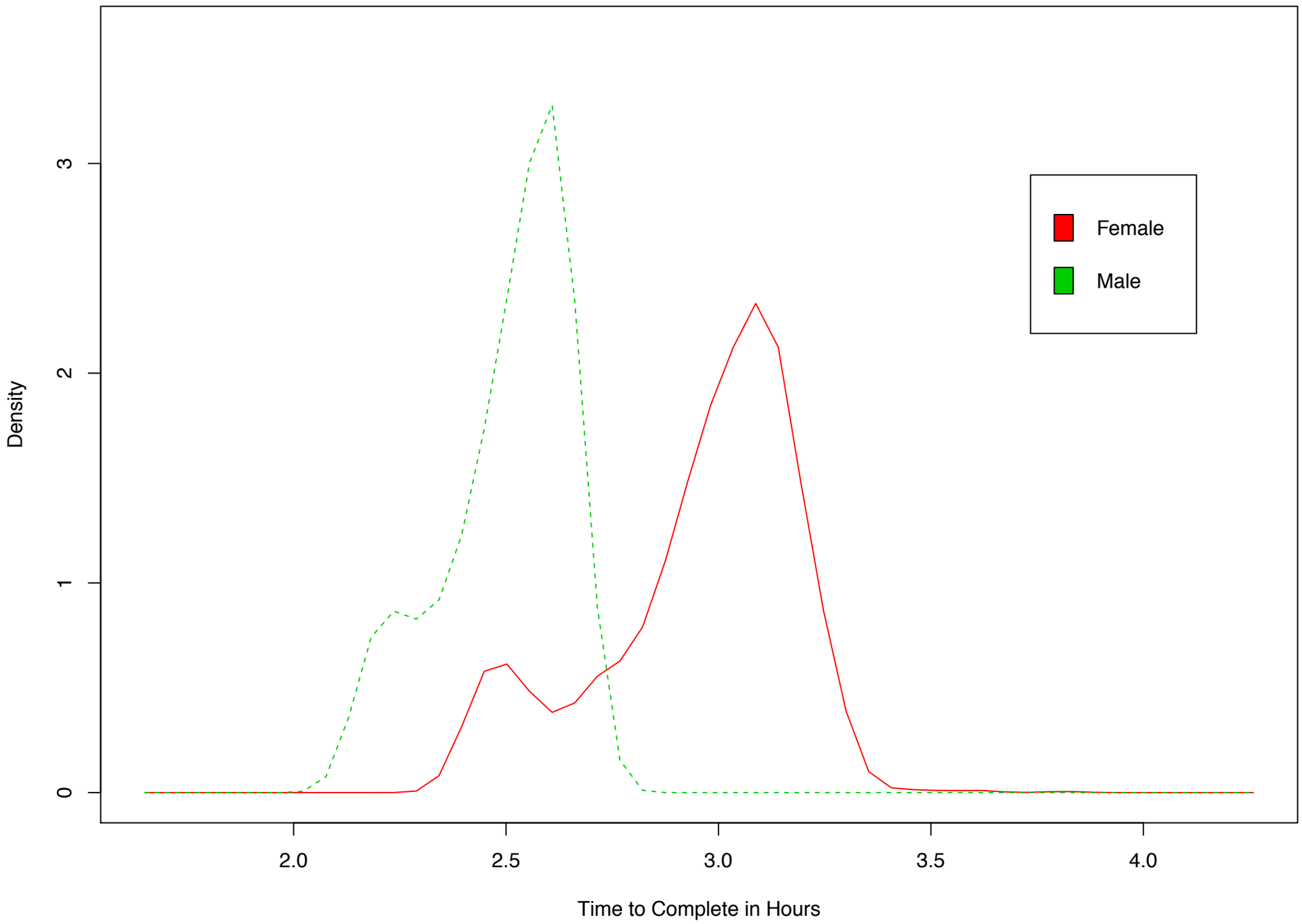
- 46 years of data - 40K to 50K entrants per year. Decided to scrape elite runners only, 1999-2015. Top 100 Men, Top 100 Women

- use of Beautiful Soup, python, mechanize, R

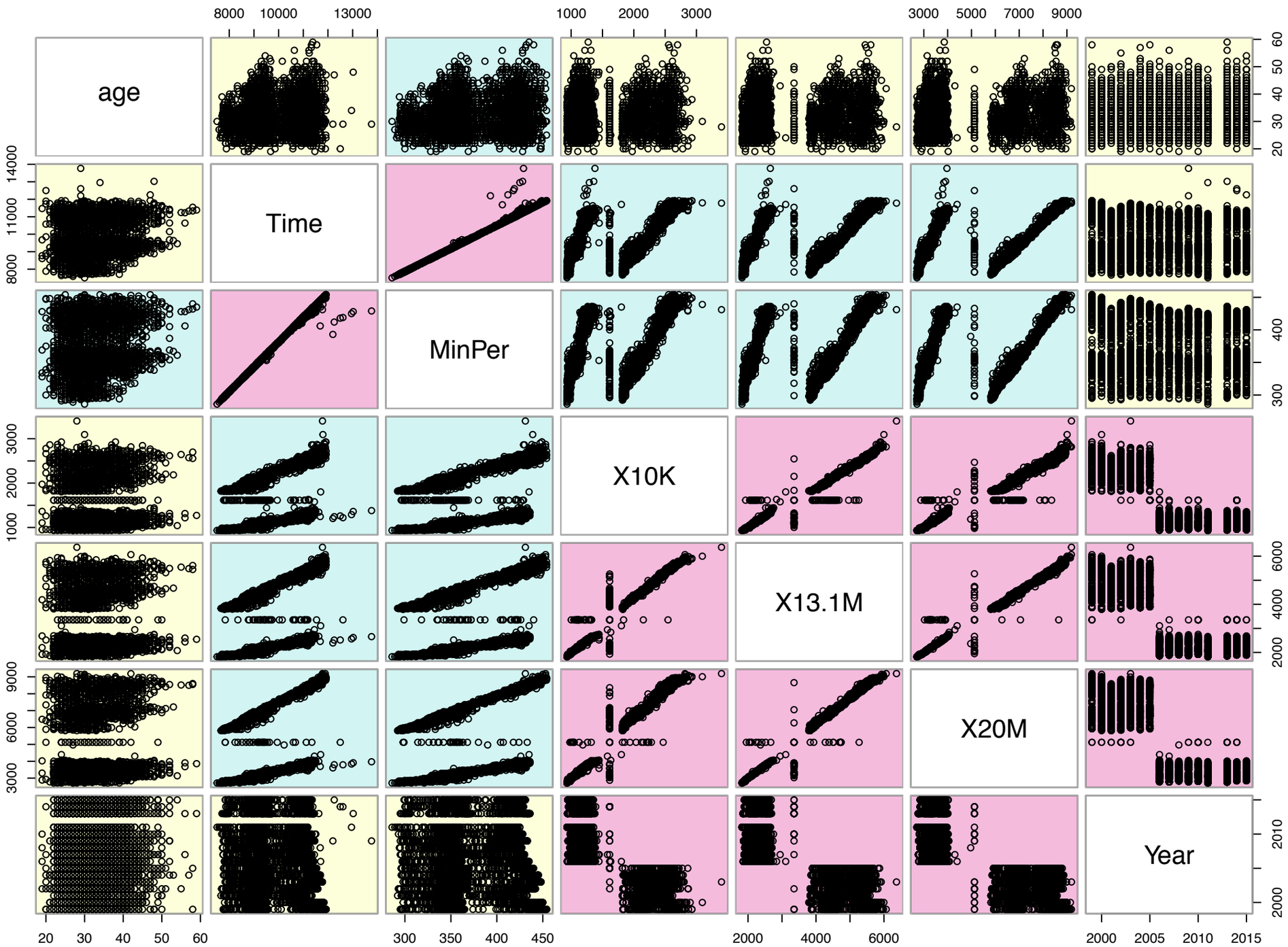
mechanize

- Stateful programmatic web browsing in Python, after Andy Lester's Perl module WWW::Mechanize.
- mechanize.Browser and mechanize.UserAgentBase implement the interface of urllib2.OpenerDirector, so:
- any URL can be opened, not just http:
- mechanize.UserAgentBase offers easy dynamic configuration of user-agent features like protocol, cookie, redirection and robots.txt handling, without having to make a new OpenerDirector each time, e.g. by calling build_opener().
- Easy HTML form filling.
- Convenient link parsing and following.
- Browser history (.back() and .reload() methods).
- The Referer HTTP header is added properly (optional).
- Automatic observance of robots.txt.
- Automatic handling of HTTP-Equiv and Refresh.

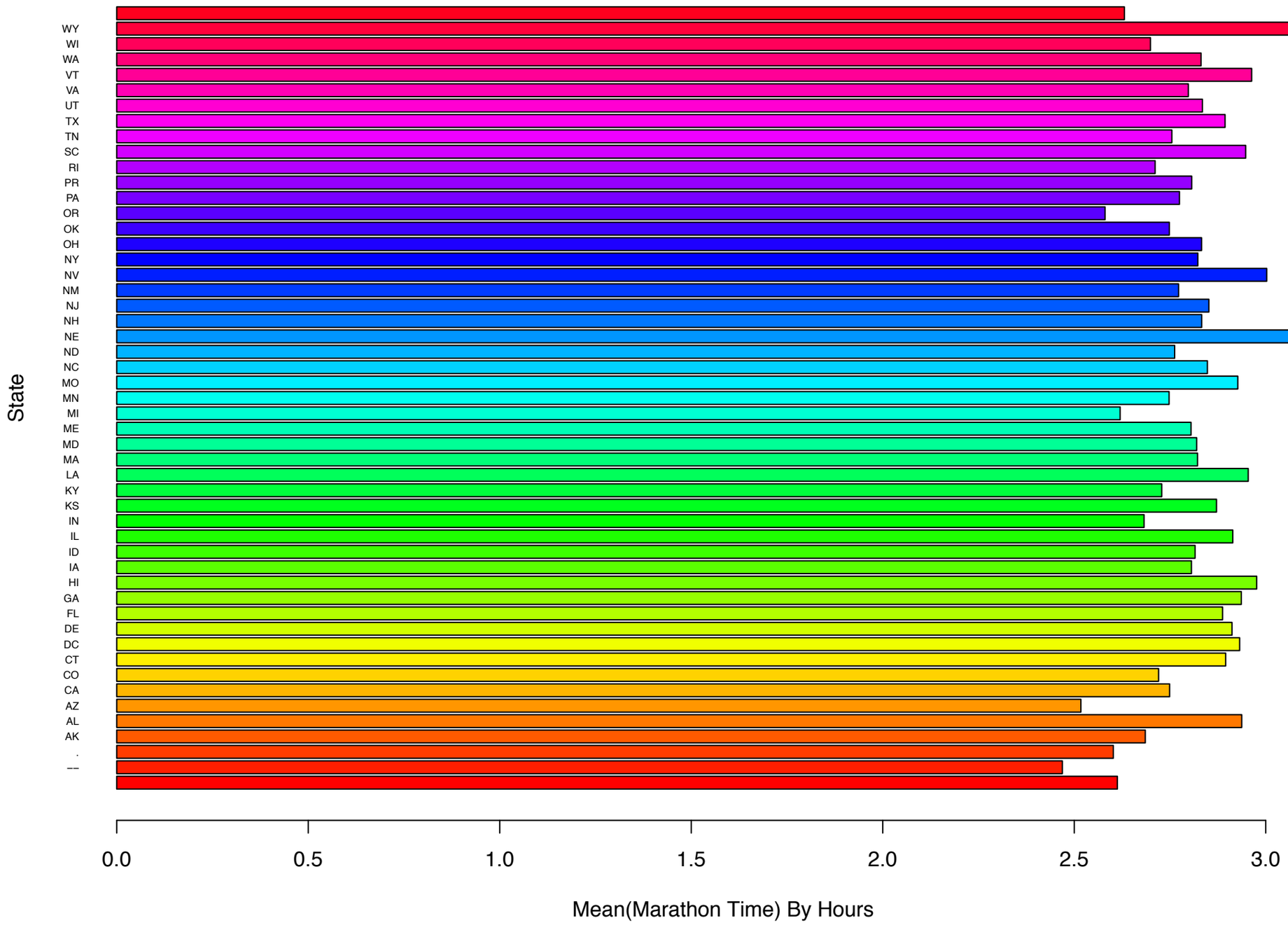
```
from bs4 import BeautifulSoup
import mechanize

url = "http://web2.nyrrc.org/cgi-bin/start.cgi/mar-programs/archive/
archive_search.html"

browser = mechanize.Browser()
browser.addheaders =
[    ('user-agent', 'Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.9.2.3) Gecko/
20100423 Ubuntu/10.04 (lucid) Firefox/3.6.3'),
    ('accept', 'text/html,application/xhtml+xml,application/xml;q=0.9,*/
*;q=0.8')]
browser.set_handle_robots(False)

# get years
browser.open(url)
browser.form = list(browser.forms())[0]
select = browser.form.controls[3]

#years = [item.name for item in select.get_items()]
#for now, hard code years

years = ['1999','2000','2001','2002','2003','2004',
        '2005','2006','2007','2008','2009','2010',
        '2011','2013','2014','2015']
gender = ['M', ‘F']

            browser.open(url)
            browser.form = list(browser.forms())[0]
            select = browser.form.controls[3]
```

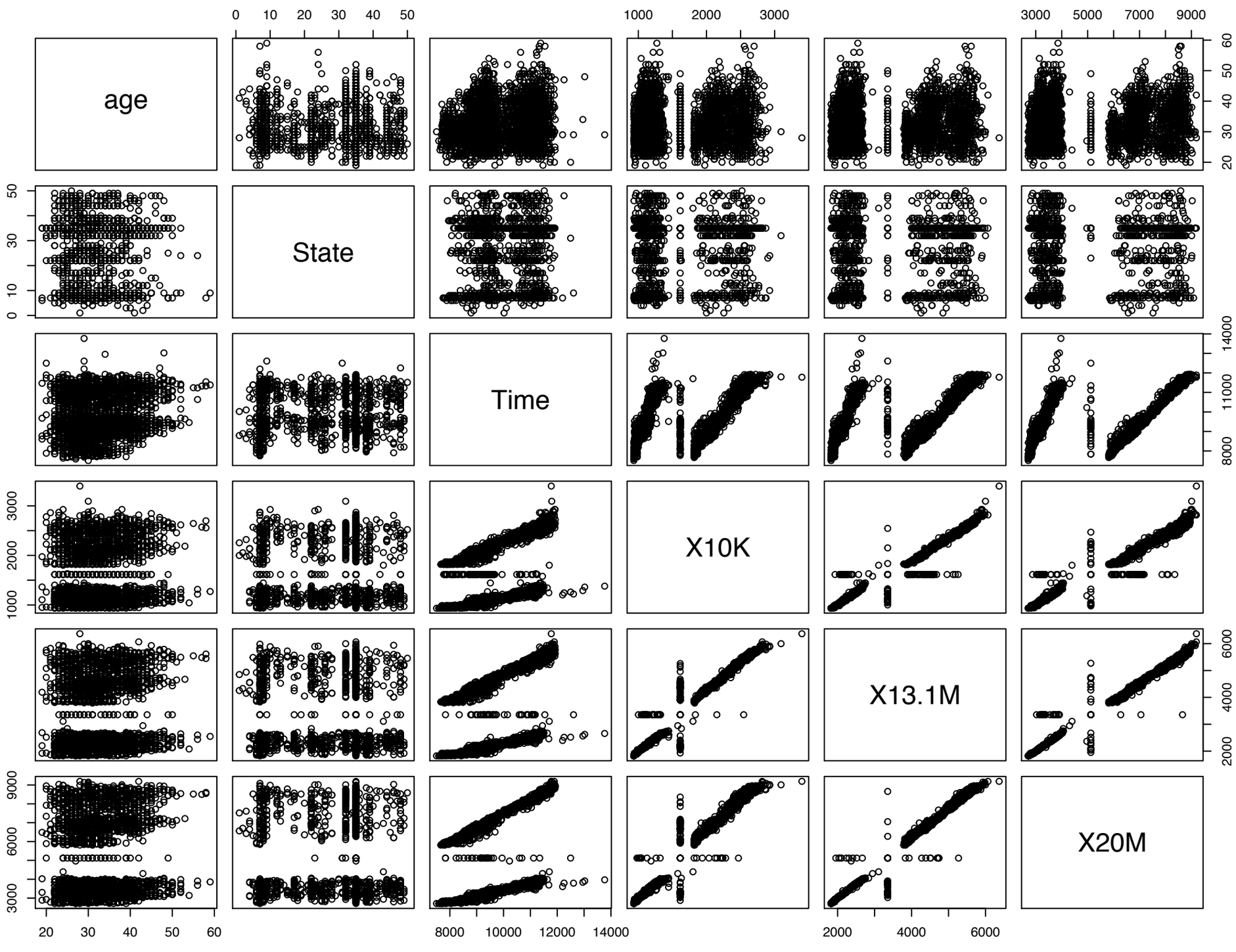

Men and Women Elite Runners
Density
0
1
2
3
2.0
2.5
3.0
3.5
4.0
Time to Complete in Hours
Female
Male

Variables Ordered and Colored by Correlation



States by Marathon Time




age
State
Time
X10K
X13.1M
X20M