COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

DSFBA: RStudio Projects and RMarkdown

Data Science for Business Analytics

Thibault Vatter

Department of Statistics, Columbia University

10/21/2020

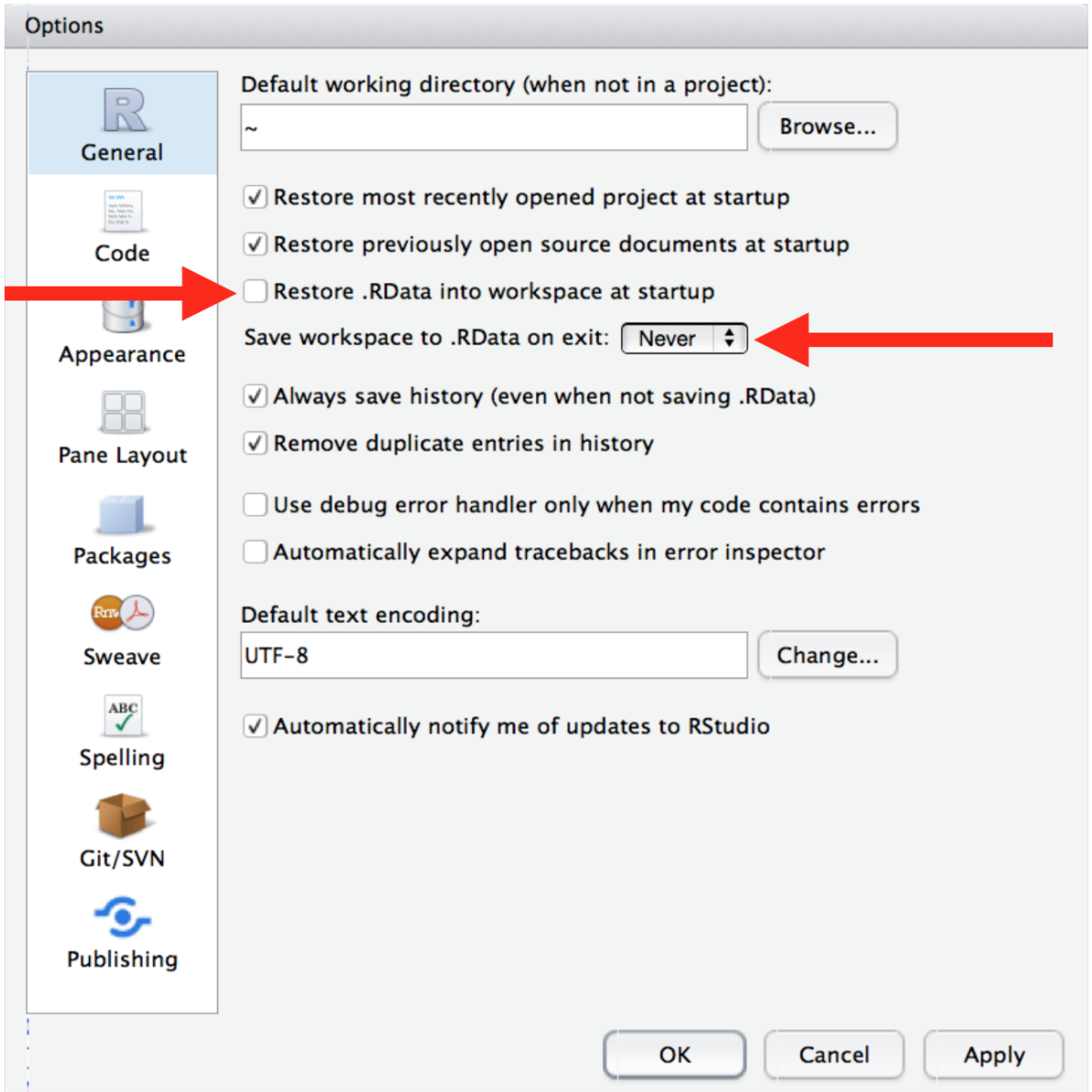
Outline

Two questions

- What about your analysis is "real"?
- Where does your analysis "live"?

What about your analysis is "real"?

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK



Where does your analysis live?

- The console
- R scripts
- RStudio projects: make it straightforward to divide your work into multiple contexts, each with their own working directory, workspace, history, and source documents.

DEMO!

The workflow with RStudio projects

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

- Create an RStudio project for each data analysis project.
- Keep data files there.
- Keep scripts there.
- Save your outputs (plots and cleaned data) there.
- Only ever use relative paths (e.g., with `here::here`), not absolute paths.

Everything you need is in one place, and cleanly separated from all the other projects that you are working on.

Outline

R Markdown

- The two components:
  - Literate programming
  - Markdown

Literate programming

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

- Motivation: helps peers understand and replicate your results, find errors and suggest enhancements
- Introduced by Donald Knuth

> *"a program is given as an explanation of the program logic in a natural language, such as English, interspersed with snippets of macros and traditional source code, from which a compilable source code can be generated [. . . It] represents a move away from writing programs in the manner and order imposed by the computer, and instead enables programmers to develop programs in the order demanded by the logic and flow of their thoughts."*
>
> — *Wikipedia*

What does this R code do?

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

```
data(women)
plot(women)
fit <- lm(weight ~ height, data = women)
abline(fit)
```

And this one?


```r
# Analysis of the 'women' dataset in R
data(women) # Load the data
attach(women) # Attach data to path
plot(weight ~ height) # Make a scatter plot
fit <- lm(weight ~ height) # Fit linear model
abline(fit) # Add a line of best fit to the plot
```

Two competing "views"

"Real programmers don't comment their code. If it was hard to write, it should be hard to understand."

— unknown

"If you can't write clearly, you probably don't think nearly as well as you think you do."

— Kurt Vonnegut

Can't we do better?

The **World Almanac and Book of Facts** (1975) includes a dataset of heights (in) and weights (lbs) of 15 American women aged 30–39. It is built into R:

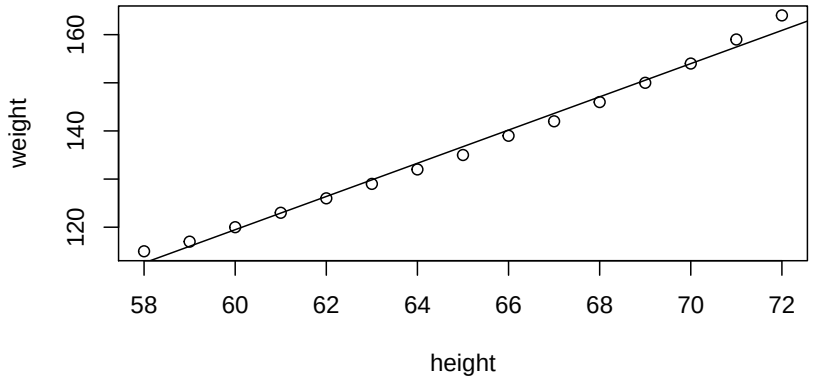
```
data(women)
```

Weight appears to increase (almost) linearly with height: every inch in height adds approximately 3.45 lbs. This was determined by fitting a simple linear regression model of weight against height:

```
fit <- lm(weight ~ height, data = women)
```

The resulting least-squares regression line can be drawn on a scatter plot of height against weight, where the models seems appropriate:

```
plot(weight ~ height, data = women)
abline(fit)
```



How did I do that?

The __World Almanac and Book of Facts__ (1975) includes a dataset of heights (in) and weights (lbs) of 15 American women aged 30-39.
It is built into R:

```{r}
data(women)
```

Weight appears to increase (almost) linearly with height: every inch in height adds approximately 3.45 lbs. This was determined by fitting a simple linear regression model of weight against height:

```{r}
fit <- lm(weight ~ height, data = women)
```

The resulting least-squares regression line can be drawn on a scatter plot of height against weight, where the models seems appropriate:

```{r}
plot(weight ~ height, data = women)
abline(fit)
```

A lightweight markup language

- **Markup:**
  - A system for annotating a document in a way that is syntactically distinguishable from the text
  - E.g., LaTeX and HyperText Markup Language (HTML)
- **Lightweight:**
  - A markup language with simple, unobtrusive syntax
  - E.g., Markdown and R markdown

Markup vs lightweight markup

Here is some text:

- in *italics*,
- in **boldface**.

In Latex:

```
Here is some text:
\begin{itemize}
\item in \textit{italics},
\item in \textbf{boldface}.
\end{itemize}
```

In Markdown:

```
Here is some text:
* in *italics*,
* in **boldface**.
```

What is R markdown?

A markdown-based literate programming system

DEMO!

Useful resources

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

- Essential: R Markdown cheat sheet
- RStudio's R markdown website
  - Tutorial (to get you started)
  - Output formats (e.g., HTML, Word documents, PDFs, presentations, etc.)
- stuff written by Yihui
  - knitr and especially [its options page](https://yihui.name/knitr/options/
  - bookdown to write technical reports
  - blogdown to even build your own website