R-eproducible workflows

1-day workshop
Afternoon practical session

Brendan Palmer,

Statistics & Data Analysis Unit,

Clinical Research Facility - Cork

University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

CRF-C
HRB Clinical Research Facility Cork

HRB
Health Research Board

A: Is too much choice good or bad?

Toothpaste & Jam: The Psychology of Choice

Something as simple as buying toothpaste can be overwhelming. Do you want the anti-tartar kind or the cavity-busting option? Sensitive enamel protection or the one with whitening? Fluoride, non-fluoride? Then there's flavor: crystal mint, intense mint, fresh mint or sparkling mint – and that's *just* mint.

Inconsistent function names, inconsistent syntax

- R is a very versatile language
  - Sometimes it can be too versatile
  - Do you want to use……
    - Names or colnames
    - row.names or rownames
    - rowSums or rowsum
    - Sys.time, system.time

  - Is it written as……
    - newobject or new.Object
    - x = 5 or x <- 5
    - mapping=aes(x,y) or mapping = aes(x, y)

Variable selection

```
summary(starwars$name)

summary(starwars$"name")

summary(starwars["name"])

summary(starwars[,"name"])

summary(starwars[1])

summary(starwars[,1])

summary(starwars[[1]])
```

- Open the script 01_pm_too_much.choice.R

Writing clearer code

- Annotation

- Object names
  - should use only lowercase letters, numbers, and "_"

- Spacing
  - Put a space before and after =
  - Put a space after a ,
  - Operators should be surrounded by spaces e.g. ==, <-, +

- For a more complete list visit
  - http://style.tidyverse.org/syntax.html

- Open the script 02_pm_good_habits.R

Everything in its right place

- Open the script 03_pm_clean_data.R

B: R Markdown

R Markdown

YAML header

```
---
title: "Diamond sizes"
date: 2016-08-25
output: html_document
---
```

Chunks of code

````
```{r setup, include = FALSE}
library(ggplot2)
library(dplyr)
smaller <- diamonds %>%
filter(carat <= 2.5)
```
````

Plain text with integrated outputs from R

```
We have data about `r nrow(diamonds)`
diamonds. Only
`r nrow(diamonds) - nrow(smaller)` are
larger than
2.5 carats. The distribution of the
remainder is shown below:
```

Chunks of code

````
```{r, echo = FALSE}
smaller %>%
ggplot(aes(carat)) +
geom_freqpoly(binwidth = 0.01)
```
````

Introduction to R Markdown

- We're now going to look at a R Markdown file that provides some of the tips and tricks you'll need yourselves

  - Code chunks

  - Formatting

  - Tables

  - Figures etc.

- Open the R Markdown file 01_RMarkdown_intro.rmd

C: R Markdown workflows

Tips:
- Ensure each notebook has a descriptive title
- If you reach a research dead end, don't delete it
    - Write a note about it. It may be useful later
- Don't modify error in the data file directly
    - Correct it within the cleaning script
- At the end of each day run a clean knit of the note book
  - If there's an error message, correct it while its still
    fresh in your mind
- If you want your code to be reproducible in the long run,
  you'll need to keep a rigorous track of the package versions
  - Consider using the **packrat** package to help with this

- For an deeper dive into R Markdown visit
  - https://bookdown.org/yihui/rmarkdown/

R Markdown workflows

- Open the R project in the example_project folder

  - Open the R Markdown file 02_all_together_now.rmd

  - Formatting

  - Tables

  - Figures etc.

- Open the R Markdown file all_together_now.rmd

Fin