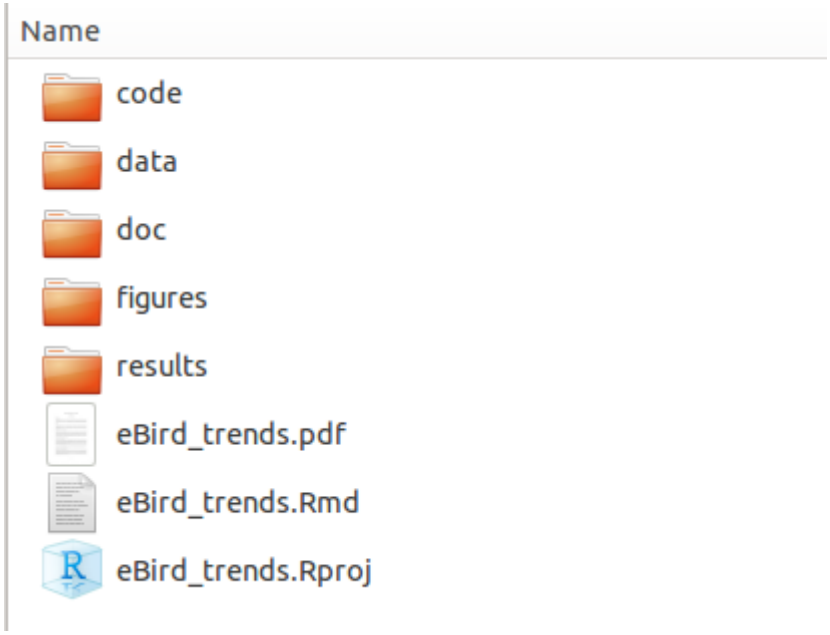
Reproducible research with R

Laura Graham
BES Macroecology, Oxford

Setting up a workflow

- R projects
- Git & GitHub
- Folder management system



- Annotated code
- Use relative file paths

Loading data

- readr for flat files (.csv, .txt)
- readxl for Excel spreadsheets
- RODBC for many types of databases
- RPostgreSQL for PostgreSQL databases
- googlesheets for interacting with Google sheets
- … and many, many more

Tidy data

Tidy datasets are all alike but every messy datset is messy in its own way

- Tidy data:
  - Observations in rows
  - Variables in columns
  - Each type of observational unit is a table
- Messy data:
  - Column headers are values, not variable names
  - Multiple variables stored in one column
  - Variables stored in both rows and columns
  - Multiple observational unit types in the same table
  - Single observational unit in multiple tables

Data tidying and manipulation tools

- tidyr: `gather()`, `separate()`, `spread()`
- plyr: split-apply-combine, `ldply()`
- dplyr: `group_by()`, `filter()`, `summarise()`, `mutate()`, `arrange()`

Messy output

```
data("mtcars")
lmfit <- lm(mpg ~ wt, mtcars)
summary(lmfit)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Tidy output tools

- broom: `tidy()`, `glance()`, `augment()`

```
library(broom)
tidy(lmfit)
```

```
##          term  estimate std.error statistic      p.value
## 1 (Intercept) 37.285126  1.877627 19.857575 8.241799e-19
## 2          wt -5.344472  0.559101 -9.559044 1.293959e-10
```

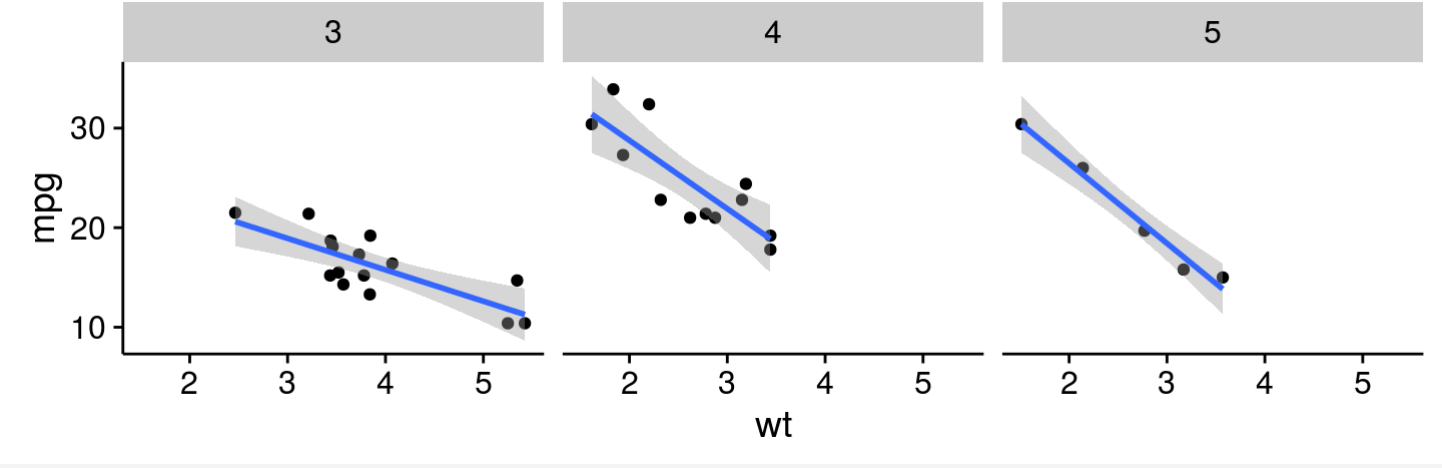
```
glance(lmfit)
```

```
##   r.squared adj.r.squared    sigma statistic      p.value df    logLik
## 1 0.7528328     0.7445939 3.045882  91.37533 1.293959e-10  2 -80.01471
##        AIC      BIC deviance df.residual
## 1 166.0294 170.4266 278.3219          30
```

Plotting

- ggplot2 is designed to work with tidy data formats

```
library(ggplot2)
library(cowplot)
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point() +
  facet_wrap(~gear) +
  geom_smooth(method = "lm")
```



RMarkdown

- Markdown + knitR + pandoc
- Outputs to PDF, Word, HTML, notebooks …
- Contains R code chunks
- Dynamic log of analysis
- Reasonably simple syntax (and lots of online resources)

Additional Resources

- RStudio cheatsheets: includes ggplot2, RMarkdown, dplyr, tidyr and more
- Swirl provides tutorials for tidyr, dplyr and much more directly in the R console
- Python pandas comparison with R
- Software Carpentry lessons
- Reproducible Research on Coursera: taught by Roger Peng, Jeff Leek and Brian Caffo at Johns Hopkins University
- These slides: https://github.com/laurajanegraham/BES-Macroecology-2016