Module 1: Big Picture Ideas

Rebecca C. Steorts (slide and course adaptation from Maria Tackett)

Announcements

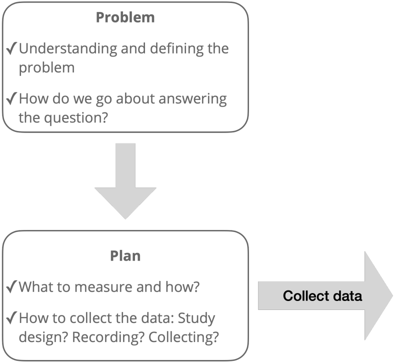
- Resources for extra R review
  - Learn R: An interactive introduction to data analysis R (focus on Chapters 4 - 6)
- Readings for next week will be posted later this week

Questions from last class?

Topics

- Data analysis life cycle
- Reproducible data analysis
- Analyzing multivariable relationships

Data analysis life cycle



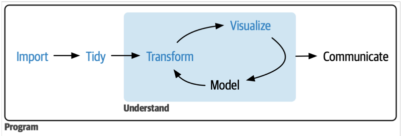


Figure 1: Source: *R for Data Science* with additions from *The Art of Statistics: How to Learn from Data*.

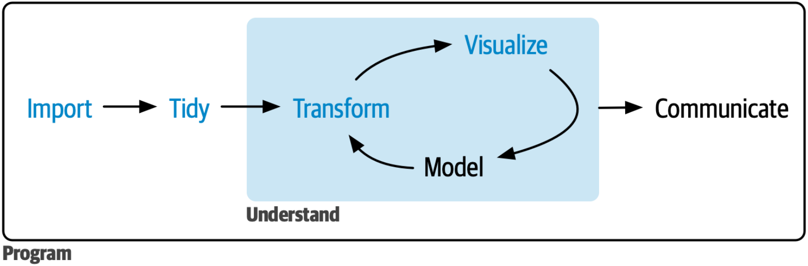


Figure 2: Source:*R for Data Science*

Reproducibility

Reproducibility checklist

What does it mean for an analysis to be reproducible?

Reproducibility checklist

What does it mean for an analysis to be reproducible?

Near term goals:

- Can the tables and figures be exactly reproduced from the code and data?
- Does the code actually do what you think it does?
- In addition to what was done, is it clear *why* it was done?

Reproducibility checklist

What does it mean for an analysis to be reproducible?

Near term goals:

- Can the tables and figures be exactly reproduced from the code and data?
- Does the code actually do what you think it does?
- In addition to what was done, is it clear *why* it was done?

Long term goals:

- Can the code be used for other data?
- Can you extend the code to do other things?

Why is reproducibility important?

- Results produced are more reliable and trustworthy [@ostblom2022]
- Facilitates more effective collaboration [@ostblom2022]
- Contributing to science, which builds and organizes knowledge in terms of testable hypotheses [@alexander2023]
- Possible to identify and correct errors or biases in the analysis process [@alexander2023]

When things go wrong

Reproducibility error	Consequence	Source(s)
Limitations in Excel data formats	Loss of 16,000 COVID case records in the UK	(Kelion 2020)
Automatic formatting in Excel	Important genes disregarded in scientific studies	(Ziemann, Eren, and El-Osta 2016)
Deletion of a cell caused rows to shift	Mix-up of which patient group received the treatment	(Wallensteen et al. 2018)
Using binary instead of explanatory	Mix-up of the intervention with the	(Aboumatar and Wise 2019)

Toolkit

- **Scriptability** $\rightarrow$ R
- **Literate programming** (code, narrative, output in one place) $\rightarrow$ Rmarkdown
- **Version control** $\rightarrow$ Git / GitHub

R and RStudio

- R is a statistical programming language
- RStudio is a convenient interface for R (an integrated development environment, IDE)

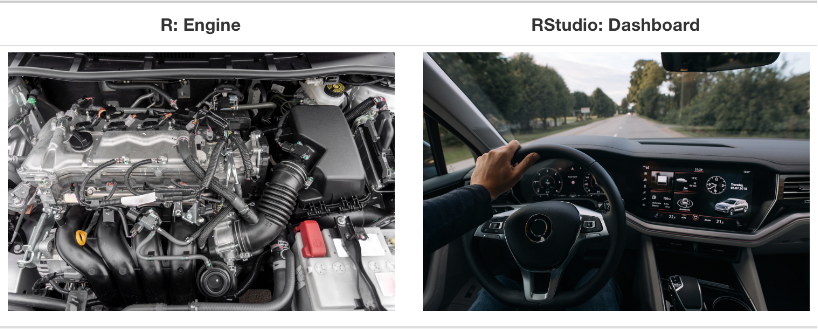
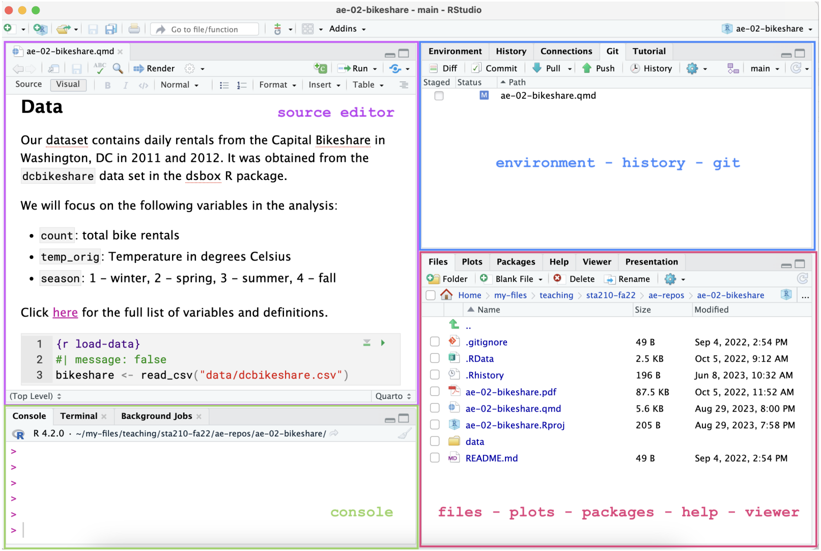


Figure 3: Source: Statistical Inference via Data Science

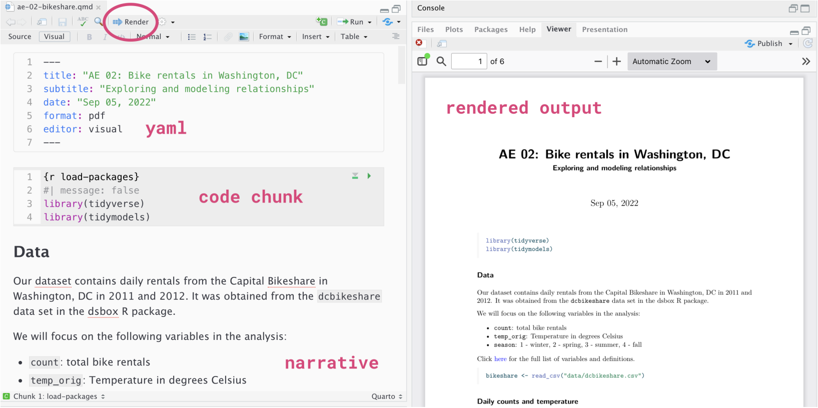
RStudio IDE


ae-02-bikeshare - main - RStudio
ae-02-bikeshare.qmd
Render
Run
Source
Visual
Normal
Format
Insert
Table
Data
source editor
Our dataset contains daily rentals from the Capital Bikeshare in Washington, DC in 2011 and 2012. It was obtained from the dcbikeshare data set in the dsbox R package.
We will focus on the following variables in the analysis:
• count: total bike rentals
• temp_orig: Temperature in degrees Celsius
• season: 1 – winter, 2 – spring, 3 – summer, 4 – fall
Click here for the full list of variables and definitions.
{r load-data}
#| message: false
bikeshare <- read_csv("data/dcbikeshare.csv")
(Top Level)
Quarto
Console
Terminal
Background Jobs
R 4.2.0 · ~/my-files/teaching/sta210-fa22/ae-repos/ae-02-bikeshare/
console
Environment
History
Connections
Git
Tutorial
Diff
Commit
Pull
Push
History
main
Staged
Status
Path
ae-02-bikeshare.qmd
environment - history - git
Files
Plots
Packages
Help
Viewer
Presentation
Folder
Blank File
Delete
Rename
Home › my-files › teaching › sta210-fa22 › ae-repos › ae-02-bikeshare
Name
Size
Modified
.gitignore 49 B Sep 4, 2022, 2:54 PM
.RData 2.5 KB Oct 5, 2022, 9:12 AM
.Rhistory 196 B Jun 8, 2023, 10:32 AM
ae-02-bikeshare.pdf 87.5 KB Oct 5, 2022, 11:52 AM
ae-02-bikeshare.qmd 5.6 KB Aug 29, 2023, 8:00 PM
ae-02-bikeshare.Rproj 205 B Aug 29, 2023, 7:58 PM
data
README.md 49 B Sep 4, 2022, 2:54 PM
files - plots - packages - help - viewer

Rmarkdown

- Fully reproducible reports – the analysis is run from the beginning each time you render
- Code goes in chunks and narrative goes outside of chunks
- Visual editor to make document editing experience similar to a word processor (Google docs, Word, Pages, etc.)

Rmarkdown/Quarto

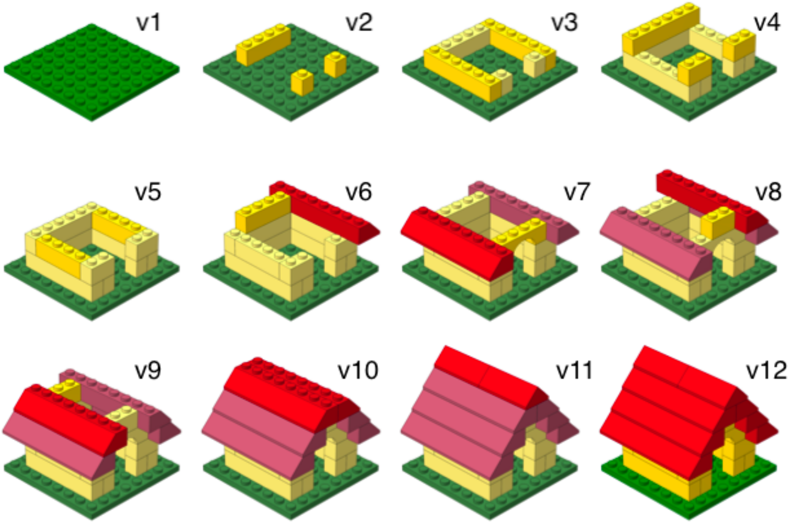


How will we use Rmarkdown?

- Every assignment is written in Rmarkdown document
- You'll have a template in Rmarkdown to start with
- If you have used Quarto before, think about what the similarities/differences are between Rmarkdown/Quarto.

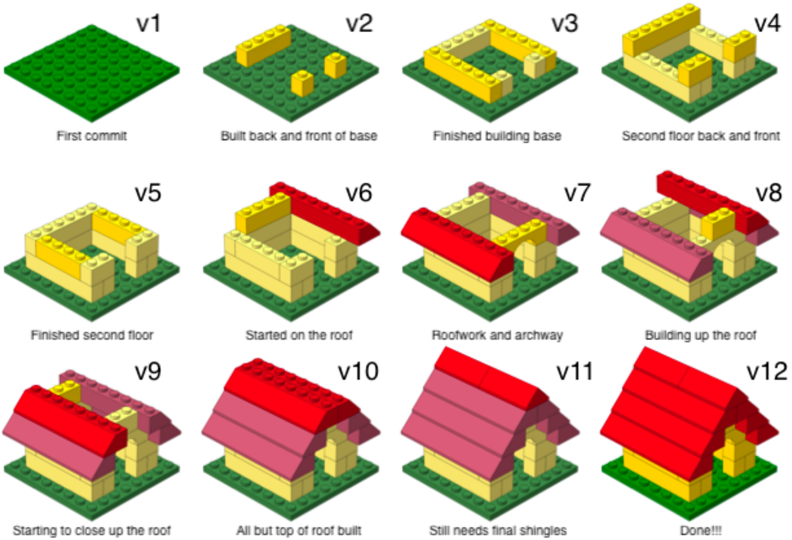
Version control with git and GitHub

What is versioning?



What is versioning?

with human readable messages



Why do we need version control?

Provides a clear record of how the analysis methods evolved. This makes analysis auditable and thus more trustworthy and reliable. [@ostblom2022]

git and GitHub

- **git** is a version control system – like “Track Changes” features from Microsoft Word.
- **GitHub** is the home for your git-based projects on the internet (like DropBox but much better).
- There are a lot of git commands and very few people know them all. 99% of the time you will use git to add, commit, push, and pull.

git and GitHub

- The entire materials for the semester can be found on github if you'd like to dig into them further at

https://github.com/resteorts/generalized-linear-models

Multivariable relationships

Carbohydrates in Starbucks food

- Starbucks often displays the total calories in their food items but not the other nutritional information.
- Carbohydrates are a body's main fuel source. The Dietary Guidelines for America recommend that carbohydrates make up 45% to 65% of total daily calories.[1]
- Our goal is to understand the relationship between the amount of carbohydrates and calories in Starbucks food items. We'd also like to assess if the relationship differs based on the type of food item (bakery, salad, sandwich, etc.)

[1] Source: Mayo Clinic

Starbucks data

- **Observations**: 77 Starbucks food items
- **Variables:**
  - `carb`: Total carbohydrates (in grams)
  - `calories`: Total calories
  - `bakery`: 1: bakery food item, 0: other food type

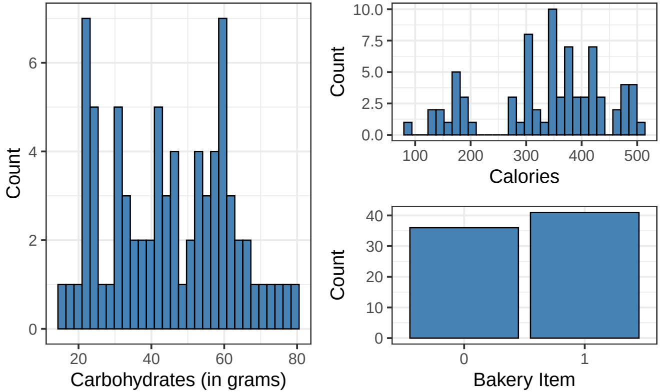
Terminology

- `carb` is the **response variable**
  - variable whose variation we want to understand / variable we wish to predict
  - also known as *outcome* or *dependent* variable

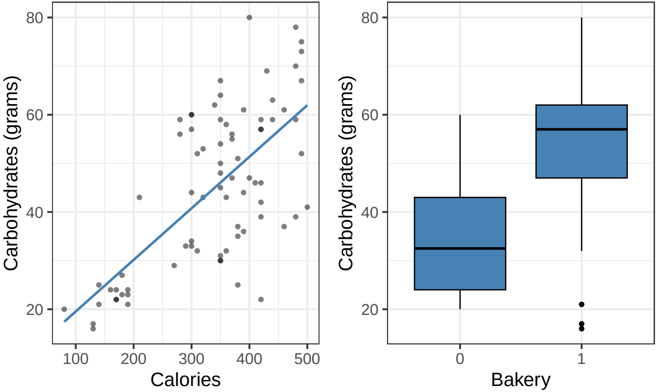
Terminology

- `carb` is the **response variable**
  - variable whose variation we want to understand / variable we wish to predict
  - also known as *outcome* or *dependent* variable
- `calories`, `bakery` are the **predictor variables**
  - variables used to account for variation in the response
  - also known as *explanatory*, *independent*, or *input* variables

Univariate exploratory data analysis



Bivariate exploratory data analysis



Function between response and predictors

$$\text{carb} = f(\text{calories}, \text{bakery}) + \epsilon$$

- **Goal**: Determine f
- How do we determine f?
  - Make an assumption about the functional form f (parametric model)
  - Use the data to fit a model based on that form

Determine f

1) Choose the functional form of f, i.e., **choose the appropriate model given the response variable**

- Suppose f takes the form of a linear model

$$y = f(\mathbf{X}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

Determine f

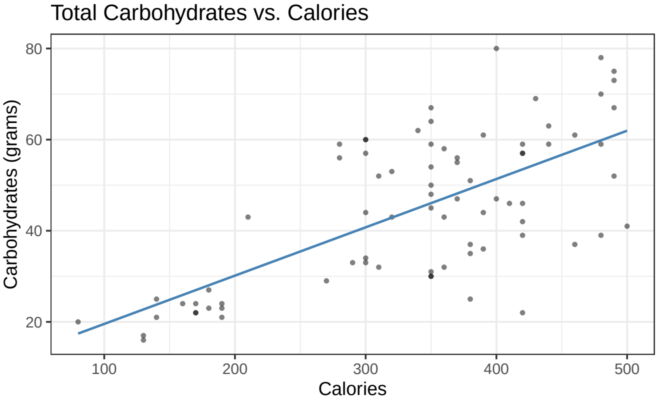
1) Choose the functional form of f, i.e., **choose the appropriate model given the response variable**

- Suppose f takes the form of a linear model

$$y = f(\mathbf{X}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

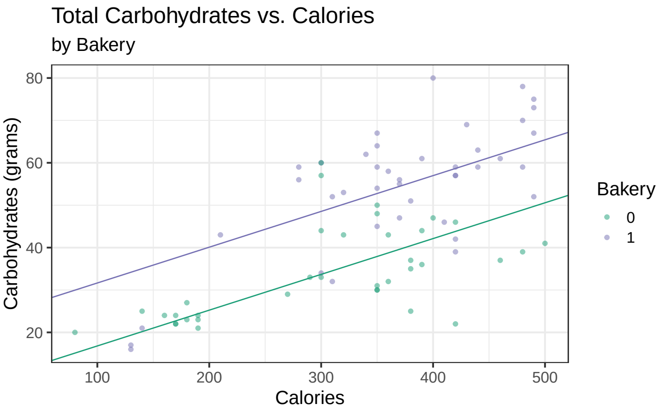
2. Use the data to fit (or train) the model, i.e, **estimate the model parameters,** $\beta_0, \beta_1, \ldots, \beta_p$

Carb vs. Calories



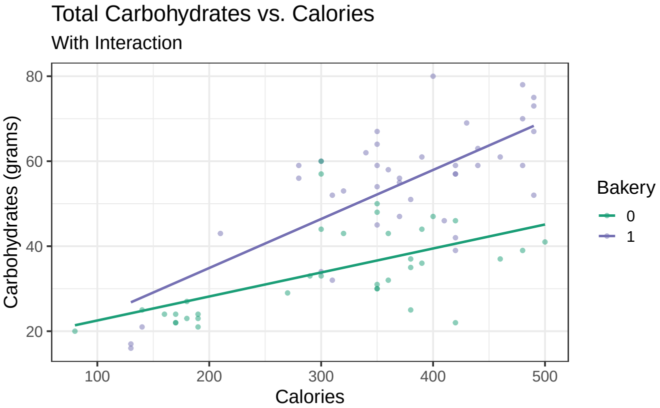
$$\text{carb} = \beta_0 + \beta_1 \ \text{calories} + \epsilon$$

Carb vs. Calories + Bakery



$$\text{carb} = \beta_0 + \beta_1 \text{ calories} + \beta_2 \text{ bakery} + \epsilon$$

Carb vs. Calories + Bakery (with interaction)



$$\text{carb} = \beta_0 + \beta_1 \text{ calories} + \beta_2 \text{ bakery} + \beta_3 \text{ calories} \times \text{bakery} + \epsilon$$

Statistical model vs. regression equation

Statistical model (also known as data-generating model)

$$\text{carb} = \beta_0 + \beta_1 \text{ calories} + \beta_2 \text{ bakery} + \beta_3 \text{ calories} \times \text{bakery} + \epsilon$$

Models the process for generating values of the response in the population (function + error)

Regression equation

Estimate of the function using the sample data

$$\hat{\text{carb}} = \hat{\beta}_0 + \hat{\beta}_1 \text{ calories} + \hat{\beta}_2 \text{ bakery} + \hat{\beta}_3 \text{ calories} \times \text{bakery}$$

Why fit a model?

- **Prediction:** Expected value of the response variable for given values of the predictor variables
- **Inference:** Conclusion about the relationship between the response and predictor variables
- What is an example of a **prediction** question that can be answered using the model of carb vs. calories and bakery?
- What is an example of an **inference** question that can be answered using the model of carb vs. calories and bakery?

Recap

- **Reproducibility**
  - It is best practice conduct all data analysis in a reproducible way
  - We will implement a reproducible workflow using R, Quarto, and git/GitHub
- **Multivariable relationships**
  - We can use exploratory data analysis to describe the relationship between two variables
  - We make an assumption about the relationship between variables when doing linear regression
  - The two main objectives for fitting a linear regression model are (1) prediction and (2) inference

References