Introduction

Luis Francisco Gomez Lopez

2023-07-22

Contents

- What you will learn
- What you would not learn
- References

What you will learn

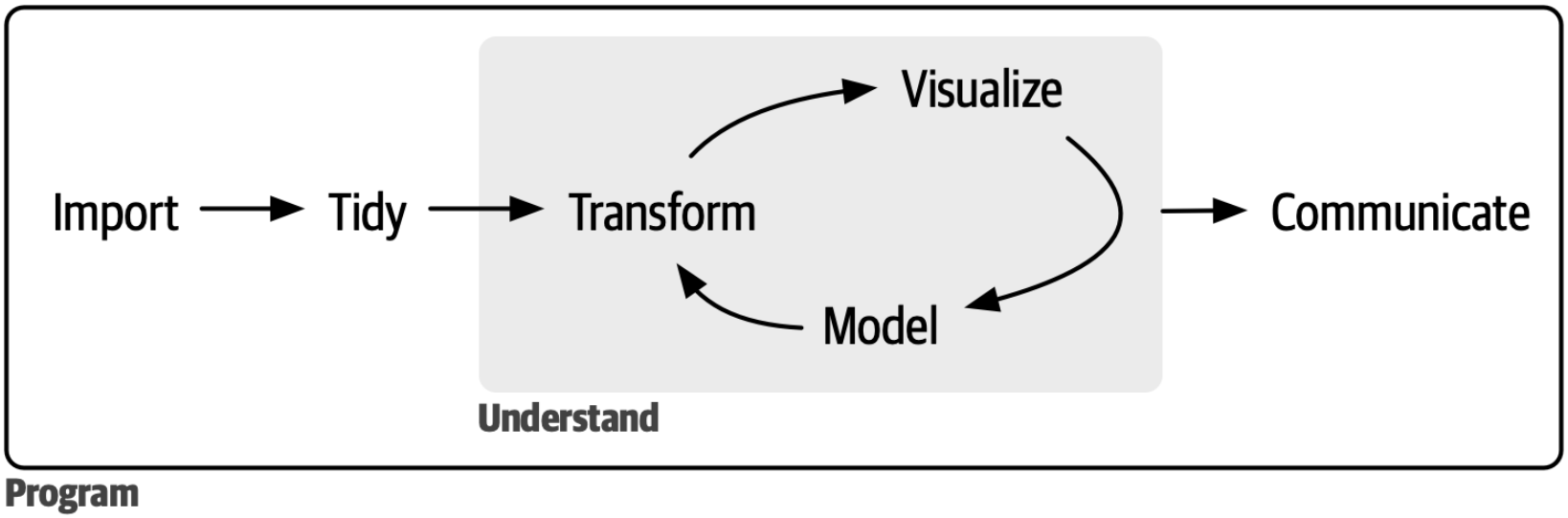

Figure 1: What you will learn (Wickham et al., 2023, fig. 1.1)

What you will learn

- **Import**
  - Take data store outside R and load it into R
    - Files
    - Databases
    - Web APIs[1]

[1] Application programming interface

What you will learn[2]

- **Tidy** (Wickham, 2014)
  - **Data structure**: rectangular tables made up of *rows* and *columns* where every value belongs to a variable and an observation
    - Each variable forms a column
    - Each observation forms a row
    - Every cell is a single value
    - Each type of observational unit forms a table
- **Transform**
  - Narrowing in on observations of interest
  - Creating new variables
  - Calculating a set of summary statistics

[2] Tidying and transforming are called **wrangling**

What you will learn

- Main engines of knowledge generation
  - **Visualization**
    - Show things not expected
    - Raise new questions
    - Identify if you are asking the wrong question
    - Identify if you need to collect different data
    - Don't scale well because it requires human brains
  - **Models**
    - You need to have clearly defined precise questions
    - They scale well because are mathematical and computational tools so they require computers
    - They are based on assumptions so they cannot question its own assumptions

What you will learn

- **Communication**
  - Make others understand your results

What you will learn

- **Programming**
  - Use in nearly every part of a data science project
  - You don't need to be an expert programmer becuase you are a data scientist not a programmer
  - However, learning more about programming pays off because becoming a better programmer allows you to automate common tasks and solve new problems

What you will learn

- **80/20 rule**
  - You can tackle about 80% of every data science project using the tools you will learn
  - You will need other tools to tackle the remaining 20%

What you would not learn

- **Modeling**
  - Use `tidymodels`
    - The tidymodels framework is a collection of packages for modeling and machine learning using tidyverse principles
    - *Tidy Modeling with R* (Kuhn & Silge, 2022)
- **Big data**
  - If you are using large data (10GB - 100GB) learn `data.table`
- **Python, Julia, and friends**
  - Data science teams use a mix of languages
  - If you learn one programming language it will be easy to learn other programming languages but first learn at least one well
  - For python start with:
    - *Python for Data Analysis 3 edition* (McKinney, 2022)
    - *Python Data Science Handbook 2 edition* (VanderPlas, 2023)

References I

Kuhn, M., & Silge, J. (2022). *Tidy modeling with R: A framework for modeling in the tidyverse*. O'Reilly Media. https://www.tmwr.org/

McKinney, W. (2022). *Python for data analysis: Data wrangling with Pandas, NumPy, and Jupyter* (Third edition). O'Reilly. https://wesmckinney.com/book/

VanderPlas, J. (2023). *Python data science handbook: Essential tools for working with data* (Second edition). O'Reilly.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10). https://doi.org/10.18637/jss.v059.i10

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for data science: Import, tidy, transform, visualize, and model data* (2nd edition). O'Reilly Media, Inc. https://r4ds.hadley.nz/