Rotman

INTRO TO R PROGRAMMING

R Tutorial (RSM358) – Session 3, 4

September 28, 2023 Prepared by Jay Cao / TDMDAL

Website: https://tdmdal.github.io/r-intro-2023-fall/

Rotman School of Management
UNIVERSITY OF TORONTO

Plan – Data Analysis with R

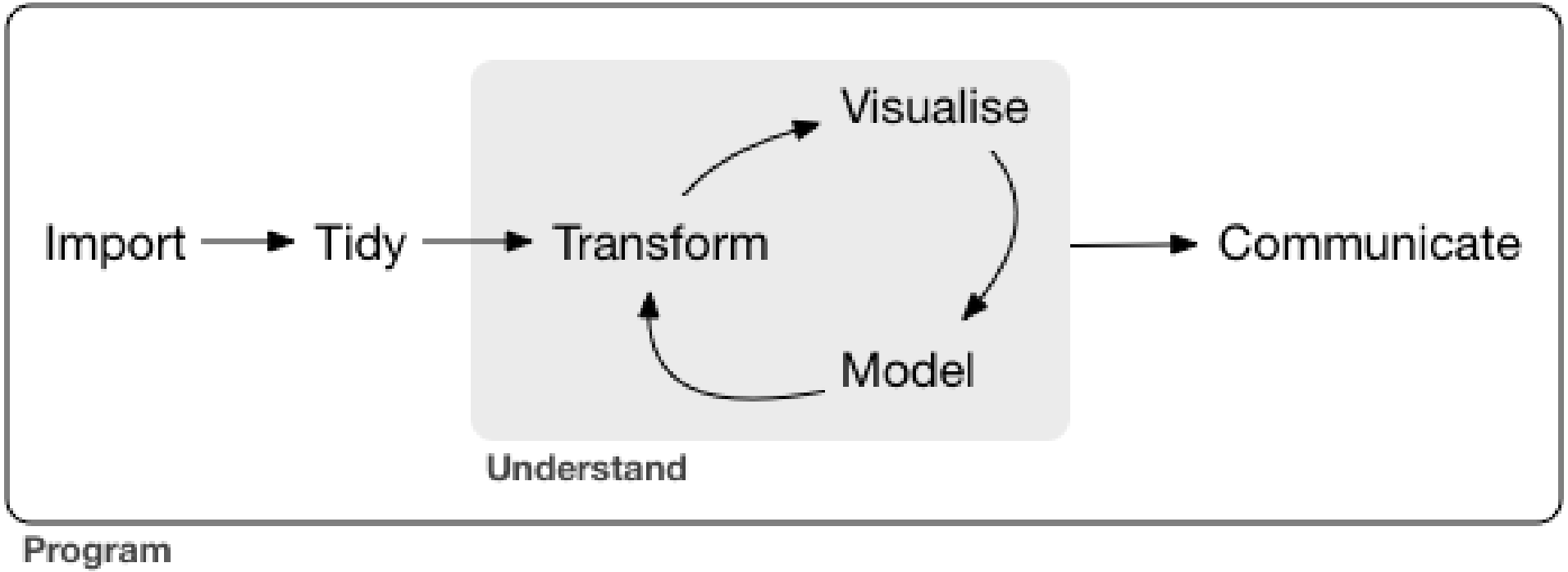
- A Typical data analysis workflow

- Choice of R packages (many choices to do the same task)
  - Data manipulation
  - Modeling

- An example: regression analysis
  - Use Base R for data processing and plotting (**textbook uses this**)
  - Use Tidyverse packages for data processing and plotting (popular choice)

Data Science/Analysis Workflow

- Use this workflow to organize your thoughts and code



https://r4ds.had.co.nz/introduction.html

Using R packages/libraries

- Install an R library (only need to install a library once)

```
install.packages("library_name")
```

- Load an R library (before you use a library)

```
library(library_name)
```

- CRAN (The Comprehensive R Archive Network)
  - CRAN Task Views

Choice of Packages

- Data manipulation (in particular, data frame; three popular choices)
  - Option 1: Base R (data frame)
    - Available since the early days of R
    - **Textbook uses this one**
  - Option 2: Tidyverse eco-system (tibble, which is 95% data frame + extra)
    - Elegant and consistent design across the eco-system -> easy to use
    - Most R users' choice these days
  - Option 3: Data table package (data.table, which is 95% data frame + extra)
    - Fast for even for huge dataset!
- Modeling
  - People converge to the same choice for simple models (e.g., `lm()`, `glm()` , etc.)
  - Many choices for advanced modeling (e.g., time series, deep learning, etc.)

An Example: Housing Price & Clean Air

- Manipulate data
  - Load data
  - Create new columns
  - Filter columns and rows
- Build models
  - Multiple linear regressions
- Report and graph
  - Plot a few graphs
  - Report regression results

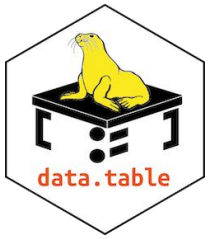
```
Obs:   506

1. price      median housing price, $
2. crime      crimes committed per capita
3. nox        nitrous oxide, parts per 100 mill.
4. rooms      avg number of rooms per house
5. dist       weighted dist. to 5 employ centers
6. radial     accessibiliy index to radial hghwys
7. proptax    property tax per $1000
8. stratio    average student-teacher ratio
9. lowstat    % of people 'lower status'
```

R Packages: Many choices, which one to use

- Often, a task can be achieved using functions in different libraries
  - R is open and extensible!

- Example: load a csv file to a data frame/tibble/data table
  - Use `read.csv()` function from the `utils` library in Base R
  - Use `read_csv()` function from the `readr` library
  - Use `vroom()` from the `vroom` library
  - Use `fread()` function from the `data.table` library



R Packages: Many choices, which one to use

- Start with the one most people use
- Choose one that is well maintained
  - check document, github, etc. for last update date
  - packages maintained by companies (e.g., RStudio Co.) or academic teams
- Choose one that suits your task
- **For RSM358**
  - Follow the examples in the R lab sections of your textbook (mostly using base R)
  - Follow Prof. Webb's notebook examples (mostly using base R)

Choice 1: the Regression Example

- Manipulate data (`Base R`)
  - Load data (`read.csv()`)
  - Create new columns (`base R data frame manipulation`)
  - Filter columns and rows (`base R data frame manipulation`)

- Build models
  - Multiple regression (`lm()` from `stats` library in R base)

- Report and graph
  - Base R plot system, `plot()`
  - Base R `summary()` function

Choice 2: the Regression Example

- Manipulate data (`tidyverse` eco-system)
  - Load data (`read_csv()` from the `readr`)
  - Create new columns (`mutate()` from `dplyr`)
  - Filter columns and rows (`select()` and `filter()` from `dplyr`)
- Build models
  - Multiple regression (`lm()` from `stats` library in R base)
- Report and graph
  - Graph using ggplot2 and some of its extensions
  - Build a publication-ready table (`huxreg()` from `huxtable` library)

Load a CSV file

- `read_csv()` from the `readr`

```
read_csv(file)
```

```
e.g. hprice <- read_csv("hprice.csv")
```

- More about `read_csv()`
- More about `readr`

Data Manipulation: `dplyr` basics

- Filter observations (rows): **`filter()`**

`filter(my_dataframe, condition1, …)`

`e.g., hprice_reg <- filter(hprice, price > 20000)`

- Create new variables: **`mutate()`**

`mutate(my_dataframe, new_var1 = expression1, …)`

`e.g., hprice_reg <- mutate(hprice_reg, lprice = log(price))`

- Select variables (columns): **`select()`**

`select(my_dataframe, var1, …)`

`e.g., hprice_reg <- select(hprice_reg, lprice, rooms)`

Ref. Base R vs dplyr data frame manipulation.

Data Manipulation: Data Pipe (%>%)

```
hprice_reg <- filter(hprice, price > 20000)
hprice_reg <- mutate(hprice_reg, lprice = log(price))
hprice_reg <- select(hprice_reg, lprice, rooms)
```

```
hprice_reg <- hprice %>%
  filter(price > 20000) %>%
  mutate(lprice = log(price)) %>%
  select(lprice, rooms)
```

Regression

- Multiple regressions: `lm()` from `stats` library in base R

```
my_model <- lm(y ~ x1 + x2, data)
```

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$$

```
my_model <- lm(y ~ x1 + x2 + I(x1 * x2), data)
```

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon_i$$

- Regression result summary: `summary()`

Ref. https://faculty.chicagobooth.edu/richard.hahn/teaching/FormulaNotation.pdf

Report

- Summary table
  - Summary for lm(): `summary(my_model)`

- publication-ready table: `huxreg()` from `huxtable` library

```
huxtable(my_model1, my_model2, …)
```

Ref. https://hughjonesd.github.io/huxtable/huxreg.html