R for Statistics Research

Young Geun Kim[1]

Department of Statistics

25 Oct, 2021 (updated: 19 Oct, 2021)

[1] ygeunkim.github.io

Introductory R Functions

Class - Vector

```
(x <- c(2.1, 4.2, 3.3, 5.4))
#> [1] 2.1 4.2 3.3 5.4
class(x)
#> [1] "numeric"
```

- Vectorized code gives the fast result
- `rowSums()`, `colSums()`, `rowMeans()`, `colMeans()` are faster than `apply()` (Wickham, 2019) 🧐

Class - Data frame

```
(df <- data.frame(x = 1:2, y = 2:1, z = letters[1:2]))
#>   x y z
#> 1 1 2 a
#> 2 2 1 b
class(df)
#> [1] "data.frame"
```

- Data for data analysis might be data frame
- Indexing in data frame connects to data transformation

```
y <- MASS::Boston[, c("medv", "lstat", "age")]
head(y)
#>   medv lstat  age
#> 1 24.0  4.98 65.2
#> 2 21.6  9.14 78.9
#> 3 34.7  4.03 61.1
```

Data Analysis using data frame

Multiple linear regression for `medv` ~ `lstat` + age:

```
lm(medv ~ ., data = y)
#>
#> Call:
#> lm(formula = medv ~ ., data = y)
#>
#> Coefficients:
#> (Intercept)        lstat          age
#>     33.2228      -1.0321       0.0345
```

- Data frame is proper to use with many `R` model functions.

Class - List

Contains any object in each element:

```
(z <- list(a = x, b = df))
#> $a
#> [1] 2.1 4.2 3.3 5.4
#>
#> $b
#>   x y z
#> 1 1 2 a
#> 2 2 1 b
class(z)
#> [1] "list"
```

Other Classes

- `matrix` and `array`
  - 2d matrix: linear algebra
  - more than 2d: used in deep learning ("tensor")
- `factor`
- Date: `POSIXct`, `POSIXt`, etc
- Time series: `ts`

```
ts(1:10, frequency = 4, start = c(1959, 2))
#>      Qtr1 Qtr2 Qtr3 Qtr4
#> 1959         1    2    3
#> 1960    4    5    6    7
#> 1961    8    9   10
```

Time Series Model

```
lh
#> Time Series:
#> Start = 1
#> End = 48
#> Frequency = 1
#>  [1] 2.4 2.4 2.4 2.2 2.1 1.5 2.3 2.3 2.5 2.0 1.9 1.7 2.
#> [20] 1.9 1.9 1.8 2.7 3.0 2.3 2.0 2.0 2.9 2.9 2.7 2.7 2.
#> [39] 2.1 3.3 3.5 3.5 3.1 2.6 2.1 3.4 3.0 2.9
class(lh)
#> [1] "ts"
```

Fit AR(1) using `arima` function:

```
arima(lh, order = c(1,0,0))
#>
#> Call:
#> arima(x = lh, order = c(1, 0, 0))
```

Tidy Data

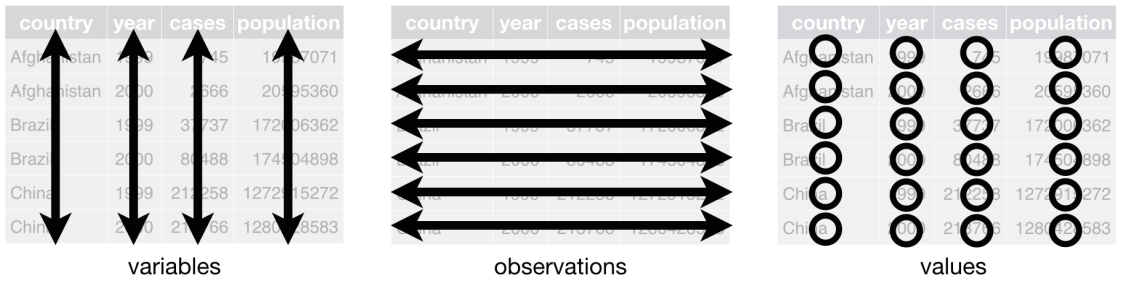


Figure 1: Tidy Data

- For easier data analysis, Wickham (2014) suggests tidy representation of dataset, called **tidy data**.

```
library(tibble)
tibble(
  x = 1:3,
  y = 1,
  z = x^2 + y
)
```

`read.csv` and `readr::read_csv`

- `read.csv(file)`: default function to import csv file
- `readr::read_csv(file)`: read csv file to `tibble`
  - Use this function 😎

Non-tidy Data

```
tidyr::table4a
#> # A tibble: 3 x 3
#>   country     `1999` `2000`
#> * <chr>        <int>  <int>
#> 1 Afghanistan    745   2666
#> 2 Brazil       37737  80488
#> 3 China       212258 213766
```

Data Wrangling

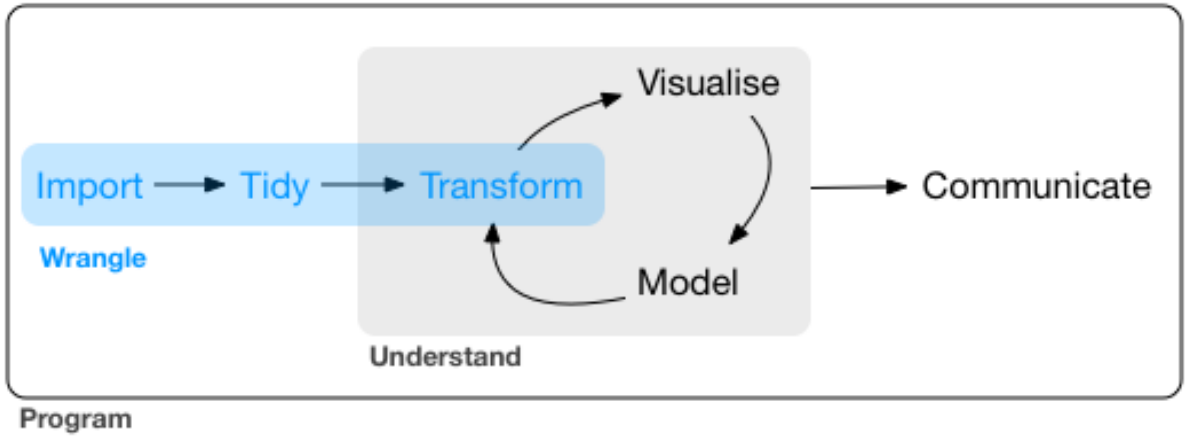


Figure 2: Steps of Data Analysis

- Since many datasets are not tidy, we need data wrangling step
- `dplyr` provides data manipulation functions
- `tidyr` package does this

dplyr

- `mutate()` adds a column
- `summarise()` summarizes variables
- et cetera

tidyr

`tidyr` package can tidy data-set. From Wickham and Grolemund (2017),

- `pivot_longer()`: Figure 3
- `pivot_wider()`: Figure 4

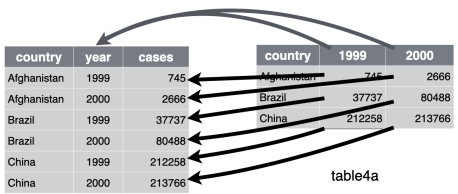


Figure 3: Gather

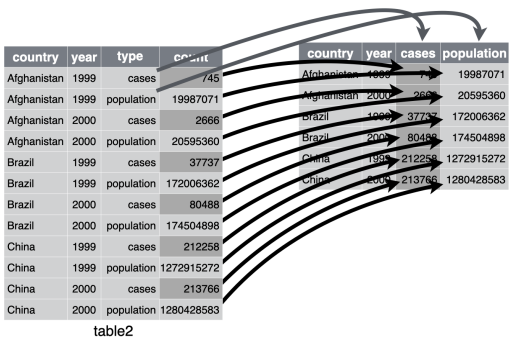


Figure 4: Spread

Large Data

- When data file is too large
- `data.table` package focuses on memory optimization
- `read.csv` or `read_csv` can mistakenly read string as factor
- but `data.table::fread()` do not

Visualization

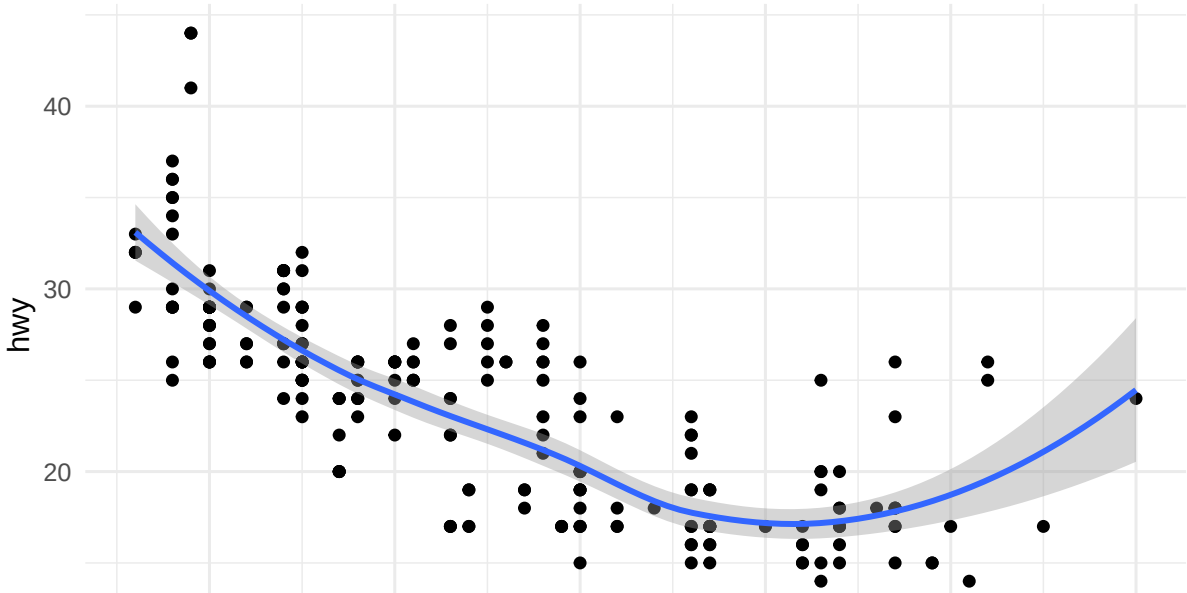
ggplot2	base plot
grammar of graphics	pen on paper model

ggplot2

- Data: data frame (`mpg`)
- Aesthetic mapping: `aes()`
  - x-axis: `displ`
  - y-axis: `hwy`
- Layers: `geom_*()` function
  - scatter plot: `geom_point()`
  - smoothing: `geom_smooth()`

Example

```
library(ggplot2)
#--------------------
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth() +
  theme_minimal()
#> `geom_smooth()` using method = 'loess' and formula 'y ~
```



🤩 Simulation 🤩

Reproducible Documents

- Data can be changed while writing the document
- R Markdown helps reproducibility by integrating Markdown and R.
- Bookdown is a package for authoring books, but it also provides a function for single document: `bookdown::*_document2`
- See Xie et al. (2018) and Xie (2016).

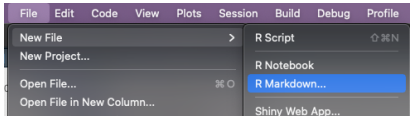


Figure 6: Making R Markdown

Figure 7: rmd Default

Style Guides

- See the `tidyverse style guide`[2]
- NOT google's

[2] https://style.tidyverse.org

Project-Oriented Workflow

Parallel Computations

parLapply, parSapply, and parApply

foreach

doMC::registerDoMC

mclapply

pvec

Reproducible Results

Parallel Options in Popular Funtions

Simulation using Parallelization

Rcpp

R Packages

Sources

- Wickham (2015)

devtools

Structure and Metadata

usethis

Documentation

Build

Check

Vignettes

Test

Data

Additional Tips

xlsx2csv

Pipe

kable and kableExtra

Why `tidyverse`?

tidyverts

Sources I

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10):1–23.

Wickham, H. (2015). *R Packages*. O'Reilly Media.

Wickham, H. (2019). *Advanced R, Second Edition*. CRC Press.

Wickham, H. and Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, Inc., 1st edition.

Xie, Y. (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. CRC Press.

Xie, Y., Allaire, J., and Grolemund, G. (2018). *R Markdown: The Definitive Guide*. CRC Press.