Data wrangling

First, we are going to load the tidyverse package .

library(tidyverse) # Loading necessary package

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

# Creating a table

tribble(~number, ~letter, ~greek,  
 1, "a", "alpha",  
 2, "b", "beta",  
 3, "c", "gamma")

## # A tibble: 3 × 3  
## number letter greek  
## <dbl> <chr> <chr>  
## 1 1 a alpha  
## 2 2 b beta   
## 3 3 c gamma

y <- tibble(number = c(1, 2, 3),   
 letter = c("a", "b", "c"),  
 greek = c("alpha", "beta", "gamma"))

# Convert a Data frame to a tibble

x <- iris # iris is still a data frame

x <- as\_tibble(x) #This will make x a tibble

class(x) # verifying if x is already a tibble

## [1] "tbl\_df" "tbl" "data.frame"

# Convert a data frame to a tibble

z <- table1

z

## # A tibble: 6 × 4  
## country year cases population  
## <chr> <int> <int> <int>  
## 1 Afghanistan 1999 745 19987071  
## 2 Afghanistan 2000 2666 20595360  
## 3 Brazil 1999 37737 172006362  
## 4 Brazil 2000 80488 174504898  
## 5 China 1999 212258 1272915272  
## 6 China 2000 213766 1280428583

class (z)

## [1] "tbl\_df" "tbl" "data.frame"

z <- as.data.frame(table1)

# INSPECT EVERY CELL OF A TIBBLE

S <- storms

View(storms)

# Spread a pair of columns into a field of cells

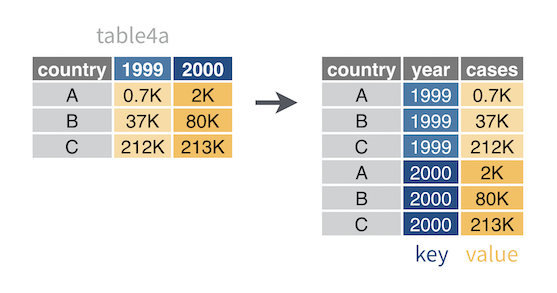
table2 # undidy data

## # A tibble: 12 × 4  
## country year type count  
## <chr> <int> <chr> <int>  
## 1 Afghanistan 1999 cases 745  
## 2 Afghanistan 1999 population 19987071  
## 3 Afghanistan 2000 cases 2666  
## 4 Afghanistan 2000 population 20595360  
## 5 Brazil 1999 cases 37737  
## 6 Brazil 1999 population 172006362  
## 7 Brazil 2000 cases 80488  
## 8 Brazil 2000 population 174504898  
## 9 China 1999 cases 212258  
## 10 China 1999 population 1272915272  
## 11 China 2000 cases 213766  
## 12 China 2000 population 1280428583

table2\_tidy <- table2 %>% # and then  
 spread(key = type, value = count)  
table2\_tidy

## # A tibble: 6 × 4  
## country year cases population  
## <chr> <int> <int> <int>  
## 1 Afghanistan 1999 745 19987071  
## 2 Afghanistan 2000 2666 20595360  
## 3 Brazil 1999 37737 172006362  
## 4 Brazil 2000 80488 174504898  
## 5 China 1999 212258 1272915272  
## 6 China 2000 213766 1280428583

# Gather a field of cells into a pair of columns



just another screenshot

table4a

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

table4a %>%   
 gather(key = "year", value = "cases", 2:3)

## # A tibble: 6 × 3  
## country year cases  
## <chr> <chr> <int>  
## 1 Afghanistan 1999 745  
## 2 Brazil 1999 37737  
## 3 China 1999 212258  
## 4 Afghanistan 2000 2666  
## 5 Brazil 2000 80488  
## 6 China 2000 213766

# Separate a column into new columns

table3

## # A tibble: 6 × 3  
## country year rate   
## \* <chr> <int> <chr>   
## 1 Afghanistan 1999 745/19987071   
## 2 Afghanistan 2000 2666/20595360   
## 3 Brazil 1999 37737/172006362   
## 4 Brazil 2000 80488/174504898   
## 5 China 1999 212258/1272915272  
## 6 China 2000 213766/1280428583

table3\_sep <- table3 %>%  
 separate(col = rate,   
 into = c("cases", "population"),   
 sep = "/",   
 convert = TRUE)  
table3\_sep

## # A tibble: 6 × 4  
## country year cases population  
## <chr> <int> <int> <int>  
## 1 Afghanistan 1999 745 19987071  
## 2 Afghanistan 2000 2666 20595360  
## 3 Brazil 1999 37737 172006362  
## 4 Brazil 2000 80488 174504898  
## 5 China 1999 212258 1272915272  
## 6 China 2000 213766 1280428583