

# Reproducible science: Module 3

## Dealing with data: Tidyverse

Gbadamassi G.O. Dossa

Xishuangbanna Tropical Botanical Garden, XTBG-CAS

(updated: 2022-11-04)

# Acknowledgements

The content of this module are based on materials from:



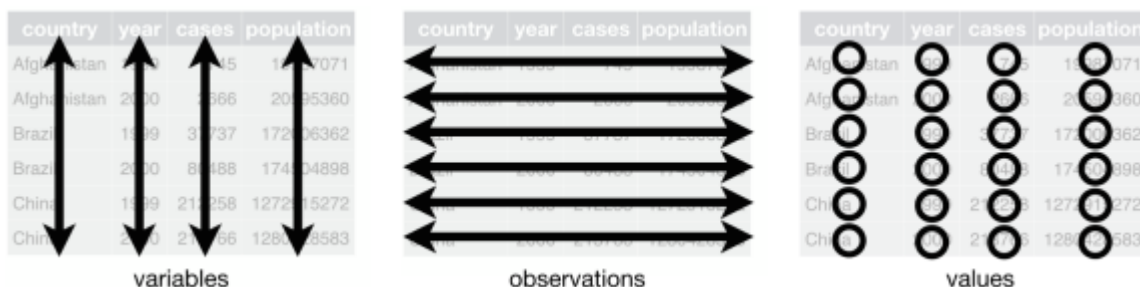
olivier gimenez's materials

# What is tidyverse and advantages?

"A framework for managing data that aims at making the cleaning and preparing steps [muuuuuuuch] easier" (Julien Barnier). Main characteristics of a tidy dataset:

- the dataset is **tibble**;
- measured variable as a column;
- an observation represents a row with each value is in a different cell.

**tidyverse** consists of a compilation of r packages for data analysis.



# Recognizing a tidy dataset

```
#> # A tibble: 12 x 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  
#> # ... with 6 more rows
```

Is this a tidy data?

No

# Recognizing a tidy dataset

```
#> # A tibble: 6 x 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
```

Is this a tidy data?

No

# Recognizing a tidy dataset

```
# Spread across two tibbles
# cases
#> # A tibble: 3 x 3
#>   country      '1999'  '2000'
#> * <chr>         <int>   <int>
#> 1 Afghanistan     745     2666
#> 2 Brazil          37737   80488
#> 3 China           212258  213766
# population
#> # A tibble: 3 x 3
#>   country      '1999'      '2000'
#> * <chr>         <int>        <int>
#> 1 Afghanistan  19987071   20595360
#> 2 Brazil       172006362  174504898
#> 3 China        1272915272  1280428583
```

Is this a tidy data?

No

# Recognizing a tidy dataset

```
#> # A tibble: 6 x 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
```

Is this a tidy data?

Yes

# Tidyverse: Multiple r packages well compiled

Allows using a consistent format for which powerful tools work.

Makes data manipulation pretty natural

- **ggplot2** - visualizing stuff;
- **dplyr**, **tidyr** - data manipulation;
- **purrr** - advanced programming;
- **readr** - import data;
- **tibble** - improved data.frame format;
- **forcats** - working with factors;
- **stringr** - working with chain of characters.

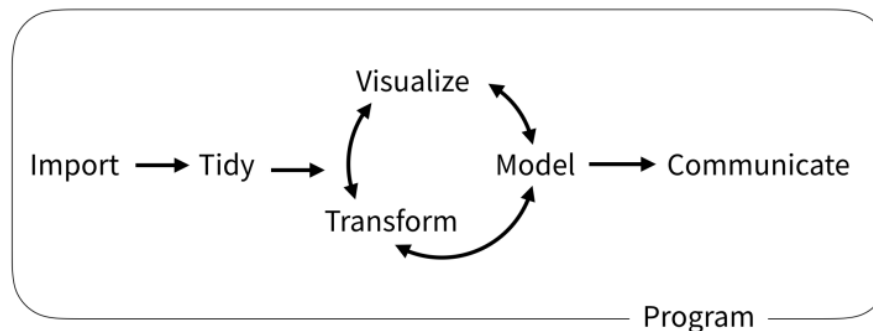


# Simplified flowchart of data science?

Any data analysis follows this typical flow:

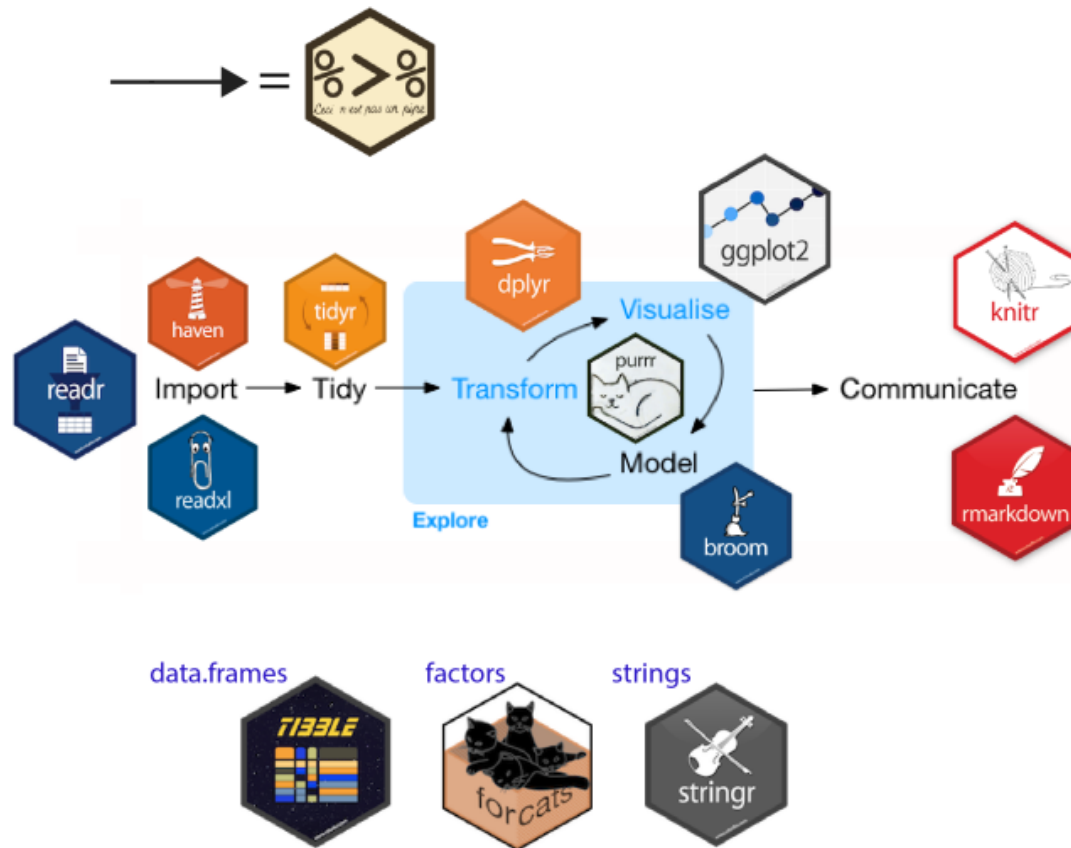
1. Import data;
2. Clean data;
3. Exploratory analysis. A cycle between:
  - Visualization;
  - modeling;
  - Transformation
4. Communicate

If these steps happen at multiple software then errors are highly inevitable.



Reproducibility equals efficient use of time

# Tidyverse saves: same flowchart in tidyverse



Reproducibility equals efficient use of time

# Practice in tidyverse “Use twitter to predict citation rate”

The screenshot shows the PLOS ONE website interface. At the top, there's a navigation bar with the PLOS ONE logo, links for PUBLISH, ABOUT, and BROWSE, a search bar, and links for plos.org, create account, and sign in. Below the navigation bar, the article is identified as an OPEN ACCESS, PEER-REVIEWED RESEARCH ARTICLE. The title is 'Twitter Predicts Citation Rates of Ecological Research'. The authors listed are Brandon K. Peoples, Stephen R. Midway, Dana Sackett, Abigail Lynch, and Patrick B. Cooney. The publication date is November 11, 2016, and the DOI is https://doi.org/10.1371/journal.pone.0166570. On the right side, there's a metrics box showing 120 Saves, 32 Citations, 18,800 Views, and 698 Shares. At the bottom, there's a tabbed interface with 'Article' selected, and other tabs for Authors, Metrics, Comments, and Media Coverage. A 'Download PDF' button is also visible.

plos.org create account sign in

PLOS ONE PUBLISH ABOUT BROWSE SEARCH advanced search

OPEN ACCESS PEER-REVIEWED RESEARCH ARTICLE

## Twitter Predicts Citation Rates of Ecological Research

Brandon K. Peoples , Stephen R. Midway , Dana Sackett , Abigail Lynch , Patrick B. Cooney

Published: November 11, 2016 • <https://doi.org/10.1371/journal.pone.0166570>

120 Save	32 Citation
18,800 View	698 Share

Article Authors Metrics Comments Media Coverage Download PDF

We will use an existing data supporting the **above publication** to learn some functions within **tidyverse**.

# Import data

`readr::read_csv` function:

- creates tibbles instead of `data.frame`;
- no names to rows;
- allows column names with special characters (see next slide);
- more clever on screen display than w/ `data.frames` (see next slide);
- no partial matching on column names;
- warning if attempt to access unexisting column;
- is incredibly fast.

# Import data

```
# Set the url from where to download the data  
url<-"https://doi.org/10.1371/journal.pone.0166570.s001"  
# name the file to be downloaded and save as destfile object  
destfile <- "twitter_cit_data.csv"  
# Apply download.file function in R to download from url  
download.file(url, destfile)  
library(tidyverse)  
# Read the data file with read_csv() and save with name "citations_raw"  
citations_raw<-read_csv(file="twitter_cit_data.csv")  
head(citations_raw)
```

# Import data

```
citations_raw
```

```
## # A tibble: 1,599 × 12
##   Journa...1 5-yea...2 Year ...3 Volume Issue Authors Colle...4 Publi...5 Numbe...6 N
##   <chr>      <dbl>    <dbl>    <dbl> <chr> <chr>    <chr>    <chr>    <dbl>
## 1 Ecology...    16.7    2014     17 12    Morin ... 2/1/20... 9/16/2...    18
## 2 Ecology...    16.7    2014     17 12    Jucker... 2/1/20... 10/13/...    15
## 3 Ecology...    16.7    2014     17 12    Calcag... 2/1/20... 10/21/...     5
## 4 Ecology...    16.7    2014     17 11    Segre ... 2/1/20... 8/28/2...     9
## 5 Ecology...    16.7    2014     17 11    Kaufma... 2/1/20... 8/28/2...     3
## 6 Ecology...    16.7    2014     17 10    Nasto ... 2/2/20... 7/28/2...    27
## 7 Ecology...    16.7    2014     17 10    Tschir... 2/2/20... 8/6/20...     6
## 8 Ecology...    16.7    2014     17 9     Barnece... 2/2/20... 6/17/2...    19
## 9 Ecology...    16.7    2014     17 9     Pinto-... 2/2/20... 6/12/2...    26
## 10 Ecology...   16.7    2014     17 9     Clough... 2/2/20... 7/17/2...    44
## # ... with 1,589 more rows, 2 more variables: `Twitter reach` <dbl>,
## #   `Number of Web of Science citations` <dbl>, and abbreviated variable r
## #   1`Journal identity`, 2`5-year journal impact factor`, 3`Year published
## #   4`Collection date`, 5`Publication date`, 6`Number of tweets`,
## #   7`Number of users`
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all var
```

# Tidy/transform: Rename columns

To rename columns, use function *rename()* new\_name=old\_name

```
citations_temp <- rename(citations_raw,
  journal = 'Journal identity',
  impactfactor = '5-year journal impact factor',
  pubyear = 'Year published',
  colldate = 'Collection date',
  pubdate = 'Publication date',
  nbtweets = 'Number of tweets',
  woscitations = 'Number of Web of Science citations')
head(citations_temp,5,6)
```

```
## # A tibble: 5 × 12
##   journal    impac...1 pubyear Volume Issue Authors coll...2 pubdate nbtwe...3
##   <chr>      <dbl>    <dbl>   <dbl> <chr> <chr>   <chr>   <chr>      <dbl>
## 1 Ecology ...    16.7    2014     17 12   Morin ... 2/1/20... 9/16/2...    18
## 2 Ecology ...    16.7    2014     17 12   Jucker... 2/1/20... 10/13/...    15
## 3 Ecology ...    16.7    2014     17 12   Calcag... 2/1/20... 10/21/...     5
## 4 Ecology ...    16.7    2014     17 11   Segre ... 2/1/20... 8/28/2...     9
## 5 Ecology ...    16.7    2014     17 11   Kaufma... 2/1/20... 8/28/2...     3
## # ... with 2 more variables: `Twitter reach` <dbl>, woscitations <dbl>, and
## # abbreviated variable names 1impactfactor, 2colldate, 3nbtweets, 15 / 58
```

# Tidy: Clean up column names

To clean columns, use function `clean_names()` from the package `janitor` from it will fill space in column names by "\_".

```
janitor::clean_names(citations_raw)
```

```
## # A tibble: 1,599 × 12
##   journa...1 x5_ye...2 year_...3 volume issue authors colle...4 publi...5 numbe...6 n
##   <chr>      <dbl>    <dbl>    <dbl> <chr> <chr>    <chr>    <chr>      <dbl>
## 1 Ecology...  16.7    2014     17 12    Morin ... 2/1/20... 9/16/2...    18
## 2 Ecology...  16.7    2014     17 12    Jucker... 2/1/20... 10/13/...    15
## 3 Ecology...  16.7    2014     17 12    Calcag... 2/1/20... 10/21/...     5
## 4 Ecology...  16.7    2014     17 11    Segre ... 2/1/20... 8/28/2...     9
## 5 Ecology...  16.7    2014     17 11    Kaufma... 2/1/20... 8/28/2...     3
## 6 Ecology...  16.7    2014     17 10    Nasto ... 2/2/20... 7/28/2...    27
## 7 Ecology...  16.7    2014     17 10    Tschir... 2/2/20... 8/6/20...     6
## 8 Ecology...  16.7    2014     17 9     Barnec... 2/2/20... 6/17/2...    19
## 9 Ecology...  16.7    2014     17 9     Pinto-... 2/2/20... 6/12/2...    26
## 10 Ecology... 16.7    2014     17 9     Clough... 2/2/20... 7/17/2...    44
## # ... with 1,589 more rows, 2 more variables: twitter_reach <dbl>,
## #   number_of_web_of_science_citations <dbl>, and abbreviated variable nam
## #   1journal_identity, 2x5_year_journal_impact_factor, 3year_published,
## #   4collection_date, 5publication_date, 6number_of_tweets, 7number_of_use
```



# Tidy: Create and modify columns

The well known function to create and modify columns is *mutate()*, This function takes first the tibble names, the new\_name= what you want to do to old column.

```
citations <- mutate(citations_temp, journal = as.factor(journal))  
#Pay attention that I store in "citations"  
citations
```

```
## # A tibble: 1,599 × 12
```

##	journal	impac... <sup>1</sup>	pubyear	Volume	Issue	Authors	colld... <sup>2</sup>	pubdate	nbtwe... <sup>3</sup>	M
##	<fct>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<chr>	<dbl>	
##	1 Ecology...	16.7	2014	17	12	Morin ...	2/1/20...	9/16/2...	18	
##	2 Ecology...	16.7	2014	17	12	Jucker...	2/1/20...	10/13/...	15	
##	3 Ecology...	16.7	2014	17	12	Calcag...	2/1/20...	10/21/...	5	
##	4 Ecology...	16.7	2014	17	11	Segre ...	2/1/20...	8/28/2...	9	
##	5 Ecology...	16.7	2014	17	11	Kaufma...	2/1/20...	8/28/2...	3	
##	6 Ecology...	16.7	2014	17	10	Nasto ...	2/2/20...	7/28/2...	27	
##	7 Ecology...	16.7	2014	17	10	Tschir...	2/2/20...	8/6/20...	6	
##	8 Ecology...	16.7	2014	17	9	Barnecc...	2/2/20...	6/17/2...	19	
##	9 Ecology...	16.7	2014	17	9	Pinto-...	2/2/20...	6/12/2...	26	
##	10 Ecology...	16.7	2014	17	9	Clough...	2/2/20...	7/17/2...	44	

## # ... with 1,589 more rows, 2 more variables: `Twitter reach` <dbl>, 17 / 58

# Tidy: Create and modify columns

Check now the levels of journal variable

```
levels(citations$journal)
```

```
## [1] "Animal Conservation"          "Conservation Letters"
## [3] "Diversity and Distributions"  "Ecological Applications"
## [5] "Ecology"                     "Ecology Letters"
## [7] "Evolution"                   "Evolutionary Applications"
## [9] "Fish and Fisheries"          "Functional Ecology"
## [11] "Global Change Biology"       "Global Ecology and Biogeography"
## [13] "Journal of Animal Ecology"   "Journal of Applied Ecology"
## [15] "Journal of Biogeography"     "Limnology and Oceanography"
## [17] "Mammal Review"              "Methods in Ecology and Evolution"
## [19] "Molecular Ecology Resources" "New Phytologist"
```

# Piping: Make your manipulations easier

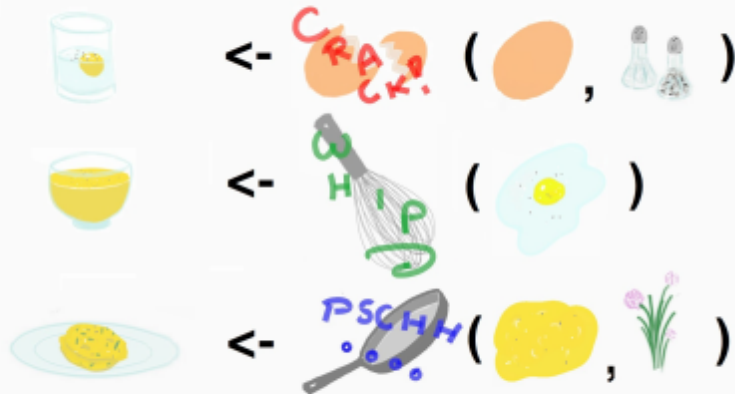
Piping was borrowed from other languages (), got incorporated into R after a question in [Pipe question](#) in 2012. Pipe which is the bar "|" on your...

# Omelette: Base r approach

You need to do complicated programming: create multiple intermediate objects; embed, needs some understanding of coding and is prone to errors.

```
white_and_yolk <- crack(egg, add_seasoning)
omelette_batter <- beat(white_and_yolk)
omelette_with_chives <- cook(omelette_batter, add_chives)
```

## Successive command lines



# Omelette: Piping approach

Simpler programming using piping. Piping consists of: taking results from previous function as a starting point of a new function; less prone to errors and consume less memory.

```
egg %>%  
  crack(add_seasoning) %>%  
  beat() %>%  
  cook(add_chives) -> omelette_with_chives
```



# Example of piping

Take the tibble "citations\_raw" **then** rename some columns then the new tibble containing the renamed tibble and *then* convert the column "journal" from current class ("character") to factor.

```
citations_raw %>%  
  rename(journal = 'Journal identity',  
         impactfactor = '5-year journal impact factor',  
         pubyear = 'Year published',  
         colldate = 'Collection date',  
         pubdate = 'Publication date',  
         nbtweets = 'Number of tweets',  
         woscitations = 'Number of Web of Science citations') %>%  
  mutate(journal = as.factor(journal))
```

Please notice every time I say **"then"** this is equal to "%>%".

# Naming final object of pipe

```
citations <- citations_raw %>%
  rename(journal = 'Journal identity',
         impactfactor = '5-year journal impact factor',
         pubyear = 'Year published',
         colldate = 'Collection date',
         pubdate = 'Publication date',
         nbtweets = 'Number of tweets',
         woscitations = 'Number of Web of Science citations') %>%
  mutate(journal = as.factor(journal))
head(citations)
```

```
## # A tibble: 6 × 12
##   journal    impac...1 pubyear Volume Issue Authors coll...2 pubdate nbtwe...3 M
##   <fct>      <dbl>    <dbl>   <dbl> <chr>  <chr>   <chr>   <chr>      <dbl>
## 1 Ecology ...    16.7    2014     17  12    Morin ... 2/1/20... 9/16/2...    18
## 2 Ecology ...    16.7    2014     17  12    Jucker... 2/1/20... 10/13/...    15
## 3 Ecology ...    16.7    2014     17  12    Calcag... 2/1/20... 10/21/...     5
## 4 Ecology ...    16.7    2014     17  11    Segre ... 2/1/20... 8/28/2...     9
## 5 Ecology ...    16.7    2014     17  11    Kaufma... 2/1/20... 8/28/2...     3
## 6 Ecology ...    16.7    2014     17  10    Nasto ... 2/2/20... 7/28/2...    27
## # ... with 2 more variables: `Twitter reach` <dbl>, woscitations <dbl>, and
## # abbreviated variable names 1impactfactor, 2colldate, 3nbtweets, 23/58
```

# Naming final object of pipe 2

```
citations_raw %>%
  rename(journal = 'Journal identity',
         impactfactor = '5-year journal impact factor',
         pubyear = 'Year published',
         colldate = 'Collection date',
         pubdate = 'Publication date',
         nbtweets = 'Number of tweets',
         woscitations = 'Number of Web of Science citations') %>%
  mutate(journal = as.factor(journal)) -> citations2
head(citations2)
```

```
## # A tibble: 6 × 12
##   journal    impac...1 pubyear Volume Issue Authors coll...2 pubdate nbtwe...3
##   <fct>      <dbl>    <dbl>   <dbl> <chr>  <chr>   <chr>   <chr>      <dbl>
## 1 Ecology ...    16.7    2014     17  12    Morin ... 2/1/20... 9/16/2...    18
## 2 Ecology ...    16.7    2014     17  12    Jucker... 2/1/20... 10/13/...    15
## 3 Ecology ...    16.7    2014     17  12    Calcag... 2/1/20... 10/21/...     5
## 4 Ecology ...    16.7    2014     17  11    Segre ... 2/1/20... 8/28/2...     9
## 5 Ecology ...    16.7    2014     17  11    Kaufma... 2/1/20... 8/28/2...     3
## 6 Ecology ...    16.7    2014     17  10    Nasto ... 2/2/20... 7/28/2...    27
## # ... with 2 more variables: `Twitter reach` <dbl>, woscitations <dbl>, and
## # abbreviated variable names 1impactfactor, 2colldate, 3nbtweets, 24/58
```



# Pipe syntax

- *Verb(Subject,Complement)* replaced by *Subject %>% Verb(Complement)*;
- No need to name unimportant intermediate variables;
- Clear syntax (readability).

If you want you can first write what you want to accomplished in a text with "then" as step wise, then code it by replace "then" by the pipe with its operator "%>%" of course.



# Other functions in Tidyverse

# Select columns

*select()* is the function one uses to select different variables in a tibble. You just need to remember that it follows a pipe operator (`%>%`), and it takes the name of columns one desires to select.

```
citations %>%  
  select(journal, impactfactor, nbtweets)
```

```
## # A tibble: 1,599 × 3  
##   journal          impactfactor nbtweets  
##   <fct>              <dbl>      <dbl>  
## 1 Ecology Letters    16.7         18  
## 2 Ecology Letters    16.7         15  
## 3 Ecology Letters    16.7          5  
## 4 Ecology Letters    16.7          9  
## 5 Ecology Letters    16.7          3  
## 6 Ecology Letters    16.7         27  
## 7 Ecology Letters    16.7          6  
## 8 Ecology Letters    16.7         19  
## 9 Ecology Letters    16.7         26  
## 10 Ecology Letters   16.7         44  
## # ... with 1,589 more rows  
## # Use `print(n = ...)` to see more rows
```

# Drop columns or deselect variables

The opposite of selecting, which is deselecting. One just need to be more logical in the writing. Would you like to guess?

```
citations %>%  
  select(-Volume, -Issue, -Authors)
```

```
## # A tibble: 1,599 × 9  
##   journal      impac...1 pubyear colld...2 pubdate nbtwe...3 Numbe...4 Twitt...5 w  
##   <fct>         <dbl>    <dbl> <chr>    <chr>    <dbl>    <dbl>    <dbl>  
## 1 Ecology Lett...    16.7    2014 2/1/20... 9/16/2...    18      16    29877  
## 2 Ecology Lett...    16.7    2014 2/1/20... 10/13/...    15      12     5997  
## 3 Ecology Lett...    16.7    2014 2/1/20... 10/21/...     5       4     1667  
## 4 Ecology Lett...    16.7    2014 2/1/20... 8/28/2...     9       8     3482  
## 5 Ecology Lett...    16.7    2014 2/1/20... 8/28/2...     3       3     1329  
## 6 Ecology Lett...    16.7    2014 2/2/20... 7/28/2...    27      23    41906  
## 7 Ecology Lett...    16.7    2014 2/2/20... 8/6/20...     6       6    12223  
## 8 Ecology Lett...    16.7    2014 2/2/20... 6/17/2...    19      18    22020  
## 9 Ecology Lett...    16.7    2014 2/2/20... 6/12/2...    26      23    23003  
## 10 Ecology Lett...    16.7    2014 2/2/20... 7/17/2...    44      42   131788  
## # ... with 1,589 more rows, and abbreviated variable names 1impactfactor,  
## # 2colldate, 3nbtweets, 4`Number of users`, 5`Twitter reach`, 6woscitati  
## # i Use `print(n = ...)` to see more rows
```

# Split a column in several columns

`separate` is the function used to split a column into several of course you need to indicate what symbol is the separator (e.g., space, -, /, etc.).

```
head(citations$pubdate)
```

```
## [1] "9/16/2014" "10/13/2014" "10/21/2014" "8/28/2014" "8/28/2014"  
## [6] "7/28/2014"
```

```
citations %>%  
  select(journal, impactfactor, nbtweets, pubdate)%>%  
  separate(pubdate, c('month', 'day', 'year'), '/')
```

```
## # A tibble: 1,599 × 6  
##   journal          impactfactor nbtweets month day   year  
##   <fct>              <dbl>      <dbl> <chr> <chr> <chr>  
## 1 Ecology Letters    16.7        18 9     16    2014  
## 2 Ecology Letters    16.7        15 10    13    2014  
## 3 Ecology Letters    16.7         5 10    21    2014  
## 4 Ecology Letters    16.7         9 8     28    2014  
## 5 Ecology Letters    16.7         3 8     28    2014  
## 6 Ecology Letters    16.7        27 7     28    2014  
## 7 Ecology Letters    16.7         6 8     6     2014
```

# Transform column in date format

Many of us work with ecological data that record date, and we find it hard to keep these on readable format in R. Within, tidyverse there is a package that specially deals with date formatting variables/columns. The package is called **lubridate**.

```
library(lubridate)
citations %>%
  mutate(pubdate = mdy(pubdate),
         colldate = mdy(colldate))%>%
  select(journal, impactfactor, nbtweets, pubdate, colldate)
```

```
## # A tibble: 1,599 × 5
##   journal          impactfactor nbtweets pubdate colldate
##   <fct>              <dbl>      <dbl> <date>   <date>
## 1 Ecology Letters    16.7         18 2014-09-16 2016-02-01
## 2 Ecology Letters    16.7         15 2014-10-13 2016-02-01
## 3 Ecology Letters    16.7          5 2014-10-21 2016-02-01
## 4 Ecology Letters    16.7          9 2014-08-28 2016-02-01
## 5 Ecology Letters    16.7          3 2014-08-28 2016-02-01
## 6 Ecology Letters    16.7         27 2014-07-28 2016-02-02
## 7 Ecology Letters    16.7          6 2014-08-06 2016-02-02
## 8 Ecology Letters    16.7         19 2014-06-17 2016-02-02
```

# For easy date format manipulation

Check out `?lubridate::lubridate` for more functions

```
library(lubridate)
citations %>%
  mutate(pubdate = mdy(pubdate),
         pubyear2 = year(pubdate))%>%
  select(journal, impactfactor, pubdate, colldate, pubyear2)
```

```
## # A tibble: 1,599 × 5
##   journal          impactfactor pubdate    colldate pubyear2
##   <fct>              <dbl> <date>      <chr>         <dbl>
## 1 Ecology Letters    16.7 2014-09-16 2/1/2016      2014
## 2 Ecology Letters    16.7 2014-10-13 2/1/2016      2014
## 3 Ecology Letters    16.7 2014-10-21 2/1/2016      2014
## 4 Ecology Letters    16.7 2014-08-28 2/1/2016      2014
## 5 Ecology Letters    16.7 2014-08-28 2/1/2016      2014
## 6 Ecology Letters    16.7 2014-07-28 2/2/2016      2014
## 7 Ecology Letters    16.7 2014-08-06 2/2/2016      2014
## 8 Ecology Letters    16.7 2014-06-17 2/2/2016      2014
## 9 Ecology Letters    16.7 2014-06-12 2/2/2016      2014
## 10 Ecology Letters   16.7 2014-07-17 2/2/2016      2014
## # ... with 1,589 more rows
```

Join tables together



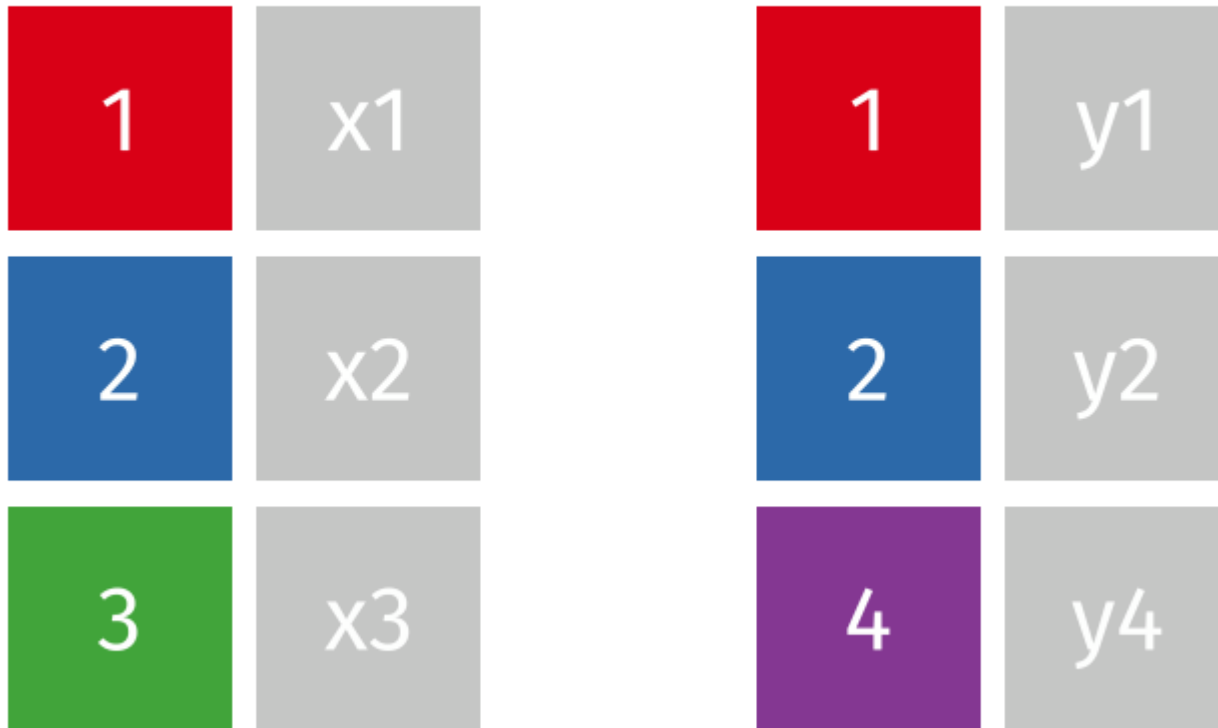
# Join two tables

Joining tables are the correspondents of merge function in base R. There is a great tutorial to all sort of joining in tidyverse made available by [Garrrick Aden-Buie](#). The joining of tables can be categorized into several types. However, we will only study the following:

- Inner join;
- Left join;
- Right join;
- Semi join;
- Union join;
- Anti join.

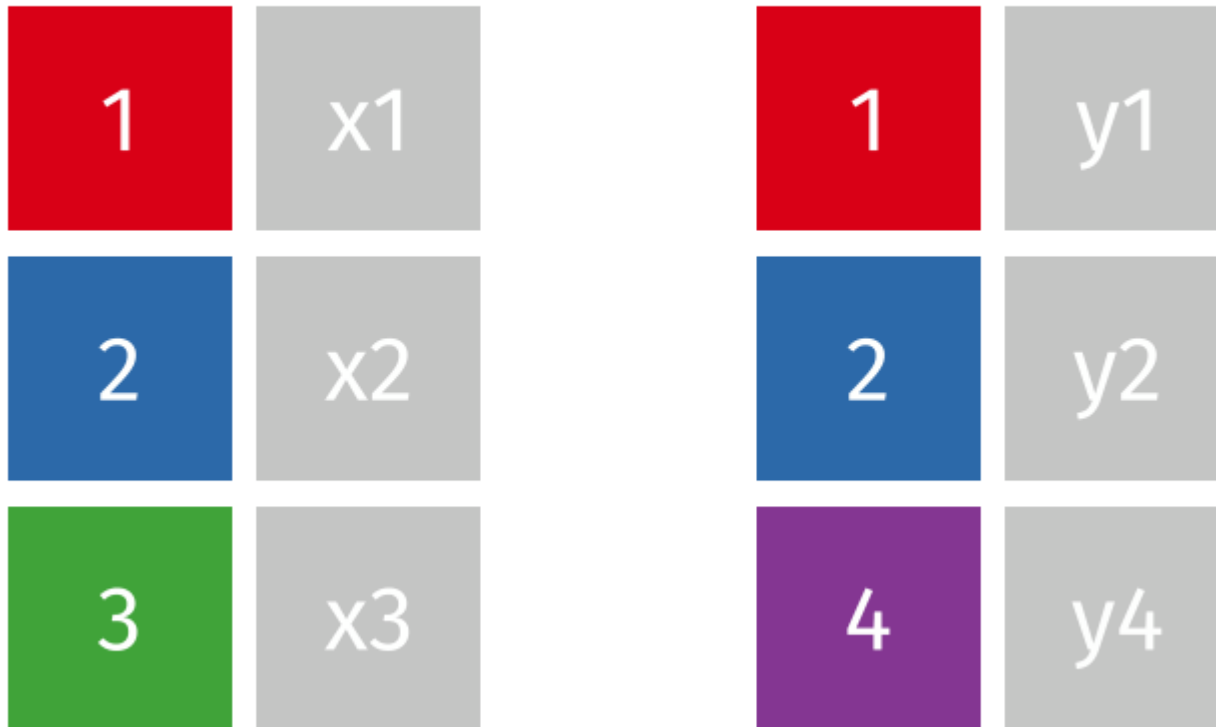
# Inner join

`inner_join(x, y)`



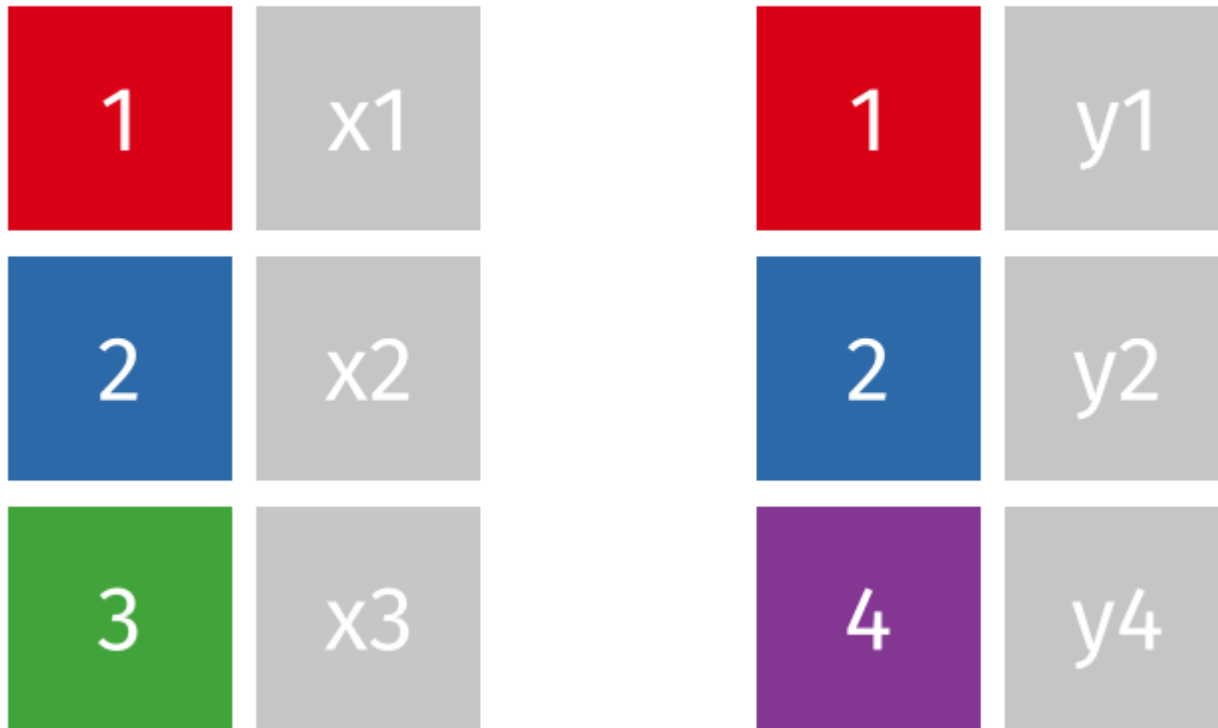
# Left join

`left_join(x, y)`



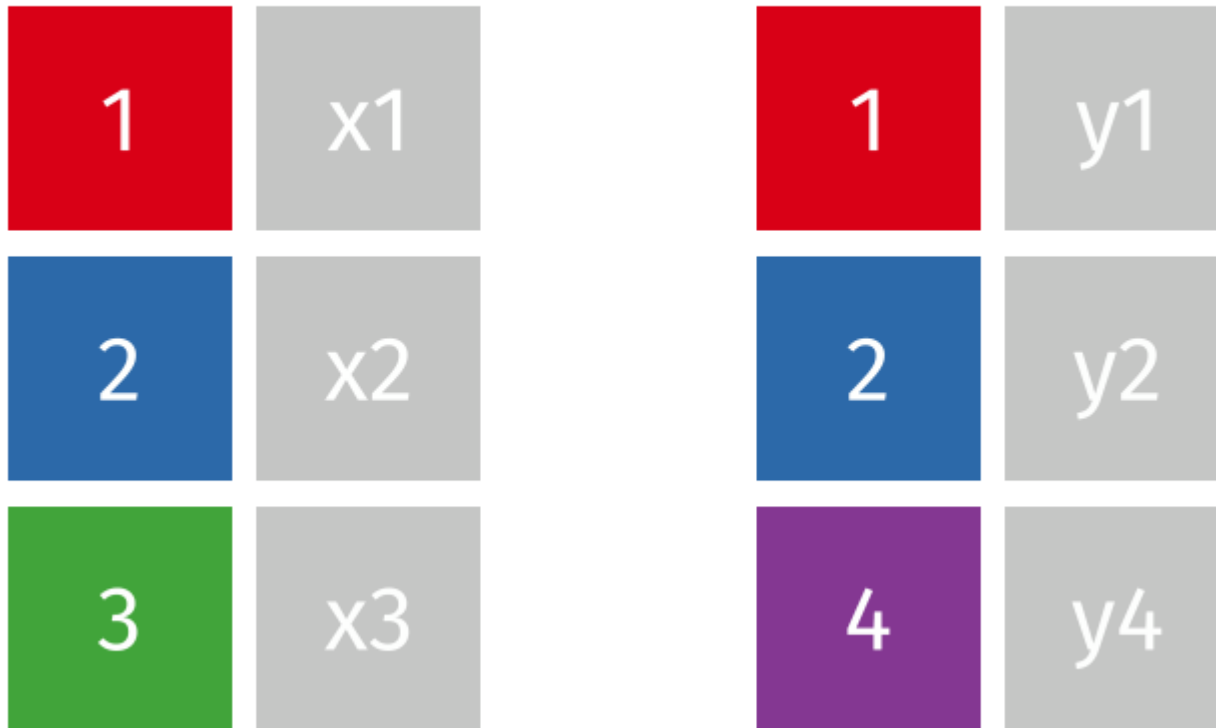
# Right join

`right_join(x, y)`



# Semi join

`semi_join(x, y)`



# Union join

# Anti join

# Character manipulation



# Select rows of papers with > 3 authors

```
citations %>%  
#str_detect() detect characters in a given column  
  filter(str_detect(Authors, 'et al'))
```

```
## # A tibble: 1,280 × 12  
##   journal impac...1 pubyear Volume Issue Authors colld...2 pubdate nbtwe...3 M  
##   <fct>      <dbl>   <dbl>   <dbl> <chr> <chr>   <chr>   <chr>      <dbl>  
## 1 Ecology...  16.7    2014    17 12  Morin ... 2/1/20... 9/16/2...    18  
## 2 Ecology...  16.7    2014    17 12  Jucker... 2/1/20... 10/13/...    15  
## 3 Ecology...  16.7    2014    17 12  Calcag... 2/1/20... 10/21/...     5  
## 4 Ecology...  16.7    2014    17 11  Segre ... 2/1/20... 8/28/2...     9  
## 5 Ecology...  16.7    2014    17 11  Kaufma... 2/1/20... 8/28/2...     3  
## 6 Ecology...  16.7    2014    17 10  Nasto ... 2/2/20... 7/28/2...    27  
## 7 Ecology...  16.7    2014    17 10  Tschir... 2/2/20... 8/6/20...     6  
## 8 Ecology...  16.7    2014    17 9    Barnece... 2/2/20... 6/17/2...    19  
## 9 Ecology...  16.7    2014    17 9    Pinto-... 2/2/20... 6/12/2...    26  
## 10 Ecology... 16.7    2014    17 9    Clough... 2/2/20... 7/17/2...    44  
## # ... with 1,270 more rows, 2 more variables: `Twitter reach` <dbl>,  
## #   woscitations <dbl>, and abbreviated variable names 1impactfactor,  
## #   2colldate, 3nbtweets, 4`Number of users`  
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all var
```

# Select columns with rows of papers with > 3 authors

```
citations %>%  
  filter(str_detect(Authors, 'et al')) %>%  
  select(Authors)
```

```
## # A tibble: 1,280 × 1  
##   Authors  
##   <chr>  
## 1 Morin et al  
## 2 Jucker et al  
## 3 Calcagno et al  
## 4 Segre et al  
## 5 Kaufman et al  
## 6 Nasto et al  
## 7 Tschirren et al  
## 8 Barnechi et al  
## 9 Pinto-Sanchez et al  
## 10 Clough et al  
## # ... with 1,270 more rows  
## # i Use `print(n = ...)`` to see more rows
```

# Select columns with rows of papers with < 3 authors

```
citations %>%  
  filter(!str_detect(Authors, 'et al')) %>% ##! for saying "not".  
  select(Authors)
```

```
## # A tibble: 319 × 1  
##   Authors  
##   <chr>  
## 1 Neutle and Thorne  
## 2 Kellner and Asner  
## 3 Griffin and Willi  
## 4 Gremer and Venable  
## 5 Cavieres  
## 6 Haegman and Loreau  
## 7 Kearney  
## 8 Locey and White  
## 9 Quintero and Weins  
## 10 Lesser and Jackson  
## # ... with 309 more rows  
## # i Use `print(n = ...)` to see more rows
```

# Select authors of columns with rows of papers with < 3 authors

```
citations %>%  
  filter(!str_detect(Authors, 'et al')) %>% ##! for saying "not".  
  pull(Authors) %>%  
  head(10)
```

```
## [1] "Neutle and Thorne" "Kellner and Asner" "Griffin and Willi"  
## [4] "Gremer and Venable" "Cavieres" "Haegman and Loreau"  
## [7] "Kearney" "Locey and White" "Quintero and Weins"  
## [10] "Lesser and Jackson"
```

# Rows of papers with less than 3 authors in journal with IF < 5

```
citations %>%
  filter(!str_detect(Authors, 'et al'), impactfactor < 5)
```

```
## # A tibble: 77 × 12
##   journal  impac...1 pubyear Volume Issue Authors collid...2 pubdate nbtwe...3 M
##   <fct>      <dbl>   <dbl>   <dbl> <chr> <chr>   <chr>   <chr>      <dbl>
## 1 Molecul...   4.9     2014     14  6    Gautier 2/27/2... 5/14/2...      2
## 2 Molecul...   4.9     2014     14  5    Gambel... 2/27/2... 3/7/20...      7
## 3 Molecul...   4.9     2014     14  4    Kekkon... 2/27/2... 3/10/2...      4
## 4 Molecul...   4.9     2014     14  3    Bhatta... 2/27/2... 12/8/2...      0
## 5 Molecul...   4.9     2014     14  1    Christ... 2/28/2... 10/25/...      0
## 6 Molecul...   4.9     2013     13  4    Villar... 2/28/2... 5/2/20...      0
## 7 Molecul...   4.9     2013     13  4    Wang      2/28/2... 4/25/2...      0
## 8 Molecul...   4.9     2012     12  1    Joly      2/28/2... 9/7/20...      3
## 9 Animal ...   3.21    2014     17  6    Plavsic 2/9/20... 4/17/2...      9
## 10 Animal ...   3.21    2014     17  Supp... Knox a... 2/11/2... 11/13/...      1
## # ... with 67 more rows, 2 more variables: `Twitter reach` <dbl>,
## #   woscitations <dbl>, and abbreviated variable names 1impactfactor,
## #   2collidate, 3nbtweets, 4`Number of users`
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all var
```

# Convert words to lowercase

```
citations %>%  
  mutate(authors_lowercase = str_to_lower(Authors)) %>%  
  select(authors_lowercase)
```

```
## # A tibble: 1,599 × 1  
##   authors_lowercase  
##   <chr>  
## 1 morin et al  
## 2 jucker et al  
## 3 calcagno et al  
## 4 segre et al  
## 5 kaufman et al  
## 6 nasto et al  
## 7 tschirren et al  
## 8 barnechi et al  
## 9 pinto-sanchez et al  
## 10 clough et al  
## # ... with 1,589 more rows  
## # i Use `print(n = ...)` to see more rows
```

# Remove all spaces in variable names

```
citations%>%  
  mutate(journal = str_remove_all(journal, " ")) %>%  
  select(journal) %>%  
  unique() %>%  
  head(5)
```

```
## # A tibble: 5 × 1  
##   journal  
##   <chr>  
## 1 EcologyLetters  
## 2 GlobalChangeBiology  
## 3 GlobalEcologyandBiogeography  
## 4 MolecularEcologyResources  
## 5 DiversityandDistributions
```

# Basic exploratory data analysis



# Count ()

This helps to count the number of occurrences.

```
citations %>%  
  count(journal, sort = TRUE) ## Embedded sorting within count()
```

```
## # A tibble: 20 × 2  
##   journal      n  
##   <fct>      <int>  
## 1 New Phytologist      144  
## 2 Ecology              108  
## 3 Evolution             108  
## 4 Global Change Biology 108  
## 5 Global Ecology and Biogeography 108  
## 6 Journal of Biogeography 108  
## 7 Ecology Letters      106  
## 8 Diversity and Distributions 105  
## 9 Animal Conservation   102  
## 10 Methods in Ecology and Evolution 90  
## 11 Evolutionary Applications 74  
## 12 Functional Ecology    54  
## 13 Journal of Animal Ecology 54  
## 14 Journal of Applied Ecology 54
```

# Count() for multiple variables

```
citations %>%  
  count(journal, pubyear)
```

```
## # A tibble: 59 × 3  
##   journal                pubyear      n  
##   <fct>                 <dbl> <int>  
## 1 Animal Conservation    2012     18  
## 2 Animal Conservation    2013     18  
## 3 Animal Conservation    2014     66  
## 4 Conservation Letters   2012     17  
## 5 Conservation Letters   2013     18  
## 6 Conservation Letters   2014     18  
## 7 Diversity and Distributions 2012     36  
## 8 Diversity and Distributions 2013     33  
## 9 Diversity and Distributions 2014     36  
## 10 Ecological Applications  2012     24  
## # ... with 49 more rows  
## # i Use `print(n = ...)` to see more rows
```

# Count sum of tweets per journal

```
citations %>%  
  count(journal, wt = nbtweets, sort = TRUE)
```

```
## # A tibble: 20 × 2  
##   journal          n  
##   <fct>          <dbl>  
## 1 Ecology Letters 1538  
## 2 Animal Conservation 1268  
## 3 Journal of Applied Ecology 1012  
## 4 Methods in Ecology and Evolution 699  
## 5 Global Change Biology 613  
## 6 Conservation Letters 542  
## 7 New Phytologist 509  
## 8 Global Ecology and Biogeography 379  
## 9 Ecology 335  
## 10 Evolution 335  
## 11 Journal of Animal Ecology 323  
## 12 Fish and Fisheries 261  
## 13 Evolutionary Applications 238  
## 14 Journal of Biogeography 209  
## 15 Diversity and Distributions 200  
## 16 Mammal Review 166
```

# Group variables to compute stats [summarise()]

```
citations %>%  
  group_by(journal) %>%  
  summarise(avg_tweets = mean(nbtweets))
```

```
## # A tibble: 20 × 2  
##   journal          avg_tweets  
##   <fct>          <dbl>  
## 1 Animal Conservation    12.4  
## 2 Conservation Letters  10.2  
## 3 Diversity and Distributions  1.90  
## 4 Ecological Applications  2.60  
## 5 Ecology               3.10  
## 6 Ecology Letters      14.5  
## 7 Evolution             3.10  
## 8 Evolutionary Applications  3.22  
## 9 Fish and Fisheries     7.25  
## 10 Functional Ecology     2.87  
## 11 Global Change Biology   5.68  
## 12 Global Ecology and Biogeography  3.51  
## 13 Journal of Animal Ecology  5.98
```

# Order stuff [arrange()]

```
citations %>%  
  group_by(journal) %>%  
  summarise(avg_tweets = mean(nbtweets)) %>%  
  # decreasing order but (without desc for increasing)  
  arrange(desc(avg_tweets)) -> arrangedat  
head(arrangedat, 10)
```

```
## # A tibble: 10 × 2  
##   journal                avg_tweets  
##   <fct>                  <dbl>  
## 1 Journal of Applied Ecology 18.7  
## 2 Ecology Letters          14.5  
## 3 Animal Conservation       12.4  
## 4 Conservation Letters     10.2  
## 5 Methods in Ecology and Evolution 7.77  
## 6 Fish and Fisheries        7.25  
## 7 Journal of Animal Ecology  5.98  
## 8 Global Change Biology      5.68  
## 9 Mammal Review             5.35  
## 10 New Phytologist          3.53
```


# Work on several columns [dplyr::across()]

**dplyr::across()**

use within `mutate()`  
or `summarize()` to  
apply function(s) to  
a selection of columns!

EXAMPLE:

```
df %>%  
  group_by(species) %>%  
  summarize(  
    across(where(is.numeric), mean)  
  )
```



species	mass_g	age_yr	range_sqmi
pika	163	2.4	0.46
marmot	1509	3.0	0.87
marmot	2417	5.6	0.62

@allison\_horst

# Compute mean across multiple variables

```
citations %>%  
  group_by(journal) %>%  
  summarize(across(where(is.numeric), mean))
```

```
## # A tibble: 20 × 8
```

##	journal	impac... <sup>1</sup>	pubyear	Volume	nbtwe... <sup>2</sup>	Numbe... <sup>3</sup>	Twitt... <sup>4</sup>	w
##	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
##	1 Animal Conservation	3.21	2013.	16.5	12.4	9.71	28345.	
##	2 Conservation Letters	6.4	2013.	6.02	10.2	8.85	23234.	
##	3 Diversity and Distrib...	5.4	2013	19	1.90	1.77	2350.	
##	4 Ecological Applicatio...	5.06	2013	23	2.60	2.5	5727.	
##	5 Ecology	6.16	2013	94	3.10	2.87	6176.	
##	6 Ecology Letters	16.7	2013.	16.0	14.5	14.0	44748.	
##	7 Evolution	5.25	2013	67	3.10	2.93	7762.	
##	8 Evolutionary Applicat...	4.6	2013.	6.05	3.22	3.07	13185.	
##	9 Fish and Fisheries	8.1	2013	14	7.25	6.19	12097.	
##	10 Functional Ecology	5.28	2013	27	2.87	2.74	3809.	
##	11 Global Change Biology	8.7	2013	19	5.68	4.94	9652.	
##	12 Global Ecology and Bi...	7.18	2013	22	3.51	3.15	8995.	
##	13 Journal of Animal Eco...	5.32	2013.	81.9	5.98	5.59	36112.	
##	14 Journal of Applied Ec...	5.93	2013	50	18.7	15.8	43839.	
##	15 Journal of Biogeograp...	4.59	2013	40	1.94	1.86	116328	

# Tidying tibbles [wide(), long()]

wide

id	wide		
	x	y	z
1	a	c	e
2	b	d	f



# Data manipulation with tidyverse: in depth study

Learn the tidyverse: books, workshops and online courses Selection of books:

- [R for Data Science](#) and [Advanced R](#);
- [Tidy Tuesdays videos](#) by D. Robinson;
- Material of the [stat545](#) course on Data wrangling, exploration, and analysis with R at the University of British Columbia;
- List of best R packages (with their description) on data import, [wrangling and visualization](#).

# Thank you for listening!

Any questions now or email me at [dossa@xtbg.org.cn](mailto:dossa@xtbg.org.cn)

Slides created via the R package **xaringan**.

The chakra comes from **remark.js**, **knitr**, and **R Markdown**.