Lecture 4: R Packages: Tidyverse

Peng Chap 4-6

Ailin Zhang

R Packages

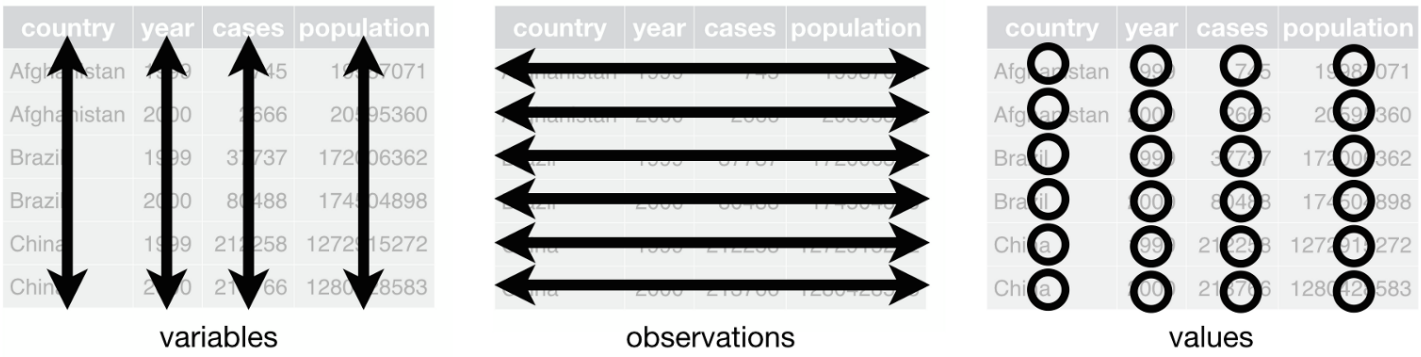
- For those new to R and arriving with interest in DS, the tidyverse is the key set of packages
- The tidyverse includes R packages that help facilitate modern stats+DS

Tidyverse is a collection of R packages

- `ggplot2` - visualization
- `dplyr`, `tidyr` - data manipulation
- `purrr` - advanced programming
- `readr` - import data
- `tibble` - improved data.frame format
- `forcats` - working w/ factors
- `stringr` - working w/ chain of characters

Tidy Data

- Main characteristics of a tidy dataset:
  - each variable is a column
  - each observation is a raw
  - each value is in a different cell



Load tidyverse

```
#install.packages("tidyverse")
library(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
- tibble is a modern reimagining of the data frame.
- fast

Import data

```
citations_raw <- read_csv('citation.csv')
citations_raw
```

```
## # A tibble: 1,599 x 12
##    Journa~1 5-yea~2 Year ~3 Volume Issue Authors Colle~4 Publi~5 Numbe~6 Numbe~7
##    <chr>      <dbl>   <dbl>  <dbl> <chr> <chr>   <chr>   <chr>     <dbl>   <dbl>
##  1 Ecology~    16.7    2014     17 12    Morin ~ 2/1/20~ 9/16/2~      18      16
##  2 Ecology~    16.7    2014     17 12    Jucker~ 2/1/20~ 10/13/~      15      12
##  3 Ecology~    16.7    2014     17 12    Calcag~ 2/1/20~ 10/21/~       5       4
##  4 Ecology~    16.7    2014     17 11    Segre ~ 2/1/20~ 8/28/2~       9       8
##  5 Ecology~    16.7    2014     17 11    Kaufma~ 2/1/20~ 8/28/2~       3       3
##  6 Ecology~    16.7    2014     17 10    Nasto ~ 2/2/20~ 7/28/2~      27      23
##  7 Ecology~    16.7    2014     17 10    Tschir~ 2/2/20~ 8/6/20~       6       6
##  8 Ecology~    16.7    2014     17 9     Barnec~ 2/2/20~ 6/17/2~      19      18
##  9 Ecology~    16.7    2014     17 9     Pinto-~ 2/2/20~ 6/12/2~      26      23
## 10 Ecology~    16.7    2014     17 9     Clough~ 2/2/20~ 7/17/2~      44      42
## # ... with 1,589 more rows, 2 more variables: `Twitter reach` <dbl>,
## #   `Number of Web of Science citations` <dbl>, and abbreviated variable names
## #   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`
```

Rename columns

```
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')
citations_temp
```

```
## # A tibble: 1,599 x 12
##    journal  impac~1 pubyear Volume Issue Authors colld~2 pubdate nbtwe~3 Numbe~4
##    <chr>      <dbl>   <dbl>  <dbl> <chr> <chr>   <chr>   <chr>     <dbl>   <dbl>
##  1 Ecology~    16.7    2014     17 12    Morin ~ 2/1/20~ 9/16/2~      18      16
##  2 Ecology~    16.7    2014     17 12    Jucker~ 2/1/20~ 10/13/~      15      12
##  3 Ecology~    16.7    2014     17 12    Calcag~ 2/1/20~ 10/21/~       5       4
##  4 Ecology~    16.7    2014     17 11    Segre ~ 2/1/20~ 8/28/2~       9       8
##  5 Ecology~    16.7    2014     17 11    Kaufma~ 2/1/20~ 8/28/2~       3       3
##  6 Ecology~    16.7    2014     17 10    Nasto ~ 2/2/20~ 7/28/2~      27      23
##  7 Ecology~    16.7    2014     17 10    Tschir~ 2/2/20~ 8/6/20~       6       6
##  8 Ecology~    16.7    2014     17 9     Barnec~ 2/2/20~ 6/17/2~      19      18
##  9 Ecology~    16.7    2014     17 9     Pinto-~ 2/2/20~ 6/12/2~      26      23
## 10 Ecology~    16.7    2014     17 9     Clough~ 2/2/20~ 7/17/2~      44      42
## # ... with 1,589 more rows, 2 more variables: `Twitter reach` <dbl>,
## #   woscitations <dbl>, and abbreviated variable names 1: impactfactor,
## #   2: colldate, 3: nbtweets, 4: `Number of users`
```

Create (or modify) columns

```
citations <- mutate(citations_temp, journal = as.factor(journal))
citations
```

```
## # A tibble: 1,599 x 12
##    journal  impac~1 pubyear Volume Issue Authors colld~2 pubdate nbtwe~3 Numbe~4
##    <fct>      <dbl>   <dbl>  <dbl> <chr> <chr>   <chr>   <chr>     <dbl>   <dbl>
##  1 Ecology~    16.7    2014     17 12    Morin ~ 2/1/20~ 9/16/2~      18      16
##  2 Ecology~    16.7    2014     17 12    Jucker~ 2/1/20~ 10/13/~      15      12
##  3 Ecology~    16.7    2014     17 12    Calcag~ 2/1/20~ 10/21/~       5       4
##  4 Ecology~    16.7    2014     17 11    Segre ~ 2/1/20~ 8/28/2~       9       8
##  5 Ecology~    16.7    2014     17 11    Kaufma~ 2/1/20~ 8/28/2~       3       3
##  6 Ecology~    16.7    2014     17 10    Nasto ~ 2/2/20~ 7/28/2~      27      23
##  7 Ecology~    16.7    2014     17 10    Tschir~ 2/2/20~ 8/6/20~       6       6
##  8 Ecology~    16.7    2014     17 9     Barnec~ 2/2/20~ 6/17/2~      19      18
##  9 Ecology~    16.7    2014     17 9     Pinto-~ 2/2/20~ 6/12/2~      26      23
## 10 Ecology~    16.7    2014     17 9     Clough~ 2/2/20~ 7/17/2~      44      42
## # ... with 1,589 more rows, 2 more variables: `Twitter reach` <dbl>,
## #   woscitations <dbl>, and abbreviated variable names 1: impactfactor,
## #   2: colldate, 3: nbtweets, 4: `Number of users`
```

Create (or modify) columns

```
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"
```

Cleaner code with "pipe" operator %>%

```
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))
```

```
## # A tibble: 1,599 x 12
##    journal  impac~1 pubyear Volume Issue Authors colld~2 pubdate nbtwe~3 Numbe~4
##    <fct>      <dbl>   <dbl>  <dbl> <chr> <chr>   <chr>   <chr>     <dbl>   <dbl>
##  1 Ecology~    16.7    2014     17 12    Morin ~ 2/1/20~ 9/16/2~      18      16
##  2 Ecology~    16.7    2014     17 12    Jucker~ 2/1/20~ 10/13/~      15      12
##  3 Ecology~    16.7    2014     17 12    Calcag~ 2/1/20~ 10/21/~       5       4
##  4 Ecology~    16.7    2014     17 11    Segre ~ 2/1/20~ 8/28/2~       9       8
##  5 Ecology~    16.7    2014     17 11    Kaufma~ 2/1/20~ 8/28/2~       3       3
##  6 Ecology~    16.7    2014     17 10    Nasto ~ 2/2/20~ 7/28/2~      27      23
##  7 Ecology~    16.7    2014     17 10    Tschir~ 2/2/20~ 8/6/20~       6       6
##  8 Ecology~    16.7    2014     17 9     Barnec~ 2/2/20~ 6/17/2~      19      18
##  9 Ecology~    16.7    2014     17 9     Pinto-~ 2/2/20~ 6/12/2~      26      23
## 10 Ecology~    16.7    2014     17 9     Clough~ 2/2/20~ 7/17/2~      44      42
## # ... with 1,589 more rows, 2 more variables: `Twitter reach` <dbl>,
## #   woscitations <dbl>, and abbreviated variable names 1: impactfactor,
## #   2: colldate, 3: nbtweets, 4: `Number of users`
```

Name object

```
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))
```

Syntax with pipe

```
df %>%
  do_this_operation %>%
  then_do_this_operation %>%
  then_do_this_operation ...
```

- The pipe operator simply feeds the results of one operation into the next operation below it.
- No need to name unimportant intermediate variables
- Clear syntax (readability)

Select columns

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

```
## # A tibble: 1,599 x 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
```

Drop columns

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

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

Split a column in several columns

```
citations %>%
  separate(pubdate,c('month','day','year'),'/')
```

```
## # A tibble: 1,599 x 14
##    journal         impac~1 pubyear Volume Issue Authors colld~2 month day   year
##    <fct>             <dbl>   <dbl>  <dbl> <chr> <chr>   <chr>   <chr> <chr> <chr>
##  1 Ecology Lette~    16.7     2014     17 12    Morin ~ 2/1/20~ 9     16    2014
##  2 Ecology Lette~    16.7     2014     17 12    Jucker~ 2/1/20~ 10    13    2014
##  3 Ecology Lette~    16.7     2014     17 12    Calcag~ 2/1/20~ 10    21    2014
##  4 Ecology Lette~    16.7     2014     17 11    Segre ~ 2/1/20~ 8     28    2014
##  5 Ecology Lette~    16.7     2014     17 11    Kaufma~ 2/1/20~ 8     28    2014
##  6 Ecology Lette~    16.7     2014     17 10    Nasto ~ 2/2/20~ 7     28    2014
##  7 Ecology Lette~    16.7     2014     17 10    Tschir~ 2/2/20~ 8     6     2014
##  8 Ecology Lette~    16.7     2014     17 9     Barnec~ 2/2/20~ 6     17    2014
##  9 Ecology Lette~    16.7     2014     17 9     Pinto-~ 2/2/20~ 6     12    2014
## 10 Ecology Lette~    16.7     2014     17 9     Clough~ 2/2/20~ 7     17    2014
## # ... with 1,589 more rows, 4 more variables: nbtweets <dbl>,
## #   `Number of users` <dbl>, `Twitter reach` <dbl>, woscitations <dbl>, and
## #   abbreviated variable names 1: impactfactor, 2: colldate
```

Transform in Date format...

```
library(lubridate)
citations %>%
  mutate(pubdate = mdy(pubdate),
         colldate = mdy(colldate))
```

```
## # A tibble: 1,599 x 12
##    journal    impac~1 pubyear Volume Issue Authors colldate   pubdate    nbtwe~2
##    <fct>        <dbl>   <dbl>  <dbl> <chr> <chr>   <date>     <date>       <dbl>
##  1 Ecology L~    16.7    2014     17 12    Morin ~ 2016-02-01 2014-09-16      18
##  2 Ecology L~    16.7    2014     17 12    Jucker~ 2016-02-01 2014-10-13      15
##  3 Ecology L~    16.7    2014     17 12    Calcag~ 2016-02-01 2014-10-21       5
##  4 Ecology L~    16.7    2014     17 11    Segre ~ 2016-02-01 2014-08-28       9
##  5 Ecology L~    16.7    2014     17 11    Kaufma~ 2016-02-01 2014-08-28       3
##  6 Ecology L~    16.7    2014     17 10    Nasto ~ 2016-02-02 2014-07-28      27
##  7 Ecology L~    16.7    2014     17 10    Tschir~ 2016-02-02 2014-08-06       6
##  8 Ecology L~    16.7    2014     17 9     Barnec~ 2016-02-02 2014-06-17      19
##  9 Ecology L~    16.7    2014     17 9     Pinto-~ 2016-02-02 2014-06-12      26
## 10 Ecology L~    16.7    2014     17 9     Clough~ 2016-02-02 2014-07-17      44
## # ... with 1,589 more rows, 3 more variables: `Number of users` <dbl>,
## #   `Twitter reach` <dbl>, woscitations <dbl>, and abbreviated variable names
## #   1: impactfactor, 2: nbtweets
```

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

Select rows corresponding to papers with more than 3 authors

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

```
## # A tibble: 1,280 x 12
##    journal  impac~1 pubyear Volume Issue Authors colld~2 pubdate nbtwe~3 Numbe~4
##    <fct>      <dbl>   <dbl>  <dbl> <chr> <chr>   <chr>   <chr>     <dbl>   <dbl>
##  1 Ecology~    16.7    2014     17 12    Morin ~ 2/1/20~ 9/16/2~      18      16
##  2 Ecology~    16.7    2014     17 12    Jucker~ 2/1/20~ 10/13/~      15      12
##  3 Ecology~    16.7    2014     17 12    Calcag~ 2/1/20~ 10/21/~       5       4
##  4 Ecology~    16.7    2014     17 11    Segre ~ 2/1/20~ 8/28/2~       9       8
##  5 Ecology~    16.7    2014     17 11    Kaufma~ 2/1/20~ 8/28/2~       3       3
##  6 Ecology~    16.7    2014     17 10    Nasto ~ 2/2/20~ 7/28/2~      27      23
##  7 Ecology~    16.7    2014     17 10    Tschir~ 2/2/20~ 8/6/20~       6       6
##  8 Ecology~    16.7    2014     17 9     Barnec~ 2/2/20~ 6/17/2~      19      18
##  9 Ecology~    16.7    2014     17 9     Pinto-~ 2/2/20~ 6/12/2~      26      23
## 10 Ecology~    16.7    2014     17 9     Clough~ 2/2/20~ 7/17/2~      44      42
## # ... with 1,270 more rows, 2 more variables: `Twitter reach` <dbl>,
## #   woscitations <dbl>, and abbreviated variable names 1: impactfactor,
## #   2: colldate, 3: nbtweets, 4: `Number of users`
```

Get column with rows corresponding to papers with more than 3 authors

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

```
## # A tibble: 1,280 x 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
```

Select rows corresponding to papers with less than 3 authors

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

```
## # A tibble: 319 x 12
##    journal  impac~1 pubyear Volume Issue Authors colld~2 pubdate nbtwe~3 Numbe~4
##    <fct>      <dbl>   <dbl>  <dbl> <chr> <chr>   <chr>   <chr>     <dbl>   <dbl>
##  1 Ecology~    16.7    2014     17 6     Neutle~ 2/15/2~ 3/17/2~       8       7
##  2 Ecology~    16.7    2014     17 5     Kellne~ 2/15/2~ 2/20/2~      18      18
##  3 Ecology~    16.7    2014     17 4     Griffi~ 2/15/2~ 1/16/2~       4       4
##  4 Ecology~    16.7    2014     17 3     Gremer~ 2/15/2~ 1/17/2~       4       4
##  5 Ecology~    16.7    2014     17 2     Cavier~ 2/15/2~ 10/17/~      16      15
##  6 Ecology~    16.7    2014     17 2     Haegma~ 2/15/2~ 12/5/2~       9       9
##  7 Ecology~    16.7    2013     16 12    Kearney 2/15/2~ 10/1/2~      13      13
##  8 Ecology~    16.7    2013     16 9     Locey ~ 2/15/2~ 7/15/2~      28      24
##  9 Ecology~    16.7    2013     16 8     Quinte~ 2/15/2~ 6/26/2~     120     120
## 10 Ecology~    16.7    2013     16 3     Lesser~ 2/15/2~ 12/22/~       9       9
## # ... with 309 more rows, 2 more variables: `Twitter reach` <dbl>,
## #   woscitations <dbl>, and abbreviated variable names 1: impactfactor,
## #   2: colldate, 3: nbtweets, 4: `Number of users`
```

Get column with rows corresponding to papers with less than 3 authors

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

```
## # A tibble: 319 x 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
```

Get column with rows corresponding to papers with less than 3 authors

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

Select rows corresponding to papers with less than 3 authors in journal with IF < 5

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

```
## # A tibble: 77 x 12
##    journal  impac~1 pubyear Volume Issue Authors colld~2 pubdate nbtwe~3 Numbe~4
##    <fct>      <dbl>   <dbl>  <dbl> <chr> <chr>   <chr>   <chr>     <dbl>   <dbl>
##  1 Molecul~    4.9     2014     14 6     Gautier 2/27/2~ 5/14/2~       2       2
##  2 Molecul~    4.9     2014     14 5     Gambel~ 2/27/2~ 3/7/20~       7       5
##  3 Molecul~    4.9     2014     14 4     Kekkon~ 2/27/2~ 3/10/2~       4       4
##  4 Molecul~    4.9     2014     14 3     Bhatta~ 2/27/2~ 12/8/2~       0       0
##  5 Molecul~    4.9     2014     14 1     Christ~ 2/28/2~ 10/25/~       0       0
##  6 Molecul~    4.9     2013     13 4     Villar~ 2/28/2~ 5/2/20~       0       0
##  7 Molecul~    4.9     2013     13 4     Wang    2/28/2~ 4/25/2~       0       0
##  8 Molecul~    4.9     2012     12 1     Joly    2/28/2~ 9/7/20~       3       3
##  9 Animal ~    3.21    2014     17 6     Plavsic 2/9/20~ 4/17/2~       9       9
## 10 Animal ~    3.21    2014     17 Supp~ Knox a~ 2/11/2~ 11/13/~       1       1
## # ... with 67 more rows, 2 more variables: `Twitter reach` <dbl>,
## #   woscitations <dbl>, and abbreviated variable names 1: impactfactor,
## #   2: colldate, 3: nbtweets, 4: `Number of users`
```

Convert words to lowercase

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

```
## # A tibble: 1,599 x 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
```

Remove all spaces in journal names

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

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

Explore stringr and regular expressions

- Check out the vignette on stringr for more examples on character manipulation and pattern matching functions.
- Check out the vignette on regular expressions which are a concise and flexible tool for describing patterns in strings.

Count

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

```
## # A tibble: 20 x 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
## 15 Limnology and Oceanography            54
## 16 Molecular Ecology Resources           54
## 17 Conservation Letters                  53
## 18 Ecological Applications               48
## 19 Fish and Fisheries                    36
## 20 Mammal Review                         31
```

Count

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

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

Group by variable to calculate stats

```
citations %>%
  group_by(journal) %>% #<<
  summarise(avg_tweets = mean(nbtweets)) %>% #<<
  head(10)
```

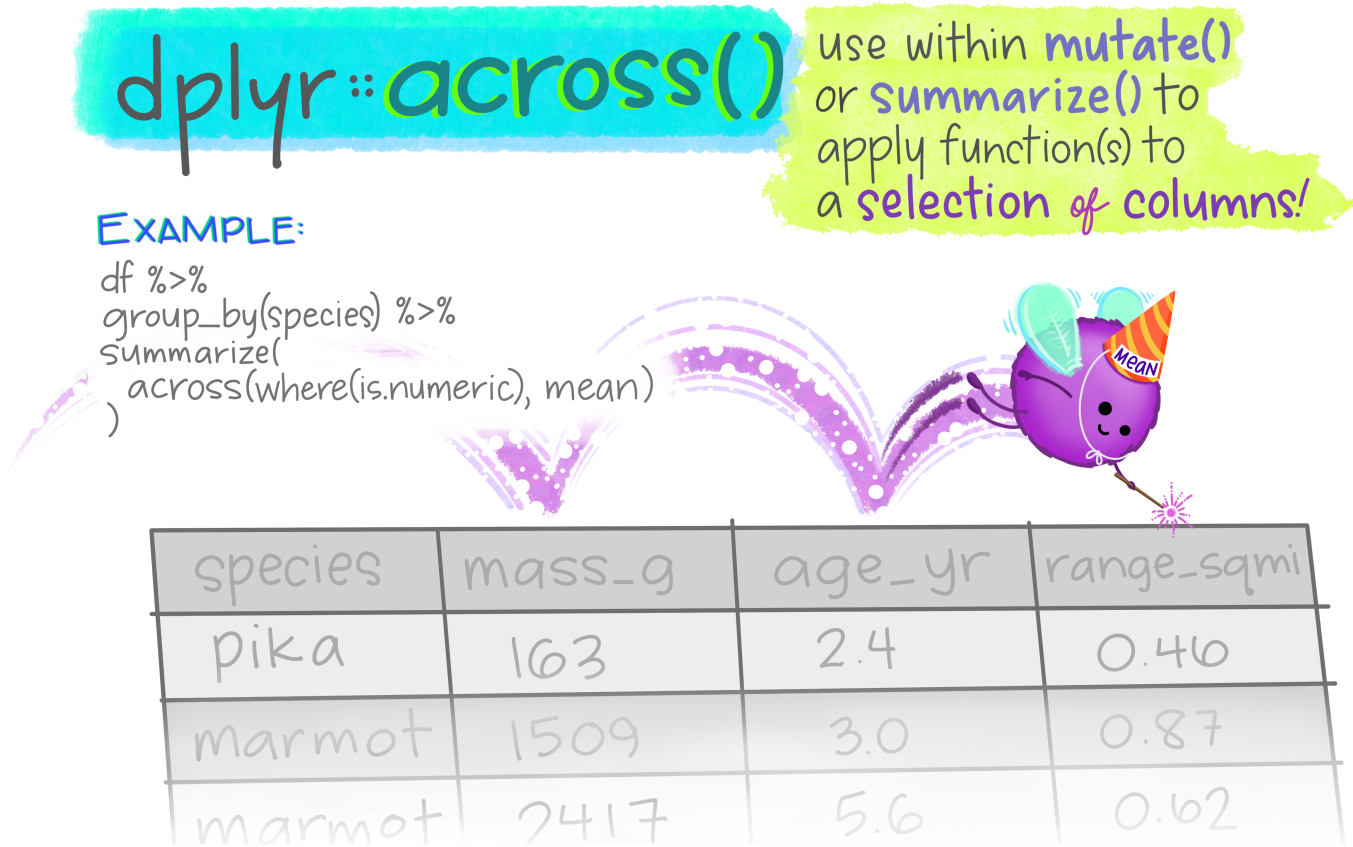
```
## # A tibble: 10 x 2
##    journal                          avg_tweets
##    <fct>                                 <dbl>
##  1 Animal Conservation                    12.4
##  2 Conservation Letters                   10.2
##  3 Diversity and Distributions             1.90
##  4 Ecological Applications                 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
```

Order stuff

```
citations %>%
  group_by(journal) %>%
  summarise(avg_tweets = mean(nbtweets)) %>%
  arrange(desc(avg_tweets)) %>% # decreasing order (wo desc for increasing) #<<
  head(10)
```

```
## # A tibble: 10 x 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
```

What if we want to work on several columns?



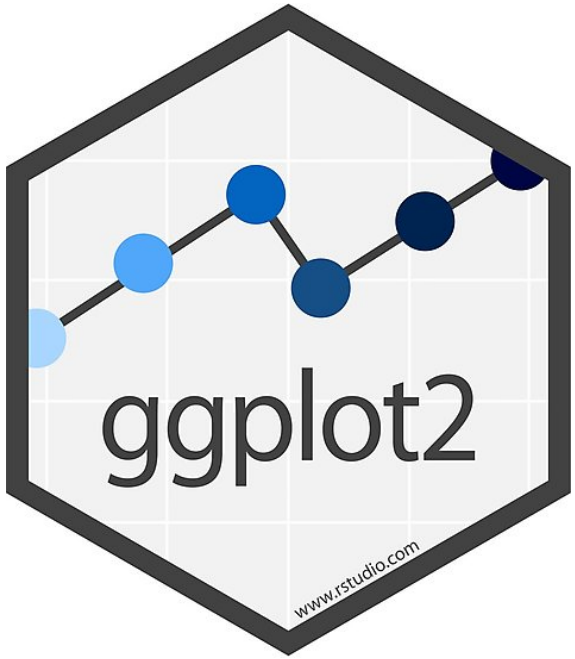
Compute mean across all numeric columns for each journal

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

```
## # A tibble: 6 x 8
##   journal                impac~1 pubyear Volume nbtwe~2 Numbe~3 Twitt~4 wosci~5
##   <fct>                    <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Animal Conservation       3.21   2013.  16.5    12.4     9.71  28345.    4.43
## 2 Conservation Letters      6.4    2013.   6.02   10.2     8.85  23234.    9.30
## 3 Diversity and Distribu~   5.4    2013   19       1.90    1.77   2350.   10.2
## 4 Ecological Applications   5.06   2013   23       2.60    2.5    5727.   10.7
## 5 Ecology                   6.16   2013   94       3.10    2.87   6176.   11.1
## 6 Ecology Letters          16.7    2013.  16.0    14.5    14.0   44748.   20.6
## # ... with abbreviated variable names 1: impactfactor, 2: nbtweets,
## #   3: `Number of users`, 4: `Twitter reach`, 5: woscitations
```

Visualization with ggplot2

- The package ggplot2 implements a **g**rammar of **g**raphics
- Operates on data.frames or tibbles, not vectors like base R
- Explicitly differentiates between the data and its representation



The ggplot2 grammar

Grammar element	What it is
Data	The data frame being plotted
Geometrics (e.g., point, boxplot, histogram)	The geometric shape that will represent th
Aesthetics (e.g., color, size, shape)	The aesthetics of the geometric object

Scatterplots

```
citations %>%  #<<
  ggplot() + #<<
  aes(x = nbtweets, y = woscitations) +
  geom_point()
```

- Pass in the data frame as your first argument

Scatterplots

```
citations %>%
  ggplot() +
  aes(x = nbtweets, y = woscitations) + #<<
  geom_point()
```

- Pass in the data frame as your first argument
- Aesthetics maps the data onto plot characteristics, here x and y axes

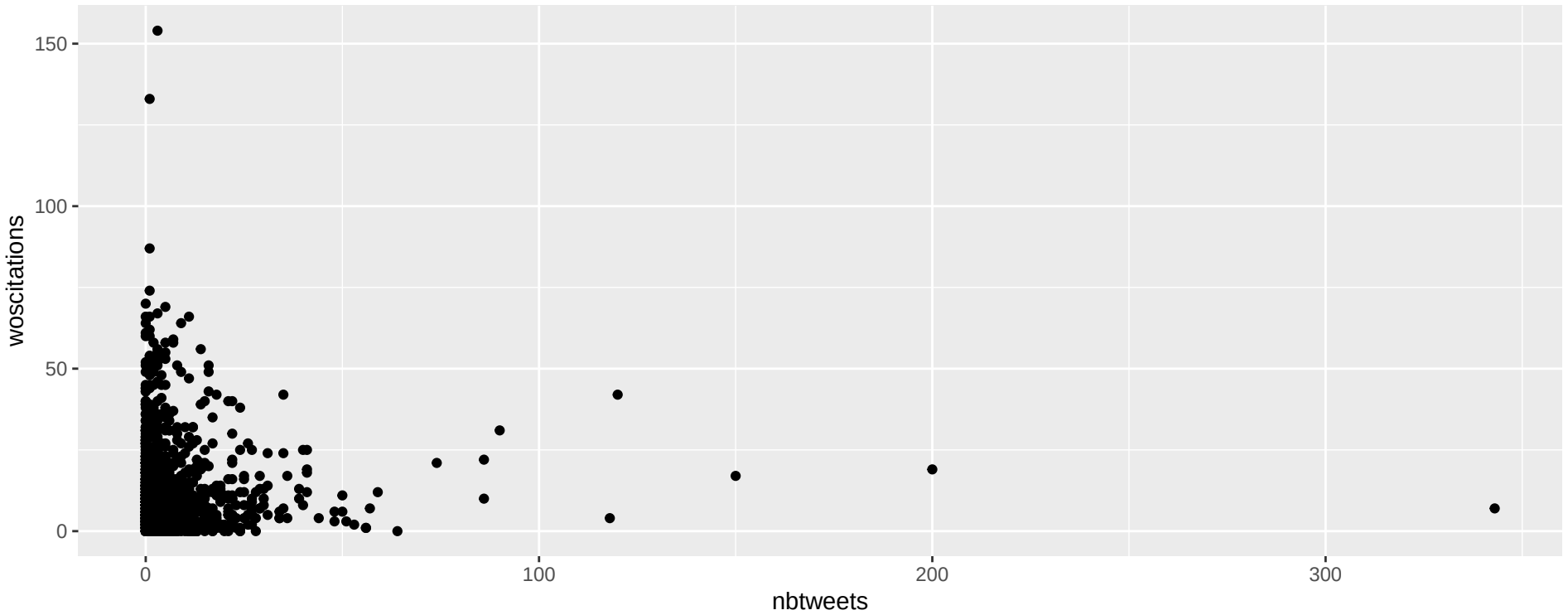
Scatterplots

```
citations %>%
  ggplot() +
  aes(x = nbtweets, y = woscitations) +
  geom_point() #<<
```

- Pass in the data frame as your first argument
- Aesthetics maps the data onto plot characteristics, here x and y axes
- Display the data geometrically as points

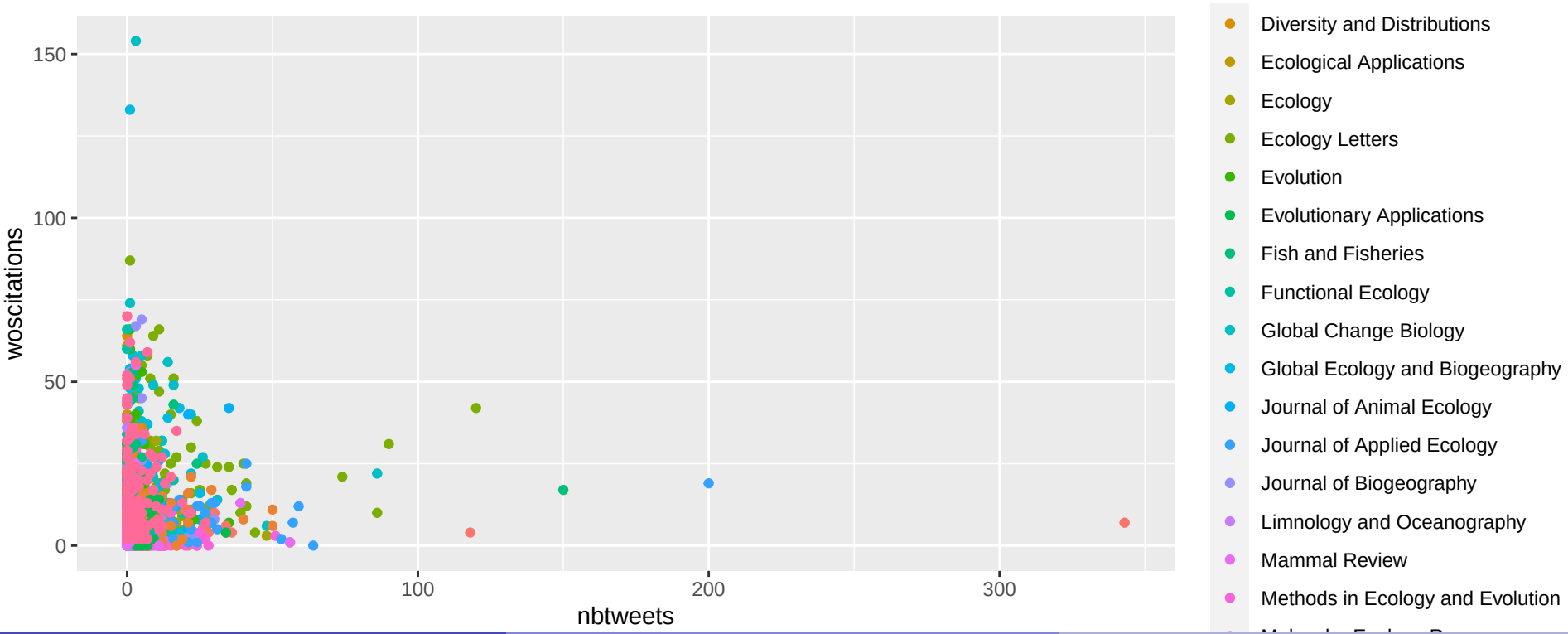
Scatterplots

```
citations %>%
  ggplot() +
  aes(x = nbtweets, y = woscitations) +
  geom_point()
```



Scatterplots, with species-specific colors

```
citations %>%
  ggplot() +
  aes(x = nbtweets, y = woscitations, color = journal) + #<<
  geom_point()
```



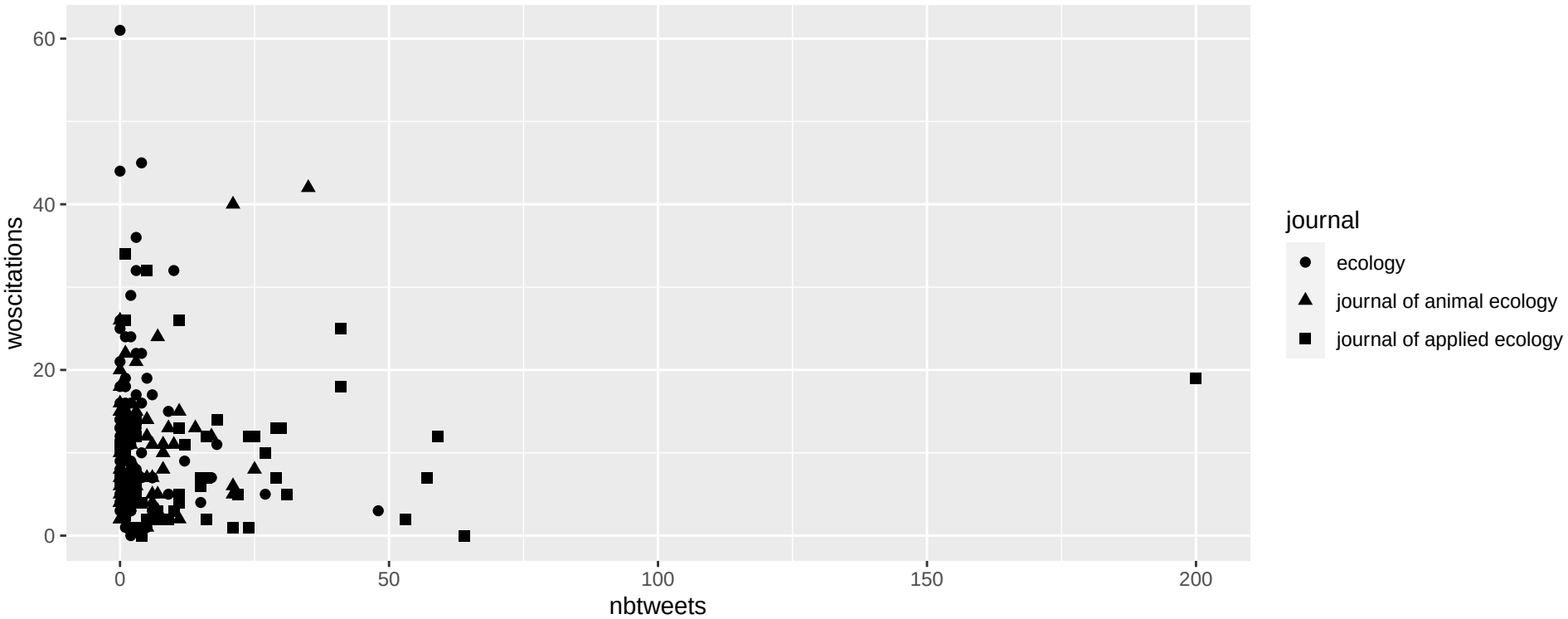
Pick a few journals

```
citations_ecology <- citations %>%
  mutate(journal = str_to_lower(journal)) %>% # all journals names lowercase
  filter(journal %in%
           c('journal of animal ecology','journal of applied ecology','ecology')) #
citations_ecology
```

```
## # A tibble: 216 x 12
##    journal impact~1 pubyear Volume Issue Authors colld~2 pubdate nbtwe~3 Numbe~4
##    <chr>      <dbl>   <dbl>  <dbl> <chr> <chr>   <chr>   <chr>     <dbl>   <dbl>
##  1 ecology     6.16    2014     95 12    Maglia~ 3/19/2~ 12/1/2~       1       1
##  2 ecology     6.16    2014     95 12    Soinen  3/19/2~ 12/1/2~       6       6
##  3 ecology     6.16    2014     95 12    Graham~ 3/19/2~ 12/1/2~       1       1
##  4 ecology     6.16    2014     95 11    White ~ 3/19/2~ 11/1/2~       9       9
##  5 ecology     6.16    2014     95 11    Einars~ 3/19/2~ 11/1/2~      15      12
##  6 ecology     6.16    2014     95 11    Haav a~ 3/19/2~ 11/1/2~       2       2
##  7 ecology     6.16    2014     95 10    Dodds ~ 3/19/2~ 10/1/2~       1       1
##  8 ecology     6.16    2014     95 10    Brown ~ 3/19/2~ 10/1/2~       1       1
##  9 ecology     6.16    2014     95 10    Wright~ 3/19/2~ 10/1/2~       0       0
## 10 ecology     6.16    2014     95 9     Ramahl~ 3/19/2~ 9/1/20~      27      25
## # ... with 206 more rows, 2 more variables: `Twitter reach` <dbl>,
## #   woscitations <dbl>, and abbreviated variable names 1: impactfactor,
## #   2: colldate, 3: nbtweets, 4: `Number of users`
```

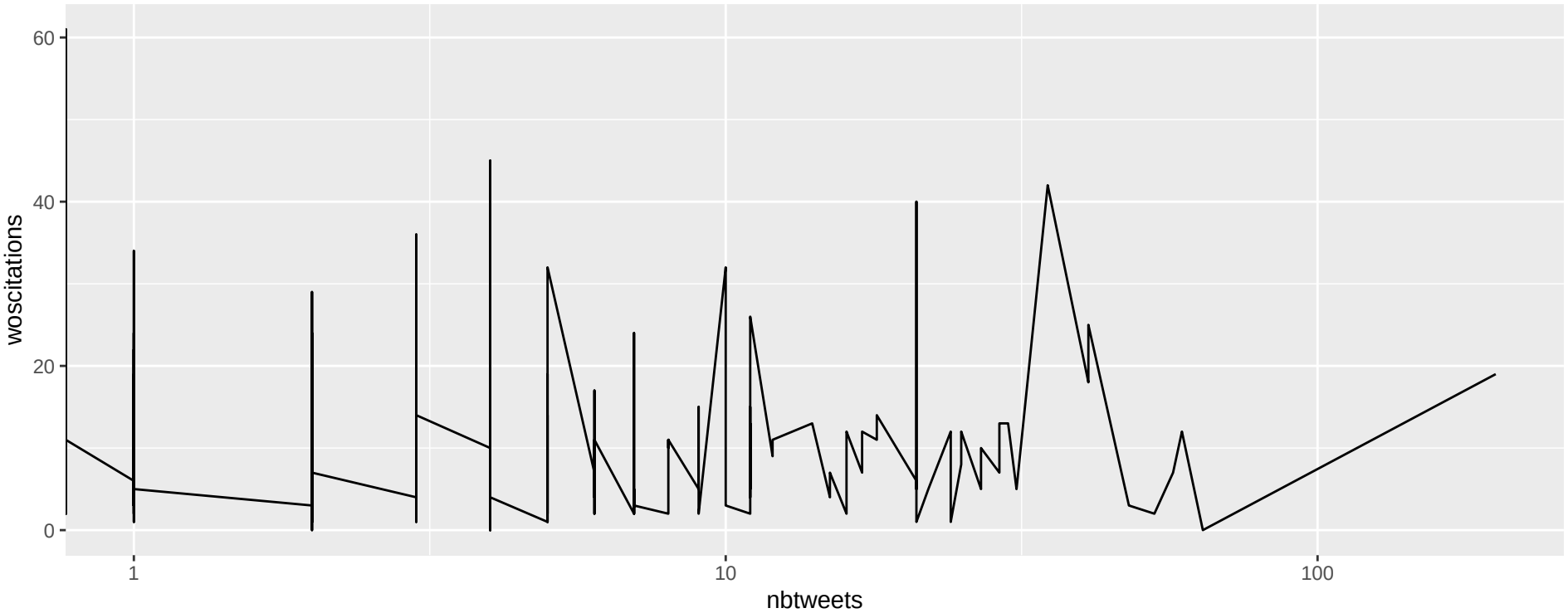
Scatterplots, with species-specific shapes

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets, y = woscitations, shape = journal) + #<<
  geom_point(size=2)
```



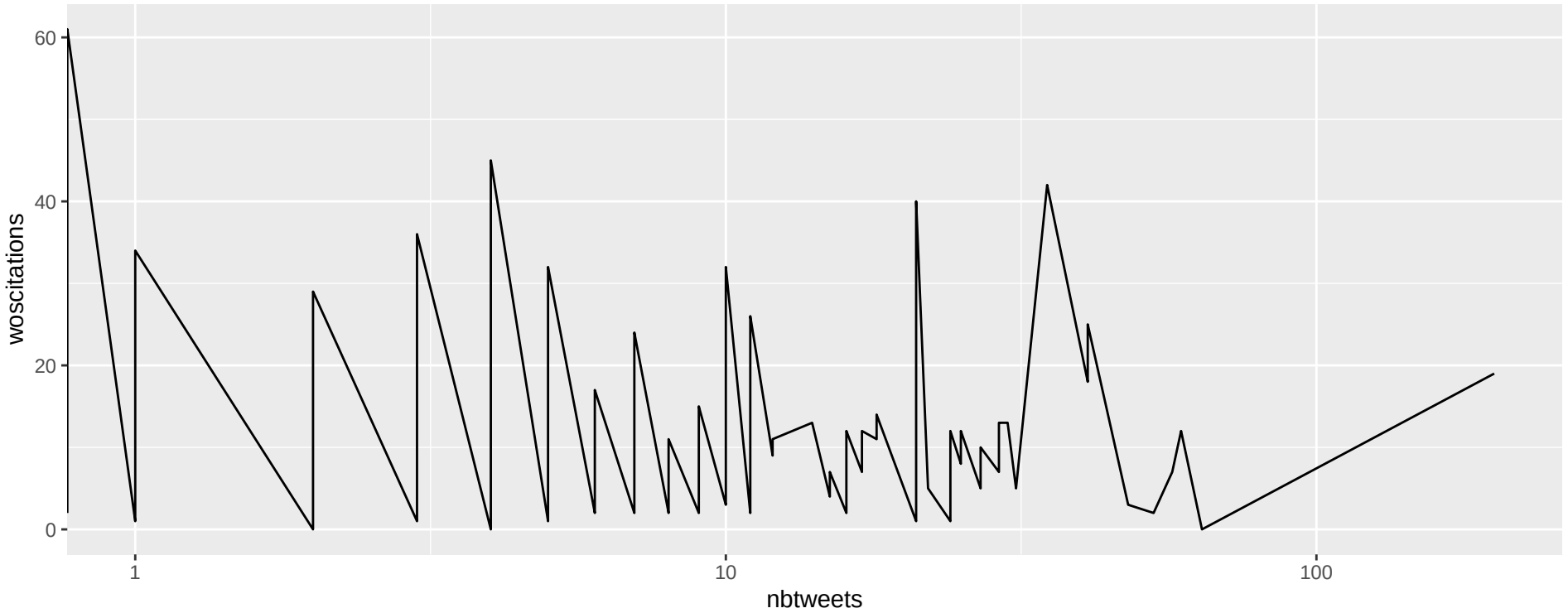
Scatterplots, lines instead of points

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets, y = woscitations) +
  geom_line() + #<<
  scale_x_log10()
```



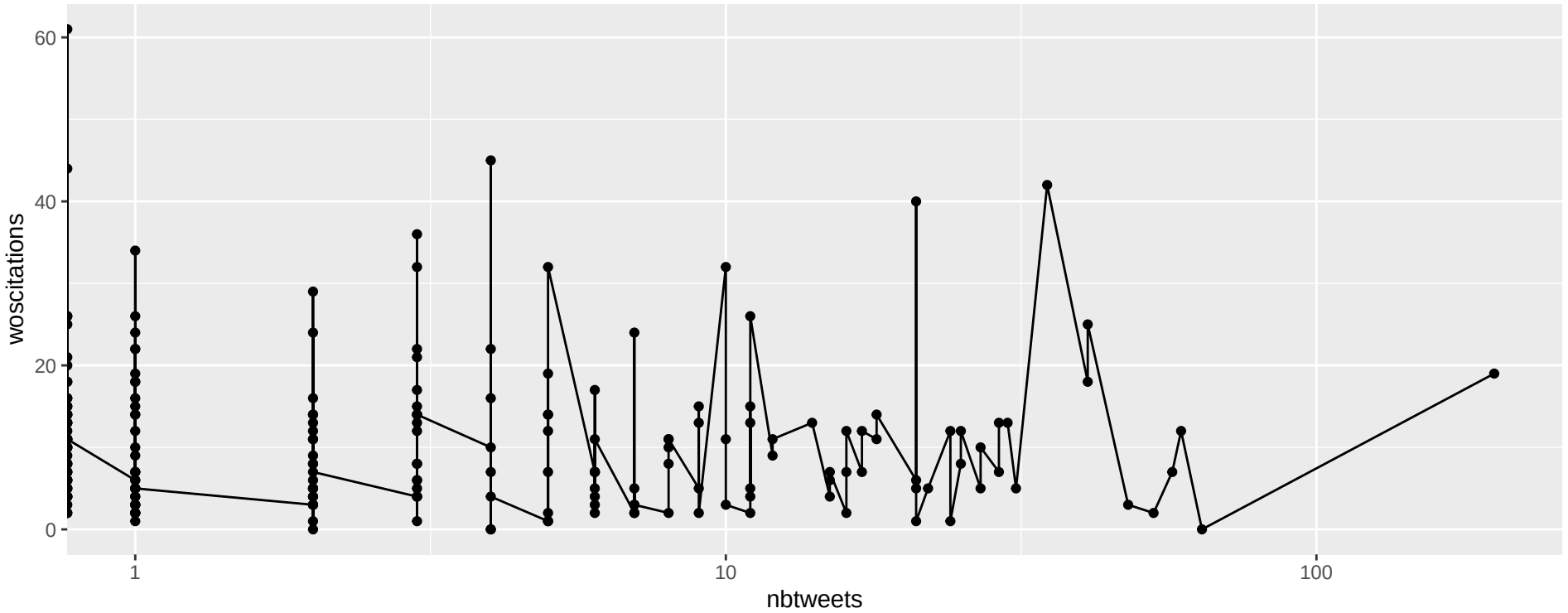
Scatterplots, lines with sorting beforehand

```
citations_ecology %>%
  arrange(woscitations) %>% #<<
  ggplot() +
  aes(x = nbtweets, y = woscitations) +
  geom_line() +
  scale_x_log10()
```



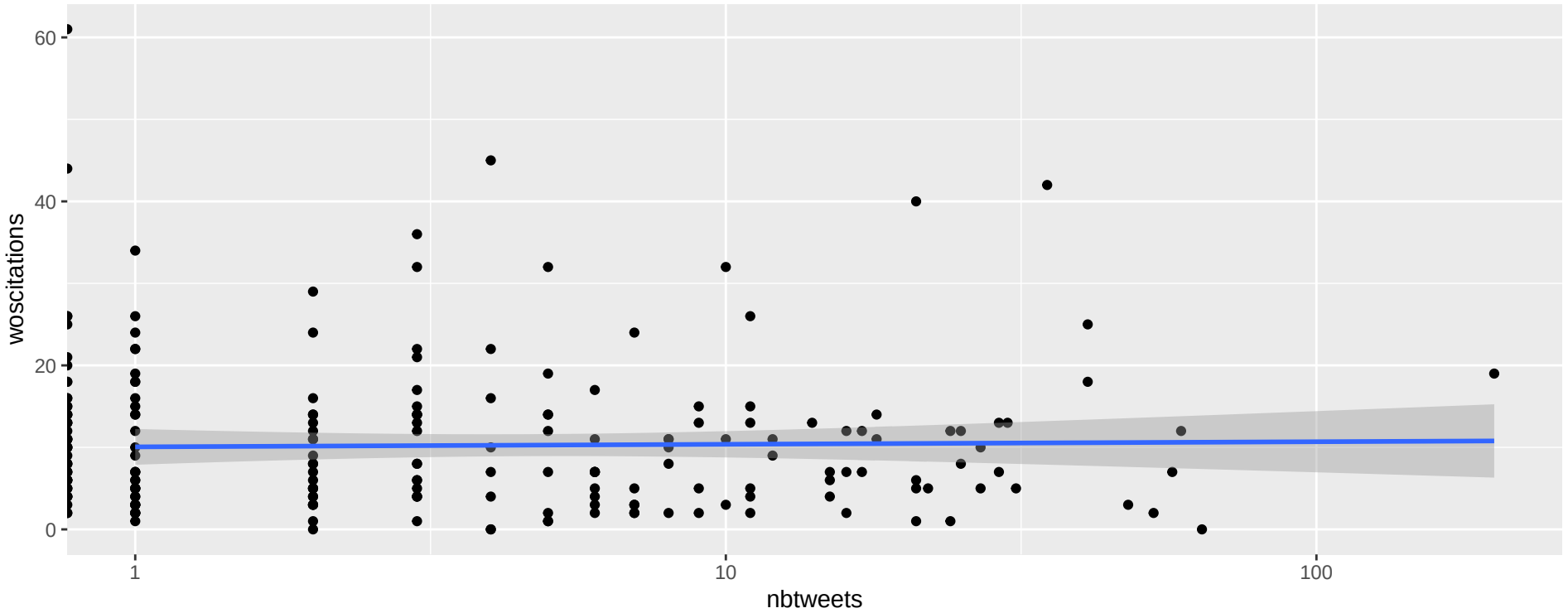
Scatterplots, add points

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets, y = woscitations) +
  geom_line() +
  geom_point() + #<<
  scale_x_log10()
```



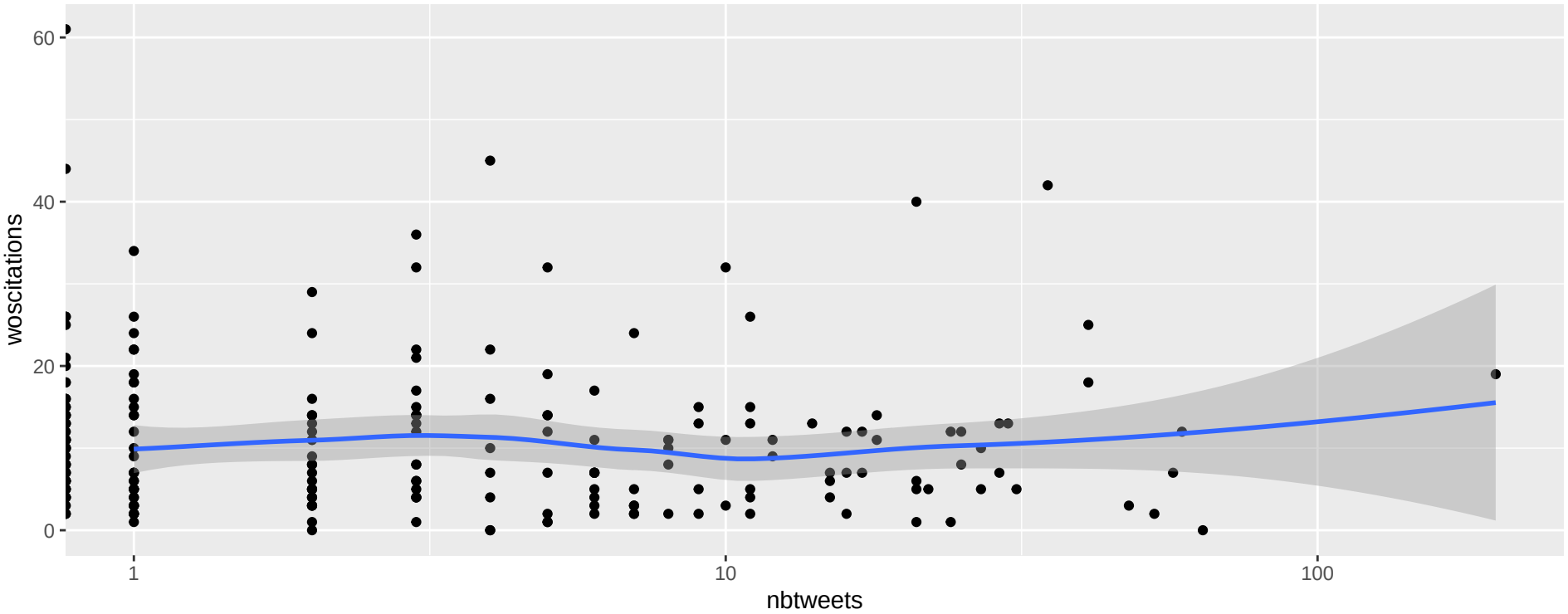
Scatterplots, add linear trend

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets, y = woscitations) +
  geom_point() +
  geom_smooth(method = "lm") + #<<
  scale_x_log10()
```



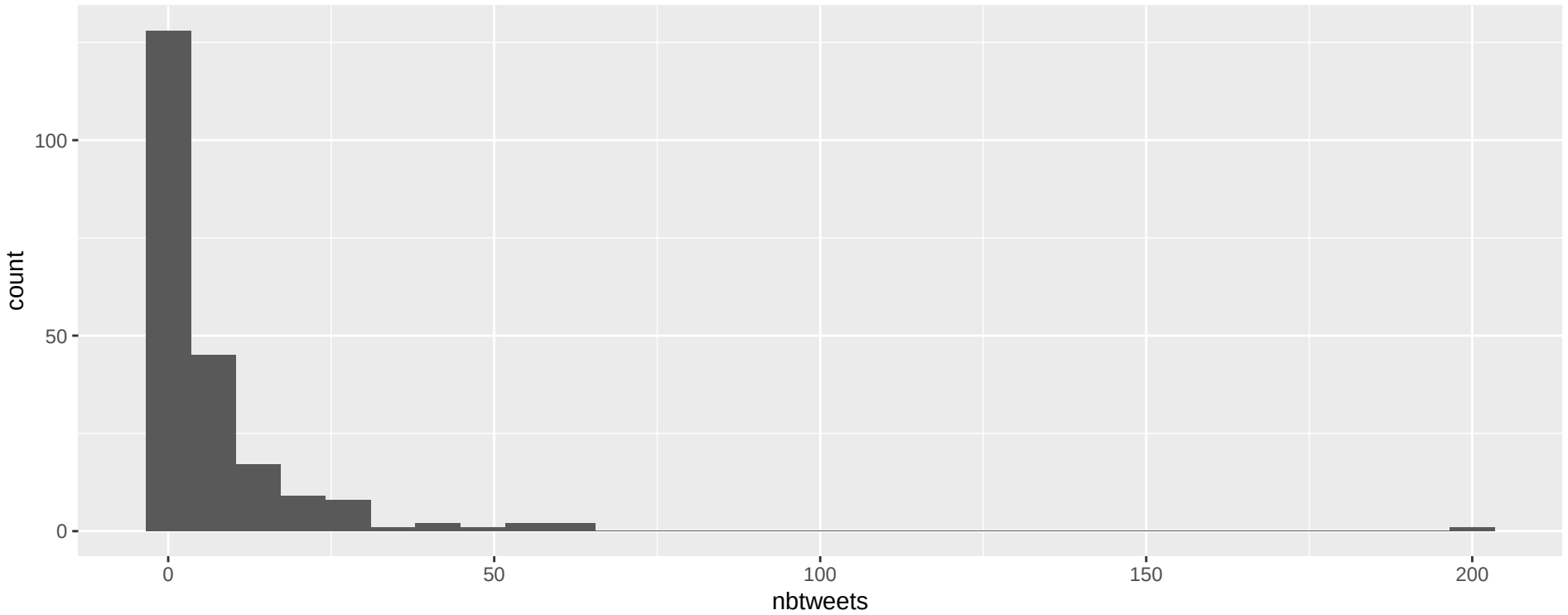
Scatterplots, add smoother

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets, y = woscitations) +
  geom_point() +
  geom_smooth() + #<<
  scale_x_log10()
```



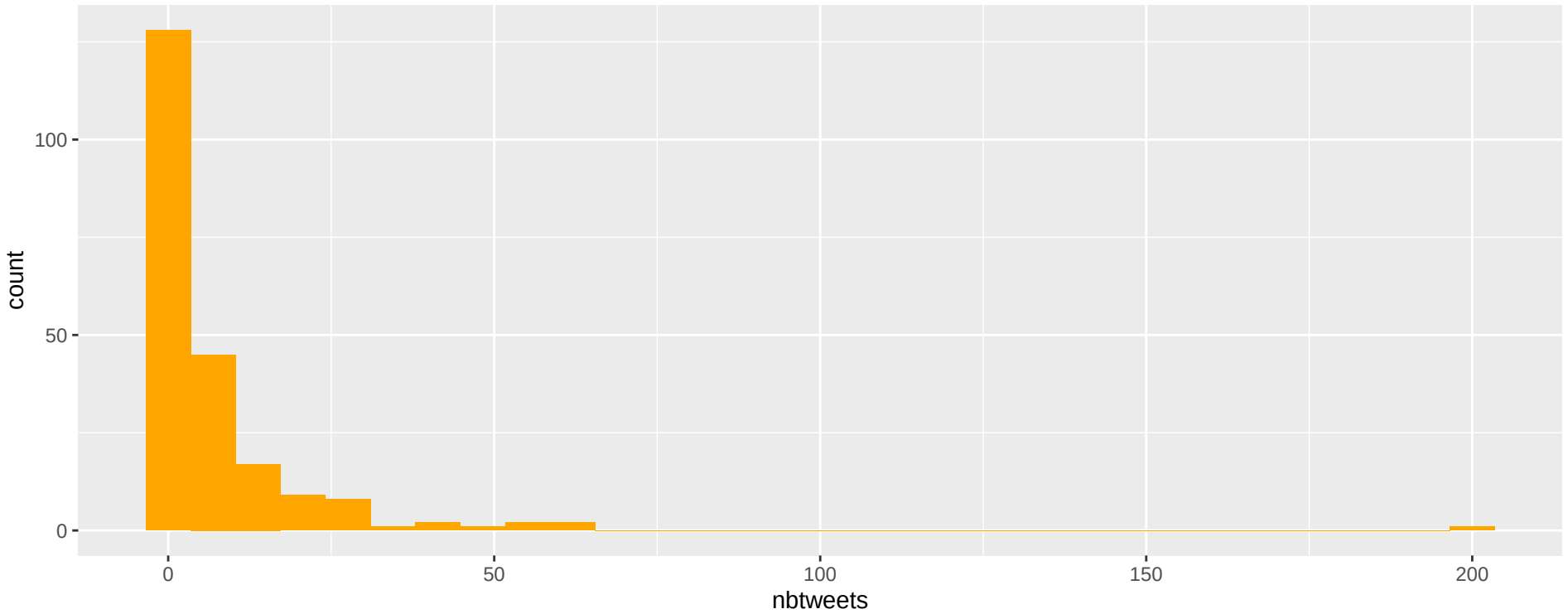
Histograms

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets) +
  geom_histogram() #<<
```



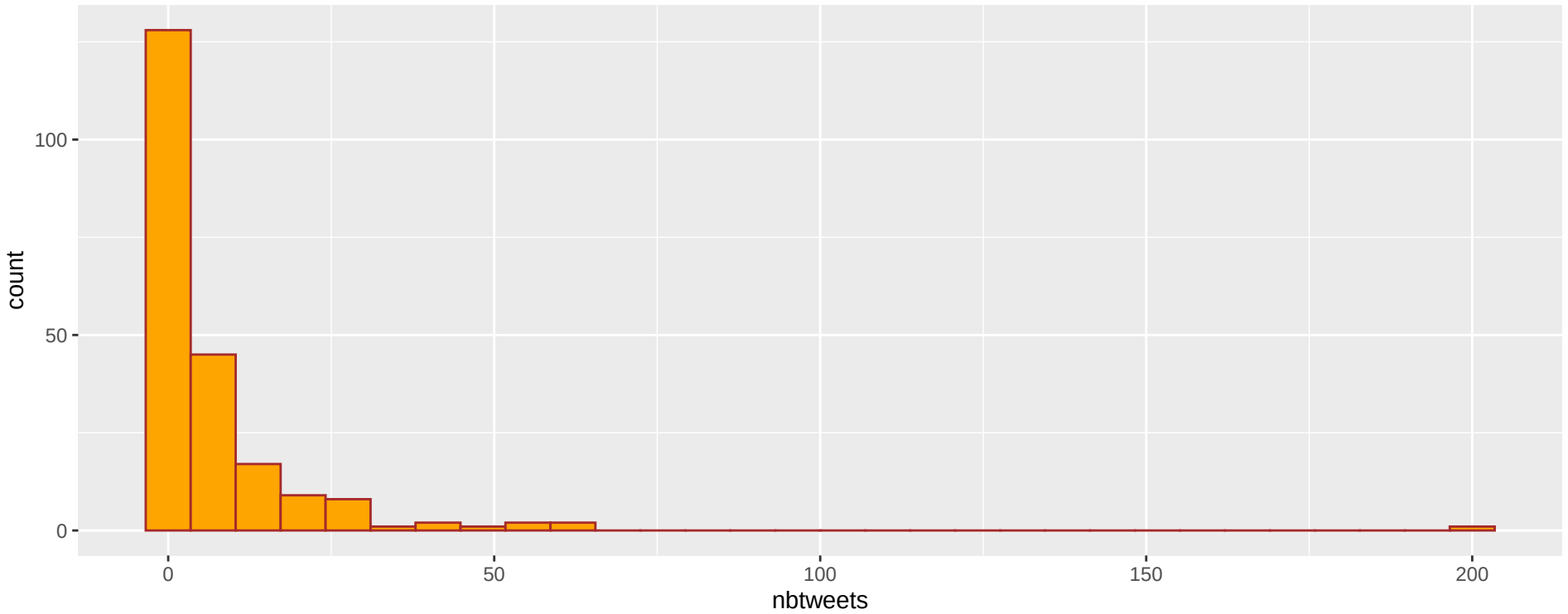
Histograms, with colors

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets) +
  geom_histogram(fill = "orange") #<<
```



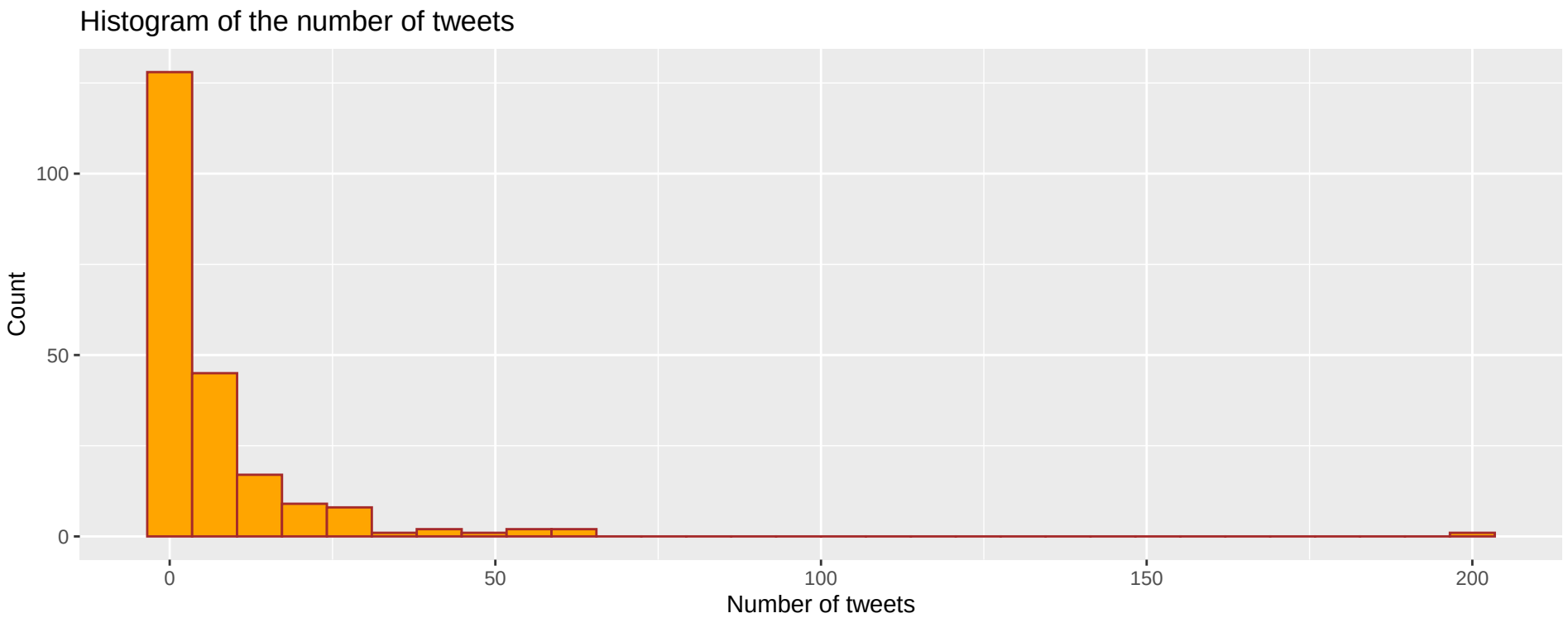
Histograms, with colors

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets) +
  geom_histogram(fill = "orange", color = "brown") #<<
```



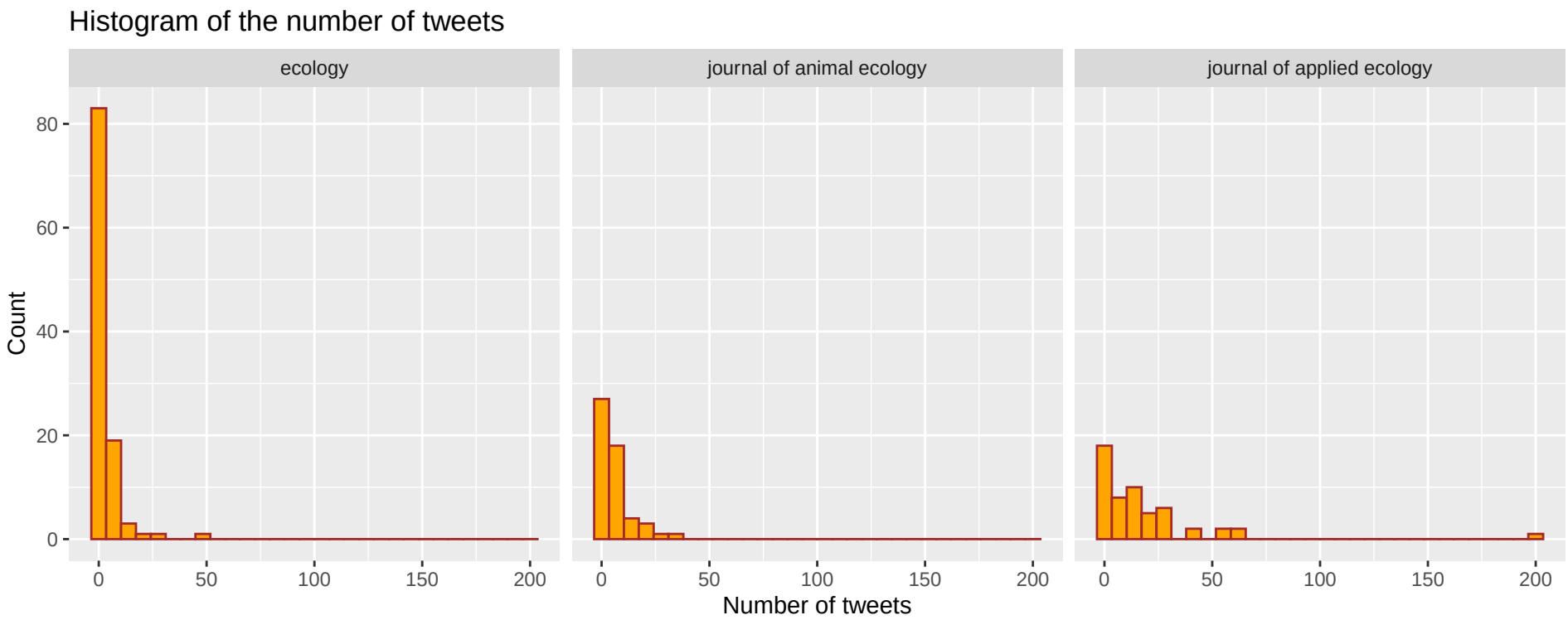
Histograms, with labels and title

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets) +
  geom_histogram(fill = "orange", color = "brown") +
  labs(x = "Number of tweets", #<<
       y = "Count", #<<
       title = "Histogram of the number of tweets") #<<
```



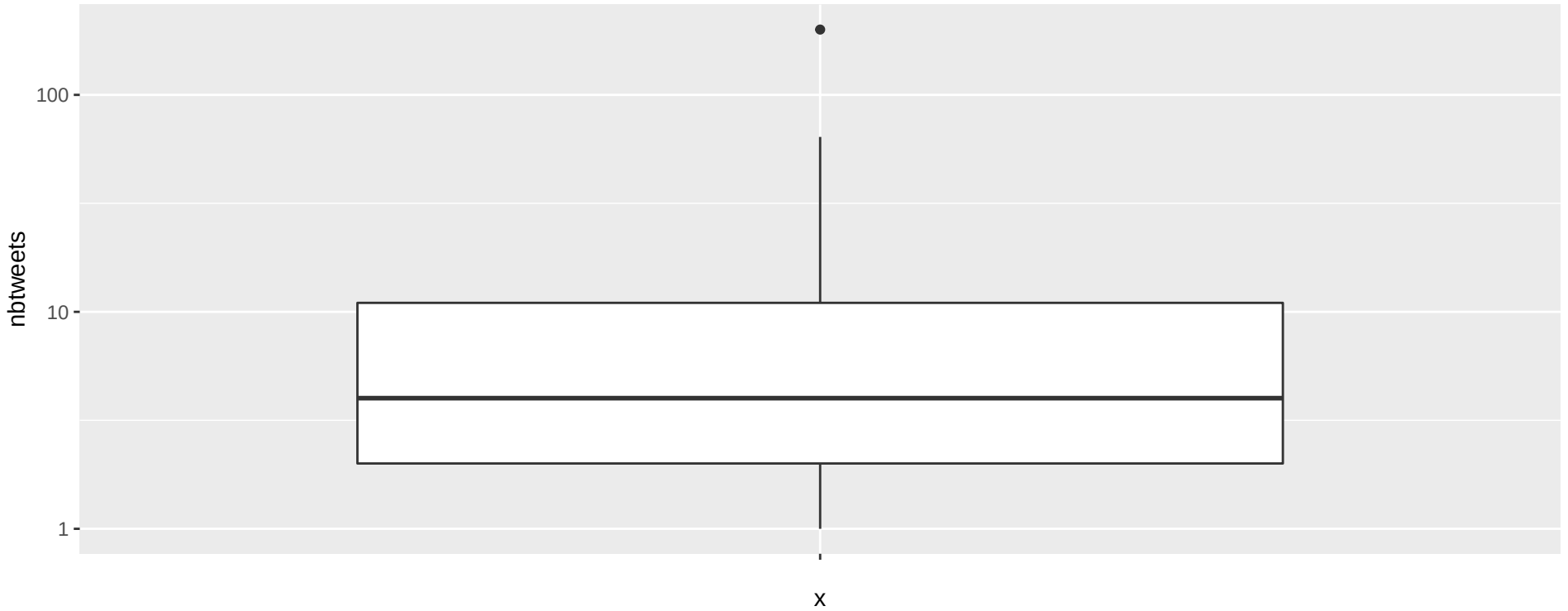
Histograms, by species

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets) +
  geom_histogram(fill = "orange", color = "brown") +
  labs(x = "Number of tweets",
       y = "Count",
       title = "Histogram of the number of tweets") +
  facet_wrap(vars(journal)) #<<
```



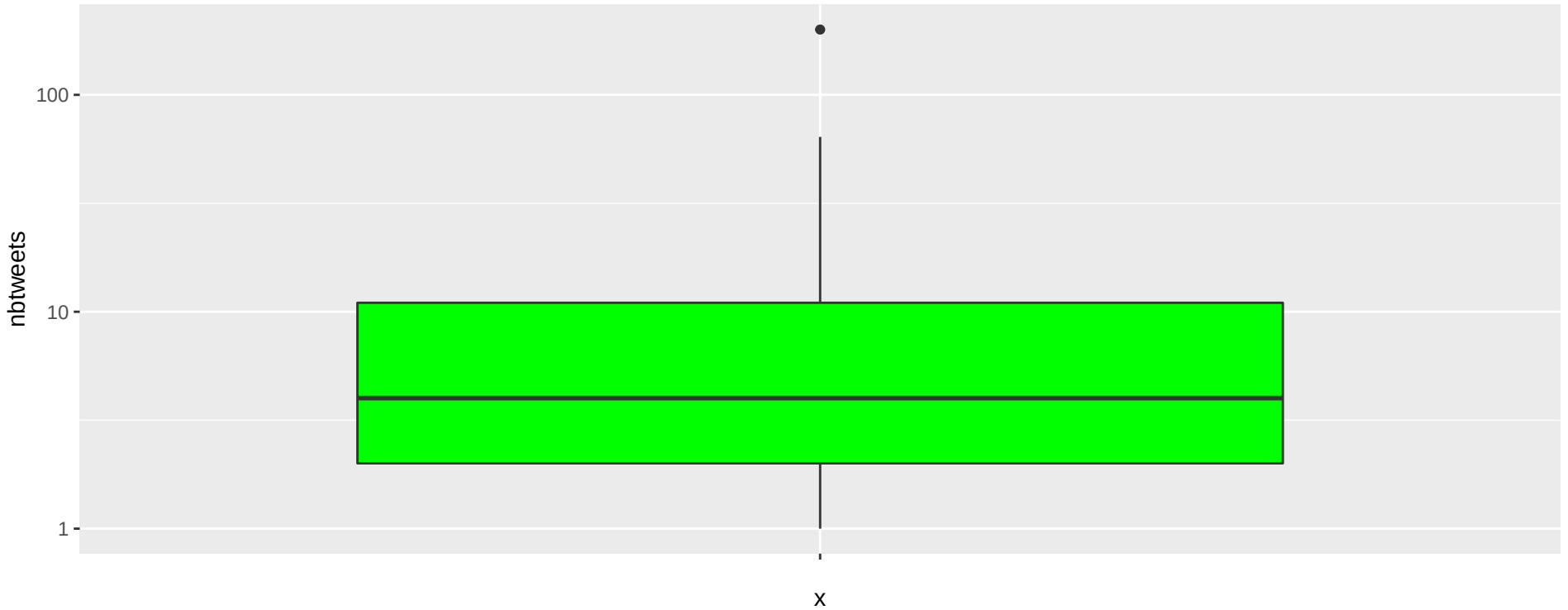
Boxplots

```
citations_ecology %>%
  ggplot() +
  aes(x = "", y = nbtweets) +
  geom_boxplot() + #<<
  scale_y_log10()
```



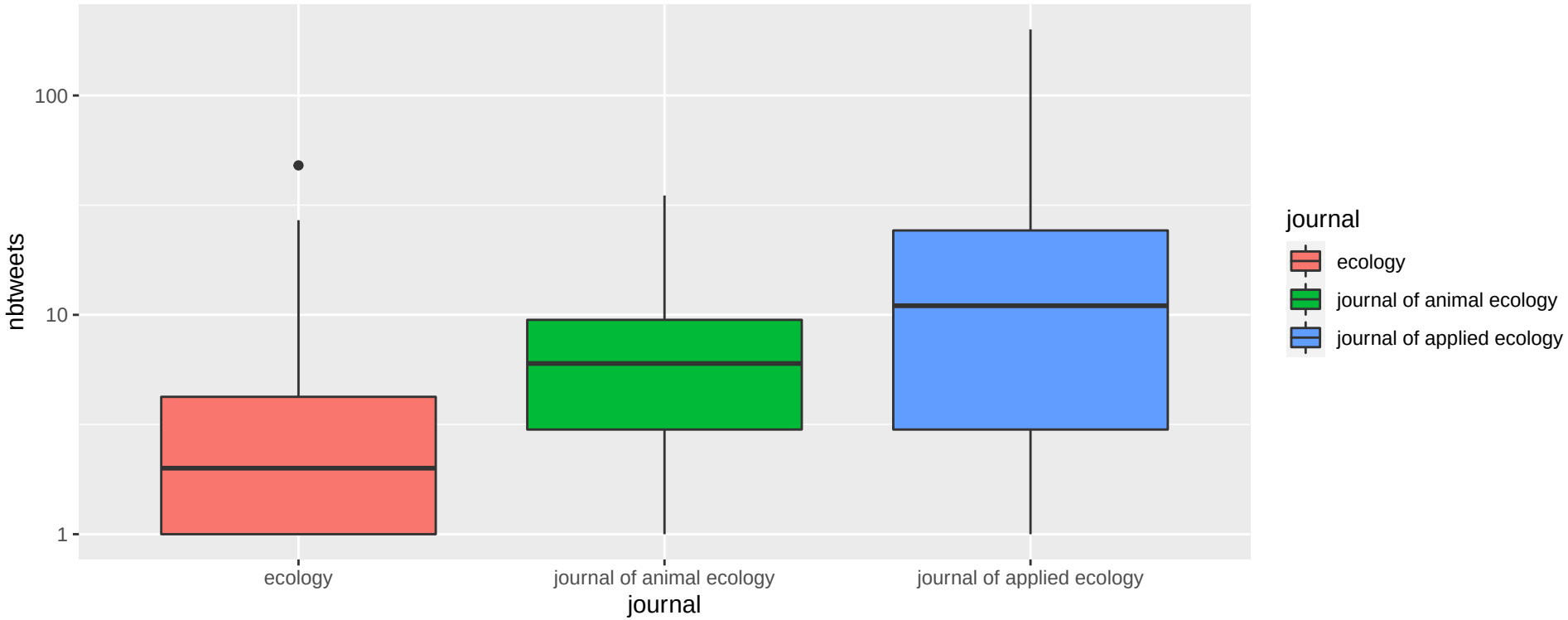
Boxplots with colors

```
citations_ecology %>%
  ggplot() +
  aes(x = "", y = nbtweets) +
  geom_boxplot(fill = "green") + #<<
  scale_y_log10()
```



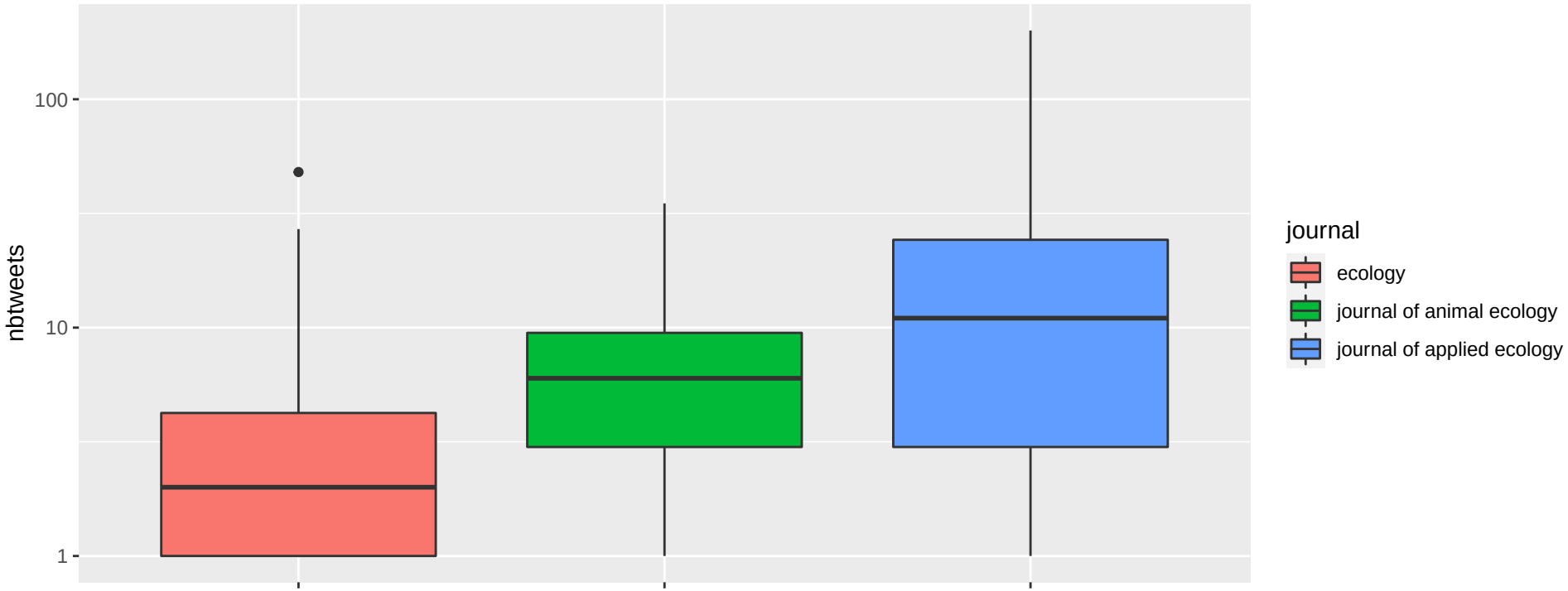
Boxplots with colors by species

```
citations_ecology %>%
  ggplot() +
  aes(x = journal, y = nbtweets, fill = journal) + #<<
  geom_boxplot() +
  scale_y_log10()
```



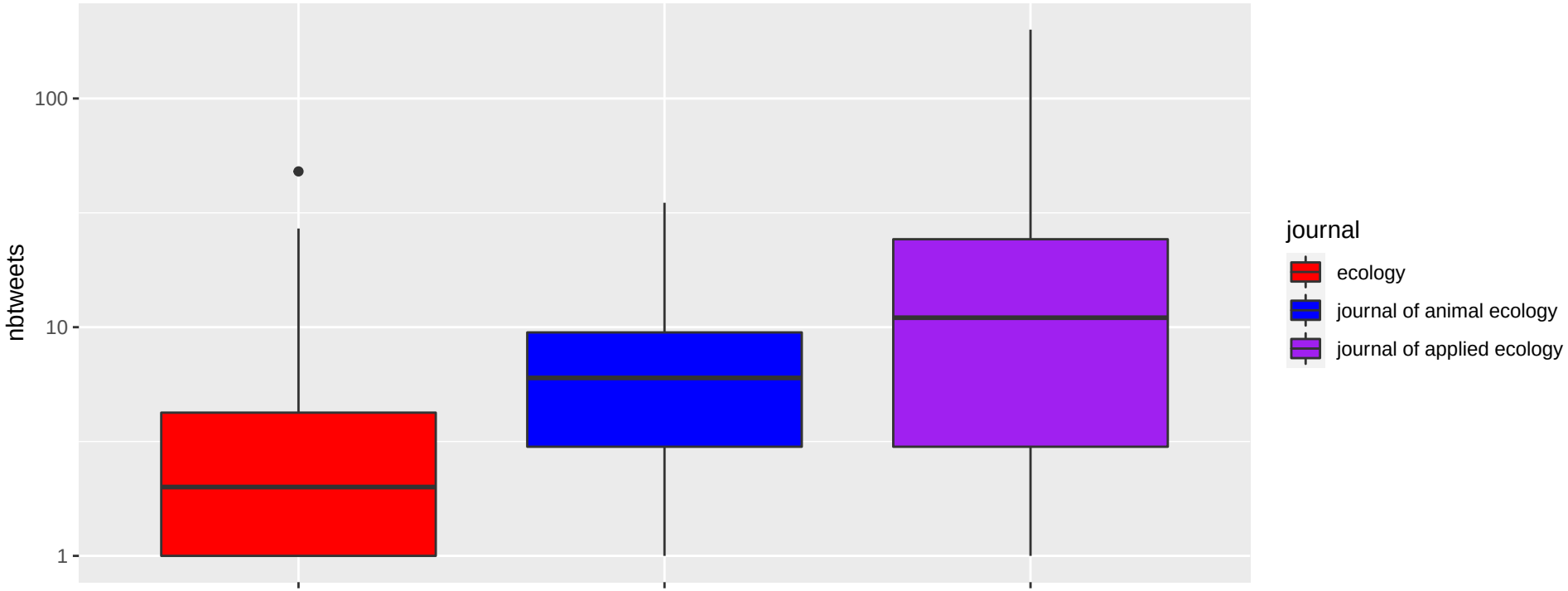
Get rid of the ticks on x axis

```
citations_ecology %>%
  ggplot() +
  aes(x = journal, y = nbtweets, fill = journal) +
  geom_boxplot() +
  scale_y_log10() +
  theme(axis.text.x = element_blank()) + #<<
  labs(x = "") #<<
```



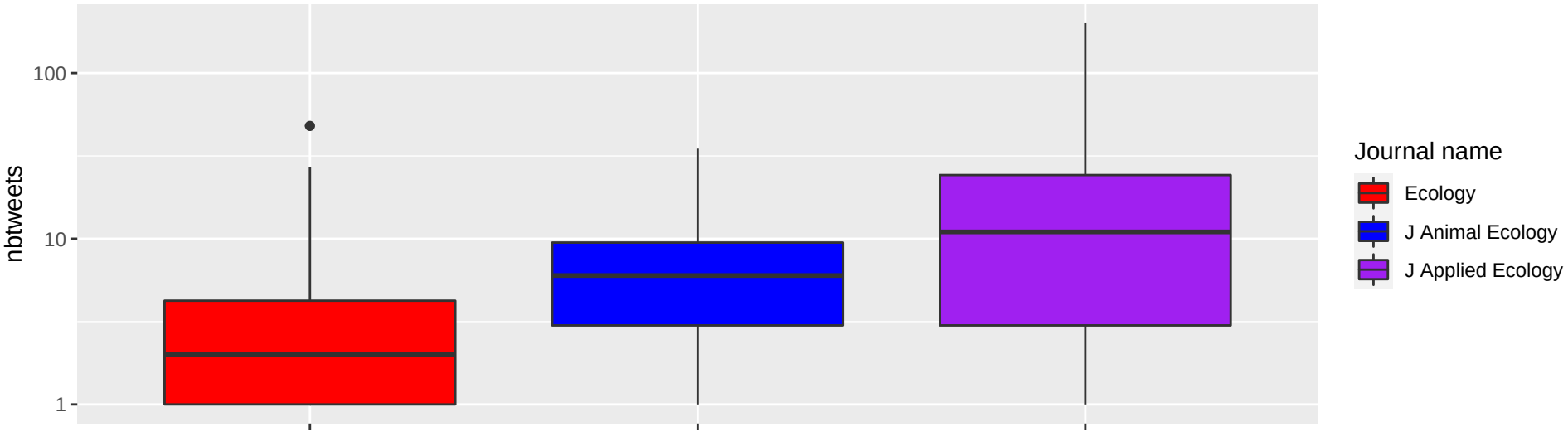
Boxplots, user-specified colors by species

```
citations_ecology %>%
  ggplot() +
  aes(x = journal, y = nbtweets, fill = journal) +
  geom_boxplot() +
  scale_y_log10() +
  scale_fill_manual( #<<
    values = c("red", "blue", "purple")) + #<<
  theme(axis.text.x = element_blank()) +
  labs(x = "")
```



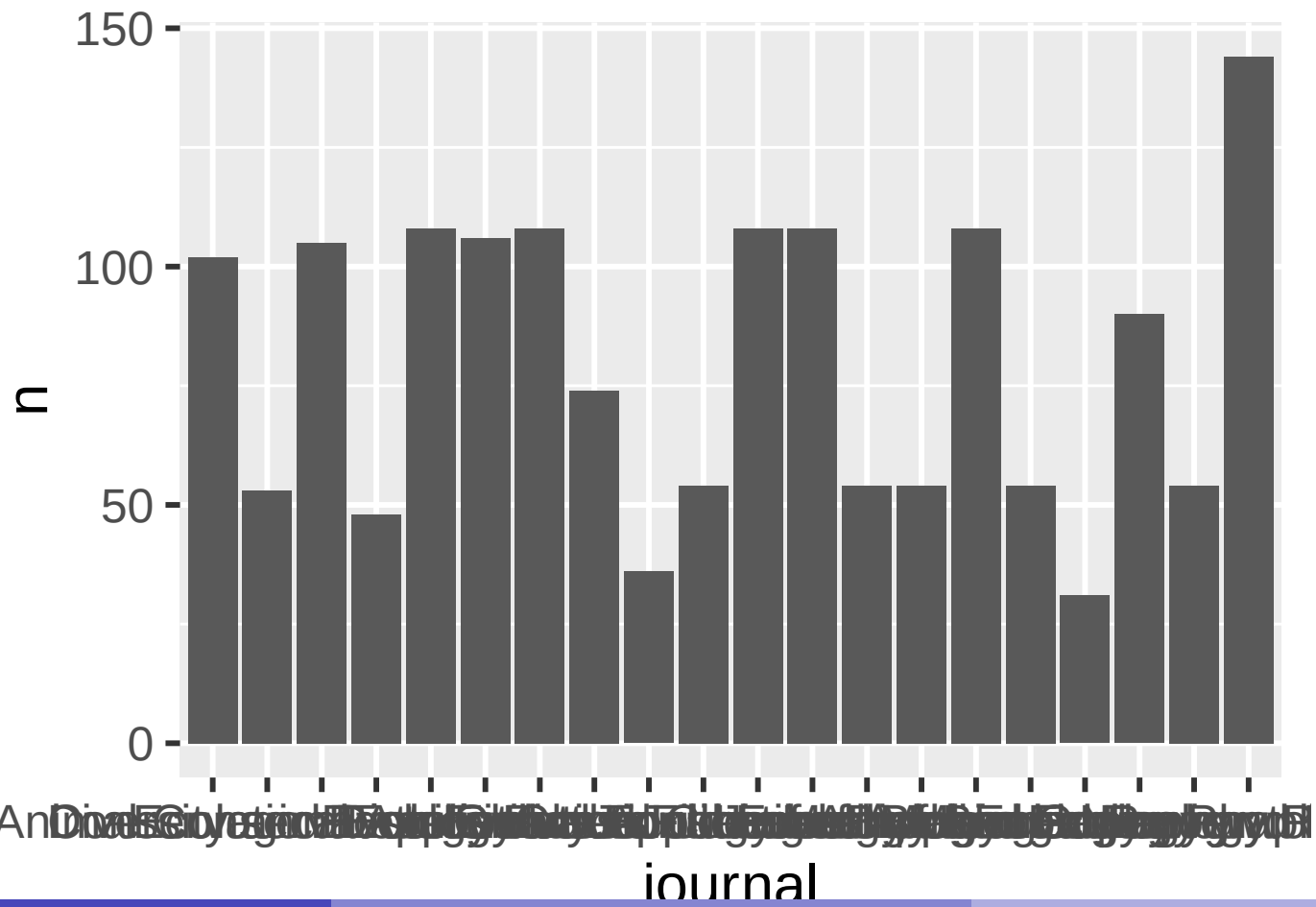
Boxplots, change legend settings

```
citations_ecology %>%
  ggplot() +
  aes(x = journal, y = nbtweets, fill = journal) +
  geom_boxplot() +
  scale_y_log10() +
  scale_fill_manual( #<<
    values = c("red", "blue", "purple"),
    name = "Journal name", #<<
    labels = c("Ecology", "J Animal Ecology", "J Applied Ecology")) + #<<
  theme(axis.text.x = element_blank()) +
  labs(x = "")
```



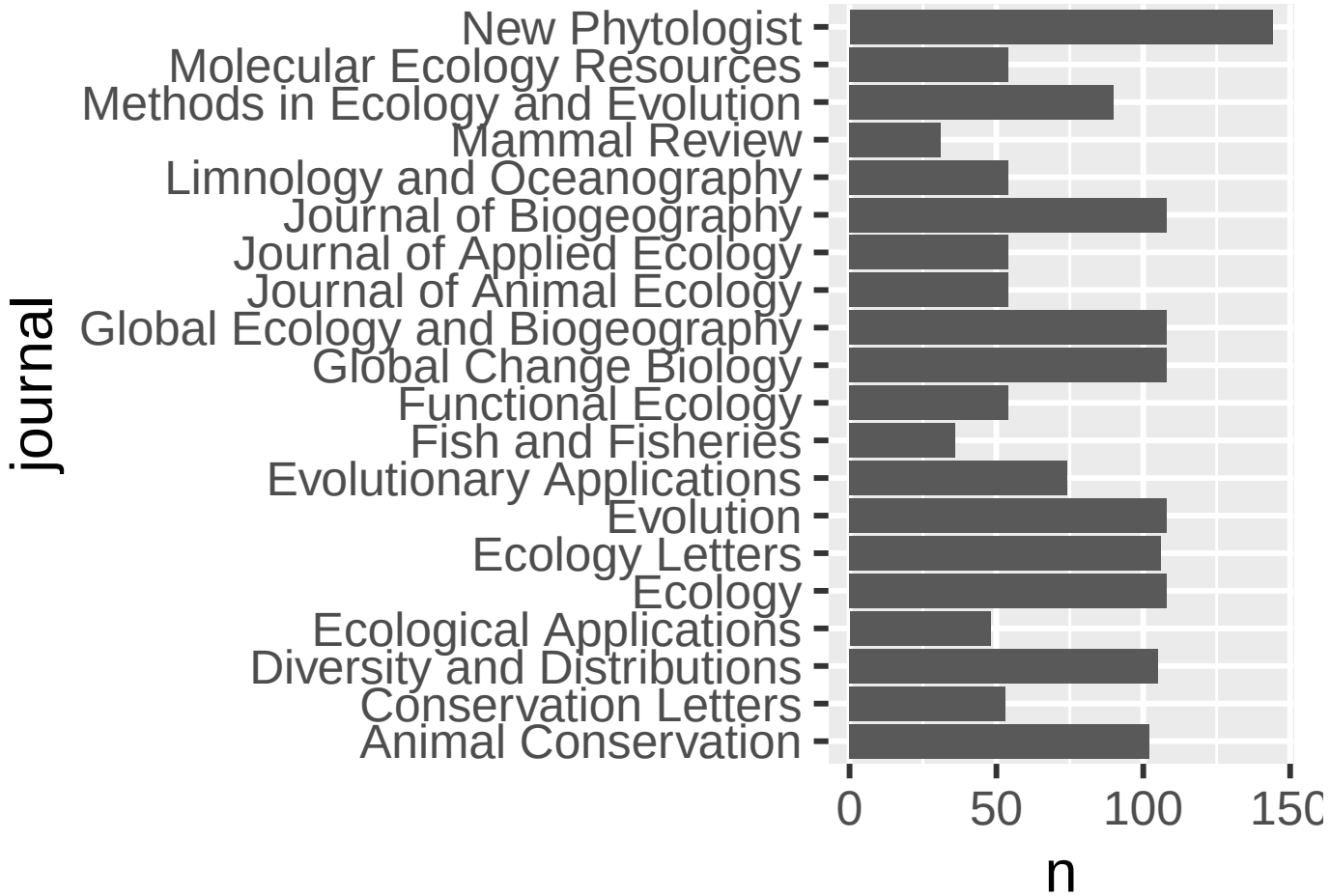
Ugly bar plots

```
citations %>%
  count(journal) %>%
  ggplot() +
  aes(x = journal, y = n) +
  geom_col() #<<
```



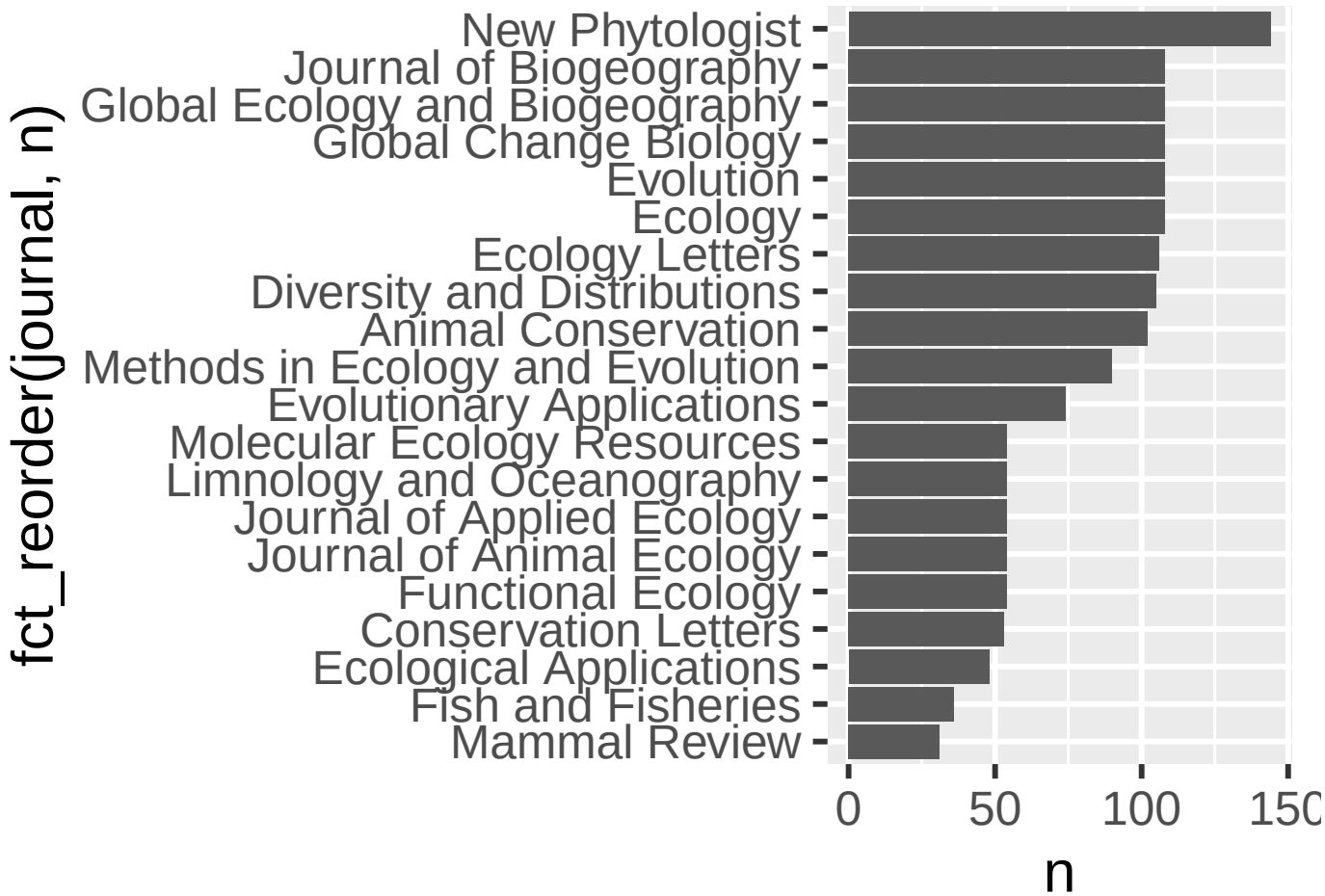
Bar plots, with flipping

```
citations %>%
  count(journal) %>%
  ggplot() +
  aes(x = n, y = journal) + #<<
  geom_col()
```



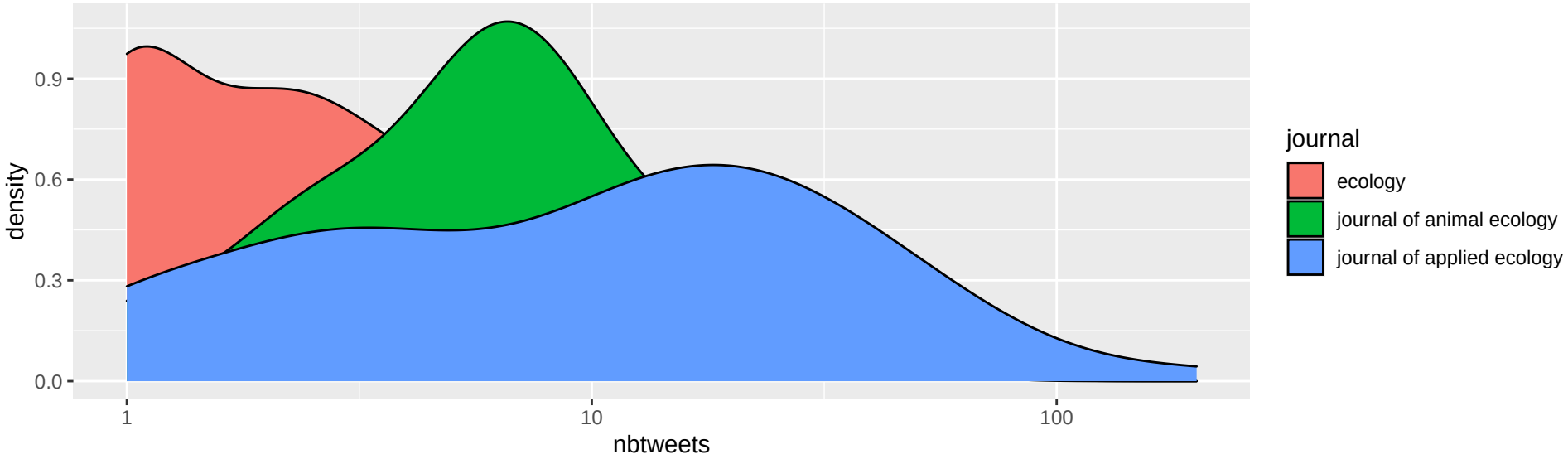
Bar plots, with factors reordering and flipping

```
citations %>%
  count(journal) %>%
  ggplot() +
  aes(x = n, y = fct_reorder(journal, n)) + #<<
  geom_col()
```



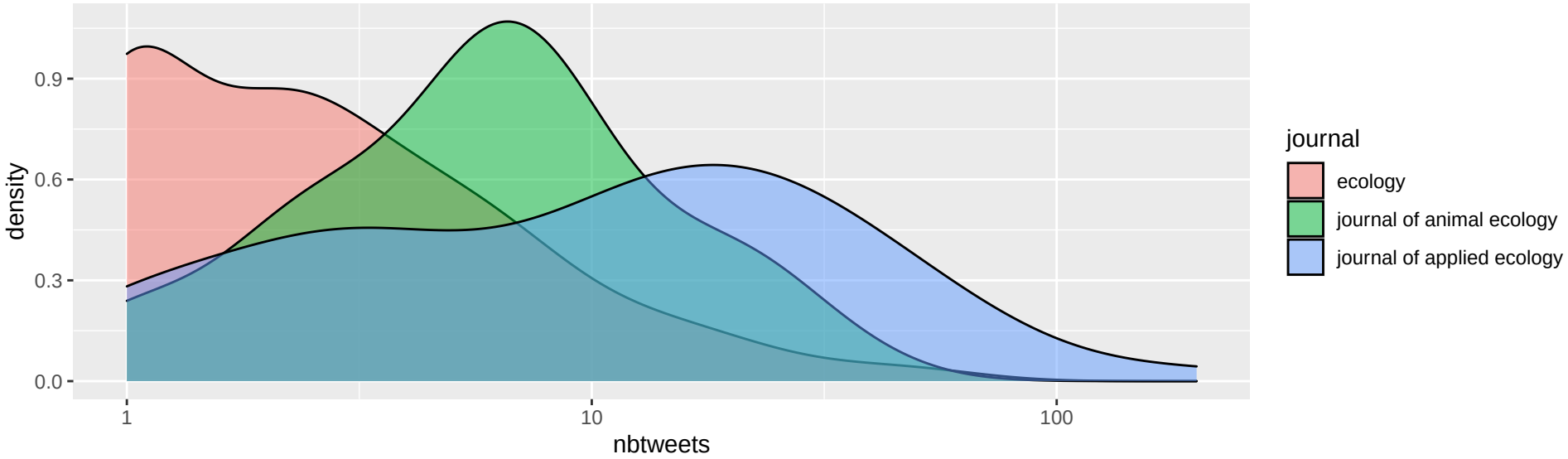
Density plots

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets, fill = journal) +
  geom_density() +#<<
  scale_x_log10()
```



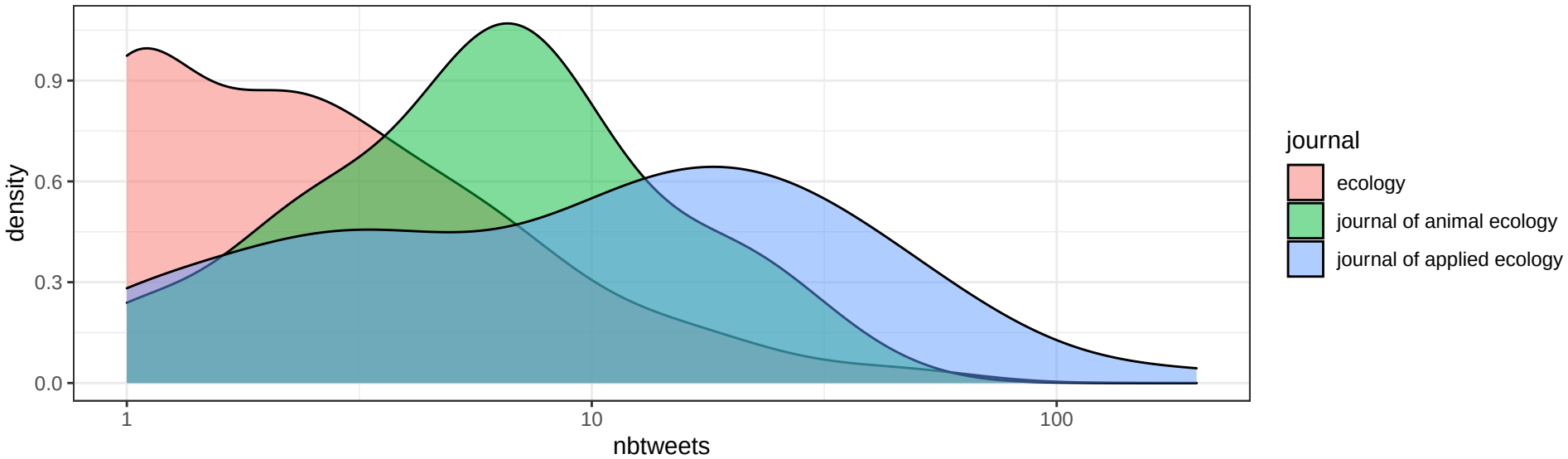
Density plots, control transparency

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets, fill = journal) +
  geom_density(alpha = 0.5) +#<<
  scale_x_log10()
```



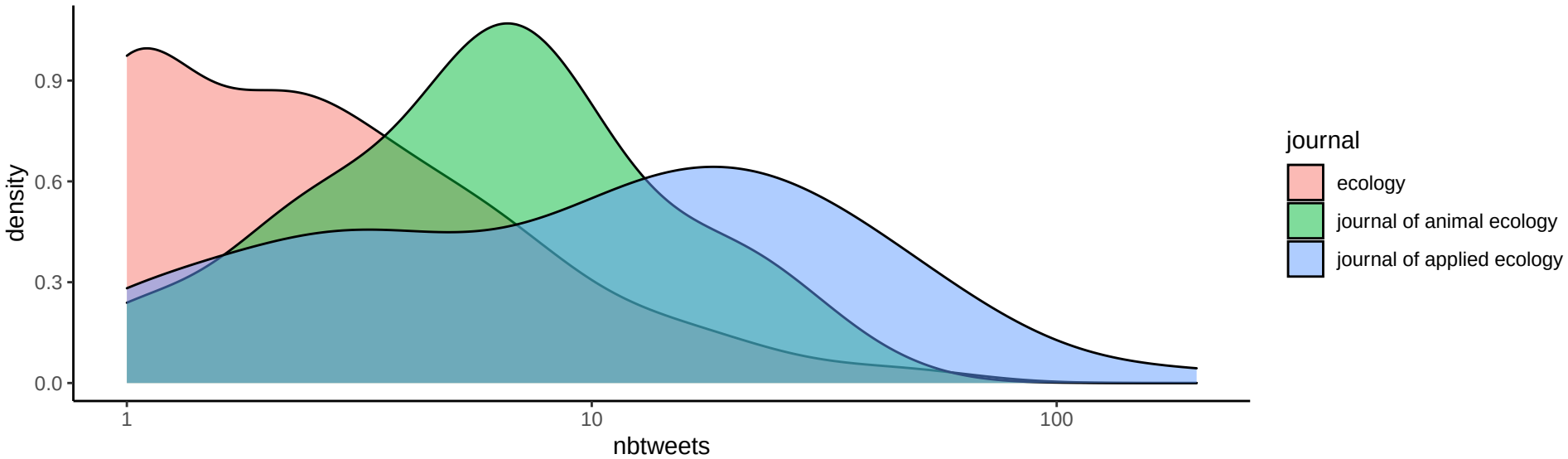
Change default background B & W theme

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets, fill = journal) +
  geom_density(alpha = 0.5) +
  scale_x_log10() +
  theme_bw() #<<
```



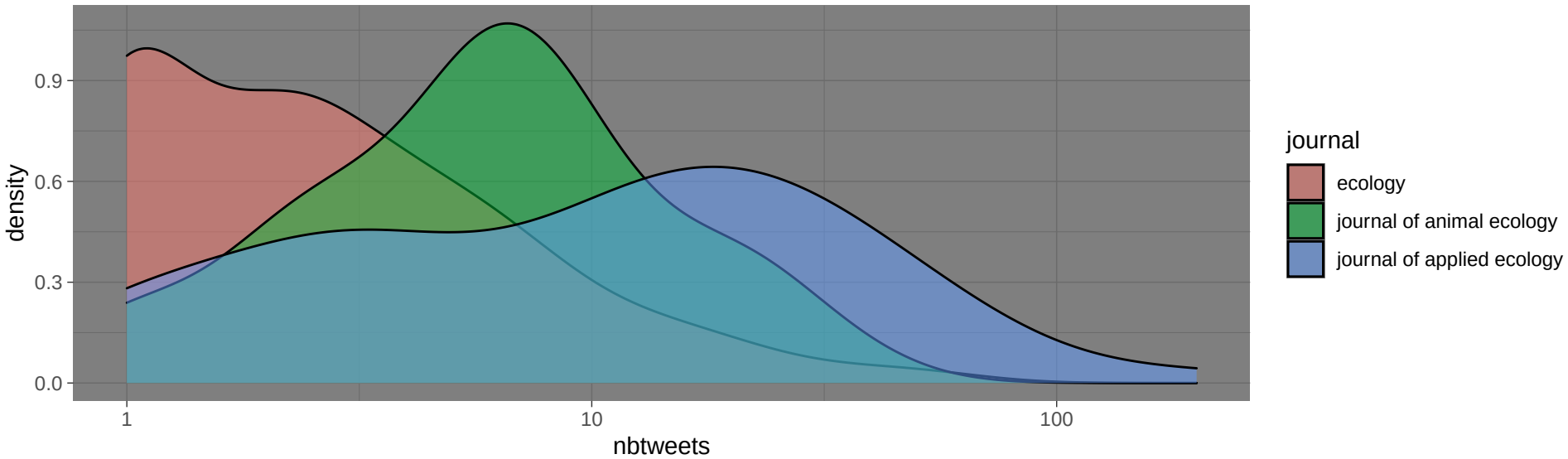
Change default background theme `classic theme`

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets, fill = journal) +
  geom_density(alpha = 0.5) +
  scale_x_log10() +
  theme_classic() #<<
```



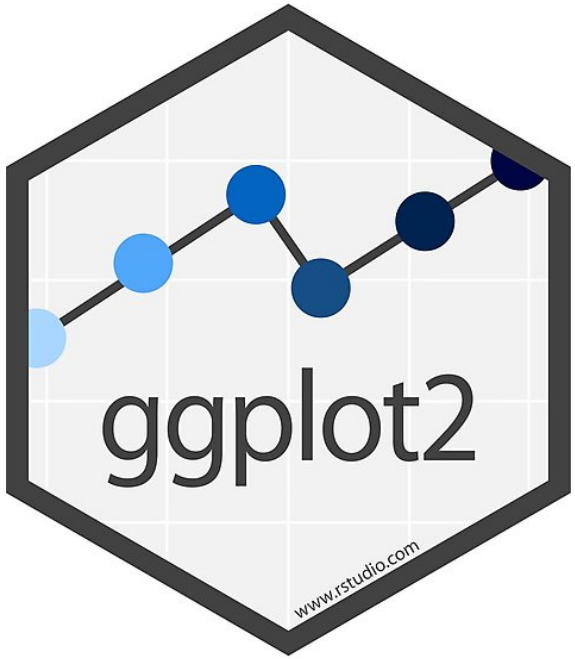
Change default background theme `dark theme`

```
citations_ecology %>%
  ggplot() +
  aes(x = nbtweets, fill = journal) +
  geom_density(alpha = 0.5) +
  scale_x_log10() +
  theme_dark() #<<
```



More on data visualisation with ggplot2

- Portfolio of ggplot2 plots
- Top ggplot2 visualizations
- Interactive ggplot2 visualizations



To dive even deeper in the tidyverse

- Learn the tidyverse: books, workshops and online courses
- Some books:
  - R for Data Science
  - Fundamentals of Data visualization
  - Data Visualization: A practical introduction
- Material of the 2-day workshop Data Science in the tidyverse held at the RStudio 2019 conference
- List of best R packages (with their description) on data import, wrangling and visualization
- How to switch from base R to tidyverse?